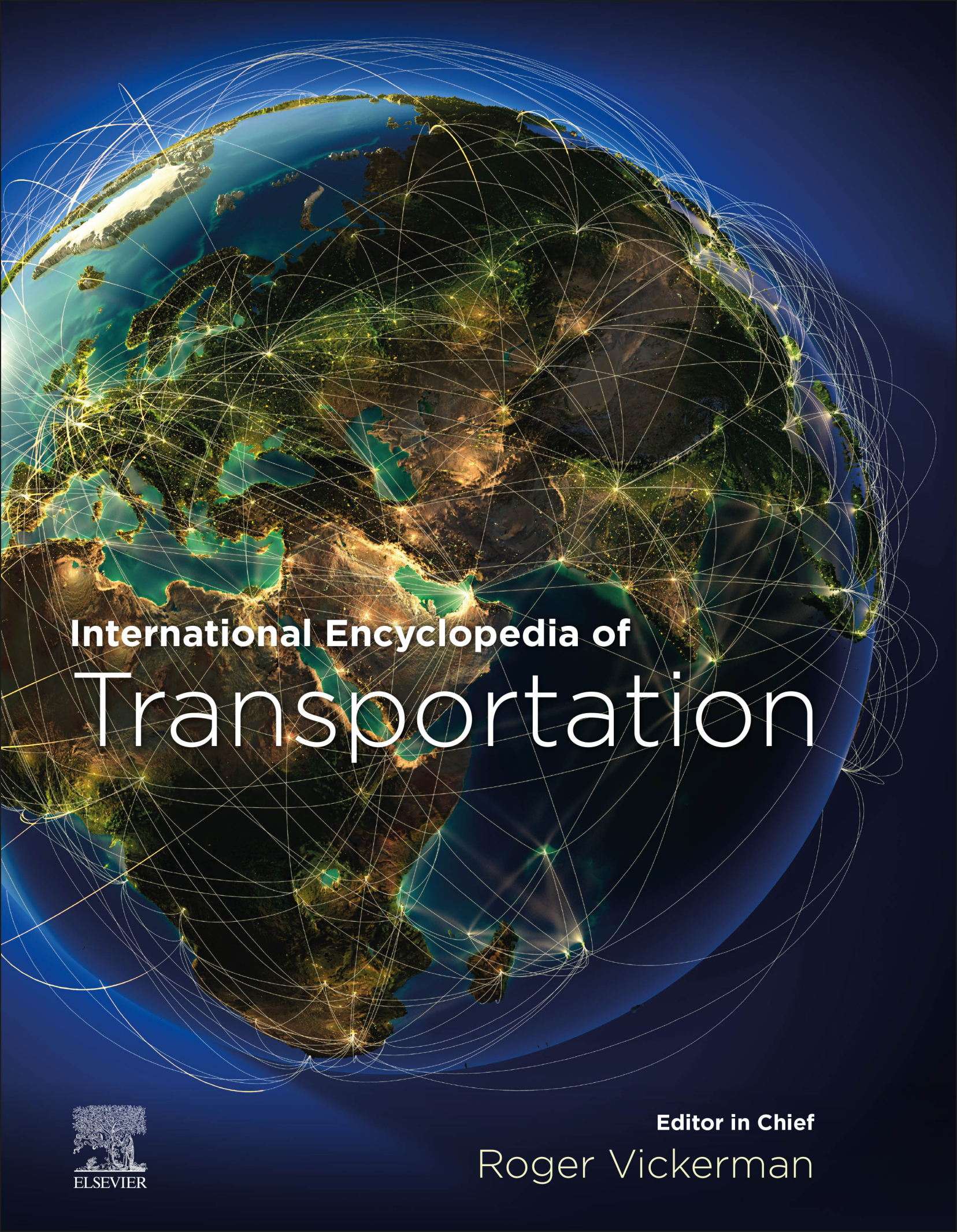




International Encyclopedia of Transportation



Editor in Chief
Roger Vickerman

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

Page left intentionally blank

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

EDITOR-IN-CHIEF

Roger Vickerman

*School of Economics, University of Kent, Canterbury, United Kingdom
and*

Transport Strategy Centre, Imperial College, London, United Kingdom

VOLUME 1

Transport Economics

SECTION EDITORS

Maria Börjesson

Professor of Economics

VTI Swedish National Road and Transport Research Institute

Affiliated Professor at Linköping University, Sweden



Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2021 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-08-102671-7

For information on all publications visit our
website at <http://store.elsevier.com>

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

Publisher: Oliver Walter
Acquisitions Editor: Oliver Walter
Content Project Manager: Natalie Lovell
Associate Content Project Manager: Manisha K and Ramalakshmi Boobalan
Designer: Matthew Limbert

EDITORIAL BOARD

Editor in Chief

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Section Editors

Maria Börjesson

Professor of Economics, VTI Swedish National Road and Transport Research Institute; Affiliated Professor at Linköping University, Sweden

Per Gårder

Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States

Prof. Kevin P.B. Cullinane

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

Prof. Edward C.S. Chung

Department of Electrical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

Chandra R. Bhat

Center for Transportation Research (CTR), The University of Texas at Austin, TX, United States

Edoardo Marcucci

Department of Political Sciences, University of Roma Tre, Rome, Italy; Department of Logistics, Molde University College, Molde, Norway

Prof. Maria Attard

Department of Geography, Faculty of Arts, University of Malta, Msida, Malta

Prof. Carlo G. Prato

School of Civil Engineering, The University of Queensland, Brisbane, Australia

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Page left intentionally blank

INTRODUCTION



Roger Vickerman

In an increasingly globalized world, despite reductions in costs and time, transportation has become even more important as a facilitator of economic and human interaction; this is reflected in technical advances in transportation systems, increasing interest in how transportation interacts with society and the need to provide novel approaches to understand its impacts. This has become particularly acute with the impact that Covid-19 has had on transportation across the world, at local, national, and international levels. This Encyclopedia brings a cross-cutting and integrated approach to all aspects of transportation from work in many disciplinary fields, engineering, operations research, economics, geography, and sociology to understand the changes taking place.

Transportation is both influenced by, and an influencer of, changes in the economy and society. Increasing speeds have reduced journey times and made the world a smaller place as globalization has affected both where people live and work and from where they source their goods and materials. Increasing volumes of traffic, often on old and life-expired infrastructure, lead to congestion and delays. Constraints on public budgets have led to increasing pressure on the private sector to fund improvements requiring new and innovative financial solutions. While there are clear differences in the nature of the pressures felt in the developed and less developed economies, there is an increasing recognition that in all societies, there is an accessibility problem such that certain groups become disadvantaged by the lack of access to reliable and cost-effective transport. While the problems are clearly multidimensional, research on transportation is often constrained by single disciplinary approaches and this carries over into the practice of transport planning and policy. The Encyclopedia cuts across these artificial boundaries by taking an approach that emphasizes the interaction between the different aspects of research and aims to offer new solutions to understand these problems. Each of the nine sections is based around a familiar dimension of work on transportation, but brings together the views of experts from different disciplinary perspectives. Each section is edited by an expert in the relevant field who has sought chapters from a range of authors representing different disciplines, different parts of the world, and different social perspectives. In this way, the work is not just a reflection of the state of the art that serves as a starting point for researchers and practitioners, but also a pointer toward new approaches, new ways of thinking, and novel solutions to problems.

The nine sections are structured around the following themes: Transport Modes; Freight Transport and Logistics; Transport Safety and Security; Transport Economics; Traffic Management; Transport Modeling and Data Management; Transport Policy and Planning; Transport Psychology; Sustainability and Health Issues in Transportation. Some of the chapters provide a technical introduction to a topic while others provide a bridge between topics or a more future-oriented view of new research areas or challenges. While there is guidance to cross-referencing between chapters, readers are encouraged to explore the tables of contents of all the sections to get a full understanding of the issues. Much of the Encyclopedia was completed before the Covid-19 pandemic and clearly this will have changed the situation in many areas covered by this work; the advantage of this type of reference work is that relevant updates will be possible in future editions.

The Encyclopedia has only been possible because of the cooperation of a large number of people. Robert Noland, Georgina Santos, Xiaowen Fu, and Dick Ettema served as an Editorial Advisory Board identifying possible editors of sections and advising on the overall structure. The Section Editors, Edoardo Marcucci, Kevin Cullinane, Per Garder, Maria Börjesson, Edward Chung, Chandra Bhat, Maria Attard, and Carlo Prato,

carried out the work of identifying potential authors of individual chapters, commissioning these, encouraging authors and reviewing drafts. More than 600 authors and co-authors of chapters are, however, are ultimately responsible for the success of this venture. Thanks are also due to the key people at Elsevier, particularly the Publishers, Andre Wolff and Oliver Walter, and the Project Managers, Sophie Harrison and Natalie Bentahar; they have shown exemplary patience over more than 3 years in bringing this work to fruition.

LIST OF CONTRIBUTORS TO VOLUME 1

José Holguín-Veras

*Department of Civil and Environmental Engineering;
Center for Infrastructure, Transportation, and the
Environment; VREF Center of Excellence for Sustainable
Urban Freight Systems, Rensselaer Polytechnic Institute,
Troy, NY, United States*

Diana G. Ramírez-Ríos

*Department of Civil and Environmental Engineering,
Rensselaer Polytechnic Institute, Troy, NY, United States*

Kathrin Goldmann

*University of Münster, Institute of Transport Economics,
Münster, Germany*

Gernot Sieg

*University of Münster, Institute of Transport Economics,
Münster, Germany*

José Manuel Vassallo

*Transport Research Centre (TRANSyT), Universidad
Politécnica de Madrid, Madrid, Spain;
Centro de Investigación del Transporte (TRANSyT), ETSI de
Ingenieros de Caminos, Canales y Puertos, Madrid, Spain*

Russell G. Thompson

The University of Melbourne, Melbourne, VIC, Australia

Marco Batarce

*Faculty of Economics and Business, Universidad San
Sebastián, Santiago, Chile*

André de Palma

*CREST, ENS Paris-Saclay, University of Paris-Saclay,
Paris, France*

Julien Monardo

*CREST, ENS Paris-Saclay, University of Paris-Saclay,
Paris, France*

Yulai Wan

*Hong Kong Polytechnic University, Hong Kong, China
Dong Yang*

Ricardo Giesen

*Department of Transport Engineering and Logistics,
Pontificia Universidad Católica de Chile, Santiago, Chile*

Darío Farren

*Department of Transport Engineering and Logistics,
Pontificia Universidad Católica de Chile, Santiago,
Chile*

Sofia F. Franco

*Department of Economics, University of California-Irvine,
Irvine, CA, United States*

John J. Bates

*Independent Consultant in Transport Economics,
Abingdon, Oxfordshire, United Kingdom*

Svante Mandell

*Swedish National Institute of Economic Research,
Stockholm, Sweden*

Katrine Hjorth

*Technical University of Denmark, Kongens Lyngby,
Denmark*

Daniel Hörcher

*Imperial College London, London, United Kingdom
Budapest University of Technology and Economics,
Budapest, Hungary*

Nathalie Picard

*Université de Strasbourg, Université de Lorraine, CNRS,
BETA, Strasbourg, France*

Bruno De Borger

University of Antwerp, Antwerp, Belgium

Stef Proost

KU Leuven, Leuven, Belgium

Stefanie Peer

*Vienna University of Economics & Business, Vienna,
Austria*

Jon P. Nelson

*Pennsylvania State University, State College, PA,
United States*

Henrik Andersson

*Toulouse School of Economics, University of Toulouse
Capitole, Toulouse, France*

Raquel Espino

Department of Applied Economic Analysis, Instituto Universitario de Desarrollo Económico Sostenible y Turismo, Universidad de Las Palmas de Gran Canaria (ULPGC), Las Palmas, Spain

Juan de Dios Ortúzar

Department of Transport Engineering and Logistics, Instituto Sistemas Complejos de Ingeniería (ISCI), Pontificia Universidad Católica de Chile, Santiago, Chile

Luis I. Rizzi

Department of Transport Engineering and Logistics, Instituto Sistemas Complejos de Ingeniería (ISCI), Pontificia Universidad Católica de Chile, Santiago, Chile

Kenneth A Small

1721 W. 104th Place, Chicago, IL, United States

Robin Lindsey

Sauder School of Business, University of British Columbia, Vancouver, BC, United Kingdom

Charles Raux

University of Lyon, CNRS, LAET, Lyon, France

Jonas Eliasson

Department of Science and Technology, Division of Communications and Transport Systems, Linköping University, Norrköping, Sweden

Ginés de Rus

*University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
University Carlos III de Madrid, Madrid, Spain
FEDEA, Madrid, Spain*

Dereje Abegaz

Department of Technology, Management and Economics, Technical University of Denmark, Lyngby, Denmark

Yili Tang

California PATH, University of California, Berkeley, California

Daniel Albalade

University of Barcelona (GiM-IREA), Barcelona, Spain

Albert Gragera

Technical University of Denmark, Copenhagen, Denmark

Andrew Daly

Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

James Fox

RAND Europe, Cambridge, United Kingdom

Bruno De Borger

University of Antwerp, Antwerp, Belgium

Ismir Mulalic

Technical University of Denmark, Kgs. Lyngby, Denmark

Jan Rouwendal

*Kraks Fond, Copenhagen, Denmark
VU University, Amsterdam, The Netherlands*

Fay Dunkerley

RAND Europe, Cambridge, United Kingdom

Charlene Rohr

RAND Europe, Cambridge, United Kingdom

Mark Wardman

Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

Stephan Lehner

Vienna University of Economics & Business, Vienna, Austria

Harald Minken

Institute of Transport Economics, Oslo, Norway

Lars Hultkrantz

School of Business, Örebro University, Örebro, Sweden

Jeremy Toner

University of Leeds, Leeds, United Kingdom

Moez Kilani

University of Littoral, Cote d'Opale, Dunkerque, UMR 9221-LEM-Lille Économie Management, Lille, France

Jose M. Grisolia

Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Ken Willis

*University of Nottingham Ningbo, Ningbo, China
University of Newcastle, Newcastle, United Kingdom*

Ioannis Tikoudis

Organisation for Economic Co-operation and Development (OECD), Paris, France

Kurt Van Dender

Organisation for Economic Co-operation and Development (OECD), Paris, France

Stephen Ison

Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

Lucy Budd

*Valeria Bernardo
University of Barcelona, TechnoCampus, Barcelona, Spain*

Xavier Fageda

University of Barcelona, IESE, Barcelona, Spain

Ricardo Flores-Fillol
Universitat Rovira i Virgili, Reus, Spain

Mogens Fosgerau
University of Copenhagen, Copenhagen, Denmark

Ninette Pilegaard
Technical University of Denmark, Kongens Lyngby, Denmark

Chau Man Fung
CIB (Centre for Industrial Management/Traffic & Infrastructure), KU Leuven, Leuven, Belgium

Morten Welde
NTNU-Norwegian University of Science and Technology, Department of Civil and Environmental Engineering, Trondheim, Norway

James Odeck
NTNU-Norwegian University of Science and Technology, Department of Civil and Environmental Engineering, Trondheim, Norway

James Laird
Institute for Transport Studies, University of Leeds, England, United Kingdom

Daniel Johnson
Peak Economics, Inverness, Scotland

Hugo E. Silva
Instituto de Economía and Departamento de Ingeniería de Transporte y Logística, Pontificia Universidad Católica de Chile, Santiago, Chile

Qianwen Guo
Department of Finance and Investment, Business School, Sun Yat-sen University, Guangzhou, China

Zhongfei Li
Department of Finance and Investment, Business School, Sun Yat-sen University, Guangzhou, China

Rafael H. M. Pereira
Institute for Applied Economic Research-Ipea, Brazil

Alex Karner
The University of Texas at Austin, United States

Niek Mouter
Delft University of Technology, The Netherlands

Daniel J. Graham
Transport Strategy Centre, Imperial College London, London, United Kingdom

Adelheid Holl
Institute of Public Goods and Policy (IPP), CSIC-Spanish National Research Council, Madrid, Spain

Jan Rouwendal
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Johan Nyström
The Swedish National Road and Transport Research Institute (VTI), Stockholm, Sweden

Zhi-Chun Li
School of Management, Huazhong University of Science and Technology, Wuhan, China

Ya-Juan Chen
School of Management, Wuhan University of Technology, Wuhan, China

Gerard de Jong
The Netherlands Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

Edoardo Marcucci
Department of Political Science, University of Roma Tre, Via Gabriello Chiabrera, Roma, Italy

Valerio Gatta
Department of Political Science, University of Roma Tre, Via Gabriello Chiabrera, Roma, Italy

Michela Le Pira
Department of Civil Engineering and Architecture, University of Catania, Via Santa Sofia, Catania, Italy

Andreas Vigren
Stockholm, Sweden

Alex Anas
Department of Economics, State University of New York at Buffalo, New York, NY, United States

Patricia C. Melo
Department of Economics, ISEG-School of Economics and Management, Universidade de Lisboa & REM/UECE, Lisbon, Portugal

Anthony J. Venables
Department of Economics, University of Oxford, Oxford, United Kingdom

Anna Matas
Universitat Autònoma de Barcelona and Barcelona Institute of Economics, Barcelona, Spain

Javier Asensio
Universitat Autònoma de Barcelona and Barcelona Institute of Economics, Barcelona, Spain

Didier van de Velde
Inno-V Consulting, Amsterdam, The Netherlands

Fabio Hirschhorn
Delft University of Technology, Delft, The Netherlands

Tiziana D'Alfonso
*Department of Computer, Control, and Management
 Engineering Antonio Ruberti, Sapienza University of
 Rome, Rome, Italy*

Giuseppe Catalano
*Department of Computer, Control, and Management
 Engineering Antonio Ruberti, Sapienza University of
 Rome, Rome, Italy*

John Armstrong
*University of Southampton, Southampton,
 United Kingdom*

John Preston
*University of Southampton, Southampton,
 United Kingdom*

David Meunier
LVMT, UMR T9403, Ecole des Ponts ParisTech, France

Emile Quinet
Paris School of Economics, Ecole des Ponts ParisTech, France

Anming Zhang
*Sauder School of Business, University of British Columbia,
 Vancouver, Canada*

Yahua Zhang
*School of Commerce, University of Southern Queensland,
 Toowoomba, QLD, Australia*

Zhibin Huang
*School of Commerce, University of Southern Queensland,
 Toowoomba, QLD, Australia*

Achim I. Czerny
*Department of Logistics and Maritime Studies, Hong Kong
 Polytechnic University, Hong Kong, China*

Hao Lang
*Department of Logistics and Maritime Studies, Hong Kong
 Polytechnic University, Hong Kong, China*

Guoquan Zhang
*School of Commerce, University of Southern Queensland,
 Toowoomba, QLD, Australia*

Colin C.H. Law
*School of Commerce, University of Southern Queensland,
 Toowoomba, QLD, Australia; Faculty of Business and
 Technology, Stamford; International University, Bangkok,
 Thailand*

Yahua Zhang
*School of Commerce, University of Southern Queensland,
 Toowoomba, QLD, Australia*

Hangjun Yang
*School of International Trade and Economics,
 University of International Business and Economics,
 Beijing, China*

Chris Nash
*Institute for Transport Studies, University of Leeds, Leeds,
 Yorkshire, United Kingdom*

Andrew Smith
*Institute for Transport Studies, University of Leeds, Leeds,
 Yorkshire, United Kingdom*

Bert van Wee
Delft University of Technology, The Netherlands

Kristofer Odolinski
*The Swedish National Road and Transport Research
 Institute, Department of Transport Economics, Stockholm,
 Sweden*

Phill Wheat
*Institute for Transport Studies, University of Leeds, Leeds,
 United Kingdom*

Siri Pettersen Strandenæs
Norwegian School of Economics, Bergen, Norway

Kevin P.B. Cullinane
University of Gothenburg, Gothenburg, Sweden

Jasmine Siu Lee Lam
*School of Civil and Environmental Engineering, Nanyang
 Technological University, Singapore*

Heike Link
*German Institute for Economic Research Berlin (DIW
 Berlin), Department Energy, Transport, Environment,
 Berlin, Germany*

Olga Ivanova
PBL, The Hague, The Netherlands

Marco Ponti
*Bridges Research Trust (Scientific Responsible), Milano,
 Italy*

Tom Worsley
*Visiting Fellow, Institute for Transport Studies, University
 of Leeds, Leeds, United Kingdom*

Ian Savage
*Department of Economics and the Transportation Center,
 Northwestern University, Evanston, IL, United States*

James Odeck
*Department of Civil and Environmental Engineering,
 NTNU-Norwegian University of Science and Technology,
 Trondheim, Norway*

Morten Welde
*Department of Civil and Environmental Engineering,
 NTNU-Norwegian University of Science and Technology,
 Trondheim, Norway*

Joel P. Franklin
KTH Royal Institute of Technology, Stockholm, Sweden

Jake Whitehead
*UQ Dow Centre for Sustainable Engineering Innovation
 & School of Civil Engineering, The University of
 Queensland, St Lucia, QLD, Australia*

Patrick Plötz
*Fraunhofer Institute for Systems and Innovation Research
 ISI, Karlsruhe, Baden-Württemberg, Germany*

Patrick Jochem
*Institute for Industrial Production (IIP),
 Chair of Energy Economics, Karlsruhe Institute of
 Technology (KIT), Karlsruhe, Baden-Württemberg,
 Germany*

Frances Sprei
*Chalmers University of Technology, Department of Space,
 Earth and Environment, Göteborg, Sweden*

Elisabeth Dütschke
*Fraunhofer Institute for Systems and Innovation Research
 ISI, Karlsruhe, Baden-Württemberg, Germany*

Pedro Cantos-Sánchez
*Department of Economic Analysis and ERI-CES,
 University of Valencia, Valencia, Spain*

Patrick M. Bösch
Verkehrsbetriebe Zürich, Zürich, Switzerland

Felix Becker
*Institute for Transport Planning and Systems, Zürich,
 Switzerland*

Henrik Becker
*Institute for Transport Planning and Systems, Zürich,
 Switzerland*

Kay W. Axhausen
*Institute for Transport Planning and Systems, Zürich,
 Switzerland*

Johannes Bröcker
*Kiel University, Institute for Environmental, Resource and
 Spatial Economics, Kiel, Germany*

Stefan Flügel
Institute of Transport Economics, Oslo, Norway

Askill H. Halse
Institute of Transport Economics, Oslo, Norway

Griet De Ceuster
*Transport & Mobility Leuven, KU Leuven,
 Diestsesteenweg, Leuven, Belgium*

Inge Mayeres
*Transport & Mobility Leuven, KU Leuven,
 Diestsesteenweg, Leuven, Belgium*

Paolo Beria
Politecnico di Milano, Milan, Italy

Jing Lu
*Nanjing University of Aeronautics and Astronautics,
 Nanjing, China*

Yucan Meng
*Nanjing University of Aeronautics and Astronautics,
 Nanjing, China*

Changmin Jiang
University of Manitoba, Winnipeg, Canada

Cheng Lv
*Nanjing University of Aeronautics and Astronautics,
 Nanjing, China*

Andrew Smith
*Institute for Transport Studies, University of Leeds, Leeds,
 United Kingdom*

Chris Nash
*Institute for Transport Studies, University of Leeds, Leeds,
 United Kingdom*

Jepppe Rich
Technical University of Denmark, Lyngby, Denmark

Patrick E.P. Jochem
*Institute for Industrial Production (IIP), Chair of Energy
 Economics, Karlsruhe Institute of Technology (KIT),
 Karlsruhe, Germany*

Jake Whitehead
*School of Civil Engineering, The University of
 Queensland, St Lucia, QLD, Australia*

Elisabeth Dütschke
*Fraunhofer Institute for Systems and Innovation Research
 ISI, Karlsruhe, Germany*

Philippe Gagnepain
Paris School of Economics-Université Paris 1, Paris, France

Marc Ivaldi
*Toulouse School of Economics-EHESS, Toulouse,
 France*

Stefan Gössling
*Department of Service Management and Service Studies,
 Lund University, Lund, Sweden*

Susan Shaheen
University of California, Berkeley, CA, United States

Adam Cohen
University of California, Berkeley, CA, United States

Georgina Santos
*School of Geography and Planning, Cardiff University,
 Cardiff, United Kingdom*

Jani-Pekka Jokinen
*Max Planck Institute for Dynamics and Self-
 Organization, Göttingen, Germany*

Leif Sörensen

*Department of Psychology, PFH Private
University of Applied Sciences Göttingen, Göttingen,
Germany*

Jan Schlüter

*Department of Economics, Georg-August-University of
Göttingen, Göttingen, Germany*

Federico Boffa

*Free University of Bozen, Faculty of Economics, Brunico,
Italy*

Alberto Iozzi

*"Tor Vergata" University of Rome, Department of
Economics and Finance, Rome, Italy*

Anna Nagurney

*Department of Operations and Information
Management, Isenberg School of Management,
University of Massachusetts, Amherst, MA, United States*

Ladimer S. Nagurney

*Department of Electrical and Computer Engineering,
University of Hartford, West Hartford, CT, United States*

CONTENTS OF ALL VOLUMES

<i>Editorial Board</i>	<i>v</i>
<i>Introduction</i>	<i>vii</i>
<i>Contributors to Volume 1</i>	<i>ix</i>

VOLUME 1

Introduction to Transportation Economics	1
Market Failures and Public Decision Making in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	2
Demand for Freight Transport <i>José Holguín-Veras and Diana G. Ramírez-Ríos</i>	7
Cost Functions for Road Transport <i>José Manuel Vassallo</i>	13
Future of Urban Freight <i>Russell G. Thompson</i>	20
Operation Costs for Public Transport <i>Marco Batarce</i>	26
Natural Monopoly in Transport <i>André de Palma and Julien Monardo</i>	30
Freight Costs: Air and Sea <i>Yulai Wan and Dong Yang</i>	36
Transport Production and Cost Structure <i>Ricardo Giesen and Darío Farren</i>	46
The Concept of External Cost: Marginal versus Total Cost and Internalization <i>Sofia F. Franco</i>	56
Value of Time <i>John J. Bates</i>	67
Valuation of Carbon Emissions <i>Svante Mandell</i>	72
Valuation of Travel Time Variability Using Scheduling Models <i>Katrine Hjorth</i>	76

Value of Crowding <i>Daniel Hörcher</i>	84
What Drives Transport and Mobility Trends? The Chicken-and-Egg Problem <i>Nathalie Picard</i>	89
Pricing Principles in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	95
Long-Run Versus Short-Run Valuations <i>Stefanie Peer</i>	102
Value of Noise <i>Jon P. Nelson</i>	106
The Value of Life and Health <i>Henrik Andersson</i>	114
The Value of Security, Access Time, Waiting Time, and Transfers in Public Transport <i>Raquel Espino, Juan de Dios Ortúzar, and Luis I. Rizzi</i>	122
Demand for Passenger Transportation <i>Kenneth A Small and Robin Lindsey</i>	127
Real-World Experiences of Congestion Pricing <i>Charles Raux</i>	134
Distributional Effects of Congestion Charges and Fuel Taxes <i>Jonas Eliasson</i>	139
The Bottleneck Model <i>Dereje Abegaz and Yili Tang</i>	146
Dynamic Congestion Pricing and User Heterogeneity <i>Kathrin Goldmann and Gernot Sieg</i>	150
Economics of Parking <i>Daniel Albalade and Albert Gragera</i>	159
Loss Aversion and Size and Sign Effects in Value of Time Studies <i>Andrew Daly</i>	165
Intertemporal Variation of Valuations <i>James Fox</i>	170
The Rebound Effect for Car Transport <i>Bruno De Borger, Ismir Mulalic and Jan Rouwendal</i>	174
Elasticities for Travel Demand: Recent Evidence <i>Fay Dunkerley, Charlene Rohr and Mark Wardman</i>	179
Parking Price Elasticities <i>Stephan Lehner</i>	185
The Pareto Criterion and the Kaldor Hicks Criterion <i>Harald Minken</i>	190
Social Discount Rates <i>Lars Hultkrantz</i>	195
Cross-Elasticities between Modes <i>Mark Wardman and Jeremy Toner</i>	201
First-Best Congestion Pricing <i>Moez Kilani</i>	209

Ethical Aspects-Can We Value Life, Health, and Environment in Money Terms? <i>Jose M. Grisolia and Ken Willis</i>	216
Car tolls, Transit Subsidies for Commuting, and Distortions on the Labor Market <i>Ioannis Tikoudis¹ and Kurt Van Dender¹</i>	221
Demand Management and Capacity Planning of Airports <i>Stephen Ison and Lucy Budd</i>	227
Dealing With Negative Externalities: Low Emission Zones Versus Congestion Tolls <i>Valeria Bernardo, Xavier Fageda, and Ricardo Flores-Fillol</i>	231
The Rule-of-a-Half and Interpreting the Consumer Surplus as Accessibility <i>Mogens Fosgerau and Ninette Pilegaard</i>	237
Producer Surplus <i>Chau Man Fung</i>	242
The Robustness of Cost-Benefit Analyses <i>Morten Welde and James Odeck</i>	249
The GDP Effects of Transport Investments: The Macroeconomic Approach <i>James Laird and Daniel Johnson</i>	256
The Mohring Effect <i>Hugo E. Silva</i>	263
Public Transport Fare and Subsidy Optimization <i>Qianwen Guo and Zhongfei Li</i>	267
Transportation Equity <i>Rafael H. M. Pereira, and Alex Karner</i>	271
Impact of Transport Cost-Benefit Analysis on Public Decision-Making <i>Niek Mouter</i>	278
Causal Inference for Ex Post Evaluation of Transport Interventions <i>Daniel J. Graham</i>	283
Transport Cost and Location of Firms <i>Adelheid Holl</i>	291
Commuting, the Labor Market, and Wages <i>Jan Rouwendal, and Ismir Mulalic</i>	297
How to Buy Transport Infrastructure <i>Johan Nyström</i>	302
Procurement of Public Transport: Contractual Regimes <i>Andrew Smith, and Chris Nash</i>	308
The Mono-Centric City Model and Commuting Cost <i>Zhi-Chun Li, and Ya-Juan Chen</i>	315
Value of Time in Freight Transport <i>Gerard de Jong</i>	321
The Economics and Planning of Urban Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	326
Incentives in Public Transport Contracts <i>Andreas Vigren</i>	332
Transportation Improvements and Property Prices <i>Alex Anas</i>	337

Transport Infrastructure Effects on Economic Output: The Microeconomic Approach <i>Patricia C. Melo</i>	347
Wider Economic Impacts of Transport Investments <i>Anthony J. Venables</i>	355
Employment Effects of Transport Infrastructure <i>Anna Matas, and Javier Asensio</i>	360
Regulatory Reforms and Competition in Public Transport <i>Didier van de Velde</i>	365
How to Finance Transport Infrastructure? <i>Tiziana D'Alfonso, and Giuseppe Catalano</i>	371
Congestion, Allocation and Competition on the Railway Tracks <i>John Armstrong, and John Preston</i>	378
Public Private Partnership <i>David Meunier, and Emile Quinet</i>	385
Airline Economics <i>Anming Zhang, Yahua Zhang, and Zhibin Huang</i>	392
Privatization and Deregulation of the Airline Industry <i>Achim I. Czerny, and Hao Lang</i>	397
Price Discrimination and Yield Management in the Airline Industry <i>Guoquan Zhang, Colin C.H. Law, Yahua Zhang, and Hangjun Yang</i>	404
Regulation and Competition in Railways <i>Chris Nash, and Andrew Smith</i>	409
Cycling Economics <i>Bert van Wee</i>	414
The Economic Rationale for High-Speed Rail <i>Ginés de Rus</i>	419
Rail Cost Functions <i>Kristofer Odolinski, and Phill Wheat</i>	425
Transport and International Trade <i>Siri Pettersen Strandenes</i>	431
Maritime Economics: Organizational Structures <i>Kevin P.B. Cullinane</i>	436
Port Planning and Investment <i>Jasmine Siu Lee Lam</i>	443
Estimating the Capital Stock of Transport Infrastructure <i>Heike Link</i>	449
The Economics of Reducing Carbon Emissions From Air and Road Transport <i>Olga Ivanova</i>	457
Regulation and Financing of Toll Roads <i>Marco Ponti</i>	464
Are Megaprojects too Transformational for Cost-Benefit Analysis? <i>Tom Worsley</i>	470
Economics of Transportation Safety <i>Ian Savage</i>	476

Cost Overruns of Transportation Infrastructure Projects <i>James Odeck, and Morten Welde</i>	483
The Downs-Thomson Paradox <i>Joel P. Franklin</i>	490
Policy Instruments for Plug-In Electric Vehicles: An Overview and Discussion <i>Jake Whitehead, Patrick Plötz, Patrick Jochem, Frances Sprei, and Elisabeth Dütschke</i>	496
Vertical and Horizontal Separation in the European Railway Sector and Its Effects on Productivity <i>Pedro Cantos-Sánchez</i>	503
How will Autonomous Vehicles Impact Car Ownership and Travel Behavior <i>Patrick M. Bösch, Felix Becker, Henrik Becker, and Kay W. Axhausen</i>	508
Policy Instruments to Reduce Carbon Emissions from Road Transport Computable General Equilibrium Analysis in Transportation Economics <i>Johannes Bröcker</i>	520
Estimation of Value of Time <i>Stefan Flügel, and Askill H. Halse</i>	527
The Taxation of Car Use in the Future <i>Griet De Ceuster, and Inge Mayeres</i>	534
Cost-Benefit Analysis and Other Assessment Techniques: Contrasts and Synergies <i>Paolo Beria</i>	540
Demand for Air Travel and Income Elasticity <i>Jing Lu, Yucan Meng, Changmin Jiang, and Cheng Lv</i>	547
Generalized Cost for Transport <i>Jepppe Rich</i>	555
The Impact of Electric Vehicles on Energy Systems <i>Patrick E.P. Jochem, Jake Whitehead, and Elisabeth Dütschke</i>	560
Uber versus Taxis <i>Georgina Santos</i>	566
Contract Efficiency in Public Transport Services <i>Philippe Gagnepain, and Marc Ivaldi</i>	572
Company Cars <i>Stefan Gössling</i>	580
Transportation Network Companies (TNCs) and the Future of Public Transportation <i>Susan Shaheen, and Adam Cohen</i>	584
Public Transport in Low Density Areas <i>Jani-Pekka Jokinen, Leif Sörensen, and Jan Schlüter</i>	589
Market Failures in Transport: Direct and Indirect Public Intervention <i>Federico Boffa, and Alberto Iozzi</i>	596
The Braess Paradox <i>Anna Nagurney, and Ladimer S. Nagurney</i>	601
VOLUME 2	
Introduction to Transportation Safety and Security <i>Per Gårder</i>	1

The Concept of “Acceptable Risk” Applied to Road Safety Risk Level <i>Claes Tingvall</i>	2
Crash Not Accident <i>Robert A. Scopatz</i>	6
Age and Gender as Factors in Road Safety <i>Marion Sinclair</i>	11
Aggressive Driving and Road Rage <i>James E.W. Roseborough, Christine M. Wickens, and David L. Wiesenthal</i>	17
Aircraft Maintenance and Inspection <i>Alan Hobbs</i>	25
Airport Security <i>Richard W. Bloom</i>	34
Transport Safety and Security: Alcohol <i>James C. Fell</i>	40
Animal Crashes <i>Michal Bil</i>	53
Attenuators <i>Simonetta Boria</i>	63
ATV, Snowmobile, and Terrain Vehicle Safety <i>David P. Gilkey, and William Brazile</i>	77
Automobile Safety Inspection <i>Subasish Das</i>	85
Aviation Safety: Commercial Airlines <i>Clarence C. Rodrigues</i>	90
Aviation Safety, Freight, and Dangerous Goods Transport by Air <i>Glenn S. Baxter and Graham Wild</i>	98
Bicycle Collision Avoidance Systems: Can Cyclist Safety be Improved with Intelligent Transport Systems? <i>Lars Leden</i>	108
Bicycle Infrastructure <i>Rock E. Miller</i>	115
Bicycles, E-Bikes and Micromobility, A Traffic Safety Overview <i>Höskuldur Kröyer</i>	125
Bicycles: The Safety of Shared Systems Versus Traditional Ownership <i>Mercedes Castro-Nuño and José I. Castillo-Manzano</i>	139
Bridge Safety <i>Per Erik Garder</i>	144
Carsharing Safety and Insurance <i>Elliot Martin and Susan Shaheen</i>	150
Carjacking <i>Terance D. Mieth, Christopher Forepaugh, Tanya Dudinskaya</i>	157
Collision Avoidance Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	164
Collision Avoidance Systems, Automobiles <i>Erick J. Rodríguez-Seda</i>	173

Connected Automated Vehicles: Technologies, Developments, and Trends <i>Azra Habibovic and Lei Chen</i>	180
Construction Zones <i>Jalil Kianfar</i>	189
Costs of Accidents <i>Ulf Persson</i>	196
Critical Issues for Large Truck Safety <i>Matthew C. Camden, Jeffrey S. Hickman, Richard J. Hanowski, and Martin Walker</i>	200
Demerit Points and Similar Sanction Programs <i>Matúš ťucha and Kristýna Josrová</i>	210
Driver State and Mental Workload <i>Dick de Waard and Nicole van Nes</i>	216
Drugs, Illicit, and Prescription <i>Rune Elvik</i>	221
Education, Training, and Licensing <i>Matúš ťucha and Kristýna Josrová</i>	228
Elderly Driver Safety Issues <i>Mark J King</i>	233
Emergency Response Systems <i>Frances L. Edwards</i>	240
Emergency Vehicles and Traffic Safety <i>Shamsunnahar Yasmin, Sabreena Anowar, and Richard Tay</i>	247
Encouragement: Awards and Incentives <i>Fred Wegman</i>	255
Enforcement and Fines <i>Matúš ťucha and Ralf Risser</i>	263
Epidemiology of Road Traffic Crashes <i>Sherrie-Anne Kaye, Judy Fleiter, and Md Mazharul Haque</i>	269
Evacuation Planning and Transportation Resilience <i>Karl Kim</i>	276
Exposure: A Critical Factor in Risk Analysis <i>Frank Gross, PhD, PE</i>	282
Passenger Ferry Vessels and Cruise Ships: Safety and Security <i>Wayne K. Talley</i>	290
Fuel Economy Standards: Impacts on Safety <i>Kenneth T. Gillingham and Stephanie M. Weber</i>	296
Hazardous Materials Transport <i>Dr. Arjan Vincent van der Vlies</i>	304
Head-on Crashes <i>John N. Ivan</i>	311
Helicopters in Emergency Medical Response <i>Stephen J.M. Sollid, M.D., PhD</i>	316
Horizontal and Vertical Geometry <i>Victoria Gitelman</i>	322

Human Factors in Transportation <i>Alison Smiley, Christina (Missy) Rudin-Brown</i>	331
In-Depth Crash Analysis and Accident Investigation <i>Yong Peng, Helai Huang, and Xinghua Wang</i>	346
Incident Detection Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	351
Inequality and Traffic Safety <i>Miles Tight</i>	358
Lighting <i>John D. Bullough</i>	361
Macroscopic Safety Analysis <i>Mohamed Abdel-Aty and Jaeyoung Lee</i>	367
Motor Vehicle Crash Reportability <i>John J. McDonough</i>	380
Nominal Safety <i>Per Erik Garder</i>	386
Parking Lots <i>Maxim A. Dulebenets</i>	392
Passenger Van Safety <i>Saksith Chalermpong and Apiwat Ratanawaraha</i>	401
Passive Prevention Systems in Automobile Safety <i>B. Serpil Acar</i>	406
Pedestrian Safety, Children <i>Mette Møller</i>	415
Pedestrian Safety, General <i>Muhammad Z. Shah, Mehdi Moeinaddini, and Mahdi Aghaabbasi</i>	420
Pedestrian Safety, Older People <i>Carlo Lui</i>	429
Visually Impaired Pedestrian Safety <i>Robert S. Wall Emerson</i>	435
Photo/Video Traffic Enforcement <i>Charles M. Farmer</i>	439
Powered Two- and Three-Wheeler Safety <i>Fangrong Chang, Helai Huang, and Md. Mazharul Haque</i>	443
Railroad Safety <i>Xiang Liu and Zhipeng Zhang</i>	451
Railroad Safety: Grade Crossings and Trespassing <i>Rahim F. Benekohal and Jacob Mathew</i>	466
Recreational Boating Safety: Usage, Risk Factors, and the Prevention of Injury and Death <i>Amy E. Peden, Stacey Willcox-Pidgeon, and Kyra Hamilton</i>	477
Refuge Islands <i>Christer Hydén</i>	487
Risk Perception and Risk Behavior in the Context of Transportation <i>Martina Raue and Eva Lerner</i>	494

Road Diets <i>Robert B. Noland</i>	500
Road Safety Audits <i>Xiao Qin</i>	508
Roadside Safety Barriers <i>Dean C. Alberson</i>	513
Road Safety Management in Selected Countries <i>Paul Boase and Brian Jonah</i>	519
Roadway Pavement Conditions and Various Summer and Winter Maintenance Strategies <i>Kamal Hossain, PhD P. Eng</i>	529
Safety of Roundabouts <i>Khaled Shaaban</i>	539
Rumble Strips, Continuous Shoulder, and Centerline <i>AnnaAnund and AnnaVadeby</i>	549
Safety Culture <i>Tor-Olav Nvestad</i>	554
Safety Data Quality Management <i>Robert A. Scopatz</i>	560
School Bus Safety <i>Yousif A. Abulhassan</i>	564
School Campus Traffic Circulation <i>Dimitrios Nalmpantis</i>	568
Sexual Violence in Public Transportation <i>Vania Ceccato</i>	576
Shared Space <i>Allison B. Duncan</i>	584
Side Area Safety and Side Slopes <i>José María Pardillo-Mayora</i>	593
Simulators <i>Nichole L. Morris, Curtis M. Craig, Jacob D. Achtemeier, and Peter A. Easterlund</i>	602
Sleep-Related Issues and Fatigue <i>Prof Marion Sinclair and Estelle Swart</i>	611
Speed Governors and Limiters <i>Christer Hydén</i>	617
Speed Limits on Rural Highways <i>Peter Tarmo Savolainen and Timothy Jordan Gates</i>	622
Speed Limits on Urban Streets <i>Anna Bray Sharpin, Claudia Adriazola-Steil, Ben Welle, and Natalia Lleras</i>	632
Speed-Reducing Measures <i>Vinod Vasudevan</i>	641
Striping, Signs, and Other Forms of Information <i>Kasem Choocharukul and Kerkritt Sriroongvikrai</i>	648
Suicides <i>Brendan Ryan</i>	656

Surrogate Measures of Safety <i>Nicolas Saunier and Aliaksei Laureshyn</i>	662
Targeting Transit: the Terrorist Threat and the Challenges to Security <i>Brian Michael Jenkins</i>	668
The Swedish Vision Zero: A Policy Innovation <i>Matts-Åke Belin</i>	675
Tire Safety <i>Saied Taheri</i>	681
Town Gates: Section on Transport Safety and Security <i>Charles Tijus</i>	685
Traffic Flow Volume and Safety <i>Athanasios Theofilatos and Apostolos Ziakopoulos</i>	692
Traffic Safety and Security of Taxis and Ride-Hailing Vehicles <i>Zhe Wang, Helai Huang, and Ye Li</i>	699
Traffic Signals and Safety <i>Andrew P. Tarko</i>	706
Tunnels, Safety and Security Issues-Risk Assessment for Road Tunnels: State-of-the-Art Practices and Challenges <i>Konstantinos Kirytopoulos, Panagiotis Ntzeremes, and Konstantinos Kazaras</i>	713
Understanding, Managing, and Learning from Disruption <i>Karl Kim</i>	719
Use/Analysis of Crash Data and Underreporting of Crashes <i>Mohammadali Shirazi and Dominique Lord</i>	726
Utility Poles <i>Lai Zheng and Tarek Sayed</i>	731
Value of Life and Injuries <i>David A. Hensher</i>	737
Effects of Weather <i>Maria Pregnolato, Amirhassan Kermanshah, and Wisinee Wisetjindawat</i>	742
Wrong-Way Driving on Motorways <i>Huaguo Zhou and Md Atiquzzaman</i>	751

VOLUME 3

Freight Transport and Logistics <i>Sharon Cullinane and Kevin Cullinane</i>	1
Expanding the Perspective of Logistics and Supply Chain Management <i>David J. Closs</i>	8
Logistics and Supply Chain Management Performance Measures <i>David B. Grant and Sarah Shaw</i>	16
Economic Regulation/Deregulation and Nationalization/Privatization in Freight Transportation <i>Wayne K. Talley</i>	24
Freight Transport Policy <i>Luca Zamparini and Aura Reggiani</i>	29

Planning and Financing Logistics Spaces <i>Nicolas Raimbault</i>	35
Supply Chain Risk Management: Creating the Resilient Supply Chain <i>Richard Wilding</i>	41
Transportation Safety and Security <i>Maria G. Burns</i>	47
Resilience in Freight Transport Networks <i>Zhuohua Qu, Chengpeng Wan, and Zaili Yang</i>	53
Environmental Sustainability in Freight Transportation <i>Lisa M. Ellram</i>	58
Sustainable Logistics, CSR in Logistics, and Sustainable Supply Chain Management <i>Maria Björklund and Maja Piecyk-Ouellet</i>	64
Information Sharing and Business Analytics in Global Supply Chains <i>Prof Usha Ramanathan and Prof Ramakrishnan Ramanathan</i>	71
Logistics Information Systems <i>Petri Helo and Javad Rouzafzoon</i>	76
Factors Affecting the Selection of Logistics Service Providers <i>Aicha AGUEZZOUL</i>	85
Logistics Service Performance <i>Kee-hung Lai, Jinan Shao, and Yongyi Shou</i>	89
The World Bank's Logistics Performance Index <i>Christina K. Wiederer cwiederer, Jean-François Arvis, Lauri M. Ojala, and Tuomas M. M. Kiiski</i>	94
Outsourcing Logistics Functions <i>Evi Hartmann, Hendrik Birkel, and Matthias Kopyto</i>	102
Freight Transport and Logistics in JIT Systems <i>James H. Bookbinder and M. Ali Aælkü</i>	107
Supply Chain Finance <i>Erik Hofmann</i>	113
Packaging Logistics <i>Jesús García-Arca, Alicia Trinidad González-Portela Garrido, and J. Carlos Prado-Prado</i>	119
The Bullwhip Effect <i>Jan C. Fransoo and Maximiliano Udenio</i>	130
Blockchain Applications in Logistics <i>Yingli Wang</i>	136
Logistics in Asia <i>Shong-lee Ivan Su</i>	143
Logistics in the Developing World <i>Charles Kunaka</i>	150
Freight Network Modeling <i>Lóránt Tavasszy and Yousef Maknoon</i>	157
National Freight Transport Models <i>Gerard de Jong</i>	162
Freight Flows in Cities <i>Genevieve Giuliano</i>	168

Urban Logistics and Freight Transport <i>Michael Browne, José Holguín-Veras, and Julian Allen</i>	178
Omni-Channel Logistics <i>Tom Van Woensel</i>	184
Humanitarian Logistics <i>Gyöngyi Kovács and Diego Vega</i>	190
Optimization of Humanitarian Logistics <i>M. Teresa Ortuño, Jose M. Ferrer, Inmaculada Flores, and Gregorio Tirado</i>	195
Event Logistics <i>Rev. Ruth Dowson and Dan Lomax</i>	201
Reverse Logistics <i>Dale S. Rogers and Ronald S. Lembke</i>	208
E-Tailing and Reverse Logistics <i>Sharon Cullinane</i>	219
Green Routing of Freight Vehicles <i>Tolga Bektaş</i>	224
Freight Mode Choice <i>Hyun Chan Kim and Alan Nicholson</i>	231
The Value of Time in Freight Transport <i>María Feo-Valero, Amaya Vega, and Bárbara Vázquez-Paja</i>	236
Behavioral Research in Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	242
Container (Liner) Shipping <i>Theo Notteboom</i>	247
Bulk Shipping Markets: An Overview of Market Structure and Dynamics <i>Manolis G. Kavussanos and Stella A. Moysiadou</i>	257
Ferries and Short Sea Shipping <i>Lourdes Trujillo and Alba Martínez-López</i>	280
Shipping and the Environment <i>Karin Andersson, Selma Brynolf, Lena Granhag, and J. Fredrik Lindgren</i>	286
Energy Efficiency of Ships <i>Harilaos N. Psaraftis</i>	294
Seaports <i>Mary R. Brooks and Geraldine Knatz</i>	299
Port Hinterlands <i>Francesco Parola, Giovanni Satta, and Francesco Vitellaro</i>	305
Seaports as Clusters of Economic Activities <i>Peter W. de Langen</i>	310
Port Efficiency and Effectiveness <i>Lourdes Trujillo, María Manuela González, Casiano Manrique-De-Lara-Peñate, and Ivone Pérez</i>	316
Container Port Automation <i>Michael G.H. Bell</i>	323
Optimizing Crane Operations in Ports Scheduling of Liner Container Shipping Services	335

<i>Yuquan Du and Qiang Meng</i>	
Dry Ports	344
<i>Gordon Wilmsmeier and Jason Monios</i>	
Arctic Shipping	349
<i>Yufeng Lin, David G. Babb, and Adolf K.Y. Ng</i>	
Airfreight and Economic Development	355
<i>Kenneth Button</i>	
Air Freight Logistics	361
<i>Keith Debbage and Neil Debbage</i>	
Air Freight Marketing	369
<i>Lucy Budd and Stephen Ison</i>	
Drones in Freight Transport	374
<i>Oliver Kunze</i>	
Duty of Care in the Selection of Motor Carriers	382
<i>Thomas M. Corsi</i>	
Carrier Selection for Less-Than-Truckload (LTL) Shipments	388
<i>Dinçer Konur, Gonca Yildirim, and Bahriye Cesaret</i>	
Decarbonizing Road Freight Transport	395
<i>Heikki Liimatainen</i>	
The Rebound Effect in Road Freight Transport	402
<i>Tooraj Jamasb and Manuel Llorca</i>	
Autonomous Goods Transport	407
<i>Heike Flømig</i>	
Rail Freight	413
<i>Dr Allan Woodburn</i>	
Rail Freight Vehicles	423
<i>Maksym Spiryagin, Qing Wu, Peter Wolfs, Colin Cole, Valentyn Spiryagin, and Tim McSweeney</i>	
Eurasia Rail Freight: Enablers and Inhibitors of Future Growth	436
<i>Hendrik Rodemann and Simon Templar</i>	
Intermodal and Synchromodal Freight Transport	456
<i>Tomas Ambra, Koen Mommens, and Cathy Macharis</i>	
Pipelines	463
<i>Matthew E. Oliver</i>	
3D Printers and Transport	471
<i>Wouter P.C. Boon and Bert van Wee</i>	
The Physical Internet and Logistics	479
<i>Eric Ballot and Shenle PAN</i>	
Bicycles for Urban Freight	488
<i>Barbara Lenz and Johannes Gruber</i>	
China's Belt and Road Initiative	495
<i>Paul Tae-Woo Lee</i>	

VOLUME 4

Introduction to Traffic Management <i>Edward C.S. Chung</i>	1
Urban Motorway Management <i>John Gaffney and Hendrik Zurlinden</i>	2
Ramp Metering Application <i>John Gaffney and Hendrik Zurlinden</i>	10
City Wide Coordinated Ramp Meters <i>John Gaffney and Hendrik Zurlinden</i>	21
Variable Speed Limits for Traffic Efficiency Improvement <i>José Ramón D. Frejo and Bart De Schutter</i>	33
Hard Shoulder Running <i>Justin Geistefeldt</i>	41
High-Occupancy Vehicle (HOV) and High-Occupancy Toll (HOT) Lanes <i>Roxana J. Javid, Jiani Xie, Lijiao Wang, Wenruifan Yang, Ramina Jahanbakhsh Javid, and Mahmoud Salari</i>	45
Reversible Lanes: Guidelines, Operation and Control, Research Directions <i>Gowri Asaithambi, Venkatesan Kanagaraj, and Madhuri Kashyap</i>	52
Electronic Toll Collection <i>Azusa Toriumi</i>	60
Road Pricing-Theory and Applications <i>Kian Keong Chin</i>	68
Road Pricing 1: The Theory of Congestion Pricing <i>Timothy D. Hau</i>	74
Road Pricing 2: Short- and Long-Run Equilibrium of Road Transportation <i>Timothy D. Hau</i>	83
Road Pricing 3: The Implications for Pricing Public Transportation <i>Timothy D. Hau</i>	90
Road Pricing 4: Case Study-The Implementation of Electronic Road Pricing in Hong Kong <i>Timothy D.</i>	103
Advanced Travelers Information Systems (ATIS) <i>Chintan Advani and Ashish Bhaskar</i>	106
Travel Time Reliability <i>Sharmili Banik, Anil Kumar, and Lelitha Vanajakshi</i>	109
Traffic Incident Management <i>Ruimin Li</i>	122
Traffic Incident Detection <i>Shuyan Chen and Yingjiu Pan</i>	128
Bottleneck <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	134
Flow Breakdown <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	143
Recurrent Congestion <i>Takahiro Tsubota</i>	154

Nonrecurrent Congestion <i>Takahiro Tsubota</i>	158
Freeway to Arterial Interfaces <i>Abolfazl Karimpour and Yao-Jan Wu</i>	162
Arterial Road Management <i>Jiaqi Ma, Yi Guo and Adekunle Adebisi</i>	169
Signalized Intersections <i>Chaitrali Shirke</i>	178
Protected Phase <i>Jiarong Yao, Yumin Cao, and Keshuang Tang</i>	185
Permitted Phase <i>Yumin Cao, Jiarong Yao, and Keshuang Tang</i>	196
Hook Turns: Implementation, Benefits, and Limitations <i>Sara Moridpour and Amir Falamarzi</i>	203
Traffic Signal Coordination <i>Rahim F. Benekohal</i>	213
Emergency Vehicle Priority (Preemption): Concept and Advancements <i>Chaitrali Shirke</i>	221
Signalized Roundabouts <i>Yetis Sazi Murat and Rui-jun Guo</i>	227
Turbo Roundabouts: Design, Capacity and Comparison With Alternative Types of Roundabouts <i>Marco Guerrieri and Raffaele Mauro</i>	238
Capacity of an Intersection <i>Ashish Verma and Milan Mathew Thomas</i>	247
Delay <i>Shinji Tanaka</i>	258
Queue Length <i>Shinji Tanaka</i>	263
Local Area Traffic Management <i>Michael A.P. Taylor</i>	268
On-Street Parking <i>Jun Chen and Guang Yang</i>	278
Off-Street Parking <i>Jun Chen and Guang Yang</i>	285
Parking Information Systems <i>Behrang Assemi and Douglas Baker</i>	289
Performance-Based Parking Management <i>Douglas Baker and Behrang Assemi</i>	299
Transit Priority <i>Wanjing Ma, Qiheng Lin, and Ling Wang</i>	307
Adaptive Bus Control <i>Monica Menendez</i>	315
Transit Fare Collection <i>Mahmoud Mesbah and Kamal Khanali</i>	325

Transit Information Systems <i>Ankit Kumar Yadav and Nagendra R Velaga</i>	331
Tram Lane Configurations and Driving Rules <i>Farhana Naznin</i>	338
Railway Crossing <i>Chunliang Wu and Inhi Kim</i>	341
Pedestrian Crossing (Crosswalk) <i>Miho Iryo-Asano, Wael K.M. Alhajyaseen, and Koji Suzuki</i>	346
Bike-Sharing System: Uncovering the “1Success Factors” <i>S.K. Jason Chang and Amanda Fernandes Ferreira</i>	355
Airspace Systems Technologies-Overview and Opportunities <i>Banavar Sridhar and Gano B. Chatterji</i>	363
Air Traffic Flow and Capacity Management <i>Li Weigang and Cristiano P Garcia</i>	380
Port Management <i>Maria G. Burns</i>	390
Port Performance Measurement from a Multistakeholder Perspective <i>Min-Ho Ha, Zaili Yang, and Young-Joon Seo</i>	396
Transport Modeling and Data Management <i>Chandra Bhat</i>	407
Computational Methods and Data Analytics <i>Bilal Farooq and David López</i>	408
Activity-Based Models <i>Renato Guadamuz and Rajesh Paleti</i>	414
Advanced Traveler Information Systems <i>Stephen D. Boyles</i>	418
Demand-Responsive Transit, Evaluation Studies <i>Sebastián Raveau</i>	423
Location Choice Models <i>Adam Wilkinson Davis</i>	428
Full Feedback and Equilibrium Modeling in Urban Travel Forecasting <i>Yu (Marco) Nie, Jun Xie, and David Boyce</i>	432
The Use and Value of Geographic Information Systems in Transportation Modeling <i>Ming Zhang</i>	440
Latent Demand and Induced Travel <i>Charisma F. Choudhury</i>	448
ICT, Virtual and In-Person Activity Participation, and Travel Choice Analysis <i>Jacek Pawlak and Giovanni Circella</i>	452
Public Transit Ridership Forecasting Models <i>Ipsita Banerjee, Deepa L, and Abdul Rawoof Pinjari</i>	459
Spatial Mismatch, Job Access, and Reverse Commuting <i>Gian-Claudia Sciarra</i>	468

Microsimulation and Agent-Based Models in Transportation <i>Milos Balac</i>	473
Choice Models in Transportation <i>Naveen Chandra Iraganaboina and Naveen Eluru</i>	477
Multi-Criteria Decision Analysis <i>Zhanmin Zhang and Srijith Balakrishnan</i>	485
The National Household Travel Survey Data Series (NPTS/NHTS) <i>Nancy McGuckin</i>	493
Route Choice and Network Modeling <i>Emma Frejinger and Maëlle Zimmermann</i>	496
Departure Time Choice Modeling <i>Khandker Nurul Habib</i>	504
Traveler Responses to Congestion <i>David T. Ory and Gayathri Shivaraman</i>	509
Origin-Destination Demand Estimation Models <i>William H.K. Lam, Hu Shao, Shuhan Cao, and Hai Yang</i>	515
Parking Demand Models <i>S.C. Wong, Zhi-Chun Li, and William H.K. Lam</i>	519
Pavement Management Systems <i>Senthilmurugan Thyagarajan</i>	524
Residential Location Choice Models <i>Shlomo Bekhor and Sigal Kaplan</i>	531
Transport Demand Management <i>Feiyang Zhang and Becky P.Y. Loo</i>	537
Traffic Flow Analysis <i>H. Michael Zhang and Jia Li</i>	544
Autonomous Vehicles and Transportation Modeling <i>Annesha Enam, Felipe de Souza, Omer Verbas, Monique Stinson, and Joshua Auld</i>	557
Ride-Hailing and Travel Demand Implications <i>Felipe F. Dias</i>	564
Transportation Modeling and Planning Software <i>Joel Freedman</i>	569
Transportation Statistics and Databases <i>Taha Hossein Rashidi</i>	574
Travel Surveys <i>Stacey G. Bricka</i>	587
Travel Demand Forecasting: Where Are We and What Are the Emerging Issues <i>Thomas F. Rossi</i>	590
Travel Model Calibration and Validation <i>Ram M. Pendyala</i>	596
Trip Chaining Analysis <i>Cynthia Chen and Yusak Susilo</i>	606
Vehicle Ownership Models <i>Dr. Sören Groth and Prof. Dr. Dirk Wittowsky</i>	612

Bicycle Sharing/Bikesharing <i>Catherine Morency and Jean-Simon Bourdeau</i>	617
Carsharing <i>Shiva Habibi and Frances Sprei</i>	623
Urban Recreational Travel <i>Long Cheng and Frank Witlox</i>	629
Place Perception and Travel Behavior <i>Kathleen Deutsch-Burgner and Konstadinos G. Goulias</i>	635

VOLUME 5

Transport Modes <i>Edoardo Marcucci</i>	1
Infrastructure Transport Investments, Economic Growth and Regional Convergence <i>Xavier Fageda and Cecilia Olivieri</i>	2
Transport Modes and an Aging Society <i>Charles B.A. Musselwhite and Theresa Scott</i>	6
Sustainable Mobility Paths <i>Erling Holden and Geoffrey Gilpin</i>	13
Vehicles that Drive Themselves: What to Expect with Autonomous Vehicles <i>Michele D. Simoni and Kara M. Kockelman</i>	19
Transport Modes and Tourism <i>Ila Maltese and Luca Zamparini</i>	26
Transport Modes and Accessibility <i>Bert van Wee</i>	32
Transport Modes and Globalization <i>Jean-Paul Rodrigue</i>	38
Transport Modes and Cities <i>Erick Guerra and Gilles Duranton</i>	45
Transport Modes and Remote Areas	51
Modeling Mode Choice in Freight Transport <i>Lóránt Tavasszy and Gerard de Jongb</i>	57
Travel Mode Choice as Reasoned Action <i>Sebastian Bamberg, Icek Ajzen, and Peter Schmidt</i>	63
Energy Consumption of Transport Modes <i>Zissis Samaras and Ilias Vouitsis</i>	71
Transport Modes and People With Limited Mobility <i>Roger L Mackett</i>	85
Transport Modes and Commuters <i>Colin G. Pooley</i>	92
Shopping and Transport Modes <i>Antonio Comi</i>	98
Transport Modes and Health <i>Jennifer S. Mindell and Sandra Mandic</i>	106

Multimodality in Transportation <i>Sören Groth and Tobias Kuhnimhof</i>	118
Transport Modes and Disasters <i>Brian Wolshon</i>	127
Big Data for Public Transport Planning <i>Jan-Dirk Schmöcker</i>	134
Active Transport: Heterogeneous Street Users Serving Movement and Place Functions <i>Regine Gerike, Stefan Hubrich, Caroline Koszowski, Bettina Schröter, and Rico Wittwer</i>	140
Electric Vehicles <i>Christine Eisenmann, Daniel Görges, and Thomas Franke</i>	147
Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service <i>Susan Shaheen, PhD and Adam Cohen</i>	155
Adoption of new travel information platforms <i>Sigal Kaplan</i>	160
ICT and Transport Modes <i>Galit Cohen-Blankshtain</i>	165
Mode Choice and Life Events <i>Joachim Scheiner</i>	171
Introduction to Air Transport <i>Milan Janić</i>	178
The History of Air Transportation <i>Richard P. Hallion</i>	192
The Geography of Air Transport <i>Lucy Budd and Stephen Ison</i>	198
The Future of Air Transport <i>Rico Merkert and James Bushell</i>	203
Air Transport and Its Territorial Implications <i>Lanfranco Senn</i>	208
Next Generation Travel: Young Adults' Travel Patterns <i>Tobias Kuhnimhof and Scott Le Vine</i>	215
Airport Network Planning and Its Integration with the HSR System <i>Francesca Pagliara, Juan Carlos Martín, and Concepción Román</i>	222
Airport Management <i>Peter Forsyth and Hans-Martin Niemeier</i>	229
Airport Regulation <i>Achim I. Czerny</i>	234
Air Route Planning and Development <i>Renan Peres de Oliveira and Gui Lohmann</i>	241
Airline Management <i>Sveinn Vidar Gudmundsson</i>	249
Air Cargo <i>Volodymyr Bilotkach</i>	258

Airline Regulation <i>Andrew R. Goetz</i>	263
Air Vehicles Classification <i>Vincenzo Torre</i>	269
Aircraft Manufacturing <i>Antonio Sollo</i>	290
A Geography of Road Transport in Cities <i>Cristian Domarchi and Juan de Dios Ortúzar</i>	300
The Future of Road Transport <i>Preston L. Schiller</i>	306
Road Transportation and Territorial Scale <i>Ana M. Condeço-Melhorado</i>	315
Road Modes: Walking <i>Kevin Manaugh, PhD Associate Professor</i>	320
Bus Public Transport Planning and Operations <i>Ehab Diab and Ahmed El-Geneidy</i>	326
Street Design for Active Travel <i>Bruce Appleyard</i>	333
Transit Planning and Management <i>Zakhary Mallett and Marlon G Boarnet</i>	349
Road Infrastructure: Planning, Impact and Management <i>Jos Arts, Wim Leendertse, and Taede Tillema</i>	360
Road Transport Planning at the Urban Scale <i>David A. King and Kevin J. Krizek</i>	373
Road Traffic Regulation: Road Pricing and Environmental Quality <i>Marco Percoco</i>	378
Car Ownership and Car Use: A Psychological Perspective <i>J.L. Veldstra, A.B. Ünal, E.M. Steg</i>	384
Road Transport: E-Scooters <i>Gysele Lima Ricci and Klaus Bogenberger</i>	391
Railway Station and Network Planning <i>Ingo Arne Hansen</i>	399
Introduction to Rail Transport <i>Chris Nash and Tony Fowkes</i>	406
The History of Rail Transport <i>Carlo Ciccarelli, Andrea Giuntini, and Peter Groote</i>	413
The Geography of Rail Transport <i>Frédéric Dobruszkes and Amparo Moyano</i>	427
Rail Transport and Territorial Scale <i>Prof. Andrés Monzón and Dr. Elena López</i>	437
Railway Management <i>Vassilios A. Profillidis</i>	444
Railway Terminal Regulation <i>Nacima Baron</i>	454

Service Network Design for Freight Railroads <i>Teodor Gabriel Crainic</i>	464
Subway Systems <i>Guillaume Monchambert, Daniel Hörcher, Alejandro Tirachini, and Nicolas Coulombel</i>	471
Railway Company Management <i>Vilius Nikitinas, Skaistė Miliauskaitė</i>	479
Regulation of Rail Infrastructure and Services <i>Javier Campos</i>	485
Rail Vehicle Classification <i>Christos Pyrgidis and Alexandros Dolianitis</i>	490
Introduction to Maritime Shipping <i>Christa Sys and Thierry Vanelslander</i>	508
The Geography of Maritime Transport <i>César Ducruet and Justin Berli</i>	517
The Future of Maritime Transport <i>Harilaos N. Psaraftis</i>	535
Maritime Transport and Territorial Scale <i>Brian Slack</i>	540
Containerization and the Port Industry <i>Hercules Haralambides</i>	545
Port Management <i>Lourdes Trujillo, Daniel Castillo Hidalgo, and Manuel Herrera</i>	557
Economic and Environmental Regulation in the Port Sector <i>Beatriz Tovar and Alan Wall</i>	563
Maritime Route Planning <i>Johan Woxenius</i>	570
Inland Waterway Transport and Inland Ports: An Overview of Synchromodal Concepts, Drivers, and Success Cases in the IWW Sector <i>Behzad Behdani, Bart Wiegmans, and Yun Fan</i>	577
Maritime Company Passenger Management/Liner Industry <i>Claudio Ferrari and Alessio Tei</i>	587
Cruise Industry <i>Athanasios A. Pallis and Aimilia A. Papachristou</i>	593
International Maritime Regulation: Closing the Gaps Between Successful Achievements and Persistent Insufficiencies <i>Laurent Fedi</i>	600
Ship Classification <i>Gareth C. Burton and Mimosa T. Miller</i>	607
The Shipbuilding Industry and its Interactions With Shipping <i>Paul William Stott</i>	617
Methods for Designing Public Transport Networks <i>Zain Ul Abedin and Avishai (Avi) Ceder</i>	625
Space Transportation <i>Mark Hempsell</i>	638

Pipelines	646
<i>Franco Cotana and Mattia Manni</i>	
Women and Transport Modes	656
<i>Priya Uteng, PhD and Yusak Susilo, PhD</i>	
Transport Modes and Big Data	665
<i>Hannah D Budnitz, Emmanouil Tranos, and Lee Chapman</i>	
Transport Modes and Inequalities	671
<i>Caroline Mullen</i>	
Railway Traffic Management	678
<i>Francesco Corman</i>	
Indoor Transportation	684
<i>Lutfi Al-Sharif</i>	
Bicycle as a Transportation Mode	697
<i>Raktim Mitra and Paul M. Hess</i>	
Urban Air Mobility: Opportunities and Obstacles	702
<i>Adam Cohen and Susan Shaheen, PhD</i>	
Transport Modes and Sustainability	710
<i>Long Cheng, Jonas De Vos, and Frank Witlox</i>	

VOLUME 6

Introduction to Transport Policy and Planning	1
<i>Maria Attard</i>	
Workplace Parking Levy	2
<i>Stephen Ison and Lucy Budd</i>	
Air Transport	7
<i>Lucy Budd and Stephen Ison</i>	
Mobility as a Service	12
<i>Milo N. Mladenović</i>	
Bicycle Sharing	19
<i>Cyrille Médard de Chardon</i>	
Light Rail	31
<i>Fiona Ferbrache</i>	
Planning Tourism Travel	39
<i>Luca Zamparini</i>	
Transport Planning and Management and its Implications in Chinese Cities	44
<i>Mengqiu Cao</i>	
Road Safety	51
<i>George Yannis and Eleonora Papadimitriou</i>	
Mobility Planning and Policies for Older People	59
<i>Charles B A Musselwhite</i>	
Electric Mobility	64
<i>Graham Parkhurst</i>	
Transport Policy and Governance	73
<i>Lisa Hansson</i>	

Transferability of Urban Policy Measures <i>Paul Martin Timms</i>	77
Accessibility Tools for Transport Policy and Planning <i>Benjamin Büttner</i>	83
Demand Responsive Transport <i>Marcus Enoch</i>	87
Transport and Climate Change <i>Robin Hickman and Christine Hannigan</i>	94
Taxicabs and Microtransit <i>David A. King</i>	101
Land-Use and Transport Planning <i>Luis A. Guzman</i>	107
Parking <i>Stephen Ison and Lucy Budd</i>	113
Transport Planning in the Global South <i>Daniel Oviedo and Mariajose Nieto-Combariza</i>	118
Planning for Children's Independent Mobility <i>E. Owen D. Waygood and Raktim Mitra</i>	125
The Politics of Mobility Policy <i>Geoff Vigar</i>	131
Technology Enabled Data for Sustainable Transport Policy <i>Susan M. Grant-Muller, Mahmoud Abdelrazek, Hannah Budnitz, Caitlin D. Cottrill, Fiona Crawford, Charisma F. Choudhury, Teddy Cunningham, Gillian Harrison, Frances C. Hodgson, Jinhyun Hong, Adam Martin, Oliver O'Brien, Claire Papaix, and Panagiotis Tsoleridis</i>	135
Car Sharing <i>Cyriac George and Tanu Priya Uteng</i>	142
Toward a More Holistic Understanding of Mega Transport Project (MTP) Success <i>John Ward</i>	147
Equity Considerations in Transport Planning <i>Karel Martens</i>	154
Planning for Rail Transport <i>Simon P. Blainey</i>	161
Connected and Autonomous Vehicles: Priorities for Policy and Planning <i>Dr. Alexandros Nikitas</i>	167
Gendered Mobility <i>Sheila Mitra-Sarkar</i>	173
Modeling and Simulation for Transport Planning <i>Michela Le Pira, Giuseppe Inturri, and Matteo Ignaccolo</i>	184
Externalities and External Costs in Transport Planning <i>Silvio Nocera</i>	191
Planning for Public Transport with Automated Vehicles <i>Gonçalo Homem de Almeida Rodriguez Correia</i>	198
Urban Congestion Charging in Transport Planning Practice <i>Ida Kristoffersson and Maria Börjesson</i>	206

Energy and Transport Planning <i>Debbie Hopkins and Christian Brand</i>	214
Customer Satisfaction as a Measure of Service Quality in Public Transport Planning <i>Laura Eboli and Gabriella Mazzulla</i>	220
Car Sharing and the Impact on New Car Registration <i>Mario Intini and Marco Percoco</i>	225
Evaluation Methods in Transport Policy and Planning <i>Niek Mouter</i>	230
Transport and Air Quality Planning and Policy <i>Dr Fabio Galatioto</i>	236
Cycling Policies <i>Esther Anaya-Boig</i>	241
Community Severance <i>Paulo Anciaes and Jennifer S. Mindell</i>	246
Planning for Bus Priority <i>Claus H. Sørensen, Fredrik Pettersson, and Joel Hansson</i>	254
High-Speed Rail and the City <i>Marie Delaplace</i>	261
Public Engagement in Transport Planning <i>Miriam Ricci</i>	266
Long-Distance Travel <i>Giulio Mattioli and Muhammad Adeel</i>	272
Urban Freight Policy <i>Laetitia Dablanç</i>	278
Regional Transport Planning <i>Chia-Lin Chen</i>	286
Transport Project Financing <i>Romeo Danielis and Lucia Rotaris</i>	292
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	298
Transitions and Disruptive Technologies in Transport Planning <i>Kate Pangbourne and Maria Attard</i>	309
Community Transport: Filling the Gaps for Those in Need of Mobility <i>Ian Shergold</i>	314
Fundamental Emerging Concepts and Trends for Environmental Friendly Urban Goods Distribution Systems <i>Sandra Melo</i>	320
Plateau Car <i>David Metz</i>	324
Emerging Trends in Transport Demand Modeling in the Transition Toward Shared Mobility and Autonomy <i>Patrizia Franco</i>	331
Policy and Planning for Walkability <i>Carlos Cañas Sanz and Maria Attard</i>	340

Public Transport Subsidy and Regulation <i>Jonathan Cowie</i>	349
Urban Regeneration and Transportation Planning <i>Thomas Vanoutrive</i>	356
Social and Distributional Impact Assessment in Transport Policy <i>Laura Walker and Angela Curl</i>	361
Mobility Planning for Healthy Cities <i>Ersilia Verlinghieri</i>	368
Transport Demand Management <i>Begoña Guirao</i>	374
Planning and the Global Movement of Goods and Commodities <i>Christopher Clott and Chris Petrocelli</i>	380
Public Transport Network Planning <i>Corinne Mulley and John D. Nelson</i>	388
A Timely Perspective on Planning for Ageing Infrastructure <i>Anthony Perl</i>	395
The role of media in transport planning and the transport policy process <i>Özgül Ardiç and J.A. Annema</i>	400
Travel Plans <i>Stephen Potter and Marcus Enoch</i>	408
Home Deliveries and their Impact on Planning and Policy <i>Rosário Macário</i>	413
Planning for Safe and Secure Transport Infrastructure <i>Per Erik Gårder</i>	418
Sensors and Data Driven Approaches in Transport <i>Mohammad Sadrani and Constantinos Antoniou</i>	426
Planning Active Travel and School Transport <i>Fahimeh Khalaj, Dorina Pojani, and Sara Alidoust</i>	432
VOLUME 7	
Introduction to Transport Psychology <i>Carlo Prato</i>	1
From Self-reports to Auto-Tech-Detect (ATD)-based Self-reports in Traffic Research <i>Türker Özkan and Timo Lajunen</i>	2
Observational Field Studies in Traffic Psychology <i>Tova Rosenbloom and Hodaya Levy</i>	8
Driving Simulators <i>Karel A. Brookhuis</i>	14
Naturalistic Driving Studies: An Overview and International Perspective <i>Johnathon P. Ehsani, Joanne L. Harbluk, Jonas Bärghman, Ann Williamson, Jeffrey P. Michael, Raphael Grzebieta, Jake Olivier, Jan Eusebio, Judith Charlton, Sjaanie Koppel, Kristie Young, Mike Lenné, Narelle Haworth, Andry Rakotonirainy, Mohammed Elhenawy, Gregoire Larue, Teresa Senserrick, Jeremy Woolley, Mario Mongiardini, Christopher Stokes, Paul Boase, John Pearson, and Feng Guo</i>	20

A Detailed Approach to Qualitative Research Methods <i>Sonja Forward and Lena Levin</i>	39
Behavioral Change <i>Sonja Haustein</i>	46
Habitual Behavior <i>Carlo G. Prato</i>	54
Driving Behavior and Skills <i>Timo Lajunen and Türker Özkan</i>	59
The Multidimensional Driving Style Inventory <i>Orit Taubman - Ben-Ari</i>	65
Risk Perception in Transport: A Review of the State of the Art <i>Trond Nordfjrn, An-Magritt Kummeneje, Mohsen F. Zavareh, Milad Mehdizadeh, and Torbjørn Rundmo</i>	74
Social-Symbolic and Affective Aspects of Car Ownership and Use <i>Birgitta Gatersleben</i>	81
Pitfalls of Statistical Methods in Traffic Psychology <i>J.C.F. de Winter and D. Dodou</i>	87
Data Analysis: Structural Equation Models <i>Marco Diana</i>	96
Data Analysis: Integrated Choice and Latent Variable Models <i>Carlo G Prato</i>	102
Explaining Data Analysis Using Qualitative Methods <i>Lena Levin and Sonja Forward</i>	107
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	113
Driver Aggression and Anger <i>Mark J.M. Sullman and Amanda N. Stephens</i>	121
Speeding: A "Tragedy of the Commons" Behavior <i>Bryan E. Porter, Thomas D. Berry, and Kristie L. Johnson</i>	130
Cycling as a Mode Choice: Motivational Psychology <i>Sigal Kaplan</i>	136
Motorcyclists <i>Narelle Haworth</i>	144
Drivers' Hazard Perception Skill <i>Mark S. Horswill and Andrew Hill</i>	151
Driver Education and Training for New Drivers: Moving beyond Current 'Wisdom' to New Directions <i>Teresa Senserrick, Oscar Oviedo-Trespalacios, David Rodwell, and Sherrie-Anne Kaye</i>	158
Road Safety Advertising: What We Currently Know and Where to From Here <i>Ioni Lewis, Barry Watson, Katherine M. White, and Sonali Nandavar</i>	165
Traffic Law Enforcement Theories and Models <i>Richard Tay</i>	171
Satisfaction with Travel and the Relationship to Well-Being <i>Tommy Gørling and Filip Fors Connolly</i>	177
	182

Electromobility: History, Definitions and an Overview of Psychological Research on a Sustainable Mobility System <i>Josef F. Krems and Isabel KreiÃŸig</i>	
Sharing: Attitudes and Perceptions <i>Yi Wen and Christopher R Cherry</i>	187
Women's Travel Patterns, Attitudes, and Constraints Around the World <i>Sandra Rosenbloom</i>	193
Driver Stress and Driving Performance <i>Lisa Dorn</i>	203
Introduction to Sustainability and Health in Transportation <i>Roger Vickerman</i>	225
Sustainable Development Goals and Health <i>Rosa Surinach</i>	226
Human Ecology <i>Roderick J. Lawrence</i>	234
Car- Free Cities <i>Haneen Khreis and Mark J. Nieuwenhuijsen</i>	240
Superblocks Base of a New Model of Mobility and Public Space. Barcelona as an Example <i>Salvador Rueda Palenzuela</i>	249
Resilience of Transport Systems <i>ErikJenelius and Lars-GöranMattsson</i>	258
Impact of Shipping to Atmospheric Pollutants: State-of-the-Art and Perspectives <i>Daniele Contini and Eva Merico</i>	268
Noise Pollution From Transport <i>Marianna Jacyna, Emilian Szczepański, Konrad Lewczuk, Mariusz Izdebski, Ilona Jacyna-Gotda, Michał, Kłodawski, Paweł, Gołda, Piotr Goł, and Ebiowski</i>	277
Visual Impacts From Transport <i>Paulo Rui Anciaes</i>	285
Light Pollution <i>John D. Bullough</i>	292
Wildlife Crossings and Barriers <i>Scott D. Jackson</i>	297
Environmental Justice, Transport Justice, and Mobility Justice <i>Devajyoti Deka</i>	305
Transport Noise and Health <i>Elisabete F. Freitas, Emanuel A. Sousa, and Carlos C. Silva</i>	311
Climate Change and Health, Related to Transport <i>Ersilia Verlinghieri</i>	320
Urban Greenspace, Transportation, and Health <i>Payam Dadvand and Mark J. Nieuwenhuijsen</i>	327
Transport Access and Health <i>Alireza Ermagun</i>	335
Social Exclusion and Health, Related to Transport <i>Roger L. Mackett</i>	341

Burden of Disease Assessment <i>David Rojas-Rueda</i>	347
Achieving a Near-Zero CO <i>Lewis M. Fulton</i>	353
Disabled Travelers <i>Bryan Matthews</i>	359
Health Impacts of Connected and Autonomous Vehicles <i>Soheil Sohrabi</i>	364
Electric Vehicles and Health <i>Kanok Boriboonsomsin</i>	372
Shared Mobility Opportunities and their Computational Challenges for Improving Health-Related Quality of Life <i>Cristiano Martins Monteiro, Cláudia Aparecida Soares Machado, Adelaide Cassia Nardocci, Fernando Tobal Berssaneti, José Alberto Quintanilha, and Clodoveu Augusto Davis</i>	376
Bike Sharing and Health <i>David Rojas-Rueda and Mark J. Nieuwenhuijsen</i>	384
E-Bikes and Health <i>Aslak Fyhri and Hanne Beate Sundfør</i>	393
<i>Index</i>	399

Introduction to Transportation Economics

Maria Börjesson, VTI Swedish National Road and Transport Research Institute, Linköping University, Linköping, Sweden

© 2021 Elsevier Ltd. All rights reserved.

A well-functioning transportation system is decisive for the modern society. High accessibility to workers, suppliers, jobs, services, and other activities is fundamental for all other sectors in the economy. It is also essential for the welfare of all citizens and their everyday lives. But the transport sector also generates negative external effects such as emissions, accidents, and noise. Moreover, huge public resources have for a long time been spent on infrastructure and transit provision. Still, the public resources allocated to the transportation system is dwarfed by the enormous resources of time and money spent on transportation by citizens and firms.

Moreover, in all countries, public decisions have a vast influence on the transportation system. Transport investments, maintenance, operations, pricing, regulations, and various policy measures have huge consequences for the economy and the welfare. Pricing instruments often under public control include vehicle and fuel taxes, transit fares, congestion charges, airport charges, waterway and port charges and railway track access charges. Public regulations and administrative policies, such as emission standards and safety regulations, also have a large impact on the transportation system.

The articles in this section span over a range of topics across the core of transportation economics: cost–benefit analysis methods; pricing and financing; industrial organization, procurement and contract analysis, as well as the interaction between the transportation system and the housing and labor markets. Some articles deal with cost–benefit analysis and wider impact of transport investments. Since transport investments and other public interventions in the sector generate both positive and negative effects, appraisal methods are essential in project and policy evaluation. Moreover, proposed mega-projects such as high-speed rail and metro extensions are often justified by their anticipated effects on regional economic growth, employment, land values, and opportunities for housing construction. Other articles cover issues of pricing, governance, regulations and institutional organization issues in the transportation sector. Such topics are receiving increasing attention, for several reasons. First, physical transport infrastructure is already well developed in many countries. Second, the resources allocated to transport infrastructure is reducing in many countries, and in other countries the prices of transportation infrastructure are increasing fast. At the same time crowding and congestion are increasing, calling for greater efficiency. Third, most of the future possibilities are not directly related to physical infrastructure: autonomous cars, a fossil-free transportation system, and real-time information to travelers. Electrification and digitalization might substantially impact the benefits of infrastructure investments. So in spite of the diversity of topics covered by the articles in this section, they are all interrelated in some way or another. Since economics is part of most domains of society, the articles are crosscutting themes with articles in other sections.

Market Failures and Public Decision Making in the Transport Sector

Bruno De Borger^{*}, Stef Proost[†], ^{*}University of Antwerp, Belgium; [†]KULeuven, Belgium

© 2021 Elsevier Ltd. All rights reserved.

Introduction	2
Market Failures	2
Public Decision Making	4
Choice of Policy Instruments by a Benevolent Planner	4
Choice of Policy Instruments by the Political Process	5
Conclusion	6
References	6
Further Reading	6

Introduction

In this chapter, we look into the motivation for public intervention in the transport sector. A market economy offers certain efficiency properties, but these can only be attained when a number of strong and unrealistic conditions is satisfied. When these conditions are not satisfied, markets may fail. These market failures are the main motivation for public intervention.

In [Section 1](#), we describe the main market failures in the transport sector, and we explain why they are a problem for efficiency. In [Section 2](#), we look at public decision making. We first consider the ideal policy intervention to resolve market failures. Then we introduce more realistic policy making institutions and discuss possible political failures.

Market Failures

To understand the concept of market failures it is instructive to start from a world without such failures, that is, an ideal world with perfectly functioning markets. In *The Wealth of Nations*, Adam Smith already argued that perfectly competitive markets would lead to highly desirable outcomes. They are nowadays typically summarized in the two theorems of welfare economics ([Johansson, 1991](#); [Salanié, 2000](#)).

The first theorem states that, when a set of basic assumptions are satisfied, a perfectly competitive equilibrium is Pareto-optimal. A Pareto optimum generates maximum efficiency, in the sense that it is impossible to improve the utility of one individual without decreasing the utility of other individuals. The second theorem shows that, assuming the government has perfect income redistribution instruments, any Pareto-optimal allocation can be achieved as a perfect competition equilibrium. This means that one can achieve an efficient equilibrium that is also equitable as a perfectly competitive equilibrium.

Although the world is far from this theoretical construct, a perfect equilibrium is a good benchmark: it allows to study the public policy questions associated with the different technical conditions that have to be satisfied (these include, among others, the absence of monopoly power, public goods and externalities, see below). When these conditions are not satisfied we say there are “market failures”: the operation of perfectly competitive markets is either impossible, or it generates inefficiencies. Market failures are of course not limited to the transport sector, they exist in many other sectors as well.

We focus on two assumptions that generate important market failures in the transport sector: increasing returns to scale (and, hence, market power) and external effects. In our discussion, we initially take the location of economic activity as given. Transport activities have also an important effect on the spatial distribution of economic activities, a topic we briefly discuss at the end of this section.

Increasing returns to scale implies that the average cost of production declines as more units are produced. Under those circumstances, it is cheaper to concentrate production in one firm or facility; splitting demand over multiple firms would only lead to higher average costs. In industries with increasing returns there will, therefore, be a limited number of producers. There, private firms no longer take market prices as given, but they understand they can exploit their market power. They can increase profit by producing less and selling at prices above the marginal production cost. This is inefficient: at the production level that is chosen, consumers’

willingness to pay (the market price) is still higher than the marginal cost, so that there remain untapped possibilities to increase the utility of consumers.

The supply of transport infrastructure is often characterized by increasing returns to scale. When one firm builds a road there is an important fixed cost, and the firm can add more lanes at a cost per lane which is approximately constant. The same holds for railway stations, railway lines, airports, ports, etc. The implication of returns to scale is that it is efficient to have only one supplier of capacity in a given region for each of these infrastructures. When this provider of infrastructure then sells access services to users of the infrastructure, he has a monopoly position. Simple profit maximization leads the provider to set prices above the marginal cost of capacity. When the price elasticity of demand by users of the infrastructure is low, prices will in fact be much higher than the marginal cost. The result is an insufficient supply of capacity: the economy as a whole could have benefited from a larger transport capacity.

Insufficient infrastructure capacity is not the only problem. When there is only one provider, there can be two other problems. The first is inappropriate selection of the quality offered by the producer. The motorway surface may or may not be smooth, it may be built for different speed levels, railway tracks may be more or less noisy, etc. The quality offered by a monopolist may be too high or too low, because his choice of quality is driven by profit maximization and not by a concern for overall welfare. A second problem is an insufficient level of innovation. Monopolists have low incentives to innovate. This has been demonstrated in the telecom sector, where the opening to competition has accelerated the innovation process despite the presence of increasing returns to scale in telecom networks.

Increasing returns are much less of a problem at the level of the operators of transport services. Consider the bus industry as an example, and interpret the number of bus trips offered as the relevant output variable. It is clear that the average cost per trip offered is approximately constant, as each trip requires a bus, some energy and a driver.

The transport sector generates important *external effects*. An external effect is present when the consumption or production activities of one economic agent decrease the utility of other agents without proper compensation. If agent A enjoys a particular consumption activity that is annoying for others, then she will consume too much of the activity and will not make efforts to reduce the negative effects imposed on others. If the number of agents is small (e.g., one agent imposes an externality on one other agent) then the externality problem can be resolved through negotiation and bilateral contracts (this is known as the [Coase conjecture](#)). However, in the case of transport externalities agents impose negative effects on many others, and each agent takes the actions of others as given.

We classify the negative external effects of the use of transport infrastructure in three groups: environmental externalities, congestion externalities, and accident externalities.

Car and bus traffic, aviation and shipping all use fossil fuels, emitting greenhouse gasses and generating *climate externalities*. The use of fossil fuels is also at the origin of other *environmental externalities* under the form of conventional air pollutants that are damaging health and ecosystems (particulates, nitrogen oxide, ozone). Some transport activities also create noise, departing airplanes or old motorcycles being the best examples. The consequence of a transport sector with important environmental externalities is that, compared to the ideal situation from a social perspective, transport volumes are too high, and not enough efforts are made to limit the damaging effects for the rest of society.

Congestion is another external effect. For example, when a road is intensively used, extra road users create delays for other users. Road use is often not directly priced, or it is priced uniformly over time via fuel excises; as a consequence, then there will be periods where a too high level of demand creates delays. Similarly, when an airport is intensively used, additional incoming flights cause delays for other passengers. When all flights are offered by a monopolist, his price will internalize *congestion* (because offering extra flights imposes delays on passengers of his other flights), but the price will be too high due to his monopoly power.

The equivalent of road congestion in public transport is the discomfort externality due to *crowding*. When public transport is supplied by a profit maximizing monopolist, one expects user prices to take account of these crowding externalities; they decrease the value and quality of his product, so the monopolist prefers to avoid this as it reduces his profit. However, the price will be higher than marginal social cost due to monopoly power. Of course, public transport is in practice not priced to maximize profit. Prices are often uniform and low (or de facto free for those who pay a fixed fee per month). Crowding can then be an important externality, as every user decreases the comfort of the other users of the same bus or train. Finally, note that there is also a positive externality associated

Table 1 Market failures in the transport sector

	Returns to scale	Environmental externalities	Congestion and crowding externalities	Accident externalities
Road infrastructure supply	De facto yes			
Road use		Climate, air pollution, noise	Delay for other cars	Debatable
Rail infrastructure	Yes, important			
Rail operation and use	Limited	Limited to noise	Discomfort for other passengers	Low
Bus supply				
Bus operation and use	Limited		Discomfort for other passengers	Low
Airports	Yes		Discomfort for other passengers	
Air transport use	Limited	Climate, air pollution, noise		
Ports	Yes			
Shipping	Limited	Climate		

with public transport. More passengers lead providers to increase frequency, and this reduces the average waiting time for passengers at bus and rail stops (this is known as the Mohring effect).

Often traffic *accidents* are also mentioned as an external effect of transport use. Accidents are more complex than the other externalities for three reasons. First, the effect on the average accident risk of adding an extra user is not necessarily positive. Second, users themselves suffer from potential accidents. Third, experience rating of insurance premiums may internalize an important part of the accident cost.

In [Table 1](#), we summarize the prevalence of increasing returns to scale and transport externalities for different transport services.

To conclude this section, note that the working of the transport sector is also important for the spatial equilibrium. Perfect competition has a hard time explaining differences in economic development across space. The presence of cities and interregional trade cannot be explained without resorting to increasing returns to scale and agglomeration externalities. Cities need commuting activities to function, and increasing returns to scale can only be fully exploited when transport costs are sufficiently low. Spatial inequality is at the origin of transport activities but, conversely, the strong decrease of freight and passenger transport costs has increased the concentration of economic activity. In this sense, changes in transport costs have long-term spatial externalities. Individual import and export decisions, as well as commuting and migration decisions, affect the location of economic activities; this in turn affects the utility of many other economic agents ([Proost and Thisse, 2019](#)). The analysis of these spatial externalities requires a more complex analysis that goes beyond the ambitions of this chapter.

Public Decision Making

There are different types of problems in public decision making. First, one has to find out what are the appropriate policies. Second, one has to make sure that politicians and the administration will select these policies. One cannot just assume that governments know the appropriate policies and implement them, so that welfare improves. There is no guarantee that this will be the case.

In this section, we first concentrate on the different types of policy instruments and briefly illustrate their advantages and disadvantages, touching also upon the information problems for the policy maker. Next, we discuss the problem of selecting good policies in a political system.

Choice of Policy Instruments by a Benevolent Planner

Most of the roads are supplied by the public sector although, in some cases, the need to finance the extension of roads has driven governments to turn to the private supply of roads. Road construction as well as pricing of road use is then *de facto* a private monopoly for that road. Neither the government nor the private sector has perfect information on the future use and revenues of the road, but the private sector may have an information advantage. There are two problems to be solved by the government. First, it has to make sure the roads have the right quality and are constructed by efficient firms. Second, prices for road use should be efficient. They should cover all marginal social costs, including the external costs of congestion and environmental damage, without additional profit margins beyond what is necessary to pay for the roads' construction. One solution would be to have an auction in which the quality of the road as well as user prices are specified. The firm that offers the highest value to the government is in principle the most efficient supplier. However, this solution is not without problems. It may suffer from the winner's curse (firms underestimating costs most have the highest probability of winning, implying the winning firm is likely to make losses and go bankrupt). Moreover, it is unclear how to adapt prices to new information.

For public transport supply, we distinguish infrastructure from operations. At the level of infrastructure, returns to scale are important so that supply by one monopolist is often optimal. The optimal supply and pricing of infrastructure for rail can be addressed using the same approach as for roads, combining auctions and restrictions on prices for the use of the infrastructure. For the supply of public transport operations (offering rail and bus services) on a given infrastructure, returns to scale are less important so that one can consider several suppliers. These will be competing for offering rail and bus services, or for airport slots. Allowing different operators to use the same rail tracks is now well accepted in most countries, but it is an important change compared to the times where the public transport infrastructure supply and operation was vertically integrated. The allocation of slots can again be done using auctions. Such auctions are, however, complex because operators have constraints that are ideally solved using combined auctions: an operator may attach a higher value to a combination of time slots that optimizes his timetable.

One distinguishes usually three types of policy instruments to address *externalities* associated to the use of a transport good: prices, tradeable permits, and regulations. The instruments can also be used in combination.

The optimal policy toward an externality has three properties. Take the example of a polluting car. First, the cost of making the car greener should be equal to the marginal benefit of reducing the environmental damages of the use of the car. Second, the user cost of a car trip (where the car has the optimal level of "greenness") should include the remaining pollution damage cost). Third, the policy should contain incentives for technological progress. The optimal environmental policy is always a combination of reducing the polluting activity, making the activity cleaner and stimulating better technologies. To reach this optimum the policy should focus as directly as possible on the externality itself.

Consider how different instruments address environmental and congestion externalities. Bear in mind that the cost of making an activity cleaner is better known by the polluters than by the policy maker.

The main environmental externalities are associated to road use, aviation, and shipping. *Climate externalities* can be addressed via excise taxes that are proportional to the carbon content of the fuel. This policy instrument can also guarantee cost efficiency across different sectors of the economy, as the origin of the carbon emissions does not matter for the damage. Therefore, the same tax on carbon should prevail in all sectors of the economy. The second type of instrument that can be used are tradable emission permits for carbon. In this case, the total quantity of emissions in the transport sector is fixed, the rights are distributed or auctioned, and a market for the trade of permits makes sure the marginal cost of emission reduction is identical across sectors. Except for the transaction costs, this instrument is as efficient as the tax instrument if the total emission reduction achieved is the same. Note that, in the case of taxes or tradable permits, the policy maker does not have to know the costs of emission reduction of the different polluters. As long as the tax or permit price is the same for all polluters, the emission reduction is organized in a cost efficient way thanks to the working of the price mechanism. The tax and permit price is also a stimulus for technological progress as long as there is a long-term commitment to keep the emission tax or permit price high enough. The third instrument is a standard on emissions; this typically takes the case of a fuel efficiency standard. However, the government lacks the information on the costs of emission reduction, so that a standard on emissions is less cost-effective to reduce emissions. Even when the fuel efficiency standard in a subsector like cars is tradable among manufacturers, it is still less efficient because the diversity in the use of vehicles is not taken into account.

Besides these three basic instruments, many other policy instruments to tackle climate change are possible (and are used): subsidies for low emission vehicles or modes, land-use policies, etc. Such instruments are generally less efficient: they do not offer the right balance between making the transport activity greener and reducing the transport activity in function of the remaining emissions.

Other pollutants can be addressed using the same instruments. The main difference is that their damage is more localized, so that damage increases with population density. Spatial differentiation of the stringency of policies is therefore needed, but this is difficult to achieve. Moreover, the emissions of conventional air pollutants are more difficult to measure. This is the reason why governments turn to second-best policies like emission standards or low emission zones, where only certain types of vehicles are allowed to operate.

External *congestion* costs require congestion fees that are differentiated across space and time. These tax or toll instruments serve to make road users pay for the additional delays they cause to other users. An alternative policy is to use tradable driving rights that are differentiated over space and time. Whenever they achieve the same allocation of traffic as the optimal congestion tax—and in the absence of uncertainty—the driving rights are as efficient as the congestion tax. Other known policies to address congestion are much less efficient, including adding road capacity, subsidizing public transport, etc.

Problems of implementation and acceptance imply that most policy instruments to address externalities are nonprice measures that largely escaped the attention of transport economists: speed bumps, speed restrictions, pedestrian zones, etc.

Choice of Policy Instruments by the Political Process

There is as yet no general theory of the political process. Political scientists and economists developed simplified models that give useful insights for specific problems, but the political process can itself be very complex.

It is commonly believed that market failures always justify government intervention. However, the decision-making process itself and the distinction between the legislative and executive branches of government imply a number of potential inefficiencies. These are referred to as political failures. We provide in this section a few illustrations of political failures based on an simplified model.

Take the case of a country with two homogeneous groups of citizens. There are λN citizens of type A and $(1 - \lambda)$ citizens of type B. Consider a transport project (road or public transport) that costs C and benefits only citizens of type A. Group A could be car users having each a benefit b of using the new road, while group B could be citizens without a car. Alternatively, A and B could represent citizens of two regions or cities in a federation. We further assume the project is paid by a uniform lump-sum tax. So every citizen pays C/N .

The project is justified in cost–benefit terms when total benefits are larger than total costs, so when

$$\lambda N \cdot b > C \quad \text{or} \quad b > C/\lambda N.$$

Consider now first simple *majority voting* to make decisions. We immediately note two potential types of political failures. First, when group A has the majority, it may accept the project even when the project is socially undesirable. This group will decide to accept the project if the benefit b for each of its citizens satisfies $b > C/N$. If the project is such that $C/N < b < C/\lambda N$, group A will accept an inefficient project. This occurs because group A benefits but the costs of the projects are spread over the whole population. Second, when group A has no majority, it will never get a good project accepted, because the citizens of group B will not agree to pay for projects that do not benefit them. The voting process is capable of accepting some of the good projects (those having large benefits for the majority) and stopping some of the bad projects (those having very low net benefits for the majority and for the minority), but the process is not very selective. Some good projects will be rejected, some bad projects will be accepted. The reason is twofold: the intensity of benefits is not measured, and the minority is discriminated against. Note that lobbying and rent-seeking behavior may cause further inefficiencies in decision making.

Many political decisions are taken in a *representative democracy* system. Each of the regions has one representative and they decide on the allocation of projects. One possible equilibrium has an agenda setter that forms a minimum winning coalition of

representatives of regions that have low cost projects so that he can select a large project for his region without losing the majority. This set up has been used to explain the allocation of federal highway funds over the different states in the US. It was found that the allocation of funds was highly inefficient: for each dollar spend on highways, there was a dollar wasted.

When public funds have to be allocated over different regions and there are spillovers between regions, one needs additional restrictions on the decision process. When constitutional restrictions can be built into the pricing and investment of transport infrastructure, the outcomes can be relatively efficient. One such restriction could be that pricing should be uniform across regions.

Introduce now *politicians* in the decision process. Decisions are taken by a majority but they are executed by professional politicians and agencies. Consider again one project benefitting citizens in A only. However, assume that the quality of the politician matters: a good politician can realize the project for a cost lower than a bad politician: $C_L < C_H$. The bad politician may be lazy or may be more interested in capturing personal benefits rather than exerting effort for the public project. Assume furthermore that the project will only pass the cost-benefit criterion when a good politician is chosen. But selecting a good politician is difficult as good and bad politicians are indistinguishable *ex ante*. Often one needs to observe them for at least one period before knowing the politician's quality. Moreover, suppose a bad politician can pretend to be a good one in the first period; they have incentives to do so because they know they can get some personal rents doing a poor job in the second period. This then gives yet another example of a political failure: a good project selected in the first period may be ruined by the bad politician in the second period.

Conclusion

Transport activities suffer from important market failures, including increasing returns that lead to market power, and a series of external costs, including climate change, pollution, congestion, and accident risks. We reviewed the socially optimal policy instruments (pricing, regulation, use of permits) needed to cope with these market failures. Moreover, we discussed the possible political failures policy makers encounter when trying to implement desirable transport policies in a democratic system. Despite problems of possible political failures due to the voting process, due to lobbying, due to common pool financing and due to bad politicians, the democratic political process is still often qualified as the most desirable (the least bad) selection process. It allows to select some really good projects and avoid some really bad projects, and it may help in selecting good politicians.

References

- Coase, R.H., 1960. The problem of social cost. *J. Law Econ.* 3, 68–111.
 Johansson, P.-O., 1991. *An Introduction to Welfare Economics*. Cambridge University Press.
 Proost, S., Thisse, J.F., 2019. What can be learned from spatial economics? *J. Econ. Lit.*
 Salanié, B., 2000. *The Micro-Economics of Market Failures*. MIT Press, Cambridge, MA.

Further Reading

- Armstrong, M., Sappington, D.E.M., 2006. Regulation, competition and liberalization. *J. Econ. Lit.* 44 (2), 325–366.
 Besley, T., 2006. *Principled Agents? The Lindahl Lectures*. Oxford University Press.
 De Borger, B., Proost, S., 2013. Traffic externalities in cities: the economics of speed bumps, low emission zones and city bypasses. *J. Urban Econ.* 76, 53–70.
 De Borger, B., Proost, S., 2016. Can we leave road pricing to the regions? The role of institutional constraints. *Reg. Sci. Urban Econ.* 60, 208–222.
 de Palma, A., Proost, S., Seshadri, R., Ben-Akiva, M., 2018. Congestion tolling—dollars versus tokens: a comparative analysis. *Transport. Res. Part B* 108, 261–280.
 Knight, B., 2004. Parochial interests and the centralized provision of local public goods: evidence from congressional voting on transportation projects. *J. Public Econ.* 88 (3–4), 845–866.
 Kolstad, C., 2011. *Environmental Economics*. Oxford University Press.

Demand for Freight Transport

José Holguín-Veras*, **Diana G. Ramírez-Ríos[†]**, ^{*}Department of Civil and Environmental Engineering; Center for Infrastructure, Transportation, and the Environment; VREF Center of Excellence for Sustainable Urban Freight Systems, Rensselaer Polytechnic Institute, Troy, NY, United States; [†]Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, NY, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	7
Urban Economies	7
Freight Generation	9
Empirical Estimates	9
Conclusion	10
References	12

Introduction

The demand for freight supplies is a physical expression of the economy, which is an obvious consequence of the fact that large portions of economic transactions entail the exchange of money for physical supplies. In this context, by transporting the supplies from points of production to points of consumption, supply chains close the loop in economic exchanges, and enable households and businesses to use the supplies in the manner desired. Since the vast majority of human or economic activities need supplies of one form or another, supply chains are pervasive. At the root of this process, one finds the demand for supplies at the receiving locations.

To gain insight into freight demand, it is important to understand the role of supply chains. Modern supply chains tie together multiple production and consumption stages that typically start with raw or recycled input materials, and end with the shipment of products for final consumption (Holguín-Veras et al., 2017). In all cases, an agent produces and/or sends supplies (the shipper) that are then consumed by a different agent (the receiver), after they are transported by the carrier. At each one of these stages supplies are consumed, transformed, produced, or stored. In modern times, with the surge of e-commerce, individuals and households have become integral parts of business-to-consumer supply chains. Compounding the inherent challenges, the logistics industry must now ensure efficient flows of supplies to both commercial establishments and households. Notwithstanding their importance, data about freight demand are very hard to get, as there are only a handful of publicly available estimates.

This paper attempts to fill this void by describing empirical evidence concerning freight demand. To do so, the authors gathered data and results from several sources representing developed and developing countries, different levels of geography (i.e., multi-national, national, and metropolitan), and a time span from the 1960s to current times. The data collected were post-processed to estimate the per-capita freight generation (FG), that is, the total amount of freight transported by type of commodity, divided by the corresponding population. The resulting values were analyzed to identify similarities and differences.

This paper has four sections in addition to this introduction. Section “Urban Economies” provides a brief overview of urban economies and their composition in terms of industry sectors. Section “Freight Generation” discusses the concepts of FG and freight trip generation (FTG). Section “Empirical Estimates” analyzes the empirical estimates of FG for the different cases considered in this paper. Section “Conclusion” summarizes the chief insights of this paper.

Urban Economies

The most straightforward way to illustrate the importance of freight activity and supply chains is to identify the sectors of the economy that, directly or indirectly, depend on supply chains to perform their activities. To this effect, one can create two major clusters of industry sectors. The first, freight-intensive sectors (FIS), corresponds to the industry sectors for which the production and consumption of freight is an indispensable component of their economic activities. The second cluster, service-intensive sectors (SIS), represents those sectors where the provision of services is the primary activity and the production or consumption of freight supplies is of secondary importance. It should be noted that both FIS and SIS produce and consume supplies, though the amounts generated by SIS are smaller than those generated by FIS. The industry sectors, defined using the North America Industry Classification System (NAICS), included in FIS and SIS are as follows:

- Freight-intensive sectors (FIS)
 - NAICS 11: Agriculture, forestry, fishing, and hunting
 - NAICS 21: Mining, quarrying, oil/gas, etc.
 - NAICS 22: Utilities
 - NAICS 23: Construction

- NAICS 31-33: Manufacturing
- NAICS 42: Wholesale trade
- NAICS 44-45: Retail trade
- NAICS 48-49: Transportation and warehousing
- NAICS 72: Accommodation and food services
- Service-intensive sectors (SIS):
 - NAICS 51: Information
 - NAICS 52: Finance and insurance
 - NAICS 53: Real estate and rental and leasing
 - NAICS 54: Professional, scientific, and technical services
 - NAICS 55: Management of companies
 - NAICS 56: Administrative, support, waste management, etc.
 - NAICS 61: Educational services
 - NAICS 62: Healthcare and social assistance
 - NAICS 71: Arts, entertainment, and recreation
 - NAICS 81: Other services
 - NAICS 92: Public administration

To illustrate the relative importance of FIS and SIS, the authors analyzed establishment and employment data by industry sector for all of the metro/micropolitan areas—defined by the US Census Bureau as “a core area containing a substantial population nucleus, together with adjacent communities having a high degree of economic and social integration with that core” (US Census Bureau, 2017)—in the United States (US Census Bureau, 2013a). Metropolitan areas are those with more than 50,000 people, while micropolitan areas have between 10,000 and 49,999 people (US Census Bureau, 2013b). The totals by industry sector are shown in **Table 1** and **Table 2** (Holguin-Veras and Aros-Vera, 2016). The results show that FIS capture 44.7% of commercial establishments (**Table 1**), and 49.4% of the employment in the United States; the rest corresponds to SIS (**Table 2**). It is worth noting that transportation and warehousing accounts for only 2.8% of the establishments and 3.6% of employment (these numbers likely underestimate the activity, as they do not include private fleets). These numbers imply that the performance of the transportation

Table 1 Establishments by industry sector

<i>Freight-intensive sectors</i>				
<i>% of establishments in United States</i>				
NAICS	Industry sector	Metro	Micro	Total
44	Retail trade	12.5%	1.5%	14.0%
72	Accommodation and food services	8.1%	0.9%	9.0%
23	Construction	7.8%	0.9%	8.7%
42	Wholesale trade	5.3%	0.4%	5.6%
31	Manufacturing	3.4%	0.4%	3.9%
48	Transportation and warehousing	2.5%	0.3%	2.8%
21	Mining, quarrying, and oil and gas extraction	0.2%	0.1%	0.3%
11	Agriculture, forestry, fishing, and hunting	0.2%	0.1%	0.2%
22	Utilities	0.2%	0.0%	0.2%
	% Total	40.2%	4.6%	44.7%
<i>Service-intensive sectors (SIS)</i>				
<i>% of establishments in United States</i>				
NAICS	Industry sector	Metro	Micro	Total
54	Professional, scientific, and technical services	11.3%	0.6%	11.9%
62	Healthcare and social assistance	10.4%	1.0%	11.4%
81	Other services, except public administration	8.7%	1.0%	9.7%
52	Finance and insurance	5.7%	0.5%	6.3%
56	Administrative and support, waste management	5.0%	0.4%	5.3%
53	Real estate and rental and leasing	4.5%	0.4%	4.9%
51	Information	1.7%	0.1%	1.8%
71	Arts, entertainment, and recreation	1.5%	0.2%	1.7%
61	Educational services	1.3%	0.1%	1.3%
55	Management of companies and enterprises	0.7%	0.0%	0.7%
99	Unclassified	0.1%	0.0%	0.2%
	% Total	51.0%	4.3%	55.3%

Table 2 Employment by number of employees per establishment

Number of employees	% of total	Freight-intensive sectors (FIS)			Service-intensive sectors (SIS)		
		% of employment in United States			% of employment in United States		
		Metro	Micro	Total	Metro	Micro	Total
1–4	7.7%	3.4%	0.4%	3.8%	5.4%	0.5%	5.9%
5–9	7.3%	3.2%	0.4%	3.6%	3.8%	0.4%	4.1%
10–19	12.2%	5.4%	0.6%	6.0%	4.9%	0.4%	5.4%
20–49	21.7%	9.7%	1.0%	10.7%	7.1%	0.5%	7.6%
50–99	15.5%	7.1%	0.6%	7.7%	5.4%	0.3%	5.7%
100–249	18.5%	8.4%	0.7%	9.1%	8.0%	0.4%	8.4%
250–499	9.2%	4.0%	0.5%	4.5%	4.7%	0.3%	5.0%
500–999	5.1%	2.2%	0.3%	2.5%	4.1%	0.3%	4.4%
>1000	2.8%	1.2%	0.1%	1.4%	3.9%	0.1%	4.0%
Average	0.1	5.0%	0.5%	5.5%	5.3%	0.4%	5.6%
SD	0.1	2.7%	0.2%	3.0%	1.4%	0.1%	1.4%
Total		44.7%	4.7%	49.4%	47.4%	3.2%	50.6%

and warehousing sector—a relatively small portion of the employment and establishments in the country—directly impacts about half the US economy, and indirectly impacts the other half. In developing economies, where the service economy is less developed, the impacts of the transportation and warehouse sector on the overall economy are even larger, because of the larger role played by the FIS. In Bangladesh, for instance, the authors estimate that the share of FIS employment is about 70% (Holguín-Veras et al., 2018).

Freight Generation

The production and consumption of cargo, or FG, should not be confused with the production and consumption of vehicle trips, or FTG. FG is an expression of the economic activity performed at businesses, by which input materials are consumed, transformed, or stored. FTG, in contrast, is the result of the logistic decisions concerning how best to transport the FG in terms of shipment size, frequency of deliveries, and vehicle/mode used (Holguín-Veras et al., 2017).

FG refers to the amount of cargo produced and consumed in the study area at the establishment or zonal level. Freight production (FP) represents the amount produced, while freight attraction (FA) is the amount of freight consumed. FG is the sum of both FA and FP. FTG is less straightforward, as it is the result of the joint decision of shipment size and frequency, which enables large business establishments to receive larger shipments, minimally increasing the amount of vehicle trips produced (Holguín-Veras et al., 2014).

Distinguishing FG from FTG makes it easier to identify what should be the primary focus of freight policy. From the economic point of view, the production and consumption of supplies are beneficial activities that help satisfy the needs of consumers and businesses. At the same time, the freight traffic that these activities generate is associated with large amounts of negative externalities—pollution, congestion, and accidents—that ought to be mitigated or eliminated altogether (Ogden, 1977). The chief objective of freight policy should be to ensure that the flows of supplies are as efficient as possible, and that the resulting freight traffic produces as little negative externalities as practically and economically possible.

Gaining insight into FG patterns is important as they provide a revealing snapshot of the corresponding economy. Given the well-known interconnections between the economy and freight activity, the amount of cargo generated within a region (the FG) provides insights into the region's economy. To facilitate comparison of results, without having to account for the difference in geographical scales and income, the authors computed a per-capita FG, which was obtained as the total amount of freight transported in a typical day divided by the corresponding population, and is expressed as kg/person per day. This simple metric provides an intuitive understanding of the amount of freight handled within a given jurisdiction that reflects the effect of income.

Empirical Estimates

This section presents estimates of per-capita FGs for different countries, geographic scales, and time periods. The source materials represent three different levels of geography, six different countries (both developed and developing), and different survey methodologies. These sources correspond to:

- Multinational: The 28 countries of the European Union (EU-28) (Eurostat, 2017)
- National: United States (US Census Bureau, 2018), China (China Transportation Department, 2013; China National Bureau of Statistics, 2018), Sweden (Transport Analysis, 2019), Colombia (Banco Interamericano de Desarrollo, 2013), Bangladesh (Holguín-Veras et al., 2018), and the Dominican Republic (Holguín-Veras, 1984)
- Metropolitan: The tri-state area (i.e., New York, New Jersey, and Connecticut) (Wood, 1970) and Medellín (Colombia) (Gonzalez-Calderon et al., 2018)

To ensure comparability of results, the original estimates were reclassified using the Standard Classification of Transported Goods (SCTG) (Bureau of Transportation Statistics, 2017). However, doing so was not straightforward, as the sources analyzed here adopted different definitions of commodity types that do not always line up well with the SCTG. Thus, the reclassification was far from perfect. Another source of discrepancies is the differences in scope of the studies; in some cases, such as the United States and EU, all modes were considered, while in other studies—Bangladesh, Colombia, and the Dominican Republic—only the highway modes were accounted for. In the latter case, for instance, the flows of coal and oil transported by water modes and pipelines, without using highways at all. Thus, the values of per-capita FGs presented here should be interpreted as order-of-magnitude estimates. The results are shown in Table 3. To facilitate referencing, the authors adopted the nicknames underlined in Table 3 to refer to specific SCTGs.

Table 3 also shows the ways in which the data were collected. “Mixed surveys” indicates that the data were obtained by means of surveys that targeted multiple agents (e.g., carriers, shippers, or receivers). “Surveys + models” refers to the technique developed by the authors and colleagues, successfully pilot-tested by Holguín-Veras et al. (2018). The most unique aspect of this approach is that the data collected are used to estimate models that, when applied to public data such as an economic census, estimate the FG for the study area (instead of estimating FG using only data collection, which requires larger samples). The net result of combining models and data is a dramatic reduction in the amount of data that needs to be collected.

The most striking feature of Table 3 is the huge difference between the per-capita FG for the large economies (i.e., United States, China, and EU-28) and the rest. The estimates show that per-capita FGs for the large economies are on average 3 times larger than those for Sweden (a high-income country with a vibrant export economy), and several times larger than those for the rest of the developing countries. This is a reflection of the size of the internal trade in these large regions, where large amounts of supplies are transported to other locations as inputs to subsequent stages of production processes. Another related factor is that the larger the geographic area under study, the larger the possibility of considering all the internal flows of cargo generated by the economy. For instance, the flows of coal and natural gas used to generate electricity end at power plants; from there, the electricity moves by power lines to rest of the country and metropolitan areas. As a result, freight surveys conducted in metropolitan areas typically do not include the transportation of the coal and gas used for electricity generation, while regional and national surveys do.

The results show that, generally speaking, the higher the income, the larger the total per-capita FG. As shown, the total per-capita FGs for the United States, EU-28, and Sweden are significantly larger than those for developing countries, which the exception is China. Again, this result confirms the interconnection between the economic activity and wealth, and the FG. A complementary perspective is provided by the per-capita FG for metropolitan areas, as they reveal the economic importance of their importance. As shown, large portions of the per-capita FG in metropolitan areas are associated with manufacturing, industrial, and construction activities.

The results for the various commodity types show that, despite the different scales and time periods, there is consistency among the estimates. In most cases, the per-capita FGs for the large economies are larger than those for the rest. Only in two cases, the per-capita FGs for metropolitan areas occupy one of the top two positions for a commodity group (i.e., base metal in Medellín and electronics in the tri-state area). In all other cases, the large economies occupy the top two positions. Reflecting the size of China’s manufacturing industry, this country has the highest per-capita FG in four out of five commodity groups (SCTGs 25-43), while the United States and EU capture the top positions in the others groups (SCTGs 01-24).

Among the developing countries studied, the per-capita FGs are in the same order of magnitude. However, and not surprisingly, the ones for Colombia—the most developed country in the group—are consistently higher than the ones for Bangladesh and the Dominican Republic. The results for the metropolitan areas are also consistent, notwithstanding the differences in levels of development and time.

Conclusion

The results presented in this paper provide a wide-ranging perspective on the demand for freight at different parts of the world, levels of geography, and time periods. Although hampered by the lack of publicly available data, differences in data collection methodologies, and unavoidable sampling and data errors, the results are remarkably consistent. Probably the most significant finding is related to the similarities between the per-capita FGs for geographic areas of similar size and income levels. As shown, most commodities exhibit rather similar values of per-capita FG. The cases where there are significant differences are typically the result of those commodities being major export products, which causes the per-capita FGs to be larger than usual.

The observation that per-capita FG generally increases with income has major implications, as it implies that, particularly in developing countries, economic improvements that lead to increasing income will produce large increases in the amount of freight produced and consumed and, subsequently, in the corresponding freight traffic. This, in turn, will aggravate congestion and environmental issues above and beyond what could be reasonably expected from the natural growth of population. Complicating matters even further, the surge of business-to-consumer transactions powered by e-commerce—accompanied by frequent deliveries of small shipments—is bound to increase congestion and emissions. There is a critical need to proactively find ways to ensure that freight activity can be conducted efficiently, but with minimal negative externalities.

Table 3 Per-capita freight generation (kg/person per day)

		<i>United States (2017)</i>	<i>China (2013)</i>	<i>EU-28 (2017)</i>	<i>Sweden (2017)</i>	<i>Colombia (2013)</i>	<i>Bangladesh (2017)</i>	<i>Dominican Republic (1982)</i>	<i>New York, New Jersey, and Connecticut (1963)</i>	<i>Medellin, Colombia (2018)</i>
<i>Geography</i>		<i>Nation</i>	<i>Nation</i>	<i>Multination</i>	<i>Nation</i>	<i>Nation</i>	<i>Nation</i>	<i>Nation</i>	<i>Metro</i>	<i>Metro</i>
<i>SCTG code</i>	<i>Methodology Commodity type</i>	<i>Shipper survey</i>	<i>Shipper survey</i>	<i>Mixed surveys</i>	<i>Mixed surveys</i>	<i>Carrier survey</i>	<i>Survey + models</i>	<i>Mixed surveys</i>	<i>Carrier survey</i>	<i>Mixed surveys</i>
01-05	Agriculture products and fish	7.31	6.40	9.34	8.47	6.02	1.17	1.60	0.97	3.10
06-09	Grains, alcohol, and tobacco	22.37	3.15	12.91	2.03	1.69	0.16	0.89	5.04	2.49
10-14	Stones, nonmetallic minerals, and metallic ores	37.85	16.14	38.12	6.71	2.92	0.92	0.24	10.67	3.85
15-19	Coal and petroleum products	28.15	12.30	0.86	7.06	0.04	-	0.09	5.93	0.25
20-24	Basic chemicals, chemical, and pharma- ceutical products	9.47	4.03	4.30	1.22	2.33	0.18	0.15	6.17	2.60
25-30	Logs, wood products, textile, and leather	4.83	22.04	4.77	3.66	0.12	0.13	-	-	1.69
31-34	Base metal and machinery	4.95	8.95	4.09	1.83	0.37	0.18	1.73	-	7.22
35-38	Electronic, motor vehicles, and precision instruments	1.69	5.90	1.91	0.67	0.35	0.06	-	3.59	1.12
39-43	Furniture, mixed freight, miscellaneous manufactured products	3.24	8.27	10.62	0.32	1.23	1.45	0.39	4.64	1.51
	Other goods, not previously specified	3.10	11.22	2.70	1.53	0.32	0.06	0.34	0.40	2.97
Totals		122.96	98.40	89.62	33.49	15.39	4.32	5.43	37.41	26.81

Note: The estimates for agriculture in the United States include grains and feeds; sugar and derivatives of sugar; soy, flaxseed, etc.; vegetables and roots; fruits and nuts; cattle, chicken, etc.; and milk and derivatives (United States Department of Agriculture (USDA) and National Agricultural Statistics Service, 2017).

References

- Banco Interamericano de Desarrollo, 2013. Transporte Carretero de Carga en America Latina y El Caribe: Estudios de Pais. BID, Bogotá, Colombia.
- Bureau of Transportation Statistics, 2017. Standard Classification of Transported Goods (SCTG) codes. Available from: https://www.bts.gov/archive/publications/commodity_flow_survey/classification.
- China National Bureau of Statistics, 2018. 2018 China statistical yearbook. Available from: <http://www.stats.gov.cn/tjsj/ndsj/2018/indexch.htm>.
- China Transportation Department, 2013. Special survey of highway and waterway transport volume, Ministry of Transport. Beijing, P.R. China.
- Eurostat, 2017. Road freight transport by group of goods, EU-28, 2013-2017. Available from: [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Road_freight_transport_by_group_of_goods_EU-28_2013-2017_\(thousand_tonnes_and_million_tonne-kilometres\)-upd.png](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Road_freight_transport_by_group_of_goods_EU-28_2013-2017_(thousand_tonnes_and_million_tonne-kilometres)-upd.png).
- Gonzalez-Calderon, C.A., et al., 2018. Characterization and analysis of metropolitan freight patterns in Medellin, Colombia. *Eur. Trans. Res. Rev.* 10 (23), 1–11, doi:10.1186/s12544-018-0290-z.
- Holguín-Veras, J., 1984. Desarrollo de un Modelo para Cuantificar la Oferta Vehicular en el Transporte de Carga. Universidad Central de Venezuela, Instituto de Urbanismo, Santo Domingo, 2 v.
- Holguín-Veras, J., Aros-Vera, F., 2016. Potential market of freight demand management. 2017 TRB Annual Meeting. Transportation Research Board, Washington, DC.
- Holguín-Veras, J. et al., 2014. Freight generation and freight trip generation models. In: Tavasszy, L., De Jong, G. (Eds.), *Modeling Freight Transport*. Elsevier, London.
- Holguín-Veras, J. et al., 2017. Using commodity flow survey and other microdata to estimate the generation of freight, freight trip generation, and service trips: guidebook. Transportation Research Board; National Cooperative Highway Research Program/National Cooperative Freight Research Program Transportation Research Board of the National Academies. Available from: <https://www.nap.edu/catalog/24602/using-commodity-flow-survey-microdata-and-other-establishment-data-to-estimate-the-generation-of-freight-freight-trips-and-service-trips-guidebook>.
- Holguín-Veras, J., et al., 2018. Bangladesh Freight Study. The World Bank, Washington, DC.
- Ogden, K.W., 1977. Modelling urban freight generation. *Traffic Eng. Control* 18 (3), 106–109.
- Transport Analysis, 2019. Swedish commodity flow survey (2016). Available from: <https://www.trafa.se/en/travel-survey/commodity-flows/>.
- United States Department of Agriculture (USDA) and National Agricultural Statistics Service, 2017. Agricultural statistics 2017. Available from: https://www.nass.usda.gov/Publications/Ag_Statistics/2017/index.php.
- US Census Bureau, 2013. Population change for metropolitan and micropolitan statistical areas in the United States and Puerto Rico (February 2013 Delineations): 2000 to 2010 (CPH-Ts). 2010 Census Population and Housing Tables. US Census Bureau.
- US Census Bureau, 2013. Revised delineations of metropolitan statistical areas, micropolitan statistical areas, and combined statistical areas, and guidance on uses of the delineations of these areas. Available from: <http://www.whitehouse.gov/sites/default/files/omb/bulletins/2013/b13-01.pdf>.
- US Census Bureau, 2017. About metropolitan and micropolitan statistical areas. Metropolitan and Micropolitan. Available from: <http://www.census.gov/population/metro/about/>.
- US Census Bureau, 2018. 2017 Commodity flow survey (CFS): data releases. Available from: <https://www.census.gov/programs-surveys/cfs/news/updates/upcoming-releases.html>.
- Wood, R.T., 1970. Measuring freight in the tri-state region. In: *The Urban Movement of Goods*. OECD, Paris, pp. 61–82.

Cost Functions for Road Transport

José Manuel Vassallo, Transport Research Centre (TRANSyT), Universidad Politécnica de Madrid, Madrid, Spain; Centro de Investigación del Transporte (TRANSyT), ETSI de Ingenieros de Caminos, Canales y Puertos, Madrid, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	13
The Concept of Cost	13
Fixed Versus Variable Costs	13
Short-Run Versus Long-Run Costs	14
Infrastructure, Vehicle Operation, and Personal Costs	14
Infrastructure Costs	14
Vehicle Operation Costs	14
Personal Costs	14
Internal, External, and Social Costs	15
Internal Costs	15
External Costs	15
Social Costs	15
Cost Functions	15
Total Cost	15
Average Versus Marginal Cost and the Impact of Congestion	16
Supply Versus Demand Equilibrium	17
Economies of Scale and Scope	18
Cost Calculation, Allocation, and Optimization	18
Cost Calculation	18
Cost Allocation	18
Life-Cycle Cost	18
References	19

Introduction

Road transportation is undoubtedly one of the most important modes in terms of the volume of passengers and freight moved all over the world, its contribution to GDP, and its impact on families' expenditure. It has also a key role in energy consumption, emissions, and contribution to climate change. Determining road cost functions is hence crucial for conducting a rational planning, and adopting policy measures aimed at improving the competitiveness of the economy and reaching a higher quality of life.

This paper begins with a brief explanation of the concept of cost and its different characteristics. Then, it makes a classification of road costs differentiating between infrastructure, vehicle operation, and personal costs. After that, it clarifies the difference between internal and external costs, and, on the basis of that difference, explains the concept of social costs. Subsequently, it describes the characteristics of the different road cost functions. This paper finishes with a definition of the concepts of economies of scale (EOSs) and scope applied to the road sector, and a description of the methods to calculate, allocate, and optimize the life-cycle cost of a road network.

The Concept of Cost

Costs are incurred when scarce resources are consumed, used, or worn. The loss or deterioration of these resources implies a reduction of utility for those who own them, and, as a consequence of that, for the society as a whole. Costs are sometimes valued using the concept of opportunity cost, which may be defined as the lost benefit of the next best activity forgone. In transportation economics, the value of travel time and reliability is often estimated on the basis of the opportunity cost ([Button and Verhoef, 1998](#)).

Fixed Versus Variable Costs

Costs can be split into fixed and variable ones. Fixed costs are those that do not depend on the output. Considering that the output of a road is the amount of vehicles that use it over its life cycle in vehicle kilometers (veh-km), and assuming that the capacity of the road is designed to provide a good level of service over its life cycle, the capital cost necessary to build the facility does not depend on the traffic volume so it may be considered a fixed cost.

Variable costs are those that depend on the output (veh-km). For instance, the wear of the pavement will depend on how many vehicles use the road, especially those with heavy axle weight. Similarly, fuel consumption and CO₂ emissions will be directly related to the amount of fuel fossil vehicles using the road. These items are classified as variable costs.

Short-Run Versus Long-Run Costs

The concept of fixed and variable costs depends on the time frame considered. The capital cost incurred is considered fixed in the short run, but not in the long run. To give an example, a highway originally built with two lanes per direction may provide a good quality of service for a period of time. However, let us say 10 years later, it may be necessary to add capacity by building an additional lane to guarantee a good traffic flow for another amount of years. Cost functions can be calculated for both the short and the long run. Long-run cost functions are usually calculated as the envelop of the short-run ones.

Infrastructure, Vehicle Operation, and Personal Costs

Getting into greater detail in the case of road costs, another distinction can be made among infrastructure, vehicle operation, and personal costs. Infrastructure costs refer to the costs necessary to keep the facilities available over the life cycle of the road. Vehicle operation costs refer to the cost directly incurred by the vehicles that use the infrastructure. Personal costs have to do with the loss of people utility associated to traveling.

Infrastructure Costs

Infrastructure costs can be divided into capital costs (usually known as CAPEX) and maintenance and operation costs (usually known as OPEX). Capital costs do not just refer to the mere construction of the facility, but also include other costs such as those related to the planning and design of the facility. These costs do not only comprise civil works, but also other auxiliary services such as intelligent transportation systems, etc. CAPEX are usually considered fixed in the short run, but they may experience changes in the long run. Capital costs are spent in a short period of time, but render service over a long period of time. This is the reason why, for accounting purposes, this cost is usually depreciated over the life span of the project.

The use of the road over the years along with weather conditions contributes to the wear and tear of the road assets. Maintenance costs intend to preserve the value of the assets over time to provide the right service to the users. They include both regular periodic maintenance and rehabilitation works. Periodic maintenance activities are those conducted on a regular basis such as, for instance, cleaning the ditches and drainages, painting signs, correcting minor deficiencies of the pavement, etc. Major rehabilitation works are conducted every several years, and they are aimed at preserving the patrimonial value of the assets ([Small et al., 1989](#)). Structural strengthening works to reinforce the pavement through additional asphalt layers and main works to repair bridges are considered major rehabilitation works.

Road operators incur other costs aimed to guarantee the availability of the infrastructure and its optimal functioning over time. These operation activities include snow removal, attention to incidents, withdrawal of objects in the road, signal management, etc.

Vehicle Operation Costs

Vehicle operation costs include the consumption of resources or damages to the environment and the society stemming from the use of the road by different types of vehicles. They comprise depreciation and financial cost of the vehicles, repairs, fuel and oil consumption, and tire wear. These costs depend not only on the characteristics of the vehicle, but also on the specific traffic flow conditions and the state of the road. For instance, fuel consumption per kilometer gets its optimal rate at a speed ranging from 40 to 90 km/h depending on the type of vehicle. Therefore, in congested roads, fuel consumption per kilometer gets higher. Similarly, deteriorated roads with high roughness levels bring about larger vehicle operation costs because of more frequent repairs, higher fuel consumption, and a more rapid depreciation of the vehicle. The HDM-IV model promoted by the World Bank provides relationships between the state of the road and the operation costs of the vehicles ([Paterson, 1987](#)).

Vehicle fleets of buses and trucks used by transport companies for business purposes also include the labor cost of drivers or other personnel necessary for carrying out their activities. These companies also incur indirect costs stemming from the managerial activities of the company. Indirect costs cannot be assigned to a specific output since they are common to all the activities of the firm.

Personal Costs

Personal costs comprise the individuals' opportunity cost of not being able to do something that has greater utility for them at that time. In the case of individual drivers or public transport passengers, this refers to the value of travel time and reliability ([De Palma et al., 2011](#)). The time spent within a vehicle cannot be used for other activities such as working, practicing sports, or going shopping. Again, the congestion of the road will imply trip delays that will directly impact the cost of time and reliability, thereby reducing users' utility.

Internal, External, and Social Costs

Internal Costs

Internal costs are defined as the costs directly borne by the users of the road—individual drivers, passengers, and transportation companies. They include most of vehicle operation costs such as fuel, vehicle depreciation, etc., and also personal costs such as travel time, reliability, etc. However, other costs, such as pollution, are not directly assumed by the users. If roads are not charged, infrastructure costs are external to the user as well. However, these costs may be internalized through pricing approaches aimed at making the users aware of them (Small, 1992).

Internal costs are also divided into costs that are perceived by the users and, as a consequence of that, influence users' decisions, and costs that, despite falling on the users, are not perceived by them. This might be for instance the case of the risk of having an accident. Even though users are aware of that risk, it is hard for them to perceive their ultimate consequences. The internal cost perceived by the users is also called the "generalized travel cost of transport," which is usually employed in transport modeling theory to explain users' behavior.

Cost borne by the users but not perceived by them usually lead to take wrong decisions. In the case a driver takes her car thinking that there is not congestion when actually there is a big bottleneck, she is taking a wrong decision not only for the society but also for herself. This is the reason why information is a key element to get to optimal solutions, notwithstanding the fact that people are not always rational. For instance, people do not necessarily use seat belts even though they know that it reduces the risk of getting hurt.

External Costs

External costs, or externalities, are those costs generated by the road system, both infrastructure and vehicles, but not borne by their ultimate producers, either users or companies. Those are for instance pollution, climate change, noise, and external damages of accidents. Externalities are recognized as a market failure that precludes optimal decisions without the right regulation. This is the reason why over the last few years, large efforts have been made to quantify these costs. In spite of all the effort conducted to that end, the values of different studies still diverge substantially (Link et al., 1999).

The cost of accidents includes material damages, administrative and medical costs, productivity losses, and the value of risk. Part of these costs is internalized through, for instance, insurance payments or damages anticipated by the users. However, a large proportion of accident costs is still considered external.

Most congestion costs are also external insofar as the damage produced by new vehicles entering a road stretch near to be congested is barely perceived by them. However, due to its specific characteristics, these costs will be described in greater detail later. Congestion costs as well as other externalities may be internalized in the short run through road pricing approaches.

Social Costs

Social costs are calculated by adding all the road costs borne by the members of the society: users, companies, government, and the rest of the population. The calculation of social costs should avoid double counting since some internal costs for a certain stakeholder may be benefits to other stakeholders. For instance, tolls are internal costs for road users, which actually influence their decisions, but, at the same time, produce revenues for the company or public authority that charges them. Similarly, taxes are not only costs for a company, but also revenues for the government.

Transferences, through prices taxes, etc., should therefore be excluded of the calculation of social costs. Determining the social cost function is of paramount importance to take right transport planning decisions by using for instance the cost-benefit analysis method.

Cost Functions

Production functions are simply equations for predicting the quantity of output as a function of all inputs' quantities such as fuel, employee hours, and vehicles. Cost functions are similar to production functions in that they predict the cost of production as a function of the output, and the prices of all inputs. A cost function represents the minimum expense necessary to produce a certain output (Jara-Díaz, 2007).

Total Cost

The total cost function shows the evolution of costs depending on a certain output. Considering that the output of a road is traffic, usually measured through veh-km, and assuming that all the vehicles that use the road have similar characteristics, the functional form of the total costs (TC) of a road in terms of traffic would resemble the shape displayed in Fig. 1.

As it was mentioned earlier, costs are different depending on the stakeholder that bears them. In the case of a road, it is crucial to focus on at least two types of costs: the social total cost (STC) painted as a continuous line in Fig. 1 and the users' total costs (UTC) painted as a broken line. The former intends to quantify the total costs borne by the whole society for a certain level of traffic, while

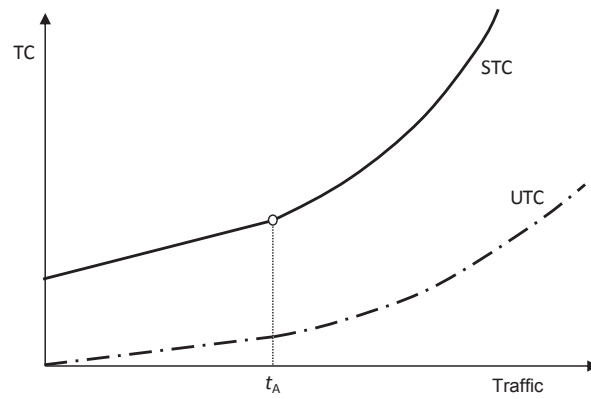


Figure 1 Total cost functions (social and users') for a road. *Source: Own elaboration of the author.*

the latter focuses just on the costs borne by road users. The first one is important for planning and project appraisal purposes. The second one is necessary to understand drivers' decisions since users decide on the basis of the costs borne and perceived by them.

The shape of the curves can be split into two different parts. From traffic zero to traffic t_A , traffic flow is fluid. As a consequence of that, total costs—both social and users'—grow linearly with traffic. The slope of the curve is steeper in the case of STC since externalities not borne by users—such as emissions, noise, and climate change—are included in this curve. The STC comprises also the fixed capital costs of building the infrastructure, which in the case of free roads are not paid by the final users. This explains the difference between STC and UTC when there is no traffic.

From t_A onwards, the road starts experiencing congestion since vehicles begin to disturb each other. According to traffic engineering principles, this effect produces a speed delay in the road, and henceforth greater travel time and a cost increase for the users and the whole society (Gómez-Ibañez et al., 1998). As long as the traffic volume goes up, the average speed in the road goes down. Therefore, the total cost curve gives up being linear to become concave. Again, the growth of the STC curve is more accelerated than the growth of the UTC curve. The reason that explains this phenomenon is that externalities are not borne by users.

Average Versus Marginal Cost and the Impact of Congestion

The total costs function provides a first understanding of the overall cost evolution in a certain road. However, for the purpose of planning, social evaluation and pricing a more detailed analysis through the calculation of the average and marginal cost curves are required.

The average costs curve is the result of dividing the total cost curve by output unit, usually defined in terms of traffic volume. The average cost can be calculated for either the social cost or the users' cost. The social average cost (SAC) shows the cost for all the society—externalities included—per traffic unit. The SAC encompasses all social costs including CAPEX and externalities produced. For low traffic levels this curve reaches very high values since the capital cost is split among few users. As Fig. 2 displays, the SAC gets lower as traffic increases till a certain traffic level t_0 where it reaches its minimum. From then on, the increasing costs due to congestion will offset the capital costs so the SAC curve will start growing. From the social point of view, the optimal traffic for the capacity of the road will be t_0 .

The average cost curve can be also calculated for the cost borne by the users (UAC). This curve is important since users take their decisions according to the average cost they bear as long as they perceive it. The UAC curve can be therefore considered the supply

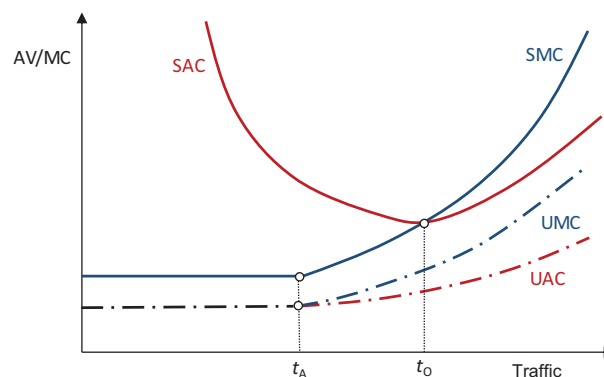


Figure 2 Average and marginal cost functions (social and users') for a road. *Source: Own elaboration of the author.*

function for equilibrium in a road facility with no pricing implemented. This curve comprises only the costs actually borne and perceived by the users such as vehicle operation (depreciation, fuel, lubricants, tire wear, etc.) plus the disutility associated to travel time, reliability, and comfort. This curve can be calculated as the tangent of the angle that joins the coordinate axis with each one of the points of the UTC curve shown in Fig. 1. Before congestion, from traffic zero to t_A , the users' average cost will remain constant. However, once the capacity of the road starts getting constrained, average speeds will go down and travel times will become larger. This is the reason why from t_A onwards the UAC curve will start growing. This basically means that when users see how the speed of the road gets reduced, they experience greater disutility.

The marginal cost curve is also very important for transport economics. The economic theory states that if the supply curve coincides with the social marginal cost curve, the assignment of resources is optimal from the social point of view since the summation of consumer and producer surpluses is optimal. The marginal cost curve reflects the additional cost of producing an additional output unit. From the mathematical point of view, this curve is calculated as the derivative of the total cost function to traffic. The marginal cost curve has the property that it finds the average cost curve at its minimum. At the point where the traffic level reaches an optimal average cost, the marginal and the average cost curves meet each other.

Again, it is possible to calculate the marginal cost curve for both the social cost and the users' cost. The social marginal cost will be the derivative of the STC curve with respect to traffic. This curve is constant from traffic zero to traffic t_A since, if there is no congestion, the additional cost of a new vehicle will be the same as the average of the rest of the vehicles. The users' marginal cost is also constant from traffic zero to t_A . However, it will be located a little below the SMC curve since it does not include the marginal cost of externalities. From t_A onwards, the marginal cost will start increasing in a rapid way. Again, the growth of the SMC is more accentuated than that of the social marginal cost.

It is worth to note that from t_A onwards the UMC curve grows more rapidly than the UAC curve. This means that, when congestion problems appear, each user perceives a cost that is lower than the marginal cost inflicted to the rests of the users of the road (Small, 1992). This is the reason why congestion is usually considered an externality, which could be internalized through road pricing strategies.

Supply Versus Demand Equilibrium

The cost curves are also crucial to understand the equilibrium between supply and demand in a certain road. Fig. 3 shows the behavior of three demand levels for the same infrastructure: low demand (D_L), medium demand (D_M), and high demand (D_H). If no price is set, the natural equilibrium will happen when the demand curve and the UAC curve meet since this curve reflects the cost actually borne by the users.

Fig. 3 shows that the traffic flows of equilibrium for the three demand levels are not optimal according to welfare maximization criteria. The optimal welfare that optimizes the summation of consumer and producer surpluses would happen when users actually perceive the social marginal cost they produce. To reach a first best optimal outcome is necessary to introduce different prices depending on the level of demand (p_L , p_M , and p_H). Those prices will imply reducing the natural traffic flows of equilibrium t_L^n , t_M^n , and t_H^n to the optimal flows of equilibrium t_L^o , t_M^o , and t_H^o . It is noticeable that, once the road begins to get congested, optimal prices get higher so optimal traffic flows diverge to a greater extent of the natural ones. This happens because, once the capacity of the road starts getting constrained, the additional cost produced to all the users by a new vehicle entering the road is higher than the average cost perceived by that driver.

Another interesting aspect is to determine to what extent optimal prices are able to cover the average costs produced which are not internalized by the users. These costs mostly include infrastructure costs (construction, maintenance, etc.) and externalities. For each traffic level, these costs are the difference between the SAC and the UAC curves.

Fig. 3 shows how, for low demand levels, the price p_L is not able to cover the costs not internalized by the users (SAC-UAC). As a consequence of that, this price will not be able to finance the infrastructure costs. The medium demand level, which for the case of

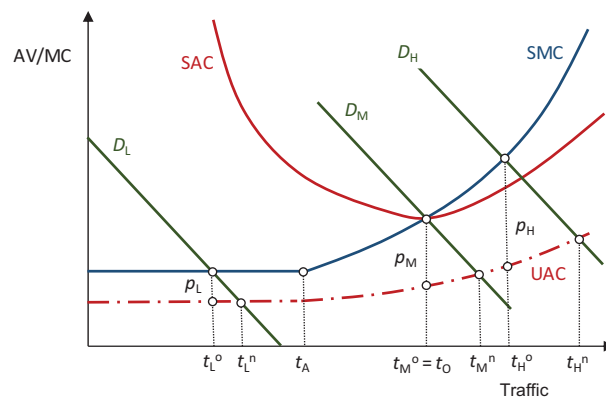


Figure 3 First best optimal prices for different demand levels. Source: Own elaboration of the author.

Fig. 3 is depicted on purpose coinciding with the minimum of the SAC curve, shows that the optimal price p_M fully covers the cost not internalized by the users. In this case, the price will be able to fully finance the infrastructure costs. It is worth to mention that this production level also coincides with the minimum average cost. Demand level D_M is therefore the optimal one for the physical characteristics of the road. At this point it is appropriate to mention that the optimal level from the economic point of view still requires a certain degree of congestion and externalities. Finally, for the high demand level D_H , the optimal price p_H exceeds the costs not borne by the users. This fact proves that, for such a demand level, the infrastructure gets small so, even though in the short term setting a price p_H to the users is a good solution, in the medium term an expansion of the capacity of the road will likely be the best solution. This is a simple explanation of what is called the “self-financing theorem” (Mohring and Harwitz, 1962).

Economies of Scale and Scope

EOS indicates what happens to average costs when scaling up production. A road can experience either (rising) EOSs, constant EOSs, or diseconomies of scale depending on how average costs change as output, such as vehicle miles traveled, increases. EOSs indicate average cost savings per unit as output goes up. According to Fig. 2, the production of the road has rising EOSs from traffic 0 to traffic t_0 . This means that traffic increases in that range will always imply lower SACs. However, from t_0 onwards, because of congestion, the road will have diseconomies of scale.

Economies of scope or cost complementarities are economic factors that make the simultaneous manufacturing of different products more cost-effective than manufacturing them on their own (Jara-Diaz, 2007). In the cost functions described earlier, a single traffic vehicle is considered. However, a road is usually utilized by different types of vehicles, such as cars, buses, and trucks, that produce a different cost per output. In most cases, the average cost of a single road that is used by both cars and trucks simultaneously, especially if there is no congestion, is lower than having two parallel roads, one for cars and one for trucks. This is a clear example of the application of the concept of economies of scope to the road sector.

Cost Calculation, Allocation, and Optimization

Cost Calculation

There are usually three methods for estimating cost functions: accounting, engineering, and econometric. The accounting method is based on taking advantage of the accounting procedures used by organizations to keep track of expenses according to detail categories. The engineering one is based on taking advantage of the knowledge of technology, operations, prices, and quantities of inputs. Engineering models can reach a high level of detail so they can be used to examine the performance of complex systems. The econometric approach is based on calibrating a model on the basis of the information available aimed at determining the impact of certain variables in cost production functions. The functional form usually implemented to calibrate these models is the translog function. This functional form provides a second-order numerical approximation to almost any underlying cost function at a given point.

Cost Allocation

As it was previously mentioned, a road network has cost complementarities for different types of vehicles such as cars and trucks. For some reasons, such as setting a fair price to each vehicle category, it is necessary to allocate road costs to different types of vehicles. This may be easy for costs that are directly attributable to the vehicle such as the wear and tear of the pavement, fuel consumption, of pollution. However, cost allocation may be more complicated for fixed costs such as CAPEX, or accidents where both a car and a truck are involved.

There are two types of allocation procedures for breaking down global costs to vehicle types: deterministic and probabilistic. The deterministic method consists of assigning costs according to certain characteristics of the vehicles on the basis of engineering criteria. For instance, pavement wear and tear costs are usually associated to the axle laden weight of the vehicle. The probabilistic or statistical method, which is usually combined with deterministic methods, is based on setting up functional equations estimated by econometric approaches.

Life-Cycle Cost

The planning process for road investment is aimed at optimizing costs and benefits over the life cycle of a certain asset. The cost component of the road is of paramount importance for the optimization process. Life-cycle cost analysis enables to determine the most cost-effective option among certain alternatives including design, construction, financing, maintenance, operation, and, if necessary, demolition of the work. All the costs are usually discounted to a present-day value.

Life-cycle cost analysis is founded on the fact that different types of cost are not independent from each other. For instance, the capital cost of the infrastructure will influence the maintenance cost. Similarly, the maintenance cost will have an impact on the operation costs of the vehicles. The goal of the life-cycle cost analysis is to find out the optimal combination of costs for the society over the life of the project.

References

- Button, K., Vehoef, E., 1998. Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility. Edward Elgar, Aldershot.
- De Pama, A., Lindsey, R., Quinet, E., Vickermann, R., 2011. A Handbook of Transport Economics. Edward Elgar, Northampton, MA.
- Gomez-Ibañez, J., Tye, B., Winston, C., 1998. Essays in Transportation Economics and Policy: A Handbook in Honour of John Meyer. Brookings Institution Press, Washington, DC.
- Jara-Díaz, S., 2007. Transport Economic Theory. Elsevier, Amsterdam.
- Link, H., Doodgson, J.S., Maibach, M., Herry, M., 1999. The Cost of Road Infrastructure and Congestion in Europe. Physica-Verlag, Heidelberg.
- Mohring, H., Harwitz, M., 1962. Highway Benefits: An Analytical Framework. Northwestern University Press, Evanston, IL.
- Paterson, W., 1987. Road deterioration and maintenance effects: model for planning and management. The Highway Design and Maintenance Standards Series. John Hopkins University Press, Baltimore, MD.
- Small, K.A., 1992. Urban Transport Economics. Harwood Academic Publishers, Chur.
- Small, K.A., Winston, C., Evans, C.A., 1989. Road Work: A New Highway Pricing and Investment Policy. The Brookings Institution, Washington, DC.

Future of Urban Freight

Russell G. Thompson, The University of Melbourne, Melbourne, VIC, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	20
Challenges	20
Population Growth	20
eCommerce	21
Safety	21
Health and Well Being	21
Consolidation	21
Adapting to the Digital Economy	21
Achieving the Sustainable Development Goals	21
City Logistics	22
Off-Hour Deliveries	22
Physical Internet	23
Parcel Lockers	23
Reservation Systems	23
On-Line Auctions	23
Shared Freight Systems	23
Technology Opportunities	24
Loading Bay Management Systems	25
3D Printing	25
Conclusions	25
References	25
Further Reading	25

Introduction

Cities and urban areas have historically been places of trade and commerce that generate a significant amount of freight. In addition to growing levels of manufacturing in urban areas, rapid urbanization in the future will create increased demand for freight relating to the distribution of food and consumer goods as well as transport of construction materials for new residences and infrastructure.

This article discusses current and future challenges confronting urban freight systems and useful approaches for addressing them such as City Logistics and the Physical Internet. Suggestions for how emerging technologies could be implemented in the future to improve the efficiency and sustainability of urban freight systems are also presented.

Challenges

Since goods are generally stored, processed, and consumed at different locations in urban areas, there is a need for goods to move to increase their monetized value for producers, manufacturers, and consumers. Consequently, freight is a derived demand, that is, it does not exist for its own sake, the primary demand is for the production, manufacturing, and consumption of goods. Freight therefore, can be considered as the economy in motion.

Population Growth

Urbanization is a global trend that is predicted to continue for some time. Over half of the world's population currently lives in cities and the urban population of the world is expected to increase by more than two-thirds by 2050 (UN, 2018). Population growth in urban areas will present substantial challenges as future urban freight systems will have to address rising levels of road congestion, that is, threatening efficiency, reliability, and sustainability.

Many cities are currently experiencing pressure on existing infrastructure. Urban freight systems will have to adapt to the limited capacity of future road and rail networks that will be similar to current levels. Associated with population growth in urban areas will be the increase in the size of metropolitan regions. As cities grow spatially, addition distances are required for transporting goods.

Logistics sprawl, the phenomena where logistics facilities, such as warehouses and distribution centers in inner or central areas of cities are relocated to the fringe of urban areas, can create increased transport when ports and high proportion of the population are

located in central areas (Aljohani and Thompson, 2016). In the future, land-use planning will need to address this by protecting and preserving logistics facilities in inner urban areas to reduce the amount of intra urban freight transport.

eCommerce

Increased level of service expectations due to eCommerce (especially B2C), just-in-time manufacturing and rapid response retail logistics is increasing the number of freight movements using vans and small trucks. There is a growing trend for receivers to be able to nominate narrow time windows that makes efficient distribution routes difficult.

Lower levels utilization levels of the load capacity of vehicles (both weight and volume) is a common trend due to higher levels of service demanded by shippers and receivers. More frequent, smaller consignments are leading to an increased number of freight vehicles on urban roads adding to emissions and congestion. Creating more resilient transport networks that can improve reliability from disruption due to incidents, extreme weather, and construction projects will be necessary in the future.

To improve urban freight systems in the future, initiatives will need to increase productivity levels and efficiency for carriers as well as provide higher levels of safety and enhance amenity for residents.

Safety

Reducing road trauma from truck-related crashes is a major challenge. In many cities, there is a recent trend toward more cycling and walking, especially in central city areas. This increases conflicts between vulnerable road users and freight vehicles.

Health and Well Being

Truck emissions are a substantial contributor to air pollution in urban areas. A major challenge is how to decarbonize urban freight to reduce emissions and improve health for residents (ITF, 2018). Although urban areas can provide high levels of liveability, many residents can be exposed to high levels of emissions and noise from freight vehicles that can be detrimental to health (WHO, 2016). Urbanization is increasing the exposure to noise and air pollution.

Traffic noise can elevate cardiovascular disease, cognitive impairment in children, sleep disturbance, tinnitus, and annoyance. Environmental noise from road traffic has been shown to increase stress levels, heart rate, blood pressure and ischaemic heart disease. In addition, high residential traffic exposure as well as road traffic noise has been associated with hypertension. Noise that disrupts sleep has been recognized as having a major health impact.

Diesel engines produce four main pollutants, carbon monoxide-CO, hydrocarbons-HC, particulate matter-PM and nitrogen oxides-NO_x. Ambient air pollution is a major factor in stroke, ischaemic heart disease, lung cancer chronic obstructive pulmonary disease and acute respiratory infection. Diesel emissions from trucks have been found to be a major contributor in urban areas.

Current methods often measure the effects of air pollution in terms of particulate matter (PM), and increases in both mortality and morbidity have been detected at existing ambient PM concentrations. Significant health impact of pollution can therefore be expected in urban centers throughout the world, as exposure to PM is ubiquitous.

PM is emitted from diesel engines. PM_{2.5} (particulate matter less than 10 µm) is believed to be a greater health threat than PM₁₀ (particulate matter less than 10 µm) as the smaller particles are more likely to be deposited deep into the lung. Eliminating diesel and adopting electric vehicles is a trend that is promising but due to the higher financial costs the uptake is slow in many cities.

Consolidation

Achieving high levels of consolidation is a growing challenge in urban freight. Just-in-time manufacturing, rapid replenishment retailing, and eCommerce driven by omni-channel logistics are all contributing to small loads being transported by trucks and vans in urban areas. Currently most freight trips only carrying a small percentage of their carrying capacity. There is also a large proportion of unladen trips traveling back to or from depots without any load. Repositioning of empty shipping containers can also account for a substantial amount of truck traffic in some cities.

Adapting to the Digital Economy

The fourth industry revolution that involves transitioning to a digital society will require a multi-disciplinary approach that will bring together a range of disciplines including engineering, computer science, and business. How supply chains and logistics adapt to the digital economy will have a profound influence on urban freight systems. A key issue for urban freight systems will be how information-based technologies can be designed and implemented to improve efficiency, productivity, and sustainability while not threatening values, such as security, privacy, and safety.

Achieving the Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide useful directions relating to how urban freight systems can support more sustainable cities as well encouraging more responsible consumption and production.

City Logistics

To address the growth in negative impacts of freight in urban areas the field of City Logistics was developed in the late 1990s. City Logistics provides a framework for developing and implementing initiatives aimed at reducing the total costs associated with freight in urban areas, including externalities, such as emissions, crashes, congestion, and noise (Taniguchi et al., 2001).

City Logistics considers urban freight as a complex system. It is important to consider that the primary purpose of an urban freight system relates to its role in providing a service for the economy. The urban freight system allows access to markets for exchanging goods. Thus, the primary objective of urban freight systems is to maximize efficiency (benefits relative to costs). However, City Logistics considers the social and environmental as well as the economic costs of urban freight.

City Logistics is based on the systems approach, where problems are identified and new solutions or schemes are designed, evaluated, and implemented. Involvement by key stakeholders: shippers, carriers, receivers, administrators, and residents are central in City Logistics. The roles and objectives of all stakeholders are considered. Partnerships between stakeholders are encouraged. The Freight Quality Partnership (FQP) program developed in the United Kingdom provides a useful structure for developing effective partnerships for solving urban freight problems (Browne et al., 2004). Such programs allow administrators and private industry to interact and work together to solve complex problems. This involves sharing information, exchanging perspectives, and developing trust between stakeholders.

City Logistics aims to strengthen the goals of cities relating to liveability, mobility, and sustainability particularly when these goals are threatened by freight. City Logistics embraces the free-market economy. Competition is considered crucial for encouraging innovation and improving efficiency. Voluntary participation in initiatives is promoted in contrast to regulation. Advances in information systems are promoted. Traffic management systems aimed at improving the network efficiency and productivity of freight vehicles, are an example of this.

A broad evaluation of a wide range of impacts for all key stakeholders is considered in City Logistics. This considers multiple criteria, encapsulating the different goals and objectives of stakeholders. Recent advances in multi-criteria evaluation methods and Multi Actor Multi Criteria Analysis (MAMCA) allow investigation of the trade-offs between stakeholders to be considered (Macharis et al., 2012).

During the last 2 decades, a number of City Logistics Solutions (CLS) have been developed and implemented in many cities throughout the world. Common solutions include, Urban Consolidation Centers, Joint Distribution Systems or Cooperative Freight Systems, and Off-Hour Deliveries. CLS provide useful initiatives for addressing future urban freight issues. It is important that City Logistics principles be adapted to solve local problems. CLS need be designed and implemented to suite local conditions, given the diverse range of social, political and technological systems that exist between cities.

Urban Consolidation Centers (UCCs) are logistics facilities that are situated in close proximity to a city center, a town or a specific site such as a shopping center. Instead of carriers delivering directly to receivers, goods are dropped off at the UCC, where the goods dropped off by carriers are sorted and consolidated to make deliveries to the final destinations, often using environmentally friendly vehicles.

UCCs address how to increase consolidation levels in freight vehicles that is a key to achieving sustainable urban goods transport. Increasing consolidation for last kilometer reduces unnecessary vehicle movements, and thus congestion and pollution. UCCs have a number of advantages including environmental and social benefits.

Future UCCs will need to be compatible with wider public policy and regulatory initiatives such as access restrictions and loading bay restrictions. For UCCs to be successful they will need to be developed with more orientation toward receiver requirements. Potential for better inventory control, product availability and customer service, and better use of resources at delivery locations as well as opportunity for carrying out value-added activities will drive their adoption.

Due to the high set up and operating costs, the public sector will need to support such initiatives by providing space and enforcing regulations for vehicles not included in consolidation schemes. Open facilities that encourage an exchange of goods between multiple carriers and different modes will be required.

A courier hub has been established in Sydney that provides cages and lockers to exchange goods and parcels for collection and distribution in the central business district. The State government provides free access to this facility for carriers. The courier hub promotes the transfer of goods between vans, bikes, and trolleys.

The Binnestadservice (inner city service) is a scheme that is growing in the Netherlands and other European countries that combines UCCs with a joint delivery service (Quak et al., 2018). When small retailers join, their suppliers are sent the UCC address for carriers to deliver goods to. This scheme deliberately focuses on small and independent retailers with goods delivered when the retailer wishes. Additional services include storage, home deliveries and reverse logistics (waste). A variety of clean vehicles, including electronic bicycles, natural gas trucks, and electric vehicles are utilized.

Off-Hour Deliveries

In future, there will be a need for road infrastructure capacity be used more imaginatively on a 24-hour basis. Recently, a number of major cities (e.g., New York City, London, and San Paulo) have promoted deliveries made during the off-hours (7:00 p.m. to 6:00 a.m.). Off-hour deliveries (OHDs) can either be staff assisted where receivers are present at the time of delivery or unassisted where unsupervised access to the establishment is provided. OHDs aim to reduce congestion and pollution during daytime hours. They have been shown to reduce congestion on main roads and provide substantial travel time savings for carriers. Although there are

potential noise problems for residents from trucks at night, these can be generally overcome by using quieter vehicle technologies as well as improved driver training.

City administrators can encourage more efficient city logistics by providing incentives for vehicles with high load factors, for example, preferential access to loading zones or reservation systems at loading docks as well as providing facilities that create opportunities for transferring goods between modes, such as trucks, vans, and bikes as well as short-term storage.

Physical Internet

The Physical Internet (PI) is an emerging paradigm for planning and managing freight and logistics networks (Montreuil, 2010). It aims to transform the way physical objects are moved, stored, realized, supplied, and used in the pursuit of global logistics efficiency and sustainability. Key elements of the PI are, open and shared networks, standardized and modular load carriers, track and trace protocols, and certificates. The PI requires new business models for sharing assets. Supply network coordination, synchromodality, and information technology are employed for improving the safety and sustainability of supply chains.

PI aims to achieve the right modes and load factors for the right loads that requires compatible load units and coordinated transfers between modes. This involves integrating vehicles, loads and transshipment facilities.

Hyperconnected City Logistics (HCL) combines concepts of the Physical Internet and City Logistics. Within cities, terminals are required for transferring loads between vehicles as well as temporary storage allow consolidated loads on vehicles. To achieve high levels of service, efficiency and productivity, a range of vehicles are used to transport goods between terminals as well as servicing local areas. Bikes, trolleys, and vans can be utilized to collect and distribute goods in areas near terminals, while rail and larger trucks can be used to transfer goods between terminals.

Parcel Lockers

Parcel lockers provide cheap, flexible, and convenient facilities for exchanging small goods as well as reducing the economic and environmental costs associated with eCommerce. Parcel lockers are likely to become more popular in the future, since they have much lower rates of delivery failure for Business-to-Consumer (B2C) and allow consolidated deliveries to locker banks for carriers as well as having lower costs of distribution for carriers in metropolitan areas. Open systems that allow multiple carriers to access locker banks will become more prevalent as they will further reduce costs for carriers.

Shared parcel lockers will also provide an opportunity for more efficient and sustainable Business-to-Business (B2B) courier deliveries in high-density areas such as Central Business Districts (CBDs). Various modes (including walking and bikes) can be integrated to reduce the financial and environmental costs of delivering and collecting parcels. Parcel lockers provide an opportunity to transfer goods between modes as well as carriers. Trucks and vans can be used to carry parcels to locker banks and these can then be picked up by couriers for delivery to the final customers within a precinct or area of the CBD. Walking and cycling will become more popular in central city areas as they are often more productive in performing the last kilometer of deliveries.

Reservation Systems

Loading dock booking or reservation systems will become more common for managing loading docks at activity hubs, such as shopping centers and residential towers. Digital and telecommunication technologies can be used to exchange information between stakeholders including shippers, carriers, receivers, facility managers, and road managers. With booking systems, supply chain communities become more visible and less time is spent manually coordinating deliveries.

Developments in the Internet of Things (IoT) can facilitate access control to dock areas and spaces, thus improving security. Booking systems save delivery firms and their customers' time and money as queuing times can be eliminated. Dramatic reductions in truck idle time at the street level can be achieved.

Booking systems lead to more efficient operation of loading docks and avoid congestion in city streets. Receiving operations for all shipments to and from a distribution location can be optimized by taking into account the type of vehicle, type and quantity of load, individual dock availability, and site operations policy.

On-Line Auctions

On-line auctions provide a means of matching supply and demand in a dynamic way to increase the utilization of vehicles and operating costs. Such systems also allow shippers to outsource to have their transport and avoid having to have their own fleets.

Shared Freight Systems

Freight systems in many metropolitan areas are typically characterized by suppliers operating their own vehicle fleets, distributing only their goods to their customers. There is an opportunity to combine freight networks to reduce the number of vehicles required as well as the distance traveled by freight vehicles. This can result in substantial savings in operating costs for carriers as well as reduced emissions and noise from freight vehicles.

There is a considerable degree of inefficiency relating to the transport of general freight within large metropolitan areas. It is common for not-for-hire and reward carriers to transport small and moderate loads across urban areas often in small vans and medium-sized trucks, with the vehicles generally not carrying freight on the return trip.

In future, high-capacity freight vehicles frequently operating between freight intensive or industrial areas with hubs created to transfer and tranship loads between vehicles to transport goods from shippers to receivers could be implemented.

A substantial amount of general freight transport occurs within metropolitan areas for production, processing, wholesaling, and retailing. General freight is typically stored in boxes and stacked on pallets when being transported. In the future, general freight from a number of shippers could be combined in large freight vehicles to increase vehicle utilization.

Shared freight networks have the potential to reduce the amount of freight vehicles traveling in metropolitan areas. Consolidated loads can often be delivered by fewer vehicles with reduced distances traveled. This can help counter the additional transshipment costs. Voluntary co-operation within specific sectors of the private sector seems to offer good potential for being a successful City Logistics scheme.

The shared urban freight systems of the future will need to be designed considering the network effect that relates to the increased benefits that are created when more users utilize a service. This has been successfully applied to urban public transport systems where hubs are used to transfer passengers between feeder services with high capacity and high frequency services. The same concept can be adapted to the movement of urban freight. The network effect is realized when more shippers and carriers utilize the shared system creating substantial operating cost savings due to the high degree of consolidated loads in vehicles transferring goods between hubs.

A higher degree of coordination between modes will be possible in the future when information can be shared easier. Synchromodality involves a flexible and sustainable allocation of cargo to various modes and routes in a network so that the shipper or forwarder is offered a real-time integrated transport solution (Tavasszy et al., 2017). This involves systemic thinking, focusing on available capacity (regardless of mode of transport). Blockchain related technologies will facilitate shared and more coordinated systems in the future.

Technology Opportunities

The digital economy will be largely influenced by advances in data, networks, and automation (DNA). Big data can be captured from the proliferation of low cost and connected sensors. Track and trace systems will allow unprecedented transparency and monitoring of freight. Developments in deep learning and cognitive computing will allow a richer understanding of demand patterns, levels of service, and impacts to be achieved.

The proliferation of sensors will create many opportunities for improved forecasting and automated systems. Artificial Intelligence (AI) based methods including machine learning will allow urban freight systems to be more efficient by anticipating demand and managing capacity in real time.

Developments in automation will allow more intelligent and autonomous systems to be developed. This will lead to lower freight and logistics costs since labor rates are high in many urban areas. Tasks such as sorting, packing, and even driving will be automated. This will lead to improvements in safety and reliability.

Although robots and drones offer many benefits for urban freight systems, their successful implementation will depend on how issues associated with security, privacy, and safety can be overcome to gain community acceptance.

Freight vehicles operating in urban areas will become more automated, connected, electric, and shared (ACES) in the future. Cleaner, low noise and more energy-efficient vehicles will improve sustainability and liveability.

The rollout of integrated information technologies will provide many opportunities for enhancing the performance of urban freight systems. Next generation traffic management systems will allow many freight-oriented services to be developed.

Advanced sensing and communication technologies such as Dedicated Short Range Communication (DSRC) allow freight vehicles to communicate with other vehicles as well as infrastructure. The ability of freight vehicles to be incorporated in future traffic management systems will enhance sustainability. Connectivity between trucks and signal systems will lead to significant improvements in efficiency, as most delays for trucks in cities occur at signalized intersections.

It is common for trucks to be delayed due to slow acceleration from stops. Signal systems currently are unable to be responsive to freight vehicles and their loads in real time since they do not communicate or respond to individual freight vehicles, leading to frequent braking, acceleration, and stopping. Future traffic signal systems will be able to adjust offsets, extend green times, and/or provide advisory speeds for drivers to ensure that stopping is minimized. Significant reductions in travel times of trucks along a corridor will be able to be realized. Savings in fuel consumption and vehicle wear and tear for carriers will be able to be achieved. Residents will also be exposed to lower noise and emission levels. Detection of trucks and vans as well as their loads could also allow priority for freight vehicles based on their utilization.

Advanced traffic management systems will be able to provide information regarding road closures and network disruptions from roadworks, signal faults, and construction zones. There will also be opportunities for improving safety of pedestrians and cyclists around trucks. It will be possible for alerts to be issued to drivers concerning vulnerable road users in vicinity of trucks.

Traffic sensor networks will permit more accurate short-term predictions of travel times between origins and destinations on urban traffic networks to be produced. Thus, predicted arrival times will be more accurate by being updated to reflect actual and predicted traffic network conditions. Updated predictions of the expected time of arrival (ETA) be available during the trip based on the estimated demand on the traffic network.

Loading Bay Management Systems

A substantial amount of time for couriers in central city areas involves conducting walking routes between vans and receivers. Future systems will be optimizing courier routes and be able to reserve loading bays to ensure that traffic congestion and delays are minimized.

3D Printing

3D printing allows specialized components and products to be made on-site and on-demand. This largely eliminates distribution and storage of goods and minimizes packaging.

Conclusions

Rapid urbanization is creating many challenges for improving the efficiency and sustainability of urban freight systems. Approaches such as City Logistics and the Physical Internet provide frameworks for ensuring that future solutions will be more integrated and address social and environmental issues. Emerging technologies have the potential to improve the performance of urban freight systems but to gain acceptance by communities they will need to ensure that values such as privacy, security, and safety are not compromised.

References

- Aljohani, K., Thompson, R.G., 2016. Impacts of logistics sprawl on the urban environment and logistics: taxonomy and review of literature. *J. Transport Geogr.* 57, 255–263.
- Browne, M., Nemoto, T., Visser, J., Whiteing, T., 2004. Urban freight movements and public-private partnerships. In: Taniguchi, E., Thompson, R.G. (Eds.), *Logistics Systems for Sustainable Cities*. Elsevier, Oxford, pp. 17–35.
- ITF, 2018. *Towards Road Freight Decarbonisation—Trends, Measures, and Policies*, International Transport Forum, OECD Publishing, Paris.
- Macharis, C., Turckin, L., Lebeau, K., 2012. Multi actor multi criteria analysis (MAMCA) as a tool to support sustainable decisions: State of use. *Decis. Support Syst.* 54, 610–620.
- Montreuil, B., 2010. *Physical internet manifesto*, [online] Available from: www.physicalinternetinitiative.org
- Quak, H., Kok, R., den Boer, E., 2018. The future of city logistics—trends and developments. Leading toward a smart and zero-emission system. In: Taniguchi, E., Thompson, R.G. (Eds.), *City Logistics 1: New Opportunities and Challenges*. Wiley, New Jersey, pp. 125–146.
- Taniguchi, E., Thompson, R.G., Yamada, T., Van Duin, R., 2001. *City Logistics—Network Modelling and Intelligent Transport Systems*. Elsevier, Oxford.
- Tavasszy, L., Behdani, B., Konings, R., 2017. Intermodality and synchronicity, Chapter 15. In: *Ports and Networks: Strategies, Operations and Perspectives*, Taylor and Francis, Abingdon, pp. 251–266.
- UN, 2018. *World Urbanization Prospects: The 2018 Revision—Key Facts*, Economic and Social Affairs, United Nations.
- WHO, 2016. *Preventing Disease Through Healthy Environments, A global assessment of the burden of disease from environmental risks*, World Health Organization, Geneva.

Further Reading

- Crainic, T.G., Montreuil, B., 2016. Physical internet enabled hyperconnected city logistics. *Transp. Res. Procedia* 12, 383–398.
- DHL, 2018. *Logistics Trend Radar Updated Report*, DHL Customer Solutions & Innovation, Deutsche Post DHL Group, Troisdorf.
- OECD, 2003. *Delivering the Goods: 21st Century Challenges to Urban Goods Transport*, Road Transport Research Programme (RTR), Directorate for Science, Technology, and Industry, Organisation for Economic Development (OECD), Paris.
- PIARC, 2012. *Public Sector Governance of Urban Freight Transport*, PIARC Technical Committee B.4, Freight Transport and Inter-Modality, World Road Association.
- Taniguchi, E., Thompson, R.G. (Eds.), 2015. *City Logistics: Mapping the Future*. CRC Press/Taylor & Francis, Boca Raton.
- Thompson, R.G., 2015. Vehicle related innovations for improving the environmental performance of urban freight systems. In: Fahimnia, B., Bell, M., Hensher, D.A., Sarkis, J. (Eds.), *Green Logistics and Transportation: A Sustainable Supply Chain Perspective*. Springer, Cham, pp. 119–129.

Operation Costs for Public Transport

Marco Batarce, Faculty of Economics and Business, Universidad San Sebastián, Santiago, Chile

© 2021 Elsevier Ltd. All rights reserved.

Introduction	26
Inputs for Public Transport Provision	26
The Output of Public Transportation	27
Economies of Scale, Density, and Scope	28
Efficiency, Contracts, and Regulation	29
References	29

Introduction

The operation cost corresponds to the production cost of transportation. In the specific case of public transportation, the operation cost is the cost incurred by the firm or public agency that provides the service and manages a fleet of either buses, trains, or a mix of both. Like any economic cost, in its definition is implicit the idea of minimization of the expenditures needed to produce a given level of transport output and input prices. This minimization of expenditures implies an optimal allocation of all the input factors used for transport production.

The economic analysis of operation costs is relevant for policy design and procurement of public transport services by governments. Usually, the design of regulatory policies needs to know the industrial structure of the public transport industry. For that purpose, the studies are led by economic theory and conducted by using statistical data and econometric methods. The usual approach is the estimation of a cost function. In this case, the objective is to identify critical parameters describing the industrial structure such as degree of scale economies, density economies or size economies, average and marginal cost, and marginal rates of substitution of inputs. This information can be used to define the type of regulation, the fare level, or the need and level of subsidies, for example.

Practitioners, however, use accounting cost studies to deal with the procurement of public transport service and contract design. These studies are not based on the economic theory and use the account cost of the firms or public agencies to establish a relationship between cost and some measures of supply-oriented output (see discussion on types of outputs next). For instance, the regulator can assume that the total cost is a linear combination of vehicle-kilometers, vehicle-hours, peak vehicles in services, and route lengths. The coefficients of this linear model are used to set the payment to the firm (or transfer to the public agency) for the provision of the service in a contract.

The formal economic analysis of the operation cost of public transport addresses three main issues (Gagnepain et al., 2011). The first issue is a description of the technology behind the industry in economic terms by measuring economies of scale, density, and scope. The second issue is a definition for the output, either supply indicators (e.g., vehicle-kilometers or seat-kilometers), demand related output measures (e.g., passenger-kilometers or the number of passengers) or multidimensional output. Finally, the third issue is the evaluation of firms' performance through a measure of the technical efficiency level.

Inputs for Public Transport Provision

The main input factors used to produce public transport are capital, labor, and energy. The capital consists of buses, trains, and bus garages. The infrastructure needed for trains is not part of the capital as it is shared with other services, like freight transport, similar to the streets for the buses. However, some authors consider the network as an input for transport production as the output depends on its structure.

The labor comprises of the human resources needed for the operation and maintenance of vehicles, stations, and depots, and the management of the firm or public agency in charge of the public transport provision. The energy consists of combustible or electricity needed for the operation of the vehicles. The importance of input factors in the total cost depends on the type of public transport and the location of the network it operates. In the case of buses, labor share is between 20% and 60%, depending on the country (White, 2018). For instance, the share of labor can reach 60% in London, around 40% in Santiago, and around 20% in India. Several factors determine the level of labor cost. For instance, the average income per capita in the network location determines the level of wage. The number of persons and the level of qualification required by the staff (especially drivers) depend on the type of vehicle (buses, trams, or trains) because the person's skills are different. The level of regulation of the industry also affects the labor cost because the unregulated networks tend to operate informally (for instance, in developing countries) with low administrative staff and unqualified drivers, and regulate networks tend to operate under strong labor rules and unions (e.g., Europe).

Capital cost comprises of vehicles and infrastructure. In the case of vehicles, the cost varies widely because the network can be operated with very different types. For instance, small, low-technology buses imply low capital cost, but high-technology trains imply high costs. Also, the technology and vehicle determine the capital cost related to infrastructure as depots and garages. The capital cost also depends on the interest rates for financing the investment, which in turn depends on the level of risk. For instance, in

developing countries with unregulated transit systems, the operating firms are small companies composed of many owners of one or two buses organized into unions. These firms get high-interest rates because of the low bargaining power and the high risk due to intense competition and weak regulatory institutions. In regulated transit systems, usually, there are one or few operating firms, which access to the low-interest rates because of the government gives subsidies or guarantees the debt. The capital cost also depends on the country's capacity to produce the vehicles and the competitiveness of the vehicle industry. This competitiveness is particularly relevant for the trains because a few countries are manufacturing it.

The cost of energy is usually related to the prices of electricity and diesel. The price of both inputs depends on the country where the public transport service operates. Some factors affecting these prices are the fuel tax policy, the level of domestic oil extraction, and the sources of electricity generation (e.g., coal, natural gas, hydroelectric, nuclear).

The operation cost also depends on the network structure (Jara-Diaz, 2007). One obvious way that the network affects the operation cost is its composition in terms of the type of vehicle and technology: buses, trains, or tramways, for instance. Another less obvious way is the network's topological structure. This structure depends on the connections of the lines of either bus, tramway, or train (subway or light rail). For instance, a bus network has a structure of direct lines if the bus routes connect most of the origins and destinations without transfers. This type of network is common in cities with deregulated public transport, where many firms deliver services independently. In turn, a network has a feeder-trunk structure if there exist several high-capacity services, such as subway or BRT, which operate in main roads and connect zones with lower demand (or lower residential density) by buses or other low-capacity vehicles. This structure is typical in cities with one or a few bus companies, where there exists an integrated fare with a unified payment system like electronic cards.

If the operating firm keeps the network structure fixed when optimize its operation allocating inputs for a given level of output, we need to recognize that the operation cost is conditional on this network structure. However, if the firm chooses the input allocation and the network structure, the cost is the global operation cost and is less or equal to the conditional operation cost.

Another essential factor for producing transport is the users' time. Indeed, the participation of users is necessary to produce final transport output, such as trips. The inclusion of users' time as input is relevant for public transport policy design. In this case, the (benevolent) transport planner takes into account the monetary cost of public transport provision and the users' cost (time). Several authors have shown that the inclusion of users' time as input led to substantial economies of density in public transport provision (Mohring, 1972) and, thus, marginal cost pricing will require subsidies. Besides, if the motivation is the study of the industrial structure to design regulatory policies, the users' time is less relevant in the operation costs as the focus is on the firms' behavior. They do not take into account users' time directly when making decisions on the operation, pricing, or merger, but through the demand, if it is elastic to either waiting, accessing or in-vehicle time.

The Output of Public Transportation

The formal economic analysis of transportation costs must consider multiproduct firms. This feature implies the output is a flow vector, with components identified by origin, destination, period, and commodity type (Jara-Diaz, 1982; Winston, 1985). Such an output disaggregation is infeasible for public transport because the number of origin-destination pairs is too large. For instance, if a bus service passes through ten different zones in a city, the vector of output should have 45 components, if there are three periods: morning peak, off-peak, and afternoon peak, the output vector should be of 135 components. Each component is the number of carried passengers from an origin to a destination. Then, the study of the operation cost needs information on the demand for every component. Moreover, the cost in one origin-destination pair depends on the output in other pairs. The complexity implied by the disaggregate output leads to study operation costs using aggregate measures of output such as total passenger, total driven kilometers, passenger-kilometer, seat-kilometer, or vehicle-kilometer.

Among the aggregate measures of output, we can distinguish two types: supply-oriented output, and demand-oriented output. The supply-oriented outputs are those measures related to the transport capacity delivered by the operators without involving the number of carried passengers. For instance, the seat-kilometers supplied in a month by a subway company, or the bus-kilometers supplied by a bus line. Other standard measures of supply-oriented output are total driven kilometers, total seats, seat-hours, vehicle-hours, and peak vehicles in service.

These output measures are also called intermediate output because to produce the final output, which is carrying passengers, it is necessary to provide seats, driven kilometers, and vehicles.

The demand-oriented output is a measure of the final output, which means the measure involves the carried passengers in some way. Standard measures of demand-oriented output are total passenger and passenger-kilometers. In regulated networks, some times, the operator's farebox revenue is a measure of the output because this information is available from the periodic reports required by regulators (for instance, Gagnepain and Ivaldi, 2002). When the information of carried passengers is not always available or is not reliable, the analysis of operation cost uses supply-oriented output or indirect measures of passenger demand. For instance, Batarce (2016) uses a transit network assignment model to compute the demand for every bus service and estimate the marginal cost in Santiago's bus transit system.

However, the aggregate measures of output are not enough to describe unambiguously the operation costs, which depend on several variables other than output and inputs. Some of these variables are related to the operation itself and some others related to exogenous conditions. For instance, among the first group of variables, the bus frequency or the headway determine the fleet and the labor (bus drivers), which may be different for the same aggregate output level. The bus capacity is another variable in the first group

that determines the operation cost. An exogenous variable is the distribution of the demand on the service route. If the demand concentrates on a few route segments, the required capacity to meet the demand is higher than the required capacity for a uniformly distributed demand. The demand concentration, in turn, depends on the population density in the city. Another exogenous variable is the distribution of the demand along the day and the concentration on the peak hours. Also, the congestion on peak hours affects the bus operation costs as the lower commercial speed, the larger the necessary fleet to fulfill the frequency and provide the capacity to meet the demand.

In consequence, operation cost estimation with aggregate output measures could result in biased estimates. To overcome this problem, [Spady and Friedlaender \(1978\)](#) proposed a hedonic output approach that defines the output as a function of a vector of aggregates. Components of this vector measure different characteristics and technological factors of output. For instance, some output characteristics that may be included in the hedonic function specification are the number of stops on the route, average passenger load, average trip length, and route length. Similarly, a way to deal with disaggregate output is allowing the cost function to depend on few aggregate measures of output and some descriptors of the operating environment (e.g., route length) (see [Basso et al., 2011](#)).

Economies of Scale, Density, and Scope

As mentioned, the analysis of firms' technology in economic terms is a relevant issue when studying the operation costs of public transportation. This technology is mainly characterized by the degree of economies of scale, scope, and density.

To understand what economies of scale and scope are, consider a vector of disaggregate output from a public transport firm $y = (y_1, \dots, y_n)$. In general, the output vector components are different products. The public transport firm's output vector is composed of passenger flows between different origin-destination pairs served. For instance, if the firm operates a bus line that serves five zones in a city in a one-way cyclic circuit, the vector would have $n = 20$ components because every zone connects all other zones. The output vector might also be composed of passenger flows in different periods or another disaggregation relevant to the firm's cost structure.

The standard definition of increasing returns to scale is that if inputs are all increased by the small factor $\lambda > 1$, then output increases by a factor larger than λ ([Panzar and Willig \(1977\)](#) show that the degree of economies of scale in multi-output production is $A = C(y) / \sum_{i=1}^n y_i C_i(y)$, where $C_i(y)$ is the marginal cost with respect to the output component y_i). For the public transport case, the definition implies that all the passenger flows increase by the same factor. Economies of scale imply decreasing ray average cost, which reduces to the familiar idea of declining average costs in the case of a single output ([Panzar and Willig, 1977](#)). In the case of public transport, decreasing ray average costs mean only that equiproportionate division of the passenger flow vector into two or more firms would increase system costs. If the division of passenger flows is not equiproportional, there is no warranty that the system costs increase. Therefore, the degree of economies of scale measures a specific effect of output increment on the costs, which is the effect of the division of passenger flows between two or more firms keeping the same distribution of flows across the origin-destination pairs.

The economies of scope relate the effect of divide the form's output vector in two or more firms with output vectors are a partition of the original vector. To define economies of scope formally, we consider a partition of the output vector. If $N = \{1, \dots, n\}$, then a partition of N is a set $T = \{T_1, \dots, T_J\}$ such that T_j is a subset of N , for all j , the union of all T_j s forms the set N , the intersection of T_j and T_k is empty, for all $j \neq k$, and $J > 1$. We denote y_{T_j} the n vector whose elements are set equal to those of y for i in T_j and 0 otherwise. Then, there are economies of scope for the partition T if $\sum_{j=1}^J C(y_{T_j}) > C(y)$ ([Panzar and Willig, 1981](#)).

For instance, in the case of the vector of passenger flows $y = (y_1, \dots, y_{20})$, a partition would be $T_1 = \{1, \dots, 10\}$ and $T_2 = \{11, \dots, 20\}$, and the vectors from this partition would be $y_{T_1} = (y_1, \dots, y_{10}, 0, \dots, 0)$ and $y_{T_2} = (0, \dots, 0, y_{11}, \dots, y_{20})$. It is worth noticing that even if the partition is theoretically admissible, in the case of a bus line, it could imply that the passengers may alight in some stops (zones) but not board or vice versa because of the origin-destination structure of the demand.

The concept of economies of scope is useful to analyze the convenience of expanding the spatial scope of the public transport firm. For instance, consider a new bus line composed by the original bus line that connects five zones but now connecting a new zone. To analyze the economies of scope, we need to compute three costs: the new bus line connecting the six zones, whose output vector has 30 components, the original bus lines with an output of 20 components, and the costs of providing public transport between the added zone and every original zone (10 components). The latter costs correspond to the extra components of the output vector needed to form a partition of the output vector from the bus line connecting the six zones. Then, there exist economies of scope if the new line's costs are less than the sum of the original bus line's costs and the costs of providing service for the extra components of output (in the most efficient way). It is worth noticing that, when applying the formal definition of economies of scope to public transport, the way to define output vector partitions is limited by the physical and operational configuration of the public transport network.

The previous definitions of economies of scale and scope need a cost function specified with disaggregate output. As explained, usually, there are not enough data to estimate such a cost function and the solution is the use of aggregate output. This solution imposes some constraints in the analysis of industrial structure because of the impossibility of computing ray average costs and defining partitions for the output vector.

So, when using aggregate output, [Small and Verhoef \(2007\)](#) point out the importance of retaining the distinction between expanding the density and expanding the spatial scale of output. The expansion of density implies keeping constant the spatial scope of the firm (e.g., route network or the number of served points) while the aggregate measure of output (e.g., passengers or passenger-kilometers) increases. If the average cost decreases when density expands, the industry exhibits economies of density. When using a demand-oriented aggregate output, the economies of density are equivalent to the multi-product economies of scale concerning the disaggregate output vector ([Jara-Díaz, 2007](#)).

The expansion of the spatial scale involves new components of the disaggregate output vector because expanding the network implies, for instance, supplying service between new origin-destination pairs. Strictly speaking, this is an expansion of scope (number of products) and must be analyzed as economies of scope ([Panzar and Willig, 1981](#)). However, when using aggregate output, the separation between scope and scale is difficult, if not impossible. Regarding this inseparability, [Small and Verhoef \(2007\)](#) say the industry exhibits economies of size when network expansion, along with aggregate output expansion, reduces the average cost. It is worth noticing that the concept of economies of size, which are usually called economies of scale in transport literature, is ambiguous because it implies new origin-destination pairs served and output expansion.

Moreover, the level of scale economies depends on the selected output regarding the supply-oriented output versus the demand-oriented one. For instance, some studies find scale diseconomies in terms of supply-related output, whereas they noticed economies of scale in terms of demand-related output. When comparing demand-oriented and supply-oriented scale economies, there seem to be higher returns to scale with demand-oriented outputs (trips, journeys, receipts per passenger, or passenger-kilometers) than with supply-oriented outputs (vehicle-kilometers or seats-kilometers) ([Croissant et al., 2013](#)). These results are also consistent with the idea that an additional passenger increases the cost very little if the supply does not increase (for instance, [Batarce and Galilea, 2018](#)). It is worth noticing, however, that the level of economies of scale depends on the selected output, control variables for heterogeneity and functional form of the cost function.

Efficiency, Contracts, and Regulation

Finally, a relevant topic in the analysis of operating costs of public transportation is the measurement of efficiency and the effects of regulation and procurement contracts on it. A standard method to measure efficiency is the stochastic frontier analysis ([Kumbhakar and Lovell, 2003](#)). This method assumes that the firms' inefficiency is a specific, unobservable component of the firm's production technology. This unobservable component is modeled as a random variable independent across firms, and its distribution is used to measure the firms' efficiency level. In transportation literature, most studies estimate the determinants of the efficiency of the firm and the level of subsidy. For instance, some authors estimate stochastic frontier models and study the mean inefficiency as a function of the type of contract, ownership regime, or subsidization mechanisms.

The results show that technical efficiency is not independent of the institutional or regulatory constraints. For instance, private operators outperform public ones in terms of efficiency ([Roy and Yvrande-Billon, 2007](#)), operators under cost-plus contracts exhibit a higher level of technical inefficiency than operators under fixed-price contracts ([Gagnepain and Ivaldi, 2002](#); [Piacenza, 2006](#)), and high-powered scheme of regulation, such as yardstick competition, significantly reduces operating costs ([Dalen and Gómez-Lobo, 2003](#)).

References

- Basso, L., Jara-Díaz, S., Waters, W.G., 2011. Cost functions for transport firms a handbook of transport economics. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *A Handbook of Transport Economics*. Edward Elgar Publishing, Cheltenham, UK, pp. 273–297.
- Batarce, M., 2016. Estimation of urban bus transit marginal cost without cost data. *Transport. Res. Part B: Methodol.* 90, 241–262.
- Batarce, M., Galilea, P., 2018. Cost and fare estimation for the urban bus transit system of Santiago. *Transport Policy* 64, 92–101.
- Croissant, Y., Roy, W., Canton, J., 2013. Reducing urban public transport costs by tendering lots: a panel data estimation. *Appl. Econ.* 45, 3711–3722.
- Jara-Díaz, S., 1982. The estimation of transport cost functions: a methodological review. *Transport Rev.* 2, 257–278.
- Jara-Díaz, S., 2007. *Transport Economic Theory*. Elsevier, UK.
- Kumbhakar, S.C., Lovell, C.K., 2003. *Stochastic Frontier Analysis*. Cambridge University Press.
- Dalen, D.M., Gómez-Lobo, A., 2003. Yardsticks on the road: regulatory contracts and cost efficiency in the Norwegian bus industry. *Transportation* 30, 371–386.
- Gagnepain, P., Ivaldi, M., 2002. Incentive regulatory policies: the case of public transit systems in France. *RAND J. Econ.* 33, 605–629.
- Gagnepain, P., Ivaldi, M., Muller-Vibes, C., 2011. The industrial organization of competition in local bus services. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *A Handbook of Transport Economics*. Edward Elgar Publishing, Cheltenham, UK, pp. 744–762.
- Mohring, H., 1972. Optimization and scale economies in urban bus transportation. *Am. Econ. Rev.* 62, 591–604.
- Panzar, J.C., Willig, R.D., 1977. Economies of scale in multi-output production. *Quart. J. Econ.* 91, 481–493.
- Panzar, J.C., Willig, R.D., 1981. Economies of scope. *Am. Econ. Rev.* 71, 268–272.
- Piacenza, M., 2006. Regulatory contracts and cost efficiency: stochastic frontier evidence from the Italian local public transport. *J. Product. Anal.* 25, 257–277.
- Roy, W., Yvrande-Billon, A., 2007. Ownership, contractual practices and technical efficiency: the case of urban public transport in France. *J. Transport Econ. Policy* 41, 257–282.
- Small, K.A., Verhoef, E.T., 2007. *The Economics of Urban Transportation*. Routledge, Abingdon, UK.
- Spady, R., Friedlaender, A.F., 1978. Hedonic cost functions for the regulated trucking industry. *Bell J. Econ.* 9, 159–179.
- Winston, C., 1985. Conceptual developments in the economics of transportation: An interpretive survey. *J. Econ. Lit.* 23, 57–94.
- White, P., 2018. Bus economics. In: Cowie, J., Ison, S. (Eds.), *The Routledge Handbook of Transport Economics*. Routledge, UK, pp. 31–47.

Natural Monopoly in Transport

André de Palma, Julien Monardo, CREST, ENS Paris-Saclay, University of Paris-Saclay, Paris, France

© 2021 Elsevier Ltd. All rights reserved.

History	30
Formal Definition of the Natural Monopoly	31
Regulating a Natural Monopoly	32
Econometrics of Natural Monopoly	33
Natural Monopolies in Transports: Panorama and Case Study	34
Acknowledgment	35
References	35

History

The concept of natural monopoly appeared with [Smith \(1776\)](#) who, without naming it, explicitly provided the main characteristics of what scholars after him refer to as “natural monopoly.” Its definition has then evolved through time and has attracted the attention of several famous scholars of the 17th–18th centuries, such as Thomas Malthus, Frédéric Bastiat, John Stuart Mill, and Léon Walras.^a

In the earliest explicit use of the concept, natural monopolies referred to as monopolies derived from natural factors of production, which are supplied in fixed quantity, with the idea that the limited supply of such factors constitutes barriers to entry.^b The first definition, however, was given by J.S. Mill: natural monopolies were “those which are created by circumstances, and not by law.” At that time, natural monopolies were therefore those created by nature, due to the presence of production factors supplied in given, and potentially limited, quantity; *natural monopolies* were thus distinguished from *artificial monopolies* created by law, that is, by government measures. For Mill, natural monopolies encompassed many situations, including, for instance, barriers to entry due to capital requirement. Mill was also the first to recognize that natural monopolies could arise due to the production process, that is, due to technological reasons.

Afterwards, natural monopolies were meant to arise due to the presence of economies of scale, that is, when the average total cost is decreasing. This happens, in particular, when there are fixed (potentially sunk) costs and low or zero marginal costs. In this situation, the cost of the incumbent firm is lower than the cost of any other firm that would wish to enter the market, and, in turn, that firm remains alone in the market. Then, price is not equal to the marginal cost, as in the case of perfect competition, since profit maximization requires the monopoly to equalize marginal revenue to marginal cost; and the monopoly produces too little with respect to the social optimum conditions, so that the government may wish to regulate it.

The current formal definition used in the academic literature is due to [Baumol \(1977\)](#) and is closely related to the subadditivity of the cost functions, that is, natural monopolies arise when the production cost associated to any set of outputs is less than the sum of the costs of producing separately all the different products in this set of outputs (see the formal definition later).

Very soon, academic scholars recognized that monopolies were unavoidable in transport networks, such as railways, roads, and highways. For Jules Dupuit, a French engineer, monopolies in transport networks are due to their need to build a large infrastructure before operations could start. This makes the entry of a new firm impossible because only a very limited number of entrepreneurs can have access to a sufficiently huge capital. Moreover, if a new firm entered the market, it would extract profits from the incumbent monopoly, making both of them unprofitable. By contrast, for Walras, monopolies arise because only the government can decide the expropriation of the lands required to build the transport networks. Note also that, in transport networks, the presence of several small businesses is inefficient: as highlighted by [Walras \(1936\)](#), “building a second network of roads in a country where there is already one that is enough for all the communications would be an absurd way of chasing economies.”

However, many monopolies we know remain unchallenged given that strong regulations often protect them. Productions of electricity, of nuclear weapons, and of military defence involve large fixed cost, and have been (and are still, for since several decades) protected by governments. By contrast, several economists belonging to the Austrian School such as Ludwig von Mises or Friedrich von Hayek have advocated that natural monopolies do not really exist (Thomas J. DiLorenzo speaks about “myth of natural monopolies”) but are often the outcome of regulation or of some kind of State protection. The libertinism of the Austrian School is here a bit confusing. What is true, for sure, is that governments often play a role in protecting some natural monopolies. However, other monopolies, even “natural,” could be challenged by firms using improved technologies.^c

^a See [Mosca \(2008\)](#) for an excellent history of the concept of natural monopoly.

^b In 1815, Malthus, in his essay *The Nature of Rent*, made the distinction between “natural” monopoly and “artificial” monopoly. For instance, he mentioned as natural monopoly the case of “certain vineyards in France, which, from the peculiarity of their soil and situation, exclusively yield wine of a certain flavour.”

^c Entry in the taxi market, for example, has been historically difficult in France, especially in Paris, Île-de-France; but Uber managed to break (more or less successfully) this market in December 2011, see https://en.wikipedia.org/wiki/Timeline_of_Uber, and started to capture customers, even when facing low network externalities, because they developed a revolutionary technology and were prepared to face (at least initially) negative profits.

Formal Definition of the Natural Monopoly

A monopoly is a market structure in which a single firm produces a good or service without any close substitutes. Monopolies may have several sources, such as legal barriers (e.g., patents), capital requirements, economies of scales, etc. One particular form of monopoly is the natural monopoly, which arises when a single firm is able to offer that good or service to an entire market at a lower cost than two or more firms could. This means that a natural monopoly can be the outcome of an unrestricted competition.

The current formal definition of the natural monopoly is due to Baumol (1977) and is closely linked to the strict subadditivity of the cost function. A cost function $C(y)$ is *strictly subadditive* if, for any vector, $(y_1, \dots, y_M)$:

$$C\left(\sum_{m=1}^M y_m\right) < \sum_{m=1}^M C(y_m),$$

where the quantities y_m are either quantities of different outputs or different quantities of the same output. A necessary and sufficient condition for a natural monopoly to exist is that the cost function is subadditive, which means that a single firm could produce at a cheaper cost compared to several firms.

The definition is also related to the concepts of economies of scale and economies of scope, which are cost efficiencies formed by quantity and by variety, respectively. *Economies of scale* correspond to a decreasing average total cost, while *economies of scope* arise when it is cheaper to produce several products together than to produce them separately.

For single product cost functions, economies of scale and economies of scope are sufficient but not necessary for subadditivity. This means that a natural monopoly arises when there are economies of scale or economies of scope over the relevant range of output (i.e., the range of output between the first unit of output produced and the output which consumers would demand at a zero price). For multiproduct cost functions, however, these conditions are neither necessary nor sufficient.^d

The subadditivity of the cost function is also related but must be put in perspective with the concept of sustainability of monopoly, which refers to “[a]n industry to which entrants are not ‘naturally’ attracted, and are incapable of survival even in the absence of ‘predatory’ measures by the monopolist” (Baumol, 1977). In particular, Faulhaber (1975) shows that subadditivity of the cost function does not imply sustainability of the monopoly, while Baumol et al. (1977) show the converse.

To illustrate what happens, consider a monopoly that produces and sells a single good or service at single price (i.e., absent of price discrimination). The monopoly produces a quantity y so as to maximize its profits $\pi(y)$ defined by the difference between its total revenues $R(y)$ and its total costs $C(y)$:

$$\pi(y) = R(y) - C(y) = p(y)y - C(y),$$

where $p(y)$ is the (decreasing) inverse demand function, which gives the price at which the quantity y can be sold.

Assuming that price and cost are differentiable and well behaved, profits will be maximum when marginal revenues equal marginal costs, that is, when $R'(y) = C'(y)$.^e Since total revenues are equal to price multiplied by demand, this first-order condition leads to the monopoly pricing formula, also known as the inverse elasticity rule:

$$\frac{p(y) - C'(y)}{C'(y)} = -\frac{1}{\varepsilon},$$

where $\varepsilon = \frac{p'(y)}{p(y)/y}$ represents the elasticity of demand. The right-hand side of the earlier equation is referred to as the Lerner index and measures the market power of the monopoly.^f

As illustrated in the top panel of Fig. 1, where marginal costs are assumed to be constant, profits are maximum at the point of intersection denoted by E , where the monopolist produces a quantity y_m and sells at a price p_m .

At this optimum, the monopoly obtains profits equal to π_m and consumers enjoy a surplus of CS_m . The society incurs a deadweight loss of DL_m : the social surplus, equal to $\pi_m + CS_m$, is lower than its socially efficient level (obtained under perfect competition), since the monopoly sets a price strictly higher than marginal cost.

For the natural monopoly, the situation gets more complicated. This is because the natural monopoly typically exhibits a decreasing total average cost, which implies that its marginal cost is lower than its average total cost. This situation is illustrated in the bottom panel of Fig. 1 for a firm with (large) fixed costs and (low or zero) constant marginal costs. In this case, the monopoly serves the entire market at a lower cost than multiple firms could achieve. For the natural monopoly, profits are still maximum at the point of intersection denoted by E , which is, however, the most undesirable situation for the society since it leads to high prices, small outputs, and a large welfare loss.

^d Consider, for example, the multiproduct cost function, when there are two outputs, 1 and 2:

$$C(y_1, y_2) = y_1 + y_2 + (y_1 y_2)^{1/3}.$$

Clearly, this cost function exhibits economies of scale when productions are strictly positive, but is never subadditive.

^e The maximum is attained provided that the second-order condition of the profit maximization program is satisfied.

^f The Lerner index is a measure of monopoly power since the higher its value, the more the firm is able to charge over its marginal cost. Lerner index ranges from 0 to 1. It cannot be negative under the assumption that negative profits are ruled out; and it cannot exceed one under profit maximization. In particular, it is equal to 0 in perfect competition where price equals marginal cost.

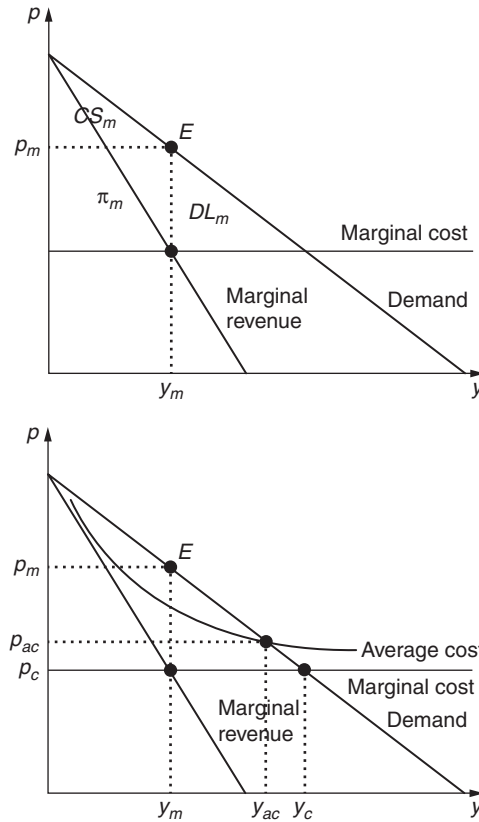


Figure 1 Monopoly and natural monopoly.

Regulating a Natural Monopoly

The inefficiency of the (natural) monopoly justifies its regulation, which aims to reduce its price and therefore increase its output. To address the inherent inefficient behavior of the monopoly, policymakers or governments can resort to regulation or public ownership (i.e., in the limiting case, they can decide to run the monopoly themselves, that is, opt for nationalization).

The choice of the regulated price is not easy. The government may want to set the price equal to the monopoly's marginal cost (marginal-cost pricing), so that efficiency is restored. However, this regulatory scheme faces two drawbacks.

First, the monopoly facing the marginal-cost pricing policy would incur losses and may, in turn, exit the market, since this policy leads to a price lower than average total cost (marginal cost being lower than average total cost). The government can address this problem, for example, by subsidizing the monopoly. However, in this case, the government incurs the loss, which can be covered by a tax that is associated itself to a deadweight loss. Alternatively, the government can allow a price higher than the marginal cost, for example, by choosing an average-cost pricing rule so that the monopoly just makes zero profit, which is associated to a lower deadweight loss.

Second, marginal-cost pricing does not provide the monopoly the incentives to reduce its costs. In a competitive market, firms can make higher profits by reducing their costs. By contrast, with the marginal-cost pricing rule, the regulated monopoly will not obtain higher profits by reducing its costs. The government can address this problem by designing a contract to induce the monopoly to reduce its cost as much as possible. Such incentives schemes are not simple to implement since effort of the monopoly is not directly observable.

To be more specific, consider the regulation of a monopoly producing M goods or services, indexed $m = 1, \dots, M$, when regulated prices are linear.⁸ In the Ramsey–Boiteux problem, the social surplus is maximized under the constraint that the firm (here the monopoly) breaks even. Let $S(y)$ denote the surplus that consumers derive from purchasing a vector of quantities $y = (y_1, \dots, y_M)$. The government solves:

$$\begin{aligned} \max_y \{ & S(y) - C(y) \} \\ \text{s.t. } & R(Y) - C(y) \geq 0, \end{aligned}$$

⁸ Linear prices are unit prices that are constant for each product and that therefore depend neither on the quantity sold (no second-degree price discrimination, involved, for example, in quantity discount), nor on the identity of the customers (no third-degree price discrimination, where customers with different characteristics pay different prices for the same good or service).

where, as earlier, $R(y)$ is total revenues and $C(y)$ is total costs. First, consider the simple case in which demands for the products are independent. The first-order conditions lead to the Ramsey–Boiteux pricing:

$$\frac{p_m(y) - C_m(y_m)}{p_m(y)} = \frac{\lambda}{1 + \lambda} \frac{1}{\eta_m}, m = 1, \dots, M,$$

where η_m denotes the own-price elasticity of good or service m , p_m its price, and C_m its marginal cost, and where $\lambda > 0$, the Lagrange multiplier represents the shadow price of the budget constraint (or the shadow cost of public funds with government transfers).

Accordingly, for each good or service, its Lerner index is inversely proportional to its own-price elasticity. However, it should be noted that the Lerner index is smaller than the inverse elasticity of the demand since $\lambda > 0$, whereas, as seen earlier, in the unregulated monopoly, the Lerner index is just equal to its corresponding inverse own-price elasticity of demand.^b

In practice, the regulator sets a price cap at the beginning of each period. The regulated price in period 1, p_1 , is given by:

$$p_1 = p_0(1 + \text{RPI} - X),$$

where p_0 is the regulated price in period 0, Retail Prices Index (RPI) is the inflation rate, and X is the *efficiency factor* (i.e., the expected efficiency improvements).ⁱ One period is typically between 3 and 5 years. The RPI can be measured by the Consumer Price Index—or RPI as used in the United Kingdom.^j The evaluation of X is trickier, since it depends on the evolution of the inputs price and on the expected change in productivity. In practice, the regulator resorts to some heuristic rules, rather than to a full econometric analysis, which may be hard to accomplish. Benchmarking is another alternative, although studies may not always be comparable, so that econometric analysis is required.

For example, Gagnepain and Ivaldi (2002) examine the impact of incentives in the case of public transportation (buses) in France. They examine how incentive compatible contracts (à la J.-J. Laffont) may induce the bus companies to lower their costs, and compare two different regulator contracts, the *cost-plus contracts* (based on observed costs and *ex-post* deficits are covered) and *fixed price contracts* (based on expected costs and expected deficits). They empirically show that fixed price contracts are more efficient to reduce costs than cost-plus contracts.

This study shed useful light on the importance of incentives in the design of the contracts. The efficiency of the contracts varies significantly according to the size of the network, the density of the customers, and the geographical characteristics. This is a common trait to many studies in transportation areas, such as airline, maritime, railroads, rail freights, or highways. Much work remains to be done to better understand the best way to regulate monopolies.

We cannot close this section without alluding to the fact that regulation may potentially reduce product innovation and process innovations. Lastly, note, as shown by Deneckere et al. (2019), that risk aversion of the principal (here the government) and the agent (here the monopoly) changes significantly the optimal contracts.

Econometrics of Natural Monopoly

With data on costs and input at hand, the cost function can be estimated to determine whether it is subadditive or not, that is, whether the industry under consideration is a natural monopoly or not. However, subadditivity is difficult to verify empirically. Fortunately, for the multiproduct case, a sufficient condition for the cost function to be subadditive is that its second partial derivatives are not positive over the relevant range of output. This condition, called “cost complementarity,” means that an increase in the production y_i of good or service i decreases the incremental cost of producing the quantity y_j of good or service j .

Cost complementarity may be hard to test empirically over the relevant range of output, but can easily be tested at the data point. Then, from an econometric point of view, we are interested in local conditions:

$$\frac{\partial^2 C}{\partial y_i \partial y_j} < 0.$$

The first step consists in assuming a functional form for the cost function that is able to identify whether or not there are cost complementarities. Flexible functional forms are usually used.^k For example, Foreman-Peck (1987) uses the generalized translog (GTL) multiproduct cost function to estimate the cost function of the British railways.

^bThis analysis can be extended to the case where products are not independent. If they are substitutes, the Ramsey–Boiteux prices are higher; if they are complements, prices are lower.

ⁱFor a discussion on how to measure efficiency, see Gagnepain and Ivaldi (2002).

^jSee http://oa.upm.es/43724/1/Mariana_Rodrigues_Brochado.pdf.

^kThe flexible functional form was introduced by Diewert (1974). A flexible cost function is able to approximate an arbitrary twice continuously differentiable cost function to the second order at the data point. This is the reason for which only a local measure of cost complementarity may be tested. See Diewert (1974) and the literature that follows for more details and for other examples of flexible functional forms.

The GTL multiproduct cost function model is defined as follows:

$$\ln CY(w) = \alpha_0 + \sum_{i=1}^N \alpha_i \ln(w_i) + \sum_{i=1}^M \beta_i Y_i + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_{ij} \ln(w_i) \ln(w_j) + \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \beta_{ij} Y_i Y_j + \sum_{i=1}^N \sum_{j=1}^M \rho_{ij} Y_j \ln(w_i) + \varepsilon,$$

where $w = (w_1, \dots, w_N)$ denotes the price vector of the N inputs, $Y = (Y^{\lambda_k} - 1)/\lambda_k$ denotes the Box-Cox transformation of the output vector of the M goods or services $y = (y_1, \dots, y_M)$, ε denotes the error term of the model, and the α s, the β s, and the γ s are parameters to be estimated.¹

Given the large number of parameters to be estimated, the statistical precision of parameter estimates can be improved by assuming that firms minimize their (input) costs to produce the exogenously predetermined levels of output. In turn, Shephard's lemma can be applied to the cost function C in order to obtain the following cost share equations:

$$S_i = \alpha_i + \sum_{j=1}^N \alpha_{ij} \ln(w_j) + \sum_{j=1}^M \rho_{ij} Y_j + \varepsilon_i, i = 1, \dots, N,$$

where ε_i denotes the error term of the model.^m The cost function can then be jointly estimated with $N - 1$ of the N share equations by using the method developed by Zellner (1962) for estimating seemingly unrelated regression models.

For the GTL multiproduct cost function model, the local cost complementarity is given by:

$$\frac{\partial^2 C}{\partial y_i \partial y_j} = \frac{C}{y_i y_j} \left[\frac{\partial \ln C}{\partial \ln y_i} \frac{\partial \ln C}{\partial \ln y_j} + \beta_{ij} y_i^{\lambda_i} y_j^{\lambda_j} \right],$$

which can be computed after estimation.

This methodology was used by Foreman-Peck (1987), to study the cost-structure of the railway industry in the United Kingdom during the 19th century in order to fuel the discussion of the efficiency of private versus public ownership.

Natural Monopolies in Transports: Panorama and Case Study

In transport, natural monopolies are important phenomena and arise, amongst other reasons, because the transport sector is capital intensive and needs large infrastructure to start producing. However, once fixed costs have been covered, the marginal cost to provide an extra unit of service is typically low. Since fixed costs are sunk, if an incumbent firm wishes to enter the market, the existing firm can easily cut prices to protect its market. On the other hand, if learning by doing is important, the incumbent firm may benefit from lower costs. Moreover, the incumbent firm can scream the market and serve the most profitable customers. For example, the intercity railways are more profitable than the regional railways, where demand is sparser.

With the opening of the market in the railway market in France (December 2019 for local train and December 2020 for intercity train), it is likely that the non-French competitors will first enter the most profitable niches. However, practice is somewhat different. It should be noted that the First European Railway Directive, which dates back to 1991, allowed open access for passengers and freight trains. In 2019, still not much competition occurs. Breaking State monopoly is in the agenda, but political and institutional barriers still remain very strong. The study of natural monopolies should not ignore their most important facet: the political economy dimension, such as electoral competition, centralized versus decentralized decisions, etc. Deregulation has been so far more successful in the airline industry or in the truck industry, even if several imperfections remain, as widely discussed by Joskow (2007).

The case of the British railways in the 19th century provides an interesting case study (Foreman-Peck, 1987). Competition has virtue to lower the price, while possibly leading to either duplication or underutilization of tracks. Moreover, competing firms may deny and make difficult interconnections. The Railway Clearing House, created in 1947, encouraged interconnection and fair competition. The estimations of Foreman-Peck (1987) suggest that before regulation, construction costs were 50% higher and national income per capita 0.75% lower than if would have been in a properly regulated market.ⁿ History teaches us that nationalization does not solve all problems. In 1911, the British railways were heavily regulated, yet the performance was poor

¹The GTL function of cost generalizes the translog function of cost by using the Box-Cox transformation, rather than the logarithm, for the output levels. It therefore allows to include zero outputs. The Box-Cox transformation reduces to the logarithm as λ_k approaches zero.

^mNote also that linear homogeneity (in input prices) of the cost function and symmetry of its Hessian matrix can be imposed by using the following linear restrictions on parameters:

$$\sum_{i=1}^N \alpha_i = 1, \sum_{j=1}^N \alpha_{ij} = 0, \sum_{j=1}^M \rho_{ij} = 0, \alpha_{ij} = \alpha_{ji}, \beta_{ij} = \beta_{ji}.$$

ⁿInterestingly, he says that in "1856 Belgian third class fares per miles were one quarter lower than the British fare and in 1883 40% less, while freight was similarly cheaper."

since competition was absent. However, privatization of British Rail, 20 years ago, was not a full success either, with high fares, low reliability, and little customer's support.^o

The study of natural monopoly is by far not completed and raises questions opened for deep debates. Possibly, the divorce of ownership and control may provide a solution to a problem that seems to never end.^p

Acknowledgment

We would like to thank Maria Börjesson for her useful comments and suggestions, as well as for her kind replies to our questions.

References

- Baumol, W., 1977. On the proper cost tests for natural monopoly in a multiproduct industry. *Am. Econ. Rev.* 67 (5), 809–822.
- Baumol, W., Bailey, E., Willig, R., 1977. Weak invisible hand theorems on the sustainability of multiproduct natural monopoly. *Am. Econ. Rev.* 67 (3), 350–365.
- Deneckere, R., de Palma, A., Leruth, L., 2019. Risk sharing in procurement. *Int. J. Ind. Organ.* 65, 173–220.
- Diewert, W.E., 1974. Applications of duality theory. In: Intriligator, M.D., Kendrick, D.A. (Eds.), *Frontiers of Quantitative Economics*, Vol. II. North-Holland Publishing Company, Amsterdam.
- Faulhaber, G., 1975. Cross-subsidization: pricing in public enterprises. *Am. Econ. Rev.* 65 (5), 966–977.
- Foreman-Peck, J., 1987. Natural monopoly and railway policy in the nineteenth century. *Oxf. Econ. Pap.* 39 (4), 699–718.
- Gagnepain, P., Ivaldi, M., 2002. Incentive regulatory policies: the case of public transit systems in France. *RAND J. Econ.* 33 (4), 605–629.
- Joskow, P.L., 2007. Regulation of natural monopoly. *Handbook Law Econ.* 2, 1227–1348.
- Mosca, M., 2008. On the origins of the concept of natural monopoly: economies of scale and competition. *Eur. J. Hist. Econ. Thought* 15 (2), 317–353.
- Smith, A., 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London.
- Walras, L., 1936. *L'Etat et le chemin de fer*. Reprinted from: *Etudes d'économie politique appliquée*. R. Pichon et R. Durand-Auzias, Paris, pp. 193–236 (1936).
- Zellner, A., 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Am. Stat. Assoc.* 57 (298), 348–368.

^o The majority of United Kingdom wish to revert the privatization of British Rail. According to the Office of Rail and Road, as of 2016 there was 62% support for public ownership of train-operating companies. See https://en.wikipedia.org/wiki/Impact_of_the_privatisation_of_British_Rail.

^p See, for example, <http://www.cirje.e.u-tokyo.ac.jp/research/dp/2012/2012cf864.pdf>.

Freight Costs: Air and Sea

Yulai Wan, Dong Yang, Hong Kong Polytechnic University, Hong Kong, China

© 2021 Elsevier Ltd. All rights reserved.

Introduction	36
Freight Cost Components: Air Versus Sea	36
Key Influential Factors of Costs	38
Non-Price Factors of Costs: Air Freight	38
Non-Price Factors of Costs: Shipping	40
Trip (Voyage) Cost Functions	41
Company-Level Cost Functions	42
Variables Included in the Cost Function of Cargo Airlines	43
Variables Included in the Cost Function of Shipping Companies	43
Economies of Scale and Density: Evidence From Company-Level Cost Functions	44
Conclusion	45
References	45

Introduction

A freight cost function is a mathematical formula which expresses the total cost of moving and/or handling freight as a function of some major influential factors, such as the amount of freight shipped, line-haul (trip or voyage) distance, network size, etc. Freight cost functions can be established to express either the cost of a single trip (or voyage) or the cost borne by a freight shipping company, such as all-cargo airlines and container shipping lines, over a period of time. This chapter focuses on freight cost functions of two transport modes, air and shipping. The reason of discussing these two modes of transport in one chapter is that both modes share lots of similarity in cost structures while at the same time each possesses some special features.

Before understanding how major factors associate with freight costs, one must understand the cost components of each mode of transport. Although capital, labor, and energy are the main inputs of providing transportation services, this classification does not associate costs with various key operations and activities and provides limited managerial insights. Thus, in Section, "Freight Cost Components: Air Versus Sea," costs are classified based on their associated functions, such as line-haul operation (flying the airplanes or sailing the ships) and ground handling/terminal operation, etc. Section, "Key Influential Factors of Costs" explains how several key influential factors affect costs in air freight transport and shipping, respectively. Trip (voyage) cost functions are presented in Section, "Trip (Voyage) Cost Functions." In Section, "Company-Level Cost Functions," company-level cost functions are discussed as well as the economies of scale and density. Section, "Conclusion" concludes this chapter.

Freight Cost Components: Air Versus Sea

There are different ways to classify costs. A common method in the economics literature is to divide the total cost into fixed and variable costs. Costs related to the fleet, equipment, and overhead are usually considered as fixed while those related to labor and energy and providers of ground handling services are usually considered as variable. On the voyage basis, in the shipping literature, capital costs of owning vessels, including the costs of ships and related financing costs, are separately listed in addition to fixed and variable costs. As a result, fixed costs only include costs paid to maintain the ships in navigable conditions and the fixed port charges. However, whether an input is fixed or variable depends on the time horizon and the type of decisions to be made. In a fleet planning scenario, fleet cost is obviously variable, but if one only investigates a particular flight or voyage with a given airplane or ship, many inputs can be fixed, even for certain part of the labor and fuel as well as airport/terminal charges. Thus, such discussion has to be grounded on specific cases and is not the focus of this chapter.

In terms of functions, air and sea shipping share lots of commonalities. According to [Table 1](#), both include three major aspects with different terms but similar meanings, that is, line-haul activities, terminal activities, and system operating activities. In the context of air transport, both ground handling and system operating costs are considered as indirect costs since they do not directly associate with operating the flights while flight operation costs are also called "direct operating costs" (DOC). Various organizations have somewhat different standards when asking airlines to report their costs, but the main ideas are similar. One major difference between these two modes comes from the treatment of airport/port charges. Airport charges levied on aircraft (i.e., landing/take-off and parking charges) are considered as terminal costs while port charges levied on ships are considered as ship (line-haul) costs [Stopford \(2009\)](#). Port charges include pilotage, mooring/unmooring fees, wharfage, towage, tonnage dues, light dues, port state, pollution control fees, etc.

Table 1 Function-based cost components

Functional areas	Air	Sea
Line-haul	Flight operation costs: costs associate with owning, renting, maintaining, and flying the aircraft. Fuel and flight crew (pilots) costs are also included.	Ship costs: ship capital and financing costs, crew costs, ship repair and maintenance costs, bunker costs, port costs
Terminal	Ground handling costs: costs associate with servicing aircraft, landing/parking aircraft at the airports, loading/unloading/processing cargo at airport terminals, as well as sales and promotion	Terminal and cargo costs: terminal costs for handling/storing containers, costs associate with maintaining and supplying containers, agency costs
System operating	Costs associate with administration, overheads, customer services, and other transport-related, such as services offered by partner airlines	Administration/overhead costs, inland transport costs, ship repositioning costs

In addition to cost items listed in **Table 1**, special cost items need to be added when ships are operated on certain routes. When the ships pass through canals, there incurs canal fees and the impact of congestion on cost should be taken into account. The costs for safety equipment and guards should be added if ships transit in zones with high risk of piracy attacks. Ice-breaking fees are sometimes required for voyages through the Arctic Ocean.

Based on the standard of US Department of Transport (DOT), **Table 2** shows cost decomposition of selected US-based cargo airlines and passenger airlines in 2017. **Table 2** compares cost decomposition of three groups of airlines: FedEx, the other three major cargo airlines in the United States, namely Atlas Air, Polar Air Cargo Airways, United Parcel Service (UPS), and selected major passenger airlines, including Alaska Airlines, American Airlines, Delta Air Lines, Envoy Air, Frontier Airlines, Hawaiian Airlines, Southwest Airlines, Spirit Air Lines, United Air Lines, and Virgin America. FedEx is separately reported, because it counts costs of ground operating/trucking services provided by its “sister” companies into transport-related costs, which is broadly defined as expenses applicable to the generation of transport-related revenues. However, the other three cargo airlines assign costs of subcontracted ground/trucking services into individual cost components based on the nature of subcontracted services instead of bundling them into transport-related costs (Onghena, 2011). This different treatment causes FedEx an unusually low share flight operation costs (21.77% compared with 82.01% of the other cargo airlines) and an unusually high share of transport-related cost (57.78% compared with 0.30% of the other cargo airlines). As a result, FedEx cost structure is very similar to passenger airlines as mainline passenger airlines rely a lot on regional airlines to provide feeding services of which the expenses are also counted in transport-related costs. However, taking transport-related costs out of the picture, one can still conclude that flight operating costs account for majority of a cargo airlines’ total operation costs of which fuel and oil, flight equipment maintenance and flight crew are the top three largest cost components. In addition, aircraft and traffic serving costs account for another large trunk.

Table 3 provides an example of cost shares of a typical voyage of a container liner service. Costs associated with ships account for 37%–54% of total costs. After excluding port costs, about 33%–47% of the total voyage costs associate with line-haul activities. This number is similar to air freight transportation. Terminal and container costs account for another 19%–26%. Since seaports are located far away from inland regions, inland intermodal transport accounts for a trunk share (17%–23%) of total container shipping

Table 2 Decomposition of total operating cost of selected US-based cargo airlines in 2017

Cost category	FedEx	Other selected cargo airlines	Selected passenger airlines
Fuel and oil	6.17%	28.16%	17.43%
Maintenance–flight equipment	6.29%	18.16%	8.81%
Flight crew	4.79%	17.22%	11.78%
Rentals and insurance–flight equipment	1.69%	11.79%	2.71%
Depreciation–flight equipment	2.55%	6.45%	3.77%
Other–flying operation	0.28%	0.23%	0.01%
<i>Total DOC (flight operating costs)</i>	<i>21.77%</i>	<i>82.01%</i>	<i>44.51%</i>
Depreciation costs–maintenance equipment	0.49%	0.00%	0.03%
Amortization–other than flight equipment	0.00%	0.08%	0.32%
Aircraft servicing costs	2.02%	11.42%	6.04%
Passenger and traffic service costs	8.03%	0.56%	18.31%
Reservation and sales costs	0.66%	0.96%	5.11%
Advertising and publicity costs	0.30%	0.00%	0.92%
General and administrative costs	7.31%	4.17%	9.15%
Maintenance and depreciation–ground	1.63%	0.51%	1.95%
<i>Total servicing, sales and general operating costs</i>	<i>20.45%</i>	<i>17.69%</i>	<i>41.84%</i>
Transport-related costs	57.78%	0.30%	13.65%

Source: US DOT Form 41 Data.

Table 3 Cost structure of a hypothetical voyage by containership sizes

Vessel size (TEU)	1200	2600	4300	6500	8500
Operating costs	8.83%	5.54%	3.62%	2.82%	2.55%
Capital costs	16.92%	16.55%	14.41%	13.78%	14.30%
Bunker costs	21.02%	18.70%	19.92%	18.98%	16.01%
Port costs	7.60%	5.46%	4.28%	3.66%	4.05%
<i>Total ship costs</i>	<i>54.37%</i>	<i>46.25%</i>	<i>42.23%</i>	<i>39.24%</i>	<i>36.91%</i>
Cost of supplying containers	1.58%	1.75%	1.82%	1.97%	2.17%
Cost of container maintenance	1.68%	1.99%	2.15%	2.25%	2.33%
Terminal costs for container handling	14.85%	17.52%	18.85%	19.81%	20.54%
Refrigeration cost for reefer containers	0.54%	0.62%	0.68%	0.72%	0.74%
<i>Total terminal and container costs</i>	<i>18.65%</i>	<i>21.88%</i>	<i>23.50%</i>	<i>24.74%</i>	<i>25.78%</i>
Administrative cost per voyage	7.20%	8.52%	9.16%	9.62%	9.98%
Inland intermodal transport cost	16.92%	19.99%	21.51%	22.59%	23.42%
Interzone repositioning	2.86%	3.36%	3.60%	3.79%	3.93%
Cargo claims	2.32%	2.71%	2.94%	3.09%	3.20%
<i>Total system operating costs</i>	<i>29.30%</i>	<i>34.59%</i>	<i>37.21%</i>	<i>39.09%</i>	<i>40.52%</i>

Source: Compiled based on [Stopford, 2009](#).

costs. The other top three cost components are bunker, ship capital, and container handling costs. The cost structure is very consistent across ship sizes, while larger ships tend to have smaller share of line-haul costs and larger share of terminal handling and inland transport costs.

Key Influential Factors of Costs

Prices of various inputs, such as crew wage rate, jet fuel, or bunker prices and prices of aircraft or ships, affect cost of freight operation. Pilot wage rate can vary significantly across countries, seniority, and airlines. Sea-going crew wage is related to the nationality of the crew, which is linked to the registration of ships. Jet fuel and bunker prices may also vary across locations and of course are affected by various external shocks mainly out of the control of freight operators. Prices of aircraft tend to be high for newly developed models and new aircraft. Old aircraft are much cheaper due to most likely lower fuel efficiency and higher maintenance costs. Prices of ships are related to the type, size, and design of the ship as well as the freight rate in the market. Given that it is almost impossible to have a clear and concise discussion on the dynamics of these input prices, our focus is on operational characteristics, which influence freight operation costs, or more precisely productivity.

Non-Price Factors of Costs: Air Freight

Direct operating costs can substantially vary across aircraft models. Cost per block hour, cost per available ton-km (ATK), and cost per revenue ton-km (RTK) are widely used to compare the costs of operating different aircrafts in different context ([Morrel, 2011](#)). Cost per block hour tells the costs of having an aircraft in use by one hour. Roughly speaking, the counting of block hour of a flight cycle starts when the aircraft leaves the gate for departure and ends when the aircraft arrives at the gate of the destination terminal. Cost per ATK and cost per RTK are usually considered as unit cost, since they not only divide the cost by the distance flown but also the amount of available cargo carrying capacity (for ATK) and the amount of actual cargo carried (for RTK), while cost per block hour does not assign costs to each ton of cargo. [Table 4](#) presents DOC per block hour, DOC per ATK and DOC per RTK of some representative freighters, that is, all-cargo aircraft, based on data reported by FedEx and UPS. In addition, [Table 4](#) also lists some major influential factors of DOC, such as average payload, average stage length (i.e., distance per flight segment), and block hours per day.

In general, for the same aircraft model, cost per block hour is highly affected by block hours per day. For example, the DOC per block hour of UPS' A300-600 is much lower than the same aircraft model of FedEx, while the block hour per day of the former is 1.59 hours more than the latter. Block hours per day tells the number of hours on average an aircraft is in use and hence it reflects aircraft utilization, that is, high block hours per day means high aircraft utilization. While the aircraft capital costs (i.e., depreciation and rentals) are associated with the time that the aircraft is owned and rented by the airline, the generation of freight transport services is linked to the time that the aircraft is in use. Note that although in a typical aircraft leasing agreement, the lessee pays for each block hour, there is usually a minimum block hour limits that the lessee must pay for. Therefore, as the utilization of an aircraft increases, its capital costs can be spread over more block hours, leading to lower cost per block hour. This can be seen by comparing the sums of depreciation and rental costs of B767-300. Although for this model FedEx has lower DOC per block hour than UPS, which might be caused by FedEx's outsourcing of maintenance costs, the sum of hourly depreciation and rental costs is still lower for UPS, which achieves higher block hours per day.

Table 4 Examples of typical freighter aircraft and their DOC (2017)

	<i>FedEx A300-600</i>	<i>FedEx B757-200</i>	<i>FedEx B767-300</i>	<i>UPS A300-600</i>	<i>UPS B757-200</i>	<i>UPS B767-300</i>	<i>FedEx B777-F</i>	<i>FedEx DC-10-30</i>	<i>FedEx MD-11</i>	<i>UPS B747-400</i>	<i>UPS MD-11</i>
Design range (km)	7,500	5,830	7,200	7,500	5,830	7,200	9200	10,620	12,455	13,490	12,455
Average available payload (tons)	53	32	63	52	38	59	116	85	96	111	90
Average stage length (km)	1,255	1,104	1,742	1,158	1,114	2,169	6,357	1,425	2,732	5,870	3,200
Total block hours	92,425	110,566	100,378	72,286	72,743	155,245	121,679	22,218	154,648	51,683	90,109
Block hours per day	3.00	2.15	4.80	4.59	3.22	9.28	10.83	3.96	6.74	14.64	8.83
Total distance flown (000 km)	53,027	49,499	54,460	43,583	40,540	89,171	62,888	13,929	97,085	27,553	52,113
Total ATK (000)	2,616,232	1,814,730	3,719,672	2,050,286	1,486,196	5,745,775	9,678,080	1,070,750	9,526,508	3,996,539	5,367,550
Total RTK (000)	1,470,998	938,168	2,222,851	1,191,848	686,470	3,771,109	5,540,725	520,762	5,220,772	2,771,313	3,381,741
US gallons fuel/hour	1,553	974	1,451	1,536	1,116	1,484	2,302	2,165	2,253	3,303	2,464
US gallons fuel/ATK	55	59	39	54	55	40	29	45	37	43	41
Fuel price (US\$ per gallon)	1.68	1.67	1.75	1.99	2.00	1.87	1.73	1.75	1.70	1.78	1.85
Daily average number of aircraft	84.50	141.19	57.33	43.11	61.94	45.85	30.78	15.38	62.91	9.67	27.96
Costs per block hour:											
Flight crew	2,368	1,811	2,050	2,319	2,300	2,343	2,578	110	3,708	2,327	2,342
Fuel	2,604	1,630	2,532	3,056	2,227	2,781	3,979	3,781	3,822	5,874	4,554
Maintenance	4,635	3,078	352	2,272	2,781	2,141	1,786	4,176	4,610	2,425	3,179
Depreciation	957	1,994	1,153	1,345	903	481	1,321	346	953	751	1,536
Aircraft rental	2,229	0	0	32	125	393	188	0	609	228	7
Other flight costs	175	141	85	103	31	70	6	129	152	56	54
DOC per block hour	12,967	8,654	6,171	9,126	8,367	8,209	9,857	8,543	13,854	11,661	11,672
Unit costs:											
DOC per ATK (US cents)	458	527	167	322	410	222	124	177	225	151	196
DOC per RTK (US cents)	815	1,020	279	553	887	338	216	364	410	217	311

Note: Payload equals to the certificated takeoff weight of an aircraft, less the empty weight, less all justifiable aircraft equipment, and less the operating load (consisting of minimum fuel load, oil, flight crew, Steward's supplies, etc.)

Source: Data from US DOT Form 41.

The average stage length also relates to costs. On one hand, aircraft utilization might be improved by flying a longer distance, as making a stop takes lots of time and longer flying distance means that fewer stops will be made during a given period of time. On the other hand, given an aircraft model, although longer flight distance suggests more fuel consumption during the cruise and more payment to flight crew, some costs are relatively independent of flight distance, such as fuel consumption during the taxiing, takeoff and landing stages, flight cycle-driven maintenance, as well as various handling and servicing costs incurred at the terminals. As a result, the costs of moving freight by one kilometer decreases as the average stage length increases, leading to a decrease in DOC per ATK and DOC per RTK. This can be seen by comparing the unit costs of MD-11: with a longer average stage length, UPS incurs lower unit costs even though its average available payload is lower than FedEx. However, according to the range-payload tradeoff, for any given aircraft type, once a certain distance is reached, further increasing stage length requires a reduction of the aircraft's freight carrying capacity (i.e., payload capacity). This is because to fly a longer distance, more fuel has to be carried on board, and as a result less freight can be carried to make sure that the aircraft does not exceed the maximum takeoff weight. Therefore, increasing stage length far above the most efficient range may lead to an increase in unit costs.

Aircraft size, or more accurately aircraft carrying capacity is another influential factor. As a large portion of a flight trip's costs is independent of the carrying capacity or the actual amount of freight carried. For example, the crew size is almost fixed regardless the size of the aircraft. Aircraft maintenance costs are affected by aircraft size, but not in a linear way. This feature is called economies of scale by some researchers while others call this economies of aircraft size or economies of traffic density, since scale in many contexts also refers to the production scale of an airline, which will be mentioned later. From [Table 4](#), one can easily observe that in general aircraft with larger average payload incurs lower DOC per ATK or DOC per RTK. Of course, larger aircraft also tend to be operated in longer routes, and thus the impact of average stage length may jointly play a role here.

In addition to carrying capacity, load factor which tells the actual proportion of the capacity utilized to carry freight affects DOC per RTK. Although more fuel needs to be consumed as more weight is carried, the amount of fuel for lifting the aircraft itself without any payload can be spread over more tons of freight carried if the load factor increases, leading to lower DOC per RTK. Scheduled flights tend to have lower load factor than charter services. Note that although each aircraft has fixed volume-metric capacity (space available for freight), the payload (weight) capacity depends on flight distance, weather condition, and many other operational considerations, such as the weight of passengers and their baggage in the case passenger aircraft. While, fuel consumption mainly relates to the weight of the cargo rather than the space it takes, and unit costs are usually calculated based on the weight of the freight. Thus, the load factor based on the weight capacity is more relevant than the volume capacity.

The above discussion applies to freighters which are mainly operated by all-cargo airlines, integrators (i.e., air express airlines) and combination airlines that operate both freighters and passenger aircraft (for example, Lufthansa, Cathay Pacific Airways). Passenger airlines, which do not operate freighters, carry cargo in the belly hold of passenger aircraft. Costing of belly cargo operation depends on how to allocate the joint costs between passengers and cargo, such as basic fuel without payload, flight crew costs, aircraft capital costs, aircraft insurance and maintenance, as well as landing fees and navigation service charges. Most passenger airlines allocate all joint costs to the passenger side of the business and assign only the incremental costs incurred by cargo-specific operations, such as cargo sales and promotion, freight insurance and additional fuel due to carriage of cargo, to the cargo services. This leads to substantially lower freight unit costs than those incurred by freighters.

Non-Price Factors of Costs: Shipping

Unlike aircraft, ships vary substantially in sizes. The capacity of a mega container ship can be 10 times as much as the capacity of a feeder container ship. In general, maritime shipping enjoys a much stronger economies of ship size. [Table 5](#) compares the unit costs of a typical voyage by containership sizes and unit costs per year by bulk ship sizes. The ship size is measured in twenty-foot equivalent unit (TEU) for containerships and deadweight tons (DWT) for bulk shipping. The total voyage cost increases as ship size increases while cost per TEU or cost per DWT declines. As the demand for freight transport is directionally imbalanced, costs can be highly different in directions. For example, eastbound voyage tends to move more containers than westbound, and therefore on per TEU basis, eastbound incurs lower cost.

Table 5 Cost per TEU or per DWT by ship sizes

Containership size (TEU)	1,200	2,600	4,300	6,500	8,500
Total voyage cost (\$'000)	2,027	3,721	5,719	8,229	10,382
Cost per TEU eastbound leg (\$)	938	795	739	703	679
Cost per TEU westbound leg (\$)	2,111	1,789	1,662	1,582	1,527
Average cost per TEU (\$)	1,299	1,101	1,023	974	940
Change in cost per TEU		-198	-78	-49	-34
% change in cost per TEU		-15.24%	-7.08%	-4.79%	-3.49%
Bulk ship size (DWT)	30,000	47,000	68,000	170,000	
Cost per DWT per year (\$)	191	143	120	74	
% change in cost per DWT per year		-25.13%	-16.08%	-38.33%	

Source: Modified based on [Stopford, 2009](#).

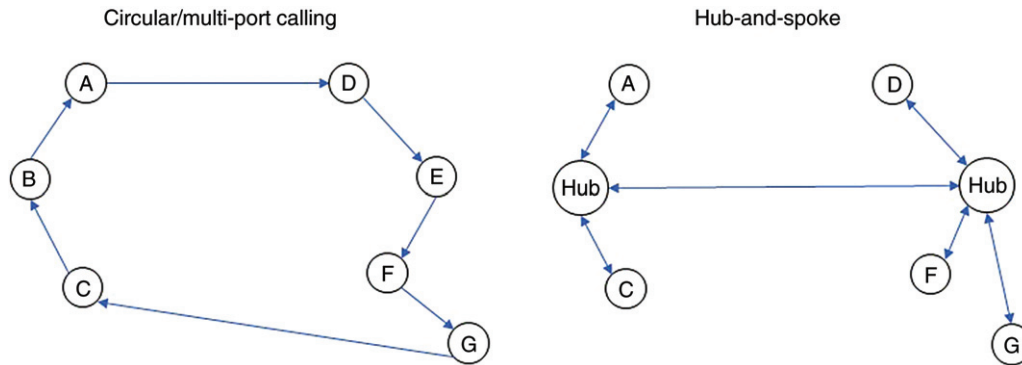


Figure 1 Major network structures of cargo airline or shipping lines.

Bunker (fuel) cost in shipping is inextricably linked to ship speed and load factor given a particular ship. It is roughly estimated that for a given ship, fuel consumption of a ship has a cubical (or exponential) relationship with the speed. Unlike aircraft, ships have a much larger room for speed changes. For example, the feasible speed of container ships may range from 12–15 knots (the lowest speed possible) to 20–25 knots (the designed optimal speed). Technically, reducing speed below this range will not bring any further reduction of fuel consumption; whilst, increasing speed above this range will lead to too much fuel consumption which is hardly compensated by reduced travel time. A 10%–20% speed cut (slow steaming) below the designed speed is very common when demand is low. To maintain stability, every ship has to reach a minimum displacement during the voyage and hence there is a minimum amount of fuel, which has to be consumed. When the load factor is below a threshold, cargo carried does not produce sufficient displacement and some ballast water has to be carried to reach the minimum displacement. As a result, the fuel consumption per unit of cargo will be increased, as cargo has to share the cost of carrying the ballast water. Once the threshold load factor is reached, although fuel consumption increases as displacement increases, ballast water is no longer needed and the fuel cost per unit of cargo will be reduced.

Shipping network also affects costs. Cargo airlines and maritime shipping companies mainly apply two network structures: circular (multi-port calling) network and hub-and-spoke network (Fig. 1). To satisfy the same shipping demand, circular network incurs higher ship costs than hub-and-spoke network mainly because the latter uses larger ships and requires fewer port calls and hence lower port costs (Imai et al., 2009). Sometimes ports can be highly congested, leading to the long waiting time and increased costs borne by carriers. However, the costs of terminal handling, feeding hub ports, and empty container management tends to be higher when hub-and-spoke network is used (Imai et al., 2009). This is caused by not only extra unloading and loading of containers at the hub for transshipment but also longer transit time and more handling of empty containers. That is, in highly directionally imbalanced trade routes, hub-and-spoke network is disadvantaged in terms of total costs.

Trip (Voyage) Cost Functions

Calculating a flight trip cost of an aircraft or a voyage cost of a ship is essential when making aircraft or vessel purchasing and deployment decisions. In practice, trip (voyage) cost function is estimated by adding up expenses of different freight transportation functions during a trip (voyage). These cost components are very similar for aviation and shipping (Table 6). The landing fees, navigation fees, and ground handling fees in aviation correspond to port charges and cargo handling fee in shipping.

Table 6 Cost components included in aviation (ATA, NASA, and AEA) and shipping

Components	Aviation			Shipping
	ATA	NASA	AEA	
Depreciation	✓	✓	✓	✓
Insurance	✓	✓	✓	✓
Flight crew (Crew cost)	✓	✓	✓	✓
Fuel	✓	✓	✓	✓
Maintenance	✓	✓	✓	✓
Landing fees (Port charges)		✓	✓	✓
Navigation fee (Port charges)		✓	✓	✓
Interest		✓	✓	✓
Cabin crew (Crew cost)		✓	✓	✓
Ground handling fee (Cargo handling fee)				✓

Source: Modified based on Ali and Al-Shamma, 2014.

Table 7 Examples of trip and unit cost functions

Aircraft type	Trip cost	Unit cost
Passenger (Swan and Adler, 2006)	Short-haul single-aisle airplanes (1000 and 5000 km) $C = (D + 722) * (S + 104) * 0.019$ Long-haul twin-aisle airplanes, 2-class seating $C = (D + 2200) * (S + 211) * 0.0115$	Regional single-aisle airplanes $c = 2.44S^{-0.40}D^{-0.25}$ Long-haul twin-aisle airplanes $c = 0.64S^{-0.345}D^{-0.088}$
Any (Ali and Al-Shamma, 2014)	Trip cost based on designed stage length $C = -4.497 \times 10^{-7} \times MTOW^2$ $+ 0.9588 \times MTOW - 33214$	
Notations	C: aircraft trip cost; c: cost per seat kilometer; D: flight distance; S: seat count	

In air transport, three common methods are used to calculate the cost of a trip. They are proposed by the Air Transportation Association of America (ATA), National Aeronautics and Space Administration (NASA), and Association of European Airlines (AEA), respectively. The NASA and AEA methods include not only DOC but also some indirect cost items, but the ATA method only takes into account DOC (Air Transport Association of America, 1967) (Table 6). The trip cost is usually estimated as a function of various parameters, such as aircraft cost, airframe cost, engine cost, airframe weight, insurance rate, block hours of the flight, maximum takeoff weight (MTOW), etc.

As the costing procedure is complicated due to various details from the engineering perspectives, sometimes for higher-level strategic decisions, one needs a simpler way to link a trip costs with only a few key factors. This type of studies is rare for freighters, but there are studies for passenger aircraft. The upper part of Table 7 lists estimated functional forms of passenger flights' trip and unit costs. The resulting trip cost functions are functions of stage length and seat capacity. Similar idea can be applied to approximate trip cost of cargo flights by replacing seat capacity with payload capacity. Alternatively, a trip cost can be estimated based on MTOW and this method is valid for any aircraft type.

Cost functions listed in Table 7 well reflect the discussion in section "Key Influential Factors of Costs." While trip costs increase in flight distance (D) and aircraft size (S), unit cost is decreasing in stage length and aircraft size. In general, the speed of cost increase slows down as the MTOW increases. Therefore, twin-aisle (wide-body) aircraft, which are mainly operated in long-haul markets are likely to incur lower unit cost than single-aisle (narrow-body) aircraft operating in short-haul markets. Moreover, based on Table 7, one can also observe that in general single-aisle aircraft are more sensitive to aircraft size and flight distance than twin-aisle aircraft.

In sea shipping, we did not find any study which establishes a simple functional form linking voyage cost and one or two key factors. Rather, voyage cost is mainly expressed and evaluated as the sum of the detailed cost components based on certain assumptions.

Company-Level Cost Functions

As many cost items, especially indirect costs, are difficult to be assigned to a particular trip or voyage, it is sometimes relevant to understand at company level, how different factors influence the cost efficiency of a freight operator. Econometric models and regression analysis are widely used to achieve this objective. Costs of airlines or shipping companies are treated as the dependent variable and various factors are treated as independent variables to establish a functional form:

$$TC = f(Q, P, Z, F, N, T),$$

where TC is the total cost incurred by an operator, Q is the total output or traffic served by the operator, P is a set of input prices, Z is a set of operation-related factors, N is network size, F is a set of firm-specific unobservable characteristics captured by dummy variables, and T is a set of time-related variables. Total output and input prices are the fundamental part of these models, while other independent variables are added according to the issues studied and the specific features of the dataset.

The Cobb-Douglas function has been widely used in the model specification for cost estimations. When only output and prices of n inputs are taken into account, the total cost can be written in the following way:

$$TC = e^{a_0} Q^{\beta_Q} \prod_{i=1}^n p_i^{a_i}.$$

Then, a log transformation can be applied so that the following form is actually estimated statistically: $\ln TC = a_0 + \beta_Q \ln Q + \sum_{i=1}^n a_i \ln p_i$

This transformation makes it straightforward to identify the individual input's cost share among total cost. Differentiating both sides of the above equation with respect to $\ln p_i$ and applying Shephard's Lemma, one can show that the cost share of input i is exactly coefficient a_i . One can also easily obtain the elasticity of total cost with respect to output (ϵ_Q) as it is equivalent to coefficient β_Q .

Therefore, log-transformed Cobb-Douglas function is usually applied instead of Cobb-Douglas function itself. A more complicated specification is the general translog specification which uses a second-order Taylor expansion to approximate to the cost function. That is,

$$\ln TC = a_0 + \beta_Q \ln Q + \sum_{i=1}^n a_i \ln P_i + \frac{1}{2} \gamma_Q (\ln Q)^2 + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \gamma_{ij} \ln P_i \ln P_j + \sum_{i=1}^n \delta_i \ln Q \ln P_i$$

where $\gamma_{ij} = \gamma_{ji}$, $\sum_{i=1}^n a_i = 1$, and $\sum_{i=1}^n \gamma_{ij} = \sum_{i=1}^n \delta_i = 0$. Similar to the log transformation of Cobb-Douglas function, one can easily obtain the cost share of each input by taking the partial derivative of the general translog function. In particular, $\frac{\partial \ln TC}{\partial \ln T_i}$ generates the cost share of input i , and the elasticity of total cost with respect to output will be

$$\epsilon_Q = \frac{\partial \ln TC}{\partial \ln Q} = \beta_Q + \sum_{i=1}^n \delta_i \ln P_i$$

In addition to output and input prices, the other control variables (Z , F , T) can be easily added into the equations with or without the log transformation, depending on the particular variable.

Variables Included in the Cost Function of Cargo Airlines

In air cargo transportation, both TC and variable cost (VC) have been used as dependent variables. Output is usually measured by total revenue ton-miles (RTM) total RTK of freight and mail. Four inputs' prices (P) are commonly considered: fuel, labor, material, and capital. In terms of vector Z , two operation-related variables are widely included: average stage length and load factor. The total number of airports served (N) is a key indicator of network size. Inclusion of network size is essential when one needs to measure and test the presence of economies of density and economies of scale for cargo airlines. As the difference between fixed and VC are relevant only when capital is fixed, when estimating VC , capital stock (K) is included as a control variable while price of capital is dropped. Capital stock can be measured by the asset value plus investment in flight equipment, ground equipment and property, capital leases, and land. After reviewing relevant studies in the literature, we list the studied cargo airlines, functional specifications, and variables in Table 8.

Variables Included in the Cost Function of Shipping Companies

Total cost, VC , and unit cost (UC) have been used as dependent variables in shipping. In bulk shipping, the output (Q) is measured in ton-miles. In container shipping, as the capacity of container ships is measured in TEU, the output (Q) is measured in TEU-miles transported by shipping lines. Different input prices (P) are used in estimations, including labor, fuel and oil, stores and materials, repairs and maintenance, capital stock, etc. Unlike air transport, size of the capital (K) is usually included and it is measured by the fleet capacity (Table 9). Some studies consider K as a control variable when estimating the functional form of VC of the short-run costs, while others include it even when estimating the functions of TC . Moreover, another focus of shipping is average ship size (S). Operation-related variables (Z) might include sailing distance, slot utilization, and freight rate.

Table 8 Selected studies on cost functions of all-cargo airlines

Authors	Airlines	Specification	Functional form	Observations	Output Q	EOD	EOS
Kiesling and Hansen (1993)	FedEx	$TC = f(Q, P, Z, N, T)$	Cobb-Douglas	Quarterly, 1986Q1–1992Q3	RTM	2.36–4.07	0.54–0.62
Onghena et al. (2014)	FedEx, UPS	$TC = f(Q, P, Z, N, F, T)$ $VC = f(Q, P, Z, N, K, F, T)$	Translog	Quarterly, 1990Q1–2010Q2	RTK	FedEx: 1.75–3.15 UPS: 2.06–3.07 After controlling for capital stock: FedEx: 2.44 UPS: 2.11	FedEx: 1.45–2.72 UPS: 2.04–3.46 After controlling for capital stock: FedEx: 2.05 UPS: 1.97
Lakew (2014)	FedEx, UPS	$TC = f(Q, P, Z, N, T)$	Translog	Quarterly, 1993Q3–2013Q4	RTM	4.525	3.077
Roberts (2014)	FedEx, UPS	$TC = f(Q, P, Z, N, F)$	Cobb-Douglas	Quarterly, 2003Q1–2011Q4	RTM	FedEx: 1.60 UPS: 3.02	FedEx: 0.87 UPS: 0.81
Balliauw et al. (2018)	UPS, FedEx, Polar, Atlas, Southern, ABX, Evergreen, Kalitta	$TC = f(Q, P, Z, N, F, T)$	Translog	Annual, 1990–2014	RTK	Integrator: 1.66 Nonintegrator: 1.34 Pooled: 1.29	Integrator: 1.63 Nonintegrator: 1.21 Pooled: 1.22

Note: EOS and EOD provided here are estimated at the sample means.

Table 9 Selected studies on cost functions of shipping companies

Authors	Shipping companies	Specification	Functional form	Observations	Output Q	EOS	ε_S	ε_K
Wu and Lin (2015)	Evergreen, Yang Ming, Wan Hai	$TC = f(Q, P, S, T)$	Translog	Annual, 1991–2012	TEU-miles	1.30	−0.24	
Tran and Haasis (2015)	Top 16 publicly traded container shipping lines	$TC = f(P, S, K, Z)$ $UC = f(P, S, K, Z)$	Cobb-Douglas	Annual, 1997–2012			No effect	TC: 0.86 UC: −0.14
Tolofari et al. (1987)	86 bulk carriers and 47 tankers	$TC = f(Q, P)$	Translog	Cross-sectional, 1982	Ton-miles	Bulk: 1.69 Tanker: 2.0		

Note: EOS and elasticities provided here are evaluated at the sample means.

Economies of Scale and Density: Evidence From Company-Level Cost Functions

After estimating the company-level cost functions, one can further test the existence of economies of scale and economies of density of the carriers' operation. In general, economies of scale refers to the cost advantage obtained by a firm when UC decreases in output. In a word, economies of scale exists if the average cost of producing one unit of output is above the marginal cost of producing one incremental output. Thus, the ratio of average cost and marginal cost, denoted as EOS, can be used to quantify economies of scale. If this ratio is above 1, economies of scale exists; if it is below 1, diseconomies of scale exists. Based on the definition of elasticity, this ratio is exactly the inverse of the cost elasticity with respect to output. That is,

$$EOS = \frac{1}{\varepsilon_Q}$$

The above definition has been modified in the context of airline operations. As most airlines operate an extensive network, output is affected by not only the amount of output on each route, but also the number of airports served, that is, network size. Therefore, in air transport, EOS describes the impact on cost by having the same proportional changes in both output and network size, while controlling for operation-related factors, such as load factor and stage length. In this case, as both output and network size change in the same proportion, density is kept constant. On the other hand, if the unit cost declines as the airline increases output in a fixed network by adding more traffic on each route, we say there exists economies of density. Therefore, in air transport, the cost function must include network size (N) as a variable, to capture the cost impact of network expansion. Then, we can test the presence of economies of scale and economies of density by calculating the following EOS and EOD respectively

$$EOS = \frac{1}{\varepsilon_Q + \varepsilon_N}, \quad EOD = \frac{1}{\varepsilon_Q}$$

where ε_N is the elasticity of cost with respect to number of airports served. When EOS (EOD) is larger than 1, economies of scale (density) exists. Intuitively, airlines are likely to enjoy economies of density, because if it is possible to generate more traffic on a particular route, larger aircraft can be used and facilities at endpoint airports can be shared by more freight and/or flights. In studies of passenger airlines, economies of density has been widely observed, but economies of scale seems to vanish once the airline reaches a certain size.

In air freight operation, **Table 8** suggests that most of the studies also find stronger density effect than scale effect. Moreover, the conclusion about the existence of economies of scale for cargo airlines might be affected by the choice of functional forms. As shown in **Table 8**, when log-transformed Cobb-Douglas function is applied, diseconomies of scale is observed. However, when translog function is used, the estimated EOS all becomes larger than 1. Another observation is that integrators (FedEx and UPS) have larger EOS and EOD than non-integrators listed in **Table 8**. This may be caused by the different network structures. Integrators rely on hub-and-spoke network (**Fig. 1**) to channel freight via their hubs. As shown in **Table 2**, this network structure may cause integrators to incur a larger share of overhead costs, which is relatively fixed. However, it allows consolidating freight from many different origin-destination markets into one flight and therefore achieving higher density on each route. Most of the non-integrator cargo airlines do not operate extensive hub-and-spoke networks. Instead, their flights are routed in a circular way to visit a number of airports in sequence (**Fig. 1**). This circular routing substantially reduces the number of markets, which can be served by one single flight and makes network expansion more costly.

In maritime shipping, network size is rarely included in the studies. Thus, the definitions of economies of scale and economies of density are not as clear as in air transport. In fact, economies of density is never mentioned in the shipping literature. Instead, EOS is calculated by $1/\varepsilon_Q$ without controlling for network size. **Table 9** suggests that despite being above 1, EOS seems much weaker in shipping industry than in air freight transport. Bulk shipping and tanker seem to enjoy more economies of scale than container shipping. The negative elasticity of ship size (−0.24) means that with fixed output, total cost reduces as ship size increases. This is consistent to the scale economy in terms of ship size. However, its impact seems unclear if fleet capacity is controlled. In other words, empirically, it is unclear what causes the scale economy in shipping. Is it caused by larger ships, larger fleet capacity or capital stock or a combination of several factors? Moreover, it is controversial whether there is a ceiling of economies of ship size. First, the water depth and width of the canals and port channels limits the number of routes that large ships can operate. Large ships may have to take detours and travel longer distances than small ones. Second, the more cargoes a ship carries, the more time it spends for loading

and unloading at port (Cullinane and Khanna, 2000). Third, cost may increase at terminal or inland transport with large ships. For example, cranes at terminals need to be upgraded with longer outreach and connection to inland transportation has to be improved to handle high traffic brought by one single ship (Mason, 2015). The above discussion is based on very limited number of studies, since this issue has not been widely studied in shipping with comparable methods and datasets.

Conclusion

In general line-haul activities account for a much larger share of total operating costs for air freight than sea shipping. Cargo handling at terminals, management of sea containers and inland transportation also account for a lion's share of total cost of moving sea cargo. Despite this cost structure difference, in both transportation modes, there exist economies of vehicle size and, in general, unit costs reduce in load factor and travel distance. Each mode also has its special features. Speed places a significant role in ships' fuel costs, but it has limited impact on air transport. As air freight can be carried either in the freighter aircraft or in the belly hold of passenger airplanes, it further complicates the comparison of air freight cost efficiency. Network structure can affect the cost of container shipping, but to our knowledge this kind of studies are rare in the context of air freight. In fact, combination airlines, which own both passenger airplanes and freighters, might operate both circular and hub-and-spoke networks for cargo, making a direct cost comparison more difficult.

At company-level, both transportation modes apply Cobb-Douglas function or the more flexible translog form to approximate the total operating costs as a function of input prices, output and various other factors. In air transport, the estimated functions are used to measure the existence and degree of economies of scale and density after controlling for network size. However, in sea shipping, network size is never included as a variable and there is no separately defined concept of economies of density. Instead, the choice of variables seems to be inconsistent across studies and so do the results. Overall, empirically it is unclear whether it is the ship size or the company size that leads to lower unit cost in sea shipping.

References

- Air Transport Association of America, 1967. Standard Method of Estimating Comparative Direct Operating Costs of Turbine Powered Transport Airplanes, The Association, Washington, D.C.
- Ali, R., Al-Shamma, O., 2014. A comparative study of cost estimation models used for preliminary aircraft design. *Glob. J. Res. Eng.* 14 (4–B), 9–18.
- Balliauw, M., Meersman, H., Van de Voorde, E., 2018. US all-cargo carriers' cost structure and efficiency: a stochastic frontier analysis. *Transp. Res. Part A: Policy and Practice* 112, 29–45.
- Cullinane, K., Khanna, M., 2000. Economies of scale in large containerships: optimal size and geographical implications. *J. Transp. Geogr.* 8 (3), 181–195.
- Imai, A., Shintani, K., Papadimitriou, S., 2009. Multi-port vs hub-and-spoke port calls by containerships. *Transp. Res. Part E: Logist. Transp. Rev.* 45 (5), 740–757.
- Kiesling, M.K., Hansen, M., 1993. Integrated air freight cost structure: the case of federal express. Working Paper UCTC No. 400, University of California Transportation Center, UC Berkeley. Available from: <https://escholarship.org/uc/item/7338517g>.
- Lakew, P.A., 2014. Economies of traffic density and scale in the integrated air cargo industry: the cost structures of FedEx Express and UPS Airlines. *J. Air Trans. Manage.* 35, 29–38.
- Mason, T., 2015. *Liner Trades*, 2015th ed. Institute of Chartered Shipbrokers, London.
- Morrel, P.S., 2011. *Moving Boxes by Air: The Economics of International Air Cargo*, Routledge, Abingdon, UK.
- Onghena, E., 2011. Integrators in a changing world. In: Macário, R., Van de Voorde, E. (Eds.). *Critical Issues in Air Transport Economics and Business*. Routledge (Chapter 7).
- Onghena, E., Meersman, H., Van de Voorde, E., 2014. A translog cost function of the integrated air freight business: the case of FedEx and UPS. *Transp. Res. Part A: Policy and Practice* 62, 81–97.
- Roberts, C.M., 2014. *Efficiency in the US Airline Industry*. Doctoral Dissertation, University of Leeds, UK.
- Stopford, M., 2009. *Maritime Economics*, third ed. Routledge, Abingdon, UK.
- Swan, W.M., Adler, N., 2006. Aircraft trip cost parameters: a function of stage length and seat capacity. *Transp. Res. Part E: Logist. Transp. Rev.* 42, 105–115.
- Tolofari, S.R., Button, K.J., Pittfield, D.E., 1987. A translog cost model of the bulk shipping industry. *Transp. Plan. Technol.* 11 (4), 311–321.
- Tran, N.K., Haasis, H.D., 2015. An empirical study of fleet expansion and growth of ship size in container liner shipping. *Int. J. Prod. Econ.* 159, 241–253.
- Wu, W.M., Lin, J.R., 2015. Productivity growth, scale economies, ship size economies, and technical progress for the container shipping industry in Taiwan. *Transp. Res. Part E: Logist. Transp. Rev.* 73, 1–16.

Transport Production and Cost Structure

Ricardo Giesen, Darío Farren, Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, Santiago, Chile

© 2021 Elsevier Ltd. All rights reserved.

Introduction	46
Transport Production and Cost Structure	46
Single Origin–Destination System	47
Three-Node System	49
Distribution From a Single Origin to Multiple Destinations	50
Economies of Scale	51
Economies of Scope	51
Road Freight Transport	51
Rail Freight Transport	53
External Costs	53
Road Transport Externalities	54
Rail Transport Externalities	54
Acknowledgments	55
References	55

Introduction

The two basic modes of land freight transport other than pipelines are road and rail, each of which has its own cost structures and external effects. In cost analyses of either mode, the relevant cost concept will depend in any given case on what it is desired to analyze. For measuring expansion of the supply of transport services, incremental cost is to be used, while for analyzing service reduction the applicable notion is avoidable cost. In short-term decisions, fixed cost plays no role; the focus is rather on variable cost. In practice, however, this latter factor is ambiguous given that what is variable in one sense may not be in another. The difficulties involved in identifying the variable costs of a particular freight movement have to do with the multiproduct nature of transport firms and indivisibilities in the production of their services.

The remainder of this paper is organized into five sections. Section “Transport Production and Cost Structure” presents a microeconomic analysis of freight transport production and cost structure, distinguishing between single-origin/single-destination systems with three different points and single-origin/multiple-destination systems. Sections “Road Freight Transport” and “Rail Freight Transport” describe certain specific aspects of road and rail transport. Finally, Section “External Costs” discusses cost externalities characteristic of the two modes.

Transport Production and Cost Structure

Analyses of production processes are based on concepts of inputs, outputs, and technological viability. This section addresses the application of these concepts to the case of transport in which a firm produces load flows between different origins and destinations over multiple time periods. Also addressed in this section are the economies of scale, scope, and density.

The production of freight transport involves the assignment of resources to the generation of trips between different points in space over multiple periods. Measuring a transport process requires a description of the load to be carried, a physical unit of reference, the load quantity (flow), and the origin and destination in space–time (Jara-Díaz, 2007). This last characteristic is what distinguishes a transport product from the traditional product concept. A transport firm uses vehicles, loading/unloading terminals, rights-of-way, energy, labor, and other resources to produce movements of goods from multiple origins to multiple destinations in different periods. The firm’s output is thus a flow quantity between an origin and a destination at a given instant in time, which can be expressed in formal terms by a vector as follows:

$$Y = \{y_{ij}^{kt}\} \in R^{K \times N \times T} \quad [1]$$

where each y_{ij}^{kt} component represents a product flow of type k moved from an origin i to a destination j in period t . Expression [1] is quite general and covers the possibility of handling various types of flow. It should be noted that transport firms produce multiple outputs not just because they may deal in various flow types but also, and primarily, because of the time and space dimensions (i.e., periods and origin–destination pairs). Unlike classical microeconomic theory, these companies are active simultaneously in various markets each having its own demand curve and marginal costs, although the latter tend to be interrelated. The space dimension, much more than time, is the key element that differentiates the transport industry from other economic activities.

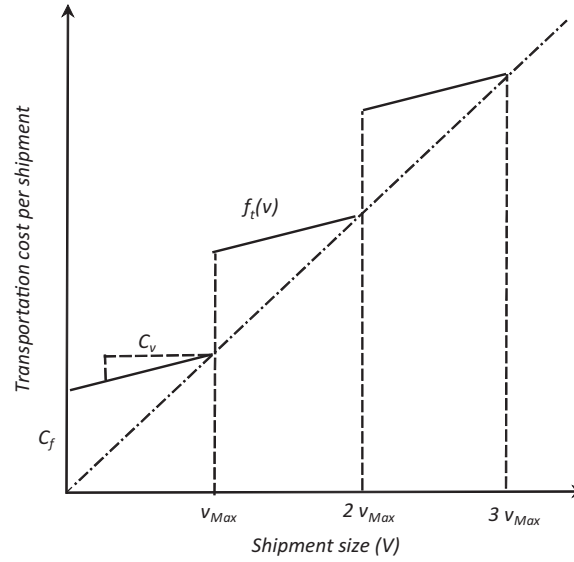


Figure 1 Relationship between transportation cost per shipment and shipment size. Source: Daganzo (2005).

For any given level of production, a transport firm must make decisions regarding the quantity and characteristics of the inputs such as the number of vehicles and the number of terminals, the latter including their loading and unloading capacities. The firm must also set the operating rules such as vehicle speeds, frequencies, and load sizes. Since transport takes place over a network the company will also have to plan a service structure, that is, the generic way in which its vehicles will visit the various destinations to produce the desired flows. This set of endogenous decisions defines the route structure, which must be chosen based on spatial information regarding demand, origin and destination locations, and the physical network.

Single Origin–Destination System

In a system consisting of a single origin producing identical products to be consumed at a single destination (Daganzo, 2005), transport costs per shipment can be broken down into a fixed cost (C_f) and a variable cost (C_v). The total mean transport cost in a given period for a sequence v_i of shipments $V = \sum_i v_i$ can be expressed as:

$$CMe = \frac{C}{V} = C_f \left(\frac{1}{V} \right) + C_v \quad [2]$$

Economies of scale arise from the fact that in Eq. [2] all the individual shipments in the sequence share the fixed cost C_f . Cost per shipment also depends on shipment size. Each time this latter factor reaches and exceeds a multiple of vehicle capacity K , an additional vehicle must be dispatched, resulting in a jump in cost per shipment as is illustrated in Fig. 1.

The foregoing implies that the per shipment cost function $f_t(v)$ must be subadditive, that is, it must satisfy the inequality $f_t(x_1 + x_2) \leq f_t(x_1) + f_t(x_2)$ for any $x_1, x_2 \geq 0$. This property is to be expected since it would be counterintuitive for the dispatching of shipments separately to reduce total transport costs.

In the same context of two nodes (1 and 2) we now assume a multi-output firm is operating a backhaul system with two flows (γ_{12}, γ_{21}) and one product in a single period (Jara-Díaz and Basso, 2003), using the same fleet of vehicles to move the two flows. Vehicle frequency (f) in both directions is given in this case by the maximum flow. If we further assume that $\gamma_{12} \geq \gamma_{21}$, the vehicle traveling from 1 to 2 will be fully loaded and the frequency will be given by:

$$f = \frac{\gamma_{12}}{K} \quad [3]$$

If the vehicles are loaded and unloaded sequentially at a rate μ , vehicle speed is v and the distances traveled between the two nodes in either direction are d_{12} and d_{21} , cycle time between 1 and 2 will then be:

$$t_c = \frac{d_{12}}{v} + \frac{2K}{\mu} + \frac{2d_{21}}{\mu} + \frac{d_{21}}{v} \quad [4]$$

Fleet size B is given by f times t_c :

$$BK = \gamma_{12} \left[\frac{d_{12}}{v} + \frac{2K}{\mu} + \frac{d_{21}}{v} \right] + \frac{2K}{\mu} \gamma_{21} \quad [5]$$

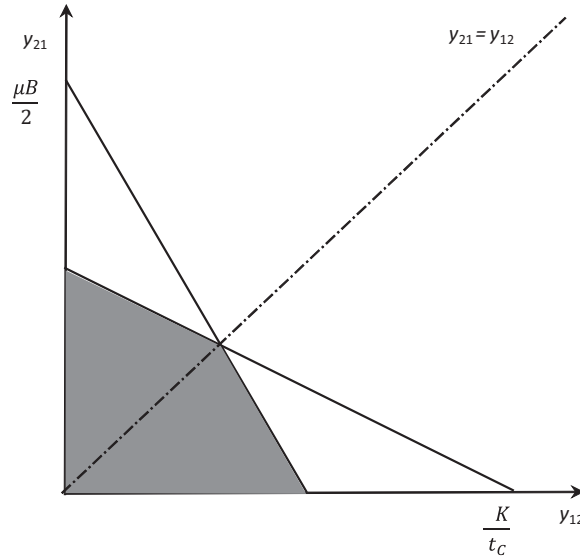


Figure 2 Production possibility frontier of a backhaul freight transport system. Source: Jara-Díaz (2007).

The production possibility frontier can be expressed as:

$$y_{21} = \frac{\mu B}{2} - \left[\left(\frac{d_{12} + d_{21}}{v} \right) \frac{\mu}{2K} + 1 \right] y_{12} \quad [6]$$

Eq. [6] gives the set of vectors (y_{12}, y_{21}) representing the output that can be efficiently produced by a fleet of vehicles B each of capacity K circulating at speed v and loading and unloading at a rate of μ . Within this frontier lie all the (y_{12}, y_{21}) combinations that are technically feasible, as shown in Fig. 2.

Total production expenses are obtained from the parameters associated with input prices, which include vehicle fuel consumption per kilometer (g), hours worked by vehicle drivers (ε) and loading/unloading site personnel (θ), wage rates (ω), fuel price (P_g), and price per hour of a capacity K vehicle (P_K) and of a loading/unloading site P_μ . Based on the foregoing, the formula for total production expenses G is given by:

$$G(y_{12}, y_{21}) = C_0 + [P_K + \omega\varepsilon]B + \frac{P_g \cdot g \cdot (d_{12} + d_{21})}{K} \cdot y_{12} + \frac{2 \cdot P_\mu + \omega\theta}{\mu} y_{21} \quad [7]$$

Thus, the optimization problem facing the firm would involve minimizing G with respect to K , μ , and v . Assuming the values of these variables are fixed, that is, the vehicles and loading/unloading sites are only available in a single size and speed is determined exogenously by technical or legal considerations, the cost function can then be formulated as follows:

$$C(y_{12}, y_{21}) = C_0 + y_{12} \cdot (d_{12} + d_{21}) \cdot \Lambda + (y_{12} + y_{21}) \cdot \Omega \quad [8]$$

where

$$\Lambda \equiv \left[\frac{P_K + \omega\varepsilon}{vK} + \frac{P_g g}{K} \right] \quad [9]$$

$$\Omega \equiv \left[\frac{P_K + \omega\varepsilon}{vK} + \frac{P_g g}{K} \right] \quad [10]$$

Note that in Eq. [8] there is a term for flow distance and another for pure flow. This latter term captures expenses incurred while the product is stationary, that is, costs arising from terminal operations, whereas the former term captures route expenses. The presence of these terms draws attention to two aspects of the description of transport firm output. First, if the cost function specification includes only factors relating to distance such as ton-kilometers and average trip length, the real transport production costs may not be properly captured. Second, since flow distance equals frequency times vehicle capacity, it is the term that determines the capacity of the transport system. This means that if the firm's output is a freight flow, for example, the term will be the total number of vehicle-kilometers it "produced." This justifies the use of such aggregates in the literature on transport cost functions. Once a firm has optimized its operations such that flows are produced at minimum cost, route expenses will be directly related to total transport capacity.

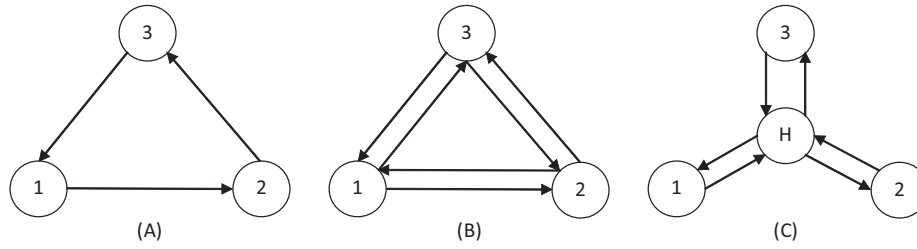


Figure 3 Service structures for a three-node network. Source: Jara-Díaz (2007).

Note, however, that the two-node system just outlined cannot give a complete demonstration of the two stages in the optimization process. The second stage, which compares cost functions that are conditional on route structure, is not applicable in this single route structure case. Nevertheless, it will be very useful for explaining the three-node system in which there is a choice of route structures and for comparing costs following an expansion of the network.

Three-Node System

Transport firms are faced with three decisions: the quantity and characteristics of the inputs, the operating rules, and the route structure. Since the third decision is a discrete one, the underlying minimization process may be seen as a sequence of two stages. In the first stage, for each possible route structure in the system the firm optimizes inputs and operating rules. A production possibility frontier is established and, on the basis of input prices, expenses are minimized to obtain a cost function conditional upon the structure that gives the minimum cost for producing a given level of output. Then, in the second stage, these conditional cost functions derived for each route structure are compared and the structure with the lowest cost is chosen.

Now consider an OD structure of six flows in a system of three nodes connected by three links of length d_{ij} . In this basic physical network there is no decision to be made on link sequence so the choice of a service structure and the choice of a route structure are one and the same, a convenient simplification. Maintaining the assumptions of our two-node model in the previous subsection, that is, the same sequential loading/unloading procedure and known values of K , μ , and ν , the objective is to find cost functions that are conditional upon the new route structure.

Three possible service structures for a three-node system with six OD pairs are illustrated in Fig. 3. Structure (A) is a simple circuit, structure (B) consists of three simple cyclical systems, and structure (C) is a hub-and-spoke structure that creates a distribution node (H in the figure), a common configuration in air transport. As regards the assignment of vehicles to fleets, in (A) only one fleet (one frequency) is possible, in (B) there are three fleets, and in (C) there can be one, two (with three alternatives), or three fleets.

In a general counterclockwise cyclical structure (Fig. 3A), which implies the use of a single fleet, the vehicle load size on each of the three segments k_{12} , k_{23} , and k_{31} in the network is defined as:

$$k_{12} = \frac{\gamma_{12} + \gamma_{13} + \gamma_{32}}{f} \quad [11]$$

$$k_{23} = \frac{\gamma_{23} + \gamma_{21} + \gamma_{13}}{f} \quad [12]$$

$$k_{31} = \frac{\gamma_{31} + \gamma_{32} + \gamma_{21}}{f} \quad [13]$$

If we assume arbitrarily that the segment with the highest load is 1–2, then the efficiency condition implies that the vehicle on this segment must carry a full load. Its frequency will thus be given by:

$$f = \frac{\gamma_{12} + \gamma_{13} + \gamma_{32}}{K} \quad [14]$$

Therefore, cycle time is expressed as:

$$t_c = \frac{d_{12} + d_{23} + d_{31}}{\nu} + \frac{2K}{\mu} + \frac{2K(\gamma_{21} + \gamma_{23} + \gamma_{31})}{\mu(\gamma_{12} + \gamma_{13} + \gamma_{32})} \quad [15]$$

The production possibility frontier for this route structure is obtained from the following formula:

$$BK = (\gamma_{12} + \gamma_{13} + \gamma_{32}) \left[\frac{d_{12} + d_{23} + d_{31}}{\nu} + \frac{2K}{\mu} \right] + \frac{2K}{\mu} (\gamma_{21} + \gamma_{23} + \gamma_{31}) \quad [16]$$

This in turn means that the conditional cost function of a cyclic route structure is:

$$C_{CG}(Y) = C_0 + (\gamma_{12} + \gamma_{13} + \gamma_{32}) \cdot (d_{12} + d_{23} + d_{31}) \cdot \Lambda + (\gamma_{12} + \gamma_{13} + \gamma_{32} + \gamma_{21} + \gamma_{31} + \gamma_{23}) \cdot \Omega \quad [17]$$

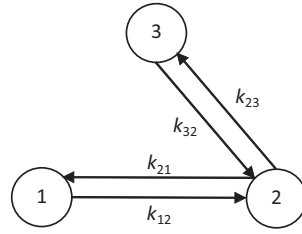


Figure 4 Hub-and-spoke route structure. Source: Jara-Díaz (2007).

This function is similar to the one obtained for the two-node system. The flow distance term obtained for the present system has the same meaning, that is, the system's capacity, given that the flows involved define the frequency. The pure flow term is generated by the loading/unloading activities. Note also that this function reduces to the two-node version if the four new flows are set to zero and $d_{23} - d_{31}$ is defined as d_{21} .

In a hub-and-spoke route structure, the hub is a node that collects and distributes all flows and is usually either the origin or the destination. Assume arbitrarily that the hub is node 2 and a fleet of vehicles operates at a single frequency. Other structures of this type could, of course, be considered, such as a two-fleet operation, one in 1–2 and the other in 2–3. Here, however, we have opted to develop the one-fleet operation since the others can be constructed using the two-node system, as shown later. Thus, a vehicle loads flows γ_{12} and γ_{13} at node 1, unloads γ_{12} and loads γ_{23} at node 2, then unloads γ_{13} and γ_{23} and loads γ_{32} and γ_{31} at node 3, returns to node 2 to unload γ_{32} and load γ_{21} , and finally goes back to node 1 to unload and start the cycle over again (Fig. 4).

Following the same procedure used in the previous case, we obtain:

$$C_{HS}(Y) = C_o + (\gamma_{12} + \gamma_{13}) \cdot (d_{12} + d_{23} + d_{32} + d_{21}) \cdot \Lambda + (\gamma_{12} + \gamma_{13} + \gamma_{32} + \gamma_{21} + \gamma_{31} + \gamma_{23}) \cdot \Omega \quad [18]$$

In this equation there again appear a flow distance term that represents system capacity and a pure flow term generated by the loading/unloading activities. We have already obtained two conditional cost functions for the three-node system, but by simple analogy we can find conditional functions for another three route structures: a clockwise cyclic system, and hub-and-spoke systems with the concentrator at nodes 1 or 3. The two-node system cost function can also be used to derive conditional cost functions for other cases: direct service with three fleets, each one serving a pair of nodes cyclically (1–2, 2–3, and 1–3) and hub-and-spoke with two fleets, each one connected by a pair of nodes with the concentrator at either of the three nodes. In this latter case, some flows will have to be loaded and unloaded twice (origin, destination, and concentrator). This increases expenses above the levels of the other cases but reduces cycle times. Finally, these examples illustrate how choosing a route structure is a key endogenous element and also show that minimum cost is associated with this choice.

Distribution From a Single Origin to Multiple Destinations

In this subsection, we address the problem of the physical distribution of goods produced at a single origin (the “depot”) to N customers spread over a region of area S with no transshipments. Assume that the N points to be visited are located uniformly and randomly over the indicated area. The expected trip distance is then:

$$L(N, S) \approx k\sqrt{NS} \quad [19]$$

The mean distance per customer is therefore:

$$\frac{k}{\sqrt{N/S}} = \frac{k}{\sqrt{\delta}} \quad [20]$$

where δ is the customer density (customers/m²) in the system. This approximation is better for large values of N . The k term is a nondimensional constant that depends on the metric used to measure distance traveled over the network. If the metric is Euclidean, $k = 0.57$ and if it is Manhattan, $k = 0.82$.

If we consider a region x_o with C stops, the total distance traveled to visit C points within the region (the tour distance) is:

$$\text{Tour distance} \approx 2\bar{r} + \left[k\delta^{-1/2}(x_o) \right] (C - 1) \quad [21]$$

where $\bar{r}$ is the average distance from the C points in region x_o to the depot along the shortest route.

The first term in Eq. [21] may be interpreted as the mean distance from the depot to the center of gravity of the points in the region, while the second term is the local distance a vehicle must travel to deliver and pick up a load. Thus, if there are N/C tours, the total distance traveled is approximated by:

$$\text{Total distance} \approx \left\{ \frac{2E[r]}{C} + kE(\delta^{-1/2}) \right\} N \quad [22]$$

If the density is uniform, that is, $E(\delta^{-1/2}) = \delta^{-1/2} = \sqrt{S/N}$, then:

$$\text{Total distance traveled} \approx \frac{2E[r]}{C}N + k\sqrt{SN} \quad [23]$$

Regardless of the points' specific locations, if cost must be estimated before they are known, Eqs. [22] and [23] will be very useful. Comparisons made by Hall et al. (1994) indicate that these two approximation formulae are quite accurate even if the number of stops is not the same for every tour.

Economies of Scale

The concept of economies of scale refers to the amount of increase in output brought about by an equiproportional increase in the amounts of all inputs. The degree of multiproduct economies of scale S can be defined as the maximum expansion of output Y , denoted $\lambda^S Y$, obtainable by expanding the input vector X to λX . In analytic terms,

$$F(\lambda X, \lambda^S Y) = 0 \quad [24]$$

where F is the production function relating outputs to inputs. The value of S may be greater than, equal to, or less than 1, indicating that returns to scale are, respectively, increasing, constant, or decreasing. It is calculated as follows:

$$S = \frac{C(Y)}{\sum_i Y_i \frac{\partial C}{\partial Y_i}} = \frac{1}{\sum_i \eta_i} \quad [25]$$

and

$$\eta_i = \frac{Y_i}{C(w, Y)} \frac{\partial C(w, Y)}{\partial Y_i} \quad [26]$$

where η_i is the cost elasticity of output i , C is the cost of production, and w is the input price vector.

Economies of Scope

The concept of economies of scope is used to analyze the economics of joint production of multiple products. The degree of economies of scope SC_R for a subset R is defined as:

$$SC_R = \frac{1}{C(Y)} [C(Y_R + C(Y_{M-R})) - C(Y)] \quad [27]$$

where Y_R represents the vector Y , M is the set of all products, and $y_i = 0, \forall i \notin R \subset M$. Then, if SC_R is positive, it is cheaper for Y to be produced by a single firm than by two different firms where one produces subset R and the other subset $M-R$.

In the sections that follow we analyze the cost structures for road and rail freight transport, including the external costs attendant upon each mode.

Road Freight Transport

Road transport is today the dominant mode for moving a wide variety of load types. The key advantage of truck haulage is that it allows loads to be picked up and delivered at locations not accessible by other modes (Webster, 2009). This section discusses road transport cost structure and vehicle capacity utilization.

Broadly speaking, the trucking industry is segmented into truckload (TL) carriers and less-than-truckload (LTL) carriers. TL shipments generally move from a single origin to a single destination on fully loaded vehicles, whereas LTL shipments occupy a certain percentage of a truck's capacity, allowing operators to consolidate multiple shipments in a single vehicle. Compared with TL, LTL shipments are more costly per ton and have longer shipping times, and the goods carried are more exposed to damage given that they are routed through various consolidation centers where loads are interchanged. In the United States, TL shipments make up 52% of the total and LTL shipments about 24% (Hooper and Murray, 2017).

A truck can cover approximately 800 km/day, of which empty returns account for about 20% in the United States and 20%–35% in the EU. Operating costs are subject to a series of underlying impacts and externalities. As a result, some expenses like fuel and tires can be measured relatively easily, but calculating labor costs may be complicated by the effects of driver experience, productivity, and differences in compensation models.

The marginal costs of road transport can be divided into two general categories. The first category is related to the trucks themselves and includes expenses such as fuel, lease or purchase payments, repairs and maintenance, vehicle insurance, and permits and special licenses. The second category covers driver expenses such as wages and fringe benefits.

Data from the US-based American Transportation Research Institute (ATRI) show that the marginal cost per mile in the United States is about US\$1.60 (Table 1) while marginal cost per hour, at an average speed of about 40 Mph, is on the order of US\$65 (Table 2).

Table 1 Summary of marginal costs per mile, 2008–16

<i>Motor carrier costs</i>	<i>2008</i>	<i>2009</i>	<i>2010</i>	<i>2011</i>	<i>2012</i>	<i>2013</i>	<i>2014</i>	<i>2015</i>	<i>2016</i>
Vehicle-based									
Fuel costs	\$0.633	\$0.405	\$0.486	\$0.590	\$0.641	\$0.645	\$0.583	\$0.403	\$0.336
Truck/trailer lease or purchase payments	\$0.213	\$0.257	\$0.184	\$0.189	\$0.174	\$0.163	\$0.215	\$0.230	\$0.255
Repair and maintenance	\$0.103	\$0.123	\$0.124	\$0.152	\$0.138	\$0.148	\$0.158	\$0.156	\$0.166
Truck insurance premiums	\$0.055	\$0.054	\$0.059	\$0.067	\$0.063	\$0.064	\$0.071	\$0.074	\$0.075
Permits and licenses	\$0.016	\$0.029	\$0.040	\$0.038	\$0.022	\$0.026	\$0.019	\$0.019	\$0.022
Tires	\$0.030	\$0.029	\$0.035	\$0.042	\$0.044	\$0.041	\$0.044	\$0.043	\$0.035
Tolls	\$0.024	\$0.024	\$0.012	\$0.017	\$0.019	\$0.019	\$0.023	\$0.020	\$0.024
Driver-based									
Driver wages	\$0.435	\$0.403	\$0.446	\$0.460	\$0.417	\$0.440	\$0.462	\$0.499	\$0.523
Driver benefits	\$0.144	\$0.128	\$0.162	\$0.151	\$0.116	\$0.129	\$0.129	\$0.131	\$0.155
Total	\$1.653	\$1.451	\$1.548	\$1.706	\$1.633	\$1.676	\$1.703	\$1.575	\$1.592

Source: Hooper and Murray (2017).

Table 2 Summary of marginal costs per hour, 2008–16

<i>Motor carrier costs</i>	<i>2008</i>	<i>2009</i>	<i>2010</i>	<i>2011</i>	<i>2012</i>	<i>2013</i>	<i>2014</i>	<i>2015</i>	<i>2016</i>
Vehicle-based									
Fuel costs	\$25.30	\$16.17	\$19.41	\$23.58	\$25.63	\$25.78	\$23.29	\$16.13	\$13.45
Truck/trailer lease or purchase payments	\$8.52	\$10.28	\$7.37	\$7.55	\$6.94	\$6.52	\$8.59	\$9.20	\$10.20
Repair and maintenance	\$4.11	\$4.90	\$4.97	\$6.07	\$5.52	\$5.92	\$6.31	\$6.23	\$6.65
Truck insurance premiums	\$2.22	\$2.15	\$2.35	\$2.67	\$2.51	\$2.57	\$2.89	\$2.98	\$3.00
Permits and licenses	\$0.62	\$1.15	\$1.60	\$1.53	\$0.88	\$1.04	\$0.76	\$0.78	\$0.88
Tires	\$1.20	\$1.14	\$1.42	\$1.67	\$1.76	\$1.65	\$1.76	\$1.72	\$1.41
Tolls	\$0.95	\$0.98	\$0.49	\$0.69	\$0.74	\$0.77	\$0.90	\$0.79	\$0.97
Driver-based									
Driver wages	\$17.38	\$16.12	\$17.83	\$18.39	\$16.67	\$17.60	\$18.46	\$19.95	\$20.91
Driver benefits	\$5.77	\$5.11	\$6.47	\$6.05	\$4.64	\$5.16	\$5.15	\$5.22	\$6.18
Total	\$66.07	\$58.00	\$61.90	\$68.21	\$65.29	\$67.00	\$68.09	\$62.98	\$63.66

Source: Hooper and Murray (2017).

A key factor impacting directly on operating costs is vehicle capacity utilization given that it is an indicator of how economic resources are being employed from the standpoint of both the operator and the customer (Ben-Akiva et al., 2013). Despite its positive contributions, road transport produces negative externalities that must be mitigated, though ideally not at the cost of economic prosperity. Improving vehicle capacity utilization would make it possible to lower the vehicle mileage required to satisfy freight transport demand and thereby also reduce the related external effects. This points up the importance of strengthening our understanding of the factors behind vehicle capacity utilization and how it influences global demand for freight transport trips.

Unlike the case of passengers, freight rarely returns to its point of origin, so freight vehicle use on return trips is less efficient. This is known as the backhaul problem. Raising the efficiency of available vehicle capacity on trips in the two directions thus depends on the extent to which the trucks travel loaded. The cost complementarities between the two legs of a round trip impel transport firms to serve multiple markets in order to minimize costs. EU statistics show that between 15% and 38% of total truck travel consisted of empty returns; estimates for developing countries range between 30% and 35%.

Studies of this topic may be divided into two groups according to their analytical approach and field of origin. The first group looks at utilization from the standpoint of economic theory, which assumes the objective of a firm is to maximize profit and examines how utilization is influenced by various characteristics of the firm and the transport market. The second group, meanwhile, proceeds from the viewpoint of the transport literature and analyzes vehicle movement and utilization in the context of transport demand models.

In studies based on economic theory, vehicle capacity is underutilized as a result of the constant challenge to equate capacity with demand due to the imbalances in the movement of goods between regions and differences between truck operators in market access costs. Freight imbalance is an external (exogenous) problem that operators can minimize, at least in the long run, only by choosing an appropriate location for their base of operations close to the main generators of traffic. In practice, however, operators have to make continuous market access decisions as part of the process of adapting specific demand levels to specific levels of capacity based on net income considerations. As a result, to the extent there exist differences in access costs not related to distance, in any given market segment there will be some vehicles carrying loads while others run empty. Recent studies in this literature highlight the

abilities of information technologies to equate capacity with demand, allowing operators to reduce market access costs and thus maintain their vehicles loaded and on the road at higher frequencies.

The transport literature, on the other hand, focuses on the relationship between the “trip chain” and the vehicle routing problem faced by operators in the context of urban freight transport, where utilization levels are lower than those experienced in long-distance operations. The trip chain approach to freight movement analysis can significantly improve modeling of demand for freight transport.

Rail Freight Transport

Rail transport is a relatively low-cost mode that is attractive for long-distance shipping and heavy loads in situations where dead time and reliability are relatively unimportant. Goods typically shipped by rail include coal, minerals, grains, and wood, as well as motor vehicles and heavy machinery (Webster, 2009).

Variable costs in rail transport are generally lower than those in truck transport, but fixed costs, consisting notably of rolling stock and stations, are higher. Railroads must also shoulder the expense of rail network maintenance. This contrasts with the case of road networks, which are often financed and maintained by governments or through tolls that trucking firms incur as a variable cost.

As with road haulage, rail transport typically offers multiple services involving multiple costs, complicating the task of identifying the cost of moving specific loads (Button and Pitfield, 1985). An example of this is the joint cost of fronthaul and backhaul. Although the supply of transport takes the form of round trips, demand for transport is one way. Many of the variable costs involved are incurred jointly, that is, in both directions, meaning the individual fronthaul and backhaul costs are inseparable. It is conceptually impossible to identify the total cost of a movement separate from the costs of other traffic. It is true that certain rail transport cost items can be uniquely attributable to specific traffic, but most of them cannot. In principle, the cost interdependence of different products requires that the optimal price be set simultaneously for all of them.

Another relevant characteristic of railroads is that their fixed capital is intensive and varied, and increments to it are necessarily discrete and indivisible. An example of such an increment is the construction of a second track to handle high-density traffic operations. The implication is that the cost to the railroad of an increase in traffic is low if there is excess network capacity, but very high if there is not. Thus, for small product increments costs are relatively invariable but for large ones they can vary greatly.

Estimating the private cost of rail freight service is inherently more complex than for similar service in road transport. Among the complicating factors are joint production between railroad companies (e.g., the interchange of track and rolling stock between operators), scale and density economies, and the lack of data on specific expenditures relating to individual movements (Forkenbrock, 2001). Since rail freight operations are highly varied, a single added value for the private cost per ton-mile would have little meaning. Thus, in what follows we present the costs for four different railroad freight scenarios.

1. Heavy unit train: It consists of 100 light (26-ton) cars, each one carrying a load of 105 tons. The trip is 1000 miles long and returns 100% empty. The train is powered by four 3000-BHK locomotives.
2. Mixed freight train: The mixed load is transported in 90 cars each weighing an average of 32 tons. The average carload is 70 tons and trip length is 500 miles, with an empty return rate of 45%. The train is powered by three 3000-BHK locomotives.
3. Intermodal train: It is made up of 120 spine cars carrying an equal number of truck semitrailers. A spine car weighs 14 tons while the average weight of a loaded semitrailer is 28 tons. Trip length is 1750 miles, with an empty return rate assumed to be 5%. The train is powered by three 3000-BHK locomotives.
4. Double-stack train: It consists of 24 light flatcars weighing an average of 80 tons each and carrying 240 containers. The empty return rate is 10%. The train is powered by four 3000-BHK locomotives.

Operating costs are estimated for the four scenarios using average operating parameters and are set out in Table 3. Both the heavy unit and mixed freight scenarios have a cost per ton-mile of approximately 1.2 cents. For the intermodal scenario, the corresponding cost figure is 2.68 cents while for the double-stack scenario it is 1.06 cents.

External Costs

In environmental impact studies of freight transport systems, it is considered a best practice to assign the impact of the system to the production of the load carried. This ensures that the consequences of the entire chain of production, including, for example, the

Table 3 Private operating costs of four railroad freight scenarios

Rail road scenario	Power	Cargo (tons)	Distance (miles)	Average cost per ton-mile (1994 cents)
Heavy unit train	Four 3000-BHP locomotives	10.500	1000	1.19
Mixed freight train	Three 3000-BHP locomotives	6.300	500	1.20
Intermodal train	Three 3000-BHP locomotives	3.360	1750	2.68
Double-stack train	Four 3000-BHP locomotives	6.720	1750	1.06

Source: Forkenbrock (2001).

Table 4 Fuel consumption for a truck with trailer per kilometer for different road types, gradients, and cargo loads

Road type	Road gradient	Cargo load factor weight	Fuel consumption L/10 km
Average road	±2%	50%	3.17
Rural road	±2%	50%	3.01
Urban road	±2%	50%	3.86
Average road	0	50%	2.71
Average road	±6%	50%	5.82
Average road	±2%	0.0%	2.27
Average road	±2%	100%	4.05

Source: Monios (2017).

Table 5 EU emission standards for trucks

Stage	Date	Co (g/kWh)	HC (g/kWh)	NO _x (g/kWh)	PM (g/kWh)
Euro I	1992	4.5	1.10	8.0	0.36
Euro II	1996	4.0	1.10	7.0	0.25
	1998	4.0	1.10	7.0	0.15
Euro III	2000	2.1	0.66	5.0	0.10
Euro IV	2005	1.5	0.46	3.5	0.02
Euro V	2008	1.5	0.46	2.0	0.02
Euro VI	2013	1.5	0.13	0.4	0.01

Source: Monios (2017).

emissions resulting from the production of the load, are taken into account along with the transport itself. The work done in transporting the load must therefore be calculated as the load mass multiplied by the transport distance (in TKM units). This allows different transport modes to be compared. Another common practice is to look at the impact in terms of 20-foot equivalent units (TEU) and the distance they are transported (TEU-km). These are sometimes given separately for loaded and empty containers (Monios, 2017).

Currently, the emissions that contribute most to global warming are greenhouse gases, the most important of which are carbon dioxide (CO₂), methane, and nitrous oxide (NO_x). CO₂ emissions are produced by burning fuels containing carbon such as gasoline, diesel fuels, natural gas, and biofuels.

Road Transport Externalities

Fuel consumption and emissions in road transport depend on vehicle speed and road topography. In urban zones or wherever traffic is congested, fuel consumption per kilometer may be significantly greater due to frequent vehicle stopping and accelerating. Another major influence on consumption is load mass. Some examples of fuel consumption are summarized in Table 4.

The main parameter determining road transport emissions is the load factor, which may vary significantly for both intermodal and unimodal systems. A high level of empty container repositioning would, of course, have a generally negative impact on environmental performance. Some examples of truck emission standards in the EU are summarized in Table 5.

Rail Transport Externalities

Since 2006, the EU has applied regulations governing diesel locomotive engine emissions of CO, HC, NO_x and PM (Table 6). However, permitted levels (measured in g/kWh) are significantly higher than those applicable to modern road vehicles. In Europe, locomotives generally run on diesel fuel of the same quality road vehicles use.

Table 6 EU emission regulations for diesel locomotive engines

Category	Stage	Net power (kW)	Date	Co (g/kWh)	HC (g/kWh)	HC + NO _x (g/kWh)	NO _x (g/kWh)	PM (g/kWh)
Rail car	IIIA	130 < P	2006	3.5		4.0		0.200
Locomotive	IIIA	130 < P < 560	2007	3.5		4.0		0.200
Locomotive	IIIA	P > 560	2009	3.5	0.50 ^a		6.0	
Rail car	IIIB	130 < P	2012	3.5	0.19		2.0	0.200
Locomotive	IIIB	130 < P	2012	3.5		4.0		0.025
Rail car	V	0 < P	2021 ^b	3.5	0.19		2.0	0.015
Locomotive	V	0 < P	2021 ^b	3.5		4.0		0.025

^aHC = 0.4 (g/kWh) and NO_x = 7.4 (g/kWh) for engines of P > 2000 kW and kW and D > 5 L/cylinder.

Table 7 Life-cycle emissions of CO_{2e} for different electricity generation technologies

<i>Electricity generation technology</i>	<i>CO₂ life-cycle emissions (g/kWh)</i>
Hydropower	4
Wind energy	12
Nuclear energy	16
Natural gas	469
Oil	840

Source: *Monios (2017)*.

Although electric motors do not produce emissions, it is common practice to consider the emissions resulting from the generation of electricity. The amounts produced can vary greatly depending on how the electricity is generated. Although the emission levels from hydro and wind power are very low, those from carbon-based sources are very high (in the case of CO₂, higher even than emissions from diesel engines). Some emissions data for different electricity generation technologies are given in **Table 7**.

Acknowledgments

We would like to thank the support by Fondecyt No. 1171049, CEDEUS, CONICYT/FONDAP 15110020, and the BRT+ Centre of Excellence funded by VREF. Also, Darío Farren thanks the support by CONICYT Doctoral Scholarship No. 21181464.

References

- Ben-Akiva, M., Meersman, H., Van de Voorde, E., 2013. Freight Transport Modelling. Emerald Group Publishing Limited, Bingley.
- Button, K.J., Pitfield, D., 1985. International Railway Economics: Studies in Management and Efficiency. Gower Publishing Company Limited, Aldershot.
- Daganzo, C.F., 2005. Logistics Systems Analysis, fourth ed. Springer-Verlag Berlin Heidelberg, Heidelberg.
- Forkenbrock, D.J., 2001. Comparison of external costs of rail and truck freight transportation. Transp. Res. Part A Policy Pract. 35 (4), 321–337, doi:10.1016/S0965-8564(99)00061-0.
- Hall, R.W., Du, Y., Lin, J., 1994. Use of continuous approximations within discrete algorithms for routing vehicles: experimental results and interpretation. Netw. Spat. Econ. 24, 43–56.
- Hooper, A., Murray, D., 2017. An analysis of the operational costs of trucking. TRB 2010 Annual Meeting, 51. Available from: <http://truckexec.typepad.com/files/atri-operational-costs-of-trucking-2013-final.pdf>.
- Jara-Díaz, S.R., 2007. Transport Economic Theory, 2007th ed. Emerald Group Publishing Limited, Bingley.
- Jara-Díaz, S.R., Basso, L.J., 2003. Transport cost functions, network expansion and economies of scope. Transp. Res. Part E Logistics Transp. Rev. 39, 271–288, doi:10.1016/S1366-5545(03)00002-4.
- Monios, J.B.R., 2017. Intermodal Freight Transport & Logistics, first ed., vol. 52. CRC Press Taylor & Francis Group, Boca Raton, FL.
- Webster, S., 2009. Principles of Supply Chain Management. Dynamic Ideas, Charlestown, MA.

The Concept of External Cost: Marginal versus Total Cost and Internalization

Sofia F. Franco*, Department of Economics, University of California-Irvine, Irvine, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	56
External Costs: Understanding the Concept and its Use	56
External Costs: Graphical Analysis	58
External Costs: Additional Remarks	59
External Costs and Travel Mode Choice	60
External Costs and Transport–Land Use Interactions	61
External Costs and Labor Market Interactions	62
Policies for Obtaining Social Optimality with External Costs	63
Internalization in Second-Best Settings	64
Conclusions	65
References	66

Introduction

When traveling to and within geographic areas, individuals use different travel modes like walking, cycling, or variants such as small-wheeled travel modes (skates, skateboards, or push scooters), public transit (bus, metro, train, or rail) or, alternatively, they may choose traveling by privately owned vehicles. However, less sustainable modal choices can generate unintended consequences ranging from traffic congestion, air pollution, climate change, noise, accidents, habitat damage, to name a few. These unintended external consequences, known as *externalities*, give rise to various costs such as time costs of delays due to traffic congestion, health costs, and ecosystem damages caused by air pollution, productivity losses due to fatalities and injuries in traffic accidents, among others (Calthrop and Proost, 1998; EC, 2019; Maibach et al., 2008; Van Essen et al., 2008; Verhoef, 1994). These *external costs* affect society at large because they fall on individuals who are not part of the decision resulting in those costs, but are not directly borne by the individual who has caused them. While the latter might be individuals who have themselves made similar decisions, unless they do so as part of a group, each transport user disregards such external effects when making travel mode choices. These external costs then create a wedge between the private costs faced by the decision maker, and the social costs incurred by society. Without policy intervention, transport users face wrong incentives, leading to inefficient outcomes and to welfare losses. Transport externalities are therefore an example of a *market failure* because in their presence, the market does not allocate resources on its own efficiently in a way that balances social costs and social benefits.

Economic theory shows that net social welfare is maximized when the marginal benefit of an activity (reflected by the price of the activity) is equal to its marginal social cost. Actual policies to achieve marginal cost pricing of transportation require measurement of the social costs related to consumption of transport (EC, 2019). While there are numerous issues related to the quantification of transportation costs, the goal of this chapter is mainly to overview key concepts related to external costs. We provide a definition of external costs, asymmetrical and reciprocal externalities, and market failure, and explain the differences between total, average, and marginal external costs. External costs provide a rationale for government intervention and pricing externalities. By internalizing these costs, externalities are made part of the decision-making process of transport users. This can be done through command and control measures or by providing the right incentives to transport users, namely with market-based instruments. Applying these instruments in an efficient way requires nevertheless detailed and reliable estimates of external costs. Because first-best scenarios are seldom met in the real world, the chapter also discusses why Pigouvian rules set to marginal external costs may not be optimal in second-best settings. For completeness, we further cover the roles of marginal, average, and total external costs (TESc) in policy analysis. The discussion of the key concepts will be illustrated with a highly stylized textbook case. It will also focus on the external costs of the use of infrastructure as a basis for market-based instruments to set transport prices right. Variable and fixed infrastructure costs and related charges are not addressed in this chapter.

External Costs: Understanding the Concept and its Use

Total social costs (TSCs) are defined as the full cost of travel within a geographical boundary. They include all costs due to the provision and use of transport infrastructure plus the external costs (congestion, accidents, and environmental costs). On the other hand, *total private costs* (TPCs) are costs directly borne by the transport user and include wear and tear and energy costs of vehicle use, own time

*The author is thankful to Maria Börjesson for her valuable comments on an earlier version of this present chapter.

Table 1 Policy goals and pricing rules

<i>Policy goal</i>	<i>Pricing rule</i>
Self-financing/cost recovery	Average cost pricing
Efficient use of infrastructures	Marginal cost pricing
Minimization of welfare losses with taxes	Ramsey pricing
Promote income distribution	Redistributive pricing

costs, and transport charges. The difference between the TSC and TPC costs represents the TEC. TECs refer then to the costs that are not borne by those who produce them, that is, to all external costs within a geographic area caused by a specific mode of transport. Thus, in order to measure them, one needs to know by whom these costs are borne in addition to the amount.

In general, social, private, or external costs can be measured in either marginal or average bases, by cost category (e.g., accidents, air pollution, climate change, noise) and by travel mode (road, buses or motorcycles, rail, and aviation) for passenger and freight transports. The *marginal cost* is defined as the incremental increase in total cost associated with an additional unit of an activity (e.g., number of drivers, road users, vehicle kilometer, road kilometer). The *average cost* is calculated by dividing total cost by total activity.

For certain externalities (air pollution, climate change), average and marginal costs are (approximately) equal to the size of the externality and do not depend on the density of the traffic flow. An extra car on a crammed traffic flow emits the same level of air pollutants as an extra car on a thin traffic flow, *ceteris paribus*. But for other externalities (accidents, noise, or congestion), the costs depend on the density of the traffic flow. An extra car on a road with free flow traffic will cause marginal external congestion costs lower than the average external congestion costs. When an extra car enters the traffic flow, at the moment, the capacity of the road is almost met; it causes marginal external congestion costs to be higher than the average costs.

The need to distinguish between private and social costs and to calculate marginal, total, and average external costs depends on the purpose of the policy analysis and on the policy instrument to be implemented (EC, 2019; Maibach et al. 2008; Van Essen et al., 2008; Verhoef, 1994). Policymakers can have different policy goals and therefore policy instruments must be set in a way to make it possible to reach specific policy objectives. Table 1 presents some examples of policy goals and its associated pricing rules. A more in-depth discussion on the merits of the alternative pricing rules to achieve a specific policy goal or even to help internalize the major transport externalities (accidents, CO₂, air pollution, and congestion) is beyond the scope of this chapter.

The concept of average external cost is typically used in policy analysis concerned with equity among different user groups (e.g., drivers vs. transit riders) or with cost recovery. The *unpaid bill to society* approach typically compares the average external costs of road transport not borne by road users (so excluding congestion costs) with the taxes paid by them to the rest of society net of the expenditures on new road infrastructure. From an equity perspective, the two should balance. But such a balance will not necessarily result in an efficient allocation of resources from an economic point of view.

When the goal is the efficiency usage of the transport system, the concept of *marginal external cost* is typically applied. Users of transport infrastructures can impose different external costs upon society. For example, users can deteriorate the road they use or they can slow down the speed at which other users travel. As such, users should pay for their marginal external costs, and only for these costs. This may be achieved by *marginal social cost pricing*. This will ensure the optimal usage of the infrastructure but not necessary a self-financed infrastructure. The reason is because when *marginal social cost pricing* is used, prices are equal to the sum of the marginal resource cost (e.g., extra cost of driver time, fuel, wear and tear of vehicle, all before taxes) and the marginal external cost (including congestion, air pollution, noise, accidents, and maintenance cost of the infrastructure), for a given infrastructure. Yet, there is no reason to expect the revenues produced by *marginal social cost pricing* to balance expenditures. *Marginal social cost pricing* is therefore likely to lead to surpluses, or, more likely, to deficits (Proost and Van Dender, 2004).

On the other hand, with *average cost pricing* production is self-financing, that is, total costs are equal to total revenues because prices are equal to the sum of financial infrastructure costs divided by its total usage volume. However, if the production of an activity (say for instance trips) is subject to constant marginal and average costs in the long run, then the efficient infrastructure facility (e.g., highway) in the long run will be the one where the collection of tolls (e.g., congestion tolls) just equals the cost of the land and capital embodied in the infrastructure.

Another challenge with *marginal social cost pricing* is that marginal external costs are difficult to estimate due to data limitations. Moreover, in the short run, marginal external costs are linked to constant infrastructure capacity, whereas long run marginal costs do take the construction of additional infrastructure into account. This implies, for example, that short-run marginal congestion costs are, in general, higher than long-run marginal congestion costs. The higher marginal short-run cost is due to the road infrastructure capacity being fixed, and therefore delay costs tend to increase rapidly as another driver enters an already crowded fixed-capacity road. Short-run marginal cost figures are typically more relevant for external cost internalization purposes and therefore for efficient pricing of existing infrastructure. On the other hand, the long-run marginal costs have to consider also the financing of infrastructure extensions.

It should be noted that both marginal and average external costs can be used for social cost benefit analyses. Whether marginal or average cost values are preferred depend on the scope of the analysis. For example, for an analysis of the construction of a new road, the noise costs can best be estimated by average cost figures because there is no existing traffic situation. On the other hand, if an

extension of a road from two to three lanes is being examined, the use of marginal cost figures is preferred because the change in an existing traffic situation is calculated.

Finally, TEC estimates have also a role in policy analysis because they allow to assess the extent of an externality and can provide a basis for comparing the burden of alternative transport modes. For example, the recent 2019 EU Handbook on the external costs of transport presents the TECs of transport for all EU member states by transport mode and cost category for the year 2016. It is shown that the overall external costs for road, rail, inland waterway transport, aviation, and maritime amount to € 987 billion, which corresponds to almost 7% of the total GDP of the 28 EU Member States. It is further revealed that the most important external cost category is accident costs equating to 29% of the TECs, followed by the congestion costs (27%). Climate change and air pollution costs both contribute to 14% of the total costs, noise costs to 7%, and habitat damage to 4% of the total costs. Another finding is that road is the largest contributor, accounting for 3/4 of TECs in absolute terms, and also the mode, which leaves the biggest amount of external cost unpaid.

External Costs: Graphical Analysis

Consider now Fig. 1 where we illustrate a motor-vehicle externality. For simplicity, we consider the case of *peak-period road congestion*. Congestion is usually the largest component of all road external costs in peak periods, whereas off-peak, air pollution, noise, and accidents have been found to be at comparable levels with congestion (EC, 2019). Peak demand generates larger traffic volume, and therefore a larger gap between private and social trip costs. In the horizontal axis, we measure traffic flow defined as the number of cars (Q) and in the vertical axis, we have trip costs.

In what follows, this road constitutes the entire transportation system. All travelers and road vehicles are homogeneous. We further assume that congestion is the only externality, all other markets are functioning perfectly. The road capacity is also fixed. Line D represents the demand for road use. The users' perception of the cost incurred when driving is given by the perceived average cost curve PPC . This curve may differ from the actual average cost curve if, for example, fuel taxes apply. Requiring each driver to pay fuel taxes increases the driver's perceived cost of a trip, raising the average cost curve. For simplicity, we assume for now that no taxes exist in this example.

In a competitive market, car users compare the extra direct benefit they get from one more trip by car, defined as the *marginal private benefit* (MPB) and represented by the demand function D , with the additional explicit private cost they pay to do that extra trip, defined as the *perceived private cost* (PPC). The MPB and the *marginal social benefit* (MSB) are equivalent in our example because external benefits from car usage (e.g., urban agglomeration benefits) are absent. Moreover, D slopes downward, incorporating the assumption that as more trips are undertaken, the value of extra trips declines. The PPC slopes upward as time cost increases with the number of cars on the road.

In addition to the private costs paid by the car users, there are also costs imposed to other users when an individual travels by car. An additional car user faces the average journey time and his presence on the road reduces the speed of all other users and increases their travel times. Consequently, this extra user imposes a negative externality on other users over and above the time costs the additional user faces. Graphically, this means that the *marginal social cost* (MSC) curve lies above the PPC curve by an amount equal to the *marginal external cost* (MEC).

In Fig. 1, we assume that the MSC is equal to the PPC up to a threshold Q_T , denoted as traffic free flow volume. This threshold represents the number of cars at which traffic has reached the assimilative capacity of the road, after which congestion begins to cause time delay damages. After Q_T has been reached, MSC exceeds PPC and grows more quickly. At low traffic flows, users can travel at the

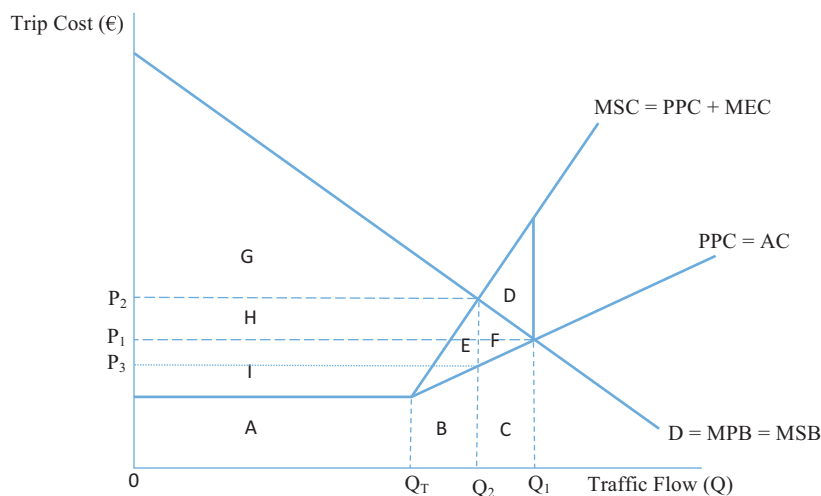


Figure 1 External costs of transportation.

free-flow speed, and the PPC is constant. After the traffic congestion develops at higher flows that cause decreases in speed, the PPC slopes upward.

Without policy intervention and in the presence of such external cost, the competitive market equilibrium, Q_1 , occurs at the intersection of the PPC and the MPB. At Q_1 , society maximizes the market surplus because the benefits to the market derived from the last car user equal the additional cost incurred. The TPC for all car users engaged in driving at Q_1 is given by area $A + B + C$, while the TSC is represented by area $A + B + C + D + E + F$. The difference between these two total cost measures, area $D + E + F$, represents the TEC (of congestion) evaluated at Q_1 . The *total social surplus* at Q_1 is, however, represented by area $G + H + I - D$ and equals the total social benefits net TST.

It is apparent from Fig. 1 that Q_1 is inefficient from a society perspective, as by moving to a quantity lower than Q_1 , say Q_2 , we raise total social surplus to $G + H + I$. In fact, Q_2 is the *social optimum equilibrium* since it maximizes the total social surplus. At Q_2 , the MSB equals MSC for the last car on the road. Because congestion costs are borne by other road users, the road user himself has no incentive to take them into account. This makes commuting on the congested road look artificially inexpensive and, excessive road usage occurs from society's perspective ($Q_2 < Q_1$). To correct this problem some traffic should be diverted to off-peak hours, when roads are less congested, and some car users should switch to an alternative travel mode.

An interesting implication from Fig. 1 is that social optimality does not necessarily imply that the externality should be set to zero. With moderate reductions in traffic flow, the marginal private losses from reducing congestion, as represented by the vertical difference between D and PPC, may outweigh at some point the MEC. Thus, part of a policy challenge is to determine the efficient level of the externality. Yet by leaving the market unregulated, society is worse off than if the activity (driving) had been restricted by regulation. The triangle defined by area D represents the net social cost of unregulated road use or the deadweight welfare loss to society. A *deadweight loss* (DWL) is a cost to society created by market inefficiency. It is an indicator of the degree of market failure caused by a distortion or an externality. The existence of triangle D provides a rationale for government intervention.

External Costs: Additional Remarks

Two remarks are now in order. First, so far congestion has been the only negative externality generated by car usage. As such, only the marginal external costs of congestion were included in the MSC curve in Fig. 1. Yet, multiple externalities coexist in the transport system. Road users also generate air pollution, noise, and accidents. In such a case, the marginal external costs of air pollution, accidents, and noise should too be added up to derive the total MSC curve.

Second, certain transport externalities (e.g., motor carbon monoxide emissions or motor vehicle noise annoyance) create damages not only to the parties producing it (e.g., car users) but also to third parties (pedestrians or nearby residents). These types of externalities are also known as *intersectoral externalities* because its external effect is posed upon society at large. As such, measures of the TEC for these types of externalities that result from multiplying the MEC by the amount of road traffic can be misleading if such costs are mostly borne by exposed individuals outside the transport system.

Take for example the cases of two important externalities generated by car users: congestion and air pollution. While air pollution is an example of an *intersectoral externality*, congestion is an example of what we call an *intrasectoral externality* because it reflects reciprocal inefficiencies in the transport system whereby road users do not account for the external effects of their decisions on other road users. Note, however, that the congestion inefficiency arises from decentralized decision-making of users. Because the producers of the congestion externality and the victims are the same (are within the same transport system), road users are actually paying the sum of external congestion costs, just not in a socially optimal way as seen in Fig. 1. This may then explain why congestion is considered an internal impact for equity analysis (because individuals bear the same amount of delay that they impose). However, for efficiency purposes, congestion should be treated as an external cost because cost bearing and decision making are separated: the sum (over all road users) of the additional delays from an individual road user can be very much greater than the average delay (experienced by each individual), which formed the basis of the decision to travel. Thus, because it is an external cost at the individual level, traffic congestion is economically inefficient. This suggests that individual payments of the car users must be adjusted to reduce the inefficiency of road capacity use due to decentralized decisions. This in turn allows the resource (road) to be put to its highest value use.

But road users also impose *unilateral or asymmetrical externalities* to other parties such as air pollution. With an asymmetrical or nonreciprocal externality, the producer and consumer of the effect can be separated. Note that in the case of air pollution, the perpetrators are still the road users but the victims comprise also the exposed individuals outside the road system. Therefore, it is questionable that road congestion can simply be added to the external costs of accidents and the environment to construct a TEC bill, which road users would have to pay in addition to the infrastructure costs. As explained earlier, in the case of congestion, because the external costs are reciprocal, road users are already paying the sum of external congestion costs. TEC that road users should cover consists of accident and environmental costs generated through their transport activity imposed on individuals outside the transport system.

While understanding who bears the external cost of an activity is important when calculating the TECs associated with a specific activity and externality, the MEC, which is important information to set policies that correct transport prices to account for their societal impacts, is not influenced by whether or not an externality is reciprocal in nature. Note from Fig. 1 that because the $MSC = MPC + MEC$, then a tax equal to the MEC at the efficient traffic level added to the MPC will internalize the congestion externality and result in the efficient traffic volume Q_2 . Finally, transport externalities interact with other markets outside the

transport sector and have feedback effects. As such, their effects should be examined in a context of dynamic efficiency. Next, we discuss how underpriced road congestion can affect modal split and the labor and land markets.

External Costs and Travel Mode Choice

We have seen that when there is a discrepancy between transport prices and social costs this leads to an inefficient demand for transport. Since these discrepancies differ between travel modes, the modal split of transport will also be suboptimal with unpriced congestion. **Fig. 2** illustrates the choice between using car or public transit in the presence of unpriced congestion.

Let total demand be given and set to 1, divided between car demand n and public transit demand $(1 - n)$. In the horizontal axis of **Fig. 2**, we measure the proportion of individuals that commute by car and by public transit. In the vertical axis, we measure trip (generalized) costs for users. In the case of the car, the trip cost includes maintenance costs and fuel consumption, and the time cost is defined by the congestion rate on the road. In the case of public transit, the trip cost is the fare price and the time cost is defined by the time on board and waiting time.

The decision whether to undertake a trip or not is determined by whether an individual considers the *generalized cost* (the sum of the monetary and nonmonetary costs of a trip) to be smaller or greater than the benefit contained in reaching his destination. Should the individual decide that he will undertake the trip, he chooses the travel mode with the lowest generalized cost.

Note that we represent the cost of public transit by a constant average cost per passenger, while the PPC of car use is upward sloping as discussed before. We further assume that crowding externalities in public transit are addressed using increases in frequency so that the average generalized cost of public transit is constant. This implies that the marginal social cost (MSC_p) and the private cost (PPC_p) for public transit are the same but, the marginal social cost of using the car (MSC_c) is higher than its perceived private cost (PPC_c).

In the absence of policy intervention, individuals only make travel mode choices based on their private costs. Therefore, the unregulated market equilibrium occurs when the PPC_c equals MSC_p . At this point, an individual is indifferent between the two travel modes. When the PPC_p is smaller than the PPC_c , the individual chooses traveling by public transit; otherwise, the car is the preferred travel mode because it exhibits the lowest generalized private cost. At this unregulated equilibrium, there are n_1 car users and $(1 - n_1)$ public transit riders. This market allocation across the two modes is nevertheless inefficient because car users do not take into account all the costs of commuting by car. Since the trip benefit is the same for both travel modes, the welfare optimum is obtained by minimizing total user costs for all travelers. This requires that the optimal allocation across the two travel modes be where the MSC_c equals MSC_p . In the social optimum, the number of car users would be lower ($n_2 < n_1$). Policy action that either disincentives car usage or increases the attractiveness of public transit can then reach n_2 .

In a first-best scenario, all travel modes should pay their marginal social cost. This suggests that car users should pay in addition to their private costs a congestion tax equal to the MEC so that the congestion externality is internalized in their travel mode choices. However, when road congestion pricing is either not feasible or politically unacceptable, public transit subsidies often emerge as a useful component of second-best policies designed to alleviate the external costs in the transport sector (De Borger and Swyssen, 1999). The rationale for public transit subsidies is that public transport is a substitute of private road use and therefore lower transit fares encourage a shift from automobile use to public transit modes, thereby reducing the social costs from congestion, air pollution, and accidents. The welfare change from the induced substitution into transit depends nevertheless on the relationship

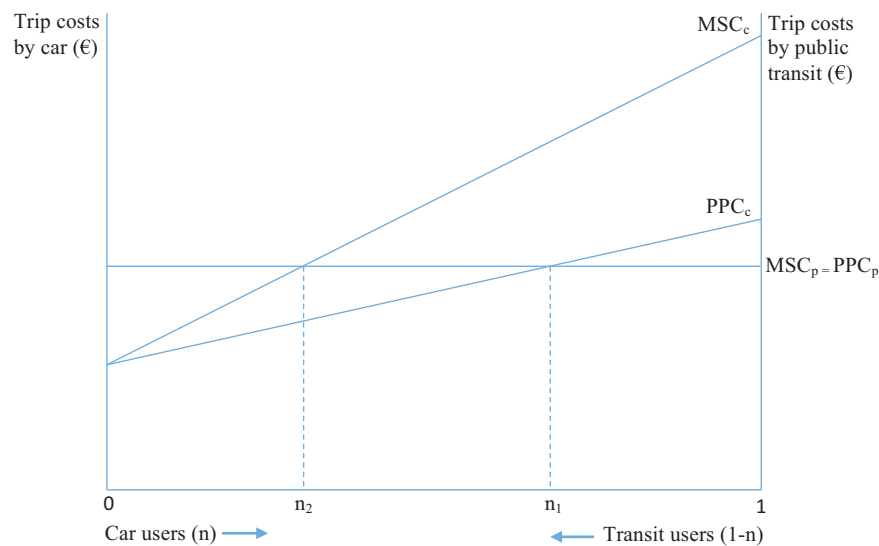


Figure 2 Modal split in the presence of external costs.

between the transit fare and the marginal social cost of service provision (Small and Verhoef, 2007). If the fare is less (greater) than the MSC_p , the increase in demand for public transit produces a welfare loss (gain).

External Costs and Transport–Land Use Interactions

One other issue associated with congestion is that the average commute distance is too long from society's perspective, and it should be shortened. Excessive long average commute means that cities are too spread out. Therefore, by causing people to commute too far, road congestion can lead indirectly to urban sprawl (Brueckner, 2000, 2007). Fig. 3 illustrates how external costs in transportation can affect land use in a simple monocentric city setup.

Assume a linear city with a single central business center (CBD) located at point zero. All workers live in the city, commute by car to the CBD to work, and face commuting costs. The horizontal axis in Fig. 3 represents distance from residence to the CBD, x . The commuting cost from a residence site from the CBD equals the commuting cost per round-trip mile times x .

Workers consume a numeraire composite good and a fixed amount of housing services (measured by acres of residential land for simplicity) taking into account their budget constraint. Given perfect mobility (zero moving costs) and identical workers, the urban equilibrium must yield the same utility level for all individuals. Spatial variation in residential land rent then allows equal utilities throughout the city. This generates a residential land bid-rent function decreasing with distance from the CBD, $R_1(x)$, as depicted in Fig. 3. Workers living far from the CBD are compensated for their long and costly commutes with a lower land rent relative to close-in locations. In a market for residential land with identical workers and no variation in land features aside from the distance from the CBD, the equilibrium residential land rent equals each worker's willing to pay for residential land at each location.

At each location from the CBD, land is allocated competitively either to residential use or agriculture. Agricultural land rent is spatially invariant and equal to R_a . Then the market city's equilibrium size, x_1 , occurs where the land bid rent curves intersect. To the right of x_1 , it is the agricultural bid rent that is higher, so that only agriculture secures land. To the left of x_1 , the bid rent for agricultural land is less than for housing so that all land up for bid is secured by urban residents. However, in the presence of road congestion, x_1 is not the optimal city size because commuting costs are underpriced.

When commuting costs increase (e.g., because of a road congestion tax), commute trips of any given length become more expensive, and as a result, close-in locations become more attractive given the original pattern of residential rents. The desire of workers to move closer to the CBD bids up central rents and reduces rents at more distant locations, causing a clockwise rotation of the residential bid land rent. This explains the location of the social residential bid land rent function, $R_2(x)$, in Fig. 3 relative to $R_1(x)$. The social optimal city size occurs where $R_2(x)$ intersects R_a yielding a smaller feasible area of residence ($x_2 < x_1$).

Thus, any policy intervention that either increases transportation costs or affects land-use practices can help curb traffic-induced externalities and push the city size to x_2 . In such a context, land-use regulations by changing population distribution emerge as another potential second-best tool when road congestion pricing has either large implementation costs or limited political acceptability. For instance, Brueckner (2007) and Anas and Rhee (2007) have investigated the efficiency of an urban growth boundary relative to a congestion toll and Kono et al. (2012) have evaluated the efficiency of regulations on building size and city size relative to the gains that can be achieved by a first-best road toll. More recently, Tikoudis et al. (2018) and Kono and Kawaguchi (2017) consider road tolls and floor-to-area ratio regulations simultaneously. All these studies have shown that in a monocentric city, residential locations should be centralized by optimal land-use regulations when there is only car commuting as in Fig. 3.

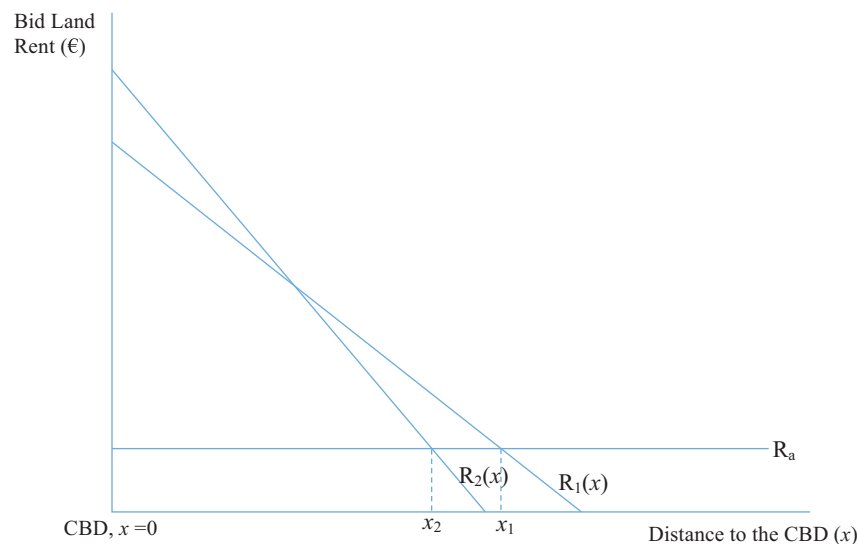


Figure 3 External costs and land use.

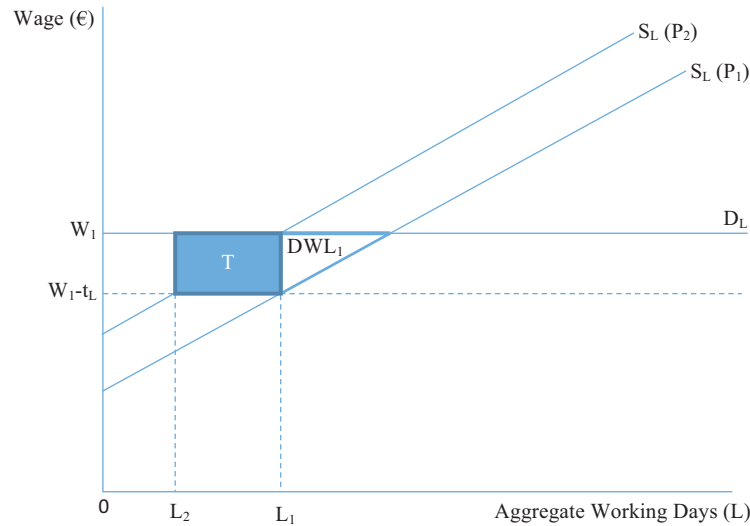


Figure 4 Transport external costs and labor market.

However, [Buyukeren and Hiramatsu \(2016\)](#), in assuming a congested car mode and an uncongested public transit mode, show that an expansionary UGB would be optimal under certain conditions.

External Costs and Labor Market Interactions

The primary means to raise revenues to fund the provision of public goods is through income taxes, which are taxes on wages that distort the labor market. The Deadweight Loss (DWL) of a tax occurs if there is no divergence between marginal private and marginal social cost. In such a situation, a tax (such as revenue-raising taxes) reduces the economic surplus that could be gained from an activity. [Parry and Bento \(2001, 2002\)](#) have examined how unpriced congestion and other road transport externalities interact with pre-existing distortions in the labor market. Next, we illustrate a simplified version of their theoretical analyses using both [Figs. 1 and 4](#).

[Fig. 4](#) represents the labor market where an initial income tax t_L applies to labor. Income tax revenues are assumed to be lump sum redistributed and the purpose of transport activity is just for people to get to work. We further assume that people commute to work using a congested road.

For simplicity, let us assume constant returns to scale in production and that labor is the only primary input, $Q(L)$. This in turn implies a perfectly elastic labor demand curve as seen in [Fig. 4](#). The labor demand curve shows the value of the marginal product of labor. With constant returns to scale the marginal product of labor is constant as total quantity changes. This makes the production function homogeneous of degree one, which can be represented as $Q = \frac{\partial Q}{\partial L}L$. Because labor is paid at a rate equal to its marginal product value, $w = p \frac{\partial Q}{\partial L}$, then the complete value of the production will be distributed to the labor factor, since $pQ(L) - wL = pQ(L) - p \frac{\partial Q}{\partial L}L = pQ(L) - pQ(L) = 0$. Hence, firm profits are zero. Moreover, since labor demand is perfectly elastic, the equilibrium amount of employment is determined by labor supply as illustrated in [Fig. 4](#). Note that in [Fig. 4](#), labor is measured by number of workdays (with daily workhours fixed) and not by number of workers. Also, a tax on labor income is a distortion because it reduces the supply of labor, resulting in a less than efficient level of employment. So, now the question is to understand how might this distortion affect the prescriptions for a road congestion tax.

Individuals face a tradeoff between work and leisure. Individuals decide the optimal number of working days by equating the private benefit from an extra day of work (the net-of-tax daily wage) with the private cost. The daily wage before taxes is represented by W_1 but after taxes is represented by $W_1 - t_L$. The private cost is the value of leisure time forgone by working and commuting an extra day plus commuting cost. Thus, if the labor supply curve is upward-sloping over some range of net wages, it means that a higher opportunity cost of leisure induces people to take less leisure and work more. If the net wage is only $W_1 - t_L$, then workers choose L_1 working days. The income tax imposes an excess burden or deadweight loss on the economy equal to the triangle area DWL_1 . The rectangle defined by $t_L \cdot L_1$ in [Fig. 4](#) represents the income tax revenues.

From [Fig. 1](#), when road users ignore the MEC of congestion, traffic flow is Q_1 at price P_1 , resulting in a DWL of area D . In contrast, if congestion was internalized, then the DWL in the transport market would be zero, traffic flow would be smaller ($Q_2 < Q_1$), and trip costs would rise from P_1 to P_2 . This change in trip costs has implications for the labor market. Note that the supply of labor depends on work commuting trip costs. A change in commuting costs shifts the labor supply curve because it affects the overall return to work effort relative to leisure.

If leisure and work are substitutes in consumption, the increase in commuting trip cost causes substitution of leisure for work (and also commuting). As a result, labor supply shifts leftward from $S(P_1)$ to $S(P_2)$, which reduces both the amount of labor supplied from L_1 to L_2 and tax revenues in an amount equal to the shaded rectangle T . Therefore, unpriced congestion generates a welfare loss in the transport market given by triangle D in Fig. 1 and an increase in income tax revenues in the labor market equal to area T in Fig. 4. This increase in income tax revenues partially offsets the external congestion cost in the transport market. The general equilibrium social loss from unpriced congestion thus appears to be less than would be anticipated on the basis of a partial equilibrium analysis where the income tax distortion is ignored. The net welfare effect of introducing a policy that would correct the road congestion externality would equal area $D-T$.

But as explained earlier, unpriced congestion also affects the land market and therefore, the general equilibrium effects of such unpriced road externality and internalization policies can be quite complex. Moreover, a labor supply model, which also allows for optimally chosen daily workhours, implies that (monetary and time) commuting costs increase the number of hours worked per day, and thus the effect on total labor supply is ambiguous. And, if individuals can only choose the optimal number of daily workhours (and thus cannot adjust the number of working days), then the income tax would not help internalizing congestion as seen earlier. Note that in order to go to work, it is necessary to commute. Thus, a labor tax and the road congestion charge (which is set per round-trip or equivalently, per workday) have the same effect because they both affect the same choice margin, whether to work or not. But if road users cannot adjust workdays then the two fees do not affect the same choice margin anymore.

Another factor to take also into account is that workers may react quite differently to an increase in monetary commuting costs than to a decrease in wages. Furthermore, commuting time may even reduce workdays if the number of working days can be adjusted. Therefore, the effect of commuting time and monetary commuting costs on total labor supply is ambiguous, as it is not clear a priori whether the effect on daily hours or workdays dominates.

All of these cases highlight the complexity of the interactions of road congestion pricing with preexisting distortions in the economy. But are also extreme examples because in reality, there is no strict proportionality between labor supply and commuting, labor markets are not perfectly competitive, not all trips are work-related trips and commuters are an heterogeneous group. In general, this means that the results discussed earlier may play a weaker impact in reality (De Borger, 2009; Van Dender, 2003). Finally, it is important to keep in mind that labor taxes exist for more than one reason, being one notable motivation a desire for income redistribution. Later in the chapter, we will discuss further more second-best issues and draw additional conclusions on how second-best analysis can help improve the practical implementation of congestion charging systems. Because the literature on the second-best analysis is vast, this chapter still makes no attempt at providing an overview, referring the interested reader to a concise discussion in Small and Verhoef (2007).

Policies for Obtaining Social Optimality with External Costs

Internalization is a way to ensure that each transport user pays the social costs associated to his individual trip. It can be done by a variety of methods and instruments. So far, our discussion has focused mostly on price instruments and in particular road congestion charges. But fuel taxes and cordon charges are other examples of the so called *market-based instruments* because these policies also use economic incentives through price-based measures to regulate the level of the externality. On the other hand, vehicle standards, fuel standards, driving restrictions, low emission zones, parking restrictions are examples of *command-and-control instruments* because they are imposed on private actions by government through regulation (such as setting standards, targets, or process requirements). A combination of these two basic types of instruments is also possible, e.g., taxes differentiated to Euro emission classes of vehicles. For the purpose of illustration of both general approaches, let us consider again the externality described in Fig. 1.

A *Pigouvian tax* is a tax named after A.C. Pigou (1920) that imposes the external cost on the perpetrator of the externality. Where MEC is represented by vertical difference in MSC and PPC in Fig. 1, a Pigouvian tax is determined by the vertical difference in these two lines at level Q_2 , P_3-P_2 . When the marginal tax equals the MEC of congestion associated with Q_2 , the road user bears the full costs of his travel mode choices, and thus the PPC in the presence of this tax (known as a *congestion tax*) reflects the MSC. In this case, $Q_1 - Q_2$ represents the number of road users who will find other ways (e.g., time or mode) of travel or cancel the trip all together. The net gain from a congestion tax is represented by triangle D in Fig. 1 and results from subtracting the deadweight loss of the tax (triangle F) from the avoided external cost ($F + D$).

Four remarks should now be made. First, the appropriate tax to internalize the externality is the marginal and not the average external cost, the latter being lower. Second, because the external cost of transport consists of several elements (e.g., noise, air pollution, accidents), which are assumed to be additive, it follows that the optimal internalization strategy prices transport activities at their total marginal social costs. Third, the coexistence of multiple externalities within the transport market and their spillover effects to markets outside the transport sector requires that a first-best analysis be conducted with a general equilibrium framework. As seen in Fig. 4, when congestion is internalized, labor force participation is discouraged at the margin because of the reduction in the wage net of taxes and commuting costs. Also, changing the modal split is not an explicit goal of a congestion internalization strategy, but such change can be induced as seen in Fig. 2. Pollution affects the level and composition of economic activity which, in turn, affects the level of pollution. Thus, feedbacks in terms of compensated demand responses affect the size of the realized benefits of an internalization policy. Fourth, because a tax is being used to internalize the externality, the government gains the area $(P_3 - P_2) \times Q_2$ in tax revenues. The tax revenues is not of itself an economic cost but it represents a transfer to the rest of society.

The use of these revenues is an integral part of the internalization policy, but the goal of internalization is not revenue raising but holding users accountable for their external effects. Under the Pigouvian tax framework, the revenues are not supposed to go to those affected by the externality as such would lower their incentive to avoid the externality below efficient level. For the same reason, it should not be used to directly reimburse those responsible for the externality.

Consumer surplus is typically used by economists to measure how well-off individuals are. It is defined as the difference between the total amount that *consumers* are willing and able to pay for a good or service (indicated by the demand curve) and the total amount that they actually pay (the market price). It is interesting to observe from Fig. 1 that drivers lose consumer surplus under the Pigouvian tax, which may explain why it is politically difficult to introduce road charging. The welfare gain of internalizing the externality is small relative to the loss in consumer surplus and the gain in tax revenues. Opposition to road charging can be even stronger when demand is more inelastic (lower responsiveness to price) as tax revenues and loss to drivers both increase and, the DWL from excessive congestion is smaller. It is then in such a context that attention in the congestion pricing literature has moved to more realistic types of second-best congestion pricing, in which various costs or constraints deter or prevent the setting of first-best tolls. Examples of second-best tolling include the use of toll cordons around cities instead of tolling each road in the network or the use of step tolls instead of smoothly time-varying tolls. The literature has also studied how other more politically feasible policies may help mitigate congestion costs in a second-best setting. One of such examples is the use of public transit subsidies.

The use of a command-and-control policy such as driving restrictions can also be represented in Fig. 1. Since Q_2 represents the social optimal number of cars in the road, the government may limit the number of cars that can circulate to Q_2 . With the imposition of such regulation, traffic flow is reduced from Q_1 to Q_2 , which again induces a price increase from P_1 to P_2 .

In an ideal world, Pigouvian taxation and command-and-control regulation would be identical as seen above. In practice, there are complications that may make taxes a more effective way of dealing with externalities. Which approach leads to the most efficient regulatory outcome depends on the heterogeneity of travelers and travel mode being regulated, the flexibility embedded in quantity regulation, and the uncertainty over the costs of externality reduction. For instance, in the case of global warming, the marginal damage is fairly constant over large ranges of emissions and thus emission reductions. If costs are uncertain, then taxation leads to a lower DWL than does regulation. On the other hand, if marginal damage is very steep and costs are still uncertain, the result is reversed.

Internalization in Second-Best Settings

As already discussed, first-best pricing requires that the MSC pricing be set throughout the whole transport network or for all competing modes. If limited to a single travel mode (e.g., car) or only a part of a network (e.g., highways), this may give rise to a shift from the priced travel modes or parts of the transport network to the other network parts (e.g., uncongested roads) or travel modes (e.g., public transit). From a welfare point of view, this could lead to lower positive welfare effects.

In addition, the MEC varies with traffic conditions, speed-flow relationships, vehicle type, capacity demand, the value of travel time, and vehicle's occupancy. Thus, a first-best Pigouvian tax system requires not only knowledge about demand, PPC and MSC so that a tax can be calculated but, it also needs to be able to differentiate price levels according to drivers for the various external costs. This requires a technological system, which may be complex or expensive to use. In the case where prices cannot differ by user group and take into account all these dimensions of heterogeneity, it is shown that the second-best tax is a weighted average of the MECs for different groups (Small and Verhoef, 2007).

The question is then to understand what should be the second-best pricing scheme, which aims to achieve efficiency under second-best scenarios. In practice, actual congestion pricing schemes (e.g., cordon schemes, partial charging like toll roads, or HOT lanes) are expensive and imperfect. Because they do not involve perfect marginal cost pricing, the gains are also smaller than under first best. Moreover, second-best settings come in several variants as already exemplified in this chapter. So, to gain additional insights of the second-best issues for road charging, we now turn the discussion to other different sources of second-best distortions in isolation.

Coexistence of multiple externalities or other types of distortions in the same market: If market power from a monopoly is taken into account, then the tax that maximizes social welfare is not the same as when the monopoly distortion is absent. A tax based only on MEC ignores the social cost of further output contraction by a monopolistic perpetrator whose output is already below an optimal level. Ideally, we could have a policy to increase the production level together with a tax to control the externality. But if the product market distortion cannot be directly corrected, then the tax must achieve an optimal second-best tradeoff of distortions (Small and Verhoef, 2007). Corrective taxes reduce external costs but they add to the DWL attributable to the monopolist's restricted output.

As discussed earlier, the coexistence of multiple externalities implies that the level of one externality (e.g., congestion) code-termines the level of the others (e.g., air pollution, accident, noise). Any policy directed toward one of these external costs will also affect the others (Bento et al., 2006; Mayeres, 2000; Parry and Bento, 2001, 2002). This in turn affects how one should calculate the welfare effects of a policy intervention and measure its effects on TECs. Take again the case of a congestion tax. A congestion tax set to the MEC internalizes the externality on the congested road but it exacerbates congestion, air pollution, and potential accidents on competing roadways. In addition, permanent shifts in the demand for travel between modes lead to changes in transport infrastructure investments. Because there are also distortions between the MSB and the MSC of investment in different travel modes, these changes induced by a policy that alters the level of an externality may give rise to significant welfare effects (net benefits) and excess burden (or gross cost).

Distortions in other markets: Similar problems can emerge if a corrective policy is implemented in a market when distortions exist in other related markets as seen in Fig. 4. The potential for a double dividend may nevertheless emerge because the tax revenues from internalizing an externality can be used to reduce preexisting distortions in a revenue-neutral way (e.g., cut labor taxes) (Parry and Bento, 2001, 2002; Proost, 2011). For instance, Arnott et al. (2005) identify a potential triple dividend from the pricing of parking in downtown areas: reduced search for on-street parking, reduced urban traffic congestion, and use of parking revenues to lower other taxes.

Long-run approaches: Most efforts to estimate the MSC and the MPC of travel have focused on short-run costs. However, transport costs and land use patterns are interlinked as seen in Fig. 3. Changes in transportation costs can induce long-run changes on where people reside and work with implications for the calculations of the MSC (Zhang and Kockelman, 2016). Over time, individuals may change their place or time of work, or move house, in order to save time (Downs and Downs, 2004), although taxes that increase the cost of choosing appropriate vehicles or moving home act against this. In general, static monocentric models show that congestion pricing, both in first and second-best settings, increase population density toward central locations. Fig. 3 provides the intuition for such an outcome. On the other hand, when firms' locations are endogenously determined, congestion pricing seems to disperse jobs, though the existing results are highly dependent on the framework and assumptions. In any case, a policy implication of such result is that to achieve city-wide welfare gains, efficient land-use regulations should allow job decentralization. Studies that have examined second-best congestion pricing in a monocentric city with distortionary, rigid regulatory mechanisms in the housing market (e.g., building height restrictions, zoning and property taxation) show that a Pigouvian toll retains its optimality under any setting with quantity restrictions in the housing market (Tikoudis et al., 2018). However, the extent of the quantity restriction determines the volume of the welfare gains in a nonmonotonic fashion. On the other hand, when congestion and misallocation of jobs within the city coexist, a Pigouvian tax strategy is not a social optimum as this strategy cannot solve two problems simultaneously.

Parking Interactions: Another major source of inefficiency in transport markets, especially in urban settings, is parking underpricing, particularly at the workplace (Franco, 2017, 2019; Shoup, 2005). Most drivers do not pay for the resource cost of their parking spot nor for the external costs of cruising for a parking space. Parking costs are nevertheless part of the generalized cost of using the car. Therefore, reducing parking supply or increasing parking prices would increase the PPC of using the car and thus, reduce car journey, increase the use of alternative travel modes and tackle road use externalities, especially peak-period congestion in high density urban core areas (Shoup, 2005). Existing studies have shown that parking fees and area-based tolls for entry into or traveling within a defined area (e.g., a cordon charge) need to be examined simultaneously (Calthrop et al., 2000; Small and Verhoef, 2007; Verhoef, 2005). In the absence of efficient parking fees, the optimal level of a cordon toll rises. The highest welfare gains are nevertheless obtained when both policies are implemented. However, efficiency gains require that the parking fee be borne by the employee and not the employer (Brueckner and Franco, 2018; Franco, 2019). One striking result from the existing literature is that the sole use of parking fees can generate greater welfare gains than the sole use of just one cordon charge, suggesting that parking policies can be used in place of road pricing. However, both policies are imperfect tolls and thus future research should be done to understand further the welfare and efficiency effects of road pricing in the presence of free and subsidize parking.

Conclusions

In the transport sector congestion, air pollution, noise, and accidents are sources of inter or intrasectoral external costs leading to welfare losses and an inefficient market equilibrium. External effects arise if marginal private user costs deviate from marginal social costs. Marginal social cost pricing means setting generalized prices equal to the sum of marginal producer costs, marginal private user costs, and marginal external costs. By charging the optimal price, external effects are internalized and taken into account by transport users. Under such an approach setting, the right prices provide the right incentives to achieve market efficiency and maximize social welfare.

Such right incentives follow from a solution of a well-established model of social welfare. The solution requires charging a tax equal to the difference between marginal social costs and private average costs of traveling to restore the socially optimal equilibrium. Since first-best conditions may not exist in reality, second-best solutions are required, and workable instruments have to be developed with respect to technology, transaction costs, and social acceptance. Insights from the first-best solution may nevertheless help to develop a second-best pricing strategy.

Another challenge is how to define an optimal bundle of instruments to achieve a desired pattern of the transport system. Since different price instruments are likely to be introduced for different types of external effects to optimize the triggers with respect to effectiveness issues, it may be the case that interdependencies between instruments either affect multiple externalities or be counterproductive to other objectives.

In addition, the introduction of charges to internalize external costs leads to revenues. Yet the aim of internalization is not revenue raising but holding transport users accountable for the externalities they create. Yet, the revenues generated by an internalization policy can be used to attain other goals. For example, applying revenues from congestion charges for providing alternatives, e.g., investment within the modes, helps in gaining public support.

Finally, good assessment and monetization of the TECs of the transport sector is also a prerequisite when it comes to target-based internalization strategies (Musso and Rothengatter, 2013). In such a context, social welfare is understood as a bundle of long-term economic, environmental, and social targets summarized by different indicators which the market fails to achieve. External

costs associated to existential risks for human life, nature, climate, or cultural heritage can help establish safe minimum requirements for target achievement. This type of external costs, however, cannot be traded-off against monetary compensation nor be integrated into a social cost function in additive way as seen in the traditional marginal social cost approach. Moreover, it is important to define the share of the different sectors (transport, energy, industry, households) in the overall target achievement (e.g., a GHG reduction target by a certain date). This requires knowledge of the TECs generated by each sector to understand their burden (e.g., transport) and the burden of their subsectors (e.g., road transport). Internalization in this context then requires an adjustment of the market mechanism in the transport sector so that such targets can be attained. The adjustment can be done through a set of instruments designed and optimized to attain such targets with minimum economic costs. This broader concept of internalization is widely followed by the European Union (EC, 2019).

References

- Anas, A., Rhee, H.J., 2007. When are urban growth boundaries not second-best policies to congestion tolls? *J. Urban Econ.* 61, 263–286.
- Arnott, R., Rave, T., Schöb, R., 2005. *Alleviating Urban Traffic Congestion*. MIT Press, Cambridge, Mass.
- Bento, A., Franco, S.F., Kaffine, D., 2006. The efficiency and distributional impacts of alternative anti-sprawl policies. *J. Urban Econ.* 59 (1), 121–141.
- Brueckner, J.K., 2000. Urban sprawl: diagnosis and remedies. *Int. Regional Sci. Rev.* 23, 160–171.
- Brueckner, J.K., 2007. Urban growth boundaries: an effective second-best remedy for unpriced traffic congestion? *J. Housing Econ.* 16 (3), 263–273.
- Brueckner, J.K., Franco, S.F., 2018. Employer-paid parking, mode choice, and suburbanization. *J. Urban Econ.* 104, 35–46.
- Buyukeren, A.C., Hiramatsu, T., 2016. Anti-congestion policies in cities with public transportation. *J. Econ. Geogr.* 16, 395–421.
- Calthrop, E., Proost, S., 1998. Road transport externalities. *Environ. Resour. Econ.* 11, 335–348.
- Calthrop, E., Proost, S., van Dender, K., 2000. Parking policies and road pricing. *Urban Stud.* 37, 63–76.
- De Borger, B., 2009. Commuting, congestion tolls and the structure of the labor market: optimal congestion pricing in a wage bargaining model. *Regional Sci. Urban Econ.* 39, 434–448.
- De Borger, B., Swysen, D., 1999. Public transport subsidies versus road pricing: an empirical analysis for interregional transport in Belgium. *Int. J. Transport Econ.* 26, 55–89.
- Downs, A., Downs, A., 2004. *Still Stuck in Traffic: Coping with Peak-Hour Traffic Congestion*. Brookings Institution Press, Washington, DC.
- European Commission (EC), 2019. 2019 Handbook on the external costs of transport. Available from: <https://ec.europa.eu/transport/sites/transport/files/studies/internalisation-handbook-isbn-978-92-79-96917-1.pdf>.
- Franco, S.F., 2017. Downtown parking supply, work-trip mode choice and urban spatial structure. *Transport. Res. Part B: Methodol.* 101, 107–122.
- Franco, S.F., 2019. Parking Prices and Availability, Mode Choice and Urban Form, OECD/International Transport Forum, ITF Round Table on Zero Car Growth? Managing Urban Traffic, OECD Publishing/ITF.
- Kono, T., Joshi, K.K., Kato, T., Yokoi, T., 2012. Optimal regulation on building size an city boundary: an effective second-best remedy for traffic congestion externality. *Regional Sci. Urban Econ.* (42), 619–630.
- Kono, T., Kawaguchi, H., 2017. Cordon pricing and land-use regulation. *Scand. J. Econ.* 119, 405–434.
- Maibach, M., Schreyer, D., Sutter, D., van Essen, H., Boon, B., Smokers, R., Schrotten, A., Doll, C., Pawlowska, B., Bak, M., 2008. Handbook on estimation of external costs in the transport sector. Internalisation Measures and Policies for All external Cost of Transport, Version 1. 1. European Commission DG TREN, Delft, CE, The Netherlands.
- Mayeres, I., 2000. The efficiency effects of transport policies in the presence of externalities and distortionary taxes. *J. Transport Econ. Policy* 34, 233–260.
- Musso, A., Rothengatter, W., 2013. Internalisation of external costs of transport—a target driven approach with a focus on climate change. *Transport Policy* 29, 303–314.
- Parry, I.W.H., Bento, A., 2001. Revenue recycling and the welfare effects of road pricing. *Scand. J. Econ.* 103, 645–671.
- Parry, I.W.H., Bento, A., 2002. Estimating the welfare effect of congestion taxes: the critical importance of other distortions within the transport system. *J. Urban Econ.* 51, 339–365.
- Pigou, A.C., 1920. *The Economics of Welfare*. MacMillan, London.
- Proost, S., 2011. Theory of External Costs, Chapters. In: *A Handbook of Transport Economics*, chapter 14. Edward Elgar Publishing.
- Proost, S., Van Dender, K., 2004. Marginal social cost pricing for all transport modes and the effects of modal budget constraints. *Res. Transport. Econ.* 9, 159–177.
- Shoup, D., 2005. *The High Cost of Free Parking*. Planners Press, Chicago.
- Small, K.A., Verhoef, E.T., 2007. *The Economics of Urban Transportation*. Routledge.
- Tikoudis, I., Verhoef, E.T., van Ommeren, J.N., 2018. Second-best urban tolls in a monocentric city with housing market regulations. *Transport. Res. Part B: Methodol.* 117, 342–359.
- Van Dender, K., 2003. Transport taxes with multiple trip purposes. *Scand. J. Econ.* 105, 295–310.
- Van Essen, H., B. Boon, A. Schrotten, M. Otten, M. Maibach, C. Schreyer, C. Doll, P. Jochem, M. Bak, and B. Pawlowska. (2008). Internalization measures and policy for the external cost of transport. Technical report, 2008.
- Verhoef, E.T., 2005. Second-best congestion pricing schemes in the monocentric city. *J. Urban Econ.* 58 (3), 367–388.
- Verhoef, E.T., 1994. External effects and social costs of transport. *Transport. Res. Part A: Policy Pract.* 28 (4), 273–287.
- Zhang, W., Kockelman, K.M., 2016. Congestion pricing effects on firm and household location choices in monocentric and polycentric cities. *Reg. Sci. Urban Econ.* 58, 1–12.

Value of Time

John J. Bates, Independent Consultant in Transport Economics, Abingdon, Oxfordshire, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	67
Microeconomic Theory—The Neoclassical Approach	67
Empirical Methods for Measuring Values of Time	69
Key Findings	70
Further Reading	71

Introduction

The value of time, more strictly the value of changes in travel time (VCTT), is a key economic concept in transport modeling and appraisal. Since many transport schemes and policies impact on travel time, the trade-off between travel time and money is of general relevance. The earliest treatment focused on “generalized cost,” which was defined as a linear combination of travel time and cost. However, while this is a convenient simplification and remains widely used, it is better seen as a reflection of the “indirect utility,” which is a feature of discrete choice models, and this has led to a more complex variation in the value of time.

While some exceptional forms of travel time may be seen as “pure leisure,” in general travel time is seen as conveying disutility, so that travelers will be willing to spend money to reduce it. In an indirect utility formulation, both travel time t and travel expenditure c will make negative contributions, and the value of time may be defined as the ratio of the partial derivatives of utility U to time and cost, so that $VCTT = \frac{\partial U / \partial t}{\partial U / \partial c}$. This formulation suggests that VCTT can vary with characteristics both of the journey (e.g., comfort) and of the individual (e.g., income).

Because of the importance of VCTT to transport appraisal, many countries have carried out value of time studies in order to inform “official” values. The earlier studies relied on revealed preference data, and required the researchers to find actual situations where travelers could be observed to trade between a faster, more expensive journey and a slower cheaper journey: these might be found with tolled bridge crossings, or express public transport services. In practice, the statistical requirements for successful studies of this kind are quite exacting, and most of these studies focused on the journey to work.

The development of stated preference (SP) techniques during the 1980s allowed the estimation of VCTT to be based on choices made between hypothetical alternatives, and, although there remain some reservations about the use of such data, all the most recent national studies have relied on SP. This has also allowed for much richer datasets and more detailed analysis of the variations in VCTT.

Microeconomic Theory—The Neoclassical Approach

The theoretical basis for time valuation is based on an expansion of the standard microeconomic demand analysis to include time as well as commodity prices. In this analysis, the individual’s total utility is expressed as a function both of quantity consumed and time expended, and this is maximized subject to various constraints, including a money budget and a time budget.

While a complete theory of time allocation would embrace all aspects of human behavior and as such would be quite unmanageable, it is necessary to make drastic simplifications, and the nature of these simplifications will affect the conclusions. Assume that an individual’s utility is a function of a vector of commodities $\mathbf{x}$, plus a vector of time spent in various activities, $\mathbf{t}$. One of these activities is assumed to be work, which for convenience is distinguished as t_w . This implies $U(\mathbf{x}, \mathbf{t}, t_w)$.

Various constraints are then introduced into the maximization problem. As budget constraints, the total expenditure ($\mathbf{p} \cdot \mathbf{x}$) cannot exceed the available income, which can be represented as $w \cdot t_w + \gamma$ where w is the wage rate (here assumed constant regardless of time worked, and, by implication, net of tax) and γ is the amount of income available from nonwork sources. In addition, the total time spent in activities cannot exceed the total time available, T . It would be possible to exclude from T certain essential requirements (e.g., sleeping, essential eating, etc.); note that the choice of T (24 h, say) implies certain units: for example, the amount of the income from other sources, γ , must be commensurate with the period assumed.

The standard assumption—that working hours are infinitely flexible, and can be freely varied by the consumer—is replaced by a slightly more realistic constraint—that it is necessary to work a minimum number of hours, t_w^* (again commensurate with the assumptions about T). Further, allowance is made for the fact that for certain activities (of which traveling to work is a good example), individuals are compelled to spend more time than they would ideally wish. Given the choice of period T , it is therefore assumed that each element t_i in the time vector has associated with it a minimum t_i^* (which may of course be zero); if required, these t_i^* could be taken as functions of further properties of the system, etc. Here, they act as exogenous constraints in exactly the same way as the minimum working hours hypothesis.

The model can thus be described as a maximization problem subject to a number of constraints, where, for clarification, the associated Lagrangian multiplier is included in square brackets after each constraint.

$$\begin{aligned} & \text{Max } U(\mathbf{x}, \mathbf{t}, t_w) \\ & \text{subject to} \\ & w \cdot t_w + \gamma \geq \mathbf{p} \cdot \mathbf{x} \quad [\lambda] \\ & T \geq \sum_i t_i + t_w \quad [\mu] \\ & t_w \geq t_w^* \quad [\phi] \\ & t_i \geq t_i^* \quad [\psi_i] \forall i \end{aligned}$$

An important distinction can be drawn between those activities for which the minimum time requirements are binding and those for which they are not. In the latter case, individuals are freely willing to commit more time to these activities than is strictly required. These activities can be referred to as “pure leisure activities,” with those activities where the time constraints do bind as “intermediate activities.” It is clear that with this definition, most types of traveling will be an intermediate activity. However, as we shall see, even in the case where travel is viewed as “pure leisure,” it will not lead to a negative value of time.

Formulating the Lagrangian as:

$$L = U(\mathbf{x}, \mathbf{t}, t_w) + \lambda \cdot (w \cdot t_w + \gamma - \mathbf{p} \cdot \mathbf{x}) + \mu \cdot \left(T - \sum_i t_i - t_w \right) + \phi \cdot (t_w - t_w^*) + \sum_i \psi_i \cdot (t_i - t_i^*)$$

the first order conditions for a maximum are obtained by differentiating with respect to $\mathbf{x}$, $\mathbf{t}$, and t_w (as well as the Lagrangian multipliers, which deliver the constraints), giving:

$$\begin{aligned} \partial U / \partial x_i - \lambda \cdot p_i &= 0 \\ \partial U / \partial t_j - \mu + \psi_j &= 0 \\ \partial U / \partial t_w + \lambda \cdot w - \mu + \phi &= 0 \end{aligned}$$

Defining the “marginal valuation of time spent in activity j ” as the ratio of the marginal utility of time in activity j to the marginal utility of income (λ), the second equation gives:

$$(\partial U / \partial t_j) / \lambda = \mu / \lambda - \psi_j / \lambda$$

When $\psi_j = 0$, because the time constraint does not bind, the marginal valuation of time in activity j is equal to μ / λ , the “resource value of time,” representing the consumer’s willingness to pay to have the total time budget increased (though a relaxation of the time budget constraint is not, of course, feasible in reality). Thus, for all “pure leisure” activities, the marginal valuation of time is equal to the resource value, at the optimum. This could either be because consumers are genuinely indifferent as to which leisure activity they are partaking of, implying a **constant** marginal valuation of time, or because they have rearranged their allocation of time to leisure activities so that the marginal valuations are all equal. In this latter case, there may be physical constraints (in space and time) or indivisibilities that impede such a rearrangement, so that the full value of leisure time cannot be realized. This “constrained transferability of time” is a difficult element to incorporate formally within the model, but it must be borne in mind.

While “pure leisure” time has a value (i.e., the “resource” value), in that utility is derived from it, there is no value, **at the margin**, to a **saving** in leisure time. Any time saved in one leisure activity can only be used in another leisure activity, and will have the same valuation. Thus, the consumer will not be prepared to pay to save (pure) leisure time, since he cannot increase his utility by so doing. This argument carries over, of course, to those types of travel that may be considered “pure leisure.”

For intermediate activities, however, the marginal valuation of time in the activity is less than the resource value, and indeed for most kinds of traveling will be negative. The difference between the marginal valuation of time spent in an intermediate activity i and the resource value is ψ_i / λ . By reducing the amount of time spent in activity i and transferring it to leisure, it is possible to increase utility by a unit amount equal to the difference between the marginal valuations of time spent in the activity and time spent in leisure. Hence, ψ_i / λ represents the value of saving time in activity i and transferring it to leisure, and it is this concept which is conventionally referred to as the “value of time” in transport appraisals.

That the marginal valuation of a certain kind of traveling time may be negative reflects directly that such time incurs disutility. But this is not a precondition for time savings having a value. Travel in certain conditions may be sufficiently comfortable that additional time is not directly viewed as a disbenefit. However, because of the overall constraint on time, the traveler can attain a greater total utility by transferring time from the travel activity to leisure. The only circumstance in which this will not be the case is where the travel itself is viewed as pure leisure (perhaps a Sunday drive in scenic surroundings). In this case, the constraint does not bind, so that ψ_i is zero.

Hence, the empirical interest is centered on the values of ψ_i / λ , and in terms of the theory set out here, these values are never negative, and are nonzero every time the consumer is forced to spend more time in an activity than he would ideally wish. The earlier equation can be rewritten to derive the fundamental property of time value:

Value of saving time in activity i (ψ_i / λ) = Resource value of time (μ / λ) – Marginal valuation of time spent in activity i ($(\partial U / \partial t_i) / \lambda$)

This implies that the VCTT could vary because of (1) the income of the individual (λ), (2) the extent to which the individual is time constrained (μ), and (3) the (marginal) utility of the time spent traveling ($\partial U / \partial t_i$), which will be affected by factors such as

comfort, and the opportunity to undertake other activities. In most transport problems, the marginal valuation of time is expected to be negative, because travel time contributes to disutility. However, recent technological developments (mobile phones, etc.) can be considered to have an important impact in reducing this disutility.

While this remains the generally accepted theory, it can be argued that it still lacks two other dimensions—possible variation in goods consumption through substitution of travel for other activities and the possibility of retiming activities (to deal with what was described earlier as the “constrained transferability of time”).

Further, while the economic theory outlined is strictly neoclassical in nature, there are further extensions that owe more to prospect theory and in particular the concept of “reference dependence.” This in turn leads to the phenomenon of “loss aversion” (essentially, a discontinuity in the derivative around the current “reference point”), as is discussed later.

Empirical Methods for Measuring Values of Time

In terms of empirical research, in the simplest case, respondents are asked to choose between two alternative journeys (A and B) with different costs C and times T . To avoid a “dominated” choice, the data should be such that if $C_A > C_B$, then $T_A < T_B$, and vice versa. The choice implies a “boundary value of time” (BVoT) = $-\frac{C_A - C_B}{T_A - T_B}$. If a respondent chooses the more expensive option, this implies that their VCTT > BVoT, and conversely. While there may be more than two choices, and/or the choices may be defined on more variables, the essential principles remain the same. The data are usually analyzed by discrete choice methods, and treated as a panel, since typically each respondent faces multiple choices.

There are two key issues relating to the analysis: the error specification and the utility specification. In relation to the error specification, illustrating the issue with the binary choice example just given, a conventional treatment would assume:

$$U_A = \lambda \cdot C_A + \psi \cdot T_A + \varepsilon_A, \quad U_B = \lambda \cdot C_B + \psi \cdot T_B + \varepsilon_B$$

for the conditional indirect utilities of the alternatives, where ε has the extreme value Type 1 (Gumbel) distribution, and VCTT = ψ/λ . This can also be rearranged in “difference form” as:

$$U_{A-B} = \lambda \cdot [(C_A - C_B) + \text{VCTT} \cdot (T_A - T_B)] + \eta, \quad U_0 = 0$$

where η has the logistic distribution (equivalent to the difference of two identically and independent Gumbel distributions).

However, it is also possible for the error term to be applied “multiplicatively” (or additive in the logarithms). Taking logs, after some rearrangement, we obtain:

$$U_{A-B} = \lambda' \cdot \ln \left[\left(-\frac{C_A - C_B}{T_A - T_B} \right) / \text{VCTT} \right] + \eta = \lambda' \cdot \ln [\text{BVoT}/\text{VCTT}] + \eta, \quad U_0 = 0$$

Some empirical work has found that this latter form provides a much better fit to the data.

In both cases, VCTT can be estimated as a function of appropriate covariates, as well as allowing it to have a random distribution.

The other notable complication for the analysis is the allowance for possible “sign and size effects,” which are present when the alternative choices are based on a status quo position. Respondents may value a reduction in travel time less than an increase (in line with prospect theory), and they may also have a lower unit value for small changes, which may not be seen as useful (or perceptible). Failure to allow for these “reference-dependent” effects could lead to biases and results which are overly dependent on the design of the SP.

One possible approach is to introduce nonlinear functions that allow for the possibility that size and sign effects exist, by defining a function for the *value* of a change Δx relative to the reference value x_0 of a given attribute. Since size and sign effects should be expected for both time and cost attributes, the impact on VCTT is not readily predictable. A suggested formulation for the value function is:

$$v(\Delta x) = \text{Sgn}(\Delta x) \cdot \exp(\omega \cdot \text{Sgn}(\Delta x)) \cdot |\Delta x|^\alpha$$

where

$$\begin{aligned} \Delta x &= x - x_0 \\ \alpha &= 1 - \beta - \gamma \cdot \text{Sgn}(\Delta x) \end{aligned}$$

$\text{Sgn}(\Delta x)$ is the sign function, defined for $\Delta x \neq 0$ by $\text{Sgn}(\Delta x) = \Delta x/|\Delta x|$.

ω measures the sign effect, giving the difference of gain value and loss value from an “underlying” value. It is expected that $\omega > 0$, so that the value of losses (increases in Δx) is greater than the value of gains.

β is the main measure of the size effect, allowing the impact of gains and losses to be nonlinear. If $\beta > 0$, the marginal value of changes decreases as the change increases, that is, the value is “damped” (noting the minus sign in the formula for α earlier).

γ allows for an interaction between the sign and size effects. For example, a negative value for γ (again noting the minus sign in the formula for α earlier) would mean that any damping (due to $\beta > 0$) would be smaller for increases (i.e., losses) than for decreases (i.e., gains) from the reference value.

The arguments of the value functions need to be defined in consistent units. While this is an arbitrary choice, it is sensible—given the interest in VCTT—to choose units of money, so that, if θ is the “underlying” value of time, the value of a cost change ΔC is given by $v(\Delta C)$, while the value of a time change ΔT is given by $v(\theta \cdot \Delta T)$. The resulting values calculated for v are in arbitrary units.

The value functions are assumed to be of the same general form for time and cost (and potentially for other utility components) but the parameters ω , β , and γ are specific to each utility component.

The presence of the sign function $\text{Sgn}(\Delta x)$ in the value formula introduces a discontinuity at the reference value, so that it is not appropriate to obtain VCTT from strictly marginal valuations, as would be found by differentiation. It is also a potential obstacle to its use in appraisal, as will be discussed later. Only if the parameters η and γ are zero for both time and cost, thus eliminating the discontinuity, can the formula $\text{VCTT} = \frac{\partial U}{\partial T} / \frac{\partial U}{\partial C}$ be used.

Instead, VCTT can be derived by thinking of the values of ΔC and ΔT that would maintain indifference with the base situation, where $\Delta C = \Delta T = 0$ and the total value $v(\Delta C) + v(\theta \cdot \Delta T)$ is zero. Thus, given a specific value $\Delta T'$, and the estimated parameters of the value functions v , the requirement is to find the value $\Delta C'$ such that $v(\Delta C') + v(\theta \cdot \Delta T') = 0$. The average willingness to pay per unit of time is then $\Delta C' / \Delta T'$.

In appraisal, and indeed other practical applications of VCTT, we will generally require a “reference free” value. By taking an average of the gain value and the loss value to express an “underlying” VCTT, the formulae can be simplified. For example, consider the geometric mean of $v(\Delta x)$ and $-v(-\Delta x)$:

$$\begin{aligned} & \sqrt{v(\Delta x) \cdot [-v(-\Delta x)]} \\ &= \sqrt{\left[\text{Sgn}(\Delta x) \cdot \exp(\omega \cdot \text{Sgn}(\Delta x)) \cdot |\Delta x|^{1-\beta-\gamma \cdot \text{Sgn}(\Delta x)} \right] \cdot \left[-\text{Sgn}(-\Delta x) \cdot \exp(\omega \cdot \text{Sgn}(-\Delta x)) \cdot |-\Delta x|^{1-\beta-\gamma \cdot \text{Sgn}(-\Delta x)} \right]} \\ &= \sqrt{\left[\exp(\omega \cdot \text{Sgn}(\Delta x)) \cdot |\Delta x|^{1-\beta-\gamma \cdot \text{Sgn}(\Delta x)} \right] \cdot \left[\exp(\omega \cdot \text{Sgn}(-\Delta x)) \cdot |-\Delta x|^{1-\beta-\gamma \cdot \text{Sgn}(-\Delta x)} \right]} \\ &= \sqrt{\exp(\omega \cdot [\text{Sgn}(\Delta x) + \text{Sgn}(-\Delta x)]) \cdot |\Delta x|^{1-\beta-\gamma \cdot \text{Sgn}(\Delta x) + 1-\beta-\gamma \cdot \text{Sgn}(-\Delta x)}} = |\Delta x|^{1-\beta} \end{aligned}$$

This gives a value which omits the asymmetry parameters ω and γ . However, there is no analogous argument by which β can be eliminated, so that a “size effect” may remain, with value being a function of Δx .

Using these simplified value functions $v(\Delta x) = \text{Sgn}(\Delta x) |\Delta x|^{1-\beta}$, the equation $v(\Delta C') + v(\theta \cdot \Delta T') = 0$ can be readily solved. For example, with negative $\Delta T'$ thus implying positive $\Delta C'$, the equation becomes:

$$(\Delta C')^{1-\beta_c} - (\theta \cdot |\Delta T'|)^{1-\beta_t} = 0$$

where

$$\Delta C' = (\theta \cdot |\Delta T'|)^{\frac{1-\beta_t}{1-\beta_c}}$$

This then yields $\text{VCTT} = \Delta C' / \Delta T' = (\theta)^{\frac{1-\beta_t}{1-\beta_c}} \cdot (|\Delta T'|)^{\frac{\beta_c - \beta_t}{1-\beta_c}}$. The same value is obtained if the signs of $\Delta T'$ and $\Delta C'$ are reversed.

If there is no size effect ($\beta = 0$) or the size effect is the same for both cost and time ($\beta_c = \beta_t$), then VCTT is independent of $\Delta T'$ and equal to the “underlying value” θ . However, in general the β values will not be zero and, given the expectation that β should be larger for cost than for time, VCTT will increase as the changes increase. This means that a decision is required for application in terms of what value of $\Delta T'$ is appropriate.

Key Findings

Empirical analysis has established certain key results. There is general agreement that VCTT increases with income, as would be predicted by the influence of λ , the marginal utility of income. There is also considerable evidence that VCTT increases with journey length, though the reasons for this are less clear: there is some empirical evidence that VCTT tends to increase with the journey cost, and decrease with the journey time (note that these are the opposite effects that might be expected on grounds of time and money budgets). Of course, both cost and time are expected to be positively associated with distance: the resulting association of VCTT with distance seems to be brought about by the cost effect being slightly larger than the time effect.

With regard to journey purpose, business journeys tend to have the highest VCTT (though in terms of application such values are often based directly on the wage rate), followed by commuting and then other leisure journeys. While a priori one might expect variation by travel mode to reflect different levels of comfort, conflicting results have been found. For a given mode, higher values tend to be found in conditions of congestion or crowding, but VCTT for the bus mode is typically lower than for rail, whereas on comfort grounds one would expect the opposite. It has been posited that this is due to income effects, but explicit attempts to take account of income do not affect the results materially. There may therefore be some “self-selection” issues, whereby individuals with low VCTT gravitate toward slower, cheaper modes and vice versa.

Individual studies have identified a number of plausible covariate effects, but there is no strong consistency between them. For example, VCTT for car users can be influenced by the number of passengers, but different datasets can suggest different directions.

Of course, pleasant passengers could be expected to reduce VCTT, while annoying passengers would increase VCTT. The “quality” of the passengers will not usually be known to the analyst.

There has also been recent interest in whether the VCTT associated with autonomous vehicles might be lower than that for conventional cars. There is some suggestion that this is the case, but the results need to be seen in the context of the difficulty in describing the experience of an autonomous vehicle to a respondent within an SP exercise.

In any event, after testing for the effect of all reasonable covariates, there remains a considerable random (unexplained) element in the VCTT distribution. Typically, a positively skewed distribution is found (such as the lognormal), suggesting some concentration in the lower values.

In application, a decision needs to be made as to how many of these variations should be included. In practice, relatively aggregate values are likely to be used (especially with regard to official recommendations for transport appraisal), and such values, being averages over a number of covariates, should be based on a representative sample of journeys (as might be obtained from a national travel survey).

Further Reading

Arup, ITS Leeds, Accent, 2015. Provision of market research for value of time savings and reliability. Phase 2 Report to the Department for Transport. Available from: <https://www.gov.uk/government/uploads/>.

Becker, G., 1965. A theory of the allocation of time. *Econ. J.* 75 (299), 493–517.

De Borger, B., Fosgerau, M., 2008. The trade-off between money and travel time: a test of the theory of reference-dependent preferences. *J. Urban Econ.* 64 (1), 101–115.

De Serpa, A., 1971. A theory of the economics of time. *Econ. J.* 81 (324), 828–846.

Evans, A., 1972. On the theory of the valuation and allocation of time. *Scott. J. Polit. Econ.* 19 (1), 1–17.

Jara-Díaz, S.R., 2008. Allocation and valuation of travel time savings. In: *Handbook of Transport Modelling*, second ed. Emerald Group Publishing Limited, Bingley, pp. 363–379.

MVA, TSU, ITS, 1987. *The Value of Travel Time Savings*. Policy Journals, Newbury.

Valuation of Carbon Emissions

Svante Mandell, Swedish National Institute of Economic Research, Stockholm, Sweden

© 2021 Elsevier Ltd. All rights reserved.

References

76

The monetary value to assign to carbon emissions is an important question for many policy decisions. The focus of this chapter will be on the value to use in cost-benefit analyses (CBA) for transport infrastructure investments. We address a situation in which there is a target for carbon emissions now or in the future. This target may cover the transport sector only or a larger sector of which transportation is a part. The main message of this chapter is that the value to assign to carbon emissions in these CBAs should be derived from the target. However, the value will also be affected by which policy instruments are in place in order to reach the target. A number of caveats and complications will also be addressed.

A usual way to determine a value of a nonmarket effect to use in a CBA is to estimate the marginal damage the activity causes. For instance, in order to estimate the value of noise, one approach would be to use the price of sold houses and apply a hedonic approach to isolate the effect of noise exposure, see [Wilhelmsson \(2000\)](#) or [Andersson et al. \(2010\)](#). The corresponding approach for a value of carbon would be to calculate the damage associated with emitting one extra unit of carbon. Calculating this so-called social cost of carbon is a rather complex task, not least because it is the concentration of carbon dioxide (CO₂) in the atmosphere that matters for the greenhouse effect and CO₂ will remain in the atmosphere for a long time ([Bell and Callan, 2011](#); [Farmer et al., 2015](#); [Nordhaus, 2014, 2017](#)). Thus, the future trajectory of emissions must first be established and then the costs associated with shifting this trajectory up by one marginal unit can be calculated. Several studies have attempted to calculate the social cost of carbon and have ended up with quite a wide range of estimates ([Tol, 2009, 2019](#)).

Even if a good estimate of the social cost of carbon existed, I claim that it would have little relevance to the purposes discussed here. This is because once a climate policy is in place, infrastructure investment should be aligned with this policy. After all, the role of a CBA is to provide well-founded guidance for infrastructure investments. If there is a quantitative target for greenhouse gas emissions, the CBA must take into consideration the restrictions imposed by the target. The target will create an implicit price for CO₂ emissions, a shadow price, and this price is what should be reflected in the CBA, see [Mandell \(2011, 2013\)](#) and [Isacs et al. \(2016\)](#). We will return to this below.

I will use the Swedish case as a base for the discussion. Sweden serves as an interesting example as it has one of the world's most ambitious climate targets and has a relatively long history of climate policy. Even if Sweden's climate policy differs from many other countries, the reason for not adopting the social cost of carbon applies almost universally as basically all countries have ratified the Paris Agreement. Thus, they must decide on Nationally Determined Contributions (NDCs). Therein, each country specifies how it intends to contribute to the overarching goal of the Paris Agreement—to keep temperature rise well below 2°C with a goal of 1.5°C. NDCs come in different forms, see [Denison et al. \(2019\)](#) and King and van den Bergh (2019). Roughly one-third of the countries set economy-wide reduction targets in relation to a specific base year. Almost one half set relative targets to reduce emissions below business-as-usual levels. A few countries set intensity targets that relate greenhouse gas (GHG) reductions to gross domestic product. Around one-fifth of the countries have included strategies, plans and actions for low GHG emissions development ([UNFCCC, 2016](#)). How the NDC is formulated matters, but basically, once a country has decided on a target—absolute or relative—the policy challenge it faces is how to reach that target.

Sweden implemented a CO₂ tax in 1991, making it one of the first countries to do so (Finland and the Netherlands introduced CO₂ taxes in 1990). Presently, around 30 other countries have CO₂ taxes in place or are planning to introduce them soon ([World Bank, 2019](#)). Currently, Sweden has the world's highest CO₂ tax. When it was introduced in 1991 it corresponded to approximately EUR 25 per ton (fossil) CO₂. In 2019, the CO₂ tax amounted to EUR 118 per ton (fossil) CO₂. In real terms, this corresponds to an increase just above a factor three since the introduction of the tax.

The main rule in Sweden is that CO₂ tax is levied on all activity that is not subject to the EU's Emissions Trading Scheme, EU ETS, even if various forms of tax deductions apply. As the discussion here is limited to CBAs for transport infrastructure investments, our focus is on the transport sector. Apart from the CO₂ tax, there is also an energy tax and VAT levied on fuel. The CO₂ tax is proportional to the fossil fuel content of the fuel in question. However, the energy tax is not proportional to the energy content. This is understandable as the stated purpose of the CO₂ tax is to handle CO₂ emissions whereas the stated purpose of the energy tax is broader. Initially, its purpose was mainly fiscal but over time it has become increasingly viewed as the primary tool for internalizing externalities not handled by other instruments.

The CO₂ tax and the energy tax increase annually with the consumer price index and an additional 2% so as not to erode the pressure of the taxes as the economy grows and people get richer. In addition to the CO₂ tax and the energy tax is a series of other policy instruments which, in various ways, influence CO₂ emissions from the transport sector. These include a CO₂-differentiated vehicle tax, subsidies for low-emission cars, etc. Nevertheless, the main instrument has arguably been the CO₂ tax on fuel. As indicated below, recent reforms have changed this.

In Sweden, the values to use in CBAs for CO₂ emissions, as well as all other externalities from transport activities that are handled in the analyses, are based on recommendations from a working group known as ASEK. The recommendations are updated annually although major revisions occur at around 3- to 4-year intervals.

To date, six versions of ASEK recommendations have been published. In several of the recommendations, the CO₂ tax is used as a proxy for the CO₂ value. The stated rationale is that the CO₂ tax could be viewed as revealing the value politicians place on reducing emissions by one unit, that is, how much politicians believe society would be willing to sacrifice in order to reduce emissions.

The second version of the ASEK recommendations, published in 1998, contains a different argument. At the time, Sweden had implemented an emission target specific to the transport sector stating that until 2010 it should reduce its emissions to the level it showed in 1990. Rather than basing the CO₂ value on the CO₂ tax, ASEK argued that the value should reflect the shadow price resulting from the target. The rationale remains similar: the CO₂ value should be derived from policy and reflect a “political” value, c. f. Sager (2013). However, as the CO₂ tax was probably not high enough for the transport sector to reach the target, ASEK set out to answer the following question: what CO₂ tax would be needed in order to meet the target?

In short, the approach applied to answering this question involved studying three effects of an increased CO₂ tax: less kilometres travelled, changed driving behavior (eco-driving), and a shift to more energy-efficient vehicles. The responsiveness of each effect to changes in the CO₂ tax was estimated along with how significant an impact each effect was likely to have on reducing emissions.

The analyses resulted in the CO₂ value being increased considerably, from approximately EUR 38 per ton CO₂ to around EUR 150. This value was retained in the subsequent recommendations. When it was time to release the fourth version of the recommendations, it was apparent that the target would not be met. However, as no new target had been set, ASEK opted to keep a CO₂ value of EUR 150 per ton. In the fifth version, ASEK reverted to using the CO₂ tax as a valuation and consequently reduced the value from EUR 150 to EUR 108 per ton. Applying a value corresponding to the CO₂ tax is also the recommendation in the most recent sixth revision.

Thus, the current Swedish approach is to use the value of the CO₂ tax as a kind of political revealed preference. This approach clearly has some merits. The value used in the CBA must be aligned with the climate policy—otherwise the CBA recommendations will not correspond with the needs created by the policy which, in turn, may lead to construction choices that are not well founded. In a simple world in which CO₂ tax is the only climate policy instrument and in which the tax is intended to handle climate issues only (i.e., it is not calibrated to also handle other co-benefits), it should serve as a valid CO₂ value in the CBA as long as it is set in such a way that the target is reached.

In Sweden, the CO₂ tax has never been the only climate policy instrument. However, it has arguably been the most important instrument—and the approach of using it as a proxy for the CO₂ value seems justifiable. If nothing else, it has been a transparent strategy.

However, a series of recent reforms to Swedish climate policy has changed this. Of particular importance to our discussion is a new target for CO₂ emissions from the transport sector and the implementation of a CO₂ emissions reduction obligation. The transport sector goal, adopted in 2017, states that by 2030 the Swedish transport sector—not including aviation, which is covered by the EU ETS—should have reduced its CO₂ emissions by 70% compared to 2010. In 2018, Sweden introduced a CO₂ emissions reduction obligation for transport fuels. It feeds biofuels into the fuel mix. This was previously handled by differentiated CO₂ and energy taxes to the extent that taxes on biofuels were kept sufficiently low for them to be economically viable alternatives to fossil fuels. However, this approach did not work well with EU legislation and required exemptions. It was therefore abandoned.

Both these reforms imply that the principle of applying CO₂ as a proxy for a CO₂ value in the CBA has become obsolete, but for slightly different reasons. First, the new transport sector target will probably not be reached given the CO₂ tax, even considering the annual indexation. Second, the reduction obligation changes the functioning of the CO₂ tax. The tax will still influence car usage, vehicle choice, etc., but will no longer influence the fuel mix. This is now handled by the reduction obligation. Thus, the CO₂ tax has lost one of its purposes.

As the previous approach of using CO₂ tax as a proxy for an appropriate CO₂ valuation is no longer applicable, an alternative approach is needed. The approach we will pursue here is something that is in line with what ASEK did in 1998, that is, start with an estimate of the shadow price that results from the target. There are several ways of estimating the shadow price, for example, applying a computable general equilibrium (CGE) model to calculate what CO₂ tax would be required for the transport sector target to be met by 2030. There are a number of technical problems with such an approach but we will not address these here. Rather, let us assume that we have a model capable of estimating what tax would be required to reach the target. Given this, what issues must be clarified before we apply it?

One obvious issue is to clarify the measure we are really interested in. Above, and in line with how ASEK phrases it, we state that the CO₂ tax was applicable as it gives a proxy for a political valuation. This suggests that we are looking for a measure of how many other resources society is willing to give up in order to reduce CO₂ emissions by one unit. This is clearly an interesting figure. However, given the new policy, I would argue that it is not necessarily what should be used in the CBA as a CO₂ value.

To see this, let us take as a fact that a quantitative emissions target is to be fulfilled at a specified future point in time—in a Swedish context, minus 70% by 2030. Now consider an infrastructure project, say a new road, which is projected to increase CO₂ emissions *ceteris paribus* by one unit. The CO₂ value should be a measure of the societal cost of this extra unit.

First, it should be noted that this cost has nothing to do with climate. The target is to be met with or without this project, so CO₂ emissions will remain the same. Consequently, if there is an additional unit of CO₂ emissions stemming from this project, there must be a corresponding reduction by one unit somewhere else or the target will be breached. It is the cost of achieving this one unit reduction somewhere else that should be the CO₂ value, that is, we adopt an opportunity cost approach.

Second, the relevant CO₂ value has nothing to do with the “political valuation” more than in an indirect sense. The emission target is a consequence of political decisions, but once it is in place it is the opportunity cost it creates that is of interest. This distinction is of little concern if the instrument used to reach the target is a CO₂ tax only. However, consider a situation in which two policy instruments are used: a CO₂ tax on fuel and subsidies for energy-efficient vehicles. In order for the CO₂ value to capture the political valuation, it should consider both the CO₂ tax and the subsidies, as both are associated with costs to society. However, if the CO₂ value is to capture the opportunity cost in the sense discussed above, it should only consider the CO₂ tax. If the government subsidizes energy-efficient vehicles, the applicable CO₂ value will be less than if there were no subsidies. This is because subsidies will make it easier and less costly for society to reduce emissions somewhere else in such a way that the target is still met. In this case, the political valuation—which includes the costs associated with subsidizing energy-efficient vehicles—will exceed the relevant opportunity cost—which only includes the variable cost.

In this view, not only do we need to consider the target and the shadow price in which it results. We also need to take into account what policy mix will be used to reach it. This makes the task more complicated. Even so, the outcome is in line with the purpose of the CBA—to help policymakers choose which transport infrastructure to invest in. To illustrate using a crude example, consider two alternative policy packages that reach the same emission target at some future point in time. One package relies on CO₂ taxes only. In order to reach the target, these taxes must be high, thus making it costly to travel by car. Demand for car travel therefore drops. In such a situation, spending resources on, for example, new roads may be less desirable. The other policy package combines CO₂ taxes with large subsidies for energy-efficient vehicles. The CO₂ tax required to reach the target is now lower than without the subsidies and the cost of travelling by car is consequently lower compared to the alternative policy package. In this situation, new roads may be more desirable as demand for road transport will be higher and the negative effects of CO₂ emissions will be less. Thus, in order to provide guidance on what to invest, the CBA—through the CO₂ value—must consider the policy design, not only the target.

There are several other complicating factors. One of these concerns whether it is correct to view the target as given. After all, targets are not always met. As discussed above, Sweden previously had a target for the transport sector that was abandoned. This target resulted in the CO₂ value being increased quite substantially. Thus, for a period, the CBAs calculated on a future that did not emerge, thereby resulting in erroneous recommendations. However, ignoring a politically determined target does not appear to provide a basis for well-founded analyses.

A related issue is that the arguments above rely on the idea that the CO₂ value should be derived from the adopted climate policy. But what if infrastructure investments form part of that policy package? For example, some may consider investments in railways in order to provide substitutes for less climate-friendly alternatives, such as climate policy measures. We may view this in two ways: either we invest in railways to save the climate or, when we try to save the climate, demand for climate-friendly alternatives will increase and we then need to construct more railways. If we adopt the latter view, the above arguments apply. However, if we adopt the former view, the measure we are assessing itself becomes part of the policy package. The basic problem is that if infrastructure investments are considered part of climate policy, constructing a project should entail calibrating other climate policy measures, for example, the CO₂ tax, so that emissions are kept at the target level. One way of reconciling this is to view individual projects as being too small to justify such calibrations as anything but negligible (i.e., an assumption akin to that, in the perfect competition model, no individual firm is large enough to influence price).

Thus far, we have argued that a valid CO₂ value should be based on the target if such a target exists and that we must also consider what policy measures are employed in order to reach the target. As policy design must also be taken into consideration, we must also consider the policy’s purpose. It is often claimed that climate policy is associated with substantial co-benefits, including less congestion and better air quality, as well as more jobs, more competitive businesses, etc. To the extent that such co-benefits exist, we would not want them to end up in a CO₂ value as they concern issues other than CO₂ emissions. The CBA should capture all relevant effects, but the effects should not be attributed to climate if they concern other aspects. Given the opportunity cost approach outlined above, this problem may largely disappear as it relies on the shadow price of the CO₂ target.

Another complication lies in the cost effectiveness of the actual policy mix. The approach here is to start by calculating the shadow price resulting from the target. This corresponds to calculating how large a CO₂ tax must be in order to reach the target, given that no other policy measures are being used. We have then argued that if the actual (anticipated) policy mix contains instruments that influence the opportunity cost, these should be taken into consideration. The problem is that the least costly way to reach a CO₂ emissions target is to assign a uniform price for emissions, for example, by using a uniform CO₂ tax. Thus, a policy mix that also includes other measures will increase costs. In principle, this should influence the CO₂ value. After all, we have argued that it is the opportunity cost to keep emissions at the target level that matters. If the policy is designed in a nonoptimal way, in such a way that the opportunity cost becomes higher, the CO₂ value should address this. At the same time, it will probably make it more difficult to calculate the shadow price.

The fundamental driver behind the discussion above is that the CO₂ value should take its departure in the target and the policy mix that is used to reach it. If there are several different targets covering different parts of the economy, this may generate counterintuitive outcomes.

The most obvious example is a discrepancy between the value of CO₂ emissions that occur during infrastructure construction versus emissions that occur when the infrastructure is in use. We have argued that the value to associate with the latter follows from the costs associated with having to change behavior somewhere else under the transport sector target in such a way that the target is still met. A similar argument should apply to emissions associated with construction. However, to a large extent, such emissions, for example, emissions from steel and cement production, fall under a different target, given by the EU ETS. Thus, the corresponding logic states that these emissions should be valued at the shadow price generated by the ETS. The allowance price would appear to be a

good candidate. In the Swedish case, this would imply a much lower valuation of CO₂ emissions occurring during the infrastructure construction phase than those emissions that occur from using the same.

This creates something of a problem. Consider two alternative projects: A and B. They are identical in every way except that A is associated with lower CO₂ emissions in the usage phase and higher emissions associated with construction in such a way that total emissions would be higher from project A than project B. Climate is affected in the same way regardless of whether the emissions stem from construction or usage. So, from a climate perspective, project B would appear to be preferable. However, if the CBA applies a higher CO₂ value to emissions that result from usage rather than from construction, the analyses may result in the opposite recommendations. This may seem counterintuitive. Even so, it is in line with the underlying logic that, if we accept the notion that targets should be met, the CO₂ value has nothing to do with climate. It is only driven by how the costs of reaching the targets in question are influenced by the project under study. Thus, total emissions are not influenced by whether project A or B is chosen (as both the emissions from construction and usage are subject to targets). However, the cost of reaching the targets differs between projects and a well-designed CBA should capture these differences.

A similar example considers the electrification of transport. This may take the form of constructing infrastructure in such a way that traffic shifts from road to (electrified) rail transport, or stems from major electrification of the car fleet. In both cases, emissions will most probably be reduced but will also “move” from the ESR sector, which covers emissions from transportation, to the EU ETS sector, which covers emissions from the generation of electricity. The problem is the same. A lower value should be attached to emissions under the EU ETS than those under ESR to reflect the difference in opportunity cost, even though the emissions have the same impact on climate. Again, this is due to both the ETS and the ESR being subject to targets in such a way that the climate is left unaffected. However, the (marginal) cost of reaching the targets differ between sectors, and this should be captured by the CO₂ values in the CBA. It should be noted that a recent reform of the EU ETS implies that emissions from the system are, to some extent, endogenous (Carlén et al., 2019). This complicates the question of how to view the EU ETS target but should not fundamentally change the arguments put forward here.

To summarize, in this chapter I have argued that it is important that the CO₂ value used in CBAs for transport infrastructure captures the societal costs associated with CO₂ emissions. However, if there is a binding emissions target, this cost is not associated with climate (as the target implies that emissions will be the same whether or not the project is subject to the CBA) but rather with how infrastructure construction influences the costs of reaching such a target. This perspective introduces a series of interesting questions and problems, some of which have been discussed above.

References

- Andersson, H., Jonsson, L., Ögren, M., 2010. Property prices and exposure to multiple noise sources: Hedonic regression with road and railway noise. *Environ. Resour. Econ.* 45 (1), 73–89.
- Bell, R.G., Callan, D., 2011. More than Meets the Eye: The Social Cost of Carbon in US Climate Policy. Environmental Law Institute.
- Carlén, B., Dahlqvist, A., Mandell, S., Marklund, P., 2019. EU ETS emissions under the cancellation mechanism—effects of national measures. *Energy Policy* 129, 816–825.
- Denison, S., Forster, P.M., Smith, C.J., 2019. Guidance on emissions metrics for nationally determined contributions under the Paris Agreement. *Environ. Res. Lett.* 14 (12), 124002.
- Farmer, J.D., Hepburn, C., Mealy, P., Teytelboym, A., 2015. A third wave in the economics of climate change. *Environ. Resour. Econ.* 62 (2), 329–357.
- Isacs, L., Finnveden, G., Dahllöf, L., Håkansson, C., Petersson, L., Steen, B., Swanström, L., Wikström, A., 2016. Choosing a monetary value of greenhouse gases in assessment tools: a comprehensive review. *J. Clean. Prod.* 127, 37–48.
- King, L.C., van den Bergh, J.C., 2019. Normalisation of Paris agreement NDCs to enhance transparency and ambition. *Environ. Res. Lett.* 14 (8), 084008.
- Mandell, S., 2011. Carbon emission values in cost benefit analyses. *Transp. Policy* 18 (6), 888–892.
- Mandell, S., 2013. Carbon emissions and cost benefit analyses. In: *International Transport Forum Discussion Paper, No. 2013-32*, International Transport Forum, Paris.
- Nordhaus, W., 2014. Estimates of the social cost of carbon: concepts and results from the DICE-2013R model and alternative approaches. *J. Assoc. Environ. Resour. Econ.* 1 (1/2), 273–312.
- Nordhaus, W., 2017. Revisiting the social cost of carbon. *Proc. Natl. Acad. Sci.* 114 (7), 1518–1523.
- Sager, T., 2013. The comprehensiveness dilemma of cost-benefit analysis. *Eur. J. Transp. Infrastruct. Res.* 13 (3), 169–183.
- Tol, R.S., 2009. The economic effects of climate change. *J. Econ. Perspect.* 23 (2), 29–51.
- Tol, R.S., 2019. A social cost of carbon for (almost) every country. *Energy Econ.* 83, 555–566.
- UNFCCC. (2016). Aggregate effect of the intended nationally determined contributions: an update, FCCC/CP/2016/2.
- Wilhelmsson, M., 2000. The impact of traffic noise on the values of single-family houses. *J. Environ. Plan. Manag.* 43 (6), 799–815.
- World Bank State and Trends of Carbon Pricing 2019, Washington DC, June 2019

Valuation of Travel Time Variability Using Scheduling Models

Katrine Hjorth, Technical University of Denmark, Kongens Lyngby, Denmark

© 2021 Elsevier Ltd. All rights reserved.

The Concept of Travel Time Variability	76
Microeconomic Foundation: A Simple Scheduling Model	77
The General Framework	77
The Step Model	78
The Slope Model	79
Other Specifications	80
Extensions of the Simple Model	80
Imperfect Information or Limited Rationality	80
Scheduled Services	80
Trip Chains with Multiple Activities	80
Time-Varying or Endogenous Travel Time Distributions	80
Empirical Evidence About Preferences	81
Stated Preference and Revealed Preference Data	81
The Scheduling and Reduced-Form Approaches	81
Application in Demand Prediction and Cost-Benefit Analysis	82
See Also	82
References	82
Further Reading	83

The Concept of Travel Time Variability

The travel time for a journey from A to B with a given transport mode can vary for many different reasons. Some of this variation can be anticipated by the travelers: The travel time may be longer on a Monday morning during the peak hours than on a Sunday evening, due to different levels of congestion in the network. However, some variation is unpredictable for the traveler, for instance the day-to-day variation for a trip carried out every weekday at the same time of day under seemingly identical circumstances. We refer to this unpredictable variation as travel time variability (TTV).

TTV may arise due to fluctuations in traffic demand from day to day, which are random from the travelers' point of view, or due to traffic incidents, transitory roadworks or unanticipated harsh weather conditions affecting travel speeds and road capacity. TTV is costly for the travelers, since the unpredictability causes people to either risk arriving later than desired, or to depart sufficiently early to allow for a safety margin to avoid being late, potentially wasting their time. As a cost for the travelers, TTV should in principle be accounted for, both when predicting demand in transport models and when evaluating transport and infrastructure projects in cost-benefit analyses, in order to incorporate effects of changed TTV along with effects of changes in average travel times. This requires that we know the following:

1. how to appropriately measure TTV (i.e., should we use the standard deviation of travel time, the mean delay or another measure of variability to quantify the extent of TTV?);
2. how TTV affects behavior in terms of departure time, mode, and route choice;
3. the monetary value of TTV; and
4. how to predict TTV for various scenarios.

The topic of this chapter is the theoretical foundation for (1)–(3). Two approaches coexist in the transport economic literature: One is the scheduling approach, assuming travelers derive utility from spending time in certain activities at certain times of day, and that TTV affects their expected utility through its impact on the arrival times at the activities. The other is the direct approach, which ignores the scheduling aspect and simply assumes that travelers incur a cost from TTV, whether it may stem from scheduling inconvenience or anxiety. The early literature dates back to the 1960s, and particularly the scheduling approach has developed over time with several extensions and adaptations. We do not provide a review of this literature here, but refer to the list of further readings, which contains a comprehensive review of the use of the two approaches (Carrion and Levinson, 2012), as well as research articles with recent methodological advances. For a review of the literature related to forecasting TTV for appraisal (4), we refer to de Jong and Bliemer (2015).

Here, we will present a simple scheduling model as the basic microeconomic framework for valuing TTV, based on Fosgerau and Karlström (2010), Fosgerau and Engelson (2011), and Fosgerau (2016).

Microeconomic Foundation: A Simple Scheduling Model

The General Framework

A scheduling model describes travelers' preferences for being at different locations at a given time, and for transporting themselves at a given time. Most scheduling models in the literature are partial models that only consider the problem of choosing the optimal departure time on a given journey, and take as given the travelers' preferences for being at the origin and destination. It is assumed that the departure time can be chosen continuously. Moreover, many models are pure demand-side models, taking the distribution of travel time as given and thus ignoring that travelers' choices of departure times affect the distribution. Finally, it is common to consider the mode and route choice as given, but in principle the model can be extended to allow several modes and routes.

Our simple model embodies all these restrictions. Later, we discuss what happens if some of these restrictions are relaxed. Consider a traveler having a specific origin and destination (we can think of them as home and work), between which he desires to travel during a specified period of the day. We assume that the traveler is equipped with scheduling preferences, that is, preferences for being at the origin and destination at a given time, and that he chooses his departure time to maximize his expected utility. Moreover, we assume he considers only a single mode of transport and a single route, with a given (known) distribution of travel times.

We formalize this as follows: Let $[t_{start}, t_{end}]$ be the time interval, during which the traveler wants to travel. Define $h(t)$ as the marginal utility of being at the origin minus the marginal utility of time spent traveling at time t , and define $w(t)$ as the marginal utility of being at the destination minus the marginal utility of time spent traveling at time t . Assume $h(t)$ and $w(t)$ are monotonic and bounded on $[t_{start}, t_{end}]$: $h(t)$ is weakly decreasing and $w(t)$ weakly increasing. With these definitions and assumptions, the integrals $\int_a^b h(t)dt$ and $\int_a^b w(t)dt$ are well defined for $a, b \in [t_{start}, t_{end}]$. $\int_a^b h(t)dt$ is the excess utility from being at the origin in the time period $[a, b]$ compared to being traveling. Similarly, $\int_a^b w(t)dt$ is the excess utility from being at the destination compared to traveling. Assume further, that there is a time $t_0 \in [t_{start}, t_{end}]$ such that $h(t) > w(t)$ for $t < t_0$, and $h(t) < w(t)$ for $t > t_0$. Then the traveler prefers to be at the origin before t_0 and at the destination after t_0 . This idea is expressed graphically in Fig. 1. We also need a technical assumption to ensure the interval $[t_{start}, t_{end}]$ is wide enough to include the relevant departure and arrival times.

The traveler's scheduling utility, U , is assumed to be additively separable in the total utility obtained at the origin and the total utility obtained at the destination. For departure time t_d and a specific realization of travel time T , we have:

$$U(t_d, T) = \int_{t_{start}}^{t_d} h(t)dt + \int_{t_d+T}^{t_{end}} w(t)dt. \quad (1)$$

This is a general model that accommodates a variety of specifications of $h(t)$ and $w(t)$, corresponding to different assumptions about the scheduling preferences. We shall discuss two specific popular specifications in more detail below. Note that the function in Eq. (1) can be normalized in different ways to ease computations, corresponding to scaling or shifting the utility function.

We assume that travel time T is an absolutely continuous random variable with some distribution known by the traveler, and that the traveler chooses the departure time that maximizes his expected utility $EU(t_d) \equiv E[U(t_d, T)]$. We make the following assumptions about the distribution of T :

- the mean and variance of T exist;
- T has bounded support;
- the distribution of T does not depend on the departure time.

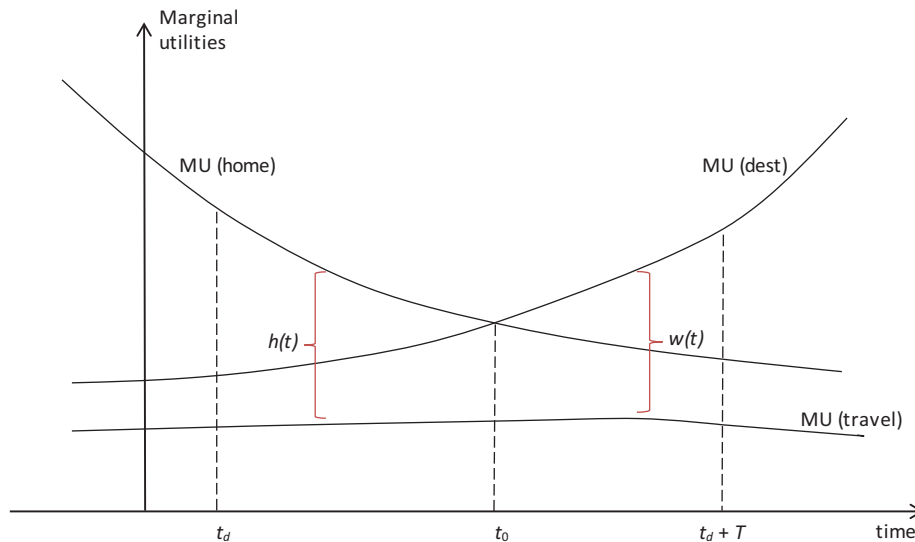


Figure 1 Marginal utilities of time spent at home and at the destination and time spent traveling.

Engelson and Fosgerau (2011) have shown that these assumptions imply that $EU(t_d)$ is concave, finite, and continuous in t_d , and that there exists an optimal departure time which maximizes $EU(t_d)$. Note that they define the traveler's departure time choice in terms of a cost minimization problem, where the cost function is equal to a constant minus $EU(t_d)$.

The assumption that T has bounded support means that we can define an interval $[t_{start}, t_{end}]$ that contains all possible departure and arrival times. This corresponds to reality, where the set of possible travel times is bounded. For modeling purposes, it may be relevant to consider cases where T does not have bounded support, such as the shifted exponential or lognormal distributions. The interested reader may consult Engelson and Fosgerau (2011) about conditions for existence of an optimal departure time in this case.

To formalize the link between scheduling costs and TTV, it is convenient to apply the decomposition $T = \sigma X + \mu$, where μ and σ are the mean and standard deviation of T , and X is a continuous random variable with zero mean and variance equal to 1. X has bounded support, and has a continuous density function ϕ . Moreover, we assume that X has an invertible distribution function Φ , such that the quantiles of the travel time distribution are uniquely defined. It is possible to assume specific distributional forms of X , if one has specific knowledge about this, but this assumption is not necessary, so we proceed without.

The assumption that the distribution of T does not depend on the departure time means both that the distribution of X does not vary over $[t_{start}, t_{end}]$ (which, according to some empirical evidence, may be acceptable as an approximation to reality), and also that μ and σ do not vary over $[t_{start}, t_{end}]$. This latter invariance assumption is restrictive, and almost certainly incorrect if $[t_{start}, t_{end}]$ includes both a congested peak period and the periods immediately before or after. We shall proceed using the invariance assumption because it provides simple and easily interpretable results, and because Fosgerau and Karlström (2010) have demonstrated, in the special case called the step model, c.f. below, that these results still hold approximately if μ and σ are linear functions of time, as long as they do not change too rapidly over time. We return to the implications of the invariance assumption in the third section.

The expected utility can be written as:

$$\begin{aligned} EU(t_d) &= \int_{t_{start}}^{t_d} h(t) dt + E\left(\int_{t_d+T}^{t_{end}} w(t) dt\right) \\ &= \int_{t_{start}}^{t_d} h(t) dt + \int\left(\int_{t_d+\mu+\sigma x}^{t_{end}} w(t) dt\right) \phi(x) dx. \end{aligned} \quad (2)$$

When $h(t)$ and $w(t)$ are continuous, $EU(t_d)$ is continuously differentiable in departure time t_d , and the unique optimal departure time t_d^* is defined by the first-order condition

$$\frac{\partial EU}{\partial t_d} = 0 \Leftrightarrow h(t_d^*) = E(w(t_d^* + T)). \quad (3)$$

For specific simple functional forms of $h(t)$ and $w(t)$, the optimal expected utility $EU^* \equiv EU(t_d^*)$ has a closed form. In that case, we can obtain VTT, the value of travel time (in utility units), as minus the derivative of EU^* with respect to μ , and VTTV, the value of TTV (in utility units), as minus the derivative of EU^* with respect to σ , σ^2 or another measure of the extent of variability in the travel time distribution.

We now take a closer look at two popular scheduling models, the step model and the slope model.

The Step Model

In the step model, $h(t)$ is a constant function equal to α , and $w(t)$ is a piece-wise constant function equal to $\alpha - \beta$ for $t < t_0$ and $\alpha + \gamma$ for $t > t_0$. The parameters α, β, γ are assumed to be positive. This type of preferences is widely used in applications and often referred to as α - β - γ -preferences. In this case, the utility function in Eq. (1) becomes

$$\begin{aligned} U(t_d, T) &= -\alpha T - \beta(t_0 - t_d - T)^+ - \gamma(t_d + T - t_0)^+ + K_1 \\ &= -(\alpha - \beta)T + \beta(t_d - t_0) - (\beta + \gamma)(t_d + T - t_0)^+ + K_1, \end{aligned} \quad (4)$$

where $z^+ \equiv \max(z, 0)$ and K_1 is a constant term which depends on $\alpha, \gamma, t_0, t_{start}$ and t_{end} . For a given value of T , this utility function is piecewise linear in t_d with a unique maximum in $t_d^* = t_0 - T$. Hence, in the case without TTV, where travel time is always equal to μ , the optimal departure time would be $t_0 - \mu$ with corresponding arrival time t_0 . When travel time is random, we assume the traveler maximizes his expected utility:

$$\begin{aligned} EU(t_d) &= -(\alpha - \beta)\mu + \beta(t_d - t_0) \\ &\quad - (\beta + \gamma) \int_{t_0 - t_d - \mu}^{\infty} (t_d + \mu + \sigma x - t_0) \phi(x) dx + K_1 \end{aligned} \quad (5)$$

Note that the integral on the right hand side is differentiable in t_d because the integrand is continuous in x and continuously differentiable in t_d , and the lower limit is continuously differentiable in t_d . The expected utility is twice continuously differentiable and strictly concave in t_d (as the second-order derivative is negative). The optimal departure time satisfies the first order condition $\frac{\partial EU}{\partial t_d} = 0$, which is equivalent to

$$t_d^* = t_0 - \mu - \sigma \Phi^{-1}\left(\frac{\gamma}{\beta + \gamma}\right), \quad (6)$$

where Φ is the distribution function of X . We see that the optimal departure time is linear in μ and σ , and depends on the distribution of X . The traveler departs earlier than in the case with certain travel time, by a safety margin $\sigma\Phi^{-1}[\gamma/(\beta + \gamma)]$. This safety margin is proportional to the standard deviation of travel time, and depends on the shape of the travel time distribution through Φ^{-1} . Naturally, it depends on the preferences for arriving earlier or later than t_0 : A higher value of γ (= a higher disutility of arriving later than t_0) implies an earlier optimal departure time. Departing at t_d^* , the traveller has probability $\frac{\beta}{\beta + \gamma}$ to arrive later than t_0 . Inserting the optimal departure time in Eq. (6) into Eq. (5) yields the following optimal expected utility:

$$\begin{aligned} EU^* &= -\alpha\mu - (\beta + \gamma)\sigma \int_{\Phi^{-1}\left(\frac{\gamma}{\beta + \gamma}\right)}^{\infty} x\phi(x)dx + K_1 \\ &= -\alpha\mu - (\beta + \gamma)\sigma \int_{\frac{\gamma}{\beta + \gamma}}^1 \frac{\gamma}{\beta + \gamma} \Phi^{-1}(s)ds + K_1. \end{aligned} \quad (7)$$

The second equality follows from applying the transformation $s = \Phi(x)$. This expression shows that the cost imposed by TTV depends on σ , the preference parameters, and the shape of the travel time distribution. Given Eq. (7), Fosgerau and Karlström (2010) suggest to measure the extent of TTV by σ , and the value of TTV (in utility units) as

$$VTTV = -\frac{\partial EU^*}{\partial \sigma} = (\beta + \gamma) \int_{\frac{\gamma}{\beta + \gamma}}^1 \Phi^{-1}(s)ds. \quad (8)$$

This value can be converted to a monetary unit by dividing with the marginal utility of income. As noted by Fosgerau (2016), the model also supports other choices of a measure of the TTV: The interquantile range of the travel time distribution, that is, the difference between two specific quantiles T_{p_2} and T_{p_1} , is often applied in the literature as a measure of TTV, and is equal to $IQR = \sigma[\Phi^{-1}(p_2) - \Phi^{-1}(p_1)]$ and hence proportional to σ , since we consider the distribution of X as fixed. Hence, Eq. (7) can be rewritten in terms of IQR , and the value of TTV can be defined in terms of $\partial EU^* / \partial IQR$. The same holds for the difference between a specific quantile and μ .

Another option, suggested by Small (2012), would be to measure TTV by $Q \equiv \sigma \int_{\frac{\gamma}{\beta + \gamma}}^1 \Phi^{-1}(s)ds$. Since the difference between the p -quantile of the travel time distribution and the mean equals $\sigma\Phi^{-1}(p)$, the quantity Q is an average over such differences for all values of p larger than $\gamma/(\beta + \gamma)$. Hence, Q is a generalization of a commonly applied measure of TTV, the difference between a specific quantile and μ . The VTTV corresponding to Q is $-(\partial EU^* / \partial Q) = \beta + \gamma$.

The first measures (σ and IQR) have the benefit that they are easy to compute and interpret; however the associated values depend on Φ and are hence not transferable from one application to another. For the second measure (Q), the opposite applies; and TTV itself depends on the preference parameters, such that it is not transferable from one group of travelers to another.

The Slope Model

In the slope model, $h(t)$ and $w(t)$ are linear functions intersecting at $t = t_0$: $h(t) = a + b(t - t_0)$ and $w(t) = a + g(t - t_0)$, where $b \leq 0$, $g \geq 0$, and $b < g$.

In this model, $h(t)$ and $w(t)$ are twice continuously differentiable, $EU(t_d)$ is strictly concave in departure time t_d , and there exists a unique optimal departure time t_d^* . This optimal departure time satisfies the first-order condition $\frac{\partial EU}{\partial t_d} = 0$, that is,

$$h(t_d^*) = E(w(t_d^* + T)) \Leftrightarrow t_d^* = t_0 - \frac{g}{g - b}\mu. \quad (9)$$

Unlike in the step model, the optimal departure time depends only on the mean travel time, and not on the level of TTV. This is a special property of the slope model, which follows from the linearity of $h(t)$ and $w(t)$. The optimal expected utility is:

$$\begin{aligned} EU^* &= \int_{t_{start}}^{t_0 - \mu \frac{g}{g - b}} \frac{g}{g - b} (a + b(t - t_0))dt + E\left(\int_{t_0 - \mu \frac{g}{g - b} + T}^{t_{end}} \frac{g}{g - b} (a + g(t - t_0))dt\right) \\ &= -a\mu + \frac{bg}{2(g - b)}\mu^2 - \frac{g}{2}\sigma^2 + K_2, \end{aligned} \quad (10)$$

which implies that the cost imposed by TTV is proportional to the variance of travel time, σ^2 , and depends only on σ^2 and the preference parameters. The constant term K_2 depends on t_0 , t_{start} , t_{end} , and on a , b , g , but not on the travel time distribution. Given Eq. (10), Fosgerau and Engelson (2011) suggest it is natural to measure the extent of TTV by σ^2 , and the value of TTV (in utility units) as

$$VTTV = -\frac{\partial EU^*}{\partial \sigma^2} = \frac{g}{2}. \quad (11)$$

Other Specifications

The step model and the slope model have the benefit that they yield closed-form expressions for the optimal expected utility and simple measures and values of TTV. However, it remains to be verified empirically how well the models represent traveler preferences. In principle there are many possible formulations of $h(t)$ and $w(t)$, but apart from the step model and the slope model, only a few have been applied in practice. Models with exponential functions also yield closed form solutions, but appear difficult to apply in practice, since the model parameters can be hard to identify from empirical data ([Hjorth et al., 2015](#)).

Extensions of the Simple Model

In this section, we briefly mention a couple of relevant extensions to the simple model. The interested reader is referred to the original research articles, provided in the literature list.

Imperfect Information or Limited Rationality

We have so far assumed that the traveler knows the travel time distribution and optimizes his expected utility. But how are the results affected if this is not the case? Generally, if the traveler has a wrong perception of the travel time distribution, it leads to suboptimal departure time choices, but the size of the additional cost, in terms of his utility loss, depends both on the nature of the misperception and on the assumed scheduling model.

Misperception could take the form of probability weighting, where the traveler systematically misperceives the probabilities of different travel time outcomes. Systematic probability misperception is consistent with observed behavior in many laboratory experiments regarding other types of random outcomes, and can be formalized using a rank-dependent utility model instead of assuming expected utility optimizing behavior. Applying a rank-dependent utility approach to the scheduling model is one way of deriving the cost of misperception: For both the step model and the slope model, [Xiao and Fukuda \(2015\)](#) have shown formally that this type of misperception implies a higher perceived cost of TTV.

Another form of misperception could be that the traveler forms his expectation about the travel time distribution partly based on a fixed perception of travel time, a so-called anchor value, and partly based on his experience of the last K trips, which he may not always remember or remembers inaccurately. [Koster et al. \(2015\)](#) have shown (for the slope model) that this may lead to suboptimal departure time choices and to a higher perceived value of TTV.

Scheduled Services

The assumption that departure time choice is continuous may at first glance seem to restrict the model to private modes of transport (cars, bikes, etc.). However, the approach can be extended to cover scheduled services as well. [Fosgerau and Karlström \(2010\)](#) have demonstrated that, compared to the basic model, the traveler incurs a larger cost when the service is scheduled, and the size of this additional cost is of course related to the service headway. For some scheduled services, it may be more relevant to consider a model where the in-vehicle travel time is known with certainty by the traveler, but the waiting time at the station is considered random. In this case, [Benezech and Coulombel \(2013\)](#) have shown that one can derive the traveler's value of mean headway and headway variability, along the lines set out in our simple model.

Trip Chains with Multiple Activities

The framework in our simple model considers a single trip, which is suitable if all journeys made by a traveler in a day are scheduled independently of each other. However, it may be the case that the arrival time at work in the morning affects the departure time from work in the afternoon, because the traveler has to work a fixed number of hours. More generally, the choices of all departure times and all activity durations for all trips and activities during the day may be interrelated. [Jenelius \(2012\)](#) has defined a trip chain scheduling model that incorporates this, and shown that it is possible, in some cases, to derive the optimal departure times using dynamic programming. For the special case with a two-trip chain between three activities, analytical values of travel time and TTV can be derived under the assumption of piecewise constant or piecewise linear marginal utilities of activity participation, which corresponds to the preferences in the step model and the slope model, respectively. The values of travel time and TTV differ between the first and the second trip, and also depend on the joint distribution of travel times on the two trips.

Time-Varying or Endogenous Travel Time Distributions

Two important assumptions in the basic scheduling model are that the travel time distribution is exogenous (the traveler takes it as given) and invariant over the period of analysis. In many real congested transport systems, however, the mean and variance of the travel time distribution vary systematically over the day, as a consequence of the departure time pattern stemming from the combined choices of all travelers. It is highly likely that travelers take the variation over time into account when choosing their departure times, which violates the invariance assumption. How well the invariance assumption approximates reality is yet unclear, and this is arguably a weak point of the model, which needs further investigation.

When it comes to measuring preferences and the value of TTV using empirical data, the exogeneity assumption might not be so important—if we are willing to assume that travelers' responses to changes in TTV do not depend on how they think other travelers react. This may be a fair assumption in many cases, such as stated preference studies. However, when it comes to predicting welfare effects of new infrastructure or policy programs, it is necessary to account for second-order effects, that is, that travelers' aggregate behavior in terms of departure time choice affects the travel time distribution. We return to this discussion in the final section.

Empirical Evidence About Preferences

Stated Preference and Revealed Preference Data

The empirical evidence about scheduling preferences under travel time uncertainty stems mainly from stated preference analyses, that is, hypothetical behavior, but there is also some evidence from observations of real behavior. Typically, the stated preference studies infer knowledge about preferences from hypothetical trade-offs between monetary costs, mean travel time, and TTV and/or departure time. In these surveys, TTV is commonly presented to the participants as a list of 2–5 possible travel times or arrival times with corresponding probabilities, or graphical representations of this. Stated preference analyses have two major benefits: First, the data are relatively cheap because surveys can be conducted over the Internet, providing large samples at low costs. Second, because the trade-offs measured in the surveys are hypothetical, it is possible to design the surveys to obtain sufficient variation in the attributes and avoid co-linearity, which is a prerequisite for statistical identification of the effects. An often mentioned disadvantage of stated preference analyses is of course that we can never be certain that respondents answer truthfully: they may have reasons for not doing so. However, in the context of scheduling preferences and TTV, it is probably just as important to consider the “observation error” stemming from respondents misinterpreting the often considerable amount of information in a choice exercise, or not making the effort to fully comprehend the consequences of the information in terms of their travel planning and activity schedule.

The revealed preference studies typically use data of route choice and experienced travel times, from which it is sometimes possible to estimate trade-offs between mean travel time and TTV, and in some cases also departure times and monetary costs. Compared to stated preference studies, the revealed preference analyses may suffer from insufficient variation in the attributes, or co-linearity of the attributes, because real-life observations of mean travel time, TTV and cost for different routes between a given origin and destination will often be highly correlated. The consequence is less precise statistical results. Until recently, data collection was also a problematic issue: The studies relied on good quality travel time data with a sufficiently fine time resolution for all relevant routes, and corresponding observation of choices and choice sets. Some studies relied on GPS-transmitters being installed in the vehicles of the participants, which was relatively expensive and usually only applied with small samples. These days, information about route choices, departure times, and travel times for very large samples of travelers has become easily available—with the potential to bring about new evidence about scheduling preferences from revealed preference studies in the near future. The issue regarding colinearity between the attributes, however, still prevails.

The Scheduling and Reduced-Form Approaches

Within the theoretical framework of our basic model, the empirical value of variability can be obtained in two different ways: One is called the scheduling approach and the other the direct or reduced-form approach. With the scheduling approach, one uses observations of trade-offs between travel time, departure time, and monetary cost. The preference parameters are estimated based on the equation for the expected utility as a function of departure time, $EU(t_d)$. There are numerous applications of the scheduling approach using the step model, both with fixed and random travel time, and the focus of these studies is usually on the estimated preference parameters. The VTTV can be derived as a function of these estimates, as set out above. In contrast to the scheduling approach, the reduced-form approach does not use information about departure times: It uses the equation for the optimal expected utility, EU^* , together with observations of trade-offs between mean travel time, TTV and monetary costs. This approach is the most common way to measure the VTTV, and usually TTV is measured by the standard deviation or variance of the travel time distribution (the mean-variance approach), or by the interquantile range. As explained, this corresponds to the reduced forms of the step model and the slope model, respectively. In theory, the scheduling and reduced-form approaches should yield similar results in terms of the VTTV, for a given underlying scheduling model, as long as we can assume travelers maximize their expected utility by choosing the optimal departure time. However, two stated preference studies by Abegaz et al. (2017) and Börjesson et al. (2012), who have compared the two types of valuations, reveal that they may be very different. Both studies find that the scheduling approach yields much lower valuations of TTV than the reduced form approach. This may be because travelers incur a disutility of variability in itself, over and above the scheduling costs, due to increased anxiety or higher planning costs, or because the underlying scheduling model is otherwise misspecified. The popular step and slope models are attractive because of their simplicity; however, it remains to be validated that they are indeed good approximations of actual behavior. And even if these simple scheduling models are good approximations of real behavior, it may well turn out that the behavior observed in stated preference trade-offs is not consistent with behavior in the real world. As mentioned earlier, this latter explanation does

seem likely, because choice tasks involving TTV are complicated to communicate without risk of misinterpretation, and complex to process in terms of the consequences for travel planning and activities. This casts some doubt on the appropriateness of using stated preference data in the context of TTV.

Application in Demand Prediction and Cost-Benefit Analysis

As mentioned earlier, application in practical demand modeling and cost-benefit analyses requires four things: (1) definition of an appropriate measure of TTV, (2) knowledge about effects of TTV on behavior in terms of departure time, mode and route choice, (3) a monetary social value of TTV, and (4) models which predict TTV for the relevant scenarios.

For cost-benefit analyses, ideally the predictions of TTV should stem from models taking both departure time choice, mode choice and route choice into account. Eventually, this may become available from the traffic models used to predict demand, but until this becomes available at a sufficiently detailed level, it is necessary to apply an ad hoc approach. Recent applications, either recently implemented or close to being implemented, involve simple TTV prediction models, that forecast the TTV based on output from traffic models, and ignore or underestimate that TTV may affect departure time patterns as well (de Jong and Bliemer, 2015).

Now, the potential errors related to using an ad hoc approach may discourage analysts and decision makers from even considering this, preferring to stick to their usual practice. To gain insight regarding the best way forward from here, we need numerical analyses of the size of the potential errors. A few theoretical or numerical illustrations are already available in the literature, but it would be useful to expand this evidence taking into account the empirical knowledge about the scheduling preferences in the population.

One such theoretical illustration stems from a bottleneck model with stochastic travel time. A bottleneck model is a highly simplified description of a transport system, usually considering a single transport corridor with a single point of congestion (the bottleneck) where travelers queue up if demand exceeds capacity. The benefit of this simple set-up is that the model can be used to derive general results analytically, whereas traffic models with large networks are much more complex and provide results that are highly context-specific and usually not easy to generalize. Using a bottleneck model where the travel time distribution is endogenous and changes over time, Xiao et al. (2017) demonstrate that ignoring the effect of TTV on departure time patterns may imply errors in project appraisal. However, the size of this error will depend on the true functional form of the scheduling preferences: If we are willing to assume scheduling preferences corresponding to the slope model or a model with a constant $h(t)$ and an exponential $w(t)$, these errors disappear altogether. In other cases, it is theoretically possible that the errors are of considerable magnitude. In other words, to make firm conclusions, we need additional empirical evidence about the scheduling preferences in the population of interest.

See Also

Cost Functions for Road Transport; Value of Time; Value of Crowding; Long- Versus Short-Run Valuations; The Value of Security, Access Time, Waiting Time, and Transfers in Public Transport; Demand for Passenger Transport; The Bottleneck Model; Dynamic Congestion Pricing and User Heterogeneity; Intertemporal Variation of Nonmarket Valuations; Elasticities for Travel Demand; Are Travel Choices Derived from the Rationality Assumption?; Estimating the Value of Time; Departure Time Choice Modeling; Stated Preference Surveys: Experimental Design and Modeling

References

- Abegaz, D., Hjorth, K., Rich, J., 2017. Testing the slope model of scheduling preferences on stated preference data. *Transp. Res. B* 104, 409–436.
- Benezech, V., Coulombel, N., 2013. The value of service reliability. *Transp. Res. B* 58, 1–15.
- Börjesson, M., Eliasson, J., Franklin, J., 2012. Valuations of travel time variability in scheduling versus mean-variance models. *Transp. Res. B* 46, 855–873.
- Carrion, C., Levinson, D., 2012. Value of travel time reliability: a review of current evidence. *Transport. Res. A* 46, 720–741.
- de Jong, G.C., Bliemer, M.C.J., 2015. On including travel time reliability of road traffic in appraisal. *Transp. Res. A* 73, 80–95.
- Engelson, L., Fosgerau, M., 2011. Additive measures of travel time variability. *Transp. Res. B* 45, 1560–1571.
- Fosgerau, M., 2016. The valuation of travel time variability. *International Transport Forum - Discussion Paper* 2016-04.
- Fosgerau, M., Engelson, L., 2011. The value of travel time variance. *Transp. Res. B* 45, 1–8.
- Fosgerau, M., Karlström, A., 2010. The value of reliability. *Transp. Res. B* 44, 38–49.
- Hjorth, K., Börjesson, M., Engelson, L., Fosgerau, M., 2015. Estimating exponential scheduling preferences. *Transp. Res. B* 81, 230–251.
- Jenelius, E., 2012. The value of travel time variability with trip chains, flexible scheduling and correlated travel times. *Transp. Res. B* 46, 762–780.
- Koster, P., Peer, S., Dekker, T., 2015. Memory, expectation formation and scheduling choices. *Econ. Transport.* 4, 256–265.
- Small, K., 2012. Valuation of travel time. *Econ. Transport.* 1, 2–14.
- Xiao, Y., Fukuda, D., 2015. On the cost of misperceived travel time variability. *Transport. Res. A* 75, 96–112.
- Xiao, Y., Coulombel, N., de Palma, A., 2017. The valuation of travel time reliability: Does congestion matter? *Transp. Res. B* 97, 113–141.

Further Reading

- Engelson, L., 2011. Properties of expected travel cost function with uncertain travel time. *Transp. Res. Rec. J. Transp. Res. Board* 2254, 151–159.
- Noland, R.B., Small, K., 1995. Travel-time uncertainty, departure time choice, and the cost of morning commutes. *Transp. Res. Rec.* 1493, 150–158.
- Tseng, Y.-Y., Verhoef, E.T., 2008. Value of time by time of day: a stated-preference study. *Transp. Res. B* 42, 607–618.

Value of Crowding

Daniel Hörcher, Imperial College London, London, United Kingdom; Budapest University of Technology and Economics, Budapest, Hungary

© 2021 Elsevier Ltd. All rights reserved.

Introduction	84
Physical and Behavioral Foundations	84
Measuring the User Cost of Crowding	85
Modeling Travel Disutility	85
Data: Declared and Revealed Preferences	86
Typical Empirical Estimates	87
Crowding in Models of Optimal Transport Supply	87
Concluding Remarks	88
See Also	88
References	88
Further Reading	88

Introduction

Crowding in a transport context refers to the state of high volumes of passengers or pedestrians sharing limited space when walking, queuing, waiting, or traveling. Empirical evidence shows that avoiding crowding has a value for public transport users, and they often do trade this source of disutility with excess travel time or monetary payments. Thus, the value of crowding can be measured in discrete choice models of mode, route, and departure time decisions. In most cases, crowding costs are expressed in terms of a multiplier function imposed on the value of in-vehicle travel time, where the argument of the multiplier may be the density of standing passengers inside the vehicle, for example. Empirical estimates of the value of crowding are applied in various contexts. First, it is an important component of demand modeling and forecasting. Second, crowding has a role in the economic appraisal of transport investments and policies; if such interventions ease crowding, then improved travel conditions generate social benefits in the form of consumer surplus. Third, the value of crowding affects short-run supply policies targeting welfare maximization, including pricing, timetabling, fleet management, and the degree of public subsidization.

Physical and Behavioral Foundations

Crowding emerges when travelers share limited floor space and their volume reaches a level that triggers discomfort for them. In case of public transport trips, one can distinguish between crowding (1) inside vehicles, (2) during boarding and alighting, and (3) during activities before boarding and after alighting and when accessing the vehicle. From a physical perspective, passengers may experience crowding both when they move (walk) and stand still. In the former case, high pedestrian density can lead to a reduction in walking speed, and thus travel delays, due to frictions and potential conflicts between pedestrians. Such delays can be especially significant in magnitude in large stations where the access time required inside the station to get to or leave the vehicles constitutes a nonnegligible part of the total journey time. Additionally, the density of crowding inside the vehicle and on the platform or bus stop can affect dwell times, and consequently cause external travel time expansion for a wide range of passengers through the delay accumulated by the vehicle itself. In extreme cases, but in densely populated urban areas absolutely not rarely, in-vehicle crowding can lead to failed boarding, which is an even more severe source of travel time delay. All these consequences of crowding may have an additional external impact on travel utility through travel time loss and unreliability (refer to Article Valuing Travel Time Variability: Scheduling and Reduced Form Models).

In crowded situations, passengers experience a number of nuisance factors stemming from close physical proximity to fellow users. This may cause disutility through physiological stress, annoyance, and ultimately various forms of frustrations and fatigue. Major sources of discomfort include intrusion into personal space, noise, smell, increased accident risk, security concerns, and lack of access to natural light and fresh air (Haywood et al., 2017). Crowding limits the extent to which passengers can move inside the vehicle and use its amenities; for example, access to washrooms and preferred seats can be hindered by passengers blocking the interior area. Access to seating is a particularly relevant consequence of high vehicle occupancy, as standing in itself is a major source of disutility that may lead to physical tiredness and pain, even if a passenger is not surrounded by other standees at all. Similarly, the difficulty in accessing the doors is a major concern for travelers in crowding, as failure to reach the doors at the desired destination station can lead to various unexpected travel costs, including a detour.

It is important to note that, in contrast to other user cost components in transport, the micro-foundations of crowding disutility have not been established in the literature so far. The value of travel time loss and the disutility of monetary payments, for example, can both be derived from time and monetary budget constraints of the utility maximizing individual. In the former case, travel time

as a resource may be costly because it impacts the amount of time available for leisure or paid work, and this disutility can be quantified separately from the inconvenience of in-vehicle crowding. By contrast, there are no known models of such thing as a *discomfort budget*, from which people's aversion to crowding and the shadow price of relieving could be derived in a theoretically sound manner. Intuition suggests that spending time in crowding may have a negative impact through fatigue on people's productivity at work as well as their need for leisure time and recovery, but the mathematical formulation of this dependency and its empirical validation is an outstanding task on the research agenda.

Measuring the User Cost of Crowding

Crowding causes disutility for travelers, and measuring the magnitude of annoyance is essential to predict user behavior with high precision and develop efficient supply-side policies to mitigate the inconvenience. Crowding disutility cannot be modeled independently of other travel attributes, such as the costs of accessing public transport, and spending time while waiting or traveling inside the vehicle, because we can hardly observe situations in which crowding varies but the remaining travel attributes are constant. From both a chronological and methodological point of view, the measurement of the value of crowding had been predominantly part of value-of-time estimation experiments (refer to Article Estimating the Value of Time).

The majority of the empirical literature estimates the cost of crowding by modeling and calibrating observed or stated discrete travel decisions, where passengers trade crowding inconvenience with monetary expenses and/or travel time loss (Li and Hensher, 2011; Wardman and Whelan, 2011). Such trade-offs happen, for example, when one selects the departure time, mode, and route before a trip. The severity of crowding, just like other travel attributes, varies by time, space, and direction in a large public transport network. Thus, the choice of mode, route, and departure time implies a comparison of relative crowding costs between available alternatives. Random utility discrete choice modeling offers a suitable method to calibrate a utility function of crowding and other discomfort factors that best explains the observed choices, conditional on the assumption that passengers select the alternative providing the highest level of utility. On this basis, the rate of substitution between crowding disutility and other nuisance factors can be revealed. When the comparison is made with monetary costs, in particular, the rate of substitution measures the marginal willingness to pay to avoid crowding.

Modeling Travel Disutility

The functional form of the underlying utility function, that is, the way how crowding is assumed to be related to other travel attributes, is a critical challenge of the estimation process. In the simplest case, a suitable measure of crowding can be an additively separable component, where its coefficient is simply the marginal utility of crowding. This approach implies, however, that crowding discomfort is completely independent of other travel attributes, including the time spent inside the vehicle. This assumption is arguably rather unrealistic, and thus the resulting estimate could not be applied in other choice situations unless travel time is constant. It is more common in the literature to assume that crowding affects disutility as a multiplier of the value of time. From the viewpoint of the specification of utility, this implies that the function needs to include a travel time attribute, and an interaction term of travel time and the measure of crowding. This is where the literature stands right now.

What do we mean by a suitable measure of crowding? The challenge here is to define quantitative or qualitative metrics of crowding that create a good association between what enters the model and what the observed passengers perceive. In early contributions to the crowding literature, the occupancy rate of seats was used for this purpose, allowing the metric to go above 100%, thus expressing the number of standees relative to seating capacity. This approach has the obvious limitation that the metric changes if we add or remove seats, keeping the total number of people on board constant. One may also express vehicle occupancy in percentage terms, assuming a theoretical capacity, including both standing and seated passengers. It is a usual industrial practice to set this theoretical threshold to four standing passengers per square meter. However, the resulting estimates still cannot be applied for vehicles with varying seat-to-standing-room ratios, as densely seated vehicles reach the theoretical capacity much earlier, while, in fact, more passengers travel more comfortably in such vehicles. To overcome the difficulties associated with interior layout, it is rational to distinguish the disutility of standing from discomfort stemming from the density of passengers. Recent contributions in the literature treat the probability of standing and the density of standing passengers as separate factors of the value of time multiplier, where standing density is normally expressed in terms of the average number of standees sharing a unit of floor area (Wardman and Whelan, 2011).

Another critical question of model specification is whether disutility is a linear function of the density of crowding, or more complicated functional forms should be applied to represent passenger experience adequately. Convex functional forms may be appealing intuitively, in line with the usual cost functions applied for modeling road congestion. However, the existing literature has not found empirical evidence of a nonlinear relationship (Wardman and Whelan, 2011), meaning that linear specifications achieve better model fit and coefficient significance in the observed range of crowding levels. Note that we cannot assume on this basis that crowding costs increase linearly without any upper bound on possible equilibrium occupancy rates. Very dense crowding leads to failed boarding, inducing considerable delays for passengers. The researcher may treat the time cost of denied boarding as part of the expected waiting time, but eventually, the occupancy rate function of the aggregate user cost must at some point become very steep, if one intends to exclude the possibility of unrealistically high in-vehicle crowding levels as a model outcome.

The issue of nonlinearity arises in case of the interaction with travel time as well. Traveling in crowding causes fatigue, especially while standing, and tired passengers may be more sensitive to the disutility of high vehicle occupancy. This implies that the crowding multiplier on the value of a unit of time should increase with the travel time itself. In the empirical literature there is a tendency toward higher crowding cost estimates for medium- and long-distance services compared to shorter urban public transport trips, but, to the best of our knowledge, there is no published study in which an endogenous, travel time dependent crowding multiplier function is estimated explicitly. This opens up room for new empirical contributions.

Finally, let us note that despite the mainstream literature tends to focus on very high vehicle occupancy rates in the context of crowding costs, the marginal trip may cause disutility under more relaxed demand conditions as well. [Wardman and Murphy \(2015\)](#) show that passengers may have strong preferences for using certain seats in public transport vehicles, for example, the ones facing forward or located close to the window. Users sometimes even decide to stand instead of using an unattractive seat. Thus, if a passenger's most preferred seat is already occupied by someone else, the marginal cost of that fellow passenger's trip can be relatively high, despite the negligible occupancy rate and the fact that no passenger is forced to stand.

Data: Declared and Revealed Preferences

To estimate the value of crowding successfully, the underlying data should comply with four essential criteria: it has to provide information on (1) what alternatives a decision-maker considered before selection, (2) which one was eventually chosen, (3) data should be available on expected travel conditions, including all attribute levels for each alternative, and (4) there must be sufficient variation between alternatives in crowding conditions as well as other attributes to make coefficient estimation possible. Criteria 1–3 depend on the data source and data quality, while the last one depends on the choice situation itself. In the upcoming discussion we review the most often cited pros and cons of stated and revealed preference (SP and RP) methods.

Survey-based data collection offers a number of advantages compared to recording RPs. As the choice situation is artificial in a survey, the researcher can be sure about what alternatives the decision-maker considers, and expectations about various attribute levels are also unambiguous. Moreover, with appropriate survey design methods, sufficient variation in attribute levels can also be guaranteed. On the other hand, the downsides of these idealized experimental conditions are numerous: responses may not be fully in line with how passengers behave in reality; only subsamples of the entire population of travelers can be involved in the experiment; and data collection is generally more expensive than using existing automated datasets.

Representing crowding in an easily perceivable manner for respondents is one of the main challenges of the SP approach. The literature suggests that users cannot associate crowding densities expressed in terms of the number of passengers per square meter with their personal experience, so some form of verbal or diagrammatic representation is required to explain or visualize occupancy rates. Plan view images, photos, and stylized axonometric drawings are the most usual forms of representation. Recent empirical results show that in case of crowding cost estimation, the actual form of visual representation has no severe influence on the estimated coefficients. Declared preference studies are in a majority among crowding cost estimation experiments. It is indeed a huge benefit of the stated preference method that *expectation formation* is not an issue in their case. That is, in a survey the researcher has perfect control over the alternatives as well as the attribute levels considered by respondents.

Choice models can be calibrated using data on the observed behavior of real passengers as well. In an RP setting it is not easily evident what alternatives were available for the user before selecting a route or mode, for instance. It is even less obvious what attribute levels she or he associates with the available alternatives. Furthermore, the actual crowding level a passenger experiences on a trip, which can be recovered from smart card data, for example, is not necessarily in line with her prior expectation. The lack of uncertainty seems like a comparative advantage of the SP approach, but in fact passengers rarely face travel decisions in which they know deterministically the exact travel conditions a priori. The cost of uncertainty can be an important component of the discomfort associated with crowding in public transport.

Occupancy rates may vary during a single journey as well. It is often the case that passengers have to stand at the beginning of a trip, but later on the chance of getting a seat increases to same extent. In the SP framework the representation of varying travel conditions becomes too complicated. Intuition suggests that the probabilistic treatment of standing and traveling in crowding is more realistic, simply because this is the only way public transport users consider crowding when selecting the mode and route of traveling. [Hörcher et al. \(2017\)](#) illustrate that the combination of smart card and vehicle movement data is a suitable and attractive source of information to perform the three main steps of crowding cost estimation: (1) recover the pattern of occupancy rates of vehicles in a network where multiple potential routes exist for at least a subset of users, (2) derive the expectation of passengers departing at a given point in time on the level of crowding along potential routes considered, and (3) infer the actual route chosen. In the currently available literature, expectation formation in step (2) is limited to the assumption that regular commuters are familiar with the mean occupancy rate of network segments used in their respective time of departure. However, future contributions may extend this approach, and model expectations, using the past experience of each regular passenger, encapsulated in a time series of smart card data. Note that even if expectations are estimated correctly, it is possible in an RP setting that crowding on alternative routes does not differ sufficiently to get reliable estimates.

The magnitude of uncertainty regarding the crowding experience prior to a trip may have an adverse effect on travel utility, in a fashion similar to travel time variability (refer to Article Valuing Travel Time Variability: Scheduling and Reduced Form Models). The micro-foundations of crowding variability, however, are less straightforward than the link between the likelihood of late arrivals and schedule delay costs.

The majority of published crowding cost estimation studies use a representative consumer approach assuming homogeneous attitude toward crowding, despite the fact that the methodological toolbox of measuring heterogeneous preferences has been available for more than two decades. A potential reason is that the measurement of heterogeneity is not crucially necessary for some of the applications, such as the cost–benefit analysis of crowding reduction projects. Recent empirical research efforts by [Tirachini et al. \(2017\)](#), built on mixed logit and latent class discrete choice models, show that heterogeneity does exist, and age, gender, and income are among the influential factors correlated with crowding avoidance.

Typical Empirical Estimates

In this short review, let us concentrate on existing estimates of the occupancy rate dependent travel time multiplier, where the rate is expressed in terms of the density of standing passengers. As a rule of thumb, the penalty for standing, measured as a value of time multiplier, is normally between 1.1 and 1.6. Specifically, this value is found to be 1.54 in major Swedish cities, 1.43 in suburban London, 1.29 in Paris, 1.26 in Hong Kong, and 1.1 in Santiago, Chile. In very dense crowding, the standing multiplier goes up to 2.21 around London, 2.13 in Swedish cities, around 2 in both Hong Kong and Santiago, and 1.6 in Paris. Empirical evidence shows that seated passengers are also negatively affected by the density of standees; their value of time multiplier peaks at 1.71 in Hong Kong, 1.67 in Santiago, 1.54 on suburban lines around London, 1.50 in Sweden, and 1.41 in Paris (for a more detailed comparison of the empirical results above, see [Tirachini et al., 2017](#)). In relation to systematic differences between the estimates of SP and RP studies, some authors state that SP methods tend to overestimate the value of crowding, especially in case of standing multipliers. However, this is not a consensual finding, nor one supported by a quantitative meta-analysis, primarily because the number of RP experiments available in the literature is still very limited.

In terms of heterogeneity in crowding valuations, for the specific case of Santiago, Chile, [Tirachini et al. \(2017\)](#) report that younger, high-income males are less sensitive to crowding discomfort, while females, the elderly, and, interestingly, low-income travelers form the other extreme of the distribution. At least in this particular city, the median value of crowding, derived in a mixed logit setting with lognormal distribution, happens to be very similar to the point estimate of a simple multinomial logit (MNL) model. This implies that the more simple MNL approach gives sufficiently reliable estimates when the goal of the analysis is to model aggregate behavior. On the other hand, heterogeneity can be important in pricing models, as we often see in the second-best infrastructure pricing literature (refer to Article Second-Best Congestion Pricing).

Crowding in Models of Optimal Transport Supply

For a given public transport capacity, the disutility experienced by the average user increases with demand. This implies that crowding has to be considered as a consumption externality. Empirical evidence suggests that the external crowding cost of peak-hour trips is a nonnegligible part of its marginal social cost. That is, unless the value of this externality appears in the price of using a public transport service, passengers tend to overconsume it, thus generating potentially significant deadweight loss for society ([Tirachini et al., 2013](#); [De Palma et al., 2015](#)).

Crowding breaks the paradigm that public transport is an uncongestible substitute of car use. That is, policy-makers have to be careful when underpricing public transport with the aim of tackling car congestion, because suboptimal fares may lead to overconsumption and welfare losses. Until the wider dissemination of crowding cost estimation techniques reviewed in the previous section, the literature of optimal public transport supply was dominated by models built on waiting time costs and the Mohring effect (refer to Article The Mohring Effect). Adding crowding costs to a public transport model is expected to lead to higher fares and lower subsidies in optimum. One can reach this intuitive conclusion from the Mohring–Harwitz cost recovery theorem as well. The theorem states that under certain assumptions, including perfectly divisible capacity, the optimal degree of self-financing equals the elasticity of operational costs with respect to output (which is one in case of neutral scale economies) plus the degree of homogeneity of the user cost function. If the user cost function contains waiting time only, then its degree of homogeneity is -1 , so that the optimal cost recovery ratio becomes zero. By contrast, if user costs are dominated by the occupancy rate dependent crowding inconvenience, then the user cost function's degree of homogeneity turns into zero, and optimal pricing results in perfect self-financing ([Hörcher and Graham, 2018](#)). Since crowding is very likely to have a bigger impact on travel utility than waiting time in high-frequency public transport services, the theorem implies that, *ceteris paribus*, the cost of capacity provision should be recovered by fare revenues to a greater extent than in low-density public transport. At the same time, density economies in operational costs (refer to Article Operation Costs for Public Transport), substitution with underpriced road congestion (refer to Article Second-Best Congestion Pricing), and positive agglomeration economies (refer to Article Wider Economic Impacts of Transport Investments) may still provide second-best motivations for subsidization, even in the case of a heavily crowded service.

Are there effective strategies in capacity provision that eliminate crowding entirely? Even if such strategies exist, they are likely to be wasteful and very inefficient from a social welfare point of view. The reason why crowding is inevitable to some extent in a public transport network is that travel demand is unbalanced in spatial, temporal, and directional terms as well ([Hörcher and Graham, 2018](#)). For technological reasons, capacity (i.e., frequency and vehicle size) cannot be varied as quickly as demand fluctuates. Therefore, a given second-best capacity often has to be in service under various demand conditions. This leads to suboptimally low occupancy rates in certain time periods and network sections, and inevitable crowding in the peak. The literature of multi-period public transport supply suggests that the optimal size of vehicles increases with the degree of demand imbalances

at the expense of frequency, and the underlying reason is crowding: if demand is concentrated in a small subset of markets served by fixed capacity, then the relevance of crowding discomfort once again increases relative to waiting, even if aggregate demand remains constant.

Due to the inflexibility of capacity, differentiated pricing remains the last resort to mitigate the crowding-related consequences of demand fluctuations. In the past, the implementation of either temporally or spatially differentiated pricing was complicated from a technological point of view, especially compared to the convenience and simplicity of flat fares and travel passes. However, the advent of the digital age and new mobile payment technologies enable public transport operators to move their tariff systems closer to marginal crowding cost pricing. With differentiated pricing, crowding can be moderated in the most densely used network sections, while capacity utilization can be improved in off-peak periods and line sections.

Concluding Remarks

Crowding research is an emerging subject in the field of transport economics. Its relevance is undeniable in large metropolitan areas where mass public transport is a primary means of mobility due to the scarcity of space, as well as in economies where car ownership is not affordable for a significant part of society. The core policy messages of transport economics apply for public transport as well. We should not expect that we can build our way out of crowding by simply adding more capacity; indeed, induced demand is even more often neglected in a public transport context than in the case of road expansion. Nevertheless, capacity expansion does have the potential to improve social welfare, even if crowding reemerges due to induced demand. Crowding cannot be eliminated completely with differentiated marginal cost pricing either, but it provides the right incentive for passengers to perform trips that generate more benefits than costs for society as a whole.

See Also

Operation Costs for Public Transport; Valuation of Travel Time Variability Using Scheduling Models; The Mohring Effect; Wider Economic Impacts of Transport Investments; Estimation of Value of Time

References

- De Palma, A., Kilani, M., Proost, S., 2015. Discomfort in mass transit and its implication for scheduling and pricing. *Transp. Res. Part B Methodol.* 71, 1–18.
- Haywood, L., Koning, M., Monchambert, G., 2017. Crowding in public transport: who cares and why? *Transp. Res. Part A Policy Pract.* 100, 215–227.
- Hörcher, D., Graham, D.J., 2018. Demand imbalances and multi-period public transport supply. *Transp. Res. Part B Methodol.* 108, 106–126.
- Hörcher, D., Graham, D.J., Anderson, R.J., 2017. Crowding cost estimation with large scale smart card and vehicle location data. *Transp. Res. Part B Methodol.* 95, 105–125.
- Li, Z., Hensher, D.A., 2011. Crowding and public transport: a review of willingness to pay evidence and its relevance in project appraisal. *Transp. Policy* 18 (6), 880–887.
- Tirachini, A., Hensher, D.A., Rose, J.M., 2013. Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. *Transp. Res. Part A Policy Pract.* 53, 36–52.
- Tirachini, A., Hurtubia, R., Dekker, T., Daziano, R.A., 2017. Estimation of crowding discomfort in public transport: results from Santiago de Chile. *Transp. Res. Part A Policy Pract.* 103, 311–326.
- Wardman, M., Murphy, P., 2015. Passengers' valuations of train seating layout, position and occupancy. *Transp. Res. Part A: Policy Pract.* 74, 222–238.
- Wardman, M., Whelan, G., 2011. Twenty years of rail crowding valuation studies: evidence and lessons from British experience. *Transp. Rev.* 31 (3), 379–398.

Further Reading

- Haywood, L., Koning, M., 2015. The distribution of crowding costs in public transport: new evidence from Paris. *Transp. Res. Part A Policy Pract.* 77, 182–201.
- Hörcher, D., Graham, D.J., Anderson, R.J., 2018. The economics of seat provision in public transport. *Transp. Res. Part E Logist. Transp. Rev.* 109, 277–292.
- Kroes, E., Kouwenhoven, M., Debrincat, L., Pauget, N., 2014. Value of crowding on public transport in Île-de-France, France. *Transp. Res. Rec. J. Transp. Res. Board* 2417, 37–45.
- Tirachini, A., Hensher, D.A., Rose, J.M., 2014. Multimodal pricing and optimal design of urban public transport: the interplay between traffic congestion and bus crowding. *Transp. Res. Part B Methodol.* 61, 33–54.

What Drives Transport and Mobility Trends? The Chicken-and-Egg Problem

Nathalie Picard, Université de Strasbourg, Université de Lorraine, CNRS, BETA, Strasbourg, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	89
Trends and Drivers of Transport Demand	90
The Basic Four-Step Model	91
Sources of Endogeneity: What Generates the Chicken-and-Egg Problem?	92
Common Determinants and Heterogeneity	92
Reversed Causality and Simultaneous Decisions	92
Interactions Between Transport and Urban Development	93
Egg, Chicken, Hen and Cock, or Family Economics	93
Conclusions	94
Acknowledgments	94
References	94

Introduction

Nowadays, in developed countries, urban passenger transport demand is dominated by daily trips between home and workplace, especially during the morning and evening peaks. The evolution of transport demand is thus mainly governed by the evolution of the location of residential units and job location, that is, by urban dynamics. We discuss below the interactions between transport and urban development, sometimes referred to, in the literature, as the “chicken and egg” problem (see, e.g., [Rodrigue, 2020](#), chapter 8.2).

Transit investments shape urban dynamics, but such investments also react to increased demand for transit induced by urban dynamics. The process of urbanization and development of public transportation goes hand to hand. However, the dynamics may vary over time and over space. Sometimes, population and job relocate before the construction of new transit infrastructures, partly because households, firms, or stakeholders want to buy property before real estate prices increase. Such anticipations usually occur when a large investment in public transit infrastructure (such as Crossrail or the Grand Paris Express) is planned. See [Picard and de Palma \(2019\)](#) for details. In other cases, the relocation of residential units or the conversion of single-family residential units to multifamily housing or to mixed development units, and more generally densification occurs after the construction of mass transit. One of the reasons is the uncertainty concerning the time to complete the transit investment and the uncertainty related to a snow bowl effect generated by urbanization, which involves a complex pattern of interactions within and between sectors.

The urban development process is complex and requires detailed econometric analysis to be quantified (see [Antoniou and Picard, 2015a](#) for the econometric methods relevant for measuring each relation in this complex system). For example, the price of land is not monotonic with respect to the distance to a railway station, to a bus station or to a metro station. This is because public transit generates both positive and negative externalities. Positive externalities are related to the positive effect of transit on accessibility. Improved accessibility in turn changes the social mix and moves jobs within and between regions, and generates agglomerations effects. Negative externalities include noise, pollution, and possibly higher crime rates. Such combination of positive and negative effects may give rise to multiple equilibria, as discussed below. Econometric models can be used to measure the mutual influences between transit investments and urban dynamics. The relationships between transit and land use is involved, and it also depends on the specific legislations in urban areas, such as zoning, reserved areas, green areas and the like.

Based on standard economic principles, all other things being equal, the local demand for dwellings or offices decreases with local real estate prices, whereas local building supply increases with local real estate prices. Local real estate prices in turn increase with local demand for dwellings or offices, and decreases with local building supply, all other things being equal. From the mathematical point of view, this means the long-term real estate prices, population and job locations, and transport demand are the solution of a fixed-point problem associated to a nonlinear system of equations. Standard tools based on [Kakutani \(1941\)](#) fixed point theory guarantee existence of a solution for systems of nonlinear equations, under quite realistic assumptions. Other tools (based e.g., on contracting mapping or on diagonal dominance like in [Ginsburgh et al., 1985](#)) guarantee uniqueness under far more restrictive and often unrealistic assumptions. For example, uniqueness is not guaranteed when positive externalities are high enough, or if population is characterized by heterogeneous preferences or firms are heterogeneous. This leaves room for policy analysis in order to uncover instruments to help the selection of the “best” equilibrium solution in terms of efficiency and sustainability.

Dynamic models offer a different view, and provide a path to stationary solution(s). In the case of multiple equilibria, dynamic models offer powerful tools to analyze how public policies may contribute to the selection of the preferred equilibrium solution. The way agents acquire, transmit, and process information may induce a time dependency of the final solution. When positive network externalities are strong enough (with respect to negative externalities), multiple solutions may arise. As a result, two forces are opposed, and the modeling of the dynamic path allows selecting the specific solution. For example, according to [Brueckner et al. \(1999\)](#), the location of the most preferred local amenities explains why the equilibrium selected concentrates the rich population in central Paris, but the poor population in downtown Detroit.

The question of which comes first, the chicken or the egg, is not only a philosophical one, it is also of primary relevance for policy evaluation. In the context of demand analysis, the chicken-and-egg problem refers to simultaneity, endogeneity, and causality. There is a simultaneity problem because the drivers of transport demand such as local economic activity, performance or growth also influence transport supply, and because transport demand depends on transport supply, and vice versa. Such interdependencies raise endogeneity problems in the measurement of transport demand functions, in particular in the computation of various elasticities.

In such a context, the relationship observed empirically between transport demand and transport supply reflects both the elasticity of transport demand to transport supply, the elasticity of transport supply to transport demand and the effect of confounding factors (drivers of transport demand) affecting both transport demand and supply.

Such problems are not specific to the context of transport, and they have given rise to a very rich literature in macroeconometrics, in relation to various economic theories. The most popular methodology to go from correlation to causality measurement relies on the dynamics of interactions, with the famous notion of [Granger \(1969\)](#) causality. The idea is to rely on the timing of variables evolution. Variable A causes variable B in the Granger sense if a change in A tends to be followed, one or a few periods later, by a change in B. However, when agents anticipate consistently, the consequence may happen before its cause. For example, when a couple is expecting a baby, they usually decide to move to a larger dwelling before the baby is born. The birth is the cause of moving, but it is generally observed after its consequence (the move).

This article is mainly focused on Western Europe, and restricted to urban passenger transport demand (expressed by individuals living in households), which excludes freight, airplanes, and high-speed trains. The next section reviews the drivers (determinants, causes) of transport demand, and their trends. Most of those determinants are also affected by transport demand, giving rise to a chicken-and-egg problem. The basic four-step model provides an illustration of interdependent decisions leading to a chicken-and-egg problem and stresses the role of anticipations. After reviewing the different sources of endogeneity in a general setup, we analyze the chicken-and-egg problem in the light of Interactions Land Use and Transport. We finally argue that taking into account gender differences in transport demand, and the interactions between individuals in the same household transforms the chicken-and-egg problem into a chicken, egg, hen, and cock problem.

Trends and Drivers of Transport Demand

[Sessa and Enei \(2009\)](#) provide a review of transport demand trends and drivers from 1995 to 2007. They stress that passenger transport demand (expressed in passenger kilometers travelled) in the 27 EU countries has decreased by 7.7% for maritime transportation, and increased by 21.4% for cars, 24.8% for powered-2-wheel, 6.9% for bus and coach, 12.7% for railway, 20.1% for tram and metro, and 70.4% for air. All modes together, the increase is 22.3% for this period, which represents an average of 1.7% per year. The car mode share, 72%, is stable over the period. The long-distance (over 100 km) trips account for 2.5% of the number of trips and 53% person kilometers.

These trends were significantly affected during the last decade by the emergence of new transport services and systems such as Mobility on Demand, Mobility as a Service, vehicle-sharing, ridesharing, or autonomous vehicles. This emergence was mainly driven by rapidly developing mobile information and communication technologies. See [Antonioni et al. \(2019\)](#) for details.

More recently, the COVID-19 pandemic has dramatically reduced transport demand, especially for air and public transport, while dramatically increasing the shares of walking, cycling and other new energy-friendly modes. According to [Sung and Monschauer \(2020\)](#), the Covid-19 crisis has changed already people's transport behaviors in dramatic ways, with large reductions in aviation and public transport use and significant growth in cycling uptake. They argue that "Evidence from previous crises shows that in the immediate aftermath of crisis events, transport behaviors will change, as people reassess the costs and benefits of different transport modes." [Abu-Rayash and Dincer \(2020\)](#) seem less optimistic about the long-term effects on the environment, possibly because their study covers a slightly longer period after the end of the lockdown policy. Indeed, their mobility index significantly increased again in May or June in Moscow, Paris, Brussels, Singapore, or Hong Kong, for example. Although it is too early to assess accurately the long-term effects of the Covid-19 pandemic, the other long-term trends can be better assessed by analyzing the trends in the usual transport demand drivers. Changing mobility behavior may influence location patterns, which may in turn call for new developments of transport infrastructure, another chicken-and-egg illustration.

Starting with **demographic drivers**, there is a clear consensus that transport demand depends on age, and that population is ageing in most of developed countries. At a given period and in a given country or region, travel demand significantly depends on age, reflecting a mix of the effect of age and of generation. On the one hand, above a certain age, a given person tends to travel less, and the car mode share decreases with age; on the other hand, as argued by [Sessa and Enei \(2009\)](#), "it can be expected that the future old people will travel more than previous generations of older people did." Moreover, the car mode share will probably be larger within the generation which will be over 60 in 20 years than within the generation currently over 60. In addition, older people tend to live farther away from the CBD than younger generations, which typically increases their need for mobility, at least until they retire. Finally, older people are richer on average, and thus have a larger demand for medium and long distance leisure trips. All in all, in the long run, population ageing will probably induce more inter-urban trips, by train, car, and air, and more urban trips by public transport and car.

International migration is the second demographic driver of transport demand mentioned by [Sessa and Enei \(2009\)](#), with nearly 100 million people expected to migrate from developing countries (mainly from Asia) to developed countries (mainly in North America, Western Europe, and Australia) between 2005 and 2050. Migration will largely offset the natural decrease in population (more deaths than births) in developed countries, whereas it will represent less than 5% of population growth in

developing countries. If migration flows remain younger, with higher fertility than natives, and mainly directed toward the outskirts of agglomerations, they will result in more short distance trips in urban areas by car and public transport. Moreover, the first generations of immigrants are expected to become richer, which will further increase future transport demand by individual means (cars, motorbikes, etc). The (chicken-and-egg) question is who comes first, international migration or urbanization?

Urbanization (from about 1/5 of European population in 2015 to about 4/5 in 2030) is expected to be associated with increased urban sprawl, that is a relative shift in the location of housing and activities toward the peripheries of the urban agglomeration, in a context where land resources are already relatively scarce. Achieving a sustainable balance between competing land uses will thus become more and more a key issue for development policies in Western Europe. Urban sprawl is usually associated with an increasing dependence on the automobile, either as the single mode for commuting, affairs and leisure trips, or in association with public transport for daily commuting trips between dwellings typically located in the suburbs and jobs typically located closer to the CBD. All in all, urbanization should result in an increase of local and short distance trips using collective transport and short-medium distance trips by car. The (chicken-and-egg) question is who comes first, urbanization, urban sprawl, or commuting decisions?

The first **economic driver** of transport demand is **gasoline price**, which depends both on international markets and on national taxation policy. Gasoline is heavily taxed in most European countries (about 70%), whereas tax represents only 21% of gasoline price in USA, 38% in Canada, 48% in Australia and New Zealand, and 53% in Japan. In international comparisons, the negative effect of gasoline price on transport demand by car (price elasticity) is difficult to estimate because of the confounding effect induced by population density and the availability of alternatives to car. [Sessa and Enei \(2009\)](#) argue that, when energy prices become very high, they become the main barrier to global trade. The possible answers to a fuel price increase are: switch to a more fuel-efficient or electric vehicle; consolidate or link trips; carpool and other modes (walking, biking) shifts, relocation of residence or activity, and (more and more in reaction to COVID-19 pandemic) teleworking. Polycentric cities offer opportunities for reducing commuting distances because they offer more possibilities to make jobs and dwellings closer, and to reduce congestion on the transport network.

The second economic driver of transport demand mentioned by [Sessa and Enei \(2009\)](#) is **GDP growth**. They report an elasticity of passenger transport to GDP of 0.9 in EU for the period 1990–2005, with a projected decrease to 0.65 for the period 2005–2030. This corresponds to a 1.4% yearly increase in passenger transport demand between 2005 and 2030 and suggests a decoupling between GDP growth and passenger transport demand. It is too early to find updated figures showing how the elasticity and the passenger transport demand reacted and will react further to the huge decrease in GDP during the pandemic. The (chicken-and-egg) question is whether insufficient transport supply slows down GDP growth, or faster GDP growth increases transport demand?

Turning to the **technical drivers** of transport demand, [Sessa and Enei \(2009\)](#) argue that, if one assumes reducing but stable economic growth, and sustained international trade and urbanization, then the technological opportunities offered by the development of **Information and Communication Technologies** may influence transport demand in the following directions. A reduction in travel frequency, but perhaps longer distance travel (individuals move further from work, due the globalization trends); substitution of work travel with other travel (with time saved by not travelling to work), due to widespread diffusion of flexible and remote working technologies. ICT may induce a modal shift toward public transport, due to new technologies (Integrated public transport planning information, e.g., real time information on bus schedules) and E-ticketing. Paradoxically, Real-time route guidance and hazard warning will obviously help saving in congestion and travel time, but may also increase distances travelled. The (chicken-and-egg) question is whether ICT technological developments were induced by the pressure of the demand to improve the efficiency of transport systems, or ICT innovations shaped transport demand?

The aim of **new transport infrastructures** is not limited to serving already existing or anticipated business-as-usual transport demand resulting from urban development. In the case of transformational projects such as the Grand Paris Express, they also aim at enhancing urban development and economic growth, by attracting the most productive workers and firms in the region. Such densification of high-productivity economic activity induces agglomeration effects, as detailed in the section analyzing Interactions between transport and urban development. This places the chicken-and-egg problem at the core of the evaluation of transport infrastructure investments.

Last but not least, **lifestyle changes** are important drivers of the changes in transport demand. For example, in most European countries, owning a car starts not to be seen much as a status symbol (at least among parts of the younger generation) and the only provider of “mobility freedom” in the younger generations. A new sustainable mobility freedom concept is emerging in the urban environment, with a mode switch toward active travel (walking and cycling), and from car (driving alone) toward carpooling and public transport. The mode switch from car to public transport was, however, more or less stopped by the COVID-19 pandemic. The chicken-and-egg problem in this context is whether lifestyle changes in residential location and transport demand are made possible by economic growth and availability of cleaner and more environment-friendly transport modes, or whether new ecological and sustainability concerns imposed enough pressure to promote technical changes and enhance a more sustainable urban development.

The Basic Four-Step Model

Several modeling approaches have been used in the literature for studying travel demand. In the light of the chicken-and-egg problem, these approaches differ with respect to their inclusiveness of choice dimensions (do these models consider other choices made simultaneously, or interacting, with transport choices) and their forward and backward consistency, as discussed below.

Urban transportation planners have traditionally used the so-called classical model, which incorporates the following interdependent four choices. (1) Trip generation (choice of number of trips by purpose, origin, and destination of each trip) produces the aggregate number of trips in the system by origin, by destination, and by purpose. Since most trips at rush hours are commuting trips, trip generation depends on the location of households and jobs. (2) Trip distribution links origins and destinations, that is produces O-D matrices of number of trips by purpose, by O-D pair. This step typically relies on gravity or an entropy model. (3) Mode choice is typically described by binary or multinomial logit model, and (4) Assignment is specific to the mode used. Assignment usually relies on minimum distance, or minimum general cost.

The classical model has been applied with some success to analyze major infrastructure investments. However, it is not well suited for evaluating policies such as road pricing that are designed to modify travel behavior in a substantial and/or comprehensive way. The first reason is related to the endogeneity of attributes in the different steps of the model, as explained in the endogeneity section below. This chicken-and-egg problem can be solved by using adequate econometric techniques such as instrumentation, fixed effect modeling or other difference in difference techniques to correct for the endogeneity of attributes. Such corrections are crucial because policy evaluation requires a precise and unbiased estimation of the determinants of travel demand. In the case of stated preferences surveys, [de Palma and Picard \(2005\)](#) argue that the best way to correct for the endogeneity of attributes is to introduce enough randomness when selecting the value of each (endogenous) attribute.

The second reason is related to the interdependence of the individual choices leading to the four-step model. As noticed by [de Palma et al. \(2005\)](#), “it would simplify matters greatly if travel decisions could be modeled as if they were made sequentially, rather than simultaneously, since the number of choice combinations to consider would be reduced from the product of the number of choices at each level to the sum.” A sequential approach, to be valid, should ensure the consistency of choices in both “forward” and “backward” directions. Consistency in the forward direction matters because upper-level choices determine lower levels opportunities. For example, before car sharing was available, and neglecting car rental, the car alone mode was not available to those who decided not to own a car. Consistency in the backward direction also matters because preferences for options available at a lower level affect the relevance choices made at upper levels. For example, if commuting by car becomes more interesting because congestion is alleviated, this increases the expected utility of buying a car. See the anticipation and nested models below for more details.

Car use and car ownership are thus closely related. The chicken-and-egg problem arises because forward and backward consistency implies that they influence each other. Is it the case that you do not drive by car because you do not own a car, or you decided not to buy a car because you anticipated you would not use it much? The answer to these questions is typically given by the Nested Logit model, which ensures the consistency of choices in both “forward” and “backward” directions. The four-step model can be interpreted as the result of individual nested decisions amenable to a Nested Logit model. The top of the individual decision tree corresponds to long-term decisions such as location and vehicle, ownership. The bottom of individual decision tree corresponds to short run parking location. The three choices considered in the road-pricing studies typically correspond to medium run decisions such as travel mode, departure time, and route.

Sources of Endogeneity: What Generates the Chicken-and-Egg Problem?

Endogeneity has attracted the attention of economists and econometricians for decades, since it represents a major stake for policy evaluation. For an easy-to-read introduction on this topic, the reader is referred to [Greene \(2017\)](#). We detail below the main sources of endogeneity in the estimation of transport demand. In the transport context, the different causes of endogeneity can be illustrated by the choice of lane experiment analyzed by [Lam and Small \(2001\)](#). See [de Palma et al. \(2005\)](#) for an explanation of their method and results to a noneconometrician audience.

Common Determinants and Heterogeneity

Departure time was traditionally treated as given in the lane-choice model. Such a specification is misleading because departure time affects the explanatory variables (expected travel time, variability in travel time, and toll) on both lanes. Since unobserved factors may influence both departure time choice and lane choice, the explanatory variables and the error terms in the lane-choice equation may be correlated, which invalidates the standard regression model. For example, if your (well-paid) job starts at a popular time and requires punctuality, you will prefer both to travel at the height of the rush hour and to use the toll lanes to reduce the chance of delay. In this case, the unobserved factor, the type of job, induces a positive correlation between departure time and use of the toll lanes. This typically leads to overestimate the coefficients of the explanatory variables, and thus the Value of Time and the Value of Reliability.

The best way to solve this endogeneity, or chicken-and-egg problem is to include all the common determinants (here, the type of job) in the model estimated, so that their effect is not relegated to the error term.

Reversed Causality and Simultaneous Decisions

A typical example of reversed causality is when demand is induced by supply, and at the same time, supply is induced by demand, which is the case when transport policies aim at serving the demand. [de Palma et al. \(2005\)](#) stress that “Until the 1980s the UK approach to combat traffic congestion was to forecast demand, and then to add enough road capacity to accommodate it.” Such

“predict and provide” strategy is naturally undermined by the tendency of new roads to fill up with additional traffic demand induced by supply. In the short-medium run, this additional demand is mainly diverted from other routes, other modes, or other times of day. In the longer run, it mainly corresponds to new trips induced by additional activities (nonnecessary activities become interesting when congestion is reduced) or household and job relocation (households can locate farther away from the CBD if they anticipate less congestion, even though it finally turns out to be a bad decision when congestion increases again after a few years).

Interactions Between Transport and Urban Development

From the point of view of public policy, the question is not which comes first, but rather which one is the most sensitive to public policies, and more broadly to what extent is optimal public policy design affected by the mutual influences between transport, land use, and economic growth.

Land Use and Transport Interaction models aim at analyzing, measuring and predicting long term interactions between transport and land use, households and firms location, real estate markets, labor markets, economic activity and more generally urban development. They have received increased attention both in research and in practice over the last decades. Researchers and decision makers highlight the ability of these integrated models to examine the combined effects of land use and transport policies, to study the endogeneity of urban development and of travel patterns, as well as to analyze and quantify the effects of transport network expansion and policies in transport, dwellings, employment, or environment. They can be used to answer a large variety of important questions such as the following. What would be the optimal timing for building the different segments of the network over the years to come? What is the list and order of magnitude of the wider economic benefits? How many jobs will be created or attracted in the region in reaction to transport infrastructure development, in the short, medium, and long run? When and where will the different economic agents (residents and businesses) relocate after transport network extensions? This last question is important since it will help local authorities to cope with the demand for housing (especially regarding collective dwellings) and for offices. How will the traffic demand evolve, both in the existing and new public transport network, and on roads?

LUTI models differ in the way they model agent behavior and market (dis)equilibrium, in the degree of agent heterogeneity they can handle, in their level of geographic aggregation level and Geographical Unit of Analysis, in the complexity of their data requirements, and in the type of policies they can reasonably simulate. Different agents make different interrelated decisions, which systematically raise chicken-and-egg problems. Individuals/households choose tenure status (owner vs renter), dwelling type (single vs collective dwelling unit), residential location, activity, job location, job type and activity sector, and daily mobility (mode choice). Firms choose establishment location, number of jobs, wages, quality and quantity of goods produced, and prices. Public authorities (State, Region, Agglomeration, City) make choices related to Transport infrastructures, building permits, urban development policy, social dwelling and fiscal competition, using different tools (public investments, contribution to PPP, regulation). Other private decision makers and stake holders (transport companies, promoters, large firms, . . .) decide private investments, contribution to PPP, and lobbying. Real estate and labor market model the interactions between supply and demand, and the role of real estate prices and wages in reducing (possibly to zero) the gap between supply and demand.

LUTI models can be classified in two main categories with respect to their focus either on the dynamics of evolution of final point reached, their assumption of (general or partial) equilibrium versus disequilibrium, and with respect to their implications in terms of unicity versus multiplicity of equilibria.

The RELU-TRANS General Equilibrium model developed by [Anas and Liu \(2007\)](#) is representative of the first category. It assumes general equilibrium and is only interested in the final point reached by the system at equilibrium. Since its complexity is a convex function of the degree of agent heterogeneity, it is usually limited to 2–3 income classes, 2–3 classes of household composition, 3–5 activity sectors, and 2–3 qualification levels. It usually models either individuals or households, or implicitly assumes that each household is made of one representative member. It requires a rather limited amount of (aggregate) data.

The Urbanism disequilibrium model developed by [Waddell \(2002\)](#) is representative of the second category. It recognizes that markets are not at equilibrium and analyses the dynamics of the evolutions of the system. It imposes no limit on the degree of agent heterogeneity it can handle. It jointly models the behaviors of households (choosing dwellings and car ownership) and individuals (making decisions relative to jobs and commuting). It is therefore adequate for targeting and for evaluating the redistributive effects of policies: who will react the most to such policies, who will benefit most from the different policies, in the short, medium, and long run? What are the long-run effects of those policies on individual and household well-being (on this topic, see [Antoniou and Picard, 2015b](#))? However, such disaggregate models requires very rich individual data, and they require significant effort to be implemented in a new region.

Egg, Chicken, Hen and Cock, or Family Economics

The chicken-and-egg problem becomes even more complex (and at the same time analyzed in a more realistic way and allowing more accurate public policy evaluation) when one recognizes the nature of the decision process within the family.

The decision tree described in the previous sections is simpler for a single man or woman than for a couple, with or without children. In the first case, the individual has to decide (jointly or sequentially) where to live, where to work, whether to own a car, how to commute, and which route to choose.

In the second case, the couple has first to agree on a common residential location, which is not easy when the man's current job or future job opportunities are typically located far away from the woman's current job or future job opportunities. Then each spouse has to choose job location and job position, which may include a trade-off between a well-suited job located in a place, which will induce long commutes and a poorly-suited job located close to the residence. See Chiappori et al. (2018) for a detailed analysis. The couple then has to agree on car ownership and on car usage by each spouse. When the couple decides to own only one car (which is common in many European dense cities such as Paris, London, or Stockholm), this induces a competition between spouses to use it. Picard et al. (2018) show that taking into account such interactions within the family significantly improves the estimation of individual values of time.

Conclusions

The chicken-and-egg problem is a very old philosophical problem, which raises specific questions not only in most disciplines of academic research, but also in public policy implementation and evaluation and in practice.

The problem is that we usually observe the word as it is today, and we wish to understand what are the causes, which led to the current situation. Usually, there are many suspects, and causality is at stake. Fortunately, several econometric tools have been developed, which help to disentangle a beam of causalities, which includes subtle ones, as the value of the initial conditions. For example, why do cars and trucks drive on the right side in some countries and on the left one in other countries?

The chicken-and-egg problem of interest in this article is related to major snowball effects governing long-run evolutions in transport and urban development. The improvement of the transport networks attracts population, firms and commercial activity from outside the region and concentrates population in the vicinity of new station or roads. The newly installed firms and households attract even more jobs, providing them inputs or services. And some firms are attracted by other firms, generating wider economic benefits.

The increased local demand for dwellings, offices and commercial buildings increases real estate prices (as well as congestion and pollution), which in turn induces promoters to increase the local supply of buildings, which itself real estate prices. The settlement of new households and firms increases transport demand up to a point such that new improvements of the transport network are required. Population and employment densification generates agglomeration effects and increases productivity, which attracts even more firms and highly qualified workers. All these processes are combined over space and time and the question is who is responsible?

Acknowledgments

This article was partially written while I was working in CY, Cergy Paris University. It benefitted from the financial support of ANR-18-JPUE-0001 project MAAT, as well as from the ANR project AFFINITE.

References

- Abu-Rayash, A., Dincer, I., 2020. Analysis of mobility trends during the COVID-19 coronavirus pandemic: Exploring the impacts on global aviation and travel in selected cities. *Energy Res. Soc. Sci.* 68.
- Anas, A., Liu, Y., 2007. A regional economy, land use, and transportation model (RELU-TRAN©): Formulation, algorithm design, and testing. *J. Reg. Sci.* 47 (3), 415–455.
- Antoniou, C., Efthymiou, D., Chaniotakis, E., 2019. Demand for Emerging Transportation Systems. Modeling Adoption, Satisfaction and Mobility Patterns. Elsevier.
- Antoniou, C., Picard, N., 2015a. In: Bierlaire, M., de Palma, A., Hurtubia, R., Waddell, P. (Eds.), *Econometric methods for land use microsimulation Integrated transport land use modeling for sustainable cities*, EPFL Press, Ch. 12.
- Antoniou, C., Picard, N., 2015b. Urban sustainability and individual/household well-being. In: Michelangeli, A. (Ed.), *Quality of Life in Cities - Equity, Sustainable Development, Happiness from a Policy Perspective*. Routledge.
- Brueckner, J., Thisse, J.-F., Zenou, Y., 1999. Why is central Paris rich and downtown Detroit poor?: An amenity-based theory. *Eur. Econ. Rev.* 43 (1), 91–107.
- Chiappori, P.A., de Palma, A., Picard, N., 2018. Couple residential location and spouses workplaces. *WPElitisme2018-02*.
- de Palma, A., Picard, N., 2005. Route choice decision under travel time uncertainty. *Transp. Res. Part A: Policy. Pract.* 39 (4), 295–324.
- de Palma, A., Lindsey, R., Picard, N., 2005. Urban passenger travel demand. In: Arnott, R.J., McMillen, D.P. (Eds.), *Companion to Urban Economics*, ch16.
- Ginsburgh, V., Papageorgiou, Y., Thisse, J.F., 1985. On existence and stability of spatial equilibria and steady-states. *Reg. Sci. Urban Econ.* 15 (2), 149–158.
- Granger, C.W.J., 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* 37 (3), 424–438.
- Greene, W., 2017. *Econometric Analysis*, 8th ed. Prentice Hall.
- Lam, T.C., Small, K.A., 2001. The value of time and reliability: measurement from a value pricing experiment. *Transp. Res. E* 37, 231–251.
- Picard, N., Dantan, S., de Palma, A., 2018. Mobility decisions within couples. *Theory Decis.* 84 (2), 149–180.
- Picard, N., de Palma, A., 2019. Le modèle Urbansim, un outil d'analyse prévisionnelle de la localisation des emplois et de la population, in *Les effets économiques du Grand Paris Express*, Economica.
- Rodrigue, J.-P., 2020. In: *The Geography of Transport Systems*. 5th ed. Routledge, London, ISBN: 978-0-367-36463-2., p. 456.
- Sessa, C., Enei, R., 2009. EU transport demand: Trends and drivers ISIS, paper produced as part of contract ENV.C.3/SER/2008/0053 between European Commission Directorate-General Environment and AEA Technology plc; see www.eurtransportghg2050.eu.
- Sung, J., Monschauer, Y., 2020. Changes in transport behaviour during the Covid-19 crisis, IEA, Paris <https://www.iea.org/articles/changes-in-transport-behaviour-during-the-covid-19-crisis>.
- Waddell, P., 2002. UrbanSim: Modeling Urban Development for Land Use, Transportation, and Environmental Planning. *J. Am. Plann. Assoc.* 68 (3), 297–314.
- Kakutani, Shizuo, 1941. A generalization of Brouwer's fixed point theorem. *Duke Math. J.* 8 (3), 457–459.

Pricing Principles in the Transport Sector

Bruno De Borger*, Stef Proost†, *University of Antwerp, Antwerp, Belgium; †KU Leuven, Leuven, Belgium

© 2021 Elsevier Ltd. All rights reserved.

Introduction	95
First-Best Pricing: The Main Principles	95
Imperfect Pricing Instruments	96
Second-Best Pricing I: Uniform Pricing	97
Second-Best Pricing II: Not All Transport Services Can Be Priced	98
Only Part of a Network Can Be Priced	98
Not All Modes Can Be Optimally Priced	99
Pricing Only Freight (But Not Passenger) Transport	99
Pricing of Public Transport Services	100
Some Further Complications	100
Commuting Transport and the Labor Market	100
Distributive Issues	100
Conclusions	101
References	101
Further Reading	101

Introduction

In this paper, we review the pricing principles as they apply in the transport sector. We focus on socially optimal pricing in first-best and various second-best environments. Although the general principles we discuss in this paper apply universally to all transport modes and transport services, we illustrate the main principles using road transport as our main example. We briefly discuss application to other modes in the concluding section.

In the remainder of this paper, we first explain first-best pricing principles. Next, we point at important restrictions on pricing instruments that prevent these principles to be applied. We then turn to a series of second-best pricing rules that take into account that pricing in the real world may not allow the required differentiation to implement first-best rules, and that it may be infeasible to implement optimal pricing for all transport services on the complete network. Before concluding, we briefly consider some additional complications.

First-Best Pricing: The Main Principles

The main principles of optimal transport pricing are based on the important work of Dupuit and Pigou in the 19th and the beginning of the 20th century. Ignoring issues of income redistribution for the time being and assuming that governments face no restrictions on the pricing instruments they can implement, the first-best pricing principle states that transport services should be priced at marginal social cost: the price of a transport service (a car trip, a freight shipment by rail, etc.) should capture the marginal production cost for the operator plus the full marginal external cost generated by the service. In the absence of externalities, pricing at marginal production cost is socially optimal because it makes sure users are willing to pay the extra production cost, so that price can fulfill its role as a signal of scarcity. Moreover, to signal to transport users that their transport choices generate undesirable side effects, socially optimal prices should also internalize all external costs of congestion, pollution, global warming, noise, and accident risks. The optimal price equals the marginal production cost plus a “tax” that captures the marginal external cost of the transport service.

It is instructive to derive the optimal tax in a simple framework; this also allows a better understanding of various second-best policies described later. Consider a transport service (e.g., the use of a particular road of given capacity during peak hours) for which the inverse demand function is given by:

$$P(X) = a - bX$$

The generalized price of use of the road consists of both monetary expenses (fuel, etc.) and time spent in traffic; this in turn depends on how much traffic X there is. We specify the generalized cost as:

$$g(X) = d + cX + t$$

In this expression, d is the money plus time cost when traffic can flow freely (literally, at a zero traffic flow), and c is the slope of the congestion function; it captures how the time cost increases when traffic levels increase. Finally, t is a congestion tax or toll (which of course may be zero).

Economists capture social welfare in this simple framework by gross consumer surplus minus the generalized price (which depends on the traffic level, hence, on congestion), plus toll revenues for the government, minus the external costs other than congestion; we assume here that all users drive the same type of car; the marginal external cost is constant and given by e . The first-best toll is then defined as the one that maximizes:

$$\int_0^X P(x)dx - g(X)X + tX - eX$$

Differentiating with respect to t and using the equality of the generalized price and generalized cost in equilibrium, straightforward algebra shows that the socially optimal toll is:

$$t = cX + e$$

The toll equals the marginal external cost of congestion (cX) plus the marginal external cost (e) of other externalities (pollution, etc.). The marginal external congestion cost of an extra trip is the effect of a marginal increase in the traffic flow on the time cost, multiplied by the traffic volume.

The earlier discussion holds for one given transport service; moreover, all drivers were assumed to have the same value of time, so that they all perceived the same generalized cost. Of course, in practice many different services can be distinguished (different times of the day, different roads, etc.), and drivers may be very heterogeneous in terms of their value of time. The principle of taxing marginal external costs then implies that first-best taxes should be differentiated in time and space, they should take into account the emission characteristics of vehicles, and they should reflect the heterogeneity in values of time.

The tax or toll should, as much as possible, vary with the level of congestion; as a minimum, it should be higher during peak times than in off-peak traffic conditions (analogous to peak-load pricing in, e.g., the electricity sector), and it should differ between different types of roads. Ideally, since congestion varies continuously over time, socially optimal pricing requires time-dependent charges in function of real traffic levels. The bottleneck model, developed initially by William Vickrey in the 1960s, has been intensively used to study such time-dependent tolling systems. It analyzes how congestion builds up and declines again over the peak period. Moreover, the model takes into account that people adapt their trip timing in function of expected congestion and that, next to a cost associated with time losses in traffic, there is also a cost of being too early or too late at the final destination. Application of the model shows that a time-varying (or “fine”) toll performs much better than the optimal uniform toll over the whole peak period. The socially optimal time-varying toll on a single road is zero at the beginning and end of the peak period, and the toll level closely follows real-time congestion levels. It eliminates all congestion but does leave a substantial scheduling cost (more precisely, the cost associated with having to travel at an undesirable time due to congestion) in equilibrium. Scheduling costs are of the same order of magnitude as the time costs of congestion (Arnott et al., 1993).

Moreover, the ideal tax should differ between vehicles to capture their exact marginal external cost of emissions. This would require the ability to directly measure and tax emissions for all vehicles. Due to the difficulties of directly taxing emissions the literature shows that, although pricing is an efficient instrument to deal with congestion (which for a given road capacity mainly depends on the traffic flow), other externalities such as pollution may be reduced more efficiently using regulatory instruments (emission restrictions on vehicles, etc.). Even if regulatory instruments are used, it should be noted that pricing remains necessary to deal with remaining external (mainly congestion) costs.

Finally, first-best pricing should take into account the heterogeneity in time values. If users have different alternatives to reach their destination, users with low time values will be channeled toward the slowest alternative. The first-best pricing rules are then straightforward extensions of those absent heterogeneity; the social optimum implies different tolls on the two alternative routes, with a higher toll on the faster route. However, it should be noted that the efficient pricing solution may be highly undesirable from an equity perspective; it may even increase travel times for people with low time values while still requiring them to pay a toll on the slower road (Verhoef and Small, 2004). As always in economics, efficient outcomes are not necessarily equitable.

The first-best principle of marginal social cost pricing applies to all transport services, including road transport, rail service, airport use, etc. Of course, applying the principles requires taking into account the specific characteristics of the transport service considered.

Imperfect Pricing Instruments

Many currently used price instruments (fuel taxes, kilometer charges, and cordon pricing) do not allow policy-makers to sufficiently differentiate in space and time or according to vehicles’ emission characteristics. In [Table 1](#) we summarize the effect of some frequently used price instruments on the different external costs of transport. In the last column, we comment on the ability of the instrument to differentiate according to congestion and environmental externalities.

Restrictions on instruments necessitate second-best deviations from the marginal social cost pricing principles outlined earlier. Moreover, deviations from first-best pricing also apply in many practical instances when governments can only tax some, but not all, transport services. We briefly discuss a number of relevant second-best pricing rules, using a simple model with linear demand and generalized costs as illustration.

Table 1 Effect of pricing instruments on external costs

Price instrument	Effect on congestion	Effect on environmental externalities	Effect on accident risks	Impact on government revenues	Comments
Electronic road pricing	++	+	+	++	First best if differentiated in time and space
Fuel taxes	+	++	0	++	No time differentiation; allows differentiation between fuel types and fuel efficiency
Kilometer charges	+	+	+	++	No time differentiation, no fuel type differentiation
Cordon pricing	++ (in cities only)	+	+	+	No differentiation according to distance, inappropriate regional differentiation
Parking charges	+	+	+	+	Very imperfect instrument to deal with congestion; little effect on emissions
Public transport subsidies	+	+	+	--	Cost of funds, second best
Ownership taxes	0	+ if differentiated to emission characteristics	0	0	Little effect on congestion, poorly related to external costs
Subsidies to clean vehicles	0	++	0	--	Costly policy, no congestion effects

Second-Best Pricing I: Uniform Pricing

Uniform pricing instruments such as fuel taxes and distance-related taxes (kilometer charges) do not allow appropriate differentiation in time and space and according to vehicle type. For purposes of concreteness, consider a setting with two time periods (e.g., peak and off-peak) with clearly different marginal external costs. Assume that the government wants to set a socially optimal uniform tax per kilometer, that is, an equal tax in both periods. To capture differences in congestion between periods of the day on a given road infrastructure, it can be shown that the optimal tax per kilometer can be written as a weighted average of the marginal social costs in the two periods. The weights reflect the relative price elasticities of demand in the two periods. To minimize distortions, the common variable price will be closer to the marginal social cost in a given period the larger the relative price sensitivity of demand in that period.

To illustrate the pricing rule, denote the peak and off-peak by subscripts p and o , respectively. Linear demands in the two periods are $P_i(X_i) = a_i - b_i X_i$; $i = p, o$. Let generalized costs be given by:

$$g(X_i) = d + cX_i + \tau; \quad i = p, o$$

Here, τ is the uniform toll per kilometer in both periods. Ignoring noncongestion externalities for simplicity, the optimal toll solves:

$$\text{Max}_{\tau} \int_0^{X_p} P_p(x_p) dx_p + \int_0^{X_o} P_o(x_o) dx_o - (g(X_p))X_p - (g(X_o))X_o + \tau(X_p + X_o)$$

One then shows that the optimal uniform toll is given by:

$$\tau = s_p(cX_p) + (1 - s_p)(cX_o)$$

where the $0 < s_p < 1$ (more precisely, $s_p = \frac{\epsilon_p X_p}{\epsilon_p X_p + \epsilon_o X_o}$; $\epsilon_i = \frac{\partial X_i}{\partial \tau} \frac{\tau}{X_i}$, $\frac{\partial X_i}{\partial \tau} = -\frac{1}{c + b_i}$). This shows the optimal time-independent toll is a weighted average of the relevant marginal extern congestion costs; the weights depend on the price sensitivities of peak and off-peak demands with respect to the toll.

Similar pricing rules govern the optimal uniform tax when the same tax applies to two types of vehicles that differ in emissions per kilometer. Consider for concreteness the optimal fuel tax, when different vehicles have different emission characteristics. Similar analysis shows that the optimal tax is a weighted average of the external costs of different cars, where the weights depend on the sensitivity of the demand for car use of each type with respect to the tax. The implication is that, if "dirty" cars are more price sensitive than "cleaner" cars then the optimal uniform tax should be higher.

Finally, as will be argued later, due to the strong complementarity of commuting transport and labor supply, governments may want to impose a lower tax on commuting transport than on noncommuting transport. Here again the same principle applies: the inability to differentiate tolls according to trip purposes implies a uniform toll on both commuting and noncommuting that is a weighted average of the optimal differentiated tolls.

Second-Best Pricing II: Not All Transport Services Can Be Priced

In many practical cases, deviations from first-best principles apply because it is not feasible to optimally tax all transport services. Suppose, for example, that only a subset of roads (highways, some major roads) can be tolled. Alternatively, some governments may want to tax freight, but not passenger, transport. Finally, one may wonder how public transport should be priced depending on whether or not car use is charged for external congestion and pollution costs.

Only Part of a Network Can Be Priced

Very commonly governments cannot toll the complete network but are limited to charging tolls on a subset of important roads. Several cases can be distinguished here.

First, consider a simple network of two parallel roads between an origin and a destination, one of which can be tolled while the other remains un-tolled. This may be due to technical difficulties when tolling the whole network, or it may be deliberate government policy: in some countries (e.g., France) tolling requires an available un-tolled alternative. A major insight is that the optimal toll on the tolled road equals the marginal external congestion cost on that road minus a fraction of the marginal external congestion costs on the un-tolled one. If congestion costs on the un-tolled road are severe, this may easily result in a negative (in practice, a zero) toll.

To illustrate, assume two roads connect a given origin and destination (e.g., two cities); the roads may differ in capacity, design, etc. Denote the tolled road as A , the un-tolled alternative as road B . The demand for trips between the two cities is given as:

$$P(X) = a - bX; \quad X = X_A + X_B$$

Generalized costs of the two alternatives are, where the toll on road A is denoted t_A :

$$g_A(X_A) = d_A + c_A X_A + t_A$$

$$g_B(X_B) = d_B + c_B X_B$$

When people select a route only in function of the generalized cost, the equilibrium conditions are:

$$a - b(X_A + X_B) = d_A + c_A X_A + t_A = d_B + c_B X_B$$

which imply the following:

$$\frac{dX_A}{dt_A} = -\frac{(b + c_B)}{\Delta} < 0; \quad \Delta = b(c_A + c_B) + c_A c_B > 0$$

$$\frac{dX_B}{dt_A} = \frac{b}{\Delta} > 0$$

The toll reduces demand on the tolled road and diverts traffic to the un-tolled road. Focusing on congestion externalities only, we solve:

$$\text{Max}_{t_A} \int_0^X P(x) dx - (g_A(X_A))X_A + t_A X_A - (g_B(X_B))X_B$$

We find the optimal toll as:

$$t_A = (c_A X_A) - \delta (c_B X_B); \quad 0 \leq \delta = \frac{b}{b + c_B} \leq 1$$

The toll on road A is below the marginal external congestion cost on that road; setting the toll lower limits diversion of traffic toward road B , which is not tolled but becomes highly congested when many drivers try to avoid the toll on A . How much the toll should be below the marginal external cost depends on the price sensitivity of demand for the use of road A (captured by the parameter b) and on the slope c_B of the congestion function in B . If demand is very elastic (b small, so that δ is small) then a small reduction in the toll on road A is sufficient to limit traffic diversion toward the un-tolled road. On the contrary, if demand is not price sensitive (large b , hence large δ), a much larger toll reduction on road A is needed to prevent too much traffic diversion toward road B . Moreover, the earlier expression implies that if congestion on B is severe it may actually be better not to toll road A at all: the optimal toll can easily become zero or negative. To see this note that, if c_B becomes large, the second term in the earlier expression may exceed the first one for two reasons: the marginal external congestion cost $c_B X_B$ on road B becomes large, and δ approaches one. It does not make sense to toll road A if the main effect is to cause very much congestion on road B .

The earlier example was very stylized, and it has been generalized in different dimensions. One extension considers a general network consisting of an arbitrarily large number of nodes and links, where some but not all of the links can be tolled. The second-best optimal toll on a given link should then be a weighted average of the sum of the generalized marginal external costs, minus the

tolls paid on other links, for the relevant path flows. The weights are increasing in the elasticity of the path flow to prices in the second-best optimum. Another extension accounts for heterogeneity in time values. This does not affect the finding that the toll is less than marginal external cost. A final extension is to use a bottleneck approach and allow for time-varying tolls. In a setting with homogenous drivers using a network that consists of a tolled road and an un-tolled alternative (say, a two-lane highway with a toll-free lane), optimal second-best pricing implies a time-varying toll on the pay lane, combined with a subsidy that does not depend on time. Intuitively, the time-varying toll is imposed to eliminate queuing, and the subsidy is granted to attract users to the road whose capacity is efficiently used. In the absence of the subsidy, the pay lane has lower social marginal costs. This makes it desirable to attract drivers from the free lane, hence the subsidy also. Allowing for heterogeneity in time values does not affect these main principles.

Second, another example of partial pricing occurs when cities implement cordon tolls, whereby drivers pay a toll when entering a well-defined cordon around the city center. In this case, part of the network (namely, the area outside the cordon) remains unpriced. To illustrate the implications for the optimal cordon toll in the simplest possible way, suppose that some people commute from the suburbs to the city center. Denote their demand for trips by X . These people all drive into the city cordon but also use the un-tolled suburban network. A second group of commuters lives closer to the city; they only use the city network but not the suburban roads. Denote their demand by Y . Let the inverse demand functions be:

$$P^X(X) = a_X - b_X X; P^Y(Y) = a_Y - b_Y Y$$

Generalized costs of using the city (subscript “city”) and suburban network (subscript “suburb”) are given by, respectively (the cordon toll is denoted τ_c):

$$\begin{aligned} g_{\text{city}}(X + Y) &= d_{\text{city}} + c_{\text{city}}(X + Y) + \tau_c \\ g_{\text{suburb}}(X) &= d_{\text{suburb}} + c_{\text{suburb}}X \end{aligned}$$

Equilibrium requires:

$$\begin{aligned} P^X(X) &= g_{\text{city}}(X + Y) + g_{\text{suburb}}(X) \\ P^Y(Y) &= g_{\text{city}}(X + Y) \end{aligned}$$

As expected, the effect of the cordon toll on both types of demand is negative; we find:

$$\begin{aligned} \frac{dX}{d\tau_c} &= \frac{-b_Y}{\Delta} < 0; \quad \Delta = (b_X + c_{\text{suburb}})(b_Y + c_{\text{city}}) + c_{\text{city}}b_Y > 0 \\ \frac{dY}{d\tau_c} &= \frac{-(b_X + c_{\text{suburb}})}{\Delta} < 0 \end{aligned}$$

Maximizing welfare, we solve:

$$\int_0^X P^X(x) dx + \int_0^Y P^Y(y) dy - (g_{\text{city}}(X + Y))(X + Y) - (g_{\text{suburb}}(X))X + \tau_c(X + Y)$$

The optimal cordon toll can be derived as:

$$\tau_c = c_{\text{city}}(X + Y) + c_{\text{suburb}}X \left(\frac{b_Y}{b_Y + b_X + c_{\text{suburb}}} \right)$$

The toll exceeds the marginal external congestion cost in the city. The reason is that it also corrects for the un-tolled suburban road; it includes a fraction of the marginal external cost on the suburban road, too.

Not All Modes Can Be Optimally Priced

When multiple modes are considered, not all of which can be optimally priced, the same principles hold as in the case only part of the network can be tolled. Consider, for example, the competition between car and public transport use. Denoting the two modes as A and B this is conceptually the same problem as the two-road problem discussed before. If for some reason car use does not appropriately charge for external congestion costs, then it is second-best optimal to also price public transport below its social cost. We return to public transport pricing later.

Pricing Only Freight (But Not Passenger) Transport

In some countries it has been politically difficult to optimally toll both passenger and freight transport. If only freight can be taxed, what is the optimal tax to be implemented? Higher taxes on freight transport will typically raise the demand for passenger

transport: freight taxes reduce demand for freight, reducing congestion and attracting more passenger cars. Under plausible conditions, if passenger transport does not fully pay its marginal external costs, then the optimal toll on freight transport is below its marginal external cost as well. Still, only raising the tax on freight transport (but not on passenger transport) will be a good idea from a welfare perspective, unless passenger transport is severely under-taxed and the tax reform attracts substantially more passenger transport.

Pricing of Public Transport Services

In the absence of first-best pricing on congestible roads, reducing public transport fares below marginal social cost can be a legitimate second-best policy. The general idea that public transport can serve as a tool to alleviate the deadweight loss of suboptimally priced road congestion is by now widely documented, and large optimal subsidies have been reported. The standard second-best argument for lower than social cost prices is reinforced by the marginal benefits resulting from Mohring effects: low fares stimulate demand and this induces operators to raise frequency; this generates an external benefit in that it reduces passengers' average waiting times at public transport stops.

However, there are several caveats to this result. First, the argument is only relevant if the cross price elasticity of the demand for car use with respect to the public transport fare is sufficiently large relative to the own price elasticity of public transport demand. Second, public transport is subject to on-vehicle crowding: this increases boarding time, slowing travel speed; it also causes discomfort to others, and this disutility implies that public transport users generate a negative externality for other users. Third, governments may be willing to charge car use for its external congestion costs. When private transport is optimally priced, public transport pricing should be based on the same principles as pricing for road use. This implies that the second-best argument for low fares disappears, and fares therefore should also take into account the marginal external cost of an additional user. When road use is correctly priced and there is important crowding in public transport, the need for large subsidies is therefore much less clear. For example, evidence suggests that congestion pricing and dedicated bus lanes (absent bus subsidies) are much more efficient than bus subsidies ([Basso and Silva, 2014](#)). Some studies even find that even at current (too low) road charges, it is best to increase peak period prices of public transport on some highly crowded links. Moreover, fares should be increased in general whenever road pricing is put in place. Some studies argue that crowding implies that during peak and off-peak periods it is strongly welfare enhancing to differentiate public transport prices and vary frequency of service. Peak pricing helps to cover the costs of supplying capacity. When peak prices are too low, there is an excess demand for capacity and with a low revenue base it is difficult to improve or expand public transport services. The best overall transport policy reform includes higher toll levels on car use, higher peak bus fares and frequency, and free off-peak bus services offered at lower frequencies. These policies therefore definitely do not necessarily imply higher bus subsidies.

Some Further Complications

Commuting Transport and the Labor Market

Many congestion problems are due to commuters driving to work. The question arises whether commuting should, for this reason, be treated differently than other trip purposes. The literature suggests that the desirability of charging marginal external congestion cost to drivers and allowing toll reductions for commuters depend both on the sensitivity of labor supply with respect to the toll and on the distortive effects of existing labor taxes ([Van Dender, 2003](#)). If labor taxes are too high from a social perspective then it is optimal to differentiate tolls on commuting and other trip purposes, with lower tolls for the journey-to-work. This holds both for car and bus travel. Depending on the level of the labor tax relative to its socially optimal value, it may be optimal to not toll commuting at all. If labor and transport taxes are jointly optimized, transport taxes on car use increase substantially; commuting is still taxed somewhat lower than other trip purposes.

These findings do not justify general subsidies to commuting. They justify commuting subsidies (in the sense of lower tolls compared to other trip purposes) only on condition that congestion is taxed and given that current labor taxes are higher than socially optimal. Absent congestion charges subsidies for commuting are strongly welfare-reducing. These results generally hold both with competitive and imperfectly competitive labor markets (e.g., in a setting where wages and/or employment are determined by bargaining between unions and employer organizations).

Distributive Issues

There is some discussion whether distributive issues should matter for setting optimal transport prices. A standard partial equilibrium result is that, conditional on external costs and price elasticities, services used to a large extent by relatively low-income groups may be priced at a lower rate. However, general equilibrium models capturing the link between transport and labor markets suggest that the level of the externality tax does not depend strongly on distribution concerns, as re-optimization of the other taxes available to the government ensures that the income distribution objective is reached.

Conclusions

In this paper, we very briefly reviewed some important principles of socially optimal pricing of transport services. Although we illustrated these principles mainly for road transport, first-best and second-best pricing rules based on marginal external costs apply to all transport services, including road transport, rail service, airport use, etc. In practice, applying the principles does require taking into account the specific characteristics of the transport service considered. Ports and airports are part of a logistic chain, and external costs may be associated with each part of the chain (e.g., hinterland congestion); rail stations and airports have limited slot capacity, and congestion may strongly depend on slot allocation mechanisms.

Finally, given distortions elsewhere in the economy, it should be noted that the welfare effects of optimal pricing strongly depend on the use of the revenues. This also largely determines the political acceptability of optimal pricing schemes (De Borger and Proost, 2012). Take road pricing as an example. Many drivers will be worse off, and unless the revenues are used to partly compensate them and to offer good alternatives (better public transport, biking paths, etc.) to road use, proposals for the adoption of optimal prices will be rejected.

References

- Arnott, R., de Palma, A., Lindsey, R., 1993. A structural model of peak-period congestion: a traffic bottleneck with elastic demand. *Am. Econ. Rev.* 83 (1), 161–179.
- Basso, L., Silva, H., 2014. Efficiency and substitutability of transit subsidies and other urban transport policies. *Am. Econ. J. Econ. Policy* 6 (4), 1–33.
- De Borger, B., Proost, S., 2012. A political economy model of road pricing. *J. Urban Econ.* 71, 79–92.
- Van Dender, K., 2003. Transport taxes with multiple trip purposes. *Scand. J. Econ.* 105 (2), 295–310.
- Verhoef, E., Small, K.A., 2004. Product differentiation on roads: constrained congestion pricing with heterogeneous users. *J. Transp. Econ. Policy* 38 (1), 127–156.

Further Reading

- Braid, R.M., 1996. Peak-load pricing of a transportation route with an unpriced substitute. *J. Urban Econ.* 40 (2), 179–197.
- Brueckner, J.K., 2002. Airport congestion when carriers have market power. *Am. Econ. Rev.* 92, 1357–1375.
- Daniel, J., 1995. Congestion pricing and concentration at large hub airports: a bottleneck model with stochastic queues. *Econometrica* 63, 327–370.
- De Borger, B., Proost, S., Van Dender, K., 2008. Private port pricing and public investment in port and hinterland capacity. *J. Transp. Econ. Policy* 42, 527–561.
- De Palma, A., Lindsey, R., 2000. Private roads: competition under various ownership regimes. *Ann. Regional Sci.* 34, 13–35.
- Knittel, C., Sandler, R., 2018. The welfare impact of indirect Pigouvian taxation: evidence from transportation. *Am. Econ. J. Econ. Policy* 10 (4), 211–242.
- Mayeres, I., Proost, S., 1997. Optimal tax and public investment rules for congestion type of externalities. *Scand. J. Econ.* 99, 261–279.
- Vandenberg, V., Verhoef, E., 2011a. Winning or losing from dynamic bottleneck congestion pricing? The distributional effects of road pricing with heterogeneity in values of time and schedule delay. *J. Public Econ.* 95, 983–992.
- Vandenberg, V., Verhoef, E., 2011b. Congestion tolling in the bottleneck model with heterogeneous values of time. *Transp. Res. Part B Methodol.* 45, 60–78.
- Vickrey, W.S., 1963. Pricing in urban and suburban transport. *Am. Econ. Rev. Papers Proc.* 53, 452–465.
- Vickrey, W.S., 1969. Congestion theory and transport investment. *Am. Econ. Rev. Papers Proc.* 59, 251–260.

Long-Run Versus Short-Run Valuations

Stefanie Peer, Vienna University of Economics & Business, Vienna, Austria

© 2021 Elsevier Ltd. All rights reserved.

Introduction	102
Distinguishing Between Short-Run, Medium-Run, and Long-Run Choices	102
Definitions	102
Literature	103
Empirical Evidence on Differences in Valuations Depending on the Temporal Scale	104
Implications	104
Discussion and Outlook	105
References	105
Further Reading	105

Introduction

Monetary valuations of travel-related attributes are crucial inputs to appraisals of infrastructure investments. In particular, the valuation of travel time savings, which indicates how much travelers are willing to pay for reducing travel time, is often a strong determinant of the benefits associated with infrastructure investments, as travel time gains usually constitute a major share of the expected economic benefits (Small, 2012). Other valuations relevant for appraisals are valuations of schedule delays, reliability, and comfort.

This chapter discusses the evidence for a divergence between valuations depending on the time frame during which travel behavior can be adjusted, and discusses possible implications and avenues for further research. Recent literature has provided evidence that travel-related valuations may depend on the temporal dimension of the underlying decision. For instance, valuations may depend on whether a travel time saving occurs only on a specific day without notification (short-run), repeatedly (medium-run), or permanently (long-run) (e.g., Peer et al., 2015; Beck et al., 2017; Peer and Börjesson, 2018).

Most transport policies and investments lead to permanent changes to which users can adapt to over long time spans. In contrast, most valuations are derived from hypothetical choices made in situations (Peer and Börjesson, 2018) with a short-run framing, leading to a divergence between short-run and long-run valuations, which have immediate policy implications. Choosing the right valuations as inputs for appraisals is crucial for obtaining unbiased results. Dependency of valuations on the temporal dimension may also explain to some extent why the valuations of trip-related attributes vary so widely across studies and data sources.

The section “Distinguishing between short-run, medium-run and long-run choices” provides an overview on how short-, medium-, and long-run choices can be distinguished, and how they are reflected in the available literature. Section “Empirical evidence on differences in valuations depending on the temporal scale” summarizes the relevant empirical findings regarding the temporal dimension being able to explain differences in valuations. Also alternative explanations for divergent findings are discussed. Section “Implications” discusses the implication of the findings. Section “Discussion and outlook” summarizes and provides some avenues for further research.

Distinguishing Between Short-Run, Medium-Run, and Long-Run Choices

Definitions

Travel-related decisions can be categorized along multiple dimensions. One dimension is the temporal dimension of a specific decision, or more precisely, the time frame during which behavior can be adjusted.

In models of industrial organization, the long run is defined as the time frame when all input factors are variable (often defined as time spans exceeding 5 years), whereas in the short run it holds that at least one input factor is fixed. Similarly, in the travel context, the long run can be defined as the time frame in which persons have the flexibility to change all travel-related attributes, most importantly the duration of the commute. Primarily, they can do so by changing their home location or work location, but also buying a car (if they have not owned one before) may have a similar effect. These decisions are not easily reversible and can under normal circumstances not be changed in the shorter run.

The medium run can be defined, as the time frame in which the outcomes of the long-run choices are fixed, while other decisions, most importantly travel routines and times for meetings that can be irregular, are still endogenous. Travel routines may for instance concern the usual schedule for the commute from home to work (and vice versa), but also a more general activity plan, which includes trip chains, such as dropping off and picking up children at/from school on the way to work. Routines may, for instance,

also refer to the usual transport mode used for a specific trip purpose. The medium run is often defined as a timeframe of more than 1 year but less than 5 years, however, depending on the context alternative delineations of the medium run from the short and long run may be more useful.

In the short run, not only the outcomes of the long-run choices but also the travel routines chosen in the medium run are fixed. Individuals thus have already made an activity plan for a specific day, and then decide on their day-specific travel behavior subject to the long- and medium-run decisions they have made. Short-run decisions are thus usually defined such that they concern a specific trip (or trip chain); for instance, a short-run decision may concern scheduling aspects (choosing for a specific departure and/or arrival time, respectively), route choice or a choice between competing transport modes. Deviations of short-run decisions from routines are often driven by updated information regarding travel time expectations (Peer et al., 2015). Short-run deviations from medium-run choices thus occur primarily if day- and/or trip-specific factors play a role. For example, bad weather or special events may raise the expectation of longer travel times for the commute. In order to still arrive at the usual arrival time at work, a person may then decide to depart earlier than usual from home. Persons are expected to trade-off the benefits resulting from deviations from their medium-run choices (e.g., routines) against the costs of the deviation from the routine.

A distinction between long-run, medium-run and short-run time frames in travel-related decisions is also consistent with findings related to the relevant demand elasticities, for instance, concerning fuel prices or transit prices: these elasticities tend to increase (in absolute terms) over time, as travelers become more flexible (Litman, 2004) in adjusting decisions. Long-run transit elasticities have been found to be 2–3 times higher than the corresponding short-run elasticities.

Finally, it should be noted that the distinction between long-, medium-, and short-run decisions is not uniquely defined, but can be context specific. For instance, the housing choice might be a medium-run choice for many students (who tend to be fairly flexible), but constitutes a long-run choice for most families.

Literature

There exists a wide body of empirical literature on short-run travel choices, such as route, departure time, and mode choices for a specific trip. They yield monetary valuations by including trade-offs between a monetary component and other attributes. In the last decades, most of valuation studies have been based on stated preference (SP) data, and hence on hypothetical choices. Although these SP studies often do not make the short-run framing explicit, the choice context tends to be implicitly short-run, as respondents are asked to consider a specific trip in their choices, including the applicable scheduling constraints and activity schedule planned for that day. This is especially true for studies published in the past decade, many of which are driven by the idea that framing SP experiments around a specific reference situation (referencing) renders it easier for respondents to take into account (scheduling) constraints. In turn, the SP choice situations are expected to be more realistic and thus easier to answer for respondents, in turn reducing the presence of hypothetical biases. This reasoning has, however, been disputed in more recent studies (e.g., Hultkrantz and Savsén, 2018). Moreover, as argued below, the “referencing” approach and the resulting short-run re-scheduling and reference-dependence might be leading cause of divergence between short-run and long-run valuations.

Medium-run choices, in particular the choice of travel routines, have received comparatively little explicit attention in the transport economics literature. Travel routines have mostly been discussed in the context of (transport) psychology, where routines are often being characterized by little deliberation rather than the outcomes of a conscious choice process, implying that the underlying trade-offs based on which valuations can be derived have rarely been modeled explicitly. Only recently, it has been shown that also routines follow a rationale that allows for the derivation of valuations. For instance, the usual arrival time at work has been shown to result from the trade-off between regular (time-of-day-specific) travel times and schedule delays relative to an intrinsically (long-term) preferred arrival time (Peer et al., 2015). Although often not explicitly referred to as medium-run valuations, it can be argued that valuations derived from revealed-preference (RP) data, for instance from trip diaries, may reflect medium-run behavior (e.g., Schmid et al., 2019). Such data often reflect the usual, equilibrium behavior of individuals in situations where long-run choices are fixed. So far, only few SP studies have explicitly adopted a medium-run perspective, for instance by explicitly stating that the choice concerns a regular trip and that the choice will only be effective at some point in time, rather than immediately. The lack of such SP studies may be related to difficulties to ensure that respondents indeed interpret the choices as intended by the designer of the SP experiment: they may ignore the medium-run context and fill in the choice experiment as if it referred to a specific trip on a specific day.

Long-run choices regarding residential and workplace location as well as car ownership have been researched extensively. Most of them find that commuting time has a significant effect on residential and employment choices. The same holds for other transport-related factors, such as the availability of alternative transport modes at the home location. The effect of travel-related factors on long-run decisions has been shown both in situations with real-life data as well as in SP choice experiments that present for instance trade-offs between wage rates and commuting time, or between characteristics of the residential/workplace location and transport-related attributes (e.g., Van Ommeren et al., 2000). Although for similar reasons as discussed in the context of medium-run choices, SP experiments are not very common. In SP experiments with a long-run framing (and to a lesser extent also in experiments with a medium-run framing) it is also uncertain how respondents interpret changes in (travel-related) costs and/or revenues (salary), in particular because income effects as well as discounting may play a role.

Empirical Evidence on Differences in Valuations Depending on the Temporal Scale

In general, evidence on differences between short-run and longer-run valuations can be derived from unrelated studies that investigate transport-related choices in different time frames. However, due to the presence of a large number of potentially confounding factors, we focus on evidence derived from studies that make a direct comparison between short-run and longer-run valuations, but we additionally provide evidence from indirect comparisons when useful.

The following paragraphs attempt to summarize some patterns that have become evident in the relevant literature. These in particular concern travel times and to a lesser extent schedule delays and reliability. Given the limited number of studies and their heterogeneity in terms of research design (among others, regarding the underlying data source, transport mode, decision type and definition of the time frame), results cannot be regarded yet as conclusive.

The main result that can be generalized across most studies that directly compare short-run and longer-run valuations is that travel time is valued higher in a longer-run context (such as the medium-run choice of a travel routine, or the long-run trade-off between wage rate and commuting time) compared to a short-run context (including trip-specific departure time and mode choices) (Peer et al., 2015; Beck et al., 2017; Peer and Börjesson, 2018). Also a comparison between SP- and RP-based valuations of travel time savings that have been elicited in a similar context, sometimes even by the same travelers, yields a pattern consistent with long-run valuations of time being higher: RP data, which tend to reflect longer-run equilibrium situations, tend to imply higher valuations than SP data, which usually have a short-run focus as discussed earlier. A plausible explanation for the higher travel time valuations in the long run may be that the time gained from reduced travel time can be used more productively when arrival and departure times are known well in advance and if the travel time reduction occurs on a regular basis. It is realistic to assume that a small, unexpected, one-off time gain in the short run is more difficult to use productively, as travel time cannot be accumulated and saved (unlike money) compared to an expected, repeatedly occurring time gain of the same size.

The lower short-run values of time are also consistent with short-run rescheduling efforts with respect to a (short-run) reference point, introduced by the short-run framing of most SP experiments (Peer and Börjesson, 2018). This is particularly true if respondents are asked to make their choices while thinking about a specific trip, including the time constraints and activity schedule planned for that day. Short-run re-scheduling and the related reference-dependence are both undesirable, as they shed strong doubt on respondents answering to SP questions in a way that uncovers their long-run reference-free stable preferences, which are essential for economic welfare analysis. This is particularly important, as usually the objective of eliciting the preferences of respondents is to obtain valuations for investments and policy appraisals with a long time horizon.

The existing evidence for schedule delays and reliability is more mixed. In some studies, schedule delays and reliability are found to be valued higher in the short run, possibly reflecting that scheduling restrictions tend to be more binding in the short run, and re-scheduling is costlier (e.g., Peer et al., 2015). Other studies have found the opposite pattern, bringing forward similar arguments as for the higher value of time in the long-run context (Peer and Börjesson, 2018). Unreliability in the long-run context may also be higher due to it causing more anxiety over a longer period of time.

A general observation is that studies comparing shorter- and longer-run valuations based on SP data may have shortcomings regarding the comparability of the valuations. These are also brought forward by the authors of these studies, who in some cases provide alternative explanations for their findings, indicating that the differences in valuations may be attributable to other factors than the temporal dimension. For instance, the sensitivity to the cost component, which may also vary between the short run and the long run, may at least partially drive differences in valuations. Differences in the sensitivity to money over time can, however, rarely be distinguished from the response scale, which may also differ between short-run and long-run experiments, but without a clear a-priori expectation whether it is higher in SP experiments with a short-run framing or those with a long-run framing: reasons for a lower response scale (and hence for choices being less deterministic from analyst's point of view) in short-run experiments may be that they are difficult to answer, in particular if they require short-run rescheduling. Longer-run experiments may have a lower scale as long-run decisions are taken only at a low frequency and respondents may hence be less familiar with the corresponding trade-offs. Moreover, long-term decisions are potentially affected by a multitude of different factors, which cannot all be included in a standard SP experiment, leading to a larger random component.

Implications

The evidence of differences between short-run, medium-run, and long-run estimates (in particular, of the value of time) implies that preferences may not be stable in time, which is problematic as stable preferences are a key assumption in welfare economics. Most transport policies and investments tend to have long time horizons of several decades, for which long-run, reference-free stable preferences are the relevant input. This is in stark contrast to the fact that most recent valuation studies are based on SP experiments that adopt a short-run framing, and are hence likely to also elicit short-run valuations. This suggests a mismatch between the way the valuations are derived and their usage in appraisals. More precisely, it suggests that valuations used in the appraisal of policies and investments should reflect whether they lead to permanent changes (e.g., road capacity expansions, construction of new routes) or whether changes occur only under certain circumstances (e.g., improvements in incident management).

Differences in valuations between short-run and longer-run choices may also have implications for the optimal design of transport policies. For instance, pricing instruments required to reach the social optimum have been shown to differ considerably

from the standard (short-run) solution of the bottleneck model once the distinction between a medium-run choice of travel routines and short-run scheduling preferences regarding departure times is taken into account (Peer and Verhoef, 2013).

Discussion and Outlook

The earlier discussed summary of the relevant literature has shown that short-run and longer-run valuations of travel attributes have been found to diverge. In particular, valuations of travel time tend to be higher (Beck et al., 2017; Peer and Börjesson, 2018) in the longer run. This has been shown in SP studies that compare valuations elicited in experiments with a short- and long-run framing. Also indirect evidence from studies that compare valuations derived from SP data, with the underlying experiment typically having a short-run focus, to those derived from RP data, which tend to represent a more longer-run equilibrium situation. Evidence for other attributes (schedule delays, reliability) is quite mixed.

The direct comparison between short-run and longer-run valuations has only gained interest by researchers in the past few years. So far, only a limited number of studies that focus on such a comparison have been published. They differ widely in terms of the underlying research design, including data source (SP vs. RP), decision type (scheduling choice; trade-off between time and cost; etc.) as well as the main transport mode (car, train) considered. As a result, the evidence on differences between long-run and short-run valuations is not conclusive.

The limited number of studies and the (at least partially) inconclusive results render the topic of this chapter an interesting and promising avenue for further research. First, future research should focus on confirming the validity of the results obtained in the existing literature by replicating existing approaches and testing whether they also hold in different contexts. Second, also methodological advancements are required. In particular, it would be important to develop research designs that have a high comparability between short-run and long-run designs. For SP-based studies, the main challenge remains to phrase and frame the experiments such that respondents interpret the choice situations in the same way as intended by the researcher (hence, either as long-run or short-run choice) and react to it as they would react in reality. The latter is of course a more generic concern related to SP data, including data that are collected for the analysis of travel behavior. In particular, the medium- and long-run framing of choices is an important avenue for future research, as most existing SP experiments adopt a short-run framing. It is challenging to design SP experiments such that they closely reflect the complex decision processes underlying, in particular, long-run decisions regarding residential and work location.

Once evidence on differences in valuations depending on the time frame of the underlying decision has converged, differences in valuations across time dimensions may also inform theoretical and simulation studies on the optimal design of policies. For instance, most of these models so far assume that successive days are replicas in terms of travel times, implying that medium-run choices of routines and short-run choices of departure time are identical and cannot be disentangled. The same is true for stochastic models when they assume that travel time expectations are constant across days, and hence no learning or expectation updating takes place (Peer et al., 2015). Also in this case, short-run choices will not change between days, and again, medium-run and short-run behavior coincides.

In the relevant literature, different time horizons of travel-related decision making have mostly been considered separately, despite the underlying inter-temporal optimization that travelers engage in. A reason for the lack of models that represent such inter-temporal optimization processes may be the inherent complexity of these processes. Future research should aim at yielding a better understanding of the inter-temporal dynamics that travel choices are based on, making explicit, how these dynamics affect the travelers' response to transport, but also environmental and land-use policies, and in turn, the characterization of appropriate first- and second-best policies. Ignoring the embedment of most decisions in a wider framework of decisions across inter-temporal dimensions might lead to wrong conclusions regarding policy design and appraisal.

References

- Beck, M.J., Hess, S., Cabral, M.O., Dubernet, I., 2017. Valuing travel time savings: a case of short-term or long term choices. *Transp. Res.* 100, 133–143.
- Hultkrantz, L., Savsin, S., 2018. Is "referencing" a remedy to hypothetical bias in value of time elicitation? Evidence from economic experiments. *Transportation* 45 (6), 1827–1847.
- Litman, T., 2004. Transit price elasticities and cross-elasticities. *J. Public Transp.* 7 (2), 3.
- Peer, S., Börjesson, M., 2018. Temporal framing of stated preference experiments: does it affect valuations. *Transp. Res.* 117, 319–333.
- Peer, S., Verhoef, E.T., 2013. Equilibrium at a bottleneck when long-run and short-run scheduling preferences diverge. *Transp. Res.* 57, 12–27.
- Peer, S., Verhoef, E., Knockaert, J., Koster, P., Tseng, Y.Y., 2015. Long-run versus short-run perspectives on consumer scheduling: evidence from a revealed-preference experiment among peak-hour road commuters. *Int. Econ. Rev.* 56 (1), 303–323.
- Schmid, B., Jokubauskaite, S., Aschauer, F., Peer, S., Hössinger, R., Gerike, R., Axhausen, K.W., 2019. A pooled RP/SP mode, route and destination choice model to capture the heterogeneity of mode and user type effects in the value of travel time savings. *Transp. Res.* 124, 262–294.
- Small, K.A., 2012. Valuation of travel time. *Econ. Transp.* 1 (1–2), 2–14.
- Van Ommeren, J., Van den Berg, G.J., Gorter, C., 2000. Estimating the marginal willingness to pay for commuting. *J. Regional Sci.* 40 (3), 541–563.

Further Reading

- Rouwendaal, J., Meijer, E., 2001. Preferences for housing, jobs, and commuting: a mixed logit analysis. *J. Regional Sci.* 41 (3), 475–505.
- Schirmer, P.M., van Eggermond, M.A., Axhausen, K.W., 2014. The role of location in residential location choice models: a review of literature. *J. Transp. Land Use* 7 (2), 3–21.

Value of Noise

Jon P. Nelson, Pennsylvania State University, State College, PA, United States

© 2021 Elsevier Ltd. All rights reserved.

Glossary	106
Overview of Noise Cost Valuation	106
Hedonic Property Value Studies of Transportation Noise	107
Contingent Valuation Studies of Transportation Noise	109
Hedonic Price and Contingent Valuation Estimates: Benefit Transfers	110
Using Cost Values for Policy Questions	110
Coming to the Nuisance: Is Transportation Noise a One-Time Cost?	111
Biography	112
See Also	112
Relevant Websites	112
References	113
Further Reading	113

Glossary

A-weighted decibel (dBA) Weighted measure of decibels that accounts for the fact that the human ear is less sensitive to low-frequency sounds.

Coase's theorem The proposition that market trade will solve externality problems provided property rights are sufficiently well defined and enforceable, and negotiation or bargaining is not limited by high transaction costs.

contingent valuation (CV) Survey-based method for valuing market and nonmarket resources such as environmental quietude obtained by asking people to state their willingness-to-pay for more of a specified resource or willingness-to-accept less of a specified resource.

day-evening-night sound level (L_{den}) Average equivalent sound level exposure over a 24-h period with a 5-dBA penalty added for the evening hours of 19:00–22:00 and a 10-dBA penalty added for the nighttime hours of 22:00–07:00. Also called the community noise equivalent level or CNEL.

day-night sound level (DNL) Average equivalent sound level exposure over a 24-h period with a 10-dBA penalty added for nighttime hours of 22:00–07:00.

decibel (dB) Unit used to measure or compare the intensity of sound on a logarithmic scale, with the capital “B” included in recognition of Alexander Graham Bell.

hedonic price (HP) Measure of the implicit monetary value of characteristics or attributes that comprise a resource such as the implicit price or willingness-to-pay for the absence of noise as embodied in the market price of a house or apartment.

noise depreciation index (NDI) Method of expressing the effect of noise on property values or apartment rents when measured as the percentage reduction in value for each additional decibel increase in noise exposure.

Schultz curve Relationship between level of noise exposure and percent of the exposed population that is highly annoyed. First proposed and derived by acoustician Theodore Schultz in 1978.

stated preference (SP) method Survey-based methods for determining the monetary value of nonmarket resources such as expressed by contingent valuation surveys.

willingness-to-accept (WTA) Monetary measure of the minimum compensation required to forego a resource or accept a price increase. As a lump-sum amount, it is the compensating variation or increase in income required to accept a change in the status quo.

willingness-to-pay (WTP) Monetary measure of the maximum amount offered to obtain a resource or forgo a price increase. As a lump-sum amount, it is the equivalent variation or decrease in income required to avoid a change in the status quo.

Overview of Noise Cost Valuation

Noise is unwanted sound and vibration. Although by no means the only source, transportation is the dominant source of noise in areas near highways, major streets, railroads, and airports. Noise is an uncompensated cost to property owners and bystanders or, in the language of economics, an external cost of market activity. Costs are incurred by occupants of residential properties and others in the form of annoyances such as sleep disturbance, speech interference, disrupted outdoor activities, productivity losses at school and work, and possible health effects such as stress and loss of hearing. Prior to the early 1970s, noise was viewed generally as an environmental

nuisance to be dealt with by local and state authorities and the courts. The shift to more centralized control is illustrated by federal legislation in the United States such as the Airport and Airway Development Act of 1970, Federal-Aid Highway Act of 1970, Noise Control Act of 1972, and the Quiet Communities Act of 1978. Federal legislation requires application of social benefit–cost analysis as the basis for environmental decision-making. Early attempts to value noise in the context of benefit–cost analysis began with the Roskill Commission on the Third London Airport (1968–70), which leads to debate among economists on the definition and valuation of noise costs (Waters, 1975). In recent years, noise valuation has expanded to include considerations of access to transportation networks and related employment, environmental justice, and efficiency of markets for real estate. Consider the economic and political issues associated with a localized improvement project such as expansion of a major highway, rail line, or airport. Users of transportation incur benefits in terms of improved access, vehicle cost savings, and safer or more reliable travel. Secondary or community-wide impacts also may occur such as employment gains, pecuniary changes in prices of related commodities, changes in market competition, and economic growth, but these effects are transfers in the context of a social benefit–cost analysis. Costs of improvements include direct resources used for expansion and any uncompensated external costs that continue after project completion, including increased noise exposure. How should negative externalities be measured and valued? What is the willingness-to-pay for noise abatement? What is the willingness-to-accept compensation for increased exposure to noise? Absent an explicit market for quietude, economists have developed two main methods to value external costs from transportation noise. Revealed preference methods use observed changes in market prices of complementary resources such as the market for residential property. Primary studies using this method are labeled hedonic price (HP) studies, where “hedonic” refers to implicit values attached to attributes of housing such as environmental quiet. Stated preference methods use people’s responses to survey questions regarding willingness-to-pay for hypothetical changes in living conditions or changes in specific activities such as hiking in a public park that is free from overflights by aircraft. Primary studies using this method are labeled contingent valuation (CV) studies, where a survey is used to create an artificial market within which individuals purchase more or less of environmental services contingent on hypothetical prices. Both empirical methods have long histories in economics and both have advantages and disadvantages for valuation of transportation noise.

Hedonic Property Value Studies of Transportation Noise

Suppose a house or apartment is either located in a noisy residential neighborhood or it is not. The current level of noise is easily noticed or experienced by all market participants, but there may be uncertainty about future levels. The difference in values of identical residences in the quiet versus noisy neighborhood yields an implicit cost of noise, which is capitalized into the market value of houses or revealed by differences in monthly apartment rents. In principle, the discount attributable to noise compensates residents of noisy houses and apartments. That is, informed market participants reveal the value of noise as a result of bids and offers in the market for real estate after accounting for quality differences among properties. The sorting of buyers and sellers in a stable real-estate market produces an equilibrium outcome in which noisy residences tend to be occupied by people with a low willingness-to-pay for quietude (imperturbables) and quiet residences tend to be occupied by those with a high willingness-to-pay. A change in the noise environment such as construction of a new airport runway alters relative supplies of noisy and quiet residences and eventually leads to new market equilibrium. In practice, environmental conditions and market adjustments are more complex. Noise exposure varies with distance to sources and residents may be able to adjust to environmental changes by the so-called averting behavior, such as installation of soundproofing windows. Furthermore, relocation is not costless, and the initial response of residents may be to voice their concerns through political processes and protests rather than voting with their feet and exiting the noisy neighborhood. Differences in housing values and noise levels yield a noise discount that falls as distance from the source increases, but care is required by empirical researchers in order to obtain unbiased estimates. A variety of estimation issues and other special concerns exist.

Sound intensity for a single event is expressed in decibels (dB), which is measured on a logarithmic scale. Perceived loudness doubles for each 10 dB increase, which is a tenfold increase in sound intensity. Noise measurements for transportation sources rely primarily on cumulative noise energy exposure measured using the day-night average sound level (DNL) or similar exposure measures such as L_{den} . For traffic and railroad noise, it is also common to use measurements from the noise distribution such as the median sound level or L_{50} . Both approaches use an A-weighted decibel (dBA) to adjust for the human ear’s sensitivity to different noise frequencies. As a cumulative measure, DNL is the equivalent energy-averaged sound level in dBA during a 24-h period with a 10-dB penalty for nighttime noise. “Equivalent” means the acoustical energy of an average steady-state sound equals the energy from actual time-varying sounds. Early research by acousticians established a relationship known as the Schultz curve between noise exposure and percentage of the population classified as “highly annoyed.” The effective origin for the Schultz curve is not an ambient sound level of zero but some higher background value such as 55 DNL, which is the threshold used presently in the United States. At a given exposure level, aircraft noise is more annoying, followed by traffic noise and then rail. However, debate continues about which aspect of the noise distribution is most important for human annoyance with other duration and event measures suggested as appropriate, such as exposure time above a sound level of 65 dBA. Experimentation with different noise measures is a feature also of some property value studies. Formally, an HP estimating equation for the value of noise is derived in the following manner (Nelson, 1980):

$$V = b_0 Z^{b_1} A^{b_2} u_1 \quad (1)$$

where V is property value or rent; Z is a set of physical, locational, and neighborhood housing characteristics; A is subjective annoyance due to noise exposure; the b -terms are coefficients; and u is a stochastic error term or residual. Examples of housing

characteristics include square footage, proximity to employment, and school quality. Annoyance is expressed by a modified Schultz curve:

$$A = c_0 e^{c_1(DNL)} u_2 \quad (2)$$

where the c -terms are coefficients and e is the natural log base. Taking logs of Eq. (1) and substituting for $\ln(A)$ from Eq. (2) yields the following estimating equation:

$$\ln V = d_0 + d_1 \ln Z + d_2 DNL + u_3 \quad (3)$$

where the d -terms are reduced-form coefficients obtained after substitution. The coefficient $d_2(100)$ is the percentage change in a property value for a decibel change in noise exposure or the so-called noise depreciation index (NDI). Differentiating Eq. (3) with respect to DNL yields the marginal HP of noise equal to $d_2(V)$, which is increasing in V . An NDI of 1% implies a \$2,000 decrease per dB in the market value of a \$200,000 property, other things being equal, which is the present value of a future stream of costs. If the difference in noise exposure for quiet versus noisy neighborhoods is 10 dB, otherwise identical homes are priced at \$200,000 and \$180,000. At an annual real discount rate of 4% and 30-year lifetime, the annualized value per 10 dB is \$1157 per year. Debate exists regarding the constancy of NDI across noise levels and its transferability over time and space. Econometric regression techniques are used to ensure control of the large number of other features of residences that determine market values. Common estimation errors include ignoring ambient thresholds for human annoyance and using a log-transformation of noise exposure levels. There are several other issues that researchers deal with that are on the frontier of applied econometric research. First, Eq. (3) yields a marginal price, rather than an inverse demand or willingness-to-pay function. An important theoretical contribution by Rosen (1974) showed that a two-stage approach is necessary to estimate the latter function, but relatively few HP studies have estimated second-stage demand functions (Day et al., 2007; von Graevenitz, 2018). Potentially this is important if the area affected by a noise abatement project is large or noise sources are not localized. Second, estimation of HP equations may require accounting for housing submarkets if residents and residences are differentiated or segmented in some important manner such as private and public housing, resulting in nonuniform HPs. Unlike many consumer products, houses are durable, infrequently traded, and short-run supplies are relatively fixed for properties and characteristics. Use of GIS methods and large data sets often yield samples that cover substantial portions of an urban area, where submarkets are more likely to be encountered. Housing submarkets also are one possible way to identify second-stage demand functions as clusters of properties may occur along several important dimensions of housing, both structurally and locationally. Furthermore, HP functions for noise and other housing attributes may be nonlinear or estimated using binary dummy variables, which complicate identification of submarkets. More recent HP studies frequently employ spatial autoregression and autocorrelation methods to avoid inefficient coefficient estimates, which also correct for interdependence produced by submarkets (Allen et al., 2015; Cohen and Coughlin, 2008). If enough data are available, hedonic analysis can be conducted for properties sold both before and after an event or project, thereby creating what is closer to a natural experiment (Boes and Nuesch, 2011; Ossokina and Verweij, 2015; Palmquist, 1982).

Beyond concerns with econometric estimation, there are several related issues with important implications for application of HP results. Access to transportation or to transportation-related employment is likely to have positive effects on property values, which means empirical studies must attempt to isolate effects of noise from effects of accessibility. If noise and access were perfectly correlated, estimation problems might be easily resolved but usually this is not the case. Failure to account for accessibility will bias noise costs downward and might lead to incorrect noise policy conclusions. Fig. 1 illustrates the problem for an airport, where elongated contours reflect noise exposure and circles represent accessibility. Some sideline properties can have a high degree of access but experience low noise levels. New highways and railroad terminals also will heighten access for travelers and employees, but external costs can be incurred by a larger number of people. The policy dilemma is clear, as net beneficiaries cannot easily compensate those who are net losers.

A related issue is the ability of property markets to adjust to major changes in transportation networks. Informational asymmetry between sellers and buyers has the effect of underpricing of higher quality homes (lemons problem). However, major new facilities

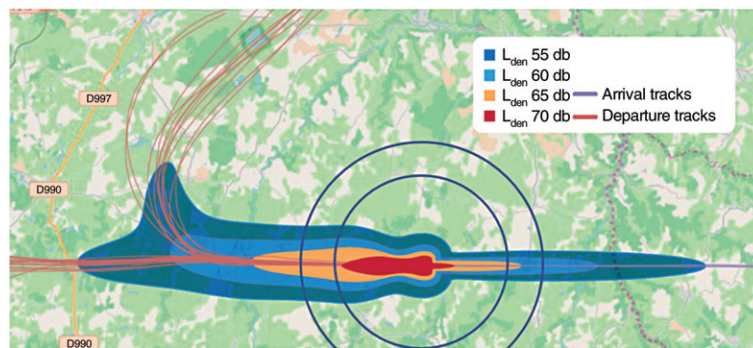


Figure 1 Airport noise contours and accessibility rings. Source: Adapted and reproduced with permission from European Aviation Safety Agency, 2016. European aviation environmental report 2016. EASA, Cologne.

usually are announced far in advance of construction, so there often is ample time for property markets to adjust to forthcoming events. For example, homebuyers have a rough idea of future noise levels from airport master plans. Many airports also use noise hotlines or have a noise ombudsman. At Melbourne's Essendon Airport, operators of aircraft are required to sign a Fly Neighborly Agreement. Some municipalities require that prospective homebuyers are provided with a noise disclosure statement. A study by Pope (2008) of pre- and post-disclosure property sales near a major airport found that the NDI increased by 55% after a disclosure requirement was mandated. The study concluded that HP estimates may be substantially biased if real-estate buyers are poorly informed about housing attributes. In a spatially efficient housing market, the Law of One Price should hold subject to allowance for relocation costs (Waights, 2018). That is, given a period of adjustment, land prices compensate for differences in residential amenities and disamenities as individuals self-select among locations according to willingness-to-pay. A study of the Zurich airport found that it took 2 years for apartment rents to stabilize at new equilibrium values after major changes in aircraft flight paths (Almer et al., 2017). However, a study of dwellings in Maarssenbroek, the Netherlands, found no relationship between noise sensitivity of residents and noise exposure, suggesting inefficient sorting (Nijland et al., 2007). Additional research is required on this critical issue.

A strength of the HP method for valuation of noise is that people's willingness-to-pay for quietude is revealed by actual choices made in market settings. Monetary values also are top down and avoid a potential double-counting problem in benefit-cost analysis that arises when summing across multiple effects of a project. One weakness is that specification of Eq. (3) for housing attributes is complex, including numerous locational variables as well as structural features. Measurement issues can arise such as correlation of noise exposure with other pollutants or with safety concerns. It is unlikely that housing prices capture long-term health effects, but fewer residences now exist at levels above 65 DNL. Average NDI values for aircraft, road traffic, and railroad noise are summarized later after discussion of CV studies.

Contingent Valuation Studies of Transportation Noise

The main alternative to HP studies is a stated preference study that uses survey-based responses to hypothetical choices between more or less of an environmental amenity or disamenity. In the first instance, focus is on maximum willingness-to-pay for an amenity and, in the second, minimum willingness-to-accept compensation for exposure to a disamenity. These potentially different monetary measures of welfare reflect differences in implied property rights assigned to either the status quo or the proposed environmental change. Many economists regard this choice as an ethical issue for society. In stated preference surveys, hypothetical choices are contingent on differences in costs associated with a payment mechanism, such as local taxes, monthly rents, or costs of travel. Hence, stated preference values are often labeled CV estimates. Thousands of CV studies have been conducted for environmental, health, transportation, cultural, and other issues. For transportation noise, there are several hundred CV studies, but most are for the United Kingdom and Scandinavia and very few for North America and other areas. Advantages of CV studies are that cost estimates can be based on both perceived and objective measures of noise; differences in costs can be examined for individual characteristics (age, income, family size, etc.); hypothetical settings can be employed that are unobserved in HP studies; and nonuse valuation is possible. For example, a CV study of traffic noise in Lisbon found that annoyance continued below noise levels used as ambient background levels for HP studies (Arsenio et al., 2006). A CV study of the Dallas-Ft. Worth Airport in the United States found that at severe noise levels many individuals were unwilling to consider a residence close to the airport and hence their willingness-to-accept dropped to zero (Feitelson et al., 1996). These results illustrate possible dangers of averaging across noise-exposed populations in HP studies. On the other hand, CV surveys lack the sorting effects that are implicit in property markets, suggesting CV estimates may be biased upwards as a consequence. The disadvantages of survey-based CV methods are well known: choices are hypothetical and may be simplistic or unrealistic; inexperienced respondents may not have well-defined preferences; and a variety of biases may distort estimates including starting point bias, framing bias, scope insensitivity, interviewer bias, and strategic bias in the form of invalid large responses or zero responses (protest zeros). The last problem is frequently dealt with by trimming the sample of responses in some manner. Furthermore, noise trade-offs are fundamentally a spatial issue and survey methods may lack this important dimension. That is, the survey should remind respondents that there are both substitutes and complements for a given abatement option. These problems might be overcome by appropriate sampling of respondents and by careful construction of the survey instrument. Combining HP and CV estimates also is possible such as concurrent estimates for both methods for a given location or through use of benefit-transfer methods. Formally, CV surveys may be phrased in terms of willingness-to-pay for improvements or willingness-to-accept deterioration of the residential environment. Consider the following questions that compare the status quo with its alternative: would you be willing to pay \$X extra each month in rent to live in a quiet environment? Or how much extra rent per month would you be willing to pay for a 50% reduction in nearby traffic noise? Or which of the following cost-amenity choices do you prefer? Alternatively, willingness-to-accept questions include: if your taxes fell by \$X would you be willing to relocate to a noisy environment? Or how much compensation would you require if the traffic volume and noise increased by 50%? Or if you could save \$X in travel costs per month would you be willing to live closer to a busy highway? These are proxy measures of noise exposure, but advocates of CV estimates sometimes claim this is an advantage of survey-based methods. Some questions are open-ended, while others follow a bidding game format. In some cases, the CV survey takes the form of a binary or yes-no choice containing only a few attributes, but complexity of the experiment may require that noise is considered along with other amenities and disamenities. Sometimes this is done to avoid disclosing the real purpose of the survey and thereby reduce concerns for strategic bias associated with local public goods, but it can also lead to embedding bias where responses on noise may include values for other attributes. Surveys have been conducted by mail, telephone, and in-person. A typical CV survey includes three sections: a description

of the environmental change, a section asking for monetary valuation of the change, and a set of debriefing questions for socio-demographic information about respondents. Following fundamental work by Carson (2012) and others, it is common to divide CV experiments into several phases that also can be used to judge the quality of existing studies: (1) pretest the survey instrument for its plausibility and understandability using focus groups or pilot studies; (2) determine that the population to be sampled and manner of payment are relevant for the proposed project or fit the institutional setting; (3) determine that the sample size is sufficient and properly weighted; (4) consider possible self-selection biases as these respondents are more likely to provide extreme estimates; and (5) include debriefing questions about why respondents answered certain questions in the manner that they did. Historically, neoclassical economists expressed doubt regarding the ability of behavioral surveys to produce truthful responses to complex questions regarding willingness-to-pay for both privately tradeable commodities and public goods. Although some objections have been addressed in CV studies, the most pressing issue is careful design of survey questionnaires where most economists lack training. For transportation noise values, this is one area where a concerted training program might yield substantial benefits. Although there is little work in this area, CV studies are one possible way to value health effects of noise or other nonuse values for which there is a willingness-to-pay such as moral sentiments and broader concerns for the environment.

Hedonic Price and Contingent Valuation Estimates: Benefit Transfers

A general problem in transportation economics is the use of values estimated for a given time and location that are applied to a policy issue in a different setting. Rather than conducting an entirely new study, investigators instead desire to use existing values by engaging in a benefit transfer (Nellthorp et al., 2007). Conducting a new primary study may be too costly, infeasible, or not readily available within allotted time. Benefit transfers thus use study results from one or more situations and extrapolate them to another similar situation. Economists have studied the circumstances under which a benefit transfer is possible and have constructed guidelines for doing so. A basic choice is between use of a single-value or point estimate and a function-based transfer. Use of function-based estimates has been shown to reduce transfer errors by a substantial margin (Kaul et al., 2013). Both types of transfers are aided by systematic literature reviews and meta-analyses. Systematic reviews attempt to answer a research question by collecting and summarizing all empirical evidence that meets a prespecified protocol for eligibility. Conducted in a critical manner, the results are useful for indicating strengths and weaknesses of each primary study, with a summary of evidence in the form of a range of values or average estimates. Attention is given also to the way the literature search is conducted with a focus on databases employed and keywords used in the search. Meta-analyses also are systematic reviews, but more advanced statistical procedures are used to summarize and analyze the primary study estimates. In the fixed-effect model, primary estimates are viewed as random draws from a common population of values and a weighted-mean is used to summarize the sample of estimates, with ideal weights defined by the inverse variance of each effect size. In the random-effects model, primary estimates are viewed as random draws from plural populations. Weighted-means again are employed, but ideal weights incorporate both the variance of each estimate and the between-effects variance due to sampling among multiple populations. Random effects thus allow for many unobserved factors that might underlie primary estimates and are preferred when heterogeneity is present but not easily explained. The best meta-analyses therefore obtain both a sample of similar effect sizes and the standard error or precision of each primary estimate. Meta-regressions can provide evidence on the extent to which estimates vary systematically according to both factual differences such as different countries or time periods and methodological differences such as different statistical methods employed in primary studies. Weighted-regressions again are appropriate. Table 1 displays estimates for costs of transportation noise from HP and CV studies obtained using systematic reviews, meta-analyses, and a few individual HP studies for rail noise.

The range of values is substantial, but this is not unexpected given the complexity of the problem. For average HP values, there is strong agreement across several reviews for each transport mode. For aircraft, there are over 50 primary HP studies dating back to 1967. The average NDI is about 0.70% per dB for aircraft noise with some suggestion of a slight increase since an early literature review in 1980. For road traffic, there are over 25 primary HP studies dating back to 1975. The average NDI is about 0.50% per dB for traffic noise with some suggestion of a slight increase since an early literature review in 1982. For rail noise, the number of HP studies is quite small and no systematic review exists. Studies for Berlin and Seoul indicate NDI values in the range 0.50%–0.65% per dB, which is not out of line with results for traffic noise. Results from CV studies also are summarized in Table 1, but these values are not directly comparable with HP estimates. Average CV estimates indicate a higher willingness-to-pay for abatement of aircraft noise compared to traffic and rail. For a smaller sample of estimates, the CV meta-analysis found that marginal values were much smaller below 55 dB and comparable for noise exposure in excess of that level. The analysis also concluded that use of decibel measures along with a description of annoyance yielded lower values than other survey approaches. In both HP and CV reviews, there are several attempts to examine variation across income levels. However, it has been demonstrated that the elasticity of income with respect to willingness-to-pay is not identical to conventional measures of income elasticity familiar to economists (Flores and Carson, 1997). This difference needs to be incorporated in future research on this issue.

Using Cost Values for Policy Questions

There are three major policy applications for cost values obtained from HP and CV studies. First, benefit–cost analyses of specific noise mitigation and abatement projects, including airport expansions, curfews and other quieting regulations, quieter aircraft and

Table 1 Transportation noise: hedonic price and contingent valuation studies

<i>Hedonic price studies and type</i>	<i>Transport mode</i>	<i>NDI values: range (no.)</i>	<i>Average NDI</i>
Kopsch (2016)—HP meta-analysis	Aircraft	0.28–2.30 (44)	0.67
Nelson (1980)—HP review	Aircraft	0.40–1.10 (12)	0.62
Nelson (2004)—HP meta-analysis	Aircraft	0.28–1.49 (33)	0.67
Nelson (2008)—HP review	Aircraft	0.20–2.44 (24)	0.74
Schipper et al. (1998)—HP meta-analysis	Aircraft	0.10–3.57 (30)	0.61
Wadud (2013)—HP meta-analysis	Aircraft	0.00–2.30 (65)	0.69
Bateman et al. (2001)—HP review	Traffic	0.08–2.22 (18)	0.55
Bertrand (1997)—HP meta-analysis	Traffic	NA (16)	0.64
Kopsch (2016)—HP meta-analysis	Traffic	0.00–2.22 (46)	0.49
Nelson (1982)—HP review	Traffic	0.08–1.05 (17)	0.44
Nelson (2008)—HP review	Traffic	0.18–1.50 (29)	0.54
Beimer and Maennig (2017)—HP Berlin	Traffic	NA	0.61
Beimer and Maennig (2017)—HP Berlin	Rail	NA	0.65
Chang and Kim (2013)—HP Seoul	Rail	NA	0.53
<i>Contingent valuation study</i>	<i>Mode</i>	<i>WTP range (no.)</i>	<i>Average</i>
Bristow et al. (2015)—CV meta-analysis	Aircraft	\$40–786 (69)	\$292
Bristow et al. (2015)—CV meta-analysis	Traffic	\$2–149 (141)	\$106
Bristow et al. (2015)—CV meta-analysis	Rail	NA (44)	\$26

Notes: NDI is noise depreciation index as a % per dB of a property value; NA indicates not available or not applicable; number of values in parentheses. Average values for CV studies are per dB per household per annum in 2009 USD.

Sources: Bateman, I.J., Day, B., Lake, I., Lovett, A., 2001. *The Effect of Road Traffic on Residential Property Values: A Literature Review and Hedonic Pricing Study*. Scottish Executive Development Department, Edinburgh; Beimer, W., Maennig, W., 2017. *Noise effects and real estate prices: a simultaneous analysis of different noise sources*. *Trans. Res. D* 54, 282–286; Bertrand, N.F., 1997. *Meta-Analysis of Studies of Willingness to Reduce Traffic Noise*. MSc. dissertation. University College, London; Bristow, A.L., Wardman, M., Chintakayala, V.P.K., 2015. *International meta-analysis of stated preference studies of transportation noise nuisance*. *Transportation* 42, 71–100; Chang, J.S., Kim, D.-J., 2013. *Hedonic estimates of rail noise in Seoul*. *Transp. Res. D* 19, 1–4; Kopsch, F., 2016. *The cost of aircraft noise—does it differ from road noise? A meta-analysis*. *J. Air Transp. Manag.* 57, 138–142; Nelson, J.P., 1980. *Airports and property values: a survey of recent evidence*. *J. Transp. Econ. Policy* 14, 37–52; Nelson, J.P., 1982. *Highway noise and property values: a survey of recent evidence*. *J. Transp. Econ. Policy* 16, 117–138; Nelson, J.P., 2004. *Meta-analysis of airport noise and hedonic property values: problems and prospects*. *J. Transp. Econ. Policy* 38, 1–27; Nelson, J.P., 2008. *Hedonic property value studies of transportation noise: aircraft and road traffic*. In: Baranzini, A., Ramirez, J., Schaefer, C., Thalmann, P. (Eds.), *Hedonic Methods in Housing Markets*. Springer, New York, pp. 57–82; Schipper, Y., Nijkamp, P., Rietveld, P., 1998. *Why do aircraft noise value estimates differ? A meta-analysis*. *J. Air Transp. Manag.* 4, 117–124; Wadud, Z., 2013. *Using meta-regression to determine noise depreciation indices for Asian airports*. *Asian Geogr.* 30, 127–141.

motor vehicles, noise barriers, improved roads and highways, and abandonment or relocation of transportation facilities. Second, overall evaluations of full social costs of transportation, which are attempted to determine all paid and unpaid costs of each mode of transport. Third, evaluation of alternative policy instruments, such as calculation of noise and congestion taxes or user charges. Continued refinement of HP and CV estimates of noise damage costs will aid these policy applications. As one example, Becker and Lavee (2003) studied social costs and benefits of highway noise reductions in Israel. They conducted an HP study for a sample of 280 sales of occupant-owned apartments in three major Israeli cities, with noise exposure measured on a continuous basis for each dwelling in the sample. NDIs were 1.2% and 2.2% per dB for urban and higher-income suburban areas, respectively, which are at the high end of HP values reported in Table 1 for traffic noise. Net total benefit values were derived for two changes in noise standards: 67 dB–64 dB and 64 dB–59 dB contingent on abatement options for road restructuring, noise barriers, or apartment insulation. Overall results were mixed, with net benefits varying importantly with respect to location, density of housing, and nearness to roads. For example, apartment buildings as close as 25 m to a noisy road never passed a benefit–cost test for abatement. The authors suggested that a uniform national standard for traffic noise may not be optimal in economic terms but might be required on normative grounds.

Coming to the Nuisance: Is Transportation Noise a One-Time Cost?

People move or relocate for various reasons including a desire to avoid external costs of various disamenities, such as transportation-generated noise. Suppose a homeowner relocates to the vicinity of an environmental nuisance and then sues the generator of external costs for damages (a classic example is building a home close to an existing pig farm). In tort law, this is known as “coming to the nuisance.” In a simple two-person case, the parties might negotiate to arrive at an economically efficient solution provided property (ownership) rights are clearly defined and enforceable. As pointed out by Coase (1960), an efficient solution may not occur if property rights are not clearly defined or transaction costs are high, which is surely the case in most environmental problems involving transportation noise. However, as previously discussed, noise costs are capitalized into property values and reflected also in lower apartment rents, which compensates movers for newly incurred costs when they locate close to a highway or airport. Is

transportation noise therefore a one-time cost that society might choose to ignore or downplay as a continuing cost? Is it efficient to focus on the first mover's use of a resource and assign complete property rights to that user? An extension of Coase's theorem indicates that an economic analysis is incomplete unless the following thought experiment is conducted (Cordato, 1998; Wittman, 1980): who should have been first given current values for alternative uses of a resource? What is the highest and best use of the resource today? Which use of land resources will maximize economic efficiency, either homeowners or the airport? Answers to these questions require benefit–cost calculations, but Coase's analysis is complicated by focus on complete allocation of resources over time for both homeowners and the airport, including costs of relocation, abatement options, and other alternative land uses. Estimates also must be discounted to present values. Some observers argue that the experiment requires a level of knowledge that is beyond capabilities of courts and governments since the information required is a moving target, that is, the ideal efficient outcome may not be obtainable due to real-world costs of analysis. Another objection is that the first-party's right to use a resource is rendered less certain, which reduces incentives for efficient resource use and investment. A first-mover rule does send a powerful signal to other parties that their choices should be made with extreme care, thereby lessening conflicting land uses and moral hazard problems. However, it is frequently argued that external costs such as noise tend to be regressive and fall more often on poorer, minority, or disadvantaged groups in society (Chakraborty, 2006; Sobotta et al., 2007). These distinctive groups may therefore object to proposed changes on equity or moral grounds, raising ethical issues regarding legal standing and environmental justice. Suppose the proposed project is a new highway that impacts an existing older neighborhood. If the homeowners have a legitimate right to an undisturbed neighborhood, then the relevant measure of potential damages is the amount of compensation required if the proposed change did occur, that is, their compensating variation in income or willingness-to-accept. A willingness-to-accept measure implicitly assigns ownership rights to the homeowners and asks how much they require to sell their rights. The measure of gain to potential users of the highway is their willingness-to-pay for construction at the proposed location. Alternatively, if rights are assigned to potential users of the highway, measurements are reversed and for homeowners the issue is their equivalent variation in income or willingness-to-pay to prevent construction. These comparisons show the close connection between the measurement of noise values for public policy decisions and the costs and benefits of noise exposure, which is in part a distributional issue. In either case, a benefit–cost analysis is required but the outcome of the analysis may depend on assignment of property rights to amenities and disamenities from transportation.

Biography

Jon P. Nelson is Professor Emeritus of Economics at the Pennsylvania State University, University Park, PA. His research focuses on issues of public policy including transportation regulation, environmental economics, antitrust and utility regulation, and public health. He is the author of 3 books and more than 100 peer-reviewed articles and book chapters.

See Also

The Concept of External Cost: Marginal versus Total Cost and Internalization; Value of Time; Pricing Principles in the Transport Sector; The Value of Life and Health; Demand for Passenger Transportation; The Pareto Criterion and the Kaldor Hicks Criterion; Social Discount Rates; Cross-Elasticities between Modes; Ethical Aspects—Can We Value Life, Health, and Environment in Money Terms?; Car tolls, Transit Subsidies for Commuting, and Distortions on the Labor Market; The Robustness of Cost–Benefit Analyses; Wider Economic Impacts of Transport Investments; Regulatory Reforms and Competition in Public Transport; How to Finance Transport Infrastructure?; Airline Economics; Computable General Equilibrium Analysis in Transportation Economics; Value of Life and Injuries; Road pricing; Location choice models; Peak period congestion and travel; Residential location models; Transport modes and accessibility; Transport modes and health; Airport regulation; Road network planning; Road traffic regulation; Railway line planning

Relevant Websites

US Airport Noise Law. Available from: <http://www.airportnoiselaw.org/>.
 European Aviation Safety Agency. Available from: <https://www.easa.europa.eu/eaer/>.
 European Commission, Noise. Available from: http://ec.europa.eu/environment/noise/index_en.htm.
 European Commission, Noise in Europe. Available from: https://ec.europa.eu/info/events/noise-europe-2017-apr-24_en.
 European Environment Agency. Available from: <https://www.eea.europa.eu/soer-2015/europe/noise>.
 US Federal Aviation Administration. Available from: <https://www.faa.gov/airports/environmental/>.
 US Federal Highway Administration. Available from: <https://www.fhwa.dot.gov/environment/noise/>.
 US Federal Highway Administration. Available from: https://www.fhwa.dot.gov/Environment/noise/regulations_and_guidance/.
 US Federal Interagency Committee on Aviation Noise. Available from: <https://fican.org/findings/>.
 US Federal Transit Administration. Available from: <https://www.transit.dot.gov/regulations-and-guidance/regulations-and-guidance>.
 World Health Organization. Available from: <http://www.euro.who.int/en/health-topics/environment-and-health/noise>.

References

- Allen, M.T., Austin, G.W., Swaleheen, M., 2015. Measuring highway impacts on house prices using spatial regression. *J. Sustain. Real Estate* 7, 83–98.
- Almer, C., Boes, S., Nuesch, S., 2017. Adjustments in the housing market after an environmental shock: evidence from a large-scale change in aircraft noise exposure. *Oxf. Econ. Pap.* 69, 918–938.
- Arsenio, E., Bristow, A.L., Wardman, M., 2006. Stated choice valuations of traffic related noise. *Transp. Res. D* 11, 15–31.
- Becker, N., Lavee, D., 2003. The benefits and costs of noise reduction. *J. Env. Planning Manage.* 46, 97–111.
- Boes, S., Nuesch, S., 2011. Quasi-experimental evidence on the effects of aircraft noise on apartment rents. *J. Urban Econ.* 69, 196–204.
- Carson, R.T., 2012. Contingent valuation: a practical alternative when prices aren't available. *J. Econ. Perspect.* 26, 27–42.
- Chakraborty, J., 2006. Evaluating the environmental justice impacts of transportation improvement projects in the US. *Transp. Res. D* 11, 315–323.
- Coase, R.H., 1960. The problem of social cost. *J. Law Econ.* 3, 1–44.
- Cohen, J.P., Coughlin, C.C., 2008. Spatial hedonic models of airport noise, proximity, and housing prices. *J. Reg. Sci.* 48, 859–878.
- Cordato, R.E., 1998. Time passage and the economics of coming to the nuisance: reassessing the Coasean perspective. *Campbell Law Rev.* 20, 273–292.
- Day, B., Bateman, I., Lake, I., 2007. Beyond implicit prices: recovering theoretically consistent and transferable values for noise avoidance from a hedonic property price model. *Environ. Resour. Econ.* 37, 211–232.
- Feitelson, E.I., Hurd, R.E., Mudge, R.R., 1996. The impact of airport noise on willingness to pay for residences. *Transp. Res. D* 1, 1–14.
- Flores, N.E., Carson, R.T., 1997. The relationship between the income elasticities of demand and willingness to pay. *J. Environ. Econ. Manage.* 33, 287–295.
- Kaul, S., Boyle, K.J., Kuminoff, N.V., Parmeter, C.F., Pope, J.C., 2013. What can we learn from benefit transfer errors? evidence from 20 years of research on convergent validity. *J. Environ. Econ. Manage.* 66, 90–104.
- Nijland, H.A., Hartemink, S., van Kamp, I., van Wee, B., 2007. The influence of sensitivity for road traffic noise on residential location: does it trigger a process of spatial selection? *J. Acoust. Soc. Amer.* 122, 1595–1601.
- Nellthorpe, J., Bristow, A.L., Day, B., 2007. Introducing willingness-to-pay for noise changes into transport appraisal: an application of benefit transfer. *Transp. Rev.* 27, 327–353.
- Nelson, J.P., 1980. Airports and property values: a survey of recent evidence. *J. Transp. Econ. Policy* 14, 37–52.
- Ossokina, I.V., Verweij, G., 2015. Urban traffic externalities: quasi-experimental evidence from housing prices. *Region. Sci. Urban Econ.* 55, 1–13.
- Palmquist, R.B., 1982. Measuring environmental effects on property values without hedonic regressions. *J. Urban Econ.* 11, 333–347.
- Pope, J.C., 2008. Buyer information and the hedonic: the impact of a seller disclosure on the implicit price for airport noise. *J. Urban Econ.* 63, 498–516.
- Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *J. Polit. Econ.* 82, 34–55.
- Sobotta, R.R., Campbell, H.E., Owens, B.J., 2007. Aviation noise and environmental justice: the barrio barrier. *J. Reg. Sci.* 47, 125–154.
- von Graevenitz, K., 2018. The amenity cost of road noise. *J. Environ. Econ. Manage.* 90, 1–22.
- Waters, A.A., 1975. *Noise and Prices*. Oxford University Press, Oxford, UK.
- Waight, S., 2018. Does the law of one price hold for hedonic prices? *Urban Stud.* 55, 3299–3317.
- Wittman, D., 1980. First come, first served: an economic analysis of 'coming to the nuisance'. *J. Legal Stud.* 9, 557–568.

Further Reading

- Freeman, A.M., 2003. *The Measurement of Environmental and Resource Values: Theory and Methods*, second ed. Resources for the Future, Washington, DC.
- Johnston, R.J., Rolfe, J., Rosenberger, R.S., Brouwer, R. (Eds.), 2015. *Benefit Transfer of Environmental and Resource Values: A Guide for Researchers and Practitioners*. Springer, New York.
- Nelson, J.P., Kennedy, P.E., 2009. The use (and abuse) of meta-analysis in environmental and natural resource economics: an assessment. *Environ. Resour. Econ.* 42, 345–377.
- Zerbe, R.O., Bellas, A.S., 2006. *A Primer for Benefit-Cost Analysis*. Edward Elgar, Cheltenham, UK.

The Value of Life and Health

Henrik Andersson, Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	114
Theory	115
Monetizing Safety Preferences	115
One-Period Model	115
Selected Predictions from the One-Period Model	117
Multiperiod Model	117
Empirical Methods	117
Revealed Preference Methods	117
Stated Preference Methods	118
The Contingent Valuation Method	119
Discrete Choice Experiments	119
Discussion	120
References	121
Further Reading	121

Introduction

Individuals as well as governments make decisions that reveal that they are prepared to forgo something else to reduce the risk of dying or being injured in traffic accidents. For instance, individuals may reduce their speed when driving with the purpose of reducing the risk of being involved in an accident, thereby increasing the time traveled, which is time lost for other activities. Governments may on their part decide to invest in safety measures that will reduce the risk of pedestrians at railway crossings, thereby reducing the amount of resources available for, for example, education. If individuals are well informed when they make their decisions, and their actions have no negative impact on others, then their actions will increase social welfare; they will only undertake actions that make them better off. Government actions are motivated by market failures, such as externalities, public good/bad, asymmetric information, etc., and since the government acts on behalf of the society its actions need to be evaluated.

We are in this chapter interested in traffic safety, but the models and methods discussed are applicable for other types of risks as well. There exist several approaches to evaluate safety measures. One approach would be to evaluate the measure based on the outcome in terms of less fatalities and injuries. That is, the measure with the largest reduction is the preferred measure. This of course ignores the cost of the measure and would risk prioritizing the most resource demanding ones, as it could be expected the more resources used the larger effects. An alternative approach that would take into account the costs would therefore be cost-effectiveness analysis (CEA). This approach would take into account the costs, relate them to the outcome, and provide an estimate of the cost per unit of output. One example would be the cost per life saved, that is, the cost of preventing one unidentified death in society. The CEA would result in a more efficient allocation of the resources since it would identify the more cost-effective safety measures or policies. However, CEA evaluates the technical efficiency of different measures/policies, but it does not take into account strength of preferences for the different policies, and cannot be used as a decision rule for resource allocation between sectors. This leads us to a third evaluation approach, which is the one usually favored by economists, that is, cost-benefit analysis (CBA).

CBA relies on both benefits and costs to be measures in a common metric, usually money. Since they are both measured in the same metric it allows for an analysis which provides information whether the policy is welfare improving for society. That is, if benefits exceed the costs, then the policy increases social welfare, if they do not then the policy will reduce social welfare. Since the foundation of CBA is based on welfare analysis, it is important that benefits and costs reflect their social value. For some of the benefits and the costs, it can be assumed that available market prices reflect their social value. However, for many other benefits and costs no market prices may exist. One example is the benefit of interest for this chapter, that is, traffic safety. As described earlier individuals value safety in the sense that they get a positive utility from being exposed to less risk (if we assume that they prefer to live compared to be dead, and prefer to be in good compared to bad health). However, there does not exist any easily available market price that reflects this value. Therefore in order for CBA to be able to take into account the benefits from reducing fatality and injury risks the value of safety needs to monetize. This monetization is done using the willingness to pay (WTP) approach. The aim of this chapter is to describe the theory behind this approach and some empirical methods used to obtain actual estimates.

The chapter is outline as follows. In the following section we describe the theory behind valuing safety by first providing a brief introduction before the main models. We then in the third section describe the main approaches and methods to elicit preferences for safety. The first part of that section focuses on methods using market data, whereas the second part focuses on methods that collect data using surveys. The chapter ends with a discussion and some suggestions for further reading.

Theory

Monetizing Safety Preferences

Before the WTP approach became widely accepted as the appropriate approach to monetize the value to society of reducing health risks the human capital (HC) approach dominated. Whereas the WTP is a preference-based approach the HC assumes that the value of an individual is reflected by his/her market productivity. For instance, if we illustrate the concept using the risk of fatality, the value of a life is estimated as the discounted expected future earnings, that is,

$$HC = \sum_{t=1}^T \frac{p_t y_t}{(1+r)^t}, \quad (1)$$

where p_t , y_t , and r are the survival probability conditional on surviving until t , the income in t , and the discount rate, respectively, and t and T define the time periods and end of life, respectively. The calculation is intuitive and straightforward, data necessary for the calculation are relatively easily available, and indeed estimates based on the HC approach are still used for policy purposes in some settings. However, its use for policy purposes is limited due to two major drawbacks: (1) nonmarket production is assigned a zero value, and (2) individual safety preferences do not enter the equation. The former suggests that unemployed and retired have a zero value. This seems both unethical and illogical, if society and its individuals also care for these groups, and it has been suggested to include also monetary values for nonmarket earnings, for example, services and goods produced in the household, and leisure time. This does not address, though, the main critique against the approach, that is, the second drawback. It has been shown that under some reasonable and plausible restrictions on the utility function the HC can serve as a lower bound of the value of a statistical life (VSL), which is a preference-based measure of the social value of preventing one statistical death in society, and hence in addition to not being a preference-based measure it will also underestimate the social value of safety. For these reasons it is out of favor among economists as the approach to elicit monetary safety values (unless for difficulty of getting necessary data to estimate preference-based measure the HC is the only approach available to obtain any monetary values).

We therefore focus on preference-based monetary safety measures in the reminder of this chapter. The concept and models described, as follows, can be used to monetize both fatality (mortality) and injury (morbidity) risks. Using fatality risk simplifies the presentation of the models, though, since it avoids the issues of severity and duration of injury - there is only one possible negative health outcome with fatality risk, that is, death - when describing the models. When describing the models, we will refer to the marginal WTP (MWTP) as the VSL. There is an ongoing discussion whether the best term for the monetary equivalent of preventing one statistical death is the VSL. Since the expression contains the word "life," it is sometimes misunderstood as measuring the value of saving an identified life. However, since VSL still is the main terminology used, we will also use it in this chapter. To stress that the VSL is not a monetary measure of saving an identified life the following example is illustrative:

Assume a region with a population of 100,000 in which six persons die each year in traffic accidents. The local government proposes a safety program that is expected to reduce the number of deaths in traffic accidents to 2 per year, that is, 4 fewer deaths per year. Suppose each individual in the region is willing to pay EUR120 annually for the proposed safety program to be implemented. This would mean that per statistical death prevented the population is willing to pay $\text{EUR}120 \times 100,000/4 = \text{EUR}3 \text{ million}$. That is, EUR12 million would be collected to save four lives, which would mean that the VSL is EUR3 million.

In the following two subsections, we first present the standard one-period model and then the multiperiod model for VSL. In each subsection, we also provide some of the main predictions of the models.

One-Period Model

Under the assumption that the individual marginal rate of substitution (MRS) and the personal risk change are uncorrelated, the VSL is the population mean of the MRS between fatality risk and wealth. The theoretical framework is a state-dependent utility model in which individuals are assumed to maximize their utility (Jones-Lee, 1974). Let p and w define the survival probability and wealth, and $u_s(w)$ the state-dependent utilities, where the subscript s defines staying alive ($s = a$) or dead ($s = d$). Individuals are then assumed to maximize,

$$EU(w, p) = pu_a(w) + (1 - p)u_d(w). \quad (2)$$

Utility functions are twice differentiable, and following the standard assumptions utility of wealth is higher if alive than dead, marginal utility of wealth is higher if alive than dead and nonnegative, and regarding financial risk individuals are weakly risk averse, that is,

$$u_a(w) > u_d(w), u'_a(w) > u'_d(w) \geq 0 \text{ and } u''_s(w) \leq 0. \quad (3)$$

Thus both the utility and the marginal utility are higher if alive than dead at any wealth level. Under these assumptions, the indifference curves over wealth and survival probability are decreasing and strictly convex. This is illustrated in Fig. 1.

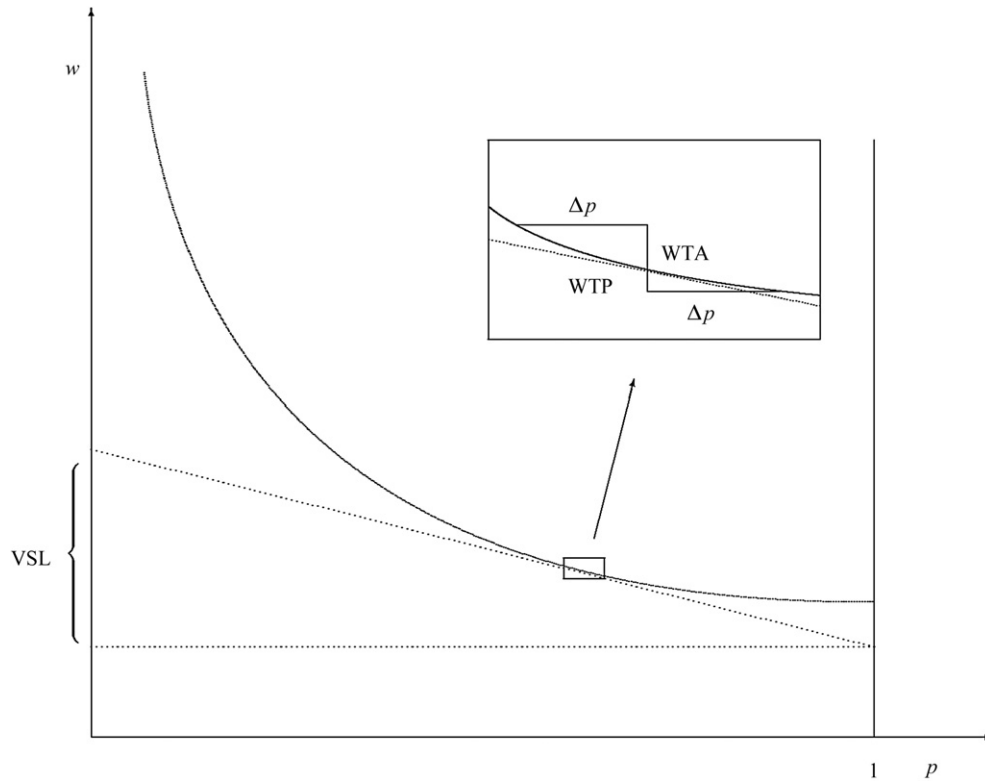


Figure 1 The value of a statistical life. Source: Lectures notes, Henrik Andersson, Toulouse School of Economics, inspired by lectures notes by James Hammitt, Harvard University.

The compensating and equivalent surplus, that is, the WTP and willingness to accept (WTA), for a change in the fatality risk $\Delta p \equiv \varepsilon$ can be derived using Eq. (2). Let Eq. (2) defines EU_0 and let the WTP for the risk reduction ε be denote by $C(\varepsilon)$, then $C(\varepsilon)$ is given by,

$$(p + \varepsilon)u_a[w - C(\varepsilon)] + (1 - p - \varepsilon)u_d[w - C(\varepsilon)] = EU_0. \quad (4)$$

The WTA for the risk increase ε can similarly be denoted by $P(\varepsilon)$, that is,

$$(p - \varepsilon)u_a[w + P(\varepsilon)] + (1 - p + \varepsilon)u_d[w + P(\varepsilon)] = EU_0. \quad (5)$$

From Eqs. (4) and (5), it is evident that the WTP and WTA will depend on the size of ε , with both increasing with the change in the risk. However, it is important to stress that since the size of ε will be small in empirical applications, we expect WTP and WTA to be close to equal and that they are near proportional to ε .

As explained, VSL is the MRS between wealth and mortality risk, and it can be obtained by taking the limit of WTP or WTA when $\varepsilon \cong 0$, and is defined as follows:

$$VSL = -\frac{dw}{dp} \Big|_{EU \text{ constant}} = \frac{u_a(w) - u_d(w)}{pu'_a(w) + (1-p)u'_d(w)}. \quad (6)$$

It is obtained by totally differentiating Eq. (2) and keeping utility constant. The numerator contains the utility difference and the denominator the expected marginal utility. The assumptions in Eq. (3) ensure that VSL is always strictly positive.

It not possible to always in empirical applications estimate WTP and WTA for a marginal change in risk. For instance, in surveys it is necessary to ask respondents about a small but finite risk reduction. VSL is then given by the ratio between the change in wealth, for example, WTP, and the change in risk, that is,

$$VSL = \frac{WTP}{\Delta p}. \quad (7)$$

As explained, WTP is near proportional to the size of Δp . Eq. (7) should therefore be interpreted as an approximation of the VSL.

Selected Predictions from the One-Period Model

The theoretical model plays an important role in examining the validity of the preference estimates in empirical studies eliciting WTP and WTA for risk reductions. In addition to the prediction of near-proportionality described earlier, the two main predictions are that VSL increases with wealth and decreases with baseline survival probability. That wealthier individuals are willing to pay more is intuitive and in Eq. (6) is driven by the numerator that increases with wealth and the denominator that is nonincreasing with wealth, as a result of the assumptions in Eq. (3). Regarding the effect from the baseline survival probability, it is sometimes referred to as the dead-anyway effect, intuition being that if at a high risk why not spend the wealth on reducing the risk, and is in Eq. (6) driven by only the denominator and a result of the assumption that $u'_a > u'_d$.

The theoretical framework earlier has been extended to examine how background risks, that is, independent or additive risks to the specific risk of interest, compared to an overall risk as presented earlier, health status, financial risk, or ambiguity aversion influence the VSL. These predictions are also important for empirical applications, but the ones described earlier are the main ones, and for brevity of this description of how to value safety, we refer to the further readings for the discussion of these other predictions (Andersson et al., 2019).

Multiperiod Model

The single-period model described earlier can be extended to a multiperiod model in which the individual is assumed to maximize the expected life-time utility given by,

$$EU_\tau = \sum_{t=\tau}^{\infty} q_{\tau,t} (1+i)^{\tau-t} u(c_t), \quad (8)$$

where τ , $u(c_t)$, i , and $q_{\tau,t} = p_\tau \dots p_{t-1}$ denote the point of reference, the utility of consumption at time t , the utility discount rate, and the probability at τ of surviving to t , respectively. The extension to a multiperiod model is of high relevance in health valuation since it allows for an examination of how latency, that is, a delay of the health effects from being exposed to a risk (e.g. air pollution) affects individuals' WTP to reduce the risk. When eliciting preferences for traffic safety, it is usually assumed that the effect is immediate, that is, there is no time delay between the exposure to the risk and the health outcome. That is, the negative effect from being involved in a car crash is immediate and an action taken to reduce the risk of the crash will be an immediate reduction in the probability of death (or injury). Therefore latency despite its relevance in health valuation in general is not of great interest in valuation of traffic safety.

The multiperiod models are still of high relevance since it also allows for the examination of how age may affect the WTP to reduce traffic risk. Intuition would suggest that WTP to reduce fatality risk declines with age, since an older individual (*ceteris paribus*) has less to gain from reducing his/her risk. Both theoretical and empirical research studies have shown that this is not necessarily true, though. Regarding the theoretical predictions, it has been shown that the relationship between age and WTP will depend on the individual's optimal consumption path over his/her life and this path will depend on the assumptions of the model. Hence, the relationship can be considered ambiguous.

The discussion earlier concerns a temporary risk reduction that lasts one time period, for example, a risk reduction for one year at the age of 50. However, a risk reduction may last over several time periods, or be permanent. Any time period, though, can be treated as a series of shorter time periods, for example, the annual risk reduction can be treated as a series of monthly risk reductions. This means that the WTP for the longer time period can be calculated as the sum of the WTP for the shorter time periods that make up the longer ones. The multiperiod model has been used to examine the effect from treating WTP as a sum of a series of WTP compared to a one-period WTP. This question is highly relevant, since empirical studies have, as a mean to make the risk reduction larger, used scenarios with longer time periods. The theoretical findings based on the multiperiod model showed that precaution should be taken since with too long time periods and a high discount rate, the difference between a one-period and a series of time-periods models can be nonnegligible (Andersson et al., 2013).

Empirical Methods

The two approaches to monetize traffic safety preferences are broadly defined as revealed preference (RP) or stated preference (SP) methods, where the former employ actual market decisions and the latter decisions in hypothetical scenarios (Freeman et al., 2014). The actual techniques used in either RP or SP studies are usually referred to as nonmarket valuation techniques, since they monetize preferences in cases, like traffic safety, where easily available market prices are nonexistent. In the following sections, we will briefly describe some of the main approaches used to monetize traffic safety preferences.

Revealed Preference Methods

Discrete choices are one example of observed behavior. These includes, for example, whether to buy and/or use a bicycle helmet, use the seat belt, and use a reflector. Individuals will buy and/or use the safety equipment only if the benefits exceed the cost. Regarding those not using the equipment available, for example a bicycle helmet, their behavior does not suggest that they

do not have preferences for safety, only that the cost of buying and the disutility of using it are higher than the benefit of the risk reduction. Eq. (7) can be rearranged to show the necessary relationship between the benefit and the cost in a discrete choice situation,

$$WTP < VSL \cdot \Delta p. \quad (9)$$

The nature of the discrete choice therefore means that the estimates obtained from individual decisions among those using the safety device reflect a lower bound of their WTP. However, it may not be a lower bound of the population's WTP since among those who decide not to use the safety device the WTP is lower than the cost of using it. Hence, it is necessary to also take into account the part of the population not using the device when estimating the value to be used for policy purposes.

Data from discrete safety choices are informative for policy purposes. The relatively simple choice situations, the facts that choices are real, and that many choices are made on more than one occasion (i.e. respondents are familiar with the good) provide strong arguments for the use of this kind of data. However, caution should be taken. One important distinction to consider is the difference between expenditure and costs. For instance, a bicycle helmet can be used over a longer time period, which means that the analyst needs to be informed about the length of life of a bicycle helmet, which is not the expected length by the producer but the expected length of usage by the buyers. Other issues of concern are whether consumers indeed make well-informed decisions, any disutility of usage (which then is a cost), and how well informed the analyst is about not only the buyer's cost for and perceived risk reduction from the device but also the intended actual usage when the decision was taken.

Another example of an RP approach is the hedonic regression technique (Rosen, 1974). This technique investigates the relationship between the price of a good and its attributes. Let P and $Q = (q_1, q_2, \dots, q_k)$ be the price of the good and its vector of attributes, then the hedonic price function can be written as follows:

$$P = P(Q). \quad (10)$$

For example, assume that Q is a car then Eq. (10) states that the price of the car depends on its attributes, such as speed, comfort, etc. The underlying theoretical model for the hedonic regression approach, which in a competitive market assumes utility maximizing individuals and profit maximizing firms, shows that the MWTP for an attribute, q_k equals the marginal price change in the market for the same attribute, that is,

$$MWTP_{q_k} = \frac{\partial P(Q)}{\partial q_k}. \quad (11)$$

As indicated by Eq. (11), the hedonic regression technique examines the effect on the price, changing the level of one of the attributes holding the other attributes constant. Hence, if the examined attribute is the safety attribute then the increase in the price from an increase in that attribute can be interpreted as a safety premium. If the attribute q_k is adequately defined as a safety attribute of the car that reduces the risk of fatality in the event of an accident then Eq. (11) can be interpreted as the VSL of Eq. (6).

Central to the regression analysis is the functional form of the price function. Theory only states that the function should not be linear, but apart from that it is something to be decided empirically. Common functional forms in the literature have been the semi-log or the log-linear forms, that is, both using the log of the dependent variable. To illustrate, using the semi log the price function could be specified as follows:

$$\ln P_i = +X_i'\beta + \gamma_1 s_i + \gamma_2 q_i + \varepsilon_i, \quad (12)$$

where P_i is the price of the car i , X_i is a vector of car attributes (in addition to the risk variables), s_i is the probability of a fatal car accident, q_i is the probability of a nonfatal car accident, and ε_i is an error term. Based on the specification of the price regression in Eq. (12), the VSL is given by:

$$VSL = -\gamma_1 P, \quad (13)$$

where the minus sign is included, since the regression in Eq. (12) is defined using risk and not safety, and hence the minus sign converts the result to a positive VSL.

Stated Preference Methods

In SP studies respondents face hypothetical choice situations and are asked to make a decision as if it was real (Johnston, et al., 2017). The main weakness with the SP approach is that choices are not real, that is, respondents do not face the consequences of their decisions. However, the approach also provides advantages compared to the RP approach such as a control of the choice situations, since designed by the analyst, and flexibility, that is, the hypothetical choice situation can be tailored to the questions the analyst wants an answer to.

There exists a large range of different SP methods. In the following two sections we will briefly describe the two main methods, the contingent valuation method (CVM) and discrete choice experiments (DCE).

The Contingent Valuation Method

In the CVM data are collected through surveys, where respondents are asked to answer questions online, face-to-face, in paper questionnaire, etc. It is standard to collect information about socioeconomics and demographics, for example, income, education, age, to examine how such characteristics may influence the WTP, which can be used to both examine the validity of answers and for policy purposes. The core of the survey is the risk scenario and the WTP questions. To obtain valid estimates of respondents' preferences, it is important that the risk scenario provides an accurate description of the risk itself and how the risk reduction will be provided, for example by a public safety measure or as a private safety device, and that the scenario is understandable to the respondents. It is well established that individuals have difficulties understanding risk changes, and hence it is important to make sure that they understand the scenarios and find the questions relevant to answer.

The respondents may be asked to state their maximum WTP for the risk reduction directly in what is usually referred to as the open-ended format. The preferred format, by most analysts, is the referendum format, though. In the referendum format, respondents are asked to either answer yes or no to a question, where the risk policy and the cost to them of implementing it are described, or are asked if they would pay a specified amount (bid) for the risk policy or safety device. One reason why the referendum format is preferred is that it more resembles both voting and market scenarios, another is that it under some circumstances is incentive compatible, that is, respondents have incentives to state their true preferences. Following is an example of a referendum-format question.

Assume that the national transport authority will invest in the road network that will reduce the number of fatalities in the road network. Would you be willing to pay EUR200 for a policy that would imply that four fewer persons will die next year as a result of road accidents?

Yes ☐ No ☐

In this example, the information obtained will reveal whether the respondents' WTP is at least, or below, EUR200 for the risk reduction. To obtain a more precise range of the respondents' WTP, a follow-up question where the bid is increased if the respondents accept the first bid and lowered if they decline it is often included.

Regarding the empirical analysis of data from open-ended CVM studies, it is straightforward to use standard regression techniques, where details of the risk scenario like the size of the risk reduction and information about individual characteristics are used as explanatory variables for the respondents' stated maximum WTP. The referendum format provides closed-ended data, which can be modeled, using a latent variable framework. Let X_i be a vector of individual characteristics and the latent (unobserved) WTP can be specified as follows:

$$WTP_i^* = \alpha + X_i' \beta + \varepsilon_i, \quad (14)$$

where ε_i is an error term. Assuming that ε_i be normally distributed with mean 0 and variance σ^2 , the probability that respondent i accepts a bid with value r_i is then given by,

$$P(WTP_i^* > r_i) = P(\alpha + X_i' \beta + \varepsilon_i > r_i) = P(\varepsilon_i > r_i - \alpha - X_i' \beta) = 1 - \Phi\left(\frac{r_i - \alpha - X_i' \beta}{\sigma}\right), \quad (15)$$

where Φ denotes the standard normal CDF. The α , β and σ parameters can be estimated by maximum likelihood. Empirical evidence suggests that model specifications may influence the estimated WTP, and hence extensive sensitivity analysis should be conducted. One option is also to analyze the data, using nonparametric methods such as the Turnbull estimator.

Discrete Choice Experiments

In DCE the respondents face hypothetical choice situations with two or more alternatives. They are asked to choose between these policies and are usually also provided the option to opt out, that is, choosing the status quo. Compared to the CVM policies in DCE usually contain more than only two attributes (in the earlier-mentioned CVM example, the two attributes were the risk reduction and its cost). To illustrate, we provide an example of a choice situation, often referred to as a choice set, in Fig. 2.

It is evident from Fig. 2 that DCE can be considered as an extended version of the referendum format CVM. That is, if respondents were only to choose between the current situation and Policy B, that is, Policy A is removed from the choice set, and if injury risk was assumed not to be affected by Policy B, that is, the attribute related to injury risk was also removed from the choice set, then the choice situation in Fig. 2 would be identical to the CVM scenario earlier. The choice situations in DCE will be more complex for the respondent, but by including more attributes and choice alternatives more information can be extracted from the respondents. For instance, in the example in Fig. 2, we would be able to investigate how individuals trade off fatality and injury risk reductions, and not only fatality risk and wealth as in the CVM example. Moreover, CVM seems to be plagued by starting-point and anchoring bias when attempting to ask respondent more than one WTP question. This may be a result of the simplicity of the question format, and usually therefore only one CVM question is asked in surveys (with often one or two follow-up questions to get a more precise range of the respondents' WTP as explained earlier). In DCE the complexity of the choice sets mitigate the risk of anchoring or strategic answering and usually respondents are asked a series of choice sets where attribute levels differ between choice sets. This again means that usually more information about respondents' preferences can be extracted from a DCE compared with a CVM with the same number of respondents.

Assume that the national transport authority will invest in the road network what will reduce the number of fatalities and severe injuries in the road network.		
	Policy A	Policy B
Number of fewer individuals who will die next year when the policy is implemented	2	4
Number of fewer individuals who will be severely injured next year when the policy is implemented	300	400
Your cost	EUR 100	EUR 200
Which option do you prefer ☐ Policy A ☐ Policy B ☐ None of the suggested policies (today's situation remains and no additional cost for you)		

Figure 2 Example DCE question.

One issue to deal with when implementing DCE is the design of the choice sets: the number of attributes, the levels of these attributes, choice alternatives, etc. There exists a rich literature on experimental design and also several software programs to aid with the design of the DCE. Regarding the empirical analysis data from DCE are typically analyzed using a random utility model framework. Based on the example in **Fig. 2**, the utility that respondent n derives from choosing alternative j in choice set t is given by,

$$U_{njt} = \beta_0 \text{sq}_{njt} + \beta_1 \text{die}_{njt} + \beta_2 \text{injury}_{njt} + \beta_3 \text{cost}_{njt} + \varepsilon_{njt}, \quad (16)$$

where die_{njt} , injury_{njt} , and cost_{njt} are the attributes of the choice set; sq_{njt} is a dummy variable for the status quo alternative; the β s are the coefficient to be estimated; and ε_{njt} is a random error term, which is assumed to be IID type I extreme value. As described earlier, the VSL is the MRS between wealth and a reduction in fatality risk and the estimation of VSL based on Eq. (16) is therefore:

$$-\frac{\partial U_{njt} / \partial \text{die}_{njt}}{\partial U_{njt} / \partial \text{cost}_{njt}} = -\frac{\beta_1}{\beta_3}. \quad (17)$$

The standard model to examine the probability that respondent n chooses alternative j in choice set t is the conditional logit (sometimes also referred to as the multinomial logit):

$$P_{njt} = \frac{\exp(\beta_0 \text{sq}_{njt} + \beta_1 \text{die}_{njt} + \beta_2 \text{injury}_{njt} + \beta_3 \text{cost}_{njt})}{\sum_j \exp(\beta_0 \text{sq}_{njt} + \beta_1 \text{die}_{njt} + \beta_2 \text{injury}_{njt} + \beta_3 \text{cost}_{njt})}, \quad (18)$$

which can be estimated using maximum likelihood.

Observed preference heterogeneity can be taken into account in the conditional logit by including, for example, interactions between the attributes of the choice sets and respondents' characteristics. However, typically not all characteristics related to preference heterogeneity are observed and models taking into account such unobserved heterogeneity, like the mixed logit and latent class models, are often used in DCE studies.

Discussion

This chapter has provided a brief introduction to the theory behind and the main empirical approaches of valuing traffic safety. Nonmonetary utility-based health measures, such as health-related quality of life and quality adjusted life years (QALYs) (Hammit, 2002), were not cover, and for brevity the models and methods were described using fatality risk. The methods and model described can also be used, though, to value injury risk.

The WTP approach adopts the concept that individuals are the best judges of their own welfare. That welfare analysis should be based on individuals' own preferences is standard in economics, and the adoption of using the WTP approach to value safety

(risk reductions) is therefore accepted as the appropriate way to monetize safety preferences. Moreover, since both individuals on a daily basis undertake activities where they tradeoff safety for other goods, and that there is a veil of ignorance who will benefit from any safety policies implemented by governments, valuing safety is not considered as unethical among economists. Despite this, many raise concern about the fact that safety (health) is given a monetary value, as it suggests that lives or injuries (illnesses) are valued per se. As this chapter has explained this is a misconception of what is being valued and hopefully the chapter has been able to clarify this.

While there is strong consensus that the WTP (and WTA) approach is the appropriate way to value safety, there is less consensus what the appropriate monetary value should be. Economists do agree that there is not a unique value of the VSL, or any equivalent for specific nonfatal risk reductions, since such values will depend on the population and the context for which the WTP is elicited. However, empirical evidence suggests that values for the same population and context often differ. Such differences may depend on the empirical approach or technique used, or the data on which the analysis were conducted. Therefore much of today's research focus of "The value of life and health" is on how to empirically elicit monetary values that can be considered valid and reliable estimates of individuals' preferences to be used for policy purposes. Thus work is still in progress but a lot of progress has been done in the last decades regarding the empirical methods and data availability.

References

- Andersson, H., Hammitt, J.K., Lindberg, G., Sundström, K., 2013. Willingness to pay and sensitivity to time framing: a theoretical analysis and an application on car safety. *Environ. Res. Econ.* 56, 437–456.
- Andersson, H., Hole, A.R., Svensson, M., 2019. Valuation of health risk. *Oxford Research Encyclopedia of Economics and Finance*. doi: 10.1093/acrefore/9780190625979.013.288.
- Freeman, A.M., Herriges, J.A., Kling, C.L., 2014. *The Measurement of Environmental and Resource Values*. RFF Press, New York, NY.
- Hammitt, J.K., 2002. QALYs versus WTP. *Risk Anal.* 22, 985–1001.
- Johnston, R.J., Boyle, K.J., Adamowicz, W., et al., 2017. Contemporary guidance for stated preference studies. *J. Assoc. Environ. Res. Econ.* 4 (2), 319–405.
- Jones-Lee, M.W., 1974. The value of changes in the probability of death or injury. *J. Polit. Econ.* 82, 835–849.
- Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *J. Polit. Econ.* 82, 34–55.

Further Reading

- Andersson, H., Treich, N., 2011. The value of a statistical life. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *Handbook in Transport Economics*. Edward Elgar, Cheltenham, UK, pp. 396–424.
- Hole, A.R., 2008. Modelling heterogeneity in patients' preferences for the attributes of a general practitioner appointment. *J. Health Econ.* 27, 1078–1094.
- Lindhjelm, H., Navrud, S., Braathen, N.A., Biaisque, V., 2011. Valuing mortality risk reductions from environmental, transport, and health policies: a global meta-analysis of stated preference studies. *Risk Anal.* 31, 1381–1407.
- Schelling, T.C., 1968. The life you save may be your own. In: Chase, S.B. (Ed.), *Problems in Public Expenditure Analysis*. Brookings Institution, Washington, DC, pp. 127–162.
- Train, K., 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, UK.
- Viscusi, W.K., 2014. The value of individual and societal risks to life and health. In: Machina, M., Viscusi, W.K. (Eds.), *Handbook of the Economics of Risk and Uncertainty*, vol. 1. North-Holland, Amsterdam, The Netherlands, pp. 385–452.
- Wijnen, W., Weijermars, W., Schoeters, A., et al., 2019. An analysis of official road crash cost estimates in European countries. *Safety Sci.* 113, 318–327.

The Value of Security, Access Time, Waiting Time, and Transfers in Public Transport

Raquel Espino*, **Juan de Dios Ortúzar†**, **Luis I. Rizzi†**, *Department of Applied Economic Analysis, Instituto Universitario de Desarrollo Económico Sostenible y Turismo, Universidad de Las Palmas de Gran Canaria (ULPGC), Las Palmas, Spain; †Department of Transport Engineering and Logistics, Instituto Sistemas Complejos de Ingeniería (ISCI), Pontificia Universidad Católica de Chile, Santiago, Chile

© 2021 Elsevier Ltd. All rights reserved.

Introduction	122
Value of Walking and Waiting Times	122
Value of Transfer	123
Value of Security	124
References	125
Further Reading	126

Introduction

The time spent in a public transport journey is a key element in user's satisfaction and modal preferences. The perception of travel time varies depending on whether it is in-vehicle time (IVT) or out-of-vehicle time (OVT). In turn, the OVT has at least two components: access to and egress from the bus stop or platform station and waiting for the service to arrive. There may be an additional time when users need to transfer between different transit services (e.g., bus to bus, bus to underground, and so on).

The perception of each travel time component (as well as any opportunities to be productive while travelling) varies as these take place in different environments (on-street, at the stop or platform, inside the vehicle) evoking different feelings (e.g., being anxious while waiting) and requiring different levels of effort (e.g., physical activity when walking). In terms of utility, the various travel time components detract from users' utility. In this regard, OVT is consistently more onerous than IVT.

Another element that influences user's satisfaction and loyalty to public transport is how secure and safe passengers feel when using transit services. By security we refer to the perceived risk of being mugged or subject to crime while waiting at a bus stop or station, or when traveling in the vehicle. Safety, on the other hand, refers to the possibility of suffering an accident while using a public transport service.

The perceptions of safety and security vary wildly in the population. In the case of safety, it is necessary to distinguish too if people are considering a fatal accident, a serious one leading to serious injuries—including losing a limb or becoming paraplegic—or just a minor injury. We will not consider these issues here, but the interested reader is referred to [Rizzi and Ortúzar \(2006\)](#) for a good review of evidence in this sense. In the latter case, there is not much work in the area of public transport related security, but interesting work has been done in terms of feeling of security when walking in potentially dangerous neighbourhoods ([Iglesias et al., 2013](#)) and we will discuss some of it below.

Value of Walking and Waiting Times

The access/egress and walking times depend on the distance between, say, origin and bus stop, and the average speed of walking. The walking distance can be estimated by asking individuals the pathway used or by assuming that transit users walk the shortest path ([Burke and Brown, 2007](#)). Another possibility is to ask for the access or walking times directly, and from this estimate the walking distance using average walking speeds for the gender/age of the individual.

Transport planners' rule of thumb was for many years that the maximum distance for walking to a bus stop was 400 m ([O'Neill et al., 1992](#); [Zhao et al., 2003](#)) and 800 m for accessing rail or underground stations ([Kuby et al., 2004](#); [Schlossberg et al., 2007](#)). However, [El-Geneidy et al. \(2014\)](#) have provided recently evidence that the 85th percentile walking distance to a bus stop is around 524 m from home (and 1259 m for rail) and suggested the importance of the revising the rules of thumb for different modes. Furthermore, [Mulley et al. \(2018\)](#) have shown that these values may also depend on service characteristics, such as frequency or comfort.

The waiting time at stops is the time (min) elapsed from the moment a person reaches the stop or platform to the moment they board the service. When headways (time between two consecutive buses or trains from the same service) are perfectly regular and patrons arrive randomly at the stop or platform, the expected waiting time $E(wt)$ is obviously half the headway (h). If services are not perfectly regular, such that the standard deviation of the headway $[SD(h)]$ is greater than zero, then the expected waiting time increases according to the following formula ([Holroyd and Scraggs, 1966](#)):

$$E(wt) = \frac{1}{2}E(h)\{1 + [SD(h)/E(h)]^2\}, \quad (1)$$

where $E(h)$ is the headway's expected value. Note that if arrivals were exponentially distributed, implying $SD(h) = E(h)$, then $E(wt) = E(h)$, that is, twice the average waiting time of a perfectly regular service. Thus regularizing services definitively contributes to lowering waiting times, and this is why public transport agencies make great efforts to increase regularity.

Another element affecting waiting times is crowding. If a bus or train service arriving at a stop or platform is very crowded, many users will not be able to board the service and will have to wait for the next arrival. Even if a service is not fully crowded, there will be some users, not pressed for time, who will prefer to wait for the next (or the one after the next) service to make a more comfortable trip. If this is the case, average waiting times will increase as effective frequency from the passenger's standpoint becomes higher than scheduled frequency (Batarce et al., 2016).

Real information about headways contributes to reducing anxiety for those waiting at a stop or platform. In addition, real time information can help transit users to better plan in advance their trips. Lu et al. (2018) studied the impact of real-time information about bus arrivals, finding that the expected waiting times were shorter when the bus users checked the bus arrival information before departure.

Regarding the values of access/egress walking time and waiting time, these are usually expressed as multipliers of the value of IVT, as evidence suggest that such multipliers are rather stable when compared across studies in different cities/regions/countries, enhancing their transferability. This gives local transport planners the freedom to estimate monetary savings without getting involved in complex currency transformations.

A large-scale review of British evidence (Wardman, 2014) recommended point estimate multiplier values for walking and waiting time (over IVT) of 1.62 and 1.68; the same study reported a multiplier of 1.93 for both walking and waiting time from nonUK evidence. In Latin America, however, these values tend to be further apart; for example, the seminal study by Gaudry et al. (1989), using carefully measured individual time and cost data for two corridors in Santiago, Chile, found that the walking time and waiting time multipliers were 1.60 and 4.34, respectively, when using their preferred Box-Cox specification. Therefore, the issue is not crystal clear. A recent study by the OECD/ITF (2014) concluded that, notwithstanding the heterogeneity within the evidence, a premium should be attached to walking and waiting time relative to IVT. Historically this premium was normally thought to be a multiplier value of 2.0, but more up-to-date evidence suggests the 2.0 may be an upper bound rather than an average, at least in the case of walking.

However, it has to be borne in mind that these point estimate multipliers may hide a great deal of variation depending on trip purpose, mode, distance, level of crowding, demographic traits, and type of data [revealed versus stated preference (SP)]. Furthermore, as they are, effectively point estimates, their confidence intervals may be (and usually are) quite large (Armstrong et al., 2001; Sillano and Ortúzar, 2005).

Another point to highlight is that frequent users' perception of waiting time is less negative than that of non-frequent users (Dell'Olio et al., 2011). This is unfortunate, because if transport planners want to encourage car users to switch to public transport by, for example, offering higher frequencies and greater spatial coverage, they will need to invest heavily in conveying this information to these nonfrequent public transport users, in imaginative ways.

Notwithstanding, if one considers official practice in social cost-benefit analysis of transport appraisal, the results are more consistent. For example, in the United Kingdom—which is a leading country worldwide because of the many studies carried out on this subject and their excellent official documentation—the Department for Transport establishes multipliers of 2.0 for the valuation of both waiting and walking time relative to IVT (UKDfT, 2017). Across the ocean, the US Department of Transport guidance on the valuation of travel time savings when evaluating competitive funding applications also requires applying a multiplier of 2.0 for both waiting and walking time relative to IVT (USDOT, 2011). These recommendations tend to be followed by most countries worldwide.

Value of Transfer

In the last decade, a large number of studies have focused on the analysis of public transport connectivity. Public transport users try to get to and from a wide variety of destinations. To do so, transferring among different public transporter services is necessary. Therefore transfers are the key elements in any public transport system, as they usually require taking time, effort, and costs; also to many users making a transfer implies subjective cost over and above the latter. Thus the total penalty associated with a transfer may be composed of four components, well differentiated: the transfer walking time, the transfer waiting time, a pure transfer penalty (associated with, for example, the anxiety and risk of missing the connection), and the variation of the latter depending on the transfer environment.

Transfer walking time is defined as the time (min) elapsed from the moment a person alights a given mode to the moment they reach the stop or platform to wait for a subsequent mode. The transfer waiting time, in turn, is defined as the time (min) elapsed from the moment a person reaches the stop or platform of a second (or third) mode, to the moment they are able to board it. The pure transfer penalty is defined as the sheer inconvenience of having to change to another mode or vehicle to reach the final destination. Its valuation is independent of the time spent in making the transfer, but it is strongly dependent on the transfer environment, that is, the quality of the interchange facilities in general (dark and empty, versus lighted and with commerce), the level of crowding experienced, the possibility to get a seat in the next mode, security and safety, if the transfer movement is at the same level or not (i.e., with escalators or elevators), and so on. It has also been found that the pure transfer penalty is influenced by the mental effort required for the "activity disruption" (i.e., having to change vehicle) involved in this situation (Wardman, 2014; Cascajo et al., 2018).

Unfortunately, most studies are not clear in distinguishing among the first three components of the transfer penalty. Some studies consider a pure transfer penalty, whereas some others interact it with the walking and waiting times involved in the transfer itself, but without properly differentiating these from the waiting and walking times in the first leg of the trip. Regarding the first type of studies, [Wardman \(2001\)](#) reports a pure transfer penalty of around 18 min for the United Kingdom. This figure increases if transfer conditions are not at their best; a transfer at its best has to be at the same level, requiring minimum physical effort, safe and secure conditions, and with enough information for users to find their way easily.

[Liu et al. \(1997\)](#) estimated the “transfer penalty” using revealed and SP data, finding that it was higher for changing from car to rail (equivalent to 15 min) than transferring among two rail services (equivalent to only 5 min). In addition, [Currie \(2005\)](#) and [Iseki and Taylor \(2009\)](#) reported a wide range of penalty values by mode, confirming that the intermodal pure transfer penalties tend to be higher than the intramodal penalties. This was also confirmed by [Navarrete and Ortúzar \(2013\)](#), who examined pure transfer penalties in the Santiago integrated public transport system, for different cases: metro–metro, bus–metro, and metro–bus, finding that users had a five times higher preference for transferring between two underground lines, than from bus to metro, and almost 2.5 times more than transferring from metro to bus.

Other studies have explicitly considered the transfer environment, as well as the walking and waiting times, and its effect on the pure transfer penalty. For the London Underground, for example, [Guo and Wilson \(2011\)](#) found that the transfer experience was better if transfers took place at the same horizontal level than if they required a vertical connection. [Douglas and Jones \(2013\)](#) arrived at the same results using a SP survey answered by bus and railways users in Sydney. [Raveau et al. \(2014\)](#) also found that the cost of the pure penalty transfer increased with the level of physical effort, using revealed preference data from both the London Underground and the Santiago Metro. Finally, [Navarrete and Ortúzar \(2013\)](#) reported that transferring between different levels was perceived as less convenient if there were no escalators.

[Raveau et al. \(2014\)](#) also studied the level of comfort (i.e., possibility of getting a seat) and crowding (likelihood of not to being able to board the first train when reaching the stop or platform to connect). With regard to comfort, they found that the greater the likelihood of getting a seat, the less costly the transfer was perceived; with regard to crowding it was the opposite, two sensible results. Finally, [Ceder et al. \(2013\)](#) found that comfort and safety made for a lower pure transfer penalty.

Some studies considered the effect of characteristics such as trip purpose (mandatory or nonmandatory trips), gender, and age. A mandatory trip conveys a higher transfer penalty than a nonmandatory trip, and elder people perceived transfers more negatively than younger people ([Wardman and Hine, 2000](#)). Despite this finding, [García-Martínez et al. \(2018\)](#) found that crowding was more penalized by younger users when transferring. In this sense, [Navarrete and Ortúzar \(2013\)](#) found that young patrons placed a higher value on the availability of information when transferring. Finally, the cost of transferring appears to be greater for females than for males ([Wardman and Hine, 2000](#); [Raveau et al., 2014](#)); females also place a higher premium on safety when deciding which route to take when making a transfer ([Cascajo et al., 2018](#); [Chowdhury, 2019](#)).

In conclusion the transfer penalty is indeed influenced by different factors such as the type of transfer, trip characteristics, the transfer environment, and the demographics of the public transport user. If transport planners want to lure car users to using public transport services more often, they need to devote substantial resources to improve the connections within the public transport system ([Chowdhury and Ceder, 2013](#); [Cascajo et al., 2018](#)). Otherwise, not only car users will not be attracted to using these modes, but it could also happen that current public transport users become less satisfied and eventually prefer to switch to the private car.

Regarding social cost-benefit analysis of transport projects, there are no official guidelines in the United Kingdom not in the United States to consider a pure transfer penalty. Official guidelines just require that walking time and waiting in transfers be valued at the same values as walking and waiting in transit trips of just one leg. However, this practice is not recommended, as it was one of the key reasons behind the monumental failure of the initial implementation of the Transantiago system in Chile ([Muñoz et al., 2009](#)).

Value of Security

Security can be defined as the freedom from being threatened by other people ([Beecroft and Pangbourne, 2015](#)). Travelling by public transport requires interacting with people on the street and inside a public transport unit (i.e., bus or train). In these instances, transport users may be subject to crime. The lack of supervision in vehicles, platforms, stops, and corridors makes public transport users vulnerable to attacks. Poor lighting and lack of visibility, when walking to and from stops or platforms, also create an environment favourable to crime. [Loukaitou-Sideris \(1997\)](#) states that a great deal of criminality occurred at bus stops in the Los Angeles mid and south-central corridors. [Smith and Clarke \(2000\)](#) described a range of crimes that may take place in public transport.

[Masoumi and Fastenmeier \(2016\)](#) cited several studies reporting perception of insecurity as a factor affecting mode choice. [Crime Concern \(2004\)](#) reported the results of a survey where public transport patronage would increase significantly (i.e., 7% in females, 10% in men, and 13 % in young people) if people were happier about their personal security. If, in addition, people who fear travelling by public transport do not have access to a car, they could be severely affected in their social life and even displaced from the job market.

Another reason for self-restrictions on the use of public transport is fear of terrorist attacks: a survey of Australians showed that some respondents avoided public transport because it was considered unsafe in this sense ([Aly, 2012](#)). In particular, [Bennetts and Charles \(2016\)](#) wrote that . . .

"Where users of passenger transport might previously have valued levels of service, cost, on-time performance and convenience as the most important considerations for the infrastructure's operation, it is possible that, with the increased specter of terrorism and asymmetric warfare, users might increasingly see personal safety and security as one of the critical prerequisites in the management of passenger transport."

Despite the relevance of crime and, more recently, terrorism affecting transit users' perception and willingness to use public transport, very little research has been carried out about the willingness to pay for improving security of public transport users. Most valuation studies related with personal security refer to safety as the impact of crashes occurring to a vehicle or an individual. The valuation of safety has a long-standing tradition and constitutes a mature field of analysis both at from a theoretical and an empirical level. This is not the case at all with security in public transport.

A literature review of security valuation studies when walking in urban areas only reveals a handful of studies. For example, Sillano et al. (2006) examined the issue of how to improve urban street design to increase the perception of security while walking in poor neighborhoods in Santiago, Chile. To this end, they designed a SP survey where respondents had to choose between different street environments for walking. Iglesias et al. (2013), improved on this work by adding better imagery and including a payment mechanism that allowed to estimate willingness-to-pay values for urban street traits related with the perception of security, in poor neighborhoods. Their SP experiment was posed for daylight conditions as, otherwise, street lighting would be the major factor hiding any other attributes that could affect the perception of security when walking. They concluded that elements associated with two metavariables, "visual control" and "natural vigilance" (being able to see and to be seen) contributed significantly to a higher perception of security.

Also inspired by Sillano et al. (2006), Borjesson (2012) designed a SP survey to determine how different physical environments influenced the valuation of walking time when accessing public transport. She found that walking in closed environments—not being able to see nor to be seen, and not been able to escape if an unforeseen threat occurred—and in darkness, induced more disutility; also that disutility was higher for women than for men. Finally, Larrañaga et al. (2018) analyzed a set of factors that contribute to facilitate utilitarian walking by means of a "best-worst" survey, responded by a sample of residents of Porto Alegre, Brazil. Related to security, they found that increasing the presence of police officers in the streets contributed to encourage walking.

In conclusion, the valuation of security with regard to public transport is in its infancy and should become a timely topic for attracting research. It should not only focus on the valuation of crime prevention, but also on the valuation of preventing terrorist attacks.

References

- Aly, A., 2012. Terror, fear and individual and community well-being. In: Webb, D., Wills-Herrera, E. (Eds.), *Subjective Well-Being and Security*. Social Indicators Research Series, vol. 46. Springer, Dordrecht.
- Armstrong, P.M., Garrido, R.A., Ortúzar, J. de D., 2001. Confidence intervals to bound the value of time. *Transp. Res. E Log. Transp. Rev.* 37, 143–161.
- Batarce, M., Muñoz, J.C., Ortúzar, J. de D., 2016. Value crowding in public transport: implications for cost-benefit analysis. *Transp. Res. A Policy Pract.* 91, 358–378.
- Beecroft, M., Pangbourne, K., 2015. Future prospects for personal security in travel by public transport. *Transp. Plann. Technol.* 38, 131–148.
- Bennetts, C., Charles, M.B., 2016. Between protection and pragmatism: passenger transport security and public value trade-offs. *Int. J. Pub. Admin.* 39, 26–39.
- Borjesson, M., 2012. Valuing perceived insecurity associated with use of and access to public transport. *Trans. Pol.* 22, 1–10.
- Burke, M., Brown, A.L., 2007. Distances people walk for transport. *Road Trans. Res.* 16, 16–29.
- Cascajo, R., Lopez, E., Herrero, F., Monzón, A., 2018. User perception of transfers in multimodal urban trips: a qualitative study. *Int. J. Sust. Transp.* 13, 393–406.
- Ceder, A., Chowdhury, S., Taghipouran, N., Olsen, J., 2013. Modelling public-transport users' behavior at connection point. *Trans. Pol.* 27, 112–122.
- Chowdhury, S., 2019. Role of gender in the ridership of public transport routes involving transfers. *Transp. Res. Rec.* 2673, 855–863.
- Chowdhury, S., Ceder, A., 2013. A psychological investigation on public-transport users' intention to use routes with transfers. *Int. J. Transp.* 1, 1–20.
- Crime Concern, 2004. People's perceptions of personal security and their concerns about crime on public transport: research findings. Report for the UK Department for Transport, Crime Concern UK, Colchester. Available from: <https://crimeconcernuk.net/>.
- Currie, G., 2005. The demand performance of bus rapid transit. *J. Pub. Transp.* 8, 41–55.
- Dell'Olio, L., Ibeas, A., Cecin, P., 2011. The quality of service desired by public transport users. *Trans. Pol.* 18, 217–227.
- Douglas, N., Jones, M., 2013. Estimating Transfer Penalties and Standardized Income Values of Time by Stated Preference Survey. Australian Transport Research Forum. Available from: http://www.atrf.info/papers/2013/2013_douglas_jones.pdf.
- El-Geneidy, A., Grimsrud, M., Wasfi, R., Tétreault, P., Surprenant-Legault, J., 2014. New evidence on walking distances to transit stops: identifying redundancies and gaps using variable service areas. *Transportation* 41, 193–210.
- García-Martínez, A., Cascajo, R., Jara-Díaz, S.R., Chowdhury, S., Monzon, A., 2018. Transfer penalties in multimodal public transport networks. *Transp. Res. A Policy Pract.* 114, 52–66.
- Gaudry, M.J.I., Jara-Díaz, S.R., Ortúzar, J. de D., 1989. Value of time sensitivity to model specification. *Transp. Res. B Methodol.* 23, 151–158.
- Guo, Z., Wilson, N.H.M., 2011. Assessing the cost of transfer inconvenience in public transport systems: a case study of the London Underground. *Transp. Res. A Policy Pract.* 45, 91–104.
- Holroyd, E.M., Scrags, D.A., 1966. Buses in central London. *Traffic Eng. Control* 8, 158–160.
- Iglesias, P., Greene, M., Ortúzar, J. de D., 2013. On the perception of safety in low income neighborhoods: using digital images in a stated choice experiment. In: Hess, S., Daly, A.J. (Eds.), *Choice Modelling: The State of the Art and the State of Practice*, Edward Elgar Publishing Ltd., Cheltenham.
- Iseki, H., Taylor, B.D., 2009. Not all transfers are created equal: towards a framework relating transfer connectivity to travel behavior. *Trans. Res.* 29, 777–800.
- Kuby, M., Barranda, A., Upchurch, C., 2004. Factors influencing light rail station boarding in the United States. *Transp. Res. A Policy Pract.* 38, 223–247.
- Larrañaga, A.M., Arellana, J., Rizzi, L.I., Strambi, O., Cybis, H.B., 2018. Using best–worst scaling to identify barriers to walkability: a study of Porto Alegre, Brazil. *Transportation* 46, 2347–2749. Available from: <https://doi.org/10.1007/s11116-018-9944-x>.
- Loukaitou-Sideris, A., 1997. Inner-city commercial strips: evolution, decay-retrofit? *Town Plann. Rev.* 68, 1–29.
- Liu, R., Pendyala, R.M., Polzin, S., 1997. Assessment of intermodal transfer penalties using stated preference data. *Transp. Res. Rec.* 1607, 74–80.

- Lu, H., Burge, P., Heywood, C., Sheldon, R., Lee, P., Barber, K., Phillips, A., 2018. The impact of real-time information on passengers' value of bus waiting time. *Trans. Res. Procedia* 31, 18–34.
- Masoumi, H.E., Fastenmeier, W., 2016. Perceptions of security in public transport systems of Germany: prospects for future research. *J. Transp. Sec.* 9, 105–116.
- Mulley, C., Ho, C., Ho, L., Hensher, D.A., Rose, J.D., 2018. Will bus travelers walk further for a more frequent service? An international study using a stated preference approach. *Trans. Pol.* 69, 88–97.
- Muñoz, J.C., Ortúzar, J. de D., Gschwendner, A., 2009. Transantiago: the fall and rise of a radical public transport intervention. In: Saleh, W., Sammer, G. (Eds.), *Travel Demand Management and Road User Pricing: Success, Failure and Feasibility*, Ashgate, Farnham.
- Navarrete, F.J., Ortúzar, J. de D., 2013. Subjective valuation of the transit transfer experience: the case of Santiago de Chile. *Trans. Pol.* 25, 138–147.
- O'Neill, W., Ramsey, D., Chou, J., 1992. Analysis of transit service areas using geographic information systems. *Transp. Res. Rec.* 1364, 131–139.
- OECD/ITF, 2014. *Valuing Convenience in Public Transport*. OECD Publishing, Brussels. Available from: <http://dx.doi.org/10.1787/9789282107683-en>.
- Raveau, S., Guo, Z., Muñoz, J.C., Wilson, N.H.M., 2014. A behavioral comparison of route choice on metro networks: time, transfers, crowding, topology and socio-demographics. *Transp. Res. A Policy Pract.* 66, 185–195.
- Rizzi, L.I., Ortúzar, J. de D., 2006. Estimating the willingness-to-pay for road safety improvements. *Trans. Rev.* 26, 471–485.
- Schlossberg, M., Agrawal, A., Irvin, K., Bekkouche, V., 2007. How far, by which route, and why? A spatial analysis of pedestrian preference. MTI Report 06-06, Mineta Transportation Institute and College of Business, San Jose State University.
- Sillano, M., Greene, M., Ortúzar, J. de D., 2006. Cuantificando la percepción de inseguridad ciudadana en barrios de escasos recursos. *Eure* 32, 17–35 (in Spanish).
- Sillano, M., Ortúzar, J. de D., 2005. Willingness-to-pay estimation with mixed logit models: some new evidence. *Environ. Plann. A Econ. Space* 37, 525–550.
- Smith, M.J., Clarke, R.V., 2000. Crime and public transport. *Crime Just.* 27, 169–234.
- UKDfT, 2017. *Transport Analysis Guidance (TAG) Unit A1.3: User and Provider Impacts*. Transport Appraisal and Strategic Modelling (TASM) Division, UK Department for Transport, London. Available from: <https://www.gov.uk/transport-analysis-guidance-webtag>.
- USDOT, 2011. *The Value of Travel Time Savings: Departmental Guidance for Conducting Economic Evaluations*. U.S. Department of Transportation, Washington, DC. Available from: www.dot.gov/sites/dot.dev/files/docs/vot_guidance_092811c.pdf.
- Wardman, M., 2001. A review of British evidence on time and service quality valuations. *Transp. Res. E Log. Transp. Rev.* 37, 107–128.
- Wardman, M., 2014. Valuing convenience in public transport. Discussion Paper No. 2014-02. OECD, Brussels. Available from: <https://doi.org/10.1787/9789282107683-en>.
- Zhao, F., Chow, L., Li, M., Ubaka, I., Gan, A., 2003. Forecasting transit walk accessibility: regression model alternative to buffer. *Trans. Res. Rec.* 1835, 34–41.

Further Reading

- Kennedy, D.M., 2008. Personal security in public transport travel in New Zealand: problems, issues and solutions. Research Report 344, Land Transport New Zealand, Wellington. Available from: www.nzta.govt.nz/assets/resources/research/reports/344/docs/344.pdf.
- Litman, T.A., Doherty, E., 2009. *Transportation Cost and Benefit Analysis: Techniques*. In: *Estimates and Implications*, second ed. Victoria Transport Policy Institute, Victoria. Available from: <https://www.vtpi.org/tca/>.
- Newton, A.D., 2014. Crime on public transport. In: *Encyclopedia of Criminology and Criminal Justice*, Springer, London. Available from: <http://eprints.hud.ac.uk/19462>.
- Ortúzar, J. de D., Willumsen, L.G., 2011. *Modelling Transport*, fourth ed. John Wiley & Sons, Chichester.
- Wardman, M., Hine, J., 2000. Costs of interchanging: a review of the Literature. ITS-WP-546, Institute for Transport Studies, The University of Leeds.
- Wardman, M., Hine, J., Stradling, S., 2001. *Interchange and travel choice*, vol 1. Scottish Executive Central Research Unit, Edinburgh.

Demand for Passenger Transportation

Kenneth A Small*, Robin Lindsey†, *1721 W. 104th Place, Chicago, IL, United States; †Sauder School of Business, University of British Columbia, Vancouver, BC, Canada

© 2021 Elsevier Ltd. All rights reserved.

Demand Analysis in Economics	127
Discrete Choices	128
Examples	130
Data Sources	131
Activity Patterns	131
Nonoptimizing Behavior	132
References	132
Further Reading	133

Demand Analysis in Economics

Classical economic analysis separates demand from supply, each a conceptually distinct relationship between quantity and price; it then looks for equilibria that are consistent with both. In that simplified world, the “demand side” expresses the preferences of consumers or institutions that might purchase a good, usually in terms of the quantity purchased at any given price. This idea naturally generalizes to cases where the quantity demanded depends not only on price but on quality attributes, which may also be endogenously determined in the market. The “supply side” of the analysis must describe what suppliers (i.e., firms, government, perhaps individuals) are willing to provide, in terms of quantities and quality levels, at a given price. For example, firms might be posited to maximize profits by choosing a quantity and various quality attributes, given a price of the good, and their knowledge of consumers’ quality valuations.

A common example in transportation is a congested highway, on which travelers care about price, average travel time, and variability of travel time—a measure of “reliability” (Brownstone and Small, 2005). Another example is use of public transit. Travelers care about price, access time, expected waiting time for a vehicle to arrive, and in-vehicle travel time. A supplier, whether a government or private firm, might choose price and those aspects of quality under its control, letting ridership be determined in equilibrium. Or the supplier might target ridership, based perhaps on the capacity of a subway tunnel, and use its knowledge of the demand side to optimize price and service frequency. Its objective might be profits, social welfare, or something else.

In both examples, the demand system reveals an implicit valuation of in-vehicle travel time, as well as valuations of waiting time or reliability. In practice, demand models reveal many valuations. For example, a model of choice of automobile might reveal valuations of size, power, and fuel efficiency. The fact that suppliers must know these valuations, perhaps in great detail, in order to design their services and products shows that the demand and supply sides cannot be as neatly separated as the armchair analyst might wish. Indeed, one of the major advances in analyzing public transit provision was to incorporate consumers’ values of waiting time into a cost function that a provider could then try to minimize (Mohring, 1972).

A classic “aggregate demand” model explains total use of a service, or total market for a product, for particular groups of consumers (e.g., those living in an urban neighborhood or those in a given demographic category). The model can be written as an equation expressing quantity x (e.g., daily passenger ridership or flow rate of automobile traffic) as some function of price and other quality attributes. If the functional form is linear in unknown parameters, and the price and quality attributes are summarized in a (column) vector z of “explanatory variables,” then we have the standard regression framework:

$$x = \alpha + \beta z + \varepsilon, \quad (1)$$

where α is a constant, β is a (row) vector of parameters, and ε is an “error term” denoting the inevitable imprecision in the model’s explanatory ability. The variables z might include price, in-vehicle time, and waiting time in the case of demand for public transit; or cost, average travel time, and variability of travel time in the case of a congested highway. They can also include nonlinear transformations or combinations of such quantities, as well as interactions between (e.g., products of) those variables and demographic characteristics; so the linearity of (1) is not very limiting. The unknown parameters α and β (the latter consisting of “coefficients” of each of the variables included in vector z) are typically determined through a statistical procedure to obtain the best fit to some data set on which the model is “estimated.” A common such procedure is “least squares,” which minimizes the sum of the squared deviations between actual and predicted values of x .

An example illustrates the relationship between a demand model and values of quality attributes. Suppose we estimate a linear regression model with three variables: price p , travel time T , and travel time multiplied by the individual’s wage rate w :

$$x = \alpha + \beta_p p + \beta_T T + \beta_{wT}(wT) + \varepsilon,$$

where β_p and β_T are expected to be negative since price and travel time are deterrents to use. Variable wT is included because in the most basic theories of travel-time valuation, time spent traveling reduces the time available for work. In this specification, travel time is implicitly valued at a rate:

$$VOT = [\beta_T + (\beta_{wT}w)]/\beta_p.$$

Formally, this “value of time” VOT is the marginal rate of substitution (within consumer demand) between travel time and price: you would give up VOT dollars in price in order to save 1 h in travel time, if the units of p and T are dollars and hours, respectively. Mathematically, it can be written as $(\partial x/\partial T)/(\partial x/\partial p)$. Empirically, it is often found to be approximately half the wage rate. In practice, studies of VOT use more elaborate specifications to account for other factors, such as work-hour constraints.

The variable “price” might include both the actual price and other quantifiable cost components, such as user-supplied fuel and operating costs of an automobile; the variable is then usually called “cost.” Furthermore, it may be convenient to incorporate some quality elements, appropriately valued, along with cost into a “generalized cost.” This works best when everyone has the same valuation. For example, if $\beta_{wT} = 0$ in the specification above, the generalized cost is

$$p_G \equiv p + (\beta_T/\beta_p)T,$$

which would replace the two variables p and T in the demand system. This simplification is often convenient when undertaking equilibrium calculations after estimating the demand model.

Naturally, most demand studies are more complex than this. An example is the demand for motorization, defined as per-capita motor vehicle ownership, across countries. Some researchers have posited a “partial adjustment” mechanism, in which actual motorization M_{it} in country i and year t adjusts from its previous year’s value toward a target value M_{it}^* :

$$M_{it} = M_{i,t-1} + (1 - a)(M_{it}^* - M_{i,t-1}) + \varepsilon_{it},$$

with the target value determined as a linear function of relevant variables such as per-capita income, population density, and urbanization. This specification implies a simple regression formula for M_{it} in which $M_{i,t-1}$ appears as an additional explanatory variable, with coefficient a . The estimated parameter a is a measure of inertia, and is positively related to the ratio of long-run to short-run responses to changes in conditions.

Discrete Choices

Many transportation decisions are made by choosing one of a set of discrete options. These choices are made by individuals (or other decision-making entities) who differ in observable characteristics of themselves or their environment, and also who differ in unobservable ways. A major change in demand analysis for transportation, beginning in the 1960s, was to develop rigorous models and statistical techniques to describe such behavior by individuals. These methods are called “discrete choice” or “disaggregate demand analysis.” A key innovation is to describe such choices probabilistically, thereby formally incorporating the unobservable determinants (McFadden, 1974, 1978). Aggregate forecasts can be made using a “forecasting sample,” of which each member can be assumed to represent a large population of decision-makers who are identical in their observable characteristics. A choice probability is calculated for each member of the forecasting sample, so aggregation amounts to assuming that this probability tells us the fraction of the subpopulation represented by this sample member who makes that choice.

The workhorse of discrete-choice analysis is the “additive random utility model” (RUM), in which a decision maker n chooses among alternatives j contained in a “choice set” $\{j = 1, \dots, J\}$, by selecting the alternative with the highest “conditional indirect utility” U_{jn} . (Note that one alternative may be not traveling.) This is one of many examples of deriving a demand function like Eq. (1) from a formal theory of utility maximization. In this case, utility is posited to have two additive components: a “systematic utility” V_{jn} that is a function of observable characteristics z , and a “random utility” ε_{jn} (also called an “error term”) expressing influences on the decisions that are unobservable to the analyst. In a linear-in-parameters specification, V_{jn} is a regression function linear in a coefficient vector, unlike (1) where it is quantity demanded that is a linear regression function:

$$U_{jn} \equiv V_{jn} + \varepsilon_{jn} = \alpha_j + \beta z_{jn} + \varepsilon_{jn}. \quad (2)$$

If K variables are included in vector z , there are $K + J - 1$ unknown coefficients: the K components of vector β , and all but one of the “alternative-specific constants” α_j . One of these constants (or any one combination of them) can be chosen arbitrarily in a “normalization” step because it is only utility differences across alternatives, not the absolute utility, that affect choice. (For the same reason, any portion of utility that is invariant across alternatives is excluded from V .)

The probabilistic aspect of choice arises because of the J random utilities $\{\varepsilon_{jn}\}$. By specifying a joint probability distribution for them, the analyst can calculate the probability of each choice once the coefficients are estimated and the observable characteristics measured. Often this distribution is assumed for simplicity to be identical for all decision makers n . Two probability distributions have been used extensively: multivariate normal and “generalized extreme value” (GEV). The former defines the “probit” model,

which has strong theoretical foundations, but the latter is more computationally tractable and thus far more widely used. The GEV cumulative distribution function is

$$F(\varepsilon_1, \dots, \varepsilon_J) = \exp[-G(e^{-\varepsilon_1}, \dots, e^{-\varepsilon_J})], \quad (3)$$

where G is a function increasing in all its arguments and satisfying certain other technical conditions, and $e \approx 2.718$ is the base of natural logarithms.

The most widely used GEV model is “multinomial logit,” sometimes called “conditional logit” or simply “logit,” in which function G is a simple sum. In that case, the random terms are independently and identically distributed (iid) with cumulative distribution function F equal to the product of J distribution functions, one for each alternative, each of the form:

$$\Pr[\varepsilon_{jn} < x] = \exp(-e^{-x}). \quad (4)$$

This univariate distribution is variously called “extreme value,” “double-exponential,” or “Gumbel.” Another common GEV model groups alternatives into “nests,” within which random terms are correlated but across which they are independent; that model is called “nested logit.”

The choice probabilities for the GEV model turn out to be rather simple in form. In what follows, we omit subscript n for simplicity. The probability of choosing alternative i is

$$P_i = \frac{e^{V_i} \cdot G_i(e^{V_1}, \dots, e^{V_J})}{G(e^{V_1}, \dots, e^{V_J})}, \quad (5)$$

where G_i denotes the i th partial derivative of function G . In the case of logit, this simplifies to

$$P_i = \frac{\exp(V_i)}{\sum_{j=1}^J \exp(V_j)}. \quad (6)$$

This probability formula has proven to be extraordinarily convenient both computationally and conceptually.

The conceptual convenience of (6) is that its denominator provides a summary of the value of the choice set to the decision maker, sometimes known as “inclusive value.” More formally, the log of the denominator is the maximum expected utility, that is, the expected value (in a probabilistic sense) of the utility achievable by a decision maker facing this particular choice set. This property of logit models permits them to be used as components of much more elaborate modeling systems. It also facilitates analysis of social welfare by, for example, calculating the benefits to any consumer segment of a specified change in one or more of the variables z . This same property extends to the GEV model, where the inclusive value is the denominator of Eq. (5).

Thanks to this property, it has become common practice in both academic and applied work to develop elaborate suites of models explaining many inter-connected decisions, linked together by these “inclusive values” within some specified nesting pattern. For example, the California High-Speed Rail Authority has forecast ridership of its planned new rail lines using models including trip generation by origin zone, trip attraction by destination zone, main mode (car, air, conventional train, or high-speed train), access modes at both origin and destination (for all the main modes except car), and route (including access station). The first two decisions are continuous choices, the others discrete. Typically, each choice occurring earlier in the above list depends on an inclusive value estimated for one or more of the later choices. However, all choices are viewed as simultaneous, not sequential, although they may incorporate time lags as in the partial-adjustment model described earlier.

The logit model has one serious limitation: the probabilities of any two alternatives have a ratio that is independent of characteristics of any other alternatives:

$$\frac{P_j}{P_k} = \exp(V_j - V_k). \quad (7)$$

This property is known as “independence from irrelevant alternatives” (IIA). It is unrealistic in many situations, which is why nested logit was first developed. The classic example is the “red bus—blue bus” problem, in which a transit agency operating a subway and a fleet of red buses adds a new choice, namely, a blue bus, that is identical to the red bus except for its color. Presumably, this addition would have little or no effect on subway ridership; yet Eq. (6) predicts that subway ridership would fall due to the expanded choice set (making the denominator larger). Nested logit alters this property by putting red and blue buses within a single nest, within which the error terms can be correlated. This allows blue buses to divert ridership preferentially from red buses rather than from subway.

Many other GEV models have been developed. We provide two examples in the next section. Many non-GEV models, still based on the random utility framework, have also been developed: for example, models using multiplicative errors, and models permitting alternatives to be complements as well as substitutes. In addition, researchers have specified models that jointly determine a discrete choice, such as what car to purchase, and an associated continuous choice, such as how much to drive that car.

Two breakthrough developments have greatly added to the usefulness of logit and other GEV models. First is “mixture models,” in which the coefficients themselves can be random (Walker and Ben-Akiva, 2011). The resulting probabilities follow, statistically, a mixture distribution that combines the assumed distribution of random coefficients (e.g., normal, log-normal, or triangular) with

the extreme-value distribution of the additive error term. The most common mixture model is “mixed logit,” in which probabilities are logit conditional on the values of the random parameters (McFadden and Train, 2000). This turns out to be very flexible, and frequently renders unnecessary the more complex mathematics of GEV or other more general models. Estimation of mixture models has become practical due to better computing power and the “method of simulated moments,” which approximates probabilities by simulating them numerically. Values for random coefficients are generated repeatedly from a process that follows their postulated joint probability distribution, choice probabilities are computed conditional on those coefficient values, and the results are averaged to approximate the true choice probabilities.

The second breakthrough is the use of “instrumental variables,” a well-known econometric technique long used to eliminate biases otherwise produced by endogeneity in explanatory variables. The prototype example endogeneity in discrete-choice analysis is the variable “price” for the study of automobile purchase. Price is often determined simultaneously with purchase choice because consumers care about quality attributes that are not measured by the analyst, but that affect the prices of vehicles being offered on the market. The innovation is to find one or more variables (instruments) that help explain the observed price but should have no direct effect on a consumer’s utility for a certain alternative—for example, a variable that summarizes the prices of other model cars offered by the same manufacturer. The instrumental variable can then be used to “purge” the variable in question (vehicle price) of its endogenous component.

Instrumental variables became popular for discrete-choice analysis when Berry et al. (1995) (often referred to as “BLP”) introduced them simultaneously with another important development: using a combination of disaggregate survey data and aggregate market-share information to estimate more efficiently a disaggregate choice model. This technique works because knowing the aggregate market shares permits an estimation algorithm to adjust for any tendency of its estimated coefficients to make forecast errors in those shares. Models that incorporate both instrumental variables and aggregate market-share data are now widely used in many branches of economics. At their core is the extreme-value distribution of a random utility term, as in Eq. (4).

Examples

In this section, we briefly review two empirical studies that used GEV models. The first is a study of potential demand for high-speed rail service in the Toronto-Montreal corridor (Koppelman and Wen, 2000). It was estimated on a sample of 2769 travelers who could choose among three modes: air, train, and car, denoted by $j = 1, 2$, and 3. (Bus was excluded because it had a very small market share, and bus-specific parameters could not be estimated reliably.) The results illustrate the flexibility of an intuitively specified GEV model to capture correlations among alternatives.

The model in question is the paired combinatorial logit model. It is based on the assumption that any pair of alternatives may show some degree of similarity in unobserved preferences. Similarity can be captured by allowing the random utility components of the alternatives to be correlated, and unlike with the nested logit model there is no need for pairs to be mutually exclusive. The generating function is a particular example of $G(\{y_j\})$ in Eq. (3):

$$G(y_1, y_2, y_3) = \sum_{j=1}^2 \sum_{k=j+1}^3 \left(y_j^{1/\rho_{jk}} + y_k^{1/\rho_{jk}} \right)^{\rho_{jk}}, \quad (8)$$

where it is required that $0 < \rho_{jk} \leq 1$ for each j and k . This model has three nests, consisting of all possible pairs of alternatives, each with its own “dissimilarity parameter” ρ_{ij} . From Eq. (5), the choice probabilities are:

$$P_i = \sum_{j \neq i} e^{V_i/\rho_{ij}} \frac{\left(e^{V_i/\rho_{ij}} + e^{V_j/\rho_{ij}} \right)^{\rho_{ij}-1}}{\sum_{j=1}^2 \sum_{k=j+1}^3 \left(e^{V_j/\rho_{jk}} + e^{V_k/\rho_{jk}} \right)^{\rho_{jk}}}.$$

Multinomial logit is the special case where $\rho_{ij} = 1$ for all three pairs i, j .

On the Toronto-Montreal data, the best-fitting model with dissimilarity parameters in the required range constrains ρ_{12} to equal one. Thus, only two dissimilarity parameters are estimated: one for the nest including air and car (ρ_{13} , estimated at 0.73), and one including train and car (ρ_{23} , estimated at 0.58). These results suggest that the travelers considered train and car to be the most similar modes, and air and train the least similar despite them being public modes that differ in access, privacy, and other characteristics from the car. One can only speculate as to why: perhaps access time at an airport is perceived differently or measured incorrectly, perhaps people especially strongly like or dislike flying, perhaps some people care especially about luggage capacity. (When the study was conducted, airport security procedures were not as stringent as they are now and are unlikely to have had a major influence on mode preferences.) The model statistically outperformed all other models tested, including multinomial logit and nested logit models. It also yielded significantly different own- and cross-price elasticities for the three modes. This illustrates that different models can produce substantially different demand forecasts for new travel modes.

The model estimation yields coefficients β_p , β_{IVT} , and β_{OVT} on variables indicating travel cost, in-vehicle time, and out-of-vehicle time, respectively. Their estimated ratios are measures of values of in- and out-of-vehicle times: $\beta_{IVT}/\beta_p = \text{C\$}19/\text{h}$ and $\beta_{OVT}/\beta_p = \text{C\$}80/\text{h}$, respectively (in 1989 Canadian dollars). The much higher estimate for out-of-vehicle time highlights the potential danger of considering only total trip time rather than accounting separately for the different stages of a trip.

The second example features the “ordered GEV” model. It is designed to handle situations where the alternatives are defined on an ordering along which one expects nearby alternatives to be most closely correlated. It was first used by Small (1987) to describe commuters’ choice of arrival time at work, taking into account that travel at the most convenient times might be slower due to congestion. The model was estimated using discrete (5-min) time intervals and a generating function with nests similar to those in Eq. (8), each containing two or more alternatives that form a sequence within the temporal ordering. A priori, arrival times that are close together (e.g., 8:00 and 8:05) are expected to be better substitutes than arrival times further apart (e.g., 8:00 and 8:30). Using the results, Small (1987) was able to estimate the disutility of arriving early or late for work, relative to the disutility of travel time. The trade-off between scheduling inconvenience and trip duration is central to the “bottleneck model” of trip-timing decisions and traffic congestion due to Vickrey (1969).

Data Sources

Discrete-choice models require extensive data to estimate. Values are needed for every characteristic of every alternative available to each individual in the sample. In particular, information is required on the characteristics of the transportation system including the costs, travel times, and reliability for every possible mode or route. Various shortcuts are possible. Nevertheless, for many practical investigations, these requirements are daunting. Consequently, models are sometimes broken up into parts, with entire teams of analysts and data-collection strategies developed for each part.

Values for the characteristics of the travel environment can be derived in several ways. Most often they are generated by some combination of observation and engineering calculation—for example as applied to a coded network describing all the roads and public-transport links in the study area. An advantage of such “objective” data is that they are not influenced by the limited knowledge, and perhaps biased perceptions, of travelers. (Although travelers’ perceptions may indeed affect their decisions, if perceptions are biased by the choices themselves, e.g., as a form of self-justification, they are unsuitable for predicting those choices.) A disadvantage of objective data is that they may not correspond to the quantities that actually influence travel decisions. For example, the point-to-point travel times on a coded network representing a public transit system may not accurately reflect the heuristics travelers use to choose a route, such as the distorted distances shown on a map of the transit system (Larcom et al., 2017), or the environment travelers may encounter at transfer points due to weather, noise, physical barriers, or lack of safety. Many models attempt to ameliorate such inaccuracies by including other variables, such as the number of transfers required for a journey and perhaps some observable characteristics of those transfers.

Another disadvantage of objective data on travel characteristics is that they may be highly correlated across the choice alternatives people face in real life, making it hard to separate the causal factors with satisfactory statistical precision.

Now let us consider the source of data on choices made. When data reflect actual choices in real-life situations, the analysis is termed one of “revealed preference” because individuals supposedly reveal the preferences that determine their choices. An alternative is for the analyst to design surveys in which respondents are presented with hypothetical choices. This is often called “stated preference” or “stated choice” analysis. It has the advantage that alternatives can be designed to minimize correlation and thus improve statistical precision. Stated preference analysis is also useful for forecasting potential demand for a new alternative that is unlike any that survey respondents have experienced. It has some disadvantages, however. First, people may not act in real life the way they say they would in a survey. A large research literature in survey design is aimed at finding survey techniques that minimize this problem. Second, if people misperceive a travel characteristic such as travel time, their response to actual changes may differ from what they think when the same changes are described hypothetically. This misperception can lead to errors in forecasting the effects of policies such as major transit investments or congestion tolls. Third, research has shown that people exhibit strong “loss aversion” to an unfavorable change in any particular characteristic relative to some subjective reference situation; this can affect stated-preference responses even though in real life individuals rarely are in a stable reference situation due to constant changes in their situations or in the travel environment.

Thus, both revealed preference and stated preference data have disadvantages. This difficulty can be partially overcome by using the two types of data together (Hensher et al., 1999).

Some analysts believe that only perceived values should be used to forecast behavior. But to make policy-relevant predictions, that approach must be supplemented with a model of perception formation so the analyst can predict how a given policy will affect perceptions, and in turn how it will affect actual behavior. Some progress has been made on this agenda, resulting in “attitudinal models” that incorporate both steps—often within a nested framework much like that described earlier.

One particular question of recent interest is whether cultural attitudes of newer generations toward driving may be causing a worldwide shift away from car ownership. There is some evidence of such shifts, but an unresolved debate about whether or not they can be explained by conventional explanatory variables such as home ownership, urbanization, income, fuel prices, and the changing characteristics of public transportation alternatives (Goodwin and Van Dender, 2013).

Activity Patterns

Travel is widely believed to be mostly a derived demand (i.e., not wanted for its own sake, but as a means to engage in desired activities distributed unevenly over time and space. A natural approach is to try to explain the underlying factors that motivate trips:

an idea that has led to “activity analysis” (Pinjari and Bhat, 2011). For example, one might consider work, shopping, errands, and entertainment as the desired activities, each with its own preferred frequency, time schedule, duration, and location. This type of analysis quickly becomes very complicated, as the number of potential combinations is enormous. Nevertheless, some practical models of activity choices have been developed and are now used by several urban planning agencies as the basis for forecasting travel behavior.

The advent of autonomous vehicles (AVs) is stimulating further research in activity analysis. Early evidence suggests that the availability of AVs will generate new trips for activities that were not previously considered practical, such as night travel by seniors who do not like to drive in the dark. AVs may also enable entirely new routing patterns since they can deliver people directly to destinations, and then park far away (Zakharenko, 2016). Preliminary analyses also suggest that AVs are likely to divert ridership from public transit, and so may exacerbate existing problems with traffic congestion and public-transit finance (Bahamonde-Birke et al., 2018).

Nonoptimizing Behavior

Most analysis of transportation demand, like most economic analyses of any kind, starts from an assumption that consumers are “rational” (i.e., that they choose optimally, in some definable sense, using a well-defined information set). This is a useful starting point for understanding human behavior, but few would accept it as a literal description. Indeed, psychologists and economists have observed numerous behavioral anomalies that suggest otherwise. The field of “behavioral economics” has begun to develop rigorous methods of accounting for many of them.

A few examples illustrate the possibilities for travel demand. First: studies tend to obtain larger estimates of the value of travel time when using revealed-preference rather than stated-preference data. One explanation is that people are prone to overestimating the time they lose due to congestion in practice.

A second example is the finding that survey respondents exhibit asymmetry to gains and losses when presented with alternatives differing in cost and travel time (De Borger and Fosgerau, 2008). In particular, they are more sensitive to losing cost or time (relative to some baseline) than to gaining the same amount of cost or time. Naturally, such so-called “loss aversion” could affect short-run responses to proposed or actual changes in the travel environment, and may be a factor in the tendency of people to oppose road pricing schemes before they are implemented but support them afterwards (Börjesson et al., 2012). For predicting long-run behavior, however, there is no obvious stable baseline for a population subject to continuous turnover in a continually evolving environment. Moreover, studies often find that loss aversion decays as people become familiar with gains and losses. Hence, asymmetry may be an important consideration for interpreting the results of stated-preference surveys, but not significant as far as affecting actual long-run behavior.

A final example is the alleged undervaluation of future fuel costs when choosing among automobiles (or other consumer durables) with different levels of energy efficiency. Much evidence, though not uncontested, suggests that such undervaluation is the norm (Gerarden et al., 2017). Several behavioral traits could account for it: lack of attention to fuel efficiency during purchase; inability to perform required optimization calculations; poor understanding of uncertainty in future fuel prices; purchasing a car for status rather than for its useful services; mistaken information about how fuel efficiency is correlated with other vehicle characteristics; lack of access to credit at competitive interest rates; and impulsiveness leading to decisions that are later regretted. There are some ways to empirically disentangle these possibilities, but such research is in its infancy.

More broadly, researchers have suggested that travel demand analysis needs to better formalize the decision-making process, with all its potentially “irrational” elements, and design data-collection and data-analysis strategies to measure all the aspects of decision-making. This is a big challenge, especially for a discipline that already has a reasonably good track record using “rational actor” models. Nevertheless, our understanding of travel behavior will be the better for measuring all the factors that affect it, whether “rational” or not.

References

- Bahamonde-Birke, F.J., Kickhöfer, B., Heinrichs, D., Kuhnimhof, T., 2018. A systemic view on autonomous vehicles: policy aspects for a sustainable transportation planning. *DisP – Plan. Rev.* 54 (3), 12–25.
- Berry, S.T., Levinsohn, J., Pakes, A., 1995. Automobile prices in market equilibrium. *Econometrica* 63, 841–890.
- Börjesson, M., Eliasson, J., Hugosson, M., Brundell-Freij, K., 2012. The Stockholm congestion charges—5 years on. Effects, acceptability and lessons learnt. *Transport Policy* 20, 1–12.
- Brownstone, D., Small, K.A., 2005. Valuing time and reliability: assessing the evidence from road pricing demonstrations. *Transport. Res. Part A: Policy Practice* 39 (4), 279–293.
- De Borger, B., Fosgerau, M., 2008. The trade-off between money and travel time: a test of the theory of reference-dependent preferences. *J. Urban Econ.* 64 (1), 101–115.
- Gerarden, T.D., Newell, R.G., Stavins, R.N., 2017. Assessing the energy-efficiency gap. *J. Econ. Lit.* 55 (4), 1486–1525.
- Goodwin, P., Van Dender, K., 2013. ‘Peak car’ — Themes and issues. *Transport Rev.* 33 (3), 243–254.
- Hensher, D.A., Louvière, J., Swait, J., 1999. Combining sources of preference data. *J. Econometr.* 89, 197–221.
- Koppelman, F.S., Wen, C.-H., 2000. The paired combinatorial logit model: properties, estimation and application. *Transport. Res. Part B: Methodol.* 34, 75–89.
- Larcom, S., Rauch, F., Willems, T., 2017. The benefits of forced experimentation: striking evidence from the London Underground network. *Quarterly J. Econ.* 132, 2019–2055.
- McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), *Frontiers in Econometrics*. Academic Press, New York, pp. 105–142.
- McFadden, D., 1978. Modelling the choice of residential location. In: Karlqvist, A., et al. (Eds.), *Spatial Interaction Theory and Planning Models*. North-Holland, Amsterdam and New York, pp. 75–96.

- McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. *J. Appl. Econometr.* 15, 447–470.
- Mohring, H., 1972. Optimization and scale economies in urban bus transportation. *Am. Econ. Rev.* 62, 591–604.
- Pinjari, A.R., Bhat, C.R., 2011. Activity-based travel demand analysis. In: de Palma, A., et al. (Eds.), *A Handbook of Transport Economics*. Edward Elgar, Cheltenham, UK, pp. 213–248.
- Small, K.A., 1987. A discrete choice model for ordered alternatives. *Econometrica* 55 (2), 409–424.
- Vickrey, W.S., 1969. Congestion theory and transport investment. *Am. Econ. Rev. (Papers and Proceedings)* 59 (2), 251–260.
- Walker, J., Ben-Akiva, M., 2011. Advances in discrete choice models: mixture models. In: de Palma, A., et al. (Eds.), *A Handbook of Transport Economics*. Edward Elgar, Cheltenham, UK, pp. 160–187.
- Zakharenko, R., 2016. Self-driving cars will change cities. *Reg. Sci. Urban Econ.* 61, 26–37.

Further Reading

- Ben-Akiva, M., McFadden, D., Train, K., 2019. Foundations of stated preference elicitation: consumer behavior and choice-based conjoint analysis. *Foundations Trends Econometr.* 10 (1-2), 1–144, doi:10.1561/08000000036 (<https://eml.berkeley.edu/~train/foundations.pdf>).
- Dargay, J., Gately, D., Sommer, M., 2007. Vehicle ownership and income growth, worldwide: 1960–2030. *Energy J.* 28 (4), 143–170.
- Delbosc, A., Currie, G., 2013. Causes of youth licensing decline: a synthesis of evidence. *Transport Rev.* 33 (3), 271–290.
- Fifer, S., Rose, J., Greaves, S., 2014. Hypothetical bias in stated choice experiments: Is it a problem? And if so, how do we deal with it?. *Transport. Res. Part A: Policy Practice* 61, 164–177.
- Haboucha, C.J., Ishaq, R., Shiftan, Y., 2017. User preferences regarding autonomous vehicles. *Transport. Res. Part C: Emerging Technol.* 78, 37–49.
- Hensher, D.A., Rose, J.M., Greene, W.H., 2015. *Applied Choice Analysis*. Cambridge University Press, Cambridge, UK.
- Johansson, M.V., Heldt, T., Johansson, P., 2006. The effects of attitudes and personality traits on mode choice. *Transport. Res. Part A: Policy Practice* 40, 507–525.
- McFadden, D., 2001. Economic choices. *Am. Econ. Rev.* 91, 351–378.
- Munger, D., L'Ecuyer, P., Bastin, F., Cirillo, C., Tuffin, B., 2011. Estimation of the mixed logit likelihood function by randomized quasi-Monte Carlo. *Transport. Res. Part B: Methodol.* 46, 305–320.
- Newman, J.P., Chorus, C., 2015. Attitudes and habits in highly effective travel models. *Transportation* 42, 3–5.
- Roorda, M.J., Carrasco, J.A., Miller, E.J., 2009. An integrated model of vehicle transactions, activity scheduling and mode choice. *Transport. Res. Part B: Methodol.* 43, 217–229.
- Small, K.A., Rosen, H.S., 1981. *Applied welfare economics with discrete choice models*. *Econometrica* 49 (1), 105–130.
- Small, K.A., Verhoef, E.T., 2007. *The Economics of Urban Transportation*. Routledge, London.

Real-World Experiences of Congestion Pricing

Charles Raux, University of Lyon, CNRS, LAET, Lyon, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	134
From Theory to Practical Guidelines	134
Case Studies of Ongoing Congestion Pricing Schemes	135
Singapore Electronic Road Pricing Scheme (ERP)	135
London Congestion Charging Scheme	136
Stockholm Congestion Tax Scheme	136
Value-Pricing in the United States	137
Other Ongoing Schemes	137
Overview	138
References	138
Further Reading	138

Introduction

Congestion on transport infrastructures, whether road, rail, or other facilities, occurs when too many vehicles are trying to use an infrastructure, which has a fixed capacity in the short term. For instance, road users driving toward employment centers at the peak morning period, trucks or buses suffer from time losses they entail to each other. This result in delays when arriving at destination, losses of productivity for freight transport, and public transport buses locked in jams.

For a long time, economists have advocated the principle of road user charging whether for financing investments in capacity (Jules Dupuit in 1849) or taxing congestion to discourage excess travel when compared to fixed capacity in the short term (Arthur Pigou in 1920). Numerous research works have since developed these ideas and most transport economists today agree on a basic prescription: road user charging should be implemented, where and when congestion is critical, in order to ensure that potential road users consider the delay they impose on other road users when entering the road. By reducing traffic, this should entail gains in speed and in travel time reliability, a more efficient use of road capacity, and hence reduce the pressure to costly increases of this capacity. Speaking of road does not mean that congestion charging is restricted to this kind of infrastructure. Obviously this applies also to public transport users (rush hours fares), rail (path reservation), airports (take-off slot reservation), and so on. However, in the following, we concentrate on road congestion pricing schemes as these are the most difficult to implement.

The basic prescription is based on obvious simplifying assumptions and models. These are due to limits in analytical tools, in knowledge and ability to take account of the interactions of transport with the rest of the economy. This renders the design of optimal congestion pricing schemes particularly difficult. Nevertheless several “congestion pricing” schemes have been implemented around the world after decades of research and debates. Their current operative performance and some time elapsed allow us to assess their relevance through two basic questions: Are they effective in improving traffic conditions or Are they efficient, that is, improving the community welfare?

First practical guidelines and basic scheme parameters derived from theory of congestion pricing are set out. Then case studies of iconic successful experiences are described. Finally, an overview of successes and failures allow drawing some general lessons for implementation.

From Theory to Practical Guidelines

From theory of congestion pricing comes the prescription that where congestion is critical road tolls should be implemented and differentiated according to the variations of the marginal external cost of congestion that each road user imposes to each other. Given the road transport technology this marginal cost varies according to the heterogeneity of users (e.g., trucks vs. light vehicles) and road links (various peak-load technologies).

This implies that tolls should vary according to travel times (e.g., days in the week or time of day), travel place (e.g., link or area), vehicle type (e.g., car or truck) and trip purpose (e.g., commuting). This feature of pricing variation specifically distinguishes congestion pricing from flat road pricing whose main purpose is to finance infrastructure.

The basic expected benefits are improvement in traffic speed and hence time savings, which make the overall improvement in community welfare. Revenues raised from pricing are a transfer from road users to the community and do not count in the basic welfare evaluation. However, they may be earmarked for transport improvements or even redistributed to the concerned population and hence contribute to acceptability of congestion pricing.

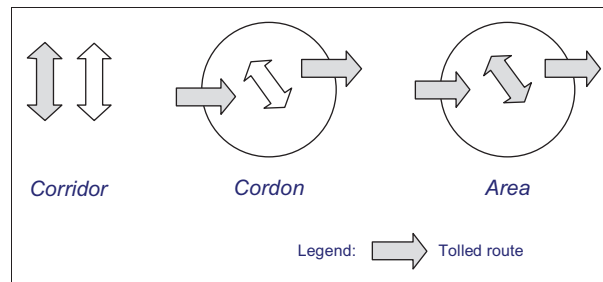


Figure 1 Various geographies of road pricing schemes.

In practice, perfect toll differentiation cannot be attained because demand and cost elasticities knowledge is insufficient. However, some differentiation level has become recently feasible due to progress in (electronic) technology.

When it goes to practical implementation current real-world congestion pricing scheme can be characterized according to three basic parameters, which are its geography, operating hours, and rate structure.

Regarding the geography of a road pricing scheme three various forms can be distinguished (Fig. 1): one is the corridor pricing where users pay for traveling on the corridor road while parallel free routes may exist (e.g., Express Lanes in the United States); a second one is a cordon encompassing an area, generally a city center, where traffic crossing the cordon must pay a toll, whether inbound (e.g., in Oslo), outbound or both (e.g., in Stockholm); the third one is area pricing where traffic entering and driving inside the area must pay the toll (e.g., in London).

The second parameter is the operating hours. Toll can be enforced every day (e.g., Express Lanes) or only on weekdays at some times.

The third parameter is the toll rate structure. The rate may be flat (e.g., in London) or varying along times of travel (e.g., in Stockholm or Singapore).

Consequently behavioral alternatives may vary with various congestion pricing schemes. When facing corridor pricing the user can switch to another route. In contrast, when facing cordon or area pricing and if the destination cannot be changed like in commuting, road user cannot escape from paying a toll unless switching to another travel mode. For other trip purposes destination change can be an option. When facing time-varying pricing the user can reschedule his trip in order to save money. Finally, trip may be avoided when possible.

When it comes to technology, toll collection can use today wireless communications (dedicated short-range communications, DSRC) between a roadside beacon (e.g., a gantry acting like a control point) and the vehicle which passes it: vehicles must be fitted with an inboard unit (transponder) generally behind the windshield, with a properly inserted stored-value or credit card (e.g., in Singapore) or linked to a prepaid account (e.g., in California). Toll collection can also use automatic number plate recognition (ANPR) by cameras located at control points with various invoicing systems (e.g., in London).

One potential additional parameter is traveled distance or time-based pricing: this could be done today crudely by tracking vehicles when passing at various control points or in the near future, more accurately, with satellite technology tracking moving vehicles.

Case Studies of Ongoing Congestion Pricing Schemes

We first present three iconic schemes, starting with the oldest one, that is, in Singapore (1975), and then more recent ones in London and Stockholm in the years 2000. To these we add the “value pricing” schemes in the United States and then end by a brief summary of other schemes.

Singapore Electronic Road Pricing Scheme (ERP)

This is a multi cordon-pricing scheme in operation in its current form since 1998. Its objective is to manage traffic through road pricing so as to maintain free flow traffic on roads (Chin, 2010). It applies to all vehicles and operates on all days except on Sundays and holidays. It follows the 1975 mono-cordon paper-based and manually enforced pricing scheme targeting vehicles entering the central business district (CBD, 6 km²). There are now (in March 2019) more than 60 gantries (control points) disseminated on major roads or expressways of the island, in entry to the CBD and in crossing two cordons inside the CBD. ERP operates with DSRC. Charges and operation hours are different on the various gantries (they may operate from 5 h 30 min to 22 h 30 min except on Sundays and holidays). For instance gantries may charge a car from zero to 6 S\$ (3.9 €) depending on the hour of passing. During the day the rate may vary every 5 min, because of the fine-tuning according to the objective of congestion management. However, these rates are known in advance and revised every three months with the objective of maintaining a speed range of 20–30 km/h on arterial roads and 45–65 km/h on expressways. Gantries are also gradually added on different points including an outer cordon, based on the observed levels of congestion.

Singapore is an island-city of 716 km² with more than 5 million inhabitants and a GDP per capita just below that of United States. Land constraints, a high density and the objective of economic leadership are since the beginning the main drivers of its urban and transport policy. The 1975 “area licensing scheme” was operating initially only in the morning peak hour, in view of maintaining good traffic conditions in the CBD and hence attracting tourists, congresses and businessmen to foster investment. To this add a high level for vehicle annual tax (e.g., 650 € for a vehicle of 2 L engine capacity) and a restrictive policy for importing vehicles (Singapore has no automobile industry). Since 1990, quotas on importations are set by the Land Transport Authority (LTA) and the potential buyer of vehicle must bid for a “certificate of entitlement” (COE) for having the right to own a vehicle. For instance, in January 2019 the COE costs 33,000 S\$ (i.e., 21,500 €) for a vehicle of 2 L engine capacity. This package of measures concerning the buying, owning, and use of automobiles has succeeded with no doubt in maintaining traffic free flow even at peak hours. Moreover the government controls firmly land use, housing, and transport developments, making their matching effective and hence public transport efficient.

The government stresses that ERP is for traffic management and not for revenue raising. Thus the other vehicle taxes have been lowered in order to maintain revenue neutrality. The efficiency of ERP is evaluated with reference to observed travel speeds on roads. These two characteristics build the acceptability of the scheme. There is no explicit economic evaluation but the LTA manages to contain the ERP operation costs. Switching the ERP technology to a new satellite-based one is planned in 2020.

London Congestion Charging Scheme

The area-pricing scheme in London is in operation since February 2003 and applies to private cars and commercial vehicles (Leape, 2006; Santos and Fraser, 2006). Its objectives are to reduce congestion on roads, improve public transit by bus and the reliability of travel time, and make urban delivery freight and other urban services more efficient. Revenues are earmarked for spending in public transport. It was applied initially on an area of 22 km² encompassing the historic city and financial district. The area was extended further in 2007 but then came back to the initial one in 2011. The current charge of 11.50 £ (13.4 € in March 2019) is a day pass applicable for driving as much as one likes inside the area from Monday to Friday between 7 and 18 h except for bank holidays and celebrations. Enforcement is achieved through ANPR by cameras disseminated at entry points and inside the area. Payment can be made in advance by internet or various modes until midnight on the day of driving. Discounts are offered to those who pay on a monthly or annual basis. Vehicles exempted are two-wheels, taxis, vehicles for the disabled, vehicles for collective transport (at least nine seats), and low energy vehicles. Residents living inside the area can apply for a package of 4 £ for five consecutive days (i.e., a 90% discount). An emission surcharge of 10 £ (marketed as “T-charge”) targeting older vehicles added to the congestion charge and following the same operating rules and exemption policy was introduced in October 2017. This emission surcharge is to be replaced in April 2019 by the ultra low emission zone with tighter emission standards and operating on a 24/7 basis.

Great Britain has virtually no road tolls. The legislation authorizing the implementation of road or parking pricing schemes by local governments was introduced in 1999–2000. At this time, the transport conditions in London were recognized as especially bad, following decades of underinvestment. The discussion about congestion pricing started with the Smeed report in 1964 and has been since ongoing, sustained with numerous researches and modeling studies. This has resulted in a broad consensus in the opinion and decision makers on the necessity to “do something” to reduce automobile traffic. This period coincided with the first election of a mayor for the Greater London. Ken Livingstone, linked to the Labour Party, announced during the election campaign that he would implement congestion charging. He won the election in 2000, implemented the scheme and was reelected in 2004 after announcing in the new election campaign an increase of the charge and an extension of the charging area.

The project was initially fiercely opposed by the conservative party, automobilists, trade unions, and inhabitants of the target area. One of the decisive factors in acceptability was the discount for inhabitants of the charging area and the exemption of vehicles for disabled. The mayor conferred abundantly with stakeholders, using various media. The discussions influenced the main parameters of the scheme, that is, the level of the charge, the hours of operation, and the area delimitation. However, it was never question of submitting the decision to a referendum. According to polls, before the launching of the scheme the opinion was half-divided between supporters and opponents of the project.

The charging area represented 1.5% of the Greater London surface and 5.2% of its population in 2003, but 26% of the jobs. At this time in the morning peak hour, less than 15% of people were taking the car to enter the planned charging area. Expected positive effects became apparent right from the start of the scheme. Vehicular traffic in charging area even decreased beyond what was expected by modeling studies, a drop not compensated by the slight increase of traffic around the area. A higher punctuality of bus services was observed resulting in an increase in patronage and a decrease in operating costs. The unexpected success in lowering traffic resulted in a negative impact on expected revenues. There seemed to be no evident effect on the economy whether negative or positive. Most of economic studies conclude to a net economic benefit despite the higher expected costs of setting up and operating the scheme. However, 4 years later congestion has reverted to pre-charging level: this can be attributed to increased roadwork and reallocation of road capacity to the detriment of cars.

Stockholm Congestion Tax Scheme

The cordon-pricing scheme in Stockholm started with a seven months trial from January 2006 and then became permanent from August 2007 (Eliasson, 2009; Börjesson et al., 2012). Its objectives are to reduce congestion and improve the environment. The cordon encompasses the Stockholm city heart (35 km², about 280,000 inhabitants). Stockholm city extends on several islands linked by bridges, which make the control of traffic easier through 20 “control stations.” The scheme operates from Monday to

Friday between 0630 and 1830 except on bank holidays and in July. A payment is due at each crossing of a control point inbound and outbound depending on the time of crossing: as of 2019 the price varies between 11 SEK at off-peak hours and 35 SEK at peak hour morning or evening (3.3 € in March 2019). However the charge is capped to a maximum of 105 SEK per day. Enforcement is achieved through ANPR at the gates and an invoice is sent to the owner of the vehicle. Vehicles exempted are two-wheels and vehicles for collective transport. Until 2008 alternative fuel vehicles were also exempted (and this exemption has proven to be the most important incentive to “clean cars” sales apart from other national incentives). The charges are tax-deductible for commuters: this amount to 60% reduction of the charges. The initial cordon was extended in 2016 to include a bypass motorway and the charge raised by 75% at peak hour.

Road pricing was initially allowed in Sweden only for financing new infrastructure. After first unsuccessful discussions in the 1990s on an investment plan for funding roads in the Stockholm region which would have been abounded by cordon pricing revenues, the idea of introducing cordon pricing on a trial basis was set forth by the municipality in 2003, based on a political alliance of the Social-Democrats and Green party at the national level. The law was changed in 2004 to allow for congestion charging, legally a tax to be approved by the national parliament. Support to the project was in a minority right before the start of trial but went in a majority after. A referendum was held in the city of Stockholm after the end of the trial (on the same day of local and national elections) and resulted in a majority for keeping the charges, while other votes organized in cities around resulted in a majority of opponents. The new government (new majority of center-right) decided to go on with the congestion charging while earmarking the revenues for road investment within an integrated package, including improvement for public transit and complemented by national funding.

The expected objectives are met since there is a decrease in vehicular traffic and pollutants emissions. These effects were still lasting at the same level five years later. Support by inhabitants of the Stockholm County is still around 60% despite a moderate decrease, following the charge increase in 2016. No effects were found in the retail sector. Most of economic studies conclude to a net economic benefit despite the high investment and operating costs of the scheme.

Value-Pricing in the United States

“Value pricing” is the official name in the United States for the federal congestion pricing pilot program (DeCorla-Souza, 2004). This gives a positive connotation to congestion pricing, with the idea that by increasing flow and speed on tolled infrastructures, their users benefit from improved service quality and this increases the infrastructure value for the community. Currently, there are three kinds of implementation: granting access to a high occupancy vehicles lane (HOV) for solo drivers accepting to pay a toll, thus making it into a high occupancy toll lane (HOT); creating lanes adjacent to existing ones and tolling them in order to finance their building; implementing variable tolling on existing (flat) tolled facilities.

While being generally successful HOV lanes are often perceived as underutilized because of the speed they offer and making them into HOT lanes is a way to increase acceptability of reserved lanes. One of the first examples is that of Interstate 15 North-South in the San Diego region (California). Initially it was an 8 miles stretch of two reversible HOV lanes. The HOT scheme started at the end of 1996 and is in operation on a 24/7 basis. The rate is variable and responsive since it depends on actual traffic in the lanes. The objective is to maintain a free flow on the HOT lane. In March 2019, the charge varies from 50 cents to a maximum of \$8 (7 €). Rate at the time of driving is displayed on signs along the freeway before entry to the HOT lane, allowing the driver to choose it or not. The lane remains freely accessible for carpools, vanpools, transit, and zero emission vehicles. The I-15 capacity has since been extended (and is currently being extended) according to the following “express lanes” principle.

The (first) Express Lanes SR 91, East-West in the same region, are four new lanes (two in each direction) added to the existing highly congested SR91 linking the Riverside and Orange counties to Los Angeles. 280,000 vehicles are traveling between the two counties on this expressway each day. The first express lanes were opened in the end of 1995 and since extended to an 18 miles stretch in 2017. These new lanes are fully funded by the tolls. Tolls operate on a 24/7 basis and vary according to the time of travel, direction, and trip length with hourly tolls between \$3.55 and \$22.25 for traveling the entire stretch (in March 2019) with reduced rates on holidays. Peak hour rates are revised every 6 months with the objective to optimize traffic at free-flowing speeds. Vehicles with at least three occupants can use the express lanes at free or with a 50% discount depending on the time and direction.

As of 2012, there were 24 pricing schemes applying variable pricing on freeways (including HOT and express lanes), highways, bridges, and tunnels in the United States. Since HOV are allowed also on new-built tolled lanes, the distinction between “express lanes” (marketing name) and “hot lanes” (technical name) becomes blurred.

Other Ongoing Schemes

Other congestion pricing schemes are in operation in the world. Italy has a tradition of limited traffic zone (ZTL), restricting the access of vehicles to historical centers in several cities. Milan switched from its ZTL (an 8 km² area) to a cordon-pricing scheme named “Ecopass” in 2008, operating on weekdays (Beria, 2016). It was environmentally oriented through a pricing structure depending on emission standards of vehicles. Due to the growing share of traffic by clean vehicles exemption from payment, the scheme evolved in 2012 after a referendum to “Area C” with a flat rate during weekdays. Durham (United Kingdom) and Valletta (Malta) also operate restricted access through pricing to their historical center (area of less than 1 km²).

Major cities in Norway have implemented cordon tolls charging inbound traffic initially with the main objective of raising revenues to complement funding for road infrastructure and public transport (Ieromonachou et al., 2006). In Oslo, the scheme was launched in 1990, despite a public opinion predominantly against, operating on a 24/7 basis with a flat rate: it was low enough to

avoid social exclusion issues. In 2008, an outer toll ring was added and recently the rate structure changed, with a higher one in peak hours (morning and afternoon) on weekdays, paving the way to congestion charging.

Like in the United States, several tolled facilities in the world apply variable pricing by adding a markup at peak periods and sometimes reducing the toll at others (e.g., three tolled motorways in Paris region).

Overview

Despite these ongoing successes, few cities in the world have actually implemented congestion charging. The study of failures to do so is also instructive for candidate cities. In Lyon, a scheme combining a new urban tolled motorway with reduction in capacity of free parallel routes was cancelled in 1998 because of fierce public opposition a few months after its launching, leading to the reopening of free routes, buying back of the new infrastructure by the community and reductions of toll rates (Raux and Souche, 2004). Hong Kong failed to actually implement electronic road pricing, despite pilot test on the electronic road pricing system between 1983 and 1985, one issue being fear about invasion of privacy. Several near real implementation projects have been rejected because of fierce public opposition explicitly expressed through ex ante referenda like in Edinburgh (2005) or Manchester (2008). Numerous cities in the world are continuously debating about road or congestion pricing schemes since they are facing excess traffic on their roads and funding shortage for their transport system.

The first lesson that can be drawn from the ongoing schemes is that they are effective in improving traffic conditions: drivers are sensitive to cost, traffic flow is smoother and its harmful effects on the environment are reduced. Most of the appraisal studies also consider that these schemes are efficient, that is to say they improve the welfare of the community. However, this improvement appears on one hand lower than expected, given the unexpected high costs of setting up and operating the charging system. On the other hand, in cases of time-varying pricing the benefits may be underestimated given the possibility of rescheduling trips, a phenomenon which is difficult to introduce in conventional appraisal.

Successes and failures also indicate some basic conditions for a successful implementation. Regarding the political context this can be a strong political will backed by a lasting debate (as in London). This can also result from agreement between political parties backed by leveraging on national funding and a significant public support of environmental policies (as in Stockholm or Oslo). However, even if a political agreement is reached this does not exempt the local authority from taking high political and financial risks, as this was obviously the case with the seven months trial in Stockholm.

The success in Stockholm (and Milan) and the failure of some proposals elsewhere indicate that a referendum before any implementation is a sure way to get a rejection by the public. The study of opinion attitudes in the successful schemes shows that opinion is highly skeptical with a rising opposition until the scheme launching and then becomes quickly supportive after the scheme delivers smooth toll collection, operation, and benefits in traffic flow. The practical experience of congestion pricing reduces the legitimate fear about something which was previously unfamiliar.

Even if political acceptability can be reached through compromises between parties, public acceptability may be more difficult to achieve. Equity issues can be highly debated before the scheme acceptance. Research shows that there is no general diagnosis since it depends on the peculiarities of the scheme (where people live and work, which socioeconomic profiles will have to pay, etc.). Earmarking revenues and redistribution are seen as critical issues.

Congestion must be at a critical level to justify the setting up of a costly toll collection system. To alleviate congestion in other cases alternative policies are available, such as parking, regulating, pricing or developing bicycling, and carpooling along with transit (Arnott et al., 2005).

References

- Arnott, R., Rave, T., Schöb, R., 2005. *Alleviating Urban Congestion*. MIT Press, Cambridge, MA.
- Beria, P., 2016. Effectiveness and monetary impact of Milan's road charge, one year after implementation. *Int. J. Sustain. Transp.* 10 (7), 657–669, doi:10.1080/15568318.2015.1083638.
- Börjesson, M., Eliasson, J., Hugosson, M., Brundell-Freij, K., 2012. The Stockholm congestion charges—5 years on. Effects, acceptability and lessons learnt. *Transp. Policy* 20, 1–12.
- Chin, K., 2010. The Singapore experience: the evolution of technologies, costs and benefits, and lessons learnt. ITF Roundtable 147. *Implementing Congestion Charges*. OECD, Paris.
- DeCorla-Souza, P., 2004. Recent U.S. experience: pilot projects. In: Santos, G. (Ed.), *Road Pricing: Theory and Evidence*. Elsevier, Oxford, UK, pp. 283–308.
- Eliasson, J., 2009. A cost-benefit analysis of the Stockholm congestion charging system. *Transp. Res. Part A* 43, 468–480.
- Ieromonachou, P., Potter, S., Warren, J.P., 2006. Norway's urban toll rings: evolving towards congestion charging? *Transp. Policy* 13, 367–378.
- Leape, J., 2006. The London congestion charge. *J. Econ. Perspect.* 20 (4), 157–176.
- Raux, C., Souche, S., 2004. The acceptability of urban road pricing: a theoretical analysis applied to experience in Lyon. *J. Transp. Econ. Policy* 38 (2), 191–216.
- Santos, G., Fraser, G., 2006. Road pricing: lessons from London. *Econ. Policy* 21 (46), 264–310.

Further Reading

- International Transport Forum, 2010. *Implementing congestion charges*. Roundtable 147. OECD, Paris.
- Santos, G., Rojey, L., 2004. Distributional impacts of road pricing: the truth behind the myth. *Transportation* 31 (1), 21–42.
- Schade, J., Schlag, B. (Eds.), 2003. *Acceptability of Transport Pricing Strategies*. Elsevier, Oxford.
- Small, K.A., Verhoef, E.T., 2007. *The Economics of Urban Transportation*. Routledge, Abingdon, UK.

Distributional Effects of Congestion Charges and Fuel Taxes

Jonas Eliasson, Department of Science and Technology, Division of Communications and Transport Systems, Linköping University, Norrköping, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	139
Methodological Questions	139
Should Revenue Recycling be Included in the Analysis?	139
Income or Expenditures as a Measure of Economic Status?	140
Must Behavioral Adaptation be Taken Into Account?	141
Must Second-Order Effects be Taken Into Account?	141
Two Examples	141
Example: Fuel Tax	141
Example: Congestion Pricing	142
A Sample of Empirical Results	143
Conclusions	144
See Also	144
References	145
Further Reading	145

Introduction

Fuel taxes and congestion charges have repeatedly been shown to be highly effective policy instruments to reduce traffic emissions and road congestion, respectively. However, a recurring argument against them is that they are claimed to fall disproportionately on the poor. This chapter analyses this argument. For brevity, fuel taxes and congestion charges are referred to as “car use taxes.” Most of the discussion in the chapter is just as relevant for other kinds of car-related taxes, such as vehicle taxes, parking charges and vehicle sales taxes.

The purpose of a car use tax matters greatly for what conclusions are drawn. Many car use taxes, in particular when fuel taxes were first introduced, have been fiscally motivated: they are simply a convenient way to raise revenues for various public expenditures. In such situations, it is clear that the taxes’ distributional burden is relevant, and should be compared to other ways to raise public revenues, such as income, sales or property taxes. But more and more, car use taxes are seen as price corrections: they are motivated by a desire to make the cost of driving better reflect its total social cost, including externalities such as carbon emissions and road congestion. In other words, this kind of car use tax adjusts the price of driving to what it really should be; without it, driving is subsidized from a social point of view. From this perspective, it is much less clear in what sense distributional effects of car use taxes are relevant, and between which situations comparisons should be made. Prices are almost always the same for everyone, regardless of income or wealth, for two good reasons. First, it lets the individual decide for herself how to allocate her resources (money, time, etc.) between different goods and services. Second, it leads to an overall efficient allocation of resources across the economy through supply-and-demand mechanisms. Desires for increased income equity is instead usually handled by taxation and social welfare systems. Accepting a default position where prices are, generally, equal for everyone (with a few deliberated exceptions), it is natural to argue that the distributional effects of corrective taxes—taxes which are introduced to make the prices “right” in the sense that they reflect full social costs—are in fact essentially irrelevant. Indeed, allowing prices of car trips to be lower than their social cost (which they will be in the absence of car use taxes) effectively constitutes subsidies from society at large to car drivers, and these implicit subsidies accrue mostly to rich groups (Eliasson, 2016).

This being said, analyzing distributional effects may still be important, partly because most car use taxes have at least some fiscal motivation as well, and partly because any change in an existing price system causes transition costs when people adapt to the new prices.

Methodological Questions

When analyzing the distributional effects of tax instruments, several methodological questions need to be considered, most of which have no clear-cut answers.

Should Revenue Recycling be Included in the Analysis?

The first question is whether the recycling of the revenues should be part of the analysis. It is important to realize that the distributional profile of a revenue-generating tax instrument is one thing, and the distributional profile of an expenditure scheme is something else. They can be analyzed either separately, or together as a single policy.

One answer is that it depends on the decision context. If the revenues will be spent on something that will be carried out in any case, it is natural to compare the distributional profiles of different possible tax instruments in isolation, leaving the distributional profile of the planned expenditure out of the analysis. On the other hand, if one is considering introducing an earmarked tax for a specific project which will not be undertaken otherwise—say, a congestion charge necessary to fund an infrastructure investment—it may be natural to consider the distributional profile of the tax and the project together, as a single policy.

However, there is a strong argument for analyzing tax instruments and expenditure schemes separately, namely that the natural and conventional definitions of what constitutes “distributionally neutral” schemes are usually defined differently for taxes and for expenditures.

The most common definition of a distributionally neutral tax instrument is one which takes an equal share of everyone’s income. A tax instrument is defined as progressive if it takes a larger share of rich people’s income than of poor people’s; the opposite is called a regressive tax. This notion is formalized in the Suits index (Suits, 1977), defined as $S = 1 - 2 \int_0^1 T(y)dy$, where y is the cumulative share of total income and $T(y)$ is the cumulative share of the total tax burden.^a The index is bounded between -1 and 1 . A flat-rate tax has Suits index 0 , a regressive tax has a negative Suits index and a progressive tax a positive index.

For public expenditures, on the other hand, the most common^b definition of a distributionally neutral scheme is one which gives an equal absolute amount (or value) to everyone, a so-called lump sum distribution. An expenditure scheme is defined as progressive, if it gives a larger amount per capita to poor people than to rich people, and regressive, if it is the other way around. This notion is formalized in the concentration index (Kakwani, 1977), defined as $CI = 1 - 2 \int_0^1 s(x)dx$ where $s(x)$ is the share of total spending accruing to the poorest $x\%$ of the population. The concentration index is also bounded between -1 and 1 , just as the Suits index. If all citizens receive the same amount (lump sum spending), the index is zero. Progressive spending (more is spent per capita on low income groups) yields a negative concentration index, and vice versa.

Now, note that the definitions of distributional neutrality are different for taxes and expenditures: a neutral tax takes an equal share of everyone’s income, while a neutral spending scheme gives an equal absolute amount to everyone. This easily leads to paradoxical results when analyzing combinations of a tax and a revenue-recycling scheme as a single policy. For example, combining a neutral tax (a fixed share of everyone’s income) with a neutral expenditure scheme (a lump sum redistribution) turns out to be a progressive policy when seen as a combined policy, not a neutral one. It follows that it is easy to construct examples where a regressive tax combined with a regressive spending scheme is defined as a progressive policy when taken together and viewed as a single scheme, and vice versa: a progressive tax and a progressive spending scheme may be regressive when taken together. This is a strong argument for analyzing distributional effects of tax schemes and expenditure schemes separately, as this prevents this kind of confusion.

It is not uncommon that studies conclude that a car use tax is regressive, but together with lump sum revenue recycling the total effect is progressive. This mixes the two different definitions of “progressive”/“regressive” explained earlier, ending up in a conclusion that is completely trivial on a closer look. Taking an equal share of everyone’s income (a neutral tax) and handing the revenues back with a lump sum distribution (neutral spending) is of course a highly progressive policy combination to start with. That a tax instrument is not regressive enough to make the combination with a lump sum redistribution regressive is hardly surprising; if such a combination was regressive, the tax has to be extremely regressive in itself, effectively taking higher equal amounts in *absolute* terms from poor people than from rich people. Again, this shows that there are strong arguments for keeping distributional analyses of public revenue sources and public expenditures separate. In certain specific decision situations, however, it may still be natural to also consider a combined tax and spending scheme; the most common example is a car use tax earmarked for a project that will for certain not be undertaken otherwise.

Income or Expenditures as a Measure of Economic Status?

The second methodological question is how to define and measure individuals’ available economic resources. One way is to simply use disposable income, that is, the net sum of after-tax wages and transfers in a month or a year. However, this ignores that many people have other sources of money available to them. People may live off their savings or other sources of wealth, be supported by relatives (parents or a spouse), or have unregistered income sources. Moreover, some people may expect to have higher earnings in the future than they currently have, leading them to behave as if they borrow against their future income; this is especially relevant for students. Finally, income varies a lot between years, especially at the extremes. For example, someone selling a house or a company one year will have a very high income that particular year, but probably not nearly as high the next year. At the other extreme, some people may have extremely low incomes one particular year because they take a year off to study, take care of children or write a book, but in such cases their income are probably considerably higher other years. All this means that disposable income, in the usual sense, is not necessarily a full and fair measure of an individual’s economic situation.

A way around this is to use individuals’ expenditures as a proxy measure of their long-run available economic resources. An obvious drawback is that such studies must be based on the survey data rather than registry data, and registry data usually gives much bigger and more precise data sets. Studies suggest that using expenditures rather than disposable income as a measure of economic

^a In applications, data is usually given in discrete form for individuals or groups. Indexing these discrete observations by i , the Suits index is approximated by:

$$S = 1 - \sum_i (T(y_i) + T(y_{i-1})) (y_i - y_{i-1})$$

^b There are studies and contexts, however, where neutral spending is defined as a scheme where each individual gets an amount proportional to her income.

resources tends to make tax instrument look more neutral—progressive taxes become less progressive, and regressive taxes less regressive (Sterner, 2012).

Must Behavioral Adaptation be Taken Into Account?

A change in car use taxes will cause behavioral changes. This means that the welfare loss of a tax change will be accurately reflected neither by the total taxes paid after the change, nor by what would have been paid ignoring behavioral adaptation. The first alternative underestimates the welfare loss of a tax increase, as it ignores the loss in utility caused by adapting behavior, and conversely the second alternative overestimates the welfare loss as it ignores the possibility to adapting and hence partly avoiding the tax. It follows that only measuring the tax incidence, that is, how much tax different groups pay, may give misleading conclusions, as this neglects the welfare loss of behavioral adaptation.

Clearly, it is preferable to use a proper welfare measure—the Marshallian or ideally the Hicksian consumer surplus (see the chapter by Harald Minken in this volume)—rather than simply using taxes paid. However, this is not always possible, as it requires forecasting behavioral adaptations to the tax. Fortunately, the error induced by neglecting adaptation costs is usually relatively small. If a tax is increased by some fraction α and the cost elasticity of demand is ϵ , the relative error of the welfare loss if adaption is ignored is $\frac{\alpha\epsilon}{2}$. So, if a tax is increased by $\alpha = 10\%$ and the cost elasticity is $\epsilon = -0.5$, the relative error is 2.5%, which is negligible in most situations. Obviously, if the change is relatively large and demand elasticities are high and different across groups, the different between welfare loss and change in taxes paid may not be negligible anymore.

Must Second-Order Effects be Taken Into Account?

A change in car use taxes may change the prices of other goods and services as a second-order effect. This could mean that even individuals who do not travel by car are affected, as the prices of goods and services they consume may change. One case where this can matter is the price of public transport in poor countries, as diesel costs make up a substantial share of transit operating costs and poor groups make a much higher share of their trips by public transport than by car (Agostini and Jiménez, 2015; Blackman et al., 2010). This means that neglecting the second-order effect of a fuel tax increase on public transport prices, only considering the direct effect on driving costs, may underestimate the impact on poor groups.

Two Examples

Consumption taxes are usually slightly regressive, as high-income groups tend to spend a smaller share of their income on consumption, and more on savings. General sales taxes typically have Suits indices in the range -0.1 to -0.2 . Whether a consumption tax on a particular good is progressive or regressive depends on whether consumption of that good increases faster or slower than proportionally to income. In other words, a consumption tax will be regressive, if the consumption elasticity with respect to income is lower than 1, and vice versa.

Broadly speaking, studies suggest that the income elasticity of car use is slightly lower than 1 in rich countries and slightly higher than 1 in poor countries. This means that car use taxes tend to be mildly regressive in rich countries but mildly progressive in poor countries. Obviously, results will differ depending on the design of the tax, the context, what type of car use is taxed and so on.

To illustrate some fairly typical results and how distributional analyses can be carried out, two empirical case studies are presented below: a fuel tax increase and a congestion charge. While the specific results obviously pertain to these specific cases, the reasoning is general, and the general findings are fairly representative for most studies.

Example: Fuel Tax

The following case study shows results for approximately 10% increase of the Swedish fuel tax (reported in Eliasson et al., 2018). Distributional impacts are calculated as welfare losses relative to disposable income, where incomes are taken from the tax registry. Driving distances are taken from the vehicle registry, and vehicles' fuel consumption from vehicle type registrations. Welfare losses are calculated using demand elasticities estimated separately for different combinations of income quartile and type of residential area (large cities, small cities, and rural areas). Elasticities are medium-term, meaning that they consider changes in vehicle kilometers driven, but not changes in residential location or changes in vehicle characteristics (fuel consumption). Revenue recycling is not considered. Results are presented for combinations of income octiles and residential area.

Fig. 1 shows that the welfare loss increases as a share of disposable income for most of the income range (octile 2-7). The pattern is different for octiles 1 and 8, however. The result for octile 1 should be treated with caution; most incomes in this group are well below the threshold for social welfare in Sweden and therefore cannot really reflect individuals' real access to money. In octile 8, incomes are so high that car use cannot reasonably increase in proportion to income. The regressivity/progressivity of the fuel tax increase is hence different across the income distribution: between octile 2 and 7 it is progressive, but in the low and high tails, it is regressive. The Suits index for the entire income range shows that the fuel tax increase is slightly regressive overall; this is caused by the result for the highest octile. This also hints at why a fuel tax is often slightly regressive in rich countries, but progressive in poor



Figure 1 Welfare loss of the fuel tax increase, relative to income, by type of residential location.

countries: in rich countries, car use in the highest income groups tend to reach a saturation level above which car use only increases slowly when income increases further. This means that a fuel tax' share of income decreases in the highest income segments.

The variation in paid fuel tax across income groups is mostly due to differences in car ownership, and not so much due to differences in car owners' driving distances or vehicles' fuel consumption. This means that if one (for some reason) would consider only car owners, a fuel tax would be strongly regressive (Fig. 1).

Fig. 1 also shows that there are considerable differences between large cities, small cities, and rural areas. A more detailed analysis of shows that residents in satellite cities, which serve as "suburbs" to a region's functional center, pay more in fuel taxes, and that such functional relationships between cities explains more of the variation than just population sizes.

However, these results only present average impact per group, which hides the fact that the variation within each income group is substantial. An income tax will, by definition, affect everyone with the same income in the same way. A car use tax is different: even if it is progressive "on average," there may still be individuals who are hurt disproportionately relative to their income. In fact, a more detailed analysis shows that the share suffering substantial welfare losses relative to their income is much higher in low-income groups than in high-income groups—despite that the average welfare loss relative to income is lower in the lower income groups. This is especially true for low-income groups in rural areas. This may explain the feeling that car use taxes hurt the poor disproportionately: not that they are regressive on average, but that the share who suffer substantial welfare losses relative to their income are higher in lower income groups. That this point seems to be underappreciated is partly a data issue: exploring the variation within groups requires large data sets, and is often impossible without access to registry data, since survey- or modeling-based data sets are usually not sufficiently large.

The argument that members of a group should be affected equally is sometimes called "horizontal equity." How this argument should be applied in the context of car use taxes depends on the purpose of the tax, as argued in the introduction, as this implies how "groups" should be defined. If the purpose is primarily to generate public revenues, it is natural to define "groups" as income segments, and consider distributional effects across and within income groups. If the purpose is to correct the price of car trips, on the other hand, it is natural to define "groups" according to how much people drive, and distributional effects across and within income segments are much less relevant.

Example: Congestion Pricing

Stockholm introduced congestion charges in 2006, first as a trial, and permanently from 2007 (Eliasson, 2008). The charging system consisted of a cordon around the inner city, with charges varying between 2€ in peak hours and 1 € before and after the peaks (nights and weekends are free of charge). The system was slightly revised in 2016, when peak charges were increased and one charging point was added, but the analysis presented here refers to the original system (Fig. 2).

Fig. 3 shows the same as a proportion of monthly income. The data comes from a travel survey (RVU 2015). Income is self-reported total household income before tax, divided by the number of adults in the household. Revenue recycling is not considered. The payment distribution is smoothed through kernel estimation (a generalization of the "moving averages" method).

Average congestion charge payments per person are almost proportional to income, except for the lowest and highest incomes. As before, results for the lowest income groups should be treated with caution, as these incomes are so low that they can hardly reflect available economic resources. In the highest income range, it is as if car use almost reaches a saturation level where it no longer increases with income, and hence payments as a share of income falls somewhat for the highest income range.

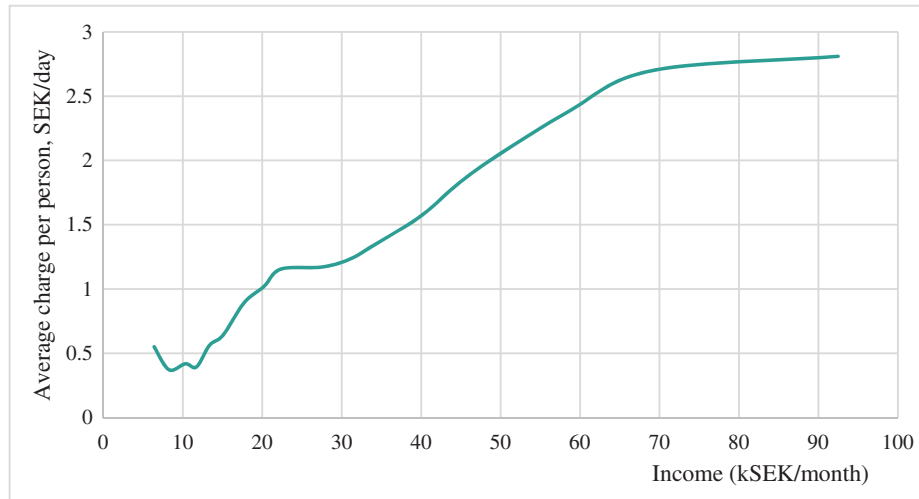


Figure 2 Average congestion charges paid per person and day (kernel estimation).

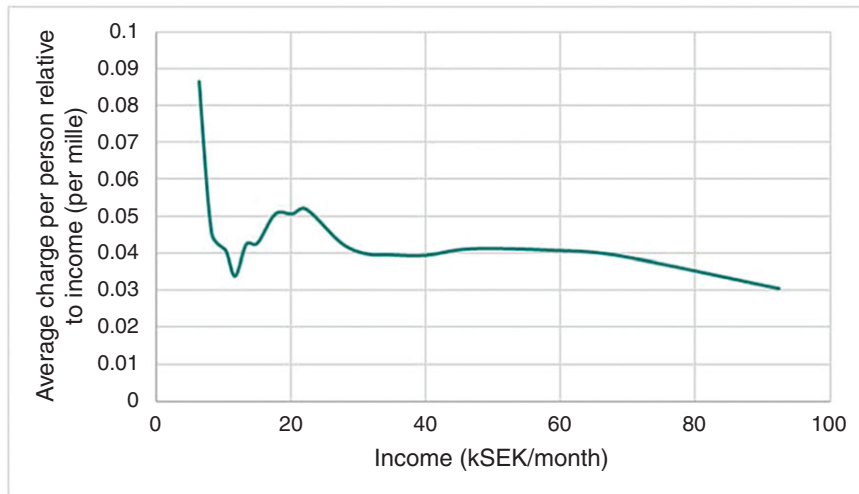


Figure 3 Average congestion charges paid per person and day relative to monthly income (per mille) (kernel estimation).

It is evident from [Fig. 3](#) here that the Stockholm charge is slightly regressive, as charge payments as a share of income falls slowly with income: the overall Suits index is -0.09 . It is also clear from the figures that the slight regressivity is almost entirely due to the results at the extremes of the income range, whereas for most of the income range, charge payments are roughly proportionally to income. Just as for the fuel tax, however, these averages obscure the fact the variation with an income group is substantial.

Analyzing charge payments geographically (not shown here) shows that the congestion charge is more regressive for residents close to the charging zone—especially for residents within the inner city—while it is progressive for residents further away from the zone. This illustrates that the geographical distribution of socioeconomic groups and travel patterns matter, and hence results will differ between cities with different socioeconomic spatial distributions.

A Sample of Empirical Results

The distributional effects of any consumption tax will depend on the local context, and car use taxes are no different. [Table 1](#) shows a sample of empirical studies of distributional effects of fuel taxes, illustrating a representative range of results summarized by Suits indices. In rich countries, fuel taxes tend to be slightly regressive, but the regressivity decreases, if expenditures are used instead of income as a measure of individuals' economic resources. In poor countries, fuel taxes tend to be progressive, and this progressivity also tends to decrease when expenditures are used as a measure of individuals' economic resources. Taking second-order effects of fuel taxes into account affect results mostly to a small extent ([Table 1](#)).

Table 1 Suits indices for fuel taxes from empirical studies

<i>Country</i>	<i>Income, excluding second-order effects</i>	<i>Income, including second-order effects</i>	<i>Expenditures, excluding second-order effects</i>	<i>Expenditures, including second-order effects</i>	<i>Source</i>
France	−0.155	−0.157	−0.021	−0.024	Stern (2012)
Germany	−0.066	−0.067	0.009	0.008	Stern (2012)
Italy			−0.110	−0.110	Stern (2012)
Serbia	0.187	0.172	0.066	0.055	Stern (2012)
Spain	−0.086	−0.086	−0.002	−0.002	Stern (2012)
Sweden	−0.171	−0.178	0.072	0.064	Stern (2012)
United Kingdom	−0.123	−0.125	−0.003	−0.004	Stern (2012)
Texas (US)	−0.25				CPPP (2007)
Costa Rica	0.09	−0.01			Blackman et al. (2010)
Sweden	−0.03				Eliasson et al. (2018)
Chile	0.05		0.17		Agostini and Jiménez (2015)

Conclusions

Car use taxes have repeatedly been shown to be very effective policies to reduce emissions and congestion. Few, if any policies can compete in terms of effectiveness, and probably none in terms of economic efficiency.

The distributional consequences of car use taxes will obviously depend on their design and the local context. It is clear, however, that rich groups will pay considerably more per person than poor groups. Considering payments as a share of income, results are more mixed, but broadly speaking, average payments tend to be approximately proportional to income, but slightly regressive in rich countries and slightly progressive in poor countries. The overall regressivity tends to be caused by the outliers: the highest and lowest income groups do not drive quite proportionally to their (registered) income level.

Variation within income groups is often substantial, however. Car use taxes also tend to place a higher burden on residents in rural areas, satellite cities, and urban peripheries, which may counteract societal goals to make such areas more attractive. If the purpose of a car use tax is to generate revenues for public expenditures, variation with income groups, higher burdens in rural areas, and slight regressivity may be viewed as serious problems. After all, it is difficult to defend that poor or rural people should contribute more than proportionally to public expenditures. In this respect, income or general sales taxes can be viewed as more fair, as these by construction take an equal amount from everyone with the same level of income or consumption, respectively.

However, it is much less clear that such distributional effects are relevant if the purpose of a car use tax is to correct the prices of car trips to make them better reflect their full social cost, by for example internalizing the cost of congestion or carbon emissions. Prices of goods and services are usually equal for everyone, for good reasons (most importantly that it leaves it up to individuals themselves to decide how to allocate their resources). Problems with inequitable income and wealth distributions are instead usually (and preferably) handled with general taxes and the social welfare system. Allowing prices of car trips to be lower than their social cost (which they will be in the absence of car use taxes) effectively constitutes subsidies to car drivers from society at large, and these implicit subsidies will overwhelmingly accrue to rich groups. From this perspective, the burden of proof from a distributional point of view lies not on those who want to introduce corrective car use taxes, but on those who defend a situation where car use is effectively subsidized by society. This is of course an even more pressing problem in countries where the price of car fuel is actually subsidized with public money.

Obviously, it can be difficult in practice to figure out whether a particular car use tax should be viewed primarily as a price correction or primarily as a source of public revenue. Nevertheless, the two perspectives are important to keep in mind when drawing conclusions from an analysis.

See Also

Pricing Principles in the Transport Sector; Real-World Experiences of Congestion Pricing; Dynamic Congestion Pricing and User Heterogeneity; Car tolls, Transit Subsidies for Commuting, and Distortions on the Labor Market; Transportation Equity; Regulation and Financing of Toll Roads; The Taxation of Car Use in the Future

References

- Agostini, C.A., Jiménez, J., 2015. The distributional incidence of the gasoline tax in Chile. *Energy Policy* 85, 243–252, doi:10.1016/j.enpol.2015.06.010.
- Blackman, A., Osakwe, R., Alpizar, F., 2010. Fuel tax incidence in developing countries: the case of Costa Rica. *Energy Policy* 38 (5), 2208–2215.
- CPPP, 2007. Center for Public Policy Priorities. Who pays taxes in Texas? (No. 287).

- Eliasson, J., 2008. Lessons from the stockholm congestion charging trial. *Transport Policy* 15 (6), 395–404.
- Eliasson, J., 2016. Is congestion pricing fair? Consumer and citizen perspectives on equity effects. *Transport Policy* 52, 1–15.
- Eliasson, J., Pyddoke, R., Swärdh, J.E., 2018. Distributional effects of taxes on car fuel, use, ownership and purchases. *Econ. Transp.* 15, 1–15, doi:10.1016/j.ecotra.2018.03.001.
- Kakwani, N.C., 1977. Applications of Lorenz curves in economic analysis. *Econometrica* 45 (3), 719–27, doi: 10.2307/1911684.
- Serner, T., 2012. Distributional effects of taxing transport fuel. *Energy Policy* 41, 75–83, doi:10.1016/j.enpol.2010.03.012.
- Suits, D.B., 1977. Measurement of tax progressivity. *Am. Econ. Rev.* 67 (4), 747–52.

Further Reading

- Arze del Granado, F.J., Coady, D., Gillingham, R., 2012. The unequal benefits of fuel subsidies: a review of evidence for developing countries. *World Develop.* 40 (11), 2234–2248, doi:10.1016/j.worlddev.2012.05.005.
- Bento, A.M., Goulder, L.H., Henry, E., Jacobsen, M.R., von Haefen, R.H., 2005. Distributional and efficiency impacts of gasoline taxes: an econometrically based multi-market study. *Am. Econ. Rev.* 95 (2), 282–287.
- Casler, S.D., Rafiqi, A., 1993. Evaluating fuel tax equity: direct and indirect distributional effects. *Nat. Tax J.* 46 (2), 197–205.
- Eliasson, J., 2016. Is congestion pricing fair? Consumer and citizen perspectives on equity effects. *Transport Policy* 52, 1–15.
- Levinson, D., 2010. Equity effects of road pricing: a review. *Transport Rev.* 30 (1), 33–57, doi:10.1080/01441640903189304.
- Santos, G., Rojey, L., 2004. Distributional impacts of road pricing: the truth behind the myth. *Transportation* 31 (1), 21–42.

The Bottleneck Model

Dereje Abegaz*, Yili Tang†, *Department of Technology, Management and Economics, Technical University of Denmark, Lyngby, Denmark;
†California PATH, University of California, Berkeley, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

The Basic Vickrey Bottleneck Model	146
Extensions and Applications	147
Road Pricing	147
Valuation of Travel Time Variability and Travel Information	148
Heterogeneity in Values of Time and Preferred Arrival Time	148
Networks	148
Other Application Areas	148
Conclusion	149
References	149

The Basic Vickrey Bottleneck Model

To introduce the basic Vickrey bottleneck model, consider a hypothetical city with N identical inhabitants who reside in one building and must each commute in their own car to the same workplace along a single road segment, which is prohibited for traffic outside the city. Every where in its stretch, the road is wide enough to serve all the N commuters at a time except at one location, the bottleneck, where at most $\phi < N$ cars can pass at a time. As individuals arrive at the bottleneck continuously in the order in which they depart from home, a queue develops behind the bottleneck from the time when the number of arrivals exceeds bottleneck capacity.

Individuals arrive at work in the order in which they depart from home. The first person to depart from home experiences the shortest travel duration. The travel time for subsequent departures increases as they queue behind the bottleneck. The travel duration in the absence of delay is normalized to zero without loss of generality as it is the same for all individuals. This means that an individual arrives at the bottleneck as soon as he or she departs from home and be at work once he or she exit the bottleneck.

Travel time is undesirable with the cost per unit time spent traveling being $\alpha > 0$. The travel time T associated with departure time d is $T(d) = Q(d)/\phi$, where $Q(d)$ is the number of cars waiting behind the bottleneck (the queue length) at the time of departure from home. That is, the travel time is the ratio of the number of cars queuing behind the bottleneck and the capacity of the bottleneck.

Individuals are assumed to have a preferred arrival time at work, t^* , and that they dislike arriving earlier or later than this time.¹ In addition to the cost of travel time, a commuter incurs $\beta > 0$ per unit time spent at work before t^* and $\gamma > 0$ per unit time after t^* . The extent of earliness and lateness relative to the preferred arrival time are respectively called schedule delay early (SDE) and schedule delay late (SDL). The marginal cost of being late for work is assumed to be higher than the marginal cost of being early, that is $\gamma > \beta$, which is needed for existence of equilibrium.

Individuals choose departure time in order to minimize the generalized cost of the commute trip, which includes the cost of travel time and the cost of schedule delays:

$$\text{cost} = \alpha(a - d) + \beta \max(0, t^* - a) + \gamma \max(0, a - t^*). \quad (1)$$

where a is the arrival time at work and hence $(a - d)$ is travel time, $\max(0, t^* - a)$ is schedule delay early and $\max(0, a - t^*)$ is schedule delay late.

While the ideal departure time from an individual's viewpoint is that which leads to on-time arrival at work without facing a queue, this is however not feasible since other individuals have an incentive to depart slightly earlier in order to avoid delays. Hence, an individual departing at the ideal departure time will spend some time waiting behind a queue and arrive late for work. Let $\hat{d}$ be the departure time that leads to on-time arrival at work. Then, an individual who departs before $\hat{d}$ will be early for work and thus incurs SDE cost. In the same vein, an individual who departs after $\hat{d}$ arrives at work late and will therefore incur SDL cost. Each individual adjusts his or her departure time so long as there is an incentive to do so (Fig. 1).

Equilibrium is obtained when departure time choices are such that no individual can benefit by unilaterally adjusting his or her departure time. Suppose the morning peak period is between times t_0 and t_1 where $t_0 < \hat{d} < t_1$. The equilibrium, which is illustrated in Fig. 1 where the vertical gap between cumulative departures and cumulative arrivals indicates the queue length while the horizontal gap between them shows travel time, has the following properties:

1. The peak period is an interval of time duration which all individuals can pass through the bottleneck, i.e., $t_1 - t_0 = \frac{N}{\phi}$.
2. The first individual to depart from home arrives at work instantly but would be early by $t^* - t_0$.

¹ While the assumption of a homogeneous preferred arrival time was not made in the original Vickrey bottleneck model, the assumption is maintained here for ease of comparison with the literature and to simplify the exposition.

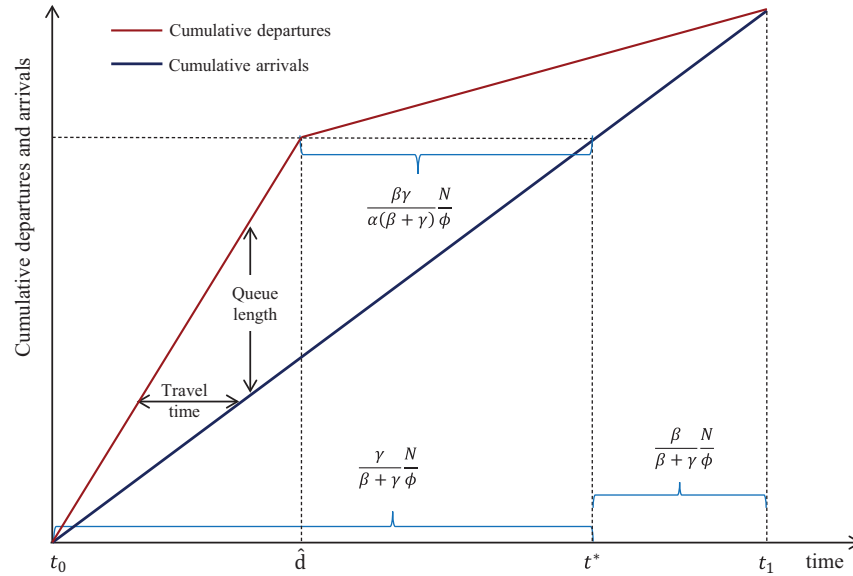


Figure 1 Illustration of the unpriced equilibrium

3. The last individual departing from home will arrive at work instantly but $t_1 - t^*$ time units later than desired.
4. If $\alpha > \beta$, everyone except the first and last individuals to depart from home will face bottleneck congestion.

Equilibrium quantities can be derived from the above conditions noting that, due to homogeneity, trip costs will be the same at all times during the peak period.² Accordingly, the trip cost of the first person to depart from home, hence everyone else by homogeneity, is $\frac{\beta\gamma}{\beta + \gamma} \frac{N}{\phi}$ while the total travel time costs and the total schedule delay costs are each equal to $\frac{\beta\gamma}{2(\beta + \gamma)} \frac{N^2}{\phi}$. It is worth to note the following:

1. While the equality the total travel time costs and the total schedule delay costs is implied only in the simple version of the model, the intuition behind is rather fundamental that travel time represents only part of the total congestion costs.
2. The total travel time costs, total schedule delay costs and hence the aggregate trip cost are each independent of the value of time. The implication of this is that the total schedule delay costs can be determined without knowledge on the value of travel time provided that the start and end times of the peak period are known.

Extensions and Applications

The basic bottleneck model has been applied in various areas and extended in various directions since Vickrey's seminar paper. Below we provide a concise overview of some of these developments. A detailed account of the literature can be found in [de Palma and Fosgerau \(2011\)](#) and [Small \(2015\)](#).

Road Pricing

In the context of the above framework, the social optimum is different from user (or unpriced) equilibrium. The bottleneck model provides an insight into the use of road pricing to manage congestion during the peak period. In the unpriced equilibrium, the waiting time is a pure dead-weight loss, and social optimum can be enforced by introducing a time-varying toll.³ Waiting times under the unpriced equilibrium can be eliminated by implementing the first-best pricing strategy. This involves the imposition of a time-varying toll that is equal to the value of waiting time. Therefore, the generalized costs remain unchanged because the optimal toll exactly replaces waiting time costs, with the collected toll revenue representing a net welfare gain. The starting and ending times of the peak period will remain the same as under the unpriced equilibrium since pricing does not affect the physical capacity of the bottleneck and commuters' work entry time.

While a time-dependent optimal toll is useful for analytical and theoretical configurations, it is difficult to implement in practice as it changes continuously over time. Second-best pricing and coarse toll are often used as an alternative for efficiency and

² We refer interested readers to consult [Arnott et al. \(1990\)](#) for the mathematical derivation of equilibrium quantities.

³ The bottleneck model has been used to analyse strategies for reducing traffic congestion that resulting in welfare loss under user equilibrium without government intervention. Some of these strategies include the introduction of tolls and parking prices at workplace.

practicality. In particular, second-best pricing has many variants that differ in terms of sources and constraints. It has less welfare gains than first-best pricing but provide implications for the attainment of multiple policy objectives relating to congestion management, cost-benefit analysis and market shares.

Valuation of Travel Time Variability and Travel Information

In the bottleneck model, equilibrium is obtained assuming that drivers have complete information about traffic situations. However, this is not often the case as travel times tend to be unpredictable from a commuter's point of view. When travel times are unpredictable, the arrival time of a particular trip cannot be known with certainty ahead of the trip. A behavioral response to account for the random variation in travel times includes leaving a safety margin at the beginning of the trip. In so doing, travelers trade the inconvenience of departing earlier and potentially being at work earlier than desired for a higher probability of arriving on-time. The higher the degree of randomness in travel times, the earlier the commuter needs to depart in order to maintain the same probability of arriving on-time.

The customary scheduling model for the valuation of travel time variability (Small, 1982) applies the same $\alpha - \beta - \gamma$ formulation of travel time and scheduling costs as that in the bottleneck model in order to analyze scheduling choices and ultimately derive a measure of the cost of travel time variability. An empirical estimate of the resulting measure of the cost of travel time variability has been used in traffic demand modeling and as input in economic appraisal of transport infrastructure and policy.

Heterogeneity in Values of Time and Preferred Arrival Time

Empirical evidence shows significant heterogeneity in trip-timing preferences across individuals. Indeed, the value of time and preferred arrival time could vary with flexibility of working hours, trip purposes, and similar other socio-economic factors. For commuting trips, empirical studies indicate that the differences in preferred arrival times come from the heterogeneity in work hours, trip-timing of daily activities, sleeping time preferences and so on. The seminal paper by Vickrey (1969) and early derivations of the bottleneck model capture the heterogeneity in preferred arrival times. For instance, Vickrey (1969) assumed a uniform distribution of preferred arrival times. Subsequent research has further generalized this for continuous and discrete nonidentical scheduling preferences. The discussion in the previous studies indicate that a certain fraction of travelers will only travel early or late for work based on the distribution of preferred arrival time and the fixed capacity.

In addition to the aforementioned factors, differences in the values of travel time and schedule delays are also attributable to heterogeneity in work hour flexibility. Elaborations on the heterogeneous values of time in the literature imply that the equilibrium arrival pattern will be sorted accordingly to their ratios $\frac{\beta}{\alpha}$ and $\frac{\gamma}{\alpha}$. For instance, suppose commuters have different unit cost of travel time but the same unit costs of schedule delay early and schedule delay late. In equilibrium, commuters arrive in different time intervals such that the group with higher unit cost of travel time (lower ratio) is traveling toward the end of the peak period.

User heterogeneity steps forward to more realistic generalizations of the bottleneck model and is important for determining and evaluating the welfare effects of toll strategies and capacity expansion. It also helps in understanding the choices of routes in a network where individual routes are characterized in terms of free flow travel time, tolls, congestion costs, and similar other features.

Networks

In fairly crowded areas, travelers might observe two or more bottlenecks in a network. In the case of configurations of upstream and downstream bottlenecks in series, travelers' departure times are socially non-optimal thus aggregate trip costs can be reduced by limiting the upstream effective capacity. This reveals the Braess paradox, in which expanding capacity to existing links in the network can result in increased aggregate travel costs. Previous studies also evaluated the use of entry-ramp metering in order to restrict the upstream flow and can thus improve efficiency. Previous research also considered parallel bottlenecks which can be different routes or different lanes of a single route. The parallel bottleneck models construct the opportunity to explore partially tolled express lanes.

The literature also analyzed a more general network with demand assignment and congestion costs among various origins and destinations. In one of the studies in this literature, the marginal values of travel time and schedule delay costs, i.e., α , β and γ are defined in relation to trip destinations and both travelers' route and departure time choices are determined. In this setting, system optimum can be achieved with time-dependent tolls by a space-time expanded network. This approach of dealing with queuing networks could provide many practical insights on traffic congestion over both space and time.

Other Application Areas

Agglomeration: In the basic bottleneck model, scheduling preferences are assumed to be exogenous and the value of time spent at the origin is assumed to be fixed irrespective of the time of the day. It is not clear how scheduling preferences arise and that the value of time spent at home may depend on when it is spent. Fosgerau and Small (2017) extend the basic bottleneck model and showed how scheduling preferences can arise as a result of agglomeration benefits at work and non-work activities performed at either ends of the trip.

METROPOLIS Model: A number of studies aim at establishing and developing simulation models to analyze a real city, such as the METROPOLIS. The METROPOLIS provides a dynamic environment to capture departure time and route choice behavior in a

large-scale network. Individuals are assumed to minimize their generalized travel cost function that depends on schedule delay costs, queuing costs and travel time cost. Route and departure time choices are thus endogenously derived at equilibrium. The system is capable of integrating additive elasticity, heterogeneity and variability in terms of capacity or demand as illustrated in the aforementioned sections, which provide tractable simulations on large-scale and real cases. The METROPOLIS system is established for both within-day and day-to-day dynamics of real network. It can provide a fully dynamic tool to evaluate strategies and pricing schemes for multiple policy objectives relating to demand management, such as smoothing out peak hour demand, increasing commuters' satisfaction and inferring commuters' travel patterns.

Parking: Parking is a growing topic due to the increasing urban population and traffic congestion. The basic bottleneck model has been augmented to analyze the choice of parking spots that differ from one another based on their location and the fee involved. In unpriced equilibrium, the first commuter passing through the bottleneck choose the closest parking spot to destination which is inefficient. In contrast, a location-dependent parking fee schedule can induce commuters to park in order of decreasing distance from destination, thereby reducing schedule delay costs. A variety of researchers also analyzed the parking setup with bottleneck model from different perspectives such as cruising, search, spot reservation and so on. The bottleneck model provides a straightforward foundation to analytically discuss and quantify the impacts of parking on urban mobility and congestion.

Land use: Land use incorporating bottleneck model is often formalized with suburb and downtown areas where travelers commute between the two areas. The situation describes commuting congestion considering the travel distance from residential locations to the bottleneck. At the equilibrium with homogeneity, commuters are sorted depending on their travel distance such that those who live closest to the bottleneck arrive to the destination first. The optimal coarse tolling benefits commuters who lives beyond a critical distance and induce a loss for nearby commuters. While with heterogeneity of values of time, commuters also sort temporally according to the ratio of unit cost of schedule delay early and late (i.e., $\frac{t}{\tau}$). Those with lower ratio value tend to arrive closer to their preferred arrival time to avoid queuing time rather than a schedule delay. Some researchers argue that when considering heterogeneous commuters, the socially optimal tolling expands the suburb populations and rent which hurts low-income communities residing in the suburb.

Autonomous vehicles: With the development of new technology and systems, the bottleneck model has new forms of settings and insights related to queuing formulation, value of time, system capacity and so on. For instance, carpooling and autonomous vehicles change user travel patterns and behaviors, resulting in different dynamic travel delay and congestion from the basic model. Recent research have used the bottleneck model for these purposes. Recently, [van den Berg and Verhoef \(2016\)](#) augmented the bottleneck model to explore the effect of autonomous vehicles on capacity, values of time and heterogeneity.

Conclusion

The bottleneck model is a fundamental tool for analyzing the phenomenon of traffic congestion and predicting the effect of transport-related policies on social welfare and traffic congestion. The model has attracted extensive attention in the last 50 years to become a standard tool for understanding traffic congestion and evaluating the effect of policies affecting it. The basic Vickrey bottleneck model has been extended in various directions and applied in numerous areas. The model continues to provide theoretical insights in economics and empirical applications. While issues discussed in the previous section can be explored in other frameworks, the bottleneck model has proven to have analytical properties to provide a formative and in-depth point of reference for researchers and practitioners in transportation analysis and modeling.

References

- Arnott, R., De Palma, A., Lindsey, R., et al., 1990. Economics of a bottleneck. *J. Urban Econ.* 27 (1), 111–130.
- de Palma, A., Fosgerau, M., 2011. Dynamic traffic modeling. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *A Handbook of Transport Economics*. Edward Elgar Publishing, Chapter 9, pp. 188–212.
- Fosgerau, M., Small, K., 2017. Endogenous scheduling preferences and congestion. *Int. Econ. Rev.* 58 (2), 585–615.
- Small, K.A., 1982. The scheduling of consumer activities: work trips. *Am. Econ. Rev.* 72 (3), 467–479.
- Small, K.A., 2015. The bottleneck model: an assessment and interpretation. *Econ. Transport.* 4 (1–2), 110–117.
- van den Berg, V.A., Verhoef, E.T., 2016. Autonomous cars and dynamic bottleneck congestion: The effects on capacity, value of time and preference heterogeneity. *Transport. Res. Part B: Methodol.* 94, 43–60.
- Vickrey, W.S., 1969. Congestion theory and transport investment. *Am. Econ. Rev.* 59 (2), 251–260.

Dynamic Congestion Pricing and User Heterogeneity

Kathrin Goldmann, Gernot Sieg, University of Münster, Institute of Transport Economics, Münster, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	150
Dynamic Congestion Pricing	150
User Heterogeneity	151
Determinants of Trip Costs	151
Real-Life Heterogeneity and Its Effects on the α , β , and γ Parameters	152
Income	152
Trip Purpose	152
Further Influences on the Value of Time	153
User Heterogeneity in the Bottleneck Model	153
Further Aspects	156
Conclusion	157
See Also	157
References	157

Introduction

The bottleneck model is a powerful tool for analyzing traffic congestion. In the chapter on the bottleneck model, it became evident that queueing in front of a bottleneck can be avoided with optimal pricing. Commuters change their departure times and, instead of wasting time stuck in a traffic jam, they pay a user charge. There is no waiting time wasted and all commuters are better off than before.

In reality, commuters differ from each other (are heterogeneous) and thus new aspects come into play. Heterogeneity means that users prefer different arrival times, have different incomes, and have different trip purposes. For this reason, several important parameters of the bottleneck model, like the desired arrival time, waiting, and schedule delay costs, are no longer equal for all commuters, but depend on individual characteristics and on trip purpose.

In this chapter, we will briefly explain dynamic congestion pricing and discuss the sources of heterogeneity among road users. If users with different characteristics and preferences are faced with congestion charges, some groups of road users are affected more than others. Understanding and mitigating these effects on different user groups is of considerable importance to increasing the political acceptability of congestion charges.

Dynamic Congestion Pricing

The first approach to analyzing the effects of traffic congestion was from [Arthur Pigou](#), who used the example of congested roads to explain external effects. Drivers assume that the travel conditions are independent of their own behavior and consequently, they only consider their average costs and not those they impose on other drivers. The no-toll equilibrium is at the intersection of the demand and the average cost functions. However, because there are also external costs, such as my decision to drive on the road increasing traffic and also slowing down other drivers, the no-toll equilibrium does not lead to the social optimum. Imposing an optimal user price p , which equals the marginal external costs in the optimum, reduces the demand to the socially optimal quantity. A graphic illustration can be found in [Santos and Verhoef \(2011\)](#). Ultimately, this charge increases the trip costs so that those drivers with the lowest willingness to pay will no longer travel and trips with a low priority will no longer be made. This Pigouvian charge efficiently reduces the traffic volume to the optimal quantity. As a result, however, before the reallocation of charges, all drivers are worse off than before, as the drivers who are priced off the road, as well as those who still use the road and pay the charge, lose utility from congestion pricing.

The Pigouvian congestion charge is based on a static model, which delivers valuable insights, but fails to depict traffic demand that varies over the course of the day. In order to capture this dynamic feature of traffic conditions, [Vickrey \(1969\)](#) introduced the so-called bottleneck model, and [Arnott et al. \(1990\)](#) developed an equilibrium model to investigate various toll regimes. This model is described in detail in the chapter on the bottleneck model. We will therefore only review those aspects of the model that are relevant to understanding the impact of user heterogeneity.

The model is based on a fixed number of homogenous commuters who drive to work during the morning rush hour. All commuters have to pass through a bottleneck with a fixed capacity s and want to start work at the time t^* ([Arnott et al., 1990](#)). Trip costs consist of α , β , and γ and, if introduced, a toll. While the parameter α displays the waiting costs, the parameters β and γ are the

schedule delay costs. β is the shadow value for being early and γ is the shadow value for being late (Arnott et al., 1993). Schedule delay costs mean that due to congestion, people do not drive at their preferred times and thus have to change their daily routines, such as getting up earlier in the morning to avoid congestion. Except for the first and the last commuters, who only incur schedule delay costs, all other commuters are faced with both schedule delay and waiting costs. At equilibrium, the trip costs for all drivers must be equal, because otherwise, individuals could change their departure time to reduce their costs.

The time-varying toll is proportional to the waiting costs and increases until t^* and decreases afterward. The optimal toll changes the departure times in such a way that there will be no queue in front of the bottleneck and consequently no waiting time. The toll transforms waiting costs into government revenues, which can be used to make a lump sum redistribution to commuters, to improve public transport, or to increase road capacity. Without any productive use of the revenues, the toll would be economically neutral. If money is invested, at least to some degree, in useful projects, everyone could be better off.

A quite new type of congestion models are the so-called bathtub models developed by Arnott (2013). Bathtub models consider the downtown urban center in an aggregated fashion. This seems to be a realistic view of urban congestion, because in heavily congested urban centers, it becomes difficult to identify single bottlenecks in the rush hour, as the whole city appears to be jammed. Both the bottleneck and the bathtub models are thus capable of handling hypercongestion. Moreover, bathtub models combine the fundamental relationship between speed, flow, and density from transportation science with important features of dynamic economic models, for example, bottleneck model.¹ This means, for instance, that total trip costs consist of travel time costs and schedule delay costs, as in the bottleneck model. In contrast to the latter, where the discharge rate of the bottleneck is not affected by the length of the queue in front of the bottleneck, in the bathtub model, however, the discharge rate of the congested downtown urban center is reduced if traffic density exceeds a critical level. This is in line with the fundamental relationship.

As in the bottleneck model, in the bathtub model, the optimal time-varying congestion charge should ease congestion. In the bottleneck model, the commuters' departure times are changed in such a way that the queue in front of the bottleneck is eliminated. In the bathtub model, the time-varying toll should keep traffic density below a critical value to avoid hypercongestion. This is especially important, as the (out)flow rate should be kept high to maintain an efficient road network performance.

The efficiency gain of tolling depends on the extent of congestion in the urban center. When congestion is low, the efficiency gain may be lower than the revenues; when congestion is strong, the efficiency gain may be substantially higher than the revenues (Arnott, 2013). If the latter is the case, congestion tolling may be beneficial even, if toll revenues are invested entirely in pointless projects. This insight for heavily congested downtown urban centers might increase the political acceptance of congestion tolling. Accordingly, this latter point also reveals an important difference to the bottleneck model, where the efficiency gain from tolling exactly equals the toll revenues.

Economic models in general, when dealing with a topic that has not been analyzed before, usually start off with a series of assumptions that make the models tractable. Models for congestion pricing initially depart from inelastic deterministic demand, no outside option like public transport, deterministic capacity, as well as homogeneous users (Arnott, 1990). While the bottleneck model has been used for decades as a tool for analyzing bottleneck congestion, it has already been extended in several respects, and various assumptions have been relaxed. Bottlenecks have, for instance, been analyzed with elastic demand (Arnott et al., 1993; Yang and Huang, 1997), stochastic capacity (Xiao et al., 2015), as well as heterogeneous commuters (Arnott et al., 1994; Takayama and Kuwahara, 2017). In the bathtub model, most of the extensions are still open for future research (Arnott, 2013; Fosgerau, 2015). Later on, we will illustrate the effects of heterogeneous users in the bottleneck model with a numerical example.

In the next section, we take a close look at the characteristics and preferences that make users differ from each other. We furthermore demonstrate how this heterogeneity is dealt with in economic models.

User Heterogeneity

Determinants of Trip Costs

In this section, we take a close look at user heterogeneity. Drivers may be shift workers, employees in IT start-ups, parents picking up their child from a table tennis lesson, or taxi drivers. We can see from these examples that there are various reasons to make car trips during the rush hour. All these users have different preferences, and thus different values of travel time and different costs of schedule delay. It is, for example, not unlikely that if we asked the drivers of 300 cars passing a highway traffic detector for their willingness-to-pay for the trip, we would observe 300 different travel time values. Although there are some important factors that impact on the value of time, it remains difficult to isolate all origins of heterogeneity (Brownstone and Small, 2005).

The value of time depends on the opportunity costs of time and the direct utility of travel time. When stuck in a traffic jam (waiting costs) or starting a trip to work earlier or later to avoid the traffic jam (schedule delay costs), you could have been doing something else during this time. If there had been no congestion, you could have been at work 1 h earlier, and the opportunity costs are your hourly wage. Alternatively, it would have been possible to start the trip to work 1 h later and use the time to sleep longer in the morning. The value of this 1 h more sleep in the morning, although this is your free time, can be substantial, as well. Consequently, it depends on the value that people assign to what else they could have been doing during that time. Although values of time are very individual, there are some factors that at least on average influence time values systematically. These are, for example, income and trip purpose (Santos and Verhoef, 2011).

¹ For more information on the fundamental relationship, refer to the Chapter "Speed-Flow-Density and Dynamics of Congestion".

User heterogeneity leads not only to different values of travel time but also to different preferences regarding the arrival time t^* . There are people working in hospitals who begin their early shift at 6.00 a.m. in the morning, and there are those working in a shop in the downtown urban center that opens at 10.00 a.m. in the morning. As we can easily see, different preferred arrival times can at least to some extent smooth the demand for road capacity and can thus also moderate traffic congestion.

Real-Life Heterogeneity and Its Effects on the α , β , and γ Parameters

Income

The value that people assign to their time turned out to depend heavily on income. That is, people with higher income usually have higher values of time (Börjesson et al., 2012; O'Flaherty, 2005). This is readily understandable when we consider that one additional hour available could have been spent at work. In the short run, however, most employees have fixed working hours and they cannot decide freely how many hours they work each day. In contrast, at least in the medium- and long-run, hours spent at work and hours of free-time are substitutes. A person who presently has long working days can change his/her job and start a new one where he/she can work 1 h less per day than before. The payment forgone due to this hour more of free-time per day is the opportunity cost. Consequently, the value of your free-time is also closely tied to your income. The higher the income, the higher the value assigned to time in general.

A thought experiment can illustrate this: If we consider two people, the first person with an annual income of 30,000€ per year and the second with an annual income of 200,000€ per year. Person 1 is paid 16€/h and person 2 receives 109€/h. If both could buy 1 h more free-time per day for 15€, person 1 will probably not accept this offer as he/she loses about the amount he/she earns in 1 h, while person 2 only has to pay about one-eighth of his/her hourly wage. When the optimal toll in the bottleneck model that avoids queuing is 15€/h, and the time lost due to congestion per day is 1 h, we can immediately see the effects of congestion pricing. Person 1 would probably prefer to pay with waiting time instead of 15€, while person 2 would probably prefer to pay 15€ to cut his/her commuting time and gain one more hour of free-time. As congestion charges are usually not tied to income, we can see that people with different incomes are affected very differently by congestion charges.

The preferences α (waiting time costs), β (time costs of being early), and γ (time costs of being late) in economic congestion models depend on the value people assign to time. As illustrated above, the value people assign to time depends on income. For this reason, we can conclude that all time cost parameters (α , β , γ) depend positively on income. If we assume that all other factors (trip purpose, sex, age, . . .) are equal, high-income individuals will have a higher value of time, compared to those with low income.

Trip Purpose

Trip purposes can be split into private and business. Private trips may entail commuting to the workplace or school, or driving to the supermarket, the zoo, or to the guitar lesson. As we can see from these examples, especially the values for schedule delay costs of being late γ differ considerably between the different trips.

The highest schedule delay costs might be those for work trips with a fixed start time followed by the guitar lesson. When you are 30 min late to a 45-min guitar lesson, the whole trip might become pointless. Workers with flex-time might be more relaxed, provided the delay is not too long. For free-time activities without a predetermined starting time t^* , in this case driving to the supermarket or to the zoo, we probably have low schedule delay costs of being late.

Being early imposes lower costs on average. The order regarding the costs of the examples for trips mentioned above probably remains the same.

Another type of trips are those during working time. Business trips are related to the job and can, for example, mean driving to the customer or to the airport to catch a plane to fly to a transport economics conference. Being late at a customer is on average bad, but being late at the airport and running the risk of missing the plane, is surely worse. The costs of being early might be the reverse. If you are a frequent flyer and can spend time being early in the airport lounge, this might not be so bad, while waiting in the car in front of the customers company might be rather unproductive time. Yet, these costs might also be perceived differently by different individuals.

Taxes and social insurance contributions split up the gross wages paid by an employer and net wages received by an employee. Waiting costs for trips with a business purpose usually comprise gross wages and are therefore much larger than waiting costs for private trips (e.g., commuting to the workplace), even if the affected person in the car is the same. People using private cars have to pay tolls from their net wages. Company car drivers usually do not pay tolls themselves, but obtain a refund from the company. As long as the company only partially refunds tolls, drivers of company cars will probably not consider changing their behavior to avoid tolls.

In contrast to the situations described above, driving may also be the purpose itself. Some people simply like cruising around or making a round trip on their motorbike. As there is then no destination to reach at a specific time t^* , there are no schedule delay costs. Regarding the waiting costs α , there are reasons for their being both low or high. One can, on the one hand, argue that such persons usually enjoy their time inside the car or on a motorbike so that waiting costs might be relatively low. On the other hand, being stuck in a traffic jam might also be annoying because driving is a lot less fun in such a situation.

As we can see from the above examples, trip purposes mainly differ from each other regarding their costs of schedule delay, β and γ . Waiting costs α do not depend so much on the trip itself, but rather on personal characteristics. This is basically the income, but can

Table 1 Summary of effects

Parameter	Trip purpose	Waiting/SDCs	Impact
High income	All	α	High
	All	β and γ	High
Trip purpose	Trip to airport to catch plane	β	Medium
		γ	Extremely high
	Visit the zoo	β	Low
		γ	Low
	Driving itself	β	None
		γ	None

also, for example, be the person's general patience or impatience. Possible influences on the α , β , and γ parameters are summarized in **Table 1**.

Further Influences on the Value of Time

Further systematic influences might also be age and sex. However, when going into detail, it becomes obvious that much of the influence can again be attributed to income and trip purpose. While younger people might have more obligations, like going to school or to work, people who are retired, do not have generally such a packed daily schedule and might thus have on average lower schedule delay costs. But elderly people can also have trip purposes with high schedule delay costs. The on average lower income compared to employed people might also influence waiting costs negatively.

A user of an autonomous car has reduced travel time and waiting costs (α), as the time inside the car is not used for driving the car and can be used productively or for entertainment. This can worsen congestion in the rush hour, as waiting is not perceived so negatively by commuters. Now, we can see that the various aspects discussed above can be combined in every conceivable way. People considered in this section could use carpooling and therefore, their waiting and schedule delay costs might be about twice as high as for single drivers. Or a 35-year-old woman with a high income, living in a suburb, may be driving to the zoo with her child. In this example, we observe a person who generally has a high value of time, but in this case, a trip purpose with low schedule delay costs. For this reason, [Van den Berg and Verhoef \(2011\)](#) argue that gains and losses from tolling are not monotonic in the value of time because there are also very different values of schedule delay costs.

User Heterogeneity in the Bottleneck Model

In this section, we take a close look at heterogeneous users in the bottleneck model, based on the paper from [Arnott et al. \(1994\)](#). We describe the effects of dynamic congestion pricing on heterogeneous users by means of a numerical example.

When we introduce heterogeneous users in congestion models, we have to consider several α , β , γ parameters. The preferred arrival time t^* is equal for all commuters at this point ([Arnott et al., 1994](#)). For illustration purposes, we distinguish between three groups of users: Group L with a low income, group M with a medium income, and group H with a high income. For our numerical example, we assume that $N = 6000$ commuters want to drive through a bottleneck with a capacity of $s = 2000$ cars/h. To pass through the bottleneck, the commuters thus need 3 h. Each group (L, M, H) consists of 2000 commuters, which means that each group requires 1 h to pass through the bottleneck. The preferred arrival time is $t^* = 8:00$ a.m. Furthermore, all model assumptions made by [Arnott et al. \(1994\)](#) are valid for this example. The parameters we assume for our numerical example are summarized in **Table 2**.

To determine when each group prefers to travel, we need to know the relative disutility of schedule delay relative to travel time (β/α) for each group. The group which does not mind waiting in the traffic jam at the center of the peak period, but seriously dislikes schedule delay will travel in the middle of the peak period. The group in our example that dislikes schedule delay most, relative to waiting, is group L ($\beta/\alpha = 0.80$). A common example is the assembly line worker, who starts his/her shift at a certain fixed time

Table 2 Parameters for three groups of commuters

Parameter	Group L	Group M	Group H
α	0.10	0.30	0.70
β	0.08	0.20	0.40
γ	0.16	0.40	0.80
β/α	0.80	0.67	0.57

Cost parameters are defined in €/min.

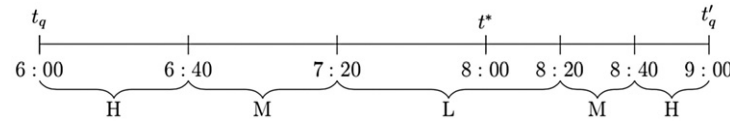


Figure 1 Departure times in the no-toll equilibrium.

(Arnott et al., 1994). If he/she is at work early, there is nothing to do until the shift starts and schedule delay costs are high, and if he/she is at work late, schedule delay costs are very high indeed, because production might come to a standstill. Of course, there are also employees with fixed working times and a high income, like doctors working in hospitals. However, it seems that there are more people with a high income who have flexible working times compared to people with a low income and flexible working times (Lott, 2016, p. 9). Therefore, it can also be assumed that on average the relative costs of schedule delay to waiting costs decrease with income ($0.80 \geq 0.67 \geq 0.57$) as displayed in the example in Table 2. As shown in Fig. 1, drivers of group L will drive in the middle of the rush hour and those of group M at adjacent times, and drivers of group H, with the lowest relative costs of schedule delay, will travel at the fringes.

We can also calculate the costs for each group in the no-toll equilibrium. The first driver arriving at the bottleneck at 6:00 a.m. incurs only schedule delay costs and no waiting costs. Schedule delay costs of group H are $\beta_H = 0.40\text{€/min}$. Because they arrive 120 min (8:00–6:00) too early, total costs are 48€. Since, in equilibrium, each point in time for the homogeneous group H must be equally attractive—otherwise people could still change their departure time to reduce their costs—the costs of all drivers of this group must be equal. The last driver of group H who is early at 6:40 a.m. also has 48€ of total cost of which 32€ are schedule delay costs and 16€ are waiting costs.² The same applies to drivers of group H who are late. The last driver is 60 min late and has schedule delay costs of 0.80€/min ($60 \text{ min } 0.80\text{€/min} = 48\text{€}$).

The first driver of group M arrives at the bottleneck at 6:40 a.m. He/she is 80 min early and has schedule delay costs of being early of 0.20€/min , which leads to schedule delay costs of 16€. As calculated in Footnote 2, waiting time in the queue at that time is approximately 23 min. 23 min multiplied by the waiting costs of drivers of group M (0.30€/min) causes waiting costs of $\approx 7\text{€}$. Thus, total costs of group M are 23€.³

As the drivers in group L have the highest relative cost of schedule delay, they thus arrive at the center of the rush hour. The first person to arrive at the bottleneck is 40 min early and has schedule delay costs of 8 cents/min and thus total costs of schedule delay of 3.20€. The waiting time at that arrival time is with 50 min⁴ relatively high and the waiting costs are: $50 \text{ min } 0.10\text{€/min} \approx 5\text{€}$. Finally, total costs of group L are 8.20€. Fig. 2 summarizes the commuting costs of the no-toll equilibrium.⁵

We can see in the no-toll equilibrium that schedule delay costs for group H are relatively high because they have the highest costs of schedule delay per minute and travel at the least attractive times.

We now assume that a toll is introduced to avoid waiting times in front of the bottleneck. In contrast to the no-toll equilibrium, the absolute schedule delay costs now matter. This is because people pay for using the road at a specific point in time, and will only travel at that time, when their schedule delay costs of traveling at another time are higher. In other words, drivers have to pay money to reduce schedule delay. As the toll is equal for all drivers, but schedule delay costs depend on income, high-income drivers will be more willing to pay the toll to avoid/reduce schedule delay.

When we now order our three groups of drivers in terms of their absolute schedule delay costs, they will change their positions (Fig. 3). Group H has the highest absolute schedule delay costs ($\beta = 0.40\text{€/min}$) and will travel in the middle of the rush hour. The adjacent times will be used by group M ($\beta = 0.20\text{€/min}$) and group L will travel at the fringes ($\beta = 0.08\text{€/min}$). In this social optimum, total schedule delay costs are minimized because the drivers with the lowest schedule delay costs per minute have the longest schedule delay and vice versa.

The example shows the main features of income heterogeneity in the bottleneck model. Whereas in the no-toll equilibrium the most attractive arrival times are used by low-income people paying with their waiting time, the toll shifts the low-income people to the fringes of the rush hour and high-income people travel at the most attractive times.

In the social optimum, group L travels at the fringes, and the total costs incurred by group L can again be calculated with the first and the last drivers. While the first driver is 120 min early and has schedule delay costs of $\beta = 0.08\text{€/min}$ he/she has costs of 9.60€ (last driver: $60 \text{ min} \cdot 0.16\text{€/min} = 9.60\text{€}$). Since the total cost incurred by group L is 9.60€, the last driver of this group being early at 6:40 a.m. has schedule delay costs of $80 \text{ min} \cdot 0.08\text{€/min} = 6.40\text{€}$. The congestion charge he/she pays is: $9.60\text{€} - 6.40\text{€} = 3.20\text{€}$. The same applies to the first driver of group L who arrives late at 8:40 a.m.

The congestion charge at 6:40 a.m. is 3.20€. The first driver of group M to arrive at 6:40 a.m. thus pays this charge as well and has schedule delay costs of $80 \text{ min} \cdot 0.20\text{€/min} = 16\text{€}$. As all drivers in this group are equal, they all incur total costs of $3.20\text{€} + 16\text{€} = 19.20\text{€}$, although the composition of costs may differ. The closer drivers of group M arrive to t^* , the lower the schedule delay costs become

² The 16€ waiting costs can be derived the following way. The last driver of group H before 8:00 a.m. is 80 min early (8:00–6:40). This amounts to schedule delay costs of $80 \text{ min} \cdot 0.4\text{€/min} = 32\text{€}$. The difference is $48\text{€} - 32\text{€} = 16\text{€}$ waiting costs. Consequently, the waiting time at 6:40 a.m. is $16\text{€}/0.7\text{€/min} \approx 23 \text{ min}$.

³ The same applies to all the members of group M who are late and where the last driver arrives at 8:40 a.m. Calculation: $40 \text{ min} \cdot 0.40\text{€/min} + 23 \text{ min} \cdot 0.30\text{€/min} = 23\text{€}$.

⁴ The waiting time can again be calculated analogous to Footnote 2 with the costs of the last early driver (or first of the group of late drivers) of group M.

⁵ In Figs. 2 and 4, schedule delay costs are labeled SDC and waiting/travel time costs are labeled WC.

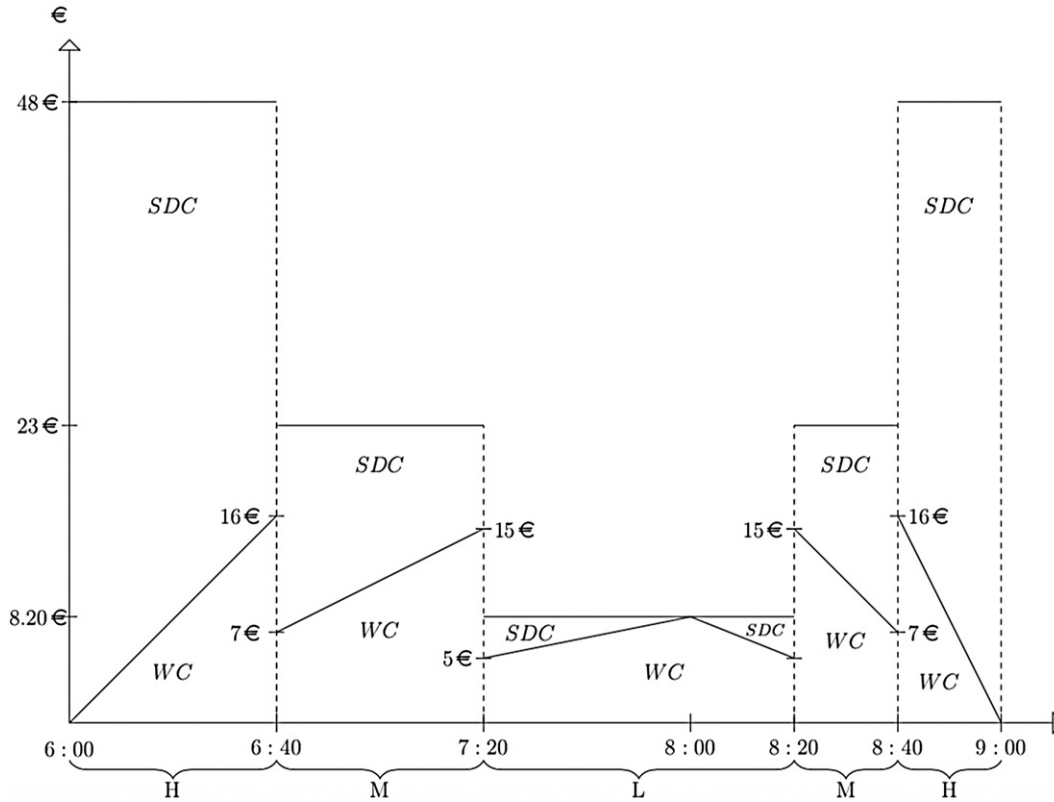


Figure 2 Departure times and costs in the no-toll equilibrium.

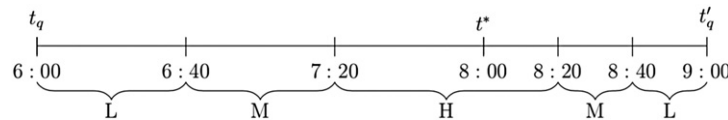


Figure 3 Departure times in the social optimum.

and the higher the toll becomes. The last driver of group M, being early at 7:20 a.m., has schedule delay costs of 8€ (40 min 0.20€/min) and pays a toll of 11.20€ (19.20€–8€). Again, the first driver of group M, being late, incurs the same composition of costs.

As the first driver of Group H who arrives at the bottleneck at 7:20 a.m. pays a toll of 11.20€, and has schedule delay costs of 40 min 0.40€/min = 16€, all drivers of group H have total costs of 27.20€. The driver of group H who arrives at 8:00 a.m. only pays the toll and has no schedule delay costs, as he/she arrives at his/her preferred time. Fig. 4 summarizes the effects.

The high-income group travels at the most attractive times and pays the highest congestion charges. Compared to Fig. 2, the areas of schedule delay costs are smaller now because the high-income group with high schedule delay costs per minute now travels close to 8:00 a.m.

Before redistribution of the toll revenues, commuting costs change as displayed in Fig. 5: Whereas the high-income group gains 48€ – 27.20€ = 20.80€ per commute, the low-income group loses 9.60€ – 8.20€ = 1.40€. A congestion charge that is personalized and, for example, tied to income could remedy this problem, but connecting travel data with income data of course raises new problems regarding data protection and privacy.

In our numerical example, total user costs would shrink from $2000 \cdot 8.20€ + 2000 \cdot 23€ + 2000 \cdot 48€ = 158,400€$ in the no-toll equilibrium to 112,000€ ($2000 \cdot 9.60 + 2000 \cdot 19.20 + 2000 \cdot 27.20$) in the optimal toll equilibrium. However, the whole picture of welfare effects is incomplete without knowing how the toll revenues are spent and including the costs for collecting the toll. The following utilizations are usually addressed: lump sum redistribution, tax cuts, improve or increase subsidies for public transport. A lump sum redistribution of the toll revenues of 56,000€ reduces user costs to 56,000€ (because the toll revenues are half of total costs), about a third of the no-toll equilibrium.

Low-income groups, fearing that redistribution will not compensate for losses, will oppose the proposed toll regime. A popular political remedy is using the toll revenues to subsidize public transport. Former car drivers shifting to public transport reduce the length of the rush hour and therefore waiting as well as schedule delay costs. We can now consider a low-income person with

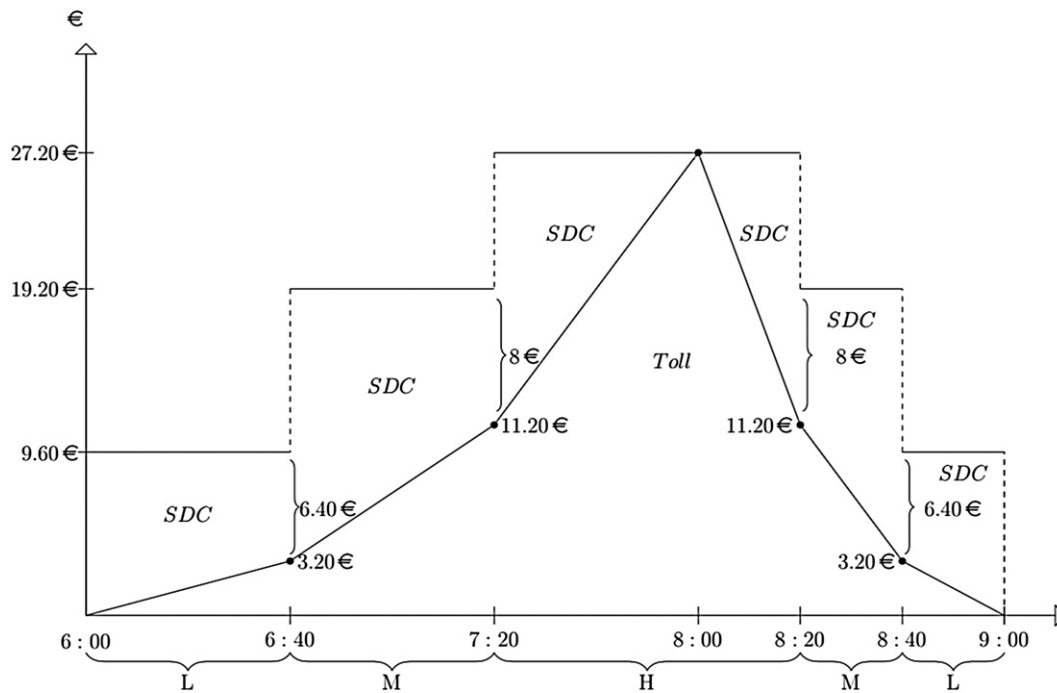


Figure 4 Departure times and costs in the optimal toll equilibrium.

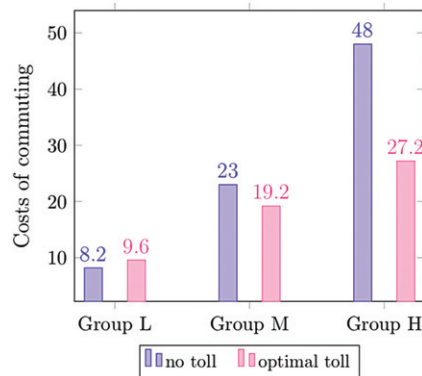


Figure 5 Commuting costs.

commuting costs of 8.20€ at the status quo no-toll equilibrium and with no opportunity to shift to public transport. In order to have no more costs than 8.20€ in the toll equilibrium, the toll must not start before 6:17:30 a.m., because at this time, schedule delay costs are $102.5 \text{ min} \times 0.08\text{€} = 8.20\text{€}$. All car drivers passing the bottleneck before 6:17:30 a.m. have to use public transport to secure nonrising commuting costs for low-income people. Comparing 17.5 new low-income minutes with 40 min shows that $17.5/40 = 43.75\%$ of the low-income group has to shift their transport mode, if public transport subsidies lead only low-income people to use public transport. Low-income group members who fear that the subsidies are not large enough to induce such a modal shift, or assume that public transport capacities are too low for the new travelers, will oppose a toll, even if the proposal earmarks the toll revenues for public transport.

For a more detailed perspective on the distributional effects of congestion charges, also see chapter “Distribution effects of congestion charges and fuel taxes.”

Further Aspects

As already pointed out in the discussion on heterogeneity, users of autonomous cars have reduced travel time values, as the time in the car can be used productively or for entertainment. [Van den Berg and Verhoef \(2016\)](#) have extended the bottleneck model for

autonomous cars and find two opposing effects. On the one hand, due to reduced travel time/waiting costs, users tend to drive closer to t^* and thereby worsen congestion (heterogeneity effect).⁶ On the other hand, autonomous cars require smaller headways and thus road capacity increases (capacity effect). This second effect eases congestion. As long as the market penetration of autonomous cars is low, the heterogeneity effect is likely to dominate and congestion in the rush hour will increase.

The bottleneck model has been extended by Yu et al. (2019) to another aspect of heterogeneity, namely carpooling. Commuters of the groups considered above with a similar origin and destination could share one car (within their group). This results, for example, in two people traveling in one car. The impact on the departure times in the bottleneck model is straightforward. It leads to carpool cars with twice the schedule delay costs, compared to a single driver of the same group. Analogously to our example, in the social optimum, commuters with the highest values of schedule delay costs will drive closest to the preferred arrival time. If income differences are large ($\beta_H > 2\beta_L$), the early arrival order will be: group L solo drivers, group L carpoolers, group H solo drivers, and group H carpoolers. If income differentials are small, ($\beta_H < 2\beta_L$), the arrival order before t^* will be: group L solo, group H solo, carpoolers of group L, carpoolers of group H. Because carpooling generally eases congestion, but a large share of benefits also accrues to solo drivers, Yu et al. (2019) propose a subsidy for carpooling.

In the long-run, preferences for living style (rural or urban) have an effect on the decision as to where to locate. The location matters when introducing congestion pricing, as people living in the central business district, in the midtown, or in the suburbs are affected differently by the toll (de Borger and Russo, 2018). Congestion pricing also affects landowners and renters differently (de Borger and Russo, 2018). How these groups are affected by congestion pricing also depends on the system introduced (e.g. cordon toll, area toll). Moreover, the distance from home to the next bus stop or railway station and the public transport's level of service is a further source of heterogeneity, which is reflected in rents or land prices, and may lead to self-selection.

Conclusion

Drivers on a congested road differ in many respects, for example trip purpose and income. Besides personal characteristics, people also differ regarding their preferences for solo-driving, carpooling, or using autonomous cars. All these aspects impact on people's schedule delay and waiting costs. Whereas for a no-toll road, the relative values of schedule delay costs and waiting costs determine at which time a driver arrives, in the optimal toll equilibrium, absolute values of schedule delay determine the arrival times. If income is the only source of heterogeneity, the low-income people use the most attractive arrival times if there are no tolls. The optimal toll shifts the low-income people to the fringes of the rush hour, and the high-income people drive at the most attractive times. Furthermore, people with a trip purpose with high schedule delay costs will also use the most attractive times. Before redistribution of the toll revenues, low-income drivers may prefer a no-toll equilibrium, because they value the gains of no more waiting time less than the disadvantage of unsuitable arrival times and toll payments. However, the toll is a potential Pareto improvement, meaning that if toll revenues are spent wisely, all commuters gain from the optimal toll regime.

See Also

Distribution Effects of Congestion Charges and Fuel Taxes;
How will Autonomous Vehicles Impact Car Ownership and Travel Behavior; The Bottleneck Model

References

- Arnott, R., 1990. Signalized intersection queuing theory and central business district auto congestion. *Econ. Lett.* 33, 197–201.
- Arnott, R., 2013. A bathtub model of downtown traffic congestion. *J. Urban Econ.* 76 (1), 110–121.
- Arnott, R., de Palma, A., Lindsey, R., 1990. Economics of a bottleneck. *J. Urban Econ.* 27, 111–130.
- Arnott, R., de Palma, A., Lindsey, R., 1993. A structural model of peak-period congestion: a traffic bottleneck with elastic demand. *Am. Econ. Rev.* 83 (1), 161–179.
- Arnott, R., de Palma, A., Lindsey, R., 1994. The welfare effects of congestion tolls with heterogeneous commuters. *J. Transport Econ. Policy* 28 (2), 139–161.
- Börjesson, M., Fosgerau, M., Algers, S., 2012. On the income elasticity of the value of travel time. *Transport. Res. Part A: Policy Practice* 46 (2), 368–377.
- Brownstone, D., Small, K.A., 2005. Valuing time and reliability: assessing the evidence from road pricing demonstrations. *Transport. Res. Part A: Policy Practice* 39, 279–293.
- de Borger, B., Russo, A., 2018. The political economy of cordon tolls. *J. Urban Econ.* 105, 133–148.
- Fosgerau, M., 2015. Congestion in the bathtub. *Econ. Transport.* 4 (4), 241–255.
- Lott, Y., 2016. Flexible Arbeitszeiten: Eine Gerechtigkeitsfrage? Hans Böckler Stiftung - Forschungsförderung Report.
- O'Flaherty, B., 2005. *City Economics*. Harvard University Press, Cambridge.
- Pigou, A., 1920. *The Economics of Welfare*. London.
- Santos, G., Verhoef, E., 2011. Road congestion pricing. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *A Handbook of Transport Economics*, Edward Elgar, Cheltenham, pp. 561–585.
- Takayama, Y., Kuwahara, M., 2017. Bottleneck congestion and residential location of heterogeneous commuters. *J. Urban Econ.* 100, 65–79.
- Van den Berg, V.A., Verhoef, E.T., 2011. Winning or losing from dynamic bottleneck congestion pricing? The distributional effects of road pricing with heterogeneity in values of time and schedule delay. *J. Public Econ.* 95 (7–8), 983–992.
- Van den Berg, V.A., Verhoef, E.T., 2016. Autonomous cars and dynamic bottleneck congestion: the effects on capacity, value of time and preference. *Transport. Res. Part B* 94, 43–60.

⁶ For a more detailed discussion on the effects of autonomous cars on travel behavior, see chapter entitled "How Will Autonomous Vehicles Impact Car Ownership and Travel Behaviour."

Vickrey, W.S., 1969. Congestion theory and transport investment. *Am. Econ. Rev.* 59, 251–261.

Xiao, L.-L., Huang, H.-J., Liu, R., 2015. Congestion behavior and tolls in a bottleneck model with stochastic capacity. *Transport. Sci.* 49 (1), 46–65.

Yang, H., Huang, H., 1997. Analysis of the time-varying pricing of a bottleneck with elastic demand using optimal control theory. *Transport. Res. Part B: Methodol.* 31 (6), 425–440.

Yu, X., van den Berg, V.A., Verhoef, E.T., 2019. Carpooling with heterogeneous users in the bottleneck model. *Transport. Res. Part B: Methodol.* 127, 178–200.

Economics of Parking

Daniel Albalade*, **Albert Gragera[†]**, *University of Barcelona (GiM-IREA), Barcelona, Spain; [†]Technical University of Denmark, Copenhagen, Denmark

© 2021 Elsevier Ltd. All rights reserved.

Introduction	159
The Economic Properties of Parking	159
Parking Demand and the Generalized Cost of Transportation	160
Cruising for Parking	160
Market Power, Spatial Competition, and Information Frictions	161
The Interplay Between Parking and Other Markets	162
Parking Regulation and Its Political Economy	162
Parking and Long-Term Decisions	163
Technology and Parking Innovations	163
References	164

Introduction

The Economics of Parking is the branch of transport economics that deals with parking, that is, the bringing to a halt of a vehicle—generally a motor vehicle—and leaving it temporarily unoccupied in a dedicated space of the urban area, which might be either at the curbside or in a facility/building off the street (either public or private garages). Unlike most transportation research and policy analyses that concern themselves with situations in which vehicles are in motion, the Economics of Parking considers the most usual state of cars, given that they are parked for about 95% of the time. As such, the issues typically addressed in the Economics of Parking are the efficient allocation of urban space to parking; parking regulation, planning, and pricing; the indirect distortions that parking produces in the functioning of other markets; and the derived externalities that may affect social welfare. Recent social and academic interest in smart and sustainable mobility in cities has boosted economic research in parking in a variety of areas.

Free or low-cost parking fuels excessive demand and creates negative externalities, and supply-oriented policies result in the massive consumption of urban space dedicated to parking, reducing parking costs and promoting the use of private vehicles by favoring their relative attractiveness. If parking is underpriced, then its costs will inevitably be hidden in the price of everything else in the city (Shoup, 2005). The social costs of underpriced parking include greater congestion and pollution, a higher number of accidents, and a greater amount of time wasted, primarily as a result of inefficient parking policies and regulations. Current concerns about sustainable mobility mean that parking is increasingly being understood as a tool for mobility while stressing the importance of accounting and addressing its market distortions. This understanding has fuelled initiatives to reorient planning toward a parking demand management approach, where parking is planned in order to serve wider urban and transport policy goals. One such goal is typically to constrain the use of cars. Here, supply and demand both need to be carefully managed to achieve efficiency and, as such, parking policy becomes a key tool for managing mobility and travel demand.

Given the close links between parking and car use, economists and planners see in parking a channel through which they can understand and tackle the challenges of sustainable mobility, improve the quality of life in cities, and promote local economic efficiency. Moreover, the current widespread adoption of parking regulations means parking is seen as a readily available, and more feasible, alternative to road pricing. In short, the Economics of Parking has emerged as a research area that is attracting the growing attention of transportation economists as well as those working in other disciplines, including urban planning and engineering.

The Economic Properties of Parking

Parking is typically classified by economists as a private intermediate good. It can be considered a private good because it possesses the properties of rivalry in consumption and excludability, where rivalry implies competition for the available, but limited, parking spaces and excludability implies the use of pricing and regulations that condition and influence the free choice of road users and parking suppliers. Despite being a private good, public sector intervention is common everywhere, being most intense in urban areas. Such intervention is usually justified by the presence of various market failures, with externalities and imperfect competition being the most frequently discussed in the literature. Interventions usually take the form of (1) public supply of parking spaces both at the curbside and off-street; and (2) regulations affecting the use of parking spaces (who is allowed to park, maximum length of stay, parking fee, etc.) or the available private supply (minimum requirements, garage concessions, quality standards, etc.).

Moreover, parking can be considered an intermediate good because parking their vehicle does not constitute the drivers' ultimate goal. Rather, they park to satisfy other needs and demands—work, shopping, visiting service facilities, leisure activities, going home, etc.—that require mobility from a point of origin to their destination and which are facilitated by the possibility of leaving their car

close to the latter. Thus, any distortion in the parking market will not only have an impact on the transportation sector and its associated externalities but it will also affect the price of almost everything else, including housing, leisure, retail, and firm location among other goods and services, as has been demonstrated and as is described later. Most economic transactions involve transportation, and parking is a compulsory component of that in the case of motor vehicles.

Parking Demand and the Generalized Cost of Transportation

Research into parking demand is extensive but has focused primarily on the impact of curbside parking regulations on commuters' travel choices based on their stated and revealed preferences. Demand is found to vary across time, space, and demand segments, and to be dependent on the relative generalized cost of transportation of motorized mobility, on the one hand, and all other possible mobility alternatives, on the other. Parking is introduced in the generalized cost of transportation by affecting the in-vehicle and out-of-vehicle travel time and its monetary components. Thus, the basic cost function can be generalized to account for parking components in the following manner:

$$\begin{aligned} C_g &= v(t + s) + vw + M \\ M &= \mu + ft + p \end{aligned}$$

where C_g is the full-generalized cost of a private motorized trip, which is a function of three components. First, the in-transit time costs component $v(t + s)$, which depends on the duration of the trip (t) from origin to destination and on the time spent searching for a free parking spot (s). Both are multiplied by the value of time (v). The second component is the time cost out of the vehicle, that is, between the point of origin and the parking space and from that space to the final destination. This component depends on the duration of the walking trip (w) multiplied also by the value of time (v). Finally, the third component is the monetary (out-of-pocket) costs of the journey (M), which includes tolls (μ), consumption of fuel (f), and payments for parking (p). As such, parking is expected to affect all three components of the generalized cost of transportation: it increases the in-vehicle travel costs if the driver needs to search for a free space at destination, which will depend on supply–demand interaction for parking; it increases the out-of-vehicle time costs if the available parking spaces are located some distance from the origin and/or destination, requiring the driver to walk; and it increases the monetary costs of the private journey if parking is not free and a fee has to be paid to park the vehicle. This amount depends on the hourly rate and the length of stay.

Empirical evidence confirms that demand for parking is negatively related to fees and its sensitivity depends on user and trip characteristics. Thus, it tends to decrease with income and to increase with the length of stay. It also depends on the purpose of the trip and increases with the availability of alternative modes of transport. Moreover, it is found to depend on the availability of parking at the point of destination and at home, with free workplace spaces acting as a major incentive to drive to work. The literature also shows negative cross-elasticities between garage and curbside parking demand. See [Marsden \(2006\)](#) and [Lehner and Peer \(2019\)](#) for a complete review of the mentioned issues.

Cruising for Parking

The Economics of Parking has paid special attention to analyzing the search externality, that is, cruising for parking. Drivers occupying a parking space impose a search cost on other drivers if the demand for parking is higher than the capacity constraint, which implies a pure deadweight loss as shown in [Fig. 1](#) (case 1). Furthermore, this market distortion is expected to aggravate other negative externalities, such as congestion, fuel consumption, and pollution, due to the longer duration of in-vehicle trips searching for a free space. This can be corrected by fixing a parking fee such that cruising is eliminated, and parking demand is equated to the parking supply capacity constraint (case 2). In this case, the cruising welfare loss is fully converted into parking revenues. If parking pricing is introduced but kept below the efficient price (underpriced), only part of the welfare loss is converted into parking revenue, as some cruising will still occur ([Inci, 2015](#)).

Initial evidence suggests that cruising for parking affects a significant number of trips because of excessive demand and that this might be the case in many urban areas, although specific figures are a source of debate in the parking literature ([Shoup, 2005](#); [Weinberger et al., 2017](#)). Indeed, cruising statistics are difficult to determine, especially because studies use different definitions and employ different methods of measurement. This is a matter, however, that will attract much attention in the future if attempts at implementing optimal pricing are to be made. The methods traditionally used for measuring cruising can be grouped into three categories: (1) observational methods, including observing cars in the traffic flow and even following them, a method that has recently incorporated GPS technologies; (2) interviews and declared cruising, a method based on asking drivers about their cruising experience; and (3) park and visit tests. Recently, new methods have been adopted to measure cruising that avoid the main shortcoming of these traditional methods: namely, that they are labor-intensive, time-consuming, expensive, and hard to replicate. They include statistical methods involving the automated observation—employing street camera recordings—of cars that pass free spaces, which allows estimates to be made of the share of traffic that is cruising based on geometric probability distributions ([Hampshire and Shoup, 2018](#)). This approach represents a quick, cheap, and approximate method to measure cruising. Alternative approaches include transaction data methods, which use the information collected by parking management authorities to estimate cruising by relating parking space occupancy ratios and the parking rate (number of cars parking per unit of time). The evidence

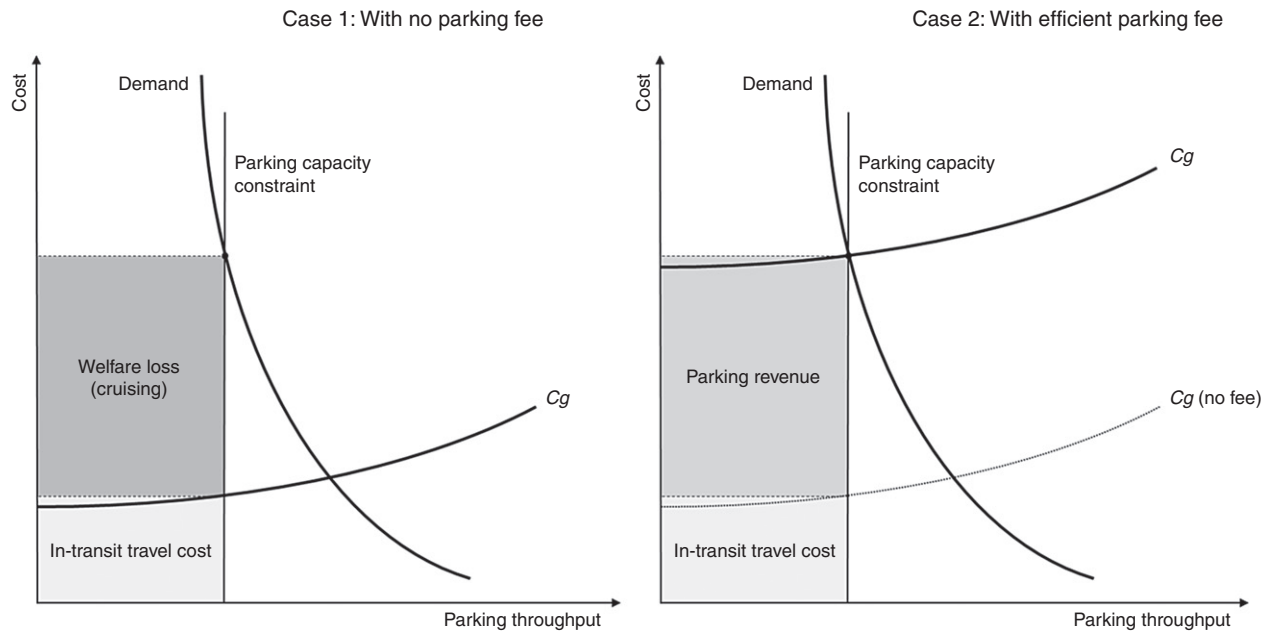


Figure 1 Welfare implications of cruising for parking and its correction through parking pricing.

suggests that the parking rate diminishes with occupancy ratios, and shows that it falls significantly with occupancy ratios close to 85%. These methods have huge potential, especially given the future availability of data collected using technologically advanced information systems.

The recent literature has also shown an interest in obtaining empirical estimates of the external costs of cruising. Although scarce, some studies provide estimates of these external costs for city residents. [van Ommeren et al. \(2011\)](#) report it to be about 1€ per day in Amsterdam; while [Inci et al. \(2017\)](#) find that the additional time visitors spend searching for each hour a driver stays parked is valued at about 15% of the average wage rate in Istanbul, which is of the same order of magnitude as the congestion costs experienced in transit from origin to destination in the city. To date, little is known about the external costs of cruising associated with raised levels of emissions during the search for a parking space and with traffic congestion. It is hoped that future research can fill this gap.

Market Power, Spatial Competition, and Information Frictions

Cruising is associated with curbside parking, but the parking market also includes off-street garages. Both goods serve the same purpose as non-perfect substitutes; thus, they have different characteristics and one might be preferred to the other depending on the specific context.^a Generally speaking, curbside parking tends to be spread more ubiquitously across a city, while public-access garages are discretely spaced due to construction scale economies. This, and drivers' walking costs to their final destination, gives garages a certain degree of localized market power.

Spatial competition models show that equilibrium in the parking market is reached when the full cost of parking at the curbside and in a garage are equated, adjusting for the variation in cruising times in the case of curbside parking (i.e., [Arnott, 2006](#); [Inci and Lindsey, 2015](#)). When garage prices are higher than those at the curbside, it makes the latter the preferred option and shifts demand to the curb while increasing cruising times. The interplay between the two means garage operators may be able to exploit their market power and take advantage of curbside congestion to charge higher markups.

The general assumption is that the higher the number of competing garages, the lower their prices will fall. However, the few empirical studies conducted to date suggest that they compete little with one other. In contrast, competition is much more intense with the curbside and a garage's dominant position in the area allows it to further exploit its localized market power.^b Additionally, recent evidence suggests that drivers might not be fully aware of the available parking alternatives and their characteristics, which introduce information frictions in the market. Drivers' knowledge in this respect seems to rely on their previous experience. Garages may have certain incentives to engage in the obfuscation of prices (complex and poorly displayed price schedules), so they can further exploit the drivers' lack of knowledge to charge higher prices and so take advantage of their localized market power and the drivers' costly search ([Albalade and Gragera, 2018](#)). The existence of this market distortion may be inferred from the recent boom in

^a Available empirical evidence suggests that curbside parking is preferred in the EU ([Kobus et al., 2013](#); [Gragera and Albalade, 2016](#)), but this can vary depending on garage facilities, safety, weather conditions, etc.

^b Most of these studies focus on the impact of mergers on garage prices, with one exception that includes the impact of curbside parking regulations on garage prices. See [Albalade and Gragera \(2017\)](#) for further description of this topic.

parking information platforms. The complex interaction between these parking market distortions remains an under-researched area requiring further study.

The Interplay Between Parking and Other Markets

As an intermediate good, the price of parking has an impact on the transport sector as a whole and on its externalities. Yet, it also means that its price affects the price of virtually any other good, given that any distortions in this market are transferred to other markets as negative welfare consequences.

When shopping malls provide free parking, they embed the parking costs in the stores' rents and they, in turn, transfer them to their retail prices. It is rational for them to use parking as a loss leader to attract customers into their facilities, and it is rationally used as insurance by shoppers who run the risk of not finding what they were looking for when visiting the mall (Ersoy et al., 2016). Downtown retailers may attempt to adopt similar strategies but with greater difficulties due to higher parking costs (parking regulation). Thus, they frequently lobby for lower parking fees to counter their being accessibility disadvantage vis-à-vis suburban shopping malls.

Likewise, many employers also provide their employees with free parking instead of paying them higher wages. This is mainly due to the tax advantages associated with fringe benefits, which reduce employers' labor costs, and to minimum parking requirements (MPRs), which ensure an abundant supply of parking. This means that employees pay less for parking than the resource costs inducing a welfare loss. van Ommeren and Wentink (2012) suggest that this loss is about 10% of parking resource costs attributable to distortionary fringe benefits taxation plus an additional 18% due to MPRs. These figures do not take into account additional welfare losses, such as those derived from higher congestion, energy consumption, and pollution resulting from encouraging workers to commute by car.

Housing prices are more or less transparently affected by on-site parking. MPRs impose construction costs on developers, which are in turn embedded into property prices, as they tend to bundle parking spaces with properties. Gabbe and Pierce (2017) suggest that the price of bundled units is 13% higher (with rents being about 17% higher) with an estimated deadweight loss for carless renters of about \$440 million per year. The relation between curbside parking and housing prices might be less apparent but is, nevertheless, relevant. If residents enjoy free curbside parking, its cost should be embedded in the property price as an amenity (given that they can use it as their own garage). Bakis et al. (2019) suggest that when the public authorities introduced paid parking this cost was unbundled from property prices, with average housing prices per square meter falling 9%. This link is also supported by evidence from Amsterdam, where the cost of waiting for parking permits (waiting list) is capitalized into housing prices. Here, estimates suggest that residents are willing to pay about 10€ per day for a parking permit (de Groote et al., 2016).

Parking Regulation and Its Political Economy

Most conventional approaches to parking policy interventions are undertaken by policymakers in the belief that it is an infrastructure that should be provided on-site to meet demand and so as to avoid any spillovers into neighboring areas (Barter, 2015). The tendency is to assume that private initiative alone will not provide an adequate supply to meet prospective demand and so impose MPRs, which also result in the private sector facilitating low-cost parking (firms for their employees, shopping malls for their customers, etc.). The unfeasibility of ever-expanding supply in areas of high demand has shifted the focus of intervention to using parking as a travel demand management measure in what has become known as an "area management approach." Taking this approach, cities have tended to keep parking prices low—lower than their social cost—and have focused on the expansion of off-street garages, while introducing controlled parking zones with different types of dedicated curbside spaces, with a clear bias toward residential permits (i.e., mixed use or resident-exclusive, rather than commercial spaces). Besides the introduction of parking pricing, many cities apply maximum stay limits. This is done to discourage long visits; yet, if limits are too lax they do not provide much cruising relief, and if too strict can cause capacity underutilization.

Policymakers take this approach in the belief that reducing the accessibility of cars might hamper the economic vitality of city centers and, at the same time, they seek to minimize political opposition to the implementation of the policy. Evidence in this field has shown that free or low-cost parking and conventional policies have a high cost for society. They encourage excessive car travel demand, impose major externalities (including cruising), and consume too much scarce urban space. This has generated support for the adoption of a "market-oriented" approach, where both supply and demand are managed in order to achieve efficiency.

Several studies have suggested different policy interventions to solve the common-property resource problem and to achieve full efficiency or, at least, to induce welfare gains. There seems to be a consensus on the need to price curbside parking appropriately so as to prevent both cruising and capacity underutilization (Russo et al., 2019). This might be achieved by setting prices at a level that ensures some parking spaces are available at all times, thus eliminating cruising. This needs to take into account garage fees, since increasing the fee differential in favor of garage parking will push drivers to cruise for curbside spaces.^c This is backed up empirically,

^c As long as curbside and garage parking are found not to be perfect substitutes, the differential between the two should be such that the perceived cost is equated. Studies suggest that this can be attained by regulating the price differential between garage and curbside parking, differentiated hourly curbside parking fees, and time-varying or uniform curbside parking fees depending on the restrictions to be faced. See Inci (2015) for further description.

as cities with curbside regulated parking spaces and a proper fee differential with respect to garages report almost no cruising levels. The right pricing also needs to take into account the variation in demand levels over time, leading to the use of dynamic or “performance-based” pricing (as implemented in cities such as San Francisco with its SFpark system).

Other studies show that it is equally important to price residential parking correctly, as the subsidy introduced by residential parking permits has a relevant negative welfare impact. MPRs in new building projects should be removed to avoid overprovision of parking at the expense of higher housing prices. Employer-paid parking is also a relevant issue, since it is a major determinant of employees’ choice of commuting mode. It has been suggested that income taxation exemptions be eliminated and schemes set up to promote alternative transport modes, either using cash-out schemes (allowing workers to choose between free parking or the parking subsidy in cash) or by taxing companies offering parking to their employees (Shoup, 2018).

Although several cities are moving in this direction, many still seem reluctant to adopt a “market-oriented” approach. The theoretical literature suggests that electoral competition and lobbying by interest groups (i.e., residents, retailers, or motorist associations) may explain why policymakers set parking fees too low and offer parking benefits to certain groups. Residents tend to lobby for higher parking fees for visitors, while policymakers are willing to provide them with parking benefits (residential permits) in order to win their electoral support. However, if parking prices negatively affect urban shops, which is a negative externality for residents, then there are incentives for them to team up with retailers to lobby for lower fees (de Borger and Russo, 2017). Parking is both a relevant source of revenue and a major investment in many cities, which means the political economy and institutional setting of parking policies will continue to attract the attention of researchers in the field.

Economic efficiency has centered most efforts when addressing parking issues but equity is also relevant from a policy perspective. Very few studies have focused their attention on equity, providing only weak evidence as to the fairness of paid parking for low-income groups. Chatman and Manville (2018) analyzed the SFpark project and found that the change in parking meter rates did not change the socioeconomic composition of curbside parkers, providing weak support for arguments that market-based fees price them out. In any case, it would be possible to use the parking revenues raised to compensate those negatively affected and, in this way, achieve a fairer outcome if needed. The relevance of this question clearly points to the need for further research.

Parking and Long-Term Decisions

Parking can also be expected to affect households’ long-term decisions—such as owning a car, choosing a residence or a job, or where to locate a firm—given that transport is known to enter the consumers’ trade-off when making such decisions. Research suggests that both owning a car and parking conditions are key elements of households’ travel decisions. If finding an empty space for your car proves costly, it makes it less attractive to own a vehicle, reinforcing parking regulation incentives toward a less car-dependent lifestyle; however, this link is still under researched. To date, only a few papers have examined this link, tending to focus on the impact of parking supply on car ownership. All of them suggest that parking availability increases motorization levels, regardless of whether they focus on free parking supply, parking norms, or residential permits. Guo (2013), for instance, reports that parking supply variables can have more than twice the impact of income on car ownership, depending on the household ownership level. However, there is not yet enough evidence on the impact of paid parking, dedicated spaces with distinct priorities being given to different consumer groups (visitors or residents) that compete for the same parking space, and specific parking regulations (fees, time limits, operating hours, etc.). Some initial evidence suggests that different types of dedicated space have different impacts on car ownership due to the way in which the allocation of parking rights facilitates competition between different demand segments for the same parking spots. As for residence location, it has been theoretically suggested that such a link exists and is relevant (Franco, 2017), but no empirical evidence is available yet.

Technology and Parking Innovations

Current technology has allowed the development of new business models in the parking market that monetize a solution for the inefficiencies discussed earlier but which are quite disruptive in what is still a traditional parking sector.

Space reservation systems have targeted both garage operators and municipalities to offer easy-to-find and quick-to-access parking spaces. The use of sensing technology allows drivers to obtain real-time data on occupancy levels while offering both guidance to parkers and the optimization of parking warden resource allocation.

Online transaction brokers are intermediaries that facilitate parking search and contract (reservation) through a single platform for a commission, as in the hotel and air transport industries. This can potentially generate additional services for both customers and parking operators through the use of gathered information on demand and supply availability. They will not only save drivers time but provide the public authorities with valuable information on how to regulate the market, and offer benefits to other stakeholders (including retailers and car manufacturers) that have a close relation with this market. In this same line, information-gathering platforms, which can also act as transaction brokers, seek to reduce drivers’ search costs in a market where search is costly.

The high sunk costs of garage parking are an incentive for owners to make the most of underutilized supply, especially when urban space is scarce and its conversion to other uses is quite limited. This has led to the emergence of “virtual” garage operators, whose business model is to condition and manage the facilities of others as public-access garages. This has led hotels and retailers to

open up their formerly private-access facilities to the public. In a few cities, some companies have gone even further and have set up parking facility networks that also open up residential private-access garages to the public.

Innovative technology-based parking management solutions are already an important niche of smart city industrial sectors. Smartphones, geolocation, and parking space sensing allow companies to set up innovative business models and encourage cities to implement integral parking management systems that help monitor occupancy levels, make payment seamless, and facilitate regulation enforcement. The capabilities that information technology systems bring to the table in terms of expanding regulatory flexibility and curbing provision costs should help the parking market reach a more efficient outcome and opens up interesting new research questions in the field. All these new systems and business models offer attractive features for addressing market distortions by fostering competition, easing information frictions, and reducing search costs, but researchers have only just started to focus on them and further development is needed.

References

- Albalade, D., Gragera, A., 2017. The determinants of garage prices and their interaction with curbside regulation. *Transp. Res. Part A Policy Pract.* 101, 86–97.
- Albalade, D., Gragera, A., 2018. Misinformation and misperception in the market for parking. *J. Transp. Econ. Policy* 52 (3), 322–342.
- Arnott, R., 2006. Spatial competition between parking garages and downtown parking policy. *Transp. Policy* 13 (6), 458–469.
- Bakis, O., Inci, E., Senturk, R.O., 2019. Unbundling curbside parking costs from housing prices. *J. Econ. Geogr.* 19, 89–119.
- Barter, P., 2015. A parking policy typology for clearer thinking on parking reform. *Int. J. Urban Sci.* 19 (2), 136–156.
- Chatman, D.G., Manville, M., 2018. Equity in congestion-priced parking: a study of SFpark, 2011 to 2013. *J. Transp. Econ. Policy* 52 (3), 239–266.
- de Borger, B., Russo, A., 2017. The political economy of pricing car access to downtown commercial districts. *Transp. Res. Part B* 98, 76–93.
- de Groot, J., van Ommeren, J., Koster, H.R.A., 2016. Car ownership and residential parking subsidies: evidence from Amsterdam. *Econ. Transp.* 6, 25–37.
- Ersoy, F.Y., Hasker, K., Inci, E., 2016. Parking as a loss leader at shopping malls. *Transp. Res. Part B Methodol.* 91, 98–112.
- Franco, S.F., 2017. Downtown parking supply, work-trip mode choice and urban spatial structure. *Transp. Res. Part B Methodol.* 101, 107–122.
- Gabbe, C.J., Pierce, G., 2017. Hidden costs and deadweight losses: bundled parking and residential rents in the metropolitan United States. *Hous. Policy Debate* 27 (2), 217–229.
- Gragera, A., Albalade, D., 2016. The impact of curbside parking regulation on garage demand. *Transp. Policy* 47, 160–168.
- Guo, Z., 2013. Does residential parking supply affect household car ownership? The case of New York City. *J. Transp. Geogr.* 26, 18–28.
- Hampshire, R.C., Shoup, D., 2018. What share of traffic is cruising for parking? *J. Transp. Econ. Policy* 52 (3), 184–201.
- Inci, E., 2015. A review of the economics of parking. *Econ. Transp.* 4 (1–2), 50–63.
- Inci, E., Lindsey, R., 2015. Garage and curbside parking competition with search congestion. *Reg. Sci. Urban Econ.* 54, 49–59.
- Inci, E., van Ommeren, J., Kobus, M., 2017. The external cruising costs of parking. *J. Econ. Geogr.* 17 (6), 1301–1323.
- Kobus, M.B., Gutiérrez-i-Puigarnau, E., Rietveld, P., van Ommeren, J., 2013. The on-street parking premium and car drivers' choice between street and garage parking. *Reg. Sci. Urban Econ.* 43 (2), 395–403.
- Lehner, S., Peer, S., 2019. The price elasticity of parking: a meta-analysis. *Transp. Res. A* 121, 177–191.
- Marsden, G., 2006. The evidence base for parking policies—a review. *Transp. Policy* 13, 447–457.
- Russo, A., van Ommeren, J., Dimitropoulos, A., 2019. The environmental and welfare implications of parking policies. Environment Working Paper No. 145, Environment Directorate—OECD, Paris.
- Shoup, D.C., 2005. *The High Cost of Free Parking*. Planners Press (Routledge), Chicago, IL.
- Shoup, D., 2018. *Parking and the City*. Planners Press (Routledge), Chicago, IL.
- van Ommeren, J., Wentink, D., 2012. The (hidden) cost of employer paid parking. *Int. Econ. Rev.* 53 (3), 965–978.
- van Ommeren, J., Wentink, D., Dekkers, J., 2011. The real price of parking policy. *J. Urban Econ.* 70 (1), 25–31.
- Weinberger, R., Millard-Ball, A., Hampshire, R.C., 2017. Parking search-caused congestion: where's all the fuss? In: *Proceedings of the 96th Transportation Research Board Annual Meeting*, Washington, DC.

Loss Aversion and Size and Sign Effects in Value of Time Studies

Andrew Daly, Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Microeconomic Theory	165
Measurement Methods	165
The Sign of Time Differences	166
The Size of Time Differences	167
The Use of Time Values	168
References	168
Further Reading	169

Microeconomic Theory

The marginal value of travel time (VTT), that is, the money equivalent per minute of travel time, is often said to be the most important number in transport economics, and its estimation has therefore been the topic of extensive academic and applied work. A substantial theory has been developed by economists (e.g. Jara-Díaz and Astroza, 2013) in which consumers are represented as optimising their utility subject to budget constraints on both time and money. This theory allows us to make the fundamental claim that time can have value, being a constrained resource like money. Further, it underpins the calculation of money amounts that would leave consumers in principle indifferent between transport policy scenarios that imply different levels of time expenditure, providing monetary adjustments are also made. This theory and these calculations are the basis for the appraisal of transport projects and policies that change travel times, giving equivalent monetary amounts that are widely used to indicate whether a proposed policy is advantageous to society as a whole.

The VTT appearing in this theory has two components. One relates to the overall time budget that we all face, 24 hours per day. If time spent travelling is reduced, time becomes available for other activities. The second component relates to the specific activity required for travel and is different, for example, between driving a car and waiting for a bus. The existence of this second component means that the specific travel activity must always be considered in stating VTT. We also see that, unlike money, time cannot be stored or borrowed, it must immediately be transferred to another activity, so that the VTT is always measured relative to the next-best activity; this property implies interpersonal variation in VTT, as the next-best activity will naturally vary between individuals.

It is important to note that in the microeconomic theory of consumer behaviour VTT is implied by an equilibrium optimum chosen by consumers in their own current circumstances, that is, this theory does not relate to the history of changes experienced by an individual. The value can vary with the *total* amounts of time and money expended by an individual, for example, as the constraints bind more or less stringently, but the notion of *change* is not relevant to determining the value, according to microeconomic theory, which describes static optimum behaviour. Therefore, in the primary application of VTT, the economic appraisal of transport projects, a single value is required, which will vary between the types of travel activity (driving, walking, etc.) and may vary between individuals, but is independent of the history of changes to the transport system.

Similarly, in the secondary use of VTT, which is to calculate generalised cost or similar time–cost composites for use in demand forecasting, the notion of change does not enter the discussion. The theory here is that each alternative (mode, destination, etc.) has a utility (negative generalised cost) composed of time and cost, brought to a common scale by the VTT. But there is no dependence of utility on previous situations, so that the notion of change is again irrelevant.

Thus, the theories on which the use of VTT are based, both for appraisal and for forecasting behaviour, require a single value for each context, independent of the consumer's history, and therefore issues of sign or size do not arise. However, when methods for measuring VTT are considered, these issues can enter the discussion.

Measurement Methods

While it indicates how VTT is to be used, the theory of time value does not indicate how empirical values are to be obtained. Historically, three approaches that have been used are as follows.

- Models can be estimated from observed (or reported) behaviour and the ratio of time and cost coefficients in those models interpreted as the value of time. Inferring VTT from revealed preferences in this way was perhaps the earliest approach for estimating values of time and corresponds to a classical economist's approach to deriving implicit values from observed behaviour in the marketplace. Here, the notion of change is not directly relevant; the models predict demand based on current travel times and costs without reference to previous times and costs.

- The cost savings approach (CSA) is applicable to time spent travelling by people in the course of their employment, such as plumbers or business representatives travelling to a distant place of work, or professional drivers. CSA then calculates the wage and additional costs incurred by the employer per minute and attributes this as the value of time. Marginal CSA values are independent of the amount of time spent and do not vary by the type of travel activity. Variants of CSA, little used but more sophisticated, such as the Hensher formula, allow for employee productivity en route and for impacts on the individual as well as the employer.
- Willingness to pay (WTP) can be estimated from travellers' stated preferences, most often implemented in the form of stated choices (SC). This approach presents each of a sample of respondents with a series of choice sets, constructed by an experimental design, and from their hypothetical choices estimates of the relative preferences for time and money are inferred. Because surveys of this type almost always present scenarios that relate closely to an actual trip made by the respondent, to preserve realism, the notion arises of *change* relative to that observed trip, often called the reference trip. In the modelling analysis of data of this type, the *sign* and *size* of this change are often found to be relevant to the value that is estimated. In current practice, the SC has come to be the most widely applied approach to estimate the VTT outside of working hours.

Thus, the importance of changes as an influence on behaviour arises only in the context of SC data, but, because of the dominance of SC in practical estimations of travel time value, the issue of the size and sign of time changes becomes relevant and, in fact, important.

For example, in the appraisal of typical transport projects, particularly for road schemes, it is often the case that a substantial part of the value difference between the proposed scheme and an alternative (e.g. 'do minimum') is made up of large numbers of people whose trips would differ in time between the scenarios by small amounts of time, often measured in seconds (e.g. as shown by [Welch and Williams, 1997](#)). When SC data are analysed, provided that the estimation process is sufficiently detailed and the data adequate, values are often found that depend on the size of the time differences being valued: specifically, small time differences are given a lower value (per minute) than larger time differences. Because of the importance in practice of small time differences, the divergence of SC results from microeconomic theory could have substantial relevance for transport policy.

Similarly, it is frequently found that losses of time, that is, increases in journey time, are valued more highly per minute in SC data than gains in time. Because transport policy aims to reduce travel time, the suggestion that time reductions could have a lower value could have important policy implications.

Clearly, the emergence of value differences by the size and size of change is awkward to reconcile with the microeconomic theory. It also presents issues for the process of appraising transport projects. For example, if a project could be implemented in two stages, each giving part of the time reductions relative to a base case, this should surely give the same total benefit as implementing the whole project, but if VTT depends on the size of time saving this will not be the case. Similarly, a temporary change in travel time which is exactly reversed should not have a long-term impact, but if VTT depends on the sign of the time change there will be a permanent loss, as the gain value will be more than outweighed by the loss value.

Essentially, the microeconomic theory and practical logic are at odds with the psychological theory that seems to be needed to explain how travellers respond to SC surveys. This article now continues to discuss these findings, how they might relate to the processes used to estimate the values and what governments have done to resolve the issues that the results raise. First we deal with the impact of the sign of time differences, then the issue of their size, followed by the approaches that have been used to resolve the issues.

The Sign of Time Differences

It is a common observation of human behaviour that people are more upset about a loss than they are happy about an equivalent gain. Politicians appear to be well aware of this effect, which is related to the concept of endowment: 'what I have I hold.' In behavioural analysis, the effect is known as loss aversion and it is an important component of Prospect Theory. There is no doubt that loss aversion is a feature of human behaviour in many contexts.

In the estimation of VTT, WTP implies paying for time, that is, a loss of money equivalent to a gain (saving) in time. In contrast, willingness to accept (WTA) implies a gain of money to compensate a loss in time. Any loss aversion relating to time or money, if it exists in this context, will imply $WTP < WTA$. For example, if there is loss aversion relating to time, then, for a given amount of time loss, the money needed as compensation (WTA) would have to be more than the money that would need to be paid (WTP) for a gain of the same amount of time. A similar argument applies to loss aversion relating to money and, more strongly, if there is loss aversion in both time and money.

The empirical evidence for loss aversion in the estimation of time values from SC data is quite strong. Typically, the monetary value of a loss (WTA) could be 50% higher than the monetary value of a gain (WTP) of the same amount of time, or more when loss aversion is found to apply to both time and money, and the statistics generally indicate that the difference is significant. Since the effect is clearly present in many SC data sets, analysing SC data without at least testing for loss aversion risks biasing other parameters in the model. The methods for making these tests are relatively straightforward.

A more difficult issue is how the effect should be handled in deriving VTT for practical use. When used in forecasting, the notion of change does not apply, since travellers will experience many changes in travel time between their current journey pattern and the

future year for which forecasts are made. They may also move home, leave school, get new jobs, etc., so that the travel pattern is completely different and a reference trip does not exist. Moreover, a fixed value is needed for all units of time, to be consistent with the microeconomic theory. The VTT for gain and for loss must somehow be reduced to a single value.

The solution that has been adopted in practice is to assume that there is some 'underlying' VTT, representing the traveller's long-term preferences, and that the gain and loss values are distorted by the elicitation process or perhaps are short-term effects only. The value that should be used in practice would then be some sort of average of the gain and loss values; typically the geometric mean is used, though the reason for preferring this to other averages (e.g. the arithmetic mean) is purely that it is more convenient in the models (De Borger and Fosgerau, 2008; Hess et al., 2017). Note that this averaging after model estimation is clearly preferable to simply estimating a single value that applies to both gains and losses, because of the potential for distortion of other parameters of the model; additionally, estimating a single value depends on the assumption that losses and gains are balanced in the data.

In summary it can be said that sign differences in time values usually exist in SC data, that they are almost certainly caused by loss aversion, which is itself well established in many contexts, they need to be represented in models estimating VTT and that they have to be handled in practice by some sort of posterior averaging of the gain and loss values.

The Size of Time Differences

Because of the importance of small time differences for many journeys between forecast scenarios, concern is raised by the common finding in analysis of SC data that small time differences, for example under 5 minutes or under 3 minutes, are relatively less highly valued or even not valued at all. The questions that are then raised include the following:

- How can such an effect be brought about?
- Does it depend on the estimation or elicitation methods?
- Is it a component of long-term traveller preferences?

In considering the possible cause of the size effect, it is relevant to think how people think about differences in time or money between alternatives. The standard assumption made by modellers is that they compare the time difference in minutes with the money difference in currency units. But it is also reasonable to consider the alternative assumption that they compare proportional differences and would prefer, for example, a 20% gain in time to a 10% loss in money. Such thinking would obviously be relevant to the comparison of small differences with large ones for trips of a given length, but it does not explain why VTT is often estimated to be less per minute for small time differences.

A possible explanation is given by the different nature of time and money. Presented with a small gain or loss in money, respondents may be able to think of this as an adjustment to their current stock of loose change and therefore to have a moderate value; however, being given a small amount of extra time may have little value, while a loss of a small amount of time may be relatively easy to accommodate, given existing plans for the day. If these mechanisms operated, it would indeed be the case that small time differences had lower value per minute, but perhaps only in the short term.

Another interpretation of the effect would be that the marginal value of time or money increases as the ultimate budget becomes more dominant. The VTT in money terms would then increase as the amount of time and money increase, as is observed, provided that the time budget was more restrictive than the money budget. Large changes in time or money would bring the respective budget limit closer. This explanation would operate in the long term as well as in the short term, providing that the relative stringency of the budgets remains unchanged. However, the empirical evidence argues against this explanation, as the marginal value of both time and cost appears to diminish as the amounts increase.

In summary it seems that we do not have a good intuitive explanation of the size effect. Perhaps respondents are simply indifferent to small amounts of time but willing to attach value to small amounts of money.

The evidence for lower values for small time differences is largely drawn from studies with relatively few observations of such differences, with little variation in the size of small time differences and with relatively simple SC experiments. These are the characteristics of the early SC studies of VTT, where experimental design was not so well developed and on-line interviewing was not possible. In many cases paper survey forms had to be pre-printed with time and money differences presented relative to the reference trip. More recent studies have been able to use more variation in the characteristics of the alternatives presented, with more focus on smaller time differences but without explicit use of the reference trip, and it appears that these changes have reduced the value-per-minute gap between small and larger time differences.

If a reduced value for small time differences is suspected in SC data, this can be investigated by estimating a separate value for each time difference: 1 minutes, 2 minutes, etc. This approach, which was used in a number of early VTT studies, gives the maximum flexibility to determine how the values for small time differences vary, but may not give stable results when the amount of data is small for some time differences. An alternative approach is to postulate a simple functional form, such as that the VTT depends on a power function of the time difference, and estimate the parameters of that function. The latter approach, used in some recent studies following the Danish example, gives more reliable parameter estimates but may miss detailed variations in value.

Because there is no clear intuitive explanation of the size effect, it is difficult to argue that it could be a component of long-term preferences. Some of the proposed explanations imply that it would be a short-term effect or even something provoked by the SC

context. It can be concluded that at present the size effect is not believed to be part of long-term preferences; it would in any case be difficult to see how it could function over the longer term, with shifting reference points.

A complication in connection with the size effect arises when the scenarios presented in the SC experiment have time, cost, or both equal to those of the reference trip. If account is taken of these cases in the modelling, it is often found that the current trip (often called 'as now') is more attractive than would naïvely be expected. Suppressing the effect in the model when it is significant in the data will cause the size effect to be overstated for time and/or for cost, so that the impact on the estimated VTT could be a bias in either direction. Other parameters might also be biased by this omission.

Whatever the fundamental cause of the size effect, it has been observed in a number of SC studies. As with the sign effect, it is therefore necessary to allow for the possibility of size effects (and 'as-now' impacts) in the analysis of SC data (e.g. as in [De Borger and Fosgerau, 2008](#); [Hess et al., 2017](#)). The effects are not always present, but whether or not they will be found in any given data set cannot be predicted without further research. It is possible that more sophisticated designs, for example, with more numerous time attributes, can avoid sign effects and there is some empirical evidence of this.

As with the sign effect, when a size effect is found in SC data a decision needs to be made about how to calculate a value of time for use in appraisal. The key decision is then what time difference should be used to give the representative VTT. Given the small time differences often arising in practical scenario appraisals, it might seem that the values indicated by small time differences might be preferred. However, governments have been reluctant to do this, because the value obtained from these small differences appear much smaller than those obtained from other estimation methods. Moreover, the values associated with small time differences often have rather wide confidence limits, depending on the estimation procedure used, which might undermine confidence in appraisals. Accordingly, an arbitrary decision has usually been taken to use the value associated with a specific moderate time difference, often 10 minutes. This choice is clearly very arbitrary and cannot be described as satisfactory. Governments are however reassured by the fact that this calculation is consistent with past work and with the procedure adopted by other governments.

The Use of Time Values

The estimation of VTT from the analysis of SC data usually gives rise to a sign effect, where losses are valued more highly than gains, and sometimes to a size effect, where small time differences are valued less than moderate or large differences. Other estimation methods, CSA or inference from revealed preference data, do not give these differences. To use VTT, however, a single value is required for consistency with economic theory and because otherwise anomalies can be created in which, for example, the separate benefits of parts of projects do not add up to the total benefit of the project.

It is necessary to include sign and size effects in the model used for estimating VTT from SC data, to avoid biasing the other parameters. However, this implies that a method must be applied for obtaining a best overall estimate of VTT from models that include sign and size effects. In this process, the size effect is considerably more difficult than the sign effect.

- The sign effect is explicable in terms of well-established loss aversion and it seems clear and generally acceptable that an average of gain and loss values can be used.
- No satisfactory intuitive explanation for the size effect seems to exist. To solve the problems it raises, most governments have chosen an arbitrary time difference, usually 10 minutes. Note that it is not reasonable to take the average of the time differences presented in the SC survey, as this would leave the VTT dependent on the survey design, though it would be possible to use the average of several arbitrary time differences. Several governments (e.g. Canada and Germany) have attributed lower value to small time differences in project appraisals and others (e.g. the United Kingdom and EU) have required reporting of small differences ([Daly et al., 2014](#)), although these governments may now have dropped this approach. But all of these approaches are unsatisfactory, as long as the basis for the size effect is not understood.

For short-term projects, an appraisal based on the changes to be experienced by travellers may be acceptable, but here there are also difficulties and these are not the key projects for which VTT is required.

The fundamental issue is that the economic theory required for appraisal excludes the behavioural effects that are required to explain the responses to SC surveys. For the future the options are to try to improve the presentation and analysis of SC surveys or to take a completely different approach to VTT estimation, such as the analysis of revealed preference data.

References

- Daly, A., Tsang, F., Rohr, C., 2014. The value of small time savings for non-business travel. *Transp. Econ. Policy* 48, 205–218.
- De Borger, B., Fosgerau, M., 2008. The trade-off between money and travel time: a test of the theory of reference-dependent preferences. *J. Urban Econ.* 64, 101–115.
- Hess, S., Daly, A., Dekker, T., Ojeda, C.M., Batley, R., 2017. A framework for capturing heterogeneity, heteroskedasticity, non-linearity, reference dependence and design artefacts in value of time research. *Transp. Res. B* 96, 126–149.
- Jara-Díaz, S.R., Astroza, S., 2013. Revealed willingness to pay for leisure. *Transp. Res. Rec.* 2382 (1), 75–82.
- Welch, M., Williams, H., 1997. The sensitivity of transport investment benefits to the evaluation of small travel-time savings. *J. Trans. Econ. Policy* 31, 231–254.

Further Reading

- Bates, J., Whelan, G., 2001. Size and sign of time savings. Institute of Transport Studies, University of Leeds. Available from: <http://eprints.whiterose.ac.uk/2065>.
- Fosgerau, M., Hjorth, K., Lyk-Jensen, S.V., 2005. The Danish value of time study. Final Report. Danmarks Transportforskning, Kongens Lyngby.
- Gunn, H.F., 2000. An introduction to the valuation of travel-time savings and losses. In: Hensher, D.A., Button, K.J. (Eds.), *Handbook of Transport Modelling*. Elsevier Science Ltd., Pergamon.
- Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47 (2), 263–291.
- Significance, VU University, John Bates Services, TNO, NEA, TNS NIPO, PanelClix, 2013. Values of time and reliability in passenger and freight transport in The Netherlands. Report for the Ministry of Infrastructure and the Environment. Significance, The Hague.

Intertemporal Variation of Valuations

James Fox, RAND Europe, Cambridge, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	170
Transferability of Cross-Sectional Models	170
Longitudinal Studies	171
Repeated Cross-Sectional VOT Studies	171
Other Evidence	172
Summary	172
References	172

Introduction

Whenever models are applied to make predictions of future behavior the issue of temporal variation of valuations is important. However, it is an issue, which has received less attention from researchers than efforts to develop models best able to explain current behavior. A closely related issue is that monetary valuations of travel time changes are expected to change over time, and different assumptions around that growth can have important impacts for forecasting. Finally, valuations in transport may change due to wider changes, for example, the use of laptops and smart phones may impact on valuations of travel time for long distance rail travelers.

There is an important distinction between cross-sectional variation in behavior and changes in behavior over time. In recent years, many academic studies have focused on developing more complex representations of travel behavior, for example, by better capturing heterogeneity in preferences between individuals or by modeling the impact of underlying variations in attitudes on traveler behavior. However, while these models can provide a more nuanced understanding of current behavior they do not necessarily yield models better able to forecast behavior over time.

To make predictions of future behavior, it is necessary to make assumptions about how the parameters change over time, if at all, and how the population changes over time. It is important to note that many forecasts are made making the implicit assumption that the model parameters are constant over time without reviewing the evidence for that assumption. Another issue for forecasting is that for some valuations, such as the variation of cost sensitivity to income, the cross-sectional elasticity differs from the longitudinal elasticity and this issue needs to be carefully considered in forecasting.

The issue of population forecasting is important when applying models, if a model incorporates segmentation of behavior across a particular variable then to apply that model there is a need to forecast the future distribution of that variable in the population. This issue is not the focus of this chapter but it is worth noting that applying more complex model forms may be more problematic because of these issues.

This review gives particular prominence to intertemporal variation in values of time (VOTs). This is because this measure is key to both transport modeling and appraisal, and given this there is considerable evidence on how these valuations vary with time.

The remaining sections of this article discuss the various approaches that can be made to assess temporal variation in valuations and summarize what researchers have found. The next section summarizes the temporal transferability literature where the temporal stability of model parameters is assessed, often by using data collected at different points in time. Next, the use of longitudinal studies, which directly estimate variations in valuations using time series data, is discussed. A number of VOT studies have been repeated over time and these sets of studies provide evidence on temporal variation of valuations. Other evidence also provides information on temporal variation including literature reviews and travel budget theory. The final section summarizes the evidence on temporal variation in valuations and sets out areas where further research would be valuable.

Transferability of Cross-Sectional Models

In the transferability literature valuations for different transport alternatives and the characteristics of those alternatives are expressed by the model parameters. The transferability of a model is its ability to predict behavior in a different context to the estimation context (Koppelman and Wilmot, 1982). Transfers may be spatial, for example, from one city to another, but in this context we are concerned with temporal transfers (i.e., transfers over time) of models within the same spatial area.

Most studies have sought to assess temporal transferability by using data collected at two or more points over time. This approach allows the predictions of a transferred model to be compared to those of a model re-estimated on the transfer data. The success of the transfer is then judged based on how well the transferred model fits the transfer data compared to the same model specification

re-estimated on the transfer data. However, in addition some model transfer studies have assessed changes in the model parameters over time and used this evidence to support or otherwise the hypothesis of temporal stability of valuations.

Many of the early model transfer studies were mode choice studies where the short-term impact of a policy intervention on mode choice was assessed (Fox and Hess, 2010). However, some transport models are applied over much longer forecasting horizons of 20–30 years and over this much longer transfer period model transferability may be less likely. In these contexts assembling consistent data suitable for assessing parameter transferability is a significant challenge. In particular, while similar travel diary surveys may be available from different points in time the availability of consistent highway and public transport network models is often an issue.

In some studies, in the absence of detailed data collected at different points in time researchers have examined the ability of models to predict observed aggregate changes in traveler behavior to make assessments of model transferability, for example Milthorpe (2005). However, in such cases it can be difficult to disentangle errors in the forecast of the model inputs from model transferability issues and furthermore such approaches are better placed to assess overall model transferability rather than the transferability of individual parameters.

Much of the transferability literature dates from the late 1970s and early 1980s when disaggregate models were being applied in transport contexts for the first time. Most studies then were mode choice studies and these were generally support of the hypothesis of model transferability over time, albeit based on relative short time scales of 5 years or less. A noteworthy finding that improving the model specification by adding socioeconomic parameters can yield a more transferable model because it helps identify better estimates of the key level-of-service and cost parameters, see for example Silman (1981).

Some of these studies also allow direct assessment of temporal variation in parameter values. By calculating the percentage change in parameter values in each study it possible to examine how the level of transferability varies between different types of parameter (Fox et al., 2014). This analysis found that the level-of-service (travel time, out-of-vehicle time, interchanges, etc.) and socioeconomic parameters were most transferable and as might be expected the constants were the least transferable parameter group.

Some more recent evidence from transfers over longer periods is available from models of either mode choice or joint mode-destination choice for urban areas. The findings from these studies are generally consistent with the earlier mode choice evidence, finding models are generally transferable over time but that the level of transferability varies substantially with parameter group.

A noteworthy finding from some of these studies was that sensitivities to in-vehicle travel were stable to within $\pm 10\%$ over periods of up to 20 years. Thus, there is evidence that valuations of travel time are stable over the long term. This finding is also important in the context of VOTs, which are probably the key valuation measure in transport modeling. In models that have separate in-vehicle time and cost parameters it suggests that adjusting these parameters for future VOT growth can reasonably be achieved by adjusting the cost parameter alone rather than adjusting both sets of parameters.

Longitudinal Studies

Longitudinal studies use data collected over time to provide direct estimates of how key valuations vary over time, often using model forms that directly estimate elasticity to the dependent variables. Typically such approaches use simpler aggregate models compared to the cross-sectional models used in transferability analysis. Their simpler form means that their data requirements are typically less onerous than those of cross-sectional approaches.

A good example of how longitudinal evidence has been used to supplement cross-sectional models is in establishing relationships between VOTs and incomes. Cost sensitivity valuations would be expected to change over time as real incomes increase. However, researchers have found that cross-sectional income elasticities are often substantially lower than corresponding longitudinal values. As such cross-sectional models estimated using data from a single point in time could not forecast changes in cost sensitivity. If data is collected at two points in time, in principle this could be used to estimate a relationship but cost is a particularly difficult variable to model accurately as factors like parking costs and public transport fares are difficult to precisely identify for a given individual, and further there is often a correlation between cost and travel times variables such that if one is weakly estimated the other parameter strengthens. A final related consideration is that the longitudinal analyses have directly estimated VOT relationships whereas in many cross-sectional models VOTs are implied from the ratio of in-vehicle time and cost parameters.

In the United Kingdom, a hybrid approach has been used whereby a longitudinal meta-analysis has been run using the results from cross-sectional VOT studies undertaken over almost 50 years. The total number of valuations included in the analysis was 1750. The meta-analysis models represented variations in valuations by distance, travel purpose, travel mode, data type (revealed versus stated preference) and other effects. The model identified a highly significant relationship between value of time and GDP per capita with an elasticity of 0.9. The model also provides long-term estimates of the ratios between in-vehicle time and other values included in the dataset including out-of-vehicle components for public transport (Abrantes and Wardman, 2011).

Repeated Cross-Sectional VOT Studies

Given their importance to transport planning national VOT studies are often repeated relatively frequently and therefore provide valuable evidence on intertemporal variation in valuations.

Evidence from Dutch VOT studies conducted in 1988 and 1997 found that increase in VOT due to income growth had been offset by a trend decline in VOT such that the real VOT remained more or less constant between the two periods (Gunn, 2001). The most recent Dutch VOT study undertaken using data collected in 2009 and 2011 also identified real VOTs essentially unchanged from the previous 1997 study. Similar findings were made when 1985 and 1994 the United Kingdom VOT studies were compared, and further when the 1994 results for car were compared to those from smaller 2006 study focused on motorway drivers only (Gunn, 2001). Researchers have suggested that increased use of technology such as mobile phones and laptops may have contributed to these findings. However, in contrast to these findings analysis of the 1994 and 2007 Swedish VOT data for car drivers found that the travel cost parameter had declined in real times while the travel time parameter had remained constant, and as such VOTs had risen due to income growth (Börjesson, 2014).

Together these repeated VOT studies raise the question as to why a number of these studies do not identify real growth in VOT whereas the United Kingdom meta-analysis identified strong growth over time. Further research is needed to understand these different findings, in principle they may still be consistent with assuming cost sensitivity decreases as a function of income if sensitivity to travel times due to technological changes also decrease to a compensating amount. The nature of the travel represented may also be relevant, evidence from cross-sectional models often comes from congested urban areas whereas the VOT studies place a greater emphasis on longer-distance travel that may be more impacted by improvements in comfort and technological changes that enable on-board time to be used more effectively.

Other Evidence

In the UK context, evidence from numerous valuation studies was assembled and then a longitudinal model was fitted to the data to test for parameterize relationships between the valuations and income. Alternatively, a literature review approach could be used to summarize evidence from different studies, make assessments of the quality of evidence available from each study, and then form conclusions as to how the valuations have evolved over time.

A separate strand of evidence comes from travel budget theory. Under travel budget theory, individuals have fixed travel budgets of time and money that limit their consumption of these goods. If these budgets remain constant over time then this would provide evidence for stability of valuations over time.

Researchers have often found that at the aggregate level travel budgets are stable between areas and over time, though at the individual level considerably greater variation is observed (Schafer, 2000). Given that the valuations captured in cross-sectional models represent average preferences for the estimation sample these findings are generally consistent with the finding from mode-destination transferability studies that sensitivities to in-vehicle time are constant over time. It is also worth noting that activity-based models place greater emphasis than traditional modeling approaches on the need to represent individual's scheduling constraints on their travel and activity patterns.

Summary

The issue of temporal variation (or stability) in valuations is key to transport modeling and appraisal but it is an area that has been somewhat neglected by researchers, in part because of the data challenges associated with investigating the issue.

An important finding from the transferability literature is that improving model specification can improve model transferability, and in particular helps ensure more transferable estimates of the key cost and time parameters that are key to the future model predictions. Another important finding, and one that has supporting evidence from VOT and travel budget literature, is that in-vehicle time parameters tend to be more transferable than other model parameters.

The issue of predicting VOT growth can be problematic. In many European countries, predictions of growth in VOT are linked to predicted growth in incomes, and in the United Kingdom, this approach is well evidenced by a large meta-analysis of historical studies. However, a number of repeated VOT studies have observed static or declining VOTs despite income growth.

There are a number of areas where further research would be valuable.

The majority of the transferability literature investigated multinomial or nested logit model forms. It would be valuable for research into more complex model forms that have been prominent in academic studies recently, such as mixed as latent class models. Some limited evidence from an investigation of mode-destination models found mixed logit models to be no more transferable than simpler nested logit models.

The issue of VOT growth over time is one that is worthy of further investigation. In particular, the relative contributions of technological and other changes on VOT need to be better understood as it is possible that these can be controlled for then the underlying VOT would still grow with income.

References

- Abrañtes, P., Wardman, M., 2011. Meta-analysis of UK values of travel time: an update. *Transp. Res. Part A* 45 (1), 1–17.
Börjesson, M., 2014. Inter-temporal variation in the travel time and travel cost parameters of transport models. *Transportation* 41 (2).

- Fox, J., Hess, S., 2010. Review of evidence for temporal transferability of mode-destination models. *Transp. Res. Rec.* 2175, 74–83.
- Fox, J., Hess, S., Daly, A., Miller, E., 2014. Temporal transferability of models of mode-destination choice for the greater toronto and hamilton area.. *J. Transp. Land. Use.* 7 (2), 65–86.
- Gunn, H., 2001. Spatial and temporal relationships between travel demand, trip cost and travel time.. *Transp. Rev. E.* 27, 163–189.
- Koppelman, F., Wilmot, C., 1982. Transferability analysis of disaggregate choice models. *Transp. Res. Rec.* 895, 18–24.
- Milthorpe, F., 2006. A comparison of long term sydney forecasts with actual outcomes.. *Australasian Transport. Res. Forum* 28.
- Schafer, A., 2000. Regularities in travel demand: an international perspective. *J. Transp. Stat.* 3, 1–31.
- Schafer, A., 1981. The time stability of a modal-split model for Tel-Aviv.. *Environ. Plan A* 13, 751–762.

The Rebound Effect for Car Transport

Bruno De Borger*, Ismir Mulalic[†], Jan Rouwendal[‡], *University of Antwerp, Antwerp, Belgium; [†]Copenhagen Business School, Copenhagen, Denmark; [‡]VU University, Amsterdam, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	174
The Direct Rebound Effect	174
How Large is the Direct Rebound Effect?	176
Estimating the Rebound Effect	176
Empirical Estimates of the Rebound Effect	176
Conclusion	177
References	178
Further Reading	178

Introduction

The transportation sector is responsible for a significant share of the world's energy use, and it contributes substantially to carbon emissions and local air pollution. One way to reduce the transport sector's energy consumption and the associated negative external effects is to enforce fuel efficiency standards on vehicle manufacturers; examples include the corporate average fuel efficiency standards in the United States and the weight-based fuel efficiency standards in the EU. However, an often-observed effect of the introduction of more stringent fuel efficiency standards is that as the fuel efficiency of cars improves, car use becomes cheaper, thereby providing an incentive to increase its use. Total fuel use thus responds less than proportionally to changes in fuel efficiency. The *direct rebound effect*, or the *pure rebound effect*, is defined as the deviation from this proportionality (Gillingham, 2014).

In the literature many other types of rebound effects have been distinguished, and the typology used is not always the same, even among specialists. One classification is the following (Gillingham et al., 2013). Apart from the direct rebound effect mentioned earlier, two other microeconomic effects are considered. The *indirect rebound effect* is defined as the increased energy consumption from changes (both substitution and income effects) in the use of other energy-using products than the one where the improvement occurs (e.g., better fuel efficiency leads people to spend less on gasoline, but the money saved may be used to buy an extra plane ticket). An *embodied energy rebound effect* accounts for the energy used to create the energy efficiency improvement. Moreover, several macroeconomic effects have been identified, including a price effect and an economic growth effect. Typically, the alternative types of rebound effects point to phenomena that make the total rebound effect larger than the direct part. It has even been suggested that under particular circumstances the rebound effect may more than fully compensate the initial effect, a possibility referred to as backfiring.

The existence of rebound effects is of course not limited to the car industry, but it has been observed in many other energy-consuming durable goods industries as well. It has first been observed by the economist William Stanley Jevons as early as 1865. He noticed that despite huge improvements in energy efficiency of coal-fired steam engines, the use of coal was actually increasing. The rebound effect was so strong that the energy efficiency improvement "backfired": the cost reductions in fact raised energy use. Based on the evidence reported later, such backfiring has not been observed for fuel efficiency improvements in the car industry.

This paper focuses solely on the direct rebound effect, and the attention is limited to vehicles; unless otherwise noted, our focus is on cars. Better fuel efficiency reduces the per kilometer cost of car use, raising the demand for driving, so that some of the energy savings that would have been realized with unchanged behavior are foregone. This behavioral response reduces the effectiveness of fuel-efficiency improvements. The empirical question therefore is to what extent does improved fuel efficiency raise the demand for kilometers driven and, therefore, reduce the fuel savings that would have been realized at constant traffic levels.

This paper argues that the direct rebound effect for cars is not trivial, so ignoring it implies that the expected benefits associated with better fuel efficiency are overestimated. A significant share of the expected reduction in emissions implied by an increase in fuel efficiency leaks away because of additional driving.

The Direct Rebound Effect

The direct rebound effect associated with an improvement in fuel efficiency for cars follows from the observation that better efficiency reduces the per kilometer cost of driving and, therefore, as long as the price elasticity of the demand for driving is nonzero, it raises the demand for driving. Consider a simple economic framework for the choice of car use by households, conditional on owning a car. Assume that the household cares for a general numeraire consumption good x and for kilometers traveled by car. Denote the demand for kilometers by q , and let the car's fuel efficiency (the distance a car can travel per liter of fuel consumed) be

given by E . Assume for simplicity that the fuel cost is the only variable cost. The fuel cost per kilometer driven is then defined as $P = P^f / E$, where P^f is the price per liter of fuel used. Conditional on owning a car, the consumer's short-run problem of choosing the optimal number of kilometers and optimal spending on other goods is:

$$\max_{x,q} (q, x) \text{ s.t. } x + \frac{P^f}{E} q = I \quad (1)$$

where I is the household's income. For simplicity, we ignore the fixed annual cost of car ownership. Assuming that at the optimum $q > 0$, the short-run optimization problem then gives the demand for kilometers as a function of income and the fuel price per kilometer:

$$q(P, I) = q\left(\frac{P^f}{E}, I\right) \quad (2)$$

Note that specification in Eq. (2) implies the implicit assumption that, in absolute value, the effect of a small change in the fuel price and in the car's fuel efficiency on the demand for kilometers is identical. The rebound effect in the early literature was often estimated under this strong assumption. In that case the rebound effect can be estimated simply by the price elasticity of demand for kilometers (q) with respect to the fuel cost per kilometer (P). To see this, note that fuel consumption (denoted F) is the ratio of the demand for kilometers q and fuel efficiency E , so $F = q(P, I)/E$. Differentiating this equation with respect to E , multiplying by fuel efficiency and dividing by fuel use F , we find after simple algebra:

$$\varepsilon_{F,E} = -(1 + \varepsilon_{q,P}) \quad (3)$$

where the $\varepsilon_{i,j}$ refer to elasticities of variable i with respect to j . It follows from Eq. (3) that an increase in fuel efficiency reduces fuel use less than proportionately if the elasticity of demand with respect to the fuel cost per kilometer $\varepsilon_{q,P}$ is negative. Under the assumptions made, the direct rebound effect is simply the absolute value of this elasticity.

The assumption that the effects of fuel price and fuel efficiency on the demand for kilometers are equal in absolute value has been challenged in the empirical literature. There are several reasons why this might be the case. For example, it could be due to consumers being generally less aware of the precise fuel efficiency of the car they drive than about fuel prices, which they observe regularly. Alternatively, it may be that consumers view fuel price changes as temporary while they believe that fuel efficiency improvements are much less likely to be reversed. Unfortunately, the empirical evidence on the relative importance of fuel price and fuel efficiency changes is mixed. Some studies find that the absolute value of the effect of fuel efficiency is substantially larger than that of the fuel price, but there are a number of papers that report the opposite.

Let us allow for different effects of fuel price and fuel efficiency, so that the demand for kilometers is a function $q(P^f, E, I)$ of the fuel price and fuel efficiency separately. Differentiating the definition of fuel consumption $F = q(P^f, E, I)/E$ with respect to E and rearranging, we then find:

$$\varepsilon_{F,E} = -(1 - \varepsilon_{q,E}). \quad (4)$$

The direct rebound effect is therefore defined as the elasticity of the demand for driving with respect to a change in fuel efficiency $\varepsilon_{q,E}$. Note, indeed, that $\varepsilon_{q,E} > 0$.

The rebound effect also has important long-run implications for the effect of fuel prices on fuel use and on the demand for kilometers driven. An increase in fuel prices raises the benefits of driving a more fuel-efficient car and makes it more attractive for producers to develop such cars. This suggests that in the long run fuel efficiency is enhanced by rising fuel prices. Although, conditional on fuel efficiency, higher fuel prices reduce the short-run demand for kilometers, in the long run the fuel efficiency improvements that are the consequence of higher fuel prices generate a rebound effect. They reduce the cost per kilometer and counteract the previous effect. The ultimate long-run effect of the higher fuel prices on the number of kilometers driven is therefore less than what would be expected keeping fuel efficiency constant. To illustrate this, using previous definitions it easily follows that:

$$\varepsilon_{q,P^f} = \varepsilon_{q,P^f} \Big|_E + \varepsilon_{q,E} \varepsilon_{E,P^f},$$

where the notation $\varepsilon|_E$ refers to the relevant elasticity at unchanged fuel efficiency. The overall fuel price effect on demand is the initial effect at unchanged fuel efficiency plus a correction term, reflecting that higher fuel prices raise fuel efficiency that in turn increase demand for kilometers. A similar reasoning implies that the long-run fuel savings of a fuel price increase are smaller than the savings at constant demand for kilometers. One finds, using similar derivations and Eq. (4):

$$\varepsilon_{F,P^f} = \varepsilon_{F,P^f} \Big|_E - \varepsilon_{E,P^f} (1 - \varepsilon_{q,E}).$$

The first term on the right-hand side is the short-run effect at given fuel efficiency. The second term captures the additional effect when people switch to more fuel-efficient cars; this effect is reduced, however, by the rebound effect that induces more kilometers to be driven.

How Large is the Direct Rebound Effect?

Empirical estimates of the rebound effect for cars yield a wide variety of estimates, depending on the country and time period considered and the type of data used (aggregate data vs. observations on individual households), and depending on whether studies take a short- or the long-run perspective on the rebound effect. Moreover, results differ according to the empirical specification of the models estimated and the econometric techniques used.

Estimating the Rebound Effect

The early literature often used aggregate time series for one country, or aggregate panel data for states within a country. The most common way to estimate the rebound effect was to look at the effect of the fuel price per kilometer on the demand for kilometers driven. These studies assumed that in absolute value fuel prices and fuel efficiency affect the demand for kilometers equally. As mentioned before, this assumption may not be appropriate. There is evidence that, although many consumers believe gasoline price shocks to be permanent, there is considerable heterogeneity in these beliefs so that, for many consumers, demand will respond differently to changes in fuel prices and in fuel economy.

A number of studies are now available that estimate the demand for kilometers based on individual data. Many of these studies continue to assume (in absolute value) equal responses of consumers to fuel prices and fuel efficiency. Those that do not assume (in absolute value) equal responses of consumers to fuel prices and fuel efficiency typically make other stringent assumptions. For example, they often implicitly assume that fuel efficiency is not correlated with other characteristics of the car, such as engine power or reliability, that affect the consumer's demand for kilometers driven. Not controlling for these observed and unobserved car attributes leads to biased estimates. Indeed, it has been argued that the vehicle design process suggests that these correlations are nonzero and that, if the correlation were negative, the rebound effect is biased downward toward zero.

Another shortcoming of most empirical studies is that they ignore the interaction between multiple cars in a household. One expects that in multivehicle households the relative demand for kilometers driven in the different vehicles to some extent depends on the price of fuel and on their relative fuel efficiency. To see this, consider a household owning two cars. Conditional on car ownership, the household's short-run optimization problem now implies two demand functions for kilometers, one for each car:

$$\begin{aligned} q_1(P_1^f, P_2^f, E_1, E_2, I) \\ q_2(P_1^f, P_2^f, E_1, E_2, I) \end{aligned} \quad (5)$$

If the cars use different fuels, one fuel price can move independently of the other, and an increase in, for example, P_1^f has an impact on the number of kilometers driven by both cars. The impact of P_1^f on q_2 is caused by substitution toward the car that has become relatively cheaper. If the two cars use the same fuel type ($P_1^f = P_2^f$) then economic theory implies that the total number of kilometers driven by the household should decrease in response to an increase in the fuel price, but the household may use the most fuel-efficient car more intensively. There is evidence that when fuel prices increase households indeed to some extent substitute cars of low fuel efficiency by increasing the use of the high fuel efficiency alternative (De Borger et al. 2016b). However, the majority of papers in the literature estimate the rebound effect by treating each vehicle as an independent observation. This ignores substitution between cars and has implications for the rebound effect estimated.

Finally, with very few exceptions existing studies do not account for the cost of the fuel efficiency improvement itself. The efficiency gain is likely associated with an increase in the car's production cost that is not taken into account. Recently, one or two innovative papers have appeared that derive estimates of the rebound effect from particular policies, for example, policies that give consumers incentives to buy more fuel-efficient cars (see, e.g., Small and Van Dender, 2007). The advantage of this approach is that the characteristics of the vehicles bought can be taken into account and that price effects of the fuel efficiency improvement can be captured.

Empirical Estimates of the Rebound Effect

The transport economic literature offers a fair number of estimates of the direct rebound effect for car transport using both aggregate data and individual household data. Both types of empirical studies are of interest. The studies using aggregate data are useful because of the interest of policy-makers for the system-wide rebound effect. Nevertheless, as the phenomenon originates at the level of the individual actors (households), it is also useful to measure the rebound effect using individual household data.

The most influential studies using aggregate panel data for US states estimate the rebound effect in a system of simultaneous equations (demand for car stock, demand for mileage, and demand for fuel efficiency). They estimate the rebound effect at some 5% in the short run and about 20% in the long run (Small and Van Dender, 2007). That is, 5% of the fuel savings implied by an increase in fuel efficiency under the *ceteris paribus* condition is retaken immediately by a change in driver behavior, and in the course of time this increases to over 20%. The inertia of the adjustment process arises because of the lack of knowledge and due to the time needed to adjust planned travel behavior or to expand or contract the vehicle stock. These studies also find that the rebound effect is declining over time, and this phenomenon is attributed to increases in incomes. The income elasticity of car use is usually small and

decreases with income. Extending earlier results using a larger data set and accounting for the potential importance of congestion in explaining the decline in the rebound effect does not drastically affect the conclusions (Hymel et al., 2010). Moreover, some evidence is reported that the rebound effect may be asymmetric: it is larger in periods of fuel price increases than in years when fuel prices decline; most likely this is because drivers are more aware of the fuel costs when the fuel price rises. Lastly, the rebound effect is also greater during times of media attention for fuel price increases and fuel price volatility, for similar reasons.

Many studies based on individual household data have rejected the equality (in absolute value) of the effects of fuel efficiency and fuel prices on the demand for driving. Unfortunately, there is no unanimity on which effect is larger. Some studies find that households respond more strongly to changes in fuel prices than to fuel efficiency changes, but others find the opposite. All suggest, however, that higher fuel prices induce households to switch toward more fuel-efficient cars.

Several studies using individual data for the United States have found rebound effects toward the high end of those reported using aggregate data. An early study based on household data over the period 1979–94 estimated the rebound effect at 17%–28% (Greene et al., 1999). In a very careful analysis based on individual household data for the United States, several of the restrictive implicit assumptions mentioned earlier have been relaxed; the study allows for multiple cars per household, it allows different effects of fuel price and fuel efficiency, and it captures the correlation of improved fuel efficiency and other car attributes. Estimates for the rebound effect were in the range of 20%–40% (Linn, 2016). A study estimating the rebound effect using information on the effect of a real-world policy to induce people to switch to a more fuel efficient car reports an elasticity of driving with respect to the operating cost per mile of -0.15 .

Remarkably, estimates based on European data show much larger variability between different studies. In some countries much higher values have been reported than for the United States. For Germany, for example, values for the rebound effect of up to 60% have been estimated (Sorell et al., 2009). Using Swiss data and accounting for the endogeneity of distance, fuel intensity, and vehicle weight, even higher values of almost 75% were found. However, based on a first difference model for a very large sample of individual household data for Denmark, the most reliable estimate of the rebound effect is much smaller, amounting to some 7.5%–10% (De Borger et al., 2016a). This study finds that the fuel price sensitivity of the demand for kilometers is declining with household income, but it does not confirm earlier suggestions in the literature that the rebound effect decreases with household income. These earlier observations relied on the imposed restriction that households react identically to changes in the cost per kilometer caused by changes in fuel price and fuel efficiency. The Danish evidence suggests that the sensitivity for changes in the fuel price reduces significantly with income, but the direct rebound effect itself is unrelated to income. Formulated differently, the declining rebound effect reported in the previous literature may have been primarily driven by declining sensitivity with respect to fuel prices, whereas the rebound effect may have been more or less constant over time. Simulation results further suggest that the small pure rebound effect and the impact of adaptations in all car attributes jointly imply that higher fuel prices lead to a substantial reduction in both the demand for kilometers and in demand for fuel.

Finally, the recent literature finds important and significant substitution effects between cars within households owning multiple cars. A fuel price increase leads households to drive more in the car with better fuel efficiency and demand for driving the least fuel efficient car declines (De Borger et al., 2016b). As a consequence, elasticities of kilometer demands with respect to fuel prices that are estimated are substantially smaller than the corresponding fuel price elasticities of fuel use. This suggests that higher fuel prices not only stimulate replacing less fuel-efficient cars by more fuel-efficient cars but also in the short-run substituting vehicle use within the household.

Although it is difficult to separate the effect of individual assumptions on estimated parameters, one suspects that not allowing for multiple cars per household and the possibility of substituting between cars leads to overestimating the rebound effect.

Conclusion

Increasing environmental awareness in transport policy-making, concerns of energy security and increasing fuel prices have generated a revival of interest in the economic implications of fuel prices and fuel efficiency policies on car use. The efficiency of policies to reduce fuel consumption (fuel taxes, fuel efficiency standards, etc.) is mitigated by changes in consumer behavior. Higher fuel prices in the long run lead consumers to switch to more fuel-efficient cars. Moreover, improved fuel efficiency reduces the costs of car use, thereby providing an incentive to increase its use. This is the direct rebound effect: the savings in fuel consumption of a 10% improvement in fuel efficiency are less than 10%, because people drive more kilometers. The effectiveness of fuel efficiency policies is critically depending on the magnitude of the rebound effect.

The best estimates in the relevant empirical literature yield on average a direct rebound effect of about 10%–20%, but there is significant variation, especially among studies using European data. The interpretation is simple: 10%–20% of the fuel savings due to improvements in fuel efficiency leak away through additional driving. Earlier studies using aggregate data suggested that the rebound effect was declining over time due to rising incomes. However, recent empirical findings based on individual household data do not confirm this hypothesis. They suggest that the sensitivity of the demand for driving with respect to changes in the fuel price reduces significantly with income, but the direct rebound effect itself does not decrease with household income. Finally, many earlier measurements of the rebound effect may have to be revised to account for the possible substitution between cars in multiple car households.

Analyses of the determinants of fuel use and of the rebound effect are highly relevant for policy. Driving does not decline as much as would be the case in the absence of the rebound effect, and the additional driving due to better fuel efficiency also contributes to

congestion and traffic accident risks. The effects of policies that differentiate fixed car costs on the basis of fuel efficiency—through taxes on new cars or on car ownership—on kilometers driven and pollution are similarly affected by the rebound effect. In general, the rebound effect reduces the net benefits of fuel efficiency improvements, raising the welfare difference between first- and second-best policies. Measurements of the direct rebound effect for car transport can contribute to the ongoing debate on strengthening fuel efficiency standards and the adoption of new technologies such as car sharing platforms, electric vehicles, and autonomous vehicles.

Given its importance and the wide variation in empirical estimates, the size of the rebound effect remains an ongoing debate. Future work might want to include the supply side of the market for new cars and the role of expectations about gasoline prices in estimating the rebound effect. Moreover, with few exceptions the available studies focus on cars. A careful analysis of the rebound effect and fuel price changes in freight transport based on individual firm data, and embedding the rebound effect in general equilibrium models of the energy sector are useful extensions of the available literature.

References

- De Borger, B., Mulalic, I., Rouwendal, J., 2016a. Measuring the rebound effect with micro data. *J. Environ. Econ. Manag.* 79, 1–17.
- De Borger, B., Mulalic, I., Rouwendal, J., 2016b. Substitution between cars within the household. *Transp. Res. Part A* 85, 135–156.
- Gillingham, K., 2014. Rebound effects. In: Durlauf, S.N., Blume, L.E. (Eds.), *The New Palgrave Dictionary of Economics* (online edition).
- Gillingham, K., Kotchen, M.J., Rapson, D.S., Wagner, G., 2013. The rebound effect is overplayed. *Nature* 493, 475–476.
- Greene, D.L., Kahn, J.R., Gibson, R.C., 1999. Fuel economy rebound effect for US household vehicles. *Energy J* 10 (3), 1–31.
- Hymel, K.M., Small, K.A., Van Dender, K., 2010. Induced demand and rebound effects in road transport. *Transp. Res. B* 44, 1220–1241.
- Linn, J., 2016. The rebound effect for passenger vehicles. *Energy J* 37, 257–288.
- Small, K.A., Van Dender, K., 2007. Fuel efficiency and motor vehicle travel: the declining rebound effect. *Energy J* 28 (1), 25–51.
- Sorell, S., Dimitropoulos, J., Sommerville, M., 2009. Empirical estimates of the direct rebound effect: a review. *Energy Policy* 37 (4), 1356–1371.

Further Reading

- Chan, N.W., Gillingham, K., 2015. The microeconomic theory of the rebound effect and its welfare implications. *J. Assoc. Environ. Resour. Econ.* 2 (1), 133–159.
- Greening, L.A., Greene, D.L., Difiglio, C., 2000. Energy efficiency and consumption—the rebound effect—a survey. *Energy Policy* 28, 389–401.
- Odeck, J., Johansen, K., 2016. Elasticities of fuel and traffic demand and the direct rebound effects: an econometric estimation in the case of Norway. *Transp. Res. A* 83, 1–13.

Elasticities for Travel Demand: Recent Evidence

Fay Dunkerley*, Charlene Rohr*, Mark Wardman*, *RAND Europe, Cambridge, United Kingdom; †Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Why Use Elasticities?	179
What Do We Mean by Elasticities?	179
How are Elasticities Determined?	180
Deducing Demand Elasticities	180
Recent Evidence on Road Traffic Demand Elasticities	181
Evidence on Elasticities of Induced Demand	182
Recent Evidence on Bus Elasticities	182
Evidence on Rail Elasticities	182
Evidence on Diversion Factors	183
Summary	184
References	184

Why Use Elasticities?

Travel demand elasticities indicate the sensitivity of travel demand to changes in relevant variables like price, journey time, and income (Goodwin, 1992). They are vital to public transport operators, allowing them to make judgments about how changes to their services may impact travel demand. The Passenger Demand Forecasting Handbook (PDFH), for example, explicitly contains price elasticities that are extensively used by the railway industry in Great Britain. They also can inform demand forecasting, investment decisions, and policymaking more generally. Elasticity measures generated from observed changes in travel demand are also used as a measure to validate the sensitivity of travel demand models. In fact, the UK Department for Transport explicitly requires travel demand models developed in the United Kingdom to conform to specific elasticity measures in its modeling guidance, saying “the acceptability of the model’s responses is determined by its demand elasticities (DfT, 2019)”.

Elasticities can also provide evidence on the potential size of induced demand effects from road improvements that increase capacity, although their use in transport appraisal is at an early stage of research. Induced demand for road travel can be broadly defined as “the increment in new vehicle traffic that would not have occurred without the improvement of the network capacity” and affects the benefits estimated for a road project as well as on markets from which traffic is displaced (Dunkerley et al, 2018b).

As well as providing information on demand responses, elasticities are used in the calculation of wider economic impacts of transport projects. For example, the elasticity of productivity with respect to effective density, which measures firms’ access to economic mass, and the elasticity of labor supply with respect to wages are used in UK transport appraisal guidance for the calculation of agglomeration effects and labor market effects, respectively.

What Do We Mean by Elasticities?

Elasticity is a general economic concept, which is used to describe the responsiveness of one variable to changes in another variable. In economics and transport, elasticities are often used to determine the demand response to changes in price, income, or other relevant variables. In this chapter we refer to travel demand elasticities only, although, as noted above, a wider range of elasticities are relevant to the transport sector. These demand elasticities are defined as the percentage change in demand for each percentage change in the variable of interest and are hence unitless. Negative elasticities indicate a negative relationship between demand and the variable of interest; positive elasticities a positive relationship. For example, an own-price demand elasticity of -0.5 means that for each percentage increase in the price of traveling by a mode the total demand for that mode will decrease by half a percent (0.5).

An elasticity value of 1.0 (or -1.0) means that changes in the variable of interest causes proportional travel demand changes. Elasticities with magnitudes less than 1.0, referred to as inelastic systems, mean that the travel demand changes are less than proportional to the change in the variable of interest. Elasticities with magnitudes greater than 1.0, referred to as elastic systems, mean that the travel demand changes are larger than proportional to the change in the variable of interest.

Demand elasticities will be influenced by a number of factors in the transport system, including the type of market (or type of travel), travel alternatives that are available, time periods for action, etc. In general, transport elasticities are expected to increase in magnitude over time, as consumers have more opportunities to adjust behavior, for example, to purchase a car, or change their residential or work location.

Cross elasticities measure the sensitivity of demand for one mode relative to changes in relevant variables like price and journey time for “another mode” (Fearnley et al., 2017). For example, the elasticity of demand for rail (patronage) due to changes in petrol

prices. They represent the percentage change in demand for one mode that results from a 1% change in price on a second mode. Cross elasticities depend strongly on local circumstances, including on the availability of alternative modes. For example, if people have access to a number of modes, you might expect much higher cross elasticities.

Linked with cross elasticities are “diversion factors,” which quantify how changes in one mode impact demand for other modes and for new trips. In transport appraisal they are used to determine the source and extent of new traffic resulting from an investment. For example, a new rail line may draw users from a range of other modes, for example, 30% may come from existing bus services, 40% from car, and maybe 30% will not have made the journey before. The percentages reflect the diversion factors (translated to values between 0 and 1). Note that diversion factors say nothing about the volume of traffic that moves to a mode (a new rail line in this case), only the way it has been reallocated from other modes (including not traveling). Diversion factors are also useful because they allow cross-price elasticities, which are often difficult to determine directly from observed data, to be calculated using (own) price elasticities and relative market shares only.

While diversion factors are a useful, practical tool, they have a number of properties that are important to understand when using them. For clarity, in our work we define an “intervention mode,” that is, the mode where the intervention (e.g., fare, capacity or level of service change) is applied and “recipient/source modes,” that is, the modes that are affected by the change in demand for the intervention mode. The “recipient/source modes” terminology reflects the fact that demand may move to or from these modes depending on the change on the intervention mode, although the diversion factor will be the same in both cases. If the change is positive, for example a bus fare is reduced, the diversion factors represent the proportion of traffic switching to bus from other modes. If the change is negative, for example a bus fare increase, the diversion factor represents the proportion of traffic diverted to other modes from bus. This property can be termed equivalence. A second property of diversion factors is that they are generally not symmetric but depend on which mode is subject to a change. The bus–rail diversion factor resulting from a bus fare increase will not be the same as the bus–rail diversion factor if there was a rail fare rise instead. Finally, diversion factors like elasticities are influenced by context and may depend on trip purpose, availability of other modes, area type, etc., as well as calculation methodology.

How are Elasticities Determined?

Elasticity values are usually obtained from regression models using aggregate time-series data of observed travel demand as well as information on key variables explaining travel demand, including income, car ownership, service levels for the mode of interest and competing modes, etc. A range of approaches are used to account for effects such as endogeneity, where changes in an explanatory variable of interest, such as road capacity, may also result from changes in demand, or the lagged nature of a response. These approaches may lead to differences in estimated elasticities that cannot always be easily explained.

It is important to note that elasticities are not necessarily constant, as they may depend on the average values of the variables over which the changes are measured; for example, sensitivity to travel cost may depend on the amount actually paid. The choice of regression model used to calculate the elasticities determines the relationship between demand and the variable of interest. For example, if a log–log model is used for demand and price, the resulting price elasticity of demand will be constant for all levels of demand. A linear relationship on the other hand would mean that the elasticity would depend on both the average price and average demand. This also has implications for how they are used in demand modeling. If composite terms are used, such as generalized cost, which combines variables into an equivalent monetary amount, or generalized journey time (GJT), which conflates various time related attributes such as in-vehicle, walk and wait into a single time term, then even if the elasticity to the composite term is constant the implied elasticities to the constituent variables will not be constant. So, for example, if generalized cost is used, the implied price (time) elasticity will depend upon the proportion that price (time) forms of generalized cost.

Several research methods have been used in the literature to calculate diversion factors: observed changes in behavior based on survey data, reported best alternative to the mode that is used, stated intentions, for example, what choices travelers might make in new or different situations and transfer time and transfer price questions, for example, the price (or time) change required to change behavior. Most surveys reporting observed changes record the impact of an intervention on a particular mode, by asking users of that mode what mode they previously used. An intervention in this case could be new infrastructure or an improvement to an existing service. The other research methods are based on a second approach that involves asking transport users about potential changes in behavior. These methods determine the behavior change that would occur when a mode is no longer available, or the journey cost or time make it unacceptable to a user. This could be for a number of reasons that are related to interventions, such as closure of infrastructure or reduction of services, fare increases, or other policy measures. Although recent work does not find substantial differences in the diversion factor values across these sources, estimates from real transport changes are preferred.

Meta-analysis is a useful method for combining datasets to explain the variation in elasticities across a range of attributes. It has been applied to price and time elasticities in the transport sector.

Deducing Demand Elasticities

There are sometimes instances where it is beneficial to be able to deduce elasticities for a particular variable from available evidence relating to other variables. The most obvious case is where there is, for whatever reason, no credible evidence for a variable, but there might also be a need for supplementary evidence for benchmarking purposes. Good examples of a dearth of own-elasticity evidence are journey time elasticities for bus and car travel and the access time elasticity for rail travel. Evidence on the former tends to

be scarce since bus and car journey times exhibit relatively little variation over time whilst the models used to estimate rail elasticities tend to be based on station-to-station demand and hence do not cover the access element.

If a unit change in generalized cost has the same impact regardless of whether it stems from a change in price or journey time (or any other variable within generalized cost) then a relationship exists between the own elasticities for price and time (and any other variable within generalized cost). So, for example, the journey time elasticity for a mode can be deduced as the product of its price elasticity, its journey time weighted by the value of time, and the inverse of its price. For access time, we would replace the time elements with the level and value of access time.

A related approach where discrete variables are concerned is to estimate demand impacts from known elasticities using valuation estimates. Generally, there is little evidence on the effects of how rolling stock, information provision, station facilities and a range of such “soft” or “secondary” quality factors impact demand. The approach often adopted is to obtain a valuation of the improvement, typically using stated preference methods, and to treat the improvement as if it were an equivalent reduction in the variable against which the improvement has been valued. Therefore if the valuation indicates that the improvement is equivalent to a 3% reduction in fare, the demand effect is inferred from applying a reference fare elasticity to the 3% fare reduction.

Similarly, there is a relationship between relevant cross and own elasticities, which is useful to deduce cross elasticities where there is a lack of relevant evidence or where there is a need for benchmarking evidence. It can be shown that the cross elasticity of demand for mode i with respect to, say, the price on mode j is the product of the absolute price elasticity on mode j , the ratio of demand for mode j and mode i , and the diversion factor between mode j and mode i as defined above. Such a relationship is also useful because it allows cross elasticities to be customized to the particular application circumstances.

Recent Evidence on Road Traffic Demand Elasticities

In 2014 we undertook a rapid evidence assessment to identify literature evidence for road passenger and freight traffic elasticities with respect to key economic and demographic factors, specifically population growth, income growth and changes in fuel costs (Dunkerley et al., 2014). The main objective of the review was to identify the elasticity estimates that were available in published literature with respect to these variables and, where evidence exists, to explore how these elasticity values have changed over time, if indeed they have changed at all. The evidence review not only focused on UK evidence, but also considered international evidence, particularly if it used UK evidence alongside evidence from other countries. It also focused on studies published after 1990.

In a rapid evidence assessment a range of academic databases is searched in a systematic way using reproducible search terms. The approach differs from a full systematic review, as the literature searched can be limited in terms of language, publication data, and geography. In this case the study was limited to English language, OECD papers from 1990 on. The academic search was complemented by approaches to key contacts and web searches to identify grey literature.

We found that the range of estimated “fuel price elasticity” values was quite narrow (-0.1 to -0.5), although a variety of data types, methodologies, and fuel price definitions (prices in pence per liter or pence per km) were used to obtain elasticity estimates. In the long run, improvements in fuel efficiency will result in a smaller reduction in demand for a given price rise—the so-called rebound effect. This should be reflected in a lower fuel price elasticity. Otherwise, fuel price elasticities will be expected to vary by distance, area type and trip purpose. Price elasticities defined in terms of vehicle kilometers will be larger than those for trips because, as fuel prices vary motorists will alter the kilometers per trip in addition to the number of trips. And in turn the elasticity with respect to the amount of fuel purchased will be larger yet as a result of changes in the efficiency of driving in response to price changes.

For road passenger transport, reported “income elasticity” values were predominantly in the range 0.5 – 1.4 . The evidence indicated that car ownership had a strong, positive, and indirect effect on the income elasticity of demand. Some studies, however, do not include this and only report the direct effect of income on demand. In addition, Gross Domestic Product (GDP), household income, and expenditure are all used as proxies for income. These measures may have different impacts on demand (and elasticity measurement) due to the underlying factors they encompass. For freight transport, elasticity estimates of economic activity are mainly in the range 0.5 – 1.5 for an aggregate commodity sector but there the evidence suggests a much greater variation between sectors. Economic activity is also measured by a number of variables: GDP, GFE, and indices of industrial production.

The evidence on how fuel price elasticities of car demand have changed over time was limited. However, fuel price elasticity is expected to increase with fuel price and decrease with increasing real income, and these impacts could explain the changes observed. There is also limited evidence that income elasticities of car demand have decreased over time: two studies found that elasticity with respect to GDP fell after the year 2000. This could be explained by saturation in car ownership levels. For freight transport, the evidence appears to be mixed.

The study highlighted a number of important gaps in the evidence base:

- First, that much of the evidence on car traffic elasticities for the United Kingdom was rather old. This has implications for the use of elasticities in forecasting and strategic planning because of the mixed evidence on changes in these over time.
- We found no information on population elasticities *per se*. However, a few studies did include demographic explanatory variables, such as urban density, population density, employment, and age, from which it is possible to calculate a corresponding elasticity.
- There has been little investigation of the factors that may be responsible for the decoupling of freight transport and economic activity seen in the 2000s. Although the share of van use increased significantly over the same period, there appears to be no evidence on the impact of this on demand for freight transport.

Evidence on Elasticities of Induced Demand

Elasticities of demand with respect to road capacity expansion provide a measure of the induced demand effect that can easily be derived from econometric analysis. In a recent review, most of the evidence was found to be of this type (Dunkerley et al., 2018b). These econometric studies use observed traffic volumes and nonspecific interventions (increases in lane-km or length of road network) rather than actual responses to particular road building projects. There were not only differences in the geographical scope of the studies and the types of road included but also the approach to controlling for background traffic growth and endogeneity (road capacity may also increase as a result of traffic volumes); these differences may explain the wide range of elasticities reported. Short-run estimates range from 0.03 to 0.6; long run estimates from 0.16 to 1.39.

The elasticity evidence is consistent with the expectation that there are more sources of induced demand in the long run when changes in employment, residential location and land use may play a role than in the short run. Of course, it is not clear whether these changes in employment and residential location are transfers from other areas, which may see reductions in travel.

Studies that differentiate between urban and non-urban areas find a larger induced demand effect in urban areas. Urban areas are expected to have high initial levels of congestion and potentially higher levels of suppressed demand. However, only one study analyses the effect of a metro system on road traffic and finds a much smaller induced demand effect in cities with metro systems.

Induced demand elasticities that are close to one are associated with studies that estimate long-run elasticities for specific road types, particularly in large metropolitan areas, outside of the United Kingdom. They also mainly use the same methodological approach. As they focus on particular road types, the demand response reported in these studies generally include reassignment effects and is larger than the induced demand response.

There is no recent econometric evidence on project level investment. Findings for state level road networks in the United States and the national Dutch network indicate an elasticity of around 0.2 across the whole road network, that is, 10% increase in road capacity could lead to 2% induced demand on the network.

Recent Evidence on Bus Elasticities

In 2017 we undertook a rapid evidence review to identify evidence on bus fare and journey time elasticities and diversion factors for all modes (Dunkerley et al., 2018a). We used a systematic search procedure to identify relevant academic and gray literature through structured database searches, as well as making enquiries to experts in the field to identify material, such as unpublished studies. The study focused on material produced in or that was judged to be relevant to the United Kingdom.

Little recent evidence on bus fare elasticities (in the United Kingdom) was found—and little evidence on bus journey time elasticities generally. A key feature of the bus market in Great Britain is concessionary travel, with nationwide free bus travel available since 2008. The most recent work on bus fare elasticities in 2014 is contained in an extensive meta-analysis of UK evidence on fare elasticities (Wardman, 2014). It covered 1633 elasticities estimated between 1968 and 2010, with 377 (23%) of the observations relating to bus travel. A meta-model was estimated to explain variations in elasticities across studies and found short-run elasticities that strongly support the short-run recommendations of Toner et al. (2010), which is the most recent UK work in the area and took account of the impact of concessionary fares. Long-run commuting and leisure elasticities implied by the meta-model very much support existing official guidance.

The most extensive review of time elasticities ever conducted (Wardman, 2012) uncovered only 16 observations of in-vehicle time (IVT) elasticities. A meta-model estimated as part of that study is used to provide estimated IVT elasticities.

Recommendations for bus fare and journey time elasticities are summarized in Table 1.

Evidence on Rail Elasticities

Rail elasticities are used extensively by the railway industry in Great Britain. Rail demand elasticities for journey time, cost, and income are recommended in the PDFH. Elasticities vary by ticket type, journey type and time period. Table 2 shows illustrative rail price elasticities, again derived from the work of Wardman (2014).

Table 1 Recommended bus fare and journey time elasticities in the United Kingdom

Segment	Time period	Bus fare elasticities		Generalized journey time elasticities
		Urban	London and rural	
Commute	Short run	−0.30	−0.40	−1.15
	Long run	−0.65	−0.85	
Leisure	Short run	−0.40	−0.55	−1.05
	Long run	−0.85	−1.10	
Overall	Long run	−0.7 to −0.9		−1.1

Source: Bus fare elasticities from Wardman (2014) and Generalized journey time elasticities from Dunkerley et al. (2018a)

Table 2 Illustrative rail fare elasticities

Ticket type	Time period	Illustrative rail elasticities				
		Short rail trips			Intercity	
		PTE	London	Other	To/from London	NonLondon
Season tickets	Short run	−0.18	−0.19	−0.31	−0.31	−0.28
	LR	−0.36	−0.38	−0.61	−0.61	−0.57
	PDFH–LR	−0.60	−0.50	−0.70	−0.75	−0.90
Non-season						
First	LR	−0.65	−0.69	−1.10	−1.10	−1.02
Full	LR	−0.70	−0.74	−1.19	−1.19	−1.10
Reduced	LR	−0.94	−0.99	−1.58	−1.58	−1.47

Source: Based on the data from Wardman (2014)

Again evidence on journey time elasticities for rail are provided by Wardman (2012), who collated evidence on elasticities for travel time, GJT and service headway. GJT is a concept widely used in the railway industry in the United Kingdom and is composed of station-to-station journey time, headway, and interchange, with the latter two converted into equivalent units of time. The railway industry's PDFH then recommended GJT elasticities ranging from −0.7 to −1.1 across a wide range of flow effects (but did not provide recommendations for journey time or IVT). The Wardman meta-model indicated that long-run GJT elasticities were larger than those recommendations (0.7 to −1.1) and included a modest but sensible distance effect. The PDFH recommendations were revised on the basis of the metaanalysis evidence. Estimated time elasticities were found to be in the order of 60–75% of the GJT elasticities, and encouragingly this is very much in line with the proportion that journey time typically forms of GJT.

Travel time variability is important to travelers, with studies indicating it to be one of the most important factors for rail travelers. In recent years, there has been much more focus on the direct estimation of elasticities to late time rather than relying on deduced elasticities. The evidence from rail demand models indicates a late time elasticity of around −0.07 for commuting flows into London and −0.10 for such flows elsewhere, with noncommuting elasticities around 20% larger. The elasticity for airport flows is, as might be expected, somewhat larger at around −0.25.

Evidence on Diversion Factors

A substantial database of diversion-factor evidence was identified and collated in the Dunkerley et al.'s (2018a) study. Recommendations are provided based on analysis of the available evidence. In general, we find that the evidence on diversion factors is very diverse, covering a wide range of mainly metropolitan geographies, trip purposes, journey types and alternative transport options (Table 3).

Table 3 Recommended diversion factors in the United Kingdom

Intervention mode	Recipient/source mode					
	Bus	Car	Rail	Light rail/metro	Cycle	Walk
Bus		All trip purposes: 0.20–0.35 Commuter: 0.30–0.55	Urban areas: 0.05–0.2 Intercity: 0.45–0.65	Urban areas: 0.05–0.35	Urban areas: 0.04–0.08	Urban areas: 0.1–0.3 Urban areas: 0.10–0.20 Interurban: 0.07–0.11
Car	Urban areas: 0.20–0.40 Interurban: 0.07–0.11		Urban areas: 0.05–0.20 Interurban: 0.55–0.75	Urban areas: 0.10–0.35	Urban areas: <0.1	Urban areas: 0.05–0.15 Urban areas: 0.1–0.25 Interurban: 0.10–0.25
Rail	Urban areas: 0.25–0.4 Interurban: 0.1–0.2	Urban areas: 0.3–0.45 Interurban: 0.4–0.55		Urban areas: 0.05–0.15	Urban areas: <0.1	Urban areas: <0.05 Urban areas: 0.10–0.20 Interurban: 0.10–0.20
Light rail/metro	Urban areas: 0.25–0.4	Urban areas: 0.15–0.3	Urban areas: 0.15–0.3		Urban areas: <0.1	Urban areas: <0.05 Urban areas: 0.10–0.20
Cycle	0.150.05–0.4, higher for walk and bus					

Source: Based on the data from Dunkerley et al. (2018a)

Some may judge the diversion factor for bus interventions and car users to be high (0.30–0.55). However, we note that first the factor for commuters is higher than for other travelers. Second, the factor only describes the proportion that comes from other modes; therefore for the case of bus interventions, the proportion of the new demand for bus that would come from car is in the range 0.30–0.55. The corresponding demand depends on the baseline share of bus, which usually has a smaller mode share, relative to car and can be determined from cross-price elasticities. Third, there is scope for rebound effects. For example, if there is an intervention on bus that attracts car users, then in theory the road should be less congested, which may attract new car trips back from other modes, including bus (see induced travel discussion earlier). We do not believe that the diversion factors take account of second order effects. This would depend on the timescale over which the data were collected.

Summary

This paper has mainly focused on the United Kingdom both in terms of evidence on elasticities and how they are used in transport. Although quite there is quite a lot of evidence, there are gaps in terms of age of the data (road demand elasticities), geographical relevance (induced demand), or lack of data (time elasticities for bus). While the evidence base for rail is quite strong, this may need to be revisited, particularly for commuting, as working practices evolve.

References

- Department for Transport, 2019. Transport Analysis Guidance (TAG) Unit M2: Variable Demand Modelling.
- Dunkerley, F., Rohr, C., Daly, A., 2014. Road traffic demand elasticities: a rapid evidence assessment, RAND Research Report. Available from: http://www.rand.org/pubs/research_reports/RR888.html.
- Dunkerley, F., Wardman, M., Rohr, C., Fearnley, N., 2018a. Bus fare and journey time elasticities and diversion factors for all modes: a rapid evidence assessment, RAND Research Report. Available from: https://www.rand.org/content/dam/rand/pubs/research_reports/RR2300/RR2367/RAND_RR2367.pdf.
- Dunkerley, F., Whittaker, B., Laird, J., 2018b. Latest evidence on induced travel demand: an evidence review, Report for the UK Department for Transport. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/762976/latest-evidence-on-induced-travel-demand-an-evidence-review.pdf.
- Fearnley, N., Flügel, S., Killi, M., Gregersen, F.A., Wardman, M., Caspersen, E., Toner, J.P., 2017. Triggers of urban passenger mode shift—state of the art and model evidence. *Transp. Res. Proc.* 26, 62–80.
- Goodwin, P.B., 1992. A review of new demand elasticities with special reference to short and long run effects of price changes. *JTEP* 155–169.
- Toner, J., Smith, A., Shen, S., Wheat, P., 2010. Review of bus profitability in England—econometric evidence on bus costs—final report. Report by the Institute of Transport Studies, University of Leeds.
- Wardman, M., 2012. Review and meta-analysis of U.K. time elasticities of travel demand. *Transportation* 39 (3), 465–490.
- Wardman, M., 2014. Price elasticities of surface travel demand: a meta-analysis of UK evidence. *JTEP* 48 (3), 367–384.

Parking Price Elasticities

Stephan Lehner, Vienna University of Economics & Business, Vienna, Austria

© 2021 Elsevier Ltd. All rights reserved.

Introduction	185
Theory	185
Price Structures	186
Pricing Versus Supply Management	186
Parking, Mode, and Activity Alternatives	186
On- and Off-Street	187
Commuters	187
Subgroups	187
Data Collection and Calculation of Elasticities	188
Data	188
Calculation	188
Empirical Values	189
References	189
Further Reading	189

Introduction

Understanding how parking fees affect mobility is critical for many stakeholders, including urban planners, policy makers, garage owners, and facility managers. A too low parking price stimulates car traffic. It makes car travel attractive due to its low overall cost. It can cause cars to cruise in the hope of finding a free parking space. The share of cruising cars in total traffic was found to be as high as 74% in city centers (Shoup, 2006). Imperfect parking prices can lead to lost revenue, for the car park owner and for local companies that benefit from visitors. Not the vehicles with a high willingness to pay use the parking spaces, but those that are willing to accept long cruises, wait for a free spot or were simply lucky enough to find a free parking space first.

This chapter of the Encyclopedia discusses: the importance of parking price elasticities for designing parking policies, how strong parking fees affect parking demand, and how parking elasticities can be measured and calculated. "Theory" defines parking elasticities and discusses which factor affect them. "Data collection and calculation of elasticities"

shows how to collect data for, and calculate, parking price elasticities. "Empirical values" states and summarizes empirically measured parking price elasticities.

Theory

Every car park, be it an underground garage or on-street parking, has a certain demand structure. The demand structure of a car park is determined above all by the demand for access to the location and the attractiveness of nearby parking spaces, as well as the attractiveness and availability of other access options (Lehner and Peer, 2019). As a general rule, higher parking fees for a given facility reduce demand. Price elasticities indicate the relationship between prices and demand.

Every car owner has to deal with parking decisions. A parking decision is based on where and when a car is parked. It should be noted that not only parking related factors influence parking decisions. If the reasons for a visit to a particular place decline, the need to park there does so too. In addition, the total travel costs of all access options to the final destination influences the parking demand. A vivid example are rural commuters who use Park & Ride at the city border, instead of garages in the city center, to avoid traffic congestion.

Parking price elasticities express the responsiveness of the quantity demanded for parking and the price of that same good. The elasticities are often studied within a specific setting, defined by a shared location, price point and regulation. Regulation objectives, such as parking availability or profit maximization, vary between different settings. Managers often aim to regulate one of the following indicators, while keeping the others in an acceptable range. Parking **volume** is the total number of vehicles using a certain parking facility within a defined period. The point of reference for a parking activity can be the pull-in or pull-out event. Parking **dwelt time** presents how long cars occupy a parking space. **Occupancy** is the ratio between consumed parking and supplied parking. It displays how many parking spaces are in use (or vacant) at a specific point in time or over a period. Occupancy cannot exceed 100% and cannot be lower than 0%. Occupancy is a function of volume and dwell time. Under controlled conditions, the price elasticity of occupancy equals the price elasticity of parking volume and dwell time (Lehner and Peer, 2019). Parking **revenue** is the total income generated by a certain facility within a certain period. Depending on the price structure, revenue is a function of volume and dwell time.

Supply restrictions strongly affect parking elasticities. The costs of parking not only consist of parking fees, but also indirect costs, such as access (additional ways to reach the parking destination), search or queuing and egress (way from the parking to the final destination) time (Axhausen and Polak, 1991). As the direct cost stays constant, the indirect costs grow in relation at high occupancy levels. Parking demand cannot exceed parking supply, that is, parking occupancy cannot be higher than 100%. If parking is fully saturated, cars have to search or wait for a free spot, leading to an increase of indirect costs. In the case of positive indirect costs before or after a price change, parking demand will react differently to a price change than in the case with no indirect costs. For instance, if parking is fully saturated and parking fees increase, **latent demand** may consume freed up parking spaces (Milosavljevic and Simićević, 2014; Nourinejad and Roorda, 2017). Latent demand classifies those drivers who did not, or could not, park due to the indirect costs caused by saturated parking.

Price Structures

Drivers may pay parking fees per permit, event or based on dwell time. A permit is the right to park at a specific facility. In the case of pay per event, drivers pay a flat parking fee when they enter or exit a facility. Dwell time dependent fees increase with the parking duration. Those fees shorten the parkers dwell time, implying more short-term parking. Those fees hence have a similar effect as short-term parking zones. It is sometimes argued that parking fees should not be dwell time dependent. As, dwell time dependent parking fees can even induce traffic, in the case of high demand. This occurs as a reduction in dwell time allows more cars to park within the same period, and because a lower occupancy increases the attractiveness of parking. This is why parking policies should always be designed in line with road pricing, as long as traffic regulation is an objective. In the absence of road pricing, parking fees with an event and a dwell time component can be used.

The price structure has a strong impact on the price elasticity. The more dwell time dependent parking fees are, the stronger drivers can react to price increases by shortening their dwell time. It is easier to save fees by shortening ones stay, if you pay parking fees per minute instead of per hour. The more a parking fee varies per time of day, the easier it is for a driver to respond to increases in charges by changing her parking times. The same is true for spatially differentiated parking fees.

Pricing Versus Supply Management

Parking space management regulates the demand of car parks in the short-run. In fact, parking space management plays a central role in urban traffic management. City administrators not only use it to control parking availability, but also traffic. Governments often see parking management as an alternative to road tolls (Glazer and Niskanen, 1992; Verhoef et al., 1995). Modern parking space management uses two instruments to regulate demand: prices and supply. The first means setting the level and structure of parking fees. The second includes restrictions on access to parking spaces at certain times of the day or during weekdays, limitations of parking dwell time, and restrictions on certain groups of drivers.

Pricing works fundamentally different to supply management. To understand the difference, we need to understand the cost components of using a car park, which are:

$$C = C_p + C_Q + C_O$$

Where C is the total cost of the car park, C_p are the driver's total parking fees, C_Q are all cost linked to getting a vacant parking spot (e.g., waiting time, discomfort of not finding a spot, costs of uncertainty, fuel consumed while cruising), and C_O are all other cost (e.g., access time to the parking spot, discomfort of parking in the sun). A car park is more attractive the lower C is, that is, when drivers can easily find a vacant spot for a low price.

Pricing directly affects C_p , whereas supply management directly affects C_Q . Higher parking prices increase the generalized cost of the car park, which leads to fewer people parking there (Lehner and Peer, 2019). Lower supply of parking space effectively increases the occupancy of the car park, which leads to fewer people finding a parking space, increases latent demand and the costs of getting a vacant parking spot, and hence the generalized cost of the car park. Indirectly pricing can also affect C_Q . This is as lower or higher occupancy affects the cost of getting a vacant parking spot. Supply management has two central flaws pricing can overcome. First, prices are more transparent to drivers, whereas drivers usually cannot tell how long they have to wait for a vacant parking space before they entered the car park. Second, pricing allocates parking according to the willingness to pay. Hence, in contrast to supply management, parking is allocated to best serve each individual, while minimizing waste and inefficiency. The second argument is also true for limitations of parking dwell time and restrictions on certain groups.

If parking is saturated, parking volume may increase or decrease with rising parking prices (Nourinejad and Roorda, 2017). Decreasing indirect costs, such as lower queuing and searching times, may cause an increase in volume. In extreme cases, occupancy may even stay unchanged. Drivers who are willing to shorten their dwell time are more likely to continue parking after a rise in the hourly parking fee. If parkers are not willing to shorten their dwell time, parking alternatives affect drivers' reaction to price changes.

Parking, Mode, and Activity Alternatives

Drivers may react to parking price changes by changing the parking location or the activity (Feeney, 1989). Changes of the driver's parking location may imply a change of the transportation mode. The elasticity of parking demand is greatly affected by the

availability and quality of alternative parking facilities, access time to those facilities, and transportation modes from those facilities to the final destination. Substitutes affecting the underlying activity associated with the parking event, include changing the parking dwell time (and hence the activities' duration), changing the parking (pull-in) time (and hence the timing of the activity), changing the location of the activity, or abandoning the activity altogether. Even though empirical data show that parking elasticity is mainly affected by parking and mode alternatives, substitutes affecting the underlying activity cannot be ignored. For instance, it is a common argument that parking prices in city centers should not be increased, as this would steer shoppers towards the suburbs. Additionally, parking fees can also be used to shift peak traffic, similar to temporal road tolls, as recent research shows (Fosgerau and De Palma, 2013; Lehner et al., 2019).

On- and Off-Street

When do drivers decide to cruise for an on-street lot and when do they decide to pay for an off-street lot? Which factors influence that decision? Shoup (2006) investigated those questions.

In his paper, he assumed that two parking options, one off-street (with high prices and infinite supply) and one on-street (with low prices and limited supply), compete for parking demand. A driver will prefer on-street parking if:

$$C^*(f + nv) < t(m - p)$$

where C^* is the expected time spent cruising for parking, f are the fuel costs per time unit of cruising for parking, n are the number of people in the car, v is the value of time spent cruising, t is the parking duration, m is the off-street price and p is the on-street price.

The expected time spent cruising for a free lot is a function of occupancy, the number of cars pulling-out while cruising and the amount of other vehicles cruising. The exact form of the function depends on the specifics of the car park. The time needed strongly increases with occupancy. The cruising time increases more, when occupancy increases from 90% to 100% than when it increases from 60% to 70%. As a rule of thumb, on-street parking occupancy should be lower than 85% per street block, to guarantee a vacant parking spot for drivers and to avoid cruising (Shoup, 2006).

Assuming that fuel costs, value of time saving, the number of people in the car, parking supply and demand for accessing the location by car are inelastic and constant, the demand for the two parking options solely depends on the prices of the on-street and off-street parking options. When on-street parking is cheap and off-street parking is expensive, drivers will be willing to cruise longer, leading to an increase in on-street occupancy. Drivers with a relatively low value for time savings, low fuel costs, a long parking duration and who are alone in their car, prefer on-street parking, while the others prefer off-street parking. It should be noted, that when the price gap between on-street and off-street parking is large, the demand for on-street parking may be completely inelastic, until its price gets close or exceeds that of off-street parking. At this point, it can suddenly become very elastic.

Commuters

Commuters owning a car and traveling from home to work primarily face two decisions: where to park and when to depart (Lehner et al., 2019). On the one hand, commuters can park at work. This implies that the way from home to work is driven by car. On the other hand, commuters may park at home. This implies that they travel the total distance with an alternative mode of transportation. In all other cases, they park in facilities between their home and work, such as Park & Ride facilities. They also have to decide when to depart. Commuters try to avoid congestion and arriving either too early or late at their workplace, or leaving too early or late from home. Similarly, when they commute home, they try to avoid congestion and to appropriately time their departure from work or arrival at their next activity (Arnott et al., 1990; Zhang et al., 2008). When commuter parking is in high demand and fills-up in the morning, commuters may try to leave earlier from home to find a vacant spot and reduce their search time (Arnott et al., 1991). Parking fees reduce demand and hence ensure sufficient vacant parking spots. Thus, parking fees also impact the departure time of commuters.

Time-varying parking fees, that is, parking fees that vary by the time of day or are only charged within a certain period of the day, additionally affect the departure decision of commuters. This occurs as commuters may change their parking times in a way that avoids or reduces parking costs (Fosgerau and De Palma, 2013). Similarly, spatially differentiated parking fees, that is, parking fees that are higher at certain locations, can incentivize commuters to park outside of a specific zone. For instance, high parking prices within the city center provide incentives for commuters to park in Park & Ride facilities, outside of the city center (Qian et al., 2012).

If commuters do not have to drive to work, higher parking prices at the workplace can also lead to an increase in the number of employees working remotely. This assumes that working at home reduces commuters' parking fees. Commuters can save on parking cost, when they are paid per commute and not per month or year. Then the price structure allows for short-term savings. As a bottom line, traffic managers can use parking fees as an instrument to tackle congestion.

Subgroups

Sometimes articles report elasticities for subgroups of drivers. As with all elasticities, these subgroups have to be interpreted with care. Elasticities for each subgroup (such as commuters, shoppers and business activities) strongly correlate with the elasticities of each of the other subgroups, in particular when parking is saturated. For instance, consider a nearly fully saturated parking garage that

increases their hourly parking price from 1.00 EUR/hr to 2.00 EUR/hr. Before the price increase, commuters mainly used the garage. Due to the higher price, the parking volume of commuters decreases by 50%. The price elasticity for the subgroup commuters is -1 . As parking availability during business hours has now strongly increased, 100% more shoppers use the garage. Some might argue that the price elasticity of shoppers is positive. The increase in shoppers, however, is not due to the higher price, but due to the increase in parking availability. Similarly, it is possible that the higher parking availability attracts new commuters. To calculate reliable elasticities, the effect of parking availability on demand has to be disentangled from the effect of the parking price on demand.

Data Collection and Calculation of Elasticities

Data

Data on parking demand can be collected by observing the parking space or the driver. A sensor installed at the entrance of a car park or at each parking lot, can detect how many spots are occupied. Sales data are a good proxy of how many parking hours drivers consume. However, often the amount of paid parking hours strongly diverges from the amount of consumed parking hours, as drivers under or overpay for parking. If no sensors are installed, the number of occupied lots can also be counted, whether manually or with technical equipment. Questionnaires can reveal not only how drivers behaved, but also how they intend to behave. However, the stated behavior of a responded can deviate greatly from the actual behavior. Data from devices connected to the internet, can measure where and how long drivers park. Since the number of connected devices is increasing rapidly at the time of writing, this data source might well be of greater relevance in the coming years. Each data source has their advantages and disadvantages. Generally, sensors give detailed information about parking volume, dwell time and occupancy within a parking facility. Data based on counts, reveal the occupancy level of the investigated parking facility, however, they often provide less accurate or no information on volume and dwell time. In contrast to parking space based data, questionnaires give insights into why and which drivers park at a certain location. Data based on connected devices, can also provide this information. Similarly, to questionnaires, data based on connected devices might also reveal the alternative behavior chosen, when drivers decide to stop parking altogether. When predictions are made on data based on drivers, the underlying distribution of the population should be known. This knowledge allows for correcting of sample biases.

Calculation

There are many possibilities on how to calculate price elasticities. The applied method depends on the underlying data. Revealed preference data is data stemming from an actual change in parking prices. For instance, in Vienna the hourly parking fees in the most central district increased from 1.20 to 2.00 EUR/hr. Urban planners may be interested in how the increase affected parking occupancy at noon. Hence, the occupancy levels for multiple days before and after the price increase are collected. The occupancy level of on-street parking at noon is then calculated by dividing all parked cars by the total parking supply. The occupancy will probably fluctuate between days. The fluctuation decreases with the number of parking lots under consideration. In addition, parking occupancy usually fluctuates strongly between weekdays, months and time of day. Furthermore, other external factors may affect the parking occupancy in the district. Assuming that the occupancy is not fluctuating between days and no external factors affect the occupancy, the elasticity can be calculated with the arc-log elasticity equation:

$$\varepsilon = \frac{\log \frac{Q_1}{Q_0}}{\log \frac{P_1}{P_0}}$$

where ε is the elasticity, Q_0 the old demand (e.g., occupancy, parking volume), Q_1 the new demand, P_0 the old price, and P_1 the new price. The log transformation accounts for the non-linearity of the demand curve. Furthermore, the elasticity is calculated for the average percentage change in quantity and price. Hence, the elasticity is the same for the concerned observation regardless of whether the price increases or decreases. This is because the formula uses the same base for both cases. In the point elasticity the old or the new price is used as the base. A price elasticity of parking demand of -0.5 , thus indicates that a price increase of 67% leads to a drop of occupancy of 23%.

Most papers in transportation economics use econometric models. In econometric models with a continuous dependent variable, the price elasticity is the estimated regression coefficient β of the price variable. In econometric models with discrete dependent variables (i.e., binary choice models), the elasticity can be calculated with the individual or aggregate elasticity formula. The individual formula is:

$$\varepsilon = \beta \times (1 - q) \times p$$

where, q is the share of the demand (regression estimate) at the sample means and p is the sample mean of the parking price. It should be noted, that binary choice models explain the behavior of individuals, while policy makers are usually concerned with the behavior of the population of interest. The individual formula calculates the elasticity that predicts how an individual, with the characteristics of the sample mean, reacts to a price change. However, if the behavior of the broader population is of interest, the

aggregated elasticity should be used (Westin, 1974). The aggregated (or weighted) elasticity, first computes the specific distribution of characteristics over all individuals, before the aggregation takes place. Aggregate predictions take into account the frequency of each individual in the population. In most applications, the price elasticity calculated with the individual formula will be higher (in absolute terms) than the price elasticity calculated with the aggregate formula. This occurs as the majority of the driver's decisions lie on the concave part of the logic function in choice models. It should be noted, that the difference between an elasticity calculated with the individual formula and the aggregated formula, could be substantial.

Empirical Values

The majority of empirically measured parking elasticities fall in the range of zero and minus one. They are negative and inelastic. This generally implies that a rise in parking prices leads to a lower parking demand and an increase in parking revenue.

Lehner and Peer (2019) state ranges of price elasticities for commuting and non-commuting trips, based on 50 parking studies. For non-commuting trips the baseline price elasticity of occupancy is -0.63 , which is the sum of the price elasticity of dwell time of -0.30 and the price elasticity of volume of -0.32 . For commuting trips, the baseline price elasticity of occupancy of -0.32 equals the price elasticity of volume. Commuters probably do not shorten their dwell time to reduce parking costs.

The empirical results show that commuters react differently to increasing parking prices than non-commuters. In the absence of parking alternatives, non-commuters mainly shorten their parking dwell times and do not switch to other transportation modes, whereas commuters stop to park at the concerned facility (Lehner and Peer, 2019). Multiple factors might explain this difference. First, commuters cannot shorten their parking as easily as non-commuters, as the underlying activity – working – is more rigid. Second, when drivers can and are willing to shorten their parking dwell time, they are less likely to give up parking altogether. Third, as commuting is a recurrent activity, drivers are more knowledgeable about alternative transportation modes. Finally, the impact of parking fees is stronger for commuters, whereas it is only marginal for non-commuters. On average, commuters pay 13% of their daily gross income for parking, when parking is being charged. The parking price elasticity of commuters is around -0.5 based on revealed preference data and -1.1 based on stated-preference data.

Elasticities based on stated preference data (data collected in hypothetical choice experiments) are systematically lower than elasticities based on revealed preference data (data collected from real price changes). This might be due to two reasons. First, stated preference experiments suffer by hypothetical bias. Participants can indicate that they would change their behavior, but when confronted with the actual situation, they do not. Second, stated preference experiments do not account for supply restrictions or for indirect costs. Supply restriction, however, may affect parking elasticity to a substantial degree (Lehner and Peer, 2019).

References

- Arnott, R., De Palma, A., Lindsey, R., 1990. Economics of a bottleneck. *J. Urban Econ.* 27 (1), 111–130.
- Arnott, R., De Palma, A., Lindsey, R., 1991. A temporal and spatial equilibrium analysis of commuter parking. *J. Public Econ.* 45 (3), 301–335.
- Axhausen, K.W., Polak, J.W., 1991. Choice of parking: stated preference approach. *Transportation* 18 (1), 59–81.
- Feeney, B.P., 1989. A review of the impact of parking policy measures on travel demand. *Transp. Plan. Technol.* 13 (4), 229–244.
- Fosgerau, M., De Palma, A., 2013. The dynamics of urban traffic congestion and the price of parking. *J. Public Econ.* 105, 106–115.
- Glazer, A., Niskanen, E., 1992. Parking fees and congestion. *Reg. Sci. Urban Econ.* 22 (1), 123–132.
- Lehner, S., Peer, S., 2019. The price elasticity of parking: a meta-analysis. *Transp. Res.* 121, 177–191.
- Lehner, S., Peer, S., Gren, M., Koller, H., Dragaschnig, M., Brändle, N., Sengupta, R., 2019. Innovative pricing policies for urban traffic: a field experiment. In: Goulias, K., Davis, A. (Eds.), *Mapping the Travel Behavior Genome*. first ed. Elsevier.
- Milosavljević, N., Simićević, J., 2014. Revealed preference off-street parking price elasticity. *Proceedings of Fifth Annual Meeting of the Transportation Research Arena*, pp. 1–10.
- Nourinejad, M., Roorda, M.J., 2017. Impact of hourly parking pricing on travel demand. *Transp. Res.* 98, 28–45.
- Qian, Z.S., Xiao, F.E., Zhang, H.M., 2012. Managing morning commute traffic with parking. *Transp. Res.* 46 (7), 894–916.
- Shoup, D.C., 2006. Cruising for parking. *Transp. Policy* 13 (6), 479–486.
- Verhoef, E., Nijkamp, P., Rietveld, P., 1995. The economics of regulatory parking policies: the (im)possibilities of parking policies in traffic regulation. *Transp. Res.* 29 (2), 141–156.
- Westin, R.B., 1974. Predictions from binary choice models. *J. Econ.* 2 (1), 1–16.
- Zhang, X., Huang, H.J., Zhang, H.M., 2008. Integrated daily commuting patterns and optimal road tolls and parking fees in a linear city. *Transp. Res.* 42 (1), 38–56.

Further Reading

- Inci, E., 2015. A review of the economics of parking. *Econ. Transp.* 4 (1–2), 50–63.
- Millard-Ball, A., Weinberger, R.R., Hampshire, R.C., 2014. Is the curb 80% full or 20% empty? Assessing the impacts of San Francisco's parking pricing experiment. *Transp. Res. Part A Policy Pract.* 63, 76–92.

The Pareto Criterion and the Kaldor Hicks Criterion

Harald Minken, Institute of Transport Economics, Oslo, Norway

© 2021 Elsevier Ltd. All rights reserved.

The Pareto Criterion and the Kaldor–Hicks Criterion	190
Definitions	190
Definitions Concerning the Pareto Criterion	190
Definitions Concerning the Kaldor–Hicks Criterion	190
The Pareto Criterion	191
History	191
Transport Applications of the Pareto Criterion	191
The Kaldor–Hicks Criterion	192
Theoretical Problems	192
Quasi-linear Utility: Problem Solved, But Reappearing	193
The Kaldor–Hicks Criterion and Standard Transport Appraisal	193
The Kaldor–Hicks Criterion and Random Utility Models	193
Alternative Approaches	193
Aggregate Willingness-to-Pay	193
Voting Theory	194
Conclusion	194
See Also	194
References	194
Further Reading	194

The Pareto Criterion and the Kaldor–Hicks Criterion

Definitions

These criteria will apply to a society (usually taken to be a country) consisting of a given number of individuals. Different policies or options are open to the society. Given the policy chosen, a given amount of goods and services will be available at the aggregate level. Either by the action of the individuals themselves or government, or by the interaction of both, each individual will get a certain share of it. An *allocation* is a distribution of the total amount of goods and services to each of the individual members of society. Each policy results in an allocation.

A *reallocation* is a measure taken to change the allocation after a policy has been implemented.

This highly abstract view of society abstracts from time (time only exists in the form of before and after a policy is implemented), and it says nothing about the mechanisms that bring about a certain allocation, be it production, trade, or the individuals' own choices of how much to work, what to consume and how to spend their time in general. In particular, utility maximization and profit maximization have not been assumed.

The individuals are, however, thought to be able in all circumstances to tell if a bundle of goods and services is better, worse, or just as good as another. That is, they have a *complete and transitive preference ordering* over such bundles.

Definitions Concerning the Pareto Criterion

The Pareto criterion provides the following sufficient condition for an allocation to be better than another:

- An allocation is better than another (constitutes an improvement from the base case allocation) if at least one member of society is better off and nobody is worse off.

If this is the case, the allocation is called a *Pareto improvement* and is deemed worth carrying out. Note that if only one single person is worse off, the allocation is not a Pareto improvement, even if all others are overjoyed by it. In such situations — certainly the majority of cases — the Pareto criterion cannot decide whether or not a policy should be implemented.

When no Pareto improvements are possible, that is, nobody can be made better off without making someone else worse off, we have a *Pareto maximum*.

Definitions Concerning the Kaldor–Hicks Criterion

A *potential Pareto improvement* is a situation where those that are made better off by the policy measure, can in principle compensate those who are made worse off (so that they are at least as well off as before), and still be better off themselves. Put otherwise, in the situation after the policy has been implemented, there is at least one reallocation that can make at least one individual better off, while making no one worse off than before the policy was implemented.

The Kaldor–Hicks criterion simply says:

- A policy which brings about a potential Pareto improvement is worth carrying out.

We note that the concept of economic efficiency as defined and computed in transport cost benefit analysis, is essentially the same as to satisfy the Kaldor–Hicks criterion.

The Pareto Criterion

History

To the early utilitarianists, like Jeremy Bentham and John Stuart Mills, it seemed obvious that the best action was the action that produces the greatest well-being to the greatest number of people, and that somehow, the level of attainment of this goal was measurable.

The Italian economist and sociologist Vilfredo Pareto (1848–1923) thought otherwise. To him, the utility of one individual and that of another could never legitimately be added or compared. Every person is indeed able to compare, rank, and choose among different outcomes for herself, but interpersonal comparisons make no sense.

So how can we pass judgment on policies that affect more than one person? Pareto came up with a criterion that does not require more than that, each affected individual is able to rank any set of policies presented to her (her preferences induces a complete and transitive ordering on the set of policy options), and that she prefers more rather than less of any good. This is the Pareto criterion, and it may be shown that it is indeed the strongest possible criterion that does not involve interpersonal comparisons.

Utility may be measured on an ordinal or cardinal scale. Measurements are on an ordinal scale if they are used for ranking purposes only. Cardinal measures, on the other hand, are based on an underlying scale which makes it meaningful to do computations. Your utility is twice the size of mine, the difference between yours and mine is 23, etc.

Pareto's concept of utility was ordinal—utility differences exists, but the level of utility, the size of utility differences, and the concept of marginal utility have no intrinsic meaning. Consequently, interpersonal comparisons are also meaningless. Other important economists, such as [Hicks \(1939\)](#), [Samuelson \(1956\)](#), and [Sen \(2018\)](#) came to exactly the opposite conclusion: Even if a cardinal and measurable concept of utility is not needed to derive the demand functions of individuals, interpersonal comparisons not only make sense, but they are also necessary for policy purposes. If a cardinal utility concept is needed for such purposes, so be it.

Both approaches have lived on, but it is fair to say that the cardinal approach has lost much ground in academic circles, while the ordinal approach has become more and more formalized. Perhaps to the point where, as K. J. Lancaster once remarked, it now stands as an example of how to extract a minimum of results from a minimum of assumptions. In the applied fields of economics, however, and perhaps more than anywhere else in transport economics, the cardinal approach has spread and developed over the last half century in the form of cost benefit analysis.

Transport Applications of the Pareto Criterion

The Pareto principle is seldom applied in transport economics. The main reason is that major transport infrastructure projects are nearly always financed by taxes. Among the taxpayers, there will usually be many that will make no use of the new infrastructure, and so there is a large group of losers that cannot possibly be compensated without undermining the financial basis for the whole project. The possibility of applying the Pareto principle might be a little better if we consider a more comprehensive urban or national transport plan, consisting not only of road projects, but also projects concerning public transport, walking, and cycling. But even then, there will be many people who stand to lose.

Research on the political processes that produces such transport plans have shown that there is a tendency to distribute projects among districts in such a way that those left out of the current plan will be first in line for projects in the next plan. In the end, every taxpayer will get something back, it is thought (and sometimes even openly said or formally agreed). Such processes are often accused of wasting taxpayers' money by including a lot of economically inefficient projects. Could it possibly be, however, that in the best of cases, they do at least produce a Pareto improvement? If the best plan according to a cost-benefit analysis (the Kaldor–Hicks criterion) is infeasible because the losers understand that compensation could not possibly be paid (or at least will not actually be paid) there might still be some room for rational bargaining for a Pareto improvement.

If we turn from infrastructure plans to pricing and regulation, the potential for actual Pareto improvements increases considerably. More efficient ways of regulating traffic, providing information to users, and utilizing existing infrastructure and rolling stock may cost little and (in the best of cases) impose no extra inconveniences on anybody. Manuals with ideas for such "small projects" exist.

Congestion charging and other forms of marginal cost pricing split the users into three groups: The first group opts out, and will incur a loss, the size of which depends on the best remaining alternative. The second group also stands to lose but finds that paying the price and staying in is a lesser evil than the other remaining alternatives. Members of the third group are the only ones who find themselves better off in the new situation, for instance because they have a very high willingness to pay for time savings on the road, for getting a seat in the bus, etc.

A fourth group — the second winning group — is the public sector, who gets the revenue from the charges. It is almost a matter of definition of marginal cost pricing that this last group could reimburse the three others and still have something left. (It will probably not be able to save some money by identifying the members of the third group.) But doing so would be meaningless: If all travelers understand that they will get their money back regardless, they will not change their travel habits. The potential Pareto improvement in this case actually could not be made a reality, at least not in the straightforward way by reimbursing the tolls. The compensation

must either be made in kind, for instance by improving public transport, or paid out with an equal amount to all inhabitants in the district. Either way, there will still be losers, but not so many, and not by so much.

We see that the Pareto principle has a role to play in transport as a complement to the Kaldor–Hicks criterion. There is a moral reason for this: We feel obliged to reduce the number of losers from a policy and the severity of their loss. Potentiality becomes a hollow phrase if it cannot ever be transformed to reality. And there is a practical side: All our congestion charging plans tend to be turned down if we do not pay attention to compensating the losers.

The Kaldor–Hicks Criterion

Theoretical Problems

To apply the Kaldor–Hicks criterion, we must be able to express the losses of some and the gains of others on a common scale, preferably money. According to microeconomic consumer theory dating back to [John Hicks \(1939\)](#), this can be done in two different ways: the equivalent variation (EV), and the compensating variation (CV). The first is the answer to the question of how much money one would be willing to pay to get the benefits of a proposed policy, and the second answers the question of how much one must have to forego these benefits.

There is also a third way, commonly associated with Alfred Marshall, but actually dating back to the French engineer [Jacques Dupuit \(1844\)](#). This is the consumer surplus (CS), consisting of the area under the ordinary demand curve for the good that the policy provides. This measure lies in between the two Hicksian measures, both of which coincide with the CS under special circumstances.

It turns out that there are theoretical problems with each of these. For instance, it could be shown that using EV or CV, it might sometimes happen that starting from the “do nothing” situation, doing “something” is judged to be an improvement, while starting from the situation where this “something” has already been implemented, going back to the original situation is also judged to be an improvement (Scitovsky reversals). The CS, on the other hand, lacks the straightforward interpretation and general applicability that EV and CV have.

In the fifties, it was shown by W. F. [Gorman \(1953\)](#) that for consistent interpersonal comparisons to be made, and consequently for computing the aggregate losses of the losers and the aggregate gains of the winners and comparing them with one another, the utility functions of all individuals must be of a particular mathematical form, the so-called Gorman Polar Form.

Let $\mathbf{p}$ be the vector of prices of the goods and services that concern us here, and let R_h be the income of individual h . The indirect utility function $V_h(\mathbf{p}, R_h)$ of individual h is of the Gorman Polar Form if and only if it takes the form

$$V_h(\mathbf{p}, R_h) = a_h(\mathbf{p}) + b(\mathbf{p})R_h$$

The function $a_h(\mathbf{p})$ is specific for individual h , while the function $b(\mathbf{p})$ is the same for all individuals. Summing over individuals, we get the (utilitarian) welfare function:

$$V(\mathbf{p}, R) = \sum_h a_h(\mathbf{p}) + b(\mathbf{p})R$$

where R is the sum of all individual incomes. It is seen that this welfare function has the same form as the individual indirect utilities. That the utility functions of all are of the Gorman Polar Form is not only the condition for consistent interpersonal comparisons to be made, it is also the condition for there to be a representative consumer, that is, for the aggregate decisions of all consumers to correspond to consumer theory applied at the aggregate level.

It is evident from the formula that welfare, as measured in this way, stays the same regardless of changes in the income distribution. Another property of this function is that even if the individuals differ in their income and tastes, they are all equal in the sense that if they get a marginal increase in income, they will all spend it in the same way. (The derivative of the demand functions with respect to income are the same for all individuals.)

It has been shown by [Chipman and Moore \(1979, 1994\)](#) that to apply the Kaldor–Hicks criterion in a situation where reversals cannot exist and to apply the consumer surplus (CS) in the situation where a representative consumer exists is basically one and the same thing. Chipman and Moore called the Kaldor–Hicks criterion restricted to situations without reversals of any kind for the Kaldor–Hicks–Samuelson criterion (KHS-criterion for short).

Where does that leave us?

It might seem that the conditions for applying the Kaldor–Hicks criterion in a consistent way are so special that they almost never will be fulfilled. But given the models that we have, this is not necessarily true. Take general equilibrium models as an example — they almost always apply representative consumers, and even if there are more than one of them, they may be sufficiently similar to be compared.

Partial equilibrium analyses are often even more close to the conditions for the Kaldor–Hicks criterion to apply. Because we only model a small part of total consumption, we abstract from the prices of most goods and services, and often from the income effect in the markets of interest. That is, demand in these markets is not supposed to be a function of income. This is really to apply the Gorman Polar Form.

It may be objected that even if the necessary conditions apply to the model, they certainly do not apply to reality. But if we find the models useful and enlightening, there is no intrinsic reason why their consumer surpluses cannot be useful and enlightening too. They will never be entirely accurate, but they may at least be good indications of the true consumer preferences.

Quasi-linear Utility: Problem Solved, But Reappearing

Setting $b(\mathbf{p}) = 1$ in the equations above, we get the quasi-linear indirect utility function. The direct utility function in this case is $U = z + u(\mathbf{x})$, where z is some composite good and $\mathbf{x}$ is a vector of the goods and services we want to model in more detail. This model is of the Gorman Polar Form, a representative consumer exists, the indirect utilities of all individuals can be added, and the Kaldor–Hicks criterion can legitimately be applied. Furthermore, the model divides the economy nicely in a sector of interest where a diverse set of goods $\mathbf{x}$ belong (the transport sector, for instance), and a single composite good z representing everything else.

But this is only as long as we maximize U subject to one single constraint – the budget constraint. But often, we want to apply constraints on total available time as well (the time budget), and maybe on the various forms of time use. The minimum number on working hours, for instance. The [de Serpa \(1971\)](#) model of the allocation and valuation of time, that has formed the basis for most studies on the valuation of travel time savings in transport, is a case in point. The two constraints are reduced to one if we define total income not as R , as usual, but as $R + wt_w$, where w is the hourly wage and t_w is the number of working hours. Since t_w also figures in the time constraint, we may solve that second constraint for t_w and insert the result in the budget constraint. Maximum potential income is then $R + wT$, where T is total available time, while the trips all have generalized costs, that is, costs composed of a monetary cost and a time cost (trip time times the wage rate, possibly modified by the shadow prices of additional constraints on time uses).

The existence of the additional form of income wt_w disrupts the separation of the indirect utility function in one part that is a function of income and another part that is not. So the Gorman Polar Form no longer applies here, and conditions for the Kaldor–Hicks criterion to be used are not fulfilled.

The Kaldor–Hicks Criterion and Standard Transport Appraisal

The key issue in the appraisal of transport plans and projects is often if the time savings cover the monetary costs. To judge about that, we need to know the value of an hour of travel time savings in different modes and under different circumstances. National and international guidance on this issue is based on average values from stated preference value-of-time studies. Already here it becomes apparent that there can be no question of identifying winners and losers among the travelers based on their individual values of time. They will all be versions of the average traveler in the different settings that he experiences – short and long trips, work and leisure trips, car and public transport trips.

It might, of course, be quite clear that a given policy benefits individuals using public transport, for instance, while car users stand to lose. But this is on the average and might be different in individual cases. And indeed, the same person might win as a public transport user and lose as a car user. Apart from that, the Kaldor–Hicks criterion boils down here to the question if the “multiplied average traveler” can compensate the public sector, the transport operators, and the environment for the costs they incur.

It seems that to substitute the average traveler for the representative traveler does require some sort of justification. We are certainly not compelled for theoretic reasons to follow the average traveler everywhere he goes, but his views might nevertheless be of considerable interest. And in fact, as the distribution of individual values of time in stated preference studies is always skewed to the right (most people have lower values than the average), assuming that those who benefit from a project are also those that have to pay, more projects get selected using the average value of time than by voting (the mean is to the right of the median).

But by the very nature of things, a cost-benefit analysis using values of time from national guidebooks (that is, average values), should always be accompanied by at least some thoughts about distributional consequences – and possibly by some forms of compensation to the losers.

The Kaldor–Hicks Criterion and Random Utility Models

If the model is of the GEV family and the marginal utility of money is constant, the logsum formula is a valid welfare measure. It is basically identical to the consumer surplus (CS) welfare measure in the deterministic case. The conditions for it to apply are the same, and when the conditions are met, the EV and CV coincide with the CS, just as in the deterministic case.

The problem is, however, if the marginal utility of money is indeed constant in the case at hand, and what to do if this is not the case.

It may be shown that when the marginal utility of money is not constant, no representative individual exists, and the Kaldor–Hicks criterion does not apply. However, [Anders Karlström \(1999\)](#) has derived an exact and computable measure of expected user benefits (expected equivalent and conditional variations) that may be used for cost-benefit analysis with GEV models in which marginal utility of money is not assumed to be constant. So, if instead of deterministic values we are willing to make do with expectations; and if the concept of the average (mean) individual is accepted, then a meaningful welfare measure exists even in such models.

Alternative Approaches

Aggregate Willingness-to-Pay

Seeing that individual consumer surpluses, equivalent variations or compensating variations cannot legitimately be added without explicit or implicit assumptions on the weight of each individual in the welfare function, some economists have pointed to the following solution: If the aggregate willingness-to-pay in the concrete situation at hand is higher than the costs, the project or the plan should be carried out.

This criterion is impractical, because it requires a local valuation study to be carried out for each alternative in each project proposal. That would be expensive. And unless the stated amounts of money could actually be collected afterwards, the respondents would have every incentive to exaggerate their willingness to pay. A compromise solution could be to regularly carry out valuation

studies of a more general type at the local level. But even that would be very expensive, and the chances would be high that the incentive problems would persist.

Voting Theory

Transport project appraisal consists of a cost benefit analysis summing up the monetarized impacts plus an analysis of non-monetary impacts. If we let every impact (or suitable aggregate of impacts) be a “voter”, as it were, and let each of these voters rank the alternatives, there will be many different ways of using this information to pick a winning alternative or form an aggregate ranking. Voting theory is both a very old field of research and a very alive one (https://en.wikipedia.org/wiki/Category:Voting_theory). If one does not want to apply the Kaldor–Hicks criterion and do a cost benefit analysis, some form of voting theory might be considered, or at least it can be applied to the nonmonetary impacts. We cannot get around Arrows impossibility theorem,^a but we may for instance be able to define sets from which the winner must be chosen.

Another possible approach is to find a ranking of the alternatives that minimizes the distance to all the individual rankings, as measured, for instance by the minimum number of permutations needed to transform all individual rankings to an aggregate one.

Both voting theory and rank aggregation by permutations can be seen as forms of multicriteria analysis, but with set rules instead of subjective expert opinion.

Conclusion

The strict conditions for the aggregation over individual consumer surpluses to form a consistent measure of how much a policy is worth to society are not often met in real life. This means that the Kaldor–Hicks criterion cannot be applied without simplifying assumptions, either on the form of the demand functions or on the legitimacy of interpersonal comparisons. But even if it can, compensations to loser are often not feasible, not even in theory. This implies that there is a role for the Pareto criterion to be applied in conjunction with the Kaldor–Hicks criterion, to avoid policies with adverse distributional impacts or design compensating schemes to accompany them. There might also be a role for methods based on theories with less strict assumptions than the consumer theory of economics.

See Also

Computable General Equilibrium Analysis in Transportation Economics; Demand for Passenger Transport; Generalized Cost for Transport;

References

- Chipman, J.S., Moore, J.C., 1979. On social welfare functions and the aggregation of preferences. *J. Econ. Theory* 21, 111–139.
- Chipman, J.S., Moore, J.C., 1994. The Measurement of Aggregate Welfare. In: Eichhorn, W. (Ed.), *Models and Measurement of Welfare and Inequality*. Springer Verlag, Berlin, pp. 552–592.
- de Serpa, A., 1971. A theory of the economics of time. *Econ. J.* 81, 828–845.
- Dupuit J., 1844. De la mesure de l'utilité des travaux publics. In: Reprinted 1995 *Revue française d'économie*, Volume 10 (2), pp. 55–94.
- Gorman, W.M., 1953. Community preference fields. *Econometrica* 21, 63–80.
- Hicks, J.R., 1939. The foundations of welfare economics. *Econ. J.* 49, 696–712.
- Karlström, A., 1999. Four essays on spatial modeling and welfare analysis, Royal Institute of Technology, Department of Infrastructure and Planning, Stockholm.
- Samuelson, P.A., 1956. Social indifference curves. *Q. J. Econ.* 70 (1), 1–22.
- Sen, A., 2018. Social choice. In: *The New Palgrave Dictionary of Economics*, Macmillan, London.

Further Reading

- Anderson, S.P., de Palma, A., Thisse, J.F., 1992. *Discrete Choice Theory of Product Differentiation*. MIT Press, Cambridge.
- Blackorby, C., Donaldson, D., 1990. A review article: the case against the use of the sum of compensating variations in cost-benefit analysis. *Can. J. Econ.* 23 (3), 471–494.
- Chipman, J.S., 2018. Compensation principle. In: *The New Palgrave Dictionary of Economics*, Palgrave Macmillan, London.
- Dahl, G., Minken, H., 2010. A note on permutations and rank aggregation. *Math. Comput. Model.* 52 (1–2), 380–385.
- Kreps, D.M., 1990. *A course in microeconomic theory*. Harvester Wheatsheaf, New York.
- Varian, H.R., 1978. *Microeconomic Analysis*, third ed. W.W. Norton & Company, New York.

^a Arrow postulated four very reasonable properties that a social welfare function should possess, and showed that no welfare function could possess them all.

Social Discount Rates

Lars Hultkrantz, School of Business, Örebro University, Örebro, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	195
General	195
Concepts	195
Motives	196
The Cost of Waiting	196
The Cost of Uncertainty (Risk)	196
Discount Rates in Private Investment	196
Social Discount Rates	197
The Market-Based Approach	197
The Social Time Preference Approach	197
The Consumption Capital Asset Pricing Model	198
The Model	198
Approaches to Risk-Adjustment	198
Empirical Parameters	199
Market Rates	199
Ramsey-Equation Parameters	199
Risk-Adjusted Rates	200
Acknowledgment	200
References	200
Further Reading	200

Introduction

Although the value of travel time may be the shiniest star on the canopy of economic parameters in evaluations of transportation infrastructure, it often is the rate of discount that determines their bottom-line destiny. The force of exponential discounting effectively dwarfs long-term returns from such investments, to the disadvantage of objects with a long construction period and long life, such as new railroads, highways, airports, or ports. At a 6% rate, 56 dollars present value remain of a 100 dollars benefit that arises in 10 years from the present and only 5 dollars if it arises after 50 years. On the other hand, at a 1% rate such a benefit represents a present value of 91 and 60 dollars, respectively. Unfortunately for the credibility of the outcome of the economic evaluation, in spite of these huge differences in effect, both rates can be arguably plausible, based on theoretical and empirical reasoning. The precise motives for discounting and arguments for choosing a specific level or range of discount rates therefore deserve much attention.

General

Concepts

The discount rate is used to convert monetary valued benefits and cost, or the cash flow, at different points on a time line to a common metric, which is the value at a specific time, for instance the current, called the present value, or the future value at some specific later time, for instance when operation of services starts. The process of moving values from the future to the present is called discounting, while going in the opposite direction is called compounding. The conversion over the time line is made by multiplication with a discount factor $D(t_0, t)$ that often is defined as:

$$D(t_0, t) = (1 + r)^{(t_0 - t)} \quad (1)$$

in discrete time (for instance, years), where t_0 is the time used as reference and t is the time when the benefit or cost accrues. r in this formula is the rate of discount. The corresponding formula when time is given as a continuous variable is:

$$D(t_0, t) = e^{r(t_0 - t)}. \quad (2)$$

This second expression tells that the discount factor declines (and the compounding factor increases) exponentially over time at a speed that is determined by the discount rate. However, these mathematical expressions are not generally valid. Many argue that the

discount rate should, for various reasons, decline over the very long term. Expressions in Eqs. (1) and (2) are convenient and lead to time consistency (see explanation later), but may be unrealistic, at least for a long time horizon.

Motives

There are two main motives for discounting, both related to how economic value is affected by distance in time, giving rise to costs of time distance. The first is the cost of waiting and the second is the cost of uncertainty. The rate of discount is affected by both and can be split in two parts: the risk-free rate that reflects the cost of waiting and the risk (or uncertainty) premium, which captures the cost of risk (or uncertainty).

The Cost of Waiting

The cost of waiting derives from postponement of consumption or investment, or both. Let's start with consumption. Consumption gives utility to the individual and individual utility gives welfare to society (abstracting from negative externalities). Both individual utility and social welfare of a given basket of consumption goods are likely to decline with the distance in time. First, individuals may for several reasons have a genuine preference for being able to consume now rather than later. Very fundamentally, individuals are mortal, which means that the expected utility of consuming later is less than consuming now, since there is a positive probability of zero utility because of death. At the societal level, individual death is less of an issue, but still there is nonzero probability of extinction (from pandemics, nuclear war, comet hitting earth, etc.). Also, decision power belongs to the current generation that may have a preference for its own utility over that of future generations, especially of those that are much distant in time. Second, marginal utility of consumption (i.e., the additional utility from a small increment of consumption) declines with income. Therefore, if there is economic growth (increase of GDP per capita) then the individual will be richer in the future and future generations will be richer than the present. Hence, the marginal utility of future consumption will be lower, and this needs to be compensated. To put it bluntly, why should otherwise a poor current generation make sacrifices for future rich generations?

Then there is investment. Instead of consuming 1 unit today you can save and invest it so that you can consume $1 + r$ units tomorrow. Hence, there is an opportunity cost of using resources that could be used for productive investment that yields an expected positive return. To invest more in for instance infrastructure, it may be necessary to forsake investment in housing, machinery equipment, education, research, etc., that would have generated resources for future consumption.

The Cost of Uncertainty (Risk)

Future outcomes are uncertain. When possible outcomes and their probabilities are known it is called risk. The risk can be measured by one single parameter, the standard deviation, or the volatility, when future returns from investment follow a known normal probability distribution. It should be noticed though that extreme events, such as an outcome four standard deviations away from the mean, are utterly unlikely in a normal probability distribution, but in real life disasters do happen and that should affect decision-making.

Anyway, what is important to individual utility and societal welfare is not the volatility of a specific asset, but that of overall consumption possibilities as represented by wealth, which is a portfolio of assets including housing and human capital, and, on the societal level, overall consumption per capita (and its distribution). Therefore, the cost of investment in a risky asset is related to how the volatility of returns from that asset contributes to the overall volatility of wealth, or consumption per capita. The crucial aspect is therefore not total volatility of an asset but only the part that covaries with wealth or consumption.

The two aspects of the cost of time, waiting and risk, means that the rate of discount will consist of, at least, two components, the "risk-free" rate that reflects waiting and a "risk premium" that is the product of a general risk premium and a factor (beta) that reflects systematic risk.

Discount Rates in Private Investment

A basic framework in financial analysis of private capital is the Capital Asset Pricing Model, or the CAPM, developed in the 1960s by William Sharpe, John Lintner, and Jan Mossin. This model states that in equilibrium investors get a return r_i on a risky asset equal to the return r^f on a risk-free asset plus a risk premium equal to the premium $r^m - r^f$ on a fully diversified market portfolio of risky assets, called the equity premium, times β_i . These two terms thus represent both the cost of waiting and uncertainty. The "beta" is a measure of the nondiversifiable systematic risk, equal to the covariance divided by the variance, or the slope coefficient in a linear regression of the return on the specific asset with market portfolio return. Thus,

$$r_i = r^f + \beta_i (r^m - r^f) \quad (3)$$

where $\beta_i = \frac{\text{Cov}(r_i, r^m)}{\sigma_m^2}$.

A private capital investment is often funded from several sources that come with varying return on investment requirements. One reason for such differences is that the burden of the investment risk is unequally shared across sources, for instance by a lower risk on holders of bonds than on owners of stocks, the latter only having a residual claim. A second reason is that because of anticipation of future macroeconomic events, yield requirements on bonds and other securities at different terms, or maturities (i.e.,

time to repayment of principal) can vary. This is called the term structure. A third cause is different tax treatment, for instance because a firm's payment of interest rates, unlike payment of dividends, is deductible. As firms for various reasons want to spread funding over multiple sources, hence with varying rate-of-return requirements, investment projects are often evaluated by a Weighted Average Cost of Capital (WACC) rate, with each source's share of the funding as weights.

Social Discount Rates

A social planner that wants to get an appropriate discount rate for evaluation of a public funded program, that is, the social discount rate (SDR), has a basic choice between using a market-based rate or an estimate of the social rate of time preference. The main argument for the first approach is that it will reflect the actual opportunity cost of the resources used in the public program, and the main argument for the second approach is that social preferences may not coincide with market-revealed preferences.

The Market-Based Approach

The *market-based approach* uses the same rate of return requirements as for private capital. The main difficulty in implementation is the multitude of such rates to choose among. Rates come nominal or real valued, before tax or after tax, with a term structure, and vary depending on the risk of the underlying security in ways that still are not very well understood. An important aspect when navigating in this zoo is whether an investment is evaluated in a closed or in a (small) open economy. Early literature focused on the former case and suggested an approach for estimation of the SDR based on a weighted average of after-tax returns received by consumers and pretax return paid by private capital. This approach is still used for the federal guidelines for government CBA in the United States. However, in small open economies that are part of the global financial market, this approach, based on the idea that increased public spending crowds out domestic private consumption or investment, is less relevant as the opportunity cost of public spending is predominantly the (pretax) global market rate. In several countries, recommended SDRs are based on the cost of government borrowing, for instance as in New Zealand on the 5-year real risk-free government bond rate. Obviously, that rate does not consider project-specific risk. This is in contrast to the United States, where the recommended rate exceeds the rate on US government bonds, so that it, to a larger extent, corresponds to the long-term average return on a well-diversified portfolio of equities.

The Social Time Preference Approach

The social time preference (*STP approach*) has been developed along two lines in the theoretical literature: (1) the Ramsey model, which is originally based on deterministic conditions, and (2) the Consumption Capital Asset Pricing Model (CCAPM), which explicitly considers project risk. The first has its origins in a study by Frank P. Ramsey in 1928 who asked how much a nation should save. To answer this, he set up a model where social welfare is expressed as the sum of an infinite stream of instantaneous utility $u(C_t)$ from consumption C_t discounted at a constant rate δ . Maximizing social welfare specified in this way results in a first-order condition that, with some simplifying technical assumptions (iso-elastic utility and a constant rate of growth of consumption per capita), gives the *Ramsey equation*:

$$r = \delta + \gamma g, \quad (4)$$

where δ is the pure rate of time preference, γ is the elasticity of marginal utility with respect to consumption, and g is the rate of growth of consumption per capita. The first term is due to unequal weighting of utility of different generations or impatience within a generation, or both. The second term takes care of the fact that a long-term investment in a growing economy is a transfer from the poor (current generation) to the rich (subsequent generations), and adjusts the discount rate in proportion to the expected relative decrease of the marginal utility of consumption due to growth in consumption per capita. The Ramsey equation can also be derived in an overlapping generations model, which avoids the unrealistic assumption of an infinite time horizon, and leads into explicit consideration on whether δ reflects within life-cycle time preference or the current generations preferences for the utility of future generations, or both.

Several authors have derived an augmented Ramsey rule for stochastic consumption growth, thus introducing macroeconomic risk. Assuming that consumption per capita follows a geometric Brownian motion process with trend μ and independent and identically normally distributed stochastic terms with standard deviation (volatility) σ , an augmented version of this rule can be derived:

$$r = \delta + \gamma g - 0.5\gamma(\gamma + 1)\sigma^2. \quad (5)$$

The new term on the right-hand side is the precautionary effect on SDR. It reduces the SDR and thus to some extent promotes investment (in safe assets) as a means for hedging against volatile consumption (i.e., saving for a rainy day). However, the standard deviation of consumption growth in many advanced countries is small so the reduction of SDR is of second-order importance within the transportation-planning horizon. However, if growth rates shocks are persistent, because of serial correlation in error terms or regime switches (for instance, representing the possibility of catastrophic depressions), the precautionary term will not be constant

but grow over time from today's perspective. This gives rise to a substantially declining SDR, which means that the precautionary effect will be large when discounting benefits in the far future, as for instance in the case of climate policies. As shown in several works by Martin Weitzman, a similar effect arises when future SDRs are uncertain for whatever reason. The basic explanation is that the expected present value is a probability-weighted average of discount factors (see Eqs. (1) and (2)) and that these decline faster for high discount rates than for low rates. These arguments have led several countries to recommend declining SDRs (sometimes called hyperbolic discounting). A possible drawback is time inconsistency, which means that a decision-maker can come to different conclusions on the same investment prospect depending on when the decision is made. For instance, the optimal time for rehabilitation of a road may be different if it is assessed when the road is built or at a later point of time.

The Consumption Capital Asset Pricing Model

The Model

The second model, the CCAPM, is derived in a framework where a representative consumer maximizes the total expected utility from current consumption $u(c_0)$ and expected future consumption $Eu(c_1)$ discounted with the utility discount rate δ . This model adds to the augmented Ramsey equation by consideration of investment risk. Again applying some simplifying technical assumptions (iso-elastic utility, growth is a Brownian motion with trend and returns and the log of per capita consumption are jointly normally distributed) to the first-order condition gives the CCAPM equation:

$$r_i = \delta + \gamma g - 0.5\gamma(\gamma + 1)\sigma^2 + \beta_i\gamma\sigma^2 = r^f + \beta_i\pi. \quad (6)$$

Hence, this amends the augmented Ramsey equation (Eq. (5)) with a fourth term, but, as shown, the right-hand side can be nicely summarized into a sum of two terms: the risk-free rate and the risk term. The risk term is the product of the project-specific beta, β_i , and the average risk premium, π . The project beta is a measure of how the total consumption per capita risk (volatility) is affected by the project. It is defined as the covariance between the project returns and overall consumption per capita over the variance of consumption per capita. The average risk premium is the product of the marginal utility of consumption and the variance of consumption per capita. The similarity to the CAPM is no coincidence; CCAPMs (and other consumption-based approaches) are closely related to the CAPM. It can be noticed that total consumption can be regarded as the returns from all individual projects (including investment in human and natural resource capital), which implies that average beta is unity. As a large part of total consumption in modern welfare economies is public consumption, this implies that it is unlikely that project beta of an average public investment is zero (Baumstark and Gollier, 2014), as sometimes has been claimed (Arrow and Lind, 1970).

There is a seemingly close correspondence between the Ramsey and the CCAPMs in an economy where the consumer is a representative agent that therefore also can be seen as the social planner. However, the parameter γ has different interpretations in these models. In the Ramsey model $1/\gamma$ is the elasticity of intertemporal substitution, while in the CCAPM γ is the coefficient of relative risk aversion. These are separate aspects, though both relate to the curvature of utility functions.

The obvious difference between the ordinary Ramsey and CCAPMs, however, is that only CCAPM accounts for the fundamental fact that the future is not just happening at another time than the current but also that it is uncertain. In the standard framework where uncertainty is measured by standard deviation, the risk premium π is proportional to the variance of consumption per capita, which with historical data and plausible values of the coefficient of relative risk aversion implies a level of the risk premium that is substantially smaller than actual risk premia on asset markets (this is called "the equity premium puzzle"; Mehra and Prescott, 1985). Among several possible explanations, one is that investors have a strong aversion against risk in poor macroeconomic states, another is that the puzzle partly vanishes when fat-tailed properties of the probability distribution (i.e., low-probability disasters such as stock-market crashes or major political failures) are considered, and the third is that real consumption is more volatile than suggested by low-frequency expenditure data. A further aspect is the term perspective. The ex-post equity premium is considerably smaller when the risk-free investment alternative is taken to be a long-term sovereign bond until maturity than a short-term government bill. However, research on this issue has as of yet not come to a fully convincing resolution.

Approaches to Risk-Adjustment

Research on risk-adjustment of SDR is quite novel and many aspects remain to be considered. Two recent contributions give some guidance into how to derive the risk term in specific cases: the tail-hedge gamma and the elasticity-based beta.

The first of these approaches has been suggested in another contribution by Weitzman (2012, 2013). He assumes that the instantaneous net benefit of a single marginal investment project B_t can be decomposed as a linear combination of contemporary consumption, standardized with the expected value, and a project-specific random variable that is uncorrelated with consumption (which therefore can be made deterministic by diversification over a pool of projects). He defines the "real project gamma" as "the fraction of expected payoffs that on average is due to the uncertain macro-economy" (Weitzman, 2012, p. 15, italics in original). The discount rate for a project can then be computed as a weighted average of the rate of return on a risk-free asset r^f and risky equity r^m , where weights are given by the project-specific γ_t and the respective discount factors.

$$r_t = -\frac{1}{t} \ln \left[(1 - \gamma_t) e^{-r^f t} + \gamma_t e^{-r^m t} \right]. \quad (7)$$

This means that the short-term risk-adjusted rate is a gamma-weighted average of the two discount rates and then eventually falls down to the risk-free rate. Weitzman (2013) does not indicate how the gamma can be empirically estimated. This issue has later been explored by others. A useful result of that research (Mantolos and Hultkrantz, 2018) is that if projects returns and aggregate consumption follow Brownian motion and are co-integrated then gamma is unity, that is, the discount rate equals the equity rate.

The second approach has been developed by Gollier and Cherbonnier (2018). They address the task of estimating CCAPM beta assuming that investments, for instance in infrastructure, enable operation and consumption of services (transportation) that are governed by demand and supply relations that are log-linear in price and income, thus have constant elasticities. On the demand side, the consumers' instantaneous benefit of the service is:

$$B = \vartheta C^{\rho} \frac{x^{1-\alpha}}{1-\alpha}, \quad (8)$$

which gives the following demand function.

$$\ln x^d = \frac{1}{\alpha} \ln \vartheta + \frac{\rho}{\alpha} \ln C - \frac{1}{\alpha} \ln p, \quad (9)$$

where x^d is demand, C is income (or aggregate consumption), p is price, $\frac{\rho}{\alpha}$ is the income elasticity of demand, ρ is the income elasticity of the consumer valuation (willingness to pay), and $-\frac{1}{\alpha}$ is the price elasticity ($\alpha < 1$). Assuming that growth of C follows a Brownian motion with a linear trend, and variable cost follows Brownian motion, the CCAPM beta can be determined in the three special cases:

- A (perfectly inelastic supply): constant beta, $\beta = \rho$
- B (perfectly elastic supply): constant beta, $\beta = \frac{\rho}{\alpha}$
- C (limited capacity): declining beta, from $\beta = \rho$ to $\beta = \frac{\rho}{\alpha}$

In interpreting this, it should first be observed that, in a growing economy, a large-income elasticity of benefits means, on the one hand, a rapid growth of the nominal (undiscounted) benefits, and, on the other hand, a high beta, and therefore a high risk-adjusted discount rate. The net effect of an increase of this elasticity on the present value can therefore go in either direction.

The polar cases A and B have different implications for how income growth affects quantity and price of the service. In case B, with perfectly elastic supply, price will stay constant but the quantity demanded will vary proportionally to income, at a degree equal to the income elasticity of demand. In case A, quantity will be constant but price will vary with income, with the income elasticity of the willingness to pay as the factor of proportionality. Since $\alpha < 1$ the income elasticity of demand is by assumption in this model higher than the income elasticity of the willingness to pay. Thus, in the combined C case, which for instance could represent a transportation infrastructure investment where services are supplied at a constant marginal cost up to the capacity limit, beta for short maturities is close to the demand elasticity with respect to income and for long maturities falls to the willingness to pay elasticity with respect to income.

Empirical Parameters

Market Rates

Following Piketty's (2014) influential book, several studies have collected historical data on rates-of-return on various types of capital assets. For instance, Jordà et al. (2019) report long-run return on government bonds, short-term bills, equities, and housing in advanced economies from 1870, while Meyer et al. (2019) map returns on sovereign bonds for many countries traded in New York and London after the battle of Waterloo in 1815. Major findings by Jordà et al. (2019) are that real safe returns in peacetime periods have been low, in the 1%–3% range, but quite volatile, while the ex-post risk premium has been stable at about 4%–5%. That means that the risky rate on both equities and housing has also been quite stable averaging between 6% and 8%.

A large range of CAPM betas for transportation investment has been estimated for various modes, countries, and time periods. On the US market, values are regularly reported by "The iShares Transportation Average ETF" that seeks to track the investment results of an index composed of US equities in the transportation sector, with exposure to US airline, railroad, and trucking companies. This average beta usually exceeds unity. Krüger (2012) has used long time-series data for Sweden to compare transportation modes at various time scales. He found that traffic demand on railroad is more GDP-sensitive than demand for road, and freight transportation is generally more GDP-sensitive than passenger transportation. Also, traffic demand variance and GDP-sensitivity decrease with the length of time scale considered.

Ramsey-Equation Parameters

The first term of the Ramsey equation can conceptually be split in three components: pure impatience of the current generation, pure preference across generations, and the probability of extinction of the human race (for instance, because of a comet hitting planet earth). A recent survey among experts on their suggestions (Drupp et al., 2018) shows mean and median value of this parameter at 1% and 0.5%, respectively. The second term is g , the expected relative growth of consumption, or GDP per head, times γ , the elasticity

of the marginal utility of consumption. The latter parameter can be considered as reflecting both inter- and intra-generation inequality aversions. Commonly used levels of this parameter are about 1–2. Thus, a typical “Ramsey rate” without consideration of the risk cost, assuming a 1.5% growth rate, could be $0.5 + 1.5\% \times 1.5\% = 2.75\%$.

Risk-Adjusted Rates

The tail-hedge discounting model has been applied for estimation of risk-adjusted SDRs for transport infrastructure planning on historical time-series data in Sweden by [Hultkrantz et al. \(2014\)](#) and in Germany by [Goldmann \(2019\)](#). The indicated risk terms in proportion to the average equity premium are higher in Sweden (SE) than in Germany (DE) and in both countries lower for rail passenger travel than for road transportation or rail freight. For an average equity premium of 4.5% the short-term risk premia would be 2.6% (SE) and 1.0% (DE) for rail passenger travel and 4.2% (SE) and 3.6% (DE) for the other categories. The corresponding numbers for the 50 years maturity are 2.5% (SE) and 0.5% (DE) (rail passenger) and 3.7% (SE) and 1.7% (DE) (other).

Acknowledgment

I am grateful for very helpful comments by Niclas Krüger, Disa Asplund, and Maria Börjesson.

References

- Arrow, K.J., Lind, R., 1970. Uncertainty and the evaluation of public investment decisions. *Am. Econ. Rev.* 60, 364–378.
- Baumstark, L., Gollier, C., 2014. The relevance and the limits of the Arrow-Lind Theorem. *J. Nat. Resour. Policy Res.* 6 (1), 45–49.
- Drupp, M.A., Freeman, M.C., Groom, B., Nesje, F., 2018. Discounting disentangled. *Am. Econ. J. Econ. Policy* 10 (4), 109–134.
- Goldmann, K., 2019. Time-declining risk-adjusted social discount rates for transport infrastructure planning. *Transportation* 46, 17–34.
- Gollier, C., Cherbonnier, F., 2018. The economic determinants of risk-adjusted social discount rates. Working Paper 18-972. Toulouse School of Economics, Toulouse.
- Hultkrantz, L., Krüger, N., Mantalos, P., 2014. Risk-adjusted social rate of discount for transportation investments. *Res. Transp. Econ.* 47, 70–81.
- Jordà, O., Knoll, K., Kuvshinov, D., Schularick, M., Taylor, A.M., 2019. The rate of return on everything, 1870–2015. *Q. J. Econ.* 134 (3), 1225–1298.
- Krüger, N., 2012. Estimating traffic demand risk—a multiscale analysis. *Transp. Res. A Policy Pract.* 46 (10), 1741–1751.
- Mantalos, P., Hultkrantz, L., 2018. Estimating “gamma” for tail-hedge discount rates when project returns are co-integrated with GDP. *Appl. Econ.* 50 (37), 4074–4085.
- Mehra, R., Prescott, E.C., 1985. The equity premium: a puzzle. *J. Monet. Econ.* 15, 145–161.
- Meyer, J., Reinhart, C.M., Trebesch, C., 2019. Sovereign bonds since Waterloo. NBER Working Paper 25543. National Bureau of Economic Research, Cambridge, MA.
- Piketty, T., 2014. *Capital in the Twenty-First Century*. Harvard University Press, Cambridge, MA/London, England.
- Weitzman, M.L., 2012. Rare disasters, tail-hedged investments, risk-adjusted discount rates. Working Paper 18496. National Bureau of Economic Research.
- Weitzman, M.L., 2013. Tail-hedge discounting and the social cost of carbon. *J. Econ. Lit.* 51 (3), 873–882.

Further Reading

- Atkinson, G., Braathen, N.A., Groom, B., Mourato, S., 2018. Cost-benefit analysis and the environment. In: *Further Developments and Policy Use*, OECD Publishing, Paris.
- Cochrane, J., 2010. Presidential address: discount rates. *J. Finance* 66 (4), 1047–1107.
- Gollier, C., 2011. *Pricing the Future: The Economics of Discounting and Sustainable Development*. Princeton University Press, Princeton, NJ.
- Gollier, C., Hammit, J.K., 2014. The long-run discount rate controversy. *Annu. Rev. Resour. Econ.* 6, 8.1–8.23.

Cross-Elasticities Between Modes

Mark Wardman, Jeremy Toner, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	201
Cross-Elasticity Relationships	202
Estimating Inter-Modal Cross-Elasticities	203
Aggregate Demand Models	203
Discrete Choice Models	204
Deduced Cross-Elasticities	204
Cross-Elasticity Conventional Wisdom	204
Illustrative Cross-Elasticities	205
Review Evidence on Cross-Elasticities	205
Deducing Cross-Elasticities	205
Conclusions	208
References	208
Further Reading	208

Introduction

A widely used summary measure of a consumer product's demand sensitivity to changes in its price and other attributes and also in the price and attributes of rival products is provided by what are termed elasticities. These represent the proportionate change in the demand for a product in response to a proportionate change in some continuous attribute, such as price or journey time. Analogous measures can be calculated after a unit change in some categorical variable, such as the need to interchange vehicles or the introduction of new vehicles, and these are sometimes termed elasticity-type measures.

Elasticities are widely used across consumer markets because they have the attractive properties of being transparent, interpretable, transferable, and easy to apply in demand forecasting. The transport market is no different, with demand elasticities used for planning and economic appraisal purposes, although what is perhaps an unusual feature of the transport market is that nonprice related attributes, such as journey time, access/egress time, reliability and comfort, and in the case of public transport service frequency, crowding and the need to change vehicles, attract as much if not more attention as price. For standard consumer products, it is price elasticity which tends to dominate in conventional economic theory. Within transport, price can take various forms, such as fuel cost, parking cost, toll charge, and public transport fare with likely different degrees of sensitivity.

The key determinants of elasticity values, other things being equal, are likely to be:

- the availability of substitutes;
- the time period (short-, medium-, long-term) over which changes are measured;
- the need for the trip—at all, by a particular mode, to a particular destination; and
- the proportion of income spent on the trip.

These apply equally to: own-elasticities, which measure change in demand for a product when some attribute of that product changes; and cross-elasticities, which measure change in demand for a product when some attribute of a different product changes. Of course, for many practical purposes, the latter may be zero; it is unlikely that the demand for bus trips in Stockholm depends on the price of meatballs. It might therefore be expected that in many cases, cross-elasticities will be (absolutely) small, since they capture only part of the response of a change in a product attribute.

Demand elasticities in the transport market therefore cover a wide range of attributes and have long been one of the most important parameters of transport planning. Of specific interest here are cross-elasticities which deal with competition between different products which might, for example, be alternative modes of transport, routes, destinations, operators, or ticket types. And from these various choice contexts, we here address competition between modes.

Cross-elasticities between modes are important for a number of reasons:

- Policy interventions on the grounds of environmental, congestion, and energy considerations are critically dependent upon forecasting modal transfer.
- Strategic demand forecasting by public transport operators requires that the effects of changes in the prices and service quality of competing modes are identified. They have increasingly recognized that in ever more competitive and deregulated environments it is not sufficient to focus only on their own demand but also on what can be attracted from other modes and the threats posed by substitute modes.

- Competition policy and antitrust policy, particularly considering ownership/control of potentially competing modes or routes of transport such as long-distance rail versus intercity coach.
- To operationalize the theory of second-best pricing, whether this be to adjust the price of one mode in order to mitigate the failure to internalize an externality in a related mode (for example, reduce bus fares to offset the failure to introduce road pricing) or to adjust the prices of different modes (bus, rail, tram) by different proportions in order to meet an overall budget constraint for the transit operator.

Essentially, travelers now have more choices than ever before and where realistic choices exist then cross-elasticities are relevant.

Cross-Elasticity Relationships

Cross-elasticities cannot be understood in isolation. But they are also notoriously tricky to find. The purpose of this section is to examine the theoretical relationships between own- and cross-elasticities in order to see how cross-elasticity values may be deduced using the evidence on own-elasticities which is available or to see how such relationships can be used to enhance the accuracy of their estimation.

Inter-modal cross-elasticities denote the proportionate change in the demand for a mode after a proportionate change in some variable that characterizes the attractiveness of a substitute mode. Larger cross-elasticities between a pair of modes indicate greater substitutability, all other things equal.

Cross-elasticities play a central role in the economic theory of travel demand, and there are a number of relationships that exist that are important in understanding, estimating, and deducing cross-elasticities and indeed ensuring consistency of empirical evidence with economic theory. They also demonstrate clearly a key feature of cross-elasticities that they can be expected to be highly sensitive to market conditions. It is one thing to use constant own-elasticities, which is often countenanced in transport demand forecasting even though the basis for it is sometimes questionable, but it is far less acceptable to adopt constant cross-elasticities as the various relationships set out below indicate.

First, the Slutsky symmetry equation expresses the relationship between two price cross-elasticities as:

$$\eta_{ij} = \eta_{ji} \cdot \frac{D_j P_j}{D_i P_i}, \quad (1)$$

where η is the cross-elasticity between two modes i and j , D denotes their respective levels of demand and P their prices. Strictly, this rule applies to income compensated elasticities. For ordinary (Marshallian) elasticities, additional terms including the income elasticity are required. However, when the income share of expenditure is small, as will often be the case, the difference between compensated and ordinary elasticities is small, certainly at the level of the individual urban trip. In other cases, this approximation may not be valid, for example the cost of an annual public transport season ticket versus the cost of maintaining a private vehicle for commuting purposes. The prospect of large variations in cross-elasticities is here apparent.

Second, the Cournot aggregation condition specifies that:

$$\sum_{j=1}^n \omega_j \eta_{ji} + \omega_i = 0. \quad (2)$$

This states that the sum of cross-elasticities weighted by the budget share ω plus the budget share of mode i should be zero. This relationship implies that cross-elasticities are bigger (smaller) when there are fewer (more) competitive alternatives and that cross-elasticities are bigger (smaller) when the mode has a major (minor) share of expenditure and, if similar prices, trips.

The usefulness of this result is helping to fill in gaps when some, but not all, of the elasticities are known, to confirm/adjust/make consistent a set of own-elasticities and cross-elasticities. It can also be imposed as a condition as part of the estimation process to enhance accuracy and ensure consistency of results.

Third, the homogeneity condition, whereupon demand functions are homogenous of degree zero in prices and incomes. This is essentially a formal statement of the absence of money illusion. Across n modes, this is specified as:

$$\sum_{j=1}^n \eta_{ij} + \eta_{ii} = 0. \quad (3)$$

This requires that the sum of the own elasticities and cross-elasticities for a mode and its income elasticity (η_{ii}) equal zero. It is unclear whether this can ever, in restricted circumstances, be applied to time; certainly, it cannot apply to time in the long run for an individual.

These three relationships rely upon the assumptions of neo-classical micro-economic theory and, since they apply at the individual level, we must assume that aggregate behavior reflects that of the typical individual. There is though a relationship between cross and own elasticities that is true by definition:

$$\eta_{ij} = -\eta_{ji} \frac{D_j}{D_i} \cdot \delta_{ji}, \quad (4)$$

where additionally δ_{ji} is the diversion factor denoting the proportion of users of mode j who switch to or are attracted from mode i . This applies to all variables and not just price and is therefore particularly important as a means of deducing cross-elasticities where there is little or no evidence.

If we believe that a unit change in generalized cost has the same impact regardless of whether it stems from a change in price or journey time or any other variable, which many regard to be a reasonable pragmatic assumption, then the following relationship between, say, time (η_{ijt}) and price (η_{ijp}) cross-elasticities exists:

$$\eta_{ijt} = \eta_{ijp} \frac{vT_j}{P_j}, \quad (5)$$

where T denotes journey time and v is the value of time. This can be extended to other cross-elasticities that enter the generalized cost term such as walk time, headway or late time. Given that the literature on cross-elasticities tends to be limited, and might particularly be so for variables which tend not to vary much such as bus and car journey times or access times to train, this allows cross-elasticities to be deduced from evidence that does exist, with price most likely being the reference cross-elasticity used.

Hence we conclude that cross-elasticities:

- are intrinsically variable;
- should observe specific relationships (in some cases approximately); and
- can be deduced.

Estimating Inter-Modal Cross-Elasticities

We can distinguish between estimation models that are based on aggregate data, which is the collective outcome of decision makers' choices, or where the unit of observation is those individual choices. We here provide a *brief* account of the key features of the different approaches

Aggregate Demand Models

Such models are based on aggregates of modal demand, which are the outcome of choices made by the relevant population of individuals. They can operate at different levels of aggregation. At one extreme, the number of trips by a specific mode might be recorded at the national level, such as the number of car kilometers per year or international departures. More granularity and indeed more data are provided by demand recorded at the route level, such as the number of bus trips on different corridors or number of rail trips between various stations, while the least aggregated form is the total number of trips made by a household or individual in a specific period.

The models are almost invariably based on secondary data, such as those obtained from public accounts, traffic counts, national travel surveys, or the sales data of transport operators, although survey data might be collected specifically for the purposes of modeling demand at the aggregate level. Secondary data are often limited in behavioral detail, such as breakdowns of the purposes of the journeys or journey length distributions, but nonetheless provides a firm basis in actual behavior.

The models tend to be single equation, dealing with the demand for a specific mode, and to be based on data with a time series element. Variations in demand over time are explained as a function of the price, journey time and other relevant features of the mode in question, the levels of external factors such as income, population and employment which are key drivers of travel behavior, and the price and characteristics of substitute modes.

In general, aggregate models contain limited allowance for cross-elasticities, partly because obtaining reliable historic data on the times, prices, and service quality features of rival modes is not straightforward, and makes little use of the relationships set out in the second section. There can also be problems of large correlations between variables and limited variations in others, which compounds the difficulties of "hard to estimate" cross-elasticities given they tend to be relatively small. Where cross-elasticities are successfully estimated, they tend to be a single term and do not exhibit the spatial and indeed temporal variations that would be expected from the relationships earlier.

An approach, which overcomes some of these difficulties, is to explain the market shares of different modes at a specific point in time in terms of their prices, journey times, and other features. There is then no need to source historic data while spatial variations in prices and journeys times generally exceed temporal variations. However, it is rare that sufficient market share data are available for modeling purposes while market share can vary due variations in the size of the total market for reason not connected with the variables in the mode choice model. The discrete choice models discussed later provide better possibilities for investigating mode choice.

Alternatively, we might explicitly harness the relationships of economic theory in developing an empirical understanding of the demand for different modes and the competition between them. These could take the form of a set of inter-related demand functions covering the various modes, rather than a single mode specific demand function, whereupon the relationship among cross-elasticities and between cross- and own-elasticities can be used both to assist the challenges of estimating relatively small and inherently variables cross-elasticities and also ensure consistency between estimated values and economic theory. Unfortunately, there has been very little use made of such methods in the transport market.

Discrete Choice Models

In contrast to the analysis of behavioral aggregates, there is a rich strand of research that examines the choices made at the level of the decision maker, typically the individual, and mode choice is the most commonly examined real-world travel choice context. Explaining choices between modes naturally provides cross-elasticity estimates as well as other means of forecasting competition.

Such models take the market size as given and aim to explain individuals' choices as a function of the times, prices, and other relevant features of different modes. The parameters estimated to explain how time and price and other factors impact on choice allow both own and cross-elasticities to be estimated, although the former account only for the mode choice element and must be supplemented with evidence on how changes in prices and times lead to changes in the number of trips.

The models can be estimated to actual choices made in the market place, and which are referred to as revealed preference (RP) data. In contrast with aggregate methods, discrete choice methods can exploit what is typically referred to as Stated Preference (SP) data, where individuals are presented with a series of different hypothetical mode choice scenarios and asked to state which mode they would prefer.

The attractions of RP data are that it reflects what decision makers actually do, but then it is limited by what the market has to offer which might not provide a firm basis for the analysis of how price, time, and other factors impact upon mode choice.

The attractions of SP data are to a greater extent the mirror image; respondents are not committed to behave in accordance with their stated preferences but the control of the experimental context means that the data problems that can afflict RP methods are avoided and multiple choice observations per decision maker are collected.

Both methods have the attraction of being able to undertake detailed segmentation by key variables because of the availability of individual level data. So, for example, they allow cross-elasticity estimates to be obtained by journey purpose and journey distance, which can be readily implemented in application models.

A key feature of discrete choice models, again in stark contrast to the aggregate models, is that the elasticity terms are inherently variable, depending upon market share and generally the levels of the variable in question and thus not obviously immediately transferable between situations. The models calibrated need to be specified carefully in order to ensure that the resulting elasticities are consistent with economic theory; this is not always done. Mode share elasticities are very different in nature when compared with ordinary (Marshallian) or income-compensated (Hicksian) elasticities of the type more conventionally used by economists. While mode share elasticities can be derived from Marshallian or Hicksian elasticities, the reverse is in general not possible.

While discrete choice models have numerous attractions in principle, in practice SP evidence dominates and models estimated to such data tend to provide what is often regarded to be large sensitivity in the demand for one mode in response to changes on another mode. There are also issues involved in relating changes in practice to individuals' perceptions while, unlike aggregate models based on time series data, the models do not readily incorporate a dynamic aspect and lags in behavioral response.

Deduced Cross-Elasticities

Given the challenges involved in estimating cross-elasticities, and indeed the time and resources required, there has been increasing use in recent years of deducing cross-elasticities. This uses Eq. (4) earlier which consists of the relevant own-elasticity, relative demand levels, and diversion factors.

Own-elasticities are typically obtained from official guidance, conventional wisdom, or the wealth of evidence covered in the numerous reviews of own-elasticities that now exist in the literature. Ideally, these should be explicitly long run, but with some indication of the short-run impact and the path to the long-run effect.

Relative demand levels can be obtained from national travel surveys, which are conducted in many countries, and are sometimes contained in census data. Other sources might be local transportation models and studies as well as bespoke surveys and indeed "informed guesstimates." It can be more difficult to obtain robust estimates of relative demand as the level of spatial disaggregation becomes more detailed; aggregation over large areas smooths out much of the inherent variability (as is, of course, common with averaging and, indeed, in most cases, an attractive feature).

Diversion factor evidence can be obtained from reviews, as reported in (10031 – Elasticities for Travel Demand), but it might also be practical to undertake surveys of travelers to determine their best alternatives. The latter has attractions given that diversion factors can be expected to vary across contexts, not least as the relative attractiveness of different modes varies, although the issue of spatial detail again arises. In some circumstances, informed guesstimates may well be sufficient.

A further attraction of this approach of deducing cross-elasticities is that it allows context specific cross-elasticities to be used in forecasting and appraisal while also aiming for consistency with economic theory.

Cross-Elasticity Conventional Wisdom

The evidence on cross-elasticities is much less secure than that on own-price elasticities since it has been derived from fewer sources and is much more location-specific. Toner et al. (1995) found very small cross-elasticities of demand for car with respect to rail price, mostly under 0.1, for the inter-urban market. Larger cross-elasticities were found for rail demand with respect to car travel time in the business market. Dargay et al. (2010) found some evidence of larger cross-elasticities, particularly between rail and coach and for rail

and coach demand with respect to price of car, up to 0.5 being common. Cross-elasticities of demand for car with respect to public transport were, though, still very small.

In the urban context, [Lewis \(1977\)](#) reports a car use elasticity with respect to public transport prices of 0.06 in the peak in London. [Kemp \(1973\)](#) has a figure of 0.14 from his work in Boston, again for car use with respect to public transport price. [Glaister and Lewis \(1978\)](#) derive cross-elasticities of peak car use with respect to peak and off-peak bus and rail fares of approximately zero in London. [Dodgson \(1990\)](#) reports on an elasticity of bus use with respect to underground fare in London of 0.1, but a much lower elasticity with respect to surface rail fare. He cites a much higher elasticity of bus use with respect to car costs of 0.62, based on work in six towns and cities. This seemingly large figure is probably because of the relatively minor role played by bus; relatively few car drivers switch, but they are large relative to the number of existing bus users. [Goodwin \(1992\)](#) obtains an average public transport use elasticity with respect to car cost of 0.34.

[Rietveld and Daniel \(2004\)](#) find low cross-elasticity between bicycle and car, and that bicycle mainly competes with public transport. [Wardman et al. \(2007\)](#) forecast that a set of political instruments to double cycling volumes only reduces car traffic by 5%. [Börjesson and Eliasson \(2013\)](#) find similar results that only 13% of Stockholm cyclists quote car as their second-choice mode. The 1991 introduction of the Oslo toll road scheme generated some research on the effect of car user payment on public transport. The effect was deemed negligible ([Ramjerdi et al., 2004](#)).

The published evidence on cross-elasticities is thus sparse, difficult to apply, since it is location-specific, depending crucially on the mode shares in the different cities, and not obviously susceptible to a “conventional wisdom” approach.

Illustrative Cross-Elasticities

Prior to setting out some illustrative cross-elasticities, we make two important points. First, the amount of empirical evidence relating to cross-elasticities tends to be much less than for own elasticities and indeed review studies of the former are far less common than for the latter. Second, while it has been pointed out that cross-elasticities are inherently variable parameters, despite this cross-elasticities are constant in some model estimations and indeed in appraisal recommendations.

The focus here is on the output of a major review of international cross-elasticity evidence and upon deducing cross-elasticities given the challenges of estimation. A feature of both discussions is the inherent variability of cross-elasticities.

Review Evidence on Cross-Elasticities

The most extensive review of cross-elasticity evidence so far undertaken covered 1096 values from 93 studies and 18 countries spanning the years 1961 to 2017 ([Wardman et al., 2018](#)). The regression based meta-analysis explained variations in cross-elasticities according to a number of factors, and key among them were the transport variable in question, combination of mode affected and mode altered, method of analysis, whether the cross-elasticity was explicitly short run, long run or indeterminate, journey purpose, journey length, a time trend, and mode share. The mode share term essentially replicated the D_j/D_i term of Eq. (4).

Table 1 reports illustrative cross-elasticities implied by the meta-model for a range of circumstances. This is done for cost (fuel or fare as appropriate) and time cross-elasticities between the main modes of car, bus, and rail, primarily for a variety of urban trip features but with extension to inter-urban trips, and for the three main journey purposes. The implied cross-elasticities are for 2017 and are generally long run but with some short-run variants.

A key feature of the meta-model is the impact of relative demand as expected from theory. The figures used in **Table 1** are the approximate 25th, 50th, and 75th percentiles of relative demand for urban trips in the estimation data set. These figures are 4, 7 and 14 for D_C/D_B , 0.4, 0.9 and 1.2 for D_R/D_B , and 3, 10 and 16 for D_C/D_R . Of course, a more extreme set of values could be used, such as a very high ratio of D_R/D_B and a very low ratio of D_C/D_R for commuting trips into major metropolis. Implied cross-elasticities are also provided for inter-urban trips but only for the mean ratios, which were 21 for D_C/D_B , 5 for D_R/D_B , and 4 for D_C/D_R .

The illustrative cross-elasticities seem sensible, covering a large range in line with expectations but without implying extreme cross-elasticities that would cast doubts upon the estimated meta-model; there is no guarantee of such desirable properties from meta-models!

The model can be used to provide cross-elasticities customized to specific contexts. As expected, the implied cross-elasticities exhibit a considerable amount of variation, not only according to the relative demand levels but also between time and cost, short run and long run, urban and inter-urban trips, and commuting, leisure and business travel.

Car cross-elasticities tend to be low, reflecting the generally large market share of car, and are larger for time than cost. The public transport cross-elasticities are also larger for time than cost, in some cases very much so. These differentials have implications for policy measures aimed at inducing mode switch. The long-run cross-elasticities are noticeably larger than the short run, with a ratio of around 2.4., and the inter-urban cross-elasticities indicate a generally much stronger degree of competition between modes than for urban trips.

Deducing Cross-Elasticities

Since 1986, the *Passenger Demand Forecasting Handbook* (PDFH) has been unique in the world in providing a framework and associated parameters that are recommended for use in Great Britain for forecasting the effects of changes in a wide range of variables.

Table 1 Illustrative cross-elasticities implied by meta-model.

	Relative demand			Cost (Car fuel or public transport fare)						In-vehicle time					
	D_O/D_B	D_R/D_B	D_O/D_R	Bus:Car	Bus:Rail	Rail:Car	Rail:Bus	Car:Bus	Car:Rail	Bus:Car	Bus:Rail	Rail:Car	Rail:Bus	Car:Bus	Car:Rail
Commuting urban LR	4	0.4	3	0.16	0.17	0.21	0.34	0.06	0.07	0.33	0.18	0.42	0.38	0.07	0.08
Leisure urban LR	4	0.4	3	0.16	0.17	0.21	0.34	0.06	0.07	0.48	0.26	0.61	0.54	0.10	0.11
Business urban LR	4	0.4	3	0.11	0.11	0.13	0.22	0.04	0.05	0.45	0.25	0.57	0.51	0.10	0.11
Commuting urban LR	7	0.9	10	0.20	0.22	0.32	0.26	0.05	0.05	0.41	0.24	0.65	0.28	0.06	0.05
Leisure urban LR	7	0.9	10	0.20	0.22	0.32	0.26	0.05	0.05	0.58	0.35	0.93	0.40	0.08	0.07
Business urban LR	7	0.9	10	0.13	0.15	0.21	0.17	0.03	0.03	0.55	0.33	0.87	0.38	0.08	0.07
Commuting urban LR	14	1.2	16	0.25	0.25	0.37	0.23	0.04	0.04	0.52	0.27	0.77	0.26	0.05	0.04
Leisure urban LR	14	1.2	16	0.25	0.25	0.37	0.23	0.04	0.04	0.75	0.39	1.10	0.37	0.07	0.06
Business urban LR	14	1.2	16	0.17	0.16	0.24	0.15	0.03	0.03	0.70	0.36	1.03	0.34	0.06	0.06
Commuting inter LR	21	5	4	0.29	0.71	0.49	0.14	0.04	0.18	0.60	0.78	1.02	0.15	0.04	0.20
Leisure inter LR	21	5	4	0.29	0.71	0.49	0.14	0.04	0.18	0.86	1.12	1.46	0.22	0.06	0.29
Business inter LR	21	5	4	0.19	0.47	0.32	0.09	0.02	0.12	0.81	1.05	1.37	0.21	0.05	0.27
Commuting urban SR	7	0.9	10	0.08	0.09	0.13	0.11	0.02	0.02	0.17	0.10	0.28	0.12	0.02	0.02
Leisure urban SR	7	0.9	10	0.08	0.09	0.13	0.11	0.02	0.02	0.25	0.15	0.40	0.17	0.04	0.03
Business urban SR	7	0.9	10	0.05	0.06	0.09	0.07	0.01	0.01	0.23	0.14	0.37	0.16	0.03	0.03

Nonetheless, prior to the 4th edition in 2002 there was no accounting for competition from other modes in the form of cross-elasticities. That edition introduced recommended cross-elasticities between rail and the competing modes of car, bus, and air, largely stimulated by the previously unaccounted for effects of increases in fuel prices and car congestion on rail demand. This was an advance but industry analysts recognized that cross-elasticities are context dependent, whereupon the next edition (4.1) in 2005 provided a further advance by including a recommendation to deduce cross-elasticities using Eq. (4).

Car costs are a potentially important driver of rail demand not only because car is the main competitor to rail in many contexts in Great Britain but because PDFH recommends some quite high cross-elasticities to car cost in the range 0.2 to 0.4 and fuel prices have a history of varying and indeed are forecast to fall somewhat in the medium term.

While PDFH still recommends “overall” cross-elasticities by a limited number of flow types, Table 2 demonstrates how the deduced cross-elasticity of rail demand with respect to car fuel cost using Eq. (4) can vary considerably.

The long-run car fuel cost own-elasticities (η_{CC}) are taken from official Department for Transport recommendations contained in webTAG, and are -0.1 for business, -0.3 for commuting, and -0.4 for other trips. The diversion factors (δ_{CR}) are taken from a review of such evidence and reported in (10031 – Elasticities for Travel Demand) and are 0.12 for urban trips and 0.65 for inter-urban trips. While the own-elasticities might vary by distance and other factors, and the diversion factors can be expected to vary with market conditions among other things, evidence on such variation is lacking or not clear-cut. Evidence on the relative demand of car and rail (D_C/D_R) for the specific flows in question was taken from the National Travel Survey, which provides a representative account of travel behavior in Great Britain. Finally, for leisure trips, we allow for car occupancy of 1.7 when deducing the implied transfer to rail.

There are instances where the cross-elasticities are very low, generally for trips of any purpose into Central London where rail dominates given high levels of traffic congestion, the congestion charge and the high costs and limited availability of parking. When extended to Inner and Greater London, rail is less dominant since these car related problems are less and hence the cross-elasticities increase, particularly for other trips.

For longer distance trips over 100 miles not involving London, rail is in a weaker position and the cross-elasticities are accordingly higher although even here rail is in a stronger position, presumably because of the distances involved, than for the longer distances trips that are between 20 and 100 miles where we observe some very large cross-elasticities as is the case for other trips for short distance trips outside London. Car is in a much stronger position, in general, for trips not involving London.

Of course, cross-elasticities will vary across specific flows within any of these categories as market conditions vary. But the figures show much more variation than PDFH recommends.

While the example here is based around car cost cross-elasticities, the procedure readily extends to other variables and to competition between other modes therefore enabling a full set of context specific cross-elasticities to be obtained for policy and appraisal purposes.

Table 2 Illustrative cross-elasticities of rail demand with respect to fuel cost.

Flow type	Purpose	η_{CC}	δ_{CR}	D_C/D_R	η_{CR}
Long distance to central London ≤ 150 miles	Business	-0.1	0.65	0.21	0.01
	Other	-0.4	0.65	0.24	0.11
Long distance to central London > 150 miles	Business	-0.1	0.65	0.18	0.01
	Other	-0.4	0.65	0.11	0.05
Long distance to central London	Business	-0.1	0.65	0.21	0.01
	Other	-0.4	0.65	0.20	0.09
Long distance to greater London	Business	-0.1	0.65	0.56	0.04
	Other	-0.4	0.65	0.95	0.42
Non-London long > 100 miles	Business	-0.1	0.65	3.79	0.25
	Other	-0.4	0.65	1.37	0.61
Non-London long 21–100 miles	Business	-0.1	0.65	9.97	0.65
	Other	-0.4	0.65	2.44	1.08
Non-London < 20 miles within urban areas	Commute	-0.3	0.12	4.66	0.17
	Business	-0.1	0.12	9.00	0.11
	Other	-0.4	0.12	9.99	0.82
London travel card area	Commute	-0.3	0.12	0.92	0.03
	Business	-0.1	0.12	1.63	0.02
	Other	-0.4	0.12	2.62	0.22
South east to central London	Commute	-0.3	0.65	0.04	0.01
	Business	-0.1	0.65	0.28	0.02
	Other	-0.4	0.65	0.16	0.07
South east to central and inner London	Commute	-0.3	0.65	0.16	0.03
	Business	-0.1	0.65	0.53	0.03
	Other	-0.4	0.65	0.47	0.21

Conclusions

Cross-elasticities are important for a range of policy evaluations: for governments seeking to effect mode shift from more- to less-polluting modes; for operators seeking to forecast demand as prices and quality attributes of various modes change; for assessing the efficiency, in a second-best world, of transport prices.

In this chapter, we have introduced some relationships between own- and cross-elasticities. These relationships can be used: to constrain econometric analysis to enhance the accuracy of estimation; to assess the plausibility of elasticities estimated using econometric models; and to fill in gaps where some elasticities are known but others are not. These relationships should always hold for price variables, but only some are relevant for time and quality variables.

Some headline results from economic reasoning are as follows:

- The own-price elasticity should be negative and can be absolutely small or large in magnitude.
- The cross-elasticities should all be positive (for substitute goods) and smaller in magnitude than the own price elasticity.
- The more cross-elasticities there are (i.e. the more related products there are; these related products can be competing modes along the line of route or even away from the line of route. As with all analysis of competition, market definition and geographic scope are important determinants of the range of feasible alternatives), the smaller each cross-elasticity should be on average for a given own price elasticity; and/or the bigger in magnitude the own price elasticity.

In general, unless a detailed bespoke study is undertaken for a particular place/flow, we recommend deducing cross-elasticities since they are so highly context-specific. Simply transferring cross-elasticities from one situation to another is difficult unless the competitive situation and the existing mode shares are very similar, although the meta-analysis reported here does facilitate transfer by allowing for mode share effects. But to make these deductions securely, much more evidence is needed on diversion factors—more information but also how these vary by typical factors like purpose and distance, and more on the effect of the competitive situation. It is necessary to be careful when defining market shares to ensure spatial consistency.

This blend of data analysis and economic theory seems to us the most promising way to estimate a full set of own- and cross-elasticities for transport given the inherent difficulties in most contexts of simply estimating such things econometrically.

References

- Börjesson, M., Eliasson, J., 2013. The benefits of cycling: viewing cyclists as travellers rather than non-motorists. In: Parkin, J. (Ed., 2012), *Cycling and Sustainability*, Emerald, Bingley, UK (Chapter 10).
- Dargay, J.M., Clark, S.D., Johnson, D., Toner, J.P., Wardman, M., 2010. A forecasting model for long distance travel in Great Britain 12th WCTR, July 11–15, 2010, Lisbon, Portugal. Also ETC, October 13–15, 2010, Glasgow, UK.
- Dodgson, J.S., 1990. Forecasting inter-modal interaction. Paper prepared for the Department of Transport Seminar on Longer Term Issues, July 1990, London.
- Glaister, S., Lewis, D., 1978. An integrated fares policy for transport in London. *J. Public Econ.* 9 (3), 341–355.
- Goodwin, P.B., 1992. A review of new demand elasticities with special reference to short and long run effects of price changes. *J. Trans. Econ. Policy* 26 (2), 155–169.
- Kemp, M.A., 1973. Reduced fare and free urban transit services - some case studies. In: Symposium on public transport fare structure, TRRL Supplementary Report 37UC, Crowthorne.
- Lewis, D.L., 1977. Estimating the influence of public transport on road traffic levels in Greater London. *J. Trans. Econ. Policy* 11 (2), 155–168.
- Ramjerdi, F., Minken, H., Østmoe, K., 2004. The Norwegian urban tolls. In: Santos, G. (Ed.), *Road Pricing: Theory and Evidence*. Elsevier, Oxford, UK.
- Rietveld, P., Daniel, V., 2004. Determinants of bicycle use: do municipal policies matter? *Trans. Res. A* 38 (7), 531–550.
- Toner, J.P., Wardman, M., Whelan, G.A., 1995. Competitive interaction in the inter-urban travel market in Great Britain. Presented at PTRC Summer Annual Conference, September 1995, Seminar F, pp. 29–39.
- Wardman, M., Tight, M., Page, M., 2007. Factors influencing the propensity to cycle to work. *Trans. Res. A* 41, 339–350.
- Wardman, M., Toner, J.P., Fearnley, N., Flügel, S., Killi, M., 2018. Review and meta-analysis of inter-modal cross-elasticity evidence. *Trans. Res. A* 118, 662–681.

Further Reading

- Acutt, M.Z., Dodgson, J.S., 1995. Cross-elasticities of demand for travel. *Trans. Policy* 2 (4), 271–277.
- Dunkerley, F., Wardman, M., Rohr, C., Fearnley, N., 2018. Bus fare and journey time elasticities and diversion factors for all modes: a rapid evidence assessment. *RAND Res. Rep.* Available from: https://www.rand.org/content/dam/rand/pubs/research_reports/RR2300/RR2367/RAND_RR2367.pdf.
- Fearnley, N., Currie, G., Flügel, S., Gregersen, F.A., Killi, M., Toner, J.P., et al., 2018. Competition and substitution between public transport modes. *Res. Trans. Econ.*, doi:10.1016/j.retrec.2018.05.005.
- Ortúzar, J., de, D., Willumsen, L.G., 2011. *Modelling Transport*, fourth Ed. Wiley.
- Transport Research Laboratory, Centre for Transport Studies University College London, Transport Studies Unit University of Oxford, Institute for Transport Studies University of Leeds, Transport Studies Group University of Westminster, 2004. *Demand for Public Transport: A Practical Guide*. TRL Report TRL593, Crowthorne, Berkshire.
- Wardman, M., Hatfield, A., Shires, J.D., Ishtaiwi, M., 2019. The sensitivity of rail demand to variations in motoring costs: findings from a comparison of methods. *Transp. Res. A* 119, 181–199.

First-Best Congestion Pricing

Moez Kilani, University of Littoral, Côte d'Opale, Dunkerque, UMR 9221-LEM-Lille Économie Management, Lille, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	209
Static Models with a Single Road	209
Road Networks	211
A Simple Case	211
Complex Networks	212
Dynamic Models	213
Further Issues and Conclusion	214
Acknowledgment	215
References	215
Further Reading	215

Introduction

Roads are the scarce resource and their usage is a source of important external costs (congestion, emissions). Road pricing allows efficient usage of the available transportation infrastructure. First-best congestion pricing is a pricing scheme that maximizes social benefit, defined as the difference between users benefit and the social cost. When only congestion cost is considered, the social cost is equal to the aggregate users cost. But, when wider impacts, like emissions, are taken into account, the social cost is larger than the aggregate users cost since a wider population is concerned by the externality. Road pricing has now been implemented in several metropolitan areas around the world, and many others are considering its introduction under urban transportation reforms. First-best congestion pricing is rather a theoretical policy which describes the maximum attainable cost reduction from a transportation pricing reforms. It is specific to each trip and requires the availability of detailed data, which is usually difficult to obtain. Second-best policies, which require less information, are usually considered easier to implement in practice, but recent technological advances in information technologies and the availability of autonomous cars can raise interest in the implementation of first-best congestion pricing. Indeed, geolocation of vehicles and detailed description of the trips with respect to spatial (exact route) and temporal patterns (departure and arrival times) became much easier to obtain.

The interest of economists in congestion pricing is rooted in the work of Arthur Pigou who explained how the divergence between private and social costs could be a source of economic inefficiency. The advanced formulation and analysis of problems dealing with congestion pricing in transportation economics followed the progress made on two fundamental themes. The first one is the good and wide understanding of the equilibrium concept through the well-known Wardrop equilibrium (named after John Glen Wardrop). Consider a network where users have several possible routes to go from a given origin to a given destination. In a Wardrop equilibrium, travel costs on all used routes are equal and are smaller than travel costs on the unused routes. While this idea of equilibrium was expressed by Knight in the 1920s, it became widespread in the analysis of traffic equilibrium after the formulation by Wardrop in the early 1950s. It became important to develop analytical approaches to compute the Wardrop equilibrium, and this is where the contributions of authors like Martin Beckmann stand. They worked on the formulation and solution procedures of network equilibrium and the quantitative evaluation of the welfare loss due to congestion. The second important step is due to the contributions of William Vickrey and his seminal work on dynamic congestion, which brought an important community of researchers to develop modern dynamic models (Arnott et al., 1990).

This chapter is organized as follows. The section Static models with a single road is devoted to static models. We first consider the case of a single link and then extend the discussion, in section Road networks, to the case of a road network. In Section Dynamic models we consider dynamic congestion and explore the bottleneck model. We then conclude with some comments.

Static Models with a Single Road

Consider a road link with a fixed capacity. Each user entering and crossing this link gets a benefit with a magnitude that depends on the user's characteristics, the trip purpose, and the other transportation alternatives available to reach the destination. The monetary value of this benefit can be interpreted as the willingness to pay of using the road link. Consider a group of several users and sort them in a decreasing order with respect to the values of their willingness to pay. We obtain a demand function for using the road link. At the same time, each user entering the link slows down travel speed. When traffic flow is significantly below road capacity the impact is small, but it becomes important as traffic flow reaches and exceeds road capacity. So, user cost increases as the number of users increases. As long as there are users with a willingness to pay that exceed

Table 1 User equilibrium and optimum. SMC is the social marginal cost, e and o exponents refer to equilibrium and optimum levels, respectively

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Users	User benefit	Cumulative benefit	User cost	SMC	Total cost	Social benefit
n	$b(n)$	$B(n)$	$C(n)$	$S_{mc}(n)$	$nC(n)$	$W(n)$
1	52	52	12	12	12	40
2	50	102	14	16	28	74
3	48	150	16	20	48	102
4	46	196	18	24	72	124
5	44	240	20	28	100	140
6	42	282	22	32	132	150
7 ^o	40	322	24	36	168	154
8	38	360	26	40	208	152
9	36	396	28	44	252	144
10	34	430	30	48	300	130
11 ^e	32	462	32	52	352	110
12	30	492	34	56	408	84
13	28	520	36	60	468	52
14	26	546	38	64	532	14
15	24	570	40	68	600	-30

the user cost, traffic flows increases. Since users are sorted in decreasing order with respect to their willingness to pay, the equilibrium is reached when both quantities are equal; the last user who enters the road link has a willingness to pay which is equal to the average user cost. Users with smaller willingness to pay will not use the road link, they either use another alternative (other route or other mode) or they simply do not make the trip, which is likely to occur when the purpose of the trip is not so urgent.

Each user then compares his benefit (willingness to pay) and the user cost to decide whether to use the road link or not. This individual process is the origin of an excessive congestion cost. Indeed, each user will rise the average travel cost, but will ignore the impact of this increase on the other users of the road link. Under free access, the equilibrium is characterized by an excessive use with large external costs. Imposing a toll for the users of the road can restore the optimum number of users.

A numerical example is useful as a first illustration. Table 1 develops a case where users, in Column (1), are sorted in decreasing order with respect to their benefit from using the road link, given in Column (2). As the number of users increases, the (average) user cost, given in column (4), increases. As long as there are users with a benefit higher than user cost traffic flows increase. Comparing values in columns (2) and (4) the equilibrium traffic volume is obtained with 11 users. The benefit of user 11 is equal to 32, which is equal to the user cost. User 12 will not enter the road link since his benefit is 30 and he faces a cost of 32. To see how this equilibrium solution compares to the optimum we need to compute the social benefit as a function of the number of users. In column (3) we give cumulative benefits, based on values in column (2), and in column (6) total cost is computed. The social benefit is the difference between these two quantities and is given in column (7). Social benefit reaches a maximum value with 7 users, which means that the equilibrium does not correspond to the social optimum.

To understand this result, it is useful to consider values in column (5) corresponding to the social marginal cost (SMC). The values of this column are computed as follows. For the first user, the SMC is equal to the user cost, 12 in this case. For the second user, the SMC is equal to the corresponding user cost, plus the impact of his entry on the user cost of the first user, or $14 + (14 - 12) \times 1 = 16$. For the third user, the SMC corresponds to the user cost plus the impact of his entry on the two former users, or $16 + (16 - 14) \times 2$. For the general case, the SMC of user n is $S_{mc}(n) = C(n) + (C(n) - C(n-1))(n-1)$. Since user cost $C(n)$ is increasing in n , $S_{mc}(n)$ is higher than $C(n)$. A simple computation shows that the SMC can be written in the following recursive form: $W(n) = W(n-1) + b(n) - S_{mc}(n)$.

So, a new entry will increase the social benefit when the benefit of the new user is higher than the corresponding social marginal cost. Since $S_{mc}(n)$ is higher than $C(n)$, the optimum number of users is smaller than the equilibrium number. From this expression of the social benefit, an entry is socially desirable when the benefit for the user is higher than the social marginal cost, not the private marginal cost. We see then why it is optimal to have 7 users in this road link.

Under free access the road link is overused. The optimum level can be reached by imposing a road toll equal to the difference between the SMC and the user cost: $S_{mc}(n^o) - C(n^o)$, where n^o denotes the optimum number of users. In this example, a toll equal to 14 will reduce users from 11 to 7, and then decentralizes the optimum.

This example can be developed in a continuous and more tractable version. Let n now denote a continuous measure of the number of users. The user benefit in Column (2) is then a decreasing smooth function $b(n)$ as illustrated in Fig. 1. The cumulative benefit in Column (3) is $B(n) = \int_0^n b(z) dz$ and corresponds to the area between the curve of $b(n)$ and the x -axis. The user cost $C(n)$ is an increasing function of n . The user equilibrium is obtained when the user benefit is equal to the (average) user cost. So, the equilibrium number of users, denoted n^e , satisfies $b(n^e) = C(n^e)$. Total welfare is the difference between $B(n)$ and total cost $nC(n)$, that

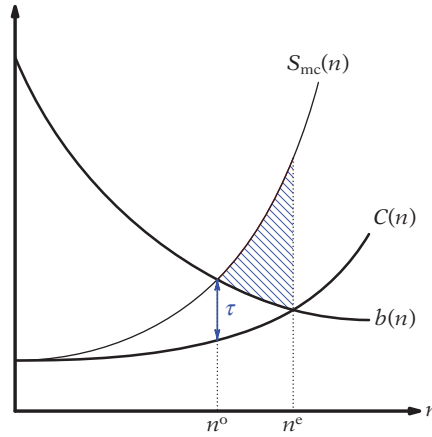


Figure 1 First-best congestion toll.

is, $W(n) = B(n) - nC(n)$. The optimum is reached when $W(n)$ is maximized over n (optimum traffic level), and the first-order condition is $W(n) = 0$. So, the optimum number of users, denoted n^o , satisfies

$$b(n^o) = \underbrace{C(n^o) + n^o C'(n^o)}_{S_{mc}(n^o)} \quad (1)$$

The equilibrium and optimum are given in **Fig. 1**, where the shaded area corresponds to the reduction of the net external cost when users are reduced from n^e to n^o . Net external cost is used to refer to the difference $nC(n) - b(n)$, that is, external costs net of the users benefits. The left hand side of eq. (1) is the benefit of user n^o and the right hand side is the sum of the private cost of the same user and the external marginal cost he generates. Both equilibrium level and the optimum level of users are uniquely defined, since $b(n)$ is decreasing and $C(n)$ is increasing. We also have $n^o \leq n^e$ because $C'(n)$ is positive, under the unpriced equilibrium there is an excessive use of the road. The decentralization of the optimum is possible through first-best congestion toll $\tau = n^o C'(n^o)$. From eq. (1), this toll will add the social cost to the private cost, or in other words it will internalize external costs.

For simple examples, a linear formulation of the user cost can be used, but for an empirical application a more realistic formulations are used. On a given road link, as we have mentioned above, travel time does not change so much when traffic flow is significantly below the road capacity, but increases substantially when traffic flow increases beyond critical level of capacity. The Bureau of Public roads (BPR) formulation is frequently used to express travel time as a function of traffic flow. In this case travel time on a given link is

$$t(n) = t_f \left(1 + a \left(\frac{n}{K} \right)^b \right), \quad (2)$$

where t_f is free-flow travel time, K the link capacity, and a and b positive parameters. In practice, $a = 0.15$ and $b = 4.0$ are the suggested values. For a given level of traffic flow, n , the nonmonetary part of travel cost is obtained by multiplying travel time [from Eq. (2)] by the value of time. The generalized travel time is then obtained by adding to this quantity the monetary travel cost. The example in **Fig. 1** has been constructed on the basis of this formulation.

Road Networks

The results obtained in the last section extend to road networks. We start with the case of a simple network to illustrate how the main purpose of the road toll is to reach an optimum distribution of users on the road network, and not necessarily to reduce urban mobility. We then comment on the case of a complex network where the computation of the optimum may be challenging.

A Simple Case

The analysis of external costs in a simple network provides a broader and more accurate description of the impacts of road pricing. In the case of a single link, a first-best congestion toll reduces traffic flow from n^e to n^o (**Fig. 1**), and it may be perceived as a policy against mobility of goods and passengers. Indeed, those who argue against road pricing frequently touch on similar arguments. By considering a network where each user chooses between several possible routes or transport modes it is described further that road pricing does not reduce the number of trips but rather induces an efficient use of the available infrastructure.

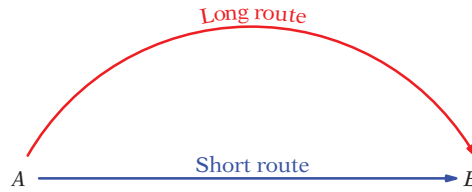


Figure 2 A simple network.

Consider the situation given in Fig. 2. A group of N drivers travel from A to B . There are two routes, a short but congested one, and a long and noncongested one (large capacity). To simplify the computation of the travel costs, we assume that only travel time matters. So, the objective of each driver is to reduce his travel time.

The shorter route is narrow and travel time significantly increases when the number of users increases. To consider a particular example, we assume that travel time on this route is $10 + n/5$ (expressed in minutes), where n is the number of users of the shorter route. The user may also travel through the longer route. This road has a large capacity and travel time does not significantly depend on the number of users and is assumed fixed and equal to 24 mins. At equilibrium, when both roads are used travel times are equal. For example, with $N = 100$ we obtain $n^e = 70$ confirming that most users choose the shorter route. The optimum number of users minimizes total cost, which is the sum of the total cost of users of the shorter road, equal to $n(10 + n/5)$, and the total cost of the users of the longer road $(N - n) \times 24$. Doing so, we easily find that total cost reaches a minimum when $n^o = 35$. In this case travel time (user cost) on the short road is 17 mins, while it is 24 mins on the longer route. Now, this optimum configuration can be decentralized by imposing a (first-best) congestion toll. The toll level can be adjusted so that only 35 drivers continue to use the short route. The corresponding value of the toll is equivalent to the monetary value of 3 mins.

By comparing the equilibrium and the optimum in this example, it is important to notice two key features in the solution. First, all users still make the trip and the first-best congestion toll is not harmful for anyone. No user has a travel cost higher than 24 mins since, for each user, the option of the longer route is always available. It is then clear that congestion toll does not prevent users from making their trips but induces an efficient distribution of flows over the network. Second, the revenue from the first-best congestion toll is equivalent to the benefit obtained by the users of the shorter route when their number is n^o . With this toll, all users have a generalized cost equal to 24, the same cost under equilibrium, but the government collects a toll revenue, which corresponds to the monetary value of the external costs produced under the unpriced regime.

Complex Networks

In a large network the computation of the equilibrium and the optimum become a challenging task. The network can be represented as a directed graph. Users are grouped with respect to their origin-destination pairs. Let g indicate the group of users, $g \in G$. Each group has one or several possible routes, a successive sequence of adjacent arcs, to travel from his origin to his destination. Let $c_l(u_l)$ denote the average user cost on link l when the number of users on that link is u_l . We have, $u_l = \sum_g u_l^g$, where u_l^g is the number of users from group g traversing link l .

The equilibrium conditions state that, for any group, generalized cost on all used routes are equal and are smaller than generalized costs on unused routes. The resulting set of equality and inequality constraints is not easy to solve, except in simple situations like the example examined in the last section. An alternative approach is to notice that the equilibrium problem is the solution of the nonlinear programming problem where the objective function given in Eq. (3) is minimized subject to the constraints that all users reach their destination and that the number of users on each link is nonnegative:

$$\min \sum_l \int_0^{u_l} c_l(z) dz. \quad (3)$$

Given a distribution of flows on the network, the total cost can be computed from $C(f) = \sum_l c_l(u_l) u_l$. The optimum is a distribution of flows $f = (u_l)$ that minimizes total cost $C(f)$. This problem can be solved as a multi-commodity problem, and it is clear that the equilibrium does not, in general, yield the optimum cost as we have shown in the simple network earlier. The optimum solution, which corresponds to an optimum distribution of flows on the network, can be decentralized through a link-specific toll similar to that described for the case of a single link. If we denote the user cost on link l by C_l and the number (flow) of users at the same link by u_l , Correa and Stier-Moses (2011) show that the social optimum is obtained when a toll

$$\tau_l = u_l C'_l(u_l) \quad (4)$$

is imposed on the user of link l . Notice that this is straightforward extension to the first-best corresponding to the case of a single link. The general idea underlying first-best congestion pricing is that each user is charged exactly the total external cost he generates.

Dynamic Models

In a dynamic model, not only travel time but also other features of the time pattern of the trip are explicitly taken into account. Users prefer a shorter travel time and avoid early or late arrivals (schedule delay) to work. They may prefer early to late arrivals. Modeling congestion in a dynamic framework is not an easy task and usually complex mathematical tools, like partial differential equations and optimal control problems with delays, are used to compute the equilibrium and the optimum (Arnott, 2013). The bottleneck model is an exception since it brings several properties entailing dynamic congestion while being mathematically tractable and intuitive. We describe here a simple version of the model that we use to derive time-varying tolls.

Let variable t denote time and consider a homogeneous group of users who prefer to arrive to destination at time t^* . We may think of the morning rush hour where t^* corresponds to the time when work starts. Generalized cost is the sum of travel time and schedule delay cost. Each unit of travel time costs α . A user arriving at $t < t^*$, early arrival, incurs a schedule delay cost equal $\beta(t^* - t)$. A user arriving at $t > t^*$, late arrival, incurs a schedule delay cost equal $\gamma(t - t^*)$. Commuters are assumed to prefer early to late arrivals. They also prefer higher travel time to later arrivals, but prefer earlier arrivals to higher travel time. These assumption are summarized in the condition $\beta < \alpha < \gamma$, and this model is frequently referred to as the $\alpha\beta\gamma$ -model.

Each user has to pass the bottleneck to reach the destination. The bottleneck has capacity of k vehicles per unit of time. When the number of users arrive at smaller rate the bottleneck is traversed instantly, otherwise k vehicles per unit of time exit the bottleneck. When arrivals to the bottleneck occur at a rate higher than k , the capacity is reached, a queue with first-in-first-out service forms. This is the only source of delays in the network. We then assume, without loss of generality, that the trip length is zero so that travel time is equal to the (possible) delay in the bottleneck. Notice that, in this simplified network, if there are no delays in the bottleneck travel time is zero. The population of the homogeneous users can be decomposed into two segments: those who arrive early and those who arrive late. Let $\tilde{t}$ denote the departure time of the user who arrives exactly on time.

The departure of the first user occurs at $t = t_0$ and the arrival of the last user occurs at $t = t_1$. Let $I = (t_0, t_1)$. Users departing at the endpoints of I do not incur delay to pass the bottleneck, but do incur a schedule delay cost. The total number of users is N , and since they all pass the bottleneck between t_0 and t_1 , we have $N = (t_1 - t_0)k$. We focus on an equilibrium where entry rates are constants and, respectively, equal to ρ_1 when $t < \tilde{t}$ and equal to ρ_2 when $t > \tilde{t}$. Given that users departing at t_0 and t_1 do not incur any delay and have the same generalized cost we have $\beta(t^* - t_0) = \gamma(t_1 - t^*)$. Using the fact that all users depart and arrive between t_0 and t_1 we have

$$t_0 = t^* - \frac{\gamma}{\beta + \gamma} \frac{N}{k}$$

$$t_1 = t^* + \frac{\gamma}{\beta + \gamma} \frac{N}{k}$$

Let $\delta = \beta\gamma(\beta + \gamma)$. These two users incur the same schedule delay cost

$$\beta \frac{\gamma}{\beta + \gamma} \frac{N}{k} = \gamma \frac{\gamma}{\beta + \gamma} \frac{N}{k} = \delta \frac{N}{k}$$

which is equal to their generalized cost. At equilibrium, this is the generalized cost for all users. At time t , the number of users in the queue is given by

$$n(t) = \begin{cases} (\rho_1 - k)(t - t_0), & \text{if } t_0 \leq t \leq \tilde{t} \\ (\rho_1 - k)(\tilde{t} - t_0) + (\rho_2 - k)(t - \tilde{t}), & \text{if } \tilde{t} < t < t_1. \end{cases} \quad (5)$$

The generalized cost for a user departing at time $t \in I$ is the sum of travel time cost and schedule delay cost. It is then given by

$$C(t) = \alpha \frac{n(t)}{k} + \beta \left(t^* - t_0 - \frac{n(t)}{k} \right) \quad (6)$$

since he has to wait for $n(t)$ users to pass the bottleneck. At equilibrium, all users have equal generalized costs. The equilibrium condition is then $C(t') = C(t'')$ for any $t', t'' \in I$. When this condition is applied for $t', t'' \leq \tilde{t}$ and $t', t'' > \tilde{t}$, respectively, it yields

$$\rho_1 = \frac{\alpha}{\alpha - \beta} k \quad \text{and} \quad \rho_2 = \frac{\alpha}{\alpha + \gamma} k. \quad (7)$$

So, at the first period, from t_0 to $\tilde{t}$ entries are higher than bottleneck capacity and a queue forms at a rate $\rho_1 - k > 0$ reaching its maximum at $t = \tilde{t}$. After this threshold, the queue starts to decrease at rate $\rho_2 - k < 0$ until it vanishes at time t_2 .

The delay caused by queuing is a source of welfare loss. This distortion can be removed, if each user arrives to the bottleneck exactly at the time where he exists the bottleneck under the equilibrium. In that case, there are no queuing and no delays. This departure pattern can be enforced through a time-varying toll given by

$$\tau(t) = \begin{cases} \beta(t - t_0) & \text{if } t \leq \tilde{t} \\ \gamma(t_1 - t) & \text{if } t > \tilde{t} \end{cases} \quad (8)$$

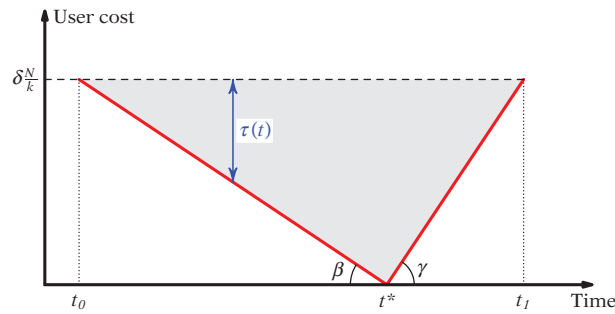


Figure 3 First-best time-varying toll.

This toll is highest during the peak period near t^* , and decreases as we move away from preferred arrival time. With this toll, travel cost of each user remains unchanged, but departure times are adjusted so that the arrival rate to the bottleneck is equal to its capacity k . First-best congestion toll removes the delays and the revenue it generates corresponds to the external cost generated under the unpriced equilibrium.

This problem is illustrated in [Fig. 3](#). For each t , the distance between the x -axis and the red curve corresponds to the schedule delay cost. The toll adds to this cost to reach a constant user cost equal to $\delta N/k$. The amount of the toll corresponds to the delay incurred by all the users in the bottleneck under unpriced equilibrium. The area above the red curve and below $\delta N/k$ line corresponds to the toll revenue (the shaded-area), which as we have noticed earlier, measures the external cost in the unpriced equilibrium. It is clear from [Fig. 3](#) that the triangle corresponding to waiting time cost (external cost) is half the total cost.

Finally, we show how dynamic tolls can be simply described through a two-player game. Each one of the two players wants to cross a limited capacity bridge. There are two time slots when crossing is possible, s_1 and s_2 , and both players prefer to cross at s_1 . Only one player can cross during each slot. Let b_i denote the benefit of crossing during slot $i = 1, 2$, and let $\bar{b} = b_1 - b_2$. If both players choose slot s_1 , only one, randomly chosen, crosses and the other waits for slot s_2 . This wait time is a social loss. If $\bar{b}$ is sufficiently high, the only equilibrium entails waiting time, since the players will choose slot s_1 . A more efficient situation is obtained when a toll, equal to $\bar{b}$, is imposed on the user who crosses during slot s_1 . Under the new equilibrium both players will be indifferent between the two slots and the configuration where each player chooses a distinct slot becomes an equilibrium. This is how the time-varying toll described above works for a continuum of players.

It is important to notice that the dynamic toll does not reduce the number of trips. Instead, it motivates the drivers to adjust their departure times. This observation is similar to the remark we have stated for the static case when we move from the single route to the case of a network where each pair of origin and destination is linked by several routes: the number of trips does not decrease but traffic is diverted to less congested routes. Most realistic situations correspond to a dynamic framework with several possible routes for each trip. In that case, the effectiveness of the first-best congestion toll will indeed depend on the flexibility of the drivers in various dimensions, especially the spatial and temporal ones.

Further Issues and Conclusion

This overview of first-best congestion pricing has not addressed all issues and possible extensions, but has deliberately focused on fundamental features of road pricing. The case of heterogeneous users has been shortly mentioned above when we discussed road pricing in a complex networks. Users have distinct origin-destination pairs and are not, by so, homogeneous. More generally, it is also possible that users have distinct desired arrival times, distinct trip purposes or may use distinct vehicles as in [de Palma et al. \(2008\)](#). In most cases, taking into account these extensions will involve a more elaborated analytical models but there are no insuperable difficulties. The essence of first-best pricing, that we have encountered along all the examined situations, is that each user should pay the external costs he generates during his trip. This is the basic ingredient for an efficient allocation of roads.

The correct value of the first-best congestion toll, depends on data related to users and the spatial and temporal patterns of the trip, among other details. In practice, this data is not easy to obtain and one may consider an approximate value of the first-best toll as in [Kilani et al. \(2014\)](#). In the dynamic case, instead of a smooth time-varying toll it is possible to consider an approximation through a step function with flat toll on given time intervals. Implementing a toll for off-peak period and peak period is the simplest approximation. As the number of intervals increases, with optimized toll levels within each interval, the social loss is reduced and at the limit we converge to the first-best situation. [de Palma et al. \(2005\)](#) show that fine tuned tolls provide a substantial benefit by comparison to flat tolls.

Some years ago the computation of the exact value of the first-best congestion toll was simply impossible, because it requires perfect information on all the trips. But, this situation is changing. Indeed, with the growing technological innovations in the geolocation and connectivity of cars, the autonomous car, it is easy to trace back the exact route, the departure time and the arrival time for any trip. The implementation of first-best congestion toll will no longer be limited by scarcity of data, but mainly by the acceptability of users and the goodwill of decision makers.

As a final comment recall that in the bottleneck problem, there are no external costs under the optimum (no delays). This is not a universal characteristic of the optimum, however. Indeed, as shown in the case of one road link and a continuum of users (**Fig. 1**), some congestion is prevalent in the optimum. The absence of congestion in the bottleneck is mainly the result of the assumption that external cost is only in the waiting queue and not on the road.

Acknowledgment

I am thankful to the Editor, Maria Börjesson, for her useful comments and to Ngagne Demba Diop for his help to transfer the document from Latex to Word.

References

- Arnott, R., 2013. 'A bathtub model of downtown traffic congestion'. *J. Urban Econ.* 76, 110–121.
- Arnott, R., de Palma, A., Lindsey, R., 1990. The economics of the bottleneck. *J. Urban Econ.* 27, 111–130.
- CORREA, José R. et STIER-MOSES, Nicolás E. Wardrop equilibria. *Wiley encyclopedia of operations research and management science*, 2011.
- de Palma, A., Kilani, M., Lindsey, R., 2005. Congestion pricing on a road network: a study using the dynamic equilibrium simulator metropolis. *Transp. Res. Part A: Policy Prac.* 39 (7–9), 588–611.
- de Palma, A., Kilani, M., Lindsey, R., 2008. The merits of separating cars and trucks. *J. Urban Econ.* 64 (2), 340–361.
- Kilani, M., Proost, S., van der Loo, S., 2014. Road pricing and public transport pricing reform in paris: complements or substitutes? *Econ. Transp.* 3 (2), 175–187.

Further Reading

- Beckmann, M., McGuire, C.B., Winsten, C.B., 1956. *Studies in the economics of transportation*. Yale University Press, New Haven, Connecticut.
- Lindsey, R., 2012. Road pricing and investment. *Econ. Transp.* 1 (1), 49–63.
- Pigou, A.C., 1912. *Wealth and Welfare*. Macmillan and Company limited, London.
- Quinet, E., Vickerman, R.W., 2004. *Principles of Transport Economics*. Cheltenham, Northampton, MA.
- Small, K.A., Verhoef, E.T., 2007. *The Economics of Urban Transportation*. Routledge, United Kingdom.
- Vickrey, W.S., 1969. Congestion theory and transport investment. *Am. Econ. Rev.* 59 (2), 251–260.

Ethical Aspects—Can We Value Life, Health, and Environment in Money Terms?

Jose M. Grisolia^{*,†}, Ken Willis[‡], ^{*}Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain; [†]University of Nottingham Ningbo, Ningbo, China; [‡]University of Newcastle, Newcastle, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	216
The Neoclassical Approach and its Ethical Problems	216
Environmental Goods	217
The Value of Statistical Life	217
The Value of Time	218
Utilitarianism Versus Deontological Ethic	218
Intrinsic or Absolute Value Versus Economic Value	218
VSL, Healthcare and Scarce Resources: A Standard is Essential	218
Altruism and Environmental Goods	219
Conclusion	220
References	220
Further Reading	220

Introduction

Cost-benefit analysis (CBA) accounts for all benefits and costs of a project, measuring every item in monetary terms. It is often the case that consequences of the project are non-market goods and, therefore, there is no price to insert in the analysis (i.e., saved time, lives, landscape, and environmental goods, which would be the main non-market goods relevant for transport appraisal). To account for a social value of these concepts it is necessary to find a link between these goods and existing market goods. Economists have developed methods to obtain these estimations applying Revealed Preferences (RP) and Stated Preference (SP) questionnaires such as Choice Experiments (CE) and Contingent Valuation (CV).

People might feel uneasy with the idea that life, environment, and time can have a value. In general, there is a tendency to reject any valuation of non-market goods, in particular the value of statistical life (VSL) and the value of the environment. The first problem arises with the mere institution of price. Bringing those concepts to the market arena seems to challenge our moral concerns because they do not seem legitimate objects with which to trade. This view is widely popular, and reflects a profound dissociation between economic analysis and society's perspective. If we are using these techniques to come up with a social value how is possible that most people reject it?

For instance, in relation to the VSL the Indian Environmental Minister Kamal Nath in 1993 stated:

"... the absurd and discriminatory Global Cost/Benefit Analysis (. . .) we unequivocally reject the theory that the monetary value of people's lives around the world is different because the value imputed should be proportional to the disparate income levels of potential victims . . . it is impossible for us to accept that which is not ethically justifiable, technically accurate or politically conducive to the interests of poor people as well as the global common good" (Grubb et al., 1999, p. 306).

In relation to environmental goods people tend to think that wildlife, biodiversity, and landscape are priceless or, at least, difficult to measure in monetary terms. It is often the case that nature is contemplated from the perspective of rights, non-related to the enjoyment and services that nature produces to human beings. This makes it more difficult to accept the idea of value. As Elton (1958), states: "There are some millions of people in the world who think that animals have a right to exist and be left alone, or at any rate that they should not be persecuted or made extinct as species. Some people will believe this even when it is quite dangerous to themselves. p. 143–45."

These and other examples demonstrate that ethical objections to monetary valuation of non-market goods are widespread. In this chapter, we will provide an answer to these concerns following this structure: first, we describe the neoclassical approach and its difficulties; second we explain why monetary value is ethically acceptable; and close with a section of conclusions.

The Neoclassical Approach and its Ethical Problems

Underlying any CBA is neoclassical economic theory, which is rooted in utilitarianism. The basic assumptions might be formulated as: individuals maximize their utility, which is obtained from a bundle of goods. The key aspect of the theory is the flexibility of the utility function, which includes all type of goods not only market goods but also environment, safety, exposure to air pollution, and any other non-market goods. At the same time, individuals are constrained by budget and time. Giving these limitations, in order to

maximize their overall utility individuals are willing to make trade-offs between different goods. This opens the space for extracting willingness to pay, which is the marginal rate of substitution between income and any other good. Hence, we are assuming that individuals are willing to trade between goods according to their preferences and constraints.

In practical terms, economists observe secondary markets that keep a substitution or complementary relation with the non-market goods. For instance, VSL could be estimated by observing wage differences for jobs that have different levels of risk of a fatal event, *ceteris paribus*. Environmental goods could be estimated by accounting for all costs incurred to visit a National Park; air pollution can be valued through house price differences and the cost of moving from polluted to non-polluted areas; or from the price of goods that prevent its consequences such as masks and air purifiers. By the same token, the value of time is indirectly observed by comparing choices of different modes of transport that comprises a trade-off between time and money. All of these are RP methods that might be combined with SP.

Although environment, life, and time are obtained based on the same principles, their ethical consequences are not exactly the same. We will present them separately

Environmental Goods

In CV experiments, it has been observed that a fraction of individuals provide protest responses. These answers could be zero bids, or infinite bids, or simply a refusal to offer any monetary value. Although protest responses would have different motivations, one possible interpretation is that these individuals are reluctant to accept any exchange value for these non-market goods. Indeed, when non-traders are asked the reason of their zero bids, often the answer is “I refuse to assess biodiversity (or any other environmental good) in monetary terms.”

Economists offer another explanation for non-traders: they might show lexicographic preferences. If people have lexicographic preferences for environment, they refuse to trade between environmental goods and market goods, because the preference for environmental goods is absolute. No matter what is the gain in terms of other market goods, they would always select the bundle with the higher level of environmental goods. These preferences cannot be represented with a utility function. It might be also the case of weak lexicographic preferences where trade is possible but only beyond certain threshold (for instance, certain level of income).

Yet it might be the case that individuals reject any compensatory rule between environmental goods and the rest of the goods in the bundle, simply because they consider that environment has *rights* and, hence, should be protected regardless of the cost or benefits for the society. In this line environmentalists usually allege that non-market valuation techniques are anthropocentric, that is, based on utility (of individuals) disregarding nature as an object that generates certain services.

For some individuals environment cannot be categorized as a good and therefore, exchange with other goods is not possible. It is simply that not all goods are exchangeable. As Throsby wrote:

“First, I may acknowledge that a good has value to me, but I cannot trade this benefit for other goods – in other words, I cannot meaningfully represent the benefit I gain from this good in monetary terms. For example, I may experience some pleasure at my sense of identity as a human being (sharing a common humanity with my fellow citizens), but neither I nor they are likely to be able to express the value of this identity in monetary terms, since it is not exchangeable for other goods” (Throsby, 2003, p. 278).

The Value of Statistical Life

The VSL is usually obtained from three sources: wage differences in riskier jobs; market for goods that increase safety (smoke detectors, safety features in a vehicle); and SP surveys that comprises the risk of a fatal event related to risky behavior (diet, smoking, pollution). Initially we should face the same problems that arise with environment: that is, the idea of using money to value life seems morally repulsive. However, if we actually understand that it is not the value of a life but the willingness to pay to prevent a fatality, this rejection might not be justified.

This does not mean that there are not many ethical issues around the concept of VSL. Firstly, the VSL depends on market conditions: this value changes with recessions, market conditions, and consumer preferences. Most of the VSL estimates come from analysis of fatalities in the labor market. As a result, estimations of VSL tend to increase with time partly because workers are more aware of the risk in their working conditions.

Secondly, since VSL is a budget limited estimation it is expected to be lower in developing countries; that is, life is less valuable in poorer countries and for poorer people. An early paper from Miller (2000) demonstrated the existence of a correlation between GDP per capita and the VSL where VSL can be approximately 120 times GDP per capita. This wide range of values suggests that the most efficient allocation of resources is to pollute in developing countries since there life has less value. As Miller points in his article: “In 1991, World Bank Vice President Lawrence Summers suggested that it might be economically efficient to transfer pollution to less developed countries where the productivity losses of ill health would be smaller” (p. 16).

Most people would agree that life should have a value per se, no matter who and where. Yet the implications of this idea are ethically unacceptable. Do we have to spend the same resources to save the life of a child versus the life of an elderly person? In transport projects, VSL is usually applied per person regardless of age and other conditions. In contrast, health care policy typically uses the quality of adjusted life year (QALY), which recognizes that the value of life decreases with age.

The Value of Time

Travel time savings is usually the most important benefit generated from any transport project. The ethical implications are, somehow different. We do not see here people opposed to the idea of measuring time with money since everybody does it –what is working after all?. However, estimations of the value of time (VT) could be *regressive* in the sense that they benefit to a minority of the population that are usually the richer. The idea is straightforward: since the VT comes from revealed and stated preferences, these are budget-limited, and higher VT will arise for those who have more money. Indeed, five decades of investigation of the VT, shows a higher WTP for faster modes of transport that are correlated with income. The greater the income, the faster the mode, the higher the VT. Hence, decisions about transport projects might end up benefiting the most advantaged group of the society. For example, lower income groups are often users of buses who have a lower WTP to save time. Hence, public investment might accrue to the development of highways and roads rather than more buses. To overcome this problem a so-called “ethical value of time,” based on average income level, has been introduced.

Utilitarianism Versus Deontological Ethic

Utilitarianism is a philosophical school founded by the English philosopher Jeremy Bentham in the 18th century. This doctrine considers the rightness of any act as dependent on its consequences—that is why it is also called consequentialism or Conditional Imperative: consequences are the only thing that matter. Thus, an action that increases the happiness of individuals is right, and vice versa. This is the origin of utility analysis and its maximization as the basis of any welfare assessment. This is also the idea behind CBA, which considers all costs and benefits of a project, that is, its *consequences*, to determine if it is good or bad, and whether the project should to go ahead or not.

In contrast, Deontology is a rule-based ethics: Any course of action is right if it is according to a *general principle of ethics*. We have to do what is right regardless of its consequences. What is “right”? What are the “rules”? It could be the Ten Commandments, but following Kant we can use his Categorical Imperative. One of its most well known tenets is this: act in a way that everybody is a goal and not a means. The corollary of this rule is that everybody counts and everybody matters equally. More or less life cannot be exchange for any other good since this would be using life as a means. Likewise for environmental goods. The whole idea of CBA would be absolutely rejected. The typical example to understand these two different doctrines is this: a terrorist has placed a bomb in a school. Torturing him and obtaining information to save the children is acceptable according to Utilitarianism; but the deontological ethic would rule against this because torture is immoral, no matter the consequences.

Intrinsic or Absolute Value Versus Economic Value

Parallel to this discussion of ethics principles is the idea of value: economic versus *intrinsic value*. Economic value stems from utility, using stated or revealed preferences. Intrinsic value is the value per se of environment or life (let us exclude time here). Environmentalists would consider that wildlife is not an instrument for human enjoyment but an end itself. In this sense, wildlife would have an intrinsic value. Another perspective is considering that an object has an intrinsic value through the properties that the good possesses: objective and not subjective measurement.

Intrinsic value comes from deontological ethics. That is, environmental goods cannot be used as exchange for something else (money) since this would mean that these good are treated as means and not goals. In this sense, any economic value would be unacceptable. Hence, the discussion could be about *intrinsic* versus *economic* (or instrumental) value.^a Which one should guide our actions as a society? Which one is ethically acceptable?

As mentioned before environmentalists consider that nature has an intrinsic value, that is, a value that is independent from any human. Even if we accept that this is the case—and precisely because this is the case, that is, a value, which is independent to human affairs—there is no reason to think that this intrinsic value brings any obligation to human beings. As O’neill (2002) states: “The defender of nature’s intrinsic value still needs to show that such value contributes to the well-being of human agents.”

Likewise, a permanent appeal to intrinsic value does not yield practical results. Again, it leaves the society without appraisal tools for any project. Indeed, during recent decades, this recurrent moral argument about the intrinsic value of nature has been unable to stop the degradation of ecosystems. In contrast, instrumental or economic value has been used to apply practical tools that have curbed, at least partially, the environmental catastrophe.

VSL, Healthcare and Scarce Resources: A Standard is Essential

Following the intrinsic value argument, we would reach the conclusion that life is ethically incommensurable and its value, if exists, would be the same for everybody. In contrast the VSL is based, mostly, on individuals’ WTP to prevent a fatality. But, what would be the result if we consider the lives of other people? Recently an article was published in *Nature* about finding rules for self-driving (Awad et al., 2018) where participants are forced to choose which life to spare in case of a hypothetical accident. It is the largest survey

^a In Kant words: “Everything has either a price or a dignity. Whatever has a price can be replaced by something else as its equivalent; on the other hand, whatever is above all price, and therefore admits of no equivalent, has a dignity. But that which constitutes the condition under which alone something can be an end in itself does not have mere relative worth, that is, price, but an intrinsic worth, that is, a dignity.”

of ethical concerns ever taken with 2.3 million people around the world. Using 13 scenarios in which someone's death was unavoidable the results shows a ranking where age, gender, and even responsibility matters.

This article reflects people's mindset about the life of others, and its results clash with the basic ethic idea, present in many constitutions, that nobody should be discriminated against based on features such as age, gender, and so on. This outcome is contrary to the Kantian ethics notion that all life matters the same, and hence no discrimination is possible. In short, this research points that, sometimes, choices are unavoidable which makes it impossible to apply intrinsic values for life.

Another way to contemplate the morality of the VSL is to recognize the contrast between the growing demand for healthcare services and the permanent scarcity in the supply. The logical outcome is searching for efficient treatments and this implies that we need *standards* to come up with the right choices. Standards such as the VSL, or QALYs. The lack of such measurements would make impossible to make choices in healthcare. In real life choices are made constantly, revealing that this is how people actually think. As Calabresi (1965) pointed

"Our society is not committed to preserving life at any cost. In its broadest sense, this rather unpleasant notion should be obvious. Wars are fought. . . . But what is more interesting . . . is that lives are used up when the quid pro quo is not some great moral principle but 'convenience.' Ventures are undertaken that, statistically at least, are certain to cost lives. . . . We take planes and cars rather than safer, slower means of travel. And perhaps most telling, we use relatively safe equipment rather than the safest imaginable because –and it is not a bad reason – the safest costs too much."

In addition, we should add that the VSL is meant to create a standard that saves lives. It does not imply benefits arising from projects that kill lives. That is why the "value of a preventable fatality" (VPF) has been proposed as a more appropriate name.

Altruism and Environmental Goods

In relation to environment and, following utilitarianism, it is possible to create a framework of different values (Krutilla, 1967). First, *use-value* is that which individuals obtain from using environmental goods, even if it is a potential use. Aside from use-value, WTP for environmental goods might reflect individuals' ethical values, beliefs, and attitudes. This would affect the so-called *non-use or passive value*. When individuals behave altruistically, they incorporate the utility of others into their decisions. Their WTP for environmental protection will reflect their desire that future generations, or others, in general enjoy environmental goods. In this respect, there is a distinction between *non-paternalistic altruism* where individuals are willing to pay for any public good that reflects public preferences, and *paternalistic altruism* where this would only happen for certain goods that reflect the payer's preferences (for instance, environmental goods). There is evidence of existence value in the support of environmental NGOs and altruistic donations, aside from thousands of stated preference experiments.

Paternalistic altruism, also known as warm-glow represents for Kahneman and Knetsch (1992) the "purchase of moral satisfaction." Likewise individuals might obtain satisfaction from an environmental good not because of altruism, but because of the knowledge that this good exists—without obtaining any services from it. Simply knowing it exists makes them feel better. This is an *existence value*. The fundamental reason for this is sentimental: As expressed by De Groot et al. (2002): "Natural ecosystems and natural elements (such as ancient waterfalls or old trees) provide a sense of continuity and understanding of our place in the universe which is expressed through ethical and heritage-values. Also religious values placed on nature (e.g., worship of holy forests, trees, or animals) fall under this function-category."

This might not convince environmentalists, who care about environment not for the services that this good provides to him/herself or to others, but because of nature *itself*, understanding nature as a subject with rights. In this case, since it is a value that arises from a non-human perspective, it can be considered an *intrinsic* value.

Nonetheless, in all these cases individuals still behave in a selfish manner in the sense that, after all, they are maximizing their utility. A step further would be *genuine altruism* where the enjoyment of others does not bring any increase on the payer's utility (Sen, 1977). Table 1 summarizes this framework of values.

Coming back to the discussion of intrinsic versus economic value there is a possible compromise here since some individuals' interpretation of existence value might reflect an intrinsic value vision. Under some flexible interpretation—a non-anthropocentric value—this could be accepted as intrinsic since nature is not used as a means but as a goal. We should expect that, at least partially, stated preferences values reflect this vision. In fact, the Millennium Ecosystem Assessment states that "many people do believe that ecosystems have intrinsic value. To the extent that they do, this would be partially reflected in the existence value they place on that ecosystem, and so would be included in an assessment of its total economic value" (Millennium Ecosystem Assessment, 2003, p. 133–134).

Table 1 Environmental non-use values

Variation in individuals utility (payer)	Beneficiary	Preferences	Motivation	Name
Positive	Others	Others	Public goods	Pure altruism
Positive	Others	Payer	Purchase of moral satisfaction	Paternalistic altruism
Positive	None	Payer	Self-transcendence	Existence value
Positive	Nature	Payer	Environmentalism	Intrinsic value
Null	Others	Payer	Commitment	

Conclusion

An intrinsic value for environment might exist and, perhaps, co-exist with exchange value, but it must not undermine the use of economic value to organize social preferences. Intrinsic value arises from environment per se and, since it is a non-anthropocentric perspective, it cannot bring any obligation to human beings. Besides, the use of intrinsic value does not yield any solution for the environmental catastrophe. It is like asking for zero pollution, which is not a feasible option. The use of economic value in the environment provides a useful framework to make choices. Using intrinsic value instead would be impractical to manage.

However, deontological ethics might play a role since individuals would consider their beliefs in their WTP and hence, at least, partially it might be possible that their WTP reflects an intrinsic value for environment. The current practice of CBA and cost-effectiveness in healthcare brings a reasonable compromise with deontological ethics—counting VSL regardless of the characteristics of the person in transport projects, and applying QALY—, which considers age and life expectancy- for healthcare treatments.

Having said that, it is true that not everything can be valued in monetary terms as Throsby argues, but we need standards to make the right choice, to allocate scarce resources, and, in this sense, Utilitarianism is essential. The resource allocation argument is “just,” because the alternative is a dead end. It is unethical not to consider values (costs and benefits), because failure to do so is to ignore alternatives, which might improve the welfare of society.

As human beings, our actions constantly support the exchange argument. In relation to the value of life for instance, zero risk does not exist. Similarly, nobody can realistically ask for zero pollution, or face all the costs and sacrifices necessary to achieve such an outcome. We make trade-offs constantly, increasing or decreasing our risks and revealing our willingness to pay. Using an absolute value would bring to the world an ethically unacceptable tenet, where we would be unable to make choices between healthcare treatments.

In relation to the value of time, most objections are about the regressive calculations that are biased toward population segments with higher WTP. This is a valid concern and there are already possible solutions to mitigate this bias.

The opposition to non-market valuation reflects objections to market itself: to the use of money as a universal measurement, as a standard. The popularity of this position is rooted in a supposedly moral superiority of deontological ethics—we follow rules, we do not count-. Yet this position is impractical, and it is refuted every day with our own behavior as choice-makers, always trading, comparing and exchanging private goods for non-market goods.

References

- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Rahwan, I., 2018. The moral machine experiment. *Nature* 563 (7729), 59.
- De Groot, R.S., Wilson, M.A., Boumans, R.M., 2002. A typology for the classification, description, and valuation of ecosystem functions, goods and services. *Ecol. Econ.* 41 (3), 393–408.
- Elton, C.S., 1958. The reasons for conservation. In: *The ecology of invasions by animals and plants*, Springer, Boston, MA, pp. 143–153.
- Grubb, M., Vrolijk, C., Brack, D., 1999. *The Kyoto Protocol: A Guide and Assessment*. Earthscan, UK.
- Kahneman, D., Knetsch, J.L., 1992. Valuing public goods: the purchase of moral satisfaction. *J. Environ. Econ. Manage.* 22 (1), 57–70.
- Krutilla, J.V., 1967. Conservation reconsidered, *Am. Econ. Rev.* 57, 777–786.
- Millennium Ecosystem Assessment, 2003. *Ecosystems and Human Well-Being: Synthesis*, Island Press, Washington, DC.
- Miller, T.R., 2000. Variations between countries in values of statistical life. *J. Transp. Econ. Policy* 169–188.
- O'Neill, J., 2002. *Ecology, Policy, and Politics: Human Well-Being and the Natural World*. Routledge, UK.
- Sen, A.K., 1977. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs* 317–344.
- Throsby, D., 2003. Determining the value of cultural goods: How much (or how little) does contingent valuation tell us? *J. Cult. Econ.* 27 (3–4), 275–285.

Further Reading

- Calabresi, G., 1965. Fault accidents and the wonderful world of Blum and Kalven. *Yale Law J.* 75, 216.
- Davidson, M.D., 2013. On the relation between ecosystem services, intrinsic value, existence value, and economic valuation. *Ecol. Econ.* 95, 171–177.

Car Tolls, Transit Subsidies for Commuting, and Distortions on the Labor Market

Ioannis Tikoudis, Kurt Van Dender, Organisation for Economic Co-operation and Development (OECD), Paris, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	221
Two Polar Cases	222
The Standard Case: Labor Supply is Independent of Road Pricing	222
The Alternative Case: Strict Complementary Between Peak Hour Car Travel and Labor Supply	222
Weakening Strict Complementarity Between Labor Supply and Commuting	224
Two Commuting Modes	224
Two Trip Purposes	224
Weaker Links Between Commuting and Labor Supply	224
Spatial Considerations	225
Discussion	225
References	226
Further Reading	226

Introduction

Transport systems in many countries are heavily congested, particularly in urbanized areas and during peak hours. Congestion leads to longer, less predictable travel times and has been shown to have a large social cost. The monetized cost of time lost due to congestion in the US has been estimated at 0.9% of the country's GDP (Schrang et al., 2012). Including the value of wasted fuel raises this number to 1.7% of GDP. On top of this, one should consider the social cost from human exposure to the air pollutants generated by the combustion of fuel due to congestion. The latter accounts for 2.9 billion gallons of fuel annually. Taking the health cost of traffic-induced air pollution into account implies social costs that may well exceed 2% of GDP. Standard economic analysis argues that, in the absence of policy interventions, the level of traffic exceeds the socially desirable level. This happens because road users base their trip decisions exclusively on their own travel times and costs. They do not account for the effect of their decisions (where, how, and when to travel) on the travel times and costs of the other road users; that is, they disregard the external cost of road use.

The external cost of road use can be very high during the peak hours, due to the large number of vehicles affected by changes in travel times and travel time reliability, and because congestion effects are worse the more dense traffic is. This suggests that reducing peak hour congestion can result in large welfare gains, at least if that is done in a way that selectively eliminates trips whose social costs exceed their social benefits.

Road tolls that vary across time and space are the best way to reduce congestion, because they allow for a wide array of behavioral changes: change of the time at which a trip is undertaken, change of route, switch from private vehicles to public transport, change of residential or workplace location—or pay tolls and continue to drive under reduced congestion. In the long run, urban road pricing may contribute to creating more compact, less car-dependent urban forms. In contrast, traffic regulations, such as partial traffic bans, risk eliminating trips with a relatively low social cost. For instance, an odd–even system of car bans was introduced during 1990s in Athens, allowing vehicles whose license plate ends in an odd (even) number to enter a specific area in odd (even) days. Variations on this type of measure, which is not relevant to the social cost of the trip, have been implemented in Latin American, Chinese, and European cities.

The standard case for congestion charging abstracts from road users' trip purpose. It treats the demand for peak hour driving as a derived demand, and the willingness to pay for trips as a measure of their full social value. As a result, there is no need to account for the precise reasons people travel. However, a considerable portion of peak hour travel is undertaken to commute to or from work. For instance, in Auckland (New Zealand) commuting kilometers make up approximately 46% of the total distance traversed by motorised vehicles during the peak hours. A tax on peak hour driving could result in an increase of the effective tax rate on labor income. What does this mean for the net welfare impact of urban road pricing, if labor markets are subject to large tax-induced distortions? How strong is the interaction between the inefficiency related to congestion during peak hours and the inefficiencies resulting from high labor taxes? What are the implications for urban road pricing policies *per se*? What are the recommendations on how to deploy the revenues from congestion charges?

This article draws from economic analysis of the past two decades to discuss if and how interactions between labor markets and congestion alter the overall case for congestion charging. It starts, in Section 2, by introducing two polar cases: one in which congestion charges are economically indistinguishable from labor taxes and one in which the bases of the two taxes do not overlap. Section 3 considers intermediate cases. Section 4 weighs the evidence, asking what practical rules of thumb are best for the design of congestion charges.

¹Views and opinions expressed in this article are the authors and not necessarily those of the OECD. We would like to thank Ian Parry for comments on an earlier version. The usual disclaimer applies



Figure 1 Schematic depiction of the interaction between a labor income tax and a road toll. Left panel: the polar case in which the bases of the labor and road tax overlap completely; middle panel: the intermediate case, in which the two taxes interact partially; right panel: the standard polar case in which interactions between labor taxes and road charges are not considered.

Two Polar Cases

The Standard Case: Labor Supply is Independent of Road Pricing

The standard case refers to a setting in which the road and the labor tax do not interact. In that case, the two tax bases are disconnected, as illustrated in the right panel of Fig. 1. Economic efficiency requires taxing road use at its marginal external cost, since all relevant sources of economic inefficiency, such as taxes and price regulations, have been assumed away. A lump sum recycling of revenue is economically efficient, since all taxes that could interact with the road tax are at their optimal values, essentially zero. This is the case because there is no predetermined level of revenue to be raised. The standard case is depicted in Fig. 2.

The Alternative Case: Strict Complementary Between Peak Hour Car Travel and Labor Supply

Suppose that all morning peak hour travel is for commuting, and the evening peak consists exclusively of trips back home from work. All workers drive a car, so there is no public transport or carpooling. Furthermore, the starting time of the workday is fixed, so workers cannot avoid congestion or charges by adjusting departure time. Also, the length of the workday is fixed. Changes in labor supply can take place through working more or fewer days, but not through working more or fewer hours per day. All workers are homogenous in skills and earn the equilibrium wage in a competitive labor market.

In this setting, all traffic results exclusively from commuting, and commuting is strictly proportional to labor supply. We illustrate this in Fig. 3, in which every unit of labor supply (lower panel) corresponds to a unit of traffic (upper panel). As a result, the base of the labor tax is identical to that of the road toll (Fig. 1, left panel). The two taxes—labor tax and road toll—differ only in name, as

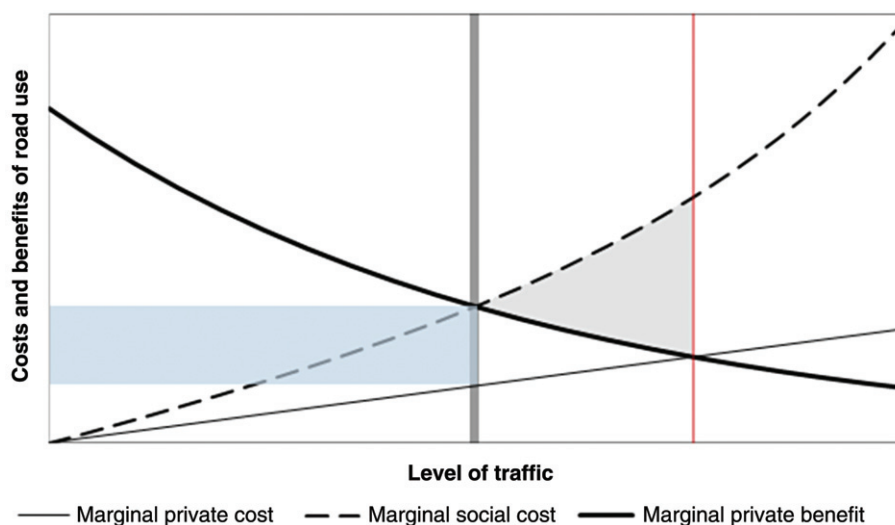


Figure 2 Efficiency-maximizing road externality taxation in absence of other relevant inefficiencies. Shaded triangular area: deadweight loss due to traffic-related externalities; Shaded rectangular area: the total revenue from road pricing; height of rectangle: the level of the efficiency-maximizing toll; thin vertical line: overall traffic level without intervention; thick vertical line: efficiency-maximizing level of traffic.

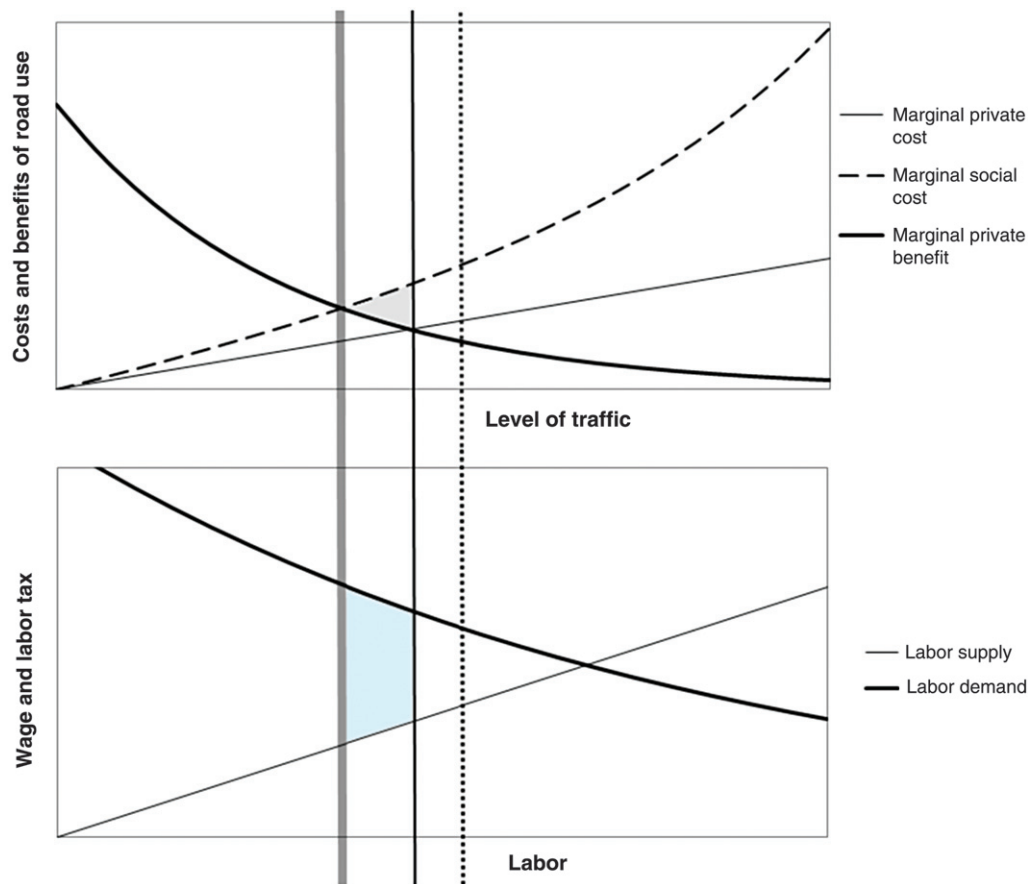


Figure 3 Road externality taxation under the perfect complementarity polar case. Upper panel: transport market, as explained in Fig. 2; Lower panel: labor market; thin vertical line: the equilibrium traffic and labor supply with a labor tax of 40%; thick vertical line: the equilibrium traffic and labor supply with a labor tax of 40% and a toll equal to the marginal external cost of road use; triangular shaded area (upper panel): gains from marginal external cost pricing on the road; trapezoid shaded area (lower panel): labor market losses from marginal external cost pricing; dashed vertical line: the efficiency-maximizing level of traffic, achieved with a negative toll on the road.

their economic effects are identical. Both affect the same behavioral margin, namely the decision on how many workdays to supply. Consequently, the sum of the road toll and the labor tax determines the size of market distortions, but the separate components do not matter. That is, increasing the road toll has no economic effect if the labor tax is reduced by the same amount.

In this case, the road toll reduces welfare if the labor tax exceeds the marginal external cost of road use, unless the labor tax is reduced to compensate for the road toll. We illustrate that negative outcome of introducing marginal external cost pricing on top of a rigid labor tax in Fig. 3. The tax generates considerable gains on the road, but these are offset by the losses in the labor market. In fact, in this polar example the efficiency-maximizing level of traffic lies above that of the no-toll case. That level can be reached only with a negative toll, which functions as an indirect way to reduce the burden in the labor market.

Labor taxes are high in many countries and likely exceed the marginal external congestion cost associated with typical commuting patterns. For instance, assuming a kilometeric external cost of €0.30 and an average commuting distance of 30 km (both ways) implies a daily external cost of €9.00. The average gross income in the OECD was approximately €45,000. Using an average annual labor supply of 1744 hours/year and assuming a working day of seven hours implies approximately 250 working days a year. In turn, this implies an average gross daily wage of €180. Multiplying that by 0.255, i.e. the average labor income tax rate in the OECD, yields a daily labor income tax of €45.9, which is more than five times larger than the generated external effect of commuting. For this reason, the result from the polar case is sometimes taken as a warning against the introduction of urban road pricing, when the latter is not combined with cuts in labor taxes. The concern is that alleviating traffic externalities may inadvertently result in larger labor market distortions that could offset the welfare benefits from abating these externalities. Combining road charges with labor tax cuts avoids these adverse effects. To what extent should that warning be taken into account in practical efforts to curb urban traffic externalities?

Weakening Strict Complementarity Between Labor Supply and Commuting

Relaxing the strong assumptions of the polar case leads to weaker interaction between the labor tax and road toll bases and reaffirms the case for road tolls.

Two Commuting Modes

Suppose that, in addition to commuting by car, workers can opt for public transport (Parry and Bento, 2002). For simplicity, public transport is not subject to economies of scale and there are no external costs associated with it, so the marginal social cost of its provision is constant. In contrast to the polar case, there is now a difference between the tax on labor income and the congestion charge. The labor tax is the same for those commuting by car or by public transport, but the congestion charge applies to car commuting only.

How should the congestion charge be set in this context? To maximize efficiency, the policy maker prices public transport at its resource cost and introduces a congestion charge on the road. With the labor tax present, and without any “sophisticated” form of revenue recycling, *i.e.* any recycling mechanism other than lump-sum transfers, the optimal road tax lies below the marginal external cost of road use. The difference depends on the degree to which public transport can substitute the use of car. In the extreme case in which substitution is perfect, road traffic is taxed optimally at its marginal external cost.

Two Trip Purposes

The polar case considers commuting transport only. How do insights change if, instead, account is taken of the fact that there are diverse trip purposes? Key observations can be deduced from a stylized setting where commuting traffic coexists with “leisure traffic” (Van Dender, 2003). The latter encompasses all non-commuting trips and is a relative complement to leisure, while the former is strictly complementary to labor supply. In this case, the tax base of the labor tax and the road toll differ. The labor tax applies to days worked only, whereas the road charge in addition applies to the non-work activities from which the demand for leisure transport is derived.

Consider first the case in which it is not possible to reduce the labor tax and the road tax cannot be differentiated between the two user types. To maximize efficiency, policy makers should set the uniform charge to a level that equates two welfare components: (1) the gains from a reduction of road externalities induced by a marginal increase of the road tax and (2) the losses in the labor market, induced by the same adjustment. If there is only non-commuting traffic, the optimal tax is the marginal external cost of road use. In the opposite case, the optimal tax is zero. Empirical observation suggests that 35%–70% of the kilometers generated within the peak hours are not related to commuting. This indicates that road pricing during commuting hours remains relevant, even in the presence of strong tax interaction effects.

Suppose next that commuters can be distinguished from the rest of the road users. Since non-working trips are relative complements to leisure, introducing a differentiated charge allows to shift the burden of the road tax from (relatively heavily taxed) labor to (less taxed) leisure, with the potential efficiency gains as described in standard optimal taxation models. This implies that efficient charges on leisure trips may lie above marginal external costs, while commuting trips are charged at a fraction of it. Differentiating charges between trip purposes is hard to do directly, as the latter are not directly observable. However, *ex post* deductibility of the road tax paid for commuting trips from income taxes may allow for reasonable approximations.

To summarize, charging marginal external costs to non-work travel is, in many cases, justified. Higher charges can also be considered if revenues are used to cut labor taxes. Optimal surcharges are hard to quantify, so sticking to the simple rule is a sensible pragmatic option. The downward pressure of road charges on real wages can be mitigated by making the costs tax-deductible, in addition to ensuring that revenues are used to keep labor taxes as low as possible.

Weaker Links Between Commuting and Labor Supply

The strict complementarity case stipulates that each workday is of fixed length and requires a commuting roundtrip. There are various reasons why there can be more flexibility in the relation between labor supply and commuting. For example, even if the length of each worker’s workday is fixed, introducing time-varying road charges could shift the time of departure for the commuting and the home-returning trip. That is, a time-varying toll offers to drivers who are—in the polar case—subject to a strong tax interaction effect an option to commute in time intervals in which the toll is lower. This weakens the negative tax interaction effect, as illustrated in the intermediate case in Fig. 1. As a result, the efficiency of a road tax scheme can improve substantially by allowing time-varying tolls, even in the presence of a distorted labor market.

Another important departure from the polar case regards the possibility of adjusting the duration of the working day. In this case, part of the decrease in labor supply that the road toll induces can be offset through longer working days. The substitution effect will be larger the slower the rate at which the marginal productivity of an additional working hour diminishes during a working day, and the lower the value of leisure time is. If the marginal productivity of labor remains constant throughout the day and lies above the marginal valuation of leisure time, the two tax bases disconnect completely (Fig. 1, right panel). Then, the optimal road tax equals the marginal external cost of road use.

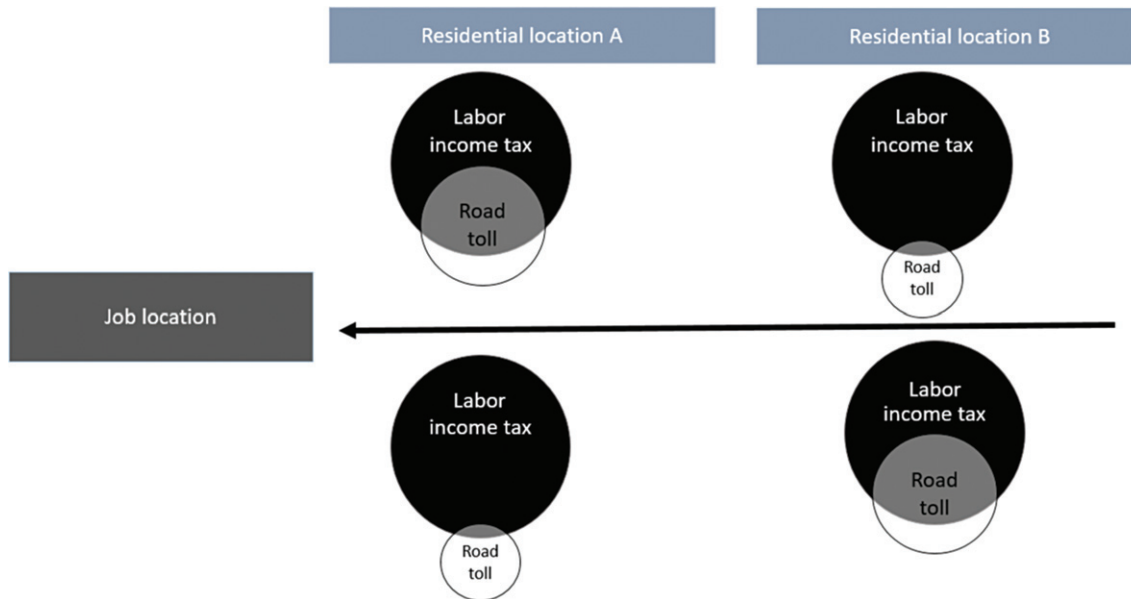


Figure 4 The interaction between a labor income tax and a road toll for different residential locations.

On top of the possibility that labor supply is elastic within a given day, teleworking constitutes another important margin that reduces the overlap between the two tax bases. If all professional groups have the possibility to work remotely with no impacts on wage or productivity, the remaining traffic is not related to commuting. Then, the optimal tax to that remaining traffic is the marginal external cost it generates (or even more, as part of a tax shift operation that we discussed above).

The upshot of these arguments is that, as flexibility to arrange working habits increases, road charges become more efficient and the tax interaction effect wanes (Fig. 2).

Spatial Considerations

So far, the discussion has abstracted from spatial heterogeneity, implicitly focusing on an average road charge across the road network. Such a charge converges to the overall marginal external cost of road use when the labor and the road tax bases are disconnected, while it approaches zero when the two bases overlap completely. What happens when the tax interaction effect is not constant across space, and the elasticity of labor supply varies with distance from the working location?

In this case, the spatial design of the charging system can be adjusted to target those segments of the labor force whose labor supply is less sensitive to changes in the commuting costs (Tikoudis et al., 2015; Tikoudis, 2020). That is possible as long as there is some distinguishable pattern of residential location, job location, or road use by these segments of the labor force. Such a spatially varying charging system may produce considerable welfare gains without a need to reduce the labor tax and even in the presence of a tax interaction effect that is, on average, substantial (Tikoudis et al., 2015).

To illustrate graphically this possibility, assume there are two households, one located close to the work location and one far away from it, as in Fig. 4. In the upper part of the figure, the tax interaction effect is weak far away from the employment hub but strong close to it. In that case, a cordon toll — which has the potential to charge particularly commuters from more remote locations — can be a more efficient option than marginal external cost pricing. The narrative is reversed in the lower panel, in which the tax interaction effect is lower closer to the employment location, possibly because that location is served well by public transport. In that case, it can make sense to leave the kilometers produced far away from the employment hub unpriced but to impose a charge close to the marginal external cost of road use near the city center. This requires that the scheme can identify (possibly using the home location of drivers) the long commuters and give them a free pass. In either case, the best results are obtained when tolls are low in road segments used by the parts of the labor force characterized by a strong tax interaction effect, and *vice versa*.

Discussion

Road pricing has the potential to improve the efficiency of transport systems by aligning the private cost of car use with its social costs. Relative scarcity of road capacity during commuting hours can result in high social costs, so the potential gains of better capacity use can be large. Road pricing can mitigate a series of traffic-related externalities more efficiently than traffic restriction schemes, as the latter risk eliminating trips whose social benefit outweighs their social cost. Mitigating congestion tends to reduce air pollution, resulting in significant additional benefits, particularly in monocentric areas with a dense urban core. During the peak

hours, air pollution in the central business district of these areas increases steeply due to car traffic, exposing a large portion of the population to high concentrations of air pollutants.

The potential welfare benefits of road pricing can be reduced by interactions with other tax instruments, such as the labor tax, whose base overlaps substantially with that of a road tax. In an extreme case, the interaction of the road and the labor tax is strong enough to render any road charge welfare-reducing, as the road tax increases the effective tax burden on labor. However, this polar case serves mainly as a conceptual device to illustrate how relaxing its underlying assumptions re-establishes the case for road charges that converge to the marginal external costs of road use.

The negative impact of road pricing on labor supply is much weaker when a substantial part of commuting trips can be rescheduled or can be undertaken by public transport. The same holds when longer days can give rise to fewer commuting trips. Also, flexibility in the length of the workday facilitates peak avoidance, and teleworking may be possible for some. For these reasons, the effect of urban road pricing on labor market inefficiencies is much weaker than what the polar case of strict complementarity suggests. If labor supply elasticity displays a clear spatial pattern, it is possible to design the charge system to tax higher the portion of traffic that is characterized by a lower elasticity of labor supply. Nevertheless, the risk that road pricing results in higher effective labor taxes when alternatives for car use are scarce and when worktime arrangements are rigid, is worth noting. In such cases, caution with high road charges is warranted.

The discussion also highlights how the economic benefits of urban road pricing depend on how the revenues from the charges are recycled to the economy. In the models discussed above, the labor tax distortion is often the main source of inefficiency in the economy. In that case, it is not surprising that using congestion charge revenues to cut labor taxes produces the most economically efficient results. Strict complementarity between labor supply and commuting implies that any other form of revenue use will possibly reduce welfare.

The more general message on revenue use is that cutting the most distortive taxes is preferable in an equal tax revenue yield analysis, and that it is crucial to ensure that road tax revenues are put to socially productive use. Socially productive spending options are many, but often include expanding public transport in urban regions. It was noted, above, that the availability of high quality public transport strengthens the case for congestion charging, as this helps keep commuting costs low. This suggests that using congestion charge revenues for improving public transport can make good economic sense. It also often makes good political sense, as public support for charging is easier when the revenues are recycled within the urban transport sector. In addition, combining urban area charges with urban area spending facilitates political ownership of congestion charges. The latter is more difficult to obtain when congestion charges are combined with labor tax cuts, given the different ministerial competencies that are typically involved in such an endeavor.

References

- Parry, I.W.H., Bento, A., 2002. Revenue recycling and the welfare effects of road pricing. *Scand. J. Econ.* 103 (4), 645–671, doi:10.1111/1467-9442.00264.
- Schrank, D., Lomax, T., Turner, S., 2012. TTI's Urban Mobility Report. Texas Transportation Institute.
- Tikoudis, I., 2020. Second-best road taxes in polycentric networks with distorted labor markets. *Scand. J. Econ.* 122 (1), 391–428, doi:10.1111/sjoe.12322.
- Tikoudis, I., Verhoef, E.T., van Ommeren, J.N., 2015. On revenue recycling and the welfare effects of second-best congestion pricing in a monocentric city. *J. Urban Econ.* 89, 32–47.
- Van Dender, K., 2003. Transport taxes with multiple trip purposes. *Scand. J. Econ.* 105 (2), 295–310, doi:10.1111/1467-9442.00010.

Further Reading

- Adler, M.W., van Ommeren, J., 2016. Does public transit reduce car travel externalities? Quasi-natural experiments' evidence from transit strikes. *J. Urban Econ.* 92, 106–119, doi:10.1016/j.jue.2016.01.001.
- De Borger, B., 2009. Commuting, congestion tolls and the structure of the labour market: optimal congestion pricing in a wage bargaining model. *Reg. Sci. Urban Econ.* 39, 434–448.
- De Borger, B., Wuyts, B., 2011. The structure of the labour market, telecommuting, and optimal peak period congestion tolls: a numerical optimization model. *Reg. Sci. Urban Econ.* 41 (5), 426–438, doi:10.1016/j.regsciurbeco.2011.02.002.
- Gutiérrez-i-Puigarnau, E., van Ommeren, J.N., 2015. Commuting and labour supply revisited. *Urban Stud.* 52 (14), 2551–2563, doi:10.1177/0042098014550452.
- Maddison, D., Johansson, O., Pearce, D.W., 1996. True Costs of Road Transport. Blueprint Series, 5. Earthscan, London, UK.
- Parry, I.W.H., Small, K.A., 2009. Should urban transit subsidies be reduced? *Am. Econ. Rev.* 99 (3), 700–724, doi:10.1257/aer.99.3.700.
- Tikoudis, I., Verhoef, E.T., van Ommeren, J.N., 2018. Second-best urban tolls in a monocentric city with housing market regulations. *Transport. Res. Part B* 117A, 342–359.

Demand Management and Capacity Planning of Airports

Stephen Ison, Lucy Budd, Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	227
Airport Capacity	227
The Airport Capacity Challenge	228
Capacity Planning and Operational Enhancements	228
Demand Management	229
Summary	230
Further Reading	230

Introduction

The principle of aviation demand management at commercial airports involves using economic or administrative mechanisms, or a combination of the two, to reduce or redistribute air traffic at airports when airline demand for an airport service (usually expressed as a wish to land or take off at a particular time of day) outstrips the ability of an aspect of that airport's infrastructure to accommodate it. In most cases, the constraining factor is the available capacity of the operational runway/s. Airport demand management has the potential, if appropriately applied, to reduce flight delays and improve on-time performance (OTP), minimize airspace and airfield congestion, lessen local environmental impacts, and ensure high levels of safety and service at capacity-constrained airports. In short, airport demand management aims to control airline demand for airport access to address imbalances between airport capacity and demand. The following sections will introduce the concepts of airport capacity and demand management as they pertain to contemporary air transport operations.

Airport Capacity

A lack of airport capacity in major cities, particularly at peak times, is arguably, along with tackling climate change, one of the most challenging issues facing the global air transport industry. Many major airports can trace their origins to the 1940s or earlier when aircraft were smaller and lighter, air travel a mode of transport for the wealthy elite, and annual passenger numbers could be measured in the thousands. Today, primarily as a result of regulatory reform, enhanced competition, and technological innovation, the financial cost of flying has fallen in real terms and more than 4.6 billion passengers undertake airline flights every year, many commencing or ending their journeys at airports that have long since exceeded their intended design capacity. The economic and social implications of capacity constraints, which frequently manifest themselves in the form of flight delays, lost productivity, constrained connection possibilities, reduced competition, and passenger dissatisfaction are often cited as justifications for expansion. Indeed, when airline demand approaches or exceeds the ability of an airport to supply a safe and efficient service then congestion and delays invariably occur.

Capacity describes the ability of an aspect of an airport's physical or technological infrastructure (which includes, but is not limited to, runways, taxiways, terminal airspace, gates, ground access, security screening, border control, baggage handling, and other passenger processing activities) to safely accommodate demand. Demand in this instance is defined as the desire by commercial airlines to operate flights from an airport at a particular time of day and year. An airport's capacity is determined by complex interdependences between internal airport-specific and external political and environmental factors. Internal factors that influence airport capacity include: the number, siting, orientation, and physical characteristics of runways and taxiways (including their length, width, surface material, available lighting and navigation aids, and pavement strength); the size and layout of airside maneuvering areas and passenger terminal facilities; the configuration and limitations of proximate airspace; the presence of local site-specific environmental/noise curfews or other operating restrictions; and the nature of airline demand (particularly as it relates to flight timings, the size and mix of aircraft using the facility, the nationality of passengers, and each airline's business model). Airport capacity is also influenced by external factors, including local meteorological conditions, Air Traffic Control provision, Rescue and Fire Fighting cover, and national safety and security regulations.

Airport capacity is usually expressed in operational terms and describes the maximum number of air traffic movements (ATMs) that can be safely accommodated at an airport within a specific time period under specified operating conditions, service standards, and continuous demand. However, capacity is notoriously difficult to define and different countries (and even individual airports within one country) adopt different approaches and there is no universally accepted standard for measuring or defining it.

The two principal types of airport capacity are declared capacity and practical capacity. Declared capacity is an administrative tool that is used for strategic scheduling purposes and slot allocation. It typically specifies the maximum number of ATMs (defined as one

landing or one takeoff) that can be scheduled per unit of time. In order to ensure operational reliability and resilience, this figure can be quite conservative. Practical capacity, in contrast, is often a truer reflection of the number of movements that can be handled under ideal operating conditions and the figure is consequently higher. Irrespective of which metric is used, as demand approaches capacity delays generally start to occur.

The Airport Capacity Challenge

In the pre-deregulation and pre-liberalization era, airline regulation served as a form of capacity management and demand distribution. In the regulated era, national governments could dictate which airlines could fly, which airports could be served, the frequency and timings of services, the airfares that could be charged, and the passenger and cargo capacity of the aircraft that could be used on individual routes. Deregulation of the US domestic airline market in 1978 and subsequent moves toward deregulation and liberalization elsewhere in the world from the mid-1990s onwards fundamentally changed the nature of passenger and airline demand. Greater freedom of market entry and exit, seen as the characteristics of a contestable market, combined with low-interventionist free-market philosophies enabled new airlines to enter the marketplace where they established new routes and introduced a range of product and service innovations. Many of the new entrant airlines chose to compete on price, not service, and their “lower cost” approach to doing business (which only supplied what was necessary for safe and reliable flight) enabled them to pass on the savings to consumers in the form of lower fares. Cheaper fares stimulated unprecedented consumer demand for flying and passenger enplanements dramatically increased.

Incumbent full-service operators reacted to this new competitive landscape by consolidating their operations at key airports and increasing the frequency of services on their most lucrative routes. The formation of “fortress hubs” and hub and spoke networks improved connectivity and flight choice for their passengers and enabled the established airlines to protect their market share while minimizing costs and maximizing profit. However, the necessary scheduling of flights in “time banks” or “waves” to maximize connection possibilities led to peak-hour congestion and growing incidents of delay. The situation became particularly acute at major US airports, many of which were also often subject to the disruptive impacts of adverse weather conditions (including intense convective activity and thunderstorms in summer and snowstorms in winter). Growing delays led to economic disbenefits in terms of lost productivity, increased costs, a reduction in consumer welfare, growing levels of passenger dissatisfaction, and increased environmental externalities. One solution to the resulting demand-capacity mismatch was to expand capacity through the construction of additional infrastructure. However, this took time and required considerable capital expenditure. Another option was to modify existing procedures to enhance operational capacity or actively intervene in the market using administrative or economic instruments to manage demand.

Capacity Planning and Operational Enhancements

Capacity planning at airports is notoriously problematic. Air travel is a derived demand and one that is both highly cyclical in nature and vulnerable to external shocks, including oil price rises, geopolitical unrest, natural disasters such as volcanic eruptions, and global economic downturns. Airport development schemes that aim to increase capacity can involve anything from constructing or extending runways, terminals, and gates at existing facilities to building entirely new airports on greenfield sites and reclaimed land. Both options are very capital intensive and, in countries where extensive public consultation and/or public inquiries form part of the national planning process, may take decades to realize and become operational. In either case, airport development projects are furthermore often controversial owing to the required land take and the social and environmental impacts of airport construction and eventual operation on local communities and ecosystems.

The most important element in the airport capacity planning process is a forecast of demand over the life span of the planned infrastructure asset. In the case of terminal buildings, this can be 25–30 years or more. Forecasting demand for airport services that far into the future involves complex calculations seeks to anticipate changes in future passenger demand, predict disruptive changes that may occur to aircraft technology and energy sources, and foresee the continued evolution of airline business models. Future traffic forecasts are thus subject to significant uncertainties and the history of airport planning and development round the world is replete with examples of airports that have either been built to serve traffic that never materialized as predicted (such as Montreal’s Mirabel airport or Spain’s Ciudad Real facility south of Madrid) or airports which have been built to plans which failed to anticipate subsequent growth in demand. To mitigate this risk, many airports produce staged development plans that afford them the flexibility to incrementally adapt future airside and landside capacity according to low, medium, and high demand scenarios. In every case, the availability of finance is crucial, and the individual regulatory regime of each country worldwide will dictate the extent to which public or private (or a combination of public and private) finance is used to fund development activities.

In situations where the physical development of infrastructure is politically problematic or financially prohibitive, it may be possible for an airport to increase capacity through operational and procedural enhancements. These interventions typically seek to reduce uncertainties in the system and provide a more robust and resilient operational environment. Examples of operational enhancements include ground delay programs, airborne speed control, and better sequencing of arriving and departing flights. In Europe, the Airport Collaborative Decision-Making (A-CDM) program aims to improve OTP and enhance network reliability through improved communications between airport stakeholders while Enhanced Time-Based Separation (ETBS) air traffic control

procedures help to improve runway capacity at capacity-constrained airports during strong headwind conditions while maintaining safe separation.

Although such operational and procedural enhancements offer some potential for airports to incrementally improve capacity, if demand continues to outstrip capacity then some form of demand management intervention may be required.

Demand Management

Demand management can offer a relatively quick (though not necessarily ideal) solution to capacity constraints at a relatively low implementation cost. It also arguably makes for better utilization of a scarce resource. The demand management interventions that are appropriate for an airport are informed by a suite of models and analytical tools, of which three basic types can be identified:

1. Administrative

- a. *Traffic diversion*: A mechanism used to redistribute air traffic from capacity-constrained airports to proximate facilities with underutilized assets. This intervention will only work in a regulated market where the state is able to dictate which airports individual airlines can use to serve certain destinations. In a liberalized market, unrealized demand for airport access is often spilled to a neighboring facility with spare capacity. In this way, low-cost operators—who value operating from cheaper and less congested airports—have been able to grow traffic at previously underutilized regional and secondary facilities.
- b. *Traffic cap*: This intervention either stipulates the maximum number of scheduled ATMs or the maximum number of passengers who are permitted at a particular airport in any given time period. These caps are usually set by the national aviation regulator. In Australia, the *Sydney Airport Demand Management Act 1997*, for example, set a cap of 80 hourly ATMs. Traffic caps may also form part of the planning conditions for airport development. For example, a cap on the annual number of ATMs permitted at Heathrow was set at 480,000 as part of the planning condition for the development of Terminal 5. On the other side of London, in February 2018 the operator of Stansted airport applied to the UK's Civil Aviation Authority to raise the cap on annual passenger numbers from 35 to 43 m. In addition, some airports impose their own voluntary (and therefore nonenforceable) caps on particular types of movements for noise or other environmental reasons.
- c. *Slot control* or slot coordination: A slot (the permission to operate an aircraft at a particular time at capacity contained airports) is a *planning* tool for rationing and managing available capacity at airports where demand for air services exceeds the capacity of the airport's runway/s and terminal/s to accommodate it. Slot control or coordination is a quantity-based approach under which airports state their declared capacity and airlines request slots that are then allocated by an independent schedule coordinator. Administrative priority rules typically specify the order in which airline requests should be considered. Incumbent operators often benefit from "grandfather rights" on their existing slots, while any unused slots are returned to the allocation pool and made available for new entrant operators. Capacity-constrained European airports practice slot coordination, but this intervention is arguably both inefficient and anticompetitive as the allocation of slots is often based on historical precedent and the associated "use it or lose it" rules deny market entry to new operators while incentivizing operators to "babysit" slots and protect them for future years. The formal procedures used to coordinate global schedules at capacity-constrained airports are detailed within the IATA Worldwide Slot Guidelines. Currently more than 200 capacity-constrained airports around the world use slot allocation and slot control to manage demand in a way that is intended to be fair, neutral, and transparent. Nevertheless, allocating slots in this way is not without its critics and debates around slot reform continue.

2. Economic

- a. *Congestion pricing*: A price-based intervention that seeks to equate the price airlines pay for airport access with the marginal social cost (MSC) of capacity utilization. A number of airports charge differential fees depending on the time of day (peak/off-peak) and season that services are operated and some, who are operating in a competitive market, offer additional incentives in the form of fee reductions or waivers to attract airline services in off-peak periods. While airport charges are undoubtedly a factor in an airline's decision to operate from one location over another, as airport costs are only a proportion of an airline's total costs and the charging differentials are small, carriers can tend to be relatively price inelastic. Consequently, congestion pricing only has a minimal impact on airline scheduling behavior. The basic economics of congestion pricing is illustrated in **Fig. 1**. This figure illustrates the cost of delay on the vertical axis and the number of aircraft movements at a particular airport on the horizontal axis. The marginal private cost (MPC) represents the airline's private cost including the cost of fuel and the MSC curve includes the cost of flight delays. The cost of delays increases as the number of aircraft movements increases. Demand represents the demand for access to the runway and can also be seen as the marginal benefit. Without congestion pricing the number of movements is M1 and the airline operator's marginal internalized costs are C0. With this level of movement, however, the social cost exceeds the marginal benefit by AB and as such economists would suggest the introduction of a congestion price equal to CD, thus resulting in M2 movements, a social optimum. Up to aircraft movements of M3 there is no divergence between MPC and MSC essentially because the airline operators are not experiencing flight delays, with demand being well within the capacity of the airport. Historically, this has been a major component of the low-cost airline model with low-cost operators actively identifying secondary underutilized airports from which to operate their services. While congestion pricing can improve economic efficiency, it is not without its difficulties, not least in terms of

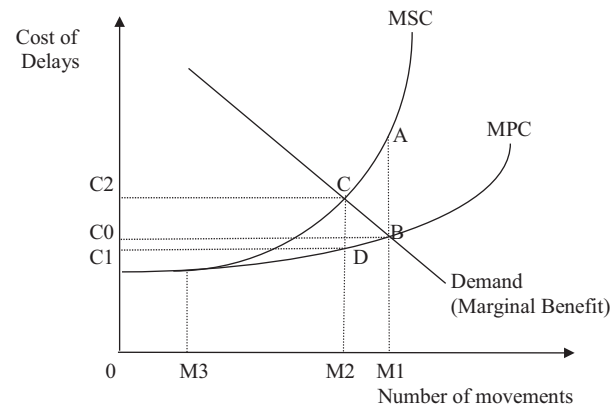


Figure 1 Delays and congestion pricing.

measuring the congestion costs and calculating the congestion price. The demand curve is also unstable, varying in relation to the peak and off-peak nature of derived demand, thus making it difficult to enact a differentiated congestion price. In addition, there is the issue of how the revenue generated from the congestion price (represented by $C1$, $C2$, and CD in **Fig. 1**) should be utilized.

- b. *Slot auctions*: Slot auctions represent another mechanism by which demand can be managed. Under this approach, airports specify a declared capacity and allocate the available slots to the highest-bidding airlines. This has the potential advantage that slots are allocated to the airlines that value them the most.

3. Hybrid

- a. *Market-based capacity allocation instruments*: These typically combine slot control and congestion pricing by allocating some slots and selling others. Secondary trading between airlines may be permitted to allow carriers to sell and purchase slots once the initial primary allocation has been completed. Hybrid mechanisms seek to enhance the efficiency of capacity allocation while addressing some of the practical challenges associated with economic interventions.

Irrespective of which approach is adopted, the aim of any demand management intervention is to reduce demand to a manageable level by modifying the temporal and/or spatial characteristics of airline demand to “de-peak” operations and/or redistribute flights to other airports.

Summary

Demand management can operate over varying timescales from the long-term strategic to the short-term tactical. Any form of demand management can deny some (potential) users free or complete access to the airport of their choice at the time of their choosing. The design of any demand management mechanism must be carefully considered and evaluated to ensure that it meets stated objectives to achieve the desired operational, managerial, and economic outcomes.

In contrast to capacity developments and operational or procedural enhancements that involve up-front investment costs but result in gains for all stakeholders, demand management imposes longer-term economic costs on airlines (in terms of scheduling options and route development) and consumers (in terms of airline choice and connectivity).

Any demand management mechanism involves a trade off between mitigating congestion and maximizing the utilization of available capacity. They must be robust and flexible enough to respond to changing operating conditions and continuous monitoring is required to improve airport demand management and capacity planning. Taken together, demand management and capacity planning attempt to address the demand-capacity mismatch that is evident at many of the world’s major airports.

Further Reading

- Czerny, A.I., 2010. Airport congestion management under uncertainty. *Transp. Res. Part B Method.* 44 (3), 371–380.
- Gillen, D., Jacquillat, A., Odoni, A.R., 2016. Airport demand management: the operations research and economics perspectives and potential synergies. *Transp. Res. Part A Policy Pract.* 94, 495–513.
- Graham, A., 2018. *Managing Airports: An International Perspective*, fifth ed. Routledge, Abingdon.
- IATA, 2017. *IATA Worldwide Slot Guidelines*, eighth ed. IATA, Geneva/Montreal.

Dealing With Negative Externalities: Low Emission Zones Versus Congestion Tolls

Valeria Bernardo*, Xavier Fageda†, Ricardo Flores-Fillol‡, *University of Barcelona, TechnoCampus, Barcelona, Spain; †University of Barcelona, IESE, Barcelona, Spain; ‡Universitat Rovira i Virgili, Reus, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	231
A Simple Model to Explain the Effect of Low Emission Zones and Congestion Tolls	232
The Excess Traffic Generated by Urban Congestion	232
Measures: Congestion Tolls Versus LEZ	233
Extension of the Model to Incorporate Pollution	233
Overall Assessment of Congestion Tolls and LEZ	233
The Effect of Urban Tolls and LEZ on Congestion and Pollution	234
Urban Tolls	234
LEZ	234
Conclusion: LEZ Versus Congestion Tolls	235
Acknowledgment	235
References	235

Introduction

The great weight that the car has as a means of mobility in cities generates two significant externalities: congestion and pollution. Congestion produces traffic jams that affect commuter drivers, but also pedestrians that find their streets blocked by vehicles producing noise and pollution. Large cities around the world suffer typically from severe congestion problems. TomTom provides data about congestion levels in cities around the world using as indicator of congestion the additional time a vehicle needs to undertake a representative trip as compared to a free flow situation. On an average, most populated cities (with more than 1 million inhabitants) report congestion levels exceeding 20%. In peak periods, such percentage may reach values close to 100%. It must be stressed that the relationship between congestion and pollution is clear, since prolonged car circulation at reduced speeds has a notable effect on polluting emissions (Barth and Boriboonsomsin, 2008; Beaudoin et al., 2015; Parry et al., 2007).

The World Health Organization (WHO) has a database that measures air quality of the 3000 most populated cities in the world in terms of PM10 and PM2.5 (WHO, 2005). This type of pollution (mostly from urban traffic) is associated with increases in the mortality of the exposed population. According to the WHO, from an average annual concentration of 10 $\mu\text{g}/\text{m}^3$, there is an association between cardiopulmonary effects and mortality due to prolonged exposure to PM2.5 (this value is 20 $\mu\text{g}/\text{m}^3$ for PM10). The most severely congested cities in the world register serious pollution records far superior to the WHO guideline values. Indeed, these emissions are the main cause of the death of 3.3 million people per year in the world (more than AIDS, malaria, and the flu together).

In short, most of large cities around the world suffer the negative consequences of congestion and pollution associated to a car-based mobility. Hence, the implementation of policies to deal with such negative externalities is usually in the agenda of local governments.

Investments in capacity are extremely expensive, involve long gestation periods, and are ineffective in areas with dense road networks (Duranton and Turner, 2011). Therefore two types of measures can be applied depending on whether they are quantity or price-based. Low emission zones (LEZ) are the most popular quantity-based measures in Europe. The EU has shown a clear determination to reduce pollution, especially since the transposition of the directives 1999/30/EC and 2008/50/EC, which has been accompanied by the establishment of the “Euro” regulatory standards for vehicles. LEZ ban polluting vehicles from city centers. Thus their primary goal is not to mitigate congestion but to abate pollution. However, their effectiveness is logically conditioned by their impact on congestion. Around 220 cities in 14 European countries have implemented LEZ, including several cities in Germany, Sweden, Italy, or Greece, and cities such as London, Paris, Prague, or Lisbon. Examples of LEZ can also be found in large Asian cities such as Beijing, Tokyo, Shanghai, or Hong Kong. The precise way LEZ are put into service may vary considerably across cities, for instance, in terms of the intervention area or the type of vehicles being affected.

Price-based measures consist in charging congestion tolls, typically to enter/exit to/from the city center during peak hours. Tolls increase drivers’ travel cost and reduce traffic consequently. They have been applied in few cities, being the most important Singapore (1975), London (2003), Stockholm (2006), Milan (2008), Gothenburg (2013), and Palermo (2016).

In this chapter, we examine the effectiveness of congestion tolls and LEZ in dealing with congestion and pollution. The rest of this chapter is organized as follows. Section “A Simple Model to Explain the Effect of Low Emission Zones and Congestion Tolls” shows how urban congestion is modeled from a microeconomic viewpoint and the theoretical effects of congestion tolls and LEZ. Section “The Effect of Urban Tolls and LEZ on Congestion and Pollution” analyzes the results in the existing literature about the

effects of tolls and LEZ on congestion and pollution. Finally, Section “Conclusion: LEZ Versus Congestion Tolls” provides some final considerations.

A Simple Model to Explain the Effect of Low Emission Zones and Congestion Tolls

We first provide a welfare analysis explaining the inefficiency associated with congestion. Then we explain the effect of tolls and LEZ. Finally, we incorporate pollution into the analysis to provide an overall assessment by suggesting four scenarios depending on the effectiveness of tolls and LEZ.

The Excess Traffic Generated by Urban Congestion

Urban congestion is an externality that appears when traffic volume exceeds a certain threshold. The textbook model (Brueckner, 2011; Cantillo and Ortúzar, 2014; Lindsey and Verhoef, 2001) assumes that the time that a vehicle needs to undertake a representative trip (t) is a function of traffic (q), so that $t = f(q)$. The total time needed by all vehicles is therefore $qt = qf(q)$. Finally, the increase in total time that represents the introduction of a new vehicle is $\partial qt / \partial q = t + q \partial t / \partial q$ where the second term of the expression identifies the externality for the rest of drivers. We can define the congestion threshold $\tilde{q}$ so that, $\partial t / \partial q = 0$ for $q < \tilde{q}$ and $\partial t / \partial q > 0$ for $q > \tilde{q}$. Graphically, as shown in Fig. 1, the externality is the vertical distance between $t = f(q)$ and $\partial qt / \partial q$.

Reinterpreting t as the generalized travel cost, $t = f(q)$ and $\partial qt / \partial q$ can be understood as the average private cost and the marginal social cost associated to the trip and can be graphically represented as the private supply (S_p) and the social supply (S_s), respectively. If we incorporate a demand for travel (D), we obtain the equilibrium q^* , t_p^* represented in Fig. 2. Congestion increases the social cost

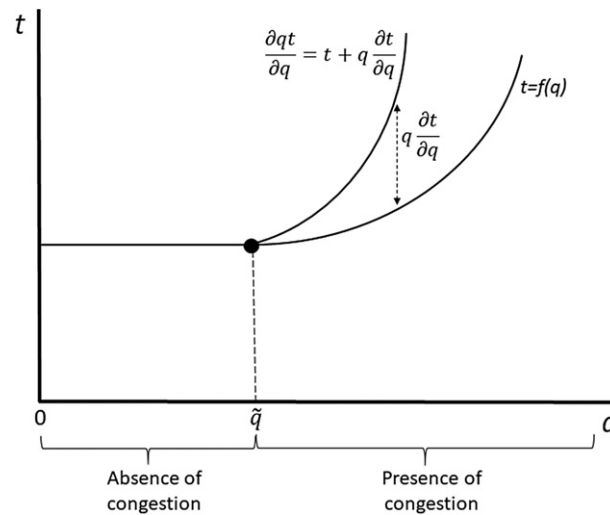


Figure 1 Congestion as a negative externality.

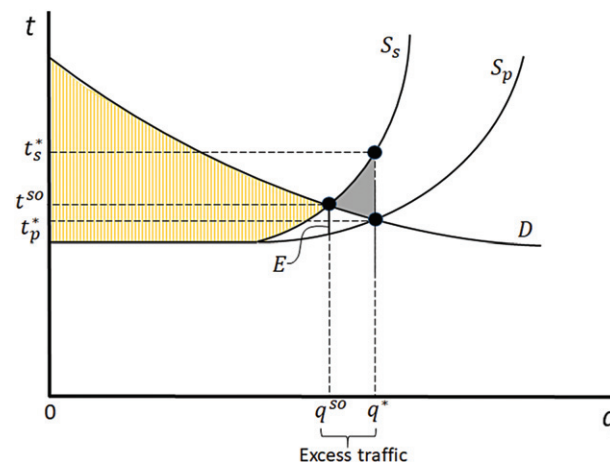


Figure 2 The equilibrium and the social optimum.

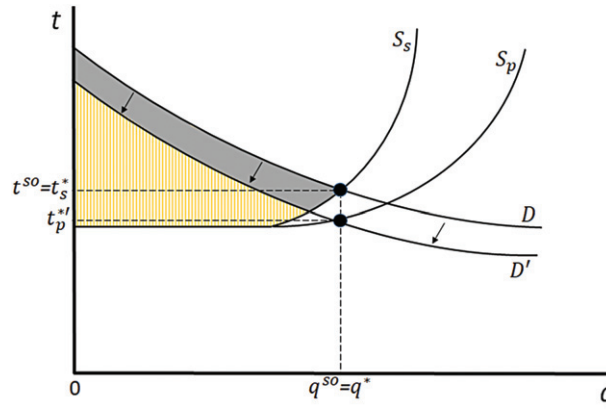


Figure 3 Quantity measures.

to t_s^* . The social optimum (q^{so} , t^{so}) is obtained from equaling the demand to the social supply and the difference $q^* - q^{so}$ indicates the excess traffic observed in equilibrium. The figure also shows the social welfare associated with car usage, which is positive (welfare gain) for $q < q^{so}$ (striped area), and negative (welfare loss) for $q > q^{so}$ (shaded area).

Measures: Congestion Tolls Versus LEZ

Two main types of measures can be applied depending on whether they are quantity or price-based.

Price measures: congestion tolls. A congestion toll increases the average private cost (t) up to the marginal social cost ($\partial qt / \partial q$). The optimal amount of the toll to recover the efficiency and eliminate the excess traffic equals the externality evaluated at the optimal number of trips $E = q^{so} \partial t / \partial q$. Such toll would internalize the externality while raising some additional revenue.

Quantity measures: LEZ. LEZ reduce the demand for travel. In case the contraction of demand would exactly eliminate the excess traffic, we would get the situation as shown in Fig. 3 where $q^* > q^{so}$.

Therefore, both congestion tolls and LEZ can be equally effective in eliminating the excess traffic if they are correctly designed. However, quantity measures generate welfare losses (new shaded area in Fig. 3) as they do not take into account drivers' valuations and are applied indiscriminately.

Extension of the Model to Incorporate Pollution

Consider now that there are two types of cars, new and old, so that the traffic volume is $q = q_{old} + q_{new}$ where $q_{old} = \lambda q$ and $q_{new} = (1 - \lambda)q$ with λ and $1 - \lambda$ being the shares of each car type. Supposing that new cars produce 0-emissions (i.e., electric cars), car pollution is given by $P = g(q_{old})$ with $g'(q_{old}) > 0$. Although both tolls and LEZ can eliminate the excess traffic generated by congestion, their effect on pollution is different with LEZ being in general more effective. The reason is that tolls do not discriminate between new and old cars, whereas LEZ only ban old cars. The relative effectiveness between both measures in mitigating pollution depends on the proportion of old cars λ , which tends to decrease over time as the fleet is renewed. In addition, as LEZ are announced some time before they are made effective, a certain accommodative behavior of local drivers can accelerate this car renewal process.

Overall Assessment of Congestion Tolls and LEZ

When assessing the performance of tolls and LEZ, we need to take into account their effect on both externalities. Looking at congestion (excess traffic), tolls are better because they are more efficient (as they take into account drivers' valuations) and at least equally effective, with the effectiveness of LEZ depending on the level of λ . Concerning pollution, the performance of LEZ is better, with the difference between the two depending again on λ . At this point, we can consider four scenarios.

- Scenario 1: $\lambda = 0$. There are no polluting cars and, therefore, LEZ produce no effects and are useless. Differently, congestion tolls can eliminate the excess traffic.
- Scenario 2: $\lambda \in (0, 1)$ low so that LEZ block every old car. In this case, LEZ cannot eliminate completely the excess traffic and are less effective than tolls in abating congestion but they are superior in mitigating pollution.
- Scenario 3: $\lambda \in (0, 1)$ high so that LEZ can eliminate the excess traffic by just banning old cars (assuming a contraction of demand that exactly eliminates the excess traffic). LEZ and tolls are equally effective in mitigating congestion and LEZ are superior in reducing pollution.
- Scenario 4: $\lambda = 1$. All cars are polluting, so that LEZ would ban all vehicle use.

The Effect of Urban Tolls and LEZ on Congestion and Pollution

Urban Tolls

From a theoretical point of view, there is a wide consensus among economists about the advantages of charging a price to get access to congested/polluted areas as a means to deal with the externalities related to car usage. However, the implementation of an optimal toll system is difficult, given that some essential decisions (such as the delimitation of the restricted area, the amount of the toll or the exempted vehicles) are inefficient as they are usually based on political considerations.

In addition, the empirical evaluation of the effects of price policies is a complex task. First, it is difficult to have an unaffected area comparable to the one affected by the toll system, as the restricted zone typically comprises a severely-congested city center. Second and most important, tolls are usually implemented along with improvements in the public transportation system. Therefore, disentangling the effect of two simultaneously applied policies that pursue the same final objective is problematic. Finally, another complication comes from the fact that drivers anticipate to some extent the effective implementation of tolls and accommodate their behavior in advance.

Taking into account these difficulties in identifying properly the effects of urban tolls, empirical evidence generally suggests that congestion pricing is an effective policy insofar as it tends to be associated with a sharp reduction in road traffic in restricted areas and, consequently, with a decrease in congestion and the emission of pollutants. The literature is composed by a series of papers that study the effect of urban tolls on individual cities by comparing traffic, congestion, and pollution levels before and after their implementation. Urban tolls are found to be effective in reducing congestion from the first year of implementation. The analyses for London and Stockholm show that urban tolls reduce congestion by 20%–30% (Börjesson et al., 2012, 2014; Eliasson, 2008; Santos and Fraser, 2006), while the impact is about 10%–15% in Milan and Gothenburg (Andersson and Nässén, 2016; Gibson and Carnovale, 2015; Percoco, 2013; Rotaris et al., 2010). In Singapore, the effectiveness of congestion pricing has been shown to be even higher as compared to European cities (Olszewski and Xie, 2005; Phang and Toh, 1997; Willoughby, 2000). Furthermore, some studies provide evidence on the effectiveness of tolls in mitigating pollution. The reduction in pollution lies between 6% and 17% in Milan (Gibson and Carnovale, 2015) and between 5% and 15% in Stockholm (Simeonova et al., 2018).

Additional positive effects associated with urban tolls have been also identified in the literature, for instance in terms of traffic accidents for London (Green et al., 2016) or in terms of children health for Stockholm (Simeonova et al., 2018).

The ultimate reason behind the underuse of urban congestion tolls has to do with their unpopularity. However, looking at the Swedish experience, this lack of social and political support seems to be a short-run effect (Börjesson et al., 2016; Eliasson, 2008). A recurrent argument against congestion charges is related with their supposedly regressive effects. Nevertheless, they are not necessarily regressive (Eliasson and Mattsson, 2006; Eliasson, 2016). In addition, funds obtained from the toll are typically used to improve public transportation and the mitigation of congestion reduces commuting times and, therefore, fuel consumption.

LEZ

While many studies have examined effects of congestion tolls, the empirical literature on LEZ is scarce and focuses on their effects on pollution. As mentioned above, studies on urban tolls usually focus on individual cities by comparing their performance through the analysis of traffic conditions before and after their implementation. Although some studies on the effectiveness of LEZ follow a similar methodology, other analyses compare pollution levels between cities with and without LEZ systems. Regarding studies on individual cities, similar identification complexities as in the case urban tolls arise. In particular, it is difficult to identify an unaffected area comparable to that affected by LEZ. As for the studies that compare cities with and without LEZ systems, the heterogeneity in the application of the policy is generally not taken into account, as the restricted area may range from a small part of the city to a wide area involving most of the city center. Undoubtedly, the aggregate effects of LEZ are conditioned by the extension of the restricted zone. Finally, as in the case of urban tolls, local drivers can anticipate the effective implementation of the policy and, consequently, adapt their habits in advance. Such accommodative behavior (typically based on a car renewal process) seems to be particularly relevant in the case of LEZ.

Keeping these limitations in mind, some studies examine the effectiveness of LEZ in abating pollution in German cities. In Germany, all vehicles (cars, buses, and trucks) are categorized into four mutually exclusive classes, depending on their PM10 emissions (as PM10 is often considered the most lethal air pollutant due to its capacity of penetration in the respiratory tract and bloodstream). Although German LEZ usually affect city centers, the boundaries of the restricted area, the implementation dates, and the precise types of banned vehicles vary across cities.

These studies for German cities adopt a diff-in-diff approach using panel data composed by cities, which are classified into two groups depending on whether the policy has been implemented. Malina and Scheffler (2015) analyze the impact of LEZ on the emission of PM10, finding a reduction of 13%. However, the authors themselves recognize the limitation of not being able to accurately measure the impact of the policy in surrounding areas. Instead, Wolff (2014) circumvents this limitation by using data at a smaller geographical scale and finds an average reduction of emissions in terms of PM10 of 9%, with a range that goes from almost zero in small cities like Tübingen up to 15% in cities like Berlin. Morfeld et al. (2014) also find a significant impact of LEZ in reducing NO, NO₂, and NO_x (limits on NO, NO₂, and NO_x were imposed in Germany since 2010 as they were proved to be a major traffic-related pollutant). However, the impact is modest, being 4% at most.

Some other studies analyze the effect of LEZ on individual cities by comparing pollution levels before and after their implementation. Panteliadis et al. (2014) study the LEZ implemented in Amsterdam, which gradually banned heavy-duty vehicles based on their emission category. They find a reduction in the concentration of different pollutants, ranging from 4% in terms of NO₂ and NO_x up to 10% in terms of PM10. Ellison et al. (2013) study the case of London, where an emission standard was imposed on trucks, coaches, and buses in an area covering most Greater London. They show that PM10 concentrations within the limits of the low emission zone dropped by 2.46%–3.07% as compared to a lower decrease of 1% in limiting areas; however, no discernible differences are found for NO_x concentrations. Cesaroni et al. (2012) analyze intervention policies in Rome, including the exclusion of all cars from the historical city center and the prohibition of old diesel vehicles within the railway ring. In the intervention area, they find a PM10 and NO₂ reduction of 33% and 58%, respectively (but the results are modest city-wide). It is important to acknowledge that the latter two studies do not employ any econometric techniques allowing controlling for potential confounders like weather.

Although the main goal of LEZ is to reduce pollution, an impact on congestion can also be expected given the strong positive relationship between both externalities. Only Bernardo et al. (2018) provide a direct test on the effect of LEZ on congestion by using data of large European urban areas over the period 2008–16. As in the mentioned studies for Germany, they adopt a diff-in-diff approach using a panel data composed by cities affected and unaffected by the policy. They find that urban tolls (and, to a lower extent, bike-share systems) can be effective in mitigating congestion. Instead, LEZ are ineffective, with the exception of urban areas having a high proportion of old cars before their implementation. Thus, LEZ seem to be applied in European cities with renovated car fleets and, consequently, they cannot have substantial effects in reducing traffic and congestion. This observation suggests that pollution (and not congestion) is the main policy objective for most European cities. An additional reason explaining the preference for LEZ would come from the unpopularity of urban tolls, which are perceived as new taxes.

Conclusion: LEZ Versus Congestion Tolls

From a theoretical viewpoint, there is a wide consensus among economists on the advantages of congestion tolls to confront car-related negative externalities. The main argument is that the price-based measures induce a more efficient use of existing infrastructures, while generating additional revenues. By contrast, LEZ can be expected to be inefficient as they reduce demand indiscriminately, that is, independently of drivers' willingness to pay.

Furthermore, LEZ can have regressive effects as they harm lower income drivers, who typically own older and more polluting cars that do not meet the emission standards. Instead, tolls may be more redistributive as they raise funds that are typically used to improve public transportation.

Regarding the effectiveness of both policies, the empirical evidence shows unambiguous effects associated with congestion tolls in mitigating both congestion and pollution. As for LEZ, the existing studies suggest that, although they may be effective in abating pollution (at least in the short term), they are not effective in reducing congestion.

All in all, urban tolls can be seen as a superior tool as they mitigate simultaneously pollution and congestion. However, tolls are applied in few cities while LEZ are massively implemented in European cities. The reason behind the underuse of tolls has to do with their unpopularity, as they are perceived as new taxes the citizens have to pay for a service that used to be free.

Acknowledgment

We acknowledge financial support from the Spanish Ministry of Economy and Competitiveness and AEI/FEDER-EU (ECO2016-75410-P and RTI2018-096155-B-I00), Generalitat de Catalunya (2017SGR770 and 2017SGR644), and RecerCaixa (2017ACUP00276).

References

- Andersson, D., Nässén, J., 2016. The Gothenburg congestion charge scheme: a pre-post analysis of commuting behavior and travel satisfaction. *J. Transp. Geog.* 52, 82–89.
- Barth, M., Boriboonsomsin, K., 2008. Real-world carbon dioxide impacts of traffic congestion. *Transp. Res. Rec.* 2058, 163–171.
- Beaudoin, J., Farzin, Y.H., Lin Lawell, C.Y., 2015. Public transit investment and sustainable transportation: a review of studies of transit's impact on traffic congestion and air quality. *Res. Transp. Econ.* 52, 15–22.
- Bernardo, V., Fageda, X., Flores-Fillol, R., 2018. How can urban congestion be mitigated? Low emission zones vs. congestion tolls.[online] SSRN. Available from: <https://ssrn.com/abstract=3289613>.
- Börjesson, M., Brundell-Freij, K., Eliasson, J., 2014. Not invented here: transferability of congestion charges effects. *Transport Policy* 36, 263–271.
- Börjesson, M., Eliasson, J., Hamilton, C., 2016. Why experience changes attitudes to congestion pricing: the case of Gothenburg. *Transp. Res. Part A* 85, 1–16.
- Börjesson, M., Eliasson, J., Hugosson, M.B., Brundell-Freij, K., 2012. The Stockholm congestion charges—5 years on. Effects, acceptability and lessons learnt. *Transport Policy* 20, 1–12.
- Brueckner, J.K., 2011. *Lectures on Urban Economics*. MIT Press, Cambridge.
- Cantillo, V., Ortúzar, J., 2014. Restricting the use of cars by license plate numbers: a misguided urban transport policy. *DYNA* 81, 75–82.
- Cesaroni, G., Boogaard, H., Jonkers, S., Porta, D., Badaloni, C., Cattani, G., Forastiere, F., Hoek, G., 2012. Health benefits of traffic-related air pollution reduction in different socioeconomic groups: the effect of low-emission zoning in Rome. *Occup. Environ. Med.* 69, 133–139.

- Duranton, G., Turner, M.A., 2011. The fundamental law of road congestion: evidence from US cities. *Am. Econ. Rev.* 101, 2616–2652.
- Eliasson, J., 2008. Lessons from the Stockholm congestion charging trial. *Transport Policy* 15, 395–404.
- Eliasson, J., 2016. Is congestion pricing fair? Consumer and citizen perspectives on equity effects. *Transport Policy* 52, 1–15.
- Eliasson, J., Mattsson, L.G., 2006. Equity effects of congestion pricing quantitative methodology and a case study for Stockholm. *Transp. Res. Part A* 40, 602–620.
- Ellison, R.B., Greaves, S.P., Hensher, D.A., 2013. Five years of London's low emission zone: effects on vehicle fleet composition and air quality. *Transp. Res. Part D* 23, 25–33.
- Gibson, M., Carnovale, M., 2015. The effects of road pricing on driver behavior and air pollution. *J. Urban Econ.* 89, 62–73.
- Green, C.P., Heywood, J.S., Navarro, M., 2016. Traffic accidents and the London congestion charge. *J. Pub. Econ.* 133, 11–22.
- Lindsey, R., Verhoef, E., 2001. Traffic congestion and congestion pricing. In: Hensher, D.A., Button, K.J. (Eds.), *Handbook of Transport Systems and Traffic Control*. Pergamon, Oxford, pp. 77–105.
- Malina, C., Scheffler, F., 2015. The impact of low emission zones on particulate matter concentration and public health. *Transp Res Part A* 77, 372–385.
- Morfeld, P., Groneberg, D.A., Spallek, M.F., 2014. Effectiveness of low emission zones: large scale analysis of changes in environmental NO₂ NO and NO_x concentrations in 17 German cities. *PLoS ONE* 9, 1–18.
- Olszewski, P., Xie, L., 2005. Modeling the effects of road pricing on traffic in Singapore. *Transp. Res. Part A* 39, 755–772.
- Panteliadis, P., Strak, M., Hoek, G., Weijers, E., van der Zee, S., Dijkema, M., 2014. Implementation of a low emission zone and evaluation of effects on air quality by long-term monitoring. *Atmos. Environ.* 86, 113–119.
- Parry, W.H., Walls, M., Harrington, W., 2007. Automobile externalities and policies. *J. Econ. Lit.* 45, 373–399.
- Phang, S.Y., Toh, R.S., 1997. From manual to electronic road congestion pricing: the Singapore experience and experiment. *Transp. Res. Part E* 33, 97–106.
- Percoco, M., 2013. Is road pricing effective in abating pollution? Evidence from Milan. *Transp. Res. Part D* 25, 112–118.
- Rotaris, L., Danielis, R., Marcucci, E., Massiani, J., 2010. The urban road pricing scheme to curb pollution in Milan, Italy: description, impacts and preliminary cost-benefit analysis assessment. *Transp. Res. Part A* 44, 359–375.
- Santos, G., Fraser, G., 2006. Road pricing: lesson from London. *Economic Policy* 21, 263–310.
- Simeonova, E., Currie, J., Nilsson, P., Walker, R., 2018. Congestion pricing, air pollution and children's health.[online] NBER working paper 24410. Available from: <https://www.nber.org/papers/w24410>.
- WHO, 2005. Air quality guidelines for particulate matter, ozone, nitrogen dioxide, and sulfur dioxide. Global update 2005, WHO Press, Geneva.
- Willoughby, C., 2000. Singapore's experience in managing motorization and its relevance to other countries. [online] *World Bank paper* TWU-43. Available from: <https://trid.trb.org/view/672945>.
- Wolff, H., 2014. Keep your clunker in the suburb: low-emission zones and adoption of green vehicles. *Econ. J.* 124, 481–512.

The Rule-of-a-Half and Interpreting the Consumer Surplus as Accessibility

Mogens Fosgerau*, Ninette Pilegaard†, *University of Copenhagen, Copenhagen, Denmark; †Technical University of Denmark, Kongens Lyngby, Denmark

© 2021 Elsevier Ltd. All rights reserved.

Introduction	237
Discrete Choice Models	237
Discrete Choice and the Additive Random Utility Model	238
Accessibility Measures	238
The Demand Curve	238
The Marshallian Consumer's Surplus	239
The Use of Consumer Surplus in Project Evaluations	239
Rule-of-a-Half	240
Logsums	240
Conclusion	241
References	241

Introduction

The direct consumer welfare impact of transport infrastructure projects is often estimated based on the outputs from a traffic model. Using the so-called rule-of-a-half (ROH), the change in consumer surplus—the user benefits—is calculated using the changes in travel times and costs and the number of trips before and after the project together with externally decided unit costs for the value of time and travel.

The change in consumer surplus measures the aggregate willingness to pay for the change in the transport system that the project generates. Often the transport project provides reduced travel times due to, for example, opening of new roads or bridges, but other types of transport projects can also be considered, for example, change in the frequency of public transport, or introduction of road taxes.

The change in consumer surplus measures the direct benefits to travelers from changes in accessibility. When markets are perfect, this is equal to the total welfare effects of accessibility changes (Jara-Diaz, 1986). Other welfare effects, for example, environmental impacts or the cost to tax payers, are typically just added.

The consumer surplus for an individual traveler is the monetary value of the utility of travel. For example, the utility of a trip is often specified as the utility associated with reaching the destination, less the utility cost of the trip. The utility cost includes both monetary and nonmonetary costs (e.g., time costs).

An alternative method of calculating the user benefits is available when the traffic model is a random utility discrete choice model. In that case, the individual consumer surplus is available within the discrete choice model. Many traffic models build on the nested logit model. In this model, the consumer surplus can be calculated in terms of the so-called logsum.

Discrete Choice Models

Traffic models typically consider the travelers' choice of destination, transport mode, and route. Each dimension is treated as a choice among a finite number of mutually exclusive alternatives, using an additive random utility discrete choice model.

Each choice alternative has an associated utility that accounts for both the attractiveness of options and travel costs. Travel costs include not only monetary costs but also different aspects of travel time. Each traveler chooses the alternative that gives him or her the highest utility.

Travelers have different preferences and therefore end up making choices that are distributed across the choice alternatives. Some heterogeneity can be accounted for via observable differences in, for example, socioeconomic characteristics. Importantly, the model also allows differences in preferences that are not related to observable characteristics. These differences in preferences are random from the perspective of the modeler. As a consequence, the consumers' choices are predicted only as probabilities.

We now proceed to presenting this in a formal model.

Discrete Choice and the Additive Random Utility Model

Consider a consumer with random utilities

$$u_i = v_i + \varepsilon_i, i \in J \quad (1)$$

associated with the alternatives in a finite choice set J , for example, comprising combinations of destinations, transport modes, and routes. v_i is a deterministic part of the utility function while ε_i is a random part determined by factors unobservable by the modeler. The random part ε_i is independent of v_i and is absolute continuous with finite means and full support. We assume for simplicity that utility is money-metric such that we can ignore the translation of utilities into monetary units, as $\frac{\partial v_i}{\partial p_i}$ is constant and equal to -1 .

The consumer chooses the alternative that maximizes his or her utility. The expected maximum utility (across a population of consumers) is given by:

$$G(v) = E \left(\max_i (u_i) \right). \quad (2)$$

$G(v)$ is the expected utility that an optimizing consumer will attain when faced with a given choice set. In the case when the choice includes alternative travel destinations, the expected maximum utility can be viewed as a measure of the accessibility of the collection of travel destinations. To understand this, we first need to discuss what an accessibility measure is.

Accessibility Measures

“Accessibility concerns physical and temporal constraints on behavior and thus is an aspect of the freedom of action of individuals” (Weibull, 1980). In a spatial context, accessibility depends on the number of opportunities for spatial interaction and on a concept of geographical distance. Accessibility can also be defined more generally in a microeconomic choice context where choice alternatives are associated with different levels of attraction and some general concept of distance or proximity. We may say that the accessibility of the set of choice alternatives increases if the alternatives become more attractive or if the distance decreases. An increase in accessibility is seen as a benefit and it is often used as a goal for, for example, planning policies.

An accessibility measure is now a function that attributes a finite and nonnegative real number to any set—or configuration—of opportunities, where the opportunities are defined as the pair of the distance to—based on the chosen distance concept—and the attraction of the alternative (Weibull, 1976). Weibull presents an axiomatic framework for basic properties that such a standard accessibility measure should have (Weibull, 1976, 1980). These properties are of technical as well as of interpretational character.

An accessibility measure should satisfy the following six conditions:

1. The order of the opportunities in a configuration does not affect the accessibility measure;
2. The accessibility measure should be nonincreasing in distance and nondecreasing in the attraction of each alternative;
3. The accessibility measure evaluated at zero distance should be continuous and increasing;
4. A single opportunity with infinite attraction at zero distance is better than any pair of opportunities with finite attractions;
5. Any opportunity with zero attraction should not contribute to the accessibility measure; and
6. When two configurations are equally accessible then adding the same new opportunity to both configurations will not change the equivalence.

Accessibility measures can have different forms. Typically, accessibility measures are separable in opportunities; each opportunity is assigned a numeric value and these are then aggregated into a single aggregate numerical value. The aggregation is often additive but other forms can also be used. The additive indicators are typically seen in the traditional gravity models. Another type of indicators is the maxitive indicators that are inspired by microeconomic theory and where the attractiveness of an alternative is the utility of the choice, and where consumers only choose one alternative: the one that maximizes his/her utility.

While the axiomatic framework of Weibull ensures that the accessibility measures have desirable properties, it does not necessarily ensure a behavioral foundation or a link to consumer benefits, and hence it can be difficult to interpret their values. To ensure interpretability, an accessibility measure should be closely related to the benefits of the choice system (e.g., in utility or monetary terms). In this respect, following Ben-Akiva and Lerman (1985), the expected maximum utility function in (2) can now be interpreted as an accessibility measure; it satisfies the six conditions and directly measures the benefits and costs of the choice system in terms of utility and, as utility is assumed money-metric, in monetary terms.

For the subutility v_i , the attraction of choosing alternative i is the benefits that this choice of, for example, travel destination will give the consumer, while the distance of the alternative is given by the costs of reaching the alternative, for example, in terms of monetary and time cost of traveling to a given destination.

The Demand Curve

Having connected discrete choice models to the notion of accessibility, we now continue with deriving the consumer surplus based on the discrete choice model in (1). It will turn out that a change in expected maximum utility can be interpreted as a change in the

Marshallian consumer surplus corresponding to the discrete choice model. We first note that, for the discrete choice model in (1), the probability that a consumer chooses a given alternative i is equal to the derivative of the maximum expected utility of the discrete choice model, that is:

$$P_i(v) = \frac{\partial G(v)}{\partial v_i} \quad (3)$$

This result is known as the Williams–Daly–Zachary theorem (McFadden, 1978).

In a population of consumers with identical observable characteristics, the choice probability is equal to the market share of the alternative i . Hence, the demand curve is given by:

$$D_i(v_i) = \frac{\partial G(v)}{\partial v_i} \quad (4)$$

Demand curves are typically expressed in prices and our demand curve can easily be translated:

$$D_i(p_i) = \frac{\partial G(v)}{\partial v_i} \cdot \frac{\partial v_i}{\partial p_i} = -\frac{\partial G(v)}{\partial v_i} \quad (5)$$

The price, p_i , is understood as the full cost of the alternative, including both monetary and nonmonetary costs and in particular time costs. The coefficient $\frac{\partial v_i}{\partial p_i}$ equals -1 as utility is assumed money-metric.

This demand curve can now be used to calculate the consumer's surplus.

The Marshallian Consumer's Surplus

The Marshallian consumer surplus is a classic benefit measure in microeconomic theory. The change in consumer surplus measures the benefit to the consumers of a change in the prices that the consumers are facing. It reflects the monetary compensation that a consumer should receive in order to be indifferent with the change in prices. The change in the Marshallian consumer surplus can be calculated as the value of the area to the left of the demand curve D for the given price change from p^0 to p^1 (Mas-Colell et al., 1995).

$$\int_{p^1}^{p^0} D(p) dp$$

This is illustrated in Fig. 1.

The Marshallian consumer surplus is a convenient welfare measure as it translates utility into willingness to pay that facilitates comparisons of projects as well as comparisons across individuals.

Small and Rosen (1981) show that calculating the change in the consumer surplus as the area to the left of the demand curve also applies when demand is given by a discrete choice model.

We can therefore rewrite the expression for the Marshallian consumer surplus in terms of our discrete choice model in (1).

The Use of Consumer Surplus in Project Evaluations

If a project improves alternative 1 by increasing v_1 from v_1^0 to v_1^1 then the change in utilities can be measured with the change in the Marshallian consumer's surplus.

$$\int_{v_1^0}^{v_1^1} D_1(v_1, v_2, \dots, v_j) dv_1 = \int_{v_1^0}^{v_1^1} \frac{\partial G(v)}{\partial v_1} dv_1 = G(v_1^1, v_2, \dots, v_j) - G(v_1^0, v_2, \dots, v_j) \quad (6)$$

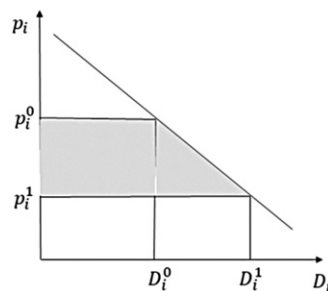


Figure 1 The change in consumer surplus as an area under the demand curve.

We see that when demand is given by the discrete choice model then the change in the Marshallian consumer surplus is the change in the expected maximum utility function, G . As we saw earlier, this change in the expected maximum utility can be interpreted as the change in accessibility, and hence the change in the Marshallian consumer surplus can be interpreted as a change in accessibility. With this formulation the measure of accessibility has a clear connection with the utility and hence is easily interpretable.

Rule-of-a-Half

In many practical applications, the change in consumer surplus following a project is computed using the following linear approximation of the demand function:

$$\int_{v_1^0}^{v_1^1} D_1(v_1, v_2, \dots, v_J) dv_1 \cong \frac{1}{2} [D_1(v_1^1, v_2, \dots, v_J) + D_1(v_1^0, v_2, \dots, v_J)] \cdot (v_1^1 - v_1^0). \quad (7)$$

This is known as the ROH.

This expression easily translates into the similar formula expressed in price changes, which is convenient for many practical applications, where it is a change in prices that is considered.

$$ROH \cong \frac{1}{2} [D_1(v_1^1, v_2, \dots, v_J) + D_1(v_1^0, v_2, \dots, v_J)] \cdot (p_1^0 - p_1^1) \quad (8)$$

The ROH has many advantages in practical applications. It only requires that the modeler knows the demand before and after the project and the change in price for the affected alternative. Hence, neither the exact utility level nor the full demand curve needs to be estimated.

The approximation in the ROH is more precise the closer the demand function is to being linear. Hence, it works best for small project changes.

With the ROH, it is quite transparent what the source of benefits is, for example, whether they derive primarily from time savings or from savings in monetary costs related to trips. Similarly, it is easy to consider distributional effects, for example, across regions or across groups of travelers with particular modes.

Another advantage of the ROH is that the planner can use standard unit costs for values of travel time and for monetary costs related to a trip, which is mandatory in many official guidelines for transport project appraisal as it facilitates comparison of appraisals of different transport projects. In these situations the price will be expressed as a function of these unit prices and the distance d_i (e.g., kilometers or time use for a given choice of destination, $p_i = d_i \cdot \text{unit price}$).

For these reasons, the ROH is common in appraisal of transport projects. However, the ROH also has some disadvantages.

The derivation of the ROH as an approximation to the user benefit is generally most precise when projects are small and the demand function is close to linear. Furthermore, using the ROH to calculate user benefits ignores income effects. When transport costs do not constitute a major part of the household budget and if the changes in transport costs are small, this is again a fair simplification. Overall, this implies that the use of the ROH to calculate the user benefits works best for smaller policy changes.

Challenges using the ROH also arise when new alternatives are added to the consumers' choice set. This is the case when, for example, a new mode or a new trip destination is introduced. In this situation, if it is not possible to define a relevant base scenario with a finite price, then it is not possible to calculate a finite price change ($p_i^0 - p_i^1$), and hence the ROH cannot be calculated directly. An alternative is then to calculate the change in consumer surplus directly by using an accessibility measure, as the expected maximum utility function G in (5), where the utility effect from a new alternative can easily be included. This is especially convenient when the choice models are of the logit type. We will illustrate this in the following.

However, while using the standard ROH assumes that the demand function can reasonably be estimated and assumed to fit the linear approximation; the use of expected maximum utility function assumes that the specification of the original utility function in (1) holds. This is more difficult to test.

Logsums

We consider now the special case, where the random utility model is a logit model. For logit models, the expected maximum utility function G has the convenient form of the so-called logsum. When this is used to calculate the change in the consumer surplus it is called the logsum approach. This will be presented in the following.

In a logit model, the market share of alternative i , the demand, is given by:

$$D_i(v) = \frac{e^{v_i}}{\sum_{j \in I} e^{v_j}}. \quad (9)$$

The expected utility of a given set of alternatives is given the logarithm of the denominator, known as the logsum:

$$G(v) = \ln \left(\sum_{j \in J} e^{v_j} \right). \quad (10)$$

The change in consumer surplus is now given by the change in the logsums:

$$G(v^1) - G(v^0) = \ln \left(\sum_{j \in J} e^{v_j^1} \right) - \ln \left(\sum_{j \in J} e^{v_j^0} \right). \quad (11)$$

The logit model is used in many traffic models. When using the logsum approach the problems of assuming linearity of the demand curve as well as the problem of considering new choice alternatives are handled. It is straightforward to add or remove alternatives from the choice set in the calculation of the logsum, so using this to compute the effect on welfare is simple. It is also easy to see that adding a new alternative will always lead to a welfare improvement.

Conclusion

Within a discrete choice model, and specifically within a logit model, a change in consumer surplus can be interpreted as a change in accessibility—the potential of opportunities for interaction. Moreover, the accessibility measure is directly linked to utility and changes in accessibility can directly be interpreted as changes in expected utility.

Relying on the discrete choice model structure to compute welfare effects has much appeal for applications. On the contrary, relying on the ROH allows one to be agnostic about the underlying model, and it facilitates the use of standard unit values and, more generally, comparisons of projects that are appraised using different traffic models.

References

- Ben-Akiva, M., Lerman, S.R., 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. The MIT Press, Cambridge, MA.
- Jara-Diaz, S.R., 1986. On the relation between user's benefits and the economic effects of transportation activities. *J. Reg. Sci.* 26 (2), 379–391.
- Mas-Colell, A., Whinston, M.D., Green, J.R., 1995. *Microeconomic Theory*. Oxford University Press, New York.
- McFadden, D., 1978. Modelling the choice of residential location. In: Karlquist, A., et al. (Eds.), *Spatial Interaction Theory and Planning Models*, North-Holland, Amsterdam.
- Small, K., Rosen, H.S., 1981. Applied welfare economics with discrete choice models. *Econometrica* 49 (1), 105–130.
- Weibull, J.W., 1976. An axiomatic approach to the measurement of accessibility. *Reg. Sci. Urban Econ.* 6, 357–379.
- Weibull, J.W., 1980. On the numerical measurement of accessibility. *Environ. Plan. A* 12, 53–67.

Producer Surplus

Chau Man Fung, CIB (Centre for Industrial Management/Traffic & Infrastructure), KU Leuven, Leuven, Belgium

© 2021 Elsevier Ltd. All rights reserved.

Introduction	242
Definition of Producer Surplus	242
Specifics in the Transport Market	243
Producer Surplus in Highways (Driving and Road Use)	244
Producer Surplus in Public Transit Supply	245
Bus	245
Train/BRT	246
(Optimal) Producer Surplus and Subsidies—Why is Negative Producer Surplus Possible?	246
Producer Surplus and Mode Choice	246
Producer Surplus and Density	247
Current Trend of Producer Surplus	247
See Also	247
References	247
Further Reading	247

Introduction

Producer surplus is a concept that is often mentioned in discussion of production and transaction. In particular, it is an important component in cost–benefit analysis (CBA). In this article, we explore producer surplus in a few steps. First, we start with giving the general definition of producer surplus commonly found in economics textbooks. Due to the nature of the transport market, it is of interest to discuss how the concept of producer surplus is applied to the transport sector. More specifically, we look into the producer surplus for transport infrastructure providers and public transport operators. The next step is to consider a bigger picture and see how producer surplus relates to transport policies and fits into the wider objective of welfare maximization. We also discuss some factors affecting producer surplus. Lastly, the future trend of producer surplus is outlined.

Note that two assumptions will hold in this article unless specified otherwise. First, it is assumed that perfect competition prevails (this holds for both transport market and other markets because it is a common practice to assume zero-profits for firms in sectors other than transportation in transport CBA; to be relaxed later); second, we focus on partial equilibrium in the transport sector and ignore the distortions in other markets.

Definition of Producer Surplus

In a market, trade can potentially occur when there is a divergence between consumers' valuation and producers' production cost of a unit of good. Assuming zero transaction costs, total gains or surpluses from trade are simply this divergence, summing across units of the good traded. The rationale is that when the downward-sloping demand (maximum willingness to pay of consumers) of the unit of good exceeds the production cost the producers have to pay, both consumers and producers will gain by trading until the marginal valuation and the marginal production cost equals. This output is the competitive equilibrium level of output, which maximizes total surplus. And this total surplus is shared between consumers and producers. How this total surplus is shared between both parties will depend on the market price of the good, and the market price will in turn depend on the form of demand and cost functions.

A second way to express the concept of producer surplus is that it is the producer's revenue minus the production cost. Algebraically, it is $PQ - TC(Q)$, where P is the price of the good, Q is the quantity of the good, and TC is the total production cost of Q units of the good. $TC(Q)$ can be represented in two ways: (1) the product of average cost and quantity of the good $AC \cdot Q$, and (2) summing the marginal production cost over the units of the good $\int_0^Q MC(f)df$. Diagrammatically (**Fig. 1**), the most common representation of producer surplus in a demand-supply diagram is the shaded inverted triangle that corresponds to the expression $PQ - \int_0^Q MC(f)df$ (while the mirror triangle under the demand curve is the consumer surplus). Alternatively, the rectangular area outlined in blue corresponding to the expression $PQ - AC \cdot Q$ also represents the producer surplus.

The expressions of producer surplus consist of revenue and cost. The revenue is mainly driven by market structure that the producers face. On the other hand, the cost is governed by the cost structure of the production of the good or service. In **Fig. 1**, the marginal cost curve is upward sloping, meaning that the production cost for an extra unit of the good is increasing. The production cost might as well be decreasing or constant, at different scale of production, depending on the production function. See **Fig. 2** for example. The marginal cost is the slope of the production cost, while the slope of the ray (connecting the origin and the point on production function) is the average cost of production. In this example, the marginal cost of production is decreasing at lower

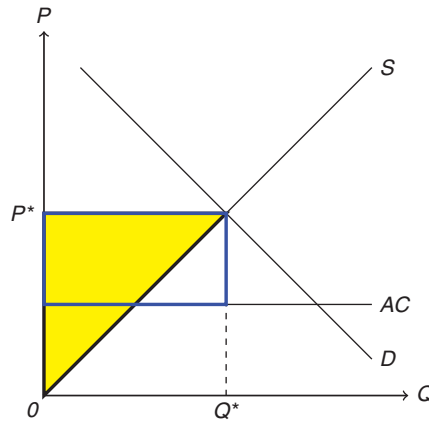


Figure 1 Producer surplus (horizontal AC).

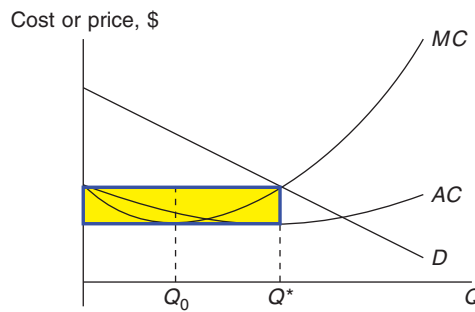


Figure 2 Producer surplus (U-shaped AC).

production level (below Q_0) and increasing at higher production level (above Q_0). The corresponding producer surplus at production level Q^* is shaded.

In fact, whenever there is a divergence between price and marginal cost (given that perfect competition prevails), a Pareto improvement is possible such that the total surplus (producer surplus, PS + consumer surplus, CS) can be increased. However, it does not necessarily mean that a total surplus increase will lead to a Pareto improvement because the loss of one party can be compensated by the gain of the other. This is inconsistent with the definition of Pareto improvement.

It is clear that the size of producer surplus depends on the cost of production. However, what is unclear is what components of production costs are included in the calculation of producer surplus. Producer's optimization problem is to maximize producer surplus, denoted by $\text{Maximize}[PQ - TC(Q)]$. Assuming no entry and exit of the market, the decision to be made is the production level. By differentiating the expression with respect to production level and set the derivative to zero, we obtain the well-known expression of $MR = MC$, which gives the optimal level of production. This equality holds regardless of the components of the total cost of production. It follows that including a fixed cost of production (which is independent of the production level) does not affect producer's decision on production level as long as he produces but the size of producer surplus is affected. The literature is not very clear on the inclusion of fixed costs. One way out is to differentiate the short-term producer surplus (gross profits without fixed costs) and long-term producer surplus (gross profits minus fixed costs). See the entries "cost functions for road transport" and "operation costs for public transport" for more discussion on production costs in a transport context.

Note that the definition of producer surplus is equivalent to the definition of profit here. We use the two terms, producer surplus and profit, interchangeably. While profit is defined as being windfall in nature in some economics texts, we do not follow this definition in our discussion.

Specifics in the Transport Market

The following discussion specific to the transport sector is valuable because the transport market is different in a few ways: first, the "product" in the transport market is trips, instead of a physical product; second, external costs of transport trips play an important role in determining the equilibrium in the transport market; third, perfect competition does not often prevail in the transport sector (in many cases of transport supply in reality, such as public transit provision and the provision of highway use, a large capital cost is invested to supply a certain capacity without extra variable cost. This results in natural monopoly due to decreasing average costs. Besides, some monopolies are results of government regulations); fourth, the investment in extra capacity does not necessarily

correspond to higher revenue. These differences, especially the third and fourth ones, make it worthwhile to discuss producer surplus in the transport sector specifically.

Producer Surplus in Highways (Driving and Road Use)

In theory, the decision of how much to produce in road transport is treated as a continuous variable that is completely flexible. In reality, this might not be the case as government regulations could be specific on specifications of road infrastructure. Here, we first consider the theoretical point of view and discuss the issues in practice.

Producer's optimization problem can be represented by $\text{Maximize}[\tau Q - \rho K(V_K)]$, where τ is the toll, V_K is road capacity, and the function K is the capital cost of capacity. Since road infrastructure lasts for many years, this fixed cost is annualized with the factor ρ . Analogous to a standard profit maximization problem, the producer chooses the level of capacity to maximize producer surplus but this is different from the standard profit maximization problem for a producer: the quantity of production (capacity) does not directly equate the quantity of consumption (trips). Instead, the capacity affects the number of trips through the generalized price of a trip $gp = P = c(Q, V_K) + \tau$, where c is the cost of the trip as a function of number of trips and capacity. Suppose that the producer can freely choose the toll. The producer's optimization problem requires the choice of both the price (toll) and the capacity. As a result, a pricing rule and an investment rule are to be satisfied to give the maximum producer surplus. This complicates the standard problem where usually each quantity corresponds to one price so deciding either will suffice.

In practice, the cost of infrastructure supply is often independent from the number of trips (users) made on the highway, and capacity choice is not entirely flexible (due to indivisibility). Once a highway has been constructed, serving an extra vehicle on the highway does not incur any extra cost (except for maintenance cost that is correlated to traffic volume). This is true until a certain limit that is constrained by the capacity of the highway. Recall that the cost of production that is constant over output is termed fixed cost. Since road infrastructure lasts for many years, this fixed cost is annualized, and compared to the revenue that the producer receives. Diagrammatically, the producer surplus is represented by a rectangle (in blue) bounded by the toll earnings (rectangle outlined in black, τQ^*) and the annualized infrastructure cost for capacity V_1 (in yellow) (Fig. 3). Note that, however (unlike the standard goods market), the cost of infrastructure (production) does not directly determine the equilibrium number of trips or the corresponding cost of road use. The cost of road use incurred by the users (AC) consists of the money cost of using the highway, and the nonmonetary cost including time cost of the trip. This is independent from the production or construction cost of the highway unless tolls are charged such that the revenues can cover the costs.

It is also common that negative producer surpluses are found in road infrastructure projects. Users are either not charged for road use or the tolls are not high enough to cover the infrastructure cost. This has been dealt with in different ways: (1) some countries (China, for instance) have a "maintain road by road" rationale so car users are charged in different ways such as road maintenance fees and registration fees. These revenues are used to pay for the negative producer surplus incurred in road projects. (2) Some countries (Germany, for instance) have earmarked other car use-related tax revenues such as gasoline taxes for road projects. Table 1 gives some examples on how different countries deal with car taxes (Fung and Proost, 2017).

It is worth pointing out that the United States has an increasing concern with decreasing gasoline tax revenues both at the federal level and the state level. On top of the fact that federal gasoline tax has not been adjusted since 1997, this is mainly due to the growth of electric vehicles and due to the better fuel efficiency of gasoline vehicles (Fisher and Wassmer, 2014). The gasoline tax revenues may not be able to cover the negative producer surpluses incurred in highways.

Certain variable costs are incurred in the production of trips on highways. An example is the maintenance cost of highways. Since maintenance cost increases with the frequency of road use by heavy trucks (number of trips), we call this a variable cost. In the

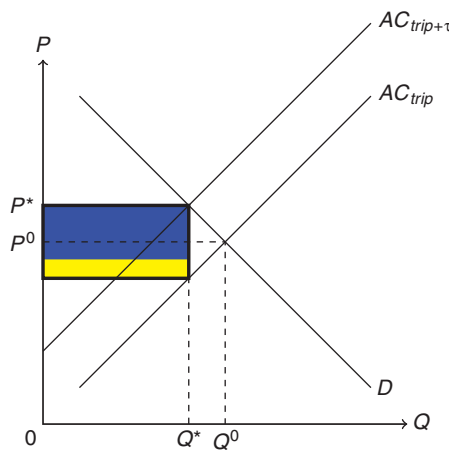


Figure 3 Producer surplus (highways).

Table 1 Car taxes

Country/ region	Federal gas tax	Federal car ownership taxes or other taxes	State gas tax	Parking fees and local tolls	Use of federal gas tax money
United States	Low—not changed since 1997	Low; taxes on tires, vehicles	Varies by state	Parking fees Occasional for public works	Redistributed to states; earmarked for highways and mass transit
China	Increased in 2009 but still low	Low; vehicle purchase tax (10%)	No	Parking fees Tolls on commercially operated toll roads; tolls on government roads in the western provinces	Redistributed to regions; by law reserved for infrastructure
EU	None but high minimum excise	None	Varies but tax competition limits increases	In some countries and cities (maximum average toll for member states)	Varies by country
Germany	High	Low; VAT (19%)	No	Parking fees Occasional urban traffic restrictions for air pollution—no tolls	By law reserved for infrastructure
Switzerland	High +	Federal vignette for highway use; VAT (8%), acquisition tax (4%)	No	Parking fees; provincial annual motor vehicle tax	Earmarked for road construction and maintenance

Source: Modified from [Fung and Proost, 2017](#).

computation of producer surplus, both fixed costs and variable costs are considered. But sometimes fixed costs are excluded because fixed costs do not affect the equilibrium; only the value of producer surplus is affected.

Producer surplus of highway capacity provision may not be relevant in many cities as the government is often the producer and its objective is often not maximizing producer surplus but instead maximizing welfare. In this case, producer surplus is then at the same footing as the consumer surplus.

Producer Surplus in Public Transit Supply

Producer surplus in public transit supply has its similarity to producer surplus in road infrastructure provision: the capacity that the producer invested in does not have a direct connection with the number of transport trips. Instead of investing in road capacity, public transport service providers invest in capital such as vehicles and drivers that results in service frequencies. The public transport service provider's optimization problem can be formulated as $\text{Maximize}[\tau Q - Kf]$ where τ is the fare, f represents the frequency of the service, and K represents the cost of service provision per frequency. It follows that the PT service provider chooses both service frequency and fare to maximize producer surplus.

The fixed cost in public transit varies among the modes: for buses, the fixed cost contains hardware such as capital cost of bus stops and capital cost of vehicles (this does not change with the number of passengers within its capacity); for other modes such as train, metro, or BRT, the fixed costs are considerably higher because the infrastructure cost such as costs of tracks are included. Just as road infrastructure, these costs for infrastructure that are durable are annualized over the period of use.

Bus

The surplus of a bus service provider is simply the total fare revenue minus the cost of bus service provision. The fare revenue depends on the fare structure such as the differentiation between peak and off-peak fares and the availability of concession schemes and monthly passes. On the other hand, the cost of service provision includes the cost of bus stops, wage of drivers, and fuel cost (fuel tax included). The service provider can set the fares and frequencies to maximize its surplus but there are often constraints in reality: there can be government regulations on the spacing of bus stops and bus accessibility of residents so frequency has to be above a certain level. It is also not uncommon for government to intervene in the fare setting of buses and avoiding time-varying fares to avoid confusion. Moreover, wages of drivers and service frequency can be influenced by unions, as they demand a certain income level for the union members. The authority may pressure the PT provider to reach a higher patronage for political reasons. These constraints affect the resulting producer surplus and producer surplus might even be negative when the costs cannot be covered by the revenues.

Analogous to the number of car trips, the number of bus trips is only determined by the service provision indirectly through the generalized price of a bus trip. The components of the generalized price of a bus trip such as in-vehicle time cost (augmented by discomfort) and waiting time are affected by frequency while the access cost to bus stops is affected by the supply of bus stops (spacing).

Train/BRT

Most of the earlier discussion on producer surplus of bus service provider is applicable to train or BRT providers but there are exceptions. First, the costs involved in providing rail or BRT service are much higher than those for bus service. The infrastructure cost of train stations, tracks, and rolling stocks are often so large that breaking even is almost impossible over the service period. In Belgium, a diabolito fee is added on top of regular train fares for some routes to pay for the construction of rail infrastructure. In some countries such as Sweden, a track charge is imposed on train operators according to the distance and weight of the trains. The former increases producer surplus, while the latter reduces producer surplus of the train operators. Note that in reality it is barely the case that the rail markets are under perfect competition; they have some monopoly power due to government regulations or natural monopoly. As a result, producer surpluses are often positive for these rail operators.

(Optimal) Producer Surplus and Subsidies—Why is Negative Producer Surplus Possible?

When we consider the expression of producer surplus (=revenue – cost), it is obvious that the negative of this expression (cost – revenue) is actually the subsidy of service provision. But at this point, a question arises: why would the producer surplus be negative in the first place? Why is the producer willing to provide the service for a negative producer surplus as the service provider is better off by exiting the market? The answer is the availability of subsidy so the negative producer surplus (loss) can be compensated. These subsidies are often provided by the government.

Another question follows: why do we need subsidies for transport service provision? The reason is that subsidy can be justifiable from a welfare point of view. Instead of having a profit-maximizing producer making the decision on service provision and the prices, we compare the profit-maximizing level of service provision and prices with the welfare-maximizing level of service provision and prices. The welfare concerning the transport market can be represented by $W = CS + PS - \text{ext} - \text{MCPF}(\text{subsidy})$, which is the sum of consumer surplus, producer surplus minus the externality and the marginal cost of public funds related to the use of subsidy. The optimal level of service provision and the optimal prices can be obtained by maximizing welfare. We could see that producer surplus is only one of the terms of the welfare expression and therefore it is unlikely that the profit-maximizing level of service provision and prices coincides with the welfare-maximizing levels. With the optimal prices and level of service determined, the optimal amount of subsidy can be calculated. In other words, there is a corresponding amount of producer surplus that is optimal (and maximizes welfare). This is the approach in [Börjesson et al. \(2017\)](#).

Optimal level of subsidies can also be perceived as a policy instrument. The authority sets the percentage of subsidies (of operating cost) and optimizes the other variables such as fares and frequencies. This is the approach in [Basso and Silva \(2014\)](#).

Note that in the previous discussion, we implicitly assume that transport service provision is managed by an independent profit-maximizing provider. This is not entirely unrealistic; in Britain, the rail service is operated by private companies. In Hong Kong, some tunnels adopt the BOT (Build-Operate-Transfer) model where the private companies build the tunnel, operate the tunnel and charge the use of the tunnel up to a certain contract period, and transfer the ownership back to the government. This is one of the ways for the government to not lose all control over the infrastructure and yet ensure that the private companies can cover their initial investment (which is probably very high) by operating and charging the service involved.

Despite the simple economic principle behind optimal subsidies, different governments tend to have very different policies and thus actual subsidy amounts vary greatly. [Table 2](#) shows the actual subsidies found in major cities (recovery rate of public transport operating expenditure by farebox revenue).

Producer Surplus and Mode Choice

Another crucial determinant of optimal producer surplus (subsidy) is mode choice. One of the common arguments for providing public transit subsidies is to divert drivers from their cars to public transit so as to reduce congestion. Public transit subsidy is justifiable as long as congestion is serious enough and enough drivers are diverted to public transit. The measure of cross-price elasticity of demand is a good indicator of how many drivers will be diverted to public transit when the fare decreases. A similar indicator is the diversion factor of cars when public transit fare changes. This diversion factor is a fraction that represents, for instance, how many of an extra bus passenger would have been car driver previously. It ranges from 0.29 to 0.25 for rural areas and metropolitan areas, respectively ([Wardman et al., 2018](#)).

Table 2 Recovery rate of public transport operating expenditure by farebox revenue

City	London	Manchester	Milan	Munich	Oslo	Paris	Stockholm
Urban population density	54.9	40.4	71.7	52.2	26.1	40.5	18.1
Recovery rate	81.2	96	41.7	64.4	63	45.5	54.3

Source: Modified from EU database.

It is important to note that producer surplus of the public transit may not be equivalent to the subsidy of the transport sector as the public transit may be cross-subsidized by the revenues from charging car use, such as tolls and fuel tax revenues.

Producer Surplus and Density

It is believed that different cities have varying amount of subsidies and one of the reasons is the difference in population density. Population density affects both the total fare revenues and the cost of public transit provision. It is logical that public transit is more intensively used and has better coverage for areas with higher population density so the patronage is higher. But it is unclear whether the actual public transit fare is higher in more densely populated areas. There is no evidence that fares in low-density areas and high-density areas vary greatly. If the optimal public transit fare is considered, it can be shown that the optimal fare is higher in more densely populated areas. This is due to the higher level of externalities such as congestion in-vehicle (crowding) and congestion on roads.

The cost of public transit provision is higher in absolute amount in densely populated areas because of the higher frequencies needed. But it is likely that the per passenger cost is lower in densely populated areas. With a higher per passenger fare and lower per passenger cost, it is clear that there is a lower public transit subsidy needed in a densely populated area. More specifically, in an optimal setting, where public transit fare and frequency are optimized, the optimal subsidy is indeed lower for densely populated areas. The reason is that the optimal frequency is governed by the square root rule so public transit provision cost increases less rapidly than the optimal public transit revenue. In reality, in a suboptimal equilibrium, it is unclear that there is a strong positive relationship between population density and public transit subsidies.

Current Trend of Producer Surplus

With the recent technological advancement in the transport sector, we do not know how producer surplus will change. Developments such as the emergence of autonomous and electric vehicles (automobiles and buses) will likely lower the production costs because both the cost of drivers (wages) and the fuel tax payment will be lower (given that the capital cost increase is not too high, as these new vehicles may be replacing existing fleet that needs replacing anyway). On the other hand, the revenue of public transit supply may be higher or lower. This is due to the following changes.

First, the rise of sharing economy such as the rapid gain in popularity of services such as Uber (share) or ride-sharing apps on smartphones definitely provides a better match for consumers in terms of the origin-destination and timing of their rides. The trips made using these new services can be generated trips or trips that are diverted from more traditional modes such as cars and public transit. The amount of public transit fare revenue will decrease and the producer surplus for public transit operators can increase or decrease depending on the relative magnitude of revenue decrease and cost decrease.

Second, the growth in popularity of modes such as cycling, shared bike, and shared scooter has a similar effect as the growth in shared cars in terms of the diversion of car trips and public transit trips. However, there are two differences: first, the patronage of public transit might increase or decrease because the convenience and accessibility of bikes and scooters can be complementary to public transit access. More passengers can access public transit more easily by using the shared bikes and scooters. Second, the rise of these new modes is likely to decrease the transport-related taxes (for cross-subsidization of public transit) because bikes and scooters are cleaner modes. Charging these modes will be almost infeasible politically and practically.

See Also

Cost Functions for Road Transport; Operation Costs for Public Transport; Natural Monopoly in Transport; The Concept of External Cost: Marginal versus Total Cost and Internalization; Long-Run Versus Short-Run Valuations

References

- Basso, L.J., Silva, H.E., 2014. Efficiency and substitutability of transit subsidies and other urban transport policies. *Am. Econ. J. Econ. Policy* 6, 1–33, doi:10.1257/pol.6.4.1.
- Börjesson, M., Fung, C.M., Proost, S., 2017. Optimal prices and frequencies for buses in Stockholm. *Econ. Transp.* 9, 20–36, doi:10.1016/j.ecotra.2016.12.001.
- Fisher, R.C., Wassmer, R.W., 2014. Perception of gasoline taxes and driver cost: implications for highway finance. *SSRN Electronic J.*, doi: 10.2139/ssrn.2537642.
- Fung, C.M., Proost, S., 2017. Can we decentralize transport taxes and infrastructure supply? *Econ. Transp.* 9, 1–19, doi:10.1016/j.ecotra.2016.10.003.
- Wardman, M., Toner, J., Fearnley, N., Flügel, S., Killi, M., 2018. Review and meta-analysis of inter-modal cross-elasticity evidence. *Transp. Res. Part A Policy Pract.* 118, 662–681.

Further Reading

- Dahlby, B., 2008. *The Marginal Cost of Public Funds: Theory and Applications*. MIT Press, Cambridge, MA.
- De Borger, B., Proost, S., 2001. *Reforming Transport Pricing in the European Union*. Edward Elgar Publishing Limited, Northampton, MA.

- Jehle, G.A., 2010. *Advanced Microeconomic Theory*. Financial Times/Prentice Hall, Harlow.
- Kreps, D., 1990. *A Course in Microeconomic Theory*. Princeton University Press, Princeton, NJ.
- Litman, T., 2004. Transit price elasticities and cross-elasticities. *J. Public Transp.* 7 (2), 37–58.
- Parry, I.W.H., Small, K.A., 2009. Should urban transit subsidies be reduced? *Am. Econ. Rev.* 99, 700–724, doi:10.1257/aer.99.3.700.
- Small, K.A., Verhoef, E.T., Lindsey, R., 2007. *The Economics of Urban Transportation*. Routledge, New York.
- Tirole, J., 1998. *The Theory of Industrial Organization*. MIT Press, Cambridge, MA.
- Varian, H.R., 2016. *Microeconomic Analysis*. Norton, New York.

The Robustness of Cost–Benefit Analyses

Morten Welde, James Odeck, NTNU—Norwegian University of Science and Technology, Department of Civil and Environmental Engineering, Trondheim, Norway

© 2021 Elsevier Ltd. All rights reserved.

Introduction	249
The Workings of CBA	249
The Robustness of CBA is Contextual	250
Major Sources of Uncertainties in CBA	251
Uncertainties in Parameter Values/Unit Prices	252
Uncertainties in Cost Estimates	252
Uncertainties in Traffic Forecasts	253
CBA as a Source of Bias in Itself?	253
Improving the Robustness of CBAs Through Ex-Post Evaluations	253
Do CBAs Answer the Right Questions?	254
The Robustness of CBA in a Future Perspective	254
Concluding Remarks	255
References	255
Further Reading	255

Introduction

Cost–benefit analyses (CBAs) are the most popular method for assessing the economic merits of public projects in general and infrastructure projects in particular. The purpose of a CBA is to inform decision-makers of the economic merits of a given intervention/project relative to the status quo (i.e., a “do nothing” scenario) and against other competing interventions/projects. The popularity of CBAs among decision-makers in the public sector is due to the ability to aggregate benefits and costs into a single measure in monetary terms. Money is an aggregate measurement that everyone (i.e., the public at large and the decision-makers themselves) can easily relate to and agree upon.

Despite the fact that CBA is a popular method for assessing the economic merits of infrastructure projects, the question remains as to whether it is robust as a decision-making tool. Here, robustness is defined as the ability of a CBA to withstand or overcome adverse conditions and/or variations in input parameters. The problem of CBAs’ robustness becomes an issue when one considers how it is used in the decision-making processes. Take for example, the case in which CBAs are used to rank transportation projects for resource allocation purposes. In such a case, a league table is produced in which projects are ranked according to their economic merits, whereby those with the highest economic returns are ranked highest, which is consistent with the economic theory of maximizing the net benefits of projects. The problem is that the rankings (league tables) are done on the basis of point estimates of CBA alone. Therefore, uncertainties regarding those estimates need to be revealed to ensure that CBA-based decision-making is robust. If the range of uncertainties is wide, it might have been possible to rank projects in completely different order than those obtained from point estimates. It is therefore imperative that if informed decisions are to be made on the basis of CBAs, analysts must attempt to estimate the level of the inherent uncertainties. Otherwise, CBA will lose its reputation as a robust decision-making tool. Furthermore, decision-makers, analysts, transportation planners, and the public at large all need knowledge about CBAs’ degree of robustness as a decision-making tool.

In this paper, we address the robustness of CBA as a decision-making tool in the transportation sector. In the next section, we summarize the workings of CBA. In the third section, we assert that the robustness of CBA depends on the context in which it is to be used and that there are different forms of CBA. We present a matrix to explain for what types of decision the different forms of CBA can be considered robust. In the fourth section, we address the most common and major sources of uncertainties in CBAs. In the fifth section, we briefly discuss the paradox that a CBA can be a source of bias itself. In the sixth section, we argue that ex-post evaluation is a good way of testing the robustness of CBAs. In the seventh section, we discuss the robustness of CBA from the perspective of the future, and, in the eighth and final section, we present some concluding remarks.

The Workings of CBA

Decision-makers in the public sector and the transportation sector in particular rely on the expected benefits and costs that a given intervention (e.g., the realization of a given transportation project) should generate throughout its lifetime when making their decisions. It is in this respect that CBA became popular as a decision-making tool because it aggregates benefits and costs into a single

measure of intervention/project worthiness and that measure is in monetary terms—a rule of measurement to which everyone can easily relate.

A CBA proceeds by first evaluating the expected change in the benefits and costs of an undertaking compared with the “do nothing” or “do the minimum” option, and all of the benefits and costs are measured in monetary terms. For example, in the case of a new road project, the monetary benefits and costs include travel-time savings, reductions in accident costs, vehicle operating costs, environmental impacts such as increased/reduced noise from vehicles, and the investment and maintenance/operational costs of the road network. The next step in a CBA is to compare the discounted monetized benefits with the discounted costs. The result of such a comparison is the net present value (NPV). If the NPV is positive, the project will be considered to be profitable from a socioeconomic perspective because its benefits will exceed its costs; otherwise, the project will be deemed unprofitable. Formally, the NPV is expressed as:

$$NPV = \sum_{i=1}^t \frac{\text{Benefits} - \text{Costs}}{(1+r)^i} \quad (1)$$

where *Benefits* is the sum of changes (measured in monetary terms) in travel time, number of accidents, vehicle operating costs, and so forth; *Costs* are the sum of investment costs and changes in operational and maintenance costs; *t* is the economic life period of the project; and *r* is the discount rate. Hence, $(1+r)^t$ is the discount factor.

Another advantage of CBA as a decision-making tool is that it can be used in the selection of the most profitable projects from a pool of projects when government investment funds are limited. In such cases, the CBA rule states that projects should be ranked according to the value of their NPV divided by the financial costs of the project provided through government funds/budgets (NPV–cost ratio). Formally, NPV–cost ratios are calculated as:

$$NPV - \text{Cost ratio} = \frac{NPV}{\text{Budget cost}} \quad (2)$$

If, for example, the NPV–cost ratio is 0.20, the interpretation will be that the return from government/societal investment in the project will be 20%. In other words, for each dollar/euro invested, there will be a return of 20 cents.

Although CBAs are performed in the transportation sector in almost all countries in the Western world, the results are often criticized on the grounds that they may depend on uncertain assumptions about the future and other parameters included in the calculations. There seems to be a general concern among decision-makers and planners that the results of CBAs are very sensitive to small changes in parameters to the extent that they often lead to different policy recommendations and/or differences in project ranking. All of these concerns relate to how robust CBA results are relative to when and how they are used.

The Robustness of CBA is Contextual

The robustness of CBAs must be evaluated from the context in which they are used in relation to the different types of CBAs available. The contexts in which CBAs are used are normally grouped into the different planning and decision-making stages of a given project. In the transportation sector, the different planning/decision-making phases are: (1) a feasibility study of the project in which the decision to continue planning for a future final decision is made, (2) the selection of an appropriate project plan (alignment/alternative) to compete for funds with other projects, (3) resource allocation between competing projects from a pool of projects where resources are limited, (4) learning about the actual outcomes of the ex-ante CBA estimates used in the decision-making stage, (5) contributing to the improvement of ex-ante CBA, and (6) learning about the accuracy of the unit prices of impacts.

There are two main types of CBAs: ex-ante and ex-post CBAs. An ex-ante CBA is performed on the basis of forecasts before a project is realized (i.e., before the actual outcomes are known). By contrast, an ex-post CBA is performed at some point after a project has been implemented, when the outcomes are known to some extent. Both ex-ante and ex-post CBAs can be further divided into two basic classes depending on when they are conducted. For the ex-ante case, the two classes are the early project phase, when the decision to go ahead with planning is made, and in the detailed project phase, when the decision to select the appropriate alignment to follow and/or the decision to build is made. For the ex-post case, the two classes involve an evaluation at some point after the project has been implemented, which is also known as in medias res evaluation, and after the project has been implemented and terminated, which is the full ex-post evaluation.

Thus, there are many different planning and decision-making stages and several different types of CBAs, which indicates that one type of CBA may be robust as a decision-making tool in one situation and not in another situation. For example, in a detailed project, an ex-ante CBA may be robust with respect to the selection of project alignments/alternatives but not for learning about actual outcomes. In the latter case, an early ex-post (in media) CBA will be more robust. Hence, the robustness of CBA is contextual, meaning that it depends both on the type of decision to be made and the type of CBA available.

We have developed a matrix for comparing the robustness of the different types/classes of ex-post CBAs according to the different planning/decision-making stages. The matrix is presented in [Table 1](#).

To aid readers in understanding when in a planning/decision-making stage a given CBA is most robust, the gray-shaded cells in [Table 1](#) indicate when the different CBAs are robust as decision-making tools. For example, it is clear that the early project phase ex-ante CBA is not robust as a decision-making tool at any stage. This is because, at early stage CBA, very little is known about the

Table 1 The robustness of CBA by planning and decision-making stages

		Types of BCA			
		Ex-ante BCA		Ex-post BCA	
		(1)	(2)	(3)	(4)
No.	Type of decision	Early project phase ex-ante BCA	Detailed project phase ex-ante BCA	Early ex-post (in medias res) BCA	Full ex-post BCA
1	Feasibility of a project; decision to go ahead planning for different options for the same project	Not robust at all; in the early project phase many input variables are uncertain	More robust; more information is available and CBA is of improved quality	N/A: the project is already built and robustness of CBA is of no relevance at this stage of decision-making	N/A: the project is already built and robustness of CBA is of no relevance at this stage/ type of decision-making
2	Selection of alignments/ alternatives to compete for funds with others of a different project	Not robust at all; contains too little or no information of potential alignments. Uncertainties are too many with respect to input variables	Robust: more information is available and uncertainties are reduced. CBA is robust for selecting best alignment, especially if the uncertainties are expected to be the same across alignments	Very robust: if low sunk costs, CBA may recommend a change of alignment. If high sunk costs, not very useful as the recommendation is proceed with the alignment already chosen	N/A: it is too late as the alignment is already chosen and built. Robustness of CBA is of no longer relevant for this type of decision-making
3	Resource allocation between competing projects	Not robust; CBA contains too little detailed information; hence, uncertainties are large	Robust; CBA is very useful for allocating resources between competing projects. Uncertainties are expected to be equally distributed across projects	N/A: the resources are already allocated; hence, considering robustness of CBA is of little relevance	N/A: it is too late and the project is already built. Robustness of CBA is of no longer relevant for this type of decision-making
4	Learning about the actual outturn of ex-ante BCA estimates used at the final decision-making	N/A; no outturn data available	Not robust; high uncertainty about future benefits and costs	Robust; some outturn data are available—uncertainties are reduced	Very robust; uncertainties are at their minimum—but some difficulties in tracing changes that occurred may be a problem
5	Contributing to the improvement of Ex-ante BCA framework	N/A; too many uncertainties	Robust; but only as a basis for comparisons to ex-post CBA	Robust; but robustness increases if ex-post CBA is performed later	Very robust; but some errors may still go undetected
6	Learning about accuracy of unit prices of impacts that enter BCA	N/A; the accuracy of unit prices is not considered for this type of CBA	N/A; does not consider the accuracy of unit prices	Robust; only if the accuracy of unit prices is evaluated	Robust; only if the accuracy of unit prices is also evaluated

parameters that will be included in it and therefore it is full of uncertainties, rendering it not robust at all. By contrast, the detailed project phase ex-ante CBA is regarded as robust in many situations, as testified by the fact that it is among the most frequently performed CBA in the transportation sector. We find that it is robust in: (1) the feasibility study phase, because it assumes that detailed data are available, (2) selecting alignments/alternatives, (3) resource allocation between projects, and (4) contributing to the improvement of ex-ante CBAs, but only as a basis for comparison with ex-post CBAs. However, the detailed project phase ex-ante CBA is not robust with regard to learning about the actual outturns and learning about the accuracies of unit prices. The rest of **Table 1** is self-explanatory. We therefore conclude that no general conclusions can be made with regard to the robustness of CBAs, as robustness is contextual with regard to the problem in question and the CBA type available. Regardless of this conclusion, CBAs have certain major sources of uncertainties that may jeopardize the robustness of ex-ante CBAs, the type commonly used for decision-making. The major sources of uncertainties in CBAs are addressed in the next section.

Major Sources of Uncertainties in CBA

Having established that the robustness of CBA is contextual, we next examine the major sources of uncertainties in CBAs. Here, uncertainty is defined as the difference between the information (CBA results) that is required to make a reliable decision and the information available at the time of the decision-making. Thus, if the difference is small, the CBA will be robust.

Bearing in mind our matrix in **Table 1**, uncertainties mostly concern the ex-ante CBAs, whereas ex-post CBAs parameters and input data are known with more certainty. It should also be noted that the literature has demonstrated that ex-ante CBAs are the most

commonly performed type of CBA, whereas ex-post CBAs are seldom performed. The same literature has also classified the sources of uncertainties in CBA into the following three main categories: (1) uncertainties in parameters (unit prices), (2) uncertainties in cost estimates, and (3) uncertainties in traffic forecasts. In the next three sections, we discuss each of these uncertainties in turn, and relate them to the two typical decisions made in the transportation sector: determining the socioeconomic merit of a project and choosing a portfolio of projects from a pool of projects.

Uncertainties in Parameter Values/Unit Prices

In a CBA, parameters values/unit prices are the monetary values that are multiplied by volumes to derive impacts in monetary terms. An example is the monetary value of time that is multiplied by the volume of time savings to derive the monetary value of the total time savings. CBA also includes fixed parameters such as the discount rate that is used to discount future benefits.

Uncertainties in parameter/unit values can seriously jeopardize decision-making. As an example, assume a CBA is being performed according to Eq. (1) and in which benefits is the product of the value of time multiplied by the change in travel time for all car user. Assume further that the following values are yielded for a project: value of time per vehicle/hour = EUR 15; change in travel time per day = 10 min; traffic volume per day = 4000 vehicles; total investment cost = EUR 50,000,000; t (the economic life period of the project) = 40 years; and r (the discount rate) = 4%. Then, the annual value of benefits is calculated as: $15 \times (\frac{10}{60}) \times 4000 \times 364 = 3,640,000$ euros. According to these values, the NPV of the project will be:

$$NPV = \sum_{i=1}^t \frac{\text{Benefits}_i - \text{costs}_i}{(1+r)^i} = 19.8 \times 3,640,000 - 50,000,000 = 22,072,000 \quad (3)$$

where 19.8 is the discounting factor, with r (the discount rate) = 4% and $t = 40$. The project is regarded as profitable in socioeconomic terms because its NPV is positive, which implies that the benefit returns net of costs will be 22 million euros/dollars.

From the earlier results, the NPV-cost ratio according to Eq. (2) is:

$$NPV - \text{Cost ratio} = \frac{22,072,000}{50,000,000} = 0.44$$

This now implies that for every dollar/euro invested there will a return of 44%.

Next, assume that at the time of decision-making, the parameter value for time/value was falsely taken as EUR 15 when its correct value was EUR 10. In that case, reliable information was not provided for decision-making, since the NPV should have been: $NPV = 19.8 \times 2,426,667 - 50,000,000 = -1,952,000$ and the project should have been judged as not profitable. The NPV-cost ratio should have been -0.04 implying that the project will lead to a loss of 4% and not a gain/return of 44% per dollar/euro invested.

The question of how the uncertainties of unit values illustrated earlier would impact on the decision should be considered. In the case of judging the economic merit of a single project, uncertainties in unit prices would render a CBA ineffective as a decision-making tool, as it might lead to false conclusions regarding the merits of a project (i.e., whether the project would be profitable or not). In the case of selecting/ranking project portfolios from a pool of projects, uncertainties in unit prices might not distort the ranking/selection because unit prices used in CBAs are uniform across projects. In such a case, uncertainties in unit prices would act as shift parameter on all of the CBA results. Thus, the only problem in the second of the two cases is that nonprofitable projects may be falsely included in a portfolio, when transportation budgets are appropriated to a sector and must be used up. In a study that used Swedish data on transportation projects it was found that changes in unit prices did not alter project rankings (Asplund and Eliasson, 2016).

Uncertainties in Cost Estimates

The costs of constructing a new project are by far the biggest cost element in the appraisal of all transportation infrastructure projects. Hence, the estimation of costs requires careful consideration. If actual costs deviate from estimates, this will have a direct impact on the project's value for money as measured by NPV and/or the NPV-cost ratio. The problem with such deviations is that they lead to ineffective decision-making because false conclusions can be drawn about project worthiness. With regard to the selection of a project from a pool of projects, uncertainties in cost estimates are particularly problematic because they may lead to the selection of economically unviable projects and consequently to inefficient resource allocation. Compared with unit prices, uncertainties in cost estimates are project-specific and therefore they will have an impact on both the decision-making with regard to determining the economic merits of projects and the selection of a portfolio of projects. Hence, if the costs of some projects and not others are overestimated or underestimated, this can affect the ranking of projects.

A number of high-profile scandals have led many to question the reliability of cost estimates used in the CBAs of transportation projects: Denver International Airport opened 16 months late and cost the city USD 560 over budget; in Sweden, the Hallandsås Tunnel opened 19 years too late and cost 12 times more than originally planned; and the Edinburgh tram service opened GBP 375 million over budget and 3 years late. There are many other examples of transportation megaproject failures.

In some cases, overruns may be justified in order to accommodate changing needs that were not known ex-ante (i.e., at the time of decision-making, such as the necessary capacity expansions to cater for increased traffic). In such cases, increased costs may be accompanied by a corresponding increase in benefits. In other cases, overruns threaten both the economic viability of the project in question and the future of other public infrastructure projects. Thus, it may be concluded that the robustness of ex-ante CBAs depends greatly on the robustness of cost estimates. The robustness of cost estimates is also addressed elsewhere in this Encyclopedia.

Uncertainties in Traffic Forecasts

In addition to construction costs, the other main avenues through which bias can enter the CBAs in transportation projects are the traffic forecasts. If actual traffic deviates significantly from forecasts, this will ultimately affect the estimated economic benefits and, potentially, the ranking of projects, similar to the way uncertainties in cost estimates have an impact on CBAs.

The consequences of inaccurate traffic forecasts depend on the context within which the new facility is built. Historically, transportation projects have been justified on the basis of existing traffic alone, but large-scale road and rail projects may result in changes in travel mode, frequency of travel, distance traveled, destination, or a combination of all of these factors. In addition, in the long term, transportation projects may result in changes in land use through residential and commercial development. Omitting induced traffic from benefit calculations may thus lead to underestimations of the user benefits from a new transportation facility. In many cases, such as high-speed rail, fixed link projects or urban light rail, the purpose of investment is precisely to induce more people to travel in order to realize agglomeration benefits or urban regeneration.

However, if congestion is a problem or will become a problem during the appraisal period, underestimation of traffic may imply a shorter period of relief from congestion and hence an overestimation of benefits. Omitting induced demand can therefore bias the assessments of the economic viability of proposed projects, especially when there is a latent demand for more capacity.

Despite the crucial role of traffic forecasts, data limitations have meant that ex-post studies are relatively rare, but most of the studies that have been carried out have found that road projects have typically experienced more traffic than expected (Nicolaisen and Driscoll, 2014). A recent example is the widening of the M25 around Greater London, where hard shoulders on parts of the orbital motorway were converted to permanent additional running lanes in 2014. Within 1 year of opening, traffic growth was 10%–13%. An evaluation report found that the traffic increase over parts of the widened sections was greater than before the capacity improvements were made. However, if the scheme had not been built, journey times probably would have deteriorated further (Highways England, 2018). By contrast, the mean inaccuracy for rail projects is usually negative, as ridership often fails to meet expectations. A well-known example is the Channel Tunnel Rail Link, which connects the United Kingdom and France. The tunnel not only exceeded its budget by 80%, but passenger traffic was only one-third of what the project's promoters had expected. Thus, as in the case of cost estimates, it may be concluded that the robustness of ex-ante CBAs depends greatly on the robustness of traffic forecasts.

CBA as a Source of Bias in Itself?

It has been contended in the literature that, paradoxically, one of the main sources of bias and inaccuracy in transportation appraisal may be the CBA itself. For example, the Transport Planning Society in the United Kingdom has suggested that the strong reliance on NPV–cost ratios in appraisal and decision-making leads project promoters and their consultants to abuse the system by artificially boosting the values of positive elements in the CBA and downplaying the costs. This suggestion is in line with Oxford economist Bent Flyvbjerg's allegations that transportation planners routinely engage in strategic behavior and rent-seeking to justify their own projects or those of their clients (Flyvbjerg, 2009). However, the claims and allegations have not been proven by scholars, nor have they been supported by any empirical data. Moreover, not all countries require positive binding corporate rules (BCRs) for projects to be implemented, and the results from different studies vary with regard to forecast accuracy. Thus, there is insufficient evidence to accuse transportation planners in general of fraudulent behavior, as there is no statistical evidence to prove that CBA values are subject to tampering.

Improving the Robustness of CBAs Through Ex-Post Evaluations

Apart from the criticisms of CBAs being inappropriate tools for decision-making because they do not include all relevant factors worth considering in decision-making and some important impacts are not valued in monetary terms, there is criticism related to robustness, which the transportation literature has addressed to a lesser extent. The criticism is that CBAs are rarely conducted ex-post (i.e., at some point after a project has been implemented).

Ex-post evaluations, if conducted regularly, can improve the robustness of CBA as decision-making tool in the following ways: they can (1) provide input for the development of CBA techniques over time, (2) provide important lessons for CBA users who are looking for practical evidence regarding delivering transportation projects, (3) show results to policymakers who want to know whether schemes deliver planned benefits and whether transportation policy is as effective as intended, and (4) show communities whether their concerns are being addressed effectively.

Ex-post CBA evaluations are conducted by substituting ex-ante estimated values by ex-post actual values as far as data availability allows. Although there has been an increase in ex-post evaluations of transportation projects over the last 10–15 years, very few countries have standardized and mandatory frameworks for such analyses. Among the few countries that carry out annual ex-post CBA evaluations are the United Kingdom, Norway, France, and New Zealand. In the United Kingdom, Highways England carries out annual Post Opening Project Evaluation (POPE) 1 and 2 years after the opening of selected projects. France has “permanent observatories” that are established by law and used to collect data to facilitate a detailed evaluation of major schemes. Norway and New Zealand both conduct ex-post assessments of some of the impacts of a small sample of projects every year. In addition, the European Union regularly initiates ex-post studies of EU-funded projects. Other countries, including Sweden, carry out ex-post evaluations of individual projects from time to time (Nicolaisen and Driscoll, 2016).

The CBAs in all of the aforementioned countries are based on uncertainties to varying degrees. However, several Swedish studies have shown that CBA rankings are robust against uncertainties in input parameters and that even omitted demand due to land use effects have had a limited effect on project rankings. The uncertainties cause negligible losses of total net benefits, and project selections based on CBAs deliver far higher total benefits than projects for which selections are made at random (Börjesson et al., 2014).

Do CBAs Answer the Right Questions?

Regardless of the inaccuracies discussed in the preceding sections, the robustness and usefulness of a CBA will depend primarily on whether it is able to capture the policy aspirations of decision-makers. A number of studies have criticized CBA as an ex-ante appraisal tool because unacceptable environmental effects may have a very limited impact on a project’s NPV. Others have criticized the ethical foundations of CBA, in particular the lack of emphasis on equity and distributional effects. There may also be an intergenerational equity problem, due to the use of a discount rate, and there is no guarantee that a project with a positive NPV will be acceptable from a sustainability perspective.

In practice, decision-makers may have goals (both short term and long term) that may be at odds with the results of an economic appraisal. For example, in many CBA frameworks, the travel times of motorists are valued higher than the travel times of transit passengers. This may favor car-based strategies for urban areas instead of strategies based on improving options for transit, walking, and cycling. In other cases, reduced travel time through better access roads to cities may encourage deagglomeration. Better roads encourage people to buy larger and potentially cheaper houses farther away from the city. This leads to increases in traffic, which is valued as induced traffic in the CBA, but in practice the amount of time a motorist spends traveling may be unchanged while the effects feed into the economy through increased property prices. In many road projects time savings may benefit individuals, but offer little benefit to society. We need to recognize that not all transportation investments improve the economy, and not all projects meet local and national objectives, such as reducing peak-hour congestion and emissions. Thus, from a societal point of view, the economic impacts as measured by the CBA may be both positive and negative.

A CBA answers merely one of the questions that decision-makers may be interested in, namely the value for money generated by a scheme. The appraisal of projects should be based on a broad framework, which includes a strategic perspective that outlines the rationale for a project. The economic case, as measured by the CBA, is there to support the strategic case, and not to dictate decisions. The strategic case should act as a filter to prevent projects with impacts that are counter to current policy, even if they have a positive NPV. Thus, appraisal should be based on an analysis of multiple objectives and not just on monetized costs and benefits as measured by the CBA. Otherwise, CBA may not be regarded as useful or robust in solving the problems that decision-makers face.

The Robustness of CBA in a Future Perspective

Despite all of its flaws and weaknesses, CBA has been a useful tool for transportation agencies and decision-makers for many decades. It has aided in the selection of projects that improve the economic welfare of societies and it has helped to prevent economic waste. Despite its often-limited practical use, CBA has added transparency to planning and decision-making processes in which a wealth of detail can obscure the overall picture.

However, in a future that may look dramatically different from the past, the old ways of appraisal may no longer be fit for purpose. We have discussed changing objectives and the role of economic appraisal in a broader framework of project governance, but today we are also faced with changes and challenges that may leave planners with less solid ground on which to stand.

We live in a transitional age. Although vehicles have remained relatively unchanged for 50 years, the age of electric and autonomous vehicles is upon us. In Norway, the market share of electric vehicles as a proportion of all sold vehicles increased from just over 0% in 2010 to over 30% in 2018. In other countries, the change has been less significant, but there is no doubt that vehicle technologies are about to change dramatically and take on forms that we may not be able to foresee today.

New technologies and new behaviors are disrupting long-established patterns. Services such as Uber are breaking down the divide between vehicle ownership and the transportation industry. Connected and possibly autonomous vehicles allow us to utilize road space more efficiently. App-based services encourage new forms of mobility that may be difficult to predict. Additionally, attitudes to travel are changing. While traveling the world might have been seen a sign of success in the past, “flight shame” is set to become the buzzword of 2019 in Northern Europe. All of these new trends translate into a major challenge for transportation

appraisals. Accurate forecasting of outcomes is set to become even more challenging than in the past, and the answers give today may not be appropriate in the future. In an age of change, the risk of misallocation of resources through spending on projects or services that may prove irrelevant in the future increases considerably.

The answers to a more uncertain future should not be to abandon ex-ante appraisal and CBA altogether. Instead, we need to acknowledge that there may not be one correct solution to a problem, and that transportation planning needs to incorporate multiple scenarios and solutions. We also need to take a broad view of possible outcomes and aim policy toward the direction that is desirable for society.

Concluding Remarks

In this chapter, we have discussed the robustness of CBA with regard to its usefulness in decision-making, given uncertainties about crucial input parameters and uncertainties about future needs and priorities.

We have argued that the main purpose of CBA is to rank projects in a project portfolio or to select project alternatives, such as road alignment, in a given project. In this context, CBA is robust, as inevitable inaccuracies of costs and benefits are unlikely to alter project rankings. However, a range of studies has demonstrated a high degree of uncertainty regarding costs and traffic in transportation projects worldwide, and therefore there is potential for improvement in ex-ante estimates and methods. We have argued that ex-ante CBA should be combined with a framework for ex-post evaluation so that we can learn about actual project performance, and increase transparency and accountability.

Although widespread, the often-limited use of CBA results suggests that decision-makers may have other priorities than those that can be summarized in monetary terms. There is a need to recognize that CBA is merely one element in a framework for project governance and that the economic case for a project should be one element in a broad strategic appraisal.

Project uncertainty does affect not only individual projects, but also the context in which projects are selected and implemented. In a world of changing needs, priorities, and technologies, we must recognize that CBA may not represent all the answers for future scenarios.

References

- Asplund, D., Eliasson, J., 2016. Does uncertainty make cost-benefit analyses pointless? *Transp. Res. Part A* 92, 195–205.
- Börjesson, M., Eliasson, J., Lundberg, M., 2014. Is CBA ranking of transport investments robust? *J. Transp. Econ. Policy* 48 (2), 189–204.
- Flyvbjerg, B., 2009. Survival of the unfittest: why the worst infrastructure gets built—and what we can do about it. *Oxf. Rev. Econ. Policy* 25 (3), 344–367.
- Highways England, 2018. Smart motorway: all lane running. M25 J5-7 Monitoring Third Year Report. Highways England, 30p.
- Nicolaisen, M.S., Driscoll, P.A., 2014. Ex-post evaluations of demand forecast accuracy: a literature review. *Transp. Rev.* 34 (4), 540–557.
- Nicolaisen, M.S., Driscoll, P.A., 2016. An international review of ex-post project evaluation schemes in the transport sector. *J. Environ. Assess. Policy Manag.* 18 (1), doi:10.1142/S1464333216500083.

Further Reading

- Anguera, R., 2006. The channel tunnel—an ex post economic evaluation. *Transp. Res. Part A* 40 (4), 291–315.
- Eliasson, J., Börjesson, M., Odeck, J., Welde, M., 2015. Does benefit–cost efficiency influence transport investment decisions? *J. Transp. Econ. Policy* 49 (3), 377–396.
- Love, P.E.D., Sing, M.C.P., Lavagnon, I.A., Newton, S., 2019. The cost performance of transportation projects: the fallacy of the planning fallacy account. *Transp. Res. Part A* 122, 1–20.
- Lyons, G., 2018. Is transport planning fit for purpose? *Proceedings of the 16th Annual Transport Practitioners Meeting*, 5–6 July, Oxford.
- Odeck, J., 2004. Cost overruns in road construction—what are their sizes and determinants? *Transp. Policy* 11, 43–53.
- Odeck, J., Kjekreit, A., 2019. The accuracy of benefit-cost analyses (BCAs) in transportation: an ex-post evaluation of road projects. *Transp. Res. Part A* 120, 277–294.

The GDP Effects of Transport Investments: The Macroeconomic Approach

James Laird^{*,†}, Daniel Johnson^{*}, ^{*}Institute for Transport Studies, University of Leeds, England, United Kingdom; [†]Peak Economics, Inverness, Scotland, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	256
Neoclassical Approaches	256
The Production Function and the Solow–Swann Model of Regional Growth	256
Gains From Trade and Comparative Advantage	258
Migration and Displacement	258
Endogenous Growth and New Economic Geography	259
An Introduction to Endogenous Growth and New Economic Geography	259
Intracity Transport Investment: Economic Growth and Displacement	259
Interregional Transport Investment: Economic Growth and Displacement	259
Transport Investment and Long-Run Growth	260
Transport-Economy Frictions	260
Evidence of Transport Investment's Impact on the Economy	260
Conclusions	262
References	262
Further Reading	262

Introduction

The gross domestic product (GDP) impact of transport investment has a strong resonance with policy. Investment in transport infrastructure is seen as one of the ways of growing economies and reducing spatial disparities. Economic growth is an objective of transport investment by international agencies such as the World Bank, the Asian Development Bank, as well as by the European Commission (EC) and national governments. The EC's Cohesion Fund for 2014–2020 is €63.4 billion and has the objective of reducing spatial disparities in the EU and will either finance transport or renewable energy-oriented projects. Economic growth, particularly in the north thereby reducing spatial disparities, is a key objective of the High Speed 2 (HS2) and the Northern Powerhouse Rail high-speed lines in England.

When thinking about the GDP effects of such transport investment at a regional or country level, we need to consider how transport investments feed into the regional and national economies. In this paper, we therefore take the reader through the key macroeconomic mechanisms starting from neoclassical theory through to the more modern New Economic Geography (NEG) theory in Sections “Neoclassical Approaches,” “Endogenous Growth and New Economic Geography,” and “Transport-Economy Frictions.” In particular, we draw out the implications of transport investment on productivity growth and the displacement of economic activity. In Section “Evidence of Transport Investment's Impact on the Economy,” we then present and discuss the evidence on how transport investment has effected the economy historically. We draw out variations by estimation approach, industry, mode, country, region, and income levels, and examine whether historical timing is of any relevance, before concluding in Section “Conclusions.”

Neoclassical Approaches

The Production Function and the Solow–Swann Model of Regional Growth

Classical approaches to understanding economic growth, whilst ultimately insufficient to give us a full understanding of the role of transport in growing an economy, are very enlightening as they describe some of the key building blocks which more complete theories utilize. The starting point invariably is a production function, which relates aggregate output (Q), to inputs of labor (L) and capital (K):

$$Q = f(L, K) \quad (1)$$

Imposing a Cobb–Douglas form on this function gives us an equation of the form:

$$Q = AL^\alpha K^\beta \quad (2)$$

The key implication of this model is that transport infrastructure investment gives a one-off shift in productivity. This is because at time t_1^* growth returns to the long-run rate of growth. The second important implication is that the productivity growth occurs in two stages: an immediate jump in productivity followed by a gradual, but diminishing, rate of growth. Over this period of time growth exceeds the long-run rate of growth in the economy (here depicted by the rate x).

From a policy perspective the time interval between t_1 and t_1^* is important. A short time interval would mean that it would be possible to focus exclusively on the level of productivity and output once the economy has returned to its long-run rate of growth, whilst a long interval would mean it is necessary to understand the dynamics of this change. Evidence suggests this adjustment period is very long with the gap to steady state narrowing by between 1.5% and 3% per year (Barro and Sala-i-Martin, 2004). This implies that 50% of the gap is narrowed (half life) between 22 and 45 years, and 75% of gap is closed between 45 and 91 years. Ultimately therefore, despite being a one-off shift, the effects of a transport investment on regional growth is likely to be felt over a very long time.

This model also informs us that if the performance of the transport networks worsens over time, for example, for reasons of say demographic change, technological change, or income growth leading to higher levels of congestion, then the locational efficiency of a region (the A term) will decrease. This will result in a contraction of the production function. Some transport investment may therefore be required to maintain existing productivity levels if transport networks are expected to become more congested in the future.

Gains From Trade and Comparative Advantage

Transport infrastructure investment also facilitates trade between regions and between countries. Davide Ricardo in the early 19th century was the first to describe the gains that can be made from trading. He showed that if transport costs (and other costs to trade) fall sufficiently to permit trading, then countries will specialize in the industry for which they have comparative advantage. In his example Portugal was more efficient at producing both wine and cloth than England. It therefore has absolute advantage over England. Its comparative advantage though is in producing wine, as it uses less inputs to produce wine than it does to produce cloth. In his example England's comparative advantage was in producing cloth, rather than wine—as England used less inputs to produce a unit of cloth than it did to produce a unit of wine. He then showed that if trade costs, of which transport costs are part, fall sufficiently to allow trade between the countries then both countries will specialize in the industry in which they hold comparative advantage. This would be wine for Portugal and cloth for England. In doing so, more cloth and wine would be produced for the same inputs of labor. That is by specializing in the industry of comparative advantage and trading: productivity increases, incomes increase, and society is better off. This is also known as gains from trade.

Ricardo's model is very simple and it requires that labor and capital will not migrate between countries or regions, and that different regions or countries hold comparative advantage in different industries. Nonetheless, the model is illustrative of the productivity benefits we gain from specializing in the tasks or industries we are best at. However, by specializing there is a need to trade otherwise we cannot produce all the commodities our region, city, and locality require. Transport is therefore part of this story of specialization. How important this ability to trade is to our overall productivity is a good question that is difficult to answer. We can draw to an extent off the situation which Gaza found itself in during the blockade of 2007–10 which is estimated to have reduced Gaza's productivity by 23% (Etkes and Zimring, 2015).

We can therefore see that transport investment, as well as directly boosting productivity through an outward shift in the production function, can permit increases in productivity through specialization and increased interregional trade.

Migration and Displacement

Labor and capital also migrate between countries and regions. In a simple neoclassical model, workers migrate to maximize their return to providing labor, and owners of capital invest in other regions to maximize their return on investment. This migration of both labor and capital between regions leads to an equalization in both returns on capital and labor between regions. This is because the return on labor decreases in the region it migrates to, as labor supply in the recipient region increases. Simultaneously, the return on labor increases in the donor region as labor supply contracts. The movement of labor between the regions eventually leads to an equalization of wage rates between regions. The same is true for capital. If we introduce migration into the Solow–Swann model presented earlier it can be shown that the speed of convergence to the steady state speeds up. That is, the economy will grow more quickly, up until the point that growth reverts to the rate at which the population grows.

In reality of course there are many barriers to full factor mobility, which prevent rates of return equalizing between regions. There are distance, information, and money costs. There are different risks between regions, and full factor mobility is not usually possible (e.g., different segments of the labor market are more mobile than others and some economic activities are fixed in location). Notwithstanding these criticisms, this simple model is useful as it illustrates how a productivity shock, such as a transport investment, destabilizes the equilibrium. Following a productivity enhancing transport investment, mobile factors of production will shift to the more productive region expanding the supply of capital and labor in this region, and reducing the supply in the donor regions. At the new equilibrium the rates of return on capital and labor will be higher in all affected regions than before the transport investment. Furthermore, activity will have been displaced to the region that received the productivity shock. This displacement of economic activity between regions by transport investment is a key economic response, and arguably gives rise to the most notable change in economic output (GDP) in the vicinity of the transport investment.

Endogenous Growth and New Economic Geography

An Introduction to Endogenous Growth and New Economic Geography

A key critique of neoclassical growth theories is their inability to adequately explain, within the model, why disparities remain between regions and countries. Long-run growth in developed economies has been sustained over many decades, and developing economies have not “caught up.” Within countries, some regions “lag” behind. Theories on endogenous growth and the field of NEG offer insights here, and give a more nuanced view as to the role of transport investment in boosting economic performance.

Endogenous growth emphasizes the role of human capital as a source of long-run economic growth through the creation and diffusion of knowledge. Education, research and development, and new ways of working, both within business and as a consequence of the institutional setting, all form mechanisms by which long-run growth can occur.

NEG links agglomeration economies, internal economies of scale, and transport costs as sources of imperfect competition. This imperfect competition then permits the spatial variation in economic performance between regions that neoclassical models cannot predict. NEG emphasizes that there exist centripetal forces that will centralize economic activity (market size effects, thick markets, and knowledge spillovers and other pure external economies), and centrifugal forces that will disperse economic activity (immobile factors, land rents, congestion, and other pure diseconomies). A policy shock, such as a transport investment, will then disturb the existing balance and the relative strengths of the different centripetal and centrifugal economic forces will determine whether economic activity will centralize or disperse.

Later we explore the implications of this on the economic performance in the context of an intraregional (and urban) transport investment, an interregional (or intercity) transport investment, and the implication of transport for long-run economic growth.

Intracity Transport Investment: Economic Growth and Displacement

As mentioned earlier and discussed elsewhere in this volume there are external benefits to clustering—agglomeration economies. An intracity transport investment would be expected to raise productivity by the mechanisms outlined earlier, and in so doing will displace economic activity to the city. This will raise city productivity further, and in so doing will induce a circular-and-cumulative causation process that continues to displace economic activity to the city until a new equilibrium is reached. The key point from a transport investment is that the presence of agglomeration economies can therefore increase the productivity of a city and the effect of regional displacement beyond that expected by the simple neoclassical model.

Interregional Transport Investment: Economic Growth and Displacement

Possibly the largest contribution of NEG with respect to the economic impact of transport investment is with respect to interregional or intercity transport investment. If economic activity is mobile then the choice of which region to locate production can be shown to vary with interregional transport costs. With high transport costs production will occur in every region; with medium transport costs production will centralize to those with the largest markets, and with low transport costs production will disperse to regions with low factor costs. That is, as transport costs fall spatial inequalities between regions increase, but as they continue to fall this trend will reverse and spatial inequalities between regions start to diminish. This is the bell-shaped curve of NEG.

Noble Laureate Paul Krugman illustrated this process with a simple two-region model. In this model there are two regions: the core and the periphery. The core is larger than the periphery, whilst the periphery has lower factor costs (land rents, wages, etc.). There are no transport costs within a region, but there are transport costs between the regions. Firms also benefit from internal economies of scale in production. Populating this model with some hypothetical data (Table 1) Krugman shows that with high transport costs firms will locate a factory in both regions and there will be no trade between regions. This is because total costs of production of a single large factory in the Core is 13 units ($=10 + 3$), a single large factory in the periphery is 16 units ($=8 + 8$), whilst having two smaller factories one in each region is lowest at 12 ($=12 + 0$). With medium shipping costs it becomes cost effective to close one factory and expand output at the other factory. The factory at which production centralizes is located in the largest home market to minimize transport costs—which is the Core in this scenario. Under medium shipping costs total costs of production in the Core is 11.5 units. That is, a lowering of transport costs leads to a trade between regions and a centralization of economic activity to the largest regions. A further reduction in interregional transport costs will eventually lead to production locating in the regions

Table 1 Economies of scale exist in production (either internal or external)

<i>Production in</i>	<i>Production costs</i>	<i>High shipping costs</i>	<i>Medium shipping costs</i>	<i>Low shipping costs</i>
Core	10	3	1.5	0
Periphery	8	8	4	0
Both	12	0	0	0

Source: Krugman (1991).

with lowest factor costs—which in this model is the Periphery. Here, with low shipping costs the total costs of production in the Periphery is 8 units ($=8 + 0$).

Reorganization effects (i.e., displacement) at a regional level in response to the transport cost reductions are therefore substantial, and in this hypothetical example dwarf the productivity gains. The key contribution relative to the neoclassical model is that economic activity may locate to either end of the interregional transport link depending on the existing level of development in the country. Arguably therefore investment in interregional transport links in developed countries with mature transport networks and relatively low transport costs will lead to a dispersion of economic activity. In contrast transport investment in developing countries with immature transport networks with associated high transport costs will lead to a centralization of economic activity.

Transport Investment and Long-Run Growth

In the neoclassical model transport investment gives a one-off productivity increase (with associated increase in economic output), albeit this is likely to take several generations to fully materialize. After this period, growth returns to the pre-intervention long-run growth rate. Endogenous growth theory emphasizes that the source of long-run growth is through the creation and diffusion of knowledge. Obviously transport investment does not directly influence this source of growth, but it may indirectly do so by strengthening agglomerations. This is because one of the sources of agglomeration economies is learning effects relating to the creation, diffusion, and accumulation of knowledge. Transport investments, by indirectly strengthening learning effects within a regional economy, may therefore affect the long-run rate of growth. The original NEG models were primarily concerned with the spatial organization of economic activity, but in more recent NEG models spatial organization and growth are jointly determined. These theoretical models suggest changes in location of economic activity are associated with faster growth in all regions. These theoretical models are, however, relatively recent, and more research on them and their empirics is needed.

Transport-Economy Frictions

The preceding discussions have been on the presumption that factor markets are not distorted and will clear. If there exist structural weaknesses in the economy that prevent firms accessing a labor force with appropriate skills, land suitable for development, or capital to finance investment, then the growth and reorganization effects described may not occur. Furthermore, in a developed country with a mature transport network, it is viewed that transport is a complement to more important factors of economic growth—the most important of which is an available workforce (Bannister and Berechman, 2001). This is a cautionary note as transport's ability to stimulate the economies of lagging regions or countries will be limited if they there are structural weaknesses present in those regions/countries. As an intermediate good, the pass-through mechanism needs to be working well for transport to influence economic growth and boost GDP.

Evidence of Transport Investment's Impact on the Economy

The evidence of transport's impact on the economy is largely from studies focused on estimation of Cobb–Douglas production functions based on aggregate country or state level data: these yield an elasticity capturing the sensitivity of output to changes in the stock of public capital stock, of which transport infrastructure is a key component. Referring back to Eq. (3), such studies cannot distinguish between effects through capital stock K and effects through transport services influencing the A term: they both involve estimating coefficients on public capital.

We see in national and regional economic data that there is a clear correlation between output measures (GDP) and transport investment levels. However, in understanding the evidence a key concern is to identify causality, that is, does transport actually drive economic growth or is it vice versa or is there no direct relationship at all? There are a host of studies in the literature to draw on here. There is a separate paper in this Encyclopedia that addresses the issue of causal identification and we do not therefore discuss it here, beyond saying that inadequate methods to address causality can result in overly large estimates of the impact of transport investment on economic output.

An early key study was conducted by David Aschauer (Aschauer, 1989) who sought to explain the impact of public sector investments on US output using an annual time series dataset of aggregate public capital stock and total factor productivity measures from 1949 to 1985. His work identified a strong relationship between real output and the stock of core infrastructure (including highways, mass transit, airports but also utilities) during this time period which was quantified in an output elasticity of 0.24; that is, an increase in the stock of core infrastructure of 10% would lead to an increase in real output of 2.3% ($=1.1^{0.24}$). From this evidence Aschauer attributed the decline in productivity in the United States since the 1970s to underinvestment in public infrastructure.

Whilst Aschauer's work offered compelling evidence, there were a number of questionable aspects that prompted a host of subsequent studies using aggregate country or state level data and more sophisticated estimation techniques. Key findings of these studies are best summarized through consideration of meta-analysis study findings (Melo et al., 2013; Holmgren and Merkel, 2017).

Meta-analysis studies in this area seek to explain variation between many separate studies that examine the relationship between transport investment and the economy in order to understand whether there is consistent evidence of such a relationship underpinning the studies.

One such study was conducted in 2013 by Patricia Melo, Daniel Graham, and Ruben Brage-Ardao looking at a dataset of 33 studies that focused on the output elasticity of transport. The mean value of the output elasticity from these studies was 0.06 with a median of 0.016: substantially lower than Aschauer's estimates and with a considerable variation in the values as captured by a high standard deviation. They used the meta-analysis to capture the effect of various study characteristics on the resultant estimates with the aim of explaining the wide-ranging results from the literature and offering guidance to policy-makers on returns to transport investment.

Whilst some of these constituent studies were based on time series approaches, others included panel data analysis using repeated observations over time on spatial units (e.g., US states). Analysis of panel data allows studies to control for estimation bias arising from unobserved time invariant heterogeneity. Other examined sources of variation included mode, sector, country, spatial scale, income level, and time frame.

Estimates such as Aschauer's, based on pure time series, yielded higher values than subsequent studies: this may be partly due to spurious associations between output and infrastructure stock which both grow over time but may not be causally linked. Subsequent studies (Holtz-Eakin, 1994) that controlled for unobserved heterogeneity through panel data approaches yielded lower elasticity values. This may be because more prosperous areas are likely to spend more on public capital leading to a positive correlation between area-specific effects and public capital: if not controlled for this will manifest itself in a higher coefficient on transport infrastructure investment. Other studies investigate potential bias in the elasticity estimate by employing methods which correct sources of reverse causality: governments may choose to invest more in growing regions to promote further growth or conversely to invest more in lagging regions to harmonize economic fortunes across regions. The idea here is to find a proxy for transport infrastructure purged of endogeneity with the output measure—often based on long lags of infrastructure or related measures that cannot be linked to current levels of output (Duranton and Turner, 2012). Melo and her coauthors find that controlling for this reverse causality yields higher elasticity measures suggesting transport investment on average across their sample is targeting toward lagging regions. More urbanized areas are associated with higher levels of transport infrastructure and output and if a measure of urbanization is not included Melo and her coauthors' findings suggest it can lead to upward bias in output elasticities. Conversely they find studies that do not control for congestion yield lower estimates and this may hamper productivity benefits from agglomeration.

In terms of other sources of variation, there are higher elasticities for manufacturing industries relative to other sectors, reflecting that these sectors are more transport intensive: Melo and her coauthors report average values of 0.082 for manufacturing sectors. They also find higher elasticities are evident for road-based infrastructure (0.088) compared to other modes (e.g., rail and air are 0.037 and 0.027, respectively), although this may reflect the higher stock of roads than other modes, that is, a 10% increase in road infrastructure is a larger absolute increase than a 10% increase in rail infrastructure. There are also higher values for the US-based studies possibly reflecting the higher use of roads in the United States.

As an aside it is useful to understand what these elasticities might mean in terms of rates of return on investments. For example, Grice (2016) in recent work on national accounting measures estimates UK transport stock to be worth £569 billion in 2014. If we take an example of a £1 billion transport investment: this implies an increase of 0.18% in the infrastructure stock. An elasticity of 0.06 applied to this proportional increase in the infrastructure stock amounts to an increase in GDP of 0.01%. Applied to a real level of GDP of around £2 trillion this would give a GDP increase of 211 million per year, which is a rate of return of 21.1%. That is the infrastructure pays for itself within 5 years.

The evidence is more mixed for non-European studies. However, Canning and Bennathan (2001) in a report for the World Bank, estimate output elasticities for paved roads at 0.05 for countries in the lowest income quartile and 0.04 for countries in the highest income quartile compared to a median value of 0.09. This contrasts with the theoretical models presented earlier, which suggested that poorer countries should experience the highest returns. However, in some poorer countries productivity is likely to be constrained by low levels of human capital and disparate and disconnected infrastructure networks thus creating frictions that prevent the economy exploiting the improved infrastructure. In richer countries diminishing returns have set in terms of benefits from expanding networks beyond a critical level of capacity (such as the US interstate highway network), which is why we expect and see low elasticities there. Middle-income countries may sustain higher elasticities due to gains from trade and higher levels of other inputs such as physical and human capital (linking to endogenous growth theory).

Studies using national data yield higher estimates; these studies are more likely to pick up spillover effects than more regionally disaggregate ones. In terms of time there are higher elasticities in the long run (more than 5 years) than the short run (less than 5 years). In the time frames observed, such studies do not tell us whether we are adjusting back to a steady state or moving to a new long-run rate of growth (endogenous growth). Clearly it is difficult to make comparisons against the theory when the adjustments require much longer time series than data currently permits.

Their mean value of elasticity of 0.06 in Melo's study, whilst considerably lower than Aschauer's, is itself thus possibly an overestimate given the older studies do not address sources of upward bias. Another similar meta-analysis exercise found that studies centered around 1965 or earlier have significantly higher estimates than more recent ones (Holmgren and Merkel, 2017). Further, earlier studies may be based on older US data where large network effect benefits from the development of the interstate highway network in the 1950s and 1960s connecting previously isolated states, were a one-off benefit.

Conclusions

There are strong theoretical reasons to believe that transport will affect the size of an economy. It can shift the production function outwards by affecting the stock of capital in the economy and by making the economy more efficient. This is a one-off shift, albeit taking several generations to materialize, and a change in long-run growth will only occur if transport investment can change the rate of knowledge creation and diffusion—the source of long-run economic growth. One potential avenue it may do this is by affecting the effective size of agglomerations. Transport investment is a destabilizing force, as by changing transport costs it changes the balance between regions and countries. This can lead to productivity gains from agglomeration and specialization, but is also likely to lead to reorganization and displacement effects. Whether economic activity centralizes or disperses is a function not only of existing transport costs, but also of the relative strengths of centrifugal and centripetal economic forces.

The empirical evidence backs up the theory to the extent that it clearly shows a linkage between output and transport investment as captured by positive output elasticities for changes in the stock of transport infrastructure. There is no “one-size-fits-all” elasticity, with notable variations by industry, region, spatial scope, and income levels. More recent work tends to yield lower elasticities and suggests that early literature presented overestimates due to issues with data and estimation approaches and also due to the one-off gains associated with development of networks. The datasets are not long enough to make clearer linkages with the types of growth that is enabled and this is an area where more evidence is needed, although other empirical studies on agglomeration impacts may provide more insight as to the extent to which such growth may be endogenous.

References

- Aschauer, D.A., 1989. Is public expenditure productive? *J. Monet. Econ.* 23 (2), 177–200.
- Banister, D., Berechman, Y., 2001. Transport investment and the promotion of economic growth. *J. Transp. Geogr.* 9 (3), 209–218.
- Barro, R.J., Sala-i-Martin, X., 2004. *Economic Growth*, second ed. MIT Press, Cambridge, MA.
- Canning, D., Bennathan, E., 2001. The social rate of return on infrastructure investments. World Bank Research Project on “Infrastructure and Growth: A Multicountry Panel Study”. World Bank, Washington, DC.
- Duranton, G., Turner, M.A., 2012. Urban growth and transportation. *Rev. Econ. Studies* 79 (4), 1407–1440.
- Etkes, H., Zimring, A., 2015. When trade stops: lessons from the Gaza blockade 2007–2010. *J. Int. Econ.* 95 (1), 16–27.
- Grice, J., 2016. National accounting for infrastructure. *Oxf. Rev. Econ. Policy* 32 (3), 431–445.
- Holmgren, J., Merkel, A., 2017. Much ado about nothing? A meta-analysis of the relationship between infrastructure and economic growth. *Res. Transp. Econ.* 63, 13–26.
- Holtz-Eakin, D., 1994. Public-Sector Capital and the Productivity Puzzle. *The Review of Economics and Statistics* Vol. 76 (No. 1), pp. 12–21.
- Krugman, P., 1991. *Geography and Trade*. MIT Press, Cambridge, MA.
- Melo, P.C., Graham, D.J., Brage-Ardao, R., 2013. The productivity of transport infrastructure investment: a meta-analysis of empirical evidence. *Reg. Sci. Urban Econ.* 43, 695–706.

Further Reading

- Acs, Z., Sanders, M., 2014. Endogenous growth theory and regional extensions. In: Fischer, M.M., Nijkamp, P. (Eds.), *Handbook of Regional Science*, Vol. 1. Springer, London, pp. 193–211, Chapter 11.
- Ferrari, C., Bottasso, A., Conti, M., Tei, A., 2018. *Economic Role of Transport Infrastructure: Theory and Models*. Elsevier, Amsterdam.
- Lafourcade, M., Thisse, J-F., 2011. New economic geography: the role of transport costs. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *A Handbook of Transport Economics*. Edward Elgar, Cheltenham, pp. 67–96, Chapter 4.
- Lakshmanan, T.R., 2011. The broader economic consequences of transport infrastructure investments. *J. Trans. Geogr.* 19 (1), 1–12.

The Mohring Effect

Hugo E. Silva, Instituto de Economía and Departamento de Ingeniería de Transporte y Logística, Pontificia Universidad Católica de Chile, Santiago, Chile

© 2021 Elsevier Ltd. All rights reserved.

Introduction	263
The Mohring Effect in a Simple Framework	263
The Mohring Effect in the Original Framework	264
Empirical Evidence on the Mohring Effect	265
Conclusions	265
Acknowledgment	266
References	266
Further Reading	266

Introduction

Herbert Mohring (1928–2012) was one of the most influential economists in the field of transportation economics. He was one of the first economists to study land values and its relationship with transportation costs in the monocentric model. He is also well known for his contribution to the theory of cost recovery for highway construction, especially for the joint work with Mitchell Harwitz that studied the conditions under which road provision is exactly financed with optimal road tolls. He has also contributed to the urban transportation policy analysis by studying the efficiency of dedicating road capacity exclusively to buses and comparing its benefits with those from welfare maximizing pricing policies. The present article discusses his contribution on economies of scale and subsidization in public transportation.

The Mohring Effect is the result that an increase in the demand for public transportation induces a decrease in the waiting time costs for all users when it is dealt with an increase in the frequency of the services. Its origin is in the article titled “Optimization and Scale Economies in Urban Bus Transportation” published in the *American Economic Review* in 1972. The term Mohring Effect has also been used in a broader way and has included other positive indirect effects of increased demand, such as reductions in access time costs from increased route density. When the average costs, including those of the users, decrease with the demand, the social welfare maximizing public transportation fare does not cover operating costs. Therefore, the Mohring Effect, which is in a way analogous to a positive externality, is the most important argument from an economic efficiency standpoint for public transport subsidization. The other arguments for subsidization, such as reductions of car travel externalities and distributional concerns, may be better addressed with instruments directly targeted at those objectives.

Recent estimations using data from large and congested cities have shown that the Mohring Effect is substantial and that it plays a key role in justifying the presence of public transport subsidization. It has also been shown that the Mohring Effect is more significant for buses than rail and subways and that is more significant for off-peak periods than for peak periods. The intuition behind these results is that the Mohring Effect is stronger for low frequencies that imply significant waiting times, and frequencies are higher in peak periods and urban rail services.

Section “The Mohring Effect in a Simple Framework” of this article discusses a simplified version of Mohring’s framework that illustrates the nature of the Mohring Effect and why it is an efficiency argument for subsidizing public transportation. Section “The Mohring Effect in the Original Framework” summarizes Mohring’s original model and his numerical results. Section “Empirical Evidence on the Mohring Effect” reviews the recent empirical evidence on the size of Mohring Effect and estimations on how much of current and optimal subsidies in some cities are due to the Mohring Effect. It also discusses more generally under which conditions the effect should be strong. Finally, Section “Conclusions” concludes providing a summary and some present challenges related to this topic.

The Mohring Effect in a Simple Framework

Mohring, in his pioneer work published in 1972, started the literature on the microeconomics of urban bus transportation. The aim of his paper, which gave rise to what is now known as the Mohring Effect, was to study scale economies in the provision of public transportation. In other words, [Mohring \(1972\)](#) studied the short- and long-run cost functions to characterize the nature of increasing returns to scale in bus operations and thus to assess the nature of the subsidy required if marginal cost pricing is in place.

Transportation differs from the classic analysis of pricing because consumers (i.e., travelers) play a producing role by providing their own time as a resource for production. In using bus services, they must supply their time, distributed at least into access

(walking), waiting, and in-vehicle time. These time costs borne by the travelers can be substantial: estimates for the value of time for commute trips are typically around 50% of the wage rate, and the values of waiting and access time are 2 or 3 times larger.

Mohring's model considers a bus corridor or line, where the demand is spatially distributed. As the focus is on costs, demand is treated parametrically. He examines four components of costs: (1) bus company operating costs, (2) passengers' walking time costs, (3) passengers' waiting time costs, and (4) passengers' in-vehicle time costs. The sum of these components is sometimes called in the literature as the total value of resources consumed.

The variables that the bus agency optimizes are the frequency of the buses and the spacing of bus stops. A key assumption in the model, which is rooted in transportation engineering, is that the waiting time is a function of the headway between buses. Therefore, as the headway is the inverse of the frequency, the average waiting time costs are inversely related to the frequency of buses.

A primary result of the analysis, inspired in William Vickrey's previous work, is that, when bus speed is constant, the frequency that minimizes the expenditure is proportional to the square root of the demand for bus services. This has become a result commonly known as the "square root formula" for the optimal frequency of bus services, and it has been extended in various directions and found to be fairly robust (Jara-Díaz and Gschwender, 2003).

In his paper, Mohring modeled in a detailed fashion the realized bus speed as a function of frequency, the distance between bus stops, and free-flow speed. However, the Mohring Effect can be understood without this complexity. This section considers that bus speed is constant and therefore the square root formula is valid, and it assumes for simplicity that the number of bus stops as well as their location are fixed. The following section deals with more complex models.

The implications behind the square root formula are intuitive. As the average waiting time is inversely related to the frequency, and the frequency is proportional to the square root of the demand, the average waiting cost decreases with total demand. This result is the essence of the Mohring Effect: an increase in demand reduces waiting costs for all passengers, as it should be dealt with an increase in frequency.

The Mohring Effect is, therefore, one source of increasing returns to scale in bus operations because waiting costs grow less than proportionally with demand. In other words, due to the Mohring Effect, the marginal cost could be less than the average cost and the welfare maximizing bus fare does not cover operating costs. As a consequence, the Mohring Effect is a driving force of the efficiency of public transport subsidization.

The strength of the Mohring Effect and, therefore, to what extent it justifies public transport subsidization depends mainly on the frequency implied by the level of demand. Consider a simple example of a bus route with low demand for which the optimal frequency is two buses per hour. Assume for simplicity that the average waiting time is half the headway so that in this example is 15 min. If demand is doubled and the optimal frequency follows the square root formula, it increases by a factor of the square root of 2, which yields a new frequency of approximately three buses per hour. The increase in frequency from two to three buses per hour reduces the average waiting time for all passengers from 15 min to 10 min. While this can be substantial, as demand grows and the frequency is higher, the effect becomes smaller. For a frequency of 30 buses per hour, the average waiting time is 1 min, and the potential decrease in waiting costs is limited. Naturally, the real effect is complex as more interrelationships come into play. We now turn to discuss some of them.

The Mohring Effect in the Original Framework

As explained in the previous section, Mohring's model included bus company operating costs and passengers' walking, waiting, and in-vehicle time costs. The previous section assumes that buses travel at a constant speed for the ease of exposition, but, in Mohring's model, the speed that buses achieve is endogenous, and the analysis is more detailed. We summarize the main features of the model and his numerical results.

First, in the model, the speed of a bus depends on the number of passengers that board and alight each bus along the route and, therefore, it depends on the number of passengers and the frequency. Furthermore, the time spent in starting and stopping at the bus stops is modeled together with a probability of buses not stopping at all bus stops.

Second, Mohring studies two types of spatial distributions of demand. The "steady state" route in which demand is uniformly distributed along the bus line, and the "feeder" route in which the same average number of people per hour board buses but they all alight at the end of the route. For both travel demand patterns, the number of equidistant bus stops per mile is an optimization variable.

As a result of the complexity of the interaction between these features of the model, it is not possible to obtain closed-form expressions for the frequency and spacing of bus stops. Nevertheless, it is clear that, in contrast to the simplified version of the previous section, an increase in the bus frequency decreases both waiting and in-vehicle time. It also increases bus-operating expenditures but less than proportionally as the travel times are reduced. Finally, increasing the number of bus stops increases in-vehicle time for users and operating costs, but reduces walking costs.

To shed light on the degree of scale economies and the size of the optimal subsidy for urban buses, Mohring performed numerical analyses based on data that reflect approximately the Minneapolis–Saint Paul metropolitan area. The results are intended to cast light into the magnitude of optimal subsidies if the effects of bus pricing on highway congestion and the marginal cost of public funds are ignored.

The numerical analysis performed by Mohring is detailed, but the result that arguably better conveys the conclusion is as follows. For conditions that reflected the situation at the time in the Minneapolis–Saint Paul metropolitan area, results are that the weighted

average gap between long-run marginal costs and average costs is 57% of the bus company operating costs. Therefore, the optimal subsidy at those conditions is large and justified by the reduction in user time costs of increased frequency.

The combination of the theoretical and numerical analysis summarized earlier gave rise to the Mohring Effect, which, in a broad sense, is the result that as demand increases the average user costs decrease if the public transport supply responds optimally. This could be due to increases in frequency that lead to a reduction in waiting time and in-vehicle time due to lower boarding and alighting time. It could also be due to increases in bus route density, which leads to decreased access time. The Mohring Effect, thus, is a source of the presence of economies of scale in public transport provision and one of the main efficiency arguments for subsidizing public transport.

Empirical Evidence on the Mohring Effect

The vast majority of the empirical literature on public transportation pricing focuses on the optimality of public transport subsidization (Proost and Van Dender, 2008; Basso and Silva, 2014; Börjesson et al., 2017). In doing so, the studies have commonly estimated an aggregation of all the effects that could give rise to subsidization, and little can be said about specific effects. For example, the size of the optimal subsidy is a function of the strength of the Mohring Effect but is also determined by the degree of scale economies of the bus company operation. The subsidy also heavily depends on the unpriced externalities in urban travel. If car travel is not priced in a welfare maximizing way, lowering the public transport fare may decrease car usage and the unpriced negative externalities such as road congestion and pollution. This reduction potentially implies the need for higher subsidies. Finally, there could be negative externalities from bus travel as well, such as crowding, pollution, and congestion that could make the subsidies not efficient.

The primary empirical evidence about the size of the Mohring Effect and its role in the efficiency of public transport subsidization is a study with data from Washington, Los Angeles, and London by Parry and Small (2009). The study considers most of the relevant features that need to be taken into account in determining the efficiency of public transport subsidies. It includes several modes for traveling, demand substitution across modes and times of day, transit supply that is responsive to changes in ridership, negative externalities from motor vehicles (congestion, pollution, and accident), in-vehicle crowding in public transport, and transit-user wait and access costs. Importantly, the transit agency can adjust the route density, service frequency, vehicle size, and load factor.

The available estimations are not exactly the size of the Mohring Effect, but the marginal welfare gain of increasing the subsidy that is due to the Mohring Effect. More precisely, it is the marginal benefit minus the marginal cost of increasing the frequency and the route density in welfare maximizing fashion as a response to the increased demand that is a consequence of the fare reduction. The marginal benefit arises from the reduction in wait costs from increased service frequency and the reduction in access costs from increased route density. The marginal cost is from increased agency supply costs from increased vehicle size and the increase in passengers' crowding costs from higher load factors. Therefore, while the estimations are not the Mohring Effect per se, they are a relevant measure of how important it is in shaping the optimal pricing policy.

At the baseline situation of the study, which already had subsidies between 50% and 80%, the size of the Mohring Effect is significant. For peak operations, the marginal welfare gains due to the Mohring Effect from increasing the subsidy in 1 cent per mile in peak periods are 0.34 cents per mile in Washington, 0.29 in Los Angeles, and 0.19 in London. In off-peak periods, where frequencies are lower, the gains due to the Mohring Effect are substantially larger: 2.00 cents per mile in Washington, 1.73 in Los Angeles, and 1.74 in London. For rail, the marginal welfare gains due to the Mohring Effect are smaller and vary between 15% and 50% of the gains estimated for buses. The only exception in the study is peak-rail in London, in which higher costs fully offset the benefits due to the Mohring Effect. This is most likely because London's subway is already very frequent, crowded, and its network of lines is dense.

Parry and Small (2009) also estimate what the optimal subsidy would be and what percentage is due to the Mohring Effect. For off-peak bus services, the optimal subsidy is above 90% of operating costs for the three cities and the proportion of the optimal subsidy that is due to the Mohring Effect is between 55% and 61%. For peak bus services, the results are case specific. For London, the optimal subsidy is above 90% of the operating costs, and only 15% of that subsidy is due to the Mohring Effect. For Los Angeles, the optimal subsidy is 74%, and the proportion due to the Mohring Effect is 22%, while in Washington it is 46% and the proportion is also 46%. For rail services, the optimal subsidy in the three cities is substantial (above 78%) in all periods, but the share due to the Mohring Effect is lower. In peak periods the Mohring Effect is responsible for 0%–10% of the subsidy and for 21%–41% in off-peak periods.

In summary, the recent empirical estimations confirm Mohring's early findings that increased demand for public transportation that is met with an optimal adjustment of supply leads to a substantial reduction in user costs. This result, which is the broad definition of the Mohring Effect, justifies subsidization on efficiency grounds at least before taking into account the distortions that would be introduced in the process of raising public funds for subsidies.

Conclusions

This article argues that the concise definition of the Mohring Effect is the result that an increase in the demand for public transportation induces a decrease in the waiting time costs for all users when it is dealt with an increase in the frequency of the

services. This article provides an overview of a simple framework to understand the Mohring Effect and of the original framework developed by Mohring. This article also reviews the recent empirical estimations of the size and relevance of the effect.

The Mohring Effect can be understood in a general fashion as the result that, as the demand for public transportation increases and the public transport supply responds optimally, the average user costs decrease. This could be due to an increased frequency that leads to a reduction in average waiting times, to increased route density that leads to reduced average walking times, and route structure changes that lead to lower transfer costs, among others.

Recent estimations of the Mohring Effect in congested cities of developed countries show that it is an important driving force of the efficiency of public transportation subsidization. While results are convincing, the evidence is still scant. A natural avenue for future research is to estimate the Mohring Effect in different situations and to assess empirically to what extent it is strong enough to justify public transport subsidization.

The present article has focused on the cost-side of the problem and has not dealt with the private, profit-maximizing, provision of public transport. There has been recently a debate in the literature about whether the Mohring Effect justifies subsidization when a monopoly firm supplies the service. The answer seems to depend on the assumptions on public transportation demand, user cost heterogeneity, and operating cost functions (see [Gómez-Lobo, 2014](#), for a synthesis). Studying the relationship among the Mohring Effect, the strategic interactions between public transport providers, the regulatory framework, and the need for subsidization seems a natural place for further research.

Acknowledgment

The author gratefully acknowledges financial support from the Complex Engineering Systems ISCI (grant CONICYT FB0816) and the Center of Sustainable Urban Development CEDEUS (grant CONICYT/FONDAP 15110020).

References

- Basso, L.J., Silva, H.E., 2014. Efficiency and substitutability of transit subsidies and other urban transport policies. *Am. Econ. J. Econ. Policy* 6 (4), 1–33.
- Börjesson, M., Fung, C.M., Proost, S., 2017. Optimal prices and frequencies for buses in Stockholm. *Econ. Transp.* 9, 20–36.
- Gómez-Lobo, A., 2014. Monopoly, subsidies and the Mohring effect: a synthesis. *Transp. Rev.* 34 (3), 297–315.
- Jara-Díaz, S., Gschwender, A., 2003. Towards a general microeconomic model for the operation of public transport. *Transp. Rev.* 23 (4), 453–469.
- Mohring, H., 1972. Optimization and scale economies in urban bus transportation. *Am. Econ. Rev.* 62 (4), 591–604.
- Parry, I.W., Small, K.A., et al., 2009. Should urban transit subsidies be reduced? *Am. Econ. Rev.* 99 (3), 700–724.
- Proost, S., Van Dender, K., 2008. Optimal urban transport pricing in the presence of congestion, economies of density and costly public funds. *Transp. Res. Part A Policy Pract.* 42 (9), 1220–1230.

Further Reading

- Basso, L.J., Jara-Díaz, S.R., 2010. The case for subsidisation of urban public transport and the Mohring effect. *J. Transp. Econ. Policy* 44 (3), 365–372.
- Mohring, H., 1976. *Transportation economics*. Ballinger, Cambridge, MA.
- Mohring, H., 1979. The benefits of reserved bus lanes, mass transit subsidies, and marginal cost pricing in alleviating traffic congestion. In: Mieszkowski, P., Straszheim, M. (Eds.), *Current Issues in Urban Economics*. Johns Hopkins University Press, Baltimore, MD, pp. 165–195.
- Savage, I., Small, K.A., 2010. A comment on 'subsidisation of urban public transport and the Mohring effect'. *J. Transp. Econ. Policy* 44 (3), 373–380.
- Van Reeve, P., 2008. Subsidisation of urban public transport and the Mohring effect. *J. Transp. Econ. Policy* 42 (2), 349–359.
- Zhang, J., Lindsey, R., Yang, H., 2018. Public transit service frequency and fares with heterogeneous users under monopoly and alternative regulatory policies. *Transp. Res. Part B Methodol.* 117, 190–208.

Public Transport Fare and Subsidy Optimization

Qianwen Guo, Zhongfei Li, Department of Finance and Investment, Business School, Sun Yat-sen University, Guangzhou, China

© 2021 Elsevier Ltd. All rights reserved.

Introduction	267
Public Transit Pricing	267
Public Transit Subsidization	269
Conclusions	269
References	270

Introduction

Public transportation is an indispensable component of the multimodal transportation system, especially in the urban environment. This article introduces public transportation fare and subsidy from the lens of optimization. The analysis is limited to traditional public transportation services, which are characterized by shared vehicles, fixed routes, and predetermined timetables. Depending on the vehicle used, there are different forms of public transport, such as bus transit, rail transit, ferry, and even airline services.

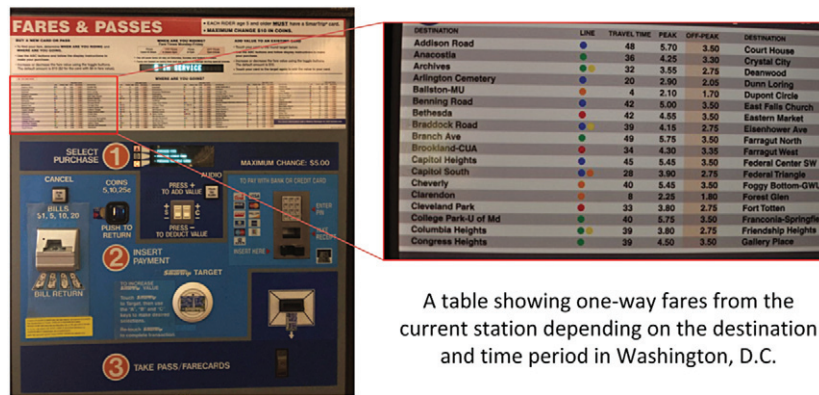
Public transportation systems across the world differ significantly in ownership, market share, fare scheme, subsidy availability, and regulatory policy. In the United States, the majority of public transport services is provided by governments at all levels (local, state and federal) with the objective of providing a viable travel alternative to driving. The market share of public transit is minimal (e.g., 5%) at the national level in the United States, while the share can be over 50% in major metropolitan areas, such as New York City (NYC). In the United States, due to the relatively low public transit demand, the farebox recovery ratio (i.e., fare revenues divided by total operating expenses) is usually quite low. For instance, the Metropolitan Transportation Authority in NYC has an agency-wide farebox recovery ratio of 35.5% in 2018 (MTA, 2018). By contrast, in major cities in East Asia, such as Hong Kong, public transit services can be operated profitably by private operators, thanks to relatively high population densities, large shares of commuting by public transit, and the innovative “rail plus property” business model (Leong, 2016). European transit systems fall between Asia and North America, facing various degrees of financial difficulty to achieve the break-even point, similar to the United States.

In the public transit pricing and subsidization literature, various demand functions, optimization objectives, constraints, and other modeling elements are used to reflect the regional characteristics. This article proceeds as follows: The second section discusses fare schemes, optimization methods for determining fares, and how optimal fares are affected by various economic and social factors; the third section analyzes whether external subsidies are needed, how such subsidies should be optimized, and how welfare losses due to subsidization can be mitigated; and the last section concludes this article with a summary.

Public Transit Pricing

In most public transport systems, passengers need to purchase tickets to take public transit. Although some transit services (such as inner-city circulators) are provided for free, it is rare to provide free transit services throughout a network or city. Nonetheless, there are examples of zero-fare public transit in the world: in Estonia's capital Tallinn public transport is free to all registered residents (Gray, 2018). When there are no fare revenues, funds from other sources such as government taxes or advertising are required to cover the transit operating expenses.

In most public transit systems across the world, the flat fare is adopted, partially because it is easy to communicate and simple to collect. If the same price is charged regardless of the travel time period, distance, direction, and quality of service (e.g., local vs. express bus), the actual operating cost of providing a passenger trip is not reflected. Intuitively, it should cost more to carry a rider over a larger distance during peak hours. To account for the operating cost differences, differentiated fares can be adopted based on travel distance or time of day. Cervero (1981) evaluate fairness in public transit pricing by comparing flat and differentiated fares. In Cervero (1981), it is defined to be fair if the costs of providing transit services are equitably and efficiently distributed among all transit users, despite other definitions of fairness in the economics literature. Based on data from three transit agencies in California, this paper finds long-haul riders are heavily crosssubsidized by those short-haul, off-peak transit users. It is suggested that differentiated fares should replace flat fares due to their advantages in improving equities among transit users. However, several obstacles to differentiated pricing especially political acceptability and institutional barriers are discussed. Cervero (1986) focuses on differentiated pricing by time-of-day involving transit agencies both in the United States and across the world. The main purpose of time-of-day differentials is to shift demand from peak hours to off-peak periods; however, empirical evidence suggests the degree of success is quite low. Off-peak travelers are much more price sensitive to discounts than peak travelers to surcharges. Ling (1998)



A table showing one-way fares from the current station depending on the destination and time period in Washington, D.C.

Figure 1 Fare table on a fare vending machine in Washington, DC.

further compares both pricing strategies by a few criteria including total revenue, ridership, passenger-km, and consumer surplus. Several key factors are identified that influence the choice between flat and differentiated pricing, which are the decision objective of the interested party, the initial market demand, and the demand elasticity.

Flat and differentiated pricing structures coexist in the world. For example, Washington Metropolitan Area Transit Authority (WMATA), the operator of the rail transit system in the Washington, DC area, charges a higher fare if the travel occurs in peak periods and the travel distance is larger. The fare also increases as the travel distance grows. **Fig. 1** shows a fare table from which a passenger is able to determine the fare needed to travel from the current station to a desired destination in a travel time period. By contrast, a flat fare is adopted by the New York City Subway, regardless of the travel distance or time-of-day. Different fare collection methods are used to support the flat and differentiated fares. In Washington, DC, a rail transit rider needs to tap in at the origin and tap out at the destination, while in NYC, card swiping is required only at the origin subway station.

A few theoretical studies on transit fare pricing optimization are reviewed first. Since Mohring's seminal work ([Mohring, 1972](#)) on urban bus transportation, the microeconomic modeling framework has been widely adopted and adapted to study transit pricing policies. In early studies (e.g., [Newell, 1979](#); [Wirasinghe and Ghoneim, 1981](#); [Chang & Schonfeld, 1991](#)), both the network structure and demand pattern are greatly simplified in order to solve the resulting model analytically. In such studies, fare is usually jointly optimized with other operational parameters, such as headway, bus size, route spacing, zone size, etc. When demand is assumed to be fixed, which means demand does not vary with fare or service level, usually the total cost including both the suppliers' cost and users' cost is minimized. When demand is sensitive to fare and other service parameters, the optimizing objective is to maximize the social welfare or the total profit, depending on the type of the transit operator, public, or private. Social welfare is defined as the sum of user surplus and profit. [Jara-Díaz and Gschwender \(2003\)](#) provide a review of the microeconomic models for public transit operations.

The setting of fare is also affected by the organizational structure. [Sun and Schonfeld \(2015\)](#) design a simple optimization model to show that a public operator with the objective of social welfare maximization would set fare to be equal to the marginal operating cost, which yields an operating deficit due to the fixed operating cost. The deficit is then covered by external subsidies to maintain financial feasibility. In the benchmark model, the public operator is unaware of the possibility of a negative profit and its implications. In an extended model, [Sun et al. \(2016\)](#) consider a bilevel modeling structure where the upper-level government agency provides subsidies to and imposes regulatory policies on the lower-level transit operator. Since the government agency is aware of the cost of public funds (which are mainly from tax revenues and are used to cover potential operating deficits), the passengers' surplus could be compromised to cover the fixed component of the operating cost, depending on how expensive the public funds are. The resulting fare is higher than the one adopted by a public operator, which maximizes social welfare only. [Sun et al. \(2016\)](#) also present and compare fares in other cases such as total profit maximization and break-even with 0 subsidies. [Sun and Zhang \(2018\)](#) enhance the existing literature by considering a many-to-many demand pattern (meaning transit riders travel from multiple origins to multiple destinations) and multiple user groups with different values of time and price elasticities. Through both analytical and numerical analyses, they analyze the effect of crosssubsidization: travelers who directly contribute to the capacity bottleneck are subsidized by others who do not.

In addition to theoretical analyses, there are many studies involving real-world public transit systems. [Proost and Dender \(2008\)](#) calculate the optimal transport price structure and conduct welfare analyses using data from Brussels and London. Their analyses suggest charging nearly zero transit fares in peak hours does not lead to significant welfare improvements. [Savage \(2010\)](#) conducts a time-series analysis of the Chicago Transit Authority's bus operations from 1953 to 2005 to examine the evolutions of fare and service frequency over time. One of the major findings is that more services are supplied, and higher fares are charged, than the socially optimally ones. [Farber et al. \(2014\)](#) conduct a numerical analysis to evaluate the social equity impacts of introducing distance-based fare in place of flat fare by the Utah Transit Authority. Their analyses show that low-income, senior and nonwhite individuals could benefit from the new fare structure, although the impact is not uniform geographically. As illustrated in the above studies, to determine the fare level properly, both economic, and social factors should be considered.

Public Transit Subsidization

Transit subsidies have been controversial for at least a few decades, although it is fairly common across the world to finance public transit systems in cases of operating deficits through external subsidies. In the literature, there is no consensus on the justification of transit subsidies. The so-called “Mohring effect” (Mohring, 1972) is frequently referenced as the rationale for the transit subsidy. The Mohring effect highlights the economics of scale on the user side. As more passengers use transit, additional vehicle runs are added, leading to a decreased waiting time. As more passengers might result in denser transit routes and stops, the access time may also drop. Therefore the increase in passenger demand leads to a lower travel cost for existing travelers, contrary to the usual traffic congestion effect, where additional road users decrease the overall travel speed and service quality for existing users. While transit operators provide vehicles, users also contribute to the “production” by investing their own time in travel (Savage, 2010). Because of the increasing returns to scale, the marginal cost of operating transit services is lower than the average cost and a subsidy is thus required to achieve the marginal cost pricing. Gómez-Lobo (2014) provides a synthesis of the Mohring effect.

Transit subsidies are also justified on other grounds such as positive externalities. Public transport is believed to bring a range of benefits, such as traffic congestion relief, environmental impact mitigation, and transportation emission reduction. One more rationale for transit subsidy is the second-best pricing to automobile externalities, because the use of private vehicles in congested areas and peak hours is priced below marginal cost (Elgar and Kennedy, 2005).

Despite the above-mentioned rationales, there are conflicting conclusions in both the theoretical and empirical literature regarding the necessity of transit subsidy. For further discussion on the controversy of transit subsidy see Sun et al. (2016). Even when the transit subsidy is deemed necessary, merely paying the subsidy does not guarantee that the benefits of subsidy are realized. As indicated by Else (1985), the subsidy’s effectiveness and efficiency are affected by the form of the subsidy, the operator’s objective, and its relation to the subsidizing body (e.g., government).

A few studies are then reviewed on how the subsidization scheme affects the social welfare. As a centralized optimization framework is unable to incorporate the decision interactions between multiple decision makers (subsidizer and transit operators), Sun and Schonfeld (2015) propose a bilevel optimization framework to study the effect of external subsidies and policy constraints on the transit operations. Three subsidy schemes are considered, the first two of which are based on the supply side and the third is based on the demand side. In Scheme 1, a fixed amount of subsidy is provided to the operator for each passenger trip, as a means to incentivize the operator to sustain the production level; in Scheme 2, subsidies are used only to compensate for the operating cost on the basis of passenger trips; and in Scheme 3, subsidies are provided to passengers to encourage the transit use. Three analytical models are thus formulated and solved. Such three schemes perform quite differently depending on the subsidy level and two important thresholds for distinguishing them are identified. The over-supply of services is identified under both Schemes 1 and 2, indicating inefficiencies of subsidies. Sun and Schonfeld (2015) also report that subsidization without regulation yields socially worse outcomes than the combination of subsidization and regulation. Sun et al. (2016) extend Sun and Schonfeld (2015) by analyzing the implications of the cost of public funds. They find the optimal profit is zero at optimality in their model, which means the transit operators break even after subsidies and positive profits reduce the social welfare. They also find when the cost of public funds exceeds the shadow price of the financial constraint, the subsidy is unjustified.

Next, a few empirical studies on transit subsidies are reviewed. While most other studies concluded that public transit subsidies can comprise a system’s efficiency and overall performance, despite their positive impacts on a system’s effectiveness, Karlaftis and McCarthy (1998) draw a different conclusion that there do not appear to exist a general relation between public transit operating subsidies and transit performance, based on the empirical data from 39 transit systems in Indiana, United States. In particular, the effect of transit subsidies on transit performance depends on the funding source (federal, state or local) as well as the system size (large, medium, or small). Therefore they suggest that the effect subsidies on transit performance should not be generalized. Another empirical study on fare subsidies is Parry and Small (2009), where a general framework for evaluating fare subsidies and potential pricing changes based on data from three metropolitan areas: Washington, DC, Los Angeles, and London. They find increasing subsidies beyond 50% of the operating costs can improve social welfare across modes (rail and bus), periods (peak and off-peak) and cities (all three cities mentioned above). This finding is robust to alternative assumptions about model parameters and behaviors of agencies. In short, the empirical analyses by Parry and Small (2009) demonstrate that the substantial operating subsidies for transit systems are justified based on economic efficiency. In addition, they also note that pricing automobile externalities properly could lead to significantly larger welfare improvements than transferring subsidies to public transit services. The empirical findings from Parry and Small (2009) are consistent with the findings from Nelson et al. (2007), namely subsidies to the transit systems in the Washington, DC region are well justified, because the combined benefits of rail and bus transit far exceed the transit subsidies.

It should be noted that the findings from the above reviewed empirical studies may not necessarily be generalized to other regions due to differences in regional characteristics. For instance, both Parry and Small (2009) and Nelson et al. (2007) justify the transit subsidies in Washington, DC, while Tscharaktschiew and Hirte (2012) conclude that the optimal subsidy should be small or zero in a German metropolitan area. Therefore the developed analysis frameworks should be adapted to evaluate the effectiveness and efficiency of transit subsidies based on local data.

Conclusions

Pricing and subsidization policies are central to the financial sustainability of public transit services. This article briefly discusses how public transit fare and subsidy should be optimized, covering both theoretical and empirical studies. Due to regional differences,

public transit fare and subsidy policies vary significantly. In the theoretical literature about fare optimization, a microeconomic analysis framework is usually adopted following the seminal work (Mohring, 1972). However, to facilitate the analytical solution, simplifying assumptions regarding demand pattern, price elasticity, user heterogeneity, etc. are usually made, which might not necessarily hold in the real world. In the empirical studies, in addition to various economic factors, social factors such as equity are usually incorporated in making fare policies. Transit subsidies have been controversial as conflicting conclusions usually appear in the literature, although a few rationales for subsidy have been clearly identified. There is some empirical evidence supporting the substantial transit subsidies; however, such a finding is usually difficult to generalize to other regions.

The following suggestions are made to improve the transit fare and subsidy studies:

1. Some unrealistic assumptions (for example, transit riders travel from multiple origins while to the same destination, that is, a many-to-one travel demand pattern is assumed) in theoretical studies should be relaxed for more realism.
2. Regulatory policies should be jointly considered when subsidy decisions are made, because subsidies in the absence of regulations may be counterproductive, for example, leading to wasteful overproduction.
3. Game-theoretical approaches should be adopted to fully capture the decision interactions between multiple stakeholders (regulator, subsidizer, transit operator, etc.) on the same or different levels in transit policy decision making, as the centralized decision-making process cannot fully capture the decision-making complexities.
4. Regional characteristics should be accounted for before a fare or subsidy policy is transferred across regions, as one policy that works in one region may not work in another region due to regional differences.

References

- Cervero, R., 1981. Flat versus differentiated transit pricing: what's a fair fare? *Transportation* 10 (3 p), p.211–232.
- Cervero, R., 1986. Time-of-day transit pricing: comparative US and international experiences. *Trans. Rev.* 6 (4), 347–364.
- Chang, S.K., Schonfeld, P.M., 1991. Optimization models for comparing conventional and subscription bus feeder services. *Transport. Sci.* 25 (4 P), 281–298.
- Elgar, I., Kennedy, C., 2005. Review of optimal transit subsidies: comparison between models. *J. Urban Plan. Dev.* 131 (2), 71–78.
- Else, P.K., 1985. Optimal pricing and subsidies for scheduled transport services. *J. Trans. Econ. Policy* 19 (3), 263–279.
- Farber, S., Bartholomew, K., Li, X., Pérez, A., Habib, K.M.N., 2014. Assessing social equity in distance based transit fares using a model of travel behavior. *Transport. Res. A Policy Pract.* 67, 291–303.
- Gómez-Lobo, A., 2014. Monopoly, subsidies and the Mohring effect: a synthesis. *Trans. Rev.* 34 (3), 297–315.
- Gray, A., 2018. Estonia is Making Public Transport Free. Available from: <https://www.weforum.org/agenda/2018/06/estonia-is-making-public-transport-free/>.
- Jara-Díaz, S., Gschwender, A., 2003. Towards a general microeconomic model for the operation of public transport. *Trans. Rev.* 23 (4), 453–469.
- Karlafitis, M.G., McCarthy, P., 1998. Operating subsidies and performance in public transit: an empirical study. *Transport. Res. A Policy Pract.* 32 (5), 359–375.
- Leong, L., 2016. The 'Rail plus Property' Model: Hong Kong's Successful Self-Financing Formula. Available from: <https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/the-rail-plus-property-model>.
- Ling, J.H., 1998. Transit fare differentials: a theoretical analysis. *J. Adv. Trans.* 32 (3), 297–314.
- Mohring, H., 1972. Optimization and scale economies in urban bus transportation. *Am. Econ. Rev.* 62 (4), 591–604.
- MTA, 2018. MTA 2018 Adopted Budget February Financial Plan 2018–2021. (Metropolitan Transportation Authority. Available from: http://web.mta.info/mta/budget/pdf/MTA-2018-AdoptedBudgetFebruaryFinancialPlan_2018-21.pdf accessed on May 15, 2019.
- Nelson, P., Baglino, A., Harrington, W., Safirova, E., Lipman, A., 2007. Transit in Washington, DC: current benefits and optimal level of provision. *J Urban Econ.* 62 (2), 231–251.
- Newell, G.F., 1979. Some issues relating to the optimal design of bus routes. *Transport. Sci.* 13 (1), 20–35.
- Parry, I.W., Small, K.A., 2009. Should urban transit subsidies be reduced? *Am. Econ. Rev.* 99 (3), 700–724.
- Proost, S., Van Dender, K., 2008. Optimal urban transport pricing in the presence of congestion, economies of density and costly public funds. *Transport. Res. A Policy Pract.* 42 (9), 1220–1230.
- Savage, I., 2010. The dynamics of fare and frequency choice in urban transit. *Transport. Res. A Policy Pract.* 44 (10), 815–829.
- Sun, Y., Guo, Q., Schonfeld, P., Li, Z., 2016. Implications of the cost of public funds in public transit subsidization and regulation. *Transport. Res. A Policy Pract* 91, 236–250.
- Sun, Y., Schonfeld, P.M., 2015. Optimization models for public transit operations under subsidization and regulation. *Transport. Res. Rec.* 2530 (1), 44–54.
- Sun, Y., Zhang, L., 2018. Microeconomic model for designing public transit incentive programs. *Transport. Res. Rec.* 2672 (4), 77–89.
- Tscharaktschiew, S., Hirte, G., 2012. Should subsidies to urban passenger transport be increased? A spatial CGE analysis for a German metropolitan area. *Transport. Res. A Policy Pract.* 46 (2), 285–309.
- Wirasinghe, S.C., Ghoneim, N.S., 1981. Spacing of bus-stops for many to many travel demand. *Transport. Sci.* 15 (3), 210–221.

Transportation Equity

Rafael H. M. Pereira*, Alex Karner†, *Institute for Applied Economic Research–Ipea, Brazil; †The University of Texas at Austin, Austin, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Definitions	271
Equity as Part of a Broader Idea of Justice	271
Why is Transportation Equity Important?	272
Theorizing Transportation Equity	272
Distribution of What: Benefits and Burdens	272
Guiding Moral Principles	274
What is a Fair Distribution?	274
For Whom?	275
The Future of Transportation Equity	276
References	277

Definitions

Transport and urban policies have implications for multiple dimensions of social life; adequate transportation services and/or well-located housing can mean the difference between getting and keeping a job, accessing healthy food, getting to school on time, or reaching needed medical care. Decisions to build or not build specific types of transportation infrastructure in specific places can result in changes to community cohesion, exposure to air pollution, and road safety risks. The moral concern with transportation equity encompasses the multiple channels through which transport and land use policies can create conditions that allow different people to flourish, to satisfy their basic needs and lead a meaningful life.

In cities around the world, transport and mobility policies create equity issues on a daily basis. Consider the following examples:

- Bus rapid transit systems that reallocate road space from private vehicles to public transport in cities like Bogota, Dar es Salaam, and Beijing disadvantage drivers while benefiting public transit users.
- Parking and congestion pricing in London and Singapore disproportionately burden low-income drivers. But mitigating policies, like reinvesting revenue in public transit, can potentially offset any apparent disadvantage.
- An \$80-million (USD) bridge built in Sao Paulo dedicated exclusively for private motor vehicles, where public transport, cyclists, and pedestrians are prohibited exclusively benefits drivers while siphoning funds away from other modes.
- The precarious state of the pedestrian and cycling infrastructure in Nairobi, Cairo, and countless other cities across the Global South routinely creates life-threatening situations for nondrivers.
- Automobile dependency in cities like Atlanta or Houston forces low-income people and others who would prefer not to drive into automobile ownership so that they can access affordable housing as well as health and education services.
- Low-income communities and communities of color are disproportionately exposed to externalities like traffic fatalities and pollution because of their residential settlement patterns, segregation, gentrification, and displacement.

Each of these examples highlights a common aspect of transport decision making: planners and policy makers have to prioritize the allocation of scarce resources, meaning that plans and policies inevitably benefit some groups at the expense of others. Evaluating the nature and extent to which benefits and burdens differ across groups is the central core of distributive justice.

Transportation equity is a way to frame distributive justice concerns in relation to how social, economic, and government institutions shape the distribution of transportation benefits and burdens in society. It focuses on the evaluative standards used to judge the outcomes of policies and plans, asking who benefits from and is burdened by them and to what extent.

Equity as Part of a Broader Idea of Justice

Transportation equity is a crucial part of a broader concern with transport and mobility justice. Transport justice encompasses moral and political concerns related to equity, democracy, and diversity in the pursuit of more just cities and mobility systems. The concern with *equity* (distributive justice) relates to how the institutions and rules that govern society shape social and economic inequalities among its members. It focuses on the evaluative standards used to judge the outcomes of policies, asking who benefits from and is burdened by them and to what extent. Meanwhile, the concern with *democracy* relates to the fairness of the political processes related to participation in decision making. It involves the challenge of moving beyond periodic voting and overcoming technocratic top-down planning practices, and it is based on the core ideal that everyone's opinion should be equally heard in the decision-making processes that form institutions, policies, and community organizations that shape the built environment (Fainstein, 2010). Finally,

the concern with *diversity* involves the constant struggles over which rights and entitlements should be recognized and enforced. It requires the recognition of group-based differences and identities and promoting diversity in decision-making processes, acknowledging that participatory democracy is constantly marked by structural inequalities of wealth and power that marginalize certain groups and favor others in their ability to influence policy decisions (Young, 1990). Related to the concerns with democracy and diversity is the idea of the “right to the city.” This idea is undergirded by the structural uneven geographies of power and privilege in cities and the political struggles over who has the power to influence and shape urbanized space (Enright, 2019). It is broadly understood as a claim for people to reshape the processes of urbanization through collective action and self-management to complement or challenge existing government policies, or to fill a gap where there is a lack of governmental action.

Why is Transportation Equity Important?

Questions of transportation equity provide a valuable moral compass that help assess the fairness of the distribution of the benefits and burdens of transport policies (van Wee, 2011). Equity, nonetheless, is not so much about the unequal allocation of resources (transport services, investments, infrastructure) across space per se. Rather it is about how policy decisions shape societal levels of environmental externalities and what groups are more or less exposed to them, just as how they affect the lives of different groups in terms of their ability to access life-enhancing opportunities such as employment, healthcare, and education (Pereira et al., 2017).

A comprehensive understanding of transport and mobility justice requires a multidimensional effort to grasp the distributive implications of urban and transportation policies while critically examining the political processes and institutional context that determine those distributive outcomes. A research focus on transportation equity, nonetheless, is valuable in its own right. Policy choices often reflect societal biases and can widen and deepen social and spatial inequalities, and the study of social and spatial inequalities can help make visible the structural biases underlying the organization of societies and governments. Even when governments have general redistribution systems, for example through taxation policies and subsidies, implementing regressive transport policies that disproportionately benefit the well-off is not justified. Because transportation needs vary widely depending on one’s personal constraints (including mode availability, residential location, and workplace), transportation equity requires spatially targeted policies and interventions, as opposed to those that are more general and society-wide. Furthermore, public policies have a crucial role to play in social democracies through the provision of public goods and services, including public transit and large infrastructure projects. In this sense, public transport services and investments remain one of the key drivers that can shape spatial inequalities of development and opportunities in cities, making the study of the distributional effects of transport policies particularly important. Finally, even a genuinely diverse, democratic, and inclusive participatory planning process is not a sufficient condition to ensure that the outcomes of transport policies are distributed equitably. The call for transport justice goes beyond distributive concerns, and yet justice cannot be achieved without equity.

Theorizing Transportation Equity

Researchers reflecting on transportation equity are often faced with four interrelated questions about distributive justice that cannot be addressed in isolation from one another:

1. What—that is, which benefits and burdens—should be distributed?
2. On which moral principles should distributions be based?
3. What is the fairest distribution?
4. Distribution among whom?¹

Different theories of justice in the political philosophy literature (including utilitarianism, Rawls’s egalitarianism, capability approaches, and critical theories of justice) can lead to different answers to these questions. These theories are receiving increasing attention in the academic transport and mobilities literature (Davoudi and Brooks, 2014; Martens, 2016; Pereira et al., 2017). There are some general agreements in the literature about how these questions should be addressed in the context of transport policies.

Distribution of What: Benefits and Burdens

The transport literature usually focuses on four types of transport-related benefits and burdens that are related to people’s well-being: transport resources, mobility as in observed travel behavior, accessibility levels, and transport externalities (Lucas et al., 2019). Analyses of the distribution of transport resources and travel behavior can be problematic. The focus on the distribution of resources, such as car ownership or proximity to transport infrastructure, for example, can be misleading because it does not account for people’s needs and preferences, and it does not reflect people’s capability to use such resources to move in space and reach desired destinations. In turn, analyses of differences in travel behavior, such as in trip frequency or commute times, often cannot untangle how much of those inequalities emerge from individuals’ tastes and preferences (voluntary choice) or from contextual constraints outside individual control. While travel patterns are closely linked to individuals’ levels of well-being and participation in society,

¹ Questions related to the decision-making process and who has a voice in it are important but related to diversity and democracy concerns in transportation justice; they are beyond the scope of this chapter on transportation equity.

travel behavior data alone do not provide information about whether long commute times reflect constrained housing options in poor and distant locations or whether they reflect preferences for suburban living or other “expensive tastes”. Transport decisions are not always, if rarely, a matter of individual rational choice as it is often assumed in the literature.

Transportation equity studies have found more fertile ground when looking at accessibility and transport externalities. The primary benefit of transportation infrastructure and services is to improve people’s accessibility, or the ease with which they can reach key destinations including employment, healthcare, and educational opportunities. Multiple authors have argued that the concept of accessibility should be at the center of understanding transportation benefits and equity for various reasons (Martens, 2012; Pereira et al., 2017). Some minimum level of accessibility to key destinations is necessary for people to satisfy their basic needs. It also has instrumental importance to support out-of-home activity participation and the freedom of choice that allows people to develop their capabilities and flourish. Moreover, understanding the causes of disparities in access to opportunities reveals the spatial dimension in moral concerns over inequality of opportunities, which is a central theme across theories of justice but which has been largely addressed as a nonspatial problem by political philosophers and social scientists.

Accessibility is also attractive as a planning goal because it brings together transportation and land use (Manaugh et al., 2015). Because people travel to meet opportunities that are widely dispersed in space, bringing origins, and destinations closer together affects the demand for travel. These types of land use policies also support urban environments conducive to reducing automobile use and increasing walking and cycling. Such policies can also create situations where land prices and rents increase, displacing long-time residents to suburban and exurban areas. These types of secondary and tertiary effects of transportation and land use policies must be considered when attempting to meet equity goals.

One challenge with accessibility as a concept is that there is no consensus on how it should be measured (Lucas et al., 2019; Neutens et al., 2010). Most commonly in the literature accessibility is calculated by summing up the number of activities (e.g., jobs or schools) reachable from an origin location using either a defined travel time threshold (e.g., 45 min of public transit travel time) or by weighting closer opportunities more heavily. These types of measures are attractive because they can be easily calculated using readily available data. Nevertheless, a broad sense of accessibility also involves issues such as affordability, security, safety, universal design, and social practices. A more rigorous account of transportation equity should consider how accessibility levels are influenced by personal characteristics such as age, race, gender, or physical and cognitive disabilities and how such characteristics relate to local contexts of social and built environments.

Ideas about transportation system benefits have certainly shifted throughout the latter half of the 20th century in academia, but these changes are slow to diffuse to practice. The situation is most challenging in the automobile-dominated economies of the United States, Canada, Australia, and many nations in the Global South. Automobile-centric planning values travel time and specifically congestion mitigation over accessibility. The focus on reducing travel times and congestion often leads to inequitable and ineffective policies (Martens, 2016; van Wee, 2011). This is in part because the monetary valuation of travel time savings that are commonly conducted in conventional project appraisal methods (such as cost–benefit analysis) implicitly favors transport investments that primarily benefit higher-income groups. Moreover, attempts to tackle congestion and reduce travel times by adding roadway capacity ultimately leads to more congestion through a process known as “induced demand.” With reduced travel times, people make more trips or shift trips from other routes to locations that have new capacity. In the medium term, this leads to land use changes. This realization suggests that congestion is not solvable in the aggregate, but providing alternatives to driving can improve the situation for individuals. An accessibility perspective makes this clear.

Transportation equity concerns also relate to the distribution of transportation burdens, who causes those burdens and who gets exposed to them (Feitelson, 2002; Schweitzer and Valenzuela, 2004). Transportation burdens include various health-related concerns stemming from exposure to air pollution, sedentary lifestyles, and increased risk of injury and death from collisions. Most of these burdens have historically been inequitably distributed, with low-income people and people of color overrepresented in terms of their population shares of proximity to heavily trafficked roads, respiratory illnesses, and rates of injury and death. Disadvantaged populations also tend to reside in areas with lower-quality pedestrian and cycling infrastructure and where traffic conditions are more dangerous, subjecting them to increased risk of death or severe injury when collisions occur. On the other hand, studies have shown that high-income populations are disproportionately responsible for flights and private vehicle use, which are responsible for overwhelming shares of transport pollution (Mattioli, 2016). Efforts to mitigate these disparities have been uneven. In Global North countries, air quality has been steadily improving over time even as levels of driving continue to increase. In the Global South, on the other hand, rising automobility, the use of older vehicles for longer periods of time, and unfavorable geography, have all contributed to dramatically worse air quality in places like Beijing, New Delhi, and Santiago.

Efforts to mitigate these health concerns can sometimes have unintended consequences. For example, a popular approach to increasing traffic safety includes adopting “Vision Zero” policies that have as their goal reducing or eliminating traffic deaths. A popular method for achieving Vision Zero in cities around the world involves increasing police presence in locations where traffic incidents are concentrated. But these also tend to be locations disproportionately populated by disadvantaged populations. Cyclists in particular are vulnerable when police presence increases because they are often violating laws (either knowingly or unknowingly) as a matter of necessity or because of financial constraints. Riding on the sidewalk to avoid heavily-trafficked streets is one example. Riding without lights or a helmet is another. While lights are often required while riding at night, they are somewhat costly to maintain because they require batteries and are easy targets for theft. These minor legal violations give police an opportunity to stop, question, and potentially search cyclists. Depending on demographics, the presence of a cycling culture, and infrastructure, the types of people bicycling will differ substantially. Attitudes toward cycling will also tend to differ. In many locations in the United States,

for example, cycling is a deeply marginalized choice and this is reflected in the demographics of those who choose to ride. Increased policing and enforcement can produce a very dangerous situation for these cyclists.

Guiding Moral Principles

Regardless of the specific benefits and burdens that are considered as part of an inquiry into transportation equity, applying different moral principles will result in different understandings about how such benefits and burdens should be distributed. Two broad types of moral concerns that are commonly applied in the transportation equity literature but rarely explicitly articulated come from sufficientarian and egalitarian perspectives.

A sufficientarian approach to justice considers whether the distribution of goods and bads in society allow everyone to meet a generally acceptable standard of living. Stated differently, absolute levels matter more than relative inequalities for sufficientarianism. Adherents would not be concerned with differences across groups, as long as all achieve above a certain threshold of, say, transport accessibility or air quality (Martens, 2016).

Sufficientarian concerns are reflected in various ideas that permeate the transportation literature and practice. Much attention has been given to issues related to environmental justice and public health, looking for example at whether certain ethnic minority neighborhoods and other vulnerable groups are exposed to exceedingly high levels of transport air pollution. A growing body of research also focuses on the ideas of “needs gaps” and “transit deserts.” These studies seek to quantify a “gap” between the level of public transport service (supply) and the needs of a particular population (demand), and identify as “deserts” those locations where demand outpaces supply levels. Most of these researches are not explicitly grounded in theoretical frameworks on justice or basic needs, even though they take an implicitly sufficientarian perspective without naming it as such. These studies depart from the idea that there exists a minimum level of transport resources, services, safety, or accessibility that is necessary for people to fulfill basic needs and lead a meaningful life, or conversely that there is a maximum amount of transport externalities such as air pollution that is deemed acceptable. A corollary to this idea of a minimum standard is that public action is necessary to guarantee the needs of people who fall below that threshold.

In contrast to sufficientarians, egalitarians are fundamentally concerned with the relative distribution of goods and bads among members of society. From an egalitarian perspective, for example, even if all individuals in a city had above acceptable levels of access to job opportunities by public transport, for example, this situation would still be deemed unfair if a population group had relatively lower access as a result from a discriminatory policy against them. This is a strong reason why transport justice problems cannot be defined solely based on sufficientarian concerns. Egalitarianism encompasses the idea that all individuals have equal moral worth and should be treated as equals in some respect, even though different theories of justice might disagree about what kinds of equality (resources, life outcomes, well-being, opportunities, etc.) are more morally relevant.

Transport inequalities are a common egalitarian concern found across the transport equity literature, even though the majority of this research is not explicitly grounded on any theory of justice. Typically, studies in this tradition investigate, for example, the extent to which certain groups or neighborhoods are disproportionately more exposed to transport externalities or why certain groups are better served by public transport and have higher transport accessibility levels than others (e.g., Bullard and Johnson, 1997). The implication is that these inequalities could reflect deeper structural factors such as historic discriminatory policies and other disadvantages that undermine the transport experience of certain groups and further exacerbate injustices.

A fair transport policy is guided by both sufficientarian and egalitarian concerns. It aims to improve, for example, overall levels of transport safety or accessibility and reduce transport externalities so that everyone is above acceptable thresholds, while at the same time prioritize improving the conditions of disadvantaged groups and reduce inequalities. However, promoting minimum levels of transport goods/bads and reducing inequalities, is not sufficient to guarantee a just urban and transport policy. A necessary condition for a policy to be considered fair, is that it cannot not override the basic rights and liberties of individuals and minority groups even if it promotes a greater good to a greater number of people (Pereira et al., 2017). Particularly in Global South, it is still more often the norm than the exception that governmental officials use this type of utilitarian discourse to justify the eviction of poor communities in order to expand road and transport infrastructure.

In summary, a fundamental characteristic of an equitable transport policy or project is that its implementation respects people’s basic rights and liberties, such as the physical and psychological liberty and integrity of the person. Furthermore, it should both prioritize improving the conditions of disadvantaged groups and reduce transport-related inequalities, and at the same time account for people’s basic needs ensuring that all individuals achieve above a certain threshold of essential transport goods/bads, whatever those thresholds may be given a set of environmental constraints. On the whole, this sets a theoretical framework that is flexible enough to accommodate universalist concerns about the protection of basic rights, the satisfaction of basic needs, and promotion of equality of opportunities without losing sight of context-specific issues regarding the particularities of each urban context and that the notions of basic needs and minimum thresholds are historically and culturally dependent.

What is a Fair Distribution?

These moral concerns of egalitarianism and sufficientarianism suggest whether relative or absolute levels of goods and bads should be considered, but these principles alone do not determine whether a specific distributional outcome is judged to be equitable. Studies that follow a sufficientarian perspective often adopt a technocratic approach and assume implicitly the minimum levels that require policy intervention can be defined through a “technical” decision based on empirical data analyses alone, without any moral

or political judgment. Nonetheless, the definition of what an “adequate” level of transport accessibility or externality means is ultimately a political decision that deeply reflects the vision of a just city and mobility system each society aspires to build and which is highly dependent on local and historical context. For policy purposes, setting those thresholds ultimately requires a legitimate political and democratic process.

Studies that take an egalitarian stand, on the other hand, generally do not specify why the transport inequalities they analyze should be deemed unfair and implicitly assume the mere existence of any inequality is put forth as evidence of unfairness. One standard that could be used when defining whether a given transport inequality should be deemed inequitable is whether an observed disparity results from the unfair treatment of disadvantaged groups and whether such inequality could undermine/compromise their life chances. It is not the level of inequality per se that determines whether a given distribution is inequitable, although oftentimes this magnitude can be very telling of deeper injustices.

Moreover, studies that focus on transport inequalities, as a rule, do not clearly state how far policies should go to mitigate them. At the outset, it is important to note that a situation of perfect equality (whether of transport resources, services, accessibility, etc.) is not possible because of how societies are (spatially) organized, even if some might consider full equality to be desirable. Nonetheless, there is a growing consensus in both transport planning practice and in the academic literature that a fair transportation policy should prioritize improving the conditions of people in the worst-off positions, helping to pull them above acceptable minimum thresholds and reducing inequalities (Pereira et al., 2017). Despite various differences of ideas between and within egalitarian and sufficientarian views of justice, it is commonly accepted that equality does not necessarily equate with equity and that in some cases the pursuit of justice requires that individuals or groups be treated differently to compensate for unfair conditions and inequalities. This is one of the reasons why it becomes increasingly important to move beyond cross-sectional descriptive analysis of transport poverty and transport inequalities, and look more closely at how and to what extent the implementation of various policy interventions can contribute to promote fairer and more inclusive cities and mobility systems.

For Whom?

Each of the prior considerations can be set without consideration of specific population groups for whom outcomes should be assessed and compared. The literature and practice have very clearly focused on population groups known to be disadvantaged because of low income, ability, and historical discrimination (Lucas, 2019). For example, relevant groups include youth, older adults, women, low-income people, zero-vehicle households, people with disabilities, single-parent households, refugees, and racial/ethnic minorities. There is a common understanding that these are morally arbitrary characteristics, beyond individuals’ control, but which often undermine people’s transport experience because of historic discriminatory policies and other disadvantages. This emphasis in the literature also arises in part from the sense that these groups have been historically excluded from the transportation planning process as well as legal frameworks that require transportation planners to explicitly consider their needs (Grengs, 2002; Karner and Niemeier, 2013).

Transportation disadvantages can manifest in various ways (Lucas and Jones, 2012; Mullen et al., 2014). They occur when there are not enough transport resources/services available to an individual or household for them to participate in desired activities. They can also occur when there are enough resources/services but these are inadequate for people with disabilities. Transport disadvantages can also occur when vulnerable groups are disproportionately exposed to environmental externalities, or when the transport experience, physical and mental health of a person or group are undermined, limited or put at risk because they are from an ethnic minority or gender in a given social and built environment. While a few studies use the term “modal equity” to refer to inequalities between transport modes, equity is about people. Disparities between transport modes (say of road space, or accessibility, etc.) only become equity issues because of the systematic differences in the socio-demographic characteristics of those who predominantly reap the benefits and bear the externalities of different modes.

Although there is no absolute standard for when a particular situation rises to the level of disadvantage, researchers have identified that *social exclusion* can result if transportation needs are not met. This situation can result in trips being foregone with attendant poor outcomes including difficulty finding and keeping a job, inability to reach medical care, difficulty accessing educational opportunities, and limited social connections. Historically disadvantaged groups in a transportation equity context are disadvantaged because of their risk of social exclusion (Lucas, 2019).

A transportation equity perspective that focuses attention on the needs of these groups is important because transportation planning has historically responded to the needs of the most mobile and has sought to further improve their mobility. This emphasis is most clear in the automobile-oriented transportation planning that has commonly occurred in Canada, Australia, and the United States and across countries in the Global South.

Legal frameworks in the United States that respond to that country’s unique and problematic history of racism, unequal treatment, and discrimination against people of color require public agencies to ensure that their policies and practices are not discriminatory (Bullard et al., 2004), although the methods agencies use to demonstrate nondiscrimination can often be inconsistent with best practices developed in the academic literature (Karner and Niemeier, 2013). The Americans with Disabilities Act (ADA) also requires agencies to provide equal accommodations for people with disabilities, but this often manifests in paratransit services that are viewed as a lower tier than fixed-route options. In the early 1990s, the United Kingdom made strides in efforts to combat social exclusion, forming an official government arm to implement policy across a number of issue domains, including transport. Difficulties with developing appropriate transport-specific metrics and tying them to real-world planning efforts meant that that effort did not result in material gains on social exclusion.

The Future of Transportation Equity

The pace of technological change in transportation is creating profound disruption and dislocation across the industry that challenges existing frameworks related to transportation equity. Questions regarding the appropriate role for private organizations in providing mobility, the effects of connected and automated vehicles on public transit provision, the desirability of “micro” transit and on-demand transit services, and the role of bike share and other “micromobility” solutions on travel outcomes all warrant increasing attention.

In the coming decades, the transport industry will increasingly rely on algorithms to support the planning, management and allocation of automated vehicles, and mobility services, for example with the use of artificial intelligence and big data predictive modeling to develop flexible routing and scheduling of services. There is a growing awareness of the ways in which big data and algorithms often reflect the biases underlying broader social and political processes. The lack of critical reflection and algorithmic transparency raises the risk of decision-making processes that inadvertently reflect those biases and end up reinforcing injustices. While much of the current research on automated vehicles and shared mobility systems focuses on how to make them commercially viable, more research is needed to understand how these changes in technology and governance will reshape who gets access, when, and how, and what implications they might have for questions of equity, public health, and environmental justice.

In the meantime, the rise of transportation network companies (TNCs)² is already leading directly to declines in public transit ridership in cities around the world. TNCs operate best when there is a high density of origins, destinations, and potential riders, so that wait times are minimized. Locations where these conditions prevail are also those where public transit is most likely to succeed. The practical result has been diminishing public transit ridership and increasing congestion in cities where TNCs have been active the longest. Some observers have hailed this development as another example of successful “disruption” of an outdated public transit business model. What these observers fail to realize is that public transit provides essential lifeline mobility to those who need it most in locations that would otherwise be unprofitable to serve. Private mobility providers simply cannot be relied upon to serve the diversity of markets that public transit serves. Indeed, public transit operators intentionally provide service in locations that they know will not generate a profit so that residents have some access to transportation services. They also operate large vehicles to accommodate peak travel demand, minimize maintenance challenges, and streamline driver licensing and certification.

The influx of private mobility providers into markets previously served by public transit has led to a push for increasing “microtransit” or on-demand service. The problem with these calls is that microtransit does not scale well. Where public transit is most effective, operators save time by not deviating from a fixed route to reach every user at their origin and destination. Instead, they collect riders at common origins and destinations, with riders required to complete the first and last mile. A microtransit system that provides every customer with door-to-door service would create substantial unreliability and unpredictability at relatively modest levels of demand. It may provide a viable option in certain types of first/last-mile situations and in areas where public transit demand is relatively low.

A perennial issue with TNCs as a transportation equity measure involves the problem of un/underbanked populations and smartphone access. Access to credit is often a prerequisite for using the smartphone apps that enable TNC access and many lower income people simply do not have access. While smartphone penetration rates continue to increase across all population groups, limitations on data plans are more likely to affect low-income people and thus their ability to access private mobility providers. Spatial discrimination also continues to be a problem, with evidence of TNC drivers avoiding certain neighborhoods out of real or perceived fear.

One response to the challenges of public transit in general is to provide automobiles to those experiencing transportation disadvantages in an effort to mitigate their social exclusion by encouraging automobility. Multiple researchers have examined the effect of such a policy on outcomes and the results are promising. Providing low-income people with cars increases the likelihood of finding and keeping a job and can lead to long-term increases in earnings. The problem, unfortunately, is that this solution also does not scale and come at an environmental cost (Mattioli, 2016). Giving all low-income people cars will saddle already vulnerable households with additional debt and monthly costs that may or may not be offset by increased revenue, increase risks of injury and death, and accelerate the climate crisis. Additionally, the cities that have grown up around automobility run contrary to the basic principles of city building that have accumulated over decades, increasing car dependency and thus limiting mobility and accessibility choices for nondrivers.

Indeed, urban planners are concerned with understanding the types of places people want and how to create them. Encouraging more driving cannot be the solution to transportation equity issues (Mattioli, 2016). More promising alternatives that address equity without exacerbating the climate crisis and leading to unintended health consequences are possible but they involve looking beyond transport to consider land use, housing, and institutional barriers in an integrated manner. Despite the knowledge that travel demand is intimately related to land use, transport practitioners still rarely consider housing and land use solutions to transport problems. Creating places where people do not have to use the car to access and hold a job, where they can make healthy transport decisions, and where their share of household income allocated to transport and housing is held to a reasonable level is a more promising—and proven—strategy for achieving transportation equity that does not rely on the benevolence of private entities nor does it require us to rely on the game-changing promise of new technology. To achieve transportation equity, we need to ensure that all residents of our cities and regions can safely reach the opportunities that they need to thrive.

² Various referred to as ridehailing, ridesourcing, and (incorrectly) ridesharing organizations, these are companies that provide users with the ability to summon a vehicle to serve a trip on demand.

References

- Bullard, R.D., Johnson, G.S. (Eds.), 1997. *Just Transportation: Dismantling Race & Class Barriers to Mobility*, New Society Publishers, Gabriola Island, BC.
- Bullard, R.D., Johnson, G.S., Torres, A.O. (Eds.), 2004. *Highway Robbery: Transportation Racism & New Routes to Equity*, South End Press, Cambridge, MA.
- Davoudi, S., Brooks, E., 2014. When does unequal become unfair? Judging claims of environmental injustice. *Environ. Plann. A* 46 (11), 2686–2702.
- Enright, T., 2019. Transit justice as spatial justice: learning from activists. *Mobilities* 14 (5), 665–680.
- Fainstein, S.S., 2010. *The Just City*. Cornell University Press, Ithaca.
- Feitelson, E., 2002. Introducing environmental equity dimensions into the sustainable transport discourse: issues and pitfalls. *Transport. Res. D: Transport Environ.* 7 (2), 99–118, doi:10.1016/S1361-9209(01)00013-X.
- Grengs, J., 2002. Community-based planning as a source of political change: the transit equity movement of Los Angeles' bus riders union. *J. Am. Plann. Assoc.* 68 (2), 165–178.
- Karner, A., Niemeier, D., 2013. Civil rights guidance and equity analysis methods for regional transportation plans: a critical review of literature and practice. *J. Transport Geogr.* 33, 126–134.
- Lucas, K., 2019. A new evolution for transport-related social exclusion research? *J. Transport Geogr.* 81, 102529, doi:10.1016/j.jtrangeo.2019.102529.
- Lucas, K., Jones, P., 2012. Social impacts and equity issues in transport: an introduction. *J. Transport Geogr.* 21, 1–3, doi:10.1016/j.jtrangeo.2012.01.032.
- 1st edition. Lucas, K., Martens, K., Ciommo, F.D., Dupont-Kieffer, A. (Eds.), 2019. *Measuring Transport Equity*, 1st edition, Elsevier, Cambridge, MA.
- Managh, K., Badami, M.G., El-Geneidy, A.M., 2015. Integrating social equity into urban transportation planning: a critical evaluation of equity objectives and measures in transportation plans in North America. *Transport Policy* 37, 167–176, doi:10.1016/j.tranpol.2014.09.013.
- Martens, K., 2012. Justice in transport as justice in accessibility: applying Walzer's 'spheres of justice' to the transport sector. *Transportation* 39 (6), 1035–1053.
- Martens, K., 2016. *Transport Justice: Designing Fair Transportation Systems*. Routledge, London.
- Mattioli, G., 2016. Transport needs in a climate-constrained world. A novel framework to reconcile social and environmental sustainability in transport. *Energy Res. Soc. Sci.* 18, 118–128, doi:10.1016/j.erss.2016.03.025.
- Mullen, C., Tight, M., Whiteing, A., Jopson, A., 2014. Knowing their place on the roads: What would equality mean for walking and cycling? *Transport. Res. A: Policy Pract.* 61, 238–248.
- Neutens, T., Schwanen, T., Witlox, F., Maeyer, P.D., 2010. Equity of urban service delivery: a comparison of different accessibility measures. *Environ. Plann. A* 42 (7), 1613–1635, doi:10.1068/a4230.
- Pereira, R.H.M., Schwanen, T., Banister, D., 2017. Distributive justice and equity in transportation. *Transport Rev.* 37 (2), 170–191, doi:10.1080/01441647.2016.1257660.
- Schweitzer, L., Valenzuela, A., 2004. Environmental injustice and transportation: the claims and the evidence. *J. Plann. Liter.* 18 (4), 383–398, doi:10.1177/0885412204262958.
- van Wee, B., 2011. *Transport and Ethics: Ethics and the Evaluation of Transport Policies and Projects*. Edward Elgar Pub, Cheltenham.
- Young, I.M., 1990. *Justice and the Politics of Difference*. Princeton, Princeton University Press.

Impact of Transport Cost-Benefit Analysis on Public Decision-Making

Niek Mouter, Delft University of Technology, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Cost-Benefit Analysis	278
Formal Role of CBA	278
Impact of Transport Cost-Benefit Analysis on Public Decision Making	278
Opportunistic and Symbolic Use	279
Barriers Hampering Politicians' Use of CBA When Forming Their Judgments	279
What Explains the Positive Attitude of Politicians Toward CBA?	280
Biography	281
References	281

Cost-Benefit Analysis

Cost-Benefit Analysis (CBA) is a widely used economic appraisal method, which aims to support policy makers in making decisions about projects and policies (Boardman et al., 2013). In virtually all western countries CBA is mandatory when national funding is asked for large transport projects (Mackie et al., 2014). The theoretical foundations for CBA are provided by welfare economics, which is a branch of economics that investigates the social desirability of alternative economic situations (Boadway, 2006). The CBA is built on the Kaldor-Hicks efficiency criterion, which recommends projects where the sum of monetary gains outweigh the sum of monetary losses and winners can potentially compensate the losers (The Pareto and Kaldor-Hicks criteria; Minken).

In a CBA, positive and negative impacts of government projects are valued by estimating private willingness to pay (WTP). Transport projects are typically intertemporal in nature, so the benefits and costs occur over a number of periods. Because people tend to prefer a present dollar over a future dollar—even after a correction for inflation—future impacts of the project are discounted and presented as so-called (net) present values (Mouter, 2018) (The social discount factor; Hultkrantz). Although the core principles of CBA are universal across countries research shows that there is a considerable spread in the price tags in the appraisal guidelines of OECD countries (e.g., the value of time (VOT), the value of a statistical life (VSL), and the valuation of CO₂) (Mackie et al., 2014).

Formal Role of CBA

The formal role of CBA differs between countries. In practices, such as Chile, the United Kingdom, and the Netherlands, CBA provides information for decision-making about the extent to which funding is approved for a specific transport project (Gomez-Lobo, 2012; Mouter et al., 2013). On the other hand, in practices, such as Sweden and Norway CBA is formally used to rank large numbers of transport investments against each other (Eliasson et al., 2015). In these countries, CBA is primarily used for choosing investments from a shortlist of suggested investments given a total available budget. Hence, it is the relative ranking of investments that is of importance, rather than the absolute level of the net benefits.

The demand forecasts, cost estimates, benefit valuations, and effect assessments that are conducted as part of CBAs are all subject to various degrees of uncertainty. Hence, the question is, given such uncertainties, how robust the policy conclusions of CBA are. Using simulations based on real data on national infrastructure plans in Sweden and Norway, scholars studied how investment selection and total realized benefits change when decisions are based on CBA assessments subject to several different types of uncertainty (Asplund and Eliasson, 2016; Börjesson et al., 2014) (Robustness of CBA; Welde and Odeck). Their results indicate that realized benefits and investment selection are surprisingly insensitive to all studied types of uncertainty, even for high levels of uncertainty.

Impact of Transport Cost-Benefit Analysis on Public Decision Making

Several researchers tried to uncover the extent to which CBA actually impacts decision-making by investigating the statistical relation between the results of CBA studies and political decisions on investments in transport infrastructure using quantitative analyses (Annema et al., 2017; Eliasson et al., 2015; Fridstrøm and Elvik, 1997; Odeck, 2010). The broad picture is that these studies show that there is no significant statistical relation between the monetized effect estimations in CBA studies and political decisions. The exception is the Swedish Transport Administration's selection in the construction of the Investment Plan 2010–21, which seems to be strongly affected by CBA results (Eliasson et al., 2015). However, although decisions of Swedish civil servants were strongly affected by the CBA outcome of transport projects, politicians' decisions were only weakly affected, and only for small projects.

Furthermore, several studies have analyzed how politicians use CBA in the context of transport investment decisions by interviewing politicians (Mouter, 2017a,b; Nyborg, 1998; Sager and Ravlum, 2005; Sager and Sørensen, 2011). These qualitative studies conclude that CBA is at best one of the factors that influences politicians' judgments. None of the politicians interviewed in these studies stated that they solely base their judgment on the results of CBA studies. When politicians use CBA in forming their standpoint, it is most likely that the results affect their viewpoint about the desirability of different alternatives of a specific transport project (Mouter 2017a). Especially when a politician supports a project, but does not have a strong preference for one of the alternatives, there is a chance that the CBA affects the desirability judgment of the politician regarding this particular decision. In countries in which CBA is formally institutionalized as a ranking tool, politicians use CBA as a screening device to pick projects requiring closer political attention, but few seemed to actually use it to rank projects (Nyborg, 1998).

Various studies investigate the role that the CBA has played in decision-makers' choices between light rapid transit (LRT) and bus rapid transit (BRT) (Nicolaisen et al., 2017). These studies establish that both in Denmark, France, and the United Kingdom decision makers choose for LRT systems instead of BRT systems even though LRT systems score relatively poorly in CBA studies. Decision makers justify their choice for LRT systems in terms of the branding value, positive image for public transportation and the perceived provision of viable, affordable and attractive alternatives to the automobile, while contributing to more liveable and sustainable cities.

Opportunistic and Symbolic Use

It is more likely that politicians use a CBA as ammunition in discussions with other politicians than as an input for their desirability judgment of transport projects (Mouter, 2017a; Nyborg, 1998). When the CBA does not support a politician's opinion she will criticize the study and she will emphasize the importance of CBA when the results support her opinion, even when she did not use CBA in forming her opinion at all. The literature also identifies that politicians use CBA to make themselves and their decisions look more rational which is also called: "symbolic use" (Sager and Ravlum, 2005; Mouter, 2017a). The institutionalization of CBA has symbolic value for politicians, as the search for and processing of information may itself send out signals that will enhance the status of the political body. In Norway, for instance, researchers observed that the main function of CBA—and analytic planning input in general—was to legitimize the Norwegian Transport Plan and the political process related to it (Sager and Sørensen, 2011). Politicians must be able to show the public that the output of expert analysis was available to them when they made their decisions, so it can be credibly stated—should the need arise—that expert advice was considered as part of the policy-making.

Particularly executives (e.g., ministers) use CBA as a means for depoliticizing the political debate. In this case, CBA is used as a "rational argument" to support the wishes of the executive. In the Netherlands, it was found that top-level civil servants pursued to produce a set of rational arguments, which supported all of the executive's preferred decisions in a consistent way, which makes it difficult for the political opposition to challenge the consistency of the executive's decisions during a debate (Mouter, 2017a). These civil servants aimed to avoid any inconsistencies in argumentation as this can force an executive into revealing her real (irrational) argument for (not) supporting a project, which is an unwelcome situation, as in general, rational arguments are more convincing than emotional arguments.

Barriers Hampering Politicians' Use of CBA When Forming Their Judgments

The scientific studies in which politicians were interviewed about their use of CBA identified several barriers, which need to be rectified to increase politicians' use of CBA when forming their judgments. A first important barrier, which limits the use of CBA by politicians, is that they don't have enough trust in CBA's impartiality (Mouter, 2017a). Politicians do not seem to think that results of CBAs are deliberately manipulated, but they have the idea that CBA analysts implicitly and even unconsciously make political choices while carrying out a CBA which influences (the communication of) the results (Mouter, 2017a; Nicolaisen, 2012). For instance, politicians believe that in cases in which analysts can make a choice between multiple assumptions which all are defensible, they can be tempted to choose the assumption that best fits the interests of the institution who orders the CBA (Mouter, 2017a). Moreover, politicians believe that analysts also have their hobby horses (e.g., positive about public transport), which can unconsciously bias the assumptions they make when conducting a CBA. Note that politicians who genuinely trust CBA's impartiality may criticize the method on this ground when the CBA does not support their pet project (see Section, Opportunistic Use) and vice versa it is possible that politicians who genuinely distrust may use the results of a CBA in a public debate when the study supports a project they want to push through the system.

A second important barrier is that politicians often receive the CBA too late in the process to (substantially) influence their viewpoint (Mouter, 2017a). Generally, key decisions regarding transport projects—and sometimes also political promises—are made in the early stage of the policy cycle. In this phase, it is highly likely that decision makers merely receive information about the transport project and the problem it needs to resolve from lobbyists or from a first-hand look during a site visit. Decision makers are presented with CBA results at the end stage of the policy cycle. In the Netherlands, it sometimes even happens that Members of Parliament receive a CBA study a few days before the decisive parliamentary debate regarding the transport project (Mouter, 2017a). In those cases it is highly unlikely that a CBA would have any impact on political decisions. Among Dutch politicians, there is a consensus that a CBA should be published 2 months before a debate to have any impact on the decisions being made during the

debate. This will give politicians enough time to discuss with their colleagues whether the results of a CBA should lead to a reconsideration of a viewpoint. Moreover, the publication of a CBA long enough before a political debate enables politicians to ask a confidante to verify the study, which increases their trust in the results.

A third barrier is that politicians assign a relatively low weight to the results of a CBA in their judgments because they contest CBA's normative premises (Mouter, 2017a; Nyborg, 1998). Politicians are of the view that as a result of the discrepancy between the premises of CBA and their belief system, projects which coincide with their own belief system score relatively poorly in CBAs.

The first normative decision analysts need to make when conducting a CBA concerns the individuals that are (not) included in the analysis, which is also known as the question of "standing." CBA generally adopts a welfarist approach to social evaluation, which means that the preferences of individual citizens form the basis of a CBA (Sen, 1979). This implies that costs and benefits for animals, and nature in general, only count when humans value them (Baum, 2009). This normative choice might not be in line with the belief system of politicians from Green Parties.

A second normative decision concerns the weighting of the preferences of different individuals when establishing the social welfare effect of the project. Economists use a social welfare function, as a formal representation of the value judgments regarding the emphasis society should place on the interests of different citizens. In essence, the question is whether the social welfare function is utilitarian (an equally large weight is attached to everybody's utility changes regardless of their current situation) or non-utilitarian (utility changes of citizens receive a different weight when determining the social welfare effect of a government project). The common approach in practice is to postulate a utilitarian social welfare function. The fact that CBA generally postulates a utilitarian social welfare function is often criticized in the academic literature for ignoring distributive justice considerations that are important for political decision-making (Nyborg, 2014).

Studies investigating political decision-making regarding transport projects established that politicians particularly regard "spatial equality" of transport investments as a key consideration in their decisions involving the allocation of investments in a national transport program for infrastructure investments (Mouter et al., 2017). For instance, politicians perceive that transport projects in densely populated areas perform better in CBAs than projects in sparsely populated areas as a result of the normative assumption in conventional CBAs that equal weight should be assigned to the utility effects of every individual in a society. These politicians, however, argue that it is fair to balance transport investments across the country to some extent, because all over the country citizens pay taxes, which makes it justifiable to improve citizens' mobility all over the country. More formally, politicians assign a higher (lower) weight to benefits of transport investments in regions, which receive relatively less (much) benefits from the national transport investment program in their social welfare function.

Analysts could deal with the inevitable value judgments implicit in CBA by communicating these judgments in a transparent way. Interestingly, politicians seem to prefer, on the one hand, that all calculations in the CBA are done in an impartial way and, on the other hand, they want the inherent partiality of the method to be recognized by making it explicitly clear which value judgments are made (Mouter, 2017b). Hence, the ideal situation seems to be that politicians can trust that analysts executed the calculations in an impartial way, while being aware of the value judgments implicit in the CBA. Apart from improving transparency regarding normative assumptions in CBA, scholars argue that a CBA report can serve politicians who disagree with the value judgments through providing supplementary information. For instance, politicians can be provided with information regarding the spatial distribution of transport benefits accruing from an investment program to better enable them to consider both the aggregate benefits of the investment program and the distribution of benefits across regions in their decisions.

What Explains the Positive Attitude of Politicians Toward CBA?

Although the results of a CBA do not seem to substantially impact decision-making in a political environment, academic studies universally find that politicians and civil servants have a positive attitude toward the institutionalization of the method (Nyborg, 1998; Mouter, 2017b). Studies which analyzed politicians' and policy makers' attitudes toward CBA in the domains of flood protection and environmental regulation find the same results (Dehnhardt, 2013; Hahn and Tetlock, 2008). The institutionalization of CBA can bring several benefits to decision-making on transport projects. The literature distinguishes at least eight categories of positive features of CBA.

First, CBA is based on a rigorous theoretical framework being welfare economics that allows for the trade-off between money and social impacts. The principles of welfare economics provide CBA researchers and users with a very clear frame of reference when reflecting on the impacts of policy measures that should (not) be included in a CBA, and how these impacts could be measured and monetized. This is a clear advantage compared to methods such as multi-criteria analysis, which are not build on such a rigorous theoretical framework (CBA versus multicriteria analysis; Beria).

Second, CBA enhances the attention given to citizens' interests in the political process. Two important value judgments underlying welfare economics and CBA are "individualism" and "non-paternalism." In combination, these postulates assert that the welfare impacts of individual members of society resulting from the project form the basis for establishing the societal welfare effect (individualism) and individuals are conceived to be the best judge of their own welfare (non-paternalism). Because impacts for citizens and firms, and not the interests of stakeholders, academics or policy makers, are the focal point of a CBA analysis, the instrument is also known as the "tax payers only model of representation at the political negotiation table."

Third, CBA can be an antidote to overcome cognitive limitations and biases from causing policy-makers to neglect vital aspects of proposed policies. This is also known as "the cognitive argument for CBA." (Sunstein, 2000) Politicians may face hundreds of

projects, and it is simply not possible to completely process all these options. In such situations, humans are bound to use simple heuristics and appraisal tools such as CBA make it easier for politicians to structure information and remember and consider all or most aspects of a suggested project. Politicians think that this is advantageous because it can prevent them from forgetting to consider important consequences for citizens and firms in the decision-making process.

Fourth, CBA is considered to be a useful building block for forming an opinion regarding public projects because the method provides insights into the order of magnitude of positive and negative impacts of a project by translating these effects into money. This provides guidance when making decisions. When the societal costs of a project are (substantially) higher than the benefits this can alarm politicians to not support a project. Fifth, due to standardization and the fact that the final indicators of a CBA (e.g. the benefit-cost ratio) communicate very clearly, CBA makes projects comparable. Sixth, CBA can enhance the sharpness of political debates and the underpinning of political decisions. That is, politicians have found a need to argue in a more precise way about why they want a project despite a negative CBA, or why they don't want a project despite a positive CBA. Dutch politicians argue that without a CBA, quite frequently, the necessity of a government project is underpinned in a very general way. The result of a very negative CBA is that these general arguments will be contested in political debates by politicians opposing the project. For instance, with regard to the negative CBA results of light rapid transit (LRT) it was found that decision makers made great efforts to emphasize and inscribe the strategic values of the projects, not included in the CBAs, into diagrams, reports, maps, etc., in order to visualize and make these values "real" for decision makers (Nicolaisen et al., 2017).

Seventh, civil servants use CBAs to optimize infrastructure projects in the early phases of their planning process. Finally, CBA can act as a filter (gatekeeper) to prevent weak projects proceeding very far through the planning process. For instance, civil servants from higher tier governments use a negative CBA as an argument in discussions with civil servants from lower tier governments to clarify that it would be better not to have any high expectations about receiving a national contribution for the project because of the poor CBA score (Mouter, 2017b). Hence, in this process, many projects are terminated before they even reach national executives.

Biography



Niek Mouter is an assistant professor in Infrastructure Project Appraisal at Delft University of Technology, the Netherlands, faculty Technology, Policy, and Management. His research primarily revolves around two research lines: (1) The inclusion of equity and other ethically important considerations in the appraisal of infrastructure projects in general and Cost-Benefit Analysis in particular; (2) Improving the usability of Cost-Benefit Analysis for policy makers. He (co)developed a novel assessment approach called "Participatory Value Evaluation."

References

- Annema, J.A., Frenken, K., Koopmans, C.C., Kroesen, M., 2017. Relating cost-benefit analysis results with transport project decisions in the Netherlands. *Let. Spat. Resour. Sci* 10 (1), 109–127.
- Asplund, D., Eliasson, J., 2016. Does uncertainty make cost-benefit analyses pointless? *Transp. Res. Part A* 92, 195–205.
- Baum, S.D., 2009. Description, prescription, and the choice of discount rates. *Ecol. Econ* 69 (1), 197–205.
- Boardman, A. E., Greenberg, D. H., Vining, A. R., Weimer, D. L., 2013. *Cost-Benefit Analysis Concepts and Practice*. Harlow, Pearson.
- Boadway, R., 2006. Principles of cost-benefit analysis. *Public Policy Rev* 2 (1).
- Börjesson, M., Eliasson, J., Lundberg, M., 2014. Is CBA ranking of transport investments robust? *J. Transp. Econ. Policy* 48 (2), 189–204.
- Dehnhardt, A., 2013. Decision-makers' attitudes towards economic valuation & a case study of German water management authorities. *J. Environ. Econ. Policy* 2 (2), 201–221.
- Eliasson, J., Lundberg, M., 2012. Do cost-benefit analyses influence transport investment decisions? Experiences from the Swedish transport investment plan 2010–2021. *Transport Rev* 32 (1), 29–48.
- Eliasson, J., Börjesson, M., Odeck, J., Welde, M., 2015. Does benefit-cost efficiency influence transport investment decisions? *J. Transport Econ. Policy* 49, 377–396.
- Fridström, L., Elvik, R., 1997. The barely revealed preference behind road investment priorities. *Public Choice* 92, 145–168.
- Gomez-Lobo, A., 2012. Institutional safeguards for cost benefit analysis: lessons from the Chilean national investment system, *J. Cost&Benefit Anal.* 3.
- Hahn, R.W., Tetlock, P.C., 2008. Has economic analysis improved regulatory decisions? *J. Econ. Perspect* 22 (1), 67–84.
- Mackie, P.J., Worsley, T., Eliasson, J., 2014. Transport appraisal revisited. *Res. Transp. Econ* 47, 3–18.
- Mouter, N., 2017a. Dutch politicians' use of cost-benefit analysis. *Transp* 44 (5), 1127–1145.
- Mouter, N., 2017b. Attitudes of Dutch politicians towards cost-benefit analysis. *Transport Policy* 54, 1–10.

- Mouter, N., Annema, J.A., Van Wee, B., 2013. Attitudes towards the role of cost-benefit analysis in the decision-making process for spatial-infrastructure projects: a Dutch case study. *Transp. Res. Part A* 58, 1–18.
- Mouter, N., van Cranenburgh, S., van Wee, B., 2017. An empirical assessment of Dutch citizens' preferences for spatial equality in the context of a national transport investment plan. *J. Transport Geogr* 60, 217–230.
- Mouter, N., 2018. A critical assessment of discounting policies for transport cost-benefit analysis in five European practices. *Eur. J. Transport and Infrastruct. Res* 18 (4), 389–412.
- Nicolaisen, M.S., 2012. Forecasts: Fact or Fiction? Uncertainty and Inaccuracy in Transport Project Evaluation. Department of Development and Planning, Aalborg University, Denmark.
- Nicolaisen, M.S., Olesen, M., Olesen, K., 2017. Vision vs. evaluation -case studies of light rail planning in Denmark. *Eur. J. Spat. Dev* 65 .
- Nyborg, K., 1998. Some Norwegian politicians's use of cost-benefit analysis. *Public Choice* 95, 381–401.
- Nyborg, K., 2014. Project evaluation with democratic decision-making: What does cost-benefit analysis really measure? *Ecol Econ* 106, 124–131.
- Odeck, J., 2010. What determines decision-makers's preferences for road investments? evidence from the Norwegian road sector. *Transport Rev* 30 (4), 473–494.
- Sager, T., 2016. Why don't cost-benefit results count for more? The case of Norwegian road investment priorities. *Urban Plann. Transport Res* 4 (1), 101–121.
- Sager, T., Ravlum, I.A., 2005. The political relevance of planners's analysis: the case of a parliamentary standing committee. *Plann. Theory* 4 (1), 33–65.
- Sager, T., Sørensen, C.H., 2011. Planning and political steering with new public management. *Eur. Plann. Stud* 19 (2), 217–241.
- Sen, A.K., 1979. Utilitarianism and welfarism. *J. Philos* 76 (9), 463–489.
- Sunstein, C.R., 2000. Cognition and cost-benefit analysis. *J. Legal Stud* 29 (52), 1059–1103.

Causal Inference for Ex Post Evaluation of Transport Interventions

Daniel J. Graham, Transport Strategy Centre, Imperial College London, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	283
Transports Interventions and Outcomes	283
Ex Post Evaluation and the Potential Outcome Framework for Causal Inference	284
Ex Post Evaluation via Model-Based Adjustment for Confounding	285
Outcome Regression	286
Propensity Score Methods	286
Doubly Robust Estimation	287
Ex Post Evaluation Under a Nonignorable Treatment Assignment	287
Instrumental Variables	287
Difference-in Differences	288
Regression Discontinuity Designs	289
Summary	289
References	290
Further Reading	290

Introduction

The objective of ex post evaluation is to quantify retrospectively the impacts that interventions have had on defined outcomes of interest. In the transport domain, such evaluations can be made in relation to almost any type of “intervention,” including physical projects, regulations, price changes, or other policy initiatives.

Our presupposition is that evidence from ex post evaluation is useful only when it points to a *causal*, rather than merely associational, relationship between intervention and outcome. Ultimately, we want to be able to identify effects that arose due to a course of action taken, relative to some other course of action that was not pursued (often no intervention). For transport evaluation, data are observed rather than generated experimentally and this limits our ability to compare outcomes under different treatment regimes. Furthermore, transport interventions are typically nonrandomly assigned and this obscures causality further. To infer cause-effect relationships in these settings, we require an appropriate causal inferential framework for identification and statistical modeling methods appropriate to the data in hand.

In this chapter we review statistical approaches for ex post evaluation of transport interventions that can be used to infer cause-effect relationships from observational data. The aim of these *causal inference* techniques is to quantify impacts that have explicitly occurred due to intervention (or “treatment”). Comprehensive reviews of the statistical principles and methods underpinning causal inference are provided by van der Laan and Robins (2003), Imbens and Wooldrige (2009), Imbens and Rubin (2015), Pearl et al. (2016), and Abadie and Cattaneo (2018). Through ex post evaluation we can assess how well resources have been allocated in the past according to some defined metrics of interest, and crucially we can also construct informed forecasts of the impacts that transport interventions may have in the future.

The chapter is structured as follows. Transports interventions and outcomes defines key concepts in causal inference. Ex post evaluation and the potential outcome framework for causal inference introduces the potential outcome framework for causal inference as a working model within which to conceptualize ex post evaluation. Ex post evaluation via model-based adjustment for confounding and Ex post evaluation under a nonignorable treatment assignment review techniques for treatment effect estimation under ignorable and nonignorable treatment assignment conditions respectively. Summary summarizes and concludes.

Transports Interventions and Outcomes

Ex post evaluation aims to quantify the causal effect of a transport intervention, or set of interventions, on an outcome of interest. Typically, we seek to compare the intervention outcome to other “counterfactual” scenarios, perhaps involving no intervention or some alternative intervention. It is important at the outset to define these two key concepts of *intervention* and *outcome*.

Interventions, termed “treatments” in the literature, are defined in the broadest sense to encompass any conceivable manipulation of the transport system. For instance, a treatment could involve the construction of a new link, an expansion in transport capacity, the imposition of a new regulation, a change in transport prices, a change in the frequency or quality of service, and so on. Treatment variables can thus be binary, multivalued, or continuous. Table 1 gives examples of transport interventions classified as treatment variables.

Table 1 Transport interventions classified as treatment variables

<i>Binary treatment</i>	<i>Multivalued treatment</i>	<i>Continuous treatment</i>
+ Δ capacity/no Δ capacity	Number + Δ capacity routes	Length of + Δ capacity
New network link	Number of new links	Distance to new link
New station	Number of new stations	Station accessibility metric
Tolled/untolled route	Number of tolled routes	+ Δ % price on tolled routes
Introduction of speed camera	Number of speed cameras	Speed cameras per km ²
Train service/no train service	Train service headway	Train pax. capacity
Peak service/off-peak service	Number of peak services	% of service in peak

Outcomes represent phenomena that we believe a-priori will respond to manipulation of the transport system. Outcomes of interest could measure traffic conditions (e.g., speeds, flow, safety, congestion), economic characteristics (e.g., output, productivity, growth), mode shares, environmental consequences, social concerns, and so on.

The relationship between outcomes and interventions are analyzed for *units* of study, normally defined by the data available for analysis. Units could be geographical zones, network links, people, households, firms, or cities.

In addition to outcomes and interventions, for ex post evaluation we often make use of a third source of data on measured characteristics of units. While our fundamental concern is on the effect of the treatment on the outcome, we also recognize that the units under study are heterogeneous and that this could be relevant in making comparisons. So, we use data on unit characteristics to measure and adjust for relevant differences between units.

Ex Post Evaluation and the Potential Outcome Framework for Causal Inference

Using the definitions just provided, we can represent the typical set up for ex post analyses as one in which the data available for evaluation are realizations of a random vector, $Z_i = (Y_i, D_i, X_i)$, where for our n units of observation, $i = (1, \dots, n)$, Y_i denotes a response (or outcome) of interest, D_i is the treatment (or exposure) received (i.e., a transport intervention), and X_i is a vector of pretreatment covariates describing characteristics of the units under study. As mentioned previously, treatment variables can be binary, (i.e., $D \in \{0, 1\}$); multivalued, in which dose d can take values in m categories $D \equiv (d_0, d_1, \dots, d_m)$; or continuous with dose d taking values in $D \subseteq \mathbb{R}$.

The objective of ex post evaluation is to quantify the effect of treatment on outcome. A useful working model for this type of inferential problem is offered by the *potential outcomes* framework for causal inference. Under this approach, we define our target of inference not simply in relation to the outcomes that we actually observe, but also in relation to those that could potentially have been observed under different courses of action (counterfactuals). Specifically, we define the Average Treatment Effect (ATE) as the estimand of interest, which in the case of binary treatments, is the difference in average outcomes under intervention and nonintervention, $\tau(1) = E[Y_i(1)] - E[Y_i(0)]$, where, $Y_i(1)$ and $Y_i(0)$ are the *potential outcomes* for unit i under treated and control status, respectively. In the case of continuous and multivalued treatments the ATE takes the form $\tau(d) = E[Y_i(d)] - E[Y_i(0)]$, for any dose of interest d .

The observed data reveal only actual, not potential, outcomes. Thus, in the case of binary treatments we observe the random variable $Y_i = Y_i(1)I_1(D_i) + Y_i(0)(1 - I_1(D_i))$, where, $I_1(D_i)$ is the indicator function for receiving the treatment, but we do not observe the joint density, $f(Y_i(1), Y_i(0))$, since the two outcomes never occur together. Similarly, for a continuous or multivalued treatment we observe only $Y_i(D_i)$, and outcome at all other levels, $d \neq D_i$, are unobserved or *counterfactual outcomes*.

The fact that we cannot observe all potential outcomes means that *individual causal effects*, for example, $Y_i(1) - Y_i(0)$ or $Y_i(d) - Y_i(0)$ for each unit i , are not quantifiable and this is sometime referred to as *the fundamental problem of causal inference*. Instead, the potential outcomes framework focuses the inference problem on estimation of averages in the form of ATEs, and determines the conditions under which these can be identified validly from the data.

One option for ATE estimation would be to simply take mean outcomes for groups of units defined by treatment status and compare (i.e., in the binary case we could compare the mean outcome for treated units with the mean outcome for untreated units). However, a key feature of transport interventions is that they are rarely randomly assigned, and when unit characteristics influence both treatment assignment and outcome simultaneously, bias will ensue in ATE estimation based on naïve comparisons of group means. This implications of nonrandom assignment is illustrated in Fig. 1.

Note that under randomization the characteristics of units in the sample, denoted X , have no influence on the treatment received (i.e., on D). Consequently, outcomes are unconditionally independent of the treatment assignment mechanism (i.e., all units in the sample have an equal probability of being assigned to treatment), and a simple comparison of mean outcomes by treatment groups will yield valid inference about the ATE for the population. Under a nonrandomized assignment, the allocation of the treatment depends on characteristics X , which are themselves important in determining outcome Y . Consequently, simple comparisons of mean responses across different treatment groups will not in general reveal a causal effect. Under these conditions, we refer to X as *confounders* and describe the treatment effect estimation problem as *confounded*. A confounder, therefore, is simply a random variable

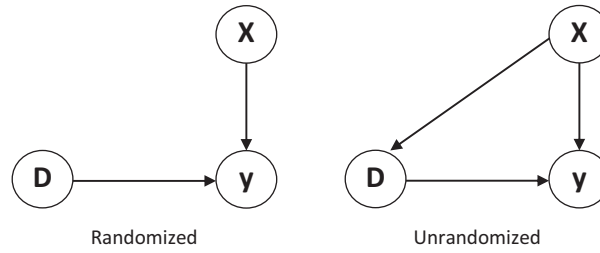


Figure 1 Directed acyclic graph of randomized and nonrandomized treatment assignment.

that influences the outcome of interest, but that is also important in determining (nonrandom) assignment to treatment. It is worth noting that *confounding* can arise dynamically, with past outcomes or treatments of units serving as baseline confounders (e.g., X can contain lagged values of D and Y).

Transport interventions are typically nonrandomly assigned, for example, because they are used to induce improvements in congested locations, in accessibility constrained locations, or in locations with poor economic performance. This implies that locations that receive transport interventions will tend to be in some sense different from those that do not. If this is the case, *confounding* is said to be present and this makes it hard to separate the effect of the treatment from other influences through simple comparisons based on treatment status alone. For ex post evaluation, the implication is that in order to obtain valid unbiased causal estimates of the ATE we need to somehow account for differences in confounding characteristics.

Broadly speaking there are two way of doing this. First, through model-based adjustment for confounding, in which differences between units in characteristics X , are measured and included within a model to obtain marginal causal effects (i.e., ATE net of confounding influences from X). Second, by developing models which exploit sources of exogenous variance to obtain causal estimates without explicit representation of X in the model. In the following two sections we consider each of these approaches in turn.

Ex Post Evaluation via Model-Based Adjustment for Confounding

A key insight of the potential outcomes approach is that to estimate the ATE we do not need to observe all potential outcomes, even under a nonrandom treatment assignment. Observed data are sufficient to conduct model-based adjustment for confounding as long as three key assumptions hold.

1. **Conditional independence:** The potential outcomes for unit i must be conditionally independent of the treatment assignment given a (sufficient) set of observed covariates X_i . For binary treatments this assumption is written $Y_i(0), Y_i(1) \perp D_i = 1 | X_i$, and for multivalued or continuous treatments $Y_i(d) \perp D_i = d | X_i$, for all d . This condition simply says that, if we can measure the X vector that influences nonrandom assignment in Fig. 1, then we can use this to adjust for bias in ATE estimation.
2. **Common support:** Conditional on covariates X , the probability of assignment to the treatment must be strictly positive in infinite samples. For binary treatments, the common support assumption is written as $0 < \Pr(I_1(D_i) = 1 | X_i = x) < 1, \forall x$. For multivalued or continuous treatments, we require common support by treatment status in the covariate distributions within some region of dose, $C \subseteq D$, say. A sufficient condition is that for any subset of C , $A \subseteq C$, then $\Pr(D_i \in A | X_i = x) > 0, \forall x$. The intuition behind the common support, or overlap, assumption is that if some sub-populations observed in X have zero probability of receiving (or not receiving) the treatment, then it does not make sense in these cases to talk of a treatment effect since the counterfactual does not exist in the observed data.
3. **Stable unit treatment value assumption (SUTVA):** The relationship between observed and potential outcomes must comply with the SUTVA, which requires that observed outcomes under a given treatment allocation must be equivalent to potential outcomes under that allocation. For binary treatments this implies that $Y_i = I_1(D_i)Y_i(1) + (1 - I_1(D_i))Y_i(0), \forall i = 1, \dots, n$; and for multivalued or continuous treatments SUTVA requires that $Y_i = I_d(D_i)Y_i(d), \forall d \in D, i = 1, \dots, n$. A key implication of the SUTVA is that the outcome for each unit should be independent of the treatment status of other units, or in other words, there should be no interference in treatment effects across units. SUTVA also implies that there are no different versions of the treatment. The no interference assumption is generally satisfied when the units are physically distinct and have no means of contact. Violations of the assumption can occur when proximity of units allows for contact and this presents a particular concern for transport applications.

The three assumptions defined above, which are together referred to as *strong ignorability*, provide the conditions for identification of causal effects from observational data. This can be demonstrated as follows. For binary treatments, the ATE can be derived as

$$\begin{aligned} \tau(1) &= E[Y_i(1)] - E[Y_i(0)] \\ &= E_X[E(Y_i(1)|X_i = x) - E(Y_i(0)|X_i = x)] \end{aligned}$$

$$\begin{aligned}
&= E_X[E(Y_i(1)|X_i = x, D_i = 1) - E(Y_i(0)|X_i = x, D_i = 0)] \\
&= E_X[E(Y_i|X_i = x, D_i = 1) - E(Y_i|X_i = x, D_i = 0)],
\end{aligned}$$

where, the equality of the second and third lines follows from conditional independence, the SUTVA allows the substitution of observed for potential outcomes to give line 4, and overlap ensures that the population ATE is estimable since there are units in both the treated and untreated groups. Thus, in place of unobservable $E[Y_i(1)]$ and $E[Y_i(0)]$, we are able to substitute conditional expectations of observable quantities.

Similarly, for continuous or multivalued treatments the Average Potential Outcome (APO), or *dose-response function*, under a given dose d is defined, $\mu(d) = E[Y_i(d)]$, and can be derived as,

$$\mu(d) = E[Y_i(d)] = E_X[E(Y_i(d)|X_i = x)] = E_X[E(Y_i(d)|X_i = x, D_i = d)] = E_X[E(Y_i|X_i = x, D_i = d)]$$

where the second equality follows from conditional independence, the third from the SUTVA, and the overlap assumption ensures that the APO is estimable since there are comparable units across treatment levels. Note that the ATE is the difference between APOs, that is, $\tau(d) = E[Y_i(d)] - E[Y_i(0)] = \mu(d) - \mu(0)$.

To proceed we need to estimate the relevant conditional expectations in the equations above. There are three approaches that are commonly used in the literature to do this: outcome regression (OR), propensity score (PS) models, and mixed or Doubly Robust (DR) models.

Outcome Regression

Outcome regression models can be used to estimate ATEs under nonrandom assignment. This can be achieved, for instance, via a Generalized Linear Model (GLM) or some parametric or semiparametric variant thereof. If the OR model, denoted $E[Y_i|D_i, X_i] = \psi^{-1}\{m(D_i, X_i, \beta)\}$ for known link function ψ , regression function $m(\cdot)$, and unknown parameter vector β ; is correctly specified for the mean response then the ATE can be consistently estimated in the binary case using

$$\hat{\tau}_{OR}(1) = \frac{1}{n} \sum_{i=1}^n [\psi^{-1}\{m(1, X_i, \hat{\beta})\} - \psi^{-1}\{m(0, X_i, \hat{\beta})\}]$$

And in the case of continuous or multivalued treatments using

$$\hat{\tau}_{OR}(d) = \frac{1}{n} \sum_{i=1}^n [\psi^{-1}\{m(d, X_i, \hat{\beta})\} - \psi^{-1}\{m(0, X_i, \hat{\beta})\}].$$

Note that covariate vector X is included in the regression models to adjust for confounding and thus to invoke conditional independence.

Propensity Score Methods

The propensity score (PS) is a random variable that measures the conditional probability of being assigned to some value of treatment given background confounding characteristics. For unit i we denote the PS by $\pi_i = \Pr(D|X_i, \alpha)$. As noted above, if observed covariates are sufficient to represent sources of confounding then they can be used to establish conditional independence between outcomes and treatment assignment. An important property of the PS is that conditional independence can be established conditional on the scalar PS rather than on the full covariate vector X . A key advantage in using the PS is that we avoid the need to condition on a potentially high dimensional covariate vector, and it is this dimension reducing property that allows for effective implementation of a number of flexible estimators for ATEs.

PS are not observed but are calculated by estimating the relationship between D and X using a regression model $E[D_i|X_i] = \psi^{-1}\{m(X_i, \alpha)\}$ for link function ψ , regression function $m(\cdot)$, and unknown parameter vector α . The estimated parameters of this model are then used to compute propensity scores, $\hat{\pi}_i = \hat{\pi}(D|X_i; \hat{\alpha})$. Once we have the estimated PS, they can be used to form a number of different nonparametric and semiparametric estimators via weighting, matching, stratification, blocking, and regression.

To illustrate, we will consider here one of the simplest PS-based estimators: the inverse PS weighting estimator. Under inverse PS weighting estimation of the ATE for binary treatments takes the form

$$\hat{\tau}(1) = \frac{1}{n} \sum_{i=1}^n \left[\frac{I_1(D_i) \cdot Y_i}{\hat{\pi}(D|X_i; \hat{\alpha})} - \frac{(1 - I_1(D_i)) \cdot Y_i}{1 - \hat{\pi}(D|X_i; \hat{\alpha})} \right],$$

and for continuous and multivalued treatments

$$\hat{\tau}(d) = \frac{1}{n} \sum_{i=1}^n \left[\frac{I_d(D_i) \cdot Y_i}{\hat{\pi}(D|X_i; \hat{\alpha})} - \frac{(1 - I_d(D_i)) \cdot Y_i}{1 - \hat{\pi}(D|X_i; \hat{\alpha})} \right],$$

where $I_1(D_i)$ and $I_d(D_i)$ are indicator for receipt of treatment.

The principle of IPW estimation is creation of a *pseudo-sample* to simulate random assignment by using the conditional probabilities to mimic the sample representation of units that would occur under randomization. The intuition is as follows. We know that under a random assignment the probability of assignment to treatment is uniform across unit in the sample. By measuring how baseline characteristics cause deviations from this uniform probability, we can weight units to effectively alter their representation in the sample such that we get back to an assignment, that is, for all intents and purposes, *as good as random*. Note that the consistency of ATE estimation under this approach relies on the PS model being correctly specified (i.e., the estimated PS must provide an accurate measure of the conditional probability of assignment to treatment).

Doubly Robust Estimation

Doubly robust (DR) estimators combine both an OR model and a PS model such that valid causal estimates of ATEs can be obtained when either the OR or PS model is correctly specified, but we do not require both to be correct. It is therefore useful when the analyst can formulate two models for evaluation but is unsure a-priori as to which best represents the link between treatment assignment and outcome.

DR estimation is typically achieved by weighting or augmenting the OR model with inverse PS covariates. For example, a DR Augmented Outcome Regression (AOR) model has the form

$$E[Y_i | D_i, X_i \hat{\kappa}(D_i, X_i)] = \Psi^{-1}\{m(D_i, X_i, \hat{\kappa}(D_i, X_i), \xi)\}$$

where $\hat{\kappa}(D_i, X_i)$ is a vector of inverse PS covariates estimated from observed data. In the case of binary treatments

$$\hat{\kappa}(D_i, X_i) = \frac{I_1(D_i)}{\hat{\pi}(D|X_i; \hat{\alpha})} + \frac{(1 - I_1(D_i))}{1 - \hat{\pi}(D|X_i; \hat{\alpha})},$$

and for multivalued and continuous treatments a discretization of the treatment into Q strata is imposed, $q = (1, \dots, Q)$, and

$$\hat{\kappa}(D_i, X_i) = \sum_{q=1}^Q \frac{I_q(D_i)}{\hat{\pi}(D|X_i; \hat{\alpha})}.$$

The ATEs are estimated as for binary treatments as

$$\hat{\tau}_{DR}(1) = \frac{1}{n} \sum_{i=1}^n [\Psi^{-1}\{m(1, X_i, \hat{\kappa}(D_i, X_i), \hat{\xi})\} - \Psi^{-1}\{m(0, X_i, \hat{\kappa}(D_i, X_i), \hat{\xi})\}],$$

and in the case of continuous or multivalued treatments as

$$\hat{\tau}_{DR}(d) = \frac{1}{n} \sum_{i=1}^n [\Psi^{-1}\{m(d, X_i, \hat{\kappa}(D_i, X_i), \hat{\xi})\} - \Psi^{-1}\{m(0, X_i, \hat{\kappa}(D_i, X_i), \hat{\xi})\}],$$

The key feature of DR estimators is that APO and ATE estimates are consistent and asymptotically normal when either the OR or PS model are correctly specified, but we do not require both models to be correct. Thus, the analyst effectively has two chances at obtaining valid causal inference. Of course, if the measure of X available in the observed data is insufficient to guarantee conditional independence, then both the OR and PS models will fail and DR estimation will not improve matters.

Ex Post Evaluation Under a Nonignorable Treatment Assignment

The validity of the evaluation methods discussed above requires us to maintain that *strong ignorability* holds (see section Ex Post Evaluation via Model-Based Adjustment for Confounding). In practice, this is often untenable, either because there are insufficient measured covariates to establish conditional independent, or because other sources of endogeneity (e.g., reverse causality or measurement error) are at play inhibiting a causal interpretation of the data.

There are a number of popular estimators that are used in this setting to obtain causal estimates of the ATE. Some use additional variables (instruments) to extract exogenous variation in treatments, while others exploit quasi-experimental conditions for identification. Here we briefly review three of the most commonly used approaches in economic evaluation: instrumental variables (IV), difference-in-differences (DID), and regression discontinuity designs (RDD).

Instrumental Variables

The IV estimator is well known and widely used in empirical studies across diverse fields of study. The key principles of IV estimation is to use exogenous variation, in the form of instruments, to nullify the bias due to confounding, measurement error, or reverse causality. The logic of the IV method is as follows:

1. Find a set of instruments which are exogenous to the outcome but highly correlated with the treatment.
2. Use the instruments to enforce orthogonality between the error term and an instrument transformed design matrix.

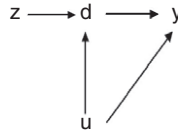


Figure 2 Relationships in instrumental variables estimation.

The relationships assumed in IV estimation are shown graphically in Fig. 2 in the context of the linear regression model $y = D\tau + u$ with instrument matrix Z .

The defining characteristics of the IV model are that changes in z are associated with changes in d , but do not lead to changes in y other than through d . Thus, z is causally associated with d but *definitely* not with u . A common method used to obtain IV estimates is two-stage Least Squares (2SLS), which comprises two steps:

1. Regress each column of D on the instrument matrix Z .
2. Regress y on the predicted values from the first stage.

IV can be used to establish causal effects under a nonignorable treatment assignment and is particularly useful when bidirectionality causality is present. However, it is crucial that the two key assumptions of exogeneity and relevance are met, and in practice such instruments can be hard to find. When instruments are only weakly correlated with the endogenous regressors, or when the instruments themselves are correlated with the error term, IV estimation can produce biased and inconsistent estimates. This problem is further compounded by the fact that the available diagnostic statistics do not provide a full proof means for detecting an inadequate instrument specification.

When panel data are available, but exogenous instruments are not, one commonly adopted route forward is to use the dynamic panel Generalized Method of Moments (GMM) estimator for panel data. Dynamic GMM applies differencing to remove the unobserved individual effects and uses the time series nature of the data to derive a set of instruments from lagged levels which are assumed correlated with the differenced covariates but orthogonal to the errors. A set of moment conditions can then be defined and solved within a GMM framework which will be satisfied at the true value of the parameters to be estimated.

Difference-in Differences

Differences-in-differences is a “before and after” treatment effect estimation approach that is applicable when the effect of treatment on units can be represented as a binary variable (i.e., “treated” or “control”). It can reveal impacts associated with exposure to an intervention relative to nonexposure (control), but it cannot tell us about impacts by scale or “dose” of intervention.

A problem in identifying treatment effects via OR, PS, or DR models is that there may be *unobserved* differences between the treated and untreated units which affect outcomes and are also influential in treatment assignment (i.e., unobserved confounders). In addition, there may be temporal trends that affect the outcome variable due to events that are unrelated to the treatment.

The DID estimator addresses such potential sources of bias by using information for both treated and control groups in both pre and post treatment periods. The DID estimator approximates the expression

$$\tau_{DID} = \{E[Y_i(1)|D_i = 1] - E[Y_i(1)|D_i = 0]\} - \{E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]\}.$$

The “double-differencing” of the DID estimator removes two potential sources of bias. Firstly, it eliminates biases in second period comparisons between the treated and control groups that could arise from time invariant characteristics. Secondly, it corrects for time varying biases in comparisons over time for the treated group that could be attributable to time trends unrelated to the treatment.

An estimate of τ_{DID} can be obtained via linear regression. For instance, we can estimate the model

$$Y_{i,t} = \mu + X'_{i,t}\beta + \alpha D_{i,t} + \delta^*t + \tau D_{i,1} + \varepsilon_{i,t}$$

for units of observation i , $i = (1, \dots, n)$ in binary time periods $t \in \{0, 1\}$, with $t = 0$ representing the pretreatment period and $t = 1$ the posttreatment period. In this model $D_{i,t}$ is the treatment indicator variable such that $D_{i,t} = 1$, if unit i has been exposed to the treatment prior to period t and $D_{i,t} = 0$ otherwise, δ is a time specific component, and $\varepsilon_{i,t}$ is a potentially autoregressive error with mean zero in each time period. The effect of the treatment is captured by the parameter τ .

It is important to note two potential limitations with the DID approach. First, it relies on the strong identifying assumption that the average outcomes for the treated and control groups would have followed parallel paths over time in the absence of the treatment. Adding covariates to the linear DID regression (i.e., X) can help in satisfying the parallel trend assumption because it is then assumed to hold conditional on those covariates, thus accommodating heterogeneity in outcome dynamics between the two groups.

Second, the model is sensitive to error specification, and in particular, it has been shown that the existence of correlation within groups or over time periods can adversely affect the performance of the DID estimator.

Regression Discontinuity Designs

RDD methods can estimate ATEs under nonignorability when a given covariate, referred to as the forcing or running variable, partly or completely determines assignment to the treatment. Under a so-called “sharp” RDD design, the conditional probability of receiving the treatment is of size one at some given threshold of the forcing variable, while under a “fuzzy” design the probability change at the threshold is less than one. The RDD method exploits this discontinuity in treatment assignment to study the conditional distribution of the outcome either side of the threshold of the forcing variable. A discontinuity in outcome is interpreted as evidence of a causal effect of the treatment.

For illustration, we will consider here identification under a sharp RDD design in which treatment assignment is a deterministic function of the forcing variable, which we denote by T . The treatment status of unit i is given by

$$D_i = 1 [T_i \geq c]$$

where $1 [T_i \geq c]$ is an indicator function that takes a value of one if the statement in brackets is true or zero otherwise. We seek to estimate the expression

$$\tau_{SRD} = E[Y_i(1) - Y_i(0)|T_i = c] = E[Y_i(1)|T_i = c] - E[Y_i(0)|T_i = c]$$

which is equivalent to the population ATE, if the treatment effect is constant. We cannot observe both expectations because we have unobserved potential outcomes. Instead, we assume continuity of the expectations in T such that

$$E[Y_i(0)|T_i = c] = \lim_{t \uparrow c} E[Y_i(0)|T_i = t] = \lim_{t \uparrow c} E[Y_i|T_i = t]$$

Implying that

$$\tau_{SRD} = \lim_{t \downarrow c} E[Y_i|T_i = t] - \lim_{t \uparrow c} E[Y_i|T_i = t].$$

The ATE we estimate via the sharp RDD design is the difference in the conditional expectation of the outcome either side of the discontinuity. A valid RDD design will provide consistent causal estimates of the ATE without the need to condition on baseline covariates, but it can however be useful to include them anyway to reduce the sampling variability of the estimator and improve precision.

Summary

In this chapter, we have reviewed methods that seek to draw causal inference from observed data and have indicated how they can be applied to undertake ex post evaluation of transport projects. We argue that a causal inference framework, based on potential outcomes, is highly suitable for ex post evaluation because it is specifically designed for instances in which “treatments” are nonrandomly assigned and experimentation is not possible; circumstances that characterize the allocation of transport interventions.

While the methods reviewed here can be used to successfully derive inference about the causal effects of transport interventions, there are several challenges that should be acknowledged.

First, is that for some of the methods, especially those reviewed in section Ex Post Evaluation via Model-Based Adjustment for Confounding, all potential sources of confounding must be measured to obtain valid inference. This onerous data requirement must be met to satisfy the so-called “selection on observables,” or conditional independence, assumption. In practice, we typically face situations in which confounders are unobserved, or measured with error, and it is then difficult to defend methods that require conditional independence to hold. The problem is exacerbated by the fact that there are no reliable diagnostic procedures to comprehensively test for violation of this key assumption.

Second, there are other methods, including all of those reviewed in section five, which perhaps require less data but instead impose stringent identifying assumptions. In the case of IV estimation instruments must be relevant and exogenous, for DID models a parallel trend must hold, while continuity of outcomes and non-manipulation of the treatment threshold must hold for valid identification under RDD estimation. It can be hard to establish empirically whether these assumptions hold in practice and again diagnostics are not always clear cut.

Finally, another key challenge for ex post evaluation relates to the SUTVA condition mentioned above, which applies to all methods reviewed in this chapter. A key implication of the SUTVA is that the outcome for each unit must be independent of the treatment status of other units, or in other words, there should be no “interference” in treatment effects between units. In evaluations based on spatial data, the assumption of no interference is often satisfied naturally when the units are physically distinct and have no means of contact. But violations of the assumption can occur when proximity of units allows for contact or network-based interactions are present. This presents a particular concern for evaluation of transport interventions, which are assigned within networks in which improvements on one link or node can affect outcomes for other spatiotemporal units throughout the network.

References

- Abadie, A., Cattaneo, M.D., 2018. Econometric methods for program evaluation. *Annu. Rev. Econ.* 10, 465–503.
- Imbens, G.W., Rubin, D.B., 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY, USA.
- Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* 47, 5–86.
- Pearl, J., Glymour, M., Jewell, N.P., 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons Ltd., West Sussex, UK.
- van der Laan, M., Robins, J.M., 2003. *Unified Methods for Censored Longitudinal Data and Causality*. Springer, Berlin.

Further Reading

- Athey, S., Imbens, G.W., 2017. The state of applied econometrics: causality and policy evaluation. *J. Econ. Perspect.* 31 (2), 3–32.
- Carbo, J.M., Graham, D.J., Casas, A.D., Melo, P.C., 2019. Evaluating the causal economic impacts of transport investments: evidence from the Madrid-Barcelona high-speed rail corridor. *J. Appl. Stat.* 46, 1714–1723.
- Graham, D.J., McCoy, E.J., Stephens, D.A., 2014. Quantifying causal effects of road network capacity expansions on traffic volume and density via a mixed model propensity score estimator. *J. Am. Stat. Assoc.* 109, 1440–1449.
- Graham, D.J., McCoy, E.J., Stephens, D.A., 2016. Approximate Bayesian inference for doubly robust estimation. *Bayesian Anal.* 11, 47–69.
- Li, H., Graham, D.J., Liu, P., 2017. Safety effects of the London cycle superhighways on cycle collisions. *Accid. Anal. Prev.* 99, 90–101.
- Li, H., Graham, D.J., 2016. Quantifying the causal effects of 20 mph zones on road casualties in London via doubly robust estimation. *Accid. Anal. Prev.* 93, 65–74.

Transport Cost and Location of Firms

Adelheid Holl, Institute of Public Goods and Policy (IPP), CSIC—Spanish National Research Council, Madrid, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	291
Theoretical Approaches to Explain Firm Location	291
Empirical Studies on the Role of Transport Costs for Firm Location	292
The Measurement of Transport Cost	293
Geographical Level of Analysis	293
Types of Firms	293
Types of Transport Infrastructure	294
Trends and a Brief Look Forward	294
Concluding Remarks	295
References	296

Introduction

There exists now a large body of theoretical and empirical literature that shows that transport costs for goods and people affect firm location in multifaceted ways and that changes in transport cost have important implications for the reorganization of firms across space. This article first reviews some of the main theoretical approaches to explain firm location and the role of transport cost. This is followed by a review of empirical studies on transport cost and firm location distinguishing how transport costs are measured, which geographical level is the unit of analysis, and which type of firms and transport infrastructure is analyzed. Transportation technology has changed profoundly over the course of history and with it transport costs and the location of firms. New developments will certainly also influence the location patterns of firms. Extent and direction of changes will, however, depend among other things also on the interactions between changes in transport cost for goods and people, the cost of exchanging complex knowledge and information, and the opportunity cost of time.

Theoretical Approaches to Explain Firm Location

Economists and geographers have long been concerned about the influence of transport cost on the location of firms and the resulting spatial distribution of economic activity. Transport plays a crucial role in the overcoming of the frictions of distance. It enables spatial interaction and economic transactions over distance.

Early models of firm location such as the models by [Von Thünen \(1826\)](#) and [Weber \(1929\)](#) have been centered on transport costs and location. Von Thünen focused on agricultural location and the transport of agricultural produce to a central market. In his model, differences in transport costs to bring the produce to the central market determine what is produced at different distances from the market. Falling transport costs allow production of a particular produce further away from the market center. In the industrial location model of Weber, both transport costs for raw materials and the final good are taken into account. Weber assumed that raw materials are allocated across space but buyers are located also at the market center. Firms search for the least-cost location that minimizes the transport cost for raw materials and that allows for the lowest delivered cost of the final good to the market center. Depending on the transport costs involved in the transportation of the inputs and outputs relatively, the least-cost location is either near the raw material source or at the market center. However, the role of input and output transportation costs is not independent of the production technology adopted as firms can adapt to changes in transportation costs by changing the relative proportions of inputs as well as the characteristics of outputs. Other models have centered on the market area of firms. [Lösch \(1954\)](#), for example, developed a model where buyers are distributed evenly over a market area.

Starting with the work of [Krugman \(1991\)](#), New Economic Geography (NEG) regional models center on increasing returns to scale and transport costs. NEG approaches concentrate on the role of market-size effects and forward and backward linkages giving rise to pecuniary externalities (centripetal forces) that foster geographical concentration, on the one side, and the opposing force of competition in product and factor markets (centrifugal forces) working against such concentration. These models point to complex mechanisms by which transport costs affect firm location. With high transport costs, firms are distributed evenly across regions because of the need to be close to customers to keep transport costs of final goods low. The decrease of interregional transport costs together with increasing returns sets in a process of concentration in core regions as transport cost reductions lower the cost of firms to sell output to customers at greater distance. This extends the market area and facilitates agglomeration in core regions. Greater market size of core regions allows for higher productivity. Firms in core regions thus can pay higher wages that attract workers from peripheral regions. Core regions also have a higher diversity of goods that increases consumer's utility, increasing further the

attractiveness of core regions to labor. The increase in labor raises demand in the goods market that attracts again other firms. In this case, a fall in transport costs leads to the weakening of centrifugal forces and encourages agglomeration in core regions. This, however, also raises factor competition and consequently the price of labor and land.

Further reductions in transport costs could then set in again a process of dispersion as lower transport costs open up new sites for production where firms can take advantage of cheaper labor and land. Transport cost reductions thus change the trade-off between different location factors and consequently the distribution of firms between core and peripheral areas. Whether transport costs reductions are agglomerative or dispersive depends on several factors. Firms that have high relative transport costs may respond to transport cost reduction by concentrating, but firms with lower transport costs may disperse to more peripheral locations to take advantage of lower factor costs. This means that transport cost reductions can in theory reduce but also increase regional inequalities due to the different responses of firms in changing the location of their activities over space (Holl, 2007). NEG points out this potential ambiguity in the impact of transport cost reductions on firm location.

Within the NEG framework, there have also been studies that have shown that interregional and intraregional transportation costs have different effects on firm location. Reducing intraregional transport costs generally leads to attracting firms to that region. However, reductions of interregional transportation costs in a poor region can induce relocation of firms away from this region. Other extensions introduce commuting costs and how together with the cost to transport goods they influence firm location.

Transport costs also play a prominent role in models of urban economics and the related firm location within urban and metropolitan areas. In the classical monocentric city models of Alonso, Mills, and Muth firms are located at the city center close to central rail hubs and ports (Mieszkowski and Mills, 1993). These were the locations that minimized firms' transportation costs before the introduction of cars and trucks, as they were the most accessible locations from the point of view of the transportation of goods as well as people. Regarding the latter, spatial differences in commuting costs of workers are compensated through the housing market where land rents adjust as a function of distance to the center.

Transport cost reductions due to innovations and improvements in transportation can lead to urban employment decentralization. Historically, cheaper and better transportation such as the introduction of cars and trucks and the construction of highways have freed firms from central city locations. By not relying on the central transportation hub for the shipping of goods, firm could find it advantageous moving to suburban locations in search of lower land costs and labor cost savings. Thus, the reduction in transportation cost for goods can promote changes also in the urban spatial structure.

In addition, once firms are distributed within the urban area and no longer concentrated in the center, compensation for differences in commuting costs can also happen through wages where firms in locations that are difficult to access have to pay a wage premium for higher commuting costs. Wages thus constitute another source through which transport costs influence firm location.

Traditional location theory deals with the firm as a single unit, but the typical large modern firm is organized in multiple units, where, for example, production facilities and different business administration offices are often separated in space. This spatial separation of different activities of the firm has been facilitated by the developments in information and communication technology (ICT) that have allowed reducing the cost of information transmission. Standard production activities and tasks have been decentralized and moved to more peripheral locations to take advantage of lower land and labor costs, while management tasks that require the exchange of tacit knowledge and complex information tend to concentrate in central business districts where face-to-face contacts are facilitated. Thus, transport costs influence the location behavior of firms depending on their functional specialization.

Empirical Studies on the Role of Transport Costs for Firm Location

A firm's location is the result of a consideration of multiple factors. There is a very broad literature that has analyzed the importance of different location determinants, but only some of them have examined explicitly the role of transport costs for firm location.

Empirical approaches to study the role of transport costs for firm location have broadly followed two main approaches: (1) survey approaches, and (2) econometric studies using firm location data (Holl, 2006, 2007).

Survey approaches investigate the role of transport among other factors in firms' location decision from the viewpoint of the managers making the choice. Two types of survey approaches can be distinguished. First are those studies that inquire about the determinants behind the firms' actual location choice. Such surveys involve a series of questions relating to firm characteristics, the actual location choice, and different location factors, including how transport influenced the choice. Entrepreneurs are asked to assess the factors that determined their choice of location usually on a Likert scale. These studies show that transport cost consideration do influence firms' location choice. Although many firms perceive transport considerations as very important or at least as important in their location choice, there are also other factors that influence the location decision. For example, the literature has documented home bias in location choices that can be related to personal preferences of entrepreneurs and existing networks and relationships of trust and local knowledge that facilitate business relationships and that are costly to replicate outside the home location. Other relevant factors are usually local amenities and quality-of-life issues or access to skilled labor and labor costs. Findings of survey studies show that lower transport costs relax location constraints, allowing firms to examine a wider range of locations. Second, there have also been survey studies that have inquired about how specific transportation improvements have impacted on firms in different locations. Firms in the vicinity of new transportation improvements are often reported to have benefited in terms of reduced input ordering times, improved output delivery, and customer service.

Survey studies refer to a specific research context and provide information about entrepreneurs' perceptions about location factors. A different approach is to ask how transport costs make a location more or less attractive for firm location compared to other locations. In this regard, transport has been considered explicitly in a range of econometric firm location studies usually framed within the discrete choice approach. Four empirical issues distinguish the studies in this strand from each other: (1) how transport costs are measured, (2) which geographical level is the unit of analysis, (3) which type of firms, and (4) which transport infrastructure is analyzed.

The Measurement of Transport Cost

Measuring transport costs is challenging. The transport costs that influence firm location include both the transport cost for goods and people and both monetary and nonmonetary costs. Data on the actual transport costs incurred by firms are rarely available. Even if there is information on firms' direct transportation expenditure, for example, from balance sheet information, the full distance-related costs are not observable directly. Part of the full distance-related transportation costs that a firm faces in a given location are also embedded in input and output prices and part of it is reflected in the opportunity cost of time for goods as well as for people. Such data constraints mean that empirical studies in general have reverted to a range of proxies for transport costs and their spatial variations.

In the trade literature, a common approach is to use distances between trading partners to estimate a gravity equation in order to derive an approximation of transport costs from the parameter estimates, but these measures can also capture other factors beyond transport costs, that is, cultural and institutional differences. Nevertheless, even with additional controls, these studies indicate that most shipments occur over relatively short distances and flows decline relatively rapidly with increasing distance, reflecting the increasing transport costs. Some studies have also adjusted distances with the quality of infrastructure. This shows that trade flows not only decline with increasing distance, but that a deterioration of infrastructure raises transport costs as well and also reduces trade volumes. Fewer studies have used commodity flow data to calculate *ad valorem* transport costs, as the availability of such data is generally limited (as a notable example, see [Behrens et al., 2018](#)).

As a result of the difficulty to obtain consistent estimates for firms' generalized transport costs, most studies in the firm location literature have focused on the role of transport infrastructure. After all, firms' transport costs arise from the need of moving of inputs and outputs and from personal travel, and this requires transportation infrastructure. Transport costs are hence closely related to availability and quality of different transportation infrastructures. Nevertheless, a limitation of infrastructure-based measures is that they do not provide goods- or industry-specific variations ([Behrens et al., 2018](#)). Yet, the importance of transport costs will vary with the specific characteristics and values of the goods that need to be transported or services that need to be provided over distance.

Among firm location studies, some have used indicators of transport infrastructure endowment in the region where the firm is located such as kilometers of rail lines, kilometers of highways, and number of airports or ports. Other studies have proxied transport costs with different accessibility indicators. In practice, a wide range of accessibility measures have been used such as network access measure (i.e., distance to the nearest highway, railway, port, or airport), travel time measure, and potential accessibility measures or market potential measures. Accessibility indicators measure the relative position of the location within the transport network and take into account the network character of the infrastructure.

Geographical Level of Analysis

Regarding the spatial scale of firm location choices, they have been studied at different levels ranging from the international level (i.e., looking at foreign direct investment decisions), to the interregional level and the intraregional level. Related to the latter, there have been specifically a range of studies that have looked at urban and metropolitan areas and how transport costs affect firm location within those areas. Transport cost reductions show different impacts according to the spatial scale of analysis, that is, intraregional versus intraregional transport costs and transport costs within and between urban areas. The question of the spatial scale of analysis is of course not unrelated to the type of firm and the type of transportation infrastructure whose influence on firm location is studied, that is, air transport is rather taken into account in studies of international firm location, while subways or urban railways are studied naturally at the urban and metropolitan level.

Types of Firms

Transport costs depend on the specific firm's product or service characteristics and the production and distribution systems or organization that is implemented in the company. Clearly, some firms and consequently sectors will be more dependent on transportation and they will therefore also be more sensitive in their location choice to transport cost. Research for the United States has shown that vehicle-intensive industries experience greater productivity growth when roads are improved ([Fernald, 1999](#)). Thus, transport-intensive industries are also more likely to be influenced more strongly in their location choice by roads ([Holl and Mariotti, 2018](#); [Gibbons et al., 2019](#)).

Most research on the role of transport costs on firm location has focused on the manufacturing sector. However, there are also important differences within manufacturing. For example, some research has shown that durable goods producers react differently to transport infrastructure improvements than perishable goods producers ([Rothenberg, 2013](#)). High transport costs for perishable goods mean that they need to be produced close to consumers and firms are hence dispersed. In contrast, durable goods have lower

transport costs and their production is more concentrated. The evidence shows that with road improvements that further reduce transport costs, the latter firms disperse to locations with lower land and labor costs, but still in the vicinity of the large urban agglomerations. Along the product life cycle, the benefits from locating in urban locations versus more peripheral locations also change. In the early stages, the firm finds greater learning opportunities in more diversified urban areas. The firms' needs for complex information and knowledge transmitted through face-to-face contacts play a greater role. However, when the production process becomes standardized firms have an incentive to relocate to specialized cities. Thus, start-ups and relocating firms are attracted by different location characteristics and there is also some empirical evidence that manufacturing plants that relocate are more strongly attracted by the proximity to highways compared to start-ups (Holl, 2004a).

Fewer studies have looked at the service sector. However, the service sector is now the most important sector in developed economies. For those firms, the role of transport stems mainly from the need of personal travel. For corporate office and headquarter location choice, the cost, time, and reliability of transportation for executive travel is an important factor. Face-to-face communication with clients and other businesses are crucial and thus they tend to locate in central business districts. It is thus also likely that front office and back office services show different location patterns and are influenced differently by transport cost considerations, as the latter process standardized information and need less face-to-face contacts. However, to date there is very little empirical evidence beyond specific case studies on firm locations that distinguishes between front office and back office services and how transport costs influences differ. Nevertheless, the limited empirical evidence also indicates that back offices are attracted by the proximity to highways, while front office by city center locations with good accessibility. This reflects that different types of transport cost considerations come into play for the different location choices.

Types of Transport Infrastructure

Regarding the type of transportation infrastructure, roads are in general the most important type. Roads constitute the bulk of infrastructure capital and the backbone of transportation systems and the majority of passenger and freight transportation is carried on roads in most countries. It is thus not surprising that most firm location studies have also focused on roads and specifically on highways. These studies have shown that there is a tendency of firms to move closer to highways in order to benefit from improved accessibility and hence lower transport costs. Most studies on the economic impacts of road transport infrastructure investment have found positive growth effects with higher density of firms and greater economic activity in general. However, new roads not only contribute to economic growth but also, as shown by NEG, cause the reorganization of economic activity (Redding and Turner, 2015). While highways increase the attractiveness of locations for firm location in their vicinity, this can also come at the detriment of other areas. There is empirical evidence that shows that highways attract firms into their corridors but at the same time reduce the attractiveness of locations adjacent and beyond their corridors (Chandra and Thompson, 2000; Holl, 2004b; Gibbons et al., 2019). Within the urban and metropolitan context there are furthermore studies that have shown that new highways have led to suburbanization (Baum-Snow, 2007; Garcia-Lopez et al., 2015). Thus, highways can also negatively affect city center locations. It has not only been people but also certain types of firms that have been induced to move to urban fringes with road improvements that lowered transport costs: especially manufacturing jobs and standardized office jobs. Furthermore, there is also some limited evidence that new highway connections can hurt peripheral regions that are crossed by the new highways while leading to greater concentration of firms in core regions.

Rail accounts for a small fraction of goods movements in most countries. Still, rail can play a role for the location of firms that require bulky commodities that need to be transported over medium to long distances. In the urban context, there is some evidence that access to light rail influences office firm location and retail firm location and that the expansion of subway systems has also led to the dispersion of economic activity, similar to the findings on urban highways. In several countries, recent years have seen a shift from infrastructure investment in roads to high-speed rail (HSR). While HSR is usually not used for the transportation of goods, it can affect the location of service sector firms and particularly business service firms and also headquarter location. Most studies conclude that new HSR stations increase the attractiveness of cities but it also generates disadvantages for cities in-between that are not served. The evidence for intermediate stops in smaller cities is generally also positive in terms of attracting new firms.

Studies concerned with the impact of rail on firm location face the challenge that accesses to not only such infrastructure counts but also the level of service that is provided. The same challenge faces studies that are concerned with the impacts of ports or airports for firm location. Ports and airports play a potentially important role particularly for long distance and international transport. Access to these facilities influences global trade patterns. With the ongoing globalization process, fast and reliable long-distance transport has also become ever more important. In several time sensitive sectors, airfreight is part of logistics strategies and thus access to air terminals can be an important location factor for global business. Some studies have also documented the importance of airports, and in particular the availability of nonstop intercontinental flights, for headquarter location choice (Bel and Fageda, 2008). Regarding ports, there is some evidence that in countries where domestic transport costs are still very high, such as for example in African countries, distance to ports matters for city growth (Storeygard, 2016).

Trends and a Brief Look Forward

Transportation technology has changed profoundly over the course of history and each dominant form of transportation has had its own imprint on the spatial location and organization of economic activity in general and firms in particular (Redding and Turner,

2015). The introduction, for example, of railways, cars and trucks, containerization, or air transportation has contributed to substantial reductions in transportation cost over time. Recent decades have also witnessed important improvements in transportation. Moreover, developments in ICT have enabled economic relations over greater geographical distances.

This has led some authors to suggest that the role of transport cost for firm location is declining. But the role of transport could also have become more important. Reductions in transport costs together with the advances in ICT have not only facilitated economic relations over greater distances but also led to important changes in industrial organization. Modern production and distribution systems have become more transport dependent with a rise in “just-in-time” (JIT) production and distribution, vertical disintegration and outsourcing, and globalization (McCann and Shefer, 2004; Holl, 2006).

Regarding inventory holdings, for example, the optimal level of inventories held within the production plant depends on transport costs, but also the speed, and reliability of transportation. There is evidence that when transportation systems are improved and transport costs decline, firms are reducing traditional inventory holdings, moving toward JIT and mobile inventories. Thus, for firms there is a clear trade-off between transport cost and inventory costs. With the rising importance of the factor time, reductions in the unit cost of transport can actually lead to higher total transport costs and a rise in total logistics costs (McCann and Shefer, 2004). In this context, the rising importance of the time cost of transportation can make transport considerations more important for firm location decisions.

Transport costs have declined primarily as far as the movement of goods is concerned, but the transport costs for people remain important, especially in terms of the opportunity cost of time, not only for the commuting of workers but also importantly in service provision and for face-to-face meetings (Glaeser and Kohlhase, 2004). With rising economic development, the opportunity cost of time is also rising. Thus, it is important to consider a broad notion of transport cost that includes the full opportunity costs of time. Moreover, in modern knowledge economies, the moving of goods is becoming relatively less important compared to the transmission and exchange of knowledge and complex information. Despite of the developments in ICT this is still strongly dependent on personal face-to-face contact, especially when it comes to tacit knowledge and highly complex information. This means that the relative relevance of different spatial transactions is changing and a broader notion of transport cost should also incorporate the costs for the transmission of information and knowledge.

Present day technological developments in the field of transportation such as the transition to autonomous vehicles have the potential to reduce traveling and transportation costs because the time used in traveling can be used otherwise. For firms it can mean that the time spent of workers traveling can be used more productively thus reducing their transportation costs and consequently increasing the accessibility of a given firm location. For example, in the road transport sector, personnel expenditure—mainly the driver costs—can account of up to over 70% of the total cost, especially in high wage countries.

As with transport innovation in the past, new transport technologies will certainly also influence the location patterns of firms, but a priori it is not clear if new innovation in transportation will stimulate a more decentralized pattern of firm location or if they will further strengthen spatial concentration. This will depend not only on the changes in transport costs and times, but also on advances in ICT, and the need for face-to-face meetings for knowledge exchange. In any case, it is likely that different changes in location patterns happen at different geographical scales and in different types of firms and economic sectors.

Autonomous vehicles have furthermore the potential to free up urban space from parking for alternative developments and to change the spatial structure of cities by making locations at the urban periphery better accessible as well as urban locations not served by public transport services. This can make new locations attractive for firm location. Finally, transport costs and firm location will also be influenced by public policies regarding air pollution and advancements in energy transition and the development of alternative fuel vehicles.

Concluding Remarks

Firm location is an important topic. Attracting new firms is an important source of concern for policy-makers at the city level, the regional level, as well as the national level. The capacity to attract firms and to keep the existing ones is an important determinant of future growth and competitiveness. New transport technology and investment in transportation infrastructure reduce the transport cost of goods and people and change the relative location advantages and consequently lead to the spatial reorganization of firms. The available empirical evidence suggests that reductions of transport costs can lead to a more dispersed pattern of firm location within metropolitan areas but they can also lead to greater concentration of certain sectors in core regions and to greater regional specialization.

It is important to bear in mind that the effect of transport costs on firm location should not be considered in isolation. Firm location is determined in equilibrium also with land rents and wages and transport costs influence both and depend itself on input–output linkages and production and distribution decisions. Modern production and distribution technologies are increasingly time-sensitive and, in this new scenario, rather than transport costs per se, what matters more are total logistics costs. Moreover, in a modern knowledge economy, the weight is shifting from the cost to transport goods to the cost of acquisition and transmission of complex information and knowledge.

The demand for transport derives from the need of spatial interaction of economic agents. Any changes in the location of firms will also impact transport demand. Thus, a better understanding of how transport affects firm location is also essential for predicting future transport demand and travel flows. Understanding the relationships between transport and location of firms is hence critical for tackling the challenges for sustainable transport solutions, especially in urban and metropolitan areas.

References

- Baum-Snow, N., 2007. Did highways cause suburbanization? *Q. J. Econ.* 122 (2), 775–805.
- Behrens, K., Bougna, T., Brown, M.W., 2018. The world is not yet flat: transport costs matter! *Rev. Econ. Stat.* 100 (4), 712–724.
- Bel, G., Fageda, X., 2008. Getting there fast: globalization, intercontinental flights and location of headquarters. *J. Econ. Geogr.* 8 (4), 471–495.
- Chandra, A., Thompson, E., 2000. Does public infrastructure affect economic activity? Evidence from the rural interstate highway system. *Reg. Sci. Urban Econ.* 30 (4), 457–490.
- Combes, P.-Ph., Mayer, T., Thisse, J.-F., 2008. *Economic Geography: The Integration of Regions and Nations*. Princeton University Press, Princeton, NJ.
- Fernald, J.G., 1999. Roads to prosperity? Assessing the link between public capital and productivity. *Am. Econ. Rev.* 89 (3), 619–638.
- García-López, M.-Á., Holl, A., Viladecans-Marsal, E., 2015. Suburbanization and highways in Spain when the Romans and the Bourbons still shape its cities. *J. Urban Econ.* 85, 52–67.
- Gibbons, S., Lyytikäinen, T., Overman, H., Sanchis-Guarner, R., 2019. New road infrastructure: the effects on firms. *J. Urban Econ.* 110, 35–50.
- Glaeser, E.L., Kohlhase, J.E., 2004. Cities, regions and the decline of transport costs. *Pap. Reg. Sci.* 83, 197–228.
- Holl, A., 2004a. Start-ups and relocations: manufacturing plant location in Portugal. *Pap. Reg. Sci.* 83 (4), 649–668.
- Holl, A., 2004b. Manufacturing location and impacts of road transport infrastructure: empirical evidence from Spain. *Reg. Sci. Urban Econ.* 34 (3), 341–363.
- Holl, A., 2006. A review of the firm-level role of transport infrastructure with implications for transport project evaluation. *J. Plann. Lit.* 21 (1), 3–14.
- Holl, A., 2007. Transport network development and the location of economic activity. In: Coto-Millán, P., Inglada, V. (Eds.), *Essays on Transport Economics*. Physica-Verlag, Heidelberg, pp. 341–362.
- Holl, A., Mariotti, I., 2018. The geography of logistics firm location: the role of accessibility. *Net. Spat. Econ.* 18 (2), 337–361.
- Krugman, P., 1991. *Geography and Trade*. MIT Press, Cambridge, MA.
- Lösch, A., 1954. *The Economics of Location*. Yale University Press, New Haven, CT.
- McCann, P., Shefer, D., 2004. Location, agglomeration and infrastructure. *Pap. Reg. Sci.* 83 (1), 177–196.
- Mieszkowski, P., Mills, E.S., 1993. The causes of metropolitan suburbanization. *J. Econ. Perspect.* 7 (3), 135–147.
- Redding, S.J., Turner, M.A., 2015. Transportation costs and the spatial organization of economic activity. In: Duranton, G., Henderson, V., Strange, W. (Eds.), *Handbook of Urban and Regional Economics*, Vol. 5. Elsevier, Amsterdam.
- Rothenberg, A., 2013. Transport Infrastructure and Firm Location Choice in Equilibrium: Evidence from Indonesia's Highways. Working Paper, RAND.
- Storeygard, A., 2016. Farther on down the road: transport costs, trade and urban growth in sub-Saharan Africa. *Rev. Econ. Studies* 83 (3), 1263–1295.
- Von Thünen, J.H., 1826. *Der Isolierte Staat in Beziehung auf Landwirtschaftslehre und Nationalökonomie*, Hamburg, English translation. Pergamon Press, Oxford 1966.
- Weber, A., 1929. *Theory of the Location of Industry*. Chicago University Press, Chicago.

Commuting, the Labor Market, and Wages

Jan Rouwendal^{*,†}, Ismir Mulalic[‡], ^{*}Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; [†]Tinbergen Institute, Amsterdam, The Netherlands; [‡]Copenhagen Business School, Copenhagen, Denmark

© 2021 Elsevier Ltd. All rights reserved.

Introduction	297
Insights From the Monocentric Model	297
The Muth Condition	297
Transportation Cost	298
Heterogeneous Workers	298
Decentralized Employment and Excess Commuting	298
Efficient Commuting Patterns	298
Insights From Job Search Theory	299
Is Excess Commuting Productive?	299
Gravity Equations and Discrete Choice	299
The Logit Model	299
Amenities	300
Commuting and Local Labor Markets	300
Shocks	300
Wages and Commutes	300
Conclusion	301
References	301
Further Reading	301

Introduction

Commuting implies the need to spend time that could have been used in a more productive way traveling from home to work and vice versa. It is therefore tempting to regard it primarily as a burden for which workers require compensation. This aspect is emphasized in the monocentric model, which focuses on the connection between commuting and housing consumption around employment centers. The labor market implied by this model is an extremely simple one, and more realistic models in which labor and housing markets are imperfect and jobs and workers heterogeneous offer important complementary insights into the value of separating employment and residential locations in dense urban areas. Lower commuting costs provide more space for exploiting agglomeration benefits while offering workers the possibility to combine working at highly productive locations with living in attractive residential areas. Dense urban labor markets allow for assortative matching of workers and jobs, resulting in higher productivity and wages. With overlapping local labor markets the immediate response to labor demand shocks may be an adjustment in the incoming and outgoing commuting flows through which the shock expands over space. In many respects, commuting thus plays a useful role in the functioning of urban and regional labor markets. This conclusion is entirely consistent with the fact that workers dislike long commutes and require significant compensation for accepting them.

Insights From the Monocentric Model

The Muth Condition

The monocentric model is the workhorse of urban economic analysis. It describes the location and housing choices of workers employed in a single center located on a featureless plain. The central issue of the model is that housing requires land, while workers dislike commuting. All workers would therefore like to live close to the employment center, but there is clearly not enough space to realize this. It is therefore inevitable that for most workers residential and work locations are spatially separated. The model elaborates on this issue, and therefore on the determination of commutes.

The classic treatment of worker location choice in the monocentric model assumes that workers maximize a utility function, $u(h, c)$ with h denoting housing (lot size, or housing services) and c other consumption, subject to a budget constraint: $p(x)h + c = \gamma - tx$, where $p(x)$ denotes the price of housing at location x , γ is income and t commuting cost per unit of distance.^a Location x is measured as the distance to the CBD. Since the employment center is—by assumption—located on a featureless plain, this distance is the only thing that matters to the worker. This is just the standard two good model of elementary microeconomics with two elements added: the price of one commodity, housing, depends on its location, as does the net income $\gamma - tx$ of the worker.

^a Note that the price of other consumption has been scaled at 1.

Note that in this simple version of the model time costs of commuting is not made explicit and commuting time does not directly affect utility.

The choice variables are the amounts consumed of housing and other goods and the location. First order conditions with respect to the first two variables are standard, but the third one—which is known as the [Muth \(1969\)](#) condition, is specific for this model:

$$\frac{\partial p}{\partial x} = -\frac{t}{h(x)}$$

With a homogeneous population of workers, this condition determines the curvature of the price of housing as a function of the distance to the center. It suggests a strong relationship between house prices on the one hand and housing consumption and commuting costs on the other. More specifically, commuting cost is a determining factor of the spatial pattern of house prices and of the density of housing. With high commuting costs, house prices decline sharply with distance from the employment center. Since house prices determine land prices, this implies that land prices in the employment center will be a steeply increasing function of the number of jobs in the center. This may impose limits on the size of such centers. With low commuting costs it is easier to give a large number of workers access to decent living space without land prices in the center getting extremely large.

Transportation Cost

Muth's condition suggests that high transportation costs impose important limits on the size of urban areas. Lowering these costs mitigates this constraint and can lead to larger cities with more housing consumption. Of course, commutes are also longer in such areas, but the important point is that the burden of commuting is, nevertheless, eased. There is indeed abundant evidence that lower transportation costs lead to expanding urban areas. For instance, suburbanization has been facilitated by the Philadelphia street car ([LeRoy and Sonstelie, 1983](#)), the London underground ([Heblich et al., 2018](#)) and 20th century highway networks ([Baum-Snow, 2007](#)). Recently the point has also been emphasized that lower commuting costs provide better possibilities for separating employment and work locations, which offers better possibilities for the realization of agglomeration economies. After the introduction of the London underground the center specialized rapidly in production activities, while workers choose to live in the quieter and less expensive outskirts of the city. At the same time, the lower population density may imply a loss of agglomeration benefits in consumption.

Heterogeneous Workers

With heterogeneous workers the Muth condition describes the slope of the bid rent function, while the actual house price function is the envelope of these bid functions. The highest bidder uses the land and the model this predicts separation of groups of workers. In this generalized monocentric model the slope of the bid rent functions is the driver of the location of different groups within the urban area. The simple rule is that the group with the steeper bid rent curve lives closer to the employment center.

This slope is the ratio of the transportation cost and housing consumption. To see the implications of the model, we turn to a more general setup than the one used above and take into account the time cost of commuting. In fact, transportation costs are dominated by time costs, and the value of time is usually thought to be proportional to the wage, which implies that it is approximately proportional to income. As the income elasticity of housing consumption is often found to be between 0 and 1, this suggests that the bid rent curves of households with higher incomes will be steeper. The implication that higher income groups live closest to employment centers was for a long time at odds with 20th century reality in the US cities, which was somewhat discomfiting for the theory. In the 19th century the United States and in many European cities also in the 20th century many inhabitants of central city residential areas had high incomes. More recently, the popularity of (central) urban living among high-income households has been increasing worldwide. Apart from the good accessibility of employment, appreciation of amenities like historical centers, the presence of theatres and a variety of restaurants appear to play a role.^b A recent study for Denmark confirms that the income elasticity of commuting distances is negative ([Gutiérrez-i-Puigarnau et al., 2016](#)).

Decentralized Employment and Excess Commuting

Efficient Commuting Patterns

The monocentric model explains important aspects of cities that are in fact far from monocentric. Somewhat paradoxically, it is unable to explain the commuting patterns in actual cities. The logic of the model discussed in the previous section implies that with multiple employment centers, the urban area splits up into a number of labor market areas. That is, for each center there is a catchment area that is spatially separated from those of the other centers.^c If there is dispersed employment outside the centers, the logic of the model implies that the jobs outside these centers are filled by workers commuting in the direction of the dominant employment center, while wages for the dispersed job are such that these workers are indifferent between their actual job and one

^b We will return to amenities below.

^c The situation is perfectly analogous to the Christaller model in which each central place has its own market area and market areas do not overlap.

located in the nearby employment center. The implied star-like commuting pattern can be shown to be efficient in that it minimizes to total commuting distance traveled.

Actual commuting patterns differ significantly from this prediction. It was found that at a small geographical scale random allocation of workers to jobs provided a better approximation to reality than the efficient commuting pattern. Although these findings were initially considered to be alarming with respect to the validity of the logic underlying the monocentric model, it has become apparent that modifications of the model allow for a reconciliation of theory and empirics.

Insights From Job Search Theory

The labor market described by the monocentric model is an extremely simplified one. It assumes perfect information and perfectly homogeneous labor. It has been pointed out that the realization of the efficient commuting pattern requires that every mutually beneficial swap of jobs should be realized. That is, if there exists a pair of workers in the city who can both realize a shorter commute by switching jobs, they should do so. If not, excess commuting remains present. However, even if we look at homogeneous groups of workers, it seems highly unlikely that all such swaps will be realized. The way these workers interact with colleagues and their specific experiences in the past causes heterogeneity that prohibits the realization of such swaps. Even if the workers find them attractive, their employers may disagree (and vice versa).

The economic model of job search that has been developed since the 1970s can be placed in a spatial setting to study the consequences of limited availability of vacancies for unemployed workers for the acceptance of commutes. The “reservation wage” property of the optimal search strategy then translates into a “reservation commute” property that implies that a wide range of commutes can be accepted, depending on the arrival rate of job offers (Rouwendal, 1998). Although the implied commuting patterns are inefficient in the sense that they do not minimize total commuting distances, they result—in such models—from an optimal trade-off between accepting an available job and waiting until a better one is offered. Hence the excess commuting is a consequence of more fundamental imperfections of labor markets.

The discussion above refers only to the labor market, while adjustments can also take place via the housing market. However, it is clear that similar problems are present there. Consider two families of which the two workers can realize shorter commutes if they “swap” their houses. In many cases one, and probably both, will be unwilling to realize this exchange even if the housing market would have no imperfections. In fact, it shows similar imperfections as the labor market possesses. We may conclude that seemingly excessive commuting is at least partly caused by idiosyncratic preferences for particular jobs or houses and plays a useful role in mitigating the imperfections of the housing and labor markets.

Is Excess Commuting Productive?

Excess commuting can, from an alternative point of view, be interpreted as an indicator of the density and specialization of workers and jobs in local labor markets. Workers that reside in a large metropolitan area with many employment opportunities have the possibility to accept many jobs from a given residential location, whereas those living in an isolated town have little choice. The relevance of housing and labor market imperfections thus suggests that the presence of close-to-random matching of workers to jobs as far as the residential and employment locations are concerned may in fact signal the existence of a well-functioning labor market in which workers can easily switch from one job to another while avoiding the substantial cost of residential mobility. A measure of labor market density that can also be regarded as an indicator of wasteful commuting has indeed been shown to have substantial explanatory power in wage regressions for the United States (Gautier and Teulings, 2003). Idiosyncratic differences in the quality of matches between workers and jobs can more easily be exploited in dense urban labor markets and may be incorrectly interpreted as excessive. In line with this, it has recently been shown that in larger urban areas assortative matching between highly productive workers and jobs is an important source of higher wages in German cities.

Gravity Equations and Discrete Choice

The Logit Model

Physical analogies have been popular among economic geographers and the gravity equation has been found especially relevant as a model for spatial interaction. Commuting is no exception and a typical formulation is:

$$C_{ij} = R_i E_j f(d_{ij})$$

Here C_{ij} denotes the number of workers commuting from i to j , R_i the number of residents in i , and E_j the number of jobs in j , while f is a decreasing function of the distance d_{ij} between these two locations. Popular specifications of f are the power and exponential functions. An economic interpretation of this model can be found by assuming that a worker’s utility over combinations of residential and work locations is an additive function:^d

$$u_{ij} = a_i + b_j + c(d_{ij}) + \varepsilon_{ij}$$

^d An alternative, but equivalent, derivation starts from a utility function that is multiplicative.

where a_i denotes the utility of living in i , b_j , that of working in j and $c(d_{ij})$ indicates the quality of the match, which is typically a function of the travel distance d_{ij} , while ε_{ij} denotes the idiosyncratic utility that a worker derives from the combination. If it is assumed that the ε_{ij} are independent and identical extreme value type I distributed, utility maximizing behavior leads to the logit model. That is, the probability that the combination of residential location i and employment location j will be chosen equals:

$$\pi_{ij} = \frac{e^{a_i + b_j + c(d_{ij})}}{\sum_k \sum_l e^{a_k + b_l + c(d_{kl})}}$$

where the index k runs over all residential locations and l over all work locations. With alternative specific constants incorporated in the utilities of the residential and work locations, maximum likelihood estimation of the model guarantees $\sum_j \pi_{ij} = \frac{R_i}{\sum_k R_k}$ and $\sum_i \pi_{ij} = \frac{E_j}{\sum_l E_l}$.

The logit model has been heavily criticized for its independence of irrelevant alternatives, which is closely related to the assumed distribution of the idiosyncratic utilities. However, it has received new popularity because of its relationship to the gravity equation. Nevertheless, it seems plausible that the ε_{ij} are correlated. For instance, if a worker has a strong idiosyncratic preference for living in a particular neighborhood, this will be reflected in all commutes that have this zone as the residence. However, these strong a priori arguments notwithstanding, the multinomial logit model is hard to beat because with a full set of origin- and destination-specific constants its generalizations like nested and mixed logit are not identified.

Amenities

The gravity equation and the related discrete choice models can provide good approximations to actual commuting patterns. They can also easily incorporate amenities at the residential and work locations. That is, the utilities of those locations need not be determined completely by the wage and the housing price, but can also be affected by the presence of shops and restaurants, nice architecture, etc. For instance the model can explain “reverse commuting:” workers living in the center because of the attractive urban amenities, while working in the suburbs, where more suitable jobs are available.

Although the model has a prominent place for commuting costs, it is also consistent with the view that separation between employment and residential locations is a potential benefit, because employment locations are not necessarily the best places to live. Easy commutes make it possible to combine attractive residential locations with the most suitable jobs, which can improve welfare. For instance, it has been argued that the decrease in commuting costs occurring in the United States between 1990 and 2010 implied an increase in welfare of around 3.3%, which is of the same order of magnitude as opening the closed economy of a medium-sized country to international trade (Monte et al., 2018).

Commuting and Local Labor Markets

Shocks

A sudden increase in local labor demand is difficult to handle with an isolated local labor market since the housing stock is difficult to adjust in the short run. However, if local labor markets are overlapping the additional workers may be attracted from neighboring locations and the shock may be much easier to absorb. It has indeed been shown that the substantial diversity in the reactions of local labor markets to a positive shock in demand is related to the possibility to increase the number of incoming commuters. The local shock thus causes a ripple effect as the shortage of labor in one location is shifted to contiguous labor markets. Commuting thus plays an important role in local labor market adjustments (Manning and Petrongolo, 2017).^e

Overlapping local labor markets have been shown to be consistent with strong resistance against long commutes and effects of place based policies that reach much further than the average commute. The reason is a ripple effect. A shortage of labor expands over space as workers from nearby locations are pulled toward the location with high demand.

The resistance against accepting long commutes is strong. Although—what at first sight was thought to be—excess commuting suggested that workers do not really care about commuting, empirical investigations almost invariably show the contrary: workers need substantial compensation for accepting longer home-to-work distances.

Wages and Commutes

In the frictionless models, such as a standard monocentric city model, the wage is equal to the marginal productivity at the workplace location but does not depend on the length of the commute of the worker. In such markets we would not expect firms to (partially) compensate the commuting costs of their employees.

In imperfect labor markets with job search frictions,^f workers, and employers bargain about the wage conditional on the commuting distance and the surplus of the match is shared between workers and employers. Wages will be higher for workers

^e Of course, long run adjustments may be more in line with Rosen-Roback models in which the housing market adjusts so as to accommodate the larger number of workers at shorter average commutes.

^f Labor market characterized by incomplete information about job offers.

with a longer commute because a long distance makes the job match less attractive to workers. Moreover, if firms enjoy some monopsony power in the labor market, which allows them to pay wages below marginal productivity, they will compensate workers for longer commutes. Hence the difference between wage and marginal productivity will be largest for workers living next to the firm. The main reason is that opportunity costs of staying with the firm for workers who live close to the firm are less than those with longer commutes.

There is recent quasi-experimental evidence that workers who face longer commutes because of firm relocations receive a wage increase. More specifically, a one percent increase in commuting distance induces a wage increase of about 0.02%. The implied hourly wage compensation for an additional hour of commuting is of the order of 15%–20% of the net hourly wage (Mulalic et al., 2014).

This and related findings give second thoughts about the appropriateness of the standard assumption that employers do not compensate workers for commuting costs. The empirical evidence just referred to suggests that employers may at least partly compensate their workers for increasing commuting costs due, for instance, to the introduction of road pricing or other policy measures.⁸

Conclusion

Commuting is still a burden for the many workers who experience traffic congestion on each working day. But it also implies the significant advantages associated with separating residential and work locations. Agglomeration benefits in production and consumption can be better exploited in this way. Moreover, dense labor markets offer workers access to larger numbers of jobs from a given residential location and firms to a large number of workers for their jobs, without having to relocate. On top of that, commuting can play an important role in labor market adjustments following local shocks. All this implies that commutes play an important role in the functioning of local and regional labor markets. These markets should be regarded as imperfect with respect to the availability of information about matching opportunities and competitiveness. The seemingly random commuting patterns observed at small geographical scales do not prove that commuting is wasteful, but result from the fact that commuting choices do much more than minimize the distance between home and work. Commuting costs are high and the social benefits of reducing them through improvements in transportation infrastructure remain large.

References

- Baum-Snow, N., 2007. Did highways cause suburbanization? *Quarterly J. Econ.* 122, 775–805.
 Gautier, P., Teulings, C., 2003. An empirical index for labor market density. *Rev. Econ. Statist.* 85, 901–908.
 Gutiérrez-i-Puigarnau, E., Mulalic, I., van Ommeren, J.N., 2016. Do rich households live farther away from their workplaces? *J. Econ. Geog.* 16, 177–201.
 Heblich, S., Redding, S., Sturm, D., 2018. The making of modern metropolis: evidence from London. NBER working paper. Available from: <https://www.nber.org/papers/w25047>.
 LeRoy, S., Sonstelie, J., 1983. Paradise lost and regained: transportation, innovation and residential location. *J. Urban Econ.* 13, 67–89.
 Manning, A., Petrongolo, B., 2017. How local are labor markets? Evidence from a spatial job search model. *Am. Econ. Rev.* 107, 2877–2907.
 Monte, F., Redding, S.J., Rossi-Hansberg, E., 2018. Commuting, migration and local employment elasticities. *Am. Econ. Rev.* 108, 3855–3890.
 Mulalic, I., van Ommeren, J.N., Pilegaard, N., 2014. Wages and commuting: quasi-natural experiments' evidence from firms that relocate. *Econ. J.* 124, 1086–1105.
 Muth, R.F., 1969. *Cities and Housing: The Spatial Pattern of Urban Residential Land Use*. University of Chicago Press, Chicago.
 Rouwendal, J., 1998. Search theory, spatial labor markets, and commuting. *J. Urban Econ.* 43, 1–22.

Further Reading

- Brueckner, J.K., 1987. The structure of urban equilibria: a unified treatment of the Muth-Mills model. In: Mills, E.S. (Ed.), *Handbook of Regional and Urban Economics*, Vol. II. Elsevier, Oxford.
 Dauth, W., Findeisen, S., Moretti, E., Suedekum, J., 2019. Matching in cities. IZA discussion paper. Available from: <https://www.iza.org/publications/dp/12278/matching-in-cities>.
 Manning, A., 2003. The real thin theory: monopsony in modern labor markets. *Labor Econ.* 10, 105–131.
 Rouwendal, J., 1999. Spatial job search and commuting distances. *Reg. Sci. Urban Econ.* 29, 491–517.
 Small, K.A., Song, S., 1992. "Wasteful" commuting: a resolution. *J. Polit. Econ.* 100, 888–898.

⁸ However, note that road pricing may also decrease commuting costs, notably for workers with a high value of travel time, by reducing congestion.

How to Buy Transport Infrastructure

Johan Nyström, The Swedish National Road and Transport Research Institute (VTI), Stockholm, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	302
Three Contracting Forms to Buy Transport Infrastructure	303
The Traditional Way of Building Transport Infrastructure—DBB	303
A Refined Model DBB Contracts	304
Cost Overruns in Infrastructure Project	304
Strategically Set Prices	305
The Push for DB Contracting to Enhance Productivity	306
PPP is Not a Way to Raise Funding	306
Pros and Cons of Contracting Forms to Buy Transport Infrastructure	307
Summary	307
Acknowledgment	307
References	307
Further Reading	307

Introduction

Transport infrastructure could in principle be provided by private companies charging users when using the system. However, transport infrastructure is generally owned by the state. This is justified by the transport infrastructure having elements of natural monopoly, that is, there are economies of scale and scope in having the state providing this service. Monopoly pricing can be a significant burden on efficiency.

The setup with a public client entails a decision to build and maintain the infrastructure in-house using publicly employed staff or to contract out the work to private companies (Shleifer, 1998). This can be modeled as a public *make-or-buy* decision.^a There are pros and cons to both alternatives, but over time most western countries have come to contract out transport infrastructure construction and maintenance.

Given that the public client chooses to procure the work, the question is how to design such a contract. Most countries have some version of a public procurement law, stipulating that the public client should accept the lowest price or the most economically advantageous tender.

The *raison d'être* of public procurement laws are to secure efficient spending of tax money and deter nepotism. Although desirable features, the legislation reduces the public client's possibilities to procure products. Private clients are free to choose any contractor they want without having to justify it. This includes using a self-enforcing and infinite contract, meaning that a contractor doing a good job can also be awarded the next job without any external restriction on how the second contract was won. This creates an incentive for the contractor to do a good job (Gibbons, 2005). The self-enforcing contract cannot be copied in full by the public client, as each contract needs to be procured in a transparent matter. Without the option of signing self-enforcing contracts, the contract design for the public sector becomes more complex.

The challenge when tendering transport infrastructure is how to design contracts that minimize cost subject to a prespecified dimensions of quality. One could in principle maximize quality subject to a budget constraint, but that is not how it usually works in reality, as quality is hard to specify and monitor.

This paper concerns the buying and construction of transport infrastructure project, that is, a cost-efficiency perspective. Socioeconomic efficiency is an issue for the planning stage before the procurement. The pros and cons of the three most common contracting forms will be described.

Three Contracting Forms to Buy Transport Infrastructure

Over time, three types of contracts have been developed for buying transport infrastructure in the public sector. The taxonomy differs between countries, but a common way of describing the three contracting forms is design-bid-build (DBB), design-build (DB), and public-private partnerships (PPP).

^aThe make-or-buy decision refers to the theory of the firm, granted with two Nobel prizes—Coase (1991) and Williamson (2009), where the size of the firm is decided by using the market to buy inputs or make the product in-house.

In a DBB contract, the client is responsible for the design and the contractor for the construction. If a bridge collapses because of an under-dimensioned pillar in the design, the client is accountable. However, if it breaks down due to the contractor not building in accordance to the design, it is the contractor's responsibility.

The DB contract makes the contractor responsible for both design and construction. Instead of giving specific instructions on how to build, the client expresses the transport infrastructure project in terms of functional requirements (e.g., travel times between cities A and B) and then the contractor comes up with the design and undertakes the construction.

PPP can be described as a version of the DB contract, but where the contractor apart from construction also will be responsible for funding and maintenance. As the contract includes maintenance, the contract duration is typically longer. Payment is received after the completion of the investment, entailing that the contractor needs to fund the project.

The Traditional Way of Building Transport Infrastructure—DBB

The DBB is the traditional and still most common contracting form in the construction industry. It is procured with the client providing a bill of quantities. This is a list of the tasks required to build the infrastructure and their respective quantities, for example, square meters of asphalt, cubic meters of gravel, number of signs, etc. These quantities are connected to technical descriptions and manuals. Contractors then submit bids in the form of price vectors, one unit price for each quantity. The lowest vector product of prices and quantities, that is, the lowest total price, is awarded the contract and takes legal responsibility for the construction. The payment within a DBB can be fixed, cost-plus, or incentive scheme with a target cost.

The fixed price contract stipulates that the client pays the contractor a predetermined price regardless of the final project cost. The price p coincides with lowest tender b , as:

$$p_i = b_i, \text{ where } b_i = \min\{B_1, B_2, \dots, B_n\}, \quad (1)$$

where B_{1-n} are the independent bids. In this setup, the contractor has a strong incentive to cut costs in order to maximize profit. Under the assumption of ex-post information asymmetry, this can entail shirking on quality. However, the difference between the price and the actual cost may also be positive. Both deviations from the price are carried by the contractor who bears all risk under a fixed price contract.

On the other side of the spectrum, there is the cost-plus contract with the client bearing all the risk. The price, p , that the contractor is paid coincides with his cost, c , as:

$$p_1 = c_1, \text{ for Contractor 1} \quad (2)$$

Here, the incentive for the contractor to hold back costs is weak but there is no reason to shirk on quality.

The third contract is a hybrid of the earlier versions and is called an incentive or a profit-sharing contract. It establishes a target price b obtained from the lowest tender. Any deviations depending on the winning contractor's costs in c is split between the parties by a predetermined weighting parameter α , as:

$$p_i = b + \alpha(c_i - b) \quad (3)$$

The weighting parameter, α , is a matter of negotiation, depending on both the risk aversion of the parties and whether the client favors high quality or low cost in terms of contractor incentives. This contract requires open accounting and therefore more transaction costs.

As the contractor carries risk in the fixed price and incentive contracts, these are expected to have a higher price due to a risk premium.

How the contracting forms affect ex-post cost is, however, not that obvious. Two opposing mechanisms can be distinguished. In theory, cost-plus contract's lack of incentive to hold back cost during the contract could outweigh the risk premium of the fixed price contracts. Empirical studies comparing outcome of the payment schemes are lacking.

A Refined Model DBB Contracts

An important feature of applied research, such as transport economics, is that the results need a link to real-life situations and stakeholders' decisions. A problem with the standard descriptions of the DBB contract as described earlier [Eqs. (1–3)] is that there are no such clear-cut versions in reality. Everyday contracting is much more complex. The real design of the payment scheme in a DBB is the unit price contract (UPC).

A UPC consists of unit prices from the contractor and adjustable (q_i) and nonadjustable ($\bar{q}$) quantities, which are specified in ex ante in the tendering document by the client. This gives a new cost function, as follows:

$$p = b = c = \sum_{i=1}^n ap_i q_i + \sum_{j=1}^m ap_j \bar{q}_j \quad (4)$$

where $i = 1, \dots, n$ indicates adjustable quantities (q_i) and $j = 1, \dots, m$ indicates nonadjustable quantities ($\bar{q}$) and their unit price vectors ap_i and ap_j .

Applying this more detailed cost function [Eq. (4)] into the cost-plus contracts [Eq. (2)] gives:

$$p = \sum_{i=1}^n a p_i q_i + \sum_{j=1}^m a p_j \bar{q}_j \quad (5)$$

The second part of Eq. (5), regarding $(\bar{q})$, adds a fixed element to the cost-plus contract by pushing over risk to the contractor.

The introduction of adjustable quantities also changes the fixed price contracts [Eq. (1)] as the client will take some risk. Formally this can be described by b no longer being fixed ex post as:

$$p = b = c = \sum_{i=1}^n a p_i q_i + \sum_{j=1}^m a p_j \bar{q}_j \quad (6)$$

Hence, by introducing adjustable and nonadjustable quantities, the fixed-price and the cost-plus contracts turn into the same specification, as follows:

$$p = b = c = \sum_{i=1}^n a p_i q_i + \sum_{j=1}^m a p_j \bar{q}_j \quad (7)$$

Eq. (7) is a combination of [Eqs. (1) and (2)], with the adjustable quantities representing the cost-plus contract and the nonadjustable quantities representing the fixed price. A UPC with only fixed quantities or only adjustable quantities would collapse [Eq. (7)] to a fixed price [Eq. (1)] and cost-plus [Eq. (2)] contract, respectively.

Including adjustable and nonadjustable quantities to these models improves the connection to reality and enables a better understanding of how these contracts work in order to make better projections of outcome. The extended model opens up for analysis concerning problems discussed in the industry regarding, for example, the problem with cost overruns in transport infrastructure projects.

Cost Overruns in Infrastructure Project

In a world of costless and complete information with no uncertainty, designing a contract would be easy as every eventuality could be specified in detail. However, full information is not a fair assumption and complete contracts cannot be written. This is because (1) all contingencies cannot be foreseen, and, even if they could, it would be (2) infinitely expensive to write all of them down. Even if both (1) and (2) would be fulfilled, then (3) the language is not clear enough to describe everything in such a way that there would be no problems of interpretation and enforcement.

Hence, in a world of incomplete contracts, problems occur when new information arrives. One well-documented problem when building transport infrastructure is cost overruns. Infrastructure projects tend to swell both in size and cost after the political decision to build has been taken. This is due to the somewhat trivial but theoretical important explanation that complete contracts do not exist in reality.

There are three hypotheses to explain cost overruns (Flyvbjerg, 2009):

1. Psychological explanations—indicating an optimism bias from the people working with the project during the planning stage.
2. Political-economic explanations—indicating a deliberate push for underestimating negative and overstating positive aspects of the project.
3. Technical explanations—indicating that some technical aspects such as type of geology or rock type were misestimated.

There are two categories of cost overruns. The first refers to a deviation between the final cost and the estimated cost when the political decision to build was made. The second type compares the procured price and the final cost. Although most transport infrastructure projects exceed both planned and procured cost, there are examples of project going under budget and ex-ante price.

All of the hypotheses earlier concern the planning stage of a transport infrastructure project and the first type of cost overrun. Focusing on second type cost overrun, the contract design has to be analyzed which originates in technical explanations (3). Building on technical explanations, the causes for cost overruns can be further developed into the following three different aspects:

- 3a. Misestimations of stipulated quantities in the tendering document. This relates to deviations in Eq. (7) regarding q .
- 3b. Ex-post adaptation, where the cost will grow due to quantities that was not included in ex-ante tendering documents.
- 3c. Strategic pricing from the contractor making use of misestimations, either of type (3a) or (3b).

The former two explanations are due to client mistakes, but the latter concerns deliberate strategic behavior from the contractor based on superior information.

Cost overruns because of the client (3a–b) are solved by more planning and better designs. In order to mitigate cost overruns due to contractors setting strategic prices (3c), the contract design needs to be analyzed.

Strategically Set Prices

An example of strategically set prices in UPCs is unbalanced bidding (Ewerhart and Fieseler, 2003). This pricing strategy is done by raising unit prices on underestimated quantities and vice versa to compensate for a competitive bid. The contractor will attain rents due to asymmetric information and the skewing of unit prices, for example, by the following example.

Table 1 Ex-ante bill of quantities and bids

<i>Ex ante</i>	<i>Bill of quantities</i>	<i>Contractor 1's bid (uninformed)</i>	<i>Contractor 2's bid (informed)</i>
Provision of gravel	100 m ³	10	12
Pavement	150 m ³	10	8.5
Total bid		2500	2475

Table 2 Ex-post quantities and revenue

<i>Ex post</i>	<i>Actual of quantities</i>	<i>Contractor 1's bid (uninformed)</i>	<i>Contractor 2's bid (informed)</i>
Provision of gravel	110 m ³	10	12
Pavement	145 m ³	10	8.5
Final cost for the client		2550	2553

Table 3 Contractor 2 being risk-neutral

<i>Ex ante</i>	<i>Actual of quantities</i>	<i>Contractor 1's bid (uninformed)</i>	<i>Contractor 2's bid (informed)</i>
Provision of gravel	110 m ³	10	25
Pavement	145 m ³	10	0
Final cost for the client		2500	2750

Assume that there are two inputs to building a road, provision of gravel and pavement. The ex-ante bill of quantities for the project estimates 100 m³ of gravel and 150 m² of pavement. Assume further that the contractors differ in costs and information, where Contractor 2 has a higher marginal cost on both inputs in comparison to Contractor 1. However, Contractor 2 also has private information on the amount of gravel and pavement, which Contractor 1 does not. Contractor 1 bids her marginal cost at unit prices of 10. Contractor 2 can use the superior information regarding the ex-post quantities and skew unit prices accordingly. As depicted in **Table 1**, Contractor 2 submits the lower total bid and wins the contract.

The project starts and Contractor 2's prediction that the quantities of gravel will increase and pavement will decrease turns out to be correct. As seen in **Table 2**, Contractor 2's skewing of prices, based on her expectation of changing quantities, enables her to win the contract and earn higher revenue.

Due to unbalanced bidding, the most efficient contractor with the lowest marginal cost will, in this example, not win the contract and the client ends up paying an information rent to Contractor 2.

However, assuming that the contractor is risk neutral, the optimal way of skewing the bid is to hand in zero-unit prices expect on the most underestimated quantity, as seen in **Table 3**.

Such bidding behavior would maximize the ex-post profit. Both European and US data indicate that this problem exists but that the amount of skewing is low.

In the general discussion regarding unbalanced bidding, firms are often framed as morally reprehensible. Such a framing could be criticized, as it would not be possible without the deficiencies in the bill of quantities produced by the client. And if the Contractor 1 would not exploit these deficiencies, then Contractor 2 would. Hence, superior information and strategically set prices is a way to compete, that is, best mitigated by improved design of the contracts.

Another version of strategically set prices is when the contractor exploits the fact that the client has missed some quantities in the ex-ante tendering documents, referring to 3b earlier. The client can use this mistake to lower the total price of the bid, by, for example, lowering all unit prices with 10%, in order to win the contract and raise the price on expected extra work. This requires that the contractor has better information than the client regarding what is required to build the transport infrastructure.

The Push for DB Contracting to Enhance Productivity

There is a general view that the productivity in the construction industry is lagging compared to other industries. Empirical macro-studies, mostly on labor productivity but also using data on capital and total factor productivity, support this view. Arguments that it is hard to deflate quality improvements in construction compared to other industries have been put forward as a way to nuance the low productive.

In the quest for higher productivity, there is a trend both in the United States and in Europe to push for using DB contracts at the expense of the traditional and still most commonly used DBB contracting. DB contracting gives the contractor more degrees of freedom in the design, which enables them to come up with new solutions and innovations. Such incentives are missing in a DBB contract, where the client provides the design and the contractor builds accordingly.

The theoretical underpinning for DB contracting providing incentives to innovate is rather straightforward, but there is a lack of statistical empirical studies supporting this claim. A hypothesis for the lack of empirical studies is the difficulty of getting a representative dataset regarding final cost on project level. There seems to be a generic problem within transport infrastructure construction with a lack interest for ex-post evaluations of costs and projects.

In the absence of empirical facts, public clients in Europe and the United States, equipped with theoretical arguments, have pushed for DB contracts. This push has been undertaken by setting up quantitative targets for the introduction of DB contracts. Examples include that 50% of all contracts were to be DB before 2018. These quantitative targets were fulfilled by relabeling DBB contracts as DB without any substantial change. The push for DB contract with the purpose to promote innovation is still present in many countries, but the change is slow.

PPP is Not a Way to Raise Funding

The usage of PPP to provide transport infrastructure is far from a new concept. An old Swedish example dates back to 1886, where a contractor was given a concession to build a tunnel for walking in central Stockholm and the right to take out a toll.

The definition of a PPP is the bundling of construction and maintenance into one long contract (at least 15 years), with the contractor putting the funding. Payment is received by tolls, shadow tolls paid by the public client or a combination of the two (Hart, 2003).

There are two theoretical arguments for PPPs. The first one is the incentive for the contractor to optimize the life cycle cost (LCC) perspective through bundling, as it is also responsible for the maintenance. In comparison to DBB and DB contracts, the PPP provides the contractor with the opportunity to build at a high initial cost to save money on maintenance. The second argument coincides with a DB contract, as a PPP gives the contractor degrees of freedom to find innovative solutions.

An argument against PPP is that the public client often has a lower capital cost than the private contractor. This is true for countries with sound public finances, but the lower capital cost must be weighed against the potential efficiency gains in terms of innovative solutions and optimal LCC that would not have been incurred otherwise. For countries with weak governmental finances, PPP becomes more interesting.

A pragmatic but insidious argument for PPP with shadow tolls is that it pushes the cost for new infrastructure onto future state budgets. This is handy in a shortsighted political perspective with new infrastructure paid by future taxpayers and still financial room for reforms today. Such reasoning is not a valid argument, as the transport infrastructure must be paid sooner or later, which encumbrance future state budgets. However, it is a real argument from a political perspective as the decision-makers might not be around when the bill is due.

Another argument against PPP, based on anecdotal evidence from southern Europe after the financial crisis in 2008 and more recently with regional Chinese public clients, is the lack of negotiating skills of the public side. Big construction companies with venture capital negotiate these contracts with commercial lawyers that the public client often cannot match. There are quite a few examples, where the public clients have signed contracts paying too much ex post.

Pros and Cons of Contracting Forms to Buy Transport Infrastructure

Like all problems in economics, there are pros and cons with all three contracting forms that are used to buy transport infrastructure. Statistical empirical analysis of contracting in the construction sector does not provide a lot of guidance. The studies that do exist are primarily based on road construction data in the United States. These studies focus on different subsections of the contracting decision, such as strategically set prices in DBB, the possibility to renegotiate contracts, the effect of having worked together earlier, etc., but not the overhauling question on which contracting form to use depending on different circumstances.

For this question, theoretical conclusions are still of guidance. To summarize there are theoretical pros and cons to each contracting form according to Table 4.

Table 4 Pros and cons on contracts in transport infrastructure construction

	<i>Pros</i>	<i>Cons</i>	<i>Comment</i>
DBB	Transparent and competitive procurement	Minor incentives for innovations	The most commonly used contracting form
DB	Incentives for innovations	Higher risk for contractor, therefore more expensive and less competitive	Problems with the so-called DB contracts that do not include degrees of freedom for the contractors
PPP	LCC perspective	More complex contracting negotiations	Political risk of being used as an “easy” way to finance infrastructure and sending the bill to future taxpayers

Summary

Most countries buy construction and maintenance of transport infrastructure from the market instead of undertaking in-house. Three ways of contracting can be distinguished. DBB is the traditional and still most commonly used way of tendering, DB gives the contractor incentives to innovate, while PPP improves the possibility to optimize LCC perspective. Empirical studies on which contract to prefer depending on external conditions are lacking. The future of this field is to model the contracts in more detail as the real contracts used do not coincide with the textbook definitions and gather data on final cost and quality to evaluate the efficiency of the contracts under different circumstances.

Acknowledgment

The author would like to acknowledge Jan-Eric Nilsson, Roger Pyddoke, and Andreas Vigren for valuable comments.

References

- Ewerhart, C., Fieseler, K., 2003. Procurement auctions & unit-price contracts. *Rand J. Econ.* 34 (3), 569–581.
 Flyvbjerg, B., 2009. Survival of the unfittest: why the worst infrastructure gets built and what we can do about it. *Ox. Rev. Econ. Policy* 25 (3), 344–367.
 Gibbons, R., 2005. Four formal(izable) theories of the firm? *J. Econ. Behav. Org.* 58 (2), 200–245.
 Hart, O., 2003. Incomplete contracts and public ownership: remarks and an application to public-private partnerships. *Econ. J.* 119, 69–76.
 Shleifer, A., 1998. State versus private ownership. *J. Econ. Persp.* 12 (4), 133–150.

Further Reading

- Abdel-Wahab, M., Vogl, B., 2011. Trends of productivity growth in the construction industry across Europe, U.S., and Japan. *Constr. Manage. Econ.* 29 (6), 635–644.
 Bajari, P., Houghton, S., Tadelis, S., 2014. Bidding for incomplete contracts: an empirical analysis of adaptation costs. *Am. Econ. Rev.* 104 (4), 1288–1319.
 Gil, R., Marion, J., 2013. Self-enforcing agreements and relational contracting: evidence from California highway procurement. *J. Law Econ. Organ.* 29 (2), 239–277.
 Hart, O., Shleifer, A., Vishny, R., 1997. The proper scope of government: theory and an application to prisons. *Q. J. Econ.* 112 (4), 1126–1161.
 Nyström, J., Nilsson, J.E., Lind, H., 2016. Degrees of freedom and innovations in construction contracts. *Transp. Policy* 47, 119–126.
 Odolinski, K., Smith, A.S.J., 2016. Assessing the cost impact of competitive tendering in rail infrastructure maintenance services: evidence from the Swedish reforms (1999–2011). *J. Transp. Econ. Policy* 50 (1), 93–112.

Procurement of Public Transport: Contractual Regimes

Andrew Smith, Chris Nash, Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	308
Theoretical Perspectives	308
Issues in Contract Design	309
Franchising Authority	310
Net or Gross Cost Contracts	310
Size and Scope	310
Length	310
Assets	311
Incentives	311
Treatment of Staff	311
Vertical Separation	311
Experience in Rail Transport	312
Experience in the Bus Sector	313
Conclusions	313
References	314
Further Reading	314

Introduction

The last 30 years have seen a global trend toward the involvement of the private sector in providing and funding a wide range of public services. The introduction of competition has formed a key part of the associated reforms alongside privatization of different degrees and types. Reforms to introduce private sector involvement and competition in the provision of public transport can therefore be seen as part of a much wider set of reforms, covering the utilities, and also public services in, for example, health, education, and a wide range of local government funded services.

There is a strong case for the involvement of government in the provision of public transport or at the very least in specifying the level of service required. It is therefore frequently considered that competition for the market (or ex ante competition) is a more appropriate means of introducing competition in public transport than competition in the market (or “on-road/on-track” competition). Under the competition for the market model, the government can specify what it wants to provide and then invite the market to bid for the exclusive right to provide these services on a given route or routes. This form of competition avoids the possible loss of economies of scale and/or density implied by competition in the market, while at the same time capturing the benefits of competition through the bidding process; also avoiding the costs of government ownership/provision or regulation. Where physical infrastructure is key to delivering public services, for example in rail, this form of competition may require some form of vertical separation (infrastructure from operations). Indeed, the notion of allowing competition between operators on a common infrastructure represents a significant rethink of regulatory policy in respect of network industries; where regulation of vertically integrated monopolies was previously considered the only option.

As discussed below, in most cases around the world public transport (rail and bus) is provided either by state-run bodies with no competition, or via competition for the market. In some cases, where there is no competition, this has occurred by a deliberate choice to allow an incumbent, who previously won the market based on a competition, to retain the market based on a negotiated settlement for some specified period. In other cases it is because public services have not yet been opened up to competition. There are also some examples of on-road competition in the bus market, provided on commercial terms, most notably in terms of developed markets in Great Britain (outside London); as well as in some developing countries and for long distance bus (and in a few cases rail) services more generally.

The article is structured as follows. The second section provides a brief introduction to the theoretical literature on anticipated benefits and issues associated with introducing competition for the market. In the third section the issue of how best to design competitive tendering exercises and contracts is considered. The fourth and fifth sections consider in turn experience of procurement in rail and bus sectors and the last section reaches conclusions.

Theoretical Perspectives

The benefits of competition in driving efficient outcomes have been long established. In the regulatory sphere, however, the notion of using competition for the market, in place of competition in the market, is relatively new (Demsetz, 1968). The general argument

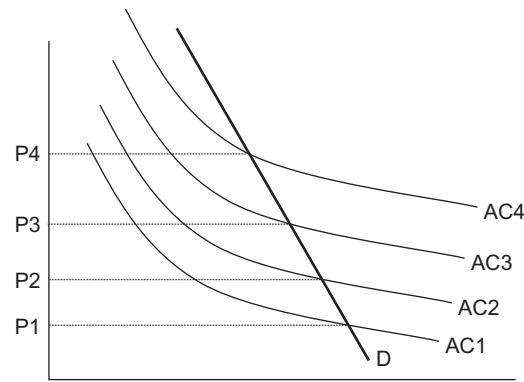


Figure 1 Illustration of competitive bidding.

for this form of competition is that in the presence of natural monopoly, especially concerning infrastructure and also extending to service provision, means that competition may be hard to achieve. In public transport, there are further particular considerations. It can also be difficult for a sensible timetable to emerge from a competitive equilibrium, with the tendency for service bunching (Hotelling, 1929). Profit maximizing operators may have incentives to schedule services close to those of its competitors to gain passengers and as a longer-term strategy to drive competitors out of the market completely. This occurred in the British bus reforms in the 1980s, but such competition was rarely sustained, usually one or other operator left the market. There was also much use of anticompetitive practices such as running free buses in front of their services to drive out rivals, but this was tackled by tightening competition laws.

Moreover, the search costs involved in selecting one public transport operator over another on the same route mean that price competition may not emerge in some transport markets, at least on local services, where the convenience of getting on the first bus or train outweighs any saving in cost. Governments of course also perceive a central role in ensuring appropriate service levels in situations where services would not be profitable for a private firm, and in ensuring good quality, appropriately priced, and accessible public transport services. Improved public transport is also seen as a way of tackling the problems of congestion and pollution caused by excessive car traffic, particularly in large cities (Goeurden et al., 2006).

The alternatives, namely economic regulation of a private monopolist or state ownership, are also widely perceived to possess disadvantages. Competition for the market, in which firms compete for the monopoly right to supply a particular market, enables a competitive process to deliver good outcomes for consumers, while ensuring that only a single provider operates at any given time. Thus the optimal cost structure is retained, while avoiding problems of monopolistic practices. Further, there is no need for economic regulation, provided that the price of the service is the basis for selection in the competitive process.

In theory the competitive process therefore ensures that firms will bid no more than their average cost, thus preventing monopoly profits, while also ensuring that the most efficient firm wins (Fig. 1). Dynamic efficiency is also encouraged since once the contract is signed, firms have strong incentives to reduce costs and grow the market. It should be noted that in most public transport tendering processes, the degree of subsidy required (or in some profitable markets, premium paid) is the basis for selection. In this scenario it is necessary to specify the price regulation framework that will be implemented (otherwise firms, having won, could exploit consumers). More widely a contract is needed between government and the firm to specify, inter alia, the service required and its quality.

Overall competition for the market appears to offer the best of all worlds. However, its success depends on a number of critical factors. First there needs to be sufficient number of bidders, and the avoidance of collusion. Information asymmetries may also interrupt an efficient outcome—in particular, incumbent advantage could deter other companies from bidding. There is also the problem of winner's curse which may either mean that the winning bidder overpays and cannot fulfill the contract—putting the problem back to government—or that bidders bid too conservatively because of fear of winner's curse. Information is crucially important here to mitigate against such problems. To the extent that firms believe they can bid aggressively and then renegotiate—or in the case of naive bidding—the winning bidder may not be the best firm to run the services.

The ability to specify and enforce a contract that sets out what the government wants in terms of service level and quality can also be a problem, particularly where quality is hard to define, or where there is uncertainty about what the government will want in the future. Contract design issues are discussed in the third section. Finally, there may be issues regarding asset handover when services change to new operators following a competitive tender. All of these factors greatly complicate the theoretical predictions, and could undermine the potential benefits of this form of competition.

Issues in Contract Design

If the decision is taken to use competitive tendering as a way of procuring public transport services, a number of crucial questions must be considered. This section considers them, and the options that will be faced (Nash et al., 2019).

Franchising Authority

A country intending to use competitive tendering for procurement of public transport must establish a tendering authority with the skills and resources to draw up appropriate contracts, monitor performance and take enforcement action should the franchisee fail to deliver. That authority is usually an agency of central, regional, or local government, although sometimes it may be a joint authority covering several local authorities.

In many cases (e.g., regions in Sweden, states in Germany) the same authority is responsible for both bus and rail services and may also have wider responsibilities regarding planning and road infrastructure. This enables them to integrate decisions across the modes and with wider considerations. The authority also needs to be large enough to exploit economies of scale in the size of tenders (if economies of scale are exhausted at a smaller scale than the authority, then it may let more than one tender, as do many of the German federal states). But such a policy may remove provision of local bus services from the layer of government best able to determine needs, as local bus services are very much a local concern.

Net or Gross Cost Contracts

A key decision in letting contracts is whether to go for net cost contracts, in which the operator bids for a certain level of subsidy and takes all the risks regarding both revenue and costs, or a gross cost contract in which the operator bids for a payment to reimburse the costs of operating the service and pays the revenue to the authority, so bearing cost risk only.

There is often a strong presumption among economists in favor of net cost contracts, as these give the operator a strong incentive to boost revenue. However, net cost contracts may have disadvantages. They expose the operator to many risks outside their control, such as relating to the state of the economy and local employment levels, which are strong determinants of demand (while there are cost risks outside the control of the operator, such general wage trends and fuel prices, these are generally less critical than demand). But contracts may partially or fully protect the operator against externally determined cost risks as well). Sometimes such demand side risks are shared, for instance by paying the contractor a share of the revenue or by taking specific risks such as those associated with levels of employment or GDP.

Exposure to a high level of risk may discourage some operators, especially smaller ones who cannot pool the risks from a multiplicity of contracts, from bidding at all, and encourage those who do bid to build in a large risk margin to their bids. It may also lead to a high level of franchise failure, where operators become bankrupt or exercise rights they have under the contract to withdraw early, giving the authority the cost and challenge of keeping services running (sometimes at short notice) and holding a premature competition for a new contract.

The choice of net versus gross contracts is often linked to the issue of how far operators are left to plan services and set fares. If operators are left a lot of freedom under the contract, subject to providing a minimum service level, then it is particularly important to give strong incentives to increase patronage and revenue, so a net cost contract may be most appropriate. If on the other hand a franchising authority itself wishes to determine timetables and fares, then the ability of the operator to influence revenue is very limited, and a gross cost contract with some quality incentives (see further) may be most appropriate.

Size and Scope

The issue here is geographical size of the contract and whether it covers a single mode or all public transport modes in the area. Linked to this is the question of whether both regional- and long-distance traffic should be included in franchises or whether they should be franchised separately (or indeed whether long distance transport should be left to commercial operators without a franchise). Practice ranges from contracts to cover all modes for metropolitan areas, down to individual bus routes. The size needs to be large enough to exploit any economies of scale (but not beyond the level at which diseconomies set in; see the fourth section). As noted earlier the size needs to be compatible with the area covered by the tendering authority and—particularly where planning is left to the train operator—compatible with the sensible planning of services. Thus, for instance, rail services into cities often fulfill a regional role and could not sensibly be divided into franchises covering individual cities.

There are advantages to covering all modes in a single tender, especially if the operator is responsible for planning services, but this has the disadvantage that evidence appears to suggest that the minimum efficient size for rail tenders is much bigger than bus, and the resulting large tender size may prevent smaller bus operators from competing. The problem does not arise if the tendering authority completely plans services and fares as it may then let separate contracts of differing size for bus and rail.

Length

There are also difficult trade-offs to make regarding contract length, which again interact with other aspects of the contract. Sometimes very short contracts, even as short as a year, are used. These maximize flexibility for the contracting authority in terms of future service changes and maximize competition by granting a monopoly for a very short period. But they may not be attractive to operators, because of bidding and start-up costs which must be recovered in a very short time, and thus may only attract a small number of bids which may not be very competitive. They clearly give no incentive for investment, whether in physical assets (which may be minimized depending how the ownership of assets is organized—discussed below) or in marketing, planning or improving the efficiency of working practices. They also involve all the costs of competition, for authority and bidders, at very frequent intervals.

On the other hand, it is possible to go for a long franchise—franchises exist of up to 25 years. But long franchises also have their problems. They grant a monopoly for a long period of time, while break points may be built in under which continuation of the contract is dependent on performance, if these are too rigorous they may lead the operator to regard the contract as a series of shorter ones from the point of view of risk and investment. They inevitably require a process to be determined for changes in services in the light of developments, as these cannot sensibly be specified for the full contract period, and for how this impacts on payments under the contract. This may be a time-consuming and contentious issue, as both parties may try to take advantage of the situation to achieve outcomes favorable for themselves. In such circumstances the question as to whether a standard regulatory model might be preferred to competitive tendering is pertinent.

Assets

As noted earlier, ownership of assets—rolling stock, depots, terminals—is a key issue in considering contract length. If the operator is required to supply all of these, then clearly a long contract, or some other way of relieving the operator of the residual value risk (such as guaranteeing the taking over of the assets by the succeeding operator at a fair price) is needed. Moreover, ownership of the assets will give the incumbent a strong advantage in any bidding process.

Thus for expensive assets, such as terminals, depots and rail rolling stock, there is a trend toward ownership by the franchising authority. That may also have disadvantages, in that the operator may be deemed to have a stronger interest in and ability for the efficient choice of assets. Sometimes assets, especially rolling stock, may be leased by third parties, but that does not fully resolve the problem, as they may still perceive a residual risk in short leasing agreements, particularly in the case of specialized equipment.

The much lower cost and more active second hand market in buses means these issues are much less significant in bus contracts than rail. Thus even where the operator is required to provide the buses, bus contracts of 5 years or less are common.

Incentives

With a net cost contract, operators should have strong incentives on both cost and demand sides, although the demand side incentives will be weaker on heavily subsidized services (where revenue is small relative to costs) than on more commercial ones. As noted earlier, the need for demand side incentives is much stronger with gross cost contracts than net, although they may be applied to both.

Usually contracts will include a set of key performance indicators covering issues such as crowding and reliability, which may be linked directly to financial incentives, or which may simply be contractual requirements, failure to achieve which would put the operator in breach of contract. This may lead to the requirement for remedial action, with the possibility of termination of the contract if the action is deemed inadequate.

The extent to which quality is a factor in the award of the contract also varies. Usually, there will be a prequalification procedure, which rules out any bidders who are not deemed capable of providing the services at the required quality, or who have inadequate finance. Particularly where the operator is responsible for planning timetables and procuring assets, their proposals regarding these factors may be given a weighting in the decision as to who should win the contract, alongside the financial attractiveness of the bid.

It may also be thought that in some cases additional incentives are necessary to reduce costs, particularly where the costs of challenging the cost base are disproportionate to the length of contract. In Great Britain, relatively short franchises, combined with a strong union, performance related penalties and loss of farebox revenue from a prolonged industrial dispute, have meant that operators have not sought to challenge the cost base as might be hoped. This is particularly so because at franchise replacement a new bidder takes over an existing company with the same staff and rolling stock. It may therefore be thought necessary to mandate the implementation of particular initiatives, such as driver only operation, in the franchise agreement, backed up by the introduction of rolling stock with the necessary technology. To date, however, trade unions in Great Britain have been able to resist this cost saving initiative.

Treatment of Staff

An important issue in franchising is the treatment of the staff of the existing operator. The ability of an incoming operator to recruit its own staff, setting its own salaries and working conditions, may be important in achieving the cost reductions often ascribed to competitive tendering. But this creates great uncertainty for the existing staff, and may worsen staff salaries and working conditions. Moreover it may lead to difficulties in terms of the new operators being able to recruit the staff they need, particularly if tendering is occurring on a large scale. Thus in many cases the new operator is required to take on the existing staff at existing salaries and working conditions.

Vertical Separation

Competitive tendering is often associated with vertical separation of infrastructure from operations, for rail as well as road based public transport. But arrangements are then needed to ensure that infrastructure and operations are planned to optimize the system as a whole. This issue is considered in another article and is therefore not pursued here.

Experience in Rail Transport

Europe offers the richest evidence on the impact of competition for the market for passenger rail services (Nash et al. 2019), though there have been examples elsewhere, for example in Australia and South America. Within Europe, competitive tendering has advanced furthest in Britain, Germany, and Sweden, though other countries such as the Netherlands, Norway, and Poland have also introduced some competitive tendering. The latest reform impulse from the European Commission, the 4th Railway Package, requires all public service contracts to be subject to competitive tendering by 2023, unless good reason for a direct award can be established. For commercial services, open access must be introduced by 2020 unless this threatens the financial equilibrium of a public service contract (this has been the arrangement for international services since 2010). Long distance services in Europe to date have typically been operated by the incumbent rail operator, but with new entry being permitted on an open-access basis in some cases. Such entry has occurred on main routes in Sweden, Italy, Austria, and the Czech Republic and to a very limited extent, in Germany and Great Britain. In respect of domestic long distance services, Great Britain is the exception, having chosen to franchise almost all long distance services, with a very small level of open-access competition alongside the franchise model.

For regional rail services, competitive tendering has been the chosen model where liberalization has occurred. Great Britain franchised all such services and importantly did not allow the former state-owned operator, British Rail to bid (British Rail effectively ceased to exist). This approach differs from that taken in Germany and Sweden where the incumbent operators were permitted to bid. In Sweden almost all such services have been subject to competitive tendering. In Germany only part of the regional services market has been opened up to competition. As of 2015, the incumbent (DB) still operated 70% of regional passenger train-km (Nash et al., 2019).

In respect of patronage growth, the European reforms can be seen to be very successful, though there is always the problem of establishing the counterfactual (Nash et al., 2013). Very strong growth in passenger-km was achieved in the three most liberalized markets (Great Britain, and regional services in Germany and Sweden). Great Britain, for example, saw usage more than double between 1995 and 2013. However, econometric models suggest that only part of this growth can be attributed to franchising (Preston and Robbins, 2013; Wardman, 2006). Rationalization—transfer of responsibility for specifying services to regional government—may also be associated with some of the growth in Sweden and Germany, while also growing the market in France even with a state-owned, monopoly provider. While net cost contracts are widely considered to have been successful at growing the market in Great Britain (though noting the franchise failure issues discussed below), Sweden, and Germany have both achieved strong patronage increases despite greater reliance on gross cost contracts. There is no clear evidence as to whether longer franchises work best in this respect, as a mix of franchise lengths have been used across all countries.

Turning to cost, tendering has been successful in Germany and Sweden, with reported savings of the order of 20%–30% per train-km. However, one issue is that these savings relate to subsidies rather than costs, and where net cost contracts are used the reductions could reflect revenue growth rather than cost savings (though Germany and especially Sweden have tended to use gross cost contracts to a much greater extent than Great Britain so subsidy reductions are likely to be more reflective of cost falls). In Great Britain unit costs have instead risen by between 16% and 25% (per vehicle-km and train-km, respectively) over the period since franchising (Smith, 2016). Costs per passenger-km have fallen however because of the sharp rise in demand (average loading has risen). Overall subsidies per train-km (infrastructure and operations) were lower in 2015 than at the time of the reforms, though having almost doubled in the intervening period. It may therefore be said that tendering in Sweden and Germany has been relatively successful, whereas cost control in Great Britain has been a significant problem.

The reasons for the different cost experiences are complex. One possible explanation is that the greater use of gross cost contracts in Sweden and Germany has encouraged focus on cost control, whereas in Great Britain there has been a strong emphasis on revenue growth. Also in Great Britain, when a franchise changes hands the incoming winner of the competition takes over an existing company, thus taking on the existing staff at existing wages and conditions. It has been argued that this approach, combined with relatively short franchises, and the penalties and lost revenue associated with industrial disputes, has created little incentive to operators to challenge the cost base. Elsewhere in Europe, the entrant has been free to recruit new staff at new wages and conditions, the existing staff having the option to remain with the previous operator (although this is no longer the case in Germany).

The most recent evidence suggests that some of Britain's rail franchises are too big from a cost perspective (likewise some of the larger German franchises). While longer franchises have been suggested to encourage longer term thinking in terms of cost reduction there is little evidence to suggest that this has been effective in Great Britain and a range of franchise lengths has been used across the different countries. There is some evidence that longer franchises have encouraged better rolling stock deals in Germany. In Great Britain problems of cost control and operational issues have been linked to vertical separation and numerous efforts have been made to enable greater coordination.

A significant problem in Great Britain has been franchise failure (this has been much less of a problem in other European countries, but it also occurred in Melbourne and South America). In the early period after franchising a number of companies failed after making too optimistic projections particularly about the degree to which they could reduce costs. More recent problems have emerged repeatedly on one of Britain's most profitable inter-city routes, the East Coast Main Line with three different companies failing in succession and withdrawing early. Thus it appears that winner's curse has been a problem to contend with.

For profitable routes that can be provided by the market, net cost (rather than gross-cost) contracts are clearly the most appropriate, assuming tendering is used as opposed to open-access competition. However, there are risks that lie outside the control of the operators, given the importance of the state of the economy in particular on trends in passenger journeys. For large franchises, these risks can be substantial. Various risk sharing approaches have been tried (linking franchise payments to revenue or GDP) and

also increasing financial penalties/capital requirements in the event of franchise failure. None of these approaches has yet been successful (if success is judged by seeing a franchisee fulfill their contract). In other European countries commercial services are provided by the incumbent railway, with open-access competition. Increased use of open-access is under consideration in Great Britain as well.

Experience in the Bus Sector

As noted earlier, local public transport is very often in the hands of a public sector monopoly operator owned by the local authority; less often there is a private sector regulated monopoly. Some countries, mostly in developing countries, leave public transport up to unregulated private companies competing on the road; in developed countries this approach is more often used for inter urban and long distance services. Finally, competitive tendering is in fairly common use ([Hensher and Wallis, 2005](#)), for instance for all services in Sweden and many in Germany, Norway, Denmark, and the Netherlands as well as some cities in the United States, Australia, and New Zealand. Most often competitive tendering takes place at a route or corridor level but sometimes (for instance in France and Spain) the entire public transport system of a city is let as a single contract.

Britain is unique in having simultaneously in 1986 introduced competition for the market in London, and for unprofitable services elsewhere, and competition in the market for profitable services outside London. In its monitoring exercise, the government collected data on costs in both sectors. In the ten years following deregulation, cost per bus km fell in real terms by 45% both in London and elsewhere, reversing a long-term upward trend ([Nash, 1993](#)). It should be noted that this reduction is somewhat exaggerated by the fact that some costs (e.g., of timetabling, publicity and provision of bus stops/shelters) remained with local authorities, and also there was a trend to smaller buses, but that there was a very substantial reduction, only partly at the expense of the wages and conditions of staff, is clear ([Heseltine and Silcock, 1990](#)). Both fares and service levels rose under both regimes, although in London it was a deliberate decision by the local transport authority, whereas elsewhere it was a market reaction (outside London services are provided on commercial terms in most cases). With few exceptions it appears that price competition was relatively ineffective in the local bus industry; passengers tended to take the first bus to come regardless of price, so service frequency was more important to competitors than price. But it should be noted that on the street competition has been comparatively rare; most of Britain has bus services provided by unregulated private monopolies ([Competition Commission, 2011](#)), but subject to the threat of entry.

However, there has been one big difference in the experience of on and off the street competition. In London, there was a modest rise in bus patronage over this period (more recently, the growth has been much stronger but in the context of increased subsidies to improve the fares and service level mix). But elsewhere in Britain bus patronage continued its downward trend. It is of course the case that as by far the largest city in Britain London offers a more favorable environment for bus operation, with denser population and lower car ownership than elsewhere. But [Fairhurst and Edwards \(1996\)](#) find that partly this more favorable trend in London was due to better integrated services; failure elsewhere to plan around regular interval services and common fare structures meant that patronage was below what might otherwise have been achieved with the service frequency and fares observed. In the light of this experience, there has been a move to secure better quality and better planned services through the use of quality partnerships between bus companies and local authorities, in which both parties agree to cooperate to raise the quality of bus services, and recent legislation makes the adoption of quality contracts (essentially the London model) elsewhere in Great Britain easier to implement ([White, 2018](#)).

In general, the experience of competitive tendering elsewhere is that it has also achieved significant cost savings ([Hensher and Wallis, 2005](#)), although generally lower than in Britain, and quite variable with circumstances (such as whether the incumbent was publicly or privately owned). What has been observed, however, is that usually competitive tendering after the initial round does not produce more cost savings, and indeed often some of those from the first round are lost. There seem to be a variety of reasons for this. In part it reflects tighter quality standards over time (e.g., new or better vehicles). But Hensher and Wallis argue that it also reflects a learning process on the part of bidders (such as a decline in the number of unrealistically low bids) and possibly also a decline in the amount of competition over time. They argue that, therefore, once an efficient operator has been assigned to a route, negotiated contracts, based on benchmarking to ensure the continued efficiency of the operator, and with performance incentives built into the contract, might have advantages over further competitive tendering. For instance, they avoid the costs of competitive tendering (although obviously having costs of their own) and the disruption of a possible change of operator, and might be more effective in encouraging innovation and the increase in traffic levels.

Conclusions

There are many choices to be made when implementing a system of competitive tendering and the choice between them involves difficult trade-offs. Decisions are also interdependent, as some packages make more sense than others. For instance, there may be a choice between long net cost franchises with ownership of assets and considerable freedom over planning of services and fares by operators (most suitable for commercial services), and short gross cost franchises with the franchising authority providing the assets and tightly specifying timetables and fares. The best approach is likely to vary greatly with circumstances such as the geography of the services, the modes involved and the level of subsidy. What follows is that franchising is a skilled business on the buyer side,

requiring staff with appropriate ability and experience. Where it is undertaken by many different local or regional bodies, sharing data and experience will be important to ensuring that it is undertaken efficiently.

Both bus and rail tendering are widespread in Europe and also to be found elsewhere. Overall it would appear that rail tendering has been successful in increasing patronage on public transport rail services in Europe, though other factors have also contributed to the growth, including rationalization. It also seems to have been successful in general in bringing about cost reductions across a number of countries, though Great Britain is an important counter case, having seen substantial increases in cost per train-km. Subsidies per train-km and passenger-km for tendered services have fallen across Europe, this being the case even in Great Britain, in the latter case driven by strong revenue growth. There is evidence that some of Britain's franchises are too big from a cost perspective (along with some of the larger German franchises), but while longer franchises have been discussed as a way of encouraging cost reductions, there is little evidence of their effectiveness in that regard, and longer franchises increase risk and reduce flexibility for funders.

A key lesson seems to be that gross cost contracts, perhaps combined with patronage incentives, work well for regional and suburban services where there is a regional government body with responsibility for setting fares and marketing services (perhaps supported by a public body to own and manage the rolling stock). Where gross cost contracts have been applied good performance on the cost and demand side has been achieved with few problems of franchise failure. Net cost contracts have been associated with high incidence of franchise failure, particularly in Great Britain. Part of the problem may lie in the choice that Great Britain made to use franchising even for commercially viable long distance services (in contrast to the rest of Europe) and where the size of the franchises increases risk and financial exposure to operators. Where competition has been introduced for commercial services elsewhere, it has usually been competition in the market, and this approach is under consideration in Great Britain.

The evidence from bus tendering likewise supports the theoretical prediction that competitive tendering will bring about cost reductions, although the savings achieved are quite variable and tend to be lower when the incumbent is already a private company. In the one case where local bus services are produced by competition in the market, Great Britain outside London, it appears that competition for the market has worked better in terms of patronage, because of better coordination of services, although the obvious differences between London and the rest of Great Britain make that conclusion tentative.

However, the evidence is that successive waves of tendering do not achieve further savings compared with the first one, and indeed that some of the initial gains are lost. That has led to the suggestion that negotiated contracts, with benchmarking to check costs and strong contractual incentives, may be a good option after the initial round of tendering. The evidence on this point is less clear in rail, although there has been a decline in the number of bidders over time and also a flight of private capital, with most bidders now being state-owned incumbents from the large European rail markets. But consideration of the application of negotiated, direct awards, supported by benchmarking, could be a useful dynamic in the rail sector also, especially in Great Britain where cost rises have been an issue.

References

- Competition Commission, 2011. Local bus services market investigation. Final Report. www.competition-commission.org.uk.
- Demsetz, H., 1968. Why Regulate Utilities? *J. Law Econ.* 11, 55–65.
- Fairhurst, M., Edwards, D., 1996. Bus travel trends in the UK. *Trans. Rev.* 16 (2), 157–167.
- Goeverden, C., Rietveld, P., Koelemeijer, J., Peeters, P., 2006. Subsidies in public transport. *Eur. Trans.* 32, 5–25.
- Hensher, D.A., Wallis, I.P., 2005. Competitive tendering as a contracting mechanism for subsidising transport. *J. Trans. Econ. Policy* 39, 295–321.
- Heseltine, P.M., Silcock, D.T., 1990. The effects of bus deregulation on costs. *J. Trans. Econ. Policy* 24.
- Hotelling, H., 1929. Stability in competition. *Econ. J.* 39 (153), 41–57.
- Nash, C., Smith, A., Crozet, Y., Link, H., Nilsson, J.-E., 2019. How to liberalise rail passenger services? Lessons from European experience. *Trans. Policy* 79 (2019), 11–12.
- Nash, C.A., 1993. British bus deregulation. *Econ. J.* 103 (419), 1042–1049.
- Preston, J.M., Robbins, D., 2013. Evaluating the long term impacts of transport policy: the case of passenger rail privatisation. *Res. Transp. Econ.* 39 (1), 14–20.
- Smith, A.S.J., 2016. Liberalisation of passenger rail services – Case Study: Britain, report for CERRE.
- Wardman, M., 2006. Demand for rail travel and the effects of external factors. *Transp. Res. E* 42 (3), 129–148.
- White, P., 2018. Prospects in Britain in the light of the Bus Services Act 2017. *Res. Transp. Econ.* 69, 337–343.

Further Reading

- Wheat, P.E., Smith, A.S.J., 2015. Do the usual results of railway returns to scale and density hold in the case of heterogeneity in outputs: a hedonic cost function approach. *J. Trans. Econ. Policy* 49 (1), 35–47.

The Mono-Centric City Model and Commuting Cost

Zhi-Chun Li*, Ya-Juan Chen[†], *School of Management, Huazhong University of Science and Technology, Wuhan, China; [†]School of Management, Wuhan University of Technology, Wuhan, China

© 2021 Elsevier Inc. All rights reserved.

Introduction	315
The Classic Mono-Centric City Model	315
Some Extensions of the Classic Mono-centric City Model	317
In the Presence of Traffic Congestion	317
Mode Choice	317
Heterogeneous Households in Terms of Income Level	317
Asymmetric Urban Structure	318
Polycentric City	318
Policy Evaluation	318
Infrastructure Investment	319
Congestion Pricing	319
Commuting Subsidies	320
Acknowledgment	320
References	320
Further Reading	320

Introduction

In urban economics, the mono-centric city model is a basic concept in modeling the spatial distribution of population in a city. It has been widely recognized as a useful tool in addressing the relationship between commuting cost, housing price, and housing consumption. It was developed in the 1960s and 1970s, largely through the work of William Alonso, Richard Muth, and Edwin Mills (Alonso, 1964; Muth, 1969; Mills, 1972). Although the modern cities have become increasingly polycentric, the mono-centric city model is still an important cornerstone for development of various city models (e.g., polycentric city models). In this chapter, we begin with the classic mono-centric city model, then describe some extensions of the classic mono-centric city model, and finally discuss some applications of the model in policy evaluation.

The Classic Mono-Centric City Model

In the classic mono-centric city model, the city is assumed to be on a featureless plain and transportation in the city is possible in all directions; all employment and goods and services are concentrated in a highly compact central business district (CBD); each location is only characterized by its distance from the CBD, and the travel time from any location to the workplace located in the CBD is a linear function of the distance between that location and the workplace, that is, the transport system in the city is radial and dense in all directions, and congestion-free; all households are homogeneous, implying that their income levels and utility functions are identical.

The household income is spent on transportation, housing and non-housing goods. Each household aims to maximize its own utility by determining the residential location, the consumption of housing and the amount of non-housing goods within a capital budget constraint. The household utility maximization problem can be represented as

$$\max_{z,g} u(z,g), \quad (1)$$

$$\text{subject to } Y - tx = z(x) + p(x)g(x), \quad (2)$$

where $u(z,g)$ is the household utility function, $g(x)$ is the housing consumption per household at location x , and $z(x)$ is the composite non-housing goods consumption per household at location x , for which the price is normalized to 1. Y is the average annual household income, tx is the average annual round-trip cost for households living at location x , and $p(x)$ is the average annual housing price at location x .

When the household residential location choice equilibrium state is reached, all households in the city enjoy the same utility level regardless of their residential locations. Denote u as such indifferent utility level. Then, solving the maximization problem (1)-(2) and setting the equilibrium utility $u(z,g) = u$, one can obtain the equilibrium housing rental price p , equilibrium housing consumption g , and non-housing goods consumption z , as functions of location x and the common utility u , represented as

$$p = p(x,u), \quad (3)$$

$$g = g(x, u), \quad (4)$$

$$z = z(x, u). \quad (5)$$

This basic model is useful in explaining the mechanism of trade-off between housing rental price and its proximity to the CBD. The housing rental price must decrease with x so as to offset the reduction in net income ($Y - tx$) due to an increase in the commuting cost.

We now look at the supply side of the housing market. The property developers determine the land and capital investment so as to maximize their net profit. Denote $H(x)$ as the housing production at location x . It is given by a constant returns-to-scale production function $H(x) = H(M(x), L(x))$, where $M(x)$ and $L(x)$ denote the inputs of capital and land at location x , respectively. Let $r(x)$ be the rental price of land at location x , and k be the price of capital (i.e., the interest rate). The net profit, $\pi(x)$, at location x can thus be given as

$$\pi(x) = p(x)H(M(x), L(x)) - (r(x)L(x) + kM(x)). \quad (6)$$

The first term on the right-hand side of Eq. (6) is the total revenue from housing rents, and the final two terms in the bracket are the land rent cost and the capital cost, respectively. With the assumption of the constant returns-to-scale housing production function, Eq. (6) can be converted to

$$\pi(x) = L(x) \left(p(x)H(M(x)/L(x), 1) - r(x) - k \frac{M(x)}{L(x)} \right). \quad (7)$$

Let $S(x) = M(x)/L(x)$. It represents the capital investment per unit of land area at location x , namely, the capital investment intensity. Eq. (7) can then be rewritten as

$$\pi(x) = L(x)(p(x)h(S(x)) - r(x) - kS(x)), \quad (8)$$

where $h(S(x)) = H(S(x), 1)$, i.e., the housing output per unit of land area at location x . Thus, maximizing the net profit at location x in Eq. (8) is equivalent to maximizing the net profit of unit land input, expressed as

$$\max_S \pi(x) = p(x)h(S(x)) - r(x) - kS(x). \quad (9)$$

The first-order optimality condition of the maximization problem (9) yields

$$\frac{\partial \pi}{\partial S} = p(x)h'(S) - k = 0. \quad (10)$$

Substituting Eq. (3) into Eq. (10) yields the optimal capital investment intensity S as a function of location x and the common utility u , as follows.

$$S(x) = S(x, u). \quad (11)$$

The residential density, denoted as $n(x)$, at location x can thus be calculated by

$$n(x) = \frac{h(S(x, u))}{g(x, u)}. \quad (12)$$

Note that under perfect competition, the property developers earn zero profit. Thus, setting $\pi(x) = 0$ yields the land rental price $r(x)$ as a function of location x and common utility u , as follows.

$$r(x) = r(x, u). \quad (13)$$

So far, the model gives rise to the functions $p(x, u)$, $g(x, u)$, $z(x, u)$, $S(x, u)$, $n(x, u)$, and $r(x, u)$, which are all functions of utility u .

Finally, we define the housing demand-supply equilibrium conditions. On one hand, the equilibrium rent per unit of land area devoted to housing at the city boundary equals the exogenous agricultural rent or the opportunity cost of the land; that is,

$$r(B, u) = r_a, \quad (14)$$

where r_a is the constant agricultural rent, and B is the city size.

On the other hand, the housing demand-supply equilibrium requires that all households fit exactly inside the city boundary, i.e.,

$$\int_0^B 2\pi n(x, u)xdx = N, \quad (15)$$

where N is the city's population size.

There are unknown variables B and u (or N), depending on whether the city is “open” or “closed”. For a closed city, the city’s population size N is fixed exogenously, whereas the utility u and city size B are determined endogenously by Eqs. (14) and (15). For an open city, the value of u is given exogenously, which is the base utility level of the economy, whereas the population size N and city size B are determined endogenously by Eqs. (14) and (15).

Some Extensions of the Classic Mono-centric City Model

The classic mono-centric city model has been extended in numerous ways, such as by considering traffic congestion effects, competition between transport modes, heterogeneity of households in terms of income level, asymmetric urban structure, and polycentric urban structure.

In the Presence of Traffic Congestion

Commuting cost plays an important role in shaping the urban spatial structure. The classic mono-centric city model assumes that the commuting cost depends only on the distance between the residential location and the workplace, and is thus independent of the actual residential density or travel demand. That’s to say, there is no traffic congestion in the urban system. The congestion-free assumption may cause a significant bias in the prediction capability of the model and thus restricts its applications in practice.

In order to consider the congestion effects, transportation cost at any location depends not only on the distance traveled, but also on the density of traffic from that location to the CBD. Let $N(x)$ be the number of residents living beyond location x , which can be determined by integrating the residential density from location x to the city boundary. Suppose that there is one commuter per household, and thus $N(x)$ is the total number of commuters traversing location x . Let $L_T(x)$ be the amount of land devoted to transport usage around location x . Accordingly, the marginal travel time per unit of distance at location x can be calculated by a function of the traffic-land ratio, $N(x)/L_T(x)$. The annual travel cost of each household at location x , denoted by $T(x)$, can thus be calculated by

$$T(x) = \int_0^x a \left(\frac{N(w)}{L_T(w)} \right)^m dw, \quad (16)$$

where a and m are positive parameters, which can be calibrated by survey data.

Mode Choice

The classic mono-centric city model involves an auto-only urban system, and the interaction and substitution effects between auto and transit modes (e.g., bus) are usually ignored. In reality, auto and bus transit usually share the same roadway. They have impacts on each other and thus on the level of congestion in the urban system. The congestion interaction between them directly affects the cost of travel by a particular mode and thus the travelers’ mode choices, which further influence service strategy of the transit operator (e.g., frequency and/or fare). Conversely, a change in the transit service frequency and/or fare can cause a change in the modal split of auto and bus. As a result, the level of traffic congestion on the road also changes. It is, therefore, of great importance to incorporate congestion interaction and substitution effects between modes and the mode choice behavior of travelers in the city models (Li et al., 2013; Li and Wang, 2018). One can adopt Wardrop’s user equilibrium or multinomial logit model to model the travelers’ mode choice behavior.

Heterogeneous Households in Terms of Income Level

The classic urban model usually assumed that all the households in the city system are homogenous in terms of income level. However, income heterogeneity is universal in reality, leading to a strong effect on the bidding price for housing, and thus the household residential distribution. It has currently been observed in some large cities across the world that high-income households and low-income households are spatially segregated, which is referred to as the residential segregation phenomenon. For example, high-income households in China prefer to reside in the urban central areas, whereas low-income households tend to live in the suburban areas. However, a reverse residential pattern (i.e., high-income households reside in the suburbs, whereas low-income households reside in urban central areas) occurs in some other countries or areas, such as Detroit, the United States (Brueckner et al., 1999). Accordingly, the homogeneity assumption may cause a significant bias in the prediction capability of the model in the urban spatial structure. The effects of income heterogeneity on the household residential location choices and the urban spatial structure should thus be considered. A recent study has identified the conditions under which either the rich or the poor live in the urban central area while the other prefer the suburb (Li and Peng, 2016). The study has also shown that a high emission tax can drive the poor to migrate from suburbs to urban central areas, and the rich to migrate from urban central areas to suburbs.

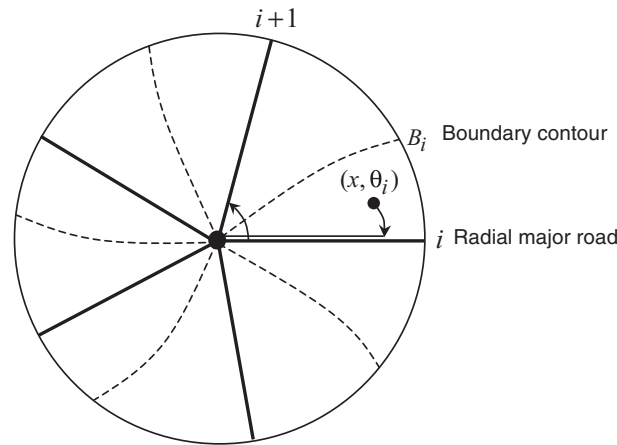


Figure 1 An asymmetric city with radial major roads.

Asymmetric Urban Structure

In the literature, the urban spatial equilibrium analysis of a mono-centric city usually assumed that the urban system was a perfectly divisible dense system consisting of an infinite number of radial roads. That is to say, they mainly focused on a symmetric urban structure with evenly distributed radial roads along mono-centric circular city. A more realistic case is to relax the “symmetry” assumption to consider an asymmetric two-dimensional mono-centric city with a small number of radial major roads (e.g., these radial major roads are not uniformly distributed along the city and/or have different capacities) connecting the CBD to its suburban areas, as shown in [Fig. 1](#). In such an urban system, a ring-radial travel method can be applied, namely, commuters first travel along a minor ring road to get to a radial major road, and then proceed along the radial major road to reach the city center ([Li et al., 2013](#)). Referring to [Fig. 1](#), two adjacent alternative radial major roads in the two-dimensional mono-centric city compete for travel demand. A watershed boundary (or market area boundary) exists that divides the area between these adjacent radial roads into two sub-areas. When an equilibrium state is reached, no commuter in the city has an incentive to change his/her choice of radial major road when traveling to the CBD. This means that the cost of traveling from a point on the boundary contour to the CBD using two adjacent radial roads is identical. It should be pointed out that the traffic flows on the radial major roads are interdependent: an increase in the flow on one radial road leads to a decrease in the flow on another. In addition, the single-mode asymmetric mono-centric city model can be extended to a multi-mode scenario ([Li and Wang, 2018](#)).

Polycentric City

Since Alonso (1964), Muth (1969), and Mills’s (1972) theory of urban spatial structure, the mono-centric city models have become widely accepted in urban economics. The mono-centric city models assume that all job opportunities cluster in a CBD area. The impractical assumption of a single employment center has suffered some criticisms. Polycentric models have been developed since the 1970s. The models focus on how workers decide where to live and work and the resulting spatial patterns of land and housing rents, population densities and commuting regions ([Anas and Kim, 1996](#)). The polycentrism favors employment decentralization and polycentric development, which may improve commuting patterns by shortening individual commuting distances and time, and the polycentric city could promote spatial equity through increased opportunities to access jobs nearby housing leading to positive effects on productivity, livability and efficiency. In the polycentric city models, the locations of the employment centers may be exogenously given or endogenously determined, and the workplace for a two-worker household may be situated at one or multiple locations. The polycentric city models can also be extended to investigate the competition and collaboration between cities in an urban agglomeration, and between urban agglomerations, particularly in the context of China’s Belt and Road Initiative.

Policy Evaluation

In an urban system, there are multiple stakeholders, for example, the government, property developers, and households. A strong interaction exists between urban policies implemented and the urban spatial structure. For example, implementation of a transport policy has a direct impact on location accessibility and thus individuals’ commuting cost, which in turn affects households’ residential location choices and property developers’ housing production, which further affects the urban spatial structure in terms of residential distribution and housing market (housing rental prices and housing space). Inversely, urban land use pattern, which involves the spatial distribution of residents over the city through the housing supply-demand balance, can govern the travel

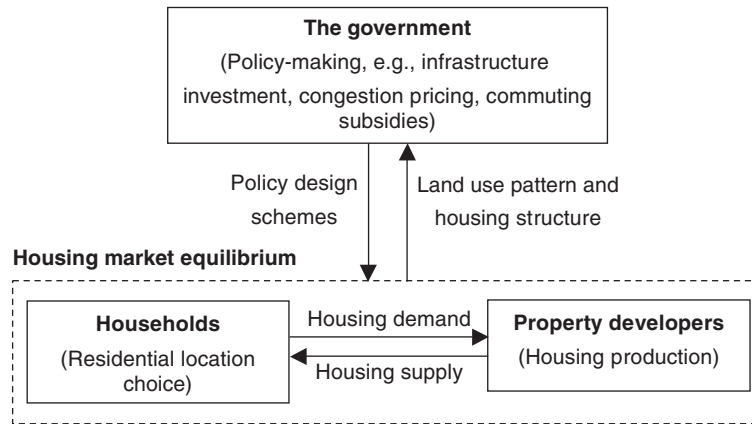


Figure 2 Interactions among stakeholders in urban systems.

demand distribution across the city, which affects the design and optimization of urban policies for achieving sustainable urban development. The interactions among the government, property developers, and households are shown in Fig. 2. Generally, urban policies include supply-side policies (e.g., infrastructure investment) and demand-side policies (e.g., congested road/parking pricing, commuting subsidies), which are discussed in the following sections.

Infrastructure Investment

In response to the rapid growth in the number of motorized vehicles and traffic congestion, local authorities have launched a number of transportation infrastructure investment projects to mitigate auto-related problems (e.g., road traffic congestion, parking congestion, and environmental issues). Some typical infrastructure investment strategies include: (1) constructing new radial major roads that connect the city's CBD to its suburban areas (Li et al., 2013) and/or new parking facilities (e.g., parking garages, on-street parking lots); (2) implementing integrated rail and railway station property development, in which such variables as rail line length, number and spacing of stations, headway, and fare can be optimized (Li et al., 2012); (3) introducing transit-oriented development (TOD) project (Peng et al., 2017), which refers to medium- and high-density housing along with complementary public uses, jobs, retail, and services in mixed-use development around transit stations. In the TOD projects, there is a need to determine the optimal location, number and size of the TOD zones on a rail transit line such that the TOD investment projects are economically viable and cost-effective from an input-output perspective. The infrastructure investment can cause changes in the urban spatial pattern in terms of the residential location choices, property values, and housing market. These externality effects should be incorporated in the infrastructure investment decisions.

Congestion Pricing

The under-pricing of auto travel is regarded as an important source of market distortion (or failure) and resource waste in urban transportation systems. As a result, auto is overused and road/parking congestion is becoming an increasingly serious problem, particularly in densely populated large cities, such as Beijing and Hong Kong. Congestion pricing has been widely proposed as a feasible solution to the growing problem of traffic congestion because of its potential to internalize congestion and environmental externalities and correct market distortion (Verhoef, 2005; Chen et al., 2018). The ongoing rapid development of information and communication technologies has aided and supported the practical implementation of congestion pricing schemes.

Congestion pricing schemes include the first-best and second-best schemes. The first-best scheme is also referred to as social optimum scheme or marginal cost pricing scheme. In the first-best scheme, the congestion tolls equal the congestion externalities that an additional auto trip imposes on the existing vehicles in this system. There are various second-best tolling schemes, such as cordon-based tolling scheme (i.e., each auto user passing through a specified cordon is charged a fixed toll) and distance-based tolling scheme (i.e., toll is proportional to the distance traveled). Tradable credit scheme, as a variation of congestion pricing schemes, has recently received considerable attention. In the tradable credit scheme, the authorities initially allocate the credits to all households in the urban system, and the households pay a certain amount of credits according to the externalities they generate during the course of their trips.

It has been seen that congestion pricing has an effect on the urban form through changing household residential distribution and housing market. Such effect is closely related to the redistribution schemes of the congestion tolls. For example, the congestion tolls may be redistributed to the urban residents or kept by the authority (e.g., the government). With different redistribution schemes, congestion toll may lead to a more compact or a more decentralized urban structure.

Commuting Subsidies

Commuting subsidy policies have been implemented in many countries in different forms, such as work-related deductions (e.g., France or Italy), a single allowance for commuters (e.g., Germany, the Netherlands, Denmark, Switzerland), or income-dependent tax credits (e.g., Japan) (Potter et al., 2006). Commuting subsidies aim to improve efficiency in the labor market by encouraging workers to increase their radius of job search and to commute further for a better match. From a perspective of social welfare, encouraging people to increase their job search radius in order to find better matches can be seen as positive, but longer commutes also entail negative externalities. The social welfare depends on the trade-off between better matches and the negative externalities.

Regarding the topic of commuting subsidies, there are many issues that deserve to be studied. For example, whether commuting costs/reimbursements should be made deductible/exempt from the income tax, and if so, whether full or only partial costs should be deducted; how to design a commuting subsidy scheme to reach an efficient level of job search and commuting; how to redistribute the commuting subsidies by vehicle ownership, income, or commuting distance, and what are the re-distributional effects of the commuting subsidies on the urban system.

Acknowledgment

The work described in this paper was jointly supported by the NSFC-JPI Urban Europe research project (71961137001) and the NSFC project (71890974/71890970; 71501150).

References

- Alonso, W., 1964. *Location and Land Use*. Harvard University Press, Cambridge, Massachusetts.
- Anas, A., Kim, I., 1996. General equilibrium models of polycentric urban land use with endogenous congestion and job agglomeration. *J. Urban Econ.* 40, 232–256.
- Brueckner, J.K., Thisse, J.F., Zenou, Y., 1999. Why is central Paris rich and downtown Detroit poor? An amenity-based theory. *Eur. Econ. Rev.* 43 (1), 91–107.
- Chen, Y.J., Li, Z.C., Lam, W.H.K., 2018. Cordon toll pricing in a multi-modal linear monocentric city with stochastic auto travel time. *Transportmetrica A* 14 (1–2), 22–49.
- Li, Z.C., Chen, Y.J., Wang, Y.D., Lam, W.H.K., Wong, S.C., 2013. Optimal density of radial major roads in a two-dimensional monocentric city with endogenous residential distribution and housing prices. *Reg. Sci. Urban Econ.* 43 (6), 927–937.
- Li, Z.C., Lam, W.H.K., Wong, S.C., Choi, K., 2012. Modeling the effects of integrated rail and property development on the design of rail line services in a linear monocentric city. *Transp. Res.* 46 (6), 710–728.
- Li, Z.C., Peng, Y.T., 2016. Modeling the effects of vehicle emission taxes on residential location choices of different-income households. *Transp. Res.* 48, 248–266.
- Li, Z.C., Wang, Y.D., 2018. Analysis of multimodal two-dimensional urban system equilibrium for cordon toll pricing and bus service design. *Transp. Res.* 111, 244–265.
- Mills, E.S., 1972. *Urban Economics*. Scott Foresman, Glenview, Illinois.
- Muth, R.F., 1969. *Cities and Housing*. University of Chicago Press, Chicago.
- Peng, Y.T., Li, Z.C., Choi, K., 2017. Transit-oriented development in an urban rail transportation corridor. *Transp. Res.* 103, 269–290.
- Potter, S., Enoch, T., Black, C., Ubbels, B., 2006. Tax treatment of employer commuting support: an international review. *Transp. Rev.* 26, 221–237.
- Verhoef, E.T., 2005. Second-best congestion pricing schemes in the monocentric city. *J. Urban Econ.* 58 (3), 367–388.

Further Reading

- Anas, A., 1982. *Residential Location Markets and Urban Transportation*. Academic Press, New York.
- Anas, A., Moses, L.N., 1979. Mode choice, transport structure and urban land use. *J. Urban Econ.* 6 (2), 228–246.
- Anderson, S.P., de Palma, A., 2004. The economics of pricing parking. *J. Urban Econ.* 55 (1), 1–20.
- Arnott, R., Inci, E., 2006. An integrated model of downtown parking and traffic congestion. *J. Urban Econ.* 60 (3), 418–442.
- Borck, R., Wrede, M., 2005. Political economy of commuting subsidies. *J. Urban Econ.* 57 (3), 478–499.
- Borck, R., Wrede, M., 2008. Commuting subsidies with two transport modes. *J. Urban Econ.* 63 (3), 841–848.
- Brueckner, J.K., 1987. The structure of urban equilibria: a unified treatment of the Muth-Mills model. In: Mills, E.S., (Ed.), *Handbook of Regional and Urban Economics*, Vol. 2, Urban Economics, Elsevier Science, Amsterdam, pp. 821–845.
- Brueckner, J.K., 2005. Transport subsidies, system choice, and urban sprawl. *Reg. Sci. Urban Econ.* 35 (6), 715–733.
- Chen, Y.J., Li, Z.C., Lam, W.H.K., Choi, K., 2016. Tradable location tax credit scheme for balancing traffic congestion and environmental externalities. *Int. J. Sust. Transp.* 10 (10), 917–934.
- DeSalvo, J.S., Huq, M., 2005. Mode choice, commuting cost, and urban household behavior. *J. Reg. Sci.* 45 (3), 493–517.
- Fujita, M., 1989. *Urban Economic Theory*. Cambridge University Press, Cambridge, U.K.
- Kraus, M., 2006. Monocentric cities. In: Arnott, R.J., McMillan, D.P. (Eds.), *A Companion to Urban Economics*. Blackwell Publishing, Oxford, pp. 96–108.
- Kwon, Y., 2003. The effect of a change in wages on welfare in a two-class monocentric city. *J. Reg. Sci.* 43 (1), 63–72.
- Li, Z.C., Guo, Q.W., Lam, W.H.K., Wong, S.C., 2015. Transit technology investment and selection under urban population volatility: a real option perspective. *Transp. Res.* 78, 318–340.
- Sasaki, K., 1990. Income class, modal choice, and urban spatial structure. *J. Urban Econ.* 27 (3), 322–343.
- Su, Q., DeSalvo, J.S., 2008. The effect of transportation subsidies on urban sprawl. *J. Reg. Sci.* 48 (3), 567–594.
- Wheaton, W.C., 1998. Land use and density in cities with congestion. *J. Urban Econ.* 43, 258–272.

Value of Time in Freight Transport

Gerard de Jong, Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	321
Value of Time in Freight Transport: What is it?	321
Why do Transport Time Savings Have a Value?	321
What is it Used for?	322
Methods Used to Determine the Freight VTTS	322
Factor Costs	323
Modeling Studies	323
Some Outcomes for the VTTS in Freight	324
Conclusions on the Value of Time in Freight Transport	325
References	325
Further Reading	325

Introduction

Value of Time in Freight Transport: What is it?

The value of time (VOT) in freight transport is the marginal rate of substitution between the time it takes to transport goods on the one hand and money on the other hand. A related concept is the value of transport time savings (VTTS) in freight transport, which is the amount of money a firm or society, is willing to pay to save one unit of transport time (reduce the transport time by one unit). Even though these concepts are called “values” transport time actually represents a disutility or cost: more transport time (for the same goods flows) is a negative and when time has a higher VOT, this negative impact is increased. In this context, one could also talk about the cost of transport time increases (CTTI). There is value and there are benefits in reducing transport time.

Why do Transport Time Savings Have a Value?

Reductions in transport time in freight can lead to different benefits for different decision-makers in freight transport (Table 1). The transport itself is for a large part carried out by specialized firms, the carriers that have their own (or sometimes leased) transport vehicles and staff to drive and supervise these vehicles. If the transport time would be reduced (e.g., as a result of a new road link), the vehicles and drivers would be used for a shorter duration, and the carriers could either perform more transport assignments or do the same assignment with fewer vehicles and staff. In the short run the possibilities or benefit from this might be limited, but in the long run they can reschedule their services and production factors. So, the long-term gain of a time saving could be as high as the sum of the time-related transport costs for that amount of time. These are often taken to be all the transport costs except the purely distance-related costs (such as fuel costs, energy costs, and user charges for the use of rail infrastructure).

The shipping firms (producers or traders of commodities) want to send their products to their clients and so they have a demand for transport services. A part of this demand might be supplied by the shippers themselves (this is called own account transport), but in most cases the transport is contracted out to carrier firms or intermediaries, for which the shipper pays an agreed rate. Benefits that a carrier gets when time is reduced may be transferred to the shippers through reductions in the price of transport (depending on the market conditions). But even when no transport cost benefits are passed on to the shipper, shippers can have a benefit from transport time reductions: the fact that the goods themselves are away for a shorter time can:

- lead to a reduction in the interest costs on the capital invested in the goods during the time that the goods are transported, and to
- the reduction in the value of perishable goods during transit, but also in the possibility that the production process is disrupted by missing inputs or that customers cannot be supplied due to lack of stock.

Shippers with own account transport can be expected to see both the vehicle and staff benefits as well as the cargo-related benefits.

The carrier's VOT typically relates to the driver and the transport equipment; drivers are paid wages and the equipment has to be depreciated (or their lease paid). A carrier incurs these costs irrespective of the type of goods being carried (or even whether the vehicle is loaded)—these costs are not commodity-specific.

The specific VOT of the shippers on the other hand relates to the contents of the shipment, the market value of the goods, their depreciation, their tendency to degrade, the risk of theft, and also to the wider logistics system of which the transport operation is a part.

Sometimes, not the senders but the receivers of the goods are responsible for the transport and already considered the owners of the goods during the transport. In such cases these receiving firms will be the ones where the cargo-related benefits of time savings may accrue.

Table 1 Different benefits from time savings for different decision-makers in freight transport

	<i>Benefits related to the cargo</i>	<i>Benefits related to the vehicles and staff (transport costs)</i>
Carrier	Not relevant	Relevant
Own account shipper	Relevant	Relevant
Shipper that contract out	Relevant	Not relevant

For society, both types of benefits (cargo-related and transport cost related) matter, so in a societal evaluation, like a social cost-benefit analysis (SCBA), both types of benefits should be taken into account. This does not necessarily imply that both should be included in the transport time benefits through the VTTS, but this is definitely one possibility. We will come back to the use of freight VTTS in SCBA later in this chapter.

What is it Used for?

Similarly to the VOT in passenger transport, the freight VOT is used for two different purposes. On the one hand, the freight and the passenger VTTS are inputs into the SCBA of infrastructure projects, making it possible to compare time savings against investment cost and other effects of the project. On the other hand, both the passenger and freight transport VOTs are also used in traffic forecasting models to calculate the “generalized cost,” an important explanatory variable in the model, defined as a linear combination of travel time and cost. Values of time for SCBA and modeling may differ in the treatment of taxes. In the further discussions, we focus on the use of the freight VTTS in SCBA.

Regarding the use of the freight VTTS in SCBA, we can say that there are two distinct “schools” or approaches in terms of what is included in the VTTS (**Fig. 1**). The first school defines VTTS as the cargo component only ([Vierth, 2013](#)) and includes impacts of projects on staff and vehicle time saved in the SCBA through the transport cost savings (together with the distance-based cost, including energy and infrastructure use charges, that should not be in the VTTS in either approach).

The second group uses a VTTS that contains both the cargo and the transport services component ([de Jong et al., 2014](#)); in the SCBA all time benefits in freight are expressed through the VTTS and transport costs benefits only refer to reductions in the trip lengths (if any).

There is no absolute truth in the choice between approaches A and B. Both methods can be carried out in such a way that no benefits are forgotten and no benefits are counted twice. If VTTS is the cargo component only, one should, in evaluating time saving projects, take care not to forget reducing the time dependent transport costs (as part of the cost savings). If both components are in the VTTS one should take care not to double count the distance-dependent transport cost. Approach A with a low VTTS is used in practice in for instance Sweden and Approach B with a high VTTS in the Netherlands.

Methods Used to Determine the Freight VTTS

The methods to find proper freight VTTS for project appraisal or forecasting can be classified into factor-cost methods and modeling studies.

Factor Costs

Factor cost refers to the cost of all input factors into some production process, in this case the production of freight transport services. The factor-cost method in VTTS research tries to find the cost of the input factors that will be saved when transport time is saved. A reduction in transport time could release production factors (e.g., labor, vehicles), which could be used in other shipments. These

Approach A	Approach B
Time savings:	Time savings:
Cargo time saved	Cargo time saved Staff time saved Vehicle time saved
VTTS	VTTS
Transport cost savings:	Transport cost savings:
Distance cost saved Staff cost saved Vehicle cost saved	Distance cost saved

Figure 1 Approaches to time/cost benefits in SCBA.

cost reductions can be calculated using data on wages and vehicle prices, taxes, insurances, and maintenance. A more difficult issue is whether overheads and non-transport inventory and logistic cost should be included. This could be analyzed using the other type of method, that is, the modeling studies. Some researchers have argued that not all labor cost should be used in the VTTS, since some of the time gains cannot be used productively. This too can be analyzed by modeling decisions in freight transport and focusing on the implied time-cost trade-offs. The issue of which cost items to include also depends on the time horizon: in the long run, more items will be time-dependent and the long VTTS will be higher than the short run one.

When applying the factor cost method, the researcher also has to decide which costs will be attributed to the impact of transport time itself and which costs to the impact of transport time reliability. In a model it might be possible to separate out the cost related to the average transport time on the one hand and the extra cost of delivering too late (too early) or the cost on the variability on the other hand. In a factor cost calculation that uses a simple transport cost function, it is very difficult to separate out the impact of transport time variability. However, if one would use a full logistics cost function, both the VTTS and the VTTV (value of transport time variability) can be calculated as the derivatives to transport time and variability (expressed as the standard deviation or variance of transport time). This full logistics costs function would then include components for the value of the goods, the deterioration of the goods, and the cost of keeping a safety stock (as a function of, among other things, transport time variability).

Modeling Studies

The modeling studies that seek to determine the VTTS can first of all be distinguished according to the type of data that is used in the modeling:

- revealed preference (RP) studies;
- stated preference (SP) studies.

There have also been a few joint RP/SP models in freight transport.

The general definition of RP studies is that they use data on observed real-world behavior. In freight transport these are data on the choices that shippers, carriers, intermediaries, or drivers actually made in practice. In SP studies, the researcher defines experimental choice alternatives. These are presented to the respondents and they are asked to make a choice or give their most preferred alternative. The researcher who wants to model VTTS using RP data first has to look for real world choice situations where these decision-makers are trading off time against cost. Some examples of such situations might be:

- the choice of mode between a fast and expensive mode and a slower and cheaper mode;
- the choice of route between a fast toll route and a congested toll-free route.

The VTTS in the literature that come from RP data are mostly based on mode choice data (e.g., road versus rail, rail versus inland waterways), which could be (Tavasszy and de Jong, 2014) aggregate (transport flows by commodity type between zones), or disaggregate (individual shipments between firms). The inputs on transport time and costs for the chosen and unchosen modes usually come from transport networks. Both in aggregate and disaggregate RP studies, researchers often have to deal with the high degree of correlation between transport time and cost, which makes it difficult to estimate significant coefficients for both attributes. In an SP experiment, the researcher can control the correlation, so this problem can be avoided. Nevertheless, there have been various RP studies where it proved possible to estimate both time and cost coefficients (e.g., the Dutch BasGoed model). Aggregate models are much more common in freight transport than disaggregate models since there are not many data sources on actual choices (e.g., mode and route choice) at the individual shipment level; there are many more travel surveys in passenger transport than shipper surveys in freight.

RP-based VTTS are usually the by-product of the development of a national or regional freight transport forecasting study. The national studies that have been undertaken in recent decades with the explicit aim of determining VTTS have almost all been SP studies (e.g., in Norway, the Netherlands, Germany, Belgium, UK, Australia). In an SP freight VTTS study, decision-makers (in practice almost exclusively shippers or carriers) are asked to elicit their preferences for hypothetical alternatives, which usually refer to shipments/transportations they have been involved in, and which will have different attribute levels for transport time and cost, and possibly also for other attributes of the shipment. In practice, the RP-based values often differ from the SP ones, but a recent meta-analysis of freight VTTS (Binsuwardan et al., 2019) has not found a significant difference between SP and RP values.

These SP experiments can be about the choice of mode or route, as in RP. However, for VTTS research favorable experience has been obtained in abstract time versus cost experiments in which all alternatives that are presented refer to the same mode and the same route. In an abstract time versus cost experiment the alternatives have different scores on travel time, travel cost, and possibly other attributes, but the alternatives are not given a mode or route label, such as "rail transport" or "motorway with toll."

A specific problem in finding the VTTS for freight, as opposed to the passenger VTTS, is that some of the information in goods transport, especially on transport cost and logistic cost, may be commercially sensitive. Firms in freight transport may be reluctant to share this information with client, competitors, and the public. SP data has some advantages in the case of freight transport modeling, in particular as it may be possible to obtain data (e.g., on costs and rates) which would be difficult to acquire by other methods.

A difficult issue in SP surveys on freight service valuation is who to interview. The best advice probably is to interview shippers (that contract transport out) to obtain estimates of the cargo VTTS and carriers to get the transport cost VTTS. However, when doing

this, it is important to make clear to the shippers and carriers which considerations they could take in mind and which aspects we are asking from other decision-makers.

For models that are built for the sole purpose of deriving monetary valuations, the mixed logit specification is nowadays the prevalent type of model. Among models that are estimated for use as forecasting models, multinomial, and nested logit models are dominant. This is caused by the fact that forecasting models are run many times after they have been estimated (and this would take very long with mixed logit because of the repeated simulation that is involved) whereas for a monetary valuation just doing model estimation is sufficient.

There is considerable heterogeneity in passenger transport, but even more in freight transport. The size of the shipment may vary from a small parcel to the contents of a bulk vessel. The value of a truckload of waste material is vastly different from that of the contents of a full money transport vehicle. One might then conclude that the value of freight travel time savings is so heterogeneous that it cannot be established, but this would go too far. Especially the transport costs component of the VTTS, which usually is by far the largest component, is largely independent of the type of goods, as was argued above. Furthermore, heterogeneity in the value of the cargo-related component can be taken into account by applying a proper segmentation (e.g., by mode, type of good). Also, it is important to use proper scaling for the VTTS (e.g., using a value for a typical shipment size or a value per ton).

Some Outcomes for the VTTS in Freight

Table 2 contains a new summary of values from the literature that refer to road and rail (studies that include other modes than road and rail are very scarce). It only used studies that yield values expressed per ton or that can be expressed as such. We explicitly make the distinction between values for the goods (cargo) component, the transport services (costs) component, and the sum of both components. All the values that we used here come from Europe.

In terms of numerical values for the VTTS, the picture that emerges from **Table 2** is that in almost all studies, the value of the goods component in the VTTS is rather low relative to the values found for the sum of both components. If we take the example of rail transport, the value of the cargo component is between 0 and 0.7 euro/ton/hour for all goods together, with a central value of about 0.3, whereas the values found for the sum of both components are around 2.7 euro/ton/hour. For road the cargo component is between 0 and 2.5 euro/ton/hour, with a central value of around 1 euro. The sum of both components is about 7 euro/ton/hour. Large values for the goods component of the VTTS are only found for specific high-value commodities (e.g., automotive, vehicles and machines, container, finished goods, and especially express goods). For both road and rail transport, the goods component appears to be the minor component in the VTTS.

The range that we obtain in **Table 2** for the cargo component is relatively large. Apart from methodological differences between studies and in income levels between countries, this can be explained by variation between commodity types. The total VTTS for rail per ton (2.7) is clearly lower than for road (which amounts to about 7 euro/ton).

A key result is that the transport service component of the VTTS will be (especially in the long run) more or less equal to the cost of producing the transport services per hour (the sum of the staff and vehicle cost per hour including overheads, but not including distance-dependent cost). It is therefore not really needed to do new SP research in some country to get these values, one can simply use the factor costs method to find this component. This component will hardly or not varies between commodity types, but it will vary between modes.

Table 2 Value of transport time savings (VTTS) in goods transport (in 2010 euro/ton/hour)

<i>Publication</i>	<i>Country</i>	<i>Data</i>	<i>Method</i>	<i>VTTS</i>
<i>The goods (cargo) component in the VTTS:</i>				
Publications between 2000 and 2015	Finland, Italy, Switzerland, Belgium, Denmark, Sweden, Switzerland, Norway, The Netherlands, France, Germany and the UK	SP, with one or two RP studies	Mainly MNL and mixed logit, some ordered probit, nested logit and weighted regression MNL	0–0.7 for rail; central value: 0.3 0–2.5 for road; central value: 1
<i>The transport services (costs) component in the VTTS:</i>				
Publication from 2013	The Netherlands	Transport cost functions	Factor cost	2.4 for rail 6.3 for road
<i>Both components in the VTTS:</i>				
Publication from 2013	The Netherlands	Transport cost functions and SP	Factor cost and MNL models	2.7 for rail 7.0 for road

For passenger transport, so many VTTS are available that various meta-regressions have been carried out, that try to explain the VTTS obtained from attributes of the respective countries and study methods used. For freight transport, the number of VTTS available is somewhere near the margin of what is minimally needed for a meta-regression.

The cargo component of the VTTS cannot so straightforwardly be derived from the factor cost. If possible, specific SP surveys are recommended. If these are not possible, one could use for the cargo component of the VTTS in freight a fraction, for example, 10%–20%, of the full transport cost (including time- and distance-dependent cost); this fraction is based on the results from the Netherlands (de Jong et al., 2014). Variation between commodity types (which one would expect for the cargo component) can be derived from the results of recent French (CGSP, 2013), German (BVU and TNS Infratest, 2014), or UK studies (Fowkes, 2006) on VTTS.

Conclusions on the Value of Time in Freight Transport

The value of transport time savings (VTTS) is one of the most used and researched concepts in transport economics. Most research refers to the VTTS in passenger transport, but the freight VTTS is also often used in the social cost-benefit analysis (SCBA) of transport projects (some projects even especially cater for freight transport benefits) and in transport forecasting models.

As regards the freight VTTS it is important to distinguish between the transport costs component and the goods component. The former is usually more important and can be derived from SP interviews with carriers, but the long run value can also be obtained by calculating the non-distance-related transport costs per hour. The cargo component usually comes from SP interviews with shippers, which can also contain transport time reliability as an attribute so that both monetary values can be derived from the same study.

In SCBA, the VTTS can be defined as only this cargo component, with the transport costs component being included in transport cost savings. Another approach, different but equally valid, is to include both components in the VTTS and therefore in the transport time benefits, so that the transport costs savings only refer to reductions in the distance-related costs.

References

- Binsuwanan, J., de Jong, G.C., Batley, R.P., Wheat, P., 2019. The value of travel time savings in freight transport: a meta-analysis. Paper presented at UTSG 2019. University of Leeds, UK.
- BVU and TNS Infratest, 2014. Entwicklung eines Modells zur Berechnung von modalen Verlagerungen im Güterverkehr für die Ableitung konsistenter Bewertungsansätze für die Bundesverkehrswegeplanung, Vorläufiger Endbericht. Freiburg/München: BVU/TNS Infratest.
- CGSP, 2013. Cost-benefit assessment of public investments, report of the mission chaired by Emile Quinet, summary and recommendations. Available from: <https://www.strategie.gouv.fr/>.
- de Jong, G.C., Kouwenhoven, M., Bates, J., Koster, P., Verhoef, E., Tavasszy, L., Warffemius, P., 2014. New SP-values of time and reliability for freight transport in the Netherlands. *Transp. Res. Part E* 64, 71–87.
- Fowkes, A.S., 2006. The design and interpretation of freight stated preference experiments seeking to elicit behavioral valuations of journey attributes. ITS, University of Leeds, UK.
- Tavasszy, L.A., de Jong, G.C., 2014. Modeling Freight Transport. Elsevier Insights Series, London/Waltham, Elsevier.

Further Reading

- Danielis, R., Marcucci, E., Rotaris, L., 2005. Logistics managers' stated preferences for freight service attributes. *Transp. Res. Part E* 41, 201–215.
- Feo-Valero, M., Garcia-Menendez, L., Garrido-Hidalgo, R., 2011. Valuing freight transport time using transport demand modeling: a bibliographical review. *Transp. Rev.* 201, 1–27.
- Halse, A.H., Samstad, H., Killi, M., Flügel, S., Ramjerdi, F., 2010. Valuation of freight transport time and reliability (in Norwegian), TØI report 1083/2010, Institute of Transport Economics, Oslo.
- de Jong, G.C., 2008. Value of freight travel-time savings, revised and extended chapter. In: Hensher, D.A., Button, K.J., (Eds.) *Handbook of Transport Modeling*, Handbooks in Transport, Vol. 1, Oxford/Amsterdam, Elsevier, pp. 649–663.
- Maggi, R., Rudel, R., 2008. The value of quality attributes in freight transport: evidence from an SP-experiment in Switzerland. In: Ben-Akiva, M.E., Meersman, H., van der Voorde, E. (Eds.), *Recent Developments in Transport Modeling, Lessons for the Freight Sector*. Bingley, UK: Emerald.
- Significance, VU University, John Bates Services, TNO, NEA, TNS NIPO, PanelClix, Values of time and reliability in passenger and freight transport in the Netherlands, Report for the Ministry of Infrastructure and the Environment. The Hague, 2013.
- Vierth, I., 2013. Valuation of transport time savings and improved reliability. In: Ben-Akiva, M.E., Meersman, H., van der Voorde, E. (Eds.), *Recent Developments in Transport Modeling: Lessons for the Freight Sector*. Bingley, UK: Emerald.
- Zamparini, L., Reggiani, A., 2007. Freight transport and the value of travel time savings: a meta-analysis of empirical studies. *Transp. Rev.* (27–5), 621–636.

The Economics and Planning of Urban Freight Transport

Edoardo Marcucci^{*,†}, Valerio Gatta[†], Michela Le Pira[‡], ^{*}Faculty of Logistics, Molde University College, Britvegen, Molde, Norway;

[†]Department of Political Science, University of Roma Tre, Via Gabriello Chiabrera, Roma, Italy; [‡]Department of Civil Engineering and Architecture, University of Catania, Via Santa Sofia, Catania, Italy

© 2021 Elsevier Ltd. All rights reserved.

Economics of Urban Freight Transport	326
The Need for UFT Planning to Foster Sustainability	327
UFT Externalities	327
Private and Public Coordination	327
Complexity of Supply Chains	328
UFT Planning	328
Conclusion	329
Acknowledgment	329
Biographies	330
References	330
Further Reading	331

Economics of Urban Freight Transport

The majority of EU population (74%) lives in city, towns and suburbs, and this number is expected to increase in the next years (UN, 2019). Urbanisation implies that people gather in a place, remote from sources of food, consumer products, and waste disposal opportunities (Ogden, 1992). Under this respect, freight transport demand is *derived*, but *essential*.

Urban freight transport (UFT), urban logistics, last mile logistics, city logistics are all different terms referring to the processes connected to freight movement at the urban level. UFT has several definition, one of the most cited being the one by Taniguchi et al. (1999, p. 17): “the process for totally optimising the logistics and transport activities by private companies in urban areas while considering the traffic environment, the traffic congestion and energy consumption within the framework of a market economy”. UFT is a complex world characterized by scarce knowledge and heterogeneous stakeholders, interacting and often competing: some of them (shippers, transport providers, retailers) aiming more at *efficiency* while some others (citizens, consumers, policy-makers) at *sustainability* (Gatta et al., 2017). In order to understand the economics of urban freight, it is useful to recall basic economic theory concepts and then apply them to UFT (Button and Pearman, 1981).

Economics essentially focuses on how stakeholders make choices; specifically, how they allocate scarce resources to alternative forms of production, how they distribute production to consumption between groups, how they choose between present and future consumption.

Production is linked to supply, while consumption to demand. Relevant information and understanding derive from a formal analysis of supply and demand interaction. Consumer behaviour relates to demand, while to study supply one needs to focus on costs. Firm behaviour is interlinked both with market type and structure. The overall purpose being the investigation of social welfare and its relationship with production, consumption and exchange of goods and services.

Supply, demand, market structures and stakeholders’ objectives are extremely heterogeneous when one looks close enough to UFT. In fact, there are different market characteristics even when, at a first glimpse, they might look very similar. The market power a generic transport provider has in UFT (inferred by the average number of employees) is much lower with respect to one operating in the building sector. Shippers are interested in getting the goods safely delivered to their customers in the city, while transport providers would like to minimize their operating costs, and retailers are interested in getting the goods delivered on time to their shop. While this is typically true, its level of representativeness might drastically change in different supply chains. What is true for frozen food might be less true for fresh ones, just to mention two apparently similar types of delivery services. These characteristics are reflected in the high level of market segmentation, customers’ preferences (rich and poor) and suppliers’ cost functions (e-tailers or omni-channel retailers). Furthermore, this substantial level of heterogeneity is further reinforced when considering other city characteristics, such as: 1) customer density (people/km²), 2) land prices, 3) road (time windows) or 4) other (supermarket dimension) regulations in place, etc.

In conclusion, one could say that the standard economic theory, no doubt, also applies to UFT; however, its practical use is much hampered both by the level of heterogeneity in this market, the complexity and articulation of market relationship among these actors and the non-negligible role interaction effects play in this specific sector. This points to the need of a public intervention to regulate UFT, especially because of the negative externalities it produces at a city level. Next section will present an overview of UFT externalities and the need for UFT planning to guarantee efficiency and sustainability of freight deliveries.

The Need for UFT Planning to Foster Sustainability

UFT Externalities

An efficient distribution system is of major importance for the competitiveness of a city and is, in itself, an important element of the urban economy (Browne and Allen, 1999). However, various negative externalities show that the effects of continued growth in freight transport is not sustainable in the long term, impacting on (1) planet, (2) people, and (3) profit. The extant freight transport and, in particular, UFT negatively impacts the planet because of the:

- pollutant emissions, including global (e.g. CO₂) and local pollutants (e.g. CO, NO_x, PM₁₀, VOC_s);
- use of non-renewable natural resources (e.g. fossil-fuel);
- waste products (e.g. tyres, oil and other materials).

Current UFT also negatively impacts on people due to the:

- physical consequences of pollutant emissions on public health (e.g. death, illness);
- injuries and death resulting from traffic accidents;
- increase in nuisance (e.g. noise disturbance, visual intrusion, stench, and vibration);
- reduction in quality of life elements (e.g. loss of greenfield sites and open spaces in urban areas due to transport infrastructure, decrease of attractiveness of a city centre);
- damage to buildings and infrastructure (e.g. Coliseum in the city of Rome).

The negative impacts UFT provokes on profit are ascribable to the:

- inefficiency and waste of resources;
- decrease in journey reliability and delivery punctuality, potentially resulting in lower quality of services to consumers;
- decrease in economic development (e.g. reduced market areas);
- congestion and decreasing city accessibility.

In order to have an idea of the effects produced by UFT, Table 1 summarizes some data taken from CIVITAS (2015).

Private and Public Coordination

UFT affects the community as a whole. However, most freight activity occurs outside of the public realm. The public sector is responsible for planning, owning, and maintaining infrastructure and to create the economic framework in which private entities operate by issuing various regulations. In other words, it is responsible for regulating/pricing UFT to optimize the allocation of scarce urban space and the externalities from transport. The private sector, in turn, largely makes operating decisions as well as company-specific investment decisions (National Academies of Sciences, Engineering, and Medicine, 2009). Public sector decision-making regarding UFT can influence or constrain the course of action and the future of freight transport. It should pay great attention to the many factors linked to competing demands, due to the heterogeneity characterising UFT (National Academies of Sciences, Engineering, and Medicine, 2015). Private sector decision-making is also driven by several factors, influenced by the need of companies to survive in a competitive marketplace, generate a return, and satisfy customers, while operating within a given legal/regulatory framework. Under this respect, public and private decisions overlap and interact in many areas, leading sometimes to cooperation, and other times to conflicts and inefficiencies.

The National Academies of Sciences, Engineering, and Medicine report on “Public and Private Sector Interdependence in Freight Transportation Markets” (2009) focused on projects involving both public and private sectors, deriving some key lessons for successful cooperation between public and private entities. Under this respect, it is important and critical to:

Table 1 Urban freight transport in numbers (retrieved from CIVITAS, 2015)

km travelled	10–15% of total km travelled
GHG emissions	≈ 6% of all transport-related GHG emissions
Workforce employed	2%–5% of the total workforce employed in urban areas
Urban land	3%–5% reserved to logistics activities
Flows	≈ 20–25% of freight_veh/km related to goods leaving urban areas, 40–50% to incoming goods and the remaining to internal flows
Delivery/pick-up per person per day	0.1
Delivery/pick-up per economic activity per week	1
Freight vehicle trips per 1000 persons per day	300–400
Tons per person per year	30–50

- build and maintain communication and cooperation among the many private and public stakeholders;
- educate the public sector on the benefits of freight projects to overcome any opposition;
- be aware of how a joint public and private process works;
- maintain key companies and officials who have undertaken an initiative;
- manage new multijurisdictional freight infrastructure projects through a governing agency with responsibility for the design and construction of the project;
- clearly identify the public and private project benefits to cement the desire for both sides to make a project work;
- make the public sector understand the private requirements for funding and the timing of financial flows to make public–private partnerships work better.

Complexity of Supply Chains

A supply chain consists of a system that includes activities, people, technologies, information and resources targeted at transferring a product or a service from the provider of raw materials to the end customer. In the last years, supply chains have become more complex and dispersed due to ever-lower transportation costs as firms along the chain substitute transport for lower cost inputs and higher product velocity (National Academies of Sciences, Engineering, and Medicine, 2013). Freight traffic is particularly growing in cities for different reasons, namely: (1) decline of inventory holding and increase in both the number/variety of goods sold and express transport services; (2) increase in logistics facilities in large metropolitan cities, since modern supply chains require close proximity to consumer markets and trans-shipment services; (3) warehouses and distribution increasingly decentralized to the periphery of metropolitan areas, which leads to increased truck travelled kilometres.

Since supply chains are highly complex and influenced by many different factors, efforts to change them are challenging. At the urban level, important attributes of a supply chain relate to delivery frequency, time and size, and to the possibility of consolidation with other product supply chains (Danielis et al., 2012).

Supply chain also needs coordination, both in terms of long-terms and short-term contracts among different partners, or vertical coordination. A lack of coordination could lead to inefficiency linked to inaccurate forecast, low capacity utilization, high inventory costs, inadequate customer service, bad quality of service and customer satisfaction.

Understanding the underlying mechanisms of supply chains and adopting a supply chain approach is fundamental to understand how urban freight distribution works and what will be the impact of the various potential urban transport policies.

UFT Planning

Before focusing on how public decisions should be made to foster sustainable and efficient UFT, it can be useful to trace back some fundamental aspects linked to transport planning and policy-making. Under this respect, transport planning has traditionally focused on passenger transport, often neglecting freight issues. This lack of attention can be ascribed to a scarce understanding of freight transport, even when local authorities acknowledge its importance for the local economy (Le Pira et al., 2017b). On the contrary, passenger transport solutions mostly rely on solid knowledge and on widely available data. Passenger and freight transport are also different in terms of stakeholders involved. Stakeholders related to passenger transport typically pertain to the public domain, i.e. public transport users, policy-makers, citizens. Vice versa, freight transport mainly involves the private interests of retailers, shippers and transport providers, determining an imbalance due to their economic interests (Gatta and Marcucci, 2016). Many of the challenges impacting on the freight system (e.g. congestion, land use conflicts, community acceptance) typically arise in metropolitan areas (National Academies of Sciences, Engineering, and Medicine, 2015).

Notwithstanding the evident importance of policy-making and knowledge related to UFT, research and several innovation initiatives have been carried out only in the last years. In particular, several projects have proposed solutions to tackle the problems caused by urban freight deliveries (see, e.g. CIVITAS I, II, PLUS, PLUS II), while others have been devoted to collecting and deploying UFT best practices (see, e.g. BESTUFS I, II, BESTFACT, TIDE, SUGAR). Nevertheless, research is still needed to reduce the lack of knowledge and understanding of UFT processes, commercial dynamics and behaviour, which are among the main obstacles to the achievement of sustainable and efficient UFT (Gatta et al., 2017). Besides, a fair number of UFT-related programmes has been characterised by a non-negligible failure rate. This is mainly attributable to the insufficient commitment from relevant stakeholders, which can be related to a lack of incentives (Le Pira et al., 2017a). Under this respect, it is fundamental to actively involve stakeholders from the beginning of the policy-making process, by making them analyse the problem, assume (if any) responsibilities for negative externalities in the logic of “all breakages must be paid for”, and discuss on possible solutions. By doing this, it is likely that decisions will be stakeholder-driven and, thus, more acceptable. There are different ways to involve stakeholders. For example, Marcucci et al. (2017a) describe a successful initiative based on a new collaborative governance model, where public and private stakeholders are appropriately engaged, which proved to be both efficient and effective, since they all felt involved in decision-making resulting in a specific contract where they undertook to respecting the duties assumed during the negotiation phase. Another example is reported in Gatta et al. (2019a) where an interactive multi-criteria approach is employed to properly plan with stakeholders alternative off-hour delivery solutions.

UFT policies can be classified according to specific issues/problems. This suggests why no easy solution is at hand and cannot be seamlessly employed in different cities/contexts. Additionally, no single policy can address and solve all UFT problems, while an integrated policy package approach is needed (Gatta and Marcucci, 2014).

Under this respect, the planning guide proposed by the National Academies of Sciences, Engineering, and Medicine (2015) is intended to introduce a complete catalogue of UFT initiatives that could be considered by public agencies. Starting from the review of more than 150 references, 54 initiatives were selected and classified into eight major groups, namely: (1) infrastructure management - improve infrastructures to enhance freight mobility and make it adequate to current market needs; (2) parking/loading areas management - plan and design both on-street and off-street parking/loading to increase service efficiency and reduce congestion; (3) vehicle-related strategies - implement technologies and programs, such as emission standards and low-noise deliveries, to reduce negative externalities produced by freight vehicles; (4) traffic management - carry out traffic engineering and control measures to improve traffic conditions; (5) pricing, incentives, and taxation - adopt monetary policy interventions to reach the targets set by the public sector; (6) logistical management - adopt strategies to facilitate cargo consolidation, intelligent transport systems and last-mile practices; (7) freight demand/land use management - modify the nature of freight demand and apply land use policy by both relocating large traffic generators and including freight issues in the urban land use planning process; (8) stakeholder engagement - increase the understanding of freight issues within the public sector and effectively involve the private sector to secure commitments to common solutions. It is worthy of notice that the effectiveness of these strategies depends on the specific context as well as on the mode of implementation and, even more important, that a combination of different strategies in policy packages has much more potential for success.

A first logical distinction is between supply and demand initiatives, the former linked to infrastructure management, the latter to freight demand and land use management. However, one should bear in mind that it is often practically difficult to classify a given action to either supply or demand given the interaction among stakeholders and considering the implications each policy implemented might have.

Another important policy principle is that public sector interventions are only justified when a market failure is preventing the economy from reaching its most efficient outcome. Under this respect, while a well-designed public sector intervention could lead the system to greater economic efficiency, it could also make things worse if there is no market failure, causing a “government failure” (Holguín-Veras et al., 2014). In the specific case of UFT, since market failures are always present, it is thus important to study how the public sector should intervene. When deciding how to address the problem, city agencies must carefully take into account this risk, and consider the potential impacts of the solution. Some examples are related to time-access restrictions for all trucks, which could lead to an increase of both carrier costs and freight traffic externalities (Quak and de Koster, 2009) or zero tolerance policies against parking violations, which may require large enforcement expenses, not always compensated by the benefits in congestion reduction (Holguín-Veras et al., 2014).

In any case, planning for sustainable and efficient UFT should account for behavioural issues and stakeholder engagement processes (Marcucci et al., 2017b). A deep understanding of the behavioural aspects linked to stakeholder decision-making is crucial to define and implement effective policies or solutions, capable of accommodating both private and public objectives. Along this line, Marcucci et al. (2015) investigate, at a strategic level, stakeholders preferences for alternative parking and pricing policies and highlight the need for a collaborative approach between local policy makers and transport providers (e.g. living lab) so to avoid “decide and defend” strategies. Marcucci et al. (2018) propose an advanced user-centred approach to appropriately design gamification and maximize its probability of success in fostering engagement and behaviour change in UFT. Gatta et al. (2019b), explore people acceptance towards public transport-based crowdshipping, one of the most promising logistics solutions dealing with the requirements of the on-demand economy, by performing a joint supply and demand analysis.

Urban freight policies/solutions should, therefore, emerge from a collaborative/participatory approach where new behavioural analyses should complement technical analyses, by considering both solution feasibility and acceptability. This important and understudied topic will be discussed in the chapter on “Behavioural Research in Freight Transport” of the Encyclopedia of Transportation.

Conclusion

Economics of urban freight is multifaceted due to the high level of heterogeneity in this market, the complexity and articulation of market relationship among these actors and the non-negligible role interaction effects play. An efficient distribution system is of major importance for the competitiveness of a city. However, various negative impacts show that the effects of continued growth in freight transport is not sustainable in the long term. Economic efficiency and sustainability of UFT solutions can be guaranteed only via a public intervention and appropriate planning processes. Under this respect, private-public coordination based on a proper stakeholder engagement process is a key issue. It will be effective only if supported by appropriate incentives for stakeholders, who should understand and bear (if any) responsibility for the negative UFT externalities produced. Besides, technical and behavioural analyses, together with a good knowledge of supply chains at a city level and of the variety of initiatives that one could adopt, are fundamental to plan for sustainable and efficient UFT.

Acknowledgment

We would like to thank all those who have collaborated in the activities performed by the Transport Research Lab at Roma Tre University (TRELab – www.trelab.it) that supported most of the research at the basis of the present paper.

Biographies



Edoardo Marcucci is Full Professor of Transport Economics at both Molde University College and Roma Tre University, where he is co-Director of *Transport Research Lab*. He is currently co-Editor in Chief of Research in Transportation Economics. Author of several articles published in international top-journals, his research interests mainly focus on urban freight distribution, stated preference, discrete choice modeling, and interaction effects in group decision making. He has been involved in several international, national research projects and evaluation committees.



Valerio Gatta is Researcher of Transport Economics and co-Director of *Transport Research Lab* at Roma Tre University. He has an extensive research experience in innovative transport solutions, especially in urban freight, linked to decision making processes and sustainability, with particular reference to stated preference survey designs and discrete choice modelling techniques. Advanced methods and models for policy acceptability and behaviour change analysis are at the core of his academic path.



Michela Le Pira is Researcher of Transport Planning at the University of Catania (Italy). She holds a PhD in Transport Engineering. As a PhD student, she focused on public participation in transport planning supported by agent-based modelling and multi-criteria decision methods. As a Post-doc researcher at University of Roma Tre, she worked on participatory decision-support methods for sustainable urban freight transport policy-making. As a research fellow at the University of Catania, she is currently investigating new mobility services for both passengers and freight.

References

- Browne, W., Allen, J., 1999. The impact of sustainability policies on urban freight transport and logistics systems. In: Meesman, H., Van De Voorde, E., Winkelmanns, W. (Eds.), *World Transport Research Vol 1: Transport Modes and Systems*. Elsevier, Oxford, pp. 505–518.
- Button, K.J., Pearman, A.D., 1981. *The Economics of Urban Freight Transport*. Springer.
- CIVITAS, 2015. Smart Choices for Cities Making Urban Freight Logistics More Sustainable. Available from: https://civitas.eu/sites/default/files/civ_pol-an5_urban_web.pdf.
- Danielis, R., Maggi, E., Rotaris, L., Valeri, E., 2012. Urban Supply Chains and Transportation Policies. Working Paper SIET 2012. Available from: http://www.sietitalia.org/wpsiet/Rotaris%202012_1.pdf.
- Gatta, V., Marcucci, E., 2014. Urban freight transport and policy changes: Improving decision makers' awareness via an agent-specific approach. *Trans. Policy* 36, 248–252.
- Gatta, V., Marcucci, E., 2016. Stakeholder-specific data acquisition and urban freight policy evaluation: evidence, implications and new suggestions. *Trans. Rev.* 36 (5), 585–609.
- Gatta, V., Marcucci, E., Le Pira, M., 2017. Smart urban freight planning process: integrating desk, living lab and modelling approaches in decision-making. *Eur. Trans. Res. Rev.* 9 (3), 32.
- Gatta, V., Marcucci, E., Delle Site, P., Le Pira, M., Carrocci, C.S., 2019a. Planning with stakeholders: analysing alternative off-hour delivery solutions via an interactive multi-criteria approach. *Res. Transport. Econ.*, <https://doi.org/10.1016/j.retrec.2018.12.004>.

- Gatta, V., Marcucci, E., Nigro, M., Serafini, S., 2019b. Sustainable urban freight transport adopting public transport-based crowdshipping for B2C deliveries. *Eur. Trans. Res. Rev.* 11 (1), 13. <https://doi.org/10.1186/s12544-019-0352-x>.
- Holguin-Veras, J., Wang, C., Browne, M., Hodge, S.D., Wojtowicz, J., 2014. The New York City off-hour delivery project: lessons for city logistics. *Proc. Soc. Behav. Sci.* 125, 36–48.
- Le Pira, M., Marcucci, E., Gatta, V., Ignaccolo, M., Inturri, G., Pluchino, A., 2017a. Towards a decision-support procedure to foster stakeholder involvement and acceptability of urban freight transport policies. *Eur. Trans. Res. Rev.* 9 (4), 54.
- Le Pira, M., Marcucci, E., Gatta, V., Inturri, G., Ignaccolo, M., Pluchino, A., 2017b. Integrating discrete choice models and agent-based models for ex-ante evaluation of stakeholder policy acceptability in urban freight transport. *Res. Transport. Econ.* 64, 13–25.
- Marcucci, E., Gatta, V., Le Pira, M., 2018. Gamification design to foster stakeholder engagement and behavior change: an application to urban freight transport. *Transport. Res. A Pol. Pract.* 118, 119–132.
- Marcucci, E., Gatta, V., Marciani, M., Cossu, P., 2017a. Measuring the effects of an urban freight policy package defined via a collaborative governance model. *Res. Transport. Econ.* 65, 3–9.
- Marcucci, E., Gatta, V., Scaccia, L., 2015. Urban freight, parking and pricing policies: an evaluation from a transport providers' perspective. *Transport. Res. A Policy Pract.* 74, 239–249.
- Marcucci, E., Le Pira, M., Gatta, V., Inturri, G., Ignaccolo, M., Pluchino, A., 2017b. Simulating participatory urban freight transport policy-making: Accounting for heterogeneous stakeholders' preferences and interaction effects. *Transport. Res. E Logist. Transport. Rev.* 103, 69–86.
- National Academies of Sciences, Engineering, and Medicine, 2009. Public and Private Sector Interdependence in Freight Transportation Markets. The National Academies Press, Washington, DC <https://doi.org/10.17226/14285>.
- National Academies of Sciences, Engineering, and Medicine, 2013. Synthesis of Freight Research in Urban Transportation Planning. The National Academies Press, Washington, DC <https://doi.org/10.17226/22573>.
- National Academies of Sciences, Engineering, and Medicine, 2015. Improving Freight System Performance in Metropolitan Areas: A Planning Guide. The National Academies Press, Washington, DC <https://doi.org/10.17226/22159>.
- Ogden, K.W., 1992. Urban Goods Movement: A Guide to Policy and Planning. Ashgate Pub.
- Quak, H.J., de Koster, M.B.M., 2009. Delivering goods in urban areas: How to deal with urban policy restrictions and the environment. *Transport. Sci.* 43 (2), 211–227.
- Taniguchi, E., Thompson, R. G., Yamada, T., 1999. Modelling city logistics. In: Taniguchi, E., Thompson, R.G. (Eds.), *City Logistics I: 1st International Conference on City Logistics*. Institute of Systems Science Research, Kyoto, pp.3–37.
- UN, 2019. World Urbanization Prospects 2018: Highlights. United Nations, Department of Economic and Social Affairs, Population Division, New York.

Further Reading

- Cherrett, T., Allen, J., McLeod, F., Maynard, S., Hickford, A., Browne, M., 2012. Understanding urban freight activity - key issues for freight planning. *Journal of Transport Geography* 24, 22–32.
- Dablanc, L., 2007. Goods transport in large European cities: Difficult to organize, difficult to modernize. *Transportation Research Part A: Policy and Practice* 41 (3), 280–285.
- Marcucci, E., Gatta, V., 2017. Investigating the potential for off-hour deliveries in the city of Rome: Retailers' perceptions and stated reactions. *Transportation Research Part A: Policy and Practice* 102, 142–156.
- Sanchez-Diaz, I., Browne, M., 2018. Accommodating urban freight in city planning. *European Transport Research Review* 10 (2), 55.
- Stathopoulos, A., Valeri, E., Marcucci, E., Gatta, V., Nuzzolo, A., Comi, A., 2011. Urban freight policy innovation for Rome's LTZ: A stakeholder perspective. *City Distribution and Urban Freight Transport: Multiple Perspectives* 75–100.

Incentives in Public Transport Contracts

Andreas Vigren, Stockholm, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	332
Contracting	332
Objectives and Targets for the Incentive	333
Monetizing and Pricing	333
Monitoring and Credibility	334
Operator Risk Preference and Adverse Effects	335
Summary	335
See Also	336
Further Reading	336

Introduction

Incentives of various sorts are present everywhere in the transport sector and other industries. They intend to affect behavior on some party so that some action is taken, or an outcome achieved. Incentives will always have the potential to create both positive and negative effects, and an important task is to balance these with respect to moral hazard and information asymmetries. Furthermore, with a public buyer, incentives would try to align the society's marginal cost for some action with the private seller's private marginal cost. In this chapter, incentives related to public transport contracts are discussed, although aspects touched upon here can well be generalized to other contexts.

The point of departure is a buyer-seller relationship of a public transport service. Two parties will be involved in a business relationship ruled by a contract. The buyer will be referred to as the public transport authority (PTA), who wants to maximize social welfare, and the seller as the (profit-maximizing) operator. The discussion will mainly revolve around incentives that are explicitly and on purpose built into the contract to make the operator strive towards some operational or societal target. This is done because the PTA, for example, through competitive tendering has left some control of the traffic over to a private operator. The incentives discussed will also give some monetary consequence; either through the production revenue, or as penalties and rewards. Alternative ways of incentivizing exist, for example, contract termination, contract extension, or a reputational mechanism that affects the operator in future procurements. However, monetary incentives are the focus here. Furthermore, the chapter's focus will not be on more indirect incentives stemming from, for example, the choice of procuring a service, different contractual forms, or the effects institutional settings and regulations can have. Some of these issues are covered in other chapters of this book.

Contracting

The contract terms govern the responsibilities of the parties for service to be run, not seldom up to 10–15 years. Contracts are mostly incomplete in the sense that they do not cover all possible events and scenarios that might occur. The contractual parties' objectives might differ and, not seldom, be in conflict with each other. Put very simply, the PTA is trying to achieve high social welfare, which often relates to some level and definition of service quality, and the operator wants to make profits, meaning it could cut costs by reducing service quality below the PTA's expected level. This means that when the PTA buys a service, it will have a need to affect, or control, the operator to some extent so that its behavior will align with the PTA's objectives. The PTA could do so by stating very detailed in the contract how the operator must run the service. This gives more control, but will also limit the operator's possibilities to operate efficiently, thus risking that the service becomes poor. One could also question, given a detailed contract, the need for the PTA to buy the service altogether as an in-house operator run by the PTA could give more control over the service. However, the PTA have good reason to buy the service from commercial operators as these are often more efficient than public in-house regimes, thus producing more public transport for the same money. By making operators compete for running the service, the PTA can benefit from efficiency gains and innovative pressure created through the competitive environment, in which the commercial operators exist. The task for the PTA is then to incentivize the operator to align with its societal objectives, which could be viewed as a try to maximize social welfare.

Arguably, in contracts where the consequences of an incomplete contract are minor and quality is easy to infer and verify over the contract period, incentives are less important to ensure a good service delivery. However, the more complex the contracting situation and the bigger consequences poor quality has on the service, objectives must be aligned. One or more well-designed incentives is one way of accommodating this.

Objectives and Targets for the Incentive

The PTA must determine what societal objectives are important for the service, and which one(s) the incentive should focus on. One objective could, for example, be a high market share for public transport, high environmental standards, or high service quality. However, these are subjective objectives. The PTA and the operator might not share the definition on what is service quality, or what is *high* service quality, as it can range a wide array of factors: travel time reliability, cleanliness, safeness, brand perception, easiness of use, and much more. Even with a subjective objective, there is a need to formulate a more concrete target for the incentive with levels the operator must fulfill or some focus it should have. An example of the former could be that the operator must complete at least 95% of the departures each month, and an example of the latter is that the PTA establishes a focus for the operator to increase public transport ridership. The target forms the incentive facing the operator and serves as a proxy for the objective, but is sometimes hard to formulate and can risk not serve as a perfect proxy for the objective. Thus, this is a balancing act. Depending on the objective, more than one target could be needed. The target can be defined only by the PTA, or in collaboration with the operator or public transport industry.

The target must be verifiable and possible to monitor, and it must affect the relevant societal objective. Furthermore, it is important to consider what measures the operator can use to reach or affect the target. Preferably, the operator should be able to take as direct measures as possible. If only indirect measures can be used, the PTA cannot expect very good target fulfillment. The PTA must therefore consider what measures it should leave to the operator to affect the service, and thus the target. This could involve letting the operator own the vehicles, be in charge of the traffic supply, or set fares; all depends on the target selected. Many measures might even lie outside the control of both parties, and can thus not necessarily be contracted easily. However, the more control the PTA hands over to the operator, naturally it loses control itself. It is therefore necessary that the PTA is aware of the power it is giving up. For example, would the power of determining supply be left fully to the operator in order to increase ridership, the PTA cannot reduce supply to affect the route network or cope with troubles in public finances. On the other hand, if the operator cannot use some, or any, efficient measures to affect the target, one could question the need and power of the incentive all together. The PTA can have good reason to trust the operator with controlling measures as more efficient planning or innovative solutions could arise, which would benefit the traffic and service quality. Also, the degree of control exhibited by either party is, of course, not binary. One could easily imagine a situation where the parties collaborate on the supply to be run, for example, by letting the PTA define a minimum service supply. But again, this will affect the operator's possibilities to affect the target.

The possibility to predict the future development of a measure, target, or the service as a whole is also important to consider. Predicting, for example, ridership up to 10 years into the future is hard for either party, which causes uncertainty and risk. Therefore, risk should ideally be placed on the actor that is best suited to manage it.

The difficulties in aligning the PTA's and operator's objectives through targets will also depend on the level of trust between the parties. With more trust, the operator will arguably be more inclined to conform more to the PTA's priorities, although probably not fully. That could channel through, for example, better cooperation in planning the traffic or resolving disputes more easily, which could make the incentive design process easier. Less trust would, in contrast, probably involve more detailed contracts and incentives, more adverse effects, and a potentially worse outcome for both parties. Procurement legislation will also restrict how much trust can be utilized in the contract, not the least because of the principle of equal treatment in procurements.

Monetizing and Pricing

The incentive must affect the operator's revenue in some way, or it will not be prioritized. This could be done with varying complexity. Typically, the incentive would affect revenue through the production revenue function, or as a reward or penalty "on top" of the revenue function, and relate to the fulfillment or focus of the target. These could be combined, as well as used separately. However, the way the incentive is monetized will, arguably, not matter for a rational operator as it will calculate the expected value of all matters affecting its profits, which in turn gives the cost of service for the PTA.

The production revenue function, as referred to here, is the revenue from some output to the transport service produced by the operator. A typical contract would contain a combination of a fixed sum, per kilometer, and per hour revenue, and the operator's risk here lying primarily in the cost of running the bus. The revenue function needs, however, not only be linked to supply related output measures, but also to others such as demand-related ones meaning that the revenue components related to supply could be coupled, or entirely replaced, with a revenue per boarding passenger or passenger kilometer. A demand related revenue component could serve as an incentive to increase ridership and help steer toward some societal objective, for example, improving the market share of public transport, but would also impose demand risk on the operator in addition to the production risk. It also illustrates how the production revenue function design can, implicitly, incorporate incentives, and risk, whether or not the PTA is aware of that.

Another way of monetizing an incentive is to put a revenue on top of the production revenue function. Put simply, the incentive can be designed to give the operator a reward (carrot, or bonus) or penalty (stick, or fine). In theory, there should be little difference between a penalty or reward; a fully informed operator will, again, internalize both. However, the selected design could have a signaling effect. Because a negative incentive is like taking away money from the operator when it does not meet the target, it could be regarded as harsher than a reward. A reward could, on the other hand, be perceived as an opportunity for the operator to earn additional money on top of other revenue.

Irrespective of monetizing the incentive through the production revenue function or by a reward/penalty, the incentive must be priced. The PTA could choose the price itself, let the operator bid on the price, or use a combination of the two. The price must not be so low that the operator disregards the incentive. That is, if the incentive does not affect the financial situation for the operator in the contract, it will not steer toward the target. One reason for this could be poor cost-recovery for the measures it needs to use to affect the target. A too high price, on the other hand, could make the operator “care too much” about the target so that other parts of the contract is neglected, causing adverse effects. A very high price might also put the operator in financial difficulties when performing poorly on the target.

Ideally one would want to use the society's valuation of improving/reducing the target with one unit. However, this value is probably hard to infer. But in determining the price of the incentive, the PTA must still consider its marginal valuation with respect to the incentive. Differentiating the price over, for example, time and route might help fulfilling the objective and prioritize activities to the right places, but might also impose risk by making the contract more complex and monitoring more expensive.

Monitoring and Credibility

Monitoring the operator's performance is necessary to measure and verify the incentive and what financial consequences should face the operator. In essence, it is about verifying quality and contract fulfillment of both parties. Arguably, quantitative targets are easier to monitor, and the more concrete these are the more reliable and cheaper the monitoring is likely to be (of course depending on the technological requirements). One downside is, however, as noted earlier, that it is potentially not a good proxy for the societal objective.

Monitoring should be agreed upon before the contract starts. That way, there is no ambiguity in, for example, how measurements are made, how often, and with what technology. This is also an important aspect when choosing the proper target as monitoring should not be more extensive than it needs to with respect to evaluating the fulfilment of the target and the degree of trust between the operator and PTA. With high trust, the PTA might even be able to rely on self-reported information from the operator, or the other way around. One quickly realizes, however, that this creates an information disadvantage for the party that is the receiver of the information, something that might not play out well in the long term if the reporting party at some point in time decides to cheat. Thus, to make sure that both parties honor the contractual agreement, good monitoring is necessary.

Monitoring is often costly. Data collecting equipment, inspections, and administrative work will increase costs and decrease economic efficiency. Furthermore, the less is agreed upon on beforehand, the harder it gets for the PTA to, with credibility, verify performance. However, the technological advancements made until today should give good opportunity to achieve better outcomes on this through more automated (routine-based) solutions, which could also have the potential to lower administrative costs. More routine-based monitoring also gives better transparency on what and how things are measured, as less ambiguity is feeded into the system. It could also be the case that the relevant data for measuring the target is already being collected through passenger counts (e.g., smart card validations) or real-time info, meaning that the monitoring capacity is already in place.

With the monitoring in place comes the task for the PTA to actually pay, or penalize, according to the conditions in the contract. This is a crucial part for making the incentive work. Even with proper monitoring and a clear incentive, for example, that every departure delayed with more than 5 min at its final stop will result in a penalty, conflicting views and situations will arise. The easier ones to resolve are probably those where the cause of the incident is in some control of some of the parties. If a bus breaks down due to bad maintenance by the operator, the delay is caused by the operator who should face the financial consequences. However, many situations are outside the control of the contractual parties. For example, roadworks next to a bus stop is not an uncommon issue and could easily cause delays, especially in urban environments. The delay would here still be on the operator, but it has not been the cause of the delay and could, arguably, not have done anything about it. However, the PTA have not received the quality required and should have the right to act upon that in accordance with the contract. To avoid this situation, the parties could have agreed in the contract that roadworks are not causing penalties for delays, but there exist many similar situations. Again, listing them would quickly become tedious and make monitoring more expensive. Instead, trust between the parties could benefit the situation. One possible scenario, if the PTA for some reason would choose not to levy the financial consequence, is a negotiation process between the parties to resolve the situation. Such negotiations would, however, also risk bringing more ambiguity to the contract relation, and becomes complex if the PTA needs to make similar negotiations with different operators with different levels of trust. The credibility of the PTA, and the incentive, could be at stake. Especially for penalties, the lower the incentive's share of the PTA's total expenses, arguably its motivation for to levy the penalty is lower because of low cost-recovery of the execution. A paradoxical situation could also arise when an operator is performing badly because of a bad financial situation, and the reduced quality results in high penalties. The PTA wants improvement and enforces the incentive to achieve this, but levying the penalty could risk causing bankruptcy and terminate the service all together. This puts the PTA in a tough spot. Qualification criteria in the awarding process to sort out poor operators in the contract awarding stage can be one way of avoiding such a situation.

Enforcing the contract and being credible is thus essential for the PTA, as it otherwise risks that operators will not respond to the incentive. That would also risk the targets and societal goals set up. This issue is especially important when there is competition for running the service. A PTA that does not fully enforce contracted incentive will impose information asymmetry in the bidding process. Exemplifying with a penalty, because full information amongst all parties is crucial in the bidding stage, less risk-averse operators could leave out parts of its expected penalties and place a lower bid (given a lowest-price auction) if they know that the PTA is reluctant to impose penalties. This could also serve as an entry barrier for operators who are not doing business with the PTA as

they do not have this knowledge thus making them less probable to win the procurement. All this have the potential of reducing competition and lower quality.

Operator Risk Preference and Adverse Effects

The overall design of the incentives will affect both the operator's behavior in the contract, and how it values the contract itself, as noted earlier. The latter is channeled through, in the case of a procurement, the bidding process and will affect the PTA's cost for the contract. When deciding what to bid for the contract, a rational operator will take into consideration the full incentive design and what measures it can use to affect the target. In the case of a penalty incentive, to cope with the expected penalties the operator will incur during the contract term, it will add a risk premium. A reward or a component in the production revenue function would be treated analogously. The greater share of the total revenues, and the more risk or uncertainty, is associated with the incentive, the larger this premium will be.

Contractual relations will create some adverse, and sometimes perverse, effects. Incentives can both accommodate and accelerate these, as well as creating new ones, and an important part of designing contracts lies in minimizing and understanding those effects. However, risk assessment is often a complex and costly task to take on for both parties. For this reason, a very low-powered incentive will probably affect the bid and performance only marginally as the operator is not prepared to make a full-on analysis because that the incentive does not affect profits. This also means the incentive is probably not very powerful either. However, the more profits are affected by the incentive, arguably, the operator will spend more resources on analyzing the risks associated with it. And the more efficient measures the operator can use to affect the incentive, the smaller the relative risk premium is likely to be.

The incentive is intended to add to the operator's behavior. For example, if its revenue is only tied to demand and the operator is paid per boarding passenger, ridership will certainly be a strategic planning variable, albeit not necessarily used in the way the PTA sought to. With a poorly designed incentive, adverse effects could arise whose consequences are not the least dependent on the price and the degrees of freedom the operator has in the contract. For example, if the operator has full control over supply planning, which is not unlikely as supply is an important measure to affect demand, this demand-related incentive could lead to shorter routes and more transfers, which might not favor the overall transport system and reduce social welfare. This was an unfortunate and unforeseen consequence of the initial contract design in Santiago de Chile's Transantiago reform, which was very focused on catering demand. The contract was re-negotiated three times to accommodate the adverse effects of the incentive mechanisms in the contract.

Some incentives are such that high target fulfillment have few apparent negative effects. However, an incentive that on its own serves to improve on specific targets can, in combination with other incentives, create a perverse incentive. For example, it is not uncommon to penalize operators for cancelled departures. Furthermore, to ensure travelers meet an integrated and recognizable public transport system, operators are sometimes penalized when running vehicles with a coloring scheme different from the PTA's. In this case of a contract that uses both incentives, a perverse incentive is created when the penalty for cancelling a departure is smaller than running the departure with a wrong colored vehicle, which is indeed the situation in some Swedish bus contracts. Although not the intention, if a rational operator temporarily have access only to a vehicle that, for some reason, has the wrong coloring, the operator would rather cancel the departure than running it. In this example, the society would have been better off with this departure running, irrespective of bus color. However, it is not necessarily apparent which, if any, of the incentives should be revoked or adjusted. Here, trust between the parties can be beneficial so that similar situations are avoided, or the incentives be adjusted to accommodate the situation.

Summary

This chapter has discussed some important aspects to consider when constructing incentives in public transport contracts. The contractual parties' objectives will not always be aligned, and using incentives is one way of accommodating this discrepancy to achieve good service quality. The level of trust between the parties is important to consider when designing the incentive. The incentive must be powerful enough to make the operator change its behavior in line with the PTA's objectives, and the PTA must also make sure the operator has the proper measures to affect the incentive's target. Similarly, the PTA needs to be aware of what control it gives away, what implication this has for its remaining possibilities to control the public transport system, and what potential adverse effects can arise following the incentive. The PTA must also make sure it is credible when it comes to enforcing the incentive. For that, proper monitoring is important. Ideally, the PTA should also evaluate the incentives it introduces, in order to infer whether they are effective or not. This is something that puts high requirements on data availability and adequate evaluation strategies. Examples of the latter includes introducing the incentive (randomly) at different points in time for different lines, and/or differentiating the price of the incentive across the same dimensions.

In addition to incentives, the PTA has other ways of ensuring quality in the contract. Ex-ante qualification criteria can help sort out poor operators, thus improving service quality, and reputational mechanisms, where the operator's performance during the contract term affects its chances of winning the next contract, could increase the operator's efforts to deliver good quality. One must also acknowledge that some tools of improving public transport lies outside the contract. Depending on the institutional settings, decisions on infrastructure, taxes, or environmental standard might lie outside the PTA's and operator's domain, thus making it necessary to collaborate with other parties than just the operator.

See Also

Operation Costs for Public Transport; The Value of Security, Access Time, Waiting Time, and Transfers in Public Transport; Demand for Passenger Transportation; Public Transport Fare and Subsidy Optimization; Procurement of Public Transport: Contractual Regimes; Regulation and Competition in Railways; Contract Efficiency in Public Transport Services

Further Reading

- Albano, G., Calzolari, G., Dini, F., Iossa, E., Spagnolo, G., 2006. Procurement contracting strategies. In: Dimitri, N., Piga, G., Spagnolo, G. (Eds.), *Handbook of Procurement*. Cambridge University Press, Cambridge, pp. 82–120.
- Bajari, P., Tadelis, S., 2006. Incentives and award procedures: competitive tendering vs. negotiations in procurement. In: Dimitri, N., Piga, G., Spagnolo, G. (Eds.), *Handbook of Procurement*. Cambridge University Press, Cambridge, pp. 121–140.
- Fearnley, N., Bekken, J.T., Norheim, B., 2004. Optimal performance-based subsidies in Norwegian intercity rail transport. *Int. J. Transp. Manag.* 2 (1), 29–38.
- Gómez-Lobo, A., Briones, J., 2014. Incentives in bus concession contracts: a review of several experiences in Latin America. *Transp. Rev.* 34 (2), 246–265.
- Hensher, D.A., Stanley, J., 2008. Transacting under a performance-based contract: the role of negotiation and competitive tendering. *Transportation Research Part A: Policy and Practice* 42 (9), 1143–1151.
- Hensher, D.A., Yvrande-Billon, A., Macário, R., Preston, J., White, P., Tyson, B., Orrico Filho, R.D., 2007. Delivering value for money to government through efficient and effective public transit service continuity: some thoughts. *Transp. Rev.* 27 (4), 411–448.
- Hensher, D. A., Ho, C., Knowles, L., 2016. Efficient contracting and incentive agreements between regulators and bus operators: the influence of risk preferences of contracting agents on contract choice. *Transp. Res. Part A Policy Pract.* 87, 22–40.
- Laffont, J.J., Tirole, J., 1993. *A Theory of Incentives in Procurement and Regulation*. MIT Press, United States.
- Pyddoke, R., Lindgren, H., 2018. Outcomes from new contracts with “strong” incentives for increasing ridership in bus transport in Stockholm. *Res. Transp. Econ.* 69, 197–206.
- Stanley, J., van de Velde, D., 2008. Risk and reward in public transport contracting. *Res. Transp. Econ.* 22 (1), 20–25.
- Taylor, B.D., Fink, C.N.Y., 2013. Explaining transit ridership: what has the evidence shown? *Transp. Lett.* 5 (1), 15–26.

Transportation Improvements and Property Prices

Alex Anas, Department of Economics, State University of New York at Buffalo, New York, NY, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	337
Firms Located Relative to an Export Hub	338
Residential Location in an Economy Open to Population	339
Residential Location in an Economy Closed in Population	341
Transport Improvement in the Monocentric City	342
Congestion and Congestion Pricing	344
The Empirical Evidence	344
Before and After Comparisons	344
Hedonic Regression Analysis	344
Market Equilibrium Simulations	345
References	346

Introduction

Transportation improvements reduce the monetary or time costs of transportation. Improvements arise from technological advances in transport technology or from the expansion of transport infrastructure capacity; or come about because transport congestion is reduced, or from the pricing of transportation such as congestion pricing, gasoline taxation, parking taxes, or public transportation fares. There can also be improvements in the reliability of travel times, in the rates of traffic accidents or in the level of pollution and noise associated with transportation. In this article we will focus solely on how improvements in the monetary cost and time of travel affect property values.

The transport and property values connection has drawn the attention not only of urban economists, but also of urban planners, real estate professionals, and transportation planners and engineers. There are several reasons why it is useful to know how and how much property prices change because of transport improvements:

1. Aggregate property values are an important tax base, so changes in values affect the fiscal health of urban or suburban communities. For example, in the United States, property taxes are most often used to finance public schools.
2. Increases in property values can be potentially captured by special property tax levies, becoming revenues that can defray the cost of making the transport improvements that caused the property values to increase in the first place.
3. Predictions of changes in property values can indicate profitable investments in real estate markets surrounding a transportation project.
4. Changes in property prices are a part of the nonuser benefits of transport improvements and thus they are an important component of a cost-benefit analysis of a transportation project or of transport policy.

Of the above, Reason 4 is the most important to economists and deserves more comment here. It is often assumed that increases in property prices caused by a public transport investment equal the full benefits of the investment, but—as we shall see—this holds only when the utility levels of the consumers are not altered by the transport investment. More generally, we will see that a transport improvement can cause property prices to rise, to remain unchanged, or to fall. When prices rise (fall), consumers of property lose (benefit) while property owners benefit (lose). The cost-benefit analysis of transport improvements need to take into account the consumer surplus change experienced by households, the change in profits experienced by firms and the change in property prices accruing to property owners.

Noneconomists have often misconstrued the effect of transportation on property prices. For example, Melvin Webber, a Berkeley urban planning professor ruminating about the effect of the newly established BART (Bay Area Rapid Transit) system on real estate values, made a bold claim in 1976:

“ . . . land economists agree that rising values in one location will be about equally matched by declines elsewhere, rather as levels in an air mattress rise and fall as one section or another is squeezed or released. Unless the whole metropolitan economy has been caused to expand, the total of land values will remain essentially constant.”

(Webber, 1976)

Economic theory finds no reason for Webber’s “air mattress theory” and the empirical evidence does not support it. Land economists do not agree now nor did they agree back in 1976 or at any time that changes in aggregate property values caused by transport improvements are essentially zero unless the urban economy expands. Fifteen years before Webber’s conjecture, Mohring (1961) had proposed a simple but economically sound model of the real estate market’s connection with transportation. In

Mohring's model, a drop in transportation costs unambiguously increases or decreases aggregate urban property values, as we shall see later in this article, depending on whether the urban economy is open or closed in population. Extensions of Mohring's model since its publication have greatly advanced our understanding of the relationship between transportation improvements and property values and the relationship is much more complex than that imagined by Webber.

Transport improvements cause economic activity to shift among different locations altering the demands for floor space and land, causing new equilibrium prices for property. To understand the causal relationships, we will consider several theoretical treatments. In the first, we will see how improvements affect competitive firms located at different distances from an export hub, for example a harbor, to which they must ship their output. We will then consider how consumer-workers located at different distances from their jobs are affected in a residential land market. Results differ depending on whether the urban economy is closed or open to population movements. Next, we will review what the literature on the monocentric urban model of urban economics, in which all jobs are located at a single center, tells us about the transport and property prices relationship. Within the framework of the monocentric city, we also consider how road traffic congestion and its pricing affect property prices; the relative degree to which prices of floor spaces and of land are affected by transport improvements; and the effect of building durability on how property prices are affected. In the final section of this article, we consider the empirical evidence that has accumulated since the 1930s, distinguishing among three strands of such studies: before and after observations of the effects of specific transport projects; hedonic multivariate regression studies of how accessibility improvements are valued in real estate markets; and equilibrium models of the real estate markets in which the demand for and supply of real estate are separately identified.

Firms Located Relative to an Export Hub

Consider a firm that ships its output to the world market, by transporting it a distance x miles from the firm's production location to the harbor. Suppose that the firm produces Q tons of output and that the cost of transporting the output is t per each ton-mile. If the world price of the output at the harbor is p per ton, then the net (or effective) price per ton is $p - tx$. Suppose that the firm's production function is $Q = F(N, L)$ and constant returns to scale where the inputs are N units of labor and L units of land. The wage rate is w , and the land rent is R . Labor is assumed to be perfectly elastically supplied to any location. Hence, the wage rate is invariant by location. Suppose that the firm is operating in a competitive market, thus taking all prices and the transport cost as given. The firm maximizes profit with respect to the two input quantities:

$$\text{Max}_{N,L} \Pi = (p - tx)F(N, L) - wN - RL. \quad (1a)$$

Because of constant returns, the aforementioned can be normalized to a unit of land and maximized with respect to labor input per unit of land, or labor density $n \equiv (N/L)$:

$$\text{Max}_n \Pi = (p - tx)F(n, 1) - wn - R. \quad (1b)$$

The first order condition says that the labor density at x miles from the center is such that the value of the marginal product of labor employed at that location equals the wage rate:

$$(p - tx)F_n - w = 0. \quad (1c)$$

This is illustrated in Fig. 1.

By differentiating this with respect to t and n we get:

$$\frac{dn}{dt} = \frac{F_n x}{(p - tx)F_{nn}} < 0. \quad (1d)$$

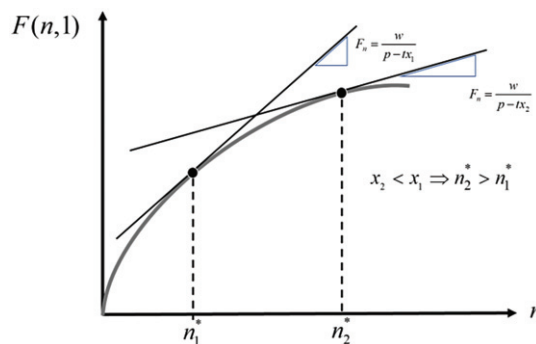


Figure 1 In a market of competitive producers, shipping outputs to a transport hub causes labor density to fall with distance from the hub.

The sign follows from diminishing marginal product, $F_{nn} < 0$. Eq. (1d) is the proof that when the unit cost of transportation improves (t decreases) then the demand for labor increases at each location. The next step is to see what happens to the land rent at location x . Competitive firms make zero profit in the long run. Therefore, we can write from Eq. (1b) that

$$R = (p - tx)F(n, 1) - wn. \quad (2a)$$

Maximizing R is equivalent to maximizing profit. The Envelope Theorem applies (because Eq. (2a) is maximized with respect to n , we can get Eq. (2b) by differentiating Eq. (2a) with respect only to the direct effect of t on R through tx , ignoring the indirect effect through n), and it is easy to see that the rent increases at each distance from the harbor as the cost of transport per mile decreases:

$$\frac{dR}{dt} = -F(n, 1)x < 0. \quad (2b)$$

Result 1: In a market of competitive firms, if the per-mile cost of transporting output to the market is reduced, then the labor density, the firm's output and the land rent increase at each location where firms are operating.

Residential Location in an Economy Open to Population

We consider an economy's residential area consisting of many locations. The economy is open to population which means that consumers can move in and out of the economy at zero cost in response to transportation improvements. If the improvements raise (lower) the level of utility within the economy, new consumers will move in (out) until the level of utility falls back to the level prevailing outside the open economy.

Consider a consumer within the economy who has income Y and chooses residence location i . Suppose that the location i entails travel from that location with total monetary cost C_i and total travel time T_i . The travel could be commuting from the residence location to the consumer's workplace (wherever that might be) or it can include a fixed bundle of work and nonwork trips from the residence location to meet the consumer's needs.

Let the utility function of the consumer be $U(X_i, H_i, D - T_i)$, where X_i is the quantity of a numeraire composite good that is an aggregate of all goods other than housing, and H_i the quantity of housing floor space that the consumer will demand if he chooses residence location i . $D - T_i$ is the leisure the consumer enjoys after the time spent on travel is deducted from D , the total time available for leisure and travel. Assume as is reasonable, that all three goods—housing, leisure, and the composite good—are normal goods. The consumer takes as given his income Y and the travel time and monetary costs of transportation C_i and T_i .

The consumer maximizes utility with respect to X_i and H_i subject to the budget constraint $X_i + P_i H_i = Y - C_i$, where P_i is the unit market price of housing floor space in i . Travel time, T_i , enters the utility as a bad (imparting disutility) because it decreases leisure. The indirect (or maximized) utility of the consumer is:

$$V(Y - C_i, P_i, D - T_i). \quad (3)$$

From Eq. (3), the marginal rate of substitution (MRS) between monetary travel cost and travel time can be derived by totally differentiating Eq. (3) with respect to T_i and C_i , while keeping the level of utility constant:

$$\frac{\partial V}{\partial C_i} dC_i + \frac{\partial V}{\partial T_i} dT_i = 0 \Rightarrow MRS(C_i, T_i) = \frac{dC_i}{dT_i} = -\frac{\partial V / \partial T_i}{\partial V / \partial C_i} < 0. \quad (4)$$

This MRS is often called the value of time (VOT) of the consumer. It tells us how much more the consumer would be willing to pay in monetary cost for a marginal reduction in travel time. The importance of Eq. (4) becomes apparent when we consider that consumers of different types, more precisely of different incomes, benefit differentially from the same transportation improvement. Higher income consumers have a high marginal disutility of travel time (high marginal utility of leisure) and a low marginal utility of disposable income. Hence, such consumers have a higher VOT than do lower income consumers. When the monetary cost of transportation does not normally vary much with a consumer's income while the MRS does vary more, any travel time improvement will generally be more valuable to the higher income consumers.

A second important property is the demands for housing floor space and the composite good, obtained by applying Roy's identity to Eq. (3) to get the former and then using the budget constraint to get the second:

$$H_i^* = -\frac{\partial V / \partial P_i}{\partial V / \partial (Y - C_i)} = H(Y - C_i, P_i, D - T_i). \quad (5a)$$

$$X_i^* = Y - C_i - P_i H_i^* \quad (5b)$$

Suppose that a transportation improvement decreases the monetary cost or the travel time or both. *Keeping income and the price of housing unchanged*, how will such a change affect the quantity of housing demanded? Because the monetary cost of transportation decreases from C_i to C_i' , disposable income increases and this works to increase the demand for the normal good H_i . But there is another effect because the transportation improvement increases leisure too. H_i^* will increase if housing and leisure are sufficiently complementary in consumption. H_i^* will decrease if composite good and leisure are sufficiently complementary in consumption.

More precisely, if the utility function is $U = u(D - T)v(X, H)$ and, hence, homothetic in X and H regardless of the level of leisure, then reducing travel time, T , does not change the quantity of X and H demanded by the consumer, but it does increase leisure and the level of utility. Instead, consider that $U = U[u(D - T, X), H]$ or $U = U[u(D - T, H), X]$. In the former leisure and consumption can be made complementary by choice of the subutility function $u(\bullet)$, and in the latter leisure and housing can be similarly made complementary.

Now, suppose that initially the population of consumers in the open economy choosing residence location i is N_i . Assume that, in the short run, N_i remains constant in response to the transport improvement. Then, the aggregate short run demand for housing at residence location i is $N_i H_i^*$ and will increase when H_i^* increases. Keeping the housing supply function unchanged but assuming that it is not perfectly elastic, the equilibrium price of housing will increase when leisure and housing are complementary enough in consumption. In the case where housing and leisure are strong enough substitutes, the aggregate demand for housing at residence location i can decrease and so will the equilibrium price of housing.

The next step is to allow mobility between different locations to gain insight about long run changes in the open economy and the effect on property prices. Assuming that all consumers have the same utility function and the same income and can change location at zero cost, what will be the long run equilibrium outcome? Consumers will relocate to those places where the transport improvement increased utility and this process will continue until the utility level is the same everywhere and there is no longer any gains from further relocations. Consider two locations i and j . Suppose that the transport improvement occurs only in location i . Thus before the improvement, we have the equilibrium condition that:

$$V(Y - C_j, P_j, D - T_j) = V(Y - C_i, P_i, D - T_i). \quad (6a)$$

After the improvement, we have that $C_i' < C_i$ and that $T_i' < T_i$, and therefore:

$$V(Y - C_j, P_j, D - T_j) < V(Y - C_i', P_i, D - T_i'). \quad (6b)$$

To bring the utility in location i back to equilibrium, the housing price must increase in the long run, so we have that $P_i' > P_i$:

$$V(Y - C_j, P_j, D - T_j) = V(Y - C_i', P_i', D - T_i'). \quad (6c)$$

From Eq. (6c), keeping the utility level constant before and after the transport improvement, we see that:

$$\frac{dP_i}{dT_i} = \frac{\partial V / \partial (D - T_i)}{\partial V / \partial P_i} < 0; \quad \frac{dP_i}{dC_i} = \frac{\partial V / \partial (Y - C_i)}{\partial V / \partial P_i} < 0. \quad (7)$$

Result 2: Either a monetary cost improvement or a travel time improvement in the transportation associated with a location unambiguously increases the housing price of that location when the level of utility is fixed.

Result 2 is illustrated in **Fig. 2**. The indifference curve bb' lies below the indifference curve aa' . Both indifference curves correspond to the same utility level, because bb' includes the effect of a lowered travel time, hence involves consumption-housing bundles with more leisure. The vertical intercept of the budget line is the disposable income and the lowered monetary transport raises disposable income, while the absolute value of the slope of the budget line is the housing price. We can see from the **Fig. 2** that the tangency between the budget line and the indifference curve bb' must occur at a higher housing price and a lower housing quantity demanded. Hence, a transport improvement which lowers travel time (increasing leisure) and/or lowers monetary travel cost, raises the price of housing when the level of utility is fixed, and reduces the demand for housing. A fixed utility level is assumed to be the condition in a city open to the in- and out-migration of population.

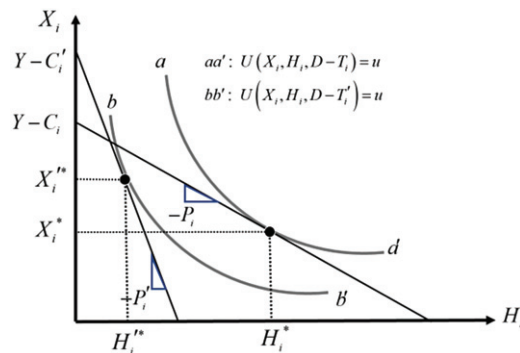


Figure 2 A transport monetary cost or travel time reduction at a location, raises the price of housing and reduces the quantity of housing demanded when the level of utility is fixed.

A second equilibrium condition makes it possible to determine the population that wants to locate at i . This is the condition that the housing market at location i clears:

$$N_i H_i^* - S_i(P_i) = 0. \quad (8)$$

$N_i H_i^*$ is the aggregate demand for housing floor space, and $S_i(P_i)$ is the upward sloping housing supply. By Result 2, the transportation improvement causes H_i^* to decrease and P_i to increase, hence N_i also increases. Hence, $\frac{dN_i}{dT_i} < 0$, and similarly $\frac{dN_i}{dC_i} < 0$.

Result 3: A lowered monetary cost or travel time associated with location i unambiguously increases the consumer population residing at location i when the level of utility is fixed.

Residential Location in an Economy Closed in Population

In an economy closed in population, the level of utility will be endogenously determined. Suppose that $N = \sum_{i=1, \dots, n} N_i$ is the total population initially distributed among n residence locations. After transportation is improved, this population will be redistributed so that $N = \sum_{i=1, \dots, n} N_i'$, where N_i' is population at i after. Intuitively, we expect that locations where transportation is improved relatively a lot, will attract population from the locations where transportation is improved relatively less or not improved at all. To obtain precise results, consider just two locations and assume that only travel time in location 2 is reduced. The equilibrium conditions are:

$$V(Y - C_1, P_1, D - T_1) = V(Y - C_2, P_2, D - T_2), \quad (9a)$$

$$N_1 H_1^* - S_1(P_1) = 0, \quad (9b)$$

$$(N - N_1) H_2^* - S_2(P_2) = 0. \quad (9c)$$

The three unknowns to be solved are P_1 , P_2 , and N_1 . By comparative static analysis of Eqs. (9a)–(9c), and assuming a utility function that is homothetic in housing and the composite good so that changes in leisure do not affect housing demands, it is shown that $\frac{dP_2}{dT_2} < 0$, $\frac{dP_1}{dT_2} > 0$, $\frac{dN_1}{dT_2} > 0$. The details of the comparative statics analysis are left as an exercise for the reader. The result generalizes to the case of an improvement in the monetary cost of transportation at location i .

Result 4: In an economy with two locations and closed in total population, a transport improvement associated with one location causes the housing price at that location to rise and the housing price at the other location to fall, while the population will be reallocated in the margin from the other location to the location where transportation improved.

Fig. 3 illustrates Result 4. For simplicity, the demand and supply functions in **Fig. 3** are drawn as linear. The demand function for floor space shifts to the right in location 2 while it shifts to the left in location 1, as the transport improvement in location 2 induces consumers in the margin to relocate from 1 to 2. How much the unit price of floor space rises in location 2 and how much it falls in location 1 is determined by the price elasticities of the two housing demand and housing supply functions.

The aggregate value of floor space at equilibrium is $\sum_{i=1,2} P_i N_i H_i(P_i) = \sum_{i=1,2} P_i S_i(P_i)$. The change in the aggregate housing price can be positive or negative, which again debunks Webber's air mattress theory. Additional information in **Fig. 3** is the aggregate land value (and its change) at each location. As is well known, aggregate land value is the producer surplus accruing to the landowner, and that is measured by the triangular area below the equilibrium price line and above the supply line. The Figure makes it clear that aggregate land value falls in location 1 and rises in location 2, while the combined aggregate again depends on the demand and supply elasticities. To sum up, what happens to the aggregate value of floor space also depends on whether the utility of residents remains unchanged. This holds, as we saw, when the urban economy is open.

As stated in the Introduction, the assumption of openness or closedness plays an important role in the cost-benefit analysis of transportation projects. Whether an economy should be modeled as open or closed is itself a nontrivial issue. Consider, for example, that transport improvements occur in some cities within a national economy. It would be appropriate to consider that in the long run population from other cities would migrate to the cities where the improvements occurred. Cities, therefore, should be modeled as open in population. We might observe higher property prices in the cities where improvements occurred while property prices in the other cities could be lowered if the national economy as a whole is closed in population.

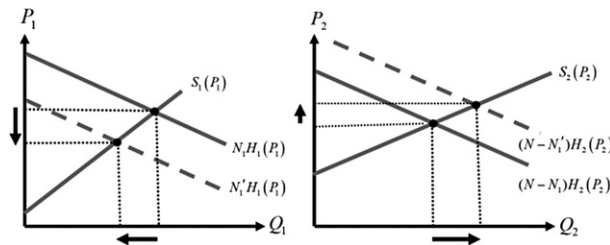


Figure 3 A transport improvement at location 2 raises the price of housing at location 2, lowering it at location 1 when the total population of the two locations is fixed but mobile between the two locations.

Transport Improvement in the Monocentric City

A powerful model used in urban economics is the monocentric model of urban land use equilibrium. The model assumes that all workers who are also consumers work at a central point, dubbed the central business district (CBD), and that the city's residential area is confined to the area of a circle around this central point. As a simplification, imagine that the roads leading to the CBD are densely distributed, so that any worker can travel to the CBD on a radius. It is assumed that the land market operates as an auction so that every location in the city goes to the consumer-worker bidding the highest rent for that location. Whether landlords own small pieces of land or one landlord owns all of the land makes no difference, because to maximize profits under perfect information, landowners must rent each parcel of land at the highest possible rent, hence to the highest bidder. The radius of the residential area is determined by arbitrage in the land market, so that at equilibrium at the urban edge the residential bid rent equals the exogenous agricultural rent.

A very simple version of the monocentric city model, mentioned earlier in this article, was proposed by [Mohring \(1961\)](#). In this model, a consumer rents a unit amount of land regardless of residential location x miles from the CBD and transport costs are linear in distance and t per mile. Mohring's simplification ignores the fact that where rents on land are lower, consumer-workers may rent more land causing the lot size per worker to increase with distance from the CBD. Yet, Mohring's model served as a very useful first step for understanding the relationship between land values and commuting costs in a monocentric city. We will now explain how the model works when the city is open or closed in population.

Assuming that all consumers working in the CBD have the same income Y , the budget constraint can be used to see that the consumption of all other goods, treated as a composite is $X(x) = Y - R(x) - tx$, where $R(x)$ is the rent at distance x and tx the cost of commuting to the CBD. For the consumers to be in equilibrium (not to want to relocate within the city), $X(x)$ —which is also a measure of the consumer's utility because lot size, the other good, is fixed—must be the same at each x . This in turn, requires that $R(x) + tx$ be the same at each x . Hence, $R(x) = R(0) - tx$. At the border of the city $x = \bar{x}$, and the equilibrium urban rent there is equal to the agricultural rent on land, r , by the no arbitrage condition: $R(\bar{x}) = r$ and $\bar{x} = [R(0) - r]/t$. In a circular city, the rent falling linearly in all radial directions at the same rate per mile, traces the surface of a cone the base of which is the surface area of the city. The volume of this cone is the aggregate rent, while the remaining volume of the cylinder with the same base within which the cone sits is the aggregate transportation cost. When t falls, both aggregate rent and aggregate transport costs increase and the open city expands in radius and area accommodating additional population, but aggregate rents always remain at one half of aggregate transport costs. If the city is closed in population, then the fixed lot size assumption implies that the city's radius and area cannot change. The arbitrage condition at the city border still requires that $R(\bar{x}) = r$ and so $R(0)$ must fall causing a higher utility level throughout the city. In this case, aggregate transport costs and aggregate rents fall, but the two-to-one relationship is still preserved at a higher level of utility.

In Mohring's simple model, when the city is open, Webber's air mattress conjecture never holds since a transport improvement always increases aggregate rents as well as aggregate transport costs. And when the city is closed, the air mattress conjecture again never holds as a transport improvement decreases aggregate rents. [Arnott and Stiglitz \(1981\)](#) showed that the two-to-one ratio is robust under various generalizations of Mohring's simple monocentric model.

The monocentric model of Mohring was extended and became much more realistic by the initial efforts of [Alonso \(1964\)](#), [Mills \(1967\)](#), [Muth \(1969\)](#), and [Strotz \(1965\)](#) all of which were available to Melvin Webber. Here, let us consider our own version of the monocentric model. Assume that all consumer-workers have identical preferences $U(X, q, D - T)$ where, as before, X is the quantity of a numeraire composite good, q is the quantity of land the consumer rents (housing is ignored here) and T is the travel time from the consumer's residence location to the CBD. As before, D is total time available and $D - T$ is leisure. The consumers all have income, Y , and incur monetary transportation cost given by the function $C(x)$. The consumers achieve the same utility level, u , at equilibrium no matter where in the city they locate. From the budget constraint, the highest rent that can be bid for a unit piece of land located at distance x from the CBD is:

$$R(x) = \frac{Y - C(x) - X[u, q(x), D - T(x)]}{q(x)}, \quad (10a)$$

where $X[u, q(x), D - T(x)]$ is the indifference surface between the composite good, the lot size and leisure. The landlord who owns land at location x is a utility taker and will maximize the rent he can charge with respect to $q(x)$. Finding the optimal value $q^*(x)$ tells the landlord how much land each of his parcels-for-rent should contain. An important issue in the model is the rent gradient, that is, the rate at which the rent changes with distance from the CBD. This can be found by applying the Envelope Theorem to Eq. (10a) (The Envelope Theorem in this case means that since the landlord maximized Eq. (10a) with respect to $q(x)$, the derivative of $R(x)$ can be found by differentiating Eq. (10a) with respect to the direct effects of x only, that is the effects through $C(x)$ and $T(x)$). Doing so gives:

$$\frac{dR(x)}{dx} = -\frac{1}{q^*(x)} \left[\frac{dC(x)}{dx} - \frac{\partial X}{\partial (D - T(x))} \frac{dT(x)}{dx} \right] < 0, \quad (10b)$$

where $\frac{dC(x)}{dx} > 0$ is the rate at which transport cost at location x increases with distance, and $q^*(x)$ is the optimal lot size at distance x . $\frac{\partial X}{\partial (D - T(x))} < 0$ is how much composite good the consumer would want to receive to be compensated for a marginal decrease in leisure keeping constant the lot size and the level of utility. Hence, this term is the consumer's "value of time" or VOT. If leisure is not in the utility function, then $VOT = 0$ and Eq. (10b) reduces to $\frac{dR(x)}{dx} = -\frac{1}{q^*(x)} \frac{dC(x)}{dx} < 0$, also known as Muth's condition ([Muth, 1969](#)). Suppose that the monetary commuting cost is linear with distance with slope t , then Muth's condition becomes $\frac{dR(x)}{dx} = -\frac{t}{q^*(x)} < 0$. The

presence of leisure in the utility function causes the rent gradient to increase, steepening the rent-distance function. As the consumer's income increases, there are actually two countervailing effects. On the one hand the *VOT* should increase because richer consumers value their travel times more highly, causing the term in the brackets in Eq. (10b) to increase; while, on the other hand, the demand for lot size increases with income causing $q^*(x)$ to increase. While the *VOT* effect steepens the rent-distance function, the lot size effect flattens it as income increases.

Wheaton (1977) showed by empirical analysis for the San Francisco Bay area using 1960 Census data that the income elasticity of the demand for lot size is bigger than the income elasticity of *VOT*. Hence, the rent-distance function gets flatter with income which supports the more peripheral location of the rich as compared to the central location of the poor in US cities. However, adequately explaining the full flattening of the rent-distance function with income requires recognizing that not only does the demand for lot size increase with income but so do also other important demands such as the demand for public services, safe neighborhoods, open spaces, and school quality. Because these attributes increase with distance from the CBD due to historical trends in the outward development of cities, the slope of the rent-distance function of Americans flattens out robustly with income.

How does the equilibrium rent-distance profile of the monocentric city respond to a lowering of the per mile monetary and/or time cost of commuting to the CBD? This question was examined in Wheaton (1974). The result, illuminated by our Fig. 2, applies to an open monocentric city: as the unit cost of commuting decreases, rent increases at each distance from the CBD because the utility level is fixed. In a city closed in population, the result is illuminated by our Fig. 3. Because the monetary and/or time cost of commuting cost decreases, a far-from-the-CBD location affords higher disposable income and/or more leisure than before. As a result, this increases the demand for peripheral locations at the expense of central ones. Moving out to the periphery where land is cheaper, consumers can rent larger lots. As a result, rents in the central locations fall while rents rise in the peripheral locations. This basic result is at the heart of why housing decentralized after the car became widely utilized and also explains why the property tax base of central cities decreased.

The monocentric model as exposit by Strotz (1965) and Alonso (1964) lacked housing as did Mohring's. Housing services were proxied by the lot-size variable, also true in our version earlier. This deficiency was corrected by Muth (1969) who showed that, at equilibrium, the structural density of buildings (housing in particular) will decrease with distance from the CBD. An important result is that the rent of housing floor space declines much more gently than does the rent on land. This property is intuitive if we recall that the amount of land at every distance from the CBD is perfectly inelastically supplied since new land cannot be created at any location, while the supply of floor space at any distance can be increased by building taller buildings. Building developers constructing closer to the CBD build taller buildings because they are confronted with higher land prices there, whereas it is plausible that the prices of structural inputs (concrete, bricks, glass, steel) are perfectly elastically supplied to any location in the city. The substitution of capital for land in the construction of floor space causes the structural density of buildings to decrease with distance from the CBD.

To illustrate this result, consider that identical consumers all have income Y , and a utility function defined over two goods: housing floor space H and all other goods X , ignoring leisure in this case. Suppose also that transport cost is linear with distance x from the CBD with per-mile transport cost, a constant t . Assume also that the utility function is Cobb–Douglas: $U = X^\alpha H^{1-\alpha}$, $0 < \alpha < 1$. Floor space is produced by competitive developers operating under constant returns to scale and combining land inputs, L , and capital inputs, K . Let this be a Cobb–Douglas production function: $H = K^a L^{1-a}$, $0 < a < 1$. Capital is perfectly elastically supplied to any location within the monocentric city. It is easy to show that under these assumptions, the equilibrium rent on floor space, $P(x)$, and the equilibrium rent on land, $R(x)$, decline with distance according to power laws as follows:

$$P(x) = C_1(Y - tx)^{\frac{1}{1-\alpha}} \quad \text{and} \quad R(x) = C_2(Y - tx)^{\frac{1}{(1-\alpha)(1-a)}}. \quad (11)$$

C_1, C_2 are constants. Plausible values are $\alpha = 0.75$ for the share of disposable income allocated to all goods other than housing, and $a = 0.9$ for the cost-share of capital in building development. Hence, the exponents in Eq. (11) are $\frac{1}{1-\alpha} = 4$ and $\frac{1}{(1-\alpha)(1-a)} = 40$. Consider now the elasticity of the two prices with respect to the unit transportation cost t . As t falls denoting a transport improvement, both prices increase, but the price on land increases 10 times more than does the price on floor space.

A final comment on the monocentric model is that, in all its versions discussed above, the city is assumed to be malleable in structural capital. Determined from scratch and, ignoring all adjustment costs, it transforms instantly into a new equilibrium when the economy changes. This condition does not hold in reality since buildings are durable and adjust slowly to changes in economic conditions. Consider, for example, the case of infinitely durable buildings and the city expanding from the CBD outwards over time, growing like a tree trunk, where a ring of new land development is added to the tree trunk's outer edge every time period. This case was examined by Anas (1978).

The basic insight of Anas (1978) can be derived from Eq. (10a) with the caveat that since buildings are inherited from the past, the lot size $q(x)$ which proxies housing will be taken as an exogenous function determined by history. Only at the current edge—the most recent ring in the tree trunk—of the growing city is $q(x)$ determined by current economic conditions. It follows that in this case, the modified Muth's condition will differ from Eq. (10b), and the Envelope theorem can no longer be used. The condition now becomes (recalling our assumption that leisure does not enter the utility function):

$$R'(x) = -\frac{C'(x)}{q(x)} - \frac{R(x) - |MRS(X, q(x))|}{q(x)} q'(x), \quad (12)$$

where $MRS(X, q) < 0$ is the marginal rate of substitution between lot size and the composite good. Only at the edge of the city, where development happens currently, does rent equal the *MRS*. In that case the second term vanishes and Muth's condition holds locally

at the edge of the city. For all other locations between the CBD and the edge of the city, the rent gradient deviates from Muth's condition in general. Consider, for example, the case where a city's income and utility have been growing over time and in such a way that $q'(x) > 0$ over all locations, that is in each period houses with larger lot sizes have been built while, once built, all older lot sizes remained unmarketable. Under such a growth profile, it is possible that $R(x) < |MRS[X, q(x)]|$. This circumstance can make the second term in Eq. (12) positive and dominating over the first term, thus causing the rent on land to be rising with distance from the CBD until at some intermediate distance it begins to decrease again. How does a transport improvement affect the rent gradient under these conditions? On the one hand, the improvement changes the gradient by making the first term in Eq. (12) more negative. On the other hand, the transport improvements occurring over time can increase $q'(x)$ too, and this reinforces the tendency for a rent that increases with distance.

Congestion and Congestion Pricing

The effect of traffic congestion on property prices is complex. In the literature, this effect has been examined mostly within the narrow setting of a monocentric city. [Strotz \(1965\)](#) was the first to model congestion in the monocentric city and the theme was picked up again in the 1970s by many authors. The common approximation is that all traffic arrives at the CBD at the same point in time each morning. This means that the commuter residing at the edge of the city departs first and commuters residing at closer-in distances join the traffic stream when it passes their location. This accumulation of traffic increases the cost of travel per mile nonlinearly with distance from the CBD, the closer the mile being to the CBD, the higher the cost of travelling it. Assuming that the presence of congestion does not disturb the location of the jobs in the CBD, residents can economize in the presence of congestion only by moving their residence locations closer to the CBD, and thus reducing the total effect of congestion on them. Rents on land and floor space near the CBD increase and so do structural and population densities closer to the CBD, while the city gets more compact (shorter radius). Pricing congestion, again assuming no changes in job locations, exacerbates this effect. There is only one margin along which commuters can blunt the effect of the congestion tolls on them. They can do so only by moving closer to the CBD which raises rents near the CBD even more, making the city even more compact than the initially unpriced congestion did.

Of course, the effect of congestion on job location is too important to be ignored. [Anas and Kim \(1996\)](#) showed that congestion and congestion pricing can cause firms and the jobs they offer to move out of the CBD and to agglomerate in new subcenters, causing the monocentric city to become polycentric. The process of subcenter formation in a city closed in population raises property prices at the subcenter locations, lowering property prices at the CBD. [Anas and Pines \(2013\)](#) showed that pricing congestion in each city of an economy with identical cities causes the number of cities to increase, each city becoming smaller and rents near the CBDs of the cities falling as new cities emerge to decongest the existing cities. This result underscores the limitation of the analysis of the solo monocentric model in which it is assumed that when congestion increases neither jobs nor residents can move to other cities. Because in the solo monocentric city, intercity mobility is assumed not to happen, density near the CBD increases and so do rents.

The Empirical Evidence

There have been three types of empirical studies that have documented that transport improvements affect property prices in cities.

Before and After Comparisons

The first type of study dates as far back as the 1930s and dealt with actual transport projects, looking to descriptively document changes in property values using before and after observations. Many of these studies tried to establish that there was a positive relationship between travel time and travel cost savings made possible by the projects and observed increases in property values. [Spengler \(1930\)](#) and [Davis \(1965\)](#) examined the effect of rapid transit projects on property values in New York and Northern Chicago, respectively. [Adkins \(1959\)](#) examined expressways in Dallas, Houston and San Antonio; [Lemly \(1959\)](#) along Atlanta's expressways; and [Golden \(1968\)](#) examined Chicago expressways. [Wootan and Haning \(1960\)](#) focused on interstate highways in Austin, Texas; and [Ashley \(1965\)](#) on interchange development along I-94. A suburban commuter railroad's effects on property values were evaluated by [Boyce et al. \(1972\)](#) who studied the Lindenwold-Camden-Philadelphia line.

Hedonic Regression Analysis

A second type of study became common in the 1970s and has continued until recent times. Authors relied on hedonic multivariate regression analyses to show positive effects of transport projects on property values while controlling for other effects. Among these are those by [Deweese \(1976\)](#) who studied the effects of subway stations on Bloor street in Toronto on property values and also of [Bajic \(1983\)](#) for Toronto. [Damm et al. \(1980\)](#) studied how the announced opening of the Washington DC METRO stations affected real estate prices near the station locations. [Voith \(1993\)](#) studied Philadelphia, [Gatzlaff and Smith \(1993\)](#) studied the Miami Metrorail, and [Hess and Almeida \(2007\)](#) studied the effects of the Buffalo light rail line on property values. Hess and Almeida also provided a thorough list of many other similar studies for various US cities that implemented new transit stations until relatively recent times. In the last decade a plethora of studies, too numerous to mention here, have implemented similar methods for Asian public transportation projects.

While studies mentioned here and many of the others found positive effects of transit stops on nearby real estate prices, in some cases the positive effects on housing prices are confounded by negative effects from the noise externalities and the commercial developments that arise as transit stations induce changes in the real estate market. Bowes and Ihlanfeldt (2001) attempted to disentangle some of these effects. They argued that while stations may raise the value of nearby properties by reducing commuting costs or by attracting retail activity to the neighborhood, countering these positive effects are the negative externalities emitted by stations and the access to neighborhoods that stations might provide to criminals. Both of these auxiliary effects which are endogenous to the transport improvement would negatively affect property prices, and more so closer to the stations.

A shortcoming of the hedonic approach is that it models real estate prices at equilibrium but cannot distinguish between the transportation project's influences on the demand for location and the supply of floor space available at that location. Suppose, for example, that the demand and the supply of floor space in a neighborhood near a transport improvement are given by the following two equations:

$$D = a + bP + cZ + dT + \varepsilon_1 \quad \text{and} \quad S = e + fP + \varepsilon_2, \quad (13a)$$

with the coefficient $b < 0$ and the coefficient $f > 0$. P is the unit price of floor space, Z are various neighborhood characteristics, and T is the travel time measure associated with the transport improvement with the coefficient $d < 0$: for example a lowering of travel time, T , increases the demand to locate in the neighborhood. Assume in this example that the Z are all exogenous so that no endogeneity bias arises. The error terms are ε_1 and ε_2 . It is assumed that the variables Z , T do not directly affect the supply side. At equilibrium, $D = S$ implies the following hedonic relationship that is to be estimated:

$$P = \underbrace{\frac{e-a}{b-f}}_{\equiv \alpha_0} - \underbrace{\frac{c}{b-f}}_{\equiv \alpha_X} Z - \underbrace{\frac{d}{b-f}}_{\equiv \alpha_T} T + \underbrace{\frac{1}{b-f}(\varepsilon_2 - \varepsilon_1)}_{\equiv \varepsilon} \quad (13b)$$

Note that $b - f < 0$. Eq. (13b) is always estimated as $P = \alpha_0 + \alpha_X Z + \alpha_T T + \varepsilon$. If $c > 0$, then $\alpha_X = -\frac{c}{b-f} > 0$, and since $d < 0$, then $\alpha_T = -\frac{d}{b-f} < 0$. While the regression captures the equilibrium effect of a $\Delta T < 0$ causing a $\Delta P > 0$, it cannot distinguish the direct demand effect of ΔT from the effect of ΔP on the supply increase. If the estimated coefficient α_T is large in absolute value, it is correctly concluded that a travel time reduction has a large effect on price, but nothing is learned about whether this large effect stems from the effect of travel time on demand or from the price inelasticity of the floor space supply function. If, for example, the supply of floor space is nearly perfectly elastic as it might be in a suburban area with plenty of land that is properly zoned and ripe for development, then the coefficient f would be large and positive resulting in the estimated α_T to obtain a very small negative value. That is, the true effect of the highway project on demand is obscured by the large supply elasticity.

Market Equilibrium Simulations

The third type of study is designed, in part, to overcome the shortcoming of the hedonic studies explained earlier. In this third type of study, microeconomic equilibrium models are used empirically to simulate the effect of projects on the real estate market. Anas (1982) and Anas and Duann (1985) presented the Chicago Area Transportation and Land Use Analysis System (CATLAS), a model which relied on empirically estimated probabilistic location choice and housing supply equations to estimate the response of the housing market to a new rail transit project planned for Chicago's Southwest Corridor. The CATLAS model was estimated using 1970 US Census data and other complementary data. 1980 Census data could not be used due to the timing of the study. So the planned project was modeled counterfactually both because it did not yet exist and because it was modeled under 1970 conditions.

A project similar to one of the alignments that was studied by the CATLAS model was actually built and opened for service about 8 years later on October 1, 1993, and is now known as the 11 miles long Orange Line which connects downtown Chicago with Midway airport. The CATLAS model divided the Southwest Corridor area into small square zones of $\frac{1}{2}$ mile edges around the proposed alternative transit alignments and into larger square zones of 1 mile edges in the suburbs. The entire Chicago MSA was covered by 1690 such zones of which 205 were located within the Southwest Corridor (137 within the City of Chicago and 68 in the suburbs). The model accounted for walking, feeder bus, and driving with and without parking ("kiss and ride" and "park and ride") as alternative methods for accessing the proposed transit stations. The model accounted for the residential location choices made by workers employed in the Chicago CBD as well as those employed in nonCBD workplaces. Nonwork trips could not be included in the model because data on such trips were not available.

The core of the CATLAS model is described by the following equation system that expresses the short run market equilibrium in the zones of the region:

$$\sum_{h=1,2} N_h \sum_{m=1,\dots,4} P_{him}(\mathbf{R}, \mathbf{T}, \mathbf{C}) = S_i \Phi_i(R_i); \quad i = 1, \dots, 1690. \quad (14)$$

In this equation, $h = 1$ denotes CBD commuters and $h = 2$ nonCBD commuters, where N_h is the given number of such consumers. $P_{him}(\mathbf{R}, \mathbf{T}, \mathbf{C})$ is a nested multinomial logit model of choice among residential zones i and the four modes of travel (car, bus, rail, and other) estimated by maximum likelihood. This choice probability depends on the rent, $\mathbf{R}$, travel time, $\mathbf{T}$, and monetary travel cost, $\mathbf{C}$, vectors of all the zone-mode combinations via a utility function that depends on various zone characteristics and zone and mode specific constants. S_i is the stock of existing housing units in zone i , and $\Phi_i(R_i)$ is the probability that a landlord will offer his housing

unit to the market than keep it vacant, estimated by maximum likelihood. Hence, $1 - \Phi_i(R_i)$ is the housing vacancy rate in zone i . $S_i\Phi_i(R_i)$ is the short run supply function of zone i .

The CATLAS model showed rent increases for the housing in the zones located near the planned stations and in some suburban zones that had good access to planned suburban stations with projected parking availability. The alternative proposed rail transit projects that were studied were found to cause commuters in the margin to shift residences from the suburbs to the central city zones and to the suburban zones that had good access to stations with parking. The alignments studied by the CATLAS model extended beyond Midway airport, but this part of the alignment was not actually constructed later. Thus rents in the central city zones increased while those in some of the suburban zones decreased, but overall aggregate rents in the corridor increased from 1.48% to 1.80% for the three alternative proposed rail alignments (from 2.14% to 2.92% in the central city part of the corridor; and from 0.70% to 0.73% in the suburban part). The highest rent increase was \$247 and the average was \$25 per housing unit per year about a 10% increase near the proposed transit line under 1970 conditions and in 1970 dollars. Estimated increases in the consumer surplus varied from 2.3% to 2.9%. The aggregate rent increase within the corridor varied from \$6.4 to \$8.2 million per year for the three alternative alignments, and discounting this with an appropriate interest rate, it was found that the aggregate rent increases, if they could be recovered by an incremental special assessment tax within the Southwest Corridor, would potentially defray from 25.8% to 36.1% of the capital costs of implementing the alternative projects studied by the model.

After the opening of the Southwest Corridor's Orange line, McDonald and Osuji (1995) and McMillen and McDonald (2004), examined how house prices actually changed in response to the anticipation of the opening and after the start of the operation of the line in 1993. The first paper pools house price data from 79 city blocks near the Orange line for which prices are observed for 1980 and 1990 before the line had opened, but its opening was announced and broadly assumed by market agents in 1990. The second study used repeat sales analysis of 4056 single family houses that sold more than once between 1983 and 1999, that is from 10 years before to 6 years after the line's opening. Both studies showed that prices began increasing prior to the opening of the line in anticipation of the future travel time savings. As for the magnitude of price increases, the results are in close agreement with those of the CATLAS model's equilibrium predictions prior to the construction of the Orange line, despite the different data sources and modeling assumptions. The CATLAS model using 1970 data predicted a 10% increase in housing prices within ½ mile of stations while McDonald and Osuji (1995) measured a 17% increase, and McMillen and McDonald (2004) using the repeat sales approach measured a nearly 7% increase for the same compared to the 10% prediction of the CATLAS equilibrium model, a prediction made 8 years prior to the project's opening and with a model calibrated with outdated data, that was nevertheless the best available at the time.

References

- Adkins, W.G., 1959. Effects of the Dallas Central Expressway on Land Values and Land Use. Texas Transportation Institute, Austin, Texas.
- Alonso, W., 1964. Location and Land Use. Harvard University Press, Cambridge, Mass.
- Anas, A., 1978. Dynamics of urban residential growth. *J. Urban Econ.* 5 (1), 66–87.
- Anas, A., 1982. Residential location markets and urban transportation. Academic Press, New York.
- Anas, A., Duann, L.S., 1985. Dynamic forecasting of travel demand, residential location and land development. *Pap. Reg. Sci. Assoc.* 56, 37–58.
- Anas, A., Kim, I., 1996. General equilibrium models of polycentric urban land use with endogenous congestion and job agglomeration. *J. Urban Econ.* 40 (2), 232–256.
- Anas, A., Pines, D., 2013. Public goods and congestion in a system of cities: how do fiscal and zoning policies improve efficiency? *J. Econ. Geogr.* 4, 649–676.
- Arnott, R.J., Stiglitz, J.E., 1981. Aggregate land rents and aggregate transport cost. *Econ. J.* 91, 331–347.
- Ashley, R.N., 1965. Interchange development along 180 miles of I-94. Highway Research Record 96. Highway Research Board.
- Bajic, V., 1983. The effect of a new subway line on housing prices in metropolitan Toronto. *Urban Stud.* 20, 147–158.
- Bowes, D.R., Ihlanfeldt, K.R., 2001. Identifying the impacts of rail transit station on property values. *J. Urban Econ.* 50, 1–25.
- Boyce, D.E., Allen, B., Mudge, R.R., Slater, P.B., Isserman, A.M., 1972. Impact of rapid transit on suburban residential property values and land development. *Transp. Res. Board* 358.
- Damm, D., Lerman, S., Lerner-Lam, E., Young, J., 1980. Responses of urban real estate values in anticipation of the Washington Metro. *J. Trans. Econ. Policy* 14, 315–355.
- Davis, J.L., 1965. The Elevated System and the Growth of Northern Chicago. *Studies in Geography*, No.10, Northwestern University, Evanston, Illinois.
- Deweese, D.N., 1976. The effect of a subway on residential property values in Toronto. *J. Urban Econ.* 3 (4), 357–369.
- Gatzlaff, D., Smith, M., 1993. The impact of the Miami Metrorail on the value of residences near station locations. *Land Econ.* 69, 54–66.
- Golden, S., 1968. Land Values in Chicago before and after Expressway Construction. The Chicago Area Transportation Study, Chicago, Illinois.
- Hess, D.B., Almeida, T.M., 2007. Impact of proximity to light rail rapid transit on station-area property values in Buffalo, New York. *Urban Stud.* 44 (5), 1041–1068.
- Lemly, J.H., 1959. Changes in land use and value along Atlanta's expressways. Highways and economic development. Highway Research Board Bulletin 277.
- McDonald, J.F., Osuji, C.I., 1995. The effect of anticipated transportation improvement on residential land values. *Reg. Sci. Urban Econ.* 25 (3), 261–278.
- McMillen, D.P., McDonald, J., 2004. Reaction of house prices to a new rapid transit line: Chicago's midway line, 1983–1999. *Real Estate Econ.* 32, 463–486.
- Mills, E.S., 1967. An aggregative model of resource allocation in a metropolitan area. *Am. Econ. Rev.* 57, 197–210.
- Mohring, H., 1961. Land values and the measurement of highway benefits. *J. Polit. Econ.* 69 (3), 236–249.
- Muth, R.F., 1969. Cities and Housing. University of Chicago Press, Chicago.
- Spengler, E.H., 1930. Land Value in New York in Relation to Transit Facilities. AMS Press, New York.
- Strotz, R.H., 1965. Urban transportation parables. In: Margolis, J. (Ed.). *The Public Economies of Urban Communities*. (Chapter 7), Resources for the Future Press, Washington D.C.
- Voith, R., 1993. Changing capitalization of CBD oriented transportation systems: evidence from Philadelphia. *J. Urban Econ.* 30, 360–372.
- Webber, M., 1976. The BART experience - what have we learned? Monograph No. 26. Institute of Urban and Regional Development and Institute of Transportation Studies. University of California, Berkeley.
- Wheaton, W.C., 1974. A comparative static analysis of urban spatial structure. *J. Econ. Theory* 9, 223–237.
- Wheaton, W.C., 1977. Income and urban residence: an analysis of consumer demand for location. *Am. Econ. Rev.* 67 (4), 620–631.
- Wootan, W., Haning, R., 1960. Changes in Land Values, Land use and Business Activity along a Section of the Interstate Highway System in Austin, Texas. Texas Transportation Institute, Austin, Texas.

Transport Infrastructure Effects on Economic Output: The Microeconomic Approach

Patricia C. Melo, Department of Economics, ISEG—School of Economics and Management, Universidade de Lisboa & REM/UECE, Lisbon, Portugal

© 2021 Elsevier Ltd. All rights reserved.

Introduction	347
Empirical Evidence From Microeconomic Firm-Level Studies	348
Measuring the Effect of Transport Improvements: From Distance to Generalized Travel Cost	348
Identification Issues and Strategies	349
Industrial Heterogeneity	350
Spatial Heterogeneity	352
Conclusions	352
Acknowledgments	353
Appendix A	353
Appendix B	353
Further Reading	354

Introduction

This article provides an overview of existing empirical evidence on the effect of transport improvements on firm-level private output and productivity. The discussion focuses on the more recent research adopting a microeconomic production function framework, while the previous article considered the older and more conventional approach based on national and regional macroeconomic analyses. These earlier estimates of the output elasticity of transport investment obtained from more aggregate-level studies tended to suggest strong economic growth benefits and were often used by politicians to justify government funding for new and improved transport infrastructure. More recent evidence shows that many of these studies identified spurious relationships by failing to address the main identification challenges in this literature, in particular, issues of firm self-selection and reverse causality between transport investment and economic performance.

Firm- and plant-level studies are relatively recent in this literature, partly due to data availability reasons. One of the main advantages of the microeconomic approach based on longitudinal firm data is that it offers the opportunity to better examine some of the mechanisms underlying performance differentials in the spatial economy. Importantly, this approach can distinguish between the effects from within-plant or firm productivity growth, firm selection effects whereby more productive firms sort into locations with better transport accessibility, and displacement effects whereby new entrants with higher productivity replace older poor performing firms.

It is not the intention of this article to provide an exhaustive discussion of the ways in which transport infrastructure investment impacts on productivity and economic growth. There are a number of valuable surveys on the economic effects of transport infrastructure, whose references are provided as further reading at the end of this article. Transport infrastructure improvements lower transport costs and improve accessibility to input (i.e., suppliers, labor, etc.) and output markets, producing several impacts on the economy. These effects include, among other factors, the expansion and integration of wider markets, leading to productivity gains from improved labor supply and specialization; higher efficiency through scale economies and economic restructuring due to firm entry and exit resulting from stronger competition; and productivity effects from spatial agglomeration economies.

The impacts of transport improvements on the economy can be directly captured by the travel time savings, or generalized travel cost savings, accruing to transport users for work and nonwork purposes and indirectly through transfer to other users such as consumers and house and landowners. However, some of the economic benefits resulting from transport infrastructure investments are thought to be additional to transport user benefits and thus are not captured in conventional cost–benefit analysis. The most prominent example of wider economic benefits refers to the productivity gains from transport-induced agglomeration externalities. By changing the way people and firms have access to economic activity, transport affects the realization of agglomeration externalities and hence the productivity effects derived from it. In contrast, however, productivity gains resulting from industry reorganization should not be counted as additional because they are captured as a transfer in transport demand and thus the conventional appraisal of transport investment projects.

In brief, this article is structured as follows. Following the outline of the main effects of transport infrastructure investment on firms' output and productivity provided in this section, Section "Empirical Evidence From Microeconomic Firm-Level Studies" discusses the current state of the art of the empirical literature in relation to the approaches used to capture changes in economic access resulting from transport improvements (Section "Measuring the Effect of Transport Improvements: From Distance to Generalized Travel Cost"), the main identification challenges and strategies adopted (Section "Identification Issues and

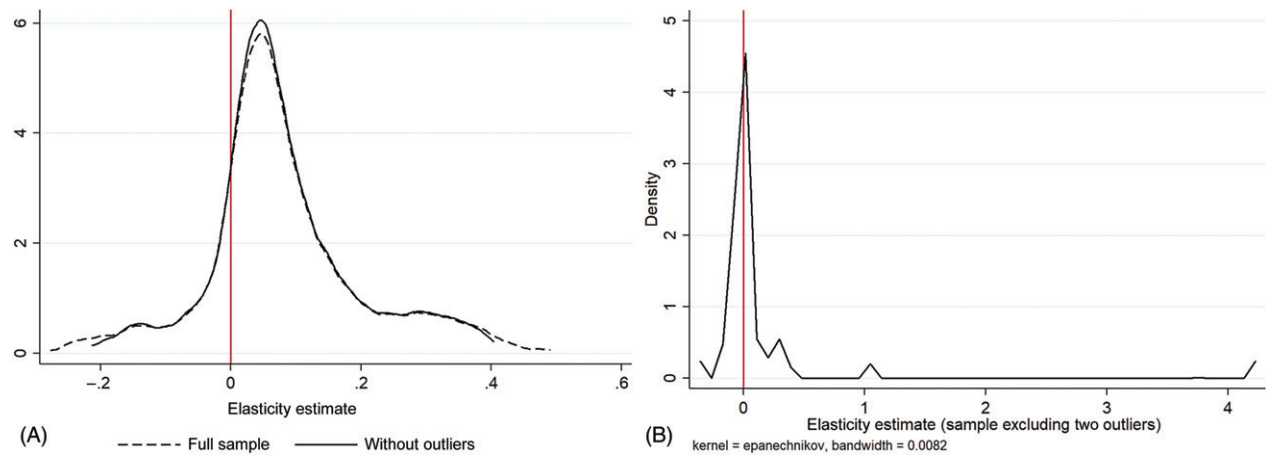


Figure 1 Histogram of transport elasticity estimates from firm microeconomic studies using market potential type measures (top panel) and distance measures (bottom panel).

Strategies”), the extent of industrial heterogeneity (Section “Industrial Heterogeneity”), and the scope for and nature of spatial effects (Section “Spatial Heterogeneity”). Section “Conclusions” provided some tentative conclusions and emphasizes some of the main challenges guiding future research.

Empirical Evidence From Microeconomic Firm-Level Studies

This section provides an overview of the current state of knowledge from empirical microeconomic studies using firm- or plant-level data on the nearly 450 elasticity estimates obtained from the 13 studies listed in Appendix A and summarized in Appendix B according to whether they use a transport measure based on market potential type accessibility indicators (12 studies, 314 estimates) or distance to the nearest access point to the transport network (2 studies, 140 estimates). Fig. 1 shows the histogram of the elasticity estimates separately for the two types of transport measure. Table B2, in Appendix B, shows the middle half of the sample of output elasticities with respect to market access has values between 0.023 and 0.127, and the mean (median) value of the output elasticity is 0.08 (0.06) indicating that an increase of 10% in market access is associated with an uplift of firm output of 0.8% (0.6%). If we consider instead the effect of reducing distance to the transport network, Table B2 shows that the middle half of the sample has values between -0.027 and 0.006 , while the mean value is substantially influenced by whether we include the top and bottom percentiles. Taking the median as the reference point, we observe that reducing the distance to the nearest access point to the transport network by 10% is associated with an increase in output of around 0.14%.

Measuring the Effect of Transport Improvements: From Distance to Generalized Travel Cost

The summary in Table B1 in Appendix B shows that the most common measure used to capture the effect of transport investment is based on the concept of market potential, for which there are several similar implementations (e.g., effective density, economic mass, etc.). The market potential of a given firm in a given location is obtained by summing the opportunities for interaction in all

Table 1 Summary of elasticity estimates by specification of (road) transport network

<i>Transport measure based on market potential type accessibility indicators</i>										
	<i>Count</i>	<i>Min</i>	<i>P10</i>	<i>P25</i>	<i>Mean</i>	<i>Median</i>	<i>P75</i>	<i>P90</i>	<i>Max</i>	<i>CV</i>
Euclidean distance	232	-0.213	-0.042	0.026	0.086	0.066	0.137	0.265	0.404	1.338
Time-invariant travel time	24	-0.003	0.008	0.013	0.025	0.023	0.029	0.056	0.059	0.701
Time-varying travel time	41	-0.213	-0.066	0.004	0.044	0.047	0.074	0.146	0.278	1.992
Generalized travel cost	9	0.042	0.042	0.114	0.215	0.226	0.329	0.399	0.399	0.588
Sample without outliers	306	-0.213	-0.033	0.024	0.080	0.060	0.121	0.251	0.404	1.405
<i>Transport measure based on distance to transport network/node/access point</i>										
	<i>Count</i>	<i>Min</i>	<i>P10</i>	<i>P25</i>	<i>Mean</i>	<i>Median</i>	<i>P75</i>	<i>P90</i>	<i>Max</i>	<i>CV</i>
Time-varying travel time	138	-0.349	-0.092	-0.027	0.054	-0.014	0.006	0.079	4.214	9.127

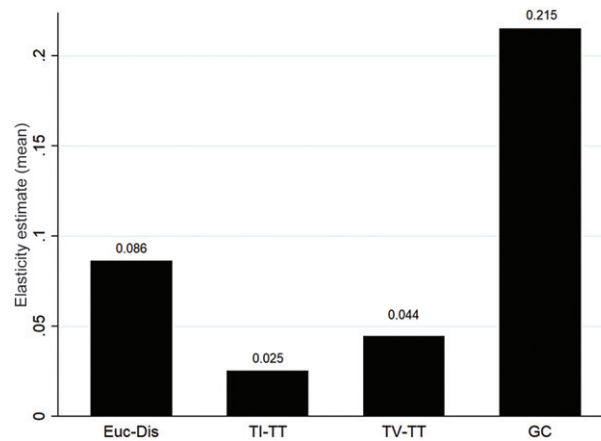


Figure 2 Summary of elasticity estimates for market potential measures and definition of transport network. *Euc-Dis*, Euclidean distance; *TI-TT*, time-invariant travel time; *TV-TT*, time-varying travel time; *GC*, generalized travel cost.

other locations and weighting them by (a function of) the distance, travel time, or generalized travel cost to access them. This approach is an improvement on the more simple measures used by macroeconomic studies, including a binary indicator of the presence or absence of a transport infrastructure within a given location (e.g., whether has railway station, whether is crossed by motorway), the amount of transport infrastructure (e.g., kilometers of railway, kilometers of motorway, number of radial roads), or transport investment. Although market potential type measures have become increasingly common in the literature, they often define transport costs using straight-line Euclidean distances that do not reflect the real transport network and cannot be used for time series analysis. Consequently, they cannot capture changes in the real transport network, but only changes in the spatial distribution of opportunities. Table 1 shows that evidence from microeconomic studies based on market potential type measures defines *access* in terms of Euclidean distances, while there is limited evidence from time-varying travel time (i.e., accounting for changes in the network) and the more comprehensive generalized travel cost. Fig. 2 shows the mean value of the output elasticity for market potential measures and definition of transport network. On average, the effect of transport accessibility on firm output appears to be stronger when based on generalized travel cost (i.e., accounting for both distance and travel time) and lowest for time-invariant travel time (i.e., assuming the transport network is fixed over time). One obvious limitation of the empirical literature is its exclusive focus on the road network, ignoring the other transport networks such as railways and urban mass transit.

Identification Issues and Strategies

There are three main identification problems faced by the empirical literature interested in the causal effect of transport accessibility on firm output and productivity. First, there can be spatial sorting bias because more productive firms may self-select into locations with better transport networks. The second problem is the nonrandom allocation of transport infrastructure across locations, that is, investment may be assigned to alleviate bottlenecks of higher productivity areas or/and encourage the growth of lower productivity areas. Finally, there can be other, omitted, factors that explain both better transport accessibility and higher firm productivity (e.g., urbanization level and type of industrial specialization). All three cases result in endogeneity bias and lead to inconsistent estimates of the transport effect.

While the use of firm fixed-effects can control for time-invariant unobserved heterogeneity (e.g., manager ability) influencing both productivity and choice of location, it does not address the problem of simultaneity bias. The latter has typically been addressed through instrumental variables (IV), whereby instruments need to be relevant (i.e., correlated with transport endowment) and orthogonal (i.e., uncorrelated with firm output and productivity). Currently, the most popular approach is to use a combination of historical instruments (e.g., Roman roads, old postal routes, planned routes, and past settlement population) with geographical and geological instruments (e.g., proximity to rivers, soil fertility, terrain ruggedness). The combination of these types of instruments can address reverse simultaneity bias between productivity and market potential because it instruments the size of the local and surrounding areas (i.e., the opportunities for interaction) and the transport cost (i.e., distance, travel time) between them.

Although less common amongst firm-level studies, there are alternative identification approaches based on the “inconsequential units” strategy. The strategy works by comparing outcomes between two groups of units receiving the transport improvement, where one of the groups consists of incidental beneficiaries (e.g., locations between large cities) for whom the outcome status is uncorrelated with the transport treatment. Recent work by Gibbons et al. (2019) using ward- and establishment-level data for Britain used the variation in the changes in accessibility amongst firms located within 1–20 km from a given transport project as the identification strategy. The argument used by the authors is that while the location of transport schemes is likely to be endogenous, the variation in accessibility experienced by firms close to the transport improvement is not. Under this assumption, they can then calculate the average causal effect of accessibility on firm labor productivity by comparing the outcomes of establishments located

Table 2 Summary of elasticity estimates for studies using IV to address simultaneity bias is transport allocation*Transport measure based on market potential type accessibility indicators*

<i>Uses IV</i>	<i>Count</i>	<i>Min</i>	<i>P10</i>	<i>P25</i>	<i>Mean</i>	<i>Median</i>	<i>P75</i>	<i>P90</i>	<i>Max</i>	<i>CV</i>
Yes	27	−0.003	0.023	0.030	0.045	0.047	0.059	0.071	0.074	0.404
No	287	−0.277	−0.052	0.019	0.083	0.064	0.141	0.278	0.491	1.537
Full sample	314	−0.277	−0.042	0.023	0.080	0.060	0.127	0.265	0.491	1.536
Yes	27	−0.003	0.023	0.030	0.045	0.047	0.059	0.071	0.074	0.404
No	297	−0.213	−0.042	0.022	0.083	0.064	0.136	0.265	0.404	1.405
Sample without outliers	306	−0.213	−0.033	0.024	0.080	0.060	0.121	0.251	0.404	1.405

Transport measure based on distance to transport network/node/access point

<i>Uses IV</i>	<i>Count</i>	<i>Min</i>	<i>P10</i>	<i>P25</i>	<i>Mean</i>	<i>Median</i>	<i>P75</i>	<i>P90</i>	<i>Max</i>	<i>CV</i>
Yes	105	−709.300	−0.124	−0.028	−6.183	−0.015	0.015	0.206	51.826	−11.234
No	35	−0.164	−0.072	−0.017	−0.019	−0.013	−0.007	0.022	0.079	−2.282
Full sample	140	−709.300	−0.096	−0.027	−4.642	−0.014	0.006	0.081	51.826	−12.956
Yes	103	−0.349	−0.099	−0.028	0.080	−0.015	0.015	0.116	4.214	7.213
No	35	−0.164	−0.072	−0.017	−0.019	−0.013	−0.007	0.022	0.079	−2.282
Sample without outliers	138	−0.349	−0.092	−0.027	0.055	−0.014	0.006	0.079	4.214	9.127

close to transport schemes, after controlling for unobservable heterogeneity for each scheme and local (i.e., ward-level) characteristics. A similar approach has been used by studies using more aggregate spatial data (e.g., municipalities in Spain, districts in India) whereby locations with motorway nodes are treated as endogenous, while non-nodal locations in the vicinity of the nodal ones are taken as exogenous and their outcomes can thus be compared with those of non-nodal locations located further away (e.g., [Martín-Barroso et al., 2015](#)).

Table 2 compares output elasticities of transport accessibility between studies depending on whether they use IV or not. A tentative conclusion from the evidence in the table is that correcting for simultaneity bias revises the magnitude of the elasticity estimates downward for market potential type accessibility indicators, while the impact is less clear for measures of transport accessibility based on the distance to nearest access point to the network. The mean (median) value of the output elasticity with respect to market potential is 0.083 (0.064) for non-IV studies compared to 0.045 (0.047) for IV studies.

Industrial Heterogeneity

The scope for productivity effects from transport improvements is also likely to vary across industries due to differences in the relative importance of proximity to input (both primary and intermediate) and output markets. Here too, there is a clear linkage to the role played by urban agglomeration economies, in particular, by raising productivity through labor market pooling, input sharing, and knowledge spillovers. Existing evidence shows that knowledge intensive and high-skill sectors benefit more from central city urban locations compared to other service sectors and manufacturing industries, and thus may also benefit comparatively more from transport improvements. **Fig. 3** shows the mean value of the output elasticity with respect to market potential for all industries and the more aggregate economic sectors, while **Table 3** considers a more disaggregate breakdown of industries. On average, and in

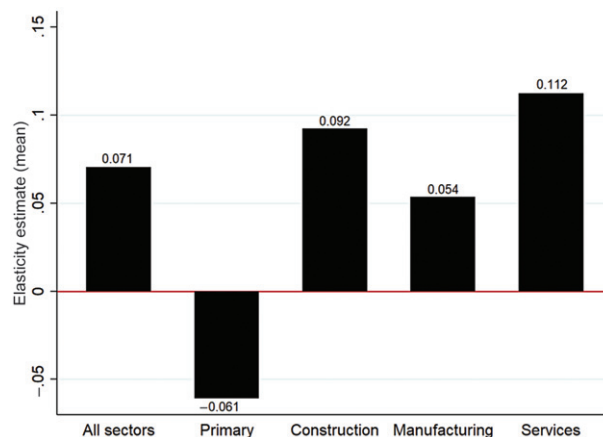
**Figure 3** Summary of elasticity estimates by main industry group for market potential type transport measures.

Table 3 Summary of elasticity estimates by industry group for market potential type and distance type transport measures*Transport measure based on market potential type accessibility indicators*

Industry	Count	Min	P10	P25	Mean	Median	P75	P90	Max	CV
All industries	13	0.016	0.025	0.041	0.071	0.049	0.066	0.105	0.278	0.953
Pooled primary	8	-0.103	-0.103	-0.071	-0.061	-0.058	-0.047	-0.033	-0.033	-0.354
Mining and electricity, gas, and water	4	-0.015	-0.015	-0.007	0.020	0.003	0.048	0.090	0.090	2.336
Construction	21	-0.163	0.023	0.032	0.092	0.095	0.164	0.212	0.243	1.032
Pooled manufacturing	53	-0.132	0.008	0.023	0.034	0.041	0.056	0.071	0.104	1.076
Manu—basic metals and alloys industries	7	-0.015	-0.015	0.004	0.038	0.060	0.061	0.063	0.063	0.852
Manu—electronics, office machinery, computers, radio, television and communication equip, and repair capital goods	13	0.115	0.133	0.168	0.272	0.317	0.356	0.382	0.404	0.393
Manu—food, beverages, and tobacco	7	-0.213	-0.213	-0.160	-0.034	-0.030	0.084	0.087	0.087	-3.389
Manu—chemical and pharmaceutical products	5	-0.008	-0.008	0.019	0.030	0.027	0.029	0.081	0.081	1.091
Manu—medical, precision and optical instruments, and watches and clocks	2	-0.213	-0.213	-0.213	-0.202	-0.202	-0.191	-0.191	-0.191	-0.077
Manu—printing and reproduction of recorded media	7	-0.016	-0.016	-0.003	0.084	0.102	0.144	0.154	0.154	0.803
Manu—pulp, paper, and paper prods	5	0.027	0.027	0.111	0.103	0.121	0.121	0.135	0.135	0.421
Manu - Rubber and plastic prods	5	-0.166	-0.166	-0.155	-0.144	-0.144	-0.135	-0.122	-0.122	-0.118
Manu—textiles, wearing apparel, dying, dressing, and shoes	9	-0.127	-0.127	0.062	0.046	0.072	0.091	0.121	0.121	1.824
Manu—wood and wood prods	5	0.069	0.069	0.078	0.094	0.084	0.092	0.148	0.148	0.332
Manu—motor vehicles and transport equip	5	0.001	0.001	0.002	0.050	0.041	0.085	0.121	0.121	1.051
Manu—other manufacturing	4	-0.019	-0.019	-0.011	0.073	0.074	0.157	0.162	0.162	1.334
Serv - Transport and storage	14	0.009	0.041	0.086	0.205	0.244	0.299	0.345	0.350	0.584
Serv—wholesale and retail trade	16	-0.003	0.009	0.033	0.070	0.062	0.081	0.167	0.203	0.821
Serv—consumer services	7	-0.008	-0.008	0.015	0.039	0.029	0.071	0.114	0.114	1.031
Serv—hotel, restaurants, and catering	11	0.014	0.041	0.044	0.111	0.123	0.151	0.178	0.224	0.580
Serv—ICT	7	-0.090	-0.090	-0.042	0.021	0.034	0.082	0.089	0.089	3.171
Serv—banking, finance, and insurance	11	0.029	0.080	0.114	0.194	0.179	0.251	0.329	0.374	0.535
Serv—business services	22	-0.135	-0.048	0.036	0.086	0.084	0.170	0.226	0.298	1.229
Serv—real estate	11	0.018	0.027	0.040	0.076	0.053	0.114	0.136	0.180	0.654
Serv—communication services	16	-0.130	-0.046	0.049	0.106	0.086	0.201	0.279	0.286	1.101
Serv—public services	13	0.026	0.043	0.096	0.215	0.292	0.318	0.368	0.399	0.648
Serv—other services	5	-0.059	-0.059	-0.013	-0.007	-0.008	0.023	0.024	0.024	-5.139
Sample without outliers	306	-0.213	-0.033	0.024	0.080	0.060	0.121	0.251	0.404	1.405

Transport measure based on distance to transport network/node/access point

Industry	Count	Min	P10	P25	Mean	Median	P75	P90	Max	CV
Pooled manufacturing	42	-0.022	-0.019	-0.017	-0.015	-0.015	-0.013	-0.010	-0.005	-0.244
Manu—basic metals and alloys industries	14	-0.027	-0.027	-0.006	0.010	0.008	0.022	0.044	0.083	2.970
Manu—electronics, office machinery, computers, radio, television and communication equip, and repair capital goods	12	-0.294	-0.284	-0.069	0.084	0.003	0.161	0.378	1.055	4.238
Manu—food, beverages, and tobacco	14	-0.164	-0.141	-0.092	-0.045	-0.028	0.001	0.009	0.022	-1.286
Manu—chemical and pharmaceutical products	4	-0.289	-0.289	-0.208	-0.131	-0.105	-0.055	-0.026	-0.026	-0.863
Manu—printing and reproduction of recorded media	6	-0.084	-0.084	-0.029	-0.022	-0.008	-0.004	0.002	0.002	-1.475
Manu—pulp, paper, and paper prods	4	-0.022	-0.022	-0.017	0.084	0.030	0.184	0.297	0.297	1.768
Manu—rubber and plastic prods	5	-0.004	-0.004	0.004	0.069	0.020	0.021	0.304	0.304	1.910
Manu—textiles, wearing apparel, dying, dressing, and shoes	11	-0.082	-0.076	-0.072	0.067	-0.038	0.222	0.318	0.461	2.810
Manu—wood and wood prods	8	-0.124	-0.124	-0.025	-0.014	-0.001	0.018	0.029	0.029	-3.594
Manu—motor vehicles and transport equip	3	-0.135	-0.135	-0.135	0.005	0.044	0.106	0.106	0.106	25.029
Manu—other manufacturing	15	-0.349	-0.161	-0.145	0.465	-0.017	-0.002	3.773	4.214	3.090
Sample without outliers	138	-0.349	-0.092	-0.027	0.055	-0.014	0.006	0.079	4.214	9.127

agreement with economic theory, the productivity gains from transport improvements are positive and larger for services, followed by construction and manufacturing, and are negative for the primary sector.

Spatial Heterogeneity

Given the network nature of transport infrastructure, there can be scope for cross-border effects affecting both connected and non-connected areas. The nature of these effects is not a priori straightforward and will depend on complex local area characteristics, which can be either supported or hindered by better transport connections. New Economic Geography (NEG) arguments based on the core-periphery model with labor mobility find support from empirical evidence showing that reductions in interregional transport costs can encourage relocation of firms and people from more peripheral and poorer regions to more developed urban regions with some variation across industries depending on the degree to which they are likely to benefit from agglomeration externalities. Consequently, the overall regional impact of transport improvements is likely to depend on the local economic structure.

There are very limited number of microeconomic studies investigating the nature and scope for spatial effects of transport improvements on firm output and productivity. [Graham and van Dender \(2011\)](#) and [Holl \(2016\)](#) are some of the exceptions. [Graham and van Dender \(2011\)](#) tested for heterogeneity in accessibility effects across subsamples for British conurbations, big urban areas (250,000 people or more), and large urban areas (100,000 or more people) using semiparametric techniques. They found evidence of nonlinearities in the productivity–accessibility relationship, and that for conurbations and big urban areas higher market accessibility was not systematically associated with increased productivity, while the relationship was positive and significant for large urban areas.

[Holl \(2016\)](#) investigated whether firm total factor productivity effects from improved highway access differed across the urban hierarchy and found no significant highway effects on productivity for firms located in rural municipalities, while in suburban municipalities firms with access to a highway within 10 km (10–20 km) had productivity gains 18% higher (16% lower) than more distant firms (i.e., more than 20 km from the nearest new highway). She interpreted this result as evidence of negative spillovers in suburban municipalities: while the productivity of highway firms (<10 km) benefited from improved highway access, the productivity of adjacent firms (10–20 km) was hindered by it. Furthermore, firms in urban municipalities also showed a positive productivity-highway access effect, but it was weaker (i.e., 11%) compared to firms in suburban municipalities.

Overall, and so far, the existing literature provides very limited knowledge on the scope for spatial spillover effects from transport improvements on firms' productivity, and the degree to which they are determined by firm selection and relocation (i.e., entry and exit) mechanisms. Evidence from the literature studying the relation between transport accessibility and the spatial organization of economic activity suggests that within large cities and metropolitan regions (intra-regional) relocation is likely to be as important as urban growth and new activity creation, while intercity and interregional effects from transport improvements are more likely to result from displacement of activities from peripheral regions to more central ones.

Conclusions

There is growing evidence from microeconomic studies on the effect from transport improvements on firm output and productivity. Compared to more aggregate macroeconomic approaches, these studies tend to use more realistic measures of the transport network and how changes made to it impact on firm-level accessibility to economic opportunities. They also present advantages in terms of the empirical identification strategies that can be implemented to address issues of spatial sorting and simultaneity bias, which hinder causal inference in this literature.

This review proposes a number of tentative conclusions about the effects of transport improvements on firm output and productivity. It is common practice to use market potential type measures of transport improvements, which reflect the extent to which changes in the transport cost improve access to economic opportunities around firms' location. However, transport cost has often been measured with physical distance or time-invariant travel time instead of real transport network travel times, and the evidence reviewed shows that this choice can affect results. Many studies have attempted to address identification issues relating to firm self-selection and reverse causality by using a mix of approaches, ranging from firm-level fixed-effects estimator, instrumentation based on historical, geographical, and geological instruments, but also, although to less extent, some form of the "inconsequential units" strategy. The evidence shows there is considerable heterogeneity both between and within main economic sectors, with overall effects higher for services, followed by construction, manufacturing, and primary sector industries. With regard to the scope for spatial effects from transport improvements, there is still limited evidence on the degree to which the effect of transport improvements on firms' productivity results from firm selection and relocation and/or growth effects and this relation play out across different areas depending on their characteristics, including the development stage of the transport system. Evidence from the related literature on the relationship between transport and the spatial organization of economic activity suggests that within large cities and metropolitan regions (intra-regional) relocation is likely to be as important as urban growth and new activity creation, while intercity and interregional effects from transport improvements are more likely to result from displacement of activities from peripheral regions to more central ones.

There are important challenges needing attention from scholars. With respect to the measurement of transport accessibility, and although there has been substantial improvement, future research needs to consider transport networks beyond roads, in particular

railway and urban mass transit networks. This presents several data-related challenges but cannot be overlooked, especially within metropolitan areas and large cities. While railways can support higher density and less dispersed urban arrangements, motorways tend to promote suburbanization, which in turn can reduce the scope for urbanization economies. While there are several studies of the effect of railways on the spatial organization of economic activity and relocation effects, we lack studies of its effects on firm output and productivity. In addition, the literature has devoted insufficient attention to the study of the spatial effects of transport improvements on firm output and productivity, and how much of the overall net effect is explained by relocation or new growth. Does better transport make all firms more productive or does it improve the productivity of the fittest (selection) and attract more productive firms which replace older less productive ones (displacement)? How do these effects play out across different industries and regions? A better understanding of these complex dynamics is important, not only from an academic perspective, but also from a policy perspective.

Acknowledgments

Support by FCT (Fundação para a Ciência e a Tecnologia), Portugal, is gratefully acknowledged. This article is part of project PTDC/EGE-ECO/28805/2017. UECE (Research Unit on Complexity and Economics) is financially supported by FCT (Fundação para a Ciência e a Tecnologia), Portugal.

Appendix A

List of microeconomic studies reviewed in this article:

- Gibbons, S., Lyytikäinen, T., Overman, H.G., Sanchis-Guarner, R., 2019. New road infrastructure: the effects on firms. *J. Urban Econ.* 110, 35–50.
- Graham, D.J., 2007a. Agglomeration, productivity and transport investment. *J. Transp. Econ. Policy* 41 (3), 317–343.
- Graham, D.J., 2007b. Variable returns to agglomeration and the effect of road traffic congestion. *J. Urban Econ.* 62, 103–120.
- Graham, D.J., 2009. Identifying urbanisation and localisation externalities in manufacturing and service industries. *Pap. Reg. Sci.* 88, 63–84.
- Graham, D.J., Kim, H.Y., 2008. An empirical analytical framework for agglomeration economies. *Ann. Reg. Sci.* 42, 267–289.
- Graham, D.J., Gibbons, S., Martin, R., 2010. The spatial decay of agglomeration economies: estimates for use in transport appraisal. Final Report. Imperial College London and LSE.
- Graham, D.J., van Dender, K., 2011. Estimating the agglomeration benefits of transport investments: some tests for stability. *Transportation* 38, 409–426.
- Holl, A., 2012. Market potential and firm-level productivity in Spain. *J. Econ. Geogr.* 12 (6), 1191–1215.
- Holl, A., 2016. Highways and productivity in manufacturing firms. *J. Urban Econ.* 93, 131–151.
- Lall, S.V., Shalizi, Z., Deichmann, U., 2004. Agglomeration economies and productivity in Indian industry. *J. Dev. Econ.* 73, 643–673.
- Le Néchet, F., Melo, P.C., Graham, D.J., 2012. The role of transport induced agglomeration effects on firm productivity in mega-city regions: evidence for Bassin Parisien. *Transp. Res. Rec. J. Transp. Res. Board* 2307, 21–30.
- Mare, D.C., Graham, D.J., 2013. Agglomeration elasticities and firm heterogeneity. *J. Urban Econ.* 75, 44–56.
- Martín-Barroso, D., Núñez-Serrano, J.A., Velázquez, F.J., 2015. The effect of accessibility on productivity in Spanish manufacturing firms. *J. Reg. Sci.* 55, 708–735.

Appendix B

Table B1 Summary of firm and plant-level microeconomic studies by type of transport measure (full sample)

<i>Transport measure based on market potential type accessibility indicators</i>											
<i>Author, year</i>	<i>Journal</i>	<i>Count</i>	<i>Min</i>	<i>P10</i>	<i>P25</i>	<i>Mean</i>	<i>Median</i>	<i>P75</i>	<i>P90</i>	<i>Max</i>	<i>CV</i>
Gibbons et al., 2019	<i>Journal of Urban Economics</i>	2	0.10	0.10	0.10	0.19	0.19	0.28	0.28	0.28	0.67
Graham and Kim, 2008	<i>Annals of Regional Science</i>	9	0.01	0.01	0.03	0.13	0.14	0.17	0.31	0.31	0.82
Graham and van Dender, 2011	<i>Transportation</i>	21	−0.16	−0.03	0.04	0.06	0.06	0.11	0.13	0.35	1.60
Graham et al., 2010	<i>Report</i>	17	0.02	0.02	0.02	0.04	0.03	0.04	0.09	0.09	0.63
Graham, 2007a	<i>Journal of Transport Economics and Policy</i>	27	−0.19	−0.04	0.03	0.11	0.09	0.19	0.30	0.38	1.28
Graham, 2007b	<i>Journal of Urban Economics</i>	19	0.04	0.04	0.09	0.20	0.21	0.27	0.35	0.40	0.54
Graham, 2009	<i>Papers in Regional Science</i>	108	−0.28	−0.12	0.01	0.10	0.08	0.19	0.35	0.49	1.66
Holl, 2012	<i>Journal of Economic Geography</i>	21	−0.08	0.04	0.04	0.05	0.05	0.07	0.07	0.10	0.69
Lall et al., 2004	<i>Journal of Development Economics</i>	18	−0.21	−0.16	−0.02	0.02	0.00	0.13	0.16	0.17	5.19
Le Néchet et al., 2012	<i>Transportation Research Record</i>	8	0.00	0.00	0.01	0.02	0.02	0.03	0.05	0.05	0.68
Mare and Graham, 2013	<i>Journal of Urban Economics</i>	48	−0.10	−0.01	0.03	0.05	0.05	0.07	0.11	0.18	1.04
Martín-Barroso et al., 2015	<i>Journal of Regional Science</i>	16	0.00	0.01	0.01	0.03	0.02	0.04	0.06	0.06	0.71
Total		314	−0.28	−0.04	0.02	0.08	0.06	0.13	0.27	0.49	1.54

Transport measure based on distance to transport network/node/access point

(continued)

Transport measure based on distance to transport network/node/access point

Author, year	Journal	Count	Min	P10	P25	Mean	Median	P75	P90	Max	CV
Author, year	Journal	Count	Min	P10	P25	Mean	Median	P75	P90	Max	CV
Holl, 2016	<i>Journal of Urban Economics</i>	122	−709.30	−0.09	−0.03	−5.32	−0.01	0.01	0.11	51.83	−12.10
Lall et al., 2004	<i>Journal of Development Economics</i>	18	−0.16	−0.14	−0.06	−0.03	−0.01	0.00	0.05	0.08	−2.44
Total		140	−709.30	−0.10	−0.03	−4.64	−0.01	0.01	0.08	51.83	−12.96

Table B2 Summary of firm- and plant-level microeconomic studies by type of transport measure (sample excluding outliers)*Transport measure based on market potential type accessibility indicators*

Author, year	Journal	Count	Min	P10	P25	Mean	Median	P75	P90	Max	CV
Gibbons et al., 2019	<i>Journal of Urban Economics</i>	2	0.099	0.099	0.099	0.189	0.189	0.278	0.278	0.278	0.671
Graham and Kim, 2008	<i>Annals of Regional Science</i>	9	0.014	0.014	0.027	0.130	0.141	0.170	0.306	0.306	0.822
Graham and van Dender, 2011	<i>Transportation</i>	21	−0.163	−0.026	0.040	0.063	0.064	0.105	0.127	0.345	1.600
Graham et al., 2010	Report	17	0.015	0.021	0.024	0.042	0.031	0.044	0.089	0.089	0.628
Graham, 2007a	<i>Journal of Transport Economics and Policy</i>	27	−0.191	−0.042	0.030	0.106	0.090	0.191	0.298	0.382	1.277
Graham, 2007b	<i>Journal of Urban Economics</i>	19	0.041	0.042	0.089	0.195	0.214	0.274	0.350	0.399	0.543
Graham, 2009	<i>Papers in Regional Science</i>	100	−0.213	−0.068	0.015	0.097	0.084	0.180	0.312	0.404	1.417
Holl, 2012	<i>Journal of Economic Geography</i>	21	−0.079	0.041	0.042	0.050	0.047	0.066	0.074	0.104	0.685
Lall et al., 2004	<i>Journal of Development Economics</i>	18	−0.213	−0.160	−0.019	0.022	0.001	0.133	0.162	0.171	5.187
Le Néchet et al., 2012	<i>Transportation Research Record</i>	8	−0.003	−0.003	0.014	0.022	0.023	0.028	0.047	0.047	0.676
Mare and Graham, 2013	<i>Journal of Urban Economics</i>	48	−0.103	−0.008	0.028	0.052	0.051	0.074	0.114	0.180	1.036
Martín-Barroso et al., 2015	<i>Journal of Regional Science</i>	16	0.002	0.008	0.013	0.027	0.024	0.042	0.057	0.059	0.710
Total		306	−0.213	−0.033	0.024	0.080	0.060	0.121	0.251	0.404	1.405

Transport measure based on distance to transport network/node/access point

Author, year	Journal	Count	Min	P10	P25	Mean	Median	P75	P90	Max	CV
Holl, 2016	<i>Journal of Urban Economics</i>	120	−0.349	−0.088	−0.027	0.067	−0.014	0.006	0.095	4.214	8.010
Lall et al., 2004	<i>Journal of Development Economics</i>	18	−0.164	−0.141	−0.057	−0.025	−0.008	−0.001	0.051	0.079	−2.442
Total		138	−0.349	−0.092	0.027	0.547	−0.014	0.006	0.079	4.214	9.127

Further Reading

- Boarnet, M.G., 1997. Highways and economic productivity: interpreting recent evidence. *J. Plan. Lit.* 11, 476–486.
- Gillen, D.W., 1996. Transportation infrastructure and economic development: a review of recent literature. *Logistics Transp. Rev.* 32, 39–62.
- Jiang, B., 2001. A review of studies on the relationship between transport infrastructure investments and economic growth. Canada Transportation Act Review Panel, Vancouver, British Columbia, Canada.
- Lakshmanan, T.R., 2011. The broader economic consequences of transport infrastructure investments. *J. Transp. Geogr.* 19, 1–12.
- Melo, P.C., Graham, D.J., Brage-Ardao, R., 2013. The productivity of transport infrastructure investment: a meta-analysis of empirical evidence. *Reg. Sci. Urban Econ.* 43, 695–706.
- Melo, P.C., Graham, D.J., Noland, R.B., 2009. A meta-analysis of estimates of urban agglomeration economies. *Reg. Sci. Urban Econ.* 39, 332–342.
- Redding, S., Turner, M., 2015. Transportation costs and the spatial organization of economic activity. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), *Handbook of Urban and Regional Economics*, Vol. 5. Elsevier, Amsterdam.

Wider Economic Impacts of Transport Investments

Anthony J. Venables, Department of Economics, University of Oxford, Oxford, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	355
The Spatial Context	355
Wider Impacts I: Connectivity and Productivity	356
Wider Impacts II: Induced Private Investment and Land-Use Change	357
Market Imperfections and “Small” Changes	357
Price Change and Transfer of Surplus	357
Agglomeration and Coordination Failure	358
Conclusions	359
References	359
Further Reading	359

Introduction

The economic impacts of transport improvements can be classified into three types. First are the direct benefits of the improvement in terms of vehicle operating costs, timesavings, and other benefits to existing and new users, together with the costs of noise and pollution produced. Second are the economic benefits deriving from better connectivity, holding constant the spatial distribution of economic activity; for example, the flow of knowledge between places may improve with benefits for productivity. Third are the changes in the location of economic activity arising as transport improvements change the spatial pattern of private sector investment; new investments may be induced in some places, possibly at the expense of other places.

A standard cost–benefit analysis (CBA) focuses on the first of these. This is not because analysts are unaware of the other impacts, but because analysis of a project is generally set in the context of a wider economy with few or no externalities or other market failures.^a In such an economy changes in quantities, such as new patterns of investment, are of zero social value; the marginal benefit of an activity is equal to its marginal cost. The terms “wider economic impacts” or “wider benefits” are used to describe the second and third types of effects and are the subject of this paper.

Analysis of the effect of transport investments has to contain two elements. One is to establish the quantity effects of the improvement; this means the changes in economic quantities—numbers of journeys made, new investments undertaken, and jobs created—arising because of the investment. This assessment may need to be quite broad because creation of a job in one place may be at the expense of a job lost somewhere else; establishing whether relocation and displacement effects of this type occur is crucially important. The other is to place a social value on these quantity changes; this is relatively straightforward if everything is traded at fixed prices in competitive markets, and more complicated otherwise.

To explore these issues this paper proceeds in three stages. The next section offers a brief review of some of the elements of spatial economics, fundamental to transport investments. Section “Wider Impacts I: Connectivity and Productivity” looks at wider impacts arising from changes in connectivity and productivity, given patterns of land-use and the spatial distribution of activity. Section “Wider Impacts II: Induced Private Investment and Land-Use Change” looks at ways in which transport improvement can induce changes in private investment and land-use, and at the value of any such changes. Throughout, the focus is on identifying and quantifying wider impacts for the purpose of ex ante appraisal of transport investments. The literature on ex-post evaluation of investments is reviewed in Redding and Turner (2015).

The Spatial Context

Economic activities are spatially “lumpy”, tending to concentrate in space. This is partly due to the minimum efficient scale of some business activities (e.g., large factories that supply a wide geographical area) and partly due to agglomeration economies. These arise if there are complementarities and increasing returns to scale operating across a range of firms and workers who are in close proximity to each other. They are driven by a number of mechanisms. Large local labor markets enable better matching of workers to firms’ skill requirements. Better communication between firms and their customers and suppliers enables knowledge spillovers, better product design, and timely production. A larger local market enables development of a larger network of more specialized suppliers. Fundamentally, larger and denser markets allow for both scale and specialization, thereby creating areas of relatively high productivity.

^aThere is an extensive literature on market failure and externalities within the direct context of the project itself. Imperfections in labor markets and in international trade policy were central to an older literature on shadow pricing in developing economies (Dreze and Stern, 1987).

Agglomeration economies typically involve externalities between firms, and between firms and workers. To take one example, the larger the market, the more likely it is to be worthwhile for an individual to specialize and hone skills in producing a particular good or service, thereby achieving high productivity. The specialist will be paid for the product or service supplied but, depending on market conditions, is unlikely to capture the full benefit created.^b Since the benefit is split between the supplier and her customers, there is a positive externality. This creates a positive feedback—more customers will be attracted to the place to receive the benefit, growing the market, further increasing the returns to specialization, and so on. This is the classic process of cluster formation.

The spatial and sectoral range of agglomeration economies varies according to the context. They are important within particular sectors, particularly those prone to cluster (e.g., financial services and high tech) and are referred to as localization, or Marshall–Romer economies. They also operate at a wider urban level, referred to as urbanization or Jacobs economies.

Agglomeration economies are a benefit, but proximity also has costs, particularly in the urban context. Clustering of firms increases commuting costs for workers who may have to travel far to employment centers such as the central business district. These costs are exacerbated by congestion—a negative externality—and other costs of dense urban living. Land becomes the scarce factor and housing consumption (floor space per household) is reduced. However, it is important to note that high land rents and land prices are not a real cost; these are a transfer payment from occupants of land to its owners, so do not use up real resources (as does time in commuting).

The location of economic activity is determined by the interplay between the benefits and costs of proximity, together with location fundamentals (including “first-nature” physical geography). Since agglomeration effects are largely externalities, location outcomes are generally not efficient, and there may be situations in which a place is stuck in a low-level equilibrium. These facts create the potential for transport improvements to be “transformational,” bringing large quantity changes and perhaps also large welfare gains. We will discuss these effects in section “Wider Impacts II: Induced Private Investment and Land-Use Change,” but note that they are, in practice, extremely hard to predict. The practical literature applying wider benefits to the practice of CBA has largely taken a more conservative route, looking at improvements in connectivity given the location of activity and the pattern of land-use. It is to this that we now turn.

Wider Impacts I: Connectivity and Productivity

Transport investments make places closer together, in economic terms, potentially increasing productivity, thereby also creating a wider impact. To capture this, a quantitative measure of the importance of agglomeration economies is needed, and many studies have produced estimates.^c A simple approach to this quantification involves two steps. First, compute a measure of the economic scale and density of a place, and, second, regress productivity on this measure. Scale and density are typically calculated by a measure such as “effective density” defined as $ED_i = \sum_j f(d_{ij}) \text{employment}_j$. This says that the effective density of place i is the sum of employment in other places, j , weighted by an inverse function of the distance between i and j , d_{ij} , with $f(\cdot)$, a decreasing function. At the second stage, the elasticity of productivity with respect to effective density is estimated, typically by regressing productivity in each place on its measure of effective density and other factors.

Doing this across cities, the first finding is that the elasticity of productivity with respect to city size is of the order of 0.05–0.1. Thus, a city of 5 million inhabitants typically has productivity of 12%–26% higher than a city of ½ million. This raw number has been refined in many directions, the literature on which is surveyed by [Rosenthal and Strange \(2004\)](#) and [Combes and Gobillon \(2015\)](#). For example, measured productivity depends on the skill and occupational mix of the labor force in each place. Controlling for observed measures of skill (e.g., education) typically brings the elasticity down by about a quarter. This might only be part of the story, as there may be sorting, meaning that people with higher innate ability (regardless of education) are disproportionately drawn to cities. The only way to observe this is to track individuals as they move in to cities or to cities of different sizes. Recent empirical work doing this suggests that the pure agglomeration effect has elasticity of 0.02–0.04. This is still a substantial number.^d

This framework has been applied to transport appraisal, notably in the WebTAG guidance of the UK Department for Transport. A transport investment that cuts journey times reduces effective economic distance between places, d_{ij} , and hence (given employment) increases the effective density of such places. This raises the productivity of *all* workers in these areas according to the estimated elasticity of productivity with respect to effective density. The productivity increment is a welfare gain that is added to the user-benefits derived from a standard transport CBA. These “wider benefits” can amount to around 15%–25% of the total net benefit of a large transport project.

The approach constitutes a well-grounded attempt to add some wider impacts in a rigorous and contained manner, although a number of further issues remain. Application of the method is sometimes mechanical, with little attention to context. There

^b This example is a pecuniary rather than a technological externality, that is, an effect transmitted through an imperfect market. The supplier will capture the full benefit only if able to perfectly price discriminate. Otherwise, the customer will also receive some consumer/user surplus on the introduction of a new product. This observation is central to the wide-range of economic models in which the number and variety of goods and services offered is endogenous.

^c Methods and data sources vary widely, and are surveyed in [Gibbons and Graham \(2018\)](#).

^d In some contexts, it may not be appropriate to impose all these controls. Suppose that someone achieves high productivity by undertaking education to acquire a skill that is in demand only in large cities. Then both the city and the education are necessary for attaining the productivity, and it would be wrong to attribute it all to education ([Combes and Gobillon, 2015](#)).

are further measurement issues, as the strength of localization economies varies by sector, and the spatial range of effects is important (i.e., the form of the function $f(\cdot)$, for example, see [Gibbons and Graham \(2018\)](#) or [Rice et al. \(2006\)](#)). Further problems arise if a transport investment only effects one mode of transport; for example, a rail improvement may only affect a small proportion of total journeys made between two places, so calculation of its impact on an overall measure of “economic distance” is controversial.

Wider Impacts II: Induced Private Investment and Land-Use Change

Large transport improvements generally change the spatial pattern of private investment and hence jobs—indeed, this is often the motivation for undertaking the improvement. Estimating these quantity changes, *ex ante*, generally requires modeling techniques such as computable spatial equilibrium models or land-use transport interaction (LUTI) models. These provide useful quantitative frameworks and can indicate a range of possible outcomes, although they are far from precise predictions. Studies of past transport improvements provide further information, but are fraught with statistical problems.^e There are problems of endogeneity (is a region booming because a road was built, or was the road built because the region was expected to boom?), selection of appropriate control areas, and of distinguishing net aggregate growth from spatial relocation. Above all, there is the problem of context specificity; what we learn from previous transport investments may or may not be relevant to the context of a particular new investment.

Given estimates of quantity changes, the task is to ascribe them social values. The benchmark is “small” changes in a “perfect” economy, in which case the net value of quantity changes is zero. We move beyond this case in several steps, progressing from “small” changes, that is, those that leave prices in affected places broadly unchanged, to large ones.

Market Imperfections and “Small” Changes

Why is it socially valuable to create a job in a particular place, even if it may involve losing a job elsewhere? One answer is that there is involuntary unemployment, so creating a new job expands the total number of people in employment, rather than just relocating jobs. Another is that there are spatial income differentials, so social welfare increases if a job moves to a low-income area. A third is that there are factors that “distort” individuals’ choices about labor-leisure decisions, or about choice of place of work.

An example of the last of these arguments, included in the UK Department of Transport appraisal procedures, is the move to more (or less) productive jobs. This can arise if there are productivity differences between places within a city (or between regions in the country) and a transport improvement causes an increase in employment in one area and a decline in another. Thus, in the intra-urban context, it may be the case that productivity is higher in the city center than at the city edge. This can be an equilibrium, as the combination of commuting costs and high city-center house prices means that individuals are indifferent between taking a high-wage and high-productivity job in the center and a lower wage job at the edge. An improvement in commuting transport infrastructure that leads to more people working in the center raises average productivity but does not necessarily create a net benefit; there is no such “wider benefit” if individuals’ job and location choices are efficient.^f The change only has nonzero value if there is some market failure or distortion that makes private choices inefficient. One such distortion is income tax. Individuals’ choice of place of work depends on posttax income, not pretax (which reflects productivity). It turns out the net social gain from moving people to higher productivity jobs is exactly equal to the additional income tax revenue raised as a consequence of the move.

Price Change and Transfer of Surplus

Transport improvements may trigger “large” changes in investment and activity in an affected place, as with the development of new urban subcenters or attempts to rebalance regional economies. Large developments will change prices, and it then becomes important to understand the effect to which such changes are, or are not, internalized in developers’ decisions. Land value appreciation in the immediate vicinity of the project can be internalized by ownership of the land, but the development might also change land prices over a wider area (perhaps reducing them elsewhere). The development might also reduce prices of goods and services if there is an increase in their supply, or bid up the wage rate if employment expands. If these changes impact on people and firms other than the project developer then there is a divergence between the private and social returns from the development; higher wages or lower prices caused by the development are a benefit to those who receive/pay them, but not to the developer. This divergence means that the development—the indirect quantity change induced by the transport improvement—has nonzero social value.

Implications are illustrated in [Fig. 1](#). The horizontal axis is the cost, t , of accessing a potential new development site (e.g., transport costs borne by customers or employees). If this is very high the development is not undertaken—both private (developer) and social returns are negative, as measured on the vertical axis and indicated by the solid lines.^g Improving access (reducing t) increases both private and social returns. Initial access costs are at point A, and they are reduced by amount Δt by a

^e A good example of this work is [Donaldson \(2018\)](#), and see [Redding and Turner \(2015\)](#) for a review of modern approaches.

^f For purposes of this argument, productivity levels are assumed to be exogenous and fixed. See [Venables \(2007\)](#) for a more general case.

^g In [Fig. 1](#), these are the social returns associated with the price and wage change not those in the standard CBA, which can be thought as being in *net* cost $C(\Delta t)$.

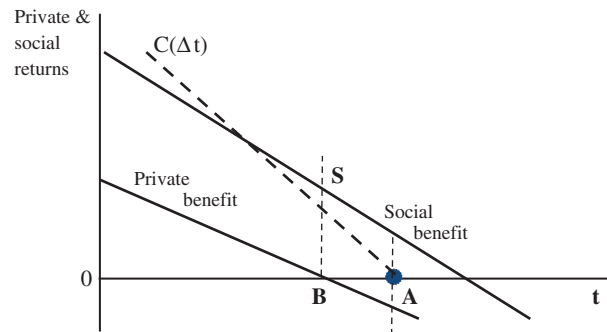


Figure 1 Transport improvement unlocking development.

publicly funded transport investment with net cost $C(\Delta t)$ (larger the greater the transport cost reduction). At initial point A the development is not undertaken; private returns are negative, although social returns are positive, the divergence between the two arising from the price or wage change effects is discussed earlier. A transport improvement reducing access costs to point B ($\Delta t = A - B$) would be just sufficient to trigger the development, yielding “wider benefit” $S - B$, an amount that should be added to a conventional narrow CBA.

In summary, the fact that the development is “large” means that the initial point A is inefficient. Transport improvement is a (second best) way of removing the inefficiency, and thereby creating social gain. This is a “wider impact” of a transport improvement that unlocks development potential.

Agglomeration and Coordination Failure

In section “Wider Impacts I: Connectivity and Productivity,” we saw that transport improvement could raise productivity by increasing effective density; even if economic activity does not move, places become closer in terms of travel time and economic distance, d_{ij} . However, transport improvements generally induce changes in the location of activity that will lead to further productivity. Using the framework set out in section “Wider Impacts I: Connectivity and Productivity,” effective density and hence productivity will change as the terms $employment_j$ in the expression for ED_i change.^h Thus, if a transport improvement—for example, an increase in commuting capacity—enables employment in a cluster of activity to grow then, in the presence of agglomeration economies, the productivity of workers in the cluster (existing workers as well as new) will increase. This has been an important argument informing decisions to improve commuting into city centers, although it is subject to caution for the usual reasons of displacement. If workers are just moving from one cluster to another then the productivity gains in the growing cluster may be cancelled out by losses in the other.ⁱ

A more fundamental issue arises as transport improvement may have as its objective not just to move investment and employment into an area, but also to trigger the establishment of a new cluster of activity. There are important market failures associated with establishing new clusters, deriving from the fact that agglomeration economies are positive reciprocal externalities; firms and workers in a cluster create and receive benefits. This creates the first-mover problem; no one wants to be the first to move to a new place unless they are confident that others will follow. Expectations of future returns are critical (particularly since location decisions are generally long-lived and incur sunk costs), and these depend on who else is (or is expected to be) there. Success requires that expectations are coordinated—so each firm knows that each other firm will move. Absent this, there is “coordination failure.” The private benefit of moving is less than the full social benefit so investment will not take place, similar to being stuck at point A on Fig. 1.

What is the contribution of transport to this? Many conditions have to be met for a place to become attractive for new investment, and it is hard to argue that reducing transport costs alone is sufficient to trigger development of a new cluster. However, a transport improvement can play an important role in coordinating expectations, over and above any transport cost reductions that it may bring. For example, consider a growing city in which it is clear that a new subcenter would be profitable, but there are many potential sites for such a center. If no one knows which site will develop, then no one will invest; the challenge is therefore to create a credible signal that will coordinate expectations. City plans are one instrument to achieve this, but their poor implementation record means that they are unlikely to be credible. Investment in infrastructure is an alternative, credible since such investments are sunk and irreversible.

^h Changes in effective density holding employment constant are sometimes referred to as “static clustering,” and those with employment changing are “dynamic clustering.”

ⁱ If there is full employment, then displacement implies $\sum_j employment_j = 0$. In some circumstances, productivity effects will exactly cancel; see Kline and Moretti (2013). However, they will not cancel out if agglomeration is driven by sector-specific localisation economies and cluster growth enables a high-agglomeration sector to expand.

Conclusions

There is often a tension between the strategic case made for a major transport improvement, and the economic case. The former is likely to be based on assertions about wider impacts, and the latter on a narrow CBA that implicitly ignores all such effects. It is important to develop a rigorous understanding, based on theory and evidence, of the mechanisms, circumstances, and magnitudes of any such wider impacts. This is needed both to make the arguments for wider impacts and to challenge spurious arguments that are sometimes made.

Progress has been made on the conceptual side, laying out the mechanisms through which wider benefits can occur, and in empirical work quantifying some key relationships such as the productivity benefits of density. More needs to be done, building on a variety of techniques including modern quantitative spatial modeling, both to get a better understanding of possible quantity changes brought about by transport investments and to ensure that these are valued in a full general equilibrium setting, where displacement effects are taken into account and the risk of double counting are minimized.

References

- Combes, P.P., Gobillon, L., 2015. The empirics of agglomeration economies. In: Duranton, G., Henderson, V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, Volume 5A. Elsevier, Amsterdam, pp. 247–348.
- Donaldson, D., 2018. Railroads of the Raj: estimating the impact of transportation infrastructure. *Am. Econ. Rev.* 108, 899–934.
- Dreze, J., Stern, N., 1987. The theory of cost-benefit analysis. In: Auerbach, A., Feldstein, M. (Eds.), *Handbook of Public Economics*, Volume 2. Elsevier, Amsterdam.
- Gibbons, S., Graham, D.J., 2018. Quantifying wider economic impacts of agglomeration for transport appraisal: existing evidence and future directions. CEP Discussion Papers dp1561, Centre for Economic Performance, LSE.
- Kline, P., Moretti, E., 2013. Agglomeration economies, and the big push: 100 years of evidence from the Tennessee Valley Authority. *Local economic development. Q. J. Econ.* 129 (1), 275–331.
- Redding, S.J., Turner, M.A., 2015. Transportation costs and the spatial organization of economic activity. In: Duranton, G., Henderson, V., Strange, W.C. (Eds.), *Handbook of Regional and Urban Economics*, volume 5B. Elsevier, Amsterdam, pp. 1339–1398.
- Rosenthal, S., Strange, W.C., 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, Volume 4. Elsevier, Amsterdam.
- Rice, P., Venables, A.J., Patacchini, E., 2006. Spatial determinants of productivity: analysis for the regions of Great Britain. *Reg. Sci. Urban Econ.* 36, 727–752.
- Venables, A.J., 2007. Evaluating urban transport improvements: cost-benefit analysis in the presence of agglomeration and income taxation. *J. Transport Econ. Policy* 41 (2), 173–188.

Further Reading

- Duranton, G., Puga, D., 2004. Microfoundations of urban agglomeration economies. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, Volume 4. Elsevier, Amsterdam.
- Mackie, P., Graham, D.J., Laird, D., 2012. Direct and wider economic benefits in transport appraisal. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *Handbook in Transport Economics*. Edward Elgar, London, pp. 501–526.

Employment Effects of Transport Infrastructure

Anna Matas, Javier Asensio, Universitat Autònoma de Barcelona and Barcelona Institute of Economics, Barcelona, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	360
Theoretical Framework	361
Measure of Accessibility	361
The Role of Distance	361
Transport Mode Choice	362
Matching Workers and Job Vacancies	362
Vacancies and Competition for Them	362
Labor Market Flexibility	362
Econometric Methodology	362
Results and Policy Implications	364
See Also	364
References	364

Introduction

It is well known that investment in transport infrastructure may have positive impacts on economic output. These arise as reductions in transport costs change the location of firms and residents, reinforcing agglomeration economies in those areas that benefit from the project, which in turn have positive effects on employment. The analysis of the impacts of transport infrastructure on the level and the localization of economic activities, as well as the discussion about whether they are a net benefit or just a transfer between areas, are dealt with in other parts of this Encyclopedia. In this chapter, we focus on the effects that a reduction in individual commuting costs may have on their labor market outcomes, particularly focusing on those workers more adversely affected by the physical disconnection from jobs. Commuting costs include both time and monetary costs of accessing the workplace.

Known as the spatial mismatch hypothesis, the analysis of the relationship between access to jobs and labor market outcomes has been a topic of debate in the North-American academic literature since Kain's (1968) seminal contribution. According to his original work, the process of employment decentralization that took place in the US cities like Chicago or Detroit since the 1950s led to a disconnection between residential and job location in the case of racial and ethnic minorities, especially for Afro-Americans, who mostly remained living in the inner city. Housing market discrimination against blacks prevented them from relocating closer to jobs. The underlying hypothesis in Kain's work is that the physical disconnection from jobs could be a key factor in explaining persistent unemployment among Afro-Americans.

Since then, a large number of empirical studies (see the surveys by Ihlandfeldt and Sjoquist, 1998 and Gobillon et al., 2007) have analyzed the relationship between different individual outcomes in the labor market (participation, hours worked, unemployment, or wages) and the characteristics of the place of residence. The latter are not limited to accessibility to jobs, but also include features related to residential segregation. The initial focus on Afro-American communities in the US cities has been extended in several directions. First, to include other disadvantaged groups such as low-educated workers, youth, and women. Second, to other metropolitan areas where sprawl has also taken place. The empirical results confirm that the spatial mismatch hypothesis is relevant in a large number of contexts. A related line of research considers that over-qualification with respect to the job requirements can also be the result, among other factors, of spatially constrained job search. Finally, while the first studies focused on the impact of car accessibility or car ownership, more recent work introduces public transport into the analysis.

Although the empirical work found evidence in favor of the existence of spatial mismatch, sustaining a causal relationship was made difficult by the lack of a theoretical framework and by different methodological problems related to the econometric estimation. More recent work has addressed these problems. On the one hand, theoretical models aimed at explaining spatial mismatch have been developed since the late 1990s. On the other hand, both the econometric methodology and access to better quality data have experienced substantial improvements that avoid biases and spurious relationships.

The rest of the chapter is organized as follows. In section, "Theoretical Framework" we summarize the theoretical framework for the spatial mismatch hypothesis, and in section, "Measure of Accessibility" we discuss the details underlying the measurement of accessibility. Section "Econometric Methodology" explains the econometric methodology and the main problems related to it. Finally, section "Results and Policy Implications" summarizes the main results and derives policy implications.

Theoretical Framework

It was not after the publication of a significant number of empirical papers that the theoretical mechanisms underlying the spatial mismatch hypothesis were developed in the context of urban labor market models (Brueckner and Zenou, 2003; Coulson et al., 2001; Wasmer and Zenou, 2002).

The development of such models made it possible to clarify how disconnection from jobs can result in poor labor market outcomes. Gobillon et al. (2007) provide a review of the mechanisms of spatial mismatch, which are related to both labor supply and demand. From the labor supply point of view, the theoretical logic can be based on either job search costs or reservation wages. Regarding the first explanation, job search efficiency may decrease with distance to jobs since information on job opportunities is lower for distant jobs. Additionally, the unemployed workers will incur increasing transport costs as distance increases and, accordingly, restrict their search to neighboring areas. The latter effect may be particularly important for those individuals dependent on limited public transport. The explanation based on reservation wages considers that potential workers may refuse job opportunities that require commuting costs that are too high in relation to their proposed wage. Therefore, reducing time and/or monetary costs of accessing to workplaces and their related information will increase the probability of finding and accepting a job.

The negative effects of distance can also be explained from the labor demand viewpoint. Given that, as shown by empirical evidence, workers that commute long distances may be less productive and have a higher rate of absenteeism, employers can be less willing to hire them, or will do so paying lower wages.

Therefore, improving accessibility to the workplaces through better transport can positively affect labor market outcomes by increasing search efficiency, reducing the reservation commuting cost, and improving labor productivity. These effects will enlarge the effective size of the labor market and have positive consequences on labor force participation rates; number of hours worked or wages levels, as well as reducing the length of unemployment duration.

However, distance is not the only spatial factor that may explain bad labor market outcomes. Residential segregation may also have an adverse effect through different channels. First, it can lead to labor market discrimination. As argued by Gobillon et al. (2007), employers may discriminate against residents in some specific areas on the basis that they presumably have bad working habits, are dishonest or criminal, although this argument is mostly related to some kind of ethnic discrimination. In the context of the US cities a large debate emerged on the role of race versus distance to jobs to explain the labor market results of the black population located in inner cities. The current consensus seems to be that both race and space contribute to spatial mismatch.

Second, residential segregation can amplify the negative labor market outcomes through neighborhood interactions. A disadvantaged neighborhood may generate lower incentives for young people to invest in education, partially as a consequence of the existence of peer effects at school level. This has labor market effects given that lower human capital investment adversely affects individual labor market outcomes. Besides, segregation causes a deterioration of information networks about jobs, which prevents people with low education and ability levels to benefit from social networking within the community. These arguments have to be understood considering that residential segregation is more related to job disconnection than to physical distance between job and residence locations. In other words, residential segregation does not necessarily increase with distance.

Finally, research on spatial mismatch shows that distance to jobs affects population groups differently, with larger effects in the case of more disadvantaged groups. The literature shows that accessibility is especially important for, among others, unskilled, female, younger, and older workers. The reason why the underlying mechanisms for spatial mismatch have a different effect on a particular group can be explained by how constrained they are by local job opportunities. For instance, in a context of decentralized employment and residence, private vehicle may sometimes be the only available option for accessing workplaces. If that is the case, those groups that have to rely on poor public transport will be constrained to accept local jobs.

In summary, after several decades of research the spatial mismatch hypothesis can be grounded on different theoretical models. However, the magnitude of the contribution of accessibility to labor market outcomes remains an empirical question. In what follows we try to shed some light about the methodological approaches used, the main challenges faced by research in this area and the results obtained.

Measure of Accessibility

There are different methodological issues related to the empirical definition of accessibility measures when testing the spatial mismatch hypothesis. As some authors have pointed out (see Houston, 2005 for a survey) methodological choices may lead to significantly different results. Following Bunel and Tovar (2014) the choices that have to be made when measuring labor market accessibility in this context can be grouped around the following issues:

The Role of Distance

When dealing with the question of which is the impact of geographical distance on job accessibility, the answers fall between the extreme cases, which consider that space is frictionless (and there is a single labor market within which all jobs are equally accessible) or regard distance as an insurmountable barrier (and the labor market is strictly local). In between those (probably irrelevant) cases, different varieties of distance-decay functions can be used to show how accessibility decreases as distance increases. Although taking straight-line distances may appear to be a logical simplification of the problem, it would in fact understate real distances

proportionally more in the case of shorter trips and, as a consequence, it would bias the results against finding evidence in support of the spatial mismatch hypothesis.

A related problem to that of modeling distance is how to deal with the so-called 'frontier effect,' which arises in empirical work when the area under study is geographically truncated, and thus makes it possible that the labor market extends beyond its limits for workers or jobs located sufficiently close to the border. Using French data, [Bunel and Tovar \(2014\)](#) compare accessibility measures obtained at the national scale with those limited to the Paris region. Although the overall measures do not differ by much, there are significant differences for border municipalities.

Transport Mode Choice

The same geographical distances will be subject to different commuting generalized costs depending on the transport modes being available and their service characteristics (including frequencies, congestion, reliability, etc.). Taking into account the ability of different individuals to use different modes leads to the need to consider mode choice models (including decisions on car ownership) that will reveal the importance of individual attributes in the valuation of each mode. The empirical literature has shown that different social groups may have different propensities to commute and that income levels will play a particularly important role in mode choice, given their correlation with values given to travel time savings.

Matching Workers and Job Vacancies

When matching jobs with workers, individual characteristics of both sides matter. Again, this can be simplified in an extreme way assuming that any job could be open to any worker. It is however more realistic to consider that jobs of specific characteristics will be available only to workers with the required educational or skills level. Of course, the level of detail of the specific job-worker matching will depend on available data on both the skill/education requirements of jobs and the workers' endowments, but ideally it would also be necessary to consider the possibility that workers access jobs for which they may be overqualified. When analyzing the spatial mismatch of particular disadvantaged groups, assumptions need to be made to identify the specific job market segments for which they compete, which need not to be the same as those of the overall population.

Vacancies and Competition for Them

Although in the previous paragraphs we have referred to access to jobs, it is clear that what matters to measure accessibility is the location of actual vacancies. However, given that the latter may not be easily observable, a simplifying assumption is often made about location of jobs and vacancies being highly correlated. Related to this, any accessibility measure for vacancies also needs to take into account the potential competition that may exist between workers to occupy them, so vacancy accessibility for a given worker will depend on the number of potential workers who may also access to it. Therefore, when computing an accessibility index, each vacancy should be indirectly weighted according to the pressure that competing workers put on it. This is done by including in the index a measure of accessibility to the vacancy by other workers, in a way that considers vacancies with more rivalry as less accessible.

Labor Market Flexibility

Given the increasing flexibility of the labor market, both the existence of multiple part-time jobs or the lack of proper definition of the location of certain vacancies creates additional challenges for defining and measuring accessibility. Workers on part-time employment looking for an additional job will face accessibility conditions that do not depend only on their residential location. Although in the case of jobs that make it possible to work from home most accessibility issues may seem to be irrelevant, there will still be some need to access certain locations for specific tasks, requiring to define accessibility accordingly.

Econometric Methodology

The most common approach to empirically test the spatial mismatch hypothesis is to relate a measure of labor market outcome to an index of job accessibility controlling for a set of additional explanatory variables, in the form of an econometric model whose parameters can be estimated. Whereas some studies use individual data, others employ aggregate data defined at some detailed spatial level. Although cross-section data is predominant, a growing number of researchers employ longitudinal (panel) data. The econometric specification of different studies vary according to the nature of the dependent variable -discrete, censored, or continuous- and the way they deal with different econometric methodological problems.

In order to identify the role that accessibility plays in terms of labor market outcomes, it is necessary to disentangle the effect of distance to jobs from other spatial and non-spatial factors that may also have labor-related impacts. The equation's specification must, therefore, include a rich set of explanatory variables to control for personal and neighborhood characteristics. The latter play a particularly important role since they control for spatial segregation, which, as already explained, may also be related to employment outcomes. Given the importance of defining these variables at small spatial units, using census tract information has become common practice.

A potentially important problem faced by the proposed econometric strategy is that of the endogeneity of the accessibility variable, which may be caused by two effects. A problem of simultaneity arises when job accessibility influences the result in the labor market, but through residential sorting processes the reverse is also true. Besides, there may be unobserved individual characteristics (for instance, different individual productivity levels) that simultaneously affect both neighborhood location and labor market performance. In both cases, the accessibility index will be correlated with the random error term of the equation and, consequently, the estimates of the effect of job accessibility will be biased. Different strategies have been proposed to deal with this endogeneity problem.

One way to obtain consistent estimates is to use instrumental variables. The underlying idea is to find a variable or set of variables strongly correlated with the explanatory variable (in our case, job accessibility) and, at the same time, with no direct effect on the dependent variable (that is, independent of the random term of the equation). In this way it is possible to extract the exogenous source of variation in the accessibility variable to assess its effect on labor market outcomes. Sometimes the lagged values of the same explanatory variables are used as instruments. However, this is not an advisable option since lagged values will generally also be correlated with the error term. The difficulties in finding instruments that fulfill the necessary requirements have led researchers to look for alternative solutions.

A second approach to deal with endogeneity is to rely on natural or quasi-natural experiments. The idea in this case is to use a situation where the residential location of individuals can be considered as exogenous. Although there are not many such situations, a growing number of researchers are exploiting this idea from different perspectives.

[Aslund et al. \(2010\)](#) make use of a refugee dispersal policy by which the Swedish government assigned refugees to neighborhoods with different degrees of geographic job accessibility. Therefore, individual residential location could be considered exogenous. Moreover, since all individual characteristics used to assign the refugees were observed by the researchers, any excluded individual variable will be uncorrelated with the measure of job accessibility. In this context, the researchers find that having been placed in a location with poor job access adversely affected employment.

[Andersson et al. \(2018\)](#) overcome the methodological limitations by using employer-employee matched data on workers searching for jobs after a mass lay-off. They focus on workers with strong labor attachment who experienced a displacement, resident in different US states adjacent to the Great Lakes. In this case, job searching reasons can be considered exogenous and not related to local job opportunities. Their rich data set make it possible to identify transitions to new jobs as well as to construct personal job accessibility measures and control for demographic and neighborhood characteristics. Their results support the spatial mismatch hypothesis: better job accessibility significantly decreases the duration of unemployment among low-medium income workers. The effect is especially important for blacks, women, and older workers.

Other papers base their quasi-experimental approach on the construction of a new transport infrastructure that can be considered exogenous to individuals' labor market outcomes. The research methodology, known as difference-in-difference (diff-in-diff) consists in comparing, before and after the construction of the infrastructure, the outcomes in the same local labor market of those who benefit from the infrastructure improvement (treated group) from those who do not (untreated group). Data requirements of this methodology are demanding. First, it is necessary to observe individual outcomes before and after the change. Second, information is needed to control for all those characteristics not related to the transport improvement that might differ between the treated and the untreated groups. The robustness of this type of analysis crucially depends on the new infrastructure being exogenous and on controlling for differences among groups other than their accessibility. Some examples of this approach are [Holzer et al. \(2003\)](#) who exploit the construction of a new railway line in the San Francisco Bay Area to evaluate accessibility improvements to jobs in the case of minority workers, [Sari \(2015\)](#)'s investigation on the impact of a new tramway in specific neighborhoods in Bordeaux and [Aslund et al. \(2017\)](#), who assess the effects of the introduction of a commuter train to the employment center in Uppsala.

Other papers restrict their analyses to sub-groups of population whose residence can be considered exogenous with respect to the characteristics of the local job market. An example are young adults still living with their parents ([Raphael, 1998](#)). However, this is an imperfect approximation since the unobserved characteristics of the young adults may be correlated with those of their parents and, therefore, be related to residential location ([Gobillon and Selod, 2014](#)).

A third possibility is the use of longitudinal data to soften the endogeneity problem caused by unobserved individual characteristics that are correlated with both the labor market outcomes and the residential location. Controlling for individual fixed effects can capture those unobserved factors that are common to the individual but do not vary over time. Using this approach [Weinberg et al. \(2004\)](#) conclude that estimates that do not correct for endogeneity overstate the effect of neighborhoods but understate the effect of job accessibility, because individuals with lower employment probabilities are located around central business districts. Given the difficulties of gathering individual longitudinal data, some researches rely on spatially aggregated data and compute average values for the variables at small neighborhood areas. However, when using longitudinal data two issues have to be considered. First, the temporal dynamic adjustment of the model must be taken into account. Second, it is crucial to separate the effects of investment in infrastructure from the effects of changes in the location of the workplaces. These are not always easy tasks.

A last methodological issue worth mentioning is the endogeneity of car ownership. As it has already been mentioned, car availability influences labor market performance but, at the same time, the decision of owning a car is clearly related to labor market outcomes. In order to deal with this simultaneous relationship some authors use instrumental variables while others estimate vehicle ownership and mode choice models.

Results and Policy Implications

After several decades of empirical and theoretical research, there is enough evidence to confirm that poor accessibility to jobs leads to worse labor market outcomes. Advances on the theoretical framework, the econometric methodology, the definition of accessibility measures and the quality of available data have resulted in more reliable results. As an example that combines these features, the recent paper by [Andersson et al. \(2018\)](#) provides robust evidence on the negative effect of accessibility on unemployment duration among low-paid displaced workers for a subset of metropolitan areas in United States, with an especial effect for blacks, women, and older workers.

From the initial studies focused on labor market performance of disadvantaged workers in the US cities, the empirical research has extended into a wider set of cities in Europe ([Patacchini and Zenou, 2005](#) for England; [Aslund et al., 2010](#) and [Norman et al., 2014](#) for Swedish cities; [Matas et al., 2010](#) for Spanish cities; [Korsu and Wenglenski, 2010](#) for Paris-Île-de-France and [Sari, 2015](#) for Bordeaux) and beyond (see the analysis by [Boisjoly et al., 2017](#) of accessibility and informality in Sao Paulo, Brazil, or [Rosbapé and Selod, 2006](#) of unemployment in Cape Town, South Africa). Moreover, whereas the initial focus was on car availability, the role that public transport may play in offsetting the adverse effects of disconnection has received increasing attention ([Holzer et al., 2003](#)).

However, not all researchers find a clear relation between accessibility and employment outcomes. [Cervero et al. \(2002\)](#), [Sanchez et al. \(2004\)](#), [Dujardin et al. \(2007\)](#), or [Aslund et al. \(2017\)](#) did not find such evidence in different contexts.

The overall conclusion that emerges out of the research outputs in this area is that transport policy can contribute to improve labor market outcomes by increasing accessibility, but it does not guarantee that just any infrastructure project will do so. Moreover, the positive effects can be expected not only from investment in private transport infrastructure but also by improving public transport provision. This last result is important on the light of the negative externalities generated by the use of car. Besides, disadvantaged workers will benefit most from public transport improvements since the effects are higher for less-skilled individuals and those with weaker labor attachment, since they are more constrained to local labor market opportunities.

See Also

Commuting, the Labor Market, and Wages; The Mono-Centric City Model and Commuting Cost; Wider Economic Impacts of Transport Investments

References

- Andersson, F., Haltiwanger, J.C., Kutzbach, M.J., Pollakowski, Winberg, D.H., 2018. Job displacement and the duration of joblessness: the role of spatial mismatch. *Rev. Econ. Stat.* 100, 203–218.
- Aslund, O., Östh, J., Zenou, Y., 2010. How important is access to jobs? Old question—improved answer. *J. Econ. Geogr.* 10, 389–422.
- Aslund, O., Blind, I., Dahlberg, M., 2017. All aboard? Commuter train access and labor market outcomes. *Reg. Sci. Urban Econ.* 67, 90–107.
- Boisjoly, G., Moreno-Monroy, A., El-Geneidy, A., 2017. Informality and accessibility to jobs by public transit: evidence from the Sao Paulo metropolitan region. *Transp. Geogr.* 64, 89–96.
- Brueckner, J., Zenou, Y., 2003. Space and unemployment: the labor-market effects of spatial mismatch. *J. Labor Econ.* 21, 242–266.
- Bunel, M., Tovar, E., 2014. Key issues in local job accessibility measurement: different models mean different results. *Urban Studies* 51, 1322–1338.
- Cervero, R., Sandoval, O., Landis, J., 2002. Transportation as a stimulus of welfare-to-work Private versus public mobility. *J. Plann. Edu. Res.* 22, 50–63.
- Coulson, E., Laing, D., Wang, P., 2001. Spatial mismatch in search equilibrium. *J. Labor Econ.* 19, 949–972.
- Dujardin, C., Selod, H., Thomas, I., 2007. Residential segregation and unemployment: the case of Brussels. *Urban Studies* 45, 89–113.
- Gobillon, L., Selod, H., 2014. Spatial mismatch, poverty, and vulnerable populations. In: Fisher, M.M., Nijkamp, P. (Eds.), *Handbook of Regional Science*, Vol. I, pp. 93–107.
- Gobillon, L., Selod, H., Zenou, Y., 2007. The mechanisms of spatial mismatch. *Urban Studies* 44, 2401–2427.
- Holzer, H., Quigley, J.M., Raphael, S., 2003. Public transit and the spatial distribution of minority employment: evidence from a natural experiment. *J. Policy Anal. Manage.* 22, 415–441.
- Houston, D.S., 2005. Methods to test the spatial mismatch hypothesis. *Econ. Geogr.* 81, 407–434.
- Ihlandfeldt, K., Sjoquist, D., 1998. The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform. *Hous. Policy Debate* 9, 849–892.
- Kain, J., 1968. Housing segregation, Negro employment, and metropolitan decentralization. *Q. J. Econ.* 82, 175–197.
- Korsu, E., Wenglenski, S., 2010. Job accessibility, residential segregation and risk of long-term unemployment in the Paris Region. *Urban Studies* 47, 2279–2324.
- Matas, A., Raymond, J.L., Roig, J.L., 2010. Job accessibility and female employment probability: the cases of Barcelona and Madrid. *Urban Studies* 47, 769–787.
- Norman, T., Börjesson, M., Anderstig, C., 2014. Labor market accessibility and unemployment. *J. Transp. Econ. Policy* 51, 47–73.
- Patacchini, E., Zenou, Y., 2005. Spatial mismatch, transport mode and search decisions in England. *J. Urban Econ.* 58, 62–90.
- Raphael, S., 1998. The spatial mismatch hypothesis and Black youth joblessness: evidence from the San Francisco Bay Area. *J. Urban Econ.* 43, 79–111.
- Rosbapé, S., Selod, H., 2006. Does city structure cause unemployment? The case study of Cape Town. Poverty and policy in post-apartheid South Africa. In: Bhorat, H., Kanbur, R. (Eds.), *Human Science Research Council*, pp. 262–287.
- Sanchez, T.W., Shen, Q., Peng, Z.R., 2004. Transit mobility, jobs access and low-income labor participation in US Metropolitan Areas. *Urban Studies* 41, 1313–1331.
- Sari, F., 2015. Public transport and labor market outcomes: analysis of the connections in the French agglomeration of Bordeaux. *Transp. Res. Part A* 78, 231–251.
- Wasmer, E., Zenou, Y., 2002. Does city structure affect search and welfare? *J. Urban Econ.* 51, 515–541.
- Weinberg, B.A., Reagan, P.A., Yankow, J.J., 2004. Do neighborhoods affect hours worked? Evidence from longitudinal data. *J. Labor Econ.* 22, 891–924.

Regulatory Reforms and Competition in Public Transport

Didier van de Velde^{*,†}, Fabio Hirschhorn[†], ^{*}inno-V Consulting, Amsterdam, The Netherlands; [†]Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	365
Why Reforming: Public Transport Performance and New Public Management	365
What Reforms: Deregulation, Liberalization, and Privatization	366
How Do Reforms Work: Typologies and Basic Policy Pathways	366
Deregulation	367
Competitive Tendering	367
Competitive Regulation	368
Reform of the Public Sector	368
Outlook: The Future of Regulatory Reforms and Competition in Public Transport	368
References	369

Introduction

Public transport—or “transit” in the North-American context—(hereafter “PT”) is an umbrella term variously defined. This chapter focuses on collective modes of land passenger transport services available to the general public on a nondiscriminatory and continuous basis, and in most cases according to predetermined routes, timetables or frequencies and fares. The analysis comprises PT services offered within the metropolitan context (frequently referred to as local and regional public transport), and intercity transportation.

The organization and provision of PT has historically been at the center of academic and policy debates, and particularly so since the 1980s, as the sector witnessed significant and continuous regulatory reforms characterized and inspired by a wave of neoliberal thinking, many times connected to the label New Public Management (here after “NPM”). These reforms aimed at improving the efficiency and effectiveness of the public sector with the introduction of managerial practices typical from the private sector. In PT, this has been mainly characterized by a growing role for the use of competition, particularly through the competitive tendering of contracts, but also via market deregulation (Van de Velde, 2019).

As a result of the changes implemented in these decades, the sector, that before—in the European context—was typically made up of de facto area monopolies with ex post subsidization, is currently growingly organized under time-limited contracts that define exclusive rights, set service requirements and include various sets of financial incentives. Furthermore, large operators, often affiliated to state-owned railway companies, are increasingly active in markets outside of their country of origin. They have appeared alongside municipally owned operators, whereas the number of small, local businesses is decreasing.

The chapter proceeds by reviewing the rationale behind reforms occurred in the last decades, describing the main institutional setup options that have appeared since then, and defining typologies and basic reference models that support the understanding of how existing institutional frameworks work in practice. To conclude, the final section proposes initial reflections on the future direction of PT regulation in view of transformations in PT triggered by the appearance of new technologies and service models such as autonomous vehicles and mobility as a service (MaaS).

Why Reforming: Public Transport Performance and New Public Management

Finding the most appropriate regulatory framework to organize and provide PT services has traditionally occupied a prominent role in academic and policy debates. Chadwick (1859), already in the 19th century, compared the potential advantages and disadvantages of employing competition for the field—namely competition for having access to a market or area where to deliver PT services—and competition within the field—related to the competition between different transport providers operating in the same market. Similar discussions, involving various other organizational elements of PT, such as the allocation of risks between authority and operators, nature and ownership of operators, or ticket and fare integration have proliferated in recent decades (Hirschhorn et al., 2019; Pucher and Kurth, 1995; Roy and Yvrande-Billon, 2007).

The backdrop for these debates is the understanding that the way in which the production and delivery of PT services are organized may influence the achievement of performance targets. That is, PT is seen as a tool to enable the realization of a number of public values—the goals society expects government to achieve in a policy area (Jørgensen and Bozeman, 2007)—and organizing the sector appropriately can support political ambitions connected to aims like cost-efficiency, sustainability, or accessibility.

In the 1970s and 1980s, the rise of neoliberal ideals giving primacy to values like efficiency and effectiveness profoundly influenced public sector services. Reforms under the New Public Management label (NPM) introduced new practices in public administration advancing horizontal specialization in the public apparatuses, with the use of contracting, and implementation of a private-sector management style with explicit performance standards and output control. At the same time, on a vertical dimension, reforms emphasized structural devolution, and the creation of specialized agencies. According to the NPM proposition, policy making is separated from policy administration and implementation, and coordination in the production and delivery of public services is to be achieved through the market's "invisible hand" (Christensen, 2012; Hood, 1991).

In particular for PT, such approach targeted problems such as productive inefficiency, the undesirable growth of public subsidies and declining market share, all highlighted by a noninnovative or bureaucratic image of the sector. These reforms are interesting in light of cross-sectoral policy objectives and the role that passenger mobility could play in supporting ambitions connected to environmental sustainability, road traffic congestion, land use, and densification policies, besides tighter public budgeting. These developments came in conjunction with reforming dynamics that may already have been present in those countries, for example, due to decentralization policies and budgetary problems. Furthermore, they were propelled by a mixture of impulses from economic theory and concrete reform experiences. Influential economic theories where in particular the theory of contestable markets (Baumol, 1982) and a renewed interest for competitive tendering (franchise bidding) as a regulatory mechanism for monopoly operations (Demsetz, 1968). Important experiences had been present in different countries and involving different modes of transport. The United States and the British long-distance coach sector were deregulated since 1980, followed by Sweden and Norway. Deregulation also spread to the local and regional bus sector outside London in 1986 with some clear influence from contestability theory. In parallel, further reform experiences based on the usage of competitive tendering grew in the bus sector in London (1984) and later in Denmark (Copenhagen in 1991), Sweden (1989), France (reform of contracting in 1981 and stricter tendering rules in 1994) and the Netherlands (2001). In the railway sector, in 1987 Japan implemented privatization reforms combined with yardstick competition and price regulation, while the United States, since 1980, opted for a far-reaching deregulation in rail freight. The British Government, rather than deregulating, introduced competitive tendering in the railway sector (known as railway "franchising") in 1994.

What Reforms: Deregulation, Liberalization, and Privatization

The regulatory reforms in PT that have appeared during the last few decades usually combine deregulation, liberalization, and/or privatization. Deregulation is characterized here by the process of reducing the number of rules to which transport operators are subject to in the market in which they operate, resulting thus in an enhanced behavioral freedom in terms of the determination of service production and characteristics. Liberalization, in turn, is defined here as the process allowing other operators, in addition to the incumbent, to get access to the market, whatever the degree of behavioral freedom allowed by regulation to operators on that market. Finally, privatization here corresponds to the process of transferring the ownership of a company or agency from the public sector (such as the national or a local government) to the private sector. It is important to note that the extent to which deregulation, liberalization, and privatization are present can substantially differ according to the reform option chosen and the preexisting organizational form and institutional configuration of the sector. Regardless of the degree of deregulation, liberalization, and privatization used, a common characteristic of most—though not all—reforms that have been implemented during the last decades is the recourse to some elements of competitive pressure or the modification of already existing competitive arrangements. The way to organize such competition, however, varied substantially across and within countries and cities.

Two main classifications help understand changes that occurred in the last decades. First, a dichotomy needs to be established to distinguish PT systems based on free market competition and systems based on competitive tendering of monopoly rights by a transport authority. Second, the main policy options for reforms within these two types of systems can be grouped into four basic models: deregulated open market, competitive tendering, competitive regulation, and public sector reform. These two classifications are detailed in the next section.

How Do Reforms Work: Typologies and Basic Policy Pathways

Two fundamentally different categories of organization of the supply of PT services can be distinguished based on market entry rules; those defining the actors that have the right to initiate PT services. In market-initiated regimes, the supply of transport services is based upon the principle of autonomous market entry. In authority initiated regimes instead, transport authorities have the legal monopoly of initiative, it being the prerogative of the authority to produce or request the production of services (Van de Velde, 1999). Regimes based on market initiative can vary from full "open entry" to more-or-less strictly regulated "authorization regimes" (the latter were often dominated by publicly owned companies, in particular after the 1960s, while being historically based on private market initiative). Regimes based on authority initiative, in turn, can vary from "public monopoly" regimes to full "private concessions" awarded by the public sector. As to the former, one can distinguish further between public management of the operations and delegated management where the assets (vehicles, garages, tunnels, etc.) are publicly owned and made available to an operator to whom the management of the services is delegated after a specific selection and awarding procedure. In principle, all of these modes of organization can make use of further subcontracting of (parts of) the operations to third operators selected, for

Table 1 Categorization of reforms

	<i>Direct competition between operators: Operators face direct competition from other operators</i>	<i>No direct competition between operators: Operators are not directly threatened by other operators</i>
<i>Regimes based on market initiative:</i> Operators are, in principle, free to take initiatives to provide services	<i>Market deregulation:</i> The possibility for direct and 'daily' competition between operators is introduced	<i>Competitive regulation of monopolistic operators:</i> Historic operators (public or private) are regulated on the basis of a mutual comparison of their performances (such as yardstick competition)
<i>Regimes based upon authority initiative:</i> A transport authority organizes transport services and/or assigns a temporary right to an operator	<i>Competitive tendering of operational rights:</i> Periodic competition between operators for temporary operational rights is introduced	<i>Public operator governance reform:</i> The governance of the publicly owned operator is reformed to instill new performance incentives

Source: Adapted from Van de Velde, D., 1999. Organisational forms and entrepreneurship in public transport. Part 1: Classifying organisational forms. *Transp. Policy* 6(3), 147–157. [https://doi.org/10.1016/S0967-070X\(99\)00016-5](https://doi.org/10.1016/S0967-070X(99)00016-5).

example, by competitive tendering. Furthermore, it should be clear that real world examples often combine different elements of these alternatives and that several regimes can coexist within one area.

An essential aspect of this distinction is that in free market regimes, whether regulated or not, in principle any entrepreneur can initiate new transport services. In regimes based on monopoly regulation this is not case, regardless if PT services are to be provided by public sector operators or through a contractual delegation of a temporary right to independent operators via competitive tendering. In authority-initiated regimes, entrepreneurial rights are concentrated in the hands of the authority and only partially in the hands of the selected operator(s), and only to the extent to which the contract gives the operator(s) some leeway in service definition.

Choices across these alternatives determine the functioning of the sector in the longer run. Changes to such core elements usually require new or revised legislation. This makes them infrequent. In other words, only significant policy reforms generate a fundamental institutional reengineering of the sector. On the other hand, decisions pertaining to the functioning and fine-tuning of regimes are less likely to be anchored in legislation and are therefore more likely to be amendable in the medium to short term. This flexibility results in the existence of very context-dependent arrangements in different PT contexts.

Reform pathways followed in PT worldwide since the 1980s can be grouped in four main reform options: market deregulation, introduction of competitive tendering, regulatory reform of monopolistic operators on the basis of indirect competition, and public operator governance reform (Van de Velde, 2016). To be sure, real world cases rarely fit perfectly in this classification, hybrid reform packages can frequently be observed for different market segments. This section discusses these four main alternatives, also depicted in Table 1, and provides some implementation examples.

Deregulation

Where true “deregulation” was introduced in PT markets, it usually included all three elements defined above: privatization of the preexisting companies owned by the authorities, liberalization of the market by allowing the entry of new operators, and a deregulation of market behavior giving the operators a larger degree of freedom in the determination of their transport services. This is the case, for instance, of the deregulation of the local bus market in Great Britain outside London implemented in 1986 and adjusted by subsequent reregulations in 2000, 2008, and 2017. Other examples include the deregulation of the long-distance coach market in Great Britain in 1980, and local and regional bus transport in New Zealand in 1989. Whilst deregulated, these systems still involve some degree of intervention by the authority. This may include safety and environmental regulation and further “rules-of-the-game” that are meant to enhance the free market outcome (such as integrated ticketing, fares coordination, or participation in integrated information systems). Further rules can add entry selection based on an evaluation of the “desirability” of the entry by a regulatory function (a civil servant or committee), as was the case in many regimes in Europe between the 1930s and the 1980s. The coach sector in Europe is another example; most countries that have allowed the long-distance coach market to develop have chosen for a regime based upon market initiative with a large degree of deregulation, mainly removing normative barriers to head-on competition on the same route and simplifying the authorization process (Beria et al., 2018).

Competitive Tendering

The introduction of competitive tendering usually involves a process of liberalization allowing new operators, both public and private, to compete in formal tendering procedures. This may also involve incumbent players. The contracts awarded in these procedures can be referred to as “service contracts,” “concessions,” or “franchises,” depending upon the legal regime in place and their characteristics in terms of risk allocation. The gist of such arrangement is that it entitles the operator to a temporary right to operate services defined in the arrangement. This right can be more or less exclusive and vary in geographical scope (a specific route

or a delimited concession area, for example). Whilst many forms of competition are allowed in Europe, the EU Regulation 1370/2007 (and later amendments) gives the framework and clearly shows a preference for the awarding of exclusive rights by competitive tendering (Van de Velde, 2008), whereas in the United States, at least so far, contracting as a whole has not become a widespread option (TransitCenter and Eno Center for Transportation, 2017).

Contracts awarded in competitive tendering procedures may vary substantially in the way they allocate service design tasks between authority and operators. In some cases, authorities employ very detailed contracts that specify tightly the services to be provided, restricting the space operators could have to propose service innovation, thus restricting their focus to the production processes and actions that increase cost-efficiency. Such contracts tend to be small in size (one or a few bus routes) and are mostly gross-cost contracts, that is, contracts where the operator carries the production cost risk but not the revenue risk, which is borne by the transport authority. This is what happens in London, Copenhagen, and Helsinki for instance. In other situations, authorities may opt to allow greater room for operators to introduce innovation in service design. Examples can be found in the most Dutch provinces and many French urban public transport contracts. In these cases, contracts tend to cover a larger territory (e.g., a town or province) and are more often net-cost contracts, i.e. the operator carries both the production and the revenue risk. It is still the authority, however, who decides which geographic area and broad markets will be serviced, as well as the core characteristics of the services to be provided (all specified in the tendering documents).

The design of gross-cost and net-cost contracts and the associated allocation of risks and responsibilities across stakeholders has become increasingly nuanced and sophisticated over time. Public private partnerships and joint-venture agreements, for example, have been gaining traction in the sector as schemes not only to offer PT services, but also to design, finance, and implement the necessary infrastructure (Pulido and Hirschhorn, 2015).

Competitive Regulation

Competitive regulation refers to situations in competition is used in the regulatory context rather than “in” or “for” the market. With yardstick competition (Shleifer, 1985) for instance, payments from the authority to operators are based on a benchmark estimated considering the cost performance of a larger set of companies. In such arrangements, regulators regulate several services in contiguous spatial markets, comparing or “benchmarking” multiple firms as a way to overcome the informational disadvantage of the regulator. Thus, an important challenge in using yardstick competition is the need of comparable data, which is scarce in the transport sector. This may be one of the reasons for the rare use of yardstick competition compared to other options. Furthermore, even when data is available, another difficulty in using yardstick competition is that regulators should always take into account variation of costs across companies that are due to legitimate differences among them (in PT this could be different route types, congestion levels, peak demand characteristics, and other exogenous influences on costs) (Estache and Gómez-Lobo, 2005). Yardstick competition, when introduced, is likely to be a light deregulation or reregulation of the sector, perhaps combined with privatization but not necessarily involving liberalization other than through the possibilities created by the privatization. Since 1987, the Japanese railway sector is submitted to, among other regulatory regimes, yardstick competition. The aim behind the system is to push the less efficient rail companies in the yardstick group to improve their efficiency (Mizutani et al., 2009).

One further indirect form of competitive regulation can take place via the “threat,” by authorities, to introduce competitive tendering if incumbent operators do not perform according to—or sufficiently evolve towards—set benchmarks. Such target level could itself be set by the measurement of the outcome of a more competitive form of regulation in other areas, typically competitive tendering. This option has, for example, been used in a few cases in the Netherlands and Switzerland.

Reform of the Public Sector

Reforming the governance of a public operator is another widespread reform option. Reregulation and light deregulation are perhaps better descriptions of this pathway, as it often involves contractual, more functional and less detailed operational guidance to operators. There is no direct competitive threat in this option, but the threat of shifting to competitive tendering can constitute a power-play element during contractual negotiations—such was the case in Amsterdam. Some benchmarking elements can be used too, to determine broad reform targets in terms of demanded efficiency gains. Liberalization, however, is not part of this reform as it is based on keeping the public monopoly in place. Neither is privatization although the transformation of a branch of a public administration into a company structure (corporatization) is a common element of such reform.

Outlook: The Future of Regulatory Reforms and Competition in Public Transport

There is no simple answer to the question of whether reforms in public administration since the 1980s have been a success. The whole body of ideals that constitute the NPM has been severely questioned, and there is a generalized perception that these reforms led to fragmentation in policy design and implementation, resulting in the delivery of lower quality public services. As such, many policy areas and administrative structures have seen a change in emphasis away from structural devolution, disaggregation, and single-purpose organizations and toward a “joined-up government” or “whole-of-government” approach, that sought to apply a more holistic strategy in government (Christensen, 2012). In the PT sector in particular, results of policy reforms are indeed mixed. In some cases, they produced a number of positive effects, such as lowered costs per bus kilometer, increased bus kilometers, and

eventually some reduction in public subsidies. However, other analyses indicate that PT also suffered with increased fare levels or lower quality of services and decline in passenger levels (e.g., Preston and Almutairi, 2013; Veeneman, 2016). All in all, it remains difficult to disentangle the effects of the reforms per se, of public budget reductions that may have occurred, and of other local economic effects. Nonetheless, some agreement exists that in view of the fragmentation in policy design and implementation, greater policy integration (both across PT policies and between PT and other areas such as land use planning), ticketing and fare integration, as well as physical and modal integration in PT delivery, is essential to remediate the negative effects of neoliberal policies (Hirschhorn et al., 2018; Hull, 2005).

The upshot is that there is probably no right or wrong approach in the regulation of PT. The use of competition or not, as well as the way to do it, are characteristics that, on their own, are insufficient to ensure that the sector performs successfully. Developments over the last decades suggest that the option between the alternative policy pathways described above increasingly becomes an option of how to develop a hybrid model, smartly combining those elements of different models that are most suitable for local and regional contexts, considering the institutional history, the skillset of authorities and civil servants, and prevailing public values.

Nevertheless, this scenario of gradual evolution and relative dynamic stability in the institutional environment of PT seems to be on the verge of a new period of profound changes. Many authors observe macro behavioral trends and technological developments that promise to disrupt the existing institutional framework of the PT sector: broad concerns with the climate crisis, younger generations less interested in owning cars, increased opportunities for tele-working, the uncertain promise of autonomous vehicles, and smartphone mobility apps are just a few examples of factors transforming how the experience of travelling and even the need for it is seen, consequently affecting mainstream PT regulation (Buehler, 2018; Mulley et al., 2018). A common characteristic of innovations in PT technologies and services is that they result from a multitude of commercial initiatives that increasingly leaves PT institutional frameworks based on authority-initiative as a shrinking island in a growing sea of market initiative. This suggests that free market initiative may play an even more revolutionary role in the future of PT than competitive tendering has played over the past decades (Van de Velde, 2019).

The next challenge for the future of regulatory reforms and competition in PT lies on the acknowledgment by policymakers and planning authorities of the growing need to develop policy tools, strategies, and human skills to govern the nascent transition in the mobility sector (Hirschhorn et al., 2019). This will require public sector actors to lead and manage networks including a potentially growing number of public and private players interacting in an increasingly complex sector, made up of interdependent elements of varied nature, such as technology, infrastructure, and values. This may demand a more flexible approach to how PT is governed, potentially requiring new governance strategies beyond traditional forms of command and control, combining these existing practices with new ones that seek collaboration with a more diverse set of actors, with various backgrounds and new and competing ideas (Hirschhorn et al., 2019). This also means that the development of a new skillset by policymakers and authorities to cleverly using existing regulatory instruments is perhaps more at the core of the evolution of PT governance than new managerial models, increased funding and cost cuts.

References

- Baumol, W.J., 1982. Contestable markets: an uprising in the theory of industry structure. *Am. Econ. Rev.* 72 (1), 1–15.
- Beria, P., Nistri, D., Laurino, A., 2018. Intercity coach liberalisation in Italy: fares determinants in an evolving market. *Res. Transp. Econ.* 69, 260–269.
- Buehler, R., 2018. Can public transportation compete with automated and connected cars? *J. Public Transp.* 21 (1), 7–18.
- Chadwick, E., 1859. Results of different principles of legislation and administration in Europe; of competition for the field, as compared with competition within the field, of service. *J. Stat. Soc. Lond.* 22 (3), 381–420.
- Christensen, T., 2012. Post-NPM and changing public governance. *Meiji J. Polit. Sci. Econ.* 1.
- Demsetz, H., 1968. Why Regulate Utilities. *J. Law Econ.* 11 (1), 55–65.
- Estache, A., Gómez-Lobo, A., 2005. Limits to competition in urban bus services in developing countries. *Transp. Rev.* 25 (2), 139–158.
- Hirschhorn, F., Paulsson, A., Sørensen, C.H., Veeneman, W., 2019. Public transport regimes and mobility as a service: governance approaches in Amsterdam, Birmingham, and Helsinki. *Transp. Res. A* 130, 178–191.
- Hirschhorn, F., Veeneman, W., Van de Velde, D., 2018. Inventory and rating of performance indicators and organisational features in metropolitan public transport: a worldwide Delphi survey. *Res. Transp. Econ.* 69, 144–156, doi:10.1016/j.retrec.2018.02.003.
- Hirschhorn, F., Veeneman, W., Van de Velde, D., 2019. Organisation and performance of public transport: a systematic cross-case comparison of metropolitan areas in Europe, Australia, and Canada. *Transp. Res. A* 124, 419–432, doi:10.1016/j.tra.2019.04.008.
- Hood, C., 1991. A public management for all seasons? *Public Admin.* 69, 3–19.
- Hull, A., 2005. Integrated transport planning in the UK: from concept to reality. *J. Transp. Geogr.* 13 (4), 318–328.
- Jørgensen, T.B., Bozeman, B., 2007. Public values an inventory. *Admin. Soc.* 39 (3), 354–381.
- Mizutani, F., Kozumi, H., Matsushima, N., 2009. Does yardstick regulation really work? Empirical evidence from Japan's rail industry. *J. Regul. Econ.* 36, 308–323.
- Mulley, C., Nelson, J.D., Wright, S., 2018. Community transport meets mobility as a service: on the road to a new flexible future. *Res. Transp. Econ.* 583–591.
- Preston, J., Almutairi, T., 2013. Evaluating the long term impacts of transport policy: an initial assessment of bus deregulation. *Res. Transp. Econ.* 39, 208–214.
- Pucher, J., Kurth, S., 1995. Verkehrsverbund: the success of regional public transport in Germany, Austria and Switzerland. *Transp. Policy* 2 (4), 279–291.
- Pulido, Daniel; Hirschhorn, Fabio. 2015. *Private participation in urban rail : a resurgence of public-private partnerships*. Transport and ICT connections note; no. 11. Washington, D.C.: World Bank Group. <http://documents.worldbank.org/curated/en/612961467991957672/Private-participation-in-urban-rail-a-resurgence-of-public-private-partnerships>
- Roy, W., Yvrande-Billon, A., 2007. Ownership, contractual practices and technical efficiency: The case of urban public transport in France. *J. Transp. Econ. Policy* 41 (2), 257–282.
- Shleifer, A., 1985. A theory of yardstick competition. *RAND J. Econ.* 16 (3), 319–327.
- TransitCenter and Eno Center for Transportation, 2017. *A Bid for Better Transit: Improving Service with Contracted Operations*.
- Van de Velde, D., 2008. A new regulation for the European public transport. *Res. Transp. Econ.* 22 (1), doi:10.1016/j.retrec.2008.05.011.

- Van de Velde, D., 1999. Organisational forms and entrepreneurship in public transport. Part 1: Classifying organisational forms. *Transp. Policy* 6 (3), 147–157, doi:10.1016/S0967-070X(99)00016-5.
- Van de Velde, D., 2016. Local public transport. In: Finger, M., Jaag, C. (Eds.), *The Routledge Companion to Network Industries*. Routledge, Oxford, pp. 241–253.
- Van de Velde, D. (2019). *Competition in Public Transport: An exploratory research in institutional frameworks in the public transport sector*. PhD Thesis, Delft University of Technology.
- Veeneman, W., 2016. Public transport governance in the Netherlands: More recent developments. *Res. Transp. Econ.* 59, 116–122.

How to Finance Transport Infrastructure?

Tiziana D'Alfonso, Giuseppe Catalano, Department of Computer, Control, and Management Engineering Antonio Ruberti, Sapienza University of Rome, Rome, Italy

© 2021 Elsevier Ltd. All rights reserved.

Public Intervention in the Financing of Transport Infrastructures	371
Externalities	372
Natural Monopoly and Infrastructure Financing with Economies and Diseconomies of Scale: Marginal Cost Versus Average Cost	
Coverage	372
Public Finance Versus Private Finance	373
Forms of Infrastructure Financing and the Public-Private Partnerships	374
The Payment Structure	375
Public Ownership Versus Private Ownership	377
References	377

Public Intervention in the Financing of Transport Infrastructures

Transport infrastructures have been fallaciously viewed as being public goods, that is nonrival and nonexcludable. With a nonrival good, one person's consumption of the service (e.g., driving miles on the highway) does not reduce the quantity that can be consumed by any other person (e.g., all other drivers can still drive as far as they want on the highway). This is the case with transportation infrastructure when congestion does not occur. A service is nonexcludable if, once produced, is accessible to all consumers; no one can be excluded from consuming such a service after it is produced (e.g., any driver can drive on the highway). However, this is not necessarily the case with transport infrastructures—typical exclusion mechanisms include toll roads or congestion pricing. For these reasons, transport infrastructures may be better conceived as club goods at least until reaching a point where congestion occurs (sometimes also called *artificially scarce goods*).

While nonexcludability and nonrivalry of pure public goods call for public sector intervention in the financing—in order to prevent the relevant form of good from being under-supplied^a—excludability and nonrivalry of goods call for public sector intervention in the regulation of access pricing, as financing of the good relies on revenues from use. In this context, why we do observe public intervention in the financing of transport infrastructure is mainly based on three arguments:

Efficiency: when the market includes failures (externalities, natural monopoly and imperfection of capital markets), the market price may not reflect the social value of the service, and the market may therefore not maximize total surplus—that is, the equilibrium may be economically inefficient.

Equity: transport infrastructure can be classified as services of general economic interest (SGEI), which are economic activities that public authorities identify as being of particular importance to citizens and that would not be supplied (or would be supplied under different conditions) in some circumstances, if there were no public intervention, for example, the case of *weak-demand areas*.

Stabilization policy: stimulus packages often highlight the important role of infrastructure investment in boosting the economy. Public infrastructure investment tends to enhance the productivity of the private sector and is thus likely to promote economic prosperity in normal times, while often offsetting falling private demand and stimulating the economy during recessions.

The implementation of an optimal public intervention can be limited by asymmetric information or incomplete contracting problems. Two problems may inhibit efficient contracting. The first is *adverse selection*, which in the context of procurement means that the contractor (the concessionaire or the agent) has better information than the owner (the contracting public sector, the principal) about the cost to complete the project. The second is *moral hazard*, which means that the contractor can take unobserved actions, such as shirking on quality, that may result in an inferior service; and the contracting public sector bears the relating risk.^b

^aTo finance an efficient level of output for a public good consumers must jointly agree that everyone contributes an amount equal to his own willingness to pay. However, since the provision of a public good is nonexclusive, everyone benefits once the public service is provided. Consequently, individuals have no incentive to pay as much as the service is really worth to them. They can behave as a free rider, paying nothing for a service while anticipating that others will contribute, which can preclude production by the private sector as no consumer would pay for it.

^bWith the standard adverse selection and moral hazard components, the principal should offer a menu of incentive contracts from which the project firm (agent) will select a contract that reveals its hidden information. In reality, such menus are not observed.

Externalities

An externality arises when the actions of any decision maker affect the benefits of other consumers or production costs of other firms in the market in ways other than through changes in prices. An externality that reduces the well-being of others is a negative externality. An externality that brings benefit to others is a positive externality.

The development of transport infrastructures may generate positive externalities on the economy (stimulation of the economic growth, increase in productivity and profitability) but can also impose internal and external costs. Internal costs (i.e., the price of driving) include gas and oil, wear and tear on the car, and any tolls, as well as the cost of time spent driving (the driver could have spent that time doing something productive). These are the costs a driver is likely to take into account when deciding whether to drive, but there are other costs (external costs, i.e., negative externalities) that she is much less likely to consider because she does not bear them herself—for instance, adding to traffic congestion and thereby increasing the travel time (and associated cost) for other drivers. The costs that a driver imposes on society include both these kinds of costs.

Other examples of negative externalities include environmental effects, noise annoyance, and accidents.

When externalities arise, the project's social returns can exceed its private returns and the classic efficiency argument for optimal regulation applies, closing the gap between private and social benefits and preventing the relevant form of infrastructure from being under-supplied.

Natural Monopoly and Infrastructure Financing with Economies and Diseconomies of Scale: Marginal Cost Versus Average Cost Coverage

A market is a natural monopoly if, for any relevant level of industry output, the total cost a single firm producing that output would incur is less than the combined total cost that two or more firms would incur, if they divided that output among them. That is, the cost function of a firm in the industry is *subadditive*.^c

Fig. 1 shows a natural monopoly market. The inverse market demand curve is D , and each firm has access to a technology that generates a long-run^d average cost curve $AC(q)$. For any output less than q_s units per year, a single firm can produce output more cheaply than two or more firms could (i.e., assume that the cost function of a firm is subadditive up to q_s). Note that, in Fig. 1, some levels of output along the demand curve can be produced more cheaply by two firms than one (e.g., $q = \tilde{q}$). However, such output levels would be demanded only at prices less than the minimum level of average cost. Thus, they would not be profitable. At all relevant levels of market demand, that is, all levels of market demand that could be profitably produced—the total cost of

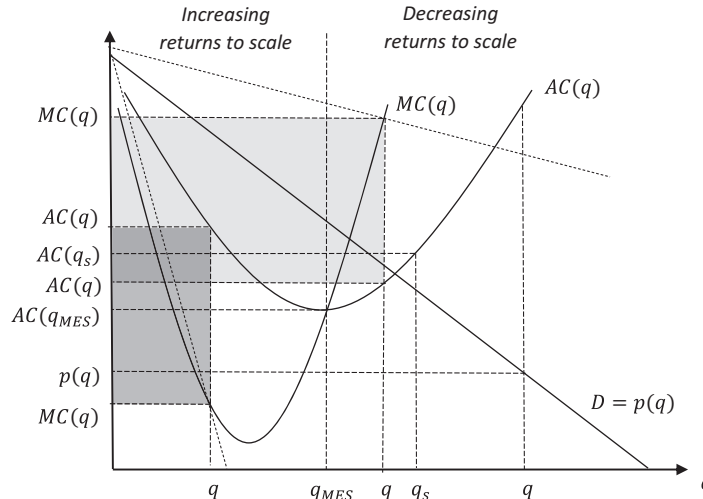


Figure 1 Marginal vs average cost coverage under economies and diseconomies of scale.

^c Following alternative definition in literature, an industry has been called a natural monopoly, if a single firm producer can find price-output combination that preclude profitable entry by others (i.e., *sustainable price-output combination*). The existence of fixed costs of sufficient magnitude generally guarantee the existence of sustainable price-output combination for a monopolist, that is, making him invulnerable against entry (Baumol and Willig, 1981)

^d Consider "the long run" to refer to a period of time sufficient for all current commitments to be liquidated. Fixed costs are those costs that are not reduced, even in the long run, by decreases in output as long as production is not discontinued altogether. But they can be eliminated in the long run by total cessation of production. Sunk costs are those costs that (in some short or intermediate run) cannot be eliminated, even by total cessation of production, that is, in the long run all sunk costs are zero. As such, once committed, sunk costs are no longer a portion of the opportunity cost of production. Fixed costs do not necessarily constitute barriers to entry; that is, they need not have undesirable welfare consequences. Sunk costs do, however, constitute barriers to entry, where an entry barrier is anything that requires an expenditure by a new entrant into an industry, but that imposes no equivalent cost upon an incumbent, and as such does reduce the sum of consumers' and producers' surplus. It follows that phenomena such as fixed costs and scale economies need not do so.

production is minimized when one firm serves the entire market. Then, whether a market is a natural monopoly depends not only on technological conditions (the shape of the AC curve) but also on demand conditions. A market might be a natural monopoly when demand is low but not when demand is high.

A sufficient condition for cost subadditivity is the existence of fixed costs of sufficient magnitude. Then, a sufficient condition for natural monopoly is that the average cost curve must decrease with output over some range. That is, natural monopoly markets might involve economies of scale.

Transport infrastructures can be interpreted as a limiting case of a fixed cost relationship. Nonrivalry implies the *supply-jointness* or *nondepletable* condition, that is, at least over a substantial range, once the good is supplied to anyone, the cost of making its benefits available to additional people is relatively small (or zero). This is because transport infrastructures are typically a class of goods whose production has a large fixed cost component and for which there is a comparatively low marginal cost (in the sense of the cost of serving another customer).

If the bulk of the total cost is fixed, the sale of the good at a price equal to its marginal cost will incur a loss. Suppose $\hat{q}$ is produced and sold within the market, with technology $C(q)$ involving economies of scale, $MC(\hat{q}) < AC(\hat{q})$, that is, $\hat{q} < q_{MES}$. The Pareto outcome ensuring allocative efficiency, that is, marginal cost pricing $p(\hat{q}) = MC(\hat{q})$, would lead the monopolist to incur a loss equal to $\pi = p(\hat{q})\hat{q} - C(\hat{q}) = MC(\hat{q})\hat{q} - AC(\hat{q})\hat{q} = \hat{q}(MC(\hat{q}) - AC(\hat{q})) < 0$ (dark-shaded area in Fig. 1). In this case, first best intervention implies subsidization of the public good production and how large must the subsidy be to lead the market to produce the efficient level of output depends on the magnitude of the loss incurred by the provider of the infrastructure. Note that the Pareto outcome is efficient in a broader general equilibrium sense only, if we ignore the costs the government incurs to raise the funds to pay subsidies. General implementation problems include limited social and political acceptability of subsidies.

In order for the transport infrastructure provider with increasing returns to break-even it appears that the prices the monopolist charges for the services it provides will have to exceed marginal cost. One way to proceed in the single product context is simply to set a single price for each unit of the product equal to its average cost ($p(\hat{q}) = AC(\hat{q})$).^c This equilibrium price, which is equal to average total cost at the quantity that clears the market, is the second-best efficient uniform price when the firm is subject to a break-even constraint. Suppose $\hat{q}$ is produced and sold within the market, with technology $C(q)$ being subadditive $\hat{q} < q_s$ and involving diseconomies of scale, $MC(\hat{q}) > AC(\hat{q})$, that is, $\hat{q} > q_{MES}$. The Pareto outcome ensuring allocative efficiency, that is, marginal cost pricing $p(\hat{q}) = MC(\hat{q})$, would lead the monopolist to incur a surplus equal to $\pi = p(\hat{q})\hat{q} - C(\hat{q}) = MC(\hat{q})\hat{q} - AC(\hat{q})\hat{q} = \hat{q}(MC(\hat{q}) - AC(\hat{q})) > 0$ (lighted-shaded area in Fig. 1).

Then, the marginal cost-pricing rule implies a deficit for the government in case of increasing returns to scale and a surplus in case of increasing returns to scale.

Public Finance Versus Private Finance

Infrastructure financing may rely on public finance, where the investment is financed by the contracting public sector (with an optimal mix of taxation or debt)^f, or private finance (be it through outside equity or debt), where the operator has full control over his access to the financial market on top of control over operations.

The viability of private finance depends on the availability of public capital for infrastructure, that is, the countries budget for infrastructure, as well as on its efficiency in comparison with the economics of financing the project with public funds.

Relying on private finance has two conflicting effects. On the one hand, it adds a new level of contracting, between the contractor and credit institutions, which can exacerbate moral hazard. On the other hand, private finance would imply two advantages:

- private capital ensures higher value for money and could bring the experience of external lenders into risk assessment. In other words, there is a willingness to pay for the infrastructure, which reduces the risk of an inefficient resource allocation. When private investors risk their own money, they will require decision making as being realistic, and do not rely on data from authorities or lobby organizations, generally too optimistic (Flyvbjerg, 2009);
- if risk allocation is right the private sector can build and operate infrastructure more efficiently.

In this regard, having full control of access to the financial market and operations could improve the more traditional way of procurement where the cost of investments is paid through taxation and investments are not supported by such a level of competence inside the public contracting sector.

The choice between public or private finance is resolved in favor of private finance when external lenders are competitive, so that the case is similar to a situation where the government itself was enjoying the external lenders' expertise to provide funds at a cheaper cost. When external lenders are specialized in infrastructure risk, they might have some market power. In such an environment, it is unlikely that the government can earn all benefits from the external lenders' expertise. There would be a trade-off between the

^c In the multiproduct case, the reader may refer to the Ramsey-Boiteux pricing problem.

^f For transport infrastructures, present expenditures will provide future benefits. When the initial outlay is large, taxpayers may not wish to assume the entire cost at once and may prefer to pay over the years as the services of the new facility are enjoyed. Musgrave's option of *pay-as-you-use* finance would apply. Using current year tax money is sometimes described as "*pay-as-you-go*" in contrast to using debt instruments called "*pay-as-you-use*" financing. When using bond financing, project beneficiaries will pay for its use. The alternative *pay-as-you-go* financing method reduces the likelihood that larger, expensive projects can be built within acceptable time periods and means that future beneficiaries will not have to pay for its construction (Levinson, 2001; 2005).

benefits of the lenders' expertise and the extra distortions that financial contracts might bring. Recourse to private finance can however improve the operator's incentives, if lenders bring their expertise in monitoring effort.

In practice, the use of private finance has also been made because it has allowed the public sector to finance the construction of infrastructure "off the balance sheet" and to accelerate delivery of projects.

Forms of Infrastructure Financing and the Public-Private Partnerships

The relationship between public authorities and private investors to build, operate, maintain, and finance transport infrastructures for public services is regulated by long-term contracts which may take the form of concession agreements, lease contracts, management contracts, and more generally public-private partnerships (PPPs, 3P or P3).

The structure used for any particular project mainly depends on:

- The functions (design, construction, financing, operation, and/or maintenance) the private sector party will perform and the risks it will assume. Common risks involved are:
 - Political risk, associated to payments received by the private sector from the contracting public sector.
 - Technical risk, associated to assets breakdown and construction.
 - Financing risk, such as foreign exchange rate risk and interest rate fluctuation, operational risk (e.g., change in the price of raw materials), demand risk (e.g., over-optimistic demand forecasts), cost-overrun risk.
- The degree of operational control the public agency wants to have over any aspect of the project.
- Whether the private sector party will own the project assets at any time during the term of the agreement. **Fig. 2** shows the extent of private sector participation depending on contract types and sources of financing.

PPP can be used to raise private money—or to increase efficiency in the building process. **Table 1** describes the scope of different PPPs models.

In the BOT (Build-Operate-Transfer) framework, the contracting public sector allows a private investor to design, build, finance and operate infrastructure assets over an extended period, that is, "the concession period" (generally up to 30–50 years).

The contracting public sector will monitor the realization of the infrastructure and specify public services' obligations, that is, public service constraints (pricing structure, safety, and environmental regulations). The private investor is left with control rights and responsibility over how to deliver the service while meeting the prespecified standards. So design, construction, and operational risk are generally substantially transferred to the private-sector party. In return, the private-sector party is paid by the public agency or from fees collected from the project's end users in order to recover all of his costs and a reasonable profit to pay the service of the debt and to provide dividends to the shareholders of the project company. Ownership of the infrastructure rests with the private company until the end of the concession. Then, all operating rights and maintenance responsibilities revert to the public sector, without any remuneration of the private entity involved.

Typically, the concessionaire creates a special purpose vehicle (SPV) entity which is capitalized through the financial contributions of the project sponsors. Because the SPV entity has only limited workforce, it generally subcontracts a third party to perform its obligations under the concession agreement.

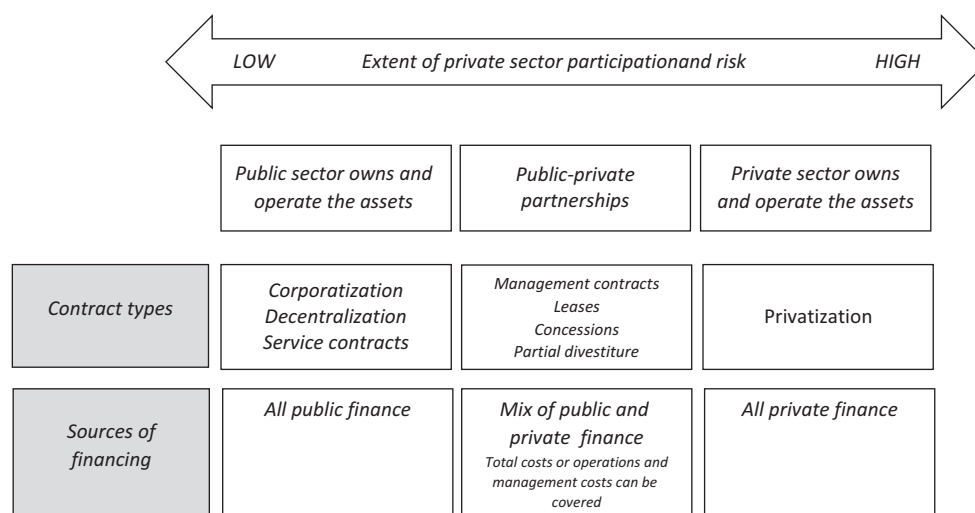


Figure 2 The relationship between public authorities and private investors.

Table 1 PPP models: functions the private sector party will perform and asset ownership

Model 1

Build-own-operate (BOO), Build-develop-operate (BDO), Design-construct-manage-finance (DCMF), Design-build-finance-operate (DBFO)

The private sector designs, develops, operates and manages an assets with no obligation to transfer ownership to the contracting public sector at the end of the operating contract.

Model 2

Buy-build-operate (BBO), Lease-develop-operate (LDO), Wrap-around addition (WAA), Build-rent-own-transfer (BROT)

The private sector buys or leases an existing asset from the contracting public sector, invests in the asset and operates the assets, with no obligation to transfer ownership to the contracting public sector at the end of the operating contract.

Model 3

Build-operate-transfer (BOT), Build-own-operate-transfer (BOOT), Build-lease-operate-transfer (BLOT), Build-transfer-operate (BTO)

The private sector designs and build an assets, operates it and then transfer ownership to the contracting public sector at the end of the operating contract (or at some other pre-determined date). The private investor may subsequently rent or lease the asset from the contracting public sector.

The financing is based on the expected cash flows of the project and it is typically on a “nonrecourse” basis, that is, the lenders will have no financial recourse for repayment of their loans against the contracting public-sector entity. Recourse is limited to the project company, that is, its assets and contractual rights.

A BOO (Build-Operate-Own) framework is a variant of the BOT and the difference is that the ownership of the newly built facility will rest with the private party.

The DBFO (Design-Build-Finance-Operate) framework is very similar to BOT, with the exception that there is not actual transfer of the ownership at the end of the entire concession period.

Fig. 3 describes a the typical relations among contracting public sector, the SPV and capital providers.

In the LDO (Lease-Develop-Operate) framework, the public-sector entity retains ownership of the newly created infrastructure facility and receives payments in terms of a lease agreement with the private sector. The lessee finance development and oversees operation. Payments received by the private investor are used in order to recover all of his costs and a reasonable profit to pay the service of the debt and to provide dividends to the shareholders of the project company. This approach is mostly followed in the development of airport facilities.

Literature finds that PPPs deliver efficiency gains when a whole-life cost approach to the project yields significant cost savings and when risk is effectively transferred to the private operator (Engel et al., 2013). Transferring design, construction, and operating risks to the contractor provides incentives for keeping project costs down and efficiently providing the service. PPPs are more beneficial when:

- a better quality of the infrastructure can significantly reduces costs (including maintenance cost);
- infrastructure quality has a great impact on the quality of the service and service demand (and service quality is verifiable);
- demand for the service is stable and easy to forecast (demand risk is low or the firm can diversify risk).

This points to the suitability of PPPs in the transport sectors, where infrastructure quality is key and demand is relatively stable.

The Payment Structure

The payments the private-sector party receives for performing its obligations under the agreement may be structured several ways. Consider the case where a contractor is in charge of both building and operation. The contractor bears costs C (excluding quality-enhancing and operating efforts) and receives a transfer t from the contracting public sector. For services where users pay, the contractor also receives revenues R . Thus, the contractor's payoff is $\pi + t$, where profits π are equal to $R - C$, if users pay, and to $-C$ if users do not pay. The following scenarios could be generally observed (Iossa and Martimort, 2015):

- *Contractible profits.* The contractor's profit $\pi = R - C$ is observable and can be contracted upon. The firm receives t from the contracting public sector, where $t = \alpha - (1 - \beta)\pi$. Under this profit-sharing rule, the contractor obtains a net profit of: $\pi + t = \alpha + \beta\pi$. In the case $\beta = 0$ the contractor receives a fixed payment and has no particular incentives to raise profits. Instead, $\beta > 0$ holds when the contractor bears profit risk. In the extreme case where $\beta = 1$, all risks are transferred to the contractor. This scheme is typically used for transport projects where users pay for the service, for instance toll roads.
- *Contractible revenues.* The contractor's revenue R is observable and can be contracted upon. The firm receives t from the contracting public sector, where $t = \alpha - (1 - \beta)R$. Under this revenue-sharing rule, the contractor obtains a net profit of: $\pi + t = \alpha + \beta R - C$.
- *User fee PPPs.* Revenue-sharing schemes are often used also in transport projects. A payment mechanism solely based on user charges corresponds to $\alpha = 0$ and $\beta = 1$ so that $t = 0$ and $\pi + t = R - C$. The contractor keeps all revenues and bears all demand risk. To mitigate these risks, private sector parties may require the public agency to guarantee a certain level of use and make payments to the private sector party if such minimum amount is not achieved, or to include in the PPP agreement extension or renewal provisions to allow the private sector party sufficient time to recoup its investment and earn a return.

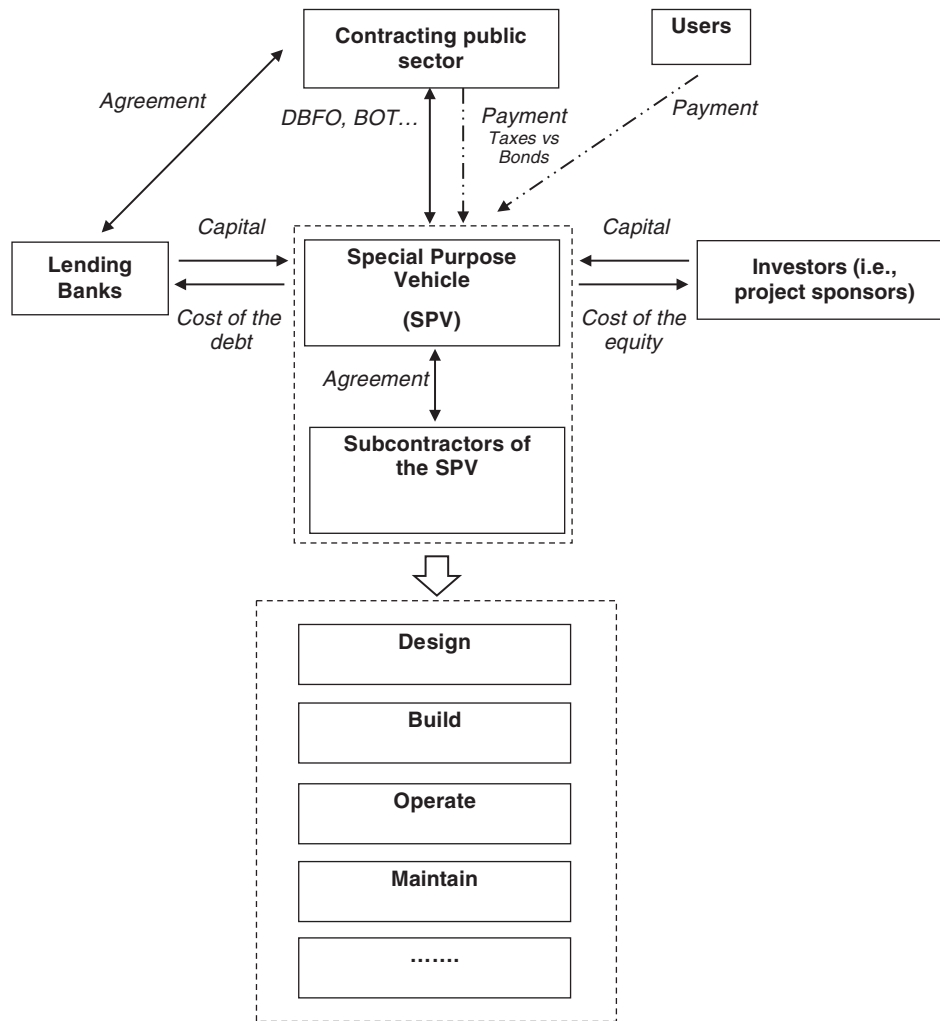


Figure 3 The structure of relations among contracting public sector, the SPV, and capital providers in PPPs.

- *Availability-based PPPs.* A payment mechanism based on availability only, corresponds instead to $\alpha > 0$ and $\beta = 0$, so that $t = \alpha$ and with $\pi + t = \alpha + R - C$. The contractor's reward is fixed and paid by the contracting public sector for making the service available to users. The public agency bears the demand risk under this structure because the amount it pays to the private sector party does not change even if the project is not used to the extent anticipated. On the other hand, the private-sector party should consider obtaining political risk insurance.
- *Shadow toll-based PPPs.* Typically used for transportation projects, shadow tolls are per vehicle amounts paid to the private sector party by the public agency and not the users of the facility ($\pi = -C$). For instance, this is used when it may not be feasible or not desired for the road to include toll facilities. It corresponds to the case where $\alpha = 0$ and $\beta = 2\gamma + 1$ so that $t = \gamma R$ and $\pi + t = \gamma R - C$, $\gamma > 0$, with γR being the total amount paid to the private sector party by the public agency. The more the road is used (i.e., higher γ), the higher the payments the public agency is required to pay to the private sector party. These payments can, therefore, impose a significant burden on the public agency's finances. To minimize this pressure, many agreements include a cap on the amount the public agency is required to pay. In this structure, the private sector party and the public sector share the demand risk.
- *Contractible costs.* The contractor's cost (excluding quality-enhancing and operating efforts) C is observable and can be contracted upon. The firm receives t from the contracting public sector, where $t = \alpha + (1 - \beta)C$. Under this cost-sharing rule, the contractor obtains a net profit of: $\pi + t = \alpha + R - \beta C$.
 - *Cost-plus contracts.* The case $\beta = 0$ corresponds to a cost-plus contract where the contractor is fully reimbursed for its own costs, that is, $\pi + t = \alpha + R$
 - *Fixed-price contracts.* The case $\beta = 1$ holds for a fixed price contract, where the contractor receives a fixed payment $t = \alpha$.

Public Ownership Versus Private Ownership

Some PPP models combine the design and operation phases, together with the private ownership of the activities for the entire duration of the contract. Traditional contracting, on the other hand, corresponds to the case in which the public contracting sector purchases an asset built by the private sector and delegates the operations to the private sector. Ownership is important to the extent that assets have a residual value at the end of the contract.

Ownership gives the owner the right to the market value of these assets and this residual value provides incentives to invest in the quality of the assets, depending on the specificity of the assets.

Consistent with the literature on incomplete contracts, the residual value of the assets cannot be specified *ex ante* in a contract, although it can be observed *ex post* and can be traded on that date. Literature finds that the hold up problem is less severe under the private ownership of the assets for the entire duration of the contract, compared to public ownership, if there is a positive externality between the building and the management phases and vice versa when the externality it is negative.⁸

Generally speaking, private ownership is optimal for infrastructures with a high-residual value.

References

- Baumol, W.J., Willig, R.D., 1981. Fixed costs, sunk costs, entry barriers, and sustainability of monopoly. *Quart. J. Econ.* 96 (3), 405–431.
- Engel, E., Fischer, R., Galetovic, A., 2013. The basic public finance of public–private partnerships. *J. Eur. Econ. Asso.* 11 (1), 83–111.
- Flyvbjerg, B., 2009. Survival of the unfittest: why the worst infrastructure gets built—and what we can do about it. *Oxford Rev. Econ. Policy* 25 (3), 344–367.
- Iossa, E., Martimort, D., 2015. The simple microeconomics of public-private partnerships. *J. Public Econ. Theory* 17 (1), 4–48.
- Levinson, D., 2001. Financing infrastructure over time. *J. Urban Plan. Develop.* 127 (4), 146–157.
- Levinson, D., 2005. Paying for the fixed costs of roads. *J. Transp. Econ. Policy* 279–294.

⁸ Positive externality (resp. negative externality) refers here to the case where a building innovation reduces (resp. increases) costs at the management stage.

Congestion, Allocation, and Competition on the Railway Tracks

John Armstrong, John Preston, University of Southampton, Southampton, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	378
Congestion	378
Capacity Allocation	382
Competition	383
Concluding Comments	384
References	384

Introduction

A conventional definition of economics is the allocation of scarce resources according to the laws of supply and demand. Where a rail system is congested, capacity becomes a scarce resource. Allocation of this capacity according to the laws of supply and demand becomes a key policy issue, which in turn determines the amount of on-track competition that is permitted. We therefore consider the key terms of congestion, capacity allocation and competition in turn, before drawing some conclusions.

Congestion

A formal definition of congestion is elusive, with most dictionaries describing the condition as a state of overcrowding or blockage. According to [The Institution of Civil Engineers \(1989\)](#), *congestion is a condition that arises because more people wish to travel at a given time than the transportation system can accommodate: a simple case of demand exceeding supply*. Moving from people to vehicles, [Vuchic and Kikuchi \(1994\)](#) note that,

when vehicular volume on a transportation facility . . . exceeds the capacity of that facility, the result is a state of congestion.

These definitions apply to all transport infrastructure and modes, for which, as demand on a network link or at a node approaches or exceeds the available capacity, congestion occurs and delays tend to increase exponentially with demand. Theoretical and empirical examples of this relationship are shown below. The empirical data are based upon recorded levels of capacity utilization (i.e., the percentage of maximum theoretical capacity that is used) and congestion-related reactionary delay (CRRD) at a critical switch (i.e., set of points) at Peterborough railway station, on the East Coast Main Line between London and York in the UK. Here, reactionary, or secondary, delay is “knock-on” delay, caused by delays to other train services ([Figs. 1 and 2](#)).

In principle, congestion should not occur on self-contained, scheduled transport systems like railways (and some commentators are critical of the use of the term congestion as a cause of delays to services as a consequence of this). In other words, using the terminology of the OnTime project ([Franke et al., 2013](#)), timetables should be feasible (trains adhere to scheduled paths under normal operating conditions without delays or conflicts), robust (they can absorb day-to-day operational variations and maintain planned schedule throughout), and stable (they can recover from primary delays so that trains regain their scheduled paths). However, while planned train service (as opposed to actual passenger) demand on railways should not exceed the normally available system capacity, there are inevitably situations where operational disruptions occur and some congestion is inevitable. These include “supply-side” problems such as train failures, switch failures, damage to electrification equipment, bridge strikes by road vehicles, speed restrictions, etc. and also “demand-side” issues, such as variations in passenger demand, sometimes causing extended alighting and boarding times, and thus extended dwell times and consequent delays. As a result of such events, capacity (and hence supply) may be temporarily reduced, and exceeded by planned demand, thus causing congestion to occur.

Ideally, sufficient spare capacity should be available in a railway system to enable the absorption of and rapid recovery from the effects of routine, “normal” levels of disruption, although this redundancy comes at a cost. It is, however, difficult to maintain and provide spare capacity when demand is high and growing, and when the cost and difficulty of providing additional infrastructure are prohibitive.

The situation is complicated further in the railway sector by the fact that there is typically no unique definition or agreed upper limit of capacity (in contrast with other transport modes). This is because the potential capacity depends not only upon the technical characteristics of the infrastructure and the trains using it, but also upon how the capacity is used or consumed, that is the mixture and sequence of trains being operated, some of which may have very different stopping patterns and average speeds. Alternate fast and slow trains on the same tracks is particularly intensive in terms of capacity use and, in such circumstances, “flights” of trains of similar speed are often scheduled together. As already indicated such capacity utilization can be calculated for both nodes and links.

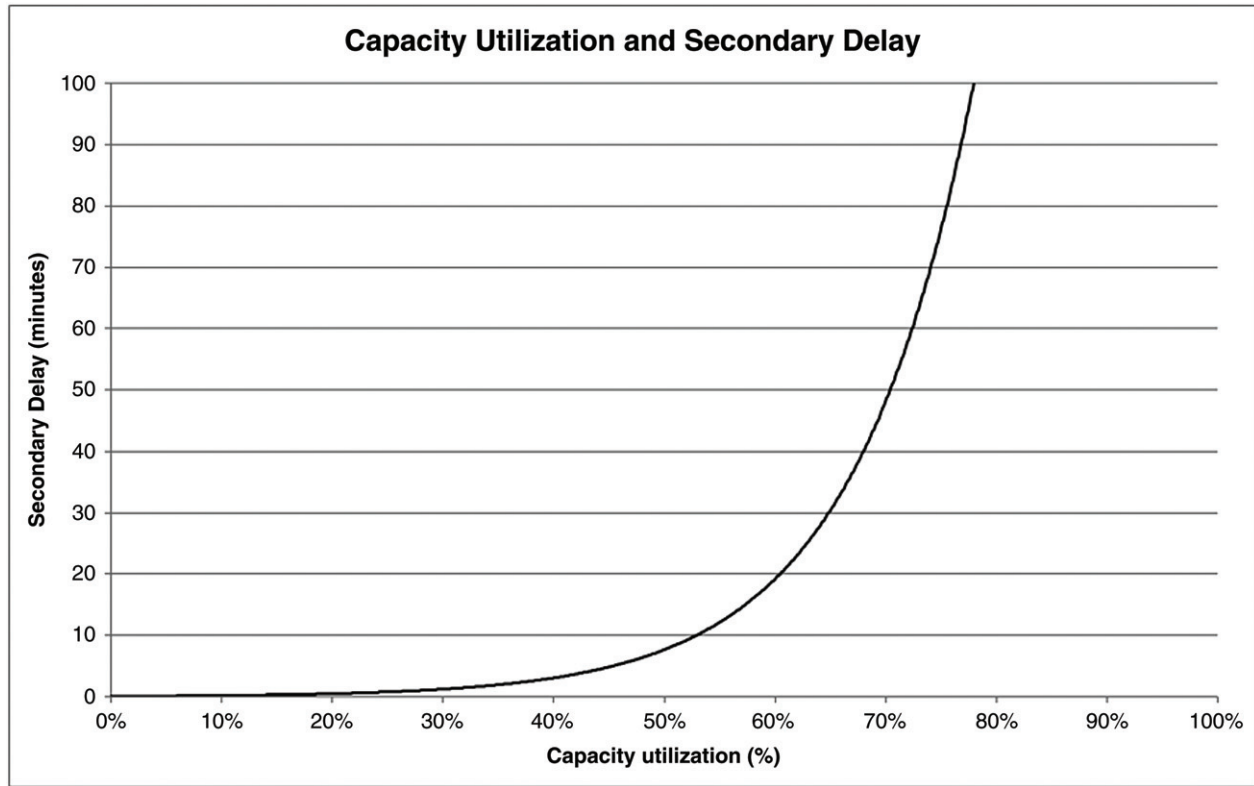


Figure 1 Theoretical relationship between capacity utilization and secondary delay.

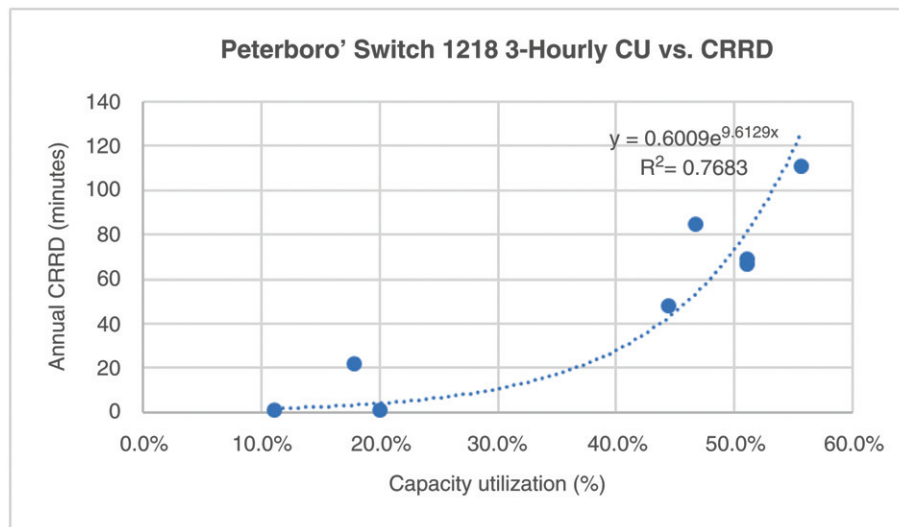


Figure 2 Relationship between recorded capacity utilisation and secondary delay (based on 3-hour time slices).

However, the upper capacity limit also depends upon the required level of performance (i.e., punctuality and reliability of train services) for individual routes (which in turn are combinations of nodes and links).

The wider relationships and interdependencies between timetable stability (i.e., performance with respect to delays), capacity, train speeds and service mix are illustrated in Fig. 3, taken from UIC's 2004 leaflet 406, "Capacity" (page 3). The yellow and black lines can be considered fixed in length, with the result that adjusting one of the four parameters shown will also affect at least one of the others. For example: as the number of trains increase, stability reduces (and vice versa); as average speed increases, stability reduces (and vice versa); and as heterogeneity increases, the number of trains reduces (and vice versa).

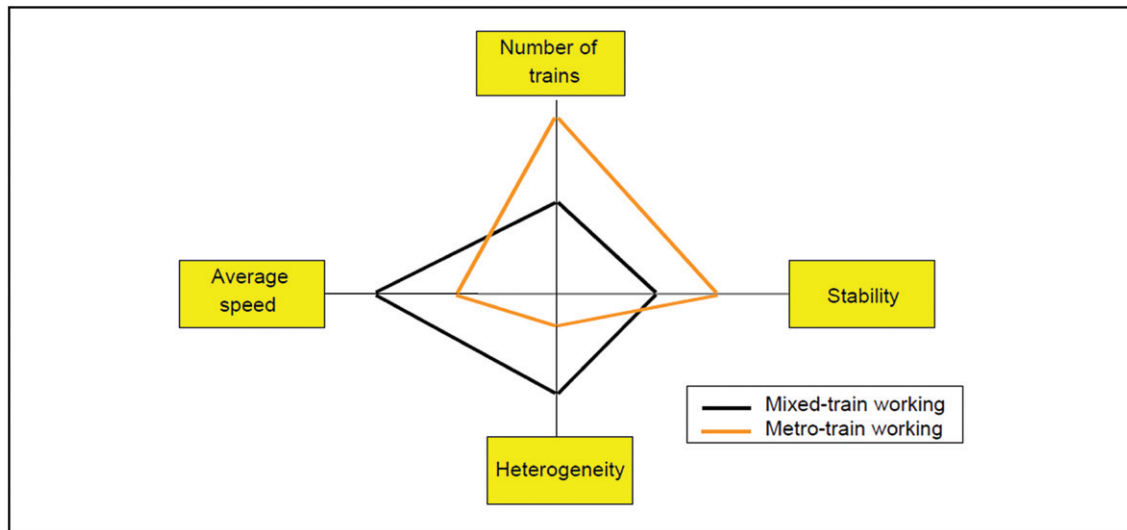


Figure 3 Operational and performance interdependencies.

However, disruptive events (such as the supply- and demand-side examples above) inevitably occur, causing temporary reductions in capacity, and thus congestion. On busy railways, operating at or near capacity, even small initial, or primary, delays to individual trains can cause widespread (through both space and time) congestion and secondary, or reactionary, delays.

The allocation (absolute and relative, i.e., total and mix of types) of train paths and capacity has a major bearing on the likelihood of congestion occurring. The more capacity that is allocated and utilized/consumed, the greater the risk of congestion and the less the scope for service recovery from perturbation. These risk are particularly great at nodes—stations and junctions (Armstrong and Preston, 2017). Also, the greater the heterogeneity, that is the variation between the speeds and stopping patterns of the allocated paths, the more capacity that is utilized/consumed, and the greater the potential for congestion.

In addition to these “macro-level” factors affecting the potential for congestion (and, of course, general infrastructure and rolling stock quality and reliability), train service quality and performance (i.e., punctuality and reliability) can be affected by a range of more detailed, “micro-level” aspects of the timetable and responses to disruptive, congestion-causing events. These include:

- The allocation (size and distribution) of running and dwell time supplements to individual trains—these help to ensure that trains can operate as scheduled, even when affected by minor variations in day-to-day operating conditions
- The allocation of “buffer times” between successive train movements—even if there is spare capacity in the overall timetable, running individual train pairs at or near their minimum headway, or time separation, increases the risk of unplanned interactions between them, and thus delay and congestion
- Timetable complexity, including variations in train stopping patterns, joining and splitting of trains en route, and the allocation and “sharing” of rolling stock and train crew across multiple routes, thus increasing the risk of delays and congestion being transmitted across the network, rather than being self-contained, or “quarantined” within the section of the network initially affected
- The availability, quality and flexibility of contingency plans available to respond to service disruptions
- The availability of signallers experienced to respond to a potentially wide range of both minor service perturbations and more serious disruptions, including the implementation of contingency plans
- The availability of sufficient human resources and quality of communication to respond quickly and effectively “on the ground” when problems occur

Figs. 4 and 5 illustrate both the macro- and some of the microlevel timetable factors, showing contrasting degrees of train service heterogeneity and complexity. Fig. 4 shows a spreadsheet-based graph of trains on London Underground’s Victoria Line between Tottenham Hale and Highbury & Islington in the morning peak, with very high levels of service frequency and uniformity, corresponding to the yellow line in Fig. 3. Fig. 5, in contrast, shows a graph of train services on the South West Main Line between Southampton Central and Eastleigh in the pre-morning peak, with a mixture of fast and local/stopping passenger services (blue and green lines, respectively), freight (red line) and empty coaching stock (ECS; black lines) trains, with varying stopping patterns and speeds. This corresponds more closely to the black line in Fig. 3.

On a busy mixed traffic railway, carrying a combination of fast, longer-distance and stopping/local passenger services, possibly in conjunction with longer-distance, but relatively slow freight services, developing a conflict-free, commercially attractive timetable that meets the needs of all its users and operators, is a major challenge. The challenge includes the provision of attractive journey times and service frequencies between origin and destination station pairs, by direct train service and/or via interchange, and

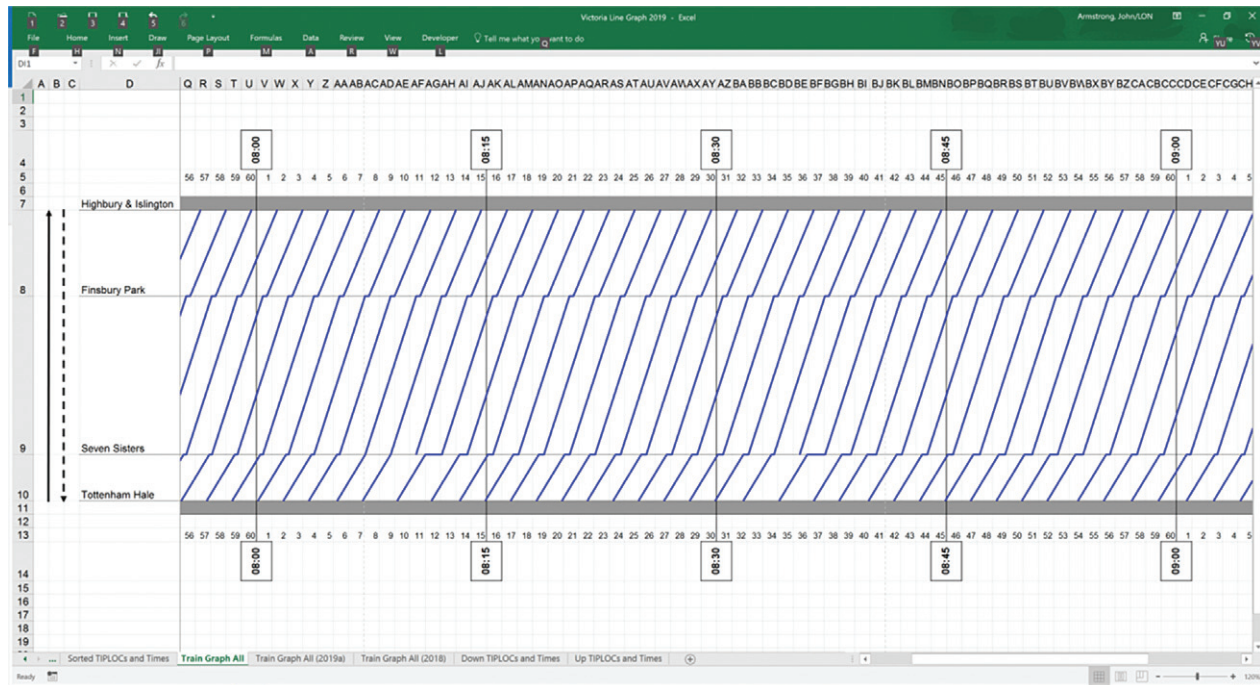


Figure 4 Victoria line train graph: homogenous services.

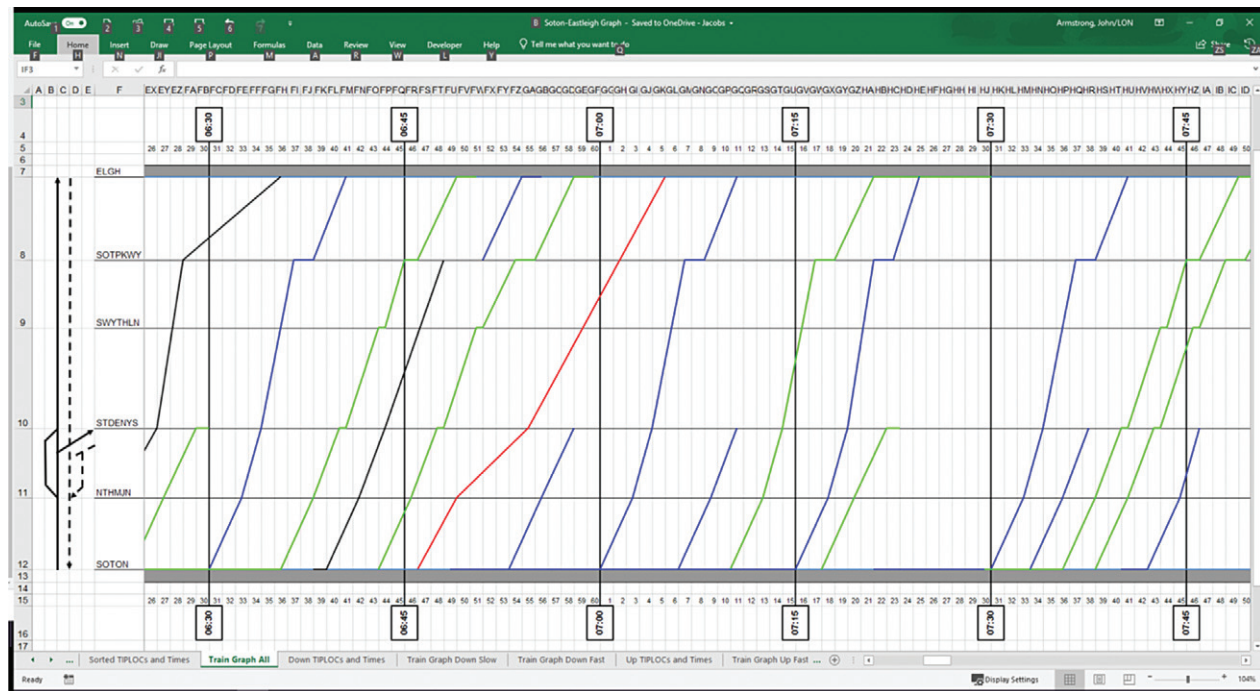


Figure 5 Mixed traffic train graph: heterogeneous services.

achieving a balanced, near-optimal outcome can be time-consuming and difficult. Train planning was described in a recent edition of *Modern Railways* magazine as a "deliciously complex nine-dimensional problem." These nine dimensions are:

- The external environment, including funding, governance and service specification, but also users and pressure/advocacy groups
- The passenger and freight markets to be served, how best to meet and balance their differing respective needs, and, particularly, how to respond to demand growth in a capacity-constrained system

- The three dimensions of space-these include the locations of and distances between the origins and destinations of demand, the gradients between them (although modern rolling stock is less affected by these, and less heavy freight tends to be carried on the network, than was previously the case), and the constraints imposed upon operational flexibility by a track-based system (railways have been described as a “single-degree-of-freedom mode of transport”, and as “operating in one-dimensional space”)
- The dimension of time-the travel times (“start to stop”, “start to pass”, “pass to pass”, and “pass to stop”) between location pairs, and the minimum times between successive train movements, are the fundamental “building blocks” (and constraints) upon which the timetable is based
- Infrastructure and rolling stock constraints - these include numbers and layout of tracks, loading gauge and axle load, numbers and layout of station platforms, signaling capability, availability and type of electrification, and rolling stock capacity and capability
- Deliverability of the timetable, including crew scheduling and changeovers, capacity utilization, reliability and punctuality
- Commercial considerations, i.e. the business case for a planned timetable, the anticipated revenue and costs of operation, and the resulting profit or subsidy requirement.

It can thus be seen that railway timetable development and capacity allocation is a highly complex, multidimensional and multistakeholder, and time-consuming process, with considerable scope for problems and error, as witnessed in the case of the May 2018 timetable update in Britain, and the effects it had on users of Northern and Govia Thameslink Railway (GTR) train services in particular.

Capacity Allocation

Partly because of the complexity of producing safe and reliable timetables, railways have traditionally been vertically integrated monopolies, with one organization responsible for both train services and infrastructure. In such circumstances, train paths are allocated using an administrative system of centralized command and control. Where there are conflicts in timetabling, these are typically resolved based on an hierarchy of service priorities, loosely based on perceived willingness to pay. For example, a high value, regular express service will typically be given priority over a low value, irregular freight train as it is believed its customers have a higher willingness to pay to avoid delays. Such priorities are often enshrined in timetable planning rules and are also used for rescheduling services in the case of major delays rather than first come, first served approaches. Perfect planning would use variants of social cost-benefit analysis to determine these priorities. In the absence of such analysis, administrative methods only work well when all transactions are internalized within the one firm as when there are many firms there will be information asymmetries between the infrastructure authorities and railway undertakings. In practice, priority allocation methods tend to favor existing services and can be criticized as stifling innovation and enshrining grandfather's rights.

Where there is a fragmented industry structure, with vertical separation (infrastructure operations separated from train service operations) and/or horizontal separation (several train service operators competing with each other), then the traditional capacity allocation methods break-down. Such a fragmented structure is becoming the norm in the Europe Union following Directive 91/440, which encouraged vertical separation, at least of accounts, and the slow emergence of the Single European Rail Area. The initiatives in Europe followed on from the deregulation of railroads in the US as a result of the Staggers Act (1980). In such cases of market liberalization, given rail infrastructure is a natural monopoly, access and the concomitant prices (track access charges) need to be regulated by a third party, such as the Office of the Rail Regulator (and its successors) in the UK or the Interstate Commerce Commission (and its successors) in the US.

One approach is the efficient component pricing rule as championed by William Baumol and used in the United States (Baumol, 1983). Access is granted to an entrant if they are prepared to pay at least as much as the revenue foregone by the incumbent as a result of entry minus the amount the incumbent charges itself for the use of the infrastructure. Where additional paths are provided (rather than the incumbent is supplanted), the entrant should be charged the incremental costs of these paths and the opportunity cost (net revenue foregone) to the incumbent of providing these paths. However, this presupposes honest and transparent accounts and that incremental costs and revenues can be easily measured, which even for additional expenditure on track maintenance has not proved straightforward.

In Europe, there have been four packages of railway reforms. The First Railway Package originated in 2001 and made provisions for open access competition in the international rail freight market, initially on the Trans European Rail Freight Network. It included directives to ensure fair track access and charging. The Second Railway Package was outlined in 2004 and extended rail freight deregulation to the whole network (realized in 2007). It also included directives for the harmonization of safety requirements and certification, and for interoperability, as well as a regulation for the creation of a European Railway Agency (ERA - renamed the European Union Agency for Railways in 2016) for safety and interoperability. The Third Railway Package was approved in 2007 and includes the liberalization of international passenger services by 2010, the harmonization of train drivers' licences and the inclusion of passenger rights and freight service quality regulations. The Fourth Railway Package was approved in 2016 and aims to complete the single European railway market by permitting open access competition for domestic services, competitive tendering for public sector contracts and the harmonization of regulation through the ERA. In addition, framework conditions were established concerning the independence of infrastructure managers, access to stations and facilities, through ticketing and inter-available ticketing.

There is thus harmonization at the European level. For example, European Union Directive 34/2012 requires infrastructure managers to grant nondiscriminatory access to railway undertakings and follow set procedures for the allocation of railway infrastructure capacity and set methods for the calculation and collection of infrastructure charges. Nonetheless, there has been a wide range of rail infrastructure charges, with at one extreme those based on short run marginal cost, related mainly to maintenance (as in Sweden) and at the other those based on fully allocated average costs (as in France for High Speed services or in Central and Eastern Europe for freight services), although only rarely have congestion costs been internalized (Nash, 2005). Hybrids might involve a two-part tariff, where an entry fee makes a contribution to fixed costs (dependent on the amount of subsidy) and a usage fee covers variable costs or some form of Ramsey pricing in which inelastic markets face higher charges than elastic markets, with the mark-up dependent on the nature of the budget constraint.

Where capacity is scarce, marginal cost charging that takes into account congestion or scarcity costs is required, although such charges are not straightforward to calculate (Johnson and Nash, 2008). A congestion charge is the appropriate capacity cost where the train in question constitutes an additional train to what would otherwise have been run. Such charges have been levied in Britain since the first periodic review of access charges in 2000 and are based loosely on the exponential relationship between additional trains and delays, as shown by Figure 1 but are charged by service group and include a large amount of averaging.

Where the train in question runs instead of some other train, the appropriate capacity cost is the opportunity cost of trains forced off the system by lack of capacity. Such scarcity costs can be estimated as the sum of the additional amount of traffic attracted to rail by the presence of this potentially additional train multiplied by the price paid, the consumer surplus to rail users as a result of the additional quality and capacity provided by the train (including reduced crowding on other services) and the savings of external costs to road users and the public at large from the train attracting passengers from road. From this should be subtracted the train operating, infrastructure and external cost savings from failing to run this train. This might be seen as a social version of Baumol's efficient component pricing rule. The practical problem is to identify and quantify the appropriate alternative services.

Another approach is to identify sections of infrastructure where capacity is constrained and to charge the long run average incremental cost of expanding capacity (Hysten, 1998). However, this is a very difficult concept to measure as the cost of expanding capacity varies enormously according to the exact proposal considered and it is not easy to relate this to the number of paths created since they depend on the precise number and order of trains run. It may be argued, however, that more appropriate incentives are given to infrastructure managers if they are allowed to charge the costs of investment they actually undertake, rather than for the scarcity resulting from a lack of investment, at least if they are commercially oriented. Short run marginal cost pricing can encourage the infrastructure manager to restrict capacity in order to keep prices high, whereas a system where a capacity charge reflected actual expenditure on expanding capacity would overcome this problem.

In Sweden, in the context of initially a low density of service and hence some spare capacity, it appears that some 30 years of using short-run marginal costs has meant that the network has been filled with subsidized local services to the detriment of more commercial longer distance services. In Great Britain, where franchised passenger train operators face a two-part tariff but open access passenger and freight operators face only a variable charge, the regulator has permitted some open access entry using the not primarily abstractive test. In this test, entrants have to demonstrate that they are generating new demand and not just abstracting demand from existing services. This has limited entry to serving peripheral markets that have not traditionally been well served by the incumbent. There has though been a substantial increase in franchised train operations, possibly in part to block open access entry but also because promises of higher service levels increase the chances of franchise bids being successful. In neither Sweden nor Great Britain has there been substantial expansion of infrastructure capacity. Instead, intensity of utilization has increased, with some deterioration of performance.

Theoreticians have proposed that auctions would be the most appropriate way to allocate rail infrastructure (Brewer and Plott, 1996). However, this would require huge combinations of space (typically access should be allocated at the signal block level) and time that would be computationally prohibitive for a reasonable scale mixed traffic railway. Furthermore, complex interdependencies between paths (and the services using them) would need to be taken into account. Solutions might involve an iterative process in which operators initially put in separate bids for train paths, with some indication of the extent that they would be prepared to be flexible. These bids would then be optimized into bundles of paths that operators would be able to rebid for. This would mimic aspect of the existing administered timetabling process (Nilsson, 2002).

Competition

Given the above, it is perhaps not surprising that competition within railways has been relatively limited and vertically integrated systems remain the norm, either in private ownership (Class I railroads in the US or the three main passenger railways in Japan) or public ownership (for example China, India and Russia). Infrastructure is allocated using internal bureaucratic processes rather than external contractual mechanisms. In some respects, this lack of within rail competition may not matter where there is intermodal competition from air, road or sea, efficient practices will be promoted. Where there has been intramodal competition, for example in North West Europe and, to an extent, in Latin America, off the track competition, in the form of concessions, franchising and tendering, is more typical than on the track competition, with infrastructure again being allocated using administrative rather than market-based measures. Where there has been on track competition, this has partly been stimulated by relatively low track access charges (Austria, Czech Republic, Italy, Sweden). There have been concerns that such oligopolistic (small group) competition may lead to too much service being provided at too high a price compared to a welfare maximizing benchmark (Preston, 2008). The

recent evidence from Europe is that services have indeed increased on competed routes but prices have come down, at least in the short run, and there has been greater product differentiation (Beria et al., 2019; Laroche et al., 2019; Tomes et al., 2016). However, this is compared to a monopoly benchmark in which there may have been too little service at too high prices, whilst it is by no means certain that the intense on-track competition being seen in some European countries can be sustained financially by the protagonists in the long-term.

Concluding Comments

Given the high degree of technical interdependencies in railway systems, safe and reliable operations require a degree of redundancy. This means that congestion effects can begin to appear at relatively low capacity utilization levels. In the absence of appropriate charges, this poses a challenge for the allocation of train paths through the infrastructure. The traditional solution has been for the railways to be vertically integrated monoliths, with little or no competition and track access allocated using centralized administrative processes. However, this is a description of the analogue railway. The digital railway promises to change the nature of some of these interdependencies. Moving block signaling, such as European Train Control System Level 3, has the potential to increase capacity by 40%, at least on plain track (Wendler, 2007), thus reducing congestion and increasing the scope for competition. Smarter timetabling and dynamic rescheduling can reduce the occurrence and impact of delays. Advances in machine intelligence and high-performance computation mean that decentralized market-based allocations of track access, through auctions, are now a practical proposition. It remains to be seen if this new digital age will recast railway operations and economics, by moving to a more open system in which on-track competition becomes the norm. It is possible that such developments have, at least for now, been thwarted by the downturn in the demand for passenger railways in many parts of the world following the emergence of COVID19 in Spring 2020.

References

- Armstrong, J., Preston, J., 2017. Capacity utilisation and performance at railway stations. *J. Rail Transport Plan. Manage.* 7, 187–205.
- Baumol, W., 1983. Some subtle issues in railroad regulation. *Int. J. Transport Econ.* 10, 341–345.
- Beria, P., Tolentino, S., Bertolin, A., Filippini, G., 2019. Fares strategies and head-on competition. The case of Italian rail market. *Proceedings of the 15th World Conference on Transport Research*. Mumbai.
- Brewer, P.J., Plott, C.R., 1996. A binary conflict ascending price (BICAP) mechanism for the decentralized allocation of the right to use railroad tracks. *Int. J. Industrial Organ.* 14, 857–886.
- Franke, B., Seybold, B., BÄcker, T., Graffagnino, T., Labermeier, H., 2013. OnTime–Network-wide analysis of timetable stability. *Proceedings of 5th International Seminar on Railway Operations Modelling and Analysis–RailCopenhagen*, 2013.
- Hyllen, B., 1998. An Examination of Rail Infrastructure Charges. Final Report for the European Commission, DG VII. Prepared by NERA, Nomisma, VTI, IVE ENPC & BPM, London.
- The Institution of Civil Engineers (I.C.E.), 1989. Congestion. Thomas Telford, London.
- Johnson, D.H., Nash, C.A., 2008. Charging for scarce rail capacity in Britain: a case study. *Rev. Netw. Econ.* 7 (1), 53–76.
- Laroche, F., Lamatkhanova, A., Grassot, L., 2019. Exploring the effects of inter- and intra- modal competition on prices and frequencies in the railway markets: Evidence from Europe. *Proceedings of the 15th World Conference on Transport Research*, Mumbai.
- Modern Railways, 2019. In Praise of the Train Planner 76 (846), 60–65.
- Nash, C.A., 2005. Rail infrastructure charges in Europe. *J. Transport Econ. Policy* 39 (3), 259–278.
- Nilsson, J-E, 2002. Towards a welfare enhancing process to manage railway infrastructure access. *Transport. Res. A* 36, 419–436.
- Preston, J.M., 2008. Competition in transit markets. *Res. Transport. Econ.* 23, 75–84.
- Tomes, Z., Kvizda, M., Jandova, M., Rederer, V., 2016. Open access passenger rail competition in the Czech Republic. *Transport Policy* 47, 203–211.
- UIC (Union Internationale Des Chemins De Fer), 2004. Leaflet 406: Capacity, 1st edition. UIC, Paris, 2nd edition published 2014.
- Vuchic, V.R., Kikuchi, S., 1994. The bus transit system: its underutilized potential, Report DOT-T-94-29. Federal Transit Administration, Washington, DC.
- Wendler, E., 2007. Influence of ETCS on Line Capacity. *UIC ERTMS World Conference* 11–13 September. Berne, Switzerland.

Public Private Partnership

David Meunier*, Emile Quinet[†], *LVMT, UMR T9403, Ecole des Ponts ParisTech, France; [†]Paris School of Economics, Ecole des Ponts ParisTech, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	385
What is a PPP?	385
Why use a PPP?	385
How to Manage a PPP?	386
Procurement Organization	386
Contract Regulation	387
Risk	387
PPP in the Transport Sector	388
PPP in Road Transport	388
PPP in Ports	389
References	390
Further Reading	390

Introduction

Public Private Partnership (PPP) is a fuzzy expression, covering a lot of different but close meanings about the association of public and private funds for the realization of essential facilities. The corresponding procedures are widely used for almost all kinds of public investments but more so in transport sector where it is shaped by several interesting features, which we will delve into after discussing some basics of PPP.

What is a PPP?

The concept of PPP covers a wide range of situations and is subject to various interpretations. Many definitions exist in the literature. According to the PPP Reference Guide ([Partnership, 2017](#)), edited by a consortium of major international organizations lead by the World Bank, PPP is “a long-term contract between a private party and a government entity, for providing a public asset or service, in which the private party bears significant risk and management responsibility and remuneration is linked to performance.” In this definition, “asset or service” means that both new and existing assets are concerned; it also includes the cases in which the private party is paid entirely by service users, and those in which a government agency makes some or all payments. It encompasses contracts in many sectors and for many services, provided there is a public interest in the provision of these services and the project involves long-life assets linked to the long-term nature of the PPP contract.

An alternative definition ([Sabot and Puentes, 2014](#)) is “a legally binding contract between a public sector entity and a private company where the partners agree to share some portion of the risks and rewards inherent in an infrastructure project.” Anyhow, the choice of the structure is embedded in the legal traditions of the countries, as it appears from the comparison of the guidelines of various countries (see, for instance, [HM Treasury, 2003](#)) and also in the current practices ([Medda et al., 2013](#)).

From this quick review it is clear that in fact PPP does not imply just funding but also other aspects such as risk management, legal responsibility, in fact, the whole management of the project, which can then cover the range between “pure public” management and “pure private” management.

Why use a PPP?

As far as public infrastructure such as essential facilities are concerned, the reasons for public intervention and motivations are numerous. Some are general and linked to the need for developing policies where private initiative is insufficient, such as the emergence of new activities, network effects and needs of coordination. Other ones are related to strategic issues, for instance military purposes. Others are more common and well known by economists; they are linked to market failures. Among them are general market power and customers’ information, which justify the role of Competition Authorities. Other market failures, frequent in transport, are externalities and market power resulting from increasing returns to scale. Increasing return to scale is a common feature of infrastructures. Externalities are most commonly negative, such as environmental externalities, congestion and safety externalities, but some can be positive like the agglomeration externalities. Other reasons of public intervention lie in distributive concerns, or local development issues, for instance in the cases of seaports and airports.

But public intervention has limits from several points of view. The first one is the commonly recognized poor public management vis-à-vis private management. Economic literature has developed this point in two different directions. The first one is the bureaucracy theory (Niskanen, 1971), which stresses the fact that public management is based on the respect of standards and budget constraints, contrary to the private management, which is based on efficiency and search for profit. The second one is the theory of organization (Williamson, 2002), which compares the forms of control in both types of management (Vickers and Yarrow, 1988). It shows that public control is looser than private control. Private firms are acknowledged to have a technological comparative advantage: they are more able to market their products and to fit better the needs—at least the bankable needs. Public authorities may want also to introduce some flexibility so as to circumvent administrative or legal constraints and this goal can be achieved through private management. Private funding is more able to use innovative financial tools. Private access to international and more competitive financing is often sought out, though states may get equal or, quite often, better financing conditions. In some countries, PPP may be a device to alleviate the public budget constraint. It is for instance the case when public authorities are not allowed to levy infrastructure charges, contrary to a private firm. It may be also the case when no infrastructure charge can be levied, but when the concessionaire receives a yearly subsidy from the public authority, as it is the case for the private finance initiative (PFI) in the United Kingdom: this process allows to shorten the implementation of a program for a given amount of money. Another often advocated motivation for PPPs is a great reduction in delays for delivering the infrastructure and related services. But ex-post analyses do not always confirm this belief (see for instance the case of Irish motorway PPPs in Ireland (Reeves, 2005)).

For some items, private and public management are complementary. For example, in the case of risk, some risks such as financial risks, are usually considered as better managed by private firms, and some other risks, such as macro-economic risks, as better managed by the public authority. These considerations draw the pertinence scope of PPP: when private motivations are large, the private firm does better; when public motivations are large, public management is better. In the middle and in the case where motivations are shared lies the PPP pertinence. Also, risk transfer to the private partner entails an inefficient risk-pricing premium that should be compensated by an additional efficiency gain (Makovsek and Moszoro, 2017). This is also true for the frequent case where the PPP generates higher financing costs and high transaction costs—public as private—due to PPP complexity.

But cooperation between these two agents is difficult as their goals are different: the private partner seeks to maximize profit; the public partner seeks to maximize welfare.^a The public authority must induce its private partner to achieve welfare goals without losing the efficiency of private management. How to choose the better PPP arrangement, giving the best public value for money, and should this arrangement be preferred to full privatization or to public management?

How to Manage a PPP?

Let us proceed through the history of a PPP, beginning by the procurement, then the contract and its life.

Procurement Organization

The first decision is whether and how to fragment the missions. From the geographical point of view, the range of possibilities goes from an isolated link or node to the entire network of the country. From the technical point of view, it is possible to put the private partner in charge of either the whole activity, from the initial design to the maintenance and operation of the infrastructure, or in charge of just a part of the process, for instance construction, or operations, or maintenance. Many possibilities exist, such as the well-known BOT (build/ operate / transfer back). A well-designed fragmentation in the technical and geographical dimensions can reduce the market imperfections and/or increase the efficiency of the management. The fragmentation has consequences on efficiency. Many reforms have led to fragmentation of the incumbent firms in order to have smoother and easier to manage organizations. It is for instance the case of the railways reforms in Japan and in South America (Thompson, 2001) where the big incumbents have been geographically split up into several smaller firms.

Fragmentation has also been used for introducing competition. For instance, the European Railways Reform (Commission of the European Community, 1991) was based on the separation between infrastructures and operations; the aim was to isolate operations, which had almost constant returns to scale and could be run in the framework of competition, and to leave apart infrastructures, which clearly have increasing returns to scale and would therefore better be run as a regulated monopoly.

Geographical fragmentation is limited by economies of scale. Building a motorway exhibits large economies of scale up to a length of about 200 km and, above this length, returns to scale are roughly constant (Fayard et al., 2005). The same happens for maintenance: a maintenance center is best fit for a length of a few hundred kilometers (Link, 2006). The geographical fragmentation is also influenced by the size and nature of externalities (Meunier and Quinet, 2007): fragmentation of two adjacent links can induce double monopolization.

Fragmentation is also a means to set up yardstick competition^b when it is possible to create several operators; and the larger the number of operators is, the more efficient the competition may be (Bouf and Lévêque, 2004). Vertical fragmentation is limited by

^a Though the public objectives may also be multiple and possibly contradictory (e.g., public debt reduction, development of transport network, environmental goals, etc.), which may make more complex the decision to choose a PPP, but also its management and regulation.

^b Yardstick competition is a regulatory means, either formal or informal, of putting pressure on a firm through comparison of its results (in terms of costs, quality, or demand) with the results of similar firms.

potential economies of scope. For instance, in the case of railways, several authors consider that economies of scope between infrastructures and operations are too large to make fragmentation interesting (Basso and Jara-Diaz, 2005). But this argument has never been presented in the case of air transport, which economies of scope are quite doubtful (Oum and Fu, 2008). The recommendations of the bodies in charge of PPP design insist on these complementarities. An important issue is to see whether there are or not complementarities between building and maintenance, since they command the issue of separating or not these activities (Iossa and Martimort, 2008).

Contract Regulation

If regulation were perfect, i.e., if the regulator or the public authority knew perfectly well the ability of the private partner and were able to perfectly observe its effort (no information asymmetries), and if there were no uncertainty, then in this first best situation, the regulation could reach the first best optimum. The real world is different. None of these conditions is fulfilled, and the contract has to minimize the gap with the first best situation, dealing with risk, information issues (adverse selection and moral hazard) and transaction costs.

Risk

The agents take a risk premium depending on the risk level of the project. We have a poor knowledge of risk premiums. Risk premiums depend also on the agent's risk aversion; from that point of view, public operators are in general less risk averse than private ones. Risks are usually classified as political risk (changes in taxes, in standards, and public decisions that may modify the profitability of PPP), technical risk (mainly during the building phase) and commercial risk (demand risk).

The general principle of risk sharing is that each risk should be shared between the agents according to two factors: the agents' respective ability to get information on and to deal with this risk, and their risk aversion. From this point of view, the demand risk of a motorway could be mainly borne by the franchiser since he can influence it by his actions (for instance through deciding to improve or not a parallel road) and the concessionaire has no large possibility to act on it and furthermore has a larger risk aversion. The situation would be different if the concessionaire had the possibility to influence the traffic level—more generally the demand—by means of some marketing action (as it may be the case for airports). Nevertheless, in case the PPP directly gets revenues from the users, the private partner may, through adjusting the price level or differentiating the prices, try to increase its profit at the expense of the users: the public authority has then to set in the contract the degree of freedom left to the private partner. Besides direct revenues stemming from the infrastructure by itself, the private partner may also have opportunities to take advantage of other assets or markets linked to the PPP (duty-free sales in airports, shops in stations, land development, etc.). The recommendations for PPP management insist on the risk sharing issues along the above lines and their consequences for productivity and costs.

Some studies have enlightened the effects of a more or less tight regulation on productivity (Cantos et al., 1999; Gagnepain and Ivaldi, 2002). A tight regulation is measured by the “distance” between the private firm and the public authority as well as the risk sharing agreement (price-cap or cost plus regulation). It appears that the tighter the regulation and the farther the agent from the public authority, the lower the production cost is; here again, the focus of these studies was on operations more than on infrastructures. These studies show that it is possible to increase the effort of the private partner through an increase of its risk exposure. But its information rent^c is then increased, as shown by theory. There is then a trade-off, widely studied by theory (Laffont and Tirole, 1993), between the effort and the rent.

Reducing this asymmetry is a major objective of the public authority, as shown for instance in the PFI recommendations (Estache and Serebrisky, 2004). Asymmetric information is one of the major causes of the regulatory inefficiency (indirect capture); this gives the operator a lever to influence the public authority, induce it to make wrong judgments and estimates and lead it to decisions that are bad for welfare but good for the operator.

The contract life is concerned with the contract duration and possible renegotiations. The contract duration must, first, be long enough to cover a sufficient proportion of the total cost through tolls or charges. Given that the cost of infrastructure is very high in general, this point advocates for long durations. Another argument advocates in the same direction: the operator must bear the fruits of its decisions, especially in case of long-term investments. This leads to proportionate the length of the contract to the lifetime of the work. Conversely, the length of the contract should not be too long in order to avoid a too large information or risk rent. Theoretically, demand risk can be coped with using endogenous duration contracts (Nombela and de Rus, 2004): in broad outline, when demand is lower than forecasted, the duration of the contract is postponed, and reversely when the demand is larger. In the case of long duration, it often happens that the risky alternatives are numerous and even difficult to enumerate, even more hard to assess; also, it is difficult for both parts to credibly commit that they will never renegotiate, even in case of the bad outcomes of the risks. In such situations, it is simpler to sign an incomplete contract with a pre-defined renegotiation process. Such is the case too when the costs of possible errors in a rigid contract are large, or when there is a large degree of trust between the partners (Athias and Saussier, 2018). The exemplary nature or the repeatability of the contract may give reasons for confidence. Furthermore, renegotiations often reveal that the initial PPP design is not always optimal; for instance, the renegotiation of some motorway PPPs in Portugal lead to transferring operational service to the public sector, at a lower marginal cost.

^c Information rent is the additional earning an agent derives from having a private information not available to the other agents with whom he is engaged in a bargain.

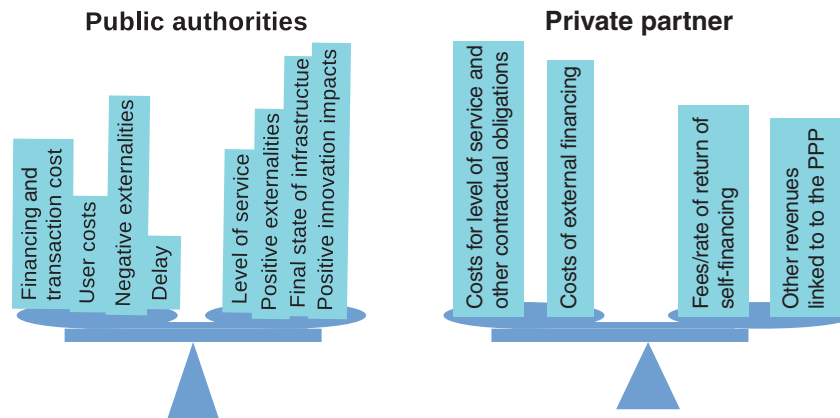


Figure 1 PPP in the transport sector.

PPP in the Transport Sector

Perhaps the most important use of PPP is in the transport sector; though there is no exhaustive statistics on the number of PPP by sector, this affirmation is supported by the records of EIB funded projects. Similarly, the World Bank records 4 G\$ in water and sanitation, 70 G\$ in transport and 37 G\$ in energy for year 2015. There are obvious reasons why transport has such a leading place in PPP (Fig. 1).

On the one hand, most transport infrastructure appear beyond the short-term fiscal capacity of public funding sources, while in many situations it is technically possible to levy charges on most of these infrastructures, though those charges are not sufficient to cover all expenses. On the other hand, such infrastructure entail externalities implying public interventions. These reasons have different weights according to the type of transport infrastructure. These points are exemplified for road, which presents simpler characteristics and more examples; some general observations are given for port infrastructure.

PPP in Road Transport

Motivations of private involvement in road mainly stem from funding, as the issues of costs, productivity, technique, innovation and financing should have decreased in developed countries. Road products and services have become standardized and know-how is widely spread and internationally available. Nevertheless, the innovations that may stem from new information and communication technologies (NICT) offer an important potential: motivations related to innovation and reactivity have thus grown in the last 10 years or so. For countries that do not have much-developed road or motorway network and lack strong competencies inside the public sector, the motivation may be high to have private firms bring their ability to optimize costs in relation to local conditions and to optimize the financing mechanisms of infrastructure projects.

Risk and uncertainty in the road sector are rather low. Construction risk is essentially the concern of private, and remains in general low, thanks to the high level of competency available internationally. Demand risk is not much in the hands of the private partner as it cannot take much action on it. The risk to demand is perceived by the infrastructure manager (IM) essentially through the risk to its revenues, and most of this uncertainty disappears once the motorway has come into operation and becomes a brownfield operation. Exogenous uncertainties tend to grow as concerns the macroeconomic trends and petrol prices. The present global economic crisis has put this to light, and this increases demand uncertainty. Social acceptability of infrastructure represents a type of risk that emerged and has been developing more specifically in the last 10–30 years, depending on the countries. It may be better treated by the public partner when, for example, social acceptability is linked to the confidence level in the collective interest of the project.

The competition framework may theoretically be organized by proper allocation of projects inside consistent programming of a meshed network to check any IM from exerting a strong market power while exploiting economies of scale. The latest are mild, since returns to scale are roughly constant above motorway lengths of about 200 km.

As regards market power, the basic idea would thus be to give to the same IM complementary links and to different IMs the links that may be in competition with each other. Though these principles are clear and simple, it is questionable how far they may be put into practice as competition between two parallel motorways is often limited when these motorways are not very close to each other; and in the rare cases where they are close, the equilibrium may be unstable (case of Mexican motorways experience). In practice though, each road link has very diverse users that come from numerous origins and want to go to numerous destinations. Relations of complementarities and competitions between road links are therefore complex, and they change with the meshing of the network.

Another question is: are there strong positive externalities of infrastructure building, operations and maintenance? For motorways these externalities do not seem to be very strong and manageable. When standard techniques are used for road works, it is

possible for the public authority to estimate average maintenance costs and then global costs, and to impose technical choices, whatever the fragmentation of PPP's missions. Bundling the two missions may present another externality for risk management. For example, the concessionaire of the Millau viaduct in France decided to finance the building phase entirely by equity and, after the viaduct was put in operation, to negotiate far better financing conditions than if he had asked outside financing from the beginning. Doing thus, the concessionaire bore the building risk and the risk on initial traffic level, and reduced the total financing costs since at this stage the risk premium asked by the markets was much lower. This decision was made possible because the concessionaire felt it could correctly manage the construction risks, had enough financial capacity for this, and had enough confidence in the robustness of traffic forecasts.

Theoretically, construction and demand risks can be alleviated through more innovative but rarely used devices, such as endogenous duration contracts (Engel et al., 2001). Quite often, demand risk is found to be more directly taken in charge by the private partner when it finds the market is mature enough (other toll motorways exist in the country and have proved to be profitable, confidence in the traffic forecasts). In such scenario, guarantees and subventions may become lower or even disappear.

Demand risk is often subject to risk sharing between public and private partners of the PPP, whether through guarantees given or by setting in the contract an incentive to voluntarily introduce optimal prices based on information which is not too costly to get: for instance, by automatic traffic counters, or by complaints of the users in case of congestion. And, rather than fixing congestion prices, it may be better to fix the required quality level (e.g., a minimal driving speed) and to let the private partner adjust prices accordingly. This is used for instance on SR91 in California and on A86 motorway near Paris, in a more limited form.

The duration of road contracts generally should come, rather naturally, from calculation of the rate of return of initial investment or the technical lifetime of infrastructure; both entail very long durations. A motorway contract's duration is generally in the range of 20–40 years (except cases of endogenous duration mentioned above). Such long durations imply that numerous unforeseen events may occur, and therefore such contracts are, by nature, incomplete contracts.

This situation entails big consequences for the design and regulation of the contracts: it becomes very difficult to include powerful incentives due to the high risk that the contract, due to its long duration, will anyway be renegotiated. Thus, contractual rules for price regulation apply during successive periods of about 5 years, introducing some flexibility in contracts with a rigid rule that would otherwise be valid for the whole duration of the contract. This depends also on demand uncertainty: empirical studies show that concession contracts with high traffic uncertainty are more likely to be flexible (Athias and Saussier, 2006).

But if revenues are linked to the investment level, for instance on the basis of guaranteed rate of return, private partner may tend to over-invest (Averch-Johnson effect). Such could be the case with endogenous duration contracts: when signing the contract, nobody knows when it will end, and lawyers, accountants, and politicians usually dislike this kind of uncertainty.

To our knowledge there is no example wherein a motorway is stopped to be used or destroyed after the end of the contract. The conditions under which the contract reaches its end will therefore focus on the final requirements for infrastructure and equipment quality, with contractual financial penalties to make these requirements credible. Generally, a new contract will be signed to ensure the continuity of motorway operation, with the same private partner or with another.

PPP in Ports

Public motivations in the port sector have varied historically. Once a major economic potential for countries and cities that were built around port activities, economic role of ports has decreased in relative terms, whereas competition between ports increased as globalization developed, especially during the last decades.

Concerns about market power of ports may depend on the context and on the geographical scale. Indeed, unlike motorways, we deal here with international and, for a good part, inter-continental traffic: a transfer of traffic or activities from one port to another one located in the same range may be seen as positive at the local level (regional) of the second port, but neutral at a higher level (country or EU). There are motivations concerning service continuity (ports are still vital for a big portion of the economy), strategic issues (energy, defence, security), and market power of "captive traffic" (industries located inside the port and very near). And, in the recent period, concerns about sustainable development have grown. Indeed, water and air pollution represent important issues on the seaside; so does biodiversity, especially in estuaries and in the large land reserves that many ports still control.

Some peculiarities of ports must be taken into account here: ports are complex interfaces involving many functions, and are only one link in a whole transport and logistics chain. Therefore, coordination is a key competitive item, and public coordination may be seen as more neutral than coordination driven by one of the actors concerned. Still, very rare have been the ports that did cover all functions by public entities: ports are by nature a place of interface between public and private interests, the question here is rather to analyze the degree of implication of the public authorities in this complex interface.

Motivations for private involvement have grown with the competition toughness, which implied high pressure on costs and productivity, but also with the development of new logistics chains and equipment, implying a need for new technical skills, the ability to innovate or adapt reactively to innovation. In this respect the business plan of a port is, as well as the business plan of an airport, much more sophisticated than the business plan of a motorway, where the single parameter at hand of the manager is the infrastructure charge. It is important also to take into account the long-lasting internationalization trend and the concentration trend in the maritime sector and in stevedoring. The container industry especially was quite a revolution, implying not only a mix of public and private funds, but also strategic views on how to develop the industry: new equipment was required, therefore costly investment; and the exploitation of scale economies driving the container development also meant progressively increased standards of capacity for ports. Indeed, the inter-continental container ships have become quite huge and do not seem to have yet reached their full

potential of scale economies. The increased volatility of traffic implied more weight was put on the commercial concerns of ports and more commercial reactivity, as it is easy to lose a client, but very difficult to get him back or to attract new ones. Private management is usually recommended to front such commercial risks. Furthermore, demand risk has grown with the degree of port competition and the loss of weight of ports in the big players' game; now macro-economic risk has added on top of that. Besides demand risk, which may be considered as the main risk, construction risk is rather low, at least for its technical components.

These considerations about the port sector evolution help to understand why we do observe many forms of public involvement in ports from the "fully integrated public port," that is nowadays kind of a white elephant, to the "landlord port" that lets all physical operations to private actors. A "landlord port" only retains basic functions such as "social regulation" (safety) and most of the time (except in the rare cases of full privatization as in the United Kingdom) control over land affectation/lease and basic infrastructure (navigation channels and berths). This is why we will use the term "port entity" (PE) from now on, rather than the more limited term "infrastructure manager" (Brooks and Cullinane, 2006). The complexity of port issues and the need for rapid reaction to (or anticipation of) market changes may advocate for more autonomy to be left to the PE together with more implication of stakeholders in port governance, which make it difficult to build a complete rigid contract as regards investment. In practice, such contractual coordination structures and organization of dialogue between stakeholders may help to reduce information asymmetry by facilitating learning of chain performance conditions (economies of scale and scope, competitive key factors, limiting factors for the port) and in assessing risks and externalities that are not easily seen unless collectively addressed (the usual problem of multi-principal studied by the literature on industrial organization).

As a conclusion, ports are a natural and complex interface between public and private concerns and actors. Though they are precisely located by nature, they intervene in complex international chains and their time scales are simultaneously very long due to their long-lasting assets and the irreversibility of many decisions, and quite short due to the necessity to react quickly to evolutions in demand or to the competitors' actions. The geographical scale of both public and private entities has major importance. The intervention of the private has been growing in a regulatory framework that tends to be enlarged and renewed: international and national legislation, environmental rules, but also contractual risk-sharing and governance of coordination between the multiple port actors.

References

- Athias, L., Saussier, S., 2006. Contractual design of toll adjustment processes in infrastructure concession contracts. Working Paper ATOM & SSRN.
- Athias, L., Saussier, S., 2018. Are public private partnerships that rigid? And why? Evidence from price provisions in French toll road concession contracts. *Transp. Res. Part A* 111, 174–186.
- Basso, L., Jara-Diaz, S., 2005. Calculation of economies of spatial scope from transport cost functions with aggregate output with an application to the Airlines industry. *J. Transp. Econ. Policy* 39, 25–52.
- Bouf, D., Lévêque, J., 2004. Yardstick competition for transport infrastructure services. ECMT round table. No. 129.
- Brooks, M., Cullinane, K., 2006. Devolution, port governance and port performance. In: Brooks, M., Cullinane, K. (Eds.), *Research in Transportation Economics*, Vol. 17, Elsevier, Oxford.
- Cantos, P., Pastor, J., Serrano, L., 1999. Productivity, efficiency and technical change in the European railways: a non-parametric approach. *Transportation* 264, 337–357.
- Commission of the European Community, 1991. Council Directive 91/440/EEC of 29 July 1991 on the development of the Community's railways. *Official Journal*. No. L 237, 24/08/1991, 25–28.
- Engel, E., Fischer, R., Galetovic, A., 2001. Least-present-value-of-revenue auctions and highway franchising. *J. Pol. Econ.* 109, 993–1020.
- Estache, A., Serebrisky, T., 2004. Where do we stand on transport infrastructure deregulation and public-private partnership? World Bank Policy Research Working Paper 3356.
- Fayard, A., Gaeta, F., Quinet, E., 2005. French motorways: experience and assessment. In: Ragazzi, G., Rothengatter, W. (Eds.), *Procurement and Financing of Motorways in Europe*. Research in Transportation Economics, Vol. 15. Elsevier.
- Gagnepain, P., Ivaldi, M., 2002. Incentive regulatory policies: the case of public transit systems in France. *RAND J. Econ.* 334, 605–629.
- HM Treasury, 2003. HM Treasury PFI Meeting the investment challenge. Available from: www.hm-treasury.gov.uk.
- Iossa, E., Martimort, D., 2008. The simple micro-economics of public-private partnerships. CEIS Tor Vergata Research Paper Series No. 139, 6 (12).
- Laffont, J.J., Tirole, J., 1993. *A Theory of Incentives in Procurement and Regulation*. The MIT Press, Cambridge.
- Link, H., 2006. An econometric analysis of motorway renewal costs in Germany. *Transp. Res. Part A* 401, 19–34.
- Makovsek, D., Moszoro, M., 2017. Risk pricing inefficiency in PPPs. *Transp. Rev.* 38 (3), 298–321, doi:10.1080/01441647.2017.1324925.
- Medda, F.R., Carbonaro, G., Davis, S.L., 2013. Public private partnerships in transportation: some insights from the European experience. *IATSS Res.* 362, 83–87.
- Meunier, D., Quinet, E., 2007. The contracting of investment and operation, and the management of infrastructure funding bodies. *Res. Transp. Econ.* 19, 81–109.
- Niskanen, W., 1971. *Bureaucracy and Representative Government*. Aldine, Chicago, IL.
- Nombela, G., de Rus, G., 2004. Flexible-term contracts for road franchising. *Transp. Res.* 38, 163–179.
- Oum, T., Fu, X., 2008. Influence des aéroports sur la concurrence dans le transport aérien. Document de référence OCDE 2008e17.
- Partnership, P.P., 2017. Reference Guide. World Bank/PPIAF.
- Reeves, E., 2005. Public private partnerships in the Irish roads sector: an economic analysis in procurement and financing of motorways in Europe. *Res. Transp. Econ.* 15, 107–120.
- Sabot, P., Puentes, R., 2014. *Private Capital, Public Good: Drivers of Successful Infrastructure Public-Private Partnerships*. Brookings, Washington, DC.
- Thompson, L., 2001. European Railways as seen from the rest of the world. Dupuit Lecture, Paris, December 5, 2001.
- Vickers, J., Yarrow, G., 1988. *Privatization: An Economic Analysis*. The MIT Press, Cambridge, MA.
- Williamson, O., 2002. The theory of the firm as governance structure: from choice to contract. *J. Econ. Perspect.* 16 (3), 171–195.

Further Reading

- Alexander, I., Estache, A., Oliveri, A., 2000. A few things transport regulators should know about risk and the cost of capital. *Utilities Policy* 9, 1e13.
- Boeuf, P., 2003. Public-private partnerships for transport infrastructure projects. Transport infrastructure development for a wider Europe seminar Paris.

- Different (User reaction and efficient differentiation of charges and tolls) Deliverable D8.3/D9.2 Report on motorway toll differentiation, 2008. Project research co-funded by the European Commission within the Sixth framework programme.
- Oum, T., Yu, C., 1994. Economic efficiency of railways and implications for public policy. A comparative study of the OECD countries' railways. *J. Transp. Econ. Policy* 28, 121–138.
- Perelman, S., Pestieau, P., 1988. Technical performance in public enterprises: a comparative study of railways and postal services. *Eur. Econ. Rev.* 32, 432–441.
- Pestieau, P., Tulkens, H., 1990. Assessing the performance of public sector activities: some recent evidence from the productivity efficiency viewpoint, Nov 1990. University of Liège. Programme d'actions de recherches concertées no 87-92/106 with CORE and no 90-94/141 with Université de Liège.
- Reis, R.F., Sarmiento, J.M., 2017. Cutting costs to the bone: the Portuguese experience in renegotiating public private partnerships highways during the financial crisis. *Transportation* 46 (1), 285–302.

Airline Economics

Anming Zhang*, Yahua Zhang[†], Zhibin Huang[†], *Sauder School of Business, University of British Columbia, Vancouver, Canada; [†]School of Commerce, University of Southern Queensland, Toowoomba, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Airline Demand	392
Airline Costs and Production	393
Airline Pricing	393
Concluding Remarks	395
References	395

Airline Demand

The model of supply and demand explains how prices are determined in a market system. It is important for an airline to understand the market demand and the supply for its own product in order to make informed strategic and operational decisions. The demand for an airline's services is influenced by the price it charges, the prices charged by its competitors, the availability of other transport modes, its service quality such as frequency, and in-flight services, its safety record, passengers' income and preferences, macroeconomic conditions, and demand shocks (Vasign et al., 2018). A change in one or more of these factors (except its own price) will shift the demand curve and lead to a new equilibrium price. In the following, some of the factors will be briefly discussed.

First, the demand for air transport is highly responsive to price. Most passengers, even some of the business travelers, are sensitive to airfares (Zhang, 2012). More and more companies have issued the travel policies to control business travel expenses. Previous demand estimations for the developed economies show that the values of air travel price elasticities range from -0.8 to -2.0 (Oum et al., 1993). A recent study on the China and India aviation markets finds that the demand elasticity is about -2.6 in India while is about -1.2 for China (Wang et al., 2018a). This suggests that the major Indian routes have a very price-elastic demand compared with other aviation markets in the world.

Income is one of the determinants of the demand for air services, which is usually proxied by gross domestic product (GDP) per capita. Air service is a normal good and thus the higher income, the higher air travel demand. Studies have shown that demand elasticity with respect to income is close to 2 (Garín-Muñoz, 2006): for instance, if the economy grows by 5%, then the air travel demand would grow by 10%; likewise, if the economy contracts by 2%, the demand would shrink by 4%. This pro-cyclical behavior has important implications for airlines' decisions on route entry and exit, and on aircraft order, purchase, and lease (Zhang and Zhang, 2018).

In addition to price and income, air travel demand is affected by the availability of other modes of transport. Since the new century, high-speed rail (HSR) has emerged as a significant transport mode in Europe and China after its success in Japan for several decades (Zhu et al., 2019). The HSR service has been shown to be a good substitute for air on short/medium-haul routes, sometimes even on long-haul routes. Here, the modal choice is based on the full opportunity cost, that is, the sum of ticket price and journey time cost. The total journey time consists of access time, travel time, and expected "schedule delay" (Behrens and Pels, 2012). For a given origin-destination (O-D) pair, while air mode has a shorter travel time than HSR, it has longer access time including the time of accessing to (and egressing from) the airport/train station, the time spent at the terminals, and the takeoff/landing time (Zhang and Zhang, 2018). Higher frequency will reduce the cost of schedule delay, a measure depicting the deviations from a passenger's preferred time of travel, and increase the visibility of a flight option (Douglas and Miller, 1974). More frequent services reduce expected schedule delays and hence is a more convenient service, which in turn increases demand.

Demand for air travel is also related to distance. The longer the travel distance involved, the fewer trips will be made. This is sometimes referred to as the gravity of law of travel demand: demand falls with the square of the distance between origin and destination, similar to the gravity law in physics. Much of the airline literature on O-D bilateral air traffic flows employs a gravity model. However, unlike the gravity models in the international-trade literature where the distance variable is consistently statistically significant, distance does not seem to be a good proxy for bilateral cost after airline deregulation (or liberalization) that first started in the United States in 1978 and then spread to other countries, especially in the markets where low-cost carriers (LCCs) are present (Zhang et al., 2018a). The average airfares have been declining in the last 30 years as a result of strong competition in the airline industry. Discounted airfares are prevalent on short-distance and particularly on long-distance routes. Therefore, geographical distance is no longer a good proxy for travel cost. That distance plays a smaller role in air transport may also be due to other reasons. First, on short-distance routes (say, less than 300–400 km) few air trips are made, owing to an increased cost/time ratio relative to alternative modes such as HSR or automobiles. Second, very long distance eliminates alternative transport modes (such as HSR and cars), which may increase the demand for air services (Zhang and Zhang, 2018).

Overall aviation safety across the world has steadily improved over the last three decades. However, in 2018 there were 15 crashes causing 556 fatalities, compared with 44 deaths in 10 accidents in 2017. Safety and security are top priorities in airline management

since both are important determinants for air travel demand. Air travel drops whenever there is a major air disaster, and so there has long been a major emphasis on safety and security regulation in aviation. Safety regulation involves policies and procedures to minimize the risks of failures in the design and operation of airlines and airports. Security concerns deliberate sabotage or other acts to cause harm to air operations and the traveling public. Since September 11, 2001, security has become the highest priority of the global aviation industry and many governments. However, increases in safety and security can only be justified when the benefits are greater than the costs. The full benefits are difficult to quantify in reality and research into the economics of safety and security is still rare in the airline literature (Vasigh, et al., 2018).

Airline Costs and Production

While consumers form one side of the market—the demand side, airlines form the supply side. As with other industries, to provide air transportation services airlines use various inputs, such as labor, capital, fuel, and materials. Therefore, the main determinants of the supply of air services include ticket price, the prices of inputs such as fuel and labor, airport charges, aircraft costs, maintenance costs, government regulations, and technology (Vasigh et al., 2018). Random factors such as weather and strikes also affect the provision of air services.

Fuel cost is the largest cost component for most airlines, which accounts for around 20%–50% of the total costs depending on the type of airlines and time (Koopmans and Lieshout, 2016). A flight sector comprises the ascent and decent phases as well as the cruise phase. During the cruise, the fuel cost is proportional to the route distance. Fuel costs can also differ substantially from one country to another, due mainly to the differences in taxes and transport costs. Airlines pay fuel prices at airports they serve, and thus the average price paid by a carrier is roughly a weighted average of fuel prices of the destinations it serves. Although the fuel price is largely out of an airline's control, some airlines are keen to use more fuel-efficient aircraft to save costs by purchasing (or leasing) new planes. Other airlines choose to replace the seats, television monitors, and even the beverage carts with newer and lighter versions (FAA, 2011). Aviation fuel is a major contributor to emissions. Interestingly, in the EU, aviation fuel has long been tax exempt and aviation fuel tax has been called for to help Europe to meet its climate targets.

As fuel cost has replaced labor cost as the largest cost component for many airlines, fuel hedging has been frequently used to reduce airline exposure to unexpected changes in fuel price. This is because these days it is difficult for airlines to pass the rise in fuel cost on to consumers in a deregulated and competitive operating environment. Interestingly, however, the benefit of fuel hedging appears controversial. Some studies report a positive impact of fuel hedging on the airlines' firm value (Sturm, 2009), while some other studies claim that hedging does not necessarily affect a firm's market value (Jin and Jorion, 2004), and that sometimes it could have a negative but insignificant impact on operating costs (Lim and Hong, 2014). This means that the actual benefit of fuel heading could be negligible. Therefore, it seems that merely engaging in fuel hedging is not sufficient to achieve cost reduction.

In the last two decades, the legacy airlines have managed to reduce their labor costs through various ways. Declaring bankruptcy is one way through which the contract conditions can be changed and the benefits and pensions can be reduced. American Airlines, Delta, Northwest, United, US Airways, and Continental all used this way to reduce the labor costs. Some of them went through bankruptcy twice. Qantas suffered huge financial losses from 2011 to 2014 and it announced to cut 5000 jobs, 15% of its workforce in 2014 as well as other cost control measures. In 2015, Qantas returned to profits. It should be noted that LCCs do not necessarily pay lower wages than full services carriers do, but they may require staff to perform multiple tasks to save costs (Vasigh, et al., 2018).

Airlines operate in a dynamic environment with a great number of uncertainties, and with airline revenues and costs being influenced heavily by overall economic activities. An important financial decision is the decision to buy or lease aircraft. Operating lease and direct purchase with bank finance are two commonly used methods of financing the aircraft acquisition. Should an airline purchase and own the aircraft? If purchased, should the airline pay 100% cash or borrow some funds for the purchase? These decisions depend on the market conditions and the airline's financial conditions (Accession Capital Corp, 2003). There is no one method that is superior to another. Interest rates, discount rates, lease rates, and depreciation are relevant costs and should be taken into account when making these decisions (Zhang and Zhang, 2018). Today, around 40% of the worldwide fleet is owned by leasing companies. The share of leased aircraft was only 20% in 2000 and is expected to reach 50% in 2020.

It is believed that airline cost per seat declines with the size of aircraft (up to some threshold). Furthermore, the cost per passenger falls as the percent of seats sold on a flight (the "load factor") rises, since much of a flight's cost is fixed regardless of the number of passengers flown (Zhang and Zhang, 2018). These relationships do not necessarily imply the existence of economies of scale in airline operations, however. In fact, some researchers claim that economies of scale are negligible or non-existent (White, 1979). This is because adding or dropping cities from an airline's network does not raise or lower unit cost. In contrast, sizeable economies of traffic density seem to exist up to fairly large volumes of traffic (Caves et al., 1984). This means that the airline industry is more likely to exhibit economies of density instead of economies of scale, and that the cost savings can be achieved through the intensive and efficient use of larger aircraft with higher load factors (Brueckner and Spiller, 1994).

Airline Pricing

For most airline markets, there are only a few airlines competing against each other. Therefore, airline markets are typically oligopolistic (Levine, 1987; Borenstein, 1992). Numerous theoretical and empirical studies have attempted to examine the pattern

of airline pricing and price dispersion, many of which have found that price discrimination, new technology and innovation, the presence of LCCs, market structure, deregulation and airline merger and alliance are key determinants of the variations in airline pricing (Borenstein, 1985; Brueckner and Spiller, 1991; Berry, 1992; Brander and Zhang, 1990, 1993; Zhang et al., 2018b).

Price discrimination is traditionally seen as one of the sources of price variation, especially in the airline industry (Cross, 1995; McGill and van Ryzin, 1999). There are three types of price discrimination: first degree, second degree, and third degree. Under first-degree price discrimination, the entire consumer surplus is transferred to the firms, which is unlikely in reality, as airlines do not usually have the ability to determine a consumer's reservation price. Second-degree price discrimination is the practices of charging consumers difference prices for different quantities of the same product consumed. Probably a frequent flyer program may fall within this type of price discrimination as the points awarded to the passengers can be used later, which is a special kind of discount and consumers do not need to pay for the points at the time of purchasing the ticket. Third-degree price discrimination refers to the charge of difference prices to different consumer groups such as male and female, children and adults. Many airlines offer corporate travel programs to companies and give cheaper contract prices, which is a practice of third-degree price discrimination.

The emergence of new technology has enhanced price discrimination. Some researchers argue that airline pricing falls somewhere between first and second-degree price discrimination because of the use of computer reservation system (CRS), later renamed as Global Distribution System (GDS) (Kons, 1999). This is because the creation of a large number of prices possible by the GDS allows the airlines to acquire more consumer surplus than second-degree discrimination offers. The GDS provides virtual real-time connectivity between thousands of suppliers of travel inventory (e.g., airlines, hotels, car rental, and tour operators) and hundreds of thousands of retail sellers of travel products (Sismanidou et al., 2009). The system has significantly improved airlines efficiency in attracting and retaining customers and driving sales through travel agents. American Airlines was a pioneer in using a sophisticated yield management program to attract both the high yield and low yield passengers in the 1980s, which gave it huge competitive advantage in the battle against the newly established carriers such as People Express Airlines following airline deregulation (Smith et al., 1992; Cross, 1995). For example, the carrier used "Ultimate Super Saver" fares to match People Express's lowest prices (Vasigh et al., 2018). This price type required 21-day advance purchase restrictions in order to draw in only the most price-sensitive travelers. The airline also controlled the number of "Ultimate Super Saver" fares available on each flight in order to make seats available to high-yield passengers (Cross, 1995). By this way, the airline was able to generate profit from both low-revenue and high-revenue passengers. In contrast, People Express could only capture low-revenue passengers with its single fare class reservations system.

In the 1990s, airlines began to reduce the commission paid to travel agencies and the distribution system by increasing their direct sales through the Internet. The Internet enables airline products accessible 24/7 and allows the airlines, the distributors, and the customers to communicate and interact with each other globally and efficiently (Timmers, 2000). In particular, LCCs took advantage of the internet to developed simple distribution strategies and bypass travel agents and GDSs, which is a key factor for the success of this business model. For example, JetBlue and Southwest Airlines sell up to 90% of their tickets directly through their websites. China's Spring Airlines has tried many ways to save costs, including the use of its own CRS and encouraging online purchases. The Internet is now its main channel of distribution and passengers who purchase tickets through a ticketing office have to pay an extra purchase fee. Its average ticket price was about 30% lower than the industry average price. With the advantage of low price ticket, Spring Airlines maintained a high load factor of about 95% through offering low airfares, well above the industry average of 70% (Zhang and Lu, 2013).

The recent advance in new technologies of the artificial intelligence, big data, and virtual reality can lead to a revolution in airline pricing and personalization concepts (Krämer et al., 2018). Dynamic pricing as a result of the capability of analyzing big data allows airlines to predict with a high degree of certainty customer demand and preferences, which makes it possible for the airlines to offer a fare that is closer to the customers' demand (Hammou et al., 2018). Based on the customers' browsing pattern, the artificial intelligence system tracks the changing volume of search requests from the airline website or other search engines. The airlines can make prompt changes to the airfare, if needed.

Airline pricing is also associated with market concentration. An increase in concentration (e.g., as a result of airline mergers and alliances) may confer the airlines with market power (Zhang and Round, 2009). Following airline deregulation there was a wave of airline mergers in the 1980s that were supported by the Reagan administration. In the 1990s, there was a proliferation of airline alliances, both international and domestic. Today there are three major global alliance groups—namely, Star Alliance, One World, and SkyTeam – make up over 60% of the world market in recent years. There are many reasons for airlines merge and form strategic alliances, including expansion of seamless service networks, traffic feeding between partners, cost efficiency (based mainly on economies of traffic density), quality improvement, various marketing advantages (e.g., frequent flyer programs, or FFP), and the advantage of market power and cooperative pricing (Oum et al., 2000). Although in many countries where antitrust laws and active enforcement have made mergers and alliances as an instrument for market power difficult, we cannot exclude market power motive as one of the important reasons pursued by airlines through airline mergers and alliances. Similarly, with the emergence of the hub-and-spoke system following deregulation in the United States, the price charged by the hub airports has been a much-debated topic, known as the "hub premium" debate (Lee and Luengo-prado, 2005). It has been found that airline yields at concentrated hubs are substantially higher than those non-hub airports, and that the dominant airlines' yields at the concentrated airports are even higher. This implies that on the one hand, cost savings and better service could be generated by the hub-and-spoke system. On the other hand, with the ability to block entry because of a dominant presence, routes associated with hub endpoint(s) are believed to have higher fares than otherwise would be the case.

In most textbooks, market power is defined as the ability of a firm or group of firms persistently to hold price above the competitive level in a profitable way without losing so many sales that the price level is unsustainable (Motta, 2004). However,

significant market power does not seem to exist in the airline market since deregulation that relaxed market entry and exit. It is a common hypothesis that the ease or difficulty of entry into an industry has a significant impact on airline pricing. The easier entry is into an industry, the more difficult it will be for collusion to be sustained, because high prices and profits will attract new firms and break the collusion. Thus entry plays a very important role in competition analysis (Shepherd, 1997).

In the airline industry, product differentiation helps incumbents lock up some consumers by establishing a reputation through their past performance, brand names and advertising, making it difficult for new entrants to attract customers (Shepherd, 1997). In this sense, product differentiation has endogenously become an entry-detering strategy. Predatory pricing is another common pricing strategic entry barrier in the airline industry. Predatory pricing occurs when a price war is waged by incumbents against new or potential entrants. The incumbent first lowers its price to drive out the rivals and then raises it. By this means, it sacrifices a short-run loss to protect its long-run interest (Carlton and Perloff, 1999).

In the 1980s, the contestable markets theory gained its popularity and many economists believe that the airlines market is contestable (Baumol et al., 1982). As airplanes can be easily shifted to other routes, entry into a market can be largely fully reversible. In addition, compared with its large fixed costs, the airline industry faces relatively low sunk costs along particular routes. Airline markets are therefore thought to be contestable. Once a market is contestable due to free entry and exit, potential competitors will become acceptable substitutes in consumers' minds for the incumbent competitors. Whenever they see any above-normal profits in the market, potential entrants would enter and stay if they could earn sufficient revenue to cover their fixed and variable costs. Alternatively, potential entrants could follow a hit-and-run strategy, given the assumptions that they can identify consumers for their products at or below the current market price, and that existing firms are unable to change their prices quickly in response to their entry. Given such a potential threat, the incumbents' pricing behavior would be restrained by the potential entry and thereby a purely competitive outcome would result.

However, the strong assumptions associated with contestable market theory have been criticized by many economists. First, sunk costs associated with entry are most likely to exist in all markets to some extent (Spence, 1983). Even in the airline industry, for example, an airline must advertise to let passengers know before opening a new route (Stiglitz et al., 1987). Every airline market manager knows that it will take some time and effort before a new route becomes profitable. Therefore in real life, perfectly contestable markets may not exist. Second, if the entry is on a small scale with no serious threat to the incumbents, incumbents do not necessarily change their market behavior (Shepherd, 1984). They can continue to charge higher prices without losing much share of the market, or choose to accommodate the entrants. Accordingly, the contestable markets theory has gradually lost its popularity since the 1990s and market structure still plays an important role in determining an airline's pricing.

Worldwide the LCCs have expanded rapidly and led to the declining in airfares (Zhang and Lu, 2013). The LCCs now carry about 30% of the total passengers. The threat of LCCs to the legacy airlines is obvious, and can be seen not only from the fall in hub premiums, but also from the response of the incumbents (most of which have established their own LCCs). The LCCs provide low fares, high frequency and point-to-point services. In the United States, the negative effect of an LCC is present not only on the fares of the routes they are operating on, but also on the fares of adjacent competitive routes because of spillover effects (Morrison, 2001). In Australia, it has also been noticed by researchers that the emergence of LCCs not only had the apparent effect of lowering the price charged by the incumbent airlines, but also had profound effects on the incumbents' performance, labor arrangements and costs and productivity, shifting towards a socially desirable direction (Forsyth, 2003). The "Southwest effect" or "Ryanair effect" is also felt in the Asian market. In India, the market share of LCCs have been above 60% for many years, which is a remarkable achievement that makes flying more affordable for low-income consumers (Wang et al., 2018b). It has been widely agreed that the presence of LCCs represents an opportunity to develop the local economy by creating new employment and contributing to the local tourism.

Concluding Remarks

This chapter has briefly introduced the factors influencing the demand and supply in the airline market. The theories and practices of airline competition and pricing are also presented. However, it should be noted that with the increasing penetration of internet as a distribution channel, airline pricing has become more transparent, making it easier for consumers to compare prices across multiple competitors and tailor travel plans to take advantage of lower fares. Such transparency will undoubtedly make the demand for each airline more elastic. Thus there is a need for airlines to develop non-price competition strategies to retain customers and a certain degree of market power. Frequent flyer programs are an example, which is a loyalty program with an aim to maintain loyalty among those who travel frequently by rewarding them with free upgrades, free tickets, additional baggage allowances, and business lounge access (Jiang and Zhang, 2016). The FFP membership has played a strong role in airline and itinerary choice, especially for passengers with elite membership. The non-pricing competition also occurs in pre-flight, in-flight, and post-flight services, which constitutes a significant research area in the ever increasingly competitive airline market.

References

- Accession Capital Corp., 2003. Commercial aircraft: lease, finance or purchase? Available from: www.connvaluation.com/caseStudies/Lease_Borrow_Purchase.pdf.
 Baumol, W.J., Panzar, J.C., Willig, R.D., 1982. *Contestable Markets and the Theory of Industry Structure*. Harcourt Brace Jovanovich, San Diego.
 Behrens, C., Pels, E., 2012. Intermodal competition in the London-Paris passenger market: high-speed rail and air transport. *J. Urban Econ* 71 (3), 278–288.

- Berry, S.T., 1992. Estimation of a model of entry in the airline industry. *Econometrica* 60 (4), 889–917.
- Borenstein, S., 1985. Price discrimination in free-entry markets. *Rand J. Econ* 16, 380–397.
- Borenstein, S., 1992. The evolution of U.S. airline competition. *J. Econ. Persp* 6, 45–73.
- Brander, J.A., Zhang, A., 1990. Market conduct in the airline industry. *Rand J. Econ* 21, 567–583.
- Brander, J.A., Zhang, A., 1993. Dynamic behaviour in the airline industry. *Int. J. Ind. Org* 11, 407–435.
- Brueckner, J.K., Spiller, P.T., 1991. Competition and mergers in airline networks. *Int. J. Ind. Org* 9, 323–342.
- Brueckner, J.K., Spiller, P.T., 1994. Economies of traffic density in the deregulated airline industry. *J. Law Econ* 37, 379–415.
- Caves, D.W., Christensen, L.R., Tretheway, M.W., 1984. Economies of density versus economies of scale: why trunk and local service airline costs differ. *Rand J. Econ* 15, 471–489.
- Carlton, D.W., Perloff, J.M., 1999. *Modern Industrial Organization*, third ed. Addison-Wesley-Longman, Boston, MA.
- Cross, R.G., 1995. An introduction to revenue management. In: D., Jenkins, (Ed.), *The Handbook of Airline Economics*. The Aviation Weekly Group of the McGraw-Hill Companies, New York, NY, 443–468.
- Douglas, D.W., Miller, J.C., 1974. Quality competition, industry equilibrium, and efficiency in the price-constrained airline market. *Am. Econ. Rev* 64 (Sept.), 657–669.
- FAA, 2011. *The Economic Impact of Civil Aviation on the US Economy*. Federal Aviation Administration. US Department of Transportation, Washington, DC.
- Forsyth, P., 2003. Low-cost carriers in Australia: experiences and impacts. *J. Air Transp. Manag* 9 (5), 277–284.
- Garín-Mun, T., 2006. Inbound international tourism to Canary Islands: a dynamic panel data model. *Tour. Manag* 27 (2), 281–291.
- Hammoud, G.A., Tawfik, H.F., Fahmy, R.S., 2018. Development of airlines' distribution capabilities. *J. Tour. Hosp. Manag* 6 (1), 66–80.
- Jiang, H., Zhang, Y., 2016. An investigation of service quality, customer satisfaction and loyalty in China's airline market. *J. Air Transp. Manag* 57, 80–88.
- Jin, Y., Jorion, P., 2004. Firm value and hedging: evidence from US oil and gas producers. *J. Financ.* 61 (2), 893–917.
- Koopmans, C., Lieshout, R., 2016. Airline cost changes: to what extent are they passed through to the passenger? *J. Air Transp. Manag* 53, 1–11.
- Kons, A., 1999. Understanding the Chaos of Airline Pricing. *The Park Place Economist* 8, 15–29.
- Krämer, A., Friesen, M., Shelton, T., 2018. Are airline passengers ready for personalized dynamic pricing? A study of German consumers. *J. Rev. Pricing Manag* 17 (2), 115–120.
- Lee, D., Luengo-Prado, M.J., 2005. The impact of passenger mix on reported "hub premiums" in the US airline industry. *South. Econ. J* 72 (2), 372–394.
- Levine, M.E., 1987. Airline competition in deregulated markets: theory, firm strategies and public policy. *Yale J. Regul* 4, 283–344.
- Lim, S.H., Hong, Y., 2014. Fuel hedging and airline operating costs. *J. Air Transp. Manag* 36, 33–40.
- McGill, J.I., van Ryzin, G.J., 1999. Revenue management: research overview and prospects. *Transp. Sci* 33, 233–256.
- Morrison, S.A., 2001. Actual, adjacent, and potential competition: estimating the full effect of Southwest Airlines. *J. Transp. Econ. Policy* 35 (2), 239–256.
- Motta, M., 2004. *Competition Policy: Theory and Practice*. Cambridge University Press.
- Oum, T.H., Zhang, A., Zhang, Y., 1993. Inter-firm rivalry and firm-specific price elasticities in deregulated airline markets. *J. Transp. Econ. Policy* 171–192.
- Oum, T.H., Park, J.H., Zhang, A., 2000. *Globalization and Strategic Alliances: The Case of the Airline Industry*. Pergamon, New York.
- Shepherd, W.G., 1984. Contestability vs. Competition. *Am. Econ. Rev* 74 (4), 572–587.
- Shepherd, W.G., 1997. *The Economics of Industrial Organisation*, fourth ed. Prentice Hall, London.
- Sismanidou, A., Palacios, M., Tafur, J., 2009. Progress in airline distribution systems: the threat of new entrants to incumbent players. *J. Ind. Eng. Manag* 2 (1), 251–272.
- Smith, B., Leimkuhler, J., Darrow, R., 1992. Yield management at American Airlines. *Interfaces* 22, 8–31.
- Spence, M., 1983. Contestable markets and the theory of industry structure: a review article. *J. Econ. Lit* 21 (3), 981–990.
- Stiglitz, J.E., McFadden, D., Peltzman, S., 1987. Technological change, sunk costs, and competition. *BPEA* 1987 (3), 883–947.
- Sturm, R.R., 2009. Can selective hedging add value to airlines? The case of crude oil futures. *Int. Rev. Appl. Financ. Issues Econ.* 1, 130–146.
- Timmers, P., 2000. *Electronic Commerce Strategies and Models for Business-to-Business Trading*. John Wiley & Sons, Chichester.
- Vasigh, B., Fleming, K., Tacker, T., 2018. *Introduction to Air Transport Economics: From Theory to Application*. Routledge, UK.
- Wang, K., Zhang, A., Zhang, Y., 2018a. Key determinants of airline pricing and air travel demand in China and India: policy, ownership, and LCC competition. *Transp. Policy* 63, 80–89.
- Wang, K., Xia, W., Zhang, A., Zhang, Q., 2018b. Effects of train speed on airline demand and price: theory and empirical evidence from a natural experiment. *Transp. Res. Part B* 114, 99–130.
- White, L.J., 1979. Economies of scale and the question of 'natural monopoly' in the airline industry. *J. Air Law Comm* 44, 545–573.
- Zhang, Y., 2012. Are Chinese passengers willing to pay more for better air services? *J. Air Transp. Manag* 25, 5–7.
- Zhang, Y., Lu, Z., 2013. Low cost carriers in China and its contribution to passenger traffic flow. *J. China Tour. Res* 9 (2), 207–217.
- Zhang, A., Zhang, Y., 2014. Airline economics and finance. In: Halpern, N., Graham, A., (Eds.), *The Routledge Companion to Air Transport Management*. Routledge, UK (Chapter 4), pp. 171–188.
- Zhang, Y., Round, D.K., 2009. Policy implications of the effects of concentration and multimarket contact in China's airline market. *Rev. Ind. Org* 34 (4), 307–326.
- Zhang, Y., Lin, F., Zhang, A., 2018a. Gravity model in air transport: a survey and an application. In: Blonigen, B.A., Wilson, W.W., (Eds.), *Handbook of International Trade and Transportation*. Edward Elgar, Northampton, MA (Chapter 4), pp. 141–158.
- Zhang, Y., Sampaio, B., Fu, X., Huang, Z., 2018b. Pricing dynamics between airline groups with dual-brand services: the case of the Australian domestic market. *Transp. Res. Part A* 112, 46–59.
- Zhu, Z., Zhang, A., Zhang, Y., 2019. Measuring multi-modal connections and connectivity radiations of transport infrastructure in China. *Transp. Sci* 15 (2), 1762–1790.

Privatization and Deregulation of the Airline Industry

Achim I. Czerny, Hao Lang, Department of Logistics and Maritime Studies, Hong Kong Polytechnic University, Hong Kong, China

© 2021 Elsevier Ltd. All rights reserved.

Introduction	397
Deregulation of the Airline Industry	397
Deregulation of Regional Airline Markets	397
United States	397
European Union	398
China	398
ASEAN	399
Africa	399
India	400
Open Skies Agreements Between Regional Airline Markets	400
Europe–United States	400
Other Open Skies Agreements	400
Airline Privatization	400
Conclusions	401
References	401

Introduction

The aviation industry is a fast-growing industry that doubles market size every 15 years (Airbus, 2018). This growth is largely determined by changes in government regulations and the present text describes how the privatization and deregulation of airline markets has contributed to this development.

We distinguish between (1) air services in regional markets and (2) air services between or among regional markets. Examples for air services in regional markets are air services in the United States (US), the countries of the European Union (EU), China, India, the countries of Africa, and the South East Asian Nations (ASEAN). These examples show that regional markets may be represented by a single country or by multiple countries. This observation is relevant because the deregulation of regional markets may involve national law only as in the cases of the United States, China, and India or changes in bilateral agreements between countries, which typically impose operational and ownership restriction on airlines. The system of bilateral air service agreements is based on the Chicago Convention of 1944. Over 3000 such bilateral agreements existed in 2007 (IATA, 2007).

The following describes the deregulation of the airline industry in regional markets and between or among regional markets where the latter are called “open skies agreements.” The privatization of airline companies is considered subsequently. The final part provides conclusions.

Deregulation of the Airline Industry

The world’s top three aviation markets in 2017 involved domestic flights in the United States, intra-Western European flights in the European Union, and domestic flights in China (Airbus, 2018). In the first step, this section explains deregulation efforts in these regional markets as well as, more briefly, deregulation efforts in the ASEAN countries, Africa, and India. In the second step, open skies agreements are considered.

Deregulation of Regional Airline Markets

United States

Aviation deregulation in the United States started with the Airline Deregulation Act of 1978. This act responded to the, at the time, growing concerns that industry regulation generally harms efficiency and innovation. Reforms were inspired by the contestability theory, which highlighted the possibility that potential market entry could exert competitive pressures even in the absence of *actual* firm rivalry (Bailey, 2002). The contestability theory seemed highly relevant for the airline industry because airplanes can easily be moved from market to market and were even qualified as “marginal costs with wings” by Alfred E. Kahn, who was one of the deregulators in charge (Bailey and Baumol, 1984). Furthermore, empirical studies indicated that airline businesses operate under economies of traffic density (Caves et al., 1984) and moderate economies of scale (Johnston and Ozment, 2013), which supports market concentration and highlights the role of competition by potential market entry.

Kahn (1988) concluded that the expectations that deregulation reduces fares and increases price-quality options as well as efficiencies had been fully vindicated. His findings have later been supported by the investigations of Morrison and Winston (1990). Their analysis revealed that fares were consistently lower in the past decade of deregulation than they would have been if the industry were still regulated. Later studies highlighted the negative effects of the deregulation. Goetz and Vowles (2009) found that deregulation had contributed to the financial and employment instability including the bankruptcy of airlines, reduced overall service quality, fewer flights, and higher fares to smaller places.

The Airline Deregulation Act of 1978 allowed airlines to freely reconfigure their networks. McShan and Windle (1989) proposed a method to quantify the change in hub-and-spoke routings, where passengers change aircrafts at hub airports before they can reach their final destinations. Hub-and-spoke networks have already been used before deregulation (Kanafani and Ghobrial, 1985). However, McShan and Windle (1989) found that the use of hub-and-spoke networks had increased by almost 50% between 1977 and 1984, that was, directly after deregulation.

The contestability theory had been used to explain the effects of deregulation. Baik et al. (2011) analyzed the effect of deregulation on management behavior. They found that airline managers were inclined to engage in income-increasing earnings management after deregulation.

A popular belief at the time was that economic deregulation and increased competitive pressures could lead to reduced expenditure on safety related items (Moses and Savage, 1990). By contrast, Kahn (1988) found that accident rates had been reduced by 35% in the period after deregulation. This finding was supported by Moses and Savage (1990), who found no evidence that safety performance had been impaired by deregulation. However, they emphasized that government actions are required to ensure safety in a dynamically growing aviation market.

European Union

The deregulation of the aviation industry in the European Union proceeded in three packages. These packages were implemented between 1987 and 1997 and led to a situation where every EU airline could basically access the whole EU aviation market. Button and Swann (1992) and Button (1996) asked whether the US lessons from aviation market deregulation can be transferred to the European Union. Button (1996) identified a set of differences between the US and the EU aviation markets at the time of deregulation in the European Union:

1. Charter: One can distinguish between scheduled and non-scheduled charter markets. The former serves leisure and business passengers, while the latter typically targets leisure passengers. Charter airlines were almost non-existent in the United States, while they have been common in the European Union.
2. Size: The geographical conditions in the United States and the European Union imply that average route lengths in the United States are longer than the average route lengths in the European Union, and airlines in the United States were generally bigger than their EU counterparts.
3. Ownership: US airlines were under private ownership, while government airline ownership was common in Europe.
4. Intermodal competition: High-speed rail was absent in the United States, while airline services and high-speed rail services provided comparable services for many trips in the European Union.

Despite these differences between the US and EU markets and despite their expectation that the European Union might not be as deregulated as the US markets, Berechman and de Wit (1996) still expected that the EU deregulation would ease market entry and exit, leading to more competitive choices of fares and service quality, and cause increase in hub-and-spoke operations. Their simulation models indicated that London Heathrow would become the dominant West European gateway hub, which appeared to be true.

Later studies showed that the deregulation increased both the number and frequency of routes, lowered fares, and pushed the use of hub-and-spoke airline networks in EU markets including the Central and Eastern European markets (e.g., Dobruskes, 2009; Derudder and Witlox, 2009; Burghouwt and de Wit, 2015; Jankiewicz and Huderek-Glapska, 2016; Lieshout et al., 2016). These developments were partly driven by charter operators that started adopting the low-cost business model (Burghouwt and de Wit, 2015). But, these studies also pointed out that deregulation created many new air routes operated by a single airline with potential market power, and that the main beneficiaries were the large cities and peripheral regions where deregulation helped in boosting tourist flows (Dobruskes, 2009). Lieshout et al. (2016) found that deregulation had a stronger effect on more remote regions in the United Kingdom, Spain, and Italy and that airline competition was still lagging behind in Germany, large parts of Scandinavia, France, and Spain.

China

Dating back to the Spring of 1950, China's civil aviation industry was established, and the Civil Aviation Bureau was formed after the country was newly founded in 1949 (Le, 1997). Between the 1950s and 1980s, China's economy, including civil aviation, was isolated from the rest of the world. During that time, the airline industry was considered as a semi-military organization with limited commercial operations under the leadership of the air force and the State Council (CAAC, 2002). The Civil Aviation Administration of China (CAAC), was at the top of the four-tier administration system operating civil aviation in China along with six regional civil aviation bureaus, 23 provincial civil aviation bureaus and 78 civil aviation stations (Zhang, 1998; Zhang and Chen, 2003; Zhang and Round, 2008). The CAAC was both a regulator and a manager of air transport services, and strictly regulated the industry with respect

to, for example, market entry, pricing and even passenger eligibility for air travel. Thus, a CAAC monopoly best describes the China's aviation industry prior to the 1980s (Le, 1997; Zhang, 1998; Lei and O'Connell, 2011).

The reform on civil aviation industry came later in the late 1970s echoing with the national opening-up policy. China was gradually moving from centralized planning to a market-oriented economy. The CAAC became independent from the air force. The six regional civil aviation bureaus became "half corporatized" in the sense that they were required to be financially independent (Zhang, 1998). To encourage operating efficiency and profitability, in 1987, the State Council ratified the "Report on Civil Aviation Reform Measures and Implementation." Consequently, between 1987 and 1991, six trunk airlines based in the different regional capital cities emerged. However, all the airlines, regional and trunk ones, were still tightly regulated by the CAAC and only offered standard, indistinguishable air services with little competition. The first noticeable deregulation in airfares occurred in 1997 when the CAAC encouraged airlines to adopt price discrimination in an attempt to attract more passengers. Airlines undercut the tickets prices without considering appropriately the costs in response to this policy change, which resulted in a heavy loss for the whole industry (e.g., Zhang and Round, 2008; Lei and O'Connell, 2011).

Later in 1998, the CAAC issued "The Decision to Enforce Supervision and Restore Order in the Air Market," in which it tried to regulate the airfares by limiting the discounts granted by airlines to passengers. But, the CAAC did not specify any penalties associated with violating these rules, which meant that airlines continued their "price wars." For understandable reasons, passengers also opposed any regulation on the prices (Li, 2001).

Ever since China joined the World Trade Organization in 2001 and embraced the market economy, China's civil aviation industry enjoyed a less regulated environment since 2000 with noticeable deregulation on the airline entry and exit in the domestic market. For instance, airlines could establish new routes without prior CAAC approval under certain criteria. New pricing rules were established where the government determined benchmark prices and floating bands, while airlines determined their prices within the specified ranges (Zhang et al., 2014). As a result of the deregulatory efforts, the first low-cost airlines of China, Spring Airlines, was established in Shanghai, which further increased the competitive pressure on the established full-service airlines (Fu et al., 2015; Jiang et al., 2017).

Altogether, the Chinese airline industry has been significantly deregulated; but it can still be characterized as mainly state-led rather than market-led. The low-cost carrier sector seems still under-developed (Fu et al., 2010, 2015). Further deregulation of, for example, aircraft purchase, and a transparent system of slot allocation and pilot recruitment, is recommended by researchers (Shaw et al., 2009; Fu et al., 2015; Wang et al., 2016).

ASEAN

Prior to 1995, air transport had been regulated on the basis of bilateral agreements in South East Asia (Findlay and Forsyth, 1992; Laplace and Latge-Roucolle, 2016). ASEAN took an initiative towards a more liberalized air traffic market as early as 1995 in the ASEAN leaders' summit held in Bangkok, Thailand (Forsyth et al., 2006; Hanaoka et al., 2014; Laplace and Latge-Roucolle, 2016; Tan, 2010). Over the next decade, a broader discussion about a greater economic integration across all sectors has taken place (Tan, 2010). One of the milestones for liberalization of ASEAN air traffic market was defined in November 2004 when the 10th ASEAN air transport ministers' meeting in Phnom Penh was held. In this meeting, the Action Plan for ASEAN Air Transport Integration and Liberalization 2005–15 was adopted (ASEAN, 2004). Together with the roadmap for integration of air travel sector, the action plan outlined the targets for an "open skies" regime by 2015. The idea was to establish unlimited third and fourth freedom flights for all designated points within ASEAN sub-regions by 2005 and unlimited fifth freedom rights by 2006 (Tan, 2010). Other agreements were signed in 2009 and 2010 with the goal to relax market entry and airline ownership restrictions and to establish a common policy for user charges, capacity, etc. (Hanaoka et al., 2014). However, not all targets were achieved due to failures of ratifying and accepting the agreements on a national level (Hanaoka et al., 2014). It appears that deregulation has not yet achieved its full potential in ASEAN (Findlay and Forsyth, 1992; Laplace and Latge-Roucolle, 2016; Tan, 2010, 2013).

Africa

In the late 1970s, many African governments started to negotiate open-skies agreements at all levels (Doganis, 2002). This led to the Yamoussoukro Declaration in 1988, which aimed at airline cooperation, integration, and air service liberalization in Africa. Later in 1997, the Banjul Accord was achieved to accelerate the implementation of the Yamoussoukro Declaration among six West African countries. The most fundamental move toward liberalization came with the Yamoussoukro Decision, which was signed in 1999. The Yamoussoukro Decision was supposed to progressively eliminate all non-physical barriers relating to the granting of traffic rights, particularly fifth traffic rights, aircraft capacity, tariff regulation, designation of airlines, and air freight operations (UNECA, 2002). Regions that implemented the Yamoussoukro Decision have enjoyed an increase in frequencies, air traffic, and aircraft movements as well as increased competition, which led to the improvement of the air services quality (Njoya, 2016). Yet, the implementation of Yamoussoukro Decision is incomplete because not all member countries accepted it (Njoya, 2016). Another main reason is that the institutional and legal frameworks required for the implementation of the Yamoussoukro Decision have not been put in place (Abate, 2016). The importance and benefit of greater liberalization has been widely recognized and it has therefore been suggested that African countries should work together toward a more deregulated aviation market (Abate, 2016; Mills and Membreno, 2007; Mills and Swantner, 2008; Myburgh et al., 2006; Njoya, 2016; Schlumberger, 2010; Surovitskikh and Lubbe, 2015).

India

China and India share many similarities because of the extraordinary sizes in terms of population and geography, and they both represent fast-growing economies. Their air transport sectors, however, exhibit substantial differences because private and low-cost airlines are the dominant players in India while the state-owned full-service airlines are dominant in China (Wang et al., 2018a). While India is likely to become the third largest aviation market in the world soon (after China and the United States), infrastructure development is considered the major challenge (IATA, 2018).

Open Skies Agreements Between Regional Airline Markets

The term “Open Skies” was introduced by Alfred E. Kahn to describe the objectives to be pursued after deregulating the US domestic market (Button, 2009). Winston and Yan (2015) estimated huge gains from US negotiated agreements for travelers, and they predicted further huge gains that could be achieved with additional open skies agreements. Based on a data set covering Canada’s bilateral service export and import data, Oum et al. (2019) found that the effect of open skies agreements on service trade, however, can be positive or negative. No strong link between open skies and tourist flows was found by Dobruszkes et al. (2016).

The following discusses specific open skies agreements in more detail.

Europe–United States

The European Union–United States open skies initiative can be considered as one of the most significant international airline deregulation efforts worldwide (Button, 2009). From 1992 onwards, a series of bilateral open skies agreements between the EU countries and the United States were established. These agreements removed restriction on fares and service levels and allowed airlines to fly between countries if the flight originates or terminates in the airline’s home country (fifth airline freedom). They were replaced by a new transatlantic agreement in 2007 (Hunnicutt, 2010).

The short-term effects of the new transatlantic agreement were changes in airline service offerings, particularly at London Heathrow airport (Pitfield, 2009). In a later study, Cosmas et al. (2010) found that these changes involved both increase in service levels for some countries and reductions in service levels for other countries. That the new (de)regulatory environment might have brought service levels down was further confirmed by the findings of Morandi et al. (2014). According to their study, direct transatlantic connections and served airport pairs have decreased after the signing of the new agreement. Zhang et al. (2018) added to these results with their finding that the US airlines are more likely to take advantage of the possibilities provided by open skies agreements. One explanation is provided by Morrison and de Wit (2019), whose study revealed that financial aid and policies have advantaged US airlines. They expressed their concerns that markets have been made more unleveled during the era of open skies agreements.

Other Open Skies Agreements

The European Union has concluded horizontal agreements where EU carriers can operate routes between any EU country and a third country with several African countries. These include the members of the West African Economic and Monetary Union, Morocco, Cape Verde, and Tunisia (Njoya et al., 2018). The European Union–Morocco open skies agreement, which was signed in December 2006, has been studied by Bernardo and Fageda (2017). Their econometric analysis revealed a 20%–35% increase in the number of seats offered on pre-existing routes and a notable increase in the number of new routes offered. These results are in line with Njoya et al.’s (2018) results. They found that the horizontal agreements between the European Union and African countries substantially increased service levels and lowered fares.

The China–US market has been studied by Lei et al. (2016). In 2004 and 2007, China and the United States negotiated air service arrangement protocols which had profound impacts on the China–US market. These protocols intensified competition, improved the Chinese airlines’ operating and financial performance and strongly increased traffic growth.

A recent study by Wang et al. (2018b) considered deregulation between China and Central-Asian countries in the context of the Belt and Road initiative of the Chinese government. The investigation suggested that great benefits could be achieved if more liberal aviation policies such as those proposed by the Belt and Road initiative of the Chinese government were introduced. The analysis suggested that if the Central Asia–China markets were regulated and operated in a similar way to the routes between Central Asia and other states, the probability of having aviation services between cities in China and Central Asia would increase substantially, with more direct flights to a larger number of Chinese cities such as Xiamen, Shenzhen, Hangzhou, Qingdao, Chengdu, and Kunming.

Airline Privatization

A change from public to private firm ownership can serve efficiency, financing, and distributive objectives (Vickers and Yarrow, 1991). These objectives are interdependent in the sense that the effect of privatization on the efficiency of firms relative to a situation with public ownership can depend on whether the firms’ environments are competitive or monopolistic. More specifically, the efficiency effects of privatization may be limited in a competitive environment (Vickers and Yarrow, 1991). The deregulation of aviation markets coincided with airline privatizations. Note that private ownership typically involves local ownerships because trans-border airline investment is often limited by the constraints embedded in the national airline establishment

regimes (Walulik, 2016). The following airline privatization studies and cases provide evidence for the effects of deregulation and airline privatization on aviation markets.

By 2016, there were around 27% and 40% of airlines (excluding Low-cost airlines) around the world that are completely state-owned and completely privatized, respectively. While the rest of airlines are mix-owned by the governments and private investors. Continent-wise, the number varies dramatically with the lowest 5.6% of completely privatized airlines in Oceania and Pacific Islands regions, and highest 69.6% of completely privatized airlines in North America, where all US airlines are completely privatized, while in Canada, a couple of airlines belong to, for example, First Nations. In Latin America and the Caribbean, the share of fully privatized airlines is 46.5%, in Europe it is 45.5%, Asia 39.5%, Middle East (including North Africa) 25% and Africa (excluding North Africa) 22.2%. This shows that a large share of the airlines worldwide is still completely or partly owned by the governments (Czerny and Lang, in press).

Several studies have analyzed the effects of airline privatization on their performance. Al-Jazzad (1999) found an overall positive evidence of private airline ownership on performance in the sense of growth in sales and net-income, total assets, capital expenditures, and dividends after privatization. Interestingly, the findings further indicated an increase in employment after privatization. Chow (2010) showed that non-state-owned airlines performed better than their state-owned counterparts. The study by Backx et al. (2002) covered mixed ownership structures in their analysis. They found that airlines with mixed ownership tend to perform better than publicly owned airlines, but worse than privately owned airlines. On the contrary, Chen et al. (2017) focused on the Chinese market and found that Chinese airlines are better off with strong privatization or majority of state ownership than with mixed ownership. The findings are consistent with Wei and Varela (2003), Shirai (2004), Wei et al. (2005), Tian and Estrin (2008), and Ng et al. (2009), which all supported that businesses can benefit from either strong state ownership or strong privatization. However, no significant negative or positive effects of public ownership have been found by Sjögren and Söderberg (2011).

The following considers specific privatization cases. Air Canada has been partly privatized in 1988 and fully privatized in 1989. Gillen et al. (1989) found that the public ownership of Air Canada has historically resulted in a substantial loss of productive efficiency due to an overexpansion of its capital. Eckel et al. (1997) analyzed the effect of privatization on the performance of British Airways. They found that airfares fell in the markets served by British Airways and stock prices of US rivals fell upon British Airways' privatization. Both effects indicated the possible pro-competitive effects and efficiency gains of airline privatization. The case of Argentina showed that the selling of publicly owned shares in airlines to private owners can be complicated. While many bids existed for the national flag airline Aerolineas Argentinas at the beginning of the process, only one bidder remained in the end, and there were allegations about corruption (Saba and Manzetti, 1997). The privatization of Kenya Airways in 1996 helped avert liquidation of the company and is considered a success story because it has helped the management to focus on commercial goals and objectives (Debra and Toroitch, 2005).

Conclusions

The historic development of the airlines' regulatory environments and ownerships across countries, regions, and continents has been described in this chapter. It was shown that deregulation in the air transport market has become a mainstream development, and that deregulation has changed aviation markets in many positive ways. Deregulation generally led to stronger competition, reduced fares, increased flight frequencies, more connections, and increased passenger numbers. This strongly indicates that deregulation is beneficial to the aviation industry and society and thus desirable, which is a conclusion that is well in line with those from other researchers (e.g., Oum et al., 2010).

Deregulation is still far from completion. In China, the market is still characterized as state-led rather than market-led. The restrictions on low-cost airlines in China result in fewer low-cost airlines and softened competition. In ASEAN and Africa, many promising plans for deregulating the aviation market have not been realized yet.

The airlines' privatization development around the world has been briefly described as well. Many studies showed that public airline ownership can be associated with major financial losses. Private ownership helped avoid such losses by improving efficiency and increasing revenues. However, a large share of today's airlines is still completely or partly government owned.

References

- Abate, M., 2016. Economic effects of air transport market liberalization in Africa. *Transp. Res. Part A* 92, 326–337.
- Al-Jazzad, M.I., 1999. Impact of privatization on airlines performance: an empirical analysis. *J. Air Transp. Manage.* 5, 45–52.
- Airbus, 2018. Global Market Forecast. Global Networks, Global Citizens 2018–2037.
- Association of Southeast Asian Nations (ASEAN), 2004. Action plan for ASEAN air transport integration and liberalization. ASEAN Document Series, pp. 221–226.
- Backx, M., Carney, M., Gedajlovic, E., 2002. Public, private and mixed ownership and the performance of international airlines. *J. Air Transp. Manage.* 8, 213–220.
- Baik, Y.-S., Kwak, B., Lee, J., 2011. Deregulation and earnings management: the case of the U.S. airline industry. *J. Account. Public Policy* 30, 589–606.
- Bailey, E.E., 2002. Aviation policy: past and present. *South. Econ. J.* 69, 12–20.
- Bailey, E.E., Baumol, W.J., 1984. Deregulation and the theory of contestable markets. *Yale J. Regul.* 1, 111–137.
- Berechman, J., de Wit, J., 1996. An analysis of the effects of European Aviation Deregulation on an airlines' network structure and choice of a primary West European hub airport. *J. Transp. Econ. Policy* 30, 251–274.
- Bernardo, V., Fageda, X., 2017. The effects of the Morocco-European Union open skies agreement: a difference-in-difference analysis. *Transp. Res. Part E* 98, 24–41.

- Burghouwt, G., de Wit, J.G., 2015. In the wake of liberalisation: long-term developments in the EU air transport market. *Transp. Policy* 43, 104–113.
- Button, K.J., 1996. Aviation deregulation in the European Union: do actors learn in the regulation game? *Contemp. Econ. Policy* 14, 70–80.
- Button, K.J., 2009. The impact of US-EU “Open Skies” agreement on airline market structures and airline networks. *J. Air Transp. Manage.* 15, 59–71.
- Button, K.J., Swann, D., 1992. Transatlantic lessons in aviation deregulation: EEC and US experiences. *Antitrust Bull.* 37, 207–255.
- Caves, D.W., Christensen, L.R., Tretheway, M.W., 1984. Economies of traffic density versus economies of scale: Why trunk and local service airline costs differ. *RAND J. Econ.* 15, 471–489.
- Chen, S.-J., Chen, M.-H., Wei, H.-L., 2017. Financial performance of Chinese airlines: does state ownership matter? *J. Air Transp. Manage.* 33, 1–10.
- Chow, C.K.W., 2010. Measuring the productivity changes of Chinese airlines: the impact of the entries of non-state-owned carriers. *J. Air Transp. Manage.* 16, 320–324.
- Civil Aviation Administration of China, 2002. China's Civil Aviation Statistics (1949–2000). Department of Planning and Development, CAAC, Beijing.
- Cosmas, A., Belobaba, P., Swelbar, W., 2010. The effects of open skies agreements on transatlantic air service levels. *J. Air Transp. Manage.* 16, 222–225.
- Czerny, A.I., Lang, H. Airline exits, entries and ownership structure: policy lessons from history. Unpublished.
- Debra, Y.A., Toroitich, O.K., 2005. The making of an African success story: the privatization of Kenya Airways. *Thunderbird Int. Bus. Rev.* 47, 205–230.
- Derudder, B., Witlox, F., 2009. The impact of progressive liberalization on the spatiality of airline networks: a measurement framework based on the assessment of hierarchical differentiation. *J. Transp. Geogr.* 17, 276–284.
- Dobruszkes, F., 2009. Does liberalisation of air transport imply increasing competition? Lessons from the European case. *Transp. Policy* 16, 29–39.
- Dobruszkes, F., Mondou, V., Ghedira, A., 2016. Assessing the impacts of aviation liberalisation on tourism: some methodological considerations derived from the Moroccan and Tunisian cases. *J. Transp. Geogr.* 50, 115–127.
- Doganis, R., 2002. *Flying Off Course—The Economics of International Airlines*, third ed. Routledge, London.
- Eckel, C., Eckel, D., Singal, V., 1997. Privatization and efficiency: industry effects of the sale of British Airways. *J. Financial Econ.* 43, 275–298.
- Findlay, C., Forsyth, P., 1992. The making of an Asian-Pacific region. *Asian-Pacific Econ. Lit.*, 6(2), 1–10.
- Forsyth, P., King, J., Rodolfo, C.L., 2006. Open skies in ASEAN. *J. Air Transp. Manage.* 12 (3), 143–152.
- Fu, X., Oum, T.H., Zhang, A., 2010. Air transport liberalization and its impacts on airline competition and air passenger traffic. *Transp. J.* 49 (4), 24.
- Fu, X., Lei, Z., Wang, K., Yan, J., 2015. Low cost carrier competition and route entry in an emerging but regulated aviation market—the case of China. *Transp. Res. Part A* 79, 3–16.
- Gillen, D., Oum, T.H., Tretheway, M.W., 1989. Privatization of air Canada: why its necessary in a deregulated environment. *Can. Public Policy* 15, 285–299.
- Goetz, A.R., Vowles, T.M., 2009. The good, the bad, and the ugly: 30 years of US airline deregulation. *J. Transp. Geogr.* 17, 251–263.
- Hanaoka, S., Takebayashi, M., Ishikura, T., Saraswati, B., 2014. Low-cost carriers versus full service carriers in ASEAN: the impact of liberalization policy on competition. *J. Air Transp. Manage.* 40, 96–105.
- Hunnicut, C.A., 2010. US-EU second stage air transport agreement: toward an open aviation area. *Georgia J. Int. Comp. Law* 39, 663–725.
- International Air Transport Association (IATA), 2007. *Airline Liberalisation*. IATA Economics Briefing No 7.
- International Air Transport Association (IATA), 2018. India's air transport sector. The future is bright...But not without challenges.
- Jankiewicz, J., Huderek-Glaska, S., 2016. The air transport market in Central and Eastern Europe after a decade of liberalisation—different paths of growth. *J. Transp. Geogr.* 50, 45–56.
- Jiang, Y., Yao, B., Wang, L., Feng, T., Kong, L., 2017. Evolution trends of the network structure of Spring Airlines in China: a temporal and spatial analysis. *J. Air Transp. Manage.* 60, 18–30.
- Johnston, A., Ozment, J., 2013. Economies of scale in the US airline industry. *Transp. Res. Part E* 51, 95–108.
- Kahn, A.E., 1988. Surprises of airline deregulation. *Am. Econ. Rev.* 78, 316–322.
- Kanafani, A., Ghobrial, A.A., 1985. Airline hubbing—some implications for airport economics. *Transp. Res. Part A* 19A, 15–27.
- Laplace, I., Latge-Roucolle, C., 2016. Deregulation of the ASEAN air transport market: measure of impacts of airport activities on local economies. *Transp. Res. Procedia* 14, 3721–3730.
- Le, T.T., 1997. Reforming China's airline industry: from state-owned monopoly to market dynamism. *Transp. J.* 36 (2), 45–62.
- Lei, Z., O'Connell, J.F., 2011. The evolving landscape of Chinese aviation policies and impact of a deregulating environment on Chinese carriers. *J. Transp. Geogr.* 19, 829–839.
- Lei, Z., Yu, M., Chen, R., O'Connell, J.F., 2016. Liberalization of China-US air transport market: assessing the impacts of the 2004 and 2007 protocols. *J. Transp. Geogr.* 50, 24–32.
- Li, L., 2001. Deregulate airfares. *Market News*, 03.04.2001, p. 6.
- Lieshout, R., Malighetti, P., Redondi, R., Burghouwt, G., 2016. The competitive landscape of air transport in Europe. *J. Transp. Geogr.* 50, 68–82.
- McShan, S., Windle, R., 1989. The implications of hub-and-spoke routing for airline costs and competitiveness. *Logist. Transp. Rev.* 25, 209–230.
- Mills, G., Membreno, L., 2007. *Air Hubs: A Checklist for Africa*. The Brenthurst Foundation, South Africa.
- Mills, G., Swantner, L., 2008. *Wings over Africa? Trends and Models for African Air Travel*. The Brenthurst Foundation, South Africa.
- Myburgh, A., Sheik, F., Fiandeiro, F., Hodge, J., 2006. *Clear skies over Southern Africa: the Importance of Air Transport Liberalisation for Shared Economic Growth*. The ComMark Trust, Woodmead, South Africa.
- Morandi, V., Malighetti, S.P., Redondi, R., 2014. EU-US open skies agreement: what is changed in the North transatlantic skies? *Transp. J.* 53, 305–329.
- Morrison, S.A., Winston, C., 1990. The dynamics of airline pricing and competition. *Am. Econ. Rev.* 80, 389–393.
- Morrison, W.G., de Wit, J., 2019. US open skies agreements and unlevel playing fields. *J. Air Transp. Manage.* 74, 30–38.
- Moses, L.N., Savage, I., 1990. Aviation deregulation and safety. *J. Transp. Econ. Policy* 24, 171–188.
- Ng, A., Yuce, A., Chen, E., 2009. Determinants of state equity ownership, and its effect on value/performance: China's privatized firms. *Pacific-Basin Finance J.* 17 (4), 413–443.
- Njoya, E.T., 2016. Africa's single aviation market: the progress so far. *J. Transp. Geogr.* 50, 4–11.
- Njoya, E.T., Christidis, P., Nikitas, A., 2018. Understanding the impact of liberalisation in the EU-Africa aviation market. *J. Transp. Geogr.* 71, 161–171.
- Oum, T.H., Zhang, A., Fu, X., 2010. Air transport liberalization and its impacts on airline competition and air passenger traffic. *Transp. J.* 49, 24–41.
- Oum, T.H., Wang, K., Yan, J., 2019. Measuring the effects of open skies agreements on bilateral passenger flow and service export and import trades. *Transp. Policy* 74, 1–14.
- Pitfield, D.E., 2009. The assessment of the EU-US Open Skies Agreement: the counterfactual and other difficulties. *J. Air Transp. Manage.* 15, 308–314.
- Saba, R.P., Manzetti, L., 1997. Privatization in Argentina: the implications for corruption. *Crime Law Soc. Change* 25, 353–369.
- Schlumberger, C.E., 2010. *Open Skies for Africa. Implementing the Yamoussoukro Decision*. The Office of the Publisher, Washington, USA.
- Shaw, S.L., Lu, F., Chen, J., Zhou, C., 2009. China's airline consolidation and its effects on domestic airline networks and competition. *J. Transp. Geogr.* 17 (4), 293–305.
- Shirai, S., 2004. Testing the three rules of equity markets in developing countries: the case of China. *World Dev.* 32 (9), 1467–1486.
- Sjögren, S., Söderberg, M., 2011. Productivity of airline carriers and its deregulation, privatisation and membership in strategic alliances. *Transp. Res. Part E* 47, 228–237.
- Surovitskikh, S., Lubbe, B., 2015. The Air Liberalisation Index as a tool in measuring the impact of South Africa's aviation policy in Africa on air passenger traffic flows. *J. Air Transp. Manage.* 42, 159–166.
- Tan, A.K.-J., 2010. The ASEAN multilateral agreement on air services: en route to open skies? *J. Air Transp. Manage.* 16, 289–294.
- Tan, A.K.-J., 2013. Toward a single aviation market in ASEAN: regulatory reform and industry challenges (No. DP-2013-22).
- Tian, L., Estrin, S., 2008. Retained state shareholding in Chinese PLCs: does government ownership always reduce corporate value? *J. Comp. Econ.* 36 (1), 74–89.
- UNECA, 2002. *Clarification Issues and Articles of the Yamoussoukro Decision*. United Nation Economic Commission for Africa, Addis Ababa.
- Vickers, J., Yarrow, G., 1991. Economic perspectives on privatization. *J. Econ. Perspect.* 5, 111–132.
- Walulik, J., 2016. At the core of airline foreign investment restrictions: a study of 121 countries. *Transp. Policy* 49, 234–251.
- Wang, J., Bonilla, D., Banister, D., 2016. Air deregulation in China and its impact on airline competition 1994–2012. *J. Transp. Geogr.* 50, 12–23.
- Wang, K., Fu, X., Czerny, A.I., Hua, G., Lei, Z., 2018. Modeling the potential for aviation liberalization in Central Asia—Market analysis and implications for the Belt and Road initiative. Unpublished.

- Wang, K., Zhang, A., Zhang, Y., 2018b. Key determinants of airline pricing and air travel demand in China and India: policy, ownership, and LCC competition. *Transp. Policy* 63, 80–89.
- Wei, Z., Varela, O., 2003. State equity ownership and firm market performance: evidence from China's newly privatized firms. *Global Finance J.* 14 (1), 65–82.
- Wei, Z., Xie, F., Zhang, S., 2005. Ownership structure and firm value in China's privatized firms: 1991e2001. *J. Financ. Quant. Anal.* 40 (1), 87–108.
- Winston, C., Yan, J., 2015. Open skies: Estimating traveler's benefits from free trade in airline services. *Am. Econ. J.: Econ. Policy* 7, 370–414.
- Zhang, A., 1998. Industrial reform and air transport development in China. *J. Air Transp. Manage.* 4 (3), 155–164.
- Zhang, A., Chen, H., 2003. Evolution of China's air transport development and policy towards international liberalization. *Transp. J.* 42, 31–49.
- Zhang, Y., Round, D.K., 2008. China's airline deregulation since 1997 and the driving forces behind the 2002 airline consolidations. *J. Air Transp. Manage.* 14 (3), 130–142.
- Zhang, Q., Yang, H., Wang, Q., Zhang, A., 2014. Market power and its determinants in the Chinese airline industry. *Transp. Res. Part A* 64, 1–13.
- Zhang, S., Derudder, B., Fuellhart, K., Witlox, F., 2018. Carriers' entry patterns under EU-US open skies agreement. *Transp. Res. Part E* 111, 101–112.

Price Discrimination and Yield Management in the Airline Industry

Guoquan Zhang*, Colin C.H. Law*,†, Yahua Zhang*, Hangjun Yang*, *School of Commerce, University of Southern Queensland, Toowoomba, QLD, Australia; †Faculty of Business and Technology, Stamford International University, Bangkok, Thailand; ‡School of International Trade and Economics, University of International Business and Economics, Beijing, China

© 2021 Elsevier Ltd. All rights reserved.

An Introduction to Price Discrimination	404
New Business Model Enhances Price Discrimination: A Case of the Australian Airline Market	405
A Short History of Yield Management	405
Conditions for the Application of Yield Management	406
Yield Management Methods	406
Big Data Analytics, Artificial Intelligence, and Yield Management	407
References	407
Further Reading	408

An Introduction to Price Discrimination

Price discrimination is a business term referring to the charging different prices to different customers, even though they are purchasing the same product produced with the same marginal cost. The concept of price discrimination was first proposed by a French economist, Jules Dupuit in 1840s, who suggested that some transportation and utility firm could benefit from engaging in the price discrimination strategy (Ekelund, 1970). It is believed that a monopoly firm's profit can be enlarged if the market can be segmented to allow the buyers to pay a price close to what they can afford (Gifford and Kudrle, 2008).

It is well known that a firm's ability to price discrimination depends on three conditions (Varian, 1989): the possession of a certain degree of market power; the ability to segment the market; the ability to stop the resale of the product or service from one buyer segment to another. The practice of price discrimination is one of the causes of price variation in the airline industry. There are three types of price discrimination: first degree, second degree, and third degree.

First-degree price discrimination is called perfect price discrimination, which charges each consumer a price equal to the consumer's willingness to pay. Although consumer surplus is completely eliminated under first-degree price discrimination, it is economically efficient and the production is at the point where marginal cost equals marginal benefit. However, in reality, this type of price discrimination is unlikely to happen because a firm usually does not have the ability to determine a consumer's reservation price. Therefore, no airline can practice first-degree price discrimination.

Second-degree price discrimination is the practice of charging consumers different prices for different quantities of a homogeneous product consumed. This type of discrimination allows the firm to take part, but not the entire consumer surplus by offering a few, well-defined pricing categories. A frequent flyer program may fall within this type of price discrimination as the points awarded to the passengers can be used later, which is a special kind of discount and consumers do not need to pay for the points at the time of purchasing the ticket.

Third-degree price discrimination refers to the charge of different prices to different consumer groups such as male and female, children and adults. That is, special discounts are offered to some groups and thus each group pays a different price for consuming the same product or service. This type of price discrimination allows the firm to expand its client base, as some consumer groups may not buy the product or service without the discount. Many airlines offer corporate travel programs to companies and give cheaper contract prices, which is a practice of third-degree price discrimination.

Third-degree price discrimination is most commonly used by the airline industry. Airlines divide consumers into different groups based on a set of characteristics and charge higher prices to those with the most inelastic demand. A typical example is to make a distinction between time-sensitive (mainly business travelers) and non-time-sensitive passengers (mainly leisure travelers), as well as between point-to-point passengers and connecting passengers. Generally, time-sensitive travelers prefer faster connections and place a lot of value on flight punctuality and frequency of service. They usually do not book their flights in advance. Their tickets need to be transferable from one flight to another in the event of changes in travel plans at short notice. As their company usually pays for their tickets, this group of passengers often buys less restricted tickets at a higher price. In contrast, non-time-sensitive travelers are interested in obtaining the lowest fares and are willing to accept longer travel times and more restrictions on the use of their tickets, for example, not refundable, not endorsable, no route and date/flight change, etc. In practice, business passengers can be identified by their inability to book in advance and by buying tickets with fewer restrictions.

Interestingly, some researchers argue that airline pricing falls somewhere between first- and second-degree price discrimination because of the use of computer reservation system (CRS) (see, e.g., Kons, 1999). The reason is that although it is impractical to obtain all the potential customers' reservation prices, which makes first-degree impossible, the creation of a large number of prices possible by the CRS allows the airlines to snatch more consumer surplus than second-degree discrimination offers.

Overall, it is generally agreed that price discrimination has helped lower the average fare these days, although it may raise the fare for some consumer groups (Vasigh et al., 2018). Compared with the single pricing strategy that may drive some price-sensitive passengers out of the airline market, price discrimination can increase load factors and economic efficiency due to the greater output produced.

New Business Model Enhances Price Discrimination: A Case of the Australian Airline Market

Qantas and Virgin have reported bumper profits in the first half of the 2017–18 financial year. The record profits for Qantas mainly came from the domestic market—Qantas and Jetstar’s domestic business recorded the highest ever first-half underlying EBIT of \$ 652 million (<https://www.qantasnewsroom.com.au/media-releases/qantas-delivers-record-first-half-profit-invests-in-aircraft-and-training/>). Virgin Australia also attributes its 142% growth in underlying profits to the strong domestic demand from business and leisure flyers (<https://www.businessinsider.com.au/virgin-australia-has-returned-to-profit-2018-2>). Strong demand and cost control are two frequently mentioned factors for their success in the domestic market. However, another factor that is less mentioned is the airline-within-airline (AinA) business model adopted by Australian major airlines (Zhang et al., 2017, 2018). This business model is another example that promotes price discrimination in the airline industry.

The AinA strategy was largely a failure in the United States and Europe, as most low-cost subsidiaries of the full-service carriers did not achieve sufficient low costs to compete against the standalone low-cost carriers (Morrell, 2005). As a result, some of the US carriers, such as American Airlines sought to segment their cabin seating and offer a basic economy seating class to those looking for cheaper prices.

Australia’s domestic airline market today is a classical duopoly (Ma et al., 2019). Qantas has successfully implemented the AinA strategy and used Jetstar as a fighting brand against other low-cost carriers in maintaining the group’s total domestic market share at around 65% in the last decade. Virgin has also successfully used this model to snatch a substantial share of Qantas’ business and corporate passengers in the last few years, and at the same time continued to use Tiger to target the leisure market.

The increasing use of Internet for airline ticket distribution has made it easier for consumers to compare prices across multiple competitors. This has limited the airlines’ ability to practice price discrimination. Price discrimination has been a common practice in the airline industry used to transfer consumer surplus to airlines’ own profits through charging different prices to different consumer groups based on the knowledge of consumers’ willingness to pay.

The AinA strategy has allowed extensive capture of consumer surplus of segmented customer groups to a maximum extent. In August 2017, of Australia’s top 144 domestic route markets in terms of passenger traffic volume (each direction of a route was counted as a separate market), Qantas and Jetstar were simultaneously present in 95 markets while Virgin and Tiger in 75 markets. The number of markets where all the four carriers were present was 71 (Ma et al., 2019).

In fact, the four carriers were simultaneously providing services on most of the heavily traveled and profitable domestic routes where each airline group can exploit consumer surplus of both business and leisure passengers. Consumers in these markets are sufficiently heterogeneous in term of their travel demand. The AinA model ensures that a variety of products are provided on these routes to meet consumers’ demand, which allows the two airline groups to maximize the potential of their revenue.

The capture and transfer of consumer surplus from passengers of each segment to airlines could substantially offset the loss resulting from the price war episodes. In addition, in many cases, price wars are only restricted to the product of a certain market segment. The prices for other segments may have little fluctuations. This is confirmed from the charts of the Domestic Air Fare Indexes published by the Department of Infrastructure, Regional Development and Cities, Australian Federal Government, in which we can see that only the best discount fares, exhibit wide fluctuations (https://bitre.gov.au/statistics/aviation/air_fares.aspx).

Price discrimination can increase total social welfare as it allows more products to be produced to a greater number of consumers, some of whom may otherwise not have access to these products. As long as consumers continue to regard Qantas (and Virgin to a less extent probably) as a premium brand and as long as it can successfully maintain such image, most of them will happily transfer part of their consumer surplus to the carrier as profits.

A Short History of Yield Management

Yield management is a pricing strategy that has been adopted by the air transport industry for decades. It is a strategy used to maximize profit by charging the consumer of different demand at a different price. This strategy is mainly popular because airline products, namely, transport services to both passenger and cargo are perishable, which cannot be stored, saved, returned, and resold after the departure of the flight (Nairn, 2005). Yield management helps the carriers to achieve high load factors to ensure high yields.

Yield management related program was first introduced by British Overseas Airways Corporation (BOAC) in the 1960s after the new jet airliner B707s was introduced and resulted in a substantial increase in the capacity that the airline could offer (Easton, 1959). Compared with the aircraft DC-7C used by BOAC, the payload of B707 was almost five times higher. This led to the launching of the excursion fare discount program known as “Early Bird.” The discount was first offered to passengers traveling between the United Kingdom and the Caribbean and later made available on other routes (Cole, 2005). The “Early Bird” discount came with some conditions. For example, passengers must book their seats three months in advance. They might be required to have a minimum stay of fourteen days at the destination.

Yield management has become popular in the United States since 1980s after the airline industry was deregulated. Throughout the period from the 1940s to 1970s, the US airline market was controlled and regulated by the Civil Aeronautics Board (CAB), which exercised its authority over entry and exit, pricing, capacity, frequency, and so on. The greatest event in the airline industry was the passage of the Airline Deregulation Act (ADA) in the United States in 1978. This led to a series of liberalizations in fares, services, and entry and exit to and from this industry across the world (Zhang and Round, 2008). Carriers were allowed to adjust fares, to service new markets, and to enter or exit as long as they met certain basic requirements.

Following the air deregulation act in 1978, many new airlines entered the market. The intensely competitive environment forced airlines to explore new strategies to cope with market change (Law et al., 2018). For example, People Express Airlines based in New York launched its operation in 1981. The airline adopted the low-cost business model and soon became a serious threat to the giant airlines including American Airlines and United Airlines. The newcomer forced both American Airlines and United Airlines matches the People Express Airlines airfare. However, matching fare with the low-cost competitors was not a long-term strategy given the different cost structure between the full service and low-cost airlines. To obtain a balance between price and payload, American Airlines developed a yield management system to compete with other carriers with the aid of CRS (Smith et al., 1992; Cross, 1995).

Airline distribution for many years has been synonymous with the CRS, later termed the Global Distribution System (GDS). The GDS provides virtual real-time connectivity between thousands of suppliers of travel inventory (airlines, hotels, car rental, tour operators, cruise lines, etc.) and hundreds of thousands of retail sellers of travel products (Sismanidou et al., 2009). The system has significantly improved airlines efficiency in attracting and retaining customers and driving sales through travel agents. These gave American Airlines numerous consumer behavior data, which allowed the airline to develop a sophisticated yield management program to attract both the high yield and low yield passengers (Sabre, 2019). In contrast, those airlines operating with a low-cost model with simple single fare structure were unable to react to this strategy. As a result, some of the small and low-cost airlines were driven out of business. Other airlines followed suit and applied the yield management strategy.

Conditions for the Application of Yield Management

Yield management is a technique used by many industries, particularly in those with capacity constraint (Kimes, 1989). The objective of applying yield management is to maximize revenue with limited resources. Air transport companies use yield management to increase their revenue for each flight by predicting the travel demand of different travelers segments between origin and destination and charge each segment with an optimal price. Airline products have the following characteristics that make them most suitable for the application of yield management: each aircraft has fixed capacity; the products are perishable and cannot be stored for the next flight; the passenger groups can be segmented into different sub-groups that allow airlines to charge different prices accordingly; passenger demand is subject to frequent fluctuations; the products can be sold in advance; the marginal cost is relatively low as compared with the fixed cost (Wirtz et al., 2003).

Successfully applying the yield management technique is not an easy task. Airlines are required to divide the customers into smaller segments and offer them the most suitable and wanted products. Given that contemporary business environment changes frequently and rapidly, airlines need to collect customers' purchasing behavior data, based on which the airline can monitor the market, change the fare and control the seat inventory (Akerkar, 2014). The objective is to ensure that not all seats are sold at a low price to reduce the airline revenue, while at the same time the seats are not sold at an expensive price that steers passengers away to the competitors, leading to aircraft departing with empty seats. An effective application of yield management requires the following conditions: an effective constraint that prohibits high yield passengers to buy low fares; a mechanism to allocate seats to maximize revenue in the face of demand fluctuations; and a reliable prediction of demand, no-shows, overbooking, etc. (Ozer and Phillips, 2012; Wensveen, 2015).

To ensure that the right price is charged based on the customers' demand; airlines have created multiple booking codes in a cabin class. To differentiate each booking code, different restrictions are set to prevent the high yield customers from buying low fares. In general, the lower fare class, the more limitations. Common restrictions (also called fences) include (Wensveen, 2015; Vasigh et al., 2018):

- Advance purchase requirement
- Restrictions on earning frequent flyer points
- Restrictions on departure days and times
- Round-trip requirement
- Minimum and maximum stays
- Flight change and refund penalties

Yield Management Methods

To maximize the profit, an airline needs to identify the amount of inventory available at each fare bucket. Fare buckets are closely connected, known as serial nesting (Vasigh et al., 2018). When a seat was sold from one fare bucket, the amount of inventory remaining in other buckets may be affected. For example, when a seat in the higher fare bucket was sold, a seat from the lower fare bucket will be reduced. Therefore, the key of airline yield management is about the seat inventory control. That is, optimally allocating the fixed number of seats to the demand before a flight departs.

The leg-based yield management system is common in the airline industry. In the single-leg seat inventory control, the booking system evaluates a booking request by one segment of flight without considering the complete itinerary of the travelers. In other words, the system shows the same availability to all travelers regardless of a single-leg or multiple-leg journey in the passengers' itinerary (Skugge, 2002).

Following air transport liberalization, many airlines have adopted the hub-and-spoke system. Today, the majority of travelers' itinerary consists of multiple flights, as most routing required the travelers to transit at a hub airport. This is facilitated by the proliferation of airline alliances and code-share agreements, which have increased the complexity of the traditional yield management system. Network seat inventory control is thus called for and an airline's network is optimized simultaneously.

The Origin and Destination (O&D) yield management system provides a solution for network seat inventory control to carriers when itineraries involve connecting and code-sharing flight legs (Millward, 2014). The O&D yield management forecast the travel volume by city pair based on the passengers' complete itinerary including the inbound and outbound connections, which allows the airline to obtain accurate demand data and implement the right pricing strategies (Skugge, 2002). Various mathematical programming models have been developed to realize the network seat inventory control (Williamson, 1992).

Upgrade bidding and airport upsell have become useful yield management tools for some airlines to maximize profits. This strategy is useful to resolve problems when there are forecast errors between supply and demand (Gallego and Stefanescu, 2009). The airline offers passengers in the lower cabin to entering an auction exercise for upgrading to premium seating by additional payment. The customers are free to offer their bidding price (Rosen, 2018). Closer to the departure date, when premium seats are still available, the airlines will issue the upgrade confirmation to those passengers who offer the highest bid. Airport up sell is another strategy similar to the upgrade-bidding program. Some airlines offer to up sell deal to passengers to upgrade to higher service cabin on the departure day with a charge at the airport check-in counter (<https://atwonline.com/blog/airlines-need-master-art-upsell>). These strategies allowed the airline to earn additional last minute earning to further maximize their profits.

Big Data Analytics, Artificial Intelligence, and Yield Management

Traditional yield management adopted by airlines for decades relied on the historical data collected by the CRS. Modern technology has changed the shape of yield management, which allows airlines to gather data before customers purchase tickets. Big data and artificial intelligence are two technologies that are closely linked and can revolutionize the traditional yield management system. Big data refer to the raw data or inputs that are uncleansed and unstructured, while artificial intelligence is the output and actions taken to resolve the issues. The development of artificial intelligence into yield management system gives the airlines a large volume of useful data of customer demand, which makes it possible for the airlines to offer a fare that is closer to the customers' demand (Corominas, 2016). The big-data based artificial intelligence system collects data from the passenger's flight search history over the internet to examine the demand for flights. Based on the customers' browsing pattern, the artificial intelligence system tracks the changing volume of search requests from the airline website or other search engines. The artificial intelligence allows the airlines to make prompt changes to the airfare, if needed (Hinke, 2018). In particular, the artificial intelligence system generates customer-centric offers and develops personalized pricing strategies to meet the needs of consumers (<https://www.triometric.net/the-ai-opportunity-in-revenue-management/>). The new technology is the key for the modern era airlines to maximize profits and succeed in the rapidly changing business environment.

References

- Akerkar, R., 2014. Analytics on big aviation data: turning data into insights. *IJCSA* 11 (3), 116–127.
- Cole, S., 2005. *Applied Transport Economics: Policy Management and Decision Making*. Kogan Page, London.
- Corominas, S.M., 2016. Is data science the future for airline revenue management? Medium. 10 July 2016. Available from: <https://medium.com/@sergiomc/is-data-science-the-future-for-airline-revenue-management-9334cf024899>.
- Cross, R.G., 1995. An introduction to revenue management. In: Jenkins, D. (Ed.), *The Handbook of Airline Economics*. The Aviation Weekly Group of the McGraw-Hill Companies, New York, NY, pp. 443–468.
- Easton, P.R., 1959. Jets in the foreign fleet. *Flying Magazine*, 44–45.
- Ekelund, R.B., 1970. Price discrimination and product differentiation in economic theory an early analysis. *Q. J. Econ.* 84, 268–278.
- Gallego, G., Stefanescu, C., 2009. Upgrades, upsells and pricing in revenue management. SSRN 1334341. Available from: <https://ssrn.com/abstract=1334341> and <http://dx.doi.org/10.2139/ssrn.1334341>.
- Gifford, D.J., Kudrle, R.T., 2008. The law and economics of price discrimination in modern economies: time for reconciliation? *Minnesota Legal Studies Research Paper No. 08-21*, pp. 1237–1292.
- Hinke, A.C., 2018. Enhancing customer experience and growing revenue with artificial intelligence. PROS. 15 May 2018. Available from: <https://resources.pros.com/blog/improve-airline-revenue-management-dynamic-pricing-ai>.
- Kimes, S.E., 1989. Yield management: a tool for capacity-considered service firms. *J. Oper. Manag.* 8 (4), 348–363.
- Kons, A., 1999. Understanding the chaos of airline pricing. *Park Place Econ* 8, 15–29.
- Law, C.C., Zhang, Y., Zhang, A., 2018. Regulatory changes in international air transport and their impact on tourism development in Asia Pacific. *Airline Economics in Asia*. Emerald Publishing Limited, Bingley, UK pp. 123–144.
- Ma, W., Wang, Q., Yang, H., Zhang, Y., 2019. An analysis of price competition and price wars in Australia's domestic airline market. *Transp. Policy* 81, 163–172.
- Millward, L., 2014. O&D Stepping Stone. Sabre. Available from: https://www.sabreairlinesolutions.com/pdfs/O&D_Stepping_Stone.pdf.
- Morrell, P., 2005. Airlines within airlines: an analysis of US network airline responses to low cost carriers. *J. Air Transp. Manag.* 11 (5), 303–312.
- Nairn, G., 2005. Airline seats have become commodities. *Financial Times*. 13 July 2015. Available from: <https://www.ft.com/content/ad233116-f2bf-11d9-8094-00000e2511c8>.

- Ozer, O., Phillips, R., 2012. *The Oxford Handbook of Pricing Management*. Oxford University Press, Oxford.
- Rosen, E., 2018. Fly business class for cheap by bidding on airline upgrades. *Forbes*. 29 August 2018. Available from: <https://www.forbes.com/sites/ericrosen/2018/08/29/fly-business-class-for-cheap-by-bidding-on-airline-upgrades/#3453dcb47a4c>.
- Sabre, 2019. *The Sabre Story*. Sabre. Available from: <https://www.sabre.com/files/Sabre-History.pdf>.
- Sismanidou, A., Palacios, M., Tafur, J., 2009. Progress in airline distribution systems: the threat of new entrants to incumbent players. *J. Ind. Eng. Manag.* 2 (1), 251–272.
- Skugge, G., 2002. Implementing airline origin and destination revenue management systems. *J. Revenue Pricing Manag.* 1 (3), 255–266.
- Smith, B., Leimkuhler, J., Darrow, R., 1992. Yield management at American Airlines. *Interfaces* 22, 8–31.
- Varian, H.A., 1989. Price discrimination. In: Armstrong, M., Porter, R. (Eds.), *Handbook of Industrial Organization*. Elsevier, North Holland, pp. 597–654.
- Vasigh, B., Fleming, K., Tacker, T., 2018. *Introduction to Air Transport Economics: From Theory to Application*. Routledge, UK.
- Wensveen, J.G., 2015. *Air Transportation: A Management Perspective*. Routledge, New York.
- Williamson, E.L., 1992. *Airline network seat inventory control: methodologies and revenue impacts*. Doctoral dissertation, Massachusetts Institute of Technology.
- Wirtz, J., Kimes, S.E., Patterson, P., Ho, J.P., 2003. Yield Management: Resolving Potential Customer and Employee Conflicts. Cornell University, SHA School Site. Available from: <http://scholarship.sha.cornell.edu/articles/849>.
- Zhang, Y., Round, D.K., 2008. China's airline deregulation since 1997 and the driving forces behind the 2002 airline consolidations. *J. Air Transp. Manag.* 14 (3), 130–142.
- Zhang, Y., Sampaio, B., Fu, X., Huang, Z., 2018. Pricing dynamics between airline groups with dual-brand services: the case of the Australian domestic market. *Transp. Res. Part A Policy Pract.* 112, 46–59.
- Zhang, Y., Wang, K., Fu, X., 2017. Air transport services in regional Australia: demand pattern, frequency choice and airport entry. *Transp. Res. Part A Policy Pract.* 103, 472–489.

Further Reading

- Belobaba, P.P., Wilson, J.L., 1997. Impacts of yield management in competitive airline markets. *J. Air Transp. Manag.* 3 (1), 3–9.
- Carrier, E., Fiig, T., 2018. Future of airline revenue management. *J. Revenue Pricing Manag.* 17 (2), 45–47.
- Cross, R.G., 1997. *Revenue Management: Hard-Core Tactics for Market Domination*. Broadway Books, New York.
- Donaghy, K., McMahon, U., McDowell, D., 1995. Yield management: an overview. *Int. J. Hosp. Manag.* 14 (2), 139–150.
- Krämer, A., Friesen, M., Shelton, T., 2018. Are airline passengers ready for personalized dynamic pricing? A study of German consumers. *J. Revenue Pricing Manag.* 17 (2), 115–120.
- Oum, T.H., 1998. Overview of regulatory changes in international air transport and Asian strategies towards the US open skies initiatives. *J. Air Transp. Manag.* 4 (3), 127–134.
- Pak, K., Piersma, N.N., 2002. *Airline revenue management: an overview of OR techniques 1982–2001*. Econometric Institute Research Papers EI 2002-03, Erasmus University Rotterdam, Erasmus School of Economics (ESE), Econometric Institute.
- Talluri, K.T., Van Ryzin, G.J., 2006. *The Theory and Practice of Revenue Management*. Springer Science & Business Media, New York.
- Wang, X.L., Yoonjoung Heo, C., Schwartz, Z., Legohérel, P., Specklin, F., 2015. Revenue management: progress, challenges, and research prospects. *J. Travel Tour Mark* 32 (7), 797–811.
- Wittman, M.D., Belobaba, P.P., 2018. Customized dynamic pricing of airline fare products. *J. Revenue Pricing Manag.* 17 (2), 78–90.

Regulation and Competition in Railways

Chris Nash, Andrew Smith, Institute for Transport Studies, University of Leeds, Leeds, Yorkshire, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	409
The US Staggers Act	409
Europe—Freight Traffic	410
Europe—Passenger Traffic	410
Europe—The New Regulatory Bodies	411
Reform in Other Countries	412
Conclusions	412
References	413
Further Reading	413

Introduction

Many of the world's railways were originally built by private companies. Whilst there were often competing main lines between the major centers, they had a degree of monopoly, and—where it existed—the competition was typically between two or at the most three companies. Since they had at the time a very strong market position in both passenger and freight transport, detailed regulation was seen as necessary, including controls on maximum fares and charges and imposition of common carrier obligations, including controls on the ability to withdraw services and close lines.

Over the years, in much of the world, the rail system became publicly owned with the pattern of a single, vertically integrated public monopoly taking control. Often regulation by a separate regulatory body continued for some time, but gradually the fact that passenger services were largely operated under public service obligations and that freight was struggling with intense competition from road made this redundant and it was removed. In North America, on the other hand, the rail system remained primarily privately owned, vertically integrated, competing over parallel routes and concentrated on freight. Because of the unprofitability of the industry, most controls on freight rates and services were removed, most notably in the US by the Staggers Act of 1980. This is considered in the next section.

In Europe, reform went in a different direction. Up until 1999, the European Commission had accepted that railways would remain publicly owned vertically integrated monopolies. Legislation had concentrated on giving them management autonomy and trying to create a level playing field with other modes of transport by relieving them of inherited non-commercial obligations. But there was much dissatisfaction with the performance of this system on behalf of both the European Commission and member states. In the 20 years from 1970 to 1990, rail freight market share in Europe went down from 20.1% to 11.0%, and passenger market share down from 10.2% to 6.6%, whilst subsidies rose. Obviously this loss of market share was due to many factors, some outside the control of the railways – for instance rising car ownership, improvements in road infrastructure and decline of heavy industry – but railways were doing poorly even in markets such as international freight which being longer distance ought to have provided more favorable prospects for rail. Thus in 1991 came the first steps in a radical shift in policy in favor of separating train services from the infrastructure, which was recognized as the true natural monopoly, and introducing competition into the operation of services. The next two sections discuss this separately in the case of freight transport and passenger.

This structure imposed a new need for regulation, partly to ensure that entry conditions did not discriminate in favor of the incumbent, but also to oversee the performance of the new monopoly infrastructure managers. This issue is discussed in section “Europe—the new regulatory bodies”.

Finally in section “Reform in other countries” we review briefly the reform of the rail systems in other parts of the world, including the major rail systems of India, China, Russia and Japan, and the role that regulation or deregulation has played in them before reaching our conclusions.

The US Staggers Act

Until the Staggers Act of 1980, the US had strict regulation of railway tariffs and services. Railways had to carry traffic at the published tariff; individually negotiated tariffs were not permitted. Unprofitable services could not be discontinued and lines closed; these had to be cross-subsidized out of the profits on the healthier traffic. This included cross subsidy of passenger services until 1970, when AMTRAK was set up as a separate government owned passenger operator, operating extensively over the track of the freight companies on a marginal cost basis. It should be said that the pattern of tight regulation of rates and continued cross subsidy of unprofitable traffic from profitable has been a common form of rail regulation, and still continues in India and China. But with the growth of road competition it has become less and less sustainable. Firstly, rail experiences a marginal cost well below average cost,

so average cost pricing may lose much traffic to road which is willing to pay its marginal cost on rail. It is necessary to price differentiate according to the willingness to pay of individual customers in order to recover as much as possible of total cost whilst retaining such traffic. For freight traffic, this is most thoroughly achieved by individually negotiated tariffs. Moreover such tariffs may impose conditions such as investment in terminals or provision of minimum volumes of traffic so that more efficient ways of operating (for instance in full train loads) may be adopted. Secondly, even with price discrimination, rail has often been unable to make sufficient profits to continue cross-subsidy.

In the US, by the 1960s the rail industry was becoming increasingly unprofitable, and even the removal of the requirement to cross-subsidize passenger services could not reverse this trend. Following the bankruptcy of Penn Central, the government realized radical change was needed, and the Staggers Act of 1980 largely swept away the existing regulation. With the ability to negotiate individual contracts, average revenue per ton kilometer halved over the succeeding two decades, but costs fell even further (Winston, 2006). Whilst some routes were abandoned, the availability of road haulage meant that this was not generally a big problem. Indeed, although some residual regulatory power to deal with possible monopolistic pricing policies by US railroads was retained, it was generally considered that the continued existence of parallel competing railroads, each with their own track, the strong competitiveness of road and in some cases water transport, and the competition from competing sources of supply meant that the monopoly power of the railroad industry now was weak. However, there has been a substantial decline in the number of separate US railways in recent years and presumably therefore a decline in intra modal competition. To a degree, this has been mitigated by requirements for the granting of running rights to other operators as a condition of merger, and there have also been voluntary running rights agreements on a purely commercial basis. Thus whilst the US system is fundamentally a vertically integrated network, around a quarter of its length is used by more than one train operator.

Europe—Freight Traffic

In Europe, as noted above, the situation was very different. Each country had a monopoly operator, and international traffic, which was always important in Europe, was handled by the operator simply handing the traffic over to the rail company of neighboring country at the border. The result was that international rail freight tended to be slow and unreliable.

In 1999, the European Commission issued a new railway policy document advocating separation of infrastructure from operations and the introduction of competition into operation of train services. The first step in this direction was introduced by way of Directive 91/990 in 1991. This required at least accounting separation of infrastructure from operations and passenger from freight, and the opening for new entry of some types of international freight. In subsequent directives, separation was to be taken further to provide independent management of infrastructure from operations. Many countries achieved this by completely separating infrastructure into a separate company or agency, although it and the major operator (at least of passenger services) remained publicly owned. (Drew and Ludewig, 2011). Others, notably Germany, Italy, Austria and later France, kept infrastructure and the main train operator as separate subsidiaries of a holding company. Only in Britain were both the infrastructure manager and all train operators privatized, although in Britain the privatized infrastructure manager became bankrupt and reverted to the public sector.

By 2007, complete freedom of entry into both domestic and international freight services for an operator based in any EU country which had the necessary license, which was to be granted solely on terms of safety of operations, was legally required. As well as non-discriminatory access to the main rail infrastructure—the tracks—entrants were given non-discriminatory rights of access to other facilities including terminals, marshalling yards and maintenance depots. New entry was slow, but by 2016 entrants had on average 39% of the European freight market (European Commission, 2019).

Some countries also sold their railway freight businesses. Indeed, there is strong evidence that horizontal separation (separation of passenger and freight operations) has led to lower costs. Invariably, these freight businesses passed into the hands of foreign railway companies. Thus, DB (German Railways) now owns the incumbent freight companies of Britain, Netherlands and Denmark, and OBB (Austrian Railways) that of Hungary. Only in Britain was there a decision to divide the incumbent into two (or more) separate companies and sell them separately to create more competition. Thus in Britain the container company Freightliner was sold separately, and in due course also entered the heavy haul market, whilst the bulk haulage company (now DB Cargo UK) entered the container market, providing competition for all types of traffic.

In general, it is considered that there is sufficient competition either within the rail sector or with other modes for rail freight now to operate without any regulation other than on safety grounds. The introduction of competition in the rail freight sector seems to have been largely successful, with the choice of operator leading to improved services and lower costs, although there is no clear evidence of an impact on mode split. An important issue here is that charging of road hauliers for the costs they impose, including externalities, whilst permitted under EU law is not required. Where road hauliers are not charged full marginal social cost, there is provision for offsetting grants to contribute to rail track access charges for competing rail services, and some countries—including Britain—make explicit use of this provision.

Europe—Passenger Traffic

Following the 1991 reforms, some countries took the opportunity to introduce competition into the passenger sector as well as freight, even before this was required by European Legislation (Nash et al., 2016). Thus, over the period 1994-97 Britain placed

virtually all its passenger services in 25 passenger franchises and let them to private companies by competitive tender (the state owned operator, British Rail, was not allowed to bid and thus disappeared as a train operating company). Sweden had already introduced competitive tendering for subsidized services and in 2010 began opening access for commercial services. Germany introduced open access for commercial services in 1994 and permitted states, who were responsible for all rail subsidies (though with funding from the federal government) to choose between competitive tendering and direct awards for subsidized rail services. These three countries are the only ones where new entrants hold a substantial share of the rail passenger market. Other countries, including the Netherlands and Denmark, practice a small amount of competitive tendering for noncore routes. Generally, the results of competitive tendering have been positive, with lower subsidies and better services.

Several other countries have seen some entry into the commercial passenger market. But it is Italy and the Czech Republic that offer the most dramatic examples. In Italy, NTV entered the market, offering a competing network of high-speed trains duplicating those of the incumbent, Trenitalia. In the Czech Republic, two entrants, Regiojet and Leo Express, entered the market competing with the incumbent Czech Railways on the most important route, gradually extending to the second key route and some international connections.

In general, open access competition has seen improvements in services and lower fares. But it has also led to duplication of services taking up scarce track capacity, and likely loss of revenue by incumbents meaning either increased subsidies, increased fares or cuts in services elsewhere. Many of the benefits of competition (without the associated problems) could potentially also be achieved by competitive tendering for franchises for commercial services, but Britain is the only country so far to have adopted this approach. However, in Britain the reforms have been accompanied by a significant increase in costs. In part this has been blamed on the fact that vertical separation has meant that the infrastructure manager and train operators each try to optimize their own costs rather than the system as a whole (although Britain does have the most complex system of track access charges and performance regimes in Europe to try to give all parties the correct incentives). There is wider evidence that vertical separation may increase costs on densely used systems (Mizutani et al., 2015), perhaps justifying placing infrastructure and the main train operator in the same holding company, which may then play a coordinating role, or implementing other measures such as alliances (agreements to work together on specific issues, with or without sharing revenues and costs) to try to overcome the misalignment of incentives. But the cost increase may also be partly the result of the way franchising was undertaken in Britain, with the existing services placed in a number of relatively large companies, control of which would be taken by the winner of the franchise competition. Thus there was no scope for new franchisees to bring their own workforce, wages and conditions. There is some evidence that the very small open-access operators have lower costs even than the winners of franchise competitions, although it is not clear whether this cost advantage would survive were open access competition to expand.

Legislation has now been passed that will completely open the rail passenger market for competition across the EU. Open entry for passenger services was first required on international routes in 2010 though with some protection for domestic services on the domestic parts of international routes. Under the 4th railway package, competition will be allowed for all commercial services from December 2019 and by means of competitive tendering for services operated under public service obligations by 2023. Again there are some exemptions; direct awards are still permitted if they can be justified to a relevant independent body such as the rail regulator, and open access for commercial services may be limited if it would adversely affect the financial stability of services run under a public service contracts. Moreover some countries have used the direct award of long (e.g. 10 year) franchises in the run up to liberalization to postpone the introduction of competitive tendering.

Europe—The New Regulatory Bodies

Whilst in North America, following the success of the Staggers Act, the general view was that regulation of the rail sector should be minimized, in Europe the reforms actually led to a need to strengthen regulation (Benedetto et al., 2017). The key reason for this difference was that North America continued to have competing vertically integrated railways, whilst in Europe competition took the form of competing train operators running over the same tracks, retaining a monopoly infrastructure manager. In most cases, the state owned the largest train operator as well as the infrastructure manager, and there were fears of discrimination favoring this train operator (obviously these fears were strengthened in those cases where the two were subsidiaries of the same holding company). Moreover, European governments were not willing to allow competition from other modes to determine what passenger services were provided, seeing political as well as economic, social and environmental reasons to subsidize unprofitable passenger services even when competing modes took much of their traffic. So the disciplinary power of inter modal competition in the passenger market was blunted and thus also the impact of this on the infrastructure manager (this is true in North America as well, but the importance of passenger traffic is very much less than in Europe). As a result, the new legislation required there to be an independent regulator, who should at least ensure that European legislation on issues such as track access charges was implemented and that non-discriminatory access (both in terms of prices and the quality of service offered to operators) was available for all. In addition, the regulator could be given the task of ensuring the efficiency of the infrastructure manager by using benchmarking to estimate efficient costs and ensuring prices were based on these (alternatively the efficiency of the infrastructure manager could be dealt with as part of a multi annual contract determining the infrastructure capability to be delivered and the funding to be made available for it between the infrastructure manager and the state).

In practice few countries made this part of the function of the regulator and only in Britain has the regulator undertaken extensive benchmarking studies as the basis for the approved prices. International benchmarking and regional benchmarking (comparing

different regions of the infrastructure manager against others) has been used, adopting statistical techniques as used by utility regulators, as well as bottom-up engineering studies. A key challenge has been addressing heterogeneity and obtaining good quality data. Further, given that the regulator's efficiency determination would impact on the infrastructure manager's funding and in turn on the call on public funds, transparent mechanisms had to be introduced to give the government an opportunity to de-specify outputs to fit within the available public subsidy.

When infrastructure is separated from operations track access charges become an essential part of the mechanism giving train operators the right incentives regarding how many trains to run, when and where to run them and what sort of rolling stock to use. In general, rail systems are subject to economies of traffic density, so charges that reflect marginal cost will not recover anything like total cost. On the other hand, prices that have significant mark-ups recover total cost but will discourage some traffic from using rail even when willing to pay marginal cost.

EU legislation requires infrastructure charges to be based on marginal cost (basically the cost of wear and tear on the track, plus congestion or scarcity where capacity is less than demand and environmental costs provided that the overall level of rail track access charges is not raised by them unless these costs are charged to other modes as well) but permits non-discriminatory mark ups where necessary to recover those costs not met by the state (Nash et al., 2018). These mark ups may be based on Ramsey pricing principles—in other words higher mark ups for traffic less sensitive to price—but may not vary between train operators for the same traffic. In practice, track access charges vary greatly between countries, with some countries such as Germany and France seeking to recover total cost (albeit with the state directly paying for investment in the case of Germany) and others (such as Sweden and Britain) basing them on marginal cost with the remaining costs being paid by the state (in Britain, there is a fixed charge paid by passenger franchisees but since this simply makes an equal reduction in the bids of those competing for franchises it may be viewed as the equivalent of being paid by the state). Importantly, there are also great variations in the extent to which charges vary by type of rolling stock and route (such variation is necessary if charges are to reflect accurately the costs caused by a particular train and therefore to give the correct incentives in terms of what trains to run and what types of rolling stock to use or for manufacturers to develop).

Finally, the level of access charges can also have an impact on the ability of competition to develop in the market, and this has been a policy issue in a number of EU countries.

Reform in Other Countries

In general no other countries have followed the European model, although some South American countries have implemented franchising for freight as well as passenger services, using long vertically integrated franchises (Thompson and Kohon, 2012), and Australia has a degree of vertical separation of infrastructure from operations, with open access for freight and some competitive tendering for passenger services. But the world's largest railway companies – those of Russia, China and India – have remained essentially vertically integrated monopolies with cross-subsidy of passenger traffic by freight. Indeed, these railways are still at the earlier stage of European reform; until relatively recently all were still run by a Ministry of Railways (Indian Railways still is) and required to run without subsidy despite tight regulation of fares and charges and of withdrawal of unprofitable services. There is currently little scope for new entry; although in India, entry is permitted into operation of containers services, entrants still have to hire locomotives and drivers from the incumbent so they are confined to planning and marketing the services and providing wagons and terminals. This change was in order to bring much needed investment into container terminals and rolling stock. In Russia some freight rolling stock was placed into the hands of a separate company that was privatized. Regulation in all three countries is essentially by a government department rather than an independent regulator, so a constraint on the growth of competition in the container market in India is that there is no independent body to whom the entrant may appeal if it believes it is being treated unfairly in terms of quality of service or even price. Although there has been some reform (for instance the transfer of Chinese Railways from the Ministry of Railways to a government owned company, China Railway Corporation) (Chen and Haynes, 2015) and more radical changes including vertical separation and introduction of competition have been proposed (Debroy Committee, 2015), reform to date in these countries has been limited.

The other interesting example of reform is Japan. Japan has always had a significant number of private railways offering regional and commuter services on a non-subsidized basis but earning a significant share of their revenue from property. However, until 1987, the main lines were all part of a national government owned vertically integrated company. In that year, the government railway was divided into 6 vertically integrated passenger companies, and 3 of these have subsequently been privatized through sale of shares (Ishida, 2011). There is still an element of fares regulation using benchmarking methods, though fares rarely change in Japan. In recent years, declining traffic for regional services due to population trends has led to some Japanese railways becoming unprofitable, thus leading to a form of "vertical separation" where the government takes financial responsibility for the infrastructure (though leases it back to the operator).

Conclusions

In the first century of its existence, the rail industry moved in many countries from a system of competing private railways to a state-owned monopoly. The principle exception to this was in North America, where private companies remained, whilst elsewhere some

rail systems had always been state owned. During this period, the railway was seen as having strong monopoly power, and was often heavily regulated, with strict controls on tariffs and a requirement to continue to operate unprofitable services and routes.

The growth of road transport as a competitor made this position unsustainable in most countries. In North America regulation was largely removed in 1980 under the Staggers Act on the basis that continued competition between privately-owned railways, with road haulage and with alternative sources of supply, made it unnecessary. US deregulation appears to have been very successful, leading to a big reduction in costs and freight rates. In Europe, earlier regulation by a separate regulator was generally removed as passenger services came to be mainly operated under public service obligations, whilst freight faced intensive competition from road. However, this left unregulated public monopolies about which there was much concern regarding efficiency. In 1991 the European Commission started a new policy of separating infrastructure from operations and introducing competition within the rail sector by opening entry to new train operating companies to compete with the incumbent either in the open market or on the basis of competitive tendering for public service contracts. This actually produced a new requirement for regulation, in that a system was needed to ensure non-discriminatory access and to check the efficiency of the monopoly infrastructure manager. Whilst the introduction of competition was generally seen as successful in improving services and reducing costs, there have been concerns that creating separate infrastructure managers and train operating companies means that there is no-one seeking to optimize the system as a whole, although track access charges and performance regimes may give some of the right incentives. There is evidence that vertical separation raises systems costs on more densely used railways because of increased challenge of dealing with co-ordination issues on busier networks.

Elsewhere, reform followed different patterns. For instance in South America, it became common to competitively tender long vertically integrated franchises for freight as well as passenger transport, whilst Australia adopted aspects of the European model. Japan split its national operator into vertically integrated companies and privatized the more profitable. Only a small number of countries, but including the very important rail systems of Russia, India and China, have so far stuck largely with the old model of tightly regulated government-owned monopolies.

Geography and the extent of demand are clearly relevant factors in this diversity of experience. For example, in Japan, population density in the major conurbations is such that passenger flows can be served by multiple, vertically-integrated routes. The dominance of a particular traffic type at high intensity, such as passenger traffic in Japan, supports the approach of having vertically-integrated passenger operations, whilst permitting access to freight services run by a separate company. Similarly, in the US the dominance of freight traffic supports vertically-integrated freight operations, with passenger traffic gaining access to these networks where required. Mixed networks such as those in Europe may warrant a different approach, where some form of vertical separation to support competition between operators on the same network, may be more appropriate, although careful consideration is needed of how to overcome problems of lack of coordination and misalignment of incentives in such a system.

It does therefore appear that there is no single model of rail reform that will clearly be preferable in all circumstances; a deep analysis of the circumstances of any individual country is needed to decide the best way forward.

References

- Benedetto, V., Smith, A.S.J., Nash, C.A., 2017. Evaluating the roles and powers of rail regulatory bodies in Europe: a survey-based approach. *Transport Policy* 59, 116–123.
- Chen, Z., Haynes, K.E., 2015. *Chinese Railways in the Era of High Speed*. Emerald Books, Bingley.
- Debroy Committee, 2015. Report of the Committee for Restructuring of Railway Ministry and Railway Board. Ministry of Railways, New Delhi.
- Drew, J., Ludewig, J. (eds), 2011. *Reforming Railways: Learning from Experience*. Community of European Railways, Brussels.
- European Commission, 2019. 51 Final report from the Commission to the European Parliament and the Council. Sixth report on monitoring development of the rail market (SWD(2019) 13 final), Brussels.
- Ishida, Y., 2011. 'Japan'. In: Drew, J., Ludewig, J. (Eds.), *Reforming railways: lessons from experience*. Eurail Press, Hamburg, pp 22–31.
- Mizutani, F., Smith, A.S.J., Nash, C.A., Uranishi, S., 2015. Comparing the costs of vertical separation, integration, and intermediate organisational structures in European and east Asian railways. *J. Transp. Econ. Policy* 49 (3), 496–515.
- Nash, C., et al., 2016. *Liberalisation of rail passenger services*. CERRE project report, Brussels.
- Nash, C., et al., 2018. *Track access charges: reconciling conflicting objectives*, CERRE project report, Brussels.
- Thompson, L.S., Kohon, G.C., 2012. Developments in rail organization in the Americas, 1990 to present and future directions. *J. Rail Transp. Plann. Manag.* Volume 2 (Issue 3), 51–56.
- Winston, C., 2006. The US: private and deregulated. In: Gómez-Ibañez, J., de Rus, G. (Eds.), *Competition in the Railway Industry*. Edward Elgar, Cheltenham, pp. 135–152.

Further Reading

- Nash, C.A., 2013. Rail Transport. In: Matthias Finger, Torben Holvad, (Eds.), *Regulating Transport in Europe*. Edward Elgar, Cheltenham.
- Pittman, R.W., 2005. Structural separation to create competition? the case of freight railways. *Rev. Netw. Econ.* 4, 181–196.
- Thompson, L.S., 2003. Changing railway structure and ownership: is anything working? *Transport Rev.* 23 (3), 311–355.
- Smith, A., Benedetto, V., Nash, C., 2018. The impacts of economic regulation on the efficiency of European railway systems. *J. Transp. Econ. Policy* 52 (2), 113–136.

Cycling Economics

Bert van Wee, Delft University of Technology, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

The Increasing Importance of Cycling in Practice and Research	414
Which are the Costs and Benefits of Cycling Policies?	414
Pitfalls in the Evaluation of Cycling Policies	415
Costs	415
Models	415
Accessibility	416
Wider Economic Effects	416
Environmental Effects	416
Health Effects	416
Safety Effects	417
Discussion	417
Biography	417
Further Reading	418

The Increasing Importance of Cycling in Practice and Research

Since around the mid of the first decade of 21st century, cycling is booming in many cities and regions in the world, for example, cities such as New York, London, Paris, Santiago (Chile), and Muenster (Germany). Other cities, such as, Copenhagen and Amsterdam have a much longer cycling tradition. The increasing interest of policy makers and planners in cycling has resulted in a rapid increase in the number of academic papers in this area and several of these papers have high download and citation rates. As an illustration, of the ten most cited papers in the journal *Transport Reviews*, eight are about cycling.

Most papers are about the effects of cycling options like infrastructure provision or bike sharing systems on cycling levels. Despite this increased interest, the literature on the economics of cycling is limited. A possible explanation might be that several cycling policies are relatively cheap. Another reason could be that at first face there is no academic added value in discussing the economics of cycling, because all major cost and benefit categories are comparable to those of other transport projects (see next section). Another reason could be that decisions on cycling policies are generally made at the level of the local municipality and cities and towns do not often finance academic research, unlike national governments. A next reason could be that the variability in cycling over time can be relatively large, compared to traveling by car or train, due to season and weather effects. A final possible reason is that many cycling policies aim to improve comfort and safety, and these effects are more difficult to assess than travel time gains.

This chapter aims to discuss the economics of cycling and has one major conclusion: in estimating the economics of cycling one can easily, "do it wrong." The emphasis is on infrastructure policies, but several of the conclusions also apply to other cycling policies, such as promotion campaigns or fiscal policies. In addition, the focus is on western countries because those dominate in the academic literature in this area. The problems and pitfalls discussed, apply even more for developing countries.

As a point of departure, it is assumed that the welfare effects are calculated via a social cost-benefit analysis (SCBA). An SCBA is an overview of all (dominant) pros and cons of a project, in quantitative terms, and next in monetary values. Final results are expressed as indicators like "benefits minus costs" or the "benefit-cost ratio." SCBA is not further explained in this chapter.

Which are the Costs and Benefits of Cycling Policies?

Table 1 summarizes the dominant costs and benefits of cycling policies and the state of the art with respect to knowledge about the economics of each category of costs and benefits, as well as the main risks in their estimations.

The overall picture that emerges is that, at first face an estimation of the costs and benefits of cycling projects is about comparable to those of other transport infrastructure projects like road and public transport projects. The categories of costs and benefits are about the same, and the estimation of welfare effects can be based on the same methods as in the case of roads and public transport projects. But the state of knowledge on cycling projects really lags behind relative to those other projects. The quality of models is generally poor, there is very limited insight in the quantitative and even more so the monetary effects of effect categories, and as far as methods are available and applied there are substantial risks in estimating effects adequately, both quantitatively and in monetary terms. We later discuss some dominant risk categories.

Table 1 Current knowledge and main risks for dominant categories of costs and benefits

	<i>Current knowledge</i>	<i>Main risks</i>
<i>Costs</i>		
Value of land needed for cycling infrastructure, construction costs.	• Generally poor, at least in the academic literature. • Some knowledge for larger projects, especially if previously built examples are available, which applies to cities and countries with a cycling tradition. • If cities implemented comparable projects before: quite well known because of experience.	• Risk of underestimating costs - worldwide transport infrastructure projects often face substantial cost overruns.
Maintenance costs	• Maintenance costs are relatively limited.	• Uncertain, but small, so not very important.
<i>Benefits</i>		
Travel behavior related impacts—quality of models	• Models for cycling are quite poor.	• Relying on “bad models”—many effect estimates rely on (model based) forecasting.
Travel time of cyclists	• Only a few studies on values of time of cyclists.	
Comfort impacts	• Poor.	• Risk of completely overlooking these benefits.
Travel times of other road vehicles	• Values of time quite well known.	
Other accessibility effects (social exclusion effects, option values)	• Poorly understood, generally ignored.	• Risk of completely overlooking these benefits.
Wider economic impacts	• Poorly understood.	
Environmental impacts – emissions related	• Cycling emissions per km zero, emissions due to increased exercise (and food intake), and for producing bikes and infrastructure.	• Emissions for the construction of infrastructure generally ignored.
Other environmental impacts	• Poorly understood, certainly not quantitatively and in monetary terms.	• Exclusion or Pro Memory in evaluations and therefore not seriously considered.
Health impacts	• Often substantial. Rapidly increasing body of knowledge, and even tools for health impact assessment. • But still large uncertainties.	• Some risk of double counting as far as health is included in travel behavior decisions. • Interaction between cycling and other forms of physical activity. • Self-selection effects.
Safety impacts	• Poor data on incidents in many countries. • Often only aggregate risk factors.	• Incorrect use of aggregate risk factors. • Ignorance of the safety-in-numbers effect.

Pitfalls in the Evaluation of Cycling Policies

Costs

There is limited information, at least in the academic literature, on costs of cycling projects, and the quality of cost estimate. Roads and rail projects often face substantial cost overruns, so estimations for cycling projects could also be biased. But the importance of the costs of the common cycling infrastructure projects is relatively limited in absolute terms because cycling projects are generally relatively cheap. This also applies relative to the benefits (benefits of bike projects are often way higher than the costs) and relative to the volume of kilometers traveled. Therefore, despite the uncertainty, the importance of costs and cost estimates is not as large as in case of (major) roads and rail projects.

Maintenance costs of cycle infrastructure projects probably are relatively small, first because the construction of cycling infrastructure is relatively cheap, and secondly because wear and tear due to the use of infrastructure by bikes is very limited. Even on motorways the impact of cars on wear and tear is limited—it is the lorries with high axle loads (especially due to overloading) that cause almost all wear and tear of motorways.

Models

The estimation of the benefits, and next their welfare effects, to a large extent relies on forecasting changes in travel behavior. This applies to travel time related impacts, other accessibility related impacts, such as, health, safety, and environmental impacts. A major problem is that cycling is generally poorly included in transport models, even in countries with cycling tradition. The poor inclusion in models might first of all result from the fact that in most countries and cities world-wide cycling is way less common than driving or even than traveling by public transport. In addition for decades transport planning has adopted the predict-and-provide paradigm, at least for roads: forecast travel behavior, explore where infrastructure capacity is not sufficient, and provide new capacity. This paradigm resulted in a focus on (capacity) problems, more than on a focus on challenges, accessibility in general, and active modes.

Accessibility

Changes in travel times and the generation of new travel are important benefit categories of transport infrastructure projects in general. This also applies to several cycling infrastructure projects. The economics of estimates are relatively easy: travelers who face reductions of travel times generally evaluate those positively. Multiplying the number of trips that travelers make by travel time reductions per trip results in an estimation of the travel time saved, and next the multiplication of travel time saved by the marginal value of time results in the monetary value of those travel timesavings. The benefits of generated travel are generally estimated via the so-called rule of half: new travelers on average face travel time gains that are half of the full travel time gains of the infrastructure projects.

This methodology also applies to cyclists. But practice is not as simple as theory. Not only the absence of adequate models is a problem, there is also limited insight into the marginal value of time of cyclists. The studies that are available suggest that these values are generally quite high, higher than those for car and public transport users.

But cycling projects can also influence values of time of other road users, both positively (e.g., if more capacity for other vehicles becomes available because some drivers switch to the bike) and negatively (e.g., because road space is reallocated to bikes). Here the problem of poor models, and consequently poor estimates of (changes in) travel times and mode choice impacts of other road users are also relevant. The marginal values of time of other road users are generally quite well known. A risk may be that cycling projects mainly influence other road users on short urban trips, and the marginal value of time for short passenger trips generally is lower than for longer trips, so the risk is an overestimation of the value of these travel time changes for other road users.

But not all accessibility related effects of bike projects are related to the travel times of cyclists (and others). At least two other accessibility effects should be added, the first one being social exclusion, referring to the fact that some people have poor access to important destinations (work, shops, schools, medical services, etc.), often because they have no car available, because of a low income, and/or because of poor public transport services. Most studies on reducing levels of social exclusion via transport policies focus on public transport services, but the bike can be a relatively cheap, flexible, and healthy alternative. Estimating the welfare effects of lower levels of social exclusion is difficult. There are academics arguing that the willingness to pay of the people being socially excluded is not the right way to estimate those benefits. An alternative could be to estimate the value of lower levels of social exclusion by looking at policy measures (not) implemented to reduce these levels, and their costs and social exclusion effects. This results in the willingness to pay of politicians to reduce levels of social exclusion.

A next potentially relevant effect category could be option values: people might value the availability of options even if they do not use them. And they might be willing to pay for this availability — this at least was found in a few studies on public transport services: car users are willing to pay for having the option available even if they do not use these services.

Wider Economic Effects

Wider economic benefits are often-debated category of impacts of transport projects in general, and the uncertainty in the estimations is still large. For cycling projects, there is hardly any literature on these effects.

Environmental Effects

Emissions effects (e.g., CO₂, NO_x, PM) are generally estimated by multiplying travel behavior changes for each mode-by-mode specific emission factors. The poor quality of models also influences the estimation of emission effects, because travel behavior changes are difficult to forecast. Emission factors for other modes than the bike are generally available but should be selected carefully. For example, average emission factors for all car kilometers do not necessarily apply to those (urban) trips that are influenced by bike policies. Emissions due to cycling are zero, but cycling takes more energy than driving or traveling by public transport. If cycling would result in additional exercise and food intake emissions due to producing food are relevant, but there is only very limited knowledge in this area. Emissions of (changes in) travel behavior on noise and noise effects and their monetary valuation are quite well understood, so the main problem is the quality of transport models.

But cycling has more environmental effects: cities that prioritize active modes over cars are often considered to be more attractive. Terms like livability, the quality of the urban environment, and spatial quality are sometimes used to label these effects. The literature on these effects, including the monetary valuation of these effects, is very limited, although some studies have tried to estimate some of those effects via housing prices.

Health Effects

Health effects of cycling projects are often very substantial, and in some cases even the dominant benefit category. To some extent people might include health in their travel behavior decisions, so there is a risk of double counting if health benefits would be fully added to those related to travel behavior benefits and their valuations. But because according to some studies, the health benefits are often way larger than the travel behavior benefits, the size of double counted health benefits must be relatively small.

A large problem is probably that we poorly understand the interaction between cycling and other forms of physical activity. Do people who cycle more, substitute other forms of physical activity for cycling? Or do they exercise more, because they have a better condition, and, for example, take the stairs and not the elevator? In other words, cycling can replace, be added to, or even increase other forms of physical activity. But there is very limited knowledge on the interaction between cycling and other forms of physical activity.

Another problem for the estimation of health effects could be self-selection effects, it could be that specific categories of people more than average are inclined to cycle, for example, people who more than average think health is important, have a healthy

lifestyle in general, or who are fit because of genetic reasons. Health effects based on cross-section data of categories of people (cyclists and others) can therefore easily overestimate the health effects of (more) cycling.

Safety Effects

At first face estimating safety effects of travel behavior changes seems to be easy: estimate kilometers per mode, and multiply these with risk factors. But several problems emerge, the first one being the poor quality of models, and thus the estimation of travel behavior changes, as discussed earlier. Second, aggregate national risk factors are often not available for cycling because of the bad quality of registrations of bike incidents, and even bike levels. And if these are available, aggregate values do a poor job for specific travel behavior changes. For example the risks for car drivers on motorways are relatively low, but bikes do not or hardly compete with car trips for which motorways are used. Excluding motorways increases risk factors for cars. The third problem, from the perspective of adequate safety effect estimations, is that cycling becomes safer if cycling levels increase. This is referred to as the, *safety-in-numbers effect*. Explanations for this effect as mentioned in the literature include first that cyclists become more experienced, and that other road users become more used to cyclists. Second, road authorities provide better and safer infrastructure for cyclists the more people cycle. Note that safety effects might be partly included in people's value of time just as health effects, if, for example, people perceive cycling as relatively unsafe, they might be willing to pay more for shorter travel times, for example, due to a new cycle line.

Discussion

The overall picture that emerges is that at first face estimating the welfare effects of cycling seems straightforward and comparable to the estimation of the welfare effects of other transport infrastructure projects. Effect categories and methods are about the same. But practice is more problematic, due to the absence of adequate models, the lack of high quality data, very limited insights in several quantitative effects and their monetary valuation (like the impact of cycling on the urban environment), and difficulties in the estimation of some other effects, in particular health and safety effects.

In theory cities could learn from other cities that implemented cycling policies. But one has to be careful—context matters and this limits the transferability of results, at least across cities and countries with a different cycling tradition. To give a few examples, whereas in more egalitarian countries cycling is not linked to specific categories of people, it might be in other countries. And a hilly area could negatively influence utilitarian cycling, but positively some forms of recreational cycling. Also substitution of drivers or public transport users to cycling depends on the quality of alternatives and current use levels, and these factors have an impact of cross-elasticities.

How important is the general lack of knowledge for practice? First of all, not all topics discussed are equally important. The lack of adequate models, and the poor estimation of safety and health effects probably are most important. Second, in general the problem seems to be somewhat limited compared to road and rail projects. Cycling projects are often very cheap, and have high benefit-cost ratios. So, despite the uncertainty they are often no-regret. In several cases assessments might not be needed at all, cities can “just do it.” And cycling projects could be implemented via guidelines for road design or the design of urban environments, like in the Netherlands. Then an assessment for each project is not needed—the guidelines are used for planning, not stand-alone assessments of candidate projects. Of course these guidelines preferably should be underpinned using the current stage of academic knowledge.

A specific problem emerges with respect to safety; sometimes (expected) safety effects are barrier for the local municipalities to implement cycling policies. If for that reasons they would not implement these policies, which would have contra-productive effects, even for health. Some studies have compared the negative impacts of higher cycling levels on health because of incident risks, exercise, and the intake of pollutants. Those studies generally found that the positive impact of more exercise on health is way larger than the negative impact of incidents (and the intake of pollutants).

Biography



Bert van Wee is professor in Transport Policy at Delft University of Technology, the Netherlands, faculty Technology, Policy, and Management. In addition, he is scientific director of TRAIL research school. His main interests are in long-term developments in transport, in particular in the areas of accessibility, land-use transport interaction, (evaluation of) large infrastructure projects, the environment, safety, policy analyses, and ethics.

Further Reading

- Banister, D., 2002. *Transport Planning*, third ed. Spon Press, London/NewYork.
- Barnes, G., Krizek, K., 2005. Tools for Predicting Usage and Benefits of Urban Bicycling, Humphrey Institute of Public Affairs, University of Minnesota. Available from: www.lrrb.org/pdf/200550.pdf.
- Börjesson, M., Eliasson, J., 2012. The value of time and external benefits in bicycle appraisal. *Transp. Res.* 46, 673–683.
- Cavill, N., Kahlmeiers, S., Rutter, H., Racioppim, F., Oja, P., 2008. Economic analyses of transport infrastructure and policies including health effects related to cycling and walking: a systematic review. *Transport Policy* 15, 291–304.
- Elvik, R., Bjørnskau, T., 2017. Safety-in-numbers: a systematic review and meta-analysis of evidence. *Saf. Sci.* 92, 274–282.
- Flyvbjerg, B., Ansar, A., Budzier, A., Buhl, S., Cantarelli, C., Garbuio, M., Glenting, C., Skamris Holm, M., Lovallo, D., Lunn, D., Molin, E., Rønneest, A., Stewart, A., Van Wee, B., 2018. Five things you should know about cost overrun. *Transp. Res.* 118, 174–190.
- Litman, T., 2014. Evaluating active transport benefits and costs: guide to valuing walking and cycling improvements and encouragement programs. *Transp. Policy* 10, 339–342.
- Lucas, K., 2012. Transport and social exclusion: where are we now? *Transp. Policy* 20, 105–113.
- Pucher, J., Böhler, R. (Eds.), 2012. *City Cycling*. MIT Press, Cambridge/London.
- Sælensminde, K., 2004. Cost-benefit analyses of walking and cycling track networks taking into account insecurity, health effects and external costs of motorized traffic. *Transp. Res.* 38, 593–606.
- Stipdonk, H., Reurings, M., 2012. The effect on road safety of a modal shift from car to bicycle. *Traffic Inj. Prev.* 13, 412–421.
- Van Wee, B., Börjesson, M., 2015. How to make CBA more suitable for evaluating cycling policies. *Transp. Policy* 44, 117–124.
- Van Wee, B., Ettema, D., 2016. Travel behaviour and health: a conceptual model and research agenda. *J. Transp. Health* 3, 240–248.
- WHO, 2014. Health economic assessment tools (HEAT) for walking and for cycling. Methods and user guide, 2014 update. Available from: <http://www.euro.who.int/en/health-topics/environment-and-health/Transport-and-health/activities/guidance-and-tools/health-economic-assessment-tool-heat-for-cycling-and-walking>.
- Woodcock, J., Givoni, M., Scott Morgan, A., 2013. Health impact modelling of active travel visions for England and Wales using an integrated transport and health impact modelling tool (ITHIM). *PLOS one* 8, e51462.

The Economic Rationale for High-Speed Rail

Ginés de Rus, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain; University Carlos III de Madrid, Madrid, Spain; FEDEA, Madrid, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	419
The Cost of a HSR Line	420
The Benefits of HSR	420
A Final Assessment	422
References	423
Further Reading	424

Introduction

The construction of high-speed rail (HSR) infrastructure is a significant component of the transport policy in many countries all over the world. From France and Germany to Japan and China. By HSR technology we mean the provision of rail services capable of speeds above 250 km/h on dedicated tracks (Albalade and Bel, 2017).

Governments of different political orientation are investing in HSR infrastructure, assigning significant amounts of public funds for its development with the declared aim of increasing the market share of railways. The demand for the new rail transport alternative comes from conventional trains, air, and road transport, and also from generated traffic, thanks to the combination of time savings, increased comfort and safety, and last but not least, subsidized prices.

The declared objectives for the implementation of HSR vary and goes from time savings, comfort and reliability to the reduction of global warming, and other negative externalities of competing modes of transport, regional development, and the like. In some countries the enthusiasm is more intense than in others. There is no a single pattern. The United Kingdom and the United States are now closer to building HSR infrastructure but until now they have been reluctant to give the definitive approval. Other countries, like France and Spain, have been keener on HSR than other European countries, like Norway or Sweden (de Rus, 2012). Spain is a peculiar case because with much less traffic density in the conventional rail network, it is the second country in the world measured in HSR kilometers, with the help of European Union (EU) funds.

Governments, the industrial lobbies, and the users, support this technological option for passenger travel. The price that users pay for HSR infrastructure and services in many routes is far from covering construction, maintenance, and infrastructure operation costs. This means that taxpayers contribute financially to the support of this transport option.

The question is not whether users, and other possible beneficiaries of HSR, would vote in favor of the construction of HSR lines. The question is whether they would be willing to pay (independently of what they actually pay) their social costs. The answer to this question varies widely depending on the local characteristics of the project: routes, urban zones crossed, bridges and tunnels required, volume of demand, per capita income, and the level of use of the capacity in competing modes.

The economic appraisal of a HSR project requires identifying the benefits and costs during its life, and calculates the social net present value. The cost-benefit analysis of HSR requires to compare the project with some “do something” alternatives. A realistic range of options has to be examined. The base case should be a “do minimum” and other likely investments should be examined as alternative “do something” options. These alternatives should be compared on an incremental basis to see whether the additional cost of moving to a more expensive option is justified. It is important to consider the main alternatives to be sure that the best alternative has been selected (de Rus, 2011).

A new HSR line in a medium distance interurban corridor changes the modal split, and this change will occur regardless of the social value of the HSR project. The final impact will heavily depend on the pricing policy of the government. The economic evaluation of HSR investment cannot be carried out without first determining which prices are going to be charged for HSR services. This is a key issue, as the market shares are very sensitive to pricing decisions taken in the public sector affecting the playing ground for intermodal competition. The public pricing decision regarding construction and maintenance costs of HSR infrastructure is crucial for the competitiveness of the railway operators, the intermodal equilibrium, and eventually the result of the appraisal of new lines.

There is considerable pressure on governments to build new HSR lines as if the investment were a kind of “now or never” decision. This does not seem to be the case with this technology. The construction of HSR infrastructure is irreversible and there is uncertainty associated with costs and demand. In these conditions the question of the right moment to invest is critical as the investment can be postponed in most cases. Hence, the optimal timing of the investment should be addressed in the case of a positive net present value. It could be profitable to build a line today and another in the future. Moreover, it is feasible to build a HSR rail track on parts of the overall line and use it for traditional trains at the same time as it is prepared for HSR services, which would operate once demand motivates building new tracks on missing links. There exist several “do something” alternatives (de Rus and Nash, 2007).

The Cost of a HSR Line

The cost of building and operating a HSR line consists of the construction and maintenance of the infrastructure, the acquisition, operation, and maintenance of the rolling stock and the external costs. The construction requires a specific design aimed at the elimination of all those technical restrictions that may limit the commercial speed below 250 km/h. These basically include roadway-level crossings, frequent stops or sharp curves unfitted for higher speeds so that, in some cases, new signaling mechanisms and more powerful electrification systems may be needed, as well as junctions and exclusive track ways in order not to share the right-of-way with freight or slower passenger trains, when there is joint use of the infrastructure.

Infrastructure costs include investments in construction and maintenance of the tracks including the sidings along the line, terminals, and stations at the ends of the line and along the line, respectively, energy supplying and line signaling systems, train controlling, and traffic management systems and equipment, etc. Infrastructure maintenance and operating costs include the costs of labor, energy, and other materials consumed by the maintenance and operation of the tracks, terminals, stations, energy supplying, and signaling systems, as well as traffic management and safety systems. Some of these costs are fixed, and depend on operations routinely performed in accordance with technical and safety standards. In other cases, as in the maintenance of tracks, the cost is affected by the traffic intensity. Similarly, the cost of maintaining electric traction installations depends on the number of trains running during operation.

Railway infrastructure also requires the construction of stations. Although sometimes it is considered that the costs of building rail stations, which are usually singular buildings with expensive architectonic design, are above the minimum required for technical operation, these costs are part of the system and the associated services provided affect the generalized cost of travel (e.g., quality of service in the stations reduces the disutility of waiting time). To these fixed costs we have to add the acquisition costs of the rolling stock and the operating and maintenance costs of running the trains.

There are other costs involved in a HSR project. For example, planning costs are associated with the technical and economic feasibility studies carried out before construction. These fixed costs, as well as those associated with the legal preparation of the land (expropriation or acquisition from current landowners), are not usually included in the published construction cost per km. There are other costs difficult to allocate to infrastructure or operation, as general administration, marketing, and internal training.

The operating costs of HSR services (train operations, maintenance of rolling stock and equipment, energy, and sales and administration) vary across rail operators depending on traffic volumes and the specific technology used by the trains. In the case of Europe, almost each country has developed its own technological specificities: each train has different technical characteristics in terms of length, composition, seats, weight, power, traction, tilting features, etc.

One of the reasons supporting the case for investing in HSR infrastructure is the alleged positive net environmental effect associated with the operation of HSR services once the intermodal substitution effects are included. HSR trains attract passengers from road and air with higher environmental externalities and, when the deviation of traffic is substantial, as it potentially the case in medium-distance corridors, the net effect on the environment of investing in HSR might be positive.

The environmental externalities of HSR point in two directions. One is positive, due to its substitution effect on air and road traffic, although it requires a significant deviation of passengers from these modes as well as high load factors in the HSR to offset the pollution associated with the production of electric power consumed by high speed trains. The other is negative: HSR lines need land, crossing areas of environmental value. The rail track creates a barrier effect in the affected territory, produces noise, and generates visual intrusion on the landscape.

The net environmental impact will depend on the reduction of environmental externalities in other modes of transport that lose traffic in favor of the new mode. Moreover, the emission of greenhouse gases during the construction period is substantial and it may take decades of HSR operation to compensate for the emissions caused by construction. The net effect depends on the value of the affected areas, the number of people affected, the benefits from diverted traffic, and so on. The net impact is very difficult to assess, without contemplating the circumstances in a particular corridor, and a key element is the volume of demand attended by the HSR.

The Benefits of HSR

Leaving aside income transfers, HSR generates different social benefits, basically timesavings, higher reliability, comfort and safety, and the reduction of congestion and accidents from alternative modes. When users shift to HSR from buses, cars, or airlines, some avoidable costs add to total benefits. Other benefits could be added to the list though they are more controversial, as the spatial effects (tunnel effect of HSR in contrast with the corridor effect of conventional rail) or the agglomeration benefits.

The direct benefits of HSR investment come from existing passenger-trips using conventional railway services in the corridor, the deviated demand from other transport modes and the induced passenger-trips after the reduction of the generalized cost of travel. The door-to-door user time invested in a trip includes access and egress time, waiting time, and in-vehicle time. The total user timesavings will depend on the conditions in the initial transport mode.

The case of road transport is quite different with access-egress and waiting time, playing a decisive role in the generalized costs and eventually in modal choice. In the case of a HSR line of a 500–600 km length, car passengers shifting to HSR benefit from travel time savings but they lose with respect to access, egress, and waiting time. Benefits are higher than costs when travel distance is long enough, as HSR runs, on average, more than twice as fast as the average car. As the travel distance gets shorter the advantage of the HSR diminishes as in-vehicle time loses weight with respect to access, egress, and waiting time. Nevertheless, in choosing between car

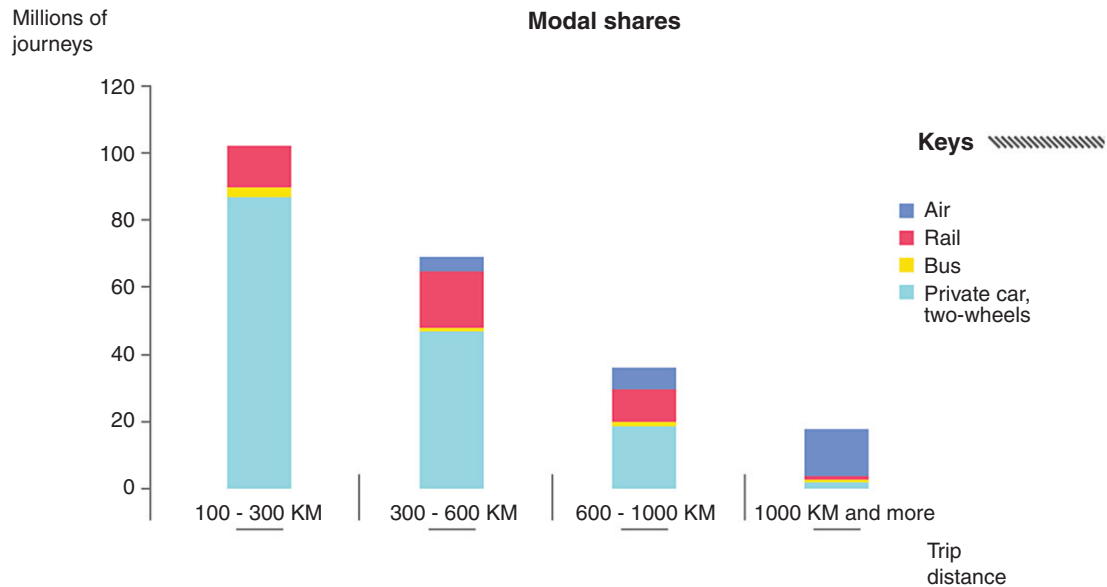


Figure 1 Market shares of rail, road, and air in France. Source: UIC, 2018.

and HSR, a key factor could be whether the traveler will need a car at his destination. This, in turn, could depend on trip purpose and the availability of mass transit at the destination. Similarly, the number of people traveling together could matter as the marginal cost of a second person traveling in a car that is already making the trip is near zero. Moreover, it is usually assumed that trip quality is higher for HSR than for auto travel. In some ways, that may be true, but not in all ways. For example, one can stop when and where one likes and it is easier to carry luggage with oneself if traveling by car.

In the case of air transport (Figs. 1 and 2), the situation is also different. In-vehicle time is longer by train but it is assumed that the user saves access, egress, and waiting time, and enjoys additional service quality. The evidence is that the money value of time of access-egress and waiting time are of the order of 2 and 1.5 times higher, respectively, than in-vehicle time and hence, even in the case of longer door-to-door travel time after the shift, the value of time savings of a passenger shifting from air to rail could be positive, as far as the values of time of access-egress and waiting time are high enough to compensate the losses in-vehicle time.

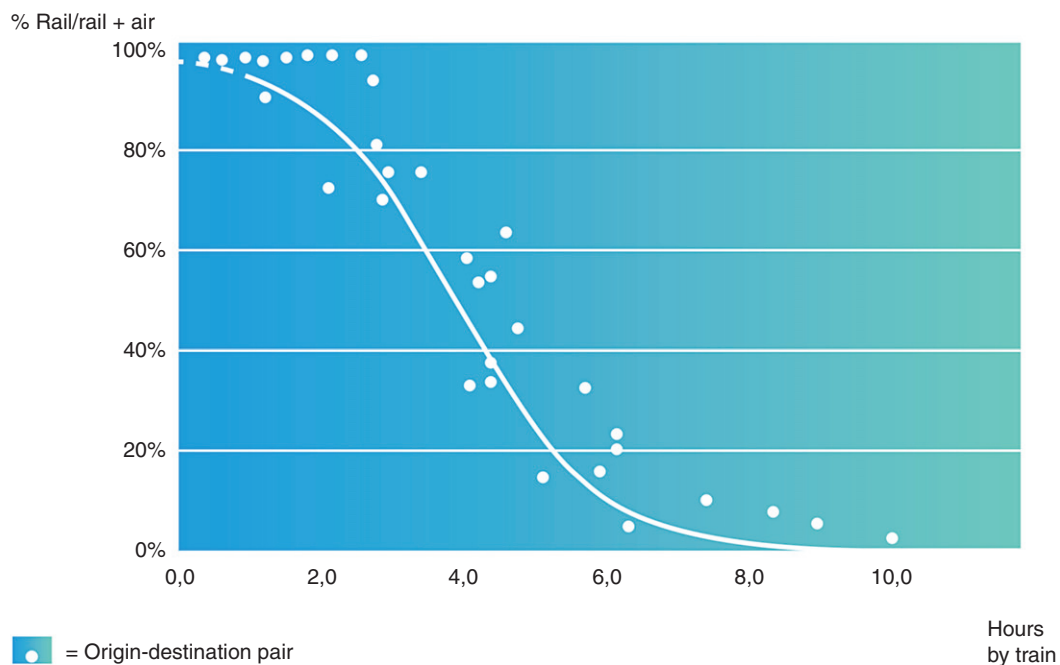


Figure 2 Rail market share on the rail + air market in France (passengers). Source: UIC, 2018.

Benefits also come from induced demand. The new passenger-trips of HSR may come from different sources: completely new generated trips; trip redistribution, i.e., trips that were made to a different destination without the project; diverted demand from other modes of transport; and finally, route or time reassignment. In the three first cases the number of passenger-trips increases for the project and so we have to account for the willingness to pay of both the deviated and the generated demand.

Promoters of HSR projects tend to emphasize the indirect effects, wider economic impacts, and regional development instead of the direct effects. It is true that transport investments produce other alleged benefits beyond the direct benefits already discussed, but it is unclear whether these additional benefits are new ones or double counting. Moreover, some genuine indirect benefits exist but in many cases they could be associated with any other infrastructure project in general, and not exclusively with high speed, so even if the benefits increase the social return on the investment in transport, they do not necessarily place HSR in a better position over other relevant alternatives (Vickerman, 2009; Preston, 2017).

Should we worry about the wider economic benefits in the case of HSR investment? Additional benefits are not expected to be very important in the case of HSR infrastructure. The reason is that freight transport does not benefit from high speed and therefore the location of the industry is not going to be affected by this type of technology. Moreover, in the case of the service industry, HSR may lead to the concentration of economic activity in the core urban centers.

Recent research suggests that agglomeration benefits in sectors such as financial services may be greater than in manufacturing. This is relevant to the urban commuting case but arguably it is also important for some HSR services (e.g., the North European network which links a set of major financial centers and may be used for a form of weekly commuting). It may be erroneous to conclude that scale economies and agglomeration economies (productivity impacts) are only found in manufacturing and freight transport.

Intermodal effects could be one of the main benefits of HSR. The critical issue is whether price is higher or lower than marginal social cost in the alternative mode of transport. When price is below marginal cost in the original transport mode, society benefits from the diversion of demand to the new transport mode (assuming price is at least equal to marginal cost in the new mode). This could happen because of the reduction of excessive congestion, or pollution. In the case of a positive externality the opposite might occur, and the indirect effect could be negative when the price is above the social marginal cost in the original transport mode, for example, if the reduction of demand in the original transport mode forces the operators to reduce the level of service (e.g., frequency), thereby increasing the generalized cost of travel.

The key point is whether the original transport mode was optimally priced. Although it has been argued that the reduction of road and airport congestion is a positive effect of HSR, this is only the case with suboptimal pricing. When road and airport congestion charges internalize the external marginal costs, there are no indirect benefits from the change in modal split. This can be viewed from another perspective. The justification of HSR investment based on indirect intermodal effects should be first compared with a “do something” approach, consisting of the introduction of optimal pricing.

Intercity bus services operate under concession contracts in many countries and so although they keep the regulated timetables unchanged in the short run, the financial difficulties will emerge later, leading to contract renegotiations. Eventually, users and/or taxpayers (or workers) will have to pay for the adjustment in the medium-term.

Airlines operate in open competition and the external shock in demand originated by the introduction of HSR services in medium distance corridors jeopardizes the financial equilibrium of the companies. The unavoidable response is to diminish frequencies when the loss of traffic is substantial. The drop in the number of flights per hour increases total travel time when passengers arrive randomly (shuttle services), or decreases utility when they choose their flight in advance within a less convenient timetable.

Regarding the spatial effects, HSR lines tend to favor central locations, so that if the aim is to regenerate the central cities, HSR investment could be beneficial. However, if the depressed areas are on the periphery, the effect can be negative. The HSR could also allow the expansion of markets and the exploitation of economies of scale, reducing the impact of imperfect competition and encouraging the location of jobs in major urban centers where there are external benefits of agglomeration. Any of these effects are most likely to be present in the case of service industries. Location effects are dependent on many factors and it is difficult to determine a priori whether the center or the periphery will benefit from the relocation of the economic activity.

There is evidence on the increase in land values or on the increase in the GDP of places where HSR stations are located. However, the problem is to disentangle genuine additional benefits from mere relocation of economic activity from other areas of the country, or from a simple reflection of the already-measured timesavings into property values.

A Final Assessment

HSR infrastructure is expensive, irreversible, and requires a high volume of traffic to be socially profitable. In spite of all the advantages of HSR investment over competing modes, there are only two lines in the world commercially viable, the Paris-Sud Est and the Tokaido. Therefore, careful consideration should be given to the introduction of HSR in new corridors given the significant amount of public funds involved.

The investment in HSR infrastructure is one of the feasible “do something” alternatives to deal with transport-capacity problems in passenger intercity corridors. It is not the only one but the economic case for this option is more likely when there are capacity constraints in the conventional rail network, roads, and airports; and the release of capacity generates additional benefits for freight, long-haul flights, and other side effects of the marginal capacity that avoid major investments. Another potential benefit of HSR

investment is the reduction of environmental externalities, though this depends on the volume of demand deviated from less environmentally friendly transport modes and whether the demand is high enough to compensate the negative externalities during construction, and also other neglected but significant impacts as the barrier effect, noise, and visual intrusion.

There are two approaches for the ex-post economic appraisal of the HSR. One is the traditional cost-benefit analysis using real data. Another is to apply statistical methods for causal inference with observed data with and without the project to estimate the economic impact of the introduction of the HSR.

The direct benefits of HSR in terms of timesavings, willingness to pay of generated traffic and avoidable costs in competing modes, are usually not enough to compensate for the fixed construction costs and the operating costs. The magnitude of direct benefits depends on the prevailing conditions without the project. The benefits also depend upon whether there is an upgraded conventional railway able to run above 150 km and convenient air services. When this is the case, it is difficult to find a socially profitable investment based exclusively on timesavings, generated traffic and cost savings in the conventional modes unless the demand is above 10 million passenger-trips. The upper bound could be much higher, particularly when the deviated traffic comes mainly from air transport in a context where this mode of transport is delivering good and competitive services (de Rus and Nombela, 2007).

Graham and Melo (2011) have estimated the potential agglomeration benefits for the HSR in the United Kingdom based on the gravity models, to deduce changes in flows resulting from travel timesavings. They estimated an upper bound of 0.19% and a lower bound of 0.02% of UK's GDP. They conclude that "even in the best-case scenario for improvement in long-distance travel times and market share of classic and high-speed rail, the potential order of magnitude of the agglomeration benefits is expected to be small."

They also estimated the spatial economic impact of the introduction of HSR Madrid-Barcelona using a difference-in-differences specification with province and year fixed effects. Although they found positive average treatment effects for provinces with stops on the HSR line (4% for economic output, 3.3% for numbers of firms, and 1.1% for labor productivity), they did not find significant effects on employment, and though they found evidence of increased economic activity in provinces with HSR stations, measured by numbers of firms, they were not able to distinguish whether these gains were relocation or growth. In a more recent (still in progress) research for the Madrid-Barcelona line they conclude (provisionally), "The results indicate that predictions of a positive impact on the economic performance of regions receiving HSR have not taken place, at least in the short to medium term. In the case of the Madrid-Barcelona HSR corridor, our results show that there are no significant differences in the pattern of regional economic growth before and after the HSR corridor between the treated and untreated provinces" (Carbo et al., 2018).

Finally, the economic evaluation of long-lived infrastructure requires a careful construction of the counterfactual and there are many assumptions that might seriously bias the results. This is the case of transport pricing during the lifespan of the project. Pricing policy needs to be explicitly treated. We need to consider how the alternative transport modes are going to be charged. For example, the government could charge air and road transport below social marginal cost and then justify a massive rail investment as a second-best policy to change the modal split, or it could optimally price all transport modes and then evaluate the optimal way to expand capacity. The final result might be quite different.

Socially profitable or not, once the HSR infrastructure is built the costs are sunk, and this irreversibility affects to more than half of the total costs (lower density lines presents a higher proportion). Once the line is built, the marginal cost of additional traffic is quite low compared with the ex-ante marginal cost. Prices well below average total costs are common in many HSR lines around the world, fostering demand and the expansion of a network in regions or countries where there were superior alternatives to solve the transport problem at stake.

Summarizing, HSR is a mass transport technology to attend medium distance intercity mobility. It requires a quite high volume of demand to compensate the high sunk costs of the construction of its infrastructure. The majority of HSR lines in the world are far from breaking even and therefore they are sustained both with commercial revenues and public funds. Under these circumstances, the development of a HSR network is not determined by market forces and therefore the economic appraisal of projects is indispensable to determine its social profitability.

References

- Albalade, D., Bel, G., 2017. *Evaluating High Speed Rail: Interdisciplinary Perspectives*. Routledge, NY.
- Carbo, J.M., Graham, D.J., Anupriya, Casas, D., Melo, P.C., 2018. Evaluating the causal economic impacts of transport investments: evidence from the Madrid-Barcelona high speed rail corridor. *J. Appl. Stat.* 46 (9), 1714–1723.
- de Rus, G., 2012. *Economic Evaluation of the High Speed Rail*. Expert Group on Environmental Studies. Ministry of Finance, Sweden.
- de Rus, G., 2011. The BCA of HSR: should the government invest in high speed rail infrastructure? *J. Benefit-Cost Anal.* 2 (1).
- de Rus, G., Nash, C.A., 2007. In what circumstances is investment in high-speed rail worthwhile? Working Paper 590, Institute for Transport Studies, University of Leeds.
- de Rus, G., Nombela, G., 2007. Is investment in high speed rail socially profitable? *JTEP* 41 (1), 3–23.
- Graham, D., Melo, P.C., 2011. Assessment of wider economic impacts of high-speed rail for Great Britain. *Transp Res. Rec.* 2261, 15–24.
- Preston, J., 2017. Direct and indirect effects of high-speed rail. In: Albalade, D., Bel, G. (Eds.), *Evaluating High Speed Rail: Interdisciplinary Perspective*. Edward Elgar, UK.
- UIC, 2018. *High Speed Rail: Fast Track to sustainable mobility*, International Union of Railways, UIC Press, Paris.
- Vickerman, R., 2009. Indirect and wider economic impacts of high speed rail. In: de Rus, G. (Ed.), *Economic Analysis of High Speed Rail in Europe*, BBVA Foundation, Madrid, Spain, pp. 89–103.

Further Reading

This paper draws on previous research of the author for the Expert Group on Environmental Studies (Ministry of Finance, Sweden) as well as on other research papers included in this reading list.

Button, K., 2012. Is there any economic justification for high-speed railways in the United States? *J. Transp. Geogr.* 22, 300–302.

Crozet, Y., 2017. High-speed rail and PPPs: between optimization and opportunism. In: Albalade, D., Bel, G. (Eds.), *Evaluating High Speed Rail: Interdisciplinary Perspective*. Edward Elgar, UK.

Rail Cost Functions

Kristofer Odolinski, Phill Wheat, The Swedish National Road and Transport Research Institute, Department of Transport Economics, Stockholm, Sweden; Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	425
Uses of Cost Analysis in the Railway Sector	425
Cost Functions	426
Modeling Approaches for Railway Costs	426
Specification of a Rail Cost Function	427
Transport Operations Costs	428
Infrastructure Provision Costs	429
References	430
Further Reading	430

Introduction

Railways have an important role in the economic development in many countries, which can explain why changes within this industry have early been one of the core subjects of economic analysis. Specifically, the railway sector has undergone organizational transformations since its early stages, where questions regarding ownership and regulations of production or pricing of goods and services have been at the center of attention. One reason for the regulations is that the railway system (especially the infrastructure) is considered to have the characteristics of a natural monopoly—that is, an industry in which the market equilibrium results in one firm producing the output. Economies of scale, scope, and density can be reasons for a natural monopoly. Other market failures, especially externalities, have also played a role in recent transport sector regulations. It should, however, be acknowledged that regulations may be the result of interest groups' influence, and not only a means for ascertaining efficient service delivery.

A rail cost function has been (and still is) an important tool in analyzing the industry's supply side, both from an economic and an engineering perspective. Specifically, a rail cost function can be used to analyze the effect of changes related to the railway system and support decision-making. One example concerns the understanding of the impact of different forms or structural alternatives for providing rail transport, such as the effects of (de)regulation, the extent of economies of scale, scope, or density, or analyzing the cost impact of innovations and investments. Results based on a proper rail cost function can thus serve as input in forming new policies, in appraisal of transport investment, renewal and maintenance, or as means for analyzing the productivity and cost efficiency of rail transport.

Uses of Cost Analysis in the Railway Sector

A rail cost function is a way to characterize the provision of rail transport. In particular, it can provide information on the existence and level of economies of scale, density, or scope, which is essential for obtaining a better or even optimal allocation of resources. Economies of density in railways are present if adding more traffic on a fixed network leads to a lower average cost, while there are economies of scale if lower average costs are generated by an increase in both network size and traffic. There can be economies of scope if different types of traffic on a railway imply lower average costs compared to running each traffic type on separate railways. Scope economies within a train operating company may also emanate from operating both passenger and freight train traffic under a common organizational framework. Information on the economics of scale, density, and scope is for example relevant for the choice between an integrated railway system vis-à-vis a vertical separation between infrastructure management and train operations, or a horizontal separation between passenger and freight train operations. For example, do economies of scope disappear from a vertical and/or horizontal separation and are economies of density lost when train operators compete on the same line, and does increased competition outweigh any negative effects? It is also relevant regarding the optimal network size that the infrastructure management's separate regions or units are responsible for.

Whether the railway services are produced in-house or contracted out may also have an impact on costs. This is referred to as competition for the market through a bidding process where the winner of a contest will deliver train services or maintenance and renewal of rolling stock or of infrastructure. Yet, another form of competition has been introduced for train operations, with an "open access" to the tracks where operators compete with each other on the same train lines in the annual timetable. Although the use of competition has increased, the railway industry is generally not a perfectly competitive market and so market forces will not necessarily drive down costs to their minimum. This has motivated cost efficiency analyses in railways, which can be useful in benchmarking studies and as input for policies that aim to reduce efficiency gaps.

Europe's vertical separation between train operations and infrastructure management that started in the 1990s has generated a need to specify rail (infrastructure) cost functions and derive marginal costs. In particular, the use of track access charges is a requirement within the European Union. According to the (short-run) marginal cost pricing principle, a track access charge that is based on the marginal cost will create an efficient use of the infrastructure. Examples of the negative externalities (third party costs) created by the train services are wear and tear of the infrastructure, noise, accidents, and congestion-related delays. In determining the wear and tear costs, both for the trains and from trains on the infrastructure, engineering methods on deterioration mechanisms together with rail cost functions have been applied. The use of rail cost functions is thus a multidisciplinary field, with advantages of combining economic and engineering approaches.

There are indeed many cases where a rail cost function can be useful. The proper specification of the function depends on the question one wishes to investigate or answer. Next, we provide a short introduction to cost functions and the modeling approaches used for railways. This is followed by a brief overview of different types of rail cost functions and their functional forms within the empirical research. Lastly, it is presented how aspects of rail transport operations and infrastructure provision can be captured in the cost function.

Cost Functions

As is the case in many other industries, modeling of the railway system involves the use of multiple inputs to produce multiple outputs and their combinations are limited by the production possibilities set of the firm(s). The properties of the production technologies can be analyzed using a production function that relates the maximum output(s) that can be obtained by the input(s). However, this requires a proper representation of the technology and data on the input quantities used. A cost function specifies the relationship between output level(s) and input price(s), and is one way to incorporate multiple outputs with unknown input quantities, with the assumption that firms try to minimize costs and that this determines the mix of inputs used in the production. Hence, the cost function only requires information on the input prices rather than measures of the inputs that are necessary for specifying a production function.

Specifically, the use of a (rail) cost function is based on the duality between production and cost functions, which means that the cost function is derived from the production function using the assumption that the firm chooses a combination of inputs that minimizes cost given input factor prices. Hence, assumptions required for duality and a well-behaved cost function are that the (railway) firm is cost minimizing and that the input prices and the outputs are exogenously determined, which is the case on a competitive market. However, it can be contested that railway firms operate on such markets, considering the public provision, subsidies, and/or regulations within the railway industry. Still, input prices are to a large extent exogenously determined, as rail undertakings often need to purchase the inputs in competitive markets. Moreover, infrastructure provision has little choice but to accommodate more or less traffic, while (passenger) train operators often have to supply a minimum train service level set by the regulator.

In addition, there is a set of properties of a well-behaved cost function that follows from cost-minimization; it should be nonnegative, nondecreasing in input prices, positively linearly homogenous in input prices, and concave in input prices (see [McFadden \(1978\)](#) for more details). Whether these conditions are met or not depends on the specification of the cost function and its functional form, where one can impose restrictions such as symmetry and homogeneity of degree one in input prices, while some specifications require checks on the concavity condition after estimation. However, before a rail cost function is specified, a modeling approach needs to be considered, which depends on the questions one wishes to investigate or answer, as well as what type of data is available.

Modeling Approaches for Railway Costs

The main approaches used in analyzing railway costs can be divided into the following:

- a top-down approach using empirical models; and
- a bottom-up approach using mechanistic models.

The models need not be in the either of these extremes; there are also hybrids or combinations of the bottom-up and top-down approaches (see, for example, [Smith et al., 2017](#)). Moreover, the modeling approaches can be either deterministic or stochastic, where the former makes predictions that exclude random variation, while the latter predict distributions of potential outcomes.

The top-down approach tries to establish relationships with factors that may explain variations in costs, while the bottom-up approach considers the mechanisms behind these relationships on a more disaggregate level. The top-down approach is often applied to analyze economies of scale, scope, or density, or the impact of organizational changes such as (de)regulation. It is also a common approach when establishing relationships between output and (aggregate or disaggregate) costs, such as the impact traffic has on costs for providing rail infrastructure.

The bottom-up (engineering) approach can also be used in establishing cost relationships but starts at a lower level to model the physical mechanisms behind these relationships. This approach has been used to estimate the relative cost impact of different vehicle types or specific characteristics of a vehicle. In short, simulations are performed based on a set of engineering models to

provide estimates of the damage on the infrastructure caused by a certain characteristic of a vehicle or train operation. Another application concerns damages done to the vehicle using an infrastructure with a set of characteristics. These damages are then linked to maintenance activities or costs that can be implemented in order to reduce the need for future maintenance. In doing this, a rail cost function is useful (and required).

Both the top-down and bottom-up approaches have its merits and disadvantages. The strength of the former is its simplicity and practical value by using actual costs to establish casual relationships between a variable and changes in the cost (while controlling for confounders). It struggles, however, in identifying causal relationships when the underlying mechanisms are complex and possibly highly correlated. Its causal explanations and predictive power can be questioned if there are underlying mechanisms or mediating variables that, when altered, could impact the established relationship. The bottom-up approach, on the other hand, explicitly models the underlying mechanisms. However, lack of data limits the use of this approach, which can imply that some mechanisms are not well understood, but also, even if they are understood, more mechanisms require more parameters that can add uncertainty to the model. For example, rail infrastructure cost data is usually more aggregated than the predictions made with a bottom-up approach: A mechanistic model can generate measures of various damage mechanisms caused by traffic; yet, to form a cost estimate, two further stages are needed. The first is to work out the appropriate remedial interventions needed to rectify the damage and the second is to cost those actions. The later stage is particularly problematic and usually uses simple unit costs that ignore such factors as economies of scale. The modeling approaches can thus complement each other in rail cost analyses, using their best features.

Specification of a Rail Cost Function

One of the first functional forms in analyzing railway costs were linear cost functions, where costs (C) were determined by a fixed component (α) and a component proportional to output (βq), that is, $C = \alpha + \beta q$. Apart from lacking the important input price (p) variable(s), a major criticism against this function is that it implies constant marginal costs and therefore assumes that there are no diminishing returns. This is not the case for the Cobb–Douglas cost function, which is dual to its production counterpart, and can be expressed as:

$$C = e^{\alpha} \left(\prod_{k=1}^m q_k^{\beta_k} \right) \left(\prod_{r=1}^n p_r^{\gamma_r} \right) \quad (1)$$

which in logarithmic form is:

$$\ln C = \alpha + \sum_{k=1}^m \beta_k \ln q_k + \sum_{r=1}^n \gamma_r \ln p_r \quad (2)$$

and can be estimated using standard linear techniques, for example, ordinary least squares, where α , β_k , and γ_r are the parameters to be estimated. Yet, this cost function assumes that the cost elasticities are constant. In other words, it assumes constant returns to scale and density as the elasticities with respect to output are constant, and a unitary elasticity of substitution between input factors (i.e., a proportional change in the input price ratio will lead to an equiproportional change in the input factor ratio). A more flexible functional form is the translog specification (Christensen et al., 1973; Christensen and Greene, 1976), which can be expressed as:

$$\ln C = \alpha + \sum_{k=1}^m \beta_k \ln q_k + \sum_{r=1}^n \gamma_r \ln p_r + \frac{1}{2} \sum_{k=1}^m \sum_{l=1}^m \beta_{kl} \ln q_k \ln q_l + \frac{1}{2} \sum_{r=1}^n \sum_{s=1}^n \gamma_{rs} \ln p_r \ln p_s + \sum_{k=1}^m \sum_{r=1}^n \delta_{kr} \ln q_k \ln p_r \quad (3)$$

This is a second-order approximation of any cost function, which relaxes the assumptions of constant returns to scale and density, as well as the unitary elasticity of substitution between inputs. Note that linear homogeneity of inputs prices can be imposed by using one of the input prices as the denominator when dividing the cost and the other input prices.

As mentioned previously, the well-behaved cost function presumes that firms or decision-making units are cost minimizing. However, they are not always successful in this objective, and models have been developed to consider cost inefficiency. In general, a component (u) for the difference between actual costs (C) and minimum costs ($C^{\min}$) is introduced in the model specification. Using the logarithmic form, the actual cost is then expressed as:

$$\ln C = \ln C^{\min} + u. \quad (4)$$

The level of cost efficiency is $C/C^{\min}$, which from Eq. (4) is e^{-u} . Modeling cost efficiency can be useful in benchmarking studies of railways, with an aim to identify the decision-making unit (i) that lies on the cost (minimizing) frontier and the level of inefficiency that the other units have.

Observations over time (t) are common and useful in (rail) cost analyses. This allows a cost specification that reflects technical change, which means that the cost frontier changes over time. One can include a (linear or quadratic) trend variable in the cost function or the less restrictive time dummy variables. Furthermore, the time dimension makes it possible to consider the dynamics in rail costs. For example, maintenance costs in one year can have an impact on the maintenance carried out in the subsequent year(s); a (sudden) change in traffic can result in a change in maintenance activities, which implies a temporary deviation from the original cost-minimizing plan that may take time to adjust back to (e.g., due to budget and/or planning

restrictions). Using the Cobb–Douglas cost function in logarithmic form, a dynamic cost model for firm i in time t can be expressed as:

$$\ln C_{it} = \alpha + \theta \ln C_{it-1} + \sum_{k=1}^m \beta_k \ln q_{kit} + \sum_{r=1}^n \gamma_r \ln p_{rit} \quad (5)$$

This is a special form of the autoregressive distributed lag model, which also includes lagged explanatory variables. In addition to the effect of for example output ($\ln q_k$) on current costs, captured by β_k , the model captures how this effect traces into subsequent year(s) through the term for lagged costs, θ . Specifically, when there is no tendency to change costs, the “equilibrium cost” is $\ln C_{it} = \ln C_{it-1} = \ln C_i^e$, and, from Eq. (5), it can then be expressed as:

$$\ln C_i^e = \frac{\alpha}{1-\theta} + \sum_{k=1}^m \frac{\beta_k}{1-\theta} \ln q_{ki} + \sum_{r=1}^n \frac{\gamma_r}{1-\theta} \ln p_{ri}. \quad (6)$$

where the effect of a change in an output ($\ln q_k$) in the current period and subsequent period(s) is $\frac{\beta_k}{1-\theta}$.

To choose between different modeling approaches and rail cost functions, the starting point is to specify the objective of the analysis. This establishes the level of detail required; is it necessary to model underlying mechanisms or are direct relationships between variables enough? Is the whole railway system considered, or parts (and subparts) of the system such as infrastructure provision (total costs or costs divided between track superstructure, track substructure, signaling equipment, etc.) and train operations (costs for freight transport or passenger transport)? This will impact the decision on which outputs and inputs to use in the specification.

Common metrics for the output of the system are train-, vehicle-, and seat-kilometers (outputs that are delivered as part of the rail system) and ton- and passenger-kilometers (outputs that have been consumed, the so-called final outputs). Both types of outputs can be important to include in the cost function. For example, to assess the efficiency of infrastructure provision, it is important to consider that a line with a high train frequency is costly to maintain (e.g., due to short and fragmented time slots available for maintenance) and that higher axle loads imply more wear and tear of the tracks. Also, for train operation efficiency, both the frequency of service (train-kilometers per route-kilometers) and seat-kilometers can be useful representations since both provide information about different types of quality aspects. A high number of seat-kilometers does not necessarily imply high frequency (it is possible to use longer trains to increase seat-kilometers) and vice versa.

It can also be relevant to consider the heterogeneity in costs between different types of passenger train services such as high-speed, interregional, regional, or suburban trains, as well as for freight train services where different commodities may imply different costs per ton-kilometer. The next section is devoted to transport operations costs and how to deal with the heterogeneity in outputs. This is followed by a section on infrastructure provision costs, which must also address a problem related to the heterogeneity of outputs: information on axle load, running gear, and the speed of the vehicles running on the network can be important for explaining variations in the wear and tear costs.

As input prices vary, the rail cost function should include these for ascertaining that the duality between the production and cost function prevails. Excluding input prices that are not constant across the observations implies a mis-specified model; there are, however, econometric techniques that can control for general (cross-section wide) input price changes. Examples of input price information that are useful are price of capital (rolling stock and/or materials for other assets used in the production), labor, and energy.

Transport Operations Costs

Transport operations refer to the part of the transport production process where vehicles are run on a network and the output may be regarded as the transport of passengers or freight. Passenger-kilometers and freight ton-kilometers are therefore a common basis for rail output measurement. Researchers and rail managers have often added these together to form a measure of output known as traffic units, although this will only be appropriate if they cost similar amounts to produce. Otherwise, increasing productivity may simply appear because the railway is moving toward producing more freight traffic and less passenger traffic, or vice versa.

Strictly, a rail transport service can be described in terms of the provision of transport from an origin to a destination at a specific point in time and of a specific quality. A rail passenger operator that runs trains between numerous stations several times per day has therefore a vast number of products on offer. However, as well as being infeasible to characterize in a cost function, this issue is only distorting if the products have significantly different cost characteristics, and traffic on them is increasing or decreasing at different rates.

Furthermore, in passenger transport, the speed of service and the stopping pattern of trains have an impact on costs. This could be measured directly, for example, average speed of trains and number of stops, or, alternatively, by splitting traffic into different categories, for example, commuting and long distance. In any event, frequency of service is an important quality attribute and railways operate trains at different frequencies to accommodate different patterns of demand. There is therefore a clear benefit in distinguishing between train-kilometers, vehicle-kilometers, and ton-kilometers. The advantage of a cost function is that multiple output measures can be incorporated within the cost function.

Freight traffic also has a multiplicity of outputs, where the transport cost for a ton of freight may vary depending on whether it is a dense product or not, and the form it is in, considering their impact on the gross weight of the wagon and on costs for loading and

unloading. This suggests that loaded-wagon-kilometers may be a better output measure than ton-kilometers. Moreover, the output measure may need to distinguish between for example bulk traffic and intermodal container traffic. If overall ton-kilometers are used, the estimated cost relationship would be distorted if there is a shift from transporting bulk aggregates to lighter containers, as has been the case in many railways.

There are indeed several challenges in specifying outputs in cost functions for transport operations arising from the multiplicity of outputs. From an econometric point of view there are two ways to view this problem. Firstly, multiple outputs are simply an issue of scope, that is, the firm produces multiple products. Thus, the issue is how disaggregate the output measure should be, trading off accuracy in specifying the production process with data and estimation accuracy constraints (similar to the issues with input prices). Alternatively, multiple outputs can be characterized by continuous attributes. So, train-kilometers could be used as the primary output, but then several characteristic variables, such as the length of trains, average speed of trains, and service type indicators, could be used. Therefore, the problem is not seen as scope, but variations in the quality or other characteristics of relatively generic high-level outputs. Both approaches, and indeed a mixture of these two approaches, are used by practitioners. Indeed, some researchers have included many output measures and characteristics of output in their cost functions using a device known as a Hedonic function, which was first used by [Spady and Friedlaender \(1978\)](#). This function has the advantage of avoiding consuming too many degrees of freedom in the estimation process, leading to greater precision in the estimated relationships.

Input price data is often difficult to come by. This is for two reasons. Firstly, available accounts from train operators may not produce a detailed breakdown in cost. The implication is that, in addition to costs for, for example, staff and rolling stock, there is often a large proportion of total cost which is allocated to “other”: a category by definition to which it is impossible to assign an input and thus construct a price. Secondly, and more fundamentally, the complexity of the production process, associated heterogeneity of inputs, and resulting problem of joint costs, means that there is a limitation to the extent that enough input prices can be constructed such that the diversity in inputs can appropriately be characterized in the cost function.

Infrastructure Provision Costs

Building and managing railway infrastructure involves using a large amount of resources that, after implementation, cannot be redeployed; costs are sunk. After an investment in infrastructure, the assets need to be maintained and eventually renewed due to wear and tear. A cost minimizing plan would imply that the infrastructure manager carries out maintenance—a cost that is increasing over time due to accumulated use—until it is cheaper to renew the infrastructure (based on, e.g., annual equivalent cost comparisons). This type of life-cycle asset management requires information on the wear and tear costs (as well as other types of costs) caused by traffic, which is also necessary for determining marginal costs of infrastructure use. Indeed, rail cost functions can be used to predict how variations in traffic affect life-cycle costs and thus the cost-minimizing plan for the asset, and also to determine the (short-run) marginal cost per train- or gross ton-kilometer. Evidence suggests that the wear and tear cost elasticities with respect to traffic is around 0.25–0.45 for maintenance, and around 0.5 for renewals ([Nash, 2018](#)), indicating that the average costs for these activities decrease with traffic.

Gross ton-kilometer is a common output measure in analyses of rail infrastructure costs as it is an important driver of the wear and tear of the infrastructure. Still, variations in the characteristics of the vehicles such as axle load, bogie type, and speed can cause different levels of damages. There is a growing interest in differentiating between vehicle characteristics and their impact on infrastructure costs, where the use of a bottom-up (mechanistic) approach is common. However, as described previously, a combination between a bottom-up and top-down (empirical) approach can be particularly useful in this case; the former is good at predicting the relative damage caused by vehicles, while the strength of the latter is to establish a relationship between these damages and actual costs using few restrictions on the elasticities of production.

To establish a relationship between traffic and costs, it is important to properly characterize the nature of the rail infrastructure. This can be done by including measures of infrastructure characteristics such as track length, switches, bridges and tunnels, as well as measures of the capability and condition of the infrastructure such as maximum line speed, maximum axle load allowed, and rail age. In general, measures that can explain variations in infrastructure costs and in traffic are important to include in the rail cost function in order to capture the short-run marginal costs of traffic. In addition, such measures are also essential in an efficiency analysis of infrastructure provision.

When analyzing infrastructure provision costs, one needs to be aware of that the data generation processes are different between maintenance and renewal costs. As described earlier, maintenance activities are carried out over the life-cycle of the asset, while a renewal is undertaken to restore the asset to its original condition. The causal structure between traffic and maintenance cost is based on the additional wear and tear caused by extra traffic, requiring an increase in maintenance activities within the same year and possibly also in subsequent year(s) to keep a certain level of performance and reliability of the infrastructure (minimize costs of the asset's life-cycle). The rail maintenance cost function is thus usually specified as traffic (and other variables) in one year explaining maintenance costs in the same year. It can also include a dynamic component (lagged maintenance costs) to analyze whether intertemporal effects are prevalent [Eqs. (5) and (6)].

The renewal marks the “death and rebirth” of the asset, which means that the time intervals between these activities on a railway line are relatively long, typically longer than 20 years. Hence, to establish a relationship between traffic (and other factors) and the lumpy renewal cost, relatively longer time series need to be used than what is required for establishing the relationship for maintenance costs.

In specifying a renewal cost function, it is important to consider that a *temporary* increase in traffic implies that future renewals will be carried out earlier than originally planned. A permanent change in traffic may also result in a change in the length of the renewal intervals. Both consequences affect the present value of renewal costs (and can also have an impact on the size of the renewal cost). One approach that has been used to capture these intertemporal aspects is survival analysis, where it is estimated how different covariates affect the risk of renewal (Andersson et al., 2016). Another approach is to specify the renewal cost function as a two-part model, where the probability of a renewal is estimated in the first part and the size of the renewal is estimated in the second part (Andersson et al., 2012).

References

- Andersson, M., Björklund, G., Haraldsson, M., 2016. Marginal railway renewal costs: a survival data approach. *Transp. Res. Part A* 87, 68–77, doi:10.1016/j.tra.2016.02.009.
- Andersson, M., Smith, A., Wikberg, Å., Wheat, P., 2012. Estimating the marginal cost of railway track renewals using corner solution models. *Transp. Res. Part A* 46, 954–964, doi:10.1016/j.tra.2012.02.016.
- Christensen, L.R., Greene, W.H., 1976. Economies of scale in U.S. electric power generation. *J. Polit. Econ.* 84 (1), 655–676.
- Christensen, L.R., Jorgenson, D.W., Lau, L.J., 1973. Transcendental logarithmic production frontiers. *Rev. Econ. Stat.* 55 (1), 28–45.
- McFadden, D., 1978. Cost, revenue, and profit functions. In: Fuss, M., McFadden, D. (Eds.), *Production Economics: A Dual Approach to Theory and Applications*. North-Holland, Amsterdam, pp. 2–109.
- Nash, C., 2018. Track access charges: reconciling conflicting objectives. Project Report. CERRE, Centre on Regulation in Europe, Brussels, May 9, 2018.
- Smith, A.S.J., Iwnicki, S., Kaushal, A., Odolinski, K., Wheat, P., 2017. Estimating the relative cost of track damage mechanisms: combining economic and engineering approaches. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* 231 (5), 620–636, doi:10.1177/0954409717698850.
- Spady, R.H., Friedlaender, A.F., 1978. Hedonic cost functions for the regulated trucking industry. *Bell J. Econ.* 9 (1), 159–179, doi:10.2307/3003618.

Further Reading

- Caves, D.W., Christensen, L.R., Swanson, J.A., 1981. Productivity growth, scale economies, and capacity utilisation in U.S. railroads 1955–74. *Am. Econ. Rev.* 71 (5), 994–1002.
- Griliches, Z., 1972. Cost allocation in railroad regulation. *Bell J. Econ. Manag. Sci.* 3 (1), 26–41, doi:10.2307/3003069.
- Ivaldi, M., McCullough, G.J., 2001. Density and integration effects on Class I U.S. freight railroads. *J. Regul. Econ.* 19 (2), 161–182, doi:10.1023/A:1011149324027.
- Jara-Diaz, S.R., 1982. The estimation of transport cost functions: a methodological review. *Transp. Rev.* 2 (3), 257–278, doi:10.1080/01441648208716498.
- Jara-Diaz, S.R., Basso, L.J., 2003. Transport cost functions, network expansion and economies of scope. *Transp. Res. Part E* 39, 271–288, doi:10.1016/S1366-5545(03)00002-4.
- Johansson, P., Nilsson, J.-E., 2004. An economic analysis of track maintenance costs. *Transp. Policy* 11 (3), 277–286, doi:10.1016/j.tranpol.2003.12.002.
- Keeler, T.E., 1974. Railroad costs, returns to scale, and excess capacity. *Rev. Econ. Stat.* 56 (2), 201–208, doi:10.2307/1924440.
- Mizutani, F., Uranishi, S., 2013. Does vertical separation reduce cost? An empirical analysis of the rail industry in European and East Asian OECD countries. *J. Regul. Econ.* 43 (1), 31–59, doi:10.1007/s11149-012-9193-4.
- Nash, C., 2005. Rail infrastructure charges in Europe. *J. Transp. Econ. Policy* 39 (3), 259–278.
- Nash, C., 2008. Passenger railway reform in the last 20 years—European experience reconsidered. *Res. Transp. Econ.* 22 (1), 61–70, doi:10.1016/j.retrec.2008.05.020.
- Oum, T.H., Waters, W.G. (II), Yu, C., 1999. A survey of productivity and efficiency measurement in rail transport. *J. Transp. Econ. Policy* 33 (1), 9–42.

Transport and International Trade

Siri Pettersen Strandenæs, Norwegian School of Economics, Bergen, Norway

© 2021 Elsevier Ltd. All rights reserved.

Introduction	431
Why Countries Trade?	431
How does International Trade Affect Transport?	432
Efficient Transport Facilitates Trade	433
Trade and Transport Costs Developments	434
References	435
Further Reading	435

Introduction

International merchandise trade implies moving cargo across borders, whereas international trade in services either requires people to travel or to get access to cross border communications. Hence, transportation derives from the volume traded and the distance between the traders. Currently, maritime transport carries more than 80% of international merchandise trade. Airlines, railways, or trucks transport the remaining cargoes in international trade (UNCTAD, 2018).

Why Countries Trade?

Countries trade with each other for the same reasons as traders within nations; to get access to goods or varieties not available in other ways and to gain economically from the exchange. International trade induces economic growth when countries produce goods and services in locations where their production requires relatively less resources and effort, and exchange them against products and services that can be produced relatively more efficiently in other countries. Ricardo (1817) first explained how economic gains from trade reflect such comparative advantages in production. The theory of comparative advantage still is valid. Any textbook on international economics explains it (Krugman et al., 2018). Comparative advantage can explain why countries trade different goods. Each country specializes in the goods it produces relatively most efficient and imports the goods produced relative more efficiently by the other country. Ricardo's classical example explains how both England and Portugal gain by exchanging wool for wine (Richardson, 1817). A more modern formulation of the background for such trade was formulated by Venables (2006) as interaction between factor intensity of goods and factor abundance of countries. Thus, countries have comparative advantage in goods that use the factors intensively that the country is rich in.

The theory of comparative advantage explains trade in raw materials such as oil, gas, iron ore, coal, and grain. These commodities are mostly undifferentiated and countries mostly either have access to them and export their excess supply, or need to import them since they are not available at home. These commodities are important cargoes in wet and dry bulk shipping.

A large share of internationally traded goods are different varieties of similar goods or services, for example, the international trade in cars. Thus, a country may both export and import cars. Most countries also exchange tourist travels both having visitors coming to the country and see its own citizens traveling abroad. Such intra-industry trade creates gains for the trading parties by enabling producers to exploit economies of scale in production, and by letting consumers choose different varieties of the goods. This new trade theory was introduced by Krugman (1979). Intra-industry trade explains trade in consumer goods and intermediaries such as parts and components. These goods are the main commodities in air cargo, break bulk shipping and container shipping. The method most widely used to measure intra-industry trade is the Grubel-Lloyd Index (Grubel, 1975).

Krugman (1991) introduced the New Economic Geography that explains localization of the different stages in the production process by combining trade theory and location theory. Production is located and concentrated based on interaction between the internal economies of scale at firm level and the transport costs.

Trade in intermediaries has opened up for fragmentation (Jones and Kierzkowski, 2001) of production lines across countries, thereby increasing international trade and transport. Following this, the share of parts and components in international trade has grown substantially. Several products are assembled from parts and components produced in a many different countries across different continents. Hence, producers outsource parts of the production. This allows for complex division of labor across the world with production of the different parts taking place in their lowest-cost locations. Such organization introduces extra costs of coordination and communication. Deregulation of the airline industry, the development of the telecom industry and the internet facilitated coordination across firms and locations. Fragmentation across locations also increases the transportation content in the final products. Both cost and timeliness of transportation are critical in such fragmented organization of the production.

The volume traded internationally has increased and continues to do so. For the transportation industry, both cargo volumes and trading distance are important, however. Changes in production and consumption locations and thereby in trade patterns may increase or decrease the average transport distance. China's stronger integration into the world economy after joining the World Trade Organization (WTO) in 2001 illustrates the effects on transport distance. China's growing export of consumer goods increased the shipping of final goods in containers to the United States and Europe. This implied both a rise in volume and average distance for container shipping. China at the same time imported more raw materials; oil, iron ore, and later coal, thereby increasing transport volumes and distance especially for the dry bulk fleet. The financial crisis and slower growth in the United States and Europe dampened the growth of these long-distance trades. At the same time, higher economic growth and outsourcing of production from China to emerging Asian economies increased the intra-Asian trade's share of world trade, thereby halting the growth in distance.

Most models in international trade theory, model trade costs including transport cost as so-called "iceberg" cost. Samuelson (1954, p 268) introduced the idea that transportation reduces the value of the goods transported by a fraction proportional to the distance traded. The iceberg formulation of transport costs also implies that valuable commodities are more costly to trade than less valuable goods, when using a fixed percentage of cargo value for transport cost by assuming that transport (iceberg) costs are proportional to cargo value.

How does International Trade Affect Transport?

Ships or aircraft move intercontinentally traded goods, whereas intracontinental cargo flows also move by rail and road. Intracontinental trade resembles regional exchange of goods except for the costs of crossing borders. Intracontinental or regional trade is not discussed in this chapter. Rather, we concentrate on the relationship between international trade and sea and air transportation.

World economic growth and international transportation develop in parallel. Fig. 1 shows the parallel development in world GDP and the volume world seaborne trade.

There is limited overlap or competition between air cargo and maritime transportation. Ships carry heavy bulky cargoes whether wet or dry. Container shipping, however, moves some of the same commodities as air cargo. The freight costs differ substantially between air and maritime transport, and implies limited competition between the two modes of containerized transportation. Air cargo carries commodities that are valuable relative to their weight and especially commodities that also require speed and timeliness. Hence, air cargo consists of time-sensitive perishable and high value commodities. The annual growth in revenue ton-kilometers for air cargo has been 2,6% per year on average in later years. Furthermore, as it is seen from Fig. 2 revenue kilometers performed by air carriers to a large degree reflect changes in industrial production. However, the average growth rate hides big geographical variations (Boeing, 2018).

Containerized maritime transportation over the same years (2007–17) rose by 4.9% per year. The share of containerized cargoes in seaborne trade also increases. This development is consistent with the changes in division of labor in the world economy, where fragmentation has increased trade in intermediaries such as parts and components (Fig. 3).

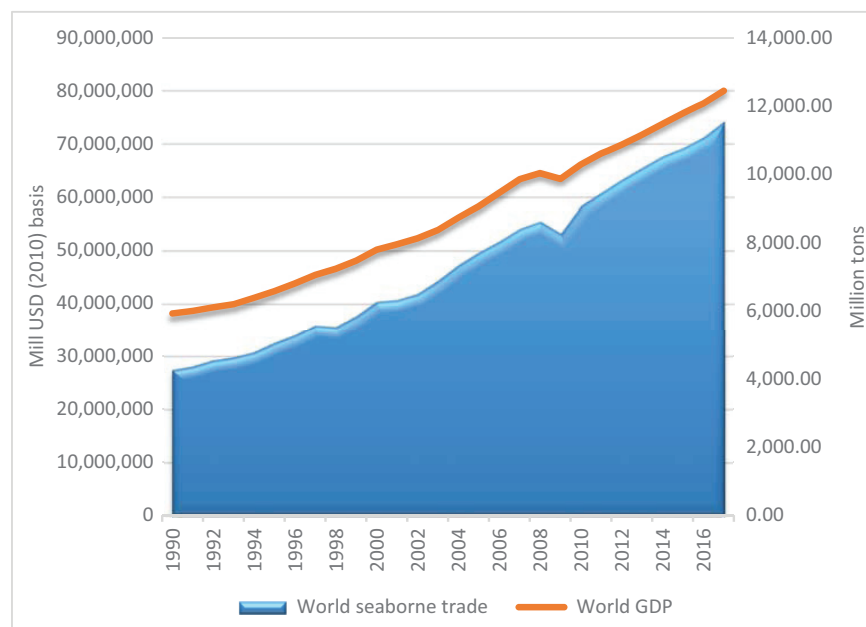


Figure 1 Seaborne trade and world economic growth. Sources: Based on data from Clarksons, 2019 and UNCTADstat, 2019

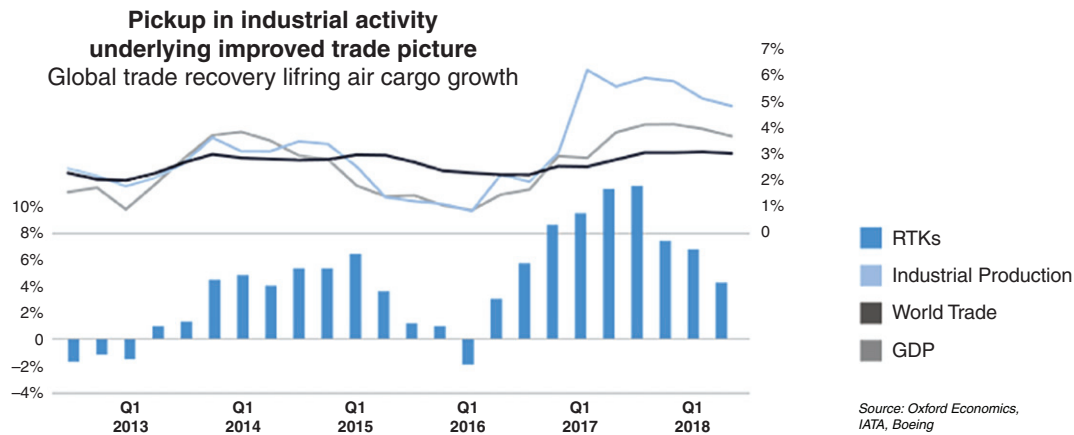


Figure 2 Development in air cargo, world GDP, and world industrial production. Source: Boeing, 2018

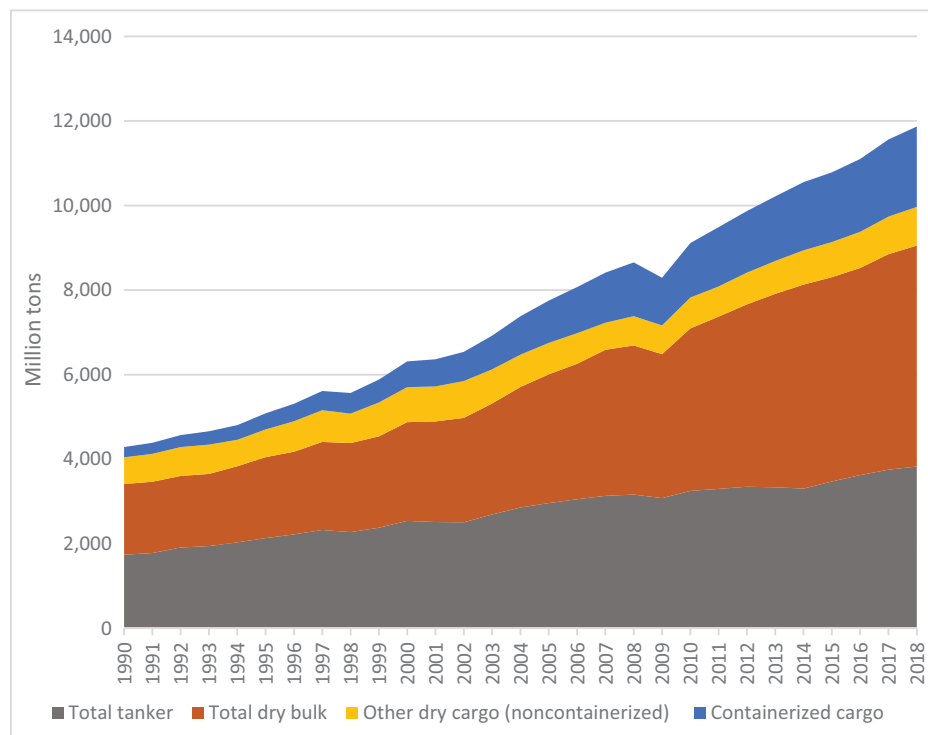


Figure 3 Containerized commodities in seaborne trade. Source: Based on data from Clarksons, 2019

Efficient Transport Facilitates Trade

Distance is costly. Adam Smith pointed out in *Wealth of Nations* 1776 that international trade relates to location as stronger division of labor and thereby more exchange typically develops along shores and navigable rivers accessible by water transport offering lower transport costs (Smith, 1776).

Hummels (2006) found that roughly half of the volume of international trade takes place between countries located within 3000 km of each other and that regression estimates suggest that doubling the distance halves the volume of trade. Hence, there is a strong relation between distance and trade volume. This distance effect is caused by the rise in transport and communication costs with distance. Hummels (2006) points out, however, that the distance effect has fallen over time. Comparing transport costs for seaborne transport in 1974 and 1998 he finds that the costs difference between long haul (9000 km) and short haul (1000 km) in 1998 was 32% compared to 59% in 1974. For air cargo, the figures were 300% in 1974 and 68% in 1998.

From a similar reason, landlocked countries face higher transport costs for their exports and imports. Economies of scale and the unit load are higher for seaborne transport than for road or rail transportation. Hence, transporting large volumes over land typically is more costly per kilometer than using seaborne transport. In addition, exports and imports from and to a landlocked country face several loading and unloading operations since the cargo must be transferred from one transport mode to another.

The transport intensity for goods differs. For countries which exports or imports transport intensive commodities, the level of transport costs becomes more important in deciding both the volume of trade and where to source imported goods and sell their exported goods. The term transport intensity may also measure the growth in freight transportation relative to the GDP growth. It furthermore, is often argued that the traditional parallel growth between GDP and freight transportation has decoupled in later years. This definition of transport intensity is, however, not used in this section. Here we focus on difference in transportation intensity among different traded goods and services.

High-value transport intensive goods typically move over longer distances than low-value goods as they can better bear the higher transport cost. For high-value products or services, transport costs will imply a lower markup than the similar transport cost for lower value products or services. Hence, high-value, high-quality goods tend to be sent over longer distances. [Hummels and Skiba, \(2002\)](#) studying the relationship between trade costs and the quality composition of trade, found that for high-quality and low-quality units of substitutable products countries tend to export the high-quality variant and sell the low-quality version in the domestic market, since the low-quality version cannot bear high transport and trade costs.

High volume of trade allow for increasing unit loads and thereby better exploitation of economies of scale in the transport sector. Hence, countries with a larger population tend to face lower unit transport costs than smaller countries. Such lower transport costs per unit induce more trade, which by itself leads to a fall in transport costs. This is so as long as there is ample capacity in the transport sector, which enables shipping the extra cargo.

Two-way trade in commodities that moves in the same kind of transport units reduces the imbalance between outbound and inbound transportation and increases the capacity utilization of the transport mode by reducing the share of ballasting in the sailing pattern for the vessels and the air carriers. This is another example of the two-way relationship between transport and trade. Better trade balance reduces transport costs and lower transport costs induce higher trade activity.

Punctuality has become more important following the international fragmentation of production processes and the emergence of just-in-time production strategies in manufacturing. Punctuality is important and more important than speed of transit for most commodities except for fresh and/or time-sensitive products. For such products or services, timeliness is essential. Hence, the quality of transportation services, both transit time, punctuality and low breakage are important for international trade. Thereby, not only the quality of vehicle operation, but also the capacity and the functioning of the infrastructure such as ports and airports, is important for world trade and economic growth. One reason often named for African countries' limited participation in international trade is the general lower capacity and quality of the continents transportation infrastructure.

Trade and Transport Costs Developments

High transport and communication costs hamper international trade. Technological development over the years has facilitated sourcing of goods and intermediaries over long distances. [Jacks and Pendakur \(2010\)](#) assess the effect on international trade from the development in maritime transport following the technological and communicative changes in the 19th century and find that the sharp reduction in transport costs appeared together with the sharp growth in global trade. On average, the freight rates fell 50% and world trade increased 400% between 1870 and 1913. They point out that the development in trade and maritime transport costs were parallel, but falling transport costs did not by itself cause the sharp growth in global trade in this period. Lower transport costs facilitate trade, but growing trade is part both of a virtuous circle of economic growth in the trading countries and of the rising trade between them.

[Hummels \(2006\)](#) gives an overview of the results from several analyses of long-term developments in transport costs in intercontinental trade in the post WWII period, that is, in the development in shipping cost for both air cargo and seaborne trade. By analyzing transport costs as a wedge between prices inclusive and exclusive of transport costs, he assesses the development in transport costs over time. The analysis covers two different price wedges (1) The import wedge assess the price differential between imported goods in the importing country and the price the producer in the exporting country receives, that is, comparing the fob (free-on-board) and cif (cost insurance freight) prices of imported commodities. This wedge does not cover transport costs to and from the border in each of the trading countries and hence does not tell the full story of transport costs for the internationally traded commodities. (2) The sourcing wedge assesses the cost difference for an importer who may import a good from either one of two countries. These wedges are not purely transport costs however. The development in other trade costs may also influence the development in these relative prices over time.

[Hummels \(2006\)](#) cautions the interpretation of these wedges, since the wedge will vary for different goods according for example, to their value and their transportation intensiveness. Some commodities may require higher quality transportation or faster transportation than others may. This has to be considered when analyzing the development in transport costs over time, since the composition of trade changes with time.

Ocean shipping and air cargo experienced substantial economies of scale in ships and aircraft respectively, thereby reducing the cost per ton-km transported. [Hummels \(2006\)](#) finds that overall the transport costs have fallen over time. The freight rates in air cargo have fallen more strongly than freight rates in ocean shipping. The reduction in aggregate transport costs has fallen further

since the composition of commodities in world trade has shifted toward lighter manufactures. Thereby a larger share of goods are shipped by air. Goods shipped by air typically also are more time sensitive than other internationally traded goods. Hence, the development reflects a change in the composition of goods in world trade. There is a reciprocal relationship here, as lower air cargo freight rates increase the distances over which time sensitive goods may be traded.

In 2001 (Baier and Bergstrand, 2001) published an analysis of the growth of world trade among OECD countries between late 1950s and late 1980s with reduced tariff levels, falling transport costs and convergence in income between countries as explanatory variables. They find that income growth is most important for increasing trade with tariff reductions as the number two element. Income convergence on the other hand had no significant effect for world trade growth in their analysis. Declining transport costs also had a significant effect, but markedly lower than the others.

We thus find that there is a close relationship between international trade and transport.

References

- Baier, S.L., Bergstrand, H., 2001. The growth of world trade: tariffs, transport costs and income similarity. *J. Int. Econ.* 53, 1–27.
- Boeing, 2018. *World Air Cargo Forecast 2018–2037*. Boeing, Seattle.
- Clarksons, 2019. *Shipping Intelligence Network*. London. Available from: <https://sin.clarksons.net/>.
- Grubel, H.G., 1975. *Intra-Industry Trade: The Theory and Measurement of International Trade in Differentiated Products*. Macmillan, London.
- Hummels, D., 2006. Transportation costs and trade over time. In: *ECMT Transport and International trade*, OECD, Paris, pp. 7–26.
- Hummels, D., and Skiba, A., 2002. Shipping the good apples out? An empirical confirmation of the Alchian-Allen Conjecture. NBER.
- Jacks, D.S., Pendakur, K., 2010. Global trade and the maritime transport revolution. *Rev. Econ. Studies* 92 (4), 745–755.
- Jones, R.W., Kierzkowski, H., 2001. A framework for fragmentation. In: Jones, R.W., Kierzkowski, H. (Eds.), *Fragmentation, New Product Patterns in the World Economy*. Oxford University Press, Oxford.
- Krugman, P., 1991. Increasing returns and economic geography. *J. Polit. Econ.* 99 (3), 483–499.
- Krugman, P.R., Obstfeld, M., Melitz, M., 2018. *International Economics; Theory and Policy*. Pearson, Essex.
- Krugman, P.R., 1979. Increasing returns, monopolistic competition, and international trade. *J. Int. Econ.* 9, 469–479.
- Richardo, D., 1817. *On the Principles of Political Economy and Taxation*. John Murray, London.
- Samuelson, P.A., 1954. The transfer problem and transport costs II: analysis of effects of trade impediments. *Econ. J.* 64, 264–289.
- Smith, A., 1776. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London.
- UNCTAD, 2018. *Review of Maritime Transport*, Geneva: UNCTAD
- UNCTADStat., 2018. Available from: <http://unctadstat.unctad.org/wds/tableViewer/tableView.aspx>.
- Venables, A., 2006. Infrastructure, trade costs and gains from international trade. In: *ECMT Transport and International Trade*, OECD, Paris.

Further Reading

- Anderson, J.E., Wincoop, Eric van, 2004. Trade costs. *J. Econ. Lit.* 42 (3), 691–751.

Maritime Economics: Organizational Structures

Kevin P.B. Cullinane, University of Gothenburg, Gothenburg, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	436
Regulation	437
The International Maritime Organization (The IMO)	437
National Maritime Administrations	437
Supranational Organizations	438
Supply-Side Structures	438
Shipowners Associations	438
Intertanko	438
Intercargo	439
World Shipping Council (WSC)	439
Alliances	439
The Demand for Shipping - Cargo Interests	440
Regional (Supranational) Shippers' Associations	440
The Global Shippers' Alliance (GSA)	440
Ancillary Institutions	440
UNCTAD	440
The Baltic Exchange	441
BIMCO	441
Biography	441
References	442

Introduction

There is a vast array of industries, which are sometimes deemed to comprise the global maritime sector. This could include an extremely diverse range of activities from aquaculture and fishing to offshore technology and pipelines (Walters and Bailey, 2013). The field of maritime economics, however, is specifically concerned with shipping as a mode of transportation for the carriage of goods and, to a lesser extent, passengers. In addition to the shipping industry itself, the field also encompasses numerous other activities, which have a direct effect upon the act, or price, of providing shipping services. The most important of these is undoubtedly the port industry, representing another major element of what constitutes the field of maritime economics (Cullinane and Talley, 2006; Talley, 2017). Also included, but to a lesser extent, are: the ship construction and demolition industries; maritime law, finance and insurance; ship broking; and a large number of ancillary activities associated with the replenishment of ships (ship agency, marine fuels, etc.).

Focusing specifically on the shipping industry, it comprises a number of sub-sectors, or market segments, which are ultimately differentiated from each other on the basis of the form of cargo or loading unit carried and the type of vessel design that facilitates the carriage of that cargo or loading unit. Thus, at an aggregate level, the two largest^a sub-sectors of the shipping industry are the *bulk* and *container* sectors. The former is recognized as having a market structure that is very close to perfect competition. However, some form of cooperation remains an important attribute of the latter, suggesting a market structure that historically has been tantamount to oligopoly (and, arguably, still remains so in some parts of the world) and all that this implies about the potential for collusive behavior (Cullinane, 2005, 2011).

This very broad dichotomous market segmentation can be further disaggregated into more specific sub-markets, each defined by their own individual set of characteristics. In general, however, the demand (from shippers of cargoes) and supply (from shipowners or operators) in virtually all shipping markets is characterized by a proliferation of private sector actors, with market structures having largely evolved organically over time as the outcomes of accumulated decisions and actions taken by private sector interests. As a consequence of this, the organizational structures that are germane to the shipping industry (and, to a large extent, also the port industry) are not the product of some grand design. These too have evolved as a response to, or reflection of, market developments. Somewhat inevitably, the skin of these organizational structures hang on the bones of a number of institutions that have developed in order to, variously, moderate the worst excesses of the free market, prevent the abuse of market power, promote free trade, and protect the interests of the many and diverse stakeholder interests at play within the shipping and port markets. The content which

^a In terms of volumes carried, ton-kms moved and revenue earned (i.e., market share).

follows, therefore, addresses the many and varied institutions which are relevant to maritime economics and which help to define its main organizational structures.

Regulation

The International Maritime Organization (The IMO)

By virtue of its inherently international nature, the shipping industry is primarily regulated at a global level by the [United Nations International Maritime Organization \(the IMO\)](#), which is ultimately responsible for ensuring both the safety of life at sea and the protection of the marine environment. In respect of the former responsibility, the IMO is supported by the work of the United Nations International Labor Organization (the ILO) in the setting of appropriate labor standards ([Branch, 2014](#)). Although many international maritime treaties had already been agreed between nations throughout history, the concept of an international organization to regulate the world's shipping industry emerged at the time the United Nations was initially created in 1948. It took until 1958, however, before the precursor of the IMO, the Inter-Governmental Maritime Consultative Organization (IMCO), was inaugurated. This organization subsequently changed its name to the IMO in 1982.

With its headquarters in London in the United Kingdom, the membership of the IMO comprises the national governments of most countries in the world. Over the years, the members of the IMO have ratified a number of conventions that have been converted into detailed technical sets of regulations, which govern the safety of ships and seafarers at sea and protect the marine environment.

The regulations that are ratified by the members of the IMO usually become embodied within a wider pre-existing convention, but can lead to the establishment of a completely new convention or code. The most relevant of these conventions or codes are as follows:

- the safety of life at sea (SOLAS)
- the avoidance of vessel collisions at sea (COLREGS)
- the allowable safe loadlines for ships under different conditions (LOADLINE)
- the reduction of marine pollution (MARPOL)
- the security of ports (ISPS Code)
- ensuring the appropriate management of ships (ISM Code)
- ensuring uniform standards of competence for seafarers (STCW)

In recent years, the IMO has very much focused on the environmental footprint of shipping. The IMO's responsibilities with respect to reducing pollution from ships are dispensed through its Marine Environment and Pollution Committee (MEPC), which also deals with reducing the greenhouse gas emissions of shipping and the industry's contribution to climate change, all of which have been addressed as relatively recently developed elements of the MARPOL convention.

From the perspective of the IMO, it is vital that shipping operates on the basis of globally uniform regulations; any local or regional deviations potentially result in regulatory conflicts, commercial distortion and administrative confusion that undermines the efficiency of shipping and, ultimately, international trade and the economic well-being of nations.

National Maritime Administrations

As implied by the fact that the membership of the IMO comprises national governments that propose, lobby for and eventually ratify regulations which impact the shipping industry globally, it should come as no surprise that those self-same governments have the main responsibility for enforcing IMO regulations within their national boundaries (including their territorial waters). This role is usually undertaken on behalf of the national government by some sort of national maritime administration, which may or may not be a part of the government or the civil service of a nation. Although the nature and name of such entities vary between countries and over time, one defining characteristic is that it administers the national register of ships and ensures that all the IMO's safety, environmental, and other international regulations are adhered to by all ships flying the national flag, irrespective of the country of residence of the shipowner or operator. As such, this responsibility is said to rest with the "flag state" and is sometimes referred to as "flag state control" ([Cullinane, 2010](#); [Branch, 2014](#)).

National maritime administrations also dispense their responsibilities through the implementation of "port state control," whereby ships flying other flags can be inspected when calling at any of the nation's ports – in order to ensure they are compliant with the international regulatory regime. Failure to meet the required standards can result in the detention of the vessel.

In implementing both flag state control and port state control, the national maritime administrations will rely heavily upon the technical expertise of marine surveyors that may be employed directly, but are more likely to be contracted from specialist organizations called Classification Societies that provide a service that validates all aspects of the quality of a vessel. The Classification Societies are fundamentally in competition worldwide for this sort of work, but twelve of the companies involved in the market (covering a 90% market share) do come together under the auspices of the [International Association of Classification Societies \(IACS\)](#) for the purpose of joint representation.

Supranational Organizations

In certain parts of the world, individual nations may be a member of a supranational political and/or economic entity (i.e., a bloc or union) and in some of these cases, a certain level of agreed authority and responsibility for shipping matters is exercised by the supranational entity. By far the most important of these is the EU, not only because of the particular legislative powers it holds over maritime matters, but also because of the economic power of the region and the prevalence of shipping in facilitating its trade, both externally and internally.

Much of the work on the development of the EU's maritime regulations is conducted by the [European Maritime Safety Agency \(EMSA\)](#) as one of the EU's decentralized agencies ([Walters and Bailey, 2013](#)). Based in Lisbon, EMSA was originally established as part of the political response to the Erika (1999) and the Prestige (2002) shipping accidents, where the resulting oil spills caused significant problems along the coasts of France and Spain. EMSA now provides technical assistance and support to the European Commission and individual EU Member States in the development and implementation of EU legislation on maritime safety, pollution by ships and maritime security. It also has operational responsibility in the region for oil pollution response, vessel monitoring, and the long-range identification and tracking of vessels.

There is no denying that the regulatory authority exercised by the EU in the maritime arena has sometimes brought it into conflict with the IMO, particularly in relation to environmental issues. In the past, the same was also true of the national government of the United States. Despite the problems that the non-uniformity of shipping regulations across the globe brings to shipowners and operators, the influence of the EU's role in such matters (as similarly was said of the United States) is widely perceived as having prompted the IMO into the implementation of appropriately more stringent regulations and speedier implementation.

Supply-Side Structures

Shipowners Associations

In nearly all of the major maritime nations, individual shipowners and operators (i.e., third party managers) are typically members of a national "Shipowners Association" which represents the interests of the national shipping industry in lobbying national maritime administrations and/or the government. Such associations also help to protect the interests of their members in relevant international fora ([Cullinane, 2010](#)). However, each national association is, in turn, likely to be a member of a supranational association representing the combined interests of a number of national shipping industries *en masse* within a particular area of the world. In the case of the EU and Norway, the [European Community Shipowners' Association \(ECSA\)](#) was founded in 1965 with a mission to promote the interests of the European shipping industry.

ECSA and other supranational associations represent a middle tier in a large pyramid structure of representation to which shipowners and operators have access. Sitting at a higher level again, the [International Chamber of Shipping \(ICS\)](#) is the world's largest organization representing the interests of shipowners and operators, having a membership of national, and other supranational, associations that together accounts for more than 80% of the world's total tonnage of merchant ships. The ICS concerns itself with the full range of regulatory, operational, and legal issues that affect shipowners or operators and has a role as a consultative body at the IMO, as well as with the other intergovernmental bodies which are relevant to the shipping industry; for example, the [World Customs Organization \(WCO\)](#), the [International Telecommunications Union \(ITU\)](#), the [United Nations Conference on Trade and Development \(UNCTAD\)](#), and the [World Meteorological Organization \(WMO\)](#). The ICS also maintains a close working relationship with a number of other trade associations that represent maritime interests, such as ports, pilotage, the oil industry, insurance, and classification societies.

The ICS claims to be unique in representing the global interests of all the different forms of shipping, encompassing the various ship types that operate within the different specific shipping markets. As this implies, there exist a number of other trade associations within the maritime arena that represent more specific interests that are usually aligned either with a particular form of cargo or a particular shipping market.

Intertanko

[INTERTANKO](#) (the International Association of Independent Tanker Owners) came into being in 1970 as a trade association representing independent tanker owners and operators at national, regional, and international levels. It has a mission to develop and disseminate best practice within the sector, in order to promote safer, cleaner, and competitive tanker operations in support of global oil, gas, and chemical product networks. This necessarily involves taking a lead on the development, acceptance and implementation of uniform, worldwide international tanker standards. In an effort to counter what is perhaps a negative public image, it also plays an important role in promoting the tanker industry by communicating its role, strategic importance and social value to both stakeholders and the public. In so doing, as a recognized non-governmental organization, INTERTANKO has a consultative role with the IMO, UNCTAD, and the [International Oil Pollution Compensation Funds \(IOPC\)](#). It also has a close working relationship with several other relevant organizations, such as the [Oil Companies International Marine Forum \(OCIMF\)](#), the [Chemical Distribution Institute \(CDI\)](#), the [Society of International Gas Tanker and Terminal Operators \(SIGTTO\)](#), the [International Association of Class Societies \(IACS\)](#), the [International Group of P&I Clubs](#), and the various maritime administrations involved in the regulation of the shipping industry.

Intercargo

Established in 1980, **INTERCARGO** is the short name for the International Association of Dry Cargo Shipowners and was created to represent the interests of a membership that primarily operate bulk cargo ships (i.e., bulk carriers) in the international dry bulk trades, such as coal, grain, iron ore, and other bulk commodities. The association seeks to achieve a safe, efficient, high quality, environmentally aware, and profitable dry cargo shipping industry within the context of free and fair competition within the market. It seeks to achieve this through the development of strategies, which enhance the interests of its members for the benefit of the dry cargo shipping sector as a whole. Much of its work toward meeting its objectives involves participation on a consultative basis at the **IMO** and, in order to avoid the costly and inefficient duplication of time and effort, close collaboration with other shipowners associations, such as the **ICS**, **ECSA**, etc. This latter is facilitated through a "Round Table of International Shipping Associations."

World Shipping Council (WSC)

The bulk-shipping sector (both wet and dry) constitutes one of the two main shipping sub-markets. The other major shipping market sector is the container-shipping sector. The interests of shipowners operating in this market are represented by the **World Shipping Council (WSC)**, the members of which comprise the world's major container (liner) shipping companies, accounting for approximately 90% of the global liner vessel capacity. The WSC was originally formed to represent the interests of the international liner shipping industry in discussions with the US government in relation to the reporting of freight rates (prices). Following the 9/11 terrorist attacks, however, the priority of the WSC became ensuring the security of international maritime commerce and the supply chains that trade depends upon, without impeding the free flow of commerce. In so doing, it forged close working relationships with the US government and the European Commission and is now based in Brussels. Its contemporary remit is to cooperate in the development of new laws, regulations and program, which are designed to enhance trade, security, the efficiency of customs procedures, and the environment, with respect to the international movement of containers. The WSC is also heavily involved in developing the laws on cargo liability and the use of IT in facilitating customs procedures for containers. These efforts are facilitated by the WSC having consultative status with the IMO as a recognized non-governmental organization. Particularly since many of its members are also the owners or operators of ports or port terminals, trucking companies, and other transportation modes, the WSC also involves itself on a worldwide basis in the development of adequate and efficient transportation infrastructure capacity.

Alliances

Aside from the formal representational role performed by WSC on behalf of container (liner) shipping companies, some form of cooperation between companies operating in this market has been a common feature of how the market operates. The precise form of cooperation has varied over time and in different geographical regions and, as a consequence, specific terms have evolved to define a certain form of cooperation. The most traditional form of cooperation takes the form of a liner shipping conference, which harks back to the 19th century and certainly, therefore, predates the containerization era. The conference system revolves around the cooperation of shipping companies operating on a specific liner trade route (usually defined in a single direction) agreeing to a predetermined tariff for the provision of shipping services. As such, many shipping companies were simultaneously party to multiple conference agreements; depending upon which routes they operated on. Due to concerns over the abuse of market power, in many parts of the world the conference system has been effectively outlawed, although it does continue to prevail in some parts of the world, notably in certain intra-Asian trade routes.

In the absence of the conference system, in an effort to preserve some form of cooperation between container (liner) shipping companies, at least at the strategic level if not operationally, a number of forms of cooperative arrangements have emerged over time. Having moved fairly swiftly through an era of container shipping consortia, we are now in an era where the shipping alliance (or ocean alliance) holds sway. This is a form of agreement where a number of container shipping companies join forces in global strategic cooperation across the various trade routes where they are jointly involved. In effect, the alliance plays exactly the same role as an individual container shipping company, but provides greater network coverage and scope. However, the individual shipping companies that constitute the membership of an alliance do remain firmly independent from each other; commercial information on cargoes, freight rates (prices), customer information etc, is not shared and arrangements between individual companies (lines) and land-based entities, such as terminals, hauliers, depots, CFS, etc. remain strictly confidential to the contractual parties involved.

Although some competitive concerns still remain about the nature of the shipping alliances, the logic underpinning their existence is that it is only by sharing resources in the form of ships, terminals in ports, and networks, that they can provide the level of shipping capacity required by the market demand. The global cooperation offered by being part of an alliance means that the variable costs of providing shipping services can be reduced to an extent that facilitates the optimal deployment of ship capacity within the container shipping networks in which a specific alliance operates. As such, the presence of alliances appears to be closely associated with fueling the development of mega-container vessels and mega-terminals to handle their cargoes. This has allowed the reaping of economies of scale in both ship and terminal operations, especially since alliances are then in a position to negotiate better tariffs at ports and push for volume discounts. These cost savings not only allow the member companies to stay competitive but also, it is argued, they can even be passed on to customers, thus refuting any possible allegations of the abuse of market power.

The deployment of ever-larger vessels to service the container trades and the responses that terminals have been forced to make in expanding their facilities, together with a tendency toward the concentration of the market via merger and acquisition activity, all point to the likely persistence and evolution of the strategic alliance model in container (liner) shipping.

The Demand for Shipping - Cargo Interests

There is no inherent desire to consume the services offered by transportation companies. The demand for shipping comes purely and simply from the demand that exists to consume the products, which are carried on ships. Thus, the demand for shipping services is defined as a “derived demand.” The ultimate source of this derived demand comes from the manufacturers, wholesalers, and retailers (collectively referred to as “shippers”) that are selling products to buyers (presumably, in order to make profits) where transportation by sea is required. In some cases, the shippers of products employ intermediaries such as freight forwarders; third party logistics service providers (LSPs) or other types of agent in order to make decisions about or to facilitate the transportation of freight. For the purposes of what follows, the all-encompassing term “shippers” can be interpreted as potentially including these agents.

Regional (Supranational) Shippers' Associations

The interests of manufacturers, wholesalers, retailers, and their agents are often represented by shippers' associations. These can and do exist at national level, but because of the multinational presence of many of these sorts of companies and the nature of international container (liner) shipping companies, regional (supranational) shippers' associations provide a more significant organization as the basis for shipper representation (Branch, 2014). The most important shippers' associations are the Asian Shippers' Alliance (ASA)^b, the [European Shippers' Council \(ESC\)](#), and the [American Association of Exporters and Importers \(AAEI\)](#). The members of regional shippers' associations comprise national shippers' associations, commodity trade associations and individual shippers who are bound together through a desire to ensure the safe delivery of their products to their customers in the right condition, at the right time, at the right price, and in the most efficient and sustainable way.

The existence of shippers' associations goes back a long way and is perceived to have significant advantages for not only shippers, but also for shipowners and operators. The most significant of the perceived advantages is the securing of low costs to members through the consolidation of purchasing and bulk buying on members' behalves. This process of consolidation has the additional benefit of increasing the buying power of smaller shippers. Other advantages are: rate stabilization and less transactions-based negotiations; reduced administrative burden on individual shippers through centralization; the use of structured agreements which facilitate members having access to more carriers and providing shipping companies with a larger pool of potential customers; easier compliance with the relevant regulations for both shippers and carriers; the sharing of market intelligence and; the use of common carriage terms and conditions, rather than on a contract by contract basis. The different regional shippers' associations have slightly different regional concerns, which are reflected in their activities. As just one example, the ESC seeks to help its members take maximum advantage of the internal European market and has vocal input into policy debates on the restrictions on the weight and dimensions of lorries, the hiring of personnel across EU member states and differences between the implementation of VAT and excise duties between EU member states.

The Global Shippers' Alliance (GSA)

The [Global Shippers' Alliance \(GSA\)](#) consists of the ASA, the ESC, and the AAEI. Its objective is to promote the beneficial evolution of regulations, markets and best practice to enhance trade and the flow of goods across the globe. It seeks to achieve this through the appropriate dissemination of information and knowledge, by facilitating better understanding between shippers and ocean carriers and by liaising with national and international authorities on maritime issues.

Ancillary Institutions

Having identified the major organizations and institutions involved on both the supply and the demand side of the overall shipping market, it is important to identify a number of other key players which have a role in facilitating the workings of the shipping markets.

UNCTAD

With its headquarters in Geneva in Switzerland, the [United Nations Conference on Trade and Development \(UNCTAD\)](#) was established in 1964 as a permanent intergovernmental body. Its purpose is to help developing nations overcome the difficulties faced in reaping the benefits from international trade. Recognizing the benefits that greater international trade has brought to

^b At the time of writing, the Asian Shippers' Alliance (ASA) does not have a website.

the world, UNCTAD is intent in ensuring a more equal distribution of those benefits (Cullinane, 2011). In its efforts to do so, UNCTAD has been pivotal in identifying the importance of transport costs and maritime connectivity in the facilitation of, and participation in, international trade, and the economic growth, which ensues from this. Working at the national, regional, and global level, this has then led to initiatives, which help to overcome the disadvantages faced in these respects by the developing world. These have encompassed the need for investment in appropriate transport infrastructure, a focus on greater efficiency in the use of infrastructure and the need for better governance structures in the management of that infrastructure; all of which have the intention of reducing the costs of maritime transport to and from developing nations and/or speeding the flow of goods. At a more structural level, UNCTAD also pursues the objective of helping developing nations move their production further up the supply chain and this too has implications and knock-on effects for its trade and the maritime transport which services it.

The Baltic Exchange

All contracts of carriage that are agreed between a shipper and a shipowner/operator contribute in some way to the current state of the shipping markets. That market may be defined very specifically in terms of a cargo or loading unit, a route and a size of shipment or, at the highest level of aggregation, as simply the bulk or container shipping market. Although some contracts of carriage are agreed directly between a shipper and a shipowner/operator (particularly in the liner trades, where containers are merely booked directly onto ships), some contracts of carriage are struck with the help of an intermediary shipbroker (particularly in the bulk trades), who may or may not be a member of a professional association called the [Institute of Chartered Shipbrokers](#), but who will almost certainly be a member of the [Baltic Exchange](#). The membership of the Baltic Exchange is responsible for a large proportion of all agreed contracts of carriage for bulk cargoes (both wet and dry), as well as the sale and purchase of merchant vessels. In all cases, the detail of the contract, particularly with respect to the agreed contract price, provides interesting and potentially useful market information to players engaged in the market (Stopford, 2009).

With its origins going back to 1744, the Baltic Exchange is based in London, with regional offices in Singapore, Shanghai, and Athens. Its membership of over 650 companies encompasses the majority of the world's major shipping players that, as members, adhere to a code of business conduct overseen by the Baltic Exchange and which is embodied in the institution's motto, "our word, our bond." With the access its members have to data on the deals agreed, primarily in the bulk shipping markets, the Baltic Exchange is in the position to aggregate this information and to disseminate it in a manner, which does not breach commercial confidentiality. As such, the Baltic Exchange is the world's leading, and only independent, source of maritime market information on the trading and settlement of both physical and derivative shipping contracts. Based on assessments made by a global panel of shipbrokers, the Baltic Exchange provides daily freight market prices and maritime shipping cost indices for a range of global shipping markets which are used as a guide or benchmark for the setting of freight rates (prices) in ongoing negotiations of contracts of carriage, as well as a settlement tool for freight derivative trades and as a general indicator of the bulk market's performance.

BIMCO

The nature of any contract of carriage agreed between a shipper and shipowner/operator will vary between particular shipping markets. In the majority of cases, the terms of the contract will be dictated by specific market characteristics and, as such, the detail contained within a specific market contract will not vary that much from deal to deal. As such, [BIMCO](#) has a major role in designing, updating and disseminating standard contract forms for use by the industry. When concluding a deal, either directly or via a shipbroker, the relevant party will start with the standard form for a contract of carriage for that specific shipping market and contract type and then edit it to encompass the fine detail as agreed between the shipper and shipowner/operator, most especially concerning the price and/or any tailored clauses that need to be included (Cullinane, 2010).

Biography

Kevin Cullinane is Professor of International Logistics and Transport Economics at the University of Gothenburg. Kevin has been a logistics adviser to the World Bank and transport adviser to the governments of Scotland, Ireland, Hong Kong, Egypt, Chile, and the United Kingdom. He holds an Honorary Professorship at the University of Hong Kong and is a visiting Professor at the VTI (the Swedish National Road and Transport Research Institute), Liverpool John Moores University, Dalian Maritime University, and Shanghai Maritime University. Prior to joining the University of Gothenburg on a permanent basis, he was appointed to the lifetime position of Honorary Visiting Professor at the University. Kevin has personally won research and consultancy projects to the value of \$US 6.5 m, many of which have informed the policies of national and/or regional governments. He has held other positions as the Director of the Transport Research Institute (TRI) at Edinburgh Napier University, Chair in Marine Transport and Management at Newcastle University, Professor and Head of the Department of Shipping and Transport Logistics at Hong Kong Polytechnic University, Head of the Center for International Shipping and Transport at Plymouth University, Postdoctoral Research Fellow at the University of Oxford Transport Studies Unit, and as Senior Partner in his own transport consultancy company, based in Cairo, Rennes, Hong Kong, Edinburgh and now Gothenburg. In terms of academic outputs, Kevin has published 14 books and over 300 journal papers and is an Associate Editor of *Transportation Research D*, *Maritime Economics & Logistics* and the *International Journal of Applied Logistics*. He is also a member of the Editorial Boards of *Transportation Research A*, *Transport Reviews*, the *Annals of Maritime Studies*, the *Journal of Shipping & Logistics* and, formerly, of the *Proceedings of IMechE Part M: Journal of Engineering for the Maritime Environment* and the *International Journal of Logistics Management*. Kevin has been a Chartered Fellow of the Chartered Institute of Logistics & Transport since 1997. His expertise has been recognized at UK national level with his appointment to REF 2014 and REF 2021; the UK's national research evaluation exercise.

References

- Branch, A., 2014. *Maritime Economics: Management and Marketing*. Routledge, London.
- Cullinane, K.P.B., *Shipping Economics*, 2005, *Research in Transportation Economics*, Vol. XII, Elsevier, Amsterdam.
- Cullinane, K.P.B., Talley, W.K., 2006. *Port Economics*. *Research in Transportation Economics*, Vol. XVI. Elsevier, Amsterdam.
- Cullinane, K.P.B., 2010. *International Handbook of Maritime Business*. Edward Elgar, Cheltenham.
- Cullinane, K.P.B., 2011. *International Handbook of Maritime Economics*. Edward Elgar, Cheltenham.
- Stopford, M., 2009. *Maritime Economics*, third ed. Routledge, London.
- Talley, W.K., 2017. *Port Economics*. Routledge, London.
- Walters, D., Bailey, N., 2013. *The Structure and Organization of the Maritime Industry*. In: *Lives in Peril*, Palgrave Macmillan, London.

Port Planning and Investment

Jasmine Siu Lee Lam, School of Civil and Environmental Engineering, Nanyang Technological University, Singapore

© 2021 Elsevier Ltd. All rights reserved.

Introduction	443
Port Planning Aspects	443
Seaward Planning	443
Terminal Planning	443
Landward Planning	444
Connecting the Three Planning Aspects	445
Port Planning and Development Process	445
Trends in Port Investment and Port Technology	446
Conclusion	447
Biography	447
References	447
Further Reading	448

Introduction

Seaports, hereafter ports, are interfaces between maritime transport and land transport. That is, ports connect these two modes of transport by providing services to the two broad groups of stakeholders. Ports facilitate trade and industries, which in turn contribute to a country's economic and social development. These fundamental roles performed by a port reflect the fact that port development affects not only the surroundings of a port, but also regional development as a whole. As such, a port is a critical coastal infrastructure which requires careful planning. The nature of this transport sector is capital-intensive, hence a key topic in port studies is port investment. This article aims to illustrate the fundamental concepts in port planning and investment. It will also discuss new trends in practice related to port planning and investment.

Port Planning Aspects

From infrastructure and physical connectivity's point of view, port planning can be broadly categorized into three aspects, namely seaward, terminal, and landward planning as summarized in **Fig. 1**. This section explains the three port planning aspects with examples from the port industry.

Seaward Planning

Seaward planning refers to all matters related to a port's maritime access in the harbor or port limit. Being an interface servicing ships, the scope of seaward planning covers infrastructure, facilities and services for ships and shipping, which include those for international ships (ocean-going vessels) and domestic ships (e.g., ferries, tug boats). This pertains to navigation channels (also known as fairways), navigation aids, locks, anchorages, pilotage, towage, breakwaters, bunkering, and shipyards. It can be seen that seaward planning entails a large spectrum of activities requiring highly specialized skills and knowledge. In fact, each activity can be considered an industry sector by itself or a subsector under the maritime industry ecosystem. For instance, bunkering is a sector for refueling ships; pilotage is a sector for guiding and navigating ships in domestic port waters.

In most ports, the entity overseeing the seaward planning is a country's maritime or port administration, such as a maritime authority. Service providers and operators can be a government organization, public enterprise or private enterprise. For example, in the case of Singapore, the Maritime and Port Authority (MPA) of Singapore, under the Ministry of Transport, is a government statutory board acting as a port planner. Navigation channels, navigation aids, anchorages, and breakwaters are provided and maintained by MPA. Pilotage, towage, bunkering, and shipyard services are provided by private companies.

Terminal Planning

Another aspect of port planning is terminal planning which refers to matters related to shore-based terminal areas. At an aggregated level, terminal planning considers strategic location selection and terminal layout design. The major components of terminal planning include berthing, loading and unloading system, storage yard, and gate system. From the planning perspective, berthing consists of quay and berth design, vessel information management, and berth allocation policies. A loading and unloading system involves stowage plan and cargo handling equipment allocation and scheduling. For storage yard planning, yard configuration,

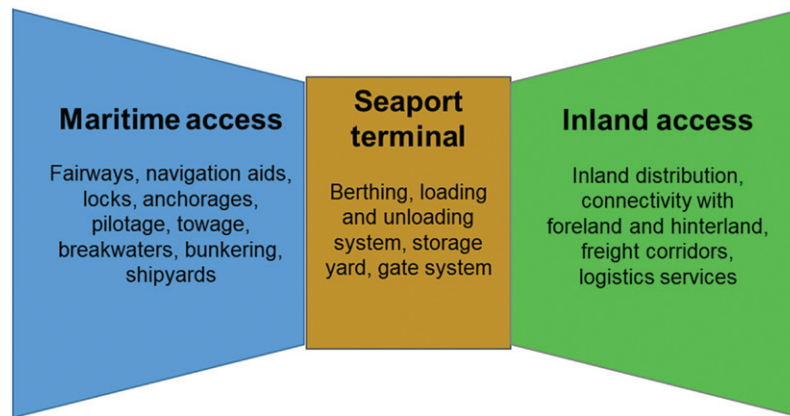


Figure 1 Three aspects of port planning and respective components.

cargo information management, as well as cargo handling equipment allocation and scheduling are required. A gate system planning includes receipt and delivery facilities, safety and security control policies, and intermodal transport connections. The four major components of terminal planning should be coherently designed to facilitate smooth terminal operations. This is the concept of integrative terminal planning. Thus, the various parts of a terminal should be well linked so any part would not become a bottleneck pulling down the terminal's performance.

From cargo's perspective, terminal planning has to cater to the characteristics of different cargo types. These include dry bulk, liquid bulk, gas and chemicals or other dangerous goods, breakbulk, general cargoes, perishables, roll-on roll-off cargoes (i.e., vehicles), project and special cargoes, as well as unitized cargoes mostly in containers. The cargoes can be very distinctive in nature, hence their handling and operating procedures are specialized. Cargo types affect every component of terminal planning as explained above. For example, berth design, cargo handling equipment, and storage facilities are all different for a dry bulk cargo terminal when compared to a liquid bulk terminal.

In terms of the party who manages terminal planning, there are various approaches in different ports due to different port ownership and governance models. The landlord model is popular in which a country's government owns the land while port superstructure, facilities (e.g., office buildings, warehouses, and cargo handling equipment) and operations are provided by a terminal operating company. The terminal operating company can be a state-owned enterprise, private firm, or hybrid entity or consortium such as joint-venture of a public organization and a private firm, that is public-private partnership (PPP) (Xiao and Lam, 2020). Therefore, quite a number of variations exist in port ownership within the landlord model. Using the same example as stated earlier, MPA of Singapore is in charge of the overall infrastructure aspect of terminal planning in Singapore while two terminal operating companies—PSA Corporation and Jurong Port Private Limited manage and operate the container terminals and multi-purpose terminals, respectively. Other ports using the landlord port model are Antwerp, New York, and Shanghai, to name a few examples. Because this article does not focus on port ownership and governance, readers are referred to the list of further reading to have a better understanding on this subject matter.

Landward Planning

Landward planning covers all issues related to a port's inland distribution and connectivity. This aspect extends from the gate system in a terminal to a port's foreland and hinterland. Ports should be well connected with maritime transport on one hand and inland transport on the other hand. For inland distribution, a port can interface with various modes of transport, including road, rail, inland waterway, air transport, pipeline, and a combination of these modes, that is intermodal transport. Associated distribution and logistics services are attributed to a port's performance. Inland transport operation is beyond the scope of a port per se, but the planning aspect is essential in order to ensure the port's accessibility. Effectively inland connectivity should be analyzed right at the start of any port planning, and it even affects the feasibility of a new port's location. Landward planning also includes the scope of freight corridors linking to industries where the cargo sources are located. Furthermore, port planning affects urban planning, or port planning is part and parcel of urban planning. Many ports and coastal cities are intrinsically linked. Planning includes, for instance, concerns on optimizing land use in view of competition for space from various sectors in an economy; codevelopment of a port and a city where the port is located (Lam and Yap, 2019).

Due to the reason that the scope of landward planning is wide and beyond the port itself, parties involved usually come from multiple organizations. That is, landward planning requires multiorganizational efforts. A country's maritime or port administration, such as a port authority may act as an overall coordinator to work with various parties including land transport authority and urban planning authority. They could jointly devise plans to develop a sustainable port city for the city and the port to support each other. Regarding service providers, typically public and private enterprises operate inland transport and logistics. For example, there are many private firms being trucking companies and third-party logistics service providers servicing port users in majority of ports in the world.

Connecting the Three Planning Aspects

Although the scope of the three aspects of port planning is very diverse, they should not be treated in isolation. The physical flow of a cargo ship calling at a terminal is briefly described here to outline how the various port planning aspects and elements are connected. When a ship enters a port through its navigation channels, the ship sails to the berth that is allocated. Depending on the port's regulation, pilotage services may be mandated for a marine pilot to navigate the ship. Before berthing, the ship may have to wait at the anchorage area if there is no available berth due to reasons such as congestion. Once a berth becomes available, the ship is towed toward the berth by a tug boat. While approaching the berth, quay cranes (or other suitable cargo handling equipment according to the cargo type) are allocated to the ship based on the volume of cargoes on board the ship to be loaded and/or unloaded. The process of cargo loading and unloading starts after berthing. For inbound cargoes, that is import, the cargoes will then be transferred to the yard area for temporary storage on the terminal before they are collected by consignees and leave the terminal for their destination, for example industrial sites. For outbound cargoes, that is export, cargoes come from the hinterland and arrive at the terminal. The cargoes will be transferred to the yard area for temporary storage before they are loaded on the ship for onward shipment. After cargo loading and unloading is completed, the ship will leave the berth, sails through the navigation channels and leaves the port eventually. Similar to ship arrival, pilotage and towage services may be required between the time of un-berthing and leaving the port limit. In the process, bunkering and other marine services like replenishing ship supplies could be performed at the same time when the ship is anchored or at berth. In other words, two or more marine services may be performed concurrently while cargo handling operation is conducted. For such cases, very good coordination among service providers is needed to ensure efficient and safe operations.

In addition, the role of ports in supply chain management has drawn growing interests in the industry. Ports are key nodes in supply chains linking various port operators, users and stakeholders including shipping companies, marine service providers, terminal operators, inland transport providers, government authorities, and urban planners. Therefore, the various port planning elements should be integrated to enable efficient physical flows and information flows. Port planning is required to take care of diversified interests of various parties, leading to its complexity. On one hand, it is a challenging task especially dealing with conflicting interests from different stakeholders. On the other hand, this can be regarded as an opportunity to differentiate from competing ports (Lam and Li, 2019). A port which is able to find win-win solutions will create and sustain value for port stakeholders, leading to a strategic competitive edge.

Therefore, port planning is a multifaceted process which entails technical issues, economic development, financial matters, environmental issues, political issues, legislative processes, urban planning, and stakeholder management. The overall environment faced by ports is constantly changing. Port planning is vital to the long-term success of a port, and it is also dynamic with various perspectives to be considered during the process. As a consequence, port planning should evolve in response to the changes and help a port to be competitive.

Port Planning and Development Process

After illustrating the three port planning aspects, this section explains the planning process and considerations of developing a new port, expanding or upgrading an existing port. This is also known as a port master plan. As a starting point, the objective of a port planning and development project should be identified. Then the planning process is designed to cater to the interests of different stakeholders as well as the development of the local economy.

1. Existing demand, supply, and pricing

Similar to any economic framework, demand, supply, and pricing are basic elements determining port planning and development. Existing demand of a port indicates the current traffic volume needs to be handled in the port, based on data such as cargo, passenger, and vessel statistics. Existing supply is the capacity already being built to accommodate and handle demand, usually depending on the capacity of infrastructure and superstructure. Analyzing demand and supply data helps to find out if there is imbalance between the two elements. A port should also review the current pricing structure to assess if any changes are required. Pricing measures can be used as a short-term solution to influence customer and supplier behavior. As for medium to longer term, insufficient supply or increasing demand can trigger the need for port development, for example, upgrading or expanding a terminal to have additional capacity to handle more ship calls.

2. Performance indicators

Port performance statistics could be generated from current demand and supply data. This indicates the performance of a port in a specific and quantitative manner, referring to the matching level of demand and supply. Some examples of port performance statistics are ship turnaround time, ship waiting time, berth utilization rate, and truck turnaround time. Performance statistics also implies the productivity level and trend of a port, which is then used to derive future demand and align with corresponding supply. Also, performance indicators reflect whether the service quality of a port has met the desired level as set by port managers and areas for improvement when needed. The information is important for identifying port planning and development projects in next steps.

3. Traffic evolution

Traffic evolution, of which data could be derived from cargo, passenger and vessel statistics, is an important factor driving the development of a port and should be considered in a port plan. In general, potential changes in production, market, trade, goods

or cargoes, and shipping lead to different trends and will influence traffic evolution (Yap and Loh, 2019). These include trends in commerce or trade patterns, trends in shipping, trends in physical appearance of goods, and trends in cargo handling. The rationale behind is that shipping is a derived demand and ports are built to support shipping activities. Future developments in trade, shipping and products will alter the demand and traffic patterns of a port. Hence, port expansion or alternation works may be required. However, a port may face competition from other ports so developments in other ports should be taken into account.

4. Development plan

On the supply side, based on the current condition and limitation in capacity of infrastructure and superstructure as well as the future desired supply capacity, a port development plan will be proposed. It is important to note that the plan should be aligned with a port's long-term goal and strategy. For instance, whether a port aims to attract transshipment traffic or mainly focus on local hinterland traffic significantly affects a development plan. At the same time, constraints and limitations faced by the port should be examined. Inherent limitations such as scarce land and resource, as well as regulatory requirements such as government policies should also be taken into consideration. A draft port development plan should go through cost and benefit analysis. After a new development plan is clearly defined, the objectives of port planning are more detailed, including capacity planning which serves as an important guide during construction. Specifically, four essential components including alternative design, financial considerations, environmental considerations, and regulatory requirements are the key determinants. They will be elaborated as follows.

5. Alternative design

This is more of a construction and engineering aspect with considerations of geography, ship–shore interface, and types of ship, cargo and passenger. Related to the above-mentioned aspects of port planning, the scope of design could be the configuration and layout of a navigation channel, harbor basin, terminal area and foreland/hinterland connection. Geotechnical and natural conditions such as coastal environment, soil, wave and weather are included in the feasibility assessment of alternative designs. For long-term development of a port, potential expansion in the future should also be accounted for.

6. Financial analysis and investment plan

Thorough financial analysis consisting of capital cost, operating expense, return on investment, infrastructure and superstructure maintenance cost, cash flow, etc. is an integral step in port planning. Simultaneously, an investment plan is required to explore and examine options of funding port projects and how to utilize suitable financial instruments such as taking loans, issuing bonds, and seeking government subsidies. Funds may be sourced from the public sector, the private sector, or a mix of the two sectors in terms of a PPP. Also, funds may come from local sources or foreign investors. In most cases, the mode of investment and imposed conditions depend on the ownership model of a port which is established by the government.

7. Environmental Impact Assessment and other regulations

Environmental considerations should also be emphasized. With reference to the Organisation for Economic Co-operation and Development (OECD), negative environmental impacts include noise, waste, emissions of particles, greenhouse gas and pollutants, and dust caused by handling substances such as grain, sand and coal (Braathen, 2011). For example, reclamation and dredging work in a port to expand port capacity can greatly affect the ecosystem on which the marine lives rely. In addition, road and rail traffic connecting from a port to hinterland also generates pollution. Therefore, Environmental Impact Assessment, usually including elements such as Initial Environmental Examination and Environmental Impact Studies, is required by many port authorities or regulators. Approval is required before conducting any port projects.

There could be other regulations that port planners have to observe, for example, those related to land use, handling of hazardous substances, port and shipping safety, security, labor, free trade zone, and antitrust law. The appropriate legislative procedures and approvals should be incorporated in the port planning process.

8. Final development project

After various steps are conducted to analyze a port project's feasibility, a port planner should integrate the analysis, consider costs, benefits, and uncertainties of different alternatives holistically, and decide on a final development project. For instance, project timeline, budget, resource, development phases, and other details should be specified.

Trends in Port Investment and Port Technology

Port development demands tremendous amounts of resources and professional expertise that may not be readily available in many port authorities. Port liberalization and privatization started in the late 1970s which have led to the prevalence of the landlord port model. In this port governance model, a government authority owns the land where infrastructure is leased to privately owned companies or a joint venture which is owned jointly by the public sector and the private sector. Therefore, it is a form of PPP which is a contractual form of collaboration between the government and the private sector. The progressive opening of the port sector worldwide presents opportunities for international terminal operating companies to enter into markets of different regions. The port industry also sees shipping companies investing in port projects. These trends are common to both developed and developing countries.

After many years of globalization and liberalization, however, protectionism takes place in some cases such as the withdrawal of the United Kingdom from the European Union, that is Brexit, and tensions between China and USA. These create undesired instability and risk in port investment. For example, USA plans to increase scrutiny of foreign direct investment in the country

from China and other foreign investors including those in seaports, claiming reasons of national security (The New York Times, 2019).

An imperative implication for port planning and investment is the need for a supportive institutional structure beyond a port by itself. A port is a kind of system in a national socio-economic-political system as well as the globalized economic system. As such, port investment matters ought to be considered in the context of structural factors of the overall systems. Port development of a country is determined by the political will of its government. Trade, investment and other policies alike shall continue to play important roles. The port industry is expected to encounter changes and uncertainties, while port investors should actively look for opportunities in growing markets.

A key area of opportunity in the 21st century is technology. It is an essential element in every aspect of port planning and operations. On the demand side, technology is a factor influencing existing demand from shipping and trade; meanwhile, the demand will trigger the development of technology. New technologies also change transport patterns by leading new trends. To cite several examples, bigger vessels increase the cargo volume to be handled and call for higher productivity from terminal operators; autonomous vessels draw the focus to new ways of handling these vessels safely; e-commerce alters consumer behavior and seaborne trade routes; blockchain technology leads to changes in shipping documentation procedures that government authorities and port operators have to cope with. On the supply side, new technologies applied in port infrastructure and superstructure can increase the capacity of a port and create more alternative designs when a port plan is proposed. Advanced technologies are also able to overcome some environmental issues as well as port limitations such as the use of full electric cargo handling equipment drawing from renewable energy sources (Iris and Lam, 2019). Together with port performance statistics and indicators, technology factors also determine productivity trends by affecting vessel and cargo handling performance (Vanelslander et al., 2019). For instance, new quay crane design, a higher level of automation, deploying Internet of things technology like sensors contribute to port productivity.

Furthermore, appropriate use of technology helps a port to become more intelligent. Data and information about vessels, cargoes, customers, assets, and other elements can be analyzed to facilitate decision making. By leveraging on data mining tools, port planners and operators may realize the benefits brought by descriptive, diagnostic, predictive, and prescriptive analytics. For example, an advanced vessel traffic management system is capable of enhancing efficiency in port waters by tracking vessel movement real-time. Collision avoidance and detection of abnormal activities lead to better safety and security of the port. Technologies are rapidly evolving in the contemporary era, those ports which make strategic moves to capture opportunities and create value for users and stakeholders will be well-positioned in the sector.

Conclusion

In times of changing trends of globalization and trade patterns, the port industry and its actors are encountered with higher market volatility and uncertainties. Combining with random shocks like negative impacts on the economy brought by political tensions among countries and pandemic result in complications for decisions in port planning and investment. Ports being very long-term critical infrastructures serve national, regional and even international interests. Hence, port planners and policy makers are recommended to embrace comprehensive strategic planning and innovation to always stay relevant.

Biography

Prof. Dr. Jasmine Siu Lee Lam is Centre Director of Maritime Energy and Sustainable Development Centre of Excellence at Nanyang Technological University, Singapore. She is also Director of Maritime Studies Programme. Her major research areas are maritime and port economics, sustainable development, logistics, and risk management. She has been invited by various organizations, port authorities, companies, and universities as a key speaker at international conferences. Leading a R&D team and working closely with the industry and government agencies, Jasmine has completed over 50 projects including consultancy to many countries and has more than 280 scientific publications including over 130 top-tier technical journal papers. She serves as the editor/board member of 10 international journals, including Maritime Policy & Management as Associate Editor. She is a Global Maritime Technology Cooperation Centres Network Stakeholders' Committee member, invited by International Maritime Organization, and the Vice President of the International Association of Maritime Economists. Furthermore, Jasmine is the recipient of many awards, including Best Paper Awards, Erasmus Mundus Faculty Scholar Award awarded by the European Commission, and National Day Award from the Prime Minister's Office in Singapore.

References

- Braathén, N. (Ed.), 2011. Environmental Impacts of International Shipping: The Role of Ports. OECD Publishing, Paris.
- Iris, C., Lam, J.S.L., 2019. A review of energy efficiency in ports: operational strategies, technologies and energy. *Renew. Sustain. Energy Rev.* 112, 170–182.
- Lam, J.S.L., Li, K.X., 2019. Green port marketing for sustainable growth and development. *Transport Policy* 84, 73–81.
- Lam, J.S.L., Yap, W.Y., 2019. A stakeholder perspective of port city sustainable development. *Sustainability* 11, 447.
- The New York Times, 2019. U.S. outlines plans to scrutinize Chinese and other foreign investment, Sept. 17, 2019.
- Vanelslander, T., Sys, C., Lam, J.S.L., Ferrari, C., Roumboutsos, A., Acciaro, M., Macário, R., Giuliano, G., 2019. A Serving innovation typology: mapping port related innovations. *Transport Rev.* 39 (5), 611–629.
- Xiao, Z., Lam, J.S.L., 2020. The impact of institutional conditions on willingness to take contractual risk in port public-private partnerships of developing countries. *Transport. Res. Part A: Policy Pract.* 133, 12–26.
- Yap, W.Y., Loh, H.S., 2019. Next generation mega container ports: implications of traffic composition on sea space demand. *Maritime Policy Manage.* 46 (6), 687–700.

Further Reading

- Brooks, M.R., Cullinane F.K., (Eds.), 2007. Devolution, Port Governance and Port Performance, Research in Transportation Economics, Volume 17. Elsevier.
- Brooks, M.R., Cullinane, K., Pallis, A.A., 2017. Revisiting port governance and port reform: a multi-country examination. *Res. Transport. Business Manage.* 100 (22), 1–10.
- Ha, M.H., Yang, Z.L., Lam, J.S.L., 2019. Port performance in container transport logistics: a multi-stakeholder perspective. *Transport Policy* 73, 25–40.
- Hales, D.N., Lam, J.S.L., Chang, Y.T., 2016. The balanced theory of port competitiveness. *Transport. J.* 55 (2), 168–189.
- Lee, P.T.W., Lam, J.S.L., 2017. A review of port devolution and governance models with compound eyes approach. *Transport Rev.* 37 (4), 507–520.
- Robinson, R., 2002. Ports as elements in value-driven chain systems: the new paradigm. *Maritime Policy Manage.* 29 (3), 241–255.
- Yap, W.Y., 2020. *Business and Economics of Port Management*. London, Routledge.
- Zhang, Y., Kim, C.W., Tee, K.F., Lam, J.S.L., 2017. Optimal sustainable life cycle maintenance strategies for port infrastructures. *J. Clean. Prod.* 142 (4), 1693–1709.

Estimating the Capital Stock of Transport Infrastructure

Heike Link, German Institute for Economic Research Berlin (DIW Berlin), Department Energy, Transport, Environment, Berlin, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	449
Indirect Method by Modeling Approaches (Perpetual Inventory Method)	450
Basic Idea	450
Gross Capital Stock and Retirement Distributions	450
Normal Distribution	450
Lognormal Distribution	451
Winfrey Distributions	451
Weibull Distributions	452
Gamma Distribution	452
Polynomial Retirement Function	453
Discussion	453
Service Lives	453
Net Capital Stock and Depreciation	453
Gross Fixed Capital Formation (GFCF)	454
Estimating an Initial Value of the Capital Stock—The Direct (Synthetic) Method	454
An Example of Capital Stock Estimates for Transport Infrastructure	455
References	455
Further Reading	456

Introduction

Transport infrastructure (e.g., roads, tracks, and stations of rail and public transport, inland waterways, ports and harbors, airports, and pipelines) plays an important role as capital input into production and the generation of wealth. It is a necessary input for the production of transport services which are in turn necessary for the exchange of final products and inputs and for broader welfare benefits such as travel time savings. For various types of economic analysis, such as productivity and efficiency analysis, cost and production function studies figures on the value of transport infrastructure, for example, the capital stock is a crucial input.

In most OECD countries, estimating the value of capital stocks of industrial sectors is a common procedure, based on the worldwide methodological standards defined within the System of National Accounts (SNA), and within its European equivalent, the European System of Accounts (ESA). Transport infrastructure is covered in different sections of the SNA classifications of assets, institutional sectors, and economic activities. It is mainly included in parts of sections H (transport, storage, and communications) and L (public administration, defense, and social security). This lumping together with other industries and the spread over different subsections of the SNA implies that an overall figure for the capital stock of transport infrastructure is in almost no country available. An exception is the road network for which several OECD countries publish road investments and capital stock values.

In general and not only related to transport infrastructure, measuring capital faces many empirical issues which arise due to the absence of observable economic transactions; they relate to questions such as not only how to measure flows and stocks, but also how to estimate volumes and prices of flows of capital services. In the transport sector, a range of institutional settings such as pure public ownership versus private and mixed public/private ownership adds complexity to the measurement of both investments and capital stocks. Capital stock calculations for transport infrastructure face additional problems due to the long service lives of assets.

As each type of capital, infrastructure capital can be viewed from two sides, from the income and wealth perspective, and from the production and productivity perspective. Depending on the chosen perspective, there are different measures of capital stock such as gross capital stock, net capital stock, and productive stock, alongside with the relevant measures of economic flows such as investment, depreciation, and capital services. The central economic relationship between these two perspectives is the so-called net present value condition according to which in a functioning market the stock value of an asset is equal to the discounted stream of future benefits (capital services) which the asset is expected to yield. For transport infrastructure, however, there are considerable difficulties in estimating future benefits that can be derived from an asset. This is the reason why alternative approaches, in particular the indirect valuation by using the perpetual inventory method (PIM), are applied. It should also be noted that the concept of capital services has been increasingly gained attention of researchers and has been suggested as an add-on in the 2008 version of SNA but has not yet formed a compulsory part of national accounts. For transport infrastructure, it is particularly difficult to apply, and given that it is a non-compulsory element of SNA, it will not be discussed further in this encyclopedia.

Capital stock accounting in company balance sheets is common practice for all private and also publicly owned companies; relevant examples are private motorway companies and rail companies. Apart from this, there are two approaches to determine the

value of capital stocks. These are the direct method, sometimes also referred to as synthetic method, and the indirect method, also referred to as cumulative or PIM. The focus in this encyclopedia will be on the PIM, which is the most common tool in the SNA of most OECD countries. Since the direct method is used as basis for deriving motorway tolls in some countries, and due to its potential role for compiling initial capital values as input for the indirect method, this encyclopedia gives a brief summary on this method. Entrepreneurial roles of calculating capital values of private and/or public companies will not be covered here, they can be found in the relevant literature in this field. It should be noted that these values are not comparable with those derived with the other two methods due to different evaluation principles, depreciation periods, etc.

Indirect Method by Modeling Approaches (Perpetual Inventory Method)

Basic Idea

The PIM is based on the idea that stocks are composed of cumulated flows of investments minus retired assets and efficiency losses of assets. This idea can be used to calculate both a gross capital stock (G_t) and a net capital stock (N_t). Starting with existing information on the gross capital stock of year t (G_t), the capital stock of the subsequent year $t + 1$ (G_{t+1}) can be calculated by adding the flow of investments ($I_{t,t+1}$, also referred to as gross fixed capital formation [GFCF]) and by deducting the retirement of assets ($A_{t,t+1}$).

$$G_{t+1} = G_t + I_{t,t+1} - A_{t,t+1} \quad (1)$$

The net capital stock (N_{t+1}) can be obtained in a similar approach, here by taking into account depreciation ($D_{t,t+1}$) of those assets which have survived from past investments.

$$N_{t+1} = N_t + I_{t,t+1} - D_{t,t+1} \quad (2)$$

The PIM requires long time series on annual investment expenditures for homogeneous groups of assets, information on service lives, and initial values on the capital stock as starting values for the PIM. The time series of investment expenditures should be sufficiently disaggregated into technically homogeneous capital goods to enable an adequate assignment of life expectancies. Depending on the asset's type and the assumed service life, these time series have to reach very long back into the past. Furthermore, an initial value for the capital stock is required except the case that the available investment time series exceeds the life expectancy of assets. Apart from the long investment time series itself, the calculation of capital stocks with the PIM requires a long time series of deflators to convert all investment vintages at the price level of a chosen reference period.

Gross Capital Stock and Retirement Distributions

The gross capital stock contains the stock of all assets surviving from past investments reevaluated at prices of new capital goods of a reference period. It can be thought of as the value of assets before deducting consumption of fixed capital (CFC), for example, economic decay of assets is ignored and past investments are considered as if they were new. For calculating the gross capital stock, a retirement or mortality profile of assets has to be assumed which models the retirement process of an asset cohort over time. The retirement function is a probability function with the underlying assumption that the life expectancies of assets within an investment vintage are dispersed around the mean value. Retirement is defined as the removal (discard) of an asset from the capital stock because it is abandoned, dismantled, sold for scrap, or exported. It should be noted that retirements are to be distinguished from disposals; this latter category also includes asset sales as secondhand goods for continued use in the production process. This latter category is of lesser importance for transport infrastructure assets, even though examples for second use of capital goods do exist (e.g., equipment, rail tracks from heavy track classes which are used at lines with lower track weight classes). The inverse of the retirement function is the survival function that describes the share of assets, which are still in use. An important assumption in capital measurement through the PIM is that assets are properly maintained and can be used until they reach their defined life expectancy. Retirements due to unforeseen events such as natural disasters are not considered within the retirement function but are treated a separate category.

The key parameter of the retirement profile is the average service life of an asset or asset group, and retirement functions can be truncated to fix a maximum service life. In practice, a range of functional forms for the retirement patterns is used (see [Table 1](#) which summarizes retirement functions, average service life assumptions, and depreciation patterns in OECD countries, and [Fig. 1](#)). Most of them are bell-shaped functions where retirements start gradually after the year of installation, build up a peak around the average service life, and decline gradually during the years after the average service life. Bell-shaped retirement patterns include the normal/lognormal, Gamma, Winfrey, and Weibull distributions.

Normal Distribution

The use of the normal distribution for retirement patterns assumes that asset retirements are symmetric with 95% of the probabilities to retire lying within two standard deviations around the mean service life. It is for example used in Hungary, in Malta (for non-dwellings), and in the United Kingdom; truncated normal distributions are applied in Israel and Italy.

Table 1 Average service life, retirement functions, and depreciation patterns for roads in official capital stock calculations of OECD countries

	<i>Service life</i>	<i>Retirement pattern</i>	<i>Depreciation pattern</i>
Australia	33	S3-Winfrey	Hyperbolic age-efficiency profiles
Austria	N/A	None	Geometric, rate = 0.030
Belgium	55	Lognormal	Linear
Canada	N/A	None	Geometric, rate = 0.106
Chile	40	S3-Winfrey	Linear
Cyprus	55	R3-Winfrey	Linear
Czech Republic	50	Lognormal	Linear
Denmark	50	Winfrey	Linear
Estonia	55	Linear	Linear
Finland	55	Weibull	Linear
France	120	Lognormal	Linear
Germany	57	Gamma	Linear
Hungary	75	Normal	Linear
Iceland	77	None	Geometric, rate = 0.030
Israel	50	Truncated normal	Linear
Italy	N/A	Truncated normal	Linear
Japan	N/A	None	Geometric, rate = 0.033
Korea	60	R3-Winfrey	Hyperbolic age-efficiency profiles
Latvia	50	Lognormal	Linear
Lithuania	60	None	Geometric, rate = 0.033
Malta	100	Normal	Linear
Mexico	60	None	Linear
Norway	60	None	Geometric, rate = 0.033
The Netherlands	25–55	Weibull	Hyperbolic age-efficiency profiles
Portugal	60	Delayed linear	Linear
Slovakia	60	None	Linear
Slovenia	50	None	Linear
Sweden	40	None	Geometric, rate = 0.0202
United Kingdom	80	Normal	Linear
United States	N/A	None	Geometric, rate = 0.0202

Source: Own compilation based on information given in Eurostat (2014).

Lognormal Distribution

In the lognormal distribution, the logarithms follow a normal distribution with the mean μ and the standard deviation σ . It is a right-skewed distribution, for example, the probability of assets to retire is zero in the first year after installation and is concentrated in later years of an asset's life whereby the right-hand tail of the distribution has arbitrarily to be set at zero. The frequency distribution—with T as the age of assets—is:

$$F_T = \frac{1}{T\sigma\sqrt{2\pi}} e^{-(\ln T - \mu)^2 / 2\sigma^2} \quad (3)$$

μ and σ are calculated as $\mu = \ln(m) - 0.5\sigma^2$ and $\sigma = \sqrt{\ln(1 + (m/s)^{-2})}$ with m and s being the mean and the standard deviation of the underlying normal distribution. With m representing the average service life of an asset, commonly used parameter values for s in capital stock estimation in the European Union (e.g., in Belgium, the Czech republic, France, Lithuania, and Cyprus) are between $m/2$ and $m/4$. Empirical support for these values comes from discard surveys performed by the statistical office in France.

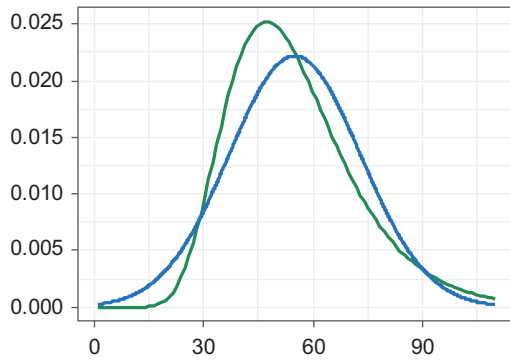
Winfrey Distributions

The group of Winfrey curves covers 18 typical empirical distributions (Winfrey, 1935). They are named after Robley Winfrey, a research engineer at the IOWA Engineering Experimentation Station, who collected information on installation and retirement dates of 176 groups of industrial assets, out of them also for railway assets and for road pavements. The Winfrey curves are denoted by L, S, and R for left modal, symmetrical, and right modal, and by the numbers 0–6 for flatter to more peaky curves. The most commonly used Winfrey curves belong to the group of symmetric S curves which is defined as:

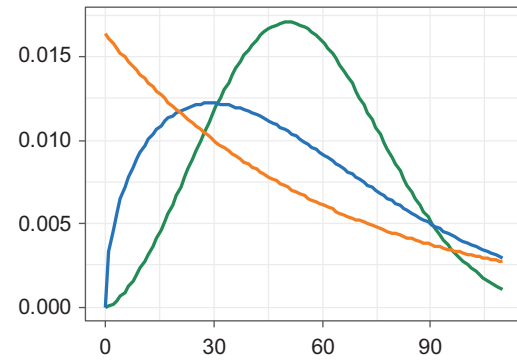
$$F_T = F_0 \left(1 - \frac{T^2}{a^2} \right)^m \quad (4)$$

F_T is the marginal probability of an asset to retire at age T where the age T is expressed as the share of the average service life (in Winfrey's work expressed in deciles). This implies that T varies from zero to infinity and F_T takes the largest value at the average

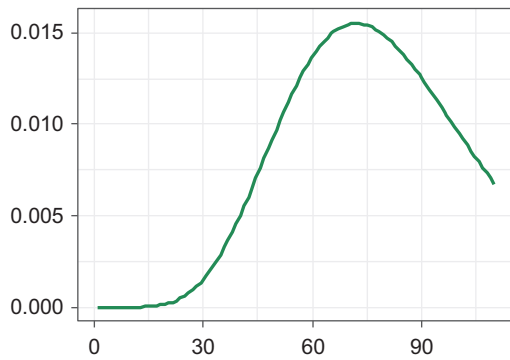
Normal and lognormal
Blue = normal, green = lognormal



Weibull
Lambda = $b_1/0.0164$, alpha: green = 2.6, blue = 1.5, orange = 1.0



Gamma
Lambda = 9, alpha = 9



Third-degree polynomial
A0 = 0.00027, A1 = -0.00011 A2 = 1.18e-05
A3 = -9.87e-08

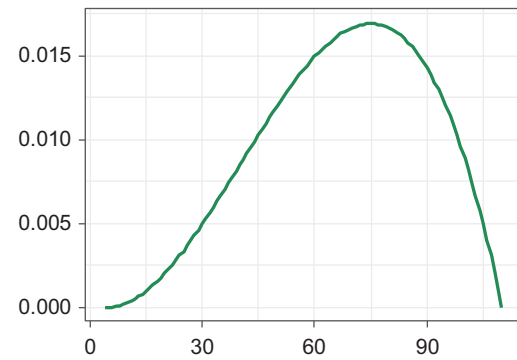


Figure 1 Different retirement functions used by OECD countries in capital stock estimation—density functions for service lives. *Source: Own presentation based on parameter values used by several OECD countries.*

service life. The most widely used Winfrey curves are the symmetrical S2 and S3 functions with the parameter values $F_0 = 11.911$, $a = 10$, and $m = 3.7$ for S2, and with $F_0 = 15.61$, $a = 10$, and $m = 6.902$ for S3. Winfrey-based retirement patterns are used for example in Sweden, Denmark, Australia, Korea, and Chile.

Weibull Distributions

The Weibull function named after the Swedish mathematician Waloddi Weibull (Weibull, 1951) is widely applied in engineering disciplines to study failure times of materials, and has also been used in modeling mortality in natural populations. The flexibility of the functional form allows accommodating similar shapes as the Winfrey curves. The Weibull frequency function, again with T as the age of the asset, is given as:

$$F_T = \alpha \lambda (\lambda T)^{\alpha-1} e^{-(\lambda T)^\alpha} \quad (5)$$

where $\alpha > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution.

Weibull retirement functions are used by the statistical offices in the Netherlands and Finland for capital stock estimations. The parameter values for the Weibull retirement functions used in the Netherlands are based on a discard survey conducted by Statistics Netherlands which yielded for buildings λ -values in a range from 0.021 to 0.05 and α -values between 0.97 and 2.21. For other tangible fixed assets similar ranges of 0.028–0.108 for λ and of 0.98–2.63 for α were obtained (Meinen et al., 1998; Rooijen-Horsten et al., 2008).

Gamma Distribution

The Gamma distribution is used by the German Statistical Bureau for modeling asset retirements. In absence of a discard survey for industrial sectors, it is empirically based on observed patterns of car registration and retirement. The Gamma retirement distribution is:

$$F_T = a p \Gamma(p)^{-1} T^{p-1} e^{-aT} \quad (6)$$

The parameters a and p which determine the shape of the curve are set equal to 9 for most capital goods in Germany (Schmalwasser and Schidlowski, 2006).

Polynomial Retirement Function

The capital stock estimation for German transport infrastructure, compiled annually by the German Institute for Economic Research (DIW Berlin) in addition to the official SNA-based capital stock calculations of the German Statistical Bureau, uses a third-degree polynomial retirement function. Under the assumption that retirements start at year t_s with $t_s > t_0$, the parameters of the polynomial

$$F_T = \beta_0 + \beta_1 T + \beta_2 T^2 + \beta_3 T^3 \quad (7)$$

can be obtained from the lower and upper limits of the service life intervals s and L , with a mean service life of $m = s + 0.6(L-s)$. This right-skewed form has found to reflect the specific long life expectancy of assets and the fact that mostly discards of such assets are concentrated at the right-hand side of the tail (see Link et al. (1999) for an application to roads in different EU countries and Link et al. (2019) for a more recent discussion on the specific properties of transport infrastructure capital).

Discussion

It should be noted that discard surveys as basis for statistical estimation of retirement functions and average service lives are almost impossible to conduct for transport infrastructure. Annual discards or sales of pavements, rail tracks, runways, etc., are in practice not relevant. Official calculations of statistical offices therefore transfer service lives and retirement functions from other sectors to the transport sector.

Service Lives

Service life of assets—one of the most crucial parameters of the PIM—is defined as the length of time that an asset remains in the capital stock. It includes both the time in the capital stock of the purchaser and in the stocks of other producers who bought the asset as secondhand good. Asset life in capital stock estimation models means the economic, not the physical or engineering notion of service life, for example, assets could physically still be in use but are discarded due to economic obsolescence. Most retirement distributions require defining the average service life, to be distinguished from the maximum service life of an asset within a cohort of assets. Sources to obtain service life information include tax authorities (tax lives), company accounts, expert estimates, and statistical surveys. For transport infrastructure, tax lives play a minor role, but can give indications on service life, in particular for equipment goods in ports, stations, or airports. Although the PIM requires the economic lifetime of assets, engineering-based information on service life is a relevant starting point for transport infrastructure, in particular regarding the relationship between traffic volume and infrastructure damage at roads and rail tracks. Table 1 gives an overview on service life assumptions for roads, together with the retirement pattern, for selected OECD countries.

Transport infrastructure has characteristics that add complexity to the definition or estimation of service lives compared to other sectors. In particular, the long lifetimes of 50 years and above imply that complete observations of service lives are often lacking, and the mortality function has to be chosen without complete observations of an empirical distribution. The cause–impact relationship between service life and various impact factors such as construction standards and construction quality, dimension (e.g., thickness of road layers, weight class of rail tracks), climate, topography, traffic volume, and traffic composition is complex. To this adds the problem that conventions on a minimum quality of service have to be defined which impact on the service life of infrastructure that can differ from a pure physical life expectancy.

Net Capital Stock and Depreciation

The net capital stock, also referred to as wealth stock, is the stock of assets that have survived from past investments minus the annual depreciations. Estimates of net capital stocks have to be considered within a broader set of capital measures that reflect the dual role of capital as storage of wealth and as the source of capital services in the production of goods and services.

For calculating net capital stocks, a depreciation function has to be derived which reflects the decrease of the value of assets during their service life. In the SNA, depreciation is defined as the decline or loss of the current value of the stock of fixed assets due to physical deterioration, normal obsolescence, or normal accidental damage during the accounting period. The decline or loss of asset value is the reason why depreciation is also referred to as CFC; in national accounting of the United States the term capital consumption is used.

Since, for transport infrastructure, information on depreciation cannot be derived from observed market transactions, assumptions are necessary. In principle, there are two approaches to define and quantify a depreciation profile. For a single homogeneous asset, the age–price profile shows how the asset price declines as a consequence of ageing, reflecting CFC due to physical deterioration (wear and tear) and normal obsolescence. In practice, this is the most common way and statistical offices often assume a linear depreciation function with constant absolute values of depreciation over time. The second approach is to consider the age–efficiency profile of the homogeneous asset that reflects the change of productive capacity over time. It provides the link to measuring capital services, for example, the asset's contribution to production. Both profiles are related and can be derived from each other. The most common depreciation patterns are linear (straight line) and geometric (see Fraumeni (1997) for an outline of both approaches).

While for a linear depreciation pattern, an asset with a service life of T loses each period a constant amount $D = 1/T$ until the asset value is zero (e.g., constant absolute values in each period), in a geometric depreciation pattern an asset loses its value at a constant rate δ each year. Geometric depreciation profiles are also called to be “convex to the origin” since the amount of absolute depreciation is highest at the beginning and declines over time. Empirical evidence that geometric profiles are appropriate for many types of assets is given in [Hulten and Wykoff \(1981\)](#) and [Koumanakos and Hwang \(1988\)](#). An overview on depreciation studies can be found in [Jorgenson \(1996\)](#). It is recommended to derive the parameters for geometric models of depreciation from econometric studies based on empirical information on prices for used assets. For transport infrastructure, but also for various other types of capital goods this is often hampered by the lack of such empirical information. A simple method for obtaining a geometric depreciation profile in such cases is the double-declining balance method where the rate of decline is given as $\delta_i = 2/\bar{T}_i$ with $\bar{T}_i$ as the average service life of asset type i . It should be noted that in many empirical studies values different from 2 are found, with no clear results supporting values below or above 2.

Most advanced capital stock estimation schemes consider cohorts of assets instead of single homogeneous assets and assume retirement distributions for these cohorts (as discussed earlier). The combination of age-efficiency profiles and retirement distributions often yield “convex to origin” depreciation patterns which are alike geometric depreciation. For example, this is the case for a bell-shaped retirement distribution for a cohort of asset and a linear depreciation for a homogeneous asset. Since geometric depreciation is supported by empirical studies, the OECD recommends the use of geometric depreciation profiles. The ESA, which is the current accounting standard for European countries, recommends linear depreciation as the standard tool, but recognizes the advantages of geometric depreciation and recommends it in cases where necessary. In practice this means that both approaches are possible, but most common seems to be the linear depreciation profile. As [Table 1](#) shows, only few countries use linear depreciation patterns without retirement functions, while for many countries linear depreciation patterns in combination with mortality distributions yield quasi-geometric depreciation profiles. Apart from linear depreciation, geometric depreciation is the second most common form. Finally, there are also few countries using other depreciation patterns. Australia and Korea apply hyperbolic age-efficiency profiles which in combination with Winfrey S3 retirement functions tend to resemble geometric depreciation patterns.

Gross Fixed Capital Formation (GFCF)

Apart from service lives, retirement distributions, and depreciation profiles, investment time series is one of the key ingredients to the PIM. In the SNA, investments are defined to contain all acquisitions of new as well as secondhand capital goods minus disposals, plus major improvements of assets (in particular of buildings, dwellings, etc.) and transfer costs (e.g., fees paid to surveyors, engineers, architects, as well as taxes within the transfer of ownership). This category is also termed as GFCF. An important link between GFCF and running expenditure is the relationship between ongoing maintenance expenditure and renewals or replacement investments. The first category does not enter the calculations of capital values due to its service life of less than 1 year, but it nevertheless impacts on the capital values in so far as inadequate maintenance reduces the expected service life of the asset. As outlined earlier, the PIM is based on the assumption of adequate maintenance of all capital goods and violations of this assumption places considerable challenges to the adequate measurement of capital stocks (e.g., problems in correcting service life figures due to under-maintenance of assets). In practice, the line between ongoing maintenance spending and renewal investments is fluent and empirically not easy to draw. The relevant criterion is the durability of the expenditure: all expenses whose service lives exceed the accounting period (usually 1 year) are treated as investments.

Investment data should be broken down to technically homogeneous types of assets in order to enable assignment of service lives and retirement distributions. For example, the German capital stock calculations for transport infrastructure distinguish four types of capital goods for roads and six types for rail. Furthermore, the time series need to be deflated by price indices for the capital goods in question. This again is an empirical matter since ideally different construction and equipment price indices are required (e.g., for roads differentiated by types of pavement constructions and by bridges and tunnels, in rail differentiated by tracks, bridges and tunnels, station buildings, signaling, etc.).

Estimating an Initial Value of the Capital Stock—The Direct (Synthetic) Method

Transport infrastructure is characterized by long-living capital goods which may in many cases not only exceed 50 or 60 years but even 100 years (e.g., structures). As a consequence, long time series of investments as ingredient to the PIM may not be available. Apart from constructing the necessary time series (e.g., by econometric analysis of the relationship between GDP and investment), it is a common practice to produce a benchmark or initial value of the capital stock. This is, as far as possible, based on statistical surveys, census, or administrative property records. The basic principle of the direct method is to produce an inventory of assets in physical units (e.g., for roads in terms of indicators such as length, width, number of lanes, tunnels, bridges, etc.) and to value these units by unit costs. The direct method requires a huge expense of time, and small errors in unit costs can contribute to considerable deviations of the estimated value from the true (unknown) value. The importance of such deviations, however, will diminish over time as the base (initial) period is left behind.

The direct method is not only used for deriving starting values for capital stock estimation with the PIM. This method is also one of the feasible methods defined in the EU directive of road toll calculations ([European Commission, 2011](#)). Germany has been using

Table 2 Investment and capital stock of transport infrastructure in Germany 2017 (€ million, at prices of 2010)

	<i>Investment spending^a</i>	<i>Capital stock</i>	
		<i>Gross^{a,b}</i>	<i>Net^{a,b}</i>
Transport infrastructure	19,171	919,098	591,327
Out of these:			
Transport ways	16,925	824,632	533,759
Railways including S-Bahn ^c	4,114	151,443	95,113
Rail-based public transport ^d	566	46,835	33,035
Roads ^e	11,417	574,640	375,230
Out of these: motorways and federal roads	5,716	242,081	162,338
Inland waterways ^f	629	47,089	27,726
Pipelines ^g	199	4,625	2,655
Stations and transshipment sites	2,246	94,466	57,568
Railways including S-Bahn ^{c,h}	775	33,897	19,911
Ports at inland waterways ⁱ	125	7,214	4,106
Seaports	371	25,681	16,579
Airports ^j	975	27,674	16,972
Total transport sector	34,711	1,099,486	688,794

^aLand purchases excluded.

^bAt the end of the year.

^cNon-DB rail companies excluded.

^dTram, metro.

^eRoad administration excluded.

^fUp to the sea border.

^gPipelines for crude oil and petrochemicals with more than 40 km length and excluding gas pipelines.

^hStation buildings, other buildings, and equipment.

ⁱExcluding ports exclusively used by private parties.

^jIncluding traffic control centers and excluding regional airports.

Source: BMVI/DIW Berlin (2018).

this method as basis for deriving depreciation and interest within the toll calculations for heavy goods vehicles on motorways since 2005 (Alfen et al., 2018).

An Example of Capital Stock Estimates for Transport Infrastructure

Table 2 gives an example of annually compiled, fully-fledged capital stock estimation for transport infrastructure for Germany (BMVI and DIW, 2018). Methodologically, it is based on the PIM compatible to the official capital stock estimations of the German Statistical Bureau for all sectors, but uses instead of the Gamma function for retirements a right-skewed third-degree polynomial distribution. Depreciation is calculated with a linear function so that in combination with the retirement function a quasi-geometric depreciation profile is obtained. The level of disaggregation varies between consideration of two groups of assets up to six groups for rail and rail-based public transport. Figures for the overall transport sectors, for example, including vehicles and transport companies, are given in the last row. Transport infrastructure makes up more than half of investments and—due to the long service lives—around 85% of capital stock. The most important mode in terms of capital value is the road network followed by rail.

References

- Alfen, Aviso, IVM, 2018. Berechnung der Wegekosten für das Bundesfernstraßennetz sowie der externen Kosten nach Maßgabe der Richtlinie 1999/62/EG für die Jahre 2018–2022. Gutachten im Auftrag des Bundesministeriums für Verkehr und digitale Infrastruktur, Weimar, Leipzig, Aachen, Münster.
- BMVI, DIW, 2018. Verkehr in Zahlen 2017/2018. 47th ed. Flensburg. Available from: <https://www.bmvi.de/SharedDocs/DE/Publikationen/G/verkehr-in-zahlen-pdf-2017-2018.html>.
- European Commission, 2011. Directive 2011/76/EU of the European Parliament and of the Council of 27 September 2011 amending Directive 1999/62/EC on the charging of heavy goods vehicles for the use of certain infrastructures. OJ L 269, 1–16. Available from: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32011L0076>.
- Eurostat, 2014. EUROSTAT-OECD Survey of National Practices in Estimating Net Stocks of Structures. Luxembourg, Paris.
- Fraumeni, B., 1997. The measurement of depreciation in the U.S. National Income and Products Accounts. Survey of Current Business, July.
- Hulten, C.R., Wykoff, F.C., 1981. The measurement of economic depreciation using vintage asset prices. J. Econ. 15, 367–396.
- Jorgenson, D., 1996. Empirical studies on depreciation. Econ. Inquiry 34, 24–42.
- Koumanakos, P., Hwang, J.C., 1988. The forms and rates of economic depreciation: the Canadian experience. Paper presented at the 50th Anniversary Meeting of the Conference on Research in Income and Wealth, Washington, DC, May 12–14, 1988.
- Link, H., Dodgson, J., Maibach, M., Herry, M., 1999. The Costs of Road Infrastructure and Congestion in Europe. Physica/Springer, Heidelberg.

- Link, H., Gaus, D., Kunert, U., 2019. Inhaltliche Aktualisierung und methodische Weiterentwicklung der Anlagevermögensrechnung für den Verkehrssektor. Endbericht. Forschungsprojekt im Auftrage des Bundesministeriums für Verkehr und digitale Infrastruktur, Berlin.
- Meinen, G., Verbiest, P., De Wolf, P.P., 1998. Perpetual Inventory Method, Service Lives, Discard Patterns and Depreciation Methods. Statistics Netherlands, Heerlen-Voorburg.
- Rooijen-Horsten, M., van den Bergen, D., de Heij, R., de Haan, M., 2008. Service lives and discard patterns of capital goods in the manufacturing industry, based on direct capital stock observations, the Netherlands. Discussion Paper 08011. Statistics Netherlands, Heerlen-Voorburg.
- Schmalwasser, O., Schidlowski, M., 2006. Kapitalstockrechnung in Deutschland, Wirtschaft und Statistik, November 2006.
- Weibull, W., 1951. A statistical distribution function of wide applicability. ASME J. Appl. Mech. Paper 18, 293–297.
- Winfrey, R., 1935. Statistical analyses of industrial property retirements. Bulletin 125, Iowa Engineering Experiment Station, Iowa State College of Agriculture and Mechanic Arts Official Publication, Vol XXXIV, 28

Further Reading

- Eurostat, 2018. European System of Accounts—ESA 2010. Available from: <https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-02-13-269>.
- OECD, 2009. Measuring Capital Manual, second ed. Available from: <http://www.oecd.org/sdd/productivity-stats/43734711.pdf>.
- UN, 2018. System of National Accounts 2008—SNA 2008. Available from: <https://unstats.un.org/unsd/nationalaccount/sna2008.asp>.

The Economics of Reducing Carbon Emissions From Air and Road Transport

Olga Ivanova, PBL, The Hague, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Increasing Importance of Transport Carbon Emissions Worldwide	457
Main Economic and Behavioral Drivers of Road and Air Transport	458
Economic Instruments for Reducing Carbon Emissions From Road and Air Transport	459
Indirect Effects of Policies: Carbon Leakage and Rebound Effect	461
Conclusions and Policy Implications	462
Acknowledgment	462
Biography	463
Relevant Websites	463
References	463
Further Reading	463

Increasing Importance of Transport Carbon Emissions Worldwide

Transport emits CO₂, the most important greenhouse gas (GHG) and if global warming crosses the critical threshold of 2 degrees Celsius that will lead to even catastrophic consequences for our planet (see IPCC report of 2004). Transport activity and transport-related carbon emissions have increased significantly in the past decades. The transport sector has been responsible for 23% of total energy-related CO₂ emissions in 2010. Transport emissions have increased by the factor of about 2.5 during the period 1970–2010 and are still growing. According to the Fifth IPCC report (IPCC, 2014), transportation by road has contributed 72.6% of the total direct transport GHG emissions in 2010. International and coastal shipping and domestic and international aviation have contributed, respectively, 9.26% and 10.62% of the total direct GHG emissions of transport (Fig. 1).

ITF Transport Outlook 2019 (OECD, 2019) is an overview of recent trends and near-term prospects for the transport sector at a global level, as well as long-term projections for transport demand to 2050. The analysis covers freight (maritime, air, and surface) and passenger transport (car, rail, and air), as well as related CO₂ emissions, under different policy scenarios. ITF Transport Outlook 2019 states with some confidence that, globally, demand for mobility will continue to grow over the next 3 decades. Passenger transport will increase nearly threefold between 2015 and 2050, from 44 trillion to 122 trillion passenger-kilometers. According to the ITF scenario, China and India will generate a third of passenger travel by 2050, compared with a quarter in 2015. Aviation passenger-kilometers in India and China alone are expected to increase almost fourfold by 2050, to 21,583 billion from an estimated 5,506 billion in 2015.

In accordance with ITF model calculations, global freight demand will triple between 2015 and 2050 based on the current demand pathway. At 4.5%, airfreight is expected to have the highest compound annual growth rate of all modes through 2050, although representing a small share of total freight tons-kilometers. More than three quarters of all freight will continue to be carried by ships in 2050, more or less unchanged from 2015.

The main challenges and barriers preventing the reduction of transport GHG emissions are incomplete international agreements related to regulation of transport as well as the current high costs of clean transport technologies. The importance of emissions associated for example with international freight transportation of industrial goods is largely overlooked in the existing climate agreements.

It is possible to bring down the (relative) costs of clean transport technologies by using the combination of taxes and subsidies on specific fuels and types of transport technology with long-term investments into transport-related R&D activities. However, this (see IPCC report of 2015) will not be sufficient for the needed drastic reduction of the amount of transport-related GHG emissions. Economic policy instruments should be supplemented by command and control actions such as closing the coal mines and coal- or gas-based electricity plants (Fig. 2).

According to the projections of ITF Transport Outlook 2019 (OECD, 2019), transport CO₂ emissions will globally remain a major challenge. The extrapolation of current policy ambitions into the future shows that these will fail to mitigate increases in transport CO₂ emissions due to the strong growth in transport demand over the coming years (especially related to the international freight transport). In an ITF baseline scenario where both current and announced mitigation policies are implemented, worldwide transport CO₂ emissions are projected to grow by 60% by 2050. The growth in transport emissions is driven mainly by increased demand for freight and nonurban passenger transport, both of which are projected to grow 225% by 2050. Emissions from urban passenger transport, on the other hand, are projected to fall in the baseline scenario by 19%, reflecting existing strong focus of current policies on urban transport.

Both freight and passenger transportation by road and air contribute significantly to the generation of GH emissions. The share of road transport is the largest for both passenger and freight transport with the use of gasoline and diesel as two main types of fuels. The

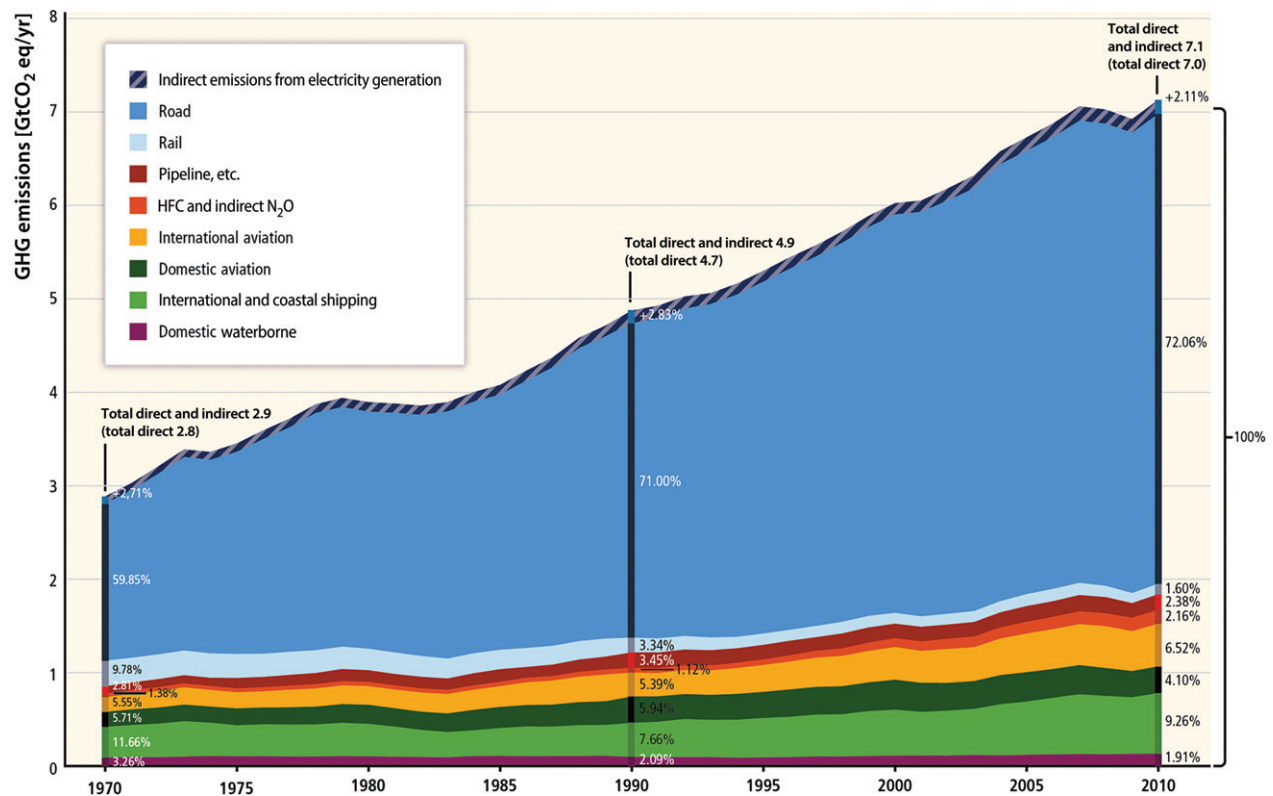


Figure 1 Development of the direct GHG emissions of the global transport sector. Source: Fifth IPCC report.

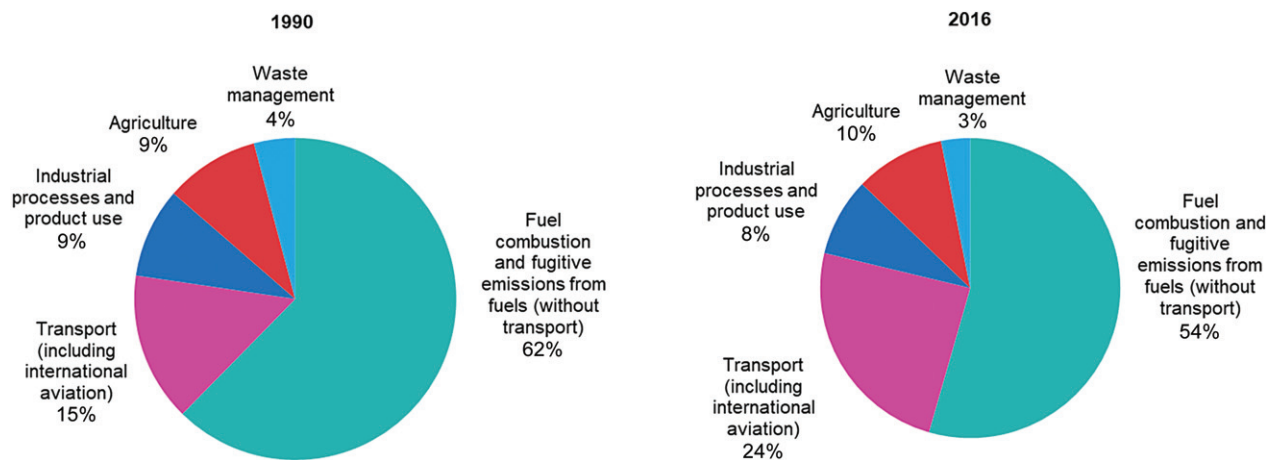


Figure 2 Share of transport as a part of total direct GH emissions in EU28. Source: Eurostat.

use of elective vehicles still remains marginal in most of the countries and globally. Rail being less emitting mode of transport is used marginally both for passenger and for freight transportation (Fig. 3).

Main Economic and Behavioral Drivers of Road and Air Transport

Transportation is an integral part of our daily lives and is a prerequisite for economic growth and prosperity. Various studies have demonstrated that good transport infrastructure and high personal accessibility lead to the economic growth and households' income by improving the access to labor market and allowing for better matching between available vacancies and persons searching for new jobs (Santos et al., 2010). Significant decrease in international freight transport costs has contributed to the ongoing process

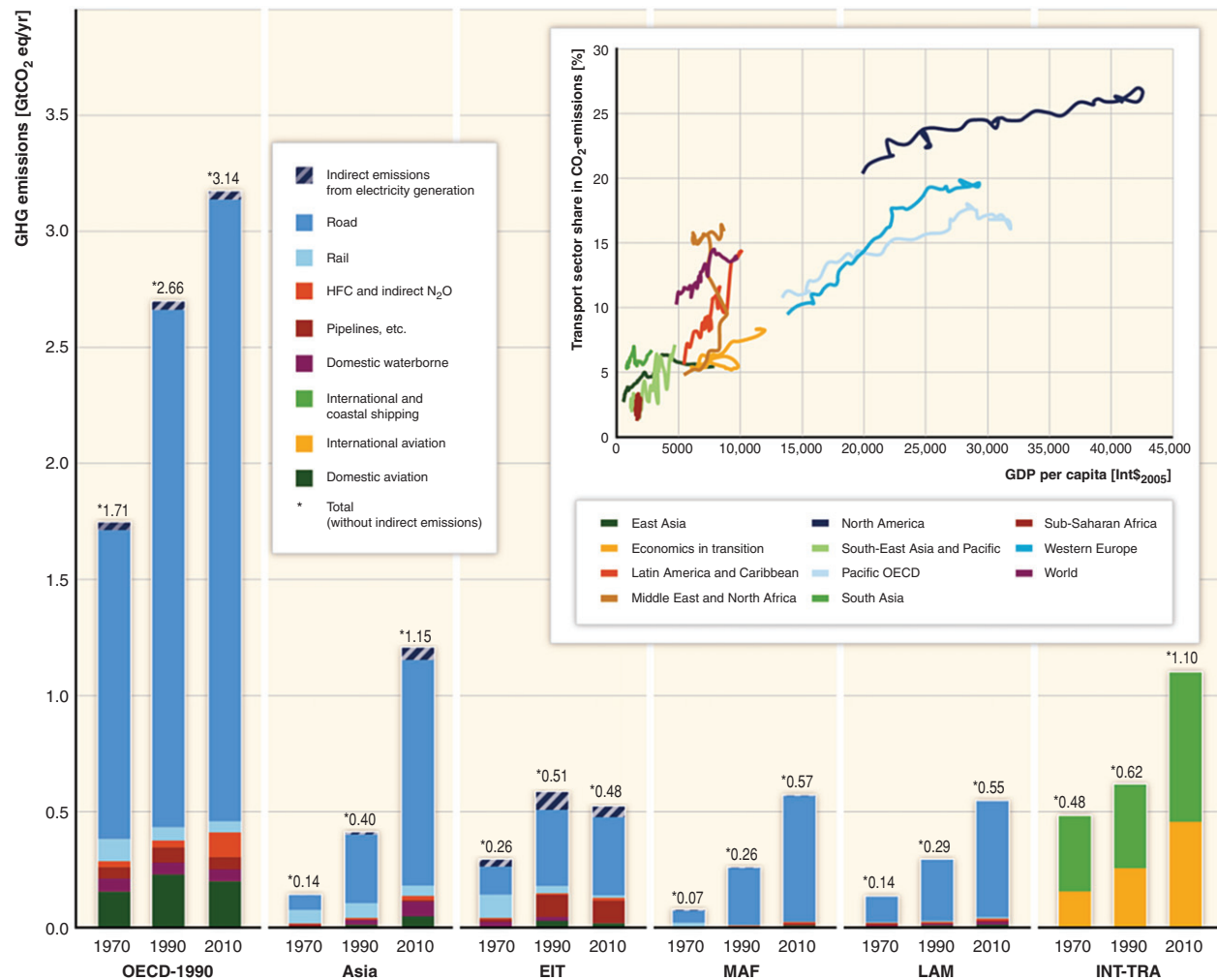


Figure 4 Regional distribution of transport GHG emissions. Source: Fifth IPCC report.

instruments because they allow reduction in emissions in the most costs effective way for both consumers and producers. Price instruments also give producers and consumers the freedom to determine themselves how they would like to reduce their emissions; this could be via investing in more fuel-efficient vehicles, traveling less, or adapting their driving style. Gasoline and diesel are heavily taxed both in EU and in Japan where the taxes are often more than 200% of the fuel price. Fuel taxes function both as a GHG emissions tax and the tax on the use of transport infrastructure, and the tax revenues from fuel taxes are often earmarked for transport infrastructure investments.

With regard to road transport, the most widely used instrument to reduce emissions is the combination of vehicle registration taxes that are related both to car's environmental performance and to its characteristics such as engine power and weight with fuel taxes. The first type of taxation is intended to discourage the purchase of more polluting cars and the second one is intended to discourage intensive driving with the car.

It is important to realize that about 50% of all new passage car sales in EU are company cars, which are often used by employees for both business and personal travel. Companies pay for both cars and fuel used by their employees and can subtract those costs from their taxable revenues. Besides this, it is often that company cars are subject to preferential taxation that results in the loss of 54 billion of Euros per year in EU according to several studies (Gossling and Cohen, 2014). Such tax arrangements provide a clear incentive for employees to by larger and heavier cars that use more fuel as well as to drive them more. It is also true that most of the super mobile individuals that drive a lot have such company cars and hence no actual incentive to change their travel behavior.

In the recent years, various countries have introduced subsidies on purchase of electric cars in the hope to reach reduction of road transport emissions. Due to the relatively high prices of electric vehicles most of these cars are currently also company cars. It is hoped both by policy-makers and by the researchers that gradual reduction in the prices of electric vehicles will make them attractive for various income groups and that current electric company cars will soon become available at the secondhand car market for more reasonable prices.

Overall price-based measures related to road and air transport do not consider individual contribution to transport emissions and also do not take into account income inequalities that affect mobility behavior that could be of importance to efficient reduction of transport-related emissions.

The environmental policy instrument that is explicitly designed to regulate GHG emissions is a tradable emissions scheme. The main advantage of this policy instrument is that the costs of emissions reduction costs between the polluters are distributed in the most efficient way and hence the GHG emissions reduction is achieved at the lowest costs. From the point of view of the environmental economic theory the introduction of ETS system should result in reduction of emissions in the most efficient and cheap way because the companies who can easily and cheaply reduce their emissions will do that and sell their remaining emissions permits to the companies that cannot reduce their emissions that are easy and cheap.

European Trade System (EU ETS) is one of the examples of the successful implementation of a tradable emissions scheme. The main idea behind EU ETS is to have a fixed amount of emissions permits that are distributed to various companies according to their historical emissions and to allow them to trade these emissions permits between each other such that the price of one permit is determined by the interplay between demand and supply of emissions permits. The fixed number of emissions permits should decrease over time that would result in the decrease of overall emissions and an increase in the price of emissions permits.

Despite the solid theoretical foundations of EU ETS, its efficiency has been undermined by too low prices of the emissions permits that did not provide airlines with sufficient incentives to invest in reducing their emissions or to increase their prices.

CO₂ emissions from aviation have been included in the EU ETS since 2012. Under the EU ETS, all airlines operating in Europe, European and non-European alike, are required to monitor, report, and verify their emissions, and to surrender allowances against those emissions. They receive tradable allowances covering a certain level of emissions from their flights per year. It is well possible to also include road transport into the road transport by requiring the permits upstream at the level of fuel distributors. EU ETS system for road transport could potentially replace the existing fuel taxes if the price of carbon is set up on the correct level.

In addition to market-based measures like the ETS and taxes, regulatory measures—such as efficiency standards, alternative fuels, and modal shift to rail and public transport—also contribute to reducing road and aviation emissions.

The use of efficiency standards next to fuel taxes is justified by the energy efficiency gap which means that consumers and producers might be to some degree myopic and they may underestimate the energy savings associated with more fuel efficient cars. The degree of the effectiveness of fuel standards depends on the degree of the myopic behavior among the consumers and producers. An alternative to fuel standards is the co-called “feebates” that represent a combination of fees on the producers of vehicles that fall below the efficiency standard with subsidies to the producers whose vehicles perform better than the efficiency standard. Feebates programs are usually designed to be revenue neutral meaning that the collected fees are equal to the paid subsidies.

Another reason to use efficiency standards is the process of technology diffusion. Introduction of efficiency standard in one large country or a group of countries will result in changes in the production technology of the vehicles and the enhanced more fuel efficient vehicles will be sold all over the world leading to reduction in GHG emissions in other parts of the world as well (in contrast to the implementation of fuel taxes).

In case of passenger road and air transport stimulating the shift from road to public transport and from air to rail transport is often seen as one of the policy options to reduce GHG emissions. High occupancy rates of public transport and rail in some countries as well as the use of alternative clean energy types make them attractive alternatives for road and short-distance air if one wants to reduce GHG emissions. The modal switch is quite a costly public policy that requires large investments into public transport infrastructure and these investments in most of the cases cannot be fully recovered via the prices of tickets. Moreover, building new railways infrastructure is in itself associated with high GHG emissions. Ticket costs of public transport and rail need to be kept low in order to promote switch from road and air transport.

Indirect Effects of Policies: Carbon Leakage and Rebound Effect

Reduction of the transport-related GHG emissions in any particular country has only small impact on the overall level of GHG emissions worldwide. The reduction of GHG emissions as a result of reduction of the fossil fuels by one particular country may be partially or even completely offset by the so-called “carbon leakages” and “rebound effects.” These two mechanisms are related to economic behavior of producers and consumers and the global nature of our economies.

Carbon leakage comes in two important types: spatial leakage and intertemporal leakage (Eliasson and Proost, 2015). Spatial leakage occurs when one country puts in place the regulation of GHG emissions and the other countries do not do that. Introduction of more stringent GHG emissions regulation may result in relocation of polluting activities to the countries with less stringent regulation which lead to less reduction in global GHG emissions as without relocation of economic activities or even their overall increase. Relocation of activities might not be as large problem for transportation sector as it is for the industrial sector but a sufficient decrease in the demand for fossil fuels may result in a decrease in their global prices and increase consumption of fossil fuels in the rest of the world. According to the empirical studies, spatial leakage can offset initial reduction of GHG emissions by 10%–30%.

Intertemporal leakage occurs when all countries in the world put in place more stringent regulation of GHG emissions. In case when the price of oil stays in the future period higher than the costs of oil extraction, the companies that produce oil will

be interested in selling their full oil reserves. This means that the introduction of more stringent regulation will result in a temporal shift in the consumption of fossil fuels without changing the total amount of their consumption and associated emissions. However, fast reduction of GHG emissions may prevent further speedup of the climate change process. Only in a particular case when there will be a new technology that is alternative to fossil fuels and that is also cheaper than the costs of oil extraction will the world have an economic incentive to switch from fossil fuels to the clean alternative. The crucial determinant of the consumption of fossil fuels and the associated GHG emissions over time is the relationship between the costs of alternative clean technology and the costs of the oil extraction. Worldwide implementation of alternative clean technology will become economically profitable only when its costs are lower than the costs of extracting oil and not when there are lower than the price of oil.

Policy and technological measures that improve fuel efficiency of road and air transport are typically offset by rebound effects that could be direct, indirect, and economy-wide. These rebound effects differ between various household groups with larger rebound effects being associated with poorer households. Rebound effects are related to the behavior of households when choosing their consumption patterns when they have more income to spend due to fuel savings while driving their car or cheaper airplane tickets.

Direct rebound effect corresponds to an observation that when driving more fuel efficient car persons will start driving more as compared to their use of less fuel efficient car because the use of car per kilometer driven is now cheaper. Indirect rebound effect describes another way in which extra money due to fuel savings can be spend by the households. Instead of driving a bit more they could choose to spend their extra money on other types of consumption such as for example food or clothes. These extra consumption in itself is related to energy use and emissions that have been used during its production process lead to an offset of an initial fuel and GHG emissions savings associated with improved fuel efficiency of vehicles. Economy-wide rebound effects describe the impacts along the whole production chain of the additional goods and services that the households have purchased with the money saved due to improved fuel efficiency and capture the notion of the economic multiplier effect. Extra production of goods and services for the purchase by households generates extra income that is spend in the economy and this extra income itself results in even larger increase in production of various goods and services and hence even higher offset of initial fuel and GHG emissions savings. The magnitude of the rebound effect ranges from 10% to 30% of the initial fuel savings.

Conclusions and Policy Implications

Taxes and changes on purchase of vehicles and fuel use are used extensively throughout the world. Most countries levy a registration tax on new vehicles and annual vehicle excise duties. In the recent years, several countries have also introduced subsidies to fuel efficient and electric vehicles, feebates, and scrappage incentives. Most countries of the world also have fuel taxes that vary substantially between them. Mexico and the United States have very low fuel taxes as compared to EU and Japan. In many of the EU countries fuel taxes represent a significant part of the governmental tax revenues (between 4% and 6%). If the transport sector would not use carbon fuels anymore, the governments would need to introduce some additional alternative taxes in order to compensate for the loss of revenues from fuel taxes.

It is important to realize that any assessment of costs and benefits of policies for reduction of road and air GHG emissions should take into account that this emissions reduction will be at least partly offset by spatial and temporal leakages and rebound effects. One should also notice that the worldwide oil reserves will be used up to the point when the costs of clean alternative are lower than the costs of oil extraction. However, reducing current demand of road and air transport for fossil fuels might be effective in postponing oil extraction and the associated GHG emissions to the point when new clean technology will become significantly cheaper.

Replacing fossil-fuel-driven cars with electric ones will only lead to reductions of GHG emissions when the used electricity is produced using renewable energy sources. If the electricity is produced using coal and gas, the switch to electric cars may result in increase of overall GHG emissions.

The rebound effect is important for determining the level of efficient policy instruments such as standards and taxes. The policy-makers should pay sufficient attention to the heterogeneity of the rebound effect. Research shows that household characteristics such as income and age as well as driving intensity lead to significant heterogeneity in the rebound effect and make uniform policy ineffective in diminishing it.

Governments can currently choose from a pallet of different policy instruments aimed at reducing GHG emissions of road and air transport in order to address the problem of climate change. The actual implementation depends on the will and commitments of the governments.

Acknowledgment

I would like to thank the editors for their useful comments on the content of the chapter.

Biography



Dr. Olga Ivanova holds a PhD degree in Applied Economics and a Master of Science degree in Environmental and Development Economics from the University of Oslo. At present, she works as a senior researcher at PBL Netherlands Environmental Assessment Agency (<http://www.pbl.nl>). She is the coordinator of Horizon 2020 project MONROE (<http://monroeproject.eu/>) that develops various macroeconomic models for the assessment of the impacts of R&I. She is also involved in Horizon 2020 project Clair-City (<http://www.claircity.eu/>) with the micro-simulation modeling for European cities. At PBL she is responsible for the development of modular Spatial Computable General Equilibrium System EU-EMS that combines regional and global geographical dimensions. In the past, she has coordinated the project on the construction of RHOMOLO regional-economic model of EU28 for DG REGIO (<http://rhomolo.jrc.ec.europa.eu/>).

Relevant Websites

International Transport Forum. Available from: <https://www.itf-oecd.org/>.

European Environmental Agency. Available from: <https://www.eea.europa.eu/>.

DG Mobility and Transport. Available from: https://ec.europa.eu/transport/home_en.

Association for European Transport. Available from: <https://aetransport.org/>.

References

- Eliasson, J., Proost, S., 2015. Is sustainable policy sustainable? *Transp. Policy* 37, 92–100.
- Gossling, S., Cohen, S., 2014. Why sustainable transport policies will fail: EU climate policy in the light of transport taboos. *J. Transp. Geogr.* 39, 197–207.
- IPCC, 2014. Fifth IPCC assessment report. Chapter on Transport. Cambridge University Press, Cambridge.
- OECD, 2019. *ITF Transport Outlook*. OECD, Paris.
- Santos, G., et al., 2010. Part I: externalities and economic policies in road transport. *Res. Transp. Econ.* 28, 2–45.

Further Reading

- Button, K., 1993. *Transport the Environment and Economic Policy*. Edward Elgar Publishing Ltd., Aldershot.
- Eftestol-Wilhelmsson, E., Sankari, S., Bask, A. (Eds.), 2019. *Sustainable and Efficient Transport: Incentives for Promoting a Green Transport Market*. Edward Elgar Publishing Ltd., UK.
- Grigolon, L., Reynaert, M., Verboven, F., 2018. Consumer valuation of fuel costs and tax policy: evidence from the European car market. *Am. Econ. J. Econ. Policy* 10, 193–225.
- Hensher, D.A., Button, K.J. (Eds.), 2003. *Handbook of Transport and the Environment (Handbooks in Transport)*. Elsevier, Amsterdam/Boston, MA.
- Midttun, A. (Ed.), Witoszek, N., 2017. *Energy and Transport in Green Transition (Routledge Studies in Sustainability)*. Routledge, New York.
- Proost, S., Van Dender, K., 2012. Energy and environmental challenges in the transport sector. *Econ. Transp.* 1, 77–87.
- Ricardo-AEA/R, 2011. Update of the handbook on external costs of transport. Report for the European Commission: DG MOVE. Ricardo-AEA, London.

Regulation and Financing of Toll Roads

Marco Ponti, Bridges Research Trust (Scientific Responsible), Milano, Italy

© 2021 Elsevier Ltd. All rights reserved.

Toll Highways: A “Natural Monopoly” with a Nonnatural Financing System	464
Efficiency, Budget Constraints and Distributive Issues	464
Regulatory Principles: Simulating Competition Pressure, Rewards and Constraints in Order to Improve Efficiency and Protect the Users, Lacking the Real Incentives of the Market	465
More on the Financing Issue: the Problem of “Asymmetry of Information”.	
The Basic “price-cap” Formula and its Foundations	465
The Traffic Risk: A Wrong Assumption in Risk Assignment	466
More Technical Issues: How Investments are Treated, the “Fair” Admitted Profit, the “Regulatory Asset Base” (RAB), Quality and Maintenance	466
Issues Concerning the Risk of the Capture of the Regulator	466
Some Selected Cases: USA, Germany, an Uncommon Spanish Decision, the Italian Autostrade Case, as an Example of Regulatory Failure	467
An Alternative, Simpler, and Less “Capture-Prone” Model, and its Financing, given also the Low Technical Content of the Road Sector, and the Need of a Consistent Planning Approach for the Entire Network	468
Further Reading	468

Toll Highways: A “Natural Monopoly” with a Nonnatural Financing System

Efficiency, Budget Constraints and Distributive Issues

Natural monopolies, differently from “pure” public goods, have no rivalry in consumption but are “excludible”, that is, the users have to pay a tariff “Pure” public goods are neither rival (no one is damaged by an additional user), nor excludible (no one can impede its use). The lighthouse is the classic example of a pure public good. Natural monopolies lack the excludability condition: one has to pay for using them). But in case of roads, to impose a tariff (named “toll”) is a pure political choice: actually, the majority of roads are without tolls, that is, are “pure” public goods. (If a road is congested, actually rivalry of consumption ensues, in the sense that each added user reduces the quality of the service provided by that road, but we will see this issue in the final point).

Let’s see first the economic rationale of imposing a toll. In theory, the maximum efficiency in the use of infrastructure is reached when the users pay for the costs they generate to the infrastructure itself, that is, wear and tear (this rationale of tolling is known as “Marginal Cost Pricing”, MCP, and obviously implies that the investment costs are paid by the State, that is, the taxpayers, and also this issue presents some problem, as we will see).

Why not making the users pay also for the investment cost? Because some users, if they have to pay also for this cost, will non travel, even if their utility in travelling is actually larger than the cost that they generate travelling, that is only, as said, wear and tear. If their benefits are higher than the costs they generate, it is efficient that they travel.

The environment and safety cost are also to be paid for (see in particular the well-known principle “polluters pay”), but not for the use of a specific infrastructure: it is simpler and more sensible to charge directly these costs where they originate, that is, with the taxation of fuels and with the insurance system.

This, seen from the basic social welfare theory (including the environmental aspects). But in general road tolls include totally or partially the investment costs, on top of the wear and tear ones. What is the motivation of this apparently inefficient policy (known as the “Average Cost Pricing”, ACP, because it tends to cover the entire cost of the infrastructure)?

In theory, it depends on the scarcity of public money. This scarcity in turn can be translated in a social opportunity cost for this resource (MOCPE, Marginal Opportunity Cost of Public Funds). A country with a balanced budget is evidently in a different position of a one heavily in debt. Therefore, a trade-off is set between the efficiency of the use of the infrastructure and the scarcity of public funds. (The exact value of this trade-off in turn depends on the elasticity of traffic demand to tolls and the magnitude of the MOCPE. Formally, it is a problem of constrained optimization. Anyhow, the higher the elasticity and the lower the MOCPE, the lower is the “optimum” toll level).

A distributive issue may also come in play: if only the users pay for an infrastructure, the taxpayers that will never or seldom use it, will not. Those who benefit will pay for their benefits.

But here a severe political problem arises, as we will see later: why quite often in Europe the users of some type infrastructure are charged at ACP (i.e., pay a lot), and others, using different transport modes, are charged at MCP (i.e., pay little)? The larger difference is between toll road and rail users, even those of HST. Political consensus (and not the environment) is often at play, but we cannot extend here the issue in more details.

Regulatory Principles: Simulating Competition Pressure, Rewards and Constraints in Order to Improve Efficiency and Protect the Users, Lacking the Real Incentives of the Market

As any monopoly, also this one requires public, independent regulation. The market incentives and pressure take care, for private goods, of incentivizing efficiency (inefficient firms will show either higher costs or inferior quality of production, and loose out against competitors), and protect consumers from excessive prices (competition tends to minimize prices). A basic question emerges here: if the toll roads are kept public as the nontoll ones are, the State will take care itself both of efficiency and of fair tolls. And here enters the issue of “capture” of the political decision makers. Efficiency tends to be in conflict with consensus, that is an explicit and legitimate political objective: no favors can be made to employees nor to suppliers, quite the contrary. Extracting rents from the users with excessive tolls is also a strong temptation, that furthermore becomes necessary if the management is inefficient. “Capture” is the synthetic term that describes a situation where the public interest succumbs to socially non acceptable goals (even plain corruption is encompassed by this term).

Therefore, it is a worldwide tendency to institute independent regulatory bodies (Ponti, 2011), endowed with technical capabilities, and with severe constraints on the behavior of the managers, that for example cannot work for many years after their appointment for former regulated companies. These bodies are in charge of periodically setting the tolls and tendering out the concessions when they expire. The controlling of the maintenance and construction activities, and the quality of the services provided by the concessionaires are generally delegated to the Ministry of Transport, since the needed technical capabilities are already there, due to the existence of a much wider network of nontoll roads.

All these issues are even more compelling since in general the toll roads are tendered out to private companies. This happens often under the assumption that private companies, especially under the pressure of competitive tendering, are more efficient than the State. This “assumption of efficiency” is frequently extended not only to the building and maintaining activities, but also to traffic managing, even if, as we will see, that this is a far less defensible assumption.

Furthermore, it is also assumed the concessions have to be very extended in time, in order for the concessionaires to have the possibility to recover, via the tolls, the investment costs, given the fact that in general these costs are included in the pricing regime (ACP).

This last assumption, as we will see, is also very questionable.

More on the Financing Issue: the Problem of “Asymmetry of Information”. The Basic “price-cap” Formula and its Foundations

So far, so good: the users have to pay, under independent public regulation, fair tolls, covering both efficient construction costs (partial or total), and efficient maintenance costs, related to the infrastructure that they are using. But the regulator, that is in charge of setting these fair tolls, ignores what are “efficient” costs, that can vary due to several different and specific causes. The concessionaire, at the contrary, knows them well (this fact is known as “information rent”). Furthermore, it has an objective, as every firm has, of maximizing its profit. And a reasonable profit has to be included in the “fair toll”, in order to induce any firm to take on whatsoever activity. It follows that the concessionaire, even if perfectly honest, has an implicit incentive to maximize its costs, being the “admitted” profit a fix percentage of these costs.

The solution for this problem is based on changing the incentive of the concessionaire. The more common strategy is to induce the firm to become efficient, setting the costs in advance after a negotiation: if it is able to save against these guaranteed revenues, it will increase its profit, and in a following round of toll setting, the regulator can have better information on the efficient costs, on which to base the level of the tolls.

The dominant practical instrument set for getting this result is slightly more complex, and it is known as the “Price cap” formula (Coco, 2008), that has to be set for a definite period constant in length (generally 5 years, and this length is known as “regulatory lag”), and then re-negotiated for each following period:

$$T + 1 = T(\text{CPI} - X + Q)$$

where T is the initial toll, based on the existing information on the annualized costs of the concessionaire including the admitted level of profit. $T + 1$ is the toll of the following year.

CPI is the Consumer Price Index (inflation has to be paid, since it is an “external” factor of cost increase). X is the annual (negotiated) increase of efficiency expected, and for this reason with a minus sign. Q is a (negotiated) quality-related factor (in toll roads, may be asphalt quality, lighting, toll-collection speed etc.).

The formula incentivates the concessionaire to become more efficient: if its costs decrease more than the negotiated X factor, its profit (that is the difference between its costs and the toll-related revenues) will increase in proportion.

In each period (“regulatory lag”) the regulator increases its information, via the negotiation itself, via some benchmark activity, and even directly observing the activities of the concessionaire during the period.

In the case that the concessionaire is so efficient to lower its cost more than is set for each period by the X factor, it will make extra-profits (a deserved prize for efficiency, as in a market context), but non forever: at the end of each regulatory lag the regulator will set

the toll level allowing for a fair level of profit, not more (This is known as the “claw back” principle, and is consistent with what happens in a market context).

The Traffic Risk: A Wrong Assumption in Risk Assignment

A basic principle of regulatory theory states that the risks left to the concessionaire have to be limited to the ones that it is able to control and manage, and therefore deserve a compensation in proportion of its results. Costs are for sure the more evident risk that it is correct to leave to it. But a firm operating in the market has to face also demand risks. It may change due to competition, the quality and the pricing strategy for its own products, the effectiveness of advertising etc. If a firm is successful, this component is at least as important as cost management. In the price-cap formula it is stated only the level of tolls, and therefore it is implicitly assumed that the higher the traffic, the higher the revenues (and vice-versa), and being the costs mainly constant with the traffic, so it goes for the total profits (or losses).

But this for toll roads has no economic sense: traffic is basically outside the control of the concessionaire. It depends mainly from the overall economic growth (GDP), demography and land use, and alternative infrastructure and transport services and prices.

And if a variable that is outside the control of an economic agent is left to it to be taken into account, it will correctly charge a high “risk price” related to the possible variations of the noncontrolled variable. This will result in a level of toll requested by the concessionaire to the regulator much higher than the efficient one, therefore generating an underutilization of the infrastructure.

It follows that the gains or the losses related to traffic variation have to be left outside the risk boundaries of the concessionaire. In reality often it is not so, due to several factors, mainly related to an historical picture of a rather constant and foreseeable traffic increase, but also to some form of “capture” of the regulator by the concessionaires.

Including in the toll a “risk price” rises its level, and the political decision makers may well have a “hidden agenda” favorable to high tolls, and if the regulator is not independent, may yield to political pressures. The above-mentioned “hidden agenda” is related to the tax revenues: the higher the profit of the concessionaire, the higher the income for the state, that for example in Europe has a level of taxation on profits in average well above the 30% threshold.

More Technical Issues: How Investments are Treated, the “Fair” Admitted Profit, the “Regulatory Asset Base” (RAB), Quality and Maintenance

Up to now, “costs” have been treated as a single concept, just mentioning the subdivision in investments and maintenance/operating costs. But in the case of infrastructure, they are basically different. The decision to build or even enlarge an infrastructure is basically a public one, and in case of roads this is made even more evident by the fact that the majority of roads is totally public, and without tolls. So what is the correct role of the actors involved, in this case? The public decision maker has to plan for a new road (hopefully based on some cost-benefit analysis) and decide if it deserves a partial public funding, and the concessionaire has to build it at minimum cost, repaying it with the corresponding tolls within a planned number of years. The possible problems here are two-faced: if the admitted profit is high, the concessionaire may “incentivate” the decision maker to build more than the necessary (a form of capture), or, in the worst case, delay the construction in order to get undue benefits from the tolls, if he is able to keep them unvaried notwithstanding the delay.

The above-mentioned admitted profit is not supposed to be arbitrary, of course.

In theory it has to be based on two components: the cost of capital (technically: the Weighted Average Capital Cost, WACC) plus an already mentioned risk premium. But in practice here there is wide space for “capture”, if the regulator is not independent, as we will show later in some example.

Another aspect concerns the scope of regulation, that is which activity of the concessionaire has to be regulated (here the technical term is Regulatory Asset Base, RAB): of course its core business, that is the road (construction, maintenance etc.). But there are important auxiliary activities that can be left to the market for their efficiency, mainly the resting areas and the connected gas stations. These activities can be auctioned periodically with an independent procedure.

Quality and maintenance are still a different issue: quality may enter in the price-cap formula, as we have seen, but its control, as for the far more crucial maintenance activities, are generally delegated, as said, to the Ministry of Transport, or to a similar technical body.

Issues Concerning the Risk of the Capture of the Regulator

Let's return on a basic question: why an independent regulator is considered in general necessary? The Ministry of Transport for sure can have all the technical resources needed to perform any required activity, both of regulation and control. The only justification is the widespread experience of “capture” of politically elected actors: they have a fully justified objective of pursuing political consensus, but this is exactly the contrary of what is needed for efficiency. Consensus is far better achieved by spending that by saving public money.

There are a set of issues that are related with the risk of capture (a risk that concerns to some extent also the independent regulator, but it is impossible here to elaborate more on this).

Let's start with the "correct" duration of the concession. The common wisdom is that its length has to permit to the concessionaire to fully amortize its investment, in order to be incentivated to build with solid technical standards, minimizing the overall maintenance costs. But there is a correspondent risk of capture linked with excessive "friendship" that may ensue, due to several years of collaboration with the technical staff of the regulator. Is it possible to define instead efficient contracts setting fair conditions of "succession", as the one resulting from a "new entrant" winning the tendering of the concession after a shorter period of time.

Related to this, is the basic issue of the assignment system. In theory, if it is made via a competitive tendering, with an extremely detailed contract, this will reduce the need of a too frequent periodic intervention of the regulator, like the price-cap formula assumes (the "regulatory lag"). But in practice the ever-changing conditions of costs, traffic, investment decisions, technologies etc. make a periodic "adjustment" indispensable.

A capture issues exists here in terms of concentration of the property of the concessionaire: in theory again, if regulation is effective and efficient, it is irrelevant the level of concentration of the property, even to the level of having a single concessionaire in each country, winning all the tenders.

But in practice, as we will see in the next point, the political clout of large or dominant firms can become such as to render very difficult or even impossible an effective regulation (we may say "too big to be regulated", in a perfect logical symmetry with the better known concept of "too big to fail"). This issue, of the risk of excessive political clout, is traduced in the principle of "minimum efficient dimension". It means that if there are economies of scale (per se a good thing in terms of costs), they have to be balanced out in terms of risks of excessive political clout and of reduced contestability of the incumbents when the concessions are tendered out when they expire. (For the Italian network, a recent econometric analysis has found a minimum efficient dimension for toll road concessions to be in the order of 300 km. The dominant concessionaire (Autostrade per l'Italia) is ten times this size. Nevertheless, no action of unbundling the dominant concessionaire ensued, being it protected by a very long lasting contract.)

Seen from the point of view of the concessionaire, there is also a symmetrical risk: a regulator too strong, and therefore able to impose rules and costs in an arbitrary and unfair way. This is known as "regulatory risk", and is something quite real, since it is mainly connected with the variability of the governments, that can well translate in variations of ideology, priorities etc.. It requires national (and sometime international) legislation, guaranteeing full respect of the signed contracts.

In turn, there is an implicit incentive for keeping this risk low, given the fact that if it is perceived by the firms involved to be high, this is correctly translated in higher requested "risk premiums" (see the "admitted level of profit").

Another risk, that concerns the political balance of the regulatory activity, may be called "the winner curse reversed" on the conceding agency. Let's make a practical example: a new, important project of a toll highway in general is charged with heavy political expectations, far over and above its technical rationale (employment, development of an isolated region, etc.).

The concessionaire in charge of its construction is perfectly aware of this, and can rather easily use this knowledge to twist the regulatory conditions in its favor, only with some hints of possible severe delays, or even worse, of a technical impossibility to deliver the project at all. In this context, it has in fact a much heightened political clout.

Some Selected Cases: USA, Germany, an Uncommon Spanish Decision, the Italian Autostrade Case, as an Example of Regulatory Failure

The nation-wide interstate American highway system is without tolls, and it is paid by public money plus by a percentage of the gasoline tax revenue. But being now the network 60 years old, these revenues have become insufficient to guarantee a steady and effective level of maintenance. This paradoxical situation is linked to the extreme reluctance of the American policy makers, at all administrative levels, to tax the car users, that are practically the entire populace, for consensus reasons (this reluctance is extended even to the environment-related taxes, negligible in the USA).

In Europe, Germany is a similar case, but now a very effective satellite-based toll collection exists for heavy trucks. It is aimed to reduce congestion, that is, to allocate traffic in an efficient way, with also a net return for the public purse (the fact that the environment has not been included in the German system is explained by the fact that this issue is fully covered by gasoline taxes, as in the rest of Europe).

Spain has also an extended network of toll roads: it deserves to be quoted here because its regulator has recently decided that an important southern stretch of the network, resulting now fully amortized by the users, has to be set free of any toll, returning within the general free road network that is planned and financed directly by the State.

Let's see now a negative example (very familiar to the author, since he has been directly involved in it): the Italian case (Ragazzi, 2008). At the beginning of the present century, the state decided to privatize the toll network, that was in charge of a public conglomerate, IRI, dissolved by the new European rules on competition and State ownership of industrial assets.

Two severe problems became immediately evident: the first is that the low technological content of the sector was not suggesting any high value-added by the private know-how implicit in any privatization. The second was linked with an equity issue: since the users at that moment had already paid for a high share of the investments, perpetuating a toll system here appeared more an arbitrary taxation than a reasonable policy.

But the worse came with the tendering procedure: nothing similar to a real competition happened, the concession was awarded to the Benetton textile company, that was for sure far away of any technical capability in road management, but very close politically with the governing coalition of that period.

The concession that was given to the Benetton group had both a very long duration, a very high admitted level of profit (WACC), and was based on a “lump sum” regulatory approach, that is, an immediate down payment to the State of 3.5 billion € (about 4 at current prices).

The traffic risk was left to the concessionaire. Several scholars deemed the contract exceptionally favorable, also because the real controlling procedures on the agreed-upon investments and maintenance activities appeared immediately rather ill-defined and weak. (Actually, these scholars had to work on incomplete documents, because the details of the contract were classified, due to a peculiar legal quibble).

The concessionaire was able to repay the down-payment in a very short time, and to reap a handsome level of profit thereafter. The unchecked maintenance generated a series of incidents, culminating in the collapse of a bridge in Genoa with 43 fatalities. The State attempted immediately to cancel the concession, but discovered a clause in the contract that guarantees the payment of the entire expected profit till the expiration of the concession, even in case of severe misbehavior on the side of the concessionaire. A real “pactum sceleris”, that obviously rises the suspect of an extreme form of “capture”, reaching probably direct or indirect corruption. The legal and political dispute now is raging, at the time of writing, and its final results high uncertain.

This example summarizes well a large set of mistakes and capture problems that may well help to suggest a radical and more simple approach to toll road regulation and financing.

An Alternative, Simpler, and Less “Capture-Prone” Model, and its Financing, given also the Low Technical Content of the Road Sector, and the Need of a Consistent Planning Approach for the Entire Network

The main emerging question of toll road regulation and financing must be: it is efficient to put tolls on some roads (given the fact that the majority of roads are not tolled)? That is, to make the users pay for some roads, and the public purse or fuel taxation for the rest of the network? No rationale appears evident here: the collection systems are per se complex (much more than fuel taxation), and even generating traffic slowing down and congestion, if not operated with the latest “free flow” or satellite-based models. Charging strategies are efficient only if limited to marginal costs, that is, wear and tear, and not in general for paying investment costs. The road network is a highly connected and inter-dependent system, that requires a consistent public planning approach. Long-distance traffic is no longer dominant, and the major problems of congestion and pollution are related with short-distance and regional traffic, now even in less-developed countries. Long concessions are prone to generate capture problems in different forms, even in the information sphere. The sector has a modest technical content, and a normally efficient public administration is capable to manage all its main aspects, as proved by the nontoll network.

This said, competition remains an essential ingredient of an effective management of the sector (there is no proof of effective and performing competition taking place in the present concessionary regimes).

The solution seems simple: abolish cost-recovering tolls entirely, re-unify the planning process within the proper administrative levels (regional if possible: both environmental and congestion problems are located mainly here, as well as the related information), and use tolls only in order to internalize the important external cost that is not covered by the gasoline taxation, that is congestion. This nowadays can be made via cheap, flexible, and efficient satellite systems, even “dialoguing” directly with the users. Keep competition pressure on construction activities (as for the nontoll network), and on maintenance by medium-term competitive tendering, avoiding undue dominant positions (see the “minimum efficient dimension” issue).

If the revenues from congestion charging result insufficient, it is really simple either to use some money from the gasoline taxation (In Europe as an average they are well above the external environmental costs in non-urban areas, that is, are by definition inefficient (IMF, 2014)), or rise these taxes up to the necessary level to cover the marginal (or even the average) road costs.

A simple, transparent, contestable system, far less prone to capture than the useless, expensive and cumbersome present concessions-based regimes.

Further Reading

- Averch, H., Johnson, L., 1962. Behavior of the firm under regulatory constraint. *Am. Econ. Rev.* 52.
- Becker, G., 1983. A theory of competition among pressure groups for political influence quarterly. *J. Econ.* 98, 371–400.
- Becker, G., 1986. The public interest hypothesis revisited: a new test of Peltzman's. *Theory Regul. Public Choice* 49 (3), 223–234.
- Buchanan, J.M., Tullock, G., 1962. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. University of Michigan Press, Ann Arbor, MI.
- Buchanan, J.M., 1965. An economic theory of clubs. *Economica* 32 (125), 1–14.
- Buchanan, J.M., 1969. *Cost and Choice: An Enquiry in Economic Theory*. Markham, Chicago.
- Coco, G., De Vincenti, C., 2008. Optimal price-cap reviews. *Utilities Policy* 16, 238–244.
- Cramton, P., Geddes, R.R., 2015. *Real-time Pricing of Road Access*, Working Paper, University of Maryland.
- Demsetz, H., 1968. Why regulate utilities. *J. Law Econ.*
- Gomez-Ibáñez, J., 2003. *Regulating Infrastructures. Monopoly, Contracts and Discretion*. Harvard University Press.
- IMF, 2014. *Getting Energy Prices Right*, Washington, DC.

- Keeler, T., 1984. Theories of Regulation and the Deregulation. *Movement Public Choice* 44 (1), 03–145.
- Kerf, M., 1998. Concessions for infrastructure – A Guide to Their Design and Award.
- Newbery, D.M., 1998. Fair and Efficient Pricing and the finance of the roads, University of Cambridge.
- Niskanen, W.A., 1971. Bureaucracy and Representative Government. Aldine-Atherton, Chicago.
- Peltzman, S., 1976. Toward a more general theory of regulation. *J. Law Econ.* 19, 211–240.
- Peltzman, S., 1989. The Economic Theory of Regulation after a Decade of Deregulation. *Microeconomics*, Brookings Papers.
- Ponti, M., 2001. The European transport policy in a “public choice” perspective, paper presented at the 9th World Conference on Transport Research, Seoul.
- Ponti, M., 2011. Competition, Regulation, and Public Service Obligation, 661-683, in *A Handbook of Transport Economics*, EEP.
- Ponti, M., Boitani, A., Ramella, F., 2013. The European transport policy: its main issues. *Case Stud. Transport Policy* 1, 53–62.
- Ragazzi, G., 2008. I Signori delle Autostrade, Bologna, Il Mulino.
- Segal, I.R., 1998. Monopoly and soft budget constraint. *Rand J. Econ.* 596–609.
- Stigler, G.I., 1971. The theory of economic regulation. *Bell J. Econ. Manage. Serv* 2 (1), 3–21.
- Transport for London, 2007. Central London Congestion Charging Impacts monitoring, Fifth Annual Report, TfL, London.
- Tullock, G., 1967. The Welfare Costs of Tariffs, Monopolies, and Theft Western. *Econ. J.* 5, 618–630.
- Winston, C., 2010. Last Exit. Privatization and Deregulation of the U.S. Transportation System. Brookings Institution, Washington.

Are Megaprojects too Transformational for Cost–Benefit Analysis?

Tom Worsley, Visiting Fellow, Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

What Constitutes a Megaproject?	470
Why have Megaprojects Become a Priority for Policy?	470
Methods for Appraising Transformational Projects	471
The Standard Cost–Benefit Approach	471
Challenges to the Conventional Approach	471
The Development of Wider Economic Impacts—Augmenting the Conventional Approach	471
Valuing Wider Impacts and Avoiding Double Counting	472
Welfare Economics and GDP in Valuing Agglomeration—Two Different Metrics	473
Estimating the Relocation of Economic Activity Using General Equilibrium and Land Use–Transport Interaction Models	473
Spatial Computable General Equilibrium Models	473
Land Use–Transport Interaction Models	474
Integration of the Broader Approach with the Partial Equilibrium Appraisal Methods	474
Conclusions	474
References	475
Further Reading	475

What Constitutes a Megaproject?

The main distinction between a megaproject and the more typical transport scheme is to be found in the outcomes that the promoters anticipate the project will deliver. Transport users are seen as the main beneficiaries of the typical transport scheme. Improvements in safety and in the local environment are also usually identified as benefits that help to make the case for such a scheme. A megaproject is different. Its objectives extend well beyond the users for whom it serves and those others who are directly affected. Megaprojects are intended as instruments of economic and spatial planning policy, with a focus on generating economic growth and creating new urban development (International Transport Forum, 2014). In some cases, the aim is to foster economic growth in the most prosperous parts of a country, maximizing the potential of infrastructure investment and attracting internationally mobile enterprises. In other cases, a megaproject is aimed at boosting growth in one or more of the less productive parts of the country with the intention of rebalancing the economy and contributing to greater social equity. A megaproject often is put forward as a flagship scheme, demonstrating the vision of its promoters and their commitment to implementing change.

Such a definition is, arguably, more relevant than one based on capital costs. It would be perfectly feasible to define all projects with a cost that exceeded some specified limit as megaprojects although there is a risk that, if those projects in this category were afforded some special treatment, scheme promoters would find ways of bundling schemes together to qualify for megaproject status. More relevant is the responsibility of the analyst in providing the decision-maker with an evidence-based assessment of the likely outcomes of the project and the extent to which the policy-maker's vision for the scheme might be realized. In other words, megaprojects are those for which the conventional cost–benefit framework is, on its own, an inadequate means of giving the policy-maker the best possible advice on the likely outcomes of the scheme that are of interest to that policy-maker. In this respect, a megaproject differs from a package of conventional smaller schemes that deliver the more familiar set of outcomes aimed at transport users and those whose health, safety, and environment are affected.

Why have Megaprojects Become a Priority for Policy?

Most megaprojects are aimed at supplying better links between conurbations or making a conurbation more accessible to its hinterland. They are primarily concerned with serving large cities. Over the last few decades, cities and the major conurbations have become increasingly dominant in most countries' economies. Changes in the structure of industry have resulted in service and other sectors that thrive in a dense urban environment delivering an increasing share of economic output. And the resumption of population growth in many cities in the developed world after a long period of decline has put further pressure on urban transport infrastructure, where construction costs tend to be high because of the complexity of urban transport networks. At the same time, the policy-makers in many cities perceive transport investment as a way enhancing the vitality of the city, fostering economic development through integration with well-designed spatial policies, and demonstrating to potential investors the opportunities offered by the city as a place to do business.

Methods for Appraising Transformational Projects

The Standard Cost–Benefit Approach

Cost–benefit analysis (CBA) was developed as a means of ensuring that transport policy-makers had a method for determining priorities between mutually exclusive options and between projects competing for funding. It had the additional advantage of providing the funding department, in most countries the ministry of finance, with evidence from the comparison of the net benefits of the preferred option with the alternative of do-nothing, the reference case, as a means of demonstrating that public money was being spent prudently. And it demonstrated to the wider community, whether parliament or the public, that the overall costs of the scheme, including those imposed on the landscape, on those whose homes were purchased to make way for the scheme, and on others affected, were outweighed by the benefits to transport users and others who provide and consume the goods and services that rely on the transport network. Policy-makers could be satisfied that the implementation of a proposed scheme with benefits in excess of its costs was in the public interest.

Most of the early applications of CBA were to highway schemes because, even in the case of tolled roads, a financial appraisal fails to reflect many of the costs and benefits that follow from the scheme. Techniques were developed over time to extend, define, and quantify the range of impacts included in the appraisal. Changes attributable to the scheme in respect of people exposed to noise, poor air quality, and visual intrusion were quantified and valued in monetary terms, while other nonmonetized environmental impacts were set out as part of the case presented to policy-makers.

Challenges to the Conventional Approach

Toward the end of the last century, the standard approach to CBA faced a number of challenges. Policy-makers wanted to know more about the likely outcomes of a project than was being provided by the standard economic welfare-based method. They wanted to know whether investing in transport would foster economic growth and how any such impacts might be incorporated into the appraisal methods so as to prioritize schemes that would increase productivity. And they also wanted to understand the spatial distribution of the costs and benefits and the role that transport might play in regional economic policy. At the same time, the shift in economic activity toward conurbations, as noted earlier, resulted in an increase in the proportion of the indicative budget to be spent on high cost urban schemes. Policy-makers wanted to use transport investment as a tool of spatial and economic planning.

In the United Kingdom, the UK government responded to these challenges by commissioning a report from the Standing Advisory Committee of Trunk Road Assessment (SACTRA) on the relationship between transport and the economy and on the adequacy of the UK's transport appraisal methods in the context of their findings ([Standing Advisory Committee of Trunk Road Assessment, 1999](#)). SACTRA noted the conditions under which the standard partial equilibrium (PE) cost–benefit methods apply. These include assumptions that prices equal marginal costs in the transport using market. SACTRA identified in their 1999 report the circumstances under which imperfect competition in transport using markets and the associated positive or negative externalities would result in the conventional approach giving biased estimates of the actual benefits.

SACTRA's review of the links between transport and the economy covered the evidence of an econometric relationship between aggregate investment in transport and in infrastructure and economic growth. More relevant to CBA was the emerging understanding of the relationship between agglomeration and productivity and of transport's role in agglomeration, a source of evidence that could be applied at a scheme-based level. SACTRAs also reviewed the recent developments in general equilibrium (GE) and land use–transport interaction (LUTI) models, which provided a possible means of analyzing the impacts of transport on the economy at a more detailed spatial level. However, the UK's Department for Transport concluded in its response to SACTRA that neither GE nor LUTI models had reached a stage at which they might provide an alternative to the PE cost–benefit method.

The Development of Wider Economic Impacts—Augmenting the Conventional Approach

A consequence of the SACTRA report and of research into causes of differences between places in the productivity of their labor force was the addition to the conventional transport appraisal methods of wider economic impacts (WEIs) ([Department for Transport, 2005](#)). The UK's WebTAG guidance, which follows similar principles to those used in many other countries' appraisal methods ([Eliasson et al., 2014](#)), identified three separate types of wider impact (<https://www.gov.uk/.../webtag-tag-unit-a2-1-wider-economic-impacts-may-2018>). Labor market effects are made up of the impact of a change in the costs of commuting. Changes in commuting costs influence people's participation in the labor force, resulting in a labor supply effect. In addition, changes in commuting costs have an effect on the choice of workplace made by people already in the labor force. A second category of impact, induced investment effects, is concerned with the role played by a scheme in changing the location and quantity of new housing and commercial development induced by the scheme and hence its effect on decisions made by firms about the supply and location of workplaces. The third category of wider impacts is made up of an assessment of the changes in productivity on account of the transport scheme resulting in a change in a measure defined as access to economic mass. These wider impacts, including all of the changes in the location of economic activity, have an impact on output and on productivity that forms the core of the additional benefits attributable to a megaproject. For most megaprojects the direct and indirect effects on access to economic mass have made up the largest component of the wider impacts.

Analysis of data on productivity—defined as output per worker—showed a causal relationship between economic mass and productivity ([Graham, 2007](#)). Better-connected cities are more productive places. Thus, investment in transport to improve

accessibility in a city increases economic mass and this results in an increase in productivity, a relationship defined in terms of an elasticity of productivity with respect to economic mass, with the elasticity varying by industrial sector. The analysis also showed that the effect of changes in economic mass on productivity declined with distance from the project, with much of the evidence suggesting minimal productivity effect in places more than around 20 km from the scheme. Improvements in urban connectivity result in more efficient, deeper labor markets with the better matching of people to jobs, opportunities for sharing knowledge between workers and more efficient markets for the goods and services that make up the city's economy.

The measure of access to economic mass is made up of two parts. The first is the change in accessibility or travel costs between that zone and all other zones, with the value of the measure for the other zones each weighted by the number of employees in that zone. This effect has been defined as the static agglomeration effect, in that it relies only on the changes in transport costs. Thus, transport costs are an integral part of the measure of economic mass.

The second is the number of workers in each zone in the city, with the zones often defined so as to correspond with the structure of zoning adopted in the transport model. Cities that have experienced an increase in economic mass induced by a transport scheme tend to benefit from a second-round effect. Workers and firms are attracted from elsewhere and relocate in the city, while existing firms expand because of both the better access to workers, suppliers, and customers, and because each expanding or relocating firm benefits from the increase in productivity from the first-round static agglomeration effect. This second-round effect is made up of two separate impacts. The first is the increase in employment in city center zones, thus increasing the economic mass of the zone. In addition, workers who move to more productive zones benefit from a placed-based effect and tend to become more productive when working in a zone with greater economic mass.

The change in productivity caused by the change in the location of employment is often defined as the dynamic agglomeration effect because it described the impact of a shift in economic activity between places with different levels of productivity.

The effect of the change in accessibility on economic mass, defined earlier as the static agglomeration impact, can be estimated directly from the transport model used to predict the reduction in travel time and congestion experienced by transport users. Estimates of the increase in employment density as a result of the relocation of activity require the use of additional modeling methods discussed in the following section.

Valuing Wider Impacts and Avoiding Double Counting

When establishing a method for valuing these wider economic benefits within the cost-benefit framework, attention has been paid to avoiding any double counting by ensuring that these additional benefits do not include some of the benefits already attributed in the appraisal to transport users. The increase in output per worker caused by the transport investment's direct effect on agglomeration and hence on productivity, the costs of which are already part of the cost-benefit calculation, is counted as an additional welfare benefit. No extra input is required of the workers who benefit from working in a better-connected city, with the opportunities that this facilitates. So the CBA includes both the additional posttax income earned by workers, who get utility from this effort-free increment to their earnings, and the tax they pay on these earnings because society as a whole benefits on the grounds that the other taxes they pay can now be reduced, albeit by a very small amount.

The benefits to workers and firms that relocate to a more productive destination form part of the conventional transport user benefits and are, in any well-conducted cost-benefit appraisal, already part of the transport user benefits. In order to earn this additional income, workers are obliged to change where they work, a change they are induced to make because of the new transport opportunities. Given that they could always have worked in the new location before the scheme opened, the maximum benefit that can be attributed to them is the full amount of the reduction in transport costs, with the most marginal person who moves getting a minimal benefit. Hence the average generated traveler is assumed to gain by half of the benefit of existing users—the well-established “rule of a half.” Their decision to move jobs is based on the trade-off they face between transport costs and posttax wages. As in the case of the first-round agglomeration impact, society benefits from the additional tax revenue on the higher earnings of those who move to a more productive job and this tax revenue provides the source of additional benefit in the appraisal of the scheme.

The method of valuing the agglomeration and other wider impacts specified earlier therefore avoids any doubled counting of these impacts with travel time savings. The agglomeration impacts are positive externalities—benefits derived by society from the actions of individuals in paying higher taxes and in the increased output that is a result of the increased accessibility making all workers more productive without any change in their behavior of the static agglomeration effect. So the PE methods of CBA can be expanded to cover these second-round impacts without compromising the theoretical basis of the method.

In addition, the approach helps to identify the likely spatial distribution of these wider impacts. The assessment of a project, which is based on an appropriate transport model, will show those zones served by the scheme where productivity is increased because of the increase in the access to economic mass. It will also show those places that experience a relative decline in economic mass as workers move to more productive locations and hence are likely to experience a loss of jobs and productivity. Policy-makers have found this information about the likely spatial distribution of the economic impacts of a major project invaluable. It enables them to formulate complementary programs of investment to facilitate the changes in the distribution of economic activity and to respond to any unwanted negative impacts, such as the two-way road effect, whereby a smaller city loses jobs to the larger one when transport links between the two are improved.

Welfare Economics and GDP in Valuing Agglomeration—Two Different Metrics

CBA is founded on the concepts of welfare economics, aimed at maximizing people's utility or the value they place on all the changes related to the transport scheme that affect them. It is not intended to measure the effect of the scheme on the national economy, defined in terms of GDP, a metric that is restricted to marketed goods and services. Impacts such as savings in travel time on leisure trips or environmental impacts figure in the cost-benefit methodology but are not part of the conventional national GDP accounting process.

However, most governments have an objective of raising productivity and so of increasing GDP as an indicator of the growth in a country's prosperity (Vickerman, 2017). One of the strengths of the appraisal of the wider impacts of a transport scheme is that these acts can be valued either in the context of a welfare cost-benefit framework as described earlier or in terms of GDP and national accounting methods. The increase in output per worker is caused by the transport investment, the costs of which are already part of the cost-benefit calculation. No additional input is required of the workers who benefit from working in a better-connected city, with the opportunities that this facilitates. As noted earlier, CBA includes both the additional posttax income earned by workers and the tax they pay on these earnings from which society as a whole benefits. The additional pretax earnings of workers which can be attributed to the scheme will be counted in a country's national accounts as a measure of increased productivity and hence as an impact on GDP. In this respect, the welfare impact and the GDP effect are identical.

The additional earnings generated by relocation are treated differently between a CBA and a GDP metric. As noted earlier, the welfare benefit gained by those who change where they work are measured through the change in transport costs they experience, with the most marginal person who moves gaining a minimal benefit and hence the average traveler who moves jobs getting half of the benefit of existing users. Added to this is the social benefit from the additional tax revenue on these higher earnings. But the GDP metric counts the entire increase in output since the additional commuting costs or the additional demands on the worker when in the higher paid job do not count as costs in the GDP calculus to be offset against the increase in productivity.

Estimates of the GDP effect have helped decision-makers understand the potential for raising funds for major projects. Such estimates can show the extent to which the initial capital costs might subsequently be recovered through tax payments on the additional income generated by the changes in the location of economic activity caused by the project. An analysis of the GDP effects of London's Crossrail scheme helped to make the case for the introduction of the business rate supplement, an additional local tax which has been levied on all large firms in London and which provides around a third of the funding for the scheme.

Estimating the Relocation of Economic Activity Using General Equilibrium and Land Use–Transport Interaction Models

Various approaches have been used to model those impacts of transport schemes that fall outside the first-round effect of a change in generalized costs experienced by transport users. There are two broad categories of approach—one based on GE modeling and the other on LUTI models. Both serve to provide estimates of the impact of a major transport project on the location of economic activity.

Spatial Computable General Equilibrium Models

The GE approach is aimed at modeling the whole economy with some level of spatial detail in the case of it being applied to transport schemes. It provides a coherent account of trade, production, investment, and employment. The approach makes projections of these key variables at a national and regional level both in a reference case and for a range of scenarios for different policy interventions and external shocks. The methods used are capable of representing key relationships, including those between accessibility, the location of employment and of households, and the impact of location on productivity. GE models can also represent the role of the public sector and the interaction of tax revenues, public expenditure, and interest rates. GE models have proved their usefulness over many years in considering policies relating to trade and taxation, accounting for the multiplier effects of income on employment and subsequent responses in the economy.

Spatial computable general equilibrium (SCGE) models are primarily aimed at providing estimates of the change in the level and broad location of economic activity as the result of a change in policy. The method has been adopted to provide decision-makers with estimates of the effect of a scheme on local and regional employment and on productivity. In this respect, the analysis has been used to give decision-makers an alternative measure of a transport scheme's expected outcomes which is relevant to wider policy objectives and provides additional context to the conventional benefit-cost ratio. GE models can deal well with the comparison of the reference case with the project: their completeness ensures that they can distinguish between effects which are a net addition to economic activity and those which are no more than a change in the location of a given level of activity. Such distinctions are often left to judgment in the PE approach framework.

SCGE models have seen only limited application in the United Kingdom. They have been used by the UK's Department for Transport in the appraisal of the Lower Thames Crossing, a scheme with an objective of reducing transport costs as a means of improving the efficiency of labor markets in north Kent and in Essex and by the UK's Airports Commission to assess the impact of runway capacity on productivity and employment. In these two examples, the model was used to provide estimates of the net impact of the project on regional and national employment and GDP impacts. It also provided estimates of the dynamic agglomeration impacts as an input into the modified PE method of estimating dynamic agglomeration impacts described in the section earlier.

Land Use–Transport Interaction Models

LUTIs have been used for many years to demonstrate the impacts of the changes to land use that are likely to result from the changes in accessibility that follow from a transport scheme ([The London Land Use and Transport Interaction Model, 2014](#)). Since such changes generally depend on land-use planning consent, the models have been used to inform scheme promoters and land-use planners about the pressures for land-use change and the consequences of such changes on traffic flows, the environment, and on local services.

LUTI models represent the responses of households, employers, and developers to changes in accessibility measured through changes in transport costs. In order to estimate such responses, the models include data on all other costs, including rents and wages, incurred by firms and households and on the interaction between agents in the model. As a consequence of this framework, the LUTI approach has been extended into a model of the local, regional, or national economy that shares many of the characteristics of the full GE approach while generally providing a greater level of spatial detail. It is not as comprehensive—for example, LUTI-based models tend to lack a government sector and so government spending does not crowd out or otherwise affect private sector investment. And, because the economic model contains less spatial detail than the transport model that serves the purpose of informing the design of the scheme, the modeling framework makes use of a simplified and cruder version of the latter.

An appropriate LUTI model will incorporate information on local planning permissions and can help to demonstrate the extent of any potentially profitable development which is dependent on the scheme, thus helping to bridge the gap between aspirations of the scheme promoters and likely market forces. The credibility of the model's predictions of the location and level of transport-induced development is strengthened by the inclusion in the model of these estimates of the costs and profitability of development.

Integration of the Broader Approach with the Partial Equilibrium Appraisal Methods

Neither these approaches provide, in their present form, a complete alternative to the PE approach for estimating welfare benefits. While they are used for estimates of the dynamic agglomeration benefits of the move to more productive jobs, the conventional transport user benefits are derived from the relevant appraisal values, such as values of time saving, and from the changes estimated in the transport model. There is therefore a potential inconsistency between the benefits derived from the conventional PE method, which is based on assumptions of fixed land use, constant returns to scale, and perfect competition in the markets for goods and services which make use of the transport network and the wider impacts estimated in the SCGE or LUTI-based supplementary economic models.

As with all such models, the assumptions made to generate these predictions of land-use change, while based on the best available estimates, remain very uncertain. The parameters in the model that represent the many responses of agents to changes in transport costs and in the associated costs of doing business in different locations are taken from a wide range of sources of differing levels of reliability and in some cases are based on the modeler's judgment. There have been cases in which predictions of a scheme's impact on GDP have exceeded, by an order of magnitude, the estimates of the welfare benefits despite the latter taking account of leisure transport user time savings and other benefits omitted from the GDP metric. Such cases tend to assume a significant level of spare capacity in the economy that will be brought into productive use through a multiplier effect acting on the agglomeration-related increase in output. Yet predictions of the level and spatial location of unemployed resources in the economy 10 or 15 years hence when eventually the scheme opens are highly speculative.

The United Kingdom has followed the practice of using spatial economic models to give estimates of the change in the location of employment as a means of estimating the dynamic agglomeration impacts of a major transport project. Assumptions about the relationship between productivity and economic mass are taken from the extended PE approach described earlier. And, as described earlier, the effects of the relocation of economic activity can be expressed either in terms of the conventional welfare economics metric or as an increase in regional or national GDP.

Conclusions

The methods described earlier go some way toward providing decision-makers with the additional information they need about the likely impacts of megaprojects. And the facility to give an estimate of a scheme's impact on regional and national GDP significantly increases the value of evidence-based analysis in the eyes of decision-makers. Nevertheless, these modifications to CBA have a number of limitations. The wider impacts approach is best suited to estimating the effects on a single agglomeration: it is less well able to model the effects of a scheme which connects two or more potentially competing urban areas by showing whether greater specialization might be a result of such a scheme. Nor does it address the potential impacts on the freight and distribution sectors of major improvements to highway networks.

The methods tend to show that the benefits of a megaproject are greatest in more prosperous and productive places. They do not go very far in helping analysts give advice to decision-makers on whether or how transport investment might promote the development of those parts of cities or regions that have experienced economic decline.

Nevertheless, structure and form of the model provides a more disciplined method of understanding likely outcomes and the forces that influence these outcomes than an assessment made without such a framework. A better understanding of the likely changes in the location of jobs and of workers makes it possible for policy-makers to plan other complementary measures to

facilitate the changes and to minimize the disruptive effects of a major scheme. For example, a LUTI-based economic model can show how a major transport project will affect the market for housing in the places it serves and the extent to which any offsetting influence of rising house prices on employment and productivity might be mitigated by the release of more land for house-building. And the approach can also show, through the use of a detailed urban transport model, the role of improved local transport links needed to distribute users of the megaproject across the urban center.

References

- Department for Transport, 2005. Transport, Wider Economic Benefits and Impacts on GDP. Department for Transport, London (modified 2006). Available from: <http://webarchive.nationalarchives.gov.uk/+/http://www.dft.gov.uk/pgr/economics/rdg/webia/webmethodology/sportwidereconomicbenefi3137.pdf>.
- Eliasson, J., Mackie, P.J., Worsley, T.E., 2014. Transport appraisal revisited. *Res. Transp. Econ.* 47, 3–18.
- Graham, D.J., 2007. Agglomeration, productivity and transport investment. *J. Transp. Econ. Policy* 41 (3), 317–343.
- International Transport Forum, 2014. Appraising Transformational Projects: The Case of the Grand Paris and Express. OECD, Paris.
- Standing Advisory Committee of Trunk Road Assessment, 1999. Transport and the Economy. TSO, Norwich, UK. Available from: http://webarchive.nationalarchives.gov.uk/20050301192906/http://dft.gov.uk/stellent/groups/dft_econappr/documents/pdf/dft_econappr_pdf_022512.pdf.
- The London Land Use and Transport Interaction Model (LonLUTI), 2014. David Simmonds Consultancy. Available from: www.davidsimmonds.com/publications.
- Vickerman, R., 2017. Beyond cost benefit analysis: the search for a comprehensive evaluation of transport investment. *Res. Transp. Econ.* 63, 5–12.

Further Reading

- Airports Commission, 2015. Wider economic impacts assessment. Available from: www.gov.uk/government/organisations/airports-commission.
- Department for Transport, 2017. WebTAG units (TAG unit A2-1 wider economic impacts appraisal; TAG unit A2-2 induced investment; TAG unit A2-3 employment effects; TAG Unit A2-4 productivity impacts; TAG Unit M5-3 supplementary economic modelling), London. Available from: <https://www.gov.uk/guidance/transport-analysis-guidance-webtag>.
- Vickerman, R., 2017. Beyond cost benefit analysis: the search for a comprehensive evaluation of transport investment. *Res. Transp. Econ.* 63, 5–12.

Economics of Transportation Safety

Ian Savage, Department of Economics and the Transportation Center, Northwestern University, Evanston, IL, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	476
Economics and Engineering	476
How Safe is Safe Enough?	477
Socially Desirable Decisions by Private Users	477
Socially Desirable Decisions for Commercial Carriers	477
Commercial Passenger Transportation	478
Commercial Freight Transportation	478
Socially Desirable Decisions by the Infrastructure Providers	478
Socially Desirable Decisions by Emergency Responders	478
Summary of Desirable Decisions	478
Why the Market Fails	478
Externalities	479
Bilateral Crashes	479
Cognitive Failures by Individuals	479
Imperfect Information	479
Carrier Myopia	479
Imperfect Competition	480
Relative Magnitude of the Market Failures	480
Market Interventions	480
Liability	480
Insurance Requirement	480
Information Provision	481
Safety Regulation	481
Closing Comments	481
References	482

Introduction

Transportation has always been risky. Our ancestors faced risks from being thrown from horses or drowning in rivers. The mechanization of transportation increased the forces exerted on the human body when something goes wrong. Mass transportation increased the number of possible casualties in any incident. Private grief becomes a public spectacle.

Highway transportation is particularly risky. The World Health Organization reports that highway crashes are the eighth leading cause of death with 1.4 million annual fatalities. (Note that safety professionals prefer the word “crash” to “accident” because the latter suggests that occurrence is due to pure fate and cannot be influenced by human decisions.) The per capita risks in low- and middle-income countries are 3 times higher than in the safest countries in northern Europe despite relatively low levels of motorization. In contrast, the risks on railways, waterways, and airways are much lower. But incidents in these modes attract considerable public attention relative to highway crashes.

Economics and Engineering

The proximate cause of crashes is an interaction between vehicle operators (riders, drivers, pilots, etc.), their vehicles, and the infrastructure (highways, railway track, waterway, airway, etc.). For highways, the matrix by Haddon (1972) categorized the factors that explain crash causation and consequences. This three-by-three matrix has the driver, the vehicle, and the highway on one axis, and factors that occur prior to a crash, during a crash, and after a crash on the other axis. This matrix along with illustrative factors is shown in Table 1. While this matrix is specific to highway crashes, clear analogies can be made to other modes.

Economics is a complement to, and not a substitute for, the work of human factors and engineering professionals. Economics is important because, long before a crash occurs, drivers have made decisions on how aggressively they drive and how well they maintain their vehicles. Bus and trucking companies have decided on employee training. Highway engineers have decided on design characteristics of the road. The legislature and police authorities have set traffic laws and decided on how aggressively to enforce these laws. Public bodies have decided on budgets for the provision of first responders and trauma centers.

Each of these actors has made decisions by comparing the benefits and costs of their actions. To an economist, the “safety problem” emerges because their decisions do not accord with the best interests of society as a whole. There are “too many” crashes.

Table 1 Haddon's matrix

	<i>Before the crash</i>	<i>During the crash</i>	<i>After the crash</i>
The driver	• Conduct (speed, etc.) • Skills • Vehicle maintenance	• Use of active safety devices (fastening of seat belt prior to the trip, use of child safety seat)	• Skills and equipment of first responders
The vehicle	• Design • Safety equipment • Collision detection • Collision avoidance	• Energy absorption • Cabin design • Passive safety devices (air bags)	• Ease of extraction from vehicle • Integrity of fuel and battery systems
The highway	• Design • Materials • Maintenance	• Guardrails • Breakaway sign posts	• Ease of access by first responders • Location of first responders and trauma centers

Economists describe this as a market failure. By understanding why actors do not act in a socially optimal way, interventions can be designed to persuade actors to modify their behavior and ameliorate the market failures.

How Safe is Safe Enough?

Crashes are undesirable, yet costly to prevent. While society cannot afford to eliminate all crashes, it certainly would prefer that there are fewer crashes than there are today. Determining exactly how many crashes are acceptable to society is difficult. However, understanding the process that actors go through in determining safety provides insights that are useful when designing interventions in the market.

Socially Desirable Decisions by Private Users

Models of safety determination fall into two broad categories. The first category is models of private users operating their own vehicles. This category includes:

- automobile drivers;
- motorcyclists;
- pedal cyclists, pedestrians, and other nonmotorized road users;
- pilots flying their private planes; and
- recreational boaters.

Private users invest money, time, and effort to reduce risks. They invest in their skills, purchase vehicles with more safety features, and pay for maintenance. They can conduct themselves cautiously, perhaps at a cost of taking longer for their trip. This can be referred to as preventive effort. Users trade off these costs against a reduced probability of a crash or increased survivability in the event that a crash occurs. In making their decisions, they adjust for the state of the infrastructure and their assumptions about the effectiveness of emergency response.

While all users prefer more safety to less safety, they may vary in their prevention costs (thrill-seeking users find it more costly to act cautiously) and the magnitude of the losses in the event of a crash (e.g., the losses incurred by a head of household vs. a single person). Consequently, users may rationally exercise different levels of prevention and hence have different safety outcomes.

The variation in safety outcomes is not necessarily a market failure. However, choices are only socially desirable if actors are (1) fully informed about the cost of preventive actions, (2) are aware of the consequences of their actions on crash occurrence and survivability, and (3) are responsible for the full consequences of any crashes that they cause, including damages to other users and bystanders.

Socially Desirable Decisions for Commercial Carriers

The second category of models describes markets where passengers and freight shippers contract with a commercial carrier. Examples include:

- taxis, buses, and other forms of public highway transportation;
- trucking;
- rail passenger and freight services;
- passenger and freight airlines;
- maritime companies; and
- pipelines.

There is a principal-agent problem in that while passengers and shippers have preferences for the level of safety they are willing to purchase, it is the carriers' conduct that produces safety outcomes (Maurino et al., 1995). Passengers and shippers have to select a carrier whose safety matches their own preferences.

Commercial Passenger Transportation

Carriers expend preventive costs to lower the safety risks. While at low safety levels, expenditures on prevention may result in a reduction in total costs due to fewer crashes; in general, carriers that offer higher safety have to charge higher fares to break even. Passengers with a high valuation of safety, and the ability to pay the higher fares, gravitate to carriers offering higher safety. A vertically differentiated market may exist with high safety—high price carriers coexisting with low safety—low price carriers. Often, a variety of safety levels in the market are seen as an indication of market failure. But this is not the case if passengers vary in their tastes for safety, and are able to recognize the safety level offered by different carriers.

Commercial Freight Transportation

Shippers of freight look to minimize the cost of transporting their goods. But in doing so, they have to trade off the costs of more safety with the consequences of any crashes that occur. Because different commodities have different costs of handling and cause different levels of harm in the event of a crash, it is likely that safety varies by type of commodity. Delicate cargoes and those that present high hazards in a crash are handled with considerable care, and more robust and benign cargoes move at lower levels of safety. As with passenger transportation, the variety of safety outcomes does not necessarily indicate a market failure. But socially desirable decisions depend on shippers being knowledgeable about the level of safety on offer, and on carriers, and hence shippers, bearing the full consequences of crashes. The latter is particularly important for hazardous materials, where releases can harm bystanders.

Socially Desirable Decisions by the Infrastructure Providers

In vertically integrated modes such as pipelines and railways, carriers also control safety of the infrastructure. In aviation and maritime, commercial and private users depend on decisions made by ports and airports, and traffic control services. But the distinction is most stark on roads. Design and maintenance of the highway is crucial to system safety, yet there is a separation between the highway authority and the users.

How should the highway authority decide on appropriate design? Economies of density and scale mean that there are a limited number of roads between two places, and a common highway authority presides over the entire network. Consequently, a “one-size-fits-all” level of safety is offered to all users irrespective of their safety preferences. In an ideal world, the authority would select a level of safety that provides the greatest amount of social benefit. This level may not sit well with users who desire a higher level of safety, and those users who would prefer the lower user fees associated with a lower level of safety. There is already a market failure inherent in road provision associated with imperfect competition or monopoly.

Socially Desirable Decisions by Emergency Responders

An often-underestimated way to improve safety is better emergency response. Prompt medical attention during what is called by emergency physicians the golden period (often about 60 minutes) after trauma occurs can be crucial in preventing fatalities. Prompt response to crashes involving hazardous materials can be crucial in mitigating harms to people and property. As with highway infrastructure provision, a “one-size-fits-all” approach is necessary recognizing that emergency services and trauma centers serve other hazards in addition to transportation crashes.

Summary of Desirable Decisions

To summarize the discussion thus far, if transportation users (1) are fully aware of the risks, (2) are fully aware of the costs and benefits of mitigating the risks, and (3) voluntarily accept the risks to gain the benefits of mobility and economic opportunity, then there is not a “safety problem.” Moreover, safety outcomes can vary between individual private users and various commercial carriers. This results from variations in willingness to pay for safety by private users and passengers, and variation in the safety prevention costs and harms caused for freight commodities. High safety may optimally coexist with lower levels of safety in vertically differentiated markets. To an economist, this is an indication that the market works and does not necessarily indicate a market failure.

Why the Market Fails

The underlying relationships that describe the socially desirable outcomes are difficult to measure. Calculating the optimal levels of safety in a particular mode is not trivial, and may be practically impossible. However, there are extensive failures in the market for safety. These lead to many actors exerting less preventive efforts than is socially optimal. Consequently, a productive role for public policy is to identify the extensive market failures and devising interventions to ameliorate or correct them. This may be more productive than agonizing over estimating what the level of safety “should be.”

The failures derive from the following three underlying characteristics of safety:

- For most actors, the beneficiaries of their preventive efforts include other actors. Unless they are altruistic, this leads the actor to underinvest in prevention because they reap only part of the benefit. Economists call this an uncompensated externality.

- The costs of prevention are incurred in the present and are incurred on every trip. The benefits of reduced crashes do not occur on every trip and only happen at randomly determined points in the future. Actors may discount these future benefits and underinvest in prevention in the present.
- Unlike characteristics such as price that are readily observable, actors may be poorly informed and may purchase a product and service that does not match their tastes in safety. This is because safety is a probabilistic attribute and may not be observable on every trip. Moreover, other actors might avariciously take advantage of poorly informed purchasers by selling a low safety product at a price that is consistent with higher levels of safety.

Externalities

Safety externalities are abundant. Injuries are sustained by innocent vulnerable road users struck by motorized vehicles. Bystanders are affected by hazardous materials releases. Legal structures have developed to assign fault and recover damages. But, they are not a panacea. There may be limitations on the harms that can be legally recovered, and victims may find that the responsible party does not have the financial resources to pay compensation.

Bilateral Crashes

While a surprisingly large number of crashes involve a single vehicle (e.g., a motor vehicle running off the roadway), collisions with other users predominate. The actions of both parties determine the probability of a crash. A complex legal literature describes how fault should be assigned to ensure that the actor who can reduce the risk at the lowest possible cost should be given incentives to take action (Shavell, 2004). Market failures abound in the adjudication of fault in multi-vehicle crashes.

Cognitive Failures by Individuals

Private users make decisions to expend effort to mitigate risk in the present, but the consequences of their actions or inactions occur in the future. Passengers have to choose between the level of fares and future crash risks. Some individuals may be myopic in ignoring or downplaying the future consequences. Others may lack self-control when trading off effort in the present with future negative consequences. More broadly, humans tend to be very poor in evaluating the probability of low-probability crashes, and thinking about the consequences of deadly events. There is considerable overconfidence, with the majority of drivers believing that they are safer than the average driver, which cannot be true statistically. Individuals also suffer from cognitive dissonance in that nearly all trips are completed safely reinforcing beliefs that a crash “will not happen to me.” As a result of this range of cognitive failures, users downplay the future probable benefits from investing in safety in the present. They take less care than they should.

Imperfect Information

While fully informed actors can have cognitive problems in processing information, the problem is compounded when actors are poorly informed. Private users may be poorly informed about the safety characteristics of their vehicles and the magnitude of the potential safety benefits from modifying their conduct. The situation is even worse for commercial transportation. Passengers and shippers have to form an opinion about the levels of safety that carriers are offering. Unlike other attributes of service such as price and speed, information on safety is difficult to observe, obtain, or understand. Crashes are rare events and tend to be a poor indicator of the safety of individual carriers.

Vertically differentiated markets require customers to sort themselves among the safety offerings, and this is difficult when information is poor. If customers are totally uninformed, all carriers offer a low level of safety. Carriers would be unable to convince customers that they are offering a high level of safety with a commensurate higher price. Therefore, differences in safety offerings indicate that customers are at least partially informed. Albeit, poor information is endemic in safety markets.

Carriers also suffer from imperfect information. Carriers rely heavily on the skills of their employees to produce safety. Yet, they are imperfectly informed about their skills at the point of hiring, and frontline employees tend to perform their duties without direct supervision. Especially in commercial road transportation, drivers with poor safety records can successfully masquerade as having higher skills in order to work at carriers who pay higher wages and wish to provide a high-quality product.

Carrier Myopia

As with private users, some carriers may be myopic in that they downplay the future consequences of their current actions. They may be very well aware of the costs of employee training but do not appreciate the beneficial effect on future crashes. This failure is particularly associated with inexperienced new entrants to the market.

The problem also extends to incumbent carriers. If passengers and freight shippers are imperfectly informed, then conditions are ripe for some carriers to cheat. Carriers that previously offered a higher level of safety could cut their safety investments (and hence their costs) while still masquerading as a higher-safety carrier and charging a high price. Such carriers can earn profits, at least until the customers find out and either shun the carrier or demand a lower price. Why might a carrier do this? Carriers close to bankruptcy

may be particularly susceptible. The carrier may hope that the cost savings provide a buffer until favorable trading conditions return. Other financially stressed carriers may discount the future consequences of crashes because they know that they do not have the finances to meet any judgments.

Economists tend to be skeptical about such arguments, because a carrier has invested to obtain a reputation for high quality. Why would such a carrier want to squander its reputation? Nevertheless, cheating tends to be very prevalent, and has severe consequences. Customers end up purchasing a service that differs from their tastes, and, in the event of a crash, the carrier may not have the financial resources to pay compensation.

Imperfect Competition

The market failures due to lack of choice in modes with economies of density and scale have already been discussed. Even in inherently competitive modes such as highways, a “one-size-fits-all” level of infrastructure safety results in a market failure. When there are few competitors, a small number of safety choices are available in the marketplace, and most passengers and shippers cannot obtain the exact level of safety that they desire. In commercial modes indivisibilities in vehicle size (airplanes, trains, and less-than-truckload trucking) mean that customers with different tastes in safety share the same vehicles, and a limited number of safety options are on offer.

Relative Magnitude of the Market Failures

The applicability and magnitude of the market failures varies significantly between modes and between private users and commercial passenger and freight services. **Table 2** provides a summary using a star rating of the relevance of the six market failures to the various modes and market segments. The more stars indicate the more prevalent market failures.

Market Interventions

Recognizing how the market failures originate and which actor(s) they affect is the basis for intelligent public policy prescription. Policy responses need to be tailored to the root causes of the problem, and the actor who can most effectively change the market outcome. No intervention is a panacea by itself, and some interventions have their own weaknesses. Consequently, these interventions should be thought of as complements and not substitutes.

Liability

Externalities and bilateral crashes are directly addressed by making responsible parties legally liable for their actions (Shavell, 2004). Liability is a powerful solution to these problems but it is not without limitations. The law may limit some losses from being recovered (e.g., emotional harm or lost profits as opposed to physical damage). It is also an “after-the-fact” market intervention. Private users and commercial carriers who are myopic may still underinvest in safety, even if they ultimately have to pay compensation.

Insurance Requirement

Insurance goes hand in hand with liability because of the concern that a responsible party may not have the resources to satisfy claims (Dionne, 2014). Insurance also tackles the problem of myopia by transforming future consequences into premiums that have

Table 2 The magnitude of the six market failures by mode

	<i>Externalities</i>	<i>Bilateral crashes</i>	<i>Individual cognitive failures</i>	<i>Imperfect information</i>	<i>Carrier myopia</i>	<i>Imperfect competition</i>
Private driving	*	***	***	**	Not applicable	*** ^a
Private aviation and boating	*	*	**	*	Not applicable	*
Commercial passenger	*	**	**	***	***	**
Road freight	***	**	Few	**	***	*** ^a
Airfreight	*	*	Few	*	***	Few
Maritime freight	**	*	Few	*	***	Few
Rail freight	**	***	Few	*	**	***
Pipelines	***	Not applicable	Few	*	**	***

Notes: *, limited failures; **, some failures; ***, substantial failures.

^aFailure comes from the provision of highway infrastructure.

to be paid in the present. Private users and commercial carriers can trade off greater safety investments against reduced premiums in the present, and there is less chance of myopia. Insurance companies also have incentives to monitor the activities of their clients, and this gives incentives to insured parties not to shirk on preventive efforts.

All insurance suffers from two main drawbacks. Moral hazard occurs when an actor covered by insurance acts in a riskier fashion than it would do otherwise because the insurance company is responsible for paying any claims. If insurance is optional, there are problems of adverse selection when low-risk clients opt not to purchase, and premiums increase as only higher-risk clients remain.

Information Provision

The obvious solution to imperfect information is making actors better informed. New private users and operating employees of commercial carriers are typically required to take classes and pass a test before receiving a license. The classes and tests focus on understanding the risks and rules of conduct. Subsequently, users are bombarded by public information campaigns warning of the consequences of fatigue, distraction, and alcohol.

Information can be provided to passengers and shippers to help them decide which carrier to select. Rating schemes can be put in place by the government or private providers. This is unambiguously a good thing, as customers become aware of which carriers offer poor service, and the customers provide a discipline on the market. However, such information cannot deal with the problem of cheating. Cheating carriers deviate from their past safety performance. Retrospective information on crashes may not be a reliable prediction of future performance. It is often argued that information should be provided on safety inputs, such as employee training, as opposed to safety outputs (crashes) as a predictor of future safety performance.

Safety Regulation

A familiar intervention is setting and enforcing minimum standards. These standards can be on user training and conduct, the safety features of vehicles, infrastructure design, emergency response, and management processes adopted within commercial carriers. These regulations only define a minimum. Users and carriers are free to adopt higher safety levels if they prefer. Therefore, safety regulation is not a good solution to the problems of imperfect information in passenger and freight transportation because it does not inform customers about the range of safety that is offered. For mode specific details, see [Elvik et al. \(2009\)](#), [Evans \(2004\)](#), [Kristiansen \(2005\)](#), [Lamm et al. \(1999\)](#), and [Savage \(1998\)](#).

A common regulatory concern is known as risk compensation ([Blomquist, 1988](#)). This is when a user or carrier compensates for regulatory action in one dimension by undertaking riskier actions in another dimension. For example, making a highway less risky by straightening curves may encourage road users to drive faster. This is not an argument that regulatory action should not be undertaken, but rather that the consequences of regulatory action may be overstated.

There is a tension between expressing the minimum as a specification standard or a performance standard. In many ways, a performance standard that specifies the required safety outcomes best addresses market failures and does so without the regulator micromanaging the production of safety. In practice, standards are often expressed as minimum technical specifications for user conduct and vehicles, as these form bright line rules for the determination of compliance and the assessment of penalties.

A regulatory approach requires a legislative framework to enact regulations, and then an enforcement strategy by the police, government inspectors, and the courts. In setting a minimum standard, there is a risk that the standard is either set too low or unreasonably high. Economists have long used a technique called benefit–cost analysis to help determine the level at which the standard should be set. This is combined with quantitative risk assessment techniques that attempt to estimate the effects of actions on the number and severity of crashes. The costs of the regulation are compared with the expected benefits. Purely financial costs and benefits are combined with monetarized equivalents of costs such as longer journey times and benefits such as reduced injuries ([Jones-Lee and Spackman, 2013](#)).

Regulations are meaningless unless they are enforced. Forming a strategy for regulatory enforcement is not trivial. For example, there is a trade-off between the probability of detecting a violation of the regulations and the size of the resulting penalty. There can either be a high probability of detection and a low penalty, or a low probability of detection and a high penalty. Users might also be rewarded for good conduct in addition to penalizing poor conduct. It is an empirical matter as to whether enforcement is effective in changing behavior and preventing recidivism.

Closing Comments

Safety is an important and contentious aspect of transportation. The annual number of fatalities and serious injuries is large. So is the amount of property damage and the potential harm to the environment. The provision of safety is riddled with cognitive and market imperfections. Corrective legal and regulatory interventions are longstanding and continue to evolve. Despite substantial improvements in the past half-century, society continues to demand that more should be done to further reduce the risks.

References

- Blomquist, G.C., 1988. Regulation of Motor Vehicle and Highway Safety. Kluwer, Boston, MA.
- Dionne, G. (Ed.), 2014. Handbook of Insurance. second ed. Springer, New York, NY.
- Elvik, R., Høye, A., Vaa, T., Sørensen, M. (Eds.), 2009. The Handbook of Road Safety Measures. second ed. Emerald, Bingley, UK.
- Evans, L., 2004. Traffic Safety. Science Serving Society, Bloomfield Hills, MI.
- Haddon Jr., W., 1972. A logical framework for categorizing highway safety phenomena and activity. J. Trauma 12, 193–207.
- Jones-Lee, M., Spackman, M., 2013. Development of road and rail transport safety valuation in the United Kingdom. Res. Transp. Econ. Econ. Transp. Safety 43, 23–40.
- Kristiansen, S., 2005. Maritime Transportation: Safety Management and Risk Analysis. Elsevier, Amsterdam.
- Lamm, R., Psarianos, B., Mailaender, T., 1999. Highway Design and Traffic Safety Handbook. McGraw-Hill, New York, NY.
- Maurino, D., Reason, J., Johnson, N., Lee, R.B., 1995. Beyond Aviation Human Factors: Safety in High Technology Systems. Ashgate, Aldershot, UK.
- Savage, I., 1998. The Economics of Railroad Safety. Kluwer Academic, Boston, MA.
- Shavell, S., 2004. Foundations of Economic Analysis of Law. Harvard University Press, Cambridge, MA.

Cost Overruns of Transportation Infrastructure Projects

James Odeck, Morten Welde, Department of Civil and Environmental Engineering, NTNU—Norwegian University of Science and Technology, Trondheim, Norway

© 2021 Elsevier Ltd. All rights reserved.

Introduction	483
The Concept of Cost Overrun	483
What are the Problems of Cost Overruns?	484
Why Cost Overruns Occur	484
Measurement of Cost Overruns	485
The Prevalence of Cost Overruns Worldwide	486
Concluding Remarks—Potential Ways of Avoiding Cost Overruns	489
References	489

Introduction

Cost overrun of transportation infrastructure projects is a topic that has attracted the attention of academics, decision makers, the media, and the public at large. For instance, in the last four decades, a plethora of scientific papers and audit reports have been written; most of them attest to large overruns (for a review of the literature, see [Odeck, 2019](#)). The media has reported hundreds of headline reports of large overruns. Interestingly, the interest in cost overruns by their stakeholders seem to be different. The academics seem to be interested in measuring the magnitudes of and finding explanations for overruns, while decision makers are more interested in finding ways to reduce the size of overruns. However, the public regards the large magnitudes of overruns as a dishonest practice, where the taxpayers' money is misused by the decision makers.

Although cost overrun is frequently addressed in the literature, there appears to be no common understanding of its basic concepts available in a single source. Such a common understanding, if available, may help in finding ways to eradicate overruns. As matter of fact, disagreements in the literature have occurred regarding both the appropriate measure of overruns and what measures can be taken to reduce their magnitudes ([Flyvbjerg et al., 2018](#); [Love and Ahiaga-Dagbui, 2018](#)).

The aim of this article is to provide a comprehensive and common understanding of the concepts of cost overrun in transportation infrastructure projects. We do so by suggesting answers to the following questions: (1) What is a cost overrun of a transportation project? (2) What problems do overruns cause for stakeholders? (3) How should overruns be measured? (4) How common and prevalent are overruns worldwide? and (5) What are the potential ways of avoiding overruns? The rationale of this article is to provide a brief and comprehensive overview of these important issues.

The remaining part of the article is organized as follows. In the next section, we define the concept of cost overrun, why overruns arise and the problems that overruns cause for stakeholders. In the third section, we describe the different alternative measures of cost overrun and suggest when these measures are useful to consider. The fourth section addresses how prevalent cost overruns are worldwide. The fifth section provides some concluding remarks and discusses potential ways of reducing overruns.

The Concept of Cost Overrun

When addressing issues surrounding cost overrun, a clear and a common definition of the concept of cost overrun must be provided because it can be understood and measured differently, depending on how it is defined. Cost overrun is a term often used in the context of any construction project when the actual cost is greater than the budgeted cost. It is thus a measure of the difference between the actual cost and some budgeted/estimated cost, given that the difference is positive. Consequently, if the difference is negative, then "cost underrun" has occurred, and if the difference is zero, then the actual cost exactly matches the budgeted/estimated cost, and neither overrun nor underrun occurs.

Note that if a cost overrun occurs and a budget was used as the base of comparison, then the cost overrun is sometimes referred to as a "budget increase," "cost increase," or "cost growth." Cost overruns should, however, be distinguished from another related term, "cost escalation," which is used to express an anticipated growth in the budgeted cost due to factors such as inflation.

So far, we have defined cost overrun as only the difference between the actual cost and some budgeted/estimated cost in absolute terms. We revert later to its different measures, for example, relative measures, and what those measures tell and which estimates/budgets should be compared to actual costs given that there are often many estimates/budgets throughout a project's planning process. Thus in the next two subsections, we first discuss (1) the problems of cost overruns and (2) why overruns occur.

What are the Problems of Cost Overruns?

Given the definition of cost overrun earlier, that is, that it is the excess of the actual cost over the estimated/budgeted cost, a natural question to address is what problems it causes for the decision makers who sanction projects, the public, transport system users, and the practitioners who make cost estimates. Although the magnitudes of cost overruns of transportation projects have been attested to be prevalent and, in some instances, exceptionally large, the real problems that they cause for stakeholders are less frequently addressed.

Large cost overruns are a problem for decision makers who, based on the cost estimates provided to them by estimators (i.e., practitioners) and given their limited budgets, rely on those estimates to allocate funds for different projects. The immediate problems that arise for decision makers if cost overruns are large and prevalent are therefore as follows: (1) they burden decision makers with unexpected expenses worth hundreds of millions euros/dollars, (2) they put the financial viability of projects at risk, where projects may be stopped and/or their implementation extended and/or delayed, which *ceteris paribus* implies that transportation infrastructure users will wait longer to realize the benefits, (3) the decision makers may be forced to crowd-out funds that could have been used for other more needy investments, and due to (1), (2), and (3) above, the decision-making process may lose legitimacy and trust in the eyes of the wider public. A fourth problem of cost overrun for decision makers, as mostly observed by transportation economists/planners, is that overruns may lead to the incorrect choice/ranking of projects against other projects competing for limited government funds. This occurs because transportation economists correctly argue that projects should be chosen according to how high their benefit-cost ratios (BCR) are, that is, the ratio of a project's net benefits to its financial costs funded by a government budget. A higher BCR results in a project being more beneficial to the society. In this respect, the problem of cost overrun is that it may initially depict a project as having a high BCR and, hence, cause it to be chosen, while if overruns (when correct estimates are provided) are accounted for, the project has a low BCR, that is, the value of the numerator should have been higher, leading to a lower BCR; therefore, the project should not have been chosen over others or not funded at all given that funds are limited. Hence, the decision makers fourth problem with cost overruns is that it misleads them into make incorrect decisions, and less beneficial projects may be chosen over those that are more beneficial. It should, however, be noted that proponents of benefit-cost analyses (BCA) with BCR as a ranking criterion do not argue that it should be the only base for decision making. They actually argue and accept that there may be other factors not captured in a BCA such as nonmonetized and distributional impacts that may matter for decision makings. What they presuppose, however, is that the use of CBA would increase the benefits for tax payers if it had some influence on the projects ranking/selection.

Note, however, that large cost underruns are also a problem for decision makers. Had the correct estimates been provided, overestimated funds could have been used to realize other profitable projects at a much earlier stage so that users of facilities would benefit much earlier. This problem is, however, not as critical as the aforementioned ones.

For the practitioners, that is, engineers and economists who make the cost estimates, the real problem with cost overruns is that these practitioners lose their credibility as experts on cost estimation.

The problem of cost overrun for transport users and the public at large has been touched upon earlier. In cases of large and frequent overruns, the public will have a perception of misuse of public funds and, consequently, may punish the decision makers, who generally are politicians, by voting them out in the next election. For transport system users, cost overruns are a serious problem in terms not being able to use the transport system in a timely manner as planned, and projects may be postponed or not built at all because the planned funds are redirected to overrun projects. Since most transport infrastructure projects are built as a means of reducing transportation costs for system users, overruns imply that system users continue incurring greater transportation costs than necessary.

The problems of cost overruns addressed earlier, if added together, should verify why academics, the media and decision makers are wary of cost overruns in the transportation sector worldwide. It should be reminded here, as previously explained that cost underruns are also a problem, although not as grave as overruns.

Why Cost Overruns Occur

The next question worth addressing, which should succeed the former question addressed earlier, is why do cost overruns occur given that they cause such problems? A plethora of papers in the academic literature have sought an explanation for why cost overruns occur, but no one conclusion has been derived; instead, several explanations exist, some of which are true, while others are more of speculations than facts. Further, we address the most frequently used arguments in the literature for why cost overruns occur.

The most frequently stated reasons for cost overruns in the transportation literature, which can be classified in three categories according to Flyvbjerg et al. (2002), but which we expand on and explain, are as follows: (1) psychological, (2) political economic, and (3) technical/economic. The psychological causes of observed cost overruns are explained as optimism bias or man's inherent optimism, that is, those who make cost estimates are optimistically biased, a phenomenon known to exist among human beings regarding future outturns of their present undertakings. In relation to infrastructure projects, this means that cost estimators tend to think highly of themselves in that their estimates are at less risk of incurring overruns than others.

The political-economic reason for observed cost overruns is that, due to pressure from interest groups, for example, politicians, project promoters, contractors, etc., those who make cost estimates are driven to deliberately underestimate cost to make the estimates more acceptable to interest groups or decision makers.

Technical/economic causes of cost overruns involve information that is not available at the time of cost estimation, such as price escalation of unit variables, changes in the design and scope of a project (change order), ever emerging new regulations (e.g., environmental demands that are costly), and pure estimation mistakes (e.g., failure to account for certain and already available information provided by engineers such as transportation costs associated with removing and depositing volumes of masses removed from tunnels/roadways or accounting for masses and their transportation costs required to fill the dugs to ascertain frost free roadway surfaces).

Another, but less cited explanation of cost overrun, is that of Eliasson and Fosgerau (2013) who claimed that *selection bias* may occur whenever ex ante predictions are related to the decisions on whether to implement a project or not. However, they argued that even if ex ante estimates are very uncertain the selected projects will turn out to perform better on average than random projects. Moreover, as they rightly pointed out, one cannot conclude based on the observation of ex post bias that there must have been bias in the prediction's ex ante. The reason for which is that the unselected projects may have been equally biased.

Although the aforementioned described reasons may be the causes of cost overruns, no clear-cut empirical evidence; a proof without doubt, has been provided in the literature to prove that reasons (1) and (2) exist. Hence, we contend that they are unsubstantiated but are possible reasons for overruns. Reason (3), however, has been proven in the literature to apply when certain information is not available at the time of cost estimation, thereby leading to cost estimates being lower than actual outturns. We discuss later how cost overruns can be avoided while reflecting on the subjective causes described earlier. However, before that, appropriate measures of the magnitudes of overruns must be addressed as well as how large and prevalent overruns are in transportation sectors worldwide.

Measurement of Cost Overruns

As previously defined, cost overrun is a term often used in the context of any construction project when the actual cost does not meet/ is not meeting the budgeted/estimated cost. Its measurement is therefore the magnitude by which the actual cost is larger/smaller than the estimated/budgeted cost.

Formally and under the definitions previously given, if Y_i represents the actual cost upon project i 's completion and if F_i represents the estimated/budgeted cost for the same project, then the cost overrun for project i (CO_i) is defined as follows:

$$CO_i = Y_i - F_i. \quad (1)$$

Hence, Eq. (1) is an absolute measure of cost overruns for example, in terms of dollars. If $CO_i = 0$, then the budgeted/estimated cost was accurate; if $CO_i < 0$, then an underrun has occurred that is, the budgeted/estimated cost was greater than actual cost; and if $CO_i > 0$, an overrun has occurred because budgeted/estimated cost because the budgeted/estimated cost was less than the actual costs.

Similarly, the relative or percentage cost overrun (PCO_i), which is useful when comparing overruns across projects is defined as:

$$PCO_i = \left(\frac{Y_i - F_i}{F_i} \right) \times 100. \quad (2)$$

Note that the actual cost could be used in the denominator such that the percentage overrun is measured as a percentage of the actual cost. We, however, propose that the percentage overrun should be measured as a percentage of the estimated/forecasted/ budgeted cost, that is, how large the overrun is relative to the estimated/forecasted/budgeted cost. Our suggestions here conform to the practices in the literature of cost overruns of transportation projects, although there are a few examples where the actual cost has been used in the denominator.

Finally, when comparing the cost overruns across different groups of projects, different mean PCO , that is, ($MPCO$) measures can be used, each giving a different result as follows:

$$MPCO = \frac{1}{n} \sum_{i=1}^n PCO_i, \quad (3)$$

$$MAPCO = \frac{1}{n} \sum_{i=1}^n |PCO_i|, \quad (4)$$

$$MPCOS^2 = \frac{1}{n} \sum_{i=1}^n (PCO_i)^2, \quad (5)$$

where $MPCO$ is the mean percentage overrun, $MAPCO$ is the mean absolute percentage overrun, and $MPCO^2$ is the mean absolute overrun squared. These measures are all useful to express changes in mean (average) overruns/underruns in different ways depending on the desired result as follows. The $MPCO$ tends to be small because the negative and positive values offset each other. It is useful as an indication of whether systematic under- or overestimation occurs across a set of projects. Furthermore, it is the most

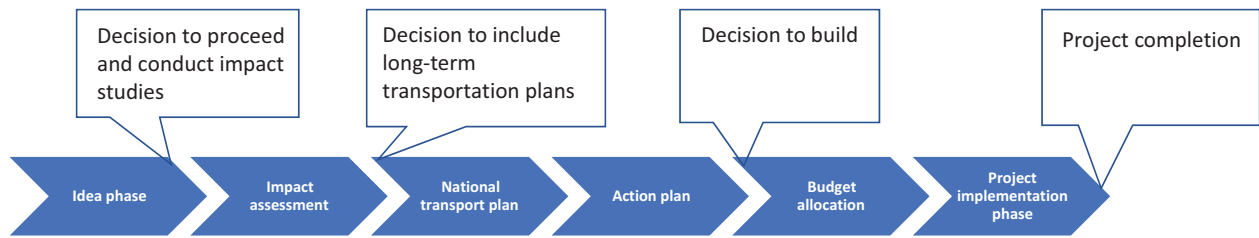


Figure 1 Different phases of planning.

commonly used measure of cost overruns in the literature. The *MAPCO*, however, assumes that all biases, irrespective of direction, are equally important to consider; hence, it uses the absolute values. Finally, because the mean *PCO* is squared, the *MPCO*² gives more weight to larger values than smaller values, but in other respects, it is similar to the *MAPCO*. Thus the *MPCO*² reflects the fact that large errors represent a much more serious problem than small errors, meaning that it is appropriate to focus primarily on larger errors to effectively reduce total errors.

Given the differences among the *MPCO*, *MAPCO*, and *MPCO*² discussed earlier, a question that still remains is which one among them is most useful to calculate/address according to different policy perspectives. The *MPCO* is most useful to consider if the policy is to infer the performance of a pool of projects as a group (or a portfolio of projects). The *MAPCO*, however, is most useful to consider if the objective is to determine the variation in performances across individual projects where overruns and underruns are assumed to be equally undesirable; it is an absolute measure. Finally, the *MPCO*² is relevant if one is specifically interested in magnifying the differences in over-/underruns between large and small projects. Some studies report all of these measures.

Another important issue that must be considered when measuring cost over-/underruns is which cost estimates to compare with actual costs at project completion. As often is the case in infrastructure planning, there are many planning phases, and different cost estimates are made in each planning phase; hence, one must be careful about estimates used to derive the magnitudes of under-/overruns. Second, it might be informative to compare estimates made at the different planning phases to obtain a measure of over-/underruns in cost estimates from one planning stage to another, assuming that the most recent estimate is more correct than prior ones. The different phases of transportation planning can generally be summarized as in [Fig. 1](#).

[Fig. 1](#) illustrates that there are different phases of planning in which cost estimates are made. Cost estimates are usually, but not always, developed throughout the different stages of project development. For example, [Cantarelli et al. \(2010\)](#) found that the cost increase from the first plans to the formal decision to build in two large Dutch rail projects was several hundred percent, but that the overruns during construction was relatively small, around 10%. Therefore, one must be careful about which estimates are being compared against each other and, importantly, compared to the actual costs; the latter comparison is the real measure of over-/underruns per the definition of over-/underruns.

[Table 1](#) explains the usefulness of measuring over-/underruns by comparing estimates made at different phases of planning against each other and against actual costs. It can be seen, for instance, that comparing an estimation made at the budget allocation phase against the actual cost is most useful as a measure of total over-/underrun as it compares the actual cost to the estimate made at the time the decision to build was made. However, comparing estimates made at the Idea Phase to those made at the National transport Plan Phase is not very meaningful. Furthermore, comparing estimates at the Idea Phase (front-end phase) to actual costs has, in the literature, been useful in the sense that it provides more information about the extent of underestimation at the front end of projects ([Welde and Odeck, 2017](#)).

The Prevalence of Cost Overruns Worldwide

Most of the literature on cost over-/underruns unanimously concludes that overruns are more prevalent than underruns. Note that this is not the same as some authors tend to conclude, for example, that overruns are always the case. We warn here that some studies have shown projects with underruns, even though the averages across samples tend to be overruns, hence the notion that overruns are more prevalent than underruns.

Next, an issue to consider when assessing the prevalence of overruns is how serious those overruns are. Normally, budgets are allocated in terms of total cost for each project. Total costs include the base cost and contingency given within a confidence interval, for example, P85 or even as plus minus x%. Thus prevalent overruns become a serious problem when overruns go beyond the confidence intervals provided at the time of budget allocations. When this occurs, projects may be stopped, leading to infrastructure users not profiting in time as promised; funds that could have been used for other profitable infrastructure and/or other sectors of the economy are out-crowded and diverted to overrun projects, and decision makers and infrastructure users alike feel deceived by those who make the estimates. These are the reasons why the prevalence of cost overruns beyond budget tolerance has caused concerns in the media and among academics worldwide.

Note that the intolerable and/or large magnitudes of cost overruns differ a great deal across countries and continents. For instance, the magnitudes of cost overruns in Norway, on average, are within tolerable limits, that is, more or less within the estimated confidence intervals ([Odeck, 2004, 2014](#)). Those studies showed that smaller projects have higher percentage overruns than larger

Table 1 The usefulness of comparing estimates made at different planning phases

Estimates at different phases of the infrastructure planning process							
		Idea Phase (front–end phase)	Impact assessment phase	National Transport Plan Phase	Action plan phase	Budget allocation phase	Project completion phase
Estimates at different phases of the infra- structure planning process	Idea Phase (front–end phase)	—	Useful for gauging how costs escalates from the Idea Phase when very little is known to the impact assessment phase when much more is known. Helps to stop a project in time.	Not very useful because reesti- mates from impact assess- ment are available.	Of little use because reestimates from impact assessment or National Transport Plan Phases are already available.	Of little use because reestimates must be conducted at the action plan level.	Useful: Estimates at the Idea Phase (front–end) are regarded to be a major catalyzer for the realization of projects. Results give insight/information that can be used in the Idea Phase of future pro- jects in estimating costs.
	Impact assess- ment phase	—	—	Very useful: Results/informa- tion can be used to revise the BCA at the National Transport Plan Phase.	Of some use: Can lead to reprioritization of projects at the action plan phase.	Useful: Results can be used to revise the BCA and, hence, lead to dif- ferent allocations and/or prioritization.	Very useful: Results on how often actual costs at the project completion phase are different as compared to estimates at the impact assessment phase can be used to revise methods of cost estimations at the impact assessment phase, which other- wise may improve the estimates of the overall socioeconomic profitability of projects.
	National Transport Plan Phase	—	—	—	Useful: Results can lead to reprioritization and changes in the action plan because actions plans are sometimes based on results from the national transport plan estimates in cases where no reestima- tions are made.	Useful: Results can lead to reallocation of funds	Useful: Indicates how prioritizations at the National Transport Plan Phase are sensi- tive to cost estimates.
	Action plan phase	—	—	—	—	Useful: Can impact budget allocations.	Useful: May indicate why one should not rely on estimates at the action plan phase and that estimates at the budget allocation phase are the most relevant to consider.
	Budget allo- cation phase	—	—	—	—	—	Very useful: The final decision to build is made at the budget allocation phase. This is the typical measure of over-/underruns in the literature.

Source: Created by the authors.

Table 2 Summary of the prevalence of cost overruns in the literature

<i>Study number</i>	<i>Publication year</i>	<i>Journal/Conference publication (1) or report (0)</i>	<i>Countries</i>	<i>Total number of projects</i>	<i>Mean percentage cost overrun (MPCO)</i>
1	2006	1	France/United Kingdom	1	99.00
2	2007	0	Australia	1	14.70
3	2007	1	Australia	1	1.60
4	2004	0	United States	2668	5.00
5	2012	1	The Netherlands	78	16.53
6	2005	1	France	5	13.30
7	2010	1	European countries	6	51.00
8	2006	0	Australia	231	28.80
9	2006	0	United States	16	30.00
10	2007	0	United States	3130	9.00
11	2003	0	United States	21	40.00
12	2003	1	Worldwide	258	27.60
13	2009	1	Zambia	8	69.00
14	2008	1	South Korea	154	29.50
15	2012	1	Australia	58	13.28
16	2011	1	Sweden	167	14.99
17	2011	1	Slovenia	36	19.00
18	1973	1	United States	66	33.21
19	1990	1	India	33	69.85
20	1999	1	European countries	9	191.00
21	1995	1	Norway	12	5.00
22	2004	1	Norway	620	8.00
23	1992	1	North	8	61.00
24	1990	0	North	0	52.00
25	2006	1	Canada	50	82.00
26	2009	0	European countries	48	21.39
27	2011	0	Sweden	28	55.00
28	1994	0	Sweden	15	53.80
29	2008		Thailand/The Philippines	129	-0.13
30	2002	0	Sweden	47	10.17
31	2010	1	India	279	55.50
32	1997	1	Denmark	7	14.00
33	2007	0	United Kingdom	56	10.28
34	2002	0	United States	1377	3.98
35	2002	0	United States	670	3.66
36	2002	0	United States	2196	8.13
37	2002	0	United States	2917	4.61
38	1991	0	United States	432	5.83
	1998	0	United States	865	10.00
39	2012	1	European countries	94	16.20
40	2014	1	Australia	73	23.00
41	2010	1	Australia	7	40.50
42	2014	1	African countries	53	12.67
43	2013	1	India	145	17.50
44	2012	1	India	14	10.26
45	2013	1	United States	236	36.22
46	2015	1	Jordan	9	214.00
47	2013	1	Tanzania	7	44.00
48	2015	0	United States	25	16.80
Mean	2003	0.54	—	354	34.12
Min.	1973	0.00	—	0	-0.13
Max.	2012	1.00	—	3130	214.00
Standard deviation	8	—	—	852	36.59
Weighted mean overrun		—	—	—	40.31
Weighted min.		—	—	—	0.00
Weighted max.		—	—	—	564.14
Weighted standard deviation		—	—	—	124.08

Source: Sampled by the author and parts was used in Odeck, J., 2019. Variation in cost overruns of transportation projects: an econometric meta-regression analysis of studies reported in the literature. *Transportation* 46, 1345–1368.

projects. Other European countries have, however, shown larger overruns. For instance, [Lundberg et al. \(2011\)](#), in the case of Sweden, found overruns in the range 10%–20%.

We revert to why there may be differences between countries in the succeeding section, where we address how to avoid overruns.

To complete this section of the article, we provide in [Table 1](#) a summary of some studies performed worldwide since 2003 that have examined the magnitudes of cost overruns among infrastructure projects. We included both scientific papers published in journals and nonpublished reports to avoid publication biases, that is, it may be that papers are published because they give certain results, and those that do not give those results are rejected/do not get published and, hence, are unpublished. The summary results in terms of means and other measures clearly show that overruns are large worldwide, which *ceteris paribus* should lead to concern, as frequently observed by academics and the media. The weighted mean overrun (weighted by number of projects in each paper/report) across all of the studies is 40%, and the standard deviation is large. Taken together, this information supports that overruns are frequent and large, and that large variation exists worldwide ([Table 2](#)).

Concluding Remarks—Potential Ways of Avoiding Cost Overruns

We addressed the explanations of what overruns are, how they occur, the problems that they cause for decision makers, how the over-/underruns should be estimated, and how prevalent overruns of infrastructure projects are worldwide. We have attempted to be concise and not engage in typical debates between opponents of the overrun issue, such as overruns are not as large as reported, the claimed causes of overruns are not true, and the method used to assess overruns is not correct. We have instead attempted to inform, as far as possible and without being partial, as is evident in each of the section of the article.

One of the clearest conclusions that we can draw from our discussions in this article is that cost overruns remain a substantial problem for decision makers and facilities users alike and that overruns are much more prevalent than the opposite, that is, underruns. As to the causes, many different explanations exist in the literature, and there are disputes as to which ones among them are the correct explanations for the causes of overruns. A further complicating issue is that many of the explanations given are not based on empirical evidence, that is, using real observed data to prove the claim. Nonetheless, we believe that the cause of overruns is a complex mixture of the reasons that have been provided in the literature.

A final question that should be attempted to be answered in a paper such as this is how can overruns be reduced and/or what are the potential strategies for reducing them for transportation infrastructure projects? Our immediate response to this question is that the strategies depend on the explanation for the causes/prevalence of overruns that have been suggested in the literature. While building on existing literature, [Siemiatycki, \(2009\)](#) offered five approaches to control for potential overruns, to which we agree but adjust, that are as follows: (1) enhance performance monitoring throughout the construction period such that total overruns can be reduced, (2) enhance accountability and responsibility, where project managers are reminded of the consequences of overruns, (3) enhance the management capabilities of construction engineers, such as being cost conscious, and the management of complex projects, (4) apply state-of-the-art cost estimation techniques, that is, those that have been proven to work with a higher degree of precision, and (5) increase completeness and rigor of early plans, where early plans are scrutinized with respect to cost.

It should be noted, however, that the five approaches earlier are complementary and not mutually exclusive. However, it is exceedingly difficult single out any of them, as the most important empirical evidence has proven that external quality assurance of estimates is a highly effective measure, see for instance [Odeck et al. \(2015\)](#). The reasons for this are that road builders/providers have a tendency to provide estimates that are too optimistic. Quality assurance implies that the estimates made by the authorities are subjected to external scrutiny to ensure that quality standards of cost estimates are being met and to detect and correct any deviations.

References

- Cantarelli, C., Flyvbjerg, B., van Wee, B., Molin, E.J.E., 2010. Lock-in and its influence on the project performance of large-scale transportation infrastructure projects: investigating the way in which lock-in can emerge and affect cost overruns. *Environ. Plan. B Plan. Des.* 37, 792–807.
- Flyvbjerg, B., Skamris Holm, M., Buhl, S., 2002. Underestimation of costs in public works projects. Error or lie? *J. Am. Plan. Assoc.* 68, 279–295.
- Flyvbjerg, B., Ansar, A., Budzier, A., Buhl, S., Cantarelli, C., Garbuio, M., Glenting, C., Skamris Holm, M., Lovallo, D., Lunn, D., Molin, E., Rønneest, A., Steward, A., van Wee, B., 2018. Five things you should know about cost overrun. *Transp. Res. A* 118, 174–190.
- Eliasson, J., Fosgerau, M., 2013. Cost overruns and demand shortfalls – Deception or selection? *Transp. Res. B* 57, 105–113.
- Love, P.E.D., Ahlaga-Dagbui, D.D., 2018. Debunking fake news in a post truth era: the plausible untruths of cost underestimation in transport infrastructure projects. *Transp. Res. A* 113, 357–368.
- Lundberg, M., Jenpanitsub, A., Pyddoke, R., 2011. Cost overruns in Swedish transport projects. CTS Working Paper 2011:11.
- Odeck, J., 2004. Cost overrun in road construction – What are their sizes and determinants? *Trans. Policy* 11, 43–53.
- Odeck, 2014. Do reforms reduce the magnitudes of cost overruns in road projects? Statistical evidence from Norway. *Transp. Res. A Policy Pract.* 65, 68–79.
- Odeck, J., 2019. Variation in cost overruns of transportation projects: an econometric meta-regression analysis of studies reported in the literature. *Transportation* 46, 1345–1368.
- Odeck, J., Welde, M., Volden, G.H., 2015. The impact of external quality assurance of costs estimates on cost overruns: empirical evidence from the Norwegian road sector. *Eur. J. Trans. Infrastruct. Res.* 15 (3), 286–303.
- Siemiatycki, M., 2009. Academics and auditors comparing perspectives on transportation project cost overruns. *J. Plan. Edu. Res.* 29 (2), 142–156.
- Welde, M., Odeck, J., 2017. Cost escalations in the front-end of projects – empirical evidence from Norwegian road projects. *Trans. Rev.* 35, 612–630.

The Downs–Thomson Paradox

Joel P. Franklin, KTH Royal Institute of Technology, Stockholm, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	490
Forces Behind the Downs–Thomson Paradox	490
The Fundamental Diagram	491
Induced Traffic	491
Mogridge's Hypothesis	491
Public Transport Attractiveness	491
The Shifting Equilibrium	492
A Numerical Example	492
Evidence of the Downs–Thomson Paradox	493
Laboratory Experiments	493
Component Effects	493
Cross-Elasticities	494
Induced Travel	494
Positive Externalities in Public Transport	494
Do the Downs–Thomson Conditions Really Apply?	494
Conclusions	494
References	494
Further Reading	495

Introduction

Ask a random stranger about transportation, and they will almost invariably offer a simple intuition about capacity analysis. If we knew how many cars and trucks would use a facility, then we would only need to design it to accommodate that. If more cars and trucks started to appear, causing congestion, then we could just increase capacity and the congestion will dissipate. The problem is that in most major cities, this intuition very often incorrect, and this gives rise to one of the most persistent frustrations in the transportation planning process: new roads and new lanes simply do not seem to make a dent in congestion in the long run.

As if this were not bad enough, there are even some cases where capacity has been increased but soon after, congestion not just remains but escalates. The reasons for this can be elusive, and they are often attributed to background trends of population growth, rising wealth, or perhaps even a general increase in travelers' tolerance for sitting in traffic—possibly because cars are getting more comfortable, and hands-free mobile phones help drivers make better use of driving time.

It would be comforting to think that our original intuition was correct, and that some or all of these other factors confounded the real outcome. But in fact, there is an alternate intuition that could explain this surprising response, at least in some contexts where a viable public transport alternative exists. This intuition is, essentially, that roadways and public transport respond very differently to additional use: roadways experience negative externalities in the form of congestion, while public transport experiences positive externalities in the form of improved service. Because of these externalities, improving the roadway can actually lead to a less efficient equilibrium as the natural efficiencies of public transport are undermined by the attractiveness of the road. A new equilibrium emerges that is more costly, on average, than before.

The intuition originally comes from [Downs \(1962\)](#) and came to be known alternatively as “Downs's Law of Peak-Hour Traffic Congestion” or “The Iron Law of Congestion.” Essentially the same intuition was formulated independently by [Thomson \(1972\)](#). Downs's Law is probably best known for its explanation of induced traffic, but in fact the paper of that title considers three separate cases. The third case, which [Mogridge et al. \(1987\)](#) later termed the “Downs–Thomson Paradox” (DTP), imagines that the roadway coexists with a segregated public transport alternative, as shown in [Fig. 1](#), and this turns out to be essential.

Forces Behind the Downs–Thomson Paradox

Let us be more specific about the context. Consider a dense urban area with heavy day-to-day traffic. Travelers have an alternative to driving: they can take a public transport mode that operates separately from the roadway. We can then imagine the sequence of forces that lead to the DTP shown in [Fig. 2](#), where the plus-signs and minus-signs indicate positive and negative correlations, respectively, between cause and effect.

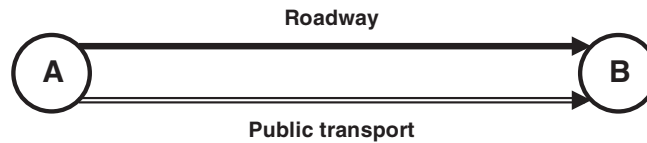


Figure 1 The abstracted transport network.

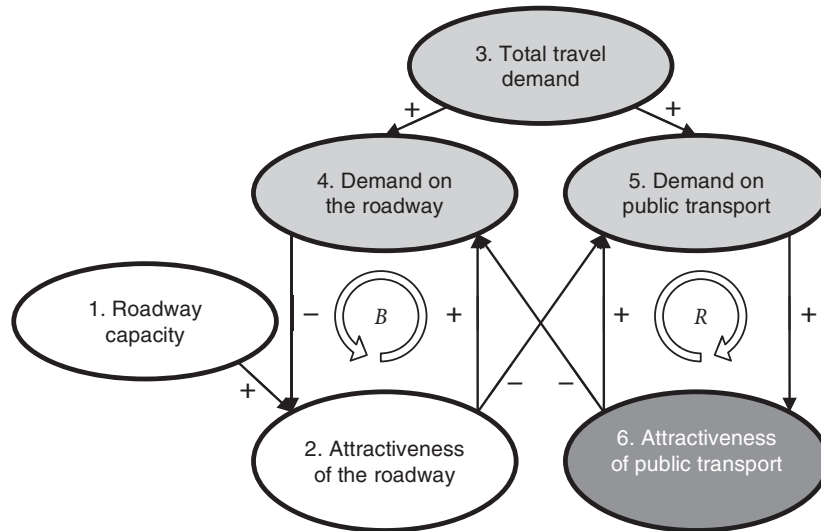


Figure 2 Causal effects behind the Downs-Thomson Paradox.

The Fundamental Diagram

Start with the white ovals: roadway capacity (1) affects roadway attractiveness (2). This encapsulates the naïve intuition about roadway expansion, and it predicts a positive causality such that higher capacity leads to the roadway being more attractive (i.e., higher speeds). This relationship rests on some theory of congestion, of which the simplest is the so-called fundamental diagram of traffic flow. In the fundamental diagram, speeds are inversely proportional to traffic volumes and directly proportional to capacity. Holding volumes constant, only the latter relationship is important.

Induced Traffic

The naïve model may be true all else being equal, but in many cases all else is not equal. If we add the light gray ovals, we no longer have fixed traffic volume, but rather a fixed total demand (3), which splits into roadway and public transport demand. So if capacity increases and this makes the road more attractive, then some additional roadway demand (4) will emerge. The loop between (2) and (4) alternates between positive and negative, creating a balancing cyclical effect (indicated by the capital “B” in the figure). In consequence, the benefits of a capacity increase will tend to be negated. This explains how induced travel can diminish the benefits of new capacity.

Mogridge's Hypothesis

Now we add demand for public transport (5). Up to now we have not speculated on the source of the new automobile traffic demand above, but according to Mogridge's hypothesis (Mogridge and Holden, 1987), as described in Holden (1989), the dynamics of automobile demand can be explained by a corollary of Wardrop's principle. Just as Wardrop's principle proposes that drivers individually vary routes until all chosen routes have the same travel time, so too could travelers in general vary their choices of mode until either the travel times by automobile and public transport are equalized or one of them is no longer chosen at all. Mogridge (1997) later reformulated his hypothesis in terms of equalized generalized cost, rather than equalized travel time alone, thereby allowing for differences in time, comfort, and out-of-pocket costs across modes. This explains cases where an equilibrium could still be reached even if observed travel times are different between automobile and public transport.

Public Transport Attractiveness

Finally, we add the single dark gray oval, indicating the attractiveness of public transport (6). Here we make an important assumption about public transport operations: that it is a scheduled service with discrete, regular departures, and that the frequency

of these departures will be adjusted to accommodate demand. In other words, if demand for public transport increases, then the frequency of departures will increase. The shorter waiting times improve the attractiveness of the service, leading to yet more demand. This phenomenon, known as the “Mohring Effect” (see Chapter 49, Mohring Effect), is central to the scale economies argued in Mohring’s (1972) classical model, and it forms an essential component of Downs’s (1962) original argument. Later variants have supposed that rather than change departure headways, per-trip fares might instead be reduced. In either case, public transport usage is assumed to have positive externalities on the service characteristics; this leads to a reinforcing cyclical effect (indicated by the capital “R” in the figure).

The Shifting Equilibrium

Sure, all the forces in Fig. 2 may exist, but what about this combination leads to an outcome where an improvement in the road worsens the situation for everyone? We can illustrate the forces above diagrammatically, as done by Mogridge (1997). In Fig. 3, the cost curves for auto and for public transport are plotted against the share of demand that chooses each option. Moving rightward along the horizontal axis, more travelers choose to travel by automobile; moving leftward, more take public transport.

There are two cost curves for auto, one before and one after improving the road, and one cost curve for public transport. Crucial to the Downs–Thomson Paradox are the principles that automobile costs increase with its mode share, while public transport costs decrease with its mode share and therefore increasing with auto mode share; hence all three curves are upward-sloping.

The two auto cost curves depict an improvement in the roadway, leading to lower costs for the same demand by automobile. This moves the equilibrium point from A to B. Were there only a fixed auto demand, this would certainly lead to a reduction in costs. But since in this case demand can choose between auto and public transport, we can see how the improvement leads to a higher auto demand, to the extent that the new average costs by auto are actually higher than before. Moreover, the remaining public transport passengers also see higher costs due to fewer riders. In other words, both modes are more costly than before.

A Numerical Example

A final illustration uses a numerical example to demonstrate how different road capacity levels could lead to different outcomes for both modes. Fig. 4 records the travel time and demand levels over varying levels of roadway design capacity. Similar illustrations can be found in earlier explanations such as Fig. 3 in Kitamura et al. (1999), Fig. 1 in Abraham and Hunt (2001), and Fig. 5 in Afimeimounga et al. (2005).

In this example, travelers have identical preferences, among which that they value travel time on either mode the same, and that is the only consideration. In other words, this model incorporates Mogridge’s hypothesis that when both modes are chosen, travel times must be equal between them. To give precision to the cost curves, the example employs a simple power-based curve to represent roadway congestion, whereby delay time is related to a ratio between demand and the road’s nominal capacity; and for public transport it assumes that vehicle headways are determined by demand. If travelers shift from public transport to automobile, then road congestion will increase while the public transport vehicles come less frequently, leading to longer average wait times.

Moving from left to right, we start with very low roadway capacity, so for the two modes to have equal travel times, only a very few travelers will drive. As roadway capacity increases, it draws more demand from public transport, to the point that travel times increase. This continues up to the point (in this case, a roadway capacity of somewhere between 600 and 650 travelers per hour) where public transport can no longer supply a service that is competitive with the automobile in travel time terms. Increasing

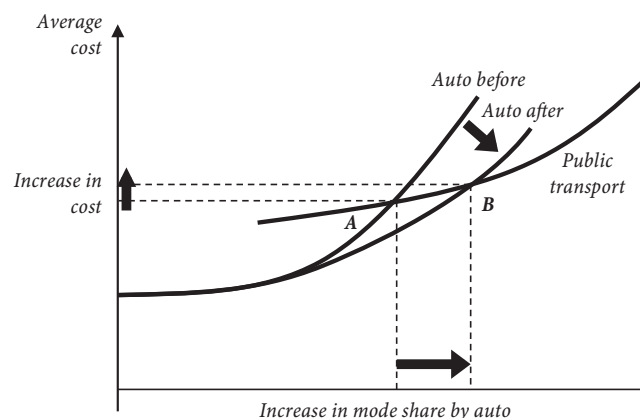


Figure 3 Diagrammatic explanation of the Downs–Thomson Paradox. Source: Based on Mogridge, M.J., 1997. The self-defeating nature of urban road capacity policy: a review of theories, disputes and available evidence. *Transp. Policy* 4, 5–23. [https://doi.org/10.1016/S0967-070X\(96\)00030-3](https://doi.org/10.1016/S0967-070X(96)00030-3).

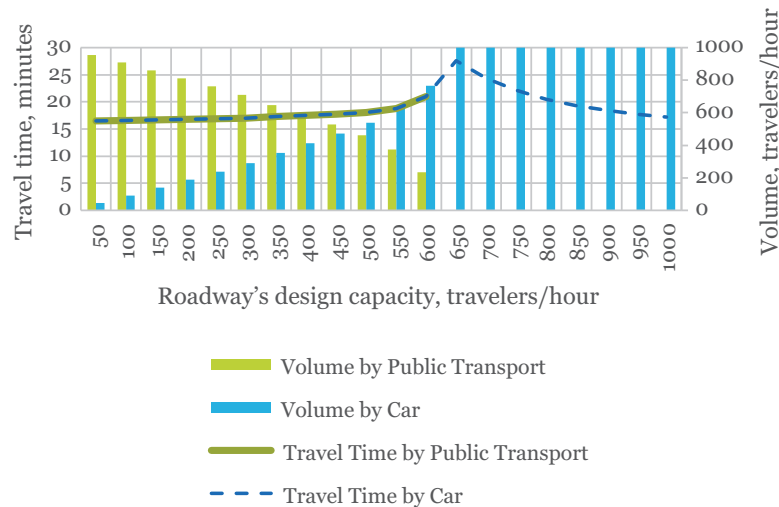


Figure 4 Numerical illustration of the Down–Thomson Paradox.

headways lead to increasing average public transport travel times that are higher than auto travel times, so no one chooses public transport any longer. At this point, there is a discontinuity in travel times: travel time for public transport is no longer relevant; and for the automobile, travel times jump as the last public transport travelers jump into their cars and there is no longer an equilibrium between modes. For the remaining road capacity levels, all 1000 travelers choose the automobile and the only question is how much congestion they endure as nominal capacity approaches 1000.

Evidence of the Downs–Thomson Paradox

The counterintuitive result of the Downs–Thomson Paradox is provocative, emphasizing the importance of viewing it in light of evidence. Unfortunately, an empirical confirmation or rejection of the DTP using real-world field data has so far eluded researchers. It is simply too difficult to isolate the relationship between roadway expansion and a deterioration in congestion. In the dense urban areas where the DTP is meant to apply, population growth, economic development, and other infrastructure changes will almost certainly confound any observations.

Instead, the evidence for the overall DTP effect has mostly come from analytical approaches, as well as a few experiments in laboratory environments. The basic analytical case was already set out in simple terms by [Mogridge et al. \(1987\)](#). [Abraham and Hunt \(2001\)](#) reformulated the problem in a discrete mode choice model, allowing some observations on when the paradox does and does not arise. [Kitamura et al. \(1999\)](#), using a cellular automata-based model, was able to reproduce the DTP, and [Afimeimounga et al. \(2005\)](#) did the same using a queuing model.

Laboratory Experiments

A few experiments in simulated settings have confirmed the DTP under controlled conditions. [Denant-Boemont and Hammiche \(2009\)](#) applied a game-theoretic approach and conducted a laboratory experiment with simulated setting, finding evidence of the DTP when increasing roadway capacity, but not when decreasing it. [Dechenaux et al. \(2014\)](#) examine two new facets: alternative pricing policies for the metro; and population size. Use game experiments. Confirm DTP for increasing capacity. Also, average cost pricing for the metro reduces road congestion.

Component Effects

In addition to the overall effect of counterproductive roadway expansion, each of the individual relationships in [Fig. 2](#) could also be tested. Some are uncontroversial: it is well understood that congestion can increase from either a reduction in capacity for fixed demand (1→2) or an increase in demand for fixed capacity (4→2), and the demand scaling relationships with modes (3→4 and 3→4) are straightforward. More interesting to examine are the following:

- Cross-elasticities of demand between public transport and roadways (6→4, 2→5)
- Induced travel demand on roadways and public transport (2→4 and 6→5)
- Positive externalities in public transport (5→6)

Cross-Elasticities

The closest efforts to examine the DTP using field-collected data appear to have probably been by [Mogridge \(1997, 1990\)](#), who examined records of highway and public transport expansion in the London Region along with historical measurements for operating speeds on both modes. The key findings were that despite substantial increases in roadway capacity over the studied time period, average speeds by both automobile and public transport steadily declined. According to [Mogridge \(1997\)](#), the parallel decline in speeds for the two modes supports the DTP, at least to the extent that there is an equilibrium process between the two modes that incorporates cross-elasticities of demand. Indeed, taking a generalized cost view, it is not actually necessary for travel times to be the same for an equilibrium process to exist, but rather it is sufficient that generalized costs are the same. However, as with most such time series data, the data do not support a robust examination of the causal relationship between capacity changes and travel speeds, since confounding factors such as population and wealth cannot easily be controlled for.

Another empirical analysis comparing mode choice to the relative travel times between auto and public transport ([Nss et al., 2001](#)) found a strong relationship. Still, this is only partial evidence of the DTP, supporting at least the notion that mode choice is subject to cross-elasticities, but it does not confirm the causal effect of capacity changes on travel times for either mode. A broader discussion of cross-elasticities can be found in Chapter 36, Cross-Elasticities Between Modes.

Induced Travel

A large literature exists on the drivers of travel demand (see Chapter 21, Demand for Passenger Transport) due to capacity increases, especially on roadways. Among these, it is notable that [Noland \(2001\)](#) made specific reference to the Downs–Thomson Paradox in setting the context for his findings. While Noland found strong evidence of induced demand in terms of vehicle-miles traveled, no results were given on whether travel time savings from the capacity increase were partially or fully offset by this induced demand. In sum, it appears that no real-world evidence so far fully confirms the Downs–Thomson Paradox.

Positive Externalities in Public Transport

An extensive literature exists on scale-economies of public transport, which extend far beyond the Mohring effect and indicate potential for both increasing and decreasing costs. Chapter 49, Mohring Effect presents a thorough discussion.

Do the Downs–Thomson Conditions Really Apply?

If the Downs–Thomson Paradox can only be confirmed in analytical studies, simulations, and laboratory experiments—all of which require some strong assumptions about the urban geography, the public transport operator, and prices—then it is worth asking whether these assumptions are realistic for real-world conditions, and what the limits are for the DTP to arise.

[Basso and Jara-Díaz \(2012\)](#) incorporated prices for both public transport and the roadway, finding that the scenario where the operator minimizes costs is not the same as the optimal scenario in socio-economic terms. [Bell and Wichensin \(2012\)](#) found that the DTP would not occur if a profit-maximizing operator does not pass the positive externalities of increased demand on to passengers. [Zhang et al. \(2014\)](#), when examining a variety of situations including different operator objectives and instruments, came to the same conclusion, but also showed that if the operator maximizes social utility, subject breakeven conditions, then the DTP does occur.

[Abraham and Hunt \(2001\)](#) also considered multiple operating policies, but with both tested policies (constant load-factor and minimum-cost), they found that the DTP would be most prominent when the transit market share is low. This contradicts the argument by [Mogridge et al. \(1987\)](#) and [Holden \(1989\)](#) that the DTP is most likely in areas with heavy transit usage.

Conclusions

The Downs–Thomson Paradox is provocative, but unfortunately there have never been any firmly confirmed sightings in the field. Moreover, the conditions of the Downs–Thomson Paradox are strict. That said, even if the required conditions are not always be present, they are certainly common enough in modern cities. Given the right conditions, the paradox has been shown to emerge in simulations, in laboratory experiments, and by analytical methods. This should be enough to give transportation planners pause: it simply cannot be assumed that a roadway expansion will lead to reduced congestion, or even that congestion will be the same after expansion. A careful examination is needed of the interaction between different transport modes and their relative economies-of-scale. Without this, there is a reasonable risk that scarce public resources are spent on infrastructure with no real social benefit.

References

- Abraham, J.E., Hunt, J.D., 2001. Transit System Management, Equilibrium Mode Split and the Downs-Thomson Paradox. Calgary, Canada. Department of Civil Engineering, University of Calgary (Preprint).
- Afimeimounga, H., Solomon, W., Ziedins, I., 2005. The Downs-Thomson Paradox: existence, uniqueness and stability of user equilibria. *Queueing Syst.* 49, 321–334, doi:10.1007/s11134-005-6970-0.

- Basso, L.J., Jara-Díaz, S.R., 2012. Integrating congestion pricing, transit subsidies and mode choice. *Transp. Res. Part Policy Pract.* 46, 890–900, doi:10.1016/j.tra.2012.02.013.
- Bell, M.G.H., Wichienso, M., 2012. Road use charging and inter-modal user equilibrium: the Downs-Thompson Paradox revisited. In: Inderwildi, O., King, S.D. (Eds.), *Energy, Transport, & the Environment: Addressing the Sustainable Mobility Paradigm*. Springer, London, pp. 373–383, doi:10.1007/978-1-4471-2717-8_20.
- Dechenaux, E., Mago, S.D., Razzolini, L., 2014. Traffic congestion: an experimental study of the Downs-Thomson paradox. *Exp. Econ.* 17, 461–487, doi:10.1007/s10683-013-9378-4.
- Denant-Boemont, L., Hammiche, S., 2009. Public Transit Capacity and User's Choice: An Experiment on Downs-Thomson Paradox (Open Archive in Humanities and Social Sciences No. halshs-00405501).
- Downs, A., 1962. The law of peak-hour expressway congestion. *Traffic Q.* 16, 393–409.
- Holden, D.J., 1989. Wardrop's third principle: urban traffic congestion and traffic policy. *J. Transp. Econ. Policy* 23, 239–262.
- Kitamura, R., Nakayama, S., Yamamoto, T., 1999. Self-reinforcing motorization: can travel demand management take us out of the social trap? *Transp. Policy* 6, 135–145, doi:10.1016/S0967-070X(99)00015-3.
- Mogridge, M.J., 1997. The self-defeating nature of urban road capacity policy: a review of theories, disputes and available evidence. *Transp. Policy* 4, 5–23, doi:10.1016/S0967-070X(96)00030-3.
- Mogridge, M.J.H., 1990. The Downs-Thomson Paradox. In: Mogridge, M.J.H. (Ed.), *Travel in Towns: Jam Yesterday, Jam Today and Jam Tomorrow?* Palgrave Macmillan, UK, London, pp. 181–212, doi:10.1007/978-1-349-11798-7_7.
- Mogridge, M.J.H., Holden, D.J., 1987. A Panacea for Road Congestion?: A Riposte. *Traffic Eng. Control* 28, 13–19.
- Mogridge, M.J.H., Holden, D.J., Bird, J., Terzis, G.C., 1987. The Downs/Thomson Paradox and the transportation planning process. *Int. J. Transp. Econ. Riv. Internazionale Econ. Dei Tras.* 14, 283–311.
- Mohring, H., 1972. Optimization and scale economies in urban bus transportation. *Am. Econ. Rev.* 62, 591–604.
- Næss, P., Mogridge, M.J.H., Sandberg, S.L., 2001. Wider roads, more cars. *Nat. Resour. Forum* 25, 147–155, doi:10.1111/j.1477-8947.2001.tb00756.x.
- Noland, R.B., 2001. Relationships between highway capacity and induced vehicle travel. *Transp. Res. Part Policy Pract.* 35, 47–72.
- Thomson, J.W., 1972. *Methods of Traffic Limitation in Urban Areas*. Organisation for Economic Cooperation and Development, Paris.
- Zhang, F., Yang, H., Liu, W., 2014. The Downs-Thomson Paradox with responsive transit service. *Transp. Res. Part Policy Pract.* 70, 244–263, doi:10.1016/j.tra.2014.10.022.

Further Reading

- Arnott, R., Small, K., 1994. The economics of traffic congestion. *Am. Sci.* 82, 446–455.
- Zhang, F., Lindsey, R., Yang, H., 2016. The Downs-Thomson paradox with imperfect mode substitutes and alternative transit administration regimes. *Transp. Res. Part B Methodol.* 86, 104–127, doi:10.1016/j.trb.2016. 01.013.

Policy Instruments for Plug-In Electric Vehicles: An Overview and Discussion

Jake Whitehead*, **Patrick Plötz†**, **Patrick Jochem‡**, **Frances Sprei§**, **Elisabeth Dütschke†**, *UQ Dow Centre for Sustainable Engineering Innovation & School of Civil Engineering, The University of Queensland, St Lucia, QLD, Australia; †Fraunhofer Institute for Systems and Innovation Research ISI, Karlsruhe, Baden-Württemberg, Germany; ‡Institute for Industrial Production (IIP), Chair of Energy Economics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Baden-Württemberg, Germany; §Chalmers University of Technology, Department of Space, Earth and Environment, Göteborg, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	496
Plug-In Electric Vehicle Policy Examples	496
Categorization of PEV Policy Instruments	497
Primary Purpose	498
Primary Type	498
Overview of Policies	498
Discussion and Conclusions	500
Biographies	501
Further Reading	502

Introduction

During the 21st United Nations Climate Change Conference in Paris, it was agreed that countries must collectively reduce long-term greenhouse gas (GHG) emissions to a level which ensures that the average global temperature increase remains below 1.5°C. Most sectors are already significantly reducing their GHG emissions, with the transportation sector being a major exception. Furthermore, the projected increase in the global car fleet over the coming decades—mainly driven by economic development in developing countries—makes a reversion of this increasing trend in GHG emissions challenging, or even impossible, if no significant changes in mode choice or technical improvements occur. Hence, the current trajectory of transport sector emissions is not congruent with international emission reduction agreements.

In light of this, plug-in electric vehicles (PEVs) are one promising technology that can play a pivotal role in reducing transport emissions. PEVs are already at a high level of technological maturity and their usage does not require significant behavioral changes from consumers, especially compared to nonmotorized transport options. Other arguments for nations to strive for electrified transport system include: lowering dependency on (foreign) fossil fuels, and strengthening local industry.

In this chapter, PEVs include both battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs), given these vehicles can be plugged-in directly to charge using electricity, and therefore take advantage of the cost and environmental benefits of switching to this alternative transport energy source.

Today, emissions from transport are highly correlated with economic development. Besides this income-effect, usage patterns of traffic participants do not change significantly over time. Consequently, significant behavioral changes in terms of substantial mode shifts in the short-term seem unrealistic. Conversely, PEVs, together with the uptake of renewable energy, demonstrate a convincing opportunity to reduce GHG emissions from future transport systems.

Although PEVs' level of technological maturity is high, and their total cost of ownership is already comparable to conventional fossil fuel vehicles in many jurisdictions, most potential users do not adopt this new technology largely due to a lack of supply (mainly few models and insufficient consultation at the point of sale), high upfront investment costs, and other systemic failures, such as limited public charging infrastructure. From an economic point of view, these perceived barriers, together with the fact that not all future costs from fossil fuel vehicles are sufficiently internalized in the cost functions of decision makers, indicates a market failure, which requires market intervention by the state.

Correspondingly, many governments already support the market uptake of PEVs by various policy instruments in order to increase overall welfare. The success of these different policy instruments relies on several factors, and no unique recommendation or tendency toward a single superior instrument can be given, yet. Nevertheless, the choice of the right instrument, or set of instruments, determines the government's investments: choosing an ineffective policy instrument can quickly lead to a significant misallocation of tax revenue. Therefore, the selection of the right policy instrument is decisive. This chapter aims to shed light on current insights into this new market, and in doing so, help to inform these strategic decisions.

Plug-In Electric Vehicle Policy Examples

Policy instruments designed to support the uptake of plug-in electric vehicles come in many shapes and forms. The following section of this chapter includes a number of contemporary examples from some of the leading, early adopting PEV markets.

China is the largest global automotive market and has had, and continues to have, a huge impact on global PEV developments. Chinese PEV policy instruments have combined strong financial incentives with a quota system, supported by nonfinancial incentives and local policies. Exemptions from taxes account for price decreases in the order of EUR 5,000 to EUR 8,500. Large Chinese cities, such as Beijing, Shanghai, or Shenzhen, also provided waivers from license plate availability restrictions. Traditionally, Chinese Government plans and targets have followed a five-year cycle, which provides a certain planning period for both consumers and industry. In the next phase, the Government plans to progressively reduce the magnitude of financial incentives and transition to a quota scheme for “New Energy Vehicles”—the term applied to PEVs in China. In that system, every manufacturer participating in the Chinese market has to comply with a predefined share of PEVs. The credits awarded through this scheme vary according to vehicle type and driving range: longer range vehicles receive more credits, and as such, BEVs receive more credits than PHEVs.

Germany, with a strong automotive industry, has had a long-term R&D industry support scheme for PEV technology. In recent years, this has been coupled with rebates and tax reductions to boost PEV sales. Additionally, local support, such as access to bus lanes or the support of strategic municipal planning, has also been available. As part of the EU, German car manufacturers are also subjected to mandates to reduce tailpipe CO₂ emissions from their vehicles sold. So far, under these provisions, PEVs have continued to receive extra credits.

Norway, which has set a target of banning fossil fuel vehicles sales by 2030, started supporting PEVs already in the 1990s by exempting them from the nation’s relatively high vehicle registration taxes. The vehicle tax exemptions amounted to around EUR 10,000 per vehicle, making PEVs cheaper to purchase than their fossil fuel counterparts, in addition to already being cheaper to operate. These financial incentives, coupled with a large number of waivers on fees, such as road tolls and ferries, provide a highly favorable environment for PEV uptake, and for BEVs in particular. This has led to Norway being the leading PEV market globally, with a new vehicle sales share of 46% in 2018. BEV taxation is planned to remain unchanged for more years, while PHEV incentives could change. Furthermore, a comprehensive network of fast charging stations has been built to facilitate long-distance trips in Norway. At a local level, the city of Oslo is also banning fossil fuel vehicles from the city by allowing parking only for PEVs. Overall, the strong incentives and support for PEVs are in line with Norway’s national target to have zero sales of fossil fuel vehicles by 2025.

Another approach that emphasizes long-term planning and creates more of a technology push are quotas, or zero-emission vehicle (ZEV) mandates, in which PEVs are included. National or regional governments, such as California in the United States or British Columbia in Canada, set long-term targets for the sales shares of ZEVs to be met by automakers. In California, the ZEV mandate seems to have accelerated industry investment in ZEV technology. California is today one of the leading PEV markets and represents approximately half of total PEV sales in the United States.

Categorization of PEV Policy Instruments

“Policies” in general describe a plan of action adopted by a specific actor to support a goal, while “public policy” refers to a plan of action that is taken by government. Policies involve strategies, instruments, and the policy process. The focus of this chapter is specifically on the policy instruments. We focus primarily on national policies as these have received the most attention, but super-national, regional, or city-level policy instruments can be important levers too.

There is a growing body of literature analyzing and defining policy instruments for sustainability transitions in general, but also for PEVs in particular. Policy instruments for the present chapter are defined as concrete tools to achieve overarching policy objectives, and in most cases, to address specific market barriers. In the case of PEVs, the three principal market barriers are:

1. **Lack of market supply:** Although PEV technology is now relatively mature, given there are significant changes required in vehicle manufacturing, particularly in regards to supply chains, many incumbent manufacturers are still repositioning their businesses to be capable of expanding the supply of PEVs to the market. A lack of supply leads to a lack of choice, particularly across vehicle classes.
2. **Higher upfront costs:** During the early stages of market development—as with most new technologies—PEVs cost more to purchase. Although PEVs are much cheaper to operate, the market has still not reached purchase price parity with fossil fuel vehicles. The higher purchase price leads to a reduced demand from potential buyers.
3. **Systemic failures:** Given the transition to PEV technology inherently requires system-wide changes, there are many existing gaps, or systemic failures, slowing the uptake of PEVs. This includes a perceived lack of charging infrastructure, leading to consumers experiencing range restrictions or anxiety, a general lack of understanding of the technology, resistance from incumbents like vehicles dealers due to uncertainty about the future, as well as a lack of standardization across suppliers. While in many developed countries we are already seeing sufficient charging infrastructure from a theoretical perspective, the perceived uncertainty is still dominating consumer opinion.

There are many different PEV policy instruments available to address these principle barriers, and many different categorizations for these instruments. Typically, authors distinguish between (1) economic or regulatory instruments, (2) between sticks, carrots, or sermons, (3) between market pull or technology push, or further much more granular distinctions. All of these distinctions capture some important aspect of the policy instrument. However, here we follow the literature of policy mixes in sustainability transitions,

Table 1 An example of PEV policy instrument categorization including examples of policy instruments

<i>Policy categorization</i>		<i>Primary purpose</i>		
		<i>Technology push</i>	<i>Demand pull</i>	<i>Systemic</i>
Primary type	Economic instruments	R&D grants; contestable funds	Purchase subsidy; income tax credit; sales tax; registration fee; annual vehicle tax; toll road fee; parking fee; emissions tax; fuel tax	Charging infrastructure subsidies; cooperative R&D grants; broader transport tax reforms
	Regulatory instruments	CO ₂ emissions standards; quotas; targets; city planning measures	Ban of fossil fuel vehicle sales; free bus/transit lane access; free public charging	Emissions zones; standardization
	Information	Vehicle dealer education	Labeling; awareness campaigns; test drive events; public exhibitions; public event integration; tourism programs	Stakeholder and public dialogues

which is well-applicable as PEVs are a sustainable transportation technology, and their introduction implies a transition for many aspects of the automotive regime, and society in general.

Given the breadth and variety of policy instruments available, and following the literature on policy mixes in sustainability transitions, the remainder of this section categorizes policy instruments based on their primary purpose and primary type; with both categories explained further below.

Primary Purpose

Categorization of policy instruments by primary purpose principally relates to the main intended outcome of the instrument. It should be emphasized that instruments can have more than one purpose, but typically the main or primary purpose is clear, particularly in regards to the specific market barrier for which the policy instrument aims to address, for example, lack of supply, high upfront costs, or systemic failures. PEV policy instruments can be categorized into the following three primary purposes:

1. **Technology push or supply side** instruments foster the development and market formation of PEVs. For example, low-carbon fuel standards; PEV quota systems.
2. **Market pull policy instrument** create or stimulate demand for PEVs. For example, purchase rebates; registration fee discounts.
3. **Systemic policies** help to improve the general systemic attributes for PEVs. For example, provision of public charging infrastructure; introduction of emissions zones; technology standardization.

Primary Type

In addition to primary purpose, PEV policy instruments can also be categorized by their primary type. The three primary types of PEV policies are:

1. **Economic instruments** directly influence the total cost of ownership of PEVs. For example, purchase price rebates, registration fee discounts, annual tax reductions.
2. **Regulatory instruments** improve market conditions in favor of PEVs, compared to incumbent alternatives, through technical specifications. For example: PEV quotas, fuel standards, fossil fuel vehicle bans, low emission zones.
3. **Information** improves awareness of the technology in the society. For example, labeling at the point of sale, public information campaigns, test drive events, public exhibitions.

Overview of Policies

By combining primary purpose and primary type, it is possible to categorize PEV policy instruments into nine distinct categories, as shown in [Table 1](#).

[Table 2](#) includes further details on each of the example PEV policy instruments categorized in [Table 1](#), including regional examples of where these instruments have currently (2019) been implemented around the world.

While there is some evidence to suggest that specific policy instruments appear to have a greater impact on PEV diffusion compared to alternative instruments, it is increasingly becoming clear that there is no “silver bullet” for encouraging the uptake of PEVs. This is in line with arguments provided by the policy mix literature, which concludes that an effective transition policy needs instruments to be in line with an overarching strategy, that is, perceived by market actors to be reliable. Furthermore, the specific design features of instruments play a role, for example, rebates lead to differential impacts, if they are provided upfront compared to if consumers need to claim money back at a later stage.

Table 2 Example list of PEV policy instruments

<i>Example PEV policy instruments</i>	<i>Primary type</i>	<i>Primary purpose</i>	<i>Description</i>	<i>Regional examples</i>
R&D grants, contestable funds	Economic instruments	Technology push	Funds established to support R&D into PEV technology, and support trials of PEV technologies, as well as use cases that would be beneficial but are not yet commercially viable.	New Zealand, Australia, Germany
Purchase subsidy		Demand pull	A point-of-sale subsidy that reduces the upfront cost of purchasing a PEV.	British Columbia (Canada)
Income tax credit			A credit that can be used to offset or reduce income tax liability, indirectly reducing the upfront cost of purchasing a PEV	US
Sales tax discount			A reduction in sales tax for PEVs.	US, Norway
Annual registration fee discount			A reduction in annual registration fees for PEVs	Germany, France
Toll road or congestion tax discount			A reduction in toll road/congestion charges for PEVs	Norway, London
Parking fee discount		Systemic	A reduction in parking fees for PEVs	Oslo, Germany
Company car tax discount			A reduction in company car tax for PEVs	Norway, Netherlands, Sweden
Emissions tax			A road or vehicle tax based on emissions rates, favoring PEVs with low or zero emissions.	Several EU countries
Fuel tax			A tax to increase the cost of using fossil fuels.	Several EU countries
Charging infrastructure			Provision of charging infrastructure to address PEV range anxiety, and bolster market confidence.	Australia, EU, Japan, US
CO2 emissions standards	Regulatory instruments	Technology push	Standards to progressively reduce CO ₂ emissions of the fleet.	US, EU, Japan
Targets or quotas			Targets or quotas used to stimulate market supply of PEVs.	Norway, New Zealand, UK, US
City planning measures			Measures to normalize the integration of PEVs into planning, including mandating charging infrastructure requirements for buildings, car parks, etc.	EU, UK, US, Australia
Ban on fossil fuel vehicle sales		Demand Pull	Bans to send a clear signal to consumers and industry about the PEV transition timeline.	France, UK, Norway, Netherlands, China
Free bus or transit lane access			Regulation that provides PEV owners with the added benefit of bus/transit lane access.	Oslo, California, UK
Free public charging			Regulation that provides PEV owners with the added benefit of free public charging.	Oslo
Emission zones			Zones restricted to low and/or zero emissions vehicles to encourage a transition to these vehicles.	Germany, UK
Standardization		Systemic	Implementation of technical standards to ensure consistency across suppliers and simplify experience for PEV users.	EU, US, Japan
Vehicle dealer education	Information	Technology push	Awareness campaigns to educate vehicle dealers on PEV technology and minimize dissemination of misleading information	US
Labeling		Demand pull	Use of labels to promote awareness of PEV benefits compared to other technologies. Also use in addition to, or as part of vehicle number plates to promote awareness in addition to facilitating regulation of PEV benefits, such as access to bus or transit lanes, PEV-exclusive parking, etc.	Norway, US
Awareness campaigns			Campaigns to improve public awareness and understanding of PEV technology.	British Columbia, Norway, US
Test drive events			Specific events held to provide the general public with opportunity to experience PEV technology, and address any questions, including potential concerns.	US
Public exhibitions			Public exhibitions to demonstrate PEV technology.	US
Public event integration		Systemic	Integration of PEV technology with public events to raise profile and general awareness.	London Olympics, Beijing Olympics
Tourism programs			Strategies to encourage EV tourism, with the ambition of address PEV technology misconceptions through real-world experiences and influence consumer preferences.	Oregon, Germany
Stakeholder and public dialogues			Increase profile and awareness of PEV technology.	Norway

This is underlined by the fact that leading PEV markets, such as Norway, have implemented a range of policy instruments spanning across the primary type and primary purpose categories outlined here, in addition to a long-term policy goal, that is, the ban of fossil fuel vehicles. This makes sense, given there are several barriers to PEV diffusion—as mentioned above—that each require differing approaches in order to be resolved.

For example, while the higher upfront investment costs of PEVs are a significant barrier to uptake that can be partly addressed through demand pull, economic instruments, such as purchase rebates, or vehicle tax discounts, these instruments will not address the PEV market barrier of a lack of charging infrastructure (a systemic failure). To address this barrier, a systemic, economic instrument, for example, a charging infrastructure subsidy to increase charging opportunities, could be combined with demand pull, information in the form of public awareness campaigns to address misconceptions about PEVs' limited driving range capabilities. Again, however, these instruments are much more powerful, if market actors perceive them as being part of a longer-term strategy that is consistently followed.

Policy-makers must also take into account regional considerations and identify a package of policy instruments that make sense for the local market, while addressing the principal barriers to PEV diffusion. For example: there is some evidence to suggest that part of the monetary benefit of vehicle purchase subsidies is captured by vehicle manufacturers—as is the case with most government subsidies. While this may make sense in regions that manufacture PEVs, such as Germany, UK, US, Japan, and China, where governments already support the local vehicle manufacturing industry, in non PEV-manufacturing nations, such as Norway, Denmark, Ireland, Australia, and New Zealand, directly subsidizing foreign manufacturers makes less sense. Instead, the same pool of available funding could potentially be more effectively used through other demand pull, economic instruments, such as vehicle tax reductions.

Discussion and Conclusions

There are a significant number of policies that decision makers can choose from to support the market diffusion of PEVs. Choosing the right policy instrument, and its adequate design, as well as the appropriate policy combination, depends on context, market conditions, as well as the national transition pathway. The evidence so far on demand pull, economic instruments, such as purchase subsidies, suggests that they foster PEV market diffusion, even if the effect is rather low. Yet, when the purchase price of PEV and conventional fossil fuel vehicles is similar due to high incentives, and as part of a broader strategy, the market can quickly tip in favor of PEVs, as has been the case in Norway.

There are differences in responses to the instruments, for example, a point of sales rebate seems to be more efficient than income tax deductions, however, more studies are needed to better discern the differences in effects. On the supply side, most research has been on the California ZEV mandate. Studies find that the mandate has had a positive impact on innovation activity for OEMs, increasing research and development.

Many of the leading PEV markets, such as Norway, have implemented several instruments to support the diffusion of PEVs. When designing a policy mix, market acceptance becomes an important factor for success to ensure market actors on the demand and supply side invest in the desired way. To increase acceptance, policy making and implementation should be coherent, and the policy mix should also be consistent. For example, when planning for a ZEV mandate it is important to understand how it interacts with other policies regulations to avoid duplications and conflicting policy mechanisms.

It is also important to keep in mind that the empirical evidence so far on the efficiency and effect of policy instruments for PEVs is caveated by the fact that most markets, with the possible exception of Norway, are still in the early phases of market diffusion, and thus the results might change as markets continue to develop. Given this, policies will need to be adjusted as markets continue to develop. Evidence from California shows that the number of “free-riders” for incentives (i.e., the number of people that would have purchased an PEV even without incentives) has decreased as the market has expanded. In Norway, as the number of PEVs on the roads have increased, local regulations, such as access to bus-lanes, have been phased out while uptake continues to grow. This has also been, in part, due to the unintended negative consequences of increased congestion from higher volumes of PEVs using these road facilities, highlighting the importance of these measures being temporary in duration, and closely monitored in order minimize negative effects.

In the case of Norway, the financial cost of PEV incentives has been high, with the benefits flowing largely to higher-income Norwegian households. In saying this, it must be recognized that almost all new technology is initially expensive, and only becomes affordable for the mainstream through economies-of-scale. The Mercedes-Benz S-Class was the first production vehicle to have an anti-lock braking system (ABS) in 1978, which retailed for 150,000 EUR/\$US 170,000 (in 2019 currency). The only reason these systems are now in vehicles retailed for under 10,000 EUR/\$US 11,000 is due to economies-of-scale. The same concept applies to PEVs, with many of these vehicles over 35,000 EUR/\$US 40,000 already approaching price parity with conventional fossil fuel vehicle equivalents. These upfront price reductions have been accelerated by policy instruments, which have led to increased consumer demand, and driven economies-of-scale across the entire PEV supply chain.

It is also important to understand the financial costs and impacts of different policy instruments, and weigh these up against the expected co-benefits of implementation, for example, lower emissions, reduced transport costs, improved air quality, etc. It is sometimes argued that other measures, such as nonmotorized (or active) modes are more cost-efficient mechanisms for emissions reductions in the transport sector, but this would require significant changes in current mobility preferences, which appears unlikely in the current market. Others argue that emissions reduction in other sectors is more cost-efficient (in terms of CO₂ abatement costs),

for example, in electricity generation, and that the electrification of road transport is a rather slow emission reduction mechanism due to required changes in vehicle production facilities and supply chains, in addition to behavioral inertia and a lack of charging infrastructure, among other factors. This is principally why strong policy support is required now to assist in accelerating the transition to electric vehicles (in conjunction with economies-of-scale leading to upfront cost reductions), in parallel to efforts being made in other sectors, to ensure that global emission reduction targets can be met as quickly as possible. Without the electrification of road transport, the rapidly increasing global car stock will lead to further increases in emissions.

As battery costs decline further, PEVs are expected to become cost-competitive with conventional fossil fuel vehicles, and thus, purchasing incentives may no longer be required. It is, however, critically important that any phase out of policy instruments is done gradually, and is clearly communicated to consumers, to avoid market shock. Even if PEV prices fall, and supply increases, there may still be rationale for continued policy support, if other systemic failures remain, including ongoing range anxiety due to a lack of charging infrastructure and low consumer awareness. As such, consumer education campaigns, and other systemic policy instruments, will continue to be important. The private sector will also play an increasingly important role in the rollout of PEV charging infrastructure, with this investment being catalyzed through initial government support and clear standards, as part of a broader PEV strategy.

A further issue that has not been addressed in this chapter is the phasing out of the incumbent fossil fuel vehicle industry. There is an additional need to support the development of industry transition strategies that consider the socioeconomic effects of these changes, particularly in regions that are currently economically-dependent on fossil fuel vehicle production, and/or the associated maintenance requirements of conventional drivetrains. If these broader impacts are not taken into account, and actively planned for by policy-makers, incumbent fossil fuel vehicle businesses, and those who they employ, may try to resist the diffusion of PEV technology, which would not only have dire environmental consequences, but could also have long-term negative economic ramifications for the regions concerned, if they are left behind in this global transition.

Biographies



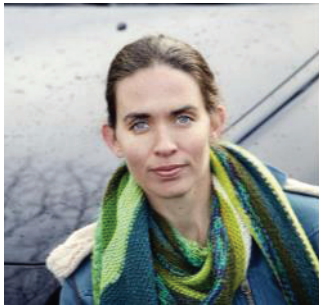
Dr. Jake Whitehead is the Founder and Secretariat of the Australia and New Zealand Clean Transport Coalition, a member of the International Electric Vehicle Policy Council, and holds two PhDs in Transport Science and Engineering. Dr Whitehead's research primarily focuses on the costs, benefits, opportunities, and risks of clean transport technologies, as well as understanding the impacts of government policies on supporting and managing the uptake of these innovations. Dr Whitehead has also undertaken research on the synergies between clean transport and energy systems, in particular through the rollout of electric vehicle-to-grid systems, as well as the effectiveness of different transport technology pathways on reaching global emission reduction targets. Dr Whitehead has worked closely with governments and businesses internationally to advise on sustainable transport policies. These efforts have included co-coordinating the development of Australia's most comprehensive electric vehicle strategy for Queensland.



Dr. Patrick Plötz studied Physics in Greifswald, St. Petersburg and Göttingen. Dissertation in Theoretical Physics on correlated electrons in one-dimensional systems. Additional studies of Philosophy and History of Science in Göttingen. Doctorate degree in Theoretical Physics from the University of Heidelberg (Institute for Theoretical Physics) on complex dynamics in cold atomic gases. From January to December 2011 researcher in the Competence Center Energy Policy and Energy Systems at the Fraunhofer Institute for Systems and Innovation Research ISI, since January 2012 in the Competence Center Energy Technology and Energy Systems.



Dr. Patrick Jochem is a post-doc at the Karlsruhe Institute of Technology (KIT) and Associate Editor of Transportation Research Part D: Transport and Environment. Since 2009 he is leading the interdisciplinary research group “Transport and Energy” at the Chair of Energy Economics and since 2012 the eMobility Lab at the Karlsruhe Service Research Institute (KSRI). He is author of more than 100 scientific peer-reviewed papers and has a strong funding record. In 2009, he received his PhD in transport economics, which was funded by the German Federal Environmental Foundation (DBU). In 2014, he was awarded by the Heidelberg Academy of Sciences and Humanities. He studied economics at the universities of Bayreuth, Mannheim and Heidelberg. His research interests are in the fields of electric mobility, energy system analysis, and ecological economics.



Frances Sprei is an Associate Professor in Sustainable Mobility. She has a PhD in Energy and Environment from Chalmers University of Technology in Sweden. Her research assess different mobility services such as electric vehicles, car-sharing, and autonomous vehicles and studies their adoption and diffusion. Her methods are interdisciplinary and she takes into consideration technical, economical, and behavioral aspects.



Dr. Elisabeth Dütschke studied psychology, business administration, and marketing at TU Darmstadt and RWTH Aachen. For her PhD thesis in organizational behavior at the university of Constance she received an award from Südwest Metall as an outstanding contribution to research. Further work experience includes consulting of private and public organizations and journalism as well as university lectures. Since June 2009 at the Fraunhofer ISI as senior scientist and project leader she has been involved in a large number of national and international research projects. Since March 2019 coordinator of the Unit “Actors and Social Acceptance in the Transition of the Energy System.” Her work focuses on the human perspective on a changing energy system. She is the main contact for societal issues around the energy transition for the institute.

Further Reading

- Richards, D., Smith, M., 2002. *Governance and Public Policy in the UK*. Oxford University Press, Oxford.
- Rogge, K.S., Reichardt, K., 2016. Policy mixes for sustainability transitions: an extended concept and framework for analysis. *Res. Policy* 45 (8), 1620–1635.
- Rogge, K., Dütschke, E., 2018. What makes them believe in the low-carbon energy transition? exploring corporate perceptions of the credibility of climate policy mixes. *Environ. Sci. Policy* 87, 74–84.
- International Energy Agency, 2019. *Global EV Outlook 2019*, IEA. Available from: www.iea.org/publications/reports/globalevoutlook2019/.
- United Nations Environmental Programme, 2018. *The Global Electric Vehicle Policy Database*, UN Environment. Available from: <https://www.unenvironment.org/resources/publication/global-electric-vehicle-policy-database>.
- Hardman, S., Chandan, A., Tal, G., Turrentine, T., 2017. The effectiveness of financial purchase incentives for battery electric vehicles—a review of the evidence. *Renew. Sustain. Energy Rev.* 80, 1100–1111, doi:10.1016/j.rser.2017.05.255.
- Mersky, A.C., Sprei, F., Samaras, C., Qian, Z., 2016. Effectiveness of incentives on electric vehicle adoption in Norway. *Transp. Res. Part D: Trans. Environ.* 46, 56–68, doi:10.1016/j.trd.2016.03.011.
- Münzel, C., Plötz, P., Sprei, F., Gnann, T., in press. How large is the effect of financial incentives on electric vehicle sales? – A global review and European analysis. *Energy Econ.*
- Whitehead, J., Washington S., Zheng, Z., Perrons, R., in press. Charging up: An analysis of the supply and demand factors influencing Australia's electric vehicle market.
- Smit, R., Whitehead, J., Washington, S., 2018. Where are we heading with electric vehicles? *Air Qual. Clim. Change* 18–27.
- Whitehead, J., Washington, S., Franklin, J., 2019. The impact of different incentive policies on hybrid electric vehicle demand and price: an international comparison. *World Electr. Veh. J.* 10 (2), 1–19, doi:10.3390/wevj10020020.
- Langbroek, J.H.M., Franklin, J.P., Susilo, Y.O., 2016. The effect of policy incentives on electric vehicle adoption. *Energy Policy* 94, 94–103, doi:10.1016/j.enpol.2016.03.050.
- Yang, Z., Slowik, P., Lutsey, N., Searle, S., 2016. *Principles for effective electric vehicle incentive design*, The International Council on Clean Transportation. Available from: https://www.theicct.org/sites/default/files/publications/ICCT_IZEV-incentives-comp_201606.pdf.
- Harman, S., 2019. Understanding the impact of reoccurring and non-financial incentives on plug-in electric vehicle adoption—a review. *Transp. Res. Part A: Policy Prac.* 119, 1–14, doi:10.1016/j.tra.2018.11.002.
- Lepitzki, J., Axsen, J., 2018. The role of a low carbon fuel standard in achieving long-term GHG reduction targets. *Energy Policy* 119, 423–440, doi:10.1016/j.enpol.2018.03.06.

Vertical and Horizontal Separation in the European Railway Sector and Its Effects on Productivity

Pedro Cantos-Sánchez, Department of Economic Analysis and ERI-CES, University of Valencia, Valencia, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	503
The Restructuring Process in the European Rail Industry	503
Restructuring Measures at Vertical Dimension	503
Restructuring Measures at Horizontal Dimension	504
Approaches to Measure the Productivity in the Rail Industry	504
An Empirical Assessment of the Rail Reforms in Europe	505
Conclusions	506
References	507
Further Reading	507

Introduction

One of the most significant results in the rail sector during the last quarter of the 20th century was the enormous decline in transport market share at the worldwide level, and mainly at the European level. In particular, a drastic fall in rail modal shares was produced both in the passenger and freight markets in Europe in the 1970s and 1980s. For the EU15 countries, the modal share for rail passenger transport fell from 10.4% in 1970 to 6.7% in 1990. The decline was even more considerable for freight with rail decreasing from 21.2% modal share in 1970 to 11.1% in 1990 (OECD, 2013).

Furthermore, this decline provoked a real deterioration in the financial accounts of many of railway networks around the world. One of the first replies to this situation was the redefinition of the traditional management system in railways. This system had consisted in public ownership and management of both infrastructure and all rail operations by a single firm. Therefore, complete integration (through the vertical dimension between infrastructure and rail operations and the horizontal level between all the different rail services) was the main economic feature of railway systems.

The first step in the restructuring of the rail industry in Europe was the Directive 91/440, which promoted the separation of accounting systems for rail operations and infrastructure. Additionally, this Directive stimulated a rail management for the industry, which was more independent of the State. The second step in this restructuring process was the launch in 1996 of the *White Paper* (known as the First Railway Package), with the objectives to favor the opening of the market, the improvement of the financial accounts and boosting market forces and an effective competition in the rail market. The subsequent Railway Packages tried to foster and create an integrated European rail area at both the legal and technical levels and at the same time to open international passenger and freight transport to competition in the European Union. Lastly, the Fourth Railway Package introduced the principle of competitive tendering for public service contracts and promoted the opening of the market for domestic passenger transport services by rail.

Our paper will be classified into three parts. Firstly a description of the different restructuring measures in the European rail industry will be presented. We differentiate between measures at vertical dimension, that is, on the degree of separation between infrastructure and rail operations, and measures at horizontal dimension, that is, those measures directed to promote competition in rail operations. Secondly, the different approaches and methodologies used to evaluate the effectiveness of these measures will be presented and described. Thirdly the evaluation of the effects of these reforms will be discussed, describing the different results obtained by the literature. Finally, we conclude with some remarks and policy recommendations.

The Restructuring Process in the European Rail Industry

Restructuring Measures at Vertical Dimension

As previously mentioned, one of the first goals of the EU in the rail markets was to promote the separation between infrastructure and rail operations. Separation implied unbundling those rail competitive activities (like rail operations) from those activities that have the features of natural monopoly. There are very detailed papers where the *pros* and *cons* of separating infrastructure and rail operations are described in depth in a detailed way. The main advantage of vertical separation was that it would allow greater competition in the rail sector. Abbott and Cohen (2017) describe that additionally vertical separation will:

- Create incentives to reduce costs and promote innovation from increased competitive pressures amongst train operators.
- Lead to greater specialization of operators and their expansion into other markets.

- Create competitive neutrality between train operators.
- Create a greater degree of transparency for policymakers in the form of information on access pricing, track usage, etc.

However, there is some evidence that scope economies can be lost in a structure which separates infrastructure and rail operations. There are some examples where coordination between infrastructure and operations can lead to deterioration. So, maintaining or upgrading the rail track and other facilities directly affects operating scheduled services and vice versa. This occurs due to the existence of a technological interdependence between infrastructure and vehicle technologies (Pittman, 2005). For this reason, a track infrastructure owner could be reluctant to bear the risks of specialized investments without long term contracts for a specific user.

Summarizing, the potential gains obtained from a separated structure in terms of greater competition should outweigh the loss of coordination and scope economies of an integrated structure. Thus, the separation degree can be implemented at different levels:

- Only at an accounting level: requiring that different accounts must be prepared for infrastructure and rail operations, but within the same entity.
- Organizational level (or “holding model”), where rail operations and infrastructure are managed by different subsidiaries within one holding company and taking independent decisions.
- Institutional separation, where rail operations and infrastructure are managed and organized by different and totally independent companies.

While some European countries have kept an integrated rail structure (such as Ireland, Hungary, Lithuania, Luxembourg, or Russia), most European countries have opted for a more disintegrated structure (United Kingdom, Sweden, Netherlands, Spain, Finland, Belgium, etc.). However, a few countries (such as Germany) keep a holding model where *DB Netze* is in charge of infrastructure management within the whole *DB* group. Other countries, like France, opted for a separated structure, where some of the tasks allocated to the infrastructure manager were allocated to the rail operator *SNCF*. In this case the model is indeed closer to a holding structure rather than a separated structure.

Restructuring Measures at Horizontal Dimension

The second dimension on the rail restructuring process is the introduction of competition in rail markets. There are basically two ways to introduce competition in rail markets:

1. The tender concession model for a particular rail market (competition *for the market*)
2. The open-access or free entry model (competition *in the market*)

The first model has been applied to many passenger regional services in some countries (especially in the United Kingdom, Sweden, Germany, Denmark, Netherlands, or the Czech Republic). Only long-distance services were offered by this model in a few countries such as the United Kingdom or Sweden. However, many of the incumbent operators are still publicly owned, and keep considerable power to influence the tendering results or to introduce barriers against potential competitors. Therefore, except in a few cases like the United Kingdom or Sweden, most of the tendering processes in Europe were awarded to the incumbent rail operators. The open access (or competition in the market) has been less used in the passenger sector. Casullo (2016) identifies 15 countries where open access is allowed, but only in 6 (Austria, Czech Republic, Germany, Italy, Sweden, and the United Kingdom) has the competition between different operators in the market been effective. As a final result, except for the United Kingdom, Poland, and Italy, the market share of competitor in the passenger market is below 20%, indicating the lack of success of these measures (European Commission, 2016; UNECE, 2018).

The open-access model has been the model developed for the freight market. In this sector, the entry of new operators has been more successful than in the passenger market. In fact, and from data provided by UNECE (2018), for seventeen European countries in 2014, the market share of competitors was in a range of 20%–55%. Furthermore, the market share of competitors in the freight market increased in 19 European countries from 2011 to 2014, decreased in only three countries and remained constant in four countries. However, except for Sweden and the United Kingdom, the market share for the incumbent remains dominant also for the freight market.

Approaches to Measure the Productivity in the Rail Industry

Productivity can be estimated as a measurement of the behavior of a firm. In particular, productivity can be expressed as the ratio of output with respect to the inputs used in the production process. When all outputs and inputs are included in the productivity measure we can define an index of total productivity. One way to measure the productivity level is to define index numbers. These indexes are very easy to calculate, because it is the simple ratio between one output and one input. They have important limitations when firms are multioutput and multiinput, as happens in the rail industry. The choice of a single output or input to define the index numbers will produce a biased result, due to the absence of the rest of inputs or outputs in the analysis. A good example of this type of technique is the paper by Oum et al. (1998).

Alternatively, more sophisticated approaches like econometric methods or Data Envelopment Analysis (DEA) can solve this last drawback and allow for the incorporation of different inputs and outputs in the analysis.

DEA is a linear programming technique that calculates technical efficiency by measuring the ratio of total inputs employed to total outputs produced for each firm or entity. DEA identifies the most efficient firms providing a vector of outputs employing the least number of possible inputs. Each firm in the sample receives an efficiency score determined by the variance in its ratio of inputs employed to outputs produced with respect to the most efficient producer in the sample. DEA has been used extensively to assess productivity and efficiency levels for the rail industry (see for instance Cantos et al., 2010, 2012; Growitsch and Wetzel, 2009; Wetzel and Growitsch, 2006). The main advantage of this method is that we do not need to consider any assumption about the distribution of deviations from the frontier or to adopt any functional form for this frontier. The disadvantages are the sensitivity to the reference sample (efficient units are those with the best performance among the observed ones, but perhaps they would not be considered efficient if the reference set were changed) and the consideration as inefficiency for any deviation from the frontier.

In particular, the DEA approach assumes that for each period t there is a set of N railway system ($i = 1, \dots, N$). Each railway system produces a vector of y_i outputs ($m = 1, \dots, M$) using x_i inputs ($k = 1, \dots, K$). The measurement of technical efficiency (input-oriented) under constant returns to scale using DEA is obtained by solving the following problem for each period and each railway system (assuming constant returns to scale):

$$\begin{aligned} \text{Min } & \theta_j \\ \text{s.t. } & \sum_i \lambda_i y_{im} \geq y_{jm} \quad \forall m \\ & \sum_i \lambda_i x_{ik} \leq \theta_j x_{jk} \quad \forall k \\ & \lambda_i \geq 0; \quad i = 1, \dots, N \end{aligned}$$

From the solution of this problem for each one of the N rail systems of the sample we obtain N optimal values for θ . Each value for the parameter is the input-oriented technical efficiency measure of each rail system which, by construction, satisfies $\theta \leq 1$. Those rail systems with $\theta < 1$ are considered technically inefficient, while those with $\theta = 1$, are cataloged as technically efficient, since they stand at the frontier.

Econometric methods can propose either the estimation of a cost or production function. The estimated function can then be used to identify changes in productivity or productive efficiency. A number of studies have applied deterministic or stochastic frontier methodologies to the rail industry. The drawback to estimate a cost function is that information on input prices and financial data is needed. In case of the rail industry, these data are difficult to achieve. This is the reason why production function is a more common analysis in the rail industry. Only physical data on outputs and inputs are required. Recent advances in econometrics have developed new approaches, such as like the distance function, which allows for the introduction of several outputs in a production function.

Assume that the production technology is defined by a vector Y of m outputs, as in the DEA approach, that can be produced by a vector X of k inputs. The output distance function is defined as $D_O(X, Y)$, which is nondecreasing, positively linearly homogenous and convex in Y , and decreasing in X . The output distance function is by definition linearly homogenous in outputs, which is imposed by dividing all outputs by one of the outputs. If we adopt a transcendental-logarithmic function (TL), the distance function to estimate can be expressed as follows:

$$-\ln y_{mi}^t = TL(x_i^t, y_i^t / y_{mi}^t, t; \beta) - \ln D_O^t(x_i^t, y_i^t)$$

Setting

$$u_{it} = -\ln D_O^t(x_{ij}, y_{it})$$

and adding a stochastic error term (v_{it}), the specification is similar to that of a parametric stochastic frontier with a decomposed error term:

$$-\ln y_{mi}^t = TL(x_i^t, y_i^t / y_{mi}^t, t; \beta) + u_{it} + v_{it}$$

where u_{it} is a nonnegative random error term representing the time-varying technical inefficiency. Finally, the technical efficiency is calculated as

$$TE_{it} = \exp(-u_{it}) = D_O^t(x_i^t, y_i^t)$$

An Empirical Assessment of the Rail Reforms in Europe

As a starting point, a first and interesting way to measure the scope of the liberalization process in the European rail industry was the elaboration of a series of indexes developed in 2002, 2004, 2007 and 2011 by IBM (2002, 2004, 2007, 2011). In these studies, different indexes on rail performance were developed according to three dimensions: the legal development of the industry (LEX), the degree of access permitted by the industry (ACCESS) and the level of competition (COM) within the rail industry. The LEX index assesses the degree in which the EU directives have been transposed to the legal system in each particular country. The ACCESS index values the degree in which a country has implemented the appropriate measures to promote the access of new rail operators in the country. Finally, the COM index measures the extent to which nonincumbent operators have managed to enter the market in an

effective way, valuing the number of new operators and their share in the rail market. The indexes are ranked between 0 and 1000, so that a higher score indicates a better performance in the area analyzed.

Analyzing the evolution of these indexes, the average value for the EU-25 for LEX and ACCESS has increased and the values for these indexes are relatively high (800 and 683, respectively) for the last year. However, the COM index has evolved negatively, with a value of 429 for the last year. The final conclusion is that most European railways have made notable progress in the implementation of the legal and procedural aspects of the European Directives. However, progress in introducing more competition in the European rail market, especially in the passenger sector, has been poor. Summarizing, the sector remains strongly concentrated and characterized by a very low number of new operators and the persistence of big market shares for incumbent operators.

Regarding the separation between infrastructure and rail operations, the findings obtained by the literature are not conclusive. Cantos et al. (2012) obtain, using DEA, that rail systems which separated infrastructure from rail operation are slightly more efficient than integrated more rail systems. Furthermore, the authors find that a higher increase in efficiency is produced when vertical separation is combined with the introduction of competition (using a tendering process or an open access system). A similar result is obtained in other works (Affuso and Newbery, 2002; Friebe et al., 2003, 2010; Jensen and Stelling, 2007; Merkert et al., 2012).

In this line, this result suggests that separation is a good measure to prepare and to promote the introduction of competition in the rail market. Similarly, countries that did not separate their industries vertically emerged as the most inefficient, although the improved efficiency in countries that only restructured vertically is very modest compared to those that have also embarked on horizontal separation. However, the scarcity of examples with a mixed approach (such as in Germany or Switzerland) where an integrated structure is combined with the introduction of competition prevents us from finding clear evidence of the benefits or otherwise of a more integrated system. This would explain the results of the studies in the United States freight industry, where separation barely leads to small efficiency gains (Bitzan, 2003; Ivaldi and McCullough, 2004, 2008). Nevertheless, other authors like Abbott and Cohen (2017) suggest that in Europe separation may have produced improvements in productivity, simply because so many national rail industries started from levels of low productivity in the first place.

However, many papers find that a separated vertical structure leads to additional costs which could be greater than any benefits derived from improved efficiency driven by competition. Clearly, close coordination between train operation and track maintenance is key to guarantee a reliable and efficient rail service. Additionally (see Pittman, 2005), there is a technological interdependence between infrastructure and vehicle technologies. In fact, rail tracks are designed depending on particular users, and these may be different whether they are used for high-speed trains, passengers in general, or freight and bulk services. In a separated structure, an infrastructure manager would be reluctant to undertake long-run and specific investments without a clear and a long run agreement with specific users. This risk is clearly minimized when the rail structure is integrated in the same entity.

The loss of these coordination effects or scope economies of vertical integration is the main danger of a separated structure. Additionally, there can be a reduction in the incentives to provide appropriate track maintenance and new capacity in the long run (see Bitzan, 2003; Growitsch and Wetzel, 2009; Ivaldi and McCullough, 2004, 2008; Merkert and Nash, 2013; Merkert et al., 2012; Wetzel and Growitsch, 2006).

Finally, some papers (Nash et al., 2014; Mizutani et al., 2015) find evidence that these coordination costs depend on the density and type of traffic. They find that these costs are notably higher if networks are highly used. Therefore, the costs of separation are significantly greater in densely populated areas. This occurs because management and scheduling costs are considerable when traffic density is also very high.

Summarizing, the final evaluation of the benefits coming from a separated structure will be an empirical task. The efficiency gains flowing from a separated structure with a higher and intense competition should outweigh the loss of coordination effects or economies of scope that the separated structure will produce. This may explain why the results obtained by the literature are not totally conclusive, stating in many cases that an integrated or separated structure is not necessarily a key factor in achieving an efficient rail system.

Finally, we present the results regarding the analysis of the measures to promote competition in the rail market. As we explained before, these measures have basically consisted in the introduction of a tendering system for some regional passenger services, specially regional or noncommercial services. However, in the freight sector the process has consisted in the direct entry of new operators, and this process has been undertaken in a larger number of countries. In general, most of the papers which have assessed these measures (Cantos et al., 2010, 2012; Driessen et al., 2006; Link, 2016) find evidence of the positive (but small) effects derived from more competition on the productivity of railway systems. The effectiveness of competitive tendering undertaken in the passenger sector largely depends on the right design of the bidding process favoring the most economically efficient operators and making sure that this efficiency gain is passed on to the society through lower subsidies. The experience of those countries who are most advanced in this measure has been positive (Germany, Sweden, The Netherlands, and United Kingdom). In the case of the entry of new operators in the freight sector, there is evidence that this process has revitalized the rail market, improving the rail market share and the productivity of the rail sector. Lastly, there is little evidence regarding the analysis of free entry (open access) in the passenger sector. New experiences of competition in the passenger market in Italy, Czech Republic and Austria have not led to conclusive results in the effectiveness of this measure (Casullo, 2016).

Conclusions

The rail industry is characterized by particular and economic features that in turn will influence their performance and technical behavior. Rail infrastructure fulfills the conditions of natural monopoly, while rail operations mostly fulfill the conditions for a

competitive market. This has been the economic rationale to favor the separation between infrastructure and rail operations in Europe. Additionally, some measures at horizontal level have been developed to promote internal competition in the European rail market. These measures have consisted in the introduction of franchising systems in the passenger sector (mostly regional or noncommercial services) and of free entry in the freight sector, although recently there are some experiences of competition in the market in some commercial passenger services. The evaluation of these measures is a key element when assessing the effectiveness of European rail policy, and for this reason we have carried out a range of different approaches and methodologies in this analysis.

A separated rail structure should favor the introduction of competition in the rail market improving the technical and financial performance within the sector. However, there is important evidence that a separated structure will increase the operational costs due to a loss of coordination effects between infrastructure and operations. Besides, there is clear risk of lower maintenance of the rail track and provision of new capacity in the long run following such a separated approach. Taking into account this tradeoff, the literature has provided mixed and far from conclusive effects on the effectiveness and productivity of this measure. Regarding the changes in the horizontal level, most of the papers show that, in general, the results have been largely positive. In the passenger sector, an appropriate design of the bidding system is key to making sure that the process can guarantee an efficient outcome. In the freight sector, a nondiscriminatory competition system between the rail incumbent and new operators is also essential in order to achieve favorable results.

References

- Abbott, M., Cohen, B., 2017. Vertical integration, separation in the rail industry: a survey of empirical studies on efficiency. *Eur. J. Transp. Infrastruct. Res.* 17 (2), 207–224.
- Affuso, L., Newbery, D., 2002. The impact of structural and contractual arrangements on a vertically separated railway. *Econ. Soc. Rev.* 33 (1), 83–92.
- Bitzan, J.D., 2003. Railroad costs and competition: the implications of introducing competition to railroad networks. *J. Transp. Econ. Policy* 37 (2), 201–225.
- Cantos, P., Pastor, J.M., Serrano, L., 2010. Vertical and horizontal separation in the European railway sector and its effects on productivity. *J. Transp. Econ. Policy* 44 (2), 139–160.
- Cantos, P., Pastor, J.M., Serrano, L., 2012. Evaluating European railway deregulation using different approaches. *Transp. Policy* 24, 67–72.
- Casullo, 2016. The Efficiency Impact of Open Access Competition in Rail Markets: The Case of Domestic Passenger Services in Europe. Discussion Paper ITF No. 2016-07.
- Driessen, G., Lijesen, M., Mulder, M., 2006. The Impact of Competition on Productive Efficiency in European Railways. CPB Discussion Paper n° n71.
- European Commission, 2016. Fifth report on monitoring developments of the rail market, Brussels.
- Friebel, G., Ivaldi, M., Vibes, C., 2003. Railway (De) regulation: A European Efficiency Comparison, IDEI Report 3. University of Toulouse, Toulouse.
- Friebel, G., Ivaldi, M., Vibes, C., 2010. Railway (De) regulation: a European efficiency comparison. *Economica* 77 (305), 77–91.
- Growitsch, C., Wetzel, H., 2009. Testing for economies of scope in European railways: an efficiency analysis. *J. Transp. Econ. Policy* 43 (1), 1–24.
- IBM, 2002, 2004, 2007, 2011. Rail Liberalisation Index. Market opening: comparison of the rail markets of the Member States of the European Union, Switzerland and Norway. A study conducted by IBM Deutschland GmbH in collaboration with Prof. Dr. h.c. Christian Kirchner, Humboldt-University Berlin, Brussels.
- Ivaldi, M., McCullough, G.J., 2004. Subadditivity Tests for Network Separation With an Application to U.S. Railroads. CEPR Discussion Paper 4392. CEPR, London.
- Ivaldi, M., McCullough, G.J., 2008. Subadditivity tests for network separation with an application to U.S. railroads. *Rev. Netw. Econ.* 7 (1), 159–171.
- Jensen, A., Stelling, P., 2007. Economic impacts of Swedish railway deregulation: a longitudinal study. *Transp. Res. E* 43 (5), 516–534.
- Kirchner, C. (2002, 2004, 2007, 2011). "Rail liberalization Index. Market opening: Comparison of Rail Markets of the Member States of the European Union, Switzerland and Norway". Study by IBM-Deutschland GmbH and Humboldt University of Berlin.
- Link, H., 2016. A two-stage efficiency analysis of Rail Passenger Franchising in Germany. *J. Transp. Econ. Policy* 50 (1), 76–92.
- Merkert, R., Smith, A.S.J., Nash, C.A., 2012. The measurement of transaction costs – evidence from European railways. *J. Transp. Econ. Policy* 46 (3), 349–365.
- Merkert, R., Nash, C.A., 2013. Investigating European railway managers' perception of transaction costs at the train operation/infrastructure interface. *Transp. Res. A: Policy Pract.* 54, 14–25.
- Mizutani, F., Smith, A., Nash, C., Uranishi, S., 2015. Comparing the costs of vertical separation, integration, and intermediate organisational structures in European and East Asian Railways. *J. Transp. Econ. Policy* 49 (3), 496–515.
- Nash, C., Smith, A.S.H., van de Velde, D., Mizutani, F., Uranishi, S., 2014. Structural reforms in the railways: incentive misalignment and cost implications. *Res. Transp. Econ.* 48, 16–23.
- OECD (2013) "Recent Developments in Rail Transportation Services", OECD Policy Roundtable on Competition Law&Policy, Paris. Document DAF/COMP(2013)24.
- Oum, T.H., Waters, W.G., Yu, C., 1998. A survey of productivity and efficiency measurement in rail transport. *J. Transp. Econ. Policy* 33 (1), 9–42.
- Pittman, R.W., 2005. Structural separation to create competition? The case of freight railways. *Rev. Netw. Econ.* 4, 181–196.
- UNECE (2018). "Railway reform in the ECE Region". United Nations Publication ECE/TRANS/261.
- Wetzel, H., Growitsch, C., 2006. Economies of Scope in European Railways: An Efficiency Analysis. Discussion Paper No. 5. Institut für Wirtschaftsforschung, Halle, Germany.

Further Reading

- Wetzel, H., (2008). "European railway deregulation: the influence of regulatory and environmental conditions on efficiency". Working Paper Series in Economics.
- Mizutani, F., Smith, A.S., Nash, C., Uranishi, S., 2015. Comparing the costs of vertical separation, integration, and intermediate organisational structures in European and East Asian railways. *J. Transp. Econ. Policy* 49 (3), 496–515.

How will Autonomous Vehicles Impact Car Ownership and Travel Behavior

Patrick M. Bösch*, Felix Becker†, Henrik Becker†, Kay W. Axhausen†, *Verkehrsbetriebe Zürich, Zürich, Switzerland; †Institute for Transport Planning and Systems, Zürich, Switzerland

© 2021 Elsevier Ltd. All rights reserved.

Introduction	508
Cost Structures for Automated Vehicles in Switzerland	508
Cost Structures in an International Comparison	511
Survey on People's Intention to Use Automated Vehicles	511
Outlook	512
References	512

Introduction

The expected arrival of automated vehicles (AVs) (SAE level 5) has the potential to drastically change the transport system. Given the 100-year experience with current vehicles, and the incremental acceptance to the new technology then, there is an urgent need to use the chance of the AV arrival to think about the regulatory framework upfront and the local goals for the transport system.

The wider discussion would have to answer at minimum the following questions:

- How do we want the transport system to be operated? At user equilibrium with all its externalities or at system optimum, but with the restrictions this would bring.
- What market structure do we want to allow? An anarchic individualistic ownership of vehicles, an oligopolistic set of fleet owners, or finally a local, maybe municipally owned, monopolist.
- How do we want to provide the transport for the low-income groups? Do we rely on the drop in costs/prices or do we still subsidize services, but how (person-based; service-based).
- Do we still need large vehicles transport large groups of persons (busses, street cars, light rail, heavy rail, etc.) and if so, where and when?
- Finally, what speeds/accessibilities do we want to achieve?

In this contribution we report results that can help in this discussion: comprehensive cost per person kilometer estimates for AV in Switzerland and worldwide; survey results on the likelihood of AV acquisition and use. These numbers and the estimation approach can help in the simulations needed to answer the questions, because the cost estimates will allow us to assess, if and how transport services could be provided sustainably by commercial firms or if subsidies are required.

The rest of the contribution is divided into three parts: first, the presentation of the cost estimation approach, then the international comparison. The third part discusses current survey results on the willingness to forgo a private vehicle in the presence of AV taxis. We conclude with an outlook.

Cost Structures for Automated Vehicles in Switzerland

The work presented by Bösch et al. (2018) provides a picture of the future cost structure of different (automated) transport modes for the Zurich, Switzerland area. It assumes that driverless vehicle technology is user ready and has already reached full market penetration.

First, the cost structures of private cars, taxi services and mass transit public transport (PT) (trains and buses) were reconstructed for the Zurich area. Where available, real data were used as a source. If not available, expert estimates were made based on aggregated data or background information. The cost effects of electrification and automation were estimated with a bottom-up approach.

For detailed cost factors (e.g., fuel or insurance) the expected changes were estimated based on literature or educated assumptions. These estimates were concluded with a calculation of variable and fixed provider costs and user prices per vehicle kilometer.

To estimate the resulting cost and price structures, however, the expected usage of the different services is required. It enables the distribution of the fixed costs, variable cost of support kilometers driven, and of variable costs per vehicle kilometer on the users. Based on the Swiss microcensus, the national travel diary survey, (Federal Statistical Office (FSO) and Federal Office for Spatial Development (ARE), 2012), and different usage studies, these values were estimated for the Zurich area. This allowed to calculate the resulting provider cost and user price per passenger kilometer.

For the already existing modes (private cars, taxi services, and mass transit public transport), these resulting prices were validated against current prices. While for private cars and public transport the estimated and observed prices matched well, for taxis it could be

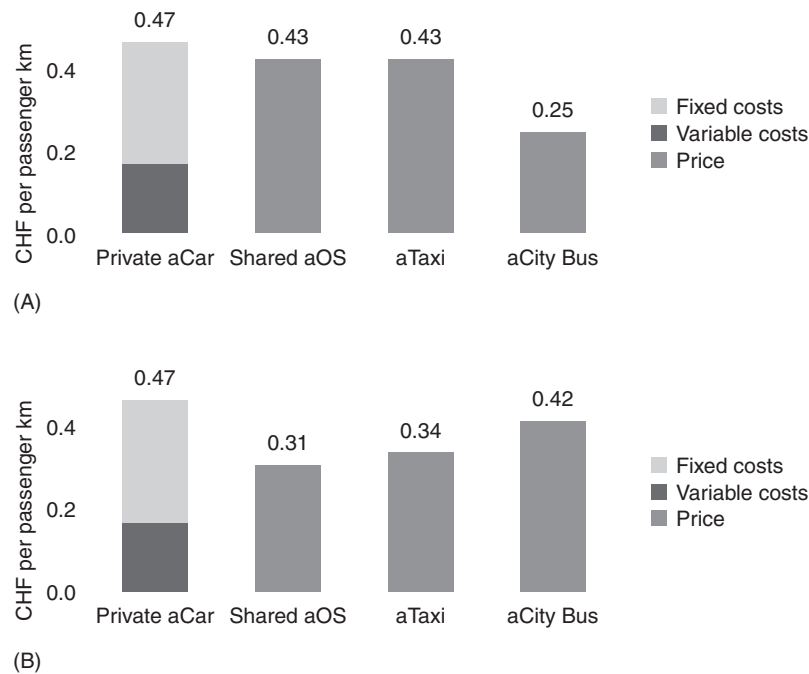


Figure 1 Future competitive situation with autonomous vehicle technology in an urban and a regional setting. (A) Future competitive situation—urban setting. (B) Future competitive situation—regional setting. Source: Bösch et al. (2018)

shown that for Zurich Uber prices are likely too low (i.e., they are likely subsidized by Uber or the drivers are not considering their full cost) and that the taxi prices are likely at the upper bound.

Results show that private cars still represent an attractive option in the era of autonomous vehicles, since out-of-pocket costs for the user (0.17 CHF^a/km, i.e., \$0.27/mile, cf. Fig. 1) are lower than for most other modes (and could be even lower if the user, e.g., produced its own electricity). This is in the range of the \$0.15/mile Burns et al. (2013) found for shared AVs and the \$0.16/mile Johnson (2015) found for purpose-built automated ride-sharing vehicles, but they neglected important cost factors—for example, cleaning—found here to amount to 30%–50% of the cost of such a service. Compared to the assumption by Fagnant and Kockelman (2015) of a \$1.00/mile price for a shared AVs or the only 30% reduction to today's taxi prices suggested by Wadud (2017), even the full cost of private vehicle ownership might be competitive.

In fact, buying an autonomous vehicle could be regarded as an investment in a private mobility robot, which can be used both for chauffeur services and for errands; this choice will therefore be even more attractive than a conventional vehicle. Hence, it can be expected that a substantial number of people would value the private use of a mobility robot and will agree to pay the associated premium. Additionally, traditional car manufacturers are strongly motivated to maintain the current emotional connection many people have to their cars (Fraedrich and Lenz, 2015). In conclusion, even as low costs for shared AVs, as estimated by Burns et al. (2013) and Johnson (2015), might not be low enough to end the reign of the private car.

In contrast to private vehicle ownership, the results suggest that current line-based public (mass) transportation will probably be subject to adjustments beyond its automation. Although its operation will become cheaper, ride-sharing schemes and other new forms of public transportation will emerge as new and serious competition. The results also indicate, however, that the situation is not as clear as suggested by Hazan et al. (2016), for example. If full cost of shared vehicles is considered, combined with the low average occupancy achievable with ride sharing even in an urban setting (International Transport Forum, 2015), mass transit public transport is still competitive—especially for urban settings and high-demand routes—and even more so, if low-demand routes can be served with more flexible services, thus increasing average occupancy. The situation gets even more interesting if one considers that today's form of public transportation in Switzerland receives subsidies for approximately 50% of its operating costs (Laesser and Reinhold, 2013). In principle, with autonomous driving technologies and constant levels of subsidies and demand, operators would be able to offer line-based mass-transit public transportation for free at the point of use.

Shared aOS vehicles are seen by some researchers as an ideal first-and-last-mile complement of mass-transit public transportation. Offering direct point-to-point service for single travelers, they promise short access and travel times. However, they are not much cheaper to operate than shared midsize vehicles (Fig. 1). Policies aside, it is therefore more likely that fleet operators opt for a homogeneous midsize or van fleet to serve individual travelers, as well as smaller groups. One vehicle size functions well, particularly because—assuming acceptably low waiting times—few relations see a demand as low as only one traveler per time interval.

^a 2016 exchange rates (ER) (Organization for Economic Co-operation and Development, 2017a) and purchasing power parities (PPP) (Organization for Economic Co-operation and Development, 2017b): Swiss Francs (CHF)/US\$: ER 1.0, PPP 1.2.

Fleets of shared AVs (ride sharing or aTaxis) are another mode often proposed as the new “jack of all trades” in transport: offered at such low prices that they will replace every known mode. As this study shows, however, this picture changes if full costs, including overhead, parking, maintenance, and cleaning, are considered. Just these factors, neglected in most studies on the topic so far, contribute more than two-thirds of the total cost of aTaxis. It is thus no surprise that our estimates of the costs of shared services are substantially higher than those in previous work (e.g., Burns et al., 2013; Hazan et al., 2016; Johnson, 2015; Stephens et al., 2016). Because shared AVs are still very economical to operate and OEMs might also have an interest in shared services (Abhishek et al., 2016), these business models might still have a bright future. Shared fleets of aTaxis also have the advantage that the usage can be controlled and (geographically) restricted—especially compared to privately owned vehicles. This minimizes the liability risk for OEMs and mobility providers as long as AVs are not yet an established and proven technology.

Ride sharing might lower prices if high average occupancies can be achieved. High occupancies, however, come with more detours, longer individual travel times, and more strangers in the same intimate environment of a car. Especially the last factor might be an obstacle to the success of midsize or van ride-sharing fleets. People usually prefer the privacy of their own vehicle to the anonymity of a bus ride, and many find sharing a ride-sharing vehicle with strangers burdensome (Schmid et al., 2016). In this respect, more research is required to better understand customer preferences and design ride-sharing vehicles accordingly, because, if successful ride sharing with AVs will definitely have a number of benefits (Friedrich and Hartl, 2016; International Transport Forum, 2015).

A further challenge for shared services will be to find a solution to maintain vehicle cleanliness. The analysis mentioned earlier revealed that even with low cleaning frequencies and costs, cleaning is the single largest contribution to the operating cost of autonomous (individual) taxi schemes. Any higher level of soiling may even endanger their competitiveness through high costs or low service standards.

The work by Bösch et al. (2018) is—to date—the most comprehensive approach to estimating operating costs of future autonomous transport services. It includes several aspects previous studies have overlooked, or assumed negligible (e.g., overhead costs or cleaning) and it draws a clearer and/or different picture of the future transportation system than earlier works (Burns et al., 2013; Fagnant and Kockelman, 2015; Hazan et al., 2016; Johnson, 2015; Litman, 2015; Stephens et al., 2016; Wadud, 2017). While this detailed approach reveals new insights, it also comes with some limitations.

For example, it is likely that vehicle automation will change the mobility behavior and the demand for certain kinds of infrastructure, such as parking, which could require the government to take measures to counterbalance negative effects. The impact of governmental policy measures has been ignored in here; attention to policies was restricted to those already implemented today. Extended services, such as non-driving personnel in the vehicles or premium vehicles, have been ignored too. While such measures, policies, and extra services are expected to substantially influence price structure and thus the competitive situation, their cost will probably be passed along to the user. It follows that they would not substantially influence the basic cost structure of different services for the provider.

While the cost structures of cars, whether private or as shared vehicles, could be analyzed with a high degree of detail, the overhead costs of shared mobility services are well-kept secrets within the respective companies. Accordingly, estimates of the overhead cost of shared services and total costs of public transport services as well as the proposed savings must be treated with caution.

The same applies for the cost effects of new technologies. Electric cars are already on the market and thus estimates of the cost effects are reliable and well grounded. The effects of automation, however, are uncertain. Admitting this, the approach taken identifies different factors resulting in the observed total cost. Then, the effect of autonomy is estimated for each cost factor separately, resulting in more precise and reliable estimates. Where available, these estimates were based on values reported in literature. Even such detailed estimates’ accuracy, however, is questionable. Correspondingly, usage values are based on current market situation and price structures. If radically lower costs are assumed, these values are likely to change substantially (e.g., lower occupancy or longer trips). Once travel behavior impacts and usage patterns of such schemes become clearer, the framework introduced in this research can be used for a scenario-based analysis of their cost structures and will probably yield more accurate results. Until then, the reader should be aware of these limitations when interpreting the results presented here.

In conclusion, it is clear that fleets of shared autonomous vehicles may become cheaper than other modes in relative terms, but, in absolute numbers, the difference will be small. Thus, there will still be competition by other modes and even by private car ownership, which may well persist beyond the dawn of autonomous vehicle technologies by offering the luxury and convenience of a personal mobility robot. On the other hand, this research was able to confirm expectations (e.g., by Meyer et al., 2017) that conventional forms of public transportation may face fierce competition in the new era. Importantly, this research also revealed that the success of shared AV fleets may well depend on a factor which has been previously ignored—cleaning efforts. According to the findings, developing viable business models for shared AV fleets will entail solutions to require that customers behave appropriately while on board (e.g., video observation of passengers, or a confirmation check by the next user on the condition of the car to identify irresponsible passengers) and/or to clean and repair vehicles efficiently and at low costs.

Based on an exclusively cost-based approach, the work presented here was able to shed some light on the potential of different future modes. It is clear, however, that many open questions remain and that further research is required. For example, the identified cost numbers represent average numbers. Individual users might use a mode more or less than assumed here and thus the resulting price per passenger kilometer might change substantially for them. For example, two-thirds of the price of a private car (0.30 of total 0.47 CHF/passenger kilometer) are fixed costs. If the car is used more often than assumed here, these fixed costs will become lower per passenger kilometer, and, if the car is used less often, it will be more expensive.

A similar observation applies to public transport vehicles. The average occupancy of a city bus today is 22.42 passengers (Federal Office of Transport, 2011), and the threshold between a minibus and a city bus is at 21 passengers. This means that in average, in the future, a city bus is still the most cost-effective mean of transport for cities. Since this is in average only, lines and usage patterns need to be differentiated spatially and temporally. A detailed analysis is required to determine where this threshold of 21 passengers is crossed, and where much smaller vehicles will be sufficient.

Another important limitation is that the actual mode choice is determined not only by cost, but also substantially by travel time and comfort (value of time), as well as other factors such as the perception of transfers, waiting times, etc.—all of which are not investigated here. Therefore, actual implementations of the proposed schemes need to be tested in a field trial or simulation approach to better understand the size of the respective market segments in a realistic environment.

Cost Structures in an International Comparison

Vehicle automation and electrification substantially change the economics of mobility services. Compared to conventional vehicles, they overcome the need for a driver, reduce marginal cost (such as fuel consumption and depreciation), but their acquisition cost is usually substantially higher (due to technology and batteries). Given the large variation in purchasing power and salaries, the net effects will depend on the local context. Hence, a simplified version of the approach by Bösch et al. (2018) was applied to 17 locations in 15 countries across all continents. The results provide insights into the potential global market for AVs.

It is assumed that any future transport service will fall into one of the four following categories:

- Private vehicles,
- Individual taxis,
- Pooled taxis/dynamic on-demand public transport, and
- Line-based public transport.

Here, *private vehicle* describes the use of vehicles similar to today's private cars, whereas *individual taxi* will be like current taxi or ride-hailing services, just without the driver. *Pooled taxis* and *dynamic on-demand public transport* will provide point-to-point trips (potentially involving a short access and egress walk), but the vehicle will be shared with a small number of other passengers for at least parts of the trip. Midsize vehicles or vans can be expected to be used for this service. Finally, *line-based public transport* is comparable to today's public transport operated with large buses, trams, or even trains. All services are considered independent of any contractual frameworks or subsidies.

The results clearly indicate that, in general, vehicle automation will have a much more profound impact on cost structures of all services than electric propulsion, which may only allow cost savings of up to 10%. As in the Swiss analysis, taxi costs are reduced substantially by vehicle automation, whereas effects on public transport are lower. For private cars, cost reductions are only marginal. In some cases, higher acquisition cost due to the new technology cannot even be compensated by savings in marginal cost (leading to higher overall cost).

Differences between the different locations can be mostly characterized by the relative importance of costs for assets (vehicles) and salaries, respectively. In developing countries, current production costs are mostly driven by the former, so that additional investments in technology (sensors, processors, batteries, etc.) have a much stronger impact. In turn, in developed countries, labor costs are the key component, outweighing costs for (acquisition and maintenance of) the vehicle. Consequently, in many industrialized countries production costs for pooled taxi services will reach a level comparable to automated public transport. Individual taxis will still be more expensive than public transport, but, in absolute terms, the difference may not be substantial. In contrast, in many developing countries, the price difference between public transport and taxis will still be substantial.

A more quantitative analysis of the cost drivers shows that the median net per-capita income can already explain more than 50% of the variation in production costs of conventional services. Moreover, taxis are about a factor 4 more expensive than buses and difference increases with growing income levels. However, in an automated-electric regime, the absolute difference in cost is much lower and independent of the respective income level. This means that the absolute cost of transportation services will be almost the same across the globe, to the advantage of higher-wage countries, where automated taxis may become an everyday service. In lower-wage countries, disruptions can be expected to take much longer to set in. Hence, both development of new business models as well as policy efforts to manage transport demand can be expected to take place in few, higher-income countries first.

Survey on People's Intention to Use Automated Vehicles

AVs are likely to play a crucial role in transport systems once the technology is developed and regulation has been adjusted. This can be reasoned with their higher comfort compared to currently available modes and an increased safety if the technology outperforms human drivers. It is therefore important to continuously monitor the public's opinion on that development and to investigate how people intend to change their mobility behavior and tools early. A survey conducted in 2018 intended to answer these questions for the Canton of Zurich, Switzerland. It focused on vehicles with SAE-automation-level 5, meaning that they are able to perform empty rides.

In line with predictions from car manufacturers such as BMW and Ford, the survey distinguishes between two different development steps. In the first step, it is assumed that vehicles are only available on-demand and cannot be bought privately. In particular, the AVs are either available as taxi automated vehicles (TAVs), that is, sequential car sharing, or pooled automated vehicles (PAVs), whose service involves simultaneous car sharing (ride sharing). Due to detours and higher occupancies, PAVs are assumed to be slower and cheaper on average. In the second step, respondents are also given the opportunity to buy AVs.

The survey has three paper-based stages and follows a pivot design. In the beginning, respondents are asked about their portfolio of mobility tools (such as cars, bikes, and season tickets) and information about a short (<50 km) and long trip (≥50 km). Afterwards they are asked how they intend to change their portfolio of mobility tools once on-demand AVs are available (development step 1). They further complete a mode-choice questionnaire for this step. The process is repeated for development step 2, which includes on-demand services as well as private AVs.

The availability of the alternatives is adjusted to the trips and mobility tools and the base levels of the attributes reflect the real choice situation. Costs for AVs are based on the cost calculations presented earlier. The possible alternatives are classic public transit (PT), public transit with an SAV feeder (PTAV), bike, walk, conventional car, SAV, and PAV.

If only on-demand AVs are available, the number of conventional cars per household drops by 30%. However, if people are also given the opportunity to buy AVs, no statistically significant effect can be observed regarding the number of cars. In contrast, half of the conventional cars were replaced with AVs.

Price changes for on-demand services in the bandwidth of ±30% do not have a significant influence on people's decisions regarding their mobility tools. For public transport subscriptions, no significant changes among the development steps could be observed.

The results from the mode-choice experiments show that people mainly switch from the conventional car to automated modes. Nevertheless, the conventional car remains the dominant mode if only on-demand AVs are available to travelers. PAVs and TAVs are each used for around 7% of the short trips. Once the respondents are able to use private AVs, conventional cars and private AVs are each used for approximately 30% of the trips. The public transport usage remains relatively stable.

The model results further show that people are less inclined to pay for reducing the travel times of AVs compared to the conventional car. While the value of travel times savings of the conventional car amounts to 30 CHF per hour, the respective values of private AVs and TAVs amount to 24 and 20 CHF per hour, respectively. One reason for lower VTTS values is an increase in comfort, since the driver can perform other tasks while being chauffeured. At 12 CHF per hour, the VTTS for public transport are still lower.

To conclude, the respondents in this sample are willing to use AVs once they become available. At this moment, the majority of those intending to use AVs are still inclined to use their own vehicle rather than sharing solutions.

Outlook

This first result indicates that commercial AV taxi-based services might have a hard time between the comparable costs of privately owned AVs and the needs to transport large numbers of persons in dense urban areas; unless, the capacity gains due to the AVs are so large, that the increasing number of vehicle kilometers can be accommodated at reasonable speeds (Loder et al., 2019).

It is clear that the work on answering the questions raised earlier has only just begun. The cost estimates will change as technology matures. The assessments by the potential users and vehicle ownership will evolve as the market continues to change.

Still, the need for a comprehensive simulation approach is clear.

References

- Abhishek, V., Guajardo, J.A., Zhang, Z., 2016. Business models in the sharing economy: manufacturing durable goods in the presence of peer-to-peer rental markets. Available from: <http://dx.doi.org/10.2139/ssrn.2891908>.
- Bösch, P., Becker, F., Becker, H., Axhausen, K., 2018. Cost-based analysis of autonomous mobility services. *Transp. Policy* 64, 76–91.
- Burns, L.D., Jordan, W., Scarborough, B., 2013. Transforming personal mobility. Technical Report. The Earth Institute, Columbia University, New York.
- Fagnant, D., Kockelman, K., 2015. Dynamic ride-sharing and optimal fleet sizing for a system of shared autonomous vehicles. Paper presented at the 94th Annual Meeting of the Transportation Research Board, Washington, DC.
- Federal Office of Transport, 2011. Key figures for regional person transport with access function. Available from: <https://www.bav.admin.ch/bav/de/home/das-bav/aufgaben-des-amtes/finanzierung/finanzierung-verkehr/personenverkehr/rpv-mit-erschliessungsfunktion/rpv-kennzahlen.html>.
- Federal Statistical Office (FSO) and Federal Office for Spatial Development (ARE), 2012. Mobility in Switzerland: Results of the microcensus mobility and transport 2010. Technical Report. Swiss Federation, Neuchâtel and Berne.
- Fraedrich, E., Lenz, B., 2015. Taking a Ride, Hitching a Ride: Autonomous Driving and Car Usage. In: Maurer, M., Gerdes, J.C., Lenz, B., Winner, H. (Eds.), *Autonomes Fahren*. Springer-Verlag, Berlin, pp. 687–708.
- Friedrich, M., Hartl, M., 2016. MEGAFON—Model results of shared autonomous vehicle fleets in local public transport Research Report. Institute for Road and Transport Science, University of Stuttgart.
- Hazan, J., Lang, N., Ulrich, P., Chua, J., Doubara, X., Steffens, T., 2016. Will autonomous vehicles derail trains? Technical Report. The Boston Consulting Group, Boston, Massachusetts.
- International Transport Forum, 2015. Urban mobility system upgrade—how shared self-driving cars could change city traffic. Technical Report. OECD, Paris, France.
- Johnson, B., 2015. Disruptive mobility. Research Report. Barclays, London.
- Laesser, C., Reinhold, S., 2013. Financing public transport in Switzerland: What is paid by the user, what by society? Technical Report. Schriftenreihe SBB Lab, St. Gallen.
- Litman, T., 2015. Autonomous vehicle implementation predictions. Technical Report. Victoria Transport Policy Institute, Victoria.

- Loder, A., Bressan, L., Dakic, I., Ambühl, L., Bliemer, M.C.J., Menendez, M., Axhausen, K.W., 2019. Modeling multi-modal traffic in cities using the 3D Macroscopic Fundamental Diagram. Paper presented at the 98th TRB Annual Meeting, Washington, DC.
- Meyer, J., Becker, H., Bösch, P.M., Axhausen, K.W., 2017. Autonomous vehicles: the next jump in accessibilities? *Res. Transp. Econ.* 62, 80–91.
- Schmid, B., Schmutz, S., Axhausen, K.W., 2016. Explaining mode choice, taste heterogeneity and cost sensitivity in a post-car world. Paper presented at the 95th Annual Meeting of the Transportation Research Board, Washington, DC.
- Stephens, T., Gonder, J., Chen, Y., Lin, Z., Liu, C., Gohlke, D., 2016. Estimated bounds and important factors for fuel use and consumer costs of connected and automated vehicles. Technical Report. National Renewable Energy Laboratory, U.S. Department of Energy, Golden, Colorado.
- Wadud, Z., 2017. Fully automated vehicles: a cost of ownership analysis to inform early adoption. *Transp. Res. Part A Policy Pract.* 101, 163–176.

Policy Instruments to Reduce Carbon Emissions from Road Transport

Davide Cerruti, Cristian Huse, Centre for Energy Policy and Economics, ETH Zürich, Zürich, Switzerland; Department of Economics, University of Oldenburg, Oldenburg, Germany

© 2021 Elsevier Ltd. All rights reserved.

Overview	514
Fuel Taxes	515
Standards	516
Vehicle Taxes	516
Rebates and Feebates	516
Vehicle Scrappage Schemes	517
Information Programs	517
Congestion Pricing and Driving Restrictions	517
Mileage (or Utilization) Tax	518
Modal Transport/Public Transport	518
Afterword	518
Biographies	519
References	519

Overview

Economic agents typically do not take into account the social cost of their choices or actions, that is, the cost they impose on society. When it comes to transport-related choices this often translates, for instance, into using private rather than public means of transportation, larger rather than smaller vehicles, more rather than less powerful engines.

In the 1920s, Economist Alfred Pigou introduced a concept that became known as Pigouvian tax, which can be seen as an attempt to discipline economic agents. Specifically, if an economic activity generates costs to third parties (external social costs), regulators should introduce a tax equal to the external social cost of such activity, or externality. This arguably simple framework is the starting point of most economic analyses of environmental problems to date.

Determining the level of the tax is, however, a nontrivial task for a number of reasons. First, one has to identify the factors generating external social costs. This could include carbon emissions from transportation, but also local pollution generated by combustion engines and tires, costs of congestion, accidents, noise and visual pollution caused by driving, wear-and-tear of roads, etc. In recent years, it has been identified that, in some jurisdictions, some external costs have been given more importance than others, which may benefit one technology in detriment of another. Concretely, the EU legislation has historically been relatively strict in terms of CO₂ emissions but more lenient with regard to local pollutants, which was documented to benefit diesel in detriment of gasoline vehicles (Miravete et al., 2018) and thus European rather than non-European manufacturers.

Second, one has to correctly quantify each of the earlier externalities. Recent evidence of the manipulation of driving tests led credence that measurement is potentially of first-order importance, especially in diesel vehicles, which led to the recent redesign of test cycles in Europe.

Third, one has to assign monetary values to each of the earlier factors. For instance, local pollution generated today might have long-lasting health effects current research was not yet able to identify.

Fourth, the passing on of taxes onto prices depends on both the competitive structure and on the characteristics of the supply and demand profiles of a particular market. In particular, firms may or may not pass on the cost of a tax to consumers.

Fifth, it assumes that consumers correctly quantify and take into account in their decisions the tax levied on their activities or the products they purchase (Huse and Koptuyug, 2018).

Despite the earlier limitations, the Pigouvian framework provides the basic tool kit used by economists and regulators in the analysis of environmental problems, also in the transport sector. If carbon emissions are the only external cost of transport imposed on society, then the Pigouvian tax is efficient (or first-best), in that it imposes on consumers the entire external costs their activity generates to society. In particular, a fuel tax is a Pigouvian tax in this specific case, as introducing it tackles both utilization (kilometers driven) and fuel consumption (how many liters a vehicle requires to drive 1 km) of a driver-vehicle combination.

An additional factor making taxes appealing is that they are cost-effective, in the sense that they satisfy the Least-Cost Theorem, an important result in environmental economics. That is, taxes are able to attain a given environmental target at least cost to the society, which results in an economically efficient allocation of resources. In more concrete terms, if a policy instrument is least cost, then it satisfies the so-called equi-marginal principle, according to which the cost of abating pollution is equalized among all polluters—in particular, economic agents for which abating pollution is cheaper will abate more than those for which abating pollution is more expensive until the cost of abatement of all agents equalize.

If consumers are not able to identify and/or quantify the extent of the taxation they are subject to, then they will undervalue the fuel tax and overutilize transportation. As a result, this is one channel through which the fuel tax loses its first-best property. Therefore, additional policy instruments are called for, such as standards or vehicle taxes. Although they do not tackle utilization, standards have an impact on the fuel economy (or fuel consumption) of a vehicle, making less efficient vehicles more expensive than their more efficient counterparts. Circulation taxes operate in a similar way, but while standards typically influence only the purchase of new vehicles, circulation taxes are charged yearly and thus affect both the new and the secondhand car market.

The distinction between new and used vehicle markets is crucial since while the former receives considerably more attention, it is only a small (typically single-digit) share of the latter. Thus, one has always to bear in mind the scope of the policy instruments aiming to tackle carbon emissions at the risk of said instruments not having a meaningful impact.

Recent events in the transport sector have brought back attention to local pollutants and shed light on a number of alternative policy instruments tackling pollution in transportation. If local pollutants are not addressed when determining the level of the fuel tax, then alternative policy instruments—especially those targeting diesel vehicles—are called for. These include congestion pricing and driving restrictions, which can occur in various forms.

Finally, considerations about optimal policies have to take into account the political constraints that often prevent a straightforward implementation of those measures. While interventions adopted by governments might not represent the most efficient practice from an economic viewpoint, they might be the only politically feasible solutions.

In what follows, we briefly describe the main policy instruments that can be used to mitigate carbon emissions and their applications.

Fuel Taxes

If the externality policy-makers aim to address is CO₂ emissions, more generally, greenhouse gases, a fuel tax typically charged on a per-liter or gallon basis is a Pigouvian tax on emissions and can thus achieve the first-best outcome in the absence of consumer undervaluation of energy (fuel) costs. Otherwise, to achieve a first-best policy one needs to combine a fuel tax with a standard and a car tax, at least in the case of a representative agent model. For alternative externalities, other instruments would be first-best. For instance, road pricing would address the congestion externality, a per-kilometer tax based on vehicle type might be first-best to address the accident externality and road wear-and-tear (Parry and Small, 2005). In case local pollutants are also considered, the design of policies becomes more complex.

Fuel taxes contribute to the reduction of vehicle CO₂ emissions in two ways: first, they induce drivers to reduce vehicle utilization; second, they push consumers to choose more energy efficient vehicles (Li et al., 2009). The timing of these two effects is different though. While adjustment in kilometers driven can occur within a short time frame from a tax change, the effect on the composition of the vehicle fleet is slower, as generally the increase in driving cost does not justify the immediate replacement of a vehicle. Nevertheless, the presence of two channels of emission reduction rather than only one makes fuel taxes often a better solution than other types of environmental regulation addressing only fuel efficiency, like standards or vehicle taxes.

The way consumers respond to fuel taxes when choosing a vehicle depends also on whether drivers understand correctly the role fuel price has on driving costs—which are also based on fuel economy and kilometer driven. Hence, the question whether a change of 100 in driving costs would be considered equivalent as a change of 100 in vehicle purchase price. Recent evidence on perception of fuel costs and fuel economy shows in general modest to no undervaluation, suggesting that consumers would be able to respond optimally to tax changes.

Compared to fluctuations in fuel prices due to supply-side shocks, fuel taxes have two relevant characteristics. First, any change is often advertised by the media and thus is more salient to consumers. Second, the taxes are more persistent over time. One important implication is that drivers might respond to changes in fuel taxes more strongly than an equivalent changes in pretax fuel price.

While reducing CO₂ emissions, fuel taxes generate considerable revenue. That led to assume the presence of a “double dividend,” in which not only fuel taxes would reduce externalities from carbon emissions (first dividend), but they would also allow to reduce other distortionary taxes like the income tax or to fund public goods (second dividend). However, the interaction between the environmental tax and other preexisting distortionary taxes can also exacerbate the efficiency loss caused by the latter. Whether this tax interaction effect dominates or not the effect of the second dividend depends on the characteristics of the fiscal system, and has important consequences on the optimal level of the fuel tax—whether lower or higher than the marginal damage linked to CO₂ emissions.

The practical implementation of fuel taxes, and carbon taxes in general, can be subject to significant political constraints. An important reason is that drivers tend to be more aware or attentive of fuel taxes than other vehicle policies (Huse and Koptuyug, 2018). Fuel taxes are often also introduced for other reasons, such as to fund infrastructure building and maintenance or as a general source of public funds. Furthermore, specific categories can be more affected than others. The impact over different income groups depends on the availability of automobiles among poorer households, and in general on the share of expenditure on fuel over the total household budget. Drivers living in urban areas can switch more easily to other transportation modes than those living in the countryside. Nevertheless, tax revenue can be used to mitigate those distributional concerns, for instance, through a lump-sum rebate, or through a cut of the income tax marginal rates (Bento et al., 2009). While politically challenging to introduce or change, fuel taxes represent one of the most direct and effective measures to deal with carbon emissions.

Fuel taxes are widespread worldwide, being adopted from EU countries to China, India, the United States, and several emerging economies. Two things are of notice, however, when looking at the levels of fuel taxes worldwide. First, according to OECD statistics for 2018, taxes on gasoline are higher than their diesel counterparts in the vast majority of the countries surveyed, sometimes substantially more so. Second, the value of the fuel tax varies considerably across countries.

Standards

Standards can equivalently be imposed on CO₂ emissions, fuel consumption, or fuel economy, as these are related quantities for a given motor fuel, given the physical relation between a vehicle's CO₂ emissions (measured in, say, gCO₂/km) and its fuel consumption (measured in, say, liters/km).

Standards impose a maximum value on the variable being regulated (a target) and, when binding, act as an implicit, revenue-neutral tax on the variable of interest (Knittel, 2012; Anderson and Sallee, 2016). The inefficiency of standards occurs due to the rebound effect. Additional externalities, such as congestion, will favor fuel tax.

That is, standards incentivize the purchase of more efficient cars but the resulting savings tend to generate an increase in vehicle utilization; dimension standards do not address. Intuitively, this can be shown as follows. Assume the only transport externality is CO₂ emissions, which are measured in grams of CO₂. A fuel tax influences both vehicle utilization (measured in vehicle-kilometers-year, say) and emission rates, whereas a standard affects only the latter. Thus, while the incidence of fuel taxes falls on all determinants of CO₂ emissions, that of standards fall only on part of them, which results in their inefficiency.

Standards apply to new vehicles and might or might not be passed on to consumers by carmakers, partly because of the highly concentrated market structure in the auto industry. Nevertheless, they incentivize the improvement of the regulated characteristic due to the change in relative prices between efficient and inefficient vehicles.

While traditionally standards applied uniformly across products, a number of countries have recently adopted attribute-based standards, according to which a target varies according to some attribute of the vehicle such as its weight or footprint (the rectangular area defined by a vehicle's tires). Typically, larger or heavier vehicles face a laxer standard, which implicitly subsidizes vehicle weight or footprint. Nevertheless, according to the US Energy Information Administration, attribute-based standards have been introduced in most of the major vehicle markets worldwide, including the United States, China, Japan, and Europe.

Vehicle Taxes

A vehicle tax is another policy instrument designed to tackle emissions-related externalities. To clarify matters, it is important to distinguish between a registration tax and a road (or vehicle circulation) tax (also known as vignette in some countries). A registration tax applies only to the first registration of a vehicle. Some prominent adopters are Spain, and South Africa, where the registration tax is set according to the CO₂ emission rate of the vehicle. By contrast, a road tax is to be paid every year by vehicle owners provided their vehicles are circulating (some countries allow vehicles to be periodically de-registered, e.g., during Winter). Thus, while a registration tax has an impact only on new vehicles, a road tax shapes the composition of the whole fleet. This distinction is important since a road tax has been found to impact vehicle retirement decisions and consumer choice in the used vehicle market. In some cases though, an introduction or a change of a road tax applied only to new registrations. That implied that older cars were still subject to previous regulations, often more lenient toward inefficient vehicles.

The vehicle tax is an explicit tax effectively paid by vehicle owners, just as it happens with the fuel tax and in contrast with standards. Typically, the tax is calculated based on vehicle attributes such as weight, fuel type CO₂ emissions, or sales price. This allows regulators some leeway in treating technologies differently. While such characteristics are already correlated with carbon emission rates, some countries based their vehicle tax system in part or fully on the CO₂ emission rate (gCO₂/km) of vehicles.

As it happens with standards, a vehicle tax does not influence utilization, and the marginal cost of driving decreases in fuel economy. For this reason, vehicle taxes are considered less cost-effective and less efficient than fuel taxes in reducing carbon emissions from the transport sector (Cerruti et al., 2019a). Nevertheless, they are often easier to implement from a political point of view, and their use is justified when consumers undervalue fuel costs—and thus fuel taxes—compared to other costs of buying and driving a car. Vehicle taxes directly or indirectly linked to CO₂ emissions have played an important role when it comes to technology adoption (gasoline vs. diesel vehicles) in the European car market, partially explaining the market share commanded by diesel vehicles in Europe (Klier and Linn, 2015; Miravete et al., 2018). Registration or sales taxes are widespread around the world, including Japan, China, India, South Africa, and most European countries.

Rebates and Feebates

Another monetary instruments available to governments to address environmental externalities are subsidies on the adoption of low-emissions vehicles, under the form of a discount on the purchase price (rebate). A rebate can be combined with a vehicle tax to form a feebate (fee + rebate), which often works in a self-financing manner. In practice, feebates are policy instruments used to alter the relative prices of high- and low-energy efficiency products. Fees are charged to vehicles emitting more than a given threshold

emissions level (the pivot point), whereas rebates are given to vehicles emitting less than said level (Adamou et al., 2014; Huse and Lucinda, 2014). Because feebates include a sales tax component, the considerations made for vehicle taxes in the previous section largely apply in this context as well. Feebates have been adopted in some European countries, noticeably France and Sweden. Rebates on low-emission vehicles are more widespread worldwide. Both types of policy have historically received substantial attention given the values involved, be it on a per-vehicle basis or in the aggregate.

When introducing rebates for low-emissions vehicles, policy-makers should pay special attention to the impact of free-riding and low pass-through rates. Free-riding occurs when part of the consumers benefiting from the rebate would have chosen a low CO₂ vehicle even without incentives, or when consumers take advantage of political or administrative borders. Such type of behavior decreases the cost effectiveness of environmental programs. In case of low pass-through rate instead, part of a rebate on final vehicle price might not benefit consumers, as car dealers and manufacturers would increase purchase prices accordingly.

Vehicle Scrappage Schemes

Governments can also incentivize the early retirement of old, inefficient, cars. Vehicle scrappage schemes offer a rebate on the purchase of a new energy efficient vehicle if a buyer trades in her old vehicle (Jacobsen and Van Benthem, 2015). After the Great Recession several countries (e.g., the United States, EU, Japan, and China) adopted these programs both as an environmental measure and as a temporary stimulus for the economy. The effectiveness of these incentives as a measure to reduce vehicle carbon emissions is still under debate. The short duration of the scrappage subsidies can exacerbate the free-riding problem. In fact, people might respond to the incentives simply by anticipating by a few months the purchase of a low CO₂ vehicle. If so, the environmental gains from the program will be likely modest compared to its costs.

Information Programs

In order for decarbonization policies to work, proper consumer information is essential. That means providing consumers with information on policy measures and vehicle characteristics, including carbon emissions, presenting such information in a clear and understandable way, and making sure that individuals are able to incorporate the information in their decision-making process. An advantage of information interventions is that they are typically less expensive to implement than other policy instruments. In several countries (e.g., the United States, EU, Chile, Japan, South Korea, and Australia), new cars on sale must display in a visible way the fundamental characteristics of the vehicle, usually contained in an “energy label.”

The way information is framed is also important. The CO₂ emissions rate (g/km) is a precise metric, but lacks a reference point for consumer to understand how a car fares in comparison to other vehicles. For this reason, some countries present CO₂ emissions according to letter rating framework (e.g., from A to G). While easier to understand, this simplified design might lead to misinterpretations, in particular when the rating is also based on other vehicle characteristics such as weight (i.e., an “attribute-based label”). Furthermore, information on CO₂ emissions and fuel economy, measured through laboratory testing, is often imprecise. Switching to better testing methodologies, like the recent replacement in the EU of the NEDC with the WLTP test, can attenuate the gap between real-world and measured values.

Although information programs are now widespread, they typically apply only to the new vehicle market. In fact, the use of outdated efficiency ratings in used vehicles, which tend to be upward-biased (e.g., overstating the efficiency of a used vehicle), might provide consumers with wrong information and thus hinder optimal decision-making.

Finally, providing comprehensive information is important not just for vehicle characteristics, but for environmental policies as well. In some instances, people might not be aware, or not have a full understanding, of the incentives in place for switching to low-emission vehicles (Cerruti et al., 2019b). For this reason, the implementation of vehicle taxes and rebates should coexist with appropriate information campaigns to clarify both the existence and the scope of those policies.

Congestion Pricing and Driving Restrictions

The policies considered so far have been implemented at the national or regional level, but they are generally precluded to local governments. Instead, municipalities have applied congestion pricing and restrictions to vehicle circulation in selected areas. In the vast majority of cases the goal of those measures is to reduce traffic congestion or local pollution, but by reducing traffic they might have an effect on carbon emissions as well.

While congestion pricing consists of charging vehicle owners a (potentially time-varying and vehicle-dependent) fee to get access to a certain area, such as the center of a city, driving restrictions only allow a subset of the vehicles to enter such area, sometimes called a low-emission zone. One variant is to restrict vehicles based on occupancy, so only vehicles with at least a number of passengers are granted access to certain lanes, called high-occupancy vehicle (HOV) lanes. While HOV lanes have been introduced mainly in North America, several cities worldwide have implemented either vehicle driving restrictions (e.g., Munich, Mexico City, Quito, and Shanghai) or congestion pricing schemes (e.g., London, Milan, and Singapore).

On the theoretical front, economic efficiency favors congestion pricing over driving restrictions; while the price mechanism inherent in the former allows flexibility and theoretically achieving a least-cost outcome, the latter can be thought of as being akin to licenses, resulting in a number of unintended consequences. Such pricing mechanism affects both distance driven and the vehicle stock, and in some cases might induce people to replace their vehicle earlier with a newer, more efficient vehicle with less driving restrictions.

In some cases, these policies might backfire. For instance, restrictions based on the last digit of the license plate number (odd or even) have prompted households to keep older vehicles longer and in some cases to have two vehicles—one ending with an odd number and another ending with an even number—to circumvent the regulation. While restrictions based on vehicles' emission rates of local pollutants have been shown to be more effective, there might be trade-offs between reducing air pollution and reducing CO₂ emissions. For instance, gasoline cars tend to emit less PM₁₀ and NO_x but more CO₂ than diesel cars.

Distributional effects are another important consideration: as typically driving restrictions and congestion charges are adopted in city centers and residents are granted free access or discounts on the charge, these policies tend to affect more people living in suburbs and commuting to the city. When driving restriction or fees are based on vehicle emission rates, low-income households who keep their older car for longer or who cannot afford to replace their vehicle would be disproportionately affected by the measures.

Mileage (or Utilization) Tax

Although not yet being implemented on a large scale in the context of passenger vehicles, interest in pricing directly vehicle distance driven has grown over the years. In part that is due to the availability of the GPS technology necessary to implement such tax, but in part it is due to concerns that, thanks to the improvements in vehicle energy efficiency, revenues from fuel taxes will decline over time. Because its simplest form does not depend on fuel efficiency, a mileage tax would have an effect on carbon emissions only through the reduction of kilometers driven, but it would not impact vehicle choice. Thus, from a climate policy standpoint it should be considered as a complement of fuel taxes, standards, and vehicle taxes, rather than a substitute (Parry et al., 2007).

Shortcomings of a mileage tax include the fact that it does not address congestion and its distributional effects, since they impose a heavier burden on rural households, who often have less access to alternative means of transportations.

Modal Transport/Public Transport

The provision of public transport services has the potential to affect vehicle use both in terms of intensive margins (how much to drive) and of extensive margins (the decision to buy a car). Moreover, mileage elasticity with respect to fuel taxes or congestion charges is affected by the availability of alternative transportation modes.

Currently, most public transport systems in the world are partially subsidized by local governments, and in Europe a number of towns have considered to introduce—if not already implemented—a zero-fare policy. The rationale for such government subsidies is that the substitution between transportation modes mitigates externalities associated with driving, including carbon emissions. Studies in different countries have indeed shown positive effects of public transport availability in terms of reduction in utilization, congestion, accidents, and air pollution (Anderson, 2014). Thus, it is arguable that positive subsidies toward mass transit are justifiable on efficiency grounds. The optimal subsidy amount likely depends on the existence of other measures aimed to reduce vehicle use, such as congestion charges or dedicated bus lanes.

Afterword

This paper briefly surveys policy instruments currently used in the transport sector. Given how mobility is believed to develop in the coming years, many new exciting questions arise. First is what will be the prevailing technology when it comes to mobility. Although electric mobility currently appears as the leading alternative to internal combustion engines, hydrogen-based alternatives should not be discarded in the long run.

Second, for a given technology to prevail, it requires charging infrastructure which itself might generate new externalities. For instance, in its current form, charging can become a challenge in cities where vehicles are parked on streets, demand for charging can eventually worsen congestion and force regulators and urban planners to rethink how cities are designed and operated.

Third, assuming electric mobility is to prevail, the precise quantification of its carbon footprint is paramount. Currently, batteries are manufactured using minerals extracted under often-unassessed conditions, and whose disposal and recycling calls for increased attention. Moreover, more should be understood about the potential impact of electric mobility on electricity prices as well as the continuing dependence of the carbon footprint of electric vehicles on the source of electricity, which can vary both across regions and even within the day for a given region.

Finally, the choice of policy instruments might be affected by how the business model of the transport sector will develop. For instance, whether vehicles will either be owned, leased, or rented, that is, if transportation will become a service and vehicles will be accessed through a subscription vehicle along the lines of a mobile phone or ISP subscription nowadays, whether batteries will become commodities and will be replaced rather than charged, and the extent to which autonomous driving will be prevalent. All

these aspects might have an impact on utilization, vehicle choice, and the speed with which vehicle owners respond to changes in policy instruments and thus on carbon emissions.

Biographies



Cristian Huse is a Professor of Economics at the Department of Economics, University of Oldenburg, Germany, where he holds the chair of Applied Microeconomics. His research interests comprise Applied Econometrics, Applied Microeconomics, Environmental, Energy, and Transport Economics, and Industrial Organization. His research has appeared in scientific journals in Economics and in the Sciences. Recent research examines whether consumers correctly value policy instruments, the choice of energy source in personal transportation, and the effectiveness of environmental policy in the transport sector. He is a graduate of the London School of Economics (PhD in Economics, MSc in Econometrics and Mathematical Economics), the Getulio Vargas Foundation (MA in Economics), and the Federal University of Rio de Janeiro (BA in Economics). His teaching experience includes courses in Econometrics, Environmental Economics, and Industrial Organization, at different levels.



Davide Cerruti is a Postdoctoral Researcher at the Centre for Energy Policy and Economics, ETH Zürich. His research interests include Applied Microeconomics, Public Economics, and Environmental, Energy and Transport Economics. His recent work evaluates the effect of the UK vehicle circulation tax on carbon emissions, and the consequences of low awareness about fiscal incentives for energy efficient vehicles. He holds a PhD in Agricultural and Resource Economics from University of Maryland and an MSc and a BSc in Economics from Università Bocconi, Italy. In the past, he worked also at the Energy Policy Research Group, University of Cambridge.

References

- Adamou, A., Clerides, S., Zachariadis, T., 2014. Welfare implications of car feebates: a simulation analysis. *Econ. J.* 124 (578), F420–F443.
- Anderson, M.L., 2014. Subways, strikes, and slowdowns: the impacts of public transit on traffic congestion. *Am. Econ. Rev.* 104 (9), 2763–2796.
- Anderson, S.T., Sallee, J.M., 2016. Designing policies to make cars greener. *Annu. Rev. Resour. Econ.* 8, 157–180.
- Bento, A.M., Goulder, L.H., Jacobsen, M.R., Von Haefen, R.H., 2009. Distributional and efficiency impacts of increased US gasoline taxes. *Am. Econ. Rev.* 99 (3), 667–699.
- Cerruti, D., Alberini, A., Linn, J., 2019a. Charging drivers by the pound: how does the UK vehicle tax system affect CO₂ emissions? *Environ. Resour. Econ.* 1–31.
- Cerruti, D., Daminato, C., Filippini, M., 2019. The impact of policy awareness: evidence from vehicle choices response to fiscal incentives. CER-ETH Economics Working Paper Series 19/316. ETH Zurich, Zurich.
- Huse, C., Koptuyg, N., 2018. Salience and policy instruments: evidence from the auto market. Working Paper. University of Oldenburg, Oldenburg.
- Huse, C., Lucinda, C., 2014. The market impact and the cost of environmental policy: evidence from the Swedish green car rebate. *Econ. J.* 124 (578), F393–F419.
- Jacobsen, M.R., Van Benthem, A.A., 2015. Vehicle scrappage and gasoline policy. *Am. Econ. Rev.* 105 (3), 1312–1338.
- Klier, T., Linn, J., 2015. Using taxes to reduce carbon dioxide emissions rates of new passenger vehicles: evidence from France, Germany, and Sweden. *Am. Econ. J. Econ. Policy* 7 (1), 212–242.
- Knittel, C.R., 2012. Reducing petroleum consumption from transportation. *J. Econ. Perspect.* 26 (1), 93–118.
- Li, S., Timmins, C., Von Haefen, R.H., 2009. How do gasoline prices affect fleet fuel economy? *Am. Econ. J. Econ. Policy* 1 (2), 113–137.
- Miravete, E.J., Moral, M.J., Thurk, J., 2018. Fuel taxation, emissions policy, and competitive advantage in the diffusion of European diesel automobiles. *RAND J. Econ.* 49 (3), 504–540.
- Parry, I.W.H., Small, K.A., 2005. Does Britain or the United States have the right gasoline tax? *Am. Econ. Rev.* 95 (4), 1276–1289.
- Parry, I.W.H., Walls, M., Harrington, W., 2007. Automobile externalities and policies. *J. Econ. Lit.* 45 (2), 373–399.

Computable General Equilibrium Analysis in Transportation Economics

Johannes Bröcker, Kiel University, Institute for Environmental, Resource and Spatial Economics, Kiel, Germany

© 2021 Elsevier Ltd. All rights reserved.

The Idea of Computable General Equilibrium Analysis	520
The Structure of CGE Models	521
Functional Forms	522
Spatial CGE and Transport	523
Economies of Scale	524
Further Extensions and Conclusion	526
Relevant Websites	526
References	526
Further Reading	526

The Idea of Computable General Equilibrium Analysis

Computable general equilibrium (CGE) analysis is a mathematical tool for studying various influences from outside on virtually any relevant economic variable of an economy. The influences to be studied may be policy interventions such as taxes, minimum wages, tariffs, or transport policies of any kind: infrastructure investments, environmental levies and congestion prices, antitrust policies, network access rules, and many more. Influences from outside may also be technology changes or developments in the rest of the world affecting the economy under study through trade or factor flows. The method is a multipurpose weapon. One-fits-all designs of CGE models abound therefore in applied research. Naturally, however, such designs need to be complex. Small models with a design focused on the specific questions at hand might therefore be preferred. The method is applicable on any spatial scale, urban, regional, national, or global. There are good examples of applications on all these scales published in respected scholarly journals, for example, [Horridge \(1994\)](#), [Anas and Liu \(2007\)](#), and [Tscharaktschiew and Hirte \(2012\)](#) on the urban scale or [Bröcker et al. \(2010\)](#) on a European scale.

The “C” in the CGE acronym means “computable” or “computational” in a double sense. First, the instrument is used for quantifying policies or other impacts by simulating the response of the economy in a computer simulation. Second, this task is doable not just by studying hypothetical examples, but also by basing the simulation on real-world data. A common feature of CGE models is data replication, meaning that solving the model for a reference situation (the “benchmark”) should reproduce accounting data recording the economic transactions covered by the model in value terms. Data sources are national accounts, including input–output data and additional accounting information needed for the problem at hand. The impact study needs to compare the benchmark with an alternative resulting from inserting a policy intervention, a technology change, or what else is to be quantified. In the CGE jargon, such exogenous changes are called “shocks.”

In brief, the CGE analysis proceeds in the following three steps:

1. Designing a theoretical model implemented on a computer.
2. Determining model parameters in such a way that the relevant real-world accounting data are replicated (calibration).
3. “Shocking” the simulated economy by a policy intervention, technology change, etc.

What does the “GE” mean in the CGE acronym? It means that the model is a version of the general equilibrium model in mainstream economics, as explicated in advanced microeconomic textbooks, with certain extensions toward imperfect markets that may be found only in specialized literature. There is, however, a fundamental—often concealed—difference between the textbook model and the applied CGE, related to the difficult issue of aggregation. While the agents in the theoretical general equilibrium model are individual firms or households, agents in a CGE model are aggregates, for example, “firms” representing entire industries or “households” representing the entire private household sector in a region or country. In the special case of firms under perfect competition this jump from micro to macro is in fact justified, because a group of firms behaves like an individual firm, but in all other cases it is not. In certain cases one can go from micro to macro by the simplifying assumption that all firms in a macro-representation of an industry are identical except for the product variety they offer (see Section “Economies of Scale”), but sometimes one just closes the theoretical eyes when jumping from micro to macro.

Similarly, while commodities in the general equilibrium model are individual items of which myriads exist in a real economy, CGE commodities are aggregates such as cars, textiles, or even more aggregated objects such as manufactures or services. The main variables to be determined in a CGE are quantities of economic transactions (goods supply and demand, labor supply and demand, and so forth) and the respective prices. How can one define a price and a quantity of something that heterogeneous as manufactures, say, in a meaningful way? The trick is the so-called LEONTIEF convention: the benchmark price of the aggregate commodity (manufactures in this case) is arbitrarily chosen, for example, equal to one. The benchmark quantity is then simply

the value in the accounting data, divided by the price. In words: one unit of manufactures is the quantity that is worth one unit of money in the benchmark. After the shock both, price and quantity, change, and the after-shock value of a transaction is its after-shock quantity times its after-shock price. Though levels of prices and quantities do depend on the arbitrary choice, percentage changes do not. This concept of a “quantity” is a striking difference between modeling philosophies in CGE analysis and transport engineering. Transport engineers look at “real” quantities such as number of passengers or tons. This is a source of confusion between the two modeling communities and causes problems for linking models from the two worlds (see Section “Spatial CGE and Transport”).

The rooting of CGE analysis in mainstream microeconomics establishes a methodological link between traditional cost-benefit analysis (CBA) and project evaluation by CGE analysis. In classical CBA, infrastructure investment (building a road, say) is evaluated by subtracting the construction costs from the money-metric benefit measured by the consumer surplus of road users. The theory behind is that this is the utility gain, translated to an amount of money by the so-called Hicks variation, if the road user is a consumer (abstracting from the negligible income effect). If it is a firm, this is just the profit gain, which eventually shows up as an income gain in the pockets of households. If a service is affected that is privately supplied, in addition the producer surplus is to be added that also measures the profit gain of the respective firms. Exactly the same valuation measures are quantified in a CGE, though not evaluated “on the road,” but based on the utility changes of the households at the end of the adjustment chain triggering through the economy.

It is now—after a long and sometimes confusing literature debate about “Wider Effects”—well understood that under perfect competition both approaches deliver the same result, provided there are no technological externalities. It does not matter how indirectly households are affected, whether they are themselves the road users or buy goods transported along the road or buy goods that use goods transported along the road as inputs and so forth.

Hence, given that results should coincide, what can CGE analysis add to traditional CBA? First, the distribution of benefits and costs can be added. Multiregional models, to be discussed in Section “Spatial CGE and Transport,” allow for identifying the regional incidence of policy intervention, which is vital for policy issues such as cost sharing of infrastructure investment. Introducing households representing different income groups allows for quantifying the incidence across social strata, which is vital for a policy debate on transport policy and social justice.

Second, CGE is the method of choice if the two restrictive assumptions mentioned, perfect competition and no technological externalities, are violated. In case of imperfect competition, a well-designed CGE can identify the deviation between consumers’ surplus and households’ benefit, and thereby give a theoretically clean and data-driven answer to the wider benefit issue. A CGE is also helpful if indirect technological externalities are at stake, externalities that are generated by a transport policy intervention elsewhere in the economy, outside the transportation sector.

The Structure of CGE Models

A CGE has two constituents, the set of agents and the set of markets. Agents sell or buy goods and production factors to or from the markets. There are two types—firms representing the behavior of industries and households representing the entire household sector or segments of it. Agents act rationally; firms maximize profits subject to technological constraints, and households maximize utility subject to their respective budget constraints requiring that what they spend equals what they receive from selling the production factors they own and possibly from sharing in profits of firms. If the state enters the scene, it is modeled as a further household collecting taxes, paying transfers, and buying goods for public consumption and investment. Tax payments and subsidy receipts then have to be subtracted or added to the firms’ revenue, and taxes and transfers have to be subtracted or added to the households’ budgets. In any case, budget constraints must hold for all agents. Black holes where receipts vanish or a Land of Cockaigne allowing agents to spend what they did not earn is not allowed.

First generation CGE models—still widely in use—assume perfect competition and constant returns to scale (CRS) (Shoven and Whalley, 1992). Firms take prices as given and do their best to adjust to those prices. CRS means that any production plan fixing the goods to sell and goods and factors to buy for use in production can be scaled up or down ad libitum. This is the traditional assumption making life technically easy, but it is obviously special, ruling out economies of scale (EOS) that are prevalent in the real world and may alter the results emerging from shocking the economy substantially. We come back to EOS in Section “Economies of Scale.”

Under perfect competition and CRS there are only two possibilities for a firm to act in equilibrium. Either prices are such that the maximum attainable profit just equals zero and the firm produces at which activity level ever market demand requires it to produce, or the firm is unable to break even and shuts down. In both cases, there are no profit shares to be transmitted from firms to households. Note that zero profit does not mean zero capital income. Capital is a production factor owned by households, such as labor. Capital income is the revenue received from selling the factor to firms.

In general, firms can produce more than one kind of output good, and one kind of good can be produced by more than one firm. Different from classical input-output analysis, no one-to-one correspondence between industries and commodities is required. Households are always assumed not to have any influence on prices. They maximize utility subject to their respective budget constraint, taking all goods and factor prices as given. Finally, in equilibrium prices and activity levels must adjust in such a way that markets clear.

Considering a closed economy without trade and income flows with the rest of the world, and incorporating the state as well as investment demand into the household sector, we end up with the following system of equations and corresponding unknowns:

$$\sum_{f \in F} x_f I_{fi}(\mathbf{p}) - \sum_{f \in F} x_f S_{fi}(\mathbf{p}) + \sum_{h \in H} D_{hi}(\mathbf{p}) = 0 \text{ for all commodities } i; \text{ unknowns } \mathbf{p}$$

$$P_f \leq 0; \quad x_f \geq 0; \quad P_f x_f = 0 \text{ for all firms } f; \text{ unknowns } x_f$$

with profit per unit of activity:

$$P_f = \sum_{i \in C} p_i S_{fi}(\mathbf{p}) - \sum_{i \in C} p_i I_{fi}(\mathbf{p}).$$

The set of firms is denoted F . For each firm f in the set of firms one has to determine a level of activity x_f . The required input of commodity i in the set of commodities C per unit of activity, given the vector of all prices $\mathbf{p}$, is $I_{fi}(\mathbf{p})$. Thus, total input demand of the firm is $x_f I_{fi}(\mathbf{p})$. Similarly, $x_f S_{fi}(\mathbf{p})$ is the firm's output supply of commodity i . Remember that we allow for multiproduct firms. Finally, $D_{hi}(\mathbf{p})$ is net demand for commodity i of household h in the set of households H . Hence, the first equation says that total net demand must be zero for all commodities i ; in other words, markets must clear. The term "commodity" here also covers factors of production such as capital and labor, possibly differentiated by skill, location, and others. Note that demand and supply for commodity i do not just depend on the price for commodity i , but also on all prices gathered in the vector $\mathbf{p}$. Note furthermore that one cannot express demand and supply of firms as a function of prices alone because firms are willing to scale supply and demand up or down at will if prices are such that they just break even. Thus, levels of activity appear as additional unknowns to be determined by the second line, expressing not an equation but a so-called complementarity. In a somewhat tricky way, it just repeats what has been verbally already explained: profits must not be positive, but may allow or not allow the firm to break even. In the former case ($P = 0$), the level of activity can be positive or zero, in the latter ($P < 0$) the firm must shut down. We thus have just the right number of equations for the unknowns to be determined. One can show that one of the equations is redundant; it must automatically hold if all others hold due to the fact that all budget constraints are fulfilled (WALRAS' Law). This delivers one degree of freedom for introducing a further equation fixing a price level, but not shown here. The formal system must be modified, if a public sector raising taxes and paying subsidies enters the scene. The modifications are tedious to describe but straightforward. In any case, one must take care that taxes paid by private agents add up to the tax revenue of the public sector and that subsidies received by private agents add up to subsidy expenditures of the public sector.

This is a nonlinear complementarity system. Equations can be understood as a special case of complementarities. Conversely, a complementarity can be transformed into an equation, but unfortunately one that is non-differentiable even if all functions involved are differentiable. Therefore, complementarities are more difficult to solve, but there exist powerful algorithms now doing the job also for large systems.

Once all technologies and utility functions are specified by concrete functional forms (see Section "Functional Forms") one solves this system twice, once for the benchmark and once for the alternative, after a shock. The benchmark solution should replicate the relevant data, which is made sure by an appropriate choice of parameters. The difference between benchmark and alternative is the impact we are looking for.

Functional Forms

In a general equilibrium model agents are characterized by technologies or preferences that determine their respective responses to market prices. In the textbook version of the model these building blocks are typically specified in an abstract way by just requiring them to fulfill certain regularity conditions such as CRS, monotonicity, and convexity necessary to make the mathematical machinery of the theory work. For computation this is not enough. We need specific parametric functional forms that can be implemented on a computer.

Microeconomics tells us that the technology and thus the input and output choices of firms (the functions I_{if} and S_{if} in the equations in Section "The Structure of CGE Models") are completely determined by the firm's cost and revenue functions (the "dual approach"). The cost function specifies the minimal attainable input cost per unit of activity, given the input prices. Similarly, the revenue function specifies the maximal revenue per unit of activity, given the output prices. The expenditure function is the analogous concept for describing the preferences of households and thus the demand functions D_{ih} .

There is a whole bunch of functional forms for these functions proposed in the literature, ranging from LEONTIEF and COBB-DOUGLAS to more and more complex constructs with strange names, constant elasticity of substitution (CES), constant elasticity of transformation (CET), LES, CDES, addilog, translog, AIDS, and more. Criteria for choice are as follows:

- Theoretical consistency: A form should fulfill certain mathematical properties implied by optimization behavior. For example, a cost function must be linear homogeneous (proportional price changes lead to proportional cost changes), monotone, and concave.
- Factual conformity: A form should conform to stylized facts about agents' behavior. For example, representing consumer behavior by a COBB-DOUGLAS function does not—though widely applied—conform to the observation that the composition of

consumer demand systematically changes with changing incomes. This can be important in transport analysis, as rich people choose for example different modes of transport than poor people, facing identical prices.

- Flexibility: A form should be flexible enough to represent different patterns of behavior.

Choosing a functional form is a bit of an art. One has to trade off flexibility against parametric parsimony. More flexible forms need more parameters to tune them to observed behavior, but there is often little evidence to assign numbers to them. Many functional forms allow for a distinction between two types of parameters, shifters (or position parameters), and elasticities. Shifters are parameters used to trim the functions in such a way that accounting data are reproduced in a benchmark equilibrium. For example, for a cost function representing input choice one needs n free shift parameters for n inputs. Elasticities tell how agents respond to price changes. From a benchmark data set one cannot infer on these parameters. They must be borrowed from econometric studies in the literature. Fixing them is guesswork. Shakiness of the elasticities is a critical point in CGE analysis.

A convenient and most popular form for the input composition is the CES function. It implies that, for any pair of goods, the percentage response of the input ratio to a 1% change of the respective price ratio is a negative constant (the so-called elasticity of substitution) that is uniform across all pairs of goods. This form is parametrically parsimonious in that one needs, beyond shift parameters, only a single further parameter, the elasticity of substitution. Obviously, however, the function is rather inflexible, not allowing for low substitutability between certain pairs of inputs and high substitutability between others.

A higher flexibility is achieved by nesting CES functions, as in the trade example in Section “Spatial CGE and Transport.” High flexibility is achievable by complex multilevel nesting structures, but at the cost of large information requirements. Consumers’ choices can be dealt with in a similar way. An analogous concept can be applied for output choice of multiproduct firms, with the CET function. It has the same mathematical form as the CES function, with the negative elasticity of substitution replaced by the positive elasticity of transformation. While the input ratio falls with an increasing price ratio, the output ratio goes up.

Spatial CGE and Transport

Some early CGE studies on transport used simple single-region national CGE model. Here, transport is just another industry whose output, the transport service, is used by firms as a production input or by households as a consumer good. A shock may be a public infrastructure investment raising the transport sector’s productivity, a levy imposed on transport for damaging emissions, a tax, a technology improvement, and so on.

This is fine as long as one is only interested in the national scale, but the strength of the CGE approach becomes apparent in particular when it comes to the spatial dimension of transport policy. In the spatial dimension, transport is not just another industry, but a core explanatory factor for understanding the subject at hand.

Conceptually, there is not much difference between a single-region and a multiregional CGE, except for the definitions of agents and commodities. Agents are assigned to regions, and commodities are distinguished by their respective place of availability. A commodity at the place of production is understood as different from the same good, after it has been transferred from the place of production to the place of consumption. This transfer is regarded as another production process generating the commodity at the place of consumption using the commodity at the place of production and a transportation service as inputs.

The nesting concept is perfect for representing this transformation process, as illustrated for trade flows from several regions of origin to one destination in Fig. 1. For transferring a good from one of the origins to the destination, the good itself is combined with a transport service to obtain the same good at the destination. The transport service and the good are combined in a fixed proportion. In other words, the respective elasticity of substitution is zero (called LEONTIEF technology). On an upper level, goods of the same kind stemming from different origins are combined to obtain a composite further used in production and final consumption in the destination region. The elasticity of substitution is positive but finite, representing the willingness of customers to substitute sources in interregional trade as a response to relative price changes. Some authors borrowed estimates of this elasticity from international trade studies using a gravity model. As one is able to estimate trade costs in international trade, one can infer on this elasticity from the estimated trade cost elasticity of the gravity equation. The application of elasticities in international trade to interregional trade

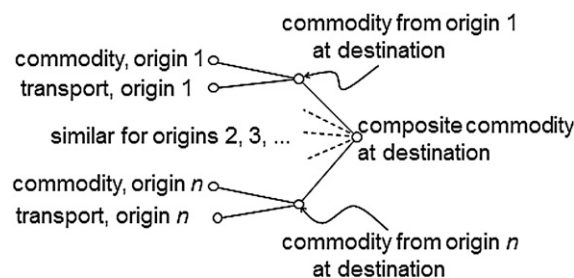


Figure 1 Nesting structure for transport and trade. In the lower nest (to the left) the commodity and the transport service are combined in fixed proportion (LEONTIEF technology), in the upper nest (to the right) commodities from different origins at the destination are combined by a CES function with positive finite elasticity of substitution.

within nations is, however, a weakness of this approach. Short-distance elasticities may be larger because less trade cost sensitive goods select into international trade than into interregional trade.

Simulating a cost-reducing infrastructure improvement needs to shock the transport technology in such a way that transport services cost less per unit after the shock than before, given unchanged input prices of the transport service producer. This induces complex adjustments through the entire economy. If transport from some origin to some destination becomes cheaper, firms and households substitute goods from the cheaper source for goods from other sources. Producers at the origin sell more, while customers at the destination get their respective input or consumption bundle at a lower price per unit of the good composed of deliveries from all sources. At the end of the chain households at the destination will pay lower prices, while households at the origin will earn higher factor incomes. But not all households at the origin and destination gain. Some may suffer from higher prices at the origin without benefiting from higher factor income, and some may suffer at the destination from facing tougher competition by producers at the origin now enjoying easier market access. The strength of a CGE is to cover all these direct and indirect channels in a single consistent framework, preventing oversimplifying conclusion that takes some channels into account and neglecting others.

Obviously, with many regions and industries, the model design becomes rather complex. Economists have invented a dirty trick simplifying it considerably, the iceberg transport cost. It means that the resource used for transport is the delivered commodity itself, instead of a special transport service as explicated earlier. A certain percentage of the commodity is simply used up for doing the transport work. It melts like an iceberg. Let's say a transport engineering study (we come to that later) tells that the transport cost for getting certain goods from an origin to the destination amounts to 10% of the good's value. This is implemented by assuming that 10% of the quantity melts, while the price per unit at the destination exceeds the price at the origin by the factor of $1/0.9$, such that the values (price times quantity) at the origin and destination coincide. While sounding strange, this greatly simplifies model designs. As long as the issue at hand is just the impact of transport cost changes and not developments in the transport industry itself, the approach is not seriously misleading.

Some authors, discontent with taking over simple aggregate cost figures from engineering studies, try to extend the CGE framework to also incorporate the different logistic decision levels such as mode choice, route choice, urban parking, and more into a unified CGE, that is, a grand overall system of complementarities simultaneously solving for production, consumption, trade, mode choice, route choice, etc. This is an interesting direction of research from a methodological point of view, but applicability to practical policy studies is still limited. The technical requirements are formidable.

Leaving the within transport sector details to the well-established transport engineers' modeling strategies and establishing an interface between the transport model and the economic CGE model turned out to be a more powerful pragmatic strategy. Nevertheless, the interface raises subtle technical questions related to different modeling philosophies of CGE modelers and transport engineers. A striking symptom of this difference is the fact that CGE modelers love CES functions for choices, for example, region of origin choices in trade (see earlier), while transport engineers love random choice models, in particular the logit model. Though these two forms have been shown to be close relatives, the difference matters for project evaluation.

Let us take mode choice as a simple example. A CGE modeler measures the quantities of transport by each mode and the composite quantity generated by combining modes according to the LEONTIEF convention (see earlier): value divided by the arbitrary benchmark price for the reference case. Adding up values by mode gives the value of the composite. After a shock (cost reduction for one mode, say), both, quantities and prices, change, but value aggregation is kept; adding the after-shock values by mode yield the value of the composite. The change of the composite price measured by the cost function is the right evaluation measure for the cost decline of one of the modes. Quantity aggregation does not hold, in general. Adding up quantities by mode does not yield the quantity of the composite.

The transport engineer measures quantities by mode in tons. Adding up tons by mode yields tons of the composite in the benchmark as well as after a shock. Using a random choice model with random components added to prices, the correct evaluation measure of a cost reduction for one mode is the expected minimal expenditure including the random components, the logsum in case of a logit model. If this measure is inserted as price of the composite, value aggregation does not hold, in general. The accounting consistency of the CGE model gets lost. In practice, some authors used therefore simply the weighted average of mode costs as composite price. This keeps value aggregation but may give wrong answers in project evaluation: average costs can increase, if one mode (a particularly costly one) becomes cheaper and none more expensive. A heterogeneous agents approach may offer a way out of this dilemma (see briefly on that in Section "Further Extensions and Conclusion").

Economies of Scale

The traditional CGE approach assumes CRS and perfect competition. These assumptions clearly contradict the evidence; EOS are relevant in many industries. Excluding them by assumption can be very misleading in project evaluation. If average cost declines with output expansion, average cost exceeds marginal cost. If prices equal marginal costs, as implied by perfect competition, firms could not cover their costs. The theory collapses.

Let us look at two ways to cope with EOS, both easily applicable in a CGE. They may be regarded as the extreme ends of an entire continuum of model specifications. Superficially looking a bit at ad hoc, they both have deeper roots in modern microeconomics. The first approach is average cost pricing illustrated in Fig. 2. If a policy intervention shifts the demand curve to the right, output expands and average cost declines, and hence the price falls. This is welfare enhancing.

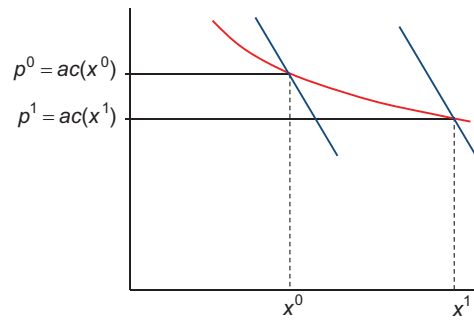


Figure 2 Response to market expansion, DIXIT-STIGLITZ version. Explanation of symbols: x = effective output; p = output price; ac = average cost; superscripts 0 and 1 denote equilibria before and after an expansion shock, respectively. The blue lines are demand schedules; the red curve is the average cost curve. Effective output includes the utility or productivity effect of increased diversity.

Simple as it is, the approach can nevertheless be derived from a popular model of monopolistic competition, the DIXIT-STIGLITZ model (Dixit and Stiglitz, 1977). According to this model, the industry output is a composite of many varieties of goods belonging to the respective industry. Producers of an individual variety have a certain degree of monopoly power to set a price above marginal cost. Entering the industry by offering an additional variety incurs a certain fixed cost, but otherwise it is free. Firms enter or leave the market until profits are driven to zero, that means until markups just cover the fixed cost. An expanding market induces market entry and thus an increase of product diversity. What is measured on the abscissa in Fig. 2 is effective output taking diversity into account. Moving to the right means more varieties, more of each variety, or a combination of both.

For implementing the model one does not need more than a single extra parameter, the elasticity of average cost with respect to industry output. It measures the percentage change of average cost per percentage change of output. In the DIXIT-STIGLITZ model this parameter is minus the inverse of the elasticity of substitution between varieties. If substitutability of varieties is low, the average cost curve declines steeply. Other justifications than the DIXIT-STIGLITZ model can also be put forward. But anyway, the mentioned elasticity is the only thing that matters empirically, whatever we accept as microeconomic background music.

At the other extreme of EOS models lies the oligopoly model illustrated in Fig. 3. As in the DIXIT-STIGLITZ model, average costs decline with an expansion of the market, but the price does not fall, as long as the marginal cost does not change (what is assumed in the figure for the sake of simplicity). Cost savings are entirely appropriated by shareholders of firms as additional profits.

A theoretical justification is the NEARY oligopoly model (Neary, 2003). As in the DIXIT-STIGLITZ model, industry output is a composite of many varieties, but in this case there is no entry or exit. The number of varieties is fixed. Each variety is the output of a tiny industry with a small number of firms interacting strategically within their respective tiny industry. They set a price according to the COURNOT oligopoly model. It depends only on the degree of substitutability between varieties and the number of firms in each tiny industry, both regarded as fixed parameters. Therefore, the price is a fixed markup over marginal cost. It does not respond to the market expansion.

To implement this approach, we also just need the elasticity of the average cost with respect to output as an extra parameter, the same information as for the former solution. In this case, this elasticity is not a constant, but we only need to know it for the benchmark. It is equal to minus the share of fixed cost in total cost. We also need the share of profits in the industry's turnover. This is available from the benchmark accounting data and thus no extra information to be taken from external sources. For both approaches, the equation system in Section "The Structure of CGE Models" needs to be modified, but we skip the technical details here.

Hence, for both approaches we need essentially the same extra information for calibration, and both lead to the same principal conclusion with regard to overall welfare effects of a demand shift: if an intervention lets the respective industry expand, this gives

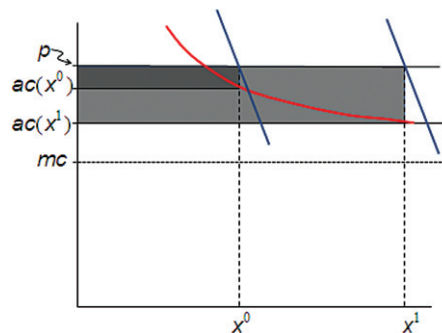


Figure 3 Response to market expansion, NEARY version. Explanation of symbols: mc = marginal cost. For the other symbols, see Fig. 2. The dark gray rectangle is benchmark profit; the total gray rectangle (dark and light) is after-shock profit.

rise to an extra welfare gain that would not occur under CRS and perfect competition. The transport policy intervention can therefore generate a wider economic benefit due to sectoral shifts induced indirectly in the economy. This sword, however, cuts both ways; the contraction of the respective industry leads to a welfare loss.

While the two approaches lead to similar (though slightly differing) overall welfare effects of a demand shift, the distributional impact is completely different. In the first approach all gains are reaped by the customers, whereas in the second approach they go entirely to the shareholders of the firms. This implies that also the regional incidence can be quite different, because customers and shareholders reside at different locations. An in-between of the two approaches can be constructed by assuming that the price declines with market expansion, but less steeply than the unit cost. This requires another additional parameter.

Further Extensions and Conclusion

Many important aspects had to be omitted in this short review. One is dynamics. Transport policies affect private investments, inducing welfare changes in the future. To account for them in a consistent way, a dynamic approach is needed. The strict neoclassical concept of dynamics is a model with forward-looking agents. Unfortunately this is not only a mathematically demanding, but also a disputable concept. Few people would accept that agents have strong beliefs about future prices and quantities that later turn out to be correct, at least on average. What speaks in favor of the approach is that dispensing with it means to also dispense with neoclassical equilibrium as a guiding principle and to get lost in arbitrariness, when it comes to investment and saving. Most applications rely on more or less plausible approaches to specify saving and investment functions linking static equilibria for consecutive years. But this is disputable as well. In any case, the theoretical strength of the CGE machinery gets lost in a dynamic context.

A more promising extension is introducing agent heterogeneity that recently is an important theme in trade theory and has a longer tradition in urban CGE models. To make sure that not all households in a city react in the same way—in contradiction to the evidence—it is assumed that they belong to a large population of individuals with varying preferences drawn from a common distribution (Anas and Liu, 2007). This leads to the stochastic choice idea that transports planners and engineers adhere to already for long.

To summarize, CGE analysis is a theoretically well-founded assessment tool in transport policy appraisal, at least as long as we do not leave the firm ground of static models. CGE models are notorious for being complex and too expensive for practical application, but this is often due to the fact that one-fits-all models are asked for by policy-makers. It is highly recommended to design tailor-made models for specific questions—parsimonious with regard to the number of agents and markets—that can be calibrated with a small and easily accessible data set. The bad reputation comes from the bad habit to let models grow bigger and bigger, which makes them necessarily more and more expensive and less and less transparent.

Relevant Websites

Markusen, J. Teaching: simulation modeling in microeconomics. Available from: <https://spot.colorado.edu/~markusen/teaching.html>.

Ferris, M.C., Munson, T.S. Path 4.7 solver for complementarity problems. Available from: https://www.gams.com/latest/docs/S_PATH.html.

References

- Anas, A., Liu, Y., 2007. A regional economy, land use, and transportation model (RELU-TRAN©): formulation, algorithm design, and testing. *J. Reg. Sci.* 47, 415–455.
- Bröcker, J., Korzhenevych, A., Schürmann, C., 2010. Assessing spatial equity and efficiency impacts of transport infrastructure projects. *Transp. Res. Part B* 44, 795–811.
- Dixit, A.K., Stiglitz, J.E., 1977. Monopolistic competition and optimum product diversity. *Am. Ec. Rev.* 67, 297–308.
- Horridge, M., 1994. A computable general equilibrium model of urban transport demands. *J. Policy Model.* 16, 427–457.
- Neary, J.P., 2003. Globalization and market structure. *J. Eur. Ec. Ass.* 1, 245–71.
- Shoven, J.B., Whalley, J., 1992. *Applying General Equilibrium*. Cambridge University Press, Cambridge, UK.
- Tscharaktschiew, S., Hirte, G., 2012. Should subsidies to urban transport be increased? A spatial CGE analysis for German metropolitan areas. *Transp. Res. Part A* 46, 285–309.

Further Reading

- Bröcker, J., Mercenier, J., 2011. General equilibrium models for transportation economics. In: Palma, A.D., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *A Handbook of Transport Economics*. Edward Elgar, Cheltenham, pp. 21–45.
- Dixon, P.B., Jorgenson, D. (Eds.), 2013. *Handbook of Computable General Equilibrium Modeling* (Vols. 1A and 1B). Elsevier Science, Amsterdam.
- Ginsburgh, V., Keyzer, M., 1997. *The Structure of Applied General Equilibrium Models*. MIT Press, Cambridge.
- Robson, E.N., Wijayarathna, K.P., Dixit, V.V., 2018. A review of computable general equilibrium models for transport and their applications in appraisal. *Transp. Res. Part A* 116, 31–53.
- Tavasszy, L.A., Thissen, M.J.P.M., Oosterhaven, J., 2011. Challenges in the application of spatial computable general equilibrium models for transport appraisal. *Res. Transp. Econ.* 31, 12–18.

Estimation of Value of Time

Stefan Flügel, Askill H. Halse, Institute of Transport Economics (TØI), Oslo, Norway

© 2021 Elsevier Ltd. All rights reserved.

Abbreviations	527
Type of Study	527
Data and Estimation Approach in National VTT Studies	528
Overview Over National Studies	528
Choice Experiments	530
Estimation Models	531
Selected Challenges	531
The Absolute Size of the Time Saving (Size Effect)	531
Reference Dependency (Sign Effect)	532
Mode and User Type Effects	532
Estimating VTT for Business Trip	532
Estimating VTT for Cycling (and Walking) for Transportation	532
Recruitment Effects	533
Estimation With Revealed Preference Data	533
Further Reading	533

Abbreviations

2aCE	2-attribute choice experiments
CAPI	Computer-assisted personal interviews
CBA	Cost–benefit analysis
CE	Choice experiment
CV	Continuous valuation
Delta T	The size of the time saving
IID	Identically and independently distributed
MNL	Multinomial logit model
MXL	Mixed logit model
OCT	Opportunity cost of time
RC	Revealed choice
SC	Stated choice
SNP	Semi-nonparametric
SP	Stated preference
VBTT	Value of business travel time
VTT	Value of travel time
WTA	Willingness-to-accept
WTP	Willingness-to-pay

Type of Study

Starting in the 1960s, the value of travel time (VTT) has been estimated empirically in an increasing number of studies. The latest meta-analysis on European evidence is based on 389 studies, and additional meta-analyses including data from the other parts of the world suggest that the total number of studies is at least 500 and likely much higher.

In this article, we focus on studies where the estimation of the VTT is the primary motivation for the study, with particular emphasis on national VTT studies. However, the VTT is also estimated explicitly or implicitly in other types of studies. **Table 1** summarizes some characteristics of typical study types.

Most studies that empirically elicit VTT are based on a similar general methodical approach, that is, to infer VTT from choice data using statistical inference—typically by means of logit models and maximum likelihood methods. The choice data can stem from stated choice (SC) experiments or/and from revealed choice (RC) data. The specific requirements for data and model specification in different studies are often motivated by the purpose of the study and the later application of VTT.

Table 1 Overview over some type of studies that estimate VTT

<i>(Some) types of project estimating VTT</i>	<i>National VTT studies</i>	<i>Other WTP studies</i>	<i>Mode choice models for transport planning</i>	<i>Consultants projects for industry</i>	<i>Academic studies</i>
Typical funding/ interest group	Transport authorities or Ministry of Finance	Various	Transport planners	Transport operators, stakeholders	Academia
General purpose	Project appraisal	Primary focus on other WTP values	Predicting travel demand	Eliciting travelers WTP and/or choice probabilities	Scientific/ methodological
Role of VTT	Primary parameter of interest	Various, may be included primary to control other WTP for VTT	Marginal rate of substitution between travel costs and time	Typically, one of many estimated parameters	Various
Typical application of VTT	As unit values for national CBA handbooks	Various	Remain implicit in utility functions underlying transport model	Optimization of prices schemes and timetables	None
Typical segmentation of VTT	Trip element, travel mode trip purpose, trip distance	Various	Trip purpose, time coefficient may be travel-mode-specific	Various, including user groups and geography	Various
Requirement for representativeness	Very important, should apply for the whole nation	Various	Less important due to possibility to recalibrate model	For specific corridors/markets	Less important (more focus on internal validity than external)
Typical data type	Stated route choice	Stated choice and other SP including CV	Revealed mode choice	Route and/or mode choice (often SC)	Various, including combined SC/RC
Typical estimation model	Mixed logit model in WTP space	Various	Nested logit model	Typically, simpler model (MNL)	Often mixed logit or other state-of-the-art

CBA, Cost–benefit analysis; *CV*, continuous valuation; *MNL*, multinomial logit model; *MXL*, mixed logit model; *RC*, revealed choice; *SC*, stated choice; *VTT*, value of travel time; *WTP*, willingness-to-pay.

Data and Estimation Approach in National VTT Studies

As stated in **Table 1**, national VTT studies are conducted with the purpose to provide unit values for cost–benefit analysis (CBA). This implies that the obtained values should be representative for the whole traveling population and relevant for (long-term) welfare analysis. It is worth mentioning that the VTT, that is relevant for welfare analysis, is not necessarily identical to the VTT, which is most relevant for predicting behavior, in particular short-term behavior.

Overview Over National Studies

The 1987 UK study was the first national VTT study of its kind. Relying on SC data and showing that the VTT can be estimated with sufficient statistical accuracy without the use of RC data, the study set the standard for national studies conducted in the 1990s. This first wave of VTT studies was completed in Europe, with the notable exception being the national VTT study of New Zealand in 1999.

In the new millennia, seven countries in Europe conducted (at least) one national VTT study (see **Table 2** for an overview). While there seems to be plenty of empirical evidence on the VTT over the whole world (e.g., there are conference papers reporting on meta-analysis based on 25 VTT studies from Africa and, another one, based on 216 VTT values from Japan), we are not aware of studies outside of Europe after year 2000 that would be characterized as national studies. This relates to the fact that CBA has a stronger tradition and is more frequently done in Europe than in the rest of the world. In some countries, such as France, the official unit value recommendations for VTT in CBA handbooks are inferred from several (minor) VTT studies, rather than one large-scale national study.

From a methodical point of view, the 2004 Danish study, in particular the work on data modeling and estimation by Mogens Fosgerau and colleagues, was very influential on consecutive studies in Sweden, Norway, and the Netherlands. These studies relied on choice experiments (CEs) with time and cost as the only attributes to estimate VTT in main transport mode [while VTT for other time components, like access- or waiting time, and other willingness-to-pay (WTP) values were derived from additional CEs with more attributes]. This so-called 2-attribute choice experiment (2aCE) received some critique over the years (see discussion later), but remains a major part of the experimental design in the 2014 UK and the 2018 Norwegian study. In the former study, the data from the 2aCE were combined with multi-attribute CEs for a pooled data estimation of VTT. Estimation based on multiple and more complex CEs is even more characteristic for the 2012-German study that used different CEs, including mode choice, route choice, and (long-term) residential/work place choice with up to 14 attributes per CE, to estimate VTT.

The German study is also prominent in the inclusion of RC data in the modeling work for VTT departing from earlier practice in national studies that relied solely on SC data.

Table 2 Overview over national VTT studies (after year 2000); characteristics refer to the estimation of the onboard VTT of motorized modes (excluding other time components and factors and excluding the VTT for walk and cycle)

<i>Country (year of data collection)</i>	<i>Central Research Institution/ researchers</i>	<i>Main type of recruitments</i>	<i>Type of interview/ questionnaire</i>	<i>Choice experiment(s) (Nr. of attributes per alternative)</i>	<i>Estimation model</i>	<i>Assumed distribution</i>
Switzerland (2002)	Institute of Transport Planning and Systems (IVT), ETH Zurich/K. Axhausen	From another survey (KEP2)	Paper self-completion questionnaires	Mode (4) and route choice (4)	Heteroscedastic MNL	Deterministic function of distance and income
Denmark (2004)	Technical University of Denmark (DTU)/M. Fosgerau	Web and phone panel	Web survey and CAPI	Route choice (2)	Integrated approach (MXL)	Lognormal with SNP-terms
Sweden (2007, 2008)	Centre for Transport Studies, KTH Royal Institute of Technology/M. Börjesson, J. Eliasson	Population register (2008) ^a	Web survey or call-back interview (2008) ^b	Route choice (2)	Integrated approach (MXL)	Lognormal
Norway (2009)	Institute of Transport economics (TØI)/F. Ramjerdi, S. Flügel	Internet panel	Web survey	Route choice (2)	Integrated approach (MXL)	Lognormal with SNP-terms
The Netherlands (2009, 2011)	Significance/M. Kouwenhoven, G. de Jong	Internet panel (2009), field (2011)	Web survey	Route choice (2)	Latent class models	Discrete distribution
Germany (2012)	IVT, ETH Zürich/K. Axhausen, I. Ehreke	Phone (nonbusiness), panel (business)	Phone (RC), pen-pencil or web (SC)	Mode (up to 11), route (up to 11) and residential/ work place choice (up to 14)	Heteroscedastic MNL	Deterministic function of distance and income
United Kingdom (2014)	University of Leeds/S. Hess, A. Daly	Intercept method (field) and telephone	Web survey and telephone interview	Route choices (2, 4, and 4)	WTP-space MXL	Log-uniform
Norway (2018)	Institute of Transport Economics (TØI)/A. Halse, S. Flügel	Internet panel, email register, and field	Web survey	Route choice (2)	Integrated approach (MXL)	Lognormal

^aNumber plate registration in 2007.

^bPhone in 2007.

Choice Experiments

The design of CEs is a crucial part of SC-based estimation of VTT. As indicated in **Table 2**, most (national) VTT studies apply route CEs. Mode CEs are less common likely because of the risk that the alternative-specific constants are confounded with the alternative-specific time coefficients underlying VTT.

To mitigate the general challenges of the hypothetical nature in stated preference studies, CEs should be designed such that choices are perceived as realistic by the respondents. There are two main approaches to try to achieve this:

1. The choice situation is framed around a reference trip that the respondent recently made. In the choice tasks, the respondent is asked to think back at the reference trip and to imagine that some (few) attributes have changed while all other travel conditions remain as in the reference trip.
2. One includes all relevant attributes (typically more than 5 attributes) in the CEs and varies them across alternatives.

(1) sets higher demands on abstraction (and imagination) on the respondents, while (2) requires higher cognitive abilities from the respondents due to the need to process many attributes and numbers.

Within type (1), we can further distinguish between those CEs with just 2 attributes, travel time, and cost (the aforementioned 2aCE) and those with more attributes (typically 3–5 attributes). **Fig. 1** shows typical examples of CEs with different number of attributes.

Besides choosing the number of attributes, a realistic and meaningful design and framing of attribute values and attribute combination is important. In the 2aCE, one alternative is typically faster but more expensive (while the other is cheaper but slower). In countries that have no road tolls, this setup may be viewed as unrealistic for car travel. Another challenge with the 2aCE might be that respondents understand that they are asked to reveal their WTP, which is likely to increase the likelihood that respondents answer strategically (i.e., not revealing their true preferences). This challenge might also apply for CE with more attributes, but there the trade-off between cost and time is more concealed.

The complexity of the underlying statistical design increases with the number of attributes. However, there are standard methods and tools to create statistical designs, both orthogonal ones where there is (close to) no correlation between attribute values and efficient designs that include an intentional correlation between attributes based on a priori expectation of parameter values.

Typical example of 2-attribute choice experiment		Option A	Option B
	One-way fuel costs	20.30 €	27.50 €
	One-way travel time by car	2 h 23 min	1 h 55 min
		☐ Option A	☐ Option B
Example of a choice experiment with 4 attributes (<i>attributes taken from the UK study; value, format, and layout adjusted by the authors</i>)		Option A	Option B
	One-way fuel costs	22.40 €	24.50 €
	Traffic conditions	1 h 22 min in heavy traffic 55 min in light traffic 23 min in free flowing traffic	22 min in heavy traffic 1 h 55 min in light traffic 58 min in free flowing traffic
		☐ Option A	☐ Option B
Example of a choice experiment with many attributes (<i>attributes taken from the German study; values, format, and layout adjusted by the authors</i>)		Option A	Option B
	Departure time	08:03 am	08:40 am
	Expected travel time of which driving of which in heavy traffic of which walking to/from car	1 h 22 min 1 h 2 min 15 min 5 min	1 h 42 min 55 min 40 min 7 min
	Expected arrival time	09:25 am (in 85% of cases) 09:00 am (in 5% of cases) 09:55 am (in 10% of cases)	10:22 am (in 50% of cases) 10:00 am (in 20% of cases) 11:05 am (in 30% of cases)
	Travel costs	23.40 €	21.50 €
		☐ Option A	☐ Option B

Figure 1 Examples of choice experiments.

Independent of the number of attributes, it is considered standard practice to pivot attribute values around reported reference values. In the 2aCE, the reference values of cost and time typically always occur in one of the two alternatives, but not always in the same alternative. This increases realism and makes it possible to test for reference-dependency [see Section “The Absolute Size of the Time Saving (Size Effect)”].

In the 2aCE, the trade-off between cost and time is “unpolluted” by other attributes, which comes at the advantage of better control about the effective range of trade-offs (bids) in the design. This may make it easier to design the CE such that one observes the whole VTT distribution.

Estimation Models

The VTT is commonly estimated based on parametric models. Nonparametric models, which are only practically feasible with the 2aCE, are sometimes used to test if the bid range in the CE design is appropriate, typically based on pilot data. In order to understand variations in VTT and in order to be able to control for certain effects prior to the inference of mean VTT values, parametric models are needed.

The state of the art of estimation models for SC data depends on the type of CE.

For the 2aCE, the so-called integrated approach to VTT estimation, introduced by the Danish national study, is the most commonly applied model. It departs from a decision rule that is verbally formulated as: “A respondent chooses the faster/more expansive alternative if his/her true VTT is larger than the offered bid.” As the subjective true VTT is not observable to the researcher, an IID logistic distributed error term ε_{nj} for respondent n and choice task t is introduced. The error term is scaled by the homogeneity parameter μ .

$$y_{nt} = 1 \text{ iff } \log VTT_{nt} + \frac{1}{\mu} \varepsilon_{nt} > \log V_{nt} \quad (1)$$

where $V_{nt} = \frac{|\Delta C_{nt}|}{|\Delta T_{nt}|}$ is the trade-off between the difference in the cost attribute and the time attribute, that is, the price to be paid for a unit of time saving.

The VTT is typically log-transformed and parameterized with covariates (A) and an additional error term (u_n) being constant for all choice tasks for a given respondent but being randomly distributed across of respondents.

$$\log VTT_{nt} = \beta_0 + \beta' X_{nt} + u_n \quad (2)$$

The beta-parameters and the variance of u_n (the mean value is captured by a constant term β_0) can be estimated with a standard mixed logit model for panel data. The distributional assumption of u_n (and thus of VTT) is a crucial assumption that can have strong effects of the inferred mean values. One can add semi-nonparametric (SNP) terms in the model that adjust the parametric distribution to the one that is nonparametrically observed.

Typical covariates in the X -vector include income, trip distance, dummies for trip purposes, and design variables, including dummies for recruitment (given mixed data collections, see Section “Recruitment Effects”), ΔT_{nt} [the “size effect,” see Section “The Absolute Size of the Time Saving (Size Effect)”] and dummies capturing the direction of the changes in cost or time relative to the reference value [the “sign effect”, see Section “Reference Dependency (Sign Effect)”]. The direct modeling of the effect of covariates on VTT, sometimes referred as modeling in the marginal-rate-of-substitution (MRS) space, makes the interpretation of beta-parameters straightforward. It is also technically easy to simulate the VTT distributions given (2) based on beta-parameters, the observed vector X and Monto Carlo draws of u_n . One can also change the X -vector based on input data from other sources or weight single observations such to obtain a more representative distribution.

For CEs with more than 2 attributes (or alternatives), the integrated approach is not feasible and model formulation must be done in utility space, where the VTT is derived as the ratio between the time and the cost parameter. It is, however, possible to rescale the utility (typically by factoring out the cost coefficient) such that the scale can be interpret in monetary term. The rescaled time coefficient (interpreted as VTT) can then also be parameterized with covariates. The 2014 UK study also suggested ways to account for size and sign effect on VTT (see later) in multiple-attribute CE. However, the required modeling work is more involved and results are more difficult to interpret compared to the integrated approach.

As mentioned earlier, the UK and the German VTT study also build estimation models on pooled data from different CEs. An important feature of these models is to allow for different utility scales (error variances) from data from different CEs.

With the 2aCE, one estimates the VTT that corresponds implicitly to the average travel quality or comfort experienced by travelers on the reference trip. To capture the effect of quality on VTT, one will need additional CEs in which the quality is presented and varies exogenously. However, the VTT from the 2aCE is not necessary the same as in multiple-attribute CE, such that inconsistency may arise. This is a disadvantage with the 2aCE compared to more complex CE and/or modeling approaches based on joint estimation.

Selected Challenges

The Absolute Size of the Time Saving (Size Effect)

A typical empirical finding is that WTP increases over-proportionally with the size of the time saving (one time saving of 20 min is more valuable than 2 time savings of 10 min) implying that the VTT increases with the absolute size of the time saving. This is obviously a challenge to the concept of unit values applied in CBA.

While this size effect is likely to reflect patterns of true preferences, and therefore may also be an issue in modeling based on RC data, the size effect is of particular concern in SC-based estimation. This is because the absolute size of the time saving (ΔT) is chosen by the researcher through the statistical design, which means that VTT results are sensitive to the SC design.

The best practice is to estimate the effect of ΔT and to simulate the VTT distribution for a predefined value of ΔT . In this case, the results get at least reproducible.

Some researchers have additionally argued that SC choices with very small time savings (i.e., ΔT of a few minutes) are unreliable; this has led to the practice adopted in many VTT studies to set a requirement on the trip duration (e.g., at least 10 min) for the reference trip. Regarding the merits of small time savings one may distinguish between internal and external validity. It might be that small time saving helps to identify the correct function relationship between VTT and ΔT in the SC-choice context, that is, in a short-term perspective, but that small time savings are not valid for long-term welfare analysis. A pervasive argument for the latter is that respondents are unable to recognize longer-term advantages of short time savings (e.g., the possibility of rescheduling activities).

Reference Dependency (Sign Effect)

Even more so than the size effect (see previous section), reference dependency (the sign effect) is a core conflict between behavioral economics and welfare economics. Behavioral economics suggest that people value the amount of goods not absolute, but relative to some internal reference value. Empirical evidence largely supports this theory and point to loss aversion, that is, that respondents value losses (negative changes compared to the reference point) stronger than gains. For the VTT, this implies that the WTP to improve travel time relative to the reference value is lower than the monetary compensation one requires to accept slower travel times (referred to as willingness-to-accept, WTA).

For welfare economics, it is problematic that the VTT is lower in WTP situations compared to WTA situations. For the estimation of VTT that is to be applied in CBA, it is therefore common practice to calculate a “reference-free” VTT. A way to achieve this is to include marginal effects for the sign of change in the cost and time attribute in the parameterization of VTT [e.g., the X -vector in Eq. (2)]. This transforms the systematic utility in value functions that can exhibit loss aversion. Parameters for the sign variables are estimated but then removed for the simulation of VTT distributions and calculation of mean VTT.

It is likely that the sign effect is amplified in SC, and in particular in pivoted CE where respondents are framed to a reference trips prior to the CEs. This makes the reference values for the time and cost explicit and easily recognizable in the choice cards, especially in the 2aCE where reference values are typically two of the four presented values.

Mode and User Type Effects

Estimating the VTT from route CEs that are based on reference trips, one infers VTT unit values that relate the current user group of the given transport mode (the VTT in bus for bus users or the VTT in train for train users). This makes it impossible to distinguish between mode effect that stems from the comfort level of travel models and the user type effect that stems from differences in opportunity cost of time (OCT) across user groups.

To disentangle the two effects one can ask respondents to answer CEs also for an alternative mode. This gives choice data for various travel mode/user group combinations. The marginal effects of this combination on the VTT can be estimated in a joint model by the inclusion of dummy variables in the X -vector in Eq. (2) in the case of the 2aCE.

It is then possible to simulate the mean VTT in different travel modes for any user group, including the “average” user group of all travel modes. (In practice, this requires some assumptions on how to weight the impact of different current user groups into the “average” user group.) This approach is adopted in the 2018 Norwegian Valuation study. It can be motivated by removing inconsistency in the VTT when considering transport projects that make travelers switch transport modes and by equity consideration for welfare analysis.

How to treat equity issues for welfare analysis is debated in the literature. The most radical practical approach is to use a generic value for all user groups and transport mode. However, this will come at the cost of lower precision.

Estimating VTT for Business Trip

The value of business travel time (VBTT) is different from the VTT for other purposes as it entails not only the private valuation of the traveler but also the employer’s valuation of time savings.

Most studies rely on (various variants of) the so-called Hensher formula that takes into account (relative) onboard work productivity (higher productive reduces the loss in the marginal product of labor) and the employee’s valuation of time savings. The former may be derived from different survey questions given that respondents are able and willing to self-report relative productivity. The latter can in principle be estimated from CE; however, the framing and interpretation of the cost attribute can be a challenge, as the employee is typically not paying for business trip for himself/herself.

Estimating VTT for Cycling (and Walking) for Transportation

As there are no out-of-pocket costs for using cycle, there is a challenge regarding the payment vehicle in valuation methods, that is, the cost attribute in CE. The typical approach is to use a mode choice setting where the marginal utility of money is inferred through the cost attribute of the paid-mode alternatives.

The VTT in cycle is likely to depend on the cycling infrastructure due to different levels of safety and cycling comfort. To avoid potential double counting in appraisal where safety improvements enter as a separated post in CBA, one would like to estimate the VTT in cycle as controlled for safety. This is most rigorously done when accident risk enters the CE as an independent attribute.

The VTT may also depend on the degree to which health effects of cycling are internalized by cyclists (the VTT becomes lower if travelers take the benefit of health into account). Isolating the health effects from the VTT is a recognized but methodologically not yet resolved challenge.

The mentioned challenges also apply to the VTT of walking for transportation.

Recruitment Effects

As mentioned in Table 1, representativeness is very important for national VTT studies. A central question is therefore how to recruit a sample of respondents that is most representative for the general travel population.

VTT-related questionnaires are typically involved and take some time to complete, typically between 10 and 25 min. A strong concern is that those that do not mind spending time answering a survey have a low OCT in general, which implies a low VTT. This sample selection bias is not rigorously testable as unobserved characteristics are likely to be main drivers behind the effect.

Empirical results from the latest Dutch and Norwegian VTT studies point to that members of internet panel have lower VTT than nonmembers, also after controlling of observables such as age and income. It is suggested that recruitment in field leads to a better representativeness compared to recruitment through internet panels.

Estimation With Revealed Preference Data

Having the multiple challenges of SC-based estimation in mind, it may surprise the reader why not more studies rely on revealed preference data when estimating VTT.

A principal challenge for the setup of the RC-based estimation model is that choice sets are not directly observed. Typically, the researcher has to assume which (other) alternatives the traveler considered and have to gather data on attributes for those alternatives from external data sources (e.g., network models). The inferred attribute values may be imprecise and/or not representative for attributes perceived by the traveler. In addition, RC is made in an uncontrolled environment, and many unobserved factors (attributes) may influence the RC.

For parameter inference, the main challenge with RC data is little variation and/or high correlation in attributes and explanatory variables. For example, for car trips and—to a lesser degree—for public transport trips, there is a very high correlation between travel cost and travel distance (and thereby travel time). This can result in statistically imprecise estimation and/or lead to that VTT estimates strongly vary with model specifications chosen by the researcher.

Traditionally, general data availability of RC data was an issue. With traditional travel survey data, for instance, the researcher observes only one (or a few) RC per respondent making parameter inference based on panel-data estimation techniques infeasible (or difficult).

The event of large-scale GPS/mobile data (possible combined with other “big data” sources), which is becoming more and more available for transportation research, may give new possibilities regarding the amount and quality of RC data. It may also be expected that advances in analyzing tools and techniques (e.g., artificial intelligent/neural networks) may help to define choice sets and infer parameters in large and complex RC data sets.

Further Reading

- Axhausen, K.W., Hess, S., König, A., Abay, G., Bates, J., Bierlaire, M., 2006. State of the art estimates of the Swiss value of travel time savings. Paper presented at the 86th Annual Meeting of the Transportation Research Board, Washington, DC, USA, 21–25 January, 2007. Available from: <http://hdl.handle.net/20.500.11850/39910>.
- Börjesson, M., Eliasson, J., 2012. The value of time and external benefits in bicycle appraisal. *Transp. Res. Part A Policy Pract.* 46 (4), 673–683, doi:10.1016/j.tra.2012.01.006.
- Börjesson, M., Eliasson, J., 2014. Experiences from the Swedish value of time study. *Transp. Res. Part A Policy Pract.* 59, 144–158, doi:10.1016/j.tra.2013.10.022.
- Flügel, S., 2014. Accounting for user type and mode effects on the value of travel time savings in project appraisal: opportunities and challenges. *Res. Transp. Econ.* 47, 50–60, doi:10.1016/j.retrec.2014.09.018.
- Fosgerau, M., 2005. Using nonparametrics to specify a model to measure the value of travel time. Paper presented at the 45th Congress of the European-Regional-Science-Association, Amsterdam, The Netherlands.
- Fosgerau, M., Bierlaire, M., 2007. A practical test for the choice of mixing distribution in discrete choice models. *Transp. Res. Part B Methodol.* 41 (7), 784–794, doi:10.1016/j.trb.2007.01.002.
- Fosgerau, M., Hjorth, K., Lyk-Jensen, S.V., 2006. An integrated approach to the estimation of the value of travel time. Paper presented at the Association for European Transport and Contributors. Danish Transport Research Institute.
- Hensher, D.A., 2001. Measurement of the valuation of travel time savings. *J. Transp. Econ. Policy* 35 (1), 71–98.
- Hess, S., Daly, A., Dekker, T., Cabral, M.O., Batley, R., 2017. A framework for capturing heterogeneity, heteroskedasticity, non-linearity, reference dependence and design artefacts in value of time research. *Transp. Res. Part B Methodol.* 96, 126–149, doi:10.1016/j.trb.2016.11.002.
- Kouwenhoven, M., de Jong, G.C., Koster, P., van den Berg, V.A.C., Verhoef, E.T., Bates, J., Warffemius, P.M.J., 2014. New values of time and reliability in passenger transport in The Netherlands. *Res. Transp. Econ.* 47, 37–49, doi:10.1016/j.retrec.2014.09.017.
- Schmid, B., Aschauer, F., Jokubauskaite, S., Peer, S., Hössinger, R., Gerike, R., Jara-Diaz, S.R., Axhausen, K.W., 2019. A pooled RP/SP mode, route and destination choice model to investigate mode and user-type effects in the value of travel time savings. *Transp. Res. Part A Policy Pract.* 124, 262–294, doi:10.1016/j.tra.2019.03.001.
- Small, K.A., 2012. Valuation of travel time. *Econ. Transp.* 1 (1), 2–14, doi:10.1016/j.ecotra.2012.09.002.
- Wardman, M., Chintakayala, V.P.K., de Jong, G., 2016. Values of travel time in Europe: review and meta-analysis. *Transp. Res. Part A Policy Pract.* 94, 93–111, doi:10.1016/j.tra.2016.08.019.
- Wardman, M., Lyons, G., 2016. The digital revolution and worthwhile use of travel time: implications for appraisal and forecasting. *Transportation* 43 (3), 507–530, doi:10.1007/s11116-015-9587-0.

The Taxation of Car Use in the Future

Griet De Ceuster, Inge Mayeres, Transport & Mobility Leuven, KU Leuven, Diestsesteenweg, Leuven, Belgium

© 2021 Elsevier Ltd. All rights reserved.

General Context	534
Current Taxation of Car Use	535
Exploring the Implications of Future Evolutions for the Taxation of Car Use	535
Demographic and Economic Growth	535
Fuel Efficiency and Electric Vehicles	536
Shared Mobility	536
Autonomous Vehicles	537
Developments in Road Pricing Technologies	538
See Also	539
References	539

General Context

The high share of car transport in OECD countries and its projected growing share in non-OECD countries is a reflection of the quality of the service, this mode offers in terms of door-to-door service, comfort, and flexibility. It also follows from the policy context in terms of prices, regulation, infrastructure provision, land use policies, etc. While car transport offers substantial benefits to its users and contributes to the smooth functioning of the economy, it nevertheless also generates costs to other transport users and society in general. The same is true for other passenger transport modes, but the level and composition of these costs differs across modes.

Travel choices involve many dimensions: the decision to own a car or not, or to become a member of a car sharing scheme or not, the choice of the car type, the decision how much to travel, where to travel to, by which mode, by which route, at what time. In all of these choices people take into account the costs and benefits to themselves. However, generally they only consider the costs to other transport users and society at large to the extent that they are confronted with them by government policies. Those other costs concern mainly congestion costs, accident costs, and environmental costs related to air pollution, greenhouse gas emissions, and noise. On the other hand, when people travel actively (by walking or cycling) they may also generate social benefits through the positive impacts on their health, to the extent they did not yet take this into account when deciding to walk or cycle. The costs and benefits that are not taken into account are referred to as external. Because of their presence the transport choices that people make are not optimal in the absence of appropriate policies: they travel too much, the modal share of the car is too high, they travel too much in the peak period, etc.

In the presence of externalities, government intervention is justified. Governments can make use of several instruments to mitigate the problems associated with transport. Among the pure transport policies, the following main categories can be distinguished: pricing, regulation, public transport supply, and infrastructure policy. Pricing refers to economic instruments such as road pricing, fuel taxation, subsidies to public transport, taxes on car registration or car ownership, insurance pricing, etc. Examples of regulatory measures include safety, emission or fuel efficiency standards for cars, access regulations to certain areas, speed limits, etc. Public transport supply concerns the type, frequency and capacity of public transport services. Infrastructure policy refers to the capacity of the infrastructure. Moreover, transport decisions are affected by policies in other domains such as land use planning or energy policy. These various policy categories can all contribute to ensuring that the full potential of transport is truly exploited while keeping its negative impacts under control. The rules for determining good policy interventions have been the topic of many analyses and are described in various other chapters of this Encyclopedia.

This chapter focuses on pricing policies for car transport. The first section points out that there is still a gap between the current pricing policies and the recommendations from transport economic theory. Second, it is explored how this gap might be affected by a number of future evolutions and what this implies for the role of pricing policies.

Inherent to future outlooks, there are still many uncertainties about relevant future evolutions as well as their impacts. This chapter should therefore be seen as exploratory. Five categories of future evolutions are considered: (1) economic and demographic changes, (2) the further increase in the fuel efficiency of cars and the growing importance of alternative technologies, (3) the role of shared mobility and of (4) innovations in connected and automated mobility and (5) developments in pricing technologies. This list is not exhaustive, but is meant to cover different aspects: changes in mobility behavior, business models, technologies and regulatory frameworks. While those various evolutions can be expected to interact, this chapter tries to distill the impacts they might have individually.

The discussion considers the range of objectives that governments want to achieve with transport policies: correcting for the externalities, as well as raising revenue, while also taking into account distributional issues. Depending on the political preferences in countries, as well as differences in external costs, the relative importance of these objectives differs from country to

country and so does their scope. For example in the case of the revenue raising objective, transport taxes and charges may serve to finance particular infrastructure or may be earmarked to spending in the transport sector, but may also serve to raise money for the general government budget.

Current Taxation of Car Use

Currently, the tax instruments for car transport that are used most are fuel taxes and taxes on vehicle registration and/or ownership. In its overview of energy taxation in 2015 in 42 countries, covering almost 80% of world energy use, the [OECD \(2018a\)](#) finds that for road transport 97% of CO₂ emissions are subject to a fuel tax (excise tax and/or carbon tax), and that respectively 50% and 47% was taxed at an effective rate of more than 30 and 50 euro per ton of CO₂. Moreover, if in addition to the excise tax and carbon tax a value added tax applies, the effective tax rate on carbon emissions is even higher. In countries with low fuel taxes, an increase in these taxes that is also in line with the carbon content of the fuels would be an efficient way to reduce carbon emissions.

In addition, most OECD countries have taxes on the purchase and registration of vehicles, as well as periodic taxes related to the ownership and use of the vehicles ([OECD, 2018b](#)). In combination with the fuel taxes, these taxes generate substantial government revenues. The purchase taxes and periodic taxes are commonly differentiated in terms of vehicle characteristics, weight, engine power, age, cylinder capacity, type of fuel or electric propulsion, emission, or fuel efficiency standard, etc. While in the past the differentiation was mainly done with efficient revenue raising and equity purposes in mind, in the last decades environmental objectives have been integrated as well.

Fuel taxes are able to directly address the CO₂ emissions that are related to fuel consumption. However, for a given fuel category they cannot make a distinction between more and less polluting vehicles for non-greenhouse gas emissions. Moreover, if countries have different ambitions regarding the reduction of CO₂ emissions, fuel tourism and tax competition complicate the matter ([Mayeres, 2003](#)). Vehicle taxes can play a role there since they can be differentiated according to the emission characteristics of the vehicles. Purchase taxes/subsidies have an immediate effect when people buy a vehicle, but in the case of taxes, the prospect of having to pay these taxes may also encourage people to hold on to their older (more polluting and less fuel efficient) vehicles longer. Periodic vehicle taxes may therefore be more efficient to lead people toward cleaner vehicles, in combination with fuel taxes.

Additional tax policies that have an impact on car use are the fiscal treatment of commuting costs and of travel for private purposes using a company car. A favorable tax treatment of commuting expenses may however lead to longer commuting distances, and may affect as well modal choice, depending on the modes that are covered by it. In countries where the private use of company cars is under taxed this may lead to social costs in various ways, depending also on the fiscal rules that apply: by leading to more and more expensive cars, by increasing the car kilometers traveled and therefore the associated external costs, and by leading to a loss in government revenues.

Many cities in the world but also interurban corridors face considerable congestion problems, as is illustrated for example by the INRIX Global Traffic Scorecard indicator that is published annually. Fuel taxes are only an imperfect instrument to tackle congestion. A pricing system that is able to differentiate prices according to the location and time of day is preferable in this case ([Anas and Lindsey, 2011](#); [Mayeres, 2003](#)). Currently, there are not many examples yet of such road pricing systems. There are a still limited number of cities that have implemented road pricing, using cordon tolls or area licensing schemes (e.g., Singapore, London, Milan, Stockholm, Gothenburg, . . .). On some corridors high occupancy toll lanes are implemented, mainly in the United States but also in some other countries. There are also some examples of cities or countries that have explored road pricing, but where plans were halted in the end (e.g., the United Kingdom, the Netherlands). Other examples of road pricing for cars apply for specific infrastructures such as bridges, tunnels, or stretches of roads but are geared towards revenue rising rather than tackling congestion.

Exploring the Implications of Future Evolutions for the Taxation of Car Use

Demographic and Economic Growth

If population grows so will, *ceteris paribus*, mobility demand. In addition, the evolution of mobility demand will depend on the age and gender structure of the population. The global population is still expected to grow in the future, though at slower rates than before. A large part of this growth is projected to take place in urban areas, thereby increasing the challenges related to urban mobility. For evaluating the impact of these changes on transport demand, one has to be careful not to project the mobility patterns of the past to the future. Indeed the mobility pattern of age classes and men/women also changes over time. For example older people now have a higher mobility demand than older people in the past. Women now also have different mobility patterns than in the past. The demographic changes are taking place together with economic changes.

Mobility demand is positively related to economic growth, *ceteris paribus*. Economic growth leads to more freight transport demand. In situations with higher economic growth, there generally is more employment and higher income, both leading to higher mobility demand. Due to better infrastructure the possibilities to travel farther are also larger. As a result of all of these evolutions the 2017 Transport Outlook of the International Transport Forum ([OECD/ITF, 2017](#)) projects a substantial increase in road transportation worldwide, especially in non-OECD countries. The main driver is said to be the economic development. The baseline scenario of the outlook, which does not yet consider a possible uptake of new mobility services and which assumes a moderate increase in oil prices, projects a doubling of the annual mileage per capita for domestic non-urban transport in non-OECD countries between 2015

and 2050, while in OECD countries this would only change marginally. Urban mobility would be almost 95% higher in 2050 than in 2015, with the increase being concentrated mostly in non-OECD countries. All of this points to a growing need for policies that can mitigate or reduce the negative impacts of this larger transport demand, with a role to play for the different policy categories, including not only pricing but also the other categories mentioned above.

The economic development also implies, *ceteris paribus*, a higher value of travel time and of the other external costs, given the positive income elasticity of these valuations. Still, there might also be evolutions with opposite effects, as discussed in the section on connected and automated mobility. In case of the health and nature impacts of transport emissions, scientific knowledge can also be expected to evolve further, potentially leading to higher valuations of the external costs.

Fuel Efficiency and Electric Vehicles

Future carbon emissions by transport depend on the evolution of mobility demand on the one hand and the evolution of fuel efficiency and the uptake of alternative vehicles and fuels on the other hand. A global overview for 2017 by the International Council on Clean Transportation (Yang and Bandivadeka, 2017) shows that markets covering together slightly less than 80% of passenger vehicle sales (Brazil, Canada, China, EU, India, Japan, Mexico, Saudi Arabia, South Korea, and the United States) have in place regulations concerning the fuel efficiency or greenhouse gas emissions of cars. Three other major markets were planning regulations at that time (Australia, Thailand, Vietnam).

It should be noted that literature has pointed out such fuel efficiency or emission standards should be used with care because they can lead to smaller welfare gains than pricing policies (Parry et al., 2014). Still, they can have a role if fuel taxes cannot be set optimally, or in the presence of misperception market failures. For the latter aspect a recent empirical study however points out that car buyers to a large extent figure in future energy savings in their purchase decisions (Grigolon et al., 2018). Another reason for using standards is that they encourage research and development in innovative technologies that take a long time to develop as they offer the possibility to create a more certain long-term policy framework than pricing policies that can be changed more easily.

The further tightening over time of these standards that is in place or being proposed for road transport is expected to have a significant effect on future global oil consumption by the sector on condition that the necessary compliance and enforcement framework is put in place. Without higher fuel efficiency and without the uptake of alternative fuels and vehicles the 2018 World Energy Outlook of the International Energy Agency projects oil demand by the road transport sector to increase by 68% between 2017 and 2040 (OECD/IEA, 2018). However, if account is taken of future improvements in fuel efficiency, the growing use of alternative fuels and natural gas and the larger uptake of electric vehicles, the increase is limited to about 10%. This lower growth in oil demand is realized to a large extent by stricter fuel efficiency and emission standards and by improvements in engines and hybrid technologies: this brings down the increase from 68% to 32%. Electric vehicles lead to a further reduction by 12% points and the rest is realized by using more alternative fuels and natural gas.

Fuel efficiency standards and commonly used policies to promote the uptake of electric vehicles such as purchase subsidies, or benefits in the form of reduced charges for parking or the allowance to use bus lanes, lead to a reduction in the variable costs of road use. *Ceteris paribus*, this can be expected to lead to a rebound effect leading to an increase in road transport demand. Because of that part of the emission reductions will be offset and the other external costs of transport that are related to transport volume are also exacerbated. This implies that road pricing will have to play a larger role in order to tackle the remaining externalities. In that case the same principles of road pricing should apply to all vehicle types (i.e., also for electric vehicles), though the actual tariffs might of course differ in terms of the external costs they create.

Another consideration is that, at present, fuel taxes are an important source of government revenue. Also in some countries the income from the fuel taxes is earmarked specifically to finance transport purposes. With improved fuel efficiency and more electric vehicles for which energy taxes are typically low, this implies that alternative sources of funding have to be explored. Road pricing then also comes more prominently in the picture as a revenue raising instrument. An example of the awareness of these implications is found in a number of US states, where there is a growing interest in mileage based user fees, as a result of the projected fall in fuel tax revenues and high financing needs for maintaining and upgrading roads.

From the public economics literature it is known that in a second-best setting in which the government has to use distortionary taxes, the optimal tax on car use deviates from the Pigouvian tax. It then consists of both revenue raising component as well as a component that corrects for any externalities that arise, with both components incorporating distributional considerations.

Shared Mobility

While car travel by a privately owned car is still dominant now in developing countries and is also expected to become more important in developing countries, developments in ICT are leading to changing concepts of mobility, including options for shared mobility. Shared mobility covers a wide range of services: better options for organizing carpooling, new options for sharing rather than owning vehicles, on-demand ride services and micro-transit. The concept also refers to the availability of apps that enable and integrate the use of these transport services. These concepts can be briefly described as follows: (Franckx and Mayeres, 2015):

- With vehicle sharing members have short-term access to vehicles. Systems exist for cars but also other vehicles such as (electric) bicycles or electric kick-scooters. These vehicles can either be available at fixed stations or free-floating. The vehicles are often owned by companies, but for cars also peer-to-peer systems exist.

- With ridesharing, or carpooling, people who want to travel together are matched with each other, with the system functioning on a non-profit basis.
- In the case of on-demand ride services Transport Network Companies (TNCs) supply apps that in real-time bring into contact passengers with drivers and determine the price of travel. They can also offer ride-splitting, where passengers share a ride with someone else.
- Micro-transit is a form of demand responsive transit that serves a market segment between that of public transport and on-demand ride services. In this system transport services are supplied by minibuses with flexible routing and/or scheduling.
- Mobility-as-a-Service (MaaS): While car travel by a privately owned car is still dominant now, the idea behind MaaS is to supply comparable travel options, in which car ownership is no longer the starting point. The travel options are designed to meet the specific travel demand of the transport users, using different combinations of modes. All information on services and prices is available in the on-line platform of a mobility broker, where the transport services can be booked and the payment can be done. Depending on the system, individual solutions for transport can be combined (or not) with public transport services.

To the extent that the shared mobility services lead to lower transport demand, to a shift to more sustainable modes or to more environmentally friendly vehicles, they may reduce the external costs of transport and therefore the need for corrective pricing. Whether that is the case and to what extent is still to be fully understood. That is because these systems are still in full development and the behavior of the early adopters cannot be generalized to the wider group of transport users. Nevertheless, one can identify several mechanisms through which shared mobility may have an impact on the external costs of transport.

With car sharing and on-ride demand services prospective transport users are confronted with the costs of their trips each time they book a trip. In addition, with car sharing it takes more effort to use a shared car than a privately owned car: the car has to be reserved, one should walk or cycle to the place where it is parked, it should be brought back in time for the next user, etc. All of this may lead to less car km by the individual car user. Moreover, by providing a solution for the transport to and from public transport stations, car sharing or on-demand ride services can increase the modal share of public transport for longer trips; this argument also holds for other shared mobility systems that improve the first and last mile transport options (shared bicycles, kick-scooters). Ridesharing and ride-splitting reduce vehicle km by increasing the occupancy rate of the vehicles. In all of these cases it should be noted however that, if prices are not correct, a reduction in car km by some car users, leading to less congestion, may encourage other people who were previously discouraged to use their car to switch to their car again. So the case for road pricing in areas with congestion remains, though the optimal road prices may be affected.

With car sharing there is an additional advantage of the possibility to “right-size” the car, which increases the average energy efficiency. The fact that the shared cars are used more intensively than privately owned cars also leads to a quicker replacement and hence newer (less polluting) cars. There are also synergies with electric vehicles, as these are more interesting with higher annual mileage, resulting from many shorter trips in areas that have a high density of charging points. This is a market segment that can be served well by car sharing. Other synergies, with automated electric vehicles, may also exist, as discussed in the next section. It is also pointed out sometimes that if the companies supplying the car services are more sensitive to fuel costs than households, then the fuel efficiency of the shared cars could be larger. However, recent empirical evidence suggests only a modest undervaluation of the future fuel savings of more fuel-efficient cars.

There are also potential negative impacts on the external costs, because car sharing or on-ride demand services lower the barrier to use a car for people who previously did not have access to the car. Moreover, by reducing (or even eliminating) parking costs and searching time for parking it can make car use more attractive. Also, if the systems are devised as a substitute rather than a complement to public transport or other sustainable modes they may also lead to a modal shift in the wrong direction.

Important to note is that in all of these cases the decisions are still taken by the individual transport users, based on a consideration of the costs and benefits for themselves. The same is the case for the companies providing the services that are aiming at profit maximization. The latter point also means that even if the companies apply peak load pricing, this will also be based on private profit maximization.

Therefore, government intervention is required in order to guide the decisions in the correct direction from the social point of view.

Another potential effect of shared mobility is related to the public acceptability of road pricing, which is commonly quite low when plans for road pricing are discussed in public debate. The intermediary role of car sharing companies, TNCs, and MaaS providers may help to improve the public acceptability of road pricing (Hensher, 2018), when the charges are paid more indirectly via these companies, compared to the situation where they are levied directly on the private car users. As users are paying a service, shared mobility will make it easier to introduce road pricing, and to adapt tariffs by time and location. The service bought by users will include all costs of ownership per km, including road pricing and will contain choices towards more environmental friendly vehicles. Even more, it will become more obvious to the user that using a car during peak hours is more costly.

Autonomous Vehicles

Autonomous vehicles (also called self-driving vehicles) can provide benefits to the user, such as providing independent mobility for people that cannot or should not drive, typically 10%–30% of the population, which increases their options for various activities. Benefits to ex-drivers are mainly found in safety and comfort: driving time, which is on average 1 hour per day, becomes now time free to spend on other activities. It is expected that autonomous vehicles will increase transport demand, both due to the increase of

ex-non-drivers now driving, and of ex-drivers now driving more. Some studies estimate that transport demand might increase by up to 14% (Schoettle and Sivak, 2015; Trommer et al., 2016). In the long run, autonomous vehicles might encourage urban sprawl and even further increase transport demand (Milakis et al., 2017).

It has to be noted that we are still far from fully autonomous vehicles. Commonly, six technology levels are defined, where level 0 has no automation and level 6 is fully automated. Levels 1 (assisted driving such as cruise control) and 2 (monitored driving, e.g., automated parking and lane keeping) already exist. Up to level 3, a driver is still required as a fall back. Level 4 has a steering wheel for some driving modes. Level 5 has no steering wheel. Experts acknowledge that significant technical progress is needed before level 5 automation is reliable, tested and approved. Even if level 5 technologies penetrate new vehicle markets in the 2020s, it will be the 2040s or 2050s before most vehicles are capable of autonomous driving.

Shared autonomous vehicles have the potential to reduce private car ownership dramatically. Cars are currently used on average 1 hour per day, which is expected to improve to 5 hours when fully automated and shared. Larger amounts are unrealistic due to the nature of travel demand showing peaks in the morning and evening. However, part of this benefit will be lost due to empty driving while floating around for passengers (Henao and Marshall, 2018).

For the user, the cost will be shifting from a high fixed cost and a low variable cost towards an almost full variable cost. Compared to current shared vehicles automated vehicles will be more convenient because they can be available at the doorstep.

It is uncertain if the average cost per km will increase or decrease. Automated technology is more expensive than manual driving, but the sharing of the vehicle, and the more efficient use might lead to a reduction in costs. In any case, automated vehicles will be cheaper than taxis and public transport that require a paid driver. Estimates vary from USD 0.1 up to USD 1 per mile, depending on the transport mode.

More important, the value of time savings for drivers might drop drastically. According to Rodier (2018) the effects vary widely, but 75%–82% of current driver values of time may be reasonable.

Despite the possible increase in traffic, autonomous vehicles that are shared and thus used more often, have the benefit of a shorter life span enabling a faster penetration of new technologies, when available (see also the section on shared mobility). However, it is at the moment unclear if autonomous vehicles will lead to lower emissions. According to the World Energy Outlook of the International Energy Agency the consequences of automation on long-term energy demand and emissions could go in different directions, as various factors influence the final outcome (OECD/IEA, 2018).

Congestion levels can go up or down (Milakis et al., 2017), mostly depending on the technology (how platooning of autonomous vehicles will affect the road capacity) and the occupancy rates (efficient sharing of vehicles). A new type of external costs that will occur is related to the availability of vehicles during peak hours. An optimum has to be found in providing enough vehicles to cope with peak demand, and the higher costs of not using such vehicles during a large part of the day.

In general effects on traffic congestion, energy consumption, pollution emissions, roadway and parking facility costs are possibly beneficial, but uncertain and depending on the increase in transport volumes. As the value of time is expected to change downwards, and impacts on emissions and congestion are uncertain, it is hard to predict the levels of road pricing needed to balance out the external effects of transport.

Developments in Road Pricing Technologies

Road user charging is now widely implemented throughout the world. The charging schemes, however, vary considerably (Walker, 2018). At the moment, most charging is either a time-based vignette, a kilometer charge on the main motorways, or a city access charge. No large area-wide kilometer charges exist yet. The reason for this is partly technical: the only feasible way to charge all roads is to use satellite navigation based on-board units. With this, each car is equipped with a device that sends its satellite positions regularly to a central system that handles the mapping of the positions on a road network, and calculates the charge. Such systems exist for truck charging on motorways and main roads in Germany since 2005, with five European countries following later. Singapore is planning to use it for its motorways for passenger cars from 2020 onwards.

The most widely used technology at the moment is short range and microwave communication, with DSRC (dedicated short-range communication) as the most popular standard. With this, a vehicle is equipped with a tag, a small chip costing EUR 5–20, that can be identified with roadside equipment. Each time a vehicle passes such a physical barrier, it is charged. A large amount of motorways across the world is equipped with such a tolling system among which the US East coast states, France, Italy, Portugal, Norway, Dubai, South Africa, and the current system in Singapore. Though this system is easy to use and cheap for simple road networks, it is not expandable to the large network rural and urban roads due to the large amount of roadside equipment it would require.

Moreover, microwave communication is even becoming less popular for motorways and urban access roads, while ANPR (automatic number plate recognition) systems are slowly becoming the standard. With this, cameras at the roadside register every passing vehicle. No device is needed in the vehicle, making this the easiest technological system to roll out. As camera quality has improved a lot since the first tolling based on ANPR was introduced in London in 2003, most urban systems are now equipped with ANPR, and even some motorway sections now use it. However, it is a costly approach for large and complicated networks, as this would require a large amount of roadside poles and portals each of them equipped with a camera.

As stated before, road charging with satellite navigation based on-board units is the future and a large amount of countries are studying its use. The main challenge is the price of the device, now costing about EUR 60–100 each, and dropping. An interesting approach would be to use the drivers' own mobile phones and/or the in-car telematics as a device, making it basically for free.

However, this would require a certified app that can meet the demanding legal standards for taxation. Such developments are expected in the near future with a possible first application around 2025.

See Also

Demand for Passenger Transport; What Drives Transport and Mobility Trends? The Chicken-and-Egg Problem; The Concept of External Cost: Marginal Versus Total Cost and Internalization; Principles of Pricing in the Transport Sector

References

- Anas, A., Lindsey, R., 2011. Reducing urban road transportation externalities: road pricing in theory and in practice. *Rev. Environ. Econ. Policy* 5 (1), 66–88.
- Franckx, L., Mayeres, I., 2015. Future trends in mobility: challenges for transport planning tools and related decision-making on mobility product and service development, Deliverable 3.3 of the MIND-Sets project and Annex, project financed by the EU Horizon 2020 program. Available from: <http://www.mind-sets.eu/>.
- Grigolon, L., Reynaert, M., Verboven, F., 2018. Consumer valuation of fuel costs and tax policy: evidence from the European car market. *Am. Econ. J.: Econ. Policy* 10 (3), 193–225.
- Henao, A., Marshall, W., 2018. The impact of ride-hailing on vehicle miles traveled. *Transp.* 45(5), 1269–1295. DOI: 10.1007/s11116-018-9923-2.
- Hensher, D.A., 2018. Tackling road congestion—What might it look like in the future under a collaborative and connected mobility model? *Transport Policy* 66, A1–A8, doi:10.1016/j.tranpol.2018.02.007.
- Mayeres, I., 2003. Taxes and transport externalities. *Public Fin. Manage.* 3 (1), 94–116.
- Milakis, D., van Arem, B., van Wee, B., 2017. Policy and society related implications of automated driving: a review of literature and directions for future research. *J. Intell. Transp. Sys.* 21 (4), 324–348, doi:10.1080/15472450.2017.1291351.
- OECD, 2018a. Taxing Energy Use 2018, Companion to the Taxing Energy Use Database, OECD Publishing, Paris, France. Available from: <http://dx.doi.org/10.1787/9789264289635-en>.
- OECD, 2018b. Consumption Tax Trends 2018, VAT/GST and excise rates, Trends and Policy Issues, OECD Publishing, Paris, France. Available from: <https://doi.org/10.1787/ctt-2018-en>.
- OECD/IEA, 2018. World Energy Outlook 2018, International Energy Agency, Paris, France.
- OECD/ITF, 2017. Transport Outlook 2017, International Transport Forum, Paris, France.
- Parry, I.W.H., Evans, D., Oates, W.E., 2014. Are energy efficiency standards justified. *J. Environ. Econ. Manage.* 67, 104–125.
- Rodier, C. 2018. Travel Effects and Associated Greenhouse Gas Emissions of Automated Vehicles, UC Davis, National Center for Sustainable Transportation White Paper, NCST and UC Davis, Institute of Transportation Studies, University of California, Davis, 2018
- Schoettle, B., Sivak, M., 2015. A Preliminary Analysis of Real-World Crashes Involving Self-Driving Vehicles, Report UMTRI-2015-34, Transportation Research Institute, University of Michigan. Available from: <http://umich.edu/~umtristw/PDF/UMTRI-2015-34.pdf>.
- TRB, 2019. Conference Proceedings 56: Socioeconomic impacts of automated and connected vehicles, Summary of the Sixth EU-U.S. Transportation Research Symposium, National Academies of Sciences, Engineering and Medicine, Washington, D.C.
- Trommer, S., Kolarova, V., Frädrieh, E., Kröger, L., Kickhöfer, B., Kuhnimhof, T., Lenz, B., Phleps, P., 2016. Autonomous Driving: The Impact of Vehicle Automation on Mobility Behavior, Institute of Transport Research, BMW, Munich, Germany.
- Walker, J. (Ed.), 2018. Road Pricing, Technologies, Economics and Acceptability, IET Transportation Series 8, The Institution of Engineering and Technology, London, UK. ISBN 978-1-78561-205-3.
- Yang, Z., Bandivadeka, A., 2017. 2017 Global Update, Light-Duty Vehicle Greenhouse Gas and Fuel Economy Standards, the International Council on Clean Transportation, Washington, D.C., US.

Cost–Benefit Analysis and Other Assessment Techniques: Contrasts and Synergies

Paolo Beria, Politecnico di Milano, Milan, Italy

© 2021 Elsevier Ltd. All rights reserved.

A Taxonomy of Appraisal Approaches	540
(Social) Cost–Benefit Analysis (CBA)	540
Cost Effectiveness Analysis (CEA)	541
Multiple-Criteria Decision-Making (MCDM) or Simply Multiple-Criteria Analysis (MCA)	541
Input Output models (I/O) or Added Value Analysis	541
Environmental Impact Assessment (EIA)	541
(Spatial) Computable General Equilibrium Models (SCGE)	541
System Dynamics Models (SD)	542
Cross-Approach Issues	543
Transparency, Complexity, and Reproducibility	543
Equity, Distribution, and Conflicts	544
Political Acceptance	544
CBA and Friends: When Interaction is Possible	545
CBA and Transport Models	545
CBA and MCA	545
CBA and CGE or LUTI	545
CBA and EIA	545
References	546

A Taxonomy of Appraisal Approaches

Cost–benefit analysis (CBA) is the most used assessment approach in public decisions concerning transport, but others exist. They differ substantially for disciplinary background, purpose, technicalities, but at the same time they share some concepts, typically the idea of weighting advantages and disadvantages of an intervention, or represent different facets of similar approaches. This translates into typical fields of applicability, but also in interesting potential cross-fertilizations. In this section, we will propose a taxonomy across assessment approaches, showing when they are really different and when they just represent different points of view.

Among the main dimensions that differentiate approaches, there are:

- The disciplinary background;
- Which markets are considered (analysis boundary);
- If the analysis is spatialized or not; and
- Who sets the weights/define what is benefit and cost.

It is important to notice that some of the following approaches have broader applications than (transport) project assessment. Consequently, looking at them as an “assessment technique” requires a redefinition and in some cases also a sort of stretching to adapt to the usual problems of transport economist deciding about one new transport infrastructure or a change in a mobility policy. For example, a CGE describes quantitatively an economy as a whole, by means of mathematical models, feedbacks, and elasticities. It is primarily a *model* and it is used to simulate what happens when something changes—a new road is build, for example, or a tax is risen. One can decide that when certain indicators are met, a project is worthwhile, making it also an assessment technique. But this is not directly an evaluation in the sense of CBA.

It is useful to start with a concise definition of the key features of the main approaches used.

(Social) Cost–Benefit Analysis (CBA)

CBA is a well-established method, founded in welfare economics, based on the quantification of all direct (e.g., time savings, investment costs, etc.) and indirect (e.g., pollution reduction benefits) effects of a transport investment, in terms of marginal utility variation, and their consequent monetization and discount. CBA assumes that the variation of utilities contributes to social surplus, whose maximization is the measure of worthwhileness. This means that the users’ benefits estimation is based on their evaluation/perception and not on an abstract measure defined by the analyst. The main use of CBA is to measure if the public intervention (including policies) improves the societal surplus and how efficient is the associated expenditure. Efficiency is clearly just one criterion among others, but this translates into a solid and univocal “meaning” of its results, often absent in other techniques. In

other words, the point of view of CBA is holistic, while other techniques tend to manage different points of view (MCA) or are not explicitly made to represent *one* point of view (any modeling exercise). Two main hypotheses lay behind CBA: the Kaldor–Hicks criterion (a measure is welfare-increasing, if benefits by winners exceeds the losses by losers) and the assumption that nontransport markets are not significantly affected (e.g., the price of energy does not change due to the project) or are perfect.

Cost Effectiveness Analysis (CEA)

CEA shares with CBA some operational features (quantification, monetization, discount, indicators), but is not assuming a superior criterion of choice, such as surplus maximization. In fact, it measures how effective is an expenditure in obtaining a result, whatever is the result and its measure. For example, CEA can either measure the cost-effectiveness of reducing CO₂ emissions, but also of increasing them (if someone aims to do that). In other words, CBA assumes intrinsically that some outcomes are benefits and other are costs, while CEA does not. Typically, they are set by the analyst or by politics (e.g., “road accidents must be halved”). Consequently, it is advisable to use it when it is clear and shared what is the desirability of the outcome measured.

Multiple-Criteria Decision-Making (MCDM) or Simply Multiple-Criteria Analysis (MCA)

MCA, as the name suggests, aims at dealing with multiple (and even conflicting) criteria, typically reflecting the objectives of different stakeholders. It is founded in operational research ([Charness and Cooper, 1961](#)) and consists in a ranking based on the optimization of criteria and relative weights. Monetization is not necessary and any quantitative or qualitative input can be in principle used. There is not one way to perform a MCA, but many methods and algorithms, each one with its characteristics, pros and cons ([De Brucker et al., 2011](#); [Macharis and Bernardini, 2015](#); [Pérez et al., 2015](#)). For example, basic MCA applications assume one decision maker, while current applications commonly take into account the multiplicity of stakeholders. Also different weighting methods and optimization algorithms exist. In MCA, what is a benefit is explicitly decided by the analyst (in simplest exercises, by the decision-maker or by involved stakeholders).

Input Output models (I/O) or Added Value Analysis

These tools, based on aggregate statistics of interdependencies between the branches of an economy or different regional economies, calculate the increase in added value of production, ignoring other effects such as externalities. Using this approach as an assessment technique means quantifying how 1 € of expenditure, which is 1 € of increase of direct production of some sectors, translates into (possibly) more than 1 € of outputs in other sectors. This apparently intuitive concept must be accompanied by some alerts. First of all, the use of this approach to evaluate the impact of one investment is basically pointless, because if substitution effects are ignored, *every* investment gives a positive added value. Secondly, the “cost” of this increase of production is not considered. If it comes from the external of the economy (foreign investment), the output could be entirely an added value, but if the expenditure comes from the inside, it is not “free” and probably requires an increase of taxation or a decrease of expenditure on other sectors. Thirdly, this technique assumes the invariance of productive structure and can therefore be used only for the very short term. Finally, it assumes a direct causality between inputs and outputs, which is not always the case.

Environmental Impact Assessment (EIA)

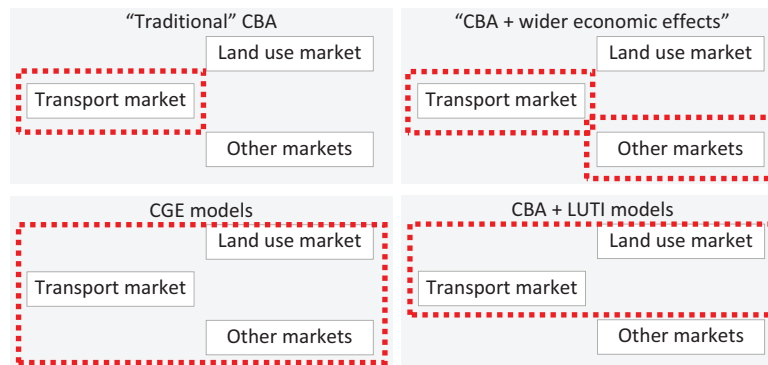
EIA is not an assessment technique in the sense of the previous ones. It describes the possible impacts of a project on environment, human activities, cultural heritage, and everything that is considered as scarce. EIA’s main output is the evidence of impact and their minimization, but this can assume the shape of an evaluation when these are considered as constraints or stakeholders’ goals. Clearly, this approach does not guarantee comparability among projects or the evaluation of trade-offs, which is the core, for example, of CBA.

(Spatial) Computable General Equilibrium Models (SCGE)

These models simulate what happens in an economy when something changes, through a series of equations expressing the production and market relations among parties ([Bröcker, 2014](#); [Bröcker et al., 2010](#)). Therefore, CGE do not intrinsically express a judgment. Using them as an assessment technique requires to decide what is the measure to be maximized and usually it is social surplus, like for CBA. The main differences with CBA are that they model all markets and can then be used to measure what happens outside of the transport market alone. They also focus on *households* and producers surplus, rather than on *users* and producers surplus. Aside to these interesting features, they require some alerts. Their formalization is not univocally defined and every CGE is different, according to author’s choices and available data, losing comparability of results. They must be calibrated with real world data to be usable, which is not simple at the scale required for transport modeling. They are very complex (and data demanding) because they must model the feedbacks between transport market and other markets, ignored by CBAs, and among regions.

Table 1 Main characteristics of assessment approaches

	<i>Foundation</i>	<i>Main point of view</i>	<i>Measure</i>	<i>Who sets weights</i>	<i>Assessment goal</i>
CBA	Microeconomics	Society as a whole	Welfare changes (via social surplus)	Users' choices	Economic efficiency of public expenditure
CEA	Microeconomics	Society as a whole	Specific performance indicators	Analyst/politician	Economic efficiency of public expenditure
MCA	Op. research	Stakeholders	Stakeholder's criteria	Stakeholders	Optimization of stakeholders' criteria
EIA	Env. sciences	Environment	(Environmental) impacts	Analyst	Minimization of environmental impacts
I/O	Macroeconomics	Country/region	Added value or GDP or employment	Cross-sector matrix	Maximization of added value, GDP, or employment
CGE	Microeconomics	Country/region	GDP, employment, welfare changes	Equilibrium	Simulate economy
SD	Op. research	Country/region	Surplus and specific indicators	Cross-sector matrix	Simulate spatial and transport assets

**Figure 1** Different boundaries of approaches.

System Dynamics Models (SD)

SD models (Rothengatter, 2017; Shepherd, 2014) may appear as CGE models, but with substantial differences. Similarly to CGE, they aim at modeling an entire economy to measure potentially any variable included (quantities, prices, welfare, etc.). Differently from most CGE, they are not economically micro-founded, they are not solved analytically but numerically and they do not imply an equilibrium, which in effect is not always true in complex phenomena like transport-economy relationship. Similarly to the other models, SD are not exactly assessment techniques based on a criterion, but being able to predict quantitatively the state of a system, they can be used to calculate indicators and/or to allow comparisons among alternatives.

The following tables schematize the main features of assessment approaches mentioned. Table 1 includes details about the disciplinary foundation of each approach, the main point of view entailed by the analysis, the measure at the basis of the assessment and the goal pursued by the approach. This last column is very important and clarifies the root of the judgment expressed apparently aseptically by an approach. For example, CBA pursues a general goal of economic efficiency of public expenditure measured through surplus changes, while MCA finds a “solution” best matching the different and contrasting points of view of the stakeholders involved. CGE and SD models do not assume a specific goal, but “simply” simulate what happens in the modeled systems. The judgment criterion is thus exogenous and in fact they are often used together with CBA.

Fig. 1 schematizes which are the relevant markets considered, which is a key feature of approaches and a powerful source of misunderstandings among experts. In fact, CBA assumes that most of effects can be measured in the transport market only. When this is not considered an acceptable proxy, wider economic effects are added. The same do LUTI models that include the land-use market into the transport model. The principle is that they include the *logsum* of transport choice models into a location choice model. This translates the fact that, in the long-run, location choices are no more independent variables, but are included in the feedback and consequently modeled and included among the impacts. CGE models instead assume that all interactions can be relevant and model an entire economy, however usually losing the spatial dimension that transport CBAs normally include.

Table 2 introduces further elements of comparison. All approaches are quantitative or allow treating also quantitative aspects. However, MCA and EIA can (and mostly) manage qualitative aspects. MCA and I/O are clearly dealing with a short-term perspective, the first because representing current stakeholders' view, the second because assuming constancy of productive structure. The other, with all the limits of long-term forecasting, aim to consider decades of effects. EIA and I/O are also unsuitable for ex-post analysis, while the remaining could—in principle—be used also to assess past choices, even if this is seldom done. The kind of output is very different. CBA, CEA, I/O, and MCA, despite different, create indicators and rankings whose function is to

Table 2 Main characteristics of assessment approaches

	<i>Quali/quantitative</i>	<i>Temporal scope</i>	<i>Ex-post</i>	<i>Output</i>	<i>Distribution matters?</i>	<i>Costs are</i>
CBA	Quantitative	Long term	Possible	Indicator	Not directly	Costs
CEA	Quantitative	Long term	Possible	Indicator	Not directly	Costs
MCA	Both	Short term	Possible	Score	Yes	Depends
EIA	Both	Long term	No	Prescriptions	Yes	Neutral
I/O	Quantitative	Short term	No	Indicator	Across industries	Benefits
CGE	Quantitative	Long term	Possible	Many	Yes	Costs
SD	Quantitative	Long term	Possible	Many	Yes	Costs

aggregate different aspects into one measure, that “automatically” inform the decision. EIA provides prescriptions based on the minimization of impacts, but no judgment or ranking of alternatives. Models are not aggregating results in one specific measure and can provide outputs on different aspects of the systems. Also for this reason they are not *exactly* assessment techniques, because not assuming one criterion.

Two other aspects are of some interest. Distribution of impacts, being one of the most relevant political criteria, is not the focus of all approaches and this—in a certain sense—explains why some approaches are more appreciated/understood by decision makers than others. CBA is intrinsically ignoring the distribution of impacts (rather summing all of them as if they were equivalent), but some recent applications split the NPV into NPV_i for the i involved parties (users, state, producers, polluted, etc.). The way costs are considered is also very interesting and explains a lot of the acceptability of assessment. Microeconomic approaches consider costs as *costs*: an expenditure enters in the assessment with a *minus* sign. Models in general and EIA are instead cost-neutral: they do not evaluate trade-offs. MCA and I/O models are more interesting. MCA assumes the criteria of the decision-makers. It is not uncommon that a local authority, for example, considers positive that an infrastructure (typically if paid by central administration) costs more, as it expects more benefits for the territory. I/O models calculate the multiplicative effect of an expenditure in the economy and then, in principle, any expenditure gives an added value. Therefore, an I/O model applied to one project alone, says that the more it is expensive, the better it is.

Cross-Approach Issues

Transparency, Complexity, and Reproducibility

There is a clear trade-off between transparency of results (i.e., how clear are the results and assumptions for a reader which is not the author of the analysis) and complexity. The apparently infinite computational capacity pushes analyses toward an extreme complexity, as if we were drawing a 1:1 map of the world. But an extreme complexity entails the impossibility to understand what is behind, in particular assumptions, limits, mathematical models, and feedback effects. This makes the assessment an apparently informative tool, but assumes a sort of *faith* in the reader. And when the readers are politicians and public opinion, faith could not be available and assessments can be refused or accepted uncritically according to one’s opinion, which is the exact opposite of the purpose of an assessment.

Therefore the key aspect—if we do not want to renounce to complexity—becomes credibility. A continuous, independent, comparative, and well-documented use of an approach makes it more credible. Ad-hoc evaluations, not documented and not reproducible, opaque, will be probably rejected by part of the readers. Here a list of elements that make an assessment exercise more accepted, whatever is its result:

1. Stable guidelines and methodology in general;
2. Continuous and comparative use, not project-specific;
3. Independency of the evaluator from the vested interests of the project assessed;
4. Separation of roles (modelist, evaluator, peer-reviewer, decision-maker);
5. Reproducibility, or at least transparency; and
6. Expost check.

Different approaches respond differently to the above factors. CBA is well-codified worldwide and if not reproducible, any CBA is at least understandable from everybody in the world. For MCA, instead, the recent approaches are more and more complex in weighting criteria and ranking solutions and many different algorithms exist, making the result more method-sensitive than any other approach. This is an advantage for acceptability, but also a problem when complexity reduces transparency and clarity of results. I/O models are quite simple and their complexity lays in the availability of reliable and sufficiently detailed data. This fact, together with the other elements mentioned, make them loved by politicians. CGE and SD, instead, make of complexity their distinctive sign. This is their main limit, and must be managed in the way listed above: stability, comparability, and credibility of the authors.

Equity, Distribution, and Conflicts

Conflicts are common when decisions must be taken under constrained resources (public money, but also environmental goods, time, and space). The main source of conflict lays in the distribution of benefits and costs associated to a project, especially in the long-run and out of the direct effects measured in the transport market (e.g., environmental or economic externalities). Moreover, these effects are also pretty hard to be measured. If everybody would gain forever and at no cost, conflicts would not exist or would be limited. However, often, a project gives large benefits to the few and spread small costs to the many (in terms of taxation, for example). Sometimes, there are even large costs to some groups, for example, for local externalities (e.g., a new highway crossing an environmentally sensitive valley). The way assessment approaches deal with conflicts, distribution of effects and ultimately equity is profoundly different.

MCA, which is conceived to manage different viewpoints, is more suitable to take into account and bring into the decision the distributive issues implicit in stakeholders' positions. On the other side, the result of MCA depends strictly on who is considered a "stakeholder." For example, in many applications, taxpayers are not sitting at the table (and hardly the elected politician can be considered as their representative) while the presence of locally impacted groups is dominating.

CBA is, in principle, inadequate to manage different points of view and conflicting objectives, simply because it assumes an aggregated measure of worthwhileness. CBA is not aiming at "being" the decision balancing the criteria at stake, but just to inform the stakeholders about the unique criterion of socioeconomic efficiency (including, under its perspective, many elements that are considered as criteria in other approaches, such as time savings, costs, and environment). However, CBA can be easily adapted to make explicit the distributive and equity impacts. In fact, instead of giving as output the aggregated indicators only, NPV can be split among involved groups. The level of such detail depends on the richness of available data, but the effect on main groups of users, polluted, state, operators, infrastructure managers, can be always computed. Interestingly, NPV can also be split spatially, when a transport model is available. It is also important to remember—and this has an impact also on transparency—that CBA assumes no weight (or, better, a weight of 1) for all impacts, once translated into the monetary units.

Another key point of CBA is the Kaldor–Hicks criterion, which assumes that parties can always compensate gains and losses or—more realistically—that from a collective point of view 1 € of loss for actor A is as important as 1 € gain for actor B. The effects of this assumption can be somewhat ignored when winners are users/citizens and losers is the State (in form of investment, taxes, etc.). But this is not always the case and different situations potentially conflict can rise. For example, some users win (the public transport users, e.g.) and others lose (the road users). Or producers gain, thanks to market power, more than what is gained by users (think for example what happens to fares when HS trains are introduced in monopoly conditions).

Distribution of impacts is also a concern of EIA, but it is treated just qualitatively and nothing is said about the equity. Other models—being models—are at best tools to evaluate distribution, but do not assume any judgment on it.

Political Acceptance

An institutional use evaluation requires to manage the complexity of multiple-objective and multi-actors political decisions, but also to limit, or at least to make more transparent, the arbitrariness of decision-makers. In this sense, is quite natural that setting a structured evaluation framework is hardly the first goal of a politician (Eliasson et al., 2015). At the same time, institutionalization means also to keep evaluation independent from politicians, not leaving them the freedom to use it or not.

However, not all approaches "limit" decisions in the same way and with the same rigidity. CBA is probably the most constraining approach, assuming one inclusive but very structured criterion. At the opposite of the range, MCA is designed exactly to maintain and manage the contradictions of different points of view. Other approaches are more perceived as a *tool* rather than a *rule*, like transport models or the other more complex models mentioned above (LUTI, CGE, etc.).

In this sense it is interesting that CBA is, despite the limited love of politicians, is worldwide the most adopted assessment approach. Reasons can be various.

1. assessment could be introduced (or often just promoted) in a legislative system from outside, for example, by international bodies (EU, WB, EIB, etc.);
2. politicians leave, while civil servants remain: in countries with a strong public administration tradition, assessment is promoted by it and just accepted by passing politicians;
3. into an administration, the payer is sometimes different from the buyer. This is the case when budget is spent by the Ministry of Transport but allocated by the Ministry of Finance, or when regions receive earmarked budget from central state. Therefore, rules are set—indirectly or not—by treasury rather than sectorial/local decision-makers;
4. a central administration could decide to introduce CBA as a strong tool to justify decisions based on efficiency rather distribution.

Actually, in understanding political acceptance of CBA as a decision tool, it is important to look at what is the role and the moment when it is used. A CBA whose results are legally binding politicians' decisions will be much less accepted (and promoted) with respect to a CBA whose role is to inform decisions, improve the performance, improve the transparency toward stakeholders. Put it simply, if $NPV < 0$ means *automatically* the rejection of a project, whatever is the opinion of the decision-maker, CBA will be less accepted (and typically never introduced). Instead, if a project can start even with a $NPV < 0$, admitting the prevalence of other justified criteria (regional distribution, equity, political strategy, etc.), the politician will be more prone to accept its result as an input for decision. The use of a MCA after a CBA is a way to rationalize this approach, which in some contexts is simply left to political debate.

CBA and Friends: When Interaction is Possible

For sure, a cross-fertilization among assessment approaches and techniques is possible and often needed to overcome reciprocal limits. At the same time, one must not ignore the deep differences mentioned above, assuming a naïve perspective of integration.

CBA and Transport Models

The first step for a CBA is to have reliable, but also detailed demand estimations, which in the case of transport must come from a transport model. In principle, CBA is suitable also for a back-of-the-envelope assessments, but these simplified exercises should be used carefully.

Transport models are needed not only to estimate the demand for an infrastructure, but also to obtain the generalized costs associated to the demand. In other words, in the typical surplus calculation, both quantities and costs derive from a multimodal equilibrium and must be used coherently to avoid paradoxical results (for example, a too small benefit with respect to large demand shifts would simply give absurd CBA results).

This apparently obvious integration brings a lot of practical complexity and even problems (Beria et al., 2018).

1. Models produce a lot of data, whose management is not effortless;
2. CBA is useful to point out local problems of a model, usually invisible to the modelist also during calibration (debugging);
3. Normally, surplus is calculated as a difference between generalized costs. But the CG should in principle be exactly the one of the model, including modal constants, calibration coefficients, and functional form. Any deviation will give a CBA incoherent with demand estimation (but not necessarily “wrong”).

Concerning this last point, it is worth noticing that sometimes this is required by guidelines. For example, when Values of Time used in the CBA must be the standard ones and not the ones derived from the calibrated model. A “perfect” integration between model and CBA is possible only using the *logsum* method, of which the *Rule of Half* is the most used proxy.

CBA and MCA

A first mating is between CBA and MCA. The relationship, however, is far from obvious (Annema et al., 2015; Beria et al., 2012), despite the interest of the literature and of decision-makers. There are three ways to integrate the two approaches, and they are very different.

1. MCA is used to quickly screen available options and exclude the worst ones, because dominated or because clearly unfeasible (e.g., one option much more complex than all other options but with similar effects). Survived options undergo a normal CBA.
2. A normal CBA is performed, but some relevant aspects remain unconsidered. For example, two roads have comparable costs and performances, but cross different landscapes or have very different authorization procedures because in two different regions. In this case, the CBA maintains its role and MCA compares the residual criteria only for worthwhile options.
3. CBA is one criterion of a MCA. This commonly used approach has many drawbacks. CBA often includes most of the relevant aspects and consequently MCA introduces unacceptable double counts. Secondly, CBA assumes a constraint (the budget) and one criterion (welfare maximization), and this is in contrast with a multi-agent approach, where criteria are different and constraints are not relevant. In common applications, MCA is de facto used to blend the results of a CBA, adding further criteria and weights. In this case, better to separate the two approaches, as the mix would lose most of credibility requirements listed above.

CBA and CGE or LUTI

A second combination is much more straight and synergic (Eliasson and Fosgerau, 2019; Rothengatter, 2017; Venables, 2007). CBA’s hypothesis of invariance of non-transport markets is often reasonable, but not always, as some investments can modify the market conditions upstream and downstream. In these cases, partial equilibrium models could be not sufficient and the obvious way to overcome this limit without leaving the CBA perspective is to “add” those effects that are additional to the direct ones (caring to avoid risky double counts). These effects are usually called “wider economic effects” and the way to calculate them is to use a CGE model including the transport system and the other relevant markets potentially influenced (labor, manufacturing, land use, etc.). The literature on this topic is still open, and well-recalled in other parts of this book, but extremely interesting in showing the complexity of such issue and—by contrast—the power of simplicity of CBA, when its hypotheses can be accepted.

CBA and EIA

EIA is not an evaluation technique in the sense of the other ones of this chapter, as it is not aimed to judge an investment or create a ranking among alternatives. It is rather used to clarify the impacts (environmental, but also societal, archeological, etc.) of a project, showing which alternatives impact less and which compensations or modifications can be imposed to the project to minimize negative impacts. In this sense, EIA shares the quantification of impacts with CBA (that monetize them), but also sets the constraints

to the project. Constraints translate into extra investment costs (for changes and compensations) and extra benefits (reduced negative impacts). It is worth noticing that a standard—like an environmental threshold that must not be passed—is, for CBA, an infinite cost.

References

- Annema, J.A., Mouter, N., Razaei, J., 2015. Cost-benefit analysis (CBA), or multi-criteria decision-making (MCDM) or both: politicians' perspective in transport policy appraisal. *Transp. Res. Proc.* 10, 788–797.
- Beria, P., Bertolin, A., Grimaldi, R., 2018. Integration between transport models and cost-benefit analysis to support decision-making practices: two applications in northern Italy. *Adv. Oper. Res.* 2018, 2018.
- Beria, P., Maltese, I., Mariotti, I., 2012. Multicriteria versus cost benefit analysis: a comparative perspective in the assessment of sustainable mobility. *Eur. Transp. Res. Rev.* 4 (3), 137.
- Bröcker, J., 2014. Spatial computable general equilibrium analysis. In: Karlsson, C., Andersson, M., Norman, T. (Eds.), *Handbook of Research Methods and Applications in Economic Geography*. Edward Elgar Publishing Ltd., Cheltenham, UK, pp. 41–66.
- Bröcker, J., Korzhenevych, A., Schürmann, C., 2010. Assessing spatial equity and efficiency impacts of transport infrastructure projects. *Transp. Res. Part B: Methodol.* 44 (7), 795–811.
- Charness, A., Cooper, W.W., 1961. *Management Models and Industrial Applications of Linear Programming*. Wiley, New York.
- De Brucker, K., Macharis, C., Verbeke, A., 2011. Multi-criteria analysis in transport project evaluation: an institutional approach. *Eur. Transp.* 47 (47), 3–24.
- Eliasson, J., Börjesson, M., Odeck, J., Welde, M., 2015. Does benefit–cost efficiency influence transport investment decisions? *J. Transp. Econ. Policy* 49 (3), 377–396.
- Eliasson, J., Fosgerau, M., 2019. Cost-benefit analysis of transport improvements in the presence of spillovers, matching and an income tax. *Econ. Transp.* 18, 1–9.
- Macharis, C., Bernardini, A., 2015. Reviewing the use of multi-criteria decision analysis for the evaluation of transport projects: time for a multi-actor approach. *Transp. Policy* 37, 177–186.
- Pérez, J.C., Carrillo, M.H., Montoya-Torres, J.R., 2015. Multi-criteria approaches for urban passenger transport systems: a literature review. *Ann. Oper. Res.* 226 (1), 69–87.
- Rothengatter, W., 2017. Wider economic impacts of transport infrastructure investments: relevant or negligible? *Transp. Policy* 59, 124–133.
- Shepherd, S.P., 2014. A review of system dynamics models applied in transportation. *Transportmetrica B: Transp. Dyn.* 2 (2), 83–105.
- Venables, A.J., 2007. Evaluating urban transport improvements: cost–benefit analysis in the presence of agglomeration and income taxation. *J. Transp. Econ. Policy* 41 (2), 173–188.

Demand for Air Travel and Income Elasticity

Jing Lu*, Yucan Meng*, Changmin Jiang[†], Cheng Lv*, *Nanjing University of Aeronautics and Astronautics, Nanjing, China; [†]University of Manitoba, Winnipeg, Canada

© 2021 Elsevier Ltd. All rights reserved.

Glossary	547
Overview of Worldwide Air Travel Demand	547
How to Calculate Income Elasticity	548
The Aggregate Method of Income Elasticity Calculation	548
The Discrete Method of Income Elasticity Calculation	548
Case Study Using the Aggregate Method	549
Time Series Income Elasticity of Air Travel Demand	549
Spatial Income Elasticity of Air Travel Demand	552
Case Study Using Discrete Method	552
Model Estimation	552
Income Elasticity of Air Travel Demand	553
Conclusion	554
References	554
Further Reading	554

Glossary

Air Travel Demand The volume of passengers who travel by air, we consider the demand using commercial airlines in this article.

Income Elasticity of Demand The ratio of percentage change in demand to the percentage change in income.

Random Utility Theory A theory measures people's choice preference on goods or services by scoring and ranking the alternative using recognized utility.

Multinomial logit model One type of logit model which models the probability of choosing different alternatives.

Overview of Worldwide Air Travel Demand

In 1903, the Wright Brothers built the first powered aircraft. Since then, people have had a new mode of transport that is both high-speed and convenient. Currently, there are approximately 12 million passengers transported by nearly 8500 flights every day, and the demand is expected to continue growing over the next 20 years. Except for trips using general aviation, most air travel starts and ends with the approximately 290 commercial airlines registered in 120 countries.

Fig. 1 shows the real-time distribution of flights covering the earth at 17:00 on June 24, 2019, as mapped by VariFlight, and demonstrates a close relationship between economic level and flight distribution. Specifically, in the figure, the departure and arrival points of flights are concentrated in several regions: Europe, East Asia, the east and west coasts of the United States, and the Middle East (mainly in Saudi Arabia). Compared to other regions in the world, all the areas mentioned above are recognized to have better economic performances evaluated by indicators such as GDP, personal income, export–import volume, etc.

As tickets for air travel are more expensive than those for land-based or marine transportation, personal income is commonly regarded to have a great impact on the purchase of air tickets and further influence air travel demand (Brons et al., 2002). **Fig. 2** describes the distribution of worldwide personal income in 2017 (provided by World Bank). In the figure, the United States, West Europe, East Asia, and Saudi Arabia are the areas with high personal income of up to \$48,600 on average per year. We find that the areas with higher income in **Fig. 2** are those that attract more flights in **Fig. 1**; hence, it is reasonable to believe that personal income is positively correlated with air travel demand (Chin, 2002).

To further explore the relationship between income and air travel demand, we collected time series data from five countries: China, the United States, the Netherlands, the Philippines, and South Africa. All data are shown in **Fig. 3**.

We first compare the income and air travel demand levels in all the countries in **Fig. 3** in 2017. Personal income in the United States is 16 times higher than that in the Philippines (58,300 USD to 3660 USD), but the difference in air travel demand is more than 12 times (857,213,313 passengers to 66,703,257 passengers). Meanwhile, personal income in the United States is approximately 6.7 times higher than that in China; however, the air travel demand difference in these two countries is less significant (857,213,313 passengers to 551,234,500 passengers). Therefore, air travel demand does not appear to increase proportionately with the growth of personal income.



Figure 1 Real-time flight distribution. Source: <http://map.variflight.com/>

Subsequently, we examine the variation of personal income and air travel demand in China from 2013 to 2017. Personal income increased by 27% in these 5 years, while air travel demand increased by 52%. Meanwhile, air travel demand seems to be more stable in the Netherlands (high personal income), the Philippines, and South Africa (low personal income). Hence, it is reasonable to suspect that the relationship between air travel demand and personal income is S shaped. In particular, the demand will increase slowly when personal income stays at a low level and then grow quickly when personal income falls within the range of quick increase, and it will finally become stable as the income reaches a relatively high level.

How to Calculate Income Elasticity

In this section, we introduce two methods to calculate the quantitative effects of personal income on air travel demand. One is a classical method of measuring the income elasticity of demand, and the other is a logit-based method that can measure the elasticity of passengers' air travel choice probability to the change of personal income. The two methods are used to evaluate income elasticity from an aggregate perspective and a discrete perspective, respectively.

The Aggregate Method of Income Elasticity Calculation

The income elasticity of demand is a classic concept in the theory of microeconomics, and it is calculated as the ratio of the percentage change in demand to the percentage change in income (Marshall, 2009). The coefficient of income elasticity of demand E is calculated as Eq. (1), in which Q measures the volume of air travel demand, ΔQ is the change of Q , I is the personal income and ΔI donates the variation of I .

$E > 1$ indicates a high elasticity, in which case the increase in personal income will lead to a relatively large increment of demand; $E = 1$ indicates a unitary elasticity, in which case the demand will increase proportionately with the growth of personal income; $0 < E < 1$ indicates a low elasticity, in which case the increase percentage of demand will be less than the increase percentage of personal income; $E = 0$ indicates a zero elasticity, in which case the change of demand occurs regardless of income growth; $E < 0$ indicates a negative elasticity, in which case the demand will not increase but decrease with the growth of personal income. Normally, positive elasticity is associated with normal goods or services, negative elasticity always appears with inferior goods or services, and zero elasticity is associated with necessities.

$$E_m = \frac{\Delta Q/Q}{\Delta I/I} \quad (1)$$

The Discrete Method of Income Elasticity Calculation

In addition to the classical method that measures income elasticity using aggregated demand data, there exists another logit-based method using discrete data. The logit model is established based on the random utility theory. Eq. (2) is the utility function of

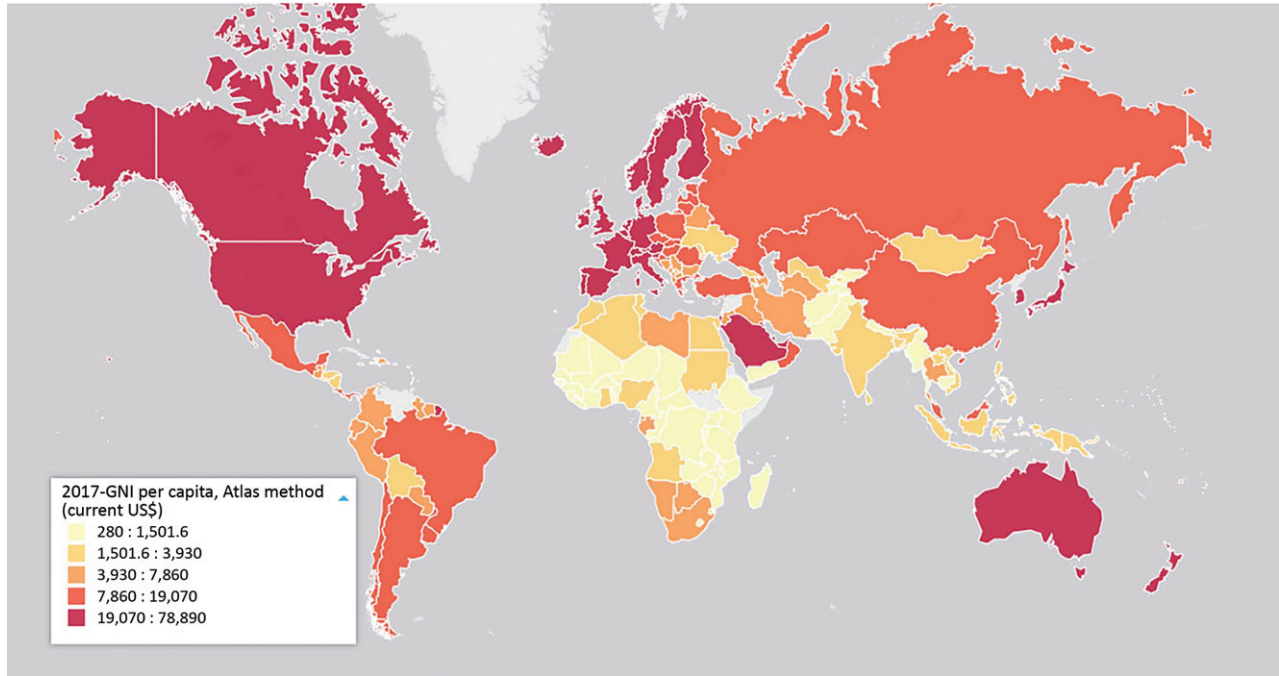


Figure 2 Distribution of personal income in the world. Source: <https://data.worldbank.org>

individual n from travel mode m , where V_{nm} is the observed portion while ϵ_{nm} stands for the random portion. In Eq.(3), V_{nm} is a function formulated by variable x_{nml} and the parameter β_{nml} , and personal income is a variable in the function of V_{nm} . ϵ_{nm} is an extreme value and is assumed to be independent of irrelevant distribution. Eq. (4) shows the specification of P_{nm} , which is the probability of individual n choosing mode m when ϵ_{nm} adopts a Gumbel distribution.

On the basis of Eq. (4), the income elasticity (E_{nm}) can be defined as the ratio between the percentage change in the propagability of choosing a specific mode (P_{nm}) and the percentage change in personal income I_n ($I_n x_{nml}$), which is demonstrated as Eq. (5). Due to the linear relationship between V_{nm} and I_n , Eq. (5) can be simplified as Eq. (6).

$$U_{nm} = V_{nm} + \epsilon_{nm} \quad (2)$$

$$V_{nm} = \sum_{l=1}^L \beta_{nml} x_{nml} \quad (3)$$

$$P_{nm} = \frac{e^{V_{nm}}}{\sum_m e^{V_{nm}}} \quad (4)$$

$$E_{mx_{nm}} = \frac{\partial P_{nm} x_{nm}}{\partial x_{nm} P_{nm}} = \frac{\partial V_{nm}}{\partial x_{nm}} P_{nm} (1 - P_{nm}) \frac{x_{nm}}{P_{nm}} = \frac{\partial V_{nm}}{\partial x_{nm}} x_{nm} (1 - P_{nm}) \quad (5)$$

$$E_{mx_{nm}} = \beta_x x_{nm} (1 - P_{nm}) \quad (6)$$

Case Study Using the Aggregate Method

Time Series Income Elasticity of Air Travel Demand

The aggregate method introduced in Section “The Aggregate Method of Income Elasticity Calculation” is applied here to measure the income elasticity of domestic air travel demand in China from 1999 to 2018, with the results being demonstrated in Table 1. The data on air travel demand and personal income are sourced from the China Statistical Yearbook, and the fourth column of the table demonstrates the income elasticity calculated with Eq. (1). In the calculation, the demand and income of the previous year are set as the baseline.

In examining the results, only one elasticity value is less than zero, indicating that the volume of domestic air travel demand in China decreased with the growth of personal income from 1998 to 1999. Referring to the Chinese air ticket pricing policy in the 1990s, it is easy to find that air tickets were fixed in 1999 as airlines suffered a great loss in prior years. Because of the increased air ticket prices, many passengers ceased traveling by air, leading to a demand decrease. Conversely, airlines were permitted to discount their tickets in 2000, hence the income elasticity calculated in this year is higher than 2. Therefore, it is reasonable to recognize air ticket price as an important factor that influences the income elasticity of air travel demand.

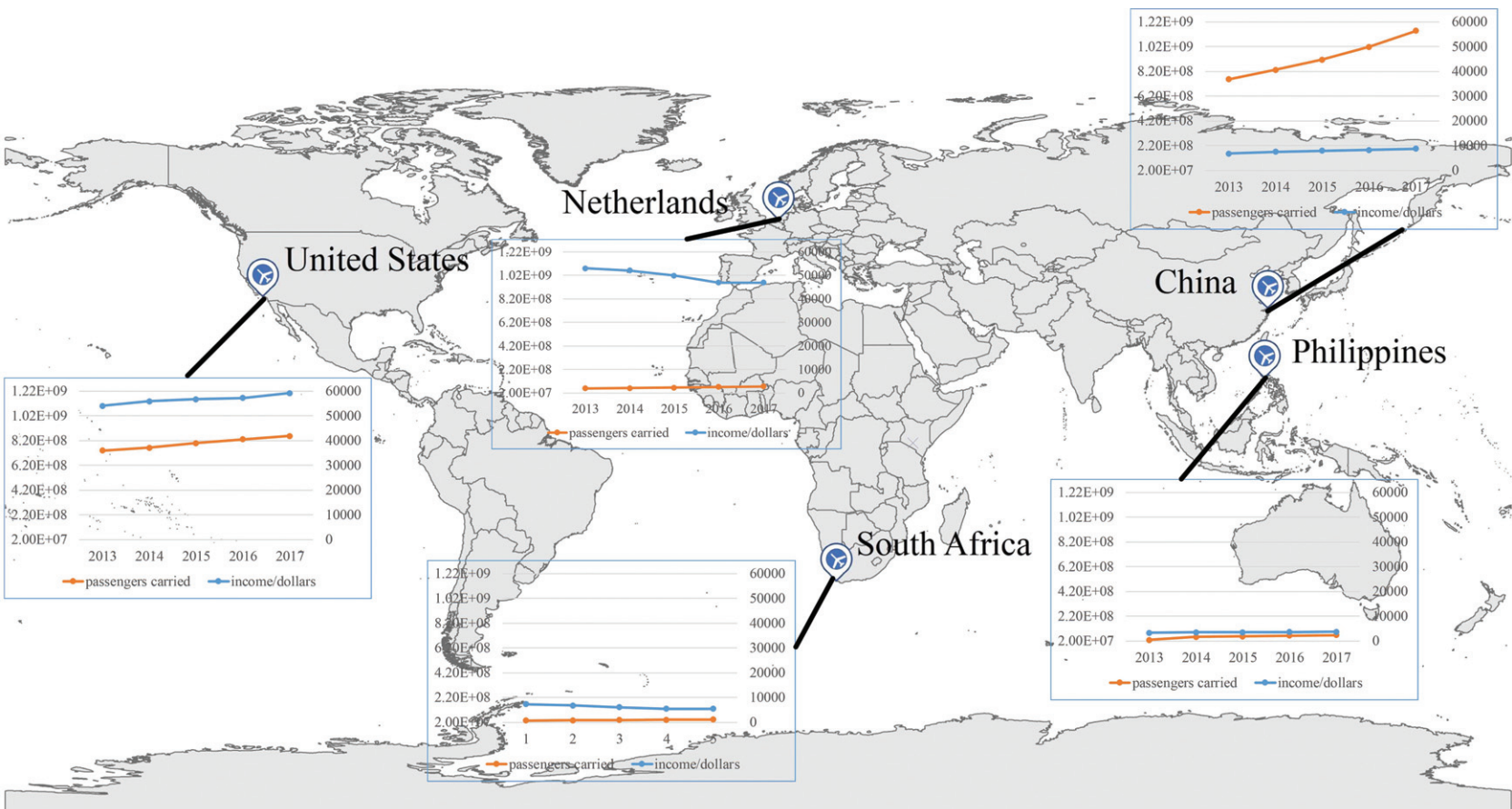


Figure 3 Relationship between personal income and air travel demand. Source: Federal Aviation Administration; Civil Aviation Administration of China; Civil Aviation Authority of the Philippines; South African Civil Aviation Authority; Eurostat

Table 1 Income elasticity of domestic air travel demand in China

Year	Air travel demand (1000 passengers)	Personal income (RMB)	Income elasticity of demand
1998	5205	5425	—
1999	5087	5854	−0.32
2000	6032	6280	2.31
2001	6832	6860	1.39
2002	7756	7703	1.09
2003	8078	8472	0.44
2004	11,046	9421	2.67
2005	12,602	10,493	1.21
2006	14,553	11,760	1.24
2007	16,884	13,786	0.94
2008	17,732	15,781	0.38
2009	21,578	17,175	2.20
2010	24,838	19,109	1.30
2011	27,199	21,810	0.70
2012	29,600	24,565	0.72
2013	32,742	26,955	1.08
2014	36,040	29,381	1.11
2015	39,411	31,790	1.13
2016	43,634	33,616	1.78
2017	49,611	36,396	1.58
2018	54,806.5	39,251	1.30

Source: China Statistical Yearbook

In addition, the average value of the calculated elasticity is 1.06 in [Table 1](#), suggesting that air transportation is a luxury service for passengers given the current income level in China. In addition, although Beijing hosted the Olympic Games in 2008, the income elasticity of domestic air travel demand was only 0.38 in that year, much lower than the average value.

Using the same method, we calculate the income elasticity of international air travel, and the results are shown in the fourth column of [Table 2](#). The averaged income elasticity in [Table 2](#) is 1.32, which is higher than domestic air travel in [Table 1](#). This is likely because international air travel is always much more expensive than domestic air travel ([Gallet and Doucouliagos, 2014](#)).

Table 2 Income elasticity of the international air travel demand in China

Year	Air travel demand (1000 passengers)	Personal income (RMB)	Income elasticity of demand
1998	526	5425	—
1999	631	5854	2.27
2000	690	6280	1.26
2001	693	6860	0.05
2002	838	7703	1.58
2003	682	8472	−2.52
2004	1077	9422	3.64
2005	1225	10,493	1.18
2006	1415	11,760	1.25
2007	1692	13,786	1.11
2008	1519	15,781	−0.90
2009	1474	17,175	−0.38
2010	1931	19,109	2.34
2011	2118	21,810	0.71
2012	2336	24,565	0.83
2013	2655	26,955	1.35
2014	3155	29,381	1.92
2015	4207	31,790	3.30
2016	5162	33,616	3.41
2017	5545	36,396	0.90
2018	6367	39,251	1.78

Source: China Statistical Yearbook

Table 3 Spatial income elasticity of air travel demand

Year	Beijing	Shanghai	Chengdu	Bangkok	Sydney	Dubai
2000	—	—	—	—	—	—
2010	1.10	1.19	1.24	1.42	3.49	6.83
2018	0.93	1.02	1.09	1.72	2.73	3.17

In the table, the income elasticities are less than zero in 2003, 2008, and 2009. The situation in 2003 was due to the outbreak of severe acute respiratory syndrome (SARS), which is an infectious disease with a high lethal rate. The subprime crisis that occurred in the United States caused the negative elasticities in 2008 and 2009. Due to the break of transnational collaboration and declined income, many business or leisure trips were cancelled during those 2 years. In addition, although the income elasticity in 2001 was positive, it was much lower than the average value in [Table 2](#). The 911 terrorist attacks may have led passengers to cancel their air travel for safety reasons.

As shown in [Table 2](#), it is obvious that the income elasticity values in 2015 and 2016 are very high. The situation is probably due to the Belt and Road initiative proposed by the Chinese government, which promotes communication and cooperation between China and other countries. Therefore, compared to the income elasticity of domestic air travel, international air travel is easily affected by policy, public emergencies and paroxysmal calamities.

Spatial Income Elasticity of Air Travel Demand

[Table 3](#) demonstrates the income elasticity of air travel demand for various international cities. In the calculation, the volume of air travel demand and the personal income in 2000 are set as the baseline. In the table, most income elasticity values for Beijing, Shanghai, and Chengdu are larger than 1, indicating that people in the three cities in China prefer the air travel mode.

However, in Bangkok, Sydney, and Dubai, the income elasticities of air travel demand are all greater than 1. For Bangkok, the Suvarnabhumi Airport opened in 2006, and the Don Mueang Airport opened its second runway in 2013, which may greatly stimulate air travel demand. In Sydney and Dubai, the reason for the high elasticity value is the rapid development of the tourism industry.

Case Study Using Discrete Method

In this section, the discrete method proposed in Section “The Discrete Method of Income Elasticity Calculation” is applied to calculate the income elasticity of air travel demand, and elasticity is here defined as the percentage change in the choice probability of air travel mode to the percentage change in personal income. In the calculation, the travel mode choice is first modeled based on a multinomial logit model; then, the data describing passengers’ choice behavior are collected using revealed preference (RP) and stated preference (SP) surveys. Finally, income elasticity is calculated on the basis of model estimation. Readers can find the fundamental theory of a multinomial logit model in the book written by [Train \(2009\)](#).

Considering substitutes for air travel, four alternatives are incorporated in the model: air, high-speed rail, normal rail, and nontravel. In the model, the passenger choice probability of each alternative is calculated by [Eq. \(4\)](#), in which the utility is formulated as in [Eq. \(2\)](#). To estimate the utility function, the RP and SP survey is designed to collect the data.

Model Estimation

The model was estimated by using the survey data; in the estimation, the survey data must be effect-coded, and the last level is set as the base level ([Cohen and Aiken, 2014](#)). Meanwhile, the full information maximum estimation method was employed for the estimation. The results are shown in [Table 4](#), and the utility function of each alternative is formulated as [Eq. \(7\)](#) to [Eq. \(10\)](#).

$$\begin{aligned}
 U(\text{Air}) = & 0.037 \cdot \text{INC_1} + 0.039 \cdot \text{INC_2} + 0.165 \cdot \text{INC_3} + 0.007 \cdot \text{PUR_1} + 0.060 \cdot \text{PUR_2} + 0.007 \cdot \text{PUR_3} - 0.003 \\
 & \cdot \text{PRICE} - 0.205 \cdot \text{TTIME} + 0.173 \cdot \text{CFEE_1} - 0.053 \cdot \text{CFEE_2} - 1.123 \cdot \text{CFEE_3} + 0.073 \cdot \text{DTIME_1} - 0.056 \\
 & \cdot \text{DTIME_2} - 0.125 \cdot \text{DTIME_3} - 0.130 \cdot \text{FTIME_1} + 0.131 \cdot \text{FTIME_2} + 0.069 \cdot \text{FTIME_3} - 0.012 \cdot \text{ATIME} + 0.033 \\
 & \cdot \text{ACTP_1} + 0.002 \cdot \text{CMFF_1} + 0.051 \cdot \text{CMFF_2} + 0.061 \cdot \text{CMFF_3} + \varepsilon_{\text{air}}
 \end{aligned} \quad (7)$$

$$\begin{aligned}
 U(\text{Normal train}) = & 0.037 \cdot \text{INC_1} + 0.039 \cdot \text{INC_2} + 0.165 \cdot \text{INC_3} + 0.007 \cdot \text{PUR_1} + 0.060 \cdot \text{PUR_2} + 0.007 \\
 & \cdot \text{PUR_3} - 0.003 \cdot \text{PRICE} - 0.205 \cdot \text{TTIME} + \varepsilon_{\text{normal}}
 \end{aligned} \quad (8)$$

$$\begin{aligned}
 U(\text{High-speed train}) = & 0.037 \cdot \text{INC_1} + 0.039 \cdot \text{INC_2} + 0.165 \cdot \text{INC_3} + 0.007 \cdot \text{PUR_1} + 0.060 \cdot \text{PUR_2} + 0.007 \\
 & \cdot \text{PUR_3} - 0.003 \cdot \text{PRICE} - 0.205 \cdot \text{TTIME} + \varepsilon_{\text{high-speed}}
 \end{aligned} \quad (9)$$

$$\begin{aligned}
 U(\text{Non-travel}) = & 0.037 \cdot \text{INC_1} + 0.039 \cdot \text{INC_2} + 0.165 \cdot \text{INC_3} + 0.007 \cdot \text{PUR_1} + 0.060 \cdot \text{PUR_2} + 0.007 \cdot \text{PUR_3} \\
 & + \varepsilon_{\text{non-travel}}
 \end{aligned} \quad (10)$$

Table 4 Results of the nested logit model

Variable	Coefficient	Variable	Coefficient
INC_1	0.0371 (0.04**)	DTIME_1	0.07268 (0.82)
INC_2	0.0387 (0.04**)	DTIME_2	−0.05564 (−0.64)
INC_3	0.1651 (0.04**)	DTIME_3	−0.12495 (−1.51)
PUR_1	0.0074 (0.04*)	FTIME_1	−0.12986 (−1.51***)
PUR_2	0.0604 (0.04*)	FTIME_2	0.13064 (1.52***)
PUR_3	0.0077 (0.04*)	FTIME_3	0.06886 (0.78***)
PRICE	−.00265 (−26.35***)	ATIME	−0.01237 (−0.23**)
TTIME	−0.20524 (−19.08***)	CATP_1	0.03341 (0.63)
CFEE_1	0.17259 (1.74*)	CMFF_1	0.00168 (0.02*)
CFEE_2	−0.05344 (−0.63*)	CMFF_2	0.05116 (0.61*)
CFEE_3	−0.12329 (−2.03**)	CMFF_3	0.06058 (0.12*)
Log Likelihood Function: −1943.28			
Pseudo R^2 (ρ^2): 0.45			

***Significant at the 1% level

**Significant at the 5% level

*Significant at the 10% level

The positive signs of the income coefficients (INC_1, INC_2, and INC_3) prove that passengers with a higher income plan more travel and accept higher travel costs. INC_3 has the highest coefficient, which means that the utility is higher for passengers with an income higher than 100,000 RMB per year.

In addition, air travel mode choice is also influenced by other variables. For example, the negative signs of PRICE and TTIME indicate that utility will decline with increasing ticket price and travel time. The coefficient of ATIME is statistically significant, and its sign shows that an increase in access time leads to a decreasing utility. The table further shows that utility is higher if the flight departs on time and then decreases with increasing flight delays, and the amplitude of decrease is related to the length of the delay. Similarly, the utility will rise when an air ticket can be changed for free (CFEE_1) but will decline if the change fee is higher than 0 (CFEE_2, CFEE_3).

In the group of flight times, FTIME_1 leads to a negative effect on utility, which shows the phenomenon that passengers do not prefer flights departing before 7:00 in the morning. The sign of the coefficient for ACTP_1 proves that passengers prefer to choose large aircrafts. In addition, the class of membership (CMFF) will lead to a positive effect: the higher the class, the more utility will be added.

Income Elasticity of Air Travel Demand

Based on the estimated utility function, the income elasticity can be calculated by Eq. (6). In the calculation, we assume that the values of other variables are fixed, and the values are demonstrated in Table 5.

Table 5 Assumed value of the variables describing the different travel modes

Modes	PRICE (yuan)	TTIME (hour)	ATIME (hour)	DTIME (hour)	FTIME	CFEE (yuan)	CMFF	CATP
Air	600	1.5	1.5	1	10:00	300	Regular	Boeing
High-speed train	400	4	1.5	0	9:00	50	—	—
Normal train	200	10	1.5	0	21:00	0	—	—

Table 6 Income elasticity of travel demand (income = 30,000 RMB per year)

	<i>Air</i>	<i>High-speed train</i>	<i>Normal train</i>
Elasticity	0.032	0.033	0.035

Table 7 Income elasticity of travel demand (income = 110,000 RMB per year)

	<i>Air</i>	<i>High-speed train</i>	<i>Normal train</i>
Elasticity	0.145	0.149	0.156

To calculate the income elasticity, we first assume that the income of passengers who travel for business is at least 30,000 RMB per year to calculate the income elasticity of air travel demand, which is shown in column 2 of **Table 6**. Then, assuming that the income of passengers who travel for business is at least 110,000 RMB per year, the income elasticity of air travel demand is calculated. The results are shown in column 2 of **Table 7**. In addition, we also obtained the income elasticity of travel demand for high-speed trains and normal trains.

As per the results in the above two tables, the income elasticity of travel demand will increase with income growth, indicating that people are willing to spend more money on travel when earning a higher salary (Jang et al., 2004). In addition, comparing the income elasticity for different travel modes, it can be found that the value of air travel is the lowest because air travel requires more expensive tickets, and passengers may switch to other modes when the air ticket prices increase (Yang and Zhang, 2012).

Conclusion

This article introduces the relationship between personal income and air travel demand and then proposes two methods of calculating the income elasticity of air travel demand from both the aggregate and the discrete perspectives. The aggregate method calculated the elasticity using the classical function provided by microeconomic theory, and the discrete method measured the elasticity based on the logit model constructed by the random utility theory. The results show the positive relationship between personal income and air travel demand and indicate that the economic background, government policy, public emergency, and paroxysmal calamity may all have significant influences on the income elasticity of air travel demand. Moreover, because of the expensive air tickets, the income elasticity of air travel demand is lower than those of other travel modes.

References

- Brons, M., Pels, E., Nijkamp, P., Rietveld, P., 2002. Price elasticities of demand for passenger air travel: a meta-analysis. *J. Air Transp. Manag.* 8 (3), 165–175.
- Chin, A.T., 2002. Impact of frequent flyer programs on the demand for air travel. *J. Air Transp.* 7 (2), 53–86.
- Cohen, P., West, S.G., Aiken, L.S., 2014. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Psychology Press.
- Gallet, C.A., Doucouliagos, H., 2014. The income elasticity of air travel: a meta-analysis. *Ann. Tour. Res.* 49, 141–155.
- Jang, S.S., Bai, B., Hong, G.S., O'Leary, J.T., 2004. Understanding travel expenditure patterns: a study of Japanese pleasure travelers to the United States by income level. *Tour. Manag.* 25 (3), 331–341.
- Marshall, A., 2009. *Principles of Economics: Unabridged eighth edition*. Cosimo, Inc.
- Train, K.E., 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press.
- Yang, H., Zhang, A., 2012. Effects of high-speed rail and air transport competition on prices, profits and welfare. *Transport. Res. B Meth.* 46 (10), 1322–1333.

Further Reading

- Jørgensen, F., Preston, J., 2009. The relationship between fare elasticity and trip length—some comments. *Int. J. Transp. Econ./Rivista internazionale di economia dei trasporti* 361–375.

Generalized Cost for Transport

Jeppe Rich, Technical University of Denmark, Lyngby, Denmark

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	555
Introduction and Understanding	555
On the Justification of Generalized Cost	556
Generalized Cost—What Can be the Problem?	557
Generalized Cost or Generalized Time?	557
Conclusion and Summary	558
Acknowledgment	559
Biography	559
References	559

Nomenclature

GC: Generalized costs

GT: Generalized time

VoT: Value-of-time

Introduction and Understanding

In transport economics and transport modeling the term “generalized cost” (GC) typically refers to a sum of monetary and non-monetary costs of a journey weighted to a common base. Depending on the definition of the different weights, it can be expressed in either monetary units, time units, or any other unit defined by the modeler. In this sense, the concept of GCs provides a mean of summing all the significant and quantifiable elements of disutility arising from a transport journey or any other transport activity involving a measure of cost. It arose because it was found convenient, rather than as part of any theory. It is usually attributed to McIntosh and Quarmby (1970).

In the simplest form, it may be expressed as:

$$GC = c + f(q) \quad (1)$$

where GC refers to the generalized cost, c to one or more variables expressed in monetary units, q to a non-monetary variable, and f to a function that maps q into monetary units.

In practice, in transport economics and when applied in transport models, a more specific form is applied as shown in Eq. (2).

$$GC_{i,j,m} = c_{i,j,m} + w_m t_{i,j,m} \quad (2)$$

In this function, i and j represent origins and destinations while m represents a particular type of transport, which could be route or choice of mode. A common generalization is when $w_m t_{i,j,m}$ is expressed as a generalized journey time $GT_{i,j,m}$. That is:

$$GC_{i,j,m} = c_{i,j,m} + VoT \times GT_{i,j,m} \quad (3)$$

where $GT_{i,j,m} = w_{m,k=1} t_{i,j,m,k=1} + \dots + w_{m,k=K} t_{i,j,m,k=K}$

Here, $w_{m,1} \dots w_{m,k}$ represents different weights for different time components. To illustrate different $GT_{i,j,m}$ expressions, the following three common examples are provided.

1. For congested networks, congestion travel time is more annoying than “freeflow travel time.” A typical formulation often applied in road route choice models is given by:

$$GT_{i,j,m} = w_f t_{\text{freeflow}} + w_c t_{\text{congestion}} + w_t t_{\text{turntime}} \quad (4)$$

Here, t_{freeflow} represents freeflow time, $t_{\text{congestion}}$ congestion time, and t_{turntime} the time incurred when turning in intersections. All of the parameters w_f , w_c , and w_t measure different degrees of annoyance in the specific choice situation. As stated, it is generally found that $w_c > w_f$ reflecting that congestion is considered more annoying than freeflow time.

2. For public transport, walking time, waiting time, and interchange time are often considered more “painful” than onboard time. A commonly used form in public assignment models is:

$$GT_{i,j,m} = w_a t_{\text{walk}} + w_o t_{\text{onboard}} + w_s N_{\text{change}} + w_w t_{\text{wait}} \quad (5)$$

Here, t_{walk} represents walk time from the origin to the station, t_{onboard} the onboard time (often referred to as in-vehicle-time), N_{change} the number of changes, and t_{wait} the time incurred while waiting at stations in between trips.

3. For freight transport, where goods are transported in long and often complicated transport chains, it can be relevant to distinguish between transport time and lateness.

$$GT_{i,j,m} = w_t t_{\text{transport}} + w_l t_{\text{lateness}} \quad (6)$$

Here, $t_{\text{transport}}$ is the journey transport time and t_{lateness} expected lateness, possibly measured as some percentile of arrival time.

Note here that sometimes variables do not measure time but something else. N_{change} is one example of a variable not measured in time units, yet it forms part of the generalized time (GT). It is just that we have grouped all non-cost elements together and valued them in time units. Transport planners are often accused of being fixated on time savings, but in this case we are merely using a particular type of time as a base to allow us to treat everything on a fair basis, with each item having its own valuation relative to a common base. Other examples of variables that cannot be expressed as either time or money include, but are not limited to, the difficulty of boarding and alighting, the variability of journey time, and the ability to work while traveling.

Even more general expressions of GCs are common in the appraisal literature when studying social cost–benefits across a range of variables that express different external effects. Typically, this is based on a combined social surplus function that includes, apart from transport cost and transport time components, variables related to environmental costs and safety costs. Where it is not possible to establish a linear weighting function for an effect, a lump-sum valuation will need to be used for each “level” of that effect. For example, in considering the building of a new road, is the value of the loss of two trees exactly twice that of the loss of one tree? And, if two trees are to be lost, is that valued equally in a place where there are only two trees to begin with as in a place where there are 100? By weighting the different components with external cost estimates (typically provided by authorities) it is possible to calculate a social surplus function for a given policy. This measure of “consumer surplus” can then be benchmarked with respect to a neutral baseline and evaluated according to the cost it may incur.

For appraisal, the unit of measurement is always monetary units. Depending on the study, there may be many different types of external costs that vary with choice of mode and possibly also where and when the transport takes place. Air pollution and traffic noise as an example is less of a problem in rural areas and during nighttime compared to urban areas during daytime. Safety, on the other hand, will often be measured at the link level in order to account for differences in risk for different types of roads. At the general level this could be expressed as a GC that embraces all of these different components. At the level of $\{i, j, m\}$ it may be stated as a linear weighted function $\widetilde{GC}_{i,j,m}$ as presented in Eq. (7).

$$\widetilde{GC}_{i,j,m} = GC_{i,j,m} + a_s X_{i,j,m}^{\text{safety}} + a_{co} X_{i,j,m}^{\text{CO}_2} + a_n X_{i,j,m}^{\text{Noise}} + a_e X_{i,j,m}^{\text{Emissions}} \quad (7)$$

In this equation, the expression of generalized transport cost from before enters the model in monetary units as $GC_{i,j,m}$, while effects related to safety, CO_2 , noise, and emissions are all represented by separate linear models.

On the Justification of Generalized Cost

For appraisal studies it is common to assume fixed marginal external costs such that these are not influenced by the specific study or the specific model. Why? Because it implicitly imposes political standards for how infrastructure should be measured and benchmarked. If different weights across different areas and individuals were allowed, it would imply that two areas with the same challenges regarding congestion and traffic would be measured differently. As an example, if the value-of-time were allowed to vary by income, infrastructure investments would tend to favor rich neighborhoods all other things equal.

In the remainder of this paper, we will focus on the justification of GC from the perspective of a transport model. In this perspective, GC should be seen, mainly as a simplification of the underlying econometric model specification, which, one way or the other, will have consequences for the way we capture preferences and behavior. Insight can be offered by considering a simple discrete choice model with utility functions $V_{n,m}$ for agent n and mode m .

$$V_{n,m} = k_m + \alpha c_{n,m} + \beta t_{n,m} \quad (8)$$

This model can be estimated from micro-data or meso-data (if n represents a meso-data classification) and such estimation will typically render maximum likelihood estimates of alternative specific constants k_m and cost and time parameters α and β . The trade-off between time and money (the value-of-time) can be formulated as changes in time relative to changes in money:

$$VoT = \frac{\partial V / \partial t}{\partial V / \partial c} = \frac{\text{utility/time unit}}{\text{utility/monetary unit}} = \frac{\text{monetary unit}}{\text{time unit}} \quad (9)$$

In the case of a simple linear utility as presented in Eq. (8) the value-of-time function simply becomes ratio $\frac{\beta}{\alpha}$.

It can happen that $c_{n,m}$ and $t_{n,m}$ are correlated up to the point where it is difficult to estimate the parameters, and thereby their ratio with a sufficient precision. In econometrics this is referred to as the “parameter identification problem,” which is a serious and not easily solved problem. It is best to avoid it by either (1) careful model specification, (2) careful selection of estimation sample, or (3) by combining revealed and stated preference surveys where, in the latter, the levels of correlation can be better controlled when design the survey.

If the problem cannot be avoided, one can attempt to reduce the model to a form that can be properly identified. Removing one or more of the variables may obviously accomplish this goal, but introducing another challenge in that any policy analysis linked to the removed variables cannot be carried out. It is obviously a problem for most transport models if it cannot address the effect of either cost- or time-related variables as it may undermine the investigation of new infrastructure. A better approach is to fix one of the parameters and subsequently estimate a common scale parameter of either GCs or GT.

Generalized Cost—What Can be the Problem?

To understand what may be the problem of using a GC, consider a simple generalized additive cost function $GC_{n,m}$ (expressed in monetary units) that enters a utility function $V_{n,m}$. That is,

$$V_{n,m} = k_m + \alpha(c_{n,m} + \omega t_{n,m}) = k_m + \alpha GC_{n,m} \quad (10)$$

If the utility function is used in a corresponding discrete choice model, by assuming that individuals maximize utility over the choices m and by further assuming an error distribution that resembles IID Gumbel distributed errors, a choice probability model of the multinomial logit type emerges. That is,

$$P_{n,m} = \frac{\exp(V_{n,m})}{\sum_m \exp(V_{n,m})}, \forall n, m \quad (11)$$

Hereby, it is possible to consider how choice elasticities with respect to the GC depend on the value-of-time. Let $E_{GC_{n,m}}$ define the elasticity of GC, $E_{t_{n,m}}$ the travel time elasticity, and $E_{c_{n,m}}$ the cost elasticity. It is then easy to show, based on the definition of elasticities for the logit model, the two following identities: $E_{GC} = E_{t_{n,m}} + E_{c_{n,m}}$ and $\frac{\omega t_{n,m}}{c_{n,m}} = \frac{E_{t_{n,m}}}{E_{c_{n,m}}}$, where ω is the value-of-time. By combining these it is straightforward to see that:

$$E_{c_{n,m}} = \frac{E_{GC_{n,m}}}{1 + \omega \frac{t_{n,m}}{c_{n,m}}} \quad (12)$$

$$E_{t_{n,m}} = \frac{E_{GC_{n,m}}}{1 + \frac{c_{n,m}}{\omega t_{n,m}}} \quad (13)$$

This exposes some of the weaknesses of using a generalized travel cost approach, namely that (1) the elasticity of GC is a linear combination of the elasticity of cost $E_{c_{n,m}}$ and the elasticity of time $E_{t_{n,m}}$, and (2) the balance between the two is strictly controlled by the value-of-time. If the value-of-time is low, time will be considered less important compared to monetary costs, whereas the opposite is true if the value-of-time is high.

If a generic value-of-time is used for all modes, it may mean that for certain expensive modes, the balance between cost and time will become skewed and not reflect the true balance. This is acknowledged to be a problem when analyzing aviation and also high-speed rail.

So, in short, the main problem of using GCs is that the imposed weights often refer to other contexts and other time periods and thereby introduce a biased assessment of the different variables (Wardman and Toner, 2018). The implication may be that the balance between model sensitivity with respect to cost and time can be biased. This can lead to situations where a given assessment of travel time benefits is overestimated on the expense of the assessment of monetary cost that is then underestimated. References also include Kono et al. (2017) who noted that GC when applied as a mean to estimate trip demand may incur bias caused by the endogenous nature of the value-of-time. While this is certainly possible it does not provide much guidance on how then to forecast trip demand.

Generalized Cost or Generalized Time?

The term “generalized transport cost” suggests that the common generalization is expressed in monetary units. While this is certainly true when doing cost–benefit appraisal analysis, it is less common when applied in transport models. Consider the two expressions of indirect utility as either GT or GC:

$$V_{n,m}^{GT} = k_m + \beta(c_{n,m}/\omega + t_{n,m}) = k_m + \beta GT_{n,m} \quad (14)$$

$$V_{n,m}^{GC} = k_m + \alpha(c_{n,m} + \omega t_{n,m}) = k_m + \alpha GC_{n,m} \quad (15)$$

The question at this point is if it makes a difference whether we choose a formulation based on *GT* or *GC*? The answer is “it depends.” If solely judged through the glasses of an econometrician who seeks an identifiable model, there is no difference as $\frac{c_{n,m}}{\omega} + t_{n,m} \approx c_{n,m} + \omega t_{n,m}$. In other words, when estimating the model there is no difference and it is easy to assure that the sensitivity with respect to cost and time will be unaffected by the choice of model when evaluated on the estimation data. However, when considered from the perspective of applying the models it may have an effect. This is relevant when applying the model for forecasting purposes.

When applying transport models for forecasting, it is common to estimate a given parameter set and then subsequently apply this set of parameter for scenarios but with changes in variables that reflect the scenario. It is also common to assume that the value-of-time can change over time and that this change can be reasonably approximated by linking it to income growth. As productivity in society increases so should the value-of-time. In practice, this is typically expressed as an “income to value-of-time” elasticity. However, recent evidence (Rich and Vandet, 2019) suggests that, although there is strong empirical evidence for a positive income to value-of-time elasticity, changes in the value-of-time over time is affected by a variety of factors which, in addition to income growth, include increasing travel length and increasing congestion.

Consider the case of $V_{n,m}^{GT}$ from Eq. (14), where all we know about the future is a value-of-time ω that reflects that people are becoming wealthier relative to a general price index. As people become wealthier, costs tend to become less important to us. So in Eq. (14) we can think of ω as a “cost-deflation” factor, which when increased will put less emphasis on costs and in general cause the *GT* expression to decrease. As $\omega > 0$ and $\beta < 0$ this will lead to an increasing demand for longer trips as we maintain β throughout the forecast and keep all other things equal.

When used in the *GC* context of $V_{n,m}^{GC}$ in Eq. (15), the value-of-time ω represents what could be referred to as comfort affects, or rather the effect of discomfort. If ω increases while keeping α constant we can interpret this as increasing discomfort for longer trips, which in turn then would lead to increasing *GC* and broadly speaking a decreasing demand. Hence, in this forecast situation where ω is the only thing we know, the effect of using either of the two expressions is inversely proportional when measured through demand. In other words, whereas the model based on *GT* will tend to increase the distance of trips when the value-of-time is increasing, the opposite will be true for a model based on *GC*. The issue was first debated in Gunn (1983) who also acknowledged the difficulty of dealing with the issue.

As discussed earlier it is generally difficult to project the value-of-time in the future as it may depend on a variety of different factors from the development of new technology to the level of income. So even though it is clear that the value-of-time may change due to either changes in cost preferences or changes in time preferences or a mixture of the two, it is often not strictly possible to judge from which specific source a change occurs in the future. What the authors hereby want to convey to the readers is that modelers, when deciding on applying either *GC* or *GT*, in ways that are expressed earlier, implicitly impose an assumption on how the balance between cost and time sensitivity is projected.

At this stage it is relevant to give attention to the microeconomic framework proposed by Train and McFadden in their seminal paper from 1978 (Train and McFadden, 1978). In this paper, by forming a simple random utility framework with budget and time constraints, it is expressed how wages w are related to the utility of traveling and more specifically the attributes related to travel time and travel cost. It turns out that the utility specification in this simple case can be expressed as:

$$V_{n,m} = k_m \alpha (w^{-\theta} c_{n,m} + w^{1-\theta} t_{n,m}) \quad (16)$$

In the model θ can take on values between zero and one while α represents a common scale parameter.

The interesting observation here is that, if wages and the value-of-time can be assumed proportional, it follows that, depending on the value of θ , *GC* and *GT* represent two out of an infinite number of affine combinations of cost and time that form the indirect utility function. The *GT* arises as a special case when $\theta = 1$, whereas $\theta = 0$ resembles the case of travel cost. Why is this relevant? It is relevant because this framework, although presenting a nonlinear estimation challenge, provides a practical way of projecting the balance between cost and time in future scenarios that are different from the extreme cases presented earlier. The early findings of Train and McFadden suggested that $\theta = 0.7$ and therefore support the *GT* specification to a greater extent than the *GC* formulation. This suggests that time preferences are more stable over time which has been supported in other more recent studies (Fox et al., 2014; Rich and Vandet, 2019).

Conclusion and Summary

Generalized transport cost is a mean to combine different elements of travel (dis)utility. It is used in a number of situations in passenger and freight transport modeling, such as in route-choice models, public transport assignment models, and transport demand models in general (where time and cost components are accumulated over many different attributes). *GC* is also widely applied in appraisal studies of infrastructure and in other scenarios in that it forms the basis for a common monetary evaluation of a given policy.

The concept of *GC* may be expressed in any unit of measurement and is not restricted to monetary units. In transport models it is common to express the *GC* in either monetary units or in time units. In the paper, it is underlined that when models are applied to forecasting, choosing one form for the other may have consequences for the results. This can happen because, as the value-of-time increases, it will either deflate cost or increase travel time annoyance depending on the formulation. It is also illustrated, by a reference to Train and McFadden, how *GC* and *GT* may be combined and how the balance between them could be projected.

Acknowledgment

The author would like to acknowledge Anthony Fowkes and Andrew Daly for valuable contributions during the preparation of this paper.

Biography



Jeppe Rich is a professor at the Technical University of Denmark in the area of transport demand modeling. His interests include model specification, model estimation, population synthesis, space–time predictions, model forecasting, and the tracking of preferences over time. He has experience from several national and international model projects.

References

- Fox, J., Daly, A., Hess, S., Miller, E., 2014. Temporal transferability of models of mode–destination choice for the Greater Toronto and Hamilton Area. *J. Transp. Land Use* 7, 41–62, doi:10.5198/jtlu.v7i2.701.
- Gunn, H.F., 1983. Generalised cost or generalised time? A note on the forecasting implications. *J. Transp. Econ. Policy* 17 (1), 91–94. http://www.bath.ac.uk/e-journals/jtep/pdf/Volume_XV11_No_1_91-94.pdf.
- Kono, T., Kishi, A., Seita, E., Yokoi, T., 2017. Limitations of using generalized transport costs to estimate changes in trip demand: a bias caused by the endogenous value of time. *Transportmetrica A Transp. Sci.* 14 (3), 192–209, doi:10.1080/23249935.2017.1363316.
- McIntosh, P.T., Quarmby, D.A., 1970. Generalised Costs and the estimation of movement costs and benefits in transport planning. M.A.U. Note 179. Department of the Environment, London.
- Rich, J., Vandet, C.A., 2019. Is the value of travel time savings increasing? Analysis throughout a financial crisis. *Transp. Res. Part A Policy Pract.* 124, 145–168, doi:10.1016/j.tra.2019.03.012.
- Train, K., McFadden, D., 1978. The goods/leisure tradeoff and disaggregate work trip mode choice models. *Transp. Res.* 12 (5), 349–353.
- Wardman, M., Toner, J., 2018. Is generalised cost justified in travel demand analysis? *Transportation* 1–34, doi:10.1007/s11116-017-9850-7.

The Impact of Electric Vehicles on Energy Systems

Patrick E.P. Jochem*, **Jake Whitehead†**, **Elisabeth Dütschke‡**, ^{*}Institute for Industrial Production (IIP), Chair of Energy Economics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany; [†]School of Civil Engineering, The University of Queensland, St Lucia, QLD, Australia; [‡]Fraunhofer Institute for Systems and Innovation Research ISI, Karlsruhe, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	560
Hazards of PEV Charging for the Energy System	561
Synergies of EV Charging and the Electricity System	561
Increasing the Degree of Capacity Utilization of Grid Infrastructure	561
Integrating (Local) Provision of Electricity From Fluctuating Renewable Energy Generation	562
Decreasing Electricity Prices and System Costs While Increasing the Efficiency of Conventional Power Plants	562
User Acceptance of Load Flexibility	562
Discussion and Conclusions	563
Biographies	564
References	564
Further Reading	565

Introduction

Given the negative impact global energy systems have had on our planet over the course of the past century, a significant turnaround is now required if the still steeply increasing energy demand of the coming decades should not result in pushing our global environment further beyond its sustainable limits. As part of this required turnaround there has been a particular focus on the reduction of greenhouse gas emissions. Globally, these emissions are still increasing, even though this trend needs to be urgently reverted in order to minimize the social, environmental, and economic impacts of anthropogenic climate change. While the industrial sector—especially in developed countries—is already contributing to this target by reducing associated emissions, other sectors, such as the built environment, agriculture, and transport, are still increasing their emissions—which is in contradiction to the overall required emissions reduction target.

In regards specifically to the transport sector, this issue has only recently appeared on the agenda of politicians around the world (Creuzig et al., 2015), as it requires significant changes for the general population, as well as for established industries. Besides many other options, electric vehicles [in the following we focus on Plug-in Electric Vehicles (PEV), i.e., Battery Electric Vehicles and Plug-in Hybrid Electric Vehicles] stand out as an attractive option, as they do not require substantial and immediate changes in mobility patterns. PEVs are more energy efficient than fossil fuel vehicles and provide significant flexibility in fuel source in terms of electricity generation. Furthermore, if the ongoing transition in electricity generation towards renewable sources continues, over time it will lead to an “autonomous” and continuing decrease in the carbon-intensity of road transport through the adoption of PEVs.

Today, the integration of PEVs into the energy system appears to be an attractive solution for assisting in meeting global emission reduction targets (Jochem et al., 2015). PEVs—similar to conventional fossil fuel vehicles—are generally parked for long periods of time, providing the opportunity to provide energy services when they are not required to fulfil mobility demand. Additionally, consumers generally expect that PEVs should be capable of driving longer distances, despite the average daily distance of consumers being less than 50 km. This significant difference between stored and required energy, on average, presents a unique opportunity for this “excess” energy capacity to be used for purposes other than simply meeting mobility demand (Babrowski et al., 2014).

While PEVs present a number of opportunities for supporting electricity grids, conversely, if several vehicles are charged in parallel, at higher rates, and during higher demand periods on the electricity grid, these additional energy loads could place additional and significant stress on energy systems (Jochem et al., 2018).

Conversely, if the integration of PEVs into the energy system is in accordance with the principles mentioned below, multiple dividends may occur: PEVs can reduce greenhouse gas emissions from the transport sector while facilitating the integration of fluctuating renewable into the grid by providing load flexibilities across the wider system.

The main requirement of assuring a synergetic integration of PEV into energy systems is the charging strategy, which can be classified as follows:

1. **Uncontrolled charging:** As soon as the car is plugged-in the charging process starts at the maximum charging rate until the battery is fully recharged (i.e., State of Charge (SoC) equals 100%).
2. **Controlled charging:** The charging process does not necessarily start right after plugging-in, but after a signal is received (from the utility or user) for recharging the battery with flexible charging rates until a certain target SoC (e.g., 80%) within a defined time horizon (e.g., until scheduled car departure the next day).

3. **Bidirectional charging or vehicle-to-grid (V2G):** This charging process is similar to controlled charging, but provides another degree of freedom for load scheduling as the associated charging hardware provides the possibility to export electricity from the vehicle's battery back to the grid, when required. This mechanism significantly increases the flexibility of PEVs energy storage capabilities, but is associated with additional hardware costs and may also be associated with higher battery degradation.

Hazards of PEV Charging for the Energy System

One obvious hazard of PEVs to the energy system is the parallel charging of several PEVs at high charging rates. Even though other electrical appliances in private households create loads of several kilowatts, their stochastic application generally leads to a broad distribution over time, that is, the average load per household in distribution grids with dozens of households, levels out to approximately 1 kW, or in case of air conditioning, up to 3 kW. Exceptions to this could be during a heat wave in warmer climates, where electricity grids are already experiencing stress today.

Hence, if a large number of households plug in their vehicles at the same time in the evening, without any form of controlled charging, this could further exacerbate existing evening peak electricity demand. With many PEVs charging for one hour at 10 kW, the simultaneousness of this electricity demand is likely to be far higher than other appliances combined. This might lead to load peaks, which have not yet been seen in electricity grids. Even at a national level, peak loads may increase, which in turn could lead to additional required investments in the energy system, including grid expansion and additional generation. If this additional generation is not sufficiently provided by renewables, there is an additional risk that these higher loads would be met using carbon intensive energy generation, deteriorating the overall emissions reductions induced through the adoption of PEVs (Jochem et al., 2015). Alternatively, if the uptake of home battery systems continues, this additional localized energy storage may go some way to alleviating some of the pressure placed on the grid due to uncontrolled charging, noting that a home battery of 14 kWh (i.e., Tesla Powerwall 2), holds enough energy to delivery 30–40 km of charge while still having 50% energy reserved for other household appliances.

For example: in Germany some PEVs are already charged at a rate of 20 kW using three-phase power. Assuming a PEV market share of 100% (i.e., 40 million cars), this could lead to in an increase in the German electricity load by 800 GW, which would overshoot the current peak load by a factor of eight. The current development for fast charging stations along highways with charging rates in excess of 350 kW may even result in higher theoretical load peaks.

Empirically, we currently see that most PEVs are charged at lower rates at home or at workplaces—which makes the likelihood of this maximum load scenario unrealistic, however, increases in peak demand are nevertheless likely, highlighting the importance of implementing some form of charging control.

Synergies of EV Charging and the Electricity System

Increasing the Degree of Capacity Utilization of Grid Infrastructure

Today the workload of electricity grids in developed countries is embarrassing. The maximum load is rarely reached and during nights the operating grid is almost negligible. Grid profiles are generally “peaky” leading to the low utilization of many assets to cater for periods of demand that are high in load, but relatively short in time. As such, a flattening of grid load profiles could substantially multiply the energy throughput and efficiency of electricity grids. In residential areas, PEVs could play a significant role in filling the night load “valleys,” that is, they are generally parked overnight and connected to the grid, and they have high-energy demands (Fig. 1). Hence, a reliable, automatic control of PEV charging loads can offer a significant load shift potential—on average,

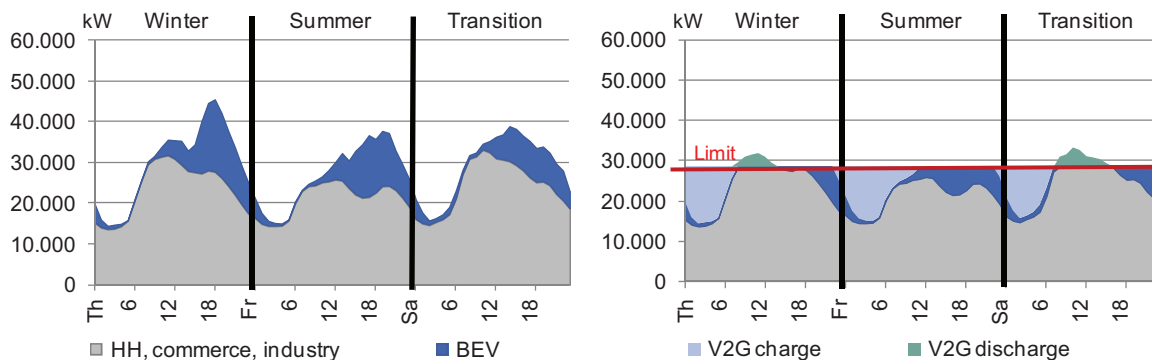


Figure 1 Exemplary V2G application for a distribution grid in Germany.

equivalent to the daily total electricity consumption of a single person household (around 8 kWh per day). Accordingly, given a high proportion of existing electricity fees in many nations is associated with the fixed costs of the grid, and the low utilization of some infrastructure, if deployed with control strategies, PEVs have the potential to increase energy demand without additional investments to the grid and, therefore, reduce electricity fees (per energy unit).

Integrating (Local) Provision of Electricity From Fluctuating Renewable Energy Generation

In many countries, renewable energy generation, particularly from solar and wind, is increasing considerably. This energy transition is going to continue over the coming decades and reverses the paradigm of “electricity supply following electricity demand” given electricity supply fluctuations will increase, leading to an increase in the share of curtailed electricity and, hence, inefficiency.

Flexible loads, such as controlled charged PEVs, can decrease the degree of curtailment. Many studies show that excess electricity from wind—especially in the night hours—can be suitably used for the controlled charging of PEVs. Furthermore, controlled charging increases the profitability of local decentralized electricity systems, such as photovoltaic solar systems, because PEVs can increase self-consumption rates (Kaschub et al., 2016; Seddig et al., 2019). As a consequence, PEVs can accelerate further investments in decentralized renewable electricity generation.

While further research is ongoing in regards to bidirectional charging, this technology has the potential to further accelerate the uptake of renewable energy given it can temporarily overcome local electricity shortages (e.g., dead calm) by exporting electricity directly from PEVs batteries. Note that a current battery capacity of an PEV is between 30 and 100 kWh, which is sufficient to cover the average two-person household electricity demand for between 3 and 12 days. This makes them also an attractive option where supply is unstable, for example, due to a weak electricity grid or for properties relying on self-supply. In many countries, if 100% of the car fleet was electric, the aggregate battery energy storage capacity would exceed the nation’s entire electricity demand for 24 hours, while still meeting the mobility needs of the fleet.

It should also be noted that used batteries from PEVs may serve as stationary energy storage through second-life applications, further supporting the economic integration and uptake of renewable energy. Overall, these co-benefits of PEVs can make electricity systems more reliable while decreasing their carbon intensity.

Decreasing Electricity Prices and System Costs While Increasing the Efficiency of Conventional Power Plants

One further incentive for encouraging load flexibility comes from the fluctuating prices of wholesale electricity markets. New business models are being developed which aggregate flexible loads for generating bids into different electricity markets, that is, lowering demand when electricity prices are high and increasing demand when they are low or even negative (Ensslen et al., 2018). In principle, the load flexibility of PEVs can be provided to all electricity markets from futures, spot, or ancillary service markets. V2G hardware further increases the potential of these services. The activities of these new energy market participants, the aggregators, and lead to decreasing average wholesale electricity prices and consequently to lower system costs. Furthermore, they maximize, on average, the utilization of the decreasing capacities of fossil fuel power plants. Again, overall, this leads to decreasing system costs.

User Acceptance of Load Flexibility

As (almost) no incentive for charging control is implemented yet, the user acceptance of these control mechanisms is unclear. An intelligent incentive for encouraging users to subscribe to those programs, and to behave correspondingly in their everyday life, is a tremendous challenge. A wrongly set incentive might even lead to undesired outcomes. Current research suggests that many PEV drivers hold positive attitudes towards such programs especially if the environmental benefits are highlighted. Many current users already combine a PEV with photovoltaic solar panels, home batteries and further smart home technologies, which underlines this interest. A frequent motivation expressed for adopting these new energy technologies is for consumers to increase their independence from the electricity system. In contrast, pilot projects have shown less enthusiasm for joining flexible electricity tariffs for other household activities and have pointed out that expectations of monetary rewards for flexibility provided usually exceed realistic market prices. In addition to this, only few very technically oriented users have a preference for manual control—most favor comfortable models which do not require expertise knowledge.

Taking these findings together, this highlights that business models need to be defined according to consumer needs: clear communication of how the system operates; the ability to override the system if needed; confidence and certainty in still obtaining the electricity required for the vehicle to fulfill the following day’s mobility services; and finally, a clear association between control strategies and lower electricity prices. However, it is also important to note that a strategy that mainly focuses on monetary motivation bears the risk to fail, as the perceived expense might be perceived as relatively high. More promising is the combination with other motives like enabling self-supply and autarchy, or a comfortable contribution to environmental goals, or additional services like a remote control of the vehicle charging in combination with anti-theft protection or a battery conservation program. It is important to note that monetary incentives do not only play a role as pure financial benefits but are also a means of communication to direct user behavior towards systemic goals.

Discussion and Conclusions

As the market share of PEVs continues to increase (Gnann et al., 2018), if they are charged without any control, electricity grids globally may be pushed toward their limits. Conversely, if the charging process can be controlled by an entity (e.g., the utility), this can lead to significant synergies between the electricity system and the decarbonization of road transport.

Today, distribution grid operators see PEVs often as a silent hazard as the market diffusion in a certain district can neither be forecasted nor observed—and after the first black-out it is already too late for grid extensions or control of the vehicles, because both solutions are time consuming. Hence, from a grid perspective the need for controlling PEV charging is key for the efficient and synergetic market integration of PEVs.

Technically, the control of PEV charging is relatively easy and has already been proven through implementation with many other technologies, such as night storage heating systems (Fig. 2). Today, the standards for communication between PEVs and the energy system are already in place: IEC 15118 assures an efficient communication between the car and the charging infrastructure (electric vehicle supply equipment, EVSE) and IEC 63110 outlines the requirements for processing the data exchange from the EVSE to the energy system. For controlled charging, communication with the EVSE seems unavoidable and therefore charging at domestic sockets will not be an appropriate long-term strategy. As such, PEV specific charging cords (cf. IEC 62196) together with intelligent EVSE are necessary. One exception might be a control signal (e.g., via GSM) to the on-board controller of the car. With all this in mind, it is important to note that to date, the controlled charging of PEVs has still only been applied in small field tests.

Besides charging control strategies, the load flexibility of PEVs also depends on consumer acceptance and individual circumstances: a charge on a long business trip may lead to minimal load flexibility while an average overnight charge at home provides much higher flexibility. User pattern analysis of conventional mobility data shows, however, that the idle time during the week at home exceeds in average 12 hours and is even significantly higher during the weekends. Hence, an average charge for 40 km (i.e., about 8 kWh) can be provided easily by a charging rate of less than 1 kW, which is far from challenging for the grid—even at a PEV market share of 100%.

Overall, the potential for mitigating greenhouse gases in road transport by PEVs is also dependent on the carbon-intensity of battery production and recycling processes. Today's production capacities are still dominated by countries with high fossil-fuel dependent electricity mixes, which can lead to undesirable environmental impacts. However, more recent production facilities already use local electricity from renewable energy resources, which makes PEVs (besides an increase in market shares of non-motorized traffic modes and car sharing) a significant opportunity for the sustainable, mobility transition.

While from a technical point of view the implementation of comprehensive charging programs mainly seems to be a question of time, user support for such a system change is also highly important. Finally, there is a promising chance that PEVs do not alone contribute to greenhouse gas reductions in the transport sector, but in the electricity sector, too. How these synergies are influenced by further developments in autonomous vehicles and car sharing is still unclear, but should also be considered in preparing for the uptake of this beneficial technology.

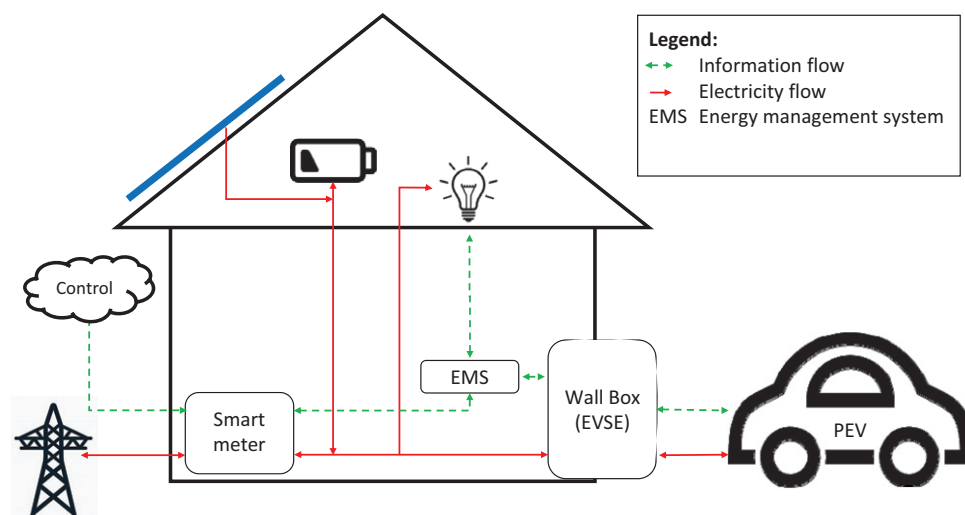


Figure 2 Stylized information and electricity flows for smart charging at home.

Biographies



Dr. Patrick Jochem is a postdoc at the Karlsruhe Institute of Technology (KIT) and Associate Editor of Transportation Research Part D: Transport and Environment. Since 2009 he is leading the interdisciplinary research group “Transport and Energy” at the Chair of Energy Economics and since 2012 the eMobility Lab at the Karlsruhe Service Research Institute (KSRI). He is author of more than 100 scientific peer-reviewed papers and has a strong funding record. In 2009, he received his PhD in transport economics, which was funded by the German Federal Environmental Foundation (DBU). In 2014, he was awarded by the Heidelberg Academy of Sciences and Humanities. He studied economics at the universities of Bayreuth, Mannheim and Heidelberg. His research interests are in the fields of electric mobility, energy system analysis, and ecological economics.



Dr. Jake Whitehead is the Founder and Secretariat of the Australia and New Zealand Clean Transport Coalition, a member of the International Electric Vehicle Policy Council, and holds two PhDs in Transport Science and Engineering. Dr Whitehead’s research primarily focuses on the costs, benefits, opportunities, and risks of clean transport technologies, as well as understanding the impacts of government policies on supporting and managing the uptake of these innovations. Dr Whitehead has also undertaken research on the synergies between clean transport and energy systems, in particular through the rollout of electric vehicle-to-grid systems, as well as the effectiveness of different transport technology pathways on reaching global emission reduction targets. Dr Whitehead has worked closely with governments and businesses internationally to advise on sustainable transport policies. These efforts have included coordinating the development of Australia’s most comprehensive electric vehicle strategy for Queensland.



Dr. Elisabeth Dütschke studied psychology, business administration and marketing at TU Darmstadt and RWTH Aachen. For her PhD thesis in organizational behavior at the university of Constance she received an award from Südwest Metall as an outstanding contribution to research. Further work experience includes consulting of private and public organizations and journalism as well as university lectures. Since June 2009 at the Fraunhofer ISI as senior scientist and project leader she has been involved in a large number of national and international research projects. Since March 2019 coordinator of the Unit “Actors and Social Acceptance in the Transition of the Energy System.” Her work focuses on the human perspective on a changing energy system. She is the main contact for societal issues around the energy transition for the institute.

References

- Babrowski, S., Heinrichs, H., Jochem, P., Fichtner, W., 2014. Load shift potential of electric vehicles in Europe: chances and limits. *J. Power Sources* 255, 283–293, doi:10.1016/j.jpowsour.2014.01.019.
- Creutzig, F., Jochem, P., Edelenbosch, O.Y., Mattauch, L., van Vuuren, D.P., McCollum, D., Minx, J., 2015. Transport—a roadblock to climate change mitigation? *Science (Policy Forum)* 350 (6263), 911–912, doi:10.1126/science.aac8033.
- Ensslen, A., Ringler, P., Dörr, L., Jochem, P., Zimmermann, F., Fichtner, W., 2018. Incentivizing smart charging: modeling charging tariffs for electric vehicles in German and French electricity markets. *Energ. Res. Soc. Sci.* 42, 112–126, doi:10.1016/j.erss.2018.02.013.
- Gnann, T., Stephens, T.S., Lin, Z., Plötz, P., Liu, C., Brokate, J., 2018. What drives the market for plug-in electric vehicles? A review of international PEV market diffusion models. *Renew. Sustain. Energ. Rev.* 93, 158–164.
- Jochem, P., Babrowski, S., Fichtner, W., 2015. Assessing CO₂ emissions of electric vehicles in Germany in 2030. *Transp. Res. A Policy Pract.* 78, 68–83, doi:10.1016/j.tra.2015.05.007.
- Jochem, P., März, A., Wang, Z., 2018. How might the German distribution grid cope with 100% market share of PEV? Impacts from PEV charging on low voltage distribution grids, Proceedings of EVS31 Conference, Kobe, Japan.

- Kaschub, T., Jochem, P., Fichtner, W., 2016. Solar energy storage in German households: profitability, load changes, and flexibility. *Energy Policy* 98, 520–532, doi:10.1016/j.enpol.2016.09.017.
- Seddig, K., Jochem, P., Fichtner, W., 2019. Two-stage stochastic optimization for cost-minimal charging of electric vehicles at public charging stations with photovoltaics. *Appl. Energy* 242, 769–781, doi:10.1016/j.apenergy.2019.03.036.

Further Reading

- Dütschke, E., Paetz, A.-G., 2013. Dynamic electricity pricing—which programs do consumers prefer. *Energy Policy* 59, 226–234, doi:10.1016/j.enpol.2013.03.025.
- Hardman, S., Jenn, A., Tal, G., Axsen, J., Beard, G., Daina, N., Figenbaum, E., Jakobsson, N., Jochem, P., Kinnear, N., Plötz, P., Pontes, J., Refa, N., Sprei, F., Turrentine, T., Witkamp, B., 2018. A review of consumer preferences of and interactions with electric vehicle charging infrastructure. *Transp. Res. Part D* 62, 508–523, doi:10.1016/j.trd.2018.04.002.
- Heinrichs, H., Jochem, P., 2016. Long-term impacts of battery electric vehicles on the German electricity system. *Eur. Phys. J. S. Top.* 225, 583–593, doi:10.1140/epjst/e2005-50115-x.
- Kazhamiaka, F., Jochem, P., Keshav, S., Rosenberg, C., 2017. On the influence of jurisdiction on the profitability of residential photovoltaic-storage systems. *Energy Policy* 107, 428–440, doi:10.1016/j.enpol.2017.07.019.
- Smit, R., Whitehead, J., Washington, S., 2018. Where are we heading with electric vehicles? *Air Quality Climate Change* 52, 18–27.
- Sovacool, B.K., Noel, L., Axsen, J., Kempton, W., 2017. The neglected social dimensions to a vehicle-to-grid (V2G) transition: a critical and systematic review. *Environ. Res. Lett.* 13 (1), 13001, doi:10.1088/1748-9326/aa9c/6d.
- Webb, J., Whitehead, J., Wilson, C., 2019. Where is the juice? the impact of electric vehicle charging infrastructure on the distribution network. In: Sioshansi, F. (Ed.), *Customer Stratification—How Innovation in Services Will Lead to Disruptions in the Electric Power Sector*. Cambridge, Elsevier, pp. 407–429.
- Whitehead, J., Washington S., Zheng, Z., Perrons, R., (under review). Charging up: an analysis of the supply and demand factors influencing Australia's electric vehicle market. *Transp. Res.*

Uber Versus Taxis

Georgina Santos, School of Geography and Planning, Cardiff University, Cardiff, United Kingdom

© 2019 Elsevier Ltd. All rights reserved.

Introduction	566
What is Uber?	566
Controversies	567
Incumbent Taxis	567
Drivers	567
Sustainable Transport	567
Regulation	568
Discussion	569
Mobility as a Service	570
Conclusions	570
References	570

Introduction

Horse-drawn carriages for hire were already present in Paris and London at the beginning of the 17th century: "It was in 1605 that hackney coaches came into use" (Gilbey, 1903, p. 26). By 1662, the number of hackney coaches in London had increased to 2490 and they were considered a "nuisance." In order to reduce their number, that year, a licensing system was introduced, with only 400 licenses granted (Gilbey, 1903, pp. 32–33). In 1637, "closed cars led by a coachman who responds to the calls of travellers" started operation in Paris, and about 20 years later, their number was also limited in Paris, through a cap of 600 licenses (Piot, 2015).

Horses were replaced with motorized transport but the hackney licenses have continued to this day, with similar systems in most cities throughout the world. At different points in time, but mostly over the 20th century, most cities also saw the emergence of private hire vehicles. Private hire vehicle companies and private hire vehicle drivers also need a license, usually granted by the local authority/municipality.

Taxis (or hackney cabs) and private hire vehicles are close substitutes. The main difference is that private hire vehicles need to be booked and cannot be hailed on the street or in a rank. The booking used to be done either in person in the company's premises or by telephone. Taxis can be booked in advance, hailed on the street, or picked up from designated taxi ranks. With the arrival of the Internet, businesses of all sorts, including taxi and private vehicle hire companies, started to build websites and in time, they also allowed bookings and sales online. As a result, by the beginning of the 2000s, it had become common for potential customers to use a search engine in order to book a taxi, a private hire vehicle, a flight, or hotel, among other goods and services. The arrival of smartphones with Internet connection signaled another milestone. The next step was, unsurprisingly, smartphone applications, that is, computer programs designed to run on mobile devices. Taxi and private hire companies then added GPS (which stands for global positioning systems) to the equation, to help dispatch, trace, and manage different taxis or private hire vehicles in a fleet, identifying drivers and passengers closest to each other, something that used to be done by radio or phone before GPS and the Internet.

This technology-enabled natural progression started well before Uber came into existence. Taxis and private hire vehicles have historically been subject to different regulations. This piece compares Uber and taxis, paying attention to various controversies and the different regulations they operate under. Uber started in 2009 and as of 2019, it claims to have at least 50% share of the private hire vehicle market in all the countries where it operates (Uber Technologies, Inc., 2019).

What is Uber?

Uber (<https://www.uber.com>) is a company that started in San Francisco, California, in 2009, and provides ride-hailing services. Over the years, it has added other services such as food delivery and bike-sharing. It is a very prominent name to the point that the emergence of the sharing economy, under which goods and services are not owned but accessed when needed, is often referred to as "uberization" of the economy.

Uber is an example of what Santos (2018) calls Model 3 of shared mobility. This Model 3 is referred to as ride-hailing, ride-sourcing, e-hailing, or transportation network companies. Under this model, companies act as brokers, matching drivers, who drive their own cars (or rented cars) and passengers. Passengers book rides through a smartphone application and are quoted a fare. When demand is high and there are not enough Uber cars to satisfy all requests, Uber fares go up. Uber calls this "surge pricing," and lets riders know the fare has gone up through the app at the time of booking.

One big competitor for Uber, worldwide, is Lyft (<https://www.lyft.com>). Other companies operating under Model 3 include Kapten (<https://www.kapten.com>), Bolt (<https://bolt.eu/>), and Xoox (<https://xoox.co.uk>), but there are many others, mostly operating in one or a few countries only.

The other models of shared mobility for cars and vans are Model 1 (peer-to-peer car rental, without a driver), Model 2 (modern car-clubs or car sharing, without drivers), and Model 4 (ride-sharing, ride-splitting, micro-transit, or new public transport on demand), all of which, like Model 3, rely on smartphone applications (Santos, 2018). Uber also offers UberPOOL, just like Lyft offers Lyft Line, both of which entail passengers sharing a ride. This type of service is cheaper because the ride and the costs are shared by passengers going in the same direction, who are picked up and dropped off at different locations on the same route. This falls under Model 4 in Santos (2018).

Controversies

Uber has been the subject of much controversy on a number of fronts. These have spanned from claims about unfair competition from the incumbent taxi industry, to employment status from the very Uber drivers, to claims that Uber and other ride-hailing companies increase traffic congestion, air pollution and CO₂ emissions, and take passengers away from public transport and active transport modes, like walking and cycling.

Incumbent Taxis

Uber and other ride-hailing companies are close substitutes of traditional taxis, as they provide transport services similar to those provided by taxi companies, and therefore compete with them. In Boston, Los Angeles, New York, San Francisco, and Seattle in the United States, the utilization rate of Uber drivers is higher than the utilization rate of taxi drivers, both when measured by time and by miles (Cramer and Krueger, 2016). Incumbent taxi drivers in the 50 largest metropolitan statistical areas in the United States have seen their relative earnings decrease by 10% after Uber's entry into the market (Berger et al., 2018). In the City of Toronto in Canada, Uber's entry is associated with a significant reduction in taxi trips (Young and Farber, 2019).

Regulations for taxis and for ride-hailing companies are different, as discussed further down. This has fuelled protests over the years from taxi drivers and taxi companies in a number of countries across all continents (BBC News, 2014, 2016).

The incumbent taxi industry argues that thousands of jobs are at risk because Uber and similar companies take their passengers away under unfair competition. We return to this point later on.

Drivers

For some time, but more so over 2018 and 2019, Uber drivers throughout the world have complained about low pay and poor working conditions (BBC, 2018a, 2019; Reuters, 2019; The Guardian, 2019). The key point is that Uber considers all Uber drivers to be independent contractors, or freelancers, rather than employees. Independent contractors are not entitled to a minimum wage or overtime, or health insurance, or scheduled breaks, or paid holidays.

As a result, in August 2013, Uber drivers filed a lawsuit in California claiming the company had to treat them as employees, not contractors, hoping to recover the tips they had never received and reimbursement for expenses such as fuel and vehicle maintenance (Lichten & Liss-Riordan, P.C., 2019). This was followed by the independent contractor status of Uber's drivers being challenged in courts and by government agencies in a number of other countries (Uber Technologies, Inc., 2019), including the United Kingdom (BBC, 2018b). The claim is that Uber drivers should be treated as employees or workers (Uber Technologies, Inc., 2019). Uber considers their drivers independent contractors because "they can choose whether, when, and where to provide services" on the Uber platform, and "are free to provide services on . . . competitors' platforms" (Uber Technologies, Inc., 2019). The matter is far from resolved and Uber "may not be successful in defending the independent contractor status of drivers in some or all jurisdictions" (Uber Technologies, Inc., 2019).

Sustainable Transport

On the one hand, ride-hailing increases traffic compared to private cars because there are extra miles on the way to pick-up (Schaller, 2018). In addition, collecting and dropping passengers also causes congestion (Erhardt et al., 2019). On the other hand, ride-hailing might be associated with lower car ownership and more frequent use of public transport and nonmotorized modes, and this could translate in a reduction in congestion (Feigon and Murphy, 2016, 2018; Smith, 2016).

Rayle et al. (2016) survey ride-hailing users in San Francisco, including those that use Uber. They find that ride-hailing could be increasing congestion. When asked what mode of transport ride-hailing users would have used had ride-hailing not been available, 8% would have not made the trip (which is evidence of induced demand for travel), 39% would have used a taxi (this is simply substitution), 33% would have used public transport (which carries more people and therefore congestion and emissions per passenger-mile are lower), and 6% would have driven their own car (which is substitution, but with the added miles on the way to pick-up). On similar lines, Clewlow and Mishra (2017) survey ride-hailing users in Boston, Chicago, Los Angeles, New York, San Francisco/Bay Area, Seattle, and Washington, DC and find that 49%–61% of ride-hailing trips would have not been made at all, or

would have been done by walking, cycling, or public transport, had ride-hailing not been available. If public transport patronage is decreasing because passengers are switching to ride-hailing, then the financial viability of public transport could be in danger. Furthermore, public transport is usually seen as the backbone of any sustainable city if congestion, air pollution and CO₂ emissions are to be minimized. [Clewlow and Mishra \(2017\)](#) also find that 91% of ride-hailing users who responded to their survey had not made any changes regarding whether to own a vehicle or not, in line with [Erhardt et al. \(2019\)](#), but in contrast with [Feigon and Murphy \(2016\)](#), who find that ride-hailing reduces the need to own a car.

Shared rides, however, could reduce congestion, air pollution, and CO₂ emissions when replacing solo trips made by car ([Furtado et al., 2017](#); [Schaller, 2018](#); [Viegas and Martinez, 2016, 2017](#)). The key here seems to be that for shared rides to yield these benefits, they need to replace solo trips, not trips by public transport. [Schwieterman and Michel \(2016\)](#) compare the different performance characteristics of UberPool and public transport services, both of which entail sharing a ride with fellow passengers, between downtown Chicago in the United States and the city's north- and northwest-side neighborhoods. Out of 50 trips that members of the research team took to and from the same location, starting at the same time, one by public transport and one by UberPOOL, 39 were faster with Uber. The average price was \$9.66 with Uber and \$2.29 with public transport. Although more expensive, UberPOOL could attract public transport passengers, especially during off-peak hours when there is no surge pricing in UberPOOL and there is no congestion on roads. This, again, would not be a good outcome for public transport because it would lose revenues. In addition, public transport in general can carry more passengers than any ride-sharing van or minibus, and thus minimize congestion, air pollution and CO₂ emissions, provided the bus/train/tram/etc in question is not (virtually) empty.

It should be highlighted that all these studies were done for US cities and may not be transferable to other regions of the world, where passengers may not switch from public transport, walking and cycling to ride-hailing (such as Uber) or to ride-sharing (such as UberPOOL).

One point that is widely recognized, however, is that ride-hailing can complement public transport. Although [Clewlow and Mishra \(2017\)](#) find that ride-hailing takes passengers away from public transport, they also find that ride-hailing is sometimes used to reach rail stations, and so, acts as a complementary mode for commuter rail services, with a 3% net increase in commuter rail use on average in Boston, Chicago, Los Angeles, New York, San Francisco/Bay Area, Seattle, and Washington, DC. These are the traditionally problematic first- and last-mile connections to and from public transport stations, where ride-hailing can provide a solution ([Erhardt et al., 2019](#), p. 1). Using data for Metropolitan Statistical Areas in the United States, [Hall et al. \(2018, p. 46\)](#) conclude that Uber complements public transport in cities where public transport patronage was low before Uber's entry, and substitutes for public transport in cities where public transport patronage was high before Uber's entry. They also find suggestive evidence that Uber is reducing commuting times for public transport users, because the first- and last-mile segments of a trip by public transport usually represent a small share of the distance traveled but a large share of the travel time, and by substituting for this segment, Uber can reduce the cost of public transport for the main segment of the trip ([Hall et al., 2018, p. 37](#)). This does, however, increase traffic congestion ([Hall et al., 2018, p. 44](#)).

On top of helping with the first- and last-mile problem, Uber and similar companies, can reach areas not well served by public transport, and operate at times when public transport does not, or at least, not frequently. Because of this, transport agencies in a number of European and North-American cities are exploring ways and even partnering with ride-hailing companies to provide first- and last-mile rides to and from public transport stations ([Crozet et al., 2019](#)).

Regulation

Competition is good. Consumers benefit from competition as they get the best services at the lowest prices, thanks to the rivalry between different suppliers. Competition pushes prices down and quality up.

The tension that has emerged between taxis and Uber is that competition may not be taking place on a level-playing field. Interestingly, while taxis complain that Uber is not subject to the same regulations, Uber complains that taxis are not subject to the same regulations. The reality is that regulations differ and the question is whether regulations should be the same for both so that they can compete on everything, or different, allowing ride-hailing companies and taxis to compete only on some features.

As highlighted in the introduction, one important difference between taxis and private hire vehicles, which include ride-hailing, such as Uber, is that taxis can be hailed on the street, on taxi ranks or booked in advance, while private hire vehicles can only be booked in advance.

Although quantity caps may reduce consumer welfare due to lower taxi availability and longer waiting times ([Competition & Markets Authority, 2017](#)), often, the number of licensed taxis is capped. This is the case in New York and Paris, for example. In contrast, the number of private hire vehicle licenses is typically not capped. New York, however, approved a cap on ride-hailing car licenses in August 2018. This was followed by the Mayor of London, Sadiq Kahn, making a similar proposal, arguing that the substantial increase in the number of private hire vehicles on London's roads in recent years had increased congestion and air pollution, and had left many drivers of such vehicles struggling to make enough money ([BBC News, 2018c](#); [The Guardian, 2018](#)). Three years before, the previous Mayor, Boris Johnson, had pushed for the same powers ([London Government, 2015](#)).

The fares that ride-hailing companies, such as Uber, charge are not capped or regulated either, whereas taxi fares tend to be heavily regulated. An argument for regulating taxi fares is that "passengers are in a relatively weak position to compare offers and negotiate prices in the hail and rank (taxi) trade" ([Competition & Markets Authority, 2017](#)), unlike in the private hire vehicle market. Taxis have a meter fitted within the vehicle, which calculates the fare for a journey, which also has a floor and a cap, and in some cities, an

extra fee during rush hour or at night time or when carrying luggage. The passenger typically knows the final price of her journey once she reaches her destination and the meter stops.

Ride-hailing companies such as Uber, on the other hand, are free to charge any fare, although the fare is shown to the customer in advance at the moment of booking the ride. Typically, fares are low when demand is low and high when demand is high. There is no meter fitted in Uber (or any other ride-hailing) vehicles. When ride-hailing companies charge very low fares, this can be expected to trickle down to the driver, who gets a low pay for her work. In December 2018, the New York Taxi and Limousine Commission passed regulations on new pay rates for drivers providing for-hire services in New York City and surrounding areas ([Uber Technologies, Inc., 2019](#)). This translated in Uber increasing the fares it charges in the affected areas. Although the regulator in this case did not impose a price floor directly, by imposing a minimum pay rate for drivers, it effectively did.

In addition to the above, in some cities such as London, taxi drivers need to pass knowledge tests demonstrating they know all the roads and back roads in the city, whereas private hire vehicle drivers need only pass a more basic test, where they are allowed to use GPS to find their way around. In the United States, Uber drivers do not need to pass any test at all; all they need is a driving license. The type of driving license required is different in the United States and in Europe. Most countries in Europe require Uber and all ride-hailing drivers to hold professional driving licenses. That is why UberPOP, which is the version of Uber with the ordinary driver, does not exist in Europe. Instead, the service offered is UberX, where the driver holds a professional driving license.

In London, black cabs, which incidentally are all wheelchair accessible, are exempt from paying the congestion charge, but private hire vehicles, including ride-hailing ones like Uber, are not, unless they are wheelchair accessible. Similarly, taxis are allowed to use bus lanes, but ride-hailing vehicles like Uber are not.

All taxi and private hire vehicle drivers, including ride-hailing drivers, are in most towns and cities throughout the world, subject to medical and criminal/background checks, need to hold insurance, and have their vehicles inspected periodically, among other matters.

Taxi and ride-hailing regulations have some similarities but they also differ. If regulations differ, taxis and Uber will only be able to compete where they can compete. Taxis in New York, for example, have extended their geographic coverage outside Manhattan, as Manhattan has been partly grabbed by Uber. This wider coverage has benefited customers ([Kim et al., 2018](#)). In London, incumbent London taxicabs accept credit card payments, which they resisted for years ([BBC News, 2017](#)). [Wallsten \(2015\)](#) also finds an association between Uber's increasing popularity and a decrease in consumer complaints about taxis in both New York and Chicago, which, he claims, may be possible to attribute to taxi drivers being more courteous to passengers as a result of the competition they face from Uber.

[Cramer and Krueger \(2016\)](#) find that in the United States, the capacity utilization rate is 30% and 50% higher for UberX drivers than taxi drivers, when measured by time and by miles, respectively, and they argue that two of the reasons for this are inefficient taxi regulations, and Uber's flexible supply model and surge pricing.

Whether regulations should be homogenized or not is a matter of debate. Homogenizing one regulation, may require homogenizing others. If fare regulation were scrapped for taxis, and these were able to fully compete on price, or if ride-hailing companies were subject to the same fare structure as taxis, then ride-hailing vehicles such as Uber would also probably be allowed to pick passengers up on streets and ranks, and neither type of vehicles would be capped or both types would receive the same number of licenses. This would make both offerings essentially identical, leaving the remaining differences to subtle details of service differentiation, such as for example, the type of knowledge test that drivers need to pass.

Discussion

After only 10 years of starting operations, Uber claims to have a market share within the private hire vehicle market of at least 50% in all the countries where it operates. The growth of the company is impressive, by any standards.

What is curious, however, is that Uber grew without offering a new, original, product. By 2009 most cities throughout the world had had private hire vehicle companies for decades. These companies matched drivers and passengers close by and accepted bookings via the Internet. Most drivers in these companies drove their own vehicles, just like Uber drivers do.

One controversial element is that originally Uber drivers did not hold professional driving licenses but normal driving licenses. This is still the case in the United States and some other countries. This may have helped charge very low fares and operate with great flexibility, to attract passengers from taxis, as well as from public transport, walking and cycling, and to induce trips that would have not taken place had Uber not been an option.

However, in many countries, Uber drivers are now required to hold professional driving licenses. Their fares are not regulated yet but New York has gone as far as setting minimum pay rates for drivers and limiting the number of private hire vehicle licenses, so the regulations under which ride-hailing companies are operating are starting to resemble those for taxis.

Capping numbers is a text-book example of a "command-and-control" policy. This may ultimately reduce the choice available to customers, but it may also contain the increase in traffic and congestion, air pollution and CO₂ emissions caused by ride-hailing companies, at least in those cities where this has been proved to be the case. In London, all ride-hailing companies, including Uber, used to be exempt from the congestion charge. This changed in April 2019. All private-hire vehicles, including Uber, pay the congestion charge now. As a result, Uber passengers are charged an extra pound for journeys going in, going out, or taking place inside the congestion charging zone, to help cover the £11.50 daily charge that drivers have to pay ([The Independent, 2019](#)). Despite ride-hailing companies paying the congestion charge in London, the Mayor is applying for powers to introduce a cap, as highlighted above.

Ride-hailing services could potentially focus on trips that increase social welfare. Three such examples are: (1) When ride-sharing replaces solo trips by car (as congestion, air pollution and CO₂ emissions per passenger-mile are lower); (2) When ride-hailing reaches areas not served by public transport, or is used when public transport is not in operation, such as for example, at nighttime; and (3) When ride-hailing facilitates first- and last mile connections to and from public transport stations. This could, however, increase congestion, air pollution and CO₂ emissions in the proximity of busy public transport stations.

Mobility as a Service

A word needs to be said about Mobility as a Service (MaaS), a concept that can be considered “the currently most widely discussed and hyped concept in the transport domain” (European Metropolitan Transport Authorities, 2019). The International Association of Public Transport (2019) defines MaaS as the “integration of, and access to, different transport services (such as public transport, ride-sharing, car-sharing, bike-sharing, scooter-sharing, taxi, car rental, ride-hailing and so on) in one single digital mobility offer, with active mobility and an efficient public transport system as its basis.” The idea is that MaaS provides mobility solutions that are consumed as a service. It offers integrated planning, booking, and payment for journeys, including multimodal journeys, through a digital platform, helping users satisfy all their travel needs without owning a car.

Uber can be part of MaaS and is so keen that it is a member of the MaaS Alliance and its Board of Directors (<https://maas-alliance.eu/maas-alliance-partners-uber-support-shared-mobility/>). The aim of the Alliance, as stated on their website, is to provide an alternative to the private car that is efficient and sustainable. Interestingly, there is no mention of public transport or active transport as being the main alternatives to the private car or the “basis” of MaaS, as defined by the International Association of Public Transport (2019).

Regardless of whether Uber complements or substitutes for public transport, Uber can increase congestion, air pollution, and CO₂ emissions simply by increasing the number of trips taken (Hall et al., 2018). However, Uber’s impact on road transport externalities will be larger if it substitutes for public transport (Hall et al., 2018). The evidence for the United States is mixed, with some work pointing towards Uber being a substitute and some work pointing towards Uber being a complement. Interestingly, when filing to go on the stock market, Uber itself publicly declared that they plan to target public transport passengers in the long run:

“We address a massive opportunity in powering movement from point A to point B... We view our market opportunity in terms of a total addressable market (‘TAM’), which we believe that we can address over the long-term . . .

Our Personal Mobility TAM consists of 11.9 trillion miles per year, representing an estimated \$5.7 trillion market opportunity in 175 countries. We include all passenger vehicle miles and *all public transportation miles in all countries** globally in our TAM, including those we have yet to enter . . . ”

(Uber Technologies, Inc., 2019, p. 10) *Emphasis added by the author.

Conclusions

Uber is a ride-hailing company that started in 2009 and 10 years later has at least 50% of the private vehicle hire market share in all the countries where it operates (Uber Technologies, Inc., 2019). It has been a subject of controversy, with discontent from the taxi industry, from its own drivers, and a debate on whether Uber and similar companies are to blame for increases in congestion, air pollution and CO₂ emissions in a number of cities in the United States.

The set of regulations under which Uber and taxis operate is different. This has caused tensions but it is unclear whether regulations should be homogenized. Under current arrangements, there are arguments for regulating taxi fares, as potential passengers hailing them on the street or at taxi ranks are not in a position to negotiate prices. Most cities also have price floors, protecting taxi drivers, and price caps, protecting passengers. These arguments do not apply to Uber or ride-hailing companies in general. However, if Uber and other ride-hailing companies were allowed to pick passengers up on the street like taxis are, then there would be an argument for homogenizing fare structures. Similarly, if ride-hailing cars were made wheelchair accessible like taxis are, there would be an argument for exempting them from the congestion charge in London.

If they remain close substitutes subject to different regulations, taxis and Uber will only be able to compete on what they can compete. Uber vehicles cannot be hailed on the street but they can attract passengers by charging very low fares. Taxis cannot compete on price, but they can be hailed on the street and they can ensure they provide a good quality service by drivers being polite and accepting card payments.

Ultimately any change in regulation should aim at helping Uber and other ride-hailing companies, as well as taxis, benefit consumers, as long as they do not substitute for public transport. Public transport and active transport, not ride-hailing or taxis, should be the backbone of any sustainable urban transport system.

References

- BBC News, 2014. Taxi and rail strikes hit European cities. 11 June. Available from: <https://www.bbc.co.uk/news/world-europe-27792942>.
 BBC News, 2016. Jakarta taxi drivers protest against Uber and Grab. 22 March. Available from: <https://www.bbc.co.uk/news/world-asia-35868396>.
 BBC News, 2017. Uber is officially a cab firm, says European court. Available from: <https://www.bbc.co.uk/news/business-42423627>.

- BBC News, 2018a. Uber drivers stage UK strikes over pay and conditions. 9 October. Available from: <https://www.bbc.co.uk/news/business-45796409>.
- BBC News, 2018b. Uber loses latest legal bid over driver rights. 19 December. Available from: <https://www.bbc.co.uk/news/business-46617584>.
- BBC News, 2018c. Mayor of London calls for power to limit minicab numbers. 15 August. Available from: <https://www.bbc.co.uk/news/uk-england-london-45196375>.
- BBC News, 2019. Uber drivers strike over pay and conditions. 8 May. Available from: <https://www.bbc.co.uk/news/business-48190176>.
- Berger, T., Chen, C., Frey, C., 2018. Drivers of disruption? Estimating the Uber effect. *Eur. Econ. Rev.* 110, 197–210.
- Clewlow, R., Mishra, G., 2017. Disruptive transportation: the adoption, utilization, and impacts of ride-hailing in the United States, Research Report UCD-ITS-RR-17-07, Institute of Transportation Studies, University of California, Davis. Available from: https://itspubs.ucdavis.edu/wp-content/themes/ucdavis/pubs/download_pdf.php?id=2752.
- Cramer, J., Krueger, A., 2016. Disruptive change in the taxi business: the case of Uber. *Am Econ Re Papers Proc* 106 (5), 177–182.
- Crozet, Y., Santos, G., Coldefy, J., 2019. Shared mobility, MaaS and the regulatory challenges of urban mobility, CERRE Report. Available from: https://www.cerre.eu/sites/cerre/files/cerre_sharedmobility_maas_report_2019.pdf.
- Erhardt, G., Roy, S., Cooper, D., Sana, B., Chen, M., Castiglione, J., 2019. Do transportation network companies decrease or increase congestion? *Sci. Adv.* 5, eaau2670.
- European Metropolitan Transport Authorities, 2019. Mobility as a service: a perspective on MaaS from Europe's transport authorities. June. Available from: <https://www.emta.com/spip.php?article1319&lang=en>.
- Feigon, S., Murphy, C., 2016. Shared mobility and the transformation of public transit, Transit Cooperative Research Program Report 188, Transportation Research Board. Available from: <https://www.eesi.org/files/APTA-Shared-Mobility.pdf>.
- Feigon, S., Murphy, C., 2018. Broadening understanding of the interplay between public transit, shared mobility, and personal automobiles, Transit Cooperative Research Program. Available from: https://www.nap.edu/login.php?action=guest&record_id=24996.
- Furtado, F., Martinez, L., Petrik, O., 2017. Shared mobility simulations for Helsinki, International Transport Forum, Paris. Available from: <https://www.itf-oecd.org/shared-mobility-simulations-helsinki>.
- Gilbey, W., 1903. Early carriages and roads. Vinton. Available from: <https://archive.org/details/earlycarriagesro00gilb>.
- Hall, J., Palsson, C., Prive, J., 2018. Is Uber a substitute or complement for public transit? *J. Urban Econ.* 108, 36–50.
- International Association of Public Transport, 2019. Mobility as a Service, Report. April. Available from: <https://www.uiip.org/report-mobility-service-maas>.
- Kim, K., Baek, C., Lee, J.-D., 2018. Creative destruction of the sharing economy in action: the case of Uber. *Trans. Res. A Policy Pract.* 110, 118–127.
- Lichten & Liss-Riordan, P.C. (2019). Uber Lawsuit. Available from: <http://uberlawsuit.com/>.
- London Government, 2015. Press release: mayor renews call for further powers over private hire trade. Available from: <https://www.london.gov.uk/press-releases/mayoral/private-hire-vehicles>.
- Piot, J.-C., 2015. Les taxis: 378 ans d'histoires et d'engueulades. Available from: <https://blog.francetvinfo.fr/deja-vu/2015/06/14/les-taxis-378-ans-dhistoires-et-dengueulades.html>.
- Rayle, L., Dai, D., Chan, N., Cervero, R., Shaheen, S., 2016. Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. *Trans. Policy* 45, 168–178.
- Reuters, 2019. Uber drivers plan protest over pay ahead of IPO. April 29. Available from: <https://www.reuters.com/article/us-uber-ipo-strike/uber-drivers-plan-protest-over-pay-ahead-of-ipo-idUSKCN1S5276>.
- Santos, G., 2018. Sustainability and shared mobility models. *Sustainability* 10(9). Available from: <https://www.mdpi.com/2071-1050/10/9/3194>.
- Schaller, B., 2018. The New Automobility: Lyft, Uber and the Future of American Cities, Schaller Consulting. <http://www.schallerconsult.com/rideservices/automobility.htm>.
- Schwieterman, J., Michel, M., 2016. Have app will travel: comparing the price & speed of fifty CTA & UberPool trips in Chicago, Chaddick Institute for Metropolitan Development at DePaul University. June. Available from: <https://las.depaul.edu/centers-and-institutes/chaddick-institute-for-metropolitan-development/research-and-publications/Documents/Have%20App%20Will%20Travel%20Uber%20-%20CTA.pdf>.
- Smith, A., 2016. Shared, collaborative and on demand: the new digital economy. May. Available from: https://www.pewresearch.org/wp-content/uploads/sites/9/2016/05/PI_2016.05.19_Sharing-Economy_FINAL.pdf.
- The Independent, 2019. Uber to increase fares in central London to help pay new congestion charge: ride-hailing app adds £1 surcharge to every trip that enters zone. Available from: <https://www.independent.co.uk/news/uk/home-news/uber-fare-new-congestion-charge-london-a8857521.html>.
- The Guardian, 2018. Sadiq Khan wants to restrict number of Uber drivers in London. 15 August. Available from: <https://www.theguardian.com/technology/2018/aug/15/sadiq-khan-wants-to-restrict-number-of-uber-drivers-in-london>.
- The Guardian, 2019. Uber drivers strike over pay and conditions. 8 May. Available from: <https://www.theguardian.com/technology/2019/may/08/uber-drivers-strike-over-pay-and-conditions>.
- Uber Technologies, Inc., 2019. Forms S-1 registration statement under the securities act of 1933, United States Securities and Exchange Commission, Washington, D.C.; 1; 20549. April 11. Available from: https://www.sec.gov/Archives/edgar/data/1543151/000119312519103850/d647752ds1.htm#toc647752_2.
- UK Competition & Markets Authority, 2017. Guidance: regulation of taxis and private hire vehicles: understanding the impact on competition. 12 July. Available from: <https://www.gov.uk/government/publications/private-hire-and-hackney-carriage-licensing-open-letter-to-local-authorities/regulation-of-taxis-and-private-hire-vehicles-understanding-the-impact-on-competition>.
- Viegas, J., Martinez, L., 2016. Shared mobility: innovation for liveable cities, Corporate Partnership Board Report, International Transport Forum, Paris. Available from: http://transitcenter.org/wp-content/uploads/2019/02/TC_WhosOnBoard_Final_digital-1.pdf.
- Viegas, J., Martinez, L., 2017. Transition to Shared Mobility: How large cities can deliver inclusive transport services, Corporate Partnership Board Report, International Transport Forum, Paris. <https://www.itf-oecd.org/transition-shared-mobility>.
- Wallsten, S., 2015. The Competitive Effects of the Sharing Economy: How is Uber Changing Taxis? Technology Policy Institute, Washington, DC, USA. <https://techpolicyinstitute.org/wp-content/uploads/2015/06/the-competitive-effects-of-the-2007713.pdf>.
- Young, M., Farber, S., 2019. The who, why, and when of Uber and other ride-hailing trips: An examination of a large sample household travel survey. *Trans. Res. A Policy Pract.* 119, 383–392.

Contract Efficiency in Public Transport Services

Philippe Gagnepain*, Marc Ivaldi[†], *Paris School of Economics-Université Paris 1, Paris, France; [†]Toulouse School of Economics-EHESS, Toulouse, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	572
Reduced Forms	572
Contract Choice is Exogenous	573
Contract Choice is Endogenous	574
Auction Versus Negotiation	575
Structural Models	576
Conclusion	578
References	578

Introduction

When a production technology is characterized by increasing returns to scale, the regulator in charge of the organization of the service prefers to keep the monopoly status of the firm in charge of the operation. The key issue is then to design the best regulatory contract that guarantees that the activity of the firm coincides with the objectives of the regulator. In particular, the question of the incentive power of the contract in terms of the reduction of the operating costs is a particularly important issue given that it is the society as a whole that will bear the costs associated with the provision of the service.

This chapter focuses more specifically on the urban transport industry, which is usually seen as a universal service in developed countries, and where operating costs are much higher than commercial revenues and therefore require that the regulator pays the operator a socially costly subsidy. Controlling operating costs is therefore an important issue for the regulator, which ideally aims at selecting the most efficient firm to provide the service. Our objective in this chapter is to present a literature review on the relationship between the productive efficiency (the cost) of a transport operator and the incentive power of the regulatory contract that defines the cost reimbursement rules of the transportation activity. The economic analysis discussed here is usually based on the estimation of a cost function or a cost frontier, which usually incorporates an inefficiency variable. We suggest that the literature usually agrees on the fact that contracts with higher incentives lead to larger cost reductions. However, we also shed light on the fact that to identify a causal effect of a type of contract on cost efficiency is a tricky task as efficiency and the choice of the type of contract are usually jointly determined, which causes problems of endogeneity.

We present in the second section a set of estimation techniques based on reduced forms that do most of the time assume that contracts are exogenous. Then, in the third section, we focus instead on more structural approaches that do explicitly represent in the structure of the estimated cost function the source of the potential endogeneity problem. These techniques are based on a tradition initiated by Leibenstein (1966) which considers that any cost expenditure should be affected by a global inefficiency index that depends on the firm's technical ability but also on the will of the managers and workers involved in the production process. On the one hand, a low-powered incentive environment due to a lack of productivity of working agents induces the inefficiency, whereas appropriate incentives can lead to significant operating cost reductions. On the other hand, since the emergence of the new theory of regulation, economists admit that, in industries where a producer is regulated by an authority, the principal-agent relationship is at the core of the question of assessing the performance of a firm. (Baron and Myerson 1982; Laffont and Tirole, 1986; Loeb and Magat, 1979). The incentive models provide a relevant framework for the analysis of operating cost in public transport activities. In addition, because such models are able to elicit the structural relationship between the observable variables and the inefficiency term, they directly provide a way to deal with the endogeneity of the inefficiency term, that is, with the stumbling block of the econometrics of frontiers. Hence technical inefficiency and effort are two unobservable parameters, which characterize the incentives faced by a firm to reduce costs and define the source of global inefficiency.

Reduced Forms

The efficiency of a firm is typically measured from a production or a cost function, as both are a dual representation of the same technology. Cost expressions have been more popular given that the public transport industry usually considers that output levels are set by transport authorities. The researcher may be interested in the direct relationship between the operator's cost and several explanatory variables or may attempt to estimate an efficiency index that is incorporated directly in the cost structure. We present in this section the different types of functional form used by the researchers and we shed light on how the contract effect is accounted for. We present a first set of cases where contracts are assumed to be exogenous. Then we discuss several papers that relax this hypothesis and attempt to explain in a simple way how the choice of contract can be identified.

Contract Choice is Exogenous

French data on urban transportation have been very popular in the economic analysis as there is quite a long tradition on collecting data on the activity of both transport operators and local regulator in each urban area of significant size. A local authority (i.e., a municipality, a group of municipalities, or a district, whose councils are elected for six-year terms) is in charge of regulating an operator, which has been selected to operate the network, that is, to provide the transport services. This framework has been recorded in a law on the organization of transport services within France enacted in 1982, which established the principle of the decentralization of these services to local authorities from the State and provided guidance and rules for their regulation. As a result, each local authority organizes its urban transportation system by setting the route structure, the level of capacity and quality of service, the fare level and structure, the conditions for subsidizing the service, the level of investment, and the nature of ownership. It may decide to operate the network directly or to require the services of a transport service provider. In this latter case, a formal contract sets the terms of reference that the operator must comply with as well as the payment and/or cost-reimbursement rules between the authority (the principal) and the operator (the agent). [Gagnepain \(1998\)](#) focuses on a panel data set covers 49 different urban transport networks over the period 1987–2001. Two types of regulatory contracts are implemented, namely, cost-plus and fixed-price schemes. Under fixed-price contracts, operators receive subsidies to finance the expected operating deficits; under cost-plus regulation, subsidies are paid to the firms to finance ex post deficits. Hence, fixed-price regimes are very high-powered incentive schemes, while cost-plus regimes do not provide any incentives for cost reduction. The identification strategy of the paper is built upon a reduced-form model, which provides insights into the effects of certain variables without requiring a complete modeling of the environment that characterizes the production activity. More in particular, a reduced Translog cost function is used to support the estimation of operating costs. A variable, which characterizes the type of regulatory scheme in operation, is introduced to study the effect of the latter on costs. The results indicate that firms regulated by fixed-price contracts have an average cost of 2.05% lower than firms regulated by cost-of-service schemes, hence suggesting that a transport company has an incentive to reduce costs depending on the type of contract that the local authority offers.

[Kerstens \(1996\)](#) uses the same type of data but works on a smaller sample (114 observations). This study also differs from the previous one as it is based on a two-step estimation procedure: First, the author estimates a production frontier for the transport operators using data envelopment analysis ([Farrell, 1957](#)). In a second step, the determinants indexes of the performance of French urban transit companies are investigated with a Tobit model. The empirical results show that higher efficiency measures are observed more often with fixed-price contracts. [Roy and Yvrande-Billon \(2007\)](#) and [Gautier and Yvrande-Billon \(2013\)](#) use the same type of data but focus instead on a stochastic cost frontier ([Aigner et al., 1977](#); [Meeusen and van Den Broeck, 1977](#)) where some of the deviations from the frontier are attributed to inefficiency, as suggested by [Battese and Coelli \(1995\)](#). Their results suggest that operators under cost-plus contracts exhibit a higher level of technical inefficiency than operators under fixed-price agreements.

[Vigren \(2016\)](#) investigates how common contract types and contract design factors in the Swedish public transport system affect cost efficiency. There are three different types of contract, which entail significant incentives, as in a fixed-price regime: On the one hand, a gross-cost contract without incentives implies that the operator receives a fixed payment based on the supply objectives set by the regulator; in this case, the operator bears only the production “risk”, that is, it is responsible for potential cost-overruns and it obtains a positive profit if the ex post cost is below the expected cost defined ex ante. On the other hand, the net-cost contract implies that the operator bears both industrial and commercial “risk”. In other words, the operator is also incentivized to increase commercial revenue and hence influence the service to a much higher degree in the form of improving service quality and punctuality. Finally, a last—not so frequent—type of contract could be viewed as a mix of the two previous contracts and is called the gross-cost contract with incentives. In this case, the operator still bears the production “risk”, and is levied some revenue “risk”. However, in the latter case, the incentives are lower than in a net-cost contract; the idea is to base some of the operator’s payment on a performance measure in order to get the operator to improve some aspects of the service. The estimation procedure considers again the same stochastic cost frontier as in [Battese and Coelli \(1995\)](#). The inefficiency index depends on several explanatory variables, namely, the population density of a contracting area, the ownership of the transport operator, a measure of the contracting term in numbers of years, and a dummy variable indicating whether the regulatory contract in place entails any incentive payment. An important assumption is that, since transport concessions are awarded through competitive tenders, more incentives in the form of a net-cost contract in place of a gross-cost contract without incentives may entail a higher “risk” for the operator. The latter may therefore require a risk premium in its bid in the form of a higher subsidy. Hence, incentives may lead to two opposite effects: The usual cost-reduction effect, which is the one on which most of the empirical literature focuses, and a cost-increase effect due to a risk premium, as emphasized here. More incentives could therefore lead to higher costs if the risk premium effect is large enough. The estimated results on the effect of the choice of contract are not significantly different from 0, which probably illustrates the fact that the ex ante risk premium impact emphasized by the author is not the unique effect to be expected in this case. The ex post effect, which guarantees that incentive contracts make the transport operator residual claimant for cost overruns and force the manager of the firm to provide a higher effort level to reduce costs, should be present as well. Since both ex ante and ex post effects go in opposite directions, it is not unreasonable that the estimation results show non-significant effect of the type of contract overall.

[Dalen and Gomez-Lobo \(2003\)](#) use the same estimation technique, but work on a slightly different regulatory context: They estimate a cost frontier model for a panel of Norwegian bus companies (1987–1997). Once again, their objective is to investigate to what extent different type of regulatory contracts affects company performance. On the one hand, regional authorities may bargain individually with transport companies on the subsidy levels to be granted during the forthcoming year; this arrangement should then be assumed to be of low power in terms of its incentive properties. On the other hand, a number of regional authorities have

adopted a more high-powered contract based on a yardstick benchmark, which consists in comparing similar regulated firms with each other: The regulator uses the cost of comparable firms to infer a firm's attainable cost target (Shleifer, 1985). This regulatory scheme is again similar in spirit to a fixed-price mechanism. In Norway the regulator and the transport operator agree upon a set of criteria for calculating the costs of operating a bus-network, namely, a cost index based on a linear model that links driver, fuel, and maintenance costs to the number of bus-kilometers produced for different categories of routes. Given fares and route schedules, this index determines the level of transfers that is granted to the operator. The important point is that the same standard cost model applies to all companies within a geographical area and transfers to the operator depend upon the cost performance of a large set of companies. The authors conclude that firms regulated under the yardstick type contract exhibit less than half the cost inefficiency compared to those firms regulated under the traditional contract, but they also acknowledge that contracts choice is endogenous, which complicates the interpretation of the results.

Similar estimation techniques are used by the following authors: Piacenza (2006) analyses a seven-year (1993–1999) balanced panel of 44 Italian municipal companies managed under cost-plus or fixed-price regulatory schemes. The stochastic inefficiency effects are supposed to be linearly related to the regulatory scheme and the commercial speed of the transit companies. The results confirm the initial hypothesis of a lower inefficiency under fixed-price schemes. Karlaftis and Tsamboulas (2012) use data from 15 European transit systems between 1990 and 2000 and test whether different efficiency assessment methodologies produce similar results and whether the impact of the contractual form are robust to the methodological specifications employed. They find that net cost contracts, where the operator bears both production and consumption "risk", involve higher efficiency indexes compared to the gross cost contract case where operators bear only production "risk". Margari et al (2007) focus on 42 Italian transport operators observed from 1993 to 1999. Their approach is based a three-stage methodology which mixes both data envelopment analysis and stochastic frontiers techniques in which they attempt to identify and separate three input-specific determinants of inefficiency, namely exogenous factors, pure managerial inefficiency and statistical noise. The so-called exogenous factors are related to the regulatory context and lie outside the managers' control but are direct decisions of regulators and policy-makers. It is found that the implementation of high-powered incentive contracts enhances production efficiency. The authors conclude that the main source for inefficiency has to be sought in contract choice or even random events, while pure managerial skills only play a minor role. Finally, an exception to all the results presented above is Zhang et al. (2018) which estimates transport efficiency for 26 different Chinese public transport operators in 13 cities across China for the period 2008–2014. Here contrary to previous predictions, the most powered incentive schemes do not entail the highest efficiency measures.

Contract Choice is Endogenous

The aforementioned studies assume that contract choice is exogenous, which is a potential drawback. As suggested for instance by Chiappori and Salanié (2003), contract endogeneity is an important issue since, in the presence of unobserved heterogeneity, the matching of transport operators to contracts is probably not independent from the unobserved heterogeneous firms' efficiency variables: If fixed-price contracts are given to efficient firms while cost-plus regimes are proposed to inefficient ones, the choice of contract is endogenous and any empirical analysis taking contracts as given will be biased. The 2.05% cost reduction emphasized in Gagnepain (1998) may not be caused by the implementation of fixed-price regimes but may illustrate instead a simple selection effect, that is, fixed-price contracts are chosen more often by efficient firms. Before turning to a presentation in the next section of a more structural analysis that allows to embody the endogenous effect in the structure of the cost expression to be estimated, we discuss in what follows a few proposals that attempt to address the endogenous bias in a reduced form.

Iseki (2010) addresses in part the endogeneity issue using data on the US bus transit services between 1992 and 2000. Transit agencies in charge of the organization of the service would typically implement three types of contracted service arrangements: 13% of the agencies contracted out a portion of service, while 21% of agencies contracted out for all service, and 66% contracted out no service. As emphasized by the author, assessing the effect of regulation on operating cost with a regression that combines contracting practices in several dichotomous variables may not be appropriate. The potential endogeneity problem is addressed by applying an instrumental variables technique which consists in predicting each agency's decision making regarding contracting: Two contracting categories are predicted using a multinomial logit model in the first-stage of the two-stage regression analysis which includes a set of explanatory variables related to agencies' revenue and governance characteristics such as labor conditions in the region, regional political climate, and state legislations governing regional and local governments. The results suggest that partial and full-contracting arrangements allow agencies to reduce operating costs.

Díaz and Charles (2016) focuses on the same French urban transport industry as in the previous studies (1995–2010). Efficiency is estimated in a first stage using a data envelopment analysis production frontier. In a second stage, the estimated efficiency indexes are regressed on a set of variables, including contract characteristics and a set of cities' fixed effects. The authors stress the role of unobserved heterogeneity in their analysis: If unobservable factors affect simultaneously both the efficiency and the choice of contracting types, a simple correlation analysis could be misleading. For instance, cities with geographical features that reduce the productivity of inputs might be the most interested in delegating the service to private firms or using certain type of contracts. The introduction of fixed-effects should at least solve the problem as long as the cities' unobserved characteristics are fixed over time, which may not be the case of congestion for instance. Moreover, unobserved characteristics on the operators' side could be relevant as well. Once again, fixed-price contracts are found to perform better than cost-plus contracts in terms of efficiency.

Auction Versus Negotiation

Before choosing a type of contract, such as fixed-price or cost-plus, a regulator must decide whether to organize a competitive tender in order to attract new firms and organize ex ante competition, or whether to negotiate the terms of a new contract with the incumbent. Such a decision has consequences as well on the future costs of the transport operator and the price paid by the regulator. The introduction of auction procedures in many developed countries aims at introducing competition for the field. An increase in the number of bidders should encourage more aggressive bidding, that is, lower bids. In a context of asymmetric information, it is well known from [Laffont and Tirole \(1987\)](#) that an optimal auction, which promotes reduces significantly the rent left to the operator compared to a situation in which a monopoly is regulated under an optimal second-best regime. In this section, we list a series of papers that attempt to provide empirical evaluation of these issues, keeping in mind that once again, regulatory arrangements are generally endogenous decisions, which complicates the analysis of the econometrician.

[Amaral et al. \(2013\)](#) investigate the relationship between the number of bidders and price reduction in the London bus market. They note that, even if competition among bidders leads them to reduce their bid, the winning bidder may, ex post, renege on the initial commitments. From an empirical perspective, this implies that assessing the impact of the number of bidders on price requires controlling for potential contractual renegotiations. The authors use a database on 806 calls for tender on London bus contracts (1999–2008) and find that a higher number of bidders is associated with a lower cost of service. A key issue is the potential endogeneity of the number of bidders, as the expected value of the winning bid potentially influences the number of participating bidders or both variables may be jointly affected by a third variable, such as the value of the service for instance. In order to get rid of potential time-varying heterogeneity, the authors add in the regression the number of auctions organized during the month as well as the number of auctions already won by the winner during the previous month. The estimation results confirm a positive and significant cost-reducing effect of competition.

[Scheffler et al. \(2013\)](#) investigates the impact of the introduction of competitive tendering in German public bus transport networks. Over the period of observation (2004 to 2009), Germany's public bus transport markets remain dominated by publicly owned companies that provide services based mainly on monopoly rights without competitive tendering. In 2009 a large majority of bus operators were fully publicly owned. European regulation (EC) No. 1370/2007, which came into force at the end of 2009 specifies that local authorities have the right to award services directly to transport companies which are under the actual control of the local authority. A direct award is also backed by EU law if the size or value of the transport service contract is small. Hence, German local authorities can generally decide on a discretionary basis as to which services they wish to open up to more competition. The methodology chosen by the authors consists in estimating a production stochastic frontier, as proposed by [Battese and Coelli \(1995\)](#), where the inefficiency term depends on a dummy variable that takes value 1 if the bus operator is chosen under a competitive tendering procedure. Once again, regulatory endogeneity is a potential issue here. Local authorities can generally decide on a discretionary basis as to which services they wish to open up to more competition. A local public might decide to open a route to competition if the publicly owned incumbent is already relatively efficient, and therefore likely to succeed in the tendering process. To address this issue, the authors use as a benchmark the Federal State of Hesse, where local authorities were obliged to put routes out to tender from 2002 to 2007; hence services tendered in this period in this area were not subject to regulatory discretion. Estimation results suggest that there is no significant difference on the influence of technical efficiency between competitive tendering in general and competitive tendering in the Federal State of Hesse, which allows the authors to discard the issue of endogeneity. Bus companies which operate in areas where competitive tendering takes place are more efficient than others. A somehow analysis is provided by [Filippini et al. \(2015\)](#) but the conclusions differ: The authors focus on 630 bus lines operated by the main Swiss company, Swiss Post, in 2009 throughout the country. Following the revision of the Swiss railways act in 1996, regional public authorities were given the choice between two different contractual regimes to procure public passenger transport services, competitive tendering versus negotiation. The paper evaluates the impact of competitive tendering and performance-based negotiation using a stochastic frontier analysis. About 10% of Postbus lines have gone through a competitive tendering process between 2005 and 2015. The estimation results show that the average levels of cost efficiency are relatively high and no significant differences are observed between competitive tendering and performance-based negotiation.

Moreover, [Mouwen and van Ommeren \(2016\)](#) aims at assessing the impact of contract renewal and competitive tendering on operational costs. They employ a panel dataset for the period 2001–2013 on the level of concession areas in the Netherlands. Contract renewal allows several operators to bid for a new contract. More precisely, in the period under study, 61 contract were renewed and 69% of these were renewed after a process of competitive tendering of which about half were awarded to a new operator. The authors estimate a cost function, which accounts for the number of contract renewals between 2001 and the period of observation. They find that contract renewal leads to a substantial reduction in operational costs: When a contract is renewed at least once, costs fall by 10%, while contracts renewed at least twice even lead to an extra costs reduction of 6%.

To conclude this section, note that the endogeneity of the decision of whether a service should be tendered or not also depends on the importance of transaction costs in the industry. [Bajari et al. \(2008\)](#) suggest that auctions perform poorly when projects are complex and contractual design is incomplete. [Hensher and Stanley \(2008\)](#) suggest that there may be high transactions costs of retendering through competitive processes, causing ex post adaptation to become an important feature of the transaction. Properly structured transparent and performance-based negotiated arrangements may avoid this problem and protect the provision of service, while meeting government service objectives. Performance-based negotiation can also avoid some drawbacks of competitive tendering, such as asymmetric information, lack of trust, and hold-up problems between the incumbent and the public authority.

Finally, [Yvrande-Billon \(2006\)](#) also suggests that problems associated with competitive tendering come from the contractual disabilities of the parties.

Structural Models

The treatment of contract choice endogeneity in reduced forms analyses such as the ones presented in the previous section typically require a two-step (sequential) estimation procedure where the inefficiency index for each transport operator is estimated first, and is then regressed on the set of contract characteristics. We turn now to more structural models which intend to embody directly in the structure of the functional form to be estimated the incentive impact of the observed contract impinging on the activity of the transport operator.

We explain briefly how the structural cost expression is constructed, starting with [Gagnepain and Ivaldi \(2002a\)](#). We consider a Cobb–Douglas technology:

$$Y = A \prod_{j=1}^n X_j^{\alpha_j} \exp[e - \theta + \epsilon_Y]. \quad (1)$$

with one output Y and n variable inputs X_j , where A and the α_j 's are parameters describing the technology. To construct the dual cost frontier, we assume that the producer seeks to minimize the cost C of producing its desired rate of output Y under technical inefficiency. The associated allocation of inputs sets the factor demand X_j , $j = 1, \dots, n$. The program of the producer is then as follows

$$\min_{X_j} \sum_{j=1}^n w_j X_j, \quad (2)$$

under the previous technological constraint, where w_j is the price of input j . The excess of factor demand above its frontier prevents the producer from reaching the theoretical level of production. Such a distortion of factor demands leads to a rise of operating costs. In logarithmic form, the associated stochastic cost frontier, $C(Y, w, e, \theta)$, is given by

$$\ln C(Y, w, e, \theta) = K + \sum_{j=1}^n \frac{\alpha_j}{r} \ln w_j + \frac{1}{r} \ln Y + \frac{\theta - e}{r} + \epsilon_c, \quad (3)$$

where $r = \sum_{j=1}^n \alpha_j$ measures returns to scale, and K is a constant. Note that the term $(\theta - e)/r$ is the total cost distortion above the cost frontier. Global inefficiency is smaller when the industry enjoys large returns to scale r . This cost frontier expression is preliminary as the cost reducing effort e is endogenous and depends on regulatory schemes or the competitive environment set by the regulator. We turn now to the cost reducing activity aspect of the problem. Consider a transport operator whose profit function is

$$\pi = R(Y) - C(Y, w, e, \theta) - \psi(e), \quad (4)$$

where $R(Y)$ denotes its revenue. Exerting effort is costly and leads to internal cost $\psi(e)$. A profit-maximizing operator determines the optimal effort level. The first order condition is given by $\psi'(e) = -C_e$, which states that the optimal effort level is chosen to equalize the marginal disutility of effort and the marginal costs savings. Let us define a specific functional form for the internal cost of effort $\psi(e) = \exp(\tau e) - 1$, where $\tau > 0$. The first order condition on the effort can then be expressed using the cost frontier and the internal cost of effort. The level of endogenous effort exerted by the operator is obtained as $e = e(K, w_j, Y, \theta, \tau, \alpha)$, and is then reintroduced in the preliminary frontiers in order to derive the final structural cost frontier to be estimated:

$$\ln C = H_C + \xi \left[\sum_{j=1}^n \frac{\ln w_j}{r} + \frac{1}{r} \ln Y - \frac{\alpha_k}{r} \ln k + \frac{1}{r} \theta \right] + \epsilon_C, \quad (5)$$

where H_C is a constant and $\xi = \tau/(\tau + 1/r)$. If the effort of the producer is nil, the structural cost frontier is given by the expression:

$$\ln C = K + \sum_{j=1}^n \frac{\ln w_j}{r} + \frac{1}{r} \ln Y - \frac{\alpha_k}{r} \ln k + \frac{1}{r} \theta + \epsilon_C, \quad (6)$$

[Gagnepain and Ivaldi \(2002b\)](#) use a database on the French urban transport services described above. The panel data set covers 59 different urban transport networks (operators) over the period 1985–1993. We stressed earlier that two types of contract are observed in practice, cost-plus and fixed-price schemes. The empirical work involves fitting the stochastic cost functions presented in Equation (5) and Equation (6) to this data set. Under fixed-price regimes, we would estimate Equation (5), while we would estimate Equation (6) under cost-plus contract.

The estimation of the system (5) and (6) allows recovering individual inefficiency and effort indexes for each local transport operator. **Table 1** lists the estimated technical inefficiency, effort levels and cost distortions over the frontier for the biggest networks included in the dataset. **Fig. 1** provides for each network, the level of the inefficiency parameter and indicates the type of contract

Table 1 Technical inefficiency, effort level, and cost distortion for some networks

Network	Technical inefficiency	Effort	Distortion
Aix	0.067	0.089	0.990
Besançon	0.318	0.000	1.153
Bordeaux	0.086	0.000	1.039
Caen	0.749	0.103	1.337
Cannes	0.646	0.000	1.337
Clermont	0.155	0.000	1.072
Dijon	0.120	0.000	1.055
Grenoble	0.083	0.114	0.986
Le Havre	0.266	0.000	1.127
Lille	0.180	0.126	1.024
Montpellier	0.131	0.110	1.009
Nantes	0.104	0.117	0.994
Nice	0.489	0.113	1.184
Nîmes	0.035	0.097	0.972
Rennes	0.484	0.000	1.243
Strasbourg	0.806	0.117	1.363
Toulon	0.064	0.000	1.029
Toulouse	0.158	0.124	1.015
Valence	0.111	0.000	1.051

Source: Gagnepain and Ivaldi (2002b)

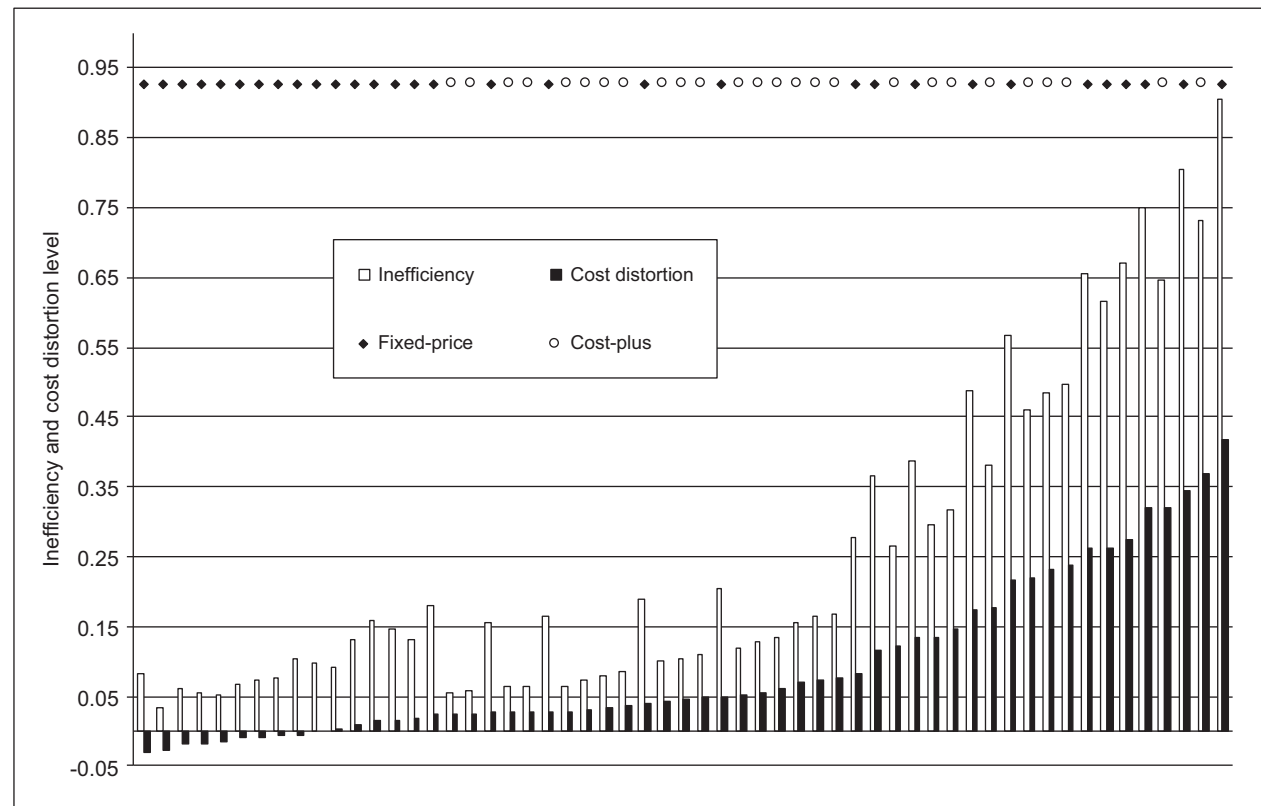


Figure 1 Inefficiency and regulatory schemes. To each network are associated three data: the inefficiency level (white bar), the cost distortion (black bar) and the type of contracts (a black diamond refers to a fixed-price contract and an empty circle indicates a cost-plus contract). Source: Based on Gagnepain and Ivaldi (2002b)

used to regulate it. Note that three groups of networks are easily detected. The first group with the lowest levels of cost distortion gathers sixteen networks, all of which are managed under a fixed-price contract. The next twenty ones can be collected in a second group as all of them (but four networks) are regulated through a cost-plus contract. Finally the last twenty networks are assembled in

a third group, almost equally shared between the two types of contract. Concerning the third group, we just conclude that technical inefficiency is so high that even a highly incentive scheme, such as a fixed-price contract, cannot cure the problem. This structural methodology has been tested successfully in other industries such as, for instance, the Norwegian bus industry (Dalen and Gomez-Lobo, 1997) or the European railroad industry (Urdanoz and Vibes, 2013).

The question of how contracts are chosen and how the information can be included in the estimation process is important and is discussed in Gagnepain and Ivaldi (2017). On the one hand, the choice of contract is determined by its impact on welfare. Welfare, however, is in turn also determined by the contract choice, which implies an endogeneity between the changes in welfare and the contract choice. To address this issue, the authors estimate a structural endogenous switching model where the welfare expression depends on the choice of contract and vice versa. In the model, a regulator chooses the regulatory mechanism (a fixed-price or a cost-plus contract) that maximizes an extended welfare function that entails the usual social welfare measure plus additional weights allocated to specific interest groups. The interest groups are the workers and the stakeholders of the regulated firm. The regulator overstates the weight of one group or another through workers' wages or firms' profits, and this creates a distortion of the regulatory contracts toward less or more powerful incentive schemes. Simultaneously, the regulatory rule affects the operator's behavior and the operating costs in one way or another. It is shown that a simple structural cost function, which does not account for the choice of contract is rejected against the political model advocated in this paper. Hence, accounting for the contract choice turns out to be adequate and fruitful, since it improves the quality of the estimates. The results suggest that ignoring the process of contract choice yields estimates that undervalue the importance of regulatory incentives for the operator's activity.

Gagnepain et al. (2013) extends this analysis into a dynamic setting. In the French urban transport industry, contracts are usually signed for 5 years periods and an extended enough database (1987–2002) allows the econometrician to observe series of contracts in the same local transport network over several periods. Interestingly, a dynamic perspective sheds light on the fact that the power of the incentives of a specific contract and the cost inefficiency type of the transport operator who chooses a specific contract are strongly related. This paper shows that very efficient firms choose series of fixed-price contract over time while very inefficient ones prefer series of cost-plus contracts. In-between, firms that are neither very efficient nor very inefficient choose a cost-plus arrangement first and switch to a fixed-price regime only if subsidies are increased.

Conclusion

This chapter focuses on the assessment of the impact of contractual regimes on the efficiency of public transport operators. The economic literature is almost unanimous on the fact that the incentive power of the contracts used by public authorities in charge of the organization of the service does indeed have a significant effect on the firms' costs. In particular, fixed-price contracts, which entail the highest possible incentive power, and which are very often implemented in the transport industry, are the ones that encourage operators to reduce their operating costs the most.

The issue of the measure of the magnitude of these effects remains, however, quite tricky to deal with since, as we have consistently pointed out in this chapter, a potential issue of selection linked to the choice of contract can create significant biases in the estimation results. The approaches used to address these biases have so far been of two types. First, in studies based on the estimation of reduced forms of cost functions, the authors choose to perform a two-step identification in which efficiency is first measured using a cost or a production frontier, and is then regressed on a set of variables that shed light on the characteristics of the productive environment. Structural studies allow integrating directly in the body of the function to estimate the structure that accounts for the contractual framework and that allows identifying the endogeneity bias.

Using incentive contracts can therefore encourage transport operators to reduce their costs. But it is also important to remember that the optimal nature of these contracts is economically guaranteed if the incentive power of each contract is adjusted to the productive efficiency of the transport operator. In other words, a fixed-price contract must be offered to a very efficient company while a cost-plus contract must be offered to a very inefficient company. Naturally, other types of contracts with intermediate incentive powers could be designed for operators characterized by average efficiency levels. The ability to design these contracts together with a good assessment of the volume of subsidies to be paid to operators (in order to avoid giving those too large rents) depends on the expertise capability of the public authorities. The role of the economist in this context is then to also encourage these authorities to equip themselves with adequate resources in order to implement the sometimes-sophisticated tools proposed by economic research.

References

- Aigner, D., Lovell, C.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. *J. Econom.* 6 (1), 21–37.
- Amaral, M., Saussier, S., Yvrande-Billon, A., 2013. Expected number of bidders and winning bids: Evidence from the London bus tendering model. *JTEP* 47 (1), 17–34.
- Bajari, P., McMillan, R., Tadelis, S., 2008. Auctions versus negotiations in procurement: an empirical analysis. *J. Law Econ. Org.* 25 (2), 372–399.
- Baron, D., Myerson, R., 1982. Regulating a Monopolist with Unknown Costs. *Econometrica* 50, 911–930.
- Battese, G.E., Coelli, T.J., 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Emp. Econ.* 20 (2), 325–332.
- Chiappori, P.-A., Salanié, B., 2003. Testing contract theory: a survey of some recent work. In: Dewatripont, M., Hansen, L., Turnovsky, S. (Eds.), *Advances in Economics and Econometrics*, vol. 1. Cambridge University Press, Cambridge, UK.
- Dalen, D.M., Gomez-Lobo, A., 1997. Estimating cost functions in regulated industries characterized by asymmetric information. *E. Econ. Rev.* 41 (3–5), 935–942.

- Dalen, D.M., Gomez-Lobo, A., 2003. Yardsticks on the road: regulatory contracts and cost efficiency in the Norwegian bus industry. *Transportation* 30 (4), 371–386.
- Diaz, G., Charles, V., 2016. Regulatory design and technical efficiency: public transport in France. *J. Reg. Econ.* 50 (3), 328–350.
- Farrell, M.J., 1957. The measurement of productive efficiency. *J. Royal Stat. Soc. Series A* 120 (3), 253–281.
- Filippini, M., Koller, M., Masiero, G., 2015. Competitive tendering versus performance-based negotiation in Swiss public transport. *Transport. Res. A Pol. Pract.* 82, 158–168.
- Gagnepain, P., 1998. Structures productives de l'industrie du transport urbain et effets des schémas réglementaires. *Économie et Prévision* 135 (4), 95–107.
- Gagnepain, P., Ivaldi, M., 2002a. Stochastic frontiers and asymmetric information models. *J. Product. Anal.* 18 (2), 145–159.
- Gagnepain, P., Ivaldi, M., 2002b. Incentive regulatory policies: the case of public transit systems in France. *RAND J. Econ.* 33 (4), 605–629.
- Gagnepain, P., Ivaldi, M., Martimort, D., 2013. The cost of contract renegotiation: Evidence from the local public sector. *Am. Econ. Rev.* 103 (6), 2352–2383.
- Gagnepain, P., Ivaldi, M., 2017. Economic efficiency and political capture in public service contracts. *J. Indust. Econ.* 65 (1), 1–38.
- Gautier, A., Yvrande-Billon, A., 2013. Contract renewal as an incentive device. An application to the French urban public transport sector. *Rev. Econ. Institut.* 4 (1), 29.
- Hensher, D.A., Stanley, J., 2008. Transacting under a performance-based contract: The role of negotiation and competitive tendering. *Transport. Res. A Pol. Pract.* 42 (9), 1143–1151.
- Iseki, H., 2010. Effects of contracting on cost efficiency in US fixed-route bus transit service. *Transport. Res. A Policy Pract.* 44 (7), 457–472.
- Karlattis, M.G., Tsamboulas, D., 2012. Efficiency measurement in public transport: are findings specification sensitive? *Transport. Res. A Policy Pract.* 46 (2), 392–402.
- Kerstens, K., 1996. Technical efficiency measurement and explanation of French urban transit companies. *Trans. Res. A Policy Pract.* 30 (6), 431–452.
- Laffont, J.J., Tirole, J., 1986. Using cost observation to regulate firms. *J. Polit. Econ.* 94 (3, Part 1), 614–641.
- Laffont, J.J., Tirole, J., 1987. Auctioning incentive contracts. *J. Polit. Econ.* 95 (5), 921–937.
- Leibenstein, H., 1966. Allocative efficiency vs “X-efficiency”. *Am. Econ. Rev.* 56 (3), 392–415.
- Loeb, M., Magat, W., 1979. A decentralized method of utility regulation. *J. Law Econ.* 22, 399–404.
- Margari, B.B., Erbetta, F., Petraglia, C., Piacenza, M., 2007. Regulatory and environmental effects on public transit efficiency: a mixed DEA-SFA approach. *J. Regul. Econ.* 32 (2), 131–151.
- Meeusen, W., van Den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *Internat. Econ. Rev.* 435–444.
- Mouwens, A., van Ommeren, J., 2016. The effect of contract renewal and competitive tendering on public transport costs, subsidies and ridership. *Trans. Res. A Policy Pract.* 87, 78–89.
- Piacenza, M., 2006. Regulatory contracts and cost efficiency: Stochastic frontier evidence from the Italian local public transport. *J. Product. Anal.* 25 (3), 257–277.
- Roy, W., Yvrande-Billon, A., 2007. Ownership, contractual practices and technical efficiency: The case of urban public transport in France. *JTEP* 41 (2), 257–282.
- Scheffler, R., Hartwig, K.H., Malina, R., 2013. The effects of ownership structure, competition, and cross-subsidisation on the efficiency of public bus transport: Empirical evidence from Germany. *JTEP* 47 (3), 371–386.
- Shleifer, A., 1985. A theory of yardstick competition. *RAND J. Econ.* 319–327.
- Urdániz, M., Vibes, C., 2013. Regulation and cost efficiency in the European railways industry. *J. Prod. Anal.* 39 (3), 217–230.
- Vigren, A., 2016. Cost efficiency in Swedish public transport. *Res. Transp. Econ.* 59, 123–132.
- Yvrande-Billon, A., 2006. The attribution process of delegation contracts in the French urban public transport sector: why competitive tendering is a myth. *Annal. Public Cooperat. Econ.* 77 (4), 453–478.
- Zhang, C., Xiao, G., Liu, Y., Yu, F., 2018. The relationship between organizational forms and the comprehensive effectiveness for public transport services in China? *Trans. Res. A Policy Pract.* 118, 783–802.

Company Cars

Stefan Gössling, Department of Service Management and Service Studies, Lund University, Lund, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	580
Company Car Benefits	580
Observed Changes in Transport Behavior	581
Environmental Outcomes	582
Conclusions	582
References	583

Introduction

Company cars, sometimes also called “benefit cars” or “take-home cars” (Engström et al., 2019), now represent a significant share of national car fleets. Cohen-Blankshtain (2008) suggests that company cars account for 10%–50% of car fleets in different countries. Company cars are usually justified as a necessary means for conducting a job requiring travel, for which a vehicle needs to be made available. More regularly, reasons for company car provisions may be tax benefits that exceed the cost of an equivalent monetary remuneration. As only a share of the taxable benefit of a company car is actually taxed, this is a financially beneficial model for companies, company car recipients, as well as car manufacturers. Company cars are a form of subsidy forwarded to consumers and producers in the automotive system.

Even though there is a considerable body of literature on company cars, it is notable that no discussion of the history of company cars is available. It seems that company car benefits were initially forwarded only to individuals at the top of companies as a form of additional financial benefit: the share of income spent on the car is significant throughout the world. Today, company car benefits are mainstreamed, with estimates that 50% of all new car registrations in the European Union are company cars (Nss-Schmidt and Winiarczyk, 2009).

Company Car Benefits

Company cars represent a subsidy forwarded to employees and company owners, because they are often used privately, or by other household members, without being fully taxed in theory and/or practice (Harding, 2014). For this reason, the fiscal and environmental implications, such as foregone tax revenue or environmental impacts, have been featured prominently in a wide range of publications (Dargay et al., 2007) and official documents by the European Union, OECD or other supranational organizations (Gutiérrez-i-Puigarnau and van Ommeren, 2011; Princen, 2017; Ramaekers et al., 2010; Shiftan et al., 2012; Van Ommeren et al., 2006; Graus and Worrell, 2008).

The legal and fiscal basis for the provision of company car benefits is that work-related travel is tax-deductible (Cohen-Blankshtain, 2008). As private trips are not related to income generation, they represent a personal benefit where they remain partially untaxed (Harding, 2014). Even though most countries do have rules regarding the private use of company cars, there is much evidence that a significant share of the cost associated with private travel is covered by employers. This would include items such as the cost of car registration, insurance, maintenance, and repairs, which private car owners have to pay for themselves; as well as fuel and operating expenses, with evidence that where these aspects are taxed, this only represents a share of the value in question (Harding, 2014).

Table 1 provides a typology of tax rules regarding the private use of company cars in OECD countries (Harding, 2014). Such taxation principles have followed a wide range of different approaches. For instance, in Spain or Switzerland, taxation is based on the consideration of the percentage of capital cost compared to cost price; in Belgium or Norway, in comparison to list price; in the United States, in comparison to fair market value. Other countries including Canada or Luxembourg consider the distances traveled for private purposes, or they consider the direct cost that may be private (Australia, Japan) vis-à-vis business related (Austria, South Africa). Finland and Sweden have lump sum approaches; Hungary and Mexico do not tax company cars at all.

Notably, commuting to work is considered a tax-deductible expense only in some countries, and even where approaches to taxing company car benefits are identical, there are often considerable differences. For example, in Germany, the recipient of a company car has to tax 1% of the list price of the car (Metzler et al., 2019), whereas in Israel, the list price tax is more than twice as high (Shiftan et al., 2012). An important implication of most taxation approaches is that the higher the car’s value, or the more the car is used, the greater is the benefit to the employee (Diekmann et al., 2011; Metzler et al., 2019).

There is a general consensus in the literature that company cars represent very significant subsidies forwarded to employees. For example, Diekmann et al. (2011) estimate that the untaxed company car benefit in Germany is worth an annual €3.3–5.5 billion. A

Table 1 Typology of tax treatments of personal use of company cars, OECD

% of capital cost			Distance			Direct cost			
Cost price	List price	Fair market value	Private	Deemed	Homework	Private	Business	Lump sum	Not taxed
Australia	Belgium	United	Canada	Italy	Germany	Australia	Austria	Estonia	Hungary
Austria	Denmark	States	Estonia			France	South Africa	Finland	Mexico
Canada	Finland		Finland			Germany		Sweden	
France	Germany		Luxembourg			Japan			
Luxembourg	Iceland		Sweden			Poland			
New Zealand	Netherlands		United States						
Portugal	Norway								
Slovakia	Sweden								
Slovenia	United								
South Africa	Kingdom								
Spain									
Switzerland									

Source: Based on [Harding \(2014\)](#)

study in Belgium the value of lost tax revenue at €3.75 billion per year, or 0.9% of the country's GDP ([Princen, 2017](#)). Another calculation for the Netherlands estimated the welfare cost of company car taxation at €2000 per car ([Gutiérrez-i-Puigarnau and van Ommeren, 2011](#)). An estimate for the European Union ([Ness-Schmidt and Winiarczyk, 2009](#)) concludes that undertaxation is a norm throughout the region, with direct revenue losses in the order of €54 billion per year, and distortions of consumer choices adding annual losses of €12 billion to €37 billion. A study of 26 OECD countries, [Harding \(2014\)](#) concluded that untaxed benefits were equivalent to a weighted average subsidy per company car in the order of €1600 per year, corresponding to an estimated €26.8 billion. Only 44%–58% of the benefits forwarded to employees are taxed in the OECD. There is thus overwhelming evidence that company car benefits represent a very significant subsidy, and on a global scale.

Observed Changes in Transport Behavior

Company car benefits change travel behavior and wider choices pertaining to the car. Most of these changes are linked to real or perceived benefits associated with company cars, and can be summarized as being related to transport mode preferences, car model choices, car availability and endowment, travel behavior, driving styles, and interest in alternative fuel technology. These changes and their implications are presented in [Table 2](#).

As an example for transport mode choice, various authors have outlined that company car availability automatically turns the car into the self-evident transport choice, to the detriment of other transport modes such as bicycle, bus, tram, or train ([Dargay et al., 2007](#); [Harding, 2014](#); [Ramaekers et al., 2010](#); [Shiftan et al., 2012](#)). Various studies have also found that company cars are regularly larger and more fuel consuming than private cars ([European Commission, 2002](#); [Graus and Worrell, 2008](#)). In light of

Table 2 Overview of changes in transport behavior induced by company car

Aspect	Change	Implication
Transport mode	Car use vis-a-vis modal split (bicycle, bus, tram, train)	Car use becomes habit and sole transport mode
Car model choice	Larger cars are provided than would have been bought by employee privately	Car size and motorization increase, as well as specific fuel use
Car age and efficiency	Company cars are newer and thus less energy consuming	Company cars change fleet age positively, are more efficient
Car availability	Other family members use company car	Car availability and perceived low cost generate additional transport demand
Car endowment	Second private car becomes affordable	Two cars increase opportunities for automobility
Travel behavior	Car use increases as a result of lower (perceived) cost	Transport demand increases
Driving style	Driving style less energy efficient as a result of lower (perceived) cost	Specific fuel use increases
Fuel use	Significantly more fuel is used by company car drivers	Overall fuel use increases
Interest in alternative fuel technology	Powerful cars with combustion engine represent norm	Attachment to cars with combustion engines grows, while other technologies become less attractive

recent research, fuel use is a more complex issue: for example, data for Germany suggest that company cars are significantly newer (4.4 years) than private cars (7.6 years) (Metzler et al., 2019). Even though company cars are more heavily motorized, at 97 kW (compared to 79 kW for private cars), fuel use was found to be almost identical to private car consumption, at 7.77 L/100km (compared to 7.80 L/100km for private cars). It is unclear, however, if fleet average fuel consumption would be lower if the fleet age was higher, yet consist of smaller and more efficient cars; this could be the case in a scenario where company car benefits did not exist.

Company cars also affect household car ownership patterns. In a study of Israeli company car owners, 64% had two cars; that is, more than the average household, suggesting that company cars are considered a saving that makes an investment in a second private car more likely (Shiftan et al., 2012). This was statistically tested in a German study (Metzler et al., 2019). Results suggest that company car households have an average of 2.0 cars, while households with private cars own an average of 1.6 cars. The study concluded that company car benefits increased household car ownership by 25%.

Travel behavior is one of the best studied variables in the context of company car ownership, and it has been suggested for decades that company cars will make employees more willing to accept longer commutes (van Ommeren et al., 2006). Quantitative studies have shown company cars to be driven 100%–200% more than private cars (Graus and Worrell, 2008; Johansson-Stenman, 2002; Ramaekers et al., 2010). Metzler et al. (2019) showed that company cars in Germany are driven an average 24,672 km per year, compared to 12,828 km per year for private vehicles. Only part of these greater distances is associated with work (Shiftan et al., 2012), as company car ownership stimulates the transport demand of other household members and increases the likelihood of leisure trips (Shiftan et al., 2012).

As perceptions of fuel use being either free or at least less costly, company car drivers have also been found to have more wasteful driving styles, characterized, for example, by idling, higher speeds, hard acceleration, or deceleration (Rutty et al., 2013). Driving styles, and in particular the greater distances traveled, translate into greater fuel use. Owners of company cars in a study in German consumed an average 1170 L of gasoline or 2361 L diesel per year, depending on the respective cars' engine type. Corresponding values for private car owners were significantly lower at 830 L (gasoline) and 1157 L (diesel) per year (Metzler et al., 2019).

Finally, company cars also have implications for the interest in alternative fuel technologies, such as electric engines or fuel cells, though outcomes are partially contradictory. Research in the Netherlands (Koetse and Hoen, 2014) found that company car drivers preferred conventional combustion engines, as a result of more limited driving ranges, recharge times, and refueling infrastructure limitations. Hybrid and flexifuel cars were perceived as more suitable for company car driver needs, and electric and fuel cell cars were more commonly bought by drivers with low mileage. Engström et al. (2019) found that company car drivers did prefer alternative fuels to combustion engines, though these choices also resembled larger, more heavily motorized and expensive cars.

Environmental Outcomes

As has been repeatedly highlighted in the literature, the existence of company car benefits has a wide range of social, economic, and environmental implications (Whitelegg, 1984). Company cars are often driven by high-income takers (Börjesson and Kristoffersson, 2018), and tax benefits thus mostly accrue to the upper income classes. This is particularly relevant in the context of climate change and emissions from transportation, as there is growing evidence that there are very significant differences in individual contributions to climate change. As highlighted in the preceding section, company cars initiate and support behavioral change and transport choices that contradict societal goals regarding accidents, space requirements for transportation, noise, air pollution, as well as climate change mitigation.

Climate change stabilization is the most important international environmental goal. The Paris Agreement focuses on stabilizing temperatures at a maximum increase of 2°C, compared to preindustrial levels, and foreseeing steep cuts in emissions. Transportation contributes about 23% to total global energy-related CO₂ emissions, and almost three quarters of transport emissions (72.1%) are caused by road transport (IPCC, 2014). How frequently vehicles are used, and which energy intensity these vehicles have, is thus of central importance for emission outcomes. Few studies have investigated the impact of company car benefits for all changes in vehicle choices and travel behavior. In one study for Germany, Metzler et al. (2019) found that private car owners emitted 2.0 t to 3.0 t CO₂ per year, while the corresponding amount was 2.8 t CO₂ to 6.2 t CO₂ for company car drivers. A significant share of the difference can be assumed to be a result of company car benefits, though it is difficult to determine the exact difference.

Conclusions

Company cars represent a form of subsidy forwarded to drivers and companies, who may seek to compensate employees indirectly with car-related benefits. Company cars are also a subsidy to car manufacturers, as they increase the number of cars per household. As specific car types are in higher demand - more powerful, larger, and expensive cars -, company car benefits also support the development of specific car model markets. These concerns over company cars have been raised for decades, with authors (Börjesson and Kristoffersson, 2018; Metzler et al., 2019) and organizations calling for reform (Harding, 2014).

There is no shortage of specific recommendations regarding the design of such company car tax reforms. Proposals have included limiting company car numbers (Cohen-Blankshtain, 2008), as well as changes in taxation, specifically with regard to private use (Shiftan et al., 2012). In practice, none of these changes will lead to a situation in which foregone tax revenues are recaptured, and they are unlikely to prevent other negative outcomes of tax benefits. In most countries, there is a simultaneous lack of equal

treatment of company cars with forms of micromobility; this is, while (private) automobility is deductible, it is impossible to deduct taxes for walking, carpooling or bicycling. This is relevant, because automobility represents a social cost, while bicycling or walking constitutes a benefit (Harding, 2014). It has thus been suggested that bicycling should be rewarded economically, while commuting by car should not be tax deductible at all (Laine and van Steenberg, 2017).

It is evident that transport demand can be reduced by eliminating the preferential tax treatment of company cars (Börjesson and Kristoffersson, 2018; Metzler et al., 2019). In view of the changes in car model choice and transport behavior, and the limited effects of company car purchases in terms of introducing alternative fuels, it seems difficult to see company cars in any other way than an overcame tax relief forwarded to car manufacturers, companies and drivers. All studies on company car externalities conclude that such tax reliefs are no longer timely. Even though no estimate of the global scale of the company car tax relief has been published, it is fair to assume that it is significantly higher than the €66–€91 billion calculated by Nass-Schmidt and Winiarczyk (2009) for the European Union.

As any change in company car benefit structures will affect many people, and with significant monetary losses, resistance to policy change will be significant. Notably, company car benefits are perceived as more attractive than additional pension payments, life insurances or savings plans (Macharis and Witte, 2012). However, in light of the negative externalities incurred by company cars, such change is paramount. Abolishing company car policies will lead to changes in transport systems that would include greater interest in alternative transport modes, smaller cars, and new mobilities. These are associated with further reductions in negative externalities, such as the very significant health cost of air pollution (Lancet Commission, 2017).

Environmentally, an elimination of company car benefits also has importance for climate change mitigation. Where CO₂-based vehicle taxes or the taxation of emissions are combined with the abolishment of company car benefits (Dineen et al., 2018), these two measures alone could bring EU (and worldwide) passenger transportation back on track to achieving the climate mitigation goals as outlined in the Paris Agreement. Currently, there is very limited evidence of climate policies for road transport that will change the status quo. Overall, findings thus suggest that irrespective of perspective—social welfare, environment, and even business economics—there is thus no good reason to maintain company car tax benefits.

References

- Börjesson, M., Kristoffersson, I., 2018. The Swedish congestion charges: ten years on. *Transport. Res. A Policy Pract.* 107, 35–51.
- Cohen-Blankshtain, G., 2008. Institutional constraints on transport policymaking: the case of company cars in Israel. *Transportation* 35 (3), 411–424.
- Dargay, J., Gately, D., Sommer, M., 2007. Vehicle ownership and income growth, worldwide: 1960–2030. *Ener. J.* 28 (4), 143–170.
- Diekmann, L., Klinski, S., Schmidt, S., Gerhards, E., Meyer, B., Thöne, M., 2011. *Steuerliche Behandlung von Firmenwagen in Deutschland*.ifo Institute for Public Economics, University of Cologne, Germany.
- Dineen, D., Ryan, L., Ó Gallachóir, B., 2018. Vehicle tax policies and new passenger car CO₂ performance in EU member states. *Clim. Policy* 18 (4), 396–412.
- Engström, E., Algers, S., Hugosson, M.B., 2019. The choice of new private and benefit cars vs. climate and transportation policy in Sweden. *Transport. Res. D Trans. Environ.* 69, 276–292.
- European Commission, 2002. Fiscal measures to reduce CO₂ emissions from new passenger cars: main report. https://ec.europa.eu/taxation_customs/sites/taxation/files/docs/body/co2_cars_study_25-02-2002.pdf.
- Graus, W., Worrell, E., 2008. The principal–agent problem and transport energy use: case study of company lease cars in the Netherlands. *Ener. Policy* 36 (10), 3745–3753.
- Gutiérrez-i-Puigarnau, E., Van Ommeren, J.N., 2011. Welfare effects of distortionary fringe benefits taxation: the case of employer-provided cars. *Inte. Econ. Rev.* 52 (4), 1105–1122.
- Harding, M., 2014. Personal tax treatment of company cars and commuting expenses: estimating the fiscal and environmental costs, OECD Taxation Working Papers, No. 20, OECD Publishing. <http://dx.doi.org/10.1787/5f214c61s7vi-en>. (accessed 1 May 2018).
- IPCC, 2014. Summary for policymakers. In: Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Farahani, E., Kadner, S., Seyboth, K., Adler, A., Baum, I., Brunner, S., Eickemeier, P., Kriemann, B., Savolainen, J., Schlömer, S., von Stechow, C., Zwickel, T., Minx, J.C. (Eds.), *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge.
- Johansson-Stenman, O., 2002. Estimating individual driving distance by car and public transport in Sweden. *Appl. Econ.* 34, 959–967.
- Koetse, M.J., Hoen, A., 2014. Preferences for alternative fuel vehicles of company car drivers. *Res. Ener. Econ.* 37, 279–301.
- Laine, B., Van Steenberg, A., 2017. Commuting subsidies in Belgium. *Reflets et perspectives de la vie économique* (2), 101–120.
- Landrigan, P.J., Fuller, R., Acosta, N.J., Adeyi, O., Arnold, R., Baldé, A.B., et al., 2018. The Lancet Commission on pollution and health. *The Lancet* 391 (10119), 462–512.
- Macharis, C., De Witte, A., 2012. The typical company car user does not exist: the case of Flemish company car drivers. *Trans. Policy* 24, 91–98.
- Metzler, D., Humpe, A., Gössling, S., 2019. Is it time to abolish company car benefits? An analysis of transport behaviour in Germany and implications for climate change. *Clim. Policy* 19 (5), 542–555.
- Nass-Schmidt, S., Winiarczyk, M., 2009. Company car taxation. subsidies, welfare and environment. European Commission. https://ec.europa.eu/taxation_customs/sites/taxation/files/resources/documents/taxation/gen_info/economic_analysis/tax_papers/taxation_paper_22_en.pdf (accessed 15 October 2019).
- Princen, S., 2017. Taxation of company cars in Belgium – room to reduce their favourable treatment. https://ec.europa.eu/info/publications/economic-and-financial-affairs-publications_en. (accessed 13 April 2018).
- Ramaekers, K., Wets, G., De Witte, A., Macharis, C., Cornelis, E., Castaigne, M., Pauly, X., 2010. The impact of company cars on travel behaviour. In: 12th World Conference on Transport Research (WCTR), July 11–15, 2010, Lisbon, Portugal.
- Rutty, M., Matthews, L., Andrey, J., Del Matto, T., 2013. Eco-driver training within the City of Calgary's municipal fleet: monitoring the impact. *Transport. Res. D* 24, 44–51.
- Shifan, Y., Albert, G., Keinan, T., 2012. The impact of company-car taxation policy on travel behavior. *Trans. Policy* 19 (1), 139–146.
- Van Ommeren, J., Van Der Vlist, A., Nijkamp, P., 2006. Transport-related fringe benefits: implications for moving and the journey to work. *J. Reg. Sci.* 46 (3), 493–506.
- Whitelegg, J., 1984. The company car in the United Kingdom as an instrument of transport policy. *Transport. Policy Dec. Mak.* 2, 219–230.

Transportation Network Companies (TNCs) and the Future of Public Transportation

Susan Shaheen, Adam Cohen, University of California, Berkeley, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

What are Transportation Network Companies (TNCs)?	584
TNC Impacts on Travel, Labor, and Emissions	584
TNCs and Curbside Management	585
TNCs and Social Equity	585
Future of TNCs and Vehicle Automation	586
Possible Impacts of Shared Automated Vehicles	586
Relationship Between SAVs and Public Transportation	587
Summary	587
Acknowledgment	587
Biographies	587
See Also	588
References	588
Further Reading	588

What are Transportation Network Companies (TNCs)?

Transportation network companies (TNCs) (also known as ridesourcing and ridehailing) provide prearranged and on-demand transportation services for compensation, which connect drivers of personal vehicles with passengers (Shaheen and Cohen, 2019). Smartphone applications are used for booking, ratings (for both drivers and passengers), and electronic payment. TNCs can provide both nonpooled (e.g., single-passenger) and pooled services (e.g., multiple passengers), which are referred to as ridesplitting. Ridesplitting is a type of TNC service that involves volunteering to share a for-hire ride with someone at a reduced cost (Shaheen and Cohen, 2019). Volunteering to split a ride typically includes a discount and may or may not result in a shared ride depending on demand and the origin/destination match suitability. TNCs differ from carpooling and vanpooling (also known as ridesharing) in its financial motivation; carpooling is not intended to result in a financial gain. TNCs also differ from taxicab services as taxis are permitted to pick-up street hails, where TNCs are not permitted to do so in most jurisdictions. Additionally, while taxis are often regulated to charge static fares, TNCs often use variable market-rate pricing, which can also employ “surge pricing” in which prices go up during periods of high demand to incentivize more drivers to take ride requests. Additionally, TNCs typically provide the traveler with an established fare at the beginning of a trip whereas taxis can employ either fixed pricing (e.g., an agreed upon fare between the passenger and the driver, zone-based pricing, etc.) or variable pricing (e.g., meters) where the actual cost of a journey is unknown to the passenger until the trip is complete. A common critique of metered pricing is that it can be susceptible to driver manipulation, such as “long hauling” (where a driver takes a more circuitous route to increase the fare) or “zapping” (an attachment to a taxi meter or its wiring system that illegally boosts mileage and customers’ fares during trips). As such, for-hire services that provide consumers with an established price when a traveler commences their journey, such as TNCs, taxi e-Hail, and zone-based pricing tend to offer consumers greater transparency and protection against fare manipulation. Services provided by TNCs vary in terms of the number of drivers available on the platform, geographic span, prices charged customers, and available services. Some may provide specialized services for particular market segments (e.g., services for children, older adults, people with disabilities, etc.). In addition to providing traditional for-hire services that connect customers requesting rides to drivers of cars, some TNC companies also provide bikesharing, scooter sharing, and courier network services or flexible goods delivery on their platforms. This chapter focuses on the impacts of for-hire passenger services offered by TNCs.

In recent years, TNCs have grown rapidly due to technological advancements, changing mobility patterns, and evolving perspectives toward transportation and vehicle ownership. TNCs may be able help bridge gaps in the transportation network (e.g., facilitating first- and last-mile connections to public transit, offering late-night transportation, and services scaled for lower-density communities). As such, TNCs may be able to help bridge gaps in the public transportation network by extending geographic coverage and the operating times of public transit services. Additionally, more modal options and a larger pool of travelers can create a “network effect,” where mobility options in close proximity to one another add collective value. This network effect may also be able to support pooled TNC services, sometimes referred to as ridesplitting or pooling. Ridesplitting involves a person sharing a TNC vehicle and splitting the cost of a ride with someone else taking a similar route.

TNC Impacts on Travel, Labor, and Emissions

Studies on the impacts of TNCs are emerging and have been shown to have varied impacts on vehicle miles traveled/vehicle kilometers traveled (VMT/VKT), modal choice, automobile ownership and use, and public transit ridership, which are typically

impacted by local characteristics such as: urban density, the built environment, public transit accessibility, public policy, and other factors (Aleml et al., 2018; Brown and Taylor, 2018; Clewlow and Mishra, 2017; Feigon and Murphy, 2018; Hampshire et al., 2017; Henaio, 2017; Martin et al., 2020). TNCs can have some of the following impacts:

- Increasing access and mobility for nonvehicle owners;
- Increasing for-hire vehicle service availability, particularly in the evening and on weekends and in smaller markets where taxi service is limited or unavailable; and
- Affecting labor in various ways, such as increased employment opportunities and uncertain financial impacts on drivers (e.g., opportunities for new drivers, downward wage pressure on existing taxi drivers, tax implications with respect to employee versus independent contractor relationships, and worker benefits).

TNC impacts on vehicle trips and occupancy, VMT/VKT, greenhouse gas (GHG) emissions, and other transportation modes can vary as well. TNCs produce VMT/VKT and associated GHG emissions when driving to an area with passenger demand, deadheading while awaiting a ride request, heading to pick up a passenger, and completing the ride itself. TNCs can also reduce VMT/VKT and GHGs through behavioral change, such as passengers who decide they no longer need to own a car due to TNC availability. TNCs can also have substitutive effects on existing transportation modes, such as changes in active and public transportation use. A three-city study of TNCs in Los Angeles, San Francisco, and Washington, DC found that VMT produced by TNCs was larger than the VMT reductions that occurred due to passenger behavior and vehicle ownership change in Los Angeles and San Francisco. However, in Washington, DC, the study found that the balance of impacts resulted in a net VMT and GHG reduction, which is likely due to land use and the built environment, including public transit availability and urban densities. The study also found that pooled TNC services can mitigate TNC VMT and emissions; however, this impact is highly sensitive to match rates and mode substitution (Martin et al., 2020). Pooled and privately used TNCs may be drawing from a slightly different cross-section of travelers, with pooled TNC services drawing more heavily from public transit and private TNC services. It is important to note that single passenger (not pooled) TNC services are more likely to substitute passenger vehicle modes (e.g., a personal vehicle, taxi/E-hail taxi, or carsharing vehicle). Research on TNC impacts varies based on a variety of factors such as: the built environment, public transit accessibility, urban density, and other city-specific factors. TNC studies tend to suggest that they may be drawing more people from public transit in denser cities, a phenomenon sometimes referred to as peak shedding. In less dense cities, studies indicate that TNCs more often replace private automobile trips. It is important to note that aggregated cross-city studies run the risk of obscuring city-specific differences in TNC impacts. In addition, if respondents are asked what transportation mode they would generally take in contrast to TNCs (instead of what mode they would have used for their last TNC trip), responses may be less representative of a respondent's mode replacement decision.

TNCs and Curbside Management

TNCs, depending on passenger demand and locations served, may require frequent loading and unloading of passengers. When there are not designated pick-up and drop-off locations, vehicles may stop in the roadway, shoulders, on-street parking, bike lanes, and public transit stops, which can lead to congestion, modal conflicts, and safety hazards. A number of public agencies are developing policies to establish loading zones for shared modes, such as TNC vehicles. For example, Washington, DC has established nightlife zones designated for TNC passenger pick-up and drop-off between 10 p.m. and 7 a.m., Thursday night through Sunday morning.

TNCs and Social Equity

Similar to other forms of shared mobility, TNCs can be associated with social equity challenges. Many TNC studies have examined who uses the service, when, and for what purposes. Research suggests that TNC users tend to be younger, of higher income, and with higher levels of educational attainment than the general population. TNC passengers also tend to have lower vehicle ownership rates, commute using public transit at a higher rate, and are less reliant on single-occupant vehicles for commuting. A number of studies have documented higher TNC usage on evenings, weekends, and during weekday peak periods (e.g., 7 to 11 a.m. and 5 to 9 p.m.). Restaurant/bar, social/recreational, and commuting are commonly indicated trip types.

A number of studies have suggested that TNC drivers may engage in unauthorized and prejudicial behavior such as: driving female passengers on longer than needed trips between origins and destinations (e.g., to engage in conversations to get a passenger's personal information, such as a telephone number or workplace, and/or to make unwanted verbal or physical sexual advances) and rejecting or canceling riders with minority sounding names. Six common transportation equity challenges include:

1. Affordability: "It's too expensive."
2. Predictability: "Will dynamic or surge pricing make it too expensive?"
3. Availability: "The services aren't available in my neighborhood."
4. Payability: "I don't have an acceptable payment method."
5. Accessibility: "The service isn't accessible for my medical condition."
6. Techno-ability: "I don't have a smartphone or a data plan."

Generally, these challenges can be categorized into four overarching themes:

1. **Access for People with Disabilities:** TNCs may not always provide wheelchair accessible services or be available in all markets.
2. **Un- and Under-Banked Households:** TNCs generally require debit/credit cards for payment. This can be a barrier for consumers who are under-banked or unbanked. TNCs may be able to provide alternative fare payment options to help overcome this.
3. **Low-Income Affordability and Service Equivalency:** TNC pricing can be expensive compared to other modes, particularly during periods of high demand and low supply (known as surge pricing).
4. **Digital Poverty:** TNCs typically require a smartphone and high-speed data package to access services. This can be a barrier to low-income and rural households who may not be able to afford or may lack data coverage to access TNCs. Alternatives such as digital kiosks, telephone services, and nontech access (such as street hail—that are generally prohibited by TNCs in most jurisdictions) can help overcome these challenges.

Future of TNCs and Vehicle Automation

Many experts predict the convergence of shared modes; mobile services (e.g., smartphones and wireless data); electrification; and automation. This convergence will further transform taxis, TNCs, and pooled service options. In particular, vehicle automation will be one of the most transformative developments in automotive history, as it is deployed over the coming decades. SAE International, a global mobility standards organization, has established five levels of vehicle automation. Level 1 describes vehicles that automate only one primary control function (e.g., self-parking or adaptive cruise control). Level 2 describes a vehicle with automated systems that have full control of specific vehicle functions such as, accelerating, braking, and steering. With Level 2, the driver must still monitor driving and be prepared to immediately resume control at any time. Level 3 allows the driver to engage in nondriving tasks for a limited time. With Level 3, the vehicle will handle situations requiring an immediate response; however, the driver must still be prepared to intervene within a limited amount of time when prompted. With Level 4, a human operator does not need to control the vehicle as long as the vehicle is operating within the specific conditions the vehicle was intended to function. Level 5 describes vehicles that are capable of driving in all environments without human control. By automating driving tasks, shared, automated, electric mobility services could become much more cost effective, efficient, and convenient than human-driven, privately owned vehicles. These studies predict vehicle automation and electrification, coupled with pooling, will result in shared automated vehicles (SAVs) that offer for-hire services that are less expensive on a per mile basis than privately owned vehicles (Greenblatt and Saxena, 2015). These studies suggest notable cost savings in three key areas: (1) labor cost savings associated with no longer needing a driver; (2) energy cost savings associated with electrification; and (3) cost savings associated with pooled rides (e.g., sharing fares among passengers with similar origins and destinations). Operationally, the cost savings coupled with the convenience of SAVs, pickup, drop-off, self-parking, and self-charging could be a driving force toward reducing private-vehicle ownership (particularly if coupled with curb management, street design, and pricing policies).

Many companies are beginning to explore the concept of shared and fully automated fleets. Cities and public agencies have recently begun to examine the possibility of managing SAV services. While it is too early to predict the range of service types that may exist as part of a SAV ecosystem, new vehicle types and services will develop over time. SAVs have the potential to reduce vehicle ownership and provide innovative opportunities to lower cost and offer flexible service models such as: (1) closed campus applications (e.g., business parks); (2) first- and last-mile connectivity; (3) low-density service; (4) off-peak/late-night service; and (5) paratransit services. SAV business models, along with passenger capacity, will shape their impacts.

Possible Impacts of Shared Automated Vehicles

While SAV impacts remain uncertain, many practitioners and researchers predict higher efficiency, affordability, and lower GHG emissions. A few studies have found that wait times could lower willingness to switch to SAVs, and marginal increases in cost also affect the likelihood of using pooled SAVs. Some hypothesize that traditional taxis in urban areas may be the first transportation mode to be replaced by SAVs. As such, automation could increase the ease of multimodal connections to public transit facilitated by SAVs (first and last mile). However, SAVs could also compete with public transportation but also facilitate peak shedding, alleviating peak public transit congestion.

Other studies have modeled the impacts of replacing “all” car and bus trips with a portion of trips served by SAVs that incorporate pooled rides. One study of Lisbon, Portugal forecasts when existing trips are served by a combination of SAV taxis and shuttle buses, emissions are reduced by one-third and 95% less public parking is required. Furthermore, the vehicle fleet would only need to be 3% the size of today’s light-duty vehicle and bus fleets. The study also predicts total VKT would be 37% lower than the present day during peak periods, although each vehicle would travel 10 times the total distance of current vehicles (OECD/ITF, 2016). However, SAVs could also contribute to induced travel demand due to reduced travel times and costs. More research is needed to better understand the potential of SAV induced travel demand on mode split and VMT/VKT.

Some studies have also concluded that a fleet of shared automated electric vehicles with “right sizing” of vehicles by trip, in combination with a 2030 low-carbon electricity grid, could reduce per-mile GHG emissions by a range of 63% to 82% compared to a privately owned, hybrid vehicle in 2030. The per-mile GHG reductions are 87% to 94% lower than a privately owned, gasoline-powered vehicle in 2014. Half of these emission savings are attributed to smaller right-sized vehicles based on trip needs (Greenblatt

and Saxena, 2015). However, more research may be needed to understand the impacts of SAV right-sizing on traveler waiting times as well as deadheading, particularly in lower-density built environments.

If automated vehicles (AVs) become mainstream, SAVs may constitute a sizeable portion of trips. However, travel behavior estimates are unknown. The number of personally owned AVs will likely determine, to some degree, SAV service demand. Impacts will also depend on sharing levels (concurrent or sequential) and the future modal split among public transit, shared AV fleets, and shared (or pooled) rides. It is possible that SAV fleets could become widely used without very many shared or pooled rides, and single-occupant vehicles will continue to dominate the majority of vehicle trips (e.g., users could access a shared fleet without sharing a ride). It is also feasible that shared rides could become more common, if automation makes route deviation more efficient, more cost effective, and more convenient. To date, most studies have not been able to deeply assess the propensity for pooled rides, since SAV travel behavior data currently do not exist. Travel behavior, business models, and public policy will be key components in determining how pooling and SAV impacts unfold.

Relationship Between SAVs and Public Transportation

In the future, automation could be the most transformative trend to impact public transportation since the automobile. Automation will likely result in fundamental changes to public transportation by altering the built environment, costs, commute patterns, and modal choice (Shaheen and Cohen, 2018). SAVs have the potential to create new opportunities to right-size public transit fleets and for public agencies to partner with the private sector. Reduced vehicle ownership due to SAVs could reduce parking demand and create new opportunities for infill development and increased densities. While SAVs may compete with public transit ridership, infill development could also create higher densities to support additional public transit ridership and allow for the conversion of bus transit to rail transit in urban cores. However, reduced commute stress from vehicle automation could make longer commutes more practical, reinforcing historic patterns of suburban and exurban residential choices. In the future, if more households telecommute due to an increase in flexible work schedules (at least part-time) and part-time commutes become less expensive and more productive due to AVs, today's travel time of commuting (and congestion) could be notably reduced. This could reinforce historical patterns of low-density development and private-vehicle ownership (including AVs). Additionally, SAVs could reduce public transit use, encourage lower vehicle occupancies (compared to fixed-route transit), and increase SAV trips. Just as AVs have the potential to reduce driving costs, automated transit vehicles have the opportunity to reduce operating costs and potentially pass on these savings to riders in the form of lower fares, more routes, first- and last-mile connectivity, and increased service frequency. This could make public transit more convenient and competitive with other modes, resulting in increased ridership (Shaheen and Cohen, 2018). It is important to consider this range of possible scenarios in planning for an automated vehicle future.

Summary

Studies on TNC impacts are emerging and have been shown to demonstrate varied impacts on VMT/VKT, GHGs, modal choice, automobile ownership and use, and public transit ridership, which are affected by local characteristics such as: urban density, land use, the built environment, public transit accessibility, public policy, and other factors. Many companies are beginning to explore evolving TNCs into SAV fleets. While SAV impacts remain uncertain, a number of studies suggest greater efficiency and affordability with lower GHG emissions. SAVs have the potential to impact public transportation in a variety of ways that present new opportunities for competition and complementarity. Vehicle automation could result in higher average vehicle occupancies (due to growth in shared fleets), lower vehicle occupancies (from the growth of zero occupant vehicles), or both—depending on the use case and spatial-temporal effects. While vehicle automation could transform public transportation services, leading to more cost effective and efficient service, automation could also compete with public transit, shifting riders to lower occupant modes. Policymakers should consider addressing the range of risks and opportunities associated with shared and automated mobility in planning for the future of urban mobility, including the transition to a more automated world.

Acknowledgment

The authors would like to thank the American Planning Association, the California Department of Transportation (Caltrans), the Mineta Transportation Institute, the US Department of Transportation, the Natural Resources Defense Council, and the Hewlett Foundation for their generous support of this work. Special thanks to the shared mobility operators, industry experts, and policymakers, who make this research possible. The contents of the chapter reflect the views of the authors and do not necessarily indicate sponsor acceptance.

Biographies

Susan Shaheen is a professor in the Department of Civil and Environmental Engineering at the University of California (UC), Berkeley. She is also Codirector of the Transportation Sustainability Research Center (TSRC) at the Institute of Transportation

Studies, Berkeley. She was among the first to observe, research, and write about the changing dynamics in shared mobility and the likely scenarios through which AVs will gain prominence. She was the policy and behavioral research program leader at California Partners for Advanced Transit and Highways, a Special Assistant to the Director's Office of the California Department of Transportation, and the first Honda Distinguished Scholar in Transportation at the Institute of Transportation Studies at UC Davis, where she served as the endowed chair from 2000 to 2012. She has authored numerous journal articles, reports and proceedings, book chapters, and coedited two books. She has a PhD from UC Davis and a MS from the University of Rochester.

Adam Cohen is a shared mobility researcher at TSRC at UC Berkeley. Since joining the group in 2004, his research has focused on shared mobility, mobility on demand, mobility as a service, urban air mobility, smart cities, and other innovative and emerging transportation technologies. He has coauthored numerous articles and reports on shared mobility in peer-reviewed journals and conference proceedings. His academic background is in city and regional planning and international affairs.

See Also

Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service; Shared Mobility Opportunities and their Computational Challenges for Improving Health-Related Quality of Life

References

- Alami, F., Circella, G., Handy, S., Mokhtarian, P., 2018. What influences travelers to use Uber? Exploring the factors affecting the adoption of on-demand ride services in California. *Travel Behav. Soc.* 13, 88–104. [10.1016/j.tbs.2018.06.002](https://doi.org/10.1016/j.tbs.2018.06.002).
- Brown, A., Taylor, B., 2018. Bridging the gap between mobility haves and have-nots. In: Sperling, D., *Three Revolutions: Steering Automated, Shared, and Electric Vehicles to a Better Future*, second ed., Island Press, Washington DC, pp. 131–150.
- Clewlow, R., Mishra, G., 2017. Disruptive transportation: the adoption, utilization, and impacts of ride-hailing in the United States. Institute of Transportation Studies, University of California, Davis, CA. Available from: https://itspubs.ucdavis.edu/wp-content/themes/ucdavis/pubs/download_pdf.php?id=2752.
- Feigon, S., Murphy, C., 2018. Shared mobility and the transformation of public transit. Transportation Research Board, Washington DC. Available from: <https://s3-us-west-2.amazonaws.com/cdn.sudbury.ma.us/wp-content/uploads/sites/391/2018/12/Shared-Mobility-and-the-Transformation-of-Public-Transit-2.pdf?version=adeae0a529ec275303f2ed4d8109c3d0>.
- Greenblatt, J., Saxena, S., 2015. Autonomous taxis could greatly reduce greenhouse-gas emissions of US light-duty vehicles. *Nat. Clim. Change* 860–863.
- Hampshire, R., Simek, C., Fabusuyi, T., Di, X., Chen, X., 2017. Measuring the impact of an unanticipated disruption of Uber/Lyft in Austin, TX.
- Henao, A., 2017. Impacts of ridesourcing – Lyft and Uber – on transportation including VMT, mode replacement, parking, and travel behavior. PhD. University of Colorado at Denver.
- Martin, E., Shaheen, S., Stocker, A., 2020. Impacts of transportation network companies on vehicle miles traveled, greenhouse gas emissions, and travel behavior. University of California, Berkeley.
- OECD/ITF, 2016. Shared mobility: innovation for liveable cities. ITF Corporate Partnership Board Report. Paris. Available from: <https://www.itf-oecd.org/sites/default/files/docs/shared-mobility-liveable-cities.pdf>.
- Shaheen, S., Cohen, A., 2019. Shared ride services in North America: definitions, impacts, and the future of pooling. *Trans. Rev.* 39 (4), 427–442. doi:10.1080/01441647.2018.1497728.
- Shaheen, S., Cohen, A., 2018. Is it time for a public transit renaissance? Navigating travel behavior, technology, and business model shifts in a brave New World. *J. Public Transp.* 21 (1), 67–81. doi:10.5038/2375-0901.21.1.8.

Further Reading

- Cohen, A., Shaheen, S., 2016. Planning for shared mobility. American Planning Association, Chicago. Available from: <https://www.planning.org/publications/report/9107556/>.
- Gehrke, S., Felix, A., Reardon, T., 2018. Fare choices: a survey of ride-hailing passengers in metro boston. Metro Area Planning Council, Boston, MA. Available from: <http://www.mapc.org/wp-content/uploads/2018/02/Fare-Choices-MAPC.pdf>.
- SAE International, 2018. J3163: Taxonomy and definitions for terms related to shared mobility and enabling technologies. SAE International.
- San Francisco Municipal Transportation Agency, 2017. The TNC regulatory landscape: an overview of current TNC regulation in California and across the country. SFMTA, San Francisco. Available from: https://www.sfcta.org/sites/default/files/content/Planning/TNCs/TNC_regulatory_020218.pdf.
- Shaheen, S., Cohen, A., Zohdy, I., 2016a. Shared mobility: current practices and guiding principles. U.S. Department of Transportation, Washington, DC. Available from: <https://ops.fhwa.dot.gov/publications/fhwahop16022/fhwahop16022.pdf>.
- Shaheen, S., Cohen, A., Zohdy, I., Kock, B., 2016b. Smartphone Applications to Influence Travel Choices: Practices and Policies. U.S. Department of Transportation, Washington, DC. Available from: www.ops.fhwa.dot.gov/publications/fhwahop16023/fhwahop16023.pdf.
- Shaheen, S., Cohen, A., Yelchuru, B., Sarkhili, S., 2017. Mobility on demand operational concept report. U.S. Department of Transportation, Washington, DC. Available from: <https://rosap.nhtl.bts.gov/view/dot/34258>.

Public Transport in Low Density Areas

Jani-Pekka Jokinen*, Leif Sørensen*,†, Jan Schlüter*,‡, *Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany;

†Department of Psychology, PFH Private University of Applied Sciences Göttingen, Göttingen, Germany; ‡Department of Economics, Georg-August-University of Göttingen, Göttingen, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	589
Service Provision, Pricing, and Subsidies	590
CBA and Wider Economic Benefits of Public Transport in Rural Areas	591
Decisions on Provided Transport Modes and the Role of New Technologies	592
The Role of New Transport Modes and Technologies	592
Demand Responsive Transport	592
Benefits and Challenges	593
Outlook	594
Conclusion	594
References	595
Further Reading	595

Introduction

This chapter considers the challenges and solutions for public transport (PT) in low population density areas, such as rural regions and outer suburbs, from economic perspective. For defining low-density areas, we use OECD's definition for rural areas, which is 150 inhabitants per km² (Dijkstra and Poelman, 2014). In these areas travel demand (per square kilometer) for PT is typically much lower than in densely populated urban areas for two interdependent reasons: (1) Total travel demand is lower, and (2) Market share of PT is lower, because private car is relative more competitive on the rural areas where bus frequency is typically much lower. The lower demand density causes economic challenges for providing PT due to relatively low revenues and cost recovery compared to PT provision in urban areas. The on-going global urbanization makes this challenge even more topical and difficult for rural areas, which sets decision makers to choose between lower service level (e.g., lower bus frequencies) or higher trip fares to maintain cost recovery. The third option is to raise subsidies instead of fares, which would require increasing taxes and plausible arguments for having political acceptance. For bringing clarity to these decisions and their mutual interdependencies, we consider an optimal PT provision and pricing in section "Service provision, pricing, and subsidies", which provide basis for the economic arguments for subsidization of PT, that is, economies of scale and underpricing of private car use (Fig. 1).

In addition, there are other political arguments for subsidies, such as equity and right to accessibility, for citizens. These various arguments for subsidies either intensify or diminish when demand density decrease. Moreover, there are certain specific arguments for organizing and supporting PT in rural areas, which are related to the wider benefits for economic activity in rural areas. For instance, PT can improve mobility of workers and job seekers, which consequently increase labor supply and thereby can support local industry and stimulate social and economic activity. These wider economic benefits are considered in section "CBA and wider economic benefits of public transport in rural areas".

The set of used transport modes and related infrastructures is the other crucial aspect of decision-making in PT provision. Traditionally, the most common PT modes in low-density areas are buses and paratransit services. Technological advancements in Information and Communication Technology (ICT) and intelligent transport technologies have increased interest for new flexible transport services (FTSs) such as automated demand responsive transport (DRT) and shared taxis. This development has already realized on the emergence of the new pilots and in some cases established FTSs, some of which operate in low-density areas. Moreover, autonomous vehicle technologies can be seen as a promising solution for cost pressures of rural PT as currently salary costs of bus drivers incur even 70% of operational costs. However, the potential societal impacts of new flexible services are multidirectional, and much is depending on how these services are implemented and priced. For instance, an increased flexibility of route typically increases driven vehicle kilometers compared to conventional fixed route services where walking distances to the bus stops are usually longer. On the other hand, flexible DRT services are driven only when needed and only to the requested pick-up locations using shortest routes whereas fixed route bus services follow predefined schedules and routes even if there are no passengers on the route (or more likely on the part of the route) for a certain hour. The decisions on transport modes and possibilities of new technologies are considered in section "Decisions on provided transport modes and the role of new technologies". Finally, last section " " presents conclusions and policy implications.



Figure 1 Rural automated DRT service EcoBus in central Germany 2019. The marketing slogan “Sie sind die Haltestelle” (English: “You are the bus stop”) describes the flexibility characteristics of the EcoBus, which provided a public door-to-door service. Source: ©MPIDS.

Service Provision, Pricing, and Subsidies

Decisions on provision of PT are typically done by regional or municipal transport authorities (sometimes transport operators have also significant decision-making power). However, these local transport authorities rarely (or barely ever) have a full autonomy as state legislations set boundaries on their decision-making. Moreover, government’s transport policies and investments on transport infrastructure (e.g., road and rail network) define largely the environment where the local transport authors aim to fulfil their duties. Furthermore, there are various other factors influencing on requirements and decisions of local transport authorities, such as passenger needs and demand for PT, market structure of transport service companies, citizens’ opinions, and values that influence through municipal politics. In this section, we aim to clarify these complex decisions by explaining the main results of the transport economics, which define optimal PT provision and pricing from the economic viewpoint.

The general objective for social planner (i.e., transport authority deciding on PT provision) is to maximize social welfare defined as total benefits minus costs, where the benefits are measured as total willingness to pay for using PT and the costs include the production costs of transport operator, travel time costs of passengers and external costs for society (such as pollution, noise, and congestion). With this general objective in mind, we first focus on a single bus route, which requires three main decisions: (1) A bus fleet size, (2) A bus frequency on the route, and (3) A bus ticket price. Optimal price is equal to social marginal costs, that is, incremental costs for society due to an additional passenger, which gives correct incentives for travelers’ choices, and thereby leads to an efficient allocation of resources and a social welfare maximization (Small et al., 2007; Kaddoura et al., 2015). Provision of certain bus frequency requires both buses (capital) and drivers (labor). Some essential results related to this optimization problem was provided by Herbert Mohring, who analyzed bus transportation and included also the value of travel time (i.e., travel time is an input provided by passengers) to the total costs (Mohring, 1972). He showed that to minimize costs of service provision require using that amount of capital which generate equal short run marginal costs and long run marginal costs (i.e., in the short run amount of capital is fixed). Moreover, if the long run marginal costs are decreasing in scale (i.e., decreasing with higher levels of trip supply and demand), then long run average costs are higher than the long run marginal costs, which means together with the marginal cost pricing that a subsidy (equal to the quantity of supply multiplied by the difference between average and marginal cost) is required to cover the total costs. More specifically, in bus services the main reason for the gap between average and marginal cost is that passengers’ waiting time costs decrease when bus frequency increase.

In general, the positive scale economies of PT causes that average cost of producing trips by PT are lower in urban areas where demand density is higher. Therefore, it is necessary to consider different variants of scale economies for deeper understanding economic challenges of PT provision in low-density areas. Mohring’s analysis explained so called *demand side scale economies* (decreasing waiting time costs), which is one of the main arguments for subsidizing PT. There can be also *supply side scale economies* in PT due to fixed capital costs or investments on required infrastructure (bus stops, terminals, rails, parking places or bus garages for the fleet, etc.), which are more efficiently utilized with higher scale of service provision. In addition, there are positive supply side scale economies in taxi and DRT services (which are often provided as an integral part of municipal PT), due to the shorter pickup routes of the vehicles when both the number of passengers and vehicles are simultaneously increased (and spread evenly) on geographically fixed service areas (Jokinen, 2016a).

Thus, both the supply side and demand side scale economies need to be taken into account when defining optimal PT provision and pricing. The other relevant classification of scale economies is the distinction between economies of *density* and *size*. Where the former is a form of economies of scale within a transport network when both the demand and supply of transport services are increasing and leading to lower average costs of passenger trips (as already described in the Mohring’s analysis), whereas the latter in related to (possible) scale economies of transport network extension, for example, lengthening existing bus route or adding new bus route. On the rural areas, there can be positive economies of size if, for instance, the bus route is extended for connecting two rural

villages and leading to significant increase of passenger demand with relatively small increase in costs of bus operations. On the contrary, there would be negative economies of size, if the bus route was extended to any distant location with low-demand density along the new part of the route.

In rural areas, low passenger demand and scarce resources for PT has often forced to produce other services with the same resources. For instance, bus services can also deliver mail and packets along the routes in many European and African countries, which has (at least to some extent) helped transport authors to fulfill their duties (Adams, 1981). The other example is companies that provide minibuses simultaneously for municipal school trips and for private taxi trips. In general, there are positive *economies of scope* if it is cheaper to produce several products or services in the same company than in the separate companies. In the rural areas, there can be positive economies of scope in combined provision of PT and post services (or other delivery services), because bus stops and delivery points can be combined, and these are not too densely located along the routes. Therefore, production of post or delivery service does not cause too much additional travel time costs for the bus passengers, whereas in the high-density areas these additional travel time costs would probably become too high, and consequently, leading to negative economies of scope. New technologies and digital services provide more possibilities for combined production of PT and other services in rural areas, which are considered in section "Decisions on provided transport modes and the role of new technologies."

Up to this point, we have considered optimal PT provision, pricing and subsidies without any constraints in decision-making, that is, we have considered so called *first best policies*. However, in reality decision-making is limited by many constraints related to legislation, political acceptability or budget constraints. Optimal decisions under these type of constraints are called as *second best policies*. One common such constraint is related to pricing of private car usage, and more specifically, to an incapability for adjusting prices dynamically with congestion levels. Marginal social costs of private car usage are high during congested peak hours leading to underpricing of private car due to the lack of congestion pricing scheme. Thus, in this case the welfare maximizing principle of marginal social cost pricing is violated, which causes a constraint for defining (second best) optimal PT provision and pricing. Now, if there is a positive cross elasticity between PT and private car, that is, price decrease of PT attracts private car users to shift their transport mode to PT (and vice versa), then it is optimal to subsidize PT for reducing congestion costs. This is so-called second-best argument for subsidizing PT. On rural areas congestion is a relevant challenge mainly only on highways, but it can have, however, a crucial effect of travel time costs of commuters, and consequently on local labor supply and economic activity of rural areas, especially, if a significant share of rural population commutes long distances to workplaces. Moreover, mode choices of rural commuters can have influence on the congestion of the nearest urban areas. This interaction between low- and high-density areas should be taken into account in the transport policies of larger areas (e.g., states or provinces), and it gives one argument for government subsidies for rural public transportation.

CBA and Wider Economic Benefits of Public Transport in Rural Areas

Thus far, we have considered the optimal policies of PT in low-density areas. The presented results provide understanding of economic arguments for subsidies and other related policies. However, the decisions on PT provision can be evaluated also, regardless of optimality conditions that are often (somewhat) ignored in political processes, by a more general appraisal method of cost–benefit analysis (CBA), where a proposed transport investment or policy change is evaluated by estimating expected future benefits and costs for users and society. The CBA gives an evaluation whether the benefits outweigh the costs, that is, is a certain policy or project economically reasonable. Moreover, the CBA can be used to rank alternative policies and projects in a decision-making situations where only one or some of the all proposed projects can be implemented (Börjesson et al., 2014). The CBA has been used widely for evaluating PT in rural areas. Godavarthy et al. (2014) provide a review of implemented cost–benefit analyses on rural PT in the United States. They mention three relevant areas of benefits of rural PT: (1) Transportation cost savings including internal costs like travel time costs and chauffeuring costs (e.g., parents cost for driving children to school and hobbies) and external costs such as car emission costs and traffic accident costs (which are only partially internalized); (2) Low-cost mobility benefits, which are realized from trips that would have been foregone without rural PT like medical and educational trips of noncar population; (3) Economic benefits that relates to economic activity induced by rural PT.

In general, the conventional CBA focuses on evaluating and comparing user benefits and costs. However, there can be also other substantial wider economic effects induced by transport projects. Vickerman (2008) identified four wider effects of large transport projects: (1) productivity effects, (2) agglomeration effects, (3) competition effects, and (4) labor-market effects.

The productivity effect relates to the results that infrastructure investments have positive effect on economic output especially on regions where the lack of infrastructure has constrained local economy to integrate into wider markets. For instance, rural PT can benefit local tourist businesses by making their services available for new customer segments.

The second wider effect, that is, agglomeration effect, is usually referred to the benefits from proximity of similar companies in urban areas (e.g., cooperation in R&D and larger labor pool in the area). However, the proximity can be seen as a relative concept, and therefore, even in rural areas improvements in transport services can bring companies closer and similarly advance mutual cooperation between companies.

The third effect (on competition) was first identified by Jara-Diaz (1986), and relates to the effect that transport investment can reduce market power of local monopolies and thereby reducing deadweight, and thus create positive external effect (i.e., increased competition forces companies to produce more with lower prices). The market power of local monopolies can be high in isolated rural areas. However, these type of isolated markets are likely rare in the most developed market economies (Vickerman, 2008).

Transport investments can have effects on labor-markets if travel costs of commuters decrease and companies have access to a larger labor supply, which decrease wages due to competition, and provide more skilled and productive labor. Thus, investments on rural PT can have effects on the labor-markets both on that rural area and on the nearest urban centers, because improved PT connections attract rural population to work in the nearest urban areas, and, respectively, employers in the rural area can attract skilled labor from the urban area.

In addition to these four wider benefits influencing on other markets (i.e., beyond transport markets), rural PT have other benefits for citizens that are also so called wider benefits (or values) in the sense that these are not directly related to the use of PT, that is, option values and nonuse values. Option values are based on the value of the possibility to use PT either in unusual situations when, for instance, own car is not available or in the future when changes in life cause a need for PT. For instance, rural train can become attractive transport mode due to finding a new job. In contrast to option values, nonuse values are unrelated to the use of PT, neither occasionally now nor in the future (Johnson et al., 2013). Nonuse values of rural PT can be related to altruistic values (e.g., valuing equal accessibility and social inclusion for the all citizens) and indirect user benefits. However, Johnson et al. (2013) note about scope for significant overlap between nonuse values and user benefits or externalities that are usually taken into account in the conventional CBA.

Decisions on Provided Transport Modes and the Role of New Technologies

The Role of New Transport Modes and Technologies

PT is undergoing disruptive changes mainly induced by developments in the ICT sector. After decades of political prioritizations of motorized individual transport (Wallin Andreassen, 1995), a societal and political shift toward new mobility concepts induced by demographic change, austerity policies, environmental, and climate concerns is occurring. New mobility services address the heterogeneous consumer preferences which traditional PT services could not satisfy. Thus, a change on the demand side for mobility has developed over the past years. A simultaneous shift in the car and services industry results in new providers of shared mobility services, a traditionally publicly dominated sector (or dominated by informal sector in many developing countries). The supply side of the mobility market is, therefore, also changing with so-called mobility as a service (MaaS) concepts (Aapaoja et al., 2017). Most of these new modes address the shortcomings of the traditional PT services, for example lack of flexibility and supply in rural areas. Based on the potential to reduce car dependency and car usage for people in rural areas, innovative PT can play an essential role in a modal shift or at least alleviate the economic challenges of rural PT caused by urbanization and diminishing rural population. For the rural areas, however, with lower population-densities the developments differ significantly to those of urban areas where manifold new transport modes are made available, for example, e-scooter or car-sharing services. Many transport modes are unfeasible due to a lack of economies of scale, for example, because of the small numbers of passengers a metro or tram line is inefficient in low population areas. Consequently, mainly busses, rail services and some taxi services exist in these areas. Hence, the share of PT for overall mobility is still low (Pucher and Renne, 2005). Over the past decades, however, the fixed schedule and fixed route services, for example, buses and light rail services, were complemented or substituted by more flexible and DRT (Mounce et al., 2018). While the former conventional modes are heavily regulated the latter new modes are only slowly falling under regulations. These circumstances are result as well as reason for new mobility providers in the markets, especially in urban areas, who benefit from an unregulated and innovative environment. However, new transport modes in rural areas are sparse since innovation requires time and political support to disperse toward less populated regions. Under these premises, innovative mobility services may help to reverse the low public transportation supply and demand in rural areas. For many innovations in the transportation sector, certain ICT requirements exist, which are commonly fulfilled in developed countries. In developing countries, however, the situation may differ and innovation is less dispersed, such that many of the new transport modes and technologies described here are not yet feasible in all areas worldwide.

Demand Responsive Transport

A prominent example for innovation in rural PT is DRT services. These DRT services are a technological evolution to former dial-a-ride services which were available in the 1970s (Guenther, 1970; Gustafson and Navin, 1973; Li and Quadrioglio, 2010; Roos et al., 1971). Many of formerly purpose-oriented DRT services are now being merged into services addressing the broader rural population with the goal of an efficient use of resources, for example, vehicle seats or bus drivers. DRT services are commonly positioned between regular bus and taxi services. As an example, Wang et al. (2015) describe DRT services as follows: (1) The service is publicly available, (2) it applies small busses or vans, (3) depending on demand the route and time is dynamically adapted, and (4) the fare is to be paid per passenger as it is in conventional public busses.

DRT concepts are also referred to as FTSs (Velaga et al., 2012). These FTS or DRT services can follow different structures and setups: For one, the scope of route flexibility can vary, for example it can follow specific virtual stops or is limited to pick-ups and drop-offs along in certain areas or it can come as a true door-to-door (also sometimes called curb-to-curb) service. For two, there can be time constraints on prebooking times. Some FTS allow bookings with a prebooking time of several days while other services may be limited to prebooking times of 1 h or even only instant trips. The specific technical infrastructure is mainly dependent on the envisaged form set by the operators and local transport authorities. Research on ideal setups for specific circumstances exists and indicates that certain characteristics of an operation area require consideration in the respective service setup (Velaga et al., 2012).

In combination with the terms of DRT and FTS the concepts of *ride-sharing*, *ride-pooling*, *ride-sourcing*, and *ride-hailing* are frequently mentioned (Machado et al., 2018). These concepts, which are often used interchangeably, differ significantly regarding several dimensions, for example, the supplying entity or technical setup. However, similarities exist as most concepts have comparable goals, for example, reduction of travel costs per person, or an improved environmental impact due to lower emissions compared to individual rides.

Ride-sharing generally refers to combining commutated trips in one vehicle. In comparison to *ride-pooling*, it is privately organized and also privately operated, for example, employees of the same company share a car for their commute. Hence, the driver and passenger(s) share an identical or similar destination. It has been argued that ride-sharing is motivated by other reasons than profit (Chan and Shaheen, 2012). Related terms are car-pooling, car-sharing and lift-sharing. *Ride-sourcing* can be used synonymously with ride-sharing (Jin et al., 2018). Ride-sharing initiatives often relate to periods of scarce resources and lower welfare, for example, during the World War II. Ferguson (1997) links the reduction in carpooling for the United States between 1970 and 1990 to the increased number of vehicles per household and lower costs, for example, fuel costs.

Ride-pooling requires an organizing entity that matches commutated trips. This organizing entity is nowadays an ICT system including vehicle dispatching algorithm and routing algorithm that utilize real-time location and trip request data that assigns similar trip-requests to a specific vehicle along a dynamic route. The concept of ride-pooling combines characteristics of ride-sharing and ride-hailing if, for example, a transport authority offers an on-demand mobility service where passengers demand (hail) a vehicle to be transported to a destination by a driver. If there are no further trips pooled along the route, the offered service to the customer shows all characteristics of ride-hailing aside from being publicly accessible during and along the route.

Ride-hailing refers to the traditional hailing of a taxi for transportation. In general, it refers to the traditional means of requesting transportation as an individual trip through a virtual path, that is, mobile applications. In that, it differs from the ride-pooling concept as only one trip at a time is serviced and ride-hailing does not fall under the concept of sharing a trip (see Frenken and Schor, 2019). In an analysis of vehicle miles travelled, Henao and Marshall (2019) found two prominent ride-hailing services, Uber and Lyft, to increase vehicle kilometers driven compared to the number without ride-hailing services. Their investigation is centered around the US city of Denver implying that for rural regions ride-hailing services might further increase the vehicle kilometers due to longer deadheading.

Concerning the question of the supplier of mobility services in rural areas, either public or private entities can provide ride-pooling or ride-sharing. However, considering the relatively high subsidy requirement of transportation in low-population density areas due to higher costs compared to possible revenues, public entities are likely to remain responsible for mobility provision. For example, Soder and Peer (2018) state that mobility provision through employers in an Austrian rural area is socially and economically inefficient.

Research on drivers of DRT usage are still as sparse as operational DRT services. Therefore, only few insights on relevant factors for DRT usage are available. Wang et al. (2015) categorize them into *service-related*, *area-related*, and *individual-related* factors. Waiting time, in-vehicle time, service frequency and overall time and route flexibility are service-related factors that influence DRT usage (see Tyrinopoulos and Antoniou, 2008; Takeuchi et al., 2003).

Benefits and Challenges

For a transport mode to be used by many customers, the benefits must outweigh the costs especially compared to individual transportation modes, for example, the car. Thus, flexibility and availability are commonly mentioned in this context. DRT meets these challenges and is intended to supply a flexible and financially viable transport service. At the same time, DRT services may allow to combine different transportation services by incorporating services with a specific user group, for example, trips for medical examinations. Overall, societal gains can result if the new transport modes are implemented under consideration of regional circumstances. By reducing social exclusion based on the lack of mobility it can further help to revitalize rural areas (Farrington and Farrington, 2005). Fransen et al. (2015) show that lack of mobility contributes to social disadvantages. López-Iglesias et al. (2018) highlight the potential for rural development by connecting rural hinterlands with local functional areas. Also, Hough and Taleqani (2018) conclude that technological developments are likely to revive rural areas because of higher livability and accessibility through innovative transport modes. For the long-run, analyses of the induced travel effect, that is, the creation of further travel demand by an extended mobility supply, will be of interest (Rayle et al., 2016).

In case a DRT service manages to substitute private individual car usage, environmental benefits are a direct and indirect result. A direct effect can be achieved through less vehicle emissions when ride-pooling is realized through DRT. Indirect effects relate to less traffic and, thus, less congestion, less additional cruising in search for parking and freed up parking spaces. For rural areas, the lack of parking spaces is unintuitive, provided that space is abundant and, thus, pricing of land is lower compared to densely populated areas. However, topographical circumstances can limit available space, for example, for parking, even in rural areas. For example, in a rural tourism context, the DRT service *EcoBus* in central Germany helped accessibility of mountainous skiing areas during snow periods while some streets were blocked due to the lack of parking in the Harz region. However, as Le-Klähn and Hall (2015) state, PT still only play a minor role in tourism for remote or rural areas. Thus, innovative approaches to PT such as automated DRT may help to attract tourists for rural regions. The importance of PT for sustainable rural tourism has been highlighted by Tomej and Liburd (2019).

A challenge in new mobility services is the consolidation of several services, for example, taxi services. As DRT can be positioned between regular bus transport and taxi services, the latter may be pushed out of the market. However, for specific tasks a ride-pooling

service may still be too inefficient or inadequate such that regular taxi services could account for that niche market. Eventually, some providers could shift or be included in a public DRT service while some remain to meet the excess demand. In this context, it has to be mentioned that diversity in transport services may enhance the overall PT provision and should be a desirable situation in the long run. Therefore, transport authorities are required to increase cooperation to provide integrated services to their passengers. Intuitively, new rural PT solutions should alleviate the segmentation among providers to allow for an integrated and broad service (Hough and Taleqani, 2018). However, funding (especially subsidy) structures need to adopt accordingly. Again, cooperation among providers is key for the future of innovative PT provision.

Outlook

In the future, autonomous driving will induce the development of new services in the public transportation sector. Commercial shared autonomous vehicles (SAVs) could replace PT in areas where mass transit is economically inefficient (von Mörner, 2019). Thus, SAVs inhibit high potential for rural and low-density areas where mass transportation is unviable (Herminghaus, 2019). For regular PT services and more importantly for DRT the cost of personnel often prevents a broad application of innovative services. For developed countries, the lack of drivers imposes an operational constraint and can limit the quality of service for customers. Therefore, SAVs can help to meet the shortage in labor supply and, in addition, reduce the operational costs for the service providers, for example, the transport authorities. As Litman (2009) points out, lower transportation costs per capita are achievable through ride-sharing compared to any other motorized transport. In reference to societal costs of transportation, Jokinen (2016b) identifies DRT services to inhibit lower societal costs compared to private car and taxi if demand is sufficient. Under the assumption that new mobility services attract more users and relieve the car dependency, Stiglic et al. (2018) highlight the high potential in cost efficiency for PT with integrated intermodal services.

However, in most rural areas the ICT infrastructure will pose a bottleneck for a broad adoption of autonomous vehicles (König and Neumayr, 2017). In reference to the application of technology in new mobility concepts in rural areas, the standards, practices of public-private partnerships and more importantly the respective laws and regulations need adjustments in the process (see Hensher, 2017).

Conclusion

In this chapter we have presented some essential results of the economics of PT, which define welfare optimal provision and pricing of PT under general market conditions and give economic arguments for subsidies, that is, economies of scale and underpricing of private car. Thus, the presented results explain why subsidized public transport is needed as a crucial part of efficient and sustainable transport systems. These general results are relevant also for low-density areas, but certain specific characteristics, such as lower total demand for transport and more scarce resources available for PT provision, must be taken into account in optimal policies. In addition to these results and economic arguments, we have considered the use of the CBA in evaluation of rural PT projects, and reviewed some potential wider economic benefits, which are relevant also in rural areas but can take different forms than in urban centers. The ongoing urbanization causes more cost pressures for maintaining rural PT. On the contrary, at the same time environmental concerns drive society to develop and provide other transport modes than private car. From these economic and societal viewpoints, we have considered the role of new technologies for maintaining and even improving rural PT.

As rural demographics show a negative trend for developed countries, mobility services will be necessary in the context of social inclusion through mobility or access to medical services for immobile people. Thus, it has to be recognized that PT will be a vital aspect for rural communities in the future and requires adjustments accordingly. In this context, the financial burden should be reflected upon overall societal costs and benefits, for example, the provision of innovative PT in rural regions may be costly from a community perspective but under consideration of the private costs, that is, through individual motorized transport, the environmental, congestion, and other costs are relevant for an overall evaluation. Therefore, introducing new innovative transport services in rural areas may allow for societal benefits and cost savings if applied strategically and efficiently. Close cooperation between a supervising transport authority and local stakeholders is essential in assuring that services are compliant with the demand in certain areas.

Several policy implications arise from the presented results and recent developments in PT services. One implication is that conventional bus services with fixed routes and schedules are not economically reasonable in the most depopulated areas, where travel demand for PT is too low and irregular, which can be verified by the CBA. In practice, this has led transport authorities to replace conventional PT by more FTs in many rural areas in order to ensure sufficient accessibility of citizens with lower costs. Apart from these most depopulated areas, there is need also for conventional rural PT which provides together with FTs significant benefits for users and usually enable wider economic benefits for rural areas and even for surrounding urban centers. An important aspect for the rural regions is cooperation between transport providers to realize economies of scale. For instance, the schedules of bus services and operating times of DRT should enable comfortable transfers to the rural train. The usage of integrated, multimodal smart services can only be fruitful if it is not limited by regional borders or insufficient cooperation between transport operators. The presented results and methods can improve mutual understanding of the potential benefits of rural PT and new technologies for the whole society, and thereby enhance cooperation between transport authorities, companies, politicians, and citizens.

References

- Apaaja, A., Eckhardt, J., Nykänen, L., Sochor, J., 2017. MaaS service combinations for different geographical areas. In: 24th World Congress on Intelligent Transportation Systems, vol. 29.
- Adams, D.E., 1981. Post-bus for rural passenger transportation and rural mail delivery: an idea whose time has come (abridgment). *Transp. Res. Rec.* 797, 77–79.
- Börjesson, M., Eliasson, J., Lundberg, M., 2014. Is CBA ranking of transport investments robust? *J. Transp. Econ. Policy* 48 (2), 189–204.
- Chan, N.D., Shaheen, S.A., 2012. Ridesharing in North America: past, present, and future. *Transp. Rev.* 32 (1), 93–112.
- Dijkstra, L., Poelman, H., 2014. A Harmonised Definition of Cities and Rural Areas: The New Degree of Urbanisation WP 01/2014.
- Farrington, J., Farrington, C., 2005. Rural accessibility, social inclusion and social justice: towards conceptualisation. *J. Transp. Geogr.* 13 (1), 1–12.
- Ferguson, E., 1997. The rise and fall of the American carpool: 1970–1990. *Transportation* 24, 349.
- Fransen, K., Neutens, T., Farber, S., De Maeyer, P., Deruyter, G., Wittlox, F., 2015. Identifying public transport gaps using time-dependent accessibility levels. *J. Transp. Geogr.* 48, 176–187.
- Frenken, K., Schor, J., 2019. Putting the sharing economy into perspective. In: Mont, O. (Ed.), *A Research Agenda for Sustainable Consumption Governance*. Edward Elgar Publishing, pp. 121–135.
- Godavarthy, R., Mattson, J., Ndembe, E., 2014. Cost-benefit analysis of rural and small urban transit. University of South Florida, Tampa: National Center for Transit Research.
- Guenther, K.W., 1970. Incremental implementation of dial-a-ride systems. (Special Report)In: Demand-Actuated Transportation Systems Conference.
- Gustafson, R.L., Navin, F.P., 1973. User preference for dial-a-bus. *Highways Research Board* 136, 85–93 (Special Report).
- Henao, A., Marshall, W.E., 2019. The impact of ride-hailing on vehicle miles traveled. *Transportation* 46, 2173–2194, doi:10.1007/s11116-018-9923-2.
- Hensher, D.A., 2017. Future bus transport contracts under a mobility as a service (MaaS) regime in the digital age: are they likely to change? *Transp. Res. A* 98, 86–96.
- Herminghaus, S., 2019. Mean field theory of demand responsive ride pooling systems. *Transp. Res. A* 119, 15–28.
- Hough, J., Taleqani, A.R., 2018. Future of rural transit. *J. Public Transp.* 21 (1), 4.
- Jara-Díaz, S.R., 1986. On the relation between users' benefits and the economic effects of transportation activities. *J. Reg. Sci.* 26 (2), 379–391.
- Jin, S.T., Kong, H., Wu, R., Sui, D.Z., 2018. Ridesourcing, the sharing economy, and the future of cities. *Cities* 76, 96–104.
- Johnson, D., Jackson, J., Nash, C., 2013. The wider value of rural rail provision. *Transp. Policy* 29, 126–135.
- Jokinen, J.P., 2016a. On the welfare optimal policies in demand responsive transportation and shared taxi services. *J. Transp. Econ. Policy* 50 (1), 39–55.
- Jokinen, J.P., 2016b. Economic Perspectives on Automated Demand Responsive Transportation and Shared Taxi Services - Analytical Models and Simulations for Policy Analysis. Aalto University (Doctoral dissertations, 120/2016).
- Kaddoura, I., Kickhöfer, B., Neumann, A., Tirachini, A., 2015. Optimal public transport pricing: Towards an agent-based marginal social cost approach. *J. Transp. Econ. Policy* 49 (2), 200–218.
- König, M., Neumayr, L., 2017. Users' resistance towards radical innovations: the case of the self-driving car. *Transp. Res. F: Traffic Psychol. Behav.* 44, 42–52.
- Le-Klähn, D.-T., Hall, C.M., 2015. Tourist use of public transport at destinations – a review. *Curr. Issues Tourism* 18 (8), 785–803, doi:10.1080/13683500.2014.948812.
- Li, X., Quadrioglio, L., 2010. Feeder transit services: choosing between fixed and demand responsive policy. *Transp. Res. C* 18 (5), 770–780.
- Litman, T., 2009. *Transportation Cost and Benefit Analysis*. Victoria Transport Policy Institute, 31.
- López-Iglesias, E., Peón, D., Rodríguez-Álvarez, J., 2018. Mobility innovations for sustainability and cohesion of rural areas: a transport model and public investment analysis for Valdeorras (Galicia, Spain). *J. Clean. Prod.* 172, 3520–3534.
- Machado, C.A.S., de Salles Hue, N.P.M., Berssaneti, F.T., Quintanilha, J.A., 2018. An overview of shared mobility. *Sustainability* 10 (12), 4342.
- Mohring, H., 1972. Optimization and scale economies in urban bus transportation. *Am. Econ. Rev.* 62 (4), 591–604.
- Mounce, R., Wright, S., Emele, C.D., Zeng, C., Nelson, J.D., 2018. A tool to aid redesign of flexible transport services to increase efficiency in rural transport service provision. *J. Intell. Transp. Syst.* 22 (2), 175–185.
- Pucher, J., Renne, J.L., 2005. Rural mobility and mode choice: Evidence from the 2001 National Household Travel Survey. *Transportation* 32 (2), 165–186.
- Small, K.A., Verhoef, E.T., Lindsey, R., 2007. *The Economics of Urban Transportation*. Routledge.
- Stiglic, M., Agatz, N., Savelsbergh, M., Gradisar, M., 2018. Enhancing urban mobility: integrating ride-sharing and public transit. *Comput. Oper. Res.* 90, 12–21.
- Takeuchi, R., Okura, I., Nakamura, F., Hiraishi, H., 2003. Feasibility study on demand responsive transport systems (DRTS) 5th Eastern Asia Society for Transportation Studies Conference. Journal CD-ROM.
- Tomej, K., Liburd, J.J., 2020, 222–239. Sustainable accessibility in rural destinations: a public transport network approach. *J. Sustain. Tourism* 28 (2), 222–239, doi:10.1080/09669582.2019.1607359.
- Tyrinopoulos, Y., Antoniou, C., 2008. Public transit user satisfaction: Variability and policy implications. *Transp. Policy* 15 (4), 260–272.
- Rayle, L., Dai, D., Chan, N., Cervero, R., Shaheen, S., 2016. Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. *Transp. Policy* 45, 168–178.
- Roos, D., Melone, T., Little, F., Porter, E., Wilson, N., Sussman, J., 1971. *The Dial-A-Ride Transportation System-Summary Report*.
- Soder, M., Peer, S., 2018. The potential role of employers in promoting sustainable mobility in rural areas: evidence from Eastern Austria. *Int. J. Sustain. Transp.* 12 (7), 541–551, https://doi.org/10.1080/15568318.2017.1402974.
- Velaga, N.R., Nelson, J.D., Wright, S.D., Farrington, J.H., 2012. The potential role of flexible transport services in enhancing rural public transport provision. *J. Public Transp.* 15 (1), 7.
- Vickerman, R.W., 2008. Cost-benefit analysis and the wider economic benefits from mega-projects. In: Priemus, H., Flyvbjerg, B., van Wee, B. (Eds.), *Decision Making on Mega-Projects: Cost-benefit Analysis, Planning and Innovation*. Edward Elgar Publishing Ltd., pp. 66–84.
- von Mörmel, M., 2019. Demand-oriented mobility solutions for rural areas using autonomous vehicles. In: Coppola, P., Esztergár-Kiss, D. (Eds.), *Autonomous Vehicles and Future Mobility*. Elsevier, pp. 43–56.
- Wallin Andreassen, T., 1995. (Dis) satisfaction with public services: the case of public transportation. *J. Serv. Mark.* 9 (5), 30–41.
- Wang, C., Quddus, M., Enoch, M., Ryley, T., Davison, L., 2015. Exploring the propensity to travel by demand responsive transport in the rural area of Lincolnshire in England. *Case Stud. Transp. Policy* 3 (2), 129–136.

Further Reading

- Camacho, T., et al., 2016. The role of passenger-centric innovation in the future of public transport. *Public Transp.* 8 (3), 453–475.

Market Failures in Transport: Direct and Indirect Public Intervention

Federico Boffa*, Alberto Iozzi†, *Free University of Bozen, Faculty of Economics, Brunico, Italy; †“Tor Vergata” University of Rome, Department of Economics and Finance, Rome, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	596
Market Power	597
Price Regulation Involves Two Dimensions: Level and Structure	598
Environmental Externality	598
Congestion Externalities	599
References	600

Introduction

Government is heavily involved in the transportation sector, both in the infrastructure and services provision and management. The transportation industry is characterized by several market failures, that is, situations in which the free market does not ensure an efficient outcome. Two types of market failures are particularly noteworthy: the cost structure, which exhibits declining average costs leading to natural monopoly, and the presence of several externalities.

As far as the cost structure is concerned, in most of the segments of the transport sector, a large portion of the total costs is fixed, which gives rise to economies of scale. Under these circumstances, monopoly is the most cost-efficient market structure (de Palma and Monardo, 2021). This cost structure is typical of infrastructures, including highways, railways, airports, ports. However, it is shared by many other transportation services as well, including, most prominently, public transit (Parry and Small, 2009).

Several externalities are involved in the transport sector. Two of them, particularly relevant, have to do with pollution and congestion. The transport sector is one of the greatest pollutants. In Europe, transport represents almost a quarter of Europe's greenhouse gas emissions and is the main cause of air pollution in cities. Within this sector, road transport is by far the biggest emitter accounting for more than 70% of all GHG emissions from transport in 2014, according to data from the European Energy Agency. ADD FOOTNOTE(See https://ec.europa.eu/clima/policies/transport_en.) In the United States, 29% of the greenhouse gas emissions in 2017 comes from transport (EPA, 2020).

Congestion is a negative externality that is typical of the transport sector. An increase in the utilization of an infrastructure negatively affects all the other users of that infrastructure. If the external effect is not appropriately priced or regulated, this can result in an excess congestion, above the socially optimal level (Franco, 2021).

The nature of the policies that governments can and do use depends on the source of the distortion they target. However, the different forms of government intervention to correct all these distortions may be classified on a spectrum ranging from *direct* government intervention, which directly impose a choice or a behavior to firms or to consumers, to *indirect* interventions aimed at modifying the incentives of the suppliers or of the consumers, so as to indirectly trigger a change in their decisions.

When government intervenes to mitigate market power, a straightforward form of *direct* government intervention consists in public ownership. This approach is based on the presumption that government-owned firms may efficiently operate with the objective of pursuing the welfare of the society as a whole. *Indirect* intervention, instead, involves delegation of the strategic choice to the firms, while setting up an appropriate incentive scheme, in the attempt to align their behavior to social optimality. This takes the form of incentive regulation. The most relevant example of this approach is provided by the price cap regulation for monopolistic firms, used, for instance, in the highways sector in some European countries. Under price cap, the price choice is delegated to the firm, under a set of constraints on their (average) level, which is meant to prevent the firm from exploiting its market power. A large array of instruments lies in between (and sometimes intersects with) these two forms of intervention, including rate-of-return regulation, procurement auctions, and concessions.

Finally, government intervention aimed at mitigating environmental or congestion externalities can also be broadly classified either as *direct* or *indirect* intervention. A *direct* type of intervention is for instance command-and-control regulation, which directly imposes a choice or a behavior to market participants; in other words, it directly intervenes on what market participants can and cannot do. On the other hand, *indirect* forms of government intervention are all those market-based regulation under which the government attenuates inefficiencies, for instance, by imposing an extra cost (such as a price or a tax) on those generating the externality, so as to modify their incentives, and to align their behavior to social efficiency.

Both command and control (C&C), and market-based regulation can be applied both to firms and to users. When directed to a firm, C&C may involve, for instance, setting an emission standard for pollutants in the energy sector and fuel economy standards in the automotive industry. When directed to consumers, C&C involves setting outright obligations, or, more frequently, prohibitions. Examples include precluding access to city centers to more polluting cars, or making a zone pedestrians-only. Market-based regulation applied to firms may involve, for instance, the emissions trading systems, which imposes an extra cost on the activity

of polluting firm, in an attempt to induce either less production (and thus pollution) or the adoption of cleaner technologies. When applied to users, they may involve congestion charge to disincentivize the use of congested infrastructures.

In this chapter, first start by considering government intervention related to market power. We then move to considering environmental and congestion externalities. We will keep an eye on how recent and prospective technological evolutions, in particular in terms of transition to electric and autonomous vehicles, have affected or have the potential to affect the scope and the structure of government intervention.

We argue that indirect policies are in general superior to direct policies in addressing externalities, while, when government intervention is geared at reducing market power, the relative performance of direct and indirect intervention is less clear cut and strongly depends on the institutional environment.

Market Power

Market power in the transport sector arises in almost the entire range of transportation modes since they displays scale economies, under which the unit average cost decreases as output increases.^a Typically, scale economies derive from the existence of large fixed cost, which are indeed prevalent in most transport industries. Prominent examples include highways, airport, ports. In industries where scale economies exist, there are limits to competition, in the sense that profitable operation is possible for a limited number of firms only (de Palma and Monardo, 2021).

It is well known that, in industries with such inherent limits to competition, the traditional perfectly competitive prices not only do not occur when the firm(s) operating in the market are left free to operate, but would also be impossible to apply by a social welfare maximizing entity, absent subsidies. If all costs need to be covered, prices equal marginal cost—the standard result of a perfectly competitive market—would be unable to cover all the firm's costs. A large body of literature has contributed to the identification of socially optimal prices in the presence of limits to competitions. These prices go under the heading of Ramsey prices and, in their simplest version, imply that, for each good, the departure from marginal cost pricing is larger the less elastic is the demand for that good (De Borger, 2021, ET; Guo, 2021). This pricing rule reflects the intuition that relatively higher prices are set for goods with relatively lower demand elasticity, since the little reduction in demand results in few foregone (but welfare-enhancing) trade opportunities.

The main issue to be tackled by a policy maker is then how to implement these optimal prices. The traditional solution is public provision, either by a governmental department or by a state-owned firm. Highways, railroads, ports, and airports are frequently publicly owned or publicly provided around the world.

Direct government ownership, however, is no guarantee of welfare-maximizing behavior, because of at least two types of distortions. First, principal-agent distortions. Even when the government directly manages the infrastructures, it must rely on a governmental operational department, composed of managers and of employees who are likely to have private information on the costs and benefits of the activities, and an objective function not fully in line with welfare maximization, but instead related to their private benefits. Second, political economy distortions. The government does not necessarily maximize welfare. Politicians might not respond to citizens' needs, especially when they feel their behavior is not adequately observed and scrutinized by the public (see Besley and Burgess, 2002, Strömberg, 2004, and Boffa et al., 2016).

An alternative approach to limit the distortions caused by market power is to leave the production activity in the hands of private firms and to regulate them. Economic regulation implies a much greater reliance on market forces, which may however vary across the different instruments used. It basically consists in the set of rules that (1) defines the market structure of the industry and (2) constrains the behavior of firm(s) operating in the industry.

In many industries, technological advancements have created the possibility of breaking up the vertical structure of the industry, separating the different stages of the supply chain. A typical example in the transport sector is the railway industry. While the operation of the railway infrastructure was previously seen as inherently connected to the provision of rail transportation services, these two segments of the industry are now considered to be separate.^b The effect of this change on the economic regulation of the industry is dramatic. While the infrastructure keeps the features of a natural monopoly, the technology of the commercial segment allows for the existence of an (imperfectly) competitive industry.

The government intervention in such situations may take different forms, which again show a different degree of reliance on market forces. The more traditional is clearly the one of preserving the vertical structure of the industry and to regulate it as fully integrated industry. By so doing, the government gives up the benefit of competition in some stages of the production process but, at the same time, allows the industry to exploit the existence of possible economies of scope, which make vertical integration more efficient. Alternatively, the government policy may be aimed at opening up to competition all those stages of production in which this is economically feasible. The main issue is then to define if and under what conditions the firm operating in the natural monopoly segment of the industry is allowed to operate also in the competitive segment(s). The solutions adopted here are very different, ranging from a full ban to subtler provisions. These may simply require legal or managerial separation (while maintaining common ownership) between the entities operating at the different stages of the vertical structure, possibly coupled with

^a More rigorously, fixed costs are associated to scale economies in the case of single-output technologies only. Conditions for the existence of scale economies for multiproduct firms are provided in Panzar (1989), which also provides conditions for the existence of a natural monopoly, based on subadditivity of costs, and for the existence of a natural oligopoly.

^b In the EU, see the so called Fourth Railway Package, comprising Regulations 2016/796, 2016/797, 2016/2337 and 2016/2338 and Directives 2016/798 and 2016/2370.

nondiscrimination constraint and (antitrust) controls, or impose some ex-ante pricing constraints on the essential input provided by the monopolist to the competitors in the competitive market.

Another possible approach to market design to correct market power distortions would involve relying on competitive forces not at the market stage but rather at the stage of selection of the firm(s) entitled to operate in the market (van de Delde and Hirshhorn, 2021). This is the case of the so-called competition “for the market,” as opposed to competition “in the market.” This approach is rather common in local transport, where the right to operate is awarded by means of a competitive tendering. The competitive awarding mechanism is supposed to ensure that production is undertaken by the most efficient firms and to extract through the franchisee fee the economic rent due to market power. In most cases, this approach is coupled with some additional forms of regulation, meant to limit this rent and prevent other distortions on strategic variables chosen by the franchisee that are of particular importance to the society.

Given the market structure, regulation then imposes a number of (ex-ante) constraints on the behavior of the firms(s) with market power. The vagueness of the term “behavior” is meant to include all aspects of a firm’s strategy, including prices, quality levels, product range, investments, customer relations, etc. For the sake of space, we will focus our discussion on price and quality regulation only.

Price Regulation Involves Two Dimensions: Level and Structure

The level of the regulated prices positively affects the firm’s profits but negatively consumer’s welfare. It is therefore the objective of price regulation that the level of prices is set to ensure profits just high enough to keep the firm in operation. The traditional solution of cost of service regulation (otherwise known as rate of return regulation), under which the price level is set to cover ex post all firm’s expenditures, is well known to suffer from great inefficiencies, often dubbed as gold-plating (Averch and Johnson, 1962). The “new economics of regulation,” having recognized these inefficiencies, has proposed the approach of incentive regulation. Under this approach, any rent left to the firm is used to incentivize it to pursue productive efficiency. In the case of optimal regulation, based on the truthful revelation of the firm’s private information on its costs, a high rent acts as a reward to efficient firms only (Laffont and Tirole, 1993). In the more applicable case of price cap regulation, the price level is set to prospectively ensure the firm’s breakeven for a limited period only (regulatory period), thus leaving to the firm any excess profit due to higher than anticipated efficiency. This excess profit is the cost for the society of giving to the regulate firm the incentive to efficiency (Littlechild, 1983).

The regulation of the price structure is probably the most prominent example of the difference between direct and indirect government intervention in the regulation of firms with market power. The traditional approach of cost of service regulation implied a close involvement of the regulator/public authority in the determination of the price structure. In practice, the price structure were jointly determined by the firm and the regulator/public authority, each pursuing its objectives. In the case of the regulator/public authority, these objectives were clearly manifolds, from efficiency to distributional, and often comprising evaluations of the external effects of the operation of the firm, from environmental to labor issues (Kahn, 1988).

Price cap regulation has represented a radical departure from this approach, in that the regulated firm only faces a set of price constraints imposed by the regulator. The main constraint caps the average of all prices, but it is often complemented by additional constraints on single prices or subsets of them. This leaves the firm free to decide independently its price structure. The delegation of the choice of the price structure to the firm illustrates an extreme example of reliance on market forces. Indeed, the basic motivation of this approach is that, in the absence of other distortions besides market power, the regulated firm is believed to have incentives aligned to those of the regulator as to the structure of prices. Intuitively, as in the case of Ramsey prices in the case of a public decision maker, the firms prefers to charge relatively higher prices for the goods whose demand is less elastic to prevent an excessive reduction in demand, which would reduce profits too much. (Bradley and Price, 1988; Vogelsang and Finsinger, 1979).

Another relevant object of behavior regulation is quality. As discussed before, firms with market power have a distinct incentive to provide an inadequate level of service quality. Also in this case, the regulatory provisions may rely on market forces in different ways (Sappington, 2005). At one extreme, zero reliance on market forces implies mandatory levels of service quality. In the case of transport, this is the approach typically used for some quality aspects in local transport service contracts, which include detailed standard of services, in some case coupled with damaged customer compensation when the standard is not met. At the other extreme, a quality index is included in the price cap level, as it happens, for instance, in the case of some highways concessionaires in Italy (Benfratello et al., 2009). As the level of quality provided by the price capped firm increases, the same firm is allowed to charge a higher average price level. This regulatory approach is based on market forces in that it is intended to mirror the mechanism, typical of competitive markets, under which a higher quality allows a firm to set higher prices (De Fraja and Iozzi, 2008).

Environmental Externalities

The contribution of the transportation sector to pollution is substantial. Total carbon dioxide emissions from road transportation exceed five gigatons annually, and the sector accounts for about 13% of global greenhouse gas emissions (Anderson and Sallee, 2016).

Mitigation can alternatively take the form of a reduction in the amount of transport, possibly by exploiting technological opportunities to increase work from home, or by a decrease in emissions per km traveled. In turn, decrease in emissions per km traveled can come through an environmentally-friendly improvement in the technology that moves vehicles, or through a more efficient use of the space on vehicles.

Given the current technological trends, the decrease in emissions per km traveled comes either from the improvement in mileage per unit of space of conventional fuels (unleaded and diesel), or from adoption of new technologies, among which the primary option appears to be the adoption of electric vehicles. It has to be emphasized that the impact on emissions of the development of electric vehicles crucially depends on the emissions in the electricity sector.

In this section, we will concentrate on the comparison between the predominant indirect and direct policies in the transport sector, respectively fuel tax and fuel standards, whereby the regulator directly sets a maximum mileage standard for each firm. In principle, emissions standards could also be tradable, thus becoming a hybrid between a direct and an indirect policy. However, in the current regulatory practice, they are generally nontradable, and we will consider these only in this review.

Fuel taxation has been ubiquitously adopted. Fuel standards have also been extensively used in the past. United States and Japan have had them for a long time now. More recently, other regions, including, notably the European Union, China and India, have joined them in the adoption. While the details in the implementation differ across countries, all the programs broadly impose to automakers a maximum average, usually weighted by sales, emissions rate (Davis and Knittel, 2019).

To compare fuel taxes and fuel standards, we follow closely Anderson and Saltee (2016), and we start with the analysis of the effects of a tax.

Cars, under a fuel tax, turn more efficient and shrink in size. This is consistent with empirical evidence in Busse et al. (2013) and Beresteanu and Li (2011), who show that a \$1 increase in fuel tax decreases average mileage per gallon by approximately 4%.

Second, consumers drive fewer miles. This is consistent with recent evidence in Gillingham and Nielsen (2019), who, using a rich dataset in Denmark, find that, on average, the elasticity of distance traveled to fuel prices is at about -0.30 (a value within the range provided by the literature that generally estimates an elasticity between -0.10 and -0.30, see Yang et al., 2020). Interestingly, the response to fuel prices comes primarily from very short or very long trips.

Third, consumers hold fewer cars, due to the increase in the costs of operating new cars, and they scrap inefficient cars earlier. In general, fuel prices are an important determinant of scrapping decisions. Jacobsen and Van Benthem (2015) estimate a -0.7 elasticity of scrapping to fuel prices.

We now move to fuel economy standards. While some of the effects of fuel standards are similar to those of a tax (for instance, cars turn more efficient and shrink in size), some other effects differ markedly. First, the standards generate the rebound effect. With more efficient cars, travelers are incentivized to travel more miles, which reduces the impact on emission reduction. Gillingham (2020), in his review, shows that the rebound effect eats up about 10% of the emissions reduction benefits from efficient cars. Second, fuel standards only reduce emissions for new cars, while a fuel tax, by affecting traveled distances, has an impact on used cars as well. Third, differently than with a tax, fuel standards do not provide incentives to own fewer cars, so that the automobile market size should not shrink, and, under some parametric specifications in Anderson and Saltee, might even increase. Fourth, relatively to a fuel tax, with a fuel standard consumers delay scrapping of inefficient old cars. Jacobsen and Van Benthem (2015) show that 13%–16% of the savings anticipated for new vehicles are offset by the delay in old vehicles scrapping.

All the above effects have been derived by considering an optimal (in a second-best sense) design of the fuel standard. In fact, most of the fuel standards implemented around the world are designed in a suboptimal way, which, as Anderson and Saltee point out, further induce to regard fuel tax as a more efficient form of regulation.

Possibly a result of all the above considerations, a survey administered to top economists (IMG, 2011), shows that 90% of them prefer a fuel tax over a fuel standard.

Congestion Externalities

One of the most relevant externalities in the transportation sector depends on congestion. The yearly congestion cost has been estimated to amount to more than one hundred billion dollars in the United States, and to be steadily increasing over time (Schrang et al., 2011).

However, congestion issues are even starker in developing and emerging economies. Based on real-time traffic data in 390 cities from 48 countries in 2016, TomTom Traffic Index shows that among the top 20 most congested cities (population over 0.8 million), all but 1 are located in developing and emerging economies with 8 of them being in China. Drivers in the most congested city spend much more time on average than they would have absent congestion, up to 66% or 227 hours per year per-capita in the extreme of Mexico City – the top city on the list. But, even if one considers Beijing, number 10 on the list, figures are still striking: 46% extra time on the road, or 179 hours per year per-capita due to congestion (Yang et al., 2020).

If not subject to an appropriate direct or indirect regulation, a driver does not fully consider the cost imposed by driving on a congested road. Part of this cost is external, since the driver also affects all the other drivers leaving after him. If not appropriately taken care of, congestion externality can severely affect welfare.

The mitigation of congestion, and of its resulting externality, may occur primarily in two ways. First, through an increase in road capacity, so as to accommodate the demand for cars, without creating congestion. Second, through a reduction in travel, either through an improvement in public transit, or through a better utilization of available capacity, for instance through car pooling. Observe that reduction in travel also achieves the goal of reducing environmental externality.

Based on evidence in Duranton and Turner (2011), increase in road capacity and increase in public transit usually do not prove effective tools at mitigating congestion. Increasing road capacity induces an increase in demand, originating from current residents, increases in commercial traffic, and migration that leaves congestion unaffected. Also, there is no evidence that provision of public

transportation affects travel, and, as a result, congestion. Having ruled out actions on the enhancement in infrastructures and in public transportation, we are left with the two alternatives of an indirect and direct policies to mitigate congestion, through a reduction of the number (or size) of vehicles on the road, or through their optimal allocation across time and space.

The standard indirect instrument to deal with congestion is road pricing. The simplest, and effective, form of road pricing is a Pigouvian tax equal to the non-internalized portion of the congestion cost, defined as *congestion charge*.

Pigouvian congestion charges achieve optimality when drivers do not factor in their decision the effects of their behavior on other drivers. Boffa et al. (2020) show that the taxation scheme to mitigate congestion will have to change as autonomous vehicles will replace conventional cars, and as fleets will gain prominence over privately owned vehicles. In the absence of policies or of market design interventions, congestion costs are indeed not bound to disappear when AVs will be deployed. However, when the mobility services will be provided by fleets of vehicles, the fleet operator will, at least partially, internalize the congestion costs (at least the portion that is imposed on travelers supplied by the same fleet), so the optimal tax will have to be tailored accordingly.

It is well-known that congestion charges, when optimally set, achieve optimality. However, while strongly advocated by economists, these are rarely implemented in practice (notable exceptions include London, Stockholm and Singapore), because of political economy reasons, given they are usually disliked by voters (Oberholzer-Gee and Weck-Hannemann, 2002).

When congestion charges cannot be implemented, it is possible to resort to direct policies, which, however, are certainly second best. An example is high occupancy vehicle lanes (HOV) policy, analyzed, among others, by Bento et al. (2013). They use 8 years of traffic flow data for 1700 locations in Los Angeles and show that the mean elasticity of flow with respect to fuel price for HOV is 0.136, which implies, for a 10% increase in fuel price, \$8.8 million per year in additional congestion costs for carpoolers and \$11.3 million lower costs for non carpool drivers. The estimates imply that, while HOV policy is clearly not first best, it is still preferable to a situation of no HOV lane.

Alternative direct congestion mitigation policies, such as, for instance, pedestrians-only area, are clearly suboptimal, at least from the perspective of pure congestion mitigation.

References

- Anderson, S., Sallee, J., 2016. Designing policies to make cars greener. *Ann. Rev. Resour. Econ.* 8, 157–180.
- Averch, H., Johnson, L., 1962. Behavior of the firm under regulatory constraint. *Am. Econ. Rev.* 52, 1052–1069.
- Benfratello, B., Iozzi, A., Valbonesi, P., 2009. Technology and incentive regulation in the Italian motorways industry. *J. Reg. Econ.* 35, 201–221.
- Bento, A., Hughes, J.E., Kaffine, D., 2013. Carpooling and driver responses to fuel price changes: Evidence from traffic flows in Los Angeles. *J. Urban Econ.* 77, 41–56.
- Beresteanu, A., Li, S., 2011. Gasoline prices, government support, and the demand for hybrid vehicles in the United States. *Int. Econ. Rev.* 52, 161–182.
- Besley, T., Burgess, R., 2002. The political economy of government responsiveness: Theory and evidence from India. *Quart. J. Econ.* 117 (4), 1415–1451.
- Boffa, F., Fedele, A., and Iozzi A. 2020. Congestion and Incentives in the Age of Driverless Cars. CEIS Tor Vergata Research Paper Series, 18(3), No. 484 – May 2020.
- Boffa, F., Piolatto, A., Ponzetto, G., 2016. Political centralization and government accountability. *Quart. J. Econ.* 131 (1), 381–422.
- Bradley, I., Price, C., 1988. The economic regulation of private industries by price constraints. *J. Indus. Econ.* 37, 99–106.
- Busse, M., Knittel, C., Zettelmeyer, F., 2013. Are consumers myopic? Evidence from new and used car purchases. *Am. Econ. Rev.* 103, 220–256.
- Coublucq, D., Ivaldi, M., McCullough, G.J., 2019. The static-dynamic efficiency trade-off in the US rail freight industry: Assessment of an open access policy. *Rev. Network Econ.* 17. (4).
- Davis, L., Knittel, C., 2019. Are fuel economy standards regressive? *J. Assoc. Environ. Resour. Econ.* 6 (S1), S37–S63.
- De Borger, B. 2021. Pricing principles in the Transport sector. In: *Encyclopedia of Transportation*, Elsevier.
- De Fraja, G., Iozzi, A., 2008. The quest for quality: A quality adjusted dynamic regulatory mechanism. *J. Econ. Manag. Strat.* 17 (4), 1011–1040.
- de Palma, A., Monaldo, J. 2021. Natural monopoly in Transport. in *Encyclopedia of Transportation*, Elsevier.
- Duranton, G., Turner, M., 2011. The fundamental law of road congestion: evidence from US Cities. *Am. Econ. Rev.* 101 (6), 2616–2652.
- Environment Protection Agency, 2020. Available from: <https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions>.
- EU Climate Action, 2020. Available from: https://ec.europa.eu/clima/policies/transport_en#tab-0-0.
- Franco, S.F., 2021. The concept of external cost: Marginal vs total cost and internalization. In: *Encyclopedia of Transportation*, Elsevier.
- Gillingham, K., Munk-Nielsen, A., 2019. A tale of two tails: commuting and the fuel price response in driving. *J. Urban Econ.* 109, 27–40.
- Gillingham, K., 2020. The rebound effect and the proposed rollback of US fuel economy standards. *Rev. Environ. Econ. Policy* 14 (1), 136–142.
- Guo, Q. 2021. Public transport fare and subsidy optimization. In: *Encyclopedia of Transportation*, Elsevier.
- Holland, S., Mansur, E., Muller, N., Yates, A., 2016. Are there environmental benefits from driving electric vehicles? *Imp. Local Fact.* 106 (12), 3700–3729.
- IMG, 2011. Available from <http://www.igmchicago.org/surveys/carbon-tax/>
- Initiative on Global Markets, IGM Economic Experts Panel, 2016. Available from: <http://www.igmchicago.org/igm-economic-experts-panel>.
- Institute for Policy Integrity, 2015. Expert Consensus on the Economics of Climate Change, Available from: <https://policyintegrity.org/files/publications/ExpertConsensusReport.pdf>.
- Jacobsen, M., van Benthem, A., 2015. Vehicle scrappage and gasoline policy. *Am. Econ. Rev.* 105 (3), 1312–1338.
- Kahn, A.E., 1988. *The Economics of Regulation: Principles and Institutions*. MIT Press, Boston (MA).
- Laffont, J.-J., Tirole, J., 1993. *A Theory of Incentives in Procurement and Regulation*. MIT Press, Boston (MA).
- Littlechild, S.C., 1983. Regulation of British Telecom's profitability. Report to the Secretary of State, Department of Industry, London.
- Parry, I., Small, A.K., 2009. Should urban transit subsidies be reduced? *Am. Econ. Rev.* 99 (3), 700–724.
- Oberholzer-Gee, F., Weck-Hannemann, H., 2002. Pricing road use: politico-economic and fairness considerations. *Transp. Res. Part D: Transp. Environ.* 7 (2), 357–371.
- Panzar, J.C., 1989. Technological determinants of firm and industry structure. In: Schmalensee, R., Willig, R. (Eds.), *Handbook of Industrial Organization*, Volume I. Amsterdam, North-Holland.
- Sappington, D.E.M., 2005. Regulating service quality: A survey. *J. Regulat. Econ.* 27, 123–154.
- Schrank, D., Lomax, T., and Eisele, B., 2011. TTI's 2011 Urban Mobility Report. Texas Transportation Institute, The Texas A&M University System.
- Strömberg, D., 2004. Radio's impact on public spending. *Quart. J. Econ.* 119 (1), 189–221.
- Tschofen, P., Azevedo, I., Muller, N., 2019. Fine particulate matter damages and value added in the US economy. *Proc. Natl. Acad. Sci. USA.* 116, 19587–19862.
- van de Velde D., Hirshhorn, F. 2021. Regulatory reform and competition in public transport. In: *Encyclopedia of Transportation*, Elsevier.
- Vogelsang, I., Finsinger, J., 1979. A regulatory adjustment process for optimal pricing by multiproduct monopoly firms. *Bell J. Econ.* 10 (1), 157–171.
- Yang, J., Purejav, A.O., Li, S., 2020. The marginal cost of traffic congestion and road pricing: evidence from a natural experiment in Beijing. *Am. Econ. J.: Econ. Policy* 12 (1), 418–453.

The Braess Paradox

Anna Nagurney*, **Ladimer S. Nagurney†**, *Department of Operations and Information Management, Isenberg School of Management, University of Massachusetts, Amherst, MA, United States; †Department of Electrical and Computer Engineering, University of Hartford, West Hartford, CT, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	601
The Classic Braess Paradox	602
Generalizations of the Classic Braess Paradox and Related Paradoxes in Transportation	603
Observations in Transportation Systems	605
Other Systems that May Exhibit Braess Paradox Behavior	605
References	606
Further Reading	607

Introduction

Congested urban transportation networks are examples of complex systems in which users, that is, travelers, interact with infrastructure. The behavior of the users is, typically, delineated as being either that of user-optimization or system-optimization based, respectively, on [Wardrop's, 1952](#) two principles of travel behavior. In the case of the user-optimization, each user "selfishly" selects his own optimal route of travel from an origin to a destination with an equilibrium being achieved when no user has any incentive to alter his travel path. The governing equilibrium conditions state that all used paths, that is, those with positive flow, connecting each origin/destination pair of nodes of travel will have equal and minimal travel cost. In the case of system-optimization, there exists a central controller that routes the traffic flow from origin nodes to destination nodes so that the total cost in the transportation network is minimized. Importantly, in congested transportation networks, the cost (usually reflecting travel time although the cost may be generalized to include monetary cost, risk, etc.) on a link is an increasing function of the volume of the traffic flow on the link.

In 1968, Dietrich Braess in his classic paper, written in German, identified the possibility of the occurrence of counterintuitive behavior in user-optimized transportation networks. The paper was inspired by a seminar delivered by W. Knoedel in Muenster in 1967 when Braess was 29 years old (see [Nagurney and Boyce, 2005](#)). Specifically, he constructed a transportation network example consisting of a single origin/destination pair of nodes and two parallel paths, such that, when expanded with the inclusion of an additional link, which provided another route option for the travelers, the result was an increase in travel cost for all the travelers! The demand and original link cost functions had not changed. This surprising discovery was in contrast to conventional wisdom in that the addition of a link, which yields another route option for travelers between their origin/destination pair, would make each user better-off in terms of travel cost/time. This counterintuitive phenomenon has become known as the Braess paradox.

What is also remarkable is that, as discussed in [Nagurney and Boyce \(2005\)](#), at the time that Braess described this paradoxical behavior, he was unaware of Wardrop's two principles of travel behavior and their rigorous mathematical formulation given in the book by [Beckmann et al. \(1956\)](#). Nevertheless, and this is also noteworthy, the Braess paper described two different concepts of traffic network utilization that correspond, respectively, to analogues of system-optimization and user-optimization. It is important to mention that the terms "user-optimization" and "system-optimization" were subsequently coined by [Dafermos and Sparrow \(1969\)](#). The [Braess \(1968\)](#) paper was followed by that of [Murchland \(1970\)](#), who elaborated upon the Braess paradox, reflected upon it in the context of [Beckmann et al. \(1956\)](#) and [Beckmann \(1967\)](#), and brought it to the attention of the English speaking community.

This paradox has come to fascinate researchers and practitioners in transportation and related fields, in which decentralized behavior in congested networks is relevant, such as in computer science, as in the case of the modeling of telecommunication networks and the Internet (see [Korilis et al., 1999](#); [Nagurney et al., 2007b](#); [Roughgarden, 2005](#); [Roughgarden and Tardos, 2002](#)); in electrical engineering, for the study of power systems ([Blumsack et al., 2007](#); [Cohen and Horowitz, 1991](#)) and electronic circuits (cf. [Nagurney and Nagurney, 2016](#)), as well as in physics in the case of mechanical ([Cohen and Horowitz, 1991](#)) and fluid systems ([Calvert and Keady, 1993](#)); in biology, as in metabolic networks (see [Motter, 2010](#)), ecosystems ([Sahasrabudhe and Motter, 2011](#)), and targeted cancer therapy ([Kippenberger et al., 2016](#)), and, surprisingly, in sports analytics as in the study of sports teams, where the Braess paradox analogue corresponds to the removal of a player resulting in better team performance (cf. [Gudmundsson and Horton, 2017](#); [Skinner, 2011](#)). Because of the interest from multiple disciplines, a translation of the paper from German to English was published by [Braess et al. \(2005\)](#). The foreword by [Nagurney and Boyce \(2005\)](#) to the translated paper contains additional background of how Braess came up with the counterintuitive phenomenon, along with clarifications of some of the concepts and terms.

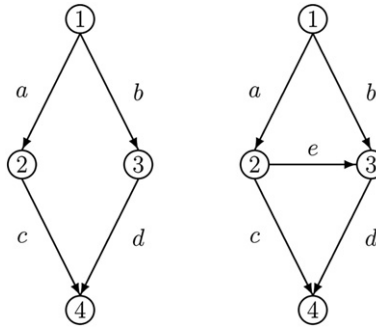


Figure 1 The Braess paradox example topology.

The Classic Braess Paradox

The classic Braess paradox example is as follows. Consider first the four-node transportation network as illustrated by the left-most network in **Fig. 1**.

There is a single origin/destination (O/D) pair of nodes $w = (1, 4)$. A traveler on this network can travel on one of two paths: on path p_1 consisting of the links: a, c or on path p_2 consisting of the links: b, d . We denote the flows on the links by: f_a, f_b , and so on, and their respective user link costs by: c_a, c_b , etc. Specifically, in this network the user link cost functions are: $c_a(f_a) = 10f_a, c_b(f_b) = f_b + 50, c_c(f_c) = f_c + 50, c_d(f_d) = 10f_d$.

Let the travel demand d_w be 6, which represents 6 vehicles of travel per unit time. We denote the flow on a path p by x_p so here we have x_{p_1} and x_{p_2} .

The conservation of flow equations ensure that the demand for each O/D pair is satisfied by the flows on paths that connect the O/D pair and that the link flows capture the flows that utilize the particular link. Specifically, the equations state that the sum of flows on paths connecting each O/D pair must be equal to the demand for that O/D pair and that the flow on a link must be equal to the sum of flows on the paths that contain/utilize that link.

Clearly, under user-optimizing behavior, the resulting equilibrium path flow for the first network is: $x_{p_1}^* = x_{p_2}^* = 3$, with incurred user path costs of: $C_{p_1} = C_{p_2} = 83$. No traveler has any incentive to switch her path since that would result in a higher cost for her. This result is apparent since the costs on the two paths are: $C_{p_1} = c_a + c_c = 10f_a + f_c + 50$ and $C_{p_2} = c_b + c_d = f_b + 50 + 10f_d$ and, therefore, the travelers at a demand of 6 will equally distribute themselves between the two paths, yielding the equilibrium path flows: $x_{p_1}^* = 3$ and $x_{p_2}^* = 3$ and incurred equilibrium flows: $f_a^* = f_b^* = f_c^* = f_d^* = 3$, with user link costs: $c_a = c_d = 30$ and $c_b = c_c = 53$.

Consider now the addition of a new link, e , joining node 2 with node 3 as in the right-most network in **Fig. 1**. Let the user link cost c_e on link e be

$$c_e(f_e) = f_e + 10.$$

A user on the expanded transportation network now has three path options: the original two paths, p_1 and p_2 , plus the new path $p_3 = (a, e, d)$. The equilibrium flow pattern on the first network will no longer yield an equilibrium for the second network. Indeed, observe that although the costs on paths p_1 and p_3 would be 83, the cost on the new path, C_{p_3} , if it is not used, that is, it has zero flow, would be 70. Clearly, some travelers, under user-optimization, would switch from paths p_1 and p_2 to path p_3 since the cost on path p_3 is less than 83.

Typically, we apply algorithms (cf. Nagurney, 1999; Patriksson, 2004) to determine the user-optimized/equilibrium flows in transportation networks, since they are usually large-scale and a priori it is difficult to identify which paths will and will not be used. In this example, because of its size, we can calculate the solution explicitly. We will assume that all paths are used. Hence, we can set up a system of equations making use of the following:

$$C_{p_1} = C_{p_2} = C_{p_3}$$

and

$$x_{p_1}^* + x_{p_2}^* + x_{p_3}^* = d_w = 6.$$

Furthermore, we know that according to the conservation of flow equations for the link and path flows:

$$f_a^* = x_{p_1}^* + x_{p_3}^*$$

$$f_b^* = x_{p_2}^*$$

$$f_c^* = x_{p_1}^*$$

$$f_d^* = x_{p_2}^* + x_{p_3}^*$$

$$f_e^* = x_{p_3}^*.$$

Hence, we can rewrite that user costs on paths as functions of path flows as follows

$$C_{p_1} = 10(x_{p_1}^* + x_{p_3}^*) + x_{p_1}^* + 50 = 11x_{p_1}^* + 10x_{p_3}^* + 50,$$

$$C_{p_2} = x_{p_2}^* + 50 + 10(x_{p_2}^* + x_{p_3}^*) = 11x_{p_2}^* + 10x_{p_3}^* + 50,$$

$$C_{p_3} = 10(x_{p_1}^* + x_{p_3}^*) + x_{p_3}^* + 10 + 10(x_{p_2}^* + x_{p_3}^*) = 10x_{p_1}^* + 10x_{p_2}^* + 21x_{p_3}^* + 10.$$

Using then the demand conservation of flow equation above and substituting, we obtain a system of equations, the solution of which yields the equilibrium flow pattern on the expanded network of: $x_{p_1}^* = x_{p_2}^* = x_{p_3}^* = 2$, with the incurred equilibrium path travel costs:

$$C_{p_1} = C_{p_2} = C_{p_3} = 92!$$

Thus, the addition of a new link makes every user worse-off in that each traveler in the expanded network incurs a higher travel cost than before!

It is important to note that the Braess paradox can only occur under user-optimization and never under system-optimization. Indeed, under system-optimization, in the second network in Fig. 1, only the original paths p_1 and p_2 would be used. It is important to emphasize that tolls can be assigned so that the system-optimizing flow pattern is at the same time user-optimizing; see Dafermos and Sparrow (1971) and Bergendorff et al. (1997). In particular, if one assigns link tolls as follows: toll on link a , $r_a=30$; toll on link b , $r_b=30$, with link tolls: $r_c=3$, $r_d=30$, and $r_e=30$, then travelers on the second network in Fig. 1 will independently distribute themselves according to the system-optimized flow pattern. The formula for tolls, in this case, due to Dafermos and Sparrow (1971), is $r_l = \hat{c}_l(f_l) - c_l(f_l)$ with the marginal total cost $\hat{c}_l$ and the user link cost c_l evaluated at the system-optimized flow pattern for all links l in the network and $\hat{c}_l = c_l \times f_l$ corresponding to the total cost on link l .

Generalizations of the Classic Braess Paradox and Related Paradoxes in Transportation

This counterintuitive example has given rise to many questions and the examination of under what conditions and scenarios the Braess paradox may arise. For example, in the classic example the travel demand $d_w=6$. Would the Braess paradox still occur under different values for the travel demand? Pas and Principio (1997) addressed this question and showed that the Braess paradox occurs only if the demand for travel falls within a certain intermediate range of values, specifically, if $2.58 < d_w < 8.89$. Interestingly, under low levels of demand only the new path p_3 is used and the paradox does not occur, whereas at higher levels of demand only the two original paths are used and the Braess paradox also does not occur. Subsequently, Nagurney et al. (2007b), utilizing connections between transportation and telecommunication networks, constructed a dynamic model of the Internet using evolutionary variational inequalities, and cast (cf. Fig. 2) the Pas and Principio results into a Braess paradox with time-varying demands, establishing the same results, but showed also that the curve of path flow equilibria is unique and that the equilibrium trajectories are continuous. Pas and Principio (1997) also showed, when examining linear, separable user link cost functions of the form as in the classic Braess paradox that, whether or not the paradox occurs, depends on the conditions of the problem; namely, the parameters of the link user cost functions.

Pas and Principio (1997) further suggested that, under higher levels of demand, the Braess paradox may not occur. Subsequently, Nagurney (2010) considered more general user link cost functions, which could be nonlinear, as well as asymmetric in that $\partial c_a(f)/\partial f_b \neq \partial c_b(f)/\partial f_a$, for all links a, b in the network, where f is the vector of link flows. In her paper, she considered the hypothesis that, in congested networks, the Braess paradox may “disappear” under higher demands, and proved this hypothesis by deriving a formula that provides the increase in demand that will guarantee that the addition of that new route will no longer increase travel cost since the new path will no longer be used. This result was established for any network in which the Braess paradox originally occurs and suggests that, in the case of congested, noncooperative networks, of which transportation networks are a prime example, a higher demand will negate the Braess paradox. At the same time, this finding shows that extreme caution should be taken in the design of network infrastructure, including transportation networks, since at higher demands, new routes/pathways may not even be used.

Steinberg and Zangwill (1983) considered linear separable user link cost functions and provided necessary and sufficient conditions, under reasonable assumptions, for the Braess paradox to occur in a general transportation network. They concluded that the Braess paradox is about as likely to occur as not occur.

While the classical example of the Braess paradox uses cost functions that are of the form: a fixed term plus a term proportional to the flow, other possible cost functions have been mathematically investigated in transportation networks, including the Bureau of Public Roads cost function (see Sheffi, 1985), which has a term quartic in the flow. The Braess paradox has been examined under such functions by LeBlanc (1975), Frank (1981), and Bloy (2007). Dafermos and Nagurney (1984a) demonstrated how, in terms of

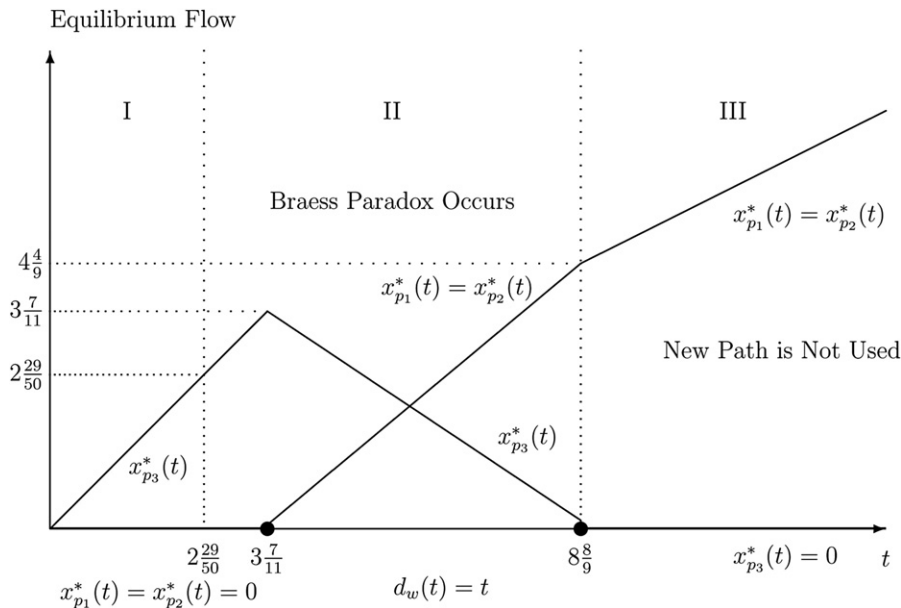


Figure 2 Equilibrium trajectories of the Braess network with time-dependent demand.

traffic networks with general (asymmetric) user link cost functions, the addition of a route connecting an origin-destination pair that shared no links with any other route in the network, could never result in the Braess paradox. Moreover, the authors provided explicit formulae for the effects of cost function and demand changes on the incurred equilibrium path flows and path travel costs. The generalization of user link cost functions from separable, as well as symmetric ones, to asymmetric ones, made use of the formulation of the governing equilibrium conditions as a variational inequality problem (Dafermos, 1980; Nagurney, 1999; Smith, 1979). User-optimized networks with symmetric user link cost functions, in contrast, could be reformulated and solved as optimization problems (cf. Beckmann et al., 1956; Dafermos and Sparrow, 1969).

Hallefjord et al. (1994), in turn, presented paradoxes in transportation networks in the case of elastic demand, rather than fixed demands as in the original Braess paradox example. The authors noted that, in the case of elastic travel demand, it is not apparent what a paradoxical situation is, and in the elastic demand case there is a need for characterizations of different paradoxes. An example is provided in which total flow (demand) decreases while travel time increases due to the addition of a new link to the network, with this being a rather extreme type of paradox. Another highlighted paradox is when the network “improvement” leads to a reduction in the social surplus.

As noted in Boyce et al. (2005), sensitivity analysis is also essential to the effective planning/design of transportation networks and the Braess paradox motivated much of the subsequent research in sensitivity analysis and networks. Dafermos and Nagurney (1984b), for example, used the variational inequality formulation of traffic network equilibrium with fixed demands to provide directional effects of link cost function changes and to demonstrate that small changes in the data yielded small changes in the resulting equilibrium link flows. For other sensitivity analysis results in the context of traffic network equilibrium problems, see, e.g., Tobin and Friesz, 1988; Frank, 1992; Yang, 1997; Patriksson, 2004, and the references therein.

It is worth pointing out additional related paradoxes to that of the Braess paradox in transportation. Sheffi and Daganzo (1978) presented a counterintuitive result that may occur when stochastic traffic assignment methods are used; see also Yao and Chen (2014). Fisk (1979) constructed examples showing that both origin to destination and global travel costs may decrease for an O/D pair as a result of an increase in travel demands for another O/D pair. Subsequently, Fisk and Pallotino (1981) provided network examples for the city of Winnipeg, demonstrating that the Braess paradox could occur in real world networks. Yang and Bell (1998) introduced a new paradox associated with network design problems via a simple network example in which the addition of a new road segment to a road network may reduce the potential capacity of the network. Nagurney (2000) identified emission paradoxes; in particular, three distinct ones that can occur in congested urban transportation networks in terms of the total emissions generated. These emission paradoxes reveal that so-called improvements to the transportation network may result in increases in total emissions generated.

Zhang et al. (2016) considered the Downs–Thomson paradox. The Downs–Thomson paradox, named after Downs (1962) and Thomson (1977), suggests that highway capacity expansion may produce counterproductive effects in a two-mode (auto and transit) transportation system. Specifically, Zhang et al. (2016) reexamined the paradox when certain assumptions are relaxed while retaining the usual assumption that there is no congestion interaction between the modes. Arnott and Small (1994) reviewed the Downs–Thomson paradox and also explicated the Pigou–Knight–Downs paradox, in which expanding road capacity may elicit its own demand with no improvement in congestion.

Observations in Transportation Systems

While it may be challenging to accurately control the demand and to measure the travel time (cost) on real road systems with actual drivers, there are several documented examples of the “inverse” of the Braess paradox being observed in a transportation network after a link was removed. In other words, travel time improves after the removal of a link. Many of these examples/instances have been written up in the popular press.

For example, [Kolata \(1990\)](#), writing in *The New York Times*, noted that, in 1990, on Earth Day, the New York City’s Transportation Commissioner decided to close 42nd Street, and to “everyone’s surprise” no historic traffic jam was generated and traffic flow actually improved. She stated that this may be a real-world example of the Braess paradox. That same year, [Cohen and Kelly \(1990\)](#) constructed an example where a Braess-type paradox occurs in a queuing network. The authors cited the paper by [Knodel \(1969\)](#) that noted that the City of Stuttgart had tried to ease downtown traffic by adding a new street. However, congestion only worsened, and hence, in desperation, the authorities closed the street. The result was that the traffic flow improved.

In 1999, according to [Vidal \(2006\)](#), one of the three main traffic tunnels in Seoul, the capital of South Korea, was closed for maintenance. Surprisingly, the result was not chaos and traffic jams, but, rather, the traffic flows improved. Inspired by their experience, Seoul’s city planners, subsequently, demolished a major motorway leading into the heart of the city and experienced the same strange result, with the added benefit of creating a 5-mile long, 1000 acre park for the local inhabitants (see also [Baker, 2009](#)).

And, in 2009, in an ambitious project in NYC, a part of Broadway in mid-Manhattan was converted to a pedestrian plaza and vehicular travel banned (see [Grynbaum, 2010](#)). This redesign of infrastructure was made permanent and has lasted even past the original mayoral administration of Michael Bloomberg. Traffic flows, in parts, improved.

Other Systems that May Exhibit Braess Paradox Behavior

There exists a plethora of realizable physical systems that may exhibit Braess paradox behavior and, in certain such physical systems, the demand (total flow) may be controllable and the laws of physics ensure that the decentralized network is, in fact, user-optimized. In contrast to congested transportation networks, users are no longer travelers, but correspond to electrons in electric power systems, or to fluid molecules (water, oil, etc.) in the case of pipeline networks, etc. Such analogues of transportation networks are quite natural and we note that [Beckmann et al. \(1956\)](#) hypothesized that electric power generation and distribution networks would behave like congested urban transportation networks. This hypothesis was substantiated by [Nagurney et al. \(2007a\)](#); see also the references therein. Moreover, fluid flow models have been developed for traffic; see, e.g., [Lighthill and Whitham \(1955\)](#), [Herman et al. \(1959\)](#), and [Herman and Prigogine \(1979\)](#).

For example, [Cohen and Horowitz \(1991\)](#) suggested that it might be possible to create mechanical, electrical, fluid, and thermal systems that exhibit counterintuitive behavior when a physical component was added. As an illustration of such a mechanical system, they showed that a weight hanging from a coupled pair of springs with safety strings can rise, rather than fall, when the taut coupling string is cut. They also demonstrated that an electrical network with a topology of the [Braess \(1968\)](#) example and consisting of ideal passive components (resistors and zener diodes) can exhibit the counterintuitive behavior of the voltage rising across the network when an additional branch (link) is added.

Details of a spring network that exhibited the Braess paradox were delineated by [Penchina and Penchina \(2003\)](#). The authors also noted that the only requirement for the spring network to exhibit the paradox is that the springs must shrink more than the safety strings stretch. [Peters and Vondracek \(2012\)](#) extended the experiments to include variation in the mass and the spring constant.

[Witthaut and Timme \(2012\)](#) studied the addition of single links in a class of oscillator networks that model modern power grids on coarse scales. They showed that, while, on the average, the additional links stabilized the network, the addition of specific new links could decrease the total grid capacity and, thereby, decrease or even destroy the stability of the grid. In addressing the question of reliability of the power grid, [Blumsack et al. \(2007\)](#) noted that in a power network with the classic Braess topology, adding a line (link) might result in another line becoming capacity limited, which would affect the optimal power dispatch and result in higher costs.

The idealized electrical circuit, as described by [Cohen and Horowitz \(1991\)](#), was converted to a real electrical circuit by [Nagurney and Nagurney \(2016\)](#), who used electric circuit theory to develop matrix equations to describe the voltage drop (equivalent to cost) as a function of the circuit elements and current flow (demand). The authors then used this formulation to build a circuit using standard electrical components in which the voltage and currents could be measured. In addition to constructing an actual physical circuit whose voltage drops were functionally similar to the cost functions suggested by Braess in his classic example, [Nagurney and Nagurney \(2016\)](#) also constructed and measured the parameters of a circuit with more general voltage drops that exhibited behavior consistent with the Braess paradox.

The concept of the Braess paradox was also investigated by [Pala et al. \(2012\)](#) in semiconductor networks, whose transport properties are governed by quantum physics. The authors demonstrated theoretically that congestion plays a key role in the occurrence of the Braess paradox in such networks.

Both macro- and microfluidic systems can exhibit behavior analogous to the Braess paradox. [Ayala and Blumsack \(2013\)](#) considered a macrofluidic system, as in natural gas distribution networks, and studied the existence of paradoxical effects, relevant to network design. In their analysis, the simple addition of a pipe to transport gas may not necessarily increase the ultimate

transmission capacity. In terms of a microfluidic system, Case et al. (2019) showed that both the flow rate in a system with the classic Braess topology and the direction of the flow in the linking channel are dependent on the input pressure.

It is clear that, more than 50 years since the publication of the paper by Braess (1968), the paradox and the associated gleaned insights into decentralized noncooperative behavior remain relevant, continue to fascinate, and to inspire research in transportation. Its applicability to practice also continues to this day, in transportation network planning and design. Finally, the Braess paradox has served as a bridge for broadening perspectives in other scientific disciplines by enabling the advancement of the theory of the behavior of complex network systems with a vast range of important applications.

References

- Arnett, R., Small, K., 1994. The economics of traffic congestion. *Am. Sci.* 82, 446–455.
- Ayala, L., Blumsack, S., 2013. The Braess paradox and its impact on natural-gas-network performance. *Oil Gas Facil. J.* 2 (3), 52–64.
- Baker, L., 2009. Removing Roads and Traffic Lights Speeds Urban Travel: Urban Travel is Slow and Inefficient, in Part because Drivers Act in Self-interested Ways. *Sci. Am.* 300 (2), 20–22.
- Beckmann, M.J., 1967. On the theory of traffic flows in networks. *Traffic Quart.* 21, 109–116.
- Beckmann, M., McGuire, C.B., Winsten, C.B., 1956. *Studies in the Economics of Transportation*. Yale University Press, New Haven, CT.
- Bergendorff, P., Hearn, D.W., Ramana, M.V., 1997. Congestion toll pricing of traffic networks. In: Pardalos, P.M., Hearn, D.W., Hager, W.W. (Eds.), *Network Optimization*, Lecture Notes in Economics, Mathematical Systems Book Series (LNE, vol. 450), Springer, Heidelberg, Germany, pp. 51–71.
- Bloy, L.A.K., 2007. *An Investigation into Braess' Paradox*. University of South Africa, Pretoria (Master of Science Thesis).
- Blumsack, S., Lave, L.B., Ilic, M., 2007. A quantitative analysis of the relationship between congestion and reliability in electric power networks. *Energy J.* 28 (4), 73–100.
- Boyce, D.E., Mahmassani, H.S., Nagurney, A., 2005. A retrospective on Beckmann, McGuire and Winsten's *Studies in the Economics of Transportation*. *Pap. Reg. Sci.* 84 (1), 85–103.
- Braess, D., 1968. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 12, 258–268.
- Braess, D., Nagurney, A., Wakolbinger, T., 2005. On a paradox of traffic planning. *Transport. Sci.* 39, 446–450.
- Calvert, B., Keady, G., 1993. Braess's paradox and power-law nonlinearities in networks. *J. Aust. Math. Soc. B* 35, 1–2.
- Case, D.J., Liu, Y., Kiss, I.Z., Angilella, J.-R., Motter, A.E., 2019. Braess's paradox and programmable behaviour in microfluidic networks. *Nature* 574, 647–652.
- Cohen, J.E., Horowitz, P., 1991. Paradoxical behaviour of mechanical and electrical networks. *Nature* 352, 699–701.
- Cohen, J.E., Kelly, F.P., 1990. A paradox of congestion in a queueing network. *J. Appl. Probab.* 27 (3), 730–734.
- Dafermos, S., 1980. Traffic equilibrium and variational inequalities. *Transport. Sci.* 14, 42–54.
- Dafermos, S., Nagurney, A., 1984a. On some traffic equilibrium theory paradoxes. *Transport. Res. B* 18, 101–110.
- Dafermos, S., Nagurney, A., 1984b. Sensitivity analysis for the asymmetric network equilibrium problem. *Math. Program.* 28, 174–184.
- Dafermos, S., Sparrow, F.T., 1971. Optimal resource allocation and toll patterns in user-optimized transport networks. *J. Transport Econ. Policy* 5 (2), 184–200.
- Dafermos, S.C., Sparrow, F.T., 1969. The traffic assignment problem for a general network. *J. Res. Natl. Bur. Stand.* 73B, 91–118.
- Downs, A., 1962. The law of peak-hour expressway congestion. *Traffic Quart.* 16, 393–409.
- Fisk, C., 1979. More paradoxes in the equilibrium assignment problem. *Transport. Res.* 13B, 305–309.
- Fisk, C., Pallotino, S., 1981. Empirical evidence for equilibrium paradoxes with implications for optimal planning strategies. *Transport. Res.* 15A, 245–248.
- Frank, M., 1981. The Braess paradox. *Math. Program.* 20, 283–302.
- Frank, M., 1992. Obtaining network cost(s) from one link's output. *Transport. Sci.* 26 (1), 27–35.
- Grynbaum M.M. Broadway is busy, with pedestrians, if not car traffic. *The New York Times*. 2010, September 5.
- Gudmundsson, J., Horton, T., 2017. Spatio-temporal analysis of team sports. *ACM Comput. Surv. (CSUR)* 50, 22, doi:10.1145/3054132.
- Hallefjord, A., Jornsten, K., Storoy, S., 1994. Traffic equilibrium paradoxes when travel demand is elastic. *Asia Pac. J. Oper. Res.* 11 (1), 41–50.
- Herman, R., Prigogine, I., 1979. A two fluid approach to town traffic. *Science* 204, 148–151.
- Herman, R., Montroll, E.W., Potts, R.B., Rothery, R.W., 1959. Traffic dynamics: analysis of stability in car following. *Oper. Res.* 7, 86–106.
- Kippenberger, S., Meissner, M., Kaufmann, R., Hrgovic, I., Zöller, N., Kleemann, J., 2016. Tumor neoangiogenesis and flow congestion: a parallel to the Braess paradox? *Circ. Res.* 119, 711–713.
- Knodel, W., 1969. *Graphentheoretische Methoden und ihre Anwendungen*. Springer-Verlag, New York.
- Kolata G. What if they closed 42d street and nobody noticed? *The New York Times*. 1990, December 25.
- Korilis, Y.A., Lazar, A.A., Orda, A., 1999. Avoiding the Braess paradox in non-cooperative networks. *J. Appl. Probab.* 36, 211–222.
- LeBlanc, L.J., 1975. An algorithm for the discrete network design problem. *Transport. Sci.* 9, 183–199.
- Lighthill, M.J., Whitham, G.B., 1955. On kinematic waves II: a theory of traffic flow on long, crowded roads. *Proc. R. Soc. Lond. Ser. A* 229, 317–345.
- Motter, A.E., 2010. Improved network performance via antagonism: from synthetic rescues to multi-drug combinations. *BioEssays* 32 (3), 236–245.
- Murchland, J.D., 1970. Braess's paradox of traffic flow. *Transport. Res.* 4, 391–394.
- Nagurney, A., 1999. *Network Economics: A Variational Inequality Approach* second and revised ed. Kluwer Academic Publishers, Boston, MA.
- Nagurney, A., 2000. Congested urban transportation networks and emission paradoxes. *Transport. Res. D* 5 (2), 145–151.
- Nagurney, A., Boyce, D., 2005. Preface to "On a paradox of traffic planning". *Transport. Sci.* 39, 443–445.
- Nagurney, L.S., Nagurney, A., 2016. Physical proof of the occurrence of the Braess paradox in electrical circuits. *Europhys. Lett.* 115, 28004.
- Nagurney, A., Liu, Z., Cojocaru, M.-G., Daniele, P., 2007a. Dynamic electric power supply chains and transportation networks: an evolutionary variational inequality formulation. *Transport. Res. E* 43, 624–646.
- Nagurney, A., Parkes, D., Daniele, P., 2007b. The Internet, evolutionary variational inequalities, and the time-dependent Braess paradox. *Comput. Manag. Sci.* 4, 355–375.
- Nagurney, A., 2010. The negation of the Braess Paradox as demand increases: The wisdom of crowds in transportation networks. *Europhys. Lett.* 91, 48002.
- Pala, M., Sellier, H., Hackens, B., Martins, F., Bayot, V., Huant, S., 2012. A new transport phenomenon in nanostructures: a mesoscopic analog of the Braess paradox encountered in road networks. *Nanoscale Res. Lett.* 7, 472.
- Pas, E.I., Principio, S.L., 1997. Braess' paradox: some new insights. *Transport. Res. B* 31 (3), 265–276.
- Patriksson, M., 2004. Sensitivity analysis of traffic equilibria. *Transport. Sci.* 38 (3), 258–281.
- Penchina, C.M., Penchina, L.J., 2003. The Braess paradox in mechanical, traffic, and other networks. *Am. J. Phys.* 71 (5), 479–482.
- Peters, S., Vondracek, M., 2012. Counterintuitive behavior in mechanical networks. *Phys. Teach.* 50, 359–362.
- Roughgarden, T., 2005. *Selfish Routing and the Price of Anarchy*. MIT Press, Cambridge, MA.
- Roughgarden, T., Tardos, E., 2002. How bad is selfish routing? *J. ACM* 49 (2), 236–259.
- Sahasrabudhe, S., Motter, A.E., 2011. Rescuing ecosystems from extinction cascades through compensatory perturbations. *Nat. Commun.* 2, 170.
- Sheffi, Y., 1985. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice-Hall, Englewood Cliffs, NJ.

- Sheffi, Y., Daganzo, C.F., 1978. Another 'Paradox' of traffic flow. *Transport. Res.* 12 (1), 43–46.
- Skinner, B., 2011. The price of anarchy in basketball. *J. Quant. Anal. Sports* 6 (1), doi:10.2202/1559-0410.1217.
- Smith, M.J., 1979. Existence, uniqueness and stability of traffic equilibria. *Transport. Res. B* 13, 259–304.
- Steinberg, R., Zangwill, W.I., 1983. Prevalence of Braess' paradox. *Transport. Sci.* 17, 301–318.
- Thomson, J.M., 1977. *Great Cities and Their Traffic*. Gollancz, London, England.
- Tobin, R.L., Friesz, T.L., 1988. Sensitivity analysis for equilibrium network flow. *Transport. Sci.* 22 (4), 242–250.
- Vidal, J., 2006, November 1. Heart and soul of the city. *The Guardian*.
- Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. *Proc. Inst. Civil Eng.* 11 1 (2), 325–378.
- Witthaut, D., Timme, M., 2012. Braess's paradox in oscillator networks, desynchronization and power outage. *New J. Phys.* 14, 083036.
- Yang, H., 1997. Sensitivity analysis for the elastic-demand network equilibrium problem with applications. *Transport. Res. B* 31 (1), 55–70.
- Yang, H., Bell, M.G., 1998. A capacity paradox in network design and how to avoid it. *Transport. Res. A* 32 (7), 539–545.
- Yao, J., Chen, A., 2014. An analysis of logit and weibit route choices in stochastic assignment paradox. *Transport. Res. B* 69, 31–49.
- Zhang, F., Lindsey, R., Yang, H., 2016. The Downs-Thomson paradox with imperfect mode substitutes and alternative transit administration regimes. *Transport. Res. B* 86, 104–127.

Further Reading

- Alvarez, J., 2015. Want Less Traffic? Build Fewer Roads! Retrieved from <https://plus.maths.org/content/want-less-traffic-build-fewer-roads>
- America Revealed, Gridlock, 2012, Retrieved from <https://www.pbs.org/video/america-revealed-gridlock/>.
- Baker, L., 2009. Removing roads and traffic lights speeds urban travel: urban travel is slow and inefficient, in part because drivers act in self-interested ways. *Sci. Am.* 300 (2), 20–22.
- Chen, W., 2016. Bad traffic? Blame Braess' paradox. *Forbes*, Retrieved from <https://www.forbes.com/sites/quora/2016/10/20/bad-traffic-blame-braess-paradox/#7a5a0ca714b5>.
- Eriksson, K., Eliasson, J., 2019. The chicken Braess paradox. *Math. Mag.* 92 (3), 213–221.
- Hayes, B., 2015. Playing in traffic. *Am. Sci.* 103, 260–263.
- Merlone, U., DalForno, A., 2016., The Braess Paradox. Retrieved from <https://youtu.be/sTQAU9TW4jM>
- Nagurney, A., The Braess Paradox. Retrieved from <https://supernet.isenberg.umass.edu/braess/braess-new.html>
- Patriksson, M., 1994. *The Traffic Assignment Problem: Models and Methods*. VSP, Utrecht, The Netherlands.
- Rapoport, A., Kugler, T., Dugar, S., Gisches, E.J., 2009. Choice of routes in congested traffic networks: experimental tests of the Braess paradox. *Games Econ. Behav.* 65, 538–571.

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

Page left intentionally blank

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

EDITOR-IN-CHIEF

Roger Vickerman

*School of Economics, University of Kent, Canterbury, United Kingdom
and*

Transport Strategy Centre, Imperial College, London, United Kingdom

VOLUME 2

Transport Safety and Security

SECTION EDITORS

Per Gårder

*Department of Civil and Environmental Engineering,
University of Maine, Orono, ME, United States*



Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2021 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-08-102671-7

For information on all publications visit our
website at <http://store.elsevier.com>



Publisher: Oliver Walter
Acquisitions Editor: Oliver Walter
Content Project Manager: Natalie Lovell
Associate Content Project Manager: Manisha K and Ramalakshmi Boobalan
Designer: Matthew Limbert

EDITORIAL BOARD

Editor in Chief

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Section Editors

Maria Börjesson

Professor of Economics, VTI Swedish National Road and Transport Research Institute; Affiliated Professor at Linköping University, Sweden

Per Gårder

Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States

Prof. Kevin P.B. Cullinane

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

Prof. Edward C.S. Chung

Department of Electrical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

Chandra R. Bhat

Center for Transportation Research (CTR), The University of Texas at Austin, TX, United States

Edoardo Marcucci

Department of Political Sciences, University of Roma Tre, Rome, Italy; Department of Logistics, Molde University College, Molde, Norway

Prof. Maria Attard

Department of Geography, Faculty of Arts, University of Malta, Msida, Malta

Prof. Carlo G. Prato

School of Civil Engineering, The University of Queensland, Brisbane, Australia

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Page left intentionally blank

INTRODUCTION



Roger Vickerman

In an increasingly globalized world, despite reductions in costs and time, transportation has become even more important as a facilitator of economic and human interaction; this is reflected in technical advances in transportation systems, increasing interest in how transportation interacts with society and the need to provide novel approaches to understand its impacts. This has become particularly acute with the impact that Covid-19 has had on transportation across the world, at local, national, and international levels. This Encyclopedia brings a cross-cutting and integrated approach to all aspects of transportation from work in many disciplinary fields, engineering, operations research, economics, geography, and sociology to understand the changes taking place.

Transportation is both influenced by, and an influencer of, changes in the economy and society. Increasing speeds have reduced journey times and made the world a smaller place as globalization has affected both where people live and work and from where they source their goods and materials. Increasing volumes of traffic, often on old and life-expired infrastructure, lead to congestion and delays. Constraints on public budgets have led to increasing pressure on the private sector to fund improvements requiring new and innovative financial solutions. While there are clear differences in the nature of the pressures felt in the developed and less developed economies, there is an increasing recognition that in all societies, there is an accessibility problem such that certain groups become disadvantaged by the lack of access to reliable and cost-effective transport. While the problems are clearly multidimensional, research on transportation is often constrained by single disciplinary approaches and this carries over into the practice of transport planning and policy. The Encyclopedia cuts across these artificial boundaries by taking an approach that emphasizes the interaction between the different aspects of research and aims to offer new solutions to understand these problems. Each of the nine sections is based around a familiar dimension of work on transportation, but brings together the views of experts from different disciplinary perspectives. Each section is edited by an expert in the relevant field who has sought chapters from a range of authors representing different disciplines, different parts of the world, and different social perspectives. In this way, the work is not just a reflection of the state of the art that serves as a starting point for researchers and practitioners, but also a pointer toward new approaches, new ways of thinking, and novel solutions to problems.

The nine sections are structured around the following themes: Transport Modes; Freight Transport and Logistics; Transport Safety and Security; Transport Economics; Traffic Management; Transport Modeling and Data Management; Transport Policy and Planning; Transport Psychology; Sustainability and Health Issues in Transportation. Some of the chapters provide a technical introduction to a topic while others provide a bridge between topics or a more future-oriented view of new research areas or challenges. While there is guidance to cross-referencing between chapters, readers are encouraged to explore the tables of contents of all the sections to get a full understanding of the issues. Much of the Encyclopedia was completed before the Covid-19 pandemic and clearly this will have changed the situation in many areas covered by this work; the advantage of this type of reference work is that relevant updates will be possible in future editions.

The Encyclopedia has only been possible because of the cooperation of a large number of people. Robert Noland, Georgina Santos, Xiaowen Fu, and Dick Ettema served as an Editorial Advisory Board identifying possible editors of sections and advising on the overall structure. The Section Editors, Edoardo Marcucci, Kevin Cullinane, Per Garder, Maria Börjesson, Edward Chung, Chandra Bhat, Maria Attard, and Carlo Prato,

carried out the work of identifying potential authors of individual chapters, commissioning these, encouraging authors and reviewing drafts. More than 600 authors and co-authors of chapters are, however, ultimately responsible for the success of this venture. Thanks are also due to the key people at Elsevier, particularly the Publishers, Andre Wolff and Oliver Walter, and the Project Managers, Sophie Harrison and Natalie Bentahar; they have shown exemplary patience over more than 3 years in bringing this work to fruition.

LIST OF CONTRIBUTORS TO VOLUME 2

Claes Tingvall
AFRY, Solna, Sweden

Anders Lie
Chalmers University of Technology, Gothenburg, Sweden

Robert A. Scopatz
VHB, Inver Grove Heights, MN, United States

James E.W. Roseborough
OCAD University, Toronto, ON, Canada

Christine M. Wickens
Institute for Mental Health Policy Research, Centre for Addiction and Mental Health; Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

David L. Wiesenthal
York University, Toronto, ON, Canada

Höskuldur Kröyer
Department of Engineering, Trafkon, Sweden

Alan Hobbs
San Jose State University Research Foundation at NASA Ames Research Center, Moffett Field, CA, United States

Richard W. Bloom
Embry-Riddle Aeronautical University, Prescott, AZ, United States, Daytona Beach, FL, United States

James C. Fell
NORC at the University of Chicago, Bethesda, MD, United States

Per Gärder
Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States

Michal Bíl
CDV-Transport Research Centre, Lišeňská Brno, Czechia

Simonetta Boria
School of Science and Technology, University of Camerino, Camerino, Italy

David P. Gilkey
Montana Technological University, Butte, MT, United States

William Brazile
Colorado State University, Fort Collins, CO, United States

Azra Habibovic
RISE Research Institutes of Sweden, Lindholmspiren Gothenburg, Sweden

Lei Chen
RISE Research Institutes of Sweden, Lindholmspiren Gothenburg, Sweden

B. Serpil Acar
Design School, Loughborough University, England, United Kingdom

Subasish Das
Texas A&M Transportation Institute, College Station, TX, United States

Clarence C. Rodrigues
Department of Industrial and Systems Engineering, Khalifa University, Abu Dhabi, UAE

Glenn S. Baxter
School of Tourism and Hospitality Management, Suan Dusit University, Hua Hin, Thailand

Graham Wild
School of Engineering and IT, UNSW, Canberra, ACT, Australia

Lars Leden
Luleå University of Technology, Luleå, Sweden

Rock E. Miller
Orange, CA, United States

Per Erik Garder
University of Maine, Orono, ME, United States

Elliot Martin
Transportation Sustainability Research Center, University of California, Berkeley, CA, United States

Susan Shaheen
Transportation Sustainability Research Center, University of California, Berkeley, CA, United States

Terance D. Miethe
Department of Criminal Justice, UNLV, Las Vegas, NV, United States

Christopher Forepaugh
Department of Criminal Justice, UNLV, Las Vegas, NV, United States

Tanya Dudinskaya
Department of Criminal Justice, UNLV, Las Vegas, NV, United States

Erick J. Rodríguez-Seda
United States Naval Academy, Annapolis, MD, United States

Jalil Kianfar
Saint Louis University, St. Louis, MO, United States

Ulf Persson
The Swedish Institute for Health Economics, IHE, Lund, Sweden

Matúš Šucha
Palacký University in Olomouc, Olomouc, Czech Republic

Kristýna Josrová
Palacký University in Olomouc, Olomouc, Czech Republic

Dick de Waard
University of Groningen, Behavioural and Social Sciences, Department of Psychology, Groningen, The Netherlands

Nicole van Nes
SWOV Institute for Road Safety Research, Den Haag, The Netherlands

Rune Elvik
Institute of Transport Economics, Oslo, Norway

Mark J King
Queensland University of Technology (QUT), Centre for Accident Research and Road Safety (CARRS-Q), QLD, Australia

Frances L. Edwards
Mineta Transportation Institute, San Jose State University, San Jose, CA, United States

Shamsunnahar Yasmin
Queensland University of Technology (QUT), Centre for Accident Research and Road Safety-Queensland (CARRS-Q), Brisbane, QLD, Australia

Sabreena Anwar
Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL, United States

Richard Tay
Department of Civil and Environmental Engineering; Department of Architectural Studies, University of Missouri, Columbia, MO, United States; RMIT University, Melbourne, VIC, Australia

Mohammadali Shirazi
Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States

Dominique Lord
Zachry Department of Civil and Environmental Engineering, Texas A&M University, TX, United States

Fred Wegman
Delft University of Technology, Delft, the Netherlands

Ralf Risser
Palacký University in Olomouc, Czech Republic

Sherrie-Anne Kaye
Queensland University of Technology, Centre for Accident Research and Road Safety-Queensland (CARRS-Q), Brisbane, QLD, Australia

Judy Fleiter
Global Road Safety Partnership, Geneva, Switzerland

Md Mazharul Haque
Queensland University of Technology, School of Civil Engineering and Built Environment, Brisbane, QLD, Australia

Karl Kim
Department of Urban and Regional Planning, University of Hawaii, Honolulu, HI, United States

Frank Gross
VHB, Raleigh, NC, United States

Wayne K. Talley
Old Dominion University, Norfolk, VA, United States

Kenneth T. Gillingham
Yale University, New Haven, CT, United States

Stephanie M. Weber
Yale University, New Haven, CT, United States

Marion Sinclair
Department of Civil Engineering, Stellenbosch University, Stellenbosch, South Africa

Paul Boase
Transport Canada, Ottawa, ON, Canada

Brian Jonah
Road Safety Canada Consulting, Ottawa, ON, Canada

Dean C. Alberson
Bulwark Design Innovations, Bryan, TX, United States

Dr. Arjan Vincent van der Vlies
Managing Consultant Safety and Crisis management, Berenschot, Utrecht; Guest staff member Institute for Security and Global Affairs, Leiden University, Leiden, the Netherlands

John N. Ivan
Civil and Environmental Engineering, University of Connecticut, Storrs, CT, United States

Stephen J.M. Sollid
Chief Medical Officer, Norwegian Air Ambulance Foundation, Oslo, Norway; Associate Professor, University of Stavanger, Faculty of Health Sciences, Stavanger, Norway

Victoria Gitelman
Transportation Research Institute, Haifa, Israel

Alison Smiley
Human Factors North Inc., Toronto, ON Canada

Christina (Missy) Rudin-Brown
Human Factors North Inc., Toronto, ON Canada

Ivan Ostroumov
National Aviation University, av. Kosmonavta Komarova 1, Kyiv, Ukraine

Nataliia Kuzmenko
National Aviation University, av. Kosmonavta Komarova 1, Kyiv, Ukraine

Yong Peng
School of Traffic & Transportation Engineering, Central South University, Changsha, China

Xinghua Wang
School of Traffic & Transportation Engineering, Central South University, Changsha, China

Miles Tight
School of Engineering (Civil), University of Birmingham, Birmingham, United Kingdom

John D. Bullough
Lighting Research Center, Rensselaer Polytechnic Institute, Troy, NY, United States

Fangrong Chang
School of Traffic and Transportation Engineering, Central South University, Changsha, China

Maxim A. Dulebenets
Department of Civil & Environmental Engineering, Florida A&M University-Florida State University (FAMU-FSU) College of Engineering, Tallahassee, FL, United States

Saksith Chalermpong
Department of Civil Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand

Apiwat Ratanawaraha
Transportation Institute, Chulalongkorn University, Bangkok, Thailand; Department of Urban and Regional Planning, Faculty of Architecture, Chulalongkorn University, Bangkok, Thailand

Mette Møller
Technical University of Denmark, Lyngby, Denmark

Carlo Luiu
The Institute for Global Innovation, University of Birmingham, Birmingham, United Kingdom

Muhammad Z. Shah
Centre for Innovative Planning and Development, Faculty of Built Environment, Universiti Teknologi Malaysia, Johor, Malaysia

Mehdi Moeinaddini
Department of Urban and Regional Planning, Faculty of Built Environment, Universiti Teknologi Malaysia, Johor, Malaysia

Mahdi Aghaabbasi
Department of Urban and Regional Planning, Faculty of Built Environment, Universiti Teknologi Malaysia, Johor, Malaysia

Robert S. Wall Emerson
Department of Blindness and Low Vision Studies, Western Michigan University, Kalamazoo, MI, United States

Vania Ceccato
Department of Urban Planning and Built Environment, KTH Royal Institute of Technology, Stockholm, Sweden

Charles M. Farmer
Research and Statistical Services, Insurance Institute for Highway Safety, Ruckersville, VA, United States

Xiang Liu
Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, United States

Zhipeng Zhang
Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, United States

Rahim F. Benekahal
University of Illinois at Urbana-Champaign, Champaign, IL, United States

Jacob Mathew
University of Illinois at Urbana-Champaign, Champaign, IL, United States

Amy E. Peden
Royal Life Saving Society-Australia, Sydney, NSW, Australia

Stacey Willcox-Pidgeon
School of Public Health and Community Medicine, University of New South Wales, Kensington, NSW, Australia

Kyra Hamilton
School of Applied Psychology, Griffith University, Brisbane, QLD, Australia

Christer Hydén
Nye Sandviksvegen, Bergen, Norway

Robert B. Noland
Alan M. Voorhees Transportation Center, Edward J. Bloustein School of Planning and Public Policy, Rutgers University, New Brunswick, NJ, United States

Kamal Hossain
Assistant Professor, Department of Civil Engineering, Memorial University of Newfoundland, St. John's, Canada

PhD P. Eng
Assistant Professor, Department of Civil Engineering, Memorial University of Newfoundland, St. John's, Canada

Khaled Shaaban
Department of Civil Engineering/Qatar Transportation and Traffic Safety Center, Qatar University, Doha, Qatar

AnnaAnund
Swedish National Road and Transport Research Institute, Linköping, Sweden

AnnaVadeby
Rehabilitation Medicine, Linköping University, Linköping, Sweden

Xiao Qin
Civil and Environmental Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

Tor-Olav Nvestad
Institute of Transport Economics, Oslo, Norway

Yousif A. Abulhassan
Department of Occupational Safety and Health, Murray State University, Murray, KY, United States

Dimitrios Nalmpantis
School of Civil Engineering, Faculty of Engineering, Aristotle University of Thessaloniki, Thessaloniki, Central Macedonia, Greece

Allison B. Duncan
Coventry University, Coventry, Warwickshire, United Kingdom

José María Pardillo-Mayora
Department of Transport Engineering, Urban and Regional Planning, Technical University of Madrid, Madrid, Spain

Nichole L. Morris
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, United States

Curtis M. Craig
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, United States

Jacob D. Achtemeier
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, United States

Peter A. Easterlund
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, United States

Prof Marion Sinclair
Department of Civil Engineering, Stellenbosch University, South Africa

Estelle Swart
Department of Civil Engineering, Stellenbosch University, South Africa

Anna Bray Sharpin
World Resources Institute, Washington, DC, United States

Claudia Adiazola-Steil
World Resources Institute, Washington, DC, United States

Ben Welle
World Resources Institute, Washington, DC, United States

Natalia Lleras
World Resources Institute, Washington, DC, United States

Peter Tarmo Savolainen
Michigan State University, East Lansing, MI, United States

Timothy Jordan Gates
Michigan State University, East Lansing, MI, United States

Vinod Vasudevan
*Department of Civil Engineering, University of Alaska
 Anchorage, Anchorage, AK, United States*

Kasem Choocharukul
*Infrastructure Management Research Unit, Department
 of Civil Engineering, Faculty of Engineering,
 Chulalongkorn University, Pathumwan, Bangkok,
 Thailand*

Kerkritt Sriroongvikrai
*Infrastructure Management Research Unit, Department of
 Civil Engineering, Faculty of Engineering, Chulalongkorn
 University, Pathumwan, Bangkok, Thailand*

Brendan Ryan
*Human Factors Research Group, University of
 Nottingham, Nottingham, United Kingdom*

Zhe Wang
*School of Traffic and Transportation Engineering, Central
 South University, Changsha, China*

Helai Huang
*School of Traffic & Transportation Engineering, Central
 South University, Changsha, China*

Ye Li
*School of Traffic and Transportation Engineering, Central
 South University, Changsha, China*

Brian Michael Jenkins
*Mineta Transportation Institute, San Jose, CA, United
 States*

Saied Taheri
*Center for Tire Research (CenTiRe), Mechanical
 Engineering Department, Virginia Tech,
 United States*

Charles Tijus
University Paris 8, Saint-Denis, France

Nicolas Saunier
*Civil, Geological and Mining Engineering Department,
 Polytechnique Montréal, Montréal, QC, Canada*

Aliaksei Lareshyn
*Traffic and Roads, Department of Technology and Society,
 Faculty of Engineering, LTH, Lund University, Lund,
 Sweden*

Athanasios Theofilatos
*Loughborough University, School of Architecture,
 Building and Civil Engineering, Loughborough, United
 Kingdom*

Apostolos Ziakopoulos
*National Technical University of Athens, Department of
 Transportation Planning and Engineering, Athens,
 Greece*

Mohamed Abdel-Aty
*Department of Civil, Environmental and Construction
 Engineering, University of Central Florida, Orlando, FL,
 United States*

Jaeyoung Lee
*Department of Civil, Environmental and Construction
 Engineering, University of Central Florida, Orlando, FL,
 United States*

Andrew P. Tarko
*Purdue University, Lyles School of Civil Engineering, West
 Lafayette, IN, United States*

Matthew C. Camden
*Virginia Tech Transportation Institute, Blacksburg, VA,
 United States*

Jeffrey S. Hickman
*Virginia Tech Transportation Institute, Blacksburg, VA,
 United States*

Richard J. Hanowski
*Virginia Tech Transportation Institute, Blacksburg, VA,
 United States*

Martin Walker
*Virginia Tech Transportation Institute, Blacksburg, VA,
 United States*

Konstantinos Kirytopoulos
*School of Natural & Built Environments, University of
 South Australia, Adelaide, SA, Australia*

Panagiotis Ntzeremes
National Technical University of Athens, Athens, Greece

Konstantinos Kazaras
National Technical University of Athens, Athens, Greece

Lai Zheng
*School of Transportation Science and Engineering, Harbin
 Institute of Technology, Harbin, Heilongjiang, China*

Tarek Sayed
*Department of Civil Engineering, University of British
 Columbia, Vancouver, BC, Canada*

David A. Hensher
*Institute of Transport and Logistics Studies,
 The University of Sydney Business School, Sydney, NSW,
 Australia*

Matts-Åke Belin
*KTH Royal Institute of Technology, Stockholm,
 Sweden; Swedish Transport Administration, Borlänge,
 Sweden*

Maria Pregnolato
*Department of Civil Engineering, University of Bristol,
 Bristol, United Kingdom*

Amirhassan Kermanshah

*Department of Civil and Environmental Engineering,
Vanderbilt University, Nashville, TN, United States*

Wisinee Wisetjindawat

*Department of Engineering Mathematics, University of
Bristol, Bristol, United Kingdom*

John J. McDonough

*National Institute for Safety Research, Inc. (NISR),
Pinehurst, NC, United States*

Mercedes Castro-Nuño

*Applied Economics and Research Group,
Universidad de Sevilla University of Seville, Seville, Spain*

José I. Castillo-Manzano

*Applied Economics & Research Group, Universidad de
Sevilla, Seville, Spain*

Huaguo Zhou

*Department of Civil Engineering, Auburn University,
Auburn, AL, United States*

Md Atiquzzaman

*Department of Civil Engineering, Auburn University,
Auburn, AL, United States*

Martina Raue

*Massachusetts Institute of Technology AgeLab,
Cambridge, MA, United States*

Eva Lerner

*LMU Center for Leadership and People Management,
Munich, Germany; FOM University of Applied
Sciences for Economics and Management, Munich,
Germany*

CONTENTS OF ALL VOLUMES

<i>Editorial Board</i>	<i>v</i>
<i>Introduction</i>	<i>vii</i>
<i>Contributors to Volume 2</i>	<i>ix</i>

VOLUME 1

Introduction to Transportation Economics	1
Market Failures and Public Decision Making in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	2
Demand for Freight Transport <i>José Holguín-Veras and Diana G. Ramírez-Ríos</i>	7
Cost Functions for Road Transport <i>José Manuel Vassallo</i>	13
Future of Urban Freight <i>Russell G. Thompson</i>	20
Operation Costs for Public Transport <i>Marco Batarce</i>	26
Natural Monopoly in Transport <i>André de Palma and Julien Monardo</i>	30
Freight Costs: Air and Sea <i>Yulai Wan and Dong Yang</i>	36
Transport Production and Cost Structure <i>Ricardo Giesen and Darío Farren</i>	46
The Concept of External Cost: Marginal versus Total Cost and Internalization <i>Sofia F. Franco</i>	56
Value of Time <i>John J. Bates</i>	67
Valuation of Carbon Emissions <i>Svante Mandell</i>	72
Valuation of Travel Time Variability Using Scheduling Models <i>Katrine Hjorth</i>	76

Value of Crowding <i>Daniel Hörcher</i>	84
What Drives Transport and Mobility Trends? The Chicken-and-Egg Problem <i>Nathalie Picard</i>	89
Pricing Principles in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	95
Long-Run Versus Short-Run Valuations <i>Stefanie Peer</i>	102
Value of Noise <i>Jon P. Nelson</i>	106
The Value of Life and Health <i>Henrik Andersson</i>	114
The Value of Security, Access Time, Waiting Time, and Transfers in Public Transport <i>Raquel Espino, Juan de Dios Ortúzar, and Luis I. Rizzi</i>	122
Demand for Passenger Transportation <i>Kenneth A Small and Robin Lindsey</i>	127
Real-World Experiences of Congestion Pricing <i>Charles Raux</i>	134
Distributional Effects of Congestion Charges and Fuel Taxes <i>Jonas Eliasson</i>	139
The Bottleneck Model <i>Dereje Abegaz and Yili Tang</i>	146
Dynamic Congestion Pricing and User Heterogeneity <i>Kathrin Goldmann and Gernot Sieg</i>	150
Economics of Parking <i>Daniel Albalade and Albert Gragera</i>	159
Loss Aversion and Size and Sign Effects in Value of Time Studies <i>Andrew Daly</i>	165
Intertemporal Variation of Valuations <i>James Fox</i>	170
The Rebound Effect for Car Transport <i>Bruno De Borger, Ismir Mulalic and Jan Rouwendal</i>	174
Elasticities for Travel Demand: Recent Evidence <i>Fay Dunkerley, Charlene Rohr and Mark Wardman</i>	179
Parking Price Elasticities <i>Stephan Lehner</i>	185
The Pareto Criterion and the Kaldor Hicks Criterion <i>Harald Minken</i>	190
Social Discount Rates <i>Lars Hultkrantz</i>	195
Cross-Elasticities between Modes <i>Mark Wardman and Jeremy Toner</i>	201
First-Best Congestion Pricing <i>Moez Kilani</i>	209

Ethical Aspects-Can We Value Life, Health, and Environment in Money Terms? <i>Jose M. Grisolia and Ken Willis</i>	216
Car tolls, Transit Subsidies for Commuting, and Distortions on the Labor Market <i>Ioannis Tikoudis¹ and Kurt Van Dender¹</i>	221
Demand Management and Capacity Planning of Airports <i>Stephen Ison and Lucy Budd</i>	227
Dealing With Negative Externalities: Low Emission Zones Versus Congestion Tolls <i>Valeria Bernardo, Xavier Fageda, and Ricardo Flores-Fillol</i>	231
The Rule-of-a-Half and Interpreting the Consumer Surplus as Accessibility <i>Mogens Fosgerau and Ninette Pilegaard</i>	237
Producer Surplus <i>Chau Man Fung</i>	242
The Robustness of Cost-Benefit Analyses <i>Morten Welde and James Odeck</i>	249
The GDP Effects of Transport Investments: The Macroeconomic Approach <i>James Laird and Daniel Johnson</i>	256
The Mohring Effect <i>Hugo E. Silva</i>	263
Public Transport Fare and Subsidy Optimization <i>Qianwen Guo and Zhongfei Li</i>	267
Transportation Equity <i>Rafael H. M. Pereira, and Alex Karner</i>	271
Impact of Transport Cost-Benefit Analysis on Public Decision-Making <i>Niek Mouter</i>	278
Causal Inference for Ex Post Evaluation of Transport Interventions <i>Daniel J. Graham</i>	283
Transport Cost and Location of Firms <i>Adelheid Holl</i>	291
Commuting, the Labor Market, and Wages <i>Jan Rouwendal, and Ismir Mulalic</i>	297
How to Buy Transport Infrastructure <i>Johan Nyström</i>	302
Procurement of Public Transport: Contractual Regimes <i>Andrew Smith, and Chris Nash</i>	308
The Mono-Centric City Model and Commuting Cost <i>Zhi-Chun Li, and Ya-Juan Chen</i>	315
Value of Time in Freight Transport <i>Gerard de Jong</i>	321
The Economics and Planning of Urban Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	326
Incentives in Public Transport Contracts <i>Andreas Vigren</i>	332
Transportation Improvements and Property Prices <i>Alex Anas</i>	337

Transport Infrastructure Effects on Economic Output: The Microeconomic Approach <i>Patricia C. Melo</i>	347
Wider Economic Impacts of Transport Investments <i>Anthony J. Venables</i>	355
Employment Effects of Transport Infrastructure <i>Anna Matas, and Javier Asensio</i>	360
Regulatory Reforms and Competition in Public Transport <i>Didier van de Velde</i>	365
How to Finance Transport Infrastructure? <i>Tiziana D'Alfonso, and Giuseppe Catalano</i>	371
Congestion, Allocation and Competition on the Railway Tracks <i>John Armstrong, and John Preston</i>	378
Public Private Partnership <i>David Meunier, and Emile Quinet</i>	385
Airline Economics <i>Anming Zhang, Yahua Zhang, and Zhibin Huang</i>	392
Privatization and Deregulation of the Airline Industry <i>Achim I. Czerny, and Hao Lang</i>	397
Price Discrimination and Yield Management in the Airline Industry <i>Guoquan Zhang, Colin C.H. Law, Yahua Zhang, and Hangjun Yang</i>	404
Regulation and Competition in Railways <i>Chris Nash, and Andrew Smith</i>	409
Cycling Economics <i>Bert van Wee</i>	414
The Economic Rationale for High-Speed Rail <i>Ginés de Rus</i>	419
Rail Cost Functions <i>Kristofer Odolinski, and Phill Wheat</i>	425
Transport and International Trade <i>Siri Pettersen Strandenes</i>	431
Maritime Economics: Organizational Structures <i>Kevin P.B. Cullinane</i>	436
Port Planning and Investment <i>Jasmine Siu Lee Lam</i>	443
Estimating the Capital Stock of Transport Infrastructure <i>Heike Link</i>	449
The Economics of Reducing Carbon Emissions From Air and Road Transport <i>Olga Ivanova</i>	457
Regulation and Financing of Toll Roads <i>Marco Ponti</i>	464
Are Megaprojects too Transformational for Cost-Benefit Analysis? <i>Tom Worsley</i>	470
Economics of Transportation Safety <i>Ian Savage</i>	476

Cost Overruns of Transportation Infrastructure Projects <i>James Odeck, and Morten Welde</i>	483
The Downs-Thomson Paradox <i>Joel P. Franklin</i>	490
Policy Instruments for Plug-In Electric Vehicles: An Overview and Discussion <i>Jake Whitehead, Patrick Plötz, Patrick Jochem, Frances Sprei, and Elisabeth Dütschke</i>	496
Vertical and Horizontal Separation in the European Railway Sector and Its Effects on Productivity <i>Pedro Cantos-Sánchez</i>	503
How will Autonomous Vehicles Impact Car Ownership and Travel Behavior <i>Patrick M. Bösch, Felix Becker, Henrik Becker, and Kay W. Axhausen</i>	508
Policy Instruments to Reduce Carbon Emissions from Road Transport Computable General Equilibrium Analysis in Transportation Economics <i>Johannes Bröcker</i>	520
Estimation of Value of Time <i>Stefan Flügel, and Askill H. Halse</i>	527
The Taxation of Car Use in the Future <i>Griet De Ceuster, and Inge Mayeres</i>	534
Cost-Benefit Analysis and Other Assessment Techniques: Contrasts and Synergies <i>Paolo Beria</i>	540
Demand for Air Travel and Income Elasticity <i>Jing Lu, Yucan Meng, Changmin Jiang, and Cheng Lv</i>	547
Generalized Cost for Transport <i>Jepppe Rich</i>	555
The Impact of Electric Vehicles on Energy Systems <i>Patrick E.P. Jochem, Jake Whitehead, and Elisabeth Dütschke</i>	560
Uber versus Taxis <i>Georgina Santos</i>	566
Contract Efficiency in Public Transport Services <i>Philippe Gagnepain, and Marc Ivaldi</i>	572
Company Cars <i>Stefan Gössling</i>	580
Transportation Network Companies (TNCs) and the Future of Public Transportation <i>Susan Shaheen, and Adam Cohen</i>	584
Public Transport in Low Density Areas <i>Jani-Pekka Jokinen, Leif Sörensen, and Jan Schlüter</i>	589
Market Failures in Transport: Direct and Indirect Public Intervention <i>Federico Boffa, and Alberto Iozzi</i>	596
The Braess Paradox <i>Anna Nagurney, and Ladimer S. Nagurney</i>	601
VOLUME 2	
Introduction to Transportation Safety and Security <i>Per Gårder</i>	1

The Concept of “Acceptable Risk” Applied to Road Safety Risk Level <i>Claes Tingvall</i>	2
Crash Not Accident <i>Robert A. Scopatz</i>	6
Age and Gender as Factors in Road Safety <i>Marion Sinclair</i>	11
Aggressive Driving and Road Rage <i>James E.W. Roseborough, Christine M. Wickens, and David L. Wiesenthal</i>	17
Aircraft Maintenance and Inspection <i>Alan Hobbs</i>	25
Airport Security <i>Richard W. Bloom</i>	34
Transport Safety and Security: Alcohol <i>James C. Fell</i>	40
Animal Crashes <i>Michal Bil</i>	53
Attenuators <i>Simonetta Boria</i>	63
ATV, Snowmobile, and Terrain Vehicle Safety <i>David P. Gilkey, and William Brazile</i>	77
Automobile Safety Inspection <i>Subasish Das</i>	85
Aviation Safety: Commercial Airlines <i>Clarence C. Rodrigues</i>	90
Aviation Safety, Freight, and Dangerous Goods Transport by Air <i>Glenn S. Baxter and Graham Wild</i>	98
Bicycle Collision Avoidance Systems: Can Cyclist Safety be Improved with Intelligent Transport Systems? <i>Lars Leden</i>	108
Bicycle Infrastructure <i>Rock E. Miller</i>	115
Bicycles, E-Bikes and Micromobility, A Traffic Safety Overview <i>Höskuldur Kröyer</i>	125
Bicycles: The Safety of Shared Systems Versus Traditional Ownership <i>Mercedes Castro-Nuño and José I. Castillo-Manzano</i>	139
Bridge Safety <i>Per Erik Garder</i>	144
Carsharing Safety and Insurance <i>Elliot Martin and Susan Shaheen</i>	150
Carjacking <i>Terance D. Mieth, Christopher Forepaugh, Tanya Dudinskaya</i>	157
Collision Avoidance Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	164
Collision Avoidance Systems, Automobiles <i>Erick J. Rodríguez-Seda</i>	173

Connected Automated Vehicles: Technologies, Developments, and Trends <i>Azra Habibovic and Lei Chen</i>	180
Construction Zones <i>Jalil Kianfar</i>	189
Costs of Accidents <i>Ulf Persson</i>	196
Critical Issues for Large Truck Safety <i>Matthew C. Camden, Jeffrey S. Hickman, Richard J. Hanowski, and Martin Walker</i>	200
Demerit Points and Similar Sanction Programs <i>Matúš ťucha and Kristýna Josrová</i>	210
Driver State and Mental Workload <i>Dick de Waard and Nicole van Nes</i>	216
Drugs, Illicit, and Prescription <i>Rune Elvik</i>	221
Education, Training, and Licensing <i>Matúš ťucha and Kristýna Josrová</i>	228
Elderly Driver Safety Issues <i>Mark J King</i>	233
Emergency Response Systems <i>Frances L. Edwards</i>	240
Emergency Vehicles and Traffic Safety <i>Shamsunnahar Yasmin, Sabreena Anowar, and Richard Tay</i>	247
Encouragement: Awards and Incentives <i>Fred Wegman</i>	255
Enforcement and Fines <i>Matúš ťucha and Ralf Risser</i>	263
Epidemiology of Road Traffic Crashes <i>Sherrie-Anne Kaye, Judy Fleiter, and Md Mazharul Haque</i>	269
Evacuation Planning and Transportation Resilience <i>Karl Kim</i>	276
Exposure: A Critical Factor in Risk Analysis <i>Frank Gross, PhD, PE</i>	282
Passenger Ferry Vessels and Cruise Ships: Safety and Security <i>Wayne K. Talley</i>	290
Fuel Economy Standards: Impacts on Safety <i>Kenneth T. Gillingham and Stephanie M. Weber</i>	296
Hazardous Materials Transport <i>Dr. Arjan Vincent van der Vlies</i>	304
Head-on Crashes <i>John N. Ivan</i>	311
Helicopters in Emergency Medical Response <i>Stephen J.M. Sollid, M.D., PhD</i>	316
Horizontal and Vertical Geometry <i>Victoria Gitelman</i>	322

Human Factors in Transportation <i>Alison Smiley, Christina (Missy) Rudin-Brown</i>	331
In-Depth Crash Analysis and Accident Investigation <i>Yong Peng, Helai Huang, and Xinghua Wang</i>	346
Incident Detection Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	351
Inequality and Traffic Safety <i>Miles Tight</i>	358
Lighting <i>John D. Bullough</i>	361
Macroscopic Safety Analysis <i>Mohamed Abdel-Aty and Jaeyoung Lee</i>	367
Motor Vehicle Crash Reportability <i>John J. McDonough</i>	380
Nominal Safety <i>Per Erik Garder</i>	386
Parking Lots <i>Maxim A. Dulebenets</i>	392
Passenger Van Safety <i>Saksith Chalermpong and Apiwat Ratanawaraha</i>	401
Passive Prevention Systems in Automobile Safety <i>B. Serpil Acar</i>	406
Pedestrian Safety, Children <i>Mette Møller</i>	415
Pedestrian Safety, General <i>Muhammad Z. Shah, Mehdi Moeinaddini, and Mahdi Aghaabbasi</i>	420
Pedestrian Safety, Older People <i>Carlo Lui</i>	429
Visually Impaired Pedestrian Safety <i>Robert S. Wall Emerson</i>	435
Photo/Video Traffic Enforcement <i>Charles M. Farmer</i>	439
Powered Two- and Three-Wheeler Safety <i>Fangrong Chang, Helai Huang, and Md. Mazharul Haque</i>	443
Railroad Safety <i>Xiang Liu and Zhipeng Zhang</i>	451
Railroad Safety: Grade Crossings and Trespassing <i>Rahim F. Benekohal and Jacob Mathew</i>	466
Recreational Boating Safety: Usage, Risk Factors, and the Prevention of Injury and Death <i>Amy E. Peden, Stacey Willcox-Pidgeon, and Kyra Hamilton</i>	477
Refuge Islands <i>Christer Hydén</i>	487
Risk Perception and Risk Behavior in the Context of Transportation <i>Martina Raue and Eva Lerner</i>	494

Road Diets <i>Robert B. Noland</i>	500
Road Safety Audits <i>Xiao Qin</i>	508
Roadside Safety Barriers <i>Dean C. Alberson</i>	513
Road Safety Management in Selected Countries <i>Paul Boase and Brian Jonah</i>	519
Roadway Pavement Conditions and Various Summer and Winter Maintenance Strategies <i>Kamal Hossain, PhD P. Eng</i>	529
Safety of Roundabouts <i>Khaled Shaaban</i>	539
Rumble Strips, Continuous Shoulder, and Centerline <i>AnnaAnund and AnnaVadeby</i>	549
Safety Culture <i>Tor-Olav Nvestad</i>	554
Safety Data Quality Management <i>Robert A. Scopatz</i>	560
School Bus Safety <i>Yousif A. Abulhassan</i>	564
School Campus Traffic Circulation <i>Dimitrios Nalmpantis</i>	568
Sexual Violence in Public Transportation <i>Vania Ceccato</i>	576
Shared Space <i>Allison B. Duncan</i>	584
Side Area Safety and Side Slopes <i>José María Pardillo-Mayora</i>	593
Simulators <i>Nichole L. Morris, Curtis M. Craig, Jacob D. Achtemeier, and Peter A. Easterlund</i>	602
Sleep-Related Issues and Fatigue <i>Prof Marion Sinclair and Estelle Swart</i>	611
Speed Governors and Limiters <i>Christer Hydén</i>	617
Speed Limits on Rural Highways <i>Peter Tarmo Savolainen and Timothy Jordan Gates</i>	622
Speed Limits on Urban Streets <i>Anna Bray Sharpin, Claudia Adriazola-Steil, Ben Welle, and Natalia Lleras</i>	632
Speed-Reducing Measures <i>Vinod Vasudevan</i>	641
Striping, Signs, and Other Forms of Information <i>Kasem Choocharukul and Kerkritt Sriroongvikrai</i>	648
Suicides <i>Brendan Ryan</i>	656

Surrogate Measures of Safety <i>Nicolas Saunier and Aliaksei Laureshyn</i>	662
Targeting Transit: the Terrorist Threat and the Challenges to Security <i>Brian Michael Jenkins</i>	668
The Swedish Vision Zero: A Policy Innovation <i>Matts-Åke Belin</i>	675
Tire Safety <i>Saied Taheri</i>	681
Town Gates: Section on Transport Safety and Security <i>Charles Tijus</i>	685
Traffic Flow Volume and Safety <i>Athanasios Theofilatos and Apostolos Ziakopoulos</i>	692
Traffic Safety and Security of Taxis and Ride-Hailing Vehicles <i>Zhe Wang, Helai Huang, and Ye Li</i>	699
Traffic Signals and Safety <i>Andrew P. Tarko</i>	706
Tunnels, Safety and Security Issues-Risk Assessment for Road Tunnels: State-of-the-Art Practices and Challenges <i>Konstantinos Kirytopoulos, Panagiotis Ntzeremes, and Konstantinos Kazaras</i>	713
Understanding, Managing, and Learning from Disruption <i>Karl Kim</i>	719
Use/Analysis of Crash Data and Underreporting of Crashes <i>Mohammadali Shirazi and Dominique Lord</i>	726
Utility Poles <i>Lai Zheng and Tarek Sayed</i>	731
Value of Life and Injuries <i>David A. Hensher</i>	737
Effects of Weather <i>Maria Pregnolato, Amirhassan Kermanshah, and Wisinee Wisetjindawat</i>	742
Wrong-Way Driving on Motorways <i>Huaguo Zhou and Md Atiquzzaman</i>	751

VOLUME 3

Freight Transport and Logistics <i>Sharon Cullinane and Kevin Cullinane</i>	1
Expanding the Perspective of Logistics and Supply Chain Management <i>David J. Closs</i>	8
Logistics and Supply Chain Management Performance Measures <i>David B. Grant and Sarah Shaw</i>	16
Economic Regulation/Deregulation and Nationalization/Privatization in Freight Transportation <i>Wayne K. Talley</i>	24
Freight Transport Policy <i>Luca Zamparini and Aura Reggiani</i>	29

Planning and Financing Logistics Spaces <i>Nicolas Raimbault</i>	35
Supply Chain Risk Management: Creating the Resilient Supply Chain <i>Richard Wilding</i>	41
Transportation Safety and Security <i>Maria G. Burns</i>	47
Resilience in Freight Transport Networks <i>Zhuohua Qu, Chengpeng Wan, and Zaili Yang</i>	53
Environmental Sustainability in Freight Transportation <i>Lisa M. Ellram</i>	58
Sustainable Logistics, CSR in Logistics, and Sustainable Supply Chain Management <i>Maria Björklund and Maja Piecyk-Ouellet</i>	64
Information Sharing and Business Analytics in Global Supply Chains <i>Prof Usha Ramanathan and Prof Ramakrishnan Ramanathan</i>	71
Logistics Information Systems <i>Petri Helo and Javad Rouzafzoon</i>	76
Factors Affecting the Selection of Logistics Service Providers <i>Aicha AGUEZZOUL</i>	85
Logistics Service Performance <i>Kee-hung Lai, Jinan Shao, and Yongyi Shou</i>	89
The World Bank's Logistics Performance Index <i>Christina K. Wiederer cwiederer, Jean-François Arvis, Lauri M. Ojala, and Tuomas M. M. Kiiski</i>	94
Outsourcing Logistics Functions <i>Evi Hartmann, Hendrik Birkel, and Matthias Kopyto</i>	102
Freight Transport and Logistics in JIT Systems <i>James H. Bookbinder and M. Ali Aelkū</i>	107
Supply Chain Finance <i>Erik Hofmann</i>	113
Packaging Logistics <i>Jesús García-Arca, Alicia Trinidad González-Portela Garrido, and J. Carlos Prado-Prado</i>	119
The Bullwhip Effect <i>Jan C. Fransoo and Maximiliano Udenio</i>	130
Blockchain Applications in Logistics <i>Yingli Wang</i>	136
Logistics in Asia <i>Shong-lee Ivan Su</i>	143
Logistics in the Developing World <i>Charles Kunaka</i>	150
Freight Network Modeling <i>Lóránt Tavasszy and Yousef Maknoon</i>	157
National Freight Transport Models <i>Gerard de Jong</i>	162
Freight Flows in Cities <i>Genevieve Giuliano</i>	168

Urban Logistics and Freight Transport <i>Michael Browne, José Holguín-Veras, and Julian Allen</i>	178
Omni-Channel Logistics <i>Tom Van Woensel</i>	184
Humanitarian Logistics <i>Gyöngyi Kovács and Diego Vega</i>	190
Optimization of Humanitarian Logistics <i>M. Teresa Ortuño, Jose M. Ferrer, Inmaculada Flores, and Gregorio Tirado</i>	195
Event Logistics <i>Rev. Ruth Dowson and Dan Lomax</i>	201
Reverse Logistics <i>Dale S. Rogers and Ronald S. Lembke</i>	208
E-Tailing and Reverse Logistics <i>Sharon Cullinane</i>	219
Green Routing of Freight Vehicles <i>Tolga Bektaş</i>	224
Freight Mode Choice <i>Hyun Chan Kim and Alan Nicholson</i>	231
The Value of Time in Freight Transport <i>María Feo-Valero, Amaya Vega, and Bárbara Vázquez-Paja</i>	236
Behavioral Research in Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	242
Container (Liner) Shipping <i>Theo Notteboom</i>	247
Bulk Shipping Markets: An Overview of Market Structure and Dynamics <i>Manolis G. Kavussanos and Stella A. Moysiadou</i>	257
Ferries and Short Sea Shipping <i>Lourdes Trujillo and Alba Martínez-López</i>	280
Shipping and the Environment <i>Karin Andersson, Selma Brynolf, Lena Granhag, and J. Fredrik Lindgren</i>	286
Energy Efficiency of Ships <i>Harilaos N. Psaraftis</i>	294
Seaports <i>Mary R. Brooks and Geraldine Knatz</i>	299
Port Hinterlands <i>Francesco Parola, Giovanni Satta, and Francesco Vitellaro</i>	305
Seaports as Clusters of Economic Activities <i>Peter W. de Langen</i>	310
Port Efficiency and Effectiveness <i>Lourdes Trujillo, María Manuela González, Casiano Manrique-De-Lara-Peñate, and Ivone Pérez</i>	316
Container Port Automation <i>Michael G.H. Bell</i>	323
Optimizing Crane Operations in Ports Scheduling of Liner Container Shipping Services	335

<i>Yuquan Du and Qiang Meng</i>	
Dry Ports	344
<i>Gordon Wilmsmeier and Jason Monios</i>	
Arctic Shipping	349
<i>Yufeng Lin, David G. Babb, and Adolf K.Y. Ng</i>	
Airfreight and Economic Development	355
<i>Kenneth Button</i>	
Air Freight Logistics	361
<i>Keith Debbage and Neil Debbage</i>	
Air Freight Marketing	369
<i>Lucy Budd and Stephen Ison</i>	
Drones in Freight Transport	374
<i>Oliver Kunze</i>	
Duty of Care in the Selection of Motor Carriers	382
<i>Thomas M. Corsi</i>	
Carrier Selection for Less-Than-Truckload (LTL) Shipments	388
<i>Dinçer Konur, Gonca Yildirim, and Bahriye Cesaret</i>	
Decarbonizing Road Freight Transport	395
<i>Heikki Liimatainen</i>	
The Rebound Effect in Road Freight Transport	402
<i>Tooraj Jamasb and Manuel Llorca</i>	
Autonomous Goods Transport	407
<i>Heike Flømig</i>	
Rail Freight	413
<i>Dr Allan Woodburn</i>	
Rail Freight Vehicles	423
<i>Maksym Spiryagin, Qing Wu, Peter Wolfs, Colin Cole, Valentyn Spiryagin, and Tim McSweeney</i>	
Eurasia Rail Freight: Enablers and Inhibitors of Future Growth	436
<i>Hendrik Rodemann and Simon Templar</i>	
Intermodal and Synchromodal Freight Transport	456
<i>Tomas Ambra, Koen Mommens, and Cathy Macharis</i>	
Pipelines	463
<i>Matthew E. Oliver</i>	
3D Printers and Transport	471
<i>Wouter P.C. Boon and Bert van Wee</i>	
The Physical Internet and Logistics	479
<i>Eric Ballot and Shenle PAN</i>	
Bicycles for Urban Freight	488
<i>Barbara Lenz and Johannes Gruber</i>	
China's Belt and Road Initiative	495
<i>Paul Tae-Woo Lee</i>	

VOLUME 4

Introduction to Traffic Management <i>Edward C.S. Chung</i>	1
Urban Motorway Management <i>John Gaffney and Hendrik Zurlinden</i>	2
Ramp Metering Application <i>John Gaffney and Hendrik Zurlinden</i>	10
City Wide Coordinated Ramp Meters <i>John Gaffney and Hendrik Zurlinden</i>	21
Variable Speed Limits for Traffic Efficiency Improvement <i>José Ramón D. Frejo and Bart De Schutter</i>	33
Hard Shoulder Running <i>Justin Geistefeldt</i>	41
High-Occupancy Vehicle (HOV) and High-Occupancy Toll (HOT) Lanes <i>Roxana J. Javid, Jiani Xie, Lijiao Wang, Wenruifan Yang, Ramina Jahanbakhsh Javid, and Mahmoud Salari</i>	45
Reversible Lanes: Guidelines, Operation and Control, Research Directions <i>Gowri Asaithambi, Venkatesan Kanagaraj, and Madhuri Kashyap</i>	52
Electronic Toll Collection <i>Azusa Toriumi</i>	60
Road Pricing-Theory and Applications <i>Kian Keong Chin</i>	68
Road Pricing 1: The Theory of Congestion Pricing <i>Timothy D. Hau</i>	74
Road Pricing 2: Short- and Long-Run Equilibrium of Road Transportation <i>Timothy D. Hau</i>	83
Road Pricing 3: The Implications for Pricing Public Transportation <i>Timothy D. Hau</i>	90
Road Pricing 4: Case Study-The Implementation of Electronic Road Pricing in Hong Kong <i>Timothy D.</i>	103
Advanced Travelers Information Systems (ATIS) <i>Chintan Advani and Ashish Bhaskar</i>	106
Travel Time Reliability <i>Sharmili Banik, Anil Kumar, and Lelitha Vanajakshi</i>	109
Traffic Incident Management <i>Ruimin Li</i>	122
Traffic Incident Detection <i>Shuyan Chen and Yingjiu Pan</i>	128
Bottleneck <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	134
Flow Breakdown <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	143
Recurrent Congestion <i>Takahiro Tsubota</i>	154

Nonrecurrent Congestion <i>Takahiro Tsubota</i>	158
Freeway to Arterial Interfaces <i>Abolfazl Karimpour and Yao-Jan Wu</i>	162
Arterial Road Management <i>Jiaqi Ma, Yi Guo and Adekunle Adebisi</i>	169
Signalized Intersections <i>Chaitrali Shirke</i>	178
Protected Phase <i>Jiarong Yao, Yumin Cao, and Keshuang Tang</i>	185
Permitted Phase <i>Yumin Cao, Jiarong Yao, and Keshuang Tang</i>	196
Hook Turns: Implementation, Benefits, and Limitations <i>Sara Moridpour and Amir Falamarzi</i>	203
Traffic Signal Coordination <i>Rahim F. Benekohal</i>	213
Emergency Vehicle Priority (Preemption): Concept and Advancements <i>Chaitrali Shirke</i>	221
Signalized Roundabouts <i>Yetis Sazi Murat and Rui-jun Guo</i>	227
Turbo Roundabouts: Design, Capacity and Comparison With Alternative Types of Roundabouts <i>Marco Guerrieri and Raffaele Mauro</i>	238
Capacity of an Intersection <i>Ashish Verma and Milan Mathew Thomas</i>	247
Delay <i>Shinji Tanaka</i>	258
Queue Length <i>Shinji Tanaka</i>	263
Local Area Traffic Management <i>Michael A.P. Taylor</i>	268
On-Street Parking <i>Jun Chen and Guang Yang</i>	278
Off-Street Parking <i>Jun Chen and Guang Yang</i>	285
Parking Information Systems <i>Behrang Assemi and Douglas Baker</i>	289
Performance-Based Parking Management <i>Douglas Baker and Behrang Assemi</i>	299
Transit Priority <i>Wanjing Ma, Qiheng Lin, and Ling Wang</i>	307
Adaptive Bus Control <i>Monica Menendez</i>	315
Transit Fare Collection <i>Mahmoud Mesbah and Kamal Khanali</i>	325

Transit Information Systems <i>Ankit Kumar Yadav and Nagendra R Velaga</i>	331
Tram Lane Configurations and Driving Rules <i>Farhana Naznin</i>	338
Railway Crossing <i>Chunliang Wu and Inhi Kim</i>	341
Pedestrian Crossing (Crosswalk) <i>Miho Iryo-Asano, Wael K.M. Alhajyaseen, and Koji Suzuki</i>	346
Bike-Sharing System: Uncovering the “1Success Factors” <i>S.K. Jason Chang and Amanda Fernandes Ferreira</i>	355
Airspace Systems Technologies-Overview and Opportunities <i>Banavar Sridhar and Gano B. Chatterji</i>	363
Air Traffic Flow and Capacity Management <i>Li Weigang and Cristiano P Garcia</i>	380
Port Management <i>Maria G. Burns</i>	390
Port Performance Measurement from a Multistakeholder Perspective <i>Min-Ho Ha, Zaili Yang, and Young-Joon Seo</i>	396
Transport Modeling and Data Management <i>Chandra Bhat</i>	407
Computational Methods and Data Analytics <i>Bilal Farooq and David López</i>	408
Activity-Based Models <i>Renato Guadamuz and Rajesh Paleti</i>	414
Advanced Traveler Information Systems <i>Stephen D. Boyles</i>	418
Demand-Responsive Transit, Evaluation Studies <i>Sebastián Raveau</i>	423
Location Choice Models <i>Adam Wilkinson Davis</i>	428
Full Feedback and Equilibrium Modeling in Urban Travel Forecasting <i>Yu (Marco) Nie, Jun Xie, and David Boyce</i>	432
The Use and Value of Geographic Information Systems in Transportation Modeling <i>Ming Zhang</i>	440
Latent Demand and Induced Travel <i>Charisma F. Choudhury</i>	448
ICT, Virtual and In-Person Activity Participation, and Travel Choice Analysis <i>Jacek Pawlak and Giovanni Circella</i>	452
Public Transit Ridership Forecasting Models <i>Ipsita Banerjee, Deepa L, and Abdul Rawoof Pinjari</i>	459
Spatial Mismatch, Job Access, and Reverse Commuting <i>Gian-Claudia Sciara</i>	468

Microsimulation and Agent-Based Models in Transportation <i>Milos Balac</i>	473
Choice Models in Transportation <i>Naveen Chandra Iraganaboina and Naveen Eluru</i>	477
Multi-Criteria Decision Analysis <i>Zhanmin Zhang and Srijith Balakrishnan</i>	485
The National Household Travel Survey Data Series (NPTS/NHTS) <i>Nancy McGuckin</i>	493
Route Choice and Network Modeling <i>Emma Frejinger and Maëlle Zimmermann</i>	496
Departure Time Choice Modeling <i>Khandker Nurul Habib</i>	504
Traveler Responses to Congestion <i>David T. Ory and Gayathri Shivaraman</i>	509
Origin-Destination Demand Estimation Models <i>William H.K. Lam, Hu Shao, Shuhan Cao, and Hai Yang</i>	515
Parking Demand Models <i>S.C. Wong, Zhi-Chun Li, and William H.K. Lam</i>	519
Pavement Management Systems <i>Senthilmurugan Thyagarajan</i>	524
Residential Location Choice Models <i>Shlomo Bekhor and Sigal Kaplan</i>	531
Transport Demand Management <i>Feiyang Zhang and Becky P.Y. Loo</i>	537
Traffic Flow Analysis <i>H. Michael Zhang and Jia Li</i>	544
Autonomous Vehicles and Transportation Modeling <i>Annesha Enam, Felipe de Souza, Omer Verbas, Monique Stinson, and Joshua Auld</i>	557
Ride-Hailing and Travel Demand Implications <i>Felipe F. Dias</i>	564
Transportation Modeling and Planning Software <i>Joel Freedman</i>	569
Transportation Statistics and Databases <i>Taha Hossein Rashidi</i>	574
Travel Surveys <i>Stacey G. Bricka</i>	587
Travel Demand Forecasting: Where Are We and What Are the Emerging Issues <i>Thomas F. Rossi</i>	590
Travel Model Calibration and Validation <i>Ram M. Pendyala</i>	596
Trip Chaining Analysis <i>Cynthia Chen and Yusak Susilo</i>	606
Vehicle Ownership Models <i>Dr. Sören Groth and Prof. Dr. Dirk Wittowsky</i>	612

Bicycle Sharing/Bikesharing <i>Catherine Morency and Jean-Simon Bourdeau</i>	617
Carsharing <i>Shiva Habibi and Frances Sprei</i>	623
Urban Recreational Travel <i>Long Cheng and Frank Witlox</i>	629
Place Perception and Travel Behavior <i>Kathleen Deutsch-Burgner and Konstadinos G. Goulias</i>	635

VOLUME 5

Transport Modes <i>Edoardo Marcucci</i>	1
Infrastructure Transport Investments, Economic Growth and Regional Convergence <i>Xavier Fageda and Cecilia Olivieri</i>	2
Transport Modes and an Aging Society <i>Charles B.A. Musselwhite and Theresa Scott</i>	6
Sustainable Mobility Paths <i>Erling Holden and Geoffrey Gilpin</i>	13
Vehicles that Drive Themselves: What to Expect with Autonomous Vehicles <i>Michele D. Simoni and Kara M. Kockelman</i>	19
Transport Modes and Tourism <i>Ila Maltese and Luca Zamparini</i>	26
Transport Modes and Accessibility <i>Bert van Wee</i>	32
Transport Modes and Globalization <i>Jean-Paul Rodrigue</i>	38
Transport Modes and Cities <i>Erick Guerra and Gilles Duranton</i>	45
Transport Modes and Remote Areas	51
Modeling Mode Choice in Freight Transport <i>Lóránt Tavasszy and Gerard de Jongb</i>	57
Travel Mode Choice as Reasoned Action <i>Sebastian Bamberg, Icek Ajzen, and Peter Schmidt</i>	63
Energy Consumption of Transport Modes <i>Zissis Samaras and Ilias Vouitsis</i>	71
Transport Modes and People With Limited Mobility <i>Roger L Mackett</i>	85
Transport Modes and Commuters <i>Colin G. Pooley</i>	92
Shopping and Transport Modes <i>Antonio Comi</i>	98
Transport Modes and Health <i>Jennifer S. Mindell and Sandra Mandic</i>	106

Multimodality in Transportation <i>Sören Groth and Tobias Kuhnimhof</i>	118
Transport Modes and Disasters <i>Brian Wolshon</i>	127
Big Data for Public Transport Planning <i>Jan-Dirk Schmöcker</i>	134
Active Transport: Heterogeneous Street Users Serving Movement and Place Functions <i>Regine Gerike, Stefan Hubrich, Caroline Koszowski, Bettina Schröter, and Rico Wittwer</i>	140
Electric Vehicles <i>Christine Eisenmann, Daniel Görges, and Thomas Franke</i>	147
Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service <i>Susan Shaheen, PhD and Adam Cohen</i>	155
Adoption of new travel information platforms <i>Sigal Kaplan</i>	160
ICT and Transport Modes <i>Galit Cohen-Blankshtain</i>	165
Mode Choice and Life Events <i>Joachim Scheiner</i>	171
Introduction to Air Transport <i>Milan Janić</i>	178
The History of Air Transportation <i>Richard P. Hallion</i>	192
The Geography of Air Transport <i>Lucy Budd and Stephen Ison</i>	198
The Future of Air Transport <i>Rico Merkert and James Bushell</i>	203
Air Transport and Its Territorial Implications <i>Lanfranco Senn</i>	208
Next Generation Travel: Young Adults' Travel Patterns <i>Tobias Kuhnimhof and Scott Le Vine</i>	215
Airport Network Planning and Its Integration with the HSR System <i>Francesca Pagliara, Juan Carlos Martín, and Concepción Román</i>	222
Airport Management <i>Peter Forsyth and Hans-Martin Niemeier</i>	229
Airport Regulation <i>Achim I. Czerny</i>	234
Air Route Planning and Development <i>Renan Peres de Oliveira and Gui Lohmann</i>	241
Airline Management <i>Sveinn Vidar Gudmundsson</i>	249
Air Cargo <i>Volodymyr Bilotkach</i>	258

Airline Regulation <i>Andrew R. Goetz</i>	263
Air Vehicles Classification <i>Vincenzo Torre</i>	269
Aircraft Manufacturing <i>Antonio Sollo</i>	290
A Geography of Road Transport in Cities <i>Cristian Domarchi and Juan de Dios Ortúzar</i>	300
The Future of Road Transport <i>Preston L. Schiller</i>	306
Road Transportation and Territorial Scale <i>Ana M. Condeço-Melhorado</i>	315
Road Modes: Walking <i>Kevin Manaugh, PhD Associate Professor</i>	320
Bus Public Transport Planning and Operations <i>Ehab Diab and Ahmed El-Geneidy</i>	326
Street Design for Active Travel <i>Bruce Appleyard</i>	333
Transit Planning and Management <i>Zakhary Mallett and Marlon G Boarnet</i>	349
Road Infrastructure: Planning, Impact and Management <i>Jos Arts, Wim Leendertse, and Taede Tillema</i>	360
Road Transport Planning at the Urban Scale <i>David A. King and Kevin J. Krizek</i>	373
Road Traffic Regulation: Road Pricing and Environmental Quality <i>Marco Percoco</i>	378
Car Ownership and Car Use: A Psychological Perspective <i>J.L. Veldstra, A.B. Ünal, E.M. Steg</i>	384
Road Transport: E-Scooters <i>Gysele Lima Ricci and Klaus Bogenberger</i>	391
Railway Station and Network Planning <i>Ingo Arne Hansen</i>	399
Introduction to Rail Transport <i>Chris Nash and Tony Fowkes</i>	406
The History of Rail Transport <i>Carlo Ciccarelli, Andrea Giuntini, and Peter Groote</i>	413
The Geography of Rail Transport <i>Frédéric Dobruszkes and Amparo Moyano</i>	427
Rail Transport and Territorial Scale <i>Prof. Andrés Monzón and Dr. Elena López</i>	437
Railway Management <i>Vassilios A. Profillidis</i>	444
Railway Terminal Regulation <i>Nacima Baron</i>	454

Service Network Design for Freight Railroads <i>Teodor Gabriel Crainic</i>	464
Subway Systems <i>Guillaume Monchambert, Daniel Hörcher, Alejandro Tirachini, and Nicolas Coulombel</i>	471
Railway Company Management <i>Vilius Nikitinas, Skaistė Miliauskaitė</i>	479
Regulation of Rail Infrastructure and Services <i>Javier Campos</i>	485
Rail Vehicle Classification <i>Christos Pyrgidis and Alexandros Dolianitis</i>	490
Introduction to Maritime Shipping <i>Christa Sys and Thierry Vanelslander</i>	508
The Geography of Maritime Transport <i>César Ducruet and Justin Berli</i>	517
The Future of Maritime Transport <i>Harilaos N. Psaraftis</i>	535
Maritime Transport and Territorial Scale <i>Brian Slack</i>	540
Containerization and the Port Industry <i>Hercules Haralambides</i>	545
Port Management <i>Lourdes Trujillo, Daniel Castillo Hidalgo, and Manuel Herrera</i>	557
Economic and Environmental Regulation in the Port Sector <i>Beatriz Tovar and Alan Wall</i>	563
Maritime Route Planning <i>Johan Woxenius</i>	570
Inland Waterway Transport and Inland Ports: An Overview of Synchromodal Concepts, Drivers, and Success Cases in the IWW Sector <i>Behzad Behdani, Bart Wiegmans, and Yun Fan</i>	577
Maritime Company Passenger Management/Liner Industry <i>Claudio Ferrari and Alessio Tei</i>	587
Cruise Industry <i>Athanasios A. Pallis and Aimilia A. Papachristou</i>	593
International Maritime Regulation: Closing the Gaps Between Successful Achievements and Persistent Insufficiencies <i>Laurent Fedi</i>	600
Ship Classification <i>Gareth C. Burton and Mimosa T. Miller</i>	607
The Shipbuilding Industry and its Interactions With Shipping <i>Paul William Stott</i>	617
Methods for Designing Public Transport Networks <i>Zain Ul Abedin and Avishai (Avi) Ceder</i>	625
Space Transportation <i>Mark Hempsell</i>	638

Pipelines	646
<i>Franco Cotana and Mattia Manni</i>	
Women and Transport Modes	656
<i>Priya Uteng, PhD and Yusak Susilo, PhD</i>	
Transport Modes and Big Data	665
<i>Hannah D Budnitz, Emmanouil Tranos, and Lee Chapman</i>	
Transport Modes and Inequalities	671
<i>Caroline Mullen</i>	
Railway Traffic Management	678
<i>Francesco Corman</i>	
Indoor Transportation	684
<i>Lutfi Al-Sharif</i>	
Bicycle as a Transportation Mode	697
<i>Raktim Mitra and Paul M. Hess</i>	
Urban Air Mobility: Opportunities and Obstacles	702
<i>Adam Cohen and Susan Shaheen, PhD</i>	
Transport Modes and Sustainability	710
<i>Long Cheng, Jonas De Vos, and Frank Witlox</i>	

VOLUME 6

Introduction to Transport Policy and Planning	1
<i>Maria Attard</i>	
Workplace Parking Levy	2
<i>Stephen Ison and Lucy Budd</i>	
Air Transport	7
<i>Lucy Budd and Stephen Ison</i>	
Mobility as a Service	12
<i>Milo N. Mladenović</i>	
Bicycle Sharing	19
<i>Cyrille Médard de Chardon</i>	
Light Rail	31
<i>Fiona Ferbrache</i>	
Planning Tourism Travel	39
<i>Luca Zamparini</i>	
Transport Planning and Management and its Implications in Chinese Cities	44
<i>Mengqiu Cao</i>	
Road Safety	51
<i>George Yannis and Eleonora Papadimitriou</i>	
Mobility Planning and Policies for Older People	59
<i>Charles B A Musselwhite</i>	
Electric Mobility	64
<i>Graham Parkhurst</i>	
Transport Policy and Governance	73
<i>Lisa Hansson</i>	

Transferability of Urban Policy Measures <i>Paul Martin Timms</i>	77
Accessibility Tools for Transport Policy and Planning <i>Benjamin Büttner</i>	83
Demand Responsive Transport <i>Marcus Enoch</i>	87
Transport and Climate Change <i>Robin Hickman and Christine Hannigan</i>	94
Taxicabs and Microtransit <i>David A. King</i>	101
Land-Use and Transport Planning <i>Luis A. Guzman</i>	107
Parking <i>Stephen Ison and Lucy Budd</i>	113
Transport Planning in the Global South <i>Daniel Oviedo and Mariajose Nieto-Combariza</i>	118
Planning for Children's Independent Mobility <i>E. Owen D. Waygood and Raktim Mitra</i>	125
The Politics of Mobility Policy <i>Geoff Vigar</i>	131
Technology Enabled Data for Sustainable Transport Policy <i>Susan M. Grant-Muller, Mahmoud Abdelrazek, Hannah Budnitz, Caitlin D. Cottrill, Fiona Crawford, Charisma F. Choudhury, Teddy Cunningham, Gillian Harrison, Frances C. Hodgson, Jinhyun Hong, Adam Martin, Oliver O'Brien, Claire Papaix, and Panagiotis Tsoleridis</i>	135
Car Sharing <i>Cyriac George and Tanu Priya Uteng</i>	142
Toward a More Holistic Understanding of Mega Transport Project (MTP) Success <i>John Ward</i>	147
Equity Considerations in Transport Planning <i>Karel Martens</i>	154
Planning for Rail Transport <i>Simon P. Blainey</i>	161
Connected and Autonomous Vehicles: Priorities for Policy and Planning <i>Dr. Alexandros Nikitas</i>	167
Gendered Mobility <i>Sheila Mitra-Sarkar</i>	173
Modeling and Simulation for Transport Planning <i>Michela Le Pira, Giuseppe Inturri, and Matteo Ignaccolo</i>	184
Externalities and External Costs in Transport Planning <i>Silvio Nocera</i>	191
Planning for Public Transport with Automated Vehicles <i>Gonçalo Homem de Almeida Rodriguez Correia</i>	198
Urban Congestion Charging in Transport Planning Practice <i>Ida Kristoffersson and Maria Börjesson</i>	206

Energy and Transport Planning <i>Debbie Hopkins and Christian Brand</i>	214
Customer Satisfaction as a Measure of Service Quality in Public Transport Planning <i>Laura Eboli and Gabriella Mazzulla</i>	220
Car Sharing and the Impact on New Car Registration <i>Mario Intini and Marco Percoco</i>	225
Evaluation Methods in Transport Policy and Planning <i>Niek Mouter</i>	230
Transport and Air Quality Planning and Policy <i>Dr Fabio Galatioto</i>	236
Cycling Policies <i>Esther Anaya-Boig</i>	241
Community Severance <i>Paulo Anciaes and Jennifer S. Mindell</i>	246
Planning for Bus Priority <i>Claus H. Sørensen, Fredrik Pettersson, and Joel Hansson</i>	254
High-Speed Rail and the City <i>Marie Delaplace</i>	261
Public Engagement in Transport Planning <i>Miriam Ricci</i>	266
Long-Distance Travel <i>Giulio Mattioli and Muhammad Adeel</i>	272
Urban Freight Policy <i>Laetitia Dablanç</i>	278
Regional Transport Planning <i>Chia-Lin Chen</i>	286
Transport Project Financing <i>Romeo Danielis and Lucia Rotaris</i>	292
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	298
Transitions and Disruptive Technologies in Transport Planning <i>Kate Pangbourne and Maria Attard</i>	309
Community Transport: Filling the Gaps for Those in Need of Mobility <i>Ian Shergold</i>	314
Fundamental Emerging Concepts and Trends for Environmental Friendly Urban Goods Distribution Systems <i>Sandra Melo</i>	320
Plateau Car <i>David Metz</i>	324
Emerging Trends in Transport Demand Modeling in the Transition Toward Shared Mobility and Autonomy <i>Patrizia Franco</i>	331
Policy and Planning for Walkability <i>Carlos Cañas Sanz and Maria Attard</i>	340

Public Transport Subsidy and Regulation <i>Jonathan Cowie</i>	349
Urban Regeneration and Transportation Planning <i>Thomas Vanoutrive</i>	356
Social and Distributional Impact Assessment in Transport Policy <i>Laura Walker and Angela Curl</i>	361
Mobility Planning for Healthy Cities <i>Ersilia Verlinghieri</i>	368
Transport Demand Management <i>Begoña Guirao</i>	374
Planning and the Global Movement of Goods and Commodities <i>Christopher Clott and Chris Petrocelli</i>	380
Public Transport Network Planning <i>Corinne Mulley and John D. Nelson</i>	388
A Timely Perspective on Planning for Ageing Infrastructure <i>Anthony Perl</i>	395
The role of media in transport planning and the transport policy process <i>Özgül Ardiç and J.A. Annema</i>	400
Travel Plans <i>Stephen Potter and Marcus Enoch</i>	408
Home Deliveries and their Impact on Planning and Policy <i>Rosário Macário</i>	413
Planning for Safe and Secure Transport Infrastructure <i>Per Erik Gårder</i>	418
Sensors and Data Driven Approaches in Transport <i>Mohammad Sadrani and Constantinos Antoniou</i>	426
Planning Active Travel and School Transport <i>Fahimeh Khalaj, Dorina Pojani, and Sara Alidoust</i>	432
VOLUME 7	
Introduction to Transport Psychology <i>Carlo Prato</i>	1
From Self-reports to Auto-Tech-Detect (ATD)-based Self-reports in Traffic Research <i>Türker Özkan and Timo Lajunen</i>	2
Observational Field Studies in Traffic Psychology <i>Tova Rosenbloom and Hodaya Levy</i>	8
Driving Simulators <i>Karel A. Brookhuis</i>	14
Naturalistic Driving Studies: An Overview and International Perspective <i>Johnathon P. Ehsani, Joanne L. Harbluk, Jonas Bärghman, Ann Williamson, Jeffrey P. Michael, Raphael Grzebieta, Jake Olivier, Jan Eusebio, Judith Charlton, Sjaanie Koppel, Kristie Young, Mike Lenné, Narelle Haworth, Andry Rakotonirainy, Mohammed Elhenawy, Gregoire Larue, Teresa Senserrick, Jeremy Woolley, Mario Mongiardini, Christopher Stokes, Paul Boase, John Pearson, and Feng Guo</i>	20

A Detailed Approach to Qualitative Research Methods <i>Sonja Forward and Lena Levin</i>	39
Behavioral Change <i>Sonja Haustein</i>	46
Habitual Behavior <i>Carlo G. Prato</i>	54
Driving Behavior and Skills <i>Timo Lajunen and Türker Özkan</i>	59
The Multidimensional Driving Style Inventory <i>Orit Taubman - Ben-Ari</i>	65
Risk Perception in Transport: A Review of the State of the Art <i>Trond Nordfjrn, An-Magritt Kummeneje, Mohsen F. Zavareh, Milad Mehdizadeh, and Torbjørn Rundmo</i>	74
Social-Symbolic and Affective Aspects of Car Ownership and Use <i>Birgitta Gatersleben</i>	81
Pitfalls of Statistical Methods in Traffic Psychology <i>J.C.F. de Winter and D. Dodou</i>	87
Data Analysis: Structural Equation Models <i>Marco Diana</i>	96
Data Analysis: Integrated Choice and Latent Variable Models <i>Carlo G Prato</i>	102
Explaining Data Analysis Using Qualitative Methods <i>Lena Levin and Sonja Forward</i>	107
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	113
Driver Aggression and Anger <i>Mark J.M. Sullman and Amanda N. Stephens</i>	121
Speeding: A "Tragedy of the Commons" Behavior <i>Bryan E. Porter, Thomas D. Berry, and Kristie L. Johnson</i>	130
Cycling as a Mode Choice: Motivational Psychology <i>Sigal Kaplan</i>	136
Motorcyclists <i>Narelle Haworth</i>	144
Drivers' Hazard Perception Skill <i>Mark S. Horswill and Andrew Hill</i>	151
Driver Education and Training for New Drivers: Moving beyond Current 'Wisdom' to New Directions <i>Teresa Senserrick, Oscar Oviedo-Trespalcacios, David Rodwell, and Sherrie-Anne Kaye</i>	158
Road Safety Advertising: What We Currently Know and Where to From Here <i>Ioni Lewis, Barry Watson, Katherine M. White, and Sonali Nandavar</i>	165
Traffic Law Enforcement Theories and Models <i>Richard Tay</i>	171
Satisfaction with Travel and the Relationship to Well-Being <i>Tommy Gørling and Filip Fors Connolly</i>	177
	182

Electromobility: History, Definitions and an Overview of Psychological Research on a Sustainable Mobility System <i>Josef F. Krems and Isabel KreiÃŸig</i>	
Sharing: Attitudes and Perceptions <i>Yi Wen and Christopher R Cherry</i>	187
Women's Travel Patterns, Attitudes, and Constraints Around the World <i>Sandra Rosenbloom</i>	193
Driver Stress and Driving Performance <i>Lisa Dorn</i>	203
Introduction to Sustainability and Health in Transportation <i>Roger Vickerman</i>	225
Sustainable Development Goals and Health <i>Rosa Surinach</i>	226
Human Ecology <i>Roderick J. Lawrence</i>	234
Car- Free Cities <i>Haneen Khreis and Mark J. Nieuwenhuijsen</i>	240
Superblocks Base of a New Model of Mobility and Public Space. Barcelona as an Example <i>Salvador Rueda Palenzuela</i>	249
Resilience of Transport Systems <i>ErikJenelius and Lars-GöranMattsson</i>	258
Impact of Shipping to Atmospheric Pollutants: State-of-the-Art and Perspectives <i>Daniele Contini and Eva Merico</i>	268
Noise Pollution From Transport <i>Marianna Jacyna, Emilian Szczepański, Konrad Lewczuk, Mariusz Izdebski, Ilona Jacyna-Gotda, Michał, Kłodawski, Paweł, Gołda, Piotr Goł, and Ebiowski</i>	277
Visual Impacts From Transport <i>Paulo Rui Anciaes</i>	285
Light Pollution <i>John D. Bullough</i>	292
Wildlife Crossings and Barriers <i>Scott D. Jackson</i>	297
Environmental Justice, Transport Justice, and Mobility Justice <i>Devajyoti Deka</i>	305
Transport Noise and Health <i>Elisabete F. Freitas, Emanuel A. Sousa, and Carlos C. Silva</i>	311
Climate Change and Health, Related to Transport <i>Ersilia Verlinghieri</i>	320
Urban Greenspace, Transportation, and Health <i>Payam Dadvand and Mark J. Nieuwenhuijsen</i>	327
Transport Access and Health <i>Alireza Ermagun</i>	335
Social Exclusion and Health, Related to Transport <i>Roger L. Mackett</i>	341

Burden of Disease Assessment <i>David Rojas-Rueda</i>	347
Achieving a Near-Zero CO <i>Lewis M. Fulton</i>	353
Disabled Travelers <i>Bryan Matthews</i>	359
Health Impacts of Connected and Autonomous Vehicles <i>Soheil Sohrabi</i>	364
Electric Vehicles and Health <i>Kanok Boriboonsomsin</i>	372
Shared Mobility Opportunities and their Computational Challenges for Improving Health-Related Quality of Life <i>Cristiano Martins Monteiro, Cláudia Aparecida Soares Machado, Adelaide Cassia Nardocci, Fernando Tobal Berssaneti, José Alberto Quintanilha, and Clodoveu Augusto Davis</i>	376
Bike Sharing and Health <i>David Rojas-Rueda and Mark J. Nieuwenhuijsen</i>	384
E-Bikes and Health <i>Aslak Fyhri and Hanne Beate Sundfør</i>	393
<i>Index</i>	399

Introduction to Transportation Safety and Security

Per Gårder, Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States

© 2021 Elsevier Ltd. All rights reserved.

The reason we travel is a need or desire to get somewhere else. People typically prioritize how to get to their destination quickly and comfortably. Most of the time, it is assumed that we will get there safely. But as parents, we may worry about our children's safety and in a snow storm we may worry about our own safety. However, as a society, we should be concerned about the lack of safety for all of us, not least the elderly and the disabled, in our road transport system. We are far from safe as pedestrians, bicyclists, or motorists. The perceived safety and the objective safety often do not coincide. Lots of people fear to fly in big jets but do not worry about driving to the airport.

This section of the encyclopedia has an emphasis on transportation safety, and how safety is managed for all modes of transport. But it also has articles that deal with personal security, such as the risk of being assaulted—from verbally abused to raped or murdered—while traveling and how to evacuate areas safely during weather events such as hurricanes and flooding, events that are becoming increasingly common because of climate change. An area that is not directly dealt with in this section is how infectious diseases, such as Covid-19, can spread from continent to continent by airplanes and ships, and how local transmission can be facilitated by public transportation. Also, breathing polluted air while traveling can certainly make us “unsafe.” Other parts of the Encyclopedia deal with such issues. In general, the overall relationship between transport and health has become increasingly acknowledged and Vision Zero programs—where the goal is to eliminate death and serious injury—are discussed in this section with respect to accidents and crashes. But for a transportation system to be safe, it also has to guarantee personal security and make sure we do not get exposed to unhealthy emissions, be they from that transportation system itself or from other sources.

The chapters, in this section present a range of views across these topics. Perspectives from and policies in North America, Europe, Asia, Australia, and Africa are brought up, but some articles are limited to practices in only a few countries. However, implications can and should be drawn to how these practices could be transferred to other regions, not least to less developed nations. All the articles are based on current state-of-the-art research. Some topics are of philosophical nature whereas others deal with actual government policies and yet others with how design details may affect the safety of users. Many of the chapters have cross-cutting themes with chapters in other sections of the Encyclopedia, especially those dealing with Transport Planning and Policy, and Sustainability of Transport Systems, but also with Psychology, and readers are encouraged to explore these to gain a fuller understanding of safety and security issues.

The Concept of “Acceptable Risk” Applied to Road Safety Risk Level

Claes Tingvall^{*,†,‡}, Anders Lie[†], ^{*}AFRY, Solna, Sweden; [†]Chalmers University of Technology, Gothenburg, Sweden; [‡]Monash University Accident Research Centre, Melbourne, VIC, Australia

© 2021 Elsevier Ltd. All rights reserved.

Background	2
The Concept of Acceptable Risk	3
Setting a Predefined AR for the Road Transport System	3
Implications and Consequences of Setting a Predefined AR Level for Road Traffic Safety	3
Discussion	4
Acknowledgments	5
References	5
Further Reading	5

Background

The concept of risk is used extensively in the community describing the possible or observed rate of an occurrence for a variety of events. Risk can apply to the probability of change to aspects relating to the economy, health, climate, safety or anything else. It can be seen as the normal or abnormal chance of a certain outcome or scenario. Risk can also denote the probability or chance of us as individuals being exposed to a road traffic crash with a certain outcome.

The risk is often described as a population risk, or a risk to be injured or maybe killed as a user of a certain kind of vehicle or mode of mobility. These two risk examples can be seen as both the societal risk and as the risk for any one citizen exposing him or her to a certain type of road traffic. We can also refer to these risks as being seen from a macro or microperspective. In the macroperspective, we “know” that a large number of people will be killed in roadway crashes in X-land in the next year. But, in the micro perspective, seen from one individual’s viewpoint, it is very unlikely that he or she will be killed in the next year.

The more intriguing or maybe philosophical question is who we address with this risk estimate. Is it society as a sort of a provider of the road transport system or is it the sum of individuals that collectively are willing or maybe are forced to take the risk?

It is clear that society consider any road user, or anyone appearing on a road, as an active individual with a legal responsibility to minimize risk. Risky behavior, irrespective of age, situation or role, is illegal from a legal standpoint according to the Vienna Convention. The Vienna Convention is an international treaty designed to facilitate international road traffic and to increase road safety by establishing standard traffic rules among the contracting parties. The convention was agreed upon at the United Nations Economic and Social Council’s Conference on Road Traffic in Vienna in 1968 and came into force in 1977. It replaces previous road traffic conventions, most notably the 1949 Geneva Convention on Road Traffic, which was ratified by 98 countries. The Vienna Convention has been ratified by 78 countries, and those who have not ratified it are still parties to the 1949 Geneva Convention if they had ratified that. Canada, Ireland, and the United States are examples of such countries. The Vienna Convention gives guideline for national road traffic rules and is detailed in what responsibilities that falls on the road user, but does not specify any rules for the providers of the system, other than signs and markings.

Any crash, in theory, can be tracked to an individual and his or her “illegal” behavior. This makes the road user a primary agent, irrespective of the technical solutions, design of roads and streets as well as vehicles. Implicitly, this means that the collective of road users are the prime source of actions and behavior resulting in crashes, injuries and deaths. Society, through the providers of the road transport system, can decide to set up requirements for vehicle standards and for road design, in order to reduce risk. One requirement for doing so is, however, that the action taken should be “cost-effective,” albeit in a broad sense and taking into account both budgetary consequences and the willingness pay or act among citizens (Elvik, 1999). It is implicitly or through explicit policies normal to consider an investment or action to be ineffective if the costs, in a broad sense, are larger than the benefits.

Despite rules of the road, the resulting consequences of the individuals’ and society’s way to deal with risk can be seen as disappointing. Each year, more than 1 million individuals are killed in road crashes across the globe (Academic Expert Group, 2019). This corresponds to an annual risk of death of around 175 deaths per million humans, or around 9000 per million lifetime risk, which is the same as 0.9% probability that an individual will be killed in a road crash. In countries with the lowest numbers of deaths, corresponding results are 20 deaths per million human, annually, and 2500 deaths per million lifetime risk. The difference between countries with the lowest risk and the highest risk seems to be in the magnitude of 10–15 times.

In other transport modes, like aviation or railway, the risks to the passengers are at a very different level compared to road traffic. In civil aviation, the level is in the magnitude of 300 deaths per year across the globe. Seen as a population risk, we have less than 0.5 death per million annually. When we measure against travel distances or number of passengers, the difference between road and aviation is smaller but still substantial. Instead of a risk ratio between road and aviation based on population (3000–4000 times) it would be in the order of 1000 times.

The Concept of Acceptable Risk

The varying risk levels in the transport sector, and the risk levels in other sectors, ask for a deeper analysis of the reasons for why this has been considered acceptable. Why has it not created a sense of human catastrophe with such risks in the road transport sector? One hypothesis would be in line with the concept of acceptable risk (AR) (Fischhoff et al., 1981; Hunter and Fewtrell, 2001).

There are two general principles for setting or accepting a level of AR (Hunter and Fewtrell, 2001). In activities where humans themselves judge their personal risk, there are clear possibilities to generate what seems to be acceptable from the participation in such activities. This is sometimes called a “revealed preference risk.” In the literature, there are a number of suggestions on what influence such acceptance. The level of self-control seems to be one of them. The consequences of a failure or mistake by the individual involved in the activity as well, and if the risk is voluntary or not. Implicitly, a revealed risk is the accumulated risk perception or risk behavior of individuals in relation to their own life and health. It seems that we use the revealed AR as the norm in road traffic management, investments, and solutions.

The other general principle is about setting a predefined AR. This is common in technical systems, medicine, occupational health and safety, nuclear power, aviation and railway, etc. A rationale for predefined AR is that the activity or system is “manmade.” Implicitly there is also an understanding of a predefined risk as a risk someone else is exposed to by a system, activity or action taken in relation to someone else’s life and health.

When predefined an AR, there are both similarities and differences between sectors. What they have in common is that a predefined risk is generally very low, and measurable. Below are a few examples from different sectors.

One example that is relevant for the road transport sector would be the railway system. Here, we can find not only an overall AR but also the AR broken down to subsystems and components.

Setting a Predefined AR for the Road Transport System

It seems to be quite simple to set a predefined AR for the road transport system by adopting the general level from systems like aviation and railroads. The more serious question is what consequences this would have on the design, function, and access to road transport.

The examples from railway and aviation are both functional and relevant down to specific solutions and locations. The EU regulation on railway safety (EU regulation 352/2009 on common safety regulation on risk evaluation) seeks to standardize risk calculations on both known and new solutions. And there are AR levels predefined as well as a definition of safety. This definition is “freedom from unacceptable risk of harm.”

The “proposer” of a new solution needs to either compare the solution to other comparable solution or to undertake an explicit risk estimation of the proposed solution. The national risk level for railroads is supposed to be smaller than 10^{-9} per operating hour for any technical failure that would be safety sensitive. A new solution may not lead to that this level is exceeded. It seems that the AR also accounts for cases where we have a man–machine interface, that is, there is a human involved in the technical system.

A risk level of 10^{-9} failures per operating hour on the transport system, for each train-set or vehicle, would lead to one failure per 114,000 years of operation, or 8.8 failures per million vehicle hours. In practice, it is very hard to evaluate if this is met, but if we look at fatalities in the European rail system, there have been very few catastrophic crashes, not even one per year for each of the 28 member states. Eight people were killed in collisions as passengers of trains in 2015 and none in a derailment. If the EU population, 508 million in 2015, each traveled by train, on average, 10 min/day, and we have an AR level of 10^{-9} fatalities per person-travel-hour, rather than operating hour, we would accept 31 fatalities per year for EU-28, and, with 8 fatalities, we met the goal for 2015. The 10 min rail travel per time may be a high guesstimate, but as long as the average rail travel time is at least $8/31 \times 10 = 2.5$ min/day, the goal is met. On the other hand, more than 1700 people were killed as third party individuals by trains that same year. The victims were mostly car occupants or pedestrians on railway crossings; or simply individuals crossing railway tracks at locations with no regular railway crossing. If we apply the same AR for the roadway system, and assume people travel for one hour per day on average, we would accept 185 fatalities per year in all of EU-28. In reality Europeans spend less than an hour a day in cars, so a slightly lower number should be accepted. However, people also walk in traffic, and if we include pedestrian fatalities in the numerator, we could also include time spent in traffic walking, and one hour per day is probably a reasonable estimate. In reality 26,100 people were killed in roadway crashes in EU-28 in 2015. So, we have around 140 times more people killed on roadways in the EU than what is considered acceptable, and met, by the railway industry. And, the EU is certainly not alone here. Let us look at the US, using 2015 data. The US had around 321 million inhabitants that year and if everyone traveled one hour per day, we would accept up to 117 fatalities in the year if we use a risk level of 10^{-9} fatalities per travel hour. The actual number of fatalities was 35,092. That means we had a fatality risk on US roads in 2015 that was almost exactly 300 times the acceptable one for passengers in the rail system.

Implications and Consequences of Setting a Predefined AR Level for Road Traffic Safety

While it might be stated that the question of whether we should apply a revealed AR perspective or a predefined AR for road traffic safety is political or even philosophical. On the other hand, there is no doubt and some quite clear examples where the design and operation of the road transport system decide the risk level, rather than the users' choice of risk taking. Using a roundabout rather than a traditional intersection is one example. Replacing centerline striping on rural 2-lane highways with cable barriers is another example, and there are of course plenty of other examples. The question is though what would follow if we change from revealed to predefined AR? There would be a number of both real consequences as well as implications if we would set a predefined acceptable level. The most obvious would be to decide at what level it should be set, and for whom it would apply. It seems natural to choose the same level as for railway and aviation. At least this would be natural for passengers in cars, buses and alike. To passengers we would also add those in an automated vehicle, including third party road users outside an automated vehicle, in essence pedestrians, cyclists, etc.

But the implications would go far wider. Both design of infrastructure and vehicles would have to be chosen on the basis on a predefined risk level, meaning that only the best possible solutions would be used. And the kinetic energy, that is, speed, would have to be based on the predefined level.

Even if we decided to choose a predefined AR, and in reality this is the case for more or less the whole world, it would still be a question of how it should be applied to different parts of society. Would policy decisions, ethical rules or something else be enough? Or do we have to use the examples from aviation and railway to bring in legislation, duty of care, and require continuous improvement to change the way the system is operated?

Discussion

In this article, it is questioned whether we should consider a predefined AR level for road transportation, and if such adoption would be not only useful but also in fact logical. If we step away from the thought that the AR in road transportation is more or less revealed, we would be offered to take quite a large step in what the society can accept in terms of risk to human life and health.

On the other hand, the UN (Academic Expert Group, 2019), and many jurisdictions including the US, the EU, Australia etc. have already adopted Vision Zero (Lindberg and Håkansson, 2017) or Safe System which are nothing but predefined AR policies. But we still build roads and streets, design vehicles and operate transport services as if we still have revealed AR as our approach.

The transport policy seems to still be based on the economic theory of balancing mobility to safety, and that investments and initiatives must be economically sound. The economic theory seems to still see the road user as the agent, and that risks taken could be considered as the correct revealed risk.

One might describe road safety countermeasures that inflicts with humans as paternalistic and a sign of a “nanny state” (Eberhard, 2006). But it could easily be mirrored by the notion that protecting other human's life is a duty that has been around for thousands of years (Hippocrates 400 BC) and codified in legislation, ethical rules, and duty of care practices, and, as shown, also in the principles of predefined AR (Hunter and Fewtrell, 2001).

What it all comes down to is if road traffic safety performance is mainly generated by individuals making informed choices of their risk, or by professional providers of a road transport system caring for the citizens' life and health? It can be shown empirically (Hauer, 2015; Krafft et al., 2008; Lindberg and Håkansson, 2017) that design solutions of roads, streets and vehicles can have a dramatic effect on the risk of death and serious injury. And it can also be shown that some stakeholders in reality can get very close to zero deaths (Rizzi et al., 2019).

If we on large accept the role of the providers as the main agents for safety, it makes sense to adopt the principles of predefined AR, and at a level close to what seems to be normal, that is 1 per million population lifetime risk. (Since people spend, on average, roughly 1 h/day in traffic, over 80 years, a person would spend around 30,000 h in traffic and if we accept a risk of 10^{-9} per hour, we would accept a lifetime risk of 30 per million people. So 1 per million people is a high standard getting close to no one getting killed which is the ultimate goal of Vision Zero.) Globally, that could be translated to a maximum of less than 100 deaths per year. In comparison to the overall current level of 1.35 million deaths per year (Academic Expert Group 2019), the difference is gigantic. But it is a clear sign of how wrong we got it from the beginning; and maybe a sign that the recent socio-economic approach to balance mobility to safety was not only ethically misused but also detrimental to the life and health of humans.

Economic theory claim that there is a marginal decrease of the return of investment, that is, to save lives will come at a gradually higher cost of the community (Elvik, 1999). Even this argument has been proven to be false many times for road traffic safety (Tingvall and Lie, 2017). Saving lives today is cheaper than ever as new innovations are implemented (Rizzi et al., 2019), something the socio-economic models applied to mobility seem to have overlooked (Tingvall and Lie, 2017).

The implications of adopting a predefined AR for road traffic would be far going. First of all, mobility would come out as a function of safety (Haddon, 1970; Tingvall and Lie, 2017). Travel speed would have to be aligned with the road infrastructure design and the vehicles used. Secondly, a new product, road, street or vehicle would have to be designed on the basis of “best practice.” None can repeat a mistake or inferior design solution (Hauer, 2015) as this would be in breach of any idea behind a predefined AR.

The final question seems to be how we as a society in theory have abandoned the economic trade-off model by adopting a new policy that road crash fatalities should reach zero in the long term, while in practice, we are still using trade-off practices (Academic Expert Group, 2019)? How come we are still building inferior road infrastructure solutions, building cars without the best seat belts,

accepting speed and speed limits well beyond a zero policy? There is no clear answer to these questions, but maybe it is time to do what railway and aviation did many years ago. To not only adopt general principles but also to build a fence of rules, regulations, standards, and management practices that gradually improve the system. And to a level that is close to the golden standard of predefined AR.

The introduction of the 2030 Agenda principles for traffic safety (Academic Expert Group, 2019) goes quite far in recommendation for the allocation of responsibility for road transport safety. In relation to multinational corporations, it is recommended that their value chains should be transparent and reported as to what they expose employees, third-party citizens, and customers to in terms of safety. This is a starting point for such stakeholders to openly declare their targets and results. And it would come as a surprise if they would not accept a “zero tolerance” policy for their impact in the road transport system. And that would probably be in line with the global financial system expectations in relation to sustainability. As would a similar adoption of “zero tolerance” be expected for public procurement, which would mean that the economic logics would take a new turn in the history of traffic safety.

Acknowledgments

The author wants to thank Dr. Per Gårder for actively contributing to the content and participating in writing a couple of the sections of this article.

References

- Eberhard, D., 2006. I Trygghetsnarkomanernas land (In the country of security addicts). Sverige och det national paniksyndromet (Sweden and its national panic syndrome) Norstedts.
- Elvik, R., 1999. Can injury prevention go too far? Some reflections on some possible implications of Vision Zero for road accident fatalities. *AAP* 3 (3), 265–286.
- Fischhoff, B., et al., 1981. *Acceptable Risk*. Cambridge University Press, New York.
- Commission., EU Commission. EU regulation no 352/2009. in press.
- Hauer, E., 2015. An Exemplum and its road safety morals. In: Presented at the International Symposium on Transportation Planning and Safety, New Delhi.
- Haddon, W., 1970. On the escape of tigers. An ecological note. *Am. J. Public Health* 60 (12), 2229–2234.
- Hunter, P.R., Fewtrell, J., 2001. In: Fewtrell, L., Bertram, J. (Eds.), *Acceptable risk*. WHO Water: Guidelines, Standards and Health. IWA Publishing, London.
- Lindberg, H., Håkansson, M., 2017. Vision Zero 20 years. ÅF.
- Krafft, M., Stigson, H., Tingvall, C., 2008. Analysis of a safe road transport system model and analysis of real life crashes and the interaction between the human, vehicles and infrastructure. In: *Proc 20th ESV Conference Lyon 2007*. TIP, 9:5 (under slightly other title), pp. 463–471.
- Rizzi, M., Hurtig, P., Sternlund, S., Lie, A., Tingvall, C., 2019. How Close To Zero Fatalities Can Volvo Cars Get By 2020? An Analysis of Fatal Crashes with Modern Volvo Passenger Cars in Sweden. In: *Proc ESV Conference*.
- Saving Lives Beyond, 2020: The Next Steps - Recommendations of the Academic Expert Group for the Third Ministerial Conference on Global Road Safety 2020. STA 2019.
- Tingvall, C., Lie, A., 2017. Traffic safety from Haddon to vision zero and beyond. In: *Blue Book of Automobile Safety*. Social Science Academic Press, China, pp. 316–333.

Further Reading

- Eurostat, 2017. *Railway Safety Statistics*.
- SafetyNet, 2009. *Cost-Benefit Analysis*.

Crash Not Accident

Robert A. Scopatz, VHB, Inver Grove Heights, MN, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction and Background	6
Are Crash and Accident Synonyms?	7
The Chain of Causality	7
What we Call Things Matters	7
An International Understanding	8
Why Not Wreck or Collision?	8
What does the Science Say?	8
Intentional Acts Should Never be Called Accidents	9
Why Bother?	9
Acknowledgments	9
References	10
Further Reading	10

Introduction and Background

This article explains the case for using crash versus accident to describe whenever a motor vehicle collides with a person, another vehicle, or a roadside appurtenance, or follows a path of solitary destruction. While it is difficult to remain neutral on the word choice, the treatment here is intentionally unbiased and matter-of-fact so that readers may draw their own conclusions.

As a behavioral safety advocate and as someone who believes that words matter, my sympathies lie strongly in favor of using the word crash and I think my colleagues, and the world at large, agree. The truth is more complex. Some colleagues resist the change in terminology, some outright reject it, and some say it does not matter and we can use both words interchangeably. Notably, very few still advocate for accident as the universally accepted term. The shift in viewpoint is traceable back to William Haddon, the first Administrator of the National Highway Traffic Safety Administration (NHTSA) in the 1960s. The story goes that Haddon would fine anyone who used the word accident in a meeting. Thus, NHTSA has, right from the start, avoided the word accident. The logic is simple and straightforward. For those who prefer to use the term crash, they point out that accidental events must meet two criteria: (1) it is unintentional, and (2) it could not have been foreseen and thus was unavoidable. Of course, not everyone agrees that both criteria can be applied—the vernacular use of accident could point to any unintentional events or consequences regardless of how avoidable they are.

We will come back to consider this point with respect to traffic crashes, but should first reflect on how common the term accident is in industrial settings. Observers have pointed out that the term accident sets up an expectation of failure. If these events are unintentional, unforeseen, and (perhaps) unavoidable, we certainly will have more of them in the future (Smith, 2018). Historian Peter Norton traces the terminology in the context of industrial safety in the early 1900s, noting that companies sought to avoid blame by suggesting that harmful events were accidental, and certainly the companies themselves were not to blame. Automakers encouraged using the word accident to avoid the perception that cars and drivers were at fault in an era when they were routinely blamed for harmful events involving motor vehicles. In the 1920s, the auto industry even organized to change semantics to cast blame on pedestrians hit by autos and lobbied legislatures to that effect.

Traffic safety has a long history of officially using the word accident in our standard documents, including one of the pillars of the field, the American National Standards Institute (ANSI) D16 Manual on the Classification of Motor Vehicle Accidents (ATSIP, 2017). Up through the 7th edition the title and all relevant labels for events used the term accident. The 8th edition of ANSI D16 (released in 2017) has a new title, Manual on the Classification of Motor Vehicle Traffic Crashes, and the word crash replaces the word accident throughout. Chapter 1 of the latest edition of ANSI D16 begins with a note that the change from accident to crash was a deliberate step to avoid the implication that these events are unpreventable. It goes on to acknowledge that there are some who still use the term accident (including some state legislatures) and that treating the terms as synonyms by simply replacing accident with crash without any attempt to separate the avoidable from unavoidable events is not the intent.

Internationally, the terminology debate has not had quite the level of attention as it has in the United States; however, the discussion has had definite impacts. Transport Canada has adopted the term crash for their official traffic safety publications. The Traffic Injury Research Foundation (TIRF)—a Canada-based, independent, charitable road safety research institution—has promoted the switch from accident to crash (TIRF & DIAD, 2017). The *British Medical Journal* (BMJ) has, for most purposes, banned the use of accident (Davis and Pless, 2001). Outside these examples, however, it is easy to draw the conclusion that the concerns over terminology are not universal. It is also clear that not all share the same two-criterion definition of what makes an accident.

Are Crash and Accident Synonyms?

The answer is it depends on whom you ask. If the word crash focuses on the preventable nature of most of these events, and the word accident implies that they are unpreventable, the terms really do mean completely different things. By this view, the change in terminology in ANSI D16 truly recognizes a shift in understanding the nature of these events—that they are preventable. One source of contention is that some feel the implementation costs and the level of effort required to change existing legislation outweigh the benefit of adopting a more precise word. If no State had ever adopted the word accident in their traffic safety laws, it is likely that the debate would have been settled long ago. Since many states did write the definition of an accident into their traffic laws, changing that to use the crash term instead meets with some resistance. One of the easiest fixes proposed in these situations is to explain to legislators that crash is the preferred term, but that crash and accident are synonymous and thus a complete rewriting of all the laws to focus on the avoidable nature of crashes is not necessary—it is simplest to change just the word (and nothing else) in the laws. Proponents of the crash terminology are not happy with this solution, but would likely agree to the word-change-only approach in lieu of continued use of accident as a term of art. Others, the most avid proponents of the term crash, still argue that it is not synonymous with accident and that, should a state's laws imply that these events are truly accidental much more might need to change in the laws beyond just this single word choice. It's not an impossible task, but the level of effort is much higher. An effort that safety advocates clearly sees as worthwhile and necessary.

The Chain of Causality

Proponents of the term crash point to studies of crash causation as proof of their non-accidental nature. If we can find at least one causal factor in a crash, it makes sense that, were that cause removed or mitigated, the crash would not have occurred. We know that in more than 90% of crashes a human did something that contributed to the crash. Human-centered contributing factors include execution errors (mistakes in control inputs), misperceptions, errors in judgment, and lapses due to inattention and distraction. Then there are the more egregious errors such as choosing to drive while impaired by drugs or alcohol, and even the purposeful use of a vehicle as a weapon (and this includes aggressive driving). There are many other potential crash causes beyond the human factor—roadway conditions, adverse weather, debris, vehicle systems failures, and more. Ultimately, the percentage of crashes that can truly be labeled as accidental is quite small. The events were not caused some external, uncontrollable force that a prudent driver could not have dealt with.

That we can almost always identify multiple contributing factors for a crash means that it is not accidental based on the two criteria listed earlier. But, there are some limitations to even this perspective. In traffic safety and crash analysis, a forensic analysis is not typically performed to uncover the specific causes. Crash reconstruction arriving at a causal attribution is performed for only the most serious crashes involving a fatality, or where a serious criminal or civil case is likely. For the vast majority of crashes, the data indicate contributing factors based on a police officer's judgment. The determination of fault is not necessarily straightforward based on the crash report alone. The officer may indicate the most-at-fault person as a matter of departmental policy, but that does not mean there are not other contributing factors. It also does not mean that we can pinpoint the exact cause even if we know which person is most likely responsible. Even with complete and accurate data on crashes, cause can be an elusive concept. If this is true, then is there still room to conclude that all such result from the accidental confluence of multiple contributing factors? If the road had been less slick, the driver's speed might not have been a problem, if the driver had not been briefly blinded by the sun refracting off a dirty windshield, and so on.

Again, the simple answer is that most crashes can be traced back to the involved parties doing something wrong, including lacking the skill to operate the vehicle safely in the conditions they encountered. Inattention is not accidental—it is a choice. Distraction is a choice. Failure to keep the vehicle in good operating condition is a choice. Failure to slow down when the road is wet is a choice. Failure to use protective devices is a choice. Ultimately, while we might all agree that people rarely intend to do harm, it is difficult to call a crash unavoidable. Some may insist on calling a crash an accident if there was no obvious human error. Contributing factors such as a slippery road surface or a vehicle system failure might make the crash appear accidental, but advocates for the term crash point to the avoidable aspects. In other words, crashes are not accidents if the people involved could have used safer behaviors in response to conditions.

What we Call Things Matters

The final argument for the term crash is the most compelling to me as a behavioral and cognitive psychologist. It is about changing attitudes through word usage. The traffic safety practice has many issues where naming conventions matter. Observations from the sciences of human factors, cognition, and social psychology are relevant. Considering early implementations of autonomous vehicle technology is instructive. In several documented incidents, the drivers were operating in the so-called self-driving mode under circumstances where its use was specifically warned against by the manufacturer. A big part of the problem is that people fail to read and understand their vehicle owner's manual, but there is also a serious concern that what the systems are called is at least partly to blame for an overreliance on the technology. In short, owners believed the car could truly operate autonomously in all situations because they believed the labels for the assistance technologies. What we call something obviously matters.

The word crash means something different from the word accident. Crash does not call to mind something unavoidable. When people hear the word crash they are apt to think of a serious event or consequential system failure. It evokes an image of a serious situation caused by a problem in how a thing was designed, constructed, or used. It even sounds like what it is in a nearly perfect example of onomatopoeia—although advocates would caution that events that do not make that noise are still counted as crashes. Conversely, if we use accident, that could be construed as if no one is at fault and the events happened for reasons beyond anyone's ability to control or predict. Industrial safety professionals have argued that the word accident should not apply to any situation involving interactions of people and systems—that its use is off target in industrial settings as much as it is for traffic safety. Some go further and argue that we live in a safety culture (specifically a traffic safety culture is meant to imply that whether we know it or not, and even if the culture is one of dangerous behavior, a safety culture exists and we all share the norms of that culture). Part of having such a culture is agreeing on the terms and definitions we will use—again, we generally agree on these things even if our safety culture is one that yields poor outcomes. By these views, the issue nearly settled—accident is too narrow and specific to encompass all of the events we are describing.

An International Understanding

I live and work in the United States. I have many international colleagues in traffic safety and I took this assignment seriously enough to find out if a similar word choice discussion has occurred in their countries. The International Standards Organization's English language versions of the relevant standards (e.g., ISO 39001 on road traffic safety management) use crash and accident interchangeably. Based solely on my own language skills, I looked first at countries where English is the dominant language. In Canada, the traffic safety publication by Transport Canada uses crash exclusively. TIRE, a research group based in Canada but with an international reach, has clearly advocated in favor of crash as the preferred term. One UK-based journal, the *BMJ* has banned the word accident. In Australia the discussion has taken place along the same lines and the safety practitioners have settled on crash. A scan of newspaper articles from these countries shows the same variability as seen in the United States, with crash, accident, and various other terms used synonymously.

For the remainder of the world, however, this discussion of terminology appears not to have taken place and it does not appear that there is a strong movement away from the equivalent term for accident. Indeed, I am informed by some colleagues outside the English-speaking parts of the world that this discussion seems quaint and baffling from their perspective. Even in places with strong safety cultures—as in Northern Europe where the focus on safety has a long history—the debate over terminology is reportedly a nonissue. In other countries there appears to be a firm commitment to the idea that crashes are accidental in nature—that nobody planned to crash and thus the event itself was an accident. This latter viewpoint, of course, dispenses with the secondary understanding (advocated for in English-speaking nations) that accident implies a degree of unpreventability. As noted earlier, this is the source of the concern for practitioners making the case for crash. But, as discussed earlier, even in the United States and Canada, preventability argument is not universally accepted. It is a weakness of this article that I do not possess the language skills needed to truly parse the discussions of terminology in multiple languages around the globe. I have relied on a set of busy international colleagues whose English is better than my skills in any of their country's native languages. I make no claims to having provided proof that similar discussions are not happening or that the non-US-English cognates for accident clearly mean only that an event was unplanned. Nor do I claim that other countries are behind the United States in their concern over traffic safety. Popular accounts of street protests in multiple countries over lax traffic safety laws or enforcement show that people do care deeply about their ability to travel safely.

Why Not Wreck or Collision?

First, let's look at collision as a substitute. It was popular for a time as it avoids the connotation of things being purely accidental. A problem arises for safety practitioners because collisions are specific types of crashes—they involve the colliding of a motor vehicle with something or someone else: another car, a barrier, a pedestrian, or a bicyclist. There are crashes that do not involve collisions such as untripped rollovers. So, collision is too narrow a term. Wreck is also not a good substitute for crash. One obvious problem is that wreck conjures up images of (perhaps even ruinous) damage to the vehicles. If a driver got in a wreck, it typically means that their vehicle sustained serious damage. Therefore, wreck is not an appropriate substitute if there was no vehicle damage. One does not get in a wreck at low speeds with a pedestrian, even if the pedestrian is injured. That is definitely a crash.

What does the Science Say?

As of the date I wrote this, there are no studies that provide direct evidence of safer behavior among those who adopt the term crash versus accident. There are studies that demonstrate an impact of word choice on things such as perceived risk and blame ([Fausey and Boroditsky, 2010](#)). Not surprisingly, passive wording convinces people that there is no blame to be apportioned, while active wording does the opposite. In the scientific literature for traffic safety there is no clear consensus. Authors studying crashes often even use both terms interchangeably. One of the top journals in the field has the word accident in its name. All of that may change in the

future. I expect that a study of word choice and perceived risk would show a clear difference between the words crash and accident, but we do not know that yet. It appears likely that journals publishing scientific articles on the topic will come to favor crash over accident in their titles and in their style guidance for authors. It is a slow process, but one that is moving in an obvious direction.

Intentional Acts Should Never be Called Accidents

Traffic safety analyses in the United States and many other countries generally exclude events involving deliberate use of a vehicle to harm oneself or others or to cause damage. ANSI D16 explicitly excludes obviously intentional events, as does the guidance from NHTSA for the Model Minimum Uniform Crash Criteria (MMUCC) and Fatality Analysis Reporting System (FARS). These are not accidents, but they are also not part of a typical crash analysis. A quick search of media coverage of intentional events shows a disturbingly frequent use of the word accident for what were obviously deliberate acts. A person using their car to purposefully run over pedestrians was not in an accident. A person committing suicide by driving off a cliff did not cause an accident. From a traffic safety perspective, analysts drop these intentionally harmful events from the databases we use for analyses.

Exclusion criteria are, however, quite narrowly defined. From a safety analysis perspective, we know some intentional choices lead to crashes or increase their severity, yet we still include them in our analyses. Logically, many contributing factors could be called intentional even if the resulting harm was not intended. A drunk driver intended to get behind the wheel while impaired. They may not have consciously decided to get into a crash and cause injury or damage, but they chose a dangerous behavior that vastly increased their risk of causing a crash. The same is true of a distracted driver, or one who drives aggressively. People decide not to use protective devices (seat belts, child safety seats, motorcycle helmets, etc.) vastly increasing risk of serious injury or death if they are involved in a crash. It is difficult to call the consequences of such deliberate choices accidental.

Accident is also a strategic word choice used by defense lawyers hoping to help their clients avoid criminal convictions or civil liability. It is not a word choice most of us would consciously agree with when we know that the involved driver actively chose to behave in an unsafe manner. The real debate over crash versus accident can safely exclude a portion of the events—those that have a measure of intent. The real question is how much intent and what intentions are required in order to avoid use of the word accident. Indeed, safety advocates argue that there really are no accidents by this criterion. People may not have set out to crash, or to injure themselves or others, or cause damage, but their choices that raised the risk of those outcomes were deliberate. Was that unplanned or simply unthinking?

Why Bother?

The grassroots advocacy organizations WeSaveLives, Transportation Alternatives, and Families for Safe Streets make the case most eloquently based on the personal experiences of their founders and members. Their loved ones did not die accidentally—they died in a car crash because of multiple reasons, predominantly human errors as well as (more rarely) unsafe conditions on the roadway or problems with the vehicle. That is truly the situation for nearly all the fatalities and injuries on our roadways—a person chose to act in a risky manner and someone was hurt or killed. If that applies to more than 90% of these incidents, calling them accidents is a disservice to the victims and their families. This point of view is simultaneously emotional and factual—a powerful combination. These safety advocacy groups and their traffic safety allies are working with a purpose. They want the laws changed to recognize the role of human choices and other recordable factors. They want the media to use a word that does not support deliberate miscasting of the events or promote fatalist attitudes about the deaths and injuries on our roadways. Crash does that in a way that accident does not.

Some still argue that accident can carry this load. Clearly, for most of the world, the equivalent word for accident works just fine. They say we can study the causes of accidents and try to avoid them in the future. The problem is that, for some, accident has two popular definitions and one of those absolves people or entities of blame. That's why the auto industry began promoting its use more than 100 years ago. The word mattered then, and it played a role in conditioning us to accept the fault-free inevitability of what are truly preventable events. While we are more savvy now and know that harmful incidents are signs of a solvable problem, there is still the lingering idea that causal attribution is too elusive or does not matter so long as we can mitigate the problem.

Safety advocates, though, have a wealth of experience to draw on. We know that grassroots efforts can drive change. We have seen it in drunk driving and seat-belt use, to cite two well-studied examples. We also know that laws can change through public pressure. Already some states (e.g., Arizona) are changing their traffic laws to eliminate the word accident. This is why the safety advocates bother asking for a word change; by making that change, they know we are more likely to affect a change in attitude that drives behavior. This is the safety culture approach—a shared set of values and beliefs that include, in this case, that crashes are almost always preventable. We have an opportunity to use the power of social norming, as we did for drunk driving and seat-belt use, to signal through our words that we do not accept the antiquated view that injuries and deaths on our roads are accidental. Truly, they are not.

Acknowledgments

I would like to thank my colleagues Pamela Shadel Fischer, Kathleen Haney, Tim Kerns, Jeff Larason, and John McDonough for helping me by sharing their time, resources, and expertise. I also thank Per Garder for insightful comments and assistance.

References

- ATSIP, 2017. Manual on the Classification of Motor Vehicle Traffic Crashes, 8th ed. American National Standards Organization. Available from: http://www.atsip.org/ANSI_Ver_2017_D16.pdf.
- Davis, R.M., Pless, B., 2001. BMJ bans "accidents". *BMJ* 322, 1320.
- Fausey, C.M., Boroditsky, L., 2010. Subtle linguistic cues influence perceived blame and financial responsibility. *Psychon. Bull. Rev.* 17 (5), 644–650.
- Smith, M., 2018. A history of traffic safety in the United States: part one. *Citysmiths*. Available from: <https://www.citysmiths.org/blog/2018/7/16/a-history-of-traffic-safety-in-the-united-states-part-one>.
- TIRF & DIAD, 2017. Let's talk about crashes. Traffic Injury Research Foundation and Drop It and Drive. Available from: <http://tirf.ca/wp-content/uploads/2017/12/Lets-Talk-About-Crashes-9.pdf>.

Further Reading

- Please note that much of the discussion on crash versus accident terminology has played out on the internet and the media. Many of the sources listed later are statements of opinion and should not be taken as indicating a consensus one way or the other. Technical documents, such as the ANSI standard, are useful in providing modern definitions of terms of art, but they do not fully engage the controversy surrounding the word choice.
- AAA Foundation for Traffic Safety, 2018. 2017 traffic safety culture index. Available from: <https://aaaafoundation.org/2017-traffic-safety-culture-index/>.
- Badger, E., 2015. When a car 'crash' isn't an 'accident'—and why the difference matters. *The Washington Post*. Available from: https://www.washingtonpost.com/news/wonk/wp/2015/08/24/when-a-car-crash-isnt-an-accident-and-why-the-difference-matters/?noredirect=on&utm_term=.9121ee094c85.
- Benderly, B.L., 2016. Disregarding a risk does not equal an accident. *Science Magazine: Careers*. Available from: <https://www.sciencemag.org/careers/2016/07/disregarding-risk-does-not-equal-accident>.
- #crashnotaccident hashtag—a search in Twitter on 2/19/2019. Available from: <https://twitter.com/search?q=%23crashnotaccident&src=tyah>.
- Crash not accident pledge. Available from: <https://www.crashnotaccident.com/>.
- Get #CrashNotAccident in AP Stylebook, TransAlt, FSS say. *Transportation Alternatives News*. Available from: <http://transalt.org/news/releases/8281>.
- Larason, J., 2016. Drop the 'A' word (library). Available from: https://droptheaword.blogspot.com/2016/07/the-drop-a-word-library-articles-on_12.html.
- Lightner et al., 2014. Letter to the Minneapolis Star-Tribune. Available from: <https://wesavelives.org/letter-to-the-minneapolis-star-tribune/>.
- Richtel, M., 2016. It's no accident: advocates want to speak of car 'crashes' instead. *New York Times*. Available from: <https://www.nytimes.com/2016/05/23/science/its-no-accident-advocates-want-to-speak-of-car-crashes-instead.html>.
- Stromberg, J., 2015. We don't say "plane accident." We shouldn't say "car accident" either. Available from: <https://www.vox.com/2015/7/20/8995151/crash-not-accident>.
- Varagur, K., 2016. We've been brainwashed into saying 'car accident'. *Huffington Post*. Available from: https://www.huffingtonpost.com/entry/car-accident-car-crash-driving-behavior_us_573e496de4b00e09e89e9a7a.
- Yagoda, B., 2017. Crash or accident? *The Chronical of Higher Education: Lingua Franca Blog*. Available from: <https://www.chronicle.com/blogs/linguafranca/2017/09/04/crash-or-accident/>.
- Zender, J.F., 2017. Accident vs. crash: the issue of responsibility. *Psychology Today: The New Normal Blog*. Available from: <https://www.psychologytoday.com/us/blog/the-new-normal/201702/accident-vs-crash>.

Age and Gender as Factors in Road Safety

Marion Sinclair, Department of Civil Engineering, Stellenbosch University, Stellenbosch, South Africa

© 2021 Elsevier Ltd. All rights reserved.

Introduction	11
Gender	11
Risk-Taking	13
Compliance With Traffic Legislation	13
Age	13
Children	14
Youth	14
Middle-Aged Road Users	15
Elderly Road Users	15
Implications for Road Safety Going Forward	15
See Also	16
References	16
Further Reading	16

Introduction

The influences of gender and age on traffic crash risk are well documented in road safety research. In spite of cultural and geographic differences between different regions of the world, the degree to which males are overrepresented in fatal crashes is uncannily similar across many different countries (around 75% traffic fatalities are male in most of the countries). Such trends are often attributed in part to the different driving behaviors of males and females, though exposure is a factor as well. Internationally, males are typically more mobile than females—as drivers, passengers, pedestrians, and cyclists—hence, more commonly exposed to traffic crash potentials. Wherever possible, however, analysis has normalized crash involvement of the genders—either by comparing rates within certain populations (e.g., within the populations of licensed drivers) or by normalizing for relative mobility by looking at miles or kilometers traveled. In all cases, the conclusion remains strong that males are more at risk of fatal crash injuries than are females.

Age as a unit of analysis is similarly noteworthy, as young people are more likely to die in road crashes across the world than middle-aged road users, while the elderly are physically frailer and so are more susceptible to death in the event of a crash than younger people. Age cannot be entirely separated from gender as age and gender combinations produce distinct patterns of crash likelihood. Young men, for example, are at heightened risk when compared with young women, but in contrast elderly females are overrepresented in injury statistics when compared with elderly males. The research confirms that gender and age are strong influences on crash risk, only marginally influenced by social norms or geography.

Gender

A significant trend in male overrepresentation in road crashes is evident from various traffic data collected from across the world. For example, in the United States, data from the National Highway Traffic Safety Administration (NHTSA) for 2015 confirmed that the crash involvement rate per 100,000 licensed drivers was 1.605 for males, and 1.274 for female drivers. For all reported crash types, licensed male drivers were 26% more likely to be involved in a crash than licensed female drivers. Even more critically, male deaths represented 74.4% of all traffic deaths in 2015 (35,297 male deaths compared with 12,172 female deaths in 2015), yet the miles driven by male drivers constituted only 62% of total miles driven. Male drivers are clearly overrepresented in fatal crashes when their miles traveled are considered.

In the United Kingdom, these percentages are similar: the number of males reported killed in traffic crashes for 2017 was 1345 versus 474 female deaths, with 74% male deaths as compared to 26% female deaths. Again, male miles driven in the United Kingdom are not comparable with their death rates—only 64% of miles driven in the United Kingdom are driven by males. Male deaths are once again higher than would be expected based on exposure levels alone.

Exposure is a combination of the time people spends on the road, mileage covered, and traffic conditions they are exposed to (Elander et al., 1993). The higher the exposure to traffic crash potential, the higher the individual's likelihood of being involved in a crash. All things being equal, we would assume that all road users with identical exposure rates would be equally represented in

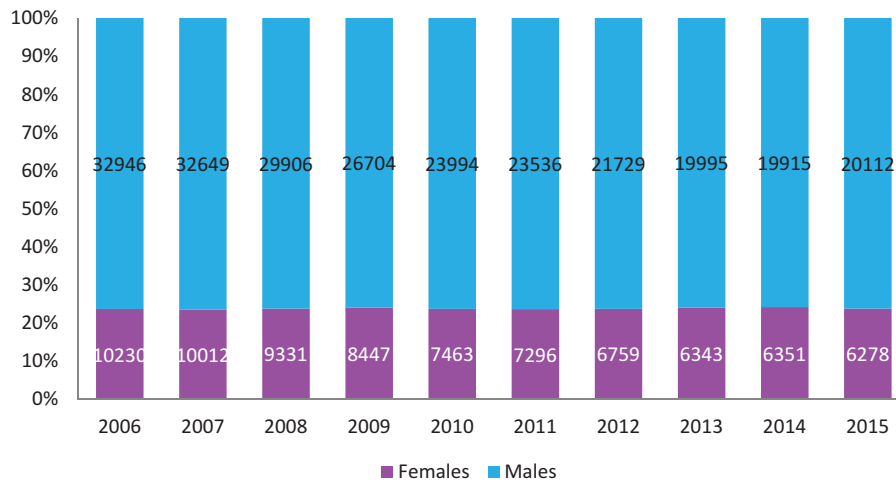


Figure 1 Annual distribution of fatalities by gender in the EU, 2006–15. Source: European Road Safety Observatory (2019).

crashes, but in the gender dimension, this is not the case, as is evident in the United States and United Kingdom examples. They are not unique.

In Europe, 76.2% of the EU region's fatal crashes involved males in 2015. Again, this exceeded the difference in miles driven by male drivers. Research in Spain, for example, indicates that the average number of kilometers traveled per day by males is only 22% higher than that of females (Ayuso et al., 2016).

Gender percentages of traffic fatalities from 2006 to 2015 show how strikingly consistent the ratio of male to female deaths has been over that decade (Fig. 1).

These patterns are similar for most regions of the world (this analysis excludes countries, such as Saudi Arabia and Iran where drivers are almost exclusively male).

Much of the analysis worldwide focuses on males as drivers, that is, when they are in control of driving decisions. However, as pedestrians and passengers, males, again are more at risk of being killed on the road than females, indicating that for every mode of transport, males are more likely to die in traffic crashes than females. In the United Kingdom, for example, female pedestrians are responsible for slightly over half pedestrian trips taken each year (52%) but 57% of pedestrians killed are males. Similarly, as shown in Fig. 2, across Europe male pedestrian deaths, and male passenger deaths, outnumber those of females.

Fatal crashes typically are the most extreme of crash types; that is, where injuries are severe enough to cause death. In slight and non-injury crashes, this pattern does not always hold true, which has interesting implications for behavioral research. For example, a Spanish study in 2017 found that male pedestrians were more likely to be involved in slight injury crashes only in the age groups of 4–13 years, and 25–44 years. In fact, from the age of 54 years and higher, females were more likely than males to be slightly injured as pedestrians in crashes.

The academic response to male overrepresentation in crashes and fatal crashes in particular has been to develop research into road user behavior and gender attributes. Behaviors exhibited by both genders, such as risk-taking, aggression, and rule violation are commonly analyzed under gender research studies.

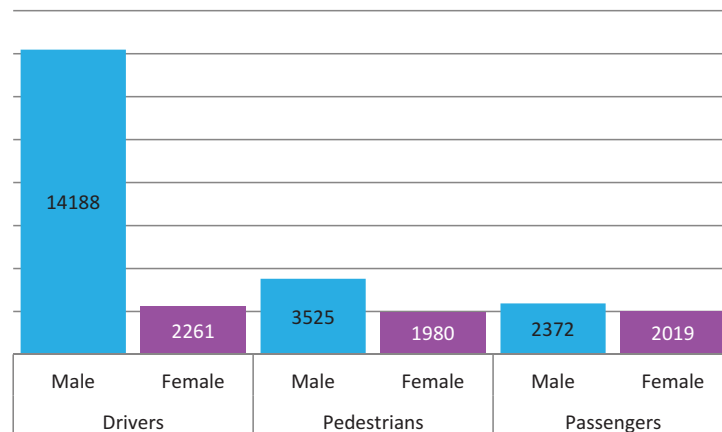


Figure 2 Road fatalities by gender and road user—Europe 2015. Source: European Road Safety Observatory, 2017. Annual Accident Report 2017. Available from: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/statistics/dacota/asr2017.pdf.

Risk-Taking

Research has showed very clearly that males have a far higher tolerance for risk in driving and in pedestrian activity, than females (DeJoy, 1992; Clarke et al., 2010). This means that they show greater inclination to drive at higher speeds (Stradling et al., 2003) and also under more difficult driving conditions, such as poor weather, or at night (Bener and Crundall, 2008), than females. In studies of pedestrian crossing behavior, males have been found to cross with smaller gap acceptance levels than females (Ishaque and Noland, 2008). Many of these are unconscious decisions. When appetite for risk is an embedded factor in an individual's personality, their propensity for intentional risky driving behavior increases. Risky behavior includes speeding for the thrill of it; following the vehicle ahead too closely, violating traffic rules, not using seatbelts, using mobile phones while driving; text messaging driving during high-risk night-time hours, and driving older vehicles. Of course risky personality traits can be found in females as well, but these are significantly less common, with anthropologists arguing that risk-taking behavior in males has been a necessary element of human evolution, with risk-taking among males typically factors in exploration and competitiveness among our early human ancestors.

Compliance With Traffic Legislation

Both age and gender have been found to be contributory factors to the likelihood of traffic violation. The likelihood of traffic violations is reported to be highest for males, young drivers and with increasing levels of mobility (González-Iglesias et al., 2012; Oppenheim et al., 2016).

Age

Until now gender has been discussed as if it affects crash risk consistently, irrespective of age. This is not, in fact, the case. Gender and age are interlinked elements in crash causation—young road users demonstrate significantly more distinct gender imbalances in crash involvement than do middle aged or elderly road users. Three European countries, for example, the Netherlands, Sweden, and the United Kingdom, reported that young males (under 25 years) have a fatal crash rate per million kilometers driven that is 3 times higher than that of females of the same age.

Common patterns of crash involvement can be identified in even a cursory examination of the age of traffic victims. In the United States, for example, crash rates per 100 million miles driven show clear differences in crash risk for all crashes, injury crashes, and fatal crashes (Table 1). Gender differences in crash risk appear to be most exaggerated among young road users—a phenomenon seen worldwide. This suggests that age is a compounding factor in road behavior and hence in crash risk.

Following the same logic as with gender, exposure, and road user behavior are the primary contributory factors in age-related crashes, with overrepresentation of certain ages again evident.

Different countries and regions divide age into comparative cohorts in different ways. Wherever possible ages are recorded in 5-yearly increments and this gives the most useful picture of crash involvement for each stage. However, such reporting can be onerous and groups with similar crash tendencies are often grouped together. In this categorization, and at the most basic level, there are four broad categories: children (typically defined as being under the age of 18 years though in countries where driving is permitted from 16 years of age this group tends to be < 16 years); young people/youth (18–24); middle aged (25–49 years), and elderly (50–64 years and above). Of course these groupings are themselves crude and distinctions can be found within them. For example, under the category of young people, novice drivers tend to be identified as the youngest among the group—those who are starting to drive, that is, typically between 16 and 21 years of old, whose risks as drivers are particularly notable. Among elderly drivers too, behavior and vulnerability change dramatically with age, with the older road users both more prone to error and more likely to die in the event of a crash than elderly drivers at the younger end of the range.

Table 1 Crashes and crash severity by age, United States rates per 100 million

Age	All crashes	Injury crashes	Fatal crashes
16–17	1432	361	3.75
18–19	730	197	2.47
20–24	572	157	2.15
25–29	526	150	1.99
30–39	328	92	1.2
40–49	314	90	1.12
50–59	315	88	1.25
60–69	241	67	1.04
70–79	301	86	1.79
80+	432	131	3.85

Source: National Highway Traffic Safety Administration (2016). Available from: <https://cdan.nhtsa.gov/tsftables/tsfar.htm>.

Children

The 2018 WHO annual status report on road safety confirms the significance of traffic injuries to young people when it reports that traffic injuries remain the leading cause of death for children and young adults aged 5–29 years.

By virtue of their small size and relative weakness, children are physically more fragile than adults and an impact that may simply injure an adult has far more potential to cause serious injury or death in children. They are, as a result of their anatomy alone, more vulnerable to serious injury than adults.

They are also more vulnerable to making errors as road users than adults, and largely dependent on adults to keep them safe. The child road user (up to the age of around 12) has physiological limitations and is exposed to higher risk within the traffic environment when compared to adults. While even young children have fairly well developed peripheral vision, the complex perceptions required to comprehend traffic risk emerge over time. In particular, children's ability to judge speed and distance of vehicles is dependent on their acquiring sufficient experience to build up a useful set of comparisons. They do not yet have the experience to be able to accurately determine which direction a sound may be coming from. To be safe on the roads, children must have advanced visual search skills, which allow them to determine very quickly what elements of the road system are dangerous and which are not. Research suggests that skills of this nature really come together in children from the age of 7 years—until then, children are unable to process all the information needed to be taken care of while crossing a road or walking along a road. Also, because they do not drive, they cannot fully anticipate what a driver's intentions might be at a junction or crossing point, and in fact this is a problem facing adults who may not be able to drive as well.

As passengers in vehicles, children are entirely dependent on adults to both navigate their vehicles safely, and to secure them in age appropriate child restraints which will give them the best possible protection in the event of a crash. Young children have specific physiological limitations that render them particularly vulnerable to serious injury in a collision. The skulls of infants are extremely malleable until around 24 months. Even low levels of force on the skulls of children of this age can result in significant deformation and brain injury. The infant/child ribcage is similarly flexible and the abdominal organs are poorly protected. Not surprisingly, given the physical limitations of children, child restraint systems have proven to be significantly better at reducing fatality risk to children than conventional seatbelts (Elliott et al., 2006).

Although airbags reduce fatal crash injuries among adult drivers and passengers, this safety technology can increase mortality risk among children, and it is increasingly uncommon for infant or child car seats to be used in the front passenger seat of a car. In countries where it is still considered acceptable for child car seats to be placed on the front passenger seat, it is always recommended that car seats be placed in a rear-facing position to prevent injury should an airbag be inflated.

Youth

As young people emerge from childhood, their exposure to potential crashes generally increases as their independent mobility grows. At this age, the young adult may have fully formed physiologies, with eyesight, hearing, and reaction times as good as they will ever be, but they are hindered by inexperience in traffic and also the personality and gender-based propensities for risk that emerge in adolescence—sensation seeking, peer pressure, and bravado are common features among young people that make them more vulnerable to error. Young (novice) drivers contribute to a sizable proportion of fatal road crashes (Oxley et al., 2014). Gender plays a key role again in this cohort, with young male drivers demonstrating a notably higher crash risk—and fatal crash risk—than female novice drivers (Whissell and Bigelow, 2003; Ivers, 2009).

Crashes involving young drivers are more likely to be single vehicle crashes, and to be speed related, than among any other age category. Alcohol use is also highest among crashes by young drivers (and lowest among elderly drivers).

The vast majority of novice drivers are still developing emotionally, psychologically, physiologically, and socially. A large body of research explains how these developmental features of young people contribute to elevated crash risk. These include personality-based factors, such as aggression, exhibitionism, recklessness, and a lower respect for authoritarian rules and regulations (hence a heightened likelihood to break traffic rules) than more mature drivers. Research has repeatedly found that young male drivers are more prone to engage in risky driving behavior, such as speeding and driving after consuming alcohol. This finding is consistent across all countries and all cultures. In addition, levels of deviant behavior or rule breaking are also much higher among younger males than among any other group of road users.

The presence of passengers has been shown to have a direct bearing on the likelihood of teenage drivers engaging in explicitly risky behaviors. In the United States, young teenage drivers are 2.5 times more likely to engage in one or more potentially risky behaviors while driving if a teenage peer were present, compared to when driving alone. The same research confirms that when driving with more than one peer, the likelihood of engaging in risky driving behaviors, increases by up to 300%.

Cognitive research studies confirm that new drivers do not yet have fully formed skills in hazard recognition. Drivers move through a complex road environment and novice drivers must learn to scan, detect, and react appropriately to hazardous road environments. As with young children gradually learning through experience as to how to safely navigate the road traffic environments, novice drivers acquire the cognitive skills necessary for safe driving over time and with growing levels of experience. Studies of first crash timing of young drivers show that there is a sharp decline in the crash risk per kilometer driven during the first few months after licensing which indicates that the months after the licensure are most critical in terms of completing their skill development (Mayhew et al., 2003). It is during this time, too, that the motor skills associated with maneuvering the vehicle are also fine-tuned—novice drivers get more technically competent over these critical months.

Middle-Aged Road Users

Almost no exceptional risk factors exist for this group. The earlier desires for sensation-seeking, peer pressure, and exhibitionism are generally under control among middle-aged road users, and the responsibilities that come with having careers and raising families tend to curb the natural risk-seeking behavior even among males. Within this cohort, males are still slightly more likely to be killed in a crash than females, but the risks are far more balanced.

Elderly Road Users

Since 1990, the global life expectancy of humans has increased by approximately 6 years. It is therefore safe to assume that the proportion of elderly people as road users also increases over time. In fact, a significant increase in the proportion of road traffic deaths in the age group older than 70 years has already become noticeable in high-income countries. The difference is most likely related to longevity in these countries, combined with the greater risk posed by reduced mobility and increased frailty.

Despite the vitality of many older persons, gradual changes in the ability to perform complex tasks, such as driving and avoiding accidents becoming evident with increasing age. Biologically, the body loses the ability to renew itself, the bodily functions weaken and vital organs become less robust. There is a decline in sensory process, perception, motor skills, and problem solving skills. Crashes involving older drivers often occur in more demanding driving environments, such as at intersections, or where important information from the road can be misread or misinterpreted (e.g., judging other vehicles speeds, or stopping distances). For the elderly, their heightened risk of being involved in a crash comes from the combined failure of initial judgment (as a result of sensory loss with age), and the failure to accommodate or modify behavior to avoid a developing incident.

The elderly are physically frailer than younger adults; their injuries will be more severe given an identical collision impact involving a younger person. The fatality rate is approximately 3 times higher for a 75-year-old motor vehicle occupant than for an 18-year-old. They are more likely to sustain fractures to all body parts in a crash, recovery time is much longer and the likelihood of long-term disability is high.

While elderly females are involved in more crashes than males (of all severity), this is largely a function of the fact that there are simply more elderly females than males in traffic. Almost all countries experience higher survivability rates of women than males among the elderly. In the United States, for example, in 1998 the number of women per 100 men was 123 at ages 65–74; 151 at ages 75–84, and 241 at ages 85 and older. More elderly women road users are involved in crashes simply there are more of them overall. However, when only drivers are considered, males remain overrepresented in crashes when compared with females (normalized against numbers of drivers of the same gender). This is true for all injury crashes as well as for fatal crashes.

The gross number of crashes involving elderly drivers is lower than might be expected. Reasons for this are generally attributed to the fact that the elderly tend to actively manage their driving risk—driving at times when there is less traffic on the roads, avoiding high volume intersections, and high-speed roads, for example, or avoiding nighttime and long-distance driving. At some point in their lives, elderly drivers choose to give up driving altogether and become more dependent on other drivers, public transport, and walking. They thus regulate their exposure to risk more actively than younger drivers.

The occurrence of pedestrian fatalities is the highest for ages older than 55 years and younger than 12 years in all European countries. A report the European Union showed that pedestrians aged 65 years and older, though representing only 15% of the total population, accounted for 45% of all pedestrian fatalities. Similar results were found in a study from Sri Lanka where 43% of pedestrian fatalities occurred in the age group of 65 years or older. The Monash University Accident Research Centre has found that the number of cycling fatalities in this age group can increase up to 70%.

Implications for Road Safety Going Forward

Risky behavior associated with males, and young people, constitute some of the biggest challenges to improving road safety across the world. Changing road user's chosen behavior relies on effective education campaigns, but traffic law enforcement could arguably be the better option for behavior that is intentionally willful and risky. Improving young driver's skills at hazard perception, and children's ability to recognize the location of hazards and to prioritize them, can be improved with age-appropriate training—much work is being done on hazard recognition as part of driver training, for example. Safety challenges associated with young drivers can be addressed in part through better training and graduated licensing programs, which are being utilized to good effect in some parts of the world.

The fact that most drivers of fleet vehicles, heavy goods vehicles, and public transport vehicles are male, means that the drivers of these vast fleets of vehicles are more at risk simply as a consequence of their gender. Gender sensitivities mean that few are likely to suggest that males should be prohibited from these professions, but there is also a tendency to avoid suggesting that special remedial programs for professional male drivers have a role to play. Given that scientific evidence clearly shows heightened crash risk among male road users, this sensitivity is undoubtedly misplaced, and going forward it seems essential for road safety initiatives to derive and prioritize gender—and age-specific programs for behavioral adaptation.

See Also

Pedestrian Safety: Children; Pedestrian Safety: Elderly; Pedestrian Safety: General Pedestrian Safety: Visually Impaired

References

- Ayuso, M., Guillen, M., Pérez-Marín, A., 2016. Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks* 4, 10. doi:10.3390/risks4020010.
- Bener, A., Crundall, D., 2008. Role of gender and driver behaviour in road traffic crashes. *Int. J. Crashworthiness* 13 (3), 331–336, doi:10.1080/13588260801942684.
- Clarke, D.D., et al., 2010. Killer crashes: fatal road traffic accidents in the UK. *Acc. Anal. Prev.* 42, 764–770, doi:10.1016/j.aap.2009.11.008.
- DeJoy, D.M., 1992. An examination of gender differences in traffic accident risk perception. *Acc. Anal. Prev.* 24, 237–246, doi:10.1016/0001-4575(92)90003-2.
- Elander, J., West, R., French, D., 1993. Behavioral correlates of individual differences in road-traffic crash risk: an examination of methods and findings. *Psychol. Bull.* 113 (2), 279–294, doi:10.1037/0033-2909.113.2.279.
- Elliott, M.R., et al., 2006. Effectiveness of child safety seats vs seat belts in reducing risk for death in children in passenger vehicle crashes. *Arch. Pediatr. Adolesc. Med.* 160 (6), 617–621, doi:10.1001/archpedi.160.6.617.
- González-Iglesias, B., Gómez-Fraguela, J.A., Luengo-Martín, M.Á., 2012. Driving anger and traffic violations: gender differences. *Transp. Res. Part F Traffic Psychol. Behav.* 15, 404–412, doi:10.1016/j.trf.2012.03.002.
- Ishaque, M.M., Noland, R.B., 2008. Behavioural issues in pedestrian speed choice and street crossing behaviour: a review. *Transp. Rev.* 28 (1), 61–85, doi:10.1080/01441640701365239.
- Ivers, R., et al., 2009. Novice drivers' risky driving behavior, risk perception, and crash risk: findings from the DRIVE study. *Am. J. Public Health.* 99 (9), 1638–1644, doi:10.2105/AJPH.2008.150367.
- Mayhew, D.R., Simpson, H.M., Pak, A., 2003. Changes in collision rates among novice drivers during the first months of driving. *Accid. Anal. Prev.* 35 (5), 683–691, doi:10.1016/S0001-4575(02)00047-7.
- Oppenheim, I., et al., 2016. Can traffic violations be traced to gender-role, sensation seeking, demographics and driving exposure? *Transp. Res. Part F Traffic Psychol. Behav.* 43, 387–395, doi:10.1016/j.trf.2016.06.027.
- Oxley, J. et al., 2014. Understanding novice driver behaviour: review of literature, 62p. Available from: <http://www.monash.edu/miri/research/reports/%5Cnhttps://trid.trb.org/view/1412546>.
- Stradling, S.G. et al., 2003. *The Speeding Driver: Who, How and Why?* Scottish Executive Central Research Unit, Edinburgh.
- Whissell, R.W., Bigelow, B.J., 2003. The speeding attitude scale and the role of sensation seeking in profiling young drivers at risk. *Risk Anal.* 23 (4), 811–820, doi:10.1111/1539-6924.00358.

Further Reading

- Byrnes, J.P., Miller, D.C., Schafer, W.D., 1999. Gender differences in risk taking: a meta-analysis. *Psychol. Bull.* 125 (3), 367–383.
- Constantinou, E., et al., 2011. Risky and aggressive driving in young adults: personality matters. *Accid. Anal. Prev.* 43 (4), 1323–1331.
- Dula, C.S., Geller, E.S., 2003. Risky, aggressive, or emotional driving: addressing the need for consistent communication in research. *J. Safety Res.* 34 (5), 559–566.
- European Road Safety Observatory, 2017. *Annual Accident Report 2017*. Available from: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/statistics/dacota/asr2017.pdf.
- Roman, G.D., et al., 2015. Novice drivers' individual trajectories of driver behavior over the first three years of driving. *Accid. Anal. Prev.* 82, 61–69.
- Scott-Parker, B., et al., 2012. They're lunatics on the road: exploring the normative influences of parents, friends, and police on young novices' risky driving decisions. *Safety Sci.* 50 (9), 1917–1928.

Aggressive Driving and Road Rage

James E.W. Roseborough*, **Christine M. Wickens^{†,*}**, **David L. Wiesenthal[§]**, ^{*}OCAD University, Toronto, ON, Canada; [†]Institute for Mental Health Policy Research, Centre for Addiction and Mental Health, Toronto, ON, Canada; [§]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada; [§]York University, Toronto, ON, Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction	17
Defining Aggressive Driving	17
Theoretical Perspectives of Aggressive Driving	17
Person-Related Contributors to Aggressive Driving	18
Demographic Factors	18
Personality Factors	18
Cognitive Factors	18
Attitudes and Social Norms	19
Mental Health Factors	19
Emotional Factors	20
Situation-Related Contributors to Aggressive Driving	20
Environmental Factors	20
Aggressive Driving Countermeasures	20
Psychological Countermeasures	20
Environmental Countermeasures	21
Enforcement Countermeasures	21
Technological Countermeasures	21
Incentive Countermeasures	21
Multimedia-based Countermeasures	22
Future Considerations	22
Efficacy of Countermeasure Research	22
Consolidation of Research	22
Conclusion	22
References	22
Further Reading	24

Introduction

Aggressive driving includes any behavior by a motorist intended to physically, emotionally, or psychologically harm another road user in the roadway environment ([Hennessy and Wiesenthal, 2002](#)). Such behavior occurs worldwide and increases the likelihood and severity of motor vehicle collisions ([Galovski et al., 2006](#); [Paleti et al., 2010](#); [Sârbescu et al., 2014](#); [Stephens and Fitzharris, 2017](#); [Stephens and Sullman, 2015](#); [Sullman et al., 2017](#)). Considerable research has been devoted to classifying types of aggressive driving behavior and identifying factors that increase and decrease the risk of its occurrence.

Defining Aggressive Driving

The driver behavior literature makes mention of multiple related terms, including aggressive driving, risky driving, road rage, retaliatory aggression, hostile aggression, dangerous driving, and instrumental aggression, among others ([Haje and Symbaluk, 2014](#); [Suhr and Dula, 2017](#); [Suhr and Nesbit, 2013](#); [Zhang et al., 2017](#)). These terms are not synonymous, necessitating careful consideration for appropriate term usage. For example, the act of speeding in order to tailgate and intimidate another motorist constitutes both risky and aggressive driving. Speeding to get to a destination sooner is risky driving, but not aggressive driving. All aggressive driving can be considered risky driving, but not all risky driving can be considered aggressive. Similarly, the term road rage is often used to describe severe behavior such as deliberately damaging another vehicle or engaging in a physical altercation with another road user. All acts of road rage can be considered aggressive driving, but not all acts of aggressive driving can be considered road rage.

Theoretical Perspectives of Aggressive Driving

Over the years, researchers have proposed several psychological theories to attempt to explain aggressive driving. The theory of planned behavior posits that attitudes, subjective norms, and perceived behavioral control interact with each other and contribute to

a person's intentions, which in turn contribute to a person's behavior (Ajzen, 1991). The frustration-aggression hypothesis posits that obstructing achievement of one's goals leads to frustration, which increases risk of subsequent aggression (Dollard et al., 1939). Both theories involve the interaction of situational events, cognitions, and emotions, influencing aggressive behavior (Efrat and Shoham, 2013; Shinar, 1998). The General Aggression Model incorporates and builds on both the theory of planned behavior and the frustration-aggression hypothesis (Anderson and Bushman, 2002). The General Aggression Model proposes that person-related factors (e.g., gender) and situation- or environment-related factors (e.g., rude gesture or heat) influence a person's emotions, cognitions, and physiological arousal, which in turn influence the decision-making process, potentially resulting in aggressive behavior. Furthermore, prior experienced events and reactions can cycle back and influence how subsequent events are experienced (Anderson and Bushman, 2002). The General Aggression Model provides a comprehensive theoretical framework outlining the emergence of aggressive driving behavior. The contributing factors subsequently discussed in this article can be placed within the framework of the General Aggression Model.

Person-Related Contributors to Aggressive Driving

Factors identified as contributors to aggressive driving generally fall into two categories: person related and situation related. Person-related factors, which represent the majority of contributing factors identified to date, are specific to the driver.

Demographic Factors

Demographic factors (e.g., age, gender, education level) are the most stable characteristics of a person; they do not change from situation to situation. Compared to females, males are more likely to engage in aggressive driving, although this difference is most consistent and most pronounced when more severe forms of aggressive driving (e.g., road rage) are examined (Hennessy et al., 2004). Younger, compared to older individuals are more likely to drive aggressively (Wickens, Mann et al., 2011). Accordingly, young male drivers tend to engage in more aggressive driving than other demographic groups (Arnett, 2002; González-Iglesias et al., 2012; Roberts and Indemaur, 2005; Wickens, Mann et al., 2012).

Personality Factors

Personality traits are a combination of a person's thoughts, feelings, and behavior. Traits influence how we respond cognitively, emotionally and behaviorally to everyday events and situations. As they are believed to be relatively stable over time, traits have been the focus of much of the aggressive driving literature. Not surprisingly, individuals with higher levels of trait aggression, the tendency to exhibit verbal or physical aggression across situations, are more inclined to engage in aggressive driving (Deffenbacher et al., 2003). Similarly, trait anger is the tendency to frequently experience anger, with varying intensity, across situations. Individuals who possess high trait anger are more prone to drive aggressively (Bogdan et al., 2016; Herrero-Fernández, 2013; Stephens and Groeger, 2009). Drivers who are more prone to experience stress behind the wheel due to time constraints, traffic congestion, or poor weather conditions are also more prone to drive aggressively (Hennessy and Wiesenthal, 1999). Narcissistic individuals are generally selfish, self-absorbed, entitled, and possess fragile self-esteem. It is believed that narcissists engage in more aggressive driving because it serves as a defense mechanism against threats toward their self-esteem or entitlement (Lustman et al., 2010). Drivers who possess greater impulsivity are more likely to behave with little or no consideration of the consequences. Impulsive drivers are more likely to retaliate against a perceived offense by another motorist, leading to the escalation of roadway aggression (Dahlen et al., 2005). Other personality traits that have been shown to increase aggressive driving include neuroticism, sensation seeking, ego defensiveness, extraversion, emotional instability, hypermasculinity, and Type-A personality (Bone and Mowen, 2006; Dahlen et al., 2005; Dahlen and White, 2006; Matthews et al., 1991; Miles and Johnson, 2003; Neighbors et al., 2002; Sümer et al., 2005). Conversely, other personality traits have been shown to reduce the risk of aggressive driving. Conscientious individuals tend to be careful and self-disciplined, which may result in driving behavior that falls within the confines of social norms and the law (Chraif et al., 2015; Jovanović et al., 2011). Agreeableness is characterized by qualities of kindness and consideration. Agreeable drivers may be more likely to be accepting of other driver's poor behavior, resulting in a reduced risk of retaliation (Chraif et al., 2015; Jovanović et al., 2011). Similarly, trait forgiveness, which is characterized by a tendency to consciously release feelings of resentment and retribution toward a wrong doer, influences aggressive driving. Drivers high in trait forgiveness are able to release their hostile and retaliatory feelings toward other motorists, reducing the likelihood of an aggressive behavioral response (Kováčsová et al., 2014; Moore and Dahlen, 2008).

Cognitive Factors

Cognition includes attributions, perceptions, and ruminations; essentially the mental process through which all stimuli from the world around us are processed. Cognitive factors influence how a person experiences, feels, and reacts to events and situations. In the driving environment, our cognition shapes the way we assess interactions with other motorists and plays a significant role in identifying and selecting our response to those roadway events.

Attributions are the means by which individuals explain the causes of behavior and events. Research has shown that the more a behavior is perceived to be intentional and caused by a specific person, the more that person is deemed to be responsible for his or her behavior (Mikula, 2003). In the driving environment, motorists are more likely to react aggressively if they perceive themselves to be victims of another motorists' negative behavior and that behavior is deemed to be intentional; the offending driver will be judged as more responsible for the negative outcome (Wickens, Wiesensthal et al., 2011). Consider a situation in which you are cut-off by another vehicle. If you believe the other motorist had control over the behavior (i.e., turning the steering wheel) and intended to cut-off your vehicle, you will attribute more responsibility to that motorist and be more likely to respond aggressively. If, however, you believe the other motorist had control over the behavior but intended only to avoid a collision with an animal on the road, you will attribute less responsibility and be less likely to respond aggressively.

Cognitive bias, an error of information processing, can also play a role in aggressive driving (Wickens et al., 2014). When interpreting a negative event caused by another motorist, we are more likely to overestimate internal factors (e.g., driver's age) than external factors (e.g., road conditions). This is known as the actor-observer bias (Jones and Nisbett, 1971). Overestimating internal factors will lead to increased attributions of responsibility (Britt and Garrity, 2006; Hennessy et al., 2005; Lennon et al., 2011). As mentioned earlier, attributing responsibility to an offending driver increases the likelihood of aggressive responses. Another cognitive bias, the illusion of control, leads motorists to overestimate their level of control over an outcome. Illusion of control is defined as the tendency to view chances for success as higher than the probability warrants (Langer, 1975). Drivers with an increased level of illusion of control will attribute chance outcomes (e.g., avoiding a collision when driving aggressively) to their own skill (Stephens and Ohtsuka, 2014). This overconfidence regarding a collision or negative consequence of reacting aggressively is what leads to increased aggressive behavior.

Rumination involves repetitive and uncontrollable thoughts about negative internal or external experiences. Increased rumination has been found to contribute to increased feelings of anger, physiological arousal, and aggressive driving (Suhr, 2016; Suhr and Dula, 2017; Suhr and Nesbit, 2013). Furthermore, increased rumination not only leads to increased aggression toward the provocateur, but it has also been shown to increase aggression toward third-party individuals (Bushman et al., 2005; Pedersen et al., 2011). For example, a motorist who ruminates about being dangerously cut-off will be more likely to react aggressively toward other motorists who were not involved in the original transgression.

Attitudes and Social Norms

Attitudes are evaluations of a target, which may be a person, place, thing, or event (Ajzen, 2005). Attitudes can be positive or negative. Positive or permissive attitudes toward aggressive driving increase an individual's likelihood of behaving aggressively while driving, whereas negative attitudes toward aggressive driving have the opposite effect. Negative attitudes toward bicyclists have also been associated with positive attitudes toward aggressive driving and engagement in aggressive driving targeting bicyclists (Fruhen and Flin, 2015). Attitudes can apply to very specific circumstances; a driver may have permissive attitudes toward aggressive driving that target other motorists but not bicyclists, or toward drivers of a particular demographic (e.g., young/old, male/female). Stereotypes held about the driving skill or driving style of certain motorists may inform these attitudes (Davies and Patel, 2005; Lawrence and Richardson, 2005).

Social norms are a broad range of behavior considered to be acceptable by society. An individual's perception of social norms or acceptable behavior can influence their subsequent behavior. Motorists who perceive aggressive behavior toward bicyclists to be socially acceptable are more likely to behave accordingly (Fruhen and Flin, 2015).

Mental Health Factors

Mental health affects our ability to cope with stress, interact with others, and make decisions, all of which are essential to safely navigating the roadway environment. Mental health issues, therefore, can impair roadway safety, including via increased risk of driver aggression. Intermittent explosive disorder, for example, is characterized by impulsive expressions of extreme anger that are out of proportion to the anger-provoking stimulus (American Psychiatric Association, 2013). Compared to nonaggressive drivers, aggressive drivers have been found to be more likely to meet the diagnostic criteria for intermittent explosive disorder (Galovski et al., 2006). Attention deficit hyperactivity disorder (ADHD) is also characterized by impulsivity, as well as inattention and hyperactivity. Individuals with ADHD are more likely to report driver anger and aggression compared to individuals without ADHD (Barkley and Cox, 2007; Richards et al., 2006). Conduct disorder is characterized by extreme externalizing behavior and a prolonged pattern of antisocial behavior that violates the rights of others or age-appropriate societal norms or rules (American Psychiatric Association, 2013). Research has found that symptoms of conduct disorder are predictive of aggressive driving behavior and collision risk (Wickens et al., 2015, 2019). Additional research has found evidence that personality disorders (e.g., antisocial personality disorder, borderline personality disorder) significantly increase the odds of perpetrated driver aggression (Galovski et al., 2002), and there is some evidence to suggest that other psychological disorders (e.g., major depressive disorder) may as well (Butters et al., 2006). Likewise, a history of traumatic brain injury (e.g., concussive injury) has been associated with an increased likelihood of threatening physical harm against other motorists or other vehicles and with increased risk of collision involvement (Ilie et al., 2015).

Mental health is also influenced by use of alcohol and recreational drugs such as cannabis, cocaine, and ecstasy/MDMA. Each of these substances has pharmacological effects, which may impact one's mood and inhibition and may increase one's risk of engaging

in driver aggression (Benavidez et al., 2013; Butters et al., 2006; Fierro et al., 2011; Richer and Bergeron, 2009; Wickens, Mann et al., 2011). It may also be that individuals who engage in driver aggression are characterized by demographic or personality traits that predispose them to risk of problematic alcohol or drug use (Benavidez et al., 2013; Mann et al., 2004). There is also evidence to suggest that users of more dangerous illicit substances (e.g., cocaine, ecstasy/MDMA, heroin) are at risk of perpetrating *serious* driver aggression or road rage (i.e., threatening or attempting to damage another vehicle or harm another driver) (Benavidez et al., 2013; Butters et al., 2006).

Emotional Factors

Emotion is a complex mental state resulting in physical and physiological changes that influence behavior. Anger is defined as an emotional state marked by subjective feelings varying in intensity from mild annoyance or irritation to intense fury and rage (Spielberger et al., 1985). Within the aggressive driving literature, anger has been researched more than any other factor, both as an independent contributor and as a mediator or moderator of other relevant factors. Anger may be particularly predictive of driver aggression because it interferes with higher-level cognitive processes that would otherwise inhibit aggression, allowing for perceived justification of retaliation (Anderson and Bushman, 2002).

Situation-Related Contributors to Aggressive Driving

Environmental Factors

The physical environment, both inside and outside the vehicle, can also affect the likelihood of driver aggression. High ambient temperatures can serve as an irritation (Kenrick and MacFarlane, 1986), whereas pleasant sounds or scents (e.g., peppermint) can soothe the driver (Raudenbush et al., 2009; Wiesenthal et al., 2000). The “weapons effect”, whereby the sight of a gun can increase aggressive behavior (Berkowitz and Lepage, 1967), has been demonstrated in the context of driving, with drivers engaging in more tailgating, speeding, horn honking, verbal aggression, and obscene gestures when a weapon was in their vehicle (Bushman et al., 2017; Hemenway et al., 2006; Miller et al., 2002). Similar findings have been demonstrated with less overtly aggressive stimuli, such as banners and billboards with aggressive text (Ellison-Potter et al., 2001). On the contrary, roadways with natural vegetation in view induce less stress than those surrounded by billboards and commercial buildings, and hence reduce the likelihood of aggressive driving (Parsons et al., 1998).

Driver aggression is typically elicited in response to a perceived offense by another driver (Wickens, Wiesenthal et al., 2011) and, depending on one’s disposition, may increase with repeated exposure to offensive roadway behavior (Chai and Zhao, 2016). Situational factors can further augment the risk of aggressive roadway retaliation. Traffic congestion can reduce frustration tolerance (Shinar, 1998), particularly when combined with a sense of time urgency (Wickens and Wiesenthal, 2005). Daily hassles and work-related stress can also impair one’s ability to tolerate a perceived injustice behind the wheel (Matthews et al., 1991; Wickens and Wiesenthal, 2005).

Visible features of the offending motorist can also influence attributions made for that motorist’s actions, affecting the likelihood of an aggressive response. Female offenders are perceived to be more careless and less aggressive than male offenders, suggesting that gender stereotypes operate in the roadway environment (Lawrence and Richardson, 2005), as has been discussed previously in this article. The relative status of vehicles involved in a roadway altercation can also increase or decrease the likelihood of expressed aggression (Doob and Gross, 1968). Tailgating in response to slow driving by a lead vehicle was greater when that lead vehicle was marked by a learner permit and less when that vehicle was an ambulance (Stephens and Groeger, 2014). Likewise, horn-honking latency in response to a middle-class vehicle stopped at a green light decreased with the relative status of the stopped vehicle, such that drivers of upper-class vehicles were quickest to honk and drivers of lower-class vehicles slowest to do so (Diekmann et al., 1996). Visibility of the “victim” motorist can also influence the likelihood of aggression. Greater anonymity associated with tinted windows, the dark of night, a hard-top vehicle (vs. a convertible), or a standard license plate (vs. a more memorable vanity plate), have been associated with more frequent roadway violations and driver aggression by an offended driver (Ellison et al., 1995; Ellison-Potter et al., 2001; Wiesenthal and Janovjak, 1992).

Aggressive Driving Countermeasures

The goal of identifying contributing factors is to develop means and methods to reduce aggressive driving. Comparatively less research, however, has focused on the effectiveness of potential countermeasures for aggressive driving.

Psychological Countermeasures

Programs to treat aggressive drivers are now being developed using cognitive-behavioral therapy, attributional retraining, and relaxation training (Galovski et al., 2006). These programs teach drivers to identify the triggers of their roadway anger and aggression, to recognize cognitive distortions that contribute to their anger, and to control their breathing and relax their muscles

when an anger-provoking event is encountered. These programs have been shown to successfully teach more effective coping strategies and alternative driving practices to help control aggressive outbursts (Galovski and Blanchard, 2004). Driver attributional retraining programs, for example, focus on making drivers aware of their cognitive biases when accounting for a perceived offense behind the wheel, encouraging drivers to focus on external factors (e.g., uncontrollable road conditions) as opposed to internal factors (e.g., intentional selfishness). One approach to attributional retraining that could prove very effective is to encourage drivers to focus on their own prior offenses; previous research has shown this focus to reduce anger in response to an offensive driving event (Takaku, 2006).

Environmental Countermeasures

Reducing aggressive behavior can also be aided by reducing the number of environmental stressors a motorist experiences such as unpleasant noises, odors, and temperature. Lavender and jasmine have been identified by aromatherapy research as calming smells, and may prove beneficial for reducing stress, irritability, and aggression while driving. The odor of peppermint has been found to reduce frustration and anxiety in a simulated driving task (Raudenbush et al., 2009). Listening to self-selected music has been found to reduce milder forms of aggressive driving such as rude hand gestures (Wiesenthal et al., 2003); however, a word of caution is necessary. While self-selected music may be beneficial, the selection of music should not distract the driver, the volume should not hinder a motorist's ability to hear other vehicles, and the tempo should be limited, as faster tempos have been associated with faster speeds and more roadway violations (Brodsky, 2002).

Enforcement Countermeasures

Several studies have identified the contribution of police enforcement to the reduction of driver aggression (Roseborough and Wiesenthal, 2014; Stanojević et al., 2018). Police presence may reduce aggressive driving for multiple reasons. Research has shown that upon witnessing an offensive driving behavior, victims experience increased happiness when a police vehicle stops the offending driver (Roseborough et al., 2018). Despite not being privy to the punishment the offending driver received (e.g., stern warning or citation), merely observing active enforcement in the driving environment seems to help reduce motorists' anger and aggression. Police presence may also stop offended drivers from retaliating aggressively by increasing the perceived likelihood of legal consequences for their own actions. Research has also contributed to enforcement efficacy by identifying more frequent or more offensively perceived aggressive driving behaviors, and when and where these acts are most commonly committed; this information is used to target limited enforcement resources more effectively (Stephens et al., 2016; Wickens, Roseborough et al., 2013; Wickens, Wiesenthal et al., 2013).

Technological Countermeasures

Technological countermeasures refer to both automobile and non-automobile technology. Recent advancements have led to the development of proximity sensors that notify a driver of an obstacle and automatically reduce the vehicle's speed when it closes in on another vehicle in order to maintain an appropriate distance; this technology could help to reduce instances of tailgating. Researchers have also suggested that automobile horn and front light control systems be modified to reduce repeated horn blowing and headlight flashing, introducing a required waiting period before the horn or lights are used again (Smart et al., 2005). While changes to vehicles may prove effective, designers and engineers must thoroughly understand the consequences of any changes; safe operation of the vehicle should never be compromised.

Despite their lack of popularity with drivers and the political controversy that results from that, both red light cameras and photo radar are widely used in Canada, Europe, and Australia and have been shown to reduce collision risk (Goldenbeld et al., 2019). In a similar vein, equipping motorways with an increased number of closed-circuit surveillance cameras designed to detect aggressive driving manoeuvres could help to reduce aggressive and risky driving behavior and assist authorities to apprehend aggressive drivers. The presence and efficacy of such cameras may have the same effect as actual law enforcement officers. Clearly identifying the presence of such cameras with signage would be essential, as would be public education about their efficacy.

Incentive Countermeasures

Some countermeasures focus on reinforcing or rewarding positive driver behavior. In 2010 speed cameras in Stockholm, Sweden were given a social twist. Drivers who obeyed the speed limit were automatically entered in a lottery with monetary prizes. During the lottery, over 24,800 vehicles passed the speed camera. Prior to the lottery, the average speed of a vehicle passing the camera was 32 km/h. After the lottery began, the average speed decreased to 25 km/h, a drop of 22%. The winner of the lottery won SEK 20,000 (approximately \$3,000 USD at the time) all of which was funded by the penalties paid by speeding motorists. As speeding is necessary for some aggressive behavior, such a campaign may help to reduce aggressive driving as well.

Recently, automobile insurance companies have been introducing rate reduction programs whereby insured drivers can elect to have their driving monitored (e.g., through a cellular phone application) and to earn reduced rates through safe driving practices. A 1-year evaluation of this type of pay-as-you-drive insurance program offered to young drivers in the Netherlands found a reduction in speed violations (Bolderdijk et al., 2011).

Multimedia-based Countermeasures

Multimedia campaigns targeting aggressive driving could also be effective. In Australia an advertising campaign was implemented targeting young male drivers who engaged in risky driving. The campaign conveyed a contemptuous sexual insult, suggesting that reckless speeding was the result of overcompensating for a deficit in young males' manhood. While previous fear-arousing campaigns were ineffective, there is evidence of this campaign's efficacy. There was a significant drop in speeding tickets and speed-related collision fatalities. In 2006, 64 young males (17–25 years of age) died in speed-related collisions. The number of fatalities dropped to 35 in 2007, the year "Pinkie" launched, and just 37 fatalities in 2008 (Roads and Traffic Authority, 2009). This campaign was likely more effective than others as it targeted a specific group of drivers (e.g., young males) and the specific motivation behind their behavior (e.g., impressing females).

Future Considerations

Efficacy of Countermeasure Research

From reading this article, it should be clear that much of what is known about aggressive driving revolves around contributing factors. Much less research has been devoted to countermeasure research. This article has discussed psychological, environmental, enforcement, technological, incentive, and multimedia-based countermeasures. Future research should employ sound methodology to assess the efficacy of such countermeasures. Furthermore, evaluation research should be conducted over an extended period to determine the longevity of any countermeasure influence on a behavior. The behavior in question should also be assessed prior to the implementation. All too often, countermeasures are implemented at the same time assessments of the negative behavior begin. In such cases, if changes in the negative behavior occur, they cannot be conclusively attributed to the countermeasure due to potential confounding variables.

Consolidation of Research

Psychology is advancing our knowledge of factors contributing to driver aggression, adding to the list of relevant variables and expanding our understanding of existing factors. Person-related and situational variables operate together; thus it is imperative that we continue to investigate how the contributions of multiple factors combine and interact to influence aggressive roadway behavior. We also need to understand the mechanisms underlying the influence of contributory factors. Personality, cognition, and affect all influence each other, and an improved assessment of the temporal order and strength of these influences is needed. Efforts to apply this information to modify driver aggression through policy, incentive-based approaches, psychotherapeutic programs (e.g., attributional retraining), and technological innovations to the vehicle and the roadway environment (e.g., electronic message boards) are in their infancy but possess great potential for impact.

Conclusion

In attempting to deal with the problem of aggressive driving, it is important to bear in mind that these behaviors have a variety of causative factors that have been summarized in this article. Given the diversity of factors, it is necessary to consider that no one suggestion is likely to reduce aggressive driving, but rather multiple approaches should be considered. When considering solutions, we urge that evaluation research be part of any program. Countermeasures need to be assessed for their efficacy, cost/benefits, and of course, political considerations.

References

- Ajzen, I., 1991. The theory of planned behavior. *Organ Behav Hum Decis Process* 50, 179–211.
- Ajzen, I., 2005. *Attitudes, Personality and Behavior*. McGraw-Hill Education, UK.
- American Psychiatric Association, 2013. *Diagnostic and Statistical Manual of Mental Disorders*, fifth ed. Author, Washington, DC.
- Anderson, C.A., Bushman, B.J., 2002. Human aggression. *Annu. Rev. Psychol.* 53, 27–51.
- Arnett, J.J., 2002. Developmental sources of crash risk in young drivers. *Injury Prevent.* 8 (Suppl. 2), ii17–ii23.
- Barkley, R.A., Cox, D., 2007. A review of driving risks and impairments associated with attention-deficit/hyperactivity disorder and the effects of stimulant medication on driving performance. *J. Safet. Res.* 38, 113–128.
- Benavidez, D.C., Flores, A.M., Fierro, I., Álvarez, F.J., 2013. Road rage among drug dependent patients. *Accid. Anal. Prevent.* 50, 848–853.
- Berkowitz, L., Lepage, A., 1967. Weapons as aggression-eliciting stimuli. *J. Pers. Soc. Psychol.* 7, 202–207.
- Bogdan, S.R., Măirean, C., Havârneanu, C.E., 2016. A meta-analysis of the association between anger and aggressive driving. *Transport. Res. F Traff. Psychol. Behav.* 42, 350–364.
- Bolderdijk, J.W., Knockaert, J., Steg, E.M., Verhoef, E.T., 2011. Effects of pay-as-you-drive vehicle insurance on young drivers' speed choice: results of a Dutch field experiment. *Accid Anal Prev* 43, 1181–1186.
- Bone, S.A., Mowen, J.C., 2006. Identifying the traits of aggressive and distracted drivers: a hierarchical trait model approach. *J. Cons. Behav.* 5, 454–464.
- Britt, T.W., Garrity, M.J., 2006. Attributions and personality as predictors of the road rage response. *Br. J. Soc. Psychol.* 45, 127–147.
- Brodsky, W., 2002. The effects of music tempo on simulated driving performance and vehicular control. *Transportat. Res. F Traff. Psychol. Behav.* 4, 219–241.

- Bushman, B.J., Bonacci, A.M., Pedersen, W.C., Vasquez, E.A., Miller, N., 2005. Chewing on it can chew you up: effects of rumination on triggered displaced aggression. *J. Pers. Soc. Psychol.* 88, 969–983.
- Bushman, B.J., Kerwin, T., Whitlock, T., Weisenberger, J.M., 2017. The weapons effect on wheels: Motorists drive more aggressively when there is a gun in the vehicle. *J. Exp. Soc. Psychol.* 73, 82–85.
- Butters, J.E., Mann, R.E., Smart, R.G., 2006. Assessing road rage victimization and perpetration in the Ontario adult population. *Can. J. Pub. Health* 97, 96–99.
- Chai, J., Zhao, G., 2016. Effect of exposure to aggressive stimuli on aggressive driving behavior at pedestrian crossings at unmarked roadways. *Accid. Anal. Prev.* 88, 159–168.
- Chraïf, M., Aniței, M., Burtăverde, V., Mihăilă, T., 2015. The link between personality, aggressive driving, and risky driving outcomes – Testing a theoretical model. *J. Risk Res.* 19, 780–797.
- Dahlen, E.R., Martin, R.C., Ragan, K., Kuhlman, M.M., 2005. Driving anger, sensation seeking, impulsiveness, and boredom proneness in the prediction of unsafe driving. *Accid. Anal. Prev.* 37, 341–348.
- Dahlen, E.R., White, R.P., 2006. The Big Five factors, sensation seeking, and driving anger in the prediction of unsafe driving. *Pers. Individ. Differ.* 41, 903–915.
- Davies, G.M., Patel, D., 2005. The influence of car and driver stereotypes on attributions of vehicle speed, position on the road and culpability in a road accident scenario. *Legal Criminol. Psychol.* 10, 45–62.
- Deffenbacher, J.L., Deffenbacher, D.M., Lynch, R.S., Richards, T.L., 2003. Anger, aggression, and risky behavior: a comparison of high and low anger drivers. *Behav. Res. Ther.* 41, 701–718.
- Diekmann, A., Jungbauer-Gans, M., Krassnig, H., Lorenz, S., 1996. Social status and aggression: a field study analyzed by survival analysis. *J. Soc. Psychol.* 136, 761–768.
- Dollard, J., Doob, L., Miller, N., Mowrer, O., Sears, R., 1939. *Frustration and Aggression*. Yale University Press, New Haven, CT.
- Doob, A.N., Gross, A.E., 1968. Status of frustrator as an inhibitor of horn-honking responses. *J. Soc. Psychol.* 76, 213–218.
- Efrat, K., Shoham, A., 2013. The theory of planned behavior, materialism, and aggressive driving. *Accid. Anal. Prev.* 59, 459–465.
- Elison, P.A., Govern, J.M., Petri, H.L., Figler, M.H., 1995. Anonymity and aggressive driving behavior: A field study. *J. Soc. Behav. Pers.* 10, 265–272.
- Ellison-Potter, P., Bell, P., Deffenbacher, J., 2001. The effects of trait driving anger, anonymity, and aggressive stimuli on aggressive driving behavior. *J. Appl. Soc. Psychol.* 31, 431–443.
- Fierro, I., Morales, C., Álvarez, F.J., 2011. Alcohol use, illicit drug use, and road rage. *J. Stud. Alcohol Drugs* 72, 185–193.
- Fruhen, L.S., Flin, R., 2015. Car driver attitudes, perceptions of social norms and aggressive driving behaviour towards cyclists. *Accid. Anal. Prev.* 83, 162–170.
- Galovski, T.E., Blanchard, E.B., 2004. Road rage: a domain for psychological intervention? *Aggr. Viol. Behav.* 9, 105–127.
- Galovski, T., Blanchard, E.B., Veazey, C., 2002. Intermittent explosive disorder and other psychiatric co-morbidity among court-referred and self-referred aggressive drivers. *Behav. Res. Ther.* 40, 641–651.
- Galovski, T.E., Malta, L.S., Blanchard, E.B., 2006. *Road Rage: Assessment and Treatment of the Angry, Aggressive Driver*. American Psychological Association, Washington, DC.
- Goldenfeld, C., Daniels, S., Schermers, G., 2019. Red light cameras revisited. Recent evidence on red light camera safety effects. *Accid. Anal. Prev.* 128, 139–147.
- González-Iglesias, B., Gómez-Fraguela, J.A., Luengo-Martín, M.Á., 2012. Driving anger and traffic violations: gender differences. *Transport. Res. F Traff. Psychol. Behav.* 15, 404–412.
- Haje, B.E., Symboluk, D.G., 2014. Personal and social determinants of aggressive and dangerous driving. *Can. J. Fam. Youth* 6, 59–88.
- Hemenway, D., Vrinotis, M., Miller, M., 2006. Is an armed society a polite society? Guns and road rage. *Accid. Anal. Prev.* 38, 687–695.
- Hennessy, D.A., Jakubowski, R., Benedetti, A.J., 2005. The influence of the actor-observer bias on attributions of other drivers, in: Hennessy, D.A., Wiesenenthal, D. (Eds.), *Contemporary Issues in Road Safety*. Nova Science Publishers, New York.
- Hennessy, D.A., Wiesenenthal, D.L., 1999. Traffic congestion, driver stress, and driver aggression. *Aggr. Behav.* 25, 409–423.
- Hennessy, D.A., Wiesenenthal, D.L., 2002. The relationship between driver aggression, violence, and vengeance. *Viol. Vict.* 17, 707–718.
- Hennessy, D.A., Wiesenenthal, D.L., Wickens, C., Lustman, M., 2004. The impact of gender and stress on traffic aggression: Are we really that different? In: Morgan, J.P. (Ed.), *Focus on Aggression Research*. Nova Science Publishers, Hauppauge, NY, pp. 157–174.
- Herrero-Fernández, D., 2013. Do people change behind the wheel? A comparison of anger and aggression on and off the road. *Trans. Res. F Traff. Psychol. Behav.* 21, 66–74.
- Ilie, G., Mann, R.E., Ialomiteanu, A., Adlaf, E.M., Hamilton, H., Wickens, C.M., Asbridge, M., Rehm, J., Cusimano, M.D., 2015. Traumatic brain injury, driver aggression and motor vehicle collisions in Canadian adults. *Accid. Anal. Prev.* 81, 1–7.
- Jones, E.E., Nisbett, R.E., 1971. *The Actor and the Observer: Divergent Perceptions of the Causes of Behavior*. General Learning Press, Morristown, NJ.
- Jovanović, D., Lipovac, K., Stanojević, P., Stanojević, D., 2011. The effects of personality traits on driving-related anger and aggressive behaviour in traffic among Serbian drivers. *Transport. Res. F Traff. Psychol. Behav.* 14, 43–53.
- Kenrick, D.T., MacFarlane, S.W., 1986. Ambient temperature and horn honking: a field study of the heat/aggression relationship. *Environ. Behav.* 18, 179–191.
- Kováčová, N., Rošková, E., Lajunen, T., 2014. Forgiveness, anger, and hostility in aggressive driving. *Accid. Anal. Prev.* 62, 303–308.
- Langer, E.J., 1975. The illusion of control. *J. Pers. Soc. Psychol.* 32, 311–328.
- Lawrence, C., Richardson, J., 2005. Gender-based judgments of traffic violations: the moderating influence of car type. *J. Appl. Soc. Psychol.* 35, 1755–1774.
- Lennon, A., Watson, B., Arlidge, C., Fraine, G., 2011. 'You're a bad driver but I just made a mistake': attribution differences between the 'victims' and 'perpetrators' of scenario-based aggressive driving incidents. *Transport. Res. F Traff. Psychol. Behav.* 14, 209–221.
- Lustman, M., Wiesenenthal, D.L., Flett, G.L., 2010. Narcissism and aggressive driving: Is an inflated view of the self a road hazard? *J. Appl. Soc. Psychol.* 40, 1423–1449.
- Mann, R.E., Smart, R.G., Stoduto, G., Adlaf, E.M., Ialomiteanu, A., 2004. Alcohol consumption and problems among road rage victims and perpetrators. *J. Stud. Alcohol* 65, 161–168.
- Matthews, G., Dorn, L., Glendon, A.I., 1991. Personality correlates of driver stress. *Pers. Individ. Differ.* 12, 535–549.
- Miles, D.E., Johnson, G.L., 2003. Aggressive driving behaviors: Are there psychological and attitudinal predictors? *Transport. Res. F* 6, 147–161.
- Miller, M., Azrael, D., Hemenway, D., Solop, F.I., 2002. 'Road rage' in Arizona: armed and dangerous. *Accid. Anal. Prev.* 34, 807–814.
- Mikula, G., 2003. Testing an attribution-of-blame model of judgments of injustice. *Eur. J. Soc. Psychol.* 33, 793–811.
- Moore, M., Dahlen, E.R., 2008. Forgiveness and consideration of future consequences in aggressive driving. *Accid. Anal. Prev.* 40, 1661–1666.
- Neighbors, C., Vietor, N.A., Knee, C.R., 2002. A motivational model of driving anger and aggression. *Pers. Soc. Psychol. Bull.* 28, 324–335.
- Paleti, R., Eluru, E., Bhat, C.R., 2010. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accid. Anal. Prev.* 42, 1839–1854.
- Parsons, R., Tassinary, L.G., Ulrich, R.S., Hebl, M.R., Grossman-Alexander, M., 1998. The view from the road: Implications for stress recovery and immunization. *J. Env. Psychol.* 18, 113–139.
- Pedersen, W.C., Denson, T.F., Goss, R.J., Vasquez, E.A., Kelley, N.J., Miller, N., 2011. The impact of rumination on aggressive thoughts, feelings, arousal, and behaviour. *Br. J. Soc. Psychol.* 50, 281–301.
- Raudenbush, B., Grayhem, R., Sears, T., Wilson, I., 2009. Effects of peppermint and cinnamon odor administration on simulated driving alertness, mood and workload. *N. Am. J. Psychol.* 11, 245–256.
- Richards, T.L., Deffenbacher, J.L., Rosen, L.A., Barkley, R.A., Rodricks, T., 2006. Driving anger and driving behavior in adults with ADHD. *J. Attent. Dis.* 10 (1), 54–64.
- Richer, I., Bergeron, J., 2009. Driving under the influence of cannabis: Links with dangerous driving, psychological predictors, and accident involvement. *Accid. Anal. Prev.* 41 (2), 299–307.
- Roads and Traffic Authority, 2009. Speeding. No one thinks big of you. 2009 Australian Effie Awards. <https://www.effie.com.au/attachments/cd29d4db-44e2-4c50-b86d-fca3acbd7c2b.pdf> (accessed 01.09.19).
- Roberts, L., Indemaur, D., 2005. Boys and road rage: driving related violence and aggression in western Australia. *Aus. N. Z. J. Criminol.* 38, 361–380.
- Roseborough, J.E., Wiesenenthal, D.L., 2018. The influence of roadway police justice on driver emotion. *Trans. Res. F Traff. Psychol. Behav.* 56, 236–244.
- Roseborough, J.E.W., Wiesenenthal, D.L., 2014. Roadway justice – making angry drivers, happy drivers. *Transport. Res. F Traff. Psychol. Behav.* 24, 1–7.

- Sărbescu, P., Stanojević, P., Jovanović, D., 2014. A cross-cultural analysis of aggressive driving: Evidence from Serbia and Romania. *Transport. Res. F Traff. Psychol. Behav.* 24, 210–217.
- Shinar, D., 1998. Aggressive driving: the contribution of the drivers and situation. *Transport. Res. F Traff. Psychol. Behav.* 1, 137–160.
- Smart, R.G., Cannon, E., Howard, A., Frise, P., Mann, R.E., 2005. Can we design cars to prevent road rage? *Int. J. Vehicl. Info. Commun. Sys.* 1, 44–55.
- Spielberger, C.D., Johnson, E.H., Russell, S.F., Crane, R.J., Jacobs, G.A., Worden, T.J., 1985. Anger and hostility in cardiovascular and behavioral disorders. In: Chesney, M.A., Rosenman, R.H. (Eds.), *Hemisphere Publishing*, Washington, DC, pp. 5–30.
- Stanojević, P., Sullman, M.J., Jovanović, D., Stanojević, D., 2018. The impact of police presence on angry and aggressive driving. *Accid. Anal. Prev.* 110, 93–100.
- Stanojević, P., Fitzharris, M., 2017. Aggressive driving on Australian Roads, Australasian Road Safety Conference, October 10–12, Perth, Australia.
- Stephens, A.N., Groeger, J.A., 2009. Situational specificity of trait influences on drivers' evaluations and driving behaviour. *Transport. Res. F Traff. Psychol. Behav.* 12, 29–39.
- Stephens, A.N., Groeger, J.A., 2014. Lead driver status moderates anger and exonerates culpability. *Transport. Res. F Traff. Psychol. Behav.* 22, 140–149.
- Stephens, A.N., Ohtsuka, K., 2014. Cognitive biases in aggressive drivers: Does illusion of control drive us off the road? *Pers. Individ. Differ.* 68, 124–129.
- Stephens, A.N., Sullman, M.J.M., 2015. Trait predictors of aggression and crash-related behaviors across drivers from the United Kingdom and the Irish Republic. *Risk Anal.* 35, 1730–1745.
- Stephens, A.N., Trawley, S.L., Ohtsuka, K., 2016. Venting anger in cyberspace: Self-entitlement versus self-preservation in #roadrage tweets. *Transport. Res. F Traff. Psychol. Behav.* 42, 400–410.
- Suhr, K.A., 2016. Mulling over anger: Indirect and conditional indirect effects of thought content and trait rumination on aggressive driving. *Transport. Res. F Traff. Psychol. Behav.* 42, 276–285.
- Suhr, K.A., Dula, C.S., 2017. The dangers of rumination on the road: predictors of risky driving. *Accid. Anal. Prevent.* 99, 153–160.
- Suhr, K.A., Nesbit, S.M., 2013. Dwelling on 'Road Rage': the effects of trait rumination on aggressive driving. *Transport. Res. F Traff. Psychol. Behav.* 21, 207–218.
- Sullman, M.J.M., Stephens, A.N., Hill, T., 2017. Gender roles and the expression of driving anger among Ukrainian drivers. *Risk Anal.* 37, 52–64.
- Sümer, N., Lajunen, T., Özkant, T., 2005. Big five personality traits as the distal predictors of road accident involvement. in: Underwood, G. (Ed.), *Traffic and Transport Psychology: Theory and Application—Proceedings of the ICTTP 2004*. Elsevier, New York, NY, pp. 215–227.
- Takaku, S., 2006. Reducing road rage: an application of the dissonance-attribution model of interpersonal forgiveness. *J. Appl. Soc. Psychol.* 36, 2362–2378.
- Wickens, C.M., Mann, R.E., Ialomiteanu, A.R., Vingilis, E., Seeley, J., Erickson, P., Kolla, N.J., 2019. The association of childhood symptoms of conduct disorder and collision risk in adulthood. *J. Trans. Health* 13, 33–40.
- Wickens, C.M., Mann, R.E., Roseborough, J.E.W., Wiesenenthal, D.L., Smart, R.G., 2014. The straight 'A's of road rage: Attribution, anger, and aggressive behaviour. In: Penrod, M.G., Paulk, S.N. (Eds.), *Psychology of anger: New research*. NOVA Science Publishers, New York, pp. 49–69.
- Wickens, C.M., Mann, R.E., Stoduto, G., Butters, J.E., Ialomiteanu, A., Smart, R.G., 2012. Does gender moderate the relationship between driver aggression and its risk factors? *Accid. Anal. Prevent.* 45, 10–18.
- Wickens, C.M., Mann, R.E., Stoduto, G., Ialomiteanu, A., Smart, R.G., 2011a. Age group differences in self-reported road rage perpetration and victimization. *Transport. Res. F Traff. Psychol. Behav.* 14, 400–412.
- Wickens, C.M., Roseborough, J.E., Hall, A., Wiesenenthal, D.L., 2013a. Anger-provoking events in driving diaries: a content analysis. *Transport. Res. F Traff. Psychol. Behav.* 19, 108–120.
- Wickens, C.M., Vingilis, E., Mann, R.E., Erickson, P., Toplak, M.E., Kolla, N.J., Seeley, J., Ialomiteanu, A.R., Stoduto, G., Ilie, G., 2015. The impact of childhood symptoms of conduct disorder on driver aggression in adulthood. *Accid. Anal. Prevent.* 78, 87–93.
- Wickens, C.M., Wiesenenthal, D.L., 2005. State driver stress as a function of occupational stress, traffic congestion, and trait stress susceptibility. *J. Appl. Biobehav. Res.* 10, 83–97.
- Wickens, C.M., Wiesenenthal, D.L., Flora, D.B., Flett, G.L., 2011b. Understanding driver anger and aggression: attributional theory in the driving environment. *J. Exp. Psychol. Appl.* 17, 354–370.
- Wickens, C.M., Wiesenenthal, D.L., Hall, A., Roseborough, J.E., 2013b. Driver anger on the information superhighway: a content analysis of online complaints of offensive driver behaviour. *Accid. Anal. Prevent.* 51, 84–92.
- Wiesenenthal, D.L., Hennessy, D.A., Totten, B., 2000. The influence of music on driver stress. *J. App. Soc. Psychol.* 30, 1709–1719.
- Wiesenenthal, D.L., Hennessy, D.A., Totten, B., 2003. The influence of music on mild driver aggression. *Transport. Res. F Traff. Psychol. Behav.* 6, 125–134.
- Wiesenenthal, D.L., Janovjak, D.P., 1992. *Deindividuation and Automobile Driving Behaviour*. LaMarsh Research Programme Report Series: Vol 46. LaMarsh Research Programme on Violence and Conflict Resolution. York University, Toronto, Canada.
- Zhang, H., Qu, W., Ge, Y., Sun, X., Zhang, K., 2017. Effect of personality traits, age and sex on aggressive driving: psychometric adaptation of the Driver Aggression Indicators Scale in China. *Accid. Anal. Prev.* 103, 29–36.

Further Reading

- Deery, H.A., Fildes, B.N., 1999. Young novice driver subtypes: relationship to high-risk behavior, traffic accident record, and simulator driving performance. *Hum. Factors* 41, 628–643.
- Simon, F., Corbett, C., 1996. Road traffic offending stress, age, and accident history among male and female drivers. *Ergonomics* 39, 757–780.
- Stanojević, P., Jovanović, D., Lajunen, T., 2013. Influence of traffic enforcement on the attitudes and behavior of drivers. *Accid. Anal. Prevent.* 52, 29–38.
- Vitaglione, G., 2012. Driving under the influence of media: a four-year examination of NASCAR and West Virginia aggressive driving accident and injuries. *J. Appl. Soc. Psychol.* 42, 488–505.

Aircraft Maintenance and Inspection

Alan Hobbs, San Jose State University Research Foundation at NASA Ames Research Center, Moffett Field, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	25
Maintenance Personnel	26
Quality Lapses in Maintenance	27
Human Factors in Maintenance	28
Human Error in Maintenance	28
Memory Failures	29
Procedural Noncompliance	29
Assumption Errors	29
Knowledge-Based Errors	29
Performance-Shaping Factors in Maintenance	29
Interventions to Improve the Quality of Maintenance and Inspection	30
Nontechnical Skills Training	30
Improved Documentation and Procedures	30
Design for Maintainability	30
Fatigue Management	31
Learning from Incidents	31
The Future of Maintenance and Inspection	32
References	32
Further Reading	33

Introduction

Without the intervention of maintenance personnel, equipment used in complex technological systems such as aviation, rail transport, and medicine would inevitably deteriorate to a point where safety and profitability would be threatened. After fuel, maintenance is the largest cost facing airlines (GRA Inc, 2007).

Despite advances such as Built-in Test Equipment (BITE), on-board sensors, and flight data monitoring, maintenance is still a largely human activity. Maintenance technicians work in an environment that is more hazardous than nearly all other jobs in the labor force. The work requires physical strength, dexterity, and balance, but also calls for clerical skills and attention to detail. The work may be carried out at heights, in confined spaces, in numbing cold or sweltering heat (Fig. 1). Communication can be difficult due to noise levels and the use of hearing protection. Although maintenance makes an essential contribution to system reliability, improper maintenance can be a major cause of system failure.

Maintenance activities can be divided into the two broad categories of preventative and corrective as shown in Fig. 2. Preventative maintenance includes lubrication, inspections, and other tasks that can be scheduled in advance. Many preventative tasks are performed regularly at airlines, sometimes daily or as part of a block of tasks such as an “A” check performed after hundreds of hours of flight. Other forms of preventative maintenance occur at greater intervals, sometimes bundled into “C” checks that may be scheduled yearly or less frequently, or heavy maintenance “D” checks. The airline industry currently outsources more than half of its preventative maintenance tasks to specialized Maintenance Repair and Overhaul (MRO) organizations.

In the early years of the airline industry, each airline developed its own preventative maintenance program that called for intensive maintenance at predetermined intervals. These programs sometimes required the complete overhaul of systems, regardless of whether the maintenance was necessary or effective. In the 1970s, there was an increasing realization that scheduled maintenance programs needed to be tailored to the needs of each component or system, based on the failure patterns of each component, and the consequences should a failure occur. Today, regulators, manufacturers, and airlines work together to develop maintenance programs for aircraft fleets based on Reliability Centered Maintenance principles (Moubray, 2001) as outlined in a document known as MSG-3 (Maintenance Steering Group). This approach ensures that systems are maintained at an appropriate level, while avoiding over-maintenance and the increased potential for error that this introduces.

Corrective maintenance, also known as unscheduled maintenance, is performed in response to unplanned operational events such as aircraft damage, component failure, or defects discovered during a scheduled check. Corrective maintenance includes fault isolation, repair, and replacement of faulty components. Although some corrective tasks are minor, others require extensive system knowledge, problem solving, and specialized skills. Some nonscheduled inspections are initiated in response to events such as a bird strikes, lightning strikes, and heavy landings. These inspections are sometimes referred to as “Chapter 5” inspections as they are found in this chapter of maintenance manuals following the industry-standard ATA numbering system. Major nonscheduled



Figure 1 A maintenance technician at work. *Source: Image courtesy of American Airlines*

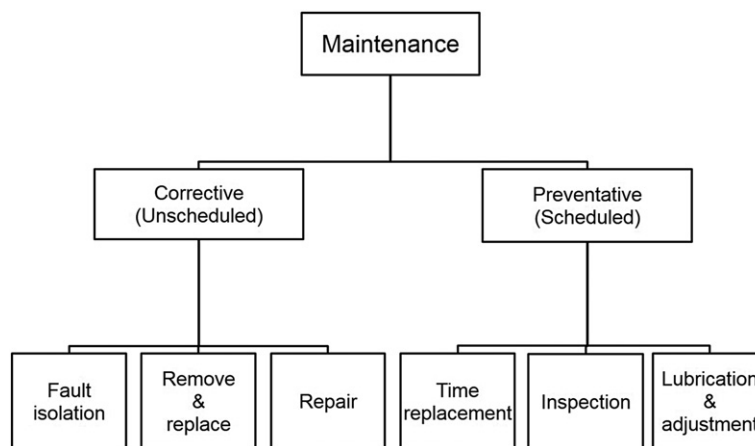


Figure 2 Major categories of maintenance. *Source: US DoD (1997)*

maintenance can be particularly disruptive and expensive for airlines. An “Aircraft on Ground” (AOG) requires an immediate response that may include flying maintenance personnel to the stranded aircraft, rapidly locating spare parts, and coordinating with the original equipment manufacturer.

Maintenance Steering Group-3 defines three levels of inspections: general visual, detailed, and special detailed. General visual inspections are carried out within touching distance of the area being examined, to detect obvious damage such as clearly visible cracks, corrosion or dents. These inspections do not require special equipment, although ladders or workstands may be used, and there may be a need to open access panels. Detailed inspections involve intense visual and/or tactile examination for less obvious defects, sometimes with the use of lenses or mirrors. There may also be a need for extensive disassembly to gain access to the area to be inspected. Lastly, special detailed inspections are examinations of an area involving the use of non-destructive inspection (NDI) techniques. These include ultrasound, thermography, X-ray, and dye penetrant techniques. The effectiveness of inspections relies on the ability of the inspector to detect the sensory signs that could indicate a defect, and then decide whether a defect is actually present. Factors such as lighting, time of day, monotony, vigilance, and the adequacy of rest breaks can all impact inspector performance (Drury and Watson, 2002).

Maintenance Personnel

In many countries, including those covered by the rules of the European Aviation Safety Agency (EASA), aircraft maintenance licenses are divided into two basic categories. The first category enables the technician to certify work on structures, engines, mechanical, and electrical systems. The second category applies to avionics (the electronic systems used on aircraft). Some personnel possess both licenses or additional qualifications providing further signature authority. In most cases a licensed technician with the appropriate type rating has the authority to certify that their own work or that of a colleague was performed correctly.

In the United States, qualified maintenance mechanics are referred to as Aviation Maintenance Technicians (AMTs) or Airframe and Powerplant technicians (A&Ps). Unlike other regulatory authorities, the US Federal Aviation Administration (FAA) does not require technicians to possess a specialized qualification to work on avionics systems. The FAA system also relies on independent inspectors as part of the quality control process. In addition to performing general and specialized inspections of structures and components, quality control inspectors certify that critical tasks have been performed correctly by AMTs.

Quality Lapses in Maintenance

According to the International Air Transport Association, improper maintenance was a factor in 17% of airline accidents in the period 2014–18 (IATA, 2019). Maintenance errors not only pose a threat to flight safety, but can also impose significant financial costs through delays, cancellations, diversions, and schedule disruptions. For example, in the case of a large wide-body airliner, a flight cancellation can cost the airline around USD \$200,000, while delays at the gate can cost over USD \$20,000 per hour. Therefore, even a small reduction in minor maintenance problems can result in major savings for the airline.

Throughout the 1980s and 1990s, a series of landmark accidents drew attention to the fact that aircraft inspection and maintenance relies on human performance. Improvements in the quality and efficiency of maintenance requires an understanding of the capabilities and limitations of maintenance personnel.

Improper maintenance was a primary factor in the world's worst single-aircraft accident, an event that claimed 520 lives. The 747-100 was on a short domestic flight in Japan when it experienced a sudden decompression due to the failure of the aft pressure bulkhead. The escaping air caused most of the vertical stabilizer and rudder to separate, and a loss of hydraulics pressure from all four systems. The pilots attempted to maneuver the aircraft using engine power, however they were unable to maintain control and after about 30 min the aircraft crashed into a mountain north west of Tokyo.

The investigators found that the aft pressure bulkhead had failed in flight due to a fatigue fracture in an area where a repair had been made seven years previously. The repair had involved replacing the lower half of the bulkhead. The new lower half should have been spliced to the upper half using a doubler plate that would have extended under three lines of rivets. However, part of the splice was made using two plates instead of a single plate as intended (Fig. 3). As a result, the join relied on only a single row of rivets. After the repair, the aircraft flew over 12,000 flights and underwent six C checks before the accident occurred (Japan Ministry of Transport, 1987).

In April 1988, an Aloha Airlines Boeing 737-200 en-route to Honolulu experienced an explosive decompression in which approximately 18 ft of cabin skin and structure separated from the aircraft. The pilots were able to make an emergency landing, however the accident resulted in one death and eight serious injuries. The NTSB determined that the accident was caused by the failure of the airline to detect the presence of significant disbonding and fatigue damage that ultimately led to the failure (National

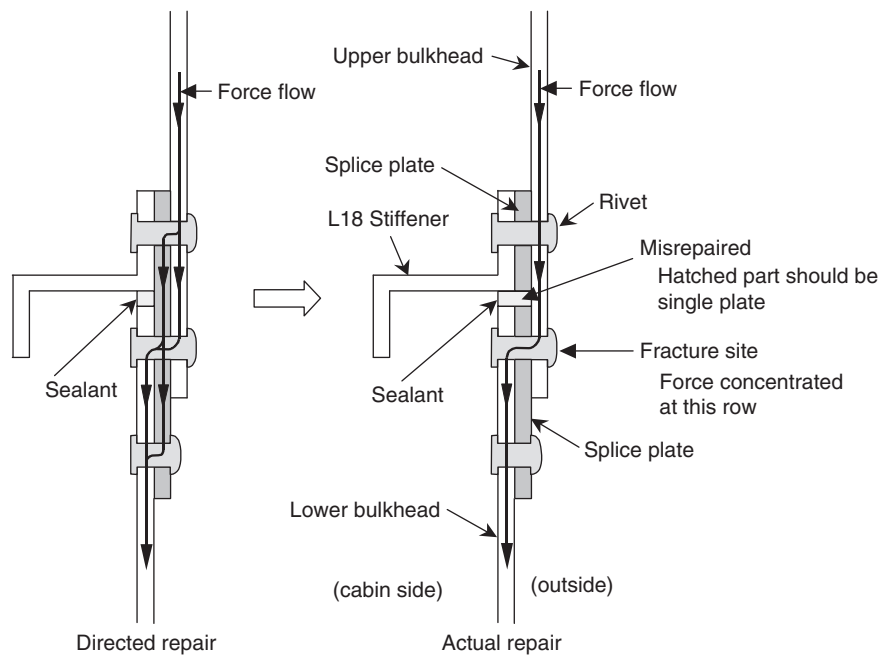


Figure 3 The repair to the aft pressure bulkhead as specified in the repair instructions (at left), and repair as actually carried out (on right). Note gap of doubler plate between top two rivets. Source: Terada and Kobayashi (2006).

Transportation Safety Board, 1989). As a result of the accident, the human factors of inspection became a major issue of concern, particularly in the United States, where the FAA sponsored an extensive research program on the topic. Details of the FAA work can be found in the recommended reading.

Little more than a year after the Aloha accident, inspection was once more the focus of attention when a United Airlines DC-10 suffered a catastrophic engine failure that damaged the aircraft's hydraulic system, resulting in a loss of flight controls. In a remarkable feat of airmanship, the crew was able to partially control the aircraft for a crash landing using engine thrust alone. 111 passengers died, however 175 passengers and 10 crewmembers survived. The engine failure occurred when the stage 1 fan rotor disk separated from the center engine as a result of an undetected fatigue crack. Fifteen months prior to the accident, the component had been inspected using a dye penetrant technique. A fluorescent fluid was applied to the component, which was then inspected under ultraviolet light to reveal the presence of cracks or discontinuities, however the inspection failed to find a 0.5 inch long fatigue crack. The [National Transportation Safety Board \(1990\)](#) concluded that the root cause of the accident was inadequate consideration of human factors limitations in the inspection and quality control procedures.

Human Factors in Maintenance

The accidents described above, and others like them, led to a major worldwide effort to understand the human factors of airline maintenance. Research sponsored by the FAA, Transport Canada, and other agencies has brought to light the unique human factors of maintenance and inspection. The lessons are now being applied by operators and regulatory authorities worldwide.

Some of the most commonly observed lapses in maintenance are:

- Equipment or part not installed (typically small parts such as O-rings or washers)
- Failure to detect damage during inspection
- Incomplete installation—often a nut or fastener left “finger tight” and not torqued
- Cross connections—electrical wiring or control cables
- Wrong parts fitted
- Parts fitted in the wrong location or orientation
- Loose objects or tools left in aircraft
- Panels, caps, or cowlings not secured.

Some maintenance quality lapses may remain undetected for months or years before discovery, or until an operational consequence occurs. Without doubt, some maintenance irregularities are never detected, and continue to fly with the aircraft until the end of its service life. Quality lapses in maintenance usually involve one or more human errors, however the culture of maintenance has tended to discourage the open reporting of minor maintenance errors. In recent years, confidential incident reporting systems and the application of “just culture” concepts has enabled the industry to gain a better understanding of the nature of maintenance error.

Human Error in Maintenance

Maintenance personnel make an invaluable and essential contribution to continuing airworthiness. As with all complex human/machine systems, however, some level of human error is inevitable, and the use of the term “error” should not automatically imply blame or fault. While individual human errors cannot be predicted, estimates of error rates are used widely in risk assessments. **Table 1** shows estimated error probabilities for maintenance tasks, based on military data. The reader should note that these numbers, although useful, should be seen as “best guesses,” to be treated with an appropriate level of skepticism. Error rates will vary according to the nature of the task, the environment, training, and other factors. Nevertheless, rates such as these can provide a general awareness of the level of risk introduced by human error. Although the probability of error for each instance of a task may be relatively low, if the task is performed numerous times, the overall chance of an error can become high.

Table 1 Estimated probabilities of maintenance errors

<i>Task</i>	<i>Estimated chance of error during task</i>
Install nuts and bolts	0.2%
Connect electrical cable	0.3%
Install O-ring	0.3%
Tighten nuts and bolts	0.4%
Read pressure gauge	1.1%
Install lock wire	3.2%
Check for error in another person's work	10.0%

Source: [Gertman and Blackman \(1993\)](#).

In some cases, particularly if a maintenance error has been discovered or reported in a timely manner, it will be possible to examine not only the observable outcome of the error (such as failing to tighten a nut), but also the thinking processes that led the technician to make the error. Human factors analysis not only gives us a powerful insight into the origins of the error, but also helps us develop ways to prevent or capture future errors. Applying human error models to maintenance discrepancies reveals that underlying these events are a limited range of cognitive error forms. More than 50% of the maintenance errors in the aviation industry can be placed in one of four categories: memory failures, procedural noncompliance, assumption errors, and knowledge-based errors (Hobbs and Williamson, 2003).

Memory Failures

One of the most common errors in maintenance incidents is memory failure. Rather than forgetting something about the past, the technician typically forgets to perform an action that they had intended to perform at some time in the future. Psychologists refer to memory for intentions as prospective memory. Two common examples are forgetting to reconnect a disconnected system at the end of a task and leaving an oil cap unsecured. Failures of prospective memory are particularly likely when a maintenance task has been interrupted and must be picked up again later.

Procedural Noncompliance

An aircraft hangar is a highly regulated workplace. Technicians are expected to carry out their duties while observing legal requirements, manufacturer's maintenance manuals, company procedures, and unwritten norms of safe behavior.

When asked about their most recent task, 34% of aircraft maintenance technicians in Europe acknowledged that they had not strictly complied with the formal procedures (McDonald et al., 2000). Examples are, signing off a task before it was completed, and performing a task without the correct tools or equipment. In most of these cases, the maintainer could probably have justified their actions, nevertheless the ubiquity of procedural noncompliance indicates a common divergence between formal procedures and actual task performance.

Assumption Errors

An assumption error occurs when a person misidentifies a situation, or fails to check that their understanding of a situation is correct. False assumptions usually occur in familiar situations where the person has the expertise to deal with the task. In maintenance, this commonly takes the form of a misunderstanding between colleagues, such as incorrectly assuming that another person has performed a task step. Careful communication is essential for effective shift handovers, when work-in-progress is continued from one shift to another.

Knowledge-Based Errors

The term "knowledge-based error" refers to mistakes arising from either failed problem-solving or a lack of system knowledge. Most maintenance technicians enjoy variety in their work, but to achieve this, they must be assigned unfamiliar tasks from time to time. Many maintenance technicians report that they experience some level of uncertainty when performing such tasks. Ambiguities encountered during the preparation stage, such as unclear procedures, may set the scene for errors that will emerge later in the task. Supervisors must ensure that technicians are able to draw on the support of experienced personnel whenever they are assigned nonroutine or challenging tasks.

Performance-Shaping Factors in Maintenance

Although some workplace errors or deviations from procedures reflect random variability in human performance, most are related to performance-shaping factors in the workplace. Some of the most common factors in maintenance incidents are time pressure, interruptions, fatigue, and documented procedures that are inadequate for the task. Memory lapses are often associated with interruptions and fatigue; procedural noncompliance is known to occur in response to both time pressure and inadequate procedures. Assumption errors are more likely when there is inadequate communication and coordination between technicians.

A widely-recognized description of error-producing conditions in maintenance is the "Dirty Dozen" developed by Gordon Dupont at Transport Canada. This list of 12 factors is commonly used in human factors training for maintenance technicians.

The Dirty Dozen Maintenance Human Factors are:

1. Lack of communication
2. Complacency
3. Lack of knowledge
4. Distraction

5. Lack of teamwork
6. Fatigue
7. Lack of resources (including documentation, parts, staffing)
8. Pressure (externally and self-imposed)
9. Lack of assertiveness
10. Stress
11. Lack of awareness
12. Norms

Interventions to Improve the Quality of Maintenance and Inspection

The remainder of this section describes several interventions that can lead to improved human performance in maintenance. These interventions are; nontechnical skills training, improved documentation and procedures, design for maintainability, fatigue management, and learning from incidents.

Nontechnical Skills Training

The International Civil Aviation Organization (ICAO), the European Aviation Safety Agency (EASA), Transport Canada, the Australian Civil Aviation Safety Authority, and many other regulatory agencies require maintenance staff to have an understanding of human factors principles. While stopping short of requiring such training, the FAA has released extensive educational material on maintenance human factors.

EASA (Part 66) requires that human factors knowledge is included among the basic initial knowledge requirements for certifying maintenance staff on commercial air transport aircraft. EASA (Part 145) requires that maintenance organizations provide regular human factors training to staff. The training is required not only for certifying staff, engineers, and technicians, but also for managers, supervisors, store personnel, and others. Human factors continuation training must occur every two years. Over 60 human factors topics are listed in the guidance material associated with the EASA regulation, including violations, peer pressure, memory limitations, workload management, teamwork, assertiveness, and disciplinary policies. The EASA human factors requirements are becoming a de-facto standard, even in regions of the world not regulated by EASA.

Improved Documentation and Procedures

According to the [FAA \(2012\)](#), aviation maintenance personnel spend between 25% and 40% of their time dealing with maintenance documentation. Poor documentation is one of the leading causes of maintenance incidents, and maintainers often report that procedures do not fully meet their needs ([Chaparro and Groff, 2002](#)). Documentation can be particularly challenging at MROs where technicians must switch between the maintenance procedures of different airlines.

There are some very basic aspects of document design that can have a major impact on the usability of a document. The Simplified Technical English specification developed by the [AeroSpace and Defence Industries Association of Europe \(2017\)](#) can help to create succinct and unambiguous text. This is particularly crucial when the reader has a first language other than English. Simplified English is used widely in the production of aerospace manuals and is now used by both Boeing and Airbus. Simplified English limits the number of words used to describe steps, and also ensures that each word only has one meaning. For example, in everyday English the word “tap” could have several different meanings including the tap above a sink, the action of removing fluid from something, or to listen in on a telephone conversation. In simplified English, “Tap” only has the meaning of to hit something, as in “tap with a hammer.”

Improving documentation is one of the most cost-effective ways to improve quality and reduce maintenance error. Maintenance organizations generally have systems to allow personnel to report problems and make suggestions for improvements in documentation. In many cases however, mechanics express frustration at the slow pace of such systems and the lack of responses to their suggestions.

Design for Maintainability

Significant improvements in cockpit design and layout have occurred since WWII. Yet the design of equipment for ease of maintenance has received less attention. Poor design is a major factor leading to maintenance problems. Examples include:

- Plumbing or electrical connections that permit cross connection
- Components that are difficult to reach, particularly where unrelated components must be disconnected to enable access
- Obstructions to vision
- Procedures that require levels of precision or force that are difficult for the technician to deliver
- Components that can be installed in the reverse sense.

Table 2 Examples of maintainability guidelines in Military Handbook 470-A

Guidelines number	Description
EC-26	Avoid using identical electrical connectors in adjacent areas.
EC-15	The removal or replacement of electronic equipment should not require the removal of any other piece of equipment.
EC-24	Use electrical connectors that incorporate alignment key-ways to reduce incidence of damage due to improper engagement.
ENG (G)-12	Provide a clear and viewable access envelope to fuel and oil filters.
CONT-06	Design all cables and brackets associated with cable installations, so they are accessible by a 75 percentile male hand.

Design standards such as the US Department of Defense Handbook 470-A contain numerous guidelines for maintainability, many of which have relevance to aircraft design (Department of Defense, 1997). Illustrative examples are listed in [Table 2](#).

Fatigue Management

Aviation maintenance personnel face a heightened risk of fatigue due to night shift work, the potential for long duty times, and the sleep disruption that can result from these working arrangements.

In recent years, comprehensive Fatigue Risk Management Systems (FRMS) have been applied to aviation maintenance (FAA, 2016). An FRMS is a data driven and scientific approach to fatigue management that utilizes a range of strategies. Some of these interventions assist the individual; examples are educational material, medical screening, and treatment of sleep disorders. Other interventions are aimed at tasks. These include avoiding the scheduling of critical tasks during times of heightened fatigue risk, and keeping the most critical tasks out of the hands of the most fatigued people (progressive restriction of responsibilities). A typical FRMS will also include a statement of policy and management commitment, a risk assessment strategy, and an incident reporting and analysis system. At the core of most FRMS are limits on the working hours of maintenance personnel.

Professor Simon Folkard has developed a set of widely-adopted working hours guidelines for airline maintenance technicians. They can be found in full on the website of the UK Civil Aviation Authority (CAA). Five key items from Folkard's guidelines are:

- There should be a 12-h limit on shift duration
- No shift should be extended beyond 13 h by overtime
- A break of at least 11 h should occur between shifts
- There should be a work break every 4 h
- A month's notice of work schedules should be provided.

Learning from Incidents

Incident reports are one of the few channels for organizations to identify organizational problems in maintenance, yet the culture of maintenance around the world has tended to discourage the open reporting of maintenance incidents. This is because the response to errors has frequently been punitive. In some companies, technicians who make errors will be punished by days without pay, or even instant dismissal. It is hardly surprising that many minor maintenance incidents are never officially reported.

A growing trend around the world is to encourage incident reporting in maintenance by giving reporters limited immunity from disciplinary action or prosecution. David Marx, an aeronautical engineer and lawyer, has promoted the idea of a "just culture" in which some unsafe acts will result in discipline, however most will not. Marx divides unsafe actions of people into four overlapping categories: (1) Human error, (2) Negligence, (3) Recklessness, and (4) Intentional rule violations. Negligence is a failure to recognize a risk that should have been recognized. Recklessness is a conscious disregard for a visible significant risk.

Part 145 of the EASA regulations requires maintenance organizations to have an internal occurrence reporting scheme that enables occurrences, including those related to human error, to be reported and analyzed.

In the United States, the FAA encourages airlines and repair stations to introduce Aviation Safety Action Programs (ASAP) that allow employees to report safety issues with an emphasis on corrective action rather than discipline. Incident reports are passed to an event review committee comprising representatives of the FAA, management, and unions.

Several investigation techniques have been developed specifically for airline maintenance events. The best known is Boeing's Maintenance Event Decision Aid (MEDA) which includes a comprehensive list of event descriptors, such as "access panel not closed" and then guides the investigator in identifying the contributing factors that led to the event. Over 70 contributing factors are listed, including fatigue, inadequate knowledge, and time constraints.

Investigators typically find that major accidents are preceded by numerous minor events. Minor everyday close-calls or incidents can serve as raw materials for a safety evaluation system. An example of such an approach is the 38-item Maintenance Environment Questionnaire (MEQ) (Hobbs, 2005).

The MEQ assesses the extent of seven error-producing conditions in maintenance workplaces: Fatigue, coordination, time pressure, knowledge, supervision, availability of parts and equipment, and procedures. The MEQ can supplement incident

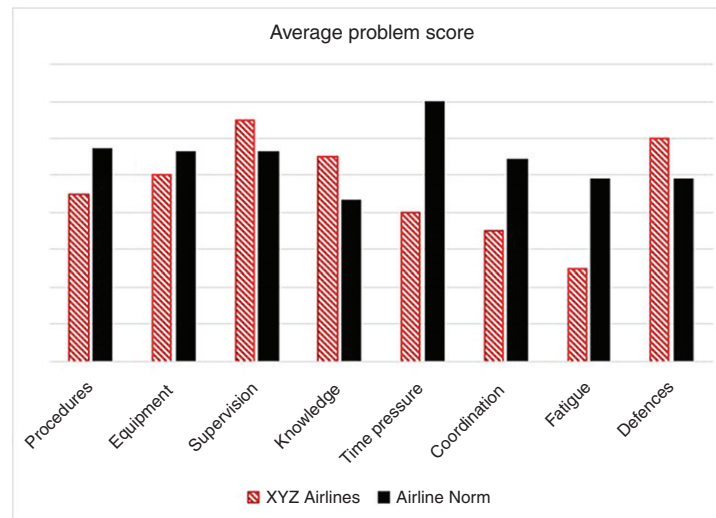


Figure 4 Example of a maintenance environment questionnaire profile.

investigations by providing a large amount of information on workplace human factors to enable comparisons with industry norms (Fig. 4).

The Future of Maintenance and Inspection

Emerging technologies have the potential to significantly change the processes involved in aircraft maintenance and inspection. Today, airline maintenance personnel routinely receive performance data from aircraft in-flight. This enables emerging problems with engines or other systems to be anticipated and diagnosed before the aircraft arrives at the gate, thereby reducing schedule disruptions.

For inspection, developments include more effective nondestructive inspection (NDI) techniques, smart materials with the ability to indicate damage through observable “bruising,” and structural health monitoring (SHM). The ability to readily transmit information over the web is enabling NDI results to be interpreted in real-time by personnel remote from the structure being examined. Several airlines are currently testing drones to assist in structural inspections. Data from permanently-placed sensors can be used as part of an SHM program. Benefits include reduced inspection times and the ability to monitor difficult-to-access areas of the airframe without the need to create access for a technician.

Despite advances in technology, the airline industry will continue to rely on the perceptual capabilities, judgment, teamwork, and communication skills of maintenance personnel. Airline travel is now the safest mode of transport. Part of the reason for this is that the industry has continuously learned and applied knowledge from the field of human factors to maximize the performance of maintenance personnel and drive down the rate of maintenance error.

References

- AeroSpace and Defence Industries Association of Europe, 2017. Simplified Technical English. Specification ASD-STE100. ASD, Brussels.
- Chaparro, A. Groff, L., 2002. Human factors survey of aviation maintenance technical manuals. Available from: www.faa.gov.
- Department of Defense, 1997. Designing and Developing Maintainable Products and Systems. Military Handbook 470-A. DoD, Washington, DC.
- Drury, C. Watson, J., 2002. Good practices in visual inspection. Available from: www.faa.gov.
- FAA, 2012. Maintenance review boards, maintenance type boards, and original equipment manufacturer and type-certificate holder recommended maintenance procedures. Advisory Circular 121-22C. Federal Aviation Administration, Washington, DC.
- FAA, 2016. Maintainer fatigue risk management. Advisory Circular AC No: 120-115. Federal Aviation Administration, Washington, DC.
- Gertman, D.I., Blackman, H.S., 1993. Human Reliability and Safety Analysis Data Handbook. John Wiley & Sons, New Jersey, United States.
- GRA Inc, 2007. Economic values for FAA investment and regulatory decisions. Available from: www.faa.gov.
- Hobbs, A., 2005. Latent failures in the hangar. Int. Soc. Air Saf. Invest. For. 38 (30), 11–13.
- Hobbs, A., Williamson, A., 2003. Associations between errors and contributing factors in aircraft maintenance. Human Fact. 45, 186–201.
- IATA, 2019. Safety report 2018, fifty fifth edition. International Air Transport Association, Montreal, Canada.
- Japan Ministry of Transport, 1987. Report on Japan Airlines 747-SR100, Guma Prefecture, August 12, 1985.
- McDonald, N., Corrigan, S., Daly, C., Cromie, S., 2000. Safety management systems and safety culture in aircraft maintenance organisations. Saf. Sci. 34, 151–176.
- Moubray, J., 2001. Reliability-Centered Maintenance. Industrial Press, New York.
- National Transportation Safety Board, 1989. Accident report NTSB/AAR-89/03.
- National Transportation Safety Board, 1990. Accident report NTSB/AAR-90/106.
- Terada, H., Kobayashi, H., 2006. Crash of Japan Airlines B-747 at mt. Osutaka. Japan Science and Technology Agency, Failure Database. Available from: www.shippai.org/fkd/en

Further Reading

FAA, 2014. Operators Manual Human Factors in Aviation Maintenance. Available from: www.faa.gov.
FAA, 2017. Maintenance Human Factors Training. Advisory Circular AC No: 120-72A. Available from: www.faa.gov.
Civil Aviation Safety Authority of Australia, n.d. Safety Behaviors for Engineers. Available from www.casa.gov.au.
Civil Aviation Authority n.d. Aviation Maintenance Human Factors. CAP 716. Available from: <https://publicapps.caa.co.uk>.
Hobbs, 2008. An overview of human factors in aviation maintenance. Australian Transport Safety Bureau, Report AR-2008-055. Available from: www.atsb.gov.au.
Kinnison, H.A., 2004. Aviation maintenance management. McGraw Hill, New York.
Reason, J., Hobbs, A., 2003. Managing Maintenance Error: A Practical Guide. Ashgate, Aldershot.

Airport Security

Richard W. Bloom, Embry-Riddle Aeronautical University, Prescott, AZ, United States

© 2021 Elsevier Ltd. All rights reserved.

Definitions and Goals	34
Laws/Policies/Regulations/Programs/Operating Instructions	34
Threats, Vulnerabilities, and Risks	35
Threats	35
Vulnerability	36
Risk	36
Airport Security Capabilities	37
Airport Security: The Future	39
References	39
Further Reading	39

Definitions and Goals

To understand airport security, let's start with definitions, which in turn, lead to goals and then to how to try and achieve these goals. *Security* commonly refers to deterring and minimizing acts intended to cause human death and injury; non-human destruction and damage; exploitation and the unauthorized modification, access, and transmittal of information, often affecting important processes; and threats to violate deterrence and minimization. Less commonly, security refers to being comfortable with how attempts to deter and minimize are being managed. Without such comfort, there's psychological injury, which in turn, can lead to noxious consequences that are political, economic, social, cultural, and even spiritual. In aviation, such consequences can lead to anything from fewer people wishing to fly to reduced use of airport concessions; less financial profit and more loss for airlines and aircraft producers, parts suppliers, and maintenance organizations and personnel; a smaller aviation-related tax base; and spreading and interactive economic, political, and socio-cultural dislocations from the local to the global.

In defining *security*, it's important to note synergistic, agonistic, and antagonistic effects between it and generic *safety*. Often, but not always, *security* and *safety* only differ in the intent of the act being deterred or minimized. Often, but not always, activities to support one will support the other. Too often, a safety problem even with a solution serves as a "free experiment" for those seeking to harm security, for example, a safety fix indicates how security can be harmed through an "unfix." Less often, a security problem may be resolved in a manner harming safety, and ironically, security, for example, stacked up automobile traffic at an inspection checkpoint to an airport entrance leading drivers to be less safe and secure. Another example of a "free experiment" comprises unintentional violations of security and safety procedures that can serve as inspiration for intentional violations and their ineluctable sequelae.

In defining *security*, it's also important to note that changes in government, business, and academic leadership can lead to different interpretations, reprioritizations, and identifications of the acts to be deterred or minimized. The same applies to what qualifies as acceptable psychological comfort to estimated degrees of deterrence and minimization. As well, to whether deterring and minimizing acts are more or less important than the actual consequences of the acts.

The same applies to changes within the global population and various general publics as to sociocultural values bearing on the differential import of various acts versus consequences, mortality, morbidity, and whether security authorities and those seeking to harm security are operating by their own choice or by factors beyond their control. (This choice versus control distinction affects beliefs in how much one can be protected and, thus, how psychologically comfortable one is in specific situations). Those who seek to harm security will be well ahead of the game, if they understand how a target's sociocultural values yield exploitable opportunities.

More briefly to the definition of *airport*. It commonly refers to a locus of incoming and outgoing aircraft—piloted fixed wing and rotary, and unmanned—with variations on size, management and servicing support, concurrent amenities, and embedding within larger physical, political, economic, and sociocultural contexts. The present analysis is intended to cover all loci of flight, especially, commercial flight. Exceptions, for example, government, military, corporate, other private, and personal aviation, will be noted as appropriate. Another important distinction is that of objective and subjective conceptions of *airport*, especially, as to boundaries. Often enough, there are differences between the legal, the operational, the security-relevant, and the mental constructs and images harbored by personnel. All have impact on what security is, what to do, and what can be done about it.

Laws/Policies/Regulations/Programs/Operating Instructions

There continue to be a patchwork of global, regional, international, national, and local documents—laws, policies, regulations, programs, and operating instructions—affecting aviation and, ineluctably, airports. This patchwork is largely created by

governmental, military, corporate, professional, and other public and private entities. This results in real and potential turf battles, ambiguities, and contradictions and differences of interpretation and application. This impedes optimal assessment, planning, implementation, and re-assessment of aviation and airport security. Specifically, there are problems with how many personnel formally involved in airport security should have a need-to-know, about what they need to know, and what they know so as to be aware of, understand, remember, and comply with and all relevant documents. Documents commonly cited with significant airport security impact—even as they vary in how often they are corrected, rescinded, or otherwise modified—include the International Civil Aviation Organization's (ICAO) *Aviation Security Manual (Doc 8973—Restricted)*, the *China Civil Aviation—Regulations & Security Check*, the *United States' Aviation and Transportation Security Act (ATSA)*, P.L. 107-71, and a myriad of airport-specific operating instructions.

Not only documents, but also definitions and goals within them can change based on new developments in basic and applied research and technology, as well as politics pertaining to and transcending aviation and transportation. These developments can change the meaning of existing words, induce neologisms, and stimulate new ways of thinking about airport security—in other words, the underlying epistemic structures of security knowledge is up for grabs.

One other vehicle of change involves *looping effects* (Hacking, 2006). As an example, security authorities develop concepts, definitions, and estimates of types of people judged more likely to harm security including how they might harm it, for example, combinations of psychological predispositions and behaviors. Airport security policies and programs are based on these types. Paradoxically, those intending to harm security can actually take on aspects of these concepts, definitions, and types—much as how criminals just beginning a life of crime and those already well advanced craft portions of their identities based on what they view in commercial films. It's as if they're being explicitly and implicitly schooled in harming security-by-security authorities. But then their psychological dispositions and behaviors begin to drift—as they develop different and more effective ways to harm security, and violate security authorities' expectations based on the original concepts, definitions, and types. So the security authorities make changes based on new information and interpretations, and an iterative cycle among authorities and those intending to harm security fade in and out of public discourse and the texts of implanting security documents. One consequence of all of this is that different security authorities may end up talking above, below, and around each other and the people with whom they provide protective and consulting services. Another is that the very changes in concepts, definitions, and types have a direct, tangible effect on the world one is trying to protect.

Threats, Vulnerabilities, and Risks

Once definitions, goals, and various implementing documents are considered as an overarching context, it becomes quickly obvious that there are infinite security needs but only finite security resources (money, people, things, and processes) to satisfy these needs. Prioritization of needs and resources is essential, and this occurs through the integration of *threat*, *vulnerability*, and *risk*. Prioritization is also difficult given the interdependence of security with the political, economic, social, and cultural aspects.

Threats

Threat refers to the, who, what, when, how, why, and where of those who seek to impede the deterrence and minimization of acts harmful to security. There are three main classes of *threat*—terrorism, non-ideological crime, and emotional and mental dispositions and disorders. These are pure types, which can often be at least partially linked and mixed for specific people and situations. Accurate threat assessment is most importantly based on intelligence collection and analysis, then the production and secure transmission of intelligence products to security personnel with a need to know. Different security personnel dependent on their responsibilities will have different needs to know and on their psychological capabilities different understandings of what they know (Elias, 2009).

Terrorism. Quite simply, *terrorism* refers to ideologically motivated acts to harm security. *Ideology* refers to a belief system of how the world works, should work, and is working—especially who should get finite resources in a world of infinite needs. Terrorist ideologies usually are religious and/or political in nature—the sacred and the secular. It is essential to note that terrorism is ultimately psychological in nature. Its ultimate purpose is not to destroy, damage, or threaten. These consequences are mere way stations to achieve some ideologically desired behavioral change in the world via psychological impact on those who become aware of the terrorist act and have the ability to make the desired behavioral change. For example, changing the religious and/or political policies and acts of others, inducing certain people judged foreign to leave a domestic areas, and to remove human and material contaminants from the world.

Terrorists are motivated ideologically through their psychologies, and without ultimately focusing on changing the content and attraction of such motivators, counter terrorists' stacking up high body counts may only induce more people to become terrorists and take the place of the dead. The dead then beget the living. Often enough with some religiously and even some secularly based terrorisms, one's own death becomes not a deterrent but an attractor, something to aspire to—for example, desired suicide through counter terrorism policy.

Finally, airports and aviation are lucrative, psychological targets for terrorists. More and more people rely on aviation for work and play—for living. There can be significant noxious effects on economies and the hearts, minds, souls, and behaviors of people from the global to the local. And even in intellectual history, the very act of flying has been interpreted as symbolic of human striving and progress, hopes, and desires—now to be dashed through terrorism. So, in a macabre fashion, airports and aviation are alluring venues to reach out and touch someone via terrorism. This allure also explains why the so-called "lone wolf" terrorist

phenomenon is bogus. A lone individual thinks and feels and imagines within a mind full of people met, people to meet, and people just envisaged and constructed. The lone individual is never alone with the right technology—and with creative thought, this can be quite primitive—some portion of aviation can be brought to its knees along with the sequelae. Perhaps controversially, governments and their military and intelligence entities, businesses, and various non-state actors from the individual through the organizational can be initiators, targets, and observers of terrorism. The airport is one locus ready for action and to be acted upon.

Non-ideological crime. Philosophers and sociologists of crime might argue that all crime is ideological in that it involves violating legal, ethical, and moral proscriptions antithetical to an ideology. But here the *non-ideological* refers to instrumental behaviors leading to material gain. For example, theft, trafficking, and catering to illicit needs of the traveling public—namely, intersections of aviation with other transportation modalities—and people throughout the world. In fact, airports are an alluring criminal venue not just for the people and commodities passing through, but as a locus to satisfy the illicit needs of the population adjacent to airports and of those who will travel to the airport for illicit satisfaction. Non-ideological crime conflates with terrorism, when airport personnel are suborned to engage in behavior to facilitate a terrorist act—not for ideology but for remuneration. An example includes, bribing someone to look the other way or provide information supporting terrorism. So is creating a spot for an individual within a non-ideological criminal enterprise to facilitate terrorist activity. Non-ideological crime can also be constituted by or intersect with corruption, mismanagement, and purposeful hiring of people without appropriate capabilities and motivations that directly or indirectly impact on security (Bunn et al., 2016).

Emotional and mental dispositions and disorders. Instead of the ideological strivings of terrorists and instrumental desires of non-ideological criminals, people may harm security as an expression of combinations of conscious and unconscious psychological conflict. Or through limited intellectual and emotional faculties, they may not understand what they're doing or be easily exploited by others. Or their behavior may be largely influenced by problematic emotional self-regulation and by personality dispositions and traits such as, sensation seeking and those termed the dark tetrad—narcissism, Machiavellianism, psychopathy, and sadism.

When unconscious psychological conflict, certain disorders, and limited psychological faculties are paramount, these people are less likely to engage in sophisticated strategic planning, coherently work with other plotters, or effectively employ any but the most primitive technology. The airport and aviation can become symbols and portals of psychological conflict, at times even psychological containers for troubling thoughts, feelings, and motives that are difficult to manage. Harming security violations become more common when the container leaks or breaks. Even when ineffectual, their attempts to violate security serve as free experiments for terrorists and the non-ideological criminal who can observe how security personnel respond.

Depending on much one is in the throes of sensation seeking and the dark tetrad; one may manifest behavior and effectiveness as if one is primarily a terrorist or non-ideological criminal. Effectiveness often is dependent on indoctrination within religious schools and gatherings, prisons, or the social clubs of gangsters and racketeers.

Finally, there are some members of mental health professions who deny the existence of emotional and mental disorders, but instead label them as sane responses to insane societies. They may be correct, but the threat remains for security personnel. One relevant implication may be that the label of an emotional or mental disorder applied to an individual may provoke the intent to harm security. Another is that national and local mental health and social services policies, for example, community versus institutional care and inevitable shortfalls—can affect intensity and frequency of this threat.

Vulnerability

Vulnerability simply refers to everything that can be wittingly or unwittingly exploited by those who seek to harm security. At issue are *people, things, process, and money*. *People* examples include, the motivations, abilities, knowledge, and personalities of security personnel. *Thing* examples include, the density of walls, the fidelity of explosive detection or remote sensing technology, the weapons carried or un-carried by security personnel, and the situatedness of an airport as to topography, topology, and adjacent populations. *Process* examples include, the take-off and landing routes of airplanes; ticketing and security rules of engagement; and human resource management procedures including education, training, and recognition programs for security personnel. (*Process* actually encompasses how any activity is structured, what it's supposed to do, and how it does it—all potentially exploitable to harm security).

The amount of *money* dedicated to support security and anything affecting security—as well as security expertise, politics, social and cultural factors (including the security cultures at an airport and organizations formally and informally impacting on airports), and resulting budget priorities often drive the potential for exploitation by those seeking to harm security. Everything identified as vulnerability, should be prioritized as to impact on security if exploited in a particular manner.

Risk

Risk refers to the matching up of *threat* to *vulnerability*. Significant *vulnerabilities* might be obviated, if the likelihood of *threat* based on intelligence collection and analysis is low. Less significant *vulnerabilities* might be strongly attended to, if the *threat* is deemed significantly high. Ultimately, these are political questions, sometime but not always informed by security expertise. This is because *risk* cannot be fully met even if all existing resources were applied to it. And, obviously, airport and aviation security is not the only game in town—cf. education, health, and various material needs of population; shareholder needs of private industry; other security needs of intelligence, security, and law enforcement organizations.

Besides attempts to assess the ideological attractiveness of airports and aviation for those who seek to harm security described in *Terrorism*, and the economic attractiveness mentioned in *Non-political crime*, quite difficult to identify is what psychologists call the *Icarus complex* and which is relevant to all three categories of *threat* including that of *Emotional and mental dispositions and disorder*. The origins of the complex's name go back to the ancient Greek myth of Daedalus and Icarus, wherein the latter flew too close to the sun and tumbled to his demise as the wax securing his wings melted. The intrinsic attractions of airports and aviation in the eyes of various bad actors both elicits motivation to harm security and can cause overshooting what's possible for an impossible dream.

Airport Security Capabilities

Layers of security. An integrated layering of money, people, things, and process constitute optimal airport security. The ideal goal is to continuously change layers as calculations of threats, vulnerabilities, and risks continuously change. What follows are common examples. One important conclusion will be that the layers should be both offensive and defensive, —for example, information operations to decrease the number of people who intend to harm security or support those who do; proactive and pre-emptive apprehension of bad actors physically distant from an airport; hardening of aircraft and airport infrastructure against explosive harm; fine-tuning of weapons, weapons components, and explosive trace detection within the airport; upping the validity of various data mining and behavioral detection programs; and, again, information operations, here to lessen the likelihood that the ultimate psychological targets of terrorism will, succumb to behavioral change—and concurrent information operations to promulgate this to the world.

Intelligence. By far the most important airport security capability comprises robust intelligence activities. As previously described, collection and analysis, and then production and transmittal, of relevant information can often elicit relevant *threat* data to be integrated with *vulnerabilities* and then yield a more useful risk assessment. In turn, airport security authorities and personnel can more likely optimize the effectiveness of layers of security. This information also can facilitate identification of actual or incipient bad actors and their support structures (Fox and Farmington, 2018). People seeking to harm security along with their supporters can be impeded or seized, and then detained, interrogated, adjudicated, and imprisoned, if appropriate.

As well, covert and clandestine activities—especially those with counterintelligence significance—can penetrate networks, shape behavior away from harming security, and even foster and guide conflict among those intending to harm security. These activities also can influence planning so that the various threats will be less likely to be effective and more likely to be matched with vulnerabilities that mitigate harmful consequences.

Overt, covert, and clandestine information operations can be crafted specific to various societies, cultures, and subpopulations to decrease the desire to harm security including airport security. The essence of this would be reinforcing more and more people with a coherent identity, a narrative to believe in, and social networks obviating the need for harming security.

More controversially, information operations can target general publics, so that their psychological reactions of fear, panic, dismay, and inclination to pressure their governments to do the wrong things are minimized. This last especially gets at the lifeblood of terrorism. If information operations by intelligence personnel against a domestic population by its government are proscribed, the role should fall to government-business-academic-civic-communications media partnerships overtly explaining the ultimate psychological purposes of terrorists, and what general publics can do not to become unwitting psychological victims. Other intelligence activities can help ameliorate political, economic, social, and cultural realities that help fuel the three major threats to airport and aviation security.

Effective intelligence activities can help prevent harm to security many miles and time zones from a specific airport. But as with all airport security capabilities and all threats to security, the best that can be done is to reduce the risk not do away with it. Like all of the oldest professions in human history, there probably are significant evolutionary phenomena impelling threat and impeding its disappearance from the human behavioral repertoire.

Planning and designing airports. As already alluded to in *vulnerability*, airport security can become much easier or much more difficult depending on where the airport is situated. Not just terrain topography, but what's adjacent to airports as to buildings, people, and everyday life. The ease or difficulty of airport security also depends on anything from the planning and design of parking, concessions, ticketing, interior layouts, runways, other infrastructure, and airport processes directly or indirectly supporting aviation. There's huge variability—often based on when, why, and how an airport was initially conceived—on how much airport security is a significant factor in planning and design or just an afterthought. The challenges of security as afterthought usually become salient after significant attempts to harm airport security and well-publicized attempts elsewhere to harm various transportation modalities and other public and even private venues.

Impediments to security-appropriate airport planning and design potentially include extra costs, trade-offs of security versus traveler comfort and flight efficiency, and comparative values as to human life. These can be anticipated and often resolved, if there's a will to concurrently manage multiple concerns. What can't always be anticipated are specific changes impacting the airport, for example, social and cultural changes from massive ingress and egress of populations; significant changes to the ratio of commercial and residential populations and the infrastructures of urbanization and de-urbanization; and new technologies that pose new security challenges including voiding what once was protection afforded by physical and natural barriers. Even the long-studied challenge of aviation-related noise pollution can have security impact through inadequate systemic analysis of trade-offs in comparative mitigation.

Surveillance and reconnaissance adjacent to and over the airport. Human assets and technical assets need to be deployed and employed to deter and minimize attacks from off-airport launched missiles; radiological and explosive weapons planted on adjacent roadways and incoming vehicles; off-airport staging areas for human attacks usually mediated with weapons; fixed wing, rotary, and unmanned aerial systems threatening aircraft on the ground or as they approach take-off or landing, as well as infrastructure, personnel, and travelers; even the surveillance and reconnaissance capabilities of those seeking to harm security. Another important technique is placing restrictions on adjacent ground areas and establishing non-fly restriction zones, as well as regulating and enforcing what can be in the air and the minimum distances between airborne aircraft. A more recent *threat* is that of unmanned aerial systems (UAS), for example, drones, as to damaging engines or other infrastructural aspects of aircraft as well as aircraft crew and passengers, and airport staff and travelers through direct kinetic impact. UAS threat also can lead to security decisions to shut down flight operations and to induce formal and informal changes in the flight paths of other aircraft. Problems in countering the UAS *threat* include devising techniques that do not pose an unacceptable risk to non-threatening people, things, and processes. The UAS *threat* also can constitute a platform for weapons employment and hostile cyber acts (see *Cyber security* later). UAS military and civilian proliferation is in such a drastic state of flux that all facets of airport infrastructure and operations are continuously being re-evaluated as to safety and security.

"Hardening" of aircraft. More against threats from terrorism and emotional and mental dispositions and disorder via decreasing *vulnerability*, attempts continue to modify and create aircraft designs, structures, materiel, and technologies to withstand explosions, crashes, and human violent behavior (see *Cyber security* later). Initiatives focus on increasing the difficulty of opening and going through cockpit doors, negotiating secondary flight deck barriers, otherwise accessing the flight deck, and preventing or distorting cabin-cockpit-air traffic control communications. In addition, there continue to be upgrades in the fidelity, real-time transmissibility, and crashworthiness of audio, video, and technical data recording. More controversial have been the utility of on-board armed security like air marshals and law enforcement; self-defense training and/or arming flight crew and attendants; and standardized, scenario-dependent education and training of passengers for scenario-dependent engagement. This latter would "aviation-ize" the common alternatives for an on-the-ground mass shooting, for example, run, hide, or fight might become do nothing, crouch in your seat, or variously engage.

Surveillance and reconnaissance at the airport. The two main kinds of capability are screening people and screening things for threat indicators. Both have problems with unacceptably high combinations of true and false positives (with positives indicating an actionable threat) and negatives (with negatives not indicating such a threat) (cf. [Harcourt, 2007](#)). Inadequate attempts to identify accurate error rates for screening occur because of policy decisions not to engage in applied research, the use of lab as opposed to field research, and field research grossly unrealistic to the environment of a daily operational environment. Also, error rates are often partially reported in a manner occluding significant shortfalls. For example, one can easily have a 100% accurate true positive rate by allowing a huge false positive rate—namely, identifying everyone as an actionable threat. Also, when applied research is accomplished, statistical analyses often are inadequate. One example of this is not allowing for the often very low base rates of attempts to harm security. Another is not explicitly making assumptions as to whether one assumes to be sampling different facets of the same population or a series of different population. Finally, statistics are chosen and interpreted without actual balancing of the priorities for large effect sizes and the likelihood of identifying true positives or true negatives. This has an impact on human, other animal, and technical screening capabilities from observation, smelling, frisking, interviewing, and various technical analytic backups—namely, algorithms programmed to spot threat indicators worthy of security action. A so far insurmountable challenge is that threat indicators can be exhibited by non-threatening people and things, no threat indicators from an actual *threat*. Finally, a common recommendation is to employ a "weak" profiling—combinations of universal screening for certain threat indicators and random sampling for other indicators.

Screening people. Whether for aircraft crew, airline personnel, airport staff, or travelers, this should begin with various data mining approaches beginning before anyone arrives at the airport. Data mining problems include incomplete, ambiguous, and partially contradictory repositories of information suggesting threat indicators; the degree of direct, indirect, and inferential linkages between threat indicators and harming security; and the transience of reliable and valid linkages ([Bloom, 2009](#)).

Once at the airport, screening entails the use of various verbal and nonverbal criteria, almost all with suspect linkage with threat. The shortfall is founded on the applied research problems described earlier. Some recent work on differential verbal response while engaged in a question answer dialogue with a trained interlocutor may show some operational promise. It is unfortunate that the quite significant body of applied psychological research on terrorism, non-political crime, and emotional and mental dispositions and disorders does not support strict use of profiles without significant error rates. It's a political decision how acceptable an estimated false positive error rate can be to secure an estimated true positive rate. Even for biometrics and facial recognition technologies, there are issues. Biometric data often facilitate validating an individual's identity, but not their intent. Although iterations of facial recognition technologies continue to resolve the challenges of ambient lighting and shading, body movement, and micro-facial expressions, the identity-intent distinction applies. However, biometrics—including facial recognition, DNA sampling, and recognition of palm prints, hand geometry, iris and retina, and odor/scent—can function as deterrent and minimization of *threat* and provide useful data for investigations and analyses after security has been harmed. And yet their value can be mitigated through having access to incomplete, poorly categorized and constructed, and difficult to access comparative data-bases.

Screening things. Here the two most highly salient *threats* bear on baggage (both carry-on and checked) and cargo.

As to baggage, whatever combinations of human, other animal, or technical approaches are employed, one problem is matching the physical and psychological capabilities of the screener with the telling physical characteristics of the baggage (and cargo). It's

fairly easy for the more sophisticated among those who intend to harm security to learn these capabilities and successfully plan by creating a mismatch with characteristics of weapons, weapon's components, explosives, chemical, biological, and radiological agents, and so forth (Press, 2008).

As to cargo, a salient issue is the often long and discontinuous chain of custody (admittedly, something that could occur with baggage as well) from point of production to final destination. Much as upscale gambling establishments are often a font of inspiration for profilers of proscribed behaviors, so are those who seek to impede multinational illicit trafficking in drugs, weapons, other commodities, and people. A recent prosecution of a famous drug trafficker yielded deceptive techniques to move cocaine varying from cans of chili to shoe boxes to tanker ships.

Cyber security. Cyber acts defined as comprising the production, storing and transmitting of information via overlapping entities of computers, information technology, and virtual reality pose an increasing threat to airport and aviation security, but also to those who seek to harm it (Zaccaro, 2016). As with all yoked enterprises of crime/counter-crime, spy/counter-spy, there's a continuing escalation in both how to reinforce security and harm it. At issue are accessing, transmitting, modifying, and impeding information with and without appropriate authorization. These can be employed to harm people, things, and processes and to steal financial assets otherwise available for security. There seems to be catastrophizing as to a comprehensive identification of threats and vulnerabilities, especially as to taking control of airborne aircraft and planting hardware/software elements to induce disaster. The two sources of *threat* and *vulnerability* in cyber security are technological and human. The human—especially the insider *threat*—is often discounted or viewed as a secondary concern. Other issues of concern include human and technical information accuracy, adequacy, fidelity, and timeliness within, between, and among air traffic management, aircraft, and airport personnel. Shortfalls can result in crises, critical incidents, and unsatisfactory management of the same.

Airport Security: The Future

Optimal testing of layers of security should increasingly comprise offensive and defensive perspectives—often called red team/blue team exercises. Layers of security may vary in looking predictable and unpredictable depending on intelligence about expectations of the *threat* and *threat-vulnerability* matches prioritized for probability of occurrence and likely impact. Significant traditional, mass, and social media descriptions and reporting on aviation and airport disasters are likely to continue to increase the attractiveness of flight as target by terrorists and those with emotional and mental dispositions and disorders, less so with non-political criminals. Thus, global, regional, national, and local policies about the communicating harms to security need to be re-assessed balancing need-to-know and need-not-to-be-harmed. Applied research needs to be even more sensitive to tradeoffs in assumptions about the realism of field experiments and about statistical assumptions used to make inferences—common labels for the latter options include Bayesian, fiduciary, frequentist, and (decision) theoretical. And the cyber security challenge still has too many unknown unknowns and—along with more efficacious intelligence activities and information operations—remains the highest priority for more resources.

Ironically, with all the technological challenges, airport security is ultimately a human problem. The irony stems from increasing observations that human nature itself may be changing in a globalizing, cyberworld (Zuboff, 2019). We don't know what this means for airport security—as to future threats, vulnerabilities, and risks. But we will speculate that many capable men and women will take up the challenge—to protect the flying public, to exploit it, and to destroy it.

References

- Bloom, R.W., 2009. A government primer: what good people should know about detecting bad people. *Government Security News*. Available from: <http://www.gsnmagazine.com>.
- Bunn, M., Sagan, S.D. (Eds.), 2016. *Insider Threats*. Cornell University Press, Ithaca, NY.
- Elias, B., 2009. *Airport and Aviation Security: U.S. Policy and Strategy in the Age of Global Terrorism*. CRC Press, Boca Raton, FL.
- Fox, B., Farrington, D.P., 2018. What have we learned from offender profiling? A systematic review and meta-analysis of 40 years of research. *Psychol. Bull.* 144 (12), 1247–1274.
- Hacking, I., 2006. Making up people. *London Rev. Books* 28 (16), 23–26.
- Harcourt, B.E., 2007. *Against Prediction: Profiling, Policing and Punishing in an Actuarial Age*. University of Chicago Press, Chicago, IL.
- Press, W.H., 2008. Strong profiling is not mathematically optimal for discovering rare malfeasors. *Proceedings of the National Academy of Sciences*. Available from: <http://www.pnas.org/content/106/6/1716.full>.
- Zaccaro, S.J. (Ed.), 2016. *Psychosocial Dynamics of Cyber Security*. Routledge, New York.
- Zuboff, S., 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, United Affairs, New York.

Further Reading

The MITRE Corporation, 2009. *Rare Events*. The MITRE Corporation: JASON Program Office, McLean, VA.

Transport Safety and Security: Alcohol

James C. Fell, NORC at the University of Chicago, Bethesda, MD, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	40
Alcohol Impairment and Blood Alcohol Concentration (BAC)	40
Relative Risk of Being Involved in a Crash by BAC	40
Setting BAC Limits for Driving	42
Zero-Tolerance Laws for Underage Drivers	42
Alcohol Regulation	42
Minimum Legal Drinking Age (MLDA) Laws	42
Driving Regulation	43
Graduated Driver Licensing for Young Drivers	43
Drinking and Driving Legislation	44
Impaired-Driving Laws	44
Administrative License Revocation/Suspension (ALR/ALS)	44
Mandating Alcohol Ignition Interlocks for all Convicted DUI Offenders	45
Enforcement of Impaired Driving Laws	45
Sobriety Checkpoints and Random Breath Testing (RBT)	45
Preventing Impaired Driving Offenders From Repeating	46
DUI/Drug Courts	46
Screening and Brief Interventions	47
Vehicle Sanction Laws	47
Alcohol Monitoring	47
Alcohol Consumption	48
Alcohol Ignition Interlocks	48
Conclusion	49
The Future	49
Driver Alcohol Detection System for Safety (DADSS) Program	49
Autonomous Vehicles	49
Acknowledgments	50
See Also	50
References	50
Further Reading	51
NHTSA Alcohol and Highway Safety Reports	51
Other Comprehensive Reviews	51

Introduction

Alcohol Impairment and Blood Alcohol Concentration (BAC)

Laboratory studies indicate that impairment in critical driving functions begins at low blood alcohol concentrations (BAC). Most subjects in laboratory studies are significantly impaired regarding visual acuity, vigilance, drowsiness, psychomotor skills, and information processing by the time they reach .05 g/dL BAC compared to their performance at .00 g/dL BAC. A review of the scientific literature on the impairment of driving-related skills at low BACs based on laboratory testing of alcohol-dosed subjects documented that impairment of driving-related skills starts at very low BACs, for some as low as .02 g/dL BAC. A review of 177 studies ([Moskowitz and Robinson, 1988](#)) clearly documented significant impairment at .05 g/dL BAC and higher, and a review of 112 more recent studies in 2000 provided even stronger evidence of impairment at .05 g/dL BAC ([Moskowitz and Fiorentino, 2000](#)). Together, these two reviews have summarized the findings of nearly 300 studies of impairment at low-BAC levels, and the findings are remarkably consistent.

Relative Risk of Being Involved in a Crash by BAC

It has been recognized that alcohol is a factor in crashes since the beginning of the 20th century. However, the mere association of alcohol with crash involvement does not demonstrate a causal influence unless all other relevant factors can be eliminated. More direct evidence of a causal relationship is provided by demonstrating that crash probability is directly related to the amount of

alcohol in the driver's body. *Relative risk* studies attempt to refine the influence of alcohol on crashes by comparing drivers in crashes with similar drivers using the road at the times and places where crashes have occurred. Such studies are titled "case-control" investigations because, for each crash-involved driver (case), one or more non-crash-involved drivers (control) are selected for comparison. The key to such studies is to ensure that the comparison driver is selected so that the only distinction between the two is the involvement in a crash. This is achieved by going to the same location where the crash occurred at the same time of day on the same day of the week and randomly selecting comparison drivers (non-crash-involved drivers) operating vehicles in the same direction on the road as the crash-involved driver. Although nine studies of this type have been conducted over the last 80 years, in practice this level of control has rarely been achieved.

The most influential of the case-control studies was conducted by Robert Borkenstein in Grand Rapids, Michigan, in which 5985 crash-involved drivers were compared with 7590 control drivers (Borkenstein et al., 1974). Between July 1952 and July 1953, research staff members traveled to the sites of crashes occurring between 6:30 p.m. and 10:30 p.m. on Monday through Saturday to collect breath samples from crash-involved drivers. Later, they collected data on four comparison drivers at each site where crashes had occurred in the previous 3 years, not the same sites as the locations at which they collected the crash driver cases. Thus, the study was not strictly a case-control study because the comparison cases were not matched to the drivers in the crash cases. Rather, comparison drivers were matched to drivers in crashes randomly sampled from police accident reports occurring during the previous 3 years. Further, no correction was made for drivers who refused to provide breath tests. Finally, the group of crash-involved drivers who were compared to the control group was modified to represent drivers judged to be "at fault" by a process of combining those in single-vehicle crashes with half of those in two-car crashes. Despite these limitations, the Grand Rapids study became the gold standard for estimating the relative risk of involvement in an alcohol-related crash based on the driver's BAC.

More recently, other case-control studies have been conducted in Huntsville, Alabama, in 1977 involving 650 drivers in injury crashes and in Adelaide, Australia, in 1980 involving 299 drivers in injury crashes. Both of those studies, like the earlier ones just described, developed relative risk curves, which showed rapid rises at BACs higher than .05 g/dL.

Because other variables as age, gender, and quantity and frequency of drinking can vary between the drivers in crashes and the comparison drivers, it is necessary to compensate for these differences by the use of covariates in computing the risk curves. A study in 2005 collected information on a substantial set of demographic variables including items, such as marital status, education, ethnicity, employment status, and vehicle type as well as age and gender for use in correcting the differences between the crash and comparison populations of drivers (Blomberg et al., 2005). Logistic regression analyses were used to tease out variables which were most important in distinguishing crash and control drivers and those that proved to be significant were used as covariates in adjusting the risk curves. The significance of adjusting those differences in calculating the risk curve is shown in Fig. 1.

That study found a statistically significant relative risk of a crash at BAC $\geq .04$ g/dL and increasing exponentially at BACs $\geq .10$ g/dL and higher. There is no evidence that lower levels of alcohol consumption actually reduce or increase the crash risk (e.g., BACs less than .04 g/dL).

A recent case-control study was conducted in Virginia Beach, Virginia, USA, during a 20 month period in 2010–11 to estimate the risk of crashes involving drivers using drugs, alcohol, or both (Compton and Berning, 2015). Biological measures were obtained on more than 3000 crash drivers at the scenes of the crashes, and 6000 control (comparison) drivers. Control drivers were recruited 1 week after the crashes at the same time, day of week, location, and direction of travel as the crash-involved drivers. Data included over 10,000 breath samples, 9,000 oral fluid samples, and 1,800 blood samples. Oral fluid and blood samples were confirmed for the presence of alcohol and drugs. The crash risk associated with alcohol and other drugs was estimated using odds ratios that indicate the probability of a crash occurring over the probability that such an event does not occur. For example, if a variable, such as alcohol

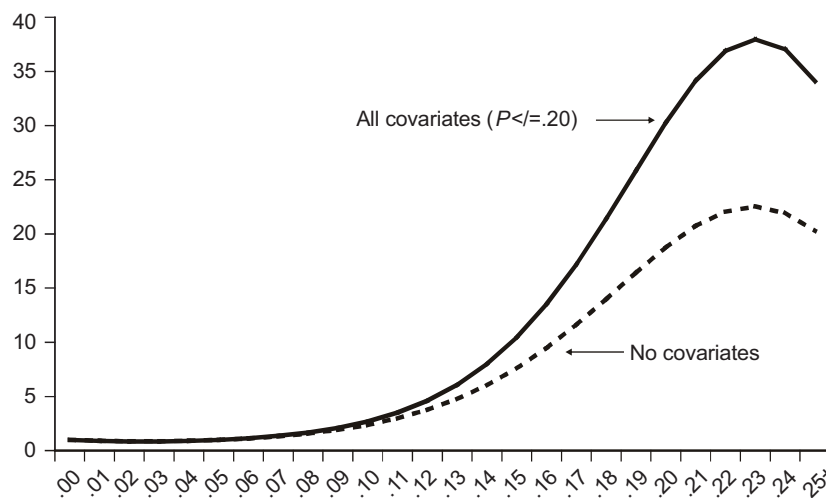


Figure 1 Relative risk of a crash by BAC with covariates and without covariates. Source: Blomberg, R.D., Peck, R.C., Moskowitz, H., Burns, M., Fiorentino, D., 2005. Crash risk of alcohol involved driving: a case-control study, final report to the National Highway Traffic Safety Administration. Dunlap and Associates, Inc., Stamford, CT.

and/or drugs is not associated with a crash, then the odds ratio for that variable will be 1.00. A higher or lower number indicates a relationship between the probability of a crash occurring and the presence of that variable. Confidence intervals (CIs) of an odds ratio indicate the range in which the true value lies—with 95% confidence. Alcohol was the largest contributor to crash risk. The crash risk estimates for alcohol indicated drivers with breath alcohol concentrations of .05 g/210 L were 2.05 times more likely to crash than drivers with BACs=.00 (no alcohol). For drivers with BACs of .08 g/210 L, the relative risk of crashing was 3.98 times that of drivers with .00 BACs.

Setting BAC Limits for Driving

A general review of studies of the effect of lowering the BAC limit in Australia, Europe, and the United States was conducted in 2006 and concluded that lowering the BAC limit to .05 would be likely to reduce alcohol-related crashes in the United States. A study in 1994 found that lowering the BAC limit from .08 to .05 g/dL in New South Wales, Australia, significantly reduced fatal crashes on Saturday by 13%. Another Australian study conducted in 1995 employed a rigorous time-series analysis of random breath testing (RBT) and .05 BAC laws in Australia, controlling for many factors including seasonal effects, weather, economic trends, road use, alcohol consumption, and day of the week. Although the primary focus of the Australian study was the impact of RBT, the findings on the effect of .05 BAC laws were also significant.

There is a strong case for setting a .05 BAC limit for driving. Laboratory and test track research shows that the vast majority of drivers, even experienced drinkers who typically reach BACs of .15 or greater, are impaired at .05 BAC and higher with regard to critical driving tasks. There are significant decrements in performance in areas, such as braking, steering, lane changing, judgment, and divided attention at .05 BAC. Some studies report that performance decrements in some of these tasks are as high as 30%–50% at .05 BAC. In addition, as shown in the previous section, the risk of being involved in a crash increases significantly at .05 BAC. The risk of being involved in a crash increases at each positive BAC level, but rises very rapidly after a driver reaches or exceeds .05 BAC compared to drivers with no alcohol in their blood systems.

Lowering the illegal per se limit to .05 BAC is a proven effective countermeasure, which has reduced alcohol-related traffic fatalities in several countries, most notably, Australia. While studies in Europe and Australia, each use a different methodology to evaluate these effects, the evidence is consistent and persuasive that fatal and injury crashes involving drinking drivers decrease on the order of at least 5%–8% and up to 18% after a country lowers their illegal BAC limit from .08 to .05 BAC. A 2017 meta-analysis of international studies on lowering the BAC limit in general found an 11.1% decline in fatal alcohol-related crashes from lowering the BAC to .05 or lower (Fell and Scherer, 2017). The study estimated that 1790 lives would be saved each year, if all states in the United States adopted a .05 BAC limit.

Zero-Tolerance Laws for Underage Drivers

By 1988, all states in the United States had enacted minimum legal drinking age (MLDA) laws making it illegal for those younger than 21 to purchase or possess alcohol. This provided a basis for implementing a zero BAC limit for drivers aged 20 and younger (generally defined as a BAC of .02 or greater). These zero-tolerance laws for youth have proven effective in reducing fatal crashes involving underage drinking drivers. A meta-analysis conducted in 2001 of the studies of zero-tolerance laws found reductions of 9%–24% in fatal crashes (Shults et al., 2001).

Alcohol Regulation

Minimum Legal Drinking Age (MLDA) Laws

Stimulated by the scientific and safety advocate support for limiting underage access to alcohol, a basic set of at least 20 laws directed at (1) control of furnishing and selling alcohol to youth, (2) possession and consumption of alcohol by youth, and (3) prevention of impaired driving by those aged 20 and younger have been adopted in the United States over the last 3 decades in many of the 50 states and the District of Columbia (DC). Evidence exists that such laws can influence underage alcohol-related traffic fatalities. From 1988—when all states had enacted MLDA-21 legislation—to 1995, alcohol-related traffic fatalities for youth aged 15–20 declined from 4187 to 2212, a 47% decrease, with wide variability in these declines between states.

Numerous studies, including a comprehensive review of literature from 1960 to 1999 (Wagenaar and Toomey, 2002), uniformly showed that increasing the minimum drinking age significantly decreased self-reported drinking by young people, the number of fatal traffic crashes, and the number of arrests for driving under the influence (DUI) involving youths aged 20 and younger. In 2001, a meta-analysis of 33 studies of the MLDA was conducted. That study showed that the MLDA resulted in changes of 10%–16% in fatal crashes: increasing fatal crashes, if the MLDA was lowered, and decreasing fatal crashes, if it was raised. In 2006, a study found that when New Zealand lowered its drinking age from 20 to 18, crash injuries among 15–19-year-olds increased.

Recent legal research involving the use of the Alcohol Policy Information System (APIS) developed under a grant from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) has indicated that there are at least 20 MLDA-21 laws that have been adopted at the state level in the United States. **Table 1** contains a brief summary of those 20 laws; in parentheses and in bold is the number of states (including DC) that have currently adopted each law (Fell et al., 2015).

Table 1 Minimum legal drinking age 21 (MLDA-21) law components and descriptions

MLDA-21 law components	Description
Core laws that apply to youth	
1. Possession	Illegal for youth under age 21 to possess alcohol (50 states + DC)
2. Purchase	Illegal for youth under age 21 to purchase or attempt to purchase alcohol (47 states + DC)
Expanded laws that apply to youth	
3. Consumption	Illegal for youth under age 21 to consume alcohol (34 states + DC)
4. Internal possession	Evidence of possession and consumption via a BAC test (9 states)
5. Use and lose	Alcohol citation for youth under age 21 results in drivers' license suspension (39 states + DC)
6. Use of fake identification	Fake ID minor—illegal for youth under age 21 to use a fake identification to purchase alcohol (50 states + DC)
Apply to youth driving	
7. Zero tolerance	ZT—illegal for a driver under age 21 to have any alcohol in their system when driving (50 states + DC)
8. Graduated driver licensing with night restrictions	GDL—youth with intermediate or provisional license prohibited from driving without an adult in the vehicle past a certain hour at night (50 states + DC)
Apply to providers	
9. Furnishing or selling	Illegal to furnish or sell alcohol to a youth under age 21 (50 states + DC)
10. Age of on-premise servers	Minimum age 21 set for selling/serving alcohol (13 states)
11. Age of on-premise bartenders	Minimum age 21 for bartenders (23 states + DC)
12. Age of off-premise sellers	Minimum age 21 set for selling/serving alcohol (23 states)
13. Keg registration	Identification number for beer keg and purchaser required (30 states + DC)
14. Responsible beverage service training	RBS—responsible beverage training mandatory or voluntary (37 states + DC)
15. Retailer support provisions for fake identification	Fake ID retailer—provisions to assist retailers in avoiding sales to youth under age 21 (45 states)
16. Social host prohibition	SHP—prohibits social hosting of underage drinking parties (28 states)
17. Dram shop liability	Action against commercial provider of alcohol (44 states + DC)
18. Social host civil liability	Action against non-commercial (private) provider of alcohol (33 states)
Apply to manufacturers or suppliers of fake identification	
19. Transfer/production of fake identification	Fake ID supplier—prohibits manufacturing and/or supplying fake identification to youth for the purposes of buying alcohol (24 states)
Apply to states concerning control of alcohol distribution	
20. State control of alcohol sales	State control—a state-run retail distribution system of alcoholic beverages (i.e., beer, wine spirits) (11 states)

Source: *Fell et al., 2015*

The 20 MLDA-21 laws used in a 2016 study were scored in a prior study for their strengths and weaknesses based on (1) the sanctions enacted for violating the law, (2) any exceptions or exemptions affecting the application and enforcement of the law, and (3) any provisions that could affect the law or its enforcement negatively or positively. The strength scores of these laws were used to assess the impact of 20 MLDA-21 laws on underage alcohol-related outcomes, including underage drinking-and-driving fatal crashes. The nine MLDA laws associated with significant decreases in fatal crash ratios of underage drinking drivers were: possession of alcohol (−7.7%), purchase of alcohol (−4.2%), use alcohol and lose your license (−7.9%), zero tolerance .02 BAC limit for underage drivers (−2.9%), age of bartender ≥21 (−4.1%), state responsible beverage service program (−3.8%), fake identification support provisions for retailers (−11.9%), dram shop liability (−2.5%), and social host civil liability (−1.7%) (*Fell et al., 2016*).

The nine laws are estimated to be currently saving approximately 1135 lives annually in the United States (*Fig. 2*). The researchers estimated that if all states adopted them, an additional 210 lives could be saved each year. Only five states have all nine effective laws: California, Colorado, New Mexico, Utah, and Wyoming.

Driving Regulation

Graduated Driver Licensing for Young Drivers

Research has shown that the first few months of licensure for young novice drivers entail the highest crash risk. To address this issue, many countries have adopted graduated driver licensing (GDL) laws that require a staged progression to full license privileges. GDL laws generally require three-staged licensing for novice drivers: (1) a learner's permit period—practice driving with a licensed driver, (2) an intermediate or provisional stage—drive solo only under certain conditions (e.g., restricts late-night driving and limits teen passengers); and (3) a full license with no restrictions (minimum age of 18 in most countries). The young driver must meet certain requirements to “graduate” to each stage. Nighttime restrictions have been found to reduce novice driver involvements in nighttime fatal crashes by an estimated 10% and young drinking drivers in nighttime fatal crashes by 13%. Passenger restrictions have been found to reduce novice driver involvements in fatal crashes with teen passengers by an estimated 9%. These results confirm the effectiveness of these provisions in GDL systems.

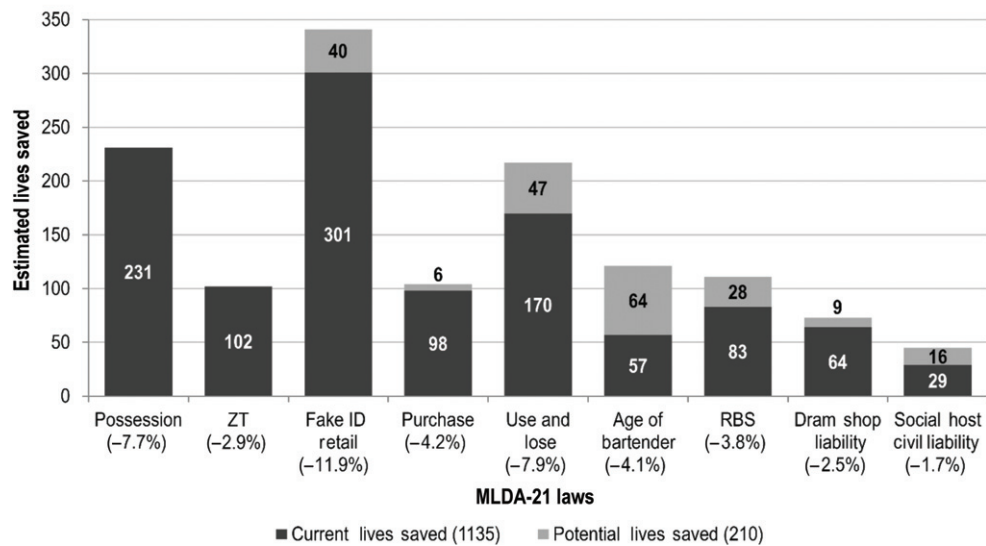


Figure 2 Estimated lives saved by nine MLDA-21 laws. Source: Fell, J.C., Scherer, M., Thomas, S., Voas, R.B., 2016. Assessing the impact of twenty underage drinking laws. *J. Stud. Alcohol Drugs* 77, 249–260.

Drinking and Driving Legislation

Impaired-Driving Laws

Over the last quarter century, much of the research in traffic safety has focused on the evaluation of proposed new laws and the development of tools for use by the police and courts in the enforcement of DUI laws. The most significant laws are listed in [Table 2](#), and recent research on key DUI laws is described in the following sections.

Administrative License Revocation/Suspension (ALR/ALS)

ALR/ALS is a proven effective sanction. ALR laws, which provide that the license of a driver with a BAC at or over the illegal limit is subject to an immediate driver's license suspension by the state or country agency in charge of driver licensing and motor vehicles, are the most widely applied example of a traffic law where the sanction rapidly follows the offense. The power of ALR laws has generally been attributed to how swiftly and how certain the sanction is applied. This licensing control measure also rapidly removes a high-risk driver from the public roadways. Several large national studies have demonstrated that the presence of an ALR law is associated with a reduction in alcohol-related crashes.

Table 2 Key DUI laws

DUI law	Description
Impaired-driving law	The traditional law against drinking and driving that depends on the officer presenting observations of the suspect's behavior that demonstrates the driver was under the influence of alcohol. It does not depend on the presentation of a BAC test result, but if the BAC level is available; it becomes presumptive evidence of intoxication.
Per se law	Operating a vehicle with a BAC at or higher than a specified illegal limit. When a BAC level is not available, the prosecution must be conducted under the impaired-driving law.
Zero-tolerance law (a "status" law based on age)	Based on the minimum legal drinking age law, the zero-tolerance law makes it an offense for a person younger than age 21 to operate a motor vehicle with any measurable amount of alcohol in their bodies; generally specified as a BAC of .02 or greater.
Implied-consent law	Establishes that, by accepting a driver's license, a driver agrees to submit to a chemical test if an officer has probable cause to make an arrest for impaired driving. In the event of a refusal, most statutes provide that no test will be given, but the suspect's license will be suspended. Alternatively, the officer can seek a warrant to require a blood test.
Administrative license revocation law	Provides that a DUI suspect whose chemical test result is higher than the legal limit (currently .08) is subject to an immediate license suspension. A provision must be made for a hearing, if requested by the offender.
High BAC laws	Provides for increased sanctions for a convicted DUI offender whose BAC level is .15 or greater.
Test refusal law	Provides for increased sanctions for DUI suspects who refuse the chemical test.
Anti-consumption, open container law	Prohibits consumption of alcohol by the driver while in charge of a vehicle or more commonly prohibits an open alcohol container in the passenger compartment of a vehicle. Passengers in commercial vehicles, such as buses, are normally exempted.

Mandating Alcohol Ignition Interlocks for all Convicted DUI Offenders

Installing alcohol ignition interlocks has been shown to have both a general deterrent and specific deterrent factor. Alcohol ignition interlocks require a negative breath test to start a vehicle's engine, and some countries have mandated some form of interlock law for drivers convicted of driving while intoxicated (DWI). A study in the United States showed that all-offender laws were associated with 16% fewer drivers with .08+ BAC involved in fatal crashes, compared to no interlock law (Teoh et al., 2018). Repeat-offender interlock laws were associated with a nonsignificant 3% reduction in impaired drivers, compared to no law. Repeat and high-BAC interlock laws were associated with an 8% reduction in impaired drivers in fatal crashes, compared to no law. Laws mandating alcohol ignition interlocks, especially those covering all offenders, are an effective impaired driving countermeasure that reduces the number of impaired drivers in fatal crashes.

Enforcement of Impaired Driving Laws

The effectiveness of traffic laws related to impaired driving depends on deterring vehicle operators from driving after drinking. The principal factor in creating deterrence is how effectively the laws are enforced. Classical deterrence theory is a psychological model designed to explain the influence of punishment on personal behavior. It holds that three factors—risk of detection, severity of the sanction, and the speed with which the sanction is applied—determine the response to laws. Deterrence is the perception of each of the three factors—rather than the actual risk of detection, the sanction severity, and sanction celerity—that controls the behavior of offenders. The basic theoretical concept has been demonstrated in many evaluations of traffic safety programs conducted in the last half century (Ross, 1973).

Sobriety Checkpoints and Random Breath Testing (RBT)

Sobriety checkpoints provide the US police departments with the closest approximation to the highly successful random breath testing (RBT) enforcement procedure used in Australia, Sweden, and other countries. RBT, as implemented in Australia, allows officers to stop any vehicle on the road at random and to take a breath test from the driver. Operators with BACs greater than .05, which is the legal limit, are transported to the police station for an evidential test. In contrast, in the United States, vehicles can only be stopped at random at specially designed "checkpoints," and the drivers cannot be required to take a breath test. Rather, the officer must conduct an interview to determine whether the driver is impaired, and if there is evidence of impairment, the officer must require the driver to perform a set of field sobriety tests to establish impairment before transporting the offender to the police station.



In a time-series analysis of four Australian states, researchers found that RBT was twice as effective as "selective" checkpoints similar to those conducted in the United States. Another study found that in Queensland, Australia, RBT resulted in a 35% reduction in fatal crashes, compared with 15% for checkpoints. The researchers estimated that every increase of 1000 in the daily RBT testing rate corresponded to a decline of 6% in all serious crashes and 19% in single-vehicle nighttime crashes. Moreover, analyses revealed a measurable continuing deterrent effect of RBT on the motorist population after the program had been in place for 10 years. On the other hand, a review of studies on the effectiveness of sobriety checkpoints in Thailand showed mixed results (some showed effectiveness, others showed no effects).

Studies of the US sobriety checkpoint procedure found that they are associated with significant decreases in alcohol-related crashes. Reviews of checkpoint programs have reported reductions of 17%–75% in alcohol-related fatalities. Two related meta-analyses studies of 15 US checkpoint programs occurring between 1985 and 1999 found that the median reduction in crashes associated with checkpoints was 20% (Elder et al., 2002). A cost-benefit study of sobriety checkpoints indicated that, for every

\$1 invested in the checkpoint strategy, the community conducting the checkpoint saved \$6. In a recent report published by the Transportation Research Board of the National Academies of Science, Engineering and Medicine, the lessons from other nations indicate that frequent and sustained use of sobriety checkpoints could save up to 3000 lives annually, if used in the United States ([Transportation Research Board, 2010](#)). If passive alcohol sensors (PAS) are used at sobriety checkpoints, this could simulate the effects of RBT.



A benefit of checkpoint operations that has not been fully evaluated is the suppression of crime beyond impaired driving. In 2003, the Fresno, California, Police Department increased impaired driving enforcement from 1 or 2 operations per year between 1998 and 2002, to 32 in 2003 and almost 130 in 2010. Between 2003 and 2011, not only were alcohol-related crashes reduced by 17%, but also burglary rates and motor vehicle theft rates per capita declined by 17% and 32%, respectively.

Preventing Impaired Driving Offenders From Repeating

License suspension was once the most effective single DUI sanction. However, recent evaluation studies have found remedial interventions (treatment and educational programs) to be more effective than traditional punitive sanctions, such as jail terms and fines when combined with license restriction, for reducing recidivism and alcohol-related crashes. A meta-analysis conducted in 1995 of 215 evaluations of drink-driving remediation (treatment) programs concluded that the best designed studies indicated that treatment can produce an additional 7%–9% reduction in drink-driving recidivism and alcohol-related crashes when compared with control groups that largely received license restrictions only (sometimes more severe than for the treatment groups) ([Wells-Parker et al., 1995](#)).

DUI/Drug Courts

Based upon the effectiveness of these drug courts, DUI courts have been established to provide intensive supervision of offenders by judges and other court officials who closely monitor compliance with court-ordered sanctions coupled with treatment. DUI courts generally involve frequent interaction between the offender and the DUI court judge, intensive supervision by probation officers, an appropriate level of treatment, random alcohol and other drug testing, community service, lifestyle changes, positive reinforcement

for successful performance in the program, and jail time for noncompliance. Mostly nonviolent offenders who have had two or more prior DUI convictions are assigned to a DUI court, if one exists in the jurisdiction (not all communities or states have DUI courts). Several studies have reported favorable results for DUI/drug courts; however, additional evaluation is required, particularly in the area of cost-effectiveness because of the heavy use of court resources.

Screening and Brief Interventions

When drink-driving prevention fails, many drivers become involved in crashes and wind up in hospital emergency rooms or regional trauma centers. This provides an opportunity outside the criminal justice system for the private health care system to intervene with high-risk drivers. The trauma of being involved in a crash is viewed as providing a “teachable moment,” when individuals may be open to undertaking a behavioral change that will reduce their risk of future crashes and injuries. Research suggests that 30%–50% of injured, crash-involved drivers admitted to emergency departments or trauma centers have BAC levels higher than the legal limit for driving. Some of these offenders may receive a treatment intervention because of a DUI arrest; however, many are not charged because they are taken to the hospital before police officers have an opportunity to examine them for impairment. Furthermore, hospital staff rarely notifies the police when they receive a high-BAC driver. When screened for alcohol use disorders, an estimated 27% of injured patients admitted to the emergency department or a trauma center test positive for alcohol abuse or dependence, indicating a large number of people entering emergency rooms are potential DUI offenders. Thus, emergency rooms and trauma centers have an important opportunity to intervene with high-risk individuals who need treatment for alcohol problems—problems that may well lead to impaired driving and alcohol-related crashes in the future. Most importantly, brief interventions effectively improve consequences related to driving, such as moving traffic violations, drink-driving violations, motor-vehicle crash involvement, crash sustained injuries, and motor-vehicle fatalities.

Vehicle Sanction Laws

Aside from the traditional license suspension sanction, several laws and programs have been developed that prevent the offender from driving. These laws provide for seizing and impounding, immobilizing, or forfeiting the offender’s vehicle, and programs for interlock-equipped vehicles prevent a drinking driver from starting a vehicle. Overall, existing studies suggest that impoundment is an effective method of reducing the recidivism of DUI and driving-while-suspended (DWS) offenders. For vehicle impoundment to be effective, the vehicle must be impounded when the offender is arrested, and a procedure must be devised to deal with nonoffender owners.

Alcohol Monitoring

An alternative to controlling the driving of DUI offenders is to control their drinking. Judges frequently admonish offenders to remain abstinent while on probation, but unless a program exists that monitors the BAC, this action has little force other than allowing the judge to impose more severe penalties, if the offender returns to court intoxicated during the probation period. In the past, offender abstinence has been monitored in several ways. Some courts have implemented closely supervised Antabuse administration, which is drug that makes the offender sick, if they consume alcohol. Others have implemented intensive supervision programs in which probation officers make surprise visits to the homes of offenders and conduct breath tests. As noted, DUI/drug courts generally provide for intensive monitoring of abstinence. Such systems are labor intensive and expensive for the courts. In the last couple of decades, innovative technological methods for collecting BAC data have received considerable attention. One of these systems, electronically supervised, home-confinement programs, has been used for some time. This program involves telephone-monitored electronic random breath testing (RBT) systems that allow frequent monitoring of the BAC level while the offender is at home.

The South Dakota 24/7 Sobriety Program is a sobriety monitoring program aimed toward repeat DUI offenders, first DUI offenders with very high BACs, and similar offenders who have had repeated convictions related to alcohol abuse. Offenders report to the Sheriff’s Office twice a day (at about 7 a.m. and 7 p.m.) for alcohol breath testing. If the offenders have any positive BACs or they miss a scheduled test, they go to jail for 12–36 h. South Dakota adopted legislation in 2007 that established the 24/7 program statewide. The program is now operating in 60 of the 66 counties in South Dakota and has been implemented in North Dakota, Wyoming, and other states. An independent evaluation of the South Dakota 24/7 program in 2013 showed a 12% reduction in repeat DUI arrests and a 9% reduction in domestic violence arrests associated with the adoption of the 24/7 program (Kilmer et al., 2013).

More direct transdermal-monitoring systems that are worn on the body and monitor the BAC level 24 h a day/7 days a week are used by the judicial system. Two devices have been studied—the SCRAM ankle bracelet and the WristAS, which is about the size of a large wristwatch. The SCRAM incorporates a system for detecting circumvention. Although laboratory evidence indicates that SCRAM is reasonably accurate in detecting a significant BAC, no adequate evaluation of the SCRAM unit has been conducted with DUI offenders to determine its effectiveness in maintaining abstinence.



A method for monitoring drinking that is just emerging in the United States but is more widely used in Europe is to analyze the biomarkers in blood, urine, or hair that result from consumption of alcohol. Although alcohol is cleared relatively rapidly from the body, usually even after heavy drinking within 8–12 h, alcohol biomarkers persist in circulation in hair and in urine long after alcohol has fallen to zero in the blood. These biomarkers provide a way to measure total consumption over an extended time span. Longer-lasting blood markers reflecting alcohol use, such as urinary ethylglucuronide (EtG), offer a window of detection of 36 h or more following drinking. Hair EtG is being used in Germany as a relicensing criterion for participants in the driver's license restitution processes following conviction for impaired driving. To be eligible for reinstatement, a person's hair EtG level must be below 3 pg/mg (picograms/milligram), which is known to reflect abstinence.

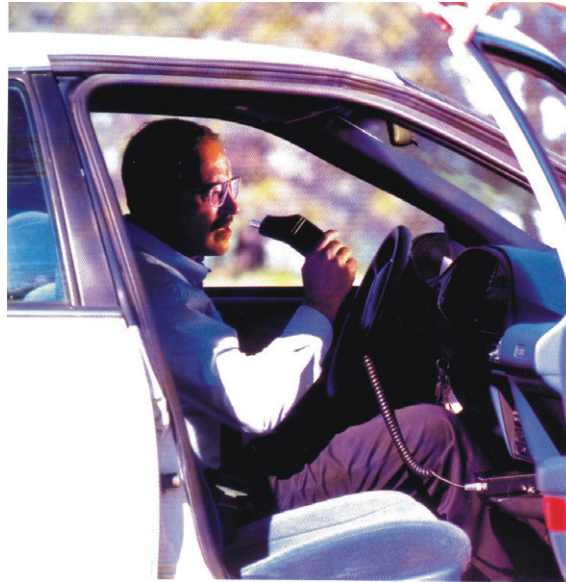
Alcohol Consumption

Generally, a country with a low rate of alcohol consumption per capita has a low alcohol involvement rate in fatal traffic crashes [e.g., Saudi Arabia (0.2 L of pure alcohol per capita), Turkey (2.4 L)]. However, that is not always the case as some countries with high alcohol consumption rates [e.g., Russia (14.5 L), Czech Republic (14.1 L), Australia (12.6 L), United Kingdom (12.0 L)] have low alcohol involvement rates in fatal crashes [e.g., Russia (5%, but most likely underreported), Czech Republic (8%), Australia (28%), United Kingdom (19%)]. The alcohol consumption rate in Sweden is 8.7 L of pure alcohol per capita and their fatal crash alcohol rate is 18%. In the United States, the alcohol consumption rate is 9.0 while the alcohol involvement rate in fatal crashes is 30%.

Countries that are highly urbanized and have adequate mass transit systems and other alternative transportation tend to have lower drinking and driving rates.

Alcohol Ignition Interlocks

The most parsimonious approach to the control of DUI offenders' impaired driving is the alcohol ignition interlock, which requires the driver to provide a breath sample when starting the car and at random times while driving. The units have four basic elements: (1) A breath-alcohol sensor that records the driver's BAC level and can be set to provide a warning, if any alcohol is detected and that will prevent starting the vehicle if the BAC is .03 or higher. (2) A rolling retest system that requires a new breath test every few minutes while driving to prevent someone else from starting the vehicle for a person who has been drinking. (3) A tamper-proof system for mounting the unit in the vehicle that is inspected every 30–60 days. (4) A data-logging system that records both the BAC tests and engine operation, thus ensuring that the offender is actually using the vehicle and not simply parking it while driving another vehicle. State-licensed service providers install the unit, inspect it regularly (generally, every 30–60 days), and provide a report on any attempt to circumvent the device to a court probation officer or a driver analyst at the department of motor vehicles. Such monitoring systems, with substantial consequences for tampering with the device, are essential for ensuring that offenders cannot drive the interlock-equipped vehicle while impaired. There is strong evidence that the recidivism of DUI offenders is substantially reduced while interlocks are on their vehicles. The offenders who install interlocks have 36%–90% lower DUI recidivism rates while the interlock is on their vehicles than similar DUI offenders who remain suspended ([Elder et al., 2011](#)). Once the interlock is removed, however, their recidivism rate does not differ from fully suspended offenders.



Three interlock laws were evaluated in 2018 by comparing alcohol-impaired passenger vehicle drivers involved in fatal crashes between 2001 and 2014 in the United States across state and time (Teoh et al., 2018). The exposure measure was the number of passenger vehicle drivers in fatal crashes that did not involve impaired drivers. Laws requiring interlocks for drivers convicted of DWI covered: (1) repeat offenders, (2) repeat offenders and high-BAC offenders, (3) all offenders, and (4) none. All offender laws were associated with 16% fewer drivers with .08+ BAC involved in fatal crashes, compared to no law. Repeat and high-BAC laws were associated with an 8% reduction in impaired drivers in fatal crashes, compared to no law. Laws mandating alcohol ignition interlocks, especially those covering all offenders, are an effective impaired driving countermeasure that reduces the number of impaired drivers in fatal crashes.

Conclusion

Alcohol-impaired driving is prevalent throughout most countries of the world. The key to controlling the problem is a systematic approach involving alcohol regulation, strong impaired driving laws, consistent, visible and frequent enforcement, and effective sanctions for DUI offenders. Success in the future involves passive alcohol detection systems in vehicles and the safety and feasibility of autonomous vehicles.

The Future

Driver Alcohol Detection System for Safety (DADSS) Program

A long-term development program was inaugurated by National Highway Traffic Safety Administration (NHTSA), automobile manufacturers, and Mothers Against Drunk Driving (MADD) to develop a system for all cars that can passively sense the BAC of the driver and prevent ignition of the vehicle's motor if the driver is over the BAC legal limit. The program is called "Driver Alcohol Detection System for Safety" (DADSS) and is currently being funded by NHTSA and most of the motor vehicle manufacturers (Zaouk et al., 2015). If such a vehicle can be produced, it would finally fulfill the objective enunciated 40 years ago—to build "Cars that Drunks Can't Drive." One of these monitors passively detects alcohol in the breath of the driver, making it unnecessary for the driver to blow into an interlock breath tube. Another developing technology measures BAC passively through the skin, providing a substitute for blowing into the interlock. These systems may have value as a method for activating a driver interlock system only when there is evidence of a drinking driver in the vehicle. Thus, the not-too-distant future offers the possibility of equipping all new vehicles with a system that would make impaired driving unlikely.

Autonomous Vehicles

Autonomous vehicles, if proven feasible and safe, could substantially change how we travel and could also eliminate drunk driving. While fully autonomous vehicles are not yet commercially available, the technology is developing rapidly and some computer-generated driving functions are already on public roads today. However, regulations are needed for enhancing safety and protecting the public interest. States will need to create stakeholder-working groups to oversee the development of state and local laws. One of the big questions will be: Can drunk drivers get home safely in autonomous vehicles?

Acknowledgments

This chapter was based, in part, on a chapter by Voas and Fell published in 2014 and entitled *Programs and Policies Designed to Reduce Impaired Driving*. It appeared in Kenneth J. Sher (Ed.), *The Oxford Handbook on Substance Use Disorders*, Volume 2, Oxford University Press. The author received permission from the Oxford University Press to use some of the material for this encyclopedia chapter.

See Also

Much more information and research is available on alcohol-impaired driving compared to other drug-impaired driving. An informative review of drugged driving research is in the article below:

Elvik, R., 2013. Risk of road accident associated with the use of drugs: a systematic review and meta-analysis of evidence from epidemiological studies. *Accid. Anal. Prev.* 60, 254–267. The most recent alcohol and highway safety report can be found here:

Voas, R.B., Lacey, J.C., 2011. *Alcohol and Highway Safety 2006: A Review of the State of Knowledge*. National Highway Traffic Safety Administration, Washington, DC.

Brief descriptions of research projects funded by NHTSA can be found in this report:

National Highway Traffic Safety Administration, 2014. *Compendium of Traffic Safety Research Projects 1985–2013*. National Highway Traffic Safety Administration, Office of Behavioral Safety Research, Washington, DC.

An earlier Transportation Research Board (TRB) report is listed below:

National Academy of Sciences, 2010. *TRB Special Report 300—Achieving Traffic Safety Goals in the United States: Lessons from Other Nations*. Committee for the Study of Traffic Safety Lessons from Benchmark Nations, Transportation Research Board. National Academies Press, Washington, DC. Available from: <http://www.nap.edu/catalog/13046.html>.

A report from Australia on an evaluation of their RBT program is below:

Ferris, J., Devaney, M., Sparkes-Carroll, M., Davis, G., 2015. *A National Examination of Random Breath Testing and Alcohol-Related Traffic Crash Rates (2000–2012)*. Foundation for Alcohol Research and Education, Australia.

A general report on alcohol use and misuse is documented below:

Rehm, J., Mathers, C., Popova, S., Thavorncharoensap, M., Teerawattananon, Y., Patra, J., 2009. Alcohol and global health 1: global burden of disease and injury and economic cost attributable to alcohol use and alcohol-use disorders. *The Lancet* 373, 2223–2233.

Technology that has potential to reduce impaired driving crashes is listed below:

Pollard, J.K., Nadler, E.D., Stearns, M.D., 2007. *Review of Technology to Prevent Alcohol-Impaired Crashes (TOPIC)*. US Department of Transportation, Research and Innovative Technology Administration, John A. Volpe National Transportation Systems Center, Cambridge MA.

Three relevant TRB Circulars are referenced below:

Transportation Research Board, 2013. *Countermeasures to Address Impaired Driving Offenders: Toward an Integrated Model*. Alcohol, Other Drugs, and Transportation Committee, Transportation Research Circular E-C174, Washington, DC. Available from: www.trb.org.

Transportation Research Board, 2007. *Traffic Safety and Alcohol Regulation: A Symposium*. Transportation Research Circular E-C123, Washington, DC. Available from: www.trb.org.

Transportation Research Board, 2005. *Implementing Impaired Driving Countermeasures: Putting Research into Action: A Symposium*. Transportation Research Circular E-C072, Washington, DC. Available from: www.trb.org.

The chapter on pedestrian safety will also provide more information on alcohol involvement in pedestrian fatalities.

References

- Blomberg, R.D., Peck, R.C., Moskowitz, H., Burns, M., Fiorentino, D., 2005. *Crash risk of alcohol involved driving: a case-control study (Final report)*, Dunlap & Associates, Stamford, CT.
- Borkenstein, R.F., Crowther, R.F., Shumate, R.F., Ziel, W.B., Zylman, R., 1974. The role of the drinking driver in traffic accidents (the Grand Rapids study). *Blutalkohol* 11 (1).
- Compton, R.P., Berning, A., 2015. *Drug and Alcohol Crash Risk (DOT HS 812 117)*, US National Highway Traffic Safety Administration, Washington, DC.
- Elder, R.W., Shults, R.A., Sleet, D.A., Nichols, J.L., Zaza, S., Thompson, R.S., 2002. Effectiveness of sobriety checkpoints for reducing alcohol-involved crashes. *Traffic Inj. Prev.* 3, 266.
- Elder, R.W., Voas, R., Beirness, D., Shults, R.A., Sleet, D.A., Nichols, J.L., Compton, R., 2011. Task force on community prevention services. Effectiveness of ignition interlocks for preventing alcohol-related crashes: a community guide systematic review. *Am. J. Prev. Med.* 40 (3), 362–376.
- Fell, J.C., Thomas, S., Scherer, M., Fisher, D., Romano, E., 2015. Scoring the strengths and weaknesses of underage drinking laws in the United States. *World Med. Health Policy* 7 (1), 28–58.
- Fell, J.C., Scherer, M., Thomas, S., Voas, R.B., 2016. Assessing the impact of twenty underage drinking laws. *J. Stud. Alcohol Drugs* 77, 249–260.
- Fell, J.C., Scherer, M., 2017. Estimation of the Potential Effectiveness of Lowering the Blood Alcohol Concentration (BAC) Limit for Driving from 0.08 to 0.05 grams per Deciliter in the United States. *Alcohol. Clin. Exp. Res.* 41 (12), 2128–2139.
- Kilmer, B., Nicosia, N., Heaton, P., Midgette, G., 2013. Efficacy of frequent monitoring with swift, certain, and modest sanctions for violations: Insights from South Dakota's 24/7 sobriety project. *Am. J. Public Health* 103, 37.
- Moskowitz, H., Robinson, C.D., 1988. *Effects of Low Doses of Alcohol on Driving-related Skills: A Review of the Evidence (DOT HS 807 280)*, US National Highway Traffic Safety Administration: Washington, DC.
- Moskowitz, H., Fiorentino, D., 2000. *A Review of the Literature on the Effects of Low Doses of Alcohol on Driving-Related Skills (DOT HS 809 028)*, US National Highway Traffic Safety Administration, Washington, DC.
- National Academies of Sciences, Engineering and Medicine, Engineering and Medicine, 2018. *Getting to Zero Alcohol-Impaired Driving Fatalities: A Comprehensive Approach to a Persistent Problem*. The National Academies Press, Washington, DC.
- Ross, H.L., 1973. Law, Science and accidents: The British road safety act of 1967. *J. Legal Stud.* 2 (1).
- Shults, R.A., Elder, R.W., Sleet, D.A., Nichols, J.L., Alao, M.O., Carande-Kulis, V.G., Zaza, S., Sosin, D.M., Thompson, R.S., 2001. Task force on community preventive services: Reviews of evidence regarding interventions to reduce alcohol-impaired driving. *Am. J. Prev. Med.* 21 (4 Suppl), 66.
- Teoh, E.R., Fell, J.C., Scherer, M., Wolfe, D.E., 2018. *State Alcohol Ignition Interlock Laws and Fatal Crashes*. Published by the Insurance Institute for Highway Safety, Arlington, VA.
- Teutsch, S.M., Naimi, T.S., 2018. Eliminating alcohol-impaired driving fatalities: what can be done? *Ann. Intern. Med.* 168, 587–589.
- Transportation Research Board, 2010. *Special Report 300: Achieving traffic safety goals in the United States: Lessons from other nations*. Washington, DC: Transportation Research Board of the National Academies of Science, Committee for the Study of Traffic Safety Lessons from Benchmark Nations.
- Wagenaar, A.C., Toomey, T.L., 2002. Effects of minimum drinking age laws: Review and analyses of the literature from 1960 to 2000. *J. Stud. Alcohol (Suppl 14)*, 206.
- Wells-Parker, E., Bangert-Drowns, R., McMillen, R., Williams, M., 1995. Final results from a meta-analysis of remedial interventions with drink/drive offenders. *Addiction* 90, 907–926.

Zaouk, A., Wills, M., Traube, E., Strassburger, R., 2015. Driver Alcohol Detection System for Safety (DADSS) – A Status Update. 24th Enhanced Safety of Vehicles Conference. Gothenburg, Sweden.

Further Reading

NHTSA Alcohol and Highway Safety Reports

The first comprehensive report on alcohol and driving in the United States was issued in 1968 in response to a Congressional requirement when the National Highway Safety Bureau (now the NHTSA) was founded. This US Department of Transportation report summarized the fairly limited research in the traffic safety field and highlighted the role of “heavy drinkers” (alcoholics and problem drinkers) in the alcohol-related crash problem. It had a substantial influence on the safety community and was an important factor in the decision by Congress to appropriate funds for the Alcohol Safety Action Program (ASAP), the first National demonstration program on drinking and driving. A follow-up to the 1968 report, *Alcohol and Highway Safety: A Review of the State of Knowledge*, was published a decade later in 1978. The report covered three main areas: (1) it defined the alcohol-crash problem, (2) it addressed the alcohol-crash problem, and (3) it provided future directions for the alcohol-crash problem. The publication contained an extensive review of individual studies in localities throughout the United States. An important addition to the 1968 report was its presentation of a “health approach” as a part of the criminal justice system. This approach focused on screening people convicted of DWI offenses and referring them to treatment.

In 1985, an update of *Alcohol and Highway Safety* was published by NHTSA. This report built on the 1978 report and included the “most clearly important studies and findings for the period from January 1978 to December 1982.” This update contained the first discussion of vehicle and roadway engineering programs as methods for reducing impaired-driving crashes. It provided the first mention of alcohol ignition interlock systems. It also discussed the topic of exposure reduction (e.g., reduction in alcohol availability), the emergence of underage drinking as a potential issue in traffic safety, and the possible value of age 21 as the MLDA limit.

In 1989, the fourth in its series of *Alcohol and Highway Safety* reports was published. Influenced in part by the *Surgeon General’s Workshop on Drunk Driving* in 1989 that was held the year before, this update gave considerable attention to controls on the availability of alcohol, reporting that 28% of the countermeasure documents reviewed related to availability. This increased focus on laws and policies related to drinking, as distinct from impaired driving, and reflected the movement to close the gap between research on traffic safety and research on public health issues surrounding alcohol abuse that was stimulated by the *Surgeon General’s Workshop*.

The fifth in the series of *Alcohol and Highway Safety* reports was published in 2001. That report found evidence for the effectiveness of four types of legislation when enforced: ALR laws, reducing the legal BAC limit from .10 to .08 g/dL, raising the MLDA to 21, and zero-tolerance laws.

Other Comprehensive Reviews

Perhaps the most significant research update was published as a product of the *Surgeon General’s Workshop on Drunk Driving* conducted in 1988. The publication of the Workshop in 1989 consisted of a compendium of 15 papers provided by experts in traffic safety and in public health research. It covered a wide range of topics, including the pricing and availability of alcohol, alcohol advertising, and marketing, as well as traditional topics, such as the epidemiology of impaired driving, law enforcement, and DWI offender treatment programs. The significance of the publication was based on the participation in the workshop by researchers in both traffic safety and public health. Bringing together leaders in these two fields served to integrate public health activities directed at reducing the health risks presented by alcohol.

In 2001, the Transportation Research Board (TRB) Committee on Alcohol, Other Drugs, and Transportation produced a report on research needs and priorities. That document, like the *Surgeon General’s Workshop Report*, consisted of a compendium of individual papers. The presenters at the conference were required to respond to a set of evaluative questions; their responses were used to develop a list of alcohol research priorities.

The latest in the series of NHTSA *Alcohol and Highway Safety* reviews was published in 2006. Several smaller, independent reviews of the alcohol safety area have appeared in the last 2 decades, among them was the *Epidemiology and Consequences of Drinking and Driving* published in 2003. That report described the groups most at risk for the involvement in alcohol-related crashes, reporting that there are 80 million trips each year in which a driver exceeds the legal BAC limit (.08). Further, although alcohol-related fatal crashes decreased in the 1980s, there has been little reduction since the mid-1990s and even a slight increase between 2000 and 2003. A report funded in part by the World Health Organization (WHO), *Alcohol: No Ordinary Commodity Research and Public Policy*, provided a broad review of alcohol consumption, patterns of drinking, and policies directed at reducing high-risk drinking. The report also contains a chapter reviewing drinking-and-driving countermeasures.

In 2018, the National Academy of Sciences, Engineering and Medicine (NASEM) released the most comprehensive report on accelerating progress to reduce alcohol-impaired driving fatalities in the United States to date (NASEM, 2018; Teutsch and Naimi, 2018). The report (written by a prestigious committee assembled to review the impaired driving problem) provides a blueprint to solving the problem by identifying evidence-based and promising policies, programs, strategies, and system changes to

increase nation progress in reducing alcohol-impaired driving traffic fatalities. Among many other recommended strategies, those pertinent to this study include:

1. Local governments should adopt and/or strengthen laws and dedicate enforcement resources to stop illegal alcohol sales (i.e., sales to already intoxicated adults and sales to underage persons).
2. Local law enforcement agencies should conduct sobriety checkpoints in conjunction with widespread publicity to promote awareness of these enforcement initiatives.
3. Municipalities should support policies and programs that increase the availability, convenience, affordability, and safety of transportation alternatives for drinkers who might drive otherwise. This includes permitting transportation network company ride sharing, enhancing public transportation options (especially during night time and weekend hours) and boosting or incentivizing transportation alternatives in rural areas.
4. Every state should implement DWI courts and these courts should include available consultation or referral for evaluation by an addiction-trained clinician.
5. All states should enact all offender alcohol ignition interlock laws. To increase effectiveness, states should consider increased monitoring periods based upon the offender's BAC at the time of arrest and past recidivism.
6. States should enact per se laws for alcohol-impaired driving at .05 BAC and accompany enactment with media campaigns and robust and visible enforcement efforts.

Animal Crashes

Michal Bíl, CDV–Transport Research Centre, Lišeňská Brno, Czechia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	53
Where Do Animal Crashes Occur?	54
Geographic Distribution of AVCs and the Species Involved	54
Clustering of AVCs Along Roads	54
When Do AVCs Occur?	55
Consequences of Animal Crashes	55
Crashes With Domestic Animals	56
Cost of Animal Crashes and Impact of Roadkill to Wildlife	57
Animal-Motor Vehicle Crash Databases	58
How to Avoid Crashes With Animals?	59
Information for Drivers	59
Measures Targeted to Animals and Their Behavior	59
Measures Suitable for Secondary Roads	61
Monitoring of Animal Behavior	61
Animal Crash Awareness Among Public and Educational Programs for Drivers	61
The Outlook for the Future	62
Acknowledgments	62
Biography	62
Relevant Websites	62
Links to Videos	62
References	62

Introduction

Mobility of people and goods is an essential need of modern societies. Land, air, and marine transportation are on the rise globally at present causing inevitable conflicts with wildlife. These interactions are the most pronounced in land transportation as road and railway infrastructures are widespread and will even be expanding in the coming years. Domestic animals are usually part of traffic and their owners have to obey the regulations. Wild animals, on the other hand, cross the road without warning for many reasons, and therefore cause serious traffic safety issues in numerous countries. Animal-vehicle crashes (AVCs) represent significant percentages of all officially registered crash records in a great number of countries and these, particularly with large mammals, are growing worldwide. The primary cause of this increase is the growing number of game species followed by rising traffic intensity as well as improved registration of these crashes.

Conflicts between transportation and animals are not only related to land transport, however, incidents between planes and flying birds or bats are very often reported. More than 13,000 bird strikes are reported annually in the United States. Despite the fact that even state-of-the-art aircraft can be endangered by such crashes, particularly when birds hit the aircraft front window or are sucked into the engines, the numbers of major accidents including human fatalities are very low. An example of this kind of crash was the collision of a passenger aircraft with a flock of birds in 2009 known as the “Miracle on the Hudson” when Airbus A320 landed safely on the river. Apart from these events, aircrafts sometimes collide with animals during take-off.

Collisions of sea mammals or large fishes with boats are often reported. The enormous size of many vessels prevents passengers from injuries. The situation is, however, different when whales collide with rapidly moving ferries or yachts. Whales even occasionally attack small yachts as a result of their erroneous determination of the floating object. Many yachts have been destroyed as a consequence of such encounters.

Collisions with animals are also frequently reported on railways. The extreme mass of railway vehicles protects passengers from injuries from AVCs. Relatively high portion of these crashes are therefore reported only when a considerably large animal is struck or when trains collide with herds of wild or domestic animals. Many videos from locomotive cabins exist on social media, which document animals’ behavior in front of an incoming train (Video 1). Deer typically use rail tracks in the winter when these are the only places without deep snow. Derailments are reported as a result of crashes with domestic animals at local railways, but also took place when a German high-speed train derailed after a crash with a herd of sheep in 2008. This article will focus on animal crashes, which take place on roads, as the road network is considerable denser than the railway network and often intersects natural habitats of wildlife.



Figure 1 Roe deer occur in high numbers in certain countries. They are habituated to traffic and thus can be seen quite often in herds resting not far from roads, Czechia. Source: Photography: Miloslav Kratochvíl.



Figure 2 Large mammals can be encountered when driving in many southern Africa countries. Source: Photography: Wendy Collinson-Jonker, South Africa.

Where Do Animal Crashes Occur?

Geographic Distribution of AVCs and the Species Involved

The road network spans the globe intersecting the habitats of native wildlife species. Crashes with wild animals thus logically occur, but in much higher numbers than people usually realize. The numbers of reported AVCs are rising over the last decades in numerous countries. Ungulates contributed the most to these statistics. The frequency of ungulate-vehicle collisions rose as a result of increase in their population, for example, white-tailed deer (*Odocoileus virginianus*) in the United States and roe deer (*Capreolus capreolus*) in many European countries (Fig. 1). The enormous growth of wild boar (*Sus scrofa*) population has been documented throughout Europe. They live in almost all of Europe today. Recently, wild boars were observed as far north as the central parts of Sweden and Finland.

Each continent and biogeographical region has its own representatives of large fauna, usually mammals, which are involved in motor vehicle crashes. Northern European countries, as well as Russia, Canada, and United States, report a number of crashes with deer, particularly moose (*Alces alces*) as the largest member of the *Cervidae* family (Sullivan, 2011; Sáenz-de-Santa-María and Tellería, 2015; Niemi et al., 2017). North Africa and Middle Eastern countries often report crashes with camels, while in Australia drivers collide frequently with kangaroos (Rowden et al., 2008). In South America, drivers meet Giant Anteaters, Capybaras, or domestic llamas on roads. Large fauna also occur on roads in South Africa (Fig. 2). Carnivores are generally less common than herbivores, but in certain countries crashes with carnivores are also often reported (e.g., Eastern European and Balkan countries report a number of crashes with wolves and bears).

Clustering of AVCs Along Roads

While the geographic distribution of roads and their density influence the overall number of AVCs, the local AVC pattern can also be investigated. It was found that AVCs occur in certain places along roads or railways with higher frequencies than in other parts. For example, 34% of all registered crashes with roe deer took place on less than 1% of the Czech road network. These places can be identified, using spatial statistic approaches applied to crash data, if AVCs were reported with a sufficient positional accuracy. Places where crashes concentrate, so-called hotspots, reveal specific local characteristics, which influence and explain the observed high concentration of crashes. These *local factors* need to be further studied, analyzed, and mitigated, where possible (Bartonička et al., 2018). Typical local factors are the view and the overall visibility for drivers, the existence of shrubs or trees

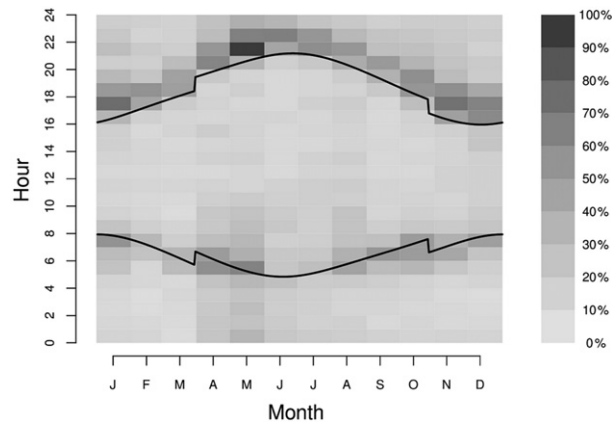


Figure 3 Relative frequency of reported AVCs involving roe deer in the Czech Republic. Lines represent the times of sunrise and sunset, respectively. Crashes with roe deer occur with the highest frequency shortly after sunset in May. Source: Data come from www.srazenazver.cz; Graphics by R. Andrášik (CDV).

close to roads, the existence of guardrails, etc. Proper identification of animal-crash hotspots is necessary for consequent effective targeting of mitigation measures. This fact is particularly important as a lot of mitigation measures are costly and their application should be focused first and foremost at the most hazardous places.

When Do AVCs Occur?

It has been determined that peaks of animal crashes occur around sunset and sunrise. This is influenced by periodic animal activity, which can be observed over years and days. Taking into account the fact that traffic intensities are usually low during sunset, the risk of AVC is therefore high. In addition, the crash frequency with animals is not stable over the year (Fig. 3). Two distinctive peaks can be found, for example, for ungulates, during the spring and autumn and another smaller one related to rutting of roe deer occurs between July and August (Steiner et al., 2014).

Consequences of Animal Crashes

More than one million deer-vehicle crashes occur every year in the United States resulting to approximately 200 fatalities (Huijser et al., 2007). According to CARE data, the European Union's road accidents database, more than 3000 human fatalities and serious injuries were recorded in Europe between 2011 and 2017. The situation is apparently the worst, in light of the absolute number of casualties, in Sweden due to the abundance of moose and in Germany and France because of their large road networks.

Despite the fact that less than 5% of wildlife-vehicle crashes result in human injury in general, geographic differences are enormous (Langbein et al., 2011). It depends on the weight and height of the animal involved in the crash (Jakobsson et al., 2015). The highest risk of fatal human consequences is with large mammals, for example, moose or camel. The moose is a large animal (Fig. 4). An adult moose weighs between 200 and 700 kg and its height in the shoulder ranges between 1.7 and 2.1 m (Fig. 5). It also has long legs. If a standard passenger car collides with a moose at high speed, the moose's large body rotates over the engine hood and lands on the windshield (Video 2). Such collisions are capable of causing substantial injury to the vehicle occupants



Figure 4 A moose crossing a road in Denali National Park (Alaska, USA). Source: Photography: Rick Price.

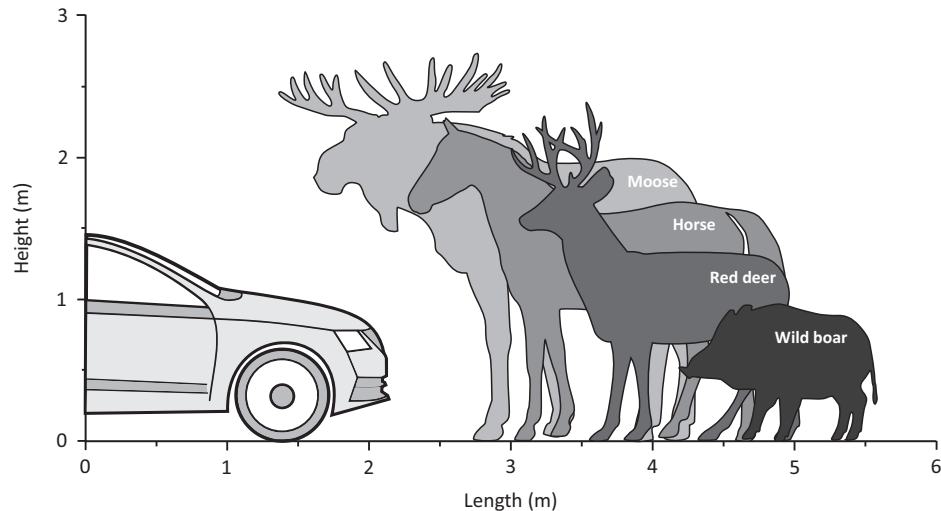


Figure 5 A comparison of relative heights of selected species which frequently collide with motor vehicles and the height of an average personal car. Source: Author: Vojtěch Nezval, CDV.



Figure 6 A comparison of car damage due to a crash with a wild boar (A) and a model of a moose (B). The smaller and lighter wild boar usually interacts with the bumper, while the larger moose interacts with the windscreen. This is often followed by a damage of car roof. Source: Photography by Generali (Czech Republic) and Jan Wenäll (VTI, Sweden), respectively.

(Fig. 6). The same mechanism can be expected with crashes with camel, horse, and red deer. Passenger car crashes are logically the most frequently reported, as these kinds of vehicles are the most abundant.

Animal crashes are, however, also reported with other types of vehicles including motorcycles and even bicycles. Motorcyclists in particular are extremely vulnerable road users as they usually drive at high speed and their stability and passive safety is significantly lower when compared to cars or trucks. Motorcyclists are therefore overrepresented among the group of injured and deceased drivers when related to crashes with animals.

In certain situations, avoiding crash with an animal is a more dangerous maneuver than a crash. This is particularly true when roads are not *forgiving*, that is, when trees are planted on road verges or other solid objects are placed there, for example, advertisement panels. Crashes into mature trees have very high fatality rates in the long run. Moreover, many fatal crashes to trees on local roads occur in night hours. It is therefore possible that some of them were a result of avoiding a collision with wildlife. Rational road verge management, focusing on forgiving roads, is necessary to avoid the negative consequences when cars travel off the paved road (Figs. 7 and 8).

Crashes With Domestic Animals

For the period of 2010–18, on average, between 300 and 500 crashes with domestic animals were registered on Czech roads annually. It is very less as compared to more than 70,000 crashes reported with wild animals over the same period (7,800 annually on average). Certain domestic animals have a different daily activity pattern than wildlife. Crashes with them therefore take place in different parts of the day than those with wildlife. Moreover, domestic animals are often utilized as engines, particularly in the developing world. Such animals, when frightened, might cause a collision with traffic. It is therefore advised to drive carefully when

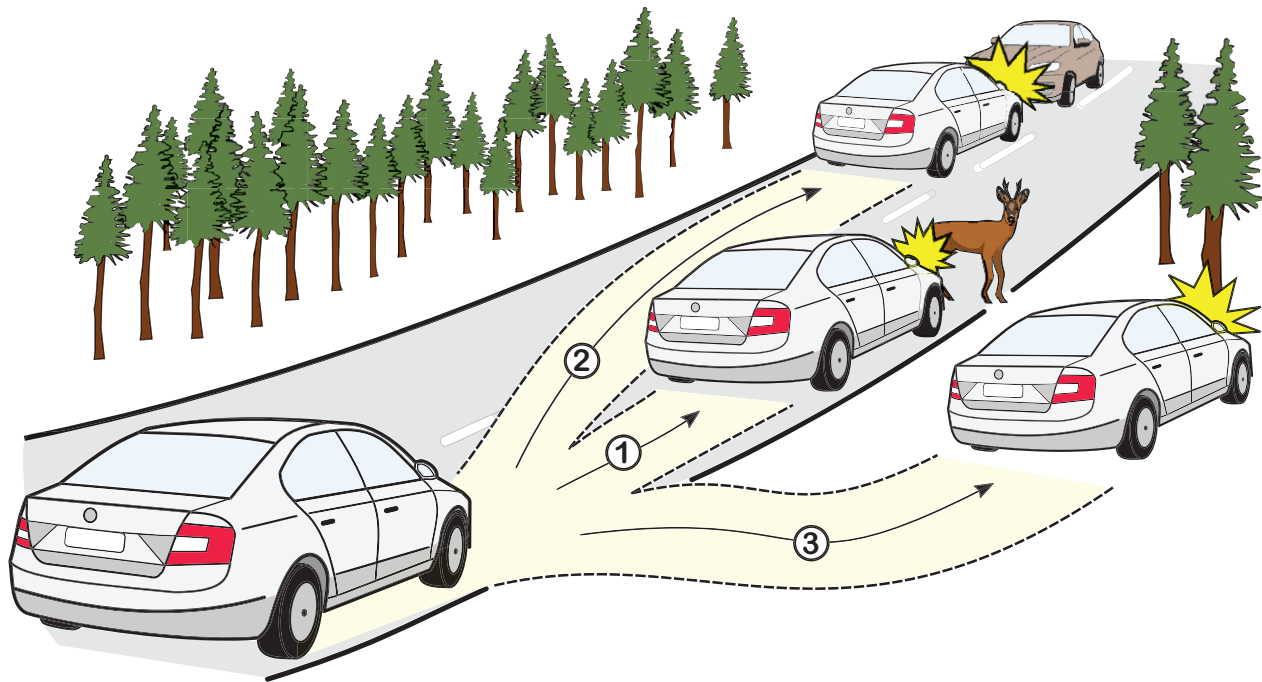


Figure 7 A sketch demonstrating that a hazardous situation can result in fatal outcomes when an animal enters the road. Whereas a crash with an animal, up to the size of a roe deer, usually means car damage (1), drivers are sometimes trying to avoid collisions, which can result in a frontal crash to an incoming vehicle (2) or to a tree (3). These situations can only occur on secondary roads without central and side guardrails or where there are no wide road verges free of trees. Source: Author: Vojtěch Nezval, CDV.



Figure 8 An example of a secondary collision due to AVC: a van hit a tapir (see the carcass of the tapir in the middle of the figure), then the driver lost control of the vehicle which consequently crashed into a large lorry. The lorry exploded and nine people died on September 2015, BR-267 Highway, Mato Grosso do Sul State, Brazil. Source: Photography: LTCI-IPE-Brazil.

passing an animal-powered vehicle. The highway law in Maine (USA) mentions these situations in point 3 (Frightened animals) as follows: When a person riding, driving, or leading an animal that appears to be frightened signals by putting up a hand or by other visible sign, an operator approaching from the opposite direction must stop as soon as possible and remain stationary as long as necessary and reasonable to allow the animal to pass. Situations when large herds, for example, sheep, cows, are grazing close to a road, where there are no fences, represent a danger to traffic. Crashes with dogs, which often move freely close to human settlements, are also quite frequent.

Cost of Animal Crashes and Impact of Roadkill to Wildlife

Overall economic losses from AVCs are enormous due to the extremely high number of these events. The majority of losses are related to car damage. Whereas domestic animals usually have an owner who can be responsible for a crash, this is not generally true for wild animals. Game species in certain countries belong to a hunting association, but costs related to crashes can only be covered

by car insurance. Vehicle-related damage, but also the loss of a domestic animal, which is being used for livelihood in developing countries, and related emotional losses are often part of the AVC outcomes.

While animal crashes mostly represent only an issue related to vehicle damages, a number of wild animals die or suffer from injuries. It was also determined that certain species can even be endangered by extinction due to roadkill, particularly for species with low population sizes and low growth rates. This threat is not, however, only directly related to roadkill. When adult animals, which are feeding youngsters, die due to roadkill, it can also affect the youngsters who consequently die as a result of starvation. The fact that the overall number of roadkill and the species involved were not known recently led to the development of wildlife-crash reporting systems in many countries.

Animal-Motor Vehicle Crash Databases

Transport agencies, motorway, and transport police record data on traffic crashes on a daily basis. National crash databases are therefore useful sources of information about this phenomenon. A well-known fact about animal crashes is their underreporting. One of the reasons is that the crash data are only gathered when the minimal damage cost is exceeded or a human injury took place. It has been estimated that the total number of collisions involving motor vehicles and large animals has been underestimated by 20%–30%, although some local officials and hunters go as far as evaluating this underestimation by as much as 50%. The situation related to small animals is, logically, even worse.

Systematic, precise, and consistent data, particularly related to species involved in AVC, are therefore needed. Many ad-hoc AVCs reporting applications, usually utilizing a citizen science approach, have recently been developed to overcome these issues. Some of them have been statewide, but the majority only regional coverage. Readers could be interested in seeing some of the websites: California roadkill observation system in the United States (<https://www.wildlifecrossing.net/california/>), Project Roadkill in Austria (<https://roadkill.at>), Srazenazver.cz (Fig. 9) in the Czech Republic (<http://www.srazenazver.cz/en/>), Taiwan Roadkill Observation network (<https://roadkill.tw/en>), Project Splatter in the United Kingdom (<https://projectsplatter.co.uk/>), and Animals under wheels in Belgium (<https://waarnemingen.be/vs/start>). These databases can be useful sources of information where AVCs occur and which species were involved. Data can then be analyzed in order to identify places where the crashes concentrate as well as other important factors including species distribution, crash circumstances and consequences, date of the crashes, etc.

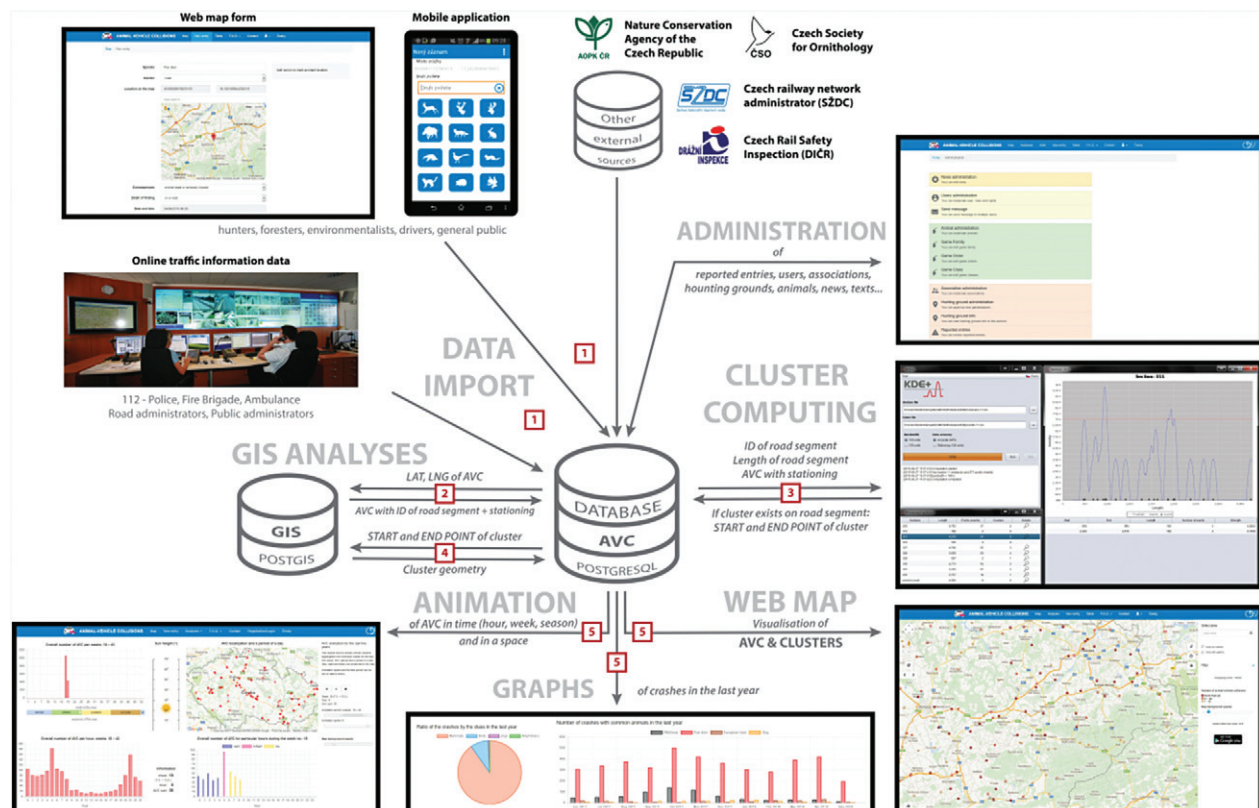


Figure 9 An example of AVC reporting system. Srazenazver.cz combines data from traffic crashes reported by Police data from volunteers and road administrators. Hotspots are identified online and graphs and maps are generated.

How to Avoid Crashes With Animals?

How do animals behave when close to roads? Researchers determined that several kinds of animal behavior related to road crossings exist. Small animals usually do not see roads and the traffic there is dangerous and are therefore consequently killed in high numbers (e.g., a great number of species of amphibians and reptiles). Other species, in contrast, react to incoming traffic with defensive behavior (e.g., hedgehogs) and are then killed as well. Other groups of species try to escape. This behavior is typical for roe deer, which are the most frequently involved in the recorded crashes in Europe. Its behavior can be seen, from the drivers' perspective, as chaotic and nonlogical. Many drivers, who collided with deer, reported later that the deer was running alongside their car and then, suddenly, jumped in front of the car.

Deer often use trees and shrubs in road verges as shelter or a source of food particularly in highly intensive agriculture landscapes where there is a lack of these sources. The availability of forage, water bodies, feeding, and resting places close to roads are likely to increase the presence of animals. There is therefore an eminent danger to traffic when animals decide to cross roads or are alarmed and escape consequently by crossing the road. High traffic volume roads may also deter certain animal species from crossing. Roads can act in these cases as a barrier to free animal movement. Many recommendations currently exist as to how to lower the number of AVCs. They can be divided into two main groups: information for drivers and approaches influencing animal behavior.

Information for Drivers

Drivers are often surprised when an animal comes closer to a road or is crossing immediately in front of a car, as has been reported many times by patients who have sustained AVCs. All drivers have to be therefore cautious when spotting animals close to roads or when going through forested areas.

Drivers should take note of road traffic signs. It has been shown earlier, however, that static warning signs can be ineffective. Driver simulator studies have demonstrated that drivers respond with a greater velocity reduction to AVCs countermeasures, which are considered novel, such as a radio-warning message (Jägerbrand and Antonson, 2016). New driver-warning systems are therefore currently being developed. Usually infrared-based systems are installed along roads in places where frequent movement of animals is expected. They are activated, that is, flashing and informing drivers about a potential danger, when animals are detected close to a road. It was also found that drivers who frequently use a road could even be habituated to these active signs, particularly when they produce a significant portion of false positive information. These systems should only be active when animals are detected close to a road. It has been shown in the United States that these temporary warning signs were well-lit at night during deer migration periods and reduced the number of deer road-kills at experimental sites compared to control sites.

Speed reduction is recommended as safety behavior for drivers, which reduces energy significantly during a crash with an animal. Speed reduction zones are therefore recommended in crash-prone areas, but they should also be dependent at the time of day. Deer often move in groups. Drivers should be ready for the other animals, particularly when one of them has just crossed the road before their cars. Breaking and then colliding with an animal is also recommended when an animal steps suddenly in front of a car. There are only a few exceptions, related to large mammals, when the crash could have worse outcomes for passengers than avoidance maneuvers. In these cases, practitioners suggest aiming at the back of the large animal, for example, a moose.

Measures Targeted to Animals and Their Behavior

Plenty of measures have been developed to influence animals and avoid animal crashes. No universal measure exists; however, as the animals differ significantly in their behavior, body sizes, and life cycles. Roads and other transportation infrastructure should be planned and designed in such a way so as to keep in mind this phenomenon of animal crashes. Animals cross the road for many reasons. Regular seasonal migration of large herds of ungulates poses a traffic safety issue in North America. Migration of pronghorn (*Antilocapra americana*) across the western United States has been studied thoroughly and many crossing structures have been built recently in order to avoid crashes with motor vehicular traffic. Such large continental migration is not typical in densely populated Central Europe. Game species (wild boar, roe deer) live within rather small areas there. The majority of crashes are caused there due to their daily movement. Daily movement of animals, which have relatively small areas, can be sufficiently affected by installing feeding places or ponds at the same side of roads where they live. Territorial behavior of animals (e.g., roe deer) occurs in only certain parts of the year and can therefore be predicted.

Animals are attracted to roads for certain reasons. Salt is often used during the winter in northern countries to help with the road maintenance. Ungulates are attracted, however, to roads and lick the salt blocks. This behavior is risky as the animals often cross the road when finding other sources of salt. Winter road maintenance should therefore use different substances than salt to de-ice road surfaces. Some animal species, for example, moose, use transport corridors for travel during periods of deep snow (Video 1) as well as for feeding.

Fruit trees or other vegetation provide both forage and shelter for animals and thus increase the attractiveness of such habitats along roads. Another source of food along roads comes from food remains, which drivers and passengers throw out of the cars in rural areas thus attracting animals close to roads. Scavengers are also attracted to roads when carcasses are left on or close to roads for a longer time. Asphalt road pavements accommodate energy over hot days and are therefore warmer than the natural neighborhood in the nighttime. Certain species lie on the surface during nights, utilizing this warm surface, and can be killed by incoming vehicles or cause a crash when drivers are avoiding collisions. These cases may result in a higher concentration of animals around the roads and the risk of crashes.

Certain wildlife species, particularly game species, are overpopulated. This is the case in many Central European countries. A number of authors have recommended targeted herd management, that is, the reduction of the animal population size through hunting to decrease the number of animals and then indirectly the number of animal crashes. In certain parts of North America, large boulders along roads are also considered as an effective preventive measure to block deer from entering roads. It has been observed that ungulates, in particular, hesitate to go over large boulders. Systems consisting of iron bars (cattle grids), allowing cars to go from roads to the adjacent service forest roads, were also already installed in Europe. These measures are effective barriers to deer, but not for wild boar as was observed.

Vegetation removal at road verges is also a recommended measure (Rea, 2003). It makes road verges safer for drivers and increases driver visibility. However, young vegetation, when road verge management is not done carefully, can attract ungulates close to roads. This is particularly true for moose, which prefer vegetation growing in open areas adjacent to forests. Animals, when resting in shelters, as is common in open or intensively used agricultural areas, can be alarmed by, for example, a free ranging dog, tourists, an agriculture vehicle, etc., and then escape by crossing a road thus causing danger to road traffic.

The most effective structures preventing AVCs are fences (Huijser et al., 2007). They can, however, create an almost absolute barrier to animal movement. Fencing should therefore be used in association with wildlife crossing structures (also called fauna passages) such as over- and underpasses (Fig. 10). Fencing ranks among the most expensive measures. Maintenance costs are considerably high when this measure is applied to a large part of the road network. It has been shown that broken fences, or fences, which are interrupted, can have an even worse effect on traffic safety than roads without them. The reason is that animals can enter the area around a road, but then they are not able to find their way out and are caught in a closed corridor (Fig. 11). Fences have to be checked regularly for consistency and overall conditions, as many animals are able to dig under or even destroy them (e.g., wild boar). Connections between fences and other structures, such as overpasses or bridges, also pose a weak part of these systems. However, poor connections with space, of about half of 1 m wide, can often be seen. This is still enough space to allow certain ungulates to go through and enter the area around roads.



Figure 10 Overpasses, when accompanied with fencing, are highly effective measures against animal-vehicle crashes. A so called green bridge “Ivaceno brdo” on “A1” highway (E71) between Zagreb and Split (Croatia). Source: Photography: Djuro Huber.



Figure 11 Fences are sometimes accompanied by ramps to allow animals (particularly deer) to escape from areas around roads. Note the mesh size of the fence. A denser grid is intended to block the movement of small animals at Zeepoort, Netherlands. Source: Photography: Edgar Van Der Griff.



Figure 12 Reflectors mounted on bollards are being used particularly in Central European countries to deter animals from crossing roads. An example from Styria (Austria). Source: Photography: Wolfgang Steiner.

Measures Suitable for Secondary Roads

In most of the countries, motorways are usually fenced nowadays and therefore safe in the view of crashes with large mammals. Rural roads and roads in certain developing countries are, on the other hand, regularly crossed by animals, thus causing potentially dangerous situations for drivers and car passengers. There is almost no controversy that the motorways should be equipped with fences accompanied with fauna passages, rural, and low traffic volume roads need less expensive but still effective solutions. Which measures can be used to follow the above-mentioned criteria of financial effectiveness? Current research has not still offered the explicit answer. Many measures were suggested, developed, studied, and tested.

For example, *olfactory repellents* are installed at high numbers across the central European countries. The odor usually consists of foam injected with scent, which mimic predators' or humans' scent. It is believed that animals, particularly ungulates, should be more cautious when crossing roads fringed with these repellents. Results from few dataset currently available are, however, not unequivocal. More research is still needed. The same is true for assessment of the effectiveness of *wildlife warning reflectors* (Fig. 12), which were in high numbers installed across the German roads recently, but only brought ambiguous results. The reflectors are supposed to deter particularly ungulates from entering the road by reflecting the lights of incoming vehicles aside, to the road verge. Some authors argue that even if these devices were not capable to deter animals off roads, and to directly decrease the number of crashes, they could serve at least as a warning for drivers. Drivers, when notice the presence of a mitigation measure of such a kind, they could slow down. *Wildlife whistles* mounted to vehicles were tested in the North America, but also here the results were rather negative. Moreover, concerns remain, if such high ultrasound will not alarm the animals and will cause their panic behavior including crossing roads.

Problem of habituation to some measures is typical for those, which should only warn animals when entering roads. Many studies focused only on data from relatively short periods. Moreover, it is not still clear how the animals behave and react on many of warning devices, as there almost were no behavioral studies conducted. Animals are, however, intelligent creatures and are to a certain extent able to understand that some measures pose no actual danger to them.

Monitoring of Animal Behavior

Knowledge of animal behavior and their daily and seasonal movement is important for predicting when and where crashes occur. Recent technological progress allowed for the development of both light and endures motion detectors, which can be, in a form of a GPS collar, attached to animals. This helps researchers understand certain basic principles of animal movement. Mitigation measures, once installed, have to be monitored regularly and repaired when necessary. Only fully functional measures are able to deter animals from crossing roads. The effectiveness of mitigation measures has to be evaluated regularly and cost-benefit analyses should always be performed to ensure that the counter measures would also be cost-effective.

Animal Crash Awareness Among Public and Educational Programs for Drivers

The probability of animal crashes can be lowered to a certain extent. Drivers should be aware that the majority of crashes with wildlife occur during twilight and night. Poor driver visibility in night hours or due to obstacles close to roads, for example, trees and shrubs, means that they will have less time to react if an animal appears in front of their cars. Therefore, drivers should travel at a lower speed when moving through forested areas particularly during the night. A lower speed also means lower energy released during a crash and this is particularly important as the kinetic energy is proportional to the speed raised to the second power.

Drivers can also learn on their own from watching instructive videos captured by dash cams mounted in many vehicles which their owners posted on the Internet and social media (e.g., YouTube). Certain common characteristics of crashes with animals can be identified from these records (Rea et al., 2018). Driving school curricula should contain information about these kinds of crashes

and the circumstances and such videos could also be a particularly useful tool. Swerving to avoid animals is often dangerous as it can result in run-off road crashes or a collision with other road users with severe consequences for car passengers.

The Outlook for the Future

Automobiles can be equipped with infrared cameras and thus inform the drivers about thermal-energy emitting objects, which cannot be detected within the visible spectra, occurring near roads during night hours or when there is a fog. Similar to the high-tech electronic systems, it is expected that the cost of these onboard warning systems will decrease once they become more commonplace on vehicles. The autonomous vehicles will also be equipped with technology which can identify animals moving along or crossing roads including animal detection systems or night vision assistants.

Acknowledgments

This work was supported by the project of the Transport R&D Centre (LO1610). The author would also like to thank many individuals who helped him providing data, photographs, or figures, namely: Richard Andrášik, Wendy Collinson-Jonker, Sandra MacDougall, Molly Grace, Edgar Van der Grift, Djuro Huber, Jiří Kasina, Miloslav Kratochvíl, David Livingstone, Patricia Medici, Vojtěch Nezval, Rick Price, Jiří Sedoník, Wolfgang Steiner, Jan Tecl, Casey Visintin, and Jan Wenäll.

Biography

Michal Bíl works at CDV—Transport Research Centre where he focuses on geographic analyses of transportation data and analyses of traffic crashes in particular.

References

- Bartonička, T., Andrášik, R., Duřa, M., Sedoník, J., Bíl, M., 2018. Identification of local factors causing clustering of animal-vehicle collisions on roads. *J. Wildl. Manage.* 82, 940–947.
- Huijser, M.P., McGowen, P., Fuller, J., Hardy, A., Kociolek, A., Clevenger, A.P., Smith, D., Ament, R., 2007. Wildlife vehicle collision reduction study. Report to Congress. U.S. Department of Transportation, Federal Highway Administration, Washington D.C., USA.
- Jägerbrand, A.K., Antonson, H., 2016. Driving behavior responses to a moose encounter, automatic speed camera, wildlife warning sign and radio message determined in a factorial simulator study. *Accid. Anal. Prevent.* 86, 229–238.
- Jakobsson, L., Lindman, M., Carlsson, H., Axelsson, A., Kling, A., 2015. Large animal crashes: the significance and challenges. Proceedings from IRCOB Conference 2015, IRC-15-42, 302–314.
- Langbein, J., Putman, R.J., Pokorny, B., 2011. Road traffic accidents involving ungulates and available measures for mitigation. In: Putman, R.J., Apollonio, M., Andersen, R. (Eds.), *Ungulate Management in Europe: Problems and Practices*. Cambridge University Press, Cambridge, UK, pp. 215–259.
- Niemi, M., Rolandsen, C.M., Neumann, W., Kukko, T., Tiilikainen, R., Pusenius, J., Solberg, E.J., Ericsson, G., 2017. Temporal patterns of moose-vehicle collisions with and without personal injuries. *Accid. Anal. Prevent.* 98, 167–173.
- Rea, R.V., 2003. Modifying roadside vegetation management practices to reduce vehicular collisions with moose *Alces alces*. *Wildl. Biol.* 9 (4), 81–92.
- Rea, R.V., Johnson, C.J., Aitken, D.A., Child, K.N., Hesse, G., 2018. Dash Cam videos on YouTube™ offer insights into factors related to moose-vehicle collisions. *Accid. Anal. Prevent.* 118, 207–213.
- Rowden, P., Steinhardt, D., Sheehan, M., 2008. Road crashes involving animals in Australia. *Accid. Anal. Prevent.* 40 (6), 1865–1871.
- Sáenz-de-Santa-María, A., Tellería, J.L., 2015. Wildlife-vehicle collisions in Spain. *Eur. J. Wildl. Res.* 61 (3), 399–406.
- Steiner, W., Leisch, F., Hackländer, K., 2014. A review on the temporal pattern of deer-vehicle accidents: impact of seasonal, diurnal and lunar effects in cervids. *Accid. Anal. Prevent.* 66, 168–181.
- Sullivan, J.M., 2011. Trends and characteristics of animal-vehicle collisions in the United States. *J. Saf. Res.* 42 (1), 9–16.

Relevant Websites

<http://www.srazenazver.cz/en/>
<https://roadkill.at/en/>
<https://www.wildlifecrossing.net/california/map/roadkill>

Links to Videos

Video 1: Behavior of a moose in front of a train.
<https://www.youtube.com/watch?v=wATuKiSf1yE#t=22>
 Video 2: A simulation of a moose-motor vehicle crash. Credit: VTI, Sweden.

Attenuators

Simonetta Boria, School of Science and Technology, University of Camerino, Camerino, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	63
Vehicle Passive Safety	64
Energy Absorbing Deformation Mechanisms	65
Design of an Impact Attenuator Made of Conventional Material	67
Impact Attenuator Definition	67
Testing Procedures	68
Numerical Simulations	68
Results and Discussion	69
Design of an Impact Attenuator Made of Composite Material	70
Impact Attenuator Definition	70
Experimental Dynamic Tests	71
Finite Element Modeling	71
Results and Discussion	73
Conclusions	75
Acknowledgments	75
References	75
Further Reading	76

Introduction

One of the most important aspects to consider in the design of any mode of transport, such as passenger cars, railways, or aircraft structures, is crash safety or crashworthiness. Crashworthiness is defined as the capability of a structure to provide adequate protection to its passengers from injuries in case of a collision. The safety requirements are outlined by agencies and organizers of a specific travel mode or competition; to meet these regulations, designers must manufacture energy absorbing systems, called impact attenuators. These can be fitted in front of bridge abutments and other unforgiving structures as well as designed into the vehicle in front of the survival area occupied by the driver and passengers. Impact attenuators are generally made up of aluminum, honeycomb, fiber-reinforced polymer (FRP) composites, or a combination of these materials that provides maximum protection. The deformation of such components should absorb the highest level of kinetic energy while transmitting low load uniformly to the rest of the vehicle and to the vulnerable components (e.g., driver and passengers) maintaining an acceptable level of crushing. The use of lightweight materials may contribute also to improve the acceleration performance and fuel economy of the vehicle.

Many research teams have investigated the behavior of empty or foam-filled aluminum tubes or honeycomb structures under axial quasi-static and dynamic loading conditions using both experimental and numerical techniques (Boria, 2010; Chirwa et al., 2003; Hanssen et al., 2000; Hsu and Jones, 2004; Langseth et al., 1999; Munusamy and Barton, 2010; Rappolt, 2015). According to the literature, honeycomb structures are relatively light in weight and absorb more energy than other combinations (i.e., empty or foam filled). However, a large block of honeycomb occupying the whole volume of the impact attenuator is a heavy and expensive option and it may be advantageous to combine the benefits of thin-walled structures and honeycombs in a single system.

More recently, the attention has shifted toward FRP materials because of their advantages of lightweight, high strength, corrosion resistance, and easy manufacturing (Davies, 2000; Fuchs et al., 2008; Smith, 1990) compared to conventional materials (Jacob et al., 2002). Consequently, the use of composite crushing absorbers is increasing and today it is more common to find them applied in many high-performance vehicle structures like in sports cars, high-speed trains, and helicopters. Initially, in order to reduce the vehicle weight, glass fiber reinforced plastic composites were adopted. Their replacement to conventional materials into specific vehicle allowed a weight reduction from 40% to 60% and better performance (Ning et al., 2007a, 2007b, 2009; Thattai parthasarthy et al., 2006). More recently, carbon fiber reinforced plastic (CFRP) composites are replacing metallic components for crashworthiness applications, as attested by extensive numerical and experimental research (Boria, 2013; Feraboli and Masini, 2004; Ghasemnejad et al., 2010; Hwang et al., 2011; Mamalis et al., 2005; Yang et al., 2007).

The main advantage of composites respect to conventional materials under crush may be attributed to the different type of crushing behavior. Contrary to conventional structures, in fact, FRP composites fail through a combination of several different fracture mechanisms (Guimard et al., 2009; Guoxing and Tongxi, 2003; Jumahat et al., 2010). These complex failure modes, that involve interlaminar failure, fiber-matrix debonding, fiber pull-out, matrix deformation/cracking, and fiber breakage, together with friction effects, contribute to the overall energy absorption and give composite structures a greater specific energy absorption (SEA) than metallic ones (Mamalis, 2005a, 2005b; Mamalis et al., 1997A). If subjected to axial loadings, such structures commonly take

the form of thin-walled tubes of circular and rectangular cross section produced with fabric prepreg obtained from the combination of fibers and resin of different nature. The structural response depends also on the technology adopted for their production. The ability to predict their behavior tends to be complicated due to the heterogeneity of structural properties and manufacturing variability. This number of effects makes the design of composite energy absorbing structures much more complex than when using metals, which are extensively used for this kind of applications.

In crashworthiness design, finite element (FE) simulations play a very important role thanks to their possibility to predict behavior in early design phases, so enabling cost reduction and shortening of the design phase. The energy dissipation by plastic deformation in ductile metallic structures is well understood and may be modeled by elastic-plastic material models in FE analysis, reproducing good levels of predictability. The modeling of composite crushing phenomena, on the other hand, is still developing given the complexity and heterogeneity of the material. There are various approaches in modeling composite structures: micro-mechanical and macromechanical ones. In the first approach, there is a high consumption of computational resources and, therefore, it cannot be applied for complex structures. It is therefore important to develop models that are simple enough to be employed in practical analysis situations, but at the same time capable to provide results with a suitable level of accuracy.

The aim of this article is to focus on the methodology to follow in order to design a specific impact attenuator made of conventional or composite materials under dynamic loading. Conventional materials are typically applied as sandwich structures, with external layers in aluminum and internal aluminum honeycomb core, with a truncated pyramidal shape, and such structures were subjected to axial dynamic loading as analyzed by [Boria \(2016a\)](#). In case of composite material, the design was focused on a CFRP structure. A prepreg obtained from high-strength carbon fibers, arranged in a canvas in a balanced manner and immersed in an epoxy resin was adopted. The frontal impact attenuator of a Formula SAE car was investigated ([Boria, 2016b](#)), whose geometry is like a truncated pyramidal structure with rounded edges to get closer to taper. Previous research has demonstrated, in fact, that cones are more suitable for impact structures than tubes since they do not require crush initiators ([Meredith et al., 2012](#)); furthermore, they more accurately reflect real-life geometry found in the automotive and motorsport arenas than simpler plate type specimens. The experimental tests should be correlated with a numerical analysis; therefore, specific FE models to simulate the collapse mode of such structures and predict their energy-absorption capability were developed using the nonlinear explicit commercial code LS-DYNA. For each analysis, the numerical results were compared with those obtained from real axial crushing tests. The overall aim is to investigate whether such structures (in conventional or composite material) can be used for designing an impact attenuator for a specific application. After the comparison between numerical and experimental data, it was observed how the proposed configurations can absorb impact energy with deceleration values coherent with the specific homologation references.

Vehicle Passive Safety

In general, it is possible to divide the vehicle safety into two disciplines: active and passive safety ([Lukaszewicz, 2014](#)). Preventing the occurrence and likelihood of a crash is under the active safety activities, while passive safety aims to reduce or mitigate the severity of a crash for the vehicle structure, the occupants, and other involved crash partners. During a crash incidence the body in white of a vehicle must protect the passengers and reduce the crash severity by allowing controlled energy dissipation through structure deformation. When evaluating real-world crashes involving at least one car, the most widespread crash events are frontal impacts, followed by side and rear ones.

In the last couple of decades, passive safety requirements have continuously increased and today vehicles need to pass a certain set of legally defined tests in order to be allowed to be sold. In addition, the implementation of new car assessment programs (e.g., NCAPs, IIHS) has started; such programs want to define tests that emulate real-world crash cases in order to allow a comparison between different vehicles.

Initially, impact attenuators were designed in conventional material, using steel or aluminum structures. The crashworthiness benefits of composite materials were first recorded for applications in helicopters and Formula 1 vehicles. In 1980 McLaren introduced a chassis as a safety cell made from aluminum sandwich with carbon-epoxy face sheets. In 1985 frontal crash tests were introduced and most Formula 1 teams adopted a composite monocoque and a nose cone to follow these new regulations. Since then, side impact tests (in 1995), rear impact tests (in 1997), and lastly side intrusion tests (in 2001) have been introduced. As a result of this, the fatality rate has decreased over time.

The crashworthiness design is based on the research of the best geometric and material configuration of the attenuators, in order to get a vehicle deceleration history during impact that has the desired characteristic of progression ([Belingardi and Chiandussi, 2011](#)). In a passenger car the structures designed to absorb energy are in the front ([Fig. 1](#)), placed between the front bumper and the passenger compartment. The front structure generally consists of four longitudinal thin-walled beams, two for each side of the vehicle; two are positioned in the upper part of the engine compartment, just below the hood, and two, generally of bigger dimensions, at an intermediate height behind the bumper. In recent vehicles, at the front extremity of these two main beams, two so called defo-box or crash boxes are placed, one at each side. These crash boxes are intended to absorb energy in case of impact at low velocity, so avoiding large structural damage to the other parts of the front structure. Just behind the front bumper there is a transverse beam, also called front end module, aiming to position and support the bumper; it is connected at its extremities to the longitudinal main beams through the defo-boxes.

Overall, the front section of a vehicle, in terms of crash performance, is governed by requirements for stiffness, energy dissipation, and controlled deformation to enable good occupant protection. All three goals can be simultaneously achieved using lightweight

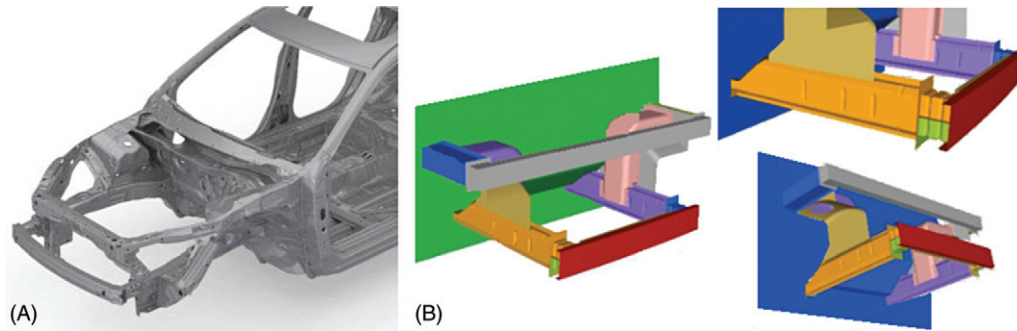


Figure 1 Front bumper assembly of a recent vehicle: (A) real, (B) model.

materials, such as aluminum, sandwich structures, or composite ones. However, it is the energy dissipation of composites that has widely been researched due to the potential improvements in performance and weight.

Energy Absorbing Deformation Mechanisms

The behavior of attenuators under axial loading conditions heavily depends on the material used. In metallic materials, energy is dissipated through plastic folding, work hardening and adiabatic losses during heating. Composites, by contrast, dissipate energy through external and internal friction, bending of the fibers and kinematic dissipation through fragmentation (**Fig. 2**).

Composite materials may offer many potential benefits for improving crashworthiness. Primarily, the SEA may be significantly higher, that is, composites may achieve the same level of crashworthiness as current metallic structures at a fraction of the weight. Lastly, composites deform by brittle fracture, unlike metals that deform by plastic crushing. During brittle fracture the whole component can be disintegrated, while for structures exhibiting plastic folding, the hinges limit the foldable length. This may have potential benefits for optimizing the deceleration of the vehicle, which can occur over a longer distance and thus resulting in a lower overall deceleration, which reduces the risk of secondary impacts (**Barnes et al., 2011**).

According to **Hull (1991)**, the principal deformation modes of energy absorbing for composite structures can be classified into the following categories:

- global buckling,
- progressive folding, and
- progressive crushing.
 - progressive splaying and
 - progressive fragmentation.

Generally, the first deformation mode absorbs very little energy; therefore it is undesirable for energy absorption applications and must be prevented for both conventional and composite materials.

Progressive folding is the typical deformation mode of metallic structures but may also be found in ductile FRPs and metal/FRP hybrid materials. It is characterized by a sequence of localized folding at the crush front, propagating away from the impact point through the structure. The SEA of a structure able to collapse into such mechanism can reach intermediate values, even if it cannot reach

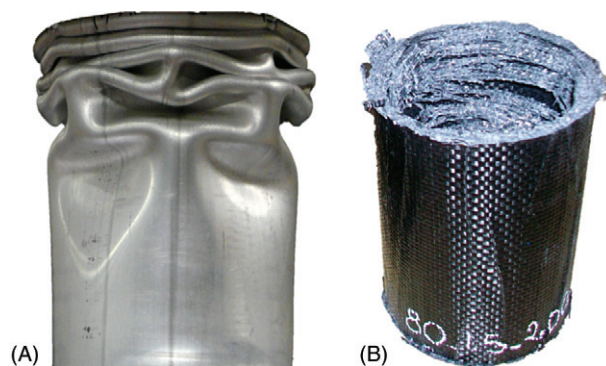


Figure 2 Comparison between plastic folding in metals (A) and brittle fracture in FRP composites (B).

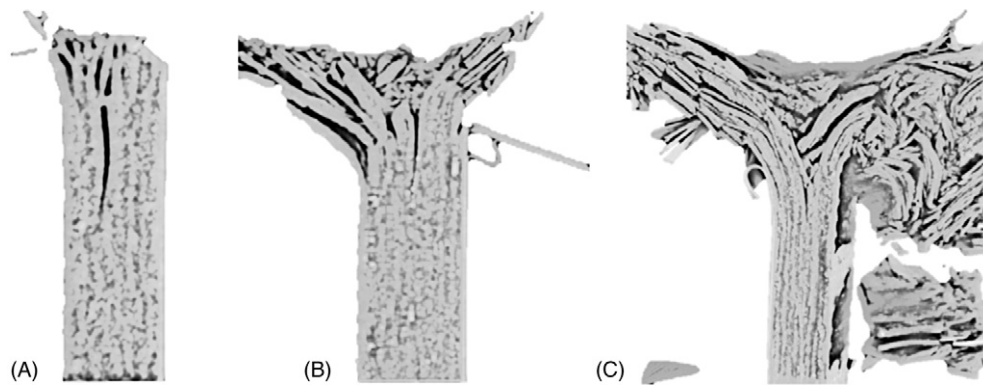


Figure 3 Sequence of deformation events during progressive crushing (Johnson and David, 2010): beginning of the crushing process (A), beginning of frond bending (B), and fully established crush front (C).

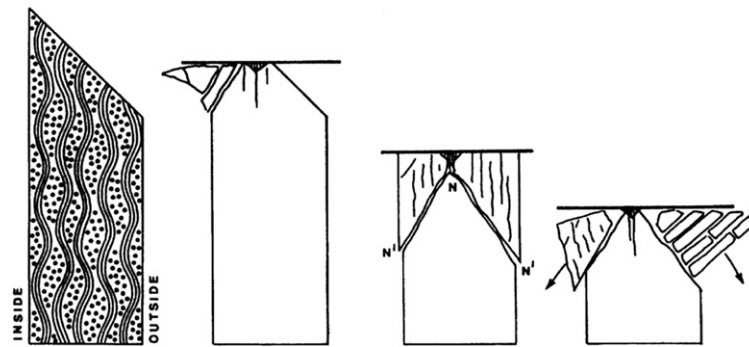


Figure 4 Schematic representation of formation of a fragmentation-mode crush zone (Hull, 1991).

the same levels as progressive crushing due to a lack of mechanisms for energy dissipation. The fact that this type of deformation was observed for some aramid and glass fiber reinforced specimens, as well as for some thermoplastic fiber reinforced composites, leads to the conclusion that a high strain to failure for the fiber and matrix is necessary for this deformation mode to occur.

Progressive crushing is the deformation mode typical of composite structures that are able to exhibit the highest SEA value. It can be divided into two main modes, progressive splaying with the formation of internal and external fronds (Fig. 3) and progressive fragmentation where a stable crush zone is established progressing down the structure, away from the point of initial contact (Fig. 4). In general, both progressive deformation modes may occur in any structure and the exact sequence of events resulting in either splaying or fragmentation it may strictly depend on the specific configuration of the specimen (e.g., fiber and matrix material, stacking sequence, manufacturing technology, geometry, trigger mechanisms) and on the test conditions (e.g., test speed, test direction, temperature).

According to Mamalis et al. (1998), the failure modes at macroscopic scale for a composite structure can be summarized into four major modes, depending on structural dimensions, material properties and testing conditions (Fig. 5). It is possible refers to them with the terminology Mode-I, Mode-II, Mode-III, and Mode-IV. Together with Mode-I and Mode-IV that correspond, respectively, to progressive crushing and progressive folding, already indicated by Hull, there are two other failure modes that can occur if the attenuator is not well designed. When the attenuator presents some edges, such as a square frustum, the undesirable failure Mode-II could occur, where it is evident a longitudinal crack progression along the edges of the structure. Moreover, it is possible to observe a centrally confined circumference crack (Mode-III) that, together with Mode-II, guarantees a low energy absorption capability and implies a catastrophic failure to avoid. Therefore it is evident that an impact attenuator should be designed to produce a Mode-I or a Mode-IV crushing mechanism, regarding mainly composite and metallic material respectively, as the structure guarantees the highest energy absorption.

Contrary to conventional materials, thin-walled composite structures under axial compression do not present a sequence of plastic hinges that in terms of force-displacement trend corresponds to low and high peaks (Fig. 6A) but are characterized by an almost constant crushing strength throughout the entire crushing process (Fig. 6B). The value of the peak load and the drop from it to the average crushing load are strongly dependent on the trigger geometry (Garner and Adams, 2008). As closer to the mean force value is the diagram, without high fluctuations, better is the designed attenuator because very similar to an ideal absorber.

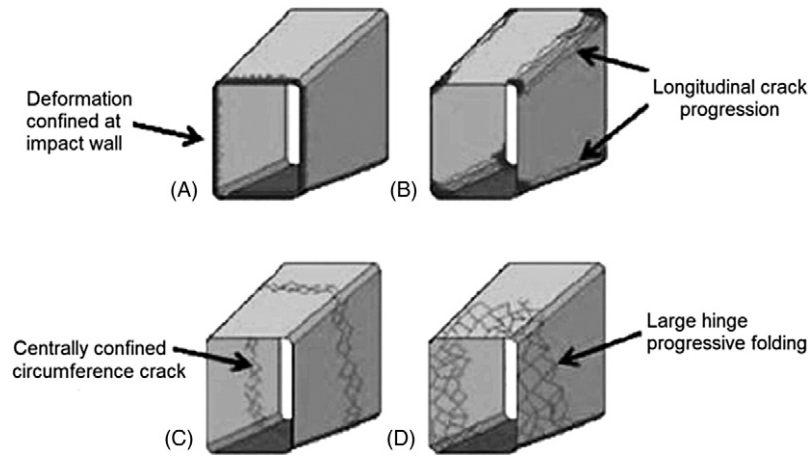


Figure 5 Collapse modes: (A) mode-I, (B) mode-II, (C) mode-III, and (D) mode-IV (Boria and Pettinari, 2014).

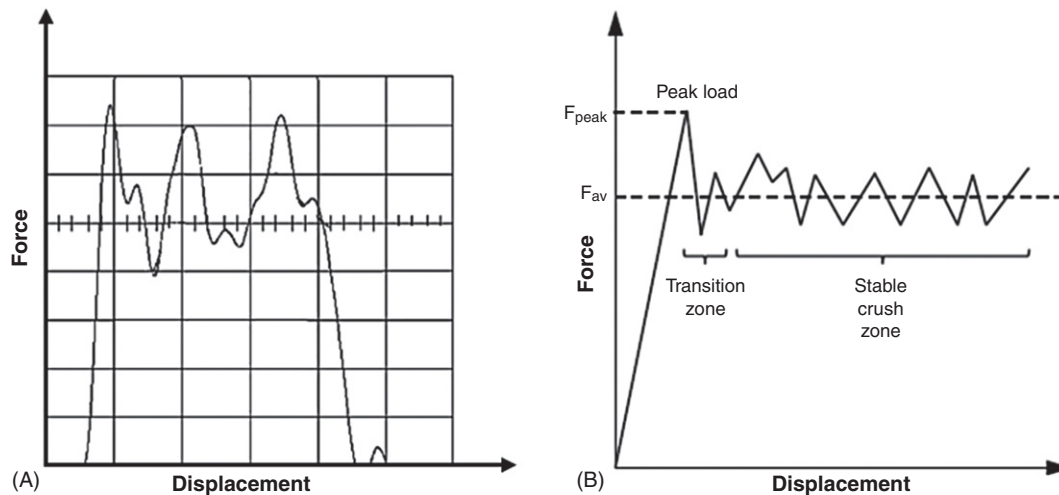


Figure 6 Typical force versus displacement trend for metallic (A) and composite (B) thin-walled structure under axial loading.

Design of an Impact Attenuator Made of Conventional Material

Impact Attenuator Definition

As mentioned before, the impact attenuator is such safety structure installed on a car that could absorb energy during frontal collision by progressive deformation. Such structures should be lightweight, resistant and deformable enough to limit the force and deceleration transmitted to the driver; moreover, it is necessary to design it according to specific technical regulations. For the FSAE rules (SAE, 2016), the impact attenuator must be installed in front of the bulkhead and must have a length, height and width of 200 mm, 100 mm and 200 mm, respectively. During the impact, it should not penetrate the front bulkhead and the total deformation must be limited to the attenuator. In the impact event of the vehicle (300 kg heavy) against a solid and nonyielding barrier at the velocity of 7 m/s, the attenuator must give an average deceleration of less than 20 g, a peak lower than 40 g and the total energy absorbed must be at least 7350 J.

In literature it is possible to find several studies on the design of impact attenuators in conventional material (Belingardi and Obradovic, 2010; Imanullah et al., 2018; Li et al., 2017; Munusamy and Barton, 2010; Santosa and Wierzbicki, 1998; Zarei and Kroger, 2006).

The impact attenuator analyzed in such section was made of aluminum sandwich material with a honeycomb core separating two aluminum thin skins. The shape chosen for the crash-box is a truncated pyramid (Fig. 7). The five sandwich panels are joined together with aluminum L-shaped profiles riveted and bonded on the skins with epoxy adhesive.

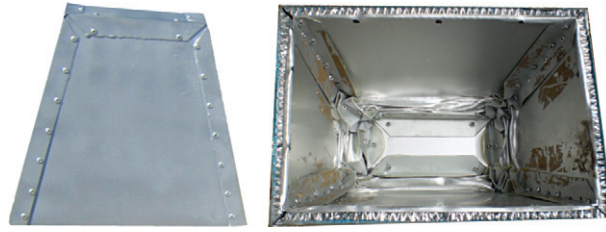


Figure 7 Impact attenuator: (A) external and (B) internal view.

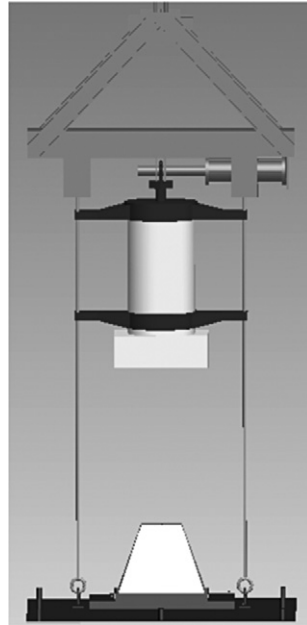


Figure 8 Schematic testing procedure.

Testing Procedures

According to the FSAE rules and in order to evaluate the crushing capability of the designed impact attenuator, dynamic experimental tests were planned and performed (Fig. 8). A mass of 300 kg is dropped from a height of 2.5 m, thus getting an impact velocity of 7 m/s, as requested by regulation. A piston driven by pressurized air operates the drop of the mass, able to fall in a controlled manner thanks to two steel ropes. The accelerometers, used to measure the mass deceleration, were placed on top of the mass. The impact attenuator was constrained by screws to a base of steel and concrete in order to reproduce the connection with the frame. A high-speed camera was used in order to capture the physical phenomenon, not visible to the naked eye; a sampling of 500 frames per second was implemented.

Numerical Simulations

In order to reproduce the crushing phenomenon numerically, a FE model was implemented. In particular, the explicit dynamic solver LS-DYNA was used employing the material properties obtained from dedicated experimental tests, such as out-of-plane compression test, three point bending test, buckling under in-plane compression and in-plane shear test. Due to the geometrical symmetry and in order to contain the computational time, only half attenuator was modeled (Fig. 9).

*MAT_PIECEWISE_LINEAR_PLASTICITY was used for the aluminum skins. The final failure of the external skins during the impact loading is characterized by a failure criterion. Failure means a loss of stiffness and it is detected when elongation of elements in any direction reaches the allowable strain. Among the different approaches available in modeling honeycomb structure, the core was simplified as an orthotropic continuum based on its effective material properties with solid elements (*MAT_HONEYCOMB). The car body and the barrier were idealized as rigid structures (*MAT_RIGID) assuming therefore that only the impact attenuator is responsible of the total kinetic energy absorption.

The total system, consisting of the impact attenuator and the car body idealized as a 300 kg mass attached to it, impacts the barrier wall at the velocity of 7 m/s, as imposed by technical regulation. Such contact was modeled using a node to surface contact, while in

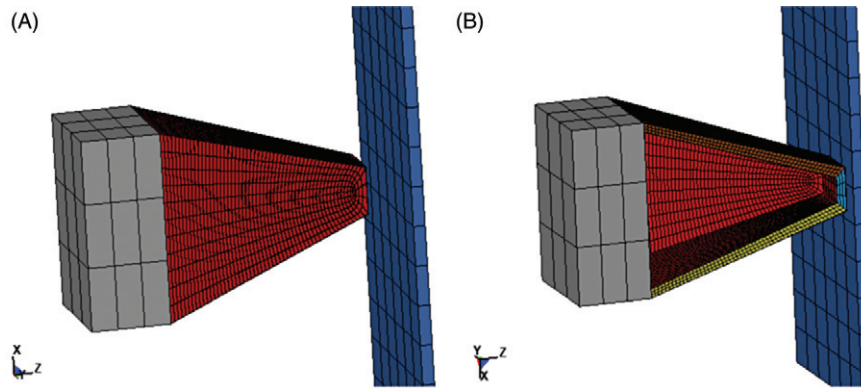


Figure 9 Numerical model of the impact attenuator: (A) external and (B) internal view.

order to prevent elements penetration of the impact attenuator during crushing a single surface contact was added. The elements representing the core were directly connected to the skin nodes at the interface using a tied break contact.

The explicit dynamic simulation using LS-DYNA can start only after fixing some control parameters, such as termination time, time-step and output requests.

Results and Discussion

Fig. 10 shows the impact attenuator crushing during the real test every hundredth of a second. From the frames it is evident that the energy absorption was guaranteed from the progressive folding mechanism; plastic hinges tended to form in succession from the top to the bottom of the structure.

The final crushing of the attenuator after the test is shown in **Fig. 11**. From the picture it is also evident the integrity of the structure not subjected to the energy absorption; only the top of the attenuator was subjected to the plastic absorption.

Together with the experimental evidence, the analysis of the attenuator under axial loading should be juxtaposed with the acceleration signal. **Fig. 12** shows the acceleration raw data versus time acquired by the accelerometer during the experimental test. Since the deceleration peak of the raw data is greater than 40 g, it was filtered with the Channel Filter Class (CFC) 60 (100 Hz), as required by the FSAE Rules. Also, the filtered data are shown in **Fig. 12**: the deceleration peak of the filtered data is about 31.1 g and the average deceleration is about 14.2 g.

The aim of the designer is to reproduce also numerically the crushing phenomenon. **Fig. 13** shows the numerical final deformation of the impact attenuator. It is evident the plastic folding at the top of the structure and a more intact configuration on the bottom of the same, as in real test.

The acceleration signal versus time can be recorded during simulation thanks to output controls. **Fig. 12** reproduces also the variation of numerical accelerations in time, and it can be observed how such trend is very close to the experimental case, except for small differences. It is due to the simplifications adopted in the numerical model compared to the real structure.

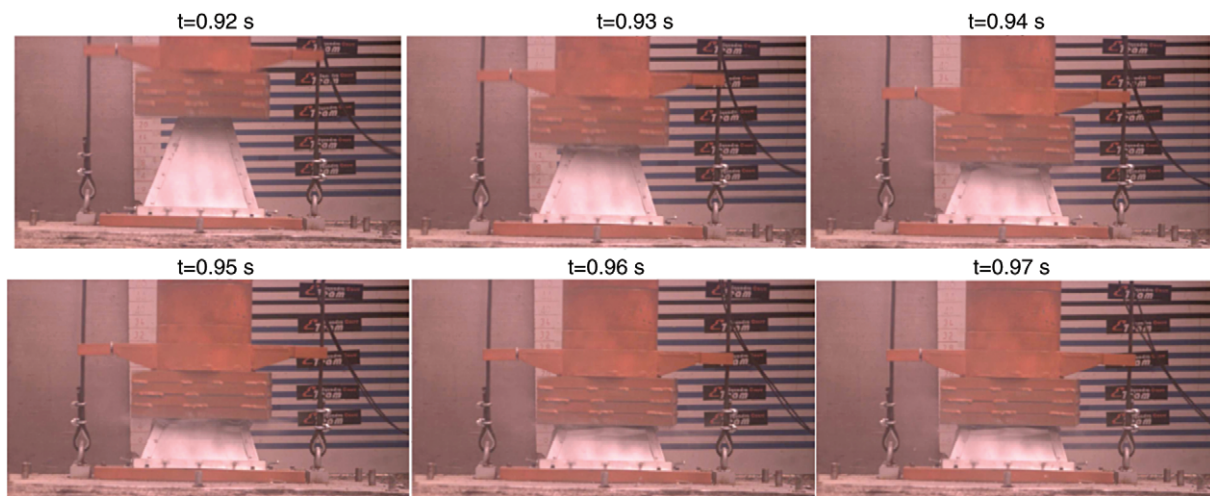


Figure 10 Frames of the impact every 10 ms (impact starts at time 0.92 s).

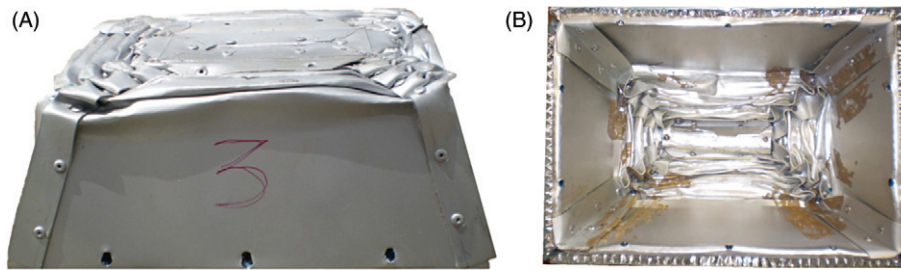


Figure 11 Impact attenuator after test: (A) external and (B) internal view.

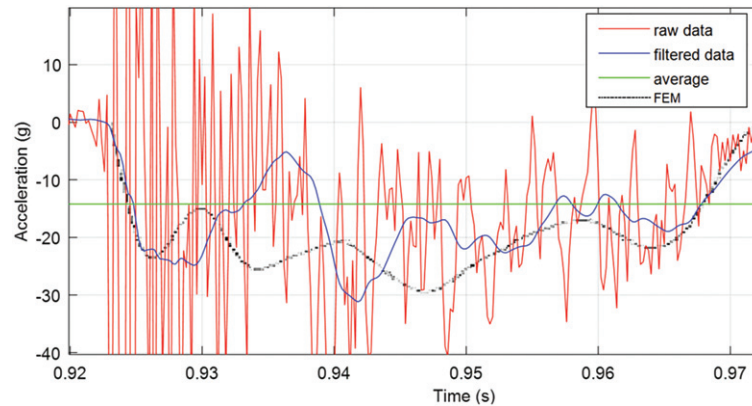


Figure 12 Acceleration-time: experimental results (raw data, filtered data, average value) and FEM.

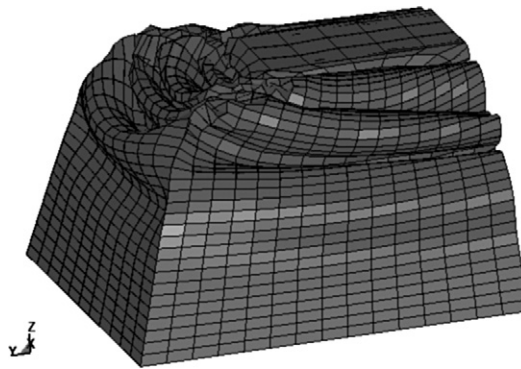


Figure 13 Final deformation of the impact attenuator from the numerical point of view.

The following table summarizes the main results of the experimental and numerical tests for such an impact attenuator with a truncated pyramidal structure (**Table 1**).

It can be concluded that the designed impact attenuator is able to satisfy the specific homologation requirements of the FSAE, therefore can be used in front of racing car bulkhead during competition.

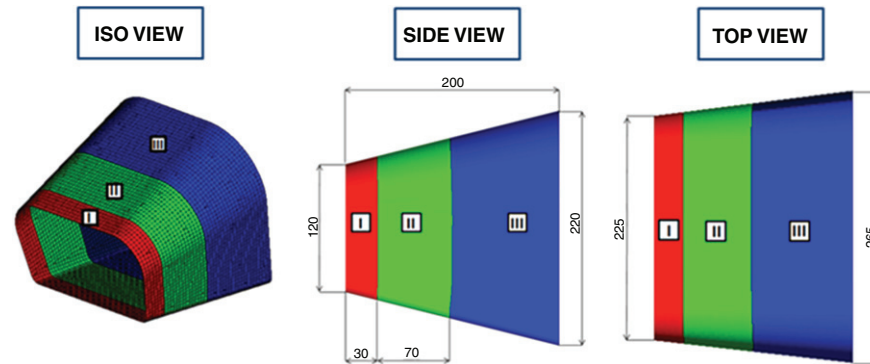
Design of an Impact Attenuator Made of Composite Material

Impact Attenuator Definition

In literature it is possible to find several studies ([Bisagni et al., 2005](#); [Heimbs et al., 2009](#); [Obradovic et al., 2012](#); [Wang et al., 2016](#)) on the design of impact attenuators in composite material, even if such research is still less respect to conventional materials and refers mainly to simple structures.

Table 1 Results of the experimental and numerical tests.

	<i>Test</i>	<i>FEM</i>
Time at the beginning of impact	0.922 s	0 s
Time at the end of the impact	0.972 s	0.048 s
Average deceleration	14.2 g	15.9 g
Peak deceleration	31.1 g	28.6 g
Final length of the attenuator	0.109 m	0.114 m

**Figure 14** Geometry of the impact attenuator.

The objective of this section is to present the results of a numerical and experimental campaign conducted on a CFRP structure in order to pass homologation test for FSAE requirement (SAE, 2016). The structure was manufactured using a carbon-epoxy preimpregnated fabric material (Obradovic et al., 2012). As regards the geometry a truncated pyramidal structure with rounded edges and various wall inclinations was chosen (Fig. 14).

Along the height of the structure three different zones were created varying only the wall thickness; such choice was dictated by the necessity to ensure a stable collapse, reduce the peak load and lighten the structure. The manufacture was obtained by hand lay up of preimpregnated composite layers and autoclave curing at 135°C and 7 bar applied pressure. Care was put on lamination process; it was avoided the overlap of the laminae at corner edges to guarantee the progressive crushing avoiding catastrophic failure under axial loading.

Experimental Dynamic Tests

The crushing behavior of the front impact attenuator was analyzed dynamically. The instrumentation consisted of a drop weight machine (Fig. 15A). Impact accelerations and velocities were acquired by a FA3403 tri-axial accelerometer with $\pm 500g$ full scale and a E3S-GS3E4 photocell, respectively (Fig. 15B and C). Moreover, a Mikrotron high-speed camera with a sampling of 1000 frames/s was used (Fig. 15D). In order to respect the FSAE regulation, the mass was fixed to 300 kg and the impact velocity to 7 m/s.

Finite Element Modeling

During the crashworthiness design is necessary to join the experimental campaign with the numerical modeling. Therefore specific FE models were done in order to reproduce, as faithfully as possible, the crushing behavior for the CFRP impact attenuator. The geometry was modeled following both the micro- and macromechanical approach. It means the implementation of solid elements plies with cohesive interface for the first methodology (Boria et al., 2015) and single shell layer model for the other one (Fig. 16).

The composite constitutive models implemented in LS-DYNA code are continuum mechanics ones. Composites are, therefore, modeled as orthotropic linear elastic materials within a failure surface. The exact shape of the failure surface is strictly dependent on the failure criterion used in the model. Beyond the failure surface, the appropriate elastic properties are degraded according to the assigned degradation laws, that can be divided into two main categories: progressive failure (for the material cards MAT 22, 54/55, 59) or continuum damage (MAT 58, 161, 162). In literature, it was shown how the progressive failure models can reproduce faithfully the brittle fracture of the composites under axial loading (Bisagni et al., 2005; Feraboli, 2008); it is therefore decided, for this study, to use the linear-elastic model *MAT_ENHANCED_COMPOSITE_DAMAGE. This material uses the Chang–Chang failure criterion (Chang and Chang, 1987) to determine individual ply failure. The deletion of the element occurs when all the layers fail.

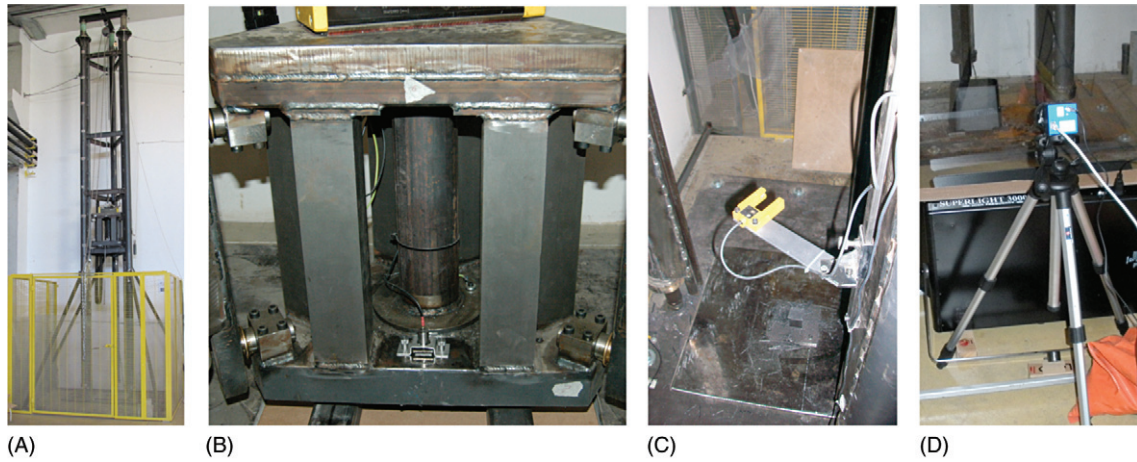


Figure 15 Drop tower and instrumentation.

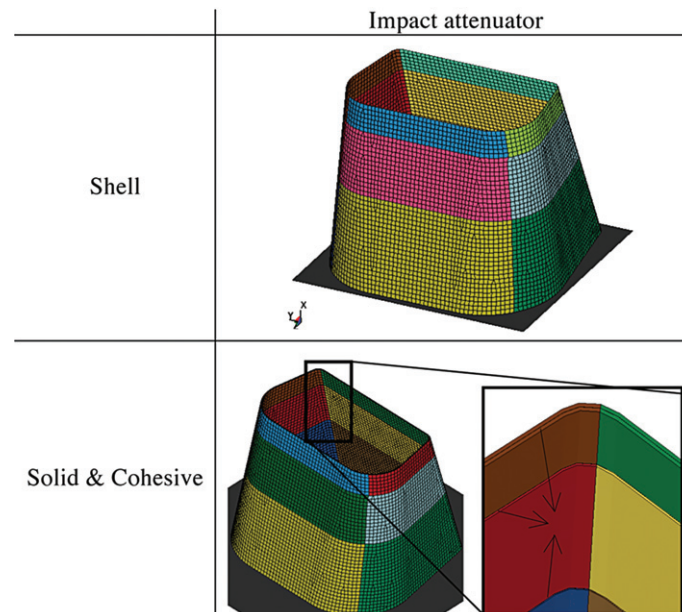


Figure 16 Different mechanical approaches used for impact attenuator modeling.

Elements which share nodes with the deleted element become “crashfront” elements and can have their strengths reduced by using the SOFT parameter (Boria and Belingardi, 2012) with TFAIL greater than zero.

With the single shell layer approach, the laminate lay-up can be defined by one integration point for each single ply with the respective ply thickness and fiber orientation angle. The card *PART_COMPOSITE permits to define this easily, defining the laminate thickness as the sum of each individual layer with the proper fiber orientation. Once all single layers of the shell element fail, the whole element is eroded, and it simply disappears from the calculation. Under-integrated shell elements of the Belytschko-Tsai type (ELFORM = 2) with the stiffness-based hourglass control (IHQ = 4) were used for the modeling. Such choice was justified by the negligible of the hourglass energy, that is less than 1% of the total energy.

The inter-laminar failure can be captured using a solid approach with cohesive elements. In particular *MAT_COHESIVE_MIXED_MODE model was used, that includes a bilinear traction-separation law with quadratic mixed mode delamination criterion and a damage formulation.

A “rigid wall planar moving forces” card with a finite mass of 300 kg and an initial velocity of 7 m/s was adopted for the impact attenuator. Master-surface to slave-node and self-contact between the impact mass and the nodes of the sacrificial structures and on the structure, respectively, were defined. As regards the boundary conditions, the same used for the real test were implemented on the numerical model.

Results and Discussion

Fig. 17 shows the impact attenuator before and after impact. A progressive crushing with the deformation confined at impact wall was observed.

The signal obtained by the acquisition was filtered with a CFC60 filter in order to appreciate the deceleration versus time trend more accurately (Fig. 18). The integration of the signal one and two times permits to obtain information about variation of velocity and crushing in time. Table 2 summarizes the principal crushing parameters, such as the effective initial velocity, the total impact time (T), the final stroke (δ), the average deceleration (d_{av}), the average load (F_{av}), and the absorbed energy (E_{abs}). The impact duration is the instant of time from the beginning of the crash to the moment when the velocity vanishes. The average deceleration is obtained from:

$$d_{av} = \frac{1}{T} \int_0^T d(t) dt, \quad (1)$$

where $d(t)$ is the progression of deceleration in time. The absorbed energy corresponds, instead, to the area under the load-shortening diagram.

Comparing the dynamic results with the static ones (Fig. 19), previously acquired, it is clear a loss of absorption capacity at higher load speed, even if lower than that observed for simplest structures (Boria, 2016b). A similar behavior in terms of force versus displacement in both cases is evident, with the presence of three peaks due to the wall thickness change, but the static condition ensures a greater capacity of energy absorption respect to the dynamic one. The reason may be due to three factors: the filtering of the

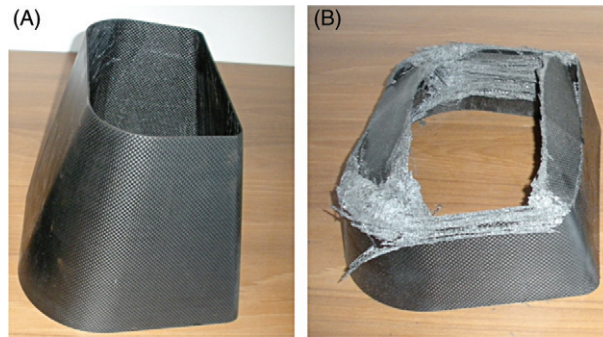


Figure 17 Attenuator before (A) and after dynamic test (B).

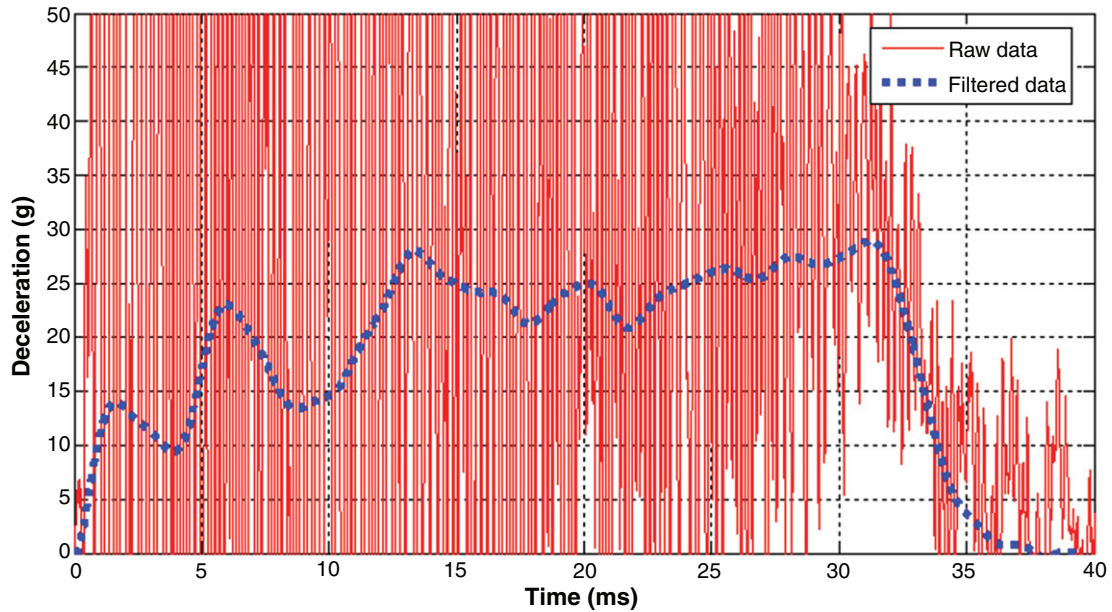
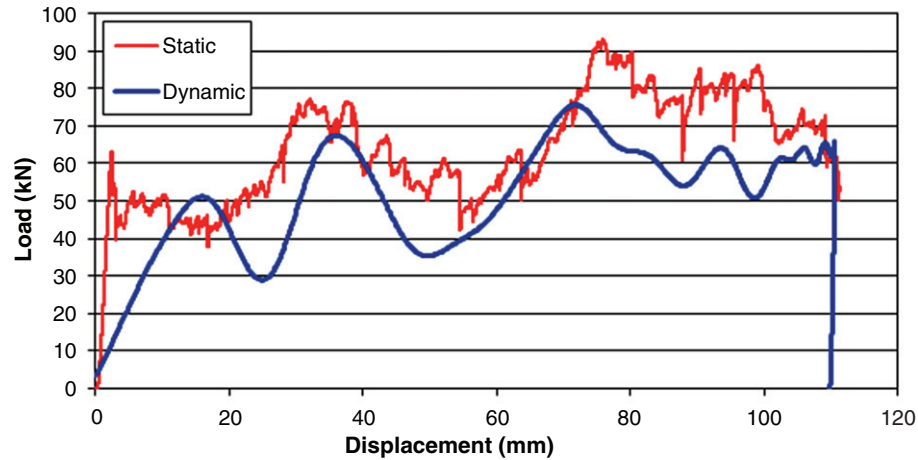


Figure 18 Raw data and filtered data for deceleration versus time.

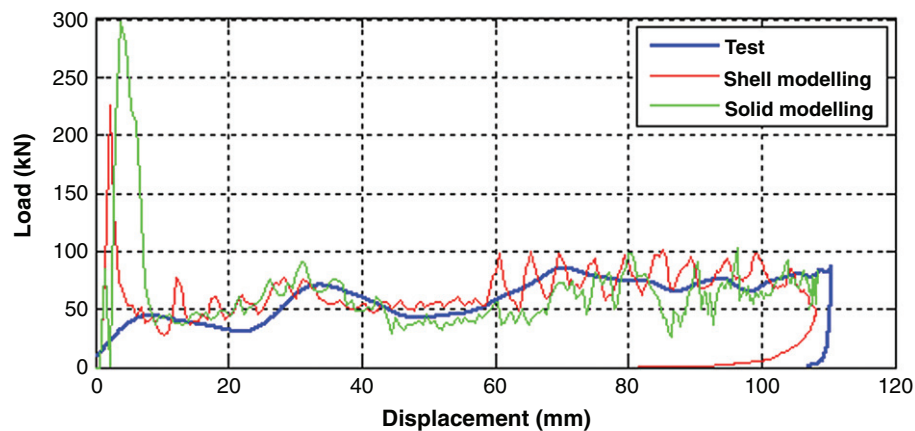
Table 2 Crushing parameters for impact attenuator under dynamic load.

Effective initial velocity (m/s)	Total impact time (ms)	δ (mm)	d_{av} (g)	F_{av} (kN)	E_{abs} (kJ)
6.8	32	112	21.6	63.5	6.9

**Figure 19** Experimental data under static and dynamic conditions for the impact attenuator.

signal, a possible strain rate dependence of the coefficient of friction and a different friction coefficient for the crushing surfaces used in the two tests (Farley, 1992; Feraboli et al., 2007; Hull, 1991; Mamalis et al., 1997B).

A comparison between numerical and experimental results was conducted using both shell and solid modeling. Fig. 20 shows such comparison in terms of load versus displacement curves under dynamic conditions. The trend highlights a good capability to capture the experimental crushing process of such structure, even though the complexity and heterogeneity of the CFRP composite material used. Such correspondence is also evident in the final crushing shape. Fig. 21A shows how the use of solid modeling can reproduce the sharp break of the laminate in a specific zone of the impact attenuator, while Fig. 21B shows that by using shell elements it is also possible to reproduce the flexion of the straighter wall under axial compression. The combination of solid and cohesive elements allows to reproduce some inter-laminar fracture phenomena that are not reproducible by using shell modeling. Unfortunately, a better crushing replication involves an increase into computing time, often not acceptable for complex systems.

**Figure 20** Force versus displacement diagram varying numerical modeling.

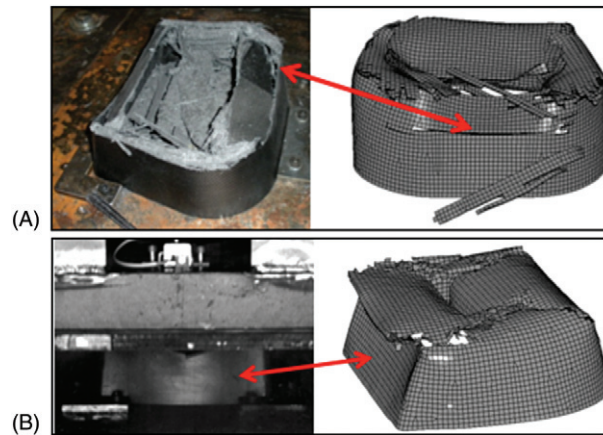


Figure 21 Experimental versus numerical deformation using (A) solid and (B) shell modeling.

Conclusions

In this article, the main crashworthiness aspects to consider during the design of an impact attenuator were presented. In particular, the attention was focused on two different materials, conventional and composite, particularly widespread into structural components. Their usage for specific attenuators was analyzed both experimentally and numerically.

Initially, a new impact attenuator design for a racing car, based on an aluminum sandwich structure was presented. Experimental test and numerical simulations of the axial dynamic crushing of such truncated structure were carried out. The results showed that the designed impact attenuator met the criteria for energy absorption, as stated by FSAE Rules, during the physical crash testing. As expected, the final deformation was characterized by plastic buckling with a series of folds that proceed from the top to the bottom of the impact attenuator with a progressive crushing. Despite the simplification adopted, the numerical model reproduced sufficiently well the crushing. A more precise numerical representation could better copy final deformation with a consequently growth of simulation times. Therefore, the numerical modeling procedure outlines in this work could be used with confidence for designing the proposed lightweight impact attenuator, using such thin-walled sandwich structure.

After, the efforts for the lightweight design and crash analysis of a specific CFRP, the frontal impact attenuator for a Formula SAE racing car were presented. A comparison with numerical modeling was done, both in terms of micro- and macro-mechanical approaches. In both cases the FE models were able to reproduce sufficiently well the crushing phenomena, despite the complexity and heterogeneity of the CFRP material used. Therefore these methodologies adopted have proven themselves accurate enough to design a specific impact attenuator made of conventional or composite material under dynamic axial loadings.

Acknowledgments

The author wishes to express her appreciation to Prof. Giovanni Belingardi, Ing. Alessandro Scattina, Ing. Jovan Obradovic of Politecnico di Torino for their support and advice. Moreover, the availability of Picchio SpA drop tower used for the experimental dynamic tests is gratefully acknowledged.

References

- Barnes, G., Coles, I., Roberts, R., Adams, D.O., Garner, D.M.J., 2011. *Crash Safety Assurance Strategies for Future Plastic and Composite Intensive Vehicles (PCIVs)*. Volpe National Transportation Systems Center, Cambridge, Massachusetts, USA.
- Belingardi, G., Obradovic, J., 2010. Design of the impact attenuator for a formula student racing car: numerical simulation of the impact crash test. *J. Serb. Soc. Comput. Mech.* 4 (1), 52–65.
- Belingardi, G., Chiandussi, G., 2011. *Vehicle Crashworthiness Design – General Principles and Potentialities of Composite Material Structures*. Springer.
- Bisagni, C., Di Pietro, G., Frascini, L., Terletti, D., 2005. Progressive crushing of fiber-reinforced composite structural components of a formula one racing car. *Comp. Struct.* 68, 491–503.
- Boria, S., 2010. Behaviour of an impact attenuator for Formula SAE car under dynamic loading. *Int. J. Vehicle Struct. Syst.* 2 (2), 45–53.
- Boria, S., Belingardi, G., 2012. Numerical investigation of energy absorbers in composite materials for automotive applications. *Int. J. Crashworth.* 17 (4), 345–356.
- Boria, S., 2013. Design solutions to improve CFRP crash-box impact efficiency for racing applications. In: Elmarakbi, A. (Ed.), *Advanced Composite Materials for Automotive Applications*. Wiley, pp. 205–226.
- Boria, S., Pettinari, S., 2014. Mathematical design of electric vehicle impact attenuators: metallic vs composite material. *Comp. Struct.* 115, 51–59.
- Boria, S., Obradovic, J., Belingardi, G., 2015. Experimental and numerical investigations of the impact behavior of composite frontal crash structures. *Comp. B* 79, 20–27.
- Boria, S., 2016a. Numerical and experimental analysis of an impact attenuator in sandwich material for racing application. *Int. J. Sci. Eng. Technol. Res.* 5 (6), 1909–1913.

- Boria, S., 2016b. Lightweight design and crash analysis of composites. In: Njuguna, J. (Ed.), *Lightweight Composite Structures in Transport: Design, Manufacturing, Analysis and Performance*. Woodhead Publishing, pp. 329–359.
- Chang, F.K., Chang, K.Y., 1987. A progressive damage model for laminated composites containing stress concentration. *J. Comp. Mat.* 21, 834–855.
- Chirwa, E.C., Latchford, J., Clavell, P., 2003. Carbon skinned aluminium foam nose cones for high performance circuit vehicles. *Int. J. Crashworth.* 8, 107–114.
- Davies, G., 2003. *Material for Automobile Bodies*. Elsevier, G. London.
- Farley, G.L., 1992. Relationship between mechanical-property and energy-absorption trends for composite tubes. NASA TP 3284, ARL-TR-29, Technical report, United States.
- Feraboli, P., Masini, A., 2004. Development of carbon/epoxy structural components for a high performance vehicle. *Comp. B Eng.* 35 (4), 323–330.
- Feraboli, P., Norris, C., McLarty, D., 2007. Design and certification of a composite thin-walled structure for energy absorption. *Int. J. Vehicle Des.* 44 (3/4), 247–267.
- Feraboli, P., 2008. Development of a corrugated test specimen for composite material energy absorption. *J. Comp. Mat.* 42 (3), 229–256.
- Fuchs, E.R.H., Field, F.R., Roth, R., Kirchain, R.E., 2008. Strategic materials selection in the automobile body: economic opportunities for polymer composite design. *Comp. Sci. Tech.* 68 (9), 1989–2002.
- Garner, D.M., Adams, D.O., 2008. Test methods for composites crashworthiness: a review. *J. Adv. Mat.* 40 (4), 5–26.
- Ghasemnejad, H., Hadavinia, H., Aboutorabi, A., 2010. Effect of delamination failure in crashworthiness analysis of hybrid composite box structures. *Mat. Desig.* 31 (3), 1105–1116.
- Guimard, J.-M., Allix, O., Pechnik, N., Thevenet, P., 2009. Energetic analysis of fragmentation mechanisms and dynamic delamination modelling in CFRP composites. *Comput. Struct.* 87 (15–16), 1022–1032.
- Guoxing, L., Tongxi, Y., 2003. *Energy Absorption of Structures and Materials*. Woodhead Publishing, Cambridge.
- Hanssen, A.G., Langseth, M., Hopperstad, O.S., 2000. Static and dynamic crushing of circular aluminium extrusions with foam filler. *Int. J. Impact Eng.* 24, 475–507.
- Heimbbs, S., Strobl, F., Middendorf, P., Gardener, S., Eddington, B., Key, J., 2009. Crash simulation of an F1 racing car front impact structures. In: *Proceedings of the 7th European LS-DYNA Conference*.
- Hsu, S.S., Jones, N., 2004. Quasi-static and dynamic axial crushing of thin-walled circular stainless steel, mild steel and aluminium alloy tubes. *Int. J. Crashworth.* 9, 195–217.
- Hwang, W.C., Cha, C.S., Yang, I.Y., 2011. Optimal crashworthiness design of CFRP hat shaped section member under axial impact. *Mat. Res. Innovat.* 15 (s1), s324–s327.
- Hull, D., 1991. A unified approach to progressive crushing of fibre-reinforced composite tubes. *Comp. Sci. Technol.* 40, 377–421.
- Imanullah, F., Ubaidillah, Prasajo, A.S., Wirawan, A.A., 2018. Experiment evaluation of impact attenuator for a racing car under static load, In: *Proceedings of the AIP Conference* 1931, 030047.
- Jacob, G.C., Fellers, J.F., Simunovic, S., Starbuck, J.M., 2002. Energy absorption in polymer composites for automotive crashworthiness. *J. Comp. Mat.* 36, 813–850.
- Johnson, A.F., David, M., 2010. Failure mechanisms in energy-absorbing composite structures. *Philosop. Mag.* 90 (31–32), 4245–4261.
- Jumahat, A., Soutis, C., Jones, F.R., Hodzic, A., 2010. Fracture mechanisms and failure analysis of carbon fibre/toughened epoxy composites subjected to compressive loading. *Comp. Struct.* 92 (2), 295–305.
- Langseth, M., Hopperstad, O.S., Berstad, T., 1999. Crashworthiness of aluminium extrusions: validation of numerical simulation, effect of mass ratio and impact velocity. *Int. J. Impact Eng.* 22, 829–854.
- Li, Z., Yu, Q., Zhao, X., Yu, M., Shi, P., Yan, C., 2017. Crashworthiness and lightweight optimization to applied multiple materials and foam-filled front end structure of auto-body. *Adv. Mech. Eng.* 9 (8), 1–21.
- Lukaszewicz, D.H.-J.A., 2014. Automotive composite structures for crashworthiness. In: Elmarakbi, A. (Eds.), *Advanced Composite Materials for Automotive Applications-Structural Integrity and Crashworthiness*, pp. 99–127.
- Mamalis, A.G., Robinson, M., Manolakos, D.E., Demosthenous, G.A., Ioannidis, M.B., Carruthers, J., 1997a. Crashworthy capability of composite material structures. *Comp. Struct.* 37 (2), 109–134.
- Mamalis, A.G., Manolakos, D.E., Demosthenous, G.A., Ioannidis, M.B., 1997b. Analytical modelling of the static and dynamic axial collapse of thin-walled fiberglass composite conical shells. *Int. J. Impact Eng.* 19 (5–6), 477–492.
- Mamalis, A.G., Manolakos, D.E., Demosthenous, G.A., Ioannidis, M.B., 1998. *Crashworthiness of Composite Thin-walled Structures*. CRC Press, Boca Raton.
- Mamalis, A.G., Manolakos, D.E., Ioannidis, M.B., Papapostolou, D.P., 2005a. On the response of thin-walled CFRP composite tubular components subjected to static and dynamic axial compressive loading: experimental. *Comp. Struct.* 69 (4), 407–420.
- Mamalis, A.G., Manolakos, D.E., Ioannidis, M.B., Papapostolou, D.P., 2005b. On the experimental investigation of crash energy absorption in laminate splaying collapse mode of FRP tubular components. *Comp. Struct.* 70 (4), 413–429.
- Meredith, J., et al., 2012. Recycled carbon fibre for high performance energy absorption. *Comp. Sci. Technol.* 72 (6), 688–695.
- Munusamy, R., Barton, D.C., 2010. Lightweight impact crash attenuators for a small Formula SAE race car. *Int. J. Crashworth.* 15 (2), 223–234.
- Ning, H., Janowski, G.M., Vaidya, U.K., Husman, G., 2007a. Thermoplastic sandwich structure design and manufacturing for the body panel of mass transit vehicle. *Comp. Struct.* 80 (1), 82–91.
- Ning, H., Vaidya, U.K., Janowski, G.M., Husman, G., 2007b. Design, manufacture and analysis of a thermoplastic composite frame structure for mass transit. *Comp. Struct.* 80 (1), 105–116.
- Ning, H., Pillay, S., Vaidya, U.K., 2009. Design and development of thermoplastic composite roof door for mass transit bus. *Mat. Design* 30 (4), 983–991.
- Obradovic, J., Boria, S., Belingardi, G., 2012. Lightweight design and crash analysis of composite frontal impact energy absorbing structures. *Comp. Struct.* 94 (2), 423–430.
- Rappolt, J.T., 2015. Analysis of a carbon fiber reinforced polymer impact attenuator for a Formula SAE vehicle using finite element analysis. Thesis presented to the Faculty of California Polytechnic State University.
- SAE, 2016. Formula SAE Rules 2016. Available from: www.fsaeonline.com.
- Santosa, S., Wierzbicki, T., 1998. Crash behavior of box columns filled with aluminum honeycomb or foam. *Comp. Struct.* 68, 343–367.
- Smith, G.F., 1990. Design and production of composites in the automotive industry. *Comp. Manufact.* 1 (2), 112–116.
- Thattaiaparthasarthy, K.B., Pillay, S., Ning, H., Vaidya, U.K., 2006. Design and development of a long fiber thermoplastic bus seat. *J. Thermop. Comp. Mat.* 19 (2), 131–154.
- Wang, J., Yang, N., Zhao, J., Wang, D., Wang, Y., Li, K., et al., 2016. Design and experimental verification of composite impact attenuator for racing vehicles. *Comp. Struct.* 141, 39–49.
- Yang, Y.Q., Nakai, A., Uozumi, T., Hamada, H., 2007. Energy absorption capability of 3d braided-textile composite tubes with rectangular cross section. *Key Eng. Mat.* 334-335, 581–584.
- Zarei, H., Kroger, M., 2006. *Optimum Honeycomb Filled Crash Absorber Design*. University of Hannover.

Further Reading

- Morello, L., Rosti Rossini, L., Pia, G., Tonoli, A., 2011. *The Automotive Body*, vols. I and II. Springer, ISSN: 0941-5122.
- Zarei, H., 2008. *Experimental and Numerical Investigation of Crash Structures Using Aluminum Alloys*. Curvillier Verlag, Göttingen.

ATV, Snowmobile, and Terrain Vehicle Safety

David P. Gilkey*, William Brazile[†], *Montana Technological University, Butte, MT, United States; [†]Colorado State University, Fort Collins, CO, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	77
ATVs/ROVs	77
Epidemiology of ATV-Related Injury and Fatality	78
Safety Measures	78
Rider Practices and Training	78
Vehicle Design	79
Laws	80
Vehicle Star Rating and Fit for Use Standards	80
Conclusions	81
List of Relevant Websites	81
Snowmobiles	81
Epidemiology of Snowmobile Injury and Fatality	82
Safety Measures	82
Rider Practices and Training	82
Design	82
Laws	83
Conclusions	83
Relevant Websites	83
References	84

Introduction

All-terrain vehicles (ATVs) and snowmobiles have become major forms of transportation on finished roads, dirt trails, and rough terrain. Both types of vehicles have been used for recreational and occupational purposes with increasing frequency. Currently, an estimated 25 million persons ride ATVs and 10 million persons ride snowmobiles annually in the United States and Canada. Of the 11 million ATVs and 1.2 million snowmobiles registered in the United States, many have been adapted for a variety of recreational and occupational uses and environments. Use in occupational environments include industry sectors, such as agriculture, logging, search and rescue, firefighting, land management, land exploration, oil and gas, mining, parks and recreation, hunting, law enforcement, military, field research, and others (Lagerstrom et al., 2015). It has been reported that 80% of the units sold, both ATV and snowmobiles are intended for recreational purposes and approximately 20% are used for occupational purposes.

In some parts of the world, such as Australia, ATVs may be referred to as “quadbikes,” where some 60,000 units are used heavily in agricultural operations. More recently, the introduction of the utility terrain vehicles (UTVs) provides users with a design alternative commonly referred to as side-by-side (SXS) vehicles, which allow riders to ride adjacent to each other on bench- or bucket-type seats. Conversely, ATVs require the rider to straddle the seat. The two distinctly different designed units comprise the larger group often identified as off-highway vehicles (OHVs) or recreational off-road vehicles (ROVs) or off-road vehicles (ORVs).

The popularity of ROVs and snowmobiles continue to climb among global consumers. Worldwide sales of ATVs/quadbikes were estimated to be more than 200,000 in 2014 while the more popular UTV sales reached 400,000 in 2016. More than 120,000 snowmobiles were purchased worldwide in 2018. The ROVs show increasing popularity and utility by meeting a variety of individual, group, and/or business transportation needs. The ROV transportation market has been estimated to reach \$8 billion in sales by 2024. The annual North American economic impact of ROVs and snowmobiles is estimated at \$66 billion and \$28 billion, respectively. These economic impacts are very positive to local communities by increasing tourism related lodging and restaurant services in addition to fuel, permits, and tax revenues.

ATVs/ROVs

ATVs arrived in the United States in the early 1970s. The vehicles had immediate application in the agricultural sector and quickly became appealing to recreational riders. The first models were three-wheeled designs with low inflation balloon tires, light weight, and steered with motorcycle handlebars. Sales soared, as did vehicle-related injuries and fatalities, catching the attention of many consumer groups. In the early 1980s, the industry manufacturers negotiated with the Consumer Product Safety Commission (CPSC)



Figure 1 Four-wheeled ATV. Source: Photo provided by Quadbar Industries.

to cease production of the three-wheeled units and were replaced by the more stable four-wheeled vehicles (quadbikes) known and used today (**Fig. 1**) ([GAO, 2010](#)). Many three-wheeled units still exist and continue to be used despite the known risks. The ATV industry continued to evolve designs resulting in the latest UTVs, also named recreational utility vehicle (RUVs), SXS, and side-by-side-vehicles (SSVs). The UTVs and ATVs are similar in some regards but differ in fundamental design attributes, such as a steering wheel instead of a motorcycle handlebars, foot pedals in place of hand levers for throttle and brake control, and bench or bucket seats replaced the traditional ATV straddle seating. A significant difference in stability stems from a wider frame design with a resulting stability that is less, or not affected by, passenger movements, which was not the case with the relatively more unstable ATVs. In addition, ATV passengers presented an increased risk for loss of control due to the added weight above the center of gravity, causing additional instability. Passengers on ATVs are not recommended as the ATV was designed for single rider use only.

Sales of UTVs have been nearly twice that of ATVs due to their enhanced stability, perceived safety, load capacity, and ability to carry more than one passenger. ATVs and UTVs now come in a variety of sizes and power ranges to accommodate user preference. The ATV sized for children are designed much smaller, weighing approximately 90 kg (200 lb) and are powered by a small 50 cc motor. Today's adult-sized ROVs may weigh as much as 680 kg (1500 lb) and are capable of speeds of more than 120 km/h (80 mi/h).

Epidemiology of ATV-Related Injury and Fatality

The CPSC started investigating ATV accidents in 1982 and has since then investigated more than 15,000 ATV/ROV-related fatalities. Major risk factors associated with loss-of-control events, collisions, injuries, and fatalities included consumption of alcohol, not wearing a helmet, riding on paved roads, carrying a passenger, excessive speed, riding up hills, down hills or cross-slope, and/or striking stationary or mobile objects. ATV/ROV sales peaked in the mid-2000s, as did injuries and fatalities. It has been estimated that more than 400,000 ROV riders are injured each year and 100,000 emergency room visits are associated with ROV accidents and loss-of-control events ([CPSC, 2019](#)). The growing use of UTVs has slowed the number of ROV fatalities resulting in a frequency ratio of 1:4, UTV to ATV (quadbike). The UTVs have wider wheelbases and are equipped with seatbelts and a roll-cage that creates a survival space, thus, providing greater protection for the user in the event of a rollover accident. Data compilation on UTV safety and risk factors continues to grow. In recent times, a great number of product recalls have occurred due to safety concerns, such as loss of control, fire, and burn hazards. For example, in 2017, the John Deere Crossover Gator utility vehicle was recalled due to a crash hazard. It was discovered that the steering shaft could separate from the steering-rack assembly resulting in loss of control. The recall affected 68,000 UTVs. The CPSC continues to investigate all ROV-related fatalities and provides publicly accessible reports to keep consumers and other interested parties informed. As the statistical safety data grow, more information about the risk factors associated with ROV injury and fatality becomes available.

Safety Measures

Rider Practices and Training

There exists an overriding belief that operating ROVs is risk interactive. This belief presumes that the operator can both increase and decrease the risk for loss-of-control events based upon choices and actions within his or her control. As a consequence of this belief, great emphasis has been placed on rider education and preparation through training and education to reduce risks and hazards associated with the vehicle's operation.

The ATV (quadbike) manufacturing industry agreed to provide rider training and formed the ATV Safety Institute (ASI) as the educational arm of the Specialty Vehicle Institute of America in the 1980s. The agreement between the ASI and CPSC required that with each new ATV sale, that the purchase be accompanied with subsidized access to the ASI ATV safety RiderCourse. The new owner would be able to use the worldwide web to identify a trainer in his or her area and procure the training, essentially free of charge. The high-quality RiderCourse was to be taught by ASI certified trainers and included a 5-h, hands-on field course to effectively enhance rider awareness of associated hazards and risks, practice skills for safer riding, reduce risks, and prevent loss-of-control events. The RiderCourse ATV safety training includes information and experiential learning about rider personal protective gear, pre-ride check, starting, stopping, turning, climbing grades, descending grades, cross-slope safety, as well as managing obstacles.

In 2017, the ASI reported that about 1 million riders had completed the training. Numerous user groups, both recreational and occupational, have advocated and even required the ASI RiderCourse for their members. In the United States, many private sector employers and government agencies require the RiderCourse before initial use and as a refresher training every 3 years for workers that use the vehicles in their job duties, and the number of employers requiring the RiderCourse training is increasing. However, very few studies support the success of rider training and additional research is warranted for the evaluation of training effectiveness. At this time, the effect of ATV safety training on post-training riding behaviors is unknown.

A large number of ATV youth-riding organizations and other types of user groups have produced an array of educational materials and training experiences to increase the safety awareness related to ROV selection, rider gear, and vehicle use. Many agricultural-focused education and research units produce ROV safety information sheets and other forms of handout materials that are disseminated person-to-person, by mail, and on the worldwide web. Most of the safety messaging appears to have been developed with the basic ASI tenants and content. Messaging commonly includes the following key items:

1. Always wear a helmet
2. Take the ASI hands-on training RiderCourse
3. Refer to the owner's manual for safe operation and sizing
4. Complete a pre-ride vehicle check before every use
5. Wear the proper riding gear—gloves, long-sleeve shirt, full-length pants, boots above the ankles, face shield or goggles, and three-fourths or full-face helmet
6. Choose the age appropriate vehicle for the rider
7. One rider only
8. Ride on permitted trails or lands only
9. Do not drink alcohol

Newer web-based interactive training is also available through a number of organizations as well as the ASI. The common goals of web-based ASI training and other ROV web-based training is to increase awareness of the risks and hazards associated with ROV use and to offer information about methods to mitigate known risks and hazards. Another major goal is to encourage new riders to take the ASI hands-on RiderCourse for the experiential learning.

Vehicle Design

Design improvements have been at the center of the ROV evolution and revolution. The migration from three-wheeled to four-wheeled ATVs and now to UTVs with rollover protection and seat belts have been major design changes driven by the demand for safer vehicles. Two different organizations developed separate voluntary safety standards for ROVs. The Recreational Off-Highway Vehicle Association (ROHVA) co-developed ANSI/ROHVA 1, American National Standard for Recreational Off-Highway Vehicles, and the Outdoor Power Equipment Institute (OPEI) co-developed ANSI/OPEI B71.9 (2012), American National Standard for Multipurpose Off-Highway Utility Vehicles. The two industry-based standards have been progressive and evolved with the growth and responsibility of the industry. The CPSC sponsored the 16 CFR Part 1422, Requirements For All Terrain Vehicles, that was adopted in 2008 to address consumer protection, information, labeling, recreation, and recreation areas pertaining to ATVs and ROVs and the import restriction of three-wheeled models.

In 2014, the CPSC proposed a design standards rule for ROVs to address continuing safety concerns. The rule did not pass, therefore leaving in place the current ANSI/ROHVA 1 for recreational ROVs and ANSI/OPEI B71.0 for utility-oriented ROVs voluntary standards (CPSC, 2014). These standards address vehicle design, configuration, and performance aspects of ROVs. Specific requirements exist for accelerator and brake controls, service and parking brake/parking mechanisms, lateral and pitch stability, lighting, tires, handholds, occupant protection system, labels, and owner's manuals. The CPSC advocated for enhanced standards related to lateral load shift stability, tilt stability, dynamic stability, steering systems, seatbelt reminder systems, and occupant retention systems. The CPSC had proposed a speed limit of 15 mi/h, if the seatbelt was not fastened. The belief was that such a limit would motivate the operator to buckle-up. Standards will continue to evolve in response to social, political, technological, and economic forces. In the meantime, retrofitting ATVs/quadbikes with rollover protective devices is a less expensive design change that has promised for reducing fatalities associated with rollover accidents.

Rollover protection systems (ROPS) or crush protection devices (CPDs) became increasingly popular in New Zealand and Australia over the past 2 decades. A number of products were developed with variations in design and materials coupled with rebate programs to incentivize retrofitting vehicles already in use. Under test conditions, the ROPS were estimated to reduce



Figure 2 ATV fitted with a Quadbar. *Source: Photo provided by Quadbar Industries.*

crash-related fatality from rollovers by approximately 40%. The ROPs, when attached to the ATV, is fitted to the vehicle behind the rider creating a survival space in the event of a rollover thus, preventing crushing by the vehicle. In 2017, New Zealand authorities recommended that ROPs be required for all quadbikes. The ROPs have not had the same level of acceptance in the United States (Fig. 2).

Laws

Various laws, statutes, rules, and regulations are pervasive, vary by state, county, or city and generally require registration at a minimum. Many states allow ROVs on roadways used by automobiles, trucks, and other vehicles. Some states allow licensing equal to other vehicles and permit full access to most road surfaces and conditions except freeways or interstate roadways.

Safety-related laws are highly variable and may include mandatory helmet use, safety education certification, age restriction for operation, headlamps, brake lighting, turn signals, passenger prohibition, width, weight and displacement limits, wheel/rim diameter limits, tire pressure requirements, wheel base requirements, braking capability, and reflectors/reflectorized material requirements. The legal landscape will continue to evolve as ROV users pressure policy makers to open access to roads and public lands while consumer safety groups and others campaign for greater restrictions to reduce injuries and save lives.

Vehicle Star Rating and Fit for Use Standards

A great deal of high quality research on ROV safety has been performed in Australia and New Zealand leading to a novel system for vehicle rating based upon its intended use (Grzebieta et al., 2015). The ATV or quadbike and SSV, "Vehicle Star Rating" (VSR) system, was developed by scientists and engineers working in the Transport and Road Safety (TARS) research center at the University of New South Wales. The research team crash-tested vehicles under various conditions to assess the probability for operator survival, considering in part, the survival space dimensions and the potential for injury. Based on the results of the performance tests, the team developed a rating scale that ranged from 0 to 5, with "5" assigned to those vehicles with the lowest risk of a rollover and superior safety performance. A rating of "0" indicates, in part, that the vehicle has no survival space and rollover results indicate high impact forces, whereas those vehicles with an ROPs, seat belts (3 point), and adequate survival space may warrant a rating of "3."

Authors of the VSR strived to reduce ROV related injury and fatality by properly classifying vehicle safety based upon intended use. The VSR applies to both the workplace and recreational environments thus increasing the number of informed consumer

choices in the selection of appropriate vehicles for a user's job or recreational needs. The TARS team also believes that the VSR will create competition among vehicle manufacturers to improve safety and seek higher ratings through enhanced design and technological changes and reduce the training and education burden currently on the operator.

Conclusions

There are known risks and hazards associated with ROV use, either recreationally or occupationally. The ROV epidemiologic data are overwhelming in identifying known risk factors associated with loss-of-control events and resulting injury and/or fatality. At minimum, riders should always wear a helmet and avoid alcohol use, if intending to operate an ROV. Safer use of ROVs may be achieved through training and education, engineering designs that maximize stability and safety performance, ROPs, regulatory protections through mandated standards, and selection of the appropriate vehicle for its intended use.

The evolution of ATVs to UTVs appears to be an effective step for ROV safety. The development of structural rollover protection, operator and passenger restraint systems, and enhanced stability, has proven to be positive steps in ROV safety to reduce the number of injuries and fatalities. The implementation of the VSR has only begun. In the long term, safety professionals are hopeful that the VSR "fit for use" system will become the consumer standard for ROV selection. It is also believed that the industry will respond by enhancing product design, engineering, and safety to provide enjoyment to riders with vehicles that meet the needs of a growing number of employers who use ROVs in their work practices.

List of Relevant Websites

- ATV Safety Institute: ATV eCourses: <https://atvsafety.org/atv-ecourse/>
- ATV Safety Institute: Hands on ATV Training Courses: <https://atvsafety.org/atv-ridercourse/>
- ATV Safety Institute: State ATV Laws and Requirements: <https://atvsafety.org/state-rules/>
- CDC and NIOSH: All-Terrain Vehicle (ATV) Safety at Work: <https://www.cdc.gov/niosh/docs/2012-167/default.html>
- Consumer Federation of America: Product Safety: <https://consumerfed.org/>
- Consumer Product Safety Commission: ATV Safety Kit: <https://www.cpsc.gov/safety-education/neighborhood-safety-network/toolkits/atv-safety>
- Consumer Product Safety Commission: ATVs and ROVs: <https://www.cpsc.gov/Research-Statistics/sports-recreation/atv-and-rovs>
- Consumer Product Safety Commission Proposed Design Standards Rule for ROVs: <https://www.federalregister.gov/documents/2014/11/19/2014-26500/safety-standard-for-recreational-off-highway-vehicles-rovs>
- High Plains Intermountain Center for Agriculture and Safety: ATV Safety: <http://csu-cvmb.colostate.edu/academics/erhs/agricultural-health-and-safety/Pages/atv-safety.aspx>
- OSHA: Hazards Associated with All-Terrain Vehicles (ATVs) in the Workplace: <https://www.osha.gov/dts/shib/shib080306.html>

Snowmobiles

Snowmobiles made their arrival in the United States in the early 1900s as motorized sleds. The modernized snowmobiles did not appear until the 1950s and quickly became immensely popular in the 1960s with the production of the lightweight, newly designed Ski Doo. The vehicle became a sensation and other manufacturers emulated the basic design of open cockpit, straddle seating, handlebar steering, and one or two seat versions. Annual sales skyrocketed peaking at an estimated 500,000 units sold in 1971. The highly versatile units found great success among the recreational users with adaptations to a number of work environments. By the early 2000s, four major manufacturers dominated the market. In 2018, an estimated 127,000 snowmobiles were sold worldwide with over 2 million registered units in the United States and Canada alone. Modern snowmobiles are designed with larger, more powerful engines ranging from 400 to 1000 cc, weighing nearly 272 kg (600 lb), and reaching speeds of 240 km/h (150 mi/h) with specialty machines capable of achieving speeds greater than 320 km/h (200 mi/h). Snowmobiles may be ridden on nearly any type of snowy terrain and are classified as ORVs as well. Their popularity adds significantly to many local economies.

It has been estimated that approximately \$2,000 per season per rider is spent (e.g., lodging, fuel, food, clothing, and equipment) in many areas with a continuing upward trend expected. The industry employs more than 100,000 individuals in North America generating in excess of \$30 billion of business in the United States and Canada. Snowmobiles have been essential vehicles for transportation of persons, supplies, and emergency response in snowbound areas throughout the most northern and southern latitudes. They have been adapted for many occupational uses, such as scientific exploration, fieldwork and sampling, materials and supplies transport, search and rescue, emergency response, enforcement, hunting, surveying, ranching, public utilities, winter sporting events, and user and passenger transportation from site-to-site when other vehicles are not able to perform. There are more than 3000 snowmobile clubs and organization worldwide. The snowmobile market is expanding into Europe and Russia, generating an estimated \$5 billion impact annually into the new sectors.

Epidemiology of Snowmobile Injury and Fatality

It has been estimated that more than 10 million people enjoy snowmobiling each year in North America where the average enthusiast rides her/his snowmobile 160–240 km (100–150 mi) per day trip and approximately 2012 km (1250 mi) per season. As snowmobiles became larger, heavier, and faster, there has been significant injury and fatality among the user population. Approximately 3% (>4000) of snowmobile riders in British Columbia, Canada, experience injuries and 36.5/100,000 are killed in events associated with vehicle use and loss-of-control events. In North America alone, it has been estimated that over 14,000 snowmobilers suffer accident-related injuries and 200 persons die annually. The estimated fatality rate in North America is 1.6 fatalities for every 100 million miles traveled $[(200 \text{ fatalities/year}) / (1250 \text{ miles traveled/rider/year} \times 10 \text{ million riders})]$. Injury events most commonly have been rollovers, collision with fixed or mobile objects, and submersion in lakes. Males are killed at twice the rate of females. Those riders between 21 and 30 years are the highest at-risk age group. Eighty percent of fatalities occurring in Canada happen off trail. Alcohol is the leading cause of loss-of-control events resulting in injury or death. Additional risk factors include not wearing a helmet, excessive speed, lack of experience operating a vehicle, lack of training, lack of good judgment, and operator error. For example, it takes at least 16 m (52 ft.) to stop a snowmobile traveling at 25 km/h (15 mi/h) and 83 m (272 ft.) for a vehicle traveling 72 km/h (45 mi/h), thus making accurate judgments about stopping are very important. Head injuries are the leading cause of death from snowmobile accidents. Common injuries include concussions, spine injuries, extremity fractures, lacerations, contusions, internal organ damage, and multisystem trauma. Extremities are the most common body part injured followed by the head and face, soft tissues, thoracic, spine, abdomen, and pelvis. Other types of injuries may occur and include burns, frostbite, hypothermia, and noise-induced hearing loss. Frostbite and hypothermia are significant risks as hazards are associated with the cold climate conditions. Snowmobiles make significant noise, and concerns have focused on preservation of quiet surroundings and protection of the riders. Hearing loss has been associated with long-term snowmobile exposure and the lack of hearing protection.

Safety Measures

Rider Practices and Training

The International Association of Snowmobile Administrators (IASA), the American Council of Snowmobile Associations (ACSA), and The International Association of Snowmobile Manufacturers Association (ISMA) support enhanced user awareness for snowmobile safety through an education program named, “Safe Riders.” This program has a number of emphasis points: no alcohol when riding, know your abilities, don’t exceed your abilities, know your machine’s capabilities, don’t exceed your machines capabilities, know your riding area, get a map, talk to local folks, learn as much as possible from organized rider groups, keep your machine in top running condition, complete a pre-ride check, follow the rules, take great care when crossing trails or roadways, dress properly, layer up, wear a helmet and goggles, gloves and boots, take a friend or ride in a group, and file a plan. Other key safety points include staying alert, becoming aware of darkness and low light conditions, be aware of water and possible ice failure, and unique mountain risks, hazards, and added dangers. If riders are interested in mountain trails and/or terrain, it is recommended to secure avalanche training when snowmobiling in mountainous areas and/or backcountry areas. The Safe Riders program also emphasize on caring for the environment and advocates staying on marked trails to preserve the habitat. For example, Minnesota maintains more than 200,000 miles of groomed trails for snowmobilers. This training program also includes training on the use of hand signals for important and common communications necessary to convey to other riders, such as stop, start, right and left turn, slowing, oncoming sleds, flowing sleds, last sled, and take care of the trail.

The Safe Riders pledge includes the following concepts and practices: I will never drink and ride, I will drive within the limits of my machine, obey the rules, be careful crossing roads and trails, keep my machine in top shape, complete a pre ride check before each ride, wear a helmet and appropriate gear, let my family or friends know my planned route and time schedule, and treat others and the out-of-doors with respect.

The American Pediatric Association has published recommendations similar to Safe Riders and added another policy warning against towing others on saucers or tubes. The ISMA also recommends using a global positioning system (GPS) when riding to ensure accurate navigation and signaling for emergency purposes.

Design

The Snowmobile Safety Certification Committee (SSCC) recently updated its safety standards for testing and design specifications in 2018. The industry organized and collaborated with the Society of Automotive Engineers (SAE) to create guidelines since 1970s. The current manufacturing and testing standards are voluntary and followed by manufacturers who sell in North America. The standards result in SSCC Certification Labeling on snowmobiles. Current SSCC certification requires adhering the SAE standard ICS1000 (Recreation Off-Road Vehicle Product Identification Numbering System) and its 18 subsections which cover definitions, lighting systems, brake systems, throttle controls, runaway prevention system, electrical systems, drive mechanisms, transmission guards, heat shields, fuel tank, operational sound levels of 78 dBA maximum, rider shields, handedness, occupant support system, capacity,

and vehicle classification. In addition, Class I type machines are designated as competitive whereas, Class II are designated for children. The standards include detailed testing requirements, such as acceleration, braking distance, and quality assurance. The voluntary standards have continued to evolve in response to political, economic, social, and technological forces. Vehicles sold outside of North America may not adhere to the same standards.

Laws

The rules, regulations, ordinances, and laws pertaining to snowmobiles are highly variable by region, state, county, and/or municipality. Half the states in the United States and all provinces of Canada have laws that regulate aspects of snowmobile ownership. For example, Alaska classifies snowmobiles as OHVs and has extensive laws governing their use. However, most jurisdictions require registration at minimum, and some states allow licensing similar to cars and require plates and OHV decals. Few states require helmets for operators and/or passengers whereas British Columbia requires helmets for both. Some states also require liability insurance for OHVs. Many groups are in favor of progressive licensing, restricting riding activities, times, and conditions until a youth becomes an adult, and a number of states legislate minimum age limits for riders ranging from 10 to 16 years. In addition, safety education certification is required in some jurisdictions to operate OHVs.

Some of the other variable rules, regulations, ordinances, and laws address traveling on roadways. In some states, roadway travel may require a valid driver's license and some allow snowmobiles on roads when traditional vehicles are not capable of operation due to snow and ice. In addition, some areas allow snowmobile travel alongside roadways with restrictions. While some states have speed limits, other states allow riders greater latitude by requiring "safe and reasonable" speeds adjusted to variable conditions. It is likely that laws will continue to change as consumers wish to expand their access to public lands, communities seek to relax restrictions on roadways, and agencies and other rider protection-oriented groups seek to strengthen safety laws to reduce the incidence of injury and fatality.

Conclusions

Snowmobiles are popular form of transportation enjoyed by many for recreational purposes as well as for essential workplace needs. Epidemiologic data have strongly related known risk factors for loss-of-control events to resulting injury or fatality. Further, sales of snowmobiles are rising as markets expand from North America to worldwide locations and as a result, injury and fatality are expected to rise. A few key safety messages remain consistent and strong:

1. Don't drink alcohol and ride
2. Don't drive too fast
3. Know your limits
4. Know the limits of the snowmobile
5. Wear the appropriate gear
6. Prepare for emergencies
7. Plan your ride
8. Let others know your plan
9. Follow the laws

The industry has collaborated uniquely to support safety and machine performance. Design and testing standards will continue to evolve. Safety training is offered by a great number of organized groups that consistently teach the tenets of the Safe Rider program. Countries, regions, states, and communities have implemented legislation to manage snowmobile use, and in many cases, the aspects of safety.

Relevant Websites

- ALASKA Snowmobile Safety Operations: http://www.snowmobileinfo.org/snowmobile-safety-docs/naoi-alaska-snowmobile-safety-operations_level-1-student-manual.pdf
- American Council of Snowmobile Associations (ACSA): <http://www.snowmobilers.org/>
- American Council of Snowmobile Association (ACSA) Hand Signals: <http://www.snowmobilers.org/snowmobiler-hand-signals.aspx>
- Canadian Council of Snowmobile Organizations: <http://www.ccsso-ccom.ca/en/homepage/>
- Canadian Safety Council: <https://canadasafetycouncil.org/product/snowmobile-operators-course/>
- Consumer Product Safety Commission (CPSC): <https://www.cpsc.gov/s3fs-public/541.pdf>
- International Association of Snowmobile Administrators (IASA): <https://www.snowiasa.org/>
- International Snowmobile Manufacturers Association (ISMA): <http://www.snowmobile.org/index.html>

- Official British Columbia Snowmobile Safety Course: <https://www.snowmobile-ed.com/britishcolumbia/>
- The Minnesota Snowmobile Safety Course: https://www.snowmobilecourse.com/usa/minnesota/?gclid=EAIaIQobChMI2t_guPW-3wIVA9VkCh3D5ADdEAAAYASAAEgIt-DBwE

References

- Consumer Product Safety Commission (CPSC), 2014. Safety standard for recreational off-highway vehicles (ROVs); Proposed rule 16 CFR Part 1422. Federal Register 79(223), 6894–6903. Available from: <https://www.federalregister.gov/documents/2009/12/22/E9-30378/standard-for-recreational-off-highway-vehicles>.
- Consumer Product Safety Commission (CPSC), 2019. 2017-annual report of ATV related death and injuries. Available from: https://www.cpsc.gov/s3fs-public/atv_annual%20Report%202017_for_website.pdf?qLMnEEqa.T8KS0dW0r8qGqpUC7gQbqEd.
- Government Accountability Office (GAO), 2010. All-terrain vehicles: How they are used, crashes, and sales of adult sized vehicles for children's use. CPSC. GAO-10-418. Available from: <https://www.gao.gov/assets/310/302950.pdf>.
- Grzebieta, R., Rechnitzer, G., McIntosh, A., Mitchell, R., Patton, D., Simmons, K., 2015. Investigation and analysis of quad bike and side by side vehicle (SSV) fatalities and injuries. University of New South Wales. Sydney, Australia: TARS. Available from: http://www.tars.unsw.edu.au/research/Current/Quad-Bike_Safety/Reports/Supplemental_Report_Exam&Analysis_Fatals&Injuries_Jan-2015.pdf.
- Lagerstrom, E., Gilkey, D., Elenbaas, D., Rosecrance, J., 2015. ATV-related injuries in Montana 2005–2012. Special issue—All-Terrain (ATVs, Quad Bikes) and Off-Highway (ROVs, UTVs, SSVs, LSVs, LUVs, MUVs, XUVs). *Safety* 1(1), 59–70; Doi: <http://dx.doi.org/10.3390/safety1010059>.

Automobile Safety Inspection

Subasish Das, Texas A&M Transportation Institute, College Station, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	85
Literature Review	85
State Laws and Effectiveness Measure	86
Vehicle Inspection Program	86
Data Sources	87
FARS Data	87
NHTSA Vehicle Complaint Data	87
Future Scope and Research Direction	88
Conclusion	88
Biography	89
Relevant Websites	89
References	89
Further Reading	89

Introduction

According to 2011–16 Fatality Analysis Reporting System (FARS), approximately 2.6% of fatal crashes in the United States involved vehicle-related faults as a primary contributing factor. Previous studies indicate that a majority of technological defects are identified during inspection of wearable parts such as lights, tires, and brakes. However, the proportion of vehicle failures that can be contributed to crash occurrences was revealed to not have changed fundamentally since the 1970s. In the United States, the National Highway Traffic Safety Administration (NHTSA) suggests that each state implement a program to inspect all registered vehicles (annually or biennially) to guarantee that unsafe vehicles are taken off the roads. Nonetheless, it is critical to acknowledge that studies lack any correlation between crash rates and inspection programs related to vehicle component failure. When legislation restricted the NHTSA's authority to reserve highway funding in 1976, the number of states with vehicle inspection programs has decreased. As of July 2015, only 16 states continue to implement the recommended regular safety inspection program.

Safety inspection regulations in Canada vary from province to province. The Ministry of Justice in Canada provides a detailed regulation in the document "Motor Vehicle Safety Act" (The Ministry of Justice, Canada, 2019). According to the directive 2014/45/EU of the European Parliament and of the Council on "periodic roadworthiness tests for motor vehicles and their trailers and repealing Directive 2009/40/EC" (issued on April 3, 2014), "all member states need to carry out periodic safety inspections for most types of motor vehicles with designed speed exceeding 40 kmph/25 mph" (Official Journal of the European Union, 2014). In Australia, each state or territory preserves the authority to set its own safety inspection regulations. Poorly designed and old vehicles contribute significantly to traffic deaths in the Asian countries. Automobile safety inspection programs are not effectively regulated through design standards or maintained through mandatory vehicle inspection schemes (World Resources Institute, 2018).

While vehicle safety inspections could likely decrease the possibility of crashes, the degree of the reduction is difficult to quantify. Earlier studies do not implicate that safety inspection programs have zero effect on the reduction of crashes; however, the results are inconclusive. Additionally, international studies were unable to determine a relation between inspection programs and crash rates. Conclusively, there is an absence of research on the effectiveness of vehicle safety inspections on crash reduction. This study provides a short overview on this topic with inclusion of research directions and future scopes.

Literature Review

Limited research has been conducted on the effects of automobile safety inspection programs. One earlier study (conducted in 1999) recorded the results of several earlier studies. Using various sources, the previous studies (conducted before 1995) displayed contradictory findings about the significance of automobile safety inspections. Using panel data from all US states during the years of 1981–93, this study found no evidence that inspections considerably decreased injury or fatality rates. The Federal Motor Carrier Safety Administration has revoked an arrangement for commercial drivers to lodge inspection reports if vehicle deficiencies or flaws are not obvious. Regardless, the regulation does not affect the responsibility of drivers to report any deficiencies or faults to the motor carrier on the condition of a vehicle.

The United States Government Accountability Office (GAO) analyzed the costs and safety benefits of operating state vehicle safety inspection schemes, obstacles faced by states in administering these schemes, and tasks that could be taken by NHTSA to assist states with these initiatives (US GAO, 2015). The study evaluated state data and data from NHTSA (2009–13) for crash trends

attributable to the failure of vehicle components; analyzed studies that reviewed relations between safety examinations and outcomes; and interviewed inspection representatives from 15 states. The GAO also interviewed representatives from five states that discontinued their safety inspection programs since 1990. The officials representing the five states and the District of Columbia cited the lack of concrete evidence to rationalize the efficiency of the system to conserve financial resources as one of the major reasons for eliminating the program. The study concluded that state officials adopted different criteria and chose not to include technological advances in their inspection systems, likely decreasing their inspection program's safety benefits. However, the study found that the states support using LED brake lights to enhance safety versus the conventional one. Similarly, the states have established more rigid system guidelines for inspection stations to be implemented to diminish fraudulent behavior such as fingerprint scanners for proper identification until inspections are carried out. The states have also begun to plan workforce reports, introduce more rigorous program guidelines, and develop digital information systems while addressing the challenges. The study suggested that to improve aid to states concerning the periodic motor vehicle inspection policy, there is a strong need to direct the NHTSA Administrator to develop and manage a communication and management framework with states to respond to questions from officials of the State Security Inspection Program and relay appropriate vehicle inspection data.

Vlahos et al. (2009) used Pennsylvania vehicle registration data. The researchers found that the failure rate of safety inspection for light-duty vehicles is between 12% and 18% (well above the often-cited rate of 2%). Older vehicles (more than 3 years old or having mileage greater than 30,000 miles) usually show higher rates. The evaluation of the new vehicles (less than or equal to one year old) shows that the failure rates of these vehicles are over zero. It is important to note the newer technologies, safer systems, and driver assistance make the vehicle fleets safer over the past few years. The current trend shows that inspection failure rate does not appear to be declining soon. This study also displayed that correct inspection data are limited, and vehicles are frequently inspected inaccurately. A similar analysis was conducted in North Carolina ([North Carolina General Assembly Program Evaluation Division, 2008](#)).

Das et al. (2018) explored 67,201 crash-related vehicle complaint reports from NHTSA database. This study found that major vehicular defects are associated with air bags, brake systems, seat belts, and speed controls. In a follow-up study, Das et al. (2019) conducted a statistical significance test to establish the efficacy of vehicle safety systems dependent on the existing states with and without safety inspection. The study used NHTSA automobile complaint data and FARS databases to evaluate the efficacy of regulatory vehicle safety components in different US states.

It is challenging to examine the advantages and disadvantages of state inspection programs, and state program officials often face challenges in program operations. However, these programs do ultimately enhance vehicle safety. It is obligatory of the administrator of NHTSA to establish and manage open communication with these state programs to relay important information and assist the states regarding the motor vehicle inspection guideline. With these enhancements, these programs could have more of an impact on improving vehicle safety in their respective states.

State Laws and Effectiveness Measure

Vehicle Inspection Program

In the United States, researchers are limited in determining the association between crash occurrences and vehicle-related failures by the lack of data. FARS data only provide fatality-related crash data at a national level. Due to the absence of national crash data with other severity levels, the crash rate comparison for inspection programs is hard to attain. Some states attribute the elimination of these programs. However, states generally do not accurately consider the costs of overseeing and managing these programs. The operating costs of inspection programs are entangled with additional programs, so determining the expense of vehicle inspection operation is not entirely possible. Also, there is typically a fee associated with safety inspections, which returns money to the state, but the amount varies between states. These funds are allocated to where each state deems fit before they are given to the inspection program. Currently, there are no states that have federally funded inspection programs ([US GAO, 2015](#)). Without properly identifying the operational costs of vehicle inspection programs, states cannot determine whether the costs of these programs outweigh the effects on crash rates. Because of the lack of evidence proving the effectiveness of the program or saving financial resources, some states have eliminated their programs.

States with inspection programs require information and guidance from the NHTSA to improve their program operations. As innovative vehicle technologies emerge, states need guidance in integrating these technologies as well as adding new requirements to their inspection programs. However, the NHTSA does not have a definite answer regarding the effectiveness of automobile inspection programs.

Table 1 displays which states have different types of automobile safety inspection regulations (i.e., annual, biennial, and random). As shown in **Table 1**, the following states have some sort of automobile safety inspection regulations: Delaware,

Table 1 Automobile inspection programs by state

Duration	States	Description
Annual	HI, LA, ME, MA, NH, NY, NC, PA, TX, VT, VA, WV	Require an annual vehicle safety inspection (LA: in areas where emissions inspections are needed)
Biennial	DE, LA, MO, RI	Require a biennial inspection (LA: in areas where emissions inspections are not needed)
Random	UT	Random roadside inspections

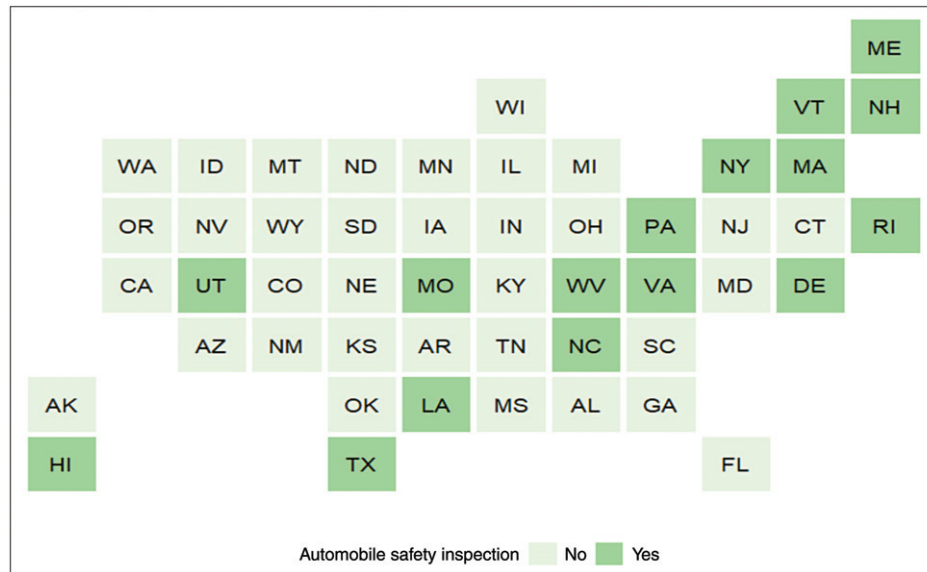


Figure 1 Vehicle inspection requirement by state.

Hawaii, Louisiana, Maine, Massachusetts, Missouri, New Hampshire, New York, North Carolina, Pennsylvania, Rhode Island, Texas, Utah, Vermont, Virginia, and West Virginia. A majority of the states have annual inspection programs and three states (Delaware, Missouri, and Rhode Island) require biennial inspections. Louisiana has both annual and biennial inspections based on the areas with emissions inspection requirement. According to the GAO report, Utah has random roadside automobile inspection programs. The NHTSA specified protocol outlines the minimum requirement for several important vehicles components such as tires, brakes, steering wheel, suspension system and wheel rotation or assemblies for safety inspections. This standard is applicable to all states with automobile inspection regulations (i.e., the states listed in [Table 1](#)). However, additional inspection requirements such as headlights, safety straps, horns, wiper blades vary based on the policies and regulations in each state ([US GAO, 2015](#)).

Fig. 1 shows the map of the United States with dark green squares indicating states that have some sort of inspection program. The light green squares indicate states with no automobile inspection program. As shown in **Fig. 1**, only 36% of the states has some sort of inspection programs.

Data Sources

FARS Data

The NHTSA has maintained the FARS database since 1975. During 1975–2017, there were 1,615,539 fatal crashes in the United States. The FARS database contains data from all 50 states, the District of Columbia, and Puerto Rico; it has information for 1.6 million fatal crashes in the United States. The major sources of this comprehensive database include police recorded crash reports, hospital entries and medical reports, emergency services, registration files, and state department of transportation records. The FARS database contains a comprehensive list of data elements in each fatal crash by characterizing crash event, involved vehicles and peoples, and event hierarchy. FARS data are a vital source for many researchers to understand the key contributing factors that lead to a fatal crash occurrence, such as roadway geometry, driver and vehicles, and ultimately working toward crash prevention.

NHTSA Vehicle Complaint Data

As an additional road safety improvement effort, the NHTSA maintains records of public provided complaints about vehicles, vehicle components, and transportation-related equipment. The database also contains vehicle and crash information to some extent. The data have been developed based on the complaints of vehicle owners or attorneys by several communication systems such as phone, fax, mail, or online submission. In these reports, the participants provide vehicle component failure information and describe the associated consequences for the failure. After receiving reports, Office of Defects Investigation analysts identify the failed component and complete the “specific component’s description” field. It is important to note that several records may exist for a single event or crash because there is a possibility of multiple failures in one event or crash ([Das et al., 2019](#)). Additionally, this database can provide insights on safety issues associated with vehicle models or components, identify safety trends for proper impact, track ongoing recalls, and order investigations for defects that may result in safety failure. The database developed till 2018 incorporates one and half million vehicle or vehicle component complaint reports in structured form with a wide list of variables. Around 8% of these reports involve fatalities or some level of injury. The complaints file contains all safety-related defect complaints received by NHTSA since January 1, 1995, as well as some incomplete rerecords for earlier years.

Das et al. (2019) applied a statistical significance test to this data to investigate the safety effectiveness of state-maintained vehicle inspections. The analysis used the “Cohen’s d” statistic to evaluate the differences in mean measures (complaints, complaint involved crashes, and fatal crashes from FARS) between the states with and without automobile inspection programs during the two time periods. This study conducted statistical significance test (using Cohen’s d measure) to determine the effectiveness of the vehicle safety inspection programs based on the states with and without automobile inspection program in place. The results from the vehicle complaint data showed that states with inspection programs expect a smaller number of monthly vehicle complaints and vehicle failure-related crashes than the states without safety inspection programs. This supports the claim that the mandatory vehicle inspection programs have a positive effect on safety. In contrary, the analysis of the FARS data showed no evidence on the positive effectiveness of safety inspection programs.

Future Scope and Research Direction

Further research concerning traditional methods such as regression modeling is required to examine the effectiveness of automobile safety inspection programs. Additionally, more advanced methods such as machine learning and deep learning can be applied to mitigate the research gap. Using innovative data source such as NHTSA vehicle complaint data is noteworthy. However, it is important to note that NHTSA may not receive all the vehicle complaints and so the results presented the recent study may not be comprehensive. There is a need for advanced analysis in reducing bias associated with the underreporting of these complaints. Furthermore, since the analysis using FARS data did not show any evidence of effectiveness of these programs, advanced tools such as machine learning models can be applied to reexamine the hypothesis of the effectiveness of the safety inspection programs using FARS data. Alternative data source such as General Estimates Systems (GES) can be used; however, precautions should be taken as the GES estimated values are not the actual counts like the FARS database. Additionally, severity specific analysis can be conducted. Moreover, several new data sources should be compiled to perform a comprehensive analysis on the effectiveness of automobile safety inspection laws. Future studies can explore the following datasets:

- **Exposure Data:** Vehicle miles traveled is the most commonly used exposure measure because it most directly captures exposure to crash occurrences. The Federal Highway Administration Highway Statistics publication provides annual estimates by roadway function class and vehicle type. The Bureau of Labor Statistics is a good source of employment-related data. The US Department of Commerce is a reliable source for gross domestic product estimates by state and year.
- **Department of Public Safety (DPS) Inspection and Citation Data:** State DPS authorities maintain roadside stop and citation data and as well the data related to the vehicle safety inspection. Inspection data are available by type of inspection certificate issued.
- **Crash Data:** State specific crash data can be beneficial in exploring the effectiveness of the automobile inspection programs. An alternative data source is GES.
- **Revenue Data:** Revenue data for the states can be collected from the state’s department of transportation and/or the comptroller of public accounts.

The evaluation of the safety inspection programs on fatal crashes is challenging because crash fatalities are relatively rare events in the context of the overall amount of travel; also, relatively few police reports attribute the cause of fatal crashes to vehicle component failure. Additionally, GES or state specific crash data can provide additional insight into the effect of inspection programs. Future studies can examine the disposition of the associated fees and the intended use of the funds for the states with automobile inspection programs. Research can also explore how similar services are funded in the other 34 states as well. Researchers can estimate revenue impacts by examining the reported proceeds in each of the identified funds, both in the aggregate and on a per capita basis.

Conclusion

Despite the belief that periodic inspection of registered vehicles can improve safety, the number of states mandating safety inspections dropped from a high of 31 in 1975 to 16 in 2015. It is widely believed that safety inspections can reduce vehicles with issues that may contribute to the improvement of roadway safety. However, research about the effects of these programs is inconclusive due to limitation of comprehensive databases. It is important to note that forms of automobile inspection such as annual and biennial only safeguard some key components of the automobiles. In many cases, these inspections are only able to identify a portion of potential automobile defects. Furthermore, the failure of inspectors to take a proper amount of care can result in ineffective safety inspection programs, which can create societal and economic loss. The costs of ineffective programs include costs related to inspection site visits, drivers’ time, inspection-related resources, and nonmandatory repairs that required for privately owned cars to pass the inspection. These costs vary based on each state’s supporting resources and density of inspection booths.

Safety inspection regulations in the countries outside of the United States widely vary. In Canada, the regulations vary from province to province. The European countries maintain safety regulation programs; however, the regulations differ in different countries. Each state or territory of Australia maintains its own safety inspection regulations. In Asia, automobile safety inspection programs are not effectively regulated through mandatory inspection programs.

Many state policymakers have questioned the effectiveness of automobile safety inspection programs. This uncertainty required a more robust scrutiny of the effectiveness of current inspection programs that will consider the inspection process by analyzing the

state of vehicles before they undergo the inspection rather than after it is conducted. The compelling findings of one of the recent studies demonstrated that safety inspection failure rates remain high, which calls into question why vehicle inspection regulations are not federally mandated like emission inspections.

The framework development and implementation of more robust data collection systems is the key to improve program efficiency by allowing for stronger oversight and improved management. The paper-based data system for inspection programs requires significant program oversight and enhancement. The system could be more efficient if it contained the functionality of electronic data collection and error checking. The fact of the matter is that few states have vehicle safety inspection programs at present, and even fewer states maintain electronic safety inspection records. Because of this, the insights on these programs are limited. More advanced data sources and analytical methods are necessary to mitigate this crucial research gap.

Biography



Dr Subasish Das is an Associate Transportation Researcher at the Texas A&M Transportation Institute (TTI). He received his Master of Science in Civil Engineering in 2012 and his PhD in Civil Engineering in 2015 both from the University of Louisiana at Lafayette. He has more than 10 years of national and international experience associated with transportation safety engineering research projects. His primary fields of research interest are roadway safety, roadway design and operation, mobility, machine learning, deep learning, and natural language processing. Dr Das is the author or coauthor of over 80 peer reviewed journal articles and research reports. He is also the author of the CRC Press book, *"Artificial Intelligence in Transportation Safety,"* which will be published in 2020.

Relevant Websites

Fatality Analysis Reporting System (FARS). <https://www-fars.nhtsa.dot.gov/Main/index.aspx>

National Highway Traffic Safety Administration (NHTSA). Flat files of Office of Defects Investigation (ODI). <https://www-odi.nhtsa.dot.gov/downloads/>

References

- Das, S., Mudgal, A., Dutta, A., Geedipally, S., 2018. Vehicle consumer complaint reports involving severe incidents: mining large contingency tables. *Trans. Res. Rec. J. Transp. Res. Board* 2672 (32), 72–82.
- Das, S., Geedipally, S.R., Dixon, K., Sun, X., Ma, C., 2019. Measuring the effectiveness of vehicle inspection regulations in different states of the U.S. *Transp. Res. Rec. J. Transp. Res. Board* 2673, 208–219.
- Official Journal of the European Union, 2014. Directive 2014/45/EU of the European Parliament and of the Council on 'periodic roadworthiness tests for motor vehicles and their trailers and repealing Directive 2009/40/EC'. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32014L0045&rid=5>.
- North Carolina General Assembly Program Evaluation Division, 2008. Doubtful return on the public's \$141 million investment in poorly managed vehicle inspection programs – Final Report to the Joint Legislative Program Evaluation Oversight Committee, Report No. 2008-12-06.
- The Ministry of Justice, Canada, 2019. Motor vehicle safety act. <https://laws-lois.justice.gc.ca/PDF/M-10.01.pdf>.
- US Government Accountability Office (GAO), 2015. Vehicle safety inspections: improved DOT communication could better inform state programs, Report No. GAO-15-705, Washington, DC.
- Vlahos, N., Lawton, S., Komanduri, A., Popuri, Y., Gaines, D., 2019. Pennsylvania's vehicle safety inspection program effectiveness study. The Pennsylvania Department of Transportation. Report No.: PA-2009-004-070609.
- World Resources Institute, 2018. Sustainable & safe: a vision and guidance for zero road deaths. <http://pubdocs.worldbank.org/en/912871516999678053/Report-Safe-Systems-final.pdf>.

Further Reading

- Christensen, P., Elvik, R., 2006. Effects on accidents of periodic motor vehicle inspection in Norway. *Accid. Anal. Prev.* 39, 47–52.
- Institute for Road Safety Research, 2012. Periodic vehicle inspection (MOT), SWOV, Leidschendam, The Netherlands. <https://www.swov.nl/file/13376/download?token=ondCj0iL>.
- Keall, M.D., Newstead, S., 2013. An evaluation of costs and benefits of a vehicle periodic inspection scheme with six-monthly inspections compared to annual inspections. *Accid. Anal. Prev.* 58, 81–87.
- Merrell, D., Poiras, M., Sutter, D., 1999. The effectiveness of vehicle safety inspections: an analysis using panel data. *South. Econ. J.* 65 (3), 571–583.
- Miller, R.E., 2014. FMCSA ends inspection reporting unless drivers find vehicle defects. *Trans. Top.* 26, 5.
- Peck, D., Matthews, H.S., Fischbeck, P., Hendrickson, C.T., 2015. Failure rates and data driven policies for vehicle safety inspections in Pennsylvania. *Trans. Res. A Policy Prac.* 78, 252–265.

Aviation Safety: Commercial Airlines

Clarence C. Rodrigues, Department of Industrial and Systems Engineering, Khalifa University, Abu Dhabi, UAE

© 2021 Elsevier Ltd. All rights reserved.

Introduction	90
Aircraft Safety	91
High-Lift Systems	91
Stopping Systems	91
Structural Safety	92
Fire	92
Human Factors Issues on the Flight Deck	92
Pilot Fatigue	92
Cabin Safety	93
Atmospheric Issues in Aviation	93
Turbulence	93
Wind Shear or Microburst	93
Volcanic Ash	93
Ice and Snow	93
Airport Safety	94
Airport Terminal Buildings	94
Hangars and Maintenance Shops	94
Ramp Operations	94
Aviation Fuel Handling	95
Deicing and Anti-Icing	95
Foreign Object Debris	95
Bird Strike	95
Runway Incursions	95
Managing Safety in Aviation	96
Conclusions	96
Relevant Websites	96
References	97

Introduction

Although they attract enormous attention, commercial aircraft crashes are rare and when one takes into consideration the number of air travelers, the chances of death from air crashes are extremely small and even less when compared to other forms of mass transportation. There is a range of estimates out there that predicts the odds of dying as a plane passenger. According to Alexandre de Juniac, IATA's Director General and CEO, an individual would need to take a flight every day for 241 years before experiencing an accident with one fatality on board. Dr. Per Garder, Professor of Civil and Environmental Engineering at the University of Maine, estimates (based on 2016 data) that an individual, on average, would need to fly once a day for 29,000 years before succumbing to a fatal crash. Finally the US National Safety Council's analysis of US Census data puts the odds of dying as a plane passenger at 1 in 205,552. That compares with odds of 1 in 4050 for dying as a cyclist, 1 in 1086 for drowning, and 1 in 102 for a car crash. Evidence suggests one is far more likely to die driving to the airport than being involved in a deadly plane accident. Air safety is improving with each passing year, with a future trend that reflects fewer deaths relative to the total number of traveling air passengers.

There are several reasons why commercial aviation remains the safest mode of mass transportation and this article will explore what commercial aviation safety is. There are four basic categories of aviation: *general aviation* (civilian/private flying that excludes scheduled passenger airlines), *corporate aviation* (air transportation for company employees and executives), *military aviation* (used for aerial warfare), and *commercial aviation* (used for paid transportation of people and cargo). This article will discuss safety in relation to *commercial aviation* only.

Commercial aviation is a type of ultra-safe high-risk industry (USHRI), such as the nuclear and chemical industry, in which there is less than one disastrous *accident* per 10 million *events*. An *incident* is something that happened during the operation of an aircraft which could affect the safety of operation but which did not rise to the severity of an accident. An incident could include a crewmember not being able to perform a normal flight duty because of injury or illness, an in-flight fire, or a flight control failure. An *accident*, however, is an occurrence that involves some degree of injury or damage related to the operation of an aircraft. While there are different variations of the definition of an accident, International Civil Aviation Organization's (ICAO) definition is

the most widely accepted in the commercial aviation sector. According to ICAO, a *major accident* occurs when an aircraft is destroyed, there are multiple fatalities, or there is a fatality coupled with a substantially damaged aircraft. A *serious accident* occurs when there is either one fatality without substantial damage to an aircraft or there is at least one serious injury and an aircraft was substantially damaged. An *injury accident* occurs when there is a nonfatal accident with at least one serious injury and without substantial damage to an aircraft, and a *damage accident* occurs when no one gets killed or seriously injured, but an aircraft receives substantial damage. The reference *Commercial Aviation Safety* (Cusick et al., 2017) reviews some of the significant accidents and actions taken (design, technology, and legislation) that have shaped the current commercial aviation industry to create a safer world for air transport.

While the aircraft is the core of the aviation transportation system, the aircraft itself is just one part of a wider air transportation system that includes airline and airport operations, maintenance, crews, security, weather services, and the air traffic control system. So aviation safety encompasses each of the aforementioned components of the air transportation system. What follows is a review of safety hazards that are encountered in the aviation transportation system and how the aviation sector manages risk associated with these hazards.

Aircraft Safety

Safety is the primary consideration of aircraft manufacturers when designing an airplane, with design standards that are often more stringent than regulatory requirements. Each critical system has at least one backup if not more. Twin-engine jets, for example, are designed to safely take off, fly, and land even if one engine fails. The airplane structure is designed to withstand 50% more load than the greatest load an aircraft might face while in service. Manufacturers build damage-tolerant airplanes that are rigorously tested to ensure they meet or exceed design standards and certification requirements, with extra margins of safety built in case of an extraordinary emergency. Aircraft structures are subjected to rigorous static and fatigue stress tests. Static tests validate the aircraft's ability to carry loads that are far greater than any load that would be faced under operational conditions. Fatigue tests subject an airplane to up to three lifetimes of normal wear and tear to help validate its durability (Boeing, 2019). Testing a new airplane design can take many months or even years and are conducted in laboratories, wind tunnels, icing tunnels, on the ground, and during flight tests. Boeing's aviation safety home page provides additional information on all of the above in greater detail.

High-Lift Systems

To accommodate the transition from piston to jet engines, high-lift systems such as movable aerodynamic surfaces on wings were developed to increase and decrease lift as a function of phase of flight. The multielement slotted trailing-edge flaps provided the required higher drag and power settings, while the addition of leading-edge flaps resulted in greater safety due to better takeoff and landing performance and greater stall margins. Leading-edge flaps on jets extend the margin of safety to even lower speeds and permit takeoff and climb out margins similar to those in prop airplanes. Today's jetliners often incorporate two-position leading-edge slats to provide low-deflection setting for low-drag takeoffs, a higher-deflection position for landing at lower speeds, and better pilot visibility from a lower deck angle. Some designs incorporate the auto slat to its configuration to enhance safety and improve stall characteristics while providing additional maneuvering margin in the case of a wind-shear encounter.

Stopping Systems

A number of safety improvements have taken place over the years in aircraft stopping systems. They include antiskid, fuse plugs, autobrakes, speed brakes, and thrust reversers.

While the need for *antiskid* braking systems was recognized due to higher takeoff and landing speeds in jet aircraft, most of the time antiskid features of brakes are not needed, as brakes are not applied hard enough to skid the tires. However, when the runway is wet and short, the antiskid system becomes very important to maximize braking effectiveness. Braking systems moved from the early antiskid systems for controlling tandem pairs of wheels in the four-wheel truck main gear configuration to controlling each wheel independently to modern digital antiskid systems that incorporate microprocessors to enhance the reliability and effectiveness. Low-melting-point *fuse plugs* in the wheels allow the tires to safely deflate during air pressure buildup when braking during high-speed rejected takeoffs instead of exploding and sending pieces of rubber flying into the airframe that occurred in earlier designs. The *automatic braking system* that enables automatic brake application on landing or during a rejected takeoff is a relatively recent feature to safety that frees the pilot to concentrate on other activities, such as applying reverse thrust and directional control of the airplane to achieve a smooth, safe stop on the runway. Another device used to enhance braking effectiveness is the *speed brake* (*wing spoiler*). These aerodynamic devices are located on the wings that serve to decrease lift and therefore increase download on the main wheels to enhance braking effectiveness. *Thrust reversers* are other devices that aid in safe stopping after touch down especially on short, slippery runways. Some designs are so effective that the airplane can stop within its required distance on a wet runway using only thrust reversers. The *stick shaker* is effective in avoiding aerodynamic stalls during both takeoff and landing. The stick shaker is a mechanical device that rapidly and noisily vibrates the control yoke of an aircraft to warn the pilot of an imminent stall and is mostly connected to the control column of jet aircraft.

Structural Safety

Through attention to detail design, manufacturing, maintenance, and inspection procedures commercial airplane designs have produced long-life, damage-tolerant structures with outstanding safety records.

The de Havilland DH 106 Comet was the world's first commercial jet airliner that had three catastrophic in-flight breakups within a year of entering airline service. One of the incidents' causative factors was dangerous concentrations of stress around some of the square windows. This led to the redesigning of windows to the current oval shape that exist on aircraft. Aircraft are designed to be damage-tolerant or fail-safe. This concept requires that a specified level of residual strength be maintained after complete failure or partial failure of a single principal structural element. A major issue facing the airlines currently is aging aircraft. By definition, an aging aircraft is one that is being operated near or beyond its originally projected design life of calendar years, flight cycles, or flight hours. The two important considerations of age are the number of flight cycles and the number of flight hours accumulated in service. Fatigue and corrosion are the primary concerns about aging aircraft. Fatigue damage to the fuselage is caused primarily by the repeated application of the pressure cycle that occurs during every flight. Fatigue damage to the wings is caused by the ground-air-ground cycle that occurs during every flight, pilot-induced maneuvers, and turbulence in the air. Thus, the goal of measuring aircraft structural safety by flight hours is more important for the wings than it is for the fuselage. With proper inspection, maintenance, and repair, the life of the aircraft structure could be unlimited.

Fire

Fire and its toxic smoke from burning aircraft materials have caused deaths. The Air Canada Flight 797 accident in 1983 had a significant impact on global aviation safety regulations, which were updated as a result of the accident. What evolved were new requirements to install smoke detectors in lavatories, strip lights marking paths to exit doors, and increased firefighting training and equipment for crew became standard across the industry. Evacuation requirements were also updated to mandate that aircraft manufacturers demonstrate that their aircraft could be evacuated within 90 seconds of the commencement of an evacuation with half the emergency exits blocked. According to studies, 90 seconds is the time needed to evacuate before there can be a large fire or explosions, or before fumes fill the cabin. Instructing passengers seated in over wing exits to assist in an emergency situation were also included as a requirement. Advanced materials for seat fabric and insulation have given between 40 and 60 additional seconds to people on board to evacuate before the cabin gets filled with fire and potential deadly fumes. In-flight fire in the cargo hold is also an issue hence cargo holds of most airliners are now equipped with automated halon fire extinguishing systems to fight fires that may occur in baggage holds.

Human Factors Issues on the Flight Deck

About 60%–80% of aviation accidents are attributed to human error. Human factors study and apply data on human capabilities and limitations to designing systems (equipment, tasks, environments, and the human–equipment interface) that are *safe and maximize human performance*. Understanding human factors is important to systems where humans interact with sophisticated equipment where human error can cause catastrophic accidents. The application of human factors is of particular importance on the flight deck where the human–machine interface often becomes a determining factor in the event of an emergency, where correct and timely decisions make the difference between life and death. Sample human factor applications on the flight deck include the comfort of the seats on the flight deck, the mental and physical preparedness of the pilots and copilots to fly, the visibility of the ramp area when looking through the windshield, the ability to blind-reach specific controls on the panels by touch alone, properly monitoring the flight instruments, the ease of interpreting and understanding audio and written information, the color-coding of annunciator lights on the display panel, ability to effectively communicate between pilots in the cockpit, ability of the flight crew to maintain situation awareness, ability to recognize and differentiate between runway and taxiway lighting against a backdrop of city lighting, and many others.

Technology improvements on flight decks have made significant contributions to improving safety in the areas of radio communication and inertial navigation, approach systems, and greatly simplified cockpit controls and displays. Some of these flight deck technology changes include crew alerting and monitoring systems, moving map displays, engine-indicating and crew-alerting system (EICAS), glass cockpit displays with color enhancement, ground-proximity warning system (GPWS), traffic collision avoidance system (TCAS), aircraft communications addressing and reporting system (ACARS), flight management system (FMS), heads-up display (HUD), and several other redundant and automated systems.

Pilot Fatigue

Fatigue is a physiological state of reduced mental or physical performance capability resulting from sleep loss or extended wakefulness, circadian phase, or workload. Fatigue is particularly prevalent among pilots because of unpredictable work hours, long duty periods, circadian disruption, and insufficient sleep. Fatigue significantly increases the chance of pilot error that can lead to accidents. While regulators attempt to mitigate fatigue by limiting the number of hours pilots are allowed to fly over varying periods of time, these restrictions may not always be sufficient.

Cabin Safety

Some of the initiatives that increase the likelihood of passenger survivability and effective evacuation in aviation accidents include:

- The NTSB recommended improvements in life preservers, passenger briefings, cabin safety data cards, emergency-evacuation slides, and floatation devices for infants, and crew postcrash survival training.
- Required protective breathing equipment for flight crews and cabin attendants, without which they could easily be overcome by fire and smoke and could not use fire extinguishers effectively.
- Fire-blocking seat cushions to prevent fire or to mitigate its effects. FAA Advisory Circular 120-80 (FAA, 2014) provides detailed guidance on how to deal with in-flight fires, emphasizing the importance of crewmembers taking immediate and aggressive action.
- Emergency floor lighting to improve the chances and speed of evacuation under conditions where there is significant smoke in the cabin.
- Stronger seats that must be able to withstand 16 times the force of gravity are required for Part 121 aircraft that were built after October 27, 2009. Seats must undergo dynamic testing with test dummies to ensure they are effective in a realistic aviation crash environment.
- Other improvements through the years include impact resistant seat frames, and airplane wings and engines designed to shear off to absorb impact forces.

Atmospheric Issues in Aviation

High-altitude flight can be significantly impacted by diverse atmospheric conditions, some of the more important ones being turbulence, wind shear, volcanic ash, ice, and precipitation.

Turbulence

Most injuries incurred due to turbulence are preventable if the passengers wear their seat belts. Studies have shown that during turbulence it was important not to change aircraft trim, but to disengage the early autopilot's auto trim feature and airspeed hold only if necessary, and to fly attitude and let the airspeed and altitude vary somewhat. Areas of concentration to address turbulence include turbulence forecasting, flight crew training, and crew procedures in areas where turbulence is anticipated. Detection and avoidance are the two main approaches addressing the turbulence issue. In this regard the FAA Aviation Weather Research program has a multidisciplinary team addressing the turbulence issue, and the National Center for Atmospheric Research (NCAR) is working on new algorithms for using data from the National Weather Service (NWS) Doppler weather radars to detect turbulence.

Wind Shear or Microburst

A wind shear is a change in wind speed and/or direction over a relatively short distance in the atmosphere. A microburst is a localized column of sinking air that drops down in a thunderstorm. Wind shear during takeoff and landing and winds-aloft hazard in mountain waves especially in high-risk areas such as the Rocky Mountains of the United States are two additional flying hazards to be contended with. Strategies to handle wind shear include training crews on avoidance of the phenomena, better detection of the conditions that can produce wind shear and alerting the crew, and getting maximum performance from the airplane (incorporating latest technologies in advanced flight decks) if the crew inadvertently encountered wind shear.

Volcanic Ash

Volcanic ash from active volcanoes is hard and abrasive, and can quickly cause significant wear to propellers and turbo compressor blades, scratch cockpit windows, contaminate the cabin, damage avionics and impair visibility. The ash can contaminate fuel and water systems, jam gears, and make engines fail. Its particles have low-melting point, so they melt in the engines' combustion chamber then the ceramic mass sticks to turbine blades, fuel nozzles, and combustors, which can lead to engines flaming out. The ICAO (2012) has set a maximum limit of 4 mg of silicate ash per cubic meter in the clouds associated with eruptions of volcanoes as the safe limit of ash concentration for ingestion by jet engines.

Ice and Snow

Ice and snow can negatively affect the airframe and engine in flight or the tire-ground contact on the runway. A small amount of icing can greatly impair the ability of a wing to develop adequate lift hence regulations prohibit ice, snow, or even frost on the wings or tail, prior to takeoff. An accumulation of ice during flight can be catastrophic as well. Prevention of in-flight ice buildup on wings and tails is done either by routing heated air from jet engines through the leading edges of the wing or by use of inflatable rubber deicing boots that expand to break off any accumulated ice. Skidding on the runway is also an issue that is addressed by prompt snow

removal and deicing from runways. Under the revised certification standards, new transport aircraft designs must incorporate methods to detect icing and to activate the airframe ice protection systems.

Airport Safety

An airport provides facilities for landing, takeoff, shelter, passengers, fuel storage and fueling, supplies, and repair of aircraft. Airport operations are complex and diverse, with hazards and their severity varying by the type of operation. Airport design and location can have a large impact on aviation safety, especially for some airports that were originally built for propeller planes and those that are in congested areas where it is difficult to meet newer safety standards. What follows are airport safety issues covering airport terminal buildings, hangars and maintenance shops, ramp operations, and specialized airport services that include aviation fuel handling, and deicing and anti-icing operations.

Airport Terminal Buildings

Inadequate layout and facility size leads to overcrowding, which can cause dangerous conditions during evacuation in an emergency (ex. fire or terrorism). Due consideration for safety should be given during the design, construction, and modification of the facilities which are covered by the design standards and requirements for building materials, fire protection, and building egress. Some of these standards/regulatory bodies include the National Fire Protection Association (NFPA), OSHA, EPA, and other local, state, and federal agencies. A summary of important safety considerations for airport terminal facilities follows.

Hangars and Maintenance Shops

Information on planning and design guidance for airport terminals is included in FAA Advisory Circular ([FAA, 2016](#)) for non-hub locations. Construction of hangars is covered by federal, state, and local building codes and many NFPA standards. Operational issues are mostly covered by OSHA and EPA requirements. Some of the safety issues include:

- Electrical—ex. aircraft docks should be permanently bonded, grounded, and secured to a permanent structure such as a hangar floor. Primary sources of electric power to hangars should be approved for the location according to the National Electrical Code, NFPA 70, and aircraft hangars, NFPA 409.
- Storing and handling of compressed gas, flammables, and hazardous and toxic substances (addressed under OSHA 29 CFR 1910 Subparts H and Z).
- Painting and stripping are hazardous activities that use toxic chemicals, and thus must be performed in well-ventilated areas (e.g., paint booths). In addition, painting and stripping generate toxic wastes that must be contained and either recycled or disposed of appropriately. Therefore, these operations are governed by both OSHA and EPA regulations.
- Material handling equipment such as overhead and gantry cranes that are used to move large aircraft sections are regulated under OSHA 29 CFR 1910.179. Slings, chains, and hoists used to move or lift smaller aircraft parts and sections (e.g., aircraft jacking) are regulated under OSHA 29 CFR 1910.184.
- The use of powered material handling equipment, such as fork trucks and hand trucks, and the use and maintenance of batteries used to power these and other equipment are regulated under OSHA 29 CFR 1910.178.
- Hoist units, such as those used to reach large aircraft tails, should comply with OSHA 29 CFR Part 1910 Subpart F, Powered Platforms, Man lifts, and Vehicle-Mounted Work Platforms.
- Welding, cutting, brazing, and other hot works that have the potential to start fires are regulated under OSHA 29 CFR Part 1910 Subpart Q.
- Noise evaluation and hearing protection for processes such as riveting, sand blasting, grinding, polishing, and other noisy operations are regulated under OSHA 29 CFR 1910.95.

Ramp Operations

The ramp area is generally designed for aircraft, not the vehicles that service and/or operate in the proximity of the aircraft. Most of the signs and markings are for aircraft. The ramp area sees a diverse collection of high-paced activities that involve aircraft, vehicles, and individuals working in close proximity to one another. Some of these activities include: aircraft ground handling that may include taxiing, towing, chocking, parking, or tie-down; aircraft refueling; aircraft servicing to include catering, cleaning, food service, etc.; and baggage and cargo handling. The earlier activities lead to several occupational hazards that include: cuts (from antennas, pitot tubes, static discharge wicks, etc.), slips and falls on the ground and from elevations, strains and sprains from baggage handling, exposure to hazardous materials (fuel), contact with moving parts (propellers) and bumps (undersurface of fuselage), electrical hazards (tools, motor, generators, etc.), biohazards from blood and other potentially infectious fluid exposures during cleaning, high-pressure air and other fluid exposures from pressurized systems, noise from engines and other equipment, injury from jet blast, and weather conditions (heat, cold, snow, ice, and rain) that can increase the risk of injury.

Aviation Fuel Handling

Fuel handling is an important safety issue to the fuelers and the operation of the aircraft. Failure to adhere to safe operating procedures when fueling aircraft and/or transporting fuel from one location to another on the airport can result in major disasters. Repeated contact with aviation fuels can cause skin irritation and dermatitis. Fuels contain additives such as benzene and can be toxic if inhaled or swallowed. Fuel contamination is a major safety issue as it can affect the operation of an aircraft. Contamination can occur by refueling an aircraft with contaminated fuel or the wrong fuel grade or type. Contaminants include rust, water, microbial growths, paint, metal, rubber, and lint. Explosions and fires during fueling or fuel transfer from sparks or static discharge are another safety hazards. Hazards from fuel spills and leaks from underground storage tanks present varying degrees of safety and environmental issues depending on the size of the spill and the kind of response required.

Deicing and Anti-Icing

Ice and snow on control, airfoil, and sensor surfaces can have serious consequences on the safe operation of the aircraft, so when freezing or near-freezing conditions exist, the aircraft should be sprayed with deicing fluid before takeoff. Hazards posed by this operation include damage to the aircraft by the deicing equipment, application of deicing fluid in areas where it should be avoided (static ports, pitot heads, angle-of-attack sensors, the engine, and other inlets), inhalation and ingestion hazards to the deicing crew, and hazards to the crew and passengers if the cabin air intakes are not shut off during deicing.

Foreign Object Debris

Foreign object debris (FOD) includes items left in the aircraft structure during manufacture, maintenance, or debris on the runway. Such items can damage engines and other parts of the aircraft. In July 2000, Air France Flight 4590 (a Concorde) ran over debris on the runway during takeoff, blowing a tire and puncturing a fuel tank. The subsequent fire and engine failure caused the aircraft to crash into a hotel after takeoff, killing all 109 people aboard and 4 more people in the hotel.

Bird Strike

Bird strikes have caused fatal accidents through engine failures following bird ingestion and breaking cockpit windshields. While jet engines are designed to withstand the ingestion of birds of a specified weight and number, ingesting birds beyond the “designed-for” limit can be catastrophic as witnessed by US Airways Flight 1549 which encountered a flock of Canada geese. In this case, pilot skill and luck led to all passengers surviving. The outcome of a bird strike depends on the number and weight of birds and where they strike the fan blade span or the nose cone. The highest risk of a bird strike occurs during takeoff and landing in the vicinity of airports, and during low-level flying. Airports use a variety of countermeasures including shotgun patrols, playing recorded sounds of predators through loudspeakers, or employing falcons, planting poisonous grass to detract birds and to insects that attract the birds, and avoiding conditions that attract birds to the area (e.g., landfills).

Runway Incursions

A runway incursion is a type of runway safety incident that involves an incorrect presence of a vehicle, person, animal, or another aircraft on the runway. The FAA classification of a runway incursion in order of increasing severity is:

- Category D—Incident that meets the definition of runway incursion such as incorrect presence of a single vehicle/person/aircraft on the protected area of a surface designated for the landing and takeoff of aircraft but with no immediate safety consequences.
- Category C—An incident characterized by ample time and/or distance to avoid a collision.
- Category B—An incident in which separation decreases and there is a significant potential for collision, which may result in a time critical corrective/evasive response to avoid a collision.
- Category A—A serious incident in which a collision was narrowly avoided.
- Accident—An incursion that resulted in a collision.

The worst aviation disaster on record was the result of two Boeing 747s colliding on a runway at Tenerife Airport in the Spanish Canary Islands in 1977. This is the most severe case of a runway incursion, one that resulted in an accident taking 583 lives. Incursions occur due to (Rodrigues, 2002):

- Operational errors: This is an action of an air traffic controller that results in a less than the recommended separation between two or more aircraft, or between an aircraft and obstacles (vehicles, equipment, personnel, etc.) on runways, or an aircraft landing or departing on a closed runway.
- Pilot deviation: This is a violation any federal aviation regulation, for example, failing to observe an ATC instruction not to cross an active runway when on an authorized route to a gate.
- Vehicle/pedestrian deviation: This is when vehicles, pedestrians, or other objects interfere with aircraft operations by entering a restricted area without authorization from air traffic control.

The airport ground operations environment is a complex system of markings, lighting, and signage with varying layouts by airport. In bad weather, visibility decreases that further increases the complexity of the environment. Night operations complicate issues additionally as airport lighting blend with background city lights. It is under these conditions that large numbers of individuals with varying levels of experience, training, and language proficiency must coordinate actions and procedures that are needed for smooth and safe operations while interfacing with large numbers of individuals, airplanes, and ground vehicles that are in close proximity to one another. It is therefore easy to understand why under these conditions, individuals are prone to error and incidents occur. Given the growth in air travel, things are going to get worse with the number of these events increasing over time. Additional information on runway incursion may be found in the reference section by [Rodrigues \(2004\)](#).

Managing Safety in Aviation

This section deals with how the aviation sector manages risk associated with the hazards outlined in the previous sections. Aviation safety management is a system in which risks associated with aviation undertakings, related directly or indirectly to support the operation of aircraft, are reduced and controlled to acceptable levels. Safety risks arise from undesired, unplanned, uncontrolled events (hazards outlined in the previous sections) that can cause adverse effects to personnel, equipment, environment, property, and reputation within a defined system. Risk is the combination of the probability of occurrence of the hazard and the severity of its effects, and is expressed as:

$$\text{RISK} = \text{Probability} \times \text{severity}$$

- Probability = Loss event/unit of time or activity
- Severity = Loss/loss event
- Risk = Expected loss/unit time or activity

Risk management is the overall process of identifying, evaluating, controlling, and accepting risks, while following all health, safety, and environmental regulations and best practices. Safety can be expressed as a level of risk associated with the operation of a system.

Safety management systems (SMS) is a term describing a standardized approach to controlling risk across an entire organization that promotes the sharing of safety data and best practices. In the United States, the SMS framework consists of 4 components or pillars and 12 elements within those components ([FAA, 2010](#)). The four components of SMS are:

- Safety policy—Establishes senior management's commitment to continually improve safety; defines the methods, processes, and organizational structure needed to meet safety goals.
- Safety risk management (SRM)—Determines the need for, and adequacy of, new or revised risk controls based on the assessment of acceptable risk.
- Safety assurance (SA)—Evaluates the continued effectiveness of implemented risk control strategies, and supports the identification of new hazards.
- Safety promotion—Includes training, communication, and other actions to create a positive safety culture within all levels of the workforce.

Conclusions

Present-day commercial aviation is a USHRI that manages to operate with a great degree of safety in a high-risk environment. In aviation, it is practically impossible to attain an environment devoid of threats to safety given that individuals interface with objects that move very fast under human control and have high potential energy. Safety is a top priority for airlines as it can affect profitability. Since efficiency is often a natural by-product of safety, a commercial aviation operator that manages safety risk adequately will often also gain in operational efficiencies. This article reviewed safety hazards that are encountered in the aviation transportation system and how the aviation sector manages risk associated with these hazards through an SMS.

Relevant Websites

Embraer safety. Available from: <http://www.embraer.com/en-us/conhecaembraer/qualidadetecnologia/pages/home.aspx>.

FAA—tips on mountain flying. Available from: https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/media/tips_on_mountain_flying.pdf.

FAA National Runway Safety Plan (2015–2017). Available from: https://www.faa.gov/airports/runway_safety/publications/media/2015_ATO_Safety_National_Runway_Safety_Plan.pdf.

Airbus safety. Available from: <http://www.airbus.com/company/aircraft-manufacture/quality-and-safety-first/>.

ASRS examples. Available from: <http://www.asrs.arc.nasa.gov/>.

Information about the growth of aviation. Available from: <http://www.iata.org>.

NASA Ames Aeronautics Research Directorate. Available from: <http://www.aeronautics.nasa.gov/>.

References

- Boeing, 2019. Boeing safety page. Available from: <http://www.boeing.com/company/about-bca/aviation-safety.page>.
- Cusick, S.K., Cortes, A.I., Rodrigues, C.C., 2017. Commercial Aviation Safety, sixth ed. McGraw Hill Publishing Co., New York.
- FAA, 2010. Safety management system implementation guide. Available from: https://www.faa.gov/about/initiatives/sms/specifics_by_aviation_industry_type/air_operators/media/sms_implementation_guide.pdf.
- FAA, 2014. Inflight fires. Available from: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_120-80A.pdf.
- FAA, 2016. Airport terminal planning and design. FAA AC 150/5360-13A. Available from: <https://crp.trb.org/acrp0715/faa-advisory-circular-1505360-13a-airport-terminal-planning-and-design/>.
- ICAO, 2012. Flight safety and volcanic ash. Available from: https://www.icao.int/publications/Documents/9974_en.pdf.
- Rodrigues, C.C., 2002. Control strategies for runway incursions, Proceedings of the International Society Conference for Occupational Ergonomics and Safety, 9–12 June, Toronto, Canada.
- Rodrigues, C.C., 2004. Runway incursions—a potential for catastrophe, Proceedings of the Intelligent Transportation Systems Safety and Security Conference, 24–25 March, Miami, FL.

Aviation Safety, Freight, and Dangerous Goods Transport by Air

Glenn S. Baxter*, Graham Wild†, *School of Tourism and Hospitality Management, Suan Dusit University, Hua Hin, Thailand; †School of Engineering and IT, UNSW, Canberra, ACT, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	98
Historical Trends in World Air Freight Growth	99
Shippers	100
Trucking Firms	100
International Air Freight Forwarders	100
General Sales Agents	101
Airports	101
Cargo Terminal Operators	101
National Customs Authority	101
Ramp Handling Agent	101
Airlines	101
Consignees	102
The Provision of Air Cargo Capacity	102
Combination Aircraft Belly-Hold Air Cargo Capacity	102
Freighter Aircraft Air Cargo Hold Capacity	102
Safe Transportation of Dangerous Goods by Air	102
Substantive Safety	104
Overview of Historical Air Cargo Safety Occurrences	104
Rates of Air Cargo Accidents	104
Case Examples	105
Conclusion	106
Acknowledgment	106
References	106

NOMENCLATURE

A18 Annex 18 of the 1944 Chicago Convention on International Civil Aviation
ASN Aviation Safety Network
DGR Dangerous goods regulation
EASA European Union Aviation Safety Agency
FAA United States Federal Aviation Administration
HMR Hazardous Materials Regulations
IATA International Air Transport Associations
ICAO International Civil Aviation Organization
ICAO9284 Technical Instructions for the Safe Transport of Dangerous Goods by Air
ULD Aircraft unit load device

Introduction

Aviation safety is the major priority for the transportation of passengers and air freight in the global airline industry. Air freight is anything that is transported in an aircraft except for mail or passenger luggage carried under a passenger ticket but including baggage shipped under the cover of an airwaybill (Hui et al., 2004). Air freight is carried by combination airlines in the lower decks of their passenger aircraft, dedicated all-cargo airlines, and by the integrators, for example, DHL Express, FedEx and United Parcel Service (UPS) (Baxter and Bardell, 2017; Dresner and Zou, 2017; Morrell and Kleing, 2019).

When transporting goods by any mode (air, sea, rail, truck), an item is considered hazardous if it is explosive, corrosive, flammable, toxic, or radioactive. Dangerous goods are articles or substances, which are capable of posing a risk to health, safety, property, or the environment and which are shown in the list of dangerous goods in the International Civil Aviation Organization (ICAO) Technical Instructions or which are classified according to those instructions (International Civil Aviation Organization, 2020b).

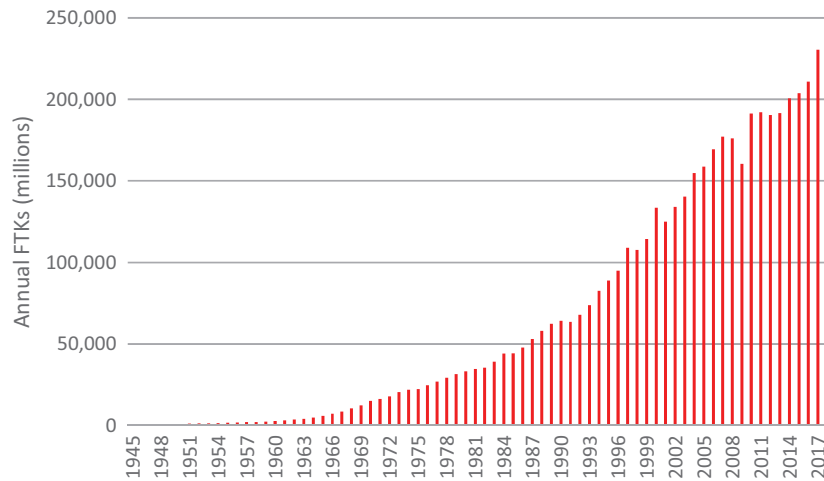


Figure 1 The annual growth in global air freight ton kilometers performed: 1945–18.

The transport of dangerous goods by air is subject to strict regulation at both the government and industry level. The aim of this article is to provide an overview of the carriage of dangerous goods by the air freight mode and the associated regulatory framework that is applicable to all actors participating in the global air freight industry.

Historical Trends in World Air Freight Growth

The world scheduled air transport industry commenced largely during the post-World War 1 period. Following the Berlin Air Lift, after the World War II, the world air freight industry began to form itself around the freight-carrying capabilities of commercial aircraft. Since those early times, the international air cargo mode has developed into a high growth industry ([Fig. 1](#)) ([Sales, 2013](#)).

Air freight was considered a sideline operation to postal and passenger traffic until 1941, when the United States “Big Four” domestic airlines [American, Eastern, Trans World Airlines (TWA), and United Airlines] formed Air Cargo Inc. to commence regular cargo flights. By the end of 1941, many of the US-based airlines, including TWA, had commenced their own independent air cargo services ([Chiavi, 2017](#)). Also, around this time Pan American World Airways an extensive airfreight operation.

Prior to the 1960s, airlines considered air freight as a way of filling spare aircraft hold capacity on what were principally narrow bodied passenger aircraft. The introduction of jet aircraft, particularly the Boeing B707 and the McDonnell Douglas DC-8 in the early 1960s, stimulated nearly 10 years of rapid air cargo growth, the annual growth rate averaged around 16% during this period ([Hayuth, 1983](#)).

During the 1970s, the annual world air freight growth averaged around 8.6%. The introduction of wide-body aircraft in the early 1970s, such as the McDonnell Douglas DC10, Lockheed L1011 Tristar, Airbus A300, and particularly the first all-freight aircraft, the Boeing B747 freighter aircraft in 1972, had a pronounced effect on the air cargo industry ([Wood et al., 2001](#)). These new wide-body aircraft enabled airlines to offer significantly greater air cargo capacity as they introduced the new aircraft type on more and more routes, while also increasing the size of air cargo consignments that could be carried. Around this time, scheduled airlines also took their air cargo services more seriously and introduced more flexible and competitive pricing in order to secure greater volumes of air cargo traffic ([United Kingdom Department of Transport, 2000](#)).

In the 1980s the world air freight growth rate averaged around 7.2% per annum. During the first half of the 1980s growth in world freight ton kilometers performed (FTKs) was stronger on scheduled passenger services. In the latter part of the 1980s dedicated freight services grew faster than on passenger services. There was a continued increase in the annual growth rate during the 1990s of 7.7% ([Doganis, 2010](#)). Since 1990 there have been variable growth levels on both scheduled passenger and dedicated freight services.

The growing importance of dedicated freighter services is due to customer demand for superior services and stricter safety regulations on lower hold cargo aboard passenger services ([Boeing Commercial Airplanes, 2006](#)). The imbalance in air cargo traffic flows is a further factor ([Kupfer et al., 2011](#)), according to Airbus. Most of the growth of world air cargo (as measured by FTKs) from 1980 to 2017 occurred in scheduled international traffic. Representing 70% in 1981, scheduled international services increased their share to around 87% in 2017 ([International Civil Aviation Organization, 2018](#)).

Probably the most significant development to occur in the world air cargo industry over the past 30 years has been the rapid growth and development of the air express sector. Express package services were pioneered by Federal Express (FedEx) in the United States in 1973 ([Doganis, 2010](#); [Smith, 2001](#)). FedEx identified a number of key product features hitherto unrecognized by the traditional air cargo industry: the requirement for door-to-door and overnight package transportation for urgent/time-important

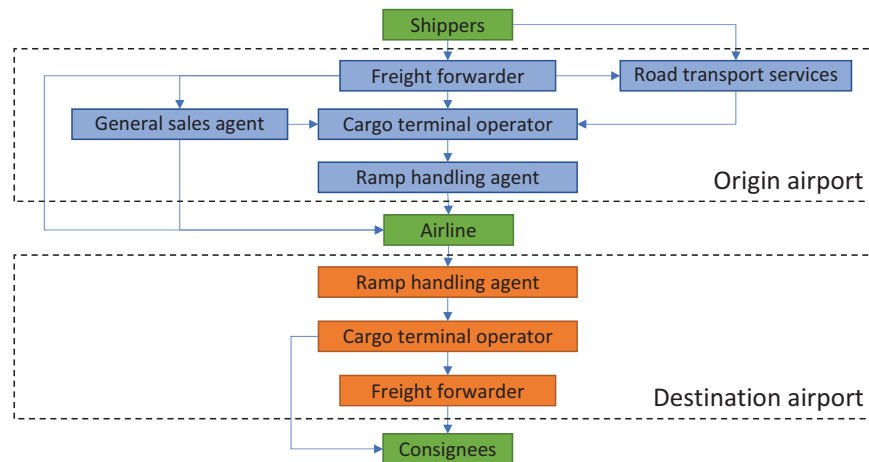


Figure 2 The air cargo supply chain, illustrating the key players from shipper to consignee.

small parcels and documents. Instead of selling services purely on the basis of weight and price, customer convenience, rapid shipment times and reliability became the principal product features (United Kingdom Department of Transport, 2000). The company developed rapidly and earned \$USD 1 billion in annual revenue in 1983. In 2000 Federal Express was renamed FedEx (Smith, 2001). The integrators services are underpinned by the use of dedicated, efficient global multimodal networks—they own and operate most of their own aircraft, smaller planes, trucks, and automated handling and storage facilities (Milenkovic, 2001). The integrators introduced guaranteed delivery times with a pricing structure to match. Hence, the integrators supplement their air services with extensive ground transport to provide time-definite delivery to their customers with continuous shipment tracking and, if necessary, logistics expertise to support just-in-time (JIT) inventory control strategies. These firms also provide value added services, such as, customs brokerage and logistics services, warehousing, and inventory control (Cavusgil et al., 2012). **The Key Players in the Global Air Cargo Supply Chain**

Air cargo service providers are a heterogeneous group that provide a variety of logistics and supply chain services and expertise (Doganis, 2010; Reynolds-Feighan, 2001). The air cargo industry consists of various commercial organizations that provide shippers and/or consignees with air cargo services. Fig. 2 shows the relationship between the major actors in the world air cargo industry. The distinction amongst these actors is often clouded. Airlines offer small-package services and freight forwarders lease or operate aircraft dedicated to their requirements. There are many contracting and subcontracting operations in which a firm in one market segment of the industry works closely with a firm in another market segment (Wood et al., 2001). All sizes of aircraft are used (Morrell and Kleing, 2019; Wraight, 2017). Airlines often offer more than one level of service, meaning that the shipper has some choices as to how quickly the cargo consignment will move and for which additional services they are willing to pay (Wood et al., 2001).

Shippers

The shipper, also known as the consignor, because they consign or entrust their consignment to the actors undertaking the physical transfer process, may be a company or an individual, exporting either general or perishable goods (The International Air Cargo Association, 1998). The shipper, the owner of the goods, establishes the initial link in the air freight chain (Damsgaard, 1999). To ship their goods to a customer, a shipper will generally contact several freight forwarders to find cargo space and obtain the best price and service offered (Beifert, 2016; Morrell and Kleing, 2019; Sales, 2016).

Trucking Firms

Trucking companies provide road transport services for producer/shippers transporting their products from their premises to their assigned freight forwarder or to the airport where the consignments are handled and stored in preparation for their designated flight.

International Air Freight Forwarders

The process of moving air cargo is more complex than for passenger traffic. For example, it involves packaging, extensive and complex documentation, insurance, pick-up and/or collection, and customs clearance. This complexity has encouraged the growth of specialist firms, the freight forwarders (Dempsey and Gesell, 1997). International air freight forwarders can be defined as ‘international trade specialists who provides a variety of functions and services to facilitate the movement of cross-border shipments’ (Murphy and Daley, 1996, p. 5). International air freight forwarders provide the full link between shippers and consignees, from surface point to surface point and also act as the interface between the shipper and the airline (Dempsey and Gesell, 1997; Ohashi et al., 2005).

International air freight forwarders contract with airlines for the physical carriage of goods, purchase block cargo space on their flights, consolidate cargo consignments from multiple shippers, and operate—or contract with—complementary surface transportation service providers (Al-Hajri, 1999). Due to economies of scale, international air freight forwarders are able to offer “consolidation” services that are often cheaper than those which the exporter could negotiate directly with the transportation provider (Corley, 2002). The freight forwarder makes a profit by consolidating shipments and obtaining a quantity or volume discount from the airline for the larger shipment size (Dempsey and Gesell, 1997; Wood et al., 2001).

Other services provided by air freight forwarders on behalf of shippers include customs brokerage, transport insurance, warehousing and distribution, loading airline air cargo containers (ULDs), documentation, (Bruins, 2006), banking requirements, trucking (pick-up or delivery), cargo space reservations for shippers, documentation preparation, and in some instances, shipment packing and cargo insurance. Air freight forwarders may also manage, under contract, some of the downstream production functions of a supply chain, offering services in warehousing, quality control, inventory control, and JIT delivery operations of firms, often large multinationals (Wan et al., 1998).

General Sales Agents

The general sales agent (GSA) is an appointed representative of the airline and can be viewed as the airline’s outsourcing partner (Hosie et al., 2012; Sales, 2016). GSAs serve as an intermediary between their client airlines and freight forwarders. GSAs handle their client airlines sales, marketing, and operational activities in the specified territory. Other functions that may be performed by GSAs include traffic handling, maintenance, catering, flight dispatch, customer service, and warehousing. GSAs typically directly market their client airlines services to the freight forwarders (Hosie et al., 2012).

Airports

Normally airports act as landlords and infrastructure providers charging landing fees and aircraft apron parking fees to airlines (their primary customers) and charging rent to other air cargo related service providers, for example, cargo terminal operators (CTOs) (Ryan, 2008; United Kingdom Department of Transport, 2000).

Cargo Terminal Operators

Cargo handling at an airport is typically performed by an airline’s own staff or is outsourced to dedicated third-party service providers or other airlines. CTOs provide the handling facilities necessary to accept air cargo consignments from international air freight forwarders and shippers, check shipment weights, and prepare aircraft load plans (Morrell and Kleing, 2019). They also store the consignment until it has been cleared for export by the customs authority or held for at least 24 hours at the terminal for safety reasons (Damsgaard, 1999). The CTO then arranges for the consignment to be loaded on the designated aircraft (Caves, 2015). At the destination, an air CTO accepts cargo from incoming flights and stores the cargo until it has been cleared and released by the receiving country’s customs authority. A freight forwarder then typically thereafter collects the consignment and arranges delivery to the consignee (Wan et al., 1998).

National Customs Authority

National Customs authorities perform two basic functions: trade facilitation and customs control (Zhang et al., 2017). Customs controls includes intellectual property rights protection, the prevention of the infiltration of prohibited imports, such as illicit drugs or other dangerous substances, and also tariff collection (Zhang, 2003). Customs authorities normally receive an application for export approval from a shipper or a freight forwarder on his behalf. After checking the application customs issues an export clearance authority. Similarly, at the receiving country, customs receive an application for import clearance and, when issued, the CTO may release the consignment into the custody of a freight forwarder (Damsgaard, 1999).

Ramp Handling Agent

When aircraft are on the ground in between flights, they require various ground handling services to be performed, for example, aircraft loading/unloading, cargo handling, lavatory services, aircraft towing, or pushback (Kazda and Caves, 2015; Wu, 2010; Thompson, 2007; Damsgaard, 1999). Operator personnel collect the freight from the CTO operator and are responsible for loading the cargo on to the correct aircraft. At the destination, they unload the aircraft and deliver the incoming consignments to the airline’s contracted CTO’s terminal.

Airlines

An airline provides transportation for air cargo consignments from the airport of origin to the destination airport (Roebuck, 2013). In the global air cargo industry there are three principal types of air cargo-carrying airlines: combination passenger airlines, dedicated all-cargo airlines, and the integrators. Combination airlines include airlines that provide both passenger and air cargo services,

transporting air cargo in the lower deck belly holds of their passenger aircraft (Dresner and Zou, 2017; Morrell and Kleing, 2019). Most international carriers fall into this category. Combination airlines typically only offer airport-to-airport services and rely on air freight forwarders to perform the remaining transport logistics. Dedicated all-cargo airlines, such as Japan's Nippon Cargo Airlines, operate dedicated scheduled freighter flights on routes where there is regular heavy air cargo demand or where the potential passenger traffic is not large. These airlines typically carry air cargo only on an airport-to-airport basis and emphasize long-haul services (Dresner and Zou, 2017). The integrators are firms that integrate the air and ground transport services traditionally performed by separate firms, for example, airlines, and freight forwarders, to provide a full door-to-door service (Merkert and Alexander, 2018). Most international carriers fall into this category. Combination airlines typically only offer airport-to-airport services and rely on air freight forwarders to perform the remaining transport logistics (Wensveen, 2016). Dedicated all-cargo service (Leinbach and Bowen, 2007). Furthermore, the integrators' business model not only focuses on the provision of a fully integrated door-to-door product offering but also on the provision of supply chain and logistics solutions, which are underpinned by highly advanced IT systems. The largest integrators are FedEx, UPS, and DHL Express. Charter airlines either operate ad hoc services when there is a special requirement, fly to unusual destinations not served by other types of airline or operate scheduled services on behalf of specific clients. The cargo space of the entire or part of the aircraft can be chartered (Roebuck, 2013; Sales, 2016).

Most air cargo is carried in containers or ULDs, with the remainder transported loose in dedicated aircraft lower deck belly-hold compartments (Morrell and Kleing, 2019). On wide-body flights the consignments are loaded into special airline containers (ULDs) normally made of aluminum or on pallets (Ashford et al., 2013; Damsgaard, 1999).

Consignees

The importer or consignee is the receiver of the goods. An importer may be a company or an individual. The importer is normally responsible for the destination processes, such as, customs clearance and product distribution (The International Air Cargo Association, 1998).

The Provision of Air Cargo Capacity

Combination Aircraft Belly-Hold Air Cargo Capacity

As a by-product of aeronautical design, space is created below the main passenger deck of aircraft where passenger's luggage, cargo, and mail can be carried. When passengers are carried on the aircraft main deck, and cargo is carried below in the lower deck "belly hold" compartments, it is described as combination aircraft (Dempsey and Gesell, 1997). Importantly, the design of passenger aircraft is dictated by passenger requirements, space for air cargo is simply what is left over in the otherwise unusable space below the main passenger deck of the aircraft that is not required for the stowage of passengers luggage and that exists simply due to the aerodynamic requirements for a tubular shape for the aircraft fuselage. Air cargo is also a by-product in the sense that it is normally carried by combination airlines to where passengers want to travel, which is not necessarily where the air cargo wants to go. Combination carriers' networks are predicated on where passengers desire to travel to. Some combination airlines that operate dedicated freighter services tend to operate these aircraft to the same airports as their passenger services due to operational and cost synergies (Tretheway and Andriulaitis, 2010).

Combination airlines air cargo capacity may come in the form of narrow bodied, single-aisle aircraft, such as, the Airbus A320 aircraft, or wide-bodied, twin aisle aircraft, such as the Airbus A330-300 aircraft. Other wide-bodied aircraft include the Boeing B787, 777, and 747 aircraft as well as the Airbus A380 aircraft.

Combination airlines cargo operations are principally long haul, with large volumes of cargo being transshipped onto shorter-haul feeder services. The high utilization of long-haul passenger aircraft often justifies the purchase (or lease) of new aircraft for these services by the combination airlines (Button and Slough, 2000).

Freighter Aircraft Air Cargo Hold Capacity

By removing all of the unnecessary passenger-related facilities, thereby saving weight, a Boeing B747-400 freighter can carry a cargo payload of 113 tons (Boeing Commercial Airplanes, 2010); the same aircraft with a main passenger deck and lower deck belly-hold cargo has a typical payload of approximately 60–70 tons. The world's freighter fleet is divided into three freighter aircraft size categories: small capacity/standard body freighters, medium size freighters, and large wide-body sized freighter aircraft (Dahl, 2009; Morrell and Kleing, 2019).

Safe Transportation of Dangerous Goods by Air

Dangerous goods are defined as any articles or substances which are capable of posing a risk to health, safety, property or the environment and which are shown in the list of dangerous goods as published in the Technical Instructions or which are classified according to those Instructions (International Civil Aviation Organization, 2020c). The provisions of Annex 18 (A18) to the 1944 Chicago Convention on International Civil Aviation govern the international transport of dangerous goods by the air freight mode (International Civil Aviation Organization, 2020a). The broad provisions of A18 are amplified by the detailed specifications of the Technical Instructions for the Safe Transport of Dangerous Goods by Air (ICAO Document 9284) (International Civil Aviation

Organization, 2020d). That is, the legal provisions are set down in A18, while the technical guidance is given in ICAO9284 (Abeyratne, 2012). In general, the transportation of dangerous goods by air is forbidden, unless it is done so in accordance with ICAO9284. The definition of dangerous goods is specified in ICAO9284 (Blumenkron, 2017). Further to this, certain goods are identified as forbidden, which can be carried under certain exceptions (e.g., infected live animals), while others can be forbidden under any circumstance (explosives) (International Civil Aviation Organization, 2011a).

Dangerous goods must be packed as stipulated in A18 and ICAO9284. Packaging used for the transport of dangerous goods as air cargo must be of good quality, well constructed, and closed securely to prevent leakage. The packaging must also be appropriate for the contents; such that they are resistant to any chemical reaction or other detrimental effect of the contents. Material and construction specifications, and testing requirements are also laid out in ICAO9284 for the packaging. Inner packaging must be used to secure and/or cushion the contents to prevent any breakage, leakage, or to control their movement within the outer packaging. Packaging cannot be reused until it has been inspected and found free from corrosion or other damage. If packaging is reused, measures must be taken to prevent contamination of subsequent contents. Also, no harmful amount of a dangerous substance should adhere to the outside of packages (International Civil Aviation Organization, 2011a, p. 23).

Packages of dangerous goods being sent by air must be labeled with appropriate labels and in accordance with the provisions set forth in ICAO9284. These packages are required to be marked with the proper shipping name of its contents and, when assigned, the appropriate UN number and such other markings as may be specified in ICAO9284. Packages manufactured to specification contained in ICAO9284 must be marked in accordance with those provisions (International Civil Aviation Organization, 2011a). To that end, packages that do not meet the appropriate packaging specification contained in ICAO9284 cannot be marked as if they do.

Prior to someone offering a package or overpack (a container with one or more packages) of dangerous goods for transport by air, they must ensure that the dangerous goods are not forbidden for transport by air and are properly classified, packed, marked, labeled, and accompanied by a properly executed dangerous goods transport document, as specified in A18 and ICAO9284. The individual who offers dangerous goods for transport by air must complete, sign, and provide a dangerous goods transport document to the operator, which must contain the information required by ICAO9284. This transport document must include a declaration signed by the individual who offered the dangerous goods for transport by air. This declaration must indicate that the dangerous goods are fully and accurately described by their proper shipping names and that they are classified, packed, marked, labeled, and in proper condition for transport by air in accordance with ICAO9284 (International Civil Aviation Organization, 2011a).

The carriage of dangerous goods by air is included in the scope of the operator's safety management system (Annex 19 to the 1944 *Convention on International Civil Aviation*). Operators also need to comply with the responsibilities outlined in A18. An operator must not accept dangerous goods for transport by air unless:

1. the dangerous goods are accompanied by a completed dangerous goods transport document (the exception being where ICAO9284 indicate that it is not required); and
2. the package containing the dangerous goods has been inspected in accordance with the acceptance procedures contained in ICAO9284 (Abeyratne, 2012; International Civil Aviation Organization, 2011a, p. 29).

Furthermore, an operator is required to develop and use an acceptance checklist to aid compliance with the provisions of A18 Chapter 8.1. Packages and overpacks containing dangerous goods and aircraft unit load devices (ULDs) containing radioactive materials must be loaded and stowed on an aircraft in accordance with ICAO9284. Packages, overpacks, and ULDs containing dangerous goods (including radioactive materials) must be inspected for evidence of leakage or damage prior to being loaded on an aircraft. Leaking or damaged packages, overpacks, or ULDs must not be loaded on an aircraft. In the case where any package containing dangerous goods on an aircraft appears to be damaged or leaking, the operator must remove this from the aircraft, or arrange for its removal by an appropriate authority or organization (for hazardous substances for example). Following this, the operator must ensure that "similar" packages are in proper condition for transport by air and have not been contaminated. After the flight, these packages, overpacks, and ULDs must be reinspected for signs of damage or leakage while being unloaded. If damage or leakage is discovered, the area where the goods were stowed on the aircraft must also be inspected for damage or contamination (International Civil Aviation Organization, 2011a).

Packages containing dangerous goods, which might react dangerously with one another, should not be stowed either next to each other or in a position that would permit them to interact in the event of leakage. Packages containing toxic and infectious substances must be stowed on an aircraft in accordance with ICAO9284. Packages comprising radioactive materials must be stowed on an aircraft so that they are separated from persons, live animals, as well as undeveloped film, in accordance with ICAO9284. When dangerous goods are loaded in an aircraft, the operator must protect the dangerous goods from being damaged. The operator must also secure such goods in the aircraft in such a manner that will prevent any movement in flight which would change the orientation of the packages. In addition, packages of dangerous goods bearing the "cargo aircraft only" label shall be loaded in accordance with ICAO9284 (International Civil Aviation Organization, 2011a, p. 30).

The International Air Transport Association (IATA), the world's peak airline body, works closely with ICAO and local governments in the development of regulations for the carriage of dangerous goods by air (International Air Transport Association, 2020a). The IATA Dangerous Goods Regulations (DGR) is the worldwide reference for shipping dangerous goods via air and is the only standard recognized by airlines. The IATA DGR is a "field manual" version of the ICAO9284. The DGR draws from the industry's most trustworthy air cargo sources to help shippers classify, pack, mark, label, and document shipments of dangerous goods. In

addition, the DGR includes international dangerous goods air regulations, as well as state and airline requirements ([International Air Transport Association, 2020b](#)).

In the United States, the Federal Aviation Administration Hazardous Materials Regulations (HMR) prescribe the minimum requirements for the safe transportation of dangerous goods by air. These regulations describe how dangerous goods are classified, communicated, handled, and stowed. The HMR is published in Subchapter C of Title 49 of the Code of Federal Regulations (49 CFR, parts 171-180). The US Department of Transportation “Pipeline and Hazardous Materials Safety Administration” is responsible for the HMR ([United States Federal Aviation Administration, 2018](#)).

Substantive Safety

Overview of Historical Air Cargo Safety Occurrences

The ASN, a service of Flight Safety Foundation, includes within its accident database a total of 222 unique safety occurrences identified as air cargo occurrences, dating from 1941 to 2018. These are then further divided into five broad occurrence types which are issues commonly associated with air cargo occurrences; specifically, overloaded aircraft, wrong center of gravity, load shift in flight, fire or smoke (from the cargo consignments), and a leak (again of the cargo consignments) ([Aviation Safety Network, 2020b](#)). [Fig. 3A](#) shows the breakdown of these occurrence types. The dominant issue for air cargo accidents is clearly overloading the aircraft, which accounts for 53% of all the occurrences.

Broadly speaking, aviation safety occurrences can either result in fatalities or not ([Hoel et al., 2008](#)). According to ICAO, the percentage of official accidents in the last decade that has resulted in fatalities was 14%. From [Fig. 3B](#), 59% of air cargo related accidents were fatal. This means that accidents that are the result of issues associated with air cargo are more than four times more likely to result in fatalities. Similarly, we can also look at the outcome of accidents for the airframe. That is, while fatality is the metric used for people onboard the aircraft, the severity of an accident will also affect the aircraft as well, in terms of the level of damage. [Fig. 3C](#) shows that the outcome for the aircraft involved in air cargo occurrences is also significant, with only 5% of aircraft in air cargo related accidents being able to be repaired.

Another important characteristic of a safety occurrence to consider is in which phase of flight it happened. Typically, accident rates see a small spike at take-off (and initial climb out) and a significant spike at landing (and final approach), referred to as the critical phases of flight. That is, the risk of an accident is moderate (before flight), then high during take-off, then lower during cruise, and the high again for final approach and landing. But as seen in [Fig. 3D](#), for air cargo occurrences the risk of occurrences starts high during take-off, remains high through the climb, and is again high during the flight, and reduced for approach and landing. While the difference is interesting, it is not surprising. The nature of the accidents being the result of issues with the cargo means probabilistically the occurrence should happen earlier during the flight (the wrong CoG would manifest itself at take-off, etc.), and also during the longest time interval which corresponds to cruise (where issues such as fire associated with the cargo consignment).

The previous four characteristics are specifically detailed by ASN in the summary data of the accidents in the database. Further to this, the accident descriptions provided by ASN were coded qualitatively, to identify the EASA safety issues, as well as the ICAO occurrence category. First, the safety issue is used by EASA to assess the contributing factors in an occurrence. These are grouped into human factors (HFs), organizational (Org), equipment problems (EP), or environmental (E), which each in turn having more subcategories. Typically, HFs are associated with 80% of all aviation accidents ([Ferguson and Nelson, 2014](#)). We note here in [Fig. 3E](#) that in air cargo occurrences HFs contributes to 95% of all occurrences. Second, according to ICAO (2011b), occurrence categories are used to classify safety occurrences broadly to enable data analysis in an effort to the support of safety initiatives. There are 34 occurrence categories. As shown in [Fig. 3F](#) the most apparent of these is the RAMP category, which is used to classify occurrences that are the result of activities on the ramp in and around the aircraft. Noting that multiple categories apply to a single occurrence, RAMP was applicable to almost 89% of occurrences. The most common “outcome” for air cargo occurrences was loss of control in flight (LOCI), which occurred in 58% of occurrences, followed by runway excursions (RE) which occurred in 13% of occurrences. In continuing order of significance the other occurrence categories associated with air cargo occurrences are fire post impact (F-POST), fire nonimpact (F-NI), controlled flight into terrain (CFIT), system component failure – powerplant (SCF-PP), collision during take-off and landing (CTOL) with an obstacle, abnormal runway contact (ARC), and icing (ICE).

Rates of Air Cargo Accidents

[Fig. 4](#) shows the rates of accidents per million departures for major US carriers. The average for the six passenger airlines is approximately 1 accident per million departures. The rate for UPS and FedEx is below this, and they are the two largest air cargo operators in the United States. Looking at Atlas Air, a small and growing air cargo operator, it has an accident rate in excess of 4 times the average for passenger airlines. As such, we cannot say that air cargo operations are less safe than passenger operations, but there is clearly a distinction that needs to be made based on the size of the organization, and then the safety culture within that organization. In this example, it is important to note that Atlas Air is being compared to airlines which are much larger. It is a well-known fact that on average the smaller an operator is, the greater the risk involved with that operation. The key example of this is that fact that accident rates are much higher in general aviation compared to regular passenger transport ([Li, 2010](#)). If we continue and include additional small US air cargo operators (from largest to smallest), Kalitta Air has an accident rate of 4.6, Lynden Air Cargo Airlines has an accident rate of 8.3, and finally Centurion Air Cargo has an accident rate of 41 per million departures. The final point here is if

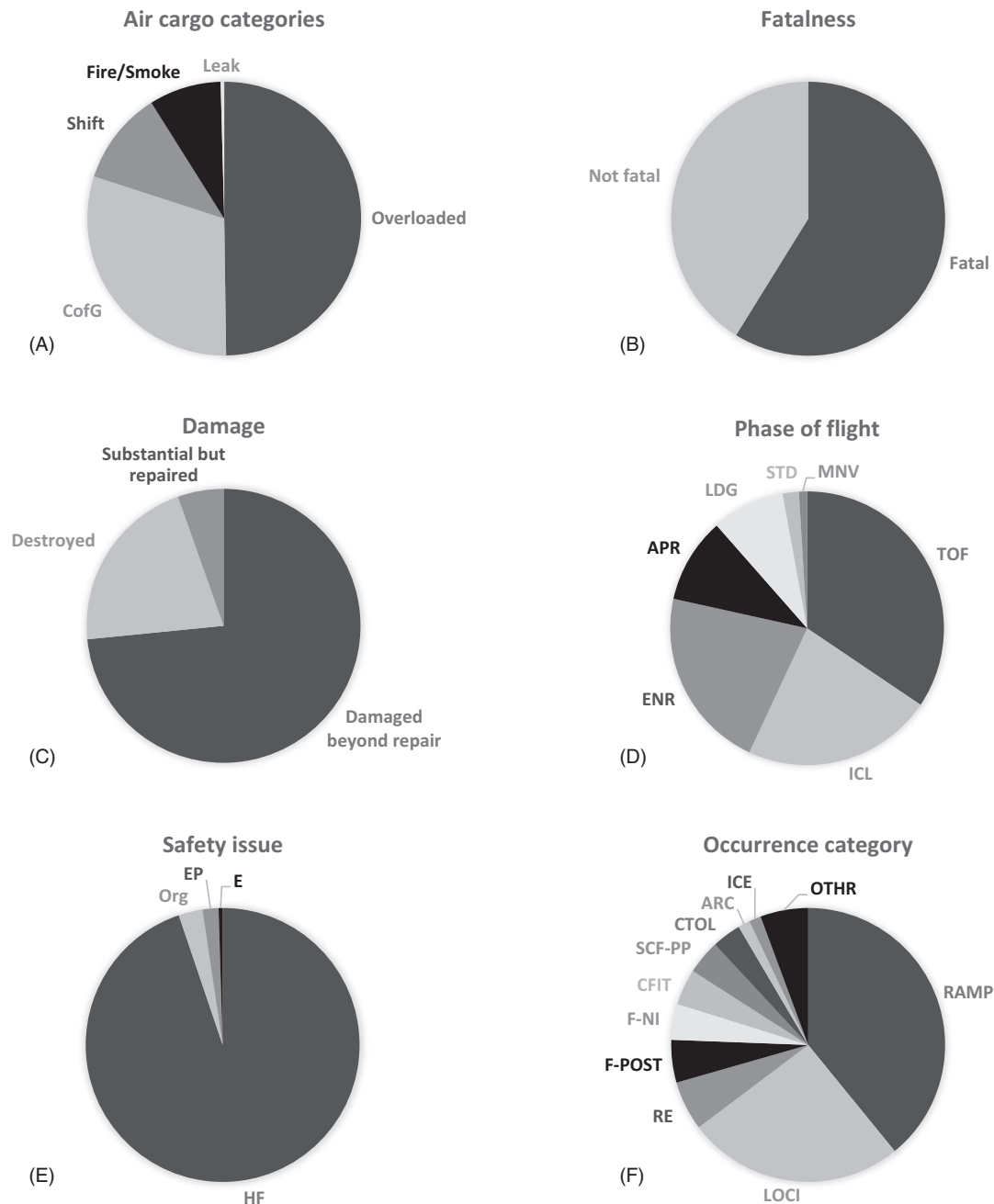


Figure 3 Pie charts showing the breakdown of cargo accidents by (A) occurrence types, (B) fatalness, (C) aircraft damage, (D) phase of flight, (E) EASA safety issue, and (F) ICAO occurrence category. Source: Adapted from [Aviation Safety Network \(2020a\)](#).

the data set is divided chronologically in half (Jan 2008 to Jun 2013, and Jul 2013 to Dec 2018), there is a dramatic improvement in accident rates over time. Specifically, for FedEx with four accidents, three occurred from 2008 to 2012, while one occurred in 2015. For UPS, two occurred from 2008 to 2012, while one occurred in the second half of 2013. Finally, for Atlas, two occurred from 2008 to 2012, and one occurred in 2018. During the same time period, the number of departures for each airline also increased, further reducing the accident rates.

Case Examples

Potentially the most infamous example of an air cargo safety occurrence is the tragic accident of National Airlines Flight 102 ([Bibel and Hedges, 2018](#)). This fatal accident was the result of human error in the securing of the cargo, which shifted during take-off, changing the aircraft center of gravity, in addition to physically damaging the aircraft controls needed to counteract the change in

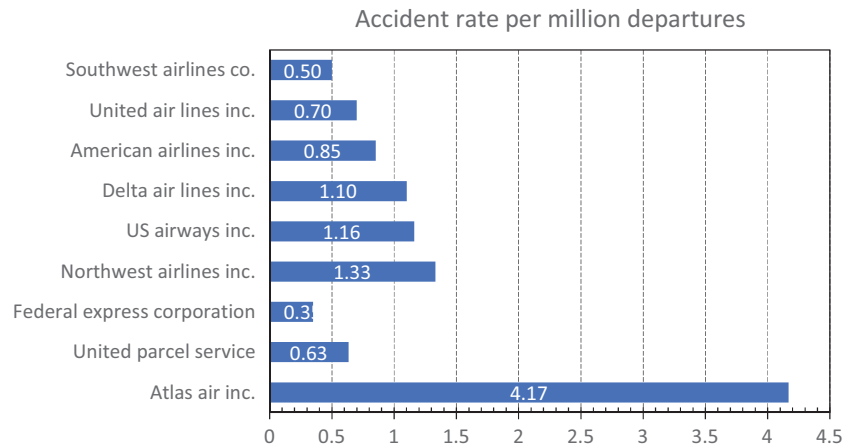


Figure 4 ICAO Official accident rates per million revenue departures for major US carriers, from 2008 to 2018 (inclusive). Accident count from ICAO (2020a). Departure data from the [US Bureau of Transportation Statistics \(2020\)](#), including all domestic and international departure to and from the United States.

stability. While not the typical example outlined above (this would be an overload case), National Airlines 102 still exemplifies the common characteristics identified earlier; they are dominated by HFs, begin at the ramp, tend to result in LOCI, occur early in the flight (take-off here), and end fatally for the crew onboard with a destroyed aircraft.

According to [Kharoufah et al. \(2018\)](#), a classic case of poor HFs resulting in an accident would involve a Russian Aircraft operated in Africa. Victoria Air 3C-LLA is an example of this from 2001. This is a typical overload example, occurring during cruise. Again, the ramp category applies as the aircraft was overloaded in contradiction to proper weight and balance requirements for the aircraft, this time with a system component failure powerplant resulting again in a loss of control inflight. The accident is fatal with the aircraft damaged beyond repair ([Aviation Safety Network, 2020b](#)).

Conclusion

Dedicated air freight operations commenced in 1941, building on the passenger and postal operations. The industry grew rapidly since the 1960s, going from less than 3 billion FTKs to over 200 billion FTKs since 2016. The evolution of aircraft used in air transport has helped facilitate the growth of that air freight market, which in turn resulted in the development of dedicated all freight aircraft and airlines, in 1972 and 1973, respectively.

There is a significant and diverse number of “links” in the global air cargo supply chain. The most obvious of these are the shipper or consignor and the receiver or consignee. Currently the transport of goods by air is not done from door to door, so road transport services will be involved, typically facilitated by trucking firms. To avoid a consignor (or consignee) needing to engage with multiple entities for a “package” to be delivered from door to door, a specialist freight forwarder is typically utilized. As airlines have a vested interest in carrying goods, they have GSAs to capture business. An airport is an essential part of the supply chain, and at the airport dedicated air cargo terminals can exist for the express purpose of handling freight. These terminals will typically have national customs authorities integrated within them to ensure all goods in and out clear customs. The interface between the terminal and the aircraft is facilitated by ramp handling agents. The aircraft itself is operated by an airline, either a dedicated cargo airline, or combination airline providing both passenger and air cargo services.

The regulatory requirements for the international transport of dangerous goods by the air freight mode are given in the provisions of A18 to the 1944 *Chicago Convention on International Civil Aviation*. These broad provisions are supported by the detailed specifications of the Technical Instructions for the *Safe Transport of Dangerous Goods by Air* (ICAO Document 9284). These documents are essential to guarantee the safe transport of dangerous goods by air, and to ensure safety in general in the aviation industry when air cargo is involved.

Acknowledgment

The authors received no support of financial assistance.

References

- Abeyratne, R., 2012. *Air Navigation Law*. Springer, Berlin.
- Al-Hajri, G., 1999. *The Impact of Sea-Air Mode on Air Cargo Transport*. A-Z Group Limited, Merstham, UK.
- Ashford, N.J., Martin Stanton, H.P., Moore, C.A., Coutu, P., Beasley, J.R., 2013. *Airport Operations*, third ed. McGraw-Hill, New York.
- Aviation Safety Network, 2020a. ASN Aviation safety database. Available from: <https://aviation-safety.net/database/>.

- Aviation Safety Network, 2020b. Safety issue list. Available from: <https://aviation-safety.net/database/events/event.php?code=CG>.
- Baxter, G.S., Bardell, N.S., 2017. Can the renewed interest in ultra-long-range passenger flights be satisfied by the current generation of civil aircraft? *Aviation* 21 (2), 42–54.
- Beifert, A., 2016. Role of air cargo and road feeder services for regional airports—Case studies from the Baltic Sea Region. *Transp. Telecommun.* 17 (2), 87–99.
- Bibel, G., Hedges, R., 2018. *Plane Crash: The Forensics of Aviation Disasters*. John Hopkins University Press, Baltimore, MD.
- Blumenkron, J., 2017. International safety requirements. In: P.S. Dempsey and R. Jakhu (Eds.), *Routledge Handbook of Public Aviation Law*, Routledge, Abingdon, UK, pp. 33–63.
- Boeing Commercial Airplanes, 2006. *World Air Cargo Forecast 2006 | 2007*. Boeing Commercial Airplanes, Seattle, WA.
- Boeing Commercial Airplanes, 2010. 747-400/-400ER Freighters. Available from: https://www.boeing.com/resources/boeingdotcom/company/about_bca/startup/pdf/freighters/747-400f.pdf.
- Bruins, G., 2006. *International Freight Forwarding in Canada*. Trade Evaluation and Analysis Division, International Markets Bureau, Agricultural and Agri-Food Canada, Ottawa, Canada.
- Button, K.J., Stough, R., 2000. *Air Transport Networks: Theory and Policy Implications*. Edward Elgar Publishing, Cheltenham, UK.
- Caves, B., 2015. Cargo. third ed. In: Kazda, A., Caves, R.E. (Eds.), *Airport Design and Operation*, third ed., Emerald Group Publishing, Bingley, UK, pp. 233–260.
- Cavusgil, S.T., Rammal, H., Freeman, S., 2010. *International Business: The New Realities*. Pearson Australia, Frenchs Forest.
- Chiavi, R., 2017. Airfreight development supporting the strategy of global logistics companies. In: Delfmann, W., Baum, H., Auerbach, S., Albers, S. (Eds.), *Strategic Management in the Aviation Industry*, Routledge, Abingdon, UK, pp. 489–515.
- Corley, W., 2002. Trade logistics 101: An introduction to forwarding. *Exp. Am.* 3 (8), 16–18.
- Dahl, R.V., 2009. Cargo jet arena retrenches. *Aviation Week Space Technol.* 170 (4), 53–56.
- Damsgaard, J., 1999. Global Logistics System Asia Co., Ltd. *J. Inform. Technol.* 14 (3), 303–314.
- Dempsey, P.S., Gesell, L.E., 1997. *Air Transportation: Foundations for the 21st Century*. Coast Aire Publications, Chandler, AZ.
- Doganis, R., 2010. *Flying Off Course: Airline Economics and Marketing*, fourth ed. Routledge, Abingdon, UK.
- Dresner, P., Zou, L., 2017. Air cargo and logistics. In: Budd, L., Ison, S. (Eds.), *Air Transport Management: An International Perspective*, Routledge, Abingdon, UK, pp. 247–264.
- Ferguson, M., Nelson, S., 2014. *Aviation Safety: A Balanced Industry Approach*. Delmar, Clifton Park, NY.
- Hayuth, Y., 1983. The evolution and competitiveness of air cargo transportation: The case of Israel's airborne trade. *Transp. Rev.* 3 (3), 265–286.
- Hoel, L.A., Garber, N.J., Sadek, A.W., 2008. *Transportation Infrastructure Engineering: A Multimodal Integration*. Nelson Publishing, Toronto.
- Hosie, P., Lim, M.K., Chng, M., 2012. The impact of general sales agents on the air cargo industry. *Int. J. Logist. Syst. Manag.* 13 (3), 393–416.
- Hui, G.W.L., Hui, Y.V., Zhang, A., 2004. Analysing China's air cargo flows and data. *J. Air Transp. Manag.* 10 (2), 125–135.
- International Air Transport Association, 2020a. Dangerous goods: Setting the standards leads to safety. Available from: <https://www.iata.org/en/programs/cargo/dgr/>.
- International Air Transport Association, 2020b. Dangerous goods: Setting the standards leads to safety. Available from: <https://www.iata.org/en/publications/dgr/>.
- International Civil Aviation Organization, 2011a. Annex 18 to the Chicago Convention on Civil Aviation: The Safe Transport of Dangerous Goods by Air, fourth ed. Montreal, ICAO.
- International Civil Aviation Organization, (2011b). *Aviation Occurrence Categories: Definitions and Usage Notes*. www.icao.int/APAC/OccurrenceCategoryDefinitions.
- International Civil Aviation Organization, (2018). *The World of Air Transport in 2017*. <https://www.icao.int/annual-report-2017/Pages/the-world-of-air-transport-in-2017.aspx>.
- International Civil Aviation Organization, 2020a. Accident Statistics. Available from: <https://www.icao.int/safety/iStars/Pages/Accident-Statistics.aspx>.
- International Civil Aviation Organization, 2020a. Annex 18. Available from: <https://www.icao.int/safety/DangerousGoods/Pages/annex-18.aspx>.
- International Civil Aviation Organization, 2020c. Dangerous Goods. Available from: <https://www.icao.int/safety/airnavigation/OPS/CabinSafety/Pages/Dangerous-Goods.aspx>.
- International Civil Aviation Organization, 2020d. Technical Instructions on The Safe Transport of Dangerous Goods by Air (Doc 9284). Available from: <https://www.icao.int/safety/DangerousGoods/Pages/technical-instructions.aspx>.
- Kazda, A., Caves, R.E., 2015. *Airport Design and Operation*, third ed. Emerald Group Publishing, Bingley, UK.
- Kharoufah, h., Murray, J., Baxter, G., Wild, G., 2018. A review of human factors causations in commercial air transport accidents and incidents: From 2000-2016. *Prog. Aerosp. Sci.* 99, 1–13.
- Kupfer, F., Meersman, H., Onghena, E., Van de Voorde, E., 2011. World air cargo and merchandise trade. In: Macário, R., Van de Voorde, E. (Eds.), *Critical Issues in Air Transport Economics and Business*, Routledge, Abingdon, UK, pp. 98–111.
- Leinbach, T.R., Bowen, J.T., 2007. Transport services and the global economy: Towards a seamless market. In: Bryson, J.R., Daniels, P.W. (Eds.), *The Handbook of Service Industries*, Edward Elgar Publishing, Cheltenham, UK, pp. 209–226.
- Li, G., 2010. Airline accidents. In: Fink, G. (Ed.), *Stress of War, Conflict and Disaster*, Academic Press, San Diego, CA, pp. 731–734.
- Merkert, R., Alexander, D., 2018. The air cargo industry. In: Halpern, N., Graham, A. (Eds.), *The Routledge Companion to Air Transport*, Routledge, Abingdon, UK, pp. 29–47.
- Milenkovic, G., 2001. Early warning of organizational crises: A research project from the international air express industry. *J. Commun. Manag.* 5 (4), 360–373.
- Morrell, P.S., Kleing, T., 2019. *Moving Boxes by Air: The Economics of International Air Cargo*, second ed. Routledge, Abingdon, UK.
- Murphy, P.R., Daley, J.M., 1996. A preliminary analysis of the strategies of international freight forwarders. *Transport. J.* 35 (4), 5–11.
- Ohashi, H., Kim, T.S., Oum, T.H., Yu, C., 2005. Choice of air cargo transshipment airport: An application to air cargo traffic to/from North East Asia. *J. Air Transp. Manag.* 11 (3), 149–159.
- Reynolds-Feighan, A.J., 2001. Air freight logistics. In: Brewer, A.M., Button, K.J., Hensher, D.A. (Eds.), *Handbook of Logistics and Supply Chain Management*, Pergamon, Amsterdam, pp. 431–439.
- Roebuck, M., 2013. The air freight supply chain. In: Sales, M. (Ed.), *The Air Logistics Handbook: Air Freight and the Global Supply Chain*, Routledge, Abingdon, UK, pp. 1–15.
- Ryan, R., 2008. Cargo: A necessary evil for airports? *J. Airport Manag.* 2 (2), 132–136.
- Sales, M., 2013. *The Air Logistics Handbook: Air Freight and the Global Supply Chain*. Routledge, Abingdon, UK.
- Sales, M., 2016. *Aviation Logistics: The Dynamic Partnership of Air Freight and Supply Chain*. Kogan Page Limited, London, UK.
- Smith, S.M., 2001. Measuring the people side of FedEx Express. *J. Org. Excel.* 20 (4), 11–18.
- The International Air Cargo Association, 1998. *The TIACA Manifesto*. The International Air Cargo Association, Miami, FL.
- Thompson, B., 2007. Ground handling opportunities for airports. *J. Airport Manag.* 1 (4), 393–397.
- Tretheway, M.W., Andriulaitis, R.J., 2010. Airport competition for freight. In: Forsyth, P., Gillen, D., Müller, J., Niemeier, H.M. (Eds.), *Airport Competition: The European Experience*, Routledge, Abingdon, UK, pp. 137–150.
- United Kingdom Department of Transport, 2000. *UK Air Freight Study Report*. Department for Transport, London, UK.
- United States Bureau of Transportation Statistics, 2020. Air Carrier Statistics (Form 41 Traffic)- U.S. Carriers. Available from: https://www.transtats.bts.gov/Tables.asp?DB_ID=110&DB_Name=Air%20Carrier%20Statistics%20%28Form%2041%20Traffic%29-%2020U.S.%20Carriers.
- United States Federal Aviation Administration, 2018. Dangerous goods regulations for air transportation. Available from: <https://www.faa.gov/hazmat/resources/regulations/>.
- Wan, Y.W., Cheung, R.K., Liu, J., Tong, J.H., 1998. Warehouse location problems for air freight forwarders: A challenge created by the airport relocation. *J. Air Transp. Manag.* 4 (4), 201–207.
- Wensveen, J.G., 2016. eighth ed. *Air Transportation: A Management Perspective*, eighth ed., Routledge, Abingdon, UK.
- Wood, D.F., Barone, A., Murphy, P., Wardlow, D.L., 2001. *International Logistics*. Kluwer Academic Publishers, Norwell, MA.
- Wraight, S., 2017. Aircraft: The role of freighters – past, present and future. second ed. In: Sales, M. (Ed.), *Air Cargo Management: Air Freight and the Global Supply Chain*, second ed., Routledge, Abingdon UK, pp. 131–138.
- Wu, Y., 2010. *Airline Operations and Delay Management Insights from Airline Economics Networks and Strategic Schedule Planning*. Routledge, Abingdon, UK.
- Zhang, A., 2003. Electronic technology and simplification of customs regulations and procedures in air cargo trade. *J. Air Transp.* 7 (2), 88–102.
- Zhang, A., Hui, G.W.L., Leung, L.C., Cheung, W., Hui, Y.V., 2017. *Air Cargo in Mainland China and Hong Kong*. Routledge, Abingdon, UK.

Bicycle Collision Avoidance Systems: Can Cyclist Safety be Improved with Intelligent Transport Systems?

Lars Leden, Luleå University of Technology, Luleå, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	108
Methodology	108
Multiple-Comfort Model	109
Multiple-Need Model	110
Environmental Adaptation with Focus on ITS Issues	110
Feeling Confident and Secure (To Leave Home and Walk and Cycle)	111
Guidance (Leading or Navigating)	111
Getting Contact with and/or Being Localized	111
ITS Improving the Safety of Cyclists	112
Conclusions: Result of Safety Impact Assessment	113
References	114

Introduction

A conventional way for bicyclists—in this chapter referred to as cyclists—to avoid having collisions has been to equip the cycle with good brakes. When riding at low speeds, a rear brake may be enough. At higher speeds, to reduce braking distance, both wheels should have brakes. Also, locking up the front wheel can lead to loss of balance and crashing so antilock brakes are desirable but typically not provided. However, rim brakes typically do not lock the wheel unless very hard pressure is applied. In order not to get hit, the cycle or cyclist also needs to be seen, meaning good reflectors and illumination is important, the latter also helps the operator to see imperfections in the pavement. A cyclist should also have a bell or some other means to get the attention of other cyclists and pedestrians, and ideally of motorists as well.

Other characteristics that are important for cycling safety is that the rider is comfortable, can reach the ground without having to stretch or jump off the seat and that the rider has an upright position so they can see the roadway ahead. These are characteristics that also apply to cycling away from closed off roads. Racing cyclists obviously do not sit upright but they also should not be mixed with regular traffic. Finally, the most important characteristic may be to keep speeds low. Riding at speeds above 30 km/h can easily lead to serious injuries, if there is an incident. In the European Union (EU), electric cycles have a maximum allowed speed of 25 km/h for motor assistance and most of them have a hybrid system that maximizes the speed to roughly that downhill as well. Conventional pedal cycles do not have speed maximizers and that can create hazardous situations.

According to a survey with senior cyclists in Sweden (Leden, 2008; Leden et al., 2008), the most commonly used safety equipment is lights, which are, or at least “were” at the time of the survey, used by 81% of the respondents. Most common are battery-powered lights followed by traditional dynamo-operated ones where the generator touches the tire. Some respondents have a dynamo in the hub. The second most common equipment is a helmet, which is used by 80% of the elderly. The remaining fifth does not own one. About two-thirds of the respondents use a bicycle-bag or basket, and the same share have reflectors mounted on the bicycle. However, reflective vests are used only by 17% of the respondents, though in rural areas the usage is close to 50%. Rear-view mirrors are used by a few respondents only, but are desired by quite many respondents (28%). Winter tires and winter cycles are desired by one-fifth of the respondents. However, more than half of the respondents stated that they do not miss any equipment or that they have no opinion.

This chapter is not looking at the above outlined conventional safety equipment but is looking ahead at what bicycle collision avoidance systems should be implemented within the next few years, or over a slightly longer time frame.

Methodology

The content here is based on literature surveys, analysis of in-depth crash data, road user questionnaires, and questionnaires with experts on Intelligent Transportation Systems (ITS) and the usage of such systems for satisfying functional needs. Mostly European users and experts were included. Surveys and interviews with children and elderly are limited to Sweden. Assessments included the potential for avoiding collisions with a focus on children and elderly as cyclists and as pedestrians, so called Vulnerable Road Users (VRUs). Children and elderly as pedestrians are included here to get a full overview of the shortcoming and needs of these groups.

For decades, the main societal rule has been to more or less neglect the needs of the VRU groups. The PhD dissertation of the author (Leden, 1989), about the safety of cycling children and the effect of the street environment was quite an exception at the time. The theme continued with the author as supervisor of a dissertation about children as pedestrians and bicyclists (Johansson, 2004).

The theses focused on making the traffic environment more manageable and understandable for children satisfying the functional needs of children. The topic was further elaborated on within the Pedestrians' Quality Needs (PQN) network of The European COST 358 Project (Leden et al., 2014), the framework of a Fulbright project, in the EU project HUMANIST (Leden, 2008), and finally in the context of EU projects like Vulnerable Road Users and ITS, VRUITS. Participation in the EU Humanist project in 2007 and 2008 had as a goal to achieve a competence concerning vulnerable road users and ITS, leading to research on needs of senior (older) cyclists. This started with exploring Finnish in-depth crash data (VALT) and testing a set of hypotheses to find out reasons behind the higher risks for senior cyclists compared to other age groups. This was then used as a background for developing bicycle collision avoidance systems targeted to aid senior cyclists.

In the HUMANIST project, exploration behind the higher risks for senior cyclists included a survey of more than 500 elderly (65 years or older) members of the Swedish Cycling Promotion organization Cykelfrämjandet, and an expert questionnaire was distributed during the Velo-city 2007 conference to get experts' views on ITS-measure actions to be taken to satisfy the requests of senior cyclists. Altogether, 14 experts participated in these in-depth interviews (Leden, 2008).

Multiple-Comfort Model

There are a lot of models that can be used to explore the safety situation in more detail. One is the *Multiple-comfort model* proposed by the Finnish researcher Summala (2005) and modified below to fit exploration of the safety of cyclists (Silla et al., 2017). According to the model, the following five issues are the most important ones for explaining road user behavior on a strategic, tactical, and operational (individual) level: Safety margins (to survive), good or expected progress of trips, rule following (according to the law and social rules), vehicle/road system (bicycle and infrastructure), and pleasure of driving and pleasure of cycling.

Safety margins: Safety margins imply a concept of available time. It is, for example, important to make cyclists and cars visible to each other, for instance through warning lights, signs or messages in the infrastructure or in-vehicle alarms to warn them about conflicting road users. Otherwise, especially in darkness, safety margins tend to be insufficient. In critical situations, combined pedestrian/bicycle detection systems and emergency braking could be needed.

Good or expected progress of trips: Good or expected progress of trips is an important issue for cyclists. Cyclists like to maintain their speed and may hesitate when it comes to braking. Therefore, there should be detectors well in advance of signalized intersections to give cyclists the possibility to get a green light without having to slow down or dismount their bicycles. And, gradients, especially downhill, are hazardous for cyclists as especially cyclists are reluctant to brake.

Rule following: The analysis of Finnish in-depth crash data which focused on fatalities revealed that 80% of the cyclists had not obeyed some rule. Though this figure is certainly biased due to the fact that the conclusions are often based on the surviving car drivers' statements, rule following is obviously critical also to cyclists. Laws and regulations should enhance and secure communication between road users. Harmonization of rules within countries such as in the United States as well as between nearby countries, such as within the EU, is an important issue. For example in Southern and Eastern Europe there is a ban on phone use when cycling; whereas in Northern Europe there is no ban, except in Denmark.

Vehicle/road system (bicycle and infrastructure): According to Summala, the vehicle/road system is designed, so that cars usually have smooth car/road performance. This is often not the case for the cycle/road system. A cycle design, for example, elderly cyclists based on new technology is lacking. One of many issues could be to implement a bicycle-to-car communication system to facilitate communication between road users. Furthermore, adequate bicycle infrastructure is often missing, even in Europe—except in the Netherlands and Denmark—and if it exists, it often does not comply with the best practice. To achieve best practice, the management and political framework is crucial. An example of a “success story” from Helsinki in Finland is shown in the Video 1. Low motor-vehicle speed achieved by Intelligent Speed Adaptation (ISA) or by other means is a prerequisite for safety. This issue will be elaborated further in the section Multiple-need model. ITS could be a way to improve the safety of cyclists for example at intersections through early detection and prioritizing of vulnerable road users. Other means to increase safety include warning signals or warning lights to warn cyclists of approaching motor vehicles or vice versa at intersections. Such warning devices could also be useful when a motor vehicle is approaching a cyclist from behind (or a cyclist is approaching a pedestrian, but then the sound has to be “gentle” so that pedestrians are not scared). ITS can be used to get better guidance for and visibility of bicyclists at nighttime, for example through LED-lights in the pavements or by increasing the intensity of street lighting at times when cycle traffic is present.

Pleasure of cycling: The pleasure of cycling is an important topic especially for senior cyclists as 84% of the respondents in the survey of the HUMANIST project stated that the joy of riding is a reason for them to cycle. Measures should keep or increase the pleasure of cycling, and the amount of cycling. Examples are green waves at coordinated signals for cyclists, information on vacancy on bicycle racks and intelligent bikes.

According to the elderly, the biggest safety problems are potholes, slipperiness, and bad snow removal; 76%, 74%, and 70% of the respondents have referred to these factors as safety problems. Other major problems are curbstones and cars going too fast.

What the elderly say would increase their cycling is linked to what they say is important for increased traffic safety. Increased safety would lead to increased cycling among the elderly. Requests dealing with the physical design of roads are especially a demand for more and better cycle tracks and cycle paths. Communication between road users expressed as more and better consideration are also perceived to increase their feeling of safety and thereby increase their cycling.

Overall, it is important to note that an increase in cycling (e.g., due to the use of ITS) could lead to more fatalities and injuries for cyclists if adequate countermeasures are not taken. However, the risk per cyclists decreases even in a given environment when we have more people riding bicycles. Often, when improving a city cycle system, there is a possibility to choose between conventional and ITS countermeasures.

Multiple-Need Model

The *Multiple-need model* was developed to explore the safety of cyclists from the perspective of satisfying needs. According to the model, the following four issues are the most important ones to explain road user behavior on a strategic, tactical, and operational (individual) level: Environmental adaptation (in this context with focus on ITS issues), confidence and security enhancements, guidance (leading or navigating) and getting contact with and/or being localized.

Environmental Adaptation with Focus on ITS Issues

Environmental adaptation with focus on ITS issues implies a change of driver behavior to the limitations of VRUs, especially for children and elderly, by informing and/or alerting the driver or the VRU of a danger. However, especially regarding children, the question has been raised, if the traffic environment ought to be relying on the use of ITS or not, and if so if ITS services should be provided to the driver, the vulnerable road user or both. Services that alert or inform children and elderly of a danger need to take the ability of those road users into account when designed and developed, introducing characteristics, which are attractive and user-friendly. The functionality of such systems could also affect pedestrian and cyclist behavior in a negative way. For example, the provision of such systems may lower the VRUs natural scanning for dangerous situations and thus lead to lower “alert” levels.

The speed of the vehicle involved in a collision is a determining factor for both the chance of an accident, and the severity of an accident involving a VRU. Examples of devices that have as a goal to reduce speeds include speed humps and other traffic-calming measures, speed-measuring systems with feedback to the driver on roadside displays, in-car systems that remind the driver of local (legal or advised) speed limits, and in-car systems that intervene to reduce the speed when someone tries to exceed the legal speed limit.

In-car systems that remind the driver of local (legal or advised) speed limits can be of different types. It is not surprising that audible warnings are recommended over visual ones. People react the quickest to touch (tactile message), second quickest to what they hear (audible message) and the slowest to what they see (visual message). This is obvious. Think about being at a busy location such as in an airport terminal, and someone touches your shoulder, someone calls out your name, or someone holds up a sign with your name. What would you be most likely to react too quickly? That audible messages are more effective than visual ones is probably true especially for children. VRUs can be warned on a device they carry with them. Such devices could be vibrating and give a tactile signal as well as an audible one.

Setting the legal speed limit to be “safe” is equally important. It has to be defined what “too fast” means when children and elderly are adjacent to or in traffic walking accompanied by and adult (Johansson and Leden, 2010). That preschool children should not encounter cars in their play areas or where they walk. Actual vehicle speeds should be below 20 km/h where children ages 7 to 12 years are crossing a street, especially if they are walking unaccompanied by an adult. This speed is often called human-powered speed, and is used in the context of “shared space” and areas marked by road signs such as “residential area.” Children above age 12 years should not cross at locations where motor vehicle speeds exceed 30 km/h. This applies to routes to school, to leisure activities, and when visiting friends. Elderly often have sight and/or hearing impairments and they move slower compared to other age groups, and sometimes they have mental-processing impairments. Therefore, older people are often as vulnerable as children. Also for adults, the actual vehicle speed should be a maximum of 20 km/h where there is a high risk of collision between vehicles and unprotected road users, for example at high volume unsignalized locations. In general, the accepted maximum safe speed where there are bicyclists mixed with cars is 20–30 km/h (Kröyer, 2015). To secure travel speeds below 20 km/h, additional measures like ISA or camera enforcement of speeds near crossing points might be needed.

Further development of intelligent “platforms,” such as speed-activated trapdoors type Edeva, is another option to secure safe speeds of 20 or 30 km/h. Ideally, a street should have a speed-reducing device that is clearly felt by drivers going above the desired speed but the street should be kept flat and comfortable for road-users traveling at or below the desired speed. This is of special benefit to standing passengers in buses and to patients traveling in ambulances.

More futuristic examples of ITS could be based on holograms aiming at shaping virtual areas or “walls.” A more conventional example of this idea is shown in Fig. 1. The virtual area is shaped with two stop lines: the first 10 m before and the second adjacent to the crosswalk. In this case, the “wall” is shaped by road markings on the ramp on an elevated crosswalk. The impression of a wall could be further developed, for example, using holograms or LED-lights to emphasize the vertical dimension and to alert about crossing VRUs.

At approaches with two lanes or more, multiple-threat conflicts occur due to vehicles overtaking stopped ones in the adjacent lane. These conflicts are a threat especially to children, as they often are hidden behind the stopped vehicle, if it has stopped close to a crosswalk. Advanced yield bars or stop lines at a distance to the crosswalk of about 10 m is recommended to alert drivers to stop and not to overtake a stopped car or truck in the adjacent lane, when a VRU is crossing (Fisher and Garay-Vega, 2012; Leden et al., 2018). However, installing yield or stop lines is not yet an option available in many countries. For example, according to Finnish and Israeli regulations among other countries, yield or stop lines cannot be installed except for at signalized crosswalks, though such lines are



Figure 1 Crosswalk protected by a virtual area and “wall” in Playa de las America, Tenerife.

already used at nonsignalized crossings in, for example, Spain, Japan, and the United States. ITS systems such as Collision Avoidance and Warning Systems and Pedestrian and Cyclist Detection System with Emergency Braking are also feasible for alerting drivers to stop and to stop early, and not to overtake a stopped car in an adjacent lane.

Feeling Confident and Secure (To Leave Home and Walk and Cycle)

Personal safety and security should be an obvious right in any civilized society; and a person's, or their parent's, fear of them having an accident or being assaulted is one of the limiting factors for leaving home to walk and cycle. In most communities, there is a higher risk of being injured in a traffic accident than from an assault. However, in some cities, the risk of becoming a victim of crime is greater than the risk of dying in a traffic accident. Children and elderly equipped with a mobile phone can call for help. Streets having video cameras or speed cameras monitored by police can improve the confidence and security even more.

Another important issue to feel confident is the design and equipment of the bike itself, an upright seating position and a low bike frame making it easy to climb on and off the bike. Some equipment facilitating turning left would be useful as many senior citizens have a stiff neck and bad balance, and are significantly ($p = 0.0012$) more involved in crashes when intending to turn left compared to other age groups. A rear-view mirror could help, as stated by senior cyclists, but improvements are also possible by designing the infrastructure, so that it becomes unnecessary to merge with motor vehicles when turning left.

Guidance (Leading or Navigating)

There are many ITS solutions that can be used for guidance (leading or navigating). For navigation, we can distinguish between two groups of devices: those based on GPS and others, the latter are normally used to complement the first in order to improve their accuracy getting a seamless positioning. The experts answering the questionnaire at the Velo-city 2007 conference mentioned two main potential guidance applications: Safe Route Guidance and Safe Crossing Guidance through warnings. Half of the experts that expressed an opinion (five out of 10) stated that route guidance systems, advising people about their safest choice, would improve safety at least for adult VRUs.

Almost all experts suggested a digital map for online route guidance when cycling, and also for trip planning before the trip starts. On-line devices like Personal Digital Assistants (PDAs) could also be used, for example, to get local weather information or to find time tables for public transport and especially to see whether it is allowed to bring the bike on a train, tram, or bus. A special design of the devices making it easy for elderly to use them was considered crucial. The following automatic types of equipment for bikes were considered important to test and further develop: automatic locking and opening, for example, at a distance by using the key as for cars, automatic gears, automatic turning on and off of bicycle lamps (with power supply from a reliable dynamo), and automatic elevating of the saddle after mounting.

Child or other VRU route might take them to a dangerous crossing point. Seven out of 10 experts suggest that VRUs are given route directions, on a GPS or cell phone, to have them walk/bike around crossing points that are dangerous with respect to vehicle traffic. But there are alternatives to avoiding dangerous locations. The crossing points can also be made less dangerous with the use of technology. Six experts suggest that streets should have video cameras that process images of objects and people in a street, and, if a child is detected, in-vehicle alarms are triggered in nearby vehicles. Five experts suggest that children carry emitting devices with signals picked up by a detector in the vehicle and that an alarm is triggered. This can either be done whenever a child is within X meters from the vehicle, where X can vary with vehicle speed (3 experts) or dangerous proximity can be determined by a GPS device when a child is already in the roadway (2 experts).

A different strategy than avoiding dangerous crossing points, or mitigating their danger, is to change the child's destination. This can obviously not always be done, but rather than trying to find the safest way to the closest convenience store or bus stop, maybe there is another one slightly further away that can be reached in a safer way, and quicker than the nearby one when following safe routes.

Getting Contact with and/or Being Localized

GPS systems permitting a person to establish contact are useful from a safety perspective especially for demented elderly people, who suddenly do not know where they are. Such systems can enable relatives to localize them. There are systems that send out SMS

messages to all designated contacts with the user's exact current position. But for children and older people in general, a regular mobile phone is by most experts regarded to be more than enough.

ITS Improving the Safety of Cyclists

Within the VRUITS project, a list of 23 ITS was drawn up (Silla et al., 2017). This included all ITS that were deemed to be near to market and to have good potential for improving the safety, mobility, or comfort of VRUs. Based on a multi-criteria analysis in the VRUITS project, five systems were estimated to have a large effect on improving the safety of cyclists. Intelligent Speed Adaptation, ISA, of cars would be the most efficient measure to provide a safe environment, if enough political support is available to implement the measure and ensure safe speeds. Recently, this seems to be the case in the European Union with suggestions to mandate ISA on new model cars starting in 2022. However this was not the case when the list of 23 ITS was drawn up.

Below follows a short description of each assessed system. We can distinguish between two kinds of systems dependent on whether the detector is mobile (goes in the vehicle/mounted on pedestrian/cyclist clothing) or fixed.

Blind spot detection (BSD) uses vehicle sensors to detect pedestrians, cyclists, and mopeds in blind spots near cars, trucks, and buses. The system addresses mainly the side areas of the car/truck/bus, but optionally also the front and rear of the car/truck/bus. After a detection, the system provides a warning to the driver. The system does not intervene. The system aims to prevent accident between cars, trucks, and buses and VRUs in the blind spot of the car/truck/bus. Note that the blind spot can be on either side of the vehicle (Fig. 2).

A *bicycle to vehicle communication (B2V)* system informs and warns the car driver about cyclists on the road in the vicinity of the car/truck/bus, and the cyclist of potential collisions with nearby cars/trucks/buses. Both the car/truck/bus and the cyclist need to be equipped with a cooperative device to send and receive messages with cyclists, and with a GPS device to determine the relative locations. Cyclists can receive information on their mobile devices (e.g., smart phones). The system does not intervene. The system aims to prevent all accidents between cars/trucks/buses and cyclists due to inattention of car/truck/bus driver or cyclist. All these road users need to be equipped for the system to work well (Fig. 3).

The *intersection safety (INS)* system assists drivers and vulnerable road users in avoiding common mistakes which may lead to typical intersection accidents. It covers the functions left- and right-turning assistance and vehicles arriving perpendicular to VRUs. Left- and right-turning assistance concerns the case where a vehicle is turning left or right into the path of the VRU. A roadside unit (RSU) detects the VRU crossing or approaching the intersection via camera or radar, assesses the risk of a collision, and sends a warning of a potential collision to the vehicle. The vehicle driver is informed or gets the warning via an on-board unit, depending on the urgency of the situation. The roadside infrastructure also informs the VRU of the danger by, for example, flashing lights and/or sound.

Vehicles arriving perpendicular to a VRU concerns the case where the vehicle drives perpendicular to the path of the cyclist. The process is the same as for the first function: an RSU detects the VRU crossing the intersection and informs the vehicle about the possible collision with the VRU. The system depends on short-range communication and roadside sensors. Optionally, the vehicle



Figure 2 Blind spot detection (BSD).



Figure 3 Bicycle to vehicle communication (B2V).

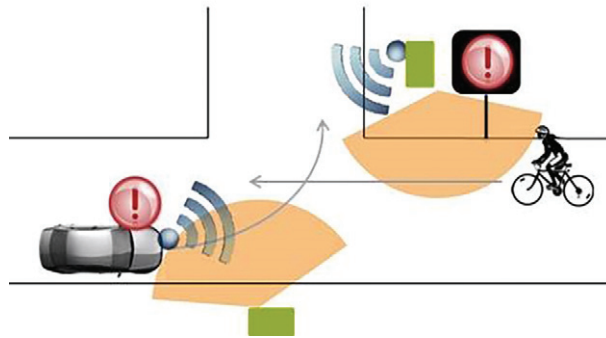


Figure 4 Intersection Safety (INS).



Figure 5 Pedestrian and cyclists detection system + emergency braking (PCDS + EBR).

also uses its own sensors. The system does not intervene. The system aims to prevent accidents between cars/trucks/buses and VRUs at signalized and nonsignalized intersections (Fig. 4).

Pedestrian and cyclist detection system can be combined with *Emergency braking (PCDS + EBR)*. PCDS + EBR is a vehicle built-in system which is aimed at preventing or reducing the severity of vehicle-pedestrian/cyclist crashes by using detection sensors that will detect pedestrians/cyclists in front of a forward-moving vehicle (also known as Automatic Emergency Braking [AEB]). If a crash is likely, the system warns the driver and if the driver fails to respond in time and the collision risk remains, the system will intervene through automatic braking. It is assumed that the system will prevent accidents in which the drivers would not have observed the pedestrian/cyclist otherwise or would not have reacted in time. Therefore, the system will mainly prevent the accidents related to the inattention of car drivers. For speeds up to 35 km/h, the system is considered to be able to prevent collision; for higher speeds (up to 50 km/h), the system should reduce the impact of the vehicle-pedestrian/cyclist crashes by reducing the car speed (Fig. 5).

The *VRU Beacon system (VBS)* consist of a tag or device carried by a VRU that sends out a signal, which is subsequently detected by a receiving device installed in vehicles. This system calculates the trajectories of the detected VRU in relation to the vehicle trajectory and assesses the possibility of a collision. The driver is then warned about a potential collision without any active intervention. Targeted vulnerable road user groups are pedestrians and bicyclists as well as Powered Two-Wheelers (PTWs). The VBS system detects not only persons/vehicles in front, but also those at the rear and side of the vehicle. The VBS system is also able to detect possible collisions when there is no line of sight between vehicle and VRU. Scenarios, which VBS is mainly addressing, are critical situations in urban areas, where motorized traffic is traveling at speeds up to 50 km/h, where obstructed views or unexpected behavior of VRUs can potentially lead to conflicts. The system aims to prevent accident between cars/trucks/buses and vulnerable road users (Fig. 6).

Conclusions: Result of Safety Impact Assessment

The author together with colleagues explored the safety impact of systems, which were estimated to have a high potential to improve the safety of cyclists, especially: Blind Spot Detection (BSD), Bicycle to Vehicle communication (B2V), Intersection safety (INS), Pedestrian and Cyclist Detection System + Emergency Braking (PCDS + EBR), and VRU Beacon System (VBS).

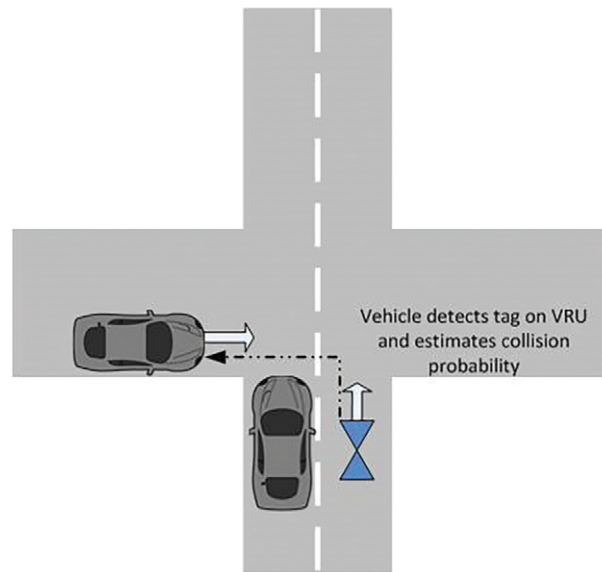


Figure 6 VRU beacon system (VBS).

The main results of the assessment showed that all investigated systems affect safety of cyclists in a positive way by preventing fatalities and injuries. The estimates considering 2012 accident data and full penetration showed that the highest effects could be obtained by the implementation of PCDS + EBR and B2V whereas VBS had the lowest effect. The estimated yearly reduction in cyclist fatalities in EU-28 varied between 77 and 286 per system. A forecast for 2030, taking into accounts the estimated accident trends and penetration rates, showed the highest effects for PCDS+EBR and BSD.

The results of this study show that the selected ITS have a high potential to improve the safety of cyclists. It should be noted that the impact depends on the penetration rates. Therefore, in order to realize the potential safety effects the deployment of most promising systems should be supported by different stakeholders and decision makers. As indicated earlier, our estimates include some uncertainty. Therefore, in order to improve the accuracy of the estimates, there is a need for better accident data (on number and details of the accidents; including also hospital records) and trials to test the functioning of the systems and their effect on road user behavior.

References

- Fisher, D., Garay-Vega, L., 2012. Advance yield markings and drivers' performance in response to multiple-threat scenarios at mid-block crosswalks.. *Accident Analysis and Prevention* 44, 35–41.
- Leden, L., Johansson, C., Rosander, P., Gitelman, V., Gärder, P., 2018. Design of crosswalks for children: a synthesis of best practice. *Trans. Transp. Sci.* 9 (1), 20–33., doi:10.5507/tots.2018.004.
- Leden, L., Gärder, P., Schirokoff, A., Monderde-i-Bort, H., Johansson, C., Basbas, S., 2014. A sustainable city environment through child safety and mobility—a challenge based on ITS? *Accid. Anal. Prev.* 62, 406–414, doi:10.1016/j.aap.2013.06.013.
- Leden, L., Johansson, C., Rosander, P., Leden, L., 2008. Elderly cyclists' opinions on safe and joyful cycling. *Proceedings of the ICTCT 21 Conference in Riga 2008*.
- Leden, L., 2008. What tools are needed to develop safe and joyful cycling for senior citizens. *Proceedings of the HUMANIST conference in Lyon April 3–4, 2008*.
- Leden, L., 2008. Safe and Joyful Cycling for Senior Citizens. Espoo, VTT Working papers 92.
- Leden L. 1989. The safety of cycling children. Effect of the street environment, VTT Publications 55, Espoo
- Johansson, C., 2004. Safety and Mobility of Children Crossing Streets as Pedestrians and Bicyclists. Doctoral Thesis 2004:27. Luleå University of Technology, Luleå.
- Johansson, C., Leden, L., 2010. Child Pedestrians' Quality Needs and how these needs relate to interventions. *Walk21, ICTCT November 17–19 The Hague*.
- Summala, H. 2005. Towards understanding driving behavior and safety efforts. *Proceedings of the international workshop on modeling driver behavior in automotive environments*.
- Silla, A., Leden, L., Rämä, P., Scholliers, J., Van Noort, M., Bell, D., 2017. Can cyclist safety be improved with intelligent transport systems? *Accid. Anal. Prev.* 105, 134–145.
- Kröyer, H.R.G., 2015. Accidents between Pedestrians, Bicyclists and Motorized Vehicles: Accident risk and injury severity. Doctoral thesis. Department of Technology and Society. Lund University. Lund.

Bicycle Infrastructure

Rock E. Miller, Orange, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	115
History of Bicycle Infrastructure	115
Forms of Infrastructure	118
Consideration for European Design Approaches	120
Level of Stress Classification System	121
Bicycling and Safety	122
Other Bicycle Infrastructure	123
Biography	124
See Also	124
References	124
Further Reading	124

Introduction

While there is no accepted definition of bicycle infrastructure, most persons consider it to be the improved travel way or the transportation facility upon which the bicyclist operates their bicycle. Infrastructure can also include support facilities such as bicycle racks, bicycle repair tool stations, and facilities to support sharing or short-term rental of bicycles. This article does not address primitively improved trails and terrain that are used for mountain bicycling, a popular recreational activity.

History of Bicycle Infrastructure

The bicycle evolved into today's shape and function in the late 1800s, when it was known as the safety bicycle. It quickly surged into popularity as a significant device for urban mobility and recreational travel soon after this date. It was affordable, safe and simple, fun to operate and maintain, and provided faster transport speeds than walking or horse-drawn transportation. This boom in bicycling contributed heavily to good roads movements in many countries. Bicyclists were among the first to advocate for paving of dirt and gravel roads and pathways to provide smoother all-weather surfaces for riding. These paved roads became the earliest forms of bicycle infrastructure, but this effort was occurring prior to the popularization of automobiles (Reid, 2015).

This bicycle boom turned into a bicycle bust by the mid-1920s in North America and in the 1950s in Western Europe. Adult urban bicycling became overwhelmed by availability, convenience, and comfort of affordable automobiles. Safety for bicyclists became compromised by collisions with cars, and bicycling was reduced to a recreational activity or a child's toy for the next 50 years in the United States and countries.

From 1920–70, specialized infrastructure for cycling in cities and other areas was largely nonexistent throughout the world. But a continued strong tradition for adult bicycling for transportation persisted in much of Scandinavia and in the Netherlands. In the 1960s, key northern European cities began to construct high-capacity transportation facilities resembling the United States freeways and high-capacity urban highway systems. These efforts to carve wide space for highways through established cities met strong local resistance, and bicycling advocates joined in opposition, in part, due to the increased dangers caused by rising automobile usage in city centers. As an alternative, bicycle advocates pushed for the development of specialized bicycle infrastructure that would allow safer bicycling in the city center as well as for more pleasant rural excursion riding. Since this time, bicycling infrastructure has been introduced and enhanced in much of Northern Europe, led by the Netherlands and its many cities, plus Copenhagen and other Scandinavian countries. Today, Copenhagen claims the largest daily bicycle commuting mode share of any Western city, and the Netherlands has the highest rate of any developed country. Figs. 1 and 2 show typical examples of bicycle infrastructure in Amsterdam and Copenhagen. While there are a variety of factors involved, the commitment of these communities to provide a comprehensive network of bicycle infrastructure gets much of the credit.

European bicycle capitals take varying approaches in developing bicycle infrastructure, but they are generally built around providing physical separation between bicyclists and motor vehicles, where appropriate and possible. Drivers and bicyclists are generally planned and expected to share road space only on low-speed, low-volume roadways. In addition, bicyclists may be separated from pedestrians, especially in densely developed or commercial areas. Where an urban roadway provides separate physical space each for motor vehicles, bicycles, and pedestrians, the area devoted to bicycles was defined or translated into English as a cycle track (The United States has also used the terms separated bikeway or protected bikeway for similar facilities). Complete European networks are also supported by bicycle trails or paths that are isolated from or placed away from motor vehicle traffic.



Figure 1 Amsterdam.



Figure 2 Copenhagen.

Embracing bicycle infrastructure was much slower to come to the United Kingdom, southern Europe, and North America. In the early 1970s, a few US college campuses began to explore the concept of developing a network to facilitate bicycling on or near their campus. The college town of Davis, California, is credited with the invention of the bicycle lane in 1968. Similar facilities may have existed in Europe at the time, but the US concept of a bicycle lane, providing separation by a painted stripe on the pavement for a wide variety of roadway speeds, traffic volumes, and access conditions, represented a departure from the European approach of physical separation.

The reluctance in the United States to introduce bicycle infrastructure comparable to Europe is often attributed to the efforts of a single individual, John Forester. He was a successful professional bicycle racer who appreciated the opportunity to ride bicycles at high speeds that could be achieved most comfortably on wide traffic lanes that were intended for cars, typically 12 feet (3.6 m). He became concerned that bicyclists might be legally forced to ride in separate “inferior” bicycle infrastructure that did not facilitate bicycle-racing speeds. He developed and popularized a bicycle riding technique in a book, *Effective Cycling*. The method was built around educating bicyclists to take control of travel lanes developed for automobiles. His teachings advocated riding in the center of the automobile lane and forcing motor vehicles to pass by changing lanes. His technique was embraced by many members of the United States League of American Wheelmen who developed classes to educate novice cyclists in its usage (Forester, 2012).

Forester and company became strong advocates against bringing European infrastructure to the United States, and their members dominated many US standards developing organizations. Forester’s philosophy was strongly opposed to bicycle lanes and to physically separated facilities that were located so close to roadways that bicyclists might be prohibited from using the nearby roadway. His teachings and methods may provide for safer operation of bicycles in a traffic environment that provides no separate accommodation for bicycling, but the technique does not appeal to most casual or inexperienced bicyclists. A similar pattern of evolution may also have occurred in the United Kingdom over the same time period.

By the mid-2000s, bicycling was rising to become a substantial component of urban transportation in the leading European bicycle capitals where high quality infrastructure was being provided. Copenhagen claimed that approximately 50% of all commuter trips were by bicycle, and the college town of Groningen, NL, claimed a mode share of over 60%, while most US cities were typically



Figure 3 Montreal.

under 1% and Davis saw the highest rate of any US city, at about 25%. Enthusiasts felt that the European approach to infrastructure with its emphasis on separation may be a strong causal factor.

While bicycle infrastructure development stagnated for decades in most of North America, Montreal, Canada, became the first North American city to begin to import bicycle infrastructure that was similar in design to Europe. Facilities resembling cycle tracks began to appear in Montreal by the early 1990s, and bicycle commuting and recreational usage became the highest of any large city in the United States or Canada (**Fig. 3**).

The US cities, led by New York City, began the process of importing European design principles and adapting them to the United States conventions, beginning in about 2005. Ninth Avenue in Manhattan became the first European inspired bikeway to be constructed in the heart of a major US city, using advice provided by Gehl Architects, a prominent international architecture and planning firm based in Copenhagen. New York continues to be among North American leaders in the deployment and perfection of the United States bicycle infrastructure (**Fig. 4**).

Many other large US cities followed the lead of New York and began introducing bicycle infrastructure in their city centers and key access routes. Canadian cities followed the lead of Montreal, and bicycle infrastructure began to spread quickly through other areas



Figure 4 NY City.



Figure 5 Seattle.



Figure 6 Budapest.

of Europe and the United Kingdom (Fig. 5). This trend continues today, and bicycle usage is found to increase with the provision of improved infrastructure virtually everywhere that it is introduced (Figs. 6 and 7).

Forms of Infrastructure

The United States has traditionally divided bicycle infrastructure into 3 classes, as defined by the American Association of State Highway and Transportation Officials (AASHTO) Bikeway Design Guide. (AASHTO, 2012). Class I facilities are defined as bike paths or trails, which are physically isolated away from roadways that serve motor vehicles. They are often built along levees, waterways, or abandoned railroad lines. They are typically 8 feet (2.4 m) wide and work best when wide enough for pairs of bicyclists to pass each other while moving in opposite directions riding side-by-side (Fig. 8). Rural bicycle trails can become popular tourist recreational facilities, and economic benefit studies for long rural trails show great promise. Urban trail networks are often found in communities with higher bicycle usage. They provide comfortable riding conditions for most users except when they become crowded with pedestrians and other users or when they have poorly designed crossings of busy roadways. But opportunities to introduce trail systems within built up communities can be limited, and they will not support increase in bicycle commuting, if they do not connect with desirable origins and destinations.

Class II facilities are known as bicycle lanes. These are areas of the roadway that are defined by striping in the street to provide limited separation from other vehicles. They can be striped next to the edge of the roadway or can be placed between motor vehicle travel lanes and parking lanes. They are minimally 5 feet (1.5 m) wide and are more comfortable to use when wider, but motorists may mistake them for automobile travel lanes when they exceed 8 feet (2.4 m). They can use a green coating, but coloring is not required by the US standards. It is generally accepted that bicycle lanes work best when placed on low-speed roadways and on roadways that do not allow auto parking (Fig. 9). Bicycle lanes can be found on many high-speed roadways, but users do not find them attractive in these settings, and they seldom produce increased bicycling usage.



Figure 7 London.



Figure 8 Bike trail.



Figure 9 Green bike lane.



Figure 10 Bike Blvd.

Class III facilities are known as bicycle routes. They are minimally roadways that also allow automobile to travel in the same space. The signs provide guidance to bicyclists and advise motorists to expect bicycle usage. They work best on quiet residential streets, but bicycle route signs have been posted on high-speed rural highways to provide for long distance bicycle travel. There are no design specifications for a Class III facility, so a wide variety of quality can be expected.

Trends in bicycle infrastructure have developed tools to enhance Class III facilities. The shared lane marking, or sharrow, is a pavement marking developed in the United States and placed within the center of the travel lane to advise motorists of bicycle activity and to encourage bicyclists to ride toward the center of the roadway. These markings were advocated by Forester and company. They are not popular with other cycling groups who advocate for strong infrastructure. They are mostly used on roadways with speed limits of 35 mph (or 50 km/h) or less.

The bicycle boulevard is another adaptation of the Class III facility. These are typically low volume residential streets that use traffic calming techniques to slow or discourage motor vehicle traffic, while optimizing the route condition for bicycling, in part by minimizing the use of stop signs. These routes tend to be popular to a wider audience of bicyclists, but city street systems can provide only limited opportunities to implement them. Portland, Oregon, and Vancouver, Canada, both have extensive networks of bicycle boulevards that have helped those cities to achieve higher than average bicycle commuting. These facilities are also known as neighborhood greenways in some communities. The bicycle boulevard concept can also be found in Europe, as shared streets or auto-free zones (Fig. 10).

Facilities comparable to European Cycle Tracks have not appeared in the most recent versions of the AASHTO guide, and this may explain in part the low rate of evolution of bicycle infrastructure in the United States. But this is expected to change in 2020 when the next edition of the AASHTO Bicycle Design Guide is planned to be released. The United States Class I, II, and III system has some recognition in Canada, but the bicycle trail, lane, route terminology is more widely used, and the design of facilities is generally like the US facilities meeting these class descriptions also can be found in Europe, but European guidance for when to best use these treatments is stronger and more precise. It discourages bicycle lanes when traffic volumes or speeds exceed specific thresholds.

Consideration for European Design Approaches

There is not agreement among leading European communities on a single best way to design a facility, but it is generally agreed that North American guidelines are primitive or inferior to European practices. The *Crow Manual*, a bicycle design guide published in the Netherlands, is a commonly used and respected design guide for applying European approaches (CROW Platform, 2016). The NACTO Bikeway Design Guide and other publications by the US National Association of City Transportation Officials (NACTO, 2014) are among the most widely recognized bicycle guides in the United States, but new publications on the subject are being completed regularly.

A tour of Europe reveals a wide variety and evolution of bicycle infrastructure as communities move from systems that were built for two travel modes, cars and pedestrians, to infrastructure that supports bicycling as a third mode. Some communities have bisected wide sidewalk areas to create an area for cycling and an area for walking, using paint or textures, as a first step toward dedicated bicycle infrastructure. Others have widened their sidewalks or constructed a new pathway area between the sidewalk and the roadway for use by bicyclists (Fig. 11).

Communities with the most experience and highest bicycle usage have generally developed design approaches that reconstruct the area of roadway to be occupied by bicyclists. They have developed intersection traffic management strategies that reduce or limit conflicts between motorists and bicyclists, often featuring roundabouts or bicycle traffic signals. They have also developed roadway design features that were crafted to address safety issues apparent in the older and more primitive designs.

A typical cycle track in the Netherlands is typically at least 2 m wide (6.5 feet) and may be separated from both the motor vehicle travel way and the sidewalk by planting strips to clearly delineate the facility. Parked cars may be found on the street side, but a raised curb is virtually always provided between the parking lane and the bikeway. Also, the bikeway surface uses red pigmentation to



Figure 11 Paint on sidewalk.



Figure 12 Plastic posts.

differentiate it from other areas of the roadway. This design approach is well suited to new development outside of the traditional city centers. Opportunities to provide this treatment may be found for routes in or near city centers, where roadways may have been widened in the 1960s to provide more travel lanes but repurposed years later to provide dedicated cycling space. In developed areas with limited space, cycle tracks are typically one way and provided on each side of two-way roadways. They can be provided as two-way facilities in more suburban environments where cross traffic is lower, and access is infrequent.

An evolution toward European configurations is visible in the United States and Canada, where many facilities use temporary plastic traffic posts and pavement marking to introduce separation. This treatment is economical, and some communities have been willing to provide temporary facilities as a demonstration test to see if the facilities will induce bicycling and not severely affect motorized traffic (**Fig. 12**). Over next 20 years many of these temporary looking facilities will likely be reconstructed with more permanent features to more closely resemble the best European facilities (**Fig. 13**).

Level of Stress Classification System

A different set of characteristics for classifying bicycle infrastructure is growing in popularity. It is based upon the level of stress or comfort felt by bicyclists when they use a facility, and not as much by its physical characteristics ([Mekuria et al., 2012](#)). Ranging from LS (Level of Stress) -1 to LS-4, it is a measure of how users will rate a facility. LS-1 facilities provide a comfortable and quality experience for all users and may encourage more persons to travel by bicycle. These can include separated facilities, paths, and bicycle boulevards. LS-4 facilities are generally not found attractive to most potential users. Many existing bicyclists will not use them, and they can discourage novice users from becoming regular users. Bicycle lanes on quiet streets can be rated LS-2, while bicycle lanes on busy, fast streets are rated LS-3.



Figure 13 Redondo Beach bikeway.



Figure 14 Cycle snake bridge.

Communities that are concentrating on building or upgrading facilities to LS-1 or LS-2 are generally finding that bicycle usage rises with the coverage of the network. The low stress facilities may have special designations, such as an All Ages and Abilities (AAA) network. A bicycle mode share of 10% of commuting trips may be reasonable where a grid of low stress facilities allows cyclists to ride comfortably to most destinations within 3–5 miles (5–7 km).

While many communities have been nervous about trading space between automobiles and bicycling, separate and new infrastructure for bicycling (and walking) is seeing substantial investment. Many bridges over roadways and waterways have been funded primarily for semi-exclusive use by bicyclists. Well-known examples include the Hovenring in Eindhoven NL, the Cycle Snake in Copenhagen, and two bicycling bridges in Calgary, Canada. Funding has been set aside for other ambitious projects, and many will likely to become landmarks when completed (**Figs. 14 and 15**).

Bicycling and Safety

Bicyclists are vulnerable road users. Riders can be susceptible to injury and death when struck by motor vehicles. Safety considerations have governed and regulated the evolution of bicycle infrastructure in many countries. Studies of early European cycle tracks suggested that they might not be as safe as other types of bicycle infrastructure. More modern designs have probably addressed concerns over the rate of occurrence of incidents, especially for bicyclists traveling at high speeds, but any potential for a death or serious injury can raise questions.

It is widely accepted that bicycling can produce health benefits, and economic studies have suggested that these health benefits greatly outweigh travel injury risks. Also, the rate of incidence of serious injury or fatality while cycling, even in automobile traffic, is comparable to the rate of incidence of driving automobiles, if measured per hour or per trip. Finally, studies have suggested that



Figure 15 Calgary bridge.

cycling may benefit from “safety in numbers.” As the number of cyclists increases, collision rates between cyclists and other road users’ trend downward. This effect has been confirmed across countries and cities, but it has not been studied as extensively as cycling increases in a specific community. Part of this effect is explained by motorists and cyclists adjusting their behaviors as they more regularly interact with each other. Another factor could stem from an overall reduction in motor vehicle travel compared to population.

Other Bicycle Infrastructure

Communities that wish to encourage bicycling have generally found the importance of having attractive and comfortable bikeways that can be used for most short trips. But other infrastructure is often needed to encourage growth in bicycling. Perhaps highest in importance is bicycle parking. Bicycle racks or storage space must be convenient, functional, and plentiful. Careful and proper planning for bicycle parking minimizes the potential for bicycle theft and keeps the urban environment orderly. Train stations in key European capitals can offer parking for 1000’s of bicycles. These can include underground facilities where users can check in their bicycles with an attendant for maximum security.

In the United States, where bicycling has not risen to levels seen in Europe, many programs to integrate bicycling with transit are in place. These include bicycle racks on buses and areas on trains dedicated to carrying bicycles with their owners. The ability to maintain these programs when bicycling rises to more than 10% of all trips may be difficult in the future.

Lack of access to a bicycle is also a limitation to increases in bicycling. Short-term bicycle rental (bike sharing) has been introduced into over 100 cities throughout the world over the past 15 years. Earlier systems required docking of bicycles at specific locations when not rented. In the past 5 years, dockless systems have become more popular (**Fig. 16**). These feature GPS location information and fare payment on the bicycles themselves.



Figure 16 Bike share.

Electrical power propulsion assist for bicycles is also becoming a new factor. E-bikes have become very popular for new bicycle sales in Europe and have attracted new riders who tend to be older or less fit users. E-bikes minimize the effect of hilly terrain and perspiration. They can make commuting more practical for many persons. E-scooters are not considered as bicycles, but their infrastructure requirements are quite like biking needs, and they have grown in popularity very quickly in large cities. Users discovering the added convenience of an electrically powered personal vehicle may migrate from scooters toward E-bikes which are more practical for longer trips.

Biography

Rock Miller is a registered Civil Engineer based in the United States with over 40 years of work experience in Traffic Engineering. He has focused on establishing standards to allow for better bikeway design in the United States and has designed many innovative facilities in California and Alberta, Canada, and has overseen facilities in other states and regions. He has focused on bikeway engineering and traffic safety for the past 10 years, and he has studied European designs to determine how to adapt them to the US practices. Rock is a Past President of the Institute of Transportation Engineers, a 15,000-member North American-based professional association. He is currently a voting member of the US National Committee on Uniform Traffic Control Devices and a member of the Transportation Research Board Bicycle Research Committee.

See Also

Bicycle Collision Avoidance Systems: Can Cyclist Safety be Improved with Intelligent Transport Systems?

References

AASHTO, 2012. *Guide to the Development of Bicycle Facilities*, American Association of State Highway and Transportation Officials, Washington DC.
CROW Platform, 2016. *Design Manual for Bicycle Traffic* (Crow Manual), Utrecht NL.
Forester, J., 2012. *Effective Cycling*, seventh ed. MIT Press.
Mekuria, M.C., Furth, P., Nixon, H., 2012. *Low Stress Bicycling and Network Connectivity*, Mineta Transportation Institute.
NACTO, 2014. *Urban Bikeway Design Guide*, second ed. Island Press.
Reid, C., 2015. *Roads Were Not Built for Cars*, Island Press.

Further Reading

Pucher, J.A., Ralph, B., 2012. *City Cycling*, MIT Press.
Bruntlett, M., Bruntlett, C., 2018. *Building the Cycling City*, Island Press.

Bicycles, E-Bikes and Micromobility, A Traffic Safety Overview

Höskuldur Kröyer, Trafkon, Sweden

© 2021 Elsevier Ltd. All rights reserved.

The Road User, the Vehicle, and the Infrastructure	125
The Road User, Children and Seniors	125
The Cycle	126
The Infrastructure	127
The Accidents	127
Accident Location and Time of Year	128
Individual Factors, Age, and Gender	129
Vehicle Type	129
Speed	129
Alcohol and Drugs	129
Visibility and Road Condition	130
Injuries Sustained by Bicyclists	130
Safety Equipment	130
The Special Case of E-Bikes and Micromobility	131
E-Bikes	131
Other Micromobility Devices	131
System View: Exposure, Risk, and Consequence	132
Exposure	132
Risk	132
Safety in Numbers	133
Consequence	134
Practical Implications	135
System Perspective and the Individual Safety	135
Relation to the Use of Bicycle Helmets	135
The Health Loss Paradox	136
References	136

The Road User, the Vehicle, and the Infrastructure

The users of bicycles are a very heterogenic group. They range in age from toddlers riding small tricycles in front of their homes, to children that use the bike as a toy as well as a travel mode to friends or school, the “regular” cyclist that uses the bike for transportation or recreation, to people well above retirement age still riding for transportation or recreation. Also, the group ranges from the “passive” child passenger in a child seat or a wagon, to cyclists that use the bicycle or tricycle for delivery services, to athletes that use it for training or competitions at high speeds. *The vehicle (the cycle)* also differs. There are child bikes that have lower statue, classical bikes, bikes with child seats (front and/or back), bicycles with child wagon connected at back, tricycles with a box at front or back for children or merchandise, tricycles for adults, mountain bikes, low riding bicycles, racing bicycles, transportation bicycles, and e-bikes. In addition, there are multiple other micromobility devices such as (e-)-kick-bikes, (e-)-skateboards etc. *The infrastructure* is used by different combination of those road users and vehicles, so you can find different combinations or road users on different parts of the road system. There are diverse design solutions, and the solution used varies both between countries and within countries and even between different locations within the same city, creating a complex system for the cyclists to navigate. To ensure a safe cycling, the infrastructure, must allow for those different combinations to fulfill their needs in safe manners, where the traffic situation is a combination of those three elements of the transport system, see Fig. 1.

The Road User, Children and Seniors

When it comes to traffic, and traffic safety, children require special attention. Children’s sight and hearing is not fully developed (DaCoTa, 2012). Their perception of the traffic environment is not always “correct,” and the youngest ones often assume that the other road user is aware of them (DaCoTa, 2012). Young children have a hard time identifying dangerous crossing sites (Ampofo-Boateng and Thomson, 1991), and do not always make safe crossing decisions (Connelly et al., 1998; Plumert et al., 2004), especially at higher speeds (Connelly et al., 1998). Children can also easily be distracted, especially the youngest ones (Demetre et al., 1992), which can cause accidents (Axelsson and Stigson, 2019). It should therefore be clear that, even though their

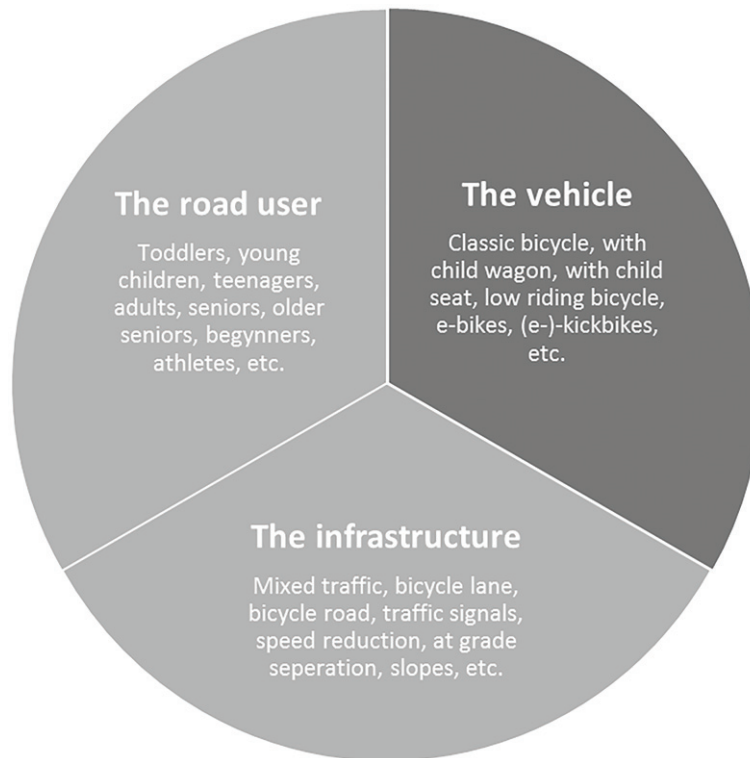


Figure 1 The three elements of the transport system.

preconditions improve with higher age, it influences young children's ability to safely navigate a complex traffic environment. Another safety related fact is that children are shorter, this means that they are more easily missed by other road users due to obstructions of sight (e.g., parked cars, railings), which is highly relevant for design of the infrastructure. We must accept that we cannot expect a child to perform in as safe manners as an adult, and therefore, adjust the transportation system so that it takes this into consideration.

Seniors is another important group for traffic safety. It is a well-known fact that with higher age, the ability to see and hear deteriorates, the body loses muscle mass and the strength of the bones is reduced. We can even see increased cognitive challenges such as Alzheimer's (Delin and Rudgren, 2007). Seniors also have highly elevated risk of both serious and fatal injuries in collisions between bicyclists and motor vehicles (e.g., Kaplan et al., 2014; Kröyer, 2015a). This creates some challenges. On one hand, we strive to increase mobility and active travel for seniors, this for improved health and quality of life. Simultaneously, by making this group cycle, we are making a relatively fragile group use a travel mode that has considerable risk of injuries and limited safety equipment. Since this group seems to be much more likely to be seriously or fatally injured and is a considerable part of those who suffer from traffic accidents, they need special focus.

The Cycle

The type of cycle has some bearing on traffic safety. There is however, rather limited research available regarding this issue. Rodgers (1997) through self-reported response, showed elevated risk for those (1) who rode an off-road trails opposed to roadways, (2) who rode racing bikes, or (3) all terrain style bicycle, compared to general purpose bikes. There are several aspects where bicycle type might influence the traffic safety, for example:

- **Visibility:** the height and structure of the bicycle relates to the cyclist sight lines and the visibility towards other road users. A low riding bicycle is more easily hidden by other elements of the traffic environment (e.g., vegetation, railings).
- **Crash mechanics:** The height and seating position of the cyclist will influence the crash mechanics. Lower stature is more likely to results in forward projection (i.e., full contact with the front of the vehicle) compared to a higher cycle. If the cyclist is sufficiently low, it might also increase the risk of an overrun accident.
- **The maneuverability and space requirement** differs between different bicycle types. We can also expect that the speed is higher on for example racing bicycles and some E-bikes compared to a box cycle. It should be noted that it is not the average speed of a bicycle trip that is of primary concern, but the speed at the time of the conflict/accident, and especially top speeds coinciding with conflict areas. Therefore, E-bikes that support up to speeds of 25 km/h might move slower downhill than traditional bicycles which often

can reach speeds around 50 km/h in hilly terrain. Higher speeds lead to more difficulty maneuvering the bicycle. Both speed and maneuverability can influence the risk of an accident.

- *Speed* is paramount for injury severity in accidents. It is well known that the speed of the motor vehicle that hits bicyclist influences the injury severity (Rosén, 2013). However, there is limited research regarding how the speed of the cyclist themselves influences the injury severity.
- *Weight*: We can speculate that the combined weight of the cycle and the cyclist is likely to influence the physics of the crash mechanism, we can speculate that this will mainly influence collisions between unprotected road users.
- *Other properties* such as braking, stability, center of weight, lights, reflection, and acceleration of e-devices can be expected to matter.

The Infrastructure

The infrastructure design influences the need for interaction as well as under which conditions this interaction is performed (e.g., speed, visibility) and thereby the risk of an accident occurring. The design itself can also create risks, for example, risk of a fall, collisions with rigid object, etc. The safety of the infrastructure is highly dependent on the design (e.g., Elvik and Vaa, 2004). There are several options to improve the safety of bicyclists through design of the infrastructure. Since that the traffic safety effects of different infrastructure measures cannot be adequately discussed in a short chapter, we refer the reader to other more comprehensive literature.

The Accidents

Bicyclist related accidents can be divided into the following five groups:

1. Collisions with motorized vehicles
2. Single accidents
3. Collisions with other bicyclists/mopeds
4. Collisions with pedestrians
5. Other collisions (e.g., animals, open doors)

The most frequent accident type are single accidents. Single accidents are often around 46%–94% of all reported bicycle accidents (Amoros et al., 2012; Boufous et al., 2011; Davidson, 2005; Eilert-Petersson and Schelp, 1997; Konkin et al., 2006; Niska and Eriksson, 2013; Scheiman et al., 2010; Simpson and Mineiro, 1992; Tin Tin et al., 2010, one study, Martínez-Ruis et al., 2013, showed lower percentage, probably due to underreporting). Among children, single accidents are in the range of 67%–94% of all bicycle accidents (Axelsson and Stigson, 2019; Boufous et al., 2011; Eilert-Petersson and Schelp, 1997; Simpson and Mineiro, 1992), the proportion of single accidents was somewhat higher among the youngest compared to older children (Axelsson and Stigson, 2019; Simpson and Mineiro, 1992). Collisions with motor vehicles constitute often between 2% and 45% of all bicycle related accidents (Amoros et al., 2012; Boufous et al., 2011; Davidson, 2005; Eilert-Petersson and Schelp, 1997; Konkin et al., 2006; Niska and Eriksson, 2013; Rosenkranz and Sheridan, 2003; Siman-Tov et al., 2012; Simpson and Mineiro, 1992; Tin Tin et al., 2010). Among children, the similar proportion are 1%–28% (Axelsson and Stigson, 2019; Boufous et al., 2011; Eilert-Petersson and Schelp, 1997; Simpson and Mineiro, 1992), somewhat lower among the youngest groups, but increased to 12%–28% among the older child groups (Axelsson and Stigson, 2019; Simpson and Mineiro, 1992). The data seems to indicate that the majority of the bicycle related accidents are single accidents, as well as the majority of the serious injuries (Niska and Eriksson, 2013; Weijermars et al., 2016). A study by Gårder (1994) of hospitalization in Maine showed that 35% of bicycle crashes leading to hospitalization involved motor vehicles and 65% were single-bicycle accidents.

However, the former required an average hospital stay of 9.2 days and the latter only 4.6 days, so the two categories account for roughly the same number of injury-days. Regarding fatal accidents, we get a clearer picture, collisions with motor vehicles stands for majority of the fatalities, often in the range of 64%–100% of all cyclist fatalities (Bil et al., 2010; Nicaj et al., 2009; Rowe et al., 1995; Rosenkranz and Sheridan, 2003).

There is far less known regarding other accident types. Collisions between two or more cyclists (or cyclists and moped drivers) have been shown to account for 1%–11% of all bicycle accidents (Davidson, 2005; Boufous et al., 2011, 2012; Eilert-Petersson and Schelp, 1997; Niska and Eriksson, 2013; Simpson and Mineiro, 1992), and 4%–8% of child bicycle accidents (Axelsson and Stigson, 2019; Boufous et al., 2011; Simpson and Mineiro, 1992). Studies of collisions between bicyclists and pedestrians indicate that those collisions make up around 1%–2% of bicycle injury accidents (Boufous et al., 2011; MSB, 2013; Nilsson and Åström, 2016; Tin Tin et al., 2010), and majority of those who were injured in those accidents were the pedestrians (Eriksson et al., 2015; Niska and Eriksson, 2013). Another known risk for cyclists is to collide with an open or opening car door. Those accidents were shown to constitute 3%–9% of bicycle accidents (Boufous et al. 2011; Isaksson-Hellman, 2012; Johnson et al., 2013). There are also cases where this accident type results in serious or even fatal injuries (Nicaj et al., 2009). There is limited research regarding the scope of collisions with animals. A survey-based study showed 3% (urban) and 4% (rural) of the self-reported accidents were collisions with animals (Kröyer, 2016a), this study was however based on people cycling for exercise in competitive ways.

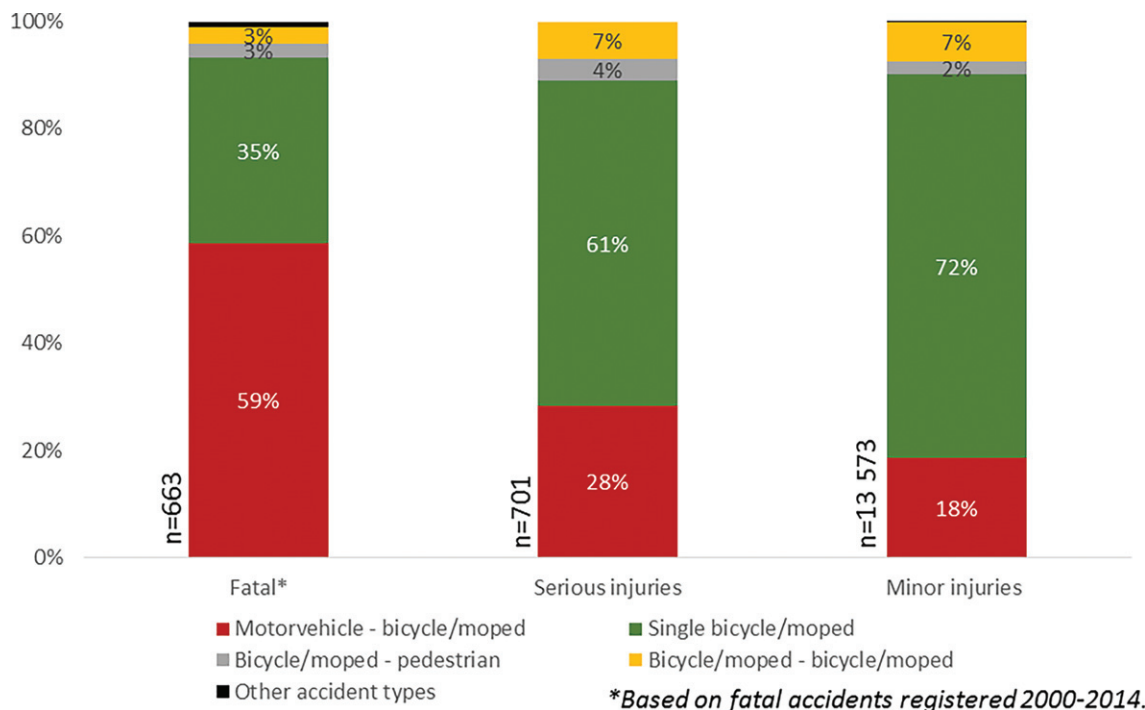


Figure 2 Classification of bicycle-related accidents that were registered in Sweden in 2013 by type and injury severity, fatal accidents are based on accidents registered in 2000–2014. Source: STRADA (2015).

If we take Sweden as an example, then in the year 2013 (injury accidents, police, and hospital registration, STRADA, 2015), 72% of the minor injury bicycle accidents were due to single accidents and a further 7% were due to collisions between bicyclists and/or moped drivers, see Fig. 2. “Only” 18% are due to collisions with motor vehicles. When it comes to the serious injuries, 61% are due to single accidents, and 28% due to collisions with motor vehicles, 7% are due to collisions between bicyclists and/or mopeds, and 4% are due to collisions between bicyclists/mopeds and pedestrians.

Since there were only 27 reported bicycle fatalities in 2013 in Sweden, we use 15 years data, 2000–14 (STRADA, 2015). Of 663 bicycle-related fatalities, at least 16 were pedestrians and 4 were other types of road users (not pedestrians, cyclists, or moped drivers). Also, 59% of the fatalities were due to collisions between a motor vehicle and a cyclist or a moped. 27% and 8% were single accident bicyclists or moped. Around 3% were collisions between bicyclists/moped and pedestrians, were in most cases it was the pedestrian who was killed. Also, 2% and 1% were collisions between two cyclists or a cyclist and a moped driver. Finally, 1% were due to other type of accidents (often related to railway).

Both the literature review and the case sample show that the two first accident types contribute to a great majority of the (recorded) serious and fatal accidents. This means that the main body of the research efforts have been aimed at these accident types. However, not all accidents are reported to authorities. If the accident is not reported it will not be registered in the accident databases, hence it will be missing in the statistics. Bicycle accidents not including a motor vehicle are frequently underrepresented in our accident statistics, and that the more severe the injuries are, the more likely they are to be included into our data (Elvik and Mysen, 1999). We can therefor speculate that there is a large under reporting of the other accident types, and they should also be considered. There are several factors that have been related to bicycle accident and/or the severity of the accidents. Some of those are discussed in the following sub sections.

Accident Location and Time of Year

Bicycle accidents are frequent during months with favorable weather conditions, that is, the summer half of the year especially in countries far away from the equator (Niska and Eriksson, 2013; Rosenkranz and Sheridan, 2003). This is related to that the number of bicycle accidents is strongly related to exposure. One study showed that a considerable part of bicycle accidents occurs during recreation or fitness, while around one in three injury accidents are trips to or from schools (Davidson, 2005). A considerable part of the child bicycle accidents has been shown to relate to trips to and from school (Axelsson and Stigson, 2019), and child bicycle accidents are not limited to residential areas (Abdel-Aty et al., 2007). Many accidents are related to the child using the bike as a toy or for performing stunts, especially among the youngest ones, but also a lack of biking skills, maintenance, or distraction (Axelsson and Stigson, 2019).

Collisions between bicyclists and motorized vehicles usually occur in urban areas (Boufous et al., 2011; Isaksson-Hellman, 2012), but it is there the bicycle exposure is greatest. Some studies however show rural accidents to be more serious

(Boufous et al., 2012; Kaplan and Prato, 2013). Many studies have shown that a majority of bicycle-motor vehicle collisions occur at intersections, often in the range of 51%–90% (Boufous et al., 2011, 2012; Isaksson-Hellman, 2012; Kaplan et al., 2014; Kim et al., 2007; Wei and Lovegrove, 2013). On the other hand, fatalities are frequently reported to have occurred away from intersections, but the proportion of all fatal bicycle accidents varies considerably, 11%–80% (Hoque, 1990; Kim et al., 2007; Nica et al., 2009).

Individual Factors, Age, and Gender

One study showed a higher injury rate for females (Blaziot et al., 2013). Kaplan and Prato (2013) showed that the gender proportions were equal in all injury accidents from collisions with motor vehicles, while other studies (Boufous et al., 2012; Kim et al., 2007; Siman-Tov et al., 2012) showed a majority of those in accidents to be males. The results are more in agreement when it comes to fatal accidents, then males are in a clear majority, often 78%–93% of those killed (Bíl et al., 2010; Kim et al., 2007; Nica et al., 2009; Rodgers, 1995; Rowe et al., 1995), and females have been shown to have a lower risk of fatal injuries (Bíl et al., 2010; Rodgers, 1995). Several studies have shown that young boys are highly overrepresented in bicycle accidents (Axelsson and Stigson, 2019; Boufous et al., 2011; Haileyesus et al., 2007; RNSA, 2014; Siman-Tov et al., 2012).

Sze et al. (2011) showed that children below 14 had elevated risk of severe and life-threatening injuries, however, the difference was not statistically significant due to large confidence intervals. Kaplan et al. (2014) showed children below the age of 10 were more likely to suffer light, severe or fatal injuries in collisions with motor vehicles compared to other children and young adults. There are, however, studies who could not find this difference (Boufous et al., 2012; Rodgers, 1995). As was discussed in the previous section, children have cognitive and physical preconditions that can work in their disfavor. We can also speculate that their travel pattern in the road network is not comparative to those of adults, where several other factors will reduce or amplify possible difference in injury severity.

Accident statistics show seniors to have a highly elevated risk of serious and fatal injuries in collisions between bicyclists and motor vehicles (Bíl et al., 2010; Kaplan et al., 2014; Kim et al., 2007; Rodgers, 1995). This elevated risk seems to start somewhere in the upper middle age years, even as low as from the age group 40–50 (Kaplan et al., 2014) or 55+ (Kim et al., 2007), however, the risk increases dramatically for the older groups (Eluru et al., 2008; Kaplan et al., 2014; Kröyer, 2015a; Sze et al., 2011; Yan et al., 2011). Seniors are also a high proportion of those who are killed in the traffic as cyclists; it varies though considerable between countries (Hagenzieker, 1996).

Vehicle Type

A great majority of vehicle-bicycle collisions are with passenger vehicles (Kaplan et al., 2014; Kim et al., 2007; Simpson and Mineiro, 1992). However, several studies have shown that the risk of serious or fatal injuries is higher if the vehicle is larger, that is, trucks, other heavy vehicle, busses, pickups and vans (Ackery et al. 2012; Eluru et al., 2008; Kaplan et al., 2014; Kim et al., 2007; Moore et al., 2011; Nica et al., 2009; Otte et al., 2012; Yan et al., 2011). Right turns by large vehicles are a serious threat for cyclists, that is, when the cyclists and the large vehicle are coming from the same direction, however, the cyclist is going straight through, while the large vehicle turns right. This is a dangerous interaction due to among other factors the blind spot of trucks, and those accidents tend to be serious when they occur (Pokorny et al., 2017; Richter and Sachs, 2017). In London, freight vehicles were involved in four of ten of all fatal bicycle accidents, and in half of those accidents this was the case (since in London you drive on the left side of the road, it were left turn accidents) (Morgan et al., 2012). The interaction between larger vehicles and cyclists should influence intersection design in particular.

Speed

Speed is a central part of traffic safety. This is due to the simple fact that (a) speed relates to time and distance travelled, hence the time the road users have to react to prevent or to mitigate an accident occurring, and (b), the speed will eventually control the forces that act on the cyclist, hence the injury severity. Collision/impact speed is strongly related to the risk of serious and fatal injuries, where the risk of injuries increases fast with higher speed (Kröyer et al., 2014; Rosén, 2013). Studies have also seen a correlation between both speed limit (Boufous et al., 2012; Kaplan et al., 2014), estimated travel speed (Kim et al., 2007), speeding (Bíl et al., 2010; Kim et al., 2007), mean travel speed (Kröyer, 2015a) and injury severity for bicyclists hit by motor vehicles. The impact speed has also been seen to influence where the head impacts the car in collisions with motor vehicles (Otte et al., 2012). To adjust speed in urban areas is therefore a central measure for bicyclist safety.

Alcohol and Drugs

It is well known that being under the influence of alcohol as the driver of a motor vehicle is a strong risk indicator, and that the risk increases quickly with higher consumption (Forsman, 2013). Several studies show driver intoxication to increase the risk of serious or fatal injuries in collisions between bicyclists and motor vehicles (Kim et al., 2007; Moore et al., 2011). One study used alcohol use for the party at fault instead (that is, either the car driver or cyclist was under influence), that did not result statistically significant difference (Bíl et al., 2010), but the definition might explain the contradicting result.

Part of the cyclists involved in an accidents were under the influence of alcohol (Eilert-Petersson and Schelp, 1997; Kaplan et al., 2014; Kim et al., 2007), however, the proportion was more frequent, 7%–21%, among the fatally injured cyclists (Kim et al., 2007; Nicaj et al., 2009; Rowe et al., 1995). Alcohol involvement was especially frequent in accidents during the evenings and nights (Davidson, 2005; Rosenkranz and Sheridan, 2003). Those who were under the influence of alcohol and/or drugs were more likely to suffer serious (Kaplan et al., 2014; Moore et al., 2011) or fatal injuries (Kaplan et al., 2014; Kim et al., 2007).

Visibility and Road Condition

In the case of a dangerous situation, the road user must understand the situation at hand and notice all critical elements of the situation to be able to decide on measures. To be able to perform this, visibility is important. In intersection collisions with motor vehicles, many cyclists (and probably car drivers) misunderstand the traffic situation and/or make an inadequate plan of actions. Habibovic and Davidsson (2012) showed that the cyclist not seeing the conflicting car due to visual obstruction was a frequent problem.

But it is not only visual obstructions that limit the visibility. Several studies have tried to estimate if the time of day and streetlights influence the number or severity of bicycle accidents. Despite fewer cyclists, a considerable part of fatal bicycle accidents occurs after dark or at dusk or dawn (Rodgers, 1995; Rowe et al., 1995). The risk of serious or fatal injuries have been shown to be higher during night than during daylight conditions if there are no streetlights (Boufous et al., 2012; Kim et al., 2007; Wang and Stamatiadis, 2011; Yan et al., 2011), and Boufous et al. (2012) showed elevated risk during nights with streetlights on. Bîl et al. (2010) showed fatal injuries to be associated with (1) nighttime without streetlights and (2) daylight conditions where the visibility was bad, compared to daylight conditions with good visibility. The study did not show statistically significant elevated risk for nighttime with streetlights. An older study showed that rear-end collisions (where the bicycle was hit from behind) were 80% of the fatal accidents during nighttime, compared to 10% of those during daytime. This study also showed that fatal accidents were frequent when the motor vehicle was overtaking the cyclist and the cyclist swerved out (Hoque, 1990). Those studies further emphasize the importance of good visibility so that the different road users can be aware of each other.

The road condition is also important for traffic safety. First, inclement weather can influence visibility and there is a relation between inclement weather and injury severity (Kim et al., 2007; Wang and Stamatiadis, 2011). Single bicycle accidents are frequently contributed to for example, winter conditions, ice/slipperiness, snow, pavement edge/curbstone, gravel, rigid objects, mechanic failures, speed, distraction and interaction with other road users (Eilert-Petersson and Schelp, 1997; Niska and Eriksson, 2013). Studies have also shown a relation between slippery surfaces and the risk of fatal injuries (Kaplan et al., 2014) and icy roads and injury severity in bicycle-car collisions (Wang and Stamatiadis, 2011). The importance of road condition should not be underestimated, given the fact that single-bicycle accidents are a considerable part of the serious injuries.

Injuries Sustained by Bicyclists

Traffic safety work aims to prevent injuries and the resulting health loss and fatalities. The types of injuries sustained in bicycle accidents is therefore important. Most of the studies that investigate the injuries of bicyclists have shown the following three regions to be the most frequently injured ones.

- Lower extremities,
- Upper extremities,
- Head/neck injuries.

Studies have shown this for: (1) all injury severities except fatalities (Axelsson and Stigson, 2019; Eilert-Petersson and Schelp, 1997; Juhra et al., 2012; Konkin et al., 2006; Maki et al., 2003; Niska and Eriksson, 2013; Otte and Haasper, 2005; Siman-Tov et al., 2012). (2) Single accidents (Amoros et al., 2011; Niska and Eriksson, 2013). (3) Collisions between motor vehicles and bicyclists (Amoros et al., 2011; Haileyesus et al., 2007; Otte et al., 2012). (4) Children (Axelsson and Stigson, 2019; Amoros et al., 2011; Boufous et al., 2011; Olofsson et al., 2012; Siman-Tov et al., 2012). It varies though between studies and injury severity which of those regions are the most frequently injured (there are studies in this list that have shown another body region among the top three).

Head injuries were registered in between 72% and 77% of the fatal injuries (Maki et al., 2003; Nicaj et al., 2009; Rowe et al., 1995), often followed by (but not necessarily in this order) chest, abdomen, torso or neck injuries (Maki et al., 2003; Nicaj et al., 2009). Here, we might add that in collisions with motor vehicles, the cyclists suffers head injuries both from impact with the car and the ground (Badea-Romero and Lenard, 2013), and that the prevalence of brain injuries, soft part injuries and serious head injuries for both children and adults are associated with higher collision speed (Otte, 1989).

Safety Equipment

An effective safety equipment is a good way to reduce the health loss of accidents. They can either be aimed at preventing the accident or to mitigate the consequences. As previously discussed, visibility is important, and there is an association between accident risk and light defects (Martínez-Ruis et al., 2013) and an older study found a high overrepresentation in night-time fatal rear-end accidents where the car driver did not observe the cyclist, where the cyclist was without rear lights (Atkinson and Hurst, 1983). Another study found an association between reflective vest and lower risk of multiparty injury accidents for cyclists (Lahrman et al., 2018). Brake

defects have also been shown to contribute to accidents (Marinez-Ruiz et al., 2013; Scheiman et al., 2010), and controlled tests have shown studded bicycle tires to provide better braking properties on ice (Hjort, 2018), something that might have positive safety effects given the frequent cause of ice/snow as cause for single accidents (in regions where that applies).

Given the importance of head injuries shown above, it should not come as a surprise that the most well known safety equipment of cyclists is the bicycle helmet. Two meta-analysis studies have recently estimated the effect of bicycle helmet. Both studies (Elvik, 2011, 2013; Høye, 2018) showed that the risk of fatal injuries, as well as head injuries and brain injuries was reduced considerably by wearing a helmet. The positive effect is both for adults and children, and in single accidents as well as in collisions with motor vehicles (Høye, 2018). Though general use of bicycle helmet will help us greatly, it cannot solve all serious injuries, since it is unlikely to have any considerable positive effect on serious injuries of other body regions.

Though it is outside the scope of this chapter, the severity of collisions with motor vehicles can also be mitigated by reducing speed, adaptation of active safety equipment, for example, automatic brakes, and measures that make the vehicle more forgiving for the hit bicyclist (inclusive lower share of motor vehicles that are related to more serious injuries).

The Special Case of E-Bikes and Micromobility

Micromobility is used to describe light travel devices that are either man powered or with electric support, where the electricity usually will not support the vehicle at higher speeds than 45 km/hr⁻¹ (ITF, 2020). This includes bicycles, kick-bikes, skates, skateboards, and several other devices. Many of those are new transport modes, and the modal share is currently low. Danish studies show that micromobility devices (exclusive bikes/e-bikes) constituted around 1.5% of the road users on bicycle roads, and 97% of those were e-kick-bikes (Færdselsstyrelsen, 2020). Increased use of those devices can however be expected to result in accidents in the future and we need to start to prevent those accidents before they cause more casualties.

E-Bikes

E-bikes are in many ways similar to a regular cycle; however, they have electric support up to for example 25, 32, or 45 km/hr⁻¹. The speed of e-cyclists has been measured to be somewhat higher compared to a regular bicyclist (Cherry and MacArthur, 2019; Steintjes, 2016), which might increase injury severity in accidents.

Accident statistics show similar patterns compared to regular bicycles. The majority of those injured are males (Papoutsis et al., 2014), females have elevated odds ratio for moderate/serious injuries (Hertach et al., 2018), and head/neck injuries are most frequently followed by injuries to upper extremities, face, chest and abdomen (Papoutsis et al., 2014). Schepers et al. (2014a) showed that when controlled for gender, age and bicycle use, there was an association between electric bicycle use and increased risk of being treated at an emergency department, while among those, the users of electric bicycle were not more likely to be further admitted to the hospital compared to regular bicyclists. Fyhri et al. (2019) also shows elevated risk for e-cyclists, however, similar risk of serious injuries as for regular bikes. Another study showed that some e-cyclists contributed their accidents at least partly to the properties of the e-bike (Hertach et al., 2018). On the other hand, Schepers et al. (2018) showed no difference in risk of injury accidents between e-bikes and regular bikes when controlled for annual kilometers travelled, that is, that study suggest that the reason for difference between e-bikes and regular bikes might be due to travelled distance. Weber et al. (2014) showed diverging results.

The most common e-bike accident type is single accidents (Papoutsis et al., 2014; Schepers et al., 2018; Weber et al., 2014). Dozza et al. (2016) by using naturalistic data showed the most frequent conflict to be with pedestrians, then light vehicles. Krøyer (2019a) studies self-reported conflicts, where single accidents were the most frequent type. Several self-reported accidents/conflicts were attributed to speed, acceleration, road damages or poor road condition, visibility, balance (among others when going on or off the bike), other road user underestimating the speed, crossing curbs, ice/slippery surface, skidding, brake problems, alcohol involvement, interaction, and the weight of the bicycle (Haustein and Møller, 2016; Hertach et al., 2018; Hiselius et al., 2013; Johnson and Rose, 2015; Krøyer, 2019a; Papoutsis et al., 2014). Those factors however also cause accidents among regular cyclists.

There is considerably less research available regarding the safety of e-bikes and e-cyclists and currently, it is hard to draw firm general conclusions if e-bikes are from traffic safety similar, better or worse compared to regular bicycles; more data and research is necessary. The current literature does however not indicate that there is a great difference in the accident statistics, this might though be biased by confounding factors.

Other Micromobility Devices

To evaluate the safety of a given travel mode requires extensive accident data due to the complexity of traffic safety. For this reason, there is much less known about those less frequent travel modes. Konkin et al. (2006) showed that against every 100 cyclists, there were around 10 skateboard user and eight inline skate users reported to an emergency room. As with cyclists, males were in great majority, and 10% respectively 15% were involved in collisions with motor vehicles. Given the recent popularity of e-micromobility devices, we can speculate that this problem has increased and will increase even further as those devices become more popular. A considerable part of the users and those injured are young people and males (APH, 2019; Færdselsstyrelsen, 2020; Stigson and Klingegård, 2020). Those injured in accidents related to e-scooters are mostly the users themselves, but 8%–13% were pedestrians or other road users (Trivedi et al., 2019; Stigson and Klingegård, 2020). Of those, half were hit by an e-scooter, while the rest were cases of fall accidents due to tripping over a parked scooter or lifting/moving a e-scooter not in use (Trivedi et al., 2019). This is in line with ITF (2020) which reported that pedestrians were “only” one in ten of the fatalities.

Single accidents seem to be the most frequent accident type, while collisions with motor vehicle account for 2%–28% of the accidents (APH, 2019; Blomberg et al., 2019; PBOT, 2019; Liew et al., 2020; Stigson and Klingegård, 2020; Trivedi et al., 2019). Majority of the fatalities involved another motor vehicle (ITF, 2020). There have been some estimates of the risk of using e-kick-bikes compared to that of other transport modes (e.g., Færdelestyrelsen, 2020), however, at this point, estimates of accident rate should be taken with some reservations due to data limitations and underreporting.

Head injuries are a considerable problem, but other body regions also suffer injuries (APH, 2019; Kobayashi et al., 2019; Stigson and Klingegård, 2020; Trivedi et al., 2019), and most of the studies report a very low helmet use (APH, 2019; Blomberg et al., 2019; Færdelestyrelsen, 2020; Kobayashi et al., 2019; Liew et al., 2020; PBOT, 2019; Stigson and Klingegård, 2020; Trivedi et al., 2019, one Australian study showed a great majority using a helmet, Haworth and Schramm, 2019), and one study showed the helmet use to be much lower for rental e-kick-bikes compared to self-owned e-kick-bikes (Færdelestyrelsen, 2020). Even though preliminary, this is somewhat concerning results, given the fact that the electric support allows for higher speeds (Cherry and MacArthur, 2019) and APH (2019) reported that excessive scooter speed contributed to 37% of the accidents. In addition, considerable part of the accidents occurred while the road user was under the influence of alcohol, though to a varying degree (APH, 2019; Blomberg et al., 2019; Kobayashi et al., 2019; Trivedi et al., 2019), and ITF (2020) report that many fatal accidents occur during the night.

Those preliminary data indicate that much of the challenges regarding micromobility are “similar” to those of bicycles and e-bikes. Helmet use is however low, and they are somewhat more sensitive to the infrastructure, for example, curbs, potholes, manholes, slippery or low friction, and even malfunctions and balance (APH, 2019; Stigson and Klingegård, 2020). We can also speculate that the interaction is somewhat more complex since the infrastructure is not designed with regards to this road user group and we have not yet established a general code of behavior using those devices. Also, we do not know how the safety of this group is regarding larger vehicles, especially right turn collisions (in countries where people drive on the right). Finally, this is a new travel mode. The users will learn, gain experience, and become better at using it, simultaneously as other road user groups will also gain experience in interacting with those devices. We would like to stress that this discussion is based on very limited accident statistics and research, and therefore subjected to changes in the coming years. This should therefore be taken with considerable reservations.

System View: Exposure, Risk, and Consequence

The state of traffic safety is often determined by counting the number of casualties due to traffic accidents, for example the number of fatalities, or fatalities per 100,000 inhabitants. But, how can we determine the traffic safety from these numbers? Is it acceptable to have 30 bicycle fatalities in Sweden per year, or is it catastrophic? Is it relatively safe to cycle in Sweden or is it dangerous? Is it safer to cycle than to use a car when fatal car accidents are considerably more numerous? To understand this and other factors related to traffic safety, we must consider factors such as *exposure*, *risk*, and *consequence*.

Exposure

Exposure describes how much travel there is, where there is a risk for an accident occurring. (1) Cycling without any interaction with other road users involves the risk of falling or colliding with a rigid object. We might also add that the risk of falling is not necessarily constant, but rather related to the risks we encounter, whether it is a point that is slippery, a badly placed sign that is in the way, or due to moment of lack of attention. (2) Interaction with other road users includes a risk of being involved in a collision or to fall in the efforts of preventing a collision.

There is a strong relation between exposure and the number of accidents. Figure 3 shows the exposure rate, and the number of bicyclist fatalities per million inhabitants for 13 European countries and the United States. It demonstrates that a region with a high level of cycling is likely to suffer more bicycle fatalities than a region with low level of cycling, everything else being equal. Even though this analysis is simplified, it demonstrates how increased exposure correlates with a higher number of fatalities. A country with high level of cycling is also likely to have a high proportion of bicycle accidents of all road accidents (Schepers et al., 2014b). This also means that a region with few bicycle fatalities does not automatically mean that it is safe to cycle there, it is possible that the explanation is simply that few cycles. Similarly, a country with more than 30 fatalities is not necessarily worse than Sweden, where it is possible that there is greater exposure due to either larger modal share of cyclists or greater population.

Risk

The risk dimension describes the probability of being involved in an accident per unit of exposure, that is, the individual risk of being involved in an accident. Several factors influence the risk, for example, the design of the infrastructure, the composition of the road user, weather conditions, modal share, traffic behavior. If we multiply the exposure and risk, we get the number of accidents that do occur. This implies that we can either reduce the number of accidents by reducing the risk, by moving trips between different travel modes or to shorten the distance travelled.

If we look at the risk of bicyclists, this is a somewhat accident-prone travel mode. Blaizot et al. (2013) showed the injury rate for cyclists per million hours to be eight times that of the injury rate for car occupants, and the rate of serious injuries were even higher. This elevated risk of cyclists has been shown in several other studies of accident statistics (Bjørnskau, 2015; Tin Tin et al., 2010).

Figure 4 is a simplified description of the risk dimension on the national level (focused on fatality rates). It shows comparison of number of bicycle fatalities per inhabitant against the bicycle exposure per inhabitant (there are considerable uncertainties regarding

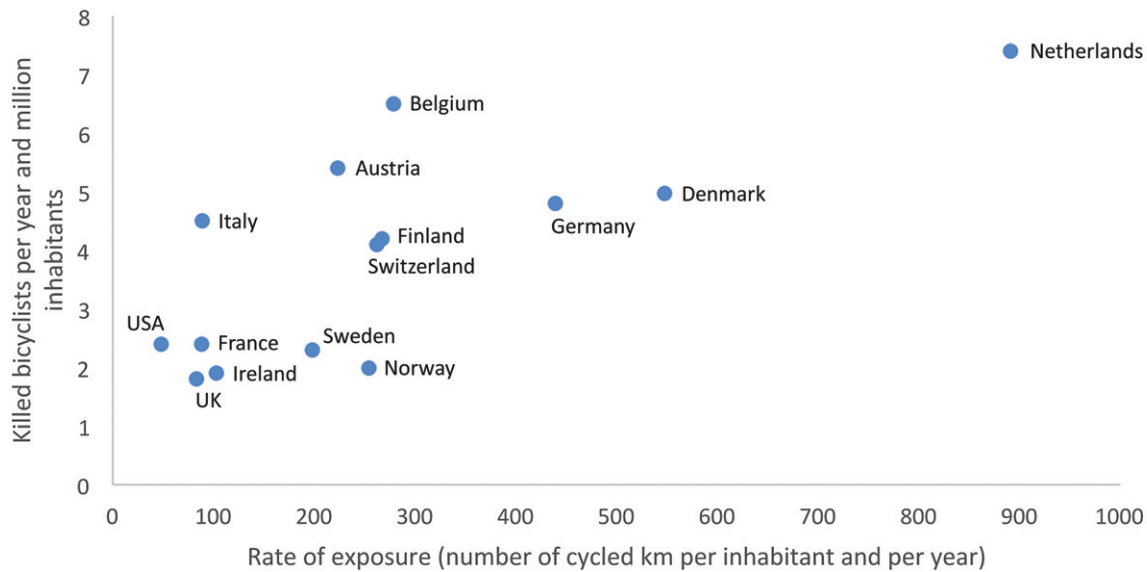


Figure 3 Comparison of bicycle exposure and the rate of killed bicyclists per million inhabitants. Source: Data: Santacreu (2018) and Castro et al. (2018), translated from Kröyer (2019b).

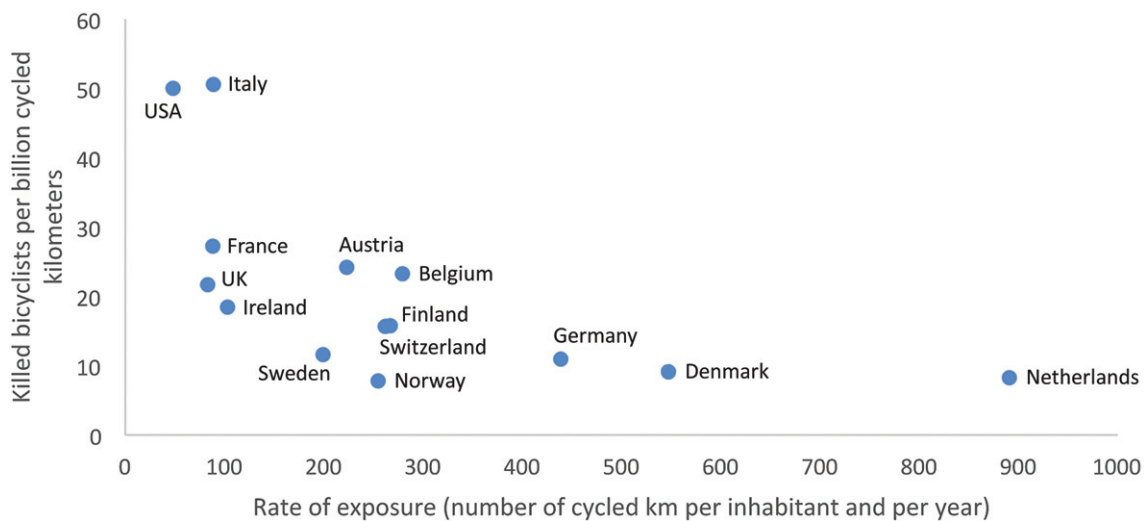


Figure 4 Comparison of the risk of fatal accidents and the rate of exposure for bicyclists. Source: Data: Santacreu (2018) and Castro et al. (2018), translated from Kröyer (2019b).

exposure of cyclists, and this way of presenting data can create a visual mathematical artifact (Brindle, 1994). This should therefore be taken with some reservations). We see that the number of fatalities per unit of exposure is considerably lower in the Netherlands, Denmark, and Germany compared to for example the USA. According to those numbers to cycle a given distance in the Netherlands is safer than cycling that same distance in the United States, given everything else being equal. It is also interesting that some of the countries that appeared to have relatively bad results when we simply focused on the number of fatalities (Netherlands, Denmark, and Germany), show relatively good results. We also see an apparent relation where higher exposure is related to lower risk. The fact is that the risk is not independent of the exposure (Elvik and Goel, 2019). This is often referred to as *safety in numbers*.

As those two simplified graphs, Figs. 3 and 4, demonstrates, the raw number of cyclist casualties misses important aspects of the situation. That there are 30 fatalities in Sweden per year is a fact, however, the raw number cannot tell us the general state of safety for cyclist, for that we need to consider both exposure and risk.

Safety in Numbers

Studies have shown that the accident rate at locations where there are many road users of a given type is usually lower than the accident rate at locations where there are few road users (Elvik and Goel, 2019), this is often referred to as *safety in numbers*. This

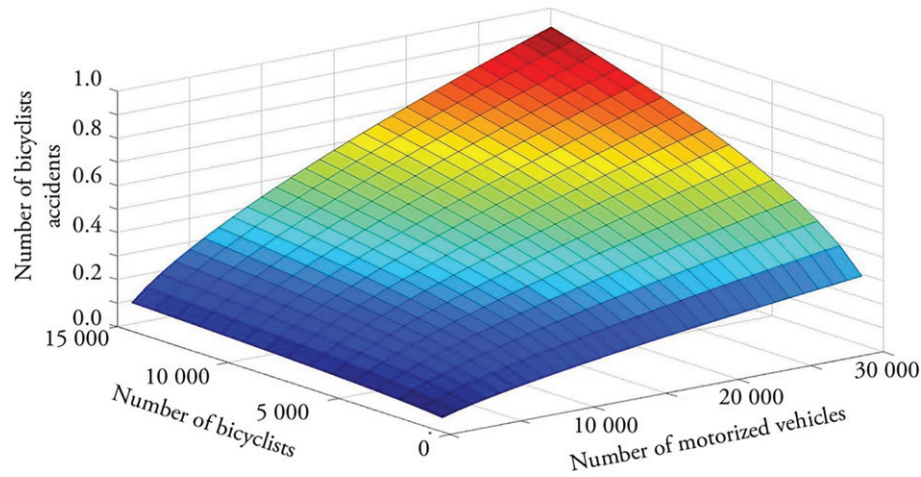


Figure 5 Number of collisions between motor vehicles and bicyclists at intersections based on number of road users. Source: Figure from Kröyer (2015b).

relation is often represented by the mathematical model shown in equation 1. N is the number of accidents, β_i are constants and E_i is the exposure of different road users.

$$N = e^{\beta_0} \prod E_i^{\beta_i} \quad (1)$$

If the constants β_i are equal to 1, that would mean that the accident rate per unit of exposure is constant, that is, simplified, a location with 10% higher exposure has 10% higher number of accidents. However, numerous studies have shown the constant (β_i) to be considerably lower than one, both for the volume of bicyclists and motor vehicles (e.g., Elvik and Goel, 2019; Kröyer, 2019b), usually in the range of 0.2–0.9 (Kröyer, 2019b). This applies to single bicycle accidents (Schepers, 2014) and collisions between bicyclists and motor vehicles (Kröyer, 2019b). There are two points we would like to stress: (1) despite that the risk per cyclists apparently is lower where the exposure is higher, the models suggest that the number of accidents is higher if the exposure is higher. (2) The exposure of motor vehicles influences the number of collisions. More bicyclists result in more accidents, and more cars results in more accidents, see Fig. 5. This means that by reducing the exposure of motor traffic we can reduce the number of collisions.

This can, with some simplification be interpreted that where there is greater volume of cyclists, each cyclist is safer. There are several theories as to why this relation exists. The most common ones are that this is due to:

1. Behavioral adaptation: more cyclists results in that those on motor vehicles are more aware of cyclists and hence adjust their behavior, resulting in fewer accidents (Brüde and Larsson, 1993).
2. More cyclists will correlate with a better infrastructure and/or maintenance for bicyclists, that is, societies that have many cyclists are perhaps more likely to have a more developed bicycle infrastructure.
3. The individual and collective experience of the road users. More cyclists will result in increased general knowledge in cycling and higher skills (Elvik, 2015; Johnson et al., 2014).
4. Numbers in safety, that is, that cyclists to greater degree choose their route from the safety of the route, or that this is partly a statistical artifact without any real relation (Bhatia and Wier, 2011; Brindle, 1994).

There is limited research regarding what is the “real” cause of the safety in numbers effect. If it is a causal effect or simply a correlation. Should the explanation be 1 or 3, then we can expect that the measure of increasing the number of cyclists would have positive effect on the safety of each cyclist. Should the explanation be 2 or 4, then the reason has no direct causal relation to the number of cyclists. Even though we cannot determine the contribution of each of those factors, there seems to be consensus that the effect, that the risk is lower if there are more cyclists, is at least partly true. We must however stress, that even if the effect would be fully causal, the model still indicate that more cyclists will increase the number of bicycle accidents. Also, there is no data available to determine if this will also apply to new micromobility devices. Practitioners should not rely on this effect “solving” the bicycle safety problem for them, but aim to improve the safety. For further reading, we refer to Bhatia and Wier (2011) or Kröyer (2015b).

Consequence

The third dimension, consequence, is how likely a road user is to be seriously or fatally injured if involved in an accident. This means that if we multiply exposure, risk (which gives the number of accidents) with the consequences we get the number of casualties. If we by some measure reduce the risk of fatality if in an accident by half, that would give the same effect on the number of fatalities as reducing the risk of all accidents occurring by a half, everything else being equal.

The consequence (the risk of injuries) varies with accident type, the crash mechanics, properties of the individual and the vehicle, the use of safety equipment, and the speed to mention a few. This dimension is “easily” influenced by those factors. Bicyclists are vulnerable road users, they can maintain considerable high speed, they interact frequently with motor traffic and have limited safety equipment compared to the occupants of cars. In the accident section, several factors that influence the consequence dimension were discussed.

Practical Implications

Cyclists are a complex group regarding traffic safety. Cyclists have relatively high risk of accidents compared to others and are among the most fragile road user groups. They are mixed with both car traffic and pedestrians, can maintain considerable speed, while they have limited safety equipment. To add to the problem, this is a highly heterogeneous group and there is considerable difference among the vehicles (cycles).

Accident statistics show that the two most common bicycle related accident types are single accidents and collisions with motor vehicles. Those two accident types also stand for the majority of reported serious and fatal injuries and thereby results in great health loss for the society. It is therefore important to both prevent those accidents and to mitigate the consequences of them. Collisions with motor vehicles are usually the greatest cause for fatalities, why our measures to improve bicycle safety must include influencing the motor traffic to reduce the threat they cause for cyclists, including reducing the impact speed in collisions and make them more bicycle “friendly”. We like to stress that even though other bicycle related accidents are not as frequent, they are also important. Also, there is usually considerable underreporting of bicycle accidents not involving motor vehicles, and in some countries they are not reported at all, hence, our accident statistics give us a biased picture of the situation, where we might underestimate the importance of those other bicycle related accidents. The accidents are also highly related to the exposure, including which groups cycles and where they cycle, why it is important practitioners study the local accident statistics.

System Perspective and the Individual Safety

There is a relation between exposure, risk, and the number of accidents. According to accident models, more cyclists and/or more cars combined and separately is likely to increase the number of collisions with bicyclists, and vice versa (Jonsson, 2005; Kröyer, 2016b). However, the total traffic safety effect is dependent on where those new cyclists come from. If the increase is due to that those cyclists are new road users (e.g., population is increasing), then this is an addition to the accident problem. However, if those new cyclists previously used motor vehicles, then the change would on one hand increase the risk of bicycle accident due to more cyclists, while reducing the risk of collisions with motor vehicles by reducing the number of motor vehicles (as well as fewer motor vehicles will reduce other accidents related to motor vehicles). Since collisions with motor vehicles are a great threat for bicyclists, then it is positive to have fewer cars.

According to the phenomenon safety in numbers, there are relatively fewer bicycle accidents were many cycles, that is, the risk of each cyclist is lower if there are many cyclists. The models also suggest that more cyclists will increase the number of bicycle accidents. However, theoretical calculations have demonstrated that transferring trips from motor vehicles towards walking or cycling may under certain conditions lead to fewer accidents (Elvik, 2009). A more thorough analysis from Schepers and Heinen (2013) demonstrated by estimation that in the Netherlands to increase the exposure of cycling by moving short trips from cars to cycling is likely to result in unchanged number of fatal accidents, however, increased number of serious injury accidents. This was due to reduced exposure of motor traffic, safety in numbers effect, and that cyclists can in urban settings take shorter and more direct routes than the motor vehicles, that is, to counteract the elevated risk per traveled kilometer, the number of kilometers (exposure) is reduced. It should also be added that to increase the number of cyclists does not mean that the number of fatalities will rise no matter what. We can influence that and prevent this increase. In the Netherlands, the cycling has increased by 40% since 1975, while yearly number of fatalities has been reduced (Schepers et al., 2017), similar results can be seen from Norway between 1983 and 2005 (Erke and Elvik, 2007).

There is also the question of individual perspective. Again, more cycling correlates with lower risk for each individual cyclist, all other things being equal. If we, for the sake of the argument, assume that the safety in numbers effect is in fact causal, then to reduce the exposure of bicyclists would increase the risk for each cyclist. For each cyclist it is not positive to reduce the number of bicycle accidents if it means that this individual will have greater risk than before of being involved in an accident. In that case, we would perhaps get an improved system safety with fewer bicycle related accidents, while simultaneously worsen the individual safety. If this is so, then one of the groups who would get worse individual safety are children, which will continue to cycle and rely on cycling for independent mobility.

Relation to the Use of Bicycle Helmets

Bicycle helmets have been shown to have considerable effect on the risk of serious head injuries, which is highly important given how frequent head injuries are in serious and fatal bicycle accidents. Just as with car occupants, where injuries from accidents are mitigated by use of seat belts, we can benefit from increased use of bicycle helmets.

There is sometimes concern that to discuss cyclists’ casualties and the real effect of bicycle helmet, or to encourage the use of helmet might be counterproductive. This due to concerns that (1) this would draw focus from improving the infrastructure for

cyclists, (2) it might lead to behavioral adaptations, where the cyclists would behave in more dangerous manners, (3) this might prevent people from cycling and thereby hinder the safety-in-number effect, or (4) lead to that those use motor vehicles instead which causes threat to other vulnerable road users.

Generally, we should give the road users our best available knowledge. Should the cyclist perceive it as unsafe to cycle, we should focus on improving the objective and subjective safety rather than to diminish that concern by over or understating the risks and effects of safety equipment. Also, to encourage road users to use helmets does not come instead of improving the safety through other measures, since the helmet will not save all and has no effect on serious non-head-injuries. A literature study by [Oliver et al. \(2018\)](#) showed that most of the literature does not support the theory of cyclist using helmets showing riskier behavior than those who do not. This study ([Oliver et al., 2018](#)) also looked at the relation between bicycle helmet legislation and bicycle exposure. Results from two studies possibly supported that the helmet law reduced cycling exposure, while 13 studies did not support it. 8 Studies had mixed results. There are methodological challenges studying this relation since several other confounding factors might influence changes in bicycle exposure. However, currently, the literature study does not support voiced concern that helmets might result in extensive reduced cycling exposure.

The Health Loss Paradox

Our traffic safety efforts are to prevent health loss in the society. Given the relative high risk of cycle accidents, lost years, and life quality, we must try to prevent those accidents or mitigate the consequences of them. However, we need to be aware of that there is a considerable positive health effect of cycling, which is considered to be even greater than the loss of health due to cycling accidents ([De Hartog et al., 2010](#)), even the use of e-bikes has been shown to increase the physical activity if it comes instead of using a car or public transportations ([Castro et al., 2019](#)) which is positive for health (we still do not know the health effect of other micromobility devices). From a health perspective, we must be careful not to reduce the health loss due to traffic accidents by perhaps causing an even greater health loss through public health, as well as limit the “safe” mobility of children that often rely on cycling for independent travel. We must therefore be aware that reduced cycling may cause counterproductive health effects for society. This supports the claim that we should aim at improving the objective safety so that we can achieve this increased public health while simultaneously minimizing the negative health effect due to traffic accidents.

References

- Abdel-Aty, M., Chundi, S.S., Lee, C., 2007. Geo-spatial and log-linear analysis of pedestrian and bicyclist crashes involving school-aged children. *J. Saf. Res.* 38, 571–579.
- Ackery, A.D., McLellan, B.A., Redelmeier, D.A., 2012. Bicyclist deaths and striking vehicles in the USA. *Injury Prev.* 2012 (18), 22–26.
- Amoros, E., Chiron, M., Martin, J.L., Thélot, B., Laumon, B., 2012. Bicycle helmet wearing and the risk of head, face and neck injury: a French case-control study based on a road trauma registry. *Injury Prev.* 18, 27–32.
- Amoros, E., Chiron, M., Thélot, B., Lumon, B., 2011. *BMC Public Health* 11, 653.
- Ampofo-Boateng, K., Thomson, J.A., 1991. Children's perception of safety and danger on the road. *Br. J. Psychol.* 82, 487–505.
- APH, 2019. Dockless electric scooter-related injuries study. *Austin Public Health*.
- Axelsson, A., Stigson, H., 2019. Characteristics of bicycle crashes among children and the effect of bicycle helmets. *Traffic Injury Prev.* 20, 2019.
- Atkinson, J.E., Hurst, P.M., 1983. Collisions between cyclists and motorizes in New Zealand. *Accid. Anal. Prev.* 15 (2), 137–151.
- Badea-Romero, A., Lenard, J., 2013. Source of head injury for pedestrians and pedal cyclists: striking vehicle or road. *Accid. Anal. Prev.* 50 (2013), 1140–1150.
- Bhatia, R., Wier, M., 2011. “Safety in numbers” re-examined: can we make valid or practical inferences from available evidence? *Accid. Anal. Prev.* 43, 235–240.
- Bil, M., Bílová, M., Müller, I., 2010. Critical factors in fatal collisions of adult cyclists with automobiles. *Accid. Anal. Prev.* 42 (2010), 1632–1636.
- Bjørnskau, T., 2015. Risiko i veitrafikken 2013/14. TØI rapport 1448/2015, Transportøkonomisk institutt, Stiftelsen Norsk senter for samferdselsforskning.
- Blaizot, S., Papon, F., Haddak, M.M., Amoros, E., 2013. Injury incidence rates of cyclists compared to pedestrians, car occupants and powered two-wheeler riders, using a medical registry and mobility data, Rhone County, France. *Accid. Anal. Prev.* 58 (2013), 35–45.
- Blomberg, S.N.F., Rosenkrantz, O.C.M., Christensen, F.L.H.C., 2019. Injury from electric scooters in Copenhagen: a retrospective cohort study. *BMJ Open* 2019, 9.
- Boufous, S., de Rome, L., Senserrick, T., Ivers, R., 2011. Cycling crashes in children, adolescents, and adults – a comparative analysis. *Traffic Injury Prev.* 12, 244–250.
- Boufous, S., de Rome, L., Senserrick, T., Ivers, R., 2012. Risk factors for severe injury in cyclists involved in traffic crashes in Victoria, Australia. *Accid. Anal. Prev.* 49 (2012), 404–409.
- Brindle, R.E., 1994. Lies, Damned Lies and “Automobile Dependence” – some hyperbolic reflections. *ATRF94*.
- Brüde, U., Larsson, J., 1993. Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit? *Accid. Anal. Prev.* 25 (5), 499–509.
- Castro, A., Gaupp-Berghausen, M., Dons, E., Standaert, A., Laeremans, M., Clark, A., Anaya-Boig, E., Cole-Hunter, T., Avila-Palencia, I., Rojas-Rueda, D., Nieuwenhuijsen, M., Gerike, R., Panis, L.I., de Nazelle, A., Brand, C., Raser, E., Kahlmeier, S., Götschi, T., 2019. Physical activity of electric bicycle users compared to conventional bicycle users and non-cyclists: Insights based on health and transport data from an online survey in seven European cities. *Transport. Res. Interdiscipl. Perspect.* 1, 100017.
- Cherry, C.R., MacArthur, J.H., 2019. E-bike safety. A review of Empirical European and North American Studies. A white paper prepared for PeopleForBikes. Lever, Light Electric Vehicle Education + Research Initiative.
- Connelly, M.L., Conaglen, H.M., Parsonson, B.S., Isler, R.B., 1998. Child pedestrians' crossing gap thresholds. *Accid. Anal. Prev.* 30 (4), 443–453.
- DaCoTa, 2012. Children in road traffic, Deliverable 4.8c, EC FP7 project DaCoTa. European Commission.
- Davidson, J.A., 2005. Epidemiology and outcome of bicycle injuries presenting to an emergency department in the United Kingdom. *Eur. J. Emerg. Med.* 12, 24–29.
- De Hartog, J.J., Boogard, H., Nijland, H., Hoek, G., 2010. Do the health benefits of cycling outweigh the risks? *Environ. Health Perspect.* 118, 1109–1116.
- Delin, O., Rudgren, Å., 2007. Geriatrik, andra upplagan. Studentlitteratur.
- Demetre, J.D., Lee, D.N., Pitcairn, T.K., Grieve, R., Thomson, J.A., Ampofo-Boateng, K., 1992. Errors in young children's decisions about traffic gaps: experiments with roadside simulations. *Br. J. Psychol.* 83, 189–202.
- Dozza, M., Piccinini, G.F.B., Werneke, J., 2016. Using naturalistic data to assess e-cyclist behaviour. *Transport Res. F: Traffic Psychol. Behav.* 41, 217–226, Par B.
- Eilert-Petersson, E., Schelp, L., 1997. An epidemiological study of bicycle-related injuries. *Accid. Anal. Prev.* 29 (3), 363–372.
- Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accid. Anal. Prev.* 40 (2008), 1033–1054.

- Elvik, R., 2009. The non-linearity of risk and the promotion of environmentally sustainable transport. *Accid. Anal. Prev.* 41 (2009), 849–855.
- Elvik, R., 2011. Corrigendum to: "Publication bias and time-trend bias in meta-analysis of bicycle helmet efficacy: A re-analysis of Attewell, Glase and McFadden, 2001". *Accid. Anal. Prev.* 43 (2011), 1245–1251.
- Elvik, R., 2013. Publication bias and time-trend bias in meta-analysis of bicycle helmet efficacy: a re-analysis of Attewell, Glase and McFadden. *Accid. Anal. Prev.* 60 (2013), 245–253, [Accid. Anal. Prev. 43 (2011) 1245–1251].
- Elvik, R., 2015. Some implications of an event-based definition of exposure to the risk of road accidents. *Accid. Anal. Prev.* 76 (2015), 15–24.
- Elvik, R., Goel, R., 2019. Safety-in-numbers: an updated meta-analysis of estimates. *Accid. Anal. Prev.* 129 (2019), 136–147.
- Elvik, R., Mysen, A.B., 1999. Incomplete accident reporting meta-analysis of studies made in 13 countries. *Transport. Res. Rec.* 1665.
- Elvik, R., Vaa, T., 2004. *The Handbook of Road Safety Measures*. Elsevier.
- Eriksson, U., Nilsson, A., Gibrand, M., Ljungberg, C., Witzell, J., Slotte, J., 2015. Trygga och säkra korsningspunkter mellan cyklister och fotgängare Rapport 2015:80, Version v1.2. Trivector Traffic, Trivector.
- Erke, A., Elvik, R., 2007. Making Vision Zero Real: Preventing Pedestrian Accidents and Making Them Less Severe. TØI report 889/2007. Institute of Transport Economics, Norwegian Centre for Transport Research.
- Forsman, Å., 2013. Riskkurva för alkohol. Studie baserad på omkomna personbilsförare i Sverige. VTI notat 25-2013. VTI.
- Fyhri, A., Johansson, O., Bjørnskau, T., 2019. Gender differences in accident risk with e-bikes – survey data from Norway. *Accid. Anal. Prev.* 132, 105–248.
- Færdelestyrelsen, 2020. Evaluering af forsøgsordningerne for små motoriserede køretøjer. Færdelestyrelsen, Denmark
- Gårder, P., 1994. Bicycle Accidents in Maine: An Analysis. Transportation Research Record—the Journal of the Transportation Research Board—No. 1438: Safety and Human Performance—Research Issues on Bicycling, Pedestrians, and Older Drivers, National Research Council, Washington, D.C., pp. 34–41.
- Habibovic, A., Davidsson, J., 2012. Causation mechanisms in car-to-vulnerable road user crashes: implications for active safety systems. *Accid. Anal. Prev.* 49 (2012), 493–500.
- Hagenzieker, M.P., 1996. International Conference, Road safety in Europe, Birmingham, 9–11 September 1996.
- Haileyesus, T., Annes, J.L., Dellinger, A.M., 2007. Cyclists injured while sharing the road with motor vehicles. *Injury Prev.* 2007 (13), 202–206.
- Haworth, N.L., Schramm, A., 2019. Illegal and risky riding of electric scooters in Brisbane. Research letter. *Med. J. Austr.* 211 (9.).
- Haustein, S., Möller, M., 2015. Sikkerhed på elcykel: Trafikantfaktorer og trafiksituationer. Selected Proceedings from the Annual Transport Conference at Aalborg University
- Hertach, P., Uhr, A., Niemann, S., Cavegn, M., 2018. Characteristics of single-vehicle crashes with e-bikes in Switzerland. *Accid. Anal. Prev.* 117, 232–238.
- Hiselius, L.W., Svensson, Å., Bondemark, A., Rye, T., 2013. I Vilken utsträckning kan elcyklar (och elmopeder) ersätta dagens biltrafik? Bulletin 288, Trafik och väg, Institutionen för Teknik och samhälle, Lunds Universitet.
- Hjort, M., 2018. Vinterdäck till cykel. VTI notat 20-2018. VTI.
- Hoque, M., 1990. An analysis of fatal bicycle accidents in Victoria (Australia) with a special reference to nighttime accidents. *Accid. Anal. Prev.* 22 (1), 1–11.
- Høye, A., 2018. Bicycle helmets – to wear or not to wear? A meta-analysis of the effects of bicycle helmets on injuries. *Accid. Anal. Prev.* 117, 85–97.
- Isaksson-Hellman, I., 2012. A study of bicycle and passenger car collisions based on insurance claims data. 56th AAAM Annual Conference, Annals of Advances in Automotive Medicine, October 14–17, 2012.
- ITF, 2020. Safe Micromobility. International Transport Forum.
- Johnson, M., Newstead, S., Oxley, J., Charlton, J., 2013. Cyclists and open vehicle doors: crash characteristics and risk factors. *Saf. Sci.* 59 (2013), 135–140.
- Johnson, M., Rose, G., 2015. Safety implications of e-bikes. RACV Research Report 15/02, Royal Automobile Club of Victoria (RACV) Ltd
- Johnson, M., Oxley, J., Newstead, S., Charlton, J., 2014. Safety in numbers investigating Australian driver behaviour, knowledge and attitudes towards cyclists. *Accid. Anal. Prev.* 70, 148–154.
- Jonsson, T., 2005. Predictive models for accidents on urban links. A focus on vulnerable road users. Bulletin 226, Institute of Technology, Department of Technology and Society, Lund University.
- Juhra, C., Wieskötter, B., Chu, K., Trost, L., Weiss, U., Messerschmidt, M., Malczyk, A., Heckwolf, M., Raschke, M., 2012. Bicycle accidents "Do we only see the tip of the iceberg?" A prospective multi-centre study in a large German city combining medical and police data. *Injury Int. J. Care Injured* 43 (2012), 2026–2034.
- Kaplan, S., Prato, C.G., 2013. Cyclist-motorist crash patterns in Denmark: a latent class clustering approach. *Traffic Injury Prev.* 14 (7), 725–733.
- Kaplan, S., Vavatsoulas, K., Prato, C.G., 2014. Aggravating and mitigating factors associated with cyclist injury severity in Denmark. *J. Saf. Res.* 50, 75–82.
- Kim, J.K., Kim, S., Ulfarsson, G.F., Porrello, L.A., 2007. Bicyclist injury severities in bicycle-motor vehicle accidents. *Accid. Anal. Prev.* 39 (2007), 238–251.
- Kobayashi, L.M., Williams, E., Brown, C.V., Emigh, B.J., Bansal, V., Badiee, J., Checchi, K.D., Castillo, E.M., Doucet, J., 2019. The e-mergin e-pidemic of e-scooters. *Trauma Surg. Acute Care Open* 2019.
- Konkin, D.E., Garraway, N., Hameed, S.M., Bown, D.R., Granger, R., Wheeler, S., Simon, R.K., 2006. Population-based analysis of severe injuries from nonmotorized wheeled vehicles. *Am. J. Surg.* 191 (2003), 615–618.
- Kröyer, H.R.G., 2015a. The relation between speed environment, age and injury outcome for bicyclists struck by a motorized vehicle – a comparison with pedestrians. *Accid. Anal. Prev.* 76, 2015.
- Kröyer, H.R.G., 2015b. Accidents between pedestrians, bicyclists and motorized vehicles. Accident risk and injury severity. Department of Technology and Society, Lund University.
- Kröyer, H., 2016a. Trafiksäkerhetsutmaningar för den cykeltäta staden. Trafiksäkerhet vid cykeltävlingar och vad kan vi lära oss av dessa? Trafkon AB.
- Kröyer, H.R.G., 2016b. Pedestrian and bicyclist flows in accident modelling at intersections. Influence of the length of observational period. *Saf. Sci.* 82, 315–324.
- Kröyer, H., 2019a. Tilraunaverkefni með rafhjól í Reykjavík. Lokaskýrsla fyrir árið 2018. Trafkon AB.
- Kröyer, H., 2019b. Öryggi fjöldans og slys á gangandi og hjólandi vegfarendum. Sambandið milli fjölda vegfarenda og fjölda slysa. Trafkon AB.
- Kröyer, H.R.G., Jonsson, T., Várhelyi, A., 2014. Relative fatality risk curve to describe the effect of change in the impact speed on fatality risk of pedestrians struck by a motor vehicle. *Accid. Anal. Prev.* 62 (2014), 143–152.
- Lahrman, H., Madsen, T.K.O., Olesen, A.V., Madsen, J.C.O., Hels, T., 2018. The effect of a yellow bicycle jacket on cyclist accidents. *Saf. Sci.* 108 (2018), 209–217.
- Liew, Y.K., Wee, C.P.J., Pek, J.H., 2020. New peril on our roads: a retrospective study on electric scooter-related injuries. *Singapore Med. J.* 61 (2), 92–95, 20202.
- Maki, T., Kajzer, J., Mizuno, K., Sekine, Y., 2003. Comparative analysis of vehicle-bicyclist and vehicle-pedestrian accidents in Japan. *Accid. Anal. Prev.* 35 (2003), 927–940.
- Martínez-Ruiz, V., Lardelli-Claret, P., Jiménez-Mejías, E., Amezcua-Prieto, C., Jiménez-Moleón, J.J., Lua del Castilla, J.D., 2013. Risk factors for causing road crashes involving cyclists: an application of a quasi-induced exposure method. *Accid. Anal. Prev.* 51, 228–237.
- Moore, D.N., Schneider, W.H., Savolainen, P.T., Farzaneh, M., 2011. Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. *Accid. Anal. Prev.* 43 (2011), 621–630.
- Morgan, A.S., Dale, H.B., Lee, W.E., Edward P.J., 2010. Deaths of cyclists in London: trends from 1992–2006. *BMC Public Health* 10, 699.
- MSB, 2013. Statistik och analys, Skadade cyklister – en studie av skadeutvecklingen över tid. Myndigheten för samhällsskydd och beredskap.
- Nicaj, L., Stayton, C., Mandel-Ricci, J., McCarthy, P., Grasso, K., Woloch, D., Kerker, B., 2009. Bicyclist Fatalities in New York City: 1996–2005. *Traffic Injury Prev.* 10, 157–161.
- Nilsson, A., Åström, J., 2016. Kollisioner mellan cyklister – en förstudie. Rapport 2016:55, Version 1.0, Trivector Traffic, Trivector.
- Niska, A., Eriksson, J., 2013. Statistik över cyklisters olyckor. Faktaunderlag till gemensam strategi för säker cykling. VTI rapport 801. VTI.
- Oliver, J., Esmailikia, M., Grzebieta, R., 2018. Bicycle helmets: systematic reviews on legislation, effects of legislation on cycling exposure, and risk compensation. University of New South Wales.
- Olofsson, E., Bunketorp, O., Andersson, A.L., 2012. Children at risk of residual physical problems after public road traffic injuries – a 1-year follow-up study. *Injury, Int. J. Care Injured* 43 (2012), 84–90.
- Otte, D., 1989. Injury Mechanism and Crash Kinematic of Cyclists in Accidents – An Analysis of Real Accidents. SAE Technical Paper 892425.

- Otte, D., Haasper, C., 2005. Technical parameters and mechanisms for the injury risk of the knee joint of vulnerable road users impacted by cars in road traffic accidents. IRCOBI Conference, Prague (Czech Republic), September.
- Otte, D., Jänsch, M., Haasper, C., 2012. Injury protection and accident causation parameters for vulnerable road users based on German In-Depth Accident Study GIDAS. *Accid. Anal. Prev.* 44 (2012), 149–153.
- Papoutsis, S., Martinolli, L., Braun, C.T., Exadaktylos, A.K., 2014. E-bike injuries: experience from an urban emergency department – a retrospective study from Switzerland. *Emerg. Med. Int.* 2014 .
- Plumert, J.M., Kearney, J.K., Cremer, J.F., 2004. Children's perception of gap affordances: bicycling across traffic-filled intersections in an immersive virtual environment. *Child Dev.* 75 (4), 1243–1253.
- Pokorny, P., Drescher, J., Pitera, K., Jonsson, T., 2017. Accidents between freight vehicles and bicycles, with a focus on urban areas World Conference on Transport Research – WCTR 2016 Shanghai. 10–15 July 2016. *Transport. Res. Procedia* 25 (2017), 999–1007.
- PBOT, 2019. 2018 E-Scooter Findings Report. PBOT, Portland Bureau of transportation.
- Richter, T., Sachs, J., 2017. Turning accidents between cars and trucks and cyclists driving straight ahead World Conference on Transport Research – WCTR 2016 Shanghai, 10–15 July 2016. *Transport. Res. Procedia* 25 (2017), 1946–1954.
- RNSA, 2014. Hjólréiðaslys á Íslandi. Rannsóknarnefnd samgönguslysa.
- Rodgers, G.B., 1995. Bicyclist deaths and fatality risk patterns. *Accid. Anal. Prev.* 27 (2), 215–223.
- Rodgers, G.B., 1997. Factors associated with the crash risk of adult bicyclists. *J. Saf. Res.* 28 (4), 233–241.
- Rosén, E., 2013. Autonomous Emergency Braking for Vulnerable Road users. IRCOBI Conference 2013.
- Rosenkranz, K.M., Sheridan, R.L., 2003. Trauma to adult bicyclists: a growing problem in the urban environment. *Injury, Int. J. Care Injured* 34 (2003), 825–829.
- Rowe, B.H., Rowe, A.M., Bota, G.W., 1995. Bicyclist and environmental factors associated with fatal bicycle-related trauma in Ontario. *Can. Med. Assoc. J.* 152 (1.), 1995.
- Santacreu, A., 2018. Cycling Safety, International Transport Forum, Paris.
- Scheiman, S., Moghaddas, H.S., Björnstig, U., Bylund, P.O., Saveman, B.I., 2010. Bicycle injury events among older adults in Northern Sweden: a 10-year population based study. *Accid. Anal. Prev.* 42 (2010), 758–763.
- Schepers, P., 2014. Does more cycling also reduce the risk of single-bicycle crashes? *Injury Prev.* 18, 240–245.
- Schepers, P., Agerholm, N., Amoros, E., Benington, R., Bjørnskau, T., Dhondt, S., de Geus, B., Hagemeister, C., Loo, B.P.Y., Niska, A., 2014b. An international review of the frequency of single-bicycle crashes (SBCs) and their relation to bicycle modal share. *Injury Prev.* 1–6, 2014.
- Schepers, J.P., Fishman, E., den Hertog, P., Wolt, K.K., Schwab, A.L., 2014a. The safety of electrically assisted bicycles compared to classic bicycles. *Accid. Anal. Prev.* 73, 174–180.
- Schepers, J.P., Heinen, E., 2013. How does a modal shift from short car trips to cycling affect road safety? *Accid. Anal. Prev.* 50, 1118–1127.
- Schepers, P., Twisk, D., Fishman, E., Fyhri, A., Jensen, A., 2017. The Dutch road to a high level of cyclings safety. *Saf. Sci.* 92 (2017), 264–273.
- Schepers, P., Wolt, K.K., Fishman, E., 2018. The Safety of E-Bikes in The Netherlands. Discussion Paper. International Transport Forum, Paris.
- Siman-Tov, M., Jaffe, D.H., Peleg, K., Israel Trauma Group, 2012. Bicycle injuries: a matter of mechanism and age. *Accid. Anal. Prev.* 44, 135–139.
- Simpson, A.H.R.W., Mineiro, J., 1992. Prevention of bicycle accidents. *Injury* 23 (3), 171–173.
- Steintjes, S.B., 2016. Comparing and analysing the behaviour of fusers of conventional bicycles and speed pedelecs: Naturalistic cycling. SWOV, Institute for Road Safety Research.
- Stigson, H., Klingegård, M., 2020. Kartläggning av olyckor med elsparkcyklar och hur olyckorna kan förhindras. Forskningsrapport, Folksam.
- STRADA, 2015. Accident data, The Swedish Traffic Accident Data acquisition.
- Sze, N.N., Tsui, K.L., Wong, S.C., So, F.L., 2011. Bicycle-related crashes in Hong Kong: is it possible to reduce mortality and severe injury in the metropolitan area? *Hong Kong J. Emerg. Med.* 18 (3), 2011.
- Tin Tin, S., Woodwards, A., Ameratunga, S., 2010. Injuries to pedal cyclists on New Zealand roads 1988–2007. *BMC Public Health* 10, 655.
- Trivedi, T.K., Liu, C., Antonio, A.L.M., Wheaton, N., Kreger, V., Yap, A., Schriger, D., Elmore, J.G., 2019. JAMA Network Open *Emerg. Med.* 2 (1) .
- Wang, C., Stamatiadis, N., 2011. Bicyclist Injury severity in Bicycle-Motor Vehicle Crashes at Unsignalized Intersections in Kentucky. Transportation Research Board 90th Annual Meeting, 2011.
- Weber, T., Scaramuzza, G., Schmitt, K.U., 2014. Evaluation of e-bike accident in Switzerland. *Accid. Anal. Prev.* 73, 47–52.
- Wei, F., Lovegrove, G., 2013. An empirical tool to evaluate the safety of cyclists: community based macro-level collision prediction models using negative binomial regression. *Accid. Anal. Prev.* 61 (2013), 129–137.
- Weijermars, W., Bos, N., Stipdonk, H.L., 2016. Serious road injuries in the Netherlands dissected. *Traffic Injury Prevention* 2016 17 (1), 73–79.
- Yan, X., Ma, M., Huang, H., Abdel-Aty, M., Wu, C., 2011. Motor vehicle-bicycle crashes in Beijing: irregular maneuvers, crash patterns, and injury severity. *Accid. Anal. Prev.* 43 (2011), 1751–1758.

Bicycles: The Safety of Shared Systems Versus Traditional Ownership

Mercedes Castro-Nuño, José I. Castillo-Manzano, Applied Economics & Research Group, Universidad de Sevilla, Seville, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction: The Value of Cycling as a Boost to Urban Mobility	139
BSS as a Public Tool to Encourage Cycling: What is the Relationship With the Private Bicycle?	139
Other Health Impacts of Cycling: Comparison of Safety Outcomes of BSS and Private Bicycles	140
Why do Public BSS Bicycles (Often Ridden by Inexperienced Cyclists) Seem so Much Safer Than Private Bicycles?	141
Does “Safety-in-Numbers” (SIN) Theory Really Work?	142
Conclusions and Policy Recommendations	142
References	143

Introduction: The Value of Cycling as a Boost to Urban Mobility

The large increase in private car motorization rates in developed countries has in many aspects contributed to significantly raising inhabitants' quality of life. But mass motorization has also generated negative externalities, such as energy dependence, traffic congestion, air pollution and noise, accidents, obesity, a sedentary lifestyle, and even social inequality challenge. These side effects underline the need to develop more sustainable and healthier transport modes. In the framework of the global environmental sustainability model currently in force, especially in urban areas in Western countries, the individual and social benefits of nonmotorized transport modes such as cycling and walking (including pedestrianization of large areas) have resulted in a revolution in the field of transportation, changing lifestyles, mobility patterns, and even the physical appearance of cities.

More specifically, the advantages of the bicycle as a synonym for health, energy savings, and efficiency are broadly confirmed by growing worldwide scientific research. In recent years, governments and public health agencies have made remarkable efforts to promote cycling not just as a leisure pursuit or sport but also for transportation as a remedy, due to its social returns: from the point of view of urban transport, the bicycle represents an opportunity to reclaim public space and create a flexible transport choice that enables easy access to denser urban areas such as city centers (Pucher et al., 2010). This mitigates the consequences of car use; from an economic perspective, cycling is often the least expensive mode and the most accessible for citizens as it minimizes public and private mobility costs, not only by fuel savings but also because it is a commuting mode that is exempt from taxation and does not require any insurance or a driver's license and has low vehicle and roadway maintenance costs. From the point of view of public health, cycling helps people lead physically active lives and counteracts major mortality risks such as a sedentary lifestyle leading to obesity. Lastly, in environmental terms, it is the most friendly and energy efficient of all transport modes, and even compared to walking.

A series of structural measures to promote the use of bicycles can be highlighted among the most common traditionally used public initiatives, including its integration into urban planning with the construction of cycle tracks and paths and the creation of bicycle parking areas; the promotion of the combined use or intermodality of the bicycle and public transport modes such as the bus and the train; the adaptation of traffic rules and road safety signs; and specific actions for the creation of facilities for cyclists, such as changing rooms and showers at their destination.

In this context, bicycle-sharing systems (BSS) have sprung up around the world during recent decades as an innovative mode of transportation for short trips. They coexist alongside traditional bicycle ownership (which may be more appropriate for longer distances) (Castillo-Manzano et al., 2016) and other road actors such as cars, motorcycles, goods vehicles, pedestrians, electric scooters, and so on, resulting in a complex urban mobility system. As cycling volumes are seen to be higher where BSS has been introduced, this study addresses the relevant issues of urban BSS user safety risks compared to cyclists using their own private bicycles.

BSS as a Public Tool to Encourage Cycling: What is the Relationship With the Private Bicycle?

Public BSS implementation has grown rapidly in many large cities around the world due to its potential to encourage cycling and so afford the benefits of active transportation and increased accessibility to multiple urban locations. In short, these systems enable users to share the use of a fleet of bicycles (through free loan systems or bicycle rental): a bicycle can be borrowed from a bicycle rack—a so-called “dock”—and returned to another dock in the same system, although some dockless systems also exist (Deighton-Smith, 2018).

Although this concept of the public bicycle has become remarkably popular in recent years, especially in certain countries in Europe, such as France and Spain, for example, it actually originated much earlier in Northern European countries with an already-established cycling culture (particularly in the Netherlands and Denmark) in the wake of Amsterdam's 1965 so-called “White Bike Plan,” which was free to users. Most of the White Bike Plan bicycles were stolen, however. Nonetheless, many years later, in 1995, a

second-generation project emerged in Copenhagen as a new public bicycle program called “Bycyklen,” with shared bicycles and a coin deposit system. The coin, worth roughly US \$1 was regained if the bicycle was placed in a dock, so this was still a free system where the bike could be kept for an unlimited time but legally not be brought outside Copenhagen, be locked to other places than a dock or taken off the street. The bicycles were, at least partially, financed and maintained through advertising plaques on the bicycles. According to the literature, relevant technological improvements were subsequently introduced during the last part of the 20th and the first part of the 21st centuries. As a result, the third generation of shared bicycles involving the use of magnetic smart cards, telecommunication systems, electronically locking racks, and mobile phone access was developed at Portsmouth University (United Kingdom) in 1996. This was followed by “LE Vêlo STAR” in Rennes (France) in 1998, “Bicing” in Barcelona and “Sevici” in Seville (Spain) in 2007, “Cycle Hire” in London (United Kingdom) in 2010, and “Citibike” in New York (United States) in 2014 to mention just a few of the systems. Apart from these automatic systems, the recent fourth generation of shared bicycles has drawn on existing experience to design an intelligent system to increase efficiency and usage. Based on smart bikes, it offers the addition of electric pedal assistance (namely, *e-bikes*) and access from mobile apps (e.g., *AirDonkey*, *Spinlister*, and *CycleSwap* in the Netherlands; *Call a Bike* in Germany; and *SoBi Hamilton* in Canada). In this case, the bicycles are connected to an integrated traffic management system and real-time information is provided by intelligent transportation technology. There are even solar-powered docking stations. In the current era of Big Data, the next challenge is currently being developed in the form of the fifth generation of BSS, with dockless bicycles provided by pioneer companies such as *Ofo* and *Mobike* in the Chinese cities of Beijing and Hangzhou since 2015, and in the Polish cities of Krakow (*Wavelo*, 2016) and Warsaw (*Acro-bike*, 2017).

As a result, BSS is one of the fastest growing modes of urban transport and data show that they are now ubiquitous on all continents. The number of cities with BSS has grown from just a small number in the late 1990s to more than 1000 cities around the world with BSS programs today and a joint fleet of over 4.5 million bicycles. BSS were adopted outside Europe for the first time in 2008 and some interesting websites, such as <http://bikes.oobrien.com/#zoom=3&lon=-60&lat=25>, currently provide a fascinating overview of BSS today, with online maps that show the location of docking stations in the main cities all around the world. For instance, at the current time, the eight leading cities in the top 100 for bicycle fleets are all in China. Next comes New York City in ninth position, followed by London, the highest-ranked European city.

Although the recent literature highlights the many benefits of BSS, some case studies analyzed in Europe and North America commonly show low BSS usage rates due to intrinsic flaws in BSS implementation that may act as deterrents for cycle users in general, thus reducing the efficiency of the public investment (or private investment in several places where there are BSS managed by private operators, for example, office campuses belonging to companies such as Google, Samsung, and Facebook in some US states, and Hong Kong, since 2017). These include, for example, congestion at peak hours, the inappropriate location of stations for intermodality, nonflexible schedules, slow loans and returns, problems with bicycle fleet redistribution from full to empty stations, low quality of service due to the lack of comfort and poor state of repair of the bicycle fleet, noncompetitive rental prices, damage due to vandalism, and so on.

In addition, public agencies and municipalities, mainly in European Mediterranean countries, often use bicycle share schemes to counteract the drawbacks of public means of transport and even as a policy to promote their city and develop local pride. Meanwhile, residents are obliged to subsidize the services by ceding public land or allowing advertising in a context of public budget austerity, which forces spending efficiency to be maximized. It is clear, therefore, that for the model to survive in the medium and long term, other complementary strategies should be promoted such as the transfer of users to bicycle ownership. According to successful experiences in specific cities such as Melbourne, Seattle, Buenos Aires, and London, a public–private partnership framework should be set up to exploit the infrastructure network created by public initiative and actions undertaken to boost the transfer to private bicycles. These would include, for example, the construction of safe public parking at places of origin/destination to allay the fear of theft, along with other facilities in large companies and public institutions (changing rooms, showers, etc.); the improvement of intermodality with other modes of urban transportation and “point-to-point” connections; and economic incentives for the purchase of bicycles that take into account the profile of new cyclists (i.e., cyclists of both sexes, not only young people with different income levels who make daily commutes and do not ride only for leisure, sport, recreation, or pleasure, but also for shopping, to travel to their place of work or study, and are committed to a healthy lifestyle).

One example of such incentives can be found in the Spanish region of Andalusia, where subsidies are provided to cofinance bicycle parking and storage areas in buildings.

To summarize, BSS and private bicycles should not be regarded as competing systems but rather as complementary ones that can be used to maximize the efficiency of public strategies and achieve a real modal shift to a more sustainable and healthy urban transportation system (Castillo-Manzano et al., 2015).

Other Health Impacts of Cycling: Comparison of Safety Outcomes of BSS and Private Bicycles

The scientific literature has consistently shown that the health benefits of cycling far outweigh the risks (Wegman et al., 2012). Nonetheless, for cycling to become accepted by more people, it is important that possible negative health effects derived from traffic accidents and hazards in urban settings are acceptable and the first step toward knowing if that is the case is to quantify the risks. It must also be taken into account that, with or without bicycles, road safety should be analyzed from different angles depending on whether urban or rural roads are being studied. The safety impact of cycling has been widely investigated in different parts of the world and has revealed the key role played by certain factors in collisions, including road geometry design, and cyclists’ decisions

regarding speed, route choice, and compliance with basic traffic safety laws. Nevertheless, a comprehensive risk comparison between BSS users and private bicycle riders is not often explored, although all experts seem to agree that the main causes of accidents are similar for both systems, that is, mainly deficient infrastructure when bicycle-only accidents are considered and risky behavior (distraction, alcohol consumption, excessive speed, etc.) for accidents involving vehicles and bicycles.

Moreover, according to certain studies carried out in the Netherlands, the inclusion of e-bikes in the urban mobility system may generate a feeling of greater safety compared to both shared and private bicycles (Schepers et al., 2014). A reason for this is that normal e-bikes in the Netherlands have a maximum speed of 25 km/h (15.5 mph). That is also the case in Sweden and most other European countries (unless the operator has a driver's license and a registered electric motorcycle or moped with insurance, etc.). If the operator tries to go faster, for example, downhill, the hybrid charger will kick in and maximize the speed to 25 km/h, whereas a traditional bicycle easily can reach speeds of 50 km/h or higher in steep downhill. This is in contrast to other countries such as China where, despite a decrease in the number of traffic fatalities from bicycle accidents, the number of e-bike accidents involving traffic casualties has increased. As of now, there were no upper speed limit for e-bikes in China but the government is planning to introduce legislation to require operators to have motorcycle licenses unless the e-bike is maximized to 30 km/h (18.6 mph). In the United States, e-bikes for unlicensed operators are maximized at 20 mph (32 km/h). Be that as it may, e-bike traffic regulation focuses on safety concerns (mandatory helmet use, maximum speed limit, etc.) and differs from one country to another. However, some BSS systems, for example, in Germany, already have e-bikes as an option to pedal bicycles.

Why do Public BSS Bicycles (Often Ridden by Inexperienced Cyclists) Seem so Much Safer Than Private Bicycles?

As has been stated previously, the recent rise of BSS has had a significant impact on urban transportation in many countries and cities worldwide. From a safety perspective, cycling with shared bicycles could present a priori some inherent characteristics basically related to the bike sharer's profile that perhaps seems to make it a more dangerous transportation mode than private bicycles: inexperienced and casual cyclists (on more than one occasion, tourists who are not sufficiently familiar with the urban environment and traffic rules) who use public BSS for short trips frequently combined with lower helmet usage (Fishman and Schepers, 2016).

Notwithstanding the earlier, relevant evidence has been found in the United States and Canada, for example, that points to lower urban accident and casualty rates for BSS than for privately owned bicycles, which could be supported by the so-called "safety-in-numbers" (SIN) effect. According to this theory, a reduction in the number of urban traffic accidents associated with the addition of BSS bicycles may be explained by a rise in driver and noncyclist (motorcycle riders, pedestrians, etc.) awareness, with these adapting their behavior, paying greater attention, and being more alert to bike share users.

Experts from international institutions such as the European Cycling Federation and the OECD go further and highlight other possible groups of factors that may explain differences in traffic accident and fatality rates between public shared bicycle and private bicycle riders.

1. Bicycle design, location, and infrastructure issues

Logically, a bicycle is a more fragile and risky vehicle than an automobile for traveling in cities but the BSS fleet is generally designed for slower and more stable riding than traditional bicycles, at least if we exclude northern Europe where most bicycles are sturdy and have upright riding styles. So compared to private bicycles in North America and southern Europe and many other places, BSS units are larger and heavier, which may mitigate dangerous or aggressive riding, while wider tires can better withstand pavement defects and powerful reflectors and bright colors improve visibility.

Regarding infrastructure, most BSS stations and racks are located in dense urban cores and in areas of traffic congestion, hence automobile speeds as well as bicycle speeds are lower and so, if an accident should occur, the consequences are, foreseeably, less severe.

2. Cycling experience

Mixed findings can be found for this point. On the one hand, some studies have revealed BSS users to frequently take greater care and ride more safely. They are more defensive in their riding due to their, on average, lower level of cycling experience. In this respect, while age and gender do not appear to be conclusive, it seems that there is a range of sociodemographic factors that indicate that private bicycle cyclists have greater cycling experience: they are people who see the bicycle as part of their healthy lifestyles, who demonstrate a greater stated willingness to purchase a bicycle, who have a higher level of education, who are ideologically more progressive, and who are city residents. BSS users are also on average younger, but typically not children, and children and older riders have higher crash rates than people in their 20s to 50s. For example, close to half of all bicyclists killed in accidents in Sweden are over age 65. And, the bicycle per capita fatality rate for people over 65 in Sweden is around 13 per million people for men and 3.5 for women, whereas the fatality rate for people aged 15–44 is around 1.5 for men and 0.8 for women. For people aged 45–64, the fatality rate for men is around 5.0 and for women around 2.5 per million people in that age group. So, if BSS systems are used mostly by people aged 15–44, we should expect lower fatality rates than for private bicycles.

On the other hand, other researchers suggest that less experienced BSS cyclists with lower levels of knowledge may be involved in a greater number of traffic accidents, which might neutralize their more cautious cycling behavior.

3. Cyclist behavior: speeding and helmet usage

It is well known that helmet usage plays a key role in the prevention of injuries and fatal consequences for riders in traffic accidents involving bicycles. This could support the implementation of helmet laws, despite the controversy that it creates among cyclists and cycling associations, which is reflected in the low rates of helmet use in most countries.

This is fully corroborated for public shared bicycles, with a majority of users reluctant to wear helmets compared to private cyclists. Among causes that could explain this can be highlighted the profile of occasional BSS users who mainly use shared bicycles for short trips; they, therefore, rarely possess their own helmets. Or, even if they own helmets they did not bring them along. The provision and maintenance of public helmets by a BSS does not seem to be the solution either, due to the risk of theft and hygiene issues.

In spite of private bicycle riders usually being more likely to wear a helmet and lower helmet usage rates being found among BSS cyclists, certain observational studies carried out in US cities do not conclude any significant differences between serious accidents concerning public BSS users and private cyclists. Notwithstanding, the promotion of helmet use through information campaigns and road safety education would be advisable, rather than by a mandatory law that could discourage BSS utilization and urban cycling in general (Woodcock et al., 2014).

Another safety issue related to cyclist behavior is cycling speed as, in general terms, greater cycling speeds may be associated with more severe traffic accidents. Empirical evidence for European cities such as Lyon shows that cycling speeds for BSS users are substantially lower than for private bicycle riders, probably due to differences in the design of the bicycles, as stated earlier.

Does “Safety-in-Numbers” (SIN) Theory Really Work?

Many authors agree that SIN is a valid argument for explaining how urban safety improves with increased numbers of cyclists (especially BSS users, who can develop a “network effect” by spreading the cyclist culture) and pedestrians on the roads, as a result of drivers’ and motorcycle riders’ greater awareness and caution shown toward cyclists and pedestrians (Elvik and Bjørnskau, 2017).

This effect might even be strengthened by, for example, public transportation agencies or municipal governments using a shared bicycle program in conjunction with other strategies, such as the provision of bicycle tracks or lanes or other facilities as tools to encourage a modal shift in urban transportation (Marqués and Hernández-Herrador, 2017), as in the case of the Spanish city of Seville (Andalusia) (Brey et al., 2017). As a result, other road users become more tolerant and regard cyclists as just another agent integrated into urban mobility rather than as a threat.

Notwithstanding, empirical data to date do not allow confirmation of the causal effect described by SIN theory, but simply point to a statistical correlation between urban traffic accident rates and the number of BSS users that suggests an opposite relationship, that is, there are large numbers of cyclists because cycling is a safer transportation mode for daily urban commutes, especially in European countries with a history of cycling such as Denmark and the Netherlands, where there is a high density of cycling facilities. In neighboring Sweden, cycling has seen a renaissance in the last few decades, especially in the larger cities. In spite of more cycling than in the 1980s, the number of fatalities involving bicyclists has decreased from around 75 per year in the late 1980s to around 20 per year in recent years. Not just SIN explain this but also lots of new bicycling facilities and separation of bicyclist from motor vehicles. Still, around 70% of Swedish bicycle accident fatalities involve motor vehicles even though only about 10% of bicycle injuries that require medical care involve motor vehicles.

Conclusions and Policy Recommendations

Considering all of the earlier, it is concluded that BSS are linked to lower accident rates than private bicycles, so implementing a shared bicycle scheme could be associated with improved urban road safety. However, there is no simple and obvious explanation for this fact. In addition to the well-known SIN effect, the specific characteristics of public shared bicycles as well as who rides them may contribute to lower accident risk (lower riding speed, greater visibility, physical design, fewer riders above age 65, etc.)

Whatever the case, this statement should take into account the fact that some limitations on analysis may skew the comparison between shared and private bicycles. For example, safety data for BSS usually come from operators (who might be underreporting outcomes), while for personal bicycles they come from police reports. BSS injury crashes are obviously also reportable by the police but the report may not show if the bicycle was privately owned or of BSS type. Furthermore, bicycle trips made using a BSS are generally shorter (among other reasons, to avoid additional fees), which leads to a lower per-trip risk rate compared to private bicycle riding. On the other hand, when local governments implement a BSS in conjunction with other strategies to encourage urban cycling (facilities, specific infrastructure, etc.), not only the safety of BSS users is expected to improve, but that of all people who choose to cycle, whether by shared or private bicycle.

This positive effect could be enhanced by the application of certain recommendations, such as the introduction of segregated bicycle tracks to prevent cyclists moving into car traffic and accidents with pedestrians (such as in Scandinavia and the Netherlands) and elimination of bicycle lanes as used in the United States and some European countries; a more consistent methodology to report accident data that separates BSS and private bicycles; and the incorporation of safety intelligent technologies into both shared and private bicycles.

And all this considering that, nowadays, there is a new urban mobility paradigm that threatens even the most organized cities, in which sustainability and efficiency present challenges in terms of new actors that share the same public space (electric scooters, hoverboards, etc.) and new conflicts that suggest more significant safety concerns requiring new regulation.

References

- Brey, R., Castillo-Manzano, J.I., Castro-Nuño, M., 2017. 'I want to ride my bicycle': delimiting cyclist typologies. *Appl. Econ. Lett.* 24 (8), 549–552.
- Castillo-Manzano, J.I., Castro-Nuño, M., López-Valpuesta, L., 2015. Analyzing the transition from a public bicycle system to bicycle ownership: a complex relationship. *Transp. Res. Part D Transp. Environ.* 38, 15–26.
- Castillo-Manzano, J.I., López-Valpuesta, L., Sánchez-Braza, A., 2016. Going a long way? On your bike! Comparing the distances for which public bicycle sharing system and private bicycles are used. *Appl. Geogr.* 71, 95–105.
- Deighton-Smith, R., 2018. The economics of regulating ride-hailing and dockless bike share. *International Transport Forum Discussion Papers*. OECD Publishing, Paris.
- Elvik, R., Bjørnskau, T., 2017. Safety-in-numbers: a systematic review and meta-analysis of evidence. *Saf. Sci.* 92, 274–282.
- Fishman, E., Schepers, P., 2016. Global bike share: what the data tells us about road safety. *J. Safety Res.* 56, 41–45.
- Marqués, R., Hernández-Herrador, V., 2017. On the effect of networks of cycle-tracks on the risk of cycling. The case of Seville. *Accid. Anal. Prev.* 102, 181–190.
- Pucher, J., Dill, J., Handy, S., 2010. Infrastructure, programs, and policies to increase bicycling: an international review. *Prev. Med.* 50, 106–125.
- Schepers, J.P., Fishman, E., Den Hertog, P., Wolt, K.K., Schwab, A.L., 2014. The safety of electrically assisted bicycles compared to classic bicycles. *Accid. Anal. Prev.* 73, 174–180.
- Wegman, F., Zhang, F., Dijkstra, A., 2012. How to make more cycling good for road safety? *Accid. Anal. Prev.* 44 (1), 19–29.
- Woodcock, J., Tainio, M., Cheshire, J., O'Brien, O., Goodman, A., 2014. Health effects of the London bicycle sharing system: health impact modelling study. *BMJ* 348, g425.

Bridge Safety

Per Erik Garder, University of Maine, Orono, ME, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	144
Structural Failures Leading to Vehicle Occupant Fatalities in the United States	144
Structural Failures Outside the United States	145
Accidents not Related to Structural Failures	145
Analysis of Fatal Bridge Crashes in the United States	146
Results	147
Fatalities on or Under the US Bridges in 2016	147
Analysis of Nonfatal Bridge Crashes in the State of Maine	147
Maine Bridges	148
Discussion	148
Biography	149

Introduction

This article is focused on safety issues for road users traveling on or under bridges. Its emphasis is on motor vehicle, bicycle, and pedestrian traffic but it also touches on safety issues related to rail and seafaring vessels. Most of what is presented is based on the author's own research and has a focus on the United States. However, what is presented in the first couple of sections, describing structural failures leading to vehicle occupant deaths, is taken from the literature, with references presented at the end of the article as further readings.

A bridge safety issue of almost similar magnitude, as traffic safety on bridges is suicides, committed by people jumping off bridges. Suicides will not be further discussed in this article since it is not primarily a transportation issue, and in the case of jumping off bridges do not take other people's lives. However, it can be concluded that suicides seem to be concentrated to fairly few, monumental bridges, and that suicides at those locations can be prevented or even eliminated by installing fencing ([Pelletier, 2007](#)). Also, suicides related to transportation facilities are further discussed by Brendan Ryan in this Encyclopedia.

Another safety issue, which will not be discussed in detail in this article, is sabotage where terrorists blow up or threaten to blow up bridges, or use a bridge as a confined area where pedestrians cannot get away from motor vehicles targeting random pedestrians to affect fear in the general population. An infamous case of this issue occurred in London, England, in 2017 when an attacker drove a car into pedestrians on the sidewalk along the south side of Westminster Bridge, injuring more than 50 people, four of them fatally. Similar events have occurred in, for example, Stockholm, Sweden; Nice, France and New York City, but those incidents were away from bridges. This type of security issue could become more common in a future and bridges may be a good place to inflict maximum effect.

Structural Failures Leading to Vehicle Occupant Fatalities in the United States

When thinking of bridge safety, the general public may limit their thoughts to structural failures of bridges. And, there were over 500 failures of bridge structures in the United States alone in the 12-year period (1989–2000) according to a study by [Wardhana and Hadipriono \(2003\)](#). The age of the failed bridges ranged from 1 to 157 years, with an average age of 53 years. The most frequent causes of bridge failures were, according to this study, attributed to floods and collisions. In most of these cases, no people were killed but they all caused delays to travelers and costly repairs.

The deadliest bridge collapse in the United States in the last 60 years did not involve motor-vehicle traffic. It occurred in 1993, when a passenger train crossing the Big Bayou Canot Bridge in Alabama derailed. The locomotive slammed into a bridge span and initiated the structure's collapse bringing the train into the river, killing 47 people onboard ([Jenkins, 2017](#)).

The second deadliest bridge failure in the United States was the 1967 Silver Bridge collapse between Point Pleasant, W.V. and Gallipolis, Ohio. The aluminum suspension bridge collapsed, plunging 32 vehicles into the Ohio River, killing 46 people.

The third deadliest collapse was the 1989 Cypress Street Viaduct failure in Oakland, California. A 6.9 magnitude earthquake caused a number of structures to collapse—including a portion of the upper tier of Cypress Street Viaduct, a section of Interstate 880. And 42 of the quake's total 67 deaths occurred as a result of the viaduct's collapse.

The fourth deadliest collapse in the United States was the 1980 failure of the Sunshine Skyway over Tampa Bay in Florida. The bridge was struck by a ship during a thunderstorm causing the structure to collapse, sending six cars, a truck, and a Greyhound bus plummeting 45 m into Tampa Bay with a total of 35 people killed.

In more recent time, the I-35 westbound bridge over the Mississippi River in Minneapolis collapsed in August 2007 killing 13 people. The National Transportation Safety Board (NTSB) cited a design flaw as the likely cause of the collapse, noting that a too thin gusset plate ripped along a line of rivets, and high weight on the bridge at the time contributed to the catastrophic failure.

Even more recently, in March 2018, a pedestrian bridge under construction collapsed in Florida killing one construction worker and five people in cars at a traffic signal below.

Another bridge collapse that took six lives occurred in 1964 when a ship collided with the Lake Pontchartrain Causeway in Louisiana. And, there were 10 deaths in 1972 in Georgia also as a result of a ship striking the bridge. Another highly publicized bridge collapse occurred in 1983 when the Mianus Bridge on the Connecticut Turnpike in Greenwich, Connecticut collapsed after one of the pins used in its construction had been sheared, killing 3 people.

Adding up these nine deadly bridge collapses gives a total of 208 fatalities in the 60-year period (1958–2018). In other words, these crashes by themselves caused, on average, 3.5 fatalities per year. If we add all other known deadly bridge collapse deaths to these, another 68 occurred in the period 1959–2018. So, we get a total of 276 deaths, or, on average, 4.6 fatalities per year over the last 60 years caused by structural failures of bridges.

Nonfatal bridge failures can also have major consequences. The collapse of a bridge on Interstate 85 in Atlanta in March 2017 meant that 250,000 commuters who travel the highway daily had to find new routes to work and school.

Besides collapses of bridges, there are other structural failures that can lead to accidents and injuries. Examples of this include when rebar cover is lost or overlays debond. It is difficult to find data on crashes resulting from such events but the number of fatalities is most likely small compared to that of bridge collapses.

Structural Failures Outside the United States

Structural bridge failures obviously happen more or less everywhere in the world. And have happened since before motorized traffic was introduced. For example, a well-known bridge failure occurred in the year 312 when the Milvian Bridge in Rome collapsed during the Battle of the Milvian Bridge. A much more recent bridge failure, and one of the deadliest in recent years, also occurred in Italy. It happened in August 2018 when sections of the Ponte Morandi collapsed. This is a series of cable-stay bridges for the motorway between Genoa and Liguria. A total of 43 people died.

Worldwide, roughly 100 people (road users and train occupants) per year have died in the last 10 years as a result of bridge collapses. It is obviously difficult to get accurate data from all countries in the world, but if 100 deaths per year is a good estimate, then the annual fatality rate is around 0.013 fatalities per million people. This is very similar to the US fatality rate from bridge collapses, which would be around 0.014 fatalities per million people if we use the figure of 4.6 fatalities per year and current population. These rates can be compared to the United States and worldwide fatality rates for “traffic.” In 2017, the United States had just over 37,000 highway fatalities giving an annual fatality rate of around 115 per million people, and if we use the last 60 years, as above for the US bridge collapses, for “highway” fatalities, we get a rate of around 170 per million people. That means that there were more than 11,000 times as many people killed in traffic accidents as in bridge collapses, putting the safety problem of bridge collapses in a perspective that few people among the public would have guessed. Worldwide, in recent years, the World Health Organization estimates that 1.25 million people die in traffic accidents every year, giving a current annual fatality rate of 160 per million people. In other words, based on a macro analysis, structural safety of bridges, with respect to risk of death, is not in the same order of magnitude as traffic safety. But, can we thereby conclude that bridge safety is not a big concern? No, even if bridges do not collapse, they may be less safe from a traffic-safety viewpoint than other roadway segments because the geometric features of bridges differ from other roadway segments.

Accidents not Related to Structural Failures

It is hard to find information on crashes related to bridges that do not involve structural failures. A literature review in 2017 came up with a published bridge crash study by [Hurt et al. \(2009\)](#) in Kansas. Its findings include, “In 2005, there were 1,919 crashes on bridges in Kansas, including 26 fatal crashes. While overall these accounted for less than 3% of all traffic crashes, they accounted for almost 7% of the total number of fatal crashes.” And, that percentage seems to refer to crashes on bridges rather than under or near bridges.

Functional obsolescence or imperfect geometric design seldom directly cause crashes. But nonoptimal designs may increase the chance of a crash and aggravate injuries once a crash occurs. But it is almost always road user behavior that is the primary factor behind a crash. Therefore, an analysis of fatal bridge injuries, with an intent to reduce that number, should look at how driver and pedestrian behavior can be influenced with, for example, information technology. But certainly, behavior can also be influenced by changes to geometric design.

A substantial subset of fatal injuries on bridges is made up of pedestrians and bicyclists—road users that often have no realistic alternative to using a functionally deficient bridge when they cross a river or an Interstate highway, as illustrated in [Fig. 1](#). Here pedestrians have a choice between a 0.2-m shoulder or a 0.5-m center island when crossing I-95 in Bangor Maine, United States. However, there is certainly not much published literature in the area of pedestrian and bicycle safety of bridges. One study that touches on the subject of pedestrian safety is a Korean study by [Tay et al. \(2011\)](#), which concludes that fatal and serious pedestrian-vehicle crashes in South Korea were associated with collisions involving pedestrians who were on bridges. However, the study does



Figure 1 Bridges over highways often have no safe space for pedestrians and bicyclists.

not go into the influence of geometric features or other characteristics. With respect to bicyclist safety, a Belgian study by [Vandenbulcke et al. \(2014\)](#) can be mentioned. They investigated risk factors for bicyclists limited to infrastructure, traffic, and environmental characteristics. The results suggest that a high risk is statistically associated with the presence of bridges without cycling facilities, such as bike tracks.

American Society of Civil Engineers (ASCE) states in their Report Card for America's infrastructure that one in nine of the nation's bridges are rated as structurally deficient. The Federal Highway Administration (FHWA) estimates that to eliminate the nation's bridge deficiency backlog by 2028, the United States would need to invest \$20.5 billion annually, while only \$12.8 billion is being spent currently. These sums do not include the cost of fixing bridges that are not structurally deficient but unsafe in some other way.

A primary goal if finding that bridges are unsafe may not be to recommend how to redesign existing bridges but how to use ITS and other information technologies to inform (warn) road users of upcoming real-time safety issues. Such a system has been sketched and presented by researchers at the University of Montana ([Vazquez, 2014](#)). At a later development stage, ITS technologies can and should be used to take over control from drivers if they approach a bridge in an unsafe manner, such as at too high speed for conditions or when they are getting too close to pedestrians or bicyclists.

There is a wealth of literature on guardrail and bridge rail design, and crash test results of different designs. And, there is a study of motorcycle fatalities from 2003 to 2008 in the state of Indiana showing that a major correlate of death is bridge-guardrails ([Nunn, 2011](#)). A specific issue with bridge piers, besides being unforgiving, is that they often cut the vision field ahead. That has been identified as a crash contributor for bicyclists in particular ([Jordan and Leso, 2000](#)).

Another specific safety issues relates to passing under low-clearance bridges with commercial vehicles. An experiment using "old-fashioned" ITS technology for reducing such crashes in Maine was presented by [Belz and Garder \(2009\)](#).

A goal when designing bridges should be to reduce human suffering and death caused by unsafe behavior at or near bridges in combination with less-than-optimal geometric design of said bridges. It is very clear that many bridges—at least in the United States—lack sidewalks as well as wide shoulders and therefore cause a special threat to vulnerable road users, as discussed in connection to [Fig. 1](#). Safety is the focus of this article. However, badly designed bridges are frequently the Achilles heel in urban and suburban pedestrian and bicycle networks and certainly have mobility and accessibility implications too.

The analysis below excludes suicides and homicides and covers only transportation-related crashes.

Analysis of Fatal Bridge Crashes in the United States

When analyzing fatal crashes in the United States, the Fatal Analysis Reporting System (FARS) database is typically a reliable source, but only for crashes that involve motor vehicles in traffic on public roads and streets. This database is maintained by the National Highway Traffic Safety Administration (NHTSA) and covers all public roads in the United States, including bridges. However, FARS does not have a field (category) indicating whether a crash occurred on a bridge or not and it is therefore difficult to identify if a crash occurred on or under a bridge or somewhere away from the bridge without going through each individual crash and using GPS coordinates. And, to make things worse, entered GPS coordinates are often inaccurate. A different approach is needed.

Fatal crashes are typically written up in local newspapers and covered by television stations. But, information from the past is sometimes not archived in a way that makes it easy to retrieve. Therefore, only 2016 fatal crashes have been analyzed here. Newspapers and other media sources from all of the United States were searched on a daily basis throughout 2016 and then the search was continued through August 2017 to get updated information on crashes that occurred in 2016. The news articles frequently show photos or videos from the crash site making it possible to identify exactly where a crash happened. This combined with the write-up of the news story typically gives a good picture of what happened, though the information may at times be biased or false, if witnesses lie or reporters misunderstand what happened. There may also be fatal crashes that neither newspapers nor TV

stations find out about, but that would be rare. Rather, most fatal crashes in the United States are covered by multiple newspapers and multiple TV stations.

Google Satellite and Street View were used to get more detailed information about the geometry of the site of the crash.

Results

As earlier mentioned, looking at recent history, bridge failures kill almost five people per year in the United States, and that is not insignificant. But there are many more people killed in crashes on bridges, or going off bridges, or hitting abutments and piers on a road passing under a bridge.

Below are the results of an analysis of “all” fatal crashes on or under bridges in the United States in 2016 and a subset of nonfatal bridge accidents in Maine.

Fatalities on or Under the US Bridges in 2016

As earlier mentioned, fatal crashes on or under bridges were found through analysis of media publications in newspapers, television, etc. It was verified that the crash had not been deemed a suicide and that it physically occurred on or under the bridge/overpass. It can be concluded that the United States in 2016 had at least 258 fatal crashes that were not deemed to be suicides or planned homicides. These crashes killed 314 people on or under bridges. There were 244 car/truck occupants, 35 MC riders, 4 bicyclists, and 31 pedestrians killed. Out of the 31 pedestrians killed, at least 7 were pedestrianized motorists who had gotten out of their motor vehicles after their vehicle had stalled, been involved in a minor crash, or other event.

Overall, 81 of the fatal crashes happened under the bridges (colliding with piers, etc.) and 174 on top of bridges. In 84 of the 174 crashes on bridges, the bridge was traversing water and in the remaining 90 cases, the bridge was passing over another highway, railroad, or similar facility.

Of the 81 crashes under bridges, 65 were single-vehicle strikes of piers or abutments and one was a truck having a raised bed that struck the bridge. Two cases involved pedestrians being hit while walking through underpasses and one case was an MC rider having stopped to shelter from rain. In three cases, an occupant drowned when driving under a flooded underpass.

Of the 174 crashes on top of bridges, 84 were single-vehicle crashes. In 21 of them, the motorist was killed as a result of striking a bridge railing or guardrail without going off the bridge. Of the multi-vehicle crashes, 17 started out as rear-end crashes, 6 as sideswipe, and 27 as head-on crashes whereas 8 of them could not be determined based on media reports. Nineteen of the crashes on bridges involved a pedestrian or pedestrianized motorist. In eight of the 19, the pedestrian was struck and killed without falling off the bridge, in five cases the pedestrian was stuck and thrown down off the bridge, and in the remaining six cases the pedestrian also fell down but it is unclear how it happened except the incidence is not deemed to be a suicide. One fatality occurred as a result of a “pedestrian” climbing a bridge and then lost their balance and fell, so this is not a traffic accident but rather a climbing accident.

Out of the 174 fatal crashes on top of bridges, 89 had a vehicle or its occupants or other road-user being thrown down onto the highway or river below the bridge with 53 of them not being preceded by a multi-vehicle crash but rather were caused by a driver losing control and going through or over bridge railings, showing the need to strengthen and/or raising these. In most of these cases, it was the occupants of the vehicle who were killed but in one case, a motorist survived whereas four “pedestrians” were killed when the vehicle landed on top of them. In 14 crashes, there was a multi-vehicle (or multi-unit) crash preceding the going off of the bridge. Five of the crashes mentioned in the previous paragraph involved a pedestrianized motorist having exited their vehicle and being hit by a passing motorist before being thrown down, and in another two, a motorist got out of their car and, seemingly by mistake, fell down when they stepped over a low railing. These two could possibly have been suicides but were deemed by police as accidents.

The media reports point toward at least 21 of the crashes (8%) involving alcohol or drugs but this may be an underestimate. Ice or snow was a definite contributor in only eight of the cases. In other words, nationwide, better snow or ice removal or real-time information about bridge friction coefficients, would not greatly influence the total number of fatalities.

Structural failures prior to the “accident” did not contribute to any fatalities on the US bridges in 2016. Debris falling from a bridge as a result of a crash took one person’s life.

Two of the 2016 fatalities involved workers doing reconstruction of bridges. One of these was a traffic accident where one worker was killed in a 50-ft. fall after a (non-construction) driver of a truck hit the boom lift, which the worker was using. The other fatal accident was not a traffic accident but involved contractors painting a bridge. The person was standing on a platform tied to the bridge by a series of ropes and cables when one cable broke and the platform suddenly tilted, throwing five workers into the water, killing one of them.

Analysis of Nonfatal Bridge Crashes in the State of Maine

Analysis of nonfatal crashes is here limited to the state of Maine. Maine Department of Transportation provided crash data and AADT information to calculate exposure. However, exposure of pedestrian and bicycle flows were not gathered. The safety analysis covered not just locations where crashes had occurred. Crash-free locations, in particular bridges, were also included.

Maine Bridges

Maine has around 2515 bridges and overpasses that are longer than 20 ft. There are also 1374 bridges with short spans of 10–20 ft. As of 2016, the State of Maine owns and manages 2744 of these bridges. For the 3-year period (2011–13), there were 765 reported crashes on, under or very near bridges in Maine with 79% of them occurring on top of the bridges and 8% under the bridges and the remaining 13% just off the bridge, but still classified as bridge related. Overall, 28% involved personal injury, which is slightly lower than for crashes away from bridges. A significantly higher percentage of crashes occurred during ice and frost conditions on bridges (19.5%) than on other segments (5.1%). Adding snow to that, 30% of bridge crashes in Maine happened when the roadway was covered by snow or had icy conditions. So, even if snow and ice is not a nationwide issue with respect to fatalities, it seems to be a significant issue in Maine, and probably in other northern states, with respect to property-damage crashes.

A detailed analysis of crash rates was done for two of the 16 counties in Maine. The crash rate per hundred million vehicle miles (HMVM) traveled for all bridges was 317 for Aroostook County and 244 for Kennebec County. This can be compared to the average statewide crash rate of around 195 for the 3 years (2011–3). The injury crash rate, including possible injuries, was on bridges 52.9 per HMVM for Aroostook County and 58.8 per HMVM for Kennebec County. The average statewide injury crash rate for all public roads in Maine was around 59 per HMVM for these years. So there were more reported crashes than “expected” but just expected number of injuries, meaning that the average severity of the crashes is lower than for surrounding areas. Maybe crashing into bridge railings cause property damage but, on average, is “safer” than going off the road and hitting utility poles, trees, and other dangerous objects.

The above numbers are averages for all roadways in these counties, and bridges may be more common on some roadway categories than others. If we look at different roadway categories in these two counties, we can see that the average crash rate on Interstate bridges was almost exactly double that of all Interstate segments (130 vs. 64 per HMVM). For other principal arterials, the average crash rate on bridges was also double that of nonbridge segment (223 vs. 114). For major collector roads, bridges also have higher rates than other segments (240 vs. 154). The difference is even greater for local roads with bridges in these two counties having an average crash rate of 1230 versus the average rate of 247 for local roads in this 3-year period. These numbers indicate that we have crashes at and away from bridges, but that the bridges have roughly double the number of crashes compared to other segments on a per mile basis.

Discussion

There is a lack of knowledge of traffic safety issues of bridges. People, including safety experts, often focus on structural concerns rather than functional ones (ASCE, 2013). This article shows that mortality-wise, traffic safety is a problem roughly 60 times greater than structural issues. To focus only on the 1.5% of fatalities that are caused by structural failures will not optimize safety. Obviously, we should not relax structural standards or safety inspections, but we should complement them with bridge roadway safety audits.

Roughly one person dies every day in crashes on or under bridges in the United States. A majority of those killed are occupants of cars or trucks but MC riders and pedestrians are also often victims. About 32% of the fatal crashes occur under the bridge, with a majority being single-vehicle strikes of piers or abutments. In other words, better protection of piers and abutments should be a priority. Of the crashes on top of bridges, almost exactly half are single-vehicle crashes, typically striking guardrails and then sometimes tumbling down from the bridge. In just over half of the crashes that start out on top of bridges, the fatality is a result of a road-user being thrown down onto the river or highway below the bridge. Stronger and higher guardrails should immediately be installed on “all” bridges. Structural failures prior to the “accident” did not contribute to any fatalities on the US bridges in 2016, but the year is a bit of an outlier, and two of the fatalities involved workers doing reconstruction of bridges.

Within a few decades, we may have transitioned to highly automatic vehicles (HAVs) when it comes to purchase of new vehicles. However, non-HAVs will probably continue to travel across bridges for the foreseeable future, at least for several decades. And, even if non-HAVs eventually are prohibited in traffic, there will be pedestrians and bicyclists using urban bridges, so the safety of non-HAVs will still be important.

Biography



Per Erik Gärder is a professor of Civil Engineering at the University of Maine, United States, since 1992. He has a PhD from Lund University in Sweden and he was at the Royal Institute of Technology (KTH) in Stockholm, Sweden from 1983 to 1992. His research interest is focused on forecasting, designing, and evaluating facilities with emphasis on the traffic safety.

Carsharing Safety and Insurance

Elliot Martin, Susan Shaheen, Transportation Sustainability Research Center, University of California, Berkeley, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction and Background	150
Carsharing and Insurance	151
Conclusion and Discussion	155
Acknowledgments	156
Biography	156
References	156

Introduction and Background

Insurance is a critical component of commercial activity in virtually every corner of the economy. It is a centuries-old industry that serves as a hedge against risk, the unknown, and the unpredictable costly events that can happen in daily life. Most people are familiar with the consumer-side of insurance products, including automotive, homeowner/rental, and life insurance policies. Some people also obtain insurance for more specialized risks, such as earthquakes or floods. Additionally, commercial entities from small businesses to corporations seek insurance to protect themselves against the liabilities that arise through the delivery of their product and services. Insurance for business can cover a broad spectrum of different risk types. The specifics of coverage depend on the nature of the business, as well as the potential costs that could be incurred through a realization of the risks it carries.

Carsharing (short-term access to a fleet of vehicles through a private company), as a transportation business, is no exception to the need for insurance. As a service that moves people, it must hedge against a variety of risk factors that can cause property damage, injury, or death. Furthermore, the shared mobility industry is predominantly comprised of private sector enterprises. Of course, many of these enterprises ultimately interface with the public sector, either in the form of curb space allocations or public transit integration. But in spite of these connections to the public right-of-way, the privatized nature of the industry has implications for insurance. This contrasts with other sectors of the surface transportation system, where insurance is either: (1) a cost born by the consumers or (2) irrelevant to vehicles not traditionally covered with insurance (e.g., bicycles) or vehicles run by public agencies (e.g., buses and trains). While public transit systems have risk management or may have supporting insurance policies, such policies and risk are generally implicitly backed by the public sector tax base. This is not a resource available to back catastrophic losses for systems that are almost fully privatized like carsharing. Modern carsharing emerged in North America during the mid-1990s with the establishment of carsharing in Montreal. A few years later, it was established in the United States, beginning in Portland, then Seattle and Boston, before spreading to multiple urban areas in the form of independent nonprofit and for-profit organizations. Because of its pioneering role, it faced a number of challenges in gaining acceptance with cities and consumers, both core to its survival. But beyond those business-related challenges, carsharing also faced major challenges with the insurance industry. The concept of carsharing in the United States was largely untested, and for an insurance industry accustomed to calculated risk and pricing that is supported by vast amounts of empirical data, the prospect of insuring it came with great risk, and thus commanded a significant premium.

At the center of insurance pricing are questions related to the safety risks of specific shared mobility services, including carsharing. These questions relate to the likelihood of injury to the user as well as injury to the general public from the service availability. Questions also relate to concerns associated with private and public property damage. This not only applies to the service assets but also to those of others (e.g., peer-to-peer carsharing in which individuals put their own vehicles into a shared fleet managed by a third-party provider), which could result in misuse or vandalism. Ultimately, questions regarding the level and costs of risk can be answered through experimentation and building empirical evidence with emerging services that depart from the traditional models. But, the early days of carsharing were fraught with challenges convincing insurance companies to cover such services. The absence of such coverage can prevent services from being able to operate and the withdrawal of such coverage can force existing operations to shut down. This latter circumstance was the experience of Buffalo CarShare in 2015, which served lower-income residents of the city, and was required to cease operations when its insurance coverage was withdrawn (Johnson, 2015). This withdrawal was the function of the lack of profits being made by the insurer and insurance laws within New York. Specifically, New York was (and still is) a "no-fault" state, meaning that insurers must cover any economic losses of the insured, including medical bills, forgone earnings, and other expenses resulting directly from an injury, regardless of who is at fault for the incident (New York State, 2020). The purpose of no-fault insurance (which is also known as personal injury protection) is to pay claims quickly. The intention of the law is to limit the time and expense of litigation that otherwise must occur to legally determine who is at fault and by what degree. Such determinations delay the compensation of those injured and lead to the associated stresses, while monetary bills from the costs of the incident accrue. Twelve states and Puerto Rico have laws requiring such insurance policies, all of which were enacted in the 1970s. Five additional states had such policies enacted during this period but later repealed them. While no-fault policies impose a

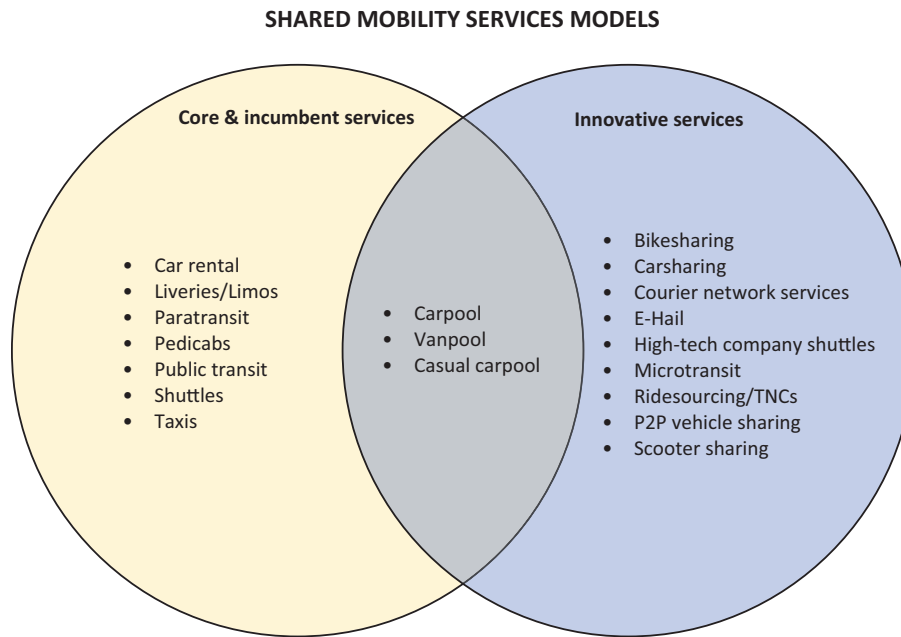


Figure 1 Shared mobility ecosystem.

guaranteed cost on the insurer for any incident, they are also designed to limit the amount of tort litigation that can occur ([Insurance Information Institute, 2020a,b](#)). The existence of these policies has its advantages and disadvantages, but in the specific case of Buffalo CarShare and its insurer, the no-fault requirements were determined to be an insurmountable disadvantage with an accident that was deemed to be *not* the fault of the carsharing driver ([Johnson, 2015](#)). The no-fault legislation increased the exposure of the insurers to a degree that was not acceptable to the industry, and Buffalo CarShare could not find another carrier. Other carsharing operations still do exist in New York state, including the non-profit Ithaca Carshare in upstate New York. Nevertheless, the carsharing landscape is increasingly dominated by larger actors.

The experience of Buffalo CarShare is illustrative of how insurance dynamics can make or break an organization, especially a smaller one within a new industry. It shows that insurance considerations, while generally among the more mundane aspects of business operations, are critically important for survival. Despite the setbacks and lessons learned, the carsharing industry served as an important pioneer for the rest of the shared mobility industry. The full scope of this industry is articulated in [Fig. 1](#), as presently consisting of micromobility, transportation network companies, taxis, microtransit, among others ([Shaheen et al., 2020](#)). As such, the shared mobility industry has evolved into new businesses and a variety of new modes, since carsharing services were first launched. In the sections that follow, the authors discuss insights from previous work as conducted in [Shaheen et al. \(2016\)](#) on carsharing safety and insurance.

Carsharing and Insurance

Previous research from [Shaheen et al. \(2016\)](#) titled: *Understanding Carsharing Risk and Insurance Claims in the United States* explored how insurance rates differed depending on the group that was covered and their crash risk. By analyzing data supplied from six different carsharing operators, conclusions can be drawn regarding carsharing risk. The data supplied included individual claims filed by the operators. The operators were all single city operators, and most operated under a nonprofit business model. The data spanned a total of 328,726 valid trips and contained 125 valid insurance claims. This amounted to 2630 trips per claim. The time span of data differed across operators, but in aggregate it spanned 2008-15 and constituted 28 operating years of data. It also comprised 733 insured vehicle years of data. In other words, each vehicle in the dataset was insured for some number of years (e.g., vehicle-years insured) and 733 is the sum of all those vehicle-years in the dataset. The total costs from claims within the dataset amounted to US \$578,801. The deductibles varied across the operators and policies and ranged from US\$0-10,000 per claim. However, most of the deductibles were on the low end of this range, with 45% of claims carrying a US\$1000 deductible, and 89% of deductibles amounting to US\$1000 or less. To standardize the computed costs to insurers across policies and operators, the cost of all claims re-calculated for this analysis assumed a deductible of US\$1000. Based on this assumption, the average cost per claim was US\$4630, and the median was US\$2189 per claim.

By comparing key risk metrics in carsharing with similar metrics of the broader automotive industry, the data can show the degree to which carsharing differed from the general population. Claim activity in the broader automobile insurance industry is profiled along with a myriad of attributes. A major curator of data in the insurance industry is the Insurance Institute for Highway Safety

(IIHS). The research institute collects and publishes data on automotive claims and additionally conducts its own crash safety tests for the industry (which are different from those of the National Highway Traffic Safety Administration (NHTSA)). The insurance industry measures the rate of claims in units of: "Claims per 100 Insured Vehicle Years." This is the number of claims that the industry receives for every 100 vehicles insured over a period of 365 days. On average, the insurance industry experienced 6.01 collision claims per every 100 vehicle years insured in 2015 (IIHS, 2020). Other types of claims were less frequent in terms of the units, including bodily injury (0.91), property damage (3.72), and comprehensive (2.72). For comparison, the claims in the carsharing dataset suggest that there were 17 claims per 100 insured vehicle-years. The higher claim frequency for carsharing vehicles is notable, but it is also logical for a few reasons. First, shared vehicles of any kind are used more intensively than personal vehicles and by a wider variety of individuals. Second, carsharing vehicles are typically used by demographics that have a higher risk profile within the general population. One example of this is age, where carsharing users are typically younger. As with traditional personal vehicle insurance, younger drivers are regularly categorized as higher risk due to inexperience and sometimes more reckless behavior. This dynamic is shown in broader industry data presented in [Fig. 2](#), which plots several key risk statistics by the age of the claimant. The figure is

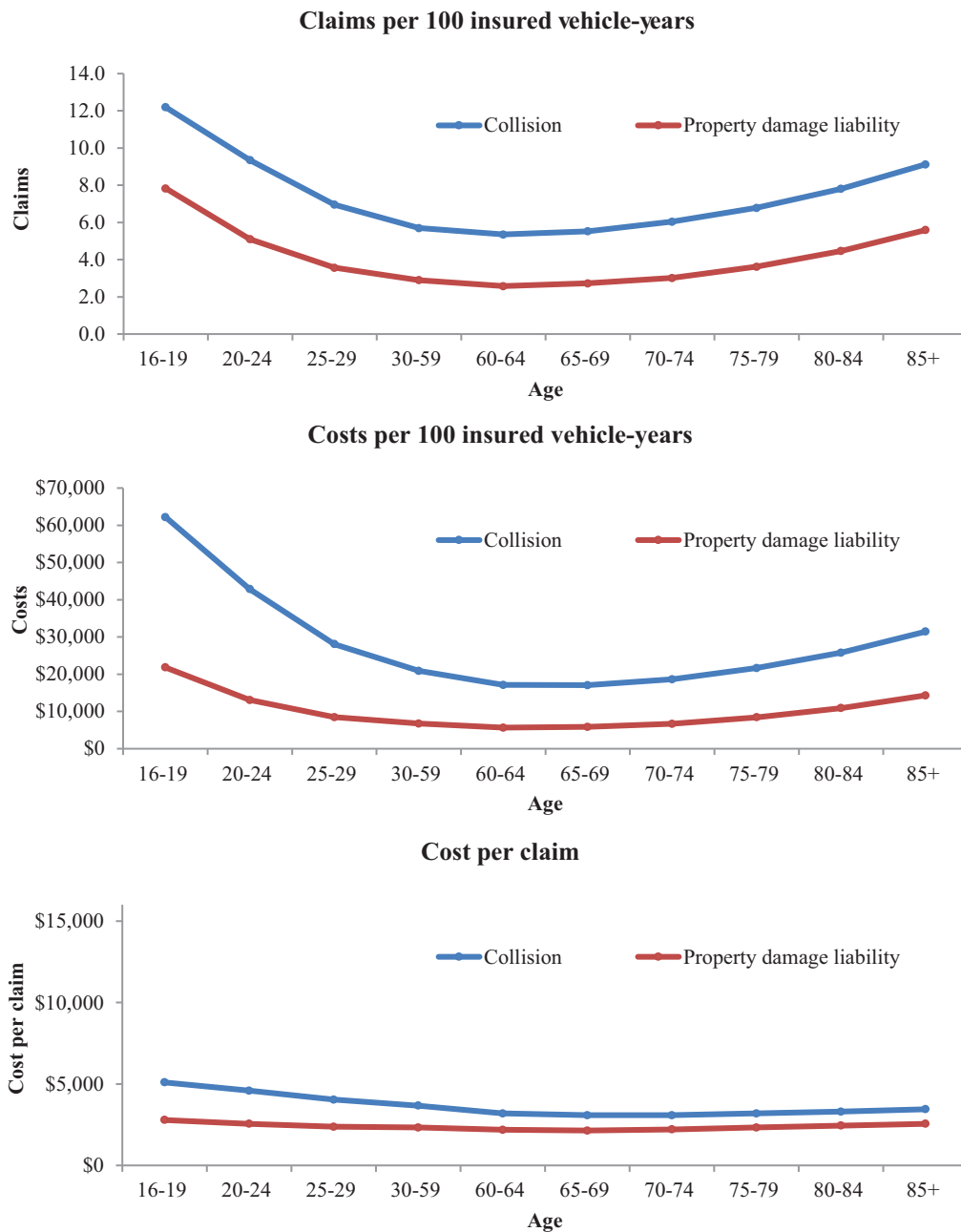


Figure 2 National claim and cost data by age.

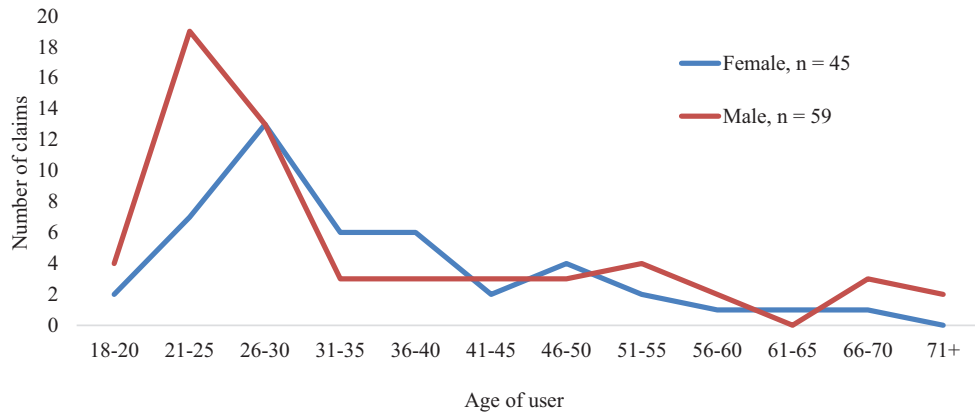


Figure 3 Number of carsharing claims by age.

derived from data assembled from IIHS for the 2002-04 model passenger cars (IIHS, 2007). The data show the relative change in claims per 100 insured vehicle years, costs per 100 insured vehicle years, and cost per claim for collision and property damage claims. Notable within the figure is the convex curve that is clearly present with claims per 100 insured vehicle-years and costs per 100 insured vehicle-years. These two measures find their minimums in the 60s age cohort and their maximums in the younger and older demographics. It is the younger demographic that is overrepresented within the carsharing population relative to the general population. For example, in a survey of North American carsharing operators, [Martin and Shaheen \(2011\)](#), found that just over 35% of respondents were younger than 30, while 31% of respondents were between the ages of 30 and 40. Hence, well over 60% of carsharing users fall within the younger cohort of drivers. These two factors (vehicle use intensity and user age) likely serve to drive up the claims per 100 vehicle years of carsharing fleet vehicles.

[Shaheen et al. \(2016\)](#) found a similar pattern in the data provided by carsharing operators, but the distribution was not as pronounced when adjusted for miles driven. But in aggregate claims, the pattern holds in that the majority of claims are filed by users who are mostly in their 20s. Different adjustments, depending on the denominator, suggest the correlation with user age is mostly due to a large number of users within the age cohort of the dataset. Of the 125 claims within the dataset, 106 had age attributes associated with the record. The distribution of claims by user age and gender is shown in [Fig. 3](#). Note that the total sample size is 104, as two of the observations with age information did not have a gender attribute.

As noted above and in previous work, adjustments to the claims significantly reduce the representation of the younger cohort in claim activity. When claims are adjusted by certain denominators that control for the level of use, the representation of the younger user cohort significantly diminishes. This is shown in [Fig. 4](#), where the claims per 1000 trips are far more level across the different user age groups.

[Fig. 4](#) shows that when the age of the user by the claim is controlled for by the number of trips taken by the age cohort, the contribution of younger users is reduced on a relative basis. This supports the concept that the higher claims per 100 insured vehicle years found within carsharing overall is more a function of the mass of younger users rather than exceptionally risky behavior of this cohort relative to the general population of the same age.

Other data can yield insights on the patterns of carsharing use across key age cohorts, which show some limited variation. The value of this assessment is that it shows that the difference in age-related patterns is generally not a function of use. All carsharing

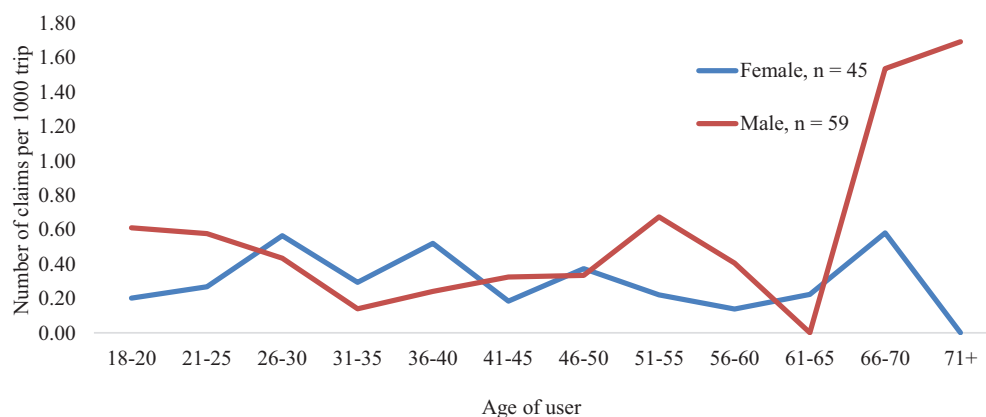


Figure 4 Age of user by number of carsharing claims per 1000 trips.

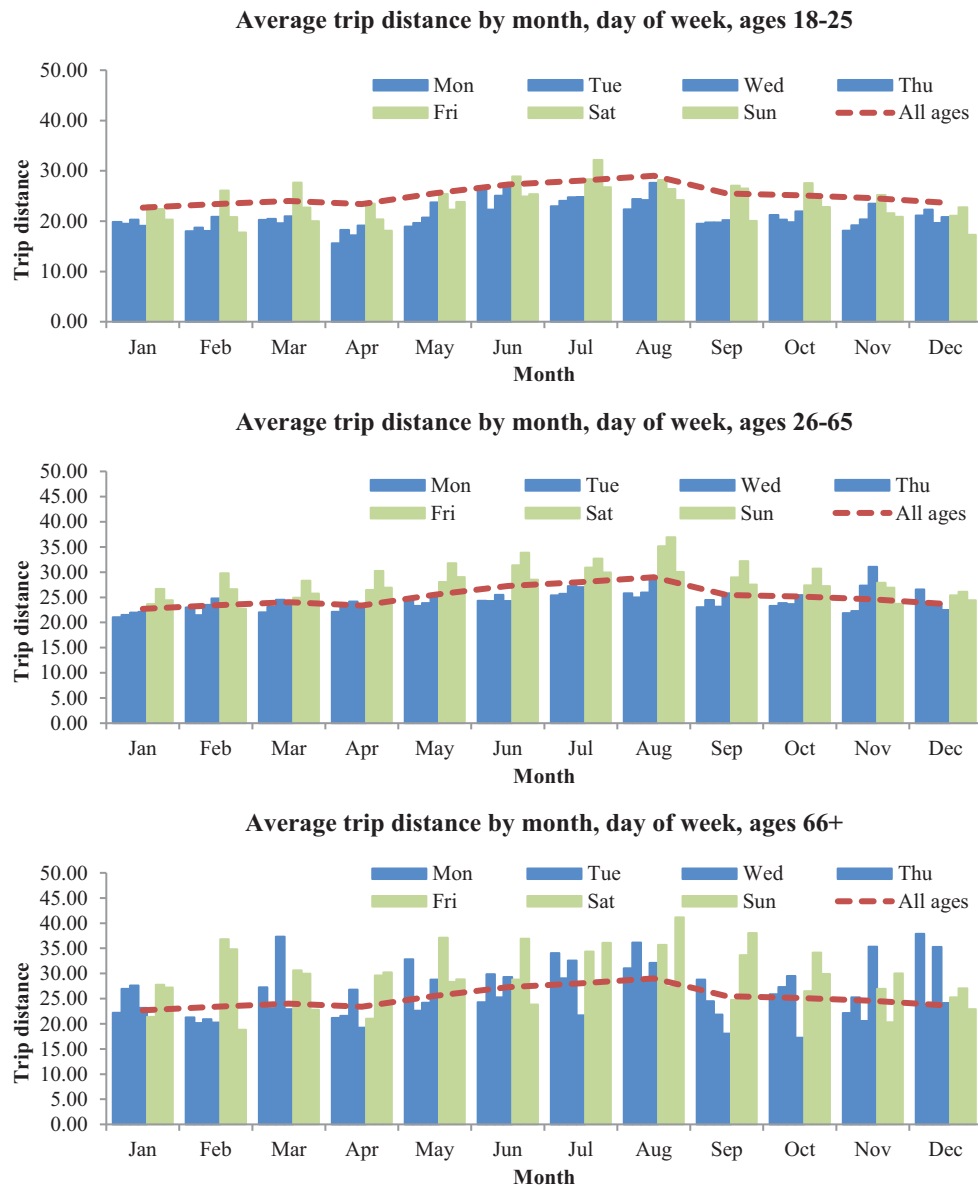


Figure 5 Carsharing trip distance by age, distance, day of the week, and month.

users appear to exhibit the same trends in the trip distance over the course of the year, where the distance of trips slightly increases during the summer and increases on weekends. The absence of key distinctions is shown in [Fig. 5](#).

Finally, usage patterns in terms of carsharing trips over the year also show limited distinctions by age cohort. For example, as shown in [Fig. 6](#), when evaluating carsharing trips by month, day of the week, and within the same age cohorts presented in [Fig. 5](#), the key distinction is a notable drop in trip activity among the younger age group of 18-25-year olds.

Here, the impact of summer vacation is clearly evident, where a drop in trip count may contribute to a drop in claims given the contribution younger members made to claim activity within the carsharing dataset.

The data presented show that the measurement of carsharing risk depends on the unit of measurement. This may be intuitive, but it is worth emphasizing when evaluating risks within the domain of shared mobility use. Under traditional industry approaches to risk measurement, namely the use of claims per 100 insured vehicle years, carsharing performs worse than the general population insuring their personal vehicles. However, this is due to the nature of the user and the system's relationship to the vehicle. Carsharing vehicles are used more intensely than personal vehicles, and a large share of vehicle users are within an age category that exhibits higher overall industry risk. When several use factors control for the amount for use, the results show that the risk of carsharing users is relatively equal on a per-trip basis across the younger age cohort.

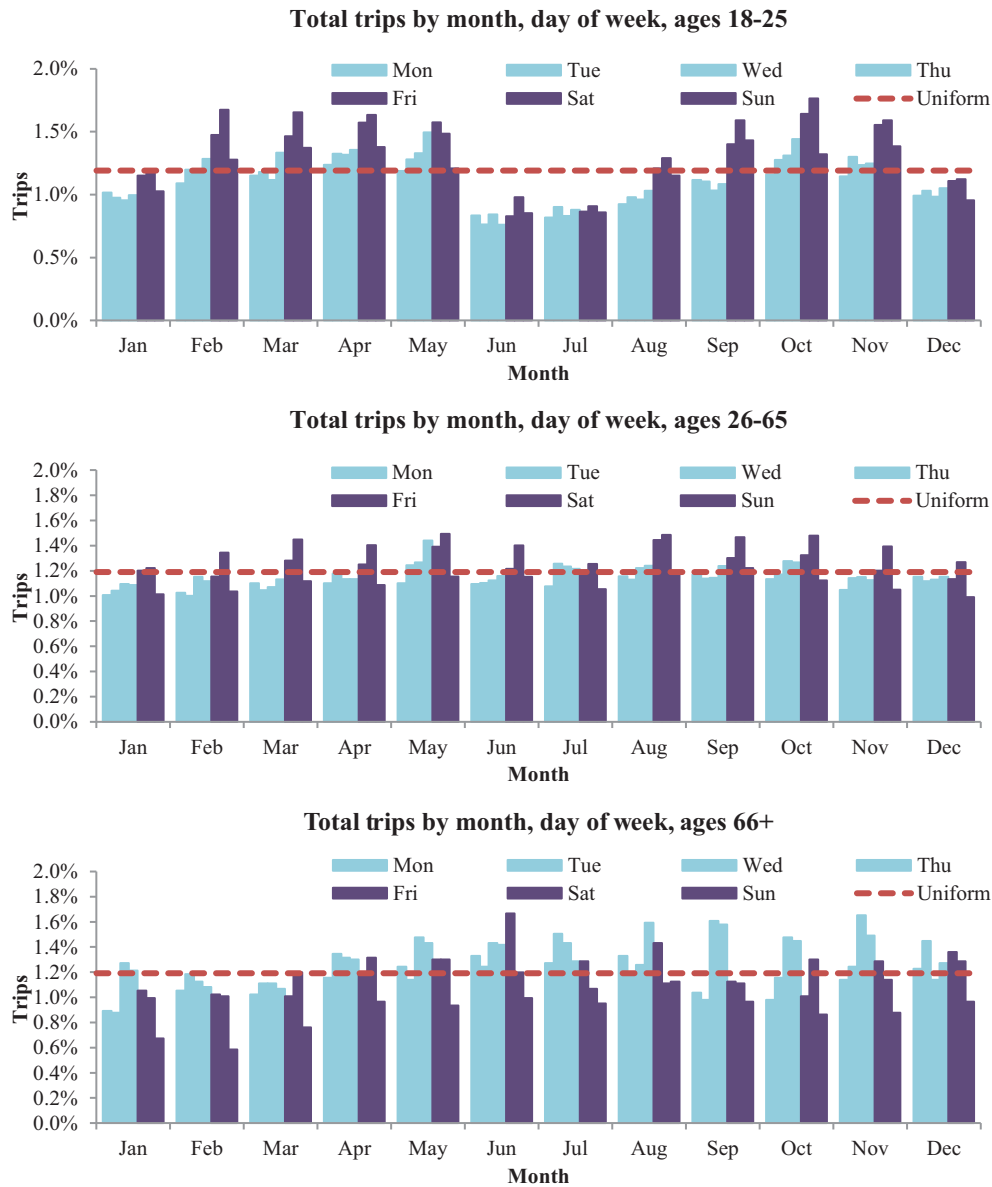


Figure 6 Trips by month, day of the week, and age by month.

Conclusion and Discussion

The study findings, which are also evaluated in the related study (Shaheen et al., 2016), suggest that metrics applied in the traditional insurance industry are not necessarily the best metrics to be applied to the carsharing industry. Carsharing safety, when controlled for by the level of use and measured by claims, did not vary across age cohorts as measured in the traditional auto insurance industry. Inevitably, the user profile of any system must be considered, and if the user profile is balanced toward riskier cohorts, then appropriate adjustments of assessment must be considered. But notably, this is a function of the cohort itself, rather than the services delivered, which otherwise aid the welfare of the cohort it is serving. Carsharing's insurance experience has been inevitably challenging, but its role in pioneering the relationship between shared mobility and the insurance industry is indisputable. Nevertheless, the evolution of the larger shared mobility ecosystem requires different metrics for each modality employed by a broader population, reflecting modal differences and unique user populations. These considerations may be relevant through the assessment of other sectors of the shared mobility industry. Such a cross-industry understanding is important. Failures to appropriately manage risk and relationships with the insurance industry can shut down an organization or significantly raise costs in industry. A better understanding of insurance pricing in carsharing and the larger shared mobility ecosystem can lead to better risk management within the industry, which will deliver greater service and cost stability to shared mobility users.

Acknowledgments

This article was based on a study conducted by [Shaheen et al. \(2016\)](#) that was originally funded by Assurant. Co-author Diwen Shen conducted data cleaning and analysis in support of the discussion and analysis within this article. Arupa Adikary also contributed background research in support of this article. Six carsharing operators provided the data necessary to conduct this research. The discussion, findings, and conclusions are the responsibility of the authors of this article.

Biography

Elliot Martin is a Research and Development Engineer at the Transportation Sustainability Research Center, University of California, Berkeley, CA, United States. His work primarily covers shared mobility, freight transportation, public transit, and parking. He conducts impact evaluation of transportation systems and advances the development of methods and data structures designed to measure how related changes in travel behavior translate to broader system impacts. He earned his PhD from UC Berkeley, his undergraduate degree from Johns Hopkins, and previously worked at the Federal Reserve Bank of Richmond.

References

- Institute for Insurance and Highway Safety, 2007. Insurance losses by rated driver age. Institute for Insurance and Highway Safety. https://www.iihs.org/media/3ae80bcd-74d9-4a1d-a669-4af65a8400d6/fEDkEQ/HLDI%20Research/Loss%20fact%20sheets/All_age_02_04.pdf.
- Insurance Information Institute, 2020a. Background On: No-fault Auto Insurance. <https://www.iii.org/article/background-on-no-fault-auto-insurance>.
- Insurance Information Institute, 2020b. Facts + Statistics: Auto Insurance. Insurance Information Institute. <https://www.iii.org/fact-statistic/facts-statistics-auto-insurance>.
- Johnson, C., 2015. Buffalo CarShare ceases operation due to New York Insurance Law. Shareable.
- Martin, Shaheen, 2011. Greenhouse gas emission impacts of carsharing in North America. *IEEE Trans. Intell. Transport. Syst.* 12 (4.).
- New York State, 2020. Minimum Auto Insurance Requirements Coverage. https://www.dfs.ny.gov/consumers/auto_insurance/minimum_auto_insurance_requirements. Accessed July 2020.
- Shaheen, S., Diwen, S., Martin, E., 2016. Understanding carsharing risk and insurance claims in the United States. *Transport. Res. Rec.* <https://doi.org/10.3141/2542-10> and <https://escholarship.org/uc/item/35j2g1b9>.
- Shaheen, S., Chan, N., Bansal, A., Cohen, A., 2020. Chapter 13: Sharing strategies: carsharing, shared micromobility (bikesharing and scooter sharing), and innovative mobility modes. *Transport. Land Use Environ. Plann.* ISBN 9780128151686, pp. 237-262. <https://doi.org/10.1016/B978-0-12-815167-9.00013-X> <https://escholarship.org/uc/item/0z9711dw>.

Carjacking

Terance D. Miethe, Christopher Forepaugh, Tanya Dudinskaya, Department of Criminal Justice, UNLV, Las Vegas, NV, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	157
Definitional Issues	157
Typologies of Carjacking	158
Level of Planning and Motive	158
Method of Acquisition	158
Bump-and-Rob	159
Trust Violations	159
Blitz Attack	159
Normalcy Illusion	159
Other Methods (Police Impersonator)	160
Prevalence and Characteristics of Carjacking	160
Prevalence	160
Socioecological Characteristics	160
Street Culture and Carjacking	160
Prevention Activities	161
Situational Crime Prevention	161
Increased Effort for Carjacking	161
Increase Risks for Carjacking	162
Reducing Rewards for Carjacking	162
Reducing Provocations	162
Removing Excuses	162
Summary and Conclusions	162
References	163

Introduction

Carjacking is a type of motor vehicle theft. However, its specific definitional elements exhibit major cross-national differences. National differences are found in (1) the type of motorized vehicle included (e.g., cars/trucks, watercraft, scooters, and/or rail transport), (2) individuals involved (e.g., strangers, non-authorized persons, or no restrictions on offender-victim relationship), (3) occupancy level (e.g., occupied vehicles, unoccupied vehicles, or both), (4) offense outcome (e.g., attempted or completed thefts), and (5) offender motivation (e.g., "joyriding," sale/trafficking of vehicles, or vehicle parts). The level of physical threat utilized and injury to the victim are other factors that influence the more general classification of carjacking as a property crime (e.g., theft without injury or threat) or violent crime (e.g., robbery, assault/battery and/or murder) across nations.

This essay examines these definitional issues around carjacking, motivations and methods of acquisition, its estimated national prevalence, and socioecological factors associated with these crimes. Situational crime prevention practices are also discussed for reducing the prevalence of carjacking and other transportation-related crimes.

Definitional Issues

Most countries do not have separate criminal statutes for carjacking and there is no uniform definition of these offenses across nations. Instead, estimates of the nature and prevalence of carjacking are often derived from statistics on other crime categories (e.g., auto theft, robberies) or through analyses of offense-related elements identified in surveys of crime victims. The United States National Crime Victimization Survey (NCVS) illustrates the use of offense-related elements by defining carjacking as "a completed or attempted robbery in which a car or other motor vehicle was taken or an attempt was made to take it and the offender was a stranger to the victim" (Klaus, 2004).

Compared to other crimes, carjacking is essentially a motor vehicle theft that involves an occupied vehicle and the threat or use of violence against the driver (Cesar and Decker, 2017). Similar to traditional robbery, carjacking is a transactional crime. Carjackers must confront victims and demand compliance through force or the threat of force. They also take advantage of an immediate opportunity and rely on speed and stealth to approach their victims. However, carjacking differs from other robbery situations because the targeted vehicle is mobile, the victim is removed from the vehicle, and there is a greater risk of physical injury and victim resistance when the vehicle is used as both a shield and a weapon against the offender (Copes et al., 2011; Jacobs, 2012). Unlike in

thefts of unoccupied vehicles, the carjacker must confront the vehicle's driver, gain compliance, and contend with a shorter window of opportunity to commit the offense.

Typologies of Carjacking

Several typologies or classification system have been used in past research to identify different types of carjacking. These include the classification of carjacking by (1) level of planning and motive and (2) method of acquisition.

Level of Planning and Motive

Young and Borzycki (2008) suggest that carjacking can be differentiated along two continuums: level of planning (i.e., organized vs. opportunistic) and motive (i.e., instrumental vs. acquisitive). The primary characteristics of these four types of carjacking based on their different levels of planning and motive are shown in **Table 1** and summarized as follows.

"Organized and instrumental carjacking" involves some planning before the offense. The vehicle is taken in order to facilitate another criminal offense. For example, the vehicle may be taken to fulfill a "smash and grab" heist (e.g., running cars into ATM machines or the front windows of convenience stores and other commercial businesses) or as temporary transportation to conduct a drive-by shooting. The vehicle itself is of no direct financial value to the offender and will likely be abandoned after serving its intended purpose.

"Organized and acquisitive carjackings" are also characterized by prior planning, whether in the type of vehicle targeted or in the method of acquisition. The goal is to obtain a vehicle for its direct financial value and sell it as soon as possible to a willing buyer or to a chop shop (who will strip the vehicle to sell its parts). For example, carjackers may target vehicles of specific makes and models because there is a market demand for such vehicles (Young and Borzycki, 2008). The market value of the particular type of vehicle selected for "jacking" could be based on the popularity of the vehicle or the usefulness of its parts.

"Opportunistic and instrumental carjackings" involve no real prior planning. As the name suggests, these incidents are entirely motivated by a potential offender recognizing a window of opportunity and acting upon it. The carjacker in this scenario is very likely unarmed and acting alone. The vehicle is taken because it caught the offender's attention, whether for its value or for the perceived opportunity to act. The stolen vehicle will be sold later or stripped for parts, but there was no specific market in mind when the vehicle was taken (Young and Borzycki, 2008).

"Opportunistic and acquisitive carjackings" are distinct because they do not involve any real planning or any financial motivation. Such events are typically precipitated by an offender or group of offenders seeing an opportunity and taking the vehicle. The vehicle may be taken for a joyride. It may be taken to boost one's social status or to punish another for flaunting his or her material wealth. However, the vehicle itself is not taken for any future purpose and is likely abandoned shortly after.

Method of Acquisition

When classifying carjackings by their method of acquisition, prior research has identified five distinct categories. Each has particular features that identify their unique signatures.

The "bump-and-rob" carjacking involves multiple offenders who initiate the criminal incident by ramming the target vehicle with another vehicle. This is usually done from behind and at a low speed. When the driver exits the target vehicle to check for

Table 1 Planning and motivation classification for carjacking

<i>Level of planning/Motivation</i>	<i>Attributes</i>
Organized/Instrumental 	Vehicle taken to facilitate other crime Vehicle has no direct financial value Vehicle abandoned after purpose is served
Organized/Acquisitive 	Vehicle targeted for direct financial benefit Vehicle sold as soon as possible Targeting centers on market demand
Opportunistic/Instrumental 	No prior planning Vehicle simply caught offender's attention Vehicle will be sold later but vehicle's market did not prompt the carjacking Likely a single, unarmed offender
Opportunistic/Acquisitive 	No future purpose intended for vehicle Vehicle is abandoned shortly after Often motivated by respect or social status

Source: Adapted from Young and Borzycki (2008)

damage or to request the at-fault driver's insurance information, one of the other offenders jumps into the target vehicle and both vehicles quickly drive away.

The "trust violation" occurs when the carjacker has been granted possession of the vehicle or allowed access to the driver seat but is expected to return the vehicle. One example of these incidents is when a carjacker test drives a vehicle for sale and forces the owner or sales agent to exit the vehicle after leaving the lot. A valet parking agent (or person pretending to be one) who takes a customer's car and does not return it is another example of these trust violations.

The "normalcy illusion" method involves a carjacker approaching the target in a manner that initially seems harmless (Jacobs, 2012). The goal is to lull the target into a false sense of security by pretending to engage in a nonthreatening task, such as asking for help, cigarettes, directions, or the time. The key is that normal people would not feel threatened by such seemingly innocuous encounters. In addition, the normalcy illusion reduces the amount of attention from witnesses who may interfere or contact law enforcement. The carjacker gets close to the driver and either draws a weapon or simply pulls the distracted driver from the vehicle.

The "blitz attack" is another method of acquisition in carjacking scenarios. In contrast to the stealth underlying the normalcy illusion method, the blitz attack utilizes speed, shock, and force (Jacobs, 2012). Blitz attacks are risky, and timing is critical to success. The goal is to catch the driver off-guard and leave him or her with no choice but to comply. Examples of blitz attacks include running up to the vehicle at a red light and brandishing a weapon. Another example is following a target until the vehicle is parked, blocking the vehicle's escape with another vehicle, and then running up to one of the windows.

"Other" methods of acquisition in carjacking situations involve an assortment of approaches to immobilize the motorist through physical or verbal means. These include such activities as running motorists off the road, pretending to be a police officer, pretending to be a hitchhiker, blocking roads with various objects (e.g., bricks, stones, oil drums, or traffic cones) and faking a car accident or flat tire in order to prey upon those who stop to help.

Police reports and interviews with actual carjackers provide qualitative data to illustrate these different methods of acquisition in carjacking situations. The following descriptive accounts of carjacking methods have been constructed and rewritten from short, abbreviated summaries of cases within these data sources:

Bump-and-Rob

- Two male suspects followed a male victim down a side-street before ramming the victim's vehicle with their own. The victim pulled over and exited his vehicle to confront the suspects. After a brief confrontation, one suspect drew a gun and fired several shots. After the victim went to the ground, the other suspect entered the victim's truck and sped off.
- A female victim traveled on an interstate highway when the suspect's vehicle flashed its headlights at her, signaling her to pull over. The suspect followed the victim off the highway and then rear-ended her vehicle. When the victim exited her vehicle, the suspect pointed a gun at the victim's head, took her keys, and drove away in her vehicle. The suspect's partner drove the ramming vehicle away.

Trust Violations

- A salesman at a car dealership presented a vehicle to the male suspect and female suspect. Both suspects sat in the vehicle before pushing the salesman away. The suspects then sped off the lot, nearly striking a person as they fled.
- A male victim drove to the entrance of an upscale restaurant. The male suspect approached and asked if the victim wanted to have the valet (i.e., parking attendant) park his vehicle. The victim handed his keys to the suspect. Later, the victim went to retrieve his vehicle and learned both that the staff had no record of his vehicle and that no one matching the suspect's description worked there.

Blitz Attack

- A male victim slowed his vehicle for a red light when the suspects' vehicle pulled in front of him. One suspect exited the vehicle while holding a gun and told the victim to "get the f*ck out of the car." The suspect then entered the victim's vehicle and fled while the second suspect drove their vehicle.
- A teenage driver waited at a traffic signal after borrowing his friend's truck. A male suspect entered the truck and began striking the teenager with his fists in order to force him to exit from the vehicle. After the driver exited the vehicle, the suspect moved into the driver's seat and drove away.

Normalcy Illusion

- An elderly victim waited in his car while his wife shopped inside a store. The male suspect approached the victim's car and asked for the time. The suspect returned shortly after to ask for spare change. The suspect then returned a third time, this time with a gun, and threatened to shoot the victim, if he did not exit the vehicle. The victim stalled but eventually complied. The suspect then drove away.
- Two male suspects approached a couple in a parking lot. One suspect asked the couple for cigarettes. One victim gave the suspect a cigarette, but the suspect then drew a gun and pointed it at the male victim. The other suspect demanded the car keys and the female victim's purse. The suspects then drove away in the vehicle.

Other Methods (Police Impersonator)

- Four victims parked outside a liquor store. The driver went inside. A male suspect approached the other three victims and claimed he was a police officer who urgently needed to respond to a bank robbery. The victims exited the car and allowed the suspect to drive away. The victims then flagged down a police officer and learned that no such bank robbery had occurred.

Prevalence and Characteristics of Carjacking

Definitional and methodological differences limit cross-national estimates of the prevalence and nature of carjacking. The problems are compounded by the general lack of national police data on carjacking and the need to extrapolate patterns from other crime categories (e.g., vehicle thefts, robberies) and incident attributes in victimization surveys. What we know about the prevalence of carjacking and the socioecological factors associated with it are summarized as follows.

Prevalence

Compared to the prevalence of motor vehicle thefts and armed robberies, available data from several nations and major cities indicate that carjackings are relatively rare. In the most comprehensive study of carjacking using the United States national victimization data, [Klaus \(2004\)](#) estimated that an average of 34,000 carjackings occurred in the United States annually from 1993–2003. The victimization rate for carjacking was 17 per 100,000 population. In contrast, rates of auto theft victimization (910 per 100,000) and personal robbery (250 per 100,000) were substantially higher during this time period ([Rennison and Rand, 2003](#)). Among US cities, the Chicago police department reported over 1000 carjacking incidents in 2017. In St. Louis in 2017, more than 250 carjacking incidents were reported to the police. News accounts estimate that Detroit had over 700 carjackings in 2014.

Estimates of the prevalence of carjacking vary widely across other nations. For example, South Africa has reportedly one of the highest rates of car “hijacking” in the world, with an estimated 11,000 of these incidents in 2014. The rate of these incidents in 2000 was 34 per 100,000 residents ([Davis, 2003](#)). Within the United Kingdom and Australia, carjacking is rare, representing less than 1% of vehicle thefts. In contrast, the majority of thefts of insured motor vehicles in Mexico are taken “with violence” ([Espinoza, 2019](#)) and the carjacking of tourists also appears to be a major concern in other nations (e.g., Brazil, Costa Rica, France, India, Spain). Across all world regions, there is a general belief that carjacking has increased over the last two decades in large part because technology is making it more difficult to enter, “hot wire,” and steal unoccupied vehicles.

Socioecological Characteristics

Similar to other crimes, carjacking is not randomly distributed over persons, places, and situations. Instead, particular types of people, places, and situational contexts are associated with higher risks of carjacking than others.

Based on previous studies, carjackers are typically young males. Black and Hispanic persons are over-represented among victims of these crimes within the United States ([Klaus, 2004](#)). Most carjacking incidents involve multiple offenders, the use of weapons to gain victim compliance, and some physical injury to the victim.

In terms of its social ecology, carjacking is spatially concentrated within densely populated urban areas. They occur disproportionately at night in open areas (e.g., on the street or near public transportation) and near commercial places (e.g., stores, gas stations). Within economically disadvantaged areas in large cities, carjacking may be especially common for fulfilling multiple extrinsic and intrinsic rewards, including economic opportunities, thrill-seeking, and status-enhancement of one’s street reputation ([Cesar and Decker, 2017](#); [Jacobs et al., 2003](#)).

Street Culture and Carjacking

Street culture plays a large role in carjacking in large urban areas within American society. The street culture emphasizes a pleasure-driven lifestyle, respect, and going with the flow ([Jacobs et al., 2003](#)). Participants must subsidize this hedonistic lifestyle, but they often have limited access to legitimate means of acquiring resources. This lifestyle centers on drug and alcohol consumption, material pursuit, reckless spending, and sexual conquests. The cash-intensive nature of these activities creates a perpetual need for money. Carjacking can help support this lifestyle by providing a means of obtaining money without having to rely on an ordinary job, but the increased resources acquired through successful carjacking may finance more intensive partying, again fueling the need for more money ([Jacobs et al., 2003](#)). Street culture encourages living in the moment. Criminal offenses might arise from circumstances beyond individuals’ immediate control and force them to participate (e.g., participating in a carjacking begun by an associate). Street culture encourages thrill-seeking, which might be accomplished through the adrenaline rush of overcoming danger or using force to get one’s way. Carjacking provides both.

Another important aspect of American street culture involves a moralism centered primarily on respect. In the streets, affronts must be met with punishment, and no affront is too trivial ([Jacobs et al., 2003](#)). Those who flaunt their material possessions run the risk of offending those who have less. Carjacking facilitates the purpose of retaliation by allowing offenders to put social violators

“back in their place” by breaking the violators’ feelings of invincibility (Jacobs et al., 2003). At the same time, flaunting material wealth is about one-upmanship. Joyriding in a jacked vehicle allows the offenders to become flaunters themselves and to reassert their dominance by flaunting their conquest to friends and neighbors alike.

Prevention Activities

Previous studies have identified several types of prevention activities to reduce individual’s personal risks of being carjacked and the offender’s motivation for committing these crimes. Examples of these control strategies include those that focus on the general methods of situational crime prevention (SCP).

Situational Crime Prevention

Situational crime prevention is a criminological perspective rooted in criminal opportunity theories. The general idea is to alter criminal opportunity in order to deter rational offenders. Carjackers are opportunistic offenders. Under SCP, these are the kind of offenders who can be deterred.

As shown in Fig. 1, SCP has five main approaches to reducing criminal opportunity (Brown, 2015; Clarke, 1995, 2010). First, increase the effort required to carry out a crime. Second, increase the risks involved in completing a crime. Third, reduce the expected rewards associated with the crime. Fourth, remove excuses that an offender would use to justify his or her actions. Finally, reduce the provocations that would tempt a person to transition from a potential offender to an actual offender. Specific SCP strategies aimed at reducing carjacking are summarized as follows.

Increased Effort for Carjacking

Increasing the effort for carrying out carjacking is one of the first steps in prevention. “Target hardening” strategies, those that make it more difficult for suspects to gaining access to vehicles and remove them without the owner’s permission, should be a priority. Establishing greater control over access points to facilities and screening entrants to various locations (e.g., parking garages, commercial business lots, gated communities) are some general ways of increasing the effort for carjackers. Particular examples of these strategies for target hardening include the following:

- Enhance the public’s knowledge of the common methods used by carjackers (e.g., bump-and-rob’s, trust violations, blitz attacks). An aware public can decrease the physical opportunity for these offenses.
- Increase the installation of remote “kill switches” (e.g., engine/ignition immobilizers) and GPS tracking devices to limit the opportunity for unauthorized use of motor vehicles.
- Improve and apply other monitoring sensors and facial/fingerprint recognition technology to restrict the pool of eligible drivers of each motorized vehicle.



Figure 1 Situational crime prevention strategies. Source: Adapted from Clarke (1995)

- Place barriers and gates in all major parking structures to reduce the unabated pathways for exiting from these structures.
- Hire security in public places (especially at night) to escort customers to their vehicles. Once you reach your vehicle, enter quickly, lock your doors, and drive away immediately.

Increase Risks for Carjacking

In addition to injury and death from the incident itself, the primary risks for carjackers are the threat of criminal punishment. From a crime deterrence perspective, the threat of punishment deters crime when it is perceived as swift, certain, and severe. Swift punishment for carjackers may derive from acts of self-protection by victims and their “strategic resistance” (Miethe and Sousa, 2009). Within many US cities, the mere threat of an armed confrontation with the vehicle’s driver may sufficiently deter these crimes. Several particular ways to increase the offender’s risks for carjacking through the swift, certain, and severe threat of criminal sanctions include the following:

- Increase the use of visual surveillance technology in public spaces (e.g., parking lots, streets/highways) to improve the speed and likelihood of carjacker’s identification and arrest.
- Increase the natural surveillance in public spaces (especially in more isolated areas at night) through enhanced lighting and environmental redesign (e.g., remove excessive foliage that decreases the visibility of these places).
- Increase the severity of criminal punishment for carjacking by prosecuting carjackers in federal court, where both the certainty of conviction and severity of punishment are higher than in most state courts. Arrest and prosecute carjackers for multiple offenses, including vehicle theft, personal robbery, and assault.

Reducing Rewards for Carjacking

Carjackers are motivated by various anticipated rewards that derive from their actions. The financial motivations associated with carjacking may be reduced through the following practices:

- Remove expensive and portable items (e.g., cell phones, briefcases, and luggage) from unoccupied vehicles that may attract economically motivated and opportunistic carjackers.
- Tighten local, state, and national controls over vehicle licensing/registration and the marketing of vehicle parts. If carjackers are less able to sell their stolen vehicles or parts from them because of these controls, the financial motivations and anticipated economic rewards may be diminished.

Reducing Provocations

Carjacking is committed by people who either lack the economic resources to purchase a motor vehicle or who seek alternative ways of getting them for material (i.e., economic) and nonmaterial motives (e.g., thrill, peer pressure, feelings of empowerment). Based on the diverse motivations, another set of strategies to reduce carjacking involves decreasing the different provocations that often underlie these crimes. Particular ways to decrease these provocations for carjacking include the following:

- Increase meaningful economic opportunities for the disadvantaged to reduce the profit-motivation for carjacking.
- Increase public service campaigns for juveniles and young adults that promote personal responsibility, decrease susceptibility to peer pressure, and encourage self-control when exposed to immediate situational opportunities for joyriding and carjacking.

Removing Excuses

Many offenders, including carjackers, are able to commit criminal acts because they can rationalize and excuse their actions. The following excuses that carjackers often use to justify their conduct may be nullified by the following actions and counter evidence:

- Reduce drugs and alcohol abuse to decrease impaired judgment caused by substance abuse and the denial of responsibility for one’s actions because of this impairment.
- Negate the excuse of “denial of injury” through educational campaigns that demonstrate the adverse economic, physical, and psychological consequences of carjacking on its victims.
- Counter the offender’s excuse of “blaming the victim” (e.g., belief that the victim was somehow responsible by being careless or provoking the incident) by increasing public knowledge that victim carelessness is not a legal excuse for criminal liability in carjacking incidents.

Summary and Conclusions

Carjacking is a type of motor vehicle theft that often has financial, physical, and psychological consequences for its victims. It is motivated by opportunity, financial gain, and nonmaterial factors such as thrill, peer pressure, and feelings of empowerment. With increase in technology that both enhance and inhibit criminal opportunities, efforts to control carjacking through prevention strategies will have to evolve over time. However, situational crime prevention approaches offer an important perspective for

reducing carjacking and other transportation-related crimes because they focus on doing something about both limiting criminal opportunities in everyday life and decreasing offender's motivations for committing these offenses.

References

- Brown, R., 2015. Crime prevention design in a vehicle registration system: a case study from Australia. *Crime Sci.* 4 (25), 1–10.
- Cesar, G.T., Decker, S., 2017. Cold-blooded and badass: A "hot/cool" approach to understanding carjackers' decisions. In: *The Oxford Handbook of Offender Decision Making*, Oxford University Press, Oxford, pp. 611–632.
- Clarke, R.V., 1995. Situational Crime Prevention. In: Tonry, M. (Ed.), *Crime and Justice: A Review of Research*, Vol. 19. University of Chicago Press, Chicago, IL, pp. 91–150.
- Clarke, R.V., 2010. Thefts of and from cars in parking facilities. *Problem-Oriented Guides for Police: Problem-Specific Guides Series*, 10. US Department of Justice Office of Community Oriented Policing Services, United States, pp. 1–56.
- Copes, H., Hochstetler, A., Cherbonneau, M., 2011. Getting the upper hand: scripts for managing victim resistance in carjackings. *J. Res. Crime Delinq.* 49 (2), 249–268.
- Davis, L., 2003. Carjacking—insights from South Africa to a new crime problem. *Aust. N. Z. J. Criminol.* 36 (2), 173–191.
- Espinoza, C., 2019. Mexican vehicle theft report, 2018. Available from: <https://www.mexinsurance.com/blog/mexican-vehicle-theft-report-2018/>
- Jacobs, B.A., 2012. Carjacking and copresence. *J. Res. Crime Delinq.* 49 (4), 471–488.
- Jacobs, B.A., Topalli, V., Wright, R., 2003. Carjacking, street life and offender motivation. *Br. J. Criminol.* 43 (4), 673–688.
- Klaus, P., 2004. Carjacking, 1993–2002. United States Bureau of Justice Statistics, Washington, DC.
- Miethe, T.D., Sousa, W.H., 2009. Carjacking and its consequences: a situational analysis of risk factors for differential outcomes. *Secur. J.* 23, 241–258.
- Rennison, C.M., Rand, M.R., 2003. *Criminal Victimization, 2002*. United States Bureau of Justice Statistics, Washington, DC.
- Young, L., Borzycki, M., 2008. Carjacking in Australia: recording issues and future directions. *Trends Issues Crime Crim. Justice* 351, 1–6.

Collision Avoidance Systems, Airplanes

Ivan Ostroumov, Nataliia Kuzmenko, National Aviation University, Kyiv, Ukraine

© 2021 Elsevier Ltd. All rights reserved.

Introduction	164
Mid-Air Collision Avoidance	164
Terrain Collision Awareness	168
Biographies	171
See Also	171
Relevant Websites	171
References	172
Further Reading	172

Introduction

An airplane operation in space is always performed under the risk of collision. An airplane is a dynamic object that usually moves and maneuvers with high speed. Equipment failures or airplane deviation from preplanned trajectory can lead to a collision with an object or other airspace users (Ostroumov and Kuzmenko, 2018). Initially, airplane flights are planned with collision-free trajectories within the network of flight routes. Air traffic management is performed under Air Traffic Controller (ATC) operation and usage of specific flight leveling. Flight levels, measured from mean sea level (MSL), guarantee safe separation between airplanes within one flight route. ATC clears specific flight level for each airspace user and is responsible for nonconflict planning of air traffic within predefined airspace volume. Pilots need to follow all ATC instructions; therefore, a mid-air collision may be the result of a human factor of an ATC or a pilot.

Ground and artificial constructions are other important threats to airplane operations. During take-off or landing phases, an airplane is located at low altitude. Thus, buildings, high-altitude artificial constructions, electrical wires, or trees may be a danger for operation in case of poor visibility. An airplane is equipped with a terrain awareness system in order to increase situational awareness of a pilot about dangerous obstacles. Nowadays, terrain awareness may be supported by the ground proximity warning system (GPWS), enhanced ground proximity warning system (EGPWS), or terrain awareness and warning system (TAWS).

Mid-Air Collision Avoidance

The constant growth of a number of airplanes in use increases the airspace demand and increases the mid-air collision risk. Traffic alert and collision avoidance systems (TCASs) are used onboard an airplane to detect and avoid near mid-air collisions between airplanes. TCAS is based on secondary surveillance radar (SSR) principle and is fully independent of any ground equipment (Livadas et al., 2000).

International regulations provide three available TCAS types, namely, TCAS I, TCAS II, and TCAS III.

TCAS I detects airspace users equipped with air traffic control radar beacon system (ATCRBS) transponders and indicates their locations to pilot in order to improve situation awareness. Basically, TCAS I provides only traffic advisory (TA) for nearby airspace users. TCAS I is widely used in general aviation and is mandated for usage onboard all aircraft with more than 10 and less than 31 passenger seats.

TCAS II provides TA and supports near mid-air collision detection and avoidance by coordinated maneuver in a vertical plane guided by resolution advisory (RA). TCAS II is mandated in European airspace for a commercial airplane with more than 19 passengers (or 30 passengers in the airspace of the United States) or a maximum take-off weight exceeding 5700 kg (or 15,000 kg in the United States).

TCAS III should support TA and RA in both vertical and horizontal planes. However, due to technical limitation of equipment, mostly in pure accuracy of bearing measurement, TCAS III is still under the development stage.

The TCAS family is based on independent surveillance function (Fig. 1). TCAS surveillance equipment generates an interrogation signal at 1030 MHz similar to secondary surveillance radar interrogation and emits them via two antennas placed on the top and bottom parts of an airplane. Both antennas support surveillance in different volumes: above and below an airplane to cover the whole airspace around them. Interrogation signals will be received by an ATCRBS antenna after a time of radio waves propagation in space. ATCRBS transponder automatically generates and transmits a reply signal at 1090 MHz after fixed 3.0 μ s delay. Reply signal in mode "S" of ATCRBS contains a digital message that includes a unique airplane code (24 bits of data), a barometrical altitude of an airplane (MSL) and some other technical data. A reply will be received by a

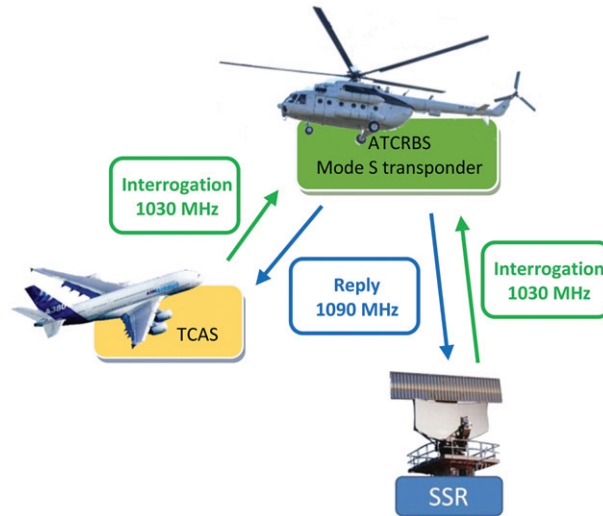


Figure 1 Surveillance function of TCAS.

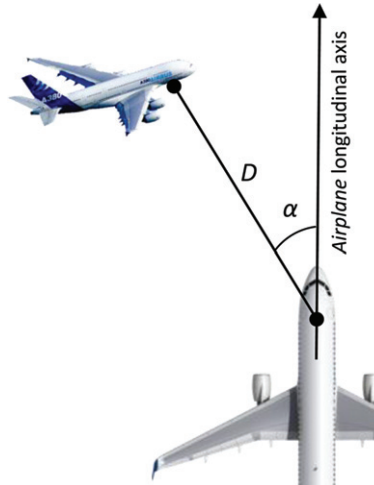


Figure 2 TCAS polar coordinate system.

directional antenna of TCAS. The time of the radio signal propagation makes it possible to calculate an exact distance to another airspace user:

$$D = 0.5c(t - t_{\text{ATCRBS}})$$

where $t_{\text{ATCRBS}} = 3.0 \mu\text{s}$ is a delay in ATCRBS transponder; t is a time between transmission interrogation signal and receiving a reply; $c = 3 \times 10^8 \text{ m/s}$ is a speed of electromagnetic wave propagation.

The directional antenna of TCAS measures the bearing of a reply signal. The bearing is the angle between an airplane longitudinal axis and direction to the ATCRBS transponder in the horizontal plane measured clockwise. Distance and bearing localize ATCRBS transponder in a polar coordinate system with a reference point at TCAS antenna locations (Fig. 2). The relative altitude is calculated as the difference between barometrical altitude from Air Data System and data of barometrical altitude from the digital message of the ATCRBS transponder. Also, surveillance equipment supports observation one time per second.

Unfortunately, the bearing measurement is not sufficiently accurate for horizontal collision avoidance. A common error of bearing measurement in TCAS does not exceed ± 5 degree, but in some cases, the configuration may exceed ± 30 degree. Low accuracy of bearing measurement makes realization of collision avoidance in a horizontal plane impossible and holds TCAS III equipment under the research stage.

TCAS II uses air traffic data for collision detection and avoidance. TCAS II algorithm tracks each airspace user-collecting airplane coordinates with the same identification code and extrapolates their trajectories (Fig. 3). TCAS II estimates closest points of approach (CPA) for both predicted trajectories and detects horizontal and vertical distances between CPA A and B. Obtained

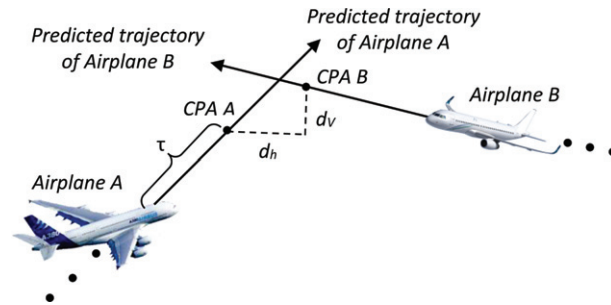


Figure 3 Conflict detection in TCAS II.

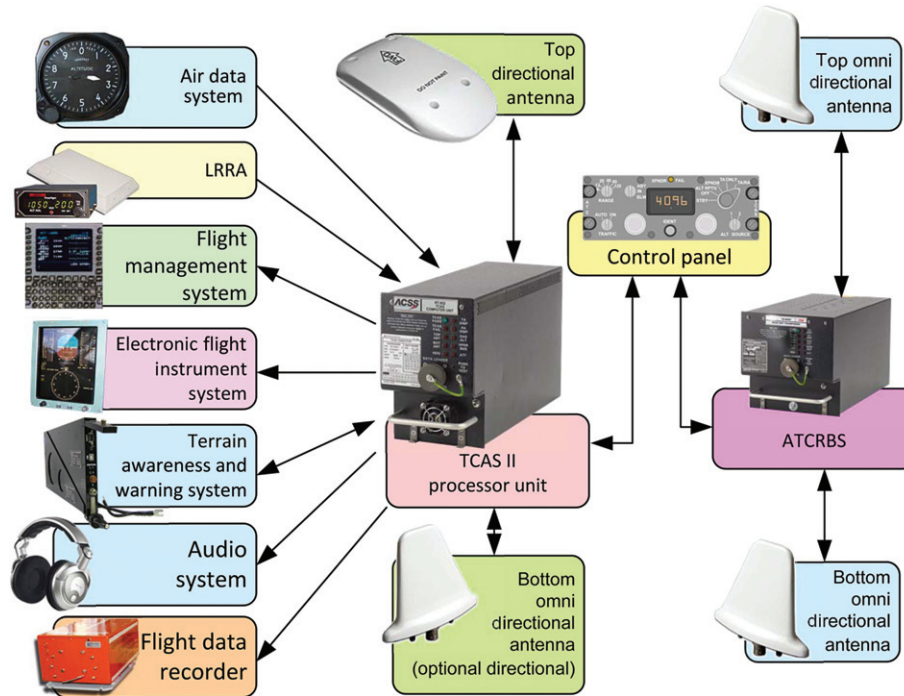


Figure 4 Structure of TCAS II.

distances are compared with the required separation minimums between airplanes. In the case of separation being reduced toward or below the required minimum, TCAS logic detects risk of mid-air collision and estimates time of flight to CPA (τ) (Rand and Eby, 2004; Munoz et al., 2013). Due to the difference in airspace structure and separation minimums, basically, there are two commercially available versions of TCAS II: version 6.04 (developed for US implementation) and version 7.1 (for European airspace). TCAS II generates two safety areas according to τ scale: Caution area within 20–48 s and Warning area 15–35 s (exact time depends on the sensitive level or airplane altitude). Location of CPA in caution area generates TA and location in warning area generates RA.

Basically, TCAS II includes a processor unit that is connected to two directional antennas (bottom antenna may be replaced with omnidirectional antenna) (Fig. 4). In common case, TCAS is connected to ATCRBS transponder of mode S and has a control panel in the cockpit. An ATCRBS transponder has two omnidirectional antennas. As input data, TCAS II uses barometric altitude and airspeed from an air data system, absolute altitude from low range radio altimeter (LRRA), heading from attitude and heading reference system, and some data from TAWS that blocks TCAS logic in case of conflict with a ground surface. Data from TCAS will be indicated on an electronic flight instrument system (EFIS) and is used in a flight management system (FMS). Aural annunciations can be shared with the audio system of the airplane via the audio control panel. Also, some TCAS data is stored in the flight data recorder.

TCAS data can be indicated by EFIS on a modern airplane cockpit design or via modified vertical speed indicator and traffic display (VSI/TDI), or radar display in an old-fashioned cockpit design. Mostly, all modern large airplanes are equipped with EFIS that includes primary flight (PFD), navigation (ND), and system displays (SD) (Fig. 5). Traffic data is indicated on ND. RA is

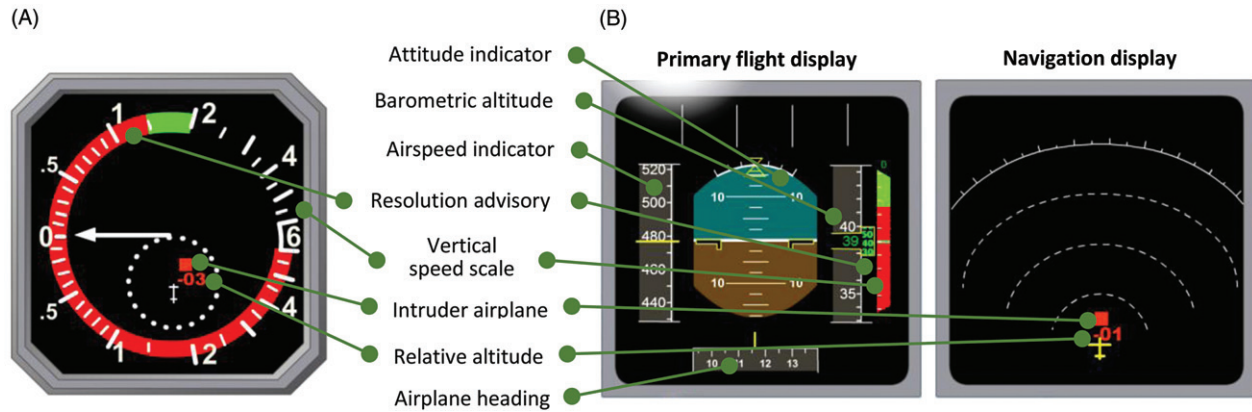


Figure 5 TCAS II indication. (A) Vertical speed indicator and traffic display (VSI/TDI). (B) Electronic Flight Instrument System.

indicated on the scale of vertical speed indicator on PFD. Air traffic is depicted on PFD using four types of symbols depending on threat value as follows:

- An unfilled diamond indicates not dangerous traffic. Predicted trajectories of these airplanes are not crossing, or crossing, but horizontal and vertical separation minimums at CPAs are more than enough for safe clearance, or airplanes are at the long distance.
- A filled diamond depicts nonconflict or proximate traffic that is located close to an airplane (within a cylinder with radius 6 NM and height ± 1200 ft. from current airplane altitude).
- A filled yellow circle shows intruder airplane. Predicted trajectories of airplanes are crossing and horizontal and vertical separations between CPAs are less than required for current airspace structure, but time of flight to CPA is within a caution area.
- A filled red square depicts intruder that caused RA and time of flight to CPA is within the warning area.

Each air traffic symbol is accompanied with relative altitude number between an airplane and airspace user in hundreds of feet above (+) or below (–) an airplane. Down or up arrow indicates a vertical speed trend for climbing or descending airplanes with the speed more than 500 ft./min.

If airspace user is detected as an intruder and time of flight to CPA is within caution area, an airplane will be depicted as a yellow circle with simultaneous aural advisory generation for pilots “Traffic, traffic”. This advisory means that we have reduced separation minimum in CPA with the depicted airplane, time of flight to CPA is less than 20–48 s and in 10–13 s RA will be generated. At this scenario, pilots need to be prepared for a possible RA and can try to make visual contact with the intruder. However, no maneuvers shall be done.

In case the CPA is located within the warning area, TCAS II will initiate a conflict avoidance algorithm that choose collision avoidance maneuver and estimates minimum required vertical speed. After approving of collision avoidance maneuver by TCAS of intruder airplane, pilots of both conflict airplanes will receive coordinated RA. RA is indicated in the form of red and green areas on the scale of vertical speed indicator. The red area indicates prohibited and green—desired vertical speed. Simultaneously with RA, the intruder is depicted as a red square symbol. Aural warnings are connected with a particular vertical speed required in RA (Table 1).

In the most cases “Climb, climb” or “Descend, descend” are initial aural alerts in TCAS II that require a pilot to initiate climbing or descending with a vertical speed not less than 1500 ft. (450 m)/min. Collision avoidance algorithm in TCAS II uses a

Table 1 Aural alerts in RA of TCAS II version 7.1

Upward sense		Downward sense	
Required vertical speed (ft./min)	Aural warning	Required vertical speed (ft./min)	Aural warning
1500	Climb, climb	–1500	Descend, descend
1500	Climb, crossing climb; Climb, crossing climb	–1500	Descend, crossing descend; Descend, crossing descend
2500	Increase climb, increase climb	–2500	Increase descent, increase descent
1500	Climb, climb NOW; Climb, climb NOW	–1500	Descend, descend NOW; Descend, descend NOW
0	Level off, level off	0	Level off, level off
No change	Monitor vertical speed	No change	Monitor vertical speed
From 1500 to 4400	Maintain vertical speed, maintain	From –1500 to –4400	Maintain vertical speed, maintain
From 1500 to 4400	Maintain vertical speed, crossing maintain	From –1500 to –4400	Maintain vertical speed, crossing maintain

cylindrical model of an airplane with a CPA in the center. In case of crossing cylindrical models in CPA, TCAS puts the word “crossing” in aural alert to indicate a threat of possible mid-air collision. If the speed of 1500 ft./min is insufficient for safe clearance in CPA, TCAS II generates RA for speed increasing up to 2500 ft. (760 m)/min, which is accompanied with aural alerts “Increase climb, increase climb” or “Increase descent, increase descent”. During conflict avoidance, TCAS may change the direction of maneuver that will be indicated in issued aural warning “Climb, climb NOW; Climb, climb NOW” or “Descend, descend NOW; Descend, descend NOW” with a minimum required vertical speed of 1500 ft./min. Warning “Level off, level off” is issued when safety separation in CPA is reached and the airplane has to stop changing vertical speed and has to start a horizontal flight. Warning “Monitor vertical speed” does not require changing vertical speed, but a pilot needs to look into vertical speed indicator and to locate some prohibited range on the scale. Alert “Maintain vertical speed, maintain” usually is connected with a wide range of possible values for vertical speed from 1500 ft. (450 m) up to 4400 ft. (1340 m)/min. Bigger values inform about closeness of location to CPA and dangerous conflict states. After successful conflict resolution, TCAS generates aural alert “Clear of conflict” and the pilot can return to the previous cleared flight level.

According to regulative documents, pilots have to respond to initial RA within 5 s and for subsequent RA within 2.5 s. After the first RA generation, pilots need to report to the ATC about TCAS II operation. The airplane will be out of ATC control within the time of RA up to aural alert “Clear of conflict.” A maneuver contrary to the RA is strongly prohibited.

At low absolute altitude, some RAs are inhibited in order to avoid airplane conflict with a ground surface:

- Alert “Increase descend” is inhibited below 1550 ft. (470 m).
- Alert “Descend, descend” is inhibited below 1100 ft. (335 m).
- All RA aural alerts are inhibited below 1000 ft. (305 m).
- All TCAS II aural alerts are inhibited below 500 ft. (150m).

Nowadays, TCAS II is mandatory onboard equipment that plays the role of “last” safety insurance system in case of a mid-air collision. Numerous disadvantages of current TCAS stimulate the development of TCAS technology. The collision detection algorithm in TCAS is grounded only on predicted trajectories and a set of geometrical tests that generate numerous false alarms in case of maneuvering airplanes. Poor accuracy of surveillance data and a relative indication of air traffic can lead to misunderstanding of conflict geometry by pilots.

Support of RA onboard some airplanes has been integrated into the autopilot system. For example, Airbus autopilot/flight director supports the airplane to automatically follow the RA issued by TCAS II. Also, an airplane may be equipped with TCAS alert prevention algorithms that have been introduced by Airbus in order to reduce the amount of false RAs.

The concept of new ACAS X is currently under the development stage by an international team of researches (Jeannin et al., 2015). Future system will use Automatic Dependent Surveillance-Broadcast data to increase surveillance performance and decrease the workload of onboard ATCRBS transponders (Kastelein and de Haag, 2014). Collision detection and avoidance will be grounded on statistical algorithms that will decrease the number of false alarms and will be safer in mid-air collision avoidance (Gardner et al., 2016; von Essen and Giannakopoulou, 2016). Also, ACAS X will be compatible with the free routes concept and will support integrated remotely piloted aircraft systems in controlled airspace.

Terrain Collision Awareness

Airplane flight at low altitude is always connected with a threat of collision with terrain. Modern airplanes are equipped with some kind of TAWS in order to reduce the risk of collision with terrain. TAWS improves situational awareness of pilots providing an aural and visual representation of potential dangerous relief and danger descend scenes of the airplane (Mozdzanowska et al., 2008).

Historically, the terrain collision awareness function has been implemented in a ground proximity warning system (GPWS). GPWS was mandated by the Federal aviation administration (FAA) since 1974 for all airplanes flying in the US airspace. GPWS uses data from various sensors onboard an airplane to detect dangerous descent of the airplane. Typical GPWS uses altitude above ground level (AGL) measured by low range radio altimeter (LRR), barometric altitude and descent rate from an air data system to detect near-accident conditions. Accident detection algorithm in GPWS compares current values of sensor measurements with allowed boundary levels. Commonly, GPWS compares two parameters simultaneously. Thus, the near-accident state is defined as a closed region (Fig. 7). Deviation of measured values to dangerous area initiates visual and aural warning to pilot about danger state of flight. Pilots should take action to move some parameters out of warning area in order to avoid collision with a ground. Also, GPWS uses landing gear status and flaps location to detect the phase of airplane flight and to reduce false alarms in case of landing. In addition, GPWS uses data from the instrument landing system (ILS) to detect significant deviation from glideslope during airplane landing. GPWS helps to decrease the number of incidents in case of controlled flight into terrain. However, GPWS has a serious limitation, because LRR cannot detect terrain or obstacles ahead of the airplane. GPWS warning is grounded on current AGL only. The solution to this problem has been found in 1990 using digital computing equipment onboard of the airplane.

Enhanced ground proximity warning system (EGPWS) includes all GPWS functionalities together with forward-looking terrain avoidance (FLTA) function (Theunissen et al., 2005; Zhang et al., 2013). Nowadays, EGPWS can be used as a synonym of TAWS. The internal memory of TAWS includes worldwide terrain, obstacles, and airport digital databases. TAWS uses airplane location from global navigation satellite system (GNSS) (or attitude and heading reference system, in case of GNSS lock) together with altitude and ground speed data to predict airplane movement (look ahead) and compares airplane flight path with available relief model and

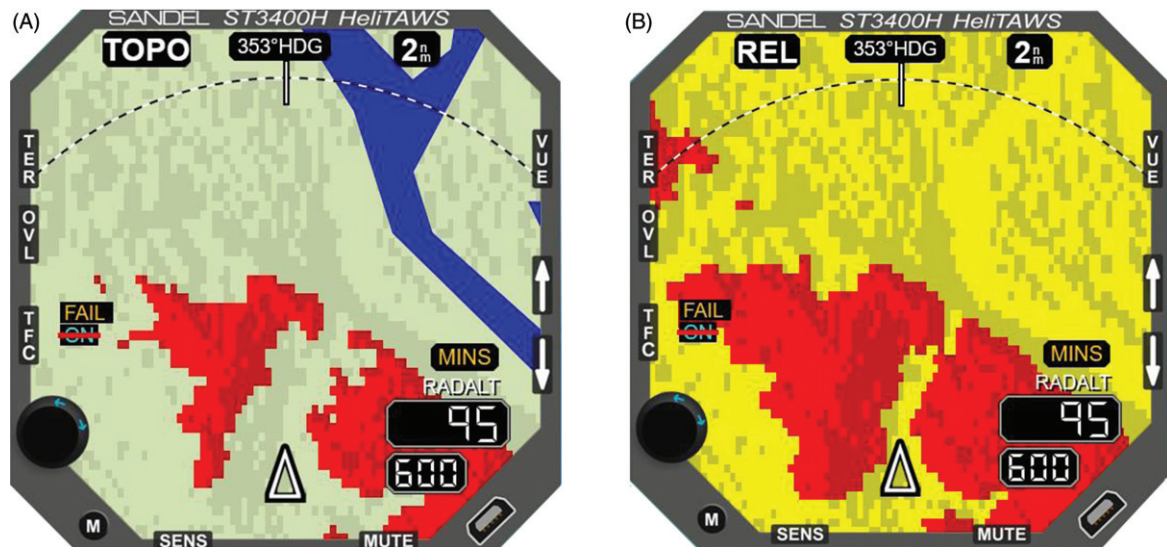


Figure 6 Topographic and relative altitudes modes of terrain indication in ST3400 TAWS. (A) Topographic mode of terrain indication. (B) Relative altitudes mode of terrain indication.

database of obstacles. In case of potential threat, TAWS provides sufficient warnings for pilots. In addition, a dangerous relief is indicated to the pilot to increase situational awareness. TAWS relief indication can be provided in two basic modes: topographic and relative altitudes. Both of these modes support the pilot with fast access to visual terrain data that clearly indicates an airplane's relation to the ground. In topographic mode, the terrain is displayed in green chart colors, except terrain above the current airplane altitude that is highlighted with red color. Relative altitude mode indicates terrain data in relative altitude to an airplane location. Relief progressively close to an airplane altitude is depicted as green, yellow, and red color table (Fig. 6).

International regulation of TAWS contains a detailed description of seven basic modes of operation:

Mode 1. Excessive rates of descent: At these modes, TAWS compares AGL from LRRA and excessive descending barometric altitude rate from air data system with predefined boundary curves (Fig. 7). In case of excessive descent rate, TAWS generates aural warning "Sink rate" or "Pull Up."

Mode 2. Excessive closure rate to the terrain: TAWS compares AGL rate from LRRA and terrain closure rate calculated by the change of altitude. In case of excessive close rate, TAWS generates "Terrain, Terrain" and "Pull Up" aural warning. Mode 2 is divided into two sub-modes, according to flight phase that is detected by flaps location: Mode 2A for en-route and approach and Mode 2B for landing. Each mode has a specific curve shape for warnings generation.

Mode 3. Negative climb rate or altitude loss after take-off: In this mode, TAWS alerts about altitude lost immediately after take-off. Alarming is grounded on comparison of AGL from LRRA and altitude loss. Value of altitude loss is calculated by integration of barometric altitude or corrected altitude from GNSS. If altitude loss is detected, caution alert "Don't sink" is generated.

Mode 4. Flight into the terrain when not in landing configuration: In this mode, TAWS compares AGL and airspeed with landing configuration of landing gear (sub-mode 4A) and flaps (sub-mode 4B). If AGL is less than 500 ft. (150 m) and landing gear is not in landing configuration, Mode 4A generates caution alert "Too low gear." If $AGL \leq 245$ ft. (75 m) and flaps are not in landing configuration, Mode 4B generates caution "Too low flaps." At the same configuration, but with high airspeed both sub-modes issue "Too low terrain" caution.

Mode 5. Excessive downward deviation from a glide path: This mode is active only in case of landing by ILS. TAWS analyses AGL and deviation below glideslope trajectory. In case of $AGL \leq 1000$ ft. (305 m) and deviation more than 1.5 dots, TAWS generates alert "Glideslope" in a soft manner, in case of $AGL \leq 300$ ft. and deviation more than 2 dots, it issues a hard alert.

Mode 6. Advisories and altitude callouts: During the landing phase or operation at low altitude, all pilots' attention is directed into instruments or setting of visual contact with runway and terrain around. Thus, voice callout of specific AGL altitude crossing improves vertical awareness of pilot, due to the terrain below of airplane. The most useful altitude callouts are 1000, 500, 300, 200, 100, 50, 40, 30, 20, and 10 ft. AGL. Some modifications of TAWS can include alarming of activation altitude of LRRA in 2500 ft. (760 m) in the form of "Radio altimeter" or "Twenty five hundred" voice callout. This message indicates that LRRA is active and AGL is available in the system. During the landing of the airplane, pilots need to exactly understand a moment of decision height crossing in the final approach procedure. The point of decision height is preselected by a pilot at relative height scale of the airplane (barometric or AGL) to runway altitude at which pilot has to make a decision about landing or fault landing and climbs in case of getting a second chance of approach. The fault of landing may be a result of pure visibility, bad weather conditions, windshear, excessive rate of descent, or other factors that cannot guaranty successful airplane landing. In Mode 6, TAWS generates voice alert "Decision height" or "Minimums," in case of crossing predefined decision height. Also, Mode 6 can include alarming about

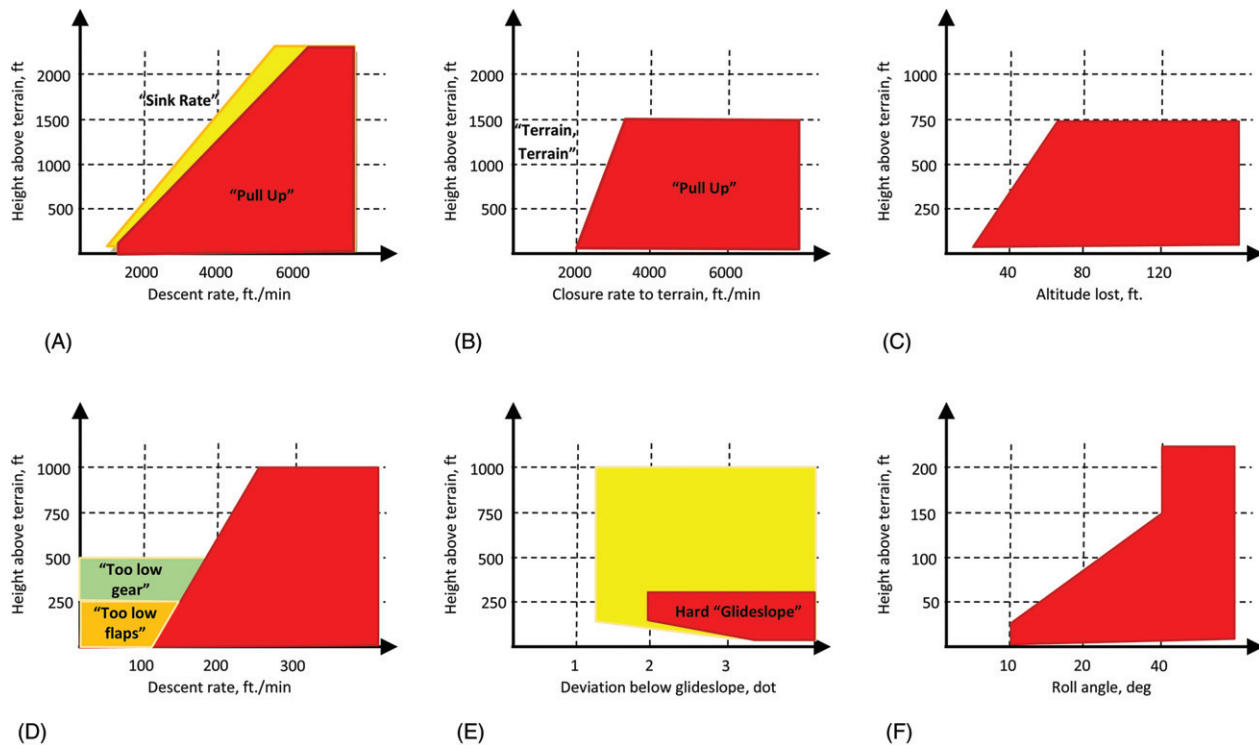


Figure 7 TAWS modes. (A) Mode 1. Excessive rates of descent. (B) Mode 2. Excessive closure rate to the terrain. (C) Mode 3. Negative climb rate or altitude loss after take-off. (D) Mode 4. Flight into the terrain when not in landing configuration. (E) Mode 5. Excessive downward deviation from a glide path. (F) Mode 6. Excessive roll angle at low AGL altitude.

overbanking conditions, in case of significant airplane maneuvering with excessive roll angle at low AGL altitude, mostly below 150 ft. (45 m) with a roll angle more than ± 10 degree.

Mode 7. Windshear detection: Excessive windshear energy may cause a big threat to airplane operation at low altitude. Windshear is a sudden change of wind direction and velocity in different planes. The appearance of wind shear at low altitude degrades the performance of the airplane and may result in sudden altitude drop. Thus, an alarm indicating high magnitude of windshear is important for pilots during landing or for airplane operation at low altitude. Presence and magnitude of windshear can be detected by various algorithms that use different input data, such as wind, air, true, and vertical speeds; angle of attack; airplane acceleration components, and other on-board sensors data. Some sensors can use specific equipment to sense the speed of air particles ahead of the airplane. Mode 7 generates aural alert "Windshear, Windshear, Windshear," in case of windshear risk detection.

Regulative documents specify various curve shapes for different modes depending on TAWS type and implementation. Also, multiple TAWS aural alert can be prohibited in general settings of TAWS in order to follow airline best practice.

FLTA algorithm of TAWS uses airplane location and speed vector to extrapolate trajectory. The extrapolated trajectory is checked with terrain and obstacle models for a potential collision. Basically, the extrapolated airplane path is based on airplane location (latitude and longitude), heading angle, and ground speed. Altitude extrapolation is grounded on current airplane altitude from GNSS and vertical speed. Caution alert "Caution, Terrain" is generated in case of conflict detection between extrapolated airplane trajectory and terrain/obstacle elevation in 20 s ahead of the airplane. If the pilot does not take appropriate action, a warning alert "Terrain, Terrain. Pull up" will be issued in 10 s ahead of the conflict.

The premature descent alert (PDA) function of TAWS compares current airplane location with a predefined three-dimensional approach trajectory in order to detect an excessive airplane deviation with further generation of aural caution "Too Low Terrain."

Regulative documents specify two basic classes of TAWS:

- Class A TAWS must provide an indication of terrain on a display system and supports Modes 1–6.
- Class B TAWS does not require data display but must support modes 1, 3, and 6 only.

Classes A and B provide FLTA and PDA functions.

Nowadays TAWS is present in the equipment list of many airplanes, but the most important role of terrain collision awareness is in helicopter implementation. Helicopter missions are mostly connected with operation at low altitude, take off, and landing from the unequipped area, hovering, operation within natural, or artificial elements (Anderson et al., 2011). For example, in most emergency operations provided by a helicopter, pilot must fly to a place where he has never been before and lands in a place that is

not intended for landing. In such kind of missions, lack of terrain data or artificial environment plays an important role in successfully mission completion.

Basically, collision avoidance on board of airplane is grounded on system assistance in a generation of a set of advisories for pilots, requiring tacking actions from pilots in order to avoid mid-air collision or conflict with the terrain. Therefore, the human factor still plays an important role in the safety of aviation.

Biographies



Ivan Ostroumov has been a faculty of Air Navigation Systems Department of the National Aviation University of Ukraine since September 2007. He obtained his PhD degree of Engineering in Navigation and Traffic Control in 2009 from National Aviation University of Ukraine. Since then, he has been a research scientist and associated professor for National Aviation University. Since 2016, he has also served as navigation instructor at “Aviation Company Ukrainian Helicopters”. In 2017/2018, he was a Fulbright scholar in the school of Aeronautics and Astronautics at Purdue University, United States. He has also participated in several international projects, including Supporting SESAR on GNSS Vulnerability Assessment by performing Space Weather Analysis (Navigation department, EUROCONTROL, Brussel), and E-learning course development (Institute of Air Navigation Services, EUROCONTROL, Luxembourg). His research theme is advanced methods for Alternative Positioning, Navigation, and Timing. Current research projects include Methods and Algorithms of positioning by multiple navigational aids, Availability and Accuracy estimation of navigation.



Nataliia Kuzmenko is a senior researcher of the National Aviation University of Ukraine. She obtained her PhD degree of Engineering in Navigation and Traffic Control in 2017 from the National Aviation University of Ukraine. Nataliia is a certified aviation security instructor by ICAO (ASTP/Basic, ASTP/Instructors). She has had a traineeship in Tool Development for support to CAA/NSA at Regulatory Division within the Directorate Single Sky (Eurocontrol, Brussels). Current research projects include aviation safety, collision detection, and avoidance, artificial intelligence, Remotely Piloted Aerial Systems, video stream object detection and recognition, kernel density estimation, and neural networks.

Relevant Websites

Eurocontrol Airborne Collision Avoidance System home page: <https://www.eurocontrol.int/acas>.

ACAS II bulletins and safety messages: <https://www.eurocontrol.int/articles/acas-ii-bulletins-and-safety-messages>.

SKYbrary: <https://www.skybrary.aero>.

ACAS X Principles: https://www.skybrary.aero/index.php/ACAS_X.

Aviation accidents reports related to GPWS: <http://www.aviation-accidents.net/tag/gpws>.

The Next Generation Air Transportation System: <https://www.faa.gov/nextgen>.

Sandel Avionics HeliTAWS™: https://www.youtube.com/watch?time_continue=1&v=JcF5EZ8AcMU.

Avionics training: <https://www.avionics.sciary.com>.

See Also

Incident detection systems, airplanes; Aviation safety, commercial airlines; Aircraft maintenance and inspection

References

- Anderson, T., Jones, W., Beamon, K., 2011. Design and implementation of TAWS for rotary wing aircraft. In: 2011 Aerospace Conference. IEEE, pp. 1–7.
- Gardner, R.W., Genin, D., McDowell, R., Rouff, C., Saksena, A., Schmidt, A., 2016. Probabilistic model checking of the next-generation airborne collision avoidance system. In: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC). IEEE, pp. 1–10.
- Jeannin, J.B., Ghorbal, K., Kouskoulas, Y., Gardner, R., Schmidt, A., Zawadzki, E., Platzer, A., 2015. A formally verified hybrid system for the next-generation airborne collision avoidance system. In: International Conference on Tools and Algorithms for the Construction and Analysis of Systems. Springer, Berlin, Heidelberg, pp. 21–36.
- Kastelein, M., de Haag, M.U., 2014. Preliminary analysis of ADS-B performance for use in ACAS systems. In: 2014 IEEE/AIAA 33rd Digital Avionics Systems Conference (DASC). IEEE, Colorado Springs, CO, USA, p. 7D3-1.
- Livadas, C., Lygeros, J., Lynch, N.A., 2000. High-level modeling and analysis of the traffic alert and collision avoidance system (TCAS). *Proc. IEEE* 88 (7), 926–948.
- Mozdzanowska, A.L., Weibel, R.E., Hansman, R.J., 2008. Feedback model of air transportation system change: implementation challenges for aviation information systems. *Proc. IEEE* 96 (12), 1976–1991.
- Munoz, C., Narkawicz, A., Chamberlain, J., 2013. A TCAS-II resolution advisory detection algorithm. In: AIAA Guidance, Navigation, and Control (GNC) Conference. AIAA, p. 4622.
- Ostroumov, I.V., Kuzmenko, N.S., 2018. An area navigation (RNAV) system performance monitoring and alerting. In: 2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC). IEEE, pp. 1–4.
- Rand, T.W., Eby, M.S., 2004. Algorithms for airborne conflict detection, prevention, and resolution. In: The 23rd Digital Avionics Systems Conference (IEEE Cat. No. 04CH37576) (Vol. 1). IEEE, pp. 521–531.
- Theunissen, E., Koeners, G.J.M., Rademaker, R.M., Jenkins, R.D., Etherington, T.J., 2005. Terrain following and terrain avoidance with synthetic vision. In: 24th Digital Avionics Systems Conference (Vol. 1). IEEE, p. 4-D.
- von Essen, C., Giannakopoulou, D., 2016. Probabilistic verification and synthesis of the next generation airborne collision avoidance system. *Int. J. Softw. Tools Technol. Transf.* 18 (2), 227–243.
- Zhang, H.M., Tuo, H.Y., Yang, C., Jing, Z.L., Li, Y.X., 2013. Forward-looking alerting threshold analysis of TAWS based on probability methods. *J. Syst. Simul.* 3, 27.

Further Reading

- Eurocontrol, 2017. ACAS Guide. Airborne Collision Avoidance.
- Federal Aviation Administration, 2011. Introduction to TCAS II Version 7.1, U.S. Department of Transportation.
- ICAO, 2007. Aeronautical Telecommunications. Annex 10 to the Convention on International Civil Aviation. Volume IV. Surveillance and Collision Avoidance Systems.
- ICAO, 2006. Airborne Collision Avoidance System (ACAS) Manual. Doc. 9863, an/461.
- Spitzer, C., 2007. Digital Avionics Handbook: Elements, Software and Functions. Avionics, 2nd ed. CRC Press, Boca Raton, FL.
- Federal Aviation Administration, 2012. Terrain Awareness and Warning System (TAWS). Technical Standard Order, TSO-C151c, Department of Transportation, Aircraft Certification Service.
- IATA, 2018. Controlled Flight Into Terrain Accident Analysis Report 2008-2017 Data.

Collision Avoidance Systems, Automobiles

Erick J. Rodriguez-Seda, United States Naval Academy, Annapolis, MD, United States

© 2021 Elsevier Ltd. All rights reserved.

A Need for Collision Avoidance Systems	173
Brief History of Collision Avoidance and Warning Systems	173
Overview and Strategies of CASs	174
Example of Current Technologies	175
ACC	175
Forward Collision Warning (FWD), Automatic Emergency Braking (AEB), and Pedestrian AEB	175
Rear Cross Traffic Alert (RCTA)	175
Blind Spot Detection (BSD)	175
Lane Departure Warning (LDW) and Lane Keeping (LK)	175
Rear Cameras and Parking Assist	176
Sensors and State-Awareness Technologies	176
Radars, Lidars, and Sonars	176
Cameras	176
GPS	176
V2V and V2X Communication	177
Challenges	177
Sensor Faults and Errors	177
Human Drivers Versus Machines	177
Human Drivers and the Effect of Automation	177
System Integration	178
Cybersecurity	178
Public Perception, Ethical Challenges, and Legal Responsibility	178
Future Developments and Considerations	178
References	179
Further Reading	179

A Need for Collision Avoidance Systems

According to the United States National Highway Traffic Safety Administration, 37,133 people died in 2017 in the United States alone due to automobile accidents (US Department of Transportation, 2018a). Worldwide estimates exceed 1.35 million fatalities each year and, based on projections by the World Health Organization, road accidents will become the fifth major cause of death by 2030 if the current trend continues. These figures stress the importance of developing and enforcing active measures to minimize car crashes. Automation and, specifically, the use of collision avoidance (World Health Organization, 2018) and warning systems (CAWSs) can help prevent road accidents. Recent studies have shown that over 90% of all vehicle crash fatalities are at least partly attributed to human-related factors, such as alcohol drinking, distraction, fatigue, and speeding (US Department of Transportation, 2018a). CAWSs can lift some of the responsibility (if not all) from the human driver and improve road safety by taking control of the automobile's actions when a collision is about to happen. Even non-autonomous measures such as pre-collision warnings, in which the driver is alerted about an impending crash event with no other action, have shown to significantly decrease the incidence and severity of accidents.

Brief History of Collision Avoidance and Warning Systems

CAWSs can be classified according to their level of autonomy. Non-autonomous systems, also known as collision warning systems (CWSs), are those that do not take control of the vehicle's driving functions but rather just alert the driver about potential collision threats via audio, visual, or haptic cues. Some CWSs can even suggest corrective measures to mitigate or avoid a crash. The first reported passenger automobile designed with a smart CWS was the Cadillac Cyclone XP-74, developed in 1959 by General Motors and seen in Fig. 1. The vehicle was equipped with two radar-based sensors near the headlights that would alert the driver of the proximity of frontal obstacles. The Cadillac Cyclone never saw mass-production and the concept of providing cars with such safety features had to wait until the early 1990s with the introduction of *Distance Warning* by Mitsubishi Debonair in Japan. This luxury car utilized a lidar to warn drivers about close-distance vehicles ahead. Other luxury cars, primarily in Japan and Europe, developed

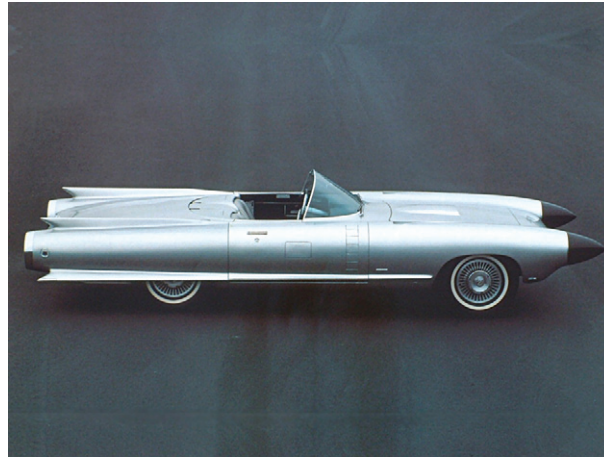


Figure 1 First reported passenger automobile with a collision warning system. Creative commons licensed (CC-BY-SA-2.0) flickr photo by mashleymorgan: <https://www.flickr.com/photos/mashleymorgan/3866256075/>.

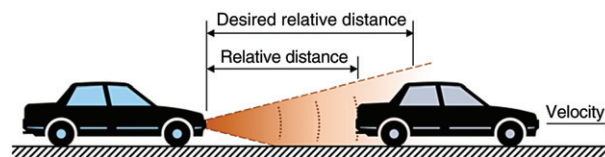


Figure 2 Schematic representation of an ACC system.

similar safety features in the late 1990s and early 2000s. Nowadays, most modern vehicles provide customers with collision warning systems based on one or multiple proximity sensors such as cameras, lasers, and radars.

Autonomous and semi-autonomous on-board CAWSs are those that can either assist or take command of the automobile's braking or steering functions without human intervention. Herein, these systems will be simply referred to as collision avoidance systems (CASs) to distinguish them from CWSs. Up to the 1980s, most of the research on automobile safety focused on passive measures to protect the driver and passengers in the case of a crash such as crumple zones, seat belts, and airbags (Akamatsu et al., 2013). Studies at the time consistently showed a decrease in collision-related fatalities but an increase in road accidents. Recognizing that human factors were the leading cause of road related accidents, private and state initiatives in Europe (particularly, the Programme for European Traffic of Highest Efficiency and Unprecedented Safety, PROMETEUS), the United States, and Japan, stimulated the development of automation in transportation systems. The introduction of CASs to the market for civilian use started with the development of the adaptive cruise control (ACC) system in the late 1990s (Vahidi and Eskandarian, 2003). As illustrated in Fig. 2, ACC systems can detect the relative distance and velocity of vehicles ahead and self-regulate the distance between a vehicle and the vehicle in front. Early ACC models required the driver to set the car into ACC mode and, therefore, the functionality was not always in place during the duration of the drive. The first mass-produced automobile for sale to constantly use active forward collision avoidance was the Volvo XC60 in 2008. Volvo XC60 utilized a CAS named *City Safety* that monitored the distance of vehicles and pedestrian ahead and could automatically stop the vehicle and avoid hard collisions when traveling under 30 km/h. Nowadays, a wide variety of vehicles do not only provide forward CASs but can also assist the vehicle when in risk of rear and lateral collisions. Moreover, several companies are currently developing and testing vehicles capable of driving safely from start to end with minimal human supervision.

It should be noted that mobile robotics research paved the way for the development of CAWSs. Several universities, scientific institutes, and research governmental agencies, such as NASA, developed and tested collision avoidance algorithms and obstacle detection methods for unmanned vehicles, primarily, in the 1970s and 1980s (Rodríguez-Seda et al., 2014). However, the implementation of these methods in the automobile industry had to wait for the robotics technologies to mature. In contrast to early studies with unmanned vehicles, automobiles need to operate in highly uncertain and dynamic environments that require a higher level of precision, robustness, and safety not viable at the time.

Overview and Strategies of CASs

Operation of CASs depends on the vehicle's level of automation (according to SAE established levels) and capabilities (US Department of Transportation, 2018a). Most CASs utilize one or multiple proximity and localization sensors such as global

positioning systems (GPSs), lidars, short and long-range radars, and cameras. They monitor traffic as well the presence of nearby vehicles, pedestrians, animals, and objects and send their information to decision making units distributed around the car. The on-board decision making units, also known as electronic control units (ECUs), process the information from multiple sensors to make predictions about potential collisions and to suggest corrective actions. In lower automation level vehicles, the CAS may start by alerting the driver. If the driver does not respond, the system may take an autonomous action such as braking or steering. In more high-level cars, such as SAE Level 3, the decision to regulate the car and avert a collision can be made without human intervention.

The methods and policies employed by the CAS may be classified as cooperative or non-cooperative. A cooperative strategy is one in which all nearby vehicles and obstacles abide to the same set of safety rules and share the common goal of avoiding a crash. The rules may include a particular protocol to deal with intersections (which vehicle can continue and which vehicle needs to stop), merging or changing lanes, and reducing the velocity or accelerating the vehicle while platooning, among many others. Due to the diversity of automobile brands and models on today's roads and the need for high level of coordination among vehicles, practically all of today's collision safety systems assume a non-cooperative strategy.

Similarly, autonomous collision avoidance strategies can be classified as planned or reactive. A planned strategy is one in which the desired avoidance trajectory of the vehicle is computed in advance. It requires prior knowledge of the road and some level of coordination among vehicles and other obstacles in order to guarantee the safest course of action. Reactive strategies, on the other hand, update the vehicle's actions on-line, based on current information about the vehicle's environment. Reactive measures are in general more robust and reliable but, due to the lack of planning, their reaction to a collision threat may be unexpected or abrupt. Today's autonomous systems are typically a combination of both strategies.

Example of Current Technologies

The following technologies are examples of current CAWSs widely available in today's passenger cars. System names are based on the US department of transportation's definitions ([US Department of Transportation, 2018b](#)) and can vary among automakers.

ACC

An ACC system is one that regulates the velocity of the car to a desired constant speed but that can also adjust the velocity based on the distance of vehicles ahead. By using proximity sensors, it aims to keep a safe distance between the vehicles and avoid forward collisions ([Fig. 2](#)).

Forward Collision Warning (FWD), Automatic Emergency Braking (AEB), and Pedestrian AEB

A FCW system is one that monitors the speed of the vehicle, the speed of any vehicle or large obstacle ahead, and the distance between both using proximity sensors. Based on the distance and speeds of both cars, the system warns the driver about a potential collision, typically by the use of visual and audio cues. If the driver does not respond and the vehicle is equipped with an AEB system, the car may automatically stop or reduce the velocity. Pedestrian AEB systems work in similar fashion by constantly looking for the presence of people crossing in front of the vehicle. It then warns the driver or automatically applies the brake, depending on the scenario, in order to avoid an accident.

Rear Cross Traffic Alert (RCTA)

RCTA systems monitor the presence of upcoming traffic and, sometimes pedestrian and cyclist, from the side as the vehicle reverses. It warns the driver about a potential collision.

Blind Spot Detection (BSD)

By using radars or side-view cameras, BSD systems aim to detect and alert the driver about the presence of vehicles in the driver's blind spots. If the driver tries to make a turn toward one of these blind spots, the system may send a stronger alert or even impede the turn. This concept is illustrated in [Fig. 3](#).

Lane Departure Warning (LDW) and Lane Keeping (LK)

LDW systems typically track the lane markings on the road using a camera and send an audio, visual, or haptic alert to the driver if the vehicle starts drifting out of its lane. A vehicle equipped with a LK system utilizes the LDW alert to prevent the vehicle from drifting away either by steering or accelerating the wheels. Both systems aim to avoid accidents with vehicles in other lanes as well as rollovers when leaving the road.

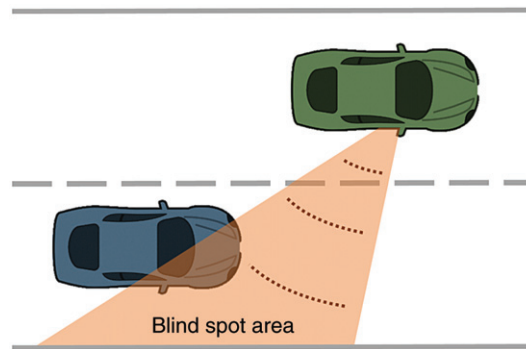


Figure 3 Blind spot detection system.

Rear Cameras and Parking Assist

Rear-view or backup cameras are typically employed in parking maneuvers and provide visual feedback to the driver when the car is in reverse mode. They also provide both visual and audio cues that alert the driver of the proximity with obstacles including people and pets.

Sensors and State-Awareness Technologies

For a CAWS to make a prediction or decision, it needs to know the state of the car (velocity and position) and detect the presence of other vehicles and obstacles. To this end, it requires the use of on-board sensors, monitoring devices, or communication systems that can provide feedback to the driver or the car. The following is a list of common sensing and communication technologies in the automobile industry.

Radars, Lidars, and Sonars

Radars, lidars, and sonars are among the most common sensors used in CAWSs (Mukhtar et al., 2015). They are examples of active sensors that emit energy in order to compute distance. Radar, which is also known as *Radio Detection and Ranging*, emits electromagnetic waves. When the electromagnetic wave encounters an object it bounces back to the sensor. The time it takes the signal to travel back and forth and the Doppler shift experienced by the signal are then used to compute the relative distance and velocity of the other object. Similar to radars, a lidar or *Light Detection and Ranging*, emits pulses in various directions. When the pulses reach an object, they bounce back to the sensor. Differences in return times and wavelengths can be then used to estimate distances as well as 2D and even 3D representations of objects. Finally, active sonar, also known as *Sound Navigation Ranging*, creates sound waves. When the waves hit an object and bounce back to the sensor, their traveling time is used to estimate the object's relative distance.

Radars and lidars, along with cameras, are the most popular sensors for collision avoidance systems and when combined with multiple units or with actuators, they can provide a 360° degree view of the surrounding environment. In general, radars are typically better for longer distances and under low visibility conditions, whereas lidars are better in detecting smaller obstacles and in providing an accurate representation of the surrounding objects. Sonars tend to have longer delays and to be more sensitive to noise. Therefore, they are mostly used for parking assist systems and slow velocities maneuvers.

Cameras

Cameras are passive optical sensors mostly used to track and identify objects surrounding the vehicle. One or multiple cameras, more commonly of stereotype, take constant images of the environment. The images are then processed by computer vision algorithms that can not only estimate the distance but also can identify and classify the objects (e.g., pedestrians, vehicles, and trees). Current research trends in machine learning and computing are helping cameras to identify and classify objects with human-like abilities. To operate at night, infrared cameras are sometimes used.

GPS

Many of today's automobiles are equipped with GPS units. GPS units require at least three radio signals coming from three different GPS satellites. These signals carry the time and position information of the satellites, which then the GPS unit uses to estimate the location of the vehicle using triangulation. Today's systems require at least four satellite signals to provide accurate position information, which have error ranges in the few meters depending on the unit, number of satellite signals, location, and noise interference.

GPSs can only provide localization for the own vehicle. Therefore, it is mostly used for navigation systems and can only provide collision information in well-known environments as well as velocity estimation. They also require a fairly open sky, which limits their usability to low rise building areas and open roads (no tunnels). However, with the development of vehicle-to-vehicle (V2V) communication and vehicle-to-environment (V2X) communication, GPSs can assist CAWSs by sharing position information among vehicles.

V2V and V2X Communication

V2V communication refers to the wireless exchange of information, such as position, trajectory, and velocity, among vehicles. V2X, also known as vehicle-to-infrastructure (V2I), refers to the communication between vehicle and surrounding infrastructure. Common wireless communication networks include radio, internet, and cellular networks.

V2V and V2X applications can increase cooperation among vehicles and can inform the driver or the CAWS about road, traffic, and weather conditions. Intelligent highways and stationary smart sensors on the roadside can help coordinate vehicular flow such as in platooning and lane merging and increase traffic capacity and efficiency. The US department of transportation estimates the use of thousands of infrastructure V2X devices already installed in the US roads and a recent study from March 2019 projects nearly six million V2X-equipped vehicles worldwide by 2022 ([US Department of Transportation, 2018a](#)). Similarly, the European Union plans to require speed-limiting technology in all new car models starting in 2022. The cars will use GPS signals and V2X communication to identify speed limits and automatically regulate the vehicles' velocity.

Challenges

Despite the current capabilities and achievements in crash prevention, CAWSs face some unique challenges ranging from technological obstacles to societal concerns. Some of these challenges are currently evolving at the same pace as smart vehicles are hitting the road.

Sensor Faults and Errors

CAWSs need to be reliable under diverse operating conditions. They need to safely operate during daylight as well as dark, under fog and low visibility, snow, and rain. Typically, such issues are addressed by redundancy (i.e., the use of multiple sensors of different modalities) and by estimation of algorithms that can provide state awareness despite noisy or limited information. These measures and advances, however, can increase the cost of the vehicle, which can only be justified and financially sustained by appropriate market demand and by large public and private investment. In addition, even state-of-the-art sensor technologies can occasionally fail and health monitoring, security, and emergency systems must be in place to recognize and recover from such state. CASs need to take into consideration the effects of sensing errors and faults without being too conservative in order to preserve efficiency ([Rodríguez-Seda et al., 2016](#)).

Human Drivers Versus Machines

Computers and automated systems present several advantages over human drivers. First, automated systems can identify and respond to unexpected threats quicker than what a human driver would. Second, CAWSs can remain in alert mode during the entire duration of the trip without being subject to fatigue or distraction. On the other hand, human drivers are typically better in making decisions under unexpected and unplanned events and may be able to classify objects better than what a computer vision or lidar system can. However, the later advantages from the human driver are only a matter of time as current advances in computer vision and computing, specifically machine learning, have preliminary shown unprecedented performance that can challenge and overpass human capabilities.

Human Drivers and the Effect of Automation

Deployment of autonomous cars requires public acceptance and trust. When it comes to transportation systems, this typically involves keeping a human driver in the loop, which can supervise and override computer commands. Moreover, human drivers may still be required as a safety measure in case of system failures. Unfortunately, the reliance on automation may conversely yield human drivers less adept to safely respond to rare events. Take, for example, the aviation industry, which has pioneered the way for self-driving transportation systems. Accidents like the Air France Flight 447 in 2009, where pilots were unable to recognize and correct a dangerous state after the plane's autopilot disengaged, have shown how automation can limit the ability of pilots to respond to unexpected and rare events. Similarly, more recent events such as the Lion Air Flight 610 in 2018 and the Ethiopian Airlines Flight 302 in 2019 highlighted the need for pilots to better understand the operation of autonomous mechanisms as well as the transition between autonomous and non-autonomous behavior. Other examples of over reliance on automation more specific to the automobile industry include the Tesla Model S fatal crash in Williston, Florida in 2016, where the autopilot system failed to recognize a tractor-trailer and drove the car into it without braking at all, and the Uber self-driving Volvo CX90, which killed a

bicyclist in Tempe, Arizona in 2018. In both cases, the drivers were determined to be distracted and were not taking proper responsibility in controlling the car. Overall, if human drivers are expected to supervise or complement autonomous mechanisms, then standards for the training and regular assessment of human driver abilities must be developed and enforced.

System Integration

In addition to a potential loss of cognitive abilities due to over reliance on automation, studies have shown that human drivers may become immune or ignore some of the warning signs of CAWSs (Akamatsu et al., 2013). Specifically, there must be different alerts according to the severity of the threat, a reduction on the false alarm rate, and coordination among other types of warning signs. Specifically, CAWSs are just one part of a larger concept known as driver assistance systems (DASs). DASs integrate a wide range of automated or smart functions that aim to ease the driver workload and improve efficiency by increasing state-awareness or taking control of some driving functions. The latter means that CAWSs will be integrated with other smart functions, which may generate their own alerts. Coordination among multiple subsystems must be in place to avoid driver's confusion and reduce mental workload.

In addition, the driver must comprehend how the vehicle's safety systems work and, when operating in an autonomous mode, the vehicle must react in a way that either emulates the driver's natural or reasonable behavior. That is, the driver may lose confidence in the CAWS if the vehicle is perceived as behaving irrationally or in unexpected ways.

Cybersecurity

The use of smart systems and wireless communication technologies can present automakers as well as the government with a new set of challenges. Researchers at different universities have shown how the increasing connectivity and integration of automobiles can make today's vehicle vulnerable to remote interference (Petit and Shladover, 2014). Intruders and criminals can easily get access to the vehicle's main controller-area-network (CAN) Bus via any of the dozens (and sometimes, hundreds) of on-board ECUs and disrupt the operation, override control commands, and disable safety systems. Even tampering with or denying the system with critical sensor data, such as speedometers, can have unsafe consequence. Automakers need to put in place measures to prevent, identify, and contain the effect of cyber attacks. Today's automobiles are examples of cyber-physical systems and require safety measures atypical of standard information technology systems. Vehicles need to be constantly monitored for cyber intrusions and control units and software need to be frequently updated to keep up with current standards. Passwords need to be customized and regularly updated for each vehicle and ECU. Vehicles need to be equipped with fault detection systems that can also identify the tampering of data and with safety systems protocols that can contain the effect of an attack.

Public Perception, Ethical Challenges, and Legal Responsibility

Successful integration of CAWSs and other autonomous features on cars will highly depend on how the public and the government embrace the use of autonomy. In contrast to the aviation industry and mass-transit systems where the driver (or the lack of a driver) is hidden from the passengers' view, the lack of human control is quite evident and present to the passenger inside the car. The environment is also different and not as predictable as for planes or trains: cars do not have dedicated lanes and have to share the roads with thousands of other vehicles and objects of different sizes and shapes, including motorcycles, animals, and trucks. Any accident involving automation tends to receive significant media attention, which in turn, can increase public concern and set back the implementation of CASs despite their safety record. Examples such as the Tesla crash in 2016 and the Uber incident in 2018 delayed the testing and release of self-driving functions for both companies.

Similarly, the use of CAWSs raises several ethical and legal questions that are yet to be addressed. Whose safety does the system should prioritize under a potential collision threat? There might be instances where the system may need to choose among the driver, other drivers, passengers, cyclists, and pedestrians. Then, there is the challenge of responsibility and liability. In the case of an accident, is the driver responsible? Or is the vehicle's manufacturer, the autopilot's software, or the company that designed the faulty sensors is the one responsible? What kind of consequences and penalties should a company or a driver receive and how insurance companies will assess responsibility? Current state and federal laws were enacted with a driver in mind and, therefore, need to be adapted to properly regulate the operation of autonomous and semi-autonomous cars. As of 2019, there are still no federal strict regulations and uniform safety tests in place for the development and operation of autonomous cars in the United States.

Future Developments and Considerations

There is a very little doubt about the positive impact that automation can have on transportation safety. As an example, one can take a look at how automation revolutionized the aviation industry by making air transportation safer (Boyd, 2017). As for the automobile industry, it has been shown in several studies that CAWSs can reduce road accidents or mitigate the severity of a crash. For instance, a recent study in 2017 demonstrated that FCW, AEB, and FWC with AEB systems can reduce rear-end striking crashes by 27%, 43%, and 50%, respectively (Cicchino, 2017). Yet, there is unarguably room for improvement. The final goal is to have fully autonomous vehicles capable of driving with better precision, safety, and efficiency than human drivers under all scenarios and circumstances. To

this end, several automakers and research institutes are working toward more robust and reliable systems, which includes newer, better, and cheaper sensing technologies as well as more precise estimation and detection algorithms. Another research trust toward safer vehicle transportation is the integration of V2V and V2X communications and the development of intelligent highways that can better support autonomous systems and enable cooperation among vehicles. Cooperation among vehicles and the infrastructure enables not only efficiency but can also increase safety by coordinating routes and informing vehicles about hazard conditions on the road. Finally, research into evasive collision avoidance must continue. Current systems command the car to stop or reduce the velocity under a crash threat. Under some circumstance, a CAS should be able to safely steer away a car and avoid a collision without the need to stop.

In the meantime, while autonomous systems share control command with a human driver (even as a supervisory role), the public needs to be better informed about the operation and shortcomings of CAWSs. Current CAWSs are not made to fully substitute the driver and the driver's attention, although reduced, cannot be neglected. Similarly, in order to continue the path toward a vision of zero-collisions, automakers and governmental entities need to invest and incentivize autonomous technologies. Specifically, the government must seek to enforce some basic, uniform autonomous measures of collision avoidance in all new cars. See also, Automobile Accidents and Passive Prevention Systems.

References

- Akamatsu, M., Green, P., Bengler, K., 2013. Automotive technology and human factors research: past, present, and future. *Int. J. Veh. Technol.* 1–27.
- Boyd, D.D., 2017. A review of general aviation safety (1984–2017). *Aerosp. Med. Human Perform.* 88 (7), 657–664.
- Cicchino, J.B., 2017. Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates. *Accid. Anal. Prev.* 99, 142–152.
- Mukhtar, A., Xia, L., Tang, T.B., 2015. Vehicle detection techniques for collision avoidance systems: A review. *IEEE Trans. Intel. Transp. Sys.* 16 (5), 2318–2338.
- Petit, J., Shladover, S.E., 2014. Potential cyberattacks on automated vehicles. *IEEE Trans. Intel. Transp. Sys.* 16 (2), 546–556.
- Rodríguez-Seda, E.J., Tang, C., Spong, M.W., Stipanović, D.M., 2014. Trajectory tracking with collision avoidance for nonholonomic vehicles with acceleration constraints and limited sensing. *Int. J. Rob. Res.* 33 (12), 1569–1592.
- Rodríguez-Seda, E.J., Stipanović, D.M., Spong, M.W., 2016. Guaranteed collision avoidance for autonomous systems with acceleration constraints and sensing uncertainties. *J. Optimiz. Theory App.* 168 (3), 1014–1038.
- US Department of Transportation, 2018a. Preparing for the future of transportation: automated vehicle 3.0, Technical report. U.S. DOT, Washington, DC.
- US Department of Transportation, 2018b. Vehicle shopper's guide: driver assistance technologies, Technical report. U.S. DOT, Washington, DC.
- Vahidi, A., Eskandarian, A., 2003. Research advances in intelligent collision avoidance and adaptive cruise control. *IEEE Trans. Intel. Transp. Sys.* 4 (3), 143–153.
- World Health Organization, 2018. Global status report on road safety 2018, Technical report. WHO Press, Geneva.

Further Reading

- Bengler, K., Dietmayer, K., Farber, B., et al. 2013. Three decades of driver assistance systems: review and future perspectives. *IEEE Intell. Transp. Syst. Mag.* 6(4), 6–22.
- Rodríguez-Seda, E.J., Stipanović, D.M., 2014. Guaranteed collision avoidance with discrete observations and limited actuation. Chen, Y., Li, L., (Eds.) *Advances in Intelligent Vehicles*, Academic Press, Oxford, UK, 89–110.
- Waschl, H., Kolmanovsky, I., Willems, F. (Eds.), 2019. *Lectures Notes in Control and Information Sciences 476: Control Strategies for Advanced Driver Assistance Systems and Autonomous Driving Functions*. Springer International Publishing, Cham, Switzerland.

Connected Automated Vehicles: Technologies, Developments, and Trends

Azra Habibovic, Lei Chen, RISE Research Institutes of Sweden, LindholmspirenGothenburg, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	180
Introduction	180
Automated Vehicles	180
A Brief History of Automated Driving	181
Categorization of Automated Vehicles	181
Technology Behind Automated Driving	181
Automated Driving Systems Under Development	182
Connected Vehicles	183
A Brief History of V2X	183
The Path of DSRC	183
The Rise of Cellular-Based V2X (C-V2X)	184
The Technology Debate	184
Connected and Automated Driving	184
The Basic V2X Applications	184
Advanced V2X Applications for Autonomous Driving	185
Key Benefits, Challenges and Development Trends	185
Key Benefits of CAV	185
Key Challenges and Development Trends	186
Brief Conclusions	187
Acknowledgment	187
Biography	187
References	187

Nomenclature

AV Automated vehicles
ADS Automated driving system
CV Connected vehicle
CAV Connected automated vehicles
V2X Vehicle-to-everything communication

Introduction

The demand for transport in society is governed by several factors, including population size, demographics, urbanization, and economic development. In recent years, we have seen an increase in transport needs around the world. In connection with this, increasing demands are placed on sustainable transports, increased accessibility and gender equality in the transport system.

To meet these demands, both passenger and freight transport are expected to shift toward electrification, automation, connectivity, and shared services in the coming decades, with the ultimate goal of increasing the transport efficiency, availability, and safety. The recent decades have thus witnessed a rapid evolution of both automated vehicles (AV) and connected vehicles (CV), and a gradual merge of these two paths into connected automated vehicles (CAVs).

This chapter reports first a brief introduction of AVs and CVs, their current development and main enabling technologies, followed by their integration into CAVs. While we acknowledge importance of all levels of automation, this chapter has an emphasis on the higher levels of automation.

Automated Vehicles

An automated vehicle (AV) is a vehicle in which an automated driving system (ADS) performs parts or entire dynamic driving task. Here, a dynamic driving task includes all real-time operational and tactical functions required to drive a vehicle in road traffic,

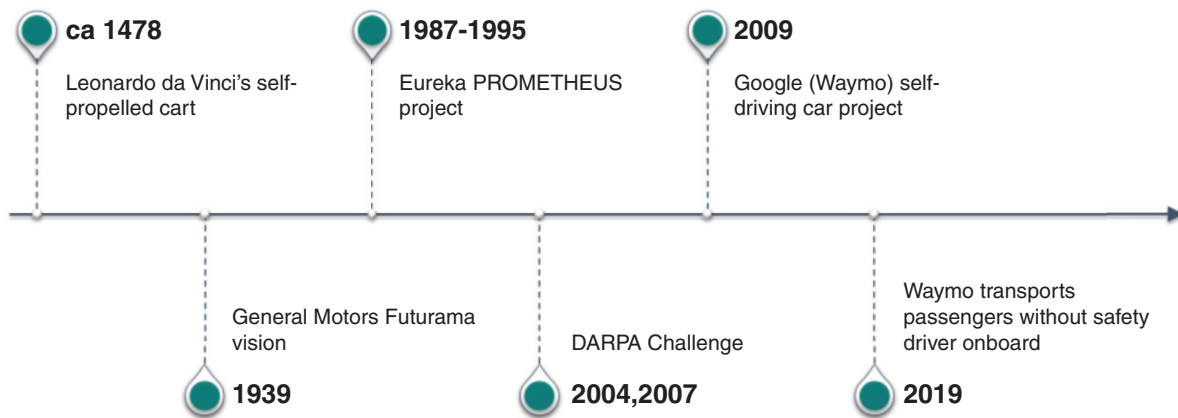


Figure 1 Major milestones in the development of automated driving.

excluding strategic functions, such as route planning. Today, various nomenclature is used for AVs, where a few of the most common ones are “autonomous,” “self-driving,” “robotic,” and “driverless” vehicles.

A Brief History of Automated Driving

Development of AVs dates back to around 1478 when Leonardo da Vinci designed a cart that could move without being pushed or pulled (Fig. 1). Centuries later, in 1939, General Motors displayed a vision of the future that included an automated highway system. In 1958, the company turned its vision into a prototype. However, it took several decades until AVs become a serious option for transportation. Automated vehicles developed within the project PROMETHEUS (1987–95) and DARPA Challenge (2004, 2007) are seen as key precursors of the AVs that we have under development today.

Nowadays, all major vehicle manufacturers are developing AVs, either on their own or in cooperation with some other stakeholders. Large technology companies, such as Alphabet (owner of Waymo), Baidu, and Yandex, are also among key stakeholders. Notably, several start-up companies (e.g., Argo AI, Cruise Automation, Aurora) have entered the scene and are often viewed as major drivers of innovation. In addition, providers of key enabling technologies for sensing and processing technologies (e.g., Intel/Mobileye, NVIDIA) play a crucial role.

Categorization of Automated Vehicles

To better define and categorize ADS, several organizations have proposed classification scales. The scale proposed by the Society of Automotive Engineers (SAE) in the SAE J3016 standard (SAE International, 2018) is today the most common. It should, however, be noted that all attempts to classify vehicles into different levels of automation have shortcomings. For example, a vehicle can contain ADS of different automation levels that are used at different occasions. Consequently, it becomes difficult to say that the vehicle belongs to a certain level of automation. A common “workaround” is to specify the operational design domain (ODD) and refer to the conditions under which the ADS operates.

The SAE scale incorporates six levels, ranging from Level 0 (fully manual) to Level 5 (fully automated), see Table 1. Largely simplified, Levels 0–3 mean that a human driver exists in the vehicle and either drives (possibly with the support of the ADS), or is prepared to take over the driving task when the system requests it. Automated vehicles belonging to the lower levels of automation (up to Level 3) are found in several types of mass-produced vehicles today (e.g., Adaptive Cruise Control (ACC), Lane Departure Warning (LDW), Lane Keeping Support (LKS), Blind Spot Detection (BSD)). Level 3 automation is currently under development and is a widely debated topic due to its requirement that the human drivers are ultimately responsible for the driving at the same time as the ADS takes care of the driving and allows them to focus on other activities.

In Levels 4 and 5, an ADS drives the vehicle and the vehicle can also handle the situation when it is not possible to drive automatically. This means that no human driver is needed at those levels. However, a vehicle with automated functions at the higher levels may also be able to drive manually (dual functions). The difference between Levels 4 and 5 is that Level 4 vehicles can only drive in certain traffic situations or in certain areas, while Level 5 vehicles can handle all situations and environments that a physical driver can handle. In practice, Level 4 is referred to as highly automated driving, while Level 5 is fully automated driving.

Technology Behind Automated Driving

Automated driving is enabled by advanced software, powerful data processing units, detailed maps, and an array of sensors that together create a 360-degree living map of the traffic environment around the vehicle. Most stakeholders are relying on a similar combination of lidars, radars, cameras, and ultrasonic sensors, with a few notable exceptions (e.g., Tesla relies on a combination of

Table 1 Automation levels as defined by the SAE standard J3016.

SAE J3016	Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
What does the human in the driver's seat have to do?	You are driving whenever these support features are engaged—even if your feet are off the pedals and you are not steering You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety			You are not driving when these automated driving features are engaged—even if you are seated in “the driver's seat” When the feature request, you must drive These automated driving features will not require you to take over driving		
What do these features do?	These features are limited to provide warnings and momentary assistance	These features provide steering OR brake/acceleration support to the driver	These features provide steering AND brake/acceleration support to the driver	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met	This feature can drive the vehicle under all conditions	

Source: Based on (SAE International, 2019).

cameras and ultrasonic sensors only). To accommodate a high level of robustness and redundancy, stakeholders are commonly using several sensors (e.g., Waymo's 5th generation of automated driving system, Waymo Driver, incorporates 5 lidars, 4 radars, and 29 cameras (Jeyachandran, 2020)). Using multiple sensors, however, poses integration challenges for system manufacturers. Consequently, several of them have chosen to place a “sensor suite” on the roof of the vehicle, which has by now become a “signature” for automated driving. The high number of sensors makes ADS rather expensive (between \$50,000 and \$250,000 (Herger, 2020)). The expectation is, however, that the cost will decrease significantly with mass-production of these systems.

On the software side, machine learning (ML) and artificial intelligence (AI) are key enablers for automated driving. These algorithms, however, rely on “learning” from reality and stakeholders are currently investing large efforts on pilots on public roads to expose the algorithms to as many traffic situations as possible. Most stakeholders have also created virtual environments (simulations) and methods that enable them to expose the software to many more situations in a cost-efficient way. The ability to capture “edge cases” is seen as a major prerequisite for safe operation of AVs (Koopman and Wagner, 2017). There is no doubt that proving safety of ADS is one of the most difficult tasks (Kalra and Paddock, 2016).

Currently, most of AVs that are piloted still require a safety driver in the vehicle to warrant safety, due to a combination of ADS limitations and regulatory constraints. To eliminate the safety driver, stakeholders are considering teleoperations technology and human remote control of the vehicles if necessary. Ideally, a remote operator would be able to control and monitor several vehicles at the time, and thereby improve cost-efficiency. With this demand, several new stakeholders have emerged offering remote control technology and services (e.g., Phantom Auto, Voysys).

Automated Driving Systems Under Development

It is possible to distinguish four broad types of ADS (corresponding to Level 3 and 4) that are currently under development for use on public roads and that are expected to become mainstream within 5–10 years. In spite of the optimistic announcements made by some stakeholders a few years ago, most forecasts agree that launch of these ADS is likely to occur at the earliest in 2022/2023.

- **Automated shuttles and taxis** addressing first- and last-mile trips are expected to revolutionize passenger mobility. They are mainly piloted in urban areas, but there are also a few examples of pilots in rural settings (e.g., in Japan and Sweden). The pilots commonly involve a few vehicles and take place in well-defined areas, at low speeds and with a safety driver onboard. In 2018, both automated shuttles and taxis were commercialized for the first time; the shuttles were integrated into regular public transport in a residential area outside Stockholm (Sweden), while Waymo started offering commercial taxi services to select customers in Phoenix. Recently, a few stakeholders got permission to pilot their services without a safety driver onboard.
- **Automated goods delivery vehicles** address the first- and last-mile urban delivery, typically from a transportation hub to another hub or the final delivery destination. A few different types of such vehicles are under development including sidewalk delivery robots, automated delivery pods and automated cars. The latter are being developed by, for example, Waymo, Ford, and General Motors, and are largely similar to conventional delivery vehicles. Sidewalk delivery robots are created by, for example, Starship Technologies, Kiwi, and Amazon. Typically, they are electric, carry one parcel at the time, travel on sidewalks at speeds up to 5–10 km/h. Electric automated delivery pods are being developed by, for example, Nuro and Neolix. They are somewhat larger than delivery robots, can deliver several parcels at the time, travel at a slightly higher speed (10–20 km/h) and use both local roads and cyclist paths. Larger delivery pods are also being developed by, for example, Einride. While some of the automated delivery

vehicles are commercially deployed, they are typically operating in non-crowded, pre-mapped areas, and require remote monitoring. Also, these activities involve a limited number of vehicles.

- **Automated truck platooning on highways** has been under development for several years and is anticipated to improve efficiency of freight transport in terms of reduced fuel consumption (4% for the platoon leader, 10% for the followers) and road utilization, among others. It is a cooperative intelligent transport system (C-ITS) application that enables automated vehicles to drive close together with short inter-vehicular distances as road trains (i.e., automated vehicle platoons) via “connected braking” between trucks enabled by vehicle-to-vehicle communication. Current commercially available platooning systems are designed to operate only on multi-lane, divided, limited access highways and involve two vehicles only with a driver in each vehicle. In the future, the expectation is that the platoon size will be increased and that only the first vehicle in the platoon will require a human driver. Key stakeholders in the area include Volvo, MAN, Scania, Peloton, and Locomotion.
- **Automated driving on highways** is under development both for passenger and freight vehicles. Such ADS would co-exist with manual controls in vehicles and would allow human drivers to remove their hands from the steering wheel and/or under certain conditions take their eyes off from the road for longer periods of time. A key distinguishing characteristic is whether the driver must remain alert and ready to take control if the system is unable to execute the driving task (Level 3), or not (Level 4). To start with, such systems are expected to be available only under certain conditions and on certain highways.

Connected Vehicles

Connected vehicles (CV) represent the paradigm where vehicles are connected and are able to “talk” with each other, pedestrians, cyclists, and infrastructure. Technologies are multi-fold including vehicle-to-vehicle (V2V), vehicle-to-pedestrian (V2P), vehicle-to-infrastructure (V2I), and vehicle-to-network (V2N). With V2V and V2P, vehicles are able to share real-time information with each other and with vulnerable road users (VRUs) such as pedestrians and cyclists, which helps to improve situational awareness and traffic safety. With V2I and V2N, vehicles become part of the cooperative intelligent transport systems (C-ITS), where vehicles share real-time information through road and telecom infrastructure to improve traffic safety, efficiency, and comfort of traveling. Vehicle-to-everything (V2X), as a general term, covers all potential areas enabled by connected vehicles.

A Brief History of V2X

There are mainly two technology streams for V2X, the Dedicated Short Range Communications (DSRC) based on Wi-Fi technology, and the C-V2X based on cellular communications. The following part gives a brief discussion on each technology and its evolution, and the on-going debate on the technology choice.

The Path of DSRC

The research on V2X can be dated back to the 1990s when Wi-Fi based DSRC started to be tested for vehicular communications. In 1990, the US Federal Communications Commission (FCC) allocated a total of 75 MHz spectrum at the 5.9 GHz band for DSRC and adopted basic operation rules. The DSRC radio access technologies, i.e., IEEE 802.11p, are based on the IEEE 802.11a Wi-Fi standard with modifications to fit the characteristics of the vehicular communication environments. It was integrated into the IEEE 802.11 standard in 2012 and became the main physical communication specifications for vehicular communications. It is designed to support relative velocities up to 200 km/h, a communication range of up to 1000 m and a response time of around 100 ms. Based on 802.11p, major V2X standards have been built including the IEEE Wireless Access in Vehicle Environments (WAVE) in the US, the ISO Communications Access for Land Mobiles (CALM), and the EU C-ITS.

The US WAVE standards are specified in the IEEE 1609 series. They define a set of protocols, services, and interfaces for secure V2X communications. The physical communication is based on IEEE 802.11p, while the network and transport layer is based on the WAVE Short Message Protocol (WSMP). A Basic Safety Message (BSM) is defined following the SAE standard J2735, which is broadcast with a frequency around 10 Hz to improve the situational awareness. The ISO standards CALM are published through the ISO 21217 series and defines wide supports of a range of access technologies including cellular, Wireless Local Area Networks (WLAN), infrared communications, and millimeter wave communications.

In the EU C-ITS framework, the V2X standards consider both DSRC and cellular, while a significant effort has been devoted to ITS-G5, which is the EU’s DSRC technology. Similar to WAVE, ITS-G5 is based on IEEE 802.11p and works at the 5.9 GHz band, which has been in place since 2008. In 2014, the first release of C-ITS standard, that is, the C-ITS R1 was finalized and released by the EU standardization organizations CEN and ETSI (Chen and Englund, 2014). Corresponding to WSMP, the EU standards introduce the Basic Transport Protocol (BTP) and Geonetworking protocol for fast distribution of short messages with geo-location awareness. Two major messages, the Cooperative Awareness Message (CAM) and the Decentralized Environmental Notification Message (DENM) for improving the situational awareness are defined. Similar to BSM, CAM is a heartbeat single-hop message that is broadcast at a frequency of 1 Hz to 10 Hz with basic vehicle information such as location and driving dynamics. DENM is triggered by events such as road works and is broadcast at a higher frequency, for example, 10 Hz and may need multi-hop delivery to cover a geographical area. In addition, the C-ITS R1 includes a basic set of applications (BSA) that are prioritized for realistic implementations. Following the release of R1, CEN, and ETSI continues the work on the second release, which is expected to support more complex and advanced applications.

The IEEE 802.11p standard was designed 2 decades ago. To leverage the advancement of Wi-Fi technologies and to provide an evolution path for IEEE 802.11p, in 2018, the IEEE 802.11 Next Generation V2X study Group was formed and later in 2019 the IEEE 802.11bd Task Group was created. The IEEE 802.11bd represents the next generation DSRC, which aims at providing improved performance and is designed to be backward compatible and can co-exist with 802.11p.

The Rise of Cellular-Based V2X (C-V2X)

Despite that the DSRC has emerged in 1990s and enormous efforts on research and development have been investigated, the realistic implementation has been slow. The rapid evolution of cellular communications, especially the 5th generation (5G) technologies, provides another solution for V2X, that is, C-V2X. In the telecom sector, the 3rd Generation Partnership Project (3GPP) has started the V2X standardization since 3GPP Release 14 (R14), aka the 4G LTE standard. The first C-V2X standard is LTE-V2X, which was included in 3GPP R14 and supports the basic V2X applications such as those defined for DSRC. V2V direct communication is supported by extending the proximity services (ProSe) that was designed for device-to-device (D2D) communications, which was introduced in 3GPP R12. Depending on the spectrum availability, direct communications can be deployed at dedicated V2X spectrum such as the 5.9 GHz. V2I and V2N is supported through the Uu interface which is the radio interface between a mobile device and the access networks. In 2016, major players in the telecom and automotive sector created the 5G Automotive Association (5GAA), with the aim to promote C-V2X. Global connected vehicle trials based on LTE-V2X also started. In 2018, 3GPP finalized the Release 15 with enhanced range and reliability on LTE-V2X. In the meanwhile, studies on C-V2X with the 5G New Radio (NR) was also started. NR-V2X represents the future evolution of C-V2X to support advanced V2X services such as for autonomous driving. However, it needs to be mentioned that NR-V2X is not designed to be backward compatible, meaning a dual-mode of LTE-V2X and NR-V2X is needed to accommodate both the technologies.

The Technology Debate

There has been a long debate regarding the most suitable technologies for vehicular communications. For many years, DSRC has been the only choice for V2V. With multiple commercial products being tested, the technology is ready to be implemented. In 2016, the US National Highway Traffic Safety Administration (NHTSA) released rulemaking processes that would require new light vehicles to include DSRC-based V2X communications. It remains a significant rulemaking and decisions are yet to be made. Similarly in the EU, the EC adopted C-ITS rules in the form of Delegated Act (C-ITS DA) in 2019 with the aim to start large-scale implementation of C-ITS as of 2020. The C-ITS DA is built on many years tests and pilots using ITS-G5, and has received oppositions from the EU member countries that prefer a more technical neutral solution. It was suggested that DA should be technically neutral and include the C-V2X technologies. In the meanwhile, the EU C-Roads project oversees the many on going C-ITS pilots and aims at harmonizing the European C-ITS deployment. Hybrid solutions have been considered to provide C-ITS services through both DSRC and the existing 4G cellular networks. From the industry side, DSRC based V2X has been in production in Japan since 2015 and in the US since 2017 for selected vehicle models. In 2019, Volkswagen announced that its Golf 8 will be equipped with DSRC based V2X, representing one of the biggest moves from the automotive industry.

On the other hand, driven by the telematics and infotainment services, a majority of today's new cars are already connected, and a number of V2X applications can already be supported. In addition, C-V2X evolves rapidly in the latest years with multiple trials and the gradually available C-V2X chips. 5GAA has been actively working with ETSI on C-V2X interoperability test, and also with C-Roads to define C-ITS communication profiles through the 4G cellular networks and LTE-V2X. In the US, the FCC considers to reframe the 5.9 GHz spectrum to open certain spectrum for normal Wi-Fi services and to allocate 20 MHz of the V2X spectrum to C-V2X. China is moving forward on C-V2X by issuing the permit for LTE-V2X testing in 2018. Since then, multiple test sites with C-V2X have been built with involvement of major OEMs and suppliers.

Connected and Automated Driving

Along with vehicle automation and electrification, V2X is one of the keys for improved traffic safety, efficiency and contributes to sustainable transport solutions. It supports both existing vehicles and the forthcoming AVs for connected and automated driving.

The Basic V2X Applications

The basic V2X applications are mostly to improve the awareness by broadcasting simple status data or alerts. Those are for the vehicles with lower automation (Levels 0–3) where human drivers are involved and messages need to be conveyed to the drivers for action. In the C-ITS R1, the BSA are categorized into three categories including traffic safety, traffic efficiency, and cooperative local services and global internet services. Those applications were further refined into Day 1 and Day 1.5 services through the C-ITS platform in 2016 (European Commission, 2016).

Day 1 services are expected to and should be available in the short term because of their expected societal benefits and the maturity of technology. Applications include emergency electronic brake light, emergency vehicle approaching, slow or stationary vehicle(s), traffic jam ahead warning, hazardous location notification, road works warning, weather conditions, in-vehicle signage, in-vehicle speed limits, probe vehicle data, shockwave damping, Green Light Optimal Speed Advisory (GLOSA)/Time To Green (TTG), signal violation/intersection safety, and traffic signal priority request by designated vehicles.

Day 1.5 services are those that are mature and highly desired by the market, though, for which specifications or standards might not be completely ready. Those include off street parking information, on street parking information and management, park and ride information, information on fueling and charging stations, traffic information, and smart routing, zone access control for urban areas, loading zone management, VRU protection, cooperative collision risk warning, motorcycle approaching indication, and wrong way driving.

LTE supported V2X were studied in 3GPP R14 in 2015. Those services to a large extent correspond to C-ITS Day 1 services and include forward collision warning, control loss warning, emergency vehicle warning, emergency stop, cooperative adaptive cruise control, V2I emergency stop, queue warning, road safety services, automated parking system, wrong way driving warning, V2V message transfer under mobile network operator (MNO) control, pre-crash sensing warning, V2X in areas outside network coverage, V2X road safety service via infrastructure, and V2N traffic flow.

Advanced V2X Applications for Autonomous Driving

Advanced V2X applications go beyond situational awareness and are enabled by rich sensor data sharing, collective perception, and cooperative driving (Chen and Englund, 2016). Those applications can support all types of AVs especially those with higher-level automation and intelligence. The Car to Car Communication Consortium (C2C-CC) (<https://www.car-2-car.org/>) summarizes the advanced V2X application as Day 2 and Day 3+ applications. Day 2 services require improved cooperative awareness, sensor information sharing, collective perception, and improved infrastructure support. Examples of such services are overtaking warning, advanced intersection collision warning, VRU protection, motorcycle protection, cooperative adaptive cruise control (CACC), CACC string, long-term road works warning, and special vehicle prioritization. Day 3 services require, for example, trajectory/maneuver sharing, coordination/negotiation, and VRU active advertisement. Example services include advanced CACC (e.g., lane change), target driving area reservation, traffic light info optimizations with V2I, automated GLOSA, transition of control notification, improved VRU protection, cooperative merging, cooperative lane change, cooperative overtaking, CACC (string) management, cooperative transition of control and automated GLOSA with I2V negotiation.

Within 3GPP, enhancement on LTE-V2X has been published in 3GPP R15 with a list of 25 advanced V2X use cases. In 3GPP R16 (3GPP, 2018), those services were further expanded to a list of 30 use cases with addition of some general use cases. Services involving autonomous driving include enhanced V2X (eV2X) support for vehicle platooning, information exchange within platoon, automated cooperative driving for short distance grouping, information sharing for limited automated platooning, information sharing for full automated platooning, changing driving-mode, Cooperative Collision Avoidance (CoCA), information sharing for limited automated driving, information sharing for full automated driving, emergency trajectory alignment, intersection safety information provisioning for urban driving, cooperative lane change (CLC) of automated vehicles, 3D video composition for V2X scenario, automotive sensor and state map sharing, collective perception of environment, video data sharing for automated driving, QoS aspect of vehicles platooning and QoS aspects of advanced driving.

To summarize, with the advancement of both connectivity and automation, the evolution of CAV is moving towards an integrated approach. Function and service development now consider requirements on both automation and connectivity. This tight integration has led to numerous new services that are not possible with only automation or connectivity. Also, with the rapid development of artificial intelligence, CAV gradually evolves into cooperative driving where vehicles are able to coordinate locally to solve complex scenarios, that is, cooperative and automated driving.

Key Benefits, Challenges and Development Trends

The developments of AVs and CVs have followed their own paths for years; however, they are now starting to merge into an integrated path, that is, CAV. In addition, CAV also interacts closely with electric vehicles and new-shared mobility concepts, forming together major trends of future mobility. Though many benefits are emerging with CAVs, challenges remain and call for technology advancement, business development, innovation, regulatory changes as well as holistic system of system thinking and inclusive design (Fig. 2).

Key Benefits of CAV

Driven by societal challenges and opportunity for profit, large-scale introduction of AVs is anticipated to bring many benefits to the society, as discussed in the following part.

Improved safety: On lower levels of automation, some favorable impacts on safety and environment have already been proven (HLDI, 2019; Leslie et al., 2019). By eliminating driver, CAVs with higher automation levels are anticipated to improve safety even more and eliminate accidents caused by, for example, distraction, driving under influence, and speeding.

Improved transport efficiency: Lower levels of automation have small effects on transport efficiency; mostly related to reduction in accident related congestions. CAVs with higher level of automation and connectivity will be integrated into an automated transport system, which helps reducing the environmental impact by smoothing the traffic flow (Papadoulis et al., 2019).

Greater access to mobility: Regarding the growing ageing population, CAVs hold the potential to improve their access to mobility. By enabling flexible, on-demand, and driverless transport, CAVs could improve quality of life and productivity of elderly as well as people with visual impairments (Owens et al., 2019).

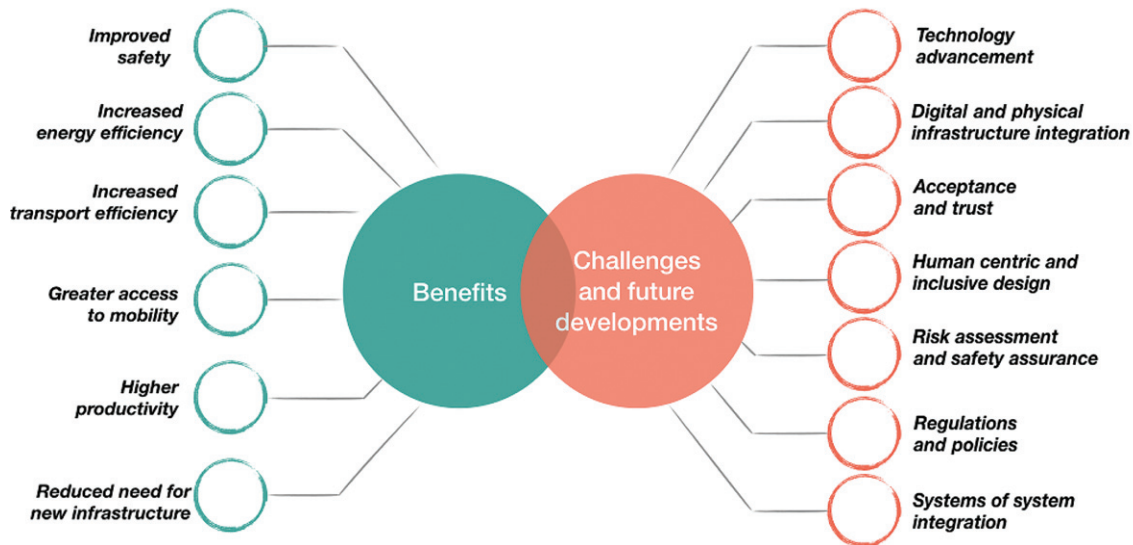


Figure 2 Key anticipated benefits and challenges of connected automated vehicles.

Higher productivity: With higher level of automation, those who have been commuting to work by car will become passengers and spend the commuting time doing something else. On the commercial side, transportation companies suffering from a serious driver shortage (e.g., mining, long-haul, and home delivery) will be able to operate their businesses with fewer drivers.

Reduced need for new infrastructure: By enabling vehicles to travel closer to each other, CAVs could increase highway capacity. Being able to be parked closer to each other, the usage of existing infrastructure can be improved and needs for new parking areas can be reduced.

Key Challenges and Development Trends

In view of the benefits, CAV continues its evolution in all aspects including technology advancement, transport system integration, social acceptance, and behavioral changes. Significant challenges remain and solutions to deal with those challenges will form the main development trends in the coming years (Litman, 2020).

Addressing technical limitations: Both automation and connectivity need improvement. For highly automated vehicles, continuous sensor technology advancement including new sensors, sensor fusion, positioning, and on-board computation remain challenging. Similarly, connectivity with high reliability and low latency need to tightly integrate with ADS for large scale testing and validation.

Digital and physical infrastructure: The digitalization is affecting all society functions including road infrastructure. It is clear that the CAV evolution is much faster and it is generally assumed that CAVs should be able to deal with existing roads without significant changes in the infrastructure. On the other hand, it is also agreed that close cooperation between CAVs and the infrastructure are key for a safe and efficient transport system. Incorporating physical and digital infrastructure with CAVs will form another major development trend.

General acceptance and trust: To reach anticipated benefits, CAVs will need to be trusted and gain societal acceptance. This means that CAVs need to safely, efficiently and seamlessly interact with other entities in the traffic system. The interaction between CAVs and conventional vehicles as well as vulnerable road users such as bicyclists and pedestrians remains an open research question. External vehicle interfaces (eHMI) have been proposed for such purposes (Habibovic et al., 2018; Dey et al., 2020), however, their role is yet to be explored. A growing interest in the area is foreseeable to improve the CAV acceptance and trust.

Human centric and inclusive design: The interaction between humans and technology is a central element of automated driving and remains challenging, and mobility solutions with CAVs need to be human centric and inclusive. While development of CAVs has attracted enormous research on human aspects from the driver and passenger perspective, it fails to address the special consideration of certain groups such as elderly and individuals with impairments. To ensure that CAVs and corresponding mobility services serve all people, stakeholders need to truly embrace a “whole journey” mindset using the universal design from early development phases (Habibovic et al., 2019).

New safety aspects: Early implementation of CAVs will be in environments where machine-based and human road users coexist. Considering unpredictability of such traffic situations, collisions might be inevitable in a foreseeable future. This may be addressed by traffic separation, such as proposed by NUMO with dedicated road network for CAVs (Chen et al., 2018). It may also be approached with in-depth understanding of behaviors and interactions. Both are challenging and call for research efforts. In addition, introducing CAVs may move safety issues from human-oriented to machine-oriented. These may include, for example, deficiencies in the software controlling the CAV, missing or incorrect map data or sensor data, and cyber-security challenges. It is then important to guarantee the CAV functions are safe and system malfunction are minimal. Currently, several relevant standards are under development (e.g., UL 4600, ISO 21448 SOTIF), and will go hand-in-hand with CAV evolution.

Regulations and policies: In recent years, it has become evident that the mainstream testing and adoption of CAVs requires significant policy and regulatory changes to govern design, construction, and performance of these vehicles. Governments around the world have thus developed, or are developing, regulatory frameworks and protocols for testing (and in some cases for deployment) of CAVs on public roads. The major challenge is that the rules vary widely between the countries (and states), making it difficult for manufacturers to navigate the regulatory landscape. Indeed, there are ongoing international discussions; however, changing international regulations is an extensive process that may take years, if not decades. Among the key challenges are risk assessment (i.e., how safe is safe enough?), type approval, liability, licensing, training, and insurance.

Transport system of systems: Transport system includes many stakeholders and components, where CAV and its corresponding stakeholders represent the major parties. With increasing connectivity for vehicles and infrastructure, emerging service innovation and collaboration, the transport system is transforming from a system of many silo systems into a system of systems (SoS) with closely interconnected and dependent systems. Introducing CAV in the system, may thus have a significant effect on the entire traffic system and its function in society. Regarding this, RISE and its partners have proposed NUMO, New Urban Mobility (<http://bit.ly/numo-urban-mobility>) (Chen et al., 2018), which is a fully electrified, connected, and automated transport system. Such an automated system requires holistic design and engineering with consideration of the technologies, humans, the society, and a continuous system evolution. It is therefore system of systems engineering (SoSE) is promising to assist the development and evolution of such a complex system (DeLaurentis, 2005). Since SoSE is also an emerging area, the application of SoSE for mobility with CAVs remains challenging.

Brief Conclusions

Recent decades have witnessed a rapid evolution of both automated vehicles and connected vehicles, and a gradual merge of these paths into connected automated vehicles (CAVs). These vehicles, in combination with other parallel trends such as artificial intelligence, electrification and shared economy, hold the key for a safer, more efficient, accessible, equal and inclusive mobility. This chapter presents a brief introduction of automated vehicles and connected vehicles, their current development, main enabling technologies, as well as the integration into CAVs. Due to the limited technological capability as well as various other reasons including trust, security, and difficulty to prove safety and acceptance in society and regulations, many challenges remain for fully reaching the benefits of CAV benefits. The chapter offers thus an overview of general development trends and the challenges to be solved.

Acknowledgment

The work was partly conducted within the projects System of Systems for Emergency Response and Urban Mobility funded by the Swedish Innovation Agency Vinnova and Trends and Market Analysis of Automated Driving (<https://omad.tech>) funded by RISE.

Biography

Azra Habibovic is senior researcher at RISE Research Institutes of Sweden and research area director for road-user behavior at the research center SAFER. She holds a PhD in Vehicle Safety Systems (2012) and an MSc in Electrical and Electronics Engineering (2006), both from Chalmers University of Technology, Sweden. Her research focuses on improving traffic safety and user experience by means of automation and connectivity.

Lei Chen is a senior researcher at the department of mobility and systems, RISE Research Institutes of Sweden. He holds a PhD on infra-informatics from Linköping University, Sweden and has many years' research experiences on mobile telecommunications, connected and automated vehicles and transport. His current research focuses are the application of information and communication technology (ICT) methods for improving traffic safety, efficiency, and sustainability.

References

- 3GPP, 2018. TR 22.886 Technical Specification Group Services and System Aspects; Study on enhancement of 3GPP Support for 5G V2X Services (Release 16).
- Chen, L., Englund, C., 2014. Cooperative ITS; EU standards to accelerate cooperative mobility. 2014 International Conference on Connected Vehicles and Expo (ICCVE). IEEE, pp. 681–686. doi: 10.1109/ICCVE.2014.7297636.
- Chen, L., Englund, C., 2016. Cooperative intersection management: a survey. IEEE Transactions on Intelligent Transportation Systems 17 (2), 570–586, doi:10.1109/TITS.2015.2471812.
- Chen, L. et al. (2018) NuMo - New Urban Mobility: New Urban Infrastructure Support for Autonomous Vehicles." RISE Research Institutes of Sweden. Available from: <http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1286741&dswid=9776> (accessed 2020-06-05)
- DeLaurentis, D., 2005. Understanding transportation as a system-of-systems design problem. 43rd AIAA Aerospace Sciences Meeting and Exhibit. AIAA.
- Dey, D. et al., 2020. Taming the eHMI jungle: a classification taxonomy to guide, compare, and assess the design principles of automated vehicles external human-machine interfaces. Transportation Research Interdisciplinary Perspectives, Submitted.
- European Commission, 2016. C - ITS Platform Final report.

- Habibovic, A., et al., 2018. Communicating intent of automated vehicles to pedestrians. *Front. Psychol.*, doi:10.3389/fpsyg.2018.01336.
- Habibovic, A., Andersson, J., Englund, C., 2019. Automated vehicles—the opportunity to create an inclusive mobility system. *Automotive World Ltd.* Available from: <https://www.automotiveworld.com/articles/automated-vehicles-the-opportunity-to-create-an-inclusive-mobility-system/>.
- Herger, M., 2020. Waymo Reveals More Details On New Sensor Hardware That Could Indicate Imminent Mass Production. *The Last Driver License Holder* . . . Available from: <https://thelastdriverlicenseholder.com/2020/03/04/waymo-reveals-more-details-on-new-sensor-hardware-that-could-indicate-imminent-mass-production/>No Title.
- HLDI, 2019., 2013-17 BMW collision avoidance features. Available from: https://www.iihs.org/media/0b76abcd-6b24-46d9-8e98-65ea962a4cb1/rNOCTg/HLDI_Research/Bulletins/hldi_bulletin_36.37.pdf.
- Jeyachandran, S., 2020. Introducing the 5th-generation Waymo Driver: Informed by experience, designed for scale, engineered to tackle more environments, Waymo. Available from: <https://blog.waymo.com/2020/03/introducing-5th-generation-waymo-driver.html>.
- Leslie, A.J. et al., 2019. Analysis of the field effectiveness of general motors production active safety and advanced headlighting systems. *Ann Arbor, MI United States.* Available from: <https://trid.trb.org/view/1654823>.
- Koopman, P., Wagner, M., 2017. Autonomous vehicle safety: an interdisciplinary challenge. *IEEE Intell. Transp. Syst. Magazine* 90–96, doi:10.1109/MITS. 2016.2583491.
- Litman, T., 2020. Autonomous Vehicle Implementation Predictions.
- Kalra, N., Paddock, S.M., 2016. Driving to Safety: How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability? RAND Corporation, Santa Monica, CA. https://www.rand.org/pubs/research_reports/RR1478.html.
- Owens, J.M., et al., 2019. Automated vehicles and vulnerable road users: envisioning a healthy, safe and equitable future. In: Meyer, G., Beiker, S. (Eds.), *Road Vehicle Automation 6*, Springer International Publishing, Cham, pp. 61–71.
- Papadoulis, A., Quddus, M., Imprialou, M., 2019. Evaluating the safety impact of connected and autonomous vehicles on motorways. *Accid. Anal. Prevent.* 124, 12–22, doi:10.1016/j.aap.2018.12.019.
- SAE International, 2018. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (J3016_201806). Available from: https://www.sae.org/standards/content/j3016_201806/.
- SAE International, 2019. SAE Standards News: J3016 automated-driving graphic update.

Construction Zones

Jalil Kianfar, Saint Louis University, St. Louis, MO, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	189
Work-Zone Temporary Traffic Control	189
Temporary Traffic Control Zone	190
Temporary Traffic Control Devices	190
Work-Zone Classification	191
Work-Zone Lane Configuration	191
Work-Zone Duration	191
Other Considerations	192
Work-Zone Safety and Mobility Impact Assessment	192
Work-Zone Safety Assessment	192
Work-Zone Mobility Impact Assessment	193
Worker Safety	194
Truck-Mounted Attenuators	194
Future Trends in Work Zones	195
References	195

Introduction

Transportation agencies regularly conduct maintenance and construction activities to preserve and rehabilitate the roadway infrastructure. These activities disrupt the normal flow of traffic on roadways and often reduce the roadway capacity, which in turn may lead to congestion and more crashes than otherwise would be expected. Transportation agencies develop work-zone traffic control plans to improve the safety of road workers and all road users and to mitigate the mobility impacts of roadwork. Work-zone traffic control aims to inform road users of the changes ahead, to guide them so that they may safely transition into the open lanes on the roadway, to ensure that road user and road workers are safely separated from each other, and then, at the end of the construction zone, to let the road users know that they have left the roadwork area.

Even though work zones occupy a relatively small fraction of a roadway system's lane miles, they are a major source of nonrecurring congestion and are considered to have higher crash rates than the parts of the roadway with no work zones. The aging of the roadway infrastructure, in countries such as the United States, means that more work zones should be expected in the next decade. Moreover, the continuous growth in the annual number of vehicle miles traveled exacerbates the safety and mobility impacts of work zones and restricts the time and locations where roadwork may be completed. This article provides a basic overview of work-zone traffic control and discusses future trends that might improve the safety of road workers and road users and mitigate the mobility impacts of roadwork. The content of this article is not intended to serve as a work-zone design and planning guide. Appropriate national and local guidelines, manuals, and standards should be followed when work zones are planned and implemented in the field.

Work-Zone Temporary Traffic Control

Work-zone temporary traffic control (TTC) aims to protect the safety of workers in the work zone while ensuring that road users move safely and efficiently through the work area. TTC encompasses seven fundamental principles that should be considered during the design phase of a roadway project and while a work zone is in place.

First, the needs of road users, including pedestrians, cyclists, and persons with disabilities, should be addressed in the work zone. The safety of road users and road workers should also be considered during the project planning, design, and construction phase, and efforts should be taken to minimize congestion because of the work zone. Second, abrupt and unnecessary changes to the travel patterns of vehicles, pedestrians, and cyclists should be avoided. Third, TTC in the work zone should be consistent with principles of positive guidance. Road users should be given adequate advanced warning about the work area and should be guided through the work zone through the use of TTC devices, such as cones and signs. Fourth, TTC devices at the work zone should be continually inspected, maintained, and modified through the life of the project. Fifth, roadside design safety should be maintained during the project, and a TTC plan should include an area for the recovery of errant vehicles. Sixth, the personnel involved in the design and implementation of the roadwork should have received appropriate training. Seventh, information related to the work zone should be communicated in advance to the public, business community, and other stakeholders (Federal Highway Administration, 2009).

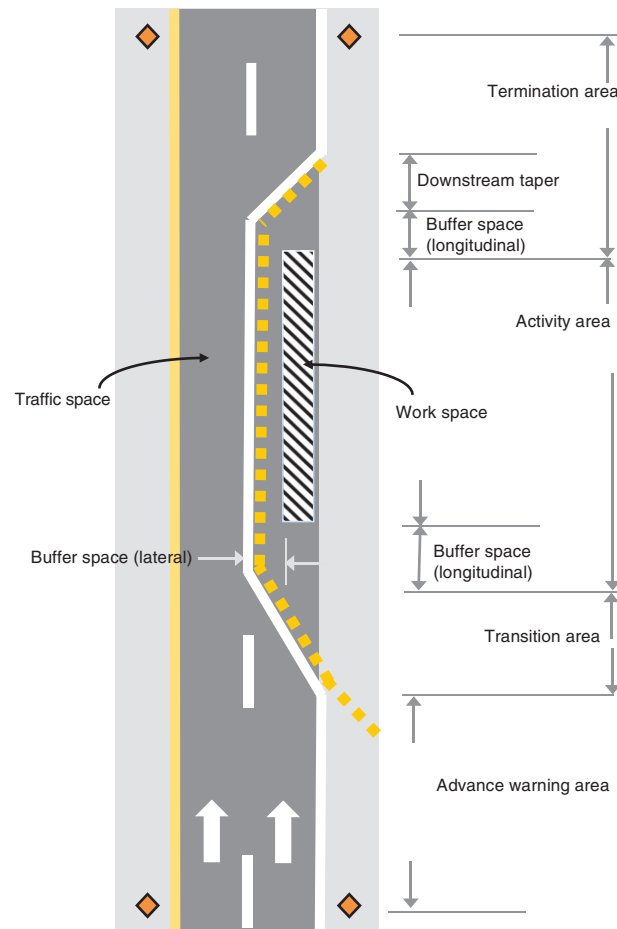


Figure 1 Example of a work-zone temporary traffic control zone.

Temporary Traffic Control Zone

A work-zone TTC zone extends beyond the location where the actual roadwork is being conducted and spans to regions upstream and downstream of the roadwork area. It could also affect parallel and other nearby roads if traffic moves to them but that is not covered in this article. The TTC zone comprises four sections: (1) an advance warning area, (2) a transition area, (3) an activity area, and (4) a termination area. A segment of roadway between the first traffic control device in the advance warning area and the last work-zone-related traffic control device on the roadway is defined as the work-zone traffic control zone. **Fig. 1** illustrates an example of a work-zone TTC zone.

In the advance warning area, road users are warned of and informed about the roadwork taking place downstream. One or a series of traffic control devices is placed in this area to suggest appropriate actions that road users should take before they encounter the changes to the roadway. In the transition area, road users must adjust their travel speed and path in anticipation of downstream roadwork. Channelizing devices are often placed on this section of the roadway to create a taper and to gradually redirect the travel path of road users so that it aligns with the travel lanes at the downstream activity area. The length of the taper is determined on the basis of the speed limit of the roadway, the work-zone lane configurations (lane closure, lane shift, etc.), and the width of the lateral shift in the vehicle travel path.

The activity area, as the name implies, is the section of roadway where maintenance and construction work are taking place. The activity area is divided into three sections: (1) the work space, (2) the buffer space, and (3) the traffic space. The work space is where equipment, material, and road workers are placed and is closed to roadway users. The maintenance and construction work is conducted in this space. The buffer space is a safety feature of work zones that aims to provide an area for errant vehicles to recover. The buffer space creates a longitudinal and/or lateral empty area between the work space and the traffic space. The traffic space is the portion of the activity area reserved for use by road users. The termination area is the section of the work zone where roadway users return to their normal path.

Temporary Traffic Control Devices

This section discusses a variety of work-zone traffic control devices that may be placed on, over, or adjacent to a roadway to regulate, warn, or guide road users, including pedestrians, cyclists, and automobiles. Specific attention should be given to the

crashworthiness of the TTC devices used in work zones ([American Association of State Highway and Transportation Officials, 2011](#)). The following warning signs shapes and colors refer to US standards. Canada uses similar signs, whereas other countries typically use different shapes and colors. In most countries, text is avoided since many drivers are not familiar with the local language, whereas in the United States, it is assumed that drivers understand English.

- Work-zone warning signs in the United States typically have an orange background with a black border and a black legend (text). Warning signs are often diamond shaped and inform road users of the roadway conditions that might not be apparent to them. These signs may be supplemented with orange flags and flashing lights. Regulatory and guide signs may be used in TTC zones as well. The regulatory signs placed in work zones should be consistent with the regulatory signs placed anywhere on the roadway system.
- Portable changeable message signs may be placed on trailers or may be mounted on trucks to provide various work-zone-related information to motorists in the advance warning area. The changeable message signs can be used for speed management, for queue warning, and to direct motorists to alternate routes.
- Arrow boards warn drivers of lane closures through the use of flashing arrows, sequential arrows, or sequential chevrons, and are required on high-speed and high-volume roads. Arrow boards are referred to as “arrow panels” and may be placed on trailers or mounted on vehicles.
- Channelizing devices, such as drums, tubular markers, cones, and barricades, are placed in a work zone to guide drivers, pedestrians, and cyclists through the work zone. The retroreflectivity, height, and type of channelizing devices and the spacing between the channelizing devices are determined on the basis of the type of work zone.
- Warning lights, including flashing or steady burn lights, are used to draw the attention of drivers to channelizing devices and advance warning signs. These devices are lightweight and portable and can easily be used to supplement the retroreflectivity of work-zone traffic control devices.
- Lighting devices, such as portable light plant towers (i.e., floodlights), balloon lighting, and roadway luminaires mounted on temporary poles, are used to provide lighting in the work zone during nighttime work. The type of roadwork activity (i.e., mobile work vs. stationary work), glare (which affects road workers and road users), light trespass onto private property near the work area, and costs are the factors that determine the minimum level of lighting and the lighting method to be used.
- Delineators, high-level warning devices or flag trees, crash cushions, temporary rumble strips, temporary traffic control signals, automated flagger assistance devices, and screens are other traffic control devices used in work zones.

Work-Zone Classification

This section discusses the work-zone duration and work-zone lane configurations, which are key factors in the development of work-zone traffic control plans.

Work-Zone Lane Configuration

The work-zone lane configuration describes if (and how) road users need to adjust their travel path to travel through the work-zone activity area. The work-zone lane configuration is sometimes referred to as the “taper type” and is broadly classified into one of the following five categories:

- Lane closure(s): One or more lanes of the roadway are closed, and vehicles must merge to an adjacent open lane. Late merge and early merge are two strategies that aim to improve mobility at work zones by encouraging early or late merges in the work zone.
- Shifting lane(s): There is no lane closure, and the number of travel lanes is not reduced; however, travel lanes are laterally shifted and often narrowed.
- Shoulder work: There is no lane reduction or lane shift, and only the shoulder is closed to traffic.
- One-lane, two-way: When a lane of a two-lane, two-way road is closed, the opposing traffic movements take turns using the available open lane and the roadway turns into an alternating one-way road.
- Rolling roadblock: This type of roadwork involves slowing down or pacing traffic with the help of shadow vehicles and law enforcement to clear an entire section of roadway for a short period. The roadwork is completed in the short period that the roadway is closed and is empty of vehicles. There is no taper in the TTC plan for a rolling roadblock.

Work-Zone Duration

The duration of a work zone determines the type of TTC devices that will be used in the work zone and is taken into account in the scheduling of the roadwork. The duration of a work zone is classified into one of the following five categories:

- A mobile work zone refers to a zone where roadwork activity moves continuously or intermittently. Roadway workers may be on foot or in a vehicle; however, they rarely stop at one location for more than few minutes. Examples of this type of roadwork include roadway striping and storm drain cleaning. As the work area is constantly moving, it is desirable to utilize traffic control devices with minimal setup and removal times for work-zone traffic control.

- Short-duration work refers to stationary roadwork that can be completed in up to 1 h. Surveying and removing graffiti from roadway signs are examples of this type of roadwork.
- Short-term stationary work refers to activities that need more than 1 h but that can be completed during daylight on a single day. Examples of this type of work include structure inspections and utility work.
- Intermediate-term stationary work refers to a type of roadwork that is completed during daylight in 1–3 days. Nighttime work that is completed in more than 1 h is considered intermediate-term stationary work as well.
- Long-term stationary work refers to work that is completed in more than 3 days.

There are trade-offs between the amount of time that roadway workers spend setting up the TTC devices and the time needed to complete the work. As the roadway workers are exposed while they are setting up the TTC devices, ideally, the amount of time spent on placing the TTC devices should not be longer than the time needed to complete the actual work. For this reason, traffic control devices that are easy to set up and remove are often used in short-duration work.

The duration of roadwork is considered in traffic impact assessments and during the scheduling of roadwork. While mobile work and short-duration work are often conducted during off-peak periods, a detailed traffic impact assessment is required for stationary work zones.

Other Considerations

The majority of work zones are planned in advance, where the planning activities consider the safety and mobility impacts of the roadwork. However, events such as utility failure and localized pavement damage may require immediate and urgent repairs. Unanticipated and emergency work presents unique challenges to the road crew, as they must set up traffic control while the emergency work is being addressed. Other factors that are considered in work-zone planning include daytime versus nighttime work, roadwork in rural versus urban areas, and work zones on high-speed facilities versus local roads.

Work-Zone Safety and Mobility Impact Assessment

Transportation agencies have traditionally conducted work-zone impact assessments to determine the impact of roadwork on mobility performance measures, such as delay and the length of queues in work zones. However, in recent years the safety of road users and roadway workers in work zones has become the primary concern of transportation agencies. This section discusses practices for work-zone safety and mobility impact assessments.

Work-Zone Safety Assessment

In the current state of practice, nominal safety is upheld in the work-zone traffic control. Nominal safety indicates that all the required guidelines and standards are met when a work-zone traffic control plan is planned and implemented on the roadway. An acceptable safety of road users and workers in the work zone is ensured by adopting a multifaceted approach of engineering, education, and enforcement (3E) toward work-zone safety.

Engineering procedures and standards are followed during the planning, design, and implementation phases of a work zone. Every aspect of work-zone TTC—from the size and placement of signs in the work zone to the work-zone speed limit—is determined to ensure that minimum safety standards are met. The *Manual on Uniform Traffic Control Devices* ([Federal Highway Administration, 2009](#)) is an example of such a standard.

In addition to the mandatory aspects of work-zone traffic control required by design standards, innovative concepts, such as variable speed limits, speed warning displays, photo-enforced speed compliance, and nighttime flashing arrows, are implemented in work zones to improve work-zone safety. During the planning phase for a work zone, mobility performance measures, such as the operating speed of the work zone and the length of queue, are determined and their potential impacts on roadway safety are investigated. For example, advance warning signs are placed on the roadway to make sure that drivers are notified of a work zone before they arrive at slow-moving traffic.

Enforcement is another strategy that transportation agencies utilize to improve roadway safety at work zones. Law enforcement officers enforce speeding, no passing, and cell phone use and distracted driving laws at work zones, as these behaviors endanger the safety of the road workers and other road users. To further deter the road users from unsafe behavior, some jurisdictions have doubled the fines incurred for traffic violations at work zones in comparison to those incurred under normal roadway conditions. Considering the potential severe injuries of roadway workers as a result of roadway crashes, some jurisdictions punish drivers who hit roadway workers with prison sentences. As the presence of law enforcement at work zones is not always practical, automated speed enforcement cameras are used in some jurisdictions to enforce speed limit in work zones.

Work-zone education and awareness campaigns are another category of activities that transportation agencies use to increase public awareness of roadway worker safety and to prevent distracted driving in work zones. These campaigns are often implemented a few weeks before the construction season begins. Dynamic message signs are also used to educate road users about work-zone safety. However, limited information on the impacts of these efforts on changing driver behavior in work zones is available.

Despite the efforts of transportation agencies to provide nominal safety in work zones, crashes continue to occur in work zones because of environmental factors, driver behavior, etc ([Campbell et al., 2012](#)). Thus, it is important to understand the expected safety

performance of a work zone. This is referred to as “substantive safety.” Over the past few decades, the importance of substantive safety has been recognized by practitioners, and safety performance functions (SPFs) and crash modification factors (CMFs) have been developed as part of methodologies to predict and assess the expected safety performance of various roadway facilities. An SPF predicts the number of crashes expected at a transportation facility that meets the base geometric design and traffic conditions. For example, the base conditions for a rural multilane divided highway segment may be lanes that are 3.6 m wide, a right shoulder that is 2 m wide and paved, a median that is 10 m wide, and no lighting and no automated speed enforcement on the roadway segment. When the condition of a transportation facility differs from the base SPF conditions, CMFs are used to adjust the SPF predictions for the base condition to the actual or proposed conditions. For instance, a CMF might be used to determine the impact of a reduction of the lane width from 2 to 1.8 m.

CMFs can be utilized to determine the impact of a work zone on roadway safety. However, only a handful of CMFs have been developed for work zones. The only work-zone-related CMF in the *Highway Safety Manual* determines the number of work-zone-related crashes on the basis of the length and duration of a work zone ([American Association of State Highway and Transportation Officials, 2010](#)). Ideally, work-zone CMFs should be developed to predict the frequency, severity, and type of crashes in the advance warning area, transition area, work area, and termination area of a work zone. This information informs designers to implement appropriate countermeasures to address the safety risks.

Similarly, there is a need for the development of CMFs for a variety of work-zone traffic control strategies. For example, what is the impact of queue warning systems on reducing the number of crashes across the work zone? Developing CMFs for work-zone traffic control devices will assist engineers with the implementation of strategies that are expected to be most effective in improving the safety of work zones.

The lack of work-zone-related CMFs is in part attributed to a lack of high-resolution crash data and work-zone record management systems. The importance of a systematic archive of work-zone data was recently recognized in the industry. For example, the US Department of Transportation recently published specifications for the Work-Zone Data Exchange, which aims to standardize data collection at work zones. This effort is important, as the collection of this information will assist safety engineers with determining the factors contributing to work-zone crashes, which in turn will lead to the adoption of strategies to mitigate such crashes.

Even if the United States still does not have good data, such data exist in other countries, for example in Sweden. The Swedish Transport Administration’s Publication 2014:128 shows an analysis of all fatal roadway crashes related to roadway construction for the 11-year period, 2003–13. There were a total of 51 fatal crashes with 56 people dying in that period. A total of 12 victims were women and 44 were men. Their average age was 50 but the age of the deceased varied from 3 to 93. Out of the 56 killed, 6 were construction workers and the other 50 were road users. This means that the overall construction-zone fatality rate, including construction workers, was 0.05 per 100,000 people and year. This can be compared to the overall roadway fatality rate in Sweden for those years of 4.2 per 100,000 people or in the United States of 12.5 per 100,000 people as an average of 2003–13. So, in Sweden, 56 people died in construction zones which is 1.3% of all 4251 traffic fatalities for those years, which is not an insignificant portion, especially since another 3000 construction-zone crashes caused injuries during these years. Out of the 51 fatal crashes in the 11-year period, 15 were single-vehicle and 12 were of rear-end type. Out of the 12 rear-end crashes, the killed person was rear-ended by someone else in 5 cases and rear-ended someone else in 7 cases. Ten fatal crashes were of opposing direction type and one was an intersection fatality. Five were pedestrians hit by motor vehicles, three were bicyclists hit by motor vehicles, one was a moped rider, and two were single-bicycle crashes. Then there was one bicycle–bicycle fatal crash and one crash involving a tractor.

Almost all the fatal crashes occurred in daylight, in good weather, on high-volume roads. That those happened in daylight is not surprising since most construction work in a country this far north is done during the months when the sun is up from well before the morning commute to well after typical evening commute. Two-thirds of the fatal crashes happened in long-term, stationary work zones, 17% in rolling or moving work zones, and 10% in intermittent work zones. A majority (34 of 51) of the fatal crashes occurred within the work zone itself but most of the rear-end fatalities happened in the zone leading up to the work zone and typically after people had passed warning signs. In almost all crashes, too high speed was a contributing factor. Of the six construction workers killed, five were killed in rolling or short-term construction areas. In total, 20 people were killed in automobiles, 9 on motorcycles, 8 in trucks, 6 on bicycles, 5 as pedestrians, 2 in construction-work vehicles, and 1 person on a moped. Of the six construction workers killed, one was traveling in a passenger car, two in construction vehicles, two in trucks, and one was not inside a vehicle and is therefore classified as a pedestrian. The primary object struck was in 24 cases a vehicle in motion, in 9 cases a stopped vehicle, in 11 cases the ground, and in 7 cases a barrier or bridge abutment. In at least six of the crashes, a person was under the influence of alcohol or drugs.

A total of 17 fatal crashes occurred on rural two-lane roads, 12 on multilane motorways (freeways), 3 on two-lane limited access highways (with one lane in each direction), 2 on rural multilane roads (more than two lanes), 5 on “narrow” rural roads (lacking centerline), 5 on two-lane urban streets, 1 on a four-lane urban street, and 3 at unclassified segments.

More than half of these Swedish fatal crashes occurred on roads that normally have a speed limit of 90 km/h (56 mph) or higher. The speed limit had been temporarily reduced in at least 19 of the 51 fatal crashes. For 18 of the fatal crashes, nominal safety was met with all signage following the protocol perfectly and, for 4 of the crashes, signage was done incorrectly. In the remaining 29 crashes, it is unclear whether plans had been followed.

Work-Zone Mobility Impact Assessment

Although motorists understand that work zones negatively affect mobility, transportation agencies conduct work-zone traffic impact assessments to ensure that delays and queues at work zones do not exceed acceptable thresholds. In general, work-zone traffic impact

assessment techniques can be classified into two broad categories: analytical methods and traffic simulation models. Analytical methods employ queuing theory and parametric and nonparametric models to estimate work-zone performance measures. These methods are coded into custom spreadsheets or are developed as stand-alone work-zone-specific analytical tools. The work-zone hourly traffic volume, work-zone lane configuration, and work-zone capacity are among the most common input parameters of these tools. Delay, length of queue, and costs to road users are among the output parameters of these models.

Traffic simulation models, on the other hand, are often computationally intensive and require more effort to code the work zone. However, a simulation model provides a detailed analysis of the impact of a work zone on corridor mobility and traffic operations at upstream on-ramps, off-ramps, and detour routes. A traffic simulation model also allows the study of complex and nonconventional work-zone configurations. However, the results of simulation models are meaningful only when the simulation model is calibrated to replicate the work-zone capacity.

Worker Safety

Even though transportation agencies attempt to guide drivers through work zones safely, errant vehicles occasionally intrude the taper and work area, which puts road workers at grave risk. Work-zone intrusion alarm systems have been developed to warn road workers about immediate hazardous conditions. These alarm systems utilize various sensor technologies to detect impacts with work-zone traffic control devices or use radar, Light Detection and Ranging (LiDAR), or other similar systems to monitor the taper, buffer space, and work area and to detect a vehicle intrusion. Workers are then warned about the intrusion through visual, audio, and/or haptic warnings. Another strategy used to protect roadway workers and errant vehicle occupants are truck-mounted attenuators (TMAs), which will be discussed in the next section.

Truck-Mounted Attenuators

Protective vehicles are placed in work zones to shield road workers from errant vehicles and to improve the safety of travelers. The protective vehicles include advance warning vehicles, barrier vehicles, and shadow vehicles. Advance warning vehicles are equipped with an arrow board to inform drivers about the downstream roadwork and are typically placed on the shoulder in the work-zone advance warning area. Barrier vehicles and shadow vehicles are placed in the activity area to protect road workers from errant vehicles and to protect errant vehicles from an impact with construction equipment. Barrier vehicles are typically used in stationary roadwork and are not occupied by drivers. Shadow vehicles, on the other hand, are used in mobile operations.

TMAs are attached to protective vehicles to reduce the risk of injury to the passengers of a vehicle that crashes head-on into the protective vehicle. **Fig. 2** illustrates examples of TMAs. TMAs function as a portable crash cushion that is attached to the rear of a truck and are designed to absorb a portion of the crash impact energy. There are two primary considerations with regard to choosing a TMA for a work zone.

The first factor is the level of crash protection provided by the TMA. The commercially available TMAs provide either Test Level (TL) II or TL III, as defined in *NCHRP Report 350: Recommended Procedures for the Safety Performance Evaluation of Highway Features* (Ross et al., 1993). TL II TMAs are designed to absorb some of the impacts of a crash when the difference between the speed of the TMA and the speed of the vehicle crashing into the TMA is 70 km/h, whereas TL III TMAs are designed to accommodate a speed differential of up to 100 km/h. Although TMAs are often designed to provide protection against rear-end crashes, some TMAs provide protection against angle collisions as well.

The second factor is the weight and type of vehicle to which the TMA is attached. Attaching the TMA to lightweight vehicles may provide protection against errant passenger cars; however, attaching the TMA to heavy vehicles reduces the risk that the protective vehicle will be pushed into the work space if a tractor trailer hits the TMA at a high speed. Manufacturer guidelines and

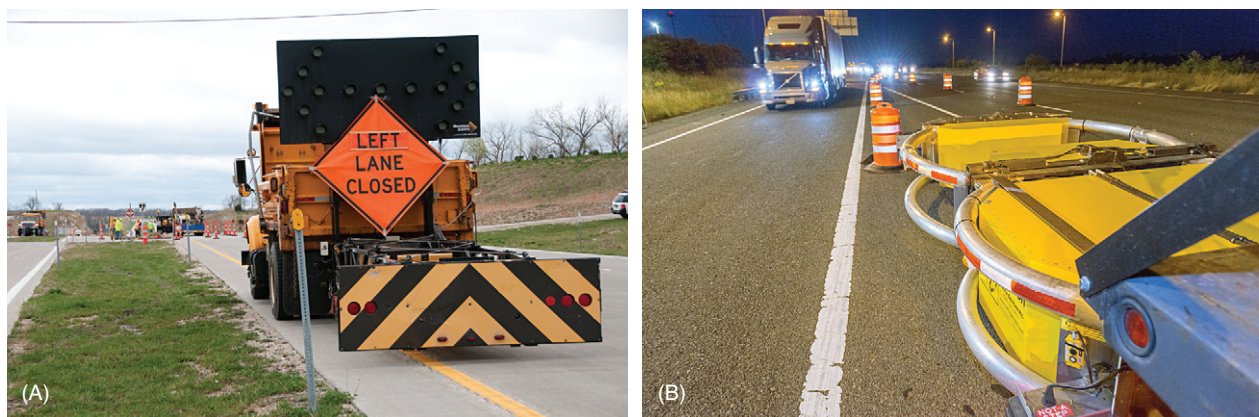


Figure 2 TMAs utilized by the Missouri and Oregon Departments of Transportation.

transportation agency manuals determine the appropriate type of TMA for a work zone and the type of vehicle to which a TMA should be connected.

When a TMA is struck by an errant vehicle, it is very likely that the TMA will be pushed forward as the result of the impact. The distance that the TMA is pushed forward is referred to as the “roll-ahead distance.” The roll-ahead distance is a function of the weight of the TMA, the weight of the errant vehicle, and the speed of impact. It is important that the roll-ahead area in front of a TMA be free of roadway equipment and road workers.

In addition, TMAs extend behind the vehicle to which they are attached, and it is essential to ensure that TMAs are highly visible to drivers, particularly during nighttime. Retroreflective materials and warning lights may be used to improve the visibility of TMAs.

In addition to TMAs that are attached to trucks, trailer-mounted TMAs are also available. The advantage of these TMAs is that the vehicles can be removed from the TMA at the work site so that the driver and the vehicle are free for other activities.

Finally, when a TMA is struck, the driver of the TMA is at risk of injury as a result of the impact. To address this situation, transportation agencies have recently experimented with semiautonomous TMAs, where a driverless TMA follows a lead vehicle so that the TMA is no longer occupied by a driver.

Future Trends in Work Zones

A key requirement for improving safety and mobility in work zones is access to high-resolution archival work-zone data. Examples of these data include detailed work activity diaries, traffic data collected at the taper and upstream of the taper, work-zone lane-configuration record, and crash records that identify the location of crashes with respect to work-zone traffic control zone areas. These high-resolution data are typically not collected at work zones, and, when collected, they are collected and stored in an arbitrary manner. The Work-Zone Activity Data (WZAD) protocol currently under development aims to improve the quality and the consistency of US data collected and the timeliness of data collection at work zones and will serve as a standard for data collection at work zones.

Another area where WZAD may be useful is mitigating the challenges associated with the travel of connected and autonomous vehicles (CAVs) through work zones. Work zones create irregularities in the road, and CAVs need to accurately process the information provided in the TTC zone to navigate the work zone safely. The accurately processing this information is challenging, as work zones and activity types vary greatly, and there is always the risk that TTC devices may be damaged or not updated to reflect changes in the work area. CAVs could receive the work-zone layout in the advance warning area through vehicle-to-infrastructure communication, communicate with the traffic control devices in the work zones to prevent work-zone intrusions, reduce the delay and the length of queue in work zones through the use of cooperation and lane changes, and issue electronic warnings to roadway workers when a potential crash is predicted. In the long term, CAVs are expected to improve safety and mobility in work zones.

References

- American Association of State Highway and Transportation Officials, 2010. Highway Safety Manual. AASHTO, Washington, DC.
- American Association of State Highway and Transportation Officials, 2011. Roadside Design Guide, fourth ed. AASHTO, Washington, DC.
- Campbell, J.L., Lichty, M.G., Brown, J.L., Richard, C.M., Graving, J.S., Graham, J., O’Laughlin, M., Torbic, D., Harwood, D., 2012. Human factors guidelines for road systems. National Cooperative Highway Research Program Report No. 600, Transportation Research Board, Washington, D.C.
- Federal Highway Administration, 2009. Manual on Uniform Traffic Control Devices. Available from: <https://mutcd.fhwa.dot.gov/>.
- Ross Jr., H.E., Sicking, D.L., Zimmer, R.A., Michie, J.D., 1993. Recommended procedures for the safety performance evaluation of highway features. National Cooperative Highway Research Program Report No. 350. Transportation Research Board, Washington, DC.

Costs of Accidents

Ulf Persson, The Swedish Institute for Health Economics, IHE, Lund, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Incidence and Prevalence	197
Data Requirements for Incidence Cost Studies	198
Cost Offsets and Values of Risk Reduction	198
Biography	199
References	199
Further Reading	199

The economic costs of accidents are a result of lost scarce resources. There are three types of economic costs: material costs, medical costs, and the value of lost production. Material costs are lost resources due to damaged vehicles, destroyed materials, time for repairs and time for police and for administration of insurances. Medical costs are lost resources in the health care sector, including time for physicians and nurses, use of medical devices, and time for caregivers. These are examples of costs related to fatal and nonfatal casualties. Material costs and health care costs are direct costs. The third type of cost is the value of lost production, due to sick leave, early retirement, and premature death. These costs are called indirect costs.

Costs of accidents can be considered from a broad societal perspective, that is, including all consequences independent of who is taken the burden of the costs. Estimates of costs from a broad societal perspective, that is the socioeconomic accident costs, can be of interest for various purposes. One of the purposes is the evaluation of the benefits of road safety measures. Cost offsets from preventing accidents are a value driver for traffic measures. Another purpose is the ex post representation of the extent of consequences in monetary terms to illustrate the burden of injuries, and compare this with such other costs, as for instance, costs for other types of injuries or diseases. A third purpose of costing is to estimate the external costs of road accidents in order to internalize various economic costs to influence road user behavior. A fourth purpose is to use estimates as a basis for judicial assessments of the sums paid in compensation for damages and other lost resources.

Estimation of accident costs involves identifying, measuring, and valuing all resources lost that occur as a result of the accidents. Valuing resources lost often rely on market prices as far as they are available. For material damage costs and administrative costs, there are usually market prices. For personal injuries there are market prices available; however, different methods must be used for direct and indirect personal casualties.

Direct health care costs are mainly based on market prices taken from the estimates of expenditures for health care and social services for home care. Indirect costs are mainly valued by the human-capital approach. Under this approach, the estimation of lost output is normally based on the wages that could have been earned by individuals if the illness or injury in question had not happened. The methodology was early applied in the 17th century by Sir William Petty (1699). Another contribution to the literature is "The Money value of a Man" by Dublin and Lotka (1946). Their purpose was to "estimate the value of a man to his dependents, that is to those who have a direct interest in his earnings." In answering this question, Dublin and Lotka estimated the discounted present value of an individual's net earnings. The net earnings were defined as the difference between discounted present value of future earnings and future consumption expenditures. It is also an answer to the question: "What is the size of a life insurance for an individual in order to guarantee the same material living standard to a family should the family provider die?" This net value is also comparable to the value of a slave on the slave market, that is, the expected value of a slave's production minus the costs to give the slave food and shelter. One of the first traffic accident cost estimations was applications with this methodology. For example, Dawson (1967) estimated the costs in Great Britain.

With this net value approach, or slave calculations, the sick or injured person's value of own consumption loss is not included. Another approach including the sick or injured individuals' consumption is the gross loss of production value. In these calculations there is no reduction of consumption value. The classical use of this approach is the estimate of the cost of war (Petty, 1964). The costs of the French–German war in 1870–71 were estimated by Giffen (1880). The US involvement in the First World War and the associated loss of human resources valued by using the value of gross lost production was estimated by Clark (1931). This is the same approach that later on have been used to estimate the value of lost production due to illness and injuries in the cost of illness studies (COI).

The seminal work in the field of COI is Dorothy P Rice's study of 1966, which presents and applies a methodology for estimating the cost of major disease categories in the United States (Rice, 1966). Additional examples of application for Sweden were Lindgren (1981), Rydenfelt (1949), and Rydenfelt (1991).

Under this approach, the economic costs associated with diseases or injuries are divided into two principle categories: direct costs and indirect costs. Direct costs represent the value of resources used for prevention, detection, treatment, rehabilitation, and long-term care due to the existence of illness or injuries. Costs are estimated by summing up the expenditures on each category attributable to the disease or injury of interest. Indirect costs represent the value of the goods and services that would have been produced if a person had not fallen ill or been injured. Often this is estimated using the "human capital" approach, whereby the lost output is

measured by the wages that could have been earned by individuals if the illness or injury in question had not happened. To estimate costs, observed market prices of goods, services, and the labor force are used.

The traffic accident costs in Sweden was published by Mattsson (1970). Hartunian et al. (1981) published the economic costs of the annual incidence of coronary heart disease, cerebrovascular disease, cancer, and motor vehicle injuries in the US 1975. Krupp and Hundhausen (1984) estimated the costs of road traffic accidents in Germany, and Jensen (1983) and Elvik (1989) the costs in Denmark and Norway, respectively. Apart from the study by Hartunian et al. (1981) there are also other US estimates (Rice et al., 1989). One international comparison of socio-economic accident costs was the project COST 313—Socioeconomic cost of road accidents (Alfaro et al., 1994). This project attempted to show how costs of accidents could be evaluated and how different methods and differences in terms of application affect the outcome.

Incidence and Prevalence

Estimates of health-care costs and value for lost production are commonly conducted using the prevalence approach, that is, the health care costs and the value of lost production during a certain time-period caused by a given injury are estimated, irrespective of when the injury emerged. However, traffic accidents often lead to costs that arise 5–10 years after the accident. By using the incidence approach, all present and future health care costs are estimated and discounted to the year in question. This approach is more commonly used when analyzing preventive programs or programs aimed at rehabilitation (Hartunian et al., 1981). Fig. 1 illustrates the difference between the two approaches.

For injuries causing short-term consequences, that is, consequences with durations shorter than one year, there is no difference between the incidence and the prevalence approach. However, for injuries resulting in long-term consequences, the two approaches will have different results even in a steady state situation, that is where the number of accidents, severity of accidents, and costs per accident are constant over time. The reason for this is that under the prevalence approach no costs will be discounted but under the incidence approach all future costs will be discounted. The present value for incidence approached costs will therefore always result in less aggregate costs than prevalence-based cost calculations.

Cost estimates under the prevalence approach are often not a “pure” application of the theoretical correct prevalence approach. In practice, there are two exceptions that have to do with the calculations of indirect costs due to premature death and early retirement. The indirect costs due to premature death and early retirement used to be estimated from the time of the accident and for subsequent time periods until the individuals are expected to die or not produce any productive work anymore, respectively. In other

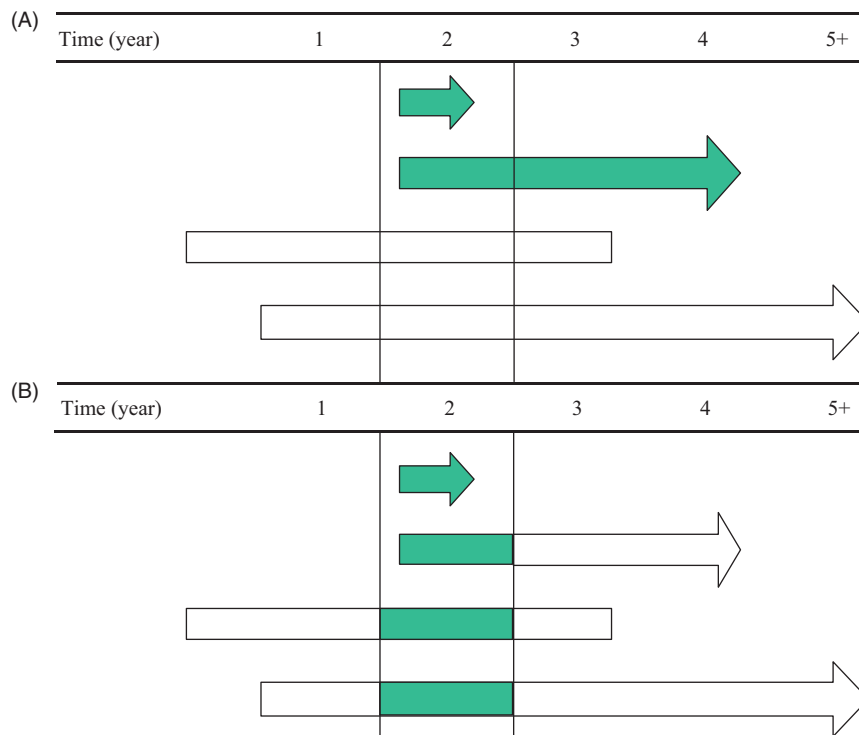


Figure 1 (A) and (B): consequences to include for estimating the material costs of accidents with the incidence and the prevalence approach, respectively. (A) Costs for consequences that should be included then estimating cost of accidents for year 2 with the incidence approach (shaded areas). (B) Costs for consequences that should be included then estimating cost of accidents for year 2 with the prevalence approach (shaded areas). Source: Own figure.

words, a usual technique for the prevalence approach is to apply a similar procedure for estimating these costs as for estimating them with the incidence approach, that is, discount future indirect costs.

The explanation for this procedure is that a pure application of the prevalence approach even for mortality and early retirement includes both data collection problems and conceptual difficulties in interpreting the results. For example, a pure prevalence application requires information about the number of individuals that should have been alive and productive during the year in question, if accidents had not happened at any year prior to the year in question. In order to avoid the problem of constructing a hypothetical model on which basis a counterfactual situation can be deducted, researchers apply the prevalence approach by making these two exceptions.

Both the prevalence and the incidence approach are theoretically correct, but results will differ for all discount rates above zero. The two approaches will be used to answer two different questions. The prevalence approach will answer the question: what is the economic burden of injuries for one year regardless of when they occurred? This is a question relevant for those interested in allocating resources to the health care sector treating injured persons during a certain budget year. The incidence approach will answer the question: what is the cost offset by reducing the number of accidents the forthcoming year? This is a question of interest for those allocating resources to accident prevention and to estimate the relationship between investments in traffic safety measures and the benefits of future costs savings.

For those interested in creating incentives for drivers to take account of economic consequences by internalizing accidents costs, for example, in motor vehicle insurance (Skogh and Katz, 2002), the incidence approach is also the most relevant approach. For creating fees for private vehicle insurances corresponding to accident costs, it is important to understand the amount of future economic burden and identify the most important cost components, the incidence approach, and applications for traffic injuries will be needed.

Data Requirements for Incidence Cost Studies

Accidents lead to healthcare costs and lost production during a certain time-period, for example the year when the injury occurred. However, accidents may also lead to costs that arise many years after the accident. By using the incidence approach, all present and future health care costs are estimated and discounted to the year in question.

Long-term follow-up of patient cohorts with severe casualties, registering overall survival and health care consumption are required for estimating health care costs in economic evaluation of preventive measures. This approach is commonly more difficult to apply than the prevalence approach due to less relevant data from cohorts available.

Two examples of studies have been used for incidence estimates of traffic accident costs in Sweden and these are used here as examples. The first Swedish study on long-term consequences due to traffic injuries were conducted by Thorsson (1975). In Thorson's study, individuals injured in road traffic accidents during the middle of the 1960s participated in a long-term medical follow-up. Investigators interviewed and examined the injured 4–5 years after the accident. The follow-up showed that more than half of the injured suffered from some sort of aftereffect of the original injury 4–5 years after the accident. The Thorson's study provides information of the long-term consequences in terms of loss of health, and this information was used for estimating cost of care (Persson, 1982).

The second study by Persson et al. (2004) collected data on resource consumption for care for 95 severe traffic casualties due to road traffic accidents occurring in a 12-month period, 1991/1992, in Sweden and over a subsequent period of 8 years. The study also brings to light the extent of employment 8 years after the accident, an important outcome of rehabilitation for individuals injured in traffic. The Persson et al.'s (2004) study shows that 2/3 of the costs arise during the period 2–8 years after the accident. More than half of the total costs consist of lost production. Moreover, the individuals themselves are affected due to loss of income. The total costs of in-patient care arise during the first year after the accident. The costs of outpatient care and cost of care at home care are fairly equally divided by the first year after the accident and the following 2–8 years.

Of the 75 injured people who were considered to be part of the labor market, one-third did not return completely to work. The majority of those who did not return completely to the labor market were women. The total costs of production losses amount to be a major part of the total costs of traffic injuries.

Because 2/3 of the costs appear 2–8 years after the accident and these are the costs that are hard to capture with conventional registrations of care cost data, specific cohort studies are needed. Capturing information of costs of home care, special forms of housing and lost production require many years follow-up after the accident.

A general problem with such long-term follow-up studies as 8 years is the risk of including cost of health care and social services that are related to increasing age, and therefore including costs that would have appeared even in the absence of a traffic accident. The data collection in the study by Persson et al. (2004) is mainly based on the subjects' own judgment of which services to include. As most subjects belongs to younger age groups, it was expected that comorbidity, that is individuals suffering from injuries and other diseases at the same time, and age-related costs might be a minor problem when studying accidents with this method. For many other types of injuries and sicknesses, particularly among elderly, long-term follow-up studies could be heavily biased by comorbidity and age-related costs.

Cost Offsets and Values of Risk Reduction

When seeking answer to the question: what is the value of reducing physical risk and reducing fatalities and nonfatal casualties? To answer this question an estimate of the cost offsets from material costs, medical costs and lost production is needed. In addition, the

value of risk reduction per se is also relevant. The value of risk reduction per se is a measure of how much society is willing to pay (WTP) in addition to the summary of health care costs, gross lost production, cost of property damage, and administration for improved traffic safety. In some countries, for example Sweden, this was earlier denoted "the human value." However, it is nowadays customary to denote the WTP estimate the value of risk reduction per se.

This value will not be discussed further here, but there is a potential risk of double counting when estimating the value of risk reduction when adding the value of risk reduction based on studies trading of income for risk to the material costs discussed here.

For example, the Swedish National Road Administration, when estimating the value of reducing the risk of one statistical fatality, assumes that individuals when answering the question of the value of risk reduction per se also include the value of lost consumption. If so the individual's value of lost consumption should be withdrawn from the total value of a statistical fatality.

Furthermore, if so, the estimate of the value of risk reduction corresponding to one statistical fatal casualty the risk reduction per se should be added to the net lost production, that is gross lost production minus consumption. This is the slave calculation, and this is the relevant material cost that should be added to the value of risk reduction per se, that is, the WTP amount estimated from stated preferences or revealed preferences.

For nonfatal casualties, without any risk of death, the risk of double counting is less and the value of consumption should not be withdrawn. The methods of contingent valuation and revealed preference, answer the question of value of safety for the individual herself and the estimates include their preferences for risk, that is risk aversion, and these methods are discussed elsewhere.

Biography

Ulf Persson, PhD, is Former Adjunct Professor at the Institute for Economic Research, Lund University, Researcher in transport economics at the Department of Technology and Society, Lund Institute of Technology, Lund University, CEO and Research Director at the Swedish Institute for Health Economics (IHE). Now he is Senior Advisor at IHE. He has 39 years research experience in the development and application of economic evaluation methods in health care. His main research areas are health economics and transport economics. He has published 260 articles, books, and reports in health economics and economics of transport safety.

References

- Alfaro J.-L., Chapuis, M., Fabre, F., 1994. European Cooperation in the field of Scientific and Technical Research, Socioeconomic Cost of Road Accidents, Transport Research COST 313, Commission of the European Communities, Brussels/Luxembourg.
- Clark, J.M., 1931. The Cost of the World War to the American People. Yale University Press, New Haven, CT, doi:10.2307/1838090.
- Dawson, R.F.F., 1967. Cost of Road Accidents in Great Britain. London Road Research Laboratory, London.
- Dublin, L.I., Lotka, A.J., 1946. The Money Value of a Man. Ronald Press, New York.
- Elvik, R., 1989. Road Accident Cost in Norway. Institute of Transport Economics, Oslo.
- Giffen, R., 1880. Essays in Finance. G Bell & Sons, London.
- Hartunian, N.S., Smart, C.I., Thompson, M.S., 1981. The Incidence and Economic Costs of Major Health Impairment. Lexington Books, Toronto.
- Jensen, F., 1983. Trafikuheldsøkonomiske omkostninger (Costs of Traffic Accidents). Laboratoriet for samfundsmedicinsk og Sundheds-Økonomisk forskning, Odense Universitet. Rådet for Trafiksikkerhedsforskning og Vejdirektoratet, Odense.
- Krupp, R., Hundhausen, G., 1984. Volkswirtschaftliche Bewertung von Personenschäden in Strassenverkehr. Bundesanstalt für Strassenwesen, Bergisch Gladbach.
- Lindgren, B., 1981. Costs of Illness in Sweden 1964-1975. Lund Economic Studies, Lund.
- Matsson, B., 1970. Vägtrafikolyckornas samhällsekonomiska kostnader (The Economic Costs of Road Accidents) Statens trafiksäkerhetsråd. rapport nr 116.
- Persson, U., 1982. vägtrafikolyckornas samhällsekonomiska kostnader (Economic costs of Road accidents). IHE meddelande 1982:4. Institutet för Hälso- och Sjukvårdsekonomi, Lund.
- Persson, U., Maraste, P., Bertman, M., Svensson, M., 2004. The economic consequences of personal injuries in road traffic accidents – an eight-year follow-up of non-fatal casualties. in: Persson, U. (Ed.), Valuing Reductions in the Risk of Traffic Accidents Based on Empirical Studies in Sweden. Lund Institute of Technology, department of Technology and Society, Traffic Engineering, Lund University. Lund.
- Petty, W., 1964. The Economic Writings of Sir William Petty 1665. Together with the observations upon the bills of mortality. In: Hull, C.H. (Ed.), Reprints of Economic Classics. Cornell University, New York.
- Rice, D.F., 1966. Estimating the Cost of Illness, Health Economic Series, No 6. US Department of Health, Education and Welfare, Washington DC.
- Rice, D.P., MacKenzie, E.J., Associates, 1989. Cost of injury in the United States. A report to the congress 1989. Institute for Health & Aging University of California and Injury Prevention Center. The John Hopkins University, San Francisco. CA.
- Rydenfelt, S., 1949. Sjukdomarnas samhällsekonomiska aspekt (The Economic Aspect of Diseases). Mimeograph. Nationalekonomiska Institutionen, Lund.
- Rydenfelt, S., 1991. Sjukdomarnas samhällsekonomiska aspekt (The Economic Aspect of Diseases). IHE Monograph. Institutet för hälso- och Sjukvårdsekonomi, Lund.
- Skogh, G., Katz, J., 2002. Rättsekonomiska aspekter på skadeståndsrätt och statlig regress. Rapport till personskadekommittén. Linköping: Ekonomiska institutionen vid Linköpings Universitet.
- Thorsson, J., 1975. Long-term Effects of Traffic Accidents. Berlingska Boktryckeriet, Lund.

Further Reading

- Felt, K.O., 1958. vägtrafikolyckornas kostnader – En samhällsekonomisk studie (The Road Traffic Accidents – An Economic Study). Meddelande nr 7. Statens trafiksäkerhets råd, Stockholm.
- Giffen, R., 1880. Essays in Finance. G. Bell & Sons, London.
- Giffen, S.R., 1904. Economic Inquiries and Studies, vol. I. George Bell and Sons, London.
- Trawén, A., Maraste, P., Persson, U., 2002. International comparison of costs of a fatal casualty of road accidents in 1990 and 1999. Accident Analysis and Prevention 34, 323–332.

Critical Issues for Large Truck Safety

Matthew C. Camden, Jeffrey S. Hickman, Richard J. Hanowski, Martin Walker, Virginia Tech Transportation Institute, Blacksburg, VA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	200
Truck Size and Weight	201
Vehicle Length	201
Vehicle Width	201
Vehicle Height	201
Vehicle Weight	201
Truck Crash Causation	202
Risk Factors for Large Truck Crashes and Effective Strategies to Improve Safety	202
Fleet Safety Culture	203
Steps to Develop a Strong Safety Culture	203
Large Truck Driver Inattention	204
Distraction From Secondary Tasks	204
Driver Fatigue and Drowsiness	205
Impairment From Drugs and Alcohol	205
Driver Health and Wellness	205
Speeding	206
Advanced Driver Assistance Systems	206
Automated Driving Systems	207
Large Truck Roadside Inspections	208
Conclusion	208
See Also	208
Relevant Websites	209
References	209
Further Reading	209

Introduction

Large trucks are the primary means of transporting cargo on roadways. Large trucks include tractors or lorries that can be paired with one or several trailers, non-articulated straight trucks, dump trucks, refuse trucks, and cement trucks. By US definition, large trucks weigh more than 26,000 lb, which equals 11,800 kg; however, a tractor with a fully loaded trailer can weigh as much as 80,000 lb (36,300 kg) or more. In many US states, an oversize permit is required for weights above 80,000 lb, whereas other US states allow heavier trucks on certain roads without a special permit. For example, Maine, New Hampshire, and Vermont allow trucks with weights up to 100,000 lb (45,400 kg) on all state roads, freeways, and interstate roads. Alaska allows the heaviest vehicle truck combinations of up to 138,000 lb (62,600 kg) without needing overweight permits. In other countries, even heavier trucks are allowed. For example, since 2015, 64 metric ton (141,000 lb) trucks are allowed on all roads in Sweden, although urban areas may restrict vehicle length and thereby gross weight. Before 2015, “only” 60 metric ton (132,000 lb) trucks were allowed in Sweden. Furthermore, in Sweden and Finland, 74 metric ton (163,000 lb) trucks are allowed on approved roads since 2018. In several other EU countries, only 40 metric ton (88,000 lb) combinations are allowed. The heaviest vehicle combinations allowed on public roads anywhere in the world, without special permit, are the Australian so-called road trains, weighing up to 200 metric tons (440,000 lb). However, these road trains are allowed only on approved roads. Thus, dependent on jurisdiction, fully loaded large trucks may be 20–200 times heavier than typical passenger vehicles.

Additionally, large trucks with trailers in the United States may, without permits, be up to 19.8 m (65 ft.) long. In Sweden, 25.25 m (82.8 ft.) trucks are allowed on all roads without special permits but city ordinances may limit lengths on urban streets. In Australia, 27.5 m (92 ft.) long combinations are allowed “everywhere” including in cities, whereas road trains that are 36.5 m (120 ft.) are only permitted to travel on approved routes. This means that large trucks can be much longer than the average passenger vehicle.

Due to the mass and length of these vehicles, large truck crashes should be disproportionately dangerous and costly. On average, large trucks account for less than 5% of the registered vehicles in the United States but are involved in more than 10% of all fatal crashes annually. But large trucks are also driven longer distances per year. In 2016, the US Federal Highway Administration (FHWA) states that Class 8 trucks (gross weights over 33,000 lb or 15,000 kg), on average, traveled 68,155 miles (109,685 km) per year, whereas passenger cars traveled only 11,244 miles (18,091 km) per year. Large truck crash involvement in Europe has similar

numbers. The large truck overall crash rate per distance driven is lower than passenger vehicle in most countries. However, large truck–passenger vehicle crashes often have more serious consequences because of large trucks’ heavier weights. Thus, in many countries the large truck fatal crash rate per distance driven is higher than passenger vehicles. For example, in 2016, passenger vehicles were involved in 1.16 fatal crashes per million miles (1.61 million km) traveled, whereas large trucks were involved in 1.29 fatal crashes per million miles (1.61 million km) traveled. This is not the case in all countries. For example, in Sweden, large trucks (defined as trucks weighing above 3500 kg or 7700 lb) are not overrepresented in fatal crashes per kilogram driven. Many other countries lack good data, but statistics suggests large truck crashes in developing countries may be deadlier due to lower standards for vehicle crashworthiness, fewer and laxer safety regulations, and higher densities of two-wheeled vehicles. And, two-wheeled riders may be disproportionately killed by heavy trucks in industrialized countries too. In 2016, 271 collisions between heavy goods vehicles (HGVs) and bicyclists in London, England, resulted in 16 riders being killed and 56 seriously injured and a further 198 injured. Although only 1.5% of cyclist casualties occurred in collisions with HGVs, this resulted in 16% of cyclist deaths. Many of them were killed by right-turning HGVs. That may also be an issue in other countries as evidenced by studies from Stockholm, Sweden. In the 5-year period 2007–11, “only” seven bicyclists were killed in the city of Stockholm, but five of the seven were killed by heavy trucks turning right at intersections.

Truck Size and Weight

In the United States, large truck size and weight restrictions are set in statute by legislation and managed by the Department of Transportation’s FHWA’s Office of Freight Management and Operations. FHWA, in coordination with the Federal Motor Carrier Safety Administration (FMCSA) and the Commercial Vehicle Safety Alliance (CVSA), oversees state enforcement of commercial motor vehicle (CMV; large truck and bus) size and weight standards in the United States. In other countries, similar agencies have similar responsibilities.

Vehicle Length

As opposed to in many other countries, in the United States, there are no federal or state overall length limits for most truck tractor-semitrailer combinations. However, the overall length of containers and trailers is regulated by federal agencies for roads that receive federal funding. In Europe, it is the bumper-to-bumper length that is regulated. This has led to very different designs of truck tractors in the United States and Europe. In the United States, the engine sticks out in front of the cab, and the cab is often long. In Europe, the engine is below the driver, and the cab is very short.

In the United States, combination vehicles (i.e., truck tractor plus semitrailer or trailer) specifically designed to carry automobiles or boats may not exceed a maximum overall vehicle length of 65 or 75 ft. (19.81 or 22.86 m), depending on the type of connection between the tractor and trailer. No state may impose a length limitation of less than 48 ft. (14.63 m) on a semitrailer (single) or less than 28 ft. (8.53 m) on a semitrailer (twin-trailer) operating in any truck tractor-semitrailer combination under federal regulations. In Europe, countries regulate the maximum length of trucks, and the maximum length varies from country to country.

Vehicle Width

In the United States, no state may impose a width limitation of less than 102 in. which is 2.59 m. That is also the widest truck allowed on most US roads without a special permit except in Hawaii where trucks can be 108 in. (2.74 m) wide. Safety devices, such as mirrors and handholds, are not included in the calculation of width. In Sweden, the maximum allowed width is 2.60 m, again allowing safety devices to extend beyond that width. In some other EU countries, vehicles are not allowed to be wider than 2.55 m without special permits.

Vehicle Height

US State standards range from 4.11 to 4.57 m (13.6–15 ft.). No US federal vehicle height limit exists. In Sweden, the clearance has to be signed if it is less than 4.5 m (14.75 ft.), so that is the de facto allowed maximum vehicle height. In other European countries, only 4.0 m (13.1 ft.) vehicles are allowed.

Vehicle Weight

Federal CMV maximum standards on the interstate highway system and highways are 20,000 lb (9,000 kg) for single axle, 34,000 lb (15,400 kg) for tandem axle, and limit maximum gross vehicle weight to 80,000 lb (36,300 kg). Additionally, the FHWA regulates the weight distribution per axle as shown in [Fig. 1](#). In 1975, FHWA introduced the bridge formula to reduce the risk of damage to highway bridges by requiring more axles, or a longer wheelbase, to compensate for increased vehicle weight. The formula may require a lower gross vehicle weight, depending on the number and spacing of the axles in the combination vehicle.

The research on oversized and overweight loads suggests that these trucks do significantly more damage to highway infrastructure, such as pavements and bridges ([Luskin and Walton, 2001](#)). Many bridges were not built to handle higher weights and should be strengthened to withstand heavier truck and trailer weights. Additionally, oversized and overweight trucks are harder to steer and control and may be at higher risk of a serious crash.

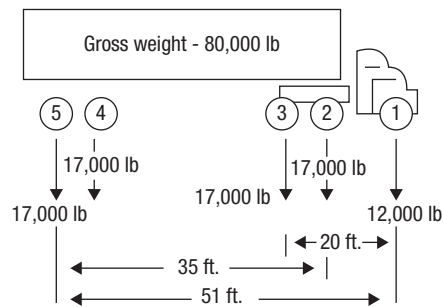


Figure 1 Commercial motor vehicle weight distribution.

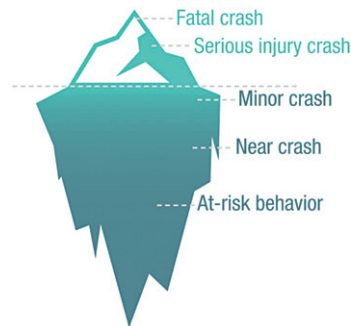


Figure 2 Crash causation pyramid.

Truck Crash Causation

Decades of research have consistently found driver behaviors to be the primary contributory factors in approximately 90% of large truck crashes, just like driver behavior contributes to at least 90% of passenger car crashes ([Federal Motor Carrier Safety Administration, 2006](#)). These driver behaviors include purposeful risky driving, such as driving faster than the speed limit, traveling too fast for conditions, and tailgating. However, other driver behaviors that contribute to crashes include inattention, poor performance, inappropriate decision-making, and medical conditions. In spite of this, in multiple-vehicle collisions involving a truck and a passenger vehicle, it is more frequently the behaviors of the passenger vehicle driver, not the large truck driver, that lead to the crash.

Prior to being involved in a crash that resulted from a particular behavior leading to a crash, the driver likely performed that same error or risky behavior at least once, if not many times, before. In fact, it is likely the driver performed the risky behavior or error hundreds or thousands of times before a crash occurred. These behaviors are the leading indicators of crashes and should be the primary focus in efforts to prevent large truck crashes. Unfortunately, these leading indicators are often not measured. Instead, lagging indicators are more commonly used to measure large truck safety. Lagging indicators are reactive and are the result of unsafe performance. Examples of lagging indicators include violations, minor crashes, injury crashes, and fatal crashes ([Fig. 2](#)).

Although lagging indicators are useful in measuring the overall effectiveness of safety countermeasures, they should not be the only measure of safety or countermeasure effectiveness. Instead, research should focus on eliminating risky behaviors in order to reduce the number of near crashes, minor crashes, and serious crashes.

Risk Factors for Large Truck Crashes and Effective Strategies to Improve Safety

There are three high-level strategies to prevent large truck crashes: education, enforcement, and engineering ([Fig. 3](#)). Education countermeasures focus on educating the large truck driver and/or large truck fleets on safe driving practices and their importance. Examples of education countermeasures include commercial driver license training programs, health and wellness programs, and individual driver coaching. Enforcement countermeasures focus on providing consequences for not adhering to safety-related laws or regulations. Examples of enforcement countermeasures include roadside vehicle inspections and the enforcement of laws and regulations related to distracted driving or speeding. Engineering countermeasures use engineering principles to redesign the roadway or vehicles to reduce crashes. Common engineering countermeasures include the development of advanced driver assistance systems (ADASs), infrastructure maintenance, and improved roadway and vehicle lighting.

The following sections provide high-level overviews of risky behaviors related to large trucks, and effective education, engineering, and enforcement strategies to prevent large truck crashes.

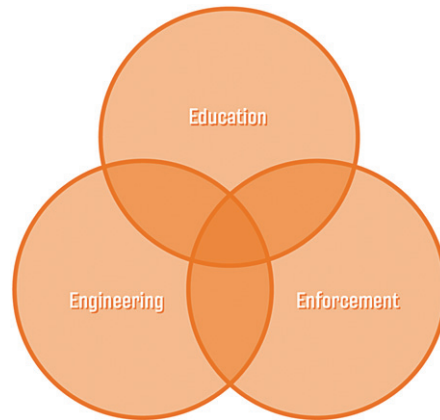


Figure 3 Three strategies to reduce crashes.



Figure 4 Impact of safety culture.

Fleet Safety Culture

Research has long shown a connection between employees' safety performance and organizational safety culture in manufacturing and industrial environments. More recently, research confirmed the connection between large truck fleets' safety culture and their safety performance (Camden et al., 2019). Each fleet has policies and programs aimed at reducing crashes and violations. These policies and programs dictate the countermeasures used to prevent crashes. If effective, these countermeasures reduce crashes, which improve fleet safety and further reinforce the safety culture of the fleet. Conversely, insufficient policies and programs related to safety may actually encourage unsafe driving, which could lead to a fleet developing poor safety culture (Fig. 4).

Although all employees in a fleet create the safety culture, safety culture begins at the top of the organization. This includes the owner, executives, and upper management. Upper management can demonstrate their commitment to safety through regular attendance at safety meetings, frequent communications centered on safety performance, the sharing of fleet-wide safety data, providing safety leadership, and investing in safe equipment and ADAS. Although a strong safety culture begins with buy-in from upper management, driver accountability for safe driving and buy-in to the safety improvement process is critical. The most effective means to generate driver buy-in is to provide positive, supportive feedback on safe and unsafe driving habits, trust that drivers will do their best to perform their job safely, actively listen to feedback provided by drivers, and allow drivers to develop their own goals related to safety.

Steps to Develop a Strong Safety Culture

Improving fleet safety culture is a lengthy and difficult process, but with dedication and effort, it is possible. Fleets should follow a four-step process to develop a strong safety culture. First, the fleet needs to assess its current safety culture. Fleets should review current safety-related policies and programs, examine how the fleet currently measures safety, identify how safety-related messages are currently communicated, and examine barriers that create challenges to improved safety. The second step involves using the information from the first step to perform a gap analysis to identify areas in need of improvement. The outcome of this analysis will serve as the fleet's road map to improved safety. The third step is to design strategies to address all areas in need of improvement. These strategies may include developing high-level safety-related goals, creating new policies and programs to communicate about safety across the fleet, developing improved driver education/training, identifying new safety-related measures, and implementing safety-related technologies. Finally, once these strategies are developed, the fleet should implement, track, and measure their effectiveness.

Large Truck Driver Inattention

Distraction from secondary tasks, fatigue, drowsiness, impairment from drugs and alcohol, and medical conditions that cause drowsiness (such as obstructive sleep apnea and narcolepsy) are all topics under the inattention umbrella. Here we define and describe distraction from secondary tasks, fatigue and drowsiness, and impairment from drugs and alcohol.

Distraction From Secondary Tasks

Secondary tasks are nondriving tasks or, more specifically, those tasks that are not related to the safe operation of the vehicle (texting, dialing a cell phone, talking/listening on a cell phone, using a dispatching device, adjusting the radio, etc.). A common sense definition of distraction is “any task that requires the driver to take attention away from the driving task.” However, this definition is misleading, as it implies that anything that takes a driver’s attention away from the driving task is dangerous. A better definition for driver distraction is “a mismatch between the current attentional resources and those necessary to safely operate the vehicle.” This definition implies that some secondary tasks are risky and others are not, a concept supported by naturalistic truck driving research. Most prior research on driver distraction focused on passenger car drivers, with a few studies including large truck drivers. A naturalistic truck driving study focused on the prevalence and risk of secondary tasks in a sample of large truck drivers found that those secondary tasks that required the driver to take their eyes away from the forward roadway increased risk, and those secondary tasks that only required cognitive effort did not increase risk (Olson et al., 2009). Fig. 5 illustrates that the blue line is the odds ratio (the higher the odds ratio, the greater the risk) and the red line is the amount of time looking away from the forward roadway. Texting while driving a large truck increased risk by 23 times, as it was associated with driving while looking away from the forward roadway for a long time. Talking/listening on a hands-free cell phone while driving a large truck actually reduced risk, as it was associated with the driver looking at the forward roadway. Interestingly, large truck drivers performed many work tasks while driving that were considered dangerous, such as interacting with a dispatching device, writing on pad/notebook, using a calculator, and reading. Research, policy, and training and education have largely focused on nomadic electronic device use (e.g., cell phone); however, it appears that this population is also prone to multitasking that involves work-related secondary tasks. Many technologies can prevent and/or mitigate the harmful effects of distracted driving, including crash avoidance systems (e.g., automatic emergency braking, lane departure warning, side collision warning, etc.) and onboard monitoring systems.

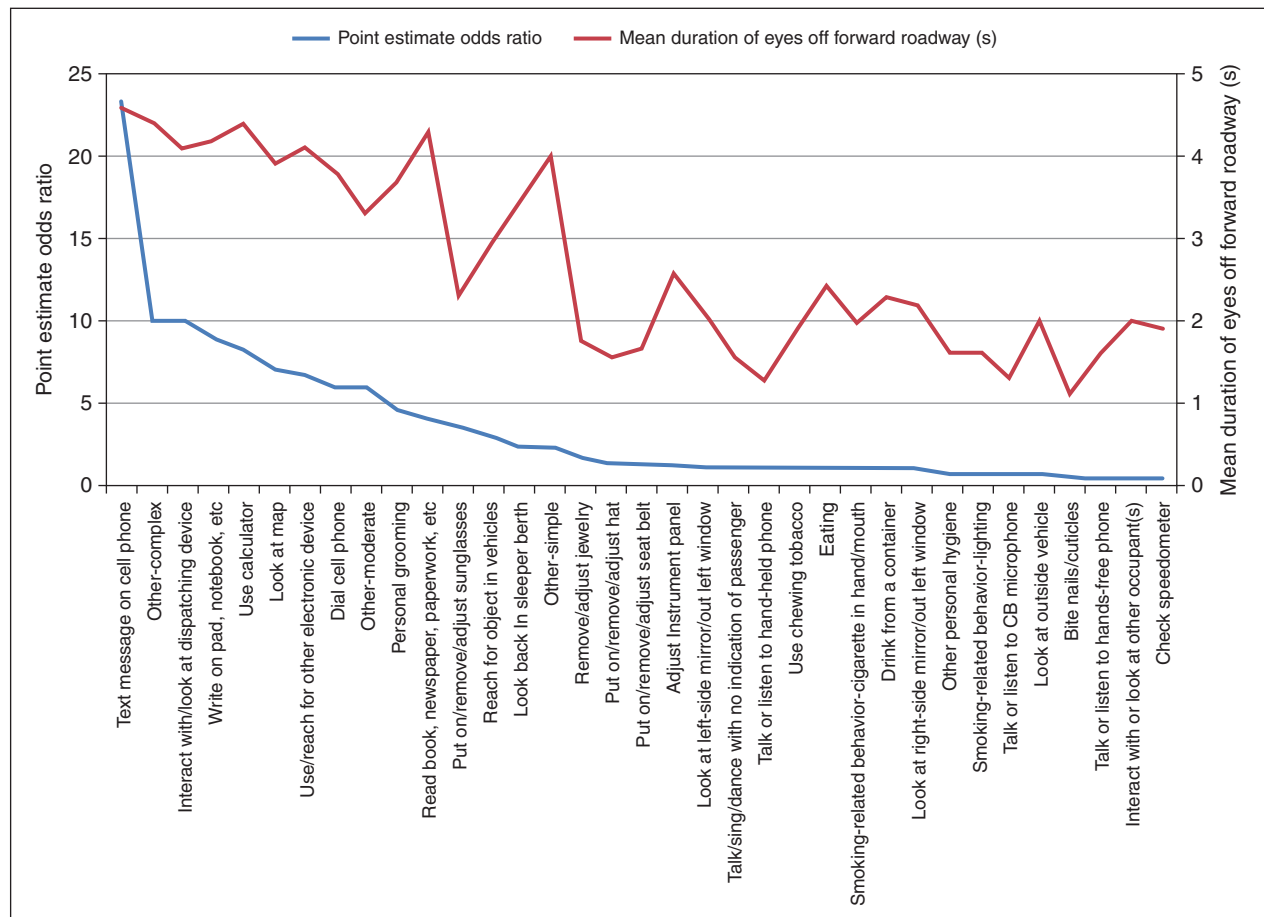


Figure 5 Risk of commercial motor vehicle driver distraction.

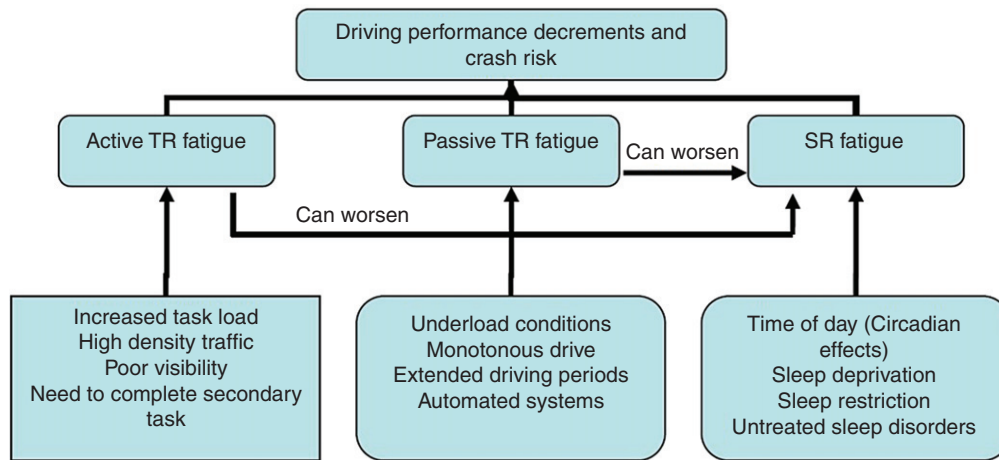


Figure 6 Driver fatigue.

Driver Fatigue and Drowsiness

The terms driver fatigue and drowsiness are often used interchangeably because their consequences are relatively similar; however, their etiology is different. Drowsiness is related to sleep quantity and quality, whereas fatigue is related to the driving task. Task-related fatigue can be passive (e.g., driving for a long time) or active (heavy loading/unloading activity). **Fig. 6** illustrates these concepts and some of the factors that influence drowsiness and fatigue. Large truck driver fatigue and drowsiness can interact, compounding their impact on driver performance (e.g., limited amount of sleep and driving for a long time). Drowsiness and fatigue are difficult to define and measure; therefore, it is difficult to assess and regulate how to avoid driving while fatigued. As fatigue is difficult to assess through postcrash reconstruction, estimates of the frequency and outcomes of fatigued driving are likely conservative (estimates are 10%–20% of fatigued driving involvement in CMV crashes). Large truck drivers may be hesitant to disclose their level of drowsiness while driving, or the severity of the crash may leave the driver too incapacitated to report this information. Fatigued drivers are more likely to be involved in crashes and are more likely to be involved in higher severity crashes, as their reaction times are often delayed and/or they do not initiate crash avoidance maneuvers (NASEM, 2016). To combat driver fatigue and drowsiness, the FMCSA places limits on drivers' working and driving hours. However, as **Fig. 6** shows, this addresses only one of the factors influencing large truck driver fatigue and drowsiness. Addressing large truck driver drowsiness and fatigue requires a comprehensive approach that includes developing a corporate culture that facilitates reduced driver fatigue; fatigue management education for drivers, drivers' families, dispatchers, carrier executives and managers, and shippers and receivers; screening for and treating sleep disorders; and driver and trip scheduling. The North American Fatigue Management Program addresses all of these topics and can be found at www.nafmp.com.

Impairment From Drugs and Alcohol

Although previous research shows that approximately 30% of all fatal crashes in the United States involved a driver impaired by drugs and/or alcohol, only 3% of these involved CMV drivers. This lower rate may be because FMCSA requires all CMV carriers to have drug and alcohol testing programs. Drug testing through urinalysis is limited to five drugs (marijuana, cocaine, amphetamines, opiates, and phencyclidine). The random testing rates are 50% for controlled substances (drugs) and 10% for alcohol, with about 1% of CMV drivers testing positive for controlled substances and 0.2% for alcohol use. Thus, it appears mandatory drug testing has largely been effective in reducing driving while drugged and/or intoxicated. However, less is known about over-the-counter, prescription, and synthetic drugs (e.g., synthetic marijuana), which may cause CMV driver impairment, but are not part of these testing programs. The Large Truck Crash Causation Study found that 30% and 19% of CMV drivers involved in serious crashes had an associated factor of prescription drug use and over-the-counter drug use, respectively (Federal Motor Carrier Safety Administration, 2006). This may seem like a lot, but to find a causal relationship would require knowing the time the drug was taken in relation to the crash; whether the drug adversely affects performance, attention, and/or decision-making capabilities; and whether the crash was related to the drug's adverse side effects (i.e., not something else, such as speeding). There is no information on the prevalence of synthetic drug use among CMV drivers. Although drug-testing programs are available to identify synthetic marijuana and amphetamines, manufacturers alter the chemical makeup of these drugs at a high rate to evade regulation and detection. Currently, drug tests are conducted through urinalysis. This method can detect drug use in the last 1–3 days. Hair testing, which is permissible for employers to require in addition to urinalysis, can detect drug use in the last few months.

Driver Health and Wellness

Large truck drivers endure a number of physical and psychological stresses inherent to their occupation. The typical lifestyle of a large truck driver may include irregular work and sleep hours, physical inactivity, prolonged occupational sitting, poor eating habits and

nutrition, continuously changing work schedules, and mental and physical stress. These combined factors have major implications for drivers' health, compromising not only their safe roadway performance but also their long-term health. A previous study of 88,246 CMV drivers found that 53.3% of CMV drivers were obese, with 26.6% being morbidly obese, and that 10.2% had diabetes (all outpacing the US prevalence rates (Thiese et al., 2015)). This poor health status of large truck drivers is commonly attributed to lifestyle and occupational factors (e.g., shift work, sedentary work, poor sleep hygiene, improper diet, inadequate exercise, etc.). Unlike passenger car drivers, large truck drivers have more stringent physical qualifications. There are certain physical abilities and medical requirements essential for driving a large truck. These drivers are required to obtain a medical examination card, performed by a certified medical examiner, prior to driving a CMV. The medical examination card can be valid for up to 2 years, though the medical examiner can recommend more frequent medical exams based on a driver's medical history. The FMCSA set the physical qualifications standards for large truck drivers to prevent individuals with certain medical conditions from operating a CMV. There are several disqualifying conditions, including epilepsy, seizure disorders, and vision and hearing impairment. Health and medical conditions, including obesity, cardiovascular disease, diabetes, obstructive sleep apnea, and musculoskeletal injuries, have been demonstrated to increase crash risk. However, treatment of these conditions typically lowers this risk to the level of non-diagnosed drivers. Thus, large truck carriers should focus on prevention and treatment of these medical conditions to increase the health of the large truck driver population (as well as reduce crash risk).

Speeding

Compared to light vehicle drivers, large truck drivers are more likely to follow posted speed limits. However, when these drivers do travel above the posted speed limit, or when they travel too fast for conditions, they are more likely to be involved in a crash. In fact, traveling too fast for conditions is the second most common contributing factor in large truck–light vehicle crashes. Furthermore, when a speed-related crash occurs between a large truck and a light vehicle, the collision will be more severe than a lower speed crash. As large trucks in the United States can weigh up to 36.28 metric tons (80,000 lb), the damage resulting from a crash increases exponentially the faster the large truck is traveling.

In an effort to address the deadly consequences of speed-related crashes involving large trucks, these trucks are typically equipped with speed limiters. Speed limiters (also known as speed governors) work to limit the speed of the large truck to a predetermined maximum speed. Although there are no current federal regulations mandating the use of speed limiters or dictating the speed to which they are set, many fleets have policies regarding their use. In a previous study, researchers evaluated the efficacy of large truck speed limiters in preventing crashes (Hanowski et al., 2012). This study analyzed data from over 150,000 large trucks and 28,000 crashes. Large trucks governed by a speed limiter had a speed-related crash rate of 1.6 crashes per 100 trucks/year, and trucks that were not governed by a speed limiter had a speed-related crash rate of 2.9 crashes per 100 trucks/year. This represented statistically significant differences where trucks governed by a speed limiter were involved in fewer speed-related crashes compared to trucks not governed by a speed limiter. Other research has also shown that telematics-based devices, when paired with driver feedback and management follow-up, can significantly reduce speeding.

Advanced Driver Assistance Systems

ADASs, sometimes referred to as adaptive driver assistance systems, are electronic systems that provide safety-related support to a driver. Though several types of ADAS technologies exist, the overall purpose of an ADAS is to promote safety. Most new vehicles manufactured today include the option to use some type of ADAS technology, typically as part of a safety or electronics package. Adaptive cruise control, forward collision warning, lane departure warning, adaptive headlights, blind spot detection, and automatic emergency braking are just a few of the ADAS options that are available. Though research is ongoing with respect to the measured safety benefits for these ADAS options, research to date has been positive (Fig. 7).

Basic ADASs provide timely information to support the driver in making safe, informed decisions. Embedded sensors focus on a specific application to support driver decision-making. For example, blind spot detection systems monitor traffic in the lateral spaces adjacent to the vehicle and provide an indicator, usually on the side mirror, informing the driver if there are vehicles in the blind



Figure 7 Blind spot detection.

spot. This information lets drivers know that it would be unsafe to change lanes at that moment (when the side-mirror indicator is “on”). More updated ADASs provide interventions (beyond information only). Automatic emergency braking systems, for example, monitor traffic in front of the vehicle, calculating time-to-collision. If the calculated time-to-collision falls beneath a predetermined threshold, which would indicate a potential forward crash scenario, a warning indicating the need for the driver to press the brakes is followed by automatic braking by the vehicle if the driver fails to respond. These more updated ADASs will, regardless of the driver’s actions, intervene if an imminent collision is detected.

Automated Driving Systems

A key aspect of these technologies is that individual ADASs provide the building blocks to make autonomous driving possible. That is, individual ADASs that are “layered” together, along with additional sensors, maps, and software, work collectively to support automated driving systems (ADSs).

Consider again the situation with automatic emergency braking: if the driver fails to recognize a forward crash imminent situation, the vehicle will brake *for* the driver. In this scenario, the driver is not in the response loop and is not needed for the brakes to be depressed. ADSs, then, represent a broad array of such technologies that provide control input to safely operate the vehicle.

SAE International has outlined various levels of automation (and control); Fig. 8 describes the six automation levels that range from “no automation” to “full automation.” This spectrum of automation levels highlights the involvement of the human driver. In all but the most advanced levels (4 and 5), the driver remains “in the loop” and, thus, has ultimate control of the vehicle’s operation.

Though many examples of lower levels of automation exist, most industry experts agree that full automation is likely many years from realization. Furthermore, beyond the technological hurdles that must be overcome, implementation of ADS in large truck fleets has many other issues that can only be resolved with the development of a concept of operations. Fig. 9 highlights some of the key contextual aspects that need to be in place before ADS can be implemented on a meaningful scale in the United States.

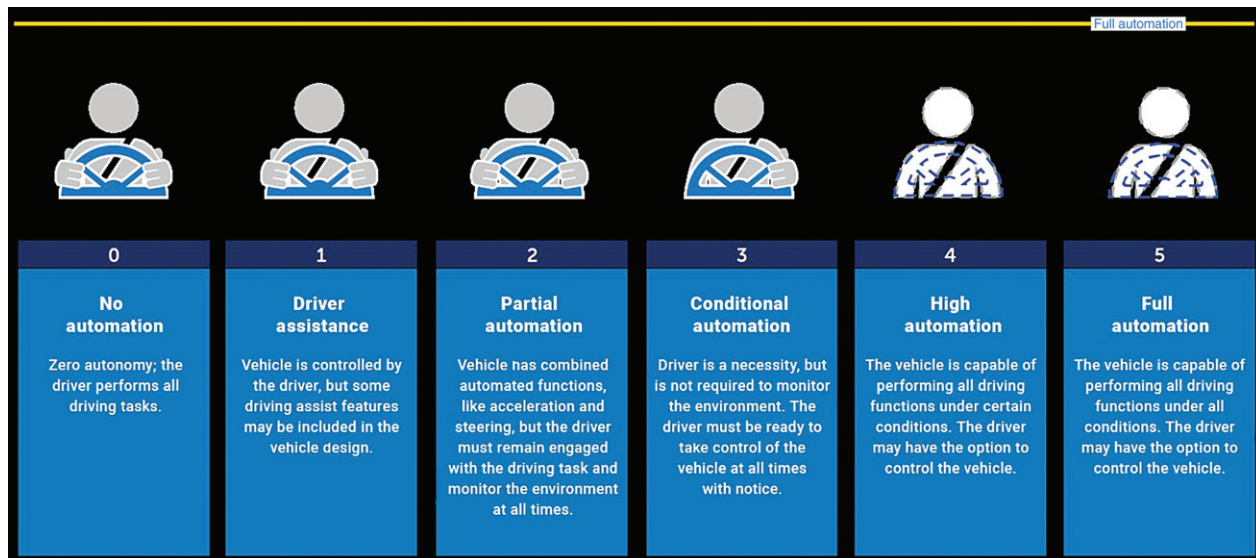


Figure 8 SAE levels of automation.

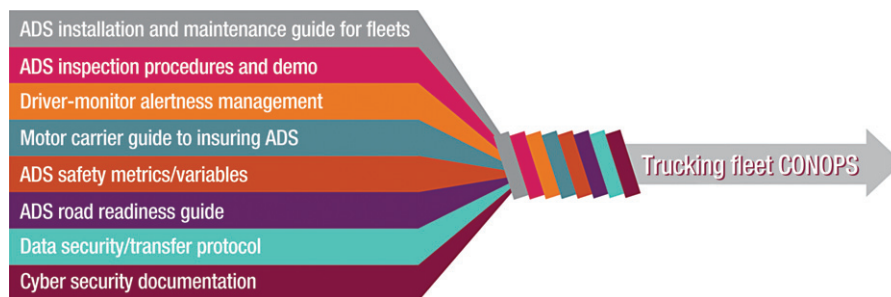


Figure 9 ADS concept of operations.

Inspection Level	Description
I	Includes a review of carrier and driver credentials, logbook or electronic logging device, an inspection of the vehicle's condition, and an inspection of all hazardous materials or dangerous goods.
II	Includes a review of driver credentials, a walk-around vehicle inspection to check components that do not require getting under the truck.
III	Includes only an inspection of driver credentials.
IV	A special inspection of a particular item and is usually conducted as part of research or to investigate a suspected trend.
V	Includes only an inspection of the vehicle that can be conducted without the driver present at any location.
VI	Includes an inspection of transuranic waste and route-controlled quantities of radioactive material.
VII	An inspection jurisdictionally mandated.
VIII	An inspection conducted electronically while the vehicle is in motion, without direct interaction.

Figure 10 CVSA commercial motor vehicle inspections.

Large Truck Roadside Inspections

In the United States, the FMCSA is the federal agency charged with reducing crashes involving large trucks and resulting injuries and fatalities. To do this, FMCSA maintains and enforces the Federal Motor Carrier Safety Regulations. FMCSA has established several tools to support its enforcement efforts, including the compliance, safety, accountability program, and the safety measurement system. The safety measurement system ranks motor carriers based on their safety performance, which is determined using data collected during roadside inspections (e.g., the frequency of different types of violations and the frequency of crashes involving the carrier). This system was developed to prioritize unsafe, high-risk motor carriers for targeted interventions.

Currently, in the United States, approximately 3.5 million inspections are conducted annually to ensure the safety of large trucks and buses operating on our roadways. Highly trained inspectors in each state inspect CMVs using inspection procedures developed by the CVSA. These procedures and criteria are part of the North American Standard Inspection Program and currently include eight levels of inspection. Level I, which is the most common, involves the examination of the driver's credentials and record of duty status along with a detailed inspection of the vehicle's mechanical condition. **Fig. 10** provides the CVSA levels of inspection and associated procedures.

Violations arising from these inspections are recorded in the Motor Carrier Management Information System. FMCSA uses this system's data to identify carriers that are out of compliance with federal regulations and good candidates for targeted safety interventions. The Motor Carrier Management Information System contains carrier registration details, information from inspections and interventions, and violation and crash data. Research on the effects of roadside inspections has shown a strong relationship between quality maintenance and inspection procedures and a decline in crashes related to vehicle defects. Research found that the defects most likely to cause crashes were those associated with brakes, tires/wheels, and lights. Additionally, vehicle inspections and application of the out-of-service criteria through the roadside inspection program have significantly decreased the rate of truck crashes in which mechanical or safety defects were cited as a primary contributing factor ([Lantz, 1993](#)).

Conclusion

There is a growing body of evidence that the safety countermeasures described earlier prevent large truck crashes and their associated injuries and fatalities. These countermeasures include management practices to improve fleet safety culture, education and training programs targeting driver health and wellness, safety-related devices to reduce large truck speeding, ADAS to assist drivers and control the vehicle, and enforcement activities to improve vehicle maintenance. Although research showed these countermeasures to be effective in preventing crashes, fleets should focus on a comprehensive mix of countermeasures rather than viewing any single countermeasure as a "quick fix."

See Also

Connected Automated Vehicles: Technologies, Developments and Trends; Driver distraction; Driver state and mental workload; Safety culture; Sleep-related Issues and fatigue; Speed governors and limiters; Behavioral research in freight transport; Road freight vehicles; Autonomous Goods Transport

Relevant Websites

Available from: www.fmcsa.dot.gov.

Available from: <https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety>.

Available from: www.nafmp.com.

Available from: <https://cvsa.org>.

References

- Camden, M.C., Hickman, J.S., Hanowski, R.J., 2019. Effective Strategies to Improve Safety: Case Studies of Commercial Motor Carrier Safety Advancement. The National Surface Transportation Safety Center for Excellence, Blacksburg, VA.
- Federal Motor Carrier Safety Administration, 2006. Report to Congress on the Large Truck Crash Causation Study (Report No. MC-R/MC-RRA). National Technical Information Service, Springfield, VA.
- Hanowski, R.J., Bergoffen, G., Hickman, J.S., Guo, F., Murray, D., Bishop, R., Johnson, S., Camden, M., 2012. Research on the Safety Impacts of Speed Limiter Device Installations on Commercial Motor Vehicles: Phase II Draft Final Report (Report No. FMCSA-RRR-12-006). Federal Motor Carrier Safety Administration, Washington, DC.
- Lantz, B.M., 1993. Analysis of Roadside Inspection Data and Its Relationship to Accident and Safety/Compliance Review Data and Reviews of Previous and Ongoing Research in These Areas. Upper Great Plains Transportation Institute, North Dakota State University, Fargo, ND.
- Luskin, D.M., Walton, C.M., 2001. Effects of Truck Size and Weights on Highway Infrastructure and Operations: A Synthesis Report (Report No. FHWA/TX-0-2122-1). Texas Department of Transportation, Austin, TX.
- National Academies of Science, Engineering, and Medicine, Engineering, and Medicine, 2016. Commercial Motor Vehicle Driver Fatigue, Long-term Health, and Highway Safety: Research Needs. The National Academies Press, Washington, DC DOI: 10.17266/21921.
- Olson, R.L., Hanowski, R.J., Hickman, J.S., Bocanegra, J., 2009. Driver Distraction in Commercial Vehicle Operations: Final Report. Contract DTMC75-07-D-00006, Task Order 3. Federal Motor Carrier Safety Administration, Washington, DC.
- Thiese, M.S., Moffitt, G., Hanowski, R.J., Kales, S.N., Porter, R.J., Hegmann, K.T., 2015. Repeated cross-sectional assessment of commercial truck driver health. *J. Occup. Environ. Med.* 57 (9), 1022–1027.

Further Reading

- Grove, K., Atwood, J., Blanco, M., Krum, A., Hanowski, R., 2017. Field study of heavy vehicle crash avoidance system performance. *SAE Int. J. Trans. Safety* 5 (1), doi:10.4271/2016-01-8011.
- Hickman, J.S., Guo, F., Camden, M.C., Hanowski, R.J., Medina, A., Mabry, J.E., 2015. Efficacy of roll stability control and lane departure warning systems using carrier-collected data. *J. Safety Res.* 52, 59–63. Available from: <https://www.sciencedirect.com/science/article/pii/S0022437514001145>.
- Krum, A., Bowman, D.S., Soccolich, S., Deal, V., Golusky, M., Joslin, S., Miller, A., Hanowski, R.J., 2015. Federal Motor Carrier Safety Administration's Advanced System Testing Utilizing a Data Acquisition System on the Highways (FAST DASH), Safety Technology Evaluation Project #2: Driver Monitoring, Final Report (FMCSA-RRR-16-002). Federal Motor Carrier Safety Administration, Washington, DC.
- Ludwig, T.D., Geller, E.S., 2001. Intervening to Improve the Safety of Occupational Driving: A Behavior-Change Model and Review of Empirical Evidence. The Hawthorne Press, Inc., Binghamton, NY.
- United States Department of Transportation, 2018. Preparing for the Future of Transportation: Automated Vehicles 3.0. USDOT, Washington, DC. Available from: <https://www.transportation.gov/av>.

Demerit Points and Similar Sanction Programs

Matúš Šucha, Kristýna Josrová, Palacký University in Olomouc, Olomouc, Czech Republic

© 2021 Elsevier Ltd. All rights reserved.

Demerit Point Systems and Their Variations	210
Demerit Point Systems as a Preventive, Selective, and Corrective Measure	210
Underlining Theories	211
Driver Improvement Measures and Intermediate Measures Within Demerit Point Systems	212
The Effects of Demerit Point Systems in Terms of Lower Rates of Accidents and Fatalities	212
Recommendations for the Implementation of a Demerit Point System	213
References	215

Demerit Point Systems and Their Variations

A demerit point system (DPS) is a traffic law enforcement practice aimed at increasing road safety by preventing and correcting dangerous behavior in traffic. The point systems implemented in many countries around the globe take different forms in each of them. Locally, a DPS can also be referred to as a point system or penalty point system. Penalty points usually accompany other types of sanctions, such as fines or the possibility of imprisonment. Every traffic law offence included in the point system is recorded within the system for each driver. Misconduct on the part of a driver leads to them receiving or losing a certain number of points for their transgression and, after the driver reaches a critical number of penalty points, penalization follows, which can ultimately lead to the suspension of a driver's license. The license can be returned after a certain period of time has passed and sometimes only after the driver meets necessary requirements, such as reeducation, assessments, or tests. The advantage of the system lies in its fairness for drivers, who are granted equal treatment, while also allowing specific high-risk groups to be focused on, so that, for example, habitual offenders can be penalized more harshly for incessant repeated offences, which can result in lower tolerance within the system (gaining more points per offence or the suspension of the license being of a longer duration each time). It must be noted that this fairness has its limitations as a result of the following facts. There might be a bias on the part of police officers toward certain drivers when charging them; for example, a police officer may be more likely to charge young drivers with speeding than older ones or decide to give oral warnings in some cases and tickets leading to demerit points in others. Additionally, the randomness of Poisson distribution has to be taken into account. This means that in reality some drivers get more or fewer fines (and points) even though they commit the same number of violations as an "average" driver (Evans, 2004). The equality of treatment has justified exceptions regarding certain groups, such as novice or professional drivers. DPSs are especially widespread in Euro-Atlantic countries, but are becoming popular worldwide. To date, such systems have been implemented in the vast majority of European Union member states, in Australia, Canada, New Zealand, and the United States of America, but also on other continents, in countries such as Brazil, South Korea, Malaysia, Singapore, Morocco, the United Arab Emirates, and South Africa. When it comes to DPSs, no two countries are alike. In some countries, drivers start with zero points and accumulate penalty points for each offence, while in other countries they are given a fixed number of points at the beginning and demerit points are then subtracted for each traffic law violation. Countries and states using DPS have lists of offences that are penalized by a given number of penalty points being assigned. In some countries, each offence has the same weight and the driver receives one point per offence, while in other countries different transgressions are penalized by different numbers of points being assigned; typically, the more dangerous the behavior (such as speeding), the more severe the punishment. In the case of speeding, in some countries, the law takes account of the speed in excess of the legal limit. Typically, the number of points given for a specific offence is set by the law, but in some countries, the number of points assigned can be flexible and might also depend on a court decision. Countries also differ in the number of demerit points resulting in drivers having their license suspended, in the types of offences subject to penalty points, in the types of rehabilitation programs that make it possible to erase or acquire points, as applicable, and in the process leading to the restoration of the driving license. Certain countries distinguish between various driver groups, such as professional and novice drivers, who come under special terms and conditions, and some even use DPS only for novice drivers (e.g., the Netherlands). In addition to the earlier criteria, countries also differ in terms of the complexity of the system, the authorities involved in the process, the transparency of the system, drivers' access to information about their current point status, and the inclusion, or exclusion, of foreigners in the DPS, as well as in whether nondrivers are included in the system.

Demerit Point Systems as a Preventive, Selective, and Corrective Measure

Road risks stem from human error rather than from the imperfection of the system, as the human factor is involved in 90%–95% of road accidents. Nevertheless, the nature of the human factor is too complex to make it possible to formulate simple conclusions and recommendations. Driving is such a complex and diverse activity that it is impossible to identify several individual qualities that

could help us distinguish safe and responsible drivers from dangerous and irresponsible ones. In addition, drivers are in no way a homogeneous group, which prevents the generalization and interpretation of available evidence. On the contrary, the number of traffic law violations is one of the best single predictors of future involvement in accidents (Elvik et al., 2009). Previous accident involvement is an even better predictor. The DPS works on the principle of improving road safety by deterring road users through the threat of punishment (on a general level) and through sanctions for those who are caught and convicted (on a specific level) (Sagberg and Sundfør, 2019). The specific deterrent effects can be evaluated by the monitoring of the reoffending rate, the general effects by the prevalence of negative phenomena among the driving community (e.g., lower rates of speeding or crashes) (Sagberg and Ingebrigtsen, 2018).

A DPS may affect violations and accidents through three mechanisms: prevention, selection, and correction. The *preventive* effect of a DPS lies in the risk of losing one's driving license if caught offending repeatedly. The general preventive effect should be found in a decrease in the number of offences for all drivers: they drive more carefully in order to avoid getting a demerit point (Sagberg and Sundfør, 2019). The other preventive effect is providing drivers with negative feedback on their risky driving (Sagberg and Sundfør, 2019). The absence of negative feedback leads to drivers having a false belief in the low probability of accidents and in their perception that they have a higher chance of avoiding a road accident in comparison to others. The absence of social negative feedback is addressed by the DPSs.

As for *selection*, if a DPS can exclude road users who often engage in dangerous behavior from participation in traffic before they could actually cause a crash, it can improve road safety (Klipp et al., 2011). But a DPS can only be an effective means of selection if demerit points or offences are indeed a good predictor of future crashes and if repeat offenders are tracked down in time. Particularly serious offences (involving many demerit points) are good predictors of future crashes. This relationship is at its strongest among young novice drivers. Systems in which drivers can have the number of their points reduced by following a driver improvement measure have an educational element that is intertwined with the preventive effect (*correction*) (Klipp et al., 2011). DPSs with educational elements can only work if it is proved that a course is effective and results in a change of behavior in a desired direction.

An added benefit of a DPS is that it takes recidivism into account: repeat violators of the rules are punished more severely.

Underlining Theories

In terms of its theoretical framework, a DPS can be designed and implemented effectively on the basis of what is referred to as the *3E* model (including its extended versions), which proposes approaching road users from three different perspectives: *Education and Training*, *Engineering*, and *Enforcement*. If we want to achieve a real change in road users' behavior (as regards better safety, or the elimination of risky behavior), we need to exert an influence on road users on all three of these levels. In terms of DPSs, the key areas are Education and Training and Enforcement.

Empirical evidence shows that effective and just enforcement has an impact on road users' behavior with regard to the reduction of risky behavior. The following conditions must be met for enforcement to be effective:

- People must know the rules (i.e., they need to know which behavior is correct and expected).
- They must be able to apply such knowledge (i.e., they need to possess relevant skills and be able to behave according to their knowledge).
- The benefits of breaking the rules must be smaller than the loss ensuing from the imminent sanction (i.e., they need to be motivated).

The positive effect of enforcement on road safety is an increasing function of certainty (that one will eventually be caught), strictness (the sanction being significant enough for the violator), and the imminent threat of being sanctioned (a person cannot avoid punishment).

The concept of deterrence is based on the principle that "prevention is better than punishment." However, deterrence is never the only and primary aim of legislation (the legal system). The substance of deterrence is the threat of punishment taking the form of a certain sanction. Deterrence is a way of gaining control on the basis of fear. If drivers commit violations without being concerned about the consequences, then deterrence is ineffective by definition. In general, deterrence involves behavior control, which occurs in a situation where a potential perpetrator is not willing to take the risk of certain behavior out of fear of its consequences. Deterrence is thus one of the mechanisms for achieving compliance.

Punishment deters the potential perpetrator from illegal behavior (a process which may be referred to as *individual prevention*) while also deterring other individuals from potential offending (*general prevention*). It can be assumed that every person knows what is permitted and what is not and is also aware of the consequences of illegal behavior. People's experience of punishment, both indirect (e.g., experienced by a friend) and direct, reduces their inclination to offend, as it enhances their perception of the risk of being sanctioned (Sagberg and Ingebrigtsen, 2018). According to the deterrence theory, punishment is regarded as a means of discouraging people from committing both administrative and criminal offences. The effects of punishment can be divided into several groups.

- The first encompasses the aforementioned theory of general prevention, which further breaks down into negative and positive categories: the negative theory of general prevention proposes that the threat of punishment would discourage the perpetrator from committing an offence (Sagberg and Sundfør, 2019). The positive theory of general prevention proposes that the inevitable

strict punishment imposed within a short time after the offence being committed will strengthen people's trust in the judicial system.

- The second group is represented by the theory of special prevention: it is assumed that strict punishment imposed on the offender according to the law would also be a sufficient deterrent for others.
- The third type is the theory of individual prevention. It is based on the notion that the harm suffered by the perpetrator as a result of the punishment they received would dissuade them from continuing to engage in illegal activities without any additional educational measures being used (Lee et al., 2018).
- The fourth is the rehabilitation theory, which proposes that a violator's illegal acts are due to various factors (including family, upbringing, social circumstances, and psychological traits), which influence their behavior and development. Here, the purpose of the punishment is to provide professional help, which will make it possible for the perpetrator to gain an insight into their characteristics, and behavior and the social influences, which resulted in their offending. The rehabilitation theory involves efforts to reintegrate the perpetrator into society on the basis of a change in their behavior occurring as a result of their understanding that such behavior is socially problematic.
- Finally, there is the restitution theory, which underlines the need to compensate for damage and harm caused by illegal conduct. In this respect, crime is perceived as a threat to society, and harm caused to the victim must be redressed.

In practice, these theories are interlinked in such a way as to ensure that traffic-related sanctions fulfill the purpose of the punitive policy of the state as regards public safety, resocialization of offenders, and compensation for damage done, as well as the deterrence of potential perpetrators.

The basic 3E model can be extended to include other "Es" with relevance to the implementation of the DPS:

Education and Encouragement—that is, raising public awareness about the purpose and implementation of the DPS and seeking to win over the public and build its commitment and engagement.

In addition, *Evaluation* is there to test whether the strategies used are working, particularly in terms of the number of traffic violations and accidents and their levels of severity.

Driver Improvement Measures and Intermediate Measures Within Demerit Point Systems

Both research and empirical studies seem consistent in their conclusions about the compatibility and synergy of DPSs and driver improvement measures. For example, the recommendations issued as part of Bestpoint, the largest European project of its kind, are clear in this respect. A DPS is effective if it incorporates a focus on specific groups, such as repeat offenders and professional drivers, by promoting driver improvement courses for perpetrators of serious violations and for those who use up all their penalty points. In addition, Bestpoint recommends driver improvement measures aimed at attitudes and behavior change rather than knowledge and skills. The key question regarding driver rehabilitation measures and DPSs is whether and why training in dealing with critical situations is a wrong path to take. While driver rehabilitation programs are intended to improve drivers' attitudes and enhance their will to abide by the rules, the basis of safe driving courses is to develop drivers' skills. These measures thus lie at opposite ends of the spectrum of measures aimed at enhancing road traffic safety. No matter how sensible it may sound that teaching people emergency braking or skid control helps in reducing the impact of hazardous situations, long-term experience from other countries indicates the opposite effect. The beginning of the criticism of this type of training dates back to the 1980s. There is a strong empirical evidence showing an increase in the number of skidding-related accidents after this type of training was made a mandatory part of the driving schools' curriculum. Following the introduction of safe driving courses, a great proportion of novice drivers' accidents occurred on slippery roads. The explanation might be that safe driving courses involved training in skills to control the car that resulted in an unreasonable increase in drivers' confidence in their own capabilities. Instead of using these capabilities in emergency situations, individuals who had completed these courses used them under regular circumstances, which led, for example, to their driving at higher speeds. Moreover, graduates of these courses overestimated their abilities; their actual skid control abilities did not differ from those possessed by drivers who had not been on the course. Evidence shows that: (1) drivers, and often also instructors, failed to understand the main point of the course, which is that safe driving under slippery conditions is about avoiding an emergency situation rather than dealing with one; (2) skills learnt in this way do not last for long; (3) the most skilled drivers show much higher accident rates, and, most importantly, (4) effective driving courses feature work with attitudes, risk perception, experience, and knowledge rather than training in skills. Generally, it can be stated that any courses based on training in skills to cope with emergency situations will eventually result in greater accident rates and worse road safety and thus they should not constitute a part of the DPS. Therefore, the European Conference of Ministers of Transport identified these courses as inappropriate for young drivers and, subsequently, the European Road Federation, drawing on long-term evaluations, issued recommendations for the European Commission that these courses should be eliminated from the driver education and training system.

The Effects of Demerit Point Systems in Terms of Lower Rates of Accidents and Fatalities

Experience clearly shows that the introduction of DPSs leads to reduced rates of traffic accidents and road fatalities (Klipp et al., 2011). This effect is particularly visible in the first years following the establishment of the measure but tends to diminish over

time. It was found that DPSs were effective in reducing the rate of road injuries and fatalities by up to 20%. Studies suggest that DPSs result in a reduction in the number of accidents and casualties. Alternatively, the effectiveness of DPSs is indicated by analyses of medical records of road accidents, which also show significant reductions following the introduction of these systems, for example, a reduction in the rate of spinal injuries suffered in road crashes. Overall, meta-analyses and reviews demonstrate positive outcomes of demerit points, especially when they are accompanied by additional measures and public campaigns. However, this effect was found to decline dramatically 18 months after the introduction of these systems. One reason for this may be the reduced effort on the part of governments to support the DPSs with accompanying measures. Without additional supporting measures, people's initial fear of losing their driver's license tends to fade quickly, as the perceived risk of being caught diminishes over time.

The effect of the DPS on preventing recidivism depends on the perception of sanctions by drivers. The most important aspect appears to be the certainty of punishment, while, regarding the general level of deterrence, an increase in the severity of the punishment has only a minor effect on people's repeated engagement in risky behavior and a change in their behavior. The severity of sanctions is an effective factor in the event that the probability of being caught and punished is very high. However, this is particularly difficult with hard-core recidivists, as these display several differences in comparison with other drivers. They consider the risk of being caught as lower, the benefits from committing traffic offences higher, and the losses related to offences lower. Consequently, a large number of drivers with custodial sentences return to prison within 3 years. Regarding drinking and driving, we can see the effect of a lifestyle associated with addictive behavior. Those who drink more also engage in drink driving more frequently, while those with low or medium alcohol consumption are better at refraining from impaired driving.

It must be noted that it is quite complex to determine the effect of a DPS. In addition to the point system, multiple other factors are often involved and these factors change over time and influence the number of crashes. Furthermore, the introduction of a DPS nearly always coincides with an (unfortunately only temporary) increase in enforcement and with public campaigns. It is therefore impossible to establish the extent to which the effects can be attributed to each factor. Accordingly, it is very difficult, in practice probably impossible, to evaluate the effects of a DPS in a way that could be regarded as scientifically rigorous. First, the introduction of a DPS is generally accompanied by significant media coverage and an increase, or proclaimed increase, in enforcement. Sometimes, it also goes together with new rules and regulations. In addition, since a DPS is always introduced on a nationwide scale, it is impossible to include a good reference group. Consequently, it will remain largely unknown whether observed changes in behavior or in the number of casualties are the result of the DPS or whether some or all of such changes should be attributed to the enforcement, publicity, or other unknown factors.

DPSs may also have some undesirable side effects that should be taken into account. The first one is driving without a driving license. If a penalty hits a motorist hard, but the enforcement of the penalty is weak, the driver will soon be inclined to ignore the penalty. In some countries, up to 40% of the drivers whose driving licenses had been suspended because of the DPS admitted in a survey to continuing to drive. The second one is hit-and-run incidents. The inclination to drive on after causing such an incident will increase out of fear of receiving additional points. No objective figures about the size of these undesirable side effects are available. Finally, there is the buying and selling of points (also referred to as "points trafficking"). If a DPS takes account of automatically detected offences as well, and the driver cannot be identified, it must be found out from the registered license holder who the driver was. This may lead to paying others with no or few points or nonactive drivers who are willing to take the blame for the offence. Objective figures about the size of this problem are lacking.

Recommendations for the Implementation of a Demerit Point System

It is important to have public support for a DPS to ensure its proper functioning. When people understand the principles and agree with them, they are more likely to be compliant with the rules as proposed. If it is otherwise, they might be repulsed by the unwanted. This can be explained by the concept of resistance, when imposed rules boost disobedience merely through one's feeling that one's freedom is being restricted. Typically, the public favors penalty systems, but, to keep the system popular, we should strive to create and keep equal and transparent conditions for all drivers. It is also essential to communicate the beneficial aspects of such systems, as this can help in sustaining their effectiveness. The system should not be too complex—all drivers should be able to understand it (it should be noted that many different people become drivers). In this way, better compliance can also be achieved when the system is viewed as fair.

Although everybody should perceive the system as fair, it is necessary to focus on those who pose the greatest threat to road safety, the minority of drivers who are more likely to take risks and violate the rules, specifically novice young drivers, habitual offenders, and those who drive under the influence of alcohol and drugs. Being responsible for a great proportion of traffic incidents, young drivers represent a high-risk group of drivers. Some countries use DPS for novice drivers only for a certain number of years after they obtain their license, while others have special, stricter conditions for this group, including a smaller number of points before a license is suspended. This measure has an important educational aspect: having to deal with stricter rules from the beginning of their driving careers, they have a better chance of developing good driving habits. Carefulness in driving aimed at not losing points can make them less likely to engage in dangerous driving. Their careful driving attitude can then serve as a good foundation that is more effective than initial leniency followed by later efforts to teach the right behavior when driving habits have already become established. Regarding professional drivers, it is recommended that offences peculiar to this group of drivers, such as compliance with rest times and good technical condition of the vehicle, are included. In addition to

novice drivers, drug and alcohol users represent another high-risk group. These should be tested for substance use and have their licenses suspended in the event that such tests show positive results. While in recovery, these drivers should only be allowed to drive cars equipped with an alcohol interlock device. It is necessary to ensure that they get specialized long-term medical and psychological treatment, as one-off interventions do not suffice. The group of habitual offenders must be dealt with similarly: they need to undergo thorough rehabilitation leading to a behavior change.

Usually, the points are awarded to car drivers or drivers of all motorized vehicles. There is nevertheless discussion among experts as to whether offences committed by other road users should also be registered by the system. Although it would make the system more complicated, the fact remains that all parties, including cyclists, motorcyclists, and pedestrians, must participate in creating a safe traffic environment and each of them should anticipate errors on the part of other road users. Shared responsibility could encourage more consideration and carefulness from drivers, cyclists, and pedestrians and might improve overall road safety. A record in the system of nondrivers having committed a serious traffic offence could be considered in the future when they apply for a driving license; their previous misconduct could even prevent them from obtaining a license.

Given that foreign drivers usually have to pay only fines for traffic offences committed abroad, they are not motivated to exercise greater caution while driving. Having an international register of the points collected by each driver could encourage drivers to become more compliant no matter where they are. The penalty points given in a certain country could be converted to the equivalent number of points which are assigned for specific offences in the driver's home country. In this way, the offence would be transferred into the driver's country of residence, with any implications this may entail.

Demerit points can be assigned for a number of violations. The Bestpoint project suggests penalizing at least those offences directly connected to safety, namely, speeding, driving under the influence of drugs and alcohol, running red lights, overtaking, crossing priority rules, not keeping the minimum distance, the use of mobile phones, dangerous behavior at pedestrian and railway crossings, driving in forbidden lanes or in the opposite direction, involvement in hit-and-run incidents, and not using seat belts, helmets, and appropriate child restraints. The number of penalty points assigned usually reflects the severity of offences, although some countries register one point for every offence. It is recommended to relate the number of points to the degree of the impact they can have on road safety, especially on crashes and the severity of their consequences. This practice would allow the predictive power of the previous violations of the rules to be used and prevent the most dangerous drivers from road use before they can cause harm while also communicating to drivers what the most dangerous behaviors are (Martensen et al., 2018).

Typically, after a certain period of time has passed, a certain number of points are erased. The lifetime of a point should correspond to the severity of an offence and the points should be erased on condition that within a given time period no new penalty point is recorded. The reward in the form of the points being deleted could motivate drivers to become more careful. Moreover, according to the self-perception theory, when drivers are freshly motivated to obey the rules, they begin to view themselves as good drivers, which strengthens their newly emerged good behavior.

When drivers accumulate a certain number of points, it is recommended to take a three-step approach to applying driver improvement and intermediate measures: send out information and warning letters, have their license temporarily withdrawn, and require them to participate in a driver improvement course aimed at bringing about a change in their attitudes and behavior before their license is renewed (Elvik et al., 2009). Sending a warning letter after a certain amount of points has been reached reduces the risk of a crash (Elvik et al., 2009). The drivers are made aware that the withdrawal of their license is a real threat, and some drivers may not even be aware of the fact that they have been given penalty points (Sagberg and Ingebrigtsen, 2018). A similar effect can be achieved by ordering drivers who have reached a certain point threshold to attend obligatory interviews where they have to present arguments explaining why their license should not be suspended. When the point threshold is reached, offenders have their driving license suspended. The suspension period should be neither too short, as it would not have the intended deterrent effect, nor too long, as it could encourage drivers to drive without a driving license. The ideal time is at least 3 months and no more than 1 year. Repeat offenders should have a longer disqualification period every time their license is suspended. License restoration should be tailored depending on the offence, especially with regard to habitual offenders, aggressive drivers, and those who drive under the influence of drugs and alcohol. The reeducation should be proved to be effective, as often it does not bring about the intended change, and the rehabilitation should aim at behavior change, which can be achieved through traffic-specific psychological counseling.

Both communication and enforcement are indispensable for a DPS to be successful and have a positive effect. It will hardly be possible to maintain the positive effects of the system without a sufficient level of enforcement and the driver's experience that the detection of offences is a real threat. The latter can be achieved by the automated detection of certain offences and increased police involvement in surveillance and testing. People should be aware that it is highly probable that their traffic violations will be detected by the police. An essential element of effective enforcement is the perceived risk of being caught when committing an offence. Obviously, the subjective chance is largely determined by an objective, real chance of being caught. In this respect, the major role played by communication campaigns and media coverage of enforcement operations and their aims and results should be noted. Enforcement should be unpredictable, so that drivers do not have the opportunity to adapt quickly to expected checks and return to their old ways in a short time. Communication means that people must be aware that there is a DPS in place and understand how it works. The campaigns should aim at setting positive social norms, as well as making it clear that offenders will be punished.

To ensure its proper functioning, each DPS must be evaluated for effectiveness. Without it, such a system might turn into a costly measure that falls short of the expected results (Martensen et al., 2018). Countries that are considering implementing a penalty point system should assess whether they are prepared for it in terms of the level of enforcement and campaigns to ensure that drivers will take the system seriously and adjust their behavior accordingly.

References

- Elvik, R., Vaa, T., Høy, A., Sørensen, M. (Eds.), 2009. *The Handbook of Road Safety Measures*. Emerald Group Publishing, Bingley.
- Evans, L., 2004. *Traffic Safety*. Science Serving Society, Bloomfield Hills, MI.
- Klipp, S., Eichel, K., Billard, A., Chalika, E., Dabrowska-Loranc, M., Farrugia, B., Larsen, L., et al., 2011. European demerit point systems: overview of their main features and expert opinions. Deliverable 1. European Commission.
- Lee, J., Park, B.J., Lee, C., 2018. Deterrent effects of demerit points and license sanctions on drivers' traffic law violations using a proportional hazard model. *Accid. Anal. Prev.* 113, 279–286.
- Martensen, H., Diependaele, K., Daniels, S., Van den Berghe, W., Papadimitriou, E., Yannis, G., Talbot, R., et al., 2018. The European road safety decision support system on risks and measures. *Accid. Anal. Prev.* 125, 344–351.
- Sagberg, F., Ingebrigtsen, R., 2018. Effects of a penalty point system on traffic violations. *Accid. Anal. Prev.* 110, 71–77.
- Sagberg, F., Sundfør, H.B., 2019. Self-reported deterrence effects of the Norwegian driver's licence penalty point system. *Transp. Res. Part F Traffic Psychol. Behav.* 62, 294–304.

Driver State and Mental Workload

Dick de Waard*, **Nicole van Nes†**, *University of Groningen, Behavioural and Social Sciences, Department of Psychology, Groningen, The Netherlands; †SWOV Institute for Road Safety Research, Den Haag, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Mental Workload	216
Mental Workload During Driving and Driver State	216
Mental Workload and Driver State in Partly Self-Driving Vehicles	217
Assessment of Mental Workload and Driver State	218
Performance	218
Self-Reports, Subjective Measures	218
Physiology	219
Interpretation of Measures	219
Conclusion	220
See Also	220
References	220
Further Reading	220

Mental Workload

Mental workload is a concept that people can relate to; at the same time, it is not clear whether all have the same concept in mind. For many people the difficulty of a task equals mental workload. In many ways that is not such a bad representation, as difficulty is very subjective. In fact, one of our main points is that; as with physical workload, mental workload depends not only on the task that one has to perform, but also on the person concerned. Therefore the level of mental workload is not constant between or within individuals.

The idea that mental workload can be quantified probably originates from its comparison with physical workload. For physical work, the force required to move an object can be defined in Newton, representing what is required to perform a task. However, it is more difficult to define mental workload. One could try to quantify the information-processing demands (for example, the number of calculations that are required to solve a problem). Crucial, however, is the individual capacity to perform these operations— for physical workload as well; moving 60 kg is easier for a physically fit person than for a fragile child. For physical workload, however, we tend to focus on task demands, while for mental workload we take the capacity to perform a task into account. In other words, with mental workload we must also consider the interaction between task demands (the task to perform) and the capacity to perform the task. The latter, also referred to as mental resources, provides a link to operator state, the background state of an individual. State differs not only between but also within individuals. The same task may require different mental resources for a novice and an experienced person—difficult for the first, easy for the second. Additionally, the capacity to perform a task differs within an individual; for example, after a bad night's sleep the task normally considered routine will be experienced as far more difficult to complete.

This brings us to another important concept in mental workload: effort. Effort is a voluntary process, which can best be described as trying hard. The result of effort can be that the performance is maintained at an acceptable level, that is performance is protected, but internally there are (energy) costs. There are two types of effort: *computational effort*, task-related effort (to deal with the increased demands of a task), and *compensatory effort*, state-related effort (to counteract a deteriorated state). In Fig 1 this is graphically illustrated: it shows the relation between task demands (x -axis) and the level of performance and mental workload (y -axis) (De Waard, 1996).

Above all these considerations is self-regulation: deciding how the different task demands are managed. This can mean accepting a lower level of performance or changing to a more efficient strategy to accomplish the goal.

In sum, task demand is a property of the task to be performed. Mental resources are the capacity to perform a task, and they depend on short- and long-term factors such as experience and operator state. With the investment of mental effort, performance can often be kept at an acceptable level. Investment of effort is an effective mechanism, but it incurs costs so it cannot be continued over long periods of time.

The next section focuses on car driving.

Mental Workload During Driving and Driver State

During a drive, a driver's task demand can vary substantially, for example, the drive can start in a busy city but then continue on a quiet highway. In addition, each individual driver has certain characteristics, some may be novice, some professional, or, the most

likely, somewhere in between. For a novice, protecting performance (driving safely without getting into a collision) can require continuous effort investment taking up most of their mental capacity, while for the professional enough mental capacity is left over to do other things, such as operating the radio. During a journey, driver state is also likely to change. Driver state is the driver's mental and physical condition, often reflected in physiological parameters such as an electroencephalogram (EEG). The environment plays an important role here. While monotonous, quiet highways at night are ideal circumstances for driver state to deteriorate, the opposite—when roads are extremely busy—could overload drivers. Imagine being overloaded by information in an unfamiliar city and trying to find your way with an uncooperative navigation device that continues to ignore the roadwork and keeps on nagging that you should turn left when turning left is not possible. On the other hand, low task demands could stimulate drivers to do things other than driving, such as using their mobile phone. The question is: how can we ensure optimal driving performance under these very different circumstances in order to secure safety and comfort? The driver who can regulate task demands also plays a role. For example, many older drivers seek to keep task demands low by avoiding complex traffic situations, such as driving at night, or adverse weather. However, if the weather suddenly changes or the preferred familiar road is closed, task demands may become very high for them, and bring them in an uncomfortable situation. The main message, however, is that people adapt to the tasks they perform in a way that in their perception is comfortable and not stressful.

Mental Workload and Driver State in Partly Self-Driving Vehicles

We are currently transitioning towards fully automated driving (Van Nes and Duivenvoorden, 2017). In the meantime, the use of driver support systems has increased, and an increasing number of vehicles are partly self-driving (also referred to as semi-automated), meaning the automation can take over part of the driving task or even that the car can drive by itself in certain conditions (the autopilot function for the highway). New systems are constantly being introduced to support the driver in the driving task. Systems such as Advanced Cruise Control (ACC) and Lane Keeping Systems (LKS) take over parts of the driving task; the ACC removes the mental load of maintaining a certain speed and keeps a safe distance from a lead vehicle, while the LKS takes care of lateral positioning. As more systems are introduced, more tasks are being taken over by automation. The role of the driver is slowly shifting from operator of the vehicle to a supervisory role, which is extremely monotonous. As a result, even though the task demands and mental workload are reduced, supervising a (partly) automated vehicle without an active role has a negative effect on driver state. From a safety perspective, the priority is to keep driver workload in the “optimal performance” window (central area in Fig. 1). State-related efforts to maintain the performance for longer periods of time should be avoided, but drops in performance (right hand side of the graph) are even worse.

Another potential risk is introduced by the temporary increase in task demand created by these systems themselves. Each system communicates to the driver in some way, giving warnings (blind spot warning, forward distance warning, etc.) and/or indicating the status of the system (on, off, or changing). With an increased number of systems, communication demands also increase. Different signals can be confusing or contradictory, leading to increased mental workload. Another increase in workload is caused by the fact that the driver has to be continuously aware whether systems are on or off and understand the systems' limitations—and check whether these may apply. It is therefore just as important to avoid too much increase in mental workload due to task-related effort requirements (Fig. 1).

Clearly, the use of driver support systems and partly self-driving vehicles impact task demand, leading to possible overload or underload, two situations that impact driver state and task demands. Thus there is the risk of leaving the optimal performance window, resulting in degraded driving performance. In the transition towards higher levels of vehicle automation, it is crucial to mitigate this risk and ensure drivers stay within the optimal performance window. Every level of automation should be designed in such a way that the optimal performance window is respected, which would require settings that can be changed.

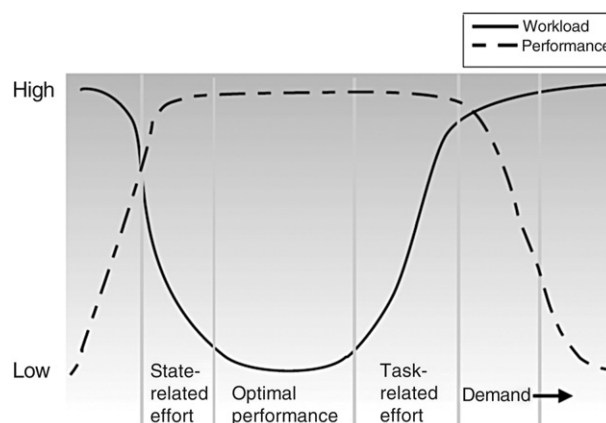


Figure 1 Increasing task demands (x-axis); task performance, and mental workload (y-axis). Source: From De Waard (1996).

Clearly, higher levels of automation offer the opportunity to mitigate driver workload. An interesting way to manage this effect would be to continuously monitor driver state during a trip to determine if it stays within the optimal performance window, and take measures if needed.

Assessment of Mental Workload and Driver State

Given the potential negative effects of mental underload/overload on driving behavior and safety, correct assessment of the driver's mental workload and state are crucial. There are three types of measures that can provide an indication of mental workload and driver state: performance, self-reporting, and psychophysiology (Paxion et al., 2014).

Performance

An assessment of mental workload and driver state should start with driving performance. In driving, primary task performance is reflected in lateral and longitudinal control. Both lateral control, that is the lateral lane position, and its variability (standard deviation of the lateral position: SDLP) have been used for a long time to reflect driver state. SDLP is sensitive to many factors, including the use of sedative drugs such as alcohol, distraction, and fatigue. Mean lateral position can reflect strategic choices: it has been shown that drivers who noticed the sedative effects of medicinal drugs drove closer to the emergency lane, where more space is available. Longitudinal control is reflected in the vehicle's speed and speed variability and distance to other vehicles (time headway). The latter measure is always in interaction with other vehicles and frequently a car following test is used where a lead car changes speed and needs to be followed at a close but safe distance. In this test, driver response time (delay in sensitivity to speed changes) can be calculated (Brookhuis et al., 1994).

In addition to primary task performance, which reflects the behavior to control the vehicle safely, there is secondary task performance. Secondary tasks are frequently used by researchers "to fill up the capacity of the driver." The idea is that with an extra task that is demanding, the driver will no longer have the spare capacity that is usually not required for normal driving. Instead, the secondary task will require this spare capacity, causing performance on either the primary driving task or the secondary task to deteriorate. In this way we are able to see, for example, which parts of a route are more demanding and resulted in higher workload. For the secondary task an artificial task is very often used. One such example is the N-back task: the driver has to remember a stimulus (a letter or a figure, the target) that was presented N stimuli back, and indicate whether it is presented again. The target has to be updated continuously. As N increases, the task can become very demanding. Another example is the peripheral detection task: drivers have to respond to a light that is presented in their peripheral vision. Both techniques are popular, but they also have major drawbacks. To start with, they are artificial. During normal driving you do not have to keep a memory set in mind and respond to it. Even though the peripheral detection task has some ecological relevance in the sense that a driver may have to respond to stimuli in the periphery, the task is still artificial because in normal traffic this need is seldom a continuous task (and the stimuli are, for example, pedestrians or other cars, not a red light). In other words, normal driving may be distorted. Moreover, we do not know which task the driver prioritizes, the secondary task or the primary task. However, there are embedded secondary tasks (that are part of normal driving): an example is mirror checking. Glancing in the rear-view mirror is a part of safe driving, but it is not a crucial task if one is not changing lanes. It has been shown in on-road experiments that the frequency and duration of mirror checks can change. For example, the frequency drops if the driver is operating a telephone or driving in a busy environment.

Thus, primary task performance (vehicle control) must be assessed when evaluating mental workload. If a secondary task is used, it should be clear that it does not affect normal driving, which in practice is difficult. An embedded secondary task is preferred.

Self-Reports, Subjective Measures

Asking how demanding a situation was for a driver has always been a popular technique to assess workload. It definitely makes sense to ask people how they experienced a situation, as this gives information that cannot be assessed in any other way. Simple, one-dimensional scales such as the Rating Scale Mental Effort can assess the experienced mental workload (and the effort invested). The popularity of self-reports may also be due to the ease and low cost of the technique. However, when assessing mental workload, self-reports are not enough (De Waard and Lewis-Evans, 2014; Matthews et al., 2019). For example, to know whether performance was protected by the investment of mental effort, one needs some objective measure of performance, in addition to the self-reported measure of workload.

Self-reports are also useful for assessing driver state: fatigue and sleepiness can be scored on a validated scale such as the Karolinska Sleepiness Scale. But again, just as with mental workload, a self-report of driver state is not enough, nor is it always accurate. Simply put, if drivers were able to evaluate their state correctly, why would accidents with drivers who fall asleep still happen? Another disadvantage of self-reports is timing: either normal behavior needs to be interrupted to obtain a rating, or ratings have to be given in retrospect after completing a (part of the) drive, which means they may be impacted by memory decay.

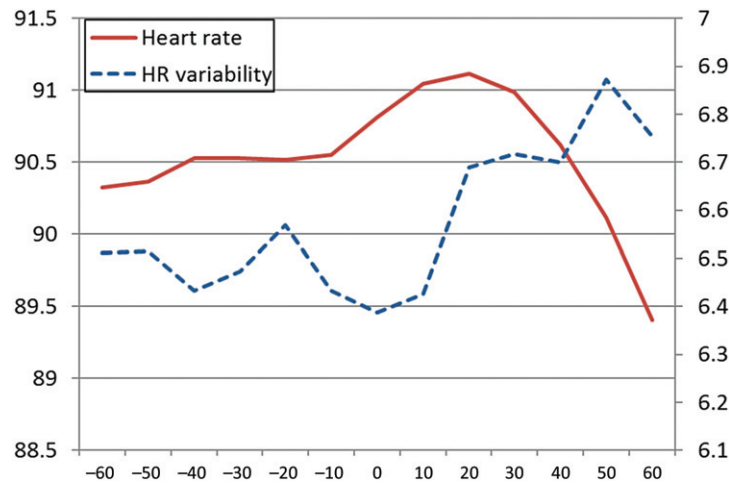


Figure 2 Example of a psychophysiological measure. Average heart rate in beats/minute (left axis) and variability of heart rate (HR Variability) in the 0.10 Hz frequency band (right axis, Ln-transformed, unit Ln^2). The averaged values of 24 participants performing a merge maneuvers (at $t = 0$ s) in their own car are depicted. On the x -axis, time is depicted in seconds. In general heart rate increases while variability decreases with increased mental effort.

Physiology

The ability to register how the driver responds physically to certain situations makes psychophysiological measures very attractive as indicators of mental workload or driver state (Hughes et al., 2019; Mehler et al., 2012). There is a broad range of measures, some requiring more advanced equipment than others. Heart rate and heart rate variability are relatively easy to assess, even though for the latter higher measurement accuracy is required. Average heart rate alone can reflect increased effort investment and reduced driver state. For driver state, measures that reflect Central Nervous System activity, such as EEG (electroencephalogram), are more accurate, but they are also more intrusive than Peripheral Nervous System activity measures, such as electro dermal activity and heart rate.

Some of these measures give the impression that they are objective and reveal the truth. In practice, this is not the case. These measures simply reflect how we respond to situations after interpreting them, and in that sense could just as well be called subjective measures. Apart from the fact that some require almost esoteric knowledge to administer and interpret them, a major disadvantage is that their reliability can be severely impacted by artifacts (disturbances that are not part of the physiological signal of interest, occurring as a result of movement, for example). On the other hand, an advantage of these measures is that many of them can be collected and interpreted continuously, so they can potentially be used in (partly) self-driving cars for monitoring purposes.

Fig. 2 illustrates how heart rate can reflect the performance of an effortful maneuver in traffic. A relatively high heart rate from the beginning reflects mental preparation; the heart rate variability is reduced in this condition, signifying higher mental effort. After the effortful maneuver, the heart rate slows down and the variability increases—both reflect less mental effort as the driving circumstances become less demanding.

Interpretation of Measures

Different measures can reflect different processes and driver states. For several reasons, it is important not to focus on one measure only. Different measures can and will diverge, which actually helps to interpret how the driver handled the situation. For example, the previously mentioned protection of performance should, by definition, show no primary task performance deterioration, while other measures such as self-reports or physiological measures can indicate an increase in mental workload. Furthermore, not all measures are equally sensitive or sensitive to the same task demands. For example, self-reports reflect investment of extra effort in conditions of both state- and task-related effort, while a physiological measure, such as heart rate variability, is only sensitive to task-related effort.

One would expect performance in the central area of Fig. 1 to be optimal; however, there are conditions related to self-regulation which could further improve performance. For example, while driving on a wide highway lane there is no need to minimize SDLP (swerve as little as possible)—however, with increased task demands drivers may actually swerve less, improving performance as a result of being extra activated. This somewhat counter-intuitive finding should be kept in mind when interpreting data: drivers have a lot of freedom to regulate their behavior, not only with regard to performance, but also at higher levels, depending on the goals they set. Self-regulation can also lead to other changes in behavior: drivers may decide to avoid situations that increase attentional demands, deciding, for example, not to drive during adverse weather conditions or on unfamiliar roads. In terms of traffic safety this can be a positive effect for drivers with limited experience (graduated licensing) and may extend mobility for drivers who suffer from mild cognitive impairment.

Conclusion

Driver state and workload demand can both affect performance, and thus driving safety. New advanced in-car systems that support, or even take over, parts of the driving task have the potential to make driving safer, but at the same time they affect driver workload and driver state. Before fully automated driving is realized, increasing levels of partial automation will continue to transform the driving task into a monotonous supervisory task, which can lead to driver underload. This state is obviously suboptimal and can jeopardize safety. On the other hand, as vehicles become more automated, if the driver needs to monitor many systems which are not well integrated, then the opposite effect—driver overload—could occur. To avoid both undesirable conditions, automation at all levels must be designed to provide optimal task demands, and this requirement must be tested prior to its introduction on the road. One promising way to optimize task demand is to dynamically adapt the driver's tasks on the basis of driver state and adaptive automation. In conditions where demands are high and overload is possible, tasks are taken over by automation to reduce the mental workload. In conditions where demands are low and underload lurks, tasks are handed back to the driver. In this way, the driver's performance is kept at the top of the inverted U in [Fig. 1](#).

Importantly, we should take multiple measures when we evaluate mental workload in experimental or naturalistic conditions, such as when investigating the effect of adaptive automation on mental workload ([Van Gent et al., 2018](#)). There is no single measure that reflects mental workload over the full range. Each measure is sensitive and not to all changes in task demand; the patterns that emerge from performance measures, self-reports (in experimental studies), and physiological parameters can provide a good indication of driver's mental workload and state.

See Also

Elderly driver safety issues; Human Factors in Transportation; Sleep-related Issues and fatigue; Education and Training of drivers; Drugs, illicit and prescription (Rune Elvik)

References

- Brookhuis, K.A., De Waard, D., Mulder, L.J.M., 1994. Measuring driving performance by car-following in traffic. *Ergonomics* 37, 427–434.
- De Waard, D., 1996. The measurement of drivers' mental workload. PhD thesis, University of Groningen. Haren: University of Groningen, Traffic Research Centre. Open access: https://www.rug.nl/research/portal/files/13410300/09_thesis.pdf.
- De Waard, D., Lewis-Evans, B., 2014. Self-report scales alone cannot capture mental workload. A reply to De Winter, Controversy in human factors constructs and the explosive use of the NASA TLX: A measurement perspective. *Cog. Technol. Work* 16, 303–305, doi:10.1007/s10111-014-0277-z.
- Hughes, A.M., Hancock, G.M., Marlow, S.L., Stowers, K., Salas, E., 2019. Cardiac measures of cognitive workload: a meta-analysis. *Hum. Factors*. DOI: 10.1177/0018720819830553.
- Matthews, G., De Winter, J., Hancock, P.A., 2019. What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures, *Theor. Issues Ergon. Sci.* DOI: 10.1080/1463922X.2018.1547459.
- Mehler, B., Reimer, B., Coughlin, J.F., 2012. Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Hum. Factors* 54, 396–412, doi:10.1177/0018720812442086.
- Paxion, J., Galy, E., Berthelon, C., 2014. Mental workload and driving. *Front. Psychol.* 5, 1344, doi:10.3389/fpsyg.2014.01344.
- Van Gent, T., Melman, H., Farah, N., van Nes, B., van, Arem., 2018. Multi-level driver workload prediction using machine learning and off-the-shelf sensors. *Transp. Res. Rec.* 2672 (37), 141–152, doi:10.1177/0361198118790372.
- Van Nes, N. van, K. Duivenvoorden, 2017. Safely towards self-driving vehicles. New opportunities, new risks and new challenges during the automation of the traffic system. R-2017-2E. SWOV, The Hague.

Further Reading

- Brookhuis, K.A., De Waard, D., Fairclough, S.H., 2003. Criteria for driver impairment. *Ergonomics* 46, 433–445.
- Schwarz, J.F., Ingre, M., Fors, C., Anund, A., Kecklund, G., Taillard, J., Philip, P., Åkerstedt, T., 2012. In-car countermeasures open window and music revisited on the real road: popular but hardly effective against driver sleepiness. *J. Sleep Res.* 21, 595–599.
- Sparrow, A.R., Lajambe, C.M., Van Dongen, H.P.A., 2019. Drowsiness measures for commercial motor vehicle operations. *Accid. Anal. Prev.* 126, 146–159.

Drugs, Illicit, and Prescription

Rune Elvik, Institute of Transport Economics, Oslo, Norway

© 2021 Elsevier Ltd. All rights reserved.

The Use of Drugs and the Risk of Road Accident: General Issues	221
Determining Dose of Drug	221
Controlling for Confounding Factors	222
Choice of Estimator of Risk	222
Publication Bias	223
Study Quality Assessment	224
Risks of Road Accident Associated With Illicit and Prescription Drugs	224
Illicit Drugs	225
Prescription Drugs	226
Concluding Comments	226
References	227
Further Reading	227

The Use of Drugs and the Risk of Road Accident: General Issues

Everybody knows that consuming alcohol before driving increases the risk of being involved in a road accident. The more you drink, the higher the risk. Does this also apply to drugs? One can easily think that the use of some drugs would increase accident risk, for example, drugs that make you sleepy, produce hallucinations, or otherwise influence cognitive functions. It has, however, proved considerably more difficult to estimate the relationship between drug use and accident risk, than to estimate the corresponding relationship for alcohol. This article reviews current knowledge regarding the relationship between drug use in drivers of motor vehicles and their risk of accident involvement (Elvik, 2013, 2015, 2018). The article does not deal with the risk of injury associated with drug use among pedestrians or cyclists, as there are almost no studies of it. The article has two main sections. The first section reviews some general issues concerning the quality of knowledge. The second section summarizes, for fourteen drugs, what is known about the relationship between use of them and the risk of involvement in a road accident.

Several general issues related to the quality and reliability of knowledge regarding the association between driver drug use and accident involvement arise when trying to summarize the current knowledge. Some of the issues arise in primary studies; others arise when trying to summarize evidence from several studies. The principal issues of concern are as follows:

1. The determination of the dose taken of a drug and the level of impairment caused by it while driving.
2. Controlling for other factors influencing accident risk in addition to the drug of primary interest.
3. How best to summarize the relationship between drug use and accident involvement, that is, either as a point estimate or as a functional relationship.
4. How to test and adjust for the potential presence of publication bias in studies of driver drug use and accident involvement.
5. How to assess study quality and its relationship to estimate the risk of accident involvement associated with drug use.

The first two of these issues arise in primary studies, the other three when summarizing the results of a set of studies.

Determining Dose of Drug

Most of the studies of the relationship between driver drug use and accident involvement are either case-control studies or culpability studies. In case-control studies, a case sample, usually injured drivers treated at hospitals, is compared to a control sample, often drivers stopped in roadside surveys. In this study design, it may be challenging to collect high quality data on the dose of drugs present in the body for both case and control drivers. Blood samples provide the best basis for determining the dose. Taking a blood sample is an invasive procedure, not easily administered roadside to drivers, most of whom are likely to be negative for drugs. To circumvent this difficulty, some studies have used less reliable sources of data to determine the dose of drugs, such as samples of saliva or urine, prescribed dose and self-reported use.

In culpability studies, drivers involved in accidents and judged to be at fault are compared to drivers involved in accidents judged not to be at fault. If both groups of drivers are recruited at medical facilities, it may be easier to collect high-quality data on drug use than in roadside surveys. There are, however, other problems in culpability studies, principally concerning the validity of assuming that drivers not-at-fault are representative of drivers in general and the interpretation of the estimator of risk usually applied in these studies (Røgeberg, 2019).

As a result of these problems, only a few of the studies of the relationship between driver drug use and accident involvement rely on high-quality data regarding drug use. In most studies, there is no precise information on the dose taken or when it was taken.

Thus, in some studies, it cannot be ruled out at that, at the time of the accident, the drug had been completely metabolized and become inactive. This is a major source of uncertainty in the results presented in the next section.

Controlling for Confounding Factors

The risk of becoming involved in a road accident depends on a host of factors. Drug use is just one of them. In order to correctly estimate the contribution to risk attributable to drug use, a study should, ideally speaking, control for all other factors that are known to influence accident risk. In practice, that is not possible. Studies vary considerably with respect to how many, and which, potentially confounding factors they control for when estimating the risk associated with a specific drug.

Easily observable factors, such as age and gender are controlled for in many studies. Quite a few studies also control for the use of other drugs than the one, which is of principal interest in the study. However, a minority of studies control for annual driving distance. This is a serious omission, as accident risk is known to be inversely related to annual driving distance (Elvik, 2011). One can easily imagine that drivers who depend on regular use of prescription drugs limit their driving, especially if the drug has side effects, which the driver notices while driving. Thus, an unbiased comparison requires that drivers using drugs should drive the same annual distance as drivers not taking drugs (either by study design or by statistical control for driving distance).

The absence of control for driving distance in most studies of the risk of accident involvement associated with drugs is a source of uncertainty in the results presented in the next section. More generally, a negative relationship has often been found between how well a study controls for confounding factors and the estimate of risk associated with a drug. The better the control for confounding factors, the lower the increase in risk attributed to the drug.

Choice of Estimator of Risk

Traditionally, the most common way of presenting the risk associated with a certain risk factor has been as a point estimate, for example, this risk factor is associated with an increase of 50% in the chance of accident involvement. For at least two reasons, presenting the risk associated with drug use as a point estimate can be uninformative. First, it seems reasonable to expect the effect of a drug to depend on the dose taken. Risk would then take the form of a dose–response curve, similar to the curves reported for blood alcohol concentration and the risk of accident involvement.

Second, it has been found, at least for illicit drugs, that there is an interaction between how common uses of a drug is in the normal population of drivers and the increase in risk associated with use of the drug. There tends to be a negative relationship: the higher the percentage of drivers in normal traffic using a drug, the lower the increase in risk associated with it. Fig. 1 gives an example.

The point estimate of risk of fatal injury associated with the use of cannabis—for the studies included in Fig. 1—is 1.37, or 37% increase in the risk. This is indicated by the horizontal line in Fig. 1. Individual estimates of risk are, however, widely and asymmetrically dispersed around the weighted mean estimate of risk. Most estimates of risk applying to studies where less than

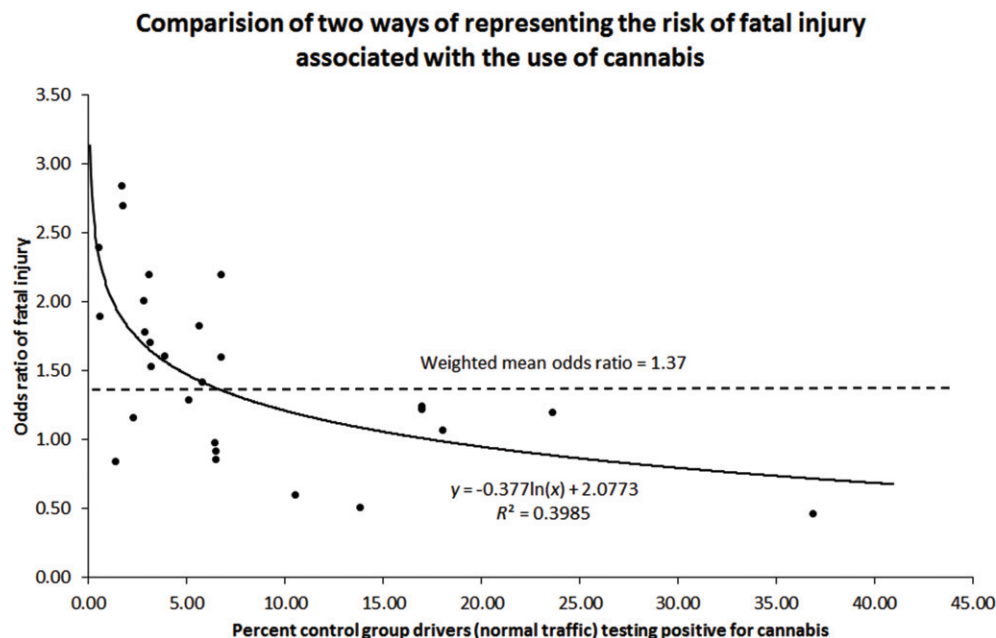


Figure 1 Point estimate of odds ratio of involvement in a fatal accident associated with the use of cannabis and functional relationship between share of control group drivers testing positive for cannabis and odds ratio of accident involvement.

10% of control group drivers tested positive for cannabis are located above the weighted mean estimate of risk (Elvik, 2018). All estimates based on studies where more than 10% of control group drivers tested positive for cannabis are located below the weighted mean estimate of risk. A curve has been fitted to the estimates. Although there is large variation in estimates around the curve, one could argue that it is a more informative summary of results than the single point estimate of risk.

Unfortunately, a curve as shown in Fig. 1 can arise in many ways. It can be a statistical artifact, but may also represent a real relationship. Many different interpretations are possible. Thus, when a negative relationship is found between the prevalence of use of a drug in normal traffic and the risk associated with it, this represents an additional source of uncertainty in knowledge, as it is rarely possible to determine why there is such a relationship.

Publication Bias

Publication bias denotes a tendency not to publish studies if the findings are not statistically significant or go in the opposite direction of what researchers expected and are therefore regarded as difficult to interpret or explain. A number of statistical techniques have been developed to probe for the potential presence of publication bias in a set of studies and to determine its magnitude. An informative technique is the trim-and-fill method. It is based on detecting asymmetry in the distribution of results plotted in a funnel plot.

A funnel plot is a scatter diagram of results. The abscissa shows the estimate of risk, the ordinate shows an indicator of the precision of the estimate, for example, its standard error. Standard error is usually plotted inversely, that is, the smallest standard errors are on top of the ordinate. Fig. 2 shows a funnel plot of estimates of the risk of becoming involved in a property-damage-only accident associated with the use of cannabis.

There were 19 estimates of risk, that is, the circular and triangular data points in Fig. 2. These were, as can be seen, very asymmetrically distributed and located far from the ordinate, suggesting publication bias with respect to both, the sign and magnitude of risk. Trim-and-fill trimmed away 15 of the 19 data points. When the mirror images of these data points are inserted, the plot becomes symmetric around the trimmed mean, but still has a hollow core. This is an example suggesting extreme publication bias.

Obviously, no statistical method provides direct evidence of publication bias. Such evidence can only be obtained by locating unpublished studies and comparing their results to published studies. The statistical methods are indicative only. Nevertheless, it is not difficult to imagine that publication bias could arise in studies of the risk of road accident associated with the use of drugs. Many researchers may expect the use of drugs to be associated with an increase in risk and may be puzzled and unable to explain an opposite finding. They may then choose to bury the study rather than try to get it published.

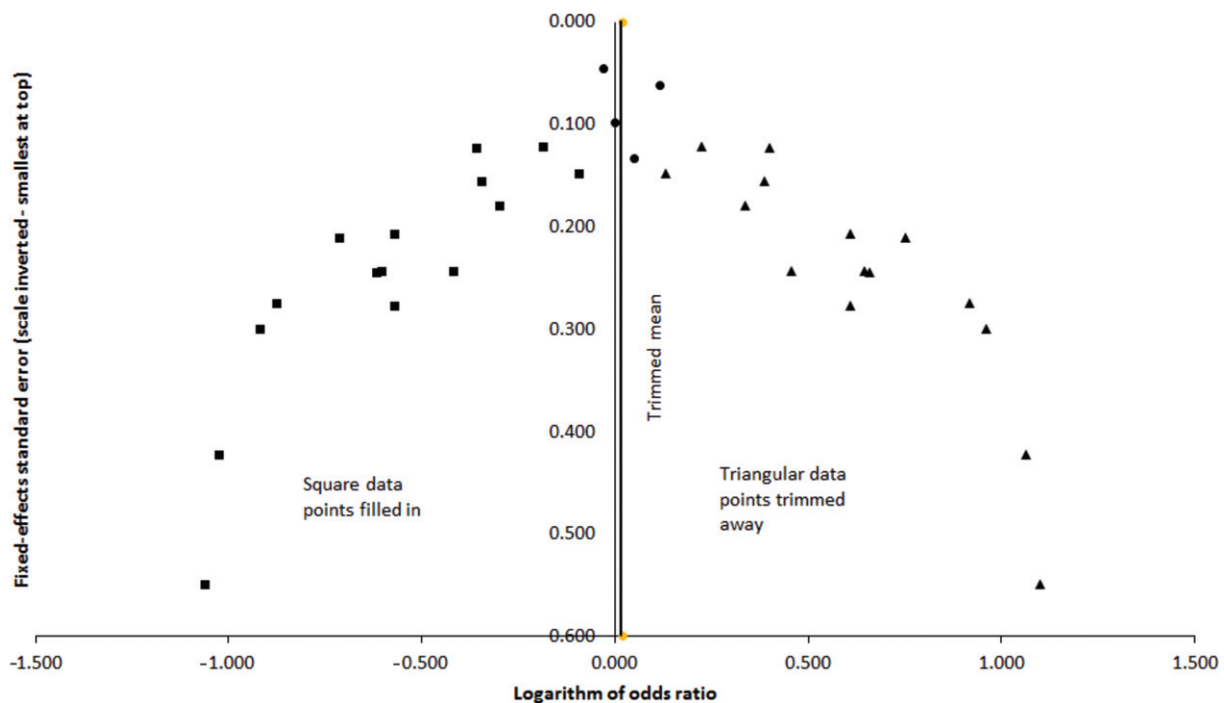


Figure 2 Trim-and-fill analysis to detect and correct for publication bias in estimates of the odds of involvement in property-damage-only accidents associated with the use of cannabis.

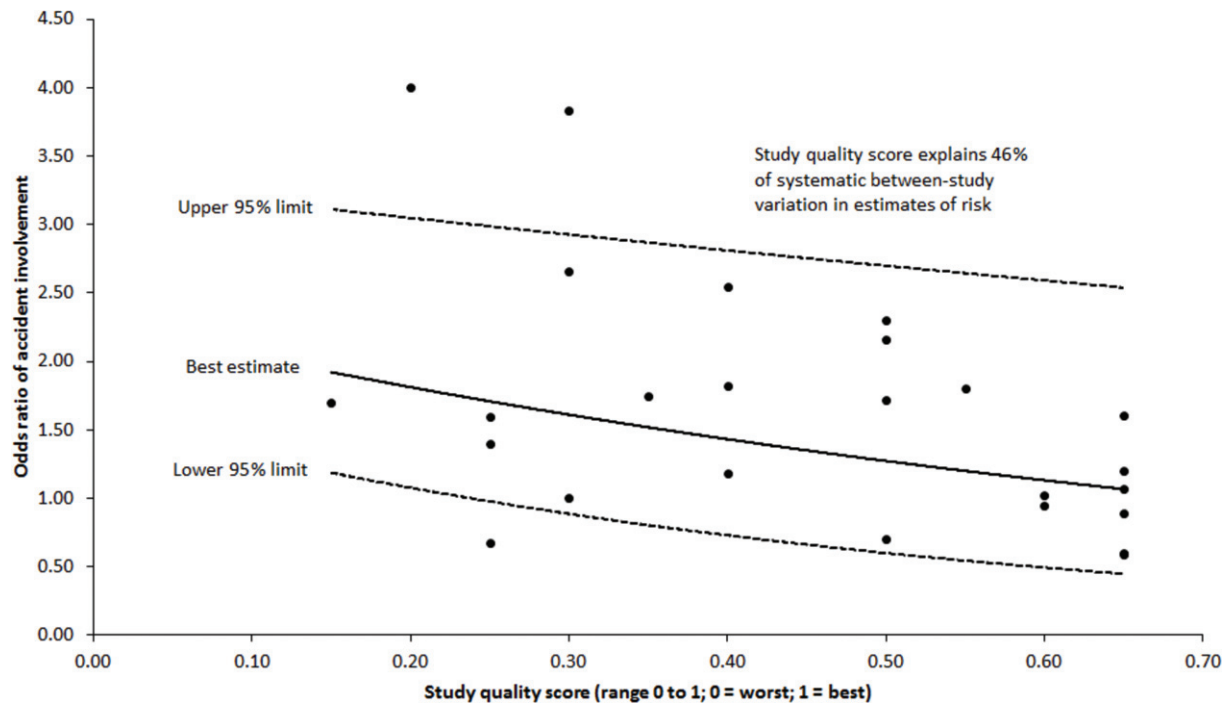


Figure 3 Relationship between study quality score and estimate of risk of accident involvement associated with the use of antidepressant drugs.

Study Quality Assessment

Study quality is a concept that does not have a standard definition and that escapes precise measurement (Greenland, 1994). However, nearly all researchers agree that studies vary in quality and that findings may be related to study quality.

Can we place equal trust in the findings of a study relying on self-reports only and not controlling for any confounding factors, as in a well-controlled study containing precise data on the dose taken of a drug that was still active during driving? Most researchers would say “no.”

Any attempt to quantify study quality contains an element of arbitrariness. An explicit and numerical scale for study quality assessment, explained in sufficient detail to be replicable, may still provide useful information. One should be reluctant to interpret numerical scores for study quality as anything more than an ordinal scale. Such a scale is nevertheless sufficient to determine, if there is any relationship between study quality and study findings. Fig. 3 shows an example.

Fig. 3 shows the relationship between quality score, on a scale ranging from 0 to 1, and estimates of the risk associated with the use of antidepressant drugs. A regression line has been fitted to the data points. The regression line is highly uncertain, as indicated by the dashed lines based on the lower and upper 95% values of the intercept and slope of the regression line.

Despite the large uncertainty, there is a systematic relationship between study quality score and estimate of risk. The higher the quality score, the lower the estimate of risk. The dispersion of estimates of risk also diminishes as study quality improves. For studies scoring 0.65, the range of estimates of risk spans from 1.61 to 0.59. For studies scoring less than 0.30, the range goes from 4.00 to 0.67. It is worth noting that the best studies scored 0.65, whereas a “perfect” study according to this scale would score 1. Thus, even the best studies have shortcomings.

Risks of Road Accident Associated With Illicit and Prescription Drugs

Drugs were included if at least three estimates of the risk of road accident associated with use of them could be found. A total of fourteen drugs met this requirement. Four of the drugs are principally used as illicit recreational drugs. Ten are prescription drugs used in the treatment of diseases. It should be noted, however, that some drugs are used both illicitly and legally. Thus, opiates are both used illegally and in analgesics for treatment of pain. Cannabis is also used both as an illicit and a prescription drug. There is a trend to legalize the recreational use of cannabis. Cannabis is by far the most commonly used drug of those that have been classified as illicit.

Table 1 describes the effects or intended use of the drugs (Marillier and Verstraete, 2019). If we make a broad distinction between stimulants and sedatives, two of the illicit drugs are stimulants (amphetamine and cocaine), two are sedatives (cannabis and opiates). The prescription drugs are used in the treatment of pain, asthma, depression, allergies, inflammations, anxiety, insomnia, diabetes, withdrawal symptoms in opiate addicts, and bacterial infections.

Table 1 List of drugs included and principal use and effects of the drugs

Drug	Effects of drug or intended use of drug
<i>Illicit</i>	
Amphetamines	A family of stimulants used both as an illicit recreational drug and as a prescription drug in the treatment of attention deficit hyperactivity disorder (ADHD). Amphetamine induces fatigue resistance and increased muscle strength. At high doses, it may impair cognitive function.
Cannabis	Drug made from a plant and used both as an illicit recreational drug and as a prescription drug. In recreational use, smoking it like a form of tobacco is common. It is a sedative. In medical use, cannabis can be used to reduce nausea and treat pain and muscle spasms.
Cocaine	Cocaine is a substance found in the coca plant. It is an illicit psychoactive drug that may cause an intense feeling of happiness or agitation. It is highly addictive, because it increases the concentration of neurotransmitters associated with reward in the brain.
Opiates	Opiates are drugs extracted from the opium plant. They are used both as an illicit recreational drug and as a prescription drug in medicine. Most opiates are sedatives. Morphine is the most common medical drug in the opiate family. It is used to reduce pain.
<i>Prescription</i>	
Analgesics	Prescription drugs used to reduce pain. Some of these drugs are opiates.
Anti-asthmatics	Prescription drugs used to treat asthma. The drugs do not cure the disease, but act to relieve the symptoms.
Anti-depressants	Prescription drugs used to treat depression. The drugs may also be used to treat anxiety disorders and some forms of chronic pain.
Anti-histamines	Drugs (over the counter or prescription) used to treat allergies. Some anti-histamines are sedative and can induce sleepiness.
Anti-inflammatory drugs	Prescriptions drugs used to treat inflammations. Some of the drugs also reduce pain.
Benzodiazepines	A large family of prescription drugs with various uses, including as a tranquilizer, to reduce anxiety, insomnia, and muscle spasms.
Insulin	Prescription drug used to treat diabetes. It replaces insulin produced by the pancreas.
Methadone	Prescription drug belonging to the opiate family, used in maintenance therapy for addicts to opiates. Maintenance means it reduces withdrawal symptoms from cessation of opium abuse.
Penicillin	A prescription antibiotic drug used to treat infections caused by bacteria. Excessive use may cause bacteria to develop resistance, which means that the drug no longer kills them.
Zopiclone	Prescription drug use to treat insomnia; a sleeping pill.

Table 2 Summary estimates of the risk of road accident associated with the use of drugs (odds ratios)

Drug	<i>Estimates of odds ratio for accident involvement—95% confidence interval in parentheses—number of studies in parentheses</i>		
	<i>Fatal accidents</i>	<i>Injury accidents</i>	<i>Property-damage-only accidents</i>
Amphetamines	5.70 (3.27; 9.95) (13)	8.98 (3.44; 23.40) (4)	8.67 (3.23; 23.33) (1)
Cannabis	1.39 (1.26; 1.54) (32)	1.62 (1.22; 2.16) (22)	1.43 (1.26; 1.63) (19)
Cocaine	2.91 (1.81; 4.67) (7)	1.59 (1.06; 2.38) (5)	1.52 (1.02; 2.26) (5)
Opiates	2.03 (1.34; 3.09) (10)	1.98 (1.55; 2.54) (20)	4.76 (2.10; 10.80) (5)
Analgesics	2.35 (1.42; 3.89) (8)	1.78 (1.35; 2.36) (17)	
Anti-asthmatics		1.33 (1.09; 1.62) (6)	
Anti-depressants		1.28 (1.07; 1.52) (25)	1.28 (0.90; 1.80) (5)
Anti-histamines	0.69 (0.18; 2.63) (2)	1.12 (1.02; 1.22) (7)	
Anti-inflammatories		1.38 (1.22; 1.56) (5)	1.53 (1.17; 2.01) (4)
Benzodiazepines	3.26 (2.30; 4.62) (10)	1.65 (1.49; 1.82) (51)	1.36 (1.04; 1.76) (4)
Insulin		1.14 (0.84; 1.54) (4)	1.01 (0.59; 1.74) (2)
Methadone		2.08 (1.47; 2.95) (3)	
Penicillin		1.12 (0.91; 1.39) (5)	
Zopiclone	2.31 (1.13; 4.70) (2)	1.42 (0.87; 2.31) (4)	4.00 (1.31; 12.21) (1)

Estimates of the odds ratio of accident involvement, which closely approximates the relative risk of accident involvement associated with the use of the drugs, are presented in **Table 2**. A distinction was made between three levels of accident severity: fatal accidents, injury accidents, and property-damage-only accidents. The reason for making this distinction was to assess whether there is a severity gradient in the risk associated with the use of drugs, analogous to that found for alcohol. Alcohol dramatically increases the risk of fatal accident, is associated with a smaller increase in the risk of injury accident and an even smaller increase in the risk of property-damage-only accidents.

Illicit Drugs

All illicit drugs included in this review have been found to increase the risk of accident, at all levels of accident severity. The increase in risk is statistically significant at the 5% level in all cases, although some confidence intervals are very wide.

The largest increase in risk is found for amphetamine. In general, relative risks do not show any clear severity gradient, possibly except for cocaine. The largest number of studies refer to cannabis, 73 in total. For three of the illicit drugs, amphetamine, cannabis and opiates, a negative relationship is found between the share of drivers in normal traffic testing positive for the drugs and the risk associated with them. For cocaine, there were too few results to test for such a relationship.

In most cases, estimates of risk associated with illicit drugs had a negative relationship to the number of potential confounding factors controlled for in a study: the more confounding factors controlled for, the lower the estimate of risk.

The consistency of evidence can be assessed by examining whether all estimates of risk indicate an increase, or whether some estimates indicate a decrease in risk. There were 17 estimates of the risk associated with the use of amphetamine. All indicated an increase in risk. For cannabis, 58 estimates indicated an increase in risk, 15 indicate a decrease. For cocaine, there were 15 estimates indicating an increase in risk, 2 indicating the opposite. Of the 31 estimates of the risk associated with the use of opiates, 27 indicated an increase in risk and 4 indicated a decrease.

Estimates of risk vary substantially for all illicit drugs. The variation in estimates of risk is substantially larger than sampling variation alone can explain. Systematic between-study variation in estimates of risk typically constitutes two-thirds of the total variation in estimates of risk, the remaining third being random sampling variation. Sources of the variation in estimates of risk are only partly known.

Prescription Drugs

Estimates of risk were found for a total of 10 prescription drugs. Evidence is less complete than for illicit drugs. Estimates of the risk of an injury accident are available for all prescription drugs. No estimates of the risk of fatal accident were found for anti-asthmatics, anti-depressants, anti-inflammatories, insulin, methadone, and penicillin. No estimates of the risk of becoming involved in a property-damage-only accident were found for analgesics, anti-asthmatics, anti-histamines, methadone and penicillin.

The increase in the risk of injury accident associated with prescription drugs is modest and not statistically significant at the 5% level in all cases. The highest risk is associated with methadone, but there are only three studies and the mean estimate of risk is highly uncertain. A severity gradient in risk is found for benzodiazepines. A total of 65 estimates of risk were identified for benzodiazepines, close to the 73 estimates found for cannabis. For most other prescription drugs, there are few estimates of the risk of road accident associated with their use.

Testing for a relationship between the share of drivers in normal traffic testing positive for a drug and the risk associated with the drug was only possible for opioid analgesics. A negative relationship was found: the larger the share of drivers in normal traffic testing positive for the drug, the lower the increase in risk associated with it.

Anti-histamines appear to be protective with respect to the risk of fatal accident. A protective association has only been found for anti-histamines that are not sedative and only one study (of the two that were found) has found a protective effect.

For most prescription drugs, there are few studies of the risk associated with them. It has therefore in most cases not been possible to test if estimates of risk are related to how well a study controlled for confounding factors. Negative relationships were found for anti-depressants and for the association between benzodiazepines and the risk of injury accident; meaning that the more potentially confounding factors a study controlled for, the lower were estimates of risk. For the association between benzodiazepines and the risk of fatal accident, a positive relationship was found: the better controlled studies indicated a larger increase in risk than the more poorly controlled studies.

Estimates of the risk of road accident associated with the use of prescription drugs are less consistent than those for illicit drugs. Thus, as an example 9 out of 30 estimates for anti-depressants indicate a decrease in risk. Two of six estimates for insulin indicate a decrease in risk. For the most widely studied class of drugs, benzodiazepines, 7 out of 65 estimates of risk indicate a decrease, the other 58 indicate an increase of risk.

There is large variation in estimates of risk, although somewhat smaller than for illicit drugs. The share of systematic between-study variation in estimates of risk is typically around 45%, but ranges from 0% to 95%. Little is known about why estimates of risk vary between studies.

Concluding Comments

The risk of road accident associated with the use of drugs is poorly known. On balance, the evidence suggests that use of illicit or prescription drugs is associated with an increased risk of accident involvement. The increase in risk associated with drugs appears to be greater for illicit drugs than for prescription drugs. On the other hand, prescription drugs are more widely used than illicit drugs and may therefore, despite the moderate increase in risk associated with them, contribute more to the population attributable risk associated with drug use than the use of illicit drugs.

There are many sources of uncertainty in the estimates of risk presented in this article. The most important sources of uncertainty are as follows:

1. Most studies have no precise information about the dose taken of a drug. Very few studies have been able to test for a dose-response relationship between the dose taken of a drug and the size of the increase in risk. This means that a key criterion for assessing the causality of the statistical associations found is largely unknown.

2. Most studies do not control for a very important potentially confounding factor, annual driving distance. It is unlikely that regular users of illicit or prescription drugs drive the same distance per year as nonusers of drugs.
3. There is evidence of a negative relationship between the share of drivers in normal traffic testing positive for a drug and the risk associated with use of the drug. Although such a relationship can have many reasons, it casts doubt on how informative it is to summarize the risk by means of a point estimate, rather than a curve showing risk as a function of the share of traffic exposed to a drug.
4. Most tests for publication bias suggest that there is publication bias. If the results of these tests are taken at face value, they imply that estimates of risk are considerably exaggerated for many drugs. Thus, to give some examples, the estimate of risk adjusted for publication bias was 2.74 for the risk of fatal accident associated with use of amphetamine; the crude estimate was 5.70. Similarly, the crude estimate for injury accidents for benzodiazepines was 1.65; the adjusted was 1.18. The corresponding estimates for the risk of injury accidents associated with analgesics were 1.78 and 1.01.
5. Estimates of risk vary considerably for all drugs. Reasons for the variation are largely unknown. Explanations can be either methodological, that is, associated with weaknesses of the studies, or substantive, that is, pointing out why drugs increase risk more in some groups (e.g., novel users of the drug) than in others. Considering the generally poor quality of studies, the presumption that methodological explanations of the variation in estimates of risk dominate must be favored.

To conclude: It is more likely that the use of drugs increases the risk of accident than the opposite. Causality of the relationship has not been established.

References

- Elvik, R., 2011. A framework for a critical assessment of the quality of epidemiological studies of driver health and accident risk. *Accid. Anal. Prev.* 43, 2047–2052.
- Elvik, R., 2013. Risk of road accident associated with the use of drugs: a systematic review and meta-analysis of evidence from epidemiological studies. *Accid. Anal. Prev.* 60, 254–267.
- Elvik, R. 2015. Risk of road traffic injury associated with the use of drugs. A paper prepared for the World Health Organization. Oslo, Institute of Transport Economics.
- Elvik, R., 2018. Interpreting interaction effects in estimates of the risk of traffic injury associated with the use of illicit drugs. *Accid. Anal. Prev.* 113, 224–235.
- Greenland, S., 1994. Invited commentary: a critical look at some popular meta-analytic methods. *Am. J. Epidemiol.* 140, 290–296, 300–302.
- Marillier, M., Verstraete, A., 2019. Driving under the influence of drugs. *WIREs Forensic Sci.* e1326. Available from: <https://doi.org/10.1002/wfs2.1326>.
- Røgeberg, O., 2019. A meta-analysis of the crash risk of cannabis-positive drivers in culpability studies—avoiding interpretational bias. *Accid. Anal. Prev.* 123, 69–78.

Further Reading

- Chihuri, S., Li, G., 2017. Use of prescription opioids and motor vehicle crashes: a meta-analysis. *Accid. Anal. Prev.* 109, 123–131.
- Monárrez-Espino, J., Möller, J., Berg, H.-Y., Kalani, M., Laflamme, L., 2013. Analgesics and road traffic crashes in senior drivers: an epidemiological review and explorative meta-analysis in opioids. *Accid. Anal. Prev.* 57, 157–164.
- Rudisill, T.M., Zhu, M., Kelley, G.A., Pilkerton, C., Rudisill, B.R., 2016. Medication use and the risk of motor vehicle collisions among licensed drivers: a systematic review. *Accid. Anal. Prev.* 96, 255–270.

Education, Training, and Licensing

Matúš Šucha, Kristýna Josrová, Palacký University Olomouc, Czech Republic

© 2021 Elsevier Ltd. All rights reserved.

Children's Traffic Education	228
Driving Education and Testing	228
Problematically Taught Skills	229
Graduated Driver Licensing (GDL)	230
Driver Education Framework	230
Continuous Education and Training	231
First Aid Training	231
Training for Senior Drivers	231
Defensive and Eco-Driving Courses	232
References	232

Children's Traffic Education

Children's traffic education is primarily aimed at improving the safety of all road users. In the course of time, the concept of children's traffic education has expanded to also include education toward safe mobility, which includes the safe utilization of public transport and the choice of a transportation mode. It is mainly provided at kindergartens and primary and middle schools, exceptionally at secondary schools. This systematic implementation reflects the practical aspect, as the target group of children, unlike that of drivers, can be reached easily. The benefit of children's school-based traffic education is its secondary impact; it plays a significant role in building children's and young people's hierarchy of values, as well as having the potential to shape the attitudes of all prospective road users.

In specific terms, children's traffic education encompasses four major domains:

- Education toward mobility as education toward safety.
- Education toward mobility as a form of social education (raising young people's awareness about vulnerable road users and developing their empathy, willingness to help others, responsibility, thoughtfulness, and sense of partnership).
- Education toward mobility as environmental education (intended to raise children's awareness about environmental issues related to motor traffic and encourage them to behave in an environmentally friendly manner).
- Education toward mobility as health education (intended to help children recognize unhealthy factors and promote measures to alleviate them, e.g., the choice of a good means of transport on their way to school).

The last of the domains includes first aid education. This is a very important aspect, as road accidents in many countries are the most frequent cause of death among people below 25 years of age. It should be noted at this point that children as young as six are capable of providing effective first aid.

However, children's traffic education cannot rely on school only. A significant part of it takes place within the family, on an informal basis. From their early years, children learn whether breaking rules or aggressive or violent behavior at the wheel are "common traffic-related predicaments" or something inappropriate. Road use-related behavior is a type of habitual behavior, which children adopt once they begin to watch the world from their car seats. Children tend to internalize the behavioral models of their parents and significant others from a very young age and incorporate them into their future behavior portfolio. For this reason, the focus on children, even at the informal level, is a key element of traffic education.

Driving Education and Testing

Motorized vehicles, whether they are cars, motorbikes, or lorries, are complex machines which are not very easy to operate and are potentially very dangerous to the drivers themselves and also to others in their vicinity. Shortly after the invention of the motor car, it became an increasingly popular means of transportation, which also meant that these vehicles caused numbers of injuries and deaths. Very soon, it became evident that drivers should be registered and educated in order to reverse the negative effects of driving. Since then, governments and local authorities have granted driver's licenses, regulated driver education, and enforced the traffic rules.

A driver should not only be able to drive a car and maintain it but also drive the vehicle in a safe manner. Driver training focuses on teaching skills, which mainly support the goal of providing learners with skills that help them to increase their mobility, while, at the same time, focusing on traffic safety. Countries and states differ in their requirements for mandatory driver education. In order to

gain a driver's license, an aspiring driver undergoes the obligatory training and then has to pass exams—both theoretical and practical. Usually, the education is composed of theoretical lessons and practical driving training. Often, the theoretical lessons, or at least a certain number of them, must precede the driving training and they are not always compulsory. Theory, which includes learning the road code, traffic signs, and in some jurisdictions also car maintenance, can be taught in various forms: lectures in driving school classrooms, online courses, or self-study of books and other materials that are provided. Instructional videos can be shown and sometimes a computer is used. Once the theoretical part of the training is completed, sometimes a theory test follows, and if the learner is successful, it allows them to proceed to driving lessons.

Driving schools sometimes let their students train first on driving simulators, which can also be a mandatory part of the training. Practical driving lessons are taught one-on-one in cars fitted with dual controls that allow the instructor to take control over the vehicle, if needed while the learner sits behind the wheel. The instructor supervises the learner's driving and monitors possible mistakes while giving instructions and corrections. The lessons often start in easier places such as tracks or car parks and progress to more difficult locations, such as city traffic and highways. The number of hours spent on certain types of roads is sometimes established by law and includes not only roads, such as highways or country roads, but also driving in specific conditions, such as nighttime driving. Learners are taught to operate a vehicle and sometimes to use procedures (e.g., MISS for changing lanes—mirror, indication, shoulder check, steering). They learn a number of essentials which they must automatize, from adjusting their mirrors and seats and checking their blind spots and moving on to forward and reverse driving, signaling, parallel or bay parking, emergency stops, three-point turns, passing, changing lanes, and driving at intersections. Once they have attended the required number of driving lessons or, in some countries, when they feel confident about their driving skills, they proceed to final testing. To obtain a driver's license, applicants in some countries must take a theoretical exam, which is typically a multiple-choice test. Learners worldwide finish their training with a practical driving test, in which the examiner usually gives basic instructions, requests certain learned actions (e.g., emergency stop or bay parking) to be performed, observes what mistakes are made and how serious they are, and decides whether the applicants are good enough drivers to get their driver's license.

There are great differences in mandatory driver education worldwide; in some cases the system can be more complex, such as graduated licensing or required specialized lessons, while in other cases only the final test is required. In most countries, one can start driving a car unsupervised at the age of 18 years, though some countries allow even younger teenagers, even as young as 14 years, to drive. Driving lessons with an instructor can take 10–40 h in some countries, but in some places it can be as little as six or even none at all—passing the exam alone suffices. In some countries, a first aid course is also a part of the mandatory education. This training aims at reducing the damage to health and loss of life in the event of a traffic accident. Drivers also typically learn to drive only either an automatic or manual car, depending on what is locally widespread, and when relocating abroad, they might not be allowed to drive a different type of vehicle or damage the vehicle by incorrect usage.

Problematically Taught Skills

Driving is a complex activity that assumes a range of skills, experience, personality prerequisites, and motivational traits. Although the most appropriate solution for specific traffic situations is the possession of basic vehicle control skills, research has shown that for long-term safe management having only those skills is not enough (Mayhew and Simpson, 2002). Safe driving is an activity that requires particular insight and self-regulation on the part of the driver—driving is a “self-paced task.” It is up to the driver (his or her decisions and behavior) to choose a safe driving style.

Modern research in traffic psychology refers not only to the performance characteristics of drivers, but also to their personal characteristics and motivational factors. Therefore, what is assessed is not only what the driver is able to do (skills), but also what the driver intends to do (motivation and personality factors).

By the time the learners pass their final test, they have merely learnt the basics of driving. Given that their practical training usually lasted for only a couple of dozen hours spread over several months, they have not had the chance to truly master this skill. It takes thousands of kilometers driven to become a confident driver, for a reason. Newly licensed drivers are sometimes obliged to visibly display a learner's plate with a large letter L or P on the rear and sometimes also the front of a car for a given period to indicate to other drivers to be more considerate around new drivers. Fresh drivers can also experience different behavior in traffic than what was taught and then adjust their own driving accordingly, simply learning by observation and then spreading those ways to others. Seeing somebody break a rule encourages others to do the same and normalizes a certain type of behavior. It is essential for drivers to learn good habits soon, as they are unlikely to change them later—in particular, those habits that are formed very early can be almost impossible to break, especially when there are few incentives to do so.

Driving is a multitasking activity, which requires us to use our focus and constantly shift it. While driving, we must simultaneously follow our chosen route, keep the vehicle moving at the intended speed, watch the traffic signs and the traffic, communicate our intentions to other drivers, and pay attention to what is going on, while also anticipating the moves of others. Within a matter of seconds one must execute a number of checks on one's surroundings. This series of activities become automatic by experience; however, adjusting never stops. A driver must accommodate his or her driving to the current conditions, including taking into account different weights of cars, the surface, the weather or his or her own tiredness. Drivers can learn this only by experience as these conditions change and simply cannot be taught.

People very often undergo driver education as soon as they legally can. The downside of this is that in young adulthood, youngsters are prone to risky behavior, to underestimating the risks, and to deliberately breaking the rules. This can be explained by

immaturity of the brain, which continues until the mid-twenties (Mayhew and Simpson, 2002). To address this problem, sometimes stricter rules apply to new drivers and some countries have introduced a graduated licensing system.

Training aims at teaching learners necessary skills, but it also easily leads to overconfidence (Hatakka et al., 2002). Most people consider themselves above-average drivers, which also makes them take more risks. Having undergone training convinces people that they have mastered certain skills, so more education can actually lead to less safe practices, which for example, happens in skid training (Mayhew and Simpson, 2002). To avoid this, instructors should let their students experience failure and demonstrate how little control drivers sometimes have over their vehicle and add discussion about the topic and experience (Peräaho et al., 2003).

One of the things typically overlooked in curricula is communication. Traffic is an organic structure and its smooth and safe functioning depends on the quality of the interaction of all its participants. When not alone in an area, one should use principles of defensive driving, anticipating the mistakes made by others, and driving safely, but when there are other people, it is necessary to let them know about our intended move and allow some buffer in case there is a lack of clarity. During training, signaling is taught but there are also individual differences in various situations, which are rather a matter of intuitive courtesy or reactions learnt by observation. Lack of communication or miscommunication of intentions can have fatal consequences.

Usually, when people are using a certain mode of transportation at any given moment, whether it involves walking, riding a bike, or driving a car, they become fully immersed in their role and although they may try to drive safely, others might not recognize their behavior as being safe. Some drivers have a lower sensitivity threshold and drive too close to others, but often people can at least know which actions on the part of others make them uncomfortable and recognize that they should not impose those on others; however, it is a bigger issue when there are people using different modes of transportation. People often do not realize that while the way they drive is completely acceptable to them, others may find it uncomfortable or dangerous. All sides see things from different perspectives and, immersed in their own world, how the other party might feel often does not even enter their mind. To empathize with people who use different forms of transport, people can try to put themselves in their shoes by simply trying to get around in a way they usually do not—by bike or on foot. Such an experience can make them live out the perceptions and emotions of those who they typically pass and become more kind to them.

Another notable fact is that roads often are not designed to encourage safe behavior. For example, straight and wide roads do not encourage safe speeds. And, even more alarming, traffic engineers assume when designing roads that a lane of a motorway will be able to safely carry well over 2000 vehicles per hour at high speed, say 120 km/h, even though drivers in driving schools are told that they at such speeds should have a minimum cushion of 3 seconds to the vehicle in front of them. A 3-second gap will lead to a capacity below 1200 vehicles per hour. And, if a driver tries to keep a 3-second gap, someone else will often enter into that gap.

Graduated Driver Licensing (GDL)

Teenagers and young drivers are more likely to be involved in a crash than any other age group. Car crashes are also the leading cause of death for teenagers. This trend led to the adoption of GDL by a number of big countries, such as most states of the United States and Australia (Shope, 2007). Graduated licensing is a system in which the learner must go through multiple stages before obtaining a full license. There are usually three stages, during which new drivers hold various kinds of provisional licenses, depending on which stage they are at. Each stage comes with certain restrictions, such as a prohibition on driving during night hours, exceeding a certain speed, driving unsupervised, transporting young passengers, or using a phone in any way while driving. These restrictions are based on the most risky factors for this age group. A number of rules apply during the entire process, but on entering a new learning stage with a more advanced type of license, some rules become freer. Drivers usually first take a theoretical test and in the last stage take a practical driving test, but meanwhile may have required hours to drive or additional tests to pass.

A traditional part of the system is the requirement to, at least at the beginning, drive only with a supervisor, a person who has held a full license for a given period. Supervisors sometimes completely replace instructors. The supervisor advises the learner driver and sometimes confirms in a logbook what drives were taken. In some countries, learners must drive a certain number of hours in the daytime and another number when it is dark. Each stage has a firmly set or minimum duration, so the entire process can sometimes take about three years to complete, unlike in traditional driving education systems which can take only months or even weeks to finish. Grading the licensing process in this way gives the learner time to learn good habits and to enter challenging conditions gradually, but it also signals to the driver that he or she is still learning and is not yet a ready, competent driver. Another important factor is delaying the time when a young person becomes a driver with full privileges (Mayhew and Simpson, 2002). This is very useful, given the inexperience and risky behavior that is typical for that age, and thanks to the fact that graduated licensing addresses specific problematic and risky behaviors, this system has proved successful in reducing the numbers of crashes involving young drivers (Shope, 2007).

Driver Education Framework

Driving education is traditionally taught with lecture-style instruction in which interaction happens only when the instructor corrects the learner's mistakes, but knowledge and skills alone do not ensure safe driving. A great number of factors which intuitively we might not even associate with driving (and which go beyond driving) in fact have an impact on the way we end up driving. There are many motives and invisible influences which determine this, such as role models, the desire to fit in and to impress others, a habit

of doing things at the last minute, and so on. Our overall motives and goals determine how we often unconsciously decide to act and yet they are not addressed in standard driver education. The GDE framework describes what influences how we drive and how to modulate each level in the education process for the best outcome. Each level affects other levels, but the highest, fourth level, called skills for life, has the greatest impact on driving. It includes goals, lifestyle, and tendencies and is established even before one starts one's driver education. Increasing the number of hours dedicated to skills training will not change how one drives unless motivation and goals are addressed because we choose our driving strategy according to our goals. Concerning the lower levels, level one describes how to maneuver the car as an independent vehicle, level two describes how to behave in traffic and in different conditions and how to manage communication with other drivers, for which we need to master the traffic code. The third level concerns individual trips and decisions regarding choosing the means of transport and location, but also when and how, for instance at what pace will we travel (Peräaho et al., 2003). To improve safety, we should not only equip learners with the necessary knowledge and train them in technical skills until they become automatized, but also aim to challenge their attitudes and perspectives (Shope, 2007). The problem is that even if there is a skill, if there is no will to practice safe driving, the skill will not help. One must be willing to practice what was taught. To achieve this, the instructor should take on the role of a tutor and realize that people are strongly influenced by their preferences, convictions, and habits, and so taking a driver's preferences into account is vital. We cannot teach by simply telling them, but should combine suitable experience with theory and engage the student to experiment and learn new perspectives. The tutor creates an engaging program, which includes exercises, feedback, self-evaluation, and group discussions, even including role play about various situations (Peräaho et al., 2003). People downplay the hazards and are constantly subjected to their biases, but discussions and evaluation from others and oneself help raise self-awareness, which is the key quality here. Drivers should learn to know their weaknesses and limitations but they should be able to learn to recognize them independently over their lifetime. The tutor must attempt to understand and respond to the learner's goals while providing stimulating content, which teaches lifelong skills of metacognition. At the moment, addressing the highest level is only rarely present and only in postlicense training, although such training brings positive outcomes (Peräaho et al., 2003).

Continuous Education and Training

A driving license for cars does not have an expiration date, so the majority of drivers undergo nonrecurring (one-off) education. Unless one seriously violates the traffic rules, or moves to a different continent, he or she does not undergo any retraining or retesting. This means that what was learnt at the beginning of one's driving career is not likely to change and people are not systematically taught about safe practices or changes in traffic rules, as well as in technologies. Most of the traffic code involves unchanging rules, but sometimes new rules are introduced, such as late merging. The only group of people who are definitely informed about changes are lawmakers, traffic professionals, and professional drivers, while most of society will only learn about them through media coverage, if at all. The problem is that the message does not reach all people and can be forgotten or ignored, and, moreover, the media can unintentionally spread misinformation.

Drivers could also benefit from learning about new technologies implemented in cars. They themselves might not use them but other drivers can adjust their driving to how the vehicle reacts. New systems, such as advanced driver-assistance systems, whether collision avoidance systems, intelligent speed adaptation, or lane assistance, alter driving and learning about them can help drivers understand what to expect from other vehicles. At the moment, drivers will mostly only learn about them, if reading specialized literature and only, if they are interested in the topic.

First Aid Training

Medical advances also modify good practices in first aid. As first aid training is attended only once or sometimes not at all, the driver will most probably not know how to respond in case of an emergency or simply forget it. Thorough training events attended at regular intervals once in a set number of years might improve knowledge and skills and teach up-to-date first aid strategies, and also serve as an opportunity to teach drivers about developments in the traffic law and similar areas which might affect traffic.

It is estimated by the Red Cross that first aid education has the potential to reduce the number of deaths resulting from road accidents by up to 10% and can also minimize the consequences of serious injuries. In addition, training in first aid can be highly effective in increasing drivers' prosocial behavior and serve as a preventive factor for risky behavior in traffic. A number of studies show that mandatory first aid courses as part of driving school significantly improve prospective drivers' knowledge and confidence. Nevertheless, neither willingness to provide first aid nor the frequency of its actual provision will increase without psychological aspects and strategies to overcome psychological barriers being incorporated in such courses.

Training for Senior Drivers

In widespread suburban areas, a private car is a necessary means of transport and giving up driving can result in a large negative impact on a person's quality of life. Thus, methods that enhance driving safety and prolong the mobility of elderly people are important.

Different studies have indicated that cognitive functioning has a significant effect on the driving safety of elderly people, both in healthy people and those with mild cognitive impairments. By focusing on this association between cognitive functioning and driving safety, many studies have introduced different types of cognitive training and investigated their transfer effects on driving safety and the maintenance of mobility. Driving is a skill that can, and should, be continually improved. Courses designed for senior drivers keep their knowledge about driving fresh and help them get the most out of their vehicle, while reducing the risk to themselves, their passengers, and others on the road. A comprehensive driving improvement course ensures that senior drivers have the most up-to-date driving techniques and understand the latest vehicle technologies.

Defensive and Eco-Driving Courses

Two of the most common types of further training courses for nonprofessional drivers focus on defensive and eco-driving. The concept of defensive driving emerged in the 1950s in the United States and is associated with the “Smith System of Defensive Driving.” Defensive driving courses are generally aimed at teaching a driving strategy, which assures a sufficient level of one’s own safety despite the effects of external factors. They consist of a theoretical and practical part. The theoretical part of the course usually covers topics such as the prevention of traffic accidents, the risks related to road traffic, the principles of defensive driving, the causes of risky situations and road accidents, and physiological and psychological limits. The practical part of the course generally involves on-road driving evaluation focusing on observation of the rules, correct driving technique, defensive driving strategy, and economical driving. In the course of time, the principles of defensive driving have been incorporated into driving school curricula and taught independently in defensive driving courses. The concept of defensive driving later became criticized to some extent. It was argued that defensive driving could make drivers engage in aggressive behavior or result in collision situations brought about by drivers asserting their need for their own safety and disregarding the context or other road users (e.g., keeping too much of a distance in heavy traffic) (Steinmetz, 2008). Finally, the term itself implies defense, that is, counteraction, which does not correspond with cooperative principles. The modern notion of defensive driving no longer involves passively aggressive elements. Rather it is based on the assumption that others may make mistakes, so one should not assume that others follow all rules all the time.

Courses in eco-driving became common in the 1990s, offered either independently or together with training in defensive driving. Courses promoting eco-driving principles involve teaching drivers how to behave in specific situations in which fuel (i.e., money), time, and the environment can be saved. Eco-driving courses are also sometimes associated with a new driving culture that highlights not only safety but also environmental aspects and the application of smart technologies. In addition, these courses respond to advances in car design (Barkenbus, 2010). Recent decades have seen dramatic improvements in technology and horsepower, while the majority of drivers have maintained their old driving styles. Eco-driving courses are thus an important component of sustainable mobility and energy-efficient car use.

References

- Peräaho, M., Keskinen, E., Hatakka, M., 2003. Driver competence in a hierarchical perspective; implications for driver education. *Report to Swedish Road Administration*. Traffic Research. p. 51.
- Hatakka, M., Keskinen, E., Gregersen, N.P., Glad, A., Hernetkoski, K., 2002. From control of the vehicle to personal self-control; broadening the perspectives to driver education. *Transp. Res. Part F: Traffic Psychol. Behav.* 5 (3), 201–215.
- Shope, J.T., 2007. Graduated driver licensing: review of evaluation results since 2002. *J. Saf. Res.* 38 (2), 165–175.
- Mayhew, D.R., Simpson, H.M., 2002. The safety value of driver education and training. *Inj. Prev.* 8 (2), ii3–ii8.
- Steinmetz, S.S., 2008. Defensive driving and the external costs of accidents and travel delays. *Transp. Res. Part B: Methodol.* 42 (9), 703–724.
- Barkenbus, J.N., 2010. Eco-driving: an overlooked climate change initiative. *Energy Policy* 38 (2), 762–769.

Elderly Driver Safety Issues

Mark J King, Queensland University of Technology (QUT), Centre for Accident Research and Road Safety (CARRS-Q), QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	233
The Aging of the World and Some Gender Differences	233
Ageing and Driving Performance	234
Ageing and Injury Severity, and the Biases Involved in Focusing on Fatality Data	234
From Driving Performance to Crash Incidence—Mitigating and Exacerbating Factors	235
Road Infrastructure Implications of Crash Patterns Among Elderly Drivers	236
Vehicle Safety Implications of Elderly Driver Crashes	237
Licensing and Medical Approaches to Elderly Driver Safety	237
Giving up Driving	238
Are Automated Vehicles the Answer?	238
References	238
Further Reading	239

Introduction

There is no standard definition of an “elderly” person, for example in most developed countries it is 60 years or more use, while for the United Nations it is 65 years or more; other definitions distinguish between being old and being elderly, with the latter indicating a degree of frailty. Age, like many other categories used in popular language, is socially constructed, and this in turn influences statistical categories. Regardless of the chosen category, population aging is occurring across the globe, and this, together with the decline in individual driving ability with age, is the basis for concerns about the safety risk that elderly drivers constitute to themselves, their passengers and other road users. At the same time, reliance on personal transport has increased in many countries, while use of alternative forms of transport by elderly people has declined. Beyond this basic picture, however, there are various levels of complexity that influence the safety of elderly drivers, including demographic, developmental, social, and behavioral factors. This entry gives a brief overview.

The Aging of the World and Some Gender Differences

For several decades high-income countries have noted that their population have been characterized by an increasing proportion of elderly people, as a result of lower birth rates and longer life expectancy. In some countries population have fallen. However, population aging is already a feature across all regions of the world, except Africa, and even Africa is projected to exhibit an increase in the proportion of elderly people across the region by 2050 (Fig. 1). Note that in Asia, which accounts for about 60% of the world’s population, the proportion aged 60 years and over will approximately double.

The implications of these changes for the road safety of elderly people, especially as drivers, vary by region. The greatest uptake in driving as countries are motorizing tends to be among people of working age, so that initially the proportion of elderly drivers is less than the proportion of elderly people. However, once licensed, drivers tend to maintain their license and continue driving as they age, so that the proportion of elderly people who drive increases progressively and draws close to the proportions observed among people of working age.

Driving varies by gender, with males being more likely to be licensed than females, sometimes by quite large amounts. The reasons relate to a range of factors including gendered division of labor, traditional gender roles (including religious and cultural traditions), income inequity, and the degree to which legislation and administration of driving is influenced by any of these factors. In Western countries, female uptake of licensing increased in the 1970s, however males still tend to drive more and be more likely to drive for employment purposes.

This gender difference in driving is important because population aging also differs by gender. Women consistently live longer than men, by about 5 years in high-income countries, and around 4 years in other countries (Table A.28, United Nations, 2019). Given marriage practices which mean that men are on average older than women at time of marriage for all countries (with the possible exception of Nauru), most married women will live several years after their husband dies and are therefore more likely to find themselves reliant on other people for transport or needing to become licensed themselves for the first time (King and Scott-Parker, 2017).

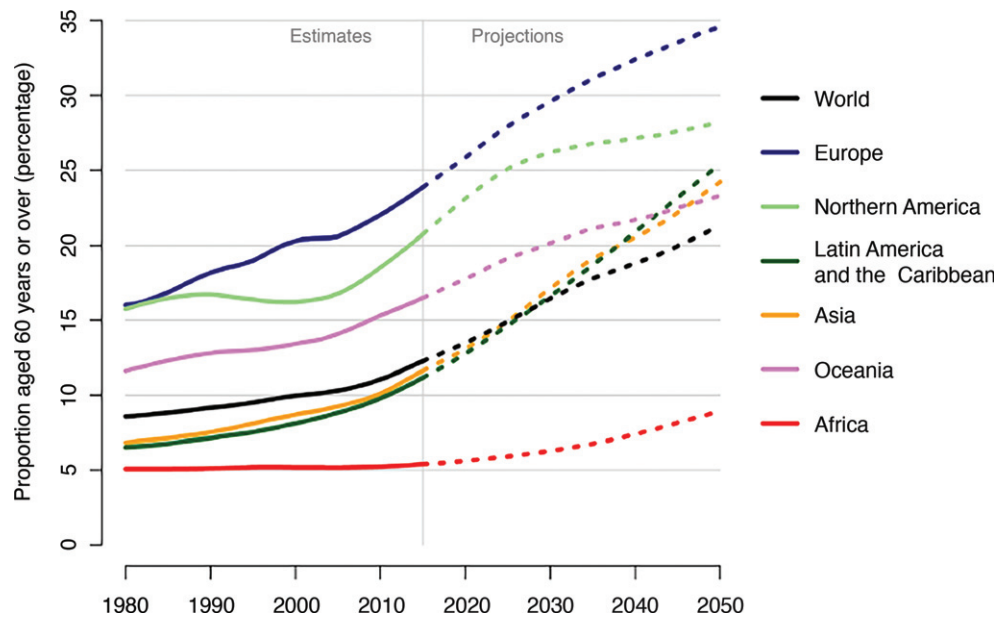


Figure 1 Percentage of population aged 60 years or over by region, from 1980 to 2050. Source: *United Nations (2017)*.

Ageing and Driving Performance

For road safety, the aging of the population is important because “normal aging” (an increasingly debated term) is associated with a decline in perceptual, cognitive, and physical abilities, as well as a decline in overall health. These changes have implications for crash risk and for crash severity.

Safe driving (i.e., lower crash risk) requires drivers to perceive the features of a constantly changing environment sufficiently well and focus on them, while aging is associated with declines in vision (especially night vision), hearing, and processing of sensory data. The driver must distinguish between relevant and irrelevant information once it has been perceived, predict likely outcomes of different motion/speed/direction combinations for one or more vehicles, and make sound decisions about safe courses of action, whereas aging is associated with declines in the ability to integrate information and make accurate decisions rapidly. Any decisions made by the driver must then be physically enacted through manipulation of the steering, acceleration, and braking functions, while aging is associated with declines in muscle strength, speed and precision of movement, and accuracy of proprioception. All of these age-related changes have been researched in laboratories, driving simulators, field studies, and (more recently) naturalistic driving studies. However, investigation of the crash risk associated with any particular effect of aging is methodologically challenging.

In addition, there is significant variability in the effects of aging from person to person, and there is a wide range of variability within any given age. *Fig. 2*, though not derived from a study on driving, gives an indication of the variability observed within people of a given age with respect to mental and physical capacity across a number of countries. The median scores remain within the range observed for young adults until the 90s are reached, where the sample size and that there is little difference between the worst performing individuals from the mid-50s on. The authors note that the trend in the mean does not match the trend for individuals, who may experience recoveries after declines. They also state that: “. . . for the population as a whole, the average reduction shown in this analysis was gradual. There was no age when people suddenly had less capacity and became ‘old’.” This has important and generally unappreciated policy implications that will be mentioned further.

Ageing and Injury Severity, and the Biases Involved in Focusing on Fatality Data

If a crash does occur, aging also contributes to greater injury severity. Ageing is associated with a loss of bone density and much slower healing, so that impact forces that lead to injury in younger people can be fatal in older people. This is most obvious when comparing the population rates at which people of different ages are killed or injured in traffic crashes (*Fig. 3*). Note that the different scales on the left and right axes of *Fig. 3* show that, up to the 50–59 years age group, there are about 55 injuries for each fatality, and the patterns are very similar, peaking for the 17–20 years age group and declining thereafter. For injuries, the rate continues to decline with age, which is consistent with a decrease in overall exposure (participation in traffic). However, from 60–69 years onward the fatality rate increases. Taken together with the decline in nonfatal injuries, this implies that while risk of injury in traffic crash decreases with age (at a population level), there is an increasing probability that the injury will be fatal from around 60–69 years.

These data came from an Australian crash database; while imperfect, the nonfatality data is reasonably reliably and comprehensive when compared to other countries. However, there is a tendency for media reports on road crash trends to focus on fatalities.

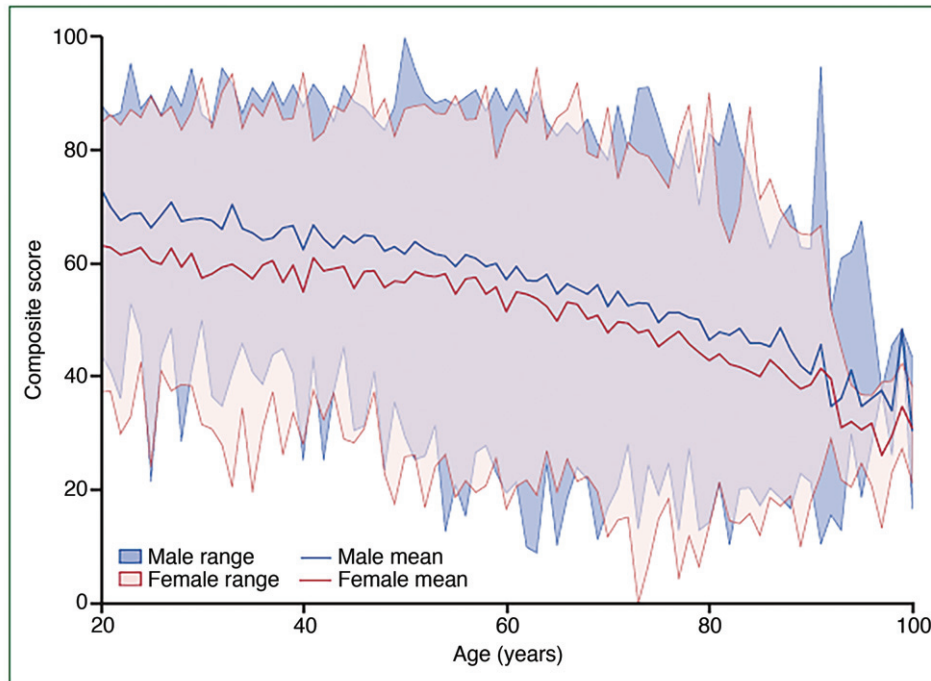


Figure 2 Range of scores on a composite of mental and physical capacities by age. Source: *Beard et al. (2015)*.

The fatality rates imply an increasing risk as people become older, consistent with expectations about how the effects of aging affect driving performance and other transport modes, but this is not borne out in injury data. Recent research from the same Australian state shows that all main forms of traffic exposure (as driver, passenger, or pedestrian) decline with age, so this decline in exposure is apparently sufficient to reduce the population rate of injury in a road crash for older people. Focusing on fatalities alone can provide a biased impression of the relationship between population aging and impact on road safety.

It is important to reiterate that *Fig. 3* is about population rates of fatality and injury, not the rate given exposure. The next section discusses mitigating factors like decreased exposure that lead to lower levels of older driver crash involvement than might be expected.

From Driving Performance to Crash Incidence—Mitigating and Exacerbating Factors

As explained earlier, an implication of population aging is that there will be a rise in fatalities as the proportion of drivers who are elderly increases, while the effects of aging on driving performance would be expected to increase the crash risk of older drivers, both of which imply that demographic change should lead to a significant road safety problem. In practice there has not been a marked change in crash involvement, a point that had been made by a number of prominent researchers over the past few decades.

However, as noted earlier, exposure in the form of miles/kilometers traveled decreases with age. When crash risk is calculated based on miles/kilometers traveled, it shows an increase in crash risk from 60 or 70 onward, even when the bias toward fatalities is taken into account. Further research has found that this is an oversimplified picture of road safety for elderly drivers that is distorted by an effect known as the “low mileage bias” (*Antin et al., 2017*). When older drivers are grouped according to how much driving they do in a year (their “mileage”), it is only the drivers with the lowest mileage that show an increase in crash risk per 100,000 km from around 60–70 years, with some studies showing that drivers in other mileage categories have a lower risk than all other drivers. Drivers who drive least are the high-risk drivers, and this group has been shown to be more likely to be frail, a term applied to older people who are more vulnerable to illness and injury, and in general terms are weaker, slower, and more prone to fatigue. Their impact on crash figures for elderly drivers is limited by their much lower mileage, and it is possible that they are more likely to give up driving voluntarily, or to be influenced to give up by doctors or family members.

Reduction of exposure by elderly drivers has been interpreted as an example of self-management of safety, reducing exposure to compensate for increased risk. For example, elderly drivers often state that they avoid driving situations such as peak hours and high-speed roads. However, it has been pointed out that this may have little to do with awareness of safety; instead, avoidance of peak hour traffic could be due to the combination of motivation to avoid frustration in traffic and flexibility in choice of when to drive. Similarly, avoidance of high-speed roads could be motivated by discomfort experienced when traveling at speed among other vehicles, since it is challenging to cope with the flow of information from the driving context at speed when a driver is older; the

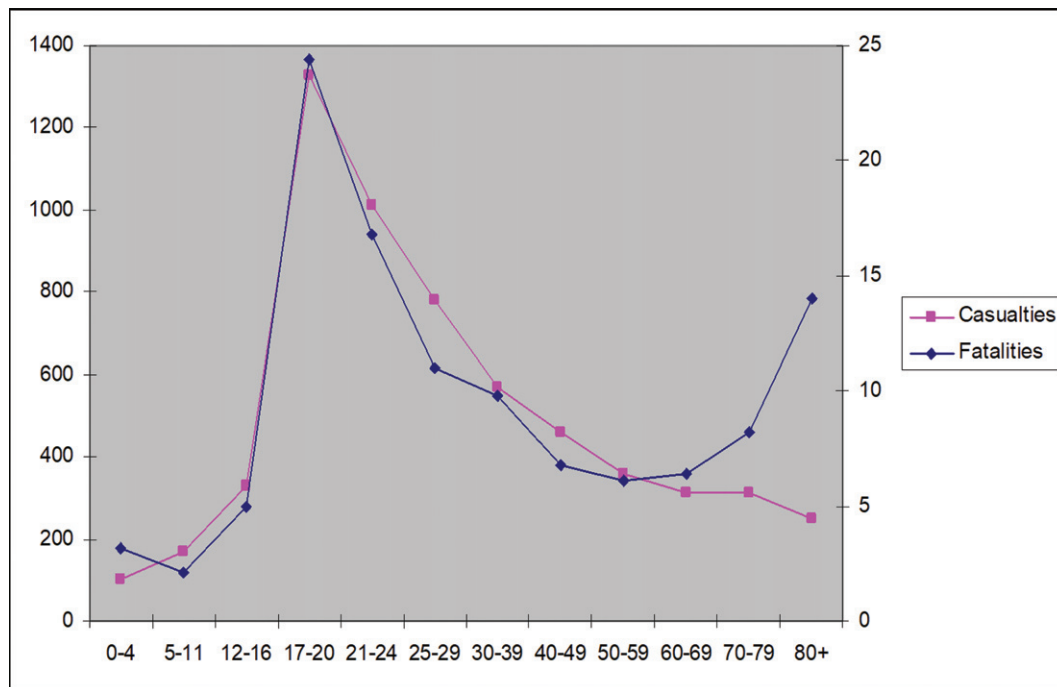


Figure 3 Median road traffic casualty and fatality rates per 105 population by age category (years), Queensland, Australia 2001-2005 (casualty rate left axis, fatality rate right axis). Source: King et al. (2007).

easiest way to deal with too high flow of driving-relevant information is to slow down to reduce the flow. On the downside, it is common for drivers in general to regard the posted speed limit almost as a mandatory speed, and elderly drivers report intimidation by other drivers (such as tailgating) to pressure them to speed up. In both cases, elderly drivers report feeling discomfort while driving, and the psychological construct of discomfort has been found to have better value in explaining older driver preferences than safety compensation.

A related issue is that self-management by elderly drivers assumes that they are aware of their declining skills, able to make decisions about how to manage the issue, and able to successfully implement the decision (Hassan et al., 2015). Since people at all ages vary in their self-awareness and may be prone to optimism bias (thinking they are “better than average” when they are not), and since the cognitive declines associated with aging may affect self-awareness, deliberate self-management (as opposed to incidental self-management via avoidance of discomfort, for example) may not be common. A complicating factor is that the final step in self-management is to be able to change behavior, but this has a contextual aspect: for example, if an elderly driver lives alone in an area with poor transport options, they may not be able to give up driving or even to reduce it much. Some countries are highly car-dependent, and research has shown that even drivers who are aware that their abilities have declined feel unable to stop driving. This is common in urban areas where one would expect more transport options to be present, and is explained by the elderly drivers believing that there are other people (family or friends) who rely on them being able to drive. Interventions aimed at familiarizing elderly drivers with their transport options and encouraging self-directed change show only short-term improvement in knowledge and intention.

Road Infrastructure Implications of Crash Patterns Among Elderly Drivers

The higher likelihood of fatalities among older drivers involved in crashes has been mentioned. There are other crash patterns that provide insights of varying value. As driver age increases, drivers are more likely to be involved in intersection crashes, although it is unclear whether this is due to greater city/suburban driving (rather than high-speed roads), the low mileage bias (since short trips will occur in the more densely populated areas around drivers’ residences) or declining ability to deal with the decision processes required to negotiate intersections (gap judgment, prediction of vehicle speed and trajectory). More important, as driver age increases, the driver is more likely to be considered “at fault” in their intersection crash and in crashes in general. The other driver age group with higher than average “at fault” crashes is young drivers in their first few years of driving, and a comparison with young drivers reveal an interesting difference. For young drivers, “fault” is more likely to be assigned because of a deliberate illegal, unsafe act, such as speeding or drink driving; for older drivers, it is more likely to be assigned for errors of judgment or poor decisions (Rakotonirainy et al., 2012). This is consistent with what is known about the effects of aging and the processes of maturation among young drivers, and as noted below, it has implications for interventions to address older driver safety.

In the 1980s and 1990s, road authorities devoted resources to understanding the implications of these crash patterns for road infrastructure design and implementation, and developed guidelines for design and construction. These have tended to be absorbed into standards and guidelines generally, although there are still gaps. For example, research indicates that assumptions about the required minimum standards for road signage, in particular the distance at which a sign can be recognized, only hold for nonelderly drivers in general, and fail to take into account the poorer contrast sensitivity and acuity of elderly drivers at night.

The greater challenges experienced by elderly drivers at intersections and in gap judgment suggests that signalized intersections with fully controlled turns would be desirable, however this has implications for traffic flow that can take priority. Roundabouts can address both safety and traffic flow objectives, provided demand is not too skewed across approaches at peak times, although elderly drivers tend to feel uncomfortable on multilane roundabouts due to the merging and lane crossing involved. A related infrastructure consideration is that the much higher rates of intersection crashes among elderly drivers means that, if they strike a fixed object, it is much more likely to be a pole than a tree; however the approach to pole installation and protection in urban areas does not approach the issue from an elderly driver perspective.

Normal aging is associated with changes in vision including lower contrast sensitivity, which at night becomes important because the visual scene is mostly perceived as various shades of gray. Elderly drivers may not be able to see pedestrians in dark clothing until they are too close to react, and like most other drivers will tend to drive at a speed that does not allow sufficient time from the appearance of a hazard in the headlights to stop in time (Wood, 2020). In many places there is a lack of road lighting to compensate, and elderly drivers are also more sensitive to glare effects. In addition to normal aging effects on vision, elderly drivers have a greater incidence of visual conditions such as glaucoma and advanced retinal maculopathy. Cataracts (a progressive clouding of the lens in the eye) are arguably part of normal aging (since most people eventually develop cataracts) and leads to both reduced visual acuity and glare sensitivity, however cataract surgery reverses its effects (Wood, 2020).

Vehicle Safety Implications of Elderly Driver Crashes

The much higher proportion of crashes that result in fatality rather than nonfatal injury for elderly drivers suggests that the common vehicle occupant protection measures that mitigate severity risk among most drivers are not as effective for elderly drivers. A common injury among fatally injured drivers is thoracic trauma, where the standard seat belt design leads to forces on the thorax that can be fatal for elderly drivers because of lower bone density. Alternative seat belt designs have been developed, though their use is not widespread.

For social and economic reasons, elderly drivers do not update their vehicles frequently, so that the benefits of more advanced safety features are not passed on as quickly. Conversely, elderly drivers experience more difficulty in adapting to new design features such as advanced driver assistance systems (ADAS), which they find distracting.

A less commonly explored area is the degree to which proprioceptive feedback from the vehicle is detected and responded to by elderly drivers. There are many crashes where an elderly driver presses the accelerator instead of the brake and does not appear to understand what has happened and how to correct it. As proprioceptive sensitivity in the legs declines with age (Lacherez et al., 2014), it is possible that elderly drivers find it hard to correctly place their feet and to detect the differences in feedback from pedal vibration and response that distinguish the accelerator and brake. There are ways of changing this feedback, but this is not part of vehicle design. An alternative and interesting line of research downplays this issue in favor of neural changes, arguing that older people's executive functioning declines, such that their ability to override or change an incorrect decision is harmed (Makizako et al., 2018).

Overall cars are not generally designed with aging in mind. In the same way that there has been a growing appreciation that standard crash test dummies do not represent the injury response of children and women, it can be argued that they do not represent the vulnerability of elderly vehicle occupants. More relevant crash testing could lead to the development of standards that ensure a much higher level of safety for elderly occupants.

Licensing and Medical Approaches to Elderly Driver Safety

Elderly drivers involved in crashes are often depicted in the media as examples of people who should not be allowed to drive because they do not meet the minimum standard required, hence arguments are advanced that older drivers need regular testing and need to conform to a strict standard. However, there is limited evidence that regular (or even targeted) testing of older drivers is an effective way of addressing elderly driver safety (Ichikawa et al., 2020). This may reflect the influence of other factors, for example, there is evidence that elderly drivers who know their ability is declining decide voluntarily to not renew their license, and that family members may directly or indirectly (through the family doctor) encourage the elderly driver to stop driving. There are also variations across countries in licensing practices for elderly drivers and the role of training and assessment: in some countries (e.g., Japan) licensing interventions for elderly drivers appear to be effective.

Also common in the media is the representation of elderly drivers as a risk to others, whereas the data on fatal crashes shows that if an elderly driver is involved, they are probably the person killed.

Often licensing approaches to elderly driver safety incorporate medical assessment. Such assessments cover the effects of normal aging as well as the wide range of health problems that are more common (though not "normal") with aging. Many of these

conditions require medications that independently impair driving—although medication can also have beneficial effects on driving. The great diversity of conditions, medications, combinations of medications, and individual differences in response make it practically impossible to apply simple rules about driving restrictions, leading to a reliance in many countries on medical practitioners to make judgments about whether their elderly client can drive safely. However, in spite of the development of guidelines to assist them, many medical practitioners do not feel comfortable making these decisions.

Giving up Driving

Ultimately, anyone who lives long enough will eventually be unable to drive safely, though there is large variation in when that might occur. As a consequence, there are many programs aimed at facilitating a shift from driving to other forms of transport. Unfortunately, the evidence shows that when a driver reaches the point where they are unable to drive safely anymore, their capability to take public transport or walk is also considerably reduced (Liddle et al., 2014). There are more subtle social and personal issues as well, since giving up driving means ceding control over part of one's life, and could be an important part of one's identity (especially among males). The loss or reduction of opportunities to socialize face-to-face, or simply to move around the transport system, is associated with poorer health and quality of life (O'Neill et al., 2019).

Are Automated Vehicles the Answer?

The advent of automated (self-driving) vehicles has been proposed as a way of solving the problem of eventual decline in safe driving ability with age: on the one hand, it is similar to a shift to standard public transport, except that it is more accessible; and on the other, the level of protection will be high. However, population aging is already proceeding at a rapid pace, while widespread use of fully automated vehicles may be many years away.

In the meantime the vehicle sector is experiencing a transition through the different levels of automation from a completely non-automated car (Level 0) to a fully automated car (Level 5). The lower levels of automation involve ADAS such as navigation systems which (as noted above) elderly drivers find distracting. For in-vehicle information systems in general, elderly drivers experience greater levels of cognitive and visual demand, and take longer to complete their interactions.

A significant issue prior to full automation is the need for drivers to take over control of an automated vehicle when unexpected circumstances arise. Although research in this area is limited, there is sufficient information to suggest that elderly drivers will find taking over control of an automated vehicle to be significantly more challenging than younger drivers. This field is evolving rapidly, however, and artificial intelligence approaches to perceiving and interpreting traffic situations may eventually be central to safe automated driving for all ages of driver.

References

- Antin, J.F., Guo, F., Fang, Y., Dingus, T.A., Perez, M.A., Hankey, J.M., 2017. A validation of the low mileage bias using naturalistic driving study data. *J. Safety Res.* 63, 115–120, doi.org/10.1016/j.jsr.2017.10.011.
- Beard, J.R., Officer, A., Araujo de Carvalho, I., Sadana, R., Pot, A.M., Michel, J.-P., Lloyd-Sherlock, P., Epping-Jordan, J.E., Peeters, G.M.E.E., Mahanani, W.R., Thiyagarajan, J.A., Chatterji, S., 2015. The World report on ageing and health: A policy framework for healthy ageing. *The Lancet* 387, 2145–2154, doi.org/10.1016/S0140-6736(15)00516-4.
- Hassan, H., King, M., Watt, K., 2015. The perspectives of older drivers on the impact of feedback on their driving behaviours - a qualitative study. *Trans. Res. Part F* 28, 25–39, https://doi.org/10.1016/j.trf.2014.11.003.
- Ichikawa, M., Inada, H., Nakahara, S., 2020. Effect of a cognitive test at license renewal for older drivers on their crash risk in Japan. *Inj. Prev.* 26, 234–L239, doi:10.1136/injuryprev-2018-043117.
- King, M.J., Scott-Parker, B.J., 2017. Older male and female drivers in car-dependent settings: how much do they use other modes, and do they compensate for reduced driving to maintain mobility? *Age. Soc.* 37 (6), 1249–1267, doi:10.1017/S0144686X15001555.
- King, M.J., Nielson, A., Larue, G., Rakotonirainy, A., 2007. Projecting the future burden of older road user crashes in Queensland. *Proceedings of the Australasian Road Safety Research, Policing and Education (ARSRPE) Conference 2007*. Available from: https://acrs.org.au/files/arsrpe/RS07043.pdf.
- Lacherez, P., Wood, J.M., Anstey, K.J., Lord, S.R., 2014. Sensorimotor and postural control factors associated with driving safety in a community-dwelling older driver population. *J. Gerontol. A Biol. Sci. Med. Sci.* 69 (2), 240–244, doi:10.1093/gerona/glt173.
- Liddle, J., Haynes, M., Pachana, N.A., Mitchell, G., McKenna, K., Gustafsson, L., 2014. Effect of a group intervention to promote older adults' adjustment to driving cessation on community mobility: a randomized controlled trial. *The Gerontol.* 54 (3), 409–422.
- Makizako, H., Shimada, H., Hotta, R., Doi, T., Tsutsumimoto, K., Nakakubo, S., Makino, K., 2018. Associations of near-miss traffic incidents with attention and executive function among older Japanese drivers. *Gerontology* 64 (5), 495–502, doi:10.1159/000486547.
- O'Neill, D., Walshe, E., Romer, D. and Winston, F., 2019. Transportation equity, health, and aging: a novel approach to healthy longevity with benefits across the life span. *NAM Perspectives*. Commentary, National Academy of Medicine, Washington, DC. doi.org/10.31478/201912a.
- Rakotonirainy, A., Steinhardt, D., Delhomme, P., Darvell, M., Schramm, A., 2012. Older drivers' crashes in Queensland, Australia. *Acci. Anal. Prev.* 48, 423–429.
- United Nations, Department of Economic and Social Affairs, Population Division (2019). *World Population Prospects 2019, Volume 1: Comprehensive Tables (ST/ESA/SER.A/426)*.
- United Nations, Department of Economic and Social Affairs, Population Division (2017). *World Population Ageing 2017 - Highlights (ST/ESA/SER.A/397)*.
- Wood, J.M., 2020. Nighttime driving: visual, lighting and visibility challenges. *Ophthalmic Physiol. Opt.* 40, 187–201, https://doi.org/10.1111/opo.12659.

Further Reading

- Austroads/National Transport Commission (2016, amended 2017). Assessing fitness to drive for commercial and private vehicle drivers. Austroads Publication AP-G56-17. Available from: https://austrroads.com.au/__data/assets/pdf_file/0022/104197/AP-G56-17_Assessing_fitness_to_drive_2016_amended_Aug2017.pdf.
- Dugan, E., Barton, K.N., Coyle, C., Lee, C.M., 2013. U. S. Policies to enhance older driver safety: A systematic review of the literature. *J. Aging Soc. Policy* 25 (4), 335–352, doi:10.1080/08959420.2013.816163.
- Guo, F., Fang, Y., Antin, J.F., 2015. Evaluation of older driver fitness-to-drive metrics and driving risk using naturalistic driving study data. Report #15-UM-036, National Surface Transportation Safety Center for Excellence, Virginia Tech Transportation Institute. Available from: <http://hdl.handle.net/10919/54824>.
- Hassan, H., King, M., Watt, K., 2017. Examination of the precaution adoption process model in understanding older drivers' behaviour: An explanatory study. *Transp. Res. Part F* 46 (A), 111–123, <https://doi.org/10.1016/j.trf.2017.01.007>.
- Reimer, B., 2014. Driver assistance systems and the transition to automated vehicles: A path to increase older adult safety and mobility? *Public Policy Aging Rep.* 24 (1), 27–31, <https://doi.org/10.1093/ppar/prt006>.
- Sangrar, R., Mun, J., Cammarata, M., Griffith, L.E., Letts, L., Vrkljan, B., 2019. Older driver training programs: A systematic review of evidence aimed at improving behind-the-wheel performance. *J. Safety Res.* 71, 295–313, <https://doi.org/10.1016/j.jsr.2019.09.022>.
- Wong, I.Y., Smith, S.S., Sullivan, K.A., Allan, A.C., 2016. Toward the multilevel older person's transportation and road safety model: A new perspective on the role of demographic, functional, and psychosocial factors. *J. Gerontol. Series B: Psycho. Sci. Social Sci.* 71 (1), 71–86., <https://doi.org/10.1093/geronb/gbu099>.

Emergency Response Systems

Frances L. Edwards, Mineta Transportation Institute, San Jose State University, San Jose, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	240
All Disasters are Local	240
World Trade Center	241
Pentagon	241
Emergency Response Structures	242
Incident Command System	242
Emergency Operations Centers	242
National Incident Management System	243
National Response Framework	243
Emergency Support Function (ESF)-1 Transportation	244
Emergency Support Function (ESF)-14 Cross Sector Business and Infrastructure	244
Critical Infrastructure	244
Sector-Specific Agencies	244
Mutual Aid	244
The Role of Transportation in Emergency Response Systems	245
References	245
Further Reading	246

Introduction

Many nations have organized emergency response systems to address natural, technological, and human-caused disasters. In most nations, the current emergency management structure began as a civil defense mechanism, often with the military in the lead, augmented by local volunteers. Many nations, such as China and Japan, still have a primary role in emergency response for the military resources. For example, after earthquakes the military units provide search and rescue and mass care services to local communities (Edwards et al., 2015). Great Britain has organized around the public safety services at the field, and administrative professionals in coordination roles, ultimately reaching up to Cabinet-level positions, using the Gold/Silver/Bronze system. Italy has adopted the Incident Command System nation-wide, developing identical training, and equipment caches in all regions to facilitate a coordinated response and effective mutual aid (Edwards and Steinhausler, 2007).

The United States of America (US) has a multilevel approach that starts with the municipality, and includes the state and the federal government in delivering services and support. The field level may have the fire service as the incident commander in a life-safety disaster, law enforcement as the incident commander in a criminal or terrorist disaster, or even public works in a flood or infrastructure-related disaster. The community's city leaders open an emergency operations center to provide coordinated support to the field forces. The US system is explained in detail, and might be a useful model for other federal nations.

The emergency response system in the United States is based on the National Incident Management System (NIMS), and the National Response Framework (NRF) and its component parts (DHS, 2019). The fourth edition of the NRF was issued in October 2019, describing the overarching systems for responding to emergencies, disasters, and catastrophes throughout the United States. It includes Emergency Support Function (ESF) #1—Transportation, and in 2019 added a new ESF #14 that is focused on cross sector issues, including supply chain transportation.

Transportation is designated as a critical infrastructure within this system. Transportation's role in providing road access, evacuation routes and capabilities, access and egress to the site of an emergency, and support of logistics activities makes transportation central to emergency response. Transportation is also the key to the national and international supply chain on which the economy depends. Transportation in turn depends on electricity for signals, pipelines for fuel delivery, and communications and information technology resources for the management of the service delivery systems.

All Disasters are Local

The current American approach to emergency response derives from lessons learned during the terrorist attacks on the United States on September 11, 2001 (9/11). The immediate emergency response to the plane crashes was the responsibility of the local governments' first responders, as fire fighters and paramedics tried to put out fires, rescue survivors, and deliver immediate medical care, while law enforcement personnel managed traffic, interviewed witnesses and survivors, and collected crime scene evidence, and transportation personnel provided traffic control and engineering services. Homeland Security Presidential Directive 5 (HSPD-5) recognizes that "Initial responsibility for managing domestic incidents generally falls on State and local authorities" (Bush, 2003, Section (6)).

World Trade Center

Two commercial airplanes were flown into World Trade Center Tower One and World Trade Center Tower Two, causing structural damage to the buildings and fires fueled by the jets' fuel, building contents, and diesel fuel stored in the buildings to operate the emergency generators. In less than 2 h, after the air crashes, both towers had collapsed (Edwards and Steinhäusler, 2007). The collapse of the two towers destabilized the plaza on which they were sitting, and caused World Trade Center Building Seven to collapse. Debris from the falling towers did irreparable damage to the rest of the World Trade Center buildings, all of which had to be demolished. The subway and Port Authority Trans Hudson (PATH) train stations under the World Trade Center were severely damaged, and the plaza had to be shored up to prevent further damage to the city's transit systems (Jenkins and Edwards-Winslow, 2002).

The scope of the World Trade Center disaster was so great that managing it effectively was difficult. The two collapsed towers left a smoldering pile of debris that included hundreds of people who had been inside. Volunteers from public safety departments across the nation arrived to help with the search and rescue, and ultimately search and recovery, work. While New York State Department of Environmental Services Police forces patrolled the perimeter to prevent unauthorized access, a lack of a formal check-in systems meant that uniformed personnel were generally permitted access to the pile. Coordination of the multiple personnel from various jurisdictions and professions was impossible (National Commission on Terrorist Attacks, 2004).

In New York City, Mayor Rudy Giuliani became known as "America's mayor" as he appeared on television giving updates on the World Trade Center response and rescue efforts, and delivering calming messages to the public. Local public health agency and fire department personnel responded to the elderly residents who were stranded when the South of Houston area was evacuated, and they were unable to leave their walk-up apartments (Jenkins and Edwards-Winslow, 2002). New York City Transit (NYCT) and Metropolitan Transportation Authority (MTA) engineers designed the shoring systems used to protect the subway systems from further damage, after the World Trade Center Towers collapsed, and the plaza on which they sat was destabilized. Iron workers and welders from NYCT and MTA joined the police and fire personnel working on "the pile"—as the Tower One and Tower Two building collapse area was known—to cut large pieces of steel that had survived the collapse to open the way for rescuers. Transit heavy equipment operators removed large debris, and took down the remains of damaged World Trade Center buildings that remained partially standing (Siano and Joseph, 2002).

Federal Emergency Management Agency (FEMA) personnel helped to organize additional assistance from outside of New York City, but the initial response in the first crucial hours was conducted by personnel from New York City and its immediate neighboring jurisdictions. FEMA leaders often affirmed that "all disasters are local", even in this catastrophic attack because, the first people to arrive at the scene and help the survivors are local public safety and public works personnel. Assistance from outside entities and federal organizations comes later, and most often in the form of specialized teams like search and rescue and medical teams, or as funding for contracting with specialized resources.

Pentagon

The Pentagon was the second site attacked by terrorists on 9/11. The building is the Department of Defense headquarters, and so served as a symbol of America's military might. It was built during World War II, and was undergoing renovation when the attack occurred. The building has five sides—thus its name—and three interior corridors that decrease in length moving toward the courtyard in the center. The five sections are called wedges. The wedge that was struck by the airplane had just completed its repairs and was starting to be repopulated, which accounts for the low occupancy of the wedge that was the site of the direct hit. The plane flew through the wedge, stopping before crossing the interior courtyard, causing a fire and partial collapse of the interior structures (National Commission on Terrorist Attacks, 2004).

The Arlington County, Virginia Fire Department was the fire and emergency medical response agency. They had practiced fire drills with Pentagon staff and had a good history of collaboration across civilian and military organizational lines using the Incident Command System (ICS) (National Commission on Terrorist Attacks, 2004), which is described in detail later.

ICS had been created in southern California in the early 1970s, after multiple fire departments worked together on large wildland urban interface fires in Malibu Canyon, Topanga Canyon, and Escondido. Based loosely on the US military's hierarchical structure, it became the western United States' proven method for multi-agency command and control in firefighting.

The National Fire Academy had begun teaching ICS as a best practice in the 1980s, and Arlington Fire Department was among those that adopted it. On 9/11, Arlington Fire Department Assistant Chief James Schwartz used ICS to manage the response to the Pentagon fire, building collapse and multiple casualty incident, coordinating the response with the Federal Bureau of Investigation (FBI) and military authorities (Titan Systems, 2002).

The ICS provides a hierarchical organization that defines the work of an emergency response—command, operations, planning/intelligence, logistics and finance/administration. It includes a safety officer within the command section and a check-in/check-out procedure in the planning/intelligence section. These structures ensure that all personnel operate safely, using the proper safety equipment, and that they are accounted for at all times. All personnel are assigned to a specific supervisor, and these assignments are recorded on a board to ensure that everyone has oversight during the operation (Edwards, Goodrich & Griffith, 2015). The Pentagon grounds are secured by the military, so controlled access to the crash site was already established and maintained by the Pentagon's military police (National Commission on Terrorist Attacks, 2004).

In the after action report on the Pentagon attack, the Incident Command System was identified as a practice “that others should emulate” (Titan Systems, 2002, p. 11). The response to the Pentagon was judged to be a comprehensively successful emergency operation, in contrast to the less structured response in New York City (National Commission on Terrorist Attacks, 2004). As a result, ICS was recognized as a successful management tool for multijurisdiction, multidisciplinary emergencies.

Emergency Response Structures

Incident Command System

ICS originated in the southern California fire service in the early 1970s. After providing mutual aid to each other in several major wildland urban interface fires, the fire chiefs of southern California, the California Division of Forestry and the U.S. Forest Service created an organization, Firefighting Resources of Southern California Organized for Potential Emergencies (FIRESCOPE) that oversees the development of ICS doctrine and training resources. They publish the *Field Operations Guide* that provides a checklist for every position of the ICS (U.S. Fire Administration, 2016). Within a few years, the ICS method had spread throughout California and to other western states. In the 1980s, it was also adopted as a best practice by the National Fire Academy.

ICS is used to manage emergencies in the field. It is based on a hierarchy that ensures unity of command. It is also flexible, allowing the Incident Commander to staff the four general staff sections as the event requires, retaining all unstaffed responsibilities. It is based on common terminology, so that public safety staff members from different communities or professions can communicate effectively during an emergency. All event-related communications are to be conducted in plain English—no codes—to ensure that all participating professions and departments have a clear understanding of what is being said.

ICS has five main functions: command at the top with four subordinate functions—operations, planning/intelligence, logistics, and finance/administration. The command leader is in overall charge of the event, and is called the Incident Commander (IC). The IC is supported by the command staff consisting of a public information officer, safety officer and liaison officer. The other four sections are collectively called the general staff. The ICS elements jointly operate out of an Incident Command Post (ICP), which may be as simple as the back of the IC’s vehicle at a house fire, or as elaborate as a portable office complex at a large scale wildland fire. The IC makes the decisions on how to attack and resolve the emergency, usually with input from the general staff. The IC determines what actions to take, within what timeframe, and with the resources at hand. The plan may include a list of additional resources that must be obtained for future operational success. These decisions, along with a map of the event, an organization chart of the staffing, and any supporting plans and materials, are collected by the Planning/Intelligence Section Chief, and turned into the Incident Action Plan (IAP). This is posted at the ICP for the use of all ICS sections. In a large event, it may be printed and distributed in writing.

The Operations Section Chief is in tactical charge of the personnel who are working to resolve the event. He takes the IAP of the IC and applies the directions to the resolution of the event. Branches under Operations usually include firefighting, search and rescue, hazardous materials, law enforcement, emergency medical services, and transportation. Operations personnel assesses the damage and analyzes conditions related to safe operations around damaged buildings and structures.

The Planning/Intelligence Section Chief oversees the collection of vital information about the event, makes maps of the event, creates situation status reports, documents all activities at the event, and develops the demobilization plan for the end of the event. Resource accountability is managed through the check-in/check-out function for personnel, and a resource tracking system for goods and services that have been requested, received and returned. Staging areas are designated for stockpiling requested goods, and holding personnel and equipment that has been received, but not yet deployed.

The Logistics Section Chief is responsible for all support services for the event. The communications unit is responsible to provide standard fire radio systems, Internet connections and cell phone services. Some jurisdictions also have interoperable radio systems that allow fire, law enforcement and emergency medical services personnel to speak on a common frequency. The Personnel Unit arranges for additional staffing, staffing for shift change, and tracks overtime costs. The food unit organizes meals for the responders. Other units may include procurement, to acquire additional resources, facilities to acquire space for activities, and transportation to acquire heavy equipment or nonfire support vehicles.

The Finance/Administration Section Chief is responsible for tracking all costs associated with the event, and to document costs for potential reimbursement from responsible parties or higher levels of government. For example, during terrorist events or catastrophic natural disasters, the federal government may reimburse up to 75% of the documented emergency response costs (FEMA, 2016). Units include compensation and claims, cost recovery, and timekeeping.

At the end of the operation, an after action review is held that involves all the section chiefs and other key participants in the field-level response. An open discussion includes what actions went well, what strategies need to be changed, and what activities failed or were not productive and should be eliminated from future responses. This AAR analysis is written as a report that includes a summation of the event, its outcomes, and an improvement matrix aimed at incorporating the lessons learned from the management of the event into organizational doctrine and plans to guide future emergency response (U.S. Fire Administration, 2017).

Emergency Operations Centers

Most jurisdictions also have an Emergency Operations Center (EOC), which may be found at the city, county, special district and state levels. The National Operations Center (NOC) is the EOC for the Department of Homeland Security. While ICS focuses on the

tactical management of an event at the field level, the EOC manages the event from a strategic perspective, considering the needs of the emergency and the unaffected portions of the community (FEMA, 2017). The EOC goes beyond the public safety services and accesses resources across all jurisdiction departments. For example, at an apartment house fire, the IC would be focused on moving victims out of harm's way, while the care and shelter unit of the EOC would collaborate with the transit agency to get busses to move the victims to shelter, and the Non-Governmental Organization (NGO) community to open a mass care center with food, sleeping accommodations, and medical surveillance. This might include social services case work and connecting survivors to community-based support services and government assistance programs.

In many states the EOC is organized using ICS. The main difference is that the top position is the Management Section headed by the EOC Director, whose role is strategic (management) rather than tactical (command) as in the field. The general staff positions are the same as in ICS, and generally fill the same roles, albeit with an eye to larger community concerns as well as emergency response needs (Edwards and Goodrich, 2012). For example, the community's fire chief might be the Operation Section Chief in the EOC, and have to consider deploying resources for the ongoing emergency, while recognizing the need to respond to developing individual emergencies, like calls for medical services from unaffected community members.

The EOC can also provide more complex support elements, such as GIS mapping, situational status reporting to higher levels of government, and updating elected officials regarding the emergency event. The Management Section Chief can also advise the governing body on whether to declare a local disaster, and begin the process of requesting assistance from higher levels of government.

Some communities prefer not to use ICS in the EOC. NIMS 2017 offers two other options. Jurisdictions might use the Incident Support model, in which the EOC is focused on information, planning and resource support. The jurisdiction's leadership retains the EOC Director role, but this model separates situational awareness functions from planning, creating two sections, merges Operations and Logistics to form a resources support section, and leaves the finance functions in a center support section. The other option is called the departmental structure. The EOC is organized around the jurisdiction's normal departments, with the day-to-day leader as the EOC Director, supported by the department heads whose resources have to be part of the emergency response (FEMA, 2017).

National Incident Management System

Following the 9/11 attacks, new structures were created to encourage the development of nationwide coordinated systems for responding to future terrorist attacks or natural hazards events. President George W. Bush issued HSPD-5: Management of Domestic Incidents (2003). HSPD-5 Section 15 directs the Secretary of the Department of Homeland Security (DHS) to create the National Incident Management System (NIMS), based on ICS and its components.

Complying with HSPD-5, DHS created NIMS in 2004, and updated it in 2008 and 2017. NIMS incorporates the use of ICS in the tactical field response, and the creation of emergency operations centers for strategic incident management at all levels of government. "NIMS applies to all incidents, from traffic accidents to major emergencies" (FEMA, 2019, n.p.). NIMS applies to government agencies at all levels, to the non-governmental sector (NGOs), and to the private sector. Thus, all elements of transportation are required to use NIMS in emergency response. This comprehensive approach is reiterated in the Federal Highway Administration's (FHWA) Manual of Uniform Traffic Control Devices (MUTCD) which states, "The National Incident Management System (NIMS) requires the use of the Incident Command System (ICS) at traffic incident management scenes" (FHWA, 2009, p. 726), where multiple public safety entities and private sector organizations may be coordinating to resolve the event.

Like ICS, NIMS is based on flexibility, standardization and unity of effort. It addresses all phases of emergency management: prevention, protection, response, recovery and mitigation. It focuses on three areas: resource management, command and coordination, and communication and information management. While ICS assigns the public information officer to the IC in the field, or the EOC Director, NIMS creates a Joint Information Center, where PIOs from the responding agencies work together to create a unified message across jurisdictions and professions based on a common operating picture. Communication across jurisdictional and professional boundaries is a key tenant of NIMS, along with the creation of the common operating picture based on incident reports and IAPs. (FEMA, 2017)

National Response Framework

The National Response Framework (NRF) is the guide that operationalizes the NIMS emergency response element. The fourth edition was issued on October 28, 2019 to add more focus on community lifelines in emergency preparedness, response and immediate recovery activities (first 72 h). As a result of lessons learned during the 2017 disasters—floods, wildland fires and hurricanes—the framework changed the orientation of the response to developing outcomes—"What exactly are we trying to achieve?" (FEMA, 2019). Recognizing that 85% of the lifeline services are privately owned—such as electrical utilities, fuel pipelines, and supply chain transportation vehicles—the fourth edition stresses the need for "a larger and more focused role" for the private sector, and includes the development of a National Business Emergency Operations Center. (DHS, 2019c, p. ii). During Presidentially declared major disasters, FEMA has daily calls to coordinate logistics requests, recognizing that "government alone cannot meet community needs", and "the private sector can solve problems better because, they do it every day" (FEMA, 2019). The lifeline construct defines seven sectors: safety and security; communications; food, water and shelter; transportation; health & medical; hazardous materials; and energy, power and fuel.

The National Response Framework includes fifteen ESFs (DHS, 2019a). The federal government organizes its response resources around these, including ESF #1—Transportation (DHS, 2016), and ESF #14—Cross Sector Business and Infrastructure (DHS, 2019b). Each ESF has a lead agency, and many other federal agencies that provide support and resources.

Emergency Support Function (ESF)-1 Transportation

The US Department of Transportation is the coordinator and lead agency for ESF #1, while 10 other federal agencies are in support. ESF #1 responds to requests from local, state, territorial and tribal governments, and in support of NGOs and the private sector, for disaster assistance. They can mobilize federal resources, or they can contract with the private sector to provide the needed services. For example, they may coordinate with the Department of Defense to provide “high water vehicles”—trucks with very large tires—to take supplies to flooded communities, or contract with a tour business for limousine busses to evacuate survivors away from a devastated community. The state receiving the assistance pays a 25% cost share for the resources during a Presidentially declared major disaster. Some states in turn share this cost with the municipality or county, where the disaster occurred.

Emergency Support Function (ESF)-14 Cross Sector Business and Infrastructure

The Department of Homeland Security’s Cyber and Infrastructure Security Agency (CISA) is the coordinator for ESF #14, with CISA and FEMA as the primary agencies, and 15 federal departments and agencies are in support. While transportation needs of communities, NGOs, and the private sector may still be met primarily through ESF #1, ESF #14 was created to ensure cross-sector coordination, such as the need for supply chain and lifeline stabilization. For example, transit systems rely on electricity for powering the trains, for providing street-level traffic signals for busses to travel safely, and for fare-related transactions like purchasing tickets or using a credit card for passenger access to the system. Supply chains rely on global positioning systems, fuel delivery, and open roadways. All of these systems rely on communications, including telephones, radios, and the Internet. The goal of ESF #14 is to ensure that these interconnected systems recover in a coordinated and expeditious manner to restore services to the impacted communities. While public transit and transportation organizations would still coordinate primarily through ESF #1, the ability to have cross sector linkages with privately owned fuel supplies and power operators could enhance response and recovery activities (FEMA, 2019).

Critical Infrastructure

Presidential Policy Directive-21 names 16 sectors as critical infrastructure. They represent assets, systems and networks whose loss would debilitate national safety and security. Transportation Systems is one of the sectors (DHS, n.d.). The Department of Homeland Security (DHS) and the Department of Transportation (DOT) partner ensure that people and goods move safely and securely to domestic and overseas destinations. Cyber security is an important part of this sector’s functioning, with its dependence on supervisory control and data acquisition (SCADA) systems for operations, computer-based traffic signals, global positioning systems (GPS) for navigation, radio frequencies for positive train control, the Internet, and cell service. The focus of transportation sector security is on understanding and managing risk to avoid emergencies (DHS, DOT, 2015), but natural hazards, technological accidents, and terrorist attacks can all impact these systems, regardless of the validity of the risk analysis and mitigation efforts. Therefore, effective emergency response to the transportation sector’s critical infrastructure is the fallback when preparedness fails.

Sector-Specific Agencies

Sector-Specific Agencies (SSA), in this case DHS and DOT, maintain plans for emergency response that conform to the requirements of the National Response Framework, in compliance with HSPD-5 and Presidential Policy Directive (PPD)-8 that calls on the whole community to collaborate for emergency preparedness and response (DHS, 2011). The transportation sector supports evacuation, rescue, medical care and incident management (DHS, DOT, 2015). Its activities are aligned with the National Infrastructure Protection Plan (DHS, 2013), and focus on preparing for terrorist attacks through “identify, deter, detect, and deny access” strategies, and on mitigating the impacts of natural hazards through redundancy and cross sector collaboration.

Mutual Aid

Not all local government jurisdictions are capable of managing all emergencies on their own. Few have the local resources to manage catastrophic events, especially, when the first responders are victims themselves. For example, the response to the 9/11 World Trade Center attack included personnel from jurisdictions all over the United States, and even volunteers from other countries. Today mutual aid is delivered through organized systems, including local and state mechanisms for coordinating assistance requested by the jurisdiction at risk. For example, in California, there is a statewide mutual aid system for police that is coordinated by the sheriffs of the counties, and a fire mutual aid program that is coordinated by the county fire chiefs. Building officials’ mutual aid is managed by volunteer coordinators who are members of the Structural Engineers Association of California (SEAOC), and emergency managers’ mutual aid is managed by the Governor’s Office of Emergency Services (California OES, 2012). In most states, the requests for mutual aid are generally coordinated through the state’s office of emergency management and its regional components, and may be delivered for free or with specified reimbursement costs.

The Emergency Management Assistance Compact (EMAC) is a national organization that coordinates state-to-state assistance. EMAC receives requests for assistance and circulates the requests to the other states. States respond to mutual aid requests with offers of resources and their costs, including overtime, equipment use fees, travel, and per diem expenses (EMAC, 2020).

The Role of Transportation in Emergency Response Systems

As demonstrated by its position as ESF #1, transportation capability underpins all other emergency response. The physical elements of the transportation system constitute critical infrastructure, which may be owned by local governments—city streets and county roads—the state or the federal government. Until the roads have been cleared of debris and the bridges have been inspected and found passable, the rest of the emergency response cannot proceed into the disaster area.

The first responders may need to be transported to the site of the emergency. Residents of the disaster area may need to be evacuated to a safer location, or even out of the state, as was the case in Hurricane Katrina. This will require not only safe roadways, but also busses from both public transit agencies and private sources. All this movement of first responders and residents requires a reliable supply of fuel—for the debris clearance vehicles, for the road and bridge inspectors' vehicles, for the transit vehicles, and for private vehicles as residents evacuate themselves. Immediate recovery requires robust transportation capability. Food, water, diapers, baby formula, and building materials need to be brought into the community to care for the survivors and begin emergency repairs and short-term recovery.

State-level transportation agencies are responsible for the state highway system and elements of the national highway system in their jurisdiction. They also coordinate intercity busses and general aviation, which may be additional assets in disaster response. They inspect the infrastructure for safety, and issue overweight permits for heavy equipment that has to travel into the disaster area, especially if the highway system roadways are impassable. The state transportation agency is often the point of contact for ESF-1 in the state's emergency operations center, tying together the needs of the disaster survivors with national assets, public and private.

References

- Bush, G.W., 2003. Homeland Security Presidential Directive/HSPD-5. Management of Domestic Incidents. The White House, Washington, DC.
- California OES. (2012). State of California Emergency Management Mutual Aid Plan. November. <https://www.caloes.ca.gov/PlanningPreparednessSite/Documents/EMMA%20PlanAnnexes%20A-F,%202012.pdf>.
- DHS. (n.d.). Critical Infrastructure Sectors. <https://www.dhs.gov/cisa/critical-infrastructure-sectors>.
- DHS, 2019a. Emergency Support Functions Retrieved from <https://www.fema.gov/media-library/assets/documents/25512>.
- DHS, 2016. Emergency Support Function 1: Transportation Retrieved from file:///C:/Users/Primary/Documents/111%20WORK/presentations%20and%20publications/2019/encyclopedia%20of%20transportation/ESF_1_Transportation_20160705_508.pdf.
- DHS. (2019b). Emergency Support Function 14: Cross-Sector Business and Infrastructure. Retrieved from https://www.fema.gov/media-library-data/1572358162675-d2c7af34a5-b5063e582ae1798b038351/ESF14AnnexFINAL508c_20191028.pdf.
- DHS. (2013). National Infrastructure Protection Plan (NIPP) 2013: Partnering for Critical Infrastructure Security and Resilience. <https://www.dhs.gov/sites/default/files/publications/national-infrastructure-protection-plan-2013-508.pdf>.
- DHS, 2019c. National Response Framework, 4th edition October 28. Retrieved from https://www.fema.gov/media-library-data/1572366339630-0e9278a0ede9ee129025182b4d0f818e/National_Response_Framework_4th_20191028.pdf.
- DHS, 2011. Presidential Policy Directive-8: National Preparedness <https://www.dhs.gov/presidential-policy-directive-8-national-preparedness#>.
- DHS, DOT. (2015). Transportation Systems Sector-Specific Plan. <https://www.dhs.gov/sites/default/files/publications/nipp-ssp-transportation-systems-2015-508.pdf>.
- Edwards, F.L., Goodrich, D.C., 2012. Introduction to Transportation Security. CRC Press, Boca Raton, FL.
- Edwards, F.L., Goodrich, D.C., Griffith, J., 2015. Incident Command System (ICS) Training for Field Level Supervisors and Staff, doi:10.17226/23411 NCHRP Web-Only Document 215.
- Edwards, F.L., Goodrich, D.C., Hellweg, M., Strauss, J.A., Eskijian, M., Jaradat, O., 2015. Great East Japan Earthquake, Jr East Mitigation Successes, And Lessons for California High-Speed Rail. Mineta Transportation Institute Report 12-37. <https://transweb.sjsu.edu/sites/default/files/1225-great-east-japan-earthquake-lessons-for-California-HSR.pdf>.
- Edwards, F., Steinhäusler, F., 2007. NATO and Terrorism - On Scene: Emergency Management After a Major Terror Attack. NATO Security through Science Series-B: Physics and Biophysics—Vol. XX. Springer Science, Dordrecht, Netherlands.
- EMAC. (2020). What is EMAC? <https://www.emacweb.org/index.php/learn-about-emac/what-is-emac>.
- FEMA. (2017). National Incident Management System. October 17. <https://www.fema.gov/national-incident-management-system>.
- FEMA, 2016. Robert T. Stafford Disaster Relief and Emergency Assistance Act. <https://www.fema.gov/media-library-data/1519395888776-af5f95a1a9237302af7e3fd5b0d07d71/StaffordAct.pdf>.
- FEMA, 2019. The National Response Framework gets an update. Episode 31. FEMA Podcasts October 31. <https://www.fema.gov/media-library/assets/audio/177436>.
- FHWA, 2009. Chapter 6i. Control of Traffic Through Traffic Incident Management Areas. Manual of Uniform Traffic Control Devices https://mutcd.fhwa.dot.gov/kno_2009r1r2.htm.
- Jenkins, B.M., Edwards-Winslow, F.L., 2002. Saving City Lifelines. Report # 02-06. Mineta Transportation Institute, San Jose, CA <https://transweb.sjsu.edu/sites/default/files/02-06.pdf>.
- National Commission on Terrorist Attacks, 2004. The 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks Upon the United States. July 17. W.W. Norton & Company, New York.
- Siano, Joseph N., PE. "Restoration of Passenger Transportation to Lower Manhattan." Presentation to American Society of Civil Engineers, North Jersey Branch, The Newark Club, February 21, 2002.
- Titan Systems, 2002. Arlington County After Action Report on the Response to the September 11 Terrorist Attack on the Pentagon. Department of Justice, Washington, DC <http://www.policefoundation.org/wp-content/uploads/2018/07/pentagonafteractionreport.pdf>.
- U.S. Fire Administration. (2017). After Action Reviews: The good, the bad and why we should care. https://www.usfa.fema.gov/current_events/111617.html.
- U.S. Fire Administration. (2016). Field Operations Guide. ICS 420-1. https://www.usfa.fema.gov/downloads/pdf/publications/field_operations_guide.pdf.

Further Reading

- DHS, 2019. 2019 National Preparedness Report https://www.fema.gov/media-library-data/1575309879997-d19134cec727cfa75c17f4051c4f88aa/2019_NPR_Final_508c_20191119.pdf.
- DHS, CISA, n.d. National Critical Functions Resources. <https://www.cisa.gov/publication/national-critical-functions-resources>.
- FEMA, n.d. Community Lifelines. <https://www.fema.gov/lifelines>.
- FEMA, 2013. Critical Infrastructure and Key Resources Support Annex <https://www.fema.gov/media-library/assets/documents/32261>.
- FEMA, 2018. National Disaster Recovery Framework <https://www.fema.gov/national-disaster-recovery-framework>.
- FEMA, various. Podcasts. <https://www.fema.gov/podcast>.
- FEMA, 2013. Private Sector Coordination Support Annex <https://www.fema.gov/media-library/assets/documents/32261>.

Emergency Vehicles and Traffic Safety

Shamsunnahar Yasmin^{*,†}, Sabreena Anowar[‡], Richard Tay[§], ^{*}Queensland University of Technology (QUT), Centre for Accident Research and Road Safety—Queensland (CARRS-Q), Brisbane, QLD, Australia; [†]Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL, United States; [‡]Department of Civil and Environmental Engineering; Department of Architectural Studies, University of Missouri, Columbia, MO, United States; [§]RMIT University, Melbourne, VIC, Australia

© 2021 Elsevier Ltd. All rights reserved.

Background	247
EV Crash Statistics	247
Objective and Scope	248
Analyzing Emergency Vehicle Crashes	248
Emergency Vehicle Crash Data	248
Analytical Contexts	249
Methodological Overview	249
Crash Contributing Factors	251
Crash Characteristics	251
Roadway Characteristics	251
Environmental Characteristics	252
Driver Characteristics	252
Vehicle Characteristics	252
Research Framework	252
Some Insights and Concluding Remarks	253
References	253
Further Reading	254

Background

Emergency response services are an integrated part of our modern society with the overarching role of ensuring public safety and security. Emergency vehicles (EVs), such as ambulances, police cars, and fire trucks, are fundamental components of most emergency response systems. EV ensures timely response to the scene of emergency and delivery of vital services during an emergency. Road traffic crashes involving EVs responding to an emergency situation may lead to significant delay in the time of emergency response. It will also reduce the capacity and efficiency of the emergency response system. Therefore, road traffic crashes involving EVs are considered to have greater societal and economic impacts than a regular crash as these unfortunate events contribute to additional difficulties for the people already in distress. The social costs of EVs involved in road traffic crashes are thus expected to be higher than other road traffic crashes (NHTSA, 2011). Moreover, if an ambulance carrying a patient gets involved in a road crash, it not only poses a threat to the patient along with EV occupants and other road users (if involved), the situation also gives rise to the violation of core medical principle of “first do no harm” (American Medical Association, 1903).

The crash-related vehicular claims (Compulsory Third-Party Claim) also impose greater liabilities for emergency agencies. Because of their liabilities, EV drivers are generally also under greater scrutiny than other occupation-related driving. Moreover, EV crashes are considered as high-profile crashes and tend to attract significant media attentions. Also, although crashes involving EVs represent a small proportion of road traffic crashes, they are reported to be more severe than other crashes mainly because of the operational differences between these two groups of road users. Therefore, road traffic crashes involving EVs are major concerns not only for road users involved but also for the relevant authorities and the wider community.

EV Crash Statistics

Motor vehicle crashes are the second leading cause of firefighter deaths in the United States (Donoughe et al., 2012). From 2000 to 2009, more than 31,600 crashes involved fire trucks in the United States (Donoughe et al., 2012); while between 2004 and 2013, 179 firefighters died as a result of road crashes (U.S. Fire Administration, 2014a). They are likely to be killed after being struck by vehicles while either directing traffic or performing roadside emergency rescues. EV crashes involving other road users also pose greater threat to the occupants of other vehicles, pedestrians, and bicyclists. From 1996 to 2012, 137 civilian fatalities and 228 civilian injuries were reported from fire truck-involved crashes (U.S. Fire Administration, 2014b). Among 107 fatalities involving fire trucks between 1997 and 2006, 94 fatalities were other road users involved in these crashes. In Europe, six Swedish firefighters died as a result of traffic crashes in the 5-year period 2012–16. They died mostly from single-vehicle

crashes but at least one firefighter was killed when hit by a motor vehicle while outside their vehicles. Although there were more fatalities in the United States, Sweden's fatality rate per population is 1.2 fatalities per year, which is higher than that in the United States. With the Swedish rate, we would expect around 38 firefighter fatalities per year in the United States instead of just 179 in ten years.

In the United States, ambulances are reported to be involved in more than 6500 crashes per year, 35% of which are associated with at least one injury and/or fatality (SFTLA, 2019). Road crashes are identified to be the first leading cause of occupation-related fatalities among emergency medical service (EMS) personnel. In fact, occupational road crash-related fatalities are 4 times higher and crash risks are 20 times higher among EMS personnel relative to other occupations (Maguire and Smith, 2013). In Great Britain, between 1999 and 2004, a total of 38 fatal crashes and 204 serious injury crashes involving ambulances were documented (Lutman et al., 2008). Crashes involving ambulances were documented to result in 64 civilian fatalities and 217 civilian injuries between 1996 and 2012 in the United States (Hsiao et al., 2018). In Sweden, 2003–13 saw 173 police-reported crashes involving ambulances, injuring 218 people in the ambulances, 26 of them seriously injured. A majority of the seriously injured were emergency medical technicians in the back of the ambulance, unable to wear seat belts while performing their duties. Two people died because of these crashes. One was a driver and one was a passenger. A recent in-depth study of injury crashes involving ambulances in a county in Sweden showed that the ambulance driver was the prime contributor (caused) 80% of the crashes. As a result of the study, a 2018/19 bill in the Swedish parliament was passed which gave the government a charge to develop a nationwide mandatory training program of ambulance drivers. Before the legislation, drivers of EVs do not need a specific license but anyone with an automobile license can drive an ambulance or police car and anyone with a truck license can drive a fire truck. However, to drive a taxi in Sweden, a driver needs extensive training beyond a normal driver's license.

In terms of police cars, in the United States, this EV is reported to be involved in 300 fatalities each year (SFTLA, 2019). Between 1998 and 2008, 50% of fatalities among law enforcement officers were reported to be from road crashes. In Australia, police car crash risk is identified to be 11.8 crashes per million kilometers traveled. It is also reported that police pursuits result in relatively higher crash risk representing one crash per 120 kilometers traveled (Symmons et al., 2005). In the United States, from 1979 through 2013, crashes involving police cars resulted in more than 2400 civilian fatalities (Hsiao et al., 2018). It has also been reported that 18% of police fatalities occur when officers are working near speeding motor vehicles, directing traffic or issuing traffic summonses either on the road or on the side of the road (Clarke and Zak, 1999). In Sweden, in the 10-year period 2003–12, at least 9 people were killed and 1457 injured in crashes involving police cars. These numbers do not include injuries and fatalities in crashes caused by police pursuits if the police car was not involved (damaged) in the crash. Comparing fatality rates on a population basis, Sweden's 0.9 fatalities per year would be comparable to 29 fatalities per year in the United States. Thus, the United States has roughly 10 times higher fatalities than Sweden, probably a reflection of police pursuits being very rare in Sweden. Also, it is rare to see a police vehicle in Sweden travel above the posted speed limit whereas that is common in the United States.

Overall, between 2001 and 2010, more than 0.3 million EVs were involved in road crashes in the United States. Among these crashes, 49.3% of the fatalities involving EVs occurred when EVs were operating in emergency mode with lights and sirens (NHTSA, 2001–2010). The national crash fatality rates, in the United States, for EV personnel are 2.5–4.8 times higher than national average for all occupations (Savolainen et al., 2009). EV crashes often also incurred in many lawsuits costing millions of dollars due to the loss of lives, the associated injuries, and property damages. It is estimated that the global cost of EV accidents is \$250 billion annually. In the United States, the cost of EVs is computed to be \$35 billion annually (SFTLA, 2019).

Objective and Scope

Despite the growing concerns and the impacts of EV crashes, there have been few studies in understanding the issues related to this crash type. These studies tend to focus on the determinants of crashes and severities, emergency operations, driver behavior, and distractions. Nevertheless, these studies have important implications in identifying countermeasures and in identifying critical areas in driver training to improve EV safety. Therefore, this paper will review the research on EV crashes, and discuss the critical factors contributing to crash risk, injury risk, and severity outcomes of crashes involving EVs. It will also discuss the different sources of EV crash data and the analytical approaches employed in existing safety literature.

The rest of this paper is organized as follows. First, this paper will discuss details of EV crash analysis while focusing on EV crash data, analytical contexts, and methodological overview. Second, a summary of the contributing factors related to EV crash analysis will be presented. The third section will present a research framework for safety analysis and examination of EV-involved crashes. Finally, insights and concluding remarks will be presented. It should be noted that the impact of different policy inventions, in terms of education, engineering, and enforcement strategies, and impacts of these interventions on EV safety are beyond the scope of this paper.

Analyzing Emergency Vehicle Crashes

Emergency Vehicle Crash Data

In evaluating critical factors contributing to EV crashes (crash risk and severity outcome), police-reported crash databases have thus far been the most prevalent data used in traffic safety research. These crash databases generally compile detailed information on crash characteristics, driver characteristics, environmental attributes, roadway attributes, and vehicle characteristics. However, these crash

databases do not generally include information on driver actions or behaviors and information on emergency operations. Therefore, by using these crash data it is often difficult to understand the impact of driver actions and emergency driving conditions on crash risk and severity outcomes.

Occasionally, researchers have used crash records along with dispatch information from different emergency department data repositories (Bui et al., 2018). These databases have greater scopes in terms of quantifying the relative risk of EV crashes with regard to their actual exposure. EVs equipped with telematics and black box are useful sources of information for monitoring driver actions and vehicle trajectories, including driving speed, acceleration, and braking actions. These secondary data sources (if available) can be used to develop more robust decision support tools to improve EV safety by incorporating driver actions and driving conditions critical to EV safety.

It is worth noting that many safety researchers continue to mainly use police-reported crash databases for EV road safety studies since the cost associated with obtaining these secondary data is much lower than collecting primary data themselves. Nevertheless, road safety analyses can still be enhanced by integrating information from emergency departments and police-reported crash databases. But such data pooling approaches, in identifying critical factors contributing to EV crashes, are far and few between.

There are also several other proactive data sources that may have greater scopes in evaluating critical EV-related crash factors, including surveys, and driving simulation. These data can be used to understand drivers' perception, EV operational influences, and behavior of civilian road users on roadways in the presence of EVs. These alternate data collections and analysis approaches have yet to be explored in EV crash safety studies.

Analytical Contexts

Existing studies have predominantly focused on identifying critical factors related to EV crash risk. Due to the lack of real-world EV exposures data, the evaluation of EV crash risk were mainly focused on the severity outcomes or the comparison of EV versus other vehicle crashes. Also, a few studies have compared crash factors under emergency versus nonemergency operation of the EVs involved in crashes (Pirrallo and Swor, 1994). Given the differences in response categories of different EVs, it can be presumed that the crash mechanisms of these vehicles will be fundamentally different. Therefore, some studies have identified differences among critical factors contributing to crashes involving ambulance, police car, and fire truck (Becker et al., 2003; Symmons et al., 2005). Moreover, given the differences in roadway environment between urban and rural crash settings, a number of studies have identified differences in factors contributing to EV crashes between two traffic environment conditions (Ray and Kupas, 2007). Evaluation of EV driver actions was also focus of few studies (Yasmin et al., 2012).

In terms of disaggregate-level crash severity analysis, few studies have analyzed the contributions of variables in EV crash severity outcomes, with major focus on EV occupant severity. The injury severity profiles of occupant in other vehicles involved in crashes with EVs are likely to be fundamentally different due to the differences in driving characteristics and vehicle compositions. However, none of the studies have identified the differences in injury severity profiles between EV and other vehicle crash victims. It is also surprising that none of the earlier studies have developed macro or planning-level models to identify critical factors contributing to aggregate-level EV crash counts over time. Outcomes of such macro-level EV crash count models might be useful to provide us with insights on EV dispatch origin and destination relative to clusters of EV crash locations (if available). Moreover, these studies will allow us to identify the plausible relations between EV crash risks and built environment.

Methodological Overview

Statistical methodologies are used by researchers in many different fields, including transportation safety, to glean information from existing data sources. Depending on the level of analysis, the studies examining EV crashes can broadly be classified into two categories: (1) aggregate-level descriptive analyses (based on univariate or bivariate associations at the aggregate level)—these studies are presented in the first row panel of Table 1, and (2) disaggregate-level multivariate analyses—these studies are presented in the second row panel of Table 1.

The majority of the previous studies fall in the first category. Using aggregated data, these studies either conduct simple qualitative or descriptive analysis or use nonparametric (distribution free) tools, such as χ^2 statistic or Fisher's exact test statistic to analyze group differences. There are two caveats associated with these types of analyses. First, aggregation of data might lead to loss of information. Second, the methods are explorative and hence, drawing causal inference is not possible. As a result, their policy implications and practical applications are quite limited.

Disaggregate-level analysis can extract richer behavioral insights that lead to more meaningful, accurate, and policy-relevant findings. Despite the clear advantages, application of multivariate methods is fairly rare in EV crash analysis. The studies in this category generally use binary logit and probit models to identify: (1) the factors associated with emergency and nonemergency pursuits or (2) the risk factors associated with fatal and injury outcomes. To differentiate between factors affecting crashes involving different EVs (police, fire truck, ambulance), multinomial logistic regression is also used. On the contrary, ordered logit model and its variants (e.g., generalized ordered logit model) have been used in occupant injury severity analyses as these are recorded in ordinal and discrete categories such as property damage only, minor injury, major injury, and fatal injury.

There has been continual and substantial progress in road safety research methodology in recent years—fueled primarily by the enormous increase in computing power and the need to overcome the limitations of the publicly available crash data. Researchers have developed methodologies that can accommodate (1) systematic and unobserved heterogeneity, (2) endogeneity

Table 1 Summary of studies on emergency vehicle-involved road crashes

<i>Study</i>	<i>Study objective(s)</i>	<i>Emergency vehicle type</i>	<i>Data source</i>	<i>Analysis framework</i>	<i>Categories of independent variables considered</i>
Univariate analysis					
Saunders and Heye (1994)	Characterize ambulance crashes	Ambulance	Ambulance crash data in San Francisco	Retrospective epidemiological analysis	Roadway factors Environmental factors Crash factors
Weiss et al. (2001)	Compare urban and rural ambulance crashes	Ambulance	Ambulance crashes in Tennessee (1993–97)	Two-tailed χ^2 test Fisher's exact test	Contextual factors Roadway factors Environmental factors Driver/occupant factors
Kahn et al. (2001)	Identify the characteristics of the fatal ambulance crashes	Ambulance	Fatal ambulance crashes in United States (1987–97)	Pearson χ^2 test of significance	Contextual factors Roadway factors Environmental factors Crash factors Occupant factors
Langham et al. (2002)	Investigate the ability of the drivers to detect parked police cars	Police vehicle	Crashes with stationary police vehicles	Laboratory experiment	Vehicle factors Driver factors
Proudfoot (2005)	Examine fatality factors for EMS ^a workers	Ambulance	Fatal ambulance crashes in United States (1991–2000)	Descriptive analysis	Roadway factors Environmental factors Crash factors Vehicle factors Driver factors
Ray and Kupas (2005)	Compare ambulance and similar-sized vehicle crashes	Ambulance	Crash data in Pennsylvania (1997–2001)	Two-tailed χ^2 test Fisher's exact test	Roadway factors Crash factors
Ray and Kupas (2007)	Compare urban and rural ambulance crashes	Ambulance	Ambulance crash data in Pennsylvania (1997–2001)	Two-tailed χ^2 test Fisher's exact test	Roadway factors Crash factors Driver factors
Lutman et al. (2008)	To investigate incidence of crash and deaths during ambulance retrieval	Ambulance	Crash data in the United Kingdom (1999–2004)	Descriptive analysis	—
Chiu et al. (2018)	Characterize the ambulance crashes	Ambulance	Ambulance crashes in Taiwan (2011–16)	Descriptive analysis Two-sample <i>t</i> -test Analysis of Variance	Contextual factors Roadway factors Environmental factors
Koski and Sumanen (2019)	Risk factors associated with emergency response driving	Emergency vehicles	—	Inductive content analysis	Environmental factors Driver factors
Multivariate analysis					
Auerbach et al. (1987)	Identify the factors associated with injury risk	Ambulance	Ambulance crashes in Tennessee (1983–86) supplemented by driver interview	Binary logit	Roadway factors Environmental factors Crash factors
Solomon and King (1995)	Examine the effect of vehicle color on fire vehicle crashes	Fire vehicle	Fire department crash data in Dallas (1984–88)	Bayesian method	Contextual factors Roadway factors
Custalow and Gravitz (2004)	Identify factors associated with ambulance crashes	Ambulance	Ambulance crash data in Denver (1989–97)	Binary logit	Roadway factors Environmental factors Crash factors Driver factors
Yasmin et al. (2012)	Identify the factors associated with crash severity	Emergency vehicle	Emergency vehicle crashes in Alberta (1999–2008)	Ordered logit Generalized ordered logit	Contextual factors Roadway factors Environmental factors Vehicle factors Driver factors
Drucker et al. (2013)	Fatality risk of emergency and civilian vehicle crash	Emergency vehicle	Fatality Analysis Reporting System data in United States (2002–10)	Binary logit	Contextual factors Roadway factors Environmental factors Crash factors Vehicle factors Driver factors
Chu (2016)	Crash risk of on-duty and routine patrol and emergency response police vehicle crash	Police vehicle	Traffic crash data in Taiwan (2005–8)	Binary probit	Contextual factors Roadway factors Environmental factors Crash factors Driver factors

(continued)

Table 1 Summary of studies on emergency vehicle-involved road crashes (*cont.*)

Study	Study objective(s)	Emergency vehicle type	Data source	Analysis framework	Categories of independent variables considered
Bui et al. (2018)	Identify driving behavior associated with fire vehicle crashes	Fire vehicle	Telematics data	Penalized logistic regression	Driver factors
Missikpode et al. (2018)	Crash risk associated with emergency vehicles	Police vehicle Ambulance Fire vehicle	Crash data in Iowa (2005–13)	Binary logit	Roadway factors Environmental factors Contextual factors Driver factors

^aDOH, Department of Health; DOT, Department of Transportation; EMS, emergency medical service; FARS, Fatality Analysis Reporting System.

bias, and (3) within crash as well as temporal and spatial correlations. However, these methods are yet to be applied to EV crash analysis although it is very likely that unobserved heterogeneity, endogeneity bias, and temporal variability might be present in the context of EV crashes. For example, driving EVs while responding to an emergency call entails driving in stressful conditions. Different drivers (even though they are highly trained) will adapt and react to these situations differently, which may result in unobserved heterogeneity in the event of a crash and its outcome. This heterogeneity can result in biased or inefficient parameter estimates. To address this concern, more generalized modeling framework such as random parameter or latent segmentation-based models need to be applied. Moreover, no study has systematically attempted to identify exogenous variables that offer time-varying effects and quantify the changes in their impact. Over the years, continual advance has been made in in-vehicle safety features, EV regulations, and driver training and monitoring initiatives. Therefore, there is much scope for methodological improvement in EV crash research.

Crash Contributing Factors

Previous research on EV crashes has predominantly focused on ambulance crashes, with very few studies focusing on police vehicle and fire truck crashes. A summary of the earlier studies is presented in Table 1. In the following subsections, factors associated with either the probability of occurrence or the outcome of EV crashes will be discussed based on the studies presented in Table 1. For ease of understanding, the factors are categorized into five groups: (1) crash characteristics, (2) roadway characteristics, (3) environment characteristics, (4) driver characteristics, and (5) vehicle characteristics.

Crash Characteristics

Location is an important factor in both the probability of a crash occurrence and its injury outcome. For instance, the incidence of ambulance crashes is higher in the urban setting. Urban roads typically carry heavier traffic and thus the probability of traffic conflicts and crashes is higher. However, due to high-posted speed limits and dark roadway conditions, rural settings are more likely to result in injuries, and the severity of these injuries is likely to be greater than in urban settings. Interestingly, opposing results are also found in the literature. The contrasting findings might be the result of different study environment, study methodology, sample size, definition of urban and rural settings, and a combination of these factors.

Ambulance crashes involving more people have a higher probability of more injuries because of non-wearing of seat belts by the accompanying members and the emergency personnel in the back of the ambulance since wearing of seat belts inhibits patient care. The presence of unattached object is another potential reason for higher injury risk of rear compartment occupants. Research has also shown that compared to the law enforcement officers and firefighters, emergency medical personnel are at a greater risk of being fatally injured.

In general, ambulances are more likely to be involved in angular crashes while loss-of-control and rollover crashes are the major collision types in which fire trucks are involved. When vehicles are deployed for emergency use, they are more likely to be involved in angular and multi-vehicle crashes. However, ambulances are also significantly more likely to have an impact at the front (resulting from head-on collisions or striking a fixed object) in the rural area crashes, whereas urban ambulances are more likely to be involved in back-end or rear-end collisions.

Roadway Characteristics

Ambulances are more likely to be involved in crashes at four-way intersections, roads with streetlights, and traffic signals in urban areas. This may be due to the other drivers who are supposed to stop and give way to the ambulance may not be stopping when the signal is green. In fact, “failed to notice emergency vehicles,” “failed to yield to emergency traffic” at an intersection, and “failing to

pull over to let a police car pass” in an emergency situation were cited to be the major faults among the civilian drivers. Crashes at intersections or driveways are also more likely to be overrepresented during emergency response.

Environmental Characteristics

Environmental factors are relatively more common in causing rural ambulance crashes. For example, snowy roadway conditions and poorly lit nighttime roads are more prevalent in rural settings whereas the urban crashes are more likely to occur in rain and on wet roads. Also, ambulance crashes are more likely to be injurious on wet pavement, under darkness, during adverse weather, and at intersection. Ambulance crashes occur with increased frequency in the evenings and during weekends. Crashes during emergency use are more likely to occur during the afternoon peak period.

Driver Characteristics

EV drivers are more likely to be passive victims of other road users’ mistakes and violations than being perpetrators of such mistakes and violations. In fact, intoxication of civilian drivers is a major contributor of injury risk in an EV crash. However, distracted driving and poor driving history (record of involvement in multiple collisions on duty) in part of the EV driver is also an important contributory factor. For example, violation of traffic signals by police officers while responding to an emergency or driving in pursuit of another vehicle significantly increases the probability of crashes causing injuries. Also, the age of police officers plays a significant role in the probability of injuries, with young officers (<30 years) being more likely to be involved in crashes causing injury than older drivers while on an emergency run. A 2009 Swedish study of 1763 police officers injured in road crashes shows that out of the 76 who were seriously injured or killed, 27 were below age 30, 28 were 30–39, 17 were 40–49, 3 were 50–59, and 1 was above age 60. Also, 59 were men and 17 were women. Without knowing exposure by age or gender, this does not express risk, but it is clear that younger men are involved in a majority of the serious crashes. However, as opposed to the data from the United States, in Sweden, the police officer was deemed to be at fault for a majority (57%) of the crashes when two vehicles collided.

Vehicle Characteristics

The color of the car and usage of lights and sirens are important vehicle-related factors associated with EV crash occurrences. For instance, the probability of a visibility-related multiple-vehicle crash or a visibility-related single-vehicle crash for a red fire pumper is 3 times greater than for a lime-yellow fire pumper. The superior visibility of lime-yellow color yields an earlier awareness of the fire pumper’s presence by the civilian driver, thereby resulting in a lower crash rate and lower likelihood of injury or tow-away damage.

The use of lights and sirens during an emergency response has a significant increased risk of being involved in a crash for police vehicles but this is likely to be due to the more risky way in which the vehicle is driven in an emergency. However, driving in emergency mode does not increase the crash risk among ambulance and fire truck drivers (Missikpode et al., 2018).

Research Framework

Traffic crash analysis involving EVs should advance along with overall road safety research. However, there are significant scopes for improvement in EV-related crash analysis in terms of employing advanced methodological approaches, exploring more policy-relevant analyses, and adopting other available crash data sources. This paper therefore presents a simple research framework for safety analysis of EV-involved crashes. The research framework is presented in Fig. 1.

In the first stage of the research framework, relevant theoretical models should be identified to develop a conceptual framework for analyzing for severity and/or likelihood of EV-involved crashes. The choice and foci of the theoretical models should be guided by the aims and objectives of the research. Crash risk theory is employed to identify attributes that result in traffic crashes and propose means to reduce the likelihood of traffic crashes. It is focused on identifying the effective countermeasure to improve the roadway design and operational attributes, vehicle designs, and road user behaviors. While reducing traffic crash occurrence is essential, it is also important to provide solutions to reduce the consequences in the unfortunate event of a traffic crash. Thus, the crash severity theory is employed to examine crash events and identifies factors that impact the crash outcome and suggests countermeasures to reduce crash-related consequences (injuries and fatalities). These associated risk factors are critical to assist decision-makers, transportation officials, insurance companies, and vehicle manufacturers to make informed decisions to improve road safety. The importance of using well-established theoretical models needs to be emphasized.

Next, a set of primary and secondary EV crash data sources should be identified which will provide the appropriate data needed to answer the research questions. In the third stage of research framework, analytical contexts focusing on severity and risk analysis should be identified. The focus is on identifying critical factors contributing to EV crash risk and severity outcomes for which data are available. It should also select the most appropriate research method to analyze the data and answers the research questions. More importantly, it should result in theory and evidence-based recommendations to improve safety. Finally, in stage four of the framework, the design, implementation and evaluation of road safety countermeasures should be conducted.

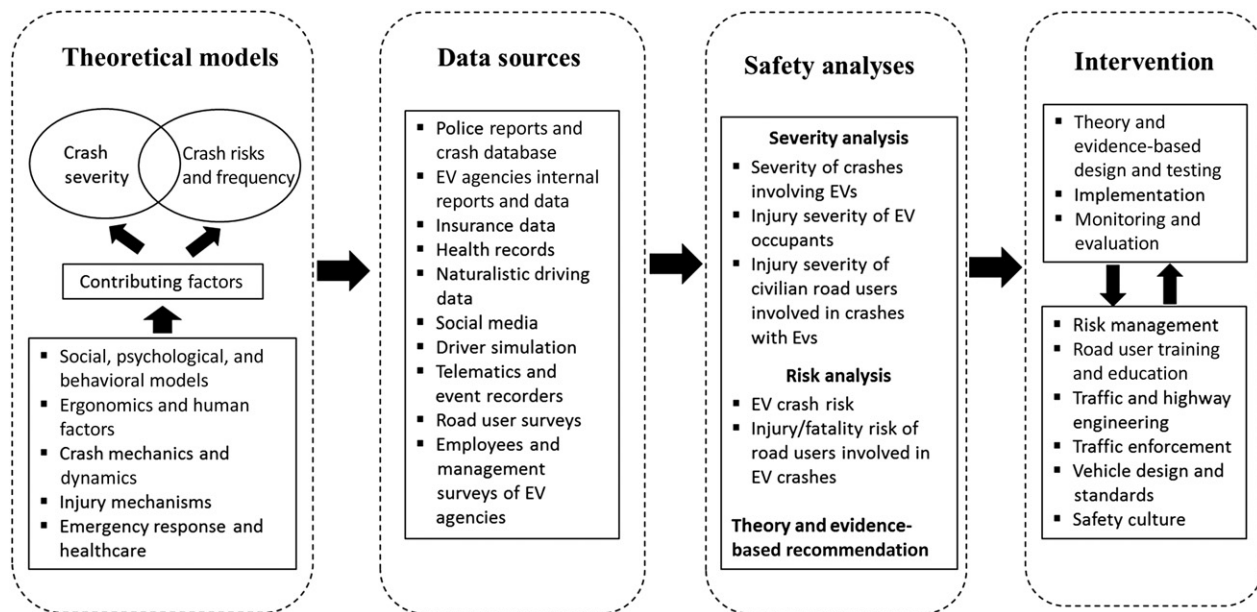


Figure 1 Research and evaluation framework for safety analysis of emergency vehicles-involved road crashes.

The presented research vision is an abstraction or simplification of analytical approaches that can be employed in identifying the critical factors and the associated countermeasures related to EV crash risk and severity outcomes. Such research vision will help up to better understand the research scope and stimulate further thoughts in facilitating future research in EV crash analysis.

Some Insights and Concluding Remarks

This paper discussed the critical factors, based on existing safety literature, contributing to crash risk, injury risk, and severity outcomes of crashes involving EVs. It also discussed the different sources of EV crash data and the analytical approaches employed in existing safety literature. Moreover, this paper presented a research framework for safety analysis of EV-involved crashes.

We found that in identifying the critical factors contributing to EV crashes (crash risk and severity outcome), police-reported crash databases have so far been the most widely used. The majority of previous studies employed univariate analysis, while only few studies adopted multivariate analytical frameworks for examining critical factors contributing to EV crashes. Most of the previous research on EV crashes has focused on ambulance crashes, with very few studies focusing on police vehicle and fire truck crashes.

In general, the factors that were attributed to higher EV crash risks include urban setting, four-way intersection, traffic signal, evening time, and weekends. The critical factors that were reported to be associated with higher injury or fatality risk outcome of EV-involved crashes include high speed limit, rural setting, dark light, and wet pavement condition.

We also found that there is significant scope for expansion in EV-related crash analysis in terms of employing advanced methodological approaches, exploring more policy-relevant analytical contexts, and adopting other available crash data sources. There has been continual and substantial progress in transportation safety research in recent years—fueled by the enormous increase in computing power and the need to overcome the limitations of conventional crash data. Research on EV crashes can utilize these recent advancements in road safety research.

While earlier research has identified the impact of several attributes contributing to EV crashes, there has been limited research in developing a comprehensive analytical framework to develop an understanding on the extent of the EV crash-related safety concerns. We proposed such a research framework in this paper.

References

- American Medical Association, 1903. Principles of Medical Ethics of the American Medical Association. American Medical Association Press, Chicago, IL.
- Auerbach, P.S., Morris, J.A., Phillips, J.B., Redlinger, S.R., Vaughn, W.K., 1987. An analysis of ambulance accidents in Tennessee. *J. Am. Med. Assoc.* 258 (11), 1487–1490.
- Becker, L.R., Zaloshnja, E., Levick, N., Li, G., Miller, T.R., 2003. Relative risk of injury and death in ambulances and other emergency vehicles. *Acc. Anal. Prev.* 35 (6), 941–948.
- Bui, D.P., Hu, C., Jung, A.M., Pollack Porter, K.M., Griffin, S.C., French, D.D., Crothers, S., Burgess, J.L., 2018. Driving behaviors associated with emergency service vehicle crashes in the US fire service. *Traffic Inj. Prev.* 19 (8), 849–855.
- Chiu, P.W., Lin, C.H., Wu, C.L., Fang, P.H., Lu, C.H., Hsu, H.C., Chi, C.H., 2018. Ambulance traffic accidents in Taiwan. *J. Formos. Med. Assoc.* 117 (4), 283–291.
- Chu, H.C., 2016. Risk factors for the severity of injury incurred in crashes involving on-duty police cars. *Traffic Inj. Prev.* 17 (5), 495–501.
- Clarke, C., Zak, M.J., 1999. Fatalities to law enforcement officers and firefighters, 1992–97. *Environments* 15 (2), 3–8.

- Custalow, C.B., Gravitz, C.S., 2004. Emergency medical vehicle collisions and potential for preventive intervention. *Prehosp. Emerg. Care* 8 (2), 175–184.
- Donoughe, K., Whitestone, J., Gabler, H.C., 2012. Analysis of firetruck crashes and associated firefighter injuries in the United States. *Ann. Adv. Automot. Med.* 56, 69–76.
- Drucker, C., Gerberich, S.G., Manser, M.P., Alexander, B.H., Church, T.R., Ryan, A.D., Becic, E., 2013. Factors associated with civilian drivers involved in crashes with emergency vehicles. *Acc. Anal. Prev.* 55, 116–123.
- Hsiao, H., Chang, J., Simeonov, P., 2018. Preventing emergency vehicle crashes: status and challenges of human factors issues. *Hum. Factors* 60 (7), 1048–1072.
- Kahn, C.A., Pirrallo, R.G., Kuhn, E.M., 2001. Characteristics of fatal ambulance crashes in the United States: an 11-year retrospective analysis. *Prehosp. Emerg. Care* 5 (3), 261–269.
- Koski, A., Sumanen, H., 2019. The risk factors Finnish paramedics recognize when performing emergency response driving. *Acc. Anal. Prev.* 125, 40–48.
- Langham, M., Hole, G., Edwards, J., O'Neil, C., 2002. An analysis of 'looked but failed to see' accidents involving parked police vehicles. *Ergonomics* 45 (3), 167–185.
- Lutman, D., Montgomery, M., Ramnarayan, P., Petros, A., 2008. Ambulance and aeromedical accident rates during emergency retrieval in Great Britain. *Emerg. Med. J.* 25 (5), 301–302.
- Maguire, B.J., Smith, S., 2013. Injuries and fatalities among emergency medical technicians and paramedics in the United States. *Prehosp. Disaster Med.* 28 (4), 376–382.
- Missikpode, C., Peek-Asa, C., Young, T., Hamann, C., 2018. Does crash risk increase when emergency vehicles are driving with lights and sirens? *Acc. Anal. Prev.* 113, 257–262.
- NHTSA, 2001–2010. Estimate of crashes involving emergency vehicles by year, emergency vehicle and emergency use. General estimates system (GES). Fatality Analysis Reporting System (FARS). National Highway Traffic Safety Administration, Washington, DC.
- NHTSA, 2011. Characteristics of law enforcement officers' fatalities in motor vehicle crashes (No. HS-811 411). U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, DC.
- Pirrallo, R.G., Swor, R.A., 1994. Characteristics of fatal ambulance crashes during emergency and non-emergency operation. *Prehosp. Disaster Med.* 9 (2), 125–132.
- Proudfoot, S.L., 2005. Ambulance crashes: fatality factors for EMS workers. *Emerg. Med. Serv.* 34 (6), 71–73.
- Ray, A.F., Kupas, D.F., 2005. Comparison of crashes involving ambulances with those of similar-sized vehicles. *Prehosp. Emerg. Care* 9 (4), 412–415.
- Ray, A.M., Kupas, D.F., 2007. Comparison of rural and urban ambulance crashes in Pennsylvania. *Prehosp. Emerg. Care* 11 (4), 416–420.
- Saunders, C.E., Heye, C.J., 1994. Ambulance collisions in an urban environment. *Prehosp. Disaster Med.* 9 (2), 118–124.
- Savolainen, P.T., Dey, K.C., Ghosh, I., Karra, T.L., Lamb, A., 2009. Investigation of emergency vehicle crashes in the state of Michigan (No. NEXTRANS Project No. 015WY01). NEXTRANS Center, USA.
- SFTLA, 2019. Statistics on Emergency Vehicle Accidents in the United States. San Francisco Trial Lawyer Association (SFTLA). Available from: <https://www.fettolawgroup.com/Personal-Injury-Blog/>.
- Solomon, S., King, J., 1995. Influence of color on fire vehicle accidents. *J. Safety Res.* 26 (1), 41–48.
- Symmons, M.A., Haworth, N.L., Mulvihill, C.M., 2005. Characteristics of police and other emergency vehicle crashes in New South Wales. Road Safety Research, Policing and Education Conference. IRB, Wellington.
- U.S. Fire Administration, 2014. Firefighter fatalities in the United States in 2013. U.S. Fire Administration (USFA), National Fire Programs (NFP) and National Fallen Firefighters Foundation (NFFF).
- U.S. Fire Administration, 2014. Emergency vehicle safety initiative. FA-336. U. S. Fire Administration (USFA).
- Weiss, S.J., Ellis, R., Ernst, A.A., Land, R.F., Garza, A., 2001. A comparison of rural and urban ambulance crashes. *Am. J. Emerg. Med.* 19 (1), 52–56.
- Yasmin, S., Anowar, S., Tay, R., 2012. Effects of drivers' actions on severity of emergency vehicle collisions. *Transp. Res. Rec.* 2318, 90–97.

Further Reading

Fahy, R.F., 2008. U.S Firefighter Fatalities in Road Vehicle Crashes—1998-2007. National Fire Protection Association, Quincy, MA.

Encouragement: Awards and Incentives

Fred Wegman, Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	255
Causes of Crashes	256
Penalizing, Nudging, Rewarding, and Incentivizing	256
Penalizing	257
Nudging	257
Rewarding and Incentivizing	258
Examples of Rewarding and Incentivizing in Road Safety	258
Safety Management in Companies	258
Insurance Rewards	259
Key Lessons Learned	260
Conclusions	260
References	261

Introduction

Julian Harvey is believed to have been the first to introduce the three E's to road safety: Education, Engineering, and Enforcement. We are talking about the year 1923. Harvey indicated that there are three different fields in which work can be done to improve road safety. To this day, the three E's are referred to. Using the familiarity of the three E's, authors have extended the three E's with more E's: for example, to five E's with the additions Encouragement and Evaluation, or even to seven E's with the additions "Economics, Emergency response, Enablement, and the overarching 'E' of Ergonomics" (Plant et al., 2018). In all these approaches, interventions are covered by one of the E's: vehicle safety, for example, is covered by Engineering, traditional police surveillance by "Enforcement" and a government campaign that aims to discourage the use of mobile devices while driving is covered by Education. Sometimes two E's (Enforcement and Education) are used simultaneously when designing interventions: for example, to supplement traffic enforcement with good communication to make an intervention more effective in terms of behavioral change.

It was not until many years later that attempts were made to place the three E's in a more conceptual framework. In the early 1970s, for example, Haddon (1972) introduced his matrix, in which he used two axes. The one axis reflects the three phases in the crash process (precrash, crash, and postcrash) and the other axis reflects the three components of road traffic: road user, road, and vehicle; hence this matrix consists of nine cells. The Haddon matrix was used to classify crash factors and to indicate crash prevention measures. The use of the Haddon matrix made it clear that other cells are important and deserve our attention in addition to the "pre-crash—road user" cell, the cell that traditionally received a lot of attention in policy and research.

In later years, more attempts were made to facilitate the analysis of crash factors and the ordering of preventive measures. Rumar (1999) described the road safety problem as a product of three dimensions: exposure to risk, crash risk, and injury risk. Adding the "exposure to risk" dimension as a way to promote safety was certainly new. A later attempt distinguished five pillars (UNRSR, unknown). These five pillars combine the components man, road and vehicle with a phase of the crash process (postcrash) and with road safety management. From a conceptual viewpoint, this is not an ideal structure, but apparently it is appealing as it is now being used in many countries (WHO, 2018).

The different phases of the crash process, the three components of road traffic, and the three E's have a (very) long tradition in road safety and they have been part of the different paradigms that were used over the years (e.g., Wegman, 2013).

The concept of "encouragement" can be applied to improve road safety, but has in fact been absent from road safety strategies. Whereas enforcement (punishment, discouragement) has played a role for almost a hundred years and is included in nearly all road safety strategies and plans, encouragement has remained virtually invisible. This is remarkable because research shows that both rewarding and punishing can effectively lead to behavioral changes and that in the dichotomy between punishing and rewarding, traditionally rewarding seems to be preferred for achieving results. What parent in the Western world is not confronted with the "fact" that it is better to change a child's behavior by rewarding good behavior and punishing a child as little as possible?

In the 1930s, the American psychologist Fred Skinner launched his theory of behaviorism which made an important contribution to the idea of rewarding desired behavior (the behavior will then be shown more often), to ignore unwanted behavior (is less frequently exhibited and extinguished) and to punish only if unwanted behavior is dangerous and should be stopped immediately. The disadvantage of punishment would be that it does not indicate what desired behavior looks like. As early as in the 1970s, the Canadian psychologist Albert Bandura introduced the phenomenon of "observational learning" or "social learning": a good example will be keenly followed. The fact that these two gurus of behavioral influence and learning both prefer reward rather than

punishment raises the question why rewarding has not been used much more to influence traffic behavior to ensure safer road traffic?

A number of ways to influence behavior will be discussed in this chapter. But it is good to give a perspective on the causes of road crashes, because such a picture is required to answer the question of how effective and efficient it is to use rewarding and penalizing to reduce traffic risks and, consequently, the number of crash casualties. After comparing different types of behavioral influencing, some examples will be given of how rewarding can be used to improve road user behavior. Specifically, this concerns safety management in companies and insurance rewards. The chapter will be concluded with a list of most important lessons learned regarding the application of rewarding.

Causes of Crashes

Crash causes are frequently discussed in road safety literature and nearly always the conclusion is that the human factor plays a (dominant) role in nearly all crashes. This is demonstrated by crash forms drawn up by the police, by crash causation studies and by naturalistic driving studies. Police reports are not a good source to assess crash causation (Hauer, 2020; Shinar, 2007), because the police form is not intended to detect crash causation, but to determine whether a person involved in a crash committed an offence and to contribute to the determination of civil liability.

In the 1970s, two “in-depth” studies were carried out into the causes of crashes, one in the United Kingdom and one in the United States. The results of these two studies (Sabey and Staunton, 1975; Treat et al., 1977) bear many similarities: in approximately 95% of the crashes the human factor was found to have contributed to the occurrence of a crash or to have had an impact on its severity. The road factor had contributed in about 30% of the crashes and vehicle factors were to have played a role in approximately 10% of the crashes. These two renowned studies, carried out by highly respected researchers have very often been cited and seem to “tell the truth.” Now, 50 years later, the results are challenged.

In his article “Crash causation and prevention,” the Canadian emeritus professor Ezra Hauer states that the above results are a “quasi-finding” (Hauer, 2020). In his very readable article, he argues that the results are the consequence of definitions of crash causes which, in his opinion, are unsuitable for crash prevention: “That quasi-finding provided and continues to provide respectability to a style of road safety management that makes the road-user the primary target of prevention actions and obscures the role of legislation, regulation, planning, road design, traffic management, vehicle design, and features in crash generation and crash severity. By so doing the quasi-finding is an obstacle to the transition to the Safe Systems style of road safety management. To support the transition to the Safe Systems style of road safety management a different definition of crash cause is needed. Hauer (and others before him) suggested the following definition: a crash cause is a circumstance or action that, were it different, the frequency of crashes and/or their severity would be different.”

The authors of the tri-level study (Treat et al., 1977) reported that “many human causes may imply the need for changes in vehicles or in the environment.” And, in line with this, Hauer states: “Saying that the human is the sole contributing factor in most crashes misleads the general public and provides succor to those who prefer the old-fashioned road user-oriented crash prevention.” Hauer emphatically links crash cause to crash prevention and this supports current thinking on road safety, the Safe System approach. A basis of this approach is that a crash is rarely caused by one single unsafe action; it is usually preceded by a whole chain of poorly attuned occurrences (Wegman, 2013). This means, as Hauer also argues, that a crash is not just caused by one or a series of unsafe road user actions; hiatuses in the traffic system also contribute to the fact that unsafe road user actions can in certain situations result in a crash. These hiatuses are also called latent errors (Reason, 1990). “Road crashes occur when latent errors in the traffic system and unsafe actions during traffic participation coincide in a sequence of time and place” (Wegman, 2013). The Safe System approach tries to eliminate “latent errors” and to reduce exposure to risk. And secondly, “it aims to make traffic safer by adjusting the environment to the human measure in such a way that people commit fewer errors, and if an error is made, the environment is designed to be forgiving for errors.”

Systematically modifying a road user’s environment by creating conditions that make safe behavior attractive and easy, and unsafe behavior unattractive and difficult is at the heart of the Safe System approach. The key question is how to achieve this in the real world and, more specifically, how to adapt the existing road and traffic environment.

Ezra Hauer’s observations that were introduced above, were in response to a dramatic crash in Saskatchewan, Canada. A bus carrying the Humboldt Broncos hockey team collided with a truck hauling two trailers at an intersection of two rural roads with speed limits of 100 km/h. Both vehicles were driving slightly slower than the specified limit. One of the roads (the one used by the truck driver) had a stop sign. Sixteen boys were killed in the crash. As long as such situations exist, crashes can occur and, unfortunately, will continue to occur. Rewarding or punishing is not the way to tackle road safety problems on intersections such as the Canadian one from a Safe System perspective. In a Safe System, it is made impossible for vehicles to ever collide at such high speeds.

Penalizing, Nudging, Rewarding, and Incentivizing

A dominant activity in road safety improvement is influencing citizen behavior and road user behavior. Traffic education and traffic enforcement (two of the three E’s) are the traditional ways to make behavior safer and to reduce the number of traffic offences. Traffic

offences are associated with a higher crash risk, traffic fines do temporarily reduce risks (Redelmeijer et al., 2003) and traffic enforcement leads to fewer traffic offences and fewer crashes (Elvik et al., 2009).

Influencing behavior does not only involve influencing actual traffic behavior, but also aims at influencing citizens' choices that have—positive or negative—road safety effects: buying a safe car or reducing exposure to risk by choosing a safe route, for example traveling a low risk route rather than the shortest route.

We have several possibilities at our disposal to make traffic behavior safer; they will briefly be introduced here. But first it is important to note that a Safe System approach attempts to prevent road users from running unnecessary risks by eliminating exposure to risk (Wegman and Aarts, 2006). Three examples will illustrate this: (1) adequate median barriers on motorways prevent head-on collisions, (2) driving under the influence of alcohol is prevented by installing an alcolock in a vehicle, and (3) well-located and well-designed speed humps reduce driving speeds and risks.

Enforcing safe conduct through legislation and monitoring is a first option, and will only be effective in reducing risks if legislation is supported by evidence and enforcement is adequate. For example, revoking a driving license is only effective if the driver concerned refrains from driving a motor vehicle. However, if the person concerned estimates that the risk of being fined for driving without a license is minimal and therefore drives anyway, such a measure is not effective. Punishing road users for dangerous behavior and for breaking the law is the first possibility of behavioral influence and will here be called *penalizing*: preventing dangerous behavior by (threatening with) punishments.

A second possibility is *nudging*: this involves moving road users towards safe traffic behavior by supporting them. Traffic-safe choices are offered without making other options unattractive or removing them. Each individual retains his freedom of choice, but subtle influence is used to stimulate a road user to choose safe options.

The third possibility is *rewarding* and *incentivizing*. Rewarding involves rewarding safe behavior. A performance has been achieved (e.g., a million kilometers travelled without being involved in a crash) and the road user is rewarded for this. In the case of incentivizing, a road user is promised a reward and is told which behavior or achievement is rewarded. An incentive should act as a motivation to only engage in safe behavior (e.g., not speeding) or to not be involved in a crash.

Let's discuss the three approaches in some more detail.

Penalizing

There is a long tradition in road safety underlying the idea that traffic behavior is influenced and made safer by using forms of police surveillance. This is a chain approach: legislation—enforcement and punishment. The underlying principles of police surveillance are rather well known: the overall preventative effects of police enforcement are generally greater if the subjective risk of the offender being caught is higher, if the penalty is more severe, if the certainty of a penalty is increased and cannot be escaped, and if the penalty is imposed more rapidly. Each of these elements constitutes a link in the enforcement chain. We have a very long tradition in road safety to work with enforcement on all sorts of risky behavior: speeding, drink-driving, red-light violations, not wearing seat belts, etc. We also learned that the subjective probability of being caught is more important than the severity of the penalty. Penalizing can be done by handing out traffic fines, by withdrawal of the driving license or by giving points in a demerit points system (Van Schagen and Machata, 2012).

Police surveillance operates on the basis of the objective and subjective probability of being caught, and the resulting penalties or measures. Roadside police checks determine the objective probability of being caught, the enforcement pressure. If road users consider the subjective probability of being caught to be sufficient, they will avoid infringements.

Traffic control or police surveillance has proved effective in reducing road risks (Elvik et al., 2009). The effectiveness estimates vary considerably and also depend on the nature of a violation and the extent and nature of the enforcement. Modern technology can assist police in carrying out their enforcement duties, for example, point to point enforcement of speeding or the use of in-vehicle technology (e.g., alcolock or Intelligent Speed Adaptation to influence speed behavior).

Nudging

In their influential book "Nudging" (2008), the authors Richard Thaler and Cass Sunstein present a practical example of influencing behavior by "nudging." They describe the course of Lake Shore Drive along Lake Michigan in Chicago. This road has a number of sharp bends and the city of Chicago tries to increase the attention of road users and to reduce their driving speed (down to 25 mph) with the help of a number of transverse rumble strips that are closer together as a bend approaches. This gives the driver the feeling that the driving speed is going up. This sensation makes road users reduce their driving speed. Thaler and Sunstein (2008) believe that other forms of behavioral influence (legislation and enforcement, public information, financial incentives) are too coercive, too paternalistic, or too expensive. They see nudging as an addition to these traditional forms of behavioral influencing. Of course, the question is whether it is a task for government or road authority to be occupied with influencing traffic behavior, but in many countries worldwide this question has been answered in the affirmative. This is reflected, for example, in the World Health Organization's Global Status Report, which shows that virtually all countries in the world have embarked on legislation and enforcement of safer behavior (e.g., driving under the influence of alcohol, or wearing seat belts). There are, of course, exceptions where the government is accused of paternalism if it were to adopt "restrictions" (such as a number of American States that have not laid down in law a helmet-carrying requirement for motorcycle riders).

The concept of “self-explaining roads” (Theeuwes and Godthelp, 1995) refers to “a traffic environment which elicits safe behavior simply by its design.” Speed behavior plays an essential role in this concept. In addition to eliminating risk to exposure, nudging of behavior (by using road characteristics recognizable for the road user to make road course and behavior of other road users predictable, thereby preventing mistakes and reducing risks) is a cornerstone of the Dutch Safe System approach (Wegman and Aarts, 2006).

Rewarding and Incentivizing

An incentive is a benefit in the form of a good or service for a specific performance intended to increase motivation. Incentives are offered by companies to employees to encourage them to show the desired behavior. Whether and how an incentive works is a careful balancing act. Financial incentives may be considered, but also gifts or outings. In essence, an incentive is a reward that is earned when a certain goal or certain behavior is achieved. Incentives can also be granted to road users to encourage them to achieve the desired level of performance or set goal.

A reward is a benefit that is provided in recognition of achievement, service, commendable behavior, etc. A reward is given to an employee only after he/she has provided evidence of his/her positive behavior and achievements. The aim of a reward is to show the employees that their work and effort are valued, and is given as an appreciation for the work already completed, as well as a motivation to continue improving their quality of work.

Despite their similarities in motivating and encouraging workers to achieve better levels of performance, there are a number of differences between rewards and incentives. The main difference lies in the timelines in which each is offered. A reward is offered after the job has been completed and after the employee has proven his worth. An incentive is offered beforehand and is aimed at improving performance of employees who do not meet expected standards or established goals.

Giving rewards or working with incentives to achieve safer traffic behavior is quite unusual in road safety, however several (small-scale) trials have been carried out in two settings: in companies and when setting insurance premiums.

Examples of Rewarding and Incentivizing in Road Safety

It is quite inconceivable that a total population of road users should be given an opportunity to participate in programs that reward safe behavior in any fashion. Nor are there any known examples. Experience has been gained in rewarding certain groups of road users. We will limit ourselves to two specific groups: road users who participate in traffic professionally and who are in an employer–employee relationship (this will be discussed in further detail in the section *Safety Management in Companies*).

Secondly, there is the relationship between a contractor and a subcontractor. Governments award concessions to public transport companies and a concession can be made to comprise road safety-important elements, including timetable requirements that prevent drivers from being placed under time pressure and thereby prevent unsafe behavior such as speeding.

Another possibility are rewards or incentives that insurance companies can give to their insured. Examples are rewards for having driven damage-free in a certain period and incentives offering premium reduction in the case of no/fewer claims for damages or no traffic violations.

Technological advances allow for monitoring and feedback of driver behavior in an inexpensive manner. In-vehicle telematics or smartphone apps enable drivers, insurers, or employers (commercial vehicles) to collect safety-specific information on a driver's on-road behavior and driving performance and to use this information as feedback to promote safer driving. This certainly made reward and incentive programs more feasible.

Safety Management in Companies

So far, road safety improvement has been strongly based on the idea that mainly the government attempts to contribute to safer traffic behavior and consequently reduce the risks of road users. There are reasons, however, to give the private sector a role (Academic Expert Group, 2019).

Work-related motor vehicle crashes are estimated to contribute at least one quarter to over one third of all work-related deaths (ERSO, 2006). And the relationships employer–employee, and contractor–subcontractor offer influencing-possibilities for safer road traffic, that do not exist in the relationship between the government and the individual road user. An employer can influence behavior in the context of the employment relationship with the employee, by punishing and rewarding. Work-related travel behavior has different characteristics than private travel behavior with private vehicles and it is claimed that the risks of work-related travel are higher than those of private travel. Some evidence has been found, but the variety of work-related trips is immense and there has been no evidence that differences in risk are due to increased risk-taking by, for example, more speeding.

An Australian study (Newnam et al., 2002) even found that work-related drivers were generally more likely to report safer driving in a work vehicle. In particular, it was found that drivers were less likely to commit violations like speeding and driving dangerously in a work vehicle than in their personal vehicle. Additional research should provide clarity on the precise explanations for these findings before drawing conclusions.

In many industrialized countries, a growing number of companies embark upon work-related road safety activities. However, very few programs have been studied (and reported in scientific journals) to establish the effectiveness of the variety of the approaches and measures adopted. This lack of publications complicates drawing firm conclusions.

Based on a literature review, [Haworth et al., \(2000\)](#) identified four areas that have potential to be effective for reducing crashes and injuries (and related costs) by private sector companies: (1) selecting safer vehicles, (2) some specific driver training and education programs, (3) incentives (not rewards), and (4) company safety programs in companies with an overall emphasis on safety. To these four fields of potential activities two more can be added. Making use of technology is another possibility that may be exploited, for example in-vehicle data recorders. Finally, the more recent concept of journey management can be added, a concept developed in the oil- and gas industry with the aim to prevent undesired security HSSE (Health, Safety, Security, Environment) consequences of land transport journeys ([NETS, 2014](#)).

A report "Advancing Road Safety. Best Practices for Companies and Their Fleets" of Together for Safer Roads ([TSR, 2016](#)), a coalition that brings together global private sector companies to collaborate on improving road safety, presents some "best practices" of incentive programs, with names such as Safe Driving Awards, Safe Driving Recognition Programs, the Million Mile Club (of collision-free drivers).

[Haworth et al. \(2000\)](#) were eager to include the results of evaluation studies in their report. But they were disappointed: "There are very few evaluations undertaken, even by best practice companies. Benchmarking is one of the few examples of evaluation, but benchmarking only hints at why some organizations may have lower crash rates or costs than others." Narelle Haworth and her colleagues do come up with a number of conclusions about the safety effects of incentive and disincentive programs reported in the literature: some programs had negative effects, incentive programs (where benefits are conditional upon future safe driving) are more effective than reward programs (where benefits are conditional upon safe driving past. Incentive programs appear to be most effective when the time period in which the desired outcome is expected is in the near future. Incentive programs may be more effective in younger drivers, and drivers with a good driving history who are given a reward either show no difference or increase their crash rate.

NETS, an employer-led advocate for global road safety and a chartered nonprofit non-governmental organization, states that successful road safety programs are resourced, led by leadership and are well-embedded in the line-organization ([NETS, 2014](#)). ERSO suggests that effective road safety programs should be based on an integrated set of data-guided measures based on a strong safety culture within the organization (ERSO, 2007). In other words, both documents argue that programs designed to make road user behavior safer through a system of incentives or disincentives, should be part of a much broader package of measures. But NETS does suggest using incentive schemes to illustrate the importance of these programs: "To reinforce the importance of corporate road safety initiatives, organizations should have incentive schemes in place to recognize good driving behavior and penalize poor performance."

If we look at the results, we can conclude that there is only limited scientific evidence that programs using incentives and disincentives can make a substantial contribution to achieving safety targets in companies. But, incentive programs are rather common in the private sector and positive outcomes are reported by stakeholders.

Insurance Rewards

Many car insurance companies offer a discount on the insurance premium for claim-free driving, also referred to as bonus-malus systems. Insurance companies also differentiate the premium by age, which means that young people (and often novice drivers) pay high premiums. These are not individually set premiums, but premiums set for the group in which someone is classified (age, large city versus rural area). A relationship with a driver's performance due to his own behavior has been abandoned and premium differentiation can therefore not really be considered a reward: own behavior has virtually no influence on the result for the group.

It is, however, reasonable to assume that strengthening the extrinsic motivation of road users by giving rewards or incentives leads to behavioral changes in the desired direction. In recent decades, insurers have focused on premium discounts as a way of influencing safe behavior with fewer crashes (and therefore fewer claims) and a lower damage burden as a result. [Hagenzieker \(1999\)](#) looked at the effects of behavior-oriented rewards and she concluded that few conclusive research results were available ("Some studies investigating the effects of on speed behavior show no conclusive results"). However, there have been developments since then.

A behavior-oriented award approach requires information on driving behavior on an individual level and fortunately modern technology offers a helping hand. First of all, in-vehicle data recorders that are connected to the diagnostic system of a vehicle can also be used to gather information about "dangerous" behavior (speeding behavior, acceleration, cornering, braking, smartphone use, close following, signal use, and brake position (see, e.g., [Horrey et al., 2012](#)). More recently, smartphones are being used to collect information, often linked to special apps (e.g., [Musicant and Lotan, 2016](#)). The first system collects information per vehicle, the second system does so per driver.

In principle, three actions can be carried out with the possibilities mentioned above: the behavior of a vehicle/driver can be monitored, feedback can be given based on the monitoring and, finally, the information collected can be used as a basis for rewarding. Actually, the idea of rewarding safe behavior is not new. As early as 1980, the New Zealand researcher Paul Hurst published a study in which this matter was addressed ([Hurst, 1980](#)).

Research is needed to determine whether monitoring (Hawthorne-effect), feedback or rewarding contribute to the road safety improvement. Experiences during the past 10–20 years have shown that this research is complicated to carry out, that coming to clear conclusions is not so easy, and that generalizing the findings (external validity) is not generally permissible.

As evidenced by a meta-analysis conducted by [Elvik \(2014\)](#), the bottom-line would be, that “most studies involving behavior-dependent (‘direct’) incentives found significant improvements in safety-relevant driving skills.” Measured changes (e.g., in speeding) can be substantial. However, the trials reported in this study included highly motivated drivers and were relatively small. This leads to the question what to expect when scaling up trials. And, is it plausible to expect a (substantial) decrease in the number of crashes and casualties in a country, when applying large scale reward programs?

Carrying out research on reward programs was not without hurdles. First of all, studies should be able to recruit sufficient subjects and the experiences in Denmark, for example, are bizarre. A study could not be carried out successfully because insufficient participants could be found ([Lahrmann et al., 2012](#)). Another problem that arose in an Austrian study, was that participants withdrew from the study ([Peer et al., 2020](#)), which can lead to an attrition bias. A bias due to self-selection could also be a problem, but according to [Peer et al. \(2020\)](#) this was not so much an issue in their research, although [Elvik \(2014\)](#) does not exclude that as a serious problem in these studies.

The next question is which if the factors do in fact explain the positive effects on behavior: is it the monitoring, the feedback or is it the rewards? Research by [Peer et al. \(2020\)](#) shows that all three do have some effect, but that the effects are more pronounced for monetary incentives. This is in line with two other studies ([Mullen et al., 2015](#); [Reagan et al., 2013](#)). It must be noted, however, that these two studies focused solely on speeding offences.

A special type of insurance incentive is PAYD (pay-as-you-drive) in which premiums are directly based on the driving behavior of policyholders. Participants (under 30 years of age) in a Dutch experiment ([Bolderdijk et al., 2011](#)) could earn a €50 reward: €30 was designated as a reward for keeping to the speed limit, €15 was for reduced mileage, and €5 was for not driving during weekend nighttime hours. Their vehicles were equipped with GPS devices allowing for the monitoring where participants were driving, at what speed they were driving and at what time, and how many kilometers they were travelling. The first reward proved to be effective, the other two were ineffective. The researchers concluded that PAYD—as applied in this experiment—leads to modest, but relevant reductions in the driving speed of young adults. If this result is then associated with the “power-law of Nilsson” (small speed changes have substantial effects on crashes/injuries/deaths), these results are promising for road safety.

In conclusion, it can be argued that reductions in insurance premiums provide extrinsic motivation to achieve safer behavior and therefore deserve to be investigated further. They can be applied to all age groups, but young drivers (relatively high risk, relatively high premium) were chosen as target audiences in many studies. An Israeli study ([Musicant and Lotan, 2016](#)) reports that when (young) participants had received the award, they stopped using the app. We also found that the usually small-scale programs required great effort to be organized. This raises the question of the cost-effectiveness of such programs. Furthermore, the extent of the potential range of these programs remains to be seen: the share of road crash casualties to be saved by using insurance rewards.

Key Lessons Learned

A number of guidelines that an effective reward program should meet are presented in a SWOV Fact sheet ([SWOV, 2011](#)). These lessons learned are summarized by Marjan Hagenzieker based on the literature and her own work (1999):

- Clearly inform participants of what they should do to receive a reward and how this is determined; use simple and clear “rules.”
- Avoid giving unexpected rewards to participants that they (could) not have anticipated on in advance.
- Use direct, immediate rewards; these may be supplemented with a lottery system with a chance of a reward.
- Make sure that the participants find the obtainable rewards attractive.
- Be careful with the value of the reward. Rewards need not be great in order to be effective. They should be large enough to induce behavioral change, but not so large that they are the only motivation for the desirable behavior.
- Consider a combination of measures. A reward program is more effective when it is combined with other interventions, such as training or police surveillance.
- Give fast and clear feedback on behavior and improvements with respect to the actual goals. This enforces the effects.
- Repeat the program at regular intervals to achieve also long(er)-term effects. However, reward programs do not need to continue very long to be effective; programs continuing for a few weeks can already be effective to bring about substantial short-term effects. However, repetition is necessary to achieve longer-term effects.
- Monitor the desirable behavior systematically.
- Measure the desirable behavior also prior to the start of the program, not only to set a realistic target, but also to be able to determine the effectiveness afterwards.

Conclusions

In road safety improvement, opportunities have been used in the fields of engineering, enforcement and education for many years; the three E's. Many effective and cost-effective measures are known and are being applied (see, e.g., [Elvik et al., 2009](#)). In many countries, implemented measures have reduced the traffic risks and reduced the number of casualties. The Safe System approach

contains the latest insights on how to make road traffic even safer. Safe System uses “man is the measure of things” as its basis: firstly, people make mistakes that can lead to road crashes and secondly, the human body has limited physical ability to tolerate crash forces before harm occurs.

The Safe System approach tries to strengthen the entire and therefore all components of the road system (human, vehicle, road) and ensures that the road user is still protected even if a component does not function properly. This also means that efforts are being made to eliminate some high-risk elements, such as four-arm intersections outside the urban area with speed limits of 100 km/h, or head-on collisions on motorways. Serious crashes can and will occur at such locations. Finally, not only does a road user have the duty to behave safely, but there is a shared responsibility with responsible road managers, vehicle manufacturers and those who take care of the organization and post-crash quality. This vision does not coincide well with the idea that is often expressed: human error plays a role in almost all crashes and therefore the performance of road users’ needs to be improved by more Enforcement and Education. A very readable article “Crash causation and prevention” by Hauer (2020) supports this idea.

This leads to the question whether giving a reward or promising an incentive for safe behavior fits into the Safe System approach and can make an additional contribution. This chapter answers this question, which also focuses on two other forms of behavioral influence: enforcement and punishing, and nudging. Steering behavior in a safe direction through legislation, monitoring behavior (by the police) and punishing traffic violations is general practice in many countries. There is also a fair amount of knowledge about how to make this effective and each country will need to develop its own chain of enforcement using general knowledge. Of course, there are practical limitations (supervision may be limited, and political priorities for the police may not lie with traffic control), but enforcement should be part of any road safety strategy.

Not as much is known about applying nudging to make traffic safer, but it is a very promising method that has proven its usefulness in other areas. It deserves attention in the development of self-explaining roads in a Safe System approach.

Finally, rewards and incentives. Giving rewards or working with incentives to solicit safer behavior is quite unusual in road safety and evidence of the effects of rewarding in road safety policies is limited. This chapter does include some examples of rewarding safe behavior, in particular rewards given in the private sector and giving discounts on insurance premiums. Occasionally, positive effects are reported, but a guarantee of positive effects cannot be given. The contribution of a reward program to achieving a national target (e.g., 25% fewer deaths in 10 years) is limited, because, as the scientific literature now shows, the scale of reward programs is considerably small. Pilot programs have not been scaled up. It is not evident how to organize large-scale applications of reward or incentive programs. It is not to be expected that replacing enforcement programs, in which dangerous behavior is punished, by programs that reward safe behavior is bringing positive road safety effects. However, the experiences gained with reward programs provide sufficient starting points to draw up a list of “key lessons learned.”

References

- Bolderdijk, J., Knockaert, J., Steg, E., Verhoef, E., 2011. Effects of pay-as-you-drive vehicle insurance on young drivers' speed choice: results of a Dutch field experiment. *Accid. Anal. Prev.* 43, 1181–1186.
- Elvik, R., 2014. Rewarding safe and environmentally sustainable driving: systematic review of trials. *Transport. Res. Rec.* 2465, 1–7.
- Elvik, R., Vaa, T., Høy, A., Sørensen, M., 2009. *The Handbook of Safety Measures*, second edition. Emerald Group Publishing.
- European Road Safety Observatory, 2006. Work-related road safety, retrieved September 14, 2007 from www.erso.eu.
- Haddon, W., 1972. A logical framework for categorizing highway safety phenomena and activity. *J. Trauma* 12 (3), 193–207.
- Hagenzieker, M.P., 1999. Rewards and road user behaviour; an investigation of the effects of reward programs on safety belt use. PhD thesis, Rijksuniversiteit Leiden.
- Hauer, E., 2020. Crash causation and prevention. *Accid. Anal. Prev.* (143), 105528.
- Haworth, N., Tingvall, C., Kowadlo, N., 2000. Review of Best Practice Road Safety Initiatives in the Corporate and/or Business Environment, Report No. 166. Monash University, Melbourne.
- Horrey, W., Lesch, M., Dainoff, M., Robertson, M., Noy, Y., 2012. On-board safety monitoring systems for driving: review, knowledge gaps, and framework. *J. Saf. Res.* 43, 49–58.
- Hurst, P., 1980. Can anyone reward safe driving? *Accid. Anal. Prev.* 12 (3), 217–220.
- Lahrmann, H., Agerholm, N., Tradisauskas, N., Naess, T., Juhl, J., Harms, L., 2012. Pay as you speed ISA with incentives for not speeding: a case of test drivers. *Accid. Anal. Prev.* 48, 10–16.
- Mullen, N., Maxwell, H., Bedard, M., 2015. Decreasing driving speeding with feedback and a token economy. *Transport. Res. F* 28, 77–85.
- Musicant, O., Lotan, T., 2016. Can novice drivers be motivated to use a smartphone-based app that monitors their behavior? *Transport. Res. F* 42, 544–557.
- NETS, 2014. NETS' comprehensive guide to road safety. Network of Employers for Traffic Safety, Vienna, Virginia.
- Newnam, S., Watson, B., Murray, W., 2002. A comparison of the factors influencing the safety of work-related drivers in work and personal vehicles. In: *Proceedings of the 2002 Road Safety Research, Policing and Education Conference*. Transport SA, Adelaide, pp. 488–495.
- Peer, S., Muermann, A., Sallinger, K., 2020. App-based feedback on safety to novice drivers: learning and monetary incentives. *Transport. Res. F* 71, 198–219.
- Plant, K., McIlroy, R., Stanton, N., 2018. Taking a '7 E's' approach to road safety in the UK and beyond. In: Charles, R. and Wilkinson, J. (Eds.), *Contemporary Ergonomics and Human Factors 2018*. Chartered Institute of Ergonomics & Human Factors.
- Reagan, I., Bliss, J., Van Houten, R., Hilton, B., 2013. The effects of external motivation and real-time automatic feedback on speeding behavior in a naturalistic setting. *Hum. Factors* 55, 218–230.
- Reason, J., 1990. *Human Error*. Cambridge University Press, Cambridge.
- Redelmeijer, D., Tibshirani, R., Evans, L., 2003. Traffic-law enforcement and risk of death from motor-vehicle crashes: case-crossover study. *Lancet* 361 (9376), 2177–2182, June 28 2003.
- Rumar, K., 1999. *Transport Safety Visions, Targets and Strategies: Beyond 2000*. European Transport Safety Council, Brussels.
- Sabey, B., Staunton, G., 1975. Interacting roles of road environment, vehicle and road user in accidents. In: *5th International conference of the International Association for Accident and Traffic Medicine*, London.
- Van Schagen, I., Machata, K., 2012. *The BestPoint Handbook: Getting the best out of a Demerit Point System*. Deliverable 3 of the EC project BestPoint. Kuratorium für Verkehrssicherheit KfV, Vienna.

- Shinar, D., 2007. *Traffic Safety and Human Behavior*. Emerald Group Publishing.
- SWOV, 2011. SWOV Fact Sheet. Rewards for Safe Road Behavior. <https://www.swov.nl/publicatie/rewards-safe-road-behaviour>.
- Thaler, R., Sunstein, C., 2008. In: *Nudge. Improving Decisions about Health, Wealth and Happiness*. Yale University Press, New Haven and London.
- Theeuwes, J., Godthelp, H., 1995. Self-explaining roads. *Saf. Sci.* 19 (2-3), 217–225.
- Treat, J., Tumbas, N., McDonald, S., Shinar, S., Hume, R., Mayer, R., Stansifer, R., Castellan, N., 1977. *Tri-Level Study of the Causes of Traffic Accidents: Final Report*. Indiana University, Bloomington Institute for Research in Public Safety and National Highway Traffic Safety Administration.
- TSR, 2016. *Advancing Road Safety. Best Practices for Companies and their Fleets. Guidelines for Developing and Managing Transportation Programs*. <https://www.togetherforsaferroads.org/>.
- United Nations Road Safety Collaboration (unknown). *Global Plan for the Decade of Action for Road Safety 2011–2020*. WHO, Geneva.
- Wegman, F., Aarts, L. (Eds.), 2006. *Advancing Sustainable Safety. National Road Safety Outlook for 2005–2020*. SWOV, Leidschendam.
- Wegman, F., 2013. Traffic safety. In: van Wee, B., Annema, J-A., Bannister, D (Eds.), *The Transport System and Transport Policy*, Edward Elgar, Cheltenham (Chapter 11).
- WHO, 2018. *Global Status Report on Road Safety 2018*. WHO, Geneva.

Enforcement and Fines

Matúš Šucha, Ralf Rissler, Palacký University in Olomouc, Czech Republic

© 2021 Elsevier Ltd. All rights reserved.

Introduction	263
Background Theories	263
3E Model of Traffic Safety	263
The Role of Norms and Traffic Safety Culture	264
Law enforcement—The deterrence theory	264
A Glimpse on Psychological Learning Theory	264
Behavioral Interventions - Educational and Therapeutic Measures	265
Implications in Road Safety Work	265
Traffic Law Enforcement is Usually Punishment	265
What to Consider When Applying Punishment (Fines)	266
Awareness Raising/Campaigning as a Complement to Police Enforcement	266
Reinforcement in Addition to Law Police Enforcement	266
Does the Size of the Fine Matter?	266
Demerit Point System and Driver Rehabilitation	267
Recommendations for the Effective Enforcement in the Road Safety Domain	267
References	267
Further Reading	268

Introduction

Traffic laws must be enforced, to sanction violations and to deter offenders and encourage responsible driving. Enforcement actions target violations like speeding, drinking and driving, not using a seat belt, failing to stop at a red traffic light, use of mobile phones and others. Traditional methods of police enforcement include on-the-spot roadside checks and the use of automated devices such as speed cameras. To be most effective, police enforcement should be publicized, and involve a mix of highly visible and low-profile activities. To maximize the road safety benefit, enforcement should be aimed at road rule violations that have been proven to increase the likelihood or severity of crashes.

Background Theories

3E Model of Traffic Safety

Regarded as a classic model used in traffic sciences, 3E proposes that road users should be perceived from three different points of view: Education and Training, Engineering, Enforcement (e.g., Evans, 2004). If we wish to achieve a real change in road users' behavior (in terms of improving safety or eliminating risky behavior), we need to exert an influence on road users at all the above levels.

Concerning education and training, we need to consider what is referred to as the mechanism of negative effects (Rothengatter, 1981). This mechanism involves the normalization of risky behavior by exposing road users to risk and increasing the level of their acceptance of such risk and their confidence. Engineering primarily involves the planning and implementation of traffic infrastructures and the development of means of transport in such a way as to make them responsive to human needs and nature, as well as promoting the desired safe behavior (Wolshon and Pande, 2016). The concept that is brought up with the greatest frequency in this respect is a self-explaining environment, which should provide intuitive guidance for people as to which behavior is appropriate and safe in a given situation (Kuiken et al., 2012). Another principle involves a forgiving traffic environment, which implies that we need to plan and build the traffic environment in such a way as to ensure that it can offset our mistakes and enable us to learn from such mistakes (Herrstedt, 2006).

The last approach under consideration involves enforcement. It has been demonstrated that this approach to traffic has a positive effect on public health (Tay, 2005; Watling and Leal, 2012), that is effective and fair enforcement influences road users' behavior in terms of reducing their risky behavior. Nevertheless, several preconditions must be met for enforcement measures to be effective. Effective enforcement requires that:

- people know the rules (i.e., they know what behavior is appropriate and expected)
- they must be able to apply their knowledge (i.e., they need to possess the necessary skills and be able to act on their knowledge)
- the benefit derived from breaking the rules must be smaller than the cost of the imminent sanction (i.e., they must be motivated).

The positive effect of enforcement on road safety is an increasing function of certainty (that a person will be sanctioned), strictness (the sanction must have a real impact on a person), and an immediate threat of punishment (a person has no chance of avoiding the sanction) (Armour, 1984).

The Role of Norms and Traffic Safety Culture

Leviakangas (1998) defined the **traffic culture** as a complex of all factors that affect attitudes, skills and behaviors of drivers as well as infrastructure and vehicles. This pushes researchers to think about the definition of term “traffic safety culture”. Edwards et al. (2014) refer to **traffic safety culture** as “the assembly of underlying assumptions, beliefs, values and attitudes shared by members of community, which interact with a community’s structures and systems to influence road safety related behavior.” The traffic culture operates at different levels. It created a background for the multilevel framework, which consists of four levels:

- micro – “individual level factors”;
- meso – “group/community level factors”;
- macro – “national level factors”;
- magno – “ecocultural socio-political level factors”.

Özkan and Lajunen (2011) also proposed the term “**traffic safety climate**” as a way to manifest the traffic safety culture. According to them, it is defined as “road users’ (e.g., drivers, bicyclists, and pedestrians) attitudes and perceptions of the traffic in a context (e.g., country) at a given point in time.”

There are a number of studies, which developed above-mentioned problematics on the cross-national level. The differences among cultures in cases of attitudes toward traffic and risk perception were pointed out by Lund and Rundmo (2009).

Law enforcement—The deterrence theory

The concept of deterrence builds on the notion that “prevention is better than punishment.” Deterrence is, however, never the only or primary purpose of legislation (the legal system). The essence of deterrence is the threat of punishment realized through a certain form of sanction. Deterrence is a way of assuming control on the basis of fear. If drivers commit violations without fearing the consequences, then, by definition, deterrence has no effect on them.

Deterrence generally involves behavioral control, which occurs when a potential perpetrator is not willing to take the risk of engaging in certain behavior out of fear of the consequences (Beyleveld, 1979). Deterrence is thus a mechanism for achieving compliance.

Punishment deters the perpetrator from forbidden behavior, which is considered individual prevention, while also deterring other individuals who might become perpetrators, which is referred to as general prevention. The assumption is that every person knows what is allowed and prohibited, as well as being aware of the possible consequences of illegal conduct. Stafford and Warr (1993) suggest that people’s experience of punishment, both indirect and direct, makes them less prone to violating the rules, as it increases the perception of the risk of punishment. According to the deterrence theory, punishment is considered a means of discouraging perpetrators (including potential perpetrators) from committing violations or offences. The effect of punishment can be divided into several groups.

The first involves the aforementioned theory of general prevention, which is further broken down into negative and positive: the negative theory of general prevention proposes that the threat of punishment will deter the perpetrator from offending. The positive theory of general prevention proposes that strict and unavoidable punishment imposed within a short time of the offence will strengthen confidence in the operation of justice and the national legal system.

The second group is constituted by the theory of special prevention: it is assumed that strict punishment imposed on the perpetrator according to the law will be a sufficient warning for others too.

Another type is the theory of individual prevention. It is based on the idea that the punishment imposed on a perpetrator will cause them harm which will deter them from continuing to engage in illegal conduct without the need for additional remedial measures.

The fourth is the rehabilitation theory. It proposes that a perpetrator’s illegal behavior results from a range of factors (family, upbringing, social circumstances, psychological preconditions, etc.) which influence their conduct and development. The purpose of punishment here is the provision of professional help which will make it possible for the offender to understand their traits, behavior, and the influence of the social environment which led to their offending. The rehabilitation theory seeks to have the perpetrators reintegrated into society on the basis of a change in their behavior achieved by means of their understanding the problematic nature of their previous behavior in the social context (Kuchta and Valkova, 2005).

The final type involves the restitution theory, which highlights the necessity of making amends for the damage and consequences resulting from their unlawful conduct. In these terms, crime is viewed as a threat to society and the damage to the victims it entails.

A Glimpse on Psychological Learning Theory

As human beings are provided with innate behaviors, it is necessary that most of the preconditions for social interactions (e.g., speech, opinions, attitudes) are acquired by experience. They need to be learned. Skinner in his learning theory (1938, 1953) coined the concepts “stimulus” and “response”. In his theory, responses are divided into two classes: reflexes and operands. Reflexes are elicited by stimuli immediately and directly. An example is the pupil reflex: the diameter of the pupil changes automatically with each change of brightness. Operands are arbitrary responses which lead to consequences that are more or less consciously strived for

Table 1 Types of reinforcements and punishments

Reinforcement	1. Positive consequences	2. Nonappearance of negative consequences
Punishment	3. Unpleasant consequences	4. Nonappearance of positive consequences

by the individual. However, they can develop the character of a reflex. For example, if they are regularly connected in time or space to certain situations or circumstances that lead to reinforcement or to negative consequences, those situations or circumstances can elicit such responses in a reflex-like way. To overtake another car according to this definition can be operant behavior. To be faster than the other car is the reinforcement, the car in front is the trigger.

Learning history – for example, frequency of such experiences – determines how strong the degree of automation (= character as a reflex) of the behavior is. The timely distance between the operant and the reinforcement is thereby important for the learning process, it has a strong impact on the speed of learning. The shorter the distance in time, the nearer the reinforcement (or the punishment) to any behavior the faster the learning process. The principles of reinforcement (= facilitation of behavior) and punishment (inhibition of behavior) are of great importance in road traffic. There are lots of rules and there are many legally envisaged sanctions and punishments. Theoretically, there are four different types of reinforcement and punishment (**Table 1**).

Behavioral Interventions - Educational and Therapeutic Measures

Behavioral interventions aim to contribute to the prevention and resolution of behavioral disorders by addressing their causes and determinants. In the context of road offenders, these measures are intended to address the underlying causes that explain the occurrence of inappropriate or deviant driving behaviors. Whether or not they apply to the context of road traffic offenses, the development of such interventions must be based on specific measures and fundamental principles. In the context of traffic offenses, three functioning components turn out to be of major importance:

- Principle of risk: it refers to the prediction of the repetition of a risk for a specific person. The principle implies that the intensity of an intervention must be proportional to the risk of repetition.
- Principle of need: it is relevant for the determinants of an offense (road) behavior. On the basis of this principle, effective interventions target the factors associated with transgressive behavior.
- Principle of the answer: it refers to the probability of a good answer from a specific person. According to this principle, the design of the program must be consistent with the skills, abilities and motivation of the participant.

Educational and/or therapeutic measures are one of the possible forms of “punishment” for traffic offenders. These are generally part of a legal context, most often for severe traffic offences, and can be imposed on offenders either in addition to or instead of a “traditional” sanction (i.e., withdrawal of driver’s license, fines imprisonment) or even to reduce its heaviness. For example, educational intervention may be imposed as a necessary condition to recover the driver’s license after withdrawal.

The presupposition of such measures is threefold: (1) driving is in essence a ‘learned’ behavior and can therefore be modified and influenced; (2) traditional punishments (e.g., fines or withdrawal of driver’s licenses) are not always sufficient to lead to behavior change; (3) these measures are necessary to restore driving ability in completeness driver’s assessment (Panosch, 2001).

Conceptually, taxonomy based on the level of empathy has identified four profiles of drivers at risk (Brion et al., 2018):

1. Anti-social drivers: the lowest level of empathy and not understanding the consequences of their behavior.
2. Careless or ignorant drivers: ignore the needs of others even if they are capable of empathy; drive irrespective of other road users but are in principle sufficiently skilled to drive safely.
3. Drivers in search of risk-taking: has a good level of empathy but has a poor ability to assess risks and drive for the excitement and fun that the road gives them.
4. Unskilled drivers: Has a good level of empathy but the level of risk is determined by limited experience, functional limitations or physical vulnerability. This group consists of older drivers, drivers with some disease, healthy drivers with an anxious driving style or those who are not assertive enough and cannot withstand social pressure, but also of young drivers with still low experience level.

Implications in Road Safety Work

In practice, above-mentioned theories become interlinked in such a way as to ensure that traffic-related sanctions meet the objectives of a national punitive policy to the effect of protecting the public, resocializing the offender, making amends for damage caused, and deterring potential perpetrators.

Traffic Law Enforcement is Usually Punishment

From a pedagogic point of view, positive reinforcement is a useful method to enhance wished-for behavior. However, it is difficult to implement in road traffic. Punishment functions, as well, but tends to have negative side effects (Schlag, 1995). For instance, it is

problematic that by providing negative consequences you stop nondesired behavior, but if no help is given, no alternative behavior will develop. The resistance of any behavior, enhanced by punishment, toward extinction is lower compared to reinforcement: the behavior is no longer carried out when nondesired behavior is not followed by punishment. Well-known examples include what happens when a speed measurement radar is taken away or when the alcohol-interlock period of an infringer ends: The probability that the nondesired behaviors—speeding, driving under the influence of alcohol—appear again when there is no more punishment is high.

Moreover, punishment can cause anxiety, or reactance towards those who punish. Such effects would imply that people do not want to learn anything from the punishing institution; for example, you don't want authorities to tell you what to do. Especially hard punishment can also cause aggression, which of course is not supporting traffic safety.

As far as police work is concerned, up to now no effective models of systematic positive reinforcement have been found that would work in connection with law enforcement on the road.

What to Consider When Applying Punishment (Fines)

The conclusion is not to have a traffic law enforcement system without punishment. Rather, it is important to consider some aspects in connection with the application of punishment (see e.g. [Zimbardo and Ruch, 2013](#)):

- Punishment should be provided in a way that it cannot be avoided (it should be *consequent*)
- Punishment should follow immediately after the behavior (it should be *contingent*). A fine several weeks after speeding does not provide any relationship of the punishment (= the fine) to the nonwished for behavior.
- Until the desired behavior is *internalized* (= is carried out habitually) punishment should follow after every erroneous behavior as far as possible (it should be *consistent*).

In sum, one has to punish consequentially, contingently and consistently in order to make the relation between nondesired behavior and fine clear and to keep it that way. But some other conditions should be fulfilled, as well:

- The punished person should be able to understand the reason for the punishment. There should be enough problem awareness to understand why to avoid the nondesired behavior.
- It is necessary that the punished persons trust in the authorities (and vice versa).

Awareness Raising/Campaigning as a Complement to Police Enforcement

To reach these later goals societal development of a certain status of awareness is needed. In this connection, awareness raising with the help of arguments and facts has turned out to be a most important complement to law enforcement measures in traffic. Not least, meta-analyses have shown that law enforcement is efficient when systematically and professionally combined with traffic safety campaigns (e.g., [Aamodt, 2004](#); [Elvik, 1997](#); [Elvik, 2016](#)). Among other things, public institutions and decision makers, supported by the media, need to display problems connected to road traffic and to mobility in an objective and evidence based way. For instance, one needs to make it clear that traffic is a system where one's own behavior has impact on others and vice versa and such a system can only work if one respects other road users. Constitutional law concerning human rights would support such an approach ([Eisenberg, 1986](#); [Halmetschlager, 2008](#); [Popper, 1971](#)). The interests of other persons are always connected to the consequences of one's own behavior as a road user, especially as a motorized one. Thus, prosocial behavior needs to be argued for, among other things.

Reinforcement in Addition to Law Police Enforcement

Even if it is difficult to envisage an enforcement system applied by the police that uses positive reinforcement, such measures can be—and are—applied by employers, or by insurance companies. [Schade et al. \(2003\)](#) mentioned the Swedish Televerket study by [Gregersen et al. \(1996\)](#) where the effectiveness of four different interventions (driver training, group discussion, information campaign, bonus program) was analyzed. Four groups that underwent such measures were compared to a control group without measures. All four groups, but not the control group, showed improvements concerning accident frequency and accident costs. The best results could be found in the groups "discussion" and "training."

Other types of reinforcement are bonus-malus systems as applied by insurance companies. Drivers who did not get involved into traffic accidents over a certain period of time do receive a reduction of their insurance fees, while involvement in accidents leads to an increase of the fee. A reduced fee can be raised again, if there is accident involvement again, and the malus can be cancelled, leading to a reduction of the fee, if no accident involvement was given for a certain period of time.

Table 2 shows examples of the four different types of reinforcement and punishment (**Table 1**) that are relevant in connection with traffic safety. Only the unpleasant consequences refer to traffic-law enforcement in the narrow sense:

Does the Size of the Fine Matter?

The impact of the size of the punishment is assessed as being low as [Björnskau and Elvik \(1992\)](#) and [Elvik and Christensen \(2007\)](#) could show. In the analysis of 2007, higher fines led to slightly positive effects concerning belt use. But the authors conclude that

Table 2 Types of reinforcements and punishments

	<i>Positive consequences</i>	<i>Unpleasant consequences</i>
Following the behavior	Reinforcement: e.g. lower insurance fee after one year of driving without accidents	Punishment: e.g. fine after an infringement; higher insurance fee after accidents
Removed after the behavior	Indirect reinforcement: removing an insurance malus after years of driving without accidents	Indirect punishment: withdrawal of driver's license; removing an insurance bonus

there was an overlap effect by the campaign to wear seat belts that was carried out at the same time. Concerning speeding, higher fines had low effect. The explanation usually is that the probability to get caught is much too low. The point of view that mainly the subjective assessment of the probability to get involved into an enforcement and punishment measure is relevant for law abiding behavior is supported by [Schlag et al. \(2012\)](#).

Demerit Point System and Driver Rehabilitation

Point demerit systems are a much-debated topic in the traffic context. In addition to commenting on general issues concerning appropriate ways of addressing drivers' transgressions, traffic psychologists engage in frontline work with those whose driving licenses were suspended as a form of punishment. Because of their improper traffic behavior, such drivers are required to undergo a special psychological assessment before they can have their license reinstated. During the assessment for psychological fitness to drive, the psychologist should determine whether the driver has learnt their lesson from the punishment. Interestingly, those drivers who had lost all their points in particular were found to show less sensitivity to punishment ([Goodwin et al., 2013](#)). It is essential to recognize, however, that no psychological assessment in itself will make a person a better driver. Other follow-up interventions and rehabilitation programs need to be embarked upon. Sanctions alone do not necessarily result in what is hoped for, the long-term rectification of risky or aberrant traffic behavior. The issue of reoffending is associated with the absence of insight into the hazardous nature of one's own behavior, the adverse arrangement of personality traits, and rigid attitudes toward road safety. A certain group of drivers sees punishment only as a warning which they ignore or cannot change their behavior to comply with. This is where rehabilitation programs may be useful in providing help to those risky drivers who are seeking to change their behavior and avoid repeating their previous mistakes in the future ([Sucha et al., 2013](#)).

Recommendations for the Effective Enforcement in the Road Safety Domain

Based on the large scale European research project ESCAPE ([Mäkinen et al., 2002](#)), we conclude the following recommendations for the effective enforcement in the road safety work:

- There is wide-spread support for enforcement among the public.
- The mechanism of enforcement is based on deterrence. Ample empirical evidence show that the deterrence principle works in experiments. Enforcement based on deterrence is cost effective.
- Enforcement needs can be crudely classified into three groups: (1) old priority areas all nations share: alcohol, seat belts, speeds; (2) special needs of individual nations, and (3) new needs: e.g. cross-border traffic, illicit drugs, road rage.
- Drink driving enforcement shows that long-term, durable and combined methods (high risk of detection, education, information) bring about good results.
- There is no proof that severe punishments are effective in suppressing reckless driving.
- Interference by the police is important in influencing speeds; perhaps more important than severe punishments.
- The need for massive deterrence-based enforcement systems targeted to catch as many offenders as possible may be questioned; it is a sign of disintegration of the transport safety system.
- The possibilities of enforcement do not reside in enforcement itself, but in integrated traffic safety work; the road infrastructure should be improved as far as possible.
- Enforcement is not solely the responsibility of the police; rather a division of enforcement responsibilities between different authorities is needed.
- Improvement of in-vehicle technology (intelligent speed adaptation; lateral and longitudinal support systems) decreases the need for enforcement in the long run.

References

- Aamodt, M.G., 2004. Research in Law Enforcement Selection. Universal-Publishers.
- Armour, M., 1984. A review of the literature on police traffic law enforcement. *Austr. Road Res.* 14 (1.).
- Beyleveld, D., 1979. Identifying, explaining and predicting deterrence. *Br. J. Criminol.* 19 (3), 205–224.

- Björnskau, T., Elvik, R., 1992. Can road traffic law enforcement permanently reduce the number of accidents? *Acc. Anal. Prevent.* 24 (5), 507–520.
- Brion, M., Meunier, J.-C., Tsapi, A., Vissers, J., Sucha, M., Drimlova, M., Kluppels, L., 2018. Educational and therapeutic approaches as responses to traffic offences: Review of literature and applicability to the Belgian context. Vias Institute - Knowledge Centre, Brussels, Belgium.
- Edwards, J., Freeman, J., Soole, D., Watson, B., 2014. A framework for conceptualising traffic safety culture. *Transp. Res. Part F Traffic Psychol. Behav.* 26, 293–302.
- Eisenberg, N., 1986. *Altruistic Emotion, Cognition and Behaviour*. Lawrence Erlbaum Associates Publishers, Hillsdale, New Jersey.
- Elvik, R., 1997. Effects on accidents of automatic speed enforcement in Norway. *Transport. Res. Rec.* 1595 (1), 14–19.
- Elvik, R., 2016. A theoretical perspective on road safety communication campaigns. *Acc. Anal. Prevent.* 97, 292–297.
- Elvik, R., Christensen, P., 2007. The deterrent effect of increasing fixed penalties for traffic offences: the Norwegian experience. *J. Saf. Res.* 38 (6), 689–695.
- Evans, L., 2004. *Traffic Safety*. Bloomfield Hills, Science Serving Society.
- Gregersen, N.P., Bremer, B., und Moren, B., 1996. Road safety improvement in larger companies. An experimental comparison of different measures. *Acc. Anal. Prevent* 28 (3).
- Goodwin, A., Kirley, B., Sandt, L., Hall, W., Thomas, L., O'Brien, N., Summerlin, D., 2013. *Countermeasures that Work: A Highway Safety Countermeasures Guide for State Highway Safety Offices*, 7th ed. National Highway Traffic Safety Administration, Washington, DC.
- Halmetschlager, M. 2008, <https://docplayer.org/22074401-Bildungspsychologie-i-wahlfachmodul.html>.
- Herrstedt, L., 2006. Self-explaining and Forgiving Roads—Speed management in rural areas. In ARRB Conference.
- Kuchta, J., Valkova, H., 2005. *Zaklady kriminologie a trestní politiky*. Praha: C. H. Beck.
- Kuiken, M., van der Horst, R., Theeuwes, J., 2012. *Designing Safe Road Systems: A Human Factors Perspective*. Ashgate Publishing, Ltd.
- Leviakangas, P., 1998. Accident risk of foreign drivers – The case of Russian drivers in south-eastern Finland. *Acc. Anal. Prevent.* 30, 245–254.
- Lund, I.O., Rundmo, T., 2009. Cross-cultural comparisons of traffic safety, risk perception, attitudes and behaviour. *Safety Sci.* 47 (4), 547–553.
- Mäkinen, T., Zaidel, D.M., et. al., 2002. ESCAPE – Traffic enforcement in Europe: effects, measures, needs and future. Final report of the ESCAPE consortium. Retrieved from https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/projects_sources/escape_final_report.pdf.
- Özkan, T., Lajunen, T., 2011. Person and environment: traffic culture. In: Porter, B. (Ed.), *Handbook of Traffic Psychology*. Elsevier, London, pp. 179–192.
- Panosch, E., 2001. The development of driver improvement and retraining in Austria. *Psychol. Austria* 21 (3), 218–228.
- Popper, K.R., 1971. *The Open Society and Its Enemies: The Spell of Plato*, Princeton University Press, Princeton, New Jersey, pp. 109–110.
- Rothengatter, J.A., 1981. The influence of instructional variables on the effectiveness of traffic education. *Acc. Anal. Prevent.* 13 (3), 241–253.
- Schade J., et al., 2003. Anreizsysteme in der Verkehrssicherheitsarbeit – eine Expertenevaluation, im Auftrag des DVR. Lehrstuhl für Verkehrspsychologie der TU Dresden.
- Schlag, B., 1995. Lern- und Leistungsmotivation. UTB Band 1855, Opladen. In: Schade J. et al., *Anreizsysteme in der Verkehrssicherheitsarbeit – eine Expertenevaluation, im Auftrag des DVR. Lehrstuhl für Verkehrspsychologie der TU Dresden*.
- Schlag, B., Röbger, L., Schade, J., 2012. Regelbefolgung - ein Modell der Einflussgrößen/A traffic rule compliance model. *Zeitschrift für Verkehrssicherheit* 58 (2).
- Skinner, B.F., 1938. *The Behaviour of Organisms: An Experimental Analysis*. B.F. Skinner Foundation, Cambridge, Massachusetts.
- Skinner, B.F., 1953. *Science and Human Behaviour*. Macmillan Company, New York.
- Sucha, M., Rehnova, V., Koran, M., Cernochova, D., 2013. *Dopravní psychologie pro praxi Vyber, vycvik a rehabilitace ridicu*. Grada Publishing, Praha.
- Stafford, M.C., Warr, M., 1993. A reconceptualization of general and specific deterrence. *J. Res. Crime Delinquency* 30 (2), 123–135.
- Tay, R., 2005. General and specific deterrent effects of traffic enforcement: do we have to catch offenders to reduce crashes? *Journal of Transport Economics and Policy (JTEP)* 39 (2), 209–224.
- Watling, C.N. Leal, N.L., 2012. Exploring perceived legitimacy of traffic law enforcement.
- Wolshon, B., Pande, A., 2016. *Traffic Engineering Hand Book*. John Wiley & Sons.
- Zimbardo, P.G., Ruch, F.L., 2013. *Psychology and Life*. Pearson International, London.

Further Reading

- Alghuson, M., Abdelghany, K., Hassan, A., 2019. Toward an integrated traffic law enforcement and network management in connected vehicle environment: conceptual model and survey study of public acceptance. *Acc. Anal. Prevent.* 133, 105300.
- Britt, J.W., Bergman, A.B., Moffat, J., 1995. Law enforcement, pedestrian safety, and driver compliance with crosswalk laws: evaluation of a four-year campaign in Seattle (with discussion and closure). *Transport. Res. Rec.* (1485).
- Castillo-Manzano, J.I., Castro-Nuño, M., López-Valpuesta, L., Pedregal, D.J., 2019. From legislation to compliance: the power of traffic law enforcement for the case study of Spain. *Transport Policy* 75, 1–9.
- Elvik, R., 2006. Are individual preferences always a legitimate basis for evaluating the costs and benefits of public policy?: The case of road traffic law enforcement. *Transport Policy* 13 (5), 379–385.
- Gardiner, J.A., 2013. *Traffic and the Police: Variations in Law-Enforcement Policy*. Harvard University Press, Cambridge, Massachusetts, United States.
- Hijar, M., Trostle, J.A., 2003. Traffic law enforcement and safety. *The Lancet* 362 (9386), 833.
- Redelmeier, D.A., Tibshirani, R.J., Evans, L., 2003. Traffic-law enforcement and risk of death from motor-vehicle crashes: case-crossover study. *The Lancet* 361 (9376), 2177–2182.
- Rothengatter, T., 1982. The effects of police surveillance and law enforcement on driver behaviour. *Curr. Psychol. Rev.* 2 (3), 349–358.
- Vishnevsky, V.M., Larionov, A., 2012, November. Design concepts of an application platform for traffic law enforcement and vehicles registration comprising RFID technology. In 2012 IEEE International Conference on RFID-Technologies and Applications (RFID-TA). IEEE, pp. 148–153.
- Weiss, A., Freels, S., 1996. The effects of aggressive policing: the Dayton traffic enforcement experiment. *Am. J. Police* 15 (3), 45–64.

Epidemiology of Road Traffic Crashes

Sherrie-Anne Kaye*, **Judy Fleiter^{†,‡}**, **Md Mazharul Haque[§]**, *Queensland University of Technology, Centre for Accident Research and Road Safety—Queensland (CARRS-Q), Brisbane, QLD, Australia; [†]Global Road Safety Partnership, Geneva, Switzerland; [‡]Queensland University of Technology, School of Psychology and Counselling, Brisbane, QLD, Australia; [§]Queensland University of Technology, School of Civil Engineering and Built Environment, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Epidemiology and Scope	269
Road Traffic Crashes—High-Income Country Context	270
Five Behavioral Risk Factors Which Contribute Significantly to Road Deaths	270
Encouraging Safer Road Use	271
Road Safety Strategies	271
Legislation and Enforcement	271
Public Education, Advertisement, and Mass Media Campaigns	272
Road Traffic Crashes—Low- and Middle-Income Country Contexts	273
Road Safety Programs and Initiatives in Low- and Middle-Income Countries	273
Legislation and Enforcement	273
Public Education, Advertisement, and Mass Media Campaigns	274
Summary and Concluding Comments	274
Relevant Websites	275
References	275
Further Reading	275

Epidemiology and Scope

Epidemiology can be defined as the study of how often diseases and health issues occur within the population and the reasoning of why these events occur. Epidemiologists study disease and health problems in many disciplines, including in the area of road safety. Serious injuries and fatalities resulting from road traffic crashes are a major public health issue that has attracted intensified global attention since the United Nations declared a Decade of Action for Road Safety (2011–20). As reported in the most recent Global Status Report published by the World Health Organization (WHO) (2018), road crashes result in 1.35 million fatalities each year and between 20 and 50 million injuries. In 2016, road injuries were reported to be the 8th leading cause of death worldwide (ranked 10th in 2000). For the same year, road crash-related injuries were reported as the 8th and 10th leading cause of death, respectively, for upper-middle-income countries and low- and lower-middle-income countries. The road trauma burden falls disproportionately across the globe—deaths are disproportionately higher in low- and middle-income countries (LMICs) than in high-income countries (HICs), in relation to population levels and size of vehicle fleet. The rate of road traffic fatalities for low-income countries (LICs) is 3 times higher than the rate of HICs. More specifically, for LICs there is an average of 27.5 deaths per 100,000 population compared to an average of 8.3 deaths per 100,000 population in HICs (WHO, 2018). Socioeconomic status has also been shown to affect road traffic deaths, with individuals from low-income classes, irrespective of country, more susceptible to road traffic deaths than those from high-income classes. In addition, the WHO (2018) reported that worldwide road injuries are the leading cause of death for individuals aged 5–24 years. More specifically, the proportion of deaths from road injuries for people aged 5–29 years is 12.5%. Further, male road users are more likely to be involved in road crashes when compared to female road users.

A broad range of policies and programs have been introduced in an effort to reduce serious injuries and fatalities associated with road crashes. Notably, at the global level, the United Nations Decade of Action for Road Safety 2011–2020 was initiated in March 2010 by the United Nations General Assembly. This global movement aimed to first stabilize, and then reduce, forecasted levels of fatalities across a 10-year period. The Decade of Action for Road Safety 2011–2020 strategy had the ambitious target to halve the number of projected road deaths by 2020 (from 2010). In order to achieve this target, countries were encouraged to implement activities within their own national road safety strategies which aligned with five pillars: Pillar 1: road safety management, Pillar 2: safer roads and mobility, Pillar 3: safer vehicles, Pillar 4: safer road users, and Pillar 5: postcrash response. The Decade of Action has produced some important momentum and has increased global attention and efforts to reduce injuries and deaths on our roads in many countries. Yet, more needs to be done, as evidenced by the information contained in the recent 2018 Global Status report.

There is substantial variation in road trauma among various road user groups according to country-level income. Generally, road traffic crashes in HICs are more likely to involve drivers, whereas in LMICs, vulnerable road users, such as pedestrians and riders of motorized and nonmotorized two and three wheelers, are overrepresented in the crash data. Given the differences in road safety-related challenges across low to HICs, this article discusses the behavioral risk factors associated with road traffic crashes for these countries separately. First, this article will discuss the five of the leading high-risk behaviors reported to account for a large proportion of injuries and deaths in HICs. These five risk behaviors include speeding, driving while impaired by alcohol and drugs, not wearing

a seat belt, fatigue (drowsy driving), and driver distraction. It is acknowledged that not wearing a helmet on motorcycles is an additional risk factor. However, and given that helmet rates for motorcycles have increased substantially, to a point where in most HICs it is now not considered a major risk factor, this behavior is not discussed later. Next, this article will discuss approaches which have been implemented to change driver behavior, including road safety strategies, legislation and enforcement, and public education, advertisement, and mass media campaigns. This article will then discuss main human behaviors that account for the road traffic crashes in LMICs and present some of the road safety initiatives which have been introduced to reduce fatalities and injuries in these countries. It is acknowledged that factors other than human behavior contribute to road traffic crashes (e.g., road infrastructure, vehicle type and condition, and weather conditions). However, a fuller discussion including nonbehavioral factors is beyond the scope of this article.

Road Traffic Crashes—High-Income Country Context

Five Behavioral Risk Factors Which Contribute Significantly to Road Deaths

Five behavioral risk factors that contribute substantially to road deaths include speeding, driving while impaired by alcohol or drugs, seat-belt use, fatigue, and driver distraction (Fig. 1). Speeding refers to exceeding the posted speed limit and has been shown to increase both crash risk and crash severity. Even a low-level speeding (e.g., 5 km/h over the posted speed limit) can increase a driver's risk of being involved in a crash. While research has shown that drivers are aware of the risks associated with speeding, speeding behavior still remains relatively socially acceptable and commonplace in HICs, even those where speed-management countermeasures have been in place for decades. Driving while impaired is another behavior shown to increase crash risk. Alcohol consumption and drug use have both been shown to significantly impair cognitive and motor abilities and have negative effects on driving performance. For instance, alcohol can reduce the drivers' ability to judge speed and distance, increase drivers' reaction time, and impair vision. While driving under the influence of alcohol or drugs is generally less socially acceptable than speeding behavior, alcohol- and drug-impaired driving crashes remain prevalent.

Seat belts have been shown to significantly reduce injury severity that might be sustained in the event of a crash. Seat belts were first introduced into vehicles from 1959, and were made compulsory between the 1970s and 1990s. Since the introduction of

Speeding	• Refers to driving at a speed that exceeds the posted speed limit. • In 2016, 27% of fatal crashes involved at least one driver who was speeding in the United States.
Driving while impaired by alcohol and drugs	• Driving while impaired by alcohol refers to driving when over the legal blood alcohol concentration (BAC) limit of 0.08g/100mL. Driving while impaired by drugs refers to driving while under the influence of prescribed illicit drugs. • In 2016, 28% of fatal crashes were associated with alcohol-impaired driving in the United States.
Seat belts	• All occupants of passenger vehicles are required to wear appropriate an restraint (e.g. seat belts or child restraint) • In 2016, 44% of fatal crashes in the United States involved occupants who were unrestrained at the time of crash.
Fatigue	• Fatigue refers to driving while tired (also referred to as drowsy driving). • The NHTSA estimates 100,000 police reported crashes involved drivers who were driving while tired each year. However, it is thought that this number is higher.
Distraction	• Driver distraction refers to any activity that takes the drivers attention away from the driving task. • In 2016, 3450 people lost their lives while distracted behind the wheel in the United States (9.2% of total road fatalities).

Figure 1 Five risk factors which contribute to road deaths based on data from the United States (NHTSA, 2017).

seat-belt law—which requires all vehicle occupants to wear seat belts—seat-belt use has been steadily increasing. From the mid-1970s/1980s in Australia, later in the United States and Europe, child restraint safety laws were also introduced requiring children under a certain age to use an approved child restraint or booster seat.

The next behavior, fatigue or driving while tired, has been shown to reduce vigilance and alertness and, as a result, leads to a decline in driving performance. The behavioral outcome of driving while tired has been reported to be similar to that of driving under the influence of alcohol. For instance, it has been reported that being awake for 17 h is equivalent to having a blood alcohol concentration (BAC) limit of 0.05 g/100 mL. However, and unlike the other behaviors discussed in this section, there are no current laws that regulate driving while tired except for heavy vehicle drivers (i.e., trucks, buses, and coach drivers) in some countries.

Finally, driver distraction refers to any activity that may take the drivers' attention away from the road (e.g., eating and drinking, talking to passengers, and phone use). In a vehicle traveling at 100 km/h, a driver who takes their eyes off the road for 2 s will have driven over 50 m without awareness of what is occurring in front of them. The focus on driver distraction has increased in recent years with the introduction of smartphones (Oviedo-Trespalacios et al., 2016). Smartphones can result in the following distractions for drivers: visual distractions (i.e., a driver is looking at their smartphone rather than motoring the road environment), physical distractions (i.e., a driver is required to take a hand off the steering wheel to access their phone), and cognitive distractions (i.e., a driver's attention is directed away from the road environment). Further, it has been estimated that drivers who use a smartphone anytime while driving are 4 times more likely to be involved in a crash compared to those drivers who do not use a smartphone when driving. The last two behavioral risk factors discussed earlier (i.e., driving while fatigued and distracted driving) are notable in that they pose specific challenges for law enforcers, making these behaviors difficult to manage.

Encouraging Safer Road Use

The following section of this article focuses upon countermeasures that have been implemented to improve road user behavior. These approaches include (1) road safety strategies, (2) legislation and enforcement, and (3) public education, advertisement, and mass media campaigns.

Road Safety Strategies

A large number of HICs have implemented national strategies aimed to reduce serious injuries and road crashes. The National Road Safety Strategy (2011–20) in Australia, New Zealand's Safer Journeys, Vision Zero from Sweden, Sustainable Safety from the Netherlands, and the European Commission: Strategic Action Plan on Road Safety are some examples of strategies which have been introduced to improve road safety. These strategies are based on the safe system approach to improving road safety. The safe system approach acknowledges that human error will continue to contribute to road crashes, but these errors should not cause serious injuries or deaths of road users. The safe system approach acknowledges that a crash can be the result of a combination of factors and, as such, consists of four main elements including safe roads and roadsides, safe speeds, safe vehicles, and safe road use. In other words, responsibility for road crashes is shared between all system designers and users, rather than the traditional approach that focused only on laying blame on the road user.

Legislation and Enforcement

As highlighted in the WHO's (2017) Save LIVES Technical Package—a compendium of evidence-based road safety countermeasures—legislation is a necessary and effective intervention to reduce road crashes and related trauma. Importantly, for legislation to be effective, it must be coupled with well-resourced, sustained enforcement activities that can assist with attitudinal, social, and behavioral change over time. In relation to the risky behaviors discussed earlier, legislation and various enforcement techniques have been implemented in various HICs to manage and reduce speeding, limit the amount of alcohol and drugs present while operating a vehicle, increase seat belt and child restraint use, and reduce driver distraction. Some examples of legislation and enforcement approaches for each of these behaviors are briefly discussed later.

Speed limits vary within and across HICs. For example, a default urban 50 km/h speed limit applies in Australia, whereas on highways the maximum speed limit is typically 110 km/h, with some state-based variation. In the United States, minimum speed limits may also apply on some roads, such as highways, and in the United Kingdom, some local authorities have adopted a 20 mph (32 km/h) speed limit for residential and urban streets. In France, there are two sets of speed limits: one that applies in dry weather and a lower speed that applies in wet conditions. A range of speed-management strategies, including speed enforcement, is used to encourage compliance with speed limits. Speed enforcement strategies vary, and can include the overt and covert use of fixed, mobile and section control/average speed cameras. Research has shown that speed cameras and appropriately administered related penalty regimes are effective in reducing speed-related crashes and associated fatalities and injuries. The use of intelligent speed adaptation (ISA) technology is increasingly being considered to assist in managing speeds, and a recent development in Europe indicates that all new models of cars sold in the European Union by 2022 will need to be fitted with ISA.

As noted earlier, driving under the influence of alcohol and drugs increases crash risk. As such, many HICs set and enforce strict BAC limits for drivers of light passenger vehicles and a zero BAC for novice and commercial drivers. BAC limits vary across countries. For example, in Australia and New Zealand the legal BAC limit is 0.05 g/100 mL for drivers who hold an open

(unrestricted) license (Australia) or for drivers 20 years and over (New Zealand). The legal BAC limit of 0.05 g/100 mL also applies to those who had a standard drivers' license in many European countries (range BAC limit of 0.00 g/100 mL–0.08 g/100 mL). In almost all of the United States and Canada, the legal BAC limit is higher, at 0.08 g/100 mL for drivers aged 21 years and older. Enforcement strategies differ across countries, primarily based on legislative power of enforcement agencies to stop and test a road user for the presence of alcohol, without direct evidence of impairment. In some jurisdictions, police are able to conduct random breath testing (RBT) which means a vehicle can be stopped at any time and the driver breath tested, without any suspicion of impairment. RBT has operated successfully in New Zealand, Australia, and some European countries (e.g., Finland and Sweden). Compulsory breath testing (CBT) is a different enforcement strategy, where all drivers who stop at CBT checkpoints are required to be tested for alcohol use. This enforcement technique is used in countries such as New Zealand. Another enforcement strategy gaining in popularity in some HICs is the legislated use of alcohol ignition interlocks, especially for recidivist offenders. An interlock device can be installed in a vehicle and is essentially a breathalyzer for individual vehicles that requires the driver to provide an alcohol-free breath sample or the vehicle will not start. Interlock devices are already prevalent in commercial vehicles in some HICs, compulsory for repeat offenders in others, and could also be mandated by the European parliament for all new vehicles from 2022. Enforcing the presence of drugs, however, is a more complicated enforcement issue when compared to alcohol. It is illegal in some HICs for drivers to drive under the influence of impairing drugs. This law includes both illicit drugs (e.g., methamphetamines) and prescription medications, which can be purchased over the counter at local pharmacies or drug stores. Similar to RBT, random roadside drug testing is one type of countermeasures used in countries, such as in Australia, to test for illicit drug use by drivers.

Legislation for the compulsory wearing of seat belts was introduced in many countries in the 1970s–90s. While there was some criticism of seat belts when they were first made mandatory, today there are high seat-belt wearing rates amongst drivers and passengers in many HICs. In Australia, the United Kingdom, and New Zealand, seat-belt rates are above 90% for drivers and passengers. Seat-belt usage rates in the United States differ by state; however, and similar to Australia, the United Kingdom, and New Zealand, has risen since the 1980s. While seat-belt use is mandatory for light vehicles, in some countries, seat belts are not required for bus occupants (unless seat belts are already installed in the bus). In addition to seat belts, child restraint laws have also been introduced in many HICs. Child restraint legislation states that it is mandatory for all children to wear an approved child restraint. The type of restraint differs based on the child's age/height, for example, and in Australia, rear-facing restraints for babies up to 6 months of age or booster cushions for children aged 4–7 years of age. Similar to adult seat belts, child restraints have been shown to reduce injuries and deaths associated with traffic crashes.

While it is important to note that driver distraction includes any activity which may distract from the driving task, more recently, the focus has been on distraction caused by mobile (cell) phone use. Given that mobile phone use is rated as one of the top driver distractions, many countries have introduced legislation that bans the use of handheld mobile phones while driving. Further, a number of countries, including France and Spain, have applied the legislation to ban the use of taking phone calls through Bluetooth headsets. For some states in Australia and the United States, novice drivers are also prohibited for using a mobile phone while driving. Despite the introduction of legislation that prevents mobile phone use while driving, many drivers still report using their phone to text, call, or engage in other online activities (e.g., social media or checking emails). Enforcement of phone use while driving can be difficult, especially if phones are used in a concealed manner, i.e., a driver hides their phone on their lap or underneath the steering wheel. Phone applications and functions, for example, the iPhones' "do not disturb while driving" function, can be used to block incoming text messages and phone calls; however, this technology relies on the driver installing and using this technology while driving. Handheld mobile phone use while driving is difficult for police to enforce. Unlike other illegal behaviors, which use camera technology, for example, speed cameras, the primary detection method used by police for handheld mobile phone is visual detection.

Compared to speeding, driving while impaired by alcohol, seat-belt use, and driver distraction, limited legislation and enforcement strategies have been introduced to combat driver fatigue. In some countries, there are fatigue-related heavy vehicle laws that prevent drivers from operating a heavy vehicle while fatigued. In Australia, for instance, work time and rest time are required to be counted in order to prevent fatigue. Further, drivers who operate heavy vehicles and who drive more than 100 km from their base are required to keep a work diary to record their work and rest breaks. For people with untreated sleep disorders, driver license restrictions may also apply in some countries. Compared to the other behaviors, driving while tired is difficult to assess. For instance, while there are speed cameras to determine speed and breath/blood testing to assess alcohol/drug impairment, there are no objective measures to assess driver fatigue. Technologies, such as using an in-vehicle camera to track eye movement activity, are currently being used in studies to assess drowsy driving. However, these technologies are not yet widely available.

Public Education, Advertisement, and Mass Media Campaigns

Public education, advertisement, and mass media campaigns are additional countermeasures used to promote safe road user behaviors. Advertisement and mass media campaigns can be used to publicize changes in legislation and/or penalties, support enforcement and advocacy activities, and to educate drivers on the negative consequences associated with risky and/or illegal behaviors and the positive consequences of safe road use (Lewis et al., 2019). In Australia, for instance, mass media campaigns were instrumental in promoting the introduction of RBT in the 1980s by highlighting that motorists could expect to be stopped and breath tested for the presence of alcohol anywhere and at any time (Cameron et al., 1993). Since then, sustained media campaigns have been used to reinforce this message. More recently, road safety campaigns have been used to promote other safe behaviors used

to protect vulnerable road users. For example, the minimum passing distance rule of 1 m (or 3 ft.) when passing a cyclist or to remind drivers of the reduce speed limits around schools when children are often seen entering or leaving the school grounds.

Road Traffic Crashes—Low- and Middle-Income Country Contexts

As noted earlier, road traffic deaths are disproportionately high in LMICs when accounting for population levels and size of vehicle fleet. The most recent WHO Global Status Report on Road Safety (2018) notes that no LIC has experienced a reduction in road-related deaths since 2013. LICs have the smallest proportion of motorized vehicles (1%) but account for 13% of global fatalities. In middle-income countries (MICs), the burden is proportionately greater—with 59% of the global vehicle fleet and 80% of global road deaths. When compared to HICs, there is also variation in the types of road users most affected by road trauma. In HICs, vehicle occupants are predominant. However, in the LMIC context, it is vulnerable road users such as motorcyclists, pedestrians, and cyclists who are overrepresented in road crashes. Further, and in LICs, young children are more likely to be involved in road crashes as vulnerable road users than children in HICs. Some parts of Asia and Africa have been, or are, experiencing dramatic increases in motorized two- and three-wheeler forms of transportation: a situation that brings unique challenges rarely faced in HICs to date. Regulating the safety standards of such vehicles, as well as regulating and enforcing their safe use on the road requires resources and systems that may not be available (e.g., a robust rider licensing system, sufficient political will to enforce laws when authorities recognize that to limit use of such vehicles would prohibit individuals from earning an income).

Countries in Africa have the highest rate of deaths (average of 26.6 deaths per 100,000 population), followed by countries in Southeast Asia (average of 20.7 deaths per 100,000 population) (WHO, 2018). In some countries in Africa, walking is the most common form of transport. In many cases, limited facilities are available for pedestrians to safely negotiate their interactions with motorized vehicles. Therefore, it is not surprising that pedestrians (the least protected road user group) are overrepresented in crashes. By comparison, in Southeast Asia, there are a higher proportion of deaths involving motorcyclists (compared to deaths involving other road users), because motorcycles are more common on the roads and are often the main family vehicle that must carry a large number of family members at one time. In LMICs, more evidence-based, best practice legislation covering key behavioral risk factors is needed. Importantly, it is not only the LMICs that must improve in this regard. Unfortunately, the most recent Global Status Report on Road Safety (2018) reports that only five countries have best practice laws for speeding, driving while impaired by alcohol, the use of helmets for motorcycles, and the use of seat belts and child restraints. This result signals the need for urgent legislative attention in almost all nations.

In the LMIC context, even when legislation is present, a lack of compliance is commonplace, as are lower levels of police enforcement and less public education, advertisement, and mass media campaigns compared to what has traditionally occurred in many HICs. As such, road users may be more likely to engage in on-road risk taking behaviors, possibly because they perceive that they are unlikely to be caught by police or because they are not aware of the risk of harm they face. These behaviors include, but are not limited to, not wearing motorcycle helmets, illegal crossing behaviors or walking on roads (due to a lack of footpaths), speeding, alcohol impaired walking and driving, drug driving, and fatigue (due to walking longer distances rather than operating a vehicle when tired). Driving under the influence of drugs is also likely to be a contributor to road trauma, although prevalence is difficult to determine due to a lack of testing regimes (the 2018 Global Status Report indicates that only 75 countries reported conducting some level of drug testing among fatally injured drivers). In addition to these risky behaviors, and compared to HICs, there is less access to appropriate and timely emergency services and medical treatment. It is well known that individuals have the greatest chance of survival if treatment is received soon after a crash. A lack of availability of adequate first response and specialized medical treatment in LMICs therefore contributes to their disproportionate representation in the global road trauma data.

Road Safety Programs and Initiatives in Low- and Middle-Income Countries

In recognition of the need for dramatic and ongoing improvements in road safety performance of LMICs across a range of issues (e.g., legislation, infrastructure, and enforcement), a number of multicountry programs have been implemented to address key road user behaviors, including the Child Health Initiative, the Botnar Child Road Safety Challenge, and the Bloomberg Philanthropies Initiative for Global Road Safety.

Legislation and Enforcement

Some of the programs listed earlier focus on strengthening risk factor-related legislation to align with best practice. The 2018 Global Status Report shows that some progress is being made on this front. For instance, since the 2015 report, 22 countries had improved laws to align with best practice recommendations for speeding, driving while impaired by alcohol, and the use of helmets, seat belts, and child restraints. However, the report also noted that the presence of best practice laws covering these five behavioral risk factors was more common in HICs than in LMICs. As a signal of just how much improvement is still required in high-, middle-, and low-income contexts, only five countries were reported as having evidence-based legislation on all five of these behavioral risk factors.

If each risk factor is examined separately, we see the same pattern according to country income status (WHO, 2018). For example, in regard to speeding, best practice laws were more common in HICs (50%) compared to MICs (37%) and LICs (13%). For driving while impaired by alcohol, best practice laws were again more common in HICs (58%) compared to MICs (40%) and LICs (2%). For

motorcycle helmet laws, 167 countries that had mandatory helmet laws for motorcyclists, 49 were identified to have best practice seat-belt laws. Even though LICs have a higher prevalence of motorcyclists when compared to HICs, of those 49 countries which met best practice, only 6% were LICs. For seat belts, of the 161 countries which had seat-belt laws, 105 were identified to have best practice seat-belt laws, with only 7% of those 105 countries meeting best practice were LICs. For child restraints, 84 countries had child restraint laws, and 33 of those countries met best practice. However, of those 33 countries, no LIC had child restraint laws that met best practice (WHO, 2018).

Enforcing legislation is a challenge in the LMIC context for a number of reasons. In some cases, no penalty provisions are present which can lead to reluctance on the part of police to enforce the law because they are unable to give a sanction to offenders to deter reoffending. In other cases, laws contain specific information about equipment quality standards (e.g., helmets or child restraints) but no provision is made regarding the agency with responsibility to monitor and enforce such standards. Other challenges can include a lack of political will to enforce the law because of public disapproval as well as limited enforcement resources (e.g., personnel and equipment) and a lack of technical expertise.

Public Education, Advertisement, and Mass Media Campaigns

Since road safety is relatively new on the agenda of many LMICs, the use of road safety awareness raising campaigns, compared to what has occurred in many HICs, is not as extensive. Initiatives have tended to focus on public education in schools as well as limited mass media communications about the risks of a particular behavior. There is a significant need to align communications so as to link awareness raising about the behavioral risk factors with police enforcement activities. Experience from the HIC context has consistently demonstrated that enforcement and education campaigns need to work together to obtain maximum road safety benefits. Given the aforementioned challenges in enforcement in LMICs, coupled with limited resources for road safety advertising, there is need to improve this aspect to improve road safety outcomes.

Summary and Concluding Comments

This article has discussed some of the human-related factors associated with traffic crashes in HICs and LMICs. It has been highlighted that various approaches including road safety strategies, legislation and enforcement, and public education, advertisement, and mass media campaigns have been effectively implemented to reduce crashes associated speeding, driving while impaired by alcohol, not wearing seat belt, driver distraction, and fatigue in HICs. It was also noted that in LMICs, while vulnerable road users are at greater risk on the roads, there is a lack of evidence-based best practice legislation for key behavioral risk factors, less compliance with road rules and limited enforcement, and a lack of public education advertisement strategies compared to HICs.

Knowledge in understanding the factors associated with road crashes is advancing. Epidemiology, which starts with looking at injuries or fatalities on a per capita basis, is an efficient way of identifying sectors, areas, or geographic regions with high risk per 100,000 people. In a road safety context, epidemiology assists with identifying the causes associated with injuries and fatalities resulting from road traffic crashes. Recently, researchers have applied a process referred to as data linkage in order to gain a greater understanding of the injuries associated with traffic crashes (Watson et al., 2015). Data linkage is the process of linking sources of data (e.g., police reported crash data to hospital data) that belong to the same person. By using multiple linked data sources, researchers are able to identify, which individuals are most at risk on our roads. Further, the findings from studies which have used data linkage may assist with providing recommendations on where Government funding and resources should be allocated in order to target those at higher risk of road traffic crashes.

In terms of reducing road traffic crashes associated with human error, continued advancements in vehicle technology may also assist with improving safety. Examples of emerging technologies include intelligent speed assistance to manage speeds, lane departure warning systems, lane keeping assist, and automated emergency braking. Lane departure warning systems are designed to provide an auditory warning to the driver when the vehicle is close to the line marking. The driver is then expected to steer the vehicle away from the line marking. Lane keeping assist, however, is designed so that the vehicle will steer itself back within the line markings (without the need of human intervention). Automated emergency braking will first warn the driver if there is an object in front of the vehicle and the vehicle will then automatically brake if there is no response from the driver. As mentioned previously, the European Union is likely to mandate ISA starting 2022. As such, ISA will assist with improving the safety of vehicles in HICs and, in the longer term, might be a way to improve the safety of vehicle fleets in LMICs if these countries import secondhand vehicles from the European Union.

In summary, it is important that the research continues to assess why road users engage in risky driving behaviors and what can be done to reduce crashes associated with human error. The information gained from decades of trial and error of countermeasures in HICs has led to the development of best practice guidance and understanding of what needs to be done to reduce road crashes and the associated trauma burden. Appropriate legislation, enforcement and education, along with continued advancements in technology will assist with reducing fatalities and injuries in all countries. Importantly, the needs of LMICs might differ from the HIC context in a number of ways (e.g., cultural practices, mix and quality of vehicle fleet, quality of road infrastructure, availability of public transport, and existence of and accessibility to quality postcrash care). However, these differences do not mean that the same countermeasures will not work. Rather, it highlights the need for ongoing efforts to better understand how to continue to promote evidence-based countermeasures in all income contexts.

Relevant Websites

- Towards Zero Foundation. The safe system. Available from: <http://www.towardszerofoundation.org/thesafesystem/>.
- World Health Organization. The top 10 causes of death. Available from: <https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death>.
- NHTSA Traffic Safety Facts. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812580>.
- Global Health Estimates. Available from: https://www.who.int/healthinfo/global_burden_disease/GHE2016_Deaths_2016-country.xls.
- Global Road Safety Partnership. Available from: <https://www.grsproadsafety.org/>.
- United Nations Road Safety Collaboration. Available from: <https://www.who.int/roadsafety/publications/en/>.

References

- Cameron, M., Haworth, N., Oxley, J., Newstead, S., Le, T., 1993. Evaluation of transport accident commission road safety television advertising. Available from: http://www.monash.edu/___data/assets/pdf_file/0019/216514/muarc052.pdf.
- Lewis, I., Forward, S., Elliott, B., Kaye, S-A., Fleiter, J., Watson, B., 2019. Designing and evaluating road safety advertising campaigns. In: Ward, N., Watson, B., Fleming-Vogl, K. (Eds.). Traffic Safety Culture: Definition, Foundation, and Application. Emerald Publishing Limited, UK, pp. 275–295.
- National Highway Traffic Safety Administration (NHTSA), 2017. Countermeasures that work: a highway safety countermeasure guide for state highway safety offices, ninth ed. Available from: https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/812478_countermeasures-that-work-a-highway-safety-countermeasures-guide-9thedition-2017v2_0.pdf.
- Oviedo-Trespalacios, Ó., Haque, M.M., King, M., Washington, S., 2016. Understanding the impacts of mobile phone distraction on driving performance: a systematic review. *Transp. Res. Part C Emerg. Technol.* 72, 360–380.
- Watson, A., Watson, B., Vallmuur, K., 2015. Estimating the under-reporting of road crash injuries to police using multiple linked data collections. *Accid. Anal. Prev.* 83, 18–25.
- World Health Organization, 2017. Save LIVES: a road safety technical package. Available from: https://www.who.int/violence_injury_prevention/publications/road_traffic/save-live-package/en/.
- World Health Organization, 2018. Global status report on road safety 2018. Available from: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/.

Further Reading

- Fleiter, J., Lewis, I., Kaye, S-A., Soole, D., Rakotonirainy, A., Debnath, A., 2016. Public demand for safer speeds: identification of interventions for trial. Research report AP-R507-16. Austroads, Sydney, Australia.
- Naci, H., Chisholm, D., Baker, T.D., 2009. Distribution of road traffic deaths by road user group: a global comparison. *Inj. Prev.* 15, 55–59.
- Tulu, G., Washington, S., King, M., Haque, M., 2013. Why are pedestrian crashes so different in developing countries? A review of relevant factors in relation to their impact in Ethiopia. In: The 2013 Proceedings of the Australasian Transport Research Forum, Brisbane, Australia. Available from: https://atrf.info/papers/2013/2013_tulu_washington_king_haque.pdf.
- World Health Organization, 2010. Global plan for the decade of action for road safety 2011–2020. Available from: https://www.who.int/roadsafety/decade_of_action/plan/en/.
- World Health Organization, 2017. Powered two- and three-wheeler safety: a road safety manual for decision-makers and practitioners. Available from: <http://www.who.int/iris/handle/10665/254759>.

Evacuation Planning and Transportation Resilience

Karl Kim, Department of Urban and Regional Planning, University of Hawaii, Honolulu, HI, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	276
Should I Stay or Go Now?	276
Walk, Bike, Drive, Carpool, Bus, or Ride Hailing Services	277
Friends and Family, Hotel, Campsite, or Public Shelter	277
All Clear, Return Home, and Back to Normalcy	278
New Technologies	278
Training and Education	279
Resilient Evacuation Planning	280
Biography	281
References	281
Further Reading	281

Introduction

Effective, safe evacuation is a critical component of transportation resilience. Resilience is the capacity to absorb shocks, to avoid disruptions, recover functionality and restore services quickly and efficiently. Transportation is an essential component of community resilience and transportation systems must also be resilient and resistant to shocks, interruptions, and disruptions (Kim et al., 2018b). Among the most important aspects of transportation is the safe, efficient movement of people away from harm. Evacuation planning involves anticipating threats and hazards and ensuring that vehicles, facilities, and services are available to move affected populations to safe locations. Perhaps the greatest “disaster” associated with Hurricane Katrina was the failure to safely evacuate thousands of residents when New Orleans was flooded.

In this article, the challenges and opportunities for evacuation planning are described. While evacuations are common, there are limited opportunities and resources available for practicing and preparing for the safe movements of people, pets, and their possessions out of harm’s way. The nature of the hazard or threat will affect the movements and behaviors of populations. There is need to understand decision-making processes as well as the physical systems and social systems for supporting both evacuation and sheltering of at-risk populations (Kim et al., 2018a). There are multiple hazards and threats, different travel modalities, many sources of information and channels for communications, and diverse populations who respond differently to warning and alerts. These factors and the potential for spontaneous actions, reactions, and behaviors complicate the planning and implementation of evacuation. Evacuation planners must be particularly sensitive to the physical, medical, social, and financial needs of evacuees. There is need for more research and better understanding of evacuation behavior and translation of this knowledge into effective plans, strategies, and training programs to build resilience. This article is structured around six aspects of evacuation planning: (1) evacuate or shelter-in-place; (2) evacuation travel modes; (3) origins and destinations for evacuees; (4) all clear and safe return; (5) technologies for detection, alert, and warning; (6) training and education.

Should I Stay or Go Now?

The choice as to whether or not to evacuate or remain in place is a function of information, risk tolerance, decision-making, and resources, including transportation assets. Evacuation involves first and foremost *mobility*, or the ability to move from one location to another, but also *accessibility*, or the use of transportation facilities, vehicles, and infrastructure to escape from harm. The evacuation decision involves the timing, source, credibility, and trustworthiness of information about a threat or hazard (Lindell and Perry, 2012). There must be sufficient time to receive, process, react, and respond to an evacuation order or to a perceived, impending threat. In the absence of official warnings and alert messages, there may be natural cues such as ground shaking, changes in atmospheric pressure or other signals that a potential hazard may strike. It is important to distinguish between short or no-notice events such as earthquakes, explosions, and sudden onset hazards and those such as distant tsunamis or approaching tropical cyclones or flooding for which there is time to prepare, mobilize, and evacuate. Other short notice hazards such as fires, industrial accidents, active shooters, or tornadoes may force evacuees to quickly evaluate threats, formulate escape plans, and move to safety. With some hazards, unfortunately, there is no time to react. A hazard can occur in the middle of the night, when most people are asleep and not aware of the danger. Other hazards such as building collapses or plane crashes can be unexpected events. There may also be rare or unusual events such as the Sarin gas attacks in the Tokyo subway system for which there is little knowledge and

experience with the hazard. For rare, unusual events, evacuees, and emergency managers may not be cognizant of the danger nor how best to respond, further delaying evacuation.

The trigger for the evacuation influences the evacuation order and the decision to evacuate. For weather and flooding events, there may be sufficient warning time so officials can issue an evacuation order. While a “mandatory” evacuation may be issued by a Mayor or emergency management official, people may or may not comply with the order. They may or may not believe that the danger is imminent. They may have experienced other evacuation orders for which harm did not materialize, sometimes referred to as the “cry wolf” effect. Other reasons for not evacuating include the fear of break-ins or burglaries or having pets and animals which need care and cannot not be easily transported. There are also households who may have mobility limitations or health conditions making it difficult to travel even under normal circumstances. Many people also serve as caregivers and either return or stay behind to assist others instead of evacuating.

While people may want to evacuate, they may lack the resources and ability to evacuate (Renne et al., 2011). They may learn of the threat too late or they may not have access to transportation modes to safely and successfully evacuate. They may have physical impairments or medical and mental conditions which prevent them from self-evacuating or taking early protective action. There may be language or communication barriers which affect reception of messages and limit knowledge of when, where, and how to travel. They may feel uncomfortable, unwelcome, or unable to travel because of their economic, social, or political status. Undocumented individuals, refugees, people of color, and those who fear law enforcement, border security and immigration officials may also be reluctant to evacuate, use public evacuation services, and shelters.

Walk, Bike, Drive, Carpool, Bus, or Ride Hailing Services

While the most common modality for evacuation involves the use of personal motorized vehicles, it is important to recognize that most people need to walk to get to their motor vehicles. Travel mode is also generally influenced by trip purpose (work, school, recreational, social, etc.) as well as other factors such as time of day, day of week, weather, and other factors. If evacuees are leaving from home or work or school, they may rely on their normal mode of travel. There has also been increased investment in bike infrastructure (bike lanes, shared bikes, other facilities) and increased awareness of health benefits associated with biking and walking which have promoted alternatives to driving, influencing mode choice both not only in terms of evacuation decision making, but also in terms of the exposure of travelers when evacuation orders are issued. It is important to realize that evacuations are also influenced by mode choice. In addition to trip purpose, the distance of the trip and the availability of fuel may also affect the modality.

Depending on both the hazard and the location, the vehicle may be used to transport other people, possessions, pets, and goods (food, water, clothing, survival supplies, etc.) needed during an evacuation. Households may use multiple vehicles and travel in groups to safety. They may coordinate evacuation plans, with rendezvous at key locations to gather people and goods, and to make collective decisions as to when to leave, where to go, and what or who to take or leave behind.

Unfortunately, there are also cases where individuals have been killed or hurt because they were trying to save their vehicles from harm, typically flooding, or because they were involved in motor vehicle collisions during the evacuation process. These effects can also hamper the movement and flow of traffic as well as create additional emergency response demands.

Carless households, low income populations, and those without extended social or familial support networks (Renne et al., 2011) may be more reliant upon public transport, taxi, paratransit or other shared ride services. Those who decide to evacuate early will have more options, resources, and destinations available to them, while those waiting until the last minute, will face greater competition from others trying to evacuate, as well as limitations imposed because of travel restrictions, closed facilities, and fewer drivers willing to travel into and out of hazard zones.

The role of public transit (Transportation Research Board, 2008) is complicated both with internal movements within an affected jurisdiction and especially if evacuees are transported outside of impacted cities, states, or countries. For a mass evacuation, there may need to be staging areas, pickup and temporary holding areas, transfer stations, and multimodal services (Kim et al., 2013) involving rescue vehicles, marine and air transport, and other vehicles and facilities pressed into service.

Friends and Family, Hotel, Campsite, or Public Shelter

Evacuation planning must also include identification of safe destinations. The most common, ideal location for most evacuees will be to travel to another family member’s or a friend’s home located outside of the hazard zone. This may entail traveling to another state or jurisdiction away from the threatened location. With a large event such as a hurricane, the homes of friends and family may also be impacted, requiring evacuees to travel further away and to seek shelter in a hotel or campsite or public shelter.

The challenge with destination planning is the multiplicity of stakeholders and the mix of public and private decision-making and use of resources. While many decisions and actions are made by individuals or households, not necessarily shared with public authorities, the impacts in terms of roadway traffic, congestion, and travel to safe locations impact public resources including law enforcement, transportation facilities, public works, utilities, availability of fuel, shelter, and other resources. Because schools, parks, and other facilities may be occupied by evacuees, normal operations and services may be impacted. A large number of evacuees from

another jurisdiction may strain physical, personnel, and fiscal resources for the receiving jurisdiction, creating disruptions, challenges, and conflict.

There may be other effects associated with a large mass evacuation including increased demand for hotel rooms, restaurant meals, as well as other goods and services associated with the movement of population to safe zones. It is useful to consider “plausible worst case scenarios” when doing evacuation planning (Kim et al., 2015). Some businesses may thrive while others in the impact zone may suffer losses due to cancellations or trip diversions and reduced spending. These impacts, dislocations, and secondary impacts suggest the need for receiving communities to also engage in evacuation planning.

One of the challenges to evacuation planning is the identification, tracking, and sharing of information regarding evacuees. Address information is useful to logistics planning and coordination of efficient transportation services. This may require the recording of names and other personal identifiers so that families can be reunited or evacuees can be transported via airlines or moved across borders and international boundaries. There are concerns with regard to privacy, identity theft, and exploitation of sensitive personal information. Evacuation planners must secure and protect confidential, sensitive, and private information about evacuees.

Evacuation planners should account for “who is left behind?” Some are left behind because they refuse to believe that the threat is real. Others are willing to accept the risk of harm, knowing that they are in a dangerous situation, but still choose to stay back. Others, such as those in institutions, prisons, and other facilities may not be free to travel and unless provisions have been made for evacuation, they may be left behind. There are caregivers and others providing important services or care for pets and livestock who may also need to stay back. There are others including those without transportation, economic resources, and ability to leave, who are left behind. Many of these people are also frail, require medical and social services, and have special needs. Others including homeless, undocumented people, minorities, nonEnglish speakers, and those who face oppression, discrimination, and stigmatization in society may also be left behind, in harms way.

All Clear, Return Home, and Back to Normalcy

For some hazards, such as man-made threats (active shooter, industrial accident, fire, etc.) the all clear or safe to return home message is a function of when the event has ended. For some weather events such as tornadoes or heavy rainfall, the duration and determination of the “all clear” is supported by detection and warning systems, satellite imagery, and other information systems. For hazards such as earthquakes, volcanoes, landslides, and complex disasters such as earthquake, tsunami, and nuclear disaster in Japan, the “all-clear” safe to return home message may not be so decisive or definitive. Moreover, even though the threat may have ended, there may be other hazards caused by a tornado or hurricane such as downed powerlines or environmental contamination that may create a lingering hazard, compelling authorities to limit access to evacuation zones. Heavily damaged structures can also be a risk to the health and safety of returning evacuees. Areas damaged by flooding or other natural hazards may also be impacted by wastewater spills or debris and require cleanup and restoration which may take time and delay the return to normalcy.

Evacuation due to social conflict may also create a lingering hazard or threat for which evacuees and those involved and the return home or resettlement may take a longer time period. Forced migration may begin with emergency evacuation, but lead to permanent relocation. In some locations with some disasters, the damage to homes and businesses is so extensive that people are forced to leave these communities permanently in search of new homes, jobs, and livelihoods. The “new normal” may mean that the community never returns to the level of population and economic activity that existed prior to the disaster and that evacuees never return home.

System restoration, therefore, is a function of the extent of damage and disruption caused by the harm-causing event. The greater the damaging forces and extent of destruction, the more widespread the evacuation and longer the duration of the evacuation. With small, localized events, households and businesses can return quickly, clean up and resume business as usual. With larger, complex, and devastating events, the return home and restoration of normalcy is delayed, protracted, or reinvented. Transportation plays a crucial role in not just evacuation but also reentry to the community (National Cooperative Highway Research Program, 2009).

For larger disasters, it may be necessary to restrict reentry to emergency repair workers, utilities, and to limit traffic to residents or those with passes or permits so that debris removal and repair of infrastructure can proceed ahead of other activities. Communities may want to impose these limitations to increase safety and security and limit outsiders, tourists, and others who may interfere with the recovery process.

The return home is a key metric for disaster resilience. How quickly, completely, and effectively residents and workers return following an evacuation is also a measure of the community’s resilience. Tracking and following evacuation behavior, therefore, provides evidence of a community’s detection, warning, and alert systems, but also the extent of preparedness and recovery capabilities. Having effective damage assessment, debris management, and restoration of services will hasten the return home and recovery from disasters.

New Technologies

Many new technologies affect evacuation behavior and evacuation planning. A useful framework for understanding the interactions between warning systems, evacuee decision making, and outcomes is the protective action decision model, which systematically considers diverse sources and channels for information, the credibility or “trustworthiness” of the information, how receivers and

evacuees process information and utilize it in decision making to formulate intended and actual actions and behaviors (Lindell and Perry, 2012) which result in outcomes (safe evacuation, sheltering from harm, or negative consequences such as death, injury, or other losses). The advent of social media, improved communications technologies, and widespread adoption of mobile computing and communications devices alongside diverse platforms and systems for disseminating, sharing, posting, hosting, and receiving information has both improved and complicated warning and detection processes. These technologies can also support the development of information systems for ensuring the safe, efficient movements of evacuees and their reunification with families and return home.

In the past, conventional media (radio, television, and print) were the primary sources of information about hazards and threats. While these sources still exist, there are many other ways in which people receive information about hazards and evacuation orders. The absence of a definitive source is also complicated by the fact that different federal, state, and local authorities are involved in the issuance of warnings and threat levels. While National Oceanic Atmospheric Administration (NOAA) is largely responsible for weather related events and while the United States Geological Survey (USGS) is responsible for earthquakes and volcanoes, there are shared responsibilities between national and regional and local centers. While tsunamis are most often triggered by earthquakes, the warnings are handled by NOAA. Flooding requires information about rainfall (NOAA) but also stream gauge data (USGS). While there have been improvements in sensors, satellites, and communications systems, there are ongoing challenges with the accuracy, timeliness, and extent of forecasts. While the prediction of storm tracks has improved, forecasts of wind speeds, storm surge, and localized damage needs additional development and testing. Advances in data integration and assimilation of information from multiple sources, ensemble forecasting, big data systems and application of machine learning and other advanced computational models support both early detection and effective response and recovery.

There is need for integration between scientific research and modeling and translation of hazards data into actionable information to support evacuation planning. In addition to the hazard and threat levels, it is important to have detailed information about the affected populations including not just their exposure to threats, but also their willingness and capabilities with respect to evacuation and sheltering. The integration of traffic monitoring systems with emergency management can help with the management of traffic during evacuations, but also provides vital information as to the location and movements of motorists in the hazard zones useful to understanding risk and exposure to disruption and the planning and design of mitigation strategies for reducing harm from natural and man-made hazards.

The improvement of tools for assessing and visualizing hazards such as (CAMEO, HAZUS, Sea Level Rise Viewer, COMMIT and other packages) needs to be accompanied with development of localized data sets and training of key personnel to integrate risk and vulnerability assessment and understanding of transportation evacuation models across highway, urban, and nonmotorized evacuation scenarios. It remains challenging to keep models, data, and personnel up-to-date with changes in local conditions and available technologies, system and software versions, and other updates.

In addition to hazard models, there are many evacuation models (Hardy et al., 2009; Murray-Tuite and Wolshon, 2013) with subroutines and packages built into or added onto systems such as TransCAD, TANSIMS, or CUBE, with different macro- (ETIS, HEADSUP, PC DYNEV), micro- (CORSIM, VISSIM, DYANMIT, Paramics) and meso-scale (DYNASMART-P), Integration 2.0, EMME/2, OREMS) models with varying input requirements that typically include not just origins and destinations but also vehicle occupancies, loading factors, roadway capacities, and directionality resulting in outputs such as travel times, speed, delay, volume, congestion, capacity utilization and emergency management outcomes such as clearance times, and cumulative proportions of those evacuees who have left the danger zone. Greater dissemination of tools and datasets, model results, and examples of how these advances have been used in evacuation plans along with evaluation of actual evacuation plans, orders, and cases would be helpful to the development of training and education.

Training and Education

The key to improved evacuation planning is training and education. There is a need for more evacuation simulations, exercises and drills, and evaluation of these efforts. Such efforts are costly and require financial resources, personnel commitments, and specialized knowledge, skills, and equipment for modeling, mapping, and making effective evacuation plans. A commitment to training and education requires strong leadership and the involvement of not just management, but also operations and field staff within and across agencies responsible for transportation, law enforcement, public facilities, and health and human services. Training and education should be coordinated not just across federal, state, and local agencies, but also with private and nongovernmental organizations involved in evacuation and sheltering. More attention to special needs populations is especially needed.

While a core element of evacuation planning involves the training and education of engineers, it is also important to include others involved in the planning, management, and operation of the facilities, vehicles, and services associated with evacuation (Matherly et al., 2013). Often those involved with large-scale special events such as hosting a Superbowl or Olympics or major convention or international gathering, have experience in multi-modal planning, managing crowds, accommodating special needs, or unusually high levels of demand for transport and other services. Those familiar with travel demand modeling with an understanding of origins and destinations, mode choice, and network constraints have a foundation for the development of evacuation plans. Moreover, those who run transit systems, paratransit operations, manage parking facilities, ferries, airports, and commuter facilities also understand the movements, capacities, and flows of critical components of transportation supply and demand. While there is need to adjust time frames and to account for complexities caused by changing conditions and the

reactions of travelers, those already in the transportation industry have the knowledge, skills, and professionalism to support successful evacuation.

The National Disaster Preparedness Training Center funded by the US Department of Homeland Security and the Federal Emergency Management Agency has developed evacuation training for state and local governments. In addition to reviewing case studies on evacuation, the course delves into the key operational strategies for making use of real-time traffic data, signage, pavement markings, barricades, coordination of traffic controls, roadway configurations, contraflow, intersection management, and other adaptive strategies for managing roadway, transit, and other assets for improving emergency evacuation. In addition to highway operations, there is also need to coordinate bus and paratransit operations and manage other transportation assets (school buses, shuttle, rail, and commuter services). Other demand management strategies include school closings, timed/staged government employee release, and encouragement of shelter-in-place strategies where appropriate. The course includes modules on estimation of the evacuee population to include not just residents and employees, but also tourists, transient populations, institutionalized populations, and other at-risk groups. Needs of diverse travelers may vary widely. Key sources of information include paratransit operators, faith-based organizations, social service providers, and the visitor industry. The course recommends dividing evacuees into three groups: (1) vehicle based, self-evacuees; (2) assisted self-evacuees (transit, taxi, rideshare, etc.); and (3) assisted evacuees (medical, paratransit, institutionalized, schools, etc.). A key element of the course involves lessons from Katrina, Ivan, Rita and other large storms with regard to messaging, information sharing, coordination between emergency management, transportation agencies, law enforcement and other key stakeholders. Other training courses have been developed in Australia (focused on wildfires) and in other countries for specific hazards such as tsunamis and volcanoes in Indonesia.

Among the most developed evacuation plans are those required by the US Nuclear Regulatory Commission in communities with nuclear power plants. These plans need to be updated periodically or when the affected population or travel conditions change significantly and include evacuation planning for schools, special facilities, and vulnerable populations. In addition to detailed evacuation time estimates, the modeling procedures also require estimates of “shadow evacuations” for those who do not need to evacuate but do so anyway, adding to network traffic, and the effects of warnings on the behavior of vulnerable populations. There are valuable lessons to be learned by examination of not just communities that have conducted recent evacuations, but also from those communities which have developed and practiced evacuation plans. There is value to studying “after action” reports lessons learned, and evaluations of exercises, tabletops, and drills carried out by state and local governments. More systematic evaluation and sharing of knowledge is needed.

Resilient Evacuation Planning

Resilient evacuation planning begins with understanding of needs, capabilities, decision making, and the social context in which evacuees receive, process, and act to take protective actions. At the core of decision making is understanding of risk, that is the likelihood of events, and the consequences of those events. There is both the *actual risk* or threat created by a particular hazard and *perceived risks* based on the information, experience, and influences from media, friends, and family, and others with whom the evacuee is connected. Evacuees may choose to ignore or accept crucial information. They may decide to delay decision making. They may follow or heed the advice of others who may or may not have full information nor possess robust decision-making capabilities. These factors and influences complicate the issuance of warnings and alerts, evacuation orders, and the efforts of emergency managers and those involved in executing evacuation plans.

As described in this article, the transportation system is comprised of many different moving parts. There are different modalities, technologies, vehicles, facilities and networks for movement and communications which influence understanding of how, when, where, and with whom evacuation occurs. There are also different hazards and threats with varying degrees of impact, physical forces, duration, extent, and disruption. Many of these hazards can also cripple transportation assets, impacting the evacuation efforts. There are also many different agencies, organizations, and assets which need to be deployed at the management and operational levels to ensure safe, efficient evacuation. Greater emphasis on the physical, medical, social, and economic needs of evacuees must be included in evacuation plans. Investments in training and the development of operational procedures supports not just knowing what to do and how to accomplish it, but also the formation of relationships and trust between emergency managers, transport operators, law enforcement, social service providers, and others who need to work collaboratively in evacuation planning.

With longer term threats, such as climate change and sea level rise and increased frequency and intensity of storms, flooding, wildfires, and other natural hazards, there are opportunities now to invest in the design, planning, engineering, financing, and construction of mitigation strategies to either lessen the need for evacuation or support positive outcomes if evacuation orders are implemented. These investments include more sensors and equipment for detecting hazards, improved communications and warning systems so that vulnerable populations receive earlier evacuation orders, and improving the sheltering-in-place and increasing neighborhood refuges and safe zones which are easily accessible to evacuees. Tsunami evacuation buildings, elevated safe havens from flooding, and structures designed to withstand various hazards need to be designed, built, and designated within hazard prone communities. By engaging communities in the planning and design of local evacuation strategies, coupled with mitigation and adaptation projects and effective training programs planners and engineers can increase awareness, capabilities, and resilience among those most at risk of harm. Communities must fully engage in equal opportunity evacuation planning so that *no one is left behind*.

Biography

Karl Kim is Professor of Urban and Regional Planning at the University of Hawaii at Manoa, and Director of the graduate program in Disaster Management and Humanitarian Assistance. He studies transportation, cities, and resilience. He served as the Chief Academic Officer (Vice Chancellor for Academic Affairs) overseeing tenure and promotion, strategic planning, international programs, and program review for the University of Hawaii. He is Executive Director of the National Disaster Preparedness Training Center (ndptc.hawaii.edu), authorized by the US Congress to develop and deliver FEMA certified training courses for underserved, at-risk communities on natural hazards, mitigation, and urban planning. The Center has trained more than 50,000 first responders, emergency managers, and leaders in over 400 different communities across the world. Dr Kim has served as the Chairman of the National Domestic Preparedness Consortium (ndpc.us). He has published many papers on disaster management, transportation, and urban planning. He is Editor-in-Chief of *Transportation Research Interdisciplinary Perspectives* (Elsevier) and is editor of a 10-volume series on disaster risk reduction and resilience (Routledge). He has been a Fulbright Scholar to Korea and the Russian Far East. Dr Kim was educated at Brown University and the Massachusetts Institute of Technology.

References

- Hardy, M., Wunderlich, K., Bunchard, J., 2009. Structural Modeling and Simulation Analyses for Evacuation Planning and Operations. U.S. Department of Transportation. Research and Innovation Technology Administration, Washington, DC.
- Lindell, M., Perry, R., 2012. The protective action decision model: theoretical modifications and additional evidence. *Risk Anal.* 32 (4), 616–631.
- Kim, K., et al., 2013. Learning from crisis: transit evacuation in Honolulu, Hawaii after tsunami warnings. *Transp. Res. Rec.* 2376, 56–62.
- Kim, K., Pant, P., Yamashita, E., 2018a. Integrating travel demand modeling and flood hazard risk analysis for evacuation and sheltering. *Int. J. Dis. Risk Reduct.* 31, 1177–1186, doi:10.1016/j.ijdr.2017.10.025.
- Kim, K., Francis, O., Yamashita, E., 2018b. Learning to build resilience in transportation systems. *Transp. Res. Rec.* 1–13, doi:10.1177/0361198118786622.
- Kim, K., Pant, P., Yamashita, E., 2015. Evacuation planning for plausible worst case inundation scenarios in Honolulu, Hawaii. *J. Emer. Manag.* 13 (2), 93–108, doi:10.5055/jem.2015.0223.
- Matherly, D., Mobley, J., Wolshon, B., Renne, J., Thomas, R., & Nichols, E., 2013. A Transportation Guide for All-Hazards Emergency Evacuation, Strategic Highway Research Program, Report 740, ISBN: 978-0-309-25901-9, Washington DC, Transportation Research Board.
- Murray-Tuite, P., Wolshon, B., 2013. Evacuation transportation modeling: an overview of research, development, and practice. *Transp. Res. C* 27, 25–45.
- National Cooperative Highway Research Program, 2009. Transportation's Role in Emergency Evacuation and Reentry: A Synthesis of Highway Practice. Transportation Research Board, Washington, DC ISBN 978-0-309-098311.
- Renne, J., Sanchez, T., Litman, T., 2011. Carless and special needs evacuation planning: a literature review. *J. Plan. Literat.* 26 (4), 420–431.
- Transportation Research Board, 2008. The role of transit in emergency evacuation. Transportation Research Board Special Report 294. Washington, DC, ISBN 978-0-309-11333-5.

Further Reading

- Lindell, M., Murray-Tuite, P., Wolshon, B., Baker S E., 2008. Large-Scale Evacuation: The Analysis, Modeling, and Management of Emergency Relocation from Hazardous Areas. Taylor and Francis, New York.
- Tierney, K., 2014. The Social Roots of Risk: Producing Disasters, Promoting Resilience. Stanford University Press, Palo Alto ISBN 9780804791397.

Exposure: A Critical Factor in Risk Analysis

Frank Gross PhD, PE, VHB, Raleigh, NC, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	282
Per Capita	282
Licensed Drivers	282
Registered Vehicles	283
Vehicle-Kilometers Traveled	283
Nonmotorized Road Users	283
Annual Average Daily Traffic (AADT)	284
Using Exposure Responsibly	284
Current Practice for Estimating Crash Risk	286
Future of Estimating Crash Risk	288
End Note	288
Biography	288
Relevant Websites	288
References	289

Introduction

Exposure is a critical factor in risk analysis and one of the primary predictors of crashes and other harmful events. Want to avoid a shark bite? Stay out of the ocean. Similarly, if you want to avoid a car crash, stay off the road. Obviously, if there were no motor vehicles on the road, then there would be no motor vehicle crashes. Since the chance of eliminating vehicles on the roadway is about the same as keeping swimmers out of the ocean, there is a need to understand how to account for exposure in estimating crash risk.

There are numerous measures of vehicle exposure such as traffic volume, segment length, and vehicle-kilometers traveled (VKT). There are also various measures of nonmotorized exposure such as pedestrian volumes, bicycle volumes, and time or distance traveled on a given facility type. The appropriate measure of exposure depends on the question at hand and the subject of interest. The following sections present different measures of exposure used to account for differences in crash risk and the applicable questions that could be answered related to traffic safety.

Per Capita

A very basic measure of exposure is the population of a given geographic area. Per capita means per population, which might be expressed per thousand or per million people. Per capita is a good measure of exposure when assessing risk in terms of overall public health. If the focus is on reducing the number of people injured or killed in motor vehicle crashes, then this is the measure to use. Per capita is not a reliable measure to predict the safety performance of a specific location because it does not account for the many factors that contribute to changes in risk such as the number of road users, type of road users, type of facilities, and distance traveled by mode and facility type.

The following are potential questions answered using the population as the measure of exposure:

- What is the difference in crash risk by population group?
- What is the difference in crash risk by geographic area?

Licensed Drivers

Another basic measure of exposure is the number of licensed drivers. This is similar to the population-based measure of exposure (per capita) but limits the population to those with a driver license. Similar to per capita, the number of licensed drivers is not a reliable measure to predict the safety performance of individual locations because it also does not account for the many factors that contribute to changes in risk such as the number of road users, type of road users, type of facilities, and distance traveled by mode and facility type. It is also not the most reliable measure of exposure when looking at this from a public health perspective. For example, the number of licensed drivers does nothing to account for the number of school-aged children who walk to school or the types of roadways the licensed drivers travel.

Registered Vehicles

Another spin on population-based measures of exposure is the vehicle population. Specifically, the number of registered vehicles is a common measure of exposure for computing and comparing crash rates over time. Similar to the number of licensed drivers, the number of registered vehicles is not a reliable measure of crash risk because it does not account for the distance traveled or the types of facilities traveled by those vehicles. For example, a car collector may have a 1956 Ford Thunderbird in their collection and drive it not more than 100 km per year, most of which are on a closed course as part of a car show; however, this vehicle would count toward the number of registered vehicles. Further, there are more registered vehicles in the United States than there are licensed drivers (225 million licensed drivers and 273 million registered vehicles in 2017; [USDOT, 2018](#)) and it is not possible to drive two vehicles at the same time.

Vehicle-Kilometers Traveled

VKT [or vehicle-miles traveled (VMT)] is applicable to crash risk related to motor vehicles. This measure of exposure accounts for the number of vehicles on the roadway and the distance traveled by those vehicles.

In the United States, data for this measure come from the Federal Highway Administration (FHWA) Highway Statistics ([USDOT, 2019](#)). The reports are based on individual state reports on traffic data counts collected through permanent automatic traffic recorders on public roadways. Data for urbanized areas are available from the FHWA Highway Statistics Series ([USDOT, 2018](#)). Some countries collect this type of data in slightly different ways. In Sweden, there are estimates of VKT, but most official data are vehicle-axle-pairs because the average axle-numbers per vehicle are not known well enough for site-level analysis.

We have already established that crash risk increases as traffic volume increases. But this is a general relationship and the change in crash risk differs for different types of vehicles, types of roads, and groups of drivers. VKT could be stratified in a number of ways such as by vehicle type (e.g., passenger cars vs. large trucks), facility type (e.g., urban vs. rural), or driver type (e.g., younger vs. older drivers) to compare crash risk.

The following are potential questions answered using VKT as the measure of exposure:

- What is the difference in crash risk by vehicle type?
- What is the difference in crash risk by facility type?
- What is the difference in crash risk by driver age?

When comparing crash risk among vehicle types, it is important to account for the number of kilometers traveled by those vehicles. For example, there are many more crashes involving passenger cars than large trucks; however, the VKT for passenger cars is also much greater than the VKT for large trucks. In 2016 in the United States, there were approximately 21,000 passenger cars (excluding motorcycles and light trucks) involved in fatal crashes and approximately 4200 large trucks involved in fatal crashes. While the number of passenger vehicles involved in fatal crashes is approximately 5 times greater than the number of large trucks involved in fatal crashes, the involvement rate of passenger cars, using VMT as the measure of exposure, is nearly identical (1.45 passenger cars involved in fatal crashes per 100 million VMT compared to 1.46 large trucks involved in fatal crashes per 100 million VMT).

Similarly, when comparing crash risk among facility types, the number of kilometers by facility type is not likely to change dramatically from one year to the next; however, crashes, injuries, and fatalities change from year to year and there is a need to understand how much of this is due to changes in exposure. In 2016 in the United States, there were 4445 fatal crashes on interstates and 4022 fatal crashes on local roads. While the number of fatal crashes is slightly higher on interstates, the fatal crash rate, using VMT as the measure of exposure, is much higher on local roads (0.55 fatal crashes per 100 million VMT on interstates compared to 0.93 fatal crashes per 100 million VMT on local roads).

When comparing VKT from year to year, it is also important to understand if there are changes in VKT by age group. For example, younger (less experienced) drivers tend to have a higher crash risk compared to older (more experienced) drivers. Therefore, it is conceivable that the overall crash risk for the system could decrease even if VKT increased or remained the same from one year to the next. This might be the case if VKT decreased for younger drivers and increased for more experienced drivers. To capture the relationship between crash risk and exposure, this scenario would require information by age group. Specifically, there would be a need to know the VKT by age group and the number of crashes by age group. With this information, one could model the relationship and predict the overall change in crash risk given a predicted change in travel by age group.

Nonmotorized Road Users

Nonmotorized users may include any person on foot, walking, running, jogging, hiking, sitting, lying down, or any person on personal conveyances such as bicycles, roller skates, inline skates, skateboards, baby strollers, scooters, Segway-style devices, wheelchairs, and scooters for those with disabilities. The crash risk for these users depends on factors such as the number of users by mode and distance traveled by user group and facility type. The exposure could also differentiate the distance traveled along

roadway and distance traveled when crossing the roadway. These numbers are much more difficult to track than motor vehicles because in the United States the typical traffic counters that track the number of vehicles on the road do not collect numbers for nonmotorized travel, and we also do not register these vehicle types as we do with motor vehicles. In some countries, bicycle and pedestrian flows are collected on a large number of facilities but in the United States they are typically counted only at a few intersections in central parts of cities, and those counts are done manually by hand. Agencies are starting to use newer video analysis and mobile applications for smartphones are beginning to track some information about pedestrian and bicycle exposure. Also, household surveys in some countries capture estimates of walking and biking distances; however, this is only for a sample of the country's population, only for one day of the year, and most countries do not execute the surveys every year. More research is needed to obtain accurate estimates. Once these estimates are available, it will be useful to generate and compare crash rates by various demographics (age and gender), locations (intersection and midblock), area types (urban and rural), and other categories. It will also be useful to model the statistical relationship between crash risk and exposure. For example, the United Kingdom and Sweden have developed models to estimate the number of crashes based on the number of vehicles crossing the crosswalk and pedestrian daily flow. A hypothetical example of this relationship is shown in Eq. (1).

$$\text{Pedestrian crashes} = \text{vehicles crossing crosswalk}^{0.5} \times \text{pedestrian daily flow}^{0.5} \quad (1)$$

Annual Average Daily Traffic (AADT)

The previous measures of exposure help to compare crash risk among general groups (e.g., populations, regions, and facility types), but how would you measure the crash risk for a single site (e.g., intersection or roadway segment)? Research has shown that traffic volume is one of the single most important predictors of crash risk for a given location (AASHTO, 2010; Hauer, 1994). Again, no traffic, no crashes. For roadway segments such as tangents, curves, and ramps, the annual average daily traffic (AADT) and segment length are common measures of exposure. Crash risk increases as traffic volume and segment length increase. The AADT is typically the two-way traffic volume, but could include directional, hourly, or sub-hourly traffic volume. For intersections, the AADT on the major and minor road, or total entering vehicles, are common measures of exposure. When using the total entering volume, the crash risk is typically expressed as crashes per million entering vehicles; however, the crash risk may change for different proportions of traffic on the major and minor road even when the total entering vehicles is the same. As such, it is often more reliable to account for differences in AADT on the major and minor road.

Using Exposure Responsibly

It is important to account for exposure, but it is just as important to do so responsibly. Traditionally, analysts have used crash rates to account for differences in exposure among sites. Crash rate is the ratio of crash frequency to exposure as shown in Eq. (2), where the crash frequency of interest could be the count of crashes by type or severity, and the measure of exposure could be population, registered vehicles, VKT, nonmotorized users, or AADT as described above.

$$\text{Crash rate} = \frac{\text{Crash frequency}}{\text{Measure of exposure}} \quad (2)$$

Crash rates implicitly assume a linear relationship between crash frequency and the measure of exposure; however, many studies have shown that the relationship between crashes and traffic volume is often nonlinear, and this relationship depends on the facility and site type (Hauer, 1997). While crash rates may be useful for high-level comparisons of crash risk among populations, vehicle types, and facility types, crash rates are not a reliable measure to rank sites for potential investment. The remainder of this section explains the limitations of using crash rate to compare the safety performance among sites as illustrated through an example.

Table 1 provides data for three hypothetical sites that an agency is considering for further study and potential investment to improve safety. Site 1 has an average of six crashes per year with an AADT of 2000 vehicles per day and corresponding crash rate of 0.0030 (crashes per AADT). Site 2 has an average of eight crashes per year with an AADT of 5000 vehicles per day and corresponding crash rate of 0.0016 (crashes per AADT). Site 3 has an average of 11 crashes per year with an AADT of 10,000 vehicles per day and corresponding crash rate of 0.0011 (crashes per AADT).

Table 1 Ranking of hypothetical sites by crash frequency and crash rate

Sites	Crash frequency	Traffic volume (exposure)	Crash rate (crashes/exposure)	Rank by frequency	Rank by rate
1	6	2000	0.0030	3	1
2	8	5000	0.0016	2	2
3	11	10,000	0.0011	1	3

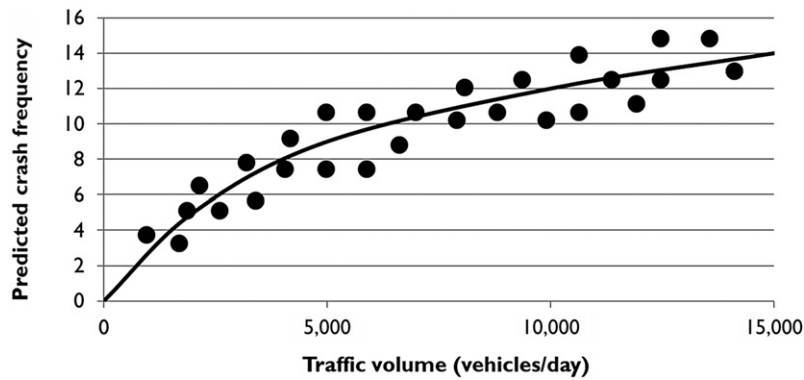


Figure 1 Hypothetical SPF relating crash frequency and traffic volume.

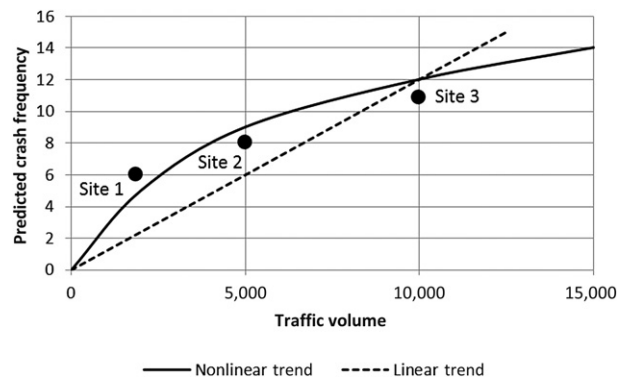


Figure 2 Example of site selection using exposure in different ways.

Consider the three hypothetical sites from [Table 1](#) and two performance measures to rank the sites for further investigation: crash frequency and crash rate. Using crash frequency, the sites are ranked 3, 2, and 1 where site 3 is the highest priority as shown in [Table 1](#). Using crash rate, the sites are ranked 1, 2, and 3 where site 1 is the highest priority as shown in [Table 1](#). Clearly, one can make an argument to account for exposure when comparing safety performance because one would expect more crashes at sites with higher traffic volumes than sites with lower traffic volumes. If the analyst does not account for differences in traffic volume among sites, then they may incorrectly identify sites with higher traffic volumes as sites with higher potential for safety improvement. On the contrary, a site with a single crash and relatively low exposure would have a relatively high, and potentially misleading, crash rate.

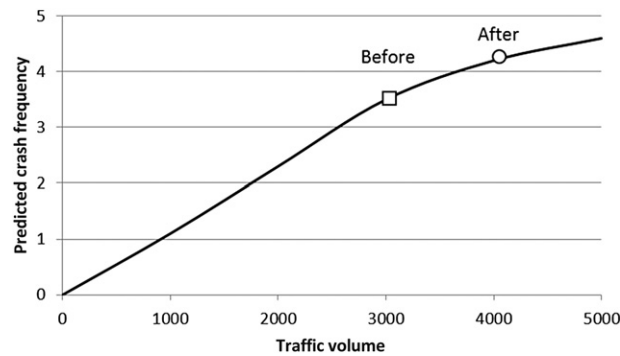
The safety literature has shown that nonlinear relationships are more appropriate than linear relationships such as crash rates to account for differences in traffic volume among sites when comparing safety performance. Specifically, safety performance functions (SPFs) are a more reliable method to account for differences in traffic volume among sites because they reflect the nonlinear relationship between crash frequency and traffic volume. The SPF is an equation, representing a best-fit model that relates annual observed crashes to the annual traffic volume for a group of sites with similar attributes. [Fig. 1](#) is a hypothetical SPF where the points represent observed crashes at specific traffic volumes for individual sites, and the solid line represents the best-fit model (i.e., the SPF). Given an SPF and the traffic volume (i.e., exposure) for a specific location, one could predict the crashes for that location. For example, at an AADT of 5000 vehicles per day, this SPF predicts approximately nine crashes.

Now, let's revisit the data from [Table 1](#) and include a linear threshold (average crash rate) and nonlinear threshold (i.e., SPF) in the mix. [Fig. 2](#) plots the hypothetical data from [Table 1](#) and illustrates the difference between a linear and nonlinear relationship between crash frequency and traffic volume. In [Fig. 2](#), the dashed line is the average crash rate for similar sites, representing a linear relationship between crash frequency and traffic volume. The solid line is the SPF for similar sites, representing a nonlinear relationship between crash frequency and traffic volume. As performance measures, these lines provide a threshold representing the average crashes for a given traffic volume. Sites above the line represent sites with the greatest potential for safety improvement and sites below the line represent sites performing relatively well with respect to other similar sites with similar traffic volumes. Note that sites below the line may still present opportunities for safety improvement.

Using the average crash rate as the threshold, assuming a linear relationship, the sites are ranked 1, 2, and 3. It is apparent that sites 1 and 2 have potential for improvement because the actual crash frequency is greater than the predicted crash frequency for that level of exposure. Site 1 is the highest priority because it shows the greatest potential for improvement. Using the SPF as the threshold,

Table 2 Before–after analysis of hypothetical curve improvement project

Performance measure	Before treatment	After treatment	Safety improvement?
Average crash frequency (crashes/year)	3.5	4.2	No
Average traffic volume treatment (vehicles/day)	3000	4000	
Crash rate (crashes/million-vehicles)	3.2	2.9	Yes

**Figure 3** Hypothetical SPF for predicting crashes on curves.

assuming a nonlinear relationship, the sites are again ranked 1, 2, and 3 where site 1 remains the highest priority; however, site 1 is the only site that shows a clear need for improvement. For sites 2 and 3, the actual crash frequency is lower than the predicted average crash frequency, indicating these sites are performing relatively well compared to similar sites with similar exposure.

If the actual relationship between crashes and traffic volume is nonlinear, and the analyst assumed a linear trend as the threshold to represent the average crashes for a given traffic volume, then they would incorrectly identify any sites between the SPF and dashed line (site 2 in this case) as priority sites for further investigation. It is important to understand that different performance measures can produce different priority rankings and the more reliable performance measures (i.e., those based on SPFs) can lead to the identification of sites with greater opportunity for safety improvement and better investment of safety improvement funds. When it is not possible to use SPF-based performance measures to identify sites with potential for safety improvement, then crash frequency can serve as a reasonable alternative (Srinivasan et al., 2016b). If crash rate is used to identify sites with potential for safety improvement, then it is important to incorporate the average crash rate as a threshold and use performance measures such as critical crash rate.

As another illustration of the potential differences in results from different performance measures, consider a before–after safety evaluation (Srinivasan et al., 2016a). Agencies regularly conduct before–after evaluations to estimate the safety impacts of completed projects. For example, let's assume an agency installed advance curve warning signs and chevrons to address curve-related crashes. Three years after implementation, the agency evaluates the effectiveness of the chevrons to determine if the treatment improved safety. Table 2 presents information gathered for the analysis. Using crash frequency as the performance measure, the agency would conclude there was no safety improvement because the number of crashes increased from 3.5 per year to 4.2 per year. Using crash rate as the performance measure, the agency would conclude there was a safety improvement because the crash rate decreased from 3.2 per million-vehicles to 2.9 per million-vehicles. Now, let's assume the agency had access to the SPF in Fig. 3 that represents similar curves, allowing them to predict crashes based on traffic volume. Plotting the actual crash frequency before and after on the same scale as the SPF, it is apparent that the actual crashes before and after treatment are similar to the predicted crashes based on the traffic volumes before and after. In this case, the agency would conclude the treatment did not impact safety performance (i.e., no change relative to other similar sites with similar traffic volumes). Given this information, the agency might decide to try something different at this location to improve safety.

Current Practice for Estimating Crash Risk

The current state of the practice for estimating crash risk is to use negative binomial regression models to explain the relationship between number of crashes and exposure. This is reflected in the predictive methods in the Highway Safety Manual and related research (AASHTO, 2010). For segments, exposure includes segment length and AADT. For intersections, exposure includes traffic volumes on the major and minor roads or total entering vehicles. The following equations form the basis of SPFs. Agencies use these functional forms with data from their jurisdiction to estimate the coefficients through regression modeling. While it is possible to calibrate existing SPFs from another agency, it is typically more reliable to develop agency-specific SPFs.

Eq. (3) shows the functional form for a typical crash prediction model for roadway segments. The predicted crashes ($N_{\text{predicted}}$) are estimated based on segment length (L) and AADT. Eq. (4) is the same equation transformed using the natural log function. This transformation allows for the estimation of the coefficients (a and b) using linear regression. In these equations, " a " is the intercept and " b " is the regression coefficient for AADT.

$$N_{\text{predicted}} = L \times e^a \times \text{AADT}^b \quad (3)$$

$$\ln(N_{\text{predicted}}) = \ln(L) + a + b \times \ln(\text{AADT}) \quad (4)$$

Eqs. (5) and (6) show a slight variation of the functional form and transformed equation for a crash prediction model for roadway segments. In Eqs. (3) and (4), segment length is treated as an offset where the predicted number of crashes is linearly related to the length of the segment. A coefficient " c " is added to Eqs. (5) and (6) to allow for a nonlinear relationship between the predicted crashes and segment length.

$$N_{\text{predicted}} = L^c \times e^a \times \text{AADT}^b \quad (5)$$

$$\ln(N_{\text{predicted}}) = c \times \ln(L) + a + b \times \ln(\text{AADT}) \quad (6)$$

Eqs. (7) and (8) show the typical functional form and transformed equation for a crash prediction model for intersections. The predicted crashes ($N_{\text{predicted}}$) are estimated based on the major road AADT (AADT_Major) and the minor road AADT (AADT_Minor).

$$N_{\text{predicted}} = e^a \times \text{AADT_Major}^b \times \text{AADT_Minor}^c \quad (7)$$

$$\ln(N_{\text{predicted}}) = a + b \times \ln(\text{AADT_Major}) + c \times \ln(\text{AADT_Minor}) \quad (8)$$

Finally, Eqs. (9) and (10) show a slight variation of Eqs. (7) and (8) where the predicted crashes ($N_{\text{predicted}}$) are estimated based on the total entering traffic volume (AADT_Total) rather than separating the AADT on the major and minor road.

$$N_{\text{predicted}} = e^a \times \text{AADT_Total}^b \quad (9)$$

$$\ln(N_{\text{predicted}}) = a + b \times \ln(\text{AADT_Total}) \quad (10)$$

The following is an illustrative example of why it may be preferable to account for both the major and minor road AADT in the model (i.e., the proportion of traffic on the major and minor road can change the estimated safety performance depending on the facility type). Consider two rural, four-legged intersections with similar geometry and the same total entering volume, say 10,000 vehicles per day. The difference is the split in traffic on the major and minor road. Intersection 1 has 9000 vehicles per day on the major road and 1000 vehicles per day on the minor road. Intersection 2 has 5000 vehicles per day on the major road and 5000 vehicles per day on the minor road. Eq. (11) represents the SPF for a rural, four-legged, two-way stop-controlled intersection from the United States Highway Safety Manual. Eq. (12) represents the SPF for a rural, four-legged signalized intersection from the United States Highway Safety Manual.

$$N_{\text{predicted}} = \exp(-8.56 + 0.60 \times \ln(\text{AADT_Major}) + 0.61 \times \ln(\text{AADT_Minor})) \quad (11)$$

$$N_{\text{predicted}} = \exp(-5.13 + 0.60 \times \ln(\text{AADT_Major}) + 0.20 \times \ln(\text{AADT_Minor})) \quad (12)$$

Table 3 shows the predicted crashes for each of the scenarios after plugging in the values for major and minor road traffic volume. With two-way stop-controlled intersection, the predicted crashes for Intersection 1 are 2.40 crashes per year and the predicted crashes for Intersection 2 are 3.72 crashes per year. It is apparent that the proportion of traffic on the major and minor road impacts the safety performance of an intersection, and, in this case, Intersection 1 is predicted to perform better than Intersection 2 with respect to safety. With signal control, the predicted crashes for Intersection 1 are 4.35 crashes per year and the predicted crashes for Intersection 2 are 4.05 crashes per year. Again, the proportion of traffic on the major and minor road impacts the safety performance of an intersection; however, in this case, Intersection 2 is predicted to perform slightly better than Intersection 1 with respect to safety.

Table 3 Predicted crashes for intersections with different proportions of major and minor volumes

Intersections	Major road volume	Minor road volume	Predicted crashes with two-way stop-control	Predicted crashes with traffic signal
Intersection 1	9000	1000	2.40	4.35
Intersection 2	5000	5000	3.72	4.05

Future of Estimating Crash Risk

The current state of the practice for estimating crash risk is based on annual measures of exposure (e.g., population, VKT, and AADT). As technology advancements allow for the collection and analysis of more disaggregate exposure data, the state of the practice will likely move toward the estimation of risk at a more disaggregate level. However, this will depend on the availability of data and the question at hand. For example, if the question at hand is related to the risk of different populations, and the resulting answer would guide decisions related to education programs that would not change more than once per year, then it would not be necessary to use disaggregate data such as sub-hourly volumes by population to estimate crash risk. However, if the question was related to the real-time risk of a roadway segment or intersection, and the answer would guide decisions related to the real-time operation of the site, then sub-hourly or real-time exposure data would be quite useful. For example, consider a freeway corridor with ramp metering in place. If an agency has access to real-time exposure data (e.g., traffic volumes on the freeway and on-ramp), then they could monitor the change in risk and activate the ramp metering when the risk exceeded a certain threshold.

End Note

When calculating and comparing crash risk, it is important to account for exposure and to do so properly. Traditional methods account for exposure through linear crash rates (simply dividing the number of crashes by the exposure). More reliable methods establish a threshold and allow for the typical nonlinear relationship between crashes and exposure. It is also important to use an appropriate measure of exposure and to consider the data required to answer the question at hand. As such, it is useful to stop and ask yourself questions such as:

- What measure of exposure should be used for the risk analysis/assessment?
- What level of detail is required in the exposure data to answer the question at hand (e.g., annual, daily, hourly, sub-hourly volumes, etc.)?

Biography



Dr. Frank Gross, PE, is a principal at VHB with over 15 years of diverse transportation research and engineering experience. As a researcher, he specializes in the development of national guidance and tools to support data-driven decision-making, including research to support the Highway Safety Manual. As a highway safety engineer, he implements and promotes the use of these national tools, supporting state and local agencies with their programs, focusing most recently on the Highway Safety Improvement Program. His national exposure keeps him informed of the latest tools and techniques. His experience at the state and local level keeps him informed of the common challenges and opportunities to overcome challenges in transportation management efforts.

Relevant Websites

FHWA Roadway Safety Data & Analysis Toolbox.

This toolbox contains resources related to collecting, managing, and analyzing safety data, including three of the four resources listed in the "Further Reading" section.

Available from: <https://safety.fhwa.dot.gov/rsdp/>.

Highway Statistics Series.

The Highway Statistics Series consists of annual reports containing analyzed statistical information on motor fuel, motor vehicle registrations, driver licenses, highway user taxation, highway mileage, travel, and highway finance.

Available from: <https://www.fhwa.dot.gov/policyinformation/statistics.cfm>.

Traffic Safety Facts.

Traffic Safety Facts presents descriptive crash statistics to help understand national and statewide trends. The crash statistics included in Traffic Safety Facts are generated from two data systems: (1) the Fatality Analysis Reporting System (FARS) and (2) the General Estimates System (GES). FARS contains data on traffic crashes in which someone was killed. GES contains data from a nationally representative sample of police-reported crashes of all severities, including those that result in death, injury, or property damage.

Available from: <https://cdan.nhtsa.gov/tsftables/tsfar.htm>.

Travel Monitoring: Traffic Volume Trends.

Traffic Volume Trends is a monthly report based on hourly traffic count data reported by the States. These data are collected at approximately 5000 continuous traffic counting locations nationwide and are used to estimate the percent change in traffic for the current month compared with the same month in the previous year. Estimates are readjusted annually to match

the vehicle kilometers of travel from the Highway Performance Monitoring System and are continually updated with additional data.
Available from: https://www.fhwa.dot.gov/policyinformation/travel_monitoring/tvt.cfm.

References

- American Association of State Highway and Transportation Officials (AASHTO), 2010. Highway Safety Manual, first ed. AASHTO, Washington, DC.
- Hauer, E., 1994. On exposure and accident rate. *Traffic Eng. Control* 36 (3), 134–138.
This paper explores the relationship between crashes and exposure. It also demonstrates the potential shortcomings of assuming a linear crash rate if the relationship between crashes and exposure is nonlinear.
- Hauer, E., 1997. *Observational Before–After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Elsevier Science, Ltd., Oxford, United Kingdom.
This book demonstrates how to properly account for exposure in estimating the safety effects of safety improvements. It also explains that the relationship between expected crash frequency and traffic flow is often nonlinear.
- Srinivasan, R., Gross, F., Bahar, G., 2016a. Reliability of safety management methods: safety effectiveness evaluation, Report No. FHWA-SA-16-040. Federal Highway Administration, Washington, DC.
This guide describes various methods and the latest tools to support safety effectiveness evaluation. The objectives of this guide are to (1) raise awareness of more reliable methods, and (2) demonstrate through examples the value of more reliable methods in safety effectiveness evaluation. This guide compares more reliable evaluation methods to traditional methods that are more susceptible to bias and may result in less reliable estimates and less effective decisions. Readers will understand the value of and be prepared to select more reliable methods in safety effectiveness evaluation.
Available from: <https://safety.fhwa.dot.gov/rsdp/toolbox-content.aspx?toolid=149>.
- Srinivasan, R., Gross, F., Lan, B., Bahar, G., 2016b. Reliability of safety management methods: network screening, Report No. FHWA-SA-16-037. Federal Highway Administration, Washington, DC.
This guide describes various methods and the latest tools to support network screening. The objectives of this guide are to (1) raise awareness of more reliable methods, and (2) demonstrate through examples the value of more reliable methods in network screening. This guide compares more reliable crash-based performance measures to traditional measures that are more susceptible to bias and may result in less reliable results and less effective decisions. Readers will understand the value of and be prepared to select more reliable performance measures in network screening.
Available from: <https://safety.fhwa.dot.gov/rsdp/toolbox-content.aspx?toolid=192>.
- USDOT, 2018. Highway Statistics Series. Federal Highway Administration, Office of Highway Policy Information. Available from: <https://www.fhwa.dot.gov/policyinformation/statistics.cfm>.
- USDOT, 2019. Travel Monitoring: Traffic Volume Trends. Federal Highway Administration, Office of Highway Policy Information. Available from: https://www.fhwa.dot.gov/policy-information/travel_monitoring/tvt.cfm.

Passenger Ferry Vessels and Cruise Ships: Safety and Security

Wayne K. Talley, Old Dominion University, Norfolk, VA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	290
Ferry Vessels	290
Types of Ferry Vessels	290
Ferry Vessel Safety	291
Ferry Vessel Accidents in the US Waters	291
Ferry Vessel Security	292
Cruise Ships	293
Cruise-Ship Sizes	293
Cruise Ship Safety	293
Cruise Ship Accidents in the US Waters	293
Cruise Ship Security	294
References	295
Further Reading	295

Introduction

In this article, types, safety, and security of ferry vessels and cruise ships are discussed.

Ferry Vessels

A ferry vessel is a passenger vessel used to transport passengers and sometimes vehicles (e.g., automobiles of ferry passengers) and cargoes of shippers from ferry marine terminals as origin to ferry marine terminals as destination (<https://en.wikipedia.org/wiki/Ferry>). Shipper cargoes are stored aboard ferry vessels: (1) individually, (2) in packages and containers, and (3) in the land freight transportation vehicles used in the transportation of the cargoes from pick-up sites to ferry marine terminals. Ferry vessels that transport vehicles are described as roll-on/roll-off ferries. The transportation of passengers by ferry vessels in urban areas, where the vessels follow fixed-water routes and fixed-time schedules in the areas, is a type of urban transit passenger transportation service (Talley, 1983). This type of ferry vessel has the distinction in being larger in size than urban transit passenger land vehicles by having the capacity to transport up to 2500 passengers per trip (Talley, 1983, p. 293).

Types of Ferry Vessels

The types of ferry vessels used in the provision of transportation services will depend upon the (1) physical characteristics of the water routes over which ferry-vessel transportation services are provided (e.g., lengths in miles and the extent to which water currents and waves exist), (2) weather conditions of the water routes (e.g., the extent of rain and snow precipitations and cold versus warm temperatures on the routes), and (3) whether vehicles and cargoes are transported by ferry vessels.

Hovercraft ferry vessels in the 1960s and 1970s were built to transport passengers and their automobiles. The SR.N4, the largest hovercraft vessel built, allows automobiles to enter its central section via ramps at both its bow and stern and was used to transport automobiles between England and France. The hovercraft is supported by a cushion of air with boundaries defined by a rubber skirt in which it sits (Wang and Mcowan, 2000). Its lift is provided by a fan system that blows air into the cushion and its propulsion is provided by engines.

Hydrofoil ferry vessels have the advantage of relatively high cruising speeds. The two main types are the fully submerged hydrofoil vessel and the surface piercing hydrofoil vessel. Hydrofoil ferry vessels replaced hovercraft ferry vessels on English Channel routes that competed with the Eurotunnel and Eurostar passenger trains using the English Channel tunnel.

Catamaran ferry vessels revolutionized ferry vessel services. A catamaran ferry vessel is a twin-hull vessel with slender hulls (or pontoons), having the advantage of stability and being less susceptible to adverse sea conditions versus hydrofoil ferry vessels. Since 1990, hovercraft and hydrofoil ferry vessel services have been replaced with Catamaran ferry vessel services. The Stena HSS class catamaran (once the largest catamaran vessel in the world) is waterjet powered and is operated by Stena Line between the United Kingdom and Ireland, accommodating 1500 passengers and 375 passenger cars. High-speed (exceeding 30 knots) ferries are technically known as high-speed crafts (HSCs). The catamaran is the most commonly used HSC by passenger ferry operators as opposed to using, for example, the hydrofoil HSC ferry (Wang and Mcowan, 2000).

Other types of ferry vessels include the (1) *monohull* ferry that has greater cargo-carrying capacity than the catamaran; (2) small waterplane area twin hull (SWATH) ferry that has a bulbous shape that allows it to rise in the water as the speed of the vessel increases, providing a dynamic lift; (3) surface effect ship (SES) ferry that combines aspects of the catamaran and the hovercraft to produce a vessel with the advantages of both vessels and the disadvantages of neither; (4) *double-ended* ferry that has interchangeable bows and sterns that allow it to shuttle back and forth between marine terminals without having to turn around; (5) *train* ferry that is fitted with railway tracks and front and/or rear doors that provide access to wharves for the transportation of rail cars; (6) *foot* ferry is a small ferry that is used to transport foot passengers and cyclists across small waterways (such as rivers); and (7) *RoPax* fast ferry that has relatively large garage intake and passenger capacities, with conventional diesel propulsion, generating vessel speeds exceeding 25 knots and in 1995, these were used to provide fast ferry service between Greece and Italy. High-speed ferries (hovercrafts and catamarans) can operate at speeds exceeding 30 knots. Ferry carriers that operate large ferries around the world include: Sealink (operating between the United Kingdom, Ireland, and Europe), Washington State Ferries (operating in the US state of Washington), and British Columbia Ferry Corporation (operating in the Canadian province of British Columbia).

Ferry Vessel Safety

"Vessel safety consists of the safety of a vessel and the safety in the operation of a vessel" (Costas book, second ed., 2010, p. 533). The safety of a vessel is concerned with whether a vessel adheres to construction, design, or technical standards and is therefore seaworthy. A vessel may become unsafe during its operation, if it incurs an accident—an unexpected and undesirable event.

"Although ferry incidents exist in North America and Europe, transportation by ferry is one of the safest forms of travel in these parts of the world" (Leach, 2014), but not necessarily true in other parts of the world. The safety of ferry transportation in developing countries is a safety concern.

In April, 2014, the MV Sewol ferry capsized traveling from Incheon (South Korea) to Jeju (South Korea)—300 of the 500 or two-thirds of the passengers died (many of whom were students at Danwon High School near Seoul, South Korea) versus one-third of the crew (<https://www.ship.technology.com/features/featuretaking-action-on-ferry-safety-4379066/>). Some of the passenger deaths were attributed to passenger confusion about the ferry's emergency procedures (a fault of the ferry's crew). Crew and other passenger deaths were attributed to the size of the ferry's cargo load (more than twice the ferry's designated limit) that was not tied down properly (a fault of the ferry's operator). A general safety concern for ferries is human error by the crew while ferries are underway, for example, inattentiveness, poor judgment, and negligence.

Vessel instability is a safety concern for roll-on/roll-off ferries, since "the vessels have large holes that allow for the loading (roll-on) and the unloading (roll-off) of vehicles, thus precluding vertical watertight bulk heads that are standard features for most commercial vessels" (Yip et al., 2015, p. 112). If the bow and stern doors of roll-on/roll-off ferries are breached while the ferries are underway, water will enter the ferries, possibly resulting in the ferries listing and then sinking, contributing to passenger and crew deaths and lost vehicles and cargo.

Ferry Vessel Accidents in the US Waters

Commercial vessel accidents that occur in the US waters and investigated by the US Coast Guard are classified into various types of accidents, that is, (1) fire/explosion (where a fire or explosion is the initiating event in causing a vessel accident); (2) collision (occurs when a vessel strikes or is struck by another vessel on the water surface); (3) allision (occurs when a vessel strikes a stationary object on the water surface); (4) grounding (occurs when a vessel comes in contact with the waterway's bottom or a bottom obstacle); (5) material-failure (typically involves equipment failure onboard a vessel); and (6) maneuverability (occurs when the maneuverability of a ferry vessel is negatively affected). Ferry vessel accidents are further discussed in Loeb et al. (1994).

The number of commercial ferry vessel accidents of a given type as a percent of the total number of commercial ferry vessel accidents (1795) that occurred in the US waters during the 11-year time period 2001–08 and investigated by the US Coast are presented in Table 1. The material-failure type ferry vessel accident has the largest percentage of total ferry vessel accidents for a given type of ferry vessel accident at 44.6%, followed by the maneuverability ferry vessel accident at 30.0% and the allision ferry vessel accident at 6.6%.

Table 1 Ferry vessel accidents in the US Waters 2001–08

Type of accident	Percent of ferry vessel accidents by type
Fire	1.2
Collision	1.4
Allision	6.6
Grounding	4.7
Material failure	44.6
Maneuverability	30.0
Other	11.5

The average number of crew injuries and passenger injuries per ferry vessel accident for commercial ferry vessel accidents that occurred in the US waters during the 11-year time period 2001–08 and investigated by the US Coast Guard were 0.028 and 0.191, respectively (Yip et al., 2015, p. 114). The former study also found that the number of ferry-vessel passenger injuries are positively related to the number of crew injuries, suggesting that enhanced and effective safety training for ferry-vessel crew members can have a positive impact on reducing the number of ferry-vessel accident passenger injuries.

Ferry Vessel Security

Actions taken by the US Coast Guard to secure the operations of ferries in the US waters include conducting boat escorts of ferries and stationing armed Coast Guard personnel in key locations aboard ferries (to ensure that ferry operators maintain control of their ferries). The US Customs and Border Protection (CBP) also inspects foreign-flag ferry vessels that arrive in or bound for the US port).

The security of BC ferries of British Columbia (Canada) is promoted by the Royal Canadian Mounted Police, Transport Canada, the Ministry of Transport and other stakeholders by (1) providing identification cards for ferry vessel employees, (2) designating restricted and prohibited ferry control areas, (3) random canine screening of ferry vehicles and baggage, (4) closed-circuit television monitoring, (5) security awareness training for all employees, and (6) waterside screening.

Ferry terminals are equipped with ramps that enable vehicles to be loaded and unloaded simultaneously from two decks of a ferry. Figs. 1 and 2 are images of ferry vessels.

In the study by Talley (2002), determinants of the number of fatal and nonfatal injuries in ferry vessel accidents that occurred in the US waters during the 11-year time period 1981–91 are investigated. The study concludes that the expected number of fatal injuries is 3.35% higher for fire/explosion than for collision, material/equipment failure, or grounding ferry vessel accidents. The expected number of nonfatal injuries is 3.60% and 4.46% higher for collision and fire/explosion ferry vessel accidents, respectively than for material/equipment failure or grounding accidents. The results suggest that “policies that reduce fire/explosion ferry-vessel accidents are likely to be efficacious in reducing both fatal and non-fatal ferry injuries” (Talley, 2002, p. 337).



Figure 1 Ferry vessel (underway). Source: <https://tugster.wordpress.com>.



Figure 2 Ferry vessel (in port). Source: <https://www.bing.com/images/search?q=ferry+vessels.thehindubusinessline.com>

Cruise Ships

A cruise ship is a passenger ship used by passengers for pleasure voyages that include the voyages themselves, ship amenities and sometimes calling at different ports as part of the passengers' experience (https://en.wikipedia.org/wiki/Cruise_ship).

When cruise ship passengers disembark from a cruise ship at the same cruise marine terminal at which they boarded the cruise ship, the cruise ship voyage is not a transportation voyage, since a distinction cannot be made between a passenger's origin cruise marine terminal and a passenger's destination cruise marine terminal. If the origin and destination cruise marine terminals differ, the cruise ship voyage is a passenger transportation voyage.

Cruise-Ship Sizes

The demand for cruise-ship voyages and the sizes of cruise ships have increased over time. The world's largest cruise ships (greater than or equal to 200,000 gross tonnage in size) in service or under construction are found in **Table 2**. Cruise ships exhibit economies of ship size at sea, that is, as cruise ships increase in size, cruise ship's transportation cost per passenger-transported declines—where the number of passengers transported is the maximum passenger carrying capacity of the ship.

Cruise Ship Safety

Cruise ships are subject to safety inspections that address fire safety, crew training, and proper functioning of all ship safety systems and lifesaving equipment. Cruise ships are required to have lifeboats (or rafts) and life preservers for every person on board. Safety regulations are monitored under the US and maritime laws. The US Coast Guard conducts safety inspections of cruise ships that sail from the US ports (Landry and Kling, 2018; <https://landryking.com>).

Cruise Ship Accidents in the US Waters

The number of ocean cruise vessel accidents of a given type as a percent of the total number of ocean cruise vessel accidents (263) that occurred in the US waters during the 11-year time period 2001–08 and investigated by the US Coast Guard are found in

Table 2 World's largest cruise ships

<i>Ship</i>	<i>Cruise line</i>	<i>Year placed in service</i>	<i>Gross tonnage^a</i>	<i>Maximum passenger capacity</i>
Harmony of the Seas	Royal Caribbean International	2016	226,963	6,360
Allure of the Seas	Royal Caribbean International	2010	225,282	6,296
Oasis of the Seas	Royal Caribbean International	2009	225,282	6,296
Quantum of the Seas	Royal Caribbean International	2014	168,666	4,905
Anthem of the Seas	Royal Caribbean International	2015	168,666	4,905
Ovation of the Seas	Royal Caribbean International	2016	168,666	4,905
Norwegian Escape	Norwegian Cruise Line	2015	165,157	N/A
Liberty of the Seas	Royal Caribbean International	2007	155,889	4,375
Norwegian Epic	Norwegian Cruise Line	2010	155,873	5,183
Freedom of the Seas	Royal Caribbean International	2006	154,407	4,375
Independence of the Seas	Royal Caribbean International	2008	154,407	4,375

^aGross tonnage $\geq$ 200,000; UC, Under construction.

Source: en.wikipedia.org/wiki/List_of_the_world%27s_largest_cruise_ships#In_service.

Table 3 Cruise ship accidents in the US Waters 2001–08

<i>Accident type</i>	<i>Percent of ocean cruise ship accidents</i>	<i>Percent of river cruise ship accidents</i>
Fire	7.6	1.2
Collision	5.7	3.6
Allision	4.6	15.7
Grounding	6.1	7.2
Material failure	26.2	32.5
Maneuverability	18.6	21.7
Other	31.2	18.1

Source: Yip, T.L., Jin, D., Talley, W.K., 2015. Determinants of injuries in passenger vessel accidents. *Accid. Anal. Prev.* 82, 112–117.

Table 3. The number of river cruise vessel accidents of a given type as a percent of the total number of river cruise vessel accidents (83) that occurred in the US waters during the 11-year time period 2001–08 and investigated by the US Coast Guard are also found in **Table 3**.

Material-failure type accidents represent the largest percent (at 26.2%) of ocean cruise ship accidents that occurred in the US waters and investigated by the US Coast Guard, followed by maneuverability accidents at 18.6% and then fire accidents at 7.6%. As for ocean cruise ship accidents, material-failure and maneuverability river cruise ship accidents (**Table 3**) also represent the largest and second largest percent, respectively, of river cruise ship accidents that occurred in the US waters and investigated by the US Coast Guard, but at the higher percentages of 32.5% and 21.7%, respectively. The percentages of allision and grounding river cruise ship accidents at 15.7% and 7.2%, respectively, also exceed those for ocean cruise ship accidents but the percentages of fire and collision river cruise accidents are less than those for ocean cruise accidents.

The number of ferry-vessel (ocean-cruise ship, river-cruise ship) passenger injuries in ferry-vessel (ocean-cruise ship, river-cruise ship) accidents occurring in the US waters (investigated by the US Coast Guard) is positively related to the number of ferry-vessel (ocean-cruise ship, river-cruise ship) crew injuries occurring in ferry-vessel (ocean-cruise ship, river-cruise ship) accidents (Yip et al., 2015, p. 116)—suggesting that enhanced safety training of ferry-vessel (ocean-cruise ship, river-cruise ship) crew members can have a positive impact on reducing the number of passenger injuries in ferry-vessel (ocean-cruise ship, river cruise ship) accidents.

Cruise Ship Security

Actions taken to secure the operations of cruise ships in the US waters include (1) pre-cruise, (2) boarding, (3) preparing-to-sail, and (4) during-the-cruise actions. Pre-cruise actions include the screening by the US authorities, for example, by the US Homeland Security and Immigration, of the names of all passengers that will be boarding cruise ships in the US waters. During the boarding, all check bags are scanned and passengers must pass through an airport-style detector. With respect to preparing-to-sail, the International Convention of Safety of Life at Sea (SOLAS) requires that cruise ships hold a muster drill within 24 h of embarkation that consists of passengers familiarizes themselves with the muster station for their particular ship level in case of an emergency. During the cruise, the ship captain and staff oversee the security operations for the ship. The Cruise Vessels Security and Safety Act requires all the US vessels to report crimes and missing persons immediately and collect evidence (following strict procedures).

References

- Leach, A., 2014. Taking action on ferry safety. Available from: <https://www.ship.technology.com/author/adam/>.
- Loeb, P., Talley, W.K., Zlatoper, T., 1994. Causes and Deterrents of Transportation Accidents: An Analysis by Mode. Quorum Books, Westport, Connecticut.
- Talley, W.K., 1983. Introduction to Transportation. South-Western Publishing Company, OH, Cincinnati.
- Talley, W.K., 2002. The safety of ferries: an accident injury perspective. *Marit. Policy Manage.* 29 (3), 331–338.
- Wang, J., Mcowan, S., 2000. Fast passenger ferries and their future. *Marit. Policy Manage.* 27 (3), 231–251.
- Yip, T.L., Jin, D., Talley, W.K., 2015. Determinants of injuries in passenger vessel accidents. *Accid. Anal. Prev.* 82 (2015), 112–117.

Further Reading

- Corbett, J.J., Farrell, A., 2002. Mitigating air pollution impacts of passenger ferries. *Transp. Res. D Transp. Environ.* 7, 197–211.
- Dragovic, B., Skuric, M., Kofjac, D., 2014. A proposed simulation-based operational policy for cruise ships in the port of kotor. *Marit. Policy Manage.* 41 (6), 560–588.
- Lawson, C.T., Weisbrod, R.E., 2005. Ferry transport: the realm of responsibility for ferry disasters in developing countries. *J. Public Transp.* 8 (4), 17–33.
- Talley, W.K., 2010. Ferry services. In: Button, K., Nijkamp, P., Vega, H. (Eds.), *Transportation Dictionary*. Edward Elgar, Cheltenham, United Kingdom, pp. 158–159.
- Talley, W.K., Jin, D., Kite-Powell, H., 2008a. Determinants of the damage cost and injury severity of ferry vessel accidents. *WMU J. Marit. Aff.* 7, 175–188.
- Talley, W.K., Jin, D., Kite-Powell, H., 2008b. Determinants of the severity of cruise vessel accidents. *Transp. Res. D Transp. Environ.* 13, 86–94.
- Vaggelas, G.K., Pallas, A.A., 2010. Passenger ports: services provision and their benefits. *Marit. Policy Manage.* 37 (1), 73–89.
- Veronneau, S., Roy, J., 2012. Cruise lines and passengers. In: Talley, W.K. (Ed.), *The Blackwell Companion to Maritime Economics*. Wiley-Blackwell Publishing, Oxford, United Kingdom, pp. 138–160.
- Wergeland, T., 2012. Ferry passenger markets. In: Talley, W.K. (Ed.), *The Blackwell Companion to Maritime Economics*. Wiley-Blackwell Publishing, Oxford, United Kingdom, pp. 161–183.

Fuel Economy Standards: Impacts on Safety

Kenneth T. Gillingham, Stephanie M. Weber, Yale University, New Haven, CT, United States

© 2021 Elsevier Ltd. All rights reserved.

Glossary	296
Introduction	296
Mechanisms	298
Safety Effects	298
Vehicle Weight	298
Vehicle Size	300
VMT Rebound Effect	300
Vehicle Age	301
Conclusion	301
See Also	302
References	302

Glossary

CAFE Corporate Average Fuel Economy Standards, the fuel economy standards implemented in the United States in 1975. Historically, the policy involved meeting separate standards for cars and light trucks, but since 2011, the standards have also been based on each vehicle's footprint.

EPA Environmental Protection Agency, one of the two organizations involved in setting and enforcing CAFE standards in the United States (the second is the Department of Transportation).

Footprint A measure of vehicle size approximately equal to the product of wheelbase (the distance between the front and back axles) and the average front and back track widths. The US standards were changed to differ by vehicle footprint in an effort to remove the incentive for manufacturers to meet CAFE standards by making their vehicles smaller. In other countries that use attribute-based regulations, the standards often differ by weight.

Gruenspecht effect The tendency for fuel economy standards to defer scrappage for used vehicles, increasing the average age of vehicles in the entire fleet. The standards raise the price of new vehicles, especially those that are largest and least efficient. With higher new vehicle prices, some potential buyers will not buy a new vehicle, but instead either hold on to their current vehicles longer or purchase used vehicles instead. This in turn also raises the price of used vehicles, which means that some used vehicle owners will find it worthwhile to repair or otherwise hold on to used vehicles for longer, raising the overall age of vehicles in the fleet.

Light trucks One of two categories of passenger vehicles regulated by fuel economy standards, where the other is cars. Light trucks are a subset of trucks with a gross vehicle weight rating of 8500 lbs or below. They are built on distinct frames from cars. Common categories of light trucks are vans, pickups, and sport utility vehicles (SUVs). They are particularly popular in the United States.

NHTSA National Highway Transportation Safety Administration, which is under the US Department of Transportation and is the other of the two organizations involved in setting and enforcing CAFE standards in the United States (with EPA).

Rebound effect The response to improved energy efficiency that tends to reduce the energy savings associated with the efficiency improvement. In this context, the rebound effect most commonly discussed refers specifically to the additional percentage increase in driving that occurs in response to increased efficiency of vehicles (and, therefore, lower fuel costs per mile of driving).

VMT Vehicle miles traveled.

Introduction

Following the 1973 oil crisis, policymakers around the world sought to reduce foreign oil dependence and began to consider regulating the fuel economy of new vehicles. First implemented in the United States in 1975 as Corporate Average Fuel Economy (CAFE) standards, similar policies were put in place soon after in Japan and later, in Australia, Canada, China, the European Union, and South Korea (Anderson et al., 2011). The policies encourage automakers to improve fuel efficiency by setting average fuel economy or fuel intensity levels (in miles per gallon or liters per kilometer) that their fleet or a subset of their fleet must meet

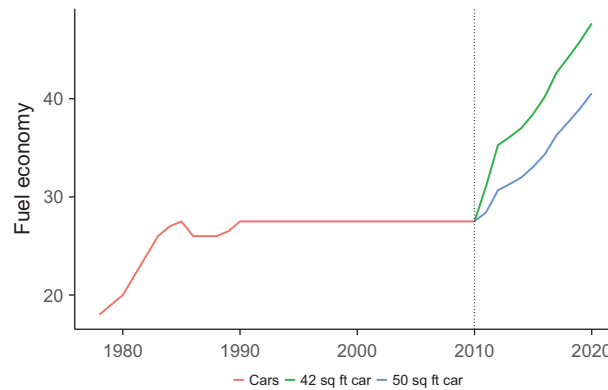


Figure 1 The red line indicates the level of the CAFE standard for cars (in miles per gallon). The dotted vertical line indicates the end of an aggregate standard for cars and the introduction of footprint-based standards. After 2010, we present two illustrative standards: the one in green shows the standard for a 42-square foot vehicle (in the 10th percentile of footprint size) and the one in blue shows it for a 50-square foot vehicle (approximately the 90th percentile).

each year. In countries where the standard is nonvoluntary, enforcement involves fines based on the extent of the manufacturer's deviation from the standard and the scale of their sales. Another commonality shared by the policies is that all have been dogged by concerns that the improved fuel economy will come at the expense of safety. Because the US policy has both the longest history and the largest body of research, the following chapter devotes the most attention to the effects of CAFE standards on safety. But we should note that many of the key findings emerging from the literature are broadly relevant to fuel economy rules worldwide.

Historically, CAFE standards were distinct for cars and light trucks, a category of vehicles that includes vans, pickups, and sport utility vehicles (SUVs). The intent behind this distinction was to differentiate between vehicles with different uses and users, though as light trucks have become increasingly popular as consumer vehicles, this feature of the policy has become more significant. For cars, the fuel economy standard increased gradually from 1978 to 1990, after which it was held constant. The 2007 Energy Independence and Security Act (EISA) introduced new and higher standards beginning with model year 2011 based on vehicle footprint. Vehicle footprint is a measure of vehicle size approximately equal to the product of wheelbase (the distance between the front and back axles) and the average front and back track widths. Larger footprints face less stringent standards. In 2018, the Trump administration proposed reductions in the 2021–25 standards. Fig. 1 shows the standards for cars since 1978.

The light truck standard has also changed over time. Between 1979 and 1982, separate two-wheel drive and four-wheel drive standards were in place, but from 1982 to 1991, firms could choose between separate standards or a combined standard. A single standard was in place from 1992 to 1996, after which it remained constant until 2005. The 2007 EISA further raised light truck standards while also incorporating footprint-based distinctions. The historical trajectory of light truck standards is shown in Fig. 2.

Other countries have designed their standards differently. In Japan, diesel and gasoline vehicles face separate standards that vary according to weight class. Both the EU and Chinese regulations do not distinguish between fuel types or vehicle classes (car versus light truck) but are differentiated by weight.

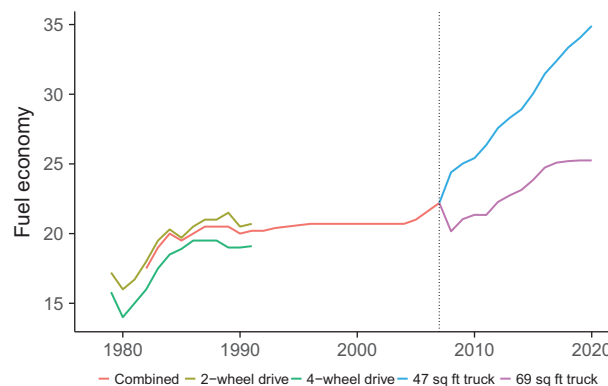


Figure 2 Initially, light-truck standards were different for two-wheel drive (shown in yellow) and four-wheel drive (shown in green), though starting in 1982, producers were allowed to choose between the separate standards and a combined standard, which eventually was the only standard. The dotted vertical line indicates the end of an aggregate standard for light trucks and the introduction of footprint-based standards. After 2010, we present two illustrative standards: the one in blue shows the standard for a 47-square foot vehicle (in the 10th percentile of footprint size) and the one in purple shows it for a 69-square foot vehicle (the 94th percentile).

Mechanisms

In order to understand how fuel economy rules can affect vehicle safety, it is helpful to first examine how automakers can respond to the regulation. A useful framework is to examine responses based on the time frame over which they can occur.

In the shortest term, vehicle producers can reduce the price of more efficient models and increase the price of less efficient models, a practice called mix shifting. This incentivizes purchasers to choose more efficient vehicles, bringing up the average fuel economy without directly changing the attributes of individual vehicles. If existing efficient vehicles tend to be smaller and lighter, this will reduce the average size and weight of vehicles sold. The downside of this approach for automakers is that it may reduce sales of their most profitable vehicles.

Over the medium run, producers can make adjustments to vehicle attributes. Moderate vehicle redesigns occur every 4–5 years, and fuel economy can be affected by trading off attributes along the existing technological frontier (Klier and Linn, 2012). For instance, it is well known that there is a tradeoff between horsepower and vehicle efficiency; reducing horsepower or making other engine modifications that reduce cylinders used at low speeds can improve vehicle efficiency without fundamental innovations. Likewise, weight and efficiency tend to be inversely correlated. Using lighter materials, removing heavy components, and redesigning the interior cabin can all reduce weight and thereby improve fuel economy. Much of the discussion around CAFE's effect on safety has long involved concerns about the consequences of this vehicle down-weighting. On the other hand, weight and footprint can be manipulated under weight- or footprint-based standards (Whitefoot and Skerlos, 2012; Gillingham, 2013; Ito and Sallee, 2018). Both direct attribute adjustment to meet a given standard and the manipulation of attributes to fall under different standards can have size and safety implications.

Finally, over the long term, automakers can develop a new technology. According to Klier and Linn (2012), this occurs on a decadal time horizon and involves changes such as implementing hybrid powertrain technology. With these kinds of innovations, it may be possible to improve efficiency while maintaining vehicle weight and horsepower, although it is possible that by directing innovation toward fuel economy, automakers may put less effort toward innovation in attributes that affect safety (positively or negatively).

Changes along each of these margins have occurred after the implementation of fuel economy standards. Goldberg (1998) estimated that CAFE was reducing the price of domestically produced compact and subcompact cars while increasing the price of all other classes. It is also well documented that vehicle weight dropped following the introduction of CAFE standards. For example, Anderson and Auffhammer (2014) note that vehicles became approximately 1000 lbs lighter between 1975, when the standards were introduced, and 1980. In one of the seminal papers on CAFE and safety, Crandall et al. (1989) estimate that 50%–80% of the observed weight reduction immediately following CAFE was attributable to the standards. Using a much longer time series, Bento et al. (2017) find that in aggregate, a 1 mpg increase in the standard reduced mean vehicle weight by 32 lbs, though the magnitude of the effect differs by initial weight. Based on a more theoretical approach, Jacobsen (2013) identifies the shifts in fleet composition implied by the stringency of the standard for different vehicle categories and finds that CAFE imposes shadow taxes on heavier vehicle categories within the car and light truck classes. Finally, technological innovation has occurred since the onset of CAFE standards. Klier and Linn (2016) find that following changes in the US and European standards, the rate of technology adoption among vehicle manufacturers increased.

Safety Effects

This discussion focuses on the effect of fuel economy standards on safety through four primary mechanisms: modification of vehicle attributes and, in particular, weight and size; changes in driving behavior in response to improved fuel economy; and changes in purchasing behavior that affect the age of vehicles on the road.

Changes in vehicle attributes can have different effects depending on the type of collision and the fatalities or injuries considered. It is useful to distinguish collisions into (at least) three types: single-vehicle collisions; multi-vehicle collisions; and collisions between vehicles and pedestrians, cyclists, and motorcyclists. In these latter two cases, one may measure the effect of a vehicle trait on safety for the occupants of that vehicle versus on safety for the individual(s) with whom the vehicle collides. Some research further distinguishes between multi-vehicle collision geometry (e.g., head-on collisions versus front-to-side collisions), though we generally abstract from that level of detail.

Vehicle Weight

The effect of vehicle weight on safety depends both on the physics of momentum and collisions and on human choices. There is no fundamental physical reason for lower weight to increase fatalities in the aggregate. The most significant factors in safety conditional on a collision are the magnitude of deceleration experienced and the penetration of the vehicle. Changes in vehicle mass have implications for the former (while changes in vehicle footprint may affect the latter). In the case of a single-vehicle collision, large vehicles are more likely to displace a smaller object (e.g., a tree or other barrier), reducing the deceleration

experienced by the occupant of the vehicle. For a simple model of a head-on, multi-vehicle collision^a, the vehicles change their velocities by

$$\Delta v_1 = \frac{m_2}{m_1 + m_2}(v_1 + v_2),$$

$$\Delta v_2 = \frac{m_1}{m_1 + m_2}(v_1 + v_2).$$

Hence, the change in velocity experienced by a vehicle is decreasing in the ratio of its own weight to the weight of the other vehicle (Joks et al., 1998; Greene and Keller, 2002). This suggests that proportional down-weighting of vehicles should not affect fatalities from multi-vehicle crashes. Relative mass may also be important in understanding collisions between a vehicle and a pedestrian, biker, or motorcyclist, where the latter categories can, as White (2004) memorably put it, be thought of as “ultralight vehicles.”

Using recorded crash data, much of the empirical literature of the past three decades has been consistent with the underlying physics. In single-vehicle crashes, there is mixed evidence that weight affects fatality risk^b. Kim et al. (2006) find no evidence, while NHTSA analyses suggest that down-weighting cars and light trucks by 100 lbs would increase fatalities from rollovers and fixed-object collisions by almost 500/year (Kahane, 1997, 2003). Bento et al. (2017) find the opposite result that lower mean weight reduces fatalities in single-vehicle collisions. In all cases though, the number of fatalities relative to the scale of the down-weighting is small.

The relationship between vehicle weight and pedestrian, bicyclist, or motorcyclist (henceforth referred to as pedestrian) safety has also been difficult to characterize. NHTSA analyses have consistently found that down-weighting the smallest cars increases pedestrian fatalities by 3+%, while weight changes across heavier cars and light trucks do not have a statistically significant effect (Kahane, 2003, 2012). There is not a fundamental reason that reducing weight among the lightest vehicles would increase danger to pedestrians, but this result may be explained by small vehicle configurations, driver sorting, or pedestrians behaving differently around smaller cars (i.e., behaving less cautiously because of lower perceived risk). Other studies, comparing pedestrian fatalities from collisions with cars versus light trucks, find that fatality rates are 77%–82% higher for light trucks (and serious injuries increase, as well) (Anderson, 2008; White, 2004). This may be due more to nonweight characteristics, however, as Anderson and Auffhammer (2014) find no statistically significant increase in pedestrian deaths for a 1000-lb weight increase when a light truck indicator is included.

With respect to multi-vehicle collisions, empirical studies generally agree that fatality risk depends on the relative weights of the colliding vehicles (Kim et al., 2006). Specifically, fatality risk to the driver is decreasing with the size of his or her own car and increasing with the size of the other car (Anderson and Auffhammer, 2014; Broughton, 2008, 2012; Evans, 2001; Tolouei and Titheridge, 2009). Notably, these results highlight a key externality: choosing a larger vehicle reduces own risk (at least in a multi-vehicle collision) but increases the danger to others. Many discussions of the effect of vehicle weight on safety wrongly conflate the internal decision-making rule with the net internal and external risks imposed by weight.

Thus far, the discussion of weight and safety has focused on how a change in weight affects the fatality risk conditional on a collision. But to understand the societal effects of a change in weight, it is necessary to calculate the expected risk, which depends on the frequency of collisions involving vehicles of different weights. That is, under a given scenario, how does the probability of a collision between vehicles of size x and size y change? The NHTSA analyses look at the effect of 100-lb weight reductions applied to each subset of vehicles, and while effects vary from study to study (as methodology and included model years vary), it generally appears that only down-weighting the smallest vehicles affects fatalities (with an observed increase of 1%–3%) (Kahane, 1997, 2003; Wenzel, 2012, 2013). More commonly, studies have examined how changes to the fleet that alter the variance of the weight distribution affect societal risk. Anderson and Auffhammer (2014) conclude that it would require a large and uneven downsizing of vehicles to increase overall fatalities by 1% or more. There are a number of studies that quantify the change in fatality risk associated with increasing (White, 2004) or decreasing (Buzeman et al., 1998) the variance of vehicle sizes within the existing fleet. In the British context, Tolouei and Titheridge (2009) examine the relationship between mass and fatalities over time and show that the individual benefits of additional mass vary over time as the size distribution of the fleet changes. Noland (2004) comes to a similar conclusion in the United States, finding that the apparent safety benefits of additional weight became less pronounced over time as the smallest vehicles stopped being produced.

To directly understand the effect of fuel economy regulations on fatalities, then, requires focusing directly on how the policies affect the distribution of vehicle weights in addition to the average effect. Jacobsen (2013) estimates the fatality risk associated with collisions between different vehicles and the effect of historical CAFE levels on market shares for different vehicle types. This is challenging because different types of drivers, such as very safe or risky drivers, may sort into different types of cars and drive in different locations and thus may encounter a different set of drivers around them as well as more or less safe road conditions (Coate and VanderHoff, 2001). Jacobsen addresses such sorting in a model and ultimately determines that each increase in mpg in the standards results in 149 additional fatalities. That is, the shift to smaller vehicles within both cars and light trucks without large shifts between the categories results in increased fatalities. However, Bento et al. (2017) directly document the effect of CAFE on both

^a If the vehicles collide at an angle, the equations are similar, but account for a two-dimensional speed vector.

^b This is distinct from the differential fatality risk associated with cars versus light trucks.

mean weight and weight dispersion and find that even though CAFE increased weight dispersion, the reduced mean weight and sorting of vehicles have a larger, mitigating effect so that in aggregate, CAFE reduced fatalities.

Vehicle Size

While much of the discussion about vehicle size focuses on mass, vehicles consist of a bundle of attributes, some of which tend to be highly correlated. Thus, empirical analyses of the effect of mass may actually be measuring the effect of these other correlated attributes. For instance, heavy vehicles tend to have a higher center of gravity, which increases the probability of a rollover (Li, 2012). Similarly, because heavier vehicles are often taller, their front ends hit the passenger compartment of smaller cars rather than the frame, hitting the upper bodies and heads of the car occupants rather than the crumple space, the area in front of the car intended to absorb impact. These same features likely affect safety outcomes in collisions with pedestrians, as well. The taller front ends of light trucks, in particular, are more likely to crush pedestrians, whereas in collisions with cars, pedestrians may be hit only in the legs and land on the softer hood (White, 2004).

Both of these factors—the high-center-of-gravity-induced rollover risk and dangerous collision geometry associated with tall front ends—would tend to result in overestimates of the risks associated with additional mass. However, perhaps the most important attribute that tends to be strongly associated with weight is footprint or size. One way that automakers can make vehicles lighter and more efficient is by making the vehicles physically smaller. Often, this involves reducing the crumple space, which manages the energy of the crash (Jokschi et al., 1998). When studies have included footprint as a distinct control variable, it tends to limit the effect of mass on fatalities, though these estimates are imprecise due to the high degree of collinearity between mass and volume (Khazzoom, 1994; Jokschi et al., 1998; Kahane, 2012). Relatedly the earlier puzzle of low-weight cars increasing pedestrian fatalities may be explained by the shorter hood; with less hood space, pedestrians' heads are more likely to hit the windshield, with adverse consequences (Kahane, 2003).

Since 2012, the US has had footprint-based standards in part to discourage shifts to smaller vehicles^c. However, this structure incentivizes firms to produce larger vehicles^d, and these incentives may be larger for light trucks than for cars (Whitefoot and Skerlos, 2012). If that is the case, the footprint-based standards would tend to increase the dispersion of vehicle weights, with adverse safety consequences. These results are not certain, however Jacobsen (2013) found that a footprint-based standard had smaller safety effects than an extension of the status quo CAFE, though the cost of meeting the standard would be higher. The precise implications of these new US standards—for safety and for the policy's stated goals—remains an open question.

At the same time that many passenger cars were becoming smaller and lighter in order to adhere to CAFE standards, light trucks were gaining a foothold as passenger vehicles. Many of the concerns about the external risks of heavy vehicles apply to light trucks (Toy and Hammitt, 2003; Gayer, 2004; White, 2004; Anderson, 2008; Fredette et al., 2008; Li, 2012), and certain features of light trucks may make them more dangerous than cars, even conditional on weight (Latin and Kasolas, 2002; Jokschi et al., 1998). However, the extent to which CAFE standards are responsible for the increased demand for light trucks is difficult to show, and there is only an imperfect correlation between increasing standards and the shift toward light trucks.

It is true that, as Figs. 1 and 2 show, the light truck standards have historically been lower than the car standards. It is thus possible for automakers to switch a car model to a light truck chassis to meet the lower standard, avoiding the imperative to improve fuel economy while introducing safety risks to other vehicles that might collide with the light truck (Bento et al., 2017). There are several high-profile examples of the use of this loophole, including the Chrysler PT Cruiser, introduced in 2001, and the Subaru Outback, modified in 2004. The former, which looks and drives like a car but is built on a light truck bed, was determined by NHTSA to be a light truck, and its sales helped bring Chrysler's light truck fleet into compliance with CAFE. The latter, originally sold as a sedan, was modified to be on a light truck chassis, so that vehicles with turbochargers, which hurt fuel economy, could be added to the Subaru fleet and the automaker would still be CAFE compliant.

VMT Rebound Effect

In addition to changing vehicle attributes in ways that may affect traffic fatalities, CAFE may also affect the number of miles people drive. By making the fleet more efficient, the regulation reduces the cost per mile of driving, incentivizing people to drive more—a rebound effect (Gillingham et al., 2016). Estimates of the magnitude of this rebound effect differ across time and context; the extent of the rebound depends on factors such as public transit availability, urban form, and the salience of changes in fuel costs. While estimates cover a wide range, Gillingham (Unpublished) finds that the average vehicle miles traveled (VMT) rebound effect in the recent literature from the United States is 8%–14% (depending on the relative weight put on more reliable odometer-based studies), which means that additional driving erodes 8%–14% of the initial fuel savings from the improved fuel economy from CAFE standards.

There are several issues with applying existing estimates of the rebound effect to the fuel economy context. First, most estimates of the rebound effect in the literature are based on how VMT change in response to changes in gasoline price. The extent to which the rebound effect with respect to fuel prices is the same as the rebound effect with respect to fuel economy is also a matter of debate.

^cThis is due to some extent to the fact that domestic vehicle manufacturers produce larger vehicles and have argued that they were unfairly hurt by CAFE standards.

^dIo and Sallee (2018) show that manufacturers respond to attribute-based efficiency standards in Japan by manipulating weight.

Additionally, these estimates are often based on short- or medium-term responses to price changes; fuel economy policy may have an impact over decades, which might allow for larger responses. Third, the extent to which driving decisions vary based on cost depends on wealth and congestion; as these factors change, the observed rebound effect may be less accurate. Fourth, as noted in the mechanisms section, fuel economy standards can be met via changes to vehicle attributes; if these changes make driving less appealing—for instance, by reducing horsepower—there may be less of a rebound. Another mechanism involves the addition of new, potentially expensive technology to vehicles to improve their performance, which would tend to increase the cost of new vehicles, producing an income effect and reducing the ability to pay for additional miles.

Despite these caveats, it is widely accepted that fuel economy standards would involve a rebound effect. This increase in driving would, all else equal, increase the probability of collisions. The simplest way^e to calculate this effect is to estimate a baseline fatality rate for each model year per cumulative VMT over the life of those vehicles and to combine this rate with the additional miles driven as a result of rebound. In their evaluation of the SAFE rule, which proposes freezing fuel economy standards at 2020 levels, NHTSA and EPA find that the incremental planned increase would result in 6340 additional deaths due to the rebound effect (under the assumption of a 20% effect; if the true effect is half of that size, it would result in 3170 additional deaths) (DOT, 2018).

On the other hand, the *ceteris paribus* assumption may not be reasonable. In response to increased risk, drivers may choose to drive more cautiously. The increased driving may contribute to congestion so that collisions, while more frequent, would occur at lower speeds and thus, with a lower probability of fatalities. This aspect has not yet been quantified to the best of our knowledge.

Vehicle Age

A final channel through which fuel economy policy may affect vehicle fatalities is by keeping old vehicles on the roads. This effect, first documented by Gruenspecht (1982), operates as follows. The standards raise the price of new vehicles, incentivizing buyers to purchase used vehicles instead, which in turn raises the price of used vehicles and can cause some owners of used vehicles to hold onto them for longer. Vehicle manufacturers may be especially likely to either raise the price or alter the attributes of their least efficient vehicles. Thus, the higher price and lower scrappage rates can be particularly pronounced among the lowest fuel economy vehicles. Jacobsen and van Benthem (2015) estimate the used vehicle price elasticity of scrappage and estimate the consequences of a hypothetical more stringent CAFE standard applied between 2010 and 2025, finding that, in terms of leakage (fraction of expected gasoline savings that are lost due to change in scrappage behavior), the Gruenspecht effect is larger than the rebound effect: on the order of 13%–16% of expected fuel savings from a fuel economy policy are lost due to the delayed scrappage of older, inefficient vehicles.

Beyond the consequences for fuel use, the Gruenspecht effect also affects safety. First, older vehicles may work less well. For example, if the brakes fail, collisions involving an older vehicle become both more likely and occur at a higher speed, increasing fatalities conditional on a collision. Second, over the history of fuel economy policy, many safety innovations have been introduced. A requirement that all vehicles include seatbelts took effect in 1968^f, meaning that at the onset of CAFE, some used vehicles that faced delayed scrappage lacked the technology. Airbags, while first installed in some cars in the 1970s, were not required in passenger cars until 1997 and in trucks until the following year. Other technologies, such as anti-locking brake systems or electronic stability control, have also become widespread or even required over the CAFE era, and technologies such as forward collision warnings or crash-imminent braking are increasingly common.

Each additional year a used vehicle lacking one of these technologies remains on the road, the risk of fatalities is increased. Recent NHTSA analysis has documented this effect: the proportion of occupants killed in fatal collisions is strictly increasing in model year age, with the proportion among the oldest model year group (18+ years) twice as high as the proportion among new (0–3) model years, though some of this is due to driver behavior (NHTSA, 2018). On the other hand, older vehicles tend to be driven less. If an older fleet translates to fewer vehicle miles traveled, this may have some offsetting safety benefits, as indicated by the earlier discussion of VMT.

Conclusion

Fuel economy standards have often been critiqued for having negative consequences on traffic safety. Concerns have primarily focused on the use of down-weighting to improve vehicle efficiency and the presumption that smaller cars are intrinsically less safe. From a societal perspective, it is not evident that fuel economy policy-induced changes in vehicle weight reduced safety. There is a general consensus that historically, CAFE contributed to a reduction in vehicle weight and that this reduction disproportionately affected smaller vehicles (though this may no longer be true with footprint-based standards). However, the net effect on traffic fatalities is ambiguous due to the directionally uncertain effect of vehicle weight on single-vehicle fatalities, and the sorting of vehicles by weight within the United States. The effect of CAFE on vehicle footprint, which is highly correlated with weight, also has

^eThis simplification is not without its limitations. The rebound effect may differ across drivers in ways that systematically relate to either their car choices or their location, which would bias the estimated fatalities.

^fThe presence of seat belts in all seating positions was mandated by NHTSA under Title 49 of the United States Code, Chapter 301, Motor Vehicle Standard No. 208. This is distinct from a requirement that passengers wear seatbelts, though regulations mandating seatbelt usage are on the books in all states except New Hampshire.

safety implications. Much of the historical down-weighting involved reductions in vehicle size, and there is some evidence that safety is increasing in vehicle footprint conditional on weight. Furthermore, the regulations may have motivated innovations in light trucks, whose popularity tended to increase the variance of vehicle weights and which may have unique safety disadvantages for those with whom they collide.

While less often discussed, two other consequences of CAFE likely reduced road safety. First, the improved fuel efficiency reduced the cost of driving, incentivizing an increase in VMT. With more driving, collisions and fatalities likely increased. However, it is important to note that this increased driving is not merely a cost—it also improves the welfare of those who are able to drive more than they previously could afford. Second, by changing the costs and attributes of new vehicles, CAFE policy affected the used vehicle market, keeping older cars on the road for longer. These older cars, which may lack the latest safety innovations, are likely less safe than the new vehicles that would have been bought in their absence.

Considerable research has documented the effect of fuel economy standards on safety via these channels. There are also several papers that examine the consequences of fuel economy improvements for traffic fatalities from a more macro viewpoint, by examining the relationship between these variables over time and space. Using state- or national-level variation and controlling, to varying degrees, for safety innovations and demographics, these papers generally did not find much of a relationship between CAFE and traffic fatalities (Noland, 2004, 2005; Ahmad and Greene, 2005). These small effects may reflect the limited safety consequences of fuel economy standards or the challenges of time series analysis with relatively little variation and many confounding factors.

Ultimately, while much is known about how historical fuel economy standards affected weight, footprint, driving behavior, and fleet age, many questions remain to be answered by future work. For instance, there is little work on how footprint-based standards affect safety by leading automakers to upsize vehicles. There is also much room for more analysis on the effect of fuel economy rules on safety outside of the United States. Such future work can help guide policymaking on fuel economy for years to come.

See Also

Automobile accidents and passive prevention systems; Collision avoidance systems, automobiles; Passenger van safety; Head-on crashes; Macroscopic safety analysis

References

- Ahmad, S., Greene, D.L., 2005. Effect of fuel economy on automobile safety: a reexamination. *Transp. Res. Rec.: J. Transp. Res. Board* 1941, 1–7.
- Anderson, M., 2008. Safety for whom? the effects of light trucks on traffic fatalities. *J. Health Econ.* 27, 973–989, doi:10.1016/j.jhealeco.2008.02.001.
- Anderson, M.L., Auffhammer, M., 2014. Pounds that kill: the external costs of vehicle weight. *Rev. Econ. Stud.* 81, 535–571, doi:10.1093/restud/rdt035.
- Anderson, S.T., Parry, I.W.H., Sallee, J.M., Fischer, C., 2011. Automobile fuel economy standards: impacts, efficiency, and alternatives. *Rev. Environ. Econ. Policy* 5 (1), 89–108, doi:10.1093/reep/req021.
- Bento, A., Gillingham, K., Roth, K., 2017. The effect of fuel economy standards on vehicle weight dispersion and accident fatalities. NBER Working Paper Series.
- Broughton, J., 2008. Car driver casualty rates in Great Britain by type of car. *Accid. Anal. Prev.* 40, 1543–1552, doi:10.1016/j.aap.2008.04.002.
- Broughton, J., 2012. The influence of car registration year on driver casualty rates in Great Britain. *Accid. Anal. Prev.* 45, 438–445, doi:10.1016/j.aap.2011.08.011.
- Buzeman, D.G., Viano, D.C., Lovsund, P., 1998. Car occupant safety in frontal crashes: a parameter study of vehicle mass, impact speed, and inherent vehicle protection. *Accid. Anal. Prev.* 30 (6), 713–722.
- Coate, D., VanderHoff, J., 2001. The truth about light trucks. *Regul.* 24 (1), 22–27.
- Crandall, R.W., John, D., Graham, J.D., 1989. The effect of fuel economy standards on automobile safety. *J. Law Econ.* 32 (1), 97–118.
- DOT, 2018. The Safer Affordable Fuel-Efficient (SAFE) Vehicles Rule for Model Year 2021–2026 Passenger Cars and Light Trucks. Preliminary Regulatory Impact Analysis. US Department of Transportation National Highway Traffic Safety Administration.
- Evans, L., 2001. Causal influence of car mass and size on driver fatality risk. *Am. J. Public Health* 91 (7), 1076–1081.
- Fredette, M., Lema, S.M., Chouinard, A., Bellavance, F., 2008. Safety impacts due to the incompatibility of SUVs, minivans, and pickup trucks in two-vehicle collisions. *Accid. Anal. Prev.* 40, 1987–1995, doi:10.1016/j.aap.2008.08.026.
- Gayer, T., 2004. The fatality risks of sport-utility vehicles, vans, and pickups relative to cars. *J. Risk Uncert.* 28 (2), 103–133.
- Gillingham, K., 2013. The Economics of Fuel Economy Standards versus Feebates. National Energy Policy Institute Working Paper.
- Gillingham, K., Unpublished. The rebound effect and the rollback of fuel economy standards, *Rev. Environ. Econ. Policy*.
- Gillingham, K., Rapson, D., Wagner, G., 2016. The rebound effect and energy efficiency policy. *Rev. Environ. Econ. Policy* 10 (1), 68–88, doi:10.1093/reep/rev017.
- Goldberg, P.K., 1998. The effects of the corporate average fuel efficiency standards in the US. *J. Indus. Econ.* 46 (1), 1–33.
- Greene, D.L., Keller, M., 2002. Appendix A—dissent on safety issues: fuel economy and highway safety. In: *Effectiveness and Impact of Corporate Average Fuel Economy (CAFE) Standards*, The National Academies Press, Washington, DC, pp. 117–24.
- Gruenspecht, H.K., 1982. Differentiated regulation: the case of auto emissions standards. *Am. Econ. Rev.* 72 (2), 328–331.
- Ito, K., Sallee, J.M., 2018. The economics of attribute-based regulation: theory and evidence from fuel economy standards. *Rev. Econ. Statist.* 100 (2), 319–336.
- Jacobsen, M.R., 2013. Fuel economy and safety: the influences of vehicle class and driver behavior. *Am. Econ. J.: Appl. Econ.* 5 (3), 1–26.
- Jacobsen, M.R., van Benthem, A.A., 2015. Vehicle scrappage and gasoline policy. *Am. Econ. Rev.* 105 (3), 1312–1338.
- Joksch, H., Dawn, M., Pichler, R., 1998. Vehicle Aggressivity: Fleet Characterization Using Traffic Collision Data. Final Report. DOT HS 808 679. US Department of Transportation National Highway Traffic Safety Administration.
- Kahane, C.J., 1997. Relationships between Vehicle Size and Fatality Risk in Model Year 1985–93 Passenger Cars and Light Trucks. NHTSA Technical Report. DOT HS 808 570. US Department of Transportation National Highway Traffic Safety Administration.
- Kahane, C.J., 2003. Vehicle Weight, Fatality Risk and Crash Compatibility of Model Year 1991–99 Passenger Cars and Light Trucks. NHTSA Technical Report. DOT HS 809 662. US Department of Transportation National Highway Traffic Safety Administration.

- Kahane, C.J., 2012. Relationships Between Fatality Risk, Mass, and Footprint in Model Year 2000-2007 Passenger Cars and LTVs. Final Report. DOT HS 811 665. US Department of Transportation National Highway Traffic Safety Administration.
- Khazzoom, J.D., 1994. Fuel efficiency and automobile safety: single-vehicle highway fatalities for passenger cars. *Energy J.* 15 (4), 49–101.
- Kim, H.S., Hyung, J.K., Bongsoo, S., 2006. Factors associated with automobile accidents and survival. *Accid. Anal. Prev.* 38, 981–987, doi:10.1016/j.aap.2006.04.001.
- Klier, T., Linn, J., 2012. New-vehicle characteristics and the cost of the corporate average fuel economy standard. *RAND J. Econ.* 43 (1), 186–213., doi:10.1111/j.1756-2171.2012.00162.x.
- Klier, T., Linn, J., 2016. The effect of vehicle fuel economy standards on technology adoption. *J. Public Econ.* 133, 41–63, doi:10.2139/ssrn.2539705.
- Latin, H., Kasolas, B., 2002. Bad designs lethal profits: the duty to protect other motorists against SUV collision risks. *Boston Univ. Law Rev.* 82 (5), 1161–1240.
- Li, S., 2012. Traffic safety and vehicle choice: quantifying the effects of the 'arms race' on American roads. *J. Appl. Econ.* 27, 34–62.
- NHTSA, 2018. Passenger Vehicle Occupant Injury Severity by Vehicle Age and Model Year in Fatal Crashes. DOT HS 812 528. Research Note. NHTSA's National Center for Statistics and Analysis.
- Noland, R.B., 2004. Motor vehicle fuel efficiency and traffic fatalities. *Energy J.* 25 (4), 1–22.
- Noland, R.B., 2005. Fuel economy and traffic fatalities: multivariate analysis of international data. *Energy Policy* 33, 2183–2190, doi:10.1016/j.enpol.2004.04.016.
- Tolouei, R., Titheridge, H., 2009. Vehicle mass as a determinant of fuel consumption and secondary safety performance. *Transp. Res. Part D* 14, 385–399, doi:10.1016/j.trd.2009.01.005.
- Toy, E.L., Hammitt, J.K., 2003. Safety impacts of SUVs, vans, and pickup trucks in two-vehicle crashes. *Risk Anal.* 23 (4), 641–650.
- Wenzel, T., 2012. Assessment of NHTSA's Report 'Relationships between fatality risk, mass, and footprint in model year 2000-2007 Passenger Cars and LTVs'. Final report prepared for the Office of Energy Efficiency and Renewable Energy. LBNL-5698E. US Department of Energy. Lawrence Berkeley National Laboratory.
- Wenzel, T., 2013. The estimated effect of mass or footprint reduction in recent light-duty vehicles on US societal fatality risk per vehicle mile traveled. *Accid. Anal. Prev.* 59, 267–276.
- White, M.J., 2004. The 'arms race' on American roads: the effect of sport utility vehicles and pickup trucks on traffic safety. *J. Law Econ.* XLVII, 333–355.
- Whitefoot, K.S., Skerlos, S.J., 2012. Design incentives to increase vehicle size created from the US footprint-based fuel economy standards. *Energy Policy* 41, 402–411, doi:10.1016/j.enpol.2011.10.062.

Hazardous Materials Transport

Dr. Arjan Vincent van der Vlies, Managing Consultant Safety and Crisis management, Berenschot, Utrecht; Guest staff member Institute for Security and Global Affairs, Leiden University, Leiden, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Hazardous Materials Transport	304
Introduction	304
Modes of Transport	304
Air	305
Rail	305
Road	305
Water	305
Rules and Regulations	306
General Arrangements for All Transport Modes	306
Arrangements Per Transport Mode	307
Challenges	308
Challenges Concerning Risk Assessment	308
Low Probability High Impact	308
Terrorism	309
Safety Culture and Human Factors	309
Energy Transition	309
Biography	309
See Also	309
References	310
Further Reading	310

Hazardous Materials Transport

Introduction

Hazardous materials (or dangerous/hazardous goods/substances) are materials that pose a risk to the health and safety of people, living organisms, and/ or the environment. A hazardous material often is a chemical that is prone to reactions with a negative external effect, such as a fire, a toxic cloud, or an explosion.

Mankind has come a long way in its use of hazardous materials, which forms a diverse mix of various materials, which are used for different applications. Hazardous materials can be used as a fuel (gasoline, natural gas), as raw materials for other products (polymers, rubber, fertilizers, plastics), or simply as day-to-day materials used in everyday life (car fuels, cleaning materials). They are produced, stored, and transported from production facilities to their clients and end-users and as such form an important part of our everyday lives and our economies. At the same time, these materials pose a risk to their surroundings due to their intrinsic hazardous nature and as such, these risks need to be managed properly. Although the American DOT and the European Eurostat on average “only” list several dozen casualties per year for their respective territories, transport of hazardous materials is something to be taken very seriously as some of the examples in this section will show.

This section will therefore touch ground on a number of elements that are relevant for transporting hazardous materials, such as the modes of transport that are used for transporting hazardous materials, their respective risks, and some basic aspects of the different transport modalities. We will also give a very general overview of international rules and regulations. Finally, we will give a non-exhaustive overview of challenges and discussions that are currently taking place in academia.

Modes of Transport

There are different modes of transport through which hazardous materials are transported. Each mode has its own specific challenges, risks, and examples of disasters. Here we will briefly discuss air, rail, road, and water transport (in alphabetical order). Obviously, hazardous materials can also be transported by pipelines, but as a mode of transport, this is not included here as there is also a dedicated section to pipeline transport of hazardous materials in this work.

Air

Air transport is not generally associated with large bulk quantities of hazardous materials, but is mostly associated with smaller quantities, or specific materials, such as infectious items, ammunition, and radioactive materials. Specific attention is paid to lithium batteries as these can develop enormous amounts of heat in seconds as well as toxic fumes when not properly handled or in case of production errors. Examples of this are the UPS airlines flight 6 crash in 2010 and the mysterious missing of flight MH370 in 2014, which was also linked to lithium ion batteries, although the wreck remains unfound as of April 2019.

Rail

Rail transport of hazardous materials has a number of challenges concerning routing, risks, and alternatives for the transport. Statistically, rail transport is a very safe mode with a very good track record. Nevertheless, there have been very serious disasters and incidents, such as in Ryonchon, North Korea (2004), Graniteville, USA (2005), Viareggio, Italy (2009) and Lac Megantic, Canada (2013). What makes this transport specifically interesting from a risk management point of view is that transport—although certainly not in every country or region to the same extent—runs on tracks that are also used for passenger transport and as such run through densely populated areas. Diverting transport around densely populated areas often is not possible as rail networks are less branched than road networks, for example. This means that alternative routes from A to B, apart from the shortest or fastest route can be scarce and less attractive from a logistical or cost point of view.

Tank wagons are roughly 2–3 times the size of tank truck and a train may transport large bulk quantities for a specific client, as well as different materials altogether (ranging from different hazardous materials to mixing hazardous materials with non-hazardous materials). Hence, the transported amounts may be enormous and could have a larger impact altogether, if several wagons are inflicted at the same time (e.g., in the Lac Megantic disaster, where 63 tank wagons of a 74 wagon train carrying crude oil exploded or burned excessively), or when the materials of one wagon (e.g., carrying flammable liquids) may inflict damage to another wagon carrying hazardous materials (e.g., flammable gases) which may cause an even greater disaster. The potential lethal effects of rail accidents with hazardous materials vary from several dozens of meters (in case of a pool fire), up to 150 m (in case of an explosion of flammable gasses) to over a kilometer (in case of a toxic gas).

Road

Road transport is a common and visible way of transporting hazardous materials. For example, for transporting fuels for cars and trucks, but also for transporting bulk of other materials in the industry, or in smaller non-bulk quantities (e.g., in boxes or drums). In history, there have been many examples of incidents with road transport of hazardous materials. In 1978, for example, over 200 people were killed in Los Alfaques, Spain after a Boiling liquid expanding vapor explosion (BLEVE) occurred at a camping site. More recently in 2018, a BLEVE occurred on a road near Bologna, Italy.

From a risk management perspective, the difference from rail transport is that road transport is often directed around cities and other populated areas. This of course leaves out the fact that supplying gas stations often leads to trucks entering densely populated areas. Also, there are many more road trucks than rail cars and road trucks interact with other traffic much more often than rail transport, which increases the potential for accidents and negative consequences. Moreover, an incident with a truck often leads to more significant spills than is the case with rail transport. A final substantial element of road transport of hazardous materials is that transport through tunnels causes significant risks for other traffic. This is due to the nature of tunnels where the confinement of space leads to other effects than in the open air. Infamous is the 1982 gasoline truck explosion in the Salang tunnel in Afghanistan, which may have caused up to 2700 fatalities and could be the deadliest road incident ever recorded. Other tunnel disasters, such as in the European Mont Blanc Tunnel (1999) and Gotthard Tunnel (2001) did not involve transport of hazardous materials (instead, the disaster in the Mont Blanc Tunnel involved flour and margarine) indicate that even when there are non-hazardous materials involved, incidents in tunnels may have enormous effects.

Water

Here we need to distinguish between inland and sea transport. Transport of hazardous materials by ships leads to different risks than by road and rail. In most cases, the distance between the transport of the hazardous materials and the adjacent area is larger than with rail and road transport. This leads to lower risks as the effect of an incident leads to less (or no) casualties. Nevertheless, bulk sea transport (e.g., tanks with toxic chemicals) or oil tankers still pose a significant environmental risk. In case of a disaster, such as a collision or sinking with a loss of containment, this may lead to large environmental disasters, such as infamous disasters the *Odyssey* and the *Exxon Valdez* spills in 1988 and 1989, respectively or more recently the *Sanchi* oil tanker collision in 2018.

These different modes of transport all have their own respective history of disasters, challenges, and risks. This had led to the development of different rules and regulations with the aim of creating a basic level of safety, as well as developing a clear regulatory framework for all involved stakeholders. We will elaborate the most important international rules and regulations in the following section.

Rules and Regulations

There are many different roles involved in transporting hazardous materials, from shippers, manufacturers, producers, carriers, governments, and emergency responders to civilians and workers. Each of them have different interests and are affected differently. People living near transport routes wish to live safely, while carriers could wish to minimize costs and governments carry responsibility for maintaining a general level of safety. Through the years and often in response to major disasters, different rules and regulation have been implemented. To ensure a level playing field and minimum requirements concerning the safety of transport, general requirements are applicable to the transport of hazardous materials. Here we will touch base of some of the important and dominant international rules and regulations.

General Arrangements for All Transport Modes

The committee of experts on the transport of dangerous goods of the United Nations Economic and Social Council (ECOSOC) has adopted the UN recommendations on the transport of dangerous goods (also known as the Orange book). The UN recommendations are contained in UN model regulations, which aim to present a basic scheme of provisions for all modes of transport (except oceanic bulk tankers) that will allow uniform development of national and international regulations governing the various modes of transport. As such, they are not mandatory, but have gained widespread international acceptance and are transcribed in national and regional legislation and regulation.

The model regulations cover principles of classification and definition of classes, listing of the principal dangerous goods, general packing requirements, testing procedures, marking, labeling or placarding, and transport documents. These include requirements concerning the construction and material and product testing.

An important aspect herein is the marking and labeling. The model regulations therefore adopted nine different classes of hazardous materials, each with different subcategories, to characterize the risks of different chemicals. The classes are:

Class 1: Explosives. Divided into six subcategories ranging from explosives with a mass explosion hazard to extremely insensitive, this class contains explosives, such as TNT, nitroglycerine, consumer fireworks, and ammunition.

Class 2: Gases. This class distinguishes between flammable (e.g., butadiene, methane, hydrogen), non-flammable (e.g., carbon dioxide, helium, nitrogen), and poisonous gases (e.g., chlorine, ammonia). These gases can be transported in different phases, such as compressed or liquefied. Apart from being flammable or poisonous, gases can also be asphyxiants or cryogenic.

Class 3: Flammable liquids. This class is subdivided into three packing groups depending on their boiling and flash point. This class contains liquids, such as gasoline, kerosene, and diesel.

Class 4: Flammable solids. Flammable solids are materials that are either flammable and ignite easily (e.g., cellulose nitrate or magnesium), spontaneously combustible and will ignite when in contact with oxygen (e.g., white phosphorus) or dangerous when wet (these are substances that will emit flammable gases or will react when in contact with water, such as potassium, cesium, or sodium).

Class 5: Oxidizing agents and organic peroxides. These substances are not necessarily combustible, but may cause, or contribute to the combustion of other material by yielding oxygen (e.g., ammonium nitrate). Organic peroxides are thermally unstable substances, which may undergo exothermic self-accelerating decomposition (acetone peroxide).

Class 6: Toxic and infectious substances. Toxic substances (class 6.1) are classified for being harmful or able to cause serious injury or death if swallowed, inhaled, or by contact with skin (e.g., pesticides). Infectious substances (class 6.2) are biohazardous substances, classified for containing pathogens, including bacteria, viruses, parasites, or other agents, which can cause disease in humans or animals upon contact. Medical waste also falls under class 6 materials.

Class 7: Radioactive substances. These are substances, such as uranium, plutonium that are subject to radioactive decay and as such emit ionizing radiation.

Class 8: Corrosive substances. These are substances, which will cause severe damage when in contact with the living tissue, or will materially damage, or even destroy, other goods, such as metals. This class is divided into an acid subclass (e.g., hydrochloric acid) and an alkalis subclass (e.g., potassium hydroxide).

Class 9: Miscellaneous. Substances and articles, which during transport present a danger, not covered by other classes, such as asbestos, seatbelt pre-tensioners, and some lithium batteries.

Each of the aforementioned classes comes with its own labels and placards (as shown in [Fig. 1](#)). In addition, hazard identification numbers (or Kemler codes) and the UN numbers are used to identify materials that are being transported. When transporting hazardous materials, tank wagons must be fitted with orange plates bearing the hazard identification number (here, "33") and the UN number ("1203"). The hazard identification number includes four numbers that indicate the nature of the transported material, while the UN number gives the exact identification of the goods. The hazard plate shown in [Fig. 2](#) indicates that a highly flammable liquid (33) is being transported, and that this liquid is petrol (1203).

This system of classification, listing, packing, marking, labeling, placarding, and documentation is meant to simplify transport, handling, and control as well as increased uniformity and clarity in dealing with incidents. This should, in general use, create more safety and let carriers, consignors, inspecting authorities, and of course the public, benefit from more safety.



Figure 1 Example of a class 1.1 label for explosives with a mass explosion hazards. Source: Available from https://commons.wikimedia.org/wiki/ADR_labels_of_danger#/media/File:Dangclass1_1.svg, User <https://commons.wikimedia.org/wiki/User:W!B>. (no changes were made).



Figure 2 Example of an orange hazard plate. Source: Available from: <https://commons.wikimedia.org/wiki/User:Masur>.

Arrangements Per Transport Mode

Hence, the UN model regulations form the template for regulating the transport of hazardous materials. Competent authorities (e.g., the Chinese Ministry of Transport, the DOT in the United States, the ACTDG in Australia, or the European Commission) may modify model recommendations for State/region specific regulations. Each modality has its own specific requirements that are described in their specific regulations:

Air transport: Individual airline and governmental requirements are incorporated by the International Air Transport Association (IATA) to produce the widely used IATA dangerous goods regulations (DGR). This manual is the global reference for shipping dangerous goods by air and the recognized standard by airlines.

Rail transport: For all ground transport, the UN model regulations are leading. However, local or regional regulations may apply. One such pan-continental regulations concerns the development by the Organization for International Carriage by Rail (Organisation intergouvernementale pour les Transports Internationaux Ferroviaires; OTIF), an intergovernmental organization of 50 European, African, and Near Eastern states, of the international regulations concerning the international carriage of dangerous goods by Rail (the Règlement concernant le transport international ferroviaire de marchandises dangereuses, or RID). The RID has two annexes that create a framework to regulate the transport of hazardous materials between member states. The two annexes include:

1. Regulations concerning the merchandise involved, notably their packaging and labels (this is called Annex A).
2. Regulations concerning the construction, equipment, and use of vehicles for the transport of hazardous materials (this is called Annex B).

Road transport: Here as well, the UN Model Regulations are leading, although local or regional regulations may apply. As such, the *Accord Européen relatif au transport international des marchandises Dangereuses par Route* (ADR) is also important to recognize. The ADR states that with the exception of certain exceptionally dangerous materials, hazardous materials may, in general, be transported internationally in wheeled vehicles, provided that two Annexes be met, which basically have the same scope as the two Annexes for rail transport: Annex A regulates the merchandise involved, notably their packaging and labels and Annex B regulates the construction, equipment, and use of vehicles for the transport of hazardous materials.

Moreover, the ADR Secretariat has defined a classification system for major tunnels in Europe. The categories are based on the idea that there are three major dangers that may cause numerous victims or serious damage to the tunnel structure, namely explosions, release of toxic gases, or the release of volatile toxic liquids and fire. Each national authority is responsible to categorize its tunnels in classes ranging from A (least restrictive), to E (most restrictive).

Water transport: In case of sea transport, the UN Model Regulations do not apply. Instead, the International Maritime Organization (IMO) has developed the International Maritime Dangerous Goods Code ("IMDG Code", part of the International Convention for the Safety of Life at Sea) for transportation of dangerous goods across seas. The IMDG Code lays down basic principles and a number of recommendations for good operational practice and on how to handle individual substances, materials, and articles. This includes advice on terminology, packing, labeling, stowage, handling, and emergency response action concerning hazardous materials. The provisions in the IMDG Code are directed at the mariner, but also affect

industries and services, such as manufacturers, packers, shippers, and port authorities. IMO member countries have also developed the hazardous and noxious substances by sea convention (HNS Convention) to provide compensation should there be a spill of dangerous goods in the sea. For inland shipping, other rules and regulations may apply. Since 2008, for example, the European agreement concerning the International Carriage of Dangerous Goods by Inland Waterways (Accord Européen relatif au transport des marchandises dangereuses par voies de navigation intérieures, or ADN) applies to inland shipping of hazardous materials. The ADN contains among others provisions concerning dangerous substances and articles, provisions concerning their carriage in packages and in bulk on board inland navigation vessels or tank vessels, as well as provisions concerning the construction and operation of such vessels.

Challenges

There are a number of challenges concerning the transport of hazardous materials. There are many publications that give interesting overviews of the challenges facing transport of hazardous materials. We will here address a number of these challenges in a non-exhaustive manner.

Challenges Concerning Risk Assessment

Each mode of transport has its challenges concerning managing the inherent risks related to the transport. It can however be stated that there is a considerable body of knowledge on quantitative risk assessment (QRA). QRA's generally follow three steps: (1) hazard and exposed receptor identification; (2) frequency analysis; and (3) consequence modeling and risk calculation (Bedford and Cooke, 2001). The result is a numerical assessment of the risk and consequences. Consequences are mostly seen in terms of casualties (number of people affected), but can also be expressed in economical or ecological damage. In terms of casualties, the outcome can be expressed in two ways: first there is the individual risk. This is the annual probability that an unprotected person will die as a result of an accident involving hazardous materials at a certain spot if that person resides there for a full year. The risk is visualized on a map by dots, which act as spatial contours. Due to the impact of a large disaster, not only individuals are affected, but also groups of people are affected. In such cases, we talk about societal risk (or group risk). This is the cumulative probability for each year that at least 10,100, or 1000 people die as a direct result of their presence in the influence area of an establishment or transport route if an incident happens with hazardous materials. This is visualized on a logarithmic scale by using the fN curve, where f represents the frequency of an accident and N the number of people expected to die as a result of that accident. This feature allows that quantitative risk assessments are often used to model a specific situation (e.g., a rail track through a dense urban area) and identify the risk level, which can then be related to norms or standards and test whether certain situations are allowed. Nevertheless, a clear problem exists in assessing the risks of transport of hazardous materials. This is due to the fact that there are relatively little accidents, which results in little reliable statistics due to little empirical reviews of accidents, their causes, effects, and dispersion of hazardous materials (Erkut et al., 2007; Erkut and Verter, 1998). Other critiques on QRAs often concern the technocratic nature, that models and people modeling the same situation generate various outcomes and that models are simplified depictions of reality that do not generate absolute truths, but should be used as decision support.

Qualitative risk assessments deal with the identification of possible accident scenarios and attempts to estimate the undesirable consequences. Qualitative methods are, for example, event tree and fault tree analysis or consultation of experts. Not only are these assessments often necessary when reliable data to estimate accident probabilities and consequence measures are scarce or missing, it also allows for more tailor-made solutions in practice. The goal is to identify events that appear to be most likely and those with the most severe consequences, and focus on them for further analysis or identify risk mitigation measures to limit the possibility of certain specific scenarios from occurring. This makes that qualitative risk assessments are often preferable to quantitative risk assessments to mitigate specific scenarios (Van der Vlies, 2011).

Low Probability High Impact

Transport of hazardous materials is often seen as low probability high impact risks. This means that the chances of an accident are slim, but the effects may be enormous (Jonkman et al., 2003). It can therefore be quite difficult for decision-makers to base policy and management on extreme scenarios that have a very small chance of occurring, but with enormous effects. One of the worst-case scenarios generally associated with the transport of hazardous materials transport, for example, is a BLEVE. In this scenario, a vessel or tank containing a liquid or pressurized gas instantaneously ruptures after the temperatures of the material have reached a level above their boiling point, for example by an external heat source (fire). In order to control the effect, a larger distance from the source to the adjacent area is effective, but also next to impossible in dense countries. These countries often implement a risk-based policy (Erkut et al., 2007; Van der Vlies, 2011). When a decision-maker wishes to have an effect based policy and minimize the effect of a potential disaster, this leads to rerouting transport axes and divert them from city centers and other dense areas. In some countries, this may be an effective policy but this also leads to large costs. As mentioned before, for water transport there often already is a larger

distance from the source to populated areas, which decreases the potential effect on humans. Of course a spill by, for example, an oil tanker may lead to a large effect area and impact marine life.

Terrorism

In recent years, a growing concern involves the use of transport of hazardous materials as a potential terrorist weapon (Erkut et al., 2007; Milazzo et al., 2009). This is not only due to general terrorist threats around the globe, but also after a devastating explosion in Toulouse France in 2001 was linked to a terrorist attack, as well as a failed terrorist attack on a French chemical plant in 2015. It is of course very difficult, if not impossible, to predict an attack. There are, however, attempts to identify risks associated to these threats and how to mitigate and predict them.

Safety Culture and Human Factors

Luckily, the standards as mentioned earlier in the international rules and regulations, have made all transport of hazardous materials modes relatively safe. Nevertheless, having good policies and standards leaves out the fact that there are humans interacting with technologies and rules. This leads to the notion that safety culture is an important concept and is often described as the environment within which individual safety attitudes develop and persist and safety behaviors are promoted. In general, safety culture can be defined as the collection of the beliefs, perceptions, and values that employees share in relation to risks within an organization, such as a workplace or community. Recent decades show a vast body of knowledge on this subject. After incident analyses of large disasters such as the space shuttle challenger disaster, the Piper Alpha oil platform explosion, and also several Korean air incidents, there are a number of risk factors that may lead to low(er) safety culture, such as a strong hierarchical organization in which subordinates do not dare to speak out to managers, profit over safety (if profit and revenue are more important than safety), fear of, for example, reprimands, and ineffective leadership (if leaders give the wrong safety examples). For more on this topic, see the Safety Culture chapter in this work.

Energy Transition

One of the major societal challenges is the transition toward a sustainable society. In the slipstream of the energy transition, there are also challenges concerning safety and transport of hazardous materials. The transition to less polluting fuels leads to a larger demand for compressed and liquefied natural gas (CNG and LNG). Although these fuels are not new, the increase of transport will lead to more risks. Especially the transition to a more hydrogen based economy and an influx in the use of hydrogen may lead to a large increase of demand and transport of hydrogen. What this will effectively mean for risks is not fully clear and also depends on the extent to which hydrogen will replace other fuels, where and how this is produced and of course how this will be transported from A to B.

Biography



Vincent van der Vlies is managing consultant at the Dutch consultancy firm Berenschot as well as guest researcher at the Institute for Security and Global Affairs of Leiden University. Vincent wrote a dissertation concerning the transport of hazardous materials by rail and how this affects urban planning in the Netherlands. In his career as a consultant, he performed dozens of studies concerning hazardous materials, as well as other safety and security-related methods ranging from incident analyses, political strategic risk management, big data safety analyses, and safety audits. To delineate this broad expertise, he studied subjects, such as hooligan-related incidents in professional football, security in cruise terminals and ships, the consequences of a no-deal Brexit and trains professionals in safety- and security-related topics.

See Also

Safety culture; Pipelines

References

- Bedford, T., Cooke, R.M., 2001. Probabilistic Risk Analysis: Foundations and Methods. Cambridge University Press, Cambridge.
- Erkut, E., Tjandra, S.A., Verter, V., 2007. Hazardous materials transportation. In: Handbooks in Operations Research and Management Science. Elsevier, Netherlands, pp. 539–621 (Chapter 9). Available from: [https://doi.org/10.1016/S0927-0507\(06\)14009-8](https://doi.org/10.1016/S0927-0507(06)14009-8).
- Erkut, E., Verter, V., 1998. Modeling of transport risk for hazardous materials. *Oper. Res.* 46 (5), 625–642.
- Jonkman, S.N., van Gelder, P.H., Vrijling, J.K., 2003. An overview of quantitative risk measures for loss of life and economic damage. *J. J. Hazard. Mater.* 99, 1–30.
- Milazzo, M.F., Ancione, G., Lisi, R., Vianello, C., Maschio, G., 2009. Risk management of terrorist attacks in the transport of hazardous materials using dynamic geoevents. *J. Loss Prev. Process Ind.* 22 (5), 625–633. Doi: 10.1016/j.jlp.2009.02.014.
- Van der Vlies, A.V., 2011. Rail transport risks and urban planning: solving deadlock situations between urban planning and rail transport of hazardous materials in the Netherlands, T2011/14, October 2011, TRAIL Thesis Series, The Netherlands.

Further Reading

- Bubbico, R., Di Cave, S., Mazzarotta, B., 2004. Risk analysis for road and rail transport of hazardous materials: a GIS approach. *J. Loss Prev. Process Ind.* 17(6), 483–488.
- Landucci, G., Tugnoli, A., Busini, V., Derudi, M., Rota, R., Cozzani, V., 2011. The Viareggio LPG accident: lessons learnt. *J. Loss Prev. Process Ind.* 24(4), 466–476. Available from: <https://doi.org/10.1016/j.jlp.2011.04.001>.
- Reilly, A., Nozick, L., Xu, N., Jones, D., 2012. Game theory-based identification of facility use restrictions for the movement of hazardous materials under terrorist threat. *Transp. Res. Part E: Logist. Transp. Rev.* 48(1), 115–131. Available from: <https://doi.org/10.1016/j.tre.2011.06.002>.
- Torretta, V., Rada, E.C., Schiavon, M., Viotti, P., 2017. Decision support systems for assessing risks involved in transporting hazardous materials: a review. *Saf. Sci.* 92, 1–9. Available from: <https://doi.org/10.1016/j.ssci.2016.09.008>.

Head-on Crashes

John N. Ivan, Civil and Environmental Engineering, University of Connecticut, Storrs, CT, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	311
Physics of Head-on Crashes	311
Road Conditions	311
Traffic Conditions	312
Severity of Head-on Crashes	312
Prevention of Head-on Crashes	313
Infrastructure Solutions	313
Vehicle Technology Solutions	314
Mitigation of Head-on Crash Severity	314
Conclusions	315
References	315

Introduction

According to the United State's National Highway Traffic Safety Administration (NHTSA), a head-on crash "refers to a collision where the front end of one vehicle collides with the front end of another vehicle while the two vehicles are traveling in opposite directions." (NHTSA, 2015). By that definition, for a head-on collision to occur, two vehicles must meet one another in opposite directions on the same roadway. This requires a two-way road with no physical separation between opposite directions of travel. This could be either a single carriageway or a dual carriageway with a narrow median separation and no physical barrier that would prevent vehicles from crossing over to the other carriageway.

Head-on crashes are among the most severe of crashes due to kinematics. Crash severity is related to the combined speed of impact (Richards and Cuerden, 2009), which for head-on crashes is the sum of the speeds of the two vehicles at impact, rather than the difference in speeds for a rear-end impact, or the speed of the ramming vehicle for a side-impact crash. This is borne out by crash fatality statistics. According to NHTSA (2015), head-on crashes represented 10.2% of all fatal crashes in the United States in 2015, compared to 2.3% of crashes overall. Due to the nature of their occurrence, head-on crashes usually involve an innocent victim driver and occupants with little control to avoid the crash, and in many cases, occupants of the not-at-fault vehicle suffer more severe injuries (Wali et al., 2018).

The rest of this article will first discuss the physics of head-on crashes, including the physical and traffic conditions that give rise to head-on crashes and the physics of how severe they are. Next will be discussion of methods that have been proven to prevent the occurrence of head-on crashes, including infrastructure and technological solutions. Then will be discussion of methods to mitigate the severity of head-on crashes, including vehicle technology and behavioral interventions. Finally will come a summary of the ideas in the article along with conclusions and perspectives for future research.

Physics of Head-on Crashes

Road Conditions

As noted previously, by its definition a head-on crash can only occur on a two-way road with no physical separation between directions of travel. One vehicle must stray out of the lanes assigned for travel in its direction into the lanes assigned for the opposite direction of travel. This can happen either intentionally, typically when a driver is passing another vehicle, or unintentionally, due to a driver either being distracted, drowsy or impaired, or traveling too fast so they lose control and cross into oncoming traffic. Accordingly, narrow lanes and shoulders offer less forgiveness for straying out of lane unintentionally and recovering and are thus usually associated with higher numbers of head-on crashes (Qin et al., 2004). Smaller radius curves are also associated with greater propensity for drivers to stray out of their lanes (Xu et al., 2018). At the same time, a FHWA study of crash causality found wider lanes and shoulders and a higher speed limit to be associated with more opposite direction crashes on horizontal curves (Porter et al., 2018). This study also found an increasing radius of curvature to be associated with fewer opposite direction crashes. While wider lanes and shoulders offer more forgiveness for brief lateral straying from the travel lane, they also can encourage drivers to travel through a horizontal curve at higher speeds than the curve can be safely negotiated.

Restricted sight distance can be more of an issue for head-on crashes than it is for other types of crashes, due to drivers needing to see farther down the road for passing on two-lane roads. AASHTO (2018) recommends a minimum of 280 m of passing sight distance for roads with speed limits of 81 km/h and 440 m for a 130 km/h speed limit. Belz and Chang (2018) conducted a simulator

study examining driver behavior when passing on two-lane two-way roads. Among other things, they found that the presence of horizontal curves affects the distance participants required to pass another vehicle. In a country with right-hand traffic, drivers were less likely to pass when the road bended to the right than if it was straight, and more likely to pass when the road bended to the left. The required sight distance is greater at higher speeds, since both the passing and oncoming vehicle close in on one another at a faster rate.

Centerline markings can increase drivers' awareness of the presence of oncoming traffic in the adjacent lane and the attendant danger in crossing out of the lane. In particular, [Harrison et al. \(2016\)](#) found that painting a 1-m wide centerline in the center of two-lane two-way roads in Queensland Australia reduced casualties by 25%. At the same time, centerline markings can also give drivers more security about where their space on the road is, leading to higher speeds, unless lane or shoulder widths are reduced.

Traffic Conditions

In order for road conditions to result in a crash, the driver must necessarily make an unsafe action, either intentionally or unintentionally. Unintentional unsafe actions frequently occur due to a driver being unable to react appropriately to unexpected traffic conditions. This often happens because drivers are overstimulated by an abundance of external stimuli in the roadway environment in different locations relative to the driver, for example heavy oncoming traffic combined with intersecting roadways or a forested area with the possibility of wildlife entering the roadway ([Thiffault and Bergeron, 2003](#); [Heslop, 2014](#)). However, drivers can also be understimulated due to driver monotony, for example, when there are few if any other vehicles on the road and little possibility of conflict with other road users or wildlife. This can apply to head-on crashes, where with low volumes, drivers do not expect or prepare for meeting oncoming traffic. With high volumes, drivers can be overstimulated and not able to react in enough time to an oncoming vehicle crossing the centerline. Multiple studies that considered head-on or opposite direction crashes on segments separately from other multiple-vehicle crashes have found the rate of these crashes (crashes per vehicle-distance) to increase with traffic volume ([Qin et al., 2004](#); [Porter et al., 2018](#)).

[Arvin et al. \(2019\)](#) conducted a study of driver maneuver and response volatility, that is, variability moment to moment in driver actions such as steering and braking. They found speed, longitudinal and lateral acceleration volatilities to be highly associated with the frequency of head-on crashes. In particular, intersections with high-lateral volatility, that is, lane keeping, have greater risk of head-on collisions.

[Oviedo-Trespalcacios et al. \(2016\)](#) conducted an extensive literature review of distracted driving and operational impacts, including safety. They reported that distracted driving tends to result in reduced speeds. The effects on lane position are less consistent, but there is a difference between distracted and non-distracted conditions, which would be a precursor to head-on crashes. Activities requiring longer off-road glances, such as text messaging rather than talking on the phone, resulted in greater crash likelihood, and the longer the glance away from the road, the higher crash risk. Given that deviation from proper lane position is a common result of distraction, reducing driver distraction of any kind would be important for reducing head-on crashes.

Severity of Head-on Crashes

Head-on crashes resulting from direct hits have higher severity than those that result from impact on an oblique angle. [Jurewicz et al. \(2016\)](#) report that head-on crashes with a direct hit, that is, when the angle between the colliding vehicles is 180 degrees, have a critical injury speed of 48 km/h (30 mi/h) for the striking vehicle, 80 km/h (50 mi/h) for rear end crashes. Opposite turning crashes, or those hitting not direct can vary up to 64 km/h (40 mi/h).

A higher change in speed during the crash results in increased severity. [Richards and Cuerden \(2009\)](#) compiled data from crash reconstruction studies in Britain, examining the relationship between the delta v (change in speed pre-crash to collision) and the severity of the outcome (fatal, severe or slight injury). For head-on crashes, a change in speed of about 30 km/h (20 mi/h) results in a fatality 10% of the time. At about 55 km/h (35 mi/h), it is 50%; and a change of about 80 km/h (50 mi/h) is 90% fatal.

[Deng et al. \(2006\)](#) investigated how physical characteristics of the roadway affect the severity of head-on crashes. They found that while a wider roadway reduces the severity, wet pavements and high density of access points both increase the severity. Nighttime crashes are also more severe than crashes during daylight hours.

Vehicle mass and mismatched bumpers are issues related to crashes between heavy vehicles and SUVs or pickup trucks with passenger cars. [Acerno et al. \(2004\)](#) studied records in the Seattle Crash Injury Research and Engineering Network database to investigate the impacts of crashes between passenger cars (PVs) and light duty trucks (LTs), including SUVs and pickup trucks. Occupants in passenger cars experiencing frontal crashes are far more likely to be killed than those in light duty trucks, due not only to the differences in vehicle mass but also bumper heights. In their sample, four out of 19 in passenger cars were killed, none in LTs were killed. Of the 15 who survived, 14 sustained serious injuries, compared to only three of the fifteen in LTs. These results are consistent with those reported by [IIHS \(1999\)](#), in which the relative risk of death for occupants of PVs involved in frontal collisions with LTVs is 3–4 times greater than those involved in similar collisions with another PV.

Prevention of Head-on Crashes

Infrastructure Solutions

A simple and long-running solution to reducing the number of head-on crashes on two-lane two-way roads is to implement no passing zones in locations with insufficient sight distance for passing. [AASHTO \(2018\)](#) provides recommended minimum sight distance requirements for passing based on design values for perception reaction time, assumed differential speed with the vehicle being passed, and a desirable buffer distance after completing the passing maneuver. No passing zones can also be implemented in locations that experience frequent head-on crashes, or when traffic volumes are high enough that passing is not practical.

A head-on crash can also occur when a driver travels the wrong way on a divided or one-way roadway. This most commonly happens when a driver inadvertently enters a freeway using an offramp rather than an onramp. Such drivers are frequently intoxicated or otherwise not fully aware of the surroundings, missing the warning signage that is required for such conditions in traffic regulations set by road jurisdictions (in all developed countries and most developing countries as well) ([FHWA, 2009](#)). [Finley and Miles \(2018\)](#) describe an evaluation of countermeasures aimed at reducing the incidence of wrong way driving, primarily due to intoxicated drivers. They conducted an experiment on a closed course to evaluate the visibility of a variety of wrong way warning options, including larger than usual signage, retroreflectivity, and flashing digital lights on the sign edges, as well as mounting the signs at both normal (2.14 m, or 7 ft, off the ground) and reduced (0.61 m or 2 ft) heights. They also evaluated retroreflective pavement arrows indicating the correct direction of travel. Participants were asked how difficult it was to identify the signs in darkness at varying levels of intoxication; they reported that the retroreflectivity and digital blinking lights improved visibility, though the effect was not significant.

Another approach for reducing head-on crashes is aimed at reducing inadvertent crossing of the centerline. Various highway agencies have tried to do this using centerline rumble strips between the lanes of opposing traffic. [Persaud et al. \(2004\)](#) describe centerline rumble strips as “either raised or grooved patterns, installed perpendicular to the direction of travel, that produce audible and tactile warnings when traversed by vehicle tires.” They analyzed data for 210 miles of treated roads in seven US states before and after installation of centerline rumble strips. They found a combined reduction of 14% for all injury crashes, and a 25% reduction in frontal opposing-direction sideswipe injury crashes. [Babineau and Fontaine \(2018\)](#) describe the findings from an implementation of centerline rumble strips by Virginia Department of Transportation. They found a reduction of 42% in head-on, opposite direction sideswipe, and fixed object, off-road crashes on curves with design speed of 55 mph or less. For all road sections irrespective of curvature, the reduction in these crashes was 35%.

In some locations, centerline rumble strips are either not practical or insufficiently effective. Rumble bars are an alternative audible warning device for these situations. Rather than depressed in the pavement-like centerline rumble strips, rumble bars are raised above the pavement, constructed of thermoplastic, and installed along with the centerline pavement markings ([Wu et al., 2018](#)). An implementation of rumble bars on 189 miles of rural two-lane highways in Texas found a reduction of 21.3% of all single vehicle run off road and opposite direction crashes, and up to a 40% reduction for injury crashes of those types.

Rumble strips and rumble bars are useful for warning drivers when they are about to cross into oncoming traffic, but they do not prevent such centerline or median breaches. More effective is to install a physical barrier between opposing lanes of traffic. A cost-effective solution is cable barriers designed not to stop the vehicle instantly, but to prevent a full breach into the opposing traffic. Cable barriers have been used effectively to reduce median crossovers on divided highways ([Kirsch et al., 2018](#); [Eustace and Almothaffar, 2018](#); [Srinivasan et al., 2017](#)). These barriers effectively eliminated virtually all median breaches and reduced 80% of fatal and injury crashes ([Eustace and Almothaffar, 2018](#)). [Srinivasan et al. \(2017\)](#) estimated a benefit cost ratio of up to 8.28. The findings indicated universally that cable barriers do their job—median crossover crashes are converted to collisions with the cable barrier, which have much lower severity.

Cable barriers have also been used on rural two-lane two-way highways to eliminate head-on crashes. These installations are generally done without substantial widening of the centerline, only enough to construct the barrier. [Xian et al. \(2019\)](#) describe an evaluation of an installation of cable barrier along the centerline on two-way two-lane expressways in rural areas in Japan. They found this treatment virtually eliminated crossover crashes, though crashes into the barrier still occurred. Widening the interior striping adjacent to the cables reduced contact with the cables. On average 6.6 poles were damaged in each crash with the cable barrier, requiring 50 min of road closure time to repair.

A drawback of cable barrier installation on rural two-lane two-way highways is that the oncoming lane can no longer be used for passing. In situations where accommodating passing is desirable and there is sufficient pavement width, the pavement can be restriped to provide two lanes in one direction and one in the other, with the two directions separated by a cable barrier to prevent vehicles from crossing over. The split between 2 and 1 lanes alternates periodically along the road. This has come to be known as a “2 + 1 road” ([TRB, 2003](#)). [Fig. 1](#) shows an example of a 2 + 1 road in Sweden, the country in which the concept was first implemented. Many rural highways had been built in Sweden with 13-m (43 ft) wide pavements over the period from 1955 to 1980, typically with two 3.5 m (11 ft) lanes and two 3 m (9.8 ft) shoulders. The shoulders were intended to serve as emergency stopping areas, but instead, drivers began to use them for overtaking, or as a place to move over when vehicles were overtaking from the opposite direction. The resulting confusion about what drivers expected from one another led to the decision to implement the 2 + 1 concept. Early trials found not only an improvement in safety but also in operations, with the setup making it easier to overtake other vehicles. As of 2000, more than 1000 km (620 mi) of these roads have been converted to the 2 + 1 concept, all with cable barriers. These 2 + 1 roads have also been implemented in Portugal, Germany (without barriers), Finland, Russia (without barriers), Romania (no barriers), Estonia, Lithuania, New Zealand, and the USA (https://en.wikipedia.org/wiki/2%2B1_road#cite_note-2).



Figure 1 2 + 1 Road near Lund, Sweden. Source: Photo credit J. Ivan

Vehicle Technology Solutions

Many motor vehicle manufacturers have begun to implement systems on their vehicles to partially automate the driving task. The intention for these systems is to assist the driver with routine aspects of the driving task, and especially to catch when the vehicle encounters a potentially hazardous situation. One such type of system is known as lane keeping assist or lane departure alarms. These systems either take over the steering of the vehicle to keep the vehicle within the lane in which it is traveling, unless a turn is signaled, or simply sounds an audible warning when the vehicle drifts out of the lane. Some systems do both.

Sternlund et al. (2017) describe an evaluation of the ability of these kinds of systems to reduce injury crashes in Sweden. They used an induced exposure method to compare crash data for Volvo passenger cars that were equipped with these systems with those that were not using them. The investigated 1853 driver injury crashes from the Swedish Traffic Accident Data Acquisition database (STRADA), including 146 cars equipped with lane departure warning and 11 equipped with lane keeping assist. They found a 53% reduction in head-on and single vehicle crashes under non ice or snow conditions for the vehicles with these systems. Cicchino (2018) performed a similar study in the United States of vehicles equipped with lane departure warning and single-vehicle, sideswipe, and head-on crashes, using crash data from 25 US States for the years 2009–15. They found that vehicles with lane departure warning had significantly lower involvement rates in crashes of all severities (18%), in those with injuries (24%), and in those with fatalities (86%).

Another potential form of driver assistance technology is passing collision warning systems. These systems assist with passing maneuvers on two-lane two-way roads that require use of the oncoming traffic lane. Hassein et al. (2018) describe such a framework and algorithm design using radar sensors to identify how far away other vehicles are and how quickly they are closing in. A simulation in MATLAB was used to develop the algorithms that direct the driver when it is safe to cross into the opposing lane and when they need to complete their maneuver.

Mitigation of Head-on Crash Severity

In addition to preventing head-on crashes from occurring, efforts have been made to reduce the severity of head-on crashes when they unfortunately occur. This started with incorporating front crumple zones into the structural design of passenger cars. This concept was patented by an engineer for Mercedes-Benz, Béla Barényi, originally in 1937 and updated in 1952 (https://en.wikipedia.org/wiki/Crumple_zone). A crumple zone works by absorbing crash energy in the outer parts of a vehicle, thus preventing it from being transferred to the passenger compartment. This reduces deformation of the passenger compartment, mitigating injuries to the occupants. These principles are incorporated into the design of nearly all vehicles today.

Front airbags and restraint systems are another important development in vehicle safety systems. Also known as “supplemental restraint systems,” these are standard equipment in all vehicles today, and deploy upon detection of a front impact on the vehicle above a certain deceleration rate. Due to the explosive nature of airbag deployment it is critical for passengers to wear three-point seatbelts to avoid contact with the air bags while they are deploying. Kuk and Shkrum (2018) compared the occurrence of fatal injuries upon deployment of airbags with and without seatbelt use using data from 110 frontal and offset frontal crashes provided by the Office of the Chief Coroner for Ontario. They found that wearing seatbelts reduced the severity of injuries when airbags deployed, except for abdominal injuries.

As noted previously, crashes between heavy vehicles, especially tractor-trailer trucks, and passenger cars, when they happen, can be extremely serious, much more likely to be fatal than crashes between two passenger cars. In an effort to reduce the severity of such crashes, Friedman et al. (2018) proposed an external radar-activated airbag system that deploys on the front of heavy trucks. They

conducted simulation to demonstrate potential for substantial reduction in damage to passenger cars colliding with a truck that deploys such an external airbag system.

Conclusions

Head-on crashes between motor vehicles are frequently extremely severe due to the physics of how they occur. Numerous infrastructure and technology approaches have been implemented to prevent the occurrence of head-on crashes, with various levels of effectiveness. As well as reducing the occurrence of head-on crashes, efforts have been made to mitigate the severity of head-on crashes when they do occur, most notably in vehicle design improvements.

The most effective way to prevent the occurrence of head-on crashes is to eliminate the possibility of them occurring, for example with physical barriers separating opposing traffic flows from one another. Where that is not practical, it is effective to warn drivers when they are about to stray out of their lane into oncoming traffic, either using audible warning devices in the road or using automatic systems in the vehicle. Continued and more extensive deployment of all of these approaches will be invaluable for reducing the incidence of deaths on roadways worldwide.

References

- Acierno, S., Kaufman, R., Rivara, F.P., Grossman, D.C., Mock, C., 2004. Vehicle mismatch: injury patterns and severity. *Accid. Anal. Prev.* 36, 761–772.
- American Association of State Highway and Transportation Officials (AASHTO). A Policy on Geometric Design of Highways and Streets, seventh ed., Washington DC, 2018.
- Arvin, Ramin, Mohsen, Kamrani, Khattak, A.J., 2019. How instantaneous driving behavior contributes to crashes at intersections: Extracting useful information from connected vehicle message data. *Accid. Anal. Prev.* 127, 118–133.
- Babicineau, S., Fontaine, M., 2018. Examining the safety effect of centerline rumble strips on curves on rural two-lane roads. 97th Transportation Research Board Annual Meeting, Paper 18-05020.
- Belz, N., Chang, K., 2018. Passing zone behavior and sight distance on rural highways: evaluation of crash risk and safety under different geometric conditions, Alaska University Transportation Center and Alaska Department of Transportation, Fairbanks, AK, USA, DOT&PF Report Number 4000143UAF, 2018.
- Cicchino, J., 2018. Effects of lane departure warning on police-reported crash rates. *J. Saf. Res.* 66, 61–70.
- Deng, Z., Ivan, J., Garder, P., 2006. Analysis of factors affecting the severity of head-on crashes on two-lane rural highways in Connecticut. *Transp. Res. Rec.* 1953, 137–146.
- Deogratias Eustace, Mohammad Almothaffar, 2018. Safety Effectiveness evaluation of median cable barriers on freeways in Ohio. The Ohio Department of Transportation, Office of Statewide Planning & Research, State Job Number 135502, December 2018, Final Report.
- Federal Highway Administration (FHWA), 2009. Manual on Uniform Traffic Control Devices, Washington, DC.
- Finley, M., Miles, J., 2018. Closed-course study to assess the conspicuity of wrong-way driving countermeasures. *Transp. Res. Rec.* 2672 (21), 41–50.
- Friedman, K., Mihora, D., Hutchinson, J., 2018. Heavy truck front-end deployable system opportunities for crash compatibility with passenger vehicles. *Int. J. Crashworth.* 23 (2), 161–172.
- Harrison, S., Atabak, S., Cheung, H., 2016. Implementation and evaluation of an innovative wide centerline to reduce cross-over-the-centerline crashes in Queensland. *Roadside Safety Design and Devices: International Workshop 2015*. Transportation Research Circular, E-C215, p. 63–69.
- Hassein, U., Diachuk, M., Easa, S.M., 2018. In-vehicle passing collision warning system for two-lane highways considering driver characteristics. *Transp. Res. Rec.* 2672 (37), 101–112.
- Heslop, S., 2014. Driver boredom: its individual difference predictors and behavioural effects. *Transp. Res. Part F* 22, 159–169.
- Insurance Institute for Highway Safety (IIHS), 1999. Putting the crash compatibility issue in perspective. *Insurance Institute for Highway Safety Status Report* 34 (9), pp. 1–11.
- Jurewicz, C., Sobhani, A., Woolley, J., Dutschke, J., Corben, B., 2016. Exploration of vehicle impact speed – injury severity relationships for application in safer road design. *Transp. Res. Proc.* 14, 4247–4256.
- Kirsch, T., Hamzeie, R., Nightingale, E., Megat-Johari, M., Savolainen, P., 2018. An examination of the cost-effectiveness of median cable barrier systems. 97th TRB Annual Meeting, Paper No. 18-00006.
- Kuk, M., Shkrum, M.J., 2018. Injury patterns sustained in fatal motor vehicle collisions with driver's third-generation airbag deployment. *J. Foren. Sci.* 63 (3), 728–734.
- NHTSA Traffic Safety Facts, 2015. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812384>.
- Oviedo-Trespalacios, O., Haque, M., King, M., Washington, S., 2016. Understanding the impacts of mobile phone distraction on driving performance: a systematic review. *Transp. Res. Part C* 72, 360–380.
- Persaud, B., Retting, R., Lyon, C., 2004. Crash reduction following installation of centerline rumble strips on rural two-lane roads. *Accid. Anal. Prev.* 36, 1073–1079.
- Porter, R., Himes, S., Musunuru, A., Le, T., Eccles, K., Peach, K., Tasic, I., Zlatkovic, M., Tatineni, K., Duffy, B., 2018. Understanding the causative, precipitating, and predisposing factors in rural two-lane crashes, Federal Highway Administration, Technical Report FHWA-HRT-17-079.
- Qin, X., Ivan, J., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accid. Anal. Prev.* 36, 183–191.
- Richards, D., Cuerden, R., 2009. The relationship between speed and car driver injury severity. *Transport Research Laboratory, Department for Transport Road Safety Web Publication* 9.
- Srinivasan, R., Bo Lan, Daniel Carter, Bhagwant Persaud, Kimberly, Eccles, 2017. Safety evaluation of cable median barriers in combination with rumble strips on divided roads, Federal Highway Administration, Report FHWA-HRT-17-070, Washington DC.
- Sternlund, S., Strandroth, J., Rizzi, M., Lie, A., Tingvall, C., 2017. The effectiveness of lane departure warning systems—A reduction in real-world passenger car injury crashes. *Traff. Inj. Prev.* Vol. 18 (2), 225–229.
- Thiffault, P., Bergeron, J., 2003. Monotony of road environment and driver fatigue: a simulator study. *Accid. Anal. Prev.* 35, 381–391.
- Transportation Research Board (TRB), 2003. Application of European 2 + 1 Roadway Designs, National Cooperative Highway Research Program, Research Results Digest, No. 275.
- Wali, B., Khattak, A., Xu, J., 2018. Contributory fault and level of personal injury to drivers involved in head-on collisions: application of copula-based bivariate ordinal models. *Accid. Anal. Prev.* 110, 101–114.
- Rumble bars - Wu, L., Geedipally, S., Pike, A., 2018. Safety evaluation of alternative audible lane departure warning treatments in reducing traffic crashes: an empirical bayes observational before-after study. *Transp. Res. Rec.* 2672 (21), 30–40.
- Xian, J., Muramatsu, T., Goto, H., Yamaguchi, D., 2019. Impact evaluation of wire rope installation on two-way two-lane expressways. *Transp. Res. Rec.* 2673 (2), 637–649.
- Xu, J., Luo, X., Shao, Y., 2018. Vehicle trajectory at curved sections of two-lane mountain roads: a field study under natural driving conditions. *Eur. Trans. Res. Rev.* 10, 12.

Helicopters in Emergency Medical Response

Stephen J.M. Sollid, M.D., PhD, Chief Medical Officer, Norwegian Air Ambulance Foundation, Oslo, Norway; Associate Professor, University of Stavanger, Faculty of Health Sciences, Stavanger, Norway

© 2021 Elsevier Ltd. All rights reserved.

The History of Life Saving Helicopters	316
Do Helicopters Save Lives?	316
Safety Record of Helicopter Emergency Medical Services	317
Pushing the Limits	317
Technical Limitations and Improvements	318
Developing Technologies	318
Balancing Flight Operations Decision Making With Medical Mission Needs	319
Fatigue	319
The Future Challenges of the Industry	320
Biography	320
Relevant Websites	320
References	320
Further Reading	321

The History of Life Saving Helicopters

The care and management of sick and wounded has always involved an element of transport. The sick and wounded must often be moved to allow for better care, and high-quality care can be brought to the patient. Choosing the right means of transport depends on many factors and can sometimes pose a challenge of its own.

Since the dawn of air transport, aircrafts have been used for medical transport to a variable degree. The first confirmed aeromedical transport dates back to 1915 and the transport of a wounded Serbian air force lieutenant from the frontline in Albania to safety. In the years to follow aeromedical transports by fixed wing aircraft occurred occasionally in different settings but often ad hoc and with a certain flare of heroism. The first organized service was established in 1928 and still exists today as the Royal Flying Doctor Service in Australia.

As with the first known transport in 1915, it was often armed forces and military conflicts that drove the development of aeromedical transport forward. The Korean War in the early 1950s marks a paradigmatic shift in that it introduced aeromedical transport by helicopter as a new and effective concept in the evacuation of wounded servicemen. This concept proved highly successful as it allowed airborne evacuation of wounded without being dependent on airfields. In the Vietnam War the addition of specially trained medical corpsmen added further improvement in survival rates. Contemporary research showed that wounded soldiers had better survival rates than civilian car accident victims on highways back in the United States, thanks to helicopter evacuation and the specially trained medics.

In other parts of the world the use of helicopters in air rescue grew out of the need for getting access to injured people in austere environments or remote areas. The Swiss Air Rescue was established as early as 1952 and was a merger of mountain rescue and aviators. The Swiss concepts of bringing care to the patient, wherever they are, developed quickly over the next years and soon became the inspiration and eventually the norm for similar initiatives in other European countries. In the 1970s, both in North America and Europe, several services of helicopter rescue were established, based on the experiences and early initiatives described earlier.

Following the early 1970s the number of helicopter emergency medical services (HEMS) grew rapidly in Europe and North America. Today HEMS is a part of the pre-hospital care system in all parts of the world and in most countries. In the United States, there are around 75 companies operating more than 1500 helicopters, and in Europe all but some of the minor countries have a HEMS in some shape or form. Let alone Great Britain has 22 different HEMS systems. The organization, staffing, and concepts vary greatly around the world. It seems, however, that no matter how HEMS is organized it plays an important role in pre-hospital care of critically ill and injured patients (Ringburg et al., 2009).

Do Helicopters Save Lives?

The early intention for establishing HEMS was to bring the patient quickly to hospital. Advanced care was only available in hospitals and quick air transport was superior to slow ground transport for the critically ill or injured. This was especially true in trauma patients, where the golden hour of trauma has been the norm for years; getting the severely injured patient to the hospital within the first hour is regarded as imperative to avoid the immediate pathophysiologic consequences of trauma and save lives. In Germany one

of the main reasons for establishing HEMS in the early 1970s was to have an extensive network of HEMS that could quickly respond to road vehicle accidents on the accident-stricken autobahn and transport the patient to a nearby trauma hospital.

With improved medical knowledge and technology, and better prevention of both causes and consequences of trauma and disease, previous lethal conditions are avoided or taken care of more effectively. The increasingly high competence of ground emergency medical services (EMS) also contributes to taking better care of patients before they arrive in hospitals. The impact and importance of the transport element in HEMS is therefore not as prominent as in the early days of HEMS. The effect that quick transport alone has on patient survival is also highly questioned in the medical community, especially after some studies have shown that trauma patients transported by civilian cars have better survival than those transported by ground EMS in some urban settings.

The positive effect HEMS has on survival in critically ill or injured patients is however well proven even in later years ([Andruszkow et al., 2016](#)). Denmark is a small country with an excellent infrastructure. For many years, the Danish pre-hospital system relied on ground EMS supplemented by physician-staffed ambulances that covered all of Denmark. The general opinion was that HEMS was superfluous in Denmark until a pilot project with a Danish HEMS showed a benefit of HEMS in certain patient groups, even in a system with excellent ground EMS cover and short transport distances to hospitals ([Hesselfeldt et al., 2013](#)). The effect is presumed to be found in quickly bringing high medical competence out to the patient in combination with quick transport to the hospital. Other studies have also pointed at the medical competence on-board as the one life saving factor that makes a difference between ground EMS and HEMS ([Lossius et al., 2002](#)).

The general opinion today is that helicopters alone do not save lives just because they represent a quick transport element. Helicopters with high medical competence onboard save lives by bringing hospital standard care quickly to the patient.

Safety Record of Helicopter Emergency Medical Services

The safety record of HEMS is tightly coupled with the safety record of the aviation industry per se. However, since the HEMS industry developed late as compared to the development of airplanes and helicopters, the technical safety standard has always been good, and very rarely have we seen accidents in HEMS caused by technical malfunction. Add to this that HEMS is often a phenomenon in high resource countries, the technical standard of the HEMS fleet is generally very good.

Pushing the limits or operating beyond the limitations of the flight equipment is, however, a more common cause that can be easily mistaken for a technical cause but is in essence a human factor issue. The two main reasons for pushing the limits seem to be the desire to do all possible to help, or to maximize profit in a system where the business plan is dependent on payment per mission or restricted by a tight budget.

Pushing the Limits

On November 24th 1987 a BO105C from the Norwegian air ambulance crashed in the Dovrefjell mountains in South Norway. The crew had been called to a mission to transfer a critically ill patient from one hospital to another in the evening after dark. Despite darkness, snow and poor visibility the crew accepted the mission and took off. Less than 10 min after take-off, the helicopter crashed in the mountains close to the base and all three crewmembers perished in the immediate fuel fire following the crash. The Accident Investigation Board concluded that the combination of darkness, snow showers and fog caused the pilot to lose visual references and eventually lose control of the aircraft. The use of strong search lights to aid orientation only contributed to the confusion in the snowy weather. Interestingly, the Accident Investigation Board did not comment on the obvious fact that the crew had accepted a mission despite poor flying conditions. In 1989 and 1991 the same HEMS base experienced two more accidents, one with fatal outcome for both the crew and the patient. In both accidents poor visibility was a contributing factor causing whiteout in the 1989 accident and collision with power lines in the 1991 accident. In the report of the Accident Investigation Board of the 1991 accident it was recommended that the company must pay attention to and follow up on attitudes of the crews on the HEMS base. The report also cited that the pilot had a "relaxed attitude towards rules and regulations that he himself did not regard as pressing." In later years it was confirmed by other crewmembers working at the same HEMS base that the general attitude was to have a low threshold for accepting missions even in poor flying conditions. The reason for this was the desire to justify the existence of the HEMS base by carrying out as many missions as possible. The operating company later introduced stricter regulations on flight operations minima to avoid similar accidents.

In the United States, the National Transportation Safety Board issued a special investigation in 2006 to address the high number of accidents in aviation emergency medical services, including HEMS and recommended operational strategies and technological improvements that would improve the safety record of the industry ([National Transportation Safety Board, 2006](#)). The three previous years before the report had seen an unusual high number of accidents in the HEMS industry with 41 crashes and 39 fatalities. Besides concerns about lack of technological equipment like terrain awareness and warning systems, the report addressed less stringent requirements for HEMS operations without patient on-board, lack of aviation flight risk evaluation programs and inconsistent flight dispatch procedures as recurring themes in the accidents investigated. In essence, the desire to provide medical care seemed to contribute to pushing the limits of the services, and internal procedures, risk management systems and equipment were not robust enough to compensate.

The high number of HEMS accidents in the United States also resulted in a considerable effort from the Federal Aviation Administration (FAA) that has resulted not only in better regulations but also better dialogue and cooperation between the

governing agencies and the industry. On reviewing all HEMS accidents in the United States between 1991 and 2010, the FAA concluded that 62 accidents could have been avoided with current regulations and industry standards potentially having saved 125 lives lost. A strong message from the FAA is that aviation safety decisions are separate from medical decisions (https://www.faa.gov/news/fact_sheets/news_story.cfm?newsId=15794). Accepting the medical need of a mission does not automatically mean accepting the mission. Flight safety cannot be compromised just because the mission has a high acuity or is of special value (e.g., pediatric trauma). Flight safety has precedence in all circumstances and the decision to fly or not must always be unaffected by the scope of the medical mission.

Technical Limitations and Improvements

The previous statement that technical issues have rarely been the cause of accidents in HEMS holds true but, in some cases, not adapting to new safety features or technical improvements in helicopter hardware have been known to cause accidents and near misses. There is still an ongoing discussion in HEMS communities on whether single engine operations are adequate or if dual engine operations should be the standard. In most industrialized countries, dual engine helicopters are the norm and, in some cases, even required for HEMS, but even in the United States there are still HEMS companies that operate with single engine aircraft. The differences between single engine and dual engine aircraft from a safety perspective are primarily in the aircraft's ability to perform safely during take-off and landing and handle engine failure in any phase of flight. Single engine aircraft are involved in the majority of HEMS accidents in the United States, today with around 75%, and this share has increased although the number of single engine aircraft in the HEMS service has decreased in the United States. This is a clear indication that dual engine aircraft are safer and should be the norm in HEMS.

Night vision goggles (NVG) were introduced in HEMS over the last 20 years to enhance operations during hours of darkness. Interestingly though it has never been shown that the number of accidents is higher during hours of darkness than during daylight. This might be contributed to the fact that not all HEMS systems operate during hours of darkness and that those who do have a higher vigilance during night flight. In this context, introducing NVG in HEMS has not contributed to reducing accident rates but has probably contributed to increasing safety margins during night flight. It may also have contributed to improving the regularity of HEMS in that NVG allows for better visual control and decision making during flight. Some pilots report that where they previously would have canceled or aborted a mission because of darkness, they now accept and continue because they have better visual control. Philosophically this poses a challenge since NVG in civilian HEMS was not supposed to push the limits of operations further but to enhance safety within existing operations. This is similar to an often-cited dilemma; do you drive faster with your seat belt on than you would without? So far, NVG does not seem to have influenced accident or incident statistics in any direction but it probably had a positive impact on regularity. Some services that previously operated only under daylight have been able to commence operations during night, thanks to NVG, thus improving the public's access to the service. HEMS operators must be attentive to the impact NVG operations have on their service and whether it increases safety margins by adding an additional safety layer to night operations or maintain safety levels by increasing regularity.

The implementation of other technical improvements in HEMS is slower than in for example, commercial fixed wing operations. HEMS operation is very often a weight versus range question where more technical equipment induces increased aircraft weight and thus reduced range. Coupled with the desire to maintain the aircraft small to allow access to patients this reduces the ability to adopt all possible and available technical improvements. New technical devices that enhance safety must therefore be weighed against their effect not only on safety but also on their impact on patient care and patient safety. With improved engineering, some equipment has become smaller and lighter thus allowing easier implementation.

Glass cockpit and electronic map systems are increasingly being implemented in new aircraft. Improved GPS technology allow for better navigational systems and autopiloting. New auto hover systems are available with zero GPS groundspeed, which is an excellent improvement to hoisting and search and rescue missions. Other technologies, such as weather radar, terrain awareness and warning systems, and traffic collision avoidance systems are also increasingly being implemented in new HEMS aircraft.

Developing Technologies

In 2014, an EC135 of the Norwegian air ambulance hit a power line on descent to an accident scene. The impact occurred at low air-speed and the installed wire cutter in the front of the aircraft did not perform as intended, but instead stretched the wire further until the main rotor impacted the wire and cut it. The wire snapped back at the aircraft damaging the main rotor pitch links thus the main rotor was in principle detached from the engine. This caused the aircraft to lose lift and fall 25 meters killing two of three crewmembers. The accident revealed an insufficient marking of all wires and lines that pose a threat to HEMS aircraft, in the available electronic map system. It also revived a demand for technology that can identify powerlines and other similar obstructions. Lines and wires are a highly relevant hazard in many HEMS systems especially in mountainous areas and have been known to cause accidents similar to the 2014 accident in Norway. The technology to identify wires is developing but still has not reached a reliable sensitivity to be implemented. In the meantime, map systems need to be accurate enough to contain all wires and threats to HEMS aircraft. This often requires manual tracking and logging that can be quite labor intensive but at the same time a necessity to maintain safe operations.

Icing is a challenge for HEMS, especially in the polar regions and also in tempered climate zones during cold season and can pose a severe restriction to operations in many areas especially during winter (Pulkkinen et al., 2019). The accumulation of ice on the

rotor during flight severely reduces the airfoil and efficiency of the rotor but the added weight of the accumulated ice on the airframe also impairs flight. In some cases, loss of view through the windshield has also been reported as a severe effect of sudden icing. Only very few accidents in HEMS have been directly related to icing. But icing can also be difficult to document post-accident. A Bell 407 operated by Mercy Air crashed near Mason City, Iowa in the United States in 2009 and icing is presumed to have been a major contributor to the accident. The airframe was consumed by fire following the crash and ice accumulation on the airframe could not be documented. However, reports from eyewitnesses at the scene confirmed that ice had accumulated on cars in the area and that roads were icy.

In risk areas the occurrence of icing is occasionally reported during flight and can often be solved by either landing or changing altitude. Being a severe hazard, it predominantly poses a restriction to service that affects the patient. For many years, deicing systems have been a standard feature of fixed wing aircraft and have eventually found its way to rotor wing aircraft. The rotor blades are the most vulnerable part of the helicopter in icing conditions and typically electrical systems have been implemented that heat up the leading edge of the rotor. These systems are however restricted to larger aircraft since they require relatively large and weight demanding hardware to operate. Most HEMS operate with small and middle-sized helicopters where this technology has not been implemented or is in a stage of development. Since the need for this technology is restricted to a very small proportion of the total number of aircraft produced, the helicopter manufacturers have also not prioritized the development of deicing systems for smaller aircraft. With developing technology and focus on demand, we should expect the development of better solutions for helicopter deicing in the future.

Balancing Flight Operations Decision Making With Medical Mission Needs

As stated previously, flight safety considerations and flight safety decisions should be made independent of the nature of the medical mission request or the development of a medical mission. On paper, this is reasonable but as history and experience has shown, this can be challenging.

HEMS crews mostly consist of one or two pilots and a medical crew. The interaction between the flight crew and the medical crew has different dynamics depending on company culture and standards.

Some companies have chosen to maintain a strict barrier between flight operations and medical operation, for example, keeping the pilot shielded from the medical mission planning. The pilots are not aware of the nature of the mission, the acuity or level of urgency and only concerns themselves with flight mission planning and execution. During flight the pilots are kept out of medical communication and do not monitor the development in the medical cabin. Some companies have taken this to the extent of physically shielding the cabin from the cockpit. However, shielding the pilots completely from the mission is not always possible and the stress of not knowing what is going on has (by few) been described as just as stressful as being involved.

In other HEMS systems members of the medical crew have flight operations duties, for example, to support the pilot during flight operations. In some countries, especially in Europe this is the core of an “integrated crew concept” where a HEMS paramedic acts as a pilot assistant during flight, and as the HEMS doctor’s assistant on ground. These HEMS paramedics have medical training but also flight operations training to enable them to assist the pilot in navigation during instrument flight rules (IFR) and night operations. This is a cost- and resource-effective way to run HEMS, as it not only avoids the need for two pilots, but also poses some potential conflicts of interest during a mission. The involvement in medical care of a critically ill or injured patient can be emotionally stressful and has the potential to influence decision making during flight. This influence is not restricted to the HEMS paramedic but also extends to the pilot. Especially in the integrated crew concept the pilot is expected to be a resource in missions on ground as long as it does not interfere with flight operations and -planning. This puts extra demands on the operating company to make sure that the medical aspects of the mission do not interfere with decision-making pre- or in-flight. Company training and culture must therefore be targeted toward this potential conflict and hazard. In Europe, the European HEMS and Air Ambulance Committee (EHAC) has developed a tailored crew resource management concept to meet this challenge; the aeromedical crew resource management course (<http://www.ehac.eu/acrm.html>). This crew resource management (CRM) course adheres to the European Aviation Safety Agency requirements for CRM training but adds the medical aspects of operations to address this factor in the decision-making and human factors.

Fatigue

Fatigue from lack of sleep or long working hours have been known to cause accidents in aviation. Fatigue causes reduced attention and reaction and can be a crucial component in an emergency situation during flight. The impact of fatigue on aviation safety has led to strict regulations on working hours and workload during duty to ensure that flight crews are rested and fit for flight. HEMS is especially vulnerable to fatigue, as there are large differences in workload in HEMS. Missions occur without warning and involve an unpredictable workload. The saying “hours of boredom and moments of terror” describes the workload or lack thereof in HEMS well. High workload can have just as much effect on the level of fatigue as long hours of standby and boredom. The aviation authorities in most countries dictate how HEMS can operate and the regulations are in most cases the same as for general civilian aviation. Whereas many HEMS systems restrict duty hours to 12-h shifts, some have longer periods of duty, in some cases up to a week. This is especially the case in remote locations or in low activity HEMS systems. In these systems, mandatory rest is imposed after a certain number of active working hours in a 24-h cycle or longer (Zakariassen et al., 2019). Recent research, however, suggests that counting only hours is not sufficient to avoid fatigue, for example, working at night time has a stronger effect on fatigue than

working during day time. Adding to this, humans are generally poor at assessing their own level of fatigue and often underestimate how tired they are. Preliminary results suggest that the effect of disturbed night sleep every night over a number of days induces a level of fatigue that requires at least 24-h of rest to achieve normal levels of attention and reaction. As a result, HEMS services that wish to maintain long on-call duty hours must implement fatigue risk management (FRSM) systems to closely monitor and act on fatigue. Crews are required to monitor themselves and have a low threshold for reporting own fatigue and act for example, going off duty or taking naps, resting, and avoiding other tasks besides mission. Part of the FRSM is also to educate and train crews to monitor their own fatigue and recognize when fatigue sets in. HEMS companies in many countries are currently challenged with finding the right balance between continuity in duty that affects mission readiness the least, and avoiding work schedules that have the potential to induce fatigue.

The Future Challenges of the Industry

The technological development in aviation is moving forward and new safety features and safety technology to both monitor and automate flight are being increasingly implemented. The rate of implementation is, however, strongly connected to the HEMS company's ability and willingness to implement new features. HEMS cannot regulate income the same way as civilian aviation by regulating airfares and often have to justify that new technology has a positive impact on patient care before air safety is addressed. Accidents are often the strongest driver to improve safety in HEMS not the prospective attitude to make flight operations as safe as possible at all times.

On the positive side, increasing regularity in HEMS seems a strong driver to implement new technology as it achieves better access to patients. GPS based points-in-space approaches are developed in many countries allowing HEMS to land at hospitals much the same way as passenger airplanes land on airports in poor weather conditions. IFR routes are increasingly implemented for HEMS specifically to allow flight between hospitals in poor weather and new technology is being developed to even allow for ad-hoc IFR landings on the scene. Whilst these new technologies improve regularity, they also push the boundaries of operations and can potentially contribute to putting HEMS in situations where the safety margins are very small and highly dependent on technology to stay safe.

Biography



Stephen J. M. Sollid is a medical doctor specialized in anesthesiology. He has worked as an HEMS physician in Norway since 2002 and is currently the chief medical officer of the Norwegian Air Ambulance Foundation. He has a PhD in patient safety and risk management from the University of Stavanger, is an Associate Professor of Prehospital Critical Care at the same university and teaches prehospital critical care at master and bachelor level at the university. He is involved in research on prehospital patient safety with over 30 peer-reviewed publications, and mentors several PhD and MSc projects.

Relevant Websites

Federal Aviation Administration (FAA) fact sheet on initiatives to improve helicopter air ambulance safety, publisher in February 2014:

https://www.faa.gov/news/fact_sheets/news_story.cfm?newsId=15794

European HEMS and Air Ambulance Committee (EHAC) information in Aeromedical Crew Resource Management course:

<http://www.ehac.eu/acrm.html>

References

- Andruszkow, H., et al., 2016. Impact of helicopter emergency medical service in traumatized patients: which patient benefits most. *PLoS One* 11, e0146897.
Hesselfeldt, R., et al., 2013. Impact of a physician-staffed helicopter on a regional trauma system: a prospective, controlled, observational study. *Acta Anaesthesiol. Scand.* 57, 660–668.

- Lossius, H.M., et al., 2002. Prehospital advanced life support provided by specially trained physicians: is there a benefit in terms of life years gained. *Acta Anaesthesiol. Scand.* 46, 771–778.
- National Transportation Safety Board, 2006. Special investigation report on emergency medical services operations, Special Investigation Report NTSB/SIR-06/01. Washington, DC.
- Pulkkinen, I., et al., 2019. Impact of icing weather conditions on the patients in helicopter emergency medical service: a prospective study from Northern Finland. *Scand. J. Trauma Resusc. Emerg. Med.* 27, 13.
- Ringburg, A.N., et al., 2009. Cost-effectiveness and quality-of-life analysis of physician-staffed helicopter emergency medical services. *Br. J. Surg.* 96, 1365–1370.
- Zakariassen, E., et al., 2019. Causes and management of sleepiness among pilots in a Norwegian and an Austrian air ambulance service—a comparative study. *Air Med. J.* 38, 25–29.

Further Reading

- Taylor, C., et al., 2012. The cost-effectiveness of physician staffed Helicopter Emergency Medical Service (HEMS) transport to a major trauma centre in NSW, Australia. *Injury* 43, 1843–1849.

Horizontal and Vertical Geometry

Victoria Gitelman, Transportation Research Institute, Haifa, Israel

© 2021 Elsevier Ltd. All rights reserved.

Introduction	322
Safety Impacts of Horizontal Curves	323
Curve Radius	323
Transition Curves	324
Superelevation	325
Bendiness	326
Safety Impacts of Vertical Alignment	327
Grade	327
Interactions Between Horizontal and Vertical Alignments	328
Evaluating Speeds	329
Future Research	329
References	329
Further Reading	330

Introduction

Knowledge about the relationship between road design and safety is continually evolving. Horizontal and vertical road geometry was always in the focus of the road design process due to its obvious implications on vehicle behavior stemming from the laws of physics. It is generally assumed that driving on a straight and flat road section is easier, while negotiating curves requires more driver efforts. Thus, road design guidelines have typically included sets of limitations on the extent of horizontal and vertical road curvature, related to road function and the design speed. Initially, the criteria for setting horizontal and vertical alignments of the roads were defined based on engineering judgment and life experience. The original approach relied on assumptions based on vehicle kinematics and drivers' ability to react to changing road conditions, and aspired to avoid extreme road design conditions that would increase the risk of loss-of-control of the vehicle and, hence, crash occurrences. Over the last decades, substantial developments took place in crash data analyses and evaluating safety effects of various road design conditions, providing quantified measures of the safety impacts of road geometry characteristics.

This paper summarizes current knowledge on safety impacts of horizontal and vertical road alignments. The horizontal alignment of a road comprises straight sections and horizontal curves that provide connections between the straight sections (tangents). Road design literature distinguishes between simple (or circular) curves, with a constant radius, and other forms of curves that combine various radii. The radius of a curve is the main characteristic of horizontal curvature that is examined in safety evaluations (other measures of curvature can be deflection angle, degree of curve, etc.). In addition, the basic concept of a horizontal curve considers forces acting on a vehicle negotiating a superelevated curve. According to the laws of mechanics, when a vehicle travels on a curve it is forced outward by a centrifugal force. This force is resisted by a superelevation force and a transverse friction force, which depend on vehicle weight and speed, side friction between the tires and the pavement, and superelevated structure of the road. Hence, a superelevation parameter can be among additional curve characteristics to be examined in the road safety context. Moreover, crash frequencies in a curve can be influenced not only by the characteristics of the curve itself but also by road characteristics in general, for example, length of a road tangent prior to the curve or overall road bendiness, which should also be examined in safety evaluations.

The vertical alignment of a road consists of straight segments (flat or inclined) connected by crest or sag vertical curves. The grade percent or grade difference are usually used to characterize the impact of vertical alignment on safety. A distinction is applied between a negative (or downhill) grade and positive (or uphill) grade. It is generally agreed that steep downhill grades increase vehicle acceleration and speed, and thus make a vehicle harder to control precisely, especially if it has a lot of inertia and increased handling difficulty, such as heavy goods vehicles (trucks). The situation can be particularly dangerous when a steep grade coincides with sharp horizontal curves, especially toward the bottom of the slope. Conversely, steep uphill grades induce additional deceleration, especially for heavy vehicles, which increases differentiation of speeds and crash risk. Thus, steep grades are often considered negatively in design, and most design manuals recommend avoiding or keeping to a minimum the use of steep slopes.

Concerning the estimation of safety effects of road design characteristics, best current knowledge on the topic is given in the form of crash modification factors (CMFs) or crash modification functions, which are based on the analyses of empirical data. A CMF indicates a relative change in crash frequency associated with a change in the design feature, that is, it represents the ratio of expected crashes with the new design element to crashes without it, assuming other road and traffic characteristics unchanged. CMFs less than 1.0 indicate that the design change is associated with safety improvement (fewer crashes).

The before–after study would be the most appropriate technique for quantifying the influence of design change on safety. However, due to the need for substantial crash counts to provide reliable estimates of CMFs and thus a consideration of many sites with a single change of the design feature, which is seldom possible in practice, most safety effects of road geometry characteristic are estimated by fitting multivariate regression models (cross-section study). Such a model predicts average crash frequency for sites with various values of the geometric feature of interest having controlled for traffic volume, segment length, and other road design characteristics.

It should be noted that the majority of current research findings with regard to safety impacts of horizontal and vertical road alignments belongs to rural (nonurban) roads and, particularly, two-lane roads, while similar results for other road types are not frequent. For example, the Highway Safety Manual in the United States (AASHTO, 2010) suggests no CMFs for horizontal or vertical curvature on rural multilane undivided and divided highways, and urban and suburban arterials.

Safety Impacts of Horizontal Curves

Typically, the crash risk in horizontal curves is greater than on straight sections. According to various estimates, the crash rate in curves is 1.5–4 times higher than in tangents, the severity of crashes in curves is high (25%–30% of all fatal crashes occur in curves), and a large proportion of crashes (60%) are single-vehicle off-road crashes (RSM, 2003). Increased crash rates are observed on horizontal curves due to the limited sight distance and the increased probability of skidding or rollover.

Curves also present road safety problems for specific types of road users. For example, motorcyclists are particularly at risk on curves, where acceleration and deceleration occur and the stability of the vehicle can be compromised. As reported (FEMA, 2012), riders are 15 times more likely to be killed than car occupants in collisions on curves (especially, when steel barriers without motorcycle protection devices are present on the roadsides). Trucks and other large vehicles are more susceptible to overturning at curves, because of their higher center of gravity. Research indicates that such overturning can occur at speeds only slightly greater than the design speed of the curve.

Curve Radius

The effect of curve radius has been extensively studied in many countries (Bonneson and Pratt, 2009; Elvik, 2013; Goldenbeld et al., 2017a). The results were consistent in showing that curves with shorter radii are associated with higher crash rates, but the results may be sensitive to other curve characteristics, for example, length of curve. In addition, the extent of the impact may change depending on the country (actually, the road design practice), road type, crash type, and other evaluation conditions. Across various models, the strongest increase in crash rates was observed for radii below 200 m. The tendency of the crash rate to increase with shorter curve radii was found not only for two-lane rural roads but also for multilane roads and access-controlled roads in rural and urban environments.

Fig. 1 presents examples of CMFs for horizontal curve radius in the United States, based on the widely used safety relationship for predicting total crashes on two-lane roads (Fig. 1A), and on the relationship between curve radius and injury (plus fatal) crash frequency on rural highways that was developed in Texas (Fig. 1B). The base condition for these CMFs is a tangent roadway. The horizontal curve radius CMF (CMF_{cr}) is estimated as:

In the case of a common US relationship for two-lane roads (AASHTO, 2010)

$$CMF_{cr} = \frac{1.55L_C + \frac{80.2}{R} - 0.012S}{1.55L_C} \quad [1]$$

in the case of rural highways in Texas (Bonneson and Pratt, 2009)

$$CMF_{cr} = 1.0 + 0.97(0.147V)^4 \frac{(1.47V)^2}{(32.2R^2)} \left(\frac{L_c}{L} \right) \quad [2]$$

where R —curve radius, feet; L_c —horizontal curve length, mile; L —segment length, mile; V —speed limit, mph; and S —the presence of spiral transitions ($S = 1$ if exist). Clearly, beside radius values, CMF_{cr} depends on other curve characteristics, for example, curve length, tangent length, and the presence of transitions. Fig. 1A indicates, for example, that shorter (sharper) curves are associated with higher crash numbers.

Fig. 2 depicts the summary crash modification function for a horizontal curve radius that was estimated by weighting the results from eight countries (Elvik, 2013). The function is:

$$\text{Relative crash rate} = 126.1658x^{-0.7099}, \quad [3]$$

where x is curve radius, in meters. The international average is limited to the upper radius of 1 km and is not suitable to account for other curve characteristics. However, it exhibits that the form of the relationship was similar in various studies, with a stronger risk of small radii.

For motorcycle crashes on curves, the impacts of small radii are even stronger. Prediction models (for motorcycle-to-barrier crashes, in the United States) showed that curves with radius of 820 ft. (250 m) or less increased the crash frequency rate by a factor

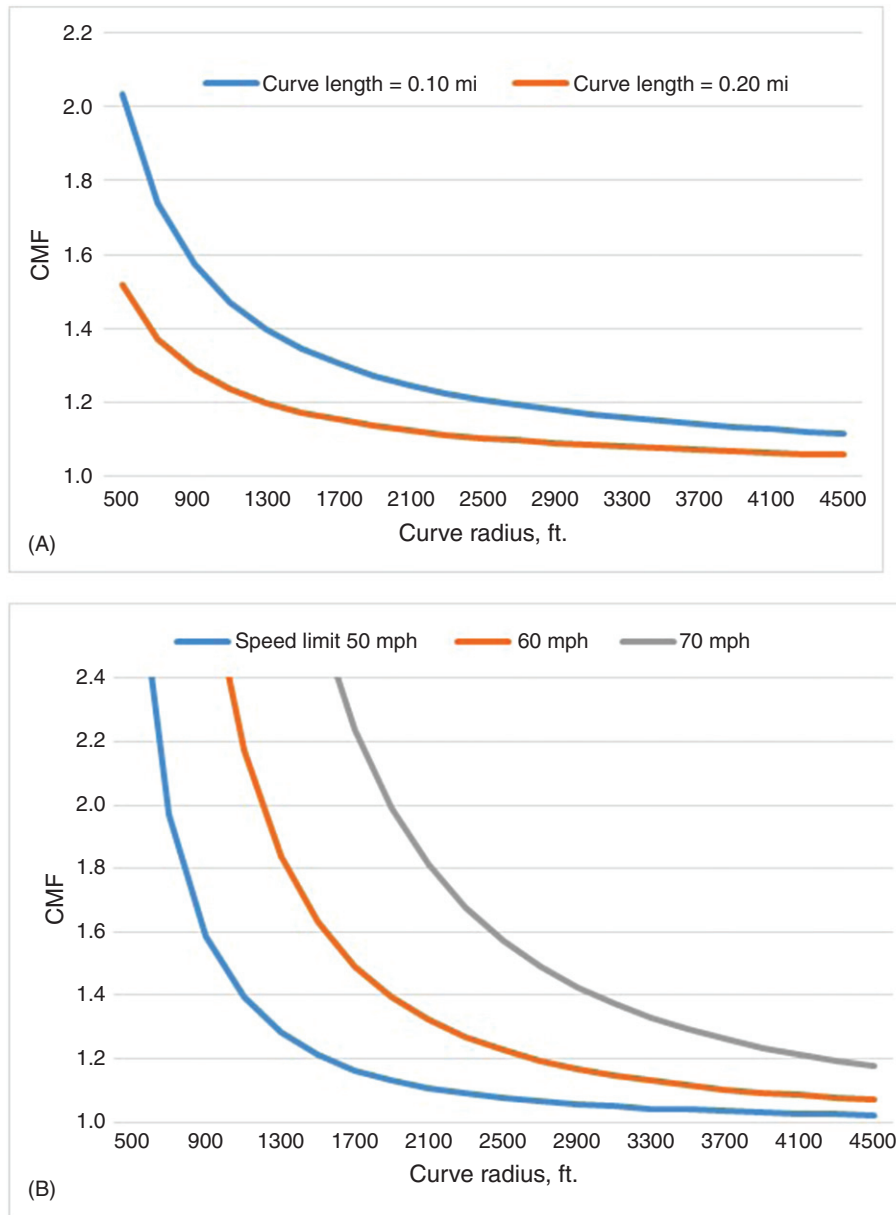


Figure 1 Examples of horizontal curve radius CMFs. (A) For crashes on two-lane highways, in the United States (cases with no spiral transitions). (B) For injury crashes on rural highways, in Texas (cases when curve length = segment length). Source: (A) Produced based on *American Association of State Highway and Transportation Officials (2010)* and (B) produced based on *Bonneson and Pratt (2009)*.

of 10 and curves with radius less than 500 ft. (150 m)—by more than 40 times relative to curves with a radius in excess of 2800 ft. (over 800 m) (*Gabauer and Li, 2015*).

Transition Curves

Transition curves (or spirals) have regular radius changes, allowing for a gradual transfer between adjacent road segments with different curve radii. A gradually changing radius of the transition curve is supposed to be consistent with the natural path drivers follow as they steer into, or out of, a horizontal curve. Experience shows that if the transition from a tangent section to a circular curve is not achieved by a transition curve, there is greater risk of drivers making abrupt movements in order to negotiate the curve.

Fig. 3 shows examples of transition curve CMFs for two-lane highway curves with a radius of 500 ft. (150 m) or more (developed in Texas) (*Bonneson and Pratt, 2009*). The base condition for this CMF is that spiral transition curves are not present. The impact of

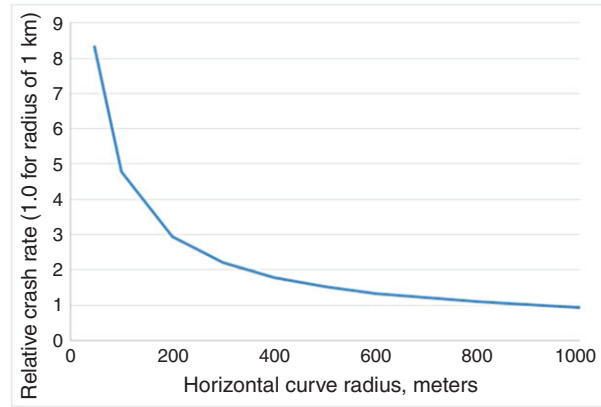


Figure 2 Summary crash modification functions for horizontal curve radius. Source: Produced based on [Elvik \(2013\)](#).

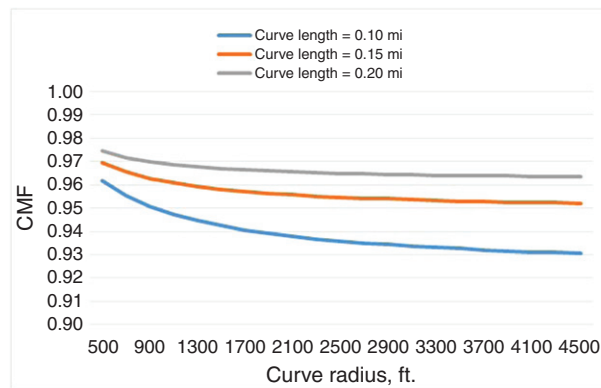


Figure 3 Examples of transition curve CMFs for two-lane highway curves (cases when curve length including spirals = segment length). Source: Produced based on [Bonneson and Pratt \(2009\)](#).

transition curves depends on horizontal curve length (L_c), segment length (L), and curve radius (R) as is given in the equation for spiral curve CMF (CMF_{sp}):

$$CMF_{sp} = 1 - \frac{0.012}{1.55L_c + \frac{80.2}{R}} \left(\frac{L_c}{L} \right) \quad [4]$$

The results indicate ([Fig. 3](#)) that safety benefits of transition curves are most notable for horizontal curves with larger radii but shorter in length (relatively sharp).

In general, safety impacts of the presence or absence of transition curves were not extensively examined in the literature ([Goldenbeld et al., 2017b](#)). Crash studies on the topic were mostly conducted in the United States; more recently, simulation and naturalistic driving studies, in various countries, explored driving behaviors while passing spiral transitions, in terms of speed, lateral acceleration, keeping desired vehicle trajectory, or similar indicators. The findings were not always consistent but sensitive to road conditions, for example, type of terrain, road width, length of spiral curves, etc. The findings generally suggest that the presence of transition curves is associated with better driving performance and lower crash rates, in flat terrain. However, some studies found negative safety effects of transition curves in rolling or mountainous area, that is, an increase in crashes.

In some studies, higher driving speeds were observed in the presence of transition curves; yet, the lateral acceleration rates were not affected in most cases. Most observation studies showed that spiral curves improved vehicle trajectories and/or resulted in less steering corrections. Hence, implementing transition curves is expected to have a positive effect on road safety (but dependent upon external factors). It should be noted that, overall, the safety effect of transition curves is not strong, indicating a crash change of a few percent only ([Fig. 3](#)).

Superelevation

Superelevation is provided on horizontal curves to offset some of the centrifugal force associated with curve driving. It reduces side friction demand and increases the margin of safety relative to vehicle slide out or rollover. It is assumed that if the superelevation provided on a curve is significantly less than the amount specified by road design guidelines, then the potential

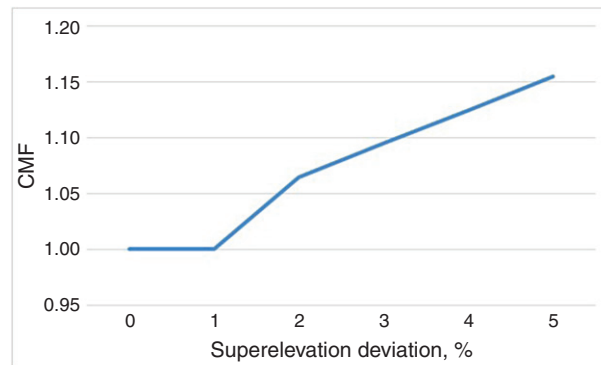


Figure 4 Superelevation CMFs for total crashes. Source: Produced based on *American Association of State Highway and Transportation Officials (2010)*.

for a crash may increase. The amount of the superelevation deficiency can be measured as a difference between the superelevation rate that is specified in the design guideline and the actual superelevation of the curve. This measure is known as superelevation variance (SV, ft.ft.⁻¹), for example, in the US Highway Safety Manual, or superelevation deviation, in percent, in other sources.

The safety effects of deficient superelevation on horizontal curves were mainly reported in US studies; reduced crash rates in curves with improved superelevation were also found in Australia. Eq. (5) shows a commonly used relationship for superelevation CMFs on horizontal curves of rural two-lane highways (CMF_{sv}):

$$\text{CMF}_{\text{sv}} = \begin{cases} 1.00, & \text{for } \text{SV} < 0.01, \\ 1.00 + 6(\text{SV} - 0.01), & \text{for } 0.01 \leq \text{SV} < 0.2 \\ 1.06 + 3(\text{SV} - 0.02), & \text{for } \text{SV} \geq 0.2 \end{cases} \quad [5]$$

where SV is superelevation variance as defined earlier. The base condition for this CMF is a horizontal curve with superelevation deficiency within 0.01 ft.ft.⁻¹ related to the recommended design values.

Fig. 4 provides a visual presentation of the relationship between superelevation deviation and total (or injury) crashes. It shows that improving the superelevation may reduce the crash numbers on curves by 5%–15%.

Bendiness

Curves are considered to be a risk factor in the design of roads. However, the safety of a horizontal curve is affected not only by features internal to it (e.g., radius, superelevation, and the presence of transition that were discussed earlier) but also by features external to a particular curve that reflect the whole road design, for example, density of curves upstream, length of the connecting tangent sections, etc. The design of preceding road sections or general bendiness of the road may have a direct effect on the drivers' level of attention and expectations with regard to the forthcoming road alignment. Hence, they might influence driver behavior and curve approach speed, in particular, and the risk of crashes.

Road bendiness is examined for long road sections, routes, or network areas. In the literature, there are different ways for measuring bendiness such as:

- Absolute number of curves on the road, or curve frequency;
- Number of curves per kilometer, or curve density;
- Sum of deflection angles (or changes in direction, in degrees) per road kilometer, or curvature change rate; and
- Length of curves as a percentage of the road length.

Most research on the risk of curves focuses on individual curves, or on two-three subsequent curves, which are related to design consistency. The number of studies that tried to relate road bendiness or curve frequency to crashes, on longer road sections or on a network level, is not big.

Findings from the research studies on the topic are inconsistent. Some studies report a higher risk of crashes on roads with a higher degree of bendiness, others find no relation, while several more recent studies report a lower risk of crashes on road network units with a higher level of bendiness (e.g., Haynes et al., 2007; Jones et al., 2012). The authors of the latter studies hypothesized that such findings might be due to a higher driver attention on the network units with a higher degree of bendiness, which might cause drivers to drive slower and, thus, reduced the crash risk. In other words, higher bendiness may have a moderating effect on driving speeds and, hence, an inverse relation to crashes. However, such speed–bendiness relationship was not examined explicitly.

The reported studies considered spatial units of different size and on various road types, both rural and urban, which make a generalization of findings difficult. Some studies that reported a positive safety impact of bendiness did not consider traffic exposure, and most studies did not refer to safety measures applied on the roads that could affect the crash risk. In general, it seems that such a

macro-consideration of the road or network curvature raises additional demands on the range of possible confounding factors to be examined.

None of the reported studies examined a comprehensive list of possible traffic, road infrastructure, spatial and environmental factors, and, thus, it can be suggested that other factors might cause the change in crashes, where an increase in bendiness was measured, rather than those that were explicitly examined by the study. Hence, the results of available studies on the risk of bendiness do not yet represent a causal relationship with crashes, but merely reflect a relation observed, given the conditions that were taken into consideration.

Based on the current research findings, there is no clear expectation concerning the effect of road bendiness on crashes as well as with regard to the impact of the measure of reducing the number of road curves, on safety. Further research on the subject is needed with a particular focus on the definition of spatial research units and corresponding measures of bendiness, and on controlling for the effects of other possible confounding factors.

Safety Impacts of Vertical Alignment

Grade

Vertical grade can indirectly influence safety by influencing the speed of the traffic stream and speed differences between various types of vehicles. Uphill grades create significant speed differences between cars and trucks that may increase the frequency of lane changes and related crashes. Downhill grades accelerate the vehicle and place additional demands on vehicle braking and maneuverability. Steep downgrades are more dangerous for two-wheelers and, particularly, for bicycles. The design and safety literature distinguishes between constant grades and vertical curves.

According to the research findings in the United States, the presence of (positive or negative) constant grades, on two-lane rural highways, is associated with an increase in crashes. The summary estimates suggest that (AASHTO, 2010):

- For level grade, up to 3%, the CMF value is equal to 1 (no safety impact);
- For moderate terrain, with a grade between 3% and 6%, CMF is equal to 1.10, that is, a 10% increase in crashes is expected; and
- For steep terrain, with a grade of 6% or more, a 16% increase in crashes is expected.

The functional form for percent grade CMF (CMF_g) on a rural two-lane highway is given by:

$$CMF_g = 1.016^{abs(g)} \quad [6]$$

where $abs(g)$ is an absolute value of percent grade. It shows an increase in total crashes when vertical grade is higher than zero, yet a tangible change in crashes (of 5% or more) is expected when the grade value is over 3%.

A slightly different form of CMF for the relationship between grade and injury (plus fatal) crash frequency was developed for rural roads in Texas, enabling to separate the estimates for multilane versus two-lane roads, as follows:

$$CMF_g = e^{0.016g} \text{ for two-lane highways,} \quad [7]$$

$$CMF_g = e^{0.019g} \text{ for multilane highways,} \quad [8]$$

where g is an absolute value of percent grade (%). Fig. 5 presents both relationships and indicates that a higher grade is associated with higher crash frequencies and also that a slightly higher crash increase, under the same grade, is expected on multilane highways. As reported, similar estimates can be applied for evaluating safety effects of vertical curves.

Safety impacts of steep vertical grades were extensively explored in the international literature, including studies conducted in various states of the United States, China, South Korea, Italy, and other countries (Ziakopoulos et al., 2016). However, the risk factor

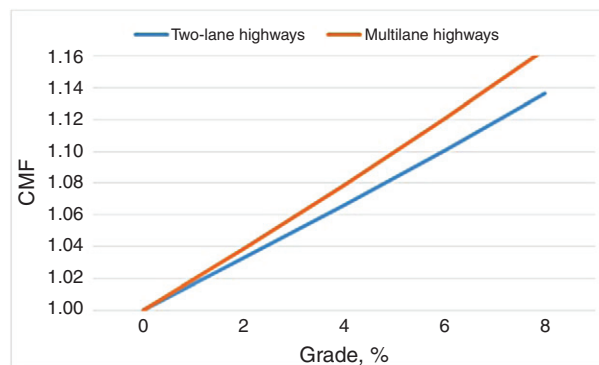


Figure 5 Examples of grade CMFs for injury crashes Source: Produced based on Bonneson and Pratt (2009).

of steep grade was typically not examined alone but appeared in conjunction with other road design features while the main purpose of the studies was to examine the safety performance of certain road types or sites. To quantify the effect of grade, frequently, a threshold approach was applied that categorized all grade values into “steep” and “not steep.” Some studies used a wider range of grade categories or continuous numeric approaches. The international results refer to various road types, including freeways, national roads, and all rural highways; some studies focused on intersections or freeway ramps only, while most studies conducted aggregate analyses of road sections.

The majority of findings were consistent indicating a detrimental effect of a steep grade on road safety, that is, increasing crash risk or injury severity associated with steep vertical grades. However, there have also been reported cases where no statistically significant effect was found for a steep grade or when a decrease in crash frequency was observed on some road segments for a corresponding one-unit increase in vertical grades.

Based on the majority of the research findings, reducing vertical grades is expected to have a mostly positive effect on road safety, although dedicated before–after studies on the topic are lacking in the literature.

Interactions Between Horizontal and Vertical Alignments

Most studies reported separate safety effects of horizontal curvature and percent grade, not accounting for the interactions between them. Such an approach assumes that the effect of a horizontal curve is the same whether it is located on a level roadway, a constant grade, or a vertical curve, or, similarly, that the effect of a constant grade is the same whether it is located on a tangent section or on a horizontal curve. Recently, several studies analyzed such interactions explicitly demonstrating the combined effects of both features.

A study based on a dataset of the State highways in Washington prepared a comprehensive classification of two-lane road segments into categories by their horizontal and vertical alignments and developed statistical models for predicting crashes, under various conditions (Bauer and Harwood, 2014). The models estimated main effects of traffic exposure, horizontal curves and vertical grades, and the interactions of both.

Table 1 presents examples of CMFs estimated for various combinations of horizontal and vertical alignments; in all cases, the base conditions belong to level tangent road sections. The results show clearly an increase in crash risk with shorter horizontal curve or higher vertical curvature, where the combined effects appear to be stronger than the values reported (previously) for separate features. Particular high crash risks were found for short horizontal curves combined with high straight grades and for steep sag

Table 1 Examples of CMFs for injury (and fatal) crashes on two-lane road sections with various horizontal and vertical alignments

A—Combinations of horizontal curves and straight grades ^a					
Grade, %	Tangent	For horizontal curve radius = 1300 ft., when horizontal curve length, mi		For horizontal curve radius = 2500 ft., when horizontal curve length, mi	
		0.10	0.30	0.10	0.30
Below 1	1.00	1.57	1.53	1.36	1.34
1	1.04	1.64	1.60	1.42	1.40
2	1.09	1.71	1.67	1.48	1.47
3	1.14	1.79	1.75	1.55	1.53
4	1.19	1.87	1.82	1.62	1.60
5	1.25	1.95	1.91	1.69	1.67
6	1.30	2.04	1.99	1.77	1.75

B—Combinations of horizontal curves and vertical curves						
K ^b	Type 1 crest vertical curves ^c with horizontal alignment of			Type 1 sag vertical curves ^d with horizontal alignment of		
	Tangent	Radius = 1300 ft.	Radius = 2500 ft.	Tangent	Radius = 1300 ft.	Radius = 2500 ft.
250	1.0	1.08	1.04	1.04	1.15	1.10
125	1.0	1.17	1.08	1.09	1.32	1.20
83	1.0	1.26	1.13	1.13	1.52	1.32
63	1.0	1.36	1.17	1.18	1.74	1.44
50	1.0	1.47	1.22	1.23	2.00	1.59

^aConstant percent grade.

^b*K* is a ratio of algebraic difference in grade and length of curve representing the sharpness of the vertical grade. *K*-values set at 250, 125, 83, 63 and 50 correspond to a grade difference of 2%, 4%, 6%, 8% and 10%, respectively, for a vertical curve length of 500 ft.

^cWith positive approach grade and negative departure grade.

^dWith negative approach grade and positive departure grade.

Source: Produced based on Bauer and Harwood (2014).

vertical curves combined with sharper horizontal curves. The results are in line with engineering judgment. The models provide useful tools for comparing safety performance of various road designs, based on the combined effects of road features. However, current findings are limited to rural two-lane highways, road segments only, and are not available for major intersection areas and for other road types. In addition, the impacts of other road features, for example, lane and shoulder width, superelevation, roadside conditions, etc., were not included yet in the combined analyses.

Evaluating Speeds

Horizontal and vertical road alignments are among the main road characteristics that impact on selecting travel speeds by drivers. It is generally assumed that, in their trip along the road, the driver relies on their experience of passing previous road segments. When abrupt changes appear in the road alignment, a conflict arises between the driver expectations and road conditions that may cause driver errors and lead to crashes. Conversely, when the road design is consistent and uniform, driver's expectations are satisfied leading to fewer driver errors and safer driving. This understanding brought to the creation of the road design consistency concept that was originated by German researchers several decades ago (Lamm et al., 1999). Operating speed is a commonly used parameter for estimating road design consistency. In the literature, a great number of models can be found for predicting operating speeds based on road geometry characteristics, while most of them were developed for driving on curves.

Design consistency is evaluated by considering speed differentials, for example, the differences between the operating speeds of consecutive road segments and between the operating and design speeds of the same road elements. A common approach suggests that good road design consistency is attained when both differences are below 10 km/h, a gap of 10–20 km/h can be accepted, whereas a difference over 20 km/h requires a redesign. Research studies provided empirical evidence supporting the above categories. It was found, for example, that the crash rate of curves from the last category was 6 times higher than that of the first group and 2 times higher than the crash rate of the second group (RSM, 2003).

As speed is considered today as a crucial factor both in crash occurrences and their consequences, road design consistency has much in common with road safety evaluations. Not surprisingly, many research studies dedicated to safety impacts of road geometry, on various road types, chose to focus the evaluation on speed indicators and their relationship with road design characteristics. Further conceptual developments lie in creating self-explaining (or self-enforcing) roads, whereby road design characteristics should deliver a clear message to drivers on the speeds appropriate for traveling on each road section. (An alternative approach is governing the actual speeds by GPS-based in-vehicle devices, e.g., intelligent speed adaptation.)

However, research findings on speed impacts of road design features are not straightforward with regard to expected safety impacts. For example, for rural two-lane roads in Israel, a moderating effect on travel speeds of the presence of horizontal curves and steep vertical grades was found, while straight and flat design was less suitable for reducing excessive speeds (Gitelman et al., 2016). Similarly, in a Canadian study, a reduced number of speeding vehicles was observed on urban roads with steep grades (Eluru et al., 2013). Such seemingly inverse effects on speed and crashes require further research for a better understanding of road geometry impacts on driver behaviors and crash occurrence.

Future Research

Evaluating safety impacts of horizontal and vertical road alignments, with consequent recommendations on safer road design, sets a great challenge for future research. Concerning the evaluation framework, the challenge relates to the high variability of crash numbers at individual sites, and the necessity to neutralize the external confounding factors. To ascertain the impacts of road geometry features, aggregated data for many road sites are required for the analyses, thus raising the need to control for other road design and traffic characteristics and their interactions. Ideally, it would be desirable to develop safety prediction models for various road types that consider the safety effects of a full range of design variables and their combinations and interactions. Such analyses will require more comprehensive databases than those currently available. In addition, further methodological developments are needed for macro-considerations of road curvature, for example, assessing safety impacts of road bendiness.

Additional research challenges lie in providing sufficient empirical findings for understanding the triangular relationship between road design features, driving speeds, and crashes, on various road types, aiming to support both safer road design and better speed management on the road network.

References

- American Association of State Highway and Transportation Officials (AASHTO), 2010. Highway Safety Manual. American Association of State Highway and Transportation Officials, Washington, DC.
- Bauer, K.M., Harwood, D.W., 2014. Safety effects of horizontal curve and grade combinations on rural two-lane highways. Report No. FHWA-HRT-13-077. Federal Highway Administration, Washington, DC.
- Bonneson, J.A., Pratt, M.P., 2009. Roadway safety design workbook. Report No. FHWA/TX-09/0-4703-P2. Texas Transportation Institute, College Station, TX.
- Eluru, N., Chakour, V., Chamberlain, M., Miranda-Moreno, L.F., 2013. Modelling vehicle operating speed on urban roads in Montreal: a panel mixed ordered probit fractional split model. *Accid. Anal. Prev.* 59, 125–134.

- Elvik, R., 2013. International transferability of crash modification functions for horizontal curves. *Accid. Anal. Prev.* 59, 487–496.
- Federation of European Motorcyclists Association (FEMA), 2012. *New Standards for Road Restraint Systems for Motorcyclists: Designing Safer Roadsides for Motorcyclists*. Federation of European Motorcyclists Association, Brussels.
- Gabauer, D.J., Li, X., 2015. Influence of horizontally curved roadway section characteristics on motorcycle-to-barrier crash frequency. *Accid. Anal. Prev.* 77, 105–112.
- Gitelman, V., Pesahov, F., Carmel, R., Bekhor, S., 2016. The identification of infrastructure characteristics influencing travel speeds on single-carriageway roads to promote self-explaining roads. *Transp. Res. Procedia* 14, 4160–4169.
- Goldenbeld, C., Schermers, G., van Petegem, J.W.H., 2017a. Low curve radius. European Road Safety Decision Support System, developed by the H2020 project SafetyCube. Available from: www.roadssafety-dss.eu.
- Goldenbeld, C., Schermers, G., van Petegem, J.W.H., 2017b. Absence of transition curves. European Road Safety Decision Support System, developed by the H2020 project SafetyCube. Available from: www.roadssafety-dss.eu.
- Haynes, R., Jones, A., Kennedy, V., Harvey, I., Jewell, Y., 2007. District variations in road curvature in England and Wales and their association with road-traffic crashes. *Environ. Plan. A* 39 (5), 1222–1237.
- Jones, A., Haynes, R., Harvey, I., Jewell, T., 2012. Road traffic crashes and the protective effect of road curvature over small areas. *Health Place* 18 (2), 315–320.
- Lamm, R., Psarianos, B., Mailaender, T., 1999. *Highway Design and Traffic Safety Engineering Handbook*. Mc-Graw Hill, New York.
- Road Safety Manual (RSM), 2003. Recommendations from the World Road Association. PIARC—World Road Association, France.
- Ziakopoulos, A., Theofilatos, A., Papadimitriou, E., Yannis, G., 2016. Alignment deficiencies—high grade. European Road Safety Decision Support System, developed by the H2020 project SafetyCube. Available from: www.roadssafety-dss.eu.

Further Reading

- DaCoTA, 2012. Roads. Deliverable 4.8q of the EC FP7 Project DaCoTA. European Commission, Brussels.
- Donell, E., Kersavage, K., Fontana Tierney, L., 2018. Self-enforcing roadways: a guidance report. Report No. FHWA-HRT-17-098. Federal Highway Administration, McLean, VA.
- Stein, W.J., Neuman, T.R., 2007. *Mitigation Strategies for Design Exceptions*. Federal Highway Administration, Washington, DC.
- Van Petegem, J.W.H., Schermers, G., 2016. Bendiness. European Road Safety Decision Support System, developed by the H2020 project SafetyCube. Available from: www.roadssafetydss.eu.

Human Factors in Transportation

Alison Smiley, Christina (Missy) Rudin-Brown, Human Factors North Inc., Toronto, ON Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction	331
Definition of Human Factors	332
Causes of Incidents and Collisions	332
Visual Characteristics and Limitations	333
Visual Acuity	333
Peripheral Vision	333
Contrast Sensitivity	333
Perception of Closing Velocity	334
Cognitive Characteristics and Limitations	335
Information Processing and Attention	335
Expectancy	335
Driver Visual Search	336
Useful Field of View (UFOV)	336
Stressors	337
Distraction	337
Fatigue	337
Medical Conditions	339
Inexperience	339
Age	339
Alcohol and Drug Effects	340
Alcohol	340
Marijuana	340
New Developments in Human Factors and Transportation	340
Automation	340
Test Methods: Naturalistic Studies	341
Human Factors Guidelines	341
Legalization of Marijuana	341
Expert Systems	342
Tools that Provide Feedback to Operators	342
Tools to Inform Planning	342
Tools that Inform Safety Countermeasures	342
References	343
Further Reading	345

Introduction

In the early days of transportation, mechanical failures were common causes of transportation crashes. Advances in technology, mechanics, occupant protection, automation, and operational environments across all modes of transportation have led to steadily improving safety in terms of fewer crashes and injuries. With that said, in all modes of transportation, human error remains the leading cause of collision. This does not necessarily mean that the cure will be to change human behavior, which is very difficult, but rather the more effective cure is to design vehicles, roads, and operational environments with human limitations in mind.

In this article on human factors and transportation, we write about human factors mainly from the perspective of road users. This is our starting point because this is our research background. Also, and more importantly, in many cases, human factors considerations apply similarly across all transportation modes. Among others, visual acuity, perception of travel speed and closing velocity, fatigue, and alcohol effects have similar characteristics in operators in all modes. There are some major exceptions and these are noted. For example, the use of checklists is a very relevant issue in aviation, but not in other modes. Automation and its attendant problems are much further advanced in aviation than in other modes like marine, rail, and road, though automation in ground transportation is developing quickly.

We begin with a definition of human factors, followed by a brief comment on the contribution of human factors to crashes and incidents and the main issues across the modes.

Definition of Human Factors

“Human factors” is the scientific discipline concerned with understanding interactions among humans and other elements of a system in order to optimize human well-being and overall system performance. Human factors (also referred to as “ergonomics”) draws on knowledge from the human sciences—for example, psychology, physiology, kinesiology, and anthropometry—and applies it to engineering design. Human factors research examines the way people interact with tools (hardware and software), equipment and workplaces, and aims to make these interactions safer, healthier, and more efficient.

The discipline of human factors developed during the Second World War, when it became evident that engineering design that did not accommodate human limitations led to errors, injuries, and fatalities. Human factors research was first widely applied in the area of aviation, leading to the generally exemplary safety record we know today. Human factors research into road operator behavior emerged in earnest during the 1950s and 1960s. Research into operator behavior in the marine and rail environments began later and, some would say, has not yet received the same level of attention.

Causes of Incidents and Collisions

Transportation crashes and incidents are very rarely the result of a single cause. Rather, they occur as a result of failures in system safety; a system that includes, at minimum, vehicle operators, equipment and infrastructure, and the natural and organizational environments. With this in mind, a review of crash causes can provide insights into the ubiquity of human error and performance in transportation incidents and crashes.

One of the largest in-depth road transportation studies is the National Vehicle Crash Causation Study ([National Highway Traffic Safety Administration, 2008](#)). This US study considered a nationally representative sample of 5,471 crashes investigated during a 2½-year period. Each investigated crash involved at least one light passenger vehicle that was towed due to damage sustained in the collision. Data were collected on at least 600 potentially contributing elements to capture information related to the drivers, vehicles, roadways, and environment involved. The “critical reason” for crashes, defined as the last event in the crash causal chain, was assigned to a driver in over 94% of crashes. In those cases, about 41% of the critical reasons were recognition errors (e.g., inattention, internal and external distractions, inadequate surveillance), 33% were decision errors (e.g., driving too fast for conditions, misjudgment of others’ speed), 11% were performance errors (e.g., overcompensation, poor directional control), and 7% were nonperformance errors (e.g., the driver having fallen asleep). About 36% of the vehicles were turning or crossing at intersections just prior to the crashes—characterized as the critical pre-crash events.

With respect to pedestrian collisions, key factors are urban areas, not at intersections and dark conditions. In one U.S. study ([National Highway Traffic Safety Administration, 2018](#)), one-fifth of pedestrians killed were children aged 10–4 and one-fifth were aged 65 and older. An estimated 33% of fatal pedestrian crashes involved a pedestrian with a blood alcohol concentration (BAC) of 0.08 g/dL or higher. Bicycle collisions involve a similar pattern of factors. Pedestrians and bicyclists appeared to be at fault most of the time, with the largest percent of errors being failure to yield right-of-way for both groups (30.6% and 34.9% pedestrians and bicyclists, respectively ([National Highway Traffic Safety Administration, 2017, 2018](#))).

Within the domain of aviation safety, a generally accepted estimate of the percentage of crashes and incidents ascribed to the human element (most often, the pilot) is two-thirds ([Hitchcock et al., 2010](#)). Research ([Endsley, 2010](#)) on the causal factors underlying aviation crashes that involve a substantial human error component reveals that most can be attributed to problems involving a failure to correctly perceive some piece(s) of information. Reasons underlying the misperception include: that data were not available (12%); the data were difficult to detect or perceive (12%); the crew failed to monitor the data despite it being available (37%); the crew misperceived the available data (9%); or the crew forgot the data (11%).

While generally more forgiving of human error than other transport modes because of lower relative speeds and broader directional tolerance, the marine safety environment presents other challenges. Often, accident investigation is limited to identifying direct and contributing causal factors, whereas wider sociotechnical context that has given rise to causal mechanisms is left unexplained ([Puisa et al., 2018](#)). The American Bureau of Shipping ([McCafferty and Baker, 2006](#)) performed a meta-analysis and developed a common taxonomy of causes across international marine accident databases and found that “human error” was the dominant factor in 80%–85% of maritime accidents worldwide. Among all human error types classified, failures of situation awareness and situation assessment predominated, being a causal factor in about 45% (offshore operations) to 70% (ships) of the recorded collisions associated with human error.

Railroading is different from other transportation modes because it is almost entirely an industrial activity and therefore does not entail the same level of individual decision-making. Regardless, and similar to the other modes, it can be difficult to ascertain the contribution of human factors issues to railway collisions, as “human factors” tends to be seen as the “cause factor of last resort” behind other technical, equipment, or service system elements ([Hill, 2007](#)). Railways assign what are called “cause codes” to indicate the primary and secondary cause of an accident. Cause codes exist for human factors (118 codes), signal defects (21 codes), track defects (65 codes), mechanical/electrical problems (143 codes), and miscellaneous (42 codes). Interestingly, no codes are related to organizational and management issues such as shift scheduling (Federal Railroad Administration, 2019).

Visual Characteristics and Limitations

It is estimated that 90% of the information drivers use while driving is visual. With the advent of semiautonomous road vehicles, auditory warnings, and automated driving capabilities are growing more prevalent, mitigating some aspects of the safety risk. While other modes of transport are associated with varying levels of vehicle automation and operator warning systems, the speed with which recent developments in automation have evolved in road vehicles remains unrivaled. Further, while operator vision is important for all modes, the nature of the operating tasks across them differs significantly. While driving a road vehicle is an essentially constant visual task, the visual aspects of flying an aircraft, driving a train, or operating a vessel are less constant over time, so visual characteristics and limitations, while important, pose less safety risk.

There are many aspects of vision that are important in vehicle operation. The most familiar is visual acuity, the ability to resolve small details, such as text, at a distance. Acuity considerations affect the design of displays and controls within the vehicle, as well as externally, for example, road traffic signs. While visual acuity is the most well-known visual characteristic, there are numerous others that are as much if not equally important. These include:

- Peripheral vision, which facilitates an operator's ability to detect intersecting vehicles and objects moving in the periphery, such as pedestrians or other vehicles.
- Contrast sensitivity, which enables detection of slight differences in luminance or reflectance between an object and its background.
- Perception of closing velocity, or the ability to estimate the speed with which one is closing on another vehicle.
- Accommodation, or the ability of an operator's eye to change focus from nearby instruments inside the vehicle to objects outside the vehicle, and vice versa.
- Adaptation, or the eye's ability to adjust in terms of light sensitivity on entering and exiting lit environments, such as streetlight areas or tunnels.
- Color vision, or the eye's ability to detect color. Color vision is particularly important to the identification of display, signal and sign colors in the rail, aviation and marine modes, but less so in the road environment because of deliberate color choices and the consistent design of traffic signals.

The first four of these visual characteristics are described in more detail next.

Visual Acuity

Of all the visual abilities, one for which vehicle operators in all modes are tested is visual acuity, which relates to how well they can read text at a distance. Under ideal conditions, in daylight, with high contrast text (black text on a white background), a person with a normal visual acuity of 6/6 (20/20) can just resolve letters at 6.8 m for each centimeter of letter height.

Road vehicle driver licensing requirements for visual acuity are generally a corrected acuity of 6/12 (20/40), or the ability to read at 6 m (20 feet) what a person with normal visual acuity can read at 12 m (40 feet) distance. For text fonts used on highway guide signs, to encompass the visual acuity of more than 95% of young drivers, 75%–85% of older drivers, and night driving conditions, design driver acuity is 4.8 m/cm (40 feet/inch) of letter height.

In other modes of transport that are predominantly commercial in nature, operator licensing requirements stipulate more stringent visual acuity. For marine positions that involve watchkeeping duties, corrected distance acuity must be at least 6/12 (20/40) in each eye, and a minimum requirement for near visual acuity. In rail, individuals occupying positions that involve movement of equipment on the track are required to have corrected or uncorrected distance visual acuity of at least 6/9 (20/30) in the better eye and at least 6/15 (20/50) in the worse eye, and minimum corrected or uncorrected near acuity. In aviation, Canadian airline and commercial pilots must have a minimum corrected or uncorrected distance acuity of 6/9 (20/30) in each eye, and minimum near acuity (corrected or uncorrected).

Peripheral Vision

The peripheral field of view is large—approximately 55 degrees above the horizontal, 70 degrees below, and 90 degrees to the left and to the right (Guyton, 1969, p. 294). However, the acuity of peripheral vision is much poorer than that of foveal (central) vision. If visual acuity for objects seen in foveal vision, along the line of sight, is 6/6 (20/20) it will fall to about 6/18 (20/60) at 5 degrees away from the line of sight (Mandelbaum and Sloan, 1947; Olson, 1988). In other words, letters would need to be 3 times bigger when they are not looked at directly but are read at 5 degrees off the line of sight. The larger the angle, the poorer the visual acuity. For this reason, drivers and vehicle operators in other modes fixate directly on an object or area of interest and then search the visual scene by rapidly moving their eyes to the next point of interest and fixating there. It is during these brief fixations that we take in information and “see.”

Contrast Sensitivity

Contrast sensitivity is important to safety. It is the ability to detect small differences in light level (or luminance) between an object and its background. The lower the level of light and the smaller the target, the more contrast is required to detect and perceive an

object such as a curb, debris on the road, an approaching train at a level crossing, or a pedestrian. Similarly, limited contrast sensitivity will make it more difficult for pilots to detect (and avoid) other aircraft in their vicinity, for locomotive engineers to correctly identify railway signals, and for mariners to detect other vessels or hazards in the water.

Good visual acuity does not necessarily imply good contrast sensitivity. For operators with 6/6 (20/20) visual acuity, the distance at which nonreflectorized objects are detected under full dark conditions can vary by a factor of 5 to 1. Therefore if using low beam headlights at night, drivers can get very close to a low contrast target before detecting it. Experimental studies show that even alerted drivers can come as close as 9 m before detecting a pedestrian in dark clothing standing on the side of the road (Olson and Sivak, 1983). Compounding the safety risk, pedestrians are unaware of how poorly drivers see them, over-estimating by a factor of two the distance at which they are seen (Allen et al., 1970).

Perception of Closing Velocity

Operator ability to estimate one's closing speed to another vehicle or object in the environment is generally very poor. One of the main cues for the road vehicle driver in determining the rapidity with which one is closing on another vehicle is the apparent change in the angle created at the eye by the rear end of the vehicle ahead, or by the front end of an oncoming vehicle. The determination of closing speed is difficult because, at a distance, the apparent size of the other vehicle is small. As the driver approaches, the size grows gradually at first and then faster and faster. This change in size is a nonlinear cue, making the driver's judgment of the rate of closing velocity very difficult. Studies suggest that alerted drivers in experimental situations do not even **begin** to be able to distinguish rapid from slow separation until the change in angle is, on average, 0.003 radians per second (Olson and Farber, 2003). Until this threshold is reached, all the driver perceives is that the gap is closing, something that happens regularly in traffic and does not precipitate emergency action. Fig. 1 shows how nonlinear the change in image size is as an object approaches closer and closer (Olson, 1996). It is this nonlinearity that greatly contributes to the difficulty observers have in making accurate estimates of other vehicles' speed.

Due to the poor ability of drivers to assess closing speed, they are at risk of several crash scenarios when in the presence of insufficient cues: rear-end crashes when they catch up to stopped or slowing vehicles, accepting too-small gaps when overtaking (Farber et al., 1967), or accepting too-small gaps when turning left across opposing traffic.

There are other challenges to the perception of closing velocity in modes of transportation other than road. For pilots, who operate in three dimensions, the basic method of visual collision avoidance follows the "see-and-avoid" principle, which is based on active scanning and a pilot's ability to detect conflicting aircraft and take appropriate measures to avoid them. This ability is affected by many factors, including physiological limitations of human visual and motor-response systems, obstructions to field of view, aircraft conspicuity, pilot scanning techniques, workload, and alerting to the presence of other aircraft. The effective practice of see-and-avoid can be influenced by limitations in what can be seen and by other activities, such as in-flight monitoring of instruments, radio communications, flight training exercises and interactions with an instructor, and navigation or conduct of simulated instrument approaches. A pilot's full attention may thus be diverted from active scanning for traffic. Research (FAA, 2016) has determined that, for pilots to have sufficient opportunity to avoid a collision, they must be able to detect a conflicting aircraft a minimum of 12.5 seconds prior to the time of impact. This delay in reaction time can and does vary depending on pilot experience, and is likely to exceed 12.5 seconds, regardless of a pilot's experience or training.

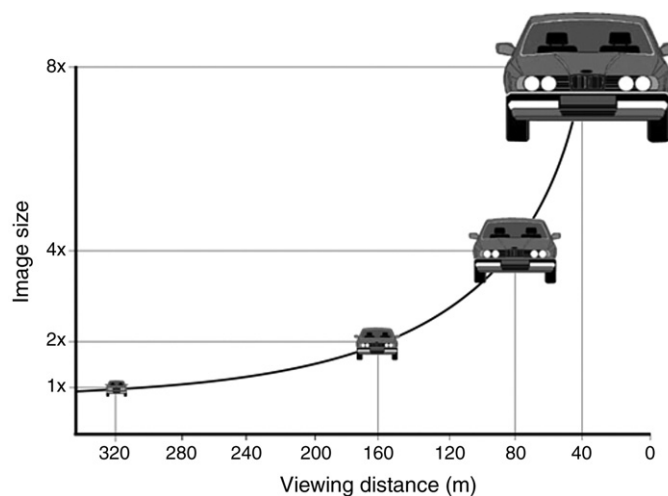


Figure 1 The nonlinear relationship between viewing distance and image size. Source: Thomas Smahel adapted from Olson, 1996.

Cognitive Characteristics and Limitations

Information Processing and Attention

Human attention and abilities in information processing are limited. These limitations can create difficulties because operating any vehicle requires the division of attention between control tasks (e.g., staying in the lane, following a planned course), guidance tasks (e.g., merging with other vehicles, avoiding other vessels or aircraft), and navigational tasks (e.g., looking for street name signs, plotting an upcoming course change). While attention can be switched rapidly from one information source to another, humans attend well to only one source at a time. Furthermore, humans can extract only a small proportion of the available information from the forward scene. It has been estimated that, out of over 1 billion bits per second of information directed at the sensory system, roughly 16 bits per second are consciously recognized (the answer to a single yes/no question provides 1 bit of information) (Grandjean, 1988). In short, the human information processing system is essentially a single channel system with limited capacity (Grandjean, 1988; Kantowitz and Sorkin, 1983). Given these limitations in information processing, it is not surprising that vehicle operators are more likely to make errors when they are faced with high demands from more than one information source (e.g., attending to a navigation task while simultaneously changing lanes), because of information overload. They also rely on design being as expected—when expectations are violated (e.g. where there is a left hand exit from a freeway, or a white light on the rear of the vehicle ahead), drivers are more likely to make errors and respond slowly.

Expectancy

Information processing load on operators is reduced by designing operational environments (e.g., roadway, cockpit, bridge) in a predictable manner, according to operator expectations. Expectation is a powerful determinant of accuracy and reaction time. To take a very simple everyday example, as we enter a dark room that we have never entered before, we expect the light switch to be about shoulder height and near the edge of the door. We also expect to move it up to turn it on. If it is placed as expected, and operates as expected, we can locate it and turn the light on quickly. If it is not placed as expected—say at waist height, and operates differently than expected, for example, it moves laterally instead of vertically, our response time will be considerably longer. Similarly, drivers will respond slowly and inaccurately when roads and traffic control devices are not designed according to their expectations. This is a central tenet of the “positive guidance” approach to highway design. This approach, based on a combination of human factors and traffic engineering, was developed in the 1970s by two psychologists, Alexander and Lunenfeld, and elaborated on in a series of documents published by the US Federal Highway Administration (Alexander and Lunenfeld, 1975).

When drivers are surprised because their expectations are violated, slowed responses and errors occur. An example of such a violation of expectations is the use of left side exits on freeways. It has been known since the 1960s that left exits are associated with over twice the crash rate of right exits. Similarly, it has been known for many years that crash rates are higher on a sharp curve that occurs after a series of gentler curves than on a sharp curve that is characteristic of the other curves on that roadway.

Expectations also affect vehicle operator performance in the other modes of transportation. For example, pilots operate in a complex environment with multiple sources and types of information to monitor. To operate an aircraft effectively, pilots must pay attention to the most meaningful information needed for the task they are performing at that time. To help them cope with the large amount of information in the environment that is available to the senses at any given time, humans have developed expectation ‘biases’ that can facilitate the processing of information. These normal biases, however, have an unintended consequence: not all of the—potentially critical—elements in the flight environment will be attended to, which can lead to poor decisions and increased accident risk. For example, when the amount of available information about a situation is limited, pilots (and other vehicle operators) may tend to rely heavily on the first piece of information that was available to them to inform subsequent situation assessments. This is known as “anchoring bias,” and can make unfolding situation assessment less accurate than otherwise. Similarly, having only limited information about a situation can increase a pilot’s tendency to look for evidence that confirms or matches one’s current assessment or decision, a phenomenon known as “confirmation bias.” These biases can make it less likely for a pilot or other vehicle operator to reassess their initial assessment and update it with new information, or lead them to hand-pick information that supports their current decision, while dismissing information that is opposite to what is expected (Tversky and Kahneman, 1982). The danger in both circumstances is that alternative outcomes will not be given an appropriate level of consideration when deciding on the best possible course of action.

Research and past accident investigations have demonstrated that, once a plan is made and committed to, it becomes increasingly difficult for an individual to recognize stimuli or conditions in the environment as necessitating a change to the plan (Orasanu et al., 1998). “Plan continuation bias” is the “deep-rooted tendency of individuals to continue their original plan of action even when changing circumstances require a new plan” (Berman and Dismukes, 2006). If a reason to change the plan is to be recognized and acted upon in a timely manner, a condition or stimulus needs to be perceived as sufficiently salient to require immediate action. When plan continuation bias interferes with a pilot’s ability to detect important cues, or if the pilot fails to recognize the implications of those cues, nonoptimal decisions that can compromise safety become more likely.

In the marine domain, bridge controls and consoles that do not operate according to a bridge crew’s expectations can lead to errors and unwanted situations. As integrated bridge and automated control systems become more common on vessels, it is important to design interfaces that operate as expected by their human operators, while at the same time providing concise and accurate feedback regarding system status.

In rail, the expectations of locomotive crews make up an integral component of train operations. For many years, the railway industry in Canada has relied on crew compliance with wayside track signals that provide them with a series of signal indications requiring actions relative to the signals displayed. In this context, train crews are expected to react to the progression of wayside signal indications. The principle of progression for a series of signals is strongly instilled in experienced train crews, as they are trained and expected to react to wayside signal indications as presented along the route. In fact, however, rail traffic controllers will minimize the amount of information passed on to operating crews to reduce the risk of anticipation that can result in crew expectation bias related to the signals ahead.

Driver Visual Search

Eye movement studies illustrate the briefness of fixations and the demanding nature of driving (Rockwell, 1988). Visual demands on pilots, mariners and rail engineers are very different from those on drivers because they are less constant over time. Further, unlike that of a road vehicle, train control does not involve a lateral element. And flying is undertaken in three, as opposed to two, dimensions, meaning visual search will be more varied than that in the other modes.

For drivers, visual search is generally concentrated at or below the horizon, and, for right hand driving jurisdictions, approximately 5 degrees to the right of the focus of expansion (the point in the distance where parallel lines appear to merge). Drivers look ahead of their vehicles by 2–3 seconds to monitor changes in the road path, and deliberately fixate particular areas, depending on the driving task. For example, they will look towards the right when searching for traffic signs, as that is where most of them are found. In contrast, pilots need not focus on the path ahead except in very specific circumstances, such as during takeoff and landing. Similarly, mariners need to focus directly ahead during berthing maneuvers and when navigating narrow passages.

In terms of eye movements, drivers make three fixations a second on average (Mourant and Rockwell, 1970) and fixations last for between 1/10 second up to about 2 seconds (Rockwell, 1988). At speed limits of 50–100 km/h, on the order of 5–0 m would be covered during an average glance duration. This means that the number of eye fixations that can be made, and the number of objects that can be identified as a driver drives through a particular area, is very limited. Because of the importance of visual search for road vehicle lane-keeping and directional control, drivers are loath to go for more than 2 seconds without getting some information from the roadway (Rockwell, 1988).

Besides looking in particular areas for information (what is known as “top-down processing”), drivers and other vehicle operators rely on their peripheral vision to detect targets of interest to be fixated (what is known as “bottom-up processing”). The targets most likely to be detected in peripheral vision—what is also known as an object’s “conspicuity”—are those that are large, moving, have good contrast with their background, and are consciously being searched for. The further a target (e.g., a pedestrian, intersecting vehicle or oncoming aircraft) is off the line of sight, the longer it will take for the driver or other vehicle operator to detect and the more likely it is to be missed.

Useful Field of View (UFOV)

Although, as noted earlier, the peripheral field of view is large, the portion of that field of view which is used while driving or operating any kind of transport vehicle is much more limited. Useful field of view (UFOV) is defined as “the total visual field in which useful information can be acquired in a single fixation.” A laboratory study (Ball et al. 1988) examined the effect of central task demand, number of distractors, and age on participants’ ability to locate a target stimulus (a cartoon of a human face) at 10, 20 or 30 degrees off the forward line of sight during a single fixation (Ball et al. 1988). The greater the central demand (e.g., indicating whether the central human face was smiling or frowning versus looking at a static central fixation point), the greater the number of distractors and the greater the eccentricity, the more frequently targets were missed. Error rates for 30 degrees were 25% or higher depending on the number of distractors that were present and the difficulty of the central task. Applying these findings to the issue in question, the more demanding the operating task and the further the target (e.g., a pedestrian, an oncoming aircraft) is off the forward line of sight, the more likely it will be missed.

An on-road driving study examined what properties, and driving circumstances, affected target detection (Cole and Hughes, 1984). One group of participants was asked to report whatever objects they noticed as they drove. The other group was asked to search for and report disc targets that had been set up by the experimenters along the roadway. The disc targets were considerably more likely to be identified when participants were asked to search for them specifically (approximately 40% of those in a shopping area were identified, rising to approximately 80% of those in a residential area). When participants merely reported what attracted their attention and were not asked to search for the disc targets, they only noticed about 6% of them in the shopping area and about 40% of them in the residential area. The more intentional the drivers were in looking for particular targets, the more likely they were to detect them. Nonetheless, even though the participants searched for the targets, many of them were missed, especially in visually cluttered urban areas. Objects that were not being deliberately searched for tended to be first noticed when they were relatively small in angular size (less than 1 degree) and at small angular eccentricities (less than 10 degrees off the driver’s line of sight).

A second on-road study exploring UFOV (Cohen, 1987) measured driver perception-reaction time to stimulus lights placed at various angles away from the forward line of sight, while drivers drove in various environments, on highway, rural road and urban streets. Drivers were required to respond to the onset of a stimulus light by touching a metal ring mounted on the steering wheel. If a light was not responded to within four seconds, it was turned off and counted as a miss. For participants, who were well-experienced police car drivers, on urban streets, 27% of the targets were missed when the angle between the target and the driver’s line of sight was

15 degrees. When the angle increased to 25 degrees, about 50% of targets were missed. Perception-reaction time for targets 15 degrees off the line of sight averaged 1.5 seconds—25%–50% longer than for targets seen up to 5 degrees off the line of sight. The author noted that perception-reaction times to more complex events than a red light illuminating for nonalerted drivers, are certain to be considerably longer than the measured values (Cohen, 1987).

Targets that are highly contrasted with the background against which they are seen (Bullough and Rea, 2000), or are moving, are more likely to be seen further off the line of sight in peripheral vision than stimuli that have poor contrast with the background or are static (Boff et al., 1986). Thus, the extent of the UFOV will be dependent in part on the conspicuity of the visual stimulus. The detectability of stimuli also depends on whether the observer is actively searching for them or not. An on-road driving study found that, during a merging task, the average peripheral detection angle for passing vehicles was 87 degrees from forward, considerably greater than the 10 degrees at which objects not being actively searched for were noticed in the study described above.

Stressors

Distraction

Operating any kind of vehicle is a complex task, involving the coordination and execution of various cognitive, physical, sensory, and psychomotor skills (Young and Regan, 2007). Despite this complexity, it is common to see road vehicle drivers engaging in other ‘secondary’ activities ranging from the seemingly benign (e.g., talking to a passenger) to the potentially hazardous (e.g., eating a bowl of cereal) (Regan et al., 2009).

Driver distraction is “the diversion of attention away from activities critical for safe driving towards a competing activity” (Lee et al., 2009). It can impair the ability to detect, recognize and respond to hazards. Distraction is typically categorized into types, depending on the source. *Visual* distraction refers to a driver neglecting to look at the road scene and instead focusing his or her visual attention on another target, such as a cell phone’s keypad or touchscreen, or on a child in the rear seat. *Physical* (or manual) distraction occurs when drivers remove one or both hands from the steering wheel to physically manipulate an object, for example, a lid on a coffee cup or handheld cell phone or music player. *Cognitive* distraction occurs when a driver’s attention is withdrawn from driving and applied instead to a nondriving related activity, such as a cell phone conversation (Strayer et al., 2013) or simply being “lost in thought” (Martens and Brouwer, 2013). Road users other than drivers, such as cyclists (Stelling and Hagenzieker, 2012) and pedestrians (Nasar et al., 2008), can also experience distraction, as can pilots, locomotive engineers and mariners.

Driver distraction is a primary contributing factor in motor vehicle crashes, and is a significant safety concern worldwide (World Health Organization (WHO), 2011). Research conducted under “naturalistic” conditions, where everyday drivers operate their own vehicles while a range of video and vehicle performance data is recorded and later analyzed, has consistently found that driver eye glances away from the forward visual scene are, not surprisingly, associated with crashes and near-crash events (Fitch et al., 2013; Klauer et al., 2006). A two-second glance away from the forward view doubles the risk of a crash.

Passengers can also be a source of cognitive (and visual) distraction (Stutts et al., 2003) for all ages of driver, and rear-seated child passengers can be particularly distracting to parent drivers (Koppel et al., 2011). Pets riding in vehicles can also lead to distraction. A recent anonymous, online, self-report survey from the United Kingdom of over 482 respondents found “interaction with pets” to be the third most common distracting behavior resulting in an accident, with 1.7% of respondents reporting having had an accident while driving with pets (Lansdown, 2012).

Risks from 50 different distracting activities were examined in an extensive naturalistic driving study involving over 3000 drivers, whose vehicles were instrumented over a one to two year period (Victo et al., 2015). Distraction involving “portable electronics visual-manual devices” had the highest odds ratio for the risk of a crash, with texting having the highest odds ratio of 5.6.

Distraction can similarly impair an operator’s performance in other modes of transportation. For example, a pilot’s ability to attend to critical stimuli within their environment will be impaired if they are distracted or inattentive, and will result in impaired situation awareness (Endsley, 1995). In light of this risk, many commercial air operators have standard operating procedures in place to limit the potential for pilot distraction. These “sterile flight deck” rules are intended to ensure that pilots experience a minimum of distraction during critical phases of flight, such as takeoff and landing, and that critical tasks are not interrupted. More specifically, the rules warn pilots to avoid unnecessary communication other than in an emergency or to communicate information essential to the safety of passengers or aircraft.

The rail mode in Canada restricts the use of communication devices to matters pertaining to railway operations, and cellular telephones are not to be used when normal railway radio communications are available (Transport Canada, 2018). The use of personal entertainment devices is also prohibited. There are presently no Canadian regulations with respect to the use of personal communication devices in the marine mode. In lieu of specific regulations, however, some pilotage associations have general requirements for marine pilots that they must apply their full attention to the task of piloting, and that any irrelevant activity that can distract from the piloting task is incompatible with this obligation.

Fatigue

Fatigue (or sleepiness) is prevalent among operators in every mode of transport, but particularly in commercial transportation operations. A US study by the National Transportation Safety Board (NTSB) of 107 single-vehicle nighttime truck crashes found that

58% included fatigue as a probable cause. The most critical factors in predicting which nighttime crashes were fatigue-related were: the duration of the driver's most recent sleep period, the amount of sleep in the past 24 hours, and split sleep patterns (due to the use of sleeper berths). The truck drivers in fatigue-related crashes were found to have obtained 2.5 hours less sleep in the last sleep period than the drivers involved in nonfatigue-related crashes. A major study conclusion was that "driving at night with a sleep deficit is far more critical in terms of predicting fatigue-related crashes than simply nighttime driving" (NTSB, 1995).

On-road, a 2007 survey found that 15% of Canadian respondents admitted that they had fallen asleep while driving during the past year (Vanlaar et al., 2008).

Fatigue can easily lead to impaired hazard detection, increased perception reaction time, or even a complete failure to react. Ultimately, it can lead to crashes, with about 20% of fatal traffic collisions in Canada involving driver fatigue (Transport Canada, 2011).

Whether an operator is fatigued is mainly dependent on the body's circadian (daily) rhythm, hours of continuous wakefulness, the quality and duration of recent and chronic sleep, and medical conditions that impair one's ability to obtain adequate restorative sleep, such as obstructive sleep apnea (OSA). The monotony of the operating task itself can also exacerbate sleepiness.

Fatigue and its counterpart, alertness, are determined by the body's circadian rhythm—a daily variation of various physiological functions such as sleep, body temperature, adrenalin production and many others. The circadian rhythm physiologically gears the body for action during the day and for sleep at night. Body temperature, for example, rises slowly throughout the day, peaking at about 8 pm. It then falls rapidly to reach a low point at approximately 3:00 a.m. (Grandjean, 1982). Along with fluctuations in body functions go fluctuations in physical capacity and mental alertness that also have a 24-hour cycle. Mental abilities peak during the daylight hours, and are poorest at night, particularly around 3:00 a.m. There is a secondary low point in the afternoon, known as the "post-lunch dip." Various studies from a number of countries show a very large increase in road accident crash risk by time of day, with the early morning hours having as much as 25 times higher crash risk than other hours of the day (Horne and Reyner, 1995). Night-shift workers are particularly vulnerable to motor vehicle crashes. A survey of 957 emergency medicine residents found that crash risk on the drive home from work was significantly higher, by a factor of six, after working night shifts than after day shifts. Further, the likelihood of having a crash increased significantly with the number of night shifts that had been worked in a sequence (Steele et al., 1999).

Fatigue can be caused by staying awake for a long period of time. In total, 22 hours of continuous wakefulness is commonly considered the point at which fatigue causes almost all aspects of human performance to decline and for "micro-sleeps" to begin (Beaumont et al., 2001). Fatigue may result from fewer hours of continuous wakefulness if these hours occur at night, rather than during the day, even for regular night workers.

Task monotony can also contribute to an operator's level of fatigue. People who experience monotonous stimuli are more likely to fall asleep than those who simply sit quietly in a dark room (Oswald, 1966). The monotony of driving on a familiar route or on a monotonous highway can contribute to fatigue and to falling asleep. Rail, marine, and long haul flights are also well recognized as soporific (sleep-inducing) environments. In aviation, flight crews can cross several time zones during a work period and face the added negative effects of circadian de-synchronization, or "jet lag." When these factors are considered in the context of the 24 hours a day/7 days a week reality of the commercial transportation industries, it is easy to see how and why fatigue is such a critical issue that needs to be addressed. In fact, the Transportation Safety Board of Canada (TSB)'s 2018 Watchlist includes "fatigue management in rail, marine and air transportation" as one of 7 key safety issues that need to be addressed to make Canada's transportation system even safer.

Finally, certain medical conditions can limit the quality and quantity of sleep that a vehicle operator obtains, which can contribute to fatigue. For example, a study of drivers who had been implicated in "falling-asleep-at-the-wheel" crashes found that many showed signs of OSA (Arbus et al., 1991), a medical condition and sleep disorder characterized by episodes when breathing ceases during sleep, resulting in frequent awakenings and chronic daytime sleepiness. Of all drivers evaluated, 62% had some form of sleep disorder.

The prevalence of OSA in the general public has been estimated to be 4% in men and 2% in women (Young et al., 1993). Some estimates indicate a higher prevalence for middle-aged men, especially if they are overweight, perhaps as high as 10% for men between the ages of 40 and 60 (Bearpark et al., 1990). It is more common in people who snore (AASM, 2001), and is reliably linked to large neck size (Chung and Elsaid, 2009), obesity (Dagan et al., 2006), and cardiovascular disease, including hypertension (Lurie, 2011).

OSA is associated with increased risk of motor vehicle crashes, which may result from lapses of consciousness, slow or inappropriate reactions, or from impaired judgment (Teran-Santos et al. 1999). OSA sufferers have a 2–10 fold increased risk of having a driving accident compared to people without OSA depending on the study cited (Ayas et al., 2006, 2014; Ellen et al., 2006; Karimi et al., 2013; Philip and Akerstedt, 2006), and are more than 6 times as likely as nonsufferers to have a traffic accident requiring hospitalization (Teran-Santos et al. 1999).

As with motor vehicle drivers, operators in other transportation modes are also at risk, though some more than others. In aviation, the US Federal Aviation Administration issued medical guidance relating to OSA in 2015 because of a 2008 incident investigated by the US NTSB. In that incident, the crew flew past their destination airport in Hawaii after both crew members fell asleep during the flight. The NTSB determined that the probable cause was the captain and first officer inadvertently falling asleep during the cruise phase of flight. Contributing to the incident was the captain's undiagnosed OSA. The NTSB concluded that efforts to identify and treat OSA in commercial pilots were needed to improve the safety of the traveling public, and recommended a

program to identify pilots who were at high risk for OSA and require that those pilots provide evidence of treatment before being granted unrestricted medical certification.

Medical guidelines in the rail mode provide a practical process by which all operating employees can be screened for OSA and subsequently diagnosed and managed appropriately. According to the guidelines, individuals with severe OSA cannot be considered fit to work until written confirmation and appropriate data have been provided by the treating physician indicating that effective treatment has been achieved and that the individual is compliant with therapy.

Medical Conditions

There are certain other medical issues that can impair human performance and vehicle operation. As discussed above, OSA is an important and common one. Based on two European studies, other 'high' risk conditions for motor vehicle crashes include alcoholism, mental disorders, neurological disorders, diabetes, and conditions resulting in mobility disorders (Sagberg, 2006; Vaa, 2003). The Ontario Ministry of Transportation has recently implemented screening for neurological disorders, such as dementia, that can impair driving performance as part of the age 80 years and up assessments. Finally, pain from chronic conditions like arthritis, and acute injuries and recent surgery can limit one's ability to operate a vehicle by limiting one's range of motion and the effectiveness of visually scanning the environment. Mental health conditions like depression can have secondary effects on performance by disrupting, for example, sleep. They can also have indirect secondary effects through medication to control symptoms like anxiety, which may have corollary negative psychoactive effects on for instance psychomotor (e.g., hand-eye coordination) control.

Inexperience

Compared to older, more experienced drivers, young novice motor vehicle drivers (i.e., those under the age of 21) are at an increased risk of having an accident, especially in the first few months of licensure (Cooper et al., 2005; IIHS, 2012; Lam et al., 2003). On average, young, novice drivers are much slower to detect hazards, and identify fewer hazards, than experienced drivers (Lee, 2007). This is especially true for hazards that are located further away, which implies more limited visual search strategies for novices.

Similarly, motorcycle riders' visual scanning patterns change with increasing riding experience, with novice riders demonstrating narrower visual scanning patterns compared to experienced riders (Hosking et al., 2010). The hazard perception ability of motorcyclists improves with experience (Liu et al., 2009), and car driving experience also appears to have positive effects in terms of hazard perception ability in motorcycle riders.

The presence and number of peer-aged passengers further increases the crash risk of young novice drivers. There is an exponential, stepwise relationship between the number of passengers and risk of having an accident (Chen et al., 2000; Tefft et al., 2012); compared to no passengers, having one passenger younger than age 21 leads to a 44% increased risk of being killed in a crash; having two passengers younger than 21, results in a doubling of risk; and having 3 or more passengers younger than 21, results in a quadrupling of a young driver's risk of death in a crash. The increased risk from carrying passengers is particularly strong when the passenger(s) is/are male (Simons-Morton et al., 2005). The mechanism for young novice drivers being at increased risk of collision when carrying peer-aged passengers likely involves the verbal interaction among them (Lam et al. 2003; White and Caird, 2010).

Age

As drivers age, many aspects of vision change, including visual acuity, contrast sensitivity, sensitivity to glare, dark adaptation, and field of view (Dewar and Olson, 2007). As a result of changes in sensitivity to contrast, older drivers have shorter seeing distances at night. For example, Sivak et al. (1981) matched groups of younger (age 18–30) and older (age 60–75) subjects in terms of daytime acuity, and found that the older subjects could read highway signs at night at an average of only about two-thirds the distance as compared to younger subjects (Sivak et al., 1981; Olson and Farber, 2003). Many aspects of cognition also change with age. Older drivers are unable to divide attention as easily and are more easily distracted from a primary task, such as driving. The UFOV test shows a decline in performance with age, suggesting greater constriction of the field of view with increasing age. Older drivers with UFOV impairment are more likely to have a crash (Owsley et al., 1991). Finally, with age comes increased likelihood of medical conditions, the most concerning of which are cognitive impairments, including dementia, which can increase crash risk by a factor of three (Diller et al., 1999).

Compared to middle-aged drivers (those aged 45–59), on a per-kilometer-driven basis, those in their late 60s have a 30% greater chance of being involved in an accident. Those aged 70–74 years are 90% more likely (Hildebrand, 2012). The risk is even higher for aging commercial truck drivers, with those drivers over the age of 70 being over 7 times more likely to experience a collision than those aged 41–50 years. This disparity among commercial drivers is likely due to higher levels of driver workload than general driving, increased complexity of the driving task and the inability for commercial drivers to "self-regulate"—voluntarily reduce their exposure to risks by making fewer, or shorter, trips.

The increase in accident risk associated with aging has led to some jurisdictions in Canada enacting mandatory retirement for some commercial (e.g., school bus) drivers, usually at the age of 65. In recent years, however, all provinces and territories repealed their mandatory retirement laws due to human rights issues. However, exceptions can be granted that permit mandatory retirement at a given age in cases where a *bona fide* occupational requirement can be demonstrated, such as for occupations that are considered

to be safety critical. For example, in Ontario in 1992, mandatory retirement of school bus drivers at age 65 was upheld because expert medical evidence indicated that, as a group, those over 65 are more likely to have crashes than younger drivers, and because it is impossible to test individually to determine who is likely to pose risks to others.

Canadian commercial aviation operators have been able to use the *bona fide* requirements to implement mandatory retirement ages for pilots. For example, some pilot collective agreements stipulate 60 as the compulsory age of retirement. However, legal challenges to the requirements have been many. Although pilots older than 60 are permitted to fly, international rules require them to be scheduled with a co-pilot who is younger. There is no mandatory retirement age for locomotive engineers in Canada. However, according to the [Seafarers International Union \(2008\)](#), as of 2008, the mandatory retirement age for seafarers was 70½.

Alcohol and Drug Effects

Alcohol

Alcohol effects on motor vehicle driving have been studied extensively, beginning as early as 1904 ([Moskowitz, 2007](#)). Alcohol consumption has been shown to impair visual search, the ability to divide attention, and information processing ([Moskowitz and Robinson, 1988](#)). Alcohol affects driving even at very low BAC levels and the effects are dose-related. With respect to driving tasks, alcohol is associated with impaired control of lane position, faster speeds, and slowed response time to subsidiary tasks ([Smiley, 1999](#)). As BAC level increases, more aspects of behavior are affected and are affected more strongly.

In Canada, it is a criminal offence to operate road, railway, or boating equipment or be in care and control of railway equipment while impaired by alcohol with a blood alcohol level of 80 milligrams of alcohol in 100 milliliters of blood (0.08%). There are lower (e.g., 0.05%) BAC limits in many provinces for which exceeding them is considered an administrative offence. This can result in, for example, temporary suspension of one's driver's license.

The International Maritime Organization recommends that countries implement national legislation prescribing a maximum of 0.08% BAC during watchkeeping duty as a minimum safety standard. International marine regulations issued under the STCW (Standards of Training, Certification and Watchkeeping for Seafarers) in 2011 include mandatory limits for alcohol consumption of not greater than 0.05% BAC or 0.25 mg/l alcohol in the breath, although individual flag states may choose to apply stricter limits.

Marijuana

Generally the effects of consuming marijuana are that it: "... impairs driving behavior". However, this impairment is mitigated in that subjects under marijuana treatment appear to perceive that they are indeed impaired. Where they can compensate, they do, for example by not overtaking, by slowing down and by focusing their attention when they know a response will be required. Such compensation is not possible, however, where events are unexpected or where continuous attention is required" ([Smiley, 1999](#)). A meta-analysis of experimental studies on the impairment of driving-relevant skills by alcohol or cannabis shows that a tetrahydrocannabinol (THC) concentration in serum of 7–10 ng/ml is correlated with an impairment comparable to that caused by a BAC of 0.05% ([Grotenhermen et al., 2007](#)). Effects have been shown to last a few hours, however more research is needed on this issue as a single study in aviation found effects still present 24 hours after dosing. A difficult problem arises with medical marijuana because the active ingredient in marijuana can stay in the body for several hours or longer, long after there is any effect on performance.

New Developments in Human Factors and Transportation

Automation

The role of operators of automated systems is to supervise the system, monitor its performance and, when required, intervene. The concept of "trust in automation" ([Sheridan, 1980](#); [Sheridan and Hennessy, 1984](#); [Sheridan et al., 1983](#)) describes individual differences in the use of, and reliance upon, automated systems. Trust evolves over time as one analyzes, compares and interprets the capabilities of a system ([Lee and See, 2004](#)). As with people, a pilot's trust—or distrust—of an automated system will develop with repeated exposure, and will be less likely to change as experience with the system increases. The extent that an individual will allow an automated system to perform and manage functions that could also be performed by the individual will depend on the amount of trust felt towards it.

If understanding of automation is poor, operators' expectations about how and when automation will intervene will tend to be inaccurate, and may be unreasonably high. The more trust an operator feels towards an automated system, the less often they will monitor it ([Muir and Moray, 1996](#)) and the "raw" information sources that it uses ([Parasuraman et al., 2008](#)), a situation that may lead to reduced vigilance and alertness, and less accurate situational awareness ([Parasuraman and Riley, 1997](#)). "Calibration" of trust is the correspondence between a person's level of trust in the automation and the automation's actual capabilities ([Lee and See, 2004](#)). Better calibration will occur when an operator has an accurate understanding, or mental model, of how an automated system actually works.

"Overtrust" in a system reflects poor calibration in which trust exceeds system capabilities ([Lee and See, 2004](#)), a phenomenon also known as "automation complacency" ([Parasuraman, 2000](#)). "Misuse" of automation refers to the failures that occur when people inadvertently violate critical assumptions and rely on automation inappropriately.

Rapidly developing sensor and tracking technology is being used to design and enable advanced driver assistance systems. Numerous aspects of the driving task have been assisted or automated, e.g., lane keeping, vehicle following (adaptive cruise control), blind spot detection, hazard detection while reversing, wayfinding, and automated braking to name the best known. As an example, one study showed that it may be possible to address some highway design problems with in-vehicle active warning devices. In total, 14 young drivers were monitored on an approach to an intersection near an arch-shaped bridge, where traffic crashes had often occurred due to poor visibility. Image and/or voice warning information was triggered by the presence of a stopped vehicle at the downhill road section of the intersection. Dynamic warning (that there was a vehicle ahead) was more effective than static warning (that there was a traffic signal ahead) in reducing decelerations greater than 0.2 g (Zhang et al., 2009).

High levels of automation have been present in aviation for some time, on occasion exceeding operators' understanding of exactly what the system is doing. The trust that aircraft pilots have in onboard automated systems determines whether they will rely on, or override, a system (Riley, 1989). A pilot's self-confidence, or evaluation of their own capabilities to perform what the automated system can do, also influences their use and trust of automation. If self-confidence is higher than trust, pilots will shift their preference from automatic to manual control, especially in risky situations. As with all transportation modes, designing interfaces and training that effectively provides pilots with accurate information regarding the purpose and performance of automation will enhance the calibration of pilots' trust.

Test Methods: Naturalistic Studies

Naturalistic driving studies represent a new experimental paradigm, which allows a greatly improved understanding of how various driver behaviors contribute to crashes, and how drivers respond to traffic control devices and interact with the roadway as well as how drivers respond to distractions such as cell phones and video advertising signs. Naturalistic studies involve large numbers of drivers having their personal vehicle instrumented with sensors and cameras to record many aspects of their driving behavior in minute detail, including near-crashes and crashes. The 100-car pilot study was the first of these (Klauer et al. 2006). Since then, a much more ambitious collection of over a year's worth of data from 3500 drivers in six states has been completed as part of the Strategic Highway Research Plan 2 administered by the Transportation Research Board (TRB). An example study using these data is one on distraction mentioned earlier (Victor et al. 2015). In addition to the driver data, roadway and roadside characteristics of about 12,000 miles traveled by study participants were collected. A roadway data collection vehicle drove selected roads at posted speeds and recorded roadway geometry (horizontal curvature information, grade, cross slope, lane, and shoulder information), speed limit signs, and intersection locations and characteristics.

Human Factors Guidelines

The US TRB's National Cooperative Highway Research Program (NCHRP) Report 600: Human Factors Guidelines for Road Systems: Second Edition provides data and insights on the extent to which road users' needs, capabilities, and limitations are influenced by the effects of age, visual demands, cognition, and influence of expectancies (Campbell et al., 2012). NCHRP Report 600 provides guidance for roadway location elements and traffic engineering elements. For example, in the section on nonsignalized intersections, considerations underlying required sight distance at right-skewed intersections are discussed, and in the section on signalized intersections, guidelines are given concerning the restriction of right turns on red to address pedestrian safety, and the accommodation of vision-impaired pedestrians at roundabouts.

Distraction-related guidelines also exist. In Canada in 2019, Transport Canada introduced their Guidelines to Limit Distraction from Visual Displays in Vehicles (Transport Canada, 2019). These guidelines followed the introduction in 2013 of the US NHTSA's voluntary Guidelines for Reducing Visual-Manual Driver Distraction during Interactions with Integrated, In-Vehicle, Electronic Devices (National Highway Traffic Safety Administration, 2013).

Legalization of Marijuana

Possession of marijuana is legal in a number of US states and, since October 2018, in Canada. This raises concerns about traffic safety and the need to, as has been done for alcohol, establish a legal limit. The active ingredient in marijuana is THC. A crash responsibility analysis was carried out for a sample of 3005 nonfatally injured motor vehicle drivers in British Columbia, Canada. There was no evidence of increased crash risk in drivers with THC < 5 ng/mL. Although nonsignificant results are not generally highlighted, the researchers report a statistically nonsignificant increased risk of crash responsibility (OR = 1.74) (95% CI = 0.59–6.36; $P = .35$) in drivers with THC $\geq$ 5 ng/mL.

A concern is the lack of correlation between the level of THC in blood, the dose that was ingested (smoking or eating) and performance impairment. Presence of THC in the blood indicates recent use except for chronic users who may show positive levels after several days of not using. There are few studies that test for extended effects of marijuana and so little guidance of how long operators should wait between consumption and return to work. This is particularly challenging in the case of operators using medical marijuana.

Pilots in Canada are restricted from using cannabis before flying. On June 3, 2019, Transport Canada announced a new policy stating that flight crew (pilots and flight engineers) and flight controllers (air traffic controllers) would be prohibited from the use of cannabis for at least 28 days before being on duty. In other transportation modes like rail, where many operating employees are

perpetually on-call, most companies have “zero tolerance” policies in place that require employees to be fit for duty, including not being impaired by THC.

Expert Systems

Recent years have seen the evolution of expert computer-based systems that have potential to improve transportation safety using various approaches, from providing direct feedback to users on their physiological state, to informing shift scheduling and work planning to optimize human performance, to proactively identifying crashes and incidents that could compromise safety and learning from their outcomes before an accident happens. Some of these systems are presented here.

Tools that Provide Feedback to Operators

Fatigue

There are wearable devices available that inform users as to their likely state in terms of sleep-related fatigue. For example, the Readiband is a wrist-worn device that measures sleep using an accelerometer that records wrist movement. Known as actigraphy, this technology estimates when a wearer is asleep *vs.* awake. While the algorithms used by the Readiband and the software that generates and provides feedback to wearers are fairly accurate, it can be challenging to differentiate between being still and being asleep. Regardless, these types of wearable technology show promise in their potential to improve safety.

Driving

Video- and sensor-based driver feedback systems encourage safe driving by providing drivers, often across company fleets but also for specified driver groups such as older or teen drivers, with in-vehicle feedback and video-based coaching. These systems can be used by drivers themselves, parents, or fleet managers to inform safety culture and provide incentives regarding driver safety. Driver rankings, data-driven rewards programs, and training are outcomes that can be used to improve overall safety.

Tools to Inform Planning

Fatigue Scheduling Systems

There are several software decision aides available today that were designed to assess and forecast performance, based on variables such as sleep restriction and time of day, to design work (and sleep) schedules to reduce the risk of fatigue and the resultant performance decrements. Based on biomathematical models developed by the military, the accuracy of the tools in predicting fatigue is significant. For example, the Fatigue Avoidance Scheduling Tool (FAST) was developed by the United States Air Force in 2000–2001 to address the problem of aircrew fatigue in aircrew flight scheduling. FAST software displays work and sleep data entry in graphic, symbolic (grid) and text formats, to shows cognitive performance effectiveness as a function of time. The calculated performance effectiveness represents composite human performance on a number of cognitive tasks, scaled from zero to 100%. The goal of the user/scheduler is to keep performance effectiveness at or above 90% by manipulating the timing and lengths of work and rest periods.

Tools that Inform Safety Countermeasures

Road

Safety Analyst is a set of software tools used by state and local highway agencies for highway safety management. Safety Analyst was developed by FHWA (US Federal Highway Administration) in cooperation with participating state and local agencies and is available through AASHTO (American Association of State Highway and Transportation Officials). Safety Analyst automates procedures to assist highway agencies in implementing the highway safety management process, including: network screening, diagnosis, countermeasure selection, and evaluation. The diagnosis tool guides the user through a series of questions answered by means of office and field investigations concerning how the driver is likely to interact with the traffic control devices and highway design, the potential for driver error and appropriate countermeasures.

Railway Level Crossing Safety

Federally regulated railway companies and road authorities in Canada are jointly responsible for the maintenance and safety of level crossings. Transport Canada’s Grade Crossing Inventory uses an internal web-based analysis tool to compare crossings against each other and identify those that are most in need of safety improvement. The tool assesses the following risk factors:

- TSB data on rail occurrences;
- the volume of road and railway traffic;
- maximum train and vehicle speeds;
- number of tracks and lanes;
- urban or rural environment; and
- warning systems in place at the crossing (i.e., gates, bells, lights).

Aviation

The US Aviation Safety Reporting System captures confidential reports, analyzes the resulting aviation safety data, and disseminates vital information to the aviation community. It brings together government, industry, and individuals in the aviation community to maintain and improve aviation safety by collecting voluntarily submitted aviation safety incident/situation reports from pilots, air traffic controllers, and others. The system acts on the information in reports by identifying system deficiencies and issuing alerting messages to people who are in a position to correct them, and by educating more generally through a regular newsletter, a scientific journal, and through its research studies.

References

- AASM, 2001. The International Classification of Sleep Disorders. Revised: Diagnostic and Coding Manual. American Academy of Sleep Medicine, Chicago, Illinois.
- Alexander, G., Lunenfeld, H., 1975. Positive guidance in traffic control. Federal Highway Administration, Washington, DC.
- Allen, M.J., Hazlett, R.D., Tacker, H.L., Graham, B.V., 1970. Actual pedestrian visibility and the pedestrian's estimate of his own visibility. *Am. J. Optom. Arch. Am. Acad. Optom.* 47, 44–49.
- Arbus, L., Tiberge, M., Serress, A., Rouge, D., 1991. Drowsiness and traffic accidents. Importance of diagnosis. *Neurophysiol. Clin.* 21 (1), 39–43.
- Ayas, N., Skomro, R., Blackman, A., et al., 2014. Obstructive sleep apnea and driving: a Canadian thoracic society and Canadian sleep society position paper. *Can. Respirat. J.* 21, 114–123.
- Ayas, N.T., Fitzgerald, J.M., Fleetham, J., et al., 2006. Cost-effectiveness of continuous positive airway pressure therapy for moderate to severe obstructive sleep apnea/hypopnea. *Arch. Int. Med.* 166, 977–984.
- Ball, K., Beard, B., Roenker, D., Miller, R.L., Griggs, D.S., 1988. Age and visual search: expanding the useful field of view. *J. Opt. Soc. Am.* 5 (12), 2210–2219.
- Bearpark, H., Fell, D., Grunstein, R.R., Leeder, S., Berthon-Jones, M., Sullivan, C., 1990. Road safety and pathological sleepiness: The role of sleep apnea. Sponsored by the Roads and Traffic Authority, NSW and the Federal Office of Road Safety; Road Safety Bureau Consultant's Report CR 3/90, Canberra, Australia.
- Beaumont, M., Batejat, D., Pierard, C., Coste, O., Doireau, P., Van Beers, P., et al., 2001. Slow release caffeine and prolonged (64-h) continuous wakefulness: Effects on vigilance and cognitive performance. *J. Sleep Res.* 10 (4), 265–276.
- Berman, B., Dismukes, R.K., 2006. Pressing the approach. *Aviat. Saf. World* 1 (6), 28.
- Boff, K.R., Kaufman, L., Thomas, J.P., 1986. Handbook of Perception and Human Performance. Volume 1: Sensory Processes and Perception. John Wiley and Sons.
- Bullough, J.D., Rea, M.S., 2000. Simulated driving performance and peripheral detection at mesopic and low photopic light levels. *Lighting Res. Technol.* 32 (4), 194–198.
- Campbell, J.L., Lichty, M.G., Brown, J.L., Richard, C.M., Graving, J.S., Graham, J., et al., 2012. Human Factors Guidelines for Road Systems, second ed. Transportation Research Board. NCHRP Report 600, Washington, DC.
- Chen, L.-H., Baker, S.P., Braver, E.R., Li, G., 2000. Carrying passengers as a risk factor for crashes fatal to 16 and 17 year old drivers. *J. Am. Med. Assoc.* 28 (12), 1578–1582.
- Chung, F., Elsaid, H., 2009. Screening for obstructive sleep apnea before surgery: Why is it important? *Current Opinion in Anaesthesiology.* 22 (3), 405–411.
- Cohen, A.S., 1987. The latency of simple reaction on highways: a field study. *Public Health Rev.* 15, 291–310.
- Cole, B.L., Hughes, P.K., 1984. A field trial of attention and search conspicuity. *Hum. Fact.* 26 (3), 299–313.
- Cooper, D., Atkins, F., Gillen, D., 2005. Measuring the impact of passenger restrictions on new teenage drivers. *Accid. Anal. Prevent.* 37 (1), 19–23.
- Dagan, Y., Doljansky, J., Green, A., Weiner, A., 2006. Body mass index (BMI) as a first-line screening criterion for detection of excessive daytime sleepiness among professional drivers. *Traff. Injury Prevent.* 7 (1), 44–48.
- Dewar, R.E., Olson, P.L., 2007. Age Differences: Drivers Old and Young, in: *Human Factors in Traffic Safety*, second ed. ISBN 1-933264-24-1. Lawyers & Judges Publishing Company, Tucson, Arizona (Chapter 8).
- Diller, E., Cook, L., Leonard, D., Dean, J.M., Reading, J., Vernon, D., 1999. Evaluating drivers licensed with medical conditions in Utah, 1992–1996. National Highway Traffic Administration. Report No. DOT HS 809 023, Washington, DC.
- Eilen, R.L., Marshall, S.C., Palayew, M., Molnar, F.J., Wilson, K.G., Man-Son-Hing, M., 2006. Systematic review of motor vehicle crash risk in persons with sleep apnea. *J. Clin. Sleep Med.* 2, 193–200.
- Endsley, M.R., 1995. Toward a theory of situation awareness in dynamic systems. *Hum. Fact.* 37 (1), 32–64.
- Endsley, M.R., 2010. Situation awareness in aviation systems. In: Wise, J.A., Hopkin, V.D., Garland, D.J. (Eds.), *Handbook of Aviation Human Factors*. CRC Press, Boca Raton (Chapter 12).
- FAA (Federal Aviation Administration), 2016. Advisory Circular 90-48D: Pilots' Role in Collision Avoidance, issued April 19, 2016.
- Farber, E., Silver, C.A., 1967. Knowledge of oncoming car speed as a determiner of drivers' passing behavior. *Highw. Res. Rec.* 195, 52–65.
- Federal Railroad Administration, 2019. Train accident cause codes - Appendix C, Office of Safety Analysis. Available from: <https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/downloads/appendixC-TrainaccidentCauseCodes.aspx?State=0>.
- Fitch, G., Soccolich, S.A., Guo, F., McClafferty, K.J., Fang, Y., Olson, R.L., et al., 2013. The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk. U.S. Department of Transportation - Report DOT HS 811 757, Washington, DC.
- Grandjean, E., 1982. Fitting the Task to the Man: An Ergonomic Approach. Taylor and Francis Ltd, London.
- Grandjean, E., 1988. Fitting the Task to the Man: A Textbook of Occupational Ergonomics. Taylor & Francis Ltd, London.
- Grotenhermen, F., Leson, G., Berghaus, G., Drummer, O.H., Krüger, H.P., Longo, M.C., et al., 2007. Developing limits for driving under cannabis. *Addiction* 102 (12), 1910–1917.
- Guyton, A.C., 1969. *Function of the Human Body*. W.B. Saunders Company, Philadelphia, PA.
- Hildebrand, E.D., 2012. Aging school bus drivers: Is mandatory retirement appropriate. In: *Proceedings of the 22nd Canadian Multidisciplinary Road Safety Conference*, Banff, Alberta, June 10–13.
- Hill, M., 2007. A study of the role of human factors in railway occurrences and possible mitigation strategies. Transport Canada report number T8080-07-0052.
- Hitchcock, L., Bourgeois-Bougrine, S., Cabon, P., 2010. Pilot Performance. In: Wise, J.A., Hopkin, V.D., Garland, D.J. (Eds.), *Handbook of Aviation Human Factors*, CRC Press, Boca Raton, Chapter 12.
- Horne, J.A., Reyner, L.A., 1995. Sleep related vehicle accidents. *Br. Med. J.* 310, 565–567.
- Hosking, S.G., Liu, C., Bayly, M., 2010. The visual search patterns and hazard responses of experienced and inexperienced motorcycle riders. *Accid. Anal. Prevent.* 42 (1), 196–202.
- IIHS, 2012. *Fatality Facts 2010*. Insurance Institute for Highway Safety, Arlington, VA.
- Kantowitz, B., Sorkin, R.D., 1983. *Human Factors: Understanding People-system Relationships*. Wiley, New York.
- Karimi, M., Eder, D.N., Eskandari, D., Zou, D., Hedner, J.A., Grote, L., 2013. Impaired vigilance and increased accident rate in public transport operators is associated with sleep disorders. *Accid. Anal. Prev.* 51, 208–214.
- Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J., Ramsey, D., 2006. The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. Report No. DOT HS 810 594. National Highway Traffic Safety Administration, Washington, DC.
- Koppel, S., Charlton, J., Kopinathan, C., Taranto, D., 2011. Are child occupants a significant source of driving distraction. *Accid. Anal. Prevent.* 43 (3), 1236–1244.

- Lam, L.T., Norton, R., Woodward, M., Connor, J., Ameratunga, S., 2003. Passenger carriage and car crash injury: a comparison between younger and older drivers. *Accid. Anal. Prev.* 35 (6), 861–867.
- Lansdown, T.C., 2012. Individual differences and propensity to engage with in-vehicle distractions - A self-report survey. *Trans. Res. F* 15 (1), 1–8.
- Lee, J.D., 2007. Technology and teen drivers. *J. Safety Res.* 38 (203), 213.
- Lee, J.D., See, K.A., 2004. Trust in automation: designing for appropriate reliance. *Hum. Fact.* 46 (1), 50–80.
- Lee, J.D., Young, K.L., Regan, M.A., 2009. Defining driver distraction. In: Regan, M.A., Lee, J.D., Young, K.L. (Eds.), *Driver Distraction: Theory, Effects, and Mitigation*, CRC Press, Boca Raton, Florida.
- Liu, C., Hosking, S.G., Lenne, M.G., 2009. Hazard perception abilities of experienced and novice motorcyclists: an interactive simulator experiment. *Trans. Res. F Traff. Psychol. Behav.* 12 (4), 325–334.
- Lurie, A., 2011. Obstructive Sleep Apnea in Adults. *Adv Cardiol. Basel, Karger*, vol 46, pp 1–42.
- Mandelbaum, J., Sloan, L.L., 1947. Peripheral visual acuity. *J. Ophthalmol.* 30, 581–588.
- Martens, M.H., Brouwer, R.F.T., 2013. Measuring being lost in thought: An exploratory driving simulator study. *Trans. Res. F Traff. Psychol. Behav.* (20), 17–28.
- McCafferty, D.B., Baker, C.C., 2016. Trending the causes of marine incidents. In: *Learning From Marine Incidents Conference*, London, UK. January 25–26.
- Moskowitz, H., 2007. Alcohol and drugs. In: Dewar, R.E., Olson, P.L. (Eds.), *Human Factors in Traffic Safety*. Lawyer & Judges Publishing Company, Tucson, Arizona (Chapter 7).
- Moskowitz, H., Robinson, C.D., 1988. Effects of low doses of alcohol on driving-related skills: a review of the evidence. SRA Technologies Inc. Report #DOT HS 807280 prepared for the National Highway Traffic Safety Administration, Washington, DC.
- Mourant, R.R., Rockwell, T.H., 1970. Mapping eye-movement patterns to the visual scene in driving: an exploratory study. *Hum. Fact.* 12 (1), 81–87.
- Muir, B.M., Moray, N., 1996. Trust in automation Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 429–460.
- Nasar, J., Hecht, P., Wener, R., 2008. Mobile telephones, distracted attention, and pedestrian safety. *Accid. Anal. Prev.* 40 (1), 69–75.
- National Highway Traffic Safety Administration, 2008. National motor vehicle crash causation survey: Report to Congress. DOT HS 811 (2008): 059, Washington, DC.
- National Highway Traffic Safety Administration, 2013. Visual-manual NHTSA driver distraction guidelines for in-vehicle electronic devices. Available from: https://www.nhtsa.gov/staticfiles/nti/distracted_driving/pdf/distracted_guidelines-FR_04232013.pdf.
- National Highway Traffic Safety Administration, 2017. Traffic safety facts - bicyclists and other cyclists - 2015 data. U.S. Department of Transportation - DOT HS 812 382. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812382>.
- National Highway Traffic Safety Administration, 2018. Traffic safety facts - pedestrians - 2016 data. U.S. Department of Transportation - DOT HS 812 493. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812493>.
- NTSB, 1995. Safety study: factors that affect fatigue in heavy truck accidents. NTSB PB95-017001-SS-95/01. U.S. National Transportation Safety Board.
- Olson, P.L., 1988. Minimum requirements for adequate nighttime conspicuity of highway signs. Final Report for the University of Michigan Transportation Research Institute (UMTRI). Report No. UMTRI-88-8, Ann Arbor, MI.
- Olson, P.L., 1996. Forensic Aspects of Driver Perception and Response. Lawyers & Judges Publishing Company, Tucson, Arizona.
- Olson, P.L., Farber, E., 2003. Forensic Aspects of Driver Perception and Response, second ed. Lawyers & Judges Publishing Company, Tucson, Arizona.
- Olson, P.L., Sivak, M., 1983. Improved low-beam photometrics. Rep. No. UMTRI-83-9. University of Michigan Transportation Research Institute, Ann Arbor, MI.
- Orasanu, J., Martin, L., Davison, J., 1998. Errors in aviation decision making: bad decisions or bad luck? In: *Fourth Conference on Naturalistic Decision Making*, NASA-Ames Research Center, Warrington, Virginia, May 29–31.
- Oswald, I., 1966. Sleep: Fourth Revised Edition. Penguin (Non-Classics) Books, London, UK.
- Owsley, C., Ball, K., Sloane, M.E., Roenker, D., Bruni, J.R., 1991. Visual/cognitive correlates of vehicle accidents in older drivers. *Psychol. Aging* 6, 401–405.
- Parasuraman, R., 2000. Designing automation for human use: empirical studies and quantitative methods. *Ergonomics* 43, 931–951.
- Parasuraman, R., Riley, V., 1997. Humans and automation: Use, misuse, disuse, abuse. *Hum. Fact.* 39, 230–253.
- Parasuraman, R., Sheridan, T., Wickens, C.D., 2008. Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *J. Cognit. Eng. Decision Making* 2 (2), 140–160.
- Philipp, P., Akerstedt, T., 2006. Transport and industrial safety. How are they affected by sleepiness and sleep restriction. *Sleep Med. Rev.* 10 (5), 347–356.
- Puisa, R., Vassalos, D., Bolbot, V., 2018. Unravelling causal factors of maritime incidents and accidents. *Safety Sci.* 110, 124–141.
- Regan, M.A., Lee, J.D., Young, K.L., 2009. Driver Distraction: Theory, Effects, and Mitigation. Taylor & Francis Group, Boca Raton, Florida.
- Riley, V., 1989. A general model of mixed-initiative human-machine systems. In: *Proceedings of the 33rd Annual Meeting of the Human Factors Society*, pp. 124–128.
- Rockwell, T.H., 1988. Spare visual capacity in driving - revisited. In: Gale, A.G., et al. (Eds.), *Vision in Vehicles II*. Elsevier Science Publishers B.V., North Holland.
- Sagberg, F., 2006. Driver health and crash involvement: a case-control study. *Accid. Anal. Prev.* 39, 28–34.
- Seafarers International Union, 2008. A guide to your benefits from the Seafarers Pension Plan. SPP 05/08. Available from: http://seafarers.org/downloads/Member%20Benefits%20and%20Resources/SPP%20Documents/SPP_SPD_2008.pdf.
- Sheridan, T., Fischhoff, B., Posner, M., Pew, R.W., 1983. Supervisory control systems. In: *Research Needs for Human Factors*, National Academy Press, Washington.
- Sheridan, T., Hennessy, R.T., 1984. Research and Modeling of Supervisory Control Behavior. National Academy Press, Washington.
- Sheridan, T.B., 1980. Computer control and human alienation. *Technol. Rev.* Oct. 61–73.
- Simons-Morton, B., Lerner, N., Singer, J., 2005. The observed effects of teenage passengers on the risky driving behavior of teenage drivers. *Accid. Anal. Prev.* 37 (6), 973–982.
- Sivak, M., Olson, P.L., Pastalan, L., 1981. The effect of driver's age on nighttime legibility of highway signs. *Hum. Fact.* 23 (1), 59–64.
- Smiley, A., 1999. Marijuana: On-road and driving simulator studies. In: Kalant, J., Corrigall, W., Hall, W., Smart, R. (Eds.), *The Health Effects of Cannabis*. The Centre for Addiction and Mental Health, Toronto, Canada (Chapter 5).
- Steele, M.T., Ma, O.J., Watson, W.A., Thomas Jr.H., Muelleman, R.L., 1999. The occupational risk of motor vehicle collisions for emergency medicine residents. *Acad. Emer. Med.* 6, 1050–1053.
- Stelling, A., Hagenzieker, M.P., 2012. Afleiding in het verkeer. Een overzicht van de literatuur (in Dutch with English summary). SWOV Institute for Road Safety Research.
- Strayer, D.L., Cooper J.M., Turrill, J., Coleman, J., Madeiros-Ward, N., Biondi, F., 2013. Measuring cognitive distraction in the automobile. AAA Foundation for Traffic Safety, Washington, DC. Available from: <https://www.aaafoundation.org/sites/default/files/MeasuringCognitiveDistractions.pdf>.
- Stutts, J.C., Faeganes, J., Rodgman, E., Hamlett, C., Meadows, T., Reinfurt, D.W., 2003. Distractions in everyday driving. AAA Foundation for Traffic Safety, Washington, DC. Available from: <http://www.aaafoundation.org/pdf/distractionsineverydaydriving.pdf>.
- Tefft, B.C., Williams, A.F., Grabowski, J.G., 2012. Teen driver risk in relation to age and number of passengers. AAA Foundation for Traffic Safety. Available from: https://www.aaafoundation.org/sites/default/files/research_reports/2012TeenDriversRiskAgePassengers.pdf.
- Teran-Santos, J., Jimenez-Gomez, A., Cordero-Guevara, J., 1999. The association between sleep apnea and the risk of traffic accidents. *N. Eng. J. Med.* 340 (11), 847–851.
- Transport Canada, 2001. Road safety in Canada - TP: 15145E. Available from: <http://www.tc.gc.ca/media/documents/roadsafety/tp15145e.pdf>.
- Transport Canada, 2018. Canadian rail operating rules, general rules A (xi, xii). Available from: <https://www.tc.gc.ca/eng/railsafety/rules-tco167.htm>.
- Transport Canada, 2019. Transport Canada guidelines to limit distraction from visual displays in vehicles. Available from: <https://www.tc.gc.ca/en/services/road/stay-safe-when-driving/guidelines-limit-distraction-visual-displays-vehicles.html>.
- Tversky, A., Kahneman, D., 1982. Causal schemas in judgments under uncertainty. In: Kahneman, D., Slovic, P., Tversky, A. (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Press Syndicate of the University of Cambridge, New York, N.Y.
- Vaa, T., 2003. Impairment, diseases, age and their relative risks of accident involvement: results from a meta-analysis. TOI Report 690.
- Vanlaar, W., Simpson, H.M., Mayhew, D.R., Robertson, R., 2008. Fatigued and drowsy driving: a survey of attitudes, opinions and behaviors. *J. Safety Res.* 39 (3), 303–309.

- Victor, T., Dozza, M., Bargman, J., Boda, C.-N., Engstrom, J., Flannagan, C., et al., 2015. Analysis of naturalistic driving study data: safer glances, driver inattention, and crash risk. SHRP 2 Report S2-S08A-RW-1, Transportation Research Board, Washington, DC.
- White, C., Caird, J.K., 2010. The blind date: the effects of change blindness, passenger conversation and gender on looked-but-failed-to-see (LBFTS) errors. *Accid. Anal. Prev.* 42, 1822–1830.
- World Health Organization (WHO), 2011. Mobile phone use: a growing problem of driver distraction, Geneva, Switzerland. Available from: https://www.who.int/violence_injury_prevention/publications/road_traffic/distracted_driving/en/.
- Young, K., Regan, M., 2007. Driver distraction: a review of the literature. In: Faulks, I.J., REgan, M., Stevenson, M., Brown, J., Porter, A., Irwin, J.D. (Eds.), *Distracted Driving*. Australasian College of Road Safety, Sydney, NSW, pp. 379–405.
- Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S., Badr, S., 1993. The occurrence of sleep-disordered breathing among middle aged adults. *N. Eng. J. Med.* 328 (17), 1230–1273.
- Zhang, J., Suto, K., Fujiwara, A., 2009. Effects of in-vehicle warning information on drivers' decelerating and accelerating behaviors near an arch-shaped intersection. *Accid. Anal. Prev.* 41 (5), 948–958.

Further Reading

- Krauss, D., 2015. *Forensic Aspects of Driver Perception and Response*, fourth ed. Lawyers & Judges Publishing Company, Inc., Tucson, AZ.
- Shinar, D., 2017. *Traffic Safety and Human Behavior*, second ed. Elsevier, Amsterdam.
- Smiley, A. (Ed.), 2015. *Human Factors in Traffic Safety*, third ed. Lawyers & Judges Publishing Company, Tucson, AZ.
- Staplin, L., Lococo, K., Byington, S., Harkey, D., 2001a. *Highway Design Handbook for Older Drivers and Pedestrians*. U.S.DOT/FHWA Publication No. FHWA-RD-01-103. U.S. Federal Highway Administration (FHWA), Washington, DC.
- Staplin, L., Lococo, K., Byington, S., Harkey, D., 2001b. *Guidelines and Recommendations to Accommodate Older Drivers and Pedestrians*. U.S.DOT/FHWA Publication No. FHWA-RD-01-051. U.S. Federal Highway Administration (FHWA), Washington, DC.
- Sussman E.D., Raslear, T.G., 2007. Railroad human factors, in: Boehm-David, D.A. (Ed.), *Reviews of Human Factors and Ergonomics*, vol. 3, pp. 148–189.

In-Depth Crash Analysis and Accident Investigation

Yong Peng, Helai Huang, Xinghua Wang, School of Traffic & Transportation Engineering, Central South University, Changsha, China

© 2021 Elsevier Ltd. All rights reserved.

In-Depth Traffic Accident Investigation	346
Accident Epidemiology Analysis	346
Traffic Accident Reconstruction	348
Injury Biomechanics Investigation Methods	348
Injury Mechanisms and Criterion	349
References	350

In-Depth Traffic Accident Investigation

In-depth traffic accident investigations are done in the airline, railroad, and shipping industry, as well as for select roadway crashes. This article focuses on roadway crashes only. An in-depth traffic accident investigation is a multi-interdisciplinary work, in which a large number of accident-related variables are collected from each level of road traffic system (Fig. 1). In-depth traffic accident data contribute to improving the road design, vehicle safety, medical services, traffic management, etc.

Governmental agencies as well as nonprofit and industry-lead investigations are done. To help find the causes of auto accidents and fatalities, Volvo established the industry's first in-depth auto accident investigation team in 1970. The Volvo safety teams annually investigate 100 accidents at the scene, immediately after they occur and another 2500 accidents are analyzed statistically each year. Already by the year 2000, Volvo's safety division has accumulated data on 28,000 crashes involving Volvo cars carrying more than 40,000 occupants. In Finland, all fatal road traffic accidents are investigated at the site by multidisciplinary Accident Investigation Teams (based on the Act on the Investigation of Road and Cross-Country Traffic Accidents from 2001). The Finnish Motor Insurers' Centre maintains the investigations. The purpose is to find out what happened in the accident, uncover risk factors, and give safety recommendations. Moreover, various in-depth traffic accident investigation projects also have been established in other countries, such as the German In-Depth Accident Study (GIDAS) project, China In-Depth Accident Study (CIDAS) project, and Crash Injury Research and Engineering Network (CIREN) project.

Haddon Matrix Model, a "three factors and three phases" theory, is often used to analyze the accident causations and corresponding prevention countermeasures (Haddon, 1968). Based on the analysis model, in in-depth traffic accident investigation, a whole traffic accident process is divided into three phases: precrash, crash, and postcrash. Precrash stage is divided into two intervals: (1) the interval between normal driving and danger detection, mainly includes participants' driving states (e.g., emotion, drinking, and visual obstacle); and (2) the interval between danger detection and crash occurrence, mainly includes participants' judgments (e.g., estimating the behaviors of other traffic participants) and participants' decisions (e.g., braking and steering). Crash stage refers to the interval between crash occurrence and eventual stopping, mainly includes initial impact parameters (e.g., impact speed and contact positions between participants), participants' decisions (e.g., braking and steering), and participants' impact responses (e.g., pedestrians' motions during a collision). Postcrash stage refers to the afterward disposal of a traffic accident, mainly includes the accident consequences (e.g., final positions of participants, vehicle damage information, and human injury information).

With the development of vehicle automation and electrification, precrash parameters, especially the driving psychology and driving behaviors, have attracted increasing attention. A series of analysis methods were proposed to further understand the driving psychology. Especially, the human failure model (HFF), allowing the classification of the driving functional failures (e.g., human errors and capacity exceeding) and of the factors of these failures, distinguishes five major functional categories within which can be identified the incapacity of a function (perceptive, diagnostic, prognostic, decision-making, and action-taking) to overcome a difficulty encountered by the traffic participant (Van Elslande and Fournier, 2017).

Accident video is an effective tool to accurately acquire precrash driving behaviors. In the view of requirement of the automobile industry, some neonatal in-depth traffic accident investigation projects are initialed, such as the Traffic Accident Investigation and Research in China (TAIRC) project—it focuses on these accidents captured on video or involving electric/autonomous vehicles. However, vehicle-mounted video cameras are often present in commercial vehicles such as buses and limousines but typically not in private vehicles in most countries. Also, video recordings of intersections are fairly common, but few segments between intersections have recordings of what actually happened in a crash. We do, therefore, have a long way to go to achieve this goal around the world.

Accident Epidemiology Analysis

Accident epidemiology analysis mainly includes: (1) the analysis of the development tendency and distribution regularities of traffic accident in a specific region. With the development of high-precision maps, the analyzable specific regions are becoming smaller and smaller, and the analyzable influencing factors are getting larger and larger; (2) the exploration of effects of various driver,

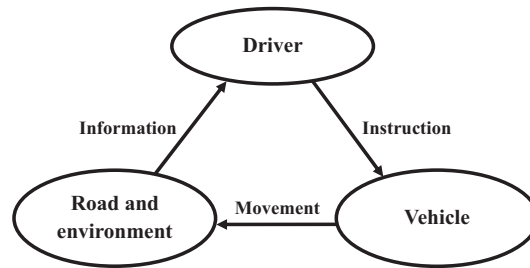


Figure 1 Road traffic system.

pedestrian, and neighborhood factors (e.g., gender, age, economy, etc.) on accident severity, in which the classification of influencing factors is the research emphasis.

Statistical analysis, based on in-depth accident data, is the most fundamental means of investigating the crash characteristics and mechanisms. Epidemiology is a subject of investigating the distribution and determinants of health-related conditions and events in a specific population, and the strategies and measures of preventing disease and improving health. This research can be divided into two categories: descriptive and analytic epidemiology. A reasonable accident epidemiology analysis can be achieved with the different types of accident data categorized based on accident type (e.g., vehicle-vehicle and vehicle-pedestrian accidents), accident region (e.g., urban and rural), road section (e.g., straight road, curve, and junction), etc. Descriptive accident epidemiology focuses on basic characteristics of any particular type of crash-related events, such as what, who, when, and where, which reflects the central tendency and discrete degree of each parameter. Analytic accident epidemiology attempts to excavate the contributing factors of any particular type of crash-related events, such as which and how, which reveals the effect of each parameter on crash severity.

Over the decades, lots of research has been conducted to analyze crash severities. In these crash-severity models, the causal factors are defined from the perspective of driver, vehicle, roadway, and environment, such as driver behaviors, vehicle performance, traffic control elements, visibility conditions, and so on. The dependent variables are either a binary outcome (e.g., fatal and nonfatal) or a multiple outcome (e.g., fatal, seriously injured, slightly injured, and noninjured). And the dependent variables with multiple outcomes can also be differentiated as ordinal outcomes (considering the order of injury-severity levels) and nominal outcomes (e.g., unordered). The appropriate statistical methodology is selected based on data characteristics: (1) underreporting of crashes, (2) ordinal nature of crash severities, (3) fixed parameters, (4) omitted variable bias, (5) small sample size, (6) endogeneity, (7) within-crash correlation, and (8) spatial and temporal correlations. The existing crash injury-severity analysis models are broadly divided into three categories: binary outcome model, ordered discrete outcome model, and unordered multinomial discrete outcome model. Lots of variations of primitive models have been conducted to achieve concrete analysis targets (Savolainen et al., 2011).

Based on simple binary outcome models, the Bayesian hierarchical binary logit model and simultaneous binary logit model are proposed to obtain more accurate parameter estimations by considering injury correlations when investigating the injury-severity levels of multiple individuals involved in same crashes. The bivariate and multivariate binary probit models are developed to deal with these cases in which explanatory variables could be endogenous in regard to injury severity, such as drivers' decisions to a vehicle with airbags or antilock brakes may be related to crash possibility and injury severity, because the endogenous variables directly as explanatory predictors will cause biased parameter estimations. Traditional ordered discrete outcome models have two obvious drawbacks: (1) they are particularly vulnerable to underreporting of crashes; and (2) they place on the way variables influence outcome probabilities. On the basis of traditional ordered probability models, a copula-based multivariate approach is applied to analyzing the injury-severity levels of multiple individuals involved in same crashes. Bivariate ordered probit model, which essentially is a layered architecture of two equations that reflect a simultaneous relationship, is used to address these issues of endogeneity. One assumption of ordered outcome models is that the error variances are homoscedastic. Heterogeneous choice model is established to solve the possibility that error terms may be heteroskedastic with respect to injury severity. Another assumption of ordered outcome models is that the parameter estimations are constant across injury-severity levels. Generalized ordered logit model is built as an alternative when the assumption is violated. Unordered multinomial discrete outcome models, which do not consider the ordinal nature of injury-severity outcomes, have attracted increasing attention, because they are not affected by these restrictions existing in traditional ordered logit and probit models. Multinomial logit models, which include three or more outcomes and do not consider the ordering of injury-severity levels, are the most basic unordered discrete outcome models. However, the models are particularly vulnerable to correlation of unobserved effects from one injury-severity level to another. The correlation results in a violation of the model's independence of irrelevant alternatives (IIA) property. Sequential logit and probit models, which essentially are generalized standard ordered logit and probit models, allow the treatment of the severity thresholds across the ordered response levels by separate parameter coefficients for explanatory variables and heterogeneity in the effects of injury-severity determinants. Nested logit models, in which the injury-severity levels are partitioned into nests that consist of injury-severity outcomes that share some unobserved elements specific to only those outcomes, overcome the IIA limitation of multinomial logit models and improve the sequential logit model by allowing for correlations among error terms across different injury-severity levels. Except these common methods, various novel methods have been used to analyze the injury-severity levels, such as artificial neural networks and classification and regression tree approach.

Accident epidemiology analysis technology is becoming more systematic, refined, and hierarchical. Establishing a decision support system by integrating traffic accident data and other multisource big data is the future development trend, which can be used to analyze the contributing factors related to traffic accidents and predict the traffic security situation in a specific region.

Traffic Accident Reconstruction

Traffic accident reconstruction, based on in-depth accident data, is an important approach of understanding the accident mechanisms and impact responses of participants. It is easy to build targeted prevention countermeasures that comparatively analyzing the differences between different types of traffic participants (e.g., pedestrian, bicyclist, and motorcyclist) in impact responses by traffic accident reconstruction (Peng et al., 2012). It can be classified into inverse accident reconstruction (calculating inputs according to known outputs) and forward accident reconstruction (calculating outputs according to tentative inputs). Inverse accident reconstruction, which is essentially theoretic calculations by mathematical equations, was originally used for accident identification. Based on accident scene information (e.g., braking distance and pedestrian throw distance), vehicle damage information (e.g., the deflection of windshield plate), and human injury information (e.g., pedestrian head injury severity), some impact parameters (e.g., vehicle crash speed) can be obtained through inverse accident reconstruction using the kinetic energy conservation law, the momentum conservation law, etc. However, there often are larger computational errors since the mechanical models used in this method are too simplistic. And the integrity of this method is poor—only obtain a certain parameter value at some point. With the development of computer technology, numerical simulation is applied to traffic accident reconstruction. Based on corresponding environment, vehicle, and human models established in numerical simulation platform, forward accident reconstruction can restore a whole accident process by tentatively entering appropriate impact parameters. Forward accident reconstruction is more reliable and integrated compared with inverse accident reconstruction. However, the reliability of results generated by this method is often detracted since some impact parameters (e.g., the position of impact point) are hard to accurately confirmed and measured at accident scene.

Traffic accident reconstruction can be broadly divided into three phases: information collection, information extraction, and accident reconstruction. To improve its reliability, lots of researches have been conducted from the earlier three aspects. For example, the Unmanned Aerial Vehicle (UAV) and Event Data Recorder (EDR) have been applied to accident investigation to collect more detailed and accurate accident information. The video parameter extraction method has been continually improved to obtain more reliable accident data. For accident reconstruction, the total uncertainty of reconstruction results has three origins: (1) the uncertainty in measurements—lots of typical measurements taken by investigators on the place of collision follow normal distribution; (2) the uncertainty in calculation refers to the scatter of the values of some parameters (e.g., the driver's response time and friction coefficient) that are not measured directly but taken from tables; and (3) the uncertainty of modeling occurs when two or more mathematical models give different analysis results with respect to same physical problem (Wach and Unarski, 2007). Especially, to understand the effects of random variable uncertainty on unknown variable uncertainty, various uncertainty analysis methods have been applied to quantify the sources of uncertainty and propagate the uncertainty from random variables to unknown variables through mathematical models of physical problems. These methods can be classified into two categories: (1) deterministic methods, such as Total Differential Method (TDM), Upper and Lower bound Method (ULM), Finite Difference Method (FDM), and Response Surface Method (RSM), by which the ranges of reconstruction results under a certain confidence level can be obtained; and (2) probabilistic methods, such as Gauss Method, Methods with the Use of Stochastic Processes, Probabilistic Perturbation Method, and Monte Carlo Method (MCM), by which the probability distributions of reconstruction results can be acquired (Cai et al., 2014).

Forward accident reconstruction is essentially a highly user-interactive and time-consuming tentative solving process, in which investigators need to constantly adjust the crash parameters to make the collision results (e.g., the contact position between participants and the final position of each participant) coincide well with the actual. Furthermore, whether the reconstruction results are acceptable is subjectively determined by investigators, which reduces the persuasion of reconstruction results largely in accident responsibility cognizance. To improve the efficiency of accident reconstruction and the reliability of reconstruction results, a viable solution is the optimization technique (Untaroiu et al., 2009). Especially, a collision optimizer has been integrated into PC-Crash—a widely used traffic accident reconstruction software (Moser et al., 2003).

The automatic driving simulation test, namely establishing a mathematical model containing static actual environment and dynamic traffic scene and then testing an automatic driving algorithm in the virtual traffic environment, is an important tool of examining the safety of autonomous vehicles. It can be easily obtained from traffic accident reconstruction that the kinematics parameters of each participant before a collision, which contributes to constructing accurate precrash accident scene. Integrating traffic accident reconstruction and automatic driving simulation may be a future development trend.

Injury Biomechanics Investigation Methods

Biomechanics is a multi-interdisciplinary subject in which the mechanics connects engineering with medicine. These methods exploring the injury biomechanics of traffic participants mainly include animal experiment, cadaver experiment, mechanical model experiment (dummy and subsystem impactor experiments), and computer numerical simulation.

Injury researches using anesthetized animals instead of human can obtain important injury data of some special parts (e.g., brain and spine). However, investigators have to consider lots of effect factors when extrapolate human injury data from animal

experiment data, since there are significant differences between human and animals in size, constitution, etc. To weaken the restriction, most animal experiments were conducted using primates most similar to human (e.g., monkey and chimpanzee). For instance, some researchers stunned 45 monkeys by accelerating their heads along the oblique, sagittal, lateral direction and found it was consistent that the symptoms between human and monkeys in coma and axonal injuries (Gennarelli et al., 1982).

Human cadaver experiment, also known as past mortal human subject test, is an important approach for investigating human injury biomechanics and verifying the validity of mechanical test model and mathematical simulation model. Compared with animal experiment, the material of a human cadaver is anthropologically identical to that of a living person. However, the death time and cadaver preparation technique have an impact on the experiment results. Additionally, the physiological response and muscle strength of a cadaver cannot be determined, and the characteristics of human tissue vary as people age. Even so, this method is still widely used to evaluate the validity of various mechanical and mathematical models. For instance, some researchers impacted head-on the frontal part of a human cadaver without preservative treatment with a rigid cylinder and comparatively analyzed the relationships between intracranial pressure and striking speed and energy under different impact conditions (Nahum et al., 1977). These experimental data have widely used to access the reliability of various head finite element (FE) models.

Compared with human cadaver experiments, mechanical model experiments have advantages of high repeatability and test convenience. At present, automobile safety evaluations around the world mainly rely on subsystem impactor experiment aiming at pedestrian safety and dummy experiment with respect to occupant safety, since these devices and dummies are similar to human in size, structure, mass distribution, dynamic characteristic, and impact kinematics, and the impact responses of some key parts have a certain degree of biomechanical fidelity. The subsystem impactor experiment was proposed by the European Enhanced Vehicle Safety Committee Working Group WG10 and WG17 to evaluate the protection performance of the vehicle's front structure for pedestrian. A legform impactor is used to investigate the impact between lower limb and bumper, which can measure the tibia acceleration, shearing displacement, and knee-bending angle. An upper legform impactor is used to investigate the impact between pelvis and bonnet's leading edge, which can measure the contact force and bending moment. And the adult (4.8 kg) and child (2.5 kg) head-form impactors are used to investigate the impact between head and bonnet, which can measure the head acceleration at the head gravity. Based on the EEVC's proposal, the International Standards Organization (ISO) also proposed a pedestrian subsystem impactor experiment, in which the adult and child head-form impactors were adjusted to 4.5 kg and 3.5 kg, respectively. To improve pedestrian safety from the government and consumer perspectives, the subsystem impactor experiment has been accepted as a regulation and new car assessment program (NCAP).

With the development of computer technology, numerical simulation has been widely applied to crash injury investigation. Currently, the commonest numerical models include: multi-body system (MBS) model, in which the segments are modeled as a rigid body and connected with several joints, and FE model, in which the human body is modeled based on the constitutive properties of human tissues. The advantages of the MBS model are low calculating cost and appropriate for accident reconstruction. For instance, the TNO models have been widely used in real-world pedestrian accident reconstruction. The disadvantages are: (1) each body part in the MBS model is represent by an approximate ellipsoid, which reduces the accuracy of impact responses; and (2) the MBS model can only be used to investigate the injury mechanisms from the perspective of kinematics, such as velocity, acceleration, and impact force, but cannot be applied to evaluating the injury severities from the level of tissues. The FE model can describe irregular human body geometry, complex anatomic structure, nonuniform nonlinear material property, variable boundary and loading conditions, and so on, which contribute to obtaining more accurate impact responses and intensive injury mechanisms. At present, the human FE model becomes increasingly subtle and anthropomorphic with the help of CT scanning and human slice techniques. And the neuromuscular responses in different crash scenarios are also considered. As the most advanced full-scaled model in the world nowadays, the THUMS Version 6 has not only detailed viscera models, but also muscle models that can simulate various strength levels when a driver is in different emotional states (e.g., tense and relaxed). The model can adapt the variety of occupant sitting posture with the popularity of autonomous vehicle, then achieves more accurate analysis.

Injury Mechanisms and Criterion

The head, chest, and lower extremities are the most frequently body parts injured in traffic accidents. To meet the need of engineering designs to reduce and even avoid accidental injuries, the injury mechanisms and corresponding tolerance levels of various body parts have been comprehensively investigated from a biomechanics point of view.

The common head injury patterns are skull fractures and brain injuries. Skull fractures include basilar and vault fractures, which are caused by the contact force. Brain injuries can be classified into two categories: diffuse and focal injuries. Diffuse injuries consist of concussion and diffuse axonal injury (DAI). Concussion is the most common injury stemming from a sharp shock or pressure. DAI refers to the disruption to the axons in the cerebral hemispheres and the subcortical white matter. Focal injuries consist of hematomas and contusions, which refer to localized damage to blood vessels and/or neurological tissue. Hematomas can be epidural, subdural, or intracerebral, while contusions can be coup or contrecoup. The noncontact rotation or translation motion of the head, which leads to a relative movement between skull and brain, often causes brain injuries. The head injury categories and corresponding mechanisms are summarized in Table 1 (Yang, 2005).

For skull fracture, the contact force has been widely adopted as a predictive indicator with different tolerance limits for different impact locations, such as 3.6–9.0 kN for frontal, 5.0–12.5 kN for temporoparietal, 6.4 kN for occipital, etc. Recently, some

Table 1 Typical head injuries and corresponding injury mechanisms

<i>Injury categories</i>	<i>Injury mechanisms</i>
Skull fracture	Contact force
<i>Brain neurological injuries due to tearing of neuronal axons in the brain tissues</i>	
Concussion	Rotational motion and relative motion between skull and brain
Diffuse axonal injury	Rotational motion
<i>Intracranial vascular injuries due to ruptured arteries and/or bridging veins</i>	
Coup contusion	Contact force
Contrecoup contusion	Pressure wave
Epidural hematoma	Contact force
Subdural hematoma	Translational and rotational acceleration

researchers found the skull internal energy was an excellent candidate to predict skull fracture. The 50% risk of skull fracture for different locations: 481 mJ for frontal, 456 mJ for temporoparietal, and 457 mJ for occipital, respectively (Sahoo et al., 2016b). For brain injuries, the predictive indicators related to brain tissue strain perform better compared with that related to translation or rotation of head. This brain injury criterion mainly contain Von Mises stress, first principal stress, Von Mises equivalent strain, first principal strain, and cumulative strain damage measure (CSDM) at three different strain levels, such as CSDM (0.10), CSDM (0.15) and CSDM (0.25), maximum axonal strain rate, and maximum axonal strain (Sahoo et al., 2016a). Thoracic injuries can be divided into three categories: ribcage fractures, lung injuries (pneumothorax or hemothorax), and injuries of the other thoracic organs. The thorax injuries are mainly from two mechanisms, namely the compression of the thorax and the viscous loading within the thorax cavity. The compressive force to the thorax can cause the rib fracture, sternum fracture, a hemothorax, and pneumothorax. The viscous and inertial loading can cause lung contusion and vessel disruption. Thorax injuries in crashes often occur in combination with combined mechanisms. The thoracic trauma index (TTI) and viscous criterion (VC) were introduced to access the thoracic injuries. A TTI value of 85 g has been proposed as the maximum exposure for adults and 60 g for children, while a VC_{max} value of 1 m/s is the recommended maximum exposure for adults.

Pelvis injuries often derive from a lateral impact with stiff bonnet edge or bonnet top. The lateral loading from the car front structure is applied to the great trochanter of the femur, resulting in compressive injuries. The injuries to the pelvis involve one or more of the bone structures, namely pubic rami, pubic symphysis, acetabulum (hip socket), femoral head, and proximal femur. An average lateral impact force caused skeletal injuries: 4 kN for 5th percentile females and 10 kN for 50th percentile males.

Typical lower limb injuries include fractures of the long bones (femur, tibia, and fibula), injuries of the knee joint (ligaments avulsion and condyle fractures), and ankle dislocation and foot bone fracture. The long bone fractures are mainly caused by lateral shearing and bending.

Knee joint injuries mainly contain femoral/tibia condyle fractures, patella fractures, and ligament tears and ruptures. These injuries are mainly caused by the force transferred through the knee joint as a result of shear loading and/or bending loadings, or a combination of both. When the knee joint is exposed to a bend loading, the medial collateral ligament (MCL) is first stretched; if the bending moment goes beyond the tolerance, the MCL could be partly or totally ruptured. Then other ligaments (anterior cruciate ligaments, posterior cruciate ligament, lateral collateral ligament) may be fractured.

References

- Cai, M., Zou, T.F., Luo, P., Li, J., 2014. Evaluation of simulation uncertainty in accident reconstruction via combining Response Surface Methodology and Monte Carlo Method. *Transp. Res. Part C Emerg. Technol.* 48, 241–255.
- Gennarelli, T.A., Thibault, L.E., Adama, J.H., et al., 1982. Diffuse axonal injury and traumatic coma in the primate. *Ann. Neurol.* 12 (6), 564–574.
- Haddon Jr., W., 1968. The changing approach to the epidemiology, prevention, and amelioration of trauma: the transition to approaches etiologically rather than descriptively based. *Am. J. Public Health Nations Health* 58 (8), 1431–1438.
- Moser, A., Steffan, H., Spek, A., Makkinga, W., 2003. Application of the Monte Carlo Methods for stability analysis within the accident reconstruction software PC-CRASH (No. 2003-01-0488). SAE Technical Paper. SAE, Warrendale, PA.
- Nahum, A.M., Smith, R., Ward, C.C., 1977. Intracranial pressure dynamics during head impact (No. 770922). SAE Technical Paper. SAE, Warrendale, PA.
- Peng, Y., Chen, Y., Yang, J.K., Willinger, R., 2012. A study of pedestrian and bicyclist exposure to head injury in passenger car collisions based on accident data and simulations. *Safety Sci.* 50 (9), 1749–1759.
- Sahoo, D., Deck, C., Willinger, R., 2016a. Brain injury tolerance limit based on computation of axonal strain. *Acc. Anal. Prev.* 92, 53–70.
- Sahoo, D., Deck, C., Yoganandan, N., Willinger, R., 2016b. Development of skull fracture criterion based on real-world head trauma simulations using finite element head model. *J. Mech. Behav. Biomed. Mater.* 57, 24–41.
- Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Acc. Anal. Prev.* 43 (5), 1666–1676.
- Untaroiu, C.D., Meissner, M.U., Crandall, J.R., Takahashi, Y., 2009. Crash reconstruction of pedestrian accidents using optimization techniques. *Int. J. Impact Eng.* 36, 210–219.
- Van Elslande, P., Fournier, J.Y., 2017. Failures of interaction between powered two-wheeler riders and car drivers in urban accidents. *Int. J. Transp. Dev. Integ.* 1 (2), 235–244.
- Wach, W., Unarski, J., 2007. Uncertainty of calculation results in vehicle collision analysis. *Forensic Sci. Int.* 167, 181–188.
- Yang, J.K., 2005. Review of injury biomechanics in car-pedestrian collisions. *Int. J. Vehicle Safety* 1, 100–116.

Incident Detection Systems, Airplanes

Ivan Ostroumov, Nataliia Kuzmenko, National Aviation University, av. Kosmonavta Komarova 1, Kyiv, Ukraine

© 2021 Elsevier Ltd. All rights reserved.

Flight Management System	353
Detection of Incidents Related to the Atmosphere	353
Runway Incursion Avoidance	353
Traffic Alert and Collision Avoidance System	354
Terrain Avoidance and Warning System	355
Biography	356
References	357
Further Reading	357
Relevant Websites	357
See Also	357

Safety plays a key role in aviation operations. From the safety point of view, an airplane can be a significant source of danger for passengers on board of the airplane, and people on the ground. During a normal flight, an airplane may be affected by numerous factors that may cause an incident and potentially lead to a catastrophe. In aviation, the safety of a flight is determined by the level of threat to human life, estimated by a probabilistic approach. The effect of each destabilizing factor is assessed as a risk of harm to the human body and is also compared with a certain level of acceptable risk for the safety of air transport. The most significant aviation event is a catastrophe, defined as an event that has led to the death of a person or their disappearance. Aviation events without human casualties but resulting in a significant damage or destruction of airplane structure, or serious to a person, are defined as an accident. An incident is an event that does not result in a catastrophe, accident, or serious incident, which took place during the airplane operation, resulting in a deviation from the normal operation of the airplane, crew, or Air Traffic Management services, and was under the influence of certain factors. Serious incidents include events that require a crew to take complicated or emergency actions that usually would not be taken during a normal flight.

The results of the analysis of aviation events indicate that, in the most cases, incidents appear due to the effect of factors that consistently degrade the flight state. The major number of incidents and catastrophes is the result of a combination of contributing factors associated with airplane performance, crew activities, and environmental conditions (Lee and Kim, 2015; Yu and Wang, 2019).

In the general case, the detection of near incidents is based on the detailed analysis of physical values that determine the airplane state (Fig. 1). Numerous sensors measure physical values that are presented in the digital form at the time of discretization. As sensor measurements contain errors, which are usually normally distributed, it is necessary to perform errors filtering in order to reduce the noise impact of further computations. During filtration, different stages can be used including data fusion algorithms, detection of outliers, statistical data processing, and methods for reduction of measurements errors. Results of filtration are used for calculation of the airplane parameters that describe its state.

The values of parameters are compared with the boundaries of a certain safety category, which serve as an indicator of near-miss detection of degrading changes in a state. During near-miss detection, simple algorithms for comparing values with permissible confidence band and complex algorithms, such as probabilistic one, grounded on risks estimation of dangerous situations, can be applied. A risk is the probability of a dangerous state, which has a threat to human life.

In probabilistic methods of risk analysis, a Bayes formula is usually used to estimate the risk of danger taking into account prior information obtained on the basis of previous experience of system operation (Ostroumov and Kuzmenko, 2018a). Detection of near-miss conditions is performed due to a simple comparison of the estimated risk with the value of acceptable risk, defined by regulative documents in aviation, or the results of statistical research, or expert assessments. Also, probabilistic methods are usually based on the maximum posteriori estimation, which provides near-miss detection with the highest probability of deviation from an acceptable level. Results of the near-miss analysis are displayed to the pilot through the Electronic Flight Instruments System (EFIS) on the scales of a parameter in the form of acceptable areas, which are highlighted with a color in accordance with the level of danger.

An oral and visual warning of an unsafe state for a pilot allows preventing further degradation of flight safety and requires a pilot to take a series of actions aimed to return the parameters to the acceptable limits. As an aircraft is a complex dynamic system, which is characterized by multiple numbers of different parameters, they are controlled automatically by built-in control systems to reduce the pilot workload.

Different systems of incident detection and avoidance are used for recognition and alarming of potential hazards in civil aviation.

An airplane engine is a complex system with a high danger of fire and explosion, which requires continuous control of numerous engine parameters. In particular, the Full Authority Digital Engine Control System (FADEC) controls gas pressure and temperature in different zones, estimates the thrust using measurements of different velocities, measures vibration parameters, and controls temperature and pressure of oil and fuel in different engine zones. Going out of the operation boundaries, deviation of parameters involves the automatic control algorithms that adjust parameters in order to reduce the risk of developing a

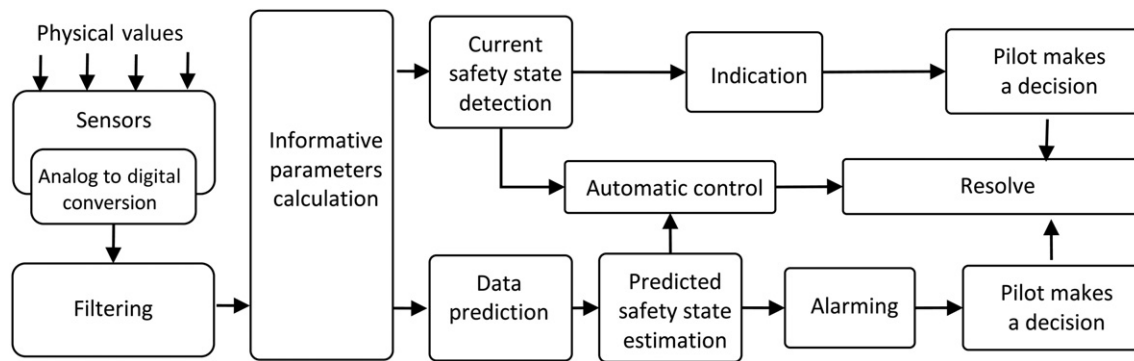


Figure 1 Near-miss detection and management.

hazardous state and minimizing degradation of engine performance. A pilot controls only the engine thrust and does not need to spend the time to monitor other engine parameters. FADEC provides measurement, near-miss analysis, and integrated control of airplane engine operation.

Fire in the engine compartment is one of the severe hazardous states, which can occur due to engine elements overheating, faults, the presence of external physical objects inside the engine, and other causes. The simple presence of a small amount of fuel or oil on a heated engine surface can lead to an instantaneous fire and engine failure. An onboard fire detection system is used for fire detection at the engine bay and adjacent compartments in order to prevent the spread of the fire to other parts of an airplane and to the fuel tanks. Numerous sensors of temperature measurements and open flame detectors are used at different engine zones in the fire protection system. The sectoral structure of the engine and sensor locations allows the system to precisely identify a zone within which the combustion is detected. In case of fire detection, a fire protection system alarms the pilot about the danger in the form of audiovisual warnings with a subsequent engine stop and interruption of fuel supply with the purpose of fire localization and prevention of fire rapidly spreading to other parts of the airplane. Depending on the fire location, a pilot initializes a fire extinguishing system that involves mixing chemical substances in containers and forms a foaming substance with its subsequent spraying in the engine burning zone.

The fuel system is another important source of the airplane safety threat. Fuel by its technical characteristics is a combustible substance, which is an additional explosive threat onboard of the airplane. Engine operation needs a large amount of fuel supply. The fuel consumption creates empty spaces in the fuel tanks, which can lead to conditions for the accumulation of hazardous gases that, at certain concentrations, may constitute a threat of explosion. Prevention of explosion is grounded on control of the pressure inside of fuel tanks and oxygen percentage to the total amount of dangerous gases with the purpose of avoiding the explosive concentration. Air separation modules or pumped inert gas can be used for reduction of explosive gas concentration.

In addition to potential explosions, fuel consumption is a threat of airplane balancing and stabilizing, which is associated with a nonuniform distribution of fuel mass in different parts of the airplane. To overcome this threat, fuel division to multiple small tanks is used. Thus, modern aircraft may include more than 20 fuel tanks distributed in a balanced way along with the airplane structure. Fuel usage in one place requires the initiation of fuel pumping in all other tanks for a uniform allocation of a fuel mass, which remains in the fuel system in order to balance the aircraft state.

Another danger for passengers is high-altitude flying in the conditions, such as rarefied air, low pressure, and low temperatures. Therefore, a passenger cabin must be completely isolated from the environment and several systems, such as air conditioning and pressurization should create the comfortable conditions for people present onboard of airplane to support a predefined temperature, air structure, humidity level, and air pressure. Cockpit and cabin design has to follow specific requirements for mechanical elements of structure and sealing of seams regarding a significant pressure difference inside and outside the airplane. A small crack or hole in the cabin body may cause a sudden airflow, which rapidly increases the load on the edge of the crack and leads to rapid destruction of the airplane. Decompression of the passenger cabin at high altitude leads to air loss and creates a serious threat to passengers and crew (Brot, 2018). Lack of sufficient oxygen supply for breath leads to the state of hypoxia that can result in fatal processes in the human body and brain. Emergency oxygen system supplies sufficient oxygen level for human surviving in case of decompression, supplying a certain mixture of oxygen through the emergency respiratory masks. The emergency oxygen system is initiated automatically after detection of a sharp reduction of pressure in the cabin. Cockpit decompression is a serious threat for pilots because the inability to wear an oxygen mask or fault of emergency oxygen system can lead to a loss of pilot's consciousness, which can be resulted in hypoxia and a loss of the ability of airplane control.

Airplane icing is another serious factor that significantly reduces the aerodynamic performance of an airplane and can cause loss of control and leads to catastrophe. An ice formation on the aerodynamic surfaces is a result of the effect of precipitation on the ground or getting chilled drops of moisture during the airplane passing through the cloud layer and the action of low temperatures. The anti-icing system contains numerous ice sensors on the main aerodynamic surfaces and the engine. In the case of ice detection, the system initiates the appropriate warning in the cockpit according to the rate of ice accumulation. Anti-ice system may utilize different approaches for reducing the ice accumulation rate, such as hot airflow through structural elements, electric heating, or

spraying liquids through certain holes on the airplane surface, which reduce the freezing point. Heating of structural elements or vibration of a certain flexible surface can be used in order to reduce the amount of already accumulated ice.

Flight Management System

Several incidents detection algorithms are included in the flight management system (FMS). FMS is an onboard computer system of airplane navigation and flight control. An internal memory of FMS contains various air-navigation databases, including navigational aids, runways, airport data, standard instrument departures, standard terminal arrivals, approaches, and routes. Each FMS has a three-dimensional trajectory of airplane motion that is continuously compared to the airplane location during the flight and is displayed to the pilot at the navigation display of EFIS. Unplanned airplane deviations from the predefined flight plan are immediately detected and displayed to the pilot with automatic calculation of parameters that are necessary to use for airplane return to a planned flight trajectory. FMS checks all of the pilot's actions of airplane navigation and guidance for possible incidents and prohibits pilots to use "wrong" commands. In addition, FMS includes an airplane mathematical model, which is used to estimate the permissible limits of a large number of flight parameters. Thus, after entering the number of passengers and weight of cargo, FMS estimates the most optimal flight level in terms of economy, evaluates the limits for airspeed, and other aircraft parameters, which prevent dangerous states of a flight. The results of FMS operation are shown on the flight and navigation displays on the corresponding scales, limiting the pilot's actions and preventing an incident appearance. Also, data processing of airplane locations measured by different sensors is another important FMS function, because the accuracy of airplane positioning is directly connected with air traffic safety. In case of Global Navigation Satellite System (GNSS) lock, FMS initiates one of the other available standby positioning systems with the criterion of maximum accuracy. For example, the Inertial Reference System or positioning by pairs of navigational aids can be used as a standby system (Kuzmenko et al., 2018; Ostroumov and Kuzmenko, 2018b).

Detection of Incidents Related to the Atmosphere

Environmental conditions also may affect the safety of airplane flight and can cause a catastrophe. Thus, bad meteorological conditions, such as wind shear, humidity, rain, or lightning, reduce the aerodynamic performance of an airplane that can lead to loss of control. Onboard weather radar and meteorological digital data from the ground are used for early identification of dangerous meteorological phenomena and informing the pilots. Aviation meteorological service supports pilots with a troposphere phenomena data, observed by ground meteorological radar, meteorological maps, a map of winds, images of the troposphere in the infrared spectrum, maps of possible icing, and turbulence areas.

Turbulence and wind shear are sources of danger associated with the factors of the troposphere state. A low-level wind shear alert system (LLWSAS) is a ground-based system used to detect wind shear and associated weather phenomena, such as microbursts, close to an airport, especially along the runway corridors. Common LLWSAS is grounded on atmosphere sensing by laser equipment (Hon and Chan, 2014). Detected dangerous areas are warned to pilots and other aerodrome services to avoid entering. Low-level wind shear is a sudden change of wind direction and velocity in different planes. The appearance of wind shear at low altitude may result in sudden airplane altitude drop. Therefore, the warning of LLWSAS helps pilots to respond appropriately. Some airplanes are equipped with a similar onboard wind shear alert systems (WSAS) that provide warnings related to the dangerous wind conditions ahead of the airplane.

Decrease in the maximum visibility range is another important factor that affects aviation safety. In particular, fog is a typical problem of airplane taxing and runway operations. Instrument landing system (ILS) supports airplane landing in case of significant visibility reduction. At the landing phase of a flight, an aircraft has to follow a glide-slope trajectory up to the touchdown point. ILS measures the deviation from the glideslope trajectory and warns pilots about the deviation side and its value. During the whole landing phase, pilots control the deviation from the glide slope to ensure a safe landing. ILS consists of ground-based radio beacons that transmit electromagnetic signals, which define glide-slope trajectory and onboard receivers of these signals for deviation measurements and indication.

An enhanced vision system (EVS) can be used in case of low visibility during airplane taxing. EVS uses infrared or night-vision sensors for visual recognition of taxiways centerlines and boundaries. Airplane location and detected data are used in EVS for warning generation related to deviations from the centerline. In addition, EVS may warn about risky maneuvers and taxi off the pavement to prevent running into obstruction side.

Runway Incursion Avoidance

Airplane runway operation and taxing are associated with the danger of a collision with other aircraft, vehicles on the ground, infrastructure elements, or simple taxi off the pavement. According to the International Civil Aviation Organization (ICAO), a runway incursion is defined as any occurrence at an aerodrome involving the incorrect presence of an aircraft, vehicle, or a person on the protected area of a surface designated for the landing and takeoff of an airplane. Incorrect airplane location can be a result of pilot or vehicle driver mistake to comply with a valid air traffic controller (ATC) clearance or their compliance with an

inappropriate ATC clearance. A system of monitoring and visualization of airplanes and vehicles onto the airport territory supports ATC service. Airplanes and vehicles are equipped with a certain type of transponder to interact with the ground multilateration system to identify localization in the airport area. Automatic Dependent Surveillance-Broadcast (ADS-B) transponders that automatically report vehicle location measured by GNSS are widely used. Vehicle location is processed and displayed on the scheme of the taxiways to support situational awareness of the ATC (Jones et al., 2009, 2010). Hazardous event identification on the ground is based on wide usage of airplane and vehicle path prediction by Kalman filter or simple sequential operations and comparing distances between objects in the closest point of approach with separation minimum taking into account type, size, and speed of moving objects. Predicted deviation out of the taxiway/runway and collisions are immediately warned to the ATC. Ground traffic data are also available to pilots in EFIS and display of vehicle drivers to increase situational awareness and to prevent hazardous operations.

Traffic Alert and Collision Avoidance System

A rapid air transport growth trend has and will continue to increase the number of operated airplanes. This leads to a constant increase in airspace and ground infrastructure load. An airplane operation within the congested airspace is connected with a severe risk of a mid-air collision. Traffic alert and collision avoidance system (TCAS) is an autonomous onboard system of mid-air collision detection and avoidance that improves flight safety (Fig. 2).

TCAS has been proposed by ICAO to be installed on board of airplane since 2000. An air traffic surveillance function of TCAS is based on the principle of secondary surveillance radar (SSR) to measure a relative location of each airspace user. TCAS antenna system transmits interrogation radio signals for an aircraft "mode S" transponder. Each of the aircraft transponders, within the TCAS service volume, generates a response signal at a certain frequency after minor delay. The response signal from airplane transponder contains a digital message that includes information about an airplane identification number and a barometric height (mean sea level). TCAS direction finder equipment measures bearing angles between an airplane centerline and each airspace user. Distances to the airspace users are measured by time of arrival method. Bearing and distance in TCAS, together with a barometric height, decoded from a digital message, allow to precisely localizing air traffic at the polar coordinate system of TCAS. There are two basic types of TCAS.

TCAS I is mandatory for installation on board of general aviation aircraft. TCAS I indicates an air space users' location on traffic display indicator (TDI) in the cockpit to increase pilots' situational awareness. Traffic Advisory on TDI includes airplane markups with a relative barometric height between the airplanes. Also, air traffic data may be visualized on the EFIS or Integrated Navigation System, depending on the equipment list of an airplane. In case of dangerous traffic, pilots cannot use TCAS surveillance data for maneuver planning, because maneuver of collision avoidance can be supported by Visual Flight Rules only in visual contact with the conflicting airplane.

According to international regulations, TCAS II has to be used on board of a large airplane. An air traffic surveillance of TCAS II is identical to TCAS I. However, TCAS II analyzes air traffic for early detection of a possible mid-air collision and automatically generates warnings to the pilots for conflict resolution. TCAS II algorithm analyses airplanes relative motion and, depending on the time of flight to the closest point of approach, generates recommendations for pilots of airplanes involved in a conflict. Conflict resolution between airplanes in TCAS II is performed by coordinated maneuver in a vertical plane. Resolution Advisory (RA) includes audio warnings associated with allowed vertical velocities range. RA increases the time of the pilot's reaction and is obligatory to follow. TCAS II complies with international initiatives of airplane separation minimum supporting in vertical, lateral, and longitudinal directions within a defined airspace structure during conflict resolution. Conflict avoidance is supported

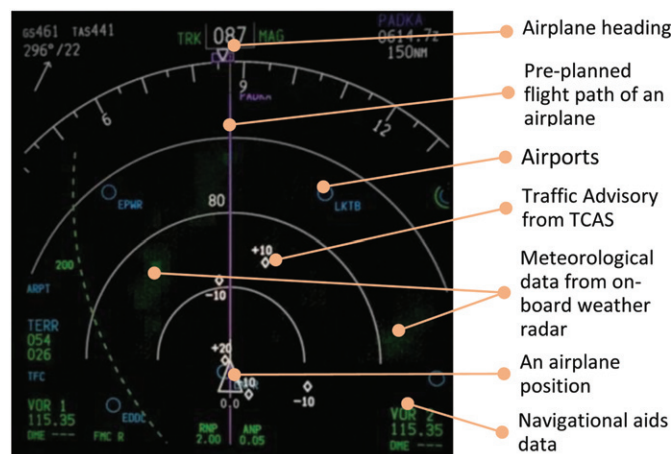


Figure 2 TCAS data on navigation display of B737.

by coordinated data processing between both airplanes with the help of digital data link communication, based on “mode S” technology.

Since 2016, each airspace user has to be equipped with the modified aircraft transponder of “mode 1090 ES” (extended squitter), which supports a periodical transmission of digital messages. A digital message includes an airplane identification code and its position. A modified transponder is one of the ways of ADS-B implementation in aviation. According to ADS-B, each user of airspace should measure its location with a certain level of accuracy and automatically share it in a broadcast mode with other users and air navigation service providers. ADS-B implementation stimulated a wide usage of digital data receivers on board of airplanes. As a result, air traffic can be displayed on EFIS in order to increase the situational awareness of pilots. In the near future, TCAS will use ADS-B data to reduce the workload on the aircraft transponders (De and Sahu, 2018).

Some general aviation airplanes may use a passive surveillance system in contrast to active surveillance in TCAS I. Portable collision avoidance system (PCAS) utilizes a passive surveillance approach. PCAS is based on the receiving both signals from ground SSR and aircraft replies to measure distances to airspace users around. PCAS antenna system uses a combination of signal amplitude and phase cancellation to detect direction data to other traffic. A relative height between the airplanes is obtained by comparison of their own barometric height with the height obtained from aircraft transponder messages. PCAS is implemented as a single block with a built-in display of air traffic, which provides a pilot with information about air traffic around them using visual and audio warnings about violation of separation minimums. Nowadays, ADS-B has gradually excluded PCAS from the market; however, a large number of general aviation airplanes is still equipped with PCAS.

Terrain Avoidance and Warning System

Airplane operation at low altitude always presents a severe risk to flight safety. Early detection and avoidance of conflict situations with the relief and artificial constructions is provided by terrain avoidance and warning system (TAWS) on board of the airplane (Breen, 1999).

TAWS implements two fundamentally different approaches for early detection of dangerous situations. At the first approach, the near-miss analysis compares flight parameters with their limits. Deviation of flight parameters out of required limits can lead to the conflict with a terrain. Boundary values of several parameters, considered simultaneously, define an area of permissible values. Each of the dangerous states corresponds to a certain area. Parameter deviations outside their limits initiate warnings to the pilots. In particular, the following pairs of parameters are analyzed: height above ground level (AGL) and vertical speed, AGL and high rate of descent, AGL and deviation from a glideslope (Fig. 3). All of these parameters identify airplane location close to the terrain and are used to generate oral and visual warnings for a pilot.

Another approach to determine potentially dangerous situations is done by the forward looking terrain alerts (FLTA). In this approach, TAWS uses airplane position and velocity to predict further flight trajectory to compare it with a global digital elevation model. Some TAWS can be equipped with a database of artificial elements to generate obstacle warnings. As GNSS is used as a common positioning system, its accuracy and resolution of the digital elevation model determine the performance of near-miss analysis at FLTA. The indication of the TAWS system represents the relative heights and dangerous elements of the relief to the pilot, according to the level of danger (Fig. 4).

There are different types of TAWS according to aircraft type. Also, there is a helicopter version of TAWS due to performance features of a controlled flight into terrain (Anderson et al., 2011).

In the general case, the majority of onboard systems are critical and the fault of a particular one can be considered as an incident that could lead to a catastrophe. To enhance system reliability, each of the important systems is duplicated. In particular, the digital data buses that ensure the data exchange between the airplane systems can be duplicated up to 5 times to reduce the number of failures. Near-miss analysis on board of an airplane is carried out at different levels to ensure the necessary safety conditions of flight. In particular, each of avionics systems has its own built-in health monitoring function and initiates the performance self-test at the start of an operation. A system provides a discrete signal transmission about the system fault to inform all other equipment in case of failure detection. In addition, each measured value is analyzed for outliers or complete inadequacy of the observed physical processes. In case of significant outliers in the sensor measurements, the built-in control system generates a discrete signal of failure, which is automatically reported to other onboard equipment for excluding faulty measurements from the computations. In order to fulfill gaps in measured data, different data recovery algorithms are used on board of a modern airplane.

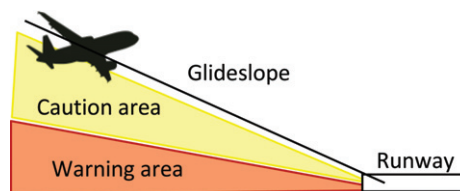


Figure 3 TAWS.

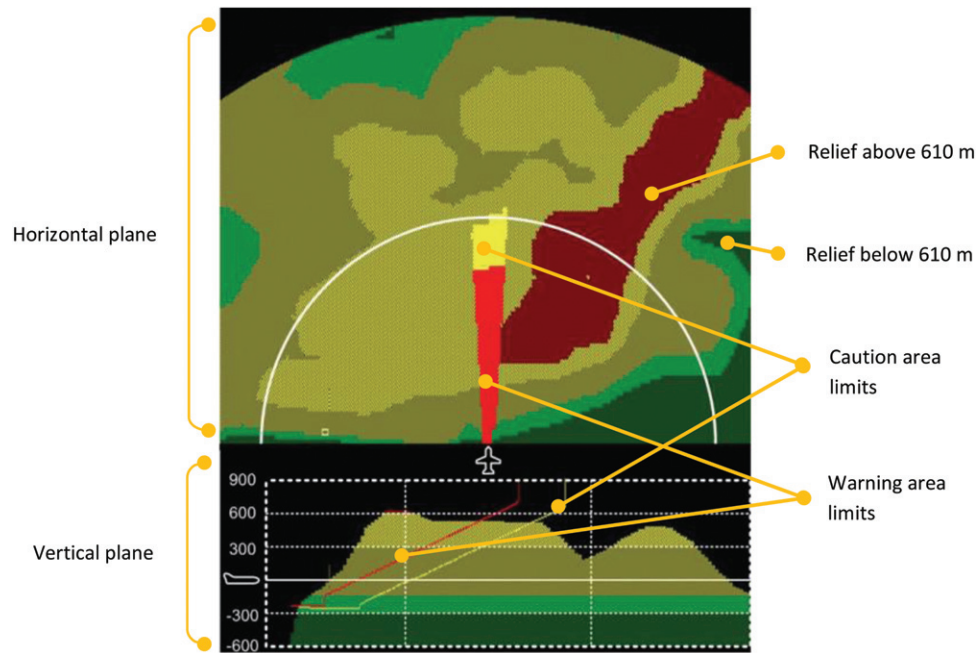


Figure 4 Visual warning about dangerous terrain in front of an airplane from TAWS on board of An148.

Flight data recording system provides registration of the flight data and the measurements of onboard sensors. The set of data is available for analysis during airplane maintenance. Software for data processing supports a near-miss analysis by recorded data, which includes equipment faults detection, risks estimation, and detection of avionics that may cause a threat due to a high risk of failure. The wide use of computerized maintenance system of onboard equipment monitoring significantly increases the safety of air transportation.

An airplane operation is controlled from the ground by the ATC unit. The ATC services support air traffic with the aim of the planning of the nonconflict airplanes flow within the controlled airspace. Similar to onboard TCAS, the algorithms of early collision detection are used in ATC unit. The functionality of the ATC unit also includes detection of airplanes deviations from the pre-planned trajectory and separation minimum. Data from various sensors about air traffic are compared with scheduled flight plan databases in ground incident detection systems.

In the framework of the controlled air traffic, an ATC must provide a conflict-free airplane flow. Thus, the TCAS onboard system plays a supporting role. For example, the appearance of resolution advisory in TCAS II may indicate an incorrect ATC clearance or pilot mistake.

Biography



Ivan Ostroumov has been a faculty of Air Navigation Systems Department of the National Aviation University of Ukraine since September 2007. He obtained his PhD degree of Engineering in Navigation and Traffic Control in 2009 from National Aviation University of Ukraine. Since then, he has been a research scientist and associated professor for National Aviation University. Since 2016, he has also served as navigation instructor at "Aviation Company Ukrainian Helicopters." In 2017/2018, he was a Fulbright scholar in the School of Aeronautics and Astronautics at Purdue. Also, he took part in several projects, including Supporting SESAR on GNSS Vulnerability Assessment by performing Space Weather Analysis (Navigation department, EUROCONTROL, Brussel) and e-learning course development (Institute of Air Navigation Services, EUROCONTROL, Luxembourg). His research theme is advanced methods for Alternative Positioning, Navigation, and Timing. His current research projects include Methods and Algorithms of positioning by multiple navigational aids, Availability and Accuracy estimation of navigation.



Nataliia Kuzmenko is a senior researcher of the National Aviation University of Ukraine. She obtained her PhD degree of Engineering in Navigation and Traffic Control in 2017 from the National Aviation University of Ukraine. She is a certified aviation security instructor by ICAO (ASTP/Basic, ASTP/Instructors). She has had a traineeship in Tool Development for support to CAA/NSA at Regulatory Division within the Directorate Single Sky (Eurocontrol, Brussels). Her current research projects include aviation safety, collision detection and avoidance, artificial intelligence, Remotely Piloted Aerial Systems, video stream object detection and recognition, kernel density estimation, and neural networks.

References

- Anderson, T., Jones, W., Beamon, K., 2011. Design and implementation of TAWS for rotary wing aircraft. In: 2011 Aerospace Conference. IEEE, Big Sky, MT, pp. 1–7.
- Breen, B.C., 1999. Controlled flight into terrain and the enhanced ground proximity warning system. *IEEE Aero. Elec. Syst. Mag.* 14 (1), 19–24.
- Brot, A., 2018. The dual threat of sudden decompression. 58th Israel Annual Conference on Aerospace Sciences, IACAS 2018, March 2018, pp. 613–629.
- De, D., Sahu, P.K., 2018. A survey on current and next generation aircraft collision avoidance system. *Int. J. Syst. Control Commun.* 9 (4), 306–337.
- Hon, K.K., Chan, P.W., 2014. Application of LIDAR-derived eddy dissipation rate profiles in low-level wind shear and turbulence alerts at Hong Kong International Airport. *Meteorol. Appl.* 21 (1), 74–85.
- Jones, D.R., Prinzel, L.J., Otero, S.D., Barker, G.D., 2009. Collision avoidance for airport traffic concept evaluation. In: 2009 IEEE/AIAA 28th Digital Avionics Systems Conference. IEEE, Piscataway, NJ, pp. 4–C.
- Jones, D.R., Prinzel, L.J., Shelton, K.J., Bailey, R.E., Otero, S.D., Barker, G.D., 2010. Collision avoidance for airport traffic simulation evaluation. In: 29th Digital Avionics Systems Conference. IEEE, Salt Lake City, UT, pp. 3–B.
- Kuzmenko, N.S., Ostroumov, I.V., Marais, K., 2018. An accuracy and availability estimation of aircraft positioning by navigational aids. *Methods and Systems of Navigation and Motion Control: MSNMC 2018 5th International Conference of IEEE*. IEEE, Kiev, Ukraine, pp. 36–40.
- Lee, W.K., Kim, S.J., 2015. Roles of safety management system (SMS) in aircraft development. *Int. J. Aeronaut. Space Sci.* 16 (3), 451–462.
- Ostroumov, I.V., Kuzmenko, N.S., 2018a. An area navigation (RNAV) system performance monitoring and alerting. *System Analysis & Intelligent Computing: SAIC 2018 First International Conference of IEEE*. IEEE, Kiev, Ukraine, pp. 211–214.
- Ostroumov, I.V., Kuzmenko, N.S., 2018b. Accuracy assessment of aircraft positioning by multiple radio navigational aids. *Telecommun. Radio Eng.* 77 (8), 705–715.
- Yu, S., Wang, H., 2019. Risk forecasting in general aviation based on sparse de-noising auto-encoder neural network. *Syst. Eng. Electronics* 41 (1), 112–117.

Further Reading

- ICAO, 2012. Manual of Aircraft Accident and Incident Investigation, Doc 9756-AN/965, ICAO.
- ICAO, 2018. Safety Management Manual, Doc 9859, ICAO.
- Moir, I., Seabridge, A., 2007. *Aircraft Systems Mechanical, Electrical, and Avionics Subsystems Integration*, third ed. John Wiley & Sons Ltd, Hoboken, NJ.
- Collinson, R., 2011. *Introduction to Avionics Systems*, third ed. Springer, New York.
- Spitzer, C., Ferrell, U., Ferrell, T., 2014. *Digital Avionics Handbook*, third ed. CRC Press, Boca Raton, FL/London.

Relevant Websites

- SKYbrary: <https://www.skybrary.aero>
- Aviation Safety Network: <https://aviation-safety.net>
- Eurocontrol Airborne Collision Avoidance System home page: <https://www.eurocontrol.int/acas>
- Federal Aviation Administration Aircraft Safety home page: <https://www.faa.gov/aircraft/safety>

See Also

The concept of “acceptable risk” applied to road safety risk level; Bicycles, E-bikes and micromobility, a traffic safety overview; Aviation Safety - Commercial Airlines; AIRCRAFT MAINTENANCE AND INSPECTION; Collision Avoidance Systems, airplanes

Inequality and Traffic Safety

Miles Tight, School of Engineering (Civil), University of Birmingham, Birmingham, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

International Differences	358
Within Country Differences—Differences Between Groups	358
The Future—Impact of Increasing Autonomy	358
References	360
Further Reading	360

International Differences

Most road traffic deaths and injuries occur in the low- and middle-income countries which are experiencing rapid levels of motorization. The death rates in low-income countries are roughly 3 times higher compared to high-income countries (the average rate in low-income countries is 27.5 deaths per 100,000 population compared to 8.3 per 100,000 population in high-income countries). Low-income countries account for around 1% of motor vehicles but 13% of traffic deaths ([World Health Organisation, 2018, 2018](#)) ([Fig. 1](#)).

The rates of road traffic deaths are highest in Africa and Southeast Asia (26.6 and 20.7 deaths per 100,000 people, respectively). In the lowest income countries in Africa, the rate is 29.3 ([Fig. 2](#)).

Within the overall global figures, vulnerable road users are particularly overrepresented in the fatality figures compared to their use of the roads—54% involve pedestrians, cyclists, and motorcycle users. [Fig. 3](#) shows the variability between different WHO regions, showing some regions, especially Africa, having very high levels of pedestrian deaths compared to elsewhere. Also of note is the high levels of motorcycle deaths in Southeast Asia.

A further aspect that varies considerably between countries is the availability and quality of postcrash care—the proportion dying before reaching hospital is over twice that in low-income countries compared to high-income countries.

Ameliorating road safety requires new implementation of context-specific solutions. Road safety literature provides strong evidence that the distribution of road traffic fatalities varies dramatically across different parts of the world. This strongly suggests that context-specific and effective prevention strategies that protect the particular at-risk road user groups should be carefully investigated ([Naci et al., 2009](#)) and that what works or is normal in one place may not work elsewhere.

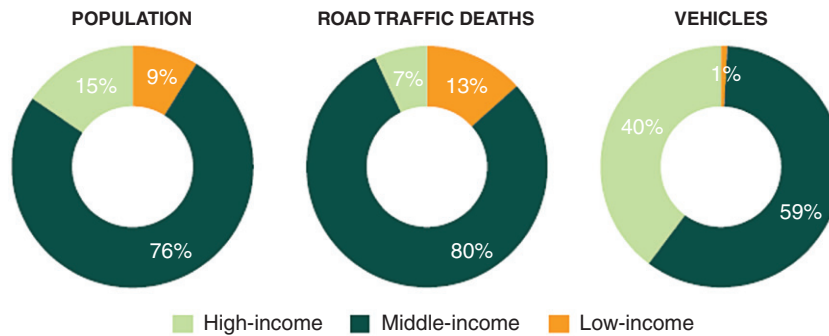
Within Country Differences—Differences Between Groups

People living in disadvantaged areas tend to live in areas where they are more exposed to traffic risk and are at greater risk of injury and death. Use of motorized transport such as cars provides a greater level of protection from injury than users less protected such as pedestrians, cyclists, and motorcyclists. People living in more deprived neighborhoods have been shown to be more likely to be killed or injured as road users (see [Ward et al., 2007](#), [Christie et al., 2007](#), or [Muir, 2013](#)). The variation in risk between different socioeconomic groups is particularly pronounced for certain road user groups such as pedestrians and especially so for children and young people ([Lucas et al., 2019](#)).

More deprived areas of cities tend to be more hazardous in terms of traffic with more through traffic, often taking shortcuts through residential roads and along roads not designed to cope with traffic volumes experienced, greater issues from things such as on-street parking, and often more complicated crossing environments. Such areas often contain little in the way of green space and children's outdoor play is more likely to occur in the street environment ([Morency et al., 2012](#)). There is some evidence that targeted improvement schemes, such as traffic calming, applied in such areas can help to mitigate socioeconomic differentials in injury rates ([Steinbach et al., 2011](#)).

The Future—Impact of Increasing Autonomy

Various future technological developments are being considered which will have an impact on road safety. Potentially the most impactful of these is the proposed developments in autonomy and independent control of vehicles which ultimately could lead to completely self-driving vehicles. The last decade or so has already seen increasing levels of technological sophistication in this area in motor vehicles, such as parking aids and cruise control systems keeping safe distance to the vehicle in front, which are now fairly standard in new vehicles, as well as improvements to safety aids and other driver support systems in vehicles. A future where autonomous vehicles are the norm could substantially change safety, removing much or all of the control from humans and making



*Income levels are based on 2017 World Bank classifications.

Figure 1 Proportion of population, road traffic deaths, and vehicles by income level 2016. Source: *World Health Organisation, 2018*.

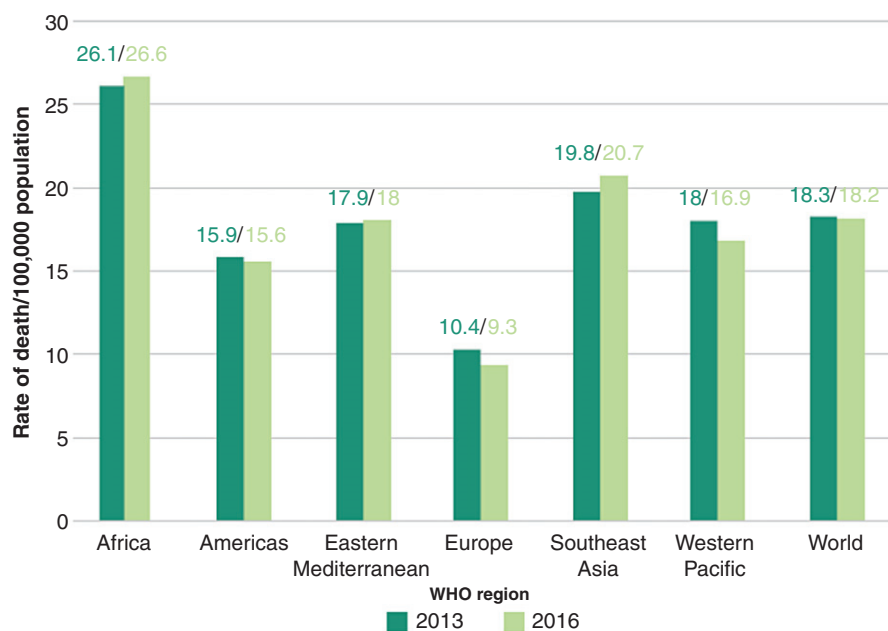


Figure 2 Rates of road traffic death per 100,000 population by WHO regions 2013, 2016. Source: *World Health Organisation, 2018*.

movement in cars and other motor vehicles considerably safer than is currently the case. However, it is not yet clear whether such safety gains will be equitably achieved amongst different road user groups. For example, how will non-motor vehicle users such as pedestrians and cyclists be affected by such a future? Arguably, they may also see safety improvements as motor vehicles will likely be more efficient at stopping or avoiding potential crash situations, but will the safety gains be as high for the nonmotorized road users as for their motorized counterparts? In part this might depend on the detail of the vehicle programming, in particular the oft-considered situations where a vehicle has a difficult choice to make between two negative outcomes in a crash situation—for example, whether to hit another car or to hit a pedestrian in the road. It is possible that inequality might even be consciously or unconsciously built into the vehicle control systems themselves.

Such new technologies are also likely to take a considerable period of time before they trickle down to all vehicles in all countries, perhaps appearing last in the lowest income countries of the world where only a very small percentage of the people will afford them. This will have the effect of increasing existing inequalities in safety between the highest and lowest income countries.

In summary, there currently exist major differences in safety between different groups of society—the differences are principally between those who have the means to afford to use safer modes of travel, such as cars, where the occupants are well protected in the event of a crash and those who are more vulnerable to injury such as walkers and cyclists. There are also differences according to the level of wealth of countries and regions and their ability to provide safe environments in which travel can take place, appropriate training and education programs to enhance user abilities and perceptions of safety, and safety standards and quality of vehicles available. However, even among countries with similar per capita GDP, some provide better safety than others do. For example,

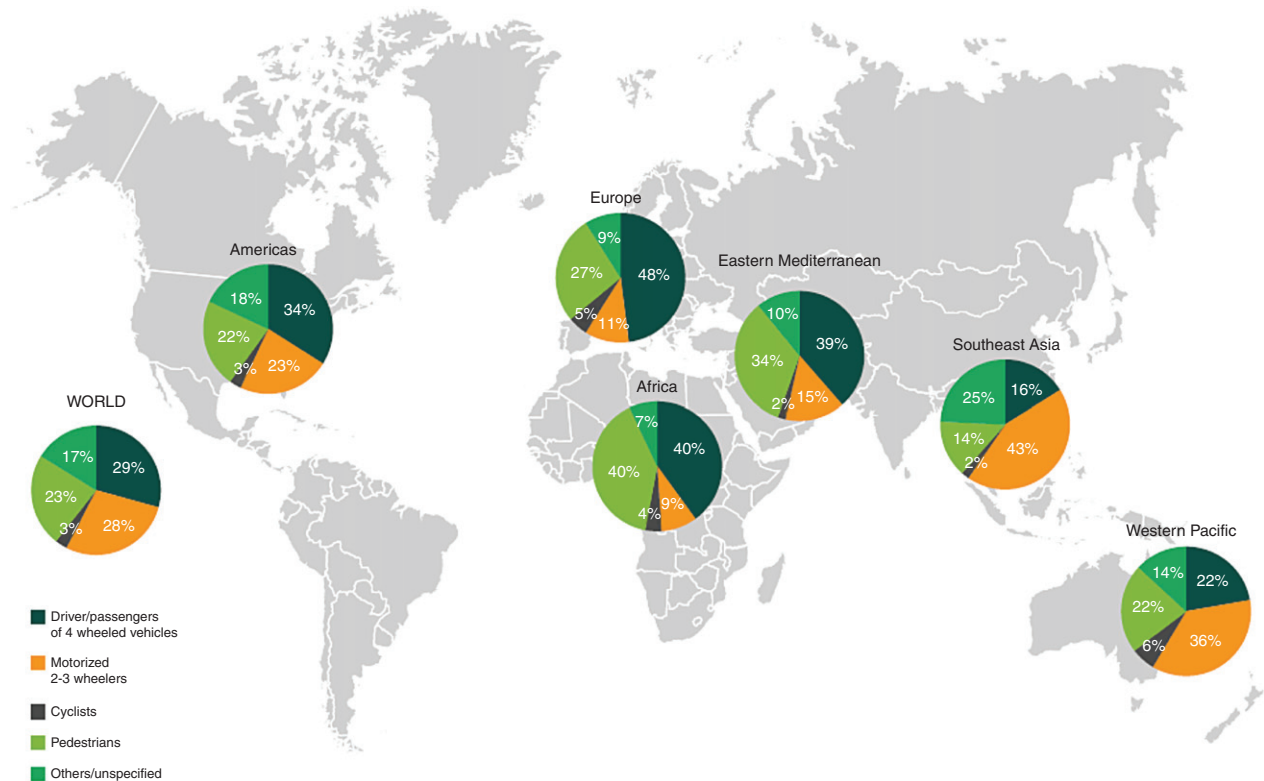


Figure 3 Distribution of deaths by road user type by WHO region. Source: *World Health Organisation, 2018*.

Sweden in 2017 had a fatality rate of 2.54 per 100,000 people—and the United Kingdom had a similar rate—whereas the United States had a rate more than 4 times higher at 11.40 per 100,000 people even though a lower percentage of people walk and ride bicycles in the United States. A part of the difference is explained by people in Europe making more trips by bus and train, which are even safer modes than the car, so people travel fewer miles by car. Another difference is the distribution of income and education level. Parts of the United States have fatality rates similar to European countries and other parts similar to low-income countries. The state with the highest education level, Massachusetts, had a per capita fatality rate of 5.1 in 2017, whereas Mississippi had a rate of 23.1 per 100,000 people. And, 42% of people in Massachusetts had a Bachelor's degree or higher in 2017, whereas only 21% did in Mississippi. In 2017, Massachusetts had a per capita personal income of almost exactly double that of Mississippi. In addition, even within these states, local traffic safety seems to vary with income and education level.

References

- Christie, N., Ward, H., Kimberlee, R., Towner, E., Sleney, J., 2007. Understanding high traffic injury risks for children in low socioeconomic areas: a qualitative study of parents' views. *Inj. Prev.* 13 (6), 394–397.
- Lucas, K., Stokes, G., Bastiaanssen, J., Burkinshaw, J., 2019. Inequalities in mobility and access in the UK transport system, 2017 to 2040. *Future of Mobility: Evidence Review*, Foresight, Government Office for Science, London.
- Morency, P., Gauvin, L., Plante, C., Fournier, M., Morency, C., 2012. Neighbourhood social inequalities in road traffic injuries: the influence of traffic volume and road design. *Am. J. Public Health* 102 (6), 1112–1119.
- Muir, H., 2013. The influence of area and person deprivation on adult pedestrian casualties. PhD thesis, University of Leeds, Leeds.
- Naci, H., Chisholm, T., Baker, T.D., 2009. Distribution of road traffic deaths by road user group: a global comparison. *Inj. Prev.* 15, 55–59, doi:10.1136/ip.2008.018721.
- Steinbach, R., Grundy, C., Edwards, P., Wilkinson, P., Green, J., 2011. The impact of 20 mph traffic speed zones on inequalities in road casualties in London. *J. Epidemiol. Community Health* 65, 921–926, doi:10.1136/jech.2010.112193.
- Ward, H., Lyons, R., Christie, N., Thoreau, R., Macey, S., 2007. *Fatal Injuries to Car Occupants: Analysis of Health and Population Data*. Department for Transport, London.
- World Health Organisation, 2018. *Global status report on road safety*. WHO, Geneva.

Further Reading

- Mullen, C., Tight, M.R., Jopson, A., Whiteing, A., 2014. Knowing their place on the roads: what would equality mean for walking and cycling? *Transp. Res. A* 61, 238–248, doi:10.1016/j.tra.2014.01.009.

Lighting

John D. Bullough, Lighting Research Center, Rensselaer Polytechnic Institute, Troy, NY, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	361
Visibility	361
Visual Performance	361
Glare	362
Lighting Technologies	362
Light-emitting Diodes	362
Vehicle Lighting	363
Forward Lighting	363
Roadway Lighting	364
Lighting and Safety	364
Roadway Lighting Practices	365
References	366

Introduction

Lighting is a critical element of the roadway visibility system. After an introduction to driver and pedestrian visual performance, this chapter discusses the design and implementation of roadway lighting systems and requirements for vehicle lighting systems, including their interactions with other visibility elements along the roadway. The chapter concludes with a brief discussion of application issues that can impact lighting practices.

Visibility

Visual Performance

The primary purpose of lighting in the roadway environment is to aid vision. An understanding of visual tasks and the factors that affect visual performance can aid in providing the appropriate level of lighting for those tasks. Visual performance, the speed, and accuracy with which a visual task can be performed is affected by several factors (Rea and Ouellette, 1991), including contrast, size, and duration.

Contrast (Fig. 1) refers to the luminance difference between the critical detail of a task and its background, for example, the clothing worn by a pedestrian against the paved surface of the roadway behind the pedestrian. A reading task such as white letters on a black sign background would have a high contrast, whereas a dark-colored animal viewed against a dark background of foliage could have a low contrast, approaching zero. As the contrast of a task increases, visual performance also increases. Once contrast is sufficiently high, further increases in contrast will have little effect on visual performance.

The size of the visual task is also important. A task with an infinitesimally small size will be invisible, with visibility increasing as size increases. As with contrast, once the size is sufficiently large, making it larger will not increase visual performance very much. For example, visual performance while reading 6-point type might be significantly better than visual performance reading 4-point type, but visual performance while reading 18-point type would only be marginally better than visual performance reading 12-point type.

The duration a visual task is presented also affects visual performance. For visual tasks, which are visible for only very brief periods of time (less than 0.1 s), the intensity and duration are traded off, such that a signal which appears for half the duration as a second signal would need to have twice the intensity to be as visible as the second signal. For presentation durations longer than 0.1 s, the duration of the visual task is much less important.

For a visual task with a specific contrast, size, and duration, lighting directly affects another factor, which helps determine the overall visual performance. The background luminance is affected by the illuminance falling on the surface surrounding the object and by the reflectance of that surface. Holding contrast, size, and duration constant, the higher the background luminance is, the better visual performance will be. Similarly, visual acuity (ability to distinguish detail of the smallest objects that can be seen) improves with increasing light levels. In addition, both speed and accuracy of visual processing improves with higher light levels. Dark surfaces, such as asphalt, will have a lower luminance than lighter surfaces, such as concrete, even for the same illuminance falling on them, because the reflectance of asphalt is lower than that of concrete.

Although the performance of a visual task improves with higher light levels, the rate of improvement decreases when the light level is high. This “plateau” effect is inherent to most visual tasks encountered at work or at home when objects in the field of

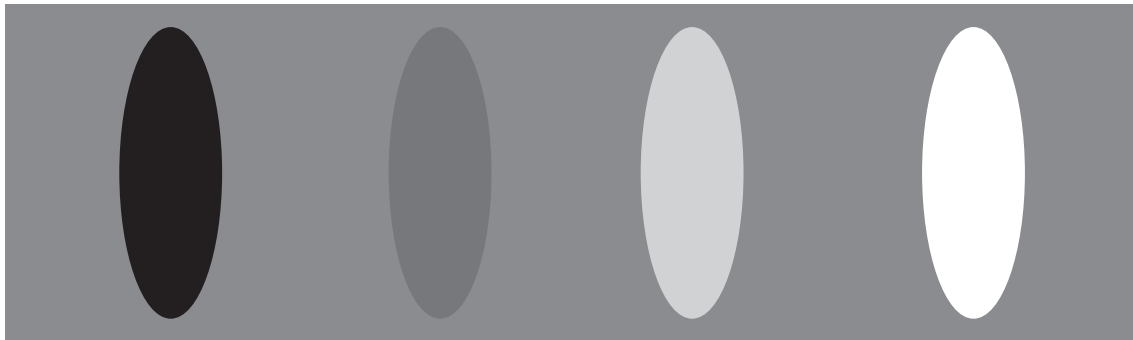


Figure 1 Illustration of contrast changes; contrast is lowest for the objects near the center.

view are well above the visual threshold. For very small, low contrast visual tasks, however, the light level will have a relatively larger impact on visual performance. The age of the occupant should also be considered when planning a lighting system. The amount of light reaching the retina of a 50-year-old person is approximately 50% of that for a 20-year-old (Rea, 2000) because the lens of the eye yellows and thickens during aging, and because the pupil of the eye tends to decrease in size as a function of age.

Glare

Although lighting is usually considered to be beneficial for visibility, excessive light can reduce visual performance by creating disability glare, which is the reduction in contrast caused by a bright light in the field of view that results in scattered light inside the eye. This scattered light creates a veiling luminance analogous to looking through a veil or haze (Fry, 1954).

Bright light sources can also cause discomfort glare, which is the annoying or even painful sensation created by the offending source. Mechanisms for discomfort glare are not well understood but differ from those for disability glare. Two lights that reduce visibility equally may be perceived as unequally uncomfortable; lights with substantial short-wavelength (“blue”) spectral content tend to be rated uncomfortable compared to “yellowish” lights (Bullough, 2009). Nonetheless, disability and discomfort glare often occur simultaneously in the nighttime lighted environment.

Lighting Technologies

Several types of light sources, or lamps, are commonly used in transportation lighting systems, consult Rea (2000) and the National Lighting Product Information Program (www.lrc.rpi.edu/programs/nlpip) for detailed technical information on many different electric light sources. Important light source performance characteristics include luminous efficacy (lm/W), rated life, color rendering characteristics, and lumen maintenance (defined as the percentage of light output that can be reasonably expected from a lamp after it has operated for approximately half of its rated life).

Light-emitting diode (LED) sources are becoming the primary light source used in transportation lighting applications. They are beginning to rapidly replace halogen, filament-based sources in vehicle lighting, and high-intensity discharge lamps in roadway illumination systems.

Light-emitting Diodes

First invented in the 1960s, LEDs have increased in efficacy, in light output, and in the range of available colors. The development of blue LEDs that emit short-wavelength light has permitting the subsequent creation of white-light LED systems in two ways: by combining red, green, and blue LEDs to create white light, and by using phosphors together with blue LEDs that convert some of the blue light into yellow light, resulting in a mixture that is perceived as white light. Newer packages for LEDs have been developed in addition to the familiar 5-mm diameter epoxy capsule LED packages used for indicator lights, which produce greater light output and wider distributions than LEDs had previously been able to exhibit. Many LED luminaires use arrays of LED packages to create the necessary distribution (Fig. 2). Some LED luminaire packages contain metal fins or slugs as heat sinks to help dissipate internal temperatures. Increased temperature inside the LED reduces its light output and over the long term, can shorten the useful life of the source. Color rendering depends upon the combination of phosphors used to generate white light, and luminous efficacies presently range from 30 to more than 60 lumens (lm) per watt (W).

While LEDs have very long operating lives, typically more than 50,000 h without failing, they do exhibit gradual reductions in light output over time. White LED indicator type packages with no heat sinking have been shown to reach half their initial light output within 10,000 h. Improved packages have exhibited much improved lumen maintenance characteristics.



Figure 2 Photograph of LED luminaires.

Vehicle Lighting

Forward Lighting

The majority of roads are not lighted; therefore, vehicle forward lighting systems are necessary for safe nighttime travel. There are two primary types of beam pattern, or distributions of luminous intensity, permissible for vehicle headlamps, known as the high beam (or driving beam) and the low beam (or passing beam).

Fig. 3 illustrates a representative beam pattern for a low beam headlamp, projected onto a vertical wall. In order to control glare, many low beam patterns exhibit relatively sharp vertical cutoff angles, above which there is relatively little light. The sharp cutoff originated with headlamps based on United Nations regulations for vehicle lighting, and it permits visual aiming of headlamps by adjusting the vertical tilt. Most headlamps in North America have followed suit since the early 2000s, using sharp vertical cutoffs in the beam pattern ([Schoettle et al., 2008](#)). For example, for the pattern in **Fig. 3**, the location of the right-side cutoff is at 0° , or in other words, at the same height as the headlamp. The cutoff on the left side, corresponding to likely locations of oncoming traffic, is often lower than the right-side cutoff to reduce glare.

The sharp cutoff of low beam headlamps can restrict the forward visibility of drivers, even on flat, straight roads, driving speeds in excess of 55–65 km/h will make it impossible for a driver to detect and stop in time to see some potential hazards ([Andre and Owens, 2001](#)). The use of high beams is warranted for these situations (except when approaching traffic is within about 150 m), but drivers almost universally underuse their high-beam headlamps ([Sullivan et al., 2004](#)).

The sharp vertical cutoff of low beam headlamps makes vertical aim critical in the performance of these lighting systems. Recent measurements of vehicles ([Skinner et al., 2010](#)) have found that the majority of vehicles measured had at least one misaimed headlamp. Upward misaim is probably the largest contributor to disability and discomfort glare from headlamps, and downward misaim can severely restrict forward visibility, because of the sharp vertical cutoff. In the United States, most states do not include



Figure 3 Photograph of the beam pattern from a typical low beam headlamp projected onto a wall.

precise checks of headlamp aim for vehicle safety inspections; a few states and many other countries do include headlamp aim in the inspection process.

High-beam headlamps are permitted by United Nations and North American regulations not only to have higher luminous intensities than low beams, but to produce light above the typical vertical cutoff location of low beam headlamps. This, of course, increases the range of visibility for the driver, and also increases the likelihood of creating disturbing glare to oncoming and preceding drivers, and may explain their underuse.

Forward lighting systems that change dynamically in response to specific driving conditions are now beginning to show promise for changing the fixed, high/low beam pattern paradigm for headlamps. One recent and dramatic development in vehicle forward lighting is the introduction of adaptive driving beam (ADB) headlamps. ADB systems are essentially matrix configurations usually using LED sources where each source produces a specific angular portion of the headlamp beam pattern. Using cameras and other sensors, the ADB systems can reliably detect and identify headlamps from oncoming vehicles and tail lamps of preceding vehicles, and will dim the intensity from the ADB system only in those directions. This has the effect of reducing glare for other drivers while maximizing visibility for the driver with the ADB system. ADB systems are legal in most countries, which use the United Nations vehicle lighting regulations. Studies have demonstrated these benefits empirically ([Bullough et al., 2016](#)) and the US Department of Transportation is considering how to permit these systems on American roads, as currently (in October, 2019), only fixed low-beam- and high-beam headlamps are permitted.

Roadway Lighting

In this section, the safety impacts of roadway lighting, the equipment used for roadway lighting, and some of the design criteria for roadway lighting are described.

Lighting and Safety

Almost all transportation agencies recognize that roadway lighting reduces nighttime crashes, and most empirical evidence is consistent with this notion. Measuring the safety impacts of roadway lighting is difficult because crashes are relatively rare, random events that require long periods of time to measure statistically reliable effects. Many factors may change along a roadway (daily traffic, road geometry, number of lanes, to name a few), so before/after studies of roadway lighting can be confounded by different periods of time. Studies comparing different locations with and without roadway lighting require very careful selection of sites to avoid confounding (a location near a shopping center would be very different with respect to crashes than an isolated rural location).

A nighttime crash reduction factor of about 30% was found in a review of more than 60 studies of roadway lighting ([Commission Internationale de l'Éclairage, 1992](#)). Although this single nighttime crash reduction factor of 30% is often used by transportation agencies to predict the benefits of roadway lighting, it is only reasonable to suppose that roadway lighting does not have the same benefit in all situations. Indeed, careful review of the literature has shown that in locations with complex roadway geometries, high potential for conflicts, or large pedestrian populations, the nighttime crash reduction factor associated with lighting appears to be larger than along straight, access-controlled highways with little to no pedestrians. Pedestrians appear to be particularly vulnerable during nighttime conditions, and reduced visibility is a likely contributor to this vulnerability.

Despite controlling for many operational and geometric factors that is inherent in the use of daytime crashes to serve as a type of control condition, observational studies using with/without or before/after methodologies can still contain biases. In part this is because roadway lighting is not assigned randomly to various roadway locations. Instead, if lighting is installed along an existing roadway, this might have been done in response to a higher than average observed number of crashes within a relatively short period of time of a few years. If the increase was spurious, it is possible that the number of crashes could undergo regression to the mean, and exhibit a relatively lower number of crashes simply because a roadway cannot experience an above-average number of crashes every year. Another explanation of bias might be that roadway lighting is often only one of many safety-related measures that might be installed along a roadway, also including improved lane markings, medians, turn lanes, or traffic signals, and these could be partially responsible, along with lighting, for any safety improvement.

A recent study to attempt to reduce such potential sources of bias ([Bullough et al., 2013](#)) used multiple nonlinear regression modeling that included not only the presence of lighting but the type of location (i.e., urban or rural), posted speed limit, percentage of heavy trucks, average daily traffic, number and width of lanes, presence of signalization and other geometric, and demographic variables in order to isolate the safety effect (if any) associated only with roadway lighting. The results were consistent with other studies ([Commission Internationale de l'Éclairage, 1992](#)) in that there was generally a reduction in nighttime crashes associated with roadway lighting, but the magnitude of the effects were smaller, as might be expected. The nighttime crash reduction factor associated with roadway lighting was closer to 10% for intersections, and smaller for highway segments with no intersections or interchanges, not 30% as commonly used by many transportation agencies.

The discussion of visual performance earlier in this chapter should reinforce the notion that improvements in visibility, the logical mechanism accounting for safety improvements associated with lighting, will depend upon the light level and other factors associated with a specific lighting configuration. A systematic examination of the visual performance benefits of roadway lighting of different light levels and spatial extents (i.e., one or two light poles at an intersection, versus a continuously lighted

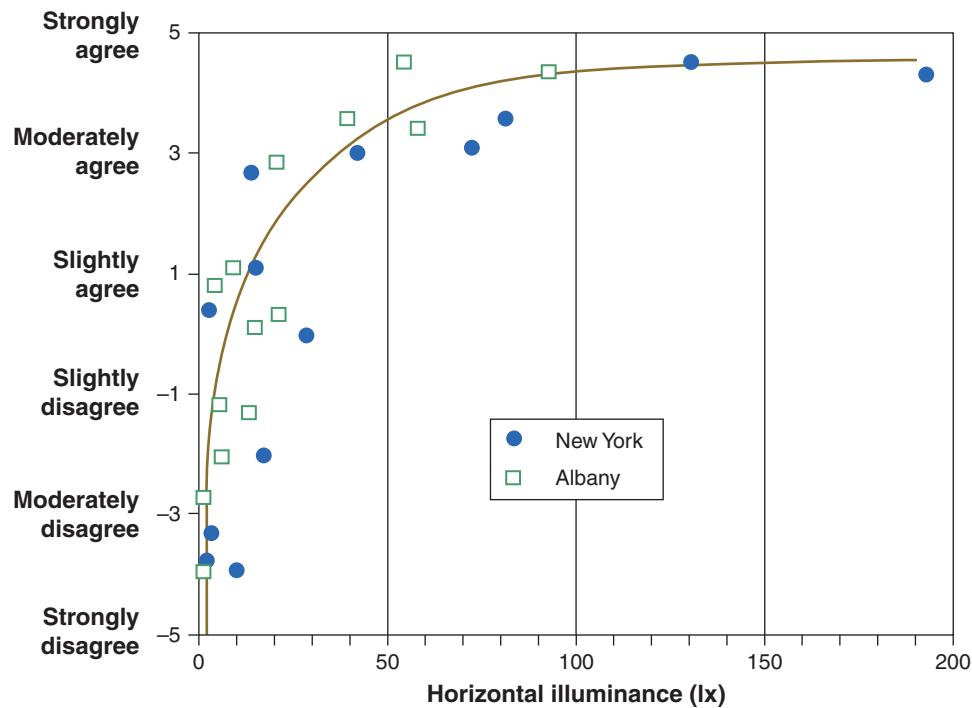


Figure 4 Agreement that outdoor lighting at night leads to perceptions of safety as a function of average illuminance on the pavement.

roadway approaching an intersection) used in conjunction with vehicle headlamps, and in locations varying in vehicle speeds, amount of ambient light from commercial properties, and for drivers of varying ages was conducted (Bullough and Rea, 2011) in conjunction with the statistical study of lighting and crashes by Bullough et al. (2013). The results of both studies yielded converging results.

For example, the benefit of roadway lighting at rural intersections was found to be negligible even though such areas are inherently dark at night, and lighting would be expected to improve visual performance, particularly when driving with low beam headlamps at speeds greater than 65 km/h, which are common on many rural highways. There were only modest visual performance improvements from “point” lighting consisting of only one or two poles at the intersection, especially for drivers along the secondary road of such an intersection. Continuous lighting of the rural highway, by comparison, would provide substantial improvement over no lighting or even “point” lighting, but is rarely used in practice (Bullough and Rea, 2011). Given these findings, drivers might be better served by lighting the approaching legs of a rural intersection and not only the immediate conflict area. It is unclear whether the practice of “beacon” lighting, where a low-wattage luminaire is mounted at a rural intersection along an otherwise unlighted road to provide an indication about the presence of the intersection, is beneficial in terms of crash reductions.

Roadway lighting can, however, improve subjective impressions of the safety of a location (Leslie and Rodgers, 1996). Surveys of exterior lighting in major urban areas as well as suburban locations found perceptions of safety and security to be positive when the average illuminance exceeded 10 lux (Fig. 4); higher light levels were not judged as feeling particularly safer.

Roadway Lighting Practices

For the most part, roadway lighting systems consist of luminaires mounted to poles, usually with individual controls consisting mainly of photocells. Poles are most commonly aluminum, steel, or wood. In many locations, existing utility poles are used to support roadway luminaires (in such cases, the lighting system is likely to be sub-optimal because utility pole locations are not determined by lighting or visibility considerations). If dedicated lighting poles are used, the utilities may be buried underground to improve the visual appearance. Cost may determine whether poles must be mounted along a single side of the road, or along both sides in an opposite or staggered arrangement.

In pedestrian crosswalks, one approach to lighting that has been investigated is the use of bollard-level lighting systems (Fig. 5) along the edges of the crosswalk (Bullough and Skinner, 2017). These systems provide illumination on the vertical surfaces of pedestrians crossing the street, in many cases making them more visible to oncoming drivers than overhead roadway lighting.

As mentioned previously, the most common control system for roadway lighting is a photocell mounted to an individual luminaire, which will switch the luminaire on and off at a specified ambient light level. Some systems will use centralized control via a single photocell or time clocks. Technological innovations are making centralized control systems attractive because these systems can also monitor the performance of individual luminaires and alert the system operator when a lamp or ballast failure has occurred or is close to occurring.



Figure 5 Photograph of bollard-level crosswalk lighting.

Standards such as those published by the [Commission Internationale de l'Éclairage \(2010\)](#) and the [Illuminating Engineering Society \(2018\)](#) are the basis for the design of continuous roadway lighting systems. These standards provide different light level recommendations as well as recommendations for uniformity (expressed as the ratio between the average and minimum light levels along the road) and the amount of disability glare that lighting systems should produce. These recommendations for uniformity and glare directly impact the spacing between poles for continuous roadway lighting. In general, busier roads and those with greater pedestrian conflict are expected to be lighted to higher light levels. The uniformity and glare recommendations are designed to help ensure that pole spacing between luminaires does not result in excessively dark portions of the illuminated roadway.

Just as vehicle lighting is beginning to implement adaptive control schemes, adaptive roadway lighting is also becoming more prevalent, based on the notion that very late in the night, there is likely to be much lower traffic and pedestrian use along many roads than there would be earlier in the evening. Accordingly, light levels may be reduced during these periods of reduced activity in order to save energy and minimize factors such as light trespass (light falling onto adjacent properties where it may be unwanted) and sky glow (brightening of the night sky making stars less visible). Such environmental impacts of roadway lighting are becoming more important to the general public.

References

- Andre, J., Owens, D.A., 2001. The twilight envelope: a user-centered approach to describing roadway illumination at night. *Hum. Fact.* 43 (4), 620–630.
- Bullough, J.D., 2009. Spectral sensitivity for extrafoveal discomfort glare. *J. Mod. Opt.* 56 (13), 1518–1522.
- Bullough, J.D., Donnell, E.T., Rea, M.S., 2013. To illuminate or not to illuminate: Roadway lighting as it affects traffic safety at intersections. *Accid. Anal. Prevent.* 53 (1), 65–77.
- Bullough, J.D., Rea, M.S., 2011. Intelligent control of roadway lighting to optimize safety benefits per overall costs. In: 14th Institute of Electrical and Electronics Engineers Conference on Intelligent Transportation Systems. Institute of Electrical and Electronics Engineers, New York, pp. 968–972.
- Bullough, J.D., Skinner, N.P., 2017. Real-world demonstrations of novel pedestrian crosswalk lighting. *Trans. Res. Rec.* 2661, 62–68.
- Bullough, J.D., Skinner, N.P., Plummer, T.T., 2016. Assessment of an adaptive driving beam headlighting system: visibility and glare. *Trans. Res. Rec.* 2555, 81–85.
- Commission Internationale de l'Éclairage, 2010. Lighting of roads for motor and pedestrian traffic, Report No. 115, Commission Internationale de l'Éclairage, Vienna.
- Commission Internationale de l'Éclairage, 1992. Road lighting as an accident countermeasure, Report No. 93, Commission Internationale de l'Éclairage, Vienna.
- Fry, G.A., 1954. A re-evaluation of the scattering theory of glare. *Illum. Eng.* 49 (2), 98–100.
- Illuminating Engineering Society, 2018. American National Standard Practice for Roadway Lighting, RP-8. Illuminating Engineering Society, New York.
- Leslie, R.P., Rodgers, P.A., 1996. *Outdoor Lighting Pattern Book*. McGraw-Hill, New York.
- Rea, M.S., (Ed.), 2000. *IES Lighting Handbook*, ninth ed. Illuminating Engineering Society, New York.
- Rea, M.S., Ouellette, M.J., 1991. Relative visual performance: a basis for application. *Light. Res. Technol.* 23 (3), 135–144.
- Schoettle, B., Sivak, M., Takenobu, N., 2008. Market-weighted Trends in the Design Attributes of Headlamps in the U.S. Society of Automotive Engineers, Warrendale, PA.
- Skinner, N.P., Bullough, J.D., Smith, A.M. 2010. Survey of the present state of headlamp aim. In: Transportation Research Board 89th Annual Meeting. Transportation Research Board, Washington.
- Sullivan, J.M., Adachi, G., Mefford, M.L., Flannagan, M.J., 2004. High-beam headlamp usage on unlighted rural roadways. *Light. Res. Technol.* 36 (1), 59–65.

Macroscopic Safety Analysis

Mohamed Abdel-Aty, Jaeyoung Lee, Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	367
Geographic Units	368
Census-Based Units	368
Traffic-Based Units	368
Traffic Analysis Zones	368
Traffic Analysis Districts	368
Traffic Safety Analysis Zones	368
Political Boundary Units	368
Postal Units (ZIP Code)	368
Developing New Geographic Units for Macroscopic Safety Analysis	369
Methodologies and Issues	370
Statistical Modeling Methods	370
Spatial Dependency	372
Boundary Crash Problem	373
Modifiable Areal Unit Problem	373
Macroscopic Contributing Factors to Traffic Crashes	374
Transportation Factors	374
Demographic Factors in Macroscopic Safety Studies	375
Socioeconomic Factors	375
Land-Use Factors	375
Climate Factors	376
Recent Research Trends	376
Simultaneous Analysis of Microscopic and Macroscopic Safety	376
Evaluating Policies for Safety	377
Integration with Transportation Planning	377
Summary and Conclusions	378
References	379
Further Reading	379

Introduction

Traffic safety is considered one of the most important areas in the transportation field. With consistent efforts of transportation engineers and researchers, both fatalities and fatality rates from traffic collisions have steadily declined between 2003 and 2011 in the United States. Nevertheless, traffic fatalities and fatality rates started to rise since 2012. And, as seen in an international perspective, the United States has gone from being among the top countries in traffic safety on a per capita as well as per-mile basis 40 years ago to nowadays being below many if not almost all other industrialized countries. In the United Kingdom along with Germany and France, traffic fatalities have been continuously decreasing since 2000.

Many researchers have conducted studies to reveal hot spots for traffic crashes and identify significant contributing factors to traffic crashes. In order to identify traffic crash prone locations or finding contributing factors for repeated traffic crash occurrence, specific traffic crash locations would be aggregated into segments, intersections, zones, etc. In general, traffic safety analysis can be classified into two categories: microscopic and macroscopic analyses. The microscopic safety analysis focuses on roadway entities such as segments, intersections, corridors, ramps, etc. The microscopic safety analysis aims at identifying contributing factors to traffic crashes from roadway geometric designs and traffic features of roadway entities, and providing specific engineering countermeasures to alleviate traffic safety problems. On the other hand, the macroscopic safety analysis concentrates on zonal-based traffic crashes with demographic, socioeconomic, land use, and zonal level traffic/roadway characteristics. Compared to the microscopic level study, macroscopic analysis provides a broad-spectrum perspective, and also it suggests long-term policy-based countermeasures such as enactments of traffic laws/rules, police enforcement, education, safety campaigns, and area-wide engineering solutions.

In the United States, Moving Ahead for Progress in the 21st Century (MAP-21) Act and Fixing America's Surface Transportation (FAST) Act proposed the requirement to incorporate traffic safety into the long-term transportation planning process. Therefore,

transportation engineers need to exert more effort to improve traffic safety and solve safety problems along with long-term transportation planning. This is the main reason that the macroscopic safety studies have become more popular during the last decade. The objective of this article is to comprehensively review and summarize macroscopic safety studies and provide researchers and practitioners with a better understanding and discuss recent trends in traffic safety research studies at the macroscopic level. This article comprises seven sections: geographic units, methodologies and issues, macroscopic contributing factors to traffic crashes, recent issues and topics, and summary and conclusions. Reference and further reading lists are provided at the end of this article.

Geographic Units

Macroscopic traffic safety studies require base geographic units. The number of crashes is aggregated into each spatial unit and the relationship between the aggregated crash counts (or crash rates) and zonal explanatory variables is explored. In the United States, which will be the focus of this description, geographic units include census-based units such as census blocks (CBs), census block groups (BGs), or census tracts (CTs). Traffic-based units such as traffic analysis zones (TAZs) or political boundary units such as counties have been used as base spatial units in some macroscopic safety studies.

Census-Based Units

CBs are the smallest geographic units used by the United States Census Bureau (USCB) for the collection and tabulation of decennial census data. However, detailed information is not available based on CBs due to confidentiality requirement. Generally, CBs are very small, especially in the urban areas. On average there are 85 people in one CB. Due to the privacy reasons, detailed information is not allowed for these small units, so CBs have not been used in macroscopic traffic safety studies. Census BGs are the next level above CBs. BGs are combinations of CBs. Each BG contains 39 CBs on average. Population in a BG ranges between 600 and 3000 people. Some macroscopic traffic safety studies adopted BGs as a base geographic unit. CTs are designed to maintain homogenous socioeconomic status in a zone. CTs are statistical subdivisions of a county that may include 2500–8000 people. Several researchers have analyzed macroscopic traffic safety based on CTs. Moreover, Census Wards, a census unit used in the United Kingdom, have been used in some macroscopic safety studies. **Fig. 1** compares CBs, BGs, and CTs in Downtown Orlando. As shown in **Fig. 1**, CBs are the smallest, whereas BGs and CTs are much larger than CBs.

Traffic-Based Units

Traffic Analysis Zones

TAZs are special-purpose geographic entities delineated by state and local transportation officials for tabulating traffic-related data, especially journey-to-work and place-of-work statistics. Since TAZs are the only traffic-related zone system, TAZs have been the most popular in the macroscopic safety literature.

Traffic Analysis Districts

Traffic analysis districts (TADs) are new, higher-level geographic entity for traffic analysis. TADs are created by aggregating existing TAZs. TADs may cross county boundaries, but they must nest within metropolitan planning organizations. Nevertheless, very few macroscopic safety studies have used TADs as a geographic unit in traffic safety analysis till now. **Fig. 2** compares TAZs and TADs in Orange County in Florida.

Traffic Safety Analysis Zones

Recently some researchers developed geographic units exclusively for traffic safety, called traffic safety analysis zones (TSAZs). TSAZs were developed using regionalization methods and estimated macroscopic safety models based on TSAZs. The development of TSAZs will be explained in detail later.

Political Boundary Units

For higher levels of macroscopic analysis, counties, states, or even nations are used as spatial units. One of the issues to analyze crashes using multinational data was the inconsistency in definition, quality, and availability of data.

Postal Units (ZIP Code)

ZIP code is a system of postal codes made and used by the United States Postal Service since 1963. As a matter of fact, ZIP codes are not a geographic unit but a collection of mail delivery routes. US census created ZIP code tabulation areas (ZCTAs), which are generalized areal representations of ZIP code service areas. ZCTA-based data are also provided from the USCB. Many studies using ZIP codes have been conducted. ZIP codes are mostly used for the residence information, instead of the crash location. This is because the residence information is only provided as a ZIP code in most cases. Many researchers have focused only on road users

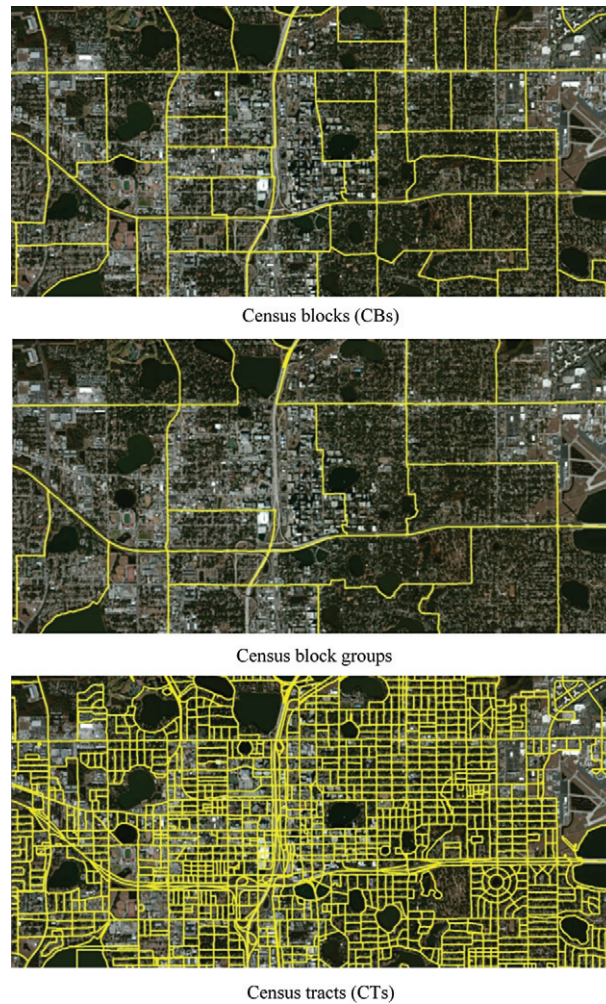


Figure 1 Census-based geographic units in Downtown Orlando, Florida.

who are involved in fatal crashes since Fatality Analysis Reporting System (FARS) offers ZIP codes of drivers involved in fatal crashes. Thus, ZIP code-based studies using FARS ZIP code data only focus on fatal crashes.

Developing New Geographic Units for Macroscopic Safety Analysis

As mentioned previously, TAZs have been most widely used in macroscopic safety studies. Nevertheless, there are possible limitations of TAZs for macroscopic safety analysis due to the zoning criteria. The general zoning criteria for TAZs are as follows (Baass, 1980):

1. Homogeneous socioeconomic characteristics for each zone's population.
2. Minimizing the number of intrazonal trips.
3. Recognizing physical, political, and historical boundaries.
4. Generating only connected zones and avoiding zones that are completely contained within another zone.
5. Devising a zonal system in which the number of households, population, area, or trips generated and attracted are nearly equal in each zone.
6. Basing zonal boundaries on census zones.

Criteria (1), (4), (5), and (6) are also sensible for the macroscopic safety modeling. Nevertheless, possible limitation of TAZs for the crash analysis arise from criteria (2) and (3). Basically, TAZs were designed to find out origin-destination pairs of trips generated from each zone. Thus, transportation planners need to minimize trips that start and end in the same zone. It is thought that minimizing intrazonal trips ends up with the small size of TAZs. On the other hand, traffic safety analysts need to analyze traffic crashes that occur inside the zone, so they are able to relate zonal characteristics with traffic crash patterns of the zone. Therefore, it is possible that TAZs are too small to analyze traffic crashes at the macroscopic level. Moreover, the small size of zones makes many zones with zero crash frequency, especially for rarely occurring crashes such as severe, fatal, or pedestrian crashes. Criterion (3) indicates that TAZs are

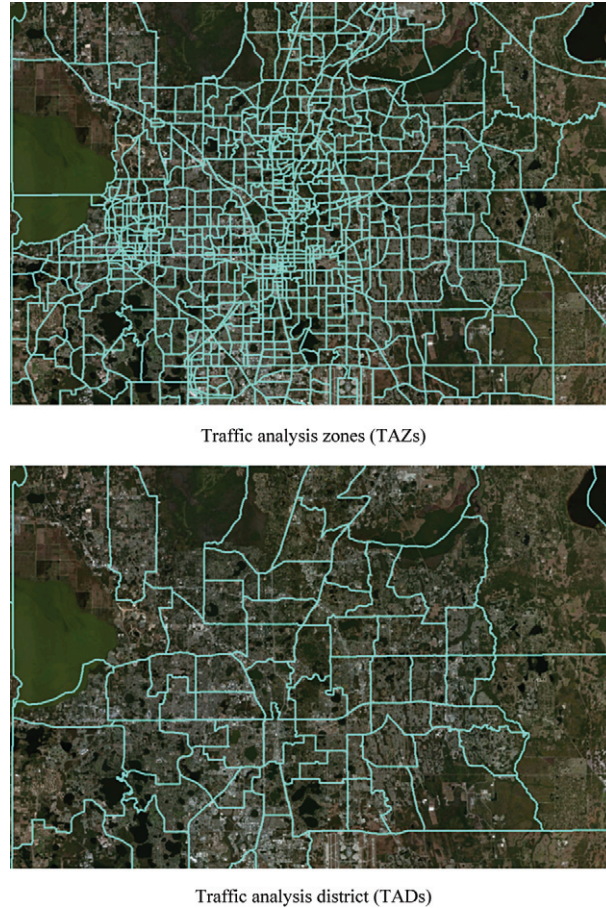


Figure 2 Traffic-based geographic units in Orange County, Florida.

usually divided based on physical boundaries, mostly arterial roadways. Considering that many of the crashes occur on arterial roads, between zones, inaccurate results will be made from relating traffic crashes on the boundary of the zone to only the characteristics of one zone. A simple way to overcome these two issues from using TAZs for the safety analysis is to aggregate TAZs into sufficiently large and homogenous traffic crash, sociodemographic, or traffic patterns (TSAZs). This process is called regionalization.

Methodologies and Issues

Statistical Modeling Methods

A wide array of statistical techniques for macroscopic safety analysis has been utilized. The statistical methods for macroscopic safety analysis are not very different from those for microscopic safety analysis. In this section, statistical models that are widely used for macroscopic safety models, including Poisson, negative binomial (NB), Poisson-lognormal, random effects, and multivariate models will be briefly addressed.

Since crash frequency data are nonnegative integers, the application of ordinary least squares regression is obviously not suitable. For crash count analyses, Poisson regression models have been used for several decades.

The equation of Poisson model is as follows:

$$P(y_i) = \exp \frac{(\lambda_i) \lambda_i^{y_i}}{y_i!} \quad (1)$$

where $P(y_i)$ is the probability of roadway entity i having y_i crashes per time period and λ_i is the Poisson parameter for the roadway entity (segment, intersection, etc.) i , which is equal to roadway entity i 's expected number of crashes per year, $E[y_i]$. Poisson regression models are estimated by specifying the Poisson parameter λ_i (the expected number of crashes per period) as a function of explanatory variables. Broadly used functional form for the Poisson regression is:

$$\lambda_i = \exp(\beta X_i) \quad (2)$$

where X_i is a vector of explanatory variables and β is a vector of estimable parameters.

The Poisson model is the most basic and also easy to estimate, but it cannot handle over-dispersion and under-dispersion. Over-dispersion is one characteristic of crash count data, which is that the variance exceeds the mean of the crash frequencies. Over-dispersion can violate the assumption of the equal mean-variance. Under-dispersion is that the mean of the crash counts is greater than the variance. Incorrect parameter is estimated in the existence of under-dispersed data. Moreover, the Poisson model is largely affected by low sample-mean and small sample size bias.

NB, or Poisson-gamma model, is an extension of the Poisson model to overcome the over-dispersion problem. NB model assumes that the Poisson parameter follows a gamma probability distribution. The NB model is derived by manipulating the relationship between the mean and the variance. The equation of NB model is as follow:

$$\lambda_i = \exp(\beta X_i + \varepsilon_i) \quad (3)$$

where i is each observation, $\exp(\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance α . The addition of variance term α allows the variance to differ from the mean as:

$$\text{VAR}[y_i] = E[y_i][1 + \alpha E[y_i]] = E[y_i] + \alpha E[y_i]^2 \quad (4)$$

Both the Poisson and NB regression models had been widely used since they can well represent crash count data. Additionally, the NB model performs better than Poisson regression model, especially when over-dispersion exists. Although NB model still cannot handle under-dispersion problem and it is also affected by low sample-mean and sample size bias, NB models are the most frequently used models in crash frequency.

In recent years, several traffic crash studies have been conducted using Poisson-lognormal models. The Poisson-lognormal model was suggested as an alternative to the NB model for crash count data. The Poisson-lognormal model is similar to the NB model, but the $\exp(\varepsilon_i)$ is a lognormal rather than gamma-distributed. Admittedly, the Poisson-lognormal can provide more flexibility compared to the NB model; it has two disadvantages: (1) model estimation is more complicated, and (2) still negatively affected by small sample sizes and low sample-mean values.

The correlation among observations could arise from spatial and/or temporal considerations. Random-effects model and fixed-effects models can be considered to account for such correlation. Random-effects model is considered where the common unobserved effects are assumed to be distributed over the spatial/temporal units according to some distribution and shared unobserved effects are assumed to be uncorrelated with explanatory variables. Fixed-effects models are considered where common unobserved effects are accounted for by indicator variables and shared unobserved effects are assumed to be correlated with explanatory variables.

Random-effects models modify the Poisson parameter as:

$$\lambda_{ij} = \exp(\beta X_{ij}) \exp(\eta_j) \quad (5)$$

where λ_{ij} is the expected number of crashes for roadway entity i belonging to group j , and η_j is a random effect for observation group j . The most common model is derived by assuming η_j is randomly distributed across groups such that $\exp(\eta_j)$ is gamma-distributed with mean 1 and variance α . The Poisson model limits the mean and variance to be equal.

$$E[y_{ij}] = \text{VAR}[y_{ij}] \quad (6)$$

And the Poisson variance to mean ratio is $1 + \lambda_{ij}/\alpha$ with random effects.

Bivariate/multivariate models become necessary when modeling multiple crash types. For instance, the number of fatal crashes cannot increase or decrease without affecting the counts of property damage only (PDO) or injury crashes. This problem can be solved using bivariate/multivariate models since they obviously consider the correlation among the severity levels or crash types for each roadway entity or zone. Multivariate models have been employed such as multivariate Poisson model, multivariate NB model, and multivariate Poisson-lognormal model.

Recently, application of the Bayesian approach became popular for traffic crash studies. Bayesian methods provide a comprehensive and robust approach to model estimation. Moreover, Bayesian models do not depend on the assumption of asymptotic normality underlying classical estimation methods as maximum likelihood. Traditional estimation methods such as the least square estimation are designed to find single estimate point. However, Bayesian estimation focuses on the entire density of parameters. For instance, in classical statistics the prediction of out-of-sample data often involves calculating moments or probabilities from the assumed likelihood for y , $p(y|\theta_m)$, which is evaluated at the selected point estimate θ_m . However, the information about θ is contained in the posterior density $p(\theta|y)$ in the Bayesian method; thus, prediction is estimated based on averaging $p(y|\theta)$ over this posterior density. It was argued that among the benefits of the Bayesian approach are a more natural interpretation of parameter intervals, often termed Bayesian credible or confidence intervals, and the freedom of obtaining true parameter density. On the contrary, maximum likelihood estimates rely on normality approximations based on large sample asymptotic. New estimation methods assist in the application of Bayesian random-effects models due to pooling strength across sets of related units.

There are several unique characteristics of the macroscopic safety analysis, which must be considered in the methodology. The first one is the spatial dependence of traffic crashes. Most statistical models assume that the values of observations in each sample are

independent or randomly distributed. However, a positive spatial autocorrelation may violate this assumption, if the samples were collected from nearby zones. Many researchers found spatial autocorrelations in the traffic crash data. The second one is regarding the boundary problem. Since TAZs are often delineated by arterial roads, most crashes occur on zone boundaries. The existence of boundary crashes may invalidate the assumptions of modeling only based on the characteristics of a zone where the crash is spatially located. The third is the modifiable areal unit problem (MAUP), which is presented when artificial boundaries are imposed on continuous geographical surfaces and the aggregation of geographic data causes the variation in statistical results. These characteristics are discussed in the following sections.

Spatial Dependency

Spatial autocorrelation is a technical term for the fact that spatial data from nearby locations are more likely to be similar than data from distant locations. The existence of spatial autocorrelation in the crash data may invalidate the assumption of the random distribution. Thus, it is required to test for the presence of spatial autocorrelations in the data set before the model estimation. If the spatial autocorrelation is detected, the crash model should account for spatial effect. Many researchers showed that spatial autocorrelations are found in the traffic crash data, and commonly observed that accounting for the spatial autocorrelations significantly improves the crash model performance. Fig. 3 shows an example of the spatial dependency in crash data. As shown in the figure, crashes are concentrated in the west and southeast counties. On the other hand, northwest counties have much less crashes compared to other regions. It is a good example of the spatial dependency that the crash frequency from near counties is more likely to be similar than that from distant locations.

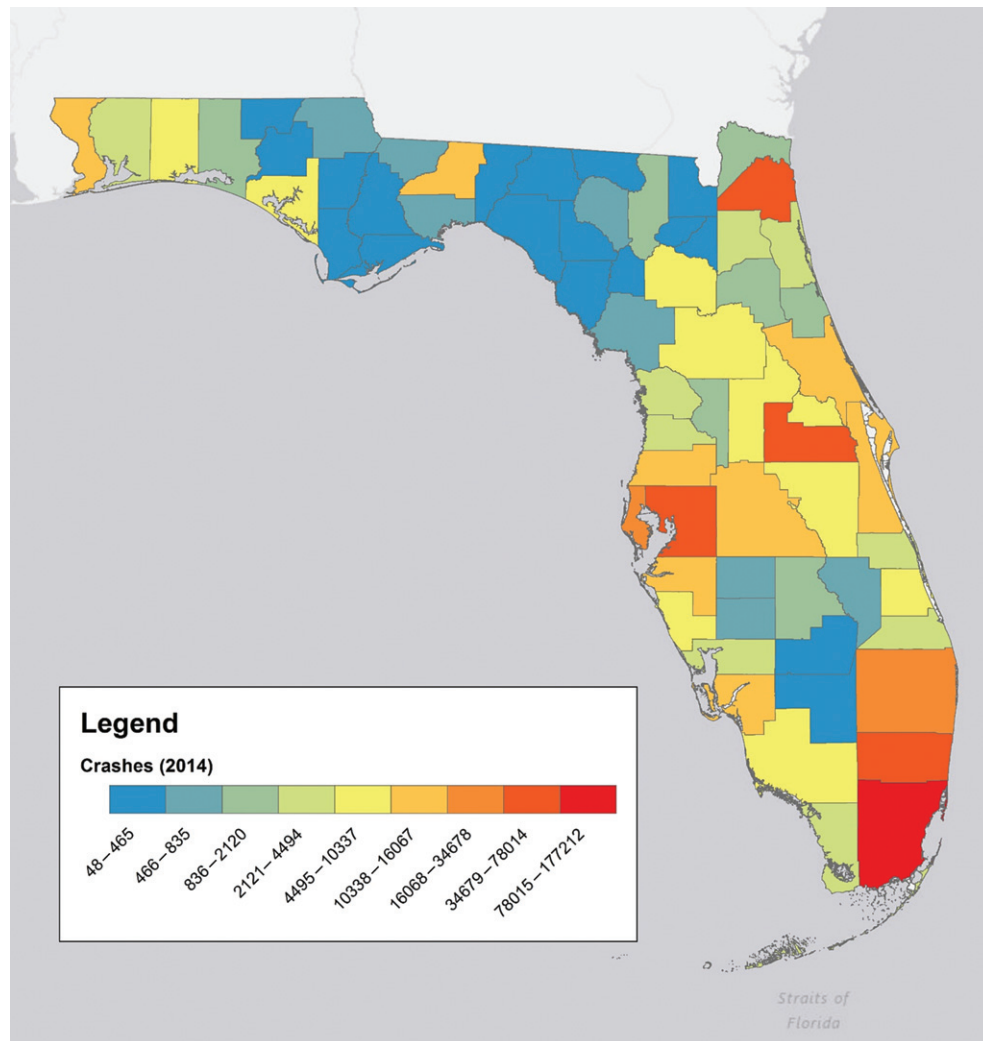


Figure 3 An example of spatial dependency in Florida.



Figure 4 An example of boundary crashes in Orlando, Florida.

Boundary Crash Problem

In the spatial analysis, boundary problems originate from ignoring the interdependences that occur from outside the boundary of zones. In the macroscopic analysis, the boundary problem is much more crucial and unique. Since TAZs are often delineated by arterial roads, most crashes occur on zone boundaries. **Fig. 4** displays an example of boundary crashes. The yellow lines in the figure are major arterials and red points show the location of crashes. As seen in the figure, the majority of crashes occur on or near the boundary of TAZs. The existence of boundary crashes may invalidate the assumptions of modeling only based on the characteristics of a zone where the crash is spatially located. One of the methods to solve the boundary crash problem is to estimate crash models separately for interior and for boundary crashes. Also, it is desirable to presume that boundary crashes are influenced by two or more adjacent zones (**Fig. 5**). Although the boundary crash issue in the macroscopic safety analysis is very important, not many research studies have addressed this issue. This is certainly an active area of further studies.

Modifiable Areal Unit Problem

The MAUP is present when artificial boundaries are imposed on continuous geographical surfaces and the aggregation of geographic data causes the variation in statistical results. Assuming that areal units in a particular study were specified differently, it is possible that very different patterns and relationships are shown. MAUP is composed of two effects: scale effects and zoning effect. Scale effects result from the different level of spatial aggregation. For example, traffic crash patterns are differently described in lower aggregation spatial units such as TAZs and higher aggregation units such as counties or states. Meanwhile, zoning effects are from the different zoning configurations at the same level of the spatial aggregation.

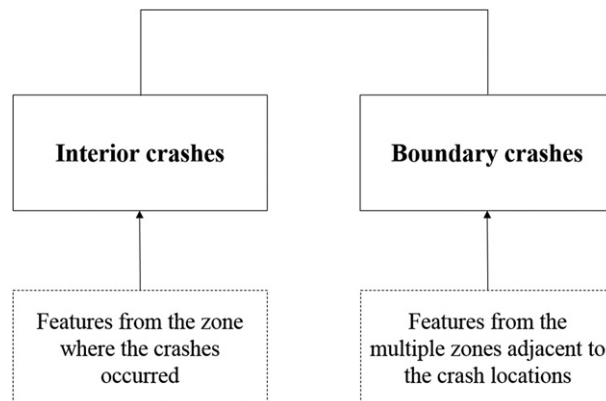


Figure 5 Separate modeling of interior and boundary crashes.

Still, few studies have addressed the MAUP on macroscopic traffic crash modeling to date. One of the important studies regarding the MAUP is conducted by [Abdel-Aty et al. \(2013\)](#). The authors conducted an interesting research regarding the effect from different zone systems. The authors compared crash models based on three different areal units BGs, CTs, and TAZs. The authors also discovered that the BG-based model had the larger number of significant variables for total and severe crashes compared to models based on other geographical units.

Macroscopic Contributing Factors to Traffic Crashes

Macroscopic crash studies that have been conducted typically have analyzed total crashes and sometimes explored crashes by severity level (i.e., total, PDO, injury, severe, fatal, etc.). Moreover, crash types/modes such as nonmotorized modes, for example, bicyclists or pedestrians, have been widely explored. In this section, contributing factors for the specific types or severity levels of the crashes are discussed based on results from previous studies carried out by [Hadayeghi et al. \(2003\)](#), [Noland and Quddus \(2004\)](#), [Kim et al. \(2006\)](#), [Naderan and Shahi \(2010\)](#), [Abdel-Aty et al. \(2013\)](#), [Lee et al. \(2014\)](#), [Lee et al. \(2015\)](#), [Chung et al. \(2018\)](#), and [Lee et al. \(2019\)](#).

Transportation Factors

Transportation data include traffic, roadway, and travel characteristics. **Table 1** summarizes the list of significant transportation factors to several key crash types (i.e., total, severe, bicycle, and pedestrian-involved crashes from the previously mentioned macroscopic safety studies). The identified factors were classified into positive or negative association categories. The positively associated factors are likely to increase crashes, while the negatively associated factors are likely to decrease the number of crashes.

“Vehicle-miles-traveled” and “total road miles” have frequently been used as a traffic exposure variable in the models. Other exposure (or surrogate exposure) variables such as bicycle-lane length and mode share of bicycles for bicycle crashes and pedestrian walking hours, and transit/pedestrian accessibility for pedestrian crashes were also found significant. Moreover, multiple studies showed that the percentage of length of high-speed roads increases total, severe, bicycle, and pedestrian-involved crashes. On the other hand, “volume-to-capacity” and “speed limit” have negative association with total and severe crashes (i.e., safer).

Work/college trip production is positively associated (i.e., more dangerous), but school/recreational trip production is negatively associated with total crashes (i.e., safer). Also, the number of intersections and intersection density are positively related to bicycle and pedestrian-involved crashes (i.e., more dangerous), respectively. It implies that both bicyclists and pedestrians are more vulnerable at intersections.

Table 1 Transportation factors in macroscopic safety studies

<i>Crash type</i>	<i>Positive association</i>	<i>Negative association</i>
Total	Vehicle-miles-traveled	
	Total road miles	
	Percentage or length of high-speed roads	Volume-to-capacity
	Speed limit	School/recreational trip production
	Work/college trip production	
Severe	Vehicle-miles-traveled	Volume-to-capacity
	Total road miles	Speed limit or average speed
	High-speed road length	Short commute time
Bicycle	Vehicle-miles-traveled	
	Bicycle-lane length	Low-speed road length
	High-speed road density	Average travel time to work
	Number of intersections	Mode share of cars
	Mode share of bicycles	
Pedestrian	Vehicle-miles-traveled	
	Pedestrian walking-hours	
	Intersection density	Provision of sidewalk
	Transit availability	People working at home
	Pedestrian accessibility	
	Percentage of high-speed road length	

Table 2 Demographic factors in macroscopic safety studies

<i>Crash type</i>	<i>Positive association</i>	<i>Negative association</i>
Total	Population Population density Minority population or percentage	Percentage of elderly people
Severe	Number of household Population density Minority population or percentage Percentage of working age people Percentage of children	Percentage of elderly people
Bicycle	Population Percentage of elderly people	Percentage of old-middle age group (45–64)
Pedestrian	Population Percentage of minority population Percentage of young-elderly Population (65–74)	Percentage of children (0–15) Percentage of elderly people

Demographic Factors in Macroscopic Safety Studies

Demographic data refer to data about the features of a population including but not limited to age, gender, and race/ethnic groups in the population. Multiple demographic factors were found significant for crashes (**Table 2**). “Population” and “population density” were positively associated with many crash types. For total, severe, and bicycle crashes, “percentage of elderly people” was negatively associated with crashes. However, it does not mean the elderly people are safer; but it might show that elderly people’s social activities are lower than young age group (15–25 years) and elderly people are less exposed to traffic crashes. It is noteworthy that the minority population (e.g., Hispanics and African Americans) is more vulnerable for total, severe, and pedestrian crashes.

Socioeconomic Factors

Socioeconomic data are defined as data of society and economy, which are related to employment, income, vehicle ownership, education level, crimes, and industry type. **Table 3** summarizes the socioeconomic factors to crashes in our study. “Total employment” was found significant and has positive effects on total and severe crashes. Also, “percentage of tertiary industry occupations” is negatively associated with pedestrian crashes. On the other hand, “percentage of the unemployed” increases the number of pedestrian crashes. It is interesting that “percentage of people working at home” is negatively associated with total crashes. This association is reasonable as people working at home do not need to commute and are thereby less exposed to traffic crashes.

“Income” is also a key factor in traffic safety. It was commonly shown that income is negatively associated with total, severe, bicycle, and pedestrian crashes. In other words, areas with higher income are likely to have a smaller number of crashes, while those with lower income tend to have a larger number of crashes.

Land-Use Factors

Land use is an important characteristic not only for traffic safety but also for transportation planning. Land-use characteristics determine the number of generated trip types by purpose, and also determine the modes of transportation. Previous macroscopic

Table 3 Socioeconomic factors in macroscopic safety studies

<i>Crash type</i>	<i>Positive association</i>	<i>Negative association</i>
Total	Total employment	Income Vehicle ownership Percentage of people working at home
Severe	Total employment	Income Education level
Bicycle	Alcohol consumption	Income Vehicle ownership
Pedestrian	Percentage of the unemployed Crime rates	Income Percentage of tertiary industry occupations

Table 4 Land-use factors in macroscopic safety studies

<i>Crash type</i>	<i>Positive association</i>	<i>Negative association</i>
Total	Mixed-use development areas Urbanization	—
Severe	Urbanization Commercial area	Residential area
Bicycle	Urbanization	—
Pedestrian	Commercial area Bars per roadway length	—

Table 5 Climate factors in macroscopic safety studies

<i>Crash type</i>	<i>Positive association</i>	<i>Negative association</i>
Total	—	—
Severe	Total precipitation Annual rain hours	Annual snow hours Annual fog hours
Bicycle	—	—
Pedestrian	Days with the maximum high temperature over 90°F	—

safety studies found several land-use factors (Table 4). For total crashes, “mixed-use development areas” and “urbanization” are positively associated with the number of crashes. Both “urbanization” and “commercial area” were positively associated with severe crashes. For bicycle crashes, “urbanization” was the only factor and it was positively associated. Pedestrian crashes have two land-use factors: “commercial area” and “bars per roadway length.” Lastly, “residential area” was negatively associated with severe crashes. In residential areas, the speed limit is relatively low and it is expected that crashes are less severe.

Climate Factors

Here, environmental factors include only one meaning: climate/weather factors. Several climate factors were found significant for crashes (Table 5). “Total precipitation” and “annual rain hours” are positively associated with severe crashes, while annual “snow hours” and “annual fog hours” are negatively associated. Lastly, it is interesting that an area with more “days with the maximum high temperature over 90°F” (i.e., very hot days) is positively associated with pedestrian crashes.

Recent Research Trends

Macroscopic safety analysis has attracted a great deal of attention in recent years as the promise of an effective tool that provides a broad-spectrum perspective to understand traffic crashes. However, the macroscopic safety analysis has shortcomings. First, although it is a useful tool to get an understanding of a broad spectrum of traffic safety influencers, specific problems (e.g., geometric design, signal, etc.) are often ignored. Second, it is not yet applied in the real world and not considered as a practical tool. Thus, the following research topics have been explored to improve the practicability of the macroscopic safety analysis: simultaneous analysis of microscopic and macroscopic safety, evaluating policies for safety, and integration with transportation planning. The studies discussed in this section are listed in the reference and further reading lists.

Simultaneous Analysis of Microscopic and Macroscopic Safety

There have been efforts to link or integrate microscopic and macroscopic safety. In one study, a comparative analysis was performed between macroscopic and microscopic level models using the same data. The study concluded that the microscopic model outperforms its counterpart (i.e., macroscopic model). Nevertheless, the macroscopic model needs less detailed data and provides suggestions not only for those directly related to traffic but also for those related to social, economic, political, and educational issues. Some researchers integrated macroscopic and microscopic analyses in the modeling stage. They commonly asserted that the modeling performance has improved when two scopes are combined since the information from each side could be shared with another (Fig. 6). Furthermore, the integrated model can be used to predict crashes at the microscopic (i.e., segment and intersection) and macroscopic levels (i.e., zone), simultaneously.



Figure 6 Information exchange between macroscopic and microscopic safety model components.

Evaluating Policies for Safety

Many policies are directly related to traffic safety but some are not directly related. For instance, motorcycle helmet law, bicycle helmet law, primary enforcement of seat-belt use, cell-phone use ban, etc., affect road users' decisions to use safety equipment and cell phones while driving. Moreover, some policies indirectly related to traffic safety include marijuana legalization, alcohol sale hours, etc. Macroscopic safety analysis is a great tool to evaluate such policies because policies and regulations are often based on political boundaries such as city, county, state, and nation.

There are two recent macroscopic safety studies that evaluated changes in policy. The first study explored the safety effects of the universal helmet law abolishment and reinstatement. The study adopted a before-and-after study with the comparison group and empirical Bayes at the county level. The study concluded that the outcomes of the helmet law changes are evident. The number of motorcycle fatal crashes increased by 26%–41% after the abolishment in Florida and decreased by 21%–26% after the reinstatement in Louisiana. The second study investigates the safety effects of the marijuana legalization. This study is state-based research as a state government determines whether marijuana is prohibited, allowed for medical purposes, or fully legalized for its jurisdiction. The results indicated that there are no significant changes after the medical marijuana legalization. However, 174.5% increase in the number of marijuana-involved fatal crashes was found after the decriminalization in Massachusetts, and 75.3% increase was observed in Connecticut after the medical legalization when decriminalization is already in place. Both Washington and Colorado experienced significant increases (31%–63%) after the full marijuana legalization.

Integration with Transportation Planning

As described earlier, there have been efforts to integrate traffic safety and transportation planning (i.e., transportation safety planning). A recent study proposed an idea to predict safety levels in the future using the estimated macroscopic safety models. This approach could become very useful to decide where we should provide countermeasures in advance to prevent serious crash problems in the future. **Fig. 7** compares the hot zones in 2010–12 and those predicted for 2040–42. Although the two figures have

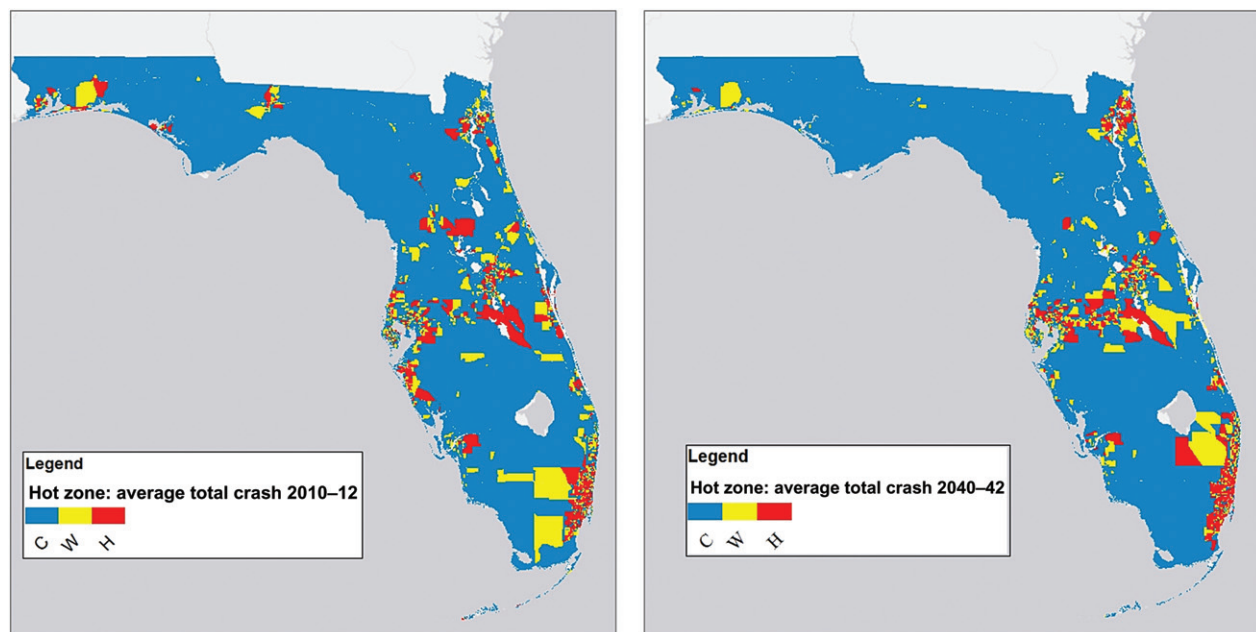


Figure 7 Comparison of hot zones in 2010–12 and 2040–42.

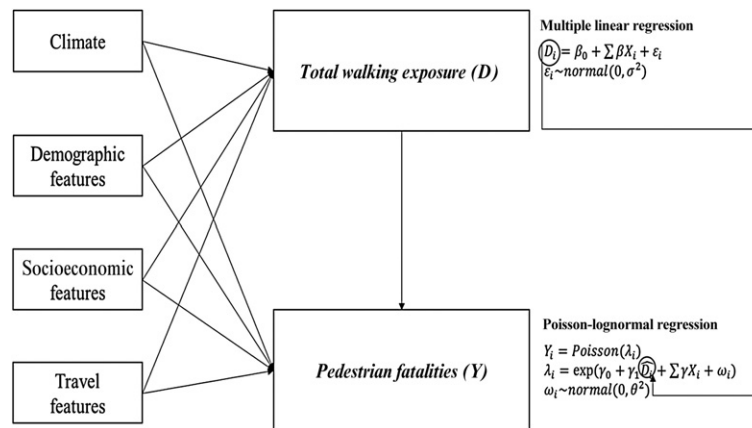


Figure 8 Schematic illustration of the integration framework.

similar hot zones in some areas, it is clear that the hot zones in the future are not exactly the same as before. Thus, more efforts and resources should be assigned to the prospective hot zones, proactively.

Meanwhile, there have been other efforts to integrate traffic safety and planning. Two recent research studies proposed a combined modeling framework that could simultaneously predicts trips and safety simultaneously. With the suggested approach, more accurate and reliable estimation is achievable.

Fig. 8 shows the schematic illustration of the integration framework. The first component estimates the pedestrian total walking hours using climate, demographic, socioeconomic, and travel features. Subsequently, the predicted pedestrian total walking hours from the first component are provided to the second component (i.e., pedestrian fatality model) as an exposure. The integrated framework could provide better performance compared to the independent trip and safety models.

Summary and Conclusions

In recent years, macroscopic safety studies have been emphasized in accordance with requirements in the MAP-21 and FAST Acts. The macroscopic safety studies focus on zonal-based traffic crashes with demographic, socioeconomic, land use, and zonal level traffic/roadway characteristics. The macroscopic studies provide a broad-spectrum perspective, and also they suggest long-term policy-based countermeasures such as enactments of traffic laws/rules, police enforcement, education, safety campaigns, and area-wide engineering treatments. Numerous researchers have conducted macroscopic safety studies since the last decade and it is necessary to summarize their efforts at this point. Thus, this article summarized and reviewed geographic units, statistical methods, crash types, contributing factors, and current issues in macroscopic safety studies.

Researchers have chosen geographic units for their studies dependent on the objectives, perspectives, data availability, and so on. The geographic units can be census-based units (i.e., BGs, CTs, etc.), traffic-based units (i.e., TAZs, TADs, TSAZs, etc.), or political boundaries (i.e., counties, states, nations, etc.). Among the geographic units, TAZs have been most widely selected as geographic units because TAZs are the only traffic/transportation-related units. Also, since TAZs are also being used for transportation planning, it is easier to incorporate traffic safety into long-term transportation plans. Lately, TADs, which is aggregation of TAZs, are developed by the USCB. It is also related to transportation but limited research studies have been conducted based on TADs. It is necessary to investigate the possibility of TADs for macroscopic safety studies. Moreover, new safety spatial units (i.e., TSAZs) could be considered, which is exclusively developed for traffic safety analysis.

Poisson models and their variants have been developed and applied for traffic crash analysis. As in the microscopic safety studies, Poisson and NB models have been popularly used for the macroscopic safety research. Poisson model is the most basic and easy to estimate; however, it is unable to handle over-dispersion. NB model was suggested to overcome the over-dispersion issue in Poisson model. Moreover, some researchers applied Poisson-lognormal models, which was suggested as an alternative to the NB model for crash frequency data. It was claimed that Poisson-lognormal models are capable to provide more flexibility than NB models. Multivariate models are required when modeling multiple crash types simultaneously. Multivariate models consider the correlation among severity levels or crash types for each roadway entity of zone. Furthermore, application of Bayesian framework has been popular in the traffic safety field. Bayesian models do not depend on the assumption of asymptotic normality. Sampling-based methods of Bayesian estimation focus on estimating the entire density of parameters, whereas the traditional classical estimation methods that are intended for finding a single point estimate using the maximum likelihood approach. Furthermore, some studies started to apply the state-of-the-art techniques such as machine learning and deep learning. It is expected that applying such new techniques will provide much higher accuracies in crash prediction, especially, with Big Data. There are three important and unique characteristics of the macroscopic safety analysis, which must be considered in the analysis: (1) spatial dependency, (2) boundary crash problem, and (3) MAUP. Without considering these three features, it is likely that the results are biased and unreliable.

As summarized in this article, there are various contributing factors to crashes, including transportation, demographic, socioeconomic, land use, and climate factors. A good understanding of the effects of the factors is the first step to reduce crashes by providing effective countermeasures. There are several important research topics in the macroscopic safety analysis. First is the simultaneous analysis of microscopic and macroscopic safety. Each scope has its own advantages and it is not desirable to rely on only one scope. From the macroscopic analysis, we can understand a broad perspective of traffic crashes of areas. Then, detailed in depth analysis about the specific problem(s) at intersections or segments in the high risk zones. Furthermore, multiple recent studies began to apply macroscopic methodologies to assess policies. In this article, we took the universal helmet law and marijuana legalization as examples. Lastly, there have been efforts to combine traffic safety and planning. We are able to predict safety levels in the future by applying the macroscopic methodologies, and also planning and safety models could be integrated at the modeling stage.

The macroscopic safety analysis is still a relatively new tool in the traffic safety field. It has advantages including but not limited to providing a broad view of traffic safety situations, identifying not only traffic/roadway factors but also demographic, land use, and socioeconomic factors, being easily integrated with planning, and suggesting policy-based long-term countermeasures. Nevertheless, there are also challenges and disadvantages such as spatial dependency, boundary crash problem, MAUP, and limited practicability. In the future, we need to overcome these challenges and disadvantages of the macroscopic safety analysis, make it more practical, and save more lives from traffic crashes.

References

- Abdel-Aty, M., Lee, J., Siddiqui, C., Choi, K., 2013. Geographical unit based analysis in the context of transportation safety planning. *Transp. Res. Part A Policy Pract.* 49, 62–75.
- Baass, K.G., 1980. Design of zonal systems for aggregate transportation planning models. *Transp. Res. Rec.* 807, 1–6.
- Chung, W., Abdel-Aty, M., Lee, J., 2018. Spatial analysis of the effective coverage of land-based weather stations for traffic crashes. *Appl. Geogr.* 90, 17–27.
- Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2003. Macrolevel accident prediction models for evaluating safety of urban transportation systems. *Transp. Res. Rec. J. Transp. Res. Board* 1840 (1), 87–95.
- Kim, K., Brunner, I.M., Yamashita, E.Y., 2006. Influence of land use, population, employment, and economic activity on accidents. *Transp. Res. Rec. J. Transp. Res. Board* 1953 (1), 56–64.
- Lee, J., Abdel-Aty, M., Jiang, X., 2014. Development of zone system for macro-level traffic safety analysis. *J. Transp. Geogr.* 38, 13–21.
- Lee, J., Abdel-Aty, M., Jiang, X., 2015. Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. *Accid. Anal. Prev.* 78, 146–154.
- Lee, J., Abdel-Aty, M., Huang, H., Cai, Q., 2019. Transportation safety planning approach for pedestrians: an integrated framework of modeling walking duration and pedestrian fatalities. *Transp. Res. Rec.*, doi:10.1177/0361198119837962.
- Naderan, A., Shahi, J., 2010. Aggregate crash prediction models: introducing crash generation concept. *Accid. Anal. Prev.* 42 (1), 339–346.
- Noland, R.B., Quddus, M.A., 2004. Analysis of pedestrian and bicycle casualties with regional panel data. *Transp. Res. Rec. J. Transp. Res. Board* 1897 (1), 28–33.

Further Reading

- Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., Wang, X., 2018. Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection crash modeling. *Anal. Methods Accid. Res.* 19, 1–15.
- Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., Abdel-Aty, M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. *J. Transp. Geogr.* 54, 248–256.
- Lee, J., Abdel-Aty, M., Cai, Q., 2017a. Intersection crash prediction modeling with macro-level data from various geographic units. *Accid. Anal. Prev.* 102, 213–226.
- Lee, J., Abdel-Aty, M., Wang, J.H., Lee, C., 2017b. Long-term effect of universal helmet law changes on motorcyclist fatal crashes: comparison group and empirical Bayes approaches. *Transp. Res. Rec.* 2637 (1), 27–37.
- Lee, J., Abdel-Aty, A., Park, J., 2018. Investigation of associations between marijuana law changes and marijuana-involved fatal traffic crashes: a state-level analysis. *J. Transp. Health* 10, 194–202.

Motor Vehicle Crash Reportability

John J. McDonough, National Institute for Safety Research, Inc. (NISR), Pinehurst, NC, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	380
Challenge of Defining Reportability	380
Narrowing the Focus	381
What is a Crash?	381
What is a Motor Vehicle?	382
What is a Motor Vehicle Crash?	383
What is a Motor Vehicle Traffic Crash?	384
Biography	385
See Also	385
Relevant Websites	385
References	385

Introduction

Crash reportability discussed here will be limited to ground transportation and the data collection on and classification of crashes involving motor vehicles. This article will present the technical classification component of what constitutes a motor vehicle traffic crash from a traffic safety perspective for statistical reporting purposes. Events involving other transportation modes, although mentioned, will not be addressed in significant detail.

The technical terminology used in the article can be found in [ATSIP \(2017\)](#). This document has been in use in the United States for decades to support uniformity in reporting, classification, and analysis of motor vehicle traffic crash data. These data become the foundation for research and decision-making concerning traffic safety at the federal, state, and local levels in the United States as well as serving to inform the public and other users of the data.

While the *Manual on Classification of Motor Vehicle Traffic Crashes, Eighth Edition*, is an American National Standard and the primary source for the terminology in this article, the terminology and definitions therein are either a direct match or provide enough detail to be relatable to similar terminology used in other traffic safety applications outside the United States. For example, the intended application and relationship between the ANSI D16.1 terms “trafficway” and “roadway” are identifiable as almost directly parallel to the terms “road” and “carriageway” used in [UNECE \(2017\)](#). The United Nations Economic Commission for Europe (UNECE) document *Statistics of Road Traffic Accidents in Europe and North America* uses terminology presented in the jointly developed *Glossary for Transport Statistics*. The *Glossary for Transport Statistics* was developed cooperatively between the UNECE, the International Transport Forum (ITF), and Eurostat. All three documents are included as additional reference with this article and terminology is referenced between the documents in select areas of discussion.

Challenge of Defining Reportability

There are many different stakeholders involved that have an interest in the reporting of motor vehicle crashes. General examples include international organizations, national governments, private industry (e.g., automobile manufacturers, insurance companies), public health organizations, local governments, law enforcement, the court system, emergency services, and many others. Like the adage “beauty is in the eye of the beholder,” it is not a stretch to suggest that crash reportability is in the eye of the stakeholder. Each stakeholder has their own domain of interest or responsibility, and what is “reportable” to one, may not be to another.

There is also complexity created by circumstances that influence what crashes are reported in practice, to whom they are reported, and what happens with the data that results. Examples include the severity of the crash, its location, the type of vehicles involved, and the resource burden and availability of the resources to collect the data. In the case of crash severity, for example, a motor vehicle crash that produces a fatality or serious injury is almost certainly going to be reported. A fatal motor vehicle crash will involve action and data collection on the part of law enforcement, emergency services, healthcare services, insurance companies, possibly the court system, news media, and representation in local, national, and even international statistics. Furthermore, depending on other factors such as the location where the crash occurred (e.g., public versus private property) and the type of vehicles involved (e.g., large truck or railway train), other entities may be included or excluded in the chain of data collection and reporting. In contrast, a motor vehicle crash involving two passenger cars where only minor vehicle damage occurs, the parties involved may agree the damage is minor enough to go their separate ways with no reporting of the incident through any formal means.

The challenge that the disparity in reporting creates when comparing data from various sources is acknowledged in many traffic safety research publications. Often, statistical methods are employed to develop estimates to account for the disparity. This challenge in comparability of data at the international level is recognized in [UNECE \(2017\)](#) in the Introduction, in a section titled *Comparability of Data* it states, “The comparability of data is also subject to differences in road accident reporting standards. Notably, the number of accidents and the number of injuries may be underreported in some countries due to administrative or practical limitations,” ([UNECE, 2017](#), p. V). In other words, although UNECE can precisely define what they intend to represent in the document’s reports, influencers like the items mentioned above have a significant effect on what appears in the end data provided by the various sources.

Narrowing the Focus

A broad discussion of reportability of motor vehicle crashes addressing all the stakeholders and influencing factors is too complex to present in a single article. Additionally, before challenges in comparing end data and the effects that various domains of interest have on an individual stakeholder’s view of what is reportable to them can be understood, it is necessary to define what constitutes a motor vehicle crash. There are precise definitions that from a traffic safety perspective can be learned as a base standard for understanding reportability of motor vehicle crashes. Even limiting the scope to just defining motor vehicle crashes for classification, data collection, and statistical reporting, the topic is still complex. To develop a full understanding, it is necessary to present the terminology and its component parts along with examples of its application with real-world examples. Circumstances that fall outside the boundaries set forth in the definitions are not reportable as motor vehicle crashes. Defining those boundaries and providing examples will be the focus of the remainder of this article.

What is a Crash?

The term motor vehicle crash has component parts. To understand what is and what is not reportable as a motor vehicle crash it is necessary to scrutinize the terminology more closely. A crash within the context of the *American National Standard (ANSI D16.1-2017)*, *Manual on Classification of Motor Vehicle Traffic Crashes, Eighth Edition* is defined as “an unstabilized situation which includes at least one harmful event,” ([ATSIP, 2017](#), p. 22). There are two parts to examine and understand in this definition within the context of a motor vehicle crash. One is the concept of an unstabilized situation and its limitations and the second part is the concept of harmful events and the exclusions in this term that may affect reportability.

Unstabilized situations are “a set of events not under human control,” ([ATSIP, 2017](#), p. 20), in other words, they are unintended sets of events. A key exclusion from the term unstabilized situation is a set of events that result from deliberate acts. That is, any set of events that is deliberately caused (e.g., homicide, suicide, injury, or damage purposely inflicted, etc.) is not classified as a crash within ANSI D16.1. These acts are intentional and thus no unstabilized situation exists. Logically then, deliberate acts involving motor vehicles are not reportable for motor vehicle crash data collection purposes. A similar exclusion for suicides is identified in the Definitions and General Notes section of [UNECE \(2017\)](#). Specific examples of deliberate intent that are not considered to be a reportable motor vehicle crash include; a driver that commits suicide by driving their vehicle off a cliff, an enraged driver that purposely collides with another vehicle, or a terrorist that intentionally strikes pedestrians on a sidewalk.

ANSI D16.1 suggests that “Normally, the medical examiner or coroner will be the final authority on matters pertaining to cause of death whether or not an autopsy is performed,” ([ATSIP, 2017](#), p. 36). The word “normally” is used because identification of deliberate acts can be challenging in some circumstances pertaining to cause of death. This is particularly true in the case of a suicide. A suicide death is one where a person takes actions intended to result in their death. In the case of a suicide by motor vehicle, the injuries to a suicide victim are not likely to be medically different than those to a fatally injured driver in a reportable motor vehicle crash. Consequently, if there is a suspected suicide by motor vehicle, the determination of intent is typically made by law enforcement as part of their investigation. In resolution of suspected suicides by motor vehicle, without specific evidence showing intent, the events are usually captured as reportable motor vehicle crashes.

It is important to note that the exclusion of deliberate acts does not extend to intentional or imprudent driver behaviors that result in unintended events from being considered a motor vehicle crash. For example, a speeding motorcyclist that loses control in a curve, leaves the roadway, and strikes a tree would be involved in a motor vehicle crash. Likewise, an automobile driver under the influence of alcohol that fails to yield at an intersection and collides with another vehicle would be involved in a motor vehicle crash. This is true in both cases even though the speeding by the motorcyclist and over consumption of alcohol by the automobile driver were intentional and imprudent behaviors. The control loss by the motorcyclist and intersection collision involving the driver under the influence were not intentional and thus are unstabilized situations. In fact, addressing unsafe driving behaviors such as the examples above are the focus of many traffic safety initiatives.

The second part of the definition of a crash to examine is what constitutes a harmful event within the context of a motor vehicle crash. The harm can be broadly grouped as an event that produces either injury or damage. The term injury is defined as “bodily harm to a person,” ([ATSIP, 2017](#), p. 18). Excluded from being considered an injury are the immediate effects of diseases. For example, if a person that has coronary artery disease experiences a heart attack while driving, the heart attack itself is not considered as an injury in classifying any crash that may occur.

Damage is defined as “harm to property that reduces the monetary value of that property”, (ATSIP, 2017, p. 19). In rough parallel to the injury exclusion, excluded from being classified as damage is a mechanical failure of the vehicle. For example, if a vehicle being driven on an expressway experiences a tire blowout, the tire blowout itself is not considered as damage in classifying any crash that may occur.

The harmful events in a crash can be used to classify crashes for reporting purposes by the harm associated with the crash as a measure of severity. As it pertains to motor vehicle crashes, severity is commonly separated into a three-category set where the classifications are mutually exclusive. Understandably, crashes that produce harm to people are considered more severe than crashes that only harm property. Motor vehicle crash severity has a significant impact on whether it is formally reported, the entities to which it is reported, the data collected, and the downstream use of the data for safety research, applicability of laws and regulations, and resources committed.

A three-category classification that separates crashes by injury severity would be as follows (ATSIP, 2017):

1. Fatal injury crash: injury crash where at least one person dies as a result of injuries sustained in the crash within 30 days of the crash occurrence.
2. Nonfatal injury crash: injury crash other than fatal injury crash
3. Property damage only crash: crash other than an injury crash

Common in crash data collection is an expansion of the nonfatal injury component so that more detailed information can be captured about the nature of the injuries. For example, the Model Minimum Uniform Crash Criteria (MMUCC) is a guideline developed cooperatively between agencies in the US Department of Transportation (USDOT), the Governors Highway Safety Association (GHSA), and the US States. MMUCC provides a set of motor vehicle crash data elements that should be collected for traffic safety applications. This guideline since its beginning in 1998 has recommended a data element that separates injury severity into five categories (P5. Injury Status in MMUCC). In the Introduction, MMUCC supports this idea when stating, “Good data about motor vehicle crashes is critical to help explain yearly fluctuations in motor vehicle deaths and injuries and guide policy makers as they consider appropriate investments to reduce those deaths and injuries,” (NHTSA, 2017, p. 3). These same additional sub-classifications exist in ANSI D16.1 and can be used to further separate the injury component in order of severity as fatal injury, suspected serious injury, suspected minor injury, possible injury, and no apparent injury.

What is a Motor Vehicle?

ANSI D16.1 defines a motor vehicle as “any motorized (mechanically or electrically powered) road vehicle not operated on rails,” (ATSIP, 2017, p. 4). Within the manual, motor vehicles fall under a broad classification of transport vehicles. Transport vehicles are devices used for transportation including the device itself, connected transport devices (e.g., truck with a trailer), and any load of the transport vehicle. The broad classification of transport vehicle includes all modes of transportation. To separate in the classification those vehicles that operate on land from those that operate in the air or on the water, the term land vehicle is applied. Next, the term road vehicle excludes land vehicles that operate on railways. Road vehicles then can include transport vehicles such as automobiles, motorcycles, bicycles, trailers, horse-drawn carriages, etc. The final refinement comes in the term motor vehicle where motorization is required.

The Glossary for Transport Statistics goes through a parallel refining process using similar terms and definitions to define “road motor vehicles.” Within this document, the terms used include vehicle, road vehicle, and road motor vehicle. These term’s definitions include the same concept embedded within the ANSI D16.1 term transport vehicle, the concept of being used for transportation. *The Glossary for Transport Statistics* identifies that road motor vehicles are “normally used for carrying persons or goods or for drawing, on the road, vehicles used for the carriage of persons or goods,” (IWG, 2009, p. 43). The process of refinement of the terminology through successive definitions employed in both documents narrows and expands the scope for data collection and statistical reporting. For example, a crash involving an automobile, a bicycle, and a train would involve three vehicles, two road vehicles (train, bicycle), and one road motor vehicle (automobile).

There are few examples of note with respect to motor vehicles that pertain to reportability of motor vehicle crashes as applied in ANSI D16.1. First, like the prior example, crashes involving only a bicycle (without a motor) are not reportable as motor vehicle crashes. A second item of note in this area is that current classification within ANSI D16.1 includes a motorized bicycle as a class of motorcycle and thus as a motor vehicle. This distinction separates motorized bicycles from human-powered bicycles. In ANSI D16.1, human-powered devices propelled by pedaling are called pedal cycles. It should be noted that this same distinction does not exist in *The Glossary for Transport Statistics* where the term bicycle includes “cycles with supportive power unit” (IWG, 2009, p. 43).

Another class of device worth noting with respect to crash reportability within the context of ANSI D16.1 is a personal conveyance. These are devices “used by a pedestrian for personal mobility assistance or recreation,” (ATSIP, 2017, p. 4). Examples include skateboards, roller skates, wheelchairs, electric scooters, and others. A personal conveyance is excluded from being a transport device, thus these devices even when motorized are not classified as motor vehicles. Pertaining to reportability this means that if a person that falls and injures themselves while riding an electric scooter or skateboard, without the involvement of a motor vehicle, it is not reportable as a motor vehicle crash. Further, if a person on a personal conveyance is involved in a motor

vehicle crash (e.g., skateboarder struck in a crosswalk by an automobile), the motor vehicle crash would be classified as involving a single motor vehicle and the person operating the personal conveyance is a special class of pedestrian.

A last important motor vehicle crash classification point to make with respect to motor vehicles being transport vehicles is the concept that any load of the transport vehicle is part of that vehicle. Loads of vehicles include persons and property upon or set in motion by the vehicle, attached to and in position to move with the vehicle, or persons boarding or alighting from the vehicle (ATSIP, 2017, p. 2). This classification distinction means that if a person riding in the back of a pickup truck on a roadway falls from the truck bed and is hurt, they are classified as an occupant (passenger) of that vehicle. In this example, there has been an unstabilized situation and harmful event involving a motor vehicle resulting in a motor vehicle crash. The same would be true of cargo that falls from a vehicle while traveling on a highway. If injury or damage results from the falling or shifting cargo to the vehicle hauling the cargo or its occupants, other vehicles, persons, or property, or even damage to the cargo itself, then the vehicle has been involved in a motor vehicle crash.

What is a Motor Vehicle Crash?

Establishing that a crash is an unstabilized situation with a harmful event and that motor vehicles are mechanically or electrically powered road vehicles, the details of motor vehicle crashes is the next item to discuss. First, a motor vehicle crash is a class of transport crash. There are some key distinctions associated with transport crashes that can affect motor vehicle crash reportability. A transport crash requires that a transport vehicle in-transport is involved. To be involved means that an in-transport motor vehicle is directly associated with a harmful event. This can be injury to its occupants, other vehicle occupants, or other persons (e.g., pedestrian), or damage to the in-transport motor vehicle, other motor vehicles, or other property (e.g., a citizen's fence). When applied to motor vehicles, a motor vehicle is in-transport when it is in motion in any location, or on the roadway (travel lanes) of a trafficway in-motion or stopped (ATSIP, 2017, p. 15). The concept of being in-transport whenever on a roadway addresses that there are normal functions of transportation involving roadways where a vehicle is stopped. For example, stopped at a traffic control, when yielding before making a turn, or in traffic back-ups.

Excluded from being in-transport are motor vehicles that are parked. Parked motor vehicles are not in-motion and not on a roadway. In other words, the opposite of in-transport. In addition, because in-transport is associated with performing a transportation function, motor vehicles that are in use for performing construction, maintenance, or utility work associated with the trafficway are classified as a specific type of not in-transport motor vehicle using the term Working Motor Vehicle. That is, when engaged in these activities, these motor vehicles are performing a work function related to the trafficway, not a transportation function. What this means with respect to motor vehicle crash reportability is that if only a working motor vehicle is involved, there is no motor vehicle crash. For example, if two motor vehicles collide that are part of a mobile work activity such as repainting of the lane line markings, then there is not a reportable motor vehicle crash (no motor vehicle in-transport involved). It is important to note that the working motor vehicle exclusion is limited strictly to construction, maintenance, or utility work associated with the trafficway and only applies during the period that the work is being performed. For example, a highway maintenance vehicle, that is, traveling from one work site to another would be classified as in-transport. Also, of note, the term working motor vehicle does not extend to other vehicles that are associated with a business function (e.g., delivery vehicles, trash trucks) or being driven by people as part of their jobs (e.g., law enforcement on patrol, taxi drivers). These vehicles are not classified as working motor vehicles. A classification of in-transport or not in-transport is about identification of vehicles serving a transportation function for crash data collection and statistical reporting purposes.

Another key distinction is that transport crashes, as events associated with transportation, exclude incidents that directly result from a cataclysm. A cataclysm, as defined in ANSI D16.1, includes an avalanche, landslide/mudslide, hurricane, cyclone, downburst, flood, torrential rain, cloudburst, lightning, tornado, tidal wave, earthquake, volcanic eruption, damage from very large hail, and any wind above the minimum speed associated with a category one hurricane (ATSIP, 2017, p. 21). The exclusion notes that the timing must be such that the cataclysm is occurring at the time of the crash. For example, if rising floodwaters sweep a vehicle from the road, the incident would not be a transport crash, but rather the result of a cataclysm. However, if after the flood waters had receded, a vehicle lost control due to a portion of the roadway being damaged during the prior flood, there would be a transport crash and a reportable motor vehicle crash.

Lastly, although other land vehicles can be involved, if any aircraft are involved it is classified as an aircraft accident. The same would be true, if any watercraft were involved; absent any aircraft involvement as aircraft accidents are given higher classification priority. For example, if an airplane or helicopter crashes into motor vehicles on a freeway, the incident is classified as an aircraft accident even though it involved motor vehicles in-transport.

Additionally, because the most likely interaction is between motor vehicles and railway trains, both land vehicles, there is special consideration given to these incidents. If there is any harmful event involving the railway train prior to the involvement of the motor vehicle in-transport, then the incident is classified as a railway accident and not a reportable motor vehicle crash. For example, if a cargo train experiences a derailment and cargo from the overturned rail cars falls on motor vehicles traveling on the adjacent highway, there is no motor vehicle crash. It is a railway accident because of the initial harmful event involving only the railway train. However, if a railway train and an in-transport motor vehicle collide at a railway crossing, there would be a reportable motor vehicle crash.

What is a Motor Vehicle Traffic Crash?

The significant impact crash severity has on whether a crash is formally reported, the entities to which it is reported, the data collected, and the downstream use of the data is mentioned in the earlier discussion of harmful events in crashes. Another very significant factor in these same areas is the classification of the motor vehicle crash based on the location where it occurs. The key term associated with location classification is with respect to involvement of a trafficway. This important component is used to separate motor vehicle crashes into two classes, motor vehicle traffic crashes and motor vehicle nontraffic crashes. The interest in this separation is identification of motor vehicle being used for transportation within the transportation infrastructure from those that are not. For example, if a citizen backing in their private driveway impacts a pedestrian in the driveway, it would be classified as a motor vehicle nontraffic crash. While there is interest in these crashes for safety research (e.g., effectiveness of back-up cameras or pedestrian detection systems), this example does not involve the shared transportation infrastructure. Additionally, given that ANSI D16.1 is a Manual on Classification of Motor Vehicle Traffic Crashes and that motor vehicle traffic crashes are predominantly what is used in research, the discussion of motor vehicle crash reportability would be incomplete without addressing this concept.

As defined in ANSI D16.1, “A traffic crash is a road vehicle crash in which (1) the unstabilized situation originates on a trafficway or (2) a harmful event occurs on a trafficway,” (ATSIP, 2017, p. 24). In classification, motor vehicle crashes and traffic crashes are both transport crashes but note that traffic crashes do not require a motor vehicle be involved. For example, if a bicyclist riding on a roadway falls and is injured after being struck by the opened door of a motor vehicle parked in a parking lane, the crash would be classified as a traffic crash. It is not a motor vehicle crash because there was no motor vehicle in-transport involved. The bicycle is a road vehicle, but not a motor vehicle. The parked motor vehicle is not in-transport. So, although both transport crashes, traffic crashes are separate from motor vehicle crashes based on the involvement of a motor vehicle in-transport. Many countries and some US states collect and report data on traffic crashes including those that do not involve a motor vehicle. For example, Sweden includes in its official statistics single bicycle crashes, if they happen on public roads allowing motor vehicles. Understanding this classification terminology is imperative to compare data between various reporting entities.

Narrowing the classification, motor vehicle crashes differ from motor vehicle traffic crashes in that the latter involve a trafficway. As defined in ANSI D16.1, “A trafficway is any land way open to the public as a matter of right or custom for moving persons or property from one place to another,” (ATSIP, 2017, p. 3). The *Glossary for Transport Statistics* has a similar definition in its term “Road” which is defined as “Line of communication (traveled way) open to public traffic, primarily for the use of road motor vehicles, using a stabilized base other than rails or air strips,” (IWG, 2009, p. 39). Note that neither definition makes mention of ownership of the trafficway. That is, there is no distinction made between public ownership versus private ownership in these terms. Just that it is a land way open to the public for transportation. Consequently, in addition to government or publicly owned and maintained trafficways, privately owned land ways open to the public for transportation are trafficways.

The most common example would be a parking lot way. Parking lot ways most frequently occur in shopping, business, or apartment home complexes where the land ways within these locations are privately owned but fully accessible to the public for transportation to the stores, businesses, or apartment homes. A parking lot way is defined in ANSI D16.1 as a “Land way which is used primarily for vehicular circulation within parking lots and for vehicular access to parking lot aisles. Parking lot ways in parking lots open to the public are trafficways,” (ATSIP, 2017, p. 29). The impact of this is that motor vehicle crashes that occur on a parking lot way are motor vehicle traffic crashes. This is true even given that most parking lot ways are owned and maintained by whatever private entity owns the land on which the parking lot way is built.

An important use for motor vehicle traffic crash data is to plan and evaluate investment by governments in the transportation infrastructure. To address this need, motor vehicle traffic crashes are typically identified in data collection as occurring on public or private property so that the ownership and maintenance responsibility of the trafficway can be identified. The interest here is to be able to separate from the body of motor vehicle traffic crashes those that involve government or publicly owned and maintained transportation infrastructure and thus require public funds.

As an example, the National Highway Traffic Safety Administration (NHTSA), the agency within the US Department of Transportation (USDOT) that is tasked with reducing deaths, injuries, and economic losses from motor vehicle crashes, separates its data collection so the distinction can be made between motor vehicle nontraffic and motor vehicle traffic crashes. Further, within the motor vehicle traffic crash data, there is data available to separate publicly from privately owned trafficways to address the interest in publicly funded transportation infrastructure. Specifically, NHTSA’s Fatality Analysis Reporting System (FARS) that has been in existence since the 1975 is a nationwide annual census of motor vehicle traffic crashes within the 50 States, the District of Columbia, and Puerto Rico. Provided within the FARS dataset of fatal motor vehicle traffic crashes is an element called Ownership. The 2019 FARS/CRSS Coding and Validation Manual identifies that “this element identifies the entity that has legal ownership of the segment of the trafficway on which the crash occurred” (NHTSA, 2019, p. 81), thus this data element provides the granularity to separate motor vehicle traffic crashes that occur on publicly owned trafficways from those that occur on privately owned property. Additionally, because NHTSA’s responsibility is not just motor vehicle traffic crashes, it maintains another database known as the Not-in-Traffic Surveillance System (NiTS). NiTS purpose is to collect and provide counts and details on fatalities and injuries that occur in nontraffic crashes and in noncrash incidents.

Biography



John McDonough is the President of the National Institute for Safety Research. He has been the principal instructor for the National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis Reporting System (FARS) and Crash Report Sampling System (CRSS) since 2007. In addition to his work with NHTSA, Mr. McDonough has served as a subject matter expert and trainer for the Federal Motor Carrier Safety Administration's (FMCSA) Data Quality Improvement Program training law enforcement in large truck and bus crash data collection. He has also assisted numerous states in crash report redesign, training law enforcement in crash classification, and by writing crash report instruction manual and training class content. Mr. McDonough has significant experience with traffic safety data standardization to include being a member of the ANSI D16.1 Manual on the Classification of Motor Vehicle Traffic Crashes 7th and 8th Edition Workgroup (2007) and Consensus Body (2017), serving on the Model Minimum Uniform Crash Criteria (MMUCC) Expert Panel (2008, 2012, 2017), and as a team member of the USDOT Data Standardization/NHTSA Data Compatibility Workgroup (2005–10).

See Also

The value of life and health; Crash not accident; Aggressive driving and road rage; Animal crashes; Automatic vehicles and connected vehicles; Connected Automated Vehicles: Technologies, Developments and Trends; Bicycle collision avoidance systems; Bridges, Traffic Safety; Cost of accidents and injury; EPIDEMIOLOGY OF ROAD TRAFFIC CRASHES; In-depth crash analysis and accident investigations; Safety data quality management

Relevant Websites

National Highway Traffic Safety Administration, National Center for Statistics and Analysis. Available from: <https://www.nhtsa.gov/research-data/national-center-statistics-and-analysis-ncsa>

NHTSA Traffic Safety Facts Annual Report Tables. Available from: <https://cdan.nhtsa.gov/tsftables/tsfar.htm>

Centers for Disease Control and Prevention. Available from: <https://www.cdc.gov/motorvehiclesafety/>

AAA Accident Reporting. Available from: <https://drivinglaws.aaa.com/tag/accident-reporting/>

Inland Transport Committee (ITC). Available from: <https://www.unece.org/trans/main/itc/itc.html>

References

Association of Transportation Safety Information Professionals (ATSIP), 2017. ANSI D16.1-2017 American National Standard, Manual on Classification of Motor Vehicle Traffic Crashes, Eighth ed.

Intersecretariat Working Group (IWG), 2009. Glossary for Transport Statistics, fourth ed.

National Highway Traffic Safety Administration (NHTSA), 2017. Model Minimum Uniform Crash Criteria, 5th ed.

National Highway Traffic Safety Administration (NHTSA), 2019, 2018 FARS/CRSS Coding and Validation Manual.

United Nations Economic Commission for Europe (UNECE), 2017. Statistics of Road Traffic Accidents in Europe and North America, Vol. LIV.

Nominal Safety

Per Erik Garder, University of Maine, Orono, ME, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	386
Types of Safety	386
Nominal Safety—Conformance to Guidelines	387
Consequences of Nominal Safety Deviating from Objective Safety	389
Discussion	390
Acknowledgment	390
Biography	391
References	391
Further Reading	391

Introduction

This article focuses on traffic safety in the roadway system but similar discussions could be targeted toward rail, airline, shipping, and other transportation sectors. In addition, personal safety in the transportation system as well as security issues relating to extreme weather events, natural disasters or terrorism could be covered by similar analogies and comparisons.

Safety can be described and measured in many ways. We often hear expressions used by the public as well as by politicians such as, “Safety is paramount,” or “It has to be completely safe.” Sometimes people do not want to see a change from what they are used to, because they feel any alternations from their daily routine may jeopardize their safety and make it less safe. For that reason, it is often assumed that our current system is “safe,” However, our current roadway system is far from safe and will remain unsafe as long as we see no or very few serious injuries and fatalities. That is definitely true if we look at safety from society’s perspective, or compare it with regards to dangers we are exposed to in most other daily activities.

One way of improving traffic safety in our roadway system is to make sure that all its components—roadways, vehicles and operators—with respect to design and operations, meet all standards and guidelines. This is often referred to as Nominal Safety. If meeting such standards and guidelines guarantee absolute safety, then having high Nominal Safety would guarantee high if not full safety. Before dwelling further into Nominal Safety, let us look at a few alternate definitions of safety.

Types of Safety

Safety and security of a transportation system can be measured and evaluated in many ways. The main ways are often classified as:

- Subjective or perceived safety
- Objective or substantive or actual safety, that is, “count” of crashes by severity
- Theoretical safety
- Nominal safety = conformance to warrants/guidelines/standards

First, let us look briefly at subjective safety. This is elaborated in the article “Risk Perception and Risk Behavior in the Context of Transportation.” But let us discuss it here as well.

Typically, subjective safety refers to an ordinary road-user’s perception of safety when being involved in an activity, such as driving a car, walking or traveling by an airplane. Subjective safety can, however, also refer to an “expert’s” opinion—that is, the opinion of a traffic engineer, police officer, motor journalist, or other professional—regarding the safety of, for example, a location or a roadway condition. When an engineer makes the assessment, it is sometimes called engineering judgment. This can be done as an estimate of its expected number of accidents but typically is done in some other way. One such other way is to evaluate whether the entity meets guidelines and standards or not. Other measures of subjective safety, as assessed by experts, may be frequency of road-user violations or number of near misses or traffic conflicts. These surrogate measures may have a direct relationship to objective safety. However, subjective safety is from here on in this article defined as the perception of ordinary road-users. This safety is by itself of interest. If a location is perceived of as unsafe, drivers may avoid it and that may cause them to use routes that are less desirable from a community’s perspective; or it may cause them to use a different mode of transportation or canceling their trip altogether. And, even if they do not change mode or avoid the location, it may cause stress and fear. For example, people who dread a dangerous location (a bridge, a high-speed traffic circle, or a narrow road), may believe there are no alternatives, so they cannot avoid the location. Will this affect the drivers’ health? Will it affect their demeanor? Will it affect the way they will interact with other people later that day? Their level of anger? Their driving once they have exited that location? The answer to all those questions may be: Yes to

some extent but probably not drastically. However, anything can, of course, be that last straw which breaks the camel's back. On the other hand, drivers who perceive a location to be unsafe are typically more careful at that site. This may lead to fewer crashes than if people view the location as safe. Some safety advocates therefore argue that engineers should try to design locations so that they are safe but look or feel unsafe. A final question: Is there a connection between perceived safety and whether a location meets nominal safety standards or not? We will not try to answer that question here. However, if most locations meet nominal safety standards, people may expect that to be the case everywhere.

Objective or substantive or actual safety refers to the final result on safety as measured through a "count" of crashes or accidents. This count is not the random outcome that a location has experienced, or will experience, but the long-term expected number of crashes if traffic volumes and all other parameters could be held constant. It would make little sense to state that the safety of a location has deteriorated by 50% from 2017 to 2018 because the number of crashes went from two to three. Also, it is not only just the number of crashes that affects whether a location has high or low actual safety, but also it is the severity of those crashes. Furthermore, exposure also matters. It makes little sense to say that riding a motorcycle has higher objective safety than riding in a car just because fewer people are injured or killed on motorcycles than in cars. For example, the US National Highway Traffic Safety Administration ([NHTSA](#)) reports that 5172 motorcyclists were killed in 2017 in the United States whereas 23,708 passenger-vehicle occupants were killed. Motorcycle crashes account for about 14% of all traffic fatalities in the United States, but motorcycle traffic represents only 0.6% of vehicle miles traveled on America's roads. So, the risk per mile is much higher for motorcyclists than for automobile occupants ([FHWA, 2019](#)). On the other hand, passenger car travel can be seen as the greater threat to public health since more people die that way.

With theoretical effect on safety we typically mean the effect a countermeasure ought to have if people behaved in a "theoretical" way. For example, if we have a factory located just off a major road and there are 50 vehicles turning off to that factory every day causing, on average, one crash per year, we could expect ten crashes per year if the number of vehicles turning off was increased tenfold to 500 a day, as long as we did not change the geometric layout or anything else. Will that be the expected result? Not necessarily. Higher turning volumes may lead to an increased awareness among drivers who go through that location on a daily basis. That, in turn, may lead to less of an increase in the number of crashes than we "theoretically" calculated. Several studies have shown that risk compensation is real (even if risk homeostasis is not), that is, subjective risk perception influences a person's behavior and, therefore, the actual safety effect cannot easily be calculated with simple mathematical formulas.

The final "type" of safety, nominal safety, is the focus of this article and will be covered next.

Nominal Safety—Conformance to Guidelines

Nominal safety is evaluated by looking at compliance with standards, warrants, guidelines, and sanctioned design procedures. Standards and guidelines vary from country to country and even within countries. So, what is nominally safe in one country may not be safe in another one. Also, state standards may not be the same as municipal standards, and municipal standards may not apply to private roads. Let us first look at some of the more known guidelines for roadways.

For road traffic in the United States, the American Association of State Highway and Transportation Officials (AASHTO) Green Book, *A policy on geometric design of highways and streets*, give geometric standards that are to be followed for major roads, the so-called National Highway System (NHS) ([AASHTO 2018](#)). These standards can also be followed on lower-priority roads but the road administrator can set their own geometric standards for those roads. Title 23 of U.S. Code §109 states, "projects (other than highway projects on the NHS) shall be designed, constructed, operated, and maintained in accordance with State laws, regulations, directives, safety standards, design standards, and construction standards." The AASHTO Green Book is updated every few years and the latest edition, the seventh, was published in September 2018 and adopted by the Federal Highway Administration (FHWA) on November 1, 2018, as stated in Federal Register/Vol. 83, No. 212/Rules and Regulations 23 CFR Part 625: "This final rule updates the regulations governing design standards and standard specifications that apply to new construction, reconstruction, resurfacing (except for maintenance resurfacing), restoration, and rehabilitation projects on the National Highway System (NHS). In issuing this final rule, FHWA incorporates by reference the latest versions of design standards and standard specifications previously adopted and incorporated by reference, and removes the corresponding outdated or superseded versions of these standards and specifications. Use of the updated standards is required for all NHS projects authorized to proceed with design activities on or after the effective date of the final rule. This final rule is effective December 3, 2018." In summary, the State highway departments, working through AASHTO develop design standards through a series of committees and task forces. FHWA contributes to the development of the design standards through membership on these working units, sponsoring and participating in research efforts, and many other initiatives. Previous AASHTO Green Books were published in 1984, 1990, 1994, 2001, 2004, and 2011. Prior to the Green Book, AASHTO or its predecessor, AASHO, published, for example, a Policy on Geometric Design of Rural Highways in 1954 and 1965; and a Policy on Design of Urban Highways and Arterial Streets in 1973. AASHTO also publishes standards and specifications for construction materials, for bridges, and other roadway transportation areas.

Other countries have similar national guidelines for geometric design. For example, Sweden has *Vägars och gators utformning* (VGU) (design of roads and streets), which is published jointly by [Trafikverket](#) (the Swedish Transport Administration) and an organization with members from the Swedish municipalities and counties. The current edition is from 2015 and replaced the 2012

edition. The design recommendations sometimes coincide with those in the AASHTO Green Book but often do not. The Swedish guidelines are considered rules and need to be followed on State roads, which currently, in 2020, make up around 98,500 km. For municipal roads, currently around 42,300 km, and private roads getting State support, around 74,000 km, VGU standards are voluntary and seen as advisory only.

In the United Kingdom (UK), there is the Design Manual for Roads and Bridges (DMRB), which is a series of 15 volumes that provide standards for the design, assessment and operation of trunk roads in the UK, and, with some amendments, the Republic of Ireland. It also forms the basis of the road design standards used in many other countries. Volume 6 gives Road Geometry and other volumes cover Highway Structures, Geotechnics and Drainage, Pavement Design and Maintenance, Traffic Signs and Lighting, Traffic Control and Communications, Environmental Design and Economic Assessment.

All guidelines for road design and operation in Germany are published by the Forschungs gesellschaft für Straßen- und Verkehrswesen (FGSV) (Road and Transportation Research Association), a non-profit organization similar to the Transportation Research Board (TRB) in the US. Guidelines are developed by technical committees, with members from state and federal governmental organizations, private industry, consultants, and universities. Guidelines are subject to continuous change and development. The first guidelines were issued after the first World Road Congress in 1908, with an emphasis put mainly on the geometry of single vehicles (motor vehicles, long horse carts). Then requirements for moving motor vehicles, with respect to driving dynamics, were added in the 1920s and 1930s. Nowadays, vehicles interacting with each other are the focus of design considerations. At all times, traffic safety has been the focus but in later years, environmental issues, and ecological sustainability, are also priorities. In the last decades, new principles for road design guidelines are being implemented. Although driving kinematics still have to be observed, principles like readability or self-explaining properties of the road are receiving higher importance. This includes trying to better understand how human factors and psycho-physiological properties of drivers should be considered in design in a systematic way (Lippold et al., 2015).

With respect to highway signs and markings, the United States has the Manual on Uniform Traffic Control Devices for Streets and Highways (MUTCD). The MUTCD has been administered by the FHWA since 1971, and is a compilation of national standards for all traffic control devices, including road markings, highway signs, and traffic signals. It is updated periodically to accommodate the nation's changing transportation needs and address new safety technologies, traffic control tools, and traffic management techniques. On December 16, 2009 a final rule adopting the 2009 Edition of the MUTCD was published in the Federal Register with an effective date of January 15, 2010. It was stated that states must adopt the 2009 National MUTCD as their legal State standard for traffic control devices within two years from the effective date. Therefore, by January 15, 2012, States were required to have adopted the National Manual or have a State MUTCD/supplement that is in substantial conformance with the National Manual. The 52 FHWA Division Offices (50 States plus Puerto Rico and Washington D.C.) reviews any State MUTCD or supplement to determine if it is in substantial conformance with the National MUTCD. As of January 2020, about 16 states have adopted the national manual, 24 have adopted the national manual along with State supplements and 10 have adopted their own State MUTCD (FHWA, 2020). The content of State MUTCDs is similar to the national one. For example, all States use double yellow centerlines to indicate that passing is a bad idea, but in some states, the striping means that it is illegal to cross the centerline to pass any vehicle while in other states it is legal to pass slow-moving vehicles and in some states, it is an advisory rule only. In some other countries, double centerline markings mean that it is illegal to cross the centerline for any purpose, such as to turn left into a driveway. So, the exact meaning of a markings varies from jurisdiction to jurisdiction and meeting nominal safety—having “correct” striping—may therefore have different implications on objective safety.

It should also be pointed out that the AASHTO Green Book typically uses “should” and sometimes “may”. The word “shall” or “must” are seldom used, so there is some flexibility in the policy. On the other hand, the current MUTCD uses the word “should” 2627 times, “shall” 3046 times, “may” 1788 times, and “must” 95 times.

Other countries have similar manuals as the MUTCD but few if any roadway signs are universal. The stop sign is the only real exception. However, most countries outside North America have signs conforming to the Vienna Convention. All European countries, except Ireland, and most countries in Asia and some in Africa and South America have adopted the Vienna Convention on Road Traffic, which is an international treaty designed to facilitate international road traffic and to increase road safety by establishing standard traffic rules among the contracting parties. The convention was agreed upon at the United Nations Economic and Social Council's Conference on Road Traffic in 1968 and came into force in May 1977. It is currently ratified by 78 countries. The convention part that contains rules on road signs and signals was amended in 1993 and 2006.

So, if different countries have different standards, why is that? Are nominal safety standards not based on thorough analysis of what is safe? Sometimes that is not the case. Hauer (1997) writes: “The field of road safety abounds with strongly held but unfounded opinions. Thus, most road users believe that traffic signals enhance safety; they go to their gut feeling. Most traffic engineers believe that four-way stops enhance safety only where warranted by the Manual of Uniform Traffic Control Devices; they go by professional folklore. In road safety, gut feeling and folklore are frequently wrong.” The latter was evaluated in a study of the effect of all-way stop as a function of percent minor volume. They found that there was very little if any correlation between meeting the guideline and objective safety (Simpson and Hummer, 2010).

Consequences of Nominal Safety Deviating from Objective Safety

When analyzing the safety of an intersection, segment of roadway, or other transportation facility, an engineer might encounter that the entity meets or does not meet one or both of the two safety types nominal safety and objective safety. Whether it meets subjective safety will not be discussed here.

Whether an entity meets nominal safety or not, can typically be evaluated in a fairly straightforward way by looking at whether standards are met or not. For example, are sight distances sufficiently long, is grade below allowed maximum value, and are radii above the minimum required? Even if the guidelines say “should” rather than “shall” it is clear that nominal safety has not been met unless the standard is met.

Whether an entity meets objective or substantive safety is somewhat harder to evaluate since we in road traffic do not have clear guidelines to what acceptable risk should be. That is discussed in a separate article written by Claes Tingvall [to be cross-referenced to article 10099: Acceptable risk by Claes Tingvall]. If we follow Vision Zero ideas, the focus of safety should be on avoiding fatalities and serious injuries. However, what is then acceptable? For the European Union (EU) rail system, it is stated that “for technical systems where a functional failure has credible direct potential for a catastrophic consequence, the associated risk does not have to be reduced further if the rate of that failure is less than or equal to 10^{-9} per operating hour (EU Commission, 2009). If we interpret this to convey that we should accept no more than one fatality per 10^9 h of travel, how many fatalities per year would we then accept in rail traffic in the EU? Rather than answer that, let us convert acceptable risk to fatalities per distance, especially since people seldom want to travel a certain time but rather a certain distance, to get to their destination. It is difficult to find estimates of average train velocities, and it can be argued that stopped time should or should not be included. Let us here assume that trains move at an average speed of 80 km/h (50 mph). That means that we would accept up to one fatality per 80 billion kilometers of travel or 0.0125 fatalities per billion kilometers traveled. In 2018, rail passenger transport in the EU was estimated at 472 billion passenger-kilometers (Eurostat, 2019). That means that we would accept up to six fatalities per year among train passengers in order to state that they are objectively safe. This may be a considerably tougher standard than intended since a failure resulting in catastrophic consequences certainly could kill more than one person per incident.

In EU-28, a total of 219 rail passengers were killed in the 5-year period 2012–16 (European Union Agency for Railway 2018), giving an average death toll of 44 per year. That includes passengers killed in collisions and derailments as well as falling on board or when alighting or boarding trains, etc. excluding only confirmed suicides. For example, in 2015, only eight people were killed in collisions as passengers of trains and none were killed in a derailment, whereas a total of 27 passengers died that year according to the numbers cited above. Still, if we did set a goal of having no more than six people per year being killed, it was not met for most years. Overall, with 44 fatalities per year, EU-28 saw around 0.09 fatalities per billion passenger km. (Besides numerous victims at grade crossings as accidents and suicides, which we will not discuss here.)

If we look at commercial airlines in the United States, to get another estimate of actual fatality risk, and include all crashes (and other accidents) since January 2002, we have had 130 people get killed on US airlines (with no one killed between February 2009 and the writing of this article in early 2020) and another three people killed traveling by a foreign airline (Asiana Airlines, crash in 2013) in the US, for a total of 133 fatalities in the US. So, from 2002 to the beginning of 2020, 18 years, we saw on average 7.4 fatalities per year. And, if using 2018 U.S. passengers travel mileage, 730.7 billion passenger-miles by plane, which is 1176 billion passenger-km, as exposure, we get an estimate of passenger flight distance. Travel volumes have increased over this time and the average travel mileage may be around 1000 billion km/year. That means we saw around 0.007 fatalities per billion km which is clearly below the desired 0.0125 fatalities per passenger-km as discussed for trains above. And, if we look at the latest 10 years only, with three fatalities in total, we would be at around 0.0003 fatalities per billion person km, which is a tiny fraction of the desired 0.0125. So, we could deem US air travel to be objectively safe.

On the other hand, if we look at automobile travel, did motor vehicles keep its occupants “safe,” meaning below 0.0125 fatalities per billion km? No, let’s look at the US, using 2017 data. The US had an official VMT of 3,212,670 million miles in 2017 according to FHWA. With an average of 1.3 occupants per vehicle, that translates to 4,176,470 million people (or passenger) miles or 6,719,940 million people-km. And, if we accept 0.0125 fatalities per billion km of travel, we would accept 84 automobile-related fatalities in the US in 2017. In reality, the US saw over 37,000 fatalities in 2017 and 23,551 of those were occupants of the motor vehicles according to NHTSA. That is 280 times what we could have as an objective required safety level, a safety level which is met by commercial airlines in the US and almost met by trains in Europe. European roads do quite a bit better than US roads, but are still far riskier than the desired level as outlined here. The EU has seen right around 12,000 occupant fatalities per year in later years (2013–16). Finally, it should be noted that there are people who travel by car in Europe as well as in the US who think US airliners seem too dangerous and do not want to fly, but we do not discuss subjective safety here.

An alternative definition of objective safety, when looking at a specific location is to say that it is safe if there has not been a single crash there in the last three or five years. However, that may sometimes be too strict, unless we have a true Vision Zero approach; and sometimes too liberal, since a location with very low traffic volumes may be objectively unsafe even if there has not been a single reported crash for many years. Note that having no crash in 3 years is not surprising if the expected number of crashes per year is very low, for example, 0.1. But imagine that this is an intersection between two driveways that serve one-family homes and it is traveled by only a few people per day. One crash every 10 years would obviously not be considered safe.

Another completely different way to define minimum required objective safety is to say that risk, as measured in crashes per some type of exposure, should not be higher than a certain number of standard deviations above the state average for such a facility. This reasoning builds on the belief that people would not venture out into traffic if safety, on average, was not acceptable, and since people do travel by car (and other modes that have even higher risks), they “obviously” think that today’s average risks are okay, but they may not find a specific location safe if it is significantly more dangerous than an average one. This reasoning makes little sense. Why would Americans be objectively safe driving through an intersection that is more likely to kill them than one in, say, the UK, that engineers in the UK find to be objectively unsafe since it is significantly more dangerous than the average intersection in the UK? Still, this is how high-crash locations are selected in many if not most countries.

If we use the evaluation of objective safety which is typically used around the world, for high-crash locations, as described in the previous paragraph, we can distinguish between four possibilities:

- Case A: nominal safety—meeting standards, objective safety—meeting standards
- Case B: nominal safety—not meeting standards, objective safety—meeting standards
- Case C: nominal safety—meeting standards, objective safety—not meeting standards
- Case D: nominal safety—not meeting standards, objective safety—not meeting standards

Case A is obviously what we desire.

In Case B, the roadway geometrics do not conform or meet the standard-based analysis. But if it still is objectively safe, it will not be a safety concern for society. However, it may be a liability concern for engineers and others when crashes do occur, even if there are very few crashes per million entering vehicles. Either the location needs to be reconstructed so that it meets the standards, or the standards should be changed. The latter is the case if we can identify many locations with similar non-conforming design that on average function with a high degree of safety.

In Case C, the roadway meets all standards and one could by mistake consider this to be a safe roadway. However, road users obviously misinterpret something or behave in unintended ways, and redesign—or enforcement or education—is needed more or less immediately.

In Case D, the roadway geometrics do not conform or meet a standards-based analysis and objective safety analysis shows that there are many crashes. It is obvious that these locations should be reconstructed, or changed in some other way, immediately. However, money is not always available and temporary measures may have to be taken, such as drastic reduction of the speed limit combined with police enforcement of the speed until reconstruction can occur.

We could add two more cases to this analysis. That is when objective safety is unknown because the location was just built or reconstructed and similar sites cannot easily be used for estimating objective safety. If that is the case, an evaluation of nominal safety may be all we can do. If nominal safety is met, we may be able to let the location be unevaluated with respect to objective safety until we have data. If nominal safety is not met, we may need to do conflict studies or evaluate objective safety in some other way to assess what is immediately required.

Discussion

Nominal safety means that standards are followed. This typically means that drivers know what to expect with respect to horizontal curve radii, sight distances, roadway widths, striping, signage, etc. Understanding rules and being able to predict what geometric characteristics will follow a hill crest or horizontal curve should help safety. Being surprised or not understanding what is expected of a driver ought to lead to lower safety. But why then does the US, which has stricter and probably better developed standards than any other country, have higher fatality rates—lower objective safety—than almost all other industrialized countries if safety is measured as fatalities per capita? And, even if measuring fatality rates as fatalities per million kilometers traveled (or driven), the US has higher rates than many other countries. For example, in 2019, Sweden had 2.62 fatalities per million vehicle km (0.42 fatalities per hundred million vehicle miles (HMVM)) whereas the US had a fatality rate that was between two and three times that. Obviously, having high nominal standards, and typically meeting them, does not guarantee objective safety.

Acknowledgment

I would like to acknowledge Dr. Eric H. Shimizu, PE, PTOE, Principal Engineer and Regional Director for Smart Cities and Connected Vehicles at DKS Associates, Seattle, Washington, for developing a first outline of an article on Nominal Safety.

Biography



Per Erik Garder is a professor of civil engineering at the University of Maine, USA, since 1992. He has a PhD from Lund University in Sweden and he was a faculty member of the Royal Institute of Technology (KTH) in Stockholm, Sweden (1983–92). His research interest is focused on forecasting, designing and evaluating facilities with emphasis on traffic safety.

References

- AASHTO, 2018. A Policy on Geometric Design of Highways and Streets, 7th Edition (referred to as Green Book). American Association of State Highway and Transportation Officials, Washington DC.
- EU Commission, 2009. Commission Regulation (EC) No 352/2009 of 24 April 2009 on the adoption of a common safety method on risk evaluation and assessment as referred to in Article 6(3)(a) of Directive 2004/49/EC of the European Parliament and of the Council. Official Journal of the European Union 29.4.2009 Page L 108/4, Accessed February 12, 2020 at https://ec.europa.eu/transport/sites/transport/files/celex_32009r0352_en_bxt_0.pdf.
- Eurostat, 2019. Railway passenger transport statistics - quarterly and annual data. Accessed on February 12, 2020 at https://ec.europa.eu/eurostat/statistics-explained/index.php/Railway_passenger_transport_statistics_-_quarterly_and_annual_data.
- FHWA, 2019. US Department of Transportation Federal Highway Administration Website: Motorcycle Safety in Research. Updated: Monday, December 2, 2019. Accessed January 26, 2020 from <https://cms7.fhwa.dot.gov/research/research-programs/safety/motorcycle>
- FHWA, 2020. US Department of Transportation Federal Highway Administration Website: MUTCDs & Traffic Control Devices Information by State. Accessed February 24, 2020 at https://mutcd.fhwa.dot.gov/resources/state_info/index.htm
- Hauer, E., 1997. Observational Before-After Studies in Road Safety. Pergamon Press, copyright Elsevier.
- Lippold, C., Lemke, K., Jährig, T., Stöckert, R., 2015. Country report Germany: The new Generation of Design Guidelines for Roads and Motorways in Germany, 5th International Symposium on Highway Geometric Design, Vancouver 2015.
- European Union Agency for Railway, 2018. Report on Railway Safety and Interoperability in the EU 2018 Accessed February 12, 2020 at https://www.era.europa.eu/sites/default/files/library/docs/safety_interoperability_progress_reports/railway_safety_and_interoperability_in_eu_2018_en.pdf.
- NHTSA, 2018. Traffic Safety Facts 2017 data Accessed on February 12, 2020 at <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812691>.
- Simpson, Carrie L., Hummer, Joseph E., 2010. Evaluation of the conversion of stop sign control to all-way stop sign at 53 locations in North Carolina. J. Transport. Saf. Security 2 (3).
- Trafikverket, 2015. Krav för Vägars och gators utformning. Trafikverkets publikation 2015:086 Available from: https://trafikverket.ineko.se/Files/sv-SE/12046/RelatedFiles/2015_086_krav_for_vagars_och_gators_utformning.pdf.

Further Reading

- Polanis, S.F., 1995. Some thoughts about traffic accidents, traffic safety and the safety management system. ITE J. 65 (10).

Parking Lots

Maxim A. Dulebenets, Department of Civil & Environmental Engineering, Florida A&M University-Florida State University (FAMU-FSU) College of Engineering, Tallahassee, FL, United States

© 2021 Elsevier Ltd. All rights reserved.

Background	392
Parking Maneuvers	393
Existing Parking Challenges	394
Vehicle Parking Policies and Regulations	394
Parking Space Cruising Impacts	395
Vehicle Parking and Cyclists	395
Vehicle Parking and Pedestrians	396
Vehicle Parking and Older Adults	396
Automation Impacts on Vehicle Parking	397
Concluding Remarks	398
Biography	399
References	399
Further Reading	399

Background

The demand for passenger and freight transportation has substantially increased over the last years (Mittal et al., 2017; Dulebenets, 2019). Such a trend can be explicated by different reasons, including increasing population, urbanization, and the existing needs to provide effective connections between different metropolitan areas. Passenger vehicles constitute the most popular mode of passenger travel in many countries. At least 80% of the hours of the day, vehicles generally spend in the parking mode (Marsden, 2006). In busy metropolitan areas that often experience congestion, parking space cruising (or parking search time) is considered as a significant part of the entire trip, except the cases when parking is reserved at the destination. Therefore, it is critical for local authorities to build a sufficient number of parking facilities with adequate capacity and in appropriate locations of a city to meet the growing demand and reduce the cruising time for parking space. Parking lots accommodate vehicles that arrive to their destinations for different purposes, including commercial, residential, industrial, shopping, entertainment, just to name a few. Many cities around the world use much of their valuable real-estate space in order to build new parking facilities, some of which have quite a substantial capacity. Table 1 shows a list of the largest parking lots in the world.

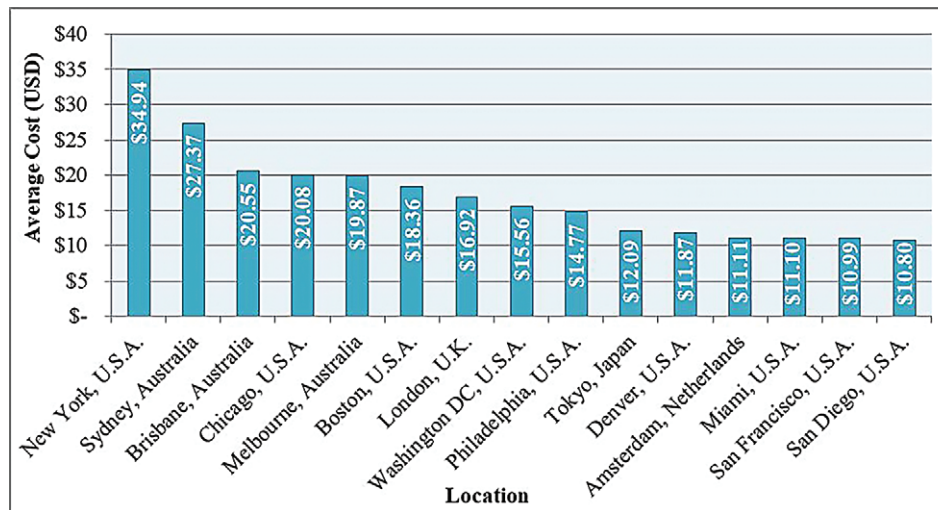
Note that the capacity of the parking lots that are listed in Table 1 may fluctuate from one year to another (e.g., additional parking spaces can be built to accommodate the demand). The West Edmonton Mall (Canada) has an enormous parking lot that is able to accommodate approximately 20,000 vehicles (DriveSpark, 2015). Some airports in the United States and Canada (such as Seattle-Tacoma International Airport, Detroit Airport, O'Hare International Airport, and Toronto Airport) also have high-capacity parking facilities, required for daily airport operations. All of these are single-purpose lots outside the city centers. In downtown areas, there are typically many smaller parking garages rather than one giant one, but even these smaller facilities can serve large numbers of customers. Due to significant parking demand in certain areas, local authorities or semi-private companies operating parking garages have to set high parking fees for on-street as well as off-street parking, especially in downtown areas of certain cities. New York (United States) is one of the cities that have prohibitively high fees for off-street parking. The average cost for 2-hour parking in New York reached approximately \$35 in 2019 – see Figure 1 (Parkopedia, 2019). Sydney (Australia), Brisbane (Australia), and Chicago (United States) are other examples of cities with parking fees that exceed \$20 for 2-hour parking in 2019.

By imposing high parking fees, local authorities are trying to encourage people to use alternative travel modes instead of passenger vehicles (e.g., use a subway to reach a destination instead of using a private vehicle and pay parking fees at the same destination). Otherwise, the existing parking capacity may not be sufficient to meet the demand, and local authorities would have to make sure that even more off-street parking facilities are provided. Building new parking facilities can be a challenge, especially in busy metropolitan areas, due to the cost and spacing constraints. Providing multistory underground parking spaces can be very expensive. The average cost of an underground parking space may vary from one city to another. Shoup (2014) reported that the average cost of an underground parking space in Honolulu (Hawaii, United States) may go up to ≈\$50,000, while the average cost of an underground parking space in San Francisco (California, United States) may go up to ≈\$40,000.

Along with meeting the growing demand for vehicle parking, local authorities as well as private businesses have to address some other issues that are associated with parking. Safety and security in parking lots are critical issues. Although travel speed is much lower in parking lots as compared to other facilities (e.g., interstates), accidents do happen in parking lots. In fact, thousands of pedestrians and cyclists are injured every year in parking lots around the world due to failure of drivers to detect them when

Table 1 The largest parking lots in the world

a/a	Parking lot	Location	Approximate capacity
1	West Edmonton Mall	Edmonton, Alberta, Canada	20,000
2	Seattle-Tacoma International Airport	Seattle-Tacoma, Washington, USA	13,000
3	Detroit Airport	Detroit, Michigan, USA	11,500
4	Disney World	Orlando, Florida, USA	11,000
5	Universal Studios	Orlando, Florida, USA	10,200
6	Disney Land	Anaheim, California, USA	10,000
7	O'Hare International Airport	Chicago, Illinois, USA	9,300
8	Toronto Airport	Toronto, Ontario, Canada	9,000
9	Baltimore Airport	Baltimore, Maryland, USA	8,400
10	Dallas Airport	Dallas, Texas, USA	8,100


Figure 1 The top cities with the most expensive 2-hour off-street parking cost in 2019.

performing a parking maneuver. In some cases, pedestrians and cyclists may be at fault and cause accidents in parking lots due to lack of attention. Collisions between pedestrians/cyclists and vehicles in parking lots may even lead to fatalities. Furthermore, according to the U.S. Bureau of Justice Statistics (2020), more than one out of ten of all the property crimes are recorded in off-street garages or parking lots. Aggravated assaults, rapes, and murder may occur in parking lots of certain neighborhoods. In some parking lots, vehicles can be stolen. This chapter focuses on safety and security issues in parking lots. In particular, different parking maneuvers will be reviewed along with their safety implications. Moreover, a number of common parking challenges that may lead to certain safety and security concerns as well as various approaches for improving safety and security in parking lots will be discussed. Finally, concluding remarks regarding vehicle parking safety and security are provided at the end.

Parking Maneuvers

Three types of maneuvers exist for drivers in order to enter a given parking space, including the following (Findley et al., 2020): (1) pull-in; (2) back-in; and (3) pull-through. On the other hand, the following maneuvers can be used to exit a given parking space: (1) pull-out; (2) back-out; and (3) pull-through. The aforementioned parking maneuvers are illustrated in Figure 2. Based on the statistical data, collected by the American Automobile Association (AAA) for the year of 2015, approximately 11.6% of drivers in the United States back-into a parking space all the time or most of the time, while another 12.8% of drivers back-into a parking space frequently (AAA, 2015). The remaining drivers (approximately 75.6%) indicated that they rarely use the back-in maneuver for parking. Furthermore, male drivers are more likely to use the back-in maneuver for parking as compared to female drivers. However, the back-in maneuver is viewed as the safest parking maneuver (AAA, 2015; Findley et al., 2020).

Based on the information provided by the National Highway Traffic Safety Administration (NHTSA, 2008), there are almost 300 fatalities in the United States every year as a result of a nonoccupant (e.g., cyclist or pedestrian) being struck by a reversing vehicle (i. e., “backover” accidents as classified by the NHTSA). Moreover, more than 18,000 injuries are reported due to backover accidents,

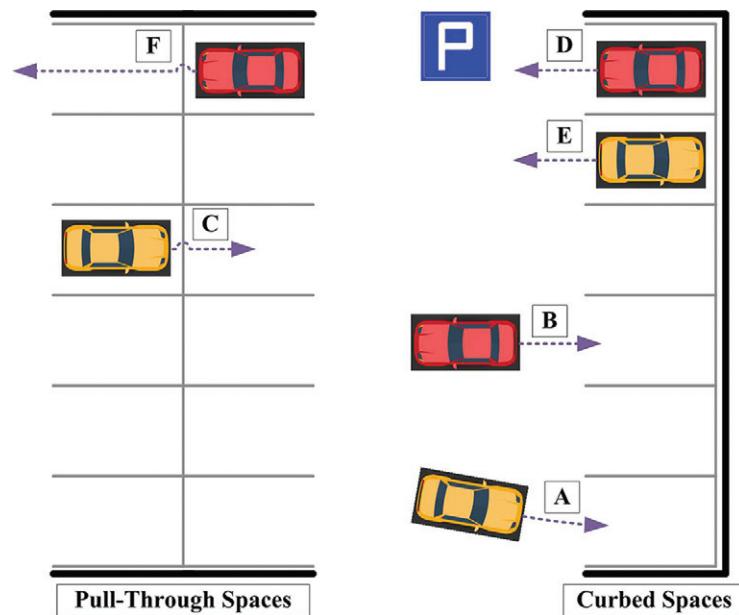


Figure 2 Parking maneuvers: [A] - pull-in; [B] - back-in; [C] - pull-through; [D] - pull-out; [E] - back-out; [F] - pull-through.

which generally occur in nonresidential parking lots. The back-in maneuver can help preventing the backover accidents. In order to execute the back-in maneuver, a driver is required to pass the desired parking space, position the vehicle accordingly, and then start backing into that parking space. While performing the aforementioned steps, the driver will have an opportunity to clearly see the desired parking space and make sure that there are no other objects and obstacles. On the other hand, the drivers performing the backing-out maneuver may have challenges with the identification of moving vehicles, objects, and obstacles in the driving aisle. Pulling out of a parking space typically allows the driver to have a clearer view of the driving aisle, especially when driving a vehicle that is longer in the back as compared to the front (e.g., trucks, SUVs).

Drivers have to be attentive not only when performing the parking maneuvers but also when entering or exiting parking lots. Parking lot entrances in many countries have crosswalks used by pedestrians, especially in downtown areas of the cities. In the United States and many other countries, there are curb cuts at parking lots and even at single-family homes (e.g., the driveway entrance to the garage or parking area has a curb cut). This means that drivers can turn into their driveways at relatively high speeds without damaging the alignments of their vehicles. In many European countries, driveways not only to single-family homes but also to commercial businesses have a raised through sidewalk across the driveway, rather than a crosswalk, and drivers entering the driveway have a steep asphalt ramp to get up onto the sidewalk. This slows down the vehicles entering and crossing the pedestrian facility, improving the safety of not just pedestrians but also of bicyclists riding along the street.

Existing Parking Challenges

There exist a number of challenges associated with vehicle parking that may lead to certain safety and security issues. The key challenges of vehicle parking will be further discussed in this chapter, including the following: (1) vehicle parking policies and regulations; (2) parking space cruising impacts; (3) vehicle parking and cyclists; (4) vehicle parking and pedestrians; (5) vehicle parking and older adults; and (6) automation impacts on vehicle parking.

Vehicle Parking Policies and Regulations

Marsden (2006) highlights that parking policies can be considered as the key link between land-use policies and transport. Parking policies have to support the major objectives of vehicle parking, including the following: (1) restraint; (2) regeneration; and (3) revenue. The aforementioned objectives can be viewed as conflicting in their nature. For example, the revenue, generated from vehicle parking facilities, contributes to the prosperity of cities. Increasing capacity of parking facilities can be used as a means of regenerating certain areas of cities. On the other hand, the attractiveness of city centers could be negatively affected with desert-like huge parking lots and lack of walkability, so parking facilities need to be kept attractive and have limited footprints on the ground level. However, the introduction of parking restraint measures may have even more negative effects on commerce. When it comes to parking restraint policies, drivers making shopping and leisure trips generally have more flexibility as compared to commuters, as they can adjust the frequency of visits and, ultimately, change their destination (Marsden, 2006).

The out-of-pocket costs (including the costs that are associated with parking fees) and walking time are recognized to be more important to drivers when comparing to in-vehicle costs. Many drivers are still willing to walk fairly long distances from free parking spaces to their respective destinations, despite the sensitivity of drivers to increasing walking time. In some cases, drivers may be willing to pay higher parking fees and select the parking spaces that are close to the destination when walking time becomes excessive. In many jurisdictions, including the United States that has the Americans with Disabilities Act (ADA) legislation, people with disabilities are often given free parking spaces that are located close to potential destinations. Furthermore, parking policies should be developed considering the fact that parking fees may actually prompt users switching from a private car to alternative modes of transportation (e.g., transit). Trip purpose, income level, travel distance, vehicle type, number of return trip stops, age, and gender are found to be significant factors that may influence mode choice under different parking fee options (Tsamboulas, 2001). Parking policies must be designed, considering local spatial and transport planning processes, to support strong and vibrant economy, ensure clean urban environment, provide good accessibility, enable safe and secure environment, and support more equitable society (Marsden, 2006). The safety issues should be adequately addressed by parking policing, taking into account high occupancy of parking spaces (especially in residential areas) that reduces the number of crossing points for pedestrians.

Parking Space Cruising Impacts

Parking space cruising may become a challenge, especially in large metropolitan areas. Off-street parking spaces (e.g., parking in dedicated garages) are often available to drivers immediately; however, they are not free and drivers are required to pay parking fees. On the other hand, curb parking can be either low-cost or completely free but cruising for such parking spaces may require significant time, as the number of available curb parking spaces is generally limited, especially during certain time periods (e.g., peak hours, weekends, particular events). Many drivers are not willing to pay for off-street parking spaces, if they can find free parking. If all the curb parking spaces are occupied by vehicles at a given moment, some drivers are willing to continue cruising until one of the occupied spaces is vacated. Typically, drivers choose cruising in case off-street parking is significantly more expensive than curb parking, fuel is cheap, parking duration is fairly long, vehicle occupancy is low, and the cost of time is fairly low (Shoup, 2006). However, cruising can be completely eliminated by imposing high parking fees for curb parking spaces (e.g., approximately equal to off-street parking fees).

Moreover, excessive cruising may result in negative externalities, including pollution of cities, increasing fuel consumption, traffic bottlenecks, and increasing accident rates. During peak hours in downtown areas of certain cities, a substantial portion of traffic is composed of drivers who are cruising for curb parking; thereby, causing congestion and increasing the risk of accidents. It can take up to 14 minutes to find a curb parking space in congested downtown areas (Shoup, 2006). Therefore, local authorities have to set the appropriate parking fees for curb parking to prevent excessive cruising of drivers that may further lead to safety concerns. Increasing curb-parking fees would prompt a certain portion of drivers using off-street parking facilities.

Vehicle Parking and Cyclists

The number of bike lanes has been increasing over the past years in many cities, especially in North America. Bicycling is encouraged around the world. Many of industrialized countries have moved away from bike lanes and provide bike tracks between the sidewalk and automobile lanes, with parking being between the travel lane and the bike track. However, a significant portion of bike lanes on urban arterials still have adjacent on-street parking spaces, and vehicles are required to cross the bike lane before entering the on-street parking spaces. As indicated earlier, the vehicles that are backing out from their parking spaces present a potential hazard for cyclists. Along with backover accidents, there are some other safety issues for cyclists that can be caused by parked vehicles (see Figure 3). "Dooring accidents" that involve opening car doors and cyclists are observed quite often in certain cities and

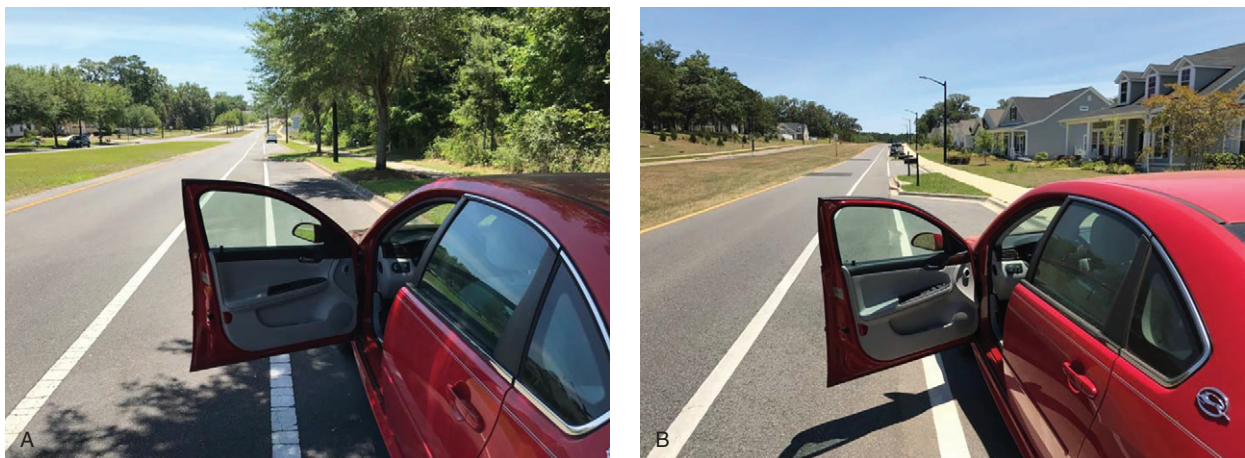


Figure 3 A bike lane adjacent to curb parking (Tallahassee, Florida, USA): (left) - Four Oaks Boulevard; (right) - Orange Avenue East.

comprise a significant percentage of cyclist-motor vehicle accidents (Schimek, 2018). Different cyclist organizations along with government agencies keep warning cyclists regarding suddenly opened doors of parked vehicles, even if cyclists are strictly using dedicated bike lanes. Nevertheless, dooring accidents continue occurring every year. Various types of injuries can be caused by dooring accidents. The cyclist can have a cutting injury after colliding with a sharp edge of the opening door. The cyclist can break the door window glass after collision at a full speed. In many cases, cyclists fall on the ground after colliding with suddenly opened doors, which further causes injuries from hitting the asphalt. Certain dooring accidents may even result in fatalities when cyclists fall in the direction of approaching vehicles, as drivers of approaching vehicles may have a very limited time to react.

The safety issues of cyclists can be addressed by increasing the bike lane width. When the minimum bike lane width standards are used, cyclists generally have to ride in the range of opening doors. However, most of the cyclists tend to ride outside of the range of opening doors after creating an additional buffer space of 3–4 feet ($\approx 0.91\text{--}1.22\text{ m}$) between the bike lane and parked vehicles. Some countries started incorporating the buffer space for bike lanes to capture the door zone in their design guidelines (e.g., North America); however, many countries are still using the minimum bike lane width standards (Schimek, 2018). Furthermore, regulations in certain countries require vehicle occupants to look around before opening the doors of a vehicle after parking. Opening doors is permitted only when it is safe to do so. Many safety guidelines also encourage cyclists to stay away from the doors of parked vehicles and use additional precautions when passing parked vehicles (e.g., reduce travel speed, so there will be more time to react in case of a suddenly opened door).

Vehicle Parking and Pedestrians

Backover accidents (i.e., a pedestrian is struck by a reversing vehicle) are one of the major safety concerns for pedestrians in parking lots. There may be different causes of backover accidents involving pedestrians. Pedestrians may not realize that a vehicle is trying to leave the parking space and walk behind a reversing vehicle. As discussed earlier, the backing-out maneuver may cause challenges for a driver with the identification of pedestrians approaching from the driving aisle. A substantial portion of backover accidents involves children (Rouse and Schwebel, 2019). The involvement of children in certain backover accidents can be explained by the fact that children may not understand the meaning of reverse lights and continue moving in the direction of a reversing vehicle. Furthermore, a driver, performing the backing-out maneuver, will have more difficulties to see children as compared to other groups of pedestrians due to the short stature of children. In some cases, children may get injured in backover accidents due to lack of adult supervision. Without constant adult supervision and guidance, many children can create risky situations in parking lots. In the meantime, young children lack certain important cognitive skills (e.g., attention and concentration), which are necessary for safe pedestrian behavior. Along with young children, older pedestrians also suffer from lack of cognitive skills that may result in backover accidents.

In order to reduce the number of backover accidents in parking lots, vehicle-manufacturing companies started installing advanced technologies in their vehicles. Rear cross-traffic alert systems, rear-view cameras, and automatic braking are some of the common technologies that are used in modern vehicles to assist drivers with safer backing decisions and prevent collisions with pedestrians (Findley et al., 2020). Many newer vehicles have rear cross-traffic alert systems that use sensors to detect objects (e.g., cyclists, pedestrians, other vehicles) that are approaching from the areas that are difficult to notice in the direction of a reversing vehicle. Once the approaching object is detected, visual alerts, audio alerts, and even automatic brakes will be applied. Rear-view cameras can be also used to identify objects when performing the backing out parking maneuver. New technologies generally provide some safety improvements when performing parking maneuvers. Nevertheless, these intelligent technologies have to be continuously tested and modified to improve their accuracy. For example, rear-view cameras show a quite poor performance under inclement weather conditions, whereas rear cross-traffic alert systems fail to detect approaching cyclists, pedestrians, and other vehicles from time to time. Also, drivers may fail to look at the monitor, so it can be helpful to combine the camera system with an audible signal, activated only when there seems to be an object behind the vehicle.

Vehicle Parking and Older Adults

Older adults typically drive less than their younger counterparts but are still required to make parking maneuvers at the end of their trips. Different sociodemographic characteristics of individuals, such as age, gender, socioeconomic belonging, driving experience, and health conditions, may substantially impact their driving and parking abilities. Age-related changes in physiological functions of individuals (i.e., hearing, vision, coordination, perception, cognition, and reaction) may cause challenges when cruising for parking space and performing various parking maneuvers in parking lots. Older adults generally have poor useful field of view and are not able to effectively identify surrounding objects, including approaching vehicles. Moreover, some older adults have a tendency of returning to the areas that have been already visually searched, which significantly slows down the speed of processing certain important information while driving. Impaired cognitive and psychomotor skills can also lead to accidents and driving violations under normal driving conditions and disruptive driving conditions (e.g., inclement weather, rush-hour traffic). Furthermore, the presence of chronic diseases (e.g., heart diseases, diabetes, dementia, Alzheimer's disease, arthritis) generally increases mental demand of older adults while performing different driving maneuvers and increases the risk of colliding with other vehicles or pedestrians.

There are different types of aberrant parking behavior that are common for older adults, including lapses, slips, anticipation errors, execution errors, and violations (Douissembekov et al., 2014). Lapses correspond to missed actions and omissions during

parking, while slips are associated with the actions that are not performed as originally planned. Anticipation errors are generally caused by lack of attention while seeking for parking space (e.g., failure to notice the vehicles leaving their parking spaces and creating room for other vehicles). Execution errors correspond to the errors that occur throughout execution of parking maneuvers. Violations represent parking actions that may lead to some legal issues. Lapses, slips, and execution errors can be explained by the declining ability of older adults to effectively survey the environment for surrounding objects and obstacles. As for anticipation errors, older adults may have difficulties noticing the presence of other drivers and anticipating their actions in parking lots. Moreover, despite the fact that older adults generally take more time to perform driving maneuvers (including parking), some of these maneuvers can be quite risky or even illegal.

Automation Impacts on Vehicle Parking

As indicated earlier, parking demand has been increasing over the years, and there is a need to better manage the existing parking space and accommodate the demand. One of the approaches to improve utilization of vehicle parking facilities is deployment of autonomous vehicles. Unlike conventional vehicles, autonomous vehicles do not require drivers and passengers to be present inside the vehicles in parking lots and perform parking maneuvers (Nourinejad et al., 2018). Drivers and passengers can exit autonomous vehicles at the entrance to a parking facility or near the designated drop-off area. The autonomous vehicle parking systems are expected to accommodate more vehicles, as the driving lanes can be narrower, elevators and staircases that are generally used by passengers can be completely removed, and the additional room that is used for opening vehicle doors will not be necessary anymore. The parking space utilization in parking facilities can be increased by parking autonomous vehicles in several rows. Typical layout examples of conventional and autonomous vehicle parking facilities are presented in Figure 4 and Figure 5.

Along with improved parking space utilization, autonomous parking has some safety benefits for passengers as well. In fact, no drivers and passengers will be present inside autonomous vehicles while these vehicles perform parking maneuvers, which eliminates the possibility of passengers being injured during parking. Moreover, since no passengers will be walking in autonomous parking lots (except the designated pick-up and drop-off areas), the possibility of backover accidents can be eliminated. However, there are some challenges with using autonomous vehicle parking facilities. In particular, an autonomous vehicle that is required to leave a parking facility may be surrounded by other vehicles (i.e., “barricaded” or “blocked” by other vehicles), and some of the parked vehicles have to be relocated in order to create the required room for the leaving vehicle (see Figure 5). This is a similar issue

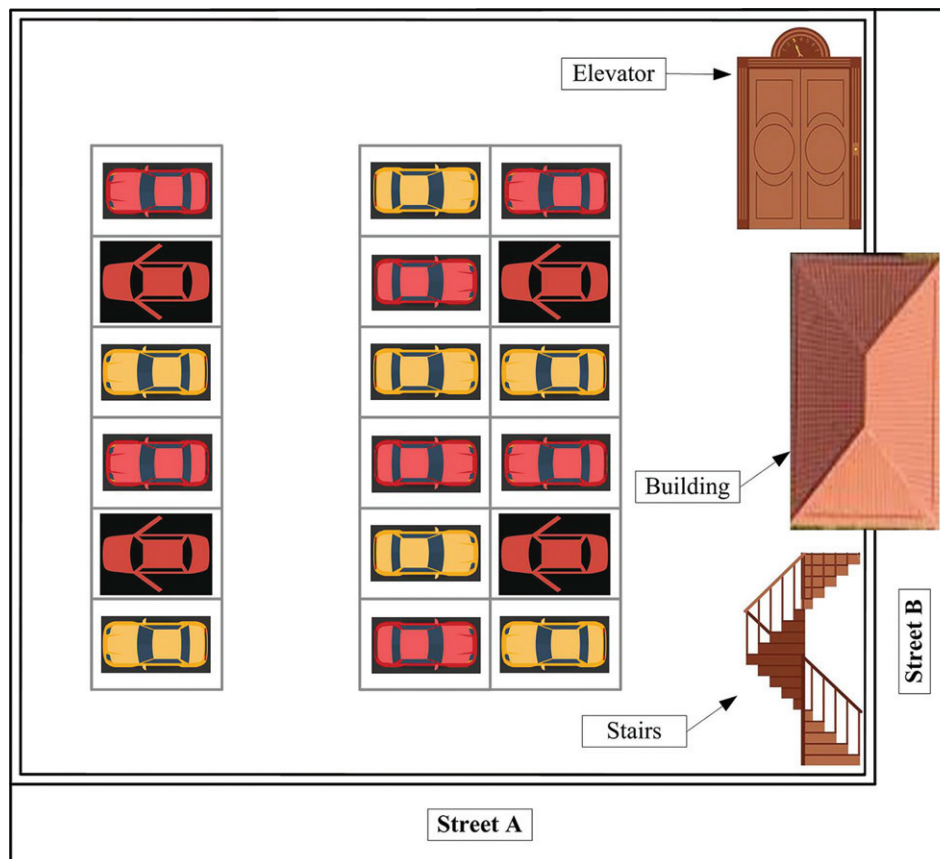


Figure 4 Conventional vehicle parking facility.

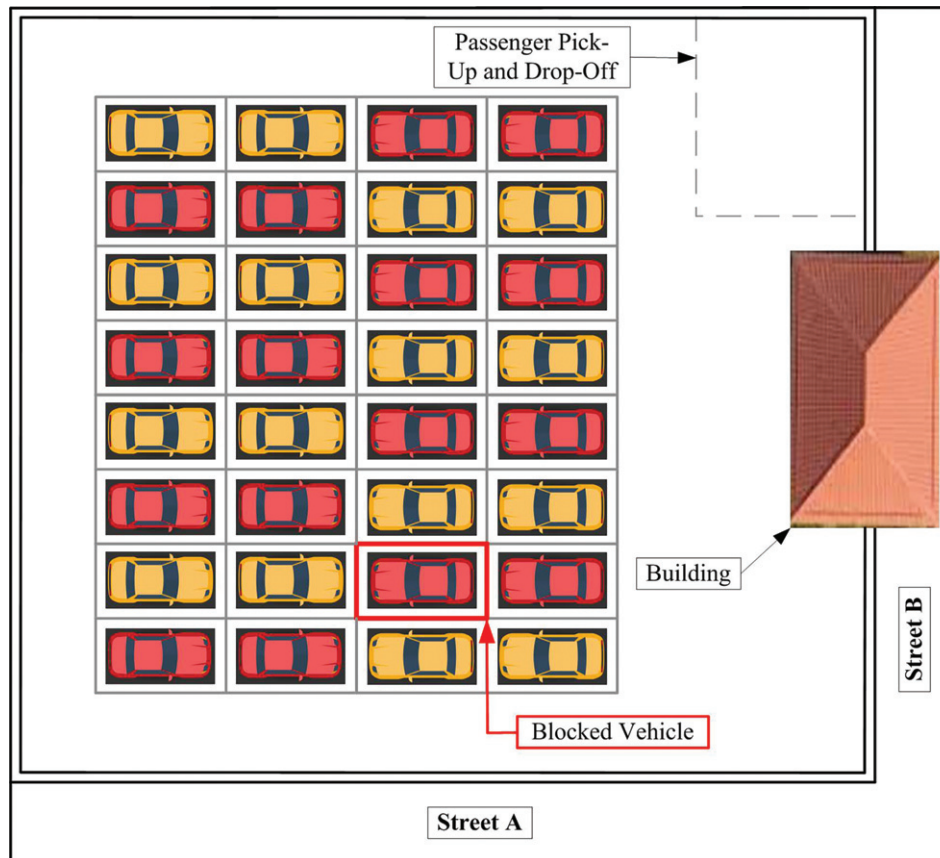


Figure 5 Autonomous vehicle parking facility.

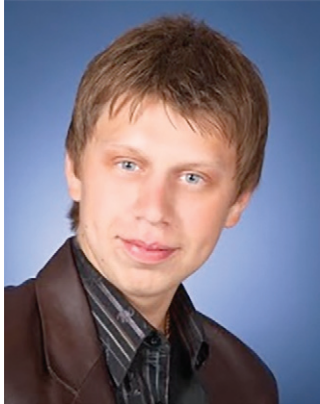
as that of stacked parking, which has been used for decades in some big cities (e.g., New York, United States). But the number of such parking facilities is not significant, and these facilities are generally managed by attendants. Autonomous parking systems may soon be standard around the world even in smaller towns. Therefore, advanced decision support systems are required to allocate the appropriate parking spaces in autonomous vehicle parking facilities. Ideally, the number of parking rows should be increased in autonomous vehicle parking facilities to improve the parking space utilization. In the meantime, the number of vehicle relocation moves should be minimized, so passengers will not have to wait for their vehicle to exit the parking facility for an extended period of time. Moreover, some special design considerations should be given to the parking facilities that serve both conventional and autonomous vehicle types.

Concluding Remarks

The demand for passenger and freight transportation has been growing in later years, even if the 2020 pandemic temporarily has halted this development. Passenger vehicles are the most popular mode of passenger travel. At the end of each trip, a driver is required to find a parking space. Considering an increasing demand for vehicle parking, local authorities of many cities have made significant investments into large parking facilities, some of which have quite a substantial capacity. Safety and security remain some of the most important issues in parking lots. Thousands of pedestrians and cyclists are injured every year in parking lots as a result of backover accidents. These accidents are primarily caused by the inability of drivers to detect approaching objects and lack of attention from pedestrians and cyclists. The number of backover accidents can be reduced if drivers start using alternative parking maneuvers (e.g., pull-out or pull-through). Many cyclists in certain cities are injured from dooring accidents that are caused by suddenly opened vehicle doors. Dooring accidents can be prevented by additional precaution measures from drivers and cyclists. Moreover, creating a buffer space between bike lanes and parked vehicles serves as an effective countermeasure against dooring accidents. The bike lane can be even relocated on the other side of the parking lane, next to the sidewalk, to effectively reduce the number of dooring accidents. Drivers are required to be more attentive when performing parking maneuvers if they see young children or older pedestrians. Due to lack of cognitive skills, young children and older pedestrians may cause risky situations in parking lots.

Some parking safety and security issues can be alleviated through the implementation of effective policies. For example, by imposing appropriate fees for curb parking spaces local authorities can reduce or even completely eliminate excessive cruising for curb parking, as excessive cruising may lead to traffic congestion and increase the risk of accidents. Furthermore, deployment of autonomous vehicles is expected to improve safety in parking lots, since no drivers and passengers will be present inside autonomous vehicles while these vehicles perform parking maneuvers. There will be no passengers walking in autonomous parking lots (except the designated pick-up and drop-off areas), which will eliminate the possibility of backover accidents. Many vehicle-manufacturing companies started installing advanced technologies (e.g., rear cross-traffic alert systems, rear-view cameras, and automatic braking) in their vehicles in order to reduce the number of accidents in parking lots. These technologies allow detecting the objects that approach in the direction of a reversing vehicle and preventing backover accidents in many cases. However, intelligent parking technologies still should be further tested and modified to improve their accuracy.

Biography



Maxim A. Dulebenets, PhD, PE is an Assistant Professor in the Department of Civil & Environmental Engineering at Florida A&M University-Florida State University (FAMU-FSU) College of Engineering. Dr. Dulebenets holds BSc and MSc degrees in Railway Construction from the Moscow State University of Railroad Engineering (Moscow, Russia), and MSc and PhD degrees from the University of Memphis (Memphis, TN, USA) in Civil Engineering. His research interests include, but are not limited to, operations research, optimization, simulation modeling, meta-heuristics, transportation engineering, and safety. Dr. Dulebenets has been involved in a variety of research projects with the overall value of ~\$5.4 million (~\$1.2 million as a Principal Investigator), sponsored by different public and private agencies, such as the United States Department of Transportation, National Science Foundation, Center for Accessibility and Safety for an Aging Population, and Florida Department of Transportation. He serves as a referee for 60+ international journals, including European Journal of Operational Research, International Journal of Production Economics, IEEE Transactions on Intelligent Transportation Systems, and Computers & Operational Research. Dr. Dulebenets is an Invited Member of the Standing Committee on International Trade and Transportation (AT020) of the Transportation Research Board of the National Academies of Sciences, Engineering, and Medicine (USA). Furthermore, Dr. Dulebenets is an Affiliated Member of the Institute for Operations Research and the Management Sciences (INFORMS) and the Institute of Electrical and Electronics Engineers (IEEE).

References

- AAA, 2015. FACT SHEET: Rear cross traffic alert systems. Accessed on 05/10/2020. Available from: <https://newsroom.aaa.com/wp-content/uploads/2015/12/RCTA-Fact-Sheet1.pdf>.
- Douissembekov, E., Gabaude, C., Rogé, J., Navarro, J., Michael, G.A., 2014. Parking and manoeuvring among older drivers: A survey investigating special needs and difficulties. *Trans. Res. Part F: Traffic Psychol. Behav.* 26, 238–245.
- DriveSpark, 2015. Top 10 Biggest Parking Lots In The World. Accessed on 05/19/2020. Available from: <https://www.drivespark.com/off-beat/top-10-biggest-car-parking-in-the-world-010286.html>.
- Dulebenets, M.A., 2019. Minimizing the total liner shipping route service costs via application of an efficient collaborative agreement. *IEEE Transact. Intel. Trans. Sys.V* 20 (1), 123–136.
- Findley, D.J., Nye, T.S., Lattimore, E., Swain, G., Bhat, S.K.P., Foley, B., 2020. Safety effects of parking maneuvers. *Trans. Res. Part F: Traffic Psychol. Behav.* 69, 301–310.
- Marsden, G., 2006. The evidence base for parking policies—a review. *Trans. Policy* 13 (6), 447–457.
- Mittal, S., Dai, H., Fujimori, S., Hanaoka, T., Zhang, R., 2017. Key factors influencing the global passenger transport dynamics using the AIM/transport model. *Trans. Res. Part D: Trans. Environ.* 55, 373–388.
- NHTSA, 2008. Fatalities and injuries in motor vehicle backing crashes. Report No. DOT HS 811, 144.
- Nourinejad, M., Bahrami, S., Roorda, M.J., 2018. Designing parking facilities for autonomous vehicles. *Trans. Res. Part B: Method.* 109, 110–127.
- Parkopedia, 2019. Global Parking Index 2019. Accessed on 05/19/2020. Available from: https://cdn2.hubspot.net/hubfs/5540406/Whitepapers_research%20reports/Parkopedia-Global-Parking-Report-2019_FINAL.pdf.
- Rouse, J.B., Schwebel, D.C., 2019. Supervision of young children in parking lots: impact on child pedestrian safety. *J. Safe. Res.* 70, 201–206.
- Schimek, P., 2018. Bike lanes next to on-street parallel parking. *Accid. Anal. Prev.* 120, 4–82.
- Shoup, D.C., 2006. Cruising for parking. *Trans. Policy* 13 (6), 479–486.
- Shoup, D.C., 2014. The high cost of minimum parking requirements Transport and Sustainability. *Parking: Issues and Policies*, 5. Emerald Group Publishing Limited, England, pp. 87–113.
- Tsamboulas, D.A., 2001. Parking fare thresholds: A policy tool. *Trans. Policy* 8 (2), 115–124.
- U.S. Bureau of Justice Statistics, 2020. The National Crime Victimization Survey. Accessed on 06/18/2020. Available from: <https://www.bjs.gov/index.cfm?ty=tp&tid=44>.

Further Reading

- Abioye, O.F., Dulebenets, M.A., Ozguven, E.E., Moses, R., Boot, W.R. and Sando, T., 2020. Assessing perceived driving difficulties under emergency evacuation for vulnerable population groups. *Socio-Economic Planning Sciences*, 100878 –in press.
- Charness, N., Boot, W., Mitchum, A., Stothart, C., Lupton, H., 2012. Aging driver and pedestrian safety: Parking lot hazards study. Final Rep. BDK 83, 977–1012.
- Gallo, M., D'Acerno, L., Montella, B., 2011. A multilayer model to simulate cruising for parking in urban areas. *Trans. policy* 18 (5), 735–744.
- Hagemeister, C., Kropp, L., 2019. The door of a parking car being opened is a risk No kerb-side parking is the key feature for perceived safety of on-road cycling facilities. *Trans. Res. Part F: Traffic Psychol. Behav.* 66, 357–367.
- Keali, M.D., Fildes, B., Newstead, S., 2017. Real-world evaluation of the effectiveness of reversing camera and parking sensor technologies in preventing backover pedestrian injuries. *Accid. Anal. Prev.* 99, 39–44.

- Kidd, D.G., Brethwaite, A., 2014. Visibility of children behind 2010-2013 model year passenger vehicles using glances, mirrors, and backup cameras and parking sensors. *Accid. Anal. Prev.* 66, 158–167.
- Kidd, D.G., Reimer, B., Dobres, J., Mehler, B., 2018. Changes in driver glance behavior when using a system that automates steering to perform a low-speed parallel parking maneuver. *Trans. Res. Part F: Traffic Psychol. Behav.* 58, 629–639.
- Marsden, G., May, A.D., 2005. Do institutional arrangements make a difference to transport policy and implementation? Lessons for Britain. *Environ. Plan. C: Govern. Policy* 24 (5), 771–789.
- Xu, C., Ding, Z., Wang, C., Li, Z., 2019. Statistical analysis of the patterns and characteristics of connected and autonomous vehicle involved crashes. *J. Safe. Res.* 71, 41–47.

Passenger Van Safety

Saksith Chalermpong^{*,†}, Apiwat Ratanawaraha[‡], ^{*}Department of Civil Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand; [†]Transportation Institute, Chulalongkorn University, Bangkok, Thailand; [‡]Department of Urban and Regional Planning, Faculty of Architecture, Chulalongkorn University, Bangkok, Thailand

© 2021 Elsevier Ltd. All rights reserved.

Introduction	401
Use	401
Accident Statistics	402
Factors Affecting Passenger Van Safety	403
Vehicle Design	403
Vehicle Safety Tests	403
Vehicle's Maintenance and Inspection	403
Driving Skills	403
Driver's Behaviors	403
Policy Initiatives, Safety Precautions, and Recommendations	404
General Safety Precautions	404
Safety Regulations in Passenger Vans Used for Public Transport	404
Biographies	404
See Also	405
References	405
Further Reading	405

Introduction

Passenger vans are medium-sized vehicles that are larger than passenger cars but smaller than buses. In this paper, the term passenger van, as generally used in North America, refers to vehicles with the seating capacity of 12–15 people, which are primarily used for transporting passengers. In other countries, passenger vans are sometimes known as minibuses, such as in Australia and New Zealand, and minibus taxis in South Africa. Vans are widely used and known as “colectivos” or shared buses/taxis in several Latin American countries.

There is a wide range of designs and body configurations of passenger vans in different countries, but most have a carrying capacity between 10 and 15 seats. Van dimensions are typically 5–6 m in length, 1.6–2 m in width, and 2–2.3 m in height. Seating arrangements also vary, but most have three to five rows of seats, with one or two passenger seats in the same row as the driver's seat, and three to four seats per row in the rows behind the driver's seat. In the passenger areas, seats in the front part are usually arranged to allow space on one side of the van for passenger movement while boarding and alighting the vans. A sliding door on this side is common, although some models have swing doors or have doors on both sides of the vehicles. The rear part of passenger vans usually has a small space for cargo storage, and a set of barn doors, which are hinged on the side or a tailgate door, which is hinged on the top. Smaller vans, with seven to eight seats, are usually referred to as minivans.

Use

Although in many countries vans are still used primarily for transporting goods, passenger vans were developed in the 1970s to cater to a group of people traveling together for various purposes, including work commute, business, recreation, and school. Since then, passenger vans have become increasingly popular in many parts of the world, due to its versatility, maneuverability, and relatively low costs. According to the National Transportation Safety Board, as many as 500,000 15-passenger vans are currently in use in the United States alone (Subramanian, 2004).

Passenger vans are used for a wide range of trip purposes, including private, commercial, and institutional activities. They are often used for school and office trips and shuttle services. People can also rent them for family and group tours in tourist destinations. Van rental services are widely available in North America, Europe, and other popular locations around the world.

In a number of developing countries, passenger vans are used for public transportation, often operated by informal owner-operators, due to lower investment costs compared with buses (Behrens et al., 2015). For example, in South Africa, minibus taxis play a major role in the country's public transport system. The South African National Taxi Council estimated in March 2017 that there were more than 200,000 minibus taxis in the country. In other African countries, passenger vans are widely used as share taxis and are known by local names, such as HiAce (Cape Verde), Matatus (Kenya), and Tro Tro (Ghana). In Thailand, approximately 15,000 passenger vans provide local and intercity public transport services, and 21,000 passenger vans are used for chartered services in 2019.

Even in developed countries, passenger van operators in some cities provide commuter services in areas with limited conventional public transport services. For instance, in the New York metropolitan areas, “dollar vans” serve a number of areas with limited subway services, particularly in low-income, immigrant communities. In Miami, Florida, passenger vans, known locally as “jitneys,” have long provided paratransit services in areas not served by buses and Metrorail.

Passenger van services range widely from ones with fixed routes, fixed stops, and fixed schedules to those without anything fixed. The level of service flexibility largely depends on the regulatory context of the city in which they operate. For example, in New Zealand, passenger vans are allowed to provide commercial service with one end of the trip ending at transport stations, such as railway stations, airports, or ferry terminals. Passenger vans are thus classified as commercial vehicles and must conform to public transport regulations including vehicle standards and driver’s qualification. In New York City and Miami, the van services generally operate on a generally fixed-route and fixed-stop basis, but they do not usually follow a fixed schedule. In South Africa, passenger vans provide services in both fixed and unfixed routes, and most, if not all, of them do not operate on a fixed-stop and fixed-schedule basis (Behrens et al., 2015). This is also the case in Thailand. Passenger vans are also available for chartered service and rental in several countries.

Passenger van services are often characterized by informality in that the operations are not initiated, supported, and, at least at the beginning, regulated by the government. They sometimes compete directly with buses and other public transport, but then they also provide complimentary services to the conventional public transport systems, serving as feeder modes in areas where the conventional modes are unavailable or as on-demand paratransit service for mobility-impaired persons. In Thailand, van operations were mostly informal at the beginning in the 1990s, but they are now systematically organized and regulated by the government. In New York City and Miami, the dollar van and jitney operations are heavily regulated by the city governments.

Accident Statistics

Van crash statistics vary among different countries. In the United Kingdom, according to the Department for Transport, between 2006 and 2015, the rates for fatality and serious injury per billion passenger-kilometers among vans (54) are much lower than cars (208) and motorcycles (4,018) (UK Department for Transport, 2017). In the United States, between 1995 and 2001, the rate of fatal single-vehicle crashes per 100,000 registered vehicles with 15-passenger vans (15.6) is substantially higher than cars (8.7), pick-up trucks (11.9), and SUVs (13.3), although the rate has steadily been decreasing (Subramanian, 2004). In Thailand, the number of fatal injuries involving motorcycles and cars far exceeds those of passenger vans, in general. However, passenger vans used for public transport purpose are notoriously unsafe compared with other types of vehicles. According to the Department of Land Transport, in 2017, despite the relatively small number of passenger vans registered in Thailand (41,202), there were 252 passenger van accidents, with 135 fatalities and 1,046 serious injuries in Thailand. The accident statistics improved significantly in 2018 thanks to the government’s required installation of GPS tracking devices, which discourage speeding and reckless driving behavior.

Similarly, in South Africa, a total of 70,000 minibus taxi crashes occur annually, according to a study by the Automobile Association of South Africa. Although minibus taxis account for only 4.5% of the total number of passenger vehicles, they are reportedly involved in 8.6% of all crashes in 2001. Almost 10% of road traffic fatalities in the country involve minibus taxi-related accidents.

The most common passenger van accidents include single-vehicle crashes, rollover crashes, and tire blowouts. In the United States, a significant proportion of fatal accidents have been found to be associated with rollovers caused by tire failures. Statistics show that half of fatalities in passenger van accidents occurred in heavily loaded vans that rolled over. The high risk of rollover crashes and tire failures in passenger vans has been attributed to the vehicles’ higher center of gravity, and consequently instability, especially when fully loaded (Potter et al., 2013). The frequency and severity of passenger van accidents may also be caused in part by the driver’s lack of training and experience in maneuvering properly in response to incidents, such as fishtailing or tire blowouts. Also, a substantial proportion (70%) of fatalities to passenger vans’ occupants have been associated with those who were ejected from the vehicles, prompting the US federal authority to issue recommendations to state authorities to revise the law to mandate safety-belt use among passenger van occupants. Recent statistics show that fatalities to van passengers have declined steadily, probably due to the awareness of van safety issues as promoted by the authorities (Subramanian, 2008). However, critics argue that the downward trend in passenger vans’ accident statistics was caused by a decrease in the use of passenger vans, rather than improvements in safety standards.

The accident risk exposure of different uses of passenger vans varies. For example, due to lower amount of mileage driven and relatively infrequent use, the risk exposure of passenger vans used for private purposes is likely to be lower than that of passenger vans used for institutional and commercial purposes, and those in public transport services. Passenger vans used for intercity public transport services, especially those on long-distance routes, are more prone to accidents than those used for urban transport services. Moreover, passenger vans in chartered services are more prone to accident than those in fixed route services since chartered van drivers tend to be less familiar with the roadway environment than those of fixed route services. For these reasons, analyses of accident statistics involving passenger vans must take such factors into account. In Thailand, for example, a study by the Department of Land Transport shows that the annual rates of fatalities per 10,000 vehicles in 2015 are 19 and 40 for passenger vans in urban and intercity routes, respectively. Likewise, the rates of serious injuries per 10,000 vehicles are 132 and 354 for passenger vans in urban and intercity routes, respectively. Therefore, measures to prevent accidents involving passenger vans must be tailored according to the vehicle’s purpose of use. Unfortunately, accident statistics by purpose in many countries are not readily available.

Factors Affecting Passenger Van Safety

Analyses of accident statistics of passenger vans in many countries show that the accident risk and severity of accidents can be attributed primarily to vehicle- and human-related factors. Vehicle-related factors include design problems, the way in which the vehicles are loaded, and roadworthiness of the vehicles. Human-related factors include the driver's skill in responding to incidents, the driver's attention and alertness, and passengers' use of restraint equipment. These factors may be interrelated. For example, due to design problems in some models, passenger vans' instability may require that drivers possess certain driving and correction skills so that potential accidents can be avoided.

Van drivers' behaviors that affect safety may also depend on the incentive structure in which they operate. For example, the low-cost nature of passenger vans also means that the vehicles can be cheaply acquired and used for providing public transport services in countries where regulations are lax. Without proper safety inspection and enforcement of traffic rules, passenger vans may be driven recklessly at high speed by small, owner-operated businesses making them prone to serious accidents. Drivers may also work for longer hours in order to earn more fare revenues, causing fatigue, and may fall asleep behind the wheel. In this section, each factor that causes safety concerns of passenger vans are discussed in turn.

Vehicle Design

Several design problems in some passenger van models are cited as causes of passenger van's poor safety records. First, due to the vehicle's length, some van models are more prone than others to fishtailing, especially when fully loaded. Second, vans can be unstable where there are crosswinds due to the vehicles' flat sides. Third, some van models have relatively short or no crumple zones, and provide little protection to occupants in the front part of the vehicles in the event of head-on collisions. Fourth, seating arrangements and the lack of escape doors in some van models may also prevent timely evacuation of occupants in the event of a stuck or damaged sliding door. Fifth, uneven distribution of load on tires on different sides of the vehicle due to unbalanced seating arrangement increases likelihood of blowouts if the tires are worn or not properly inflated. Finally, due to the high center of gravity of passenger vans, the vehicle's stability tends to be poor, posing greater risk of a rollover accident. The problem of high center of gravity is especially severe when vans are fully load or when vans are modified, such as the installation of a compressed natural gas fuel tank, which is common in some countries.

Vehicle Safety Tests

Various safety aspects of passenger vans have been tested by New Car Assessment Programs (NCAP) in different regions. For example, Euro NCAP conducted crash tests of several passenger van models, and recommended that since passenger van design is derived from commercial vans, its safety features are not updated as regularly as passenger cars or sedans, so van manufacturers should improve safety features of the vehicles more regularly. Toyota HiAce, the passenger van model that is most widely used in Thailand, has been tested by Australasian NCAP, and the result shows relatively poor passenger protection in frontal impacts.

Vehicle's Maintenance and Inspection

In addition to design problems, van safety concerns may result from poor vehicle maintenance. In developing countries where passenger vans are used for informal public transport services and official vehicle inspections are lax, operators may skip routine maintenance to save costs, resulting in unsafe vehicle conditions.

Driving Skills

Drivers' handling of passenger vans can also increase the risk of accidents. As fully loaded and lightly loaded vans handle much differently, faced with instability while driving fully loaded vans and fishtailing, drivers who are unfamiliar with passenger vans may overcorrect the steering, thereby increasing the risk of a rollover. Drivers who are not familiar with passenger vans may also not know the proper tire pressure and fail to inflate rear tires properly, thereby increasing the risk of blowouts. In developing countries where vans are used as public transport vehicles, van drivers' behavior significantly increases the risk of van accidents. Due to the vehicle's relatively small size and maneuverability compared to larger public transport vehicles, van drivers can drive at higher speed and make abrupt adjustment in directions. Where public safety regulations are loosely enforced, unsafe driving practices by passenger van drivers can translate into higher risk of accidents.

Driver's Behaviors

Inappropriate behaviors of drivers and passengers also contribute to poor safety records of passenger vans, especially in developing countries where law enforcement is lax. Due to its relatively small size, it is quite easy for van drivers to use high speed, exposing the vehicles to severe crash risks. Fare revenue incentives also induce other risky behaviors other than speeding, such as overloading of passengers, modification of seats to load more passengers than design capacity, as well as working longer hours than allowed, causing fatigue. Failure of passengers to use safety belts also contributes to serious injuries or fatalities when accidents occur.

Policy Initiatives, Safety Precautions, and Recommendations

General Safety Precautions

Due to poor safety records, governments have increasingly regulated passenger vans' operation. In the United States and Canada, transport agencies have repeatedly issued cautionary warnings about passenger van safety. The National Highway Traffic Safety Administration (NHTSA), the US safety authority, recommends specific precautions regarding the use of passenger vans, including experiences of drivers, ensuring appropriate tire pressure, seating of passenger to avoid unbalanced load, use of seat belts by all passengers, and prohibition of carrying cargo on the roof. Since the low rates of seat-belt usage among passengers of vans involved in fatal crashes were attributed to the fact that vans are treated as buses and hence passengers are not required to wear safety belts, the authority also recommended that state laws regarding mandatory use of seat belts be revised to include passenger vans.

Safety Regulations in Passenger Vans Used for Public Transport

In countries where passenger vans are used for public transport services, the government has also implemented several policy initiatives to improve the poor safety records of passenger vans. In South Africa, the national government has implemented the Taxi Recapitalization Program since 2000, which aims to replace old passenger vans with new purpose-built 18- and 35-seater vehicles (Behrens et al., 2015). In Thailand, serious van accidents prompted the government to phase out passenger vans for public transport use, starting in 2017. For those vans that are still in use for public transport services, some modifications of the vehicles are required, including the removal of some seats in the last row to allow escape in the event of an accident and to limit the number of passengers to 13 for intercity services. Installations of safety equipment such as safety belts for all seats, emergency hammer window breakers, fire extinguishers, as well as GPS tracking are also mandated for all passenger vans. However, at the time of writing, it remains unclear whether these initiatives are successful in curbing passenger van accidents and injuries.

When passenger vans are used for public transport, accidents are often caused by drivers' reckless driving behaviors and fatigue. The standard punitive responses are to fine or arrest individual drivers for violating traffic regulations, and to have their licenses suspended or revoked. But drivers' behaviors and exhaustion could be attributed to the underlying incentive structure of the business and operational models (Ratanawaraha and Chalermpong, 2018). Because passenger-van services are usually operated with little or no financial support from the government, the drivers are often without adequate salaries and social benefits. As their income depends on the number of passengers and trips, they are thus incentivized to work long hours and drive fast (Sinclair and Imaniranz, 2015). Such behaviors increase safety risks on the road. Compensation levels and methods should therefore be carefully determined in such a way that increases operational efficiency without compromising safety.

The business models are shaped by the way in which the government allocates the rights and routes to operate van services. Some licensing schemes allow open competition among operators of vans and other modes of public transport (Chalermpong et al., 2016). This often results in cutthroat, on-the-road competition on lucrative routes in high demand and an increase in safety risks, while leaving unserved other routes in low demand.

Setting and enforcing traffic and drivers' standards is one policy option to deal with safety risks associated with van drivers. In many cities, drivers of passenger vans are required to take specific safety training courses and to acquire professional/commercial driver's licenses. For instance, jitney operators in Miami are required to obtain a Passenger Motor Carrier Certificate of Transportation, while drivers must have a for-hire chauffeur registration and must attend specific training programs conducted by the For-Hire Transportation section of the Business Affairs Division. Drivers' eligibility qualifications include driving record and criminal background checks, as well as mandatory participation in drivers' training programs. The vehicles themselves are required to obtain inspection and operating permit decals issued by the Vehicle Inspection Station. Similar regulations are also applicable to dollar-van operators in the New York metropolitan areas, as well as in cities in developing countries. But the rigor with which they are enforced varies from place to place.

Another policy option is to adopt the chain-of-responsibility approach to regulating passenger van services. The basic idea is that, because each party in the transport supply chain assumes different levels of responsibility in creating and managing safety risks, they should also be held accountable differently. Accordingly, legal obligations and penalties are placed not only on the drivers themselves but also on the licensees who acquire the right to operate from the government. In the case of New York City, when an individual van driver violates a traffic rule or an operational requirement, that violation is counted cumulatively for a certain time period and also collectively for the whole group that operates under the same license.

Biographies

Saksith Chalermpong is associate professor in the Department of Civil Engineering at Chulalongkorn University, Bangkok, Thailand. He currently serves as Deputy Director of Chulalongkorn University Transportation Institute. He was a visiting researcher at Kyoto University Center for Southeast Asian Studies. His research interests include urban transportation planning and policy, public and informal transportation, and equality issues in transportation policy. He received his bachelor's degree in civil engineering from Chulalongkorn University, his master's degree from MIT, and his doctoral degree from UC Irvine, both in the field of transportation.

Apiwat Ratanawaraha is associate professor in the Department of Urban and Regional Planning, and an advisor at the Urban Design and Development Center, both at Chulalongkorn University in Bangkok, Thailand. He is specialized in urban planning and development, infrastructure finance, technology and innovation policy, and futures studies. His ongoing and recent research includes projects on the futures of urban life in Thailand, the futures of clubs and commons, globalization of land, urban citizen science, and informal mobility in Thailand. He was a visiting assistant professor at the MIT Department of Urban Studies and Planning, and a visiting scholar at the Harvard-Yenching Institute, Cambridge, MA.

See Also

Aggressive Driving and Road Rage; Automobile Safety Inspections; Driver State and Mental Workload; Education and Training of Drivers; HUMAN FACTORS IN TRANSPORTATION; Sleep-Related Issues and Fatigue

References

- Behrens, R., McCormick, D., Mfinanga, D. (Eds.), 2015. *Paratransit in African Cities: Operations, Regulation and Reform*. Routledge, New York, NY.
- Chalermping, S., Ratanawaraha, A., Sucharitkul, S., 2016. Market and institutional characteristics of passenger van services in Bangkok, Thailand. *Transp. Res. Rec.* 2581 (1), 88–94.
- Potter, T., Dubois, S., Haras, K., Bédard, M., 2013. Fifteen-passenger vans and other transportation options: a comparison of driver, vehicle, and crash characteristics. *Traffic Inj. Prev.* 14 (7), 706–711.
- Ratanawaraha, A., Chalermping, S., 2018. How operators' legal status affects safety of intercity buses in Thailand. *Transp. Res. Rec.* 2672 (31), 99–109.
- Sinclair, M., Imaniranz, E., 2015. Aggressive driving behaviour: the case of minibus taxi drivers in Cape Town, South Africa. In: *Southern African Transport Conference*. Pretoria, South Africa.
- Subramanian, R., 2004. Analysis of crashes involving 15-passenger vans (No DOT-HS-809-735). Technical Report. National Center for Statistics and Analysis (US).
- Subramanian, R., 2008. Fatalities to occupants of 15-passenger vans, 1997–2006. Report No. DOT-HS 810 947. National Highway Traffic Safety Administration, Washington, DC.
- UK Department for Transport, 2017. *Transport statistics Great Britain 2017*, Department for Transport, London, United Kingdom.

Further Reading

- Wegmann, F., Noltenius, M., 2008. *Fifteen Passenger Van Safety—Recommendations on Best Practices for Commuter and Community Transportation*. Southeastern Transportation Center, University of Tennessee, Knoxville, TN.

Passive Prevention Systems in Automobile Safety

B. Serpil Acar, Design School, Loughborough University, England, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Automobile Safety	406
Seat Belts	406
Airbags	409
Head Restraints	410
Child Restraint Systems	411
Laminated and Tempered Safety Glass for Windows	412
Crumple Zones and Safety Cell	413
Collapsible Steering Columns	413
Other Effective Safety Measures	413
Acknowledgments	413
Biography	414
See Also	414
Relevant Websites	414
References	414

Automobile Safety

Automobile safety has improved immensely in recent decades. The World Health Organization (WHO) reports that casualty numbers due to road traffic incidents are still huge; globally, 1.35 million people die and 50 million are injured in road traffic incidents every year (WHO, 2018). That is nearly 3,700 fatalities and nearly 137,000 injuries every day. Policies and laws are reasonably similar in most countries. However, law enforcement and awareness vary a great deal, often with the socioeconomic status and cultural values of the countries. The number of casualties is far greater in poorer or developing countries than in the most developed countries.

Automobile traffic accidents have happened for as long as the automobile has been around (Fig. 1). Road safety can improve through a variety of approaches. Potential incidents and casualties can be predicted and prevented in certain conditions. Road structure, such as the quality of the road or road furniture, and human factors, such as driver behavior, are important subsystems of the road safety system. Designers and manufacturers in the automotive industry continuously use the latest technology and accumulated knowledge to improve active and passive safety systems to make safer vehicles to balance the complex transport safety system. Legislation is passed and enforced by governments, and organizations try to raise awareness for public safety. The focus of this article is passive safety.

Passive safety systems aim to minimize the severity of injuries and the number of fatalities during the collision when it happens. On the contrary, active safety systems seek to avoid collisions in the first place. For example, anti-lock braking systems or electronic stability control are well-known, mature active safety systems. However, they are still not standard in automobiles in many countries. Unfortunately, even in the countries where sophisticated active safety systems are mandatory in most automobile models, accidents still happen. Therefore, reliable passive safety systems in vehicles are extremely important and will always be needed.

Some passive safety systems, such as installation of seat belts, are mandatory, and some are optional depending on automobile models and where in the world they have been manufactured, sold, or registered. The US New Car Assessment Program (NCAP) provides a five-star safety rating scheme and consumer information on new cars since 1979. Similar programs, with slightly different protocols, have been set up on other continents with extensive tests taking place in independent crash test laboratories; and new vehicles are graded according to their performance in protecting their occupants in Europe, Latin America countries, South Asian countries, Australasian countries, Japan, Korea, and China, respectively, by Euro NCAP, Latin NCAP, ASEAN NCAP, ANCAP, JNCAP, KNCAP, and C-NCAP.

The following sections cover the importance, brief history, working principles, and usage of passive safety systems, namely, seat belts, airbags, head restraints, child seats, laminated and tempered safety glass for windows, crumple zones and safety cells, collapsible steering columns, and other effective passive safety measures.

Seat Belts

Using seat belts is a highly effective passive safety method of preventing serious injuries and fatalities in automobiles. They are designed to secure the driver or passengers to a stationary object in the vehicle to protect them from sudden deceleration or stop as a



Figure 1 Motor vehicle traffic accidents have happened for as long as automobiles have been around.

result of a collision. As a body would continue to move forward within the car during deceleration of the vehicle, seat belts help to avoid the occupants from being ejected from the vehicle. Importantly they also prevent the occupants from hitting the vehicle interior and each other within the vehicle. A correctly worn seat belt also helps the occupant to take maximum advantage of other restraint systems such as airbags and head restraints.

There is evidence that vehicle seat belts were invented and used in the 19th century by a mechanical engineer George Cayley to protect his coach driver taking part in flying experiments with his glider designs. It is difficult to identify the person who first “thought about” automobile seat belts. Patents filed in 1855 and 1903 by Edward J. Claghorn and Gustave-Desire Leveauto, respectively, describe belts resembling two-point seat belts to secure people to fixed objects. An article written in 1955 by Hunter Shelden, a neuroscientist, defends prevention as the only cure for head injuries, and suggests retractable seat belts amongst other safety systems such as door locks and airbags (Shelden, 1955). Griswold and DeHaven’s seat-belt idea about a combination of lap and shoulder safety belts to restrain the occupants was patented in 1955. The three-point seat-belt design, which is in principle not very different to the seat belts that automobile occupants use today, was patented by Nils Bohlin 60 years ago (in 1959) while he was working as a mechanical engineer for Volvo (Fig. 2). He had refined the idea of restraining the occupant by a shoulder belt diagonally across the chest and a lap belt across the hip bones with a simple V-shaped design which did not move under load, and, from 1959 on, it became standard in some Volvo models on the Nordic markets. Volvo almost immediately opened up the three-point seat-belt patent to all car manufacturers.

Further developments over the years have not changed the main design but improved comfort. For example, including retractors provides convenience of easy size adjustment and gives occupants reasonable freedom of movement during the journey, whereas a sudden deceleration immediately locks the seat belt and fixes the position of the occupant. Advancement in technology is also reflected in seat belts. Pretensioners are used to preemptively limit the movement of the occupant within milliseconds of a crash. Based on similar technology, inflatable seat belts distribute the crash forces across a larger area of the body during a collision.

One of the earlier major studies that compared occupant safety when people travel with and without three-point seat belts, to analyze survival and injury rates, was conducted in Sweden in the mid-1960s. The report concluded that the seat belt offers full protection against fatal injury up to accident speeds of 60 mph and the drivers’ nonfatal injuries were reduced by 57% at lower speeds and 48% at higher speeds (Bohlin, 1967). Many global studies and meta-studies combining all earlier results have been conducted since then and they all agree that seat belts save lives. On average seat belts are estimated to prevent around 45%–50% of front seat passengers’ and drivers’ fatal injuries.

The two-point seat belt, a strap used to restrain the automobile occupant across the abdomen, was used from the mid-1950s until the three-point seat belt became the norm. Various studies show that they are not as safe as three-point seat belts and prove that in extreme conditions in automobiles it can cause severe damage to internal organs and spine and consequently fatal injuries. Five-point seat belts are commonly used in child seats. Two straps for the shoulders, two for the hips, and one between the legs are used to strap down the child to the child seat, all joined to the central buckle. Six-point seat belts are used by rally drivers and are similar to five points with the addition of a further strap between the legs.

The most prominent benefit of the seat belt is to restrain the occupant to prevent them from being thrown away from the automobile. Unbelted occupants would be thrown forward with a force up to 60 times their own weight if a collision happens while traveling at 30 mph. This is equivalent to the weight of an elephant for an average adult.

Seat-belt use rates are low, and hence the fatality rates in collisions are particularly high, in developing countries. Legislation about fittings and wearing seat belts in these countries are usually similar to developed countries; however, law enforcement varies significantly. The reasons include cultural, economic, and comfort issues and they are never simple. For public safety, widespread awareness activities, targeting all ages, is an efficient way to communicate the evidence of dangers of not wearing the seat belts.



Figure 2 Modern three-point seat belt: (A) correct use—shoulder section passing through the middle of the shoulder diagonally across the chest, lap section under the abdomen on hip bones; (B) incorrect use—shoulder section off the shoulder, lap section on the abdomen, both sections loose; (C) incorrect use—shoulder section cutting the neck lap section on the abdomen; (D) incorrect use—shoulder section under the arm, lap section on the abdomen.

The force applied on the body by seat belts during a crash can also lead to some injuries. In most cases, this is highly insignificant compared to predicted injuries if the seat belt were not worn. However, any potential injury that might be caused by the seat belt is a cause for concern for some, especially pregnant women. Many pregnant women feel the discomfort of the shoulder portion of the belt due to one of the already enlarged breasts being on its pathway. The shoulder belt and continuously riding-up lap portion of the belt also normally reside on the “bump.” Special advice, given by many transport authorities, on the correct wear of three-point seat belts during pregnancy, is based on American College of Obstetricians and Gynecologists’ advice: *The correct position for the seat belt in pregnancy is with the shoulder section passing across the shoulder, between the breasts, and around the abdomen, and the lap section passing across the hips and underneath the abdomen* (ACOG, 1999). The shoulder belt also must not cut the neck and it should not be off shoulder. The lap portion should be on the hip bones not on the thighs. Globally, correct seat-belt usage rates are very low. Research in this area suggests that a correctly positioned seat belt, without interfering with the function of the seat belt, is a very effective way of protecting the fetus and the pregnant woman. Not wearing the seat belt may have fatal consequences for both. All available patents on the seat-belt designs for pregnant women are summarized in a study by Acar and Esat (2010).

Another group of special occupants is children. In many countries it is not legal to use adult seat belts directly until they are 150 cm tall. They may be legal to use once the child is 135 cm tall in some countries; however, the shoulder section is very likely not to pass through the middle of child’s shoulder, hence to cut into the neck and not to pass across the chest. Adults in wheelchairs often use special belts to secure their wheelchairs in the vehicle and use the automobile’s three-point seat belts to secure themselves.

In future automobiles, there may be a need for seat belts with similar functionality but different shape and form that considers the safety of the occupants in a potentially different internal geometric environment and more relaxed positions and postures as in autonomous cars.

Airbags

Airbags are designed to provide a soft and strong cushion, between the occupant and hard interior surfaces of automobiles, instantaneously during collisions. Current airbags are supplemental restraint systems that work with seat belts, rather than replace them. Studies show that airbags are most effective when they are used with correctly used seat belts.

The first automobile airbags were patented in the early 1950s in Germany by Walter Linderer and in the United States by John Hetrick. They were not taken up by automobile manufacturers until at least 2 decades later; however, they inspired further research and development in the area.

The first airbags sensed the collisions through physical bumper contact (too late) and worked with compressed air (too slow). In 1968, Allen Breed invented an airbag system that detected sudden deceleration, triggering electrical ignition of sodium azide. This would in turn generate nitrogen gas, instantaneously inflating the airbag. First-generation airbags were thought to replace seat belts, especially in the United States. They were also rigid and caused secondary injuries and fatalities. In 1991, Breed co-patented an airbag that vents air as it inflates, which reduced the bags’ rigidity. The working principle of modern airbags involves sensing the collision, inflating a strong airbag from folded to fully blown state in milliseconds, then starting to deflate before the occupant hits it. This provides a soft cushion that offers protection against injuries that might otherwise occur. Currently, other explosive propellant chemicals are used rather than sodium azide, due to environmental concerns. Aluminum nitrate was also used in the past but proved to be an unsuitable propellant for airbags due to its behavior in certain atmospheric conditions such as high temperatures and humidity, which resulted in recall of large number of automobiles.

Airbags were first available for drivers and front passengers. Frontal airbags were followed by knee airbags, then side curtain airbags and side torso airbags to extend the protection to front and rear occupants from side impacts and rollovers. Rear seat passengers can also be protected by rear curtain shield and rear center airbags. Some of these airbags are dual compartment airbags providing the appropriate cushion effect according to the part of the body they contact (Fig. 3).



Figure 3 Driver airbag and side curtain airbags. Source: Photo Attribution: Aero7, licensed under the Creative Commons Attribution-Share Alike 2.5 Malaysia license, resized to fit the page.

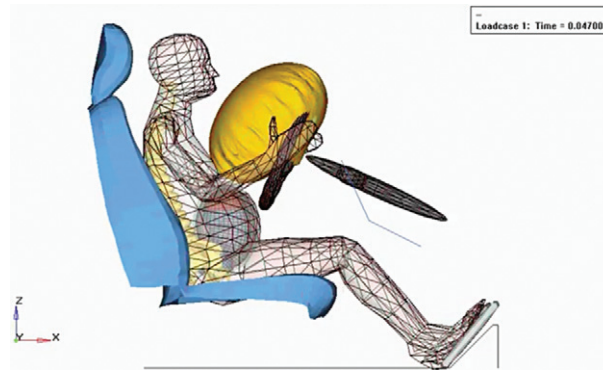


Figure 4 Simulation of 30 km/h frontal collision with 38-week pregnant driver computational model “Expecting” wearing no seat belt (video) [\[video icon\]](#).

Modern airbag control units analyze data provided by deceleration sensors and deploy the airbags if necessary. Some smart airbags also consider details such as the weight of the occupant, impact angle, and speed to adjust airbag deployment. This may be particularly useful for a group of people with weak bodies who would get injured easily and severely.

Airbags save lives. However, the latest available report by the National Center for Statistics and Analysis (NCSA) on frontal airbag-related fatalities in 2009 reveals that in the United States, from 1990 until the end of 2008, there were total of 296 confirmed airbag-related fatalities. A significant majority, 191 (65%) of them, were children. From 1990 the numbers have gradually increased and peaked in 1997–98 then started to decrease significantly until 2008. The same report estimated the gross number of lives saved by airbags in the same period as 28,244 (NCSA, 2009).

Pregnant women have been concerned about the safety of their fetus after airbag deployment. The common advice given to them is to sit 25 cm back from the center of the steering wheel. Studies suggest that the placental abruption is less likely, and the fetus is better protected with a correctly worn three-point seat belt together with a deployed airbag compared to seat belt only, airbag only, and no restraint cases. Fig. 4 shows the simulation of a low-speed frontal crash by using “Expecting,” the computational pregnant dummy with a 38-week old fetus, in the airbag only case.

There are also exterior airbags, designed to reduce the severity of collisions, limit damage, and to save lives of both the occupants and the people that the automobile comes into contact with during a collision. Exterior automobile airbags to protect pedestrians, cyclists, occupants, and automobiles in rollover crashes and to lessen the impact of colliding automobiles have been developed or patented by car manufacturers. As in seat belts, in future automobiles, for example, in autonomous cars, there will also be a need for continued innovation.

Airbags have also been developed to be worn by and protect people who might get involved in a collision with an automobile. For example, Hövding is an airbag that works as a bicycle helmet. It was invented by two female industrial design students, Anna Haupt and Terese Alstin, in Sweden in 2005. Ordinarily, it is worn around the cyclist’s neck, but in the case of an accident, it expands and the airbag covers and protects the cyclist’s head before the impact.

Head Restraints

Head restraints are predominantly used to protect the occupant from whiplash and other neck injuries. They are sometimes mistakenly called “head rests,” presumably because it was the name when the first patent was granted to Benjamin Katz in 1923 even though the purpose was still to “stabilize the head when it was subjected to jolts and irregular movements” and to “accommodate occupants of various statures.” They are not to rest, but to restrain the movement of the head in rear-end collisions during which the occupant’s head and torso move backward relative to the seat. The seat restrains the torso; however, the head suddenly rotates backward in the sagittal plane—unless it is restrained by the head restraint—then rotates forward, resembling the movement of a whip, within 125 milliseconds.

Whiplash injury is (usually) a nonlife-threatening soft tissue injury and it is the most common and one of the most costly crash injuries. There is higher risk for women than for men. It may be experienced from sudden acceleration-deceleration force on the cervical spine due to a collision in any direction but mainly after low-speed rear impacts. Some people suffer from symptoms for a long period of time, some after a delay, and diagnosis and cure are complex due to the large number of parameters including vehicle mass, velocity (ΔV), accident type, occupant’s age, gender, stature, mass, posture, and position of the head restraint.

A head restraint is either the extension of the seat (fixed) or an addition to the top of the seat, which is usually height, tilt, and sometimes depth adjustable. Many patents were filed for fixed or adjustable head restraints especially between 1960s and 1990s. Studies show that appropriate adjustment of the head restraint can be crucial to limit whiplash injuries. However, most people are not aware about the correct positioning of their head restraints. Ideally, the back of the head should be touching

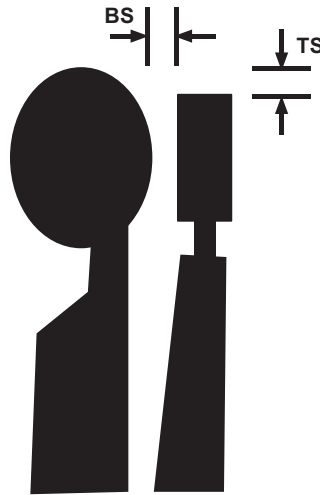


Figure 5 Backset and topset are recommended: $0 \leq BS < 7$ cm, $0 \leq TS < 6$ cm for good head restraint protection.

the head restraint and the top of the head restraint should be as high as the top of the head, but this may not be very practical. For good protection by the head restraint, BS in Fig. 5, the horizontal distance between back of the head and the head restraint (backset), should be less than 7 cm and TS, the vertical distance between top of the head and top of the head restraint, should be less than 6 cm. Unadjusted or inappropriately adjusted head restrains with $BS \geq 11$ cm or $TS \geq 10$ cm provide poor protection.

Whiplash injury represents a significant majority of personal injury claims from insurance companies. Car manufacturers have developed various mechanisms for whiplash mitigation. They can be classified into two groups. The first group is comprised of mechanisms focusing on positioning the head restraint close to the head in early stages of the rear impact. The first example of this is SAHR (SAAB Active Head Restraint) that was introduced in 1997. The system works with the occupant's weight that triggers the head restraint to move forward and upward. The NECK-PRO head restraints by Mercedes-Benz also move forward and upward but work with a sensor system which detects a rear-end collision of a predefined degree of severity. BMW's whiplash prevent system is controlled by the airbag control unit and similarly makes the head restraint move upward and forward. The second group, as in Volvo's WHIPS (Whiplash Protection System), reduces occupant acceleration by a mechanism to move the seat backward to absorb the rear-end crash energy, and upward to embrace the spine and neck, then tilts back to avoid the head rotating in the sagittal plane. Whiplash mitigating devices prevent whiplash injuries 30%–50%.

Excellent ideas to prevent whiplash injuries are developed. For example, a unique head restraint system design which adjusts its position automatically and continuously to provide the correct support for the vehicle occupant's head had been chosen as the European regional winner and the prototype was successfully demonstrated and exhibited at the 19th International Technical Conference on the Enhanced Safety of Vehicles in Washington, DC, by Loughborough University Mechanical Engineering students (LU, 2005).

Surveys reveal that most pregnant women do not adjust their head restraints and drivers in wheelchairs cannot take advantage of automobiles' head restraints.

Child Restraint Systems

Children are a very special group of occupants. They cannot directly use the seats, seat belts, and head restraints designed for adults. Hence, special seats, collectively called child restraint systems (CRSs), integrating appropriate seat belts and head restraints have been designed for their safe transport. Child Occupant protection has been included in NCAP programs.

Automobile seats for children were designed and manufactured already in the 1930s; there were multiple motivations, not including safety, as automobile safety culture was not the norm. These seats were to raise the seating level of children to enable them to see around, to limit their motions, and to be seen by the person driving the car. After the 1960s we can see that designs were evolving and making child transport safer. There are strict regulations nowadays in most countries about the manufacture and use of CRSs (Fig. 6).

Child occupant safety depends on three actions by caregivers: (1) choosing the correct CRS for the child, (2) fitting the CRS to the automobile correctly, and (3) securing the child in the CRS correctly.

For many years in the past suitable seat recommendations were based on the age and weight of the child. Lately, mainly the height is used as a measure for correct size fit. It makes sense, especially for older children, to use size rather than age. CRSs are generally recommended to be placed in rear seats. Younger children are recommended to sit in rear-facing CRSs, usually



Figure 6 Child seat from 1950s.

placed in a rear seat; however, in some countries it is legal to place them in front seats, if the passenger airbag is deactivated. Children are recommended to graduate to the next level CRS when they are too tall for their existing seats. Booster seats used to be the final seat before the adult seat; however, because boosters do not protect against side impacts, nowadays it is recommended to use CRSs with side protection. A few automobile manufacturers included integrated child seats and restraints in the rear seats.

Some portable CRSs use mechanical anchor restraints, such as ISOFIX, and some use belt installation. Comparison studies are scarce but suggest that both have advantages and disadvantages. Mechanical anchors are quicker and simpler to install correctly and hence seems to be preferred by caregivers.

Mistakes made in securing the child in the CRS also put the child in danger. For example, improperly tightened seat belts or seat belts tightened over slippery materials like some winter coats still pose hazards.

The use of CRSs has helped reduce child fatalities; however, potential errors in the three areas mentioned earlier are still a problem hindering optimum results. It is an offence not to use child seats for children in most countries, however, as always law enforcements vary.

Laminated and Tempered Safety Glass for Windows

One of the oldest effective passive safety measures which has saved many lives is the use of laminated and tempered glass in automobile windows. Regular window glass, which shatters into thin sharp pieces, was used in early automobile windshields and windows. This meant that serious injuries or fatalities after an accident were contributed to significantly by flying glass shards in addition to other consequences of the impact. Early safety glass patents were granted in 1910 in France; however, they were not used by automobile manufacturers for a long time. Laminated glass may crack under pressure and display spiderweb-like patterns but remain intact. Ford started using laminated safety glass in 1927, after refining laminated safety glass mass manufacturing (Fig. 7).

Today, all modern automobiles have laminated glass for windshields. Tempered glass is preferred for side and back windows as it is stronger than regular glass and shatters completely into tiny pebble-like dull pieces, hence eliminating injuries by broken windows during a collision and making emergency access possible.



Figure 7 Crashed laminated glass windshield.

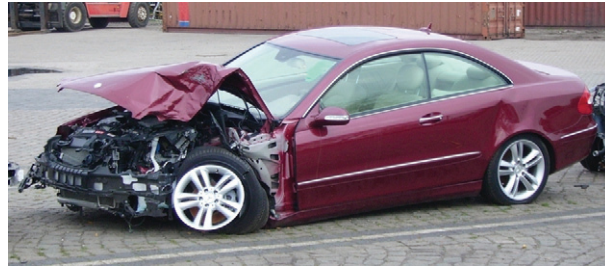


Figure 8 Crumple zone and safety cell after a serious accident. Source: Taken from public domain. Available from: https://commons.wikimedia.org/wiki/File:Crashed_Mercedes-Benz_Coupe.jpg.

Crumple Zones and Safety Cell

The body of early automobiles were uniformly rigid, did not deform in a controlled manner, and it was believed that was necessary for automobile safety. Another most effective passive safety system, Crumple Zones and Safety Cell, was invented in 1937 and improved in 1952 by Béla Barényi, an Austrian-Hungarian engineer who later worked for Mercedes-Benz. His patent was entitled “motor vehicle with body divided into three sections.”

Crumple zones are constructed at the front and back of an automobile, and envelope a nondeformable, rigid safety cell, the occupants’ compartment. During a frontal or rear crash, crumple zones absorb and distribute the kinetic energy, deforming under control and keeping the occupants safe in the strong safety cell (**Fig. 8**). The first car with crumple zones and safety cell was manufactured in 1959 by Mercedes.

A side impact protection system introduced by Volvo in 1991 distributes the side impact force to the strengthened roof, door, and floor area rather than the B pillar alone.

Rollover protection systems, especially for convertible automobiles, protect occupants by extended roll bars behind the rear head restraints once the rollover danger is detected. The mechanism works either by a gas generator or a signal sent by the airbag unit.

Collapsible Steering Columns

Another design to protect the driver in frontal collisions is to absorb the impact energy through a collapsible steering column. It is designed to prevent injuries the steering column and steering wheel can cause. In principle, a non-collapsible steering column is one long shaft connected to the gearbox. Many patents exist on collapsing mechanisms, most working like a telescope, starting to operate when the force from front-end impact is above the threshold. One of the early patents belongs to crumple zone and safety cage inventor Béla Barényi. The collapsible steering column is a commonly used passive safety measure since 1967.

Other Effective Safety Measures

Most of the current passive safety developments are based on hard work of determined researchers and advancing technology. They form parts of the “safety competition” between car manufacturers. However, in early days the introduction of simple passive safety measures made significant changes in people’s safety. Avoiding sharp edges, and protruding features, using padded dashboards and changing rear-view mirror location in automobiles can be counted among these developments. Front opening (rear-hinged) doors were quite common in the old days and they were nicknamed “suicide doors” because of potential risks of opening doors at speed. Only a few rear-hinged door (with modern locks) automobiles are available currently. There is also the child safety lock, a simple mechanism designed to avoid children (or any rear seat passenger) opening the automobile doors without authorization. When the child safety lock is activated the door can only be opened from outside. It has been a common feature in most family cars since the 1980s.

Acknowledgments

The author would like to thank EPSRC (Engineering and Physical Research Council, United Kingdom) that has funded a number of her research projects on automobile safety and Thatcham Research for its collaboration in the projects. The author also thanks people who modeled in the photographs, the Editor for his useful comments, and Baris and Baran for their constructive suggestions on the earlier draft of the article.

Biography



B. Serpil Acar is the Professor of Design for Injury Prevention at Design School, Loughborough University, United Kingdom. She received BS and MS degrees in Mathematics from the Middle East Technical University, Ankara, Turkey. After completing her PhD in the United Kingdom, she worked in close collaboration with engineering departments as well as automotive industry and clinical and academic members of the Medical Schools in the United Kingdom. Her current research interests include occupant safety, engineering design for women, and modeling the human spine. She is the principal investigator of many major Engineering and Physical Sciences Research Council funded projects investigating occupant safety. She is the winner of the 2012 Enterprise award with her invention SeatBeltPlus, a device for pregnant occupant safety.

See Also

Suggestions for cross-references to other articles within the work: Crash Not Accident; Collision Avoidance Systems, Automobiles; Elderly Driver Safety Issues; Head-on Crashes; In-Depth Crash Analysis and Accident Investigations; Wrong-Way Driving on Motorways

Relevant Websites

Insurance Institute for Highway Safety (IIHS). Available from: www.iihs.org
National Highway Traffic Safety Administration (NHTSA). Available from: www.nhtsa.gov
Road Safety Observatory (Evidence). Available from: www.roadsafetyobservatory.com
Thatcham Research. Available from: www.thatcham.org
Think! Available from: www.think.gov.uk

References

- Acar, B.S., Esat, V., 2010. Seat belt designs to protect pregnant vehicle occupants. *Recent Pat. Mech. Eng.* 3 (1), 1–10.
ACOG, 1999. Obstetric aspects of trauma management. Report 251. ACOG, Washington, DC.
Bohlin, N.I., 1967. A statistical analysis of 28,000 accident cases with emphasis on occupant restraint value. SAE Technical Paper 670925. SAE International, Anaheim, CA.
LU, 2005. Design and development of a novel head restraint system. Available from: www.lboro.ac.uk/service/publicity/news-releases/2005/30_neck_restraint.html
NCSA, 2009. Counts of Frontal Air Bag Related Fatalities and Seriously Injured Persons. U.S. Department of Transportation, National Highway Traffic Safety Administration, National Center for Statistics and Analysis Crash Investigation Division, Washington, DC.
Shelden, C.H., 1955. Prevention, the only cure for head injuries resulting from automobile accidents. *JAMA* 159 (10), 981–986.
WHO, 2018. Global status report on road safety 2018. Licence: CC BYNC-SA 3.0 IGO. World Health Organization, Geneva.

Pedestrian Safety, Children

Mette Møller, Technical University of Denmark, Lyngby, Denmark

© 2021 Elsevier Ltd. All rights reserved.

Introduction	415
Crash Involvement	415
Crash Recording	415
Crash characteristics	416
Crash Contribution	416
Crash Consequences	416
Cognitive Capacity and Skills	417
Countermeasures	417
Conclusion	418
Biography	418
References	419
Further Reading	419

Introduction

A pedestrian is a person traveling on foot from one location to another. So are people walking or jogging for recreation. People using wheelchairs and people on skateboards or similar devices are in many jurisdictions also considered to be pedestrians. Walking as a means of transportation has the potential to increase physical activity for children on a daily basis, if chosen as the main mode for the daily commute to and from school, leisure activities, or the like. Furthermore, walking provides children with the possibility for independent mobility, bonding with peers while walking, as well as improving their individual spatial and navigational skills. Thus, walking can potentially improve the quality of daily life for children due to related improvements in their physical, social, and mental health. However, despite potential individual and societal benefits, walking is also associated with a high risk of road crash involvement and related injury. This chapter provides an overview of key aspects regarding child pedestrian safety, including crash involvement, characteristic and consequences, age-related capacities for safe behavior and selected countermeasures. In this article, we define a child pedestrian to be a pedestrian below the age of 18.

Crash Involvement

Worldwide, pedestrians constitute around 22% of all police-reported road fatalities. Male pedestrians are overrepresented among all age groups and that is the case for child pedestrians as well. Furthermore, pedestrian crashes is the largest category of road traffic crashes involving children. The percentage of children in pedestrian crashes differs between countries, especially between high- and low-income countries. In high-income countries, child pedestrians constitute 5–10% of children involved in road traffic crashes, whereas in middle- and low-income countries, child pedestrians are injured or killed in 30–40% of road traffic crashes (WHO, 2008a). In Africa and Asia, where people often walk along the side of the road, the number of injured child pedestrians is particularly high. In recent years, the number of child pedestrian injuries has been reduced significantly in many high-income countries. However, despite indications of a positive development, the prevention of child pedestrian injuries remains a worldwide challenge, especially for children of the age group 5–14. In Europe, 48% of road deaths among children, involve a pedestrian. Socioeconomic factors strongly influence the risk of being killed as a child pedestrian. The risk of being killed, as a child pedestrian is four times higher for children from the lowest compared to the highest socioeconomic classes, and 20 times higher for children with unemployed parents.

Crash Recording

Several factors challenge access to knowledge about child pedestrian crashes and related injuries. First, only an injury related to an incident in which both a child pedestrian, and a vehicle (cycle, moped, motorcycle, car, etc.) are involved, is considered a road traffic crash. Consequently, injuries related to incidents involving only one child pedestrian or several pedestrians, for example stumbling in the street, slipping on an icy surface, stepping into a hole in the pavement, or the like, are not registered as road traffic crashes or road traffic crash-related injuries. This is also the case for incidents where the slipping, stumbling etc. was caused by crash avoidance maneuvers, such as trying to avoid a collision with an approaching vehicle. When considering crash involvement among child pedestrians, it is therefore important to be aware that road traffic crash only includes incidents that meet certain criteria. Second, only pedestrian crashes recorded by the police are included in the official road traffic crash statistics. Underreporting of road traffic crashes

is high, especially for vulnerable road users and less severe crashes. The degree of this underreporting differs across different countries, but in most countries, less than half of the road traffic crashes, that involve a pedestrian, are recorded by the police. In some countries, the degree of underreporting is even higher. Furthermore, recent studies indicate that the quality of the information about pedestrian road traffic crashes, recorded by the police, is sometimes low, and therefore, somewhat unreliable. Consequently, it is important to be aware that the available information about child pedestrian crashes is limited and biased in a number of ways.

Crash characteristics

Characteristics of child pedestrian crashes generally mirror their activities and mobility patterns, as well as the capacity and skills for safe road user behavior of children at different ages. Consequently, characteristics of child pedestrian crashes, involving younger children, differ from crashes involving older children, in regard to aspects such as location, road environment, and behavioral contribution to the occurrence of the crash (Petch and Henson, 2000). Regarding crash location, the majority of crashes, involving young child pedestrians, occur close to where they live. For instance, in a study from Great Britain, it was found that 70% of crashes involving a child pedestrian below the age of 5, occurred within 1 km from the child's residence. And out of these, approximately one-third of the crashes occurred right in front of their home. As the age of the child increases, the distance between the crash location and the residence of the child increases too, thereby mirroring the age-related increase in the radius of independent mobility. For older child pedestrians, the majority of crashes occur during the daily commute to/from school and leisure activities away from home. For younger child pedestrians, the crashes mostly occur to/from kindergarten or during play, either just in front of their home or very close-by. Thus, as the allowed radius of independent mobility increases with age, so does the distance between the crash location and the residence of the child.

In regard to characteristics of the road environment at child pedestrian crash locations, the prevalence of child pedestrian crashes is higher on locations with certain characteristics. In general, crash occurrence is higher on locations, where the area used by the child for walking or playing, is not separated from motor vehicles and other road users. Thus, the prevalence of child pedestrian crashes is higher on locations with limited access to playgrounds, gardens, and other green areas. At locations where such facilities are missing, children are left to play on the sidewalk or in the street, thereby increasing the risk of being hit by other road users. The number of parked vehicles is also associated with child pedestrian crash occurrence, and the crash rate is comparably high on locations with a high level of on-street parking and limited off-street parking facilities. In addition, the literature indicates a comparably high number of child pedestrian crashes on locations with long, straight, and high-speed roads, especially if they have limited pedestrian facilities and a high level of through traffic. Studies suggest that more than 90% of child pedestrian crashes occur on not dead ends streets. However, on average, the average collision speed is lower for child pedestrian crashes, compared to the average collision speed in crashes with adult or elderly pedestrians. In a retrospective Austrian survey (Mayr et al., 2003) among parents of child pedestrians ≤ 16 years old injured in a road traffic crash, it was found that 38% of the crashes happened at locations with zebra-crossings with, or without, traffic signals. Neither zebra-crossings nor signals were present at the crash location of 58% of the crashes. In addition, 37% of the pedestrians were alone at the time of the crash, parents or other adults accompanied 32%, and 24% were with other children. The majority of crashes happened during the daytime, particularly between noon and 5 p.m., and the majority of crashes happened in sunny and dry weather conditions. In general, the crash risk of child pedestrians increases with vehicle speed and traffic volume, especially on locations where the road design does not support separation of the child and the motor vehicles.

Crash Contribution

Due to the complex nature of crash occurrence, care should be taken when assessing crash causes. However, looking at the circumstances connected to child pedestrian crashes, studies indicate that the child's contribution to the crash occurrence differs depending on the age of the child. Like differences in crash location and road environment, these differences mirror the age-related differences in activities at the time of the crash. When 1–2-year-old children are involved in a road traffic crash, it is more likely to occur in a situation in which a car is backing up on a driveway or similar, while the child is playing. The child contributes to the crash occurrence merely by being present, and unfortunate behavior of the other road user—for the back-up crash, in combination with the lack of back-up cameras connected to audible warning signals—appears to be the main contributing factor to the occurrence of these crashes. From the age of around three, crash involved child pedestrians contribute more actively (although not intentionally) to the crash occurrence. For 3–9-year-old children, crashes often occur in situations in which the child plays in the street, runs out into the street from the sidewalk, or from between parked cars, or tries to cross the street at an unsafe distance from oncoming traffic. Thus, although the other road user may also contribute to the occurrence of the crash by speeding, being inattentive, being impaired by alcohol etc., the behavior of the child pedestrian also contributes to the occurrence of such crashes. This is also the case for child pedestrians aged 10–14; however, at this age the contribution of the child pedestrian also includes violations, such as not obeying a traffic signal.

Crash Consequences

Child pedestrians are vulnerable to physical injury in case of a crash, due to lack of protection from a vehicle (e.g., airbags) or personal safety equipment (e.g., helmet, protective clothing). The difference in mass, between the pedestrian and the other party, also contributes to the vulnerability. Consequently, the degree of severe injury is generally higher for pedestrian crashes, compared to

crashes among other road users, and as such, pedestrians are the most vulnerable road user group. A German study (Niebuhr et al., 2016), analyzing 1422 pedestrian–car collisions with injury, found that the severity distribution for child pedestrian crashes was different from the severity distribution for crashes involving pedestrians in other age groups. More specifically, below collision speeds of approximately 69 km/h, the risk of serious injury was lower for children below the age of 15 than for other age groups. At collision speeds above 69 km/h, the risk of being seriously injured as a child pedestrian was higher, than the risk for adult pedestrians. These results indicate that compared to pedestrians in other age groups, the risk of being injured for child pedestrians was lower at lower collision speeds but higher at higher collision speeds. Differences in force impulses, and the relation between the upper leading edge of the vehicle and the pedestrian's center of gravity may explain this. When hit by a vehicle, children are typically pushed out in front of the car, whereas adult pedestrians rotate over the vehicle. In an Austrian study, it was found that 27% of the child pedestrians injured in a road traffic crash suffered physical consequences, primarily troublesome scars, but recurring headache and pain in the lower extremities were also among the more frequently reported physical sequels. Less frequent consequences included pain in lower and upper extremities, motion restriction, tooth damage, impairment of cerebral function, and back pain. The chance of long-term physical consequences was reduced if the vehicle was braking at the time of the crash, but increased with vehicle speed and number of vehicle parts, by which the child was struck.

Traditionally, the focus is on the severity degree of the injury in terms of physical person injury. However, child pedestrian crash involvement may also have psychological consequences, such as PTSD for the child pedestrian and/or for the child's family. In the Austrian study mentioned above, 24% of the child pedestrians, injured in a road traffic crash, suffered from long-term post-traumatic behavioral and psychological disturbances, with fear of traffic being the most prominent, followed by being generally more nervous, and behaving more carefully. Other disturbances include aspects such as generally being more fearful, psychometric retardation, lack of concentration, and sleep-related disturbances. Also, 11% of families of the crash involved child pedestrians suffered from post-traumatic psychological disturbances. Fear of losing a child was the most prominent, but being more careful also occurred relatively frequently. Other aspects include feelings of emptiness, being stressed by the child's behavior, and being more contemplative. A French study on post-traumatic stress symptoms in school-aged children, injured in a road traffic crash, found, that peritraumatic distress was a robust predictor of acute PTSD symptoms, 5 weeks after the crash. Children suffering from comprised mental retardation, lifetime psychotic disorder, and life-threatening conditions were not included in the study. Of the included children, 34% were pedestrians at the time of the injury.

Cognitive Capacity and Skills

The ability to understand and predict the behavior of other road users, and to relevantly adjust one's own behavior accordingly, is a key element in child pedestrian safety. However, for child pedestrians this ability is challenged, due to age-related limitations in cognitive capacity and skills, as well as perceptual and motor abilities. While recent research indicates that human cognitive development continues until the mid-20s, some research also indicates that by the age of 10–12, children's cognitive capacity allows them to handle most traffic situations on their own, except complex ones. In short, the cognitive capacities and skills needed for safe child pedestrian behavior comprise of attention, perception, information processing, and decision-making, all of which develop with age (e.g., Schwebel et al., 2012). Small children focus spontaneously on their immediate surroundings, have difficulties controlling their attention, and even 7–8-year-old children get easily distracted in traffic. Consequently, their ability to identify relevant areas of attention, and to intentionally focus on approaching traffic, is limited. Visual and auditive information is of particular importance to child pedestrians, in order to locate other road users and anticipate traffic development. Research suggests that visual acuity is lower in children at the age of 5 compared to 8–11 year olds, as well as compared to adults. Limitations in visual acuity may, therefore, cause young children to miss critical information about safety. In regard to auditive information, the ability to locate an object develops early, whereas the ability to interpret the auditive information improves with age. However, as child pedestrians must collect and process dynamic visual information about the road environment to make safe behavioral decisions, the fact that the span of the working memory develops and improves with age is also of key importance for the cognitive processing of visual and auditive information. Due to these developmental challenges, the ability to collect and apply relevant sensory input is more challenging for younger, compared to older, child pedestrians. However, the relative contribution of limitations, related to collecting and interpreting the information, remains unclear. In regard to information processing, the child pedestrian is also challenged by age-related limitations in the ability to process and integrate multiple sources of information within a short period of time. Furthermore, the ability to predict future development in the traffic situation, and, for instance, anticipate upcoming slots for safe road crossing, is poor. Consequently, timely strategic behavioral decisions are difficult. Smaller children need more time to collect and process critical safety information than older children and adults do, and by the time the small child is ready to cross the street, it may be too late to do it safely. However, it is not clear if the process is also delayed by the process of translating the decision to cross, into actual behavior and movement of the body.

Countermeasures

Safety for child pedestrians can be improved from two directions: creating a safe environment and increasing child pedestrian safety skills. Creating a safe environment includes measures such as separating child pedestrians and motor vehicles, and reducing the speed of the motor vehicles. Examples aimed at the motor vehicles include speed bumps, moving the stop line further away from intersections and pedestrian crossings, as well as vehicle-based safety improvements, such as pedestrian friendly fronts or pedestrian

detection systems. Measures directed at the child pedestrian include measures such as overpasses and underpasses, which allow children to cross a road while spatially separated from the motor vehicles, but also other measures offering clear guidance while respecting the cognitive capacity of the child, regarding where and when to walk or play and to identify safe road crossing times and locations. However, safety measures only serve to improve child pedestrian safety when used as intended. Thus, careful consideration of the behavior, needs, and limitations of child pedestrians should be taken when implementing such measures. For example, if the location of an underpass is inconvenient, the child pedestrian is less likely to use it, and if the child perceives it as unsafe due to design issues, poor lighting, lack of overview, or other subjective perceptions, he/she may choose to cross the street not using the underpass, thereby unintentionally increasing the risk of crash involvement. In addition, signalized intersections, zebra crossings, and other measures, that provide clear guidance to the child pedestrian as where to walk or play, may provide the child with a false sense of safety, and consequently, lead it to not pay sufficient attention to approaching traffic. Thus, measures to improve child pedestrian safety should take the needs, perceptions, motivations, etc. of child pedestrians into account. In addition to designing the road environment in ways that support child pedestrian safety, child pedestrian safety can be enhanced by the safest possible usage of the existing environment. This can be done by identification of the safest routes to school, leisure activities, etc. along with parental encouragement to choose these routes. In addition, child pedestrians may benefit from the use of conspicuity aids and light-colored and retroreflective clothing, dangle tags, armbands and strips on school bags, etc.

Even though the child pedestrian's ability to engage in safe road-user behavior depends on their age-related cognitive capacity, it is possible to improve their safety skills by training. Research indicates that lack of skills and strategies, for optimal use of the actual cognitive capacity of the child, enforce the age-related limitations. Therefore, training programs targeted at improving safety skills among child pedestrians, should aim at improving the needed strategic skills, rather than focusing on knowledge or rule-based learning (Thomson et al., 2005). Thus, instead of only practicing the actual crossing of the road, training programs should aim to develop the needed critical safety skills, such as estimating time-distance, the relation between available time and needed time, and skills in terms of anticipation and planning ahead. In addition, research suggests that the presence of peers is associated with increased risk of injury for child pedestrians, and that perceptions of social norms among peers influence child pedestrian behavior. Thus, measures aimed at supporting safety enhancing social norms among peers and changing child pedestrians' perceptions of the behavioral norms of their peers, are relevant. Finally, it is relevant to encourage and prepare parents accompanying child pedestrians to use such moments, as an opportunity to teach their child about safe road user behavior, while taking the age of their child into account.

Conclusion

The risk of crash involvement and serious injury is high for child pedestrians particularly in low-income countries. Crash characteristics and circumstances change with the age of the child and mirror age-related differences in activities and mobility patterns. Factors contributing to crash involvement and injury include lack of protection from vehicle or personal safety equipment, unsafe road environments and age-related limitations in cognitive capacity and skills. Key preventive measures include spatial separation of child pedestrians and motor vehicles, reducing speeds to no more than 20 km/h of motor vehicles where separation cannot be achieved, and training focusing on safety critical strategic skills. However, measures aimed at social norms and involvement of parents are also relevant.

Biography



Mette Møller, PhD, is a Transport Psychologist. Mette Møller's main research field is road user behavior, crash analysis, crash prevention, and road safety.

References

- Mayr, J.M., Eder, C., Berghold, A., Wernig, J., Khayati, A., Ruppert-Kohlmayr, A., 2003. Causes and consequences of pedestrian injuries in children. *Eur. J. Pediatr.* 162, 184–190.
- Niebuhr, T., Junge, M., Rosén, E., 2016. Pedestrian injury risk and the effect of age. *Accid. Anal. Prevent.* 86, 121–128.
- Petch, R.O., Henson, R.R., 2000. Child road safety in the urban environment. *J. Transport Geogr.* 8, 197–221.
- Schwebel, D.C., Davis, A.L., O'Neal, E.E., 2012. Child pedestrian injury: a review of behavioural risks and preventive strategies. *Am. J. Lifestyle Med.* 6, 292–302.
- Thomson, J.A., Tolmie, A.K., Foot, H.C., Whelan, K.M., Sarvary, P., Morrison, S., 2005. Influence of virtual reality training on the roadside crossing judgments of child pedestrians. *J. Exp. Psychol.-Appl.* 11, 175–186.
- WHO, 2008a. World report on child injury prevention, World Health Organization, https://apps.who.int/iris/bitstream/handle/10665/43851/9789241563574_eng.pdf;jsessionid=B6A1D77FAE5B084B81B8D392ED5AB17D?sequence=1.

Further Reading

- Bui, E., Brunet, A., Allenou, C., Camassel, C., Raynaud, J.P., Claudet, T., Fries, F., Cahuzac, J.P., Grandjean, H., Schmitt, L., Birmes, P., 2010. Peritraumatic reactions and posttraumatic stress symptoms in school-aged children victims of road traffic accident. *Gen. Hosp. Psychiatry* 32, 330–333.
- Kovesdi, C.R., Barton, B.K., 2013. The role of non-verbal working memory in pedestrian visual search. *Transport. Res. Part F* 19, 29–31.
- Leden, L., Gärder, P., Johansson, C., 2006. Safe pedestrian crossings for children and elderly. *Accid. Anal. Prevent.* 38, 289–294.
- Methorst, R., Schepers, P., Christie, N., Dijst, M., Risser, R., Sauter, D., van Weel, B., 2017. 'Pedestrian falls' as necessary addition to the current definition of traffic crashes for improved public health policies. *J. Transp. Health* 6, 10–12.
- Morrongiello, B.A., Seasons, M., MaAuley, K., Koutsoulianos, S., 2019. Child pedestrian behaviors: Influence of peer social norms and correspondence between self-reports and crossing behaviors. *J. Saf. Res.* 68, 197–201.
- Shinar, D., 2017. Pedestrians. In: Shinar, D. (Ed.), *Traffic safety and human behavior*. 2nd ed. Emerald Publishing, United Kingdom, pp. 861–926.
- WHO, 2008b. European Report on Child Injury Prevention, World Health Organization, https://www.who.int/violence_injury_prevention/child/injury/world_report/European_report.pdf?ua=1.
- WHO, 2013. Pedestrian Safety: A Road Safety Manual for Decision-makers and Practitioners, World Health Organization, Geneva, <https://www.who.int/publications-detail/pedestrian-safety-a-road-safety-manual-for-decision-makers-and-practitioners>.
- Zeedyk, M.S., Kelly, L., 2003. Behavioural observations of adult-child pairs at pedestrian crossings. *Accid. Anal. Prevent.* 35, 771–776.

Pedestrian Safety, General

Muhammad Z. Shah*, **Mehdi Moeinaddini†**, **Mahdi Aghaabbasi†**, *Centre for Innovative Planning and Development, Faculty of Built Environment, Universiti Teknologi Malaysia, Johor, Malaysia; †Department of Urban and Regional Planning, Faculty of Built Environment, Universiti Teknologi Malaysia, Johor, Malaysia

© 2021 Elsevier Ltd. All rights reserved.

Glossary	420
Overview of Pedestrian Safety	420
The Need for Pedestrian Safety	420
Pedestrian Safety Versus Pedestrian Security	421
Causes of Accidents	422
Road User Hierarchy	424
Protection Zone for Pedestrian	424
Designing for Pedestrian Safety	425
Traffic Devices for Pedestrian Safety	427
Issues and Potential Solutions	427
Relevant Websites	428
References	428
Further Reading	428

Glossary

jaywalk The act of illegally crossing a road, either at junctions or midblock, into the traffic flow.

motorcycles Also known as motorbikes. It is a form of two-wheeler vehicle, typically using gasoline engines. It may be used to carry people as well as goods.

pathway A dedicated, paved walkway or sidewalk especially provided to separate pedestrians from motorized traffic.

pedestrian Any person, regardless of age and physical conditions, who walks either for the entire journey or part thereof.

personal mobility devices (PMD) A form of transportation that includes hoverboards, kick-scooters, and e-scooters.

Overview of Pedestrian Safety

Pedestrians are people who walk to achieve some travel goals, regardless of age and personal physical conditions. These travel goals may include, among others, commuting to school or work, shopping, recreational, sightseeing, or socializing. Walking can either be for the entire journey or part thereof. The activity may take place on a paved sidewalk or pathway on the ground, overhead bridge, underground tunnel, along a trail, or at the edge of a street or road lacking adjacent walking facilities. When pedestrians have to walk in the road or on the roadside, competing for the same road space as motorized vehicles, safety is obviously a concern. Pedestrians also include those who run, jog, sit, or lie either on the pathway or alongside a roadway. For the physically challenged people, walking may be assisted by crutches, canes, walkers, or wheelchair. Sometimes, where space is limited, or funds are not appropriated in sufficient amounts, other groups of users may share the space with pedestrians. These other groups of users include cyclists and users of personal mobility devices (PMD), such as skateboards, hoverboards, or e-scooters.

Due to different profiles of pedestrians, the mixture of users on pathways, and the need to safely share the roadway with other road users (**Photo 1**), pedestrian safety is increasingly becoming a complex issue yet one that must be tackled to ensure safe and comfortable passages to those who choose to walk. The issue of pedestrian safety may become more critical in certain cities where the city legislation allows the citizens to take legal remedy from the local government due to negligence of providing and ensuring safety along the pathway.

The Need for Pedestrian Safety

Every pedestrian has specific desires, needs, and physical limitations. For example, a disabled person in a wheelchair requires a ramp to access the pathway, whereas a senior citizen requires a longer green time to safely cross a junction due to lower walking speed.



Photo 1 Pedestrians, bicycle, and motor vehicles coexist safely in Utrecht, the Netherlands.

Risk profiles of pedestrians are different at different places and times. The risks of accidents between pedestrians and motorized vehicles are higher at a busy intersection than, for example, inside an airport terminal where only electric buggies are present. Subsequently, the risk of crossing a street may multiply many folds at night or during adverse weather. Failure to address the risk exposure of the pedestrians may consequently lead to discomfort, inconveniences, bodily injuries, or even death. Hence, it is important for city officials to plan, design, and construct pathways that guarantee not only comfort and convenience, but also safety. And, equally important, a safe walking environment will promote, motivate, and convince residents to participate actively in walking as an alternative to using private vehicles.

Pedestrian Safety Versus Pedestrian Security

The term pedestrian safety should not be confused with pedestrian security. Though both safety and security have adverse consequences to the pedestrians, they are not synonymous and are not interchangeable.

Pedestrian safety refers to situations when there are potential risks of personal injuries due to civil activities (e.g., driving or using mobile phone while walking) or due to hazards from physical obstructions and natural environments (**Photo 2**).

Pedestrian security, on the other hand, refers to risks of personal injuries due to criminal activities. Some examples of incidences leading to security risk include snatch theft, street crime, kidnapping, and terrorism as in the case when motor vehicles are used as weapons of destruction to slam into pedestrians and cyclists. Hence, protruding tree roots on the pathway is a hazard to pedestrian safety, but mugging is a risk to pedestrian security.

The remedies to mitigate both risks may be similar or different. Installing adequate and strategically placed lighting along the pathway may address both safety and security risks. However, installing a network of closed-circuit television cameras may only address security issues, but not necessarily mitigate safety risks.



Photo 2 Safety is about managing movement conflicts and physical obstructions that are risks to pedestrian safety.



Photo 3 Improper placement of fire hydrant along pathway.

Causes of Accidents

Walking, just like driving or any other form of transportation, has its associated safety risks. For pedestrians, the safety risks are extremely high and serious as pedestrians are the most vulnerable among all road users due to the absence of any form of physical protection. Understanding and profiling these risks and causes of accidents to pedestrians are imperative to city managers. Knowledge of these risk profiles helps in the formulation of strategies, action plans, and design guides to improving walkability and safety in a city. Generally, there are four factors that contribute to elevated risks to pedestrian safety—physical obstructions factors, vehicular factors, design factors, and human factors.

As pedestrians walk, they interact directly with the built and the natural environment surrounding them throughout their journey. The built environment such as sidewalks and street pavements, street furniture and buildings, and the natural environment such as trees and shrubs increase the esthetic value of the pathway as well as provide positive and enjoyable travel experiences to the pedestrians. Yet, in some cases, the built environment and the natural environment can become hazardous and turn into safety risks when they create physical obstructions. Trees become hazardous when low-hanging branches are not trimmed or when roots protruding into the sidewalk create an uneven travel surface. As for sidewalks and walking paths, they ought to be paved rather than gravel and are typically made of concrete (Stoker et al., 2015), asphalt, bricks, or tiles. If not well maintained, they may over time become cracked, broken, or loose, thus impeding smooth travel. In these circumstances, pedestrians may become injured when they accidentally trip over protruding tree roots or broken tiles. Even if accidents are avoided, pedestrians may still have to circumnavigate around the protruding tree roots, and surface bumps or holes resulting in a less than desirable walking experience. Similarly, street furniture such as lampposts or fire hydrants may be improperly placed on the walking path of the pedestrians creating man-made physical obstructions (Photo 3). For the visually impaired individuals, these physical obstructions are major hazard to their safety as well as interfere with their navigation along the pathway.

Another form of physical obstructions is commonly found around construction sites. Due to lack of guidance or enforcement by the local government and occupational safety agency, building contractors or their agents may encroach into the walkway resulting in partial or full closure of the pathway. When this happens, pedestrians are forced to use alternative paths to continue their journey by encroaching onto the roadway. As the pedestrians enter the roadway, they will directly compete for the same road space as motorized vehicles use. These conflicts between pedestrians and motorized traffic are accidents in the making. Additionally, construction sites are also a source of industrial debris that may endanger the safety of the pedestrians. Without proper roofing and scaffolding, the industrial debris may hit the pedestrians causing grievous injuries.

Motor vehicles, either moving or static, are perennial sources of hazards to pedestrians and they are ideally not meant to ever be in the same physical space. Yet, there are numerous times when conflicts happened leading to accidents that can cause serious injuries and even be fatal. Protecting pedestrian safety from the hazards of motorized vehicles means at a minimum to effectively providing an unobstructed visual line of sight for both the pedestrians and the drivers. This contributes to separating the pedestrians from the motorized traffic as well as controlling the flow of vehicles and pedestrians. One source of obstruction of the visual line of sight is parked or stopped vehicles. Parked or stopped vehicles, especially trucks, buses, trams, and other big vehicles, may prevent pedestrians from having a good, clear view of oncoming traffic. This is especially critical at an unsignalized junction lacking crosswalks where the pedestrians have to rely on visual sight to safely enter the roadway and cross the street. Similarly, drivers also need to be able to detect the presence of pedestrians attempting to cross at an intersection to be able to stop their vehicles in time and at a safe distance from the crossing pedestrians. As for moving vehicles, they may skid and encroach into the pathway possibly crashing into the pedestrians. In cities with temperate climate, skidding vehicles are a fairly common phenomenon during the winter season due to slippery road surfaces and vehicles may slide sideways onto sidewalks as well. For such locations, protecting pedestrians from skidding vehicles poses a greater challenge requiring specific measures including promoting on-street parking as a form of barrier to the pedestrians.

In other situations, speeding vehicles and vehicles driven by intoxicated drivers encroaching into pedestrian pathways are even harder to control as their occurrences are hard to predict, though preventable. Yet, there have been many cases where irresponsible drivers have killed pedestrians. And, in many jurisdictions, vehicles are allowed to travel at speeds that are not compatible with



Photo 4 Motorcycle using pedestrian pathway as alternative route.

pedestrian movements. When top speeds are kept below 30 km/h, and preferably below 20 km/h, pedestrians are typically able to avoid getting hit by vehicles. However, parallel movement of vehicles in multiple lanes can pose a risk to pedestrians even at relatively low speeds since vehicles obstruct pedestrians from each other as discussed earlier. So, the possibility of these risks on pedestrian safety should not be discounted or accepted.

Another form of hazards posed by motorized vehicles is when vehicles illegally enter the pathway and use it as an alternative route (**Photo 4**). Motorcycles (e.g., scooters and mopeds), due to their small size, may find the pathways tempting due to lack or absence of motorized traffic there, whereas the travel lanes are congested. For these motorcyclists, using the pathway as an alternative route then can save significant traveling time especially along highly congested roads. Their presence on the walkways, hence, poses serious danger to the pedestrians. Where demand of parking spaces exceed supply, some cities find that motor vehicle owners use and treat the pedestrian pathway as an additional source of supply of parking spaces, albeit illegally (**Photo 5**). In Asian cities such as Hanoi, Vietnam, and Pune, India, where motorcycles (e.g., scooters and mopeds) are a major form of transportation, the problem of encroachment into pedestrian pathways is a major issue for city officials. When moving, these motorcycles may hit pedestrians and when parked they present a physical obstruction to pedestrian mobility.

Poor land-use planning and design is another risk factor to pedestrian safety. The absence of pedestrian pathways in residential neighborhood is an example of bad design. Over the course of designing residential neighborhood layouts, town planners may put form and esthetic over function. Spaces are generously provided for landscaping, yet nothing is allocated for the provision of pedestrian pathways. This will ultimately create a highly car-dependent community, as it is too dangerous to walk around in the neighborhood. Children will then not walk to school but be taken by their parents in private vehicles. Even a short trip to the neighborhood shops will be done using private vehicles. As private vehicles become a life necessity, every adult family member will own a private vehicle, or two. With no adequate space to park their vehicles within the residential compound, the vehicles will overflow onto the streets thus escalating the risk of walking.

Similarly, the placement of land uses with high walking demand potentials, for example, schools and commercial complexes, along a busy road or next to a busy intersection, not only creates access management nightmares but also induces safety risks to



Photo 5 Parked motorcycles on pathway obstructing pedestrian movement.

pedestrians. Pedestrians who consistently are exposed to high traffic volumes make crossing the road, either at a junction or midblock, a risky and time-consuming affair, especially in jurisdictions where automobile drivers do not have the obligation to yield to pedestrians. When pedestrians assume that it will take significant amounts of time to wait for traffic to ease before attempting to cross, they become increasingly willing to take high risks and attempt crossing in short time gaps and in unguided ways—which is in fact a gamble with safety and their life. Lastly, poor design may also contribute to disconnected pedestrian networks.

Lastly, the behavior of pedestrians themselves may invite safety risks. A new phenomenon called the “zombie” is when a pedestrian is too engaged with their mobile devices (e.g., reading, texting, talking, playing games, etc.) while walking. These “zombies” are oblivious of their surroundings including potential hazards and movement conflicts. The “zombies” may bump into other pedestrians, trip over protruding tree roots, or fall into uncovered stormwater drains. Another self-destructive pedestrian behavior is jaywalking when pedestrians consciously walk into streets where motorized traffic is given the legal right-of-way. In most jurisdictions, pedestrians do not have the right-of-way in-between intersections (midblock) unless there is a marked crosswalk. In some jurisdictions, pedestrians do not have the right-of-way at intersections either unless there is a marked crosswalk. And even when the pedestrians clearly have the right-of-way, drivers may intentionally or unintentionally fail to give priority to pedestrians. So, jaywalkers as well as legally crossing pedestrians often risk their lives in return for their own and motorists’ convenience and time savings.

Road User Hierarchy

There are strategies and actions that can be implemented to mitigate the risks to pedestrian safety. These strategies and actions are normally included in citywide pedestrian master plans or pedestrian facilities design guidance. The master plan and design guidance are published after a series of detailed engineering and land-use studies: demography, travel behavior and mobility analyses, and community consultations to account for specific local issues, cultures, as well as planning and urban design philosophies. In preparing the master plan or design guidance, some cities adopt well-known principles such as the universal design, complete street, and safe city principles.

Regardless of the design principles used, the basic underlying concept is the concept of road user hierarchy. This concept establishes the priority level of every road user. In this hierarchy (Fig. 1), pedestrians typically should be given the highest priority, followed by cyclists, transit vehicles (e.g., buses and trams), and motorcycles. Private cars and trucks/lorries should be at the lowest end of the hierarchy. Adopting this hierarchy requires that city officials give guarantees, in their design, planning philosophies, and development processes, that pedestrians are given the top priority to the roadway use and all other road users must give priority to the convenience, comfort, and safety of pedestrians.

Protection Zone for Pedestrian

As the most vulnerable group of all road users, pedestrians must be protected from all potential hazards. To provide for this protection, a zone surrounding an individual pedestrian must be established (Fig. 2). The desired dimension of this protective zone is 2.0 m wide by 3.0 m high. Where space is limited, the width can be reduced to a minimum of 1.5 m.

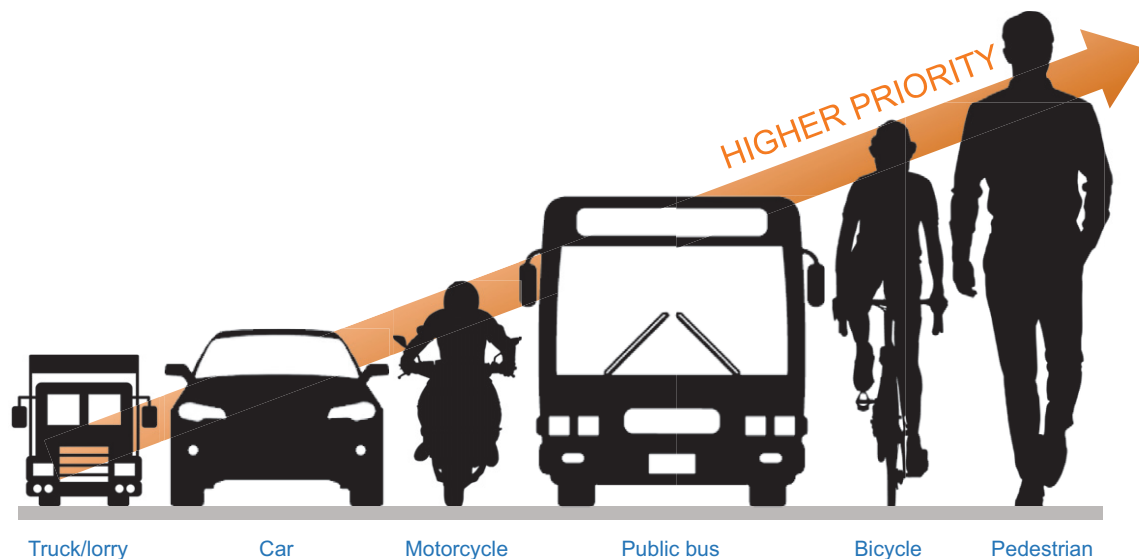


Figure 1 Hierarchy of road users.



Figure 2 Pedestrian protection zone.

Within this protective zone, there should not be any physical element (e.g., tree branches, signs, street furniture, parked bicycles, etc.) that would present a hazard obstructing safe passage of pedestrians. The establishment of this protective zone will significantly reduce the risk to pedestrian safety.

Designing for Pedestrian Safety

The main cause of accidents involving pedestrians is conflict between pedestrians and motorized vehicles. Through effective segregation, or effective integration—such as the use of shared space—conflicts between these two groups are preventable (Huybers et al., 2004). Segregation can be achieved through spatial separation acting as a buffer between pedestrians and motorized vehicles (Fig. 3), creating the spatial separator effectively dividing the right-of-way into three zones—the pedestrian zone, the buffer zone, and the traffic zone.

The effectiveness of the buffer zone as spatial separator can be enhanced through planting of trees acting as a natural physical barrier. Trees also bring the added benefits of natural shades as well as increasing the esthetics of the surrounding areas (Photo 6). However, the choice of trees must be properly selected as protruding tree roots are potential hazards to pedestrian safety. Other than as a place to plant trees, the buffer zone can also be a convenient place to install street furniture, for example, benches, fire hydrant, wayfinding signage, or streetlights, as well as a place to install public utilities.

Segregation can also be achieved through having physical barriers, for example, fences or curbs (Fig. 4). Similar to spatial separation, the effectiveness of the physical barrier can be enhanced by combining various tools, for example, combining curbs and fences. Fences, especially non-climbable ones, are effective measure to prevent, or at least discourage, jaywalking.

For the best solution, combining both spatial and physical separation is the most effective measure of separating the pedestrians from the motorized traffic (Fig. 5).

The minimum width for the pedestrian zone (i.e., pedestrian pathway) is 1.5 m. This minimum width is sufficient for two individuals to walk side by side and for easy maneuverability of wheelchairs for people with physical disabilities. At locations where the pedestrian volume is high, this minimum width is inadequate and should be increased to a minimum of 2.0 m. However, if the

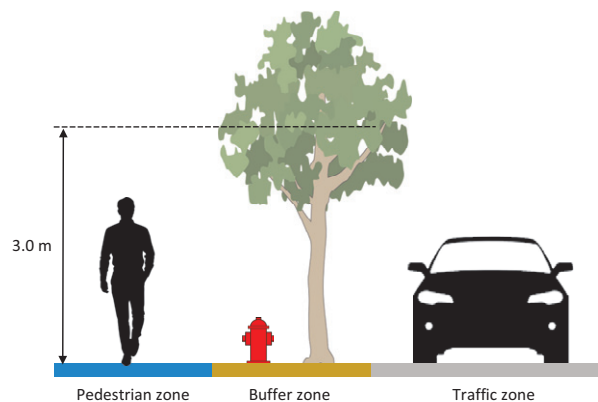


Figure 3 Buffer zone as spatial separator.



Photo 6 Combination of landscape as spatial separator and shades providing pedestrian with comfort and safety.

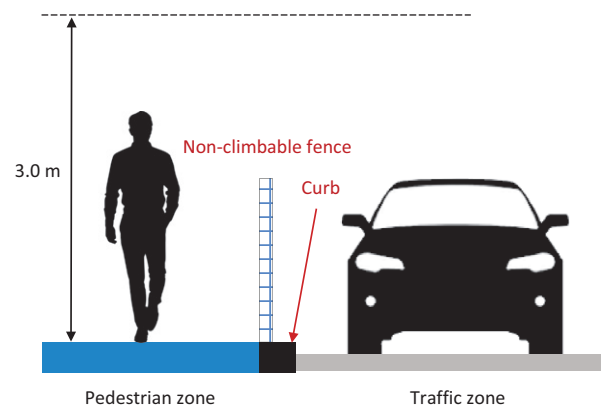


Figure 4 Fence and curb as physical separator.

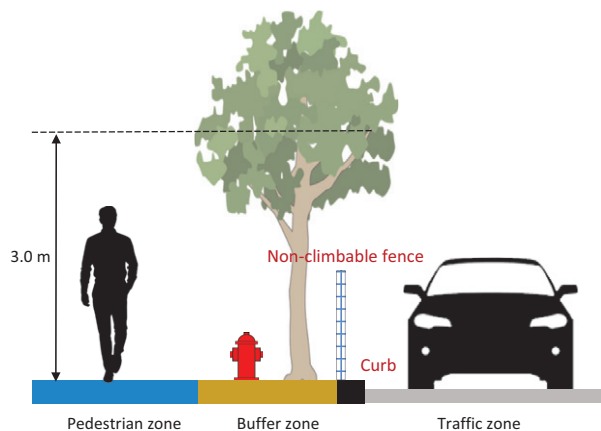


Figure 5 Combining spatial and physical separators.

pathway is to be shared with bicycles or PMD (e.g., e-scooter), then a minimum width of 2.5 m is desirable to ensure safe passing and to avoid unnecessary conflict on the pathway. Where a barrier curb is provided, the desirable height should not be less than 15 cm. As for the fence, a minimum height of 1.0 m is required to effectively discourage climbing over the fence. Some fences, however, have already incorporated non-climbing feature into its design which is preferable on streets where the probability of jaywalking is high, and it has been determined that vehicles should be given uninterrupted passage. To reduce the effect of hardscape of the fence, planting of shrubs is recommended (Photo 7).



Photo 7 Effective use landscape to soften the effect of fence.

Traffic Devices for Pedestrian Safety

Another major source of safety risk for pedestrians is where they cross at nongrade separated intersections where conflicts between pedestrians and motorized vehicles are imminent (Miranda-Moreno et al., 2011). Here, spatial and physical separations discussed earlier are not possible. Instead, separation can be achieved through effectively regulating flows of pedestrians and motorized traffic. Controlling and regulating pedestrian and traffic flows at cross-junctions can be achieved through using traffic devices such as traffic signals and signs. For the drivers, both signals and signs provide them with visual cue of impending conflicts with pedestrians as well as instructions on their vehicles' movements. Similarly, signals provide instructions to the pedestrians on when it is reasonably safe to enter the roadway to cross to the other side, especially if speeds are kept low—not above 30 km/h—so that drivers are not encouraged to run red lights near the change intervals. Combined with alarm beacons, pedestrian crossing signals provide the visually impaired audible cue to cross the streets.

Traffic calming devices can also be an effective tool to enhance pedestrian safety. These devices, for example, speed breaker and speed humps, force drivers to slow their vehicles when approaching pedestrian crossings. The speed table, a form of speed hump with a flat top surface, may double as crossing path for pedestrians and, if the ramp is not too short and not too long, and its height is at least 10 cm, they reduce speed effectively. In some locations, chokers and road diets (i.e., narrowing of streets) may be combined with speed tables not only to further enforce lower traveling speeds, but also to provide a convenient and safe crossing facility for the pedestrian. However, the suitability of each potential traffic calming measures must be extensively studied before implementation so as not to hinder smooth passage of emergency service vehicles such as ambulances and fire engines.

Issues and Potential Solutions

Ensuring safety of pedestrians is a conscious, disciplined effort and it requires deep understanding of the built environment, demography, and travel behavior (Asadi-Shekari et al., 2015; Chimba et al., 2014). Only then careful design and planning of pedestrian facilities that incorporated safety can be performed. This is a huge challenge where providing pedestrian walkways, and subsequently pedestrian safety, is not considered a priority or even somewhat important by decision-makers.

Generally, the lack of focus on pedestrian convenience and safety is more prevalent in developing and underdeveloped countries where cities are still struggling with motorization. Often, priority and resource allocation are assigned more to alleviating motor-vehicle congestion and improving travel time for motorists than providing comfortable and safe pedestrian pathways. Some cities in Africa and Central Asia, for example, are notorious for their unsafe walking environments as dedicated pathways for pedestrians are absent from most part of the city. In these cities, walking means being in the same space, and competing for the same space, with motorized vehicles. Here, people walk, not because of choice, but rather because they have to as there are limited economically viable mobility alternatives. On the contrary, in developed countries, most of their cities have already given priority to pedestrians to encourage active mobility as a means to discourage and limit motorization. Cities such as Berlin, Copenhagen, Boston, Melbourne, Paris, Kyoto, and Singapore are notable for their walkability. Here, locals and tourists walk not only because it is convenient and comfortable, but also because they feel safe doing so. Hence, the first hurdle to protecting the safety of pedestrians is to provide a dedicated space that separates the pedestrians from the motorized traffic. The next step is to narrow down the part of the streets that is open to motorized traffic and widen sidewalks in connection with lowering speed limits like Berlin has done where much of the downtown now has a 10 km/h speed limit.

A solution to guarantee the provision of pedestrian pathways is by making it mandatory for any property/land-use developer to include pedestrian pathway in any newly proposed land-use development. However, this mandatory requirement must be incorporated as a rule for development submission, not merely as a set of best practices or even standards. In other words, failure to

include provision of pedestrian pathway in the development submission will automatically render the submission noncompliant and be rejected. This rule, however, is easily incorporated into existing development law if both city officials and property developers understand the importance of providing proper facilities to support walking and sustainable development. However, in a profit-first and profit-maximization society, property developers might resent such development law requiring mandatory provision of pedestrian pathway and facilities. The developers will view such pedestrian-centric laws as costly and effectively reduce their profit margin. In such a society, city officials must be creative and resourceful to convince and educate the property developers that providing pedestrian pathways and facilities will work for them, rather than against them, by increasing the value and demand of their development. Their brand will also benefit from providing pedestrian facilities as the developers will be seen as community-friendly and sustainability-conscious.

The issue of education is not unique to property developers. The communities themselves must be educated and trained on proper behavior while walking and the importance of observing safety (Pei-Sung Lin et al., 2019). A list of dos and don'ts while walking helps to educate and remind the communities of the different type of hazards on the pathway that they may come across as well as their associated risks. The list can be in the form of pamphlets or even posted as sign along the pathway. Notwithstanding the list, the best form of education to encourage safe walking practices is when proper pedestrian knowledge and skills are integrated into the school education curriculum and system. Unfortunately, the incorporation of this subject is difficult to implement as it requires concerted effort from many parties and agencies. Teachers must be trained, curriculum must be redesigned, skills and behaviors must be assessed, and so many other issues surrounding formal education and training must be taken care of. Due to the complexities of school curriculum design/redesign, the education and training on pedestrian safety have taken the informal route. Some corporations, as part of their corporate social responsibility (CSR) programs, have taken the tasks and responsibilities onto themselves to provide the education that is so critical for the communities, especially for children and senior citizens. The success of such CSR programs may be enhanced with the participation of nongovernmental organizations and social activists. When such programs exist, sufficient incentives (e.g., subsidies, tax reliefs, etc.) must be provided by the local government, not only as a form of formal recognition but also to encourage more of such programs.

Unfortunately, there is a clear possibility that no amount of education, motivation, and promotion is sufficient to encourage or induce behavioral change. In this situation, local government should redesign streets to make them safe with existing pedestrian behavior and, if that is not practical, should introduce strict enforcement through effective policing to reprimand offenders who flaunt safety rules—be it pedestrians or motorists. For examples, motorcyclists who fail to observe no parking rules on pedestrian pathway may have their motorcycles confiscated by the city and the owners reprimanded by paying fines. Similarly, pedestrians who jaywalk should be fined as a form of hard education to force behavioral change. If we want to reach vision zero, as discussed in a different article in this encyclopedia, we need to combine engineering measures with enforcement and education. Not least is education needed to change people's attitudes to give priority to safety over mobility.

All the earlier issues on ensuring pedestrian safety are complex issues that consistently plague city officials. Nonetheless, they are not without solutions nor are they impossible to achieve. These issues can be solved with active cooperation and participation of all stakeholders for the mutual benefits of the city. Also, to be successful, there must be a strong political will to ensure that cities are safe for its people to walk.

Relevant Websites

Federal Highway Administration, US Department of Transport. Available from: <https://www.fhwa.dot.gov/>.
National Association of City Transportation Officials. Available from: <https://nacto.org/>.
Pedestrian and Bicycle Information Center. Available from: <http://www.pedbikeinfo.org/>.

References

- Asadi-Shekari, Z., Moeinaddini, M., Shah, M.Z., 2015. Pedestrian safety index for evaluating street facilities in urban areas. *Saf. Sci.* 74, 1–14.
- Chimba, D., Emaasit, D., Cherry, C.R., Pannell, Z., 2014. Patterning Demographic and Socioeconomic Characteristics Affecting Pedestrian and Bicycle Crash Frequency. Transportation Research Board 93rd Annual Meeting, Washington DC.
- Huybers, S., Houten, R., Malenfant, J.E., 2004. Reducing conflicts between motor vehicles and pedestrians: the separate and combined effects of pavement markings and a sign prompt. *J. Appl. Behav. Anal.* 37, 445–456.
- Miranda-Moreno, L., Morency, P., El-Geneidy, A., 2011. The link between built environment, pedestrian activity and pedestrian-vehicle collision occurrence at signalized intersections. *Accid. Anal. Prev.* 43 (5), 1624–1634.
- Pei-Sung, L., Rui, G., Bialkowska-Jelinska, E., Kourtellis, A., Zhang, Y., 2019. Development of countermeasures to effectively improve pedestrian safety in low-income areas. *J. Traffic Transp. Eng.* 6 (2), 162–174.
- Stoker, P., Garfinkel-Castro, A., Khayesi, M., Otero, W., Mwangi, M.N., Peden, M., Ewing, R., 2015. Pedestrian safety and the built environment. *J. Plan. Lit.* 30 (4), 377–392.

Further Reading

FHWA, 2009. Manual on Uniform Traffic Control Devices for Streets and Highways. US Federal Highway Administration, Washington, DC.
Smart Growth America, 2019. Dangerous by design. Available from: <https://smartgrowthamerica.org/resources/dangerous-by-design-2019/>.
NACTO, 2012. Global Street Design Guide. Island Press, Washington, DC.
WHO, 2013. Pedestrian Safety: A Road Safety Manual for Decision-Makers and Practitioners. World Health Organization, Geneva.

Pedestrian Safety, Older People

Carlo Luiu, The Institute for Global Innovation, University of Birmingham, Birmingham, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	429
Decline in Health Condition	429
Sensory Impairments	429
Physical Impairments	430
Cognitive Impairments	430
Built Environment Form and Design	431
Road Crossing Behavior	432
Explorative Behavior at the Crossing Environment	432
Gap Judgment and Acceptance	432
Curb Delay	432
Crossing the Road	433
See Also	433
References	433
Further Reading	434

Introduction

Walking should be considered as a valid solution to reduce the variety of problems raised by the modern car-oriented society. It is a green transport option due to no air and noise pollution involved, useful to tackle traffic congestion and parking issues, and overall more affordable and reliable than motored modes. As an active transport mode, for older people walking is at the same time a transport option and a recreational activity that provides physical activity, with consequent benefit to individuals' health and quality of life (O'Hern and Oxley, 2015; Tight, 2016).

Nonetheless, walking is not an easy activity to undertake during later life. Older pedestrians are sometimes referred as extra vulnerable road users, since all pedestrians are vulnerable to injury, and older people are even more so, due to their greater frailty and longer recovery time compared to younger people (O'Hern and Oxley, 2015). Indeed, they are over-represented in crashes at intersections, crashes when crossing at mid-sections of roads, especially on high-speed and wide or multi-lane ones (Oxley et al., 2004). Statistics from EU 27 countries show that the percentage of pedestrian fatalities is higher for those aged 65 years old and above than for any other age group, despite the number of fatalities decreased from 3459 to 2595 (25%) between 2007 and 2016. The high risk is particularly valid for those aged 80–84 years old, as the number of fatalities peaks for this age group. Moreover, the pedestrian fatality rate of older people is above the average and steadily increases from the age of 70 up to over 85 (European Commission, 2018). The United States present a similar situation, with 20% of all pedestrian fatalities (1165 out of 5903) that happened in 2017 involved people aged above 65 years old. Male older pedestrians aged 80 years old and above present the single highest fatality rate by age groups and gender at 4.55 pedestrian fatalities per 100,000 population, while the rate for the overall 65+ population is 2.29 (NHTSA, 2017).

Factors posing a risk to older pedestrians can be grouped into three main categories: (1) decline in health conditions, (2) built environment form and design, and (3) road crossing behavior.

Decline in Health Condition

It is recognized that mobility in later life is influenced by progressive changes associated with decline of health conditions. Unlike other transport modes (e.g., public transportation and car) that in some aspects can compensate for health impairments, walking can be more directly affected by deteriorating health functions (Siren and Hakamies-Blomqvist, 2005). Overall, decline in health affects sensory, physical, and cognitive functions.

Sensory Impairments

Sensory impairments are related to the loss or decline of visual and hearing functions. Older people experience these deteriorations as a consequence of advance aging, although, with considerable variation between individuals due to the heterogeneity that characterizes the older population. In general, sensory impairments are associated with the risk of falling, reduced perception of fixed and moving objects, problems in detecting approaching vehicles, and difficulties in distinguishing vehicles from other aspects of the road environment (Dunbar et al., 2004; Tournier et al., 2016).

Visual skills are required to perform some of the tasks associated with walking, such as navigating and movement detection. However, older people are more likely than other age groups to suffer from pathological age-related eye diseases, such as glaucoma, cataract or macular degeneration. Glaucoma affects peripheral vision, and the person experiencing it may be unaware of it, and difficulty in adapting to darkness. Pedestrian risks associated with this disease affect road crossing due to difficulties in detecting objects/obstacles not located straight ahead, poor adaptation to dark and detection of object/obstacles in such context, leading to a higher risk of falling and poorer self-reported mobility (Dunbar et al., 2004; Tournier et al., 2016). Presence of cataract affects dynamic and static visual acuity, contrast sensitivity, and difficulties in seeing in poor light. Suffering from cataract leads to difficulties in road crossing because of reduced ability in detecting objects, both fixed and moving, from the rest of the road. Moreover, this disease was found to be associated with risk of falling, self-reported difficulty in making distance judgments, depth perception, and navigating in dark environments (Tournier et al., 2016). Finally, macular degeneration causes blurred or distorted central vision and scotomas, with potential risk of falling, particularly in reduced light environments (Dunbar et al., 2004). Another visual impairment not linked to pathological age-related changes affecting pedestrian safety is the deterioration of visual motion sensitivity, which is found to affect road crossing decisions due to misperception of approaching vehicles at high speed (Tournier et al., 2016).

Hearing loss is frequent with advance aging, particularly for the older people aged 75 years and above. Hearing functions play an important role in pedestrian safety in terms of spatial detection and localization of objects and vehicles in the road environment. Older pedestrians with hearing impairment have difficulties at detecting approaching vehicles coming from behind them or while turning. Furthermore, hearing loss was found to affect sense of balance and consequently to increase the risk of falling, in addition to poor general mobility, physical health, and self-perceived sense of security (Dunbar et al., 2004).

Physical Impairments

Physical impairments are associated with age-related changes and decline in muscles strength and mass, bones frailty, and mobility of joints. Changes in strength and mass of leg muscles affect overall mobility in terms of walking speed, ability to maintain balance, and consequently, increased risk of falling (Oxley et al., 2004). In general, older people walk slower than younger people. Lower speed implies increased time required to cross a road and overall time to undertake a given journey (Asher et al., 2012). Similarly, physical impairments reduce head rotation movements while exploring the road environment before crossing a road, with increased risk of misjudging the evaluation of approaching vehicles.

The decline of leg strength reduces the ability to keep balance and to deal with losing balance circumstances. Indeed, older people react to loss of balance differently compared to younger cohorts and are less able to quickly recover their balance (Dunbar et al., 2004; Tournier et al., 2016). This is particularly valid when they have to cope with a secondary task while walking, or in case of suffering from sensory impairments at the same time. Therefore, older pedestrians are significantly exposed to risk of falling and the consequence of falling can be particularly serious, as they are more likely to be affected by fracture and need more time to recover than younger people (Dunbar et al., 2004). In this sense, suffering from low mass density and deterioration of bones can increase the consequences associated with falling (Dunbar et al., 2004; Tournier et al., 2016). Osteoporosis is a common systemic skeletal disease during later life, particularly among older women, associated with fractures and therefore risk of falls. Furthermore, osteoporosis is a cause of kyphosis, a postural disease that exaggerates the anterior curvature of the thoracic spine. Kyphosis affects general mobility due to the particular poor postural condition and increases the risk of falls and fractures. Moreover, it reduces the ability to avoid obstacles and climb stairs, together with visual exploration while crossing a road (Tournier et al., 2016). Likewise, arthritis is another common impairment among older people caused by inflammation of joints. Arthritis affects walking by restricting range of movements, and consequently balance, and by reducing strength and physical endurance (Oxley et al., 2004).

Cognitive Impairments

Cognitive functions also decline with advancing age. Typical age-related shortfall involves reduced performances in memory, speed, and accuracy, in addition to difficulties in spatial processing, planning, and selecting attention (Dunbar et al., 2004; Oxley et al., 2004; Tournier et al., 2016). Decline in selecting attention, executing function, and slower processing performances affect significantly multi-tasking operations. Selecting attention is a function necessary to bring information into focus and hinder inconsistent or unnecessary information. In this sense, older pedestrians report significant difficulties in dealing with situations requiring attention between more than one task, switching attention, visual search, and ignoring not useful information, especially while in movement. Risk of falling and difficulties in obstacle negotiation were found associated with loss of it (Dunbar et al., 2004). Similarly, decline in executing function was found to impact gait control and multi-task coordination, with consequent increase of risk of falling. To avoid so, older pedestrians compensate for their walking behavior by reducing their walking speed or avoiding potential distractions (e.g., talking with other people). Problems in multi-tasking operations are cause of increased risk of being involved in an accident at intersections, due to the complexity of junctions compared to other situations (Dunbar et al., 2004; Oxley et al., 2004; Tournier et al., 2016). Slow processing speed is among the factors contributing to difficulties in navigation. Moreover, it affects the decision-making process associated with road crossings by increasing curb delays (see Road Crossing Behavior section for further information).

Another necessary walking skill affected by cognitive impairments is the ability to spatially navigate in a determined environment. In general, navigation skills decline while aging and older people perform less efficiently than younger people in this sense

do. As a consequence, older pedestrians tend to commit more mistakes in finding way and take longer time in completing their walk. Decline in spatial processing and loss of memory are the two main functions impacting this skill. They both contribute in reducing the ability to navigate and orientate, particularly in an unknown context. Furthermore, older pedestrians have lower and less accurate performances in route learning and sequencing compared to younger people (Dunbar et al., 2004; Klencklen et al., 2012).

Built Environment Form and Design

The form and design of the built environment can also significantly affect older pedestrian safety. The reliance on private transport modes has contributed to the design of living environments around vehicular movement rather than human ones, with consequential difficulties for those relying on active transport modes (Matan and Newman, 2012). Urban sprawl and community severance are two phenomena affecting walking consequences of urban planning design. The former produces a dispersion of services and activities beyond a reasonable walking distance, while the latter creates a divisive effect on residential areas affecting the easiness to cross a road, street connectivity, walking ability, accessibility, and overall walking quality (Anciaes et al., 2016). In this sense, difficulties in crossing a road and associated risks of doing it are accentuated by the larger sections of road to cover. Older pedestrians are not able to cross safely because of their lower walking speed and therefore not being able to cross the road until an automobile reach their crossing point. Moreover, community severance creates a psychological barrier affecting walking quality by increasing self-perceived risk of being involved in a collision due to traffic volumes, speed, noise, and pollutant emissions associated with this phenomenon (Anciaes et al., 2016).

The design of the walking environment can also be a barrier for walking and a factor concerning older pedestrians' safety. Handling sidewalks is considered one of the most difficult tasks for older pedestrians, particularly with regard to obstacle negotiation and related risk of falling issues. This is particularly valid for those crossing environments that are not leveled and have steps (Tournier et al., 2016). The poor-quality design of footpaths and sidewalks regarding the presence of steps and pavements materials, such as the use of pebble stones or cobblestones, and general uneven and broken pavements, are all factors increasing the presence of obstacles while walking (Tournier et al., 2016). Similarly, sidewalks or nearby parallel walking paths are fundamental in order to separate pedestrians from the driving environment, but the presence of not lowered curbs at both beginning and end of footpaths can add further complications to road crossing tasks (Mitra et al., 2015; Rosenbloom, 2009; Wang et al., 2016). The presence of cars and scooters parked on, or obstructing, sidewalks, and cleanliness and lack of maintenance of footpaths are also associated with problems in obstacle negotiation and the risk of falling (Chen et al., 2015). Moreover, in contexts such as Northern countries, the presence along the pathway during the winter season of snow, ice, and mud resulting after both melt, is an additional barrier and factor of risk (Hjorthol, 2013; Rosenbloom, 2009).

Stumbling over an obstacle is considered one of the most common causes for falling. In a similar way, the task of avoiding an obstacle suggests an elevated level of risk, especially for those affected by cognitive impairments due to multi-tasking operations involved. Under such conditions, typical strategies that older pedestrians engage to negotiate with obstacles involve reducing walking speed and step length, keeping a large distance between themselves and other pedestrians and spending more time looking at their footsteps rather than straight ahead or for traffic (Dunbar et al., 2004; Oxley et al., 2004; Tournier et al., 2016). Design solutions can significantly facilitate walking and reduce the risk of being involved in accidents or getting injured. In this sense, implementations have to be particularly focused on reducing vehicle and pedestrian conflicts, especially those associated with crossing the road. Examples of such improvements include raised crosswalks, speed cushions about 10 m before the crosswalk, median islands, curb extensions, improved user-activated signal crossing devices such as puffin crossing with time regulated by sensors. Crosswalks should preferably never extend across more than one lane without having islands for resting, especially if top speeds are above 30 km/h.

The poor design of the built environment can also affect older pedestrians' self-perceived safety and older people tend to be particularly sensitive regarding this matter. Older people tend to avoid walking during particular times, such as during the night or generally when it is dark, or in areas of cities that are perceived as dangerous. A lack of adequate street lighting and the presence of dark areas contribute to preventing older people from walking, especially for the potential presence of people—either groups or individuals—hanging out in the streets (Mitra et al., 2015; Wang et al., 2016).

Another factor related to road design affecting older pedestrians' self-perceived safety is the suggestion of sharing the walking environment with other road users, especially with motor vehicles. Apart from the risk of being involved in car accidents, older pedestrians perceive that drivers fail to acknowledge the rights of other road users. Typical examples in this sense are represented by people parking cars and motorbikes on the sidewalks and car doors opening without paying attention to passing pedestrians. These negative behaviors are exacerbated in the case of narrow sidewalks (e.g., obstructing the passage) or when the elevation of the sidewalk is not sufficient to discourage car parking (Oxley et al., 2004). Finally, older pedestrians report safety problems in shared cycling/pedestrian environments. Dedicated shared cycle and walking lanes are a well-acknowledged solution to improve the safety of cyclist and pedestrians, and overall to incentivize active travel. However, striping may not be sufficient to keep cyclists out of the walking part and it is therefore common to use smooth pavement for the bicycle part and paving stones or similar material for the pedestrian part. This is also not a good solution since it leads to hard and uneven walking surfaces as discussed above. Nonetheless, continuity of cycle lanes and footpaths and problems of space invasion are perceived as a factor of risk, particularly regarding bicycle speed by younger users (Oxley et al., 2004; Vine et al., 2012).

Road Crossing Behavior

Crossing a road is among the situations in which older pedestrians face high risks to be injured or killed, as the most dangerous accidents happen while undertaking this task. Crossing a road is a complex situation involving several stages (Oxley et al., 2004). Older pedestrians have to identify a suitable place to cross the road while approaching the curb and they may have to explore the crossing environment by scanning potential approaching vehicles. Then, even though they, in most cases have the right-of-way, have to judge traffic gaps, and determine the right time to leave the curb to cross the road.

Explorative Behavior at the Crossing Environment

Not having seen an approaching vehicle that hit them while crossing a road is among the most reported causes by older pedestrians of road accidents (Dunbar et al., 2004). Consequently, identifying a suitable crossing point while approaching the curb is fundamental for older pedestrians in order to explore the road environment before crossing safely (Tournier et al., 2016). The majority of the studies investigating pedestrian crossing reveal that young and older people present a similar explorative behavior at crossing points in terms of number of head movements to check for traffic. This was found particularly valid in presence of traffic signals. Overall, older people tend to look slightly more often than younger people, especially when crossing roads with two-way traffic. However, they were found to make a smaller rotation of their head due to physical impairments, and therefore assess less efficiently approaching vehicles (Dunbar et al., 2004). In absence of traffic signals, older pedestrians tend to be more thoughtful and have a more conservative approach. The evaluation of road crossing opportunities tends to decline with advancing age and older pedestrians are likely to pause more at the curb before crossing and more often turn their heads in visual exploration, both before and while crossing the road (Oxley et al., 1997).

Gap Judgment and Acceptance

Gap acceptance is traditionally defined as the distance of a vehicle in movement from the pedestrian at the time of the first step while undertaking the action of crossing the road (Oxley et al., 1997). Where automobile drivers have the right-of-way by legal code, or de facto take that right, gap judgment by pedestrians is a necessary skill for older pedestrians in order to identify opportunities to cross the road in a safe way. Gap judgment is a particularly complex skill because older pedestrians have to take into account vehicles' speed and distance and time of arrival to the other side of the road, in addition to their own walking speed (Oxley et al., 2004). It is therefore imperative to change the behavior of drivers so that, at least at marked crosswalks, it is the automobile drivers that look for gaps in between pedestrian before they proceed rather than the traditional assumption that drivers have priority to all space.

Older pedestrians perform less safe and less efficient than younger people in terms of making gap judgments (Dunbar et al., 2004). Indeed, they, on average, do not accept short gaps as their younger counterparts, but at the same time, they take longer to cover the distance needed to cross the road. Thus, they have a tighter fit between crossing completion and the vehicle crossing their path, with consequent reduction in safety margins. Furthermore, older pedestrians have more difficulties in evaluating speed and they therefore more frequently use distance of an approaching vehicle as a determinant for when stepping off the curb (Oxley et al., 1997). Nevertheless, they also have more difficulties evaluating distances than younger people. Hence, they feel safer and prefer accepting gaps where the approaching vehicle is further away, but at the same time they may underestimate the speed of the vehicle, implying that it will arrive later than expected and accept gaps that are objectively dangerous (Oxley et al., 1997). The solution to improving safety is to secure that vehicle speeds are low. Today, many jurisdictions allow crosswalks on roads with up to 50 and even 60 km/h speed limits. Where older pedestrians or children may cross a street, crosswalks should not be marked if top speeds are above 20 km/h as that can lead to the earlier-mentioned unfavorable situation that it is pedestrians that look for gaps between cars rather than the opposite. If speeds are above that, signalization or stop signs can be used.

Curb Delay

Curb delay is an important factor explaining risk when jaywalking is occurring or even at crosswalks when automobile drivers take the right-of-way away from pedestrians. It can be defined as the time since when the last vehicle passed a pedestrian waiting to undertake the first step toward the road to be crossed (Oxley et al., 1997). Overall, older pedestrians tend to leave the curb later than younger people, with time delay identified in just under 1 s. In contrast, younger people were found to leave the curb slightly before the passage of the last vehicle (Oxley et al., 1997). Three different reasons can explain this approach difference. As previously mentioned, older people tend to be more cautious and have a more conservative attitude within the road environment. In this sense, they are more likely to pause more time at the curb, especially in case it is dark or if they have to cross a road with more than one lane (Dunbar et al., 2004). Another explanation is that older pedestrians might adopt an inefficient strategy to cross the road compared to younger people. As younger people were found to not wait for the last vehicle to pass before starting to cross the road, this might imply that they perform better in considering opportunities to cross the road (Dunbar et al., 2004; Oxley et al., 1997). Finally, increase in curb delay might be associated with slower time reaction, decision-making, and motor coordination. Crossing the road is a complex task and decline in physical and cognitive skills can reduce responding time performances, especially in complex traffic situations (Oxley et al., 1997).

Crossing the Road

Let's start with looking at jaywalking pedestrians or pedestrians crossing where drivers do not yield to them. Once older pedestrians have visually investigated the road environment and accepted a reasonable gap between them and approaching vehicles, they are ready to cross the road. Oxley et al. (1997) identify two different type of crossing styles can be identified based on the approach undertaken while crossing the road. Non-interactive crossers are identified as those pedestrians employing a safer approach and waiting until the road is clear in both directions to cross in one single action. In contrast, at locations that have more than one lane from the curb to refuge island, interactive crossers employ a less safe approach and negotiate their crossing by pausing, usually midway, modifying their walking speed and weaving through traffic. In their turn, interactive crossers can be further classified into three sub-groups: those who interact with the nearside traffic only, those who interact with the far side traffic only, and those who interact with both directions of traffic while crossing.

Older pedestrians usually belong to the interactive group, particularly by interacting with the far side of the road. Adopting this approach highlights again the complexity that crossing a road has for older pedestrians. In this light, they might encounter multi-tasking difficulties in evaluating the traffic in both directions and quickly react and adapt to the traffic conditions. As a consequence, some older pedestrians consider crossing the first half of the road without concern of the remaining half, reducing their safety margins significantly (Oxley et al., 1997). Reduction in walking speed is another factor characterizing interactive crossing. Older pedestrians might start crossing while the road is clear, but due to low walking speed, the road turns busy by the time they reach the middle of it (Dunbar et al., 2004; Oxley et al., 1997).

Decrease in walking speed can also affect crossing a road in presence of a traffic light. Several older pedestrians report difficulties and anxiety associated with the feeling of not being able to complete the crossing within the allocated time (Tournier et al., 2016). Indeed, the walking speed internationally used to base pedestrian crossing time is currently 1.2 m/s. Several studies report that older pedestrians struggle to reach such speed and findings suggest that reducing the average speed to 0.8 m/s might be more appropriate and would allow them to cross the road with higher safety (Asher et al., 2012; Oxley et al., 2004). It should also be noted that in most jurisdictions, any road user who enters on green has the right of way over perpendicular traffic until they have cleared the intersection. However, if a road has multiple lanes, a slow-walking pedestrian may not be visible to drivers and a driver may by mistake start up when the signal turns green for them. Educating drivers that they have to look for pedestrians—and not enter the intersection—in such situations is another part of ensuring safety.

In conclusion, it is of utmost importance that streets and roads where older pedestrians cross at level are kept narrow and have low speeds. Preferably, all roads should be built in such a way that older people and children can walk safely along them and cross them without having fear of getting hit by vehicles. If streets and roads are built that way, people of all age groups and with different handicaps will be reasonably safe.

See Also

Automobile accidents and passive prevention systems; Bicycle collision avoidance systems; Elderly driver safety issues; Age and Gender as factors in road safety; Inequality and traffic safety; Pedestrian safety, children; Pedestrian safety, general; Pedestrian Safety, visually impaired; Transport modes and aging society; Transport modes and health

References

- Anciaes, P.R., Jones, P., Mindell, J.S., 2016. Community severance: where is it found and at what cost? *Transp. Rev.* 36, 293–317.
- Asher, L., Aresu, M., Falaschetti, E., Mindell, J., 2012. Most older pedestrians are unable to cross the road in time: a cross-sectional study. *Age Ageing* 41, 690–694.
- Chen, Y.J., Matsuoaka, R.H., Tsai, K.C., 2015. Spatial measurement of mobility barriers: improving the environment of community-dwelling older adults in Taiwan. *J. Aging Phys. Act.* 23, 286–297.
- Dunbar, G., Holland, C.A., Maylor, E.A., 2004. Older pedestrians: a critical review of the literature. Available from: www.dft.gsi.gov.uk.
- European Commission, 2018. Traffic safety basic facts on pedestrians. Available from: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/statistics/dacota/bfs20xx_pedestrians.
- Hjorthol, R., 2013. Winter weather—an obstacle to older people's activities? *J. Transp. Geogr.* 28, 186–191.
- Klencklen, G., Després, O., Dufour, A., 2012. What do we know about aging and spatial cognition? Reviews and perspectives. *Ageing Res. Rev.* 11, 123–135.
- Matan, A., Newman, P., 2012. Jan Gehl and new visions for walkable Australian cities. *World Transp. Policy Pract.* 17 (4), 30–41.
- Mitra, R., Siva, H., Kehler, M., 2015. Walk-friendly suburbs for older adults? Exploring the enablers and barriers to walking in a large suburban municipality in Canada. *J. Aging Stud.* 35, 10–19.
- NHTSA, 2017. 2017 data: older population. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/82684>.
- O'Hern, S., Oxley, J., 2015. Understanding travel patterns to support safe active transport for older adults. *J. Transp. Heal.* 2, 79–85.
- Oxley, J., Fildes, B., Ihlen, E., Charlton, J., Day, R., 1997. Differences in traffic judgements between young and old adult pedestrians. *Accid. Anal. Prev.* 29, 839–847.
- Oxley, J., Corben, B., Fildes, B., O'Hare, M., Rothengatter, T., 2004. Older vulnerable road users—measures to reduce crash and injury risk. Report No. 218. Monash University Accident Research Centre, Clayton.
- Rosenbloom, S., 2009. Meeting transportation needs in an aging-friendly community. *Generations* 33, 33–43.
- Siren, A., Hakamies-Blomqvist, L., 2005. Sense and sensibility: a narrative study of older women's car driving. *Transp. Res. Part F Traffic Psychol. Behav.* 8, 213–228.
- Tight, M., 2016. Sustainable urban transport—the role of walking and cycling. In: *Proceedings of the Institution of Civil Engineers-Engineering Sustainability*. Thomas Telford Ltd., pp. 87–91.

- Tournier, I., Dommes, A., Cavallo, V., 2016. Review of safety and mobility issues among older pedestrians. *Accid. Anal. Prev.* 91, 24–35.
- Vine, D., Buys, L., Aird, R., 2012. Experiences of neighbourhood walkability among older Australians living in high density inner-city areas. *Plan. Theory Pract.* 13, 421–444.
- Wang, Y., Chau, C.K., Ng, W.Y., Leung, T.M., 2016. A review on the effects of physical built environment attributes on enhancing walking and cycling activity levels within residential neighborhoods. *Cities* 50, 1–15.

Further Reading

- Abou-Raya, S., ElMeguid, L.A., 2009. Road traffic accidents and the elderly, John Wiley & Sons. *Geriatr. Gerontol. Int.* 9 (3), 290–297, doi:10.1111/j.1447-0594.2009.00535.x.
- Crabtree, M., Lodge, C., Emmerson, P., 2015. A review of pedestrian walking speeds and time needed to cross the road. Available from: <https://trid.trb.org/View/1378632>.
- Ding, T., et al., 2015. Psychology-based research on unsafe behavior by pedestrians when crossing the street. *Adv. Mech. Eng.* 7 (1), 203867, doi:10.1155/2014/203867.
- ITF, 2018. Road safety annual report. Available from: www.itf-oecd.org/road-safety-annual-report-2018.
- Noh, Y., Kim, M., Yoon, Y., 2018. Elderly pedestrian safety in a rapidly aging society—commonality and diversity between the younger-old and older-old. *Traffic Inj. Prev.* 19 (8), 874–879, doi:10.1080/15389588.2018.1509209.
- Oxley, J., Corben, B., Fildes, B., Charlton, J., 2004. Older pedestrians: meeting their safety and mobility needs. In: Road Safety Research, Policing and Education Conference. Road Safety Council of Western Australia, Perth, Australia.
- Oxley, J.A., Ihsen, E., Fildes, B.N., Charlton, J.L., Day, R.H., 2005. Crossing roads safely: an experimental study of age differences in gap selection by pedestrians. *Accid. Anal. Prev.* 37, 962–971.
- WHO, 2013. Pedestrian safety: a road safety manual for decision-makers and practitioners. Available from: https://apps.who.int/iris/bitstream/handle/10665/79753/9789241505352_eng.pdf?sequence=1.

Visually Impaired Pedestrian Safety

Robert S. Wall Emerson, Department of Blindness and Low Vision Studies, Western Michigan University, Kalamazoo, MI, United States

© 2021 Elsevier Ltd. All rights reserved.

Historical Perspective	435
Orientation and Mobility	435
Challenges to Independent Travel	436
Risk in the Modern Travel Environment	436
References	438
Further Reading	438

Historical Perspective

Blind pedestrians or visually impaired have been traveling independently for millennia. For much of this time, the two primary tools available to assist in their travel were the help of other people, often as guides, or the use of some sort of walking stick. In the 1920s, dog guides became a mobility option for people who are blind and since the 1940s; the walking stick tool has been formalized as the modern white cane (sometimes referred to as a long cane and previously referred to as a typhlocane). While traveling with a sighted partner can allay many safety concerns, many blind people are not able to or prefer not to rely on other people for their travel needs. So, it is often necessary for them to travel independently. To do so, a range of tools (beyond using a dog guide or long cane) and techniques have been developed to reduce risk and increase safety.

People who are blind have been able to become accomplished independent travelers even before the advent of modern tools. James Holman, who lived from 1786 to 1857, is an excellent example of the degree of independent travel possible by someone who is blind, even under the most difficult circumstances. At a time when most of the blind people were expected to remain at home and be cared for by others, James Holman traveled the world alone, becoming an explorer, adventurer, and writer and eventually traveling to every inhabited continent. However, for most people who are blind, the scope and range of independent travel has greatly increased with the advent of modern tools, infrastructure, and instruction of travel techniques.

Orientation and Mobility

In the United States, there are approximately 65,000 legally blind children and youth and approximately 7.7 million adults over the age of 18 who have some sort of vision loss (APH, 2016; Erickson et al., 2017). When referring to a person as being “legally blind,” this means that the person’s vision is 20/200 or worse (where 20/20 is typical or normal vision) or they have a severely constricted field of view. Not all of these 7.7 million people would require the use of a dog guide or cane to get around. For example, only about 2% of blind people (about 10,000 people) use a dog guide. However, any degree of vision loss will impair a person’s ability to access information, such as street signs or to respond quickly to threats in the environment, such as vehicles running red lights at an intersection.

The set of techniques and strategies taught to pedestrians who are blind to facilitate independent travel is called orientation and mobility (O&M). The field of O&M has developed travel techniques, often expanding on natural tendencies and abilities of the people who are blind. For example, a powerful tool for blind people is using reflected sound to get a sense of what is around them. This is a skill that can be taught and refined, especially if a person becomes blind later in life and has never had to develop this skill before. The primary mobility tool, the long cane, has been improved through the use of new materials and the development of specialized tips. However, in terms of mobility tools, a greater advancement has been the development of GPS technologies and phone-based apps that allow blind people to access more information about the world around them and use this information to guide them in their travel.

Some of the main impacts on the mobility for people who are blind or with low vision is that they will tend to travel more slowly, they often need to access information tactually or through hearing (which might be less precise), they may experience more stress and anxiety when traveling, and there is an increased chance that quickly moving features of the environment will increase risk for them. For example, a signal that changes quickly at an intersection leaving pedestrians without enough time to cross the road; and vehicles that speed through an intersection, leaving little chance for auditory detection. Unfortunately, there are very few statistics on how often people who are blind are involved in accidents. In 2017, 5,977 pedestrians were killed in the United States (NHTSA, 2018). It is estimated that there are at least 70 legally blind pedestrians involved in accidents annually. Although the act of traveling without vision seems inherently more risky than traveling with vision, there are evidences that pedestrians who are legally blind are less likely to be killed or hospitalized as a result of being hit by a vehicle than a sighted pedestrian. This is most likely because there are fewer blind pedestrians, they tend to travel less, and they often use transportation modes, such as paratransit that limit exposure to the potential for serious conflicts with vehicles. Note, however, that these accident statistics are based on pulling information from

databases where disability status is not always noted. Further, blind pedestrians might be at increased risk for minor injuries from vehicles that would not show up in databases, such as the Fatality Analysis Reporting System (FARS) in the United States.

Challenges to Independent Travel

With a greater range of useful and far reaching tools for assisting in the safe travel, two of the greatest limiting factors to safe travel for blind pedestrians are public sentiment and the travel environment. Sighted parents are understandably concerned about the safety of their blind children but this can sometimes lead to over protection and a lack of sufficient movement and exploration when children are young. Lack of exploration when young can lead to lesser degrees of independent travel when older. Blind children need to be encouraged to move and explore, the same as any other child. As they grow, they will be introduced often through the work of professionals, such as O&M instructors, an array of tools and techniques that will support expansion of their independent mobility. But the blind child who is allowed to explore his or her environment will also develop their own compensatory strategies for getting around as they develop a sense of the larger world around them. Nevertheless, there are parents or general members of society who do not think people who are blind can or should travel on their own. Many people who are blind will chafe at these attitudes and have to deal with members of society that do not understand how independent they can be. Given that this is still a common perception, many aspects of the world and built environment are not designed with the intent of accessibility by all people with disabilities. As such, the greater challenge is learning how to effectively navigate the modern travel environment.

Risk in the Modern Travel Environment

The modern travel environment is full of fast moving vehicles, streams of people, and an array of signs, signals, and obstacles that might not be easily perceived or dealt with by someone who is blind. Traditionally, in the urban built environment, traffic patterns have been a consistent and useful cue used by blind pedestrians to guide their travel. Consistent traffic movements can remind a blind pedestrian how close they are to the street, whether they are approaching an intersection, and what sort of traffic control exists. One- and two-way traffic can also help indicate a pedestrian's position within a network of streets. Also, blind pedestrians use the sudden surge of parallel traffic beside them at an intersection as a strong cue for when they can cross a street (Fazzi and Barlow, 2017). Consistency and regularity of environmental characteristics and traffic patterns have been foundational in limiting the risk for blind pedestrians. Being able to predict what will happen in an environment and what different characteristics and events mean allows a blind pedestrian to more effectively deal with other aspects of the travel environment. For example, a pedestrian will encounter an assortment of poles, signs, trashcans, newspaper boxes, and other "street hardware" when walking along a downtown street. Many of these elements change from time to time and are certainly not consistent block to block. But this level of uncertainty introduces low levels of risk for a person using a cane or dog guide since the pedestrian remains on the sidewalk.

When inconsistency exists in curb height, the presence and placement of wheelchair ramps, block length, driveway interruptions to the sidewalk, the presence of medians, and especially the presence of actuated intersection control, risk is increased and safety is reduced for blind pedestrians. When wheelchair ramps first became widespread in the United States, it was noted that blind pedestrians began walking into the street without being aware of it. Tactile walking surface indicators, known as detectable warnings in the United States, were devised and installed at the bottom of wheelchair ramps in order to warn blind pedestrians that there was a transition from the walking surface to the road surface (Barlow et al., 2010). However, as ramps and detectable warnings have become more prevalent, many pedestrians have developed the misperception that detectable warnings indicate where a person should stand in order to cross the street. While this is often true, it is not always true. An apex ramp on a given intersection corner (where the ramp is placed in the middle of the corner, between two crosswalks, facing diagonally into the intersection) is one case where detectable warnings on the ramp would not be where a pedestrian should stand in order to cross either street. Further, misunderstandings of the intent of detectable warnings have led to incorrect installation of them by city engineers. Blind pedestrians cannot take a line of direction from the set of truncated domes used as detectable warnings. Therefore, aligning a rectangular section of these truncated domes with a crosswalk does nothing to help with a blind person's alignment but might hinder an accurate perception of the true boundary between the walking surface and the road surface.

One situation, where tactile warning surfaces are of particular importance, is on the edges of transit platforms. Whether in a metro system, an elevated train system, or a regular train system, when a mass transit vehicle enters a station where passengers get on and off the cars from a station platform, the edge of the platform needs to be clearly demarcated with a tactual warning (Architectural and Transportation Barriers Compliance Board, 2016). Truncated domes work well for this warning. Since subway or train cars are always situated at the same spot when a train pulls into a station, some mass transit systems have included tactual walking surface indicators to lay out tactually where pedestrians should line up in order to board the next train. In this way, both sighted and blind pedestrians are organized so that confusion is reduced and safety is increased.

While inconsistency in the built environment has increased risk and reduced safety for blind pedestrians, a recent change that has had perhaps the largest negative impact on safety has been the advent of actuated intersections. Historically, a blind pedestrian could approach an intersection and, since the traffic patterns tended to be regular and consistent, listen to these patterns and soon have a good idea of when an appropriate time to cross the street would be. However, the addition of more traffic movements, such as protected left turn phases and combinations of protected turning and straight through movements at intersections increases the level

of difficulty for a person who cannot see to understand the traffic patterns simply by listening. Larger intersections with multiple traffic movements can often trick blind pedestrians into thinking that they are hearing a parallel surge of traffic (Barlow et al., 2010). For example, at large intersections, a vehicle coming toward a pedestrian from the far parallel lanes and turning left to cross the pedestrian's path can sound like a vehicle that is going to continue moving on a parallel path since it must travel so far forward before beginning its left turning movement. If the intersection is actuated so that the length or presence of a given traffic movement depends on what traffic exists, then the traffic patterns might be very different from one time of the day to another or even from one traffic cycle to the next. This level of complexity and uncertainty increases the risk for blind pedestrians.

To some extent, this uncertainty can be reduced by the provision of accessible pedestrian signals. These are audible indicators of when the visual walk signal is lit. If an intersection is outfitted with accessible pedestrian signals, then this gives more information to the blind pedestrian about when an appropriate time to cross exists. However, it does not eliminate risk entirely. Uncertainty regarding traffic patterns would still exist and now a pedestrian would have to determine whether a pushbutton is at the intersection to activate an accessible pedestrian signal, then find the button and realign to make their crossing. Further, any inconsistency in how accessible pedestrian signals and pushbuttons are installed on different intersection corners can add to the uncertainty and confusion experienced by a blind pedestrian.

It seems that many modern developments in urban design end up making travel more difficult for blind pedestrians. Roundabout intersections can be problematic since the rounded walkways and roadways often make finding a crossing place harder (Barlow et al., 2010). Once a crossing place is found, the lack of traffic control means that pedestrians can only cross the street in a gap in traffic or in front of a yielding vehicle. However, the acoustic surroundings often make it hard to determine whether a vehicle is approaching, how far away it is, and if a vehicle has yielded. While smaller roundabout intersections are often crossable with training and perhaps minor engineering modifications, newer intersection designs, such as the diverging diamond interchange have recently introduced entirely new challenges to accessibility for blind pedestrians. While it would greatly reduce risk for all pedestrians to limit all roadways at uncontrolled crossings to one lane in each direction, with a refuge island between the lanes, such a situation is unlikely to become the norm in the United States where traffic flow is often given priority over pedestrian safety.

Another feature of the modern travel landscape that has increased risk for blind pedestrians is the advent of quieter vehicles. While many people think of this issue as relating to hybrid and electric vehicles, many internal combustion engine vehicles have also become so quiet that blind pedestrians can have difficulty in hearing whether a vehicle is near them or what that vehicle is doing. Blind pedestrians rely largely on listening to traffic to maintain a sense of position and safety, especially in urban areas. Reducing the sound level of individual vehicles makes those vehicles much more likely to surprise a pedestrian, potentially in a catastrophic manner. This is especially true when there are louder vehicles helping to mask the sound of the quieter vehicles. There are efforts being made to add sound to hybrid and electric vehicles so that they are not entirely silent but little is known about the best sounds to add and the impact of these sounds on the abilities of blind pedestrians to hear and react appropriately to vehicles equipped with the added sounds.

However, even with sounds added to quieter motor vehicles, there will remain a potential threat to blind pedestrians from bicycles. Bicycles generally are very quiet but can move very rapidly and often share space with pedestrians. Cyclists operating on sidewalks will sometimes weave between pedestrians, but can also have an expectation that pedestrians will move out of their path, if given warning, such as from a bell. A blind pedestrian, however, suddenly hearing a bell quickly approaching from an unanticipated direction, will not be able to move out of the way so quickly. Where there are bicycle lanes on the side of the roadway, cyclists often have a greater expectation of the right of way in that lane. This means that a quickly approaching cyclist may have trouble stopping for a blind pedestrian stepping out in front of their path as the pedestrian crosses the street. Since bicycle lanes are often not demarcated with a tactual indication, a blind pedestrian might not be aware of the potential threat. Some cities are experimenting with a slightly raised edge to the sides of bicycle lanes so that bicycles can easily ride over the boundary but blind pedestrians can also feel that there is a boundary they need to be aware of.

An upcoming change to the modern travel environment is the increase in automated and connected vehicles. There is very little known about how this movement will impact the mobility of blind pedestrians. Much of it depends on how automated and connected vehicles are introduced. If a given municipality introduces system-wide use of connected vehicles including buses, ride share, and other mass transit systems, and if that municipality designs the system so that individuals with disabilities are able to interact with the system effectively, then the automated and connected vehicle movement might very well improve mobility for blind pedestrians. If automated vehicles are designed in a way that would allow operation by a blind person, the opportunity for that person to get into a vehicle and go where they want, when they want, would greatly improve their mobility options (Owens et al., 2019). However, although much has been touted about how automated vehicles will remove the human error factor in driving and make vehicle-pedestrian interactions safer, there has not been evidence thus far that this utopia is easily achieved. The level of technological sophistication that would be needed to allow a blind person to operate a vehicle safely through city streets seems to be many years in the future. That being said, current technology, such as automatic braking and pedestrian detection should improve safety for all pedestrians, including those who are blind. As the principle cause of pedestrian fatalities is a failure of drivers to yield, having more vehicles that will brake automatically for pedestrians should improve the general safety landscape. It is interesting that these developments are occurring at a time when the United States is seeing an increase in pedestrian fatalities.

In general, increasing consistency in the travel environment will tend to increase safety for blind pedestrians. As long as attention is made to making public spaces accessible, if the modifications are consistent, pedestrians are able to figure out the rest. An event where this becomes critical is during times of emergency. When other people are panicking or there is smoke or fire or loud noises, the rush of people and ensuing chaos can play havoc with any pedestrian's orientation. A blind pedestrian can be at a greater

disadvantage in this situation, as emergency measures are not made accessible to them. Flashing lights do not help a blind pedestrian. Also, verbal announcements on a loudspeaker that are not linked to landmarks that make sense to a blind pedestrian are not helpful. In general, having emergency lights and audible signals that lead people along egress routes have been shown to be the most helpful in getting all pedestrians, including those with visual impairments, out of harm's way and to safety in an emergency.

References

- American Printing House for the Blind (APH), 2016. Annual report 2016: Distribution of eligible students based on the federal quota census of January 3, 2015. Available from: <https://www.aph.org/federal-quota/distribution-of-students-2016/>.
- Architectural and Transportation Barriers Compliance Board, 2016. Americans with Disabilities Act (ADA) Accessibility Guidelines for Transportation Vehicles. US Department of Transportation, Washington, DC. Available from: <https://www.federalregister.gov/documents/2016/12/14/2016-28867/americans-with-disabilities-act-ada-accessibility-guidelines-for-transportation-vehicles>.
- Barlow, J.M., Bentzen, B.L., Sauerburger, D., Franck, L., 2010. Teaching travel at complex intersections. In: Wiener, W.R., Welsh, R.L., Blasch, B.B. (Eds.), *Foundations of Orientation and Mobility, Volume II: Instructional Strategies and Practical Applications*. AFB Press, New York, NY, pp. 352–419.
- Erickson, W., Lee, C., von Schrader, S., 2017. Disability statistics from the American Community Survey (ACS). Cornell University Yang-Tan Institute, Ithaca, NY.
- Fazzi, D.L., Barlow, J.M., 2017. *Orientation and Mobility Techniques: A Guide for the Practitioner*, second ed., AFB Press, New York, NY.
- National Highway Traffic Safety Administration (NHTSA), 2018. Traffic safety facts: Research note. National Center for Statistics and Analysis, Washington, DC. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812603>.
- Owens, J.M., Sandt, L., Habibovic, A., McCullough, S., Snyder, R., Wall Emerson, R., Varaiya, P., Combs, T., Feng, F., Yousuf, M., Soriano, B., 2019. Automated vehicles & vulnerable road users: Envisioning a healthy, safe, and equitable future. *Proc. Road Vehicle Automation 2019*, San Francisco, CA July, 2018.

Further Reading

- Bentzen, B.L., Barlow, J.M., Bond, T., 2004. Challenges of unfamiliar signalized intersections for pedestrians who are blind: Research on safety. *Transp. Res. Rec.* 1878, 51–57.
- Bentzen, B.L., Barlow, J.M., Franck, L., 2000. Addressing barriers to blind pedestrians at signalized intersections. *ITE J.* 70, 32–35.
- Bourquin, E.A., Wall Emerson, R., Sauerburger, D., Barlow, J.M., 2014. Conditions that influence drivers' yielding behavior in turning vehicles at intersections with traffic signal controls. *J. Vis. Impair. Blind.* 108, 173–186.
- Bourquin, E.A., Wall Emerson, R., Sauerburger, D., Barlow, J.M., 2018. Conditions that influence drivers' behaviors at roundabouts: Increasing yielding for pedestrians who are visually impaired. *J. Vis. Impair. Blind.* 112, 61–71.
- Chanana, P., Paul, R., Balakrishnan, M., Rao, P., 2017. Assistive technology solutions for aiding travel of pedestrians with visual impairment. *J. Rehabil. Assist. Technol. Eng.* 4, 1–16.
- Hogan, C., 2008. Analysis of blind pedestrian deaths and injuries from motor vehicle crashes, 2002–2006. Available from: <https://files.meetup.com/211111/analysis-blind-pedestrian-deaths.pdf>.
- Khosravi, S., Beak, B., Larry Head, K., Saleem, F., 2018. Assistive system to improve pedestrians' safety and mobility in a connected vehicle technology environment. *Transp. Res. Rec.* 2672, 145–156.
- Kim, D.S., Emerson, R.W., Naghshineh, K., Myers, K., 2014. Influence of ambient sound fluctuations on the crossing decisions of pedestrians who are visually impaired: implications for setting a minimum sound level for quiet vehicles. *J. Vis. Impair. Blind.* 108, 368–383.
- Roberts, J., 2006. *A Sense of the World: How a Blind Man Became History's Greatest Traveler*. Harper Collins Publishing, New York, NY.
- Salamati, K., Schroeder, B., Roupail, N.M., Cunningham, C., Long, R., Barlow, J., 2011. Development and implementation of conflict-based assessment of pedestrian safety to evaluate accessibility of complex intersections. *Transp. Res. Rec.* 2264, 148–155.
- Stoker, P., Garfinkel-Castro, A., Khayesi, M., Otero, W., Mwangi, M.N., Peden, M., Ewing, R., 2015. Pedestrian safety and the built environment: a review of the risk factors. *J. Plan. Lit.* 30, 377–392.
- Wiener, W.R., Welsh, R.L., Blasch, B.B. (Eds.), 2010. *Foundations of Orientation and Mobility, Volume 1: History and Theory*. third ed. AFB Press, New York, NY.
- Wiener, W.R., Welsh, R.L., Blasch, B.B. (Eds.), 2010. *Foundations of orientation and mobility, Volume 2: Instructional Strategies and Practical Applications*. third ed. AFB Press, New York, NY.
- Zegeer, C.V., Carter, D.L., Hunter, W.W., Richard Stewart, J., Huang, H., Do, A., Sandt, L., 2006. Index for assessing pedestrian safety at intersections. *Transp. Res. Rec.* 1982, 76–83.

Photo/Video Traffic Enforcement

Charles M. Farmer, Research and Statistical Services, Insurance Institute for Highway Safety, Ruckersville, VA, United States

© 2021 Elsevier Ltd. All rights reserved.

What is Photo/Video Traffic Enforcement?	439
How Does Photo Enforcement Work?	439
The History of Photo Enforcement	439
The Effects on Driver Behavior	440
The Effects on Roadway Crashes	441
Best Practices	442
References	442

What is Photo/Video Traffic Enforcement?

Photo/video enforcement refers to the use of cameras to help enforce traffic regulations. It allows for the detection of violations without the need for human observers at the site. It is a useful supplement to standard enforcement because traffic volumes exceed the availability of officers to enforce traffic regulations. In addition, in congested areas and at intersections, there may be no place to safely pull over violators.

Examples of photo/video enforcement include photo radar for speed enforcement, red light cameras for traffic signal enforcement, virtual weigh stations for commercial vehicle weight enforcement, and automated highway toll systems. The technology can also be used to detect drivers who block intersections or fail to stop at a stop sign, drive past a stopped school bus, or disobey a railroad-crossing signal (Decina et al., 2007).

How Does Photo Enforcement Work?

Violations can be detected by sensors in the roadway, sensors on roadside architecture, and sensors within the vehicle. Most speed cameras measure the speed of a vehicle at a single spot. If a vehicle is traveling faster than a predetermined speed, the date, time, location, and speed are recorded along with photos or videos of the vehicle. Section control is an enhancement of photo speed enforcement that measures average speeds over a certain distance. Cameras located at two or more points record all vehicles that pass them. Automatic license-plate recognition is used to match individual vehicles so that average speeds between the two points can be calculated. Section control has been used to enforce speed limits in Australia, Austria, Italy, Netherlands, Norway, and the United Kingdom.

Red light cameras automatically photograph vehicles that go through red lights. The cameras are connected to the traffic signal and to sensors that monitor traffic flow. The camera captures images of any vehicle that doesn't stop within a given time after the light switches to red. Cameras record the date, time of day, time elapsed since the beginning of the red signal, vehicle speed and license plate.

Speed-on-green/red light cameras can detect both vehicles running red lights and vehicles speeding through intersections. They have been used in Belgium, Canada, and the United Kingdom.

The History of Photo Enforcement

The first speed enforcement camera was manufactured by a Dutch company Gatsometer BV, in 1964. They followed up with a red light enforcement camera 2 years later. By 1973, limited speed camera programs had been tested in Switzerland and Germany, and limited red light camera programs were established in Israel and Sweden.

A trial run of speed cameras began in Victoria, Australia in 1985. The community of Paradise Valley in the United States began using a single speed camera in September of 1987. Speed cameras came to Canada and Norway in 1988. By 1999, there were speed camera programs in Finland, Hong Kong, the Netherlands, Sweden, and the United Kingdom, and by 2009 they were common throughout the Western Europe (Delaney et al., 2005).

In 1979, several red light cameras were installed in Perth, Australia. Singapore began a program in 1986. Red light cameras were installed on the London ring road in 1990, and New York City launched its red light camera program in 1994. Red light cameras became especially prevalent in the United States. By 2011, there were more than 500 communities in the United States with red light camera programs.

Surveys conducted in jurisdictions with photo enforcement programs show a majority of drivers support them.

A 1991 survey of drivers in Victoria, Australia reported 79% support for the speed camera program. In the United Kingdom, 70% of drivers polled in 2001 said they favored speed cameras. Support was highest for cameras in school zones and at locations with a history of frequent crashes. In the Netherlands, 75% of drivers expressed support for automatic speed enforcement. About 70% said they would stick to the speed limit if enforcement occurred a few times per year, and almost all said they would stick to the speed limit if it were enforced every week.

In the US state of Arizona, 63% of drivers surveyed prior to the start of automated enforcement said speed cameras should be used on an urban freeway where camera enforcement was planned. After speed cameras were operational, 77% of drivers supported their use. A 2017 US national survey of drivers, ages 16 and older indicated that 48% supported the use of speed cameras on residential streets.

A 2001 Canadian national survey reported 65% support for the use of speed cameras in school zones, but only 39% support for speed cameras on highways. Women and older people were more likely to support photo enforcement.

A 2011 survey of drivers in 14 US cities with red light camera programs reported a 66% approval rate. A companion survey in Houston, a city that had just cancelled its red light camera program, found support for cameras still at 57% (McCartt and Eichelberger, 2012). However, 28% of those surveyed in Houston strongly opposed the cameras compared with only 18% of those in the other cities.

Enforcement cameras can be either hidden (i.e., covert) or in full view (overt). Surveys of drivers in New Zealand, where both types have been used, found that 21%–27% thought the cameras should be hidden while 15%–26% thought they should be in full view.

Photo enforcement programs have faced controversy in a number of countries. Common complaints are that photo enforcement is stricter than standard police enforcement, that there is no opportunity to explain the circumstances of the offense, and that the goal of the program is to make money rather than to improve safety. Such complaints have led to loss of support in some communities, resulting in either any change to the programs or cancellation of the programs.

The Effects on Driver Behavior

As with other forms of enhanced traffic enforcement, photo enforcement changes driver behavior, at least while the enforcement is in effect. In 1973, when speed cameras were put in place on a relatively dangerous section of the German autobahn—with a speed limit of 100 km/h—speeds in the left lane decreased by about 20 km/h. Within 6 months of speed cameras being introduced in Victoria, Australia, the number of drivers exceeding the 60 km/h speed limit by more than 15–20 km/h decreased by 50%. A similar reduction was seen for drivers exceeding the speed limit by more than 30 km/h. In British Columbia, Canada, the 1996 speed camera program was associated with a 3% reduction in mean speed on a section of roadway with an 80 km/h limit.

Photo speed enforcement has been shown to reduce speeding on a wide variety of roads. Speed cameras were tested on 2-lane rural roads in the Netherlands in 1991. The proportion of drivers exceeding the 80 km/h limit was reduced from 38% to 11%. The mean speed declined from 78 to 73 km/h and the 85th percentile speed declined from 87 to 79 km/h.

In 2001, a speed camera program was initiated in Washington, DC. On urban streets with speed limits of 25–30 mi/h, mean speeds declined by 14%, and the proportion of vehicles exceeding speed limits by more than 10 mi/h declined by 82%. In 2006, speed cameras were temporarily installed along a 65 mi/h highway in the US state of Arizona—the proportion of vehicles exceeding speed limits by more than 10 mi/h declined by 88%. Mean speeds declined by 5–7 mi/h while the cameras were in place, but rebounded after the cameras were removed.

A 2010 review examined 35 studies from various countries (Wilson et al., 2010). The authors concluded that speed cameras—including fixed, mobile, overt, and covert devices—cut average speeds by 1%–15% and the percentage of vehicles traveling above the speed limits or designated speed thresholds by 14%–65% compared with sites without cameras. There is some evidence, however, that drivers slow down at camera sites, then speed up after passing the sites (sometimes called kangaroo driving) (Hoye, 2014).

A study of automated section speed control on an 80 km/h road in Norway reported a reduction in the mean speed from 77 to 74 km/h in the direction monitored by cameras. The percentage of vehicles exceeding the speed limit decreased from 37% to 22%. In the opposite direction (without camera monitoring) there was no change in speeds.

An evaluation of red light cameras installed at five intersections in Stockholm during the 1970s reported essentially no change in red light running behavior. However, there was little effort to inform the public of the enhanced enforcement. General deterrence of aberrant driving typically is achieved through posted signs and public information campaigns advertising the enforcement program. When Singapore began its red light camera program in 1986, warning signs were placed upstream of each camera intersection. Red light running rates at these intersections were 42% lower during the first 2 years of the program compared with the prior 2 years.

The red light camera program in New York City was set up similar to that in Singapore. The number of red light violations issued per camera per day dropped steadily from 30.8 in 1994 to 7.8 in 2015. Studies in the US states of California and Virginia reported reductions in red light violation rates of about 40% soon after the 1997 introductions of red light cameras (Maccubbin et al., 2001). The effect was evident at both camera-equipped intersections and at nearby signalized intersections not equipped with red light cameras (i.e., spillover or halo effect).

A 2014 study in Virginia also found significant reductions in red light violations at camera intersections. These reductions were greater the more time had passed since the light turned red. Violations occurring at least a half second after the light turned red were

39% less likely than would have been expected without cameras. Violations occurring at least 1 s after were 48% less likely. Reviews of international red light camera studies concluded that cameras lower red light violations by 40%–60% (Retting et al., 2003).

A study of speed-on-green/red light cameras in Winnipeg, Canada concluded that the cameras reduced red light running behavior by 21% and speeding behavior by 19% (Vanlaar et al., 2014).

The Effects on Roadway Crashes

Most studies worldwide have reported reductions in crashes associated with speed camera programs—especially the more serious crashes. After Norway instituted a speed camera program in 1988, there was an associated 20% reduction in injury crashes and 12% reduction in property-damage-only crashes on photo-enforced rural two-lane roads (Elvik, 1997). The speed cameras on a highway in British Columbia, Canada were associated with a 16% reduction in all crashes. A study of speed cameras in Hong Kong reported reductions of 23% for injury crashes and 66% for fatal crashes.

A study of 62 fixed speed cameras on 30 mi/h roads throughout the United Kingdom concluded that the cameras were associated with a 25% reduction in injury crashes. However, the researchers estimated that only 20% was attributable to reductions in speed, while the other 5% was due to traffic diverting away from the cameras (Allsop, 2010). The photo speed enforcement of the 80 km/h beltway around Barcelona, Spain, was associated with a 30% reduction in injury crashes. However, the enforcement that began 2 years later on Barcelona's 50 km/h arterial roads did not have any clear effect.

A study of France's speed camera program found that the July 2002 announcement of the initiative, which was widely covered in the media and included not only the introduction of cameras but also increased penalties for traffic violations and the creation of new traffic offenses, was associated with a 12% drop in the fatality rate. When the cameras became operational, there was an additional reduction of 10%, and that effect persisted over time. The rate of nonfatal injuries also declined after the announcement and in the first month of the program, but, unlike the effect on fatalities, the effect on injuries diminished over time.

A speed camera program began in Lisbon, Portugal in 2007. Injury crashes declined by 54% within the first few years. However, crashes began to increase after that, possibly due to familiarity with the system, malfunctioning equipment, and insufficient sanctions of violators.

A speed camera program in the US state of Maryland was associated with an 8% reduction in the likelihood that a crash on a camera-eligible road was speeding-related and a 19% reduction in the likelihood that a crash involved an incapacitating or fatal injury.

A 2010 review of 28 speed camera studies reported reductions of 8%–49% for all crashes, 8%–50% for injury crashes, and 11%–44% for crashes involving fatalities and serious injuries. Reviewed studies with longer duration showed that these trends were either maintained or improved with time. A 2014 meta-analysis concluded that speed cameras reduce all crashes in the vicinity of cameras by 20% and fatal crashes by 51%. However, the crash effects are no longer evident beyond 1 km downstream of the cameras. Section control was estimated to have an even greater effect—reducing all crashes by 30% and serious/fatal crashes by 56%.

The early studies of red light cameras in Australia yielded mixed results with regard to their effect on crashes. For example, a study of red light cameras in Sydney during 1986–91 reported a 26% reduction in injury crashes and an 8% reduction in crashes overall. A study in Adelaide during 1983–93 reported a 20% reduction in injury crashes, but a 2% increase in crashes overall—including increases in both rear-end and right-angle crashes. Right-angle crashes (i.e., front-into-side) are the crashes most closely associated with red light running (Aeron-Thomas and Hess, 2005).

A study of red light cameras in London in 1996 reported an 18% reduction in injury crashes. A 1997 study in the Netherlands reported a 25%–36% reduction in injury crashes (Oei et al., 1997). An evaluation of the Singapore red light camera program reported a 10% reduction in right-angle injury crashes, a 6% increase in rear-end injury crashes, and a 9% reduction in all injury crashes.

Research in the US state of California reported crash reductions of 7% following the introduction of red light cameras, with injury crashes reduced by 29%. Right-angle crashes at these intersections declined by 32% overall, and right-angle crashes involving injuries fell 68%. Rear-end crashes increased by 3%.

A study of six jurisdictions in the US state of Virginia reported an 8%–42% decline in red-light running crashes, but a 27%–42% increase in rear-end crashes. Whether or not the overall change in crashes was positive or negative depended on the crash distribution of the jurisdiction.

A 2005 study evaluated red light camera programs in seven US cities across three states. The study found that, overall, right-angle crashes decreased by 25% while rear-end collisions increased by 15%. The authors concluded that the economic costs from the increase in rear-end crashes were more than offset by the economic benefits from the decrease in right-angle crashes (Council et al., 2005).

A study compared large US cities that turned off red light cameras with those with continuous camera programs. In 14 cities that shut down their programs during 2010–14, the rate of fatal crashes at signalized intersections was 16% higher than would have been expected if they had left the cameras on. A study comparing large cities with red light cameras to those without found the devices reduced the rate of fatal crashes at signalized intersections by 14% (Hu and Cicchino, 2017).

A study in the US city of Houston, which had a red light camera program from 2008 to 2011, found that the camera activation was associated with a 47% decline in right-angle crashes and an 18% increase in rear-end crashes. Conversely, the deactivation was

associated with a 23% increase in right-angle crashes. However, rear-end crashes also went up by 14% after the cameras were deactivated.

A review of international red light camera studies concluded that cameras reduce injury crashes by 25%–30%. A second international review estimated a 13%–29% reduction in all types of injury crashes and a 24% reduction in right-angle injury crashes. A third review concluded that fatal crashes were reduced by 17% and injury crashes were reduced by 12%, but property-damage-only crashes were increased by 3%. Rear-end crashes were increased by 39% (Hoye, 2013).

The study of speed-on-green/red light cameras in Winnipeg reported a 46% reduction in right-angle crashes, an initial 42% increase in rear-end crashes followed by a 19% decrease, and no change in speeding-related crashes. A study in Belgium reported a 6% reduction in right-angle crashes, a 44% increase in rear-end crashes, and a 5% increase in all injury crashes.

Best Practices

Although photo enforcement has generally been successful in reducing the types of driver behavior and crashes that were targeted, there have sometimes been unintended consequences—for example, an increase in rear-end crashes. The same phenomenon has been observed when traffic signals are first added to intersections. Increasing the number of vehicle stops increases the chances for a rear-end crash. It is likely that the increase in rear-end crashes at photo enforcement sites is due to drivers braking suddenly and unexpectedly. Over time this behavior may decrease as drivers become accustomed to the changes. Even so, the overall benefit of a proposed traffic enforcement camera may depend on the historical distribution of crashes at the site. If rear-end crashes greatly outnumber crashes associated with speeding or red light running, then installing an enforcement camera may not be advisable.

The most successful photo enforcement programs have concentrated their efforts on sites with a significant speeding or red light violation problem that has not been solved by traditional enforcement or traffic engineering measures. Photo enforcement is a supplement to the traditional measures—not a replacement.

Another key aspect of successful programs is widespread exposure. Drivers will not change their behavior if they are not aware of the enforcement program. Extensive media coverage, warning signs near the camera sites, highly visible cameras, and published lists of enforcement sites have been used to increase driver awareness.

Citing the owner of the vehicle with the violation (similar to a parking citation) eliminates the need to photograph and identify the driver. This means less work for those reviewing the vehicle photographs while reducing complaints of invasion of privacy.

Importantly, public support is needed if a photo enforcement program is to be sustained. Communities employing automated enforcement must be careful not to give the impression that it is a moneymaking scheme. Counting on automated enforcement fines to balance the budget or redefining the criteria for a violation is a sure way to lose public support.

References

- Aeron-Thomas, A.S., Hess, S., 2005. Red-light cameras for the prevention of road traffic crashes. *Cochrane Database Syst. Rev.* 2, CD003862.
- Allsop, R., 2010. The Effectiveness of Speed Cameras: A Review of the Evidence. Royal Automobile Club Foundation, London, England.
- Council, F., Persaud, B., Eccles, K., Lyon, C., Griffith, M., 2005. Safety evaluation of red-light cameras, Report no. FHWA-HRT-05-048, Federal Highway Administration, Washington, DC.
- Decina, L.E., Thomas, L., Srinivasan, R., Staplin, L., 2007. Automated enforcement: a compendium of worldwide evaluations of results, Report no. DOT-HS-810-763, National Highway Traffic Safety Administration, Washington, DC.
- Delaney, A., Ward, H., Cameron, M., 2005. The history and development of speed camera use, Report no. 242, Monash University Accident Research Centre, Victoria, Australia.
- Elvik, R., 1997. Effects on accidents of automatic speed enforcement in Norway. *Transp. Res. Rec.* 1595, 14–19.
- Hoye, A., 2013. Still red light for red light cameras? An update. *Accid. Anal. Prev.* 55, 77–89.
- Hoye, A., 2014. Speed cameras, section control, and kangaroo jumps—a meta-analysis. *Accid. Anal. Prev.* 73, 200–208.
- Hu, W., Cicchino, J.B., 2017. Effects of turning on and off red light cameras on fatal crashes in large U.S. cities. *J. Saf. Res.* 61, 141–148.
- Maccubbin, R.P., Staples, B.L., Salwin, A.E., 2001. Automated Enforcement of Traffic Signals: A Literature Review. Federal Highway Administration, Washington, DC.
- McCartt, A.T., Eichelberger, A.H., 2012. Attitudes toward red light camera enforcement in cities with camera programs. *Traffic Inj. Prev.* 13, 14–23.
- Oei, H.L., Catshoek, J.W.D., Bos, J.M.J., Varkevisser, G.A., 1997. Project red-light and speed (PROROS), Report no. R-97-35, SWOV Institute for Road Safety Research, Leidschendam, Netherlands.
- Retting, R.A., Ferguson, S.A., Hakkert, A.S., 2003. Effects of red light cameras on violations and crashes: a review of the literature. *Traffic Inj. Prev.* 4 (1), 17–23.
- Vanlaar, W., Robertson, R., Marcoux, K., 2014. An evaluation of Winnipeg's photo enforcement safety program: results of time series analyses and an intersection camera experiment. *Accid. Anal. Prev.* 62, 238–247.
- Wilson, C., Willis, C., Hendrikz, J.K., Le Brocq, R., Bellamy, N., 2010. Speed cameras for the prevention of road traffic injuries and deaths. *Cochrane Database Syst. Rev.* 10, CD004607.

Powered Two- and Three-Wheeler Safety

Fangrong Chang^{*,†}, Helai Huang^{*}, Md. Mazharul Haque^{*}, ^{*}School of Traffic and Transportation Engineering, Central South University, Changsha, China; [†]System Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China; [‡]School of Civil Engineering and the Built Environment, Science and Engineering Faculty, Queensland University of Technology, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Basic Introduction to PTWs	443
What are PTWs?	443
Global Distribution of PTWs	444
How PTWs are Used?	444
Traffic Safety Challenges With PTWs	444
Recent Trends in PTW Crashes	444
Road User Characteristics Associated With PTW Crashes	444
Age	444
Gender	444
PTW Users' Attitudes and Driving Behavior	444
Nonuse of Helmets	445
Lack of Conspicuity	445
Road Environment Associated With PTW Crashes	445
Mixed Traffic	445
Design of Road Infrastructure	445
Junction Type	445
Pavement Surface Quality	446
Vehicle Characteristics Associated With PTWs Crashes	446
Weather Conditions Associated With PTWs Crashes	446
A Growing Traffic Safety Challenge With E-Bikes in China	446
Interventions to Improve PTWs Traffic Safety	447
Safer Road Users	447
Addressing Road User's Illegal Behavior	447
Promoting the Use of Personal Protective Equipment	447
Safer Roads	447
Road Design	447
Exclusive Motorcycles Lanes	447
Speed Limits and Traffic Calming	448
Management of Roadside Hazards	448
Safer Vehicles	448
Antilock Brake Systems	448
Headlight On	448
Interventions Related to E-Bikes	448
Acknowledgment	449
Relevant Websites	449
References	449
Further Reading	450

Basic Introduction to PTWs

What are PTWs?

The term "PTWs" refers to motorized two- or three-wheeled vehicles, powered by either a combustion engine or rechargeable batteries. These powered vehicles can be divided into different categories, including motorcycles (street, classic, performance or super-sport, touring, custom, off-road), scooters, electric bikes, and tricycles. A powered two-wheeler is any two-wheeled vehicle propelled by any type of power, in addition to pedaling, which is divided into ultra-light moped, moped, light motorcycle, and motorcycle (MC) according to the engine displacement, maximum speed and maximum weight. E-bikes are two-wheeled bicycles propelled by human pedaling but supplemented by electrical power from a storage battery, which can be divided into bicycle-style and scooter-style. A powered three-wheeler is a three-wheeled vehicle propelled by a motor, generally used for commercial transport of passengers. Most powered three-wheelers are motor-rickshaws (also referred to as e-rickshaws which run on electric batteries) (WHO, 2017).

Global Distribution of PTWs

There is a fast-growing number of PTWs around the world. According to the global status report on road safety 2015 of the World Health Organization, the fleet of global registered PTWs was around 516 million in 2013, compared to the figure of 313 million in 2008. It should be noted that non-motorized two- or three-wheelers (such as electric bicycles), which do not need to be registered were not included in these figures. In addition, given the absence of a registration system in many low- and middle-income countries, these figures tend to be further underestimated. Nevertheless, low- and middle-income countries still account for the vast majority (about 88%) of global PTWs (WHO, 2015).

According to the statistics from the World Health Organization, the largest proportion of registered PTWs (74.5%) was in South-East Asia Region in 2013 (WHO, 2015). Within South-East Asia, China was the largest motorcycle producer and saw an increase in the number of registered motorcycles by 21% during the period from 2007 to 2013, reaching 109 million (Traffic Management Bureau of the Ministry of Public Security, 2014). In Vietnam, motorized two- or three-wheelers accounted for 95% of all registered vehicles, with the rapid growth of approximately 7,500 newly registered motorcycles each day (WHO, 2013).

How PTWs are Used?

PTWs come in diverse forms and are used for different purposes in different parts of the world. In high-income countries (e.g., United States), high-powered motorcycles are commonly used by single riders for recreation, while in low- and middle-income countries, motorcycles with lower engine capacities often carry pillion passengers and are more commonly used for mobility or commercial purposes.

Traffic Safety Challenges With PTWs

Recent Trends in PTW Crashes

The rapid growth in the sales, registrations, and activities of PTWs around the world in recent years has led to more PTW crashes, injuries, and fatalities. In addition, the increase in the world population and the number of people who realize the potential benefits to the economy, mobility, and environmental sustainability of PTWs make the traffic safety associated with PTWs more challenging.

Compared with the occupants of other motorized vehicles, PTWs users lack nearly all protections offered to car occupants, which lead to their extrusive vulnerability to injury. After controlling for per vehicle mile traveled, motorcyclists are reported to suffer a 26-fold higher risk of death in a crash than those driving other types of motor vehicles (National Highway Traffic Safety Administration, 2015). As one of the most vulnerable road-user groups, PTW users represent 28% of all road deaths around the world while in South-East Asia, Africa, and the Western Pacific, they account for a larger proportion of deaths on roads, with the figures of 43%, 40%, and 36%, respectively (WHO, 2015).

Road User Characteristics Associated With PTW Crashes

Age

The age distribution of PTW users in crashes roughly follows the age distribution of PTW user population. In low- and middle-income countries, most PTW users are aged 15–34 years, and the majority of fatalities also fall into the same age group. While in high-income countries, PTWs are widely used by people aged 35 years or older, and the mean age for PTW users killed as a result of a crash is about 55 years (WHO, 2017).

Young and senior PTW riders have a higher injury risk in a crash, compared to middle-aged riders. The increased injury risk among young riders is predominantly associated with their lack of experience and tendency to exhibit risky behavior (Vlahogianni et al., 2012). In terms of older riders, the higher probability of being involved in fatal crashes is attributed to their decreased reaction ability, reduced sensory and perceptual ability, and less physical resiliency to crashes (Chang et al., 2019).

Gender

The association between rider gender and PTW crashes has been widely explored by researchers. Overall a much higher crash risk for men than for women has been reported per mile driven. This could be explained by the differences in type of travel, risk taking, natural fragility, choice of vehicle, travel purposes between men and women, etc. (WHO, 2017). In terms of injury risks, female motorcycle riders were found more likely to be involved in serious injuries (Savolainen and Mannering, 2007a).

PTW Users' Attitudes and Driving Behavior

Motorcycle riding is considered as dangerous activity and tend to attract risk-seeking individuals, especially young people. Young riders tend to make more "attitudinal" errors, which are directly associated with their risk-taking behavior such as speeding and alcohol consumption, while young riders' risk-taking behavior have been demonstrated to be major contributing factors to road traffic crashes and injuries. As such, much research focus on exploring young PTW drivers' attitudes toward traffic safety and the associated risk-taking behavior (Vlahogianni et al., 2012).

Personality traits including anxiety, anger, sensation-seeking, and normlessness are found to have direct effects on young motorcycle riders' attitudes toward traffic safety. More specifically, riders who tend to be angry, sensation-seeking, and normless have a higher likelihood of engaging in unsafe driving behavior (Wong et al., 2010). PTW riding attitudes are also greatly affected by socio-cultural factors and socio-economic factors (Li et al., 2009). The attempt to reveal factors related to casualty risks, such as speed, gender, age, and lack of concentration could be inefficient, if social and cultural factors are not taken into consideration.

Personality traits also have indirect impacts on riders' risk-taking behavior, mediated by traffic safety attitudes. Overconfident, sensation-seeking, amiable and impatient riders tend to behave unsafely. In addition, risk-taking behavior is associated with rider age and experience (Wong et al., 2010). Although a period of absence from riding might lead to a decline in safety-related motorcycle skills, the identified specific patterns of youth behavior play a much greater role in crash involvement than inexperience (Rutter and Quine, 1996).

Nonuse of Helmets

The protective effect of helmets on users involving in crashes has been well-documented (Kim et al., 2015). For example, the use of helmets can protect riders from suffering from neck and head injuries, which are two main causes of death, severe injury and disability for PTW users. During the crash occurrence, PTW users might collide with other objects or road surface, which could lead to different types of injuries. The use of helmets prevents riders' brain from direct contact with the road surface or other objects and thus reduces the risk of serious head and neck injuries by reducing the impact of force. However, non-standard helmet and incorrect use of helmet could weaken its protection given to riders. Therefore, the use of helmet, the proper way to use it, and helmet quality should be emphasized equally to maximize the protection for riders.

Lack of Conspicuity

Due to the smaller size and rapid acceleration, PTWs are often not perceived by other drivers in time to avoid a hazardous situation. Based on some research reports, most car drivers involved in crashes admit having looked in the direction of the PTWs prior to maneuvering, but failed to detect the approaching motorcycle (Pai et al., 2009). Riders' right-of-way is often violated by other motorists due to the conspicuity issues associated with motorcycles, leading to automobile-motorcycle gap-acceptance crashes, especially at priority (i.e., stop-/yield-controlled) T-intersections (Pai et al., 2009). Therefore, reflective or fluorescent clothing and headlight operation is suggested to address conspicuity issues and reduce the potential for crashes.

Road Environment Associated With PTW Crashes

Mixed Traffic

In many countries, PTW riders share the road with other motorized vehicles, which significantly increase the likelihood of crash occurrences. In low- and middle-income countries, where there is a large PTW fleet, the increases in the interactions between PTW users and motor vehicles lead to a higher-level crash risk. While in high-income countries, many other drivers are not familiar with PTWs and thus it's difficult for them to detect PTWs or judge PTWs' speed in mixed traffic. This unfamiliarity is likely to result in dangerous conditions for PTW users (Gershon et al., 2012).

Design of Road Infrastructure

The design of road infrastructure such as road geometry and road layout is an important predictor of crash occurrence and crash-related injury severity of PTW users (Manan, 2014). The impacts of horizontal curves on PTW crashes have been underlined. Motorcycle riders tend to be involved in out-of-control crashes on horizontal curves, which is mainly due to acceleration or deceleration, or instability on curves. Both the radius and the length of horizontal curves are significantly associated with the risk of single-vehicle motorcycle crashes. Although motorcycle crash risks are found to vary with travel purpose (commuting or recreational), this influence may be explained by various road environments associated with different traffic purposes, given that a higher proportion of crashes during the recreational period were found on curves while most crashes during the commuting period occurred on straight roadways and at junctions. In addition, curve type, horizontal and vertical sight distance, and turning provision, lane and shoulder width, and surface friction were demonstrated to be significantly associated with the severity of PTW crashes (Milling and Hillier, 2015).

Crash barriers have a strong impact on the injuries associated with PTW crashes. PTW users account for 42% of all deaths resulting from guardrail crashes in the United States, and 22% of the fatalities from concrete barrier crashes (Gabler, 2007). Among collisions with guardrails, the fatality rate of PTW riders is 80 times higher than that of other vehicle users. PTW users involved in fatal crashes are reported to collide with the barriers at a relatively shallow angle of 15–45 degrees (Gibson and Benetatos, 2000). The exposed posts of the barrier are generally the main cause of injury for the crash-involved PTW users (Gibson and Benetatos, 2000).

Junction Type

Junction type plays an important role in PTW crashes. In terms of crash risk, motorcyclists are over exposed at signalized intersections (Pai and Saleh, 2007). In addition, motorcyclists' injury severities tend to increase at junctions due to their susceptibility to crash injuries which act synergistically with the complexity of conflicting movements and maneuvers between motorcycles and automobiles. More than half of motorcycle crashes causing injuries were found at T-junctions, among which motorcycle-automobile angle crashes were identified as the most common crash type (Pai et al., 2009; Pai and Saleh, 2007).

Another type of location where motorcyclists are overrepresented in fatal crashes are modern roundabouts, at least in the United States (Steyn et al., 2015). A summary that was made on May 12, 2019, shows there has so far (1990–2019) been 81 fatal crashes at modern roundabouts in the United States, killing 89 people. One pedestrian and two bicyclist have been killed but at least 30 motorcyclists are victims. (A number of the victims are shown as unknown type but motorcycles make up over a third of all fatalities). At roundabouts, it is difficult to get killed in an automobile if you drive at the design speed, typically around 18–25 mph, and automobile occupants killed are frequently high-speed single vehicle crashes. But motorcyclists can get killed at low speed, and motorcyclists also sometimes enter roundabouts at speeds way above the intended speed losing control.

Factors that are deadly to motorcyclists vary with traffic control measures employed at junctions (Pai and Saleh, 2007). Male riders were found to be more injurious than females at stop, give-way signs, markings-controlled junctions, and signalized junctions. Elderly drivers, riding in early morning or on weekend, street lights unlit, and head-on collisions appear to predispose motorcyclists to more severe injuries at stop, give-way signs or markings-controlled and uncontrolled junctions. Despite different influential factors of crashes in junctions controlled by various measures, greater motorcycle engine size, good weather, non-built-up road, and collisions with bus/coach or heavy good vehicle (HGV) result in more severe injuries regardless of the employed control measures.

Pavement Surface Quality

PTWs are more sensitive to road surface condition than other motor vehicles. Pavement surface conditions are found to significantly impact sideswipe crashes between motorcycles and other motor vehicles at junctions and at-fault motorcycle crashes at non-junctions (Haque et al., 2009; Shankar and Mannering, 1996). Nevertheless, in certain circumstances, drivers' compensation behavior can mitigate the risks, for example, lowering speed on a wet pavement surface.

Surface grip, surface irregularities and potholes, loose materials, patch repairs, and road markings are suggested to be important contributing factors in crashes. Especially, motorcycles are very sensitive to changes in friction level between the pavement surface and tires (Pearson and Whittington, 2001); once the friction between a tire and the terrain is reduced, the centrifugal force and the weight force, which are centered in the center of gravity, create a momentum leading the motorcycle to a sudden rotation and instability (Cossalter et al., 2007).

Vehicle Characteristics Associated With PTWs Crashes

PTW type has an important influence on crash risks. In a study conducted in Queensland, Australia, the crash risk of mopeds appears higher than that of motorcycles and larger scooters; moped and scooter crashes are, on average, less severe than motorcycle crashes (Blackman and Haworth, 2013). In addition, the engine size of motorcycles has been widely demonstrated to be associated with crash severity. As motorcycles with larger engine capacity are always associated with larger vehicle mass, greater power and potential speed, increased motorcycle engine size tends to result in more serious injuries for motorcycle riders (Pai and Saleh, 2007).

Weather Conditions Associated With PTWs Crashes

It seems that PTW riding is more likely influenced by weather than other motor vehicle drivers, given that PTW users are exposed to the outside environment. However, weather has rarely been reported to be a contributing factor in motorcycle crashes. Moreover, weather made no contribution in 92.7% of crashes according to a European and Australian large-scale study (ACEM, 2003; Johnston et al., 2008). A larger number of severe PTW injuries appear to occur under fine weather (Pai and Saleh, 2007). This result can partly be explained by the fact that riding is mainly a recreation activity in many parts around the world and purely influenced by weather conditions. In other countries where PTWs primarily are used as a means of transportation in daily life, users are more likely to shift to other modes of transportation (e.g., car, public transport) under adverse weather conditions.

A Growing Traffic Safety Challenge With E-Bikes in China

Electric bikes—which provide an affordable, convenient, physical labor-saving, and energy-saving personal transportation mode—are becoming increasingly popular in many countries throughout Asia and Europe, especially in China. E-bikes were introduced into China in the late 1990s. Since, the introduction of legislation to deal with e-bikes, the number of e-bikes in China has skyrocketed from 40,000 in 1998 to about 250 million in 2018 (WHO, 2017). The tremendous growth and popularity of e-bikes entail traffic safety concerns among the public as observed in crash statistics. A total of 56,200 casualty crashes involving e-bikes during the period 2013–17 resulted in 8,431 fatalities, 63,500 injuries and about 16.5 million US dollars (Equivalent to 111 million RMB) of direct property damage in China (National Bureau of Statistics of China, 2018).

E-bikes are classified as nonmotorized vehicles according to traffic laws in China. Thus, no license, insurance, or helmet use is required for e-bike riders. E-bike riders who have a driver's license were found less likely to be involved in at-fault crashes than those without a license, which imply the lack of traffic knowledge or risk perception ability for e-bike riders. This conclusion could be further supported by the phenomenon that quite a large proportion of e-bike riders travel and behave illegally in motor vehicle lanes. The most common and potentially unsafe e-bike riding behavior include speeding, carrying passengers illegally, riding against the traffic flow, violating traffic signals, and using a mobile phone while riding (Du et al., 2014).

Interventions to Improve PTWs Traffic Safety

Safer Road Users

Addressing Road User's Illegal Behavior

The improvement of road user behavior is crucial in addressing road traffic safety issues. During a quite long period, counter-measures, including training and licensing, enforcement and communication campaigns, have been generally developed for drivers' better operations on roads.

Inexperience is a significant contributor to motorcycle crashes but the training and licensing system showed limited success in ensuring the safety of motorcycle riders (Haworth and Rowden, 2010). Although the hazardous perception time decreases as PTWs users' experiences increase, it remains unclear whether there is any reduction in crashes for formally trained riders compared to those who have not undertaken a formal training course (Haworth and Mulvihill, 2005). The insignificant beneficial effects of training may be explained the improper training course setting which only focuses on basic maneuvering skills rather than changes in attitudes toward safety traffic situations or the ability to avoid emergency situations. What's more, a higher crash rate was even found for trained riders, which was attributed to the overconfidence caused by the training (Savolainen and Mannering, 2007b). The training course for motorcycle riders therefore needs to be improved for their safety traveling on roads. For more information about training, (Refer to the article 10128, Education and Training of Drivers).

Enforcement rather than education is believed to be more helpful in curbing motorized vehicle users' illegal behavior (Williams and Wells, 2004). Especially, stronger enforcement is needed to improve road users' compliance with safety rules against risky behavior such as speeding, drinking and driving, helmet use, and vehicle standards. In addition, highly visible enforcement activities should be developed sustainably to deter potential offenders.

As a common countermeasure, communication campaigns targeting specific groups or specific behavior have been demonstrated to be effective and used to promote traffic safety in many countries.

Promoting the Use of Personal Protective Equipment

The use of protective equipment such as helmet, and reflective and protective clothing is beneficial in preventing or reducing injuries of PTW users. Although wearing properly a quality-standard motorcycle helmet can reduce the risk of death by over 40% and the risk of severe injury by almost 70%, only 49 countries in the world have helmet laws that meet the best practice according to the following criteria used by WHO (WHO, 2018).

- Law mandates that helmets must be worn by all riders (drivers and passengers), on all roads and with all engine types.
- In addition, the law must specify that the helmet needs to be properly fastened.
- Law refers to a standard for helmets.

In order to reduce rider injury severity, every country should try to improve their helmet laws according to these criteria and make sure the law is strictly enforced.

Reflective and protective clothing such as jackets, pants, boots, and gloves is a promising intervention for reducing PTW injuries. By increasing the brightness, reflective clothing helps improve the visibility of PTW users and differentiate riders and passengers from surroundings. Protective clothing is specially designed to prevent, or reduce the injury severity when a crash occurs by providing adequate abrasion resistance and impact protection for users. Riders who wear a range of protective clothing are significantly less likely to be hospitalized (30%–60%) and less likely to incur injuries in crashes (de Rome et al., 2011). Despite its effectiveness, the use of protective clothing is rather limited, especially in low- and middle-income countries. Promoting and enforcing compliance is key to the successful implementation of this intervention.

Safer Roads

Road Design

As described in section Road Environment Associated with PTW Crashes, PTW users have a higher crash risk on curves, bends, slippery roads, and at junctions. Road surface and road-marking materials that provide better grip to PTWs are encouraged to ensure the stability of vehicles on roads (WHO, 2017). The radius and length of each horizontal curve should be appropriate and not require special riding skills of PTW users. Wider lanes on major and minor roads, and an increase in the number of lanes on major roads are also beneficial in reducing motorcycle crashes (Harmen et al., 2004). In terms of the higher crash risk at roundabouts, exclusive or protected (right hand) turn lanes and paved shoulders whose width is more than 1 meter may help to reduce motorcycle crashes at roundabouts. In addition, good visibility and signage around junctions and roundabouts also helps motorcyclists manage their speeds as they approach the sites. Due to the smaller size of motorcycles, signage, vegetation, other vehicle, and objects could obscure them, which should be taken into specific consideration when designing intersections and roundabouts.

Exclusive Motorcycles Lanes

Exclusive motorcycle lanes have been demonstrated effective in significantly reducing motorcycle crash frequency and severity by separating motorcycles from general traffic, especially from high-speed and heavier vehicles (Radin et al., 2000). However, setting

aside specific lanes for motorcycles is not generally economic and feasible in most countries. Segregation is likely to be significantly beneficial and acceptable to the public when the proportion of PTW users is more than 20%–30% of all vehicles on the road—as is the case in many low- and middle-income countries (Radin et al., 2000).

Speed Limits and Traffic Calming

Traffic calming measures are useful in reducing crashes for all vehicle types. But occasionally, such interventions, for example placing obstacles, including speed humps, and small vertical objects designed to minimize speed, might pose a negative effect on the safety of motorcycle riders (ITF, 2015). Other forms of traffic calming such as horizontal markings on the road used to warn motorcyclists of these obstacles are suggested. In addition, the choice of locations of traffic calming measures aiming at other vehicles should take into consideration the characteristics of PTW traffic.

Management of Roadside Hazards

PTW riders have an extremely high fatality risk, if colliding with a roadside obstacle (Daniello and Gabler, 2011). Eliminating roadside installations such as barriers, posts, utility poles, etc., can reduce the risk of a PTW rider impacting any hazardous objects as well as provide road users room to correct operation errors. Guardrails and crash barriers are often used to separate vehicles from roadside hazards but there is no consistent view about the best guardrail or crash barriers for motorcyclists. Increasing evidence showed that the position of motorcyclists impacting a guardrail may be more important than the type of guardrail. Motorcycle-friendly barriers that allow fallen motorcyclists to slide along the surface without impacting any specific component of the system are thus encouraged. In addition, it was also evidenced that collisions with fixed objects tend to result in more severe injuries to PTW users than collision with crash barriers, which support the use of barriers to prevent impacts with such objects (Bambach et al., 2013). Especially, the priority should be given to improving barriers and guardrails on curves, the proper placement and maintenance of guardrail and crash barrier systems.

In Sweden, where they have installed centerline cable barriers along over 1300 km of two-lane rural highways, motorcycle organizations were very skeptical about them, stating that the cable would act as a deadly object, if hit. However, evaluations have shown that these roads are safer for all road-users, even for MC riders, than undivided two-lane roads. A 2006 evaluation by the Swedish National Road and Transport Research Institute (VTI) shows that there would have been expected to be 67 motorcyclists killed or seriously injured in 2003–05 had the roads not had center barriers whereas the outcome was that only 22 motorcyclists were killed or seriously injured on these roads. So, the conclusion is that the benefit of not facing oncoming traffic very much outweighs the risk of injury from MC crashes involving the cable barrier (VTI, 2007).

Safer Vehicles

Antilock brake systems (ABS) and the use of headlights are two interventions demonstrated useful in improving PTWs traffic safety.

Antilock Brake Systems

Antilock brake systems have the potential to reduce the occurrence of PTW crashes and to mitigate their consequences by optimizing braking distance and helping riders sustain stability in emergency braking situations. By using ABS, the fatality rate of motorcycles could be reduced by 37%, and half of severe and fatal crashes among motorcycles over 125 cc could be avoided (Teoh, 2011). In 2016, the legislation for mandatory ABS equipment on all motorcycles (over 125 cc) was passed in the European Union. To improve the global motorcycle traffic safety, a similar law applicable for other regions is also needed.

Headlight On

The lack of conspicuity is one important factor contributing to PTW crashes. Mandatory use of headlights at daytime and night has been proven effective in increasing the conspicuity of PTWs and improving PTW safety (Radin et al., 1996). At nighttime, the use of headlamps can maximize visibility of—and the field of vision for—PTW riders, which lead to improved judgment accuracy of PTWs' approach-speed and the subsequent improvement of traffic safety. Driving with headlights on during daytime could reduce visibility-related crashes by between 29% and 40% (Radin et al., 1996). However, compliance with daytime running light legislation is impeded by various factors (e.g., energy savings) as well as by personal choice. Manufacturers play a critical role in promoting the use of daytime running lights, for example, equipping with Automatic Headlamp On at manufacturing stage to ensure the headlight is always on when the PTWs engine is running.

Interventions Related to E-Bikes

In order to improve e-bike traffic safety, a safety technical specification for electric bicycle was made inaction in 2018 and carried out from 15 April 2019 in China (Ministry of Industry and Information Technology of the People's Republic of China, 2018). The specification illustrates the standard of electric bicycles. In order to ensure the specification is implemented, State Administration for Market Regulation, Ministry of Industry and Information Technology of the People's Republic of China, the Ministry of Public Security of the People's Republic of China jointly issued the "Opinions on Strengthening the Implementation of National Standards for Electric Bicycles" which involves strict management of e-bike production, strict supervision of e-bike sales, strict registration and use management of e-bikes, dealing with the existing e-bikes that do not meet the new standard, establishment of a long-term

regulatory mechanism, and social publicity and guidance with the purpose of further standardizing the production, sales and use management of electric bicycles. The strict registration and use management of e-bikes will improve the efficiency of curbing of riders' traffic illegal behavior. Wearing a helmet when riding an electric bicycle is also suggested by these three national departments.

Acknowledgment

This work was jointly supported by: (1) the Joint Research Scheme of National Natural Science Foundation of China/Research Grants Council of Hong Kong (No. 71561167001 & N_HKU707/15); (2) National Key Research and Development Plan (No. 2018YFB1201601).

Relevant Websites

International Transport Forum

<https://apps.who.int/iris/bitstream/handle/10665/254759/9789241511926-eng.pdf;jsessionid=2101F434C7E4AE94B3D0B1B7D6593DBF?sequence=1>.

Ministry of Industry and Information Technology of the People's Republic of China

<http://www.miit.gov.cn/n1146285/n1146352/n3054355/n3057497/n3057502/c6176772/part/6176777.pdf>

OECD

<http://dx.doi.org/10.1787/9789282107942-en>.

World Health Organization

https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/.

https://www.who.int/violence_injury_prevention/road_safety_status/2015/en/.

<https://extranet.who.int/roadsafety/death-on-the-roads/#helmets>.

References

- ACEM, 2003. MAIDS—Motorcycle accident indepth study, Association des Constructeurs Europeens de Motorcycles, Brussels.
- Bambach, M.R., Mitchell, R.J., Grzebieta, R.H., 2013. The protective effect of roadside barriers for motorcyclists. *Traffic Inj. Prev.* 14 (7), 756–765.
- Blackman, R.A., Haworth, N.L., 2013. Comparison of moped, scooter and motorcycle crash risk and crash severity. *Accid. Anal. Prev.* 57, 1–9.
- Chang, F., Xu, P., Zhou, H., Lee, J., Huang, H., 2019. Identifying motorcycle high-risk traffic scenarios through interactive analysis of driver behavior and traffic characteristics. *Transport. Res. F-Traf. Psychol. Behav.* 62, 844–854.
- Cossalter, V., Aguggiaro, A., Debus, D., Bellati, A., Ambrogio, A., 2007. Real cases motorcycle and rider race data investigation: fall behavior analysis. In: 20th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration (No. 07-0342).
- Daniello, A., Gabler, H.C., 2011. Fatality risk in motorcycle collisions with roadside objects in the United States. *Accid. Anal. Prev.* 43 (3), 1167–1170.
- de Rome, L., Ivers, R., Fitzharris, M., Du, W., Haworth, N., Heritier, S., Richardson, D., 2011. Motorcycle protective clothing: protection from injury or just the weather? *Accid. Anal. Prev.* 43 (6), 1893–1900.
- Du, W., Yang, J., Powis, B., Zheng, X., Ozanne-Smith, J., Bilston, L., et al., 2014. Epidemiological profile of hospitalized injuries among electric bicycle riders admitted to a rural hospital in Suzhou: a cross-sectional study. *Injury Prevent.* 20 (2), 128–133.
- Gabler, H.C., 2007. The risk of fatality in motorcycle crashes with roadside barriers. In: *Proceedings of the Proceedings of the 20th International Technical Conference on the Enhanced Safety of Vehicles*, pp. 18–21.
- Gershon, P., Ben-Asher, N., Shinar, D., 2012. Attention and search conspicuity of motorcycles as a function of their visual context. *Accid. Anal. Prev.* 44 (1), 97–103.
- Gibson, T., Benetatos, E., 2000. Motorcycles and crash barriers. NSW Motorcycle Council Report, Australia.
- Haque, M.D.M., Chin, H.C., Huang, H., 2009. Modeling fault among motorcyclists involved in crashes. *Accid. Anal. Prev.* 41 (2), 327–335.
- Harnen, S., Umar, R.R., Wong, S.V., Hashim, W., 2004. Development of prediction models for motorcycle crashes at signalized intersections on urban roads in Malaysia. *J. Transport. Statist.* 7 (2), 27–39.
- Haworth, N., Mulvihill, C., 2005. Review of motorcycle licensing and training (Report No. 240). Monash University Accident Research Centre, Melbourne.
- Haworth, N., Rowden, P., 2010. Challenges in improving the safety of learner motorcyclists. Paper presented at the 20th Canadian Multidisciplinary Road Safety Conference, Niagara Falls.
- ITF, 2015. Improving Safety for Motorcycle, Scooter and Moped Riders. OECD Publishing, Paris.
- Johnston, P., Brooks, C., Savage, H., 2008. Fatal and Serious Road Crashes Involving Motorcyclists, Research and Analysis Report, Road Safety, Monograph 20. Department of Infrastructure, Transport, Regional Development and Local Government, Canberra, Australia.
- Kim, C., Wiznia, D.H., Averbukh, L., Feng, D., Leslie, M.P., 2015. The economic impact of helmet use on motorcycle accidents: a systematic review and meta-analysis of the literature from the past 20 years. *Traf. Inj. Prev.* 16, 1–7.
- Li, M.D., Doong, J.L., Huang, W.S., Lai, C.H., Jeng, M.C., 2009. Survival hazards of road environment factors between motor-vehicles and motorcycles. *Accid. Anal. Prev.* 41 (5), 938–947.
- Manan, M., 2014. Factors associated with motorcyclists' safety at access points along primary roads in Malaysia. *Can. J. Microbiol.* 45 (6), 520–529.
- Milling D, Hillier P., 2015. Infrastructure improvements to reduce motorcycle crash risk and crash severity. *Proceedings of the 2015 Australasian Road Safety Conference*, Gold Coast, Australia.
- Ministry of Industry and Information Technology of the People's Republic of China, 2018. Safety technical specification for electric bicycle (Chinese).
- National Bureau of Statistics of China, 2018. China Statistical Yearbook 2018. National Bureau of Statistics of China, Beijing, China.
- National Highway Traffic Safety Administration, 2015. Traffic Safety Facts 2013 Data: Motorcycle. National Highway Traffic Safety Administration, Washington, DC, USA.

- Pai, C.W., Hwang, K.P., Saleh, W., 2009. A mixed logit analysis of motorists' right-of-way violation in motorcycle accidents at priority T-junctions. *Accid. Anal. Prev.* 41 (3), 565–573.
- Pai, C.W., Saleh, W., 2007. An analysis of motorcyclist injury severity under various traffic control measures at three-legged junctions in the UK. *Safety Sci.* 45 (8), 832–847.
- Pearson, R., Whittington, B., 2001. Motorcycles and the Road Environment Road Safety: Gearing Up for the Future. Motorcycle Riders Association, Western Australia.
- Radin Sohadi, R.U., Mackay, M., Hills, B., 2000. Multivariate analysis of motorcycle accidents and the effects of exclusive motorcycle lanes in Malaysia. *J. Crush Prevent. Injury Cont.* 2 (1), 11–17.
- Radin, U.R., Mackay, M.G., Hills, B.L., 1996. Modelling of conspicuity-related motorcycle accidents in Seremban and Shah Alam. Malaysia. *Accid. Anal. Prev.* 28 (3), 325–332.
- Rutter, D.R., Quine, L., 1996. Age and experience in motorcycling safety. *Accid. Anal. Prev.* 28, 15–21.
- Savolainen, P., Mannering, F., 2007a. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. *Accid. Anal. Prev.* 39, 955–963.
- Savolainen, P., Mannering, F., 2007b. Effectiveness of motorcycle training and motorcyclists' risk taking behaviour. *Transport. Res. Rec.* 2031, 52–58.
- Shankar, V., Mannering, F.L., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. *J. Safety Res.* 27 (3), 183–194.
- Steyn, H.J., Ashleigh G., Lee A. R., 2015. Accelerating roundabout implementation in the United States - Volume IV of VII: a review of fatal and severe injury crashes at roundabouts, Publication NO, FHWA-SA-15-072, U.S. Department of Transportation-Federal Highway Administration.
- Teoh, E.R., 2011. Effectiveness of antilock braking systems in reducing motorcycle fatal crash rates. *Traf. Injury Prev.* 12 (2), 169–173.
- Traffic Management Bureau of the Ministry of Public Security, 2014. Statistical Yearbook of Road Traffic Crashes of China. Traffic Management Bureau of the Ministry of Public Security, Nanjing, Jiangsu, China.
- Vlahogianni, E.I., Yannis, G., Golias, J.C., 2012. Overview of critical risk factors in power-two-wheeler safety. *Accid. Anal. Prev.* 49, 12–22.
- VTI, 2007. The Swedish National Road and Transport Research Institute (Swedish). Available from: [https://web.archive.org/web/20130515114645/http://www.svmc.se/upload/SMC%20central/Dokument/Hur%20farlig%20%C3%A4r%20mittvajer%20f%C3%B6r%20motorcyklisterna%20\(Referat%20VTI%202007\).pdf](https://web.archive.org/web/20130515114645/http://www.svmc.se/upload/SMC%20central/Dokument/Hur%20farlig%20%C3%A4r%20mittvajer%20f%C3%B6r%20motorcyklisterna%20(Referat%20VTI%202007).pdf).
- WHO, 2013. Global status report on road safety 2013: supporting a decade of action. World Health Organization, Geneva.
- WHO, 2015. Global Status Report on Road Safety 2015. World Health Organization, Geneva.
- WHO, 2017. Powered two- and three-wheeler safety: a road safety manual for decision-makers and practitioners. World Health Organization, Geneva.
- WHO, 2018. Global Status Report on Road Safety 2018. World Health Organization, Geneva.
- Williams, A.F., Wells, J.K., 2004. The role of enforcement programs in increasing seat belt use. *J. Safety Res.* 35 (2), 175–180.
- Wong, J.T., Chung, Y.S., Huang, S.H., 2010. Determinants behind young motorcyclists' risky riding behavior. *Accid. Anal. Prev.* 42, 275–281.

Further Reading

- Haque, M.M., Chin, H.C., Huang, H., 2008. Examining exposure of motorcycles at signalized intersections. *Transp. Res. Rec.* 2048 (1), 60–65.
- Haworth, N., 2012. Powered two wheelers in a changing world—challenges and opportunities. *Accid. Anal. Prev.* 44 (1), 12–18.
- Organisation for Economic Co-operation Development (OECD)/International Transport Forum, 2015. Improving Safety for Motorcycle, Scooter and Moped Riders. OECD, Paris.

Railroad Safety

Xiang Liu, Zhipeng Zhang, Department of Civil and Environmental Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	451
Railroad Safety Statistics in the United States	452
Railroad Safety Statistics in the EU and Japan	453
Railroad Accident Causes	454
Derailment	454
Collision	456
Risk Mitigation Strategies	458
Track-Related Risk Prevention	458
Rail Inspection	458
Track Geometry Measurement and Analytics	459
Improved Materials and Manufacturing	460
Mechanical-Related Strategies	460
Manual Inspection	460
Machine Vision-Based Inspection	461
Acoustic Emission-Based Detection	462
Human Factor-Relation Strategies	462
Positive Train Control (PTC)	462
Engineer Education and Training	464
References	464
Further Reading	465

Introduction

Railroads play a vital role in transporting cargo and passengers in most countries around the world, and thereby contribute to the economy. In the United States (US), freight rail remains a crucial contributor to the movement of goods and passengers, while commuter ridership has also started to grow in recent years, balancing the nation's transportation options (FRA, 2016). The US freight rail network consists of close to 600 freight railroads, nearly 225,000 km (140,000 miles) of track with 2.8 trillion ton-km (1.74 trillion ton-miles) of traffic annually (AAR, 2018). Fig. 1 shows the current US freight network and passenger network for the United States and part of Canada. The 225,000 km in the United States equals 0.7 km per 1000 people, which is relatively typical from an international perspective. However, Africa and much of Asia have fewer miles of rail per capita. Worldwide, there were around 1,370,782 km of track in 2018. That is around 0.18 km of track per 1000 people. Conversely, most European countries have more tracks per capita. For example, Sweden has 14,475 km of active track, which equals 1.5 km per 1000 people. Also, the United States has only 1,600 km of electrified track, whereas Sweden has almost 9000 km, which is about 180 times more per capita. However, US railroads transport a lot of freight over long distances. The 2.8 trillion ton-km for the United States equals 8600 ton-km per capita. That can be compared to Swedish railroads carrying only 2,200 ton-km per capita.

The rail network of Amtrak (the National Railroad Passenger Corporation) consists of more than 34,000 route-km (FRA, 2019) and serves more than 500 destinations in 46 states plus the District of Columbia. In total, 28 commuter railroads provided over 499 million trips for passengers in 2018 in the United States (APTA, 2019). That equals roughly 1.5 trips per person per year, which is relatively low in an international comparison. For example, in Sweden, railroads carry around 230 million people per year, which equals 23 trips per capita, 15 times more than in the United States. In the United States, passenger rail carries around 10.3 billion passenger-km or 32 km per capita per year. Swedish railroads in 2016 carried people 1280 km per capita, 40 times as much as US railroads. The European Union (EU) uses a railroad network of over 130,000 miles. More specifically, Germany has the longest railway lines in the EU, with around 23,400 miles. The United Kingdom and Sweden have approximately 10,100 miles and 6800 miles, respectively (UNECE, 2019). In Asia, China and Japan have around 81,400 miles and 17,000 miles, respectively (World Bank, 2019). In these vast railroad networks, safety is of the utmost importance, since a train accident may result in numerous injuries or fatalities, substantial infrastructure and rolling stock damages, and severe environmental impacts. This chapter focuses primarily on railroad safety statistics, accident causes, and risk mitigation strategies in the United States. Some rail safety statistics are also given for the European Union (EU) and Japan.

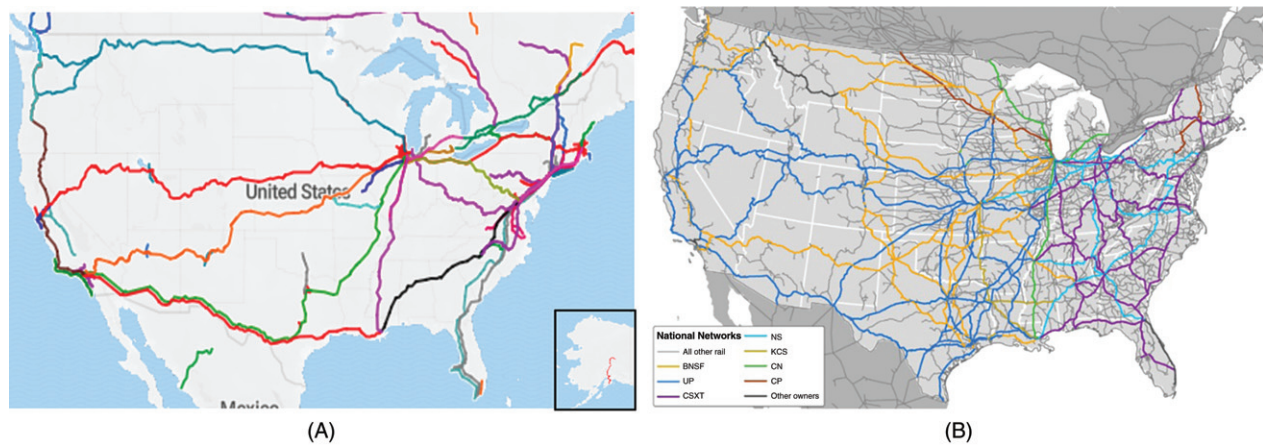


Figure 1 Railroad Network. (A) Passenger Railroads; (B) Freight Railroads. Source: (A) *Traveler*; (B) *Mississippi Export Railroad*.

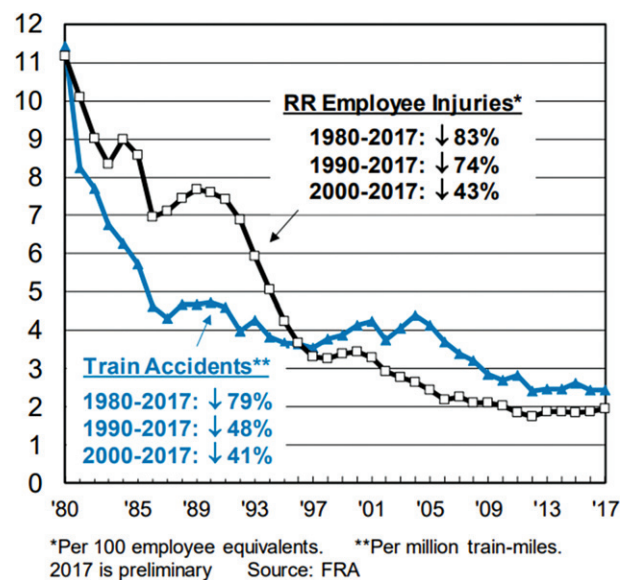


Figure 2 Rail accident and injury rates (AAR, 2018).

Railroad Safety Statistics in the United States

The Association of American Railroads (AAR) point out that nothing is more important to railroads than safety. Railroads today have lower employee injury rates than most other major industries, including trucking, airlines, agriculture, mining, manufacturing, and construction—even lower than food stores (AAR, 2018). For the rail sector, rail safety data in many countries continue to show that recent years have been the safest on record. Fig. 2 presents trends in US railroad-related injuries and accidents over several decades, namely:

- The train accident rate in 2017 was down 79% from 1980 and down 41% from 2000;
- The employee injury rate in 2017 was down 83% from 1980 and down 43% from 2000.

Over the last ten years, the declining trend in US accident rates has slowed, though railroads continue to promote rail safety. According to recent FRA data, the train accident rate, defined as the number of accidents per million train miles, has been reduced by 10% and the employee injury rate is down 16% since 2009 and was around 1.4 accidents per million train km (2.3 per million train miles) in 2017 (AAR, 2019). Fig. 3 shows the accident rate from 2007 to 2017 for all railroads and all types of track combined. Since 2012, accident rates have fluctuated and insignificant decreases indicate the never-ending challenge railroad safety presents. Thus, railroads, in cooperation with policymakers, their employees, suppliers, and customers, are constantly looking for new technologies, operational enhancements, improved training, and other ways to further improve their already excellent safety records.

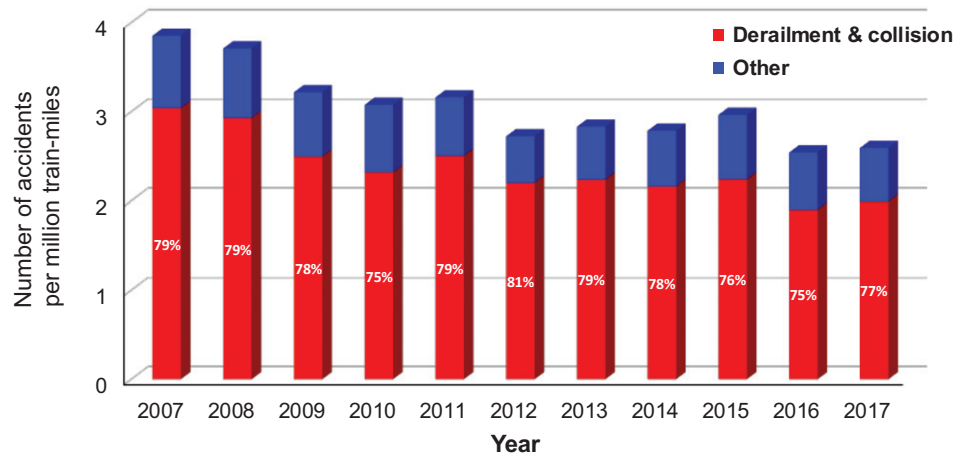


Figure 3 FRA-reportable accident rate for all railroads in the United States.

- Data source: FRA Rail Equipment Accident Database, 6180.54.
- “Accident” is the FRA-reportable accident in which reportable on-track equipment and track damage exceeds the monetary threshold for train accidents. The reporting threshold updates annually (e.g., \$7700 in 2007, \$9200 in 2010, or \$10,700 in 2017).
- “Other” includes highway-rail grade crossing accidents, obstruction, fire/violent rupture, etc.

Derailment and collision are the two major types of train accidents in the United States, accounting for over 75% of all types of accidents. Derailment, according to the FRA guide’s definition (FRA, 2011), is an accident that occurs when on-track equipment leaves the rail for a reason other than collision, explosion, highway-rail grade crossing impact, etc. A collision is defined as “an impact between on-track equipment consists while both are on rails and where one of the consists is operating under train movement rules or is subject to the protection afforded to trains.” This definition includes instances where a portion of a consist occupying a siding is fouling the mainline and is struck by an approaching train. Per the FRA Rail Equipment Accident (REA) database, 29,252 derailments and 4,000 collisions occurred in the United States from 2000 to 2017, leading to over \$4104 million and \$738 million in damage costs to infrastructure and rolling stock, respectively. On average, derailments in one year resulted in around 3 fatalities, 157 injuries, and \$228 million in damage costs. For collisions, the average annual consequence involves about 7 fatalities, 109 injuries, and \$41 million in damage costs. Since grade crossing accidents involve different accident characteristics, this article focuses on derailments and collisions, to the exclusion of grade crossing accidents, covered in a different article of this encyclopedia.

Railroad Safety Statistics in the EU and Japan

This section reviews some railroad safety statistics in the EU and Japan, with emphasis on fatalities due to train accidents. The main reason for this emphasis is that fatal accidents are mostly well-recorded, while nonfatal accidents are likely to be underreported or recorded and reported in diverse ways.

Table 1 shows traffic volumes measured in train-miles per year, fatality rates by person type, and estimated trends in fatal accident rates for the United States and European Union. In general, the EU has about three times as many fatalities of railroad passengers per billion train-miles as US railroads. The fatality rate is defined as the number of fatalities per billion train-miles and is calculated with two person types, which are railroad passengers and staff. If we instead use passenger-miles traveled as exposure, which is a more relevant measure, the result is reversed. Using 2018 passenger-miles traveled by train, the United States had 6.36 billion miles of

Table 1 Fatal railroad accident statistics in the EU and US (Evans, 2010, 2013)

	United States	European Union
Period covered	2000–09	2006–009
Billion train-miles per year	0.7497	2.5784
Fatalities per billion train-miles by person type		
Railroad passengers	9.3	27.2
Staff	35.4	13.8
Period covered	2000–17	1990–2009
Estimated rate of change in fatal accidents per billion train-miles	–7.2%	–6.3%

travel by Intercity trains and the EU had 293 billion miles (472 billion km) of travel that year, meaning the EU sees about 46 times as many miles of travel as the United States, and instead of having a risk per mile three times that of the United States, the risk per mile traveled is only 6% of that of traveling by train in the United States. This can also be expressed as the United States having a 15 times higher fatality risk per mile of travel. Another interesting safety-related statistic is that staff fatalities in the United States account for around 80% of all fatalities (covering both railroad passenger and staff, excluding trespassers) and the fatality rate is about two-and-a-half times that in the EU. The fact that the US railroad passenger fatality rate per train mile is smaller than in the EU is because US railroad operations mainly consist of freight trains, while Europe has a lot of passenger trains. Among EU countries, Sweden has the lowest fatality rate (Evans, 2013). At the end of 2009, Sweden had suffered no fatal train accidents since 1990. Since then, through October 2019, only one person has been killed onboard a train in Sweden as a result of an accident. That was on September 12, 2010, when a 24-year old woman was killed when a passenger train collided with an excavator. Furthermore, the estimated rate of change in fatal accident rates for 2000–17 US railroads and 1990–2009 EU railroads declined by 7.2% and 6.3%, respectively. Similarly, Japan also had around a 6.5% annual decline in the fatal train accident rate from 1971 to 2006 (Evans, 2010).

Railroad Accident Causes

Train accidents cause damage to infrastructure and rolling stock, as well as service disruptions, and may cause casualties and harm to the environment. Although train accidents occur as a result of many different causes, some are much more prevalent than others. Furthermore, the frequency and severity of accidents also vary widely, depending on the particular accident cause. The efficient allocation of resources to prevent accidents in the most cost-effective manner possible requires understanding which factors account for the greatest risks, and under what circumstances. Assessment of the benefits and costs of strategies to mitigate each accident cause can then be evaluated and resources can be allocated so that the greatest safety improvement can be achieved for the level of investment available. This section focuses on two major types of accident, train derailments and train collisions, which are among the most common types of accident.

Derailment

Train derailments are most common in countries with neglected track maintenance. However, derailment occurs in all countries. One of the most serious derailments in recent years occurred in Lac Megantic in Quebec, Canada in 2013. An unattended 74-car freight train carrying crude oil rolled down a 1.2% grade and derailed downtown, resulting in the fire and explosion of multiple tank cars. Forty-two people were confirmed dead and another five were missing and presumed dead. More than 30 buildings in the town's center, roughly half of the downtown area, were destroyed, and all but three of the thirty-nine remaining downtown buildings had to be demolished.

Another much publicized derailment occurred in Germany in 1998, near Eschede, when a high-speed train derailed and crashed into a road bridge, killing 101 people and injuring around 100 others on board. It remains the worst high-speed rail disaster worldwide. High-speed trains typically operate on well-maintained tracks, and Japan has operated Shinkansen trains for over 50 years, carrying over 10 billion passengers, and there have been no passenger fatalities due to train accidents such as derailments or collisions, despite frequent earthquakes and typhoons. France has been almost as successful with their high-speed TGV trains: they had a derailment when testing a train, but have not had a single fatality from derailments or collisions during operations with passengers. The service started in 1981 and TGV trains have transported 1.2 billion passengers and now transport 50 billion passenger-kilometers per year on *lignes à grande vitesse* (high-speed lines) alone.

Below, we will look more in depth at derailments on US railroads.

For the period from 2001 to 2010, 8092 train derailments occurred on Class I freight railroads, which account for approximately 68% of US railroad route miles, 97% of total ton-miles transported, and 94% of the total rail freight revenue (AAR, 2018). The seven Class I freight railroads that operate in the United States are Burlington Northern Santa Fe Railway (BNSF), CSX Transportation (CSX), Grand Trunk Corporation (CN), Kansas City Southern Railway (KCS), Norfolk Southern Combined Railroad Subsidiaries (NS), Soo Line Corporation (CP), and Union Pacific Railroad (UP).

Four types of tracks are recorded in the FRA Rail Equipment Accident (REA) database: main, siding, yard, and industry tracks. These track types are used for different operational functions and consequently have different associated accident types, causes, and consequences. Main, siding, yard, and industry account for 54.9%, 35.2%, 5.4%, and 4.7% of all freight train derailments from 2001 to 2010, respectively (Table 2). Of the four track types, freight train derailments that occurred on main lines had the greatest frequency and severity (measured in average number of cars derailed per accident). For the period 2001–2010, 4439 train derailments occurred in Class I freight railroads. These accidents led to 37,456 total cars derailed, or equivalently, one Class I freight railroad derailment resulted in an average of over 8 derailed cars per accident from 2001 to 2010.

The primary causes of derailments on Class I mainlines are divided into 49 categories or causal groups. Table 3 presents the top 10 derailment cause groups for main, siding, and yard, according to accident data from the Federal Railroad Administration. Broken rails or welds act as the most common cause for all three track types and contributed to around 16% of freight train derailments for all three track types from 2001 to 2010. In particular, broken rails and welds on main lines were more than twice as likely to have been the cause of train derailments than the second and third leading causes (track geometry and bearing failure). Broken rails or welds make up one track-related accident cause group, which includes but is not limited to transverse defect, detail fracture, broken

Table 2 Derailment frequency and severity by track type, US Class I Freight Railroads, 2001–10 (Liu et al., 2012)

Track type	Main	Yard	Siding	Industry	All
Number of freight train accidents (frequency)	4439	2848	436	369	8092
Average number of cars derailed per accident (severity)	8.4	4.7	5.7	4.3	6.8
Total number of cars derailed	37,456	13,363	2477	1593	54,889

Per the FRA guide's definitions of tracks:

- Main track: a track extending through yards or between stations, upon which trains are operated by timetable or train order or both, or the use of which is governed by a signal system.
- Yard track: a system of tracks within defined limits used for the making up or breaking up of trains, for the storage of cars, and for other purposes over which movements not authorized by timetable or by train order may be made.
- Siding: a track auxiliary to the main track used for meeting or passing trains.
- Industry track: a switching track, or series of tracks, serving the needs of a commercial industry other than a railroad.

Table 3 Top 10 accident causes of US freight train derailments by track type, 2001–10 (Liu et al., 2012)

Freight Train Derailments						
Main			Siding		Yard	
Rank	Cause group	Percentage	Cause group	Percentage	Cause group	Percentage
1	Broken rails or welds	15.3	Broken rails or welds	16.5	Broken rails or welds	16.4
2	Track geometry (excluding wide gauge)	7.3	Wide gauge	14.2	Use of switches	13.5
3	Bearing failure (car)	5.9	Turnout defects—switches	9.7	Wide gauge	13.5
4	Broken wheels (car)	5.2	Switching rules	7.7	Turnout defects—switches	11.1
5	Train handling (excluding brakes)	4.6	Track geometry (excluding wide gauge)	7.2	Train handling (excluding brakes)	6.7
6	Wide gauge	3.9	Use of switches	5.8	Switching rules	6.2
7	Obstructions	3.5	Train handling (excluding brakes)	3.5	Track geometry (excluding wide gauge)	3.6
8	Buckled track	3.4	Lading problems	2.3	Miscellaneous track and structure defects	3.4
9	Track-train interaction	3.4	Roadbed defects	2.1	Track-train interaction	3.1
10	Other axle or journal defects (car)	3.3	Miscellaneous track and structure defects	2.1	Other miscellaneous	3.0

rails at base or weld, fractures from shelling or head checks, etc. (Fig. 4). Track geometry is the second major cause of mainline derailments, while wide gauge and use of switches are the second most frequent causes in siding and yard.

Fig. 5 plots derailment frequency and severity, defined as the average number of cars derailed, with frequency on the abscissa (x -axis) and severity on the ordinate (y -axis). Of the four quadrants defined on the basis of the average derailment frequency and severity along each axis, the causes in the upper right quadrant are most likely to pose the greatest risk because they are both more frequent and more severe on average. There are five causation groups in the upper right quadrant: broken rails or welds, wide gauge, buckled track, obstructions, and mainline brake operation.

In addition, there are four other causation groups that have a notably high frequency of occurrence: track geometry (excluding wide gauge), bearing failure (car), broken wheels (car), and train handling (excluding brakes). Three other causation groups are notable because of the high average severity of the resultant derailments: rail defects at bolted joints, other rail and joint defects, and joint bar defects. These last three causes, along with the related causation group of broken rails or welds, are of particular interest, because together they account for almost 20% of all derailments and more than 30% of all derailed cars on Class I main lines.

Fig. 6 presents the primary causes for freight train derailments on Class I mainlines by the number of cars derailed. Broken rails or welds are by far the most common, accounting for 23% of total cars derailed from 2001 to 2010. Causal groups 2 through 10 account for 39%. In other words, the top 10 causal categories are responsible for 62% of cars derailed. Combined, the other 39 identified causal groups account for the remaining 38%. Since broken rails are responsible for the most derailments and cars derailed, focusing on their prevention is likely a promising risk-reduction strategy.



Figure 4 Examples of two common types of broken rail. (A) Transverse defect; (B) Detail fracture. Source: Federal Railroad Administration, 2015. *Track Inspector Rail Defect Reference Manual*; TRAINS Magazine, 2017. *Broken Rails: An Unexpected Pain*.

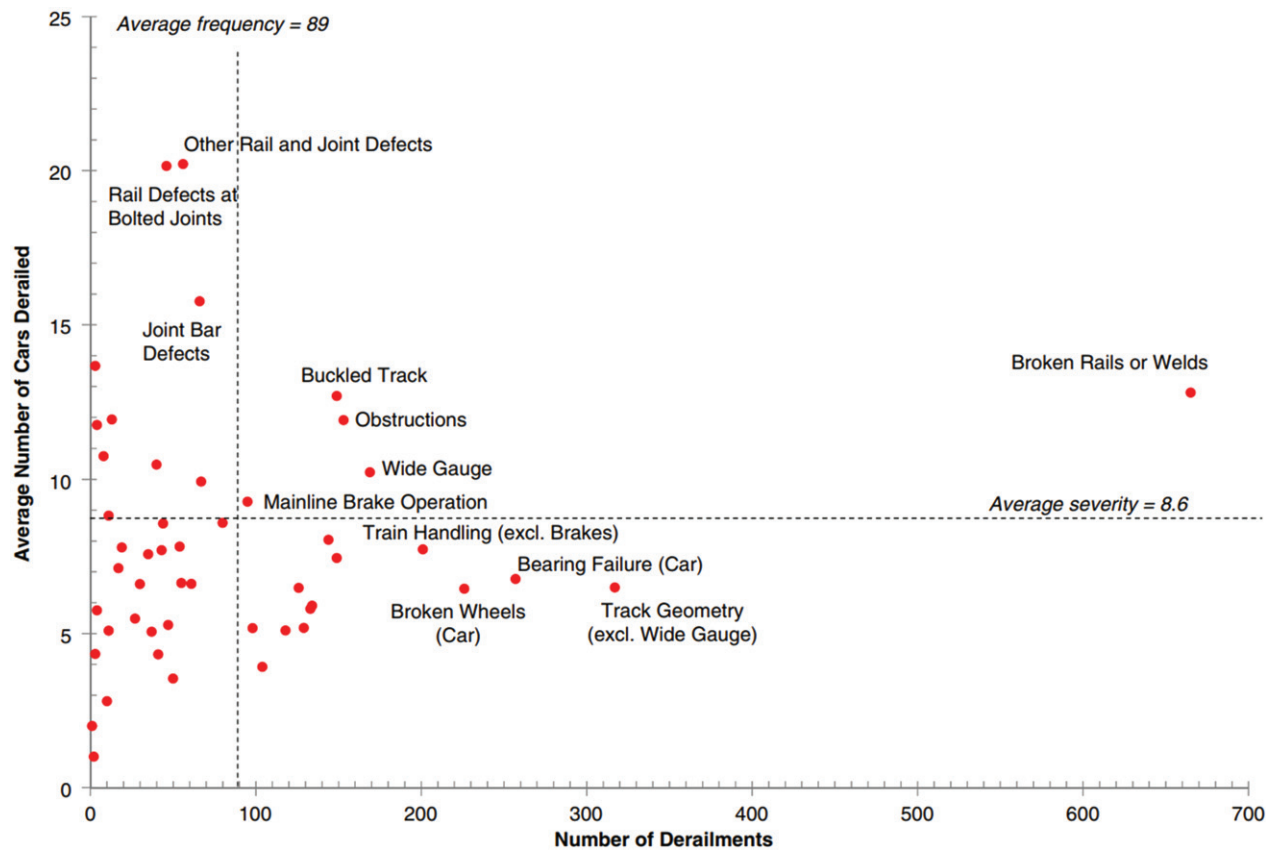


Figure 5 Derailment frequency and severity by accident cause on class I mainline, sorted by frequency (Liu et al., 2012).

Collision

Below, we will look in depth at US train collisions in recent years.

From 2001 to 2015, 394 freight train collisions occurred in the United States. Each collision led to an average of over one casualty per accident and \$0.67 million of monetary damage costs to infrastructure and rolling stock. There are 16 causes (related to human factors, signals, or miscellaneous cause groups) associated with collisions on US mainlines. This frequency analysis shows that the failure to obey or display signals, violation of train speed rules, and violation of mainline operating rules are the top three collision causes from 2001 to 2015 (Fig. 7).

Collisions caused by failures to obey or display signals resulted in an average of around 2 casualties, including both fatalities and injuries, and around \$1.23 million of monetary damage costs to infrastructure and rolling stock. The cause group designated as the failure to obey or display signals includes but is not limited to the following cases:

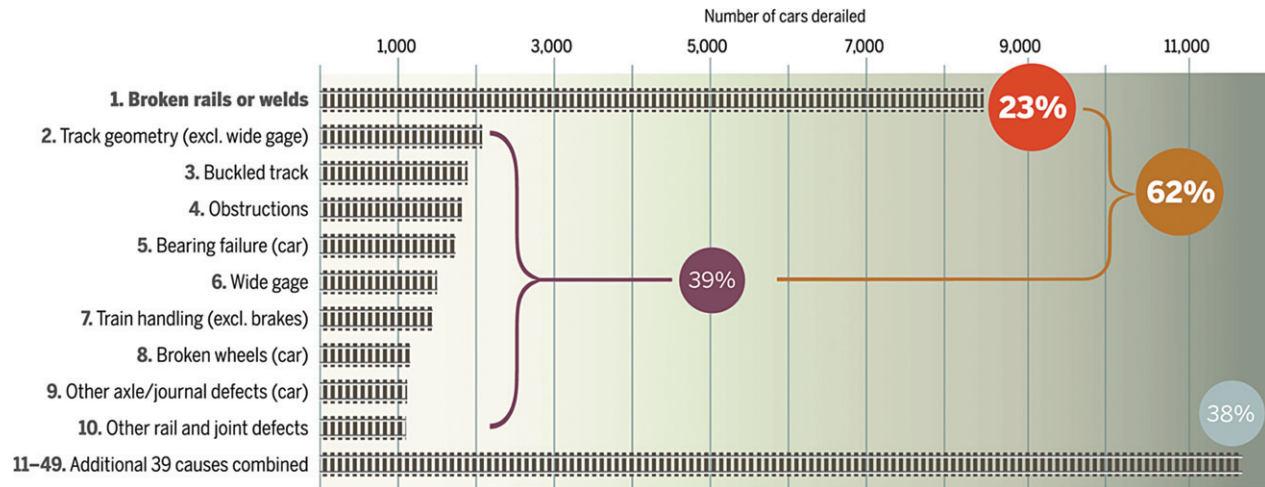


Figure 6 Primary causes for 2001–10 freight train derailments on class I mainlines by number of cars derailed. Source: Rutgers Center for Advanced Infrastructure and Transportation, 2019. RU Rail Expert's Research Chosen for Special Issue of "Transportation Research Record".

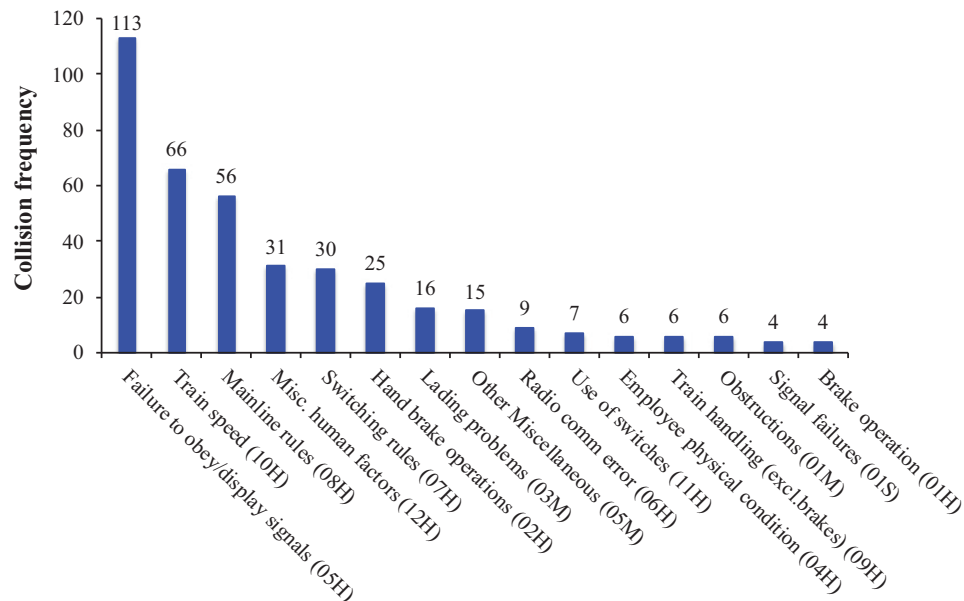


Figure 7 Number of collisions by cause, U.S. Mainlines, 2001–15 (Turla et al., 2018). H, Human factors in train operations; M, Miscellaneous causes; S, Signal and communication.

- Failure to observe hand signals given during a wayside inspection of a moving train;
- Failure to comply with failed equipment detector warnings or with applicable train inspection rules;
- Improperly displayed fixed signals (other than automatic block or interlocking signals);
- Failure to comply with fixed signals (other than automatic block or interlocking signals);
- Failure to comply with automatic block or interlocking signals;
- Absence of blue signals;
- Improper use of flagging;
- Failure to comply with flagging;
- Failure to comply with hand signals;
- Improper use of hand signals;
- Failure to give or receive hand signals.

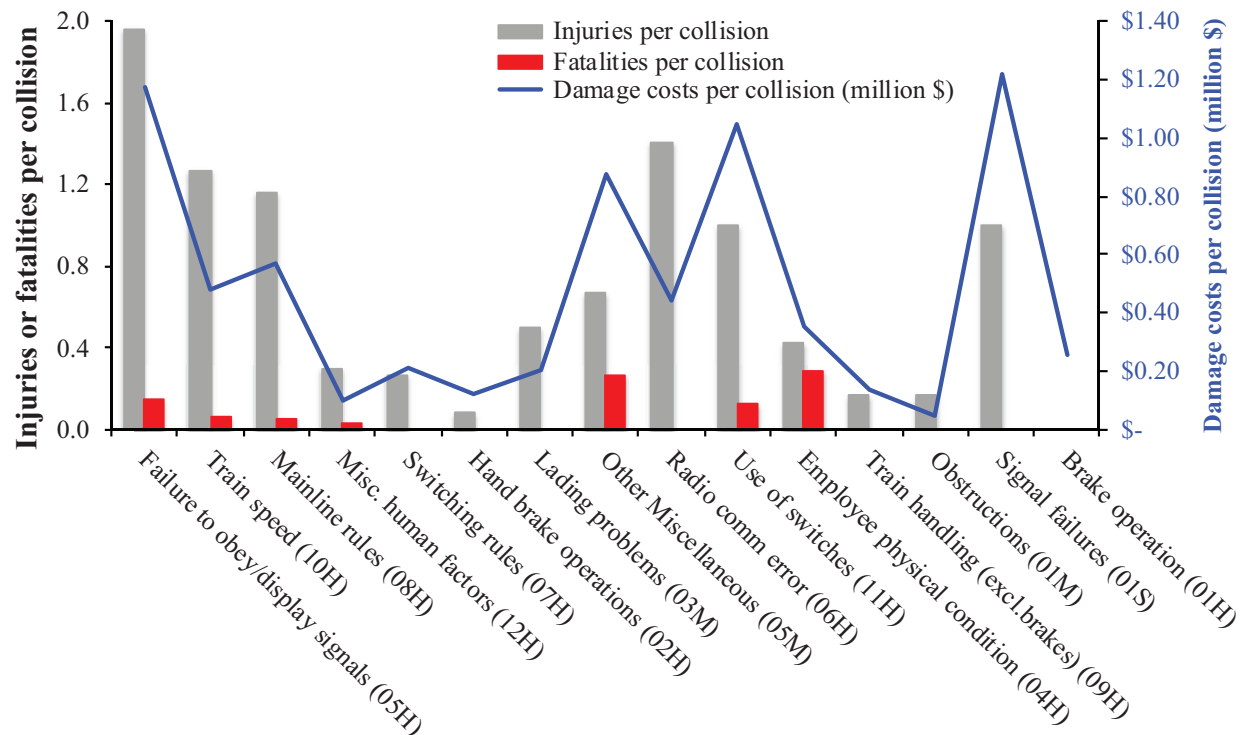


Figure 8 Collision severity by major cause, U.S. Mainlines (Turla et al., 2018).

The second most frequent collision cause group, violation of train speed rules, includes excessive coupling speed, failure to comply with restricted speeds in connection with the restrictive indication of a block or interlocking signal, trains moving at excessive speeds, and switching movement with excessive speeds. The cause group designated as the violation of mainline rules covers failure to stop the train, and failure to comply with motor car or on-track equipment rules, train orders, track warrants, track bulletins, timetable authority, etc. For the comprehensive definition of the miscellaneous cause group in freight-train collisions, refer to [ADL \(1996\)](#) and [FRA \(2011\)](#). In addition to the number of collisions by cause, the average severity in terms of injuries, fatalities, and damage costs per collision due to each cause group is also presented in [Fig. 8](#). Failure to obey/display signals, as the most frequent collision cause group, also has the highest number of injuries per collision and the greatest damage costs per collision.

Risk Mitigation Strategies

While rail safety remains at an all-time high, continuous efforts are essential to protect the public and rail workers. Continuous safety improvement requires comprehensive strategies designed to eliminate risks on railroads. The Federal Railroad Administration of the US Department of Transportation and railroad companies are collaborating to increase safety via three major domains: track, mechanical, and human factors. Safety improvements are ensured by merging proven safety approaches with performance-based measures that improve the culture around safety and harness technology and research.

Track-Related Risk Prevention

Rail Inspection

Rail inspection is the activity of examining railroad safety and serviceability, as one critical aspect of the operation of a rail network. A variety of nondestructive testing (NDT) technologies have been employed or developed ([Fig. 9](#)). For example, ultrasonic waves are employed to detect surface defects, internal railhead defects, and rail web and base defects. However, they can sometimes miss rail foot defects and can also miss internal defects and surface defects that are smaller than 4 mm at high speeds (e.g., greater than 70 km/h) ([Cheng and Bond, 2015](#)). In the detection of internal defects, radiography, laser ultrasonics, and electromagnetic acoustic transducers are three feasible technologies that overcome the limitations of most other means, although they can be adversely affected by lift-off variations.

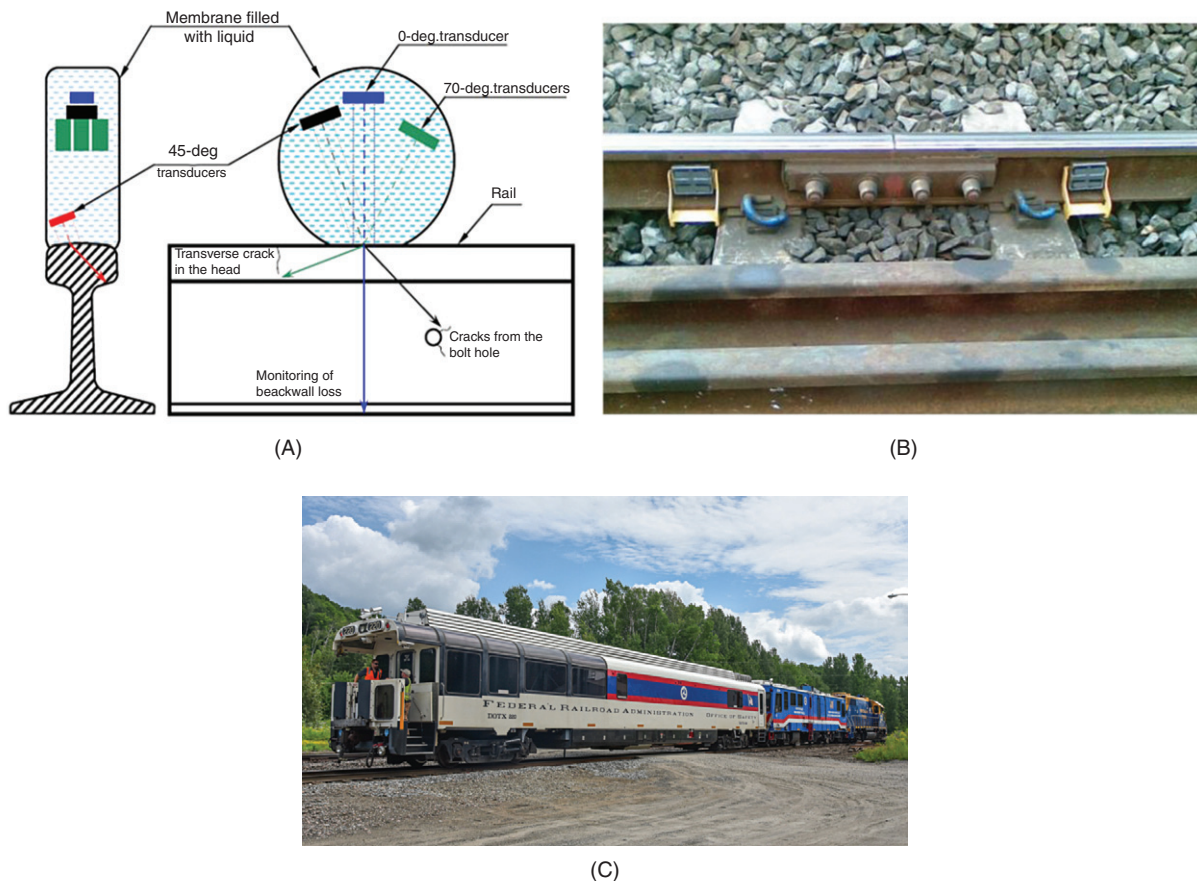


Figure 9 Rail track inspections. (A) Automated ultrasonic inspections with wheels; (B) Magnetic particles; (C) Rail Inspection vehicle. Source: (A) STARMANS, <http://www.starmans.net/applications/railway-rail-testing/>; (B) VolkerRail, <https://www.volkerrail.co.uk/en/home1/volkerrail-products/permanent-particle-rail-magnets>; (C) Richard Deuso. FRA rail inspection car. <http://photos.nerail.org/?p=245203>.

Track Geometry Measurement and Analytics

The inherent and small geometrical deviations in the position of rails from their ideal design states constitute imperfections that can have a significant impact on the safety and the rate of degradation of the rail system. Causal analysis discloses that track geometry is the second most frequent FRA-reportable derailment cause on Class I freight railroad mainlines. Track geometry measurement technologies and assessment using various statistical techniques contribute to rail track safety and maintenance decisions. Main track geometry defects include alignment, profile, gauge, cant, and twist defects.

The Autonomous Track Geometry Measurement Systems (ATGMS) technology is employed to improve rail safety by increasing the availability of track geometry data (Fig. 10). Routine use of ATGMS technology by the rail industry will eventually lead to minimized interference of inspections to revenue operations, increased inspection frequencies, and the reduced lifecycle cost of



Figure 10 Autonomous track geometry measurement system technology (ProgressiveRailroading.com/railproducts, 2016).

inspection operations, all of which will lead to improved safety and maintenance planning. There are three major components in ATGMS: onboard units, wireless communication links, and processing servers ([Carr et al., 2009](#)).

- Onboard units consist of a series of sensors, a computing platform, and location determination technology, such as Global Positioning System (GPS) technologies.
- Processing servers feature data processing capabilities, a database management system, and Geographic Information Systems (GIS) applications to facilitate the reporting of areas of interest. The server-based software performs a series of quality checks to ensure the continuity of data and subsequently converts sensor data into foot-by-foot geometry measurements.
- Communication links, constituting an important feature of any autonomous inspection system, are the manner in which data is transferred between the onboard units, the central processing servers, and the data recipients. They affect the way in which measured data is uploaded to the central processing servers, and processed information is distributed to inspectors or maintenance personnel.

The implementation of ATGMS can improve rail safety by increasing the availability of track geometry data for safety and maintenance planning purposes. The key feature of this technology is that it is a relatively low-cost, self-powered geometry measurement system for a wide range of rail vehicles, especially for freight trains.

Moreover, additional technologies and approaches have also been employed in the United States. [Andani et al. \(2015\)](#) investigated a light detection and ranging (LIDAR) sensor for the collection of horizontal alignment data, and their test, carried out on a revenue service track, indicated that LIDAR optics can provide a reliable track monitoring instrument with a wavelength range of 9.45–18.9 m for use over substantial track mileage in inclement weather and harsh track conditions with minimal operator supervision. Furthermore, the BNSF Railway has utilized unmanned aerial vehicles (UAV) to monitor track quality while safety experts monitored the drones' video feeds to precisely view rail tracks.

To characterize and predict track geometry component degradation over time, track geometry can be modeled in statistical forms. The Track Quality Index (TQI) was developed for the evaluation of the quality indicators of track geometry, in order to determine which are most conducive to accurately describing and predicting track geometry. One methodology is to employ multiple regressions to varying parameters, such as vertical and horizontal alignments, gauge, cant and twist qualities, and then filter wavelengths across curved and straight tracks. The standard deviation of the vertical alignment with a wavelength between 3 and 25 m was found to be a good parameter to describe the influence of track geometry on vehicle behavior ([Haigermoser et al., 2014](#)). In addition to the use of a TQI, Markov-based models ([Bai et al., 2015](#)), neural networks ([Sadeghi and Askarinejad, 2012](#)), and probabilistic stochastic approaches ([Andrade and Teixeira, 2011](#)) have all been employed in the study of track geometry.

Improved Materials and Manufacturing

Over the last few decades, the quality of steel manufacturing has kept improving, eliminating many fatigue failures due to internal defects. Internal defects are generally caused by inherent flaws in the rail that form when the rail is manufactured. Thus, advances in rail manufacturing can prevent the occurrence of internal defects that are nearly impossible to detect at early stages. [Smith \(2005\)](#) stated that the greatest improvement to steel manufacture in the last 30 years has been the use of welding to eliminate fish-plated gaps in the running surface, which is probably the most significant development since the introduction of the steel rail. Rail is now manufactured in strings up to 250 m long, simplifying the laying of track and reducing the number of welds. Although welds are one potential source of weakness, the thermit welding process, used in the field to join long rail strings, is continuously being improved as well. Furthermore, with the development of high-speed railways in some areas, aerospace technologies are increasingly being adapted in terms of materials (e.g., aluminums and composites), manufacturing methods, and inspection techniques ([Smith, 2005](#)).

Mechanical-Related Strategies

Manual Inspection

The Federal Railroad Administration (FRA) of the US Department of Transportation requires that every car placed in the train must receive a mechanical inspection before the train departs from a yard or terminal. In addition, trains traveling long distances must stop for inspection every 1000 miles, with an exception under special conditions that allows travel up to 3500 miles.

Traditional railcar inspection practices require a car inspector, referred to as a carman, to walk or take a vehicle along the entire length of a train, visually inspecting the mechanical components on each car ([Schlake et al., 2010](#)) ([Fig. 11](#)). In most cases, two carmen are assigned to each train so that one can inspect each side of the train. Although there is slight variance by railroad and location, the common procedures for Class I railroad inspection performed by car inspectors remain relatively consistent and can be divided into several distinct steps: (1) secure track protection, (2) inspect train, (3) make necessary minor repairs and/or identify bad-order cars, (4) perform air test, (5) release brakes, (6) verify brake release, (7) remove track protection.

Depending on the carmen's training, experience, and working conditions, these inspections can be inconsistent in their objectivity and effectiveness. As a result, a train may be inspected to the best of a particular carman's ability, yet defects may be missed ([Schlake et al., 2010](#)). There are a variety of safe, efficient, and traceable means of rolling stock inspection that automate the mechanical inspection process by employing multiple technologies. These technologies include but are not limited to: wheel impact load detectors (WILDs); truck performance detectors (TPDs); acoustic bearing detectors (ABDs); and machine vision inspection of



Figure 11 Train inspections with carmen (Schlake, 2011).

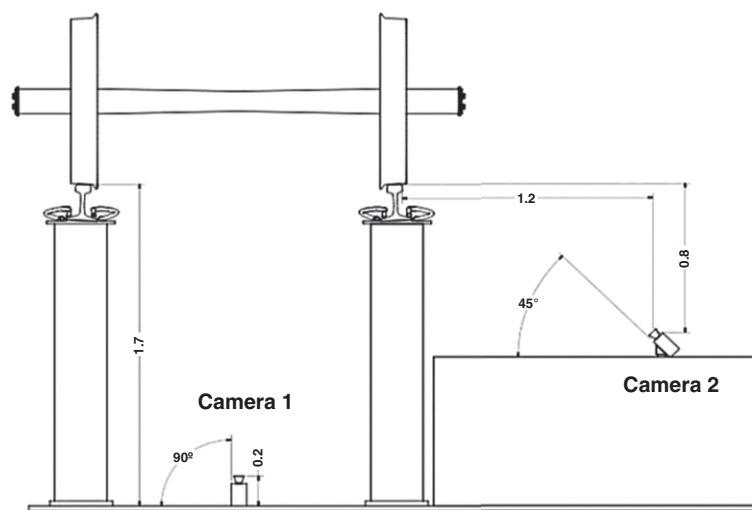


Figure 12 Architecture of measurement equipment (Schlake et al., 2010).

truck components, safety appliances, and brake shoes. The machine vision-based railcar inspection and acoustic bearing detectors are introduced below as two examples.

Machine Vision-Based Inspection

The inspection of railcar underframe components using machine vision technology is used to automate critical inspection tasks with high efficiency and effectiveness. A digital video system was developed to record images of railcar underframes, along with computer software to identify components and assess their condition. Railcar structural underframe components include the center sill, side sills, and crossbearers, which are subject to fatigue cracking due to periodic and/or cyclic loading during service and other forms of damage. Tests of the image recording system were conducted at Norfolk Southern (NS) maintenance facility in Decatur, Illinois, with two cameras (Fig. 12).

With the help of multi-scale image segmentation, pixel neighborhoods of varying size can be accessed and used for the detection and inspection of defects and cracks in structural underframe components. A crack can be modeled as a homogeneous, elongated region that appears darker than the center sill. Similarly, a break can be modeled as a dark region that represents a discontinuity in the following properties of the center sill: brightness, contiguity of the sill's contours, and co-linearity and parallelism with parts of the center sill's contours. For example, to detect a cracked center sill, the technology can identify the region that delineates the boundary of the center sill. With image segmentation, smaller subregions embedded in the region occupied by the center sill can be identified and compared to the models developed for cracks (Fig. 13). The images collected in these tests were used to develop machine vision algorithms to analyze images of railcar underframes and assess the condition of certain structural components, such as bearing failure and broken wheels, two leading causes of freight-train derailments on mainlines.

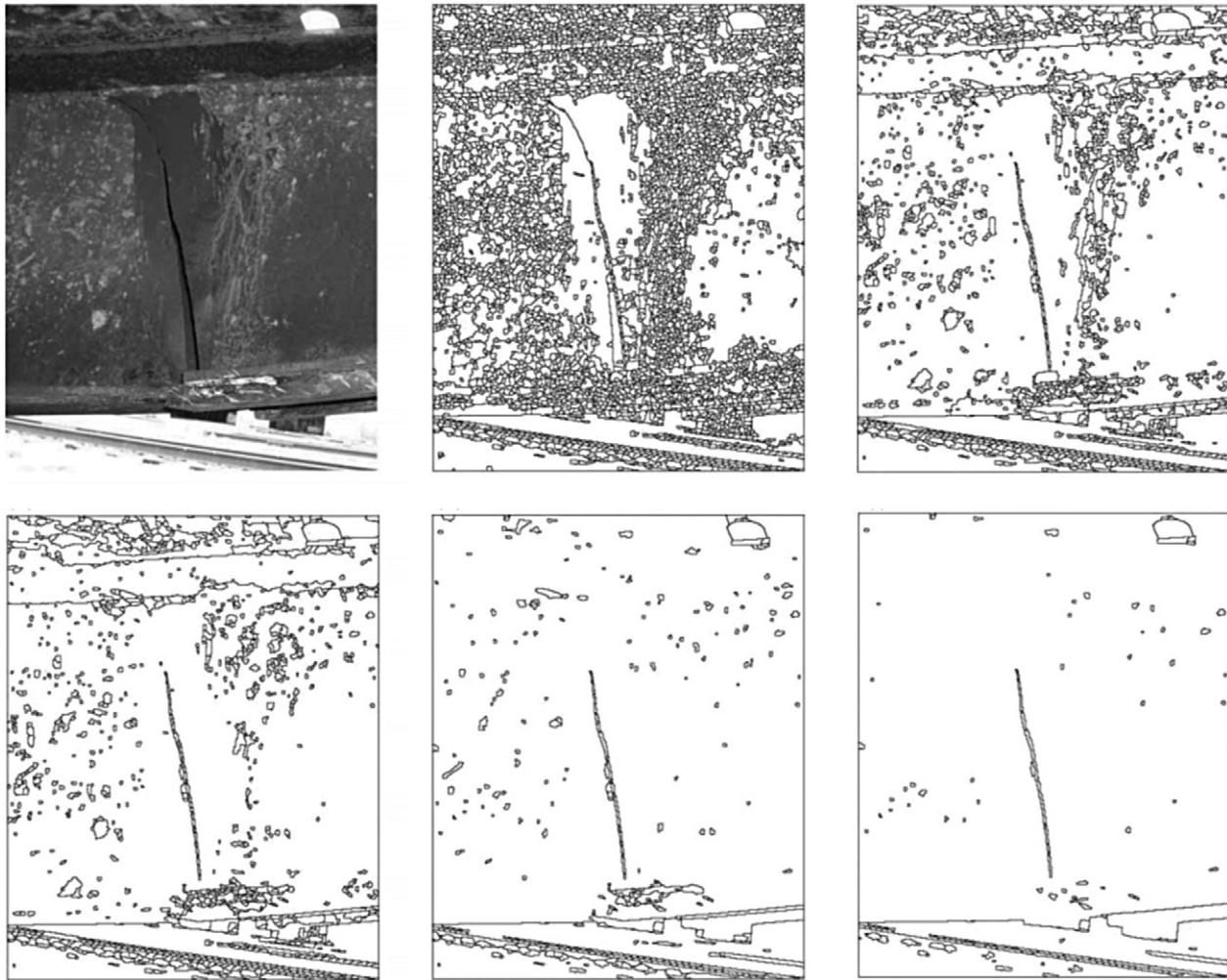


Figure 13 A cracked center sill in the original digital image and segmentation images (Schlake et al., 2010).

Acoustic Emission-Based Detection

Equipment causes, such as bearing failure and broken rails, are also high-frequency derailment accident causes in the United States. In particular, the continuous increase in train operating speeds means that failure of a wheelset can lead to serious derailments, causing loss of life and severe disruption in the operation of the network. A method of wayside detection of axle bearing defects using time spectral kurtosis analysis on high-frequency acoustic emission (AE) signals was developed by Amini et al. (2016) in order to effectively prevent risks from defects. Instead of using the hot axle box detectors, which are commonly used to detect overheating axle bearings but cannot detect damage to axle bearings at its early stages, this wayside detection technology uses high-frequency AE to collect feedback about axle bearing defects and then process the signal to remove unwanted noise produced by the engine, wheel-rail interactions, and train acceleration/deceleration with spectral kurtosis, an analytical technique that measures kurtosis values in both the frequency and time domains. Both laboratory tests and field experiments were conducted to validate the feasibility of this axle bearing wayside detection technology (Fig. 14).

Human Factor-Relation Strategies

Positive Train Control (PTC)

Positive Train Control (PTC) or similar systems such as Automatic Train Control (ATC) have been used in Japan since the 1960s and in many European countries since the 1980s. For example, in Sweden, the development of ATC started in the 1960s and was formally introduced in the early 1980s, together with high-speed trains. By 2008, 9831 km out of the 11,904 km of track maintained by the Swedish Transport Administration—the Swedish agency responsible for railway infrastructure—had ATC installed. In the United States, it is still an emerging technology and its implementation was mandated by the Rail Safety Improvement Act (RSIA) of 2008. The territory of PTC implementation and operation in the United States includes Class I railroads, main lines servicing over 5 million gross tons (MGT) annually and over which toxic or poisonous-by-inhalation hazardous materials are transported, and main lines



Figure 14 Axle bearing detection in experimental work and field-testing configuration (Amini et al., 2016).

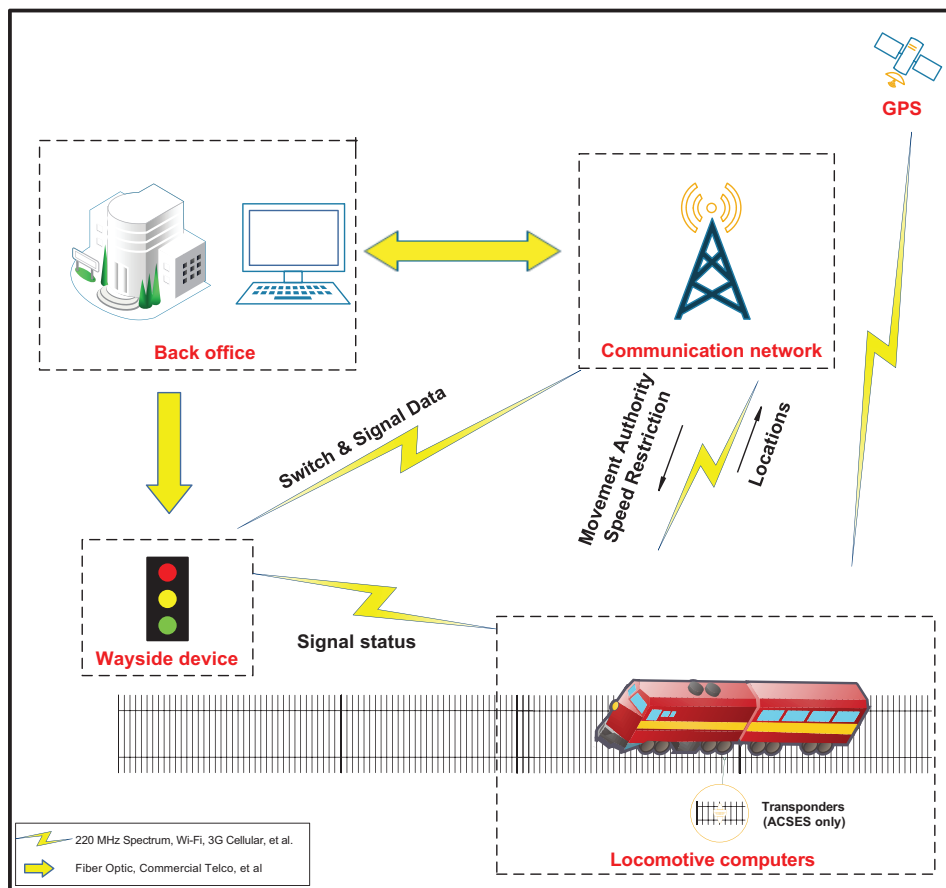


Figure 15 Architecture of PTC system (Zhang et al., 2018).

involving intercity and commuter passenger trains. PTC is a communication-based/processor-based train control technology that has the potential to improve safety because it provides a layer of additional protection beyond that provided by train crews and dispatchers (FRA, 2007b). More specifically, it can automatically prevent accidents attributable to human error by slowing or stopping trains, and is designed to prevent four major types of accidents: train-to-train collisions, derailments caused by excessive speeds, unauthorized incursions into work zones, and movements through misaligned switches.

The PTC system integrates the locomotive computer, wayside device, communication network, and back office to process collected movement authority and speed restriction data, and then compares these against the train's real conditions to ensure safety compliance (Fig. 15). With essential train information, such as real-time location, direction, and speed, if the PTC system were



Figure 16 Pulse electronics train Sentry III in an F-40 Cab (Oman and Liu, 2007). 4×3 LED array provides visual alert.

to determine the occurrence of a noncompliant train operation, then the PTC system would automatically apply the brakes and bring the train to a positive stop. Fig. 14 presents the network arrangement of various components integrated in PTC. For more technical details of the PTC system, refer to Petit (2009).

PTC describes a suite of train control standards. Railroads in the United States are allowed to install different PTC technologies in their respective systems once the types are approved and certified by the Federal Railroad Administration, including the Advanced Civil Speed Enforcement System, Interoperable Electronic Train Management System (I-ETMS), Enhanced Automatic Train Control (E-ATC), Incremental Train Control System (ITCS), Communications Based Train Control (CBTC), SafeNet System, and Sentinel System. In particular, ACSES and I-ETMS are two of the most widely implemented types of PTC system. In an ACSES-type PTC system, the ACSES system works in conjunction with the existing Automatic Train Control (ATC) system. Specifically, the ATC system ensures safe train separation and signal speed enforcement, while ACSES acts as an overlay to enforce civil speed restrictions (the maximum speed authorized for each section of track), temporary speed restrictions (e.g., temporary work zone), and positive train stops (PTS). One major feature of the ACSES-type PTC system is the employment of transponders. Differing from most types of PTC system that use GPS to identify train position, the ACSES establishes the train's exact position from the transponder sets it encounters. I-ETMS, by contrast, uses GPS to achieve train location and navigation information, and the locomotive segment monitors the train's real-time movement authority and speed, based on the collected information. The system would allow the train to proceed only if the switches are properly aligned and the speed limit is strictly observed. Unauthorized movement, overspeeding, or misaligned switches would result in penalty braking to slow the train down or even a PTS.

Engineer Education and Training

Heavy workloads, fatigue, monotony, and boredom in train engineers are found to be the issues leading to human error accidents (Dorrian et al., 2006). Fatigue is also responsible for a lack of attentiveness, which results in negligently disregarding signals or rules. Therefore, accidents caused by human errors may be reduced with periodic train operation education, a well-established working schedule, and an effective engineer fatigue monitoring program. Most incidents are linked to at least one organizational influence, which suggests that improvements in resource management, organizational climate, and organizational processes are important (Baysari et al., 2008). Train driving requires extensive knowledge of operation rules and vehicle behavior, along with the ability to integrate different static and dynamic sources of information (Giesemann, 2013); any failures of knowledge or ability may result in accidents due to violations of operation, mainline rules, or speed restrictions.

A variety of strategies and practices are being implemented on railroads. For example, a train driver training simulator was developed to train multiple drivers simultaneously and to review performance in detail (Rowe, 2013). This effective simulator includes the simulation of essential tasks (e.g., being able to look around while driving, taking power/applying brakes). Both normal and abnormal driving were documented in a full task list and analyzed with task assessment observations. To mitigate train engineer fatigue risk, a real-time online prototype driver-fatigue monitor was proposed by Ji et al. (2004). This non-intrusive monitor uses a prototype computer vision system for the real-time capture of video images of the driver and monitors the driver's vigilance. Similarly, cab alerts have been designed and implemented to alert the train crew by emitting a flashing light and an alarm (Oman and Liu, 2007) (Fig. 16).

References

- Amini, A., Entezami, M., Papaelias, M., 2016. Onboard detection of railway axle bearing defects using envelope analysis of high frequency acoustic emission signals. *Case Studies in Nondestructive Testing and Evaluation*, 6, pp. 8–16.
- Andani, M.T., Ahmadian, M., Munoz, J., O'Connor, T., Ha, D., 2015, March. On the application of LIDAR sensors for track geometry monitoring. In 2015 Joint Rail Conference (pp. V001T01A004-V001T01A004). American Society of Mechanical Engineers.

- Andrade, A.R., Teixeira, P.F., 2011. Uncertainty in rail-track geometry degradation: Lisbon-Oporto line case study. *J. Transport. Eng.* 137 (3), 193–200.
- Association of American Railroads (AAR), 2018. Overview of America's Freight Railroads, October 2018.
- Association of American Railroads (AAR), 2019. Freight Rail's Safety Record. <https://www.aar.org/issue/freight-rail-safety-record/>, Accessed by May 2019.
- Arthur D. Little, Inc. (ADL), 1996. Risk Assessment for the Transportation of Hazardous Materials by Rail, Supplementary Report: Railroad Accident Rate and Risk Reduction Option Effectiveness Analysis and Data, 2nd rev. ADL, Cambridge, Mass.
- Bai, L., Liu, R., Sun, Q., Wang, F., Xu, P., 2015. Markov-based model for the prediction of railway track irregularities. *Proc. Inst. Mech. Eng. Part F: J. Rail Rapid Transit* 229 (2), 150–159.
- Baysari, M.T., McIntosh, A.S., Wilson, J.R., 2008. Understanding the human factors contribution to railway accidents and incidents in Australia. *Accid. Anal. Prev.* 40 (5), 1750–1757.
- Carr, G., Tajaddini, A., Nejtkovsky, B., 2009, September. Autonomous track inspection systems—today and tomorrow. In AREMA 2009 Annual Conference and Exposition.
- Cheng, J., Bond, L.J., 2015, March. Assessment of ultrasonic NDT methods for high speed rail inspection. In AIP Conference Proceedings, vol. 1650, 1, pp. 605–614, AIP.
- Dorrian, J., Roach, G.D., Fletcher, A., Dawson, D., 2006. The effects of fatigue on train handling during speed restrictions. *Transport. Res. Part F* 9, 243–257.
- Evans, A.W., 2010. Rail safety and rail privatisation in Japan. *Acc. Anal. Prev.* 42 (4), 1296–1301.
- Evans, A.W., 2013. The economics of railway safety. *Res. Transport Econ.* 43 (1), 137–147.
- Federal Railroad Administration, 2011. FRA Guide for Preparing Accident/Incident Reports. Washington, DC.
- Federal Railroad Administration, 2016. Rail Safety Fact Sheet. December 16, 2016. Washington, DC.
- Federal Railroad Administration, 2019. Train Accidents and Incidents. Washington, DC. <https://safetydata.fra.dot.gov/SASOnflyoutput/EZBUTTON/DIR33030/index.html> Accessed by May 13, 2019.
- Giesemann, S., 2013. Automation effects in train driving with train protection systems—assessing person-and task-related factors. In: *Rail Human Factors. Supporting Reliability, Safety and Cost Reduction*, pp. 139–149.
- Haigermoser, A., Eickhoff, B., Thomas, D., Coudert, F., Grabner, G., Zacher, M., Kraft, S., Bezin, Y., 2014. Describing and assessing track geometry quality. *Vehicle System Dynamics* 52 (sup1), 189–206.
- Ji, Q., Zhu, Z., Lan, P., 2004. Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE Trans. Vehicular Technol.* 53 (4), 1052–1068.
- Liu, X., Saat, M.R., Barkan, C.P.L., 2012. Analysis of causes of major train derailment and their effect on accident rates. *Transport. Res. Rec. J. Transport. Res. Board* 2289, 154–163.
- Oman, C.M., Liu, A.M., 2007. Locomotive In-Cab Alert Technology Assessment. MVL Publication, 7.
- Petit, W.A., 2009. Interoperable positive train control. *IEEE Vehicular Technology Magazine* 4 (4).
- Rowe, I., 2013. Developing a tram driver route learning training simulator for Manchester's Metrolink trams. *Rail Human Factors: Supporting reliability, safety and cost reduction*, p.262.
- Sadeghi, J., Askarinejad, H., 2012. Application of neural networks in evaluation of railway track quality condition. *J. Mech. Sci. Technol.* 26 (1), 113–122.
- Schlake, B., 2011. Impact of automated condition monitoring technologies on railroad safety and efficiency.
- Schlake, B.W., Todorovic, S., Edwards, J.R., Hart, J.M., Ahuja, N., Barkan, C.P., 2010. Machine vision condition monitoring of heavy-axle load railcar structural underframe components. *Proc. Inst. Mech. Eng., Part F: J. Rail Rapid Transit* 224 (5), 499–511.
- Smith, R.A., 2005. Railway fatigue failures: an overview of a long standing problem. *Materialwissenschaft und Werkstofftechnik. Entwicklung, Fertigung, Prüfung, Eigenschaften und Anwendungen technischer Werkstoffe* 36 (11), 697–705.
- Traveler, 2019. Passenger Trains in America. https://traveler.sharemap.org/Passenger_trains_in_America#W1stMTU1LjM0MDkyLDlyLjkzMDA3SXBtLU0LjgwMDM4LDC0LjcyNzc0XV0=, Accessed in May 2019.
- Turla, T., Liu, X., Zhang, Z., 2018. Analysis of freight train collision risk in the United States. *Proc. Inst. Mech. Eng. Part F: J. Rail Rapid Transit*, 0954409718811742.
- United Nations Economic Commission for Europe (UNECE), 2019. Total Length of Railway Lines. <https://w3.unece.org/PXWeb/en/CountryRanking?IndicatorCode=42>.
- World Bank, 2019. Rail Lines from International Union of Railways (UIC). https://data.worldbank.org/indicator/is.rrs.totl.km?most_recent_value_desc=true.
- Zhang, Z.P., Liu, X., Holt, K., 2018. Positive Train Control (PTC) for railway safety in the United States: Policy developments and critical issues. *Utilities Policy* 51, 33–40.

Further Reading

- Cambridge Systematics, Inc., National Rail Freight Infrastructure Capacity and Investment Study, September 2007.
- Hartong, M., Goel, R., Wijesekera, D., 2011. Positive train control (PTC) failure modes. *J. King Saud Univ.-Sci.* 23 (3), 311–321.
- Higgins, C., Liu, X., 2018. Modeling of track geometry degradation and decisions on safety and maintenance: A literature review and possible future research directions. *Proc. Inst. Mech. Eng. Part F: J. Rail Rapid Transit* 232 (5), 1385–1397.
- Liu, X., Barkan, C.P.L., Saat, M.R., 2011. Analysis of derailments by accident cause: evaluating railroad track upgrades to reduce transportation risk. *Transport. Res. Rec. - J. Transport. Res. Board* 2261, 178–185.
- Saadat, S., Stuart, C., Carr, G., Payne, J., 2014, October. Development and use of FRA autonomous track geometry measurement system technology. In AREMA 2014 annual conference, Chicago, Illinois, pp. 28-1.
- Schafer, D.H., & Barkan, C.P. 2008. A prediction model for broken rails and an analysis of their economic impact. In Proceedings of American Railway Engineering and Maintenance of Way Association (AREMA) Annual Conference.

Railroad Safety: Grade Crossings and Trespassing

Rahim F. Benekohal, Jacob Mathew, University of Illinois at Urbana-Champaign, Champaign, IL, United States

© 2021 Elsevier Ltd. All rights reserved.

Overview of Railroad Safety	466
Safety at Highway-Rail Grade Crossings	466
Reported Accidents at Railroad Grade Crossings	468
Traffic Control Devices at Railroad Grade Crossing	468
Gates	468
Flashing Lights	468
Crossbucks	468
Operation of Active Control Devices	469
Interconnection and Preemption	471
Pre-Signal	471
Grade Crossing Elimination	471
Grade Separation	472
Potential Grade Crossing Safety Contributing Factors	472
Modeling Safety at Railroad Grade Crossings	472
Trespassing and Pedestrian Safety	472
Pedestrian Safety at Railroad Grade Crossings	473
Emerging Technologies to Improve Railroad Safety	474
Intelligent Transportation Systems in Railroads	474
Positive Train Control	475
Connected Vehicles	475
Detection Technologies for Pedestrian Detection	475
Relevant Websites	476
References	476

Overview of Railroad Safety

Accidents on rail lines may be a collision between a highway vehicle, bicyclist or pedestrian and a train at a grade crossing, a collision between two trains, incidents involving pedestrians and other trespassers on rail property away from crossings, derailments, or any other events involving on-track equipment. In order to be reportable, the damage typically needs to be above an established threshold. Railroad accidents are rare compared to roadway accidents (annually around 17,000 railroad accident compared to over 7 million highway accidents in the United States) but the proportion leading to fatalities is higher (6%) compared to roadways (0.5%).

There has been a decline in the number of railroad accidents in the last 3 decades in the United States; yet, every year there are many railroad accidents. In the recent past, there has even been an increase in the number of trespassing and pedestrian related incidents. Trespassers have the largest share of the railroad fatal accidents in the United States, followed by railroad grade crossings (Fig. 1) (Savage, 2013). The distribution of significant railroad accident in the European Union is shown in Fig. 2 (European Union Agency for Railway, 2018).

Safety at Highway-Rail Grade Crossings

Highway-rail grade crossings are locations where a highway intersects with a railroad at the same level (at grade). In the United States, there are over 200,000 railroad grade crossings of which over 130,000 are public crossings. Accidents at the US grade crossings reached the lowest in 2007 (since the beginning of record keeping) but have remained nearly at that level for the past 10 years, as shown in Fig. 3. A similar trend is also observed in the EU where the number of accidents reached a low point in 2010 and has been more or less unchanged since then.

Most railroad crossings remain accident free year after year, but a small portion of the crossings has accidents in more than 1 year. There were 6301 public crossings in the United States that had a motor vehicle accident in the past 10 years (2009–18). Table 1 shows the number of years, in the 10-year time period, a crossing had accidents and its frequency. None of the 6301 crossings had accidents in 9 or more years out of the 10 years. Only one crossing had accident in 8 of 10 years, 216 crossing had at least one accident in 3 of the 10 years, 890 crossings had at least one accidents in 2 of the 10 years, and 5083 crossings had at least one accident in 1 of those 10 years. Each year around 700 crossings that did not have an accident in the previous year is seen to have an accident.

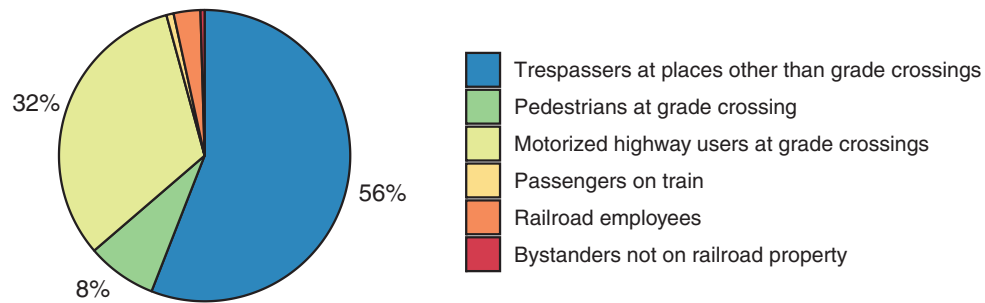


Figure 1 Percentage of mainline railroad fatalities in the United States.

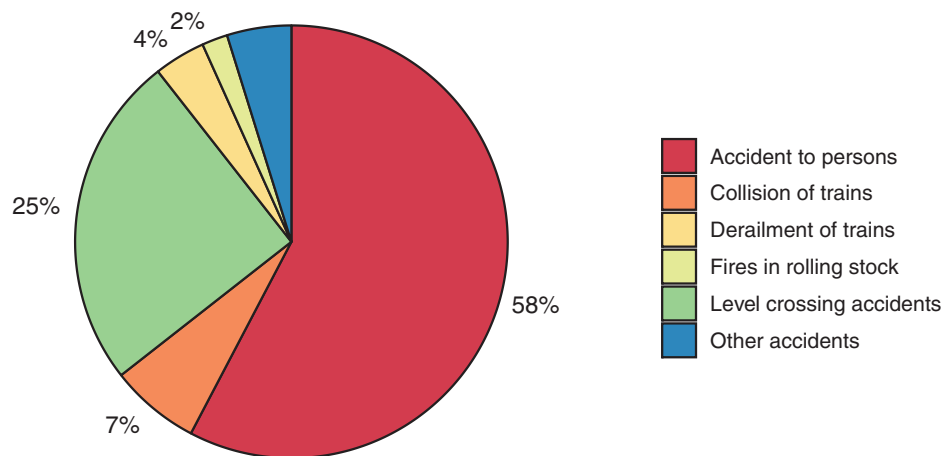


Figure 2 Percentage of significant accidents in the EU.

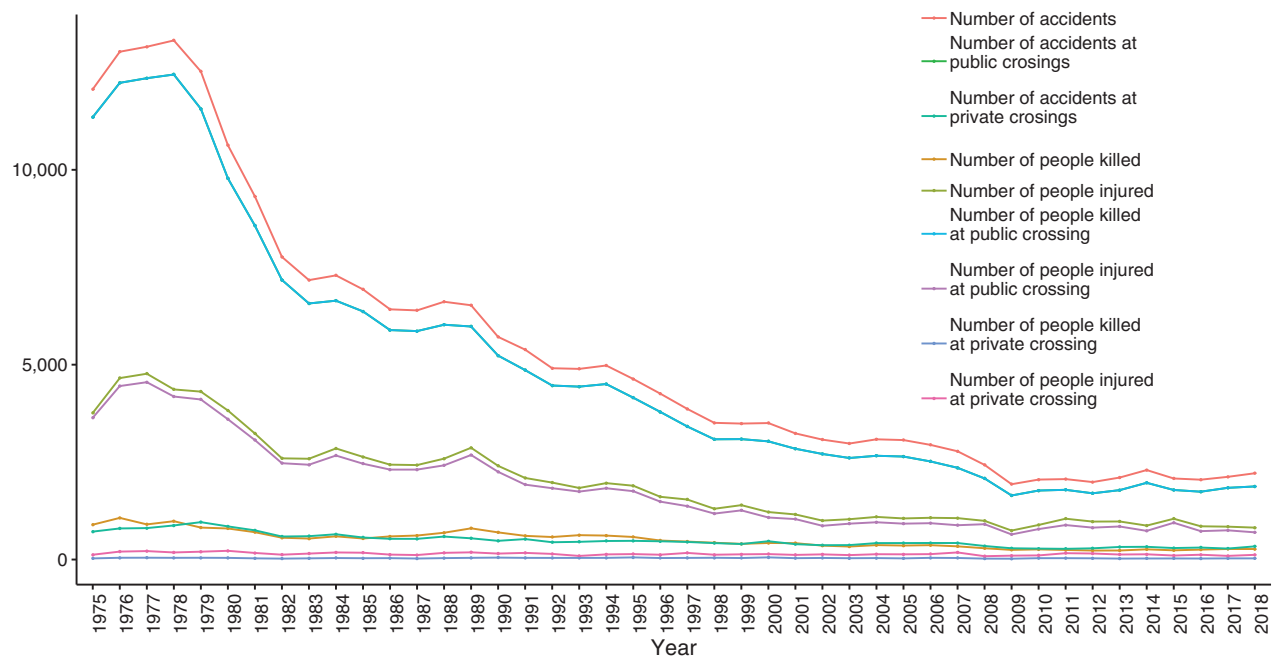


Figure 3 Accident trends at grade crossings in the United States.

Table 1 Railroad crossings, which had accidents in 10 years between 2009 and 2018

<i>Number of years crossing had accident between 2009 and 2018</i>	<i>Number of crossings</i>
1	5083
2	890
3	216
4	72
5	22
6	10
7	7
8	1
9	0
10	0

Reported Accidents at Railroad Grade Crossings

In the United States, the Federal Railroad Administration (FRA) redefined grade crossing collision in 1975 as any impact “between railroad on-track equipment and an automobile, bus, truck, motorcycle, bicycle, farm vehicle, pedestrian or other highway user at a rail-highway crossing (Ogden, 2007).” Before that, national statistics were limited to incidents involving a fatality, an injury to a person sufficient to incapacitate him or her for a period of at least 24 h, or more than \$750 in damage to railroad equipment or railroad track. This is however not the case in the EU, as the current legislation does not require the reporting of all railroad accidents by its member states. Accident reporting is limited to significant accidents or accidents defined by the EU as Common Safety Indicators (CSI).

Traffic Control Devices at Railroad Grade Crossing

The types of traffic control devices (TCDs) at a grade crossing could be broadly classified into active control devices and passive control devices. Active devices are those that warn a highway user of the train approaching or occupying the crossing. The most commonly used active TCDs include flashing light signals and gates. Other active controls include bells, wigwags, active advance warning devices, and highway traffic signals. Passive devices are those that are not activated by the train, and include pavement marking and signs. However, the trains themselves may have whistles acting as audible “active” warning signals at these locations. The number of accidents at active crossings is generally larger due to the high volume of vehicles and trains; however, the number of accidents at passive crossings normalized for exposure (or traffic moment = product of annual average daily traffic and average daily train traffic) is much higher than the rate for active crossings.

The type of warning device installed at a crossing is usually determined based on vehicle traffic count, type of vehicle using the crossing, number of daily trains and collision history at the crossing. Fig. 4 shows the number of US crossings by their warning device type. Around 50% of the public crossings are equipped with gates and 14% of crossings had flashing light warning devices without gates. Twenty five percent of the crossings have crossbucks. The remaining 11% have a variety of TCDs including stop signs, wigwags, etc. or have no warning device at all.

Gates

Gates are a type of active TCDs that function as a physical barrier across the highway lanes when a train is approaching or occupying the crossing. At a grade crossing, gates provide the highest level of traffic control and regulation. Gates are sometimes referred to as boom barriers.

Gated crossings are usually equipped with flashing light signals and crossbucks as well, as shown in Fig. 5. Most of the grade crossing incidents at gated crossings involved a motor vehicle going around the gate (around 30%) or stopped on the crossing (around 30%).

Flashing Lights

A standard flashing light signal in the United States consists of two red lights in a horizontal line flashing alternately at approaching highway traffic; the crossing is also equipped with crossbucks, as shown in Fig. 6. In some European countries, flashing light signals at crossings are sometimes also equipped with a white light which indicate to drivers and pedestrians that it is safe to cross the crossing. A high proportion (around 60%) of accidents at crossings with flashing lights involved motor vehicles that did not stop at the crossing.

Crossbucks

The railroad crossing sign or the crossbucks sign is a passive warning device that indicates a grade crossing. In the United States, crossbucks carry the word “RAILROAD CROSSING” written in black ink on a white background as shown in Fig. 7. Motorists

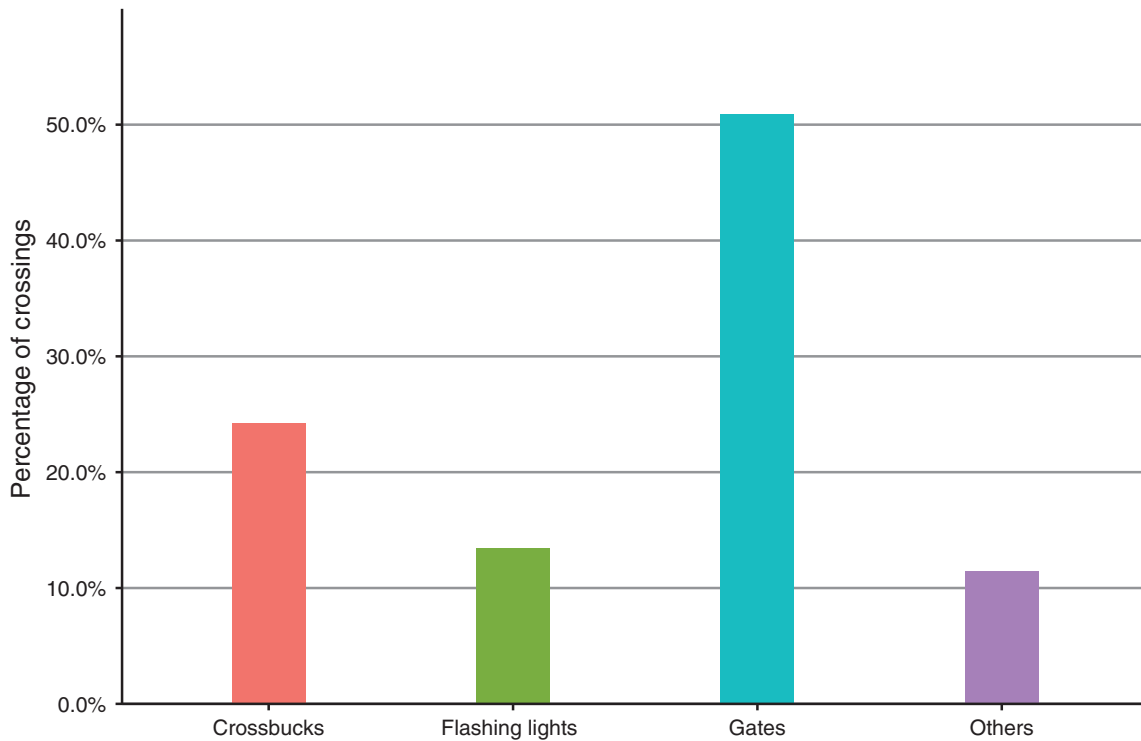


Figure 4 Percentage of crossings in the United States by warning device type. Source: FRA Office of Safety, 2019.



Figure 5 Railroad grade crossing with gates. Crossing is also equipped with flashing lights and crossbucks.

involved in accidents tend not to stop at crossings with crossbucks with over 60% of grade crossing accidents at crossbucks occurring to highway users that did not stop at the crossing.

Other warning devices at a highway-rail grade crossing include stop signs or other highway traffic signals, warning bells, wigwags, etc. Warning signs may be placed on crossing approaches to warn the highway user of an upcoming crossing known as advance warning signs, as shown in Fig. 8.

Operation of Active Control Devices

These control devices are activated by some form of train detection system. This is achieved either manually at a manned crossing or automatically using circuits within railroad tracks.



Figure 6 Railroad grade crossing with flashing lights.



Figure 7 Railroad grade crossing with only crossbucks.



Figure 8 Advance warning sign ahead of railroad grade crossing.

The train detection systems in use today include the following:

1. Direct current (DC) track circuit
2. AC–DC track circuit
3. Audio frequency overlay (AFO) track circuit
4. Motion-sensitive track circuit
5. Constant warning time track circuit

These detection systems operate in a way to provide a minimum of 20s of clearance time before the arrival of a train at the grade crossing. It is also important that the warning times provided at the crossing is not excessive as highway users could see this long duration as an equipment malfunction causing crossing violations (vehicle running around a closed gate) and accidents.

Interconnection and Preemption

If a highway–highway traffic signal exists close (within 200 ft.) to a railroad grade crossing location, the railroad and the highway traffic control may be interconnected (*Manual of Uniform Traffic Control Devices*, 2009). This allows for preemption of the highway traffic signal to increase safety at the grade crossing. Preemption of traffic signals is recommended if highway traffic queues have the potential for extending across a nearby rail crossing or if traffic backed up from a nearby downstream railroad crossing could interfere with signalized highway intersections. These recommendations are listed out by the Institute of Transportation Engineers (ITE) (ITE, 2006).

Preemption takes control of the traffic signal to provide for the safe passage of a train. Preemption of traffic signals allows for clearance of any queued vehicle on the crossing and prohibits any further movement of vehicles that would intersect the track. The preemption of the traffic signal occurs in three stages including entry into preemption, preemption hold, and exit from preemption. Entry into preemption happens when a train is detected and a call for preemption is sent to the highway signal. The mode of operation of the traffic signal is altered at this call. The preemption hold interval is active when the train is on the crossing. This remains on until the preemption input to the controller is removed. The purpose of the preemption hold is to ensure that the movements that do not interfere with the movement of the train may proceed. During the preemption hold period, all the movements that conflict with the train movement remains red while the other movements are allowed to proceed into the intersection. The traffic signal exits from preemption and enters its normal mode of operation after the train leaves the crossing.

Pre-Signal

Pre-signals are highway signals installed in front of railroad warning devices on a highway to stop the traffic before it crosses the railroad. A pre-signal shows red to an approach before the arrival of a train to prevent vehicles from entering the crossing. The vehicles already at the intersection are given the green signal so that they may clear. Pre-signals are installed in this manner to prevent vehicles from queuing across the grade crossing and finding themselves stopped on the tracks. The operation of pre-signals is integrated with the traffic signal preemption. In the United States, the *Manual on Uniform Traffic Control Devices* (MUTCD) lists out guidance for the use of pre-signals at grade crossings. **Fig. 9** shows a pre-signal.

Grade Crossing Elimination

Grade crossing elimination can be achieved by grade separating the crossing, by closing the crossing to highway and foot traffic or by closing the crossing to train traffic (abandonment of the rail line). Grade crossing elimination decisions are made by considering



Figure 9 Pre-signal indicating red to traffic approaching grade crossing.

safety, traffic operations, economic and social and environmental impacts, and costs. The benefits of eliminating a grade crossing are improved safety, reduced delay to train (trains tend to slow down at grade crossing) and highway vehicles, and avoidance of maintenance cost of crossing surface and TCDs.

Grade Separation

The highest level of traffic control and regulation for a grade crossing is the installation of crossing gates. If gates prove to be ineffective, a physical separation of the roadway and railway becomes necessary. Elimination of a grade crossing by the physical separation (grade separation) of the roadway and railway is called grade separation. Grade separation projects are costly and have long-term effects on the region and its users. Locations under consideration for grade separation should carefully be selected after considering several quantitative and qualitative factors.

Some of the techniques used to identify crossings for grade separation include multi-criteria assessment (MCA), cost benefit analysis, trade off analysis, etc. In the United States, the *Highway-Rail Grade Crossing Handbook* lists out several warrants to be considered before elimination of a grade crossing via grade separation (e.g., grade crossings should be considered for grade separation, if the maximum authorized train speed at the crossing exceeds 110 mph).

Potential Grade Crossing Safety Contributing Factors

The variables included in the model represent the traffic and geometric features of the crossing. Most hazard models developed for railroad grade crossings use annual average daily traffic (AADT) and train volume at the crossing in the original, or in a transformed form, or as a function of the two variables. Other variables include train speed, highway speed, number of highway lanes, number of train tracks, etc. Additional variables including the smallest angle between highway and rail track, distance to highway intersection, etc. are included in later models. The variables used in the hazard models could be categorized into safety-related variables, traffic-related variables, location, and crossing geometry variables. Other variables including economic development in the region, vehicle type using the crossing, community consideration (vulnerable population living near the region, use of grade crossing by emergency vehicles/school bus), etc., which could influence the hazard model, has not been included in any models so far.

Modeling Safety at Railroad Grade Crossings

Quantifying safety at highway-rail grade crossing is carried out to identify the crossings with high safety risk. Several hazard indices to identify the riskiest crossings for crossing closure or other safety improvements have been developed worldwide.

Modeling of safety at railroad grade crossings allows predicting the number of collisions, predicting the severity of accidents, prioritizing crossings for safety upgrades, or identifying patterns in accidents at an individual crossing or a group of crossings. In the United States, the USDOT accident prediction formula is used to predict the number of collisions at grade crossing locations.

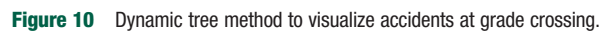
Tools like static tree and dynamic tree methods to visualize accident patterns by identifying the highest accident-contributing factor have also been developed. **Fig. 10** shows the accident visualization at an individual crossing using dynamic tree at a crossing showing all six accidents at the crossing involving an eastbound vehicle and eastbound train possibly indicating an issue due to a tight crossing angle.

Trespassing and Pedestrian Safety

Trespassing is defined as “persons who are on the part of railroad property used in railroad operation and whose presence is prohibited, forbidden or unlawful.” The motive for trespassing is usually when a person uses the railroad right of way as a shortcut, for graffiti or criminal purposes, or for suicide. Trespassing is a complicated problem to address from a community perspective because trespassing incidents only constitute around 2% of all public safety issues including homicides, illegal drugs, highway crashes, etc.

In the United States, trespassing-related railroad incidents result in higher number of fatalities than all other categories combined. In 1997, the number of trespasser fatalities exceeded the number killed in collisions at grade crossings for the first time since 1941. In 2017, there were over 500 fatalities due to trespassing incidents as compared to around 200 fatalities due to railroad grade crossing incidents. A lack of knowledge and appreciation of the dangers of trespassing accidents contributes to a very high fatality rate for trespassing incidents. Fatalities due to trespassing-related incidents may be characterized as intentional (suicide) or unintentional. A fatality due to a trespassing incident is determined to be a suicide by a coroner or a medical examiner. When the medical examiner or coroner reports that the cause of a rail fatality is undetermined, it is recorded as a trespass death.

Between 2012 and 2014, there were over 3500 trespassing incidents recorded in the United States with over 1700 of those incidents resulting in a fatality. In this time, there were an additional 700 suicides as well (as determined by a coroner). Each year, on an average there are over 450 trespassing-related fatalities, 490 trespasser injuries, and around 200 suicides on rail property.



Pedestrian Safety at Railroad Grade Crossings

The number of pedestrian accidents at public grade crossings is showing an increase from 125 accidents in 2008 to nearly 200 incidents in 2018. The number of pedestrian accidents that were fatal also saw an increase in these years. The proportion of pedestrian accidents that resulted in a fatality reached a peak in 2016.

Pedestrians can easily enter a crossing as compared to motor vehicles as they usually do not see themselves as part of the traffic. As a result, pedestrians tend to ignore warning devices. Other factors that are relevant to pedestrian violations at grade crossings, include time of day, age, gender, inattention, time pressures, the number of pedestrians present, the influence of alcohol or drugs, mental illness, and thrill-seeking behavior. Effective use of channelization devices can direct the flow of pedestrians to one point at the crossing. Such devices have shown to improve the safety of pedestrians at crossings.

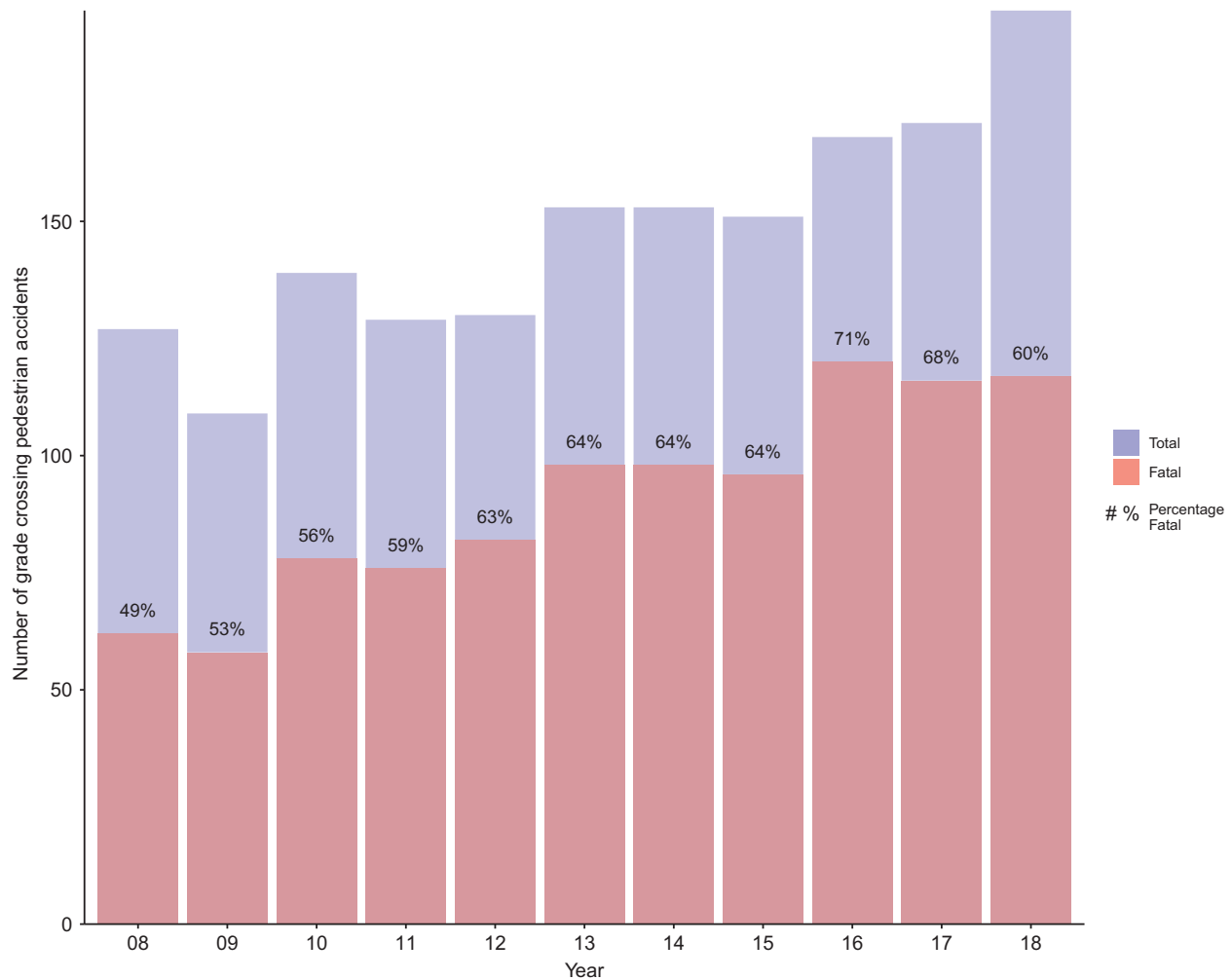


Figure 11 Pedestrian Accidents at the US grade crossings.

Most of the pedestrian incidents at railroad grade crossings involve males (around 75%), and occur during clear days (around 70%), and show a decline in winter months.

Emerging Technologies to Improve Railroad Safety

Several emerging technologies have been developed which are playing a role in improving the safety of railroad users, for example, intelligent transportation systems (ITS). In railroad applications, ITS technologies can enable improved warning of approaching trains to motorists and pedestrians. They can also be used to reroute traffic around blocked or busy grade crossings to alleviate congestion. Several projects carried out indicate that ITS technologies can have a positive impact in increasing safety and mobility at railroads.

Intelligent Transportation Systems in Railroads

ITS is the application of communication, control, and information exchange technologies, to transportation systems operation and management. Communication between vehicles and infrastructure is enabled by various wireless technologies including, dedicated short-range communication (DSRC), radio modem using ultra high frequency (UHF), wireless communication, etc. Communications between the vehicles and infrastructure can facilitate quick information exchange to warn a driver of an upcoming crossing, alert a locomotive engineer about changes in crossing signals/crossing status.

The Federal Communications Commission (FCC) has set aside 75 MHz of spectrum in the 5.9 GHz band in the United States to be used for vehicle-related safety and mobility systems ([Federal Communications Commission, 1999](#)) while the European

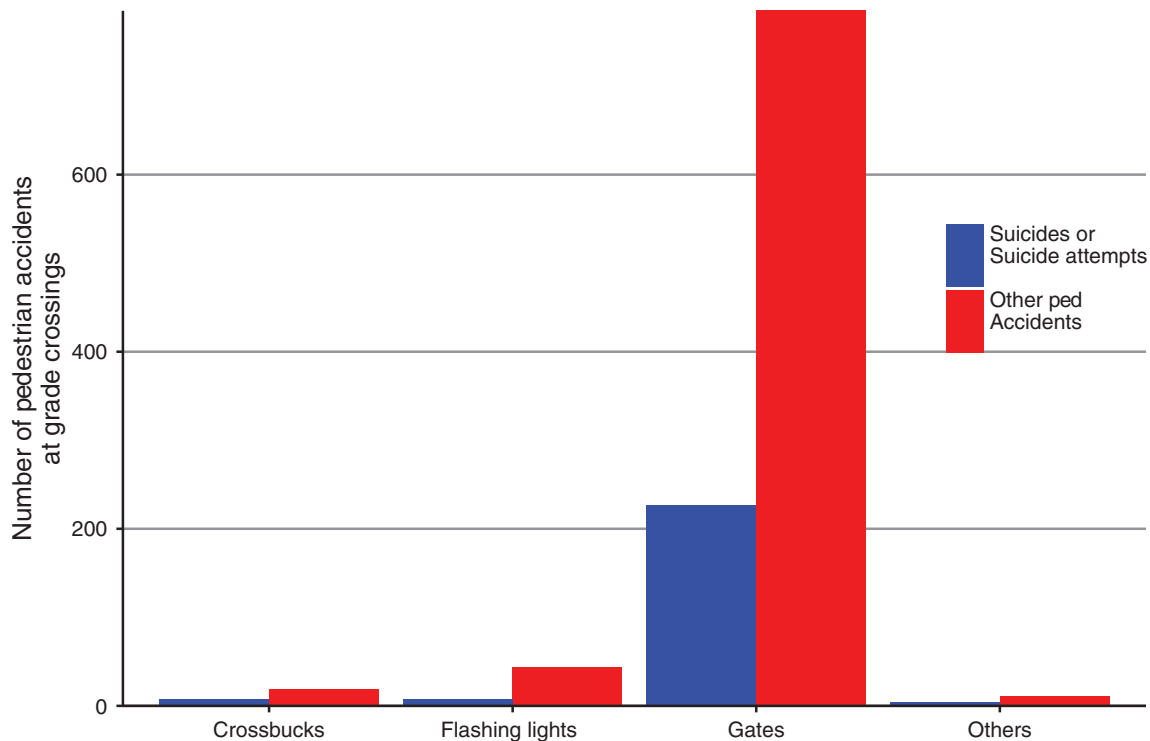


Figure 12 Number of pedestrian accidents by type of crossing warning device in the United States (2012–18).

Telecommunications Standard Institute (ETSI) has allocated 30 MHz of the 5.9 GHz. Some of the applications of ITS involve positive train control, connected vehicles, etc.

Positive Train Control

Positive train control (PTC) is a train control system which only allows train motion so as to prevent train-to-train collisions, derailments caused by excessive speed, unauthorized train movement onto sections of track where maintenance activities are taking place, movement of a train through a track switch left in the wrong position. PTC system cannot prevent grade crossing collisions (Federal Railroad Administration, 2018).

A PTC system can determine a train's location, speed, and direction. This is achieved with two basic components. They include the control unit on the locomotive and a system to communicate with the control unit about changing conditions within the track. The control unit receives information from wayside stations along the planned route of the train giving the locomotive engineer advanced warning. The system automatically applies brakes to the train, if no action is detected from the locomotive engineer.

Positive train control was recommended by the National Transportation Safety Board (NTSB) in the United States in 1990 and has been implemented more-or-less system-wide in many other countries. PTC implementation was made mandatory for most Class 1 railroads by law in United States congress via the Rail Safety Improvement Act of 2008. The system is expected to be fully interoperable by 2020.

Connected Vehicles

Communication between the crossing and a connected vehicle (a vehicle which can communicate with other vehicles or infrastructure) can alert the driver of a highway vehicle about the status of a crossing, and thereby avoid an incident.

Connected vehicles often have an on-board equipment (OBE) that provides communications, sensory, and processing functions to support the safe operation of the vehicle. The radios supporting vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) communications are a key component of the OBE to enable in-vehicle warning to the highway vehicle about the occupancy of a crossing by the train.

Detection Technologies for Pedestrian Detection

Video data has been used for surveillance of railroad property. Advances in machine learning techniques have enabled the automation of pedestrian detection from video data. Such technologies show promise in automating labor-intensive work and can support data driven research on trespassing safety.

Relevant Websites

Traffic Control Devices at Railroad Grade Crossings

https://safety.fhwa.dot.gov/hsip/xings/com_roaduser/07010/index.cfm.
https://safety.fhwa.dot.gov/hsip/xings/docs/guidance_on_traffic_control_devices.pdf.
<https://www.ite.org/pub/?id=e1dca8bc%2D2354%2Dd714%2D51cd%2Dbd0091e7d820>.

Microscopic Analysis of Railroad Grade Crossing Accidents

<https://conservancy.umn.edu/bitstream/handle/11299/201747/CTS%2019-2.pdf?sequence=1&isAllowed=y>.
<https://journals.sagepub.com/doi/abs/10.3141/2608-06>.

Trespass and Pedestrian Incidents at Railroads

<https://fas.org/sgp/crs/misc/IN10753.pdf>.
<https://fragis.fra.dot.gov/Trespassers>.
https://rosap.ntl.bts.gov/view/dot/36451/dot_36451_DS1.pdf.
<https://www.fra.dot.gov/eLib/Details/L19817>.
<https://search.informit.com.au/documentSummary;dn=729166619796355;res=IELNZC>.
<https://www.vtt.fi/inf/pdf/science/2012/S27.pdf>.

Intelligent Transportation Systems and Positive Train Control

https://rosap.ntl.bts.gov/view/dot/3820/dot_3820_DS1.pdf.
<https://www.aar.org/campaigns/ptc/>.

Video Analysis for Safety Research

<https://journals.sagepub.com/doi/pdf/10.1177/0361198118792751>.

References

- European Union Agency for Railway, 2018. Report on railroad safety and interoperability in the EU. Available from: https://www.era.europa.eu/sites/default/files/library/docs/safety_interoperability_progress_reports/railway_safety_and_interoperability_in_eu_2018_en.pdf
- Federal Communications Commission, 1999. FCC Allocates Spectrum in 5.9 GHz Range for Intelligent Transportation Systems Uses. Available from: https://transition.fcc.gov/Bureaus/Engineering_Technology/News_Releases/1999/nret9006.html
- Federal Railroad Administration, 2018. PTC System Information. Available from: <https://www.fra.dot.gov/Page/P0358>.
- FRA Office of Safety, 2019. Available from: <https://safetydata.fra.dot.gov/OfficeofSafety/publicsite/downloaddbf.aspx>.
- ITE, 2006. Preemption of Traffic Signals Near Railroad Crossings: An ITE Recommended Practice. Available from: <https://www.ite.org/pub/?id=e1dca8bc%2D2354%2Dd714%2D51cd%2Dbd0091e7d820>
- Manual of Uniform Traffic Control Devices, 2009. Flashing-Light Signals, Gates, and Traffic Control Signals. Available from: <https://mutcd.fhwa.dot.gov/htm/2009/part8/part8c.htm>
- Ogden, B.D., 2007. Highway Railway Grade Crossing Handbook. Railroad-highway grade crossing handbook. No. FHWA-SA-07-010; NTIS-PB2007106220. Federal Highway Administration, United States.
- Savage, I., 2013. Comparing the fatality risks in United States transportation across modes and over time. Res. Transp. Econ. 43 (1), 9–22.

Recreational Boating Safety: Usage, Risk Factors, and the Prevention of Injury and Death

Amy E. Peden^{*,†}, Stacey Willcox-Pidgeon^{*}, Kyra Hamilton[‡], ^{*}Royal Life Saving Society-Australia, Sydney, NSW, Australia; [†]School of Public Health and Community Medicine, University of New South Wales, Kensington, NSW, Australia; [‡]School of Applied Psychology, Griffith University, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature/Abbreviations	477
Introduction	478
Explanation of Terms Used	478
Exposure: Usage and Location	478
United States	478
Canada and Europe	478
Australia	479
Locations and Hazards	479
The Epidemiology of Death and Causes of Injury	479
United States and Canada	479
Australia	479
Sweden	479
Small Boats	480
Injuries	480
Personal Watercraft	480
Canoeing, Kayaking, and Rafting-Related Injuries	480
Propeller Strike	480
Carbon Monoxide Poisoning	480
Risk Factors	480
Injury and Fatality Risk Factors	481
Operator Inexperience, Inattention, and Poor Judgment	481
Improper Lookout	481
Excessive Speed	481
Alcohol (and Other Drugs)	481
Life Jackets	481
Weather	482
Rules Violation	482
Overloading	482
Strategies to Improve Safety	482
Safe Boating Strategies	482
Legislation and Regulatory Strategies	482
Social Psychological and Behavioral Strategies	483
Conclusion	484
Acknowledgment	484
Permissions	484
Biographies	484
See Also	485
References	485
Further Reading	486

Nomenclature/Abbreviations

BAC	Blood alcohol concentration
CO	Carbon monoxide
EPIRBs	Emergency position-indicating radio beacons
ft.	Feet
HELP	Heat escape lessening position
L	Liter
m	Meter
mg	Milligrams

NRBS	National Recreational Boating Survey
PFD	Personal flotation device
PWC	Personal watercraft
SUPs	Stand-up paddleboards
US	United States
USD	United States dollars
VHF	Very high frequency
WHO	World Health Organization

Introduction

Boating is a popular recreational activity in many high- and middle-income countries (Peden et al., 2018a). In low-income countries, boating is often used for non-recreation purposes and predominately as part of daily living, such as for transportation purposes, or used occupationally (World Health Organization, 2014). This chapter focuses on recreational boat and watercraft usage only. Despite its popularity, recreational boating is not without risks. This chapter unpacks these risks and offers solutions to improve boating safety. It first focuses on exposure through usage and location data, and then reports on the epidemiology of death and causes of injury. This chapter then moves on to discuss common risk factors associated with recreational boating including operator inexperience, improper lookout, excessive speed, overloading, alcohol and other drugs, and nonuse or unavailability of life jackets. The chapter concludes with implications for prevention to reduce boating risks including safe boating strategies, legislation and regulatory strategies, and social psychological and behavioral strategies.

Explanation of Terms Used

Boating and non-powered watercraft can be used for recreational (i.e., leisure pursuits) and non-recreational (i.e., daily living or occupational) purposes. This chapter focuses on recreational usage of boats and watercraft only. It must be noted that boating and watercraft differ. Boating is defined as the act of using water-based wind or motor-powered vessels, boats, ships, and personal watercraft (PWC) (e.g., jet skis, sailboats, yachts, and catamarans) (Australian Water Safety Council, 2016). Watercraft, on the other hand, is defined as water-based non-powered recreational equipment such as those that are rowed or paddled (e.g., rowboats, surfboards, kayaks, canoes, stand-up paddleboards [SUPs], boogie boards, windsurfers, inflatable rafts, and inflatable boats without motors) (Australian Water Safety Council, 2016). Fatal and non-fatal drownings are also discussed in this paper. Drowning is defined as “the process of experiencing respiratory impairment due to immersion/submersion in a liquid (van Beeck et al., 2005).” Outcomes of this process are defined as death (fatal drowning), morbidity or no morbidity (non-fatal drowning) (Peden et al., 2018b). Drowning can be intentional (e.g., self-harm, assault, and homicide) or unintentional (i.e., accidental) in nature. Where drowning is referred to in this paper, it refers to unintentional drowning.

Exposure: Usage and Location

Recreational boating (including the use of watercraft) is a popular activity in many high- and middle-income countries; however, global estimates of usage are difficult to derive. From reports available, estimates of the economic impact of the global recreational boating market indicate it is expected to reach an estimated \$30 billion United States Dollars (USD) by 2022 (Global News Wire, 2018).

United States

The United States (US) Coast Guard, through its National Recreational Boating Survey (NRBS), estimates there are 22.28 million recreational boats in the United States with 3.58 million boating person-hours undertaken in 2012. It is estimated that 23.9 million people in the United States undertake paddling activities per year, with canoeing (10.1 million) being the most popular activity, followed by recreational kayaking (6.2 million) (United States Coast Guard, 2012).

Canada and Europe

In Canada, it is estimated that 46% of Canadians aged 18 years and over went boating in 2014 (approximately 13.2 million Canadians), with ownership estimated at 8.6 million boats. European estimates indicate there are 36 million boaters across Europe, with 6 million boats in European waters and a further 48 million people enjoying watersports (Boating in Canada, 2019).

Australia

National estimates from Australia report 5 million recreational boaters in Australia. A total of 2 million Australians are estimated to hold a boat license. There are more than 900,000 registered boats in Australia, with as many as 900,000 watercrafts such as paddlecraft, SUPs, and sailing dinghies that do not require registration. The boating fleet is growing by 15,000 new registrations nationally per year. In Australia, PWCs are the fastest growing sector of powered vessels with 68,000 registered in Australia (as of 2017) ([Boating Industry Association, 2017](#)).

Locations and Hazards

Boating and watercraft activities can be undertaken in a range of environments, each with their own unique hazards and risks. Rivers and inland waterways such as lakes and dams are commonly known for their cold, murky water, which often conceals strong currents, submerged objects and shifting slippery banks and beds. By contrast, coastal locations can be prone to strong currents and large and unpredictable waves that are often weather dependent.

In the United States, the NRBS estimates that inland lakes are the leading location for boating, with 57% of the estimated 244 million boating days taking place on inland lakes, followed by rivers (21% of all boating days). Powerboats are most commonly used at freshwater locations, whereas sailboats are more commonly used in saltwater locations. Paddlecraft are used primarily on lakes and rivers ([United States Coast Guard, 2012](#)).

In Canada, boating and watercraft usage most commonly occurs in inland waterbodies such as lakes, ponds, rivers, creeks, streams, or waterfalls. In Europe, inland waterways, as well as offshore boating along the coast, are popular ([Boating in Canada, 2019](#)). In Australia, as in Europe, both inland and coastal waters prove popular for boating, with watercraft users, such as kayakers and canoeists more commonly preferring inland locations such as rivers and lakes ([Pidgeon and Mahony, 2016](#)).

The Epidemiology of Death and Causes of Injury

Drowning is a leading cause of death during boating- and watercraft-related activities, with other causes often injury-related ([World Health Organization, 2014](#)).

United States and Canada

In the United States and Canada, the number of transportation fatalities from recreational boating rank only second to motor vehicles compared with other modes of transportation. In 2017, the US Coast Guard counted 4291 accidents that involved 658 deaths, 2629 injuries, and approximately \$46 million USD in damage to property as a result of recreational boating accidents. When exposure is considered, the NRBS fatality rate for the United States (as calculated as a ratio of the number of deaths per 100 million exposure hours) was highest for rivers with approximately 27 deaths. By contrast, bays had the lowest fatality rate at 12 deaths per 100 million exposure hours ([United States Department of Homeland Security, 2018](#)). To put this risk in perspective, the United States saw 24,973 accidental deaths among motor-vehicle occupants in 2017 if motorcyclists are excluded. And, Americans are estimated to have traveled around 119 billion hours by cars and trucks in 2017. This gives a fatality risk of 21 per 100 million exposure hours for vehicle occupants. In other words, on average, boating seems to be a slightly more dangerous activity than traveling in a motor vehicle. In Canada, boating accounted for 34% ($n = 2323$) of the 6811 water-related fatalities in summer from 1991 to 2013 ([Red Cross Canada, 2016](#)).

Australia

In Australia, 20% of all unintentional fatal drownings are due to boating and watercraft incidents; it is the second leading cause of drowning in Australia after swimming. Across a 10-year period (2005/06–2014/15), a total of 473 people drowned while participating in boating- or watercraft-related activities. Of these, 92% were male, and incidents most commonly occurred while fishing (28%), while the vessel was underway (27%), or due to a fall overboard (12%). Ocean/harbors were the leading location for boating- and watercraft-related drowning fatalities (51.9%) ([Willcox-Pidgeon et al., 2019](#)).

Sweden

As an example of drowning deaths in Europe, we can look at Sweden, a country with 10 million inhabitants and 820,000 recreational boats. In 2018, Sweden saw 135 drowning fatalities, or 1.4 drowning deaths per million inhabitants. This number is similar to what has been experienced since the early 1990s. Before that, Sweden saw many more fatal drownings with a rate of over 4 per million inhabitants in the 1960s and around 3 per million inhabitants in the 1970s. Of the people drowning in 2018, 87% were male, 16% were above age 70, and 10% were children below age 19. Just over one-third of the drownings happened in connection to swimming, 11% by people falling through ice, and 9% (12 people) in connection to recreational boating. However, 2018 is an outlier with respect to recreational boating fatalities. The average for the last 10 years is 25 per year or about 20% of all

drownings. Less use of alcohol on boats and higher use of life jackets explain the reductions in drownings since the 1960s, but the drowning rate is still considered unacceptable by the Swedish authorities and a goal of cutting the number of drownings by 50% by 2030 is a step toward a Vision Zero for boating deaths. The previous goal was set to reduce recreational boating deaths from 40 to 25 deaths per year by 2020, and that goal seems to be reached. During the 1970s, the average was around 90 deaths per years in recreational boating accidents.

Small Boats

The World Health Organization (WHO) reports data from countries such as Australia, Canada, Germany, Finland, and the United States, and suggests the number of deaths related to incidents involving small boats (i.e., boats ranging from 5 to 8 m [16–26 ft.] in length) form a significant proportion of all drowning fatalities ([World Health Organization, 2014](#)). In Canada, powered boats 5.5 m (18 ft.) and under accounted for 47% of all summer boating-related fatalities ([Red Cross Canada, 2016](#)). In Australia, powered boats accounted for 71% of all boating- and watercraft-related drowning deaths, with the highest occurring in boats 5 m (16 ft.) and under in length (30%) ([Willcox-Pidgeon et al., 2019](#)).

Injuries

There are a range of other causes that can lead to injuries and fatalities when participating in boating- or watercraft-related activities. These include traumatic injuries and blood loss from propeller strikes ([Mann, 1980](#)), spinal injuries, and concussion among others. Carbon monoxide (CO) poisoning has also been responsible for deaths while boating ([Yoon et al., 1998](#)).

Personal Watercraft

A study of PWC-related injuries and fatalities in Arkansas identified 126 incidents involving 141 vessels and over \$156,000 USD in property damage for the years 1994 through 1997 ([Jones, 2000](#)). Almost two-fifths of PWC users were injured, mainly head trauma and fractures to the lower limbs. There were five reported fatalities. The most common injuries by type were lacerations (14%), fractures (13%), and contusions (7%). When examined by body region, injuries commonly occurred to the lower extremity (21%), followed by head and neck (14%) and the upper extremity (7%).

Canoeing, Kayaking, and Rafting-Related Injuries

An examination of injuries related to canoeing, kayaking, and rafting found aside from death, other hazards encountered include hitting objects, waterborne diseases, hypothermia from unintended submersion, blisters, muscle strain, cuts, and abrasions ([Franklin and Leggat, 2012](#)). The body areas more commonly injured when undertaking paddling-related activities were the face (33%) (including the eye, mouth, nose, and teeth), knee (15%), arm/wrist/hand (12%), and other parts of the leg/hip/foot (11%). Common contributing factors for canoe deaths and injuries recorded by the US Coast Guard included alcohol use (18%), hazardous waters (18%), improper loading (13%), operator inexperience (13%), and weather (8%). For kayak-related deaths, the most common contributing factor identified by the US Coast Guard was hazardous waters (45%) ([United States Department of Homeland Security, 2018](#)).

Propeller Strike

Propeller strike is another cause of serious injury and, in severe cases, fatality. Traumatic injury including cuts and blood loss can be caused by contact with a boat propeller and, in severe cases, amputation of limbs can occur ([Mann, 1980](#)). Most amputations occur to the lower extremities. Wounds resulting from propeller strikes are also prone to infections due to contaminated water with seawater and inland water often having different bacteria present.

Carbon Monoxide Poisoning

CO poisoning is a serious hazard associated with recreational boating. A study of 512 patients treated for acute unintentional CO poisoning identified that individuals commonly lose consciousness as a result of the poisoning. Most cases occurred aboard a boat that was older than 10 years, had an enclosable cabin, was longer than 6.7 m (22 ft.), was powered by a gasoline engine, and was without a CO detector on board ([Yoon, 1998](#)).

Risk Factors

Common risk factors contributing to boating deaths, injuries, and accidents worldwide are now discussed. Other risk factors not mentioned include preexisting medical conditions, poorly maintained vessels, older vessels, and small vessels (see small boats).

Injury and Fatality Risk Factors

A US study of boating-related injuries and fatalities in Washington State found factors, such as not wearing a life jacket, not having safety features present, being on an unpowered boat, having alcohol present (Stempski et al., 2014), boating in the ocean, boating between the hours of midnight and 5:59 a.m., incidents involving capsizing, sinking, flooding, or swamping, and incidents that involve a person leaving the boat voluntarily, ejected or due to a fall increased the chances of being fatally injured.

Operator Inexperience, Inattention, and Poor Judgment

Boating accidents are largely attributed to lack of experience, not paying attention, and poor judgment largely due to human error (United States Department of Homeland Security, 2018). Technical knowledge and experience of how to operate a boat, whether a small 5 m (16 ft.) dingy, a 3 m (10 ft.) sailing vessel, or a canoe, is essential to avoid accidents, injuries, and death on the water. This includes failure to perceive or observe danger outside or inside of the boat. Examples of incidents resulting from operator inexperience and/or inattention include hitting submerged objects; beaching on shallow water, sandbanks, or rocks; other boats or people in the water; and persons falling overboard. Of the known causes, human error is responsible for 33% of fatalities in New South Wales, Australia (Transport for New South Wales, 2018).

Improper Lookout

The operator/skipper is responsible for keeping a lookout for dangers at all times, by sight and hearing, especially in poor weather conditions, when dark or in limited visibility, and in waterways with high traffic. Most jurisdictions require having an appointed “lookout” or “observer,” when towing, for example, water skis, tubes, etc. This individual is responsible for looking out for persons overboard, other boats, or hazards, and for giving the skipper instructions in regards to speed and directions.

Excessive Speed

Excessive speed is one of the top five causes of boating-related deaths and injuries in the United States, Canada, and Australia (Red Cross Canada, 2016; Transport for New South Wales, 2018; United States Department of Homeland Security, 2018). Speed restrictions exist for vessels, similar to that of road users. For example, the five-knot rule applies when within 200 m (656 ft.) to shore, near other boats, dive buoys, or people. Similar to excessive speeding in a motor vehicle, there may be a high chance of a high-impact collision, and the vessel may capsize. Additional speed restrictions may apply when towing (e.g., water skiing, using PWC).

Alcohol (and Other Drugs)

The role of alcohol and driving a motor vehicle is well reported; however, there is less information available on alcohol and boating. Boating-related fatality statistics attribute alcohol as a leading cause of boating accidents, injuries, and deaths. Alcohol alters perception, judgment, and balance. The link between alcohol and drowning is evident, especially among men (Hamilton et al., 2018). Some medications may cause drowsiness; thus, checking medication packages for cautions and discussing with a medical professional before operating a vessel and taking medication (including mixing with alcohol) is recommended.

In Sweden, alcohol was involved in 75% of all recreational boating fatalities in 2012 where blood alcohol concentration (BAC) was measured. However, autopsies were performed in only about half of all cases, so the actual rate may be lower. By 2015, that percentage had dropped to 60%. Many of the alcohol-related fatalities occur in harbors when people walk between boats and not during the operation of boats. Since 2010, Sweden has a 0.02% drunk driving law for operating boats, but it applies only to the operator (and other people onboard responsible for passengers’ safety) and only to boats that can achieve 15 knots.

Life Jackets

Not wearing life jackets (also known as personal flotation devices [PFDs]) has been well established as a contributing factor for drowning during recreational boating, both powered and non-powered. The risk of drowning and hypothermia increases when not wearing a life jacket. Life jackets have been found to increase the chance of survival in the water by 50% (Cummins et al., 2011); however, many studies report a low level of wearing life jackets, with the exception of children. In many jurisdictions, it is mandatory to carry life jackets for everyone on board the boat, but it is not necessarily mandatory to wear them. Barriers to life jacket wear have been explored including cost, comfort, risk perception, and role modeling with children (Peden et al., 2018a). Studies have shown that children are more likely to wear life jackets than adults. In addressing comfort, inflatable slimline life jackets have been available for a number of years as an alternative to the foam block style life jackets. While these life jackets have decreased in size and cost, and increased in comfort and ability to move more freely, issues regarding the CO₂ canister have arisen, where the life jacket has not deployed due to the canister becoming unattached, rusted, or leaking. Manufacturers recommend servicing these life jackets every 12 months.

Weather

Weather influences boating activity, with favorable weather making it more attractive to be on the water for a range of activities, increasing exposure, and therefore risk (Fralick et al., 2013). Marine weather can be very different to weather on land, with conditions out at sea or over open water changing quickly and impacting water and wave conditions. Weather reports, including wind direction, wave height, and tide times, are usually broadcast over marine weather channels on very high frequency (VHF) radios, on websites, and, more recently, on boating apps. Water temperature around the world varies, and hypothermia is life threatening in the case of falling overboard or capsizing. Life jackets can help retain heat when in the water.

Rules Violation

International maritime rules are implemented worldwide (World Health Organization, 2014), and are designed to provide a safe environment for all water users and to decrease the risk of accidents, injuries, or death while on the water (e.g., life jacket carry and wear, restricted speed, and alcohol limits for operating a vessel). Local jurisdictions often add to these rules to reflect the local environmental conditions. Violating rules such as excessive speed or consuming alcohol while operating a vessel places not only the operator/skipper in danger, but also their passengers and other waterway users. The consequences of violating rules vary from fines, license suspension, and having the vessel impounded, to criminal charges being laid, especially if people are injured or killed (Willcox-Pidgeon et al., 2019).

Overloading

Overloading of boats, especially small boats, with people, cargo, or seafood, is a common occurrence worldwide. Every boat has a limit of how much weight it can safely carry without compromising stability. Overloading can unbalance the boat and cause tipping or capsizing of people and goods, especially in unfavorable weather and water conditions, and especially in small vessels that are already low to the water (World Health Organization, 2014).

Strategies to Improve Safety

This section discusses strategies to improve recreational boating safety, including safety equipment, legislation and enforcement, and behavioral strategies.

Safe Boating Strategies

Safety equipment is essential when boating regardless of experience, type of vessel, or duration on the water. Universally, the minimum recommended safety equipment includes items such as life jackets, anchor and line, distress flares, EPIRB, and marine radio (VHF) (Boating Industry Association, 2019; Stempski et al., 2014) (Table 1). Although these items are universal, recommended equipment may vary depending on type of vessel and environment.

Before going out on the water and during boating activity, it is important to check the weather and water conditions and to familiarize oneself with the general location. Logging trips on and off with local marine organizations (e.g., coast guard, marine rescue) will ensure that the vessel's general location and expected time of arrival is known. This is especially important if something unexpected were to happen.

In the event of the vessel capsizing or an unexpected fall into the water, life jackets can help retain heat when in the water. The following techniques are designed to increase survival when in the water:

- HELP: Heat escape lessening position—arms across chest and bring knees up to chest (Fig. 1).
- Huddle: Huddling close together with other people by linking arms, retaining body heat (Royal Life Saving Society - Australia, 2016).

Legislation and Regulatory Strategies

The WHO Global Report on Drowning proposes the “setting and enforcement of safe boating, shipping, and ferry regulations” as one of 10 key strategies for reducing the global drowning burden (World Health Organization, 2014).

Table 1 Recommended minimum safety equipment

<i>Recommended minimum safety items for recreational boating</i>	
Life jackets	Fire extinguisher
Anchor and line	EPIRB
Bailer or bucket	Marine radio (VHF)
Bilge pump	Waterproof torch
Distress flares	Compass
Oars/paddles	



Figure 1 The HELP position. Source: *Royal Life Saving Society - Australia, 2016*

Legislation in many countries requires carrying the appropriate number of life jackets for everyone on board. Furthermore, mandatory wearing of life jackets in boats 6 m (20 ft.) and smaller to prevent injuries and drowning in these vessels was introduced to address the high number of incidents in these small vessels. The success of such legislation is evident, in Tasmania, Australia. Tasmania was the first jurisdiction in the world to introduce this legislation in 2001 and currently reports a 90% life jacket wear rate (Willcox-Pidgeon et al., 2019). In Victoria, Australia, and Washington State, the United States, decreases in boat-related drowning deaths have been reported since introducing mandatory wearing of life jackets in small boats (Bugega et al., 2014; Chun et al., 2014). Enforcement of life jacket legislation can be done through random checks on board vessels and at boat ramps, with punishment ranging from warnings to fines and loss of license (where such licensing schemes exist) (Stempski et al., 2014).

Many developed countries require an operator or skipper to have a license to operate a recreational vessel. License testing generally includes basic boating theory and knowledge including maritime rules, speed, give-way rules for different types of vessels (powered and unpowered), safety requirements, communication, and marine weather. Examples from Australia indicate that operator licensing has been found to increase awareness and knowledge of the boating public and decrease incidents (Virk and Pikora, 2010).

Many high-income countries (such as the United States, Australia, and Canada) impose alcohol limits for people operating a recreational vessel, with some jurisdictions implementing alcohol breath testing and drug testing on the water. The limits in Australia, United States, Canada, and Europe are BAC of ≤ 0.05 mg/L. Consequences for operating a vessel under the influence may include fines, loss of license, and the vessel being impounded (Willcox-Pidgeon et al., 2019).

Social Psychological and Behavioral Strategies

It is often assumed that people engage in unsafe recreational aquatic activities, including unsafe boating behaviors such as not wearing a life jacket and consuming alcohol while boating (Hamilton et al., 2018; Peden, et al., 2018a), because of a lack of knowledge of the risks. This belief has given rise to numerous public health campaigns that have focused their attention on raising knowledge and awareness of the dangers involved in risky boating-related recreational activities. However, given the increased attention that the issue of boating safety has received in both media coverage and public health messages, particularly in developed countries, the dangers are known to many. Despite this, people continue to ignore safety warnings and carry out unsafe behaviors while boating. This highlights that having the correct information does not always translate into behavior change, suggesting that risky acts while boating are based on more than knowledge acquisition alone.

Recent research has provided emerging evidence for the social psychological and behavioral factors that may influence individuals' decisions in, on, and around water (Hamilton et al., 2016; Hamilton et al., 2014; White et al., 2018). Given psychological factors are likely to be critical in individuals' decisions to engage in risky aquatic activities, including recreational boating, it is important that interventions grounded in sound psychological and behavioral theory be adopted to modify people's risky boating behaviors. This is especially important given research has shown theory-based campaigns may be more effective in promoting health- and risk-protective behaviors compared to atheoretical ones, and evaluation of advertising countermeasures is easier and more cost-effective with theoretically devised approaches given the clearly measurable constructs (Prestwich et al., 2015; Noar, 2006).

In examining the current literature, limited research has been conducted in this area and, in particular, related to recreational boating safety. Drawing on the literature more broadly, several social psychological and behavioral factors have been identified that may influence individuals' decisions to engage in risky aquatic activities. These include past experience, attitudes (i.e., costs and benefits), social pressure and norms, self-efficacy beliefs (i.e., motivating or inhibiting factors), planning, anticipated regret (i.e., anticipated emotions experienced), and risk perceptions (i.e., weighing up the risks involved and possibility of adverse consequences) (Hamilton et al., 2016; Hamilton et al., 2014; White et al., 2018).

Given the little evidence that promoting awareness and knowledge alone results in sustained changes in behavior, it seems important to draw on these more social psychological and behavioral factors when designing campaigns aimed at improving boating safety to ensure the messages embedded within the campaigns will be effective in changing people's behavior. For example,

promotion efforts to improve safe recreational boating behavior may consider using behavior change methods such as persuasive messaging to change attitudes, building self-efficacy to overcome barriers, harnessing normative influences, and encouraging people to plan their behavior prior to going boating. It is also important to consider the framing and content of safety messages when attempting to change behavior through emphasizing the negative consequences of behaviors. For example, campaigns using *fear appeals* that seek to elicit an emotional response of fear to encourage behavior change have been shown to be mostly ineffective, and in some cases have adverse effects (Peters et al., 2013). The exception is that when fear appeals are accompanied by tangible resources to increase self-efficacy toward the behavior seeking to be changed, small effects can be attained. In contrast, gain-framed messages emphasizing the positive consequences of engaging in the desired behavior tend to be more effective because they are more readily accepted and prevent defensive and avoidant reactions.

Conclusion

Boating and watercraft use are popular activities. Tens of millions of people participate recreationally in boating and watercraft activity; however, participation is not without risk. Injuries or death, most commonly as a result of drowning, can occur and it is important that participants are mindful of the risks. Such risk factors that may impact safety include operator inexperience, improper lookout, alcohol and other drugs, excessive speed, lack of life jackets, and overloading among others.

There are a number of strategies to reduce risk and increase the safety of recreational boaters and watercraft users. These can be loosely grouped into safe boating strategies with an emphasis on safe behavior, vessel seaworthiness and safety equipment, legislation and regulatory strategies, and social psychological and behavioral strategies.

Acknowledgment

This research is supported by Royal Life Saving Society-Australia, to aid in the prevention of drowning. Research at Royal Life Saving Society-Australia is supported by the Australian Government.

Permissions

Permission to use Fig. 1 has been granted by Royal Life Saving Society-Australia as copyright owners.

Biographies



Amy Peden has extensive experience in drowning prevention research, policy, and practice. She has authored over 40 peer-reviewed articles to date, as well as over 60 professional reports. She has been a key contributor to the last three Australian Water Safety Strategies. She is currently a Senior Research Fellow with Royal Life Saving Society - Australia (RLSSA). In her previous position as National Manager of research and policy with RLSSA, her duties include the production of the National Drowning Report, maintaining the National Fatal Drowning Database, analyzing policy, and evaluating programs. She is a PhD candidate at James Cook University researching the epidemiology, risk factors, and prevention strategies for unintentional river drowning.



Stacey Willcox-Pidgeon is the Senior Research and Policy Officer with Royal Life Saving Society of Australia. She has conducted a range of drowning-related research, analysis, and program evaluation in Australia and New Zealand. She has authored drowning research publications, professional reports, and presented at a range of national and international conferences. She completed a Master of Public Health Research Thesis in 2015 investigating youth risk perception of drowning in a beach environment. She has recently commenced a PhD focusing on drowning prevention for high-risk populations in Australia.



Dr. Kyra Hamilton is an Associate Professor in health psychology and behavioral medicine in the School of Applied Psychology at Griffith University, Australia. She has both psychology and nursing qualifications and over 25 years' experience in the health field. She has particular research interests in health behavior motivation, self-regulation, and change. She has won national and international awards for her research and is Chief Investigator on national competitive, industry, and internal grant funded projects. She is the founder and director of the Health and Psychology Innovations (HaPI) research laboratory.

See Also

Alcohol; Drugs, illicit, and prescription; EPIDEMIOLOGY OF ROAD TRAFFIC CRASHES; Exposure: A Critical Factor in Risk Analysis; Ferries and short-sea shipping; HUMAN FACTORS IN TRANSPORTATION; Value of life and injuries; Weather, Effects of

References

- Australian Water Safety Council, 2016. Australian Water Safety Strategy 2016-2020, Australian Water Safety Council, Sydney, Australia. Available from: <http://www.watersafety.com.au/AustralianWaterSafetyStrategy/2016-2020Strategy.aspx> (Date accessed: 25-07-2019)
- Boating Industry Association, 2017. 2017 Annual Report, Boating Industry Association. Available from: <https://www.bia.org.au/about-us/annual-reports> (Date accessed: 25-07-2019)
- Boating Industry Association, 2019. Safety messages. Available from: <https://www.bia.org.au/community/safety-messages> (Date accessed: 25-07-2019)
- Boating in Canada 2019. Canadian Boating Statistics. Available from: <http://boating.nrc.ca/stats.html> (Date accessed: 25-07-2019).
- Bugeja, L., Cassell, E., Brodie, L., Walter, S., 2014. Effectiveness of the 2005 compulsory personal flotation device (PFD) wearing regulations in reducing drowning deaths among recreational boaters in Victoria, Australia. *Inj Prev.* 20, 387–392. <https://doi.org/10.1136/injuryprev-2014-041169>.
- Chung, C., Quan, L., Bennet, E., Kernic, M.A., Ebel, B.E., 2014. Informing Policy on Open Water Drowning Prevention: Observational Survey of Life Jacket Use in Washington State. *Inj Prev.* 20, 238–243. [doi:10.1136/injuryprev-2013-041005](https://doi.org/10.1136/injuryprev-2013-041005).
- Cummings, P., Mueller, B.A., Quan, L., 2011. Association between wearing a personal floatation device and death by drowning among recreational boaters: a matched cohort analysis of United States Coast Guard data. *Injury Prev.* 17 (3), 156–159.
- Fralick, M., Denny, C.J., Redelmeier, D.A., 2013. Drowning and the influence of hot weather. *PlosOne* 8 (8), e71689.
- Franklin, R.C., Leggat, P.A., 2012. The epidemiology of injury in canoeing, kayaking and rafting. *Med. Sport Sci.* 58, 98–111.
- Globe News Wire, 2018. Global Recreational Boating Industry Report 2017-2022. Available from: <https://www.globenewswire.com/news-release/2018/02/12/1338854/0/en/Global-Recreational-Boating-Industry-Report-2017-2022.html> (Date accessed: 25-07-2019).
- Hamilton, K., Keech, J.J., Peden, A.E., Hagger, M.S., 2018. Alcohol use, aquatic injury, and unintentional drowning: a systematic literature review. *Drug Alcohol Rev.* 37, 752–773. [doi:10.1111/dar.12817](https://doi.org/10.1111/dar.12817).
- Hamilton, K., Peden, A.E., Pearson, M., Hagger, M.S., 2016. Stop there's water on the road! Identifying key beliefs guiding people's willingness to drive through flooded waterways. *Saf. Sci.* 86, 308–314. [doi:10.1016/j.ssci.2016.07.004](https://doi.org/10.1016/j.ssci.2016.07.004).
- Hamilton, K., Schmidt, H., 2014. Drinking and swimming: investigating young Australian males' intentions to engage in recreational swimming while under the influence of alcohol. *J. Community Health* 39, 139–147. [doi:10.1007/s10900-013-9751-4](https://doi.org/10.1007/s10900-013-9751-4).
- Jones, C.S., 2000. Epidemiology of personal watercraft-related injury on Arkansas waterways, 1994-1997: identifying priorities for prevention. *Accid. Anal. Prev.* 32, 373–376.

- Mann, R.J., 1980. Propeller injuries incurred in boating accidents. *Am. J. Sports Med.* 8 (4), 280–284.
- Noar, S.M., 2006. A 10-Year Retrospective of research in health mass media campaigns: where do we go from here? *J. Health Comm.* 11 (1), 21–42, doi: 10.1080/10810730500461059.
- Peden, A.E., Demant, D., Hagger, M.S., Hamilton, K., 2018a. Personal, social, and environmental factors associated with lifejacket wear in adults and children: a systematic literature review. *PLoS One* 13 (5), e0196421, doi:10.1371/journal.pone.0196421.
- Peden, A.E., Mahony, A.J., Barnsley, P.D., Scarr, J., 2018b. Understanding the full burden of drowning: a retrospective, cross-sectional analysis of fatal and non-fatal drowning in Australia. *BMJ Open* 8.
- Peters, G.J., Ruiter, R.A.C., Kok, G., 2013. Threatening communication: a critical re-analysis and a revised meta-analytic test of fear appeal theory. *Health Psychol. Rev.* 7 (Suppl. 1), S8–S31, doi:10.1080/17437199.2012.703527.
- Pidgeon, S., Mahony, A., 2016. Boating and watercraft drowning deaths: A 10 year analysis. Royal Life Saving Society – Australia, Sydney, Australia.
- Prestwich, A., Webb, T.L., Conner, M., 2015. Using theory to develop and test interventions to promote changes in health behaviour: Evidence, issues, and recommendations. *Curr. Opin. Psychol.* 5, 1–5, doi: 10.1016/j.copsyc.2015.02.011.
- Red Cross Canada, 2016. Water-related fatality trends across Canada 1991 to 2013. Red Cross Canada, Ontario. Available from: <https://www.redcross.ca/training-and-certification/swimming-and-water-safety-tips-and-resources/drowning-research>.
- Royal Life Saving Society – Australia, 2016. Swimming and Life Saving Manual 6th ed. Royal Life Saving Society – Australia, Sydney.
- Stempski, S., Schiff, M., Bennett, E., et al., 2014. A case-control study of boat-related injuries and fatalities in Washington State. *Inj. Prev.* 20, 232–237.
- Transport for New South Wales, 2018. Maritime safety plan 2017–2021. Transport for New South Wales, Centre for Maritime Safety, Sydney. Available from: <https://maritimemanagement.transport.nsw.gov.au/> (Date accessed: 25-07-2019).
- United States Coast Guard, 2012. National Recreational Boating Survey 2012, United States Coast Guard. Available from: <https://www.uscgboating.org/statistics/national-recreational-boating-safety-survey.php> (Date accessed: 25-07-2019).
- United States Department of Homeland Security, 2018. Recreational boating statistics 2017. United States Department of Homeland Security, United States Coastguard, Washington, DC. Available from: http://uscgboating.org/statistics/accident_statistics.php. (Date accessed: 25-07-2019).
- van Beeck, E., Branche, C.M., Szpilman, D., Modell, J.H., Bierens, J., 2005. A new definition of drowning: towards documentation and prevention of a global public health problem. *Bulletin of the World Health Organisation* 83, 853–856.
- Virk, A., Pikora, T., 2010. The recreational skippers ticket and it's influence on boater behaviour. *Int. J. Aqua. Res. Edu.* 4 (2), 175–185.
- White, K.M., Zhao, X., Wihardjo, K., Hyde, M., Hamilton, K., 2018. Surviving the swim: Psychosocial influences on pool owners' safety compliance and child supervision behaviours. *Saf. Sci.* 106, 176–183, doi:10.1016/j.ssci.2018.03.020.
- Willcox-Pidgeon, S., Peden, A.E., Franklin, R.C., Scarr, J., 2019. Boating-related drowning in Australia: Epidemiology, risk factors and the regulatory environment. *J. Saf. Res.* 70, 117–125.
- World Health Organization, 2014. Global Report on Drowning. World Health Organization, Geneva, Switzerland. Available from: https://www.who.int/violence_injury_prevention/global_report_drowning/en/ (Date accessed: 25-07-2019).
- Yoon, S., Macdonald, S.C., Parrish, G., 1998. Deaths from unintentional carbon monoxide poisoning and potential for prevention with carbon monoxide detectors. *JAMA* 279 (9), 685–687.

Further Reading

- Phillips, M.T., Spitzer, N., Chow, W., Mangione, T.W., 2019. Risk factors associated with lifejacket wear among adult canoeists and kayakers in the United States, 1999–2017. *Int. J. Inj. Contr. Saf. Promot.* 26, 176–184.

Refuge Islands

Christer Hydén, Nye Sandviksvegen, Bergen, Norway

© 2021 Elsevier Ltd. All rights reserved.

Introduction	487
Pedestrian Crossings	487
Traffic Medians	488
Median Strip at Intersections	489
Median Strip Instead of Two-Way Left Turn Lanes	489
The Width of the Median Strip	489
Type of Median Strip	490
Effect on Mobility	490
The Ultimate Measure: Medians With Guardrails	490
Less Expensive but Still Effective: Centerline Rumble Strips	491
Median Barrier on 13-m Wide Roads	491
Biography	492
References	493
Further Reading	493

Introduction

Refuge comes from *refugium* in Latin, shelter or retreat place. It is often also referred to as safety island, safety area, traffic island, or refuge island.

A refuge is a solid or painted object in a road. If the island uses road markings only, without raised curbs or other physical obstructions, it is usually called a *painted island*.

Pedestrian Crossings

Well into the 1800s, pedestrians ruled our streets even if in some competition with horses and toward the end of the century bicycles. Then when steam-powered vehicles started to arrive, regulations ensured low vehicular speeds. For example, the “Red Flag Act” was passed in the United Kingdom in 1865 and not repealed until 1896. It restricted the speed of horse-less vehicles to 4 mph (6 km/h) in open countries and 2 mph (3 km/h) in towns. The act required one person to walk ahead carrying a red flag. In 1896, the speed limit was raised to 14 mph (23 km/h) and the need for the escort was removed and then in 1903 to 20 mph (32 km/h). That speed limit was kept until 1931. Similar Red Flag Acts were enacted in many US states and were kept into the 20th century, though the urban speed limit varied between different jurisdictions and was typically 5, 10, or 12 mph (8, 16, or 19 km/h). At those speeds, it was typically easy for pedestrians to safely cross streets. That all changed with mass motorization. Higher speed limits were enacted and most countries gave automobiles the right of way and pedestrians had to look for gaps between cars, especially if crossing away from marked crosswalks. And, even in crosswalks, the practice in most countries became that pedestrians had to look for gaps between cars. Putting a refuge island in the middle of the street facilitated finding such gaps. In the last couple of decades, more and more jurisdictions have given absolute priority to pedestrians in crosswalks, and, at least in theory, pedestrians should be able to walk from curb to curb blindfolded, whereas motorists should be responsible for finding gaps between pedestrians. When we reach such a situation, there will be little need for refuge islands at pedestrian crossing points, but we are certainly not yet there.

Crossing the street can be a complex task for pedestrians, especially for those who are less capable of handling traffic. Pedestrians must estimate vehicle speeds, adjust their own walking speeds, determine adequacy of gaps, predict vehicle paths, and time their crossings appropriately. Drivers must see pedestrians, estimate vehicle and pedestrian speeds, determine the need for action, and react. At night, darkness and headlamp glare make the crossing task even more complex for both pedestrians and drivers. A refuge allows pedestrians to cross one direction of vehicle traffic at a time. The refuge provides some protection from traffic in the center of the road, while the pedestrian waits for a safe gap in the second direction of traffic.

Refuge islands significantly improve amenity for pedestrians trying to cross busy streets, as they are much more likely to find two long-enough gaps in traffic rather than one situation in which gaps for both directions coincide. Since this reduces pedestrians’ average waiting time, it also improves safety, with impatient pedestrians less likely to use gaps that turn out to be too short for safe crossing. Traffic islands can also contribute to a channelization of pedestrians. As Huang and Cynecki (2001) found, more pedestrians use a channelized pedestrian crossing instead of using other crossing locations. That, however, presupposes that

pedestrians actually feel that they are helped by the crossing, for example, that motorized traffic is stopping/yielding more often at that crossing compared with another crossing site.

Generally, research has indicated that safety for pedestrians is improved when refuge islands are present. Even though there is limited evidence, it seems as if refuges, especially on streets having more than two lanes, are beneficial for pedestrians from a safety point of view. [Zegeer et al. \(2005\)](#) found that when a refuge island is installed at an existing crosswalk on a road with more than two lanes the number of pedestrian crashes drops significantly and by an estimated average of 44%. The result is controlled for both vehicular and pedestrian volumes. Another study claims that “on a road where pedestrians often cross without a crossing facility, a refuge will decrease pedestrian accidents by around 40%.”

Crossing, at locations with refuge islands, is the safest on roads with low to medium flows of vehicle traffic, and where speeds are below 50 km/h (30 mph) but they certainly help pedestrians find safer gaps at high-speed, high-volume streets as well.

One reason why islands contribute to higher safety may be speed reductions. Pedestrian refuges slow traffic because they narrow the road, and may remind drivers that pedestrians could be crossing the road. One study showed that vehicle speed at pedestrian refuges was reduced by 6%. However, this speed reduction varies with the width of the opening between the sidewalk curb and the refuge curb. If the width is 4 m or more, there is probably very little effect on speed. If the opening is 2.1 m, as used at at least a few locations in London, the resulting speed is barely crawl speed. If going to such extremes, the opening on the other side of the refuge island will need to be kept wider to allow trucks and emergency vehicles to get by the island using the wrong side of the street traveling in the opposing traffic direction.

Another reason for safety improvements may be that there might be a change in pedestrian behavior. [Lee and Abdel-Aty \(2005\)](#) showed that at crosswalks (with or without signalization) at intersections, there were 36% fewer (–45; –26) pedestrian accidents where the pedestrian was considered guilty when the main road has a median compared to when it does not.

It should be noted that pedestrian refuge island could be used both at marked crosswalks and at unmarked crossing points. Especially where speeds are high, marking crosswalks may not be safe, but refuge islands may be the most needed at such location.

To sum up, the main advantages with a median island at pedestrian crossing points are:

- Allows pedestrians to cross more easily than if there was no island.
- May help pedestrians to cross the road more quickly, as a gap is only required in one direction of traffic at a time.
- Pedestrian refuges narrow the road, which may reduce the speed of vehicles.

The main disadvantages seem to be:

- For the pedestrian to cross safely, they must have good judgment of motor vehicle speeds and gaps in vehicle traffic.
- Visually impaired people, or those with other disabilities, may find refuge island less easy to use.
- Some motor vehicle drivers act dangerously near crossing islands if a cyclist is passing through. They may squeeze past the cyclist when passing the crossing island, or swerve dangerously around the cyclist just before the crossing island. Cyclists can feel very uncomfortable with this behavior.
- In areas with heavy snowfall, island makes it harder to remove snow.

Restrictions are as follows:

- Refuge islands must be wide enough in order to accommodate baby carriages and wheelchairs with a person behind it, a minimum of 1.2 m, and the raised areas should be on both sides of a walk area that should be at the same level as the rest of the crossing.
- Normally, road widths must be at least 3.5 m on either side of the refuge. This gives a minimum curb-to-curb street width of 8.2 m. That width has to be increased if the location is on a bend, expects oversized motor vehicles, or has bicycle traffic in the travel lane and motor-vehicle speeds are so high that bicyclists cannot comfortably claim the entire lane. Widths may also need to be increased if the municipality uses snowplows that cannot accommodate 3.5-m widths.
- Parking restrictions may need to be imposed on approaches near to the refuge, though if street width allows it, using bulb-outs means that parking can be allowed close to the crossing point.

A refuge island is also used to regulate the position of motorized traffic. For instance, traffic islands can be used at partially blind intersections on backstreets to prevent cars from cutting a corner with potentially dangerous results, or to totally prevent some movements, for traffic safety or traffic calming reasons.

Traffic Medians

When traffic islands are longer, they are instead typically called traffic medians, or median islands: a strip in the middle of a road that separates opposing lanes. In urban or suburban areas, where pedestrians are permitted, traffic medians have the same mission as refuge islands in making it easier and safer for pedestrians to cross busy and wide roads, often congested with motorized traffic. On two-lane streets, they also prohibit drivers from passing slower vehicles, reducing speed and a danger to pedestrians of being hit by a car passing another one.

The [Federal Highway Administration \(2019\)](#) has summed up the effects for pedestrians in the following way.

Providing raised medians or pedestrian refuge areas at pedestrian crossings at marked crosswalks has demonstrated a 46% reduction in pedestrian crashes. At unmarked crosswalk locations, pedestrian crashes have been reduced by 39%. Installing raised pedestrian refuge islands on the approaches to non-signalized intersections has had the most impact in reducing pedestrian crashes.

One benefit of medians is that they also provide a space to install improved lighting at pedestrian crossing locations. Improved lighting has been shown to reduce nighttime pedestrian fatalities at crossings by 78%.

Once being away from urban and suburban areas, the aim of the median strip is primarily to increase the distance between the two driving directions in order to mitigate the risk that drivers unintentionally enter the wrong side of the road and have head-on collisions. The *median strip* or *central reservation* (which on some roads may allow traffic at openings making turns) is the reserved area that separates the opposing lanes of traffic on divided roadways, such as divided highways, dual carriageways, freeways, and motorways.

Separating traffic directions is a very important aim: in the United States it has been found that roadway or lane departure crashes account for well over half of all fatal roadway crashes. The situation in Europe is more or less the same. This includes vehicles crossing the centerline and either sideswiping or striking the front end of opposing vehicles. Therefore, big efforts are made in order to mitigate this problem. The most obvious measure is to prevent vehicles physically from passing the middle of the road. However, the question is how exactly should that be done?

The reserved area in the center of the road may be paved but it is commonly adapted to other functions; for example, it may accommodate decorative landscaping, trees, or some kind of light rail or railway.

Median barriers are of course the most potential way of preventing cars from entering the wrong side of the road. It is, therefore, presented separately in the next section. However, there are several options that are used without a median barrier, and those are presented first.

The area is quite complex, however, with lots of different kinds of measures, in different context. Therefore, there are not a great number of comprehensive results regarding the effects of medians. However, the Transport Economics Institute in Oslo, Norway, has produced a meta-analysis based on approximately 40 studies (https://tsh.toi.no/index.html?146863#anchor_1468639). They are based on before–after studies with control, as well as studies with and without medians, which are controlled for various confounding variables, as well as for any important publication bias, or any regression-to-the-mean effects. It has been found that newer studies indicate a more positive effect than older ones. In most cases, where it has been possible, the results shown are, therefore, from the year 2000 or later. The studies include crashes with fatalities or crashes with unspecified severity. Later, follows the findings of the analysis of median strips with curbstones but without guardrail based on before–after studies.

Median strip on segment between intersections: The establishing of a median strip with curbstones at stretches between intersections seems to reduce the number of injury crashes, but not the number of property damage only crashes. The effect is largest on the most severe crashes. The result seems independent on what type of study it is and how old it is, and it also seems to be independent of publication bias.

The reason why a median strip seems to have larger impact on more severe crashes is probably that the median strip changes the distribution of different types of crashes, as well as the severity distribution of the crashes. Gabler et al. (2005) showed that installing a median strip reduces the number of head-on collisions but increases crashes of less severity. Another study, by Saito et al. (2005), showed that the number of side-impact collisions was reduced while there were more rear-end crashes. Normally, side collisions are more severe than rear-end crashes. In a study from New Zealand it was found that there were 75% fewer head-on collisions and a lower percentage of fatal crashes (7.2% without and 2.8% with median strip) and crashes with severe injuries (21.2% without and 16.6% with median strip) on roads with median strips.

Median Strip at Intersections

These results refer to intersections between roads with versus without median strips and are valid for intersections where at least one of the intersecting roads has a median strip (median strip for channelization is not included). At intersections, the results show that median strip with curb increases the number of crashes, mostly in rural areas. The results are not statistically significant but show the same results as median strip versus two-way left turn lanes, that is, a median strip has less favorable effect at intersections compared with segments between intersections. The explanation may be that intersections with median strips are bigger and less perspicuous compared with intersections without a median strip. Besides, a median strip is an obstacle that may cause crashes by itself.

Median Strip Instead of Two-Way Left Turn Lanes

Altogether, there was a 24% reduction of all accidents when replacing two-way left turn lanes with a continuous raised median. An even larger reduction was found for crashes between intersections. However, at intersections crashes increased by 50%. The results indicate that the effect of a median strip is larger for segments between intersections compared with at intersections. The effect is probably largest for pedestrian crashes.

The Width of the Median Strip

Roads with wider median strips have fewer crashes than roads with narrower strips. On segments between intersections, however, the effect is quite small, even though statistically significant. At intersections, the effect is larger but not statistically significant. The fact that an increase of the width of the median strip would result in a larger effect at intersections was not expected based on the

results from the studies of roads with versus without median strips, because in these cases the median strip was found to have a crash reduction only along the segments between intersections, not at intersections. There is obviously a need for further research of this. One possible explanation is that the wide median makes the intersection appear as two separate intersections with one-way streets, reducing complexity for drivers, whereas a narrow refuge island does not give the driver enough time to navigate out of the first before they have entered the second.

Type of Median Strip

Quite a few studies have been dealing with different types of medians. All in all, the results indicate that both painted and lowered median may have a somewhat better effect than a median with curbstones. The reason for this is unknown, but may have to do with more hits of curbstones. In general, hitting a curbstone at high speed may result in a driver losing control of their vehicle, and curbstones should probably not be used at roads with allowed speeds above 70 km/h.

Effect on Mobility

At side roads or at intersections a median prevents cars from making left turns if the design is such that crossings are not possible. In that case, roundabouts should be used to allow U-turns at strategic locations. It has been found that the effect of medians is:

- Reducing motor vehicle crashes by 15%.
- Decreasing delays (>30%) for motorists.
- Increasing capacity (>30%) of roadways.
- Reducing vehicle speeds on the roadway.
- Providing space for landscaping within the right-of-way.
- Providing space to install additional roadway lighting, further improving the safety of the roadway.
- Providing space to allow for supplemental signage on multilane roadways.
- Costing less to build and maintain if unpaved than paved medians.
- The Federal Highway Administration (FHWA) strongly encourages the use of raised medians (or refuge areas) in curbed sections of multilane roadways in urban and suburban areas, particularly in areas where there are mixtures of a significant number of pedestrians, high volumes of traffic (more than 12,000 vehicles per day), and intermediate- or high-travel speeds.
- FHWA guidance further states that medians/refuge islands should be at least 4 feet (1.2 m) wide (preferably 8 feet (2.4 m) wide for accommodation of pedestrian comfort and safety) and of adequate length to allow the anticipated number of pedestrians to stand and wait for gaps in traffic before crossing the second half of the street.
- On refuges 6 feet (1.8 m) or wider that serve designated pedestrian crossings, detectable warning strips complying with the requirements of the Americans with Disabilities Act must be installed.
- Medians are especially important at transit stop locations. Transit stops are frequently located along busy arterials at uncontrolled crossing locations. Providing medians can make these crossings safer and more appealing to existing and potential transit users. If transit stops do not have bus turnouts but rather bulb-outs narrowing the roadway, a refuge island will prohibit drivers from passing a stopped bus which will improve not just safety but also mobility for transit users since there will be fewer cars in front of the bus once it starts up.

The Ultimate Measure: Medians With Guardrails

A guardrail in the middle of the road has the option of preventing and even eliminating vehicles from entering the wrong side of the road. However, there is a difference between theory and practice. Different types of barriers are used and implemented differently in different types of context. The main types of barriers used are: concrete barriers, steel barriers, or beam guardrails. The safety effect of guardrails by themselves is discussed in a different article of this encyclopedia, but later follows some safety results as well for guardrails placed in medians.

Following are results presented from a meta-analysis made by Transport Economics in Oslo, Norway (<https://tsh.toi.no/index.html?21836>). It is based on almost 50 different studies. The most important findings are that guardrails—any type—reduce fatalities for all types of crashes by about 15% (−33; +7). Most effective seems to be concrete barriers in the median strip on a multilane road, 38% (−69; +24) reduction, while a steel barrier in the median strip on a multilane road results in 12% (−32; +13) reduction only. Both results are nonsignificant though. There are no results presented for fatal injuries together with beam guardrails. For all crashes together with unspecified seriousness, however, there are results for all three types of rails. Generally, the reductions are much smaller than for fatalities: of crashes with unspecified seriousness for all types of guardrails the reduction is 8% (−26; +2), for concrete barriers in the median strip on a multilane road there is a reduction of 13% (−46; +40), for steel barriers in the median strip on a multilane road there is a reduction of 4% (−27; +27), and for beam guardrails the reduction is 7% (−19; +7).

These results indicate that concrete barriers are more effective in preventing serious crashes, while the difference between steel and cable barriers is small. At the same time, there are studies indicating that a resilient barrier produces fewer crashes than a non-resilient barrier. One study even found that concrete barriers produced higher risk for serious injuries than steel barriers, even though it did not compensate for the lower risk due to vehicles coming over on the wrong side.

Most results indicate that a guardrail that is less resilient is more efficient in preventing vehicles from entering the wrong side of the road. It is found that vehicles crossing the median are fewer when the guardrail is made of concrete. The results, however, are not very clear. However, resilient rails are not as efficient in preventing vehicles to enter the wrong side of the road, at the same time as the consequences of those crashes that occur is more serious. For motorcycle riders, the risk of being killed is increasing when hitting a barrier, hitting a concrete barrier being the worst case.

The crash type that is mostly reduced on roads with guardrails is crashes where the vehicle is crossing the median. Other crashes where the median is not crossed increase in numbers; crashes with unspecified seriousness have increased in the meta-analysis by 73%. The main reason is that vehicles are hitting the guardrail, and stay on that side of the road, so a potential—and more serious—crash on the other side of the median had been avoided. Other types of crashes do not seem to be changed.

Generally, it must be said that most results in the meta-analysis are not statistically reliable. Additional analyses indicate that there are real differences between outcome with regard to different injury classes and between different crash types. The differences between different rail types are, however, not statistically reliable. The differences found may therefore not be real. This can even apply to the results for different types of guardrail, both results where a direct comparison between different types of rails has been made and other studies have shown that more resilient rails have a more positive effect than less resilient ones.

Less Expensive but Still Effective: Centerline Rumble Strips

The safety effect of rumble strips, both continuous shoulder and centerline rumble strips, are addressed in a separate article of the encyclopedia, but since centerline rumble strips are a type of non-raised narrow median, they are also included here, from a slightly different perspective.

Rural two-lane roads generally lack physical measures such as wide medians or barriers to separate opposing traffic flows. Or, even if there is space enough, the cost may be too high to install it. One option then is to install rumble strips in the centerline. These strips and stripes are designed to address crashes caused by distracted, drowsy, or otherwise inattentive drivers who drift from their side of the road. With respect to fatal head-on crashes on rural two-lane roads, it has been found in studies of several jurisdictions that roughly half are caused by inattentiveness or drowsiness, and the other half by losing control and crossing the centerline because of too high speed. Only a few percent of head-on fatalities are caused by people crossing the median on purpose to pass another vehicle. In other words, continuous rumble strips should be able to affect about half of all serious head-on crashes.

Centerline rumble strips are one of the many infrastructure-related measures applied to prevent the occurrence of car accidents, especially crossover, run-off-the-road, and head-on accidents. Quite a few studies have been carried out, in different countries. There is a meta-analysis made at Transport Economics Institute in Oslo based on almost 20 studies (https://tsh.toi.no/325-forsterket-kantoppmerking.htm#anchor_136975-90). It shows that there was a 10% reduction of all crashes and as much as a 37% reduction of the target behaviors (head-on and side-impact collisions at left side and leaving the road to the left side). Both reductions are statistically significant, and similar for different degrees of severity. All studies were made on two-lane roads in non-built-up areas. The main reason for reductions seems to be either that drivers are alerted when they hit the strips, or that they are driving with somewhat longer distances to the strips (which is good but potentially could have negative effects on bicycle safety on roads lacking bicycle tracks and wide shoulders). Generally, it was not found that speeds were changed. Two studies found that nighttime speeds were lower, but that has not been confirmed in other studies.

Centerline rumble strips have been installed on many of the Danish rural roads in the last decade. As previous studies on the effectiveness of this measure were not conducted in Denmark, and many roads have centerline rumble strips installed, it was deemed necessary to implement an appropriate study. This study investigated whether centerline rumble strips really work. By using previously recorded accidents data and traffic volumes from the Road Directorate, the study evaluated the effectiveness of safety improvement on each site as well as the overall safety program. Meta-analysis method was applied to 35 sites from locations all over the country. The evaluation result complied with those from the international studies. The total number of accidents was reduced by 20% following the installations of the centerline rumble strips.

A US study presents similar conclusions. Opposing-direction crashes account for about 20% of all fatal crashes on rural two-lane roads and result in about 4500 fatalities annually in the United States. An empirical Bayes before–after procedure was employed to properly account for regression to the mean while normalizing for differences in traffic volume and other factors between the before and after periods. Overall results indicated significant reductions for all injury crashes combined (14%, 95% confidence interval (95% CI) = 5%–23%) as well as for frontal and opposing-direction sideswipe injury crashes (25%, 95% CI = 6%–44%)—the primary target of centerline rumble strips. The authors summarize: in light of their effectiveness and relatively low installation costs, consideration should be given to installing centerline rumble strips more widely on rural two-lane roads to reduce the risk of frontal and opposing-direction sideswipe crashes. The US FHWA describes Centerline Rumble Strips as a “Proven Safety Countermeasure,” with a 44%–64% reduction of head-on, opposite-direction, and sideswipe fatal and injury crashes.

Median Barrier on 13-m Wide Roads

Deaths on rural roads are a serious road safety problem. Due to the risks imposed by high speeds, multifunctionality, lower infrastructure safety, and mix of different road users, rural roads are often dangerous roads with relatively high risk levels compared to motorways. In Sweden, a special road safety concern during the 1990s was the large number of fatal and serious crashes on rural

13-m wide two-lane roads. These 13-m roads had two travel lanes and shoulders that were about 3-m wide. And, in Sweden, it is legal to enter the shoulder when driving a farm tractor or other vehicle at low speed to let faster vehicles pass. By the 1980s, a culture had developed that people going close to the speed limit often entered the shoulder to allow faster vehicles to pass. Then, faster drivers started to expect everybody to do that when passed, and if people did not move over, some drivers would cross the centerline assuming oncoming traffic would enter their shoulder, sometimes resulting in serious crashes. In 1997, the Swedish Parliament adopted the so-called Vision Zero (see separate article in this encyclopedia), a vision that nobody should be killed or seriously injured on Swedish roads. One main solution proved to be a redesign of the 13-m wide roads to 2 + 1 lanes with a median barrier. Starting in 2009, this solution was also applied on rural roads with road width of about 9–10 m. A 2 + 1 road with median barrier has a continuous three-lane cross-section with alternating passing lanes to allow defined sections of overtaking, the two directions of travel separated by a flush divider with a median barrier. The barrier is in Sweden a cable barrier. Comparing 13-m wide, 2 + 1 roads with narrow 2 + 1 roads (9 m), the main difference is the length and frequency of passing lanes. For narrow 2 + 1 roads, the share of passing lanes varies between 15% and 30% compared to about 40% for the 13-m roads. (The reason the wider roads with continuous three lanes do not give 50% passing opportunities is that the transition from two lanes to one lane needs a safety zone when no one can use the middle lane, where the cable barrier moves from one side to the other.)

The safety effect of 2 + 1 roads, 13-m wide in Sweden, has been extensively assessed. Compared with normal 13-m wide roads there is a reduction of the number of killed by more than 75%, independent of whether the speed limit is 90 km/h or 100 km/h. The effect on killed and seriously injured is between 50% and 60%, highest for a speed limit of 90 km/h, while for injury crashes there is a much smaller effect and only on 90 km/h roads, minus 13%. The result of this is that Sweden recently has stopped using the almost “universal” speed limit of 90 km/h on 13-m roads and now use 80 km/h on two-lane roads without center barriers and 100 km/h on roads with center barrier. (A few 13-m roads that have grade separated interchanges only have maintained 110 km/h.) The safety is thereby increased on both the roads with and without center barriers, compared to the traditional design with 90 km/h limit, and mobility improved on the ones with cable barriers and is reduced on the ones where the speed limit was lowered.

Crashes with the guardrail are quite frequent but less severe. On average, there are two hits per kilometer and year. This involves of course costs for repairing.

Motorcycle riders have been quite negative to the cable wires in Sweden. However, the result is a reduction of serious and fatal crashes by 40%–50% for motorcyclists. This result contradicts the earlier result built on a meta-analysis. The reason for the difference is most probably that in the meta-analysis the barrier that was used was primarily concrete barriers, while in the 2 + 1 road case cable barriers were used.

Regarding mobility, it was found that the average speed went up by 2 km/h on roads with 90 km/h speed limit (increase in the direction that went to two lanes) while it was unchanged for roads with 110 km/h speed limit. It was also found that capacity is around 1600–1700 motor vehicles per hour measured during a 15-min period, which is 300 vehicles—15%—less compared with a normal 13-m wide road.

A special evaluation of the safety effects of narrow 2 + 1 roads in Sweden has also been made. Those roads had narrow shoulders, in the before period and did not allow four informal lanes of traffic, which the 13-m two-lane roads did and therefore had different safety issues in the before period. Results from a before and after study show a number of significant effects; the total number of fatalities and seriously injured decreased by 50% and the total number of personal injury crashes decreased by 21%. Looking only at links (excluding intersections), the number of fatalities and seriously injured decreased by 63% and the personal injury crashes by 28%. For almost all of the included road sections, the speed limit was also raised from 90 km/h to 100 km/h when the road was rebuilt to 2 + 1. Regarding mobility there was no difference in efficiency compared to earlier evaluations of traditional 2 + 1 roads with 100 km/h and 40% passing lanes can be observed.

There are some special complications that have to be carefully discussed and treated: the first point is how emergency vehicles are accommodated on one-lane stretches and the second one is the risk of cars breaking down on one lane stretches. The third one is how maintenance work is done. The fourth one is the restricted access of property owners, and the fifth is public perception. In Sweden, there have also been a lot of discussions around the fact that pedestrians and bicycle riders are prohibited on these kinds of roads. In the south of Sweden, where biking is quite common, the authorities have built a bike road parallel to the 2 + 1 road. This, of course, makes the cost of reconstruction of the road considerably higher. And, in areas with little bicycle traffic, such parallel bicycle paths may not be built adding considerable distance for bicycling to certain destinations.

Although in smaller scale than in Sweden, there are 2 + 1 roads also in Ireland, Norway, Germany, and Finland. In Ireland cable barrier is used as in Sweden, while in Norway a concrete barrier is used. In both cases the effects are similar to the ones in Sweden. In Germany and Finland, the 2 + 1 roads are done by restriping wider roads. The effect is, therefore, considerably smaller than in Sweden.

Biography

Christer Hydén has been active in the traffic safety field for around 40 years. His first achievement was the development of a Traffic Conflict Technique, a way of assessing risk on the roads with the help of near accidents. He was the President of ICTCT (at first standing for “International Co-operation in Traffic Conflicts Techniques” and later for “International Co-operation on Theories and Concepts in Traffic Safety”) from 1977 to 2011. One other field of expertise is the impact of speed, and the effect of different measures, like Traffic Calming and Speed Limiters in vehicles. He is a member of the Independent Council for Road Safety International (ICORSI).

References

- Federal Highway Administration, 2019. Safety benefits of raised medians and pedestrian refuge areas. Available from: https://safety.fhwa.dot.gov/ped_bike/tools_solve/medians_trifold/medians_trifold.pdf.
- Gabler, H.C., Gabauer, D.J., Bowen, D., 2005. Evaluation of cross median crashes. Report FHWA-NJ-2005-04. Rowan University, Glassboro, NJ.
- Huang, H.F., Cynecki, M.J., 2001. The effects of traffic calming measures on pedestrian and motorist behavior. Report FHWA-RD-00-104. Federal Highway Administration, McLean, VA.
- Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle-pedestrian crashes at intersections in Florida. *Acc. Anal. Prev.* 37 (2), 775–786.
- Saito, M., Cox, D.D., Jin, T.G., 2005. Evaluation of four recent traffic and safety initiatives, Vol. II: developing a procedure for evaluating the need for raised medians. Utah Department of Transportation Research and Development Division: Final Report. UDOT, Taylorsville, UT.
- Zegeer, C.V., Stewart, J.R., Huang, H.H., Lagerwey, P.A., Feaganes, J., Campbell, B.J., 2005. Safety effects of marked versus unmarked crosswalks at uncontrolled locations. Report FHWA-HRT-04-100. University of North Carolina, Highway Research Center, Chapel Hill, NC.

Further Reading

- Carlsson, A., 2009. Evaluation of 2 + 1-roads with cable barriers: final report. Statens väg- och transportforskningsinstitut—VTI, Sweden.

Risk Perception and Risk Behavior in the Context of Transportation

Martina Raue*, Eva Lerner^{†,‡}, *Massachusetts Institute of Technology AgeLab, Cambridge, MA, United States; [†]LMU Center for Leadership and People Management, Munich, Germany; [‡]FOM University of Applied Sciences for Economics and Management, Munich, Germany

© 2021 Elsevier Ltd. All rights reserved.

Perceived Risk and Objective Risk	494
Risk Judgments	495
Bounded Rationality and the Affect Heuristic	495
Risk as Feelings and Perceived Control	496
Risk Perception and Risk Behavior	496
Mobile Phone Use While Driving	496
Learned Carelessness	496
Risk Homeostasis and Risk Compensation	497
Risk Communication	497
Acknowledgment	497
Biographies	498
References	498
Further Reading	499

Perceived Risk and Objective Risk

Risk is the probability and magnitude of harm, such as the probability and consequences of getting into an accident when driving a car. However, peoples' judgments of risks often deviate from actual probabilities of harm; as a consequence, we need to differentiate between *perceived risks* and *objective risks*. Objective risks are measurable and require analytical processing of facts such as the number of fatalities by miles driven per year. Perceived risks are based on subjective estimates for the likelihood of harm. Risk perception has been defined in various ways, including the subjective assessment of risk and the level of concern associated with the consequences (Sjöberg et al., 2004) or beliefs, attitudes, judgments, feelings and values associated with a risk (Pidgeon et al., 1992). Some social scientists argue that risk assessment is inherently subjective, because it always includes components of subjective judgment such as selection and interpretation of data and deciding the consequences. An engineer's risk estimate may be based on theoretical models, but the model is driven by assumptions and individual judgment. The subjectivity of risk assessment is also demonstrated, for example, by the choice of a measure for traffic mortality risks. Mortality risks could be expressed by deaths per mile driven or by deaths per million cars. As a result, although danger is real, the assessment of risk may never be purely objective (Slovic, 2016).

Risk means different things to different people. Assessing risks based on probabilities of harm requires cognitive analysis, which is one way that humans process information about risk. Another more common way is reliance on intuition, affective reactions, and feelings, which is quicker, easier and often more efficient than reliance on cognitive analysis (Slovic and Peters, 2006). From an evolutionary perspective, being able to quickly and automatically react to danger is essential for survival. However, while intuition leads to good decisions in most cases, it can sometimes lead people astray. This explains why risk assessments of experts and laypeople often differ. While experts (e.g., risk analysts) use cognitive analysis and rely on statistics, laypeople (e.g., consumers) are sensitive to more emotional factors such as catastrophic potential or controllability. For example, after the terrorist attacks in the United States on September 11, 2001, many Americans dreaded flying and decided to drive instead. As a result, traffic fatalities increased in the three months following the attack. In fact, the data suggest that more Americans died on the road by avoiding flying in the aftermath of the attack than the total number of passengers killed on the four flights that were attacked (Gigerenzer, 2004). This example demonstrates that people's perceptions of risk are more sensitive to the possibility of outcomes that carry strong emotional responses rather than probabilities. Because people focus so much on the negative outcome, they are less attentive to the small probability of occurrence (Loewenstein et al., 2001; Rottenstreich and Hsee, 2001; Sunstein, 2003).

In their psychometric paradigm, Fischhoff et al. (1978) demonstrated the multidimensionality of risk perception by distilling several characteristics of hazards that influence subjective risk estimates among laypeople. These characteristics include voluntariness of exposure, control over the risk, immediacy of consequences, familiarity with and knowledge about the hazard, the hazard's catastrophic potential, its dreadfulness, the severity of its consequences, and to what extent the hazard is known to science. Most of these characteristics were shown to highly correlate (e.g., risks that are rated as voluntary are often also perceived as controllable) and could be narrowed down to two main dimensions: dread risk and unknown risk (or novelty). Dread risk is characterized by lack of control, catastrophic potential and fatal consequences associated with an activity or event, while unknown risks are characterized by being unfamiliar, unobservable, poorly understood by science, and delayed in their consequences. Driving is a voluntary and familiar experience, which is often perceived as controllable. As a result, risk perception is reduced, which may at least in part explain why drivers were long reluctant to wear a seatbelt (Slovic et al., 1978) and why many do not refrain from texting while driving.

Further, studies have indicated that dread risk is a stronger predictor for people's risk perceptions than unknown risk (Slovic, 1987). As a result, people especially avoid risks that are fatal to many people at once. A hazard that may cause the same amount of fatalities over a longer period is thus less dreaded. This explains why many people generally fear flying more than driving, although from a statistical perspective, taking the car to the airport is the most dangerous part of the trip. Dread risks are perceived as less controllable, and people often have little knowledge about their actual probabilities. Additionally, dread risks that have the potential to kill many people at once may be a substantial threat to the society (Galesic and Garcia-Retamero, 2012).

Humans are driven by a need for safety, control, and predictability (van den Bos, 2009). Because the world rarely offers complete certainty, humans depend on often unconscious strategies that increase their perceptions of safety and decrease their perceptions of risk. For example, while people make quite accurate judgments about societal traffic risks (Lichtenstein et al., 1978), many do not feel that these risks apply to them personally. Most people underestimate personal risks (e.g., due to perceived control) and believe that they are safer drivers than the average driver (Svenson, 1981). One study found that after only one year of driving practice, most drivers overestimate their technical skill (e.g., reaction times, driving on slippery roads) and after three years they consider themselves above average in general driving skills (Svenson et al., 1985). These results have not been without critique, however, as data analysis has indicated that most drivers actually have few accidents while few have many accidents, and therefore most drivers are safer than the arithmetic mean (Gigerenzer et al., 2012). In addition, most people think they are less vulnerable than other people and that others are more likely to be affected by harm, which has been termed *unrealistic optimism* or *optimism bias* (Weinstein, 1980). Unrealistic optimism is associated with feelings of control over the driving situation. People acquire feelings of control over driving situations, in part, through driving experience regardless of whether they have experienced accidents or not (Svenson et al., 1985).

Risk Judgments

Bounded Rationality and the Affect Heuristic

Driving situations are complex, as various pieces of information must be considered when making judgments behind the wheel. These pieces of information include one's own speed, road and weather conditions, or predictions about another driver's behavior. In addition to these considerations, judgments and decisions in driving situations often have to be made under time pressure. However, the human mind is not able to consciously consider all the information that is relevant in a given situation, especially when under time pressure, which has been termed *bounded rationality* (Simon, 1955). Throughout their evolutionary history, humans have survived by relying on their intuition and quickly responding to threats in terms of fight-or-flight. In a driving situation, quick reactions such as braking or swerving happen without having calculated the time it takes to approach the car in front of us. Relying on intuition and feelings is often useful in situations that require quick decision making, but sometimes feelings can also get in the way. In a study on driving behavior, participants with neurological impairments who lacked emotional reactions stayed calm when hitting ice on the road, while healthy participants were more likely to panic and slam the brakes, which caused the car to skid out of control (Shiv et al., 2005).

In order to make complex decisions under time pressure, humans often rely on mental shortcuts, also termed *heuristics*, that simplify decisions, but maintain a sufficient level of accuracy (Raue and Scholl, 2018; Tversky and Kahneman, 1974). Relying on affect (i.e., good or bad feelings) to make judgments about risks is one way to simplify a complex situation and has been labeled the *affect heuristic*. Although the analysis of risks is important for decision making, relying on feelings is quicker, easier, and often more efficient (Slovic, 2016). In fact, rational decision making is generally preceded by feelings and seems to depend on prior emotional processing (Bechara and Damasio, 2005; Zajonc, 1980). According to the affect heuristic, affect that is associated with a hazard serves as a cue for the perceived likelihood of harm. Voluntary activities such as driving are often associated with positive affect—that is, it feels good to do them—which decreases risk perceptions associated with the activity. However, negative personal experiences or fear-evoking media reports may increase risk perception. For example, in the rare event of an airplane accident, the media reports extensively about it, often by using affect-laden images. In addition to the emotional reactions these reports create, media attention also makes the event easier to come to mind in the aftermath, which is an example of the *availability heuristic* (Tversky and Kahneman, 1974). People use the availability heuristic to make judgments about frequencies or probabilities; an event that comes easily to mind makes it seem more likely. Usually, more frequent events come to mind more easily, but people's perceptions about the frequency of occurrence can be biased due to, for example, vivid media coverage (Combs and Slovic, 1979).

The affect heuristic has especially been studied in the area of technology acceptance. In these studies, it was found that perceived risk and perceived benefits associated with a technology are inversely related. In general, highly risky technologies are only used if they are also highly beneficial, but in most people's minds perceived risk declines with an increase in perceived benefit and vice versa. This relationship could be explained by the affect heuristic as positive affect results in lower risk and higher benefit perception, while negative affect results in higher risk and lower benefit perceptions. This inverse relationship between risk and benefit perceptions increases under time pressure, which further supports the affect heuristic by indicating that people's judgments are not based on cognitive analysis (Finucane et al., 2000; Slovic and Peters, 2006). Recent research has suggested that affect associated with a known technology, such as traditional driving, may even inform people's risk perceptions of a new, but similar, technology such as self-driving cars (Raue et al., 2019).

Risk as Feelings and Perceived Control

Another approach is the risk-as-feelings hypothesis, which focuses on the behavioral consequences of feelings associated with risks (Loewenstein et al., 2001). Negative feelings associated with flying (e.g., due to negative experiences or intensive media coverage of very rare negative events) may prevent some people from traveling by airplane although most people *know* that flying is safe. Education about airplane safety, however, often does little to change the behavior of an anxious flyer. In fact, people's judgments about the safety of commercial aviation are quite accurate when compared with accident statistics (Slovic et al., 1979). Despite the knowledge that air travel is very safe, many feel anxious about flying. In an airplane, passengers have no control (not even an illusion of control as many people perceive when driving a car) and cannot see the actions of the pilot. Being in an airplane is an unfamiliar situation for many and most people usually have only a vague idea of how airplanes work. And while accidents with airplanes are extremely rare, consequences of accidents are almost always fatal and catastrophic as many people die in one single incident (Mauro, 2019).

Recent research has also suggested that giving up control may be a major barrier to the adoption of self-driving cars (Raue et al., 2019). The self-driving car will affect the future of transportation, but research has indicated that the public perceives the technology as riskier than conventional driving, which may hinder adoption (Brell et al., 2018; König and Neumayr, 2017; Ward et al., 2017). Having to give up control, as well as lack of familiarity, poor understanding of how the technology works, lack of trust, and unresolved ethical issues (Meder et al., 2018; Visschers and Siegrist, 2018) may explain some of the skepticism associated with self-driving cars. People initially also viewed the conventional car with skepticism, but it ultimately earned their trust and acceptance (Ladd, 2008). In fact, people's attitudes and perceptions might be more of a barrier to a successful introduction of self-driving cars than technical challenges (Coughlin et al., 2019). Experience with automated vehicle technologies already on the market has been associated with a decrease in risk perceptions associated with self-driving cars (Brell et al., 2018; Raue et al., 2019).

Risk Perception and Risk Behavior

Risk perception and risk behavior are generally negatively correlated: the higher one perceives the risk of a behavior the less likely the person will engage in that behavior (Ferrer and Klein, 2015; Mills et al., 2007). However, despite having been educated about certain risks (e.g., drinking and driving, texting and driving) and showing less unrealistic optimism (Fischhoff et al., 2010; Quadrel et al., 1993), adolescents tend to engage in more risky driving behavior than older and more experienced drivers (Gershon et al., 2018). For example, adolescents are more likely to engage in secondary tasks such as mobile phone usage or reaching for objects, which takes their eyes off the road and increases their crash risk (Gershon et al., 2019). Explanations for risky driving behavior among adolescents include lack of driving experience and high susceptibility to peer influence (Gershon et al., 2018; Frey et al., 2016). Social interactions strongly influence adolescents in various contexts and peer presence has been shown to increase their risk-taking behavior in driving simulators (Steinberg, 2008). Increased risk-taking behavior might result from adolescents' preference for trading off risks and rewards, thus, putting more weight on the benefits (e.g., peer acceptance) than the risk (Reyna et al., 2015).

Mobile Phone Use While Driving

In 2017, 9% of fatalities in motor vehicle crashes involved distracted drivers and 14% of these distractions were caused by mobile phone use (National Highway Safety Administration, 2019). Nevertheless, in a recent study conducted in Germany (Lemster and Lerner, 2019), 52% of the participants admitted using their mobile phones at least occasionally when driving. In a study with college students, 91% indicated that they occasionally text while driving despite agreeing that this behavior was dangerous (Harrison, 2011). Perceptions of being in control of the situation, mobile phone dependence and attachment, propensity for risk-taking, and impulsivity have been shown to contribute to people's use of mobile phones while driving (Beck and Watters, 2016; Weller et al., 2013). In addition, compensatory beliefs such that negative effects of risky behavior can be reduced or eliminated by showing safe behavior (e.g., using the mobile phone after having slowed down) contribute to risky behavior as well (Zhou et al., 2016). Although slowing down may decrease the actual risk of using a mobile phone while driving, it is still a distraction that may result in an accident. Specifically, among adolescents, predictors for using mobile phones while driving further include social influences such as the fear of missing out or the fear of being perceived as rude or impolite if not responding to a message immediately (Lemster and Lerner, 2019; Martha and Griffet, 2007).

Learned Carelessness

Research suggests that decreased risk perception can also be explained by repeated risky behavior without negative consequences, which results in learned carelessness (Frey et al., 2016; Raue and Schneider, 2019). According to the *theory of learned carelessness*, learning experiences based on positive consequences of behavior that are due to luck, coincidence or repeated risky behavior without negative consequences may lead to biased perceptions of risk (Frey and Schulz-Hardt, 1997). For example, if speeding or driving under the influence of alcohol has never resulted in an accident or legal consequences, a person is likely to continue the behavior and might increase their speed even more or have another beer before getting into the car. The absence of any negative experience can

lead to habituation and a biased perception of an actual risk and, consequently, to learned carelessness. Routine procedures with no encounters of critical incidents after many repetitions are especially prone to developing learned carelessness. In aviation, for example, checklists are used to ensure all precautionary measures are taken and to prevent human error. In an experimental study, airline pilots who repeatedly encountered errors while going through their flight plans were more likely to detect errors in the future than those who never encountered errors. However, no differences were found in their performance of flight inspections (Aust et al., 2011). The theory of learned carelessness suggests that when things are going well, people tend to feel safe and secure and to assume that these conditions will continue, despite a potential increase in risk. This also reflects people's general insensitivity to risk accumulation over repeated exposure (Fischhoff, 2009). People sometimes adjust their behavior to match their biased sense of safety (e.g., increasing speed, engaging in critical shortcuts in routine safety procedures) rather than adapting it to the actual risk. However, more research is needed to further validate the theory of learned carelessness.

Risk Homeostasis and Risk Compensation

One explanation for this matching process is based on risk homeostasis theory (Lermer et al., 2016; Wilde, 1982), also known as risk compensation hypothesis (Wilde, 2002). This theory states that people have the tendency to show riskier behavior (e.g. risky driving) after a perceived increase in safety (e.g. installation of a specific safety system). It is assumed that each person has a certain individual level of risk that he or she is willing to tolerate. When external factors contribute to an increase in perceived safety, this safety gain is offset by an increase in risky behavior. This means that the safer people feel, the riskier their behavior becomes. For example, after the introduction of the antilock brake system in cars, the number of accidents did not fall as expected, but increased (e.g., Biehl et al., 1987). The increase in risky behavior is not limited to safety perceptions of oneself. For instance, a field study revealed that cyclists wearing helmets are overtaken by car drivers at a shorter distance than cyclists without helmets (Robinson, 1996). Another study showed that more accidents tend to occur on easily manageable bends than on unclear stretches of the road, which drivers subjectively consider to be more dangerous. Drivers on clearly arranged routes chose significantly higher speeds (Petermann et al., 2008). However, it should be noted that the risk homeostasis theory is often controversially discussed. One counterargument is that the theory has not been supported in observations of seatbelt use among drivers (Evans et al., 1982) nor for bicycle helmet use among bicyclists. Most studies actually found that helmet use was associated with safer cycling behavior (Esmailikia et al., 2019).

While behavior may change occasionally as a result of a perceived change in risk, critiques argue that empirical evidence for risk homeostasis is still limited (Robertson and Pless, 2002). Although complete risk compensation as predicted by risk homeostasis theory lacks empirical support, there is evidence for behavioral adaption to some traffic safety measures. However, the effects and magnitude of such partial risk compensation behavior vary and not every safety measure is necessarily compensated. Because the effects of safety measures on risk behavior can rarely be studied in controlled experimental conditions, accurately capturing the multidimensionality of risk is difficult and may impede conclusive results on risk compensation (Adams, 1995; Evans, 1985; Hedlund, 2000). Various methodologies and forms of risk measurements, including self-reports, crash data, field, and laboratory studies may also explain conflicting results of studies on risk compensation (see Lermer et al., 2018 for a discussion on measuring subjective risk estimates).

Risk Communication

Most people generally struggle to understand statistical information, which is essential to make sense of risk statistics (Peters, 2012). Due to this lack of numeracy of the general public, their ability to make informed decisions based on the interpretation of statistical information about risk is limited. Because of this and their tendency to base their judgments on feelings rather than cognitive analyses, people are more susceptible to individual stories or vivid media reports than statistical information about risk, both of which can not only influence but also distort their perceptions and judgments. Therefore, alternative ways to communicate risks to the public have been suggested. In the health sector, for example, the use of visual aids has been successful in supporting people in understanding the risks and benefits of preventive screenings or medical procedures (Garcia-Retamero and Cokely, 2013). However, as shown earlier, even if people are aware of the risks associated with a behavior (e.g., texting and driving), they may not change their behavior due to unrealistic optimism, feeling in control of the situation, weighing risks and benefits and/or being influenced by their peers. Effective risk communication that not only addresses the way people perceive risk, but also acknowledges their social and psychological context is an important tool to increase personal relevance and ultimately change behavior (Bostrom et al., 2018). In the transportation sector, effective risk communication is essential on an individual and societal level. This includes preventing people from risky behavior such as speeding, texting while driving, or driving under the influence of alcohol, but also supporting them in making informed judgments about new technologies and encouraging proenvironmental behavior.

Acknowledgment

We thank Adam Felts for proofreading this chapter.

Biographies

Martina Raue is a Research Scientist at the MIT AgeLab in Cambridge, Massachusetts. She studies risk perception and decision making over the lifespan. Dr. Raue received her PhD in Social Psychology from the Ludwig-Maximilian University Munich in Germany and her Master's Degree in Psychology from the University of Basel in Switzerland.

Eva Lerner is a Professor for Business Psychology at the FOM University of Applied Sciences for Economics and Management in Munich, Germany, and Head of the Peer-to-Peer Mentoring Program at the Ludwig-Maximilian University Munich. Her research interests include risk behavior, decision making, and positive psychology. Dr. Lerner received her academic degrees from the University of Salzburg in Austria and the Ludwig-Maximilian University Munich in Germany.

References

- Adams, J., 1995. Risk. University College London Press, London, UK.
- Aust, F., Moehlenbrink, C., Jipp, M., 2011. Operationalization of learned carelessness. *Proc. Hum. Fac. Ergon. Soc. Ann. Meet.* 55 (1), 1735–1739, doi:10.1177/1071181311551360.
- Bechara, A., Damasio, A.R., 2005. The somatic marker hypothesis: a neural theory of economic decision. *Games Econ. Behav.* 52 (2), 336–372.
- Beck, K.H., Watters, S., 2016. Characteristics of college students who text while driving: Do their perceptions of a significant other influence their decisions? *Transp. Res. F Traff. Psychol. Behav.* 37, 119–128, doi:10.1016/j.trf.2015.12.017.
- Biehl, B., Aschenbrenner, M., Wurm, G., 1987. Einfluss der Risikokompensation auf die Wirkung von Verkehrssicherheitsmaßnahmen am Beispiel ABS [Influence of risk compensation on the effects of road safety measures using ABS as an example]. *Unfall-und Sicherheitsforschung Straßenverkehr* 63, 65–70.
- Bostrom, A., Böhm, G., O'Connor, B., 2018. Communicating risks: principles and challenges. In: Raue, M., Streicher, B., Lerner, E. (Eds.), *Psychological Perspectives on Risk and Risk Analysis - Theory, Models and Applications*. Springer, Chum, Switzerland.
- Brell, T., Philipsen, R., Zieffle, M., 2018. sCARY! Risk perceptions in autonomous driving: the influence of experience on perceived benefits and barriers. *Risk Anal.* 14 (6), 1085, doi:10.1111/risa.13190.
- Combs, B., Slovic, P., 1979. Newspaper coverage of causes of death. *Journal. Mass Commun. Quarter.* 56 (4), 837–849, doi:10.1177/107769907905600420.
- Coughlin, J.F., Raue, M., D'Ambrosio, L.A., Ward, C., Lee, C., 2019. Special series: social science of automated driving. *Risk Anal.* 39 (2), 293–294, doi:10.1111/risa.13271.
- Esmailikia, M., Radun, I., Grzebieta, R., Olivier, J., 2019. Bicycle helmets and risky behaviour: a systematic review. *Transp. Res. F Traff. Psychol. Behav.* 60, 299–310, doi:10.1016/j.trf.2018.10.026.
- Evans, L., Wasielewski, P., von Buseck, C.R., 1982. Compulsory seat belt usage and driver risk-taking behavior. *Hum. Fact.* 24 (1), 41–48, doi:10.1177/001872088202400105.
- Evans, L., 1985. Human behavior feedback and traffic safety. *Hum. Fact. J. Hum. Fact. Ergon. Soc.* 27 (5), 555–576, https://dx.doi.org/10.1177/001872088502700505.
- Ferrer, R.A., Klein, W.M., 2015. Risk perceptions and health behavior. *Curr. Opin. Psychol.* 5, 85–89, doi:10.1016/j.copsyc.2015.03.012.
- Finucane, M.L., Alhakami, A., Slovic, P., Johnson, S.M., 2000. The affect heuristic in judgments of risks and benefits. *J. Behav. Decis. Mak.* 13, 1–17.
- Fischhoff, B., 2009. Risk perception and communication. In: Fischhoff, B. (Ed.), *Risk Analysis and Human Behavior*. Routledge, London, doi:10.1093/med/9780199218707.003.0057.
- Fischhoff, B., de Bruin, W., Parker, A.M., Millstein, S.G., Halpern-Felsher, B.L., 2010. Adolescents' perceived risk of dying. *J. Adoles. Health* 46 (3), 265–269, doi:10.1016/j.jadohealth.2009.06.026.
- Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., Combs, B., 1978. How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sci.* 9 (2), 127–152.
- Frey, D., Schulz-Hardt, S., 1997. Eine Theorie der gelernten Sorglosigkeit [a theory of learned carelessness]. *Bericht über den 40. Kongress Der Deutschen Gesellschaft Für Psychologie*, pp. 604–611.
- Frey, D., Ullrich, B., Streicher, B., Schneider, E., Lerner, E., 2016. Theorie der gelernten Sorglosigkeit [theory of learned carelessness]. In: Frey, D., Bierhoff, H.-W. (Hrsg.) *Enzyklopädie der Psychologie - Selbst und soziale Kognition – Sozialpsychologie [Encyclopedia of Psychology – Self and Social Cognition – Social Psychology]*, 1 (S. 429–469). Göttingen: Hogrefe.
- Galesic, M., Garcia-Retamero, R., 2012. The risks we dread: a social circle account. *PLoS One* 7 (4), e32837, doi:10.1371/journal.pone.0032837.
- Garcia-Retamero, R., Cokely, E., 2013. Communicating health risks with visual aids. *Curr. Direct. Psychol. Sci.* 22 (5), 392–399.
- Gershon, P., Ehsani, J.P., Zhu, C., Sita, K.R., Klauer, S., Dingus, T., et al., 2018. Crash risk and risky driving behavior among adolescents during learner and independent driving periods. *J. Adoles. Health* 63, 568–574, doi:10.1016/j.jadohealth.2018.04.012.
- Gershon, P., Sita, K.R., Zhu, C., Ehsani, J.P., Klauer, S.G., Dingus, T.A., et al., 2019. Distracted driving, visual inattention, and crash risk among teenage drivers. *Am. J. Prevent. Med.* (56), 494–500, doi:10.1016/j.amepre.2018.11.024.
- Gigerenzer, G., 2004. Dread risk, September 11, and fatal traffic accidents. *Psychol. Sci.* 15 (4), 286–287, doi:10.1111/j.0956-7976.2004.00668.x.
- Harrison, M.A., 2011. College students' prevalence and perceptions of text messaging while driving. *Accid. Anal. Prevent.* 43 (4), 1516–1520, doi:10.1016/j.aap.2011.03.003.
- Hedlund, J., 2000. Risky business: safety regulations, risk compensation, and individual behavior. *Injury Prevent.* 6 (2), 82, https://dx.doi.org/10.1136/ip.6.2.82.
- Knoll, L.J., Magis-Weinberg, L., Speekenbrink, M., Blakemore, S.-J., 2015. Social influence on risk perception during adolescence. *Psychol. Sci.* 26 (5), 583–592, doi:10.1177/0956797615569578.
- König, M., Neumayr, L., 2017. Users' resistance towards radical innovations: the case of the self-driving car. *Transp. Res. F Traff. Psychol. Behav.* 44, 42–52, doi:10.1016/j.trf.2016.10.013.
- Ladd, B., 2008. *Autophobia: Love and Hate in the Automotive Age*. University of Chicago Press, Chicago, IL.
- Lernster, S., Lerner, E., 2019. Dark personality traits, smartphone use and driving behavior - a risky mix. Manuscript in preparation.
- Lerner, E., Raue, M., Frey, D., 2016. Risikowahrnehmung und Risikoverhalten [Risk perception and risk behavior]. In: Frey, D., Bierhoff, H.-W. (Eds.), *Enzyklopädie der Psychologie – Sozialpsychologie [Encyclopedia of Psychology – Social Psychology]* vol. 2. Hogrefe, Göttingen, pp. 535–580.
- Lerner, E., Streicher, B., Raue, M., 2018. Measuring subjective risk estimates. In: Raue, M., Lerner, E., Streicher, B. (Eds.), *Psychological Perspectives on Risk and Risk Analysis: Theory, Models, and Applications*. Springer, New York, pp. 313–327.
- Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., Combs, B., 1978. Judged frequency of lethal events. *J. Exp. Psychol. Hum. Learn. Memory* 4 (6), 551–578.
- Loewenstein, G.F., Weber, E.U., Hsee, C.K., Welch, N., 2001. Risk as feelings. *Psychol. Bull.* 127 (2), 267–286, doi:10.1037/0033-2909.127.2.267.
- Martha, C., Griffet, J., 2007. Brief report: How do adolescents perceive the risks related to cell-phone use? *J. Adoles.* 30 (3), 513–521, doi:10.1016/j.adolescence.2006.11.008.
- Mauro, R., 2019. Perception of aviation safety. In: Raue, M., Streicher, B., Lerner, E. (Eds.), *Perceived Safety – A Multidisciplinary Perspective*. Heidelberg, Germany, Springer.
- Meder, B., Fleischhut, N., Krumm, N., Waldmann, M.R., 2018. How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty. *Risk Anal.* 38 (3), 549–569, doi:10.1111/risa.13178.
- Mills, B., Reyna, V.F., Estrada, S., 2007. Explaining contradictory relations between risk perception and risk taking. *Psychol. Sci.* 19 (5), 429–433, doi:10.1111/j.1467-9280.2008.02104.x.
- National Highway Safety Administration, 2019. Distracted driving in fatal crashes, 2017. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812700.
- Petermann, I., Weller, G., Schlag, B., 2008. Beitrag des visuellen Eindrucks zur Erklärung des Unfallgeschehens in Landstraßenkurven [Contribution of the visual impression to the explanation of accidents in country road curves]. In: Schade, J., Engeln, A. (Eds.), *Fortschritte der Verkehrspsychologie*. Springer VS Verlag für Sozialwissenschaften, Wiesbaden, Germany, pp. 123–141.

- Peters, E., 2012. Beyond comprehension: the role of numeracy in judgments and decisions. *Curr. Direct. Psychol. Sci.* 21 (1), 31–35, doi:10.1177/0963721411429960.
- Pidgeon, N., Hood, C., Jones, D., Turner, B., Gibson, R., 1992. *Risk Analysis Perception and Management*. Royal Society, London.
- Quadrel, M.J., Fischhoff, B., Davis, W., 1993. Adolescent (in)vulnerability. *Am. Psychol.* 48 (2), 102, doi:10.1037/0003-066x.48.2.102.
- Raue, M., D'Ambrosio, L., Ward, C., Lee, C., Jacquillat, C., Coughlin, J., 2019. The influence of feelings while driving regular cars on the perception and acceptance of self-driving cars. *Risk Anal.* 39 (2), 358–374, doi:10.1111/risa.13267.
- Raue, M., Scholl, S., 2018. The use of heuristics in decision-making under risk and uncertainty. In: Raue, M., Streicher, B., Lerner, E. (Eds.), *Psychological Perspectives on Risk and Risk Analysis - Theory Models and Applications*. Springer, Chum, Switzerland.
- Raue, M., Schneider, E., 2019. Psychological perspectives on perceived safety: zero-risk bias, feelings and learned carelessness. In: Raue, M., Streicher, B., Lerner, E. (Eds.), *Perceived Safety - A Multidisciplinary Perspective*. Heidelberg, Germany, Springer.
- Reyna, V., Weldon, R., McCormick, M., 2015. Educating intuition reducing risky decisions using fuzzy-trace theory. *Curr. Direct. Psychol. Sci.* 24 (5), 392–398, <https://dx.doi.org/10.1177/0963721415588081>.
- Robertson, L.S., Pless, B., 2002. Does risk homeostasis theory have implications for road safety? *BMJ* 324, 1149–1152, doi:10.1136/bmj.324.7346.1149.
- Robinson, D.L., 1996. Head injuries and bicycle helmet laws. *Accid. Anal. Prevent.* 28 (4), 463–475, doi:10.1016/0001-4575(96)00016-4.
- Rottenstreich, Y., Hsee, C., 2001. Money, kisses, and electric shocks: on the affective psychology of risk. *Psychol. Sci.* 12 (3), 185–190.
- Shiv, B., Loewenstein, G., Bechara, A., Damasio, H., Damasio, A.R., 2005. Investment behavior and the negative side of emotion. *Psychol. Sci.* 16 (6), 435–439, doi:10.1111/j.0956-7976.2005.01553.x.
- Simon, H.A., 1955. A Behavioral model of rational choice. *Quart. J. Econ.* 69 (1), 99–118, doi:10.2307/1884852.
- Sjöberg, L., Moen, B.-E., Rundmo, T., 2004. Explaining risk perception. An evaluation of the psychometric paradigm in risk perception research. *Rotunde* 84, 55–76.
- Slovic, P., 1987. Perception of risk. *Science* 236 (4799), 280–285, doi:10.1126/science.3563507.
- Slovic, P., 2016. Understanding perceived risk: 1978-2015. *Environ. Sci. Pol. Sustain. Dev.* 58 (1), 25–29, doi:10.1080/00139157.2016.1112169.
- Slovic, P., Fischhoff, B., Lichtenstein, S., 1978. Accident probabilities and seat belt usage: a psychological perspective. *Accid. Anal. Prevent.* 10 (4), 281–285, doi:10.1016/0001-4575(78)90030-1.
- Slovic, P., Fischhoff, B., Lichtenstein, S., 1979. Rating the risks. *Environ. Sci. Pol. Sust. Dev.* 21, 14–39.
- Slovic, P., Peters, E., 2006. Risk perception and affect. *Curr. Direct. Psychol. Sci.* 15 (6), 322–325, doi:10.1111/j.1467-8721.2006.00461.x.
- Steinberg, L., 2008. A social neuroscience perspective on adolescent risk-taking. *Dev. Rev.* 28, 78–106.
- Sunstein, C.R., 2003. Terrorism and probability neglect. *J. Risk Uncert.* 26 (2), 121–136.
- Svenson, O., 1981. Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica* 47 (2), 143–148, doi:10.1016/0001-6918(81)90005-6.
- Svenson, O., Fischhoff, B., MacGregor, D., 1985. Perceived driving safety and seatbelt usage. *Accid. Anal. Prevent.* 17 (2), 119–133.
- Gigerenzer, G., Fiedler, K., Olsson, H., 2012. Rethinking cognitive biases as environmental consequences. In: Todd, P.M., Gigerenzer, G. (Eds.), *Ecological Rationality. Intelligence in the World.*, doi:10.1093/acprof:oso/9780195315448.003.0025, pp. 81–110.
- Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131.
- van den Bos, K., 2009. Making sense of life: the existential self trying to deal with personal uncertainty. *Psychol. Inq.* 20 (4), 197–217, doi:10.1080/10478400903333411.
- Visschers, V.H., Siegrist, M., 2018. Differences in risk perception between hazards and between individuals. In: Raue, M., Streicher, B., Lerner, E. (Eds.), *Psychological Perspectives on Risk and Risk Analysis - Theory, Models and Applications*. Springer, Chum, Switzerland.
- Ward, C., Raue, M., Lee, C., D'Ambrosio, L., Coughlin, J.F., 2017. Acceptance of automated driving across generations: the role of risk and benefit perception, knowledge, and trust. In: Kurosu, M. (Ed.), *Human-Computer Interaction. User Interface Design, Development and Multimodality, HCI 2017, Lecture Notes in Computer Science*, vol. 10271. Springer, Cham, pp. 254–266.
- Weinstein, N.D., 1980. Unrealistic optimism about future life events. *J. Person. Soc. Psychol.* 39 (5), 806–820.
- Weller, J.A., Shackelford, C., Diekmann, N., Slovic, P., 2013. Possession attachment predicts cell phone use while driving. *Health Psychol.* 32 (4), 379, doi:10.1037/a0029265.
- Wilde, G.J., 1982. The theory of risk homeostasis: implications for safety and health. *Risk Anal.* 2 (4), 209–225, doi:10.1111/j.1539-6924.1982.tb01384.x.
- Wilde, G.J., 2002. Does risk homeostasis theory have implications for road safety? *BMJ* 324, 1149–1152, doi:10.1136/bmj.324.7346.1149.
- Zajonc, R., 1980. Feeling and thinking: preferences need no inferences. *Am. Psychol.* 35 (2), 151–175, doi:10.1037/0003-066x.35.2.151.
- Zhou, R., Yu, M., Wang, X., 2016. Why do drivers use mobile phones while driving? The contribution of compensatory beliefs. *PLoS One* 11 (8), e0160288, doi:10.1371/journal.pone.0160288.

Further Reading

- Raue, M., Streicher, B., Lerner, E., 2019. *Perceived Safety - A Multidisciplinary Perspective*. Springer, Heidelberg, Germany <https://dx.doi.org/10.1007/978-3-030-11456-5>.
- Raue, M., Lerner, E., Streicher, B., 2018. *Psychological Perspectives on Risk and Risk Analysis - Theory, Models and Applications*. Springer, Chum, Switzerland <https://doi.org/10.1007/978-3-319-92478-6>.
- Slovic, P., 2015. Understanding Perceived Risk: 1978-2015. *Environ. Sci. Policy Sustain. Dev.* 58 (1), 25–29, <https://dx.doi.org/10.1080/00139157.2016.1112169>.
- Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131.

Road Diets

Robert B. Noland, Alan M. Voorhees Transportation Center, Edward J. Bloustein School of Planning and Public Policy, Rutgers University, New Brunswick, NJ, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	500
History of Road Diets	500
Typical Road Diet Configuration	500
Other Road Diet Configurations	502
Complete Streets	503
Benefits of Road Diets	503
Safety Benefits	503
Operational Effects	503
Livability Benefits	504
Ease of Implementation	504
Costs	504
Political Context and Impediments	505
Criteria for Road Diets and Assessment Procedures	505
Summary	506
References	506

Introduction

The objective of a road diet is to reduce excess lane capacity from roads that were overbuilt. The goal is improved safety, but there can also be operational benefits and other community benefits. The typical road diet conversion, at least in North America, is the reduction of a four-lane road (two-lanes in both directions) to two-travel lanes and a two-way left-turn lane (TWLTL) in the median. This has the benefit of making left-turns across the traffic stream easier and reduces blocking of traffic as vehicles wait to complete a left-turn. If the road diet occurs in an urban area, or rural area with pedestrian activity, the TWLTL may be interrupted with raised medians at every crosswalk. Average speeds in the travel lane are often reduced, as a lead vehicle will control the speed of following vehicles; weaving is also reduced. Drivers also perceive a narrower road, which reduces speeds (Burden and Lagerwey, 1999; Knapp et al., 2014).

The reduction in the number of traffic lanes allows pedestrians to more easily cross the road, as their exposure to moving traffic is less. Depending on the design, many road diets incorporate a bicycle lane or track or include a larger shoulder width that bicycles can use. There are safety benefits associated with bicycles having their own through-lane and this can improve vehicle flow as vehicles do not need to change lanes to pass a bicycle.

History of Road Diets

An obvious question is why are there so many roads with four lanes when fewer lanes could easily accommodate the traffic flow? In the United States, much of the expansion of the road network took place in an era of rapid growth in vehicle ownership and burgeoning budgets for highway development. Guidelines for roads with TWLTLs did not exist and most transportation agencies simply modified two-lane roads to four lanes, with no consideration of the impacts on communities.

While the earliest implementations of road diets can be traced back to the 1970s, these gained in popularity in the 1990s. This was largely due to advocates, such as Dan Burden, explaining the benefits to local communities throughout the country. More recently, the Federal Highway Administration (FHWA) has recognized that road diets can be effective as a safety countermeasure. While there have been some studies that have evaluated the crash reduction benefits, these are limited. In general, the guidance suggests that crash reductions ranging from 19% to 47% are feasible, depending on traffic volume and other details of the local area (Thomas, 2013). Many states and local communities are now developing policy and evaluating possible road diet conversions.

Typical Road Diet Configuration

The typical configuration for a road diet in North America is a road with two travel lanes and one two-way left-turn lane (TWLTL) in the middle of the road (Knapp et al., 2014). A schematic of this is shown in Fig. 1. Three examples of actual road diets are shown in Fig. 2 from three different locations. In these cases, the treatment of the shoulder varies. In panel-a there is a wide shoulder that can

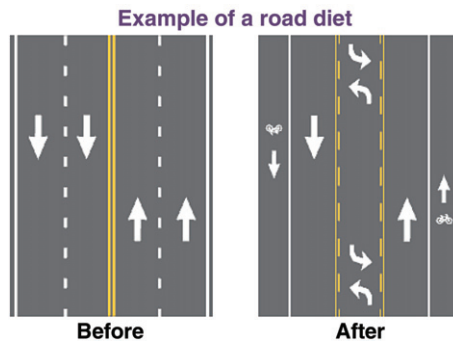


Figure 1 Schematic of a typical road diet. *Source: FHWA.*



Figure 2 Typical road diet configurations. (A) Electric Ave, Lewiston, PA; (B) Grand River Ave., East Lansing, MI; (C) N 45th St., Seattle, WA. *Source: Google Maps.*

easily function as a bicycle lane; panel-b likewise has a shoulder, but not as wide as the example in panel-a. The last example shows a road diet in an urban setting; the shoulder is used for parking and curb extensions were built at some pedestrian crossings. All of these examples were at one time four-lane roads without these amenities and with higher speed traffic.

In **Fig. 3**, an example is shown of a road without a road diet, a typical four-lane arterial in a suburban area. The figure on the right is modified to show what the road would look like after a conversion to two-lanes and a TWLTL with bicycle lanes along the side.



Figure 3 Livingston Ave, New Brunswick, NJ. Current configuration and simulated photo showing potential road diet. *Source: Author.*

Other Road Diet Configurations

While the classic and most common road diet involves reducing a four-lane road to three-lanes, other configurations are possible. These can also involve reductions in lane-width and changes in shoulder widths. For example, a four-lane road could be reconfigured with a TWLTL and maintain four-lanes of moving traffic. This is done by reducing lane and shoulder widths. While perhaps not having the safety benefits of a reduction to three-lanes, there can be operational benefits of having the turning lane. Other examples could be reducing the lane width of a current three-lane configuration to provide room for a larger shoulder or bicycle lanes.

Another design option is to introduce a hard median and converting from four-lanes to two-lanes. Since a hard median restricts turning movements, roundabouts (traffic circles) can facilitate U-turns (and also create safer road junctions). An example of this is a conversion along La Jolla Boulevard in San Diego, CA. In this case a large arterial with five lanes (including left-hand turn lanes) was reduced to two-lanes with five roundabouts along a commercial corridor (**Fig. 4**). This street handles about 22,000 vehicles per day. An additional benefit of this design was a large reduction in pedestrian crossing times, which actually reduced vehicle waiting times at crosswalks ([Logan, 2017](#)).

Road diets is largely a North American term, representing efforts to ameliorate past over-building of roads. In Europe there have been substantial efforts to reallocate road space to pedestrians and cyclists. These have ranged from pedestrian-only zones in many European cities, to extensive bicycle networks (usually separated from vehicle traffic). Traffic calming largely originated in Europe and some road diets follow similar principles of using design to reduce vehicle speeds and improve safety for all users.



Figure 4 La Jolla Blvd., Bird Rock, San Diego, CA. *Source: Google Maps.*

Complete Streets

Road diet implementations are also a method for developing *Complete Streets*. These are broadly defined as streets designed to serve not just motorized vehicles, but all modes—pedestrians, cyclists, and transit users (New Jersey Dept. of Transportation, 2017). In addition, a complete street should allow safe use by children, disabled users, and older population groups. A road diet implementation that also serves the needs of these users would be consistent with the philosophy of reallocating street space so that all users can safely access the transportation system. To make a complete street safe enough for all users, motor vehicle speeds need to be reduced especially at pedestrian crossings.

In North America, there have been a proliferation of complete street policies implemented at the state, regional, and local level (Gregg and Hess, 2019). Traditionally when new roads are constructed or existing roads undergo a major renovation, design guidelines focus on only serving motor vehicle needs. Any attempt to, for example, add a bicycle lane or create safer pedestrian crossing points is seen as an imposition on the flow of vehicles, and requires special exemptions. Complete Streets policies switch this logic around. Most of those policies implemented require road and highway agencies to justify not providing access for all users; the default policy under Complete Streets is that now the street will serve all users in a safe and equitable manner. As these policies have spread it allows local areas to implement road diets as part of their Complete Streets policy.

Benefits of Road Diets

In assessing whether to convert an existing four-lane road, the primary motivation is normally to achieve safety benefits. Other benefits can also occur, including improved vehicle flow, community development and livability, and even reductions in vehicle emissions.

Safety Benefits

The safety benefits attributable to road diets derive from both speed reductions and fewer potential conflict points for vehicles. Speed reductions can occur because lead vehicles control the speed of following vehicles. Weaving and passing movements, while still possible, are much less likely to occur as vehicles stay in a single line. Those vehicles making left turns or turning left into the road have a designated turning lane, so as to not block through traffic and reduce rear-end and side-swipe collisions (Thomas, 2013).

Average speeds are typically lower by up to 5 mph (8 km/h) after implementation. Excess speeding is also reduced. Speed differentials between lanes and traffic moving in the other direction are reduced; the median provides an enlarged buffer between opposing traffic streams also improving safety.

Pedestrian safety is improved, both from reductions in speed, but more importantly from reduced crossing distances. If curb extensions are part of the road diet, this allows pedestrians to more safely cross the street. A bicycle lane or wider shoulder, with or without parking, can provide added buffer space between vehicles and pedestrians on the sidewalk. If a bicycle lane or wide shoulder is part of the road diet, then this provides an additional safety benefit for bicyclists and may increase the likelihood of cyclists using the road. In much of the United States, a stripe separating a bicycle lane from an automobile travel lane is deemed to give acceptable safety. In much of Europe, striping is not considered enough, but a curb or pollards with reflectors are installed as well. Some cities, such as New York City, are placing bicycle lanes between the curb and a line of parked cars (though these are not necessarily associated with road diets).

While safety benefits of road diets are generally recognized, very few studies examined impacts before and after conversions. A review of studies concluded that crash reductions of between 19% and 47% are expected, depending on road type and the location of the road (Thomas, 2013). The reduction of 19% was associated with roads in large urban areas, while 47% reductions were associated with rural roads in small urban areas. It is more than likely that these estimates are inaccurate for any specific case, given the dearth of studies. Thus, while road diet conversions will likely reduce crashes, there is great uncertainty on the magnitude of the reduction. This may make it difficult for decision makers to promote road diet conversions.

It is clear that pedestrians will benefit greatly from road diets when speeds are reduced. A study of a large number of crashes in the United States found that the risk of severe injury increases with speed (Tefft, 2011). For example, a reduction of 5 mph (8 km/h) would reduce severe pedestrian injuries by about 16%.

Operational Effects

One of the main concerns expressed by those opposed to road diet conversions is that any reduction in capacity will lead to congestion. While road diets effectively take a lane of traffic away, there is actually minimal operational impact. If there are a substantial number of left-turn (cross-traffic) movements along a four-lane road it effectively is already operating with only one lane in each direction as turning vehicles block one of the through lanes. Most road diets do not affect operations up to usage of between 20,000 and 25,000 vehicles per day (Stamatiadis et al., 2011; Knapp et al., 2014). At levels greater than this, left-turning movements would likely be forbidden so that all four lanes can be effectively used. Many studies have found no increase in vehicle delay, or even reductions in vehicle delay, with road diet conversions at levels below 25,000 vehicles per day.



Figure 5 “Fire lane” marked along median, Canal Pointe Blvd., West Windsor, NJ. Conversion completed in 2017. *Source: Author.*

One issue is that if there are frequent traffic signals, a road diet may result in excess queuing at signals backing up into the area of adjacent signals. This can be mitigated by coordinating traffic signals at the posted speed limit. It typically is not an issue when peak-hour traffic volumes are below about 1750 vehicles per hour (or 875 in each direction). Traffic signals can also be replaced with single-lane roundabouts, as illustrated in [Fig. 4](#).

Emergency response vehicles also see a benefit from road diets using two-way turn lanes ([FHWA, 2017](#)). While often police and fire services oppose road diets, in reality the median lane provides an unobstructed travel lane when needed by emergency responders. Some road diet implementations actually mark the median lane as a “fire lane” ([Fig. 5](#)). The conversion shown in [Fig. 5](#) was previously a four-lane road and did not have a shoulder.

Livability Benefits

Road diets are seen as a way to improve the livability and quality of life for those living on or near the converted street ([Knapp et al., 2014](#)). This can happen when speeds are reduced and increased buffers are provided between the sidewalk and the traffic stream. Both can create a more pleasant and safer walking environment leading to more people walking along the street. Vehicle noise is also reduced as speeds are lower. Air quality improvements have been found to occur, though this may matter less as vehicles electrify ([Shu et al., 2014](#)). A smoother traffic flow reduces stop and go driving that may lead to hard accelerations and decelerations, which are associated with higher emissions, in particular particulates from brake pads and tires. In commercial districts, lower speeds, easier street crossings and a more pleasant environment can also spur increased economic activity. All these benefits can improve public health especially in traditionally underserved communities.

Ease of Implementation

Most road projects undergo large studies and detailed engineering assessments before a decision is made to build the project. These assessments are time-consuming and expensive. Road diets, on the other hand, can be implemented quickly and cheaply, assuming there is political will to improve safety. In most cases, all that is needed is the removal of existing pavement markings and replacement with new traffic stripes to delineate the lanes. If a resurfacing of the road is scheduled, the new stripes will add minimal costs to the project and maybe none at all ([FHWA, 2016](#)). If for some reason the restriping creates unexpected problems, the road diet can be quickly and cheaply reversed. This falls within what are considered tactical urbanism approaches to improve quality of urban and suburban areas. Testing of changes to street design, via restriping, can demonstrate the effectiveness of the changes without major investment.

Costs

The Federal Highway Administration provides a rough estimate of how much a four-lane to three-lane road diet conversion will cost ([FHWA, 2016](#)). Assuming that a bicycle lane is included, the total costs of removing and restriping comes to about \$100,000 per 5000 ft (~1500 m) of road. If the road surface is repaved, the addition of bicycle lanes adds between \$18,000 (simple striping) to \$46,000 (with signage and bike lane road markings) more to project costs. These estimates are based on 2015 USD estimates. Since roads are periodically restriped and repaved, additional costs will be minor as they would be part of the planned project.

The reallocation of road space achieved with a road diet is cheaper than alternative approaches that might require expanding the width of the right-of-way. For example, if sidewalks and bicycle lanes are added to an existing road without using the existing right-of-way, then additional costs for land acquisition and storm water drainage may be incurred. In many cases, this may not be feasible if space is constrained in an urban or suburban area. Adding turning lanes to a four-lane road may also require widening the right-of-way at selected intersections, adding substantially to project costs.

Political Context and Impediments

Even though road diets provide many benefits, there is often political and public opposition to converting existing four-lane roads, even when there are demonstrated safety issues. Much of this opposition is likely due to concerns that traffic congestion will increase since two travel lanes are being eliminated. As discussed earlier, operational effects are usually minor, unless the road exceeds peak-hour vehicle travel of about 1750 vehicles. Some may also complain that speeding is made more difficult as they do not believe it poses a risk; they also oppose having the freedom to pass and weave around slower moving vehicles. Political leaders and traffic engineers are also loath to experiment with untested approaches in their communities.

One solution to opposition is to propose a pilot restriping or “demonstration project.” As noted above, this is a cheap intervention that is easily reversible. If the community sees the benefits and that there is no change in congestion levels, then implementation is more likely to succeed.

Existing requirements and guidelines may also be an impediment to a fast implementation, not least in the United States. Many states may require that any change to a road layout must undergo a detailed engineering analysis, in particular to determine impacts on traffic Level of Service (LOS). This type of analysis usually requires contracting work out to an engineering firm with expertise in applying Highway Capacity Manual procedures and simulation models to evaluate changes in LOS at each intersection along the proposed road diet (TRB, 2010). Often, these studies can cost more than the actual restriping that is called for! Additional delay in project implementation also means that the safety benefits of the road diet are deferred. The United States is not alone as in many other countries mobility trumps safety.

Sweden’s Vision Zero program is an exception to this and has dramatically changed public perceptions of mobility versus safety in the Swedish road transport system. As noted in this Encyclopedia’s entry on Vision Zero, “Mobility becomes a function of safety rather than the other way around” (Belin, in press). A comparison can be made with the aviation network, “if we discover safety problems in our aviation system that might threaten people’s lives we stop the air traffic until we have fixed the problem. Luckily, we do not have to stop the road traffic in order to achieve safe mobility, since it, in most cases, is enough to reduce speeds” (Belin, in press).

Another source of delay is over-design of the road diet. For example, pedestrian curb extensions can be added and provide additional safety benefits. These obviously add costs, including additional delay in designing the infrastructure changes, and may require further changes to accommodate storm-water drainage. New signaling systems may also be required and might help with any signal retimings needed to avoid operational problems. While these might be beneficial improvements, they should not delay the restriping. The key message is that road diets are quick and cheap to implement.

Criteria for Road Diets and Assessment Procedures

In the United States, many state, county, and city agencies have developed various criteria for considering road diets. These typically include traffic volume (not exceeding 25,000 vehicles per day), speed of vehicles on the road, and number of vehicle crashes and particularly those involving pedestrians (Knapp et al., 2014). In most cases, the road is identified as having a safety problem from excess crashes and potential conflicts with pedestrians, either from inadequate space for pedestrians or speeding traffic (in excess of posted speed limits). While these criteria suggest the need for an intervention, other criteria can impact the effectiveness of a road diet. These include the number of driveways and cross-streets, signal spacing, freight traffic volumes, and bus routes along the street. A route with a large number of driveways and cross-streets with left-turn movements would likely see improved traffic flow from a road diet implementation. Buses without a turnout lane, may delay other traffic when only one travel lane is available. Signals can be retimed to mitigate any traffic problems.

One simple way to assess the benefits of a road diet is to evaluate the costs and benefits. The primary benefit is a reduction in crashes and in particular injuries and fatalities. As noted previously, conversion costs mainly consist of restriping and are trivial. If travel times and level of service in the corridor increase, then there is a cost associated with increased travel delay. Both crash costs and travel delay can be monetized to produce a cost-benefit analysis, discounted over a 20-year time period. The USDOT provides guidance on benefit-cost analysis including valuations of statistical life and valuations of travel time (Office of the Secretary, USDOT, 2018).

There are very few studies of the safety benefits of road diets, with the rule of thumb being between 19% and 47% reductions in crashes (Thomas, 2013). However, in reality, it is not possible to predict the actual crash reduction from any specific intervention. The relevant consideration is what level of crash reduction is needed to justify the costs of a road diet conversion? While many road diets will have little to no impact on corridor travel times, due to the road being a de-facto three-lane road, decision makers may still believe there will be congestion impacts. Thus, one approach is to assume large travel time increases and determine what crash

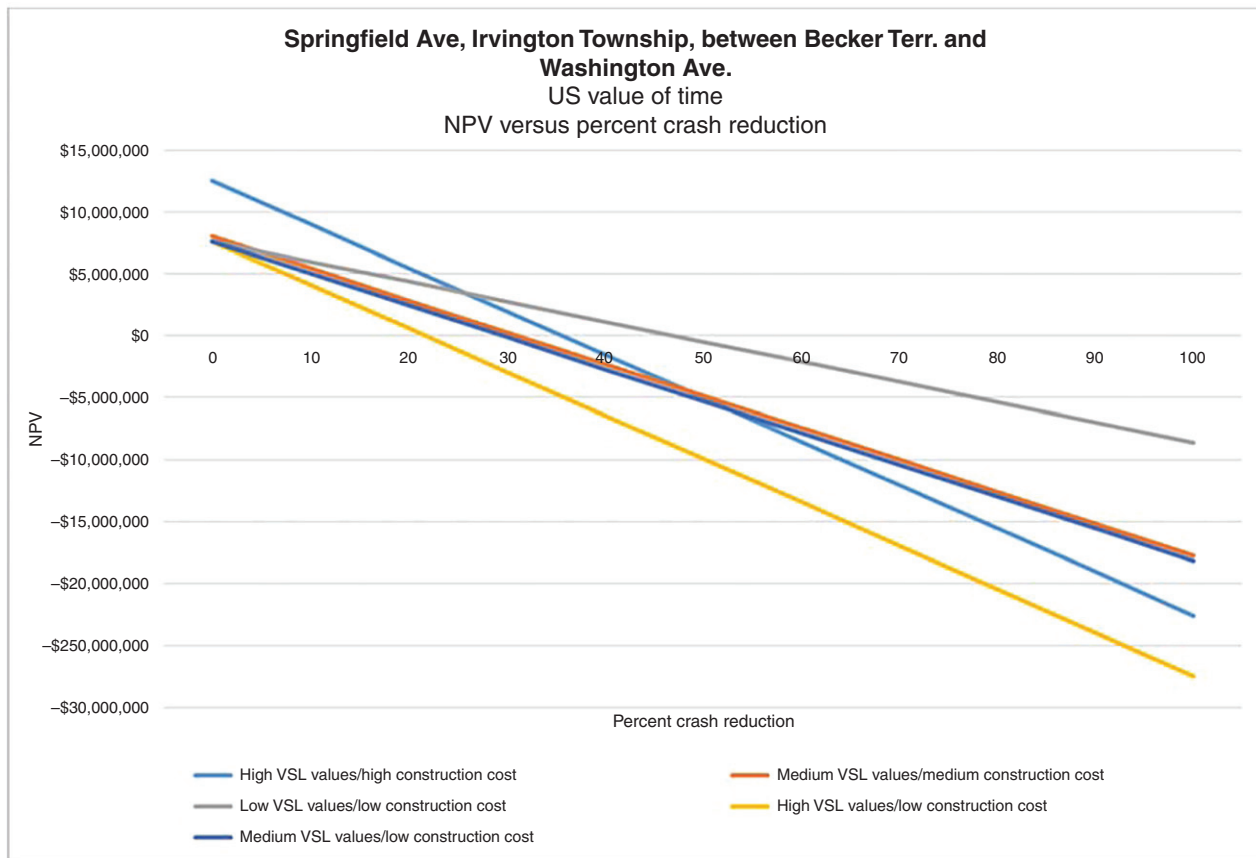


Figure 6 Graphical analysis of trade-off scenarios between travel time costs and crash reduction, showing break-even NPV values. This analysis shows that benefits exceed costs when crash reduction is between about 22% and 50%. Source: [Noland, 2018](#)

reduction is needed to have at least a zero net benefit (or benefit-cost ratio equal to 1). In some cases, the objective of the project may be to reduce speeds and speed variance, and thus, this would increase travel times. A sample graphic showing a benefit-cost trade off analysis is shown in [Fig. 6](#). This analysis only examines safety benefits and excludes many of the other benefits of a road diet, including better pedestrian access, bicycle lanes, health and livability of the area near the road diet; it also assumes a large increase in travel times from a reduction in speeds of 5 mph (8 km/h) based on liberal assumptions on average traffic flow ([Noland, 2018](#)). In this example, depending on the assumptions of the analysis, a crash reduction of anywhere from about 22% to 50% results in benefits equal to costs.

Summary

To summarize, road diets are an effective and cheap safety countermeasure that involves reducing the number of motor-vehicle travel lanes. They can be implemented quickly if agencies forego much of the detailed traffic flow analysis often required for changes in road design. There may be signalization changes needed, but these can be adjusted after the road diet is implemented. Road diets are easily reversible if not deemed successful at reducing crashes or if they introduce other problems. Any resurfacing project is an opportunity to implement a road diet. The benefits of road diets also go beyond just expected crash reductions. These include livability benefits that result in more walkable and pleasant conditions along the road, noise reductions, emissions reductions, and potential increases in commercial activity.

References

- Belin, M-A., In press. The Swedish Vision Zero – a policy innovation, *Encyclopedia of Transport*, Elsevier.
- Burden, D., Lagerwey, P., 1999. Road Diets: Fixing the Big Roads. Retrieved from Walkable Communities, Inc. Available from: https://nacto.org/wp-content/uploads/2015/04/road_diets_fixing_big_roads_burden.pdf.
- FHWA, 2016. How Much Does a Road Diet Cost? FHWA-SA-16-100. Available from: https://safety.fhwa.dot.gov/road_diets/resources/fhwasa16100/.

- FHWA, 2017. Road Diets and Emergency Response: Friends, Not Foes. Available from: https://safety.fhwa.dot.gov/road_diets/resources/pdf/fhwasa17020.pdf.
- Gregg, K., Hess, P., 2019. Complete streets at the municipal level: a review of American municipal complete street policy. *Int. J. Sustain. Transp.* 13 (6), 407–418., doi:10.1080/15568318.2018.1476995.
- Logan, J., 2017. Roundabout apostle comes full circle, revisits Bird Rock, The San Diego Union-Tribune (Feb 16 2017), Available from: <https://www.sandiegouniontribune.com/news/columnists/logan-jenkins/sd-me-20170217-story.html>.
- Knapp, Keith, Brian Chandler, Jennifer Atkinson, Thomas Welch, Heather Rigdon, Richard Retting, Stacey Meekins, Eric Widstrand, and Richard J. Porter (2014). Road diet informational guide. No. FHWA-SA-14-028. United States. Federal Highway Administration. Office of Safety, 2014. Available from: https://safety.fhwa.dot.gov/road_diets/guidance/info_guide/.
- New Jersey Department of Transportation, 2017. prepared by WSP | Parsons Brinckerhoff, 2017 State of New Jersey Complete Streets Design Guide. Available from: <http://njbikeped.org/wp-content/uploads/2017/05/Complete-Streets-Design-Guide.pdf>.
- Noland, Robert B., 2018. A streamlined method for evaluating potential road diets, New Jersey Bicycle and Pedestrian Resource Center, Rutgers University, prepared for New Jersey Department of Transportation. Available from: <http://njbikeped.org/portfolio/a-streamlined-method-for-evaluating-potential-road-diets-2018/>.
- Office of the Secretary, US Department of Transportation, 2018. Benefit-Cost Analysis Guidance for Discretionary Grant Programs. Available from: <https://www.transportation.gov/office-policy/transportation-policy/benefit-cost-analysis-guidance-2017>.
- Shu, S., Quiros, D.C., Wang, R., Zhu, Y., 2014. Changes of street use and on-road air quality before and after complete street retrofit: An exploratory case study in Santa Monica California. *Transp. Res. Part D: Transp. Environ.* 32, 387–396.
- Stamatiadis, N., Kirk, A., Wang, C., Cull, A., Agarwal, N., 2011). Guidelines for Road Diet Conversions. Kentucky Transportation Center. KTC-11-XX/SPR-415-11-1F. Available from: <http://dx.doi.org/10.13023/KTC.RR.1; 2011.19>.
- Tefft, B.C., 2011. Impact Speed and a Pedestrian's Risk of Severe Injury or Death. AAA Foundation of Traffic Safety. Available from: <https://aaaafoundation.org/impact-speed-pedestrians-risk-severe-injury-death/>.
- Thomas, Libby, 2013. Road Diet Conversions: A Synthesis of Safety Research, Pedestrian and Bicycle Information Center, White Paper Series, FHWA DTFH61-11-H-00024. Available from: https://www.researchgate.net/publication/274383847_Road_Diets_A_Synthesis_of_Safety_Research.
- Transportation Research Board (2010). Highway Capacity Manual, National Research Council, Washington, DC. <http://www.trb.org/Main/Blurbs/164718.aspx>.

Road Safety Audits

Xiao Qin, Civil and Environmental Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

© 2021 Elsevier Ltd. All rights reserved.

Background	508
The History and Evolution of Road Safety Audits	508
Benefits and Costs of Road Safety Audits	509
Steps to Perform Road Safety Audits	509
Road Safety Audits in a Road Project's Life Cycle	510
Road Safety Audit Tools	510
Case Studies	511
Transit Access Safety Audit	511
Active Work Zone Safety Audit	511
Road Weather Safety Audit	511
Conclusion	512
Biography	512
References	512
Further Reading	512

Background

Every year, transportation professionals face the increasing demand for safer roads. Various strategies have been applied to reduce traffic conflicts, injuries, and fatalities. One of those strategies, hot-spot management, has been applied extensively. A hot spot is a colloquial term for roadway segments or intersections that show a significantly higher rate or frequency of fatalities and injuries or overall crashes than others in a cohort. Even though a hot spot can be identified through a clear and treatable crash pattern, the specific countermeasures may act only as a "band-aid" solution and not address the broader safety issue. Transportation professionals should migrate from reactive hot-spot management to a more proactive approach that fixes safety deficiencies before crashes happen. One effective proactive approach is the safety audit. A road safety audit (RSA) is a formal safety examination of a future road or traffic project or an existing road by an independent, qualified, and multidisciplinary team who reports on the project's crash potential and makes recommendations for safety treatments. It is worth noting that the concept of safety audit has been applied to other transportation modes such as aviation safety audits, marine (or maritime) safety audits, and railway safety audits. However, the primary purpose of RSA in other modes is to assess the degree of compliance with safety regulatory requirements and standards.

RSAs can be conducted in any phase of a road project's life cycle, which includes planning, preliminary design, detailed design, work zone traffic control plan, project construction, preopening, and postconstruction. Although an audit's scope, purpose, focus, and complexity will vary depending on the phase, the essential elements of an RSA do not change. RSAs need to be performed by a team that is (1) independent of the project to ensure an impartial assessment, (2) multidisciplinary to ensure the audit is complete and comprehensive, and (3) capable of giving formal and constructive recommendations. This paper covers the fundamentals of RSAs, including definitions and terminologies, steps and procedures taken to perform RSAs, and the costs and benefits of conducting an RSA. Readers will also learn about the tools used during a safety audit and be introduced to case studies involving typical highway projects, pedestrian and bicyclist facilities, work zone management, and road weather safety. The content could be used as a reference guide for those interested in the concept of an RSA and as an introductory source to support further exploration.

The History and Evolution of Road Safety Audits

The concept of an RSA originated in the United Kingdom in the 1980s when some newly completed roads were experiencing a high number of crashes and injury severities that could have been prevented. By 1991, it became a national mandate in the United Kingdom for all national trunk road and freeway projects to undergo an RSA as long as resources allow. The safety audit philosophy was introduced in Australia and rapidly expanded to other countries such as Canada, Denmark, the Netherlands, Germany, Switzerland, Sweden, Norway, and South Africa. Developing countries such as Malaysia, Singapore, Bangladesh, India, Mozambique, and the United Arab Emirates have also introduced RSAs.

The first RSAs in the United States were implemented at the design stages of projects in Pennsylvania around 1997 after the Federal Highway Administration (FHWA) visited Australia and New Zealand to review their processes. The FHWA conducted 14

pilot projects in different states around the nation. Other states and local agencies followed suit and developed their own procedures and methods to carry out safety audits in pavement overlay programs, pavement rehabilitation, restoration, and resurfacing (3R) projects, and even for megaprojects (e.g., the Marquette Interchange reconstruction project in Milwaukee, Wisconsin, United States). The boom of RSA practices led to the publication of the National Cooperative Highway Research Program (NCHRP) Synthesis 321: *Roadway Safety Tools for Local Agencies: A Synthesis of Highway Practice* in 2003 (Wilson, 2003), and later Synthesis 336: *Road Safety Audits: A Synthesis of Highway Practice* in 2004 (Wilson and Lipinski, 2004) (<http://trb.org/bookstore/>). In 2006, FHWA established the *FHWA Road Safety Audit Guidelines* to promote RSAs and their integration into existing safety management systems and engineering practices (Ward, 2006). Relevant training materials and courses have been developed by the FHWA National Highway Institute (NHI) and are available through <http://safety.fhwa.dot.gov/rsa>.

The 2012 report *Road Safety Audits: An Evaluation of RSA Programs and Projects* demonstrates the effectiveness of RSAs by outlining nine RSA programs and five RSA projects in the United States funded by FHWA, as well as numerous case studies from state and local agencies and nonprofit organizations such as the Institute of Transportation Engineers and the American Automobile Association (Nabors et al., 2012). Australia's 2009 edition of the *Guide to Road Safety Part 6: Road Safety Audit* is currently being updated to include Safe System principles and account for the rapid development of predictive safety models (Austroads, 2019a; Austroads, 2019b).

Benefits and Costs of Road Safety Audits

RSAs are beneficial because it is better to be proactive and prevent crashes and fatalities rather than react to them. The long-term benefits of RSAs are substantial. From the road owners' perspective, RSAs save the potential throwaway and construction costs by correcting safety deficiencies before roads are in service. Safe roadway designs and operations also mean less potential for lawsuits. More importantly, adopting RSAs shows an agency's strong commitment to safety. Raising safety awareness at all levels is invaluable, as it leads to better decision-making that considers multimodality and human factors.

Sometimes the benefit of an RSA can be quantified by converting reduced crash frequency or severities to a monetarized value if the cost effectiveness of a safety countermeasure is known. The process of performing a benefit-cost analysis is detailed in *Chapter 7 Economic Appraisal* of the *Highway Safety Manual* (AASHTO, 2010). According to the FHWA, one of the most cited studies of the benefits of RSAs was conducted in Surrey County, United Kingdom. The study found that the average annual fatal and injury crash frequency at the audited sites decreased by 1.25 crashes per year versus only 0.26 crashes per year at unaudited sites; this suggests that RSAs are 5 times more effective in reducing fatal and injury crashes. The FHWA Road Safety Audit website (<https://safety.fhwa.dot.gov/rsa/resources/>) has published the economic analysis results for several other studies.

The cost of an RSA includes RSA team costs, design team and project owner costs, and implementation costs. The cost of RSAs varies greatly by audit project size, scope, and complexity. Typically, conducting RSAs and implementing the recommended safety countermeasures in design are estimated to be 5% of the overall engineering design fees. However, the cost ultimately depends on agencies' flexibility and creativity in integrating RSAs within existing programs and projects.

Steps to Perform Road Safety Audits

Compelling case studies and evidences have shown that timely RSAs provide cost-effective solutions to improving road safety. An RSA usually follows the eight-step audit process which has become the industry standard (Ward, 2006):

- Step 1: Identify the project to be audited. In this step, the public agency is responsible for identifying the project or existing road to be audited.
- Step 2: Select the RSA team. In this step, the public agency is responsible for selecting an independent, qualified, and multidisciplinary team of experts suitable for the project.
- Step 3: Conduct a preaudit meeting to review project information. In this step, a meeting is called between the project owner, the design team, and the audit team to discuss the context and scope of the RSA and review all available project information.
- Step 4: Perform field reviews under various conditions. In this step, the RSA team reviews information obtained from Step 3 to gain insight into the project and safety concerns and prepare for the field visit. The field visit helps the team gain further insight into the project and further verify/identify areas of safety concern.
- Step 5: Conduct an audit analysis and prepare the findings. In this step, the RSA team identifies, analyzes, and prioritizes safety concerns and locations and summarizes the findings that include recommendations for the formal RSA report.
- Step 6: Present audit findings to the project owner/design team. In this step, the RSA team orally reports the key findings to the project owner and design team.
- Step 7: Prepare the formal response. In this step, the project owner and/or design team provides a formal response to the RSA team with regard to each safety issue listed in the report and explains why some of the RSA suggestions could not be implemented.
- Step 8: Incorporate findings into the project when appropriate. This final step ensures that the corrective measures outlined in the response report are completed as described and in the time frame documented.

Road Safety Audits in a Road Project's Life Cycle

Although RSAs can be applied at every stage during the life cycle of a road project, applying audits earlier on in the planning stage and preliminary design phase offers the greatest opportunities to rectify any safety deficiencies and avoid throwaway costs. The cost for correcting or retrofitting will be higher as the project advances into later stages; however, RSAs that occur later in the process can also be beneficial as unique and new safety problems may emerge (Ward, 2006).

Three types of RSAs may be conducted during the preconstruction phase: planning or feasibility, preliminary design, and detailed design. An RSA during the planning stage focuses on items such as: vehicles to be accommodated with the proposed project, design speed, access to surrounding developments and connection with other routes and the number of spacing of intersections.

The preliminary stage audit can be performed when 30%–40% of the design is finished, and includes topics such as drainage, landscaping, accesses to developments, accommodation for maintenance and emergency vehicles, and parking provisions. It also includes design issues such as cross-section elements, alignment, sight distances, and intersection layout. The preliminary audit also considers safety issues specific to special road users (pedestrians, bicyclists, and public transport). Lighting, signage, and pavement markings are also considered during this audit.

Lastly, the detailed design audit can be conducted as soon as approximately 60%–80% of the design is finalized. The detailed design audit is critical as it is the last opportunity to review the design before construction begins. Depending on the type of roadway, the detailed design covers design features related to drainage, pavement skid resistance, geometry, medians, shoulders, barriers and guardrails, readability, traffic signals, signs, markings, special road users, and parking provisions.

Construction phase RSAs can be conducted during the construction preparation period, during the actual construction period, and during the preopening time period. The RSA team can conduct three types of RSAs using project as-let or as-build plans: work zone traffic control plan RSAs, changes in design during construction RSAs, and preopening RSAs. When performing a work zone traffic control plan RSA, the audit team should ensure that safety is adequately considered in the maintenance of traffic plan and the work zone traffic control plan. Safety items to consider for this RSA include the design of temporary roadways and transition areas, appropriateness of temporal traffic control devices, information consistency between temporary and permanent traffic controls, and possible impacts for all road users, including pedestrians, bicyclists, and the disabled. The RSA taking place at the changes in design during construction stage requires the team to envision situations where construction may lead to unforeseeable safety problems that may not be obvious. Lastly, the preopening RSA is the last chance for the team to consider any safety aspects of the design before the project is opened to the public. The team should perform a thorough field review to identify any necessary modifications related to illumination, traffic signs, pavement markings and delineation, and fixed object hazards.

Postconstruction phase RSAs take place on existing roads. The RSA team should observe all interactions between road users and roadway facilities during the day and night, note the conflicts and report any safety concerns. Crash data may become available during postconstruction RSAs, but it should not be the sole reason for the RSA. Crash data can be used to validate safety concerns identified through the RSA and to preclude potential safety issues.

Road Safety Audit Tools

The most commonly used RSA tool is the checklist, or, more appropriately, the prompt list. The word “prompt” underscores that the role of the list is to assist (not replace) knowledge and experience. Prompt lists pose questions intended to guide auditors to think “what can be done differently as a result of an RSA?” The purpose of an RSA prompt list is to help auditors identify any potential safety issues rather than check only common design standards.

General prompt lists can be organized by RSA stage in the project life cycle. Custom prompt lists have been developed for different subjects such as pedestrians and bicyclists, transit users, work zones, or specific crash types (i.e., wrong-way driving). Prompt lists can be used when reviewing project data, conducting site visits, conducting analyses, and writing RSA reports.

Prompt lists are migrating from paper forms to computerized applications. The Road Safety Audit Toolkit, developed from the Austroads checklists, is an online computer program that facilitates the audit process. The software assists auditors by prompting them with questions so that they can efficiently organize their findings in a structured manner. The toolkit automatically generates RSA reports. The software is free for download at www.rsatoolkit.com.au. The Austroads Road Safety Engineering Toolkit, which provides solutions to safety concerns, can be used in conjunction with the Road Safety Audit Toolkit. Treatments can be selected by crash patterns, or by specific category of road safety deficiencies such as signalized intersections. The engineering toolkit can be accessed at www.engtoolkit.com.au.

It is worth emphasizing that RSAs are not auditing design standards, so the checklists cannot substitute a design manual; but they can be used by roadway designers to help proactively identify safety issues during the design process. The Interactive Highway Safety Design Model (IHSDM) has been incorporated into the audit process for several studies in order to evaluate the safety and operations of a highway project. As a primary design tool, IHSDM can be useful in providing a quantitative assessment of the safety implications of various design alternatives. IHSDM is available for free download at www.ihsdm.org.

Case Studies

Although it is impossible to list every type of case study, the following cases represent three very different applications: pedestrian and bicyclist safety, work zone management, and road weather safety.

Transit Access Safety Audit

The purpose of the study was to conduct an RSA for the existing transit access roads at Sun Trans and the City of Tucson, Arizona, United States. The study site is located in an urban area with a high level of multimodal interactions among passenger vehicles, pedestrians, bicyclists, and transit vehicles. The RSA team included members from the City of Tucson, Pima Association of Governments, Sun Tran, the Tucson Police Department, and RSA consultants. Given the large size of the study site, the RSA team identified five focus areas at the intersection and segment levels. After completing the field visit and data analysis, which included crash data from the study area for the last 6 years, the team identified several safety concerns. Some concerns were unique to transit stations and stops, such as the trespassing of private vehicles into the transit center, street closures, and parking. Other concerns were related to pedestrian safety in general, such as reduced sight distance, large turning radius, long crossing distances, and compliance to traffic signals. The team recommended installing larger and more conspicuous "Do Not Enter, Authorized Vehicles Only" signs to reduce the number of trespassing vehicles, adjusting certain parking zones to eliminate violations that interfere with transit vehicles, improving pedestrian crossings through extended curbs and adding a leading pedestrian signal phasing, and applying effective pavement marking to reduce user confusion and improve operations of transit vehicles (Goughnour et al., 2016).

Specific guidelines and prompt lists for pedestrian road safety are available in the *Pedestrian Road Safety Audit Guidelines and Prompt Lists* (Nabors et al., 2007). The guide offers a better understanding of the needs of pedestrians of all abilities, and consists of a Knowledge Base and Field Manual. The Knowledge Base section discusses how to use the guide and includes the basic principles of pedestrian safety (i.e., pedestrian characteristics, pedestrian crashes, and other pedestrian considerations to include in the RSA process). The Field Manual section includes the guidelines and prompt lists.

Active Work Zone Safety Audit

The purpose of the study was to perform a work zone road safety audit (WZRSA) for an active work zone phase. The selected construction project was an urban freeway with the posted speed limit of 65 mph. The multidisciplinary RSA team included safety engineers and work zone engineers from the FHWA, contractors, RSA consultants, as well as designers, traffic and safety engineers, and construction engineers from the Florida Department of Transportation. The RSA team performed a review of two interchanges within the work zone and its adjacent areas. Upon reviewing the project information, plans, and crash data, the team visited the transportation management center to watch traffic patterns and driver behavior from surveillance cameras. Next, they visited the field to closely observe how drivers interacted with the temporal traffic control and configuration of the work zone. The team observed speeding in work zones at night when congestion had decreased, as well as vehicular conflicts at merge/yield areas. In the RSA report, the team recommended increased enforcement during off-peak hours, improving pavement markings and yield signs, and extending the merge area by shifting the barriers (Atkinson, 2013).

The case study exemplifies the application of the *WZRSA Guidelines and Prompt Lists*. The guide includes an eight-step process for performing formal work zone safety examination to improve the safety of construction workers and road users. In particular, a WZRSA assesses the temporary nature of the design that will be removed once the work zone is completed. Agencies can incorporate elements of WZRSA into their current practices by considering work zone safety and mobility at each stage of a project's life cycle (i.e., planning, preliminary design, final design, and active work zone). Mobility considerations that affect safety (i.e., queuing and congestion) can also be considered. The guide points out that impacts on safety and mobility should not be limited to the actual work zone limits. It also notes the importance of considering possible impacts to the construction project's time line, and advises transportation agencies to consider any scheduling issues before beginning the WZRSA.

Road Weather Safety Audit

Most RSAs are performed under favorable weather conditions; consequently, the impact of traffic operations and safety during adverse weather conditions (i.e., snow, rain, sleet, and fog) often are not considered. A study funded by the Wisconsin Department of Transportation explored the possibility of institutionalizing the road weather safety audit (RWSA), and in doing so the study heightened awareness of weather issues that impact infrastructure and operations. The study developed a comprehensive and formalized RWSA for Wisconsin, defined processes and procedures for RWSAs under department organizational structure, designed prompt lists, and identified and provided key data sources for different stages of the audit process (Qin et al., 2013).

Prompt lists were developed for five stages: feasibility, preliminary design (30% of the design stage), detailed design (60% of the design stage), preopening, and existing road facilities. A Weather Constraints section was included in each stage to help evaluate and improve weather-related road safety issues.

Conclusion

An RSA offers a proactive and practical approach to assessing the safety of a road any time during its life cycle. Safety deficiencies are identified by an independent, qualified, and multidisciplinary team, and corrective actions can be taken in the early stages of a project (i.e., before a road is open to the public). Through the lens of RSAs, the safety performance of a road will be evaluated comprehensively, impartially, and preemptively. Evaluating a roadway's safety, with or without crash data, yields great opportunities to address a wide range of safety problems. Over the last 4 decades, RSAs have evolved from isolated pilot projects to programmatic and systematic practices endorsed and adopted by many countries, including the United States.

RSAs are successful only when they follow a structured and formal process. RSA tools, from paper-based prompt lists to today's computer-aided software, have played a significant role in promoting and facilitating the application of RSAs. As the process has become more formalized and the general prompt lists have become more standard, custom RSAs for specific topics (i.e., pedestrian and bicyclist safety, work zone safety, and road weather safety) have been developed in response to emerging safety needs. Although safety subjects will become more diverse as new safety needs continue to grow, the essential elements of RSAs will remain intact: a formal examination by an independent team to identify existing and potential safety concerns for all road users.

An RSA is a crucial safety tool for transportation professionals. The regular use of RSAs promotes and even mandates the consideration of safety during project planning and development, design, construction, operations, and maintenance. RSAs involve all safety stakeholders, from decision-makers to practitioners. RSAs are in accordance with the safe systems approach to road safety management that uses shared responsibility and knowledge to ensure that human error does not lead to death or life-changing injuries.

Biography

Dr. Xiao Qin is a Full Professor in the Department of Civil and Environmental Engineering at the University of Wisconsin-Milwaukee, United States. He has nearly 2 decades of experience in the areas of highway safety, traffic operations, and GIS applications in transportation. He has authored more than 150 technical papers, and won several best paper awards presented by the Transportation Research Board (TRB) and the Institute of Transportation Engineers (ITE). His research has been instrumental in identifying critical safety issues in transportation systems and addressing them with effective methodologies. He is a member of several highway safety TRB committees. He is an associate editor of the *Journal of Transportation Safety & Security* and serves on the editorial board of *Accident Analysis and Prevention*. He received his PhD in Civil and Environmental Engineering from the University of Connecticut.

References

- AASHTO, 2010. Chapter 7 Economic Appraisal, Highway Safety Manual. AASHTO, Washington, DC.
- Atkinson, J., 2013. Work Zone Road Safety Audit Guidelines and Prompt Lists. The American Traffic Safety Services Association (ATSSA), Washington, DC.
- Austroroads, 2019a. Guide to Road Safety Part 6A: Implementing Road Safety Audit. Austroroads, Sydney.
- Austroroads, 2019b. Guide to Road Safety Part 6: Managing Road Safety Audit. Austroroads, Sydney.
- Goughnour, E., Revilla, J., Pitts, C., 2016. Improving Access to Transit Using Road Safety Audits: Four Case Studies (No. FHWA-SA-16-120). Federal Highway Administration, Washington, DC.
- Nabors, D., Cross, F., Moriarty, K., Sawyer, M., Lyon, C., 2012. Road Safety Audits: An Evaluation of RSA Programs and Projects (No. FHWA-SA-12-037). Federal Highway Administration, Washington, DC.
- Nabors, D., Gibbs, M., Sandt, L., Rocchi, S., Wilson, E., Lipinski, M., 2007. Pedestrian Road Safety Audit Guidelines and Prompt Lists (No. FHWA-SA-07-007). Federal Highway Administration, Washington, DC.
- Qin, X., Noyce, D.A., Cutler, C.E., Khan, G., 2013. Development of a comprehensive road weather safety audit program. ITE J. 83 (4), 31.
- Ward, L., 2006. FHWA Road Safety Audit Guidelines (No. FHWA-SA-06-06). Federal Highway Administration, Washington, DC.
- Wilson, E.M., 2003. NCHRP Synthesis 321: Roadway Safety Tools for Local Agencies. Transportation Research Board, National Research Council, Washington, DC.
- Wilson, E.M., Lipinski, M.E., 2004. NCHRP Synthesis 336: Road Safety Audits: A Synthesis of Highway Practice. Transportation Research Board, National Research Council, Washington.

Further Reading

- Nabors, D., Goughnour, E., Thomas, L., DeSantis, W., Sawyer, M., Moriarty, K., 2012. Bicycle Road Safety Audit Guidelines and Prompt Lists (No. FHWA-SA-12-018). Federal Highway Administration, Washington, DC.

Roadside Safety Barriers

Dean C. Alberson, Bulwark Design Innovations, Bryan, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

History	513
Roadside Safety Barriers	513
Acknowledgment	517
Biography	518
See Also	518
References	518
Further Reading	518

History

Roadside Safety Barriers (RSBs) have evolved as transportation modes have evolved. When horse drawn buggies and bicycles were common, traveling speeds, vehicle size, and the opportunities for injury were limited. When motorized vehicles were developed in the mid-1800s, things changed. Mass production of the Ford Model T in 1908 created affordable, higher speed transportation. This meant more people were able to drive at higher speeds and the likelihood of injury escalated. One of the first recorded automotive-related deaths occurred in 1869 when a steam-powered carriage struck a bump and one of the occupants was ejected and run over by the same vehicle (Smallwood, 2013). One of the first recorded impacts with a roadside hazard was recorded in Ohio City, Ohio, in 1891 when James William Lambert driving one of the first single cylinder gasoline automobiles struck a tree root, lost control, and subsequently smashed into a hitching post.

Early guardrail designs used wire ropes, wire mesh, or timber planking as the longitudinal beams or redirective components, and they were mostly supported on timber posts. As traffic speeds increased, 3.6 mm (10-gauge) sheet steel replaced timber beams to accommodate the higher impact energies. The states of Georgia and Missouri Transportation Departments were some of the first jurisdictions to do full-scale testing on guardrail systems in the early to mid-1930s (Division of Design Bureau of Public Roads, 1936). At this time, well over half of the world's private automobiles were located in the United States and most research on and development of roadside barriers were conducted in the United States, and that was the case for decades thereafter as well. This motivates why this article focuses on work and design in the United States.

Some of the early traffic barriers were placed on bridges. Through truss bridges most likely had railings to protect structure rather than providing safety to occupants of errant vehicles. Some open girder-type bridges did have railings, as the consequence of leaving bridge was a known danger. However, initial designs were merely more delineation than an actual longitudinal barrier.

By 1936, approaches to bridges were begin to have guardrail designs incorporated as shown in Fig. 1 from the US Department of Agriculture, Division of Design Bureau of Public Roads (Division of Design Bureau of Public Roads, 1936).

Traffic deaths escalated as traffic volumes and speeds increased. By 1930, total motor vehicle-related deaths in the United States alone had increased to approximately 30,000, more than double the number in 1920. Clearly, motor vehicle deaths were becoming something that needed to be addressed. Fig. 2 shows US motor vehicle deaths since they were first recorded in the 1920s to the current day (Bratland, 2019).

The forgiving roadside concept gained popularity in the mid-1960s and was published in the American Association of State Highway and Transportation Officials' *Highway Design and Operational Practices Related to Highway Safety*, otherwise known as the Yellow Book (American Association of State Highway Transportation Officials, 1967). The Yellow Book gained wide acceptance in the highway safety community and put forth the four steps of the forgiving roadside concept where hazards were (1) eliminated, (2) relocated, (3) made into breakaway devices, or (4) shielded. The last option is accomplished with RSBs.

Roadside Safety Barriers

RSBs such as wire rope or cable barriers, w-beam guardrail, and concrete barriers are placed parallel to the roadway to redirect errant vehicles and prevent impacts with roadside or median hazards. RSBs are used when the roadside hazard presents a greater hazard than the roadside barrier itself. RSBs are a hazard themselves. The first choice in roadway design is to remove all hazards adjacent to the roadway and grade the roadside providing a possible recovery area for drivers leaving the roadway. The obvious next question is, how far away to remove hazards from the roadway? Prior to the highway speeds being increased, the design was to provide a clear zone of recovery for at least 30 ft. (9 m) from the edge of travel lane. With today's speed limits, in the United States reaching 85 mph

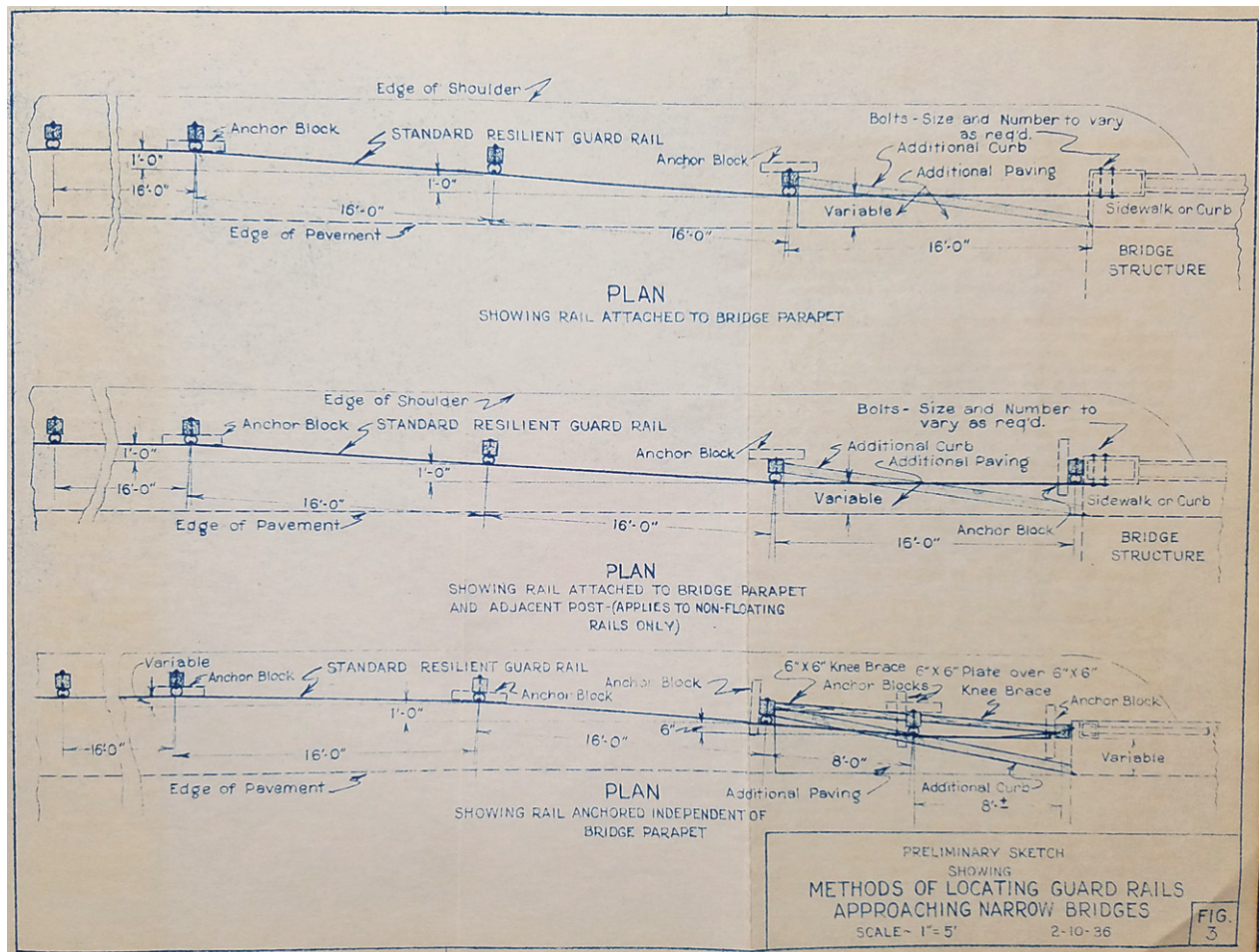


Figure 1 Design of bridge approach guardrails.

(137 km/h), clear zones of recovery can exceed 70 ft. (21 m), specifically in median applications where crossover accidents are almost always severe. These requirements often make designers consider the placement of some type of RSB.

RSBs vary in stiffness. They are considered flexible (wire rope), semirigid (w-beam), and rigid (concrete barriers). Selection of barrier type depends on site conditions, travel speeds, traffic mix, and traffic volumes. Site conditions include hazard type and location, drainage needs, topography, right-of-way extent, etc. Travel speeds and traffic mix dictate the performance level needed from the RSB and traffic volumes may influence designers and maintenance crews toward more rigid systems that require less maintenance. The most flexible systems are generally lower in total cost to install but do have maintenance costs once impacted. Rigid systems have lower maintenance costs but are more costly on the original install. As the name implies, rigid systems have low deflections but impart higher accelerations to occupants of errant vehicles. If medians have been consumed by adding lanes for increased capacity, eventually there will be an insufficient space for expected deflections of flexible or semirigid systems and rigid systems are required. Additionally, there is little space left for maintenance crews to safely service damaged RSBs. Therefore, rigid systems become the barrier of choice in these locations.

Wire rope systems are used in both roadside and median applications. These systems are considered the most flexible with deflections reaching 3.7–4.6 m (12–15 ft.) when impacted by heavier vehicles at high impact angles and speeds. This results in the lowest rates of injuries to occupants. Greater deflection equates to lower lateral deceleration values, in turn, causing fewer injuries. The roadside or median must be able to accommodate these greater deflections without a high risk of secondary impacts to objects in the deflection area. One exception to this concern was investigated in Sweden with the use of two + one roads where high-tension wire ropes replaced double yellow lines on undivided roadways with two lanes in one direction and one lane in the opposite direction. The median strip is typically 1.75 m (5.74 ft.), so only about 0.8 m (less than 3 ft.) from barrier to the oncoming lane which is 3.75 m (12.3 ft.) or 2×3.25 m (2×12.3 ft.) in the double lane direction. So, in theory there should be many accidents where oncoming vehicles are hit but in reality very few occurred. Head-on collisions are “eliminated” and overall fatalities reduced by over 75%. Obviously the number of reported crashes goes up with numerous accidents involving single-vehicle crashes into the wire ropes.

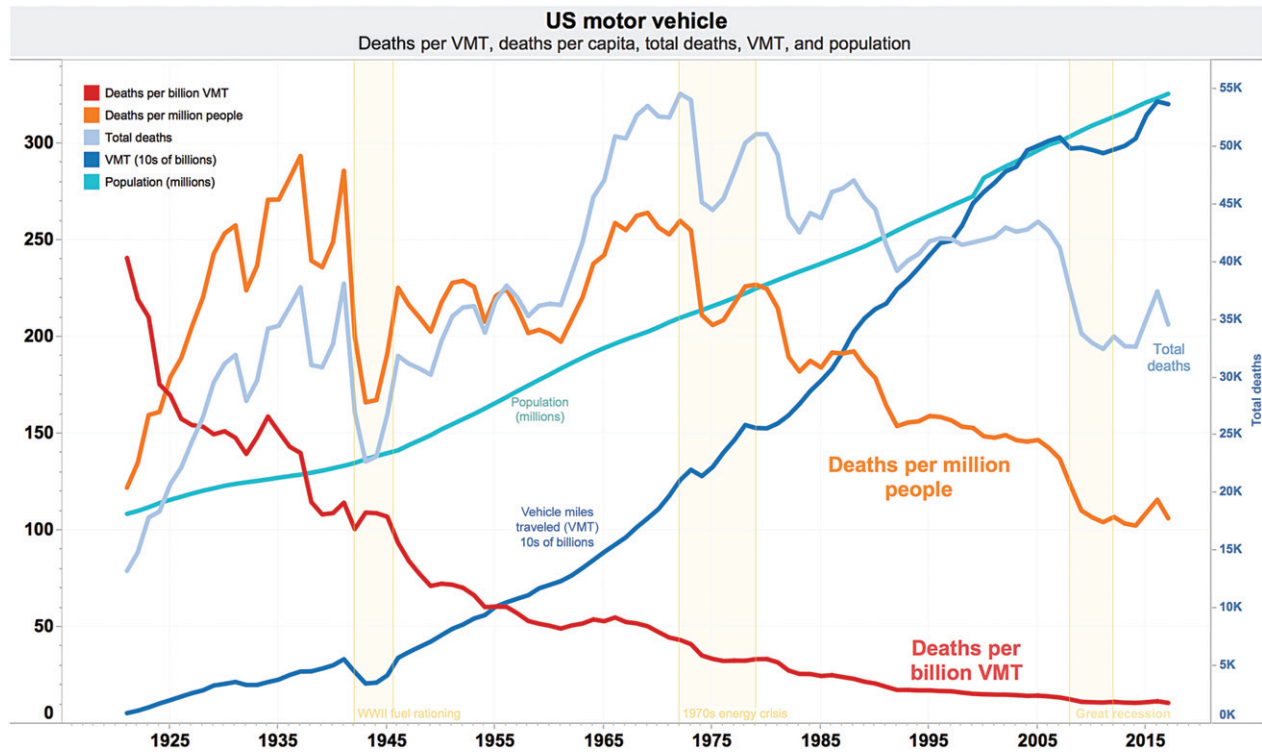


Figure 2 US motor vehicle deaths and trends. Source: Dennis Brantland, Wikimedia with permission.

Early wire rope systems were untensioned and were $\frac{1}{2}$ in. or less in diameter and were initially supported on wood posts. Subsequent designs were suspended on concrete and steel posts. By the 1930s the diameter of wire ropes increased to $\frac{3}{4}$ in. and the structure of the wire rope included three wire strands with seven wires in each strand. This is the same wire rope construction used today in the United States. The individual wires are hot dipped galvanized steel before weaving into the wire rope.

Today, all wire rope systems are mounted on small posts that readily fracture or deform when impacted by errant vehicles. All wire rope systems in the United States were low or no tension, that is, less than 2000 lb (8900 N) of tension in the wire ropes, until approximately 2000–2. A typical low-tension wire rope system is shown in Fig. 3.

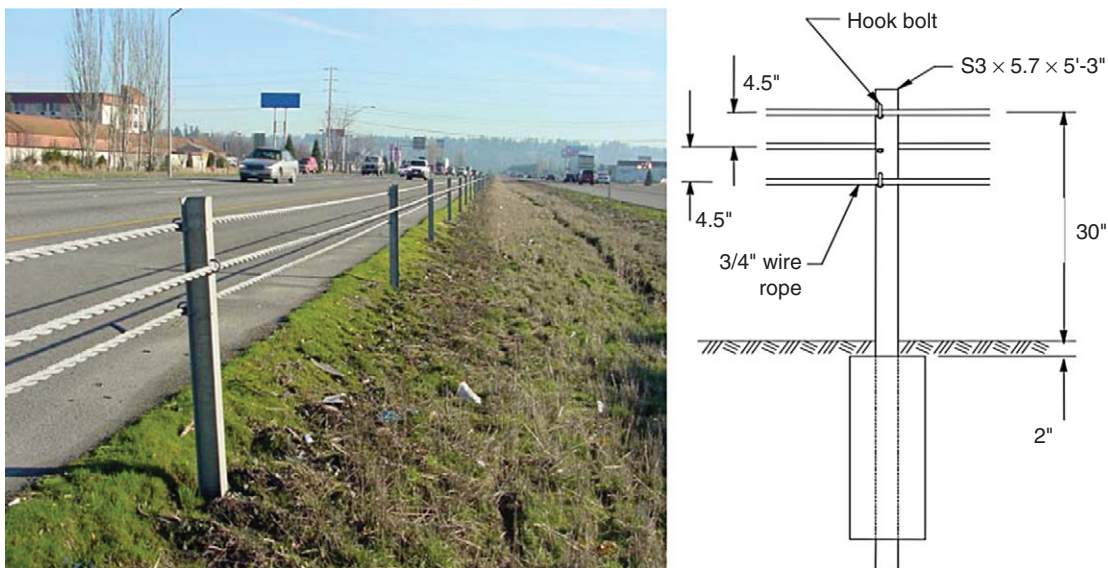


Figure 3 Typical low-tension wire rope system. Source: National Academy of Science.



Figure 4 Roadside and median guardrails. Source: Elderlee used with permission.

High-tension systems, that is, 3,000 lb (13,360 N) of tension or greater, were introduced in the United States at the turn of this century. The systems were previously developed in Europe. High-tension systems are considered to perform better over time as they have residual capacity after impacts, that is, wire ropes are often still in position, for subsequent impacts. Additionally, sagging wire ropes are less of an issue in high-tension systems with large temperature swings. Most high-tension wire rope systems today use prestretched wire ropes that also tend to relax less through thermal cycling each year. Another positive feature of wire rope systems is vehicles are trapped within the system during accidents. This prevents secondary impacts that might occur when vehicles are redirected back into traffic on more rigid systems.

The most common RSB used in the United States is the w-beam guardrails system. This system is considered a semirigid RSB, stiffer than wire ropes but more flexible than concrete barriers. Typical guardrail deflections are approximately 3–5 ft. (0.9–1.5 m) under compliance crash testing conditions.

W-beam guardrails are used in both roadside and median applications and are installed on both weak and strong post. **Fig. 4** shows typical strong post installations for roadside and median.

The weak post system is most often installed on S3x5.7 posts that are spaced at 12 ft. 6 in. (3.8 m) on center and no offset blocks are used. Strong post guardrail systems are mounted on 6x8 wood posts or a W6x8.5 or W6x9 wide flange posts that are spaced 6 ft. 3 in. (1.9 m) on center. Most strong post systems employ either an 8 in. (200 mm) deep or 12 in. (305 mm) deep offset block to minimize wheel snag on the posts during impact. The w-beam rail elements are manufactured in both 12 ft. 6 in. (3.8 m) and 25 ft. (7.6 m) lengths and are fabricated from both 10- and 12-gauge sheet metals with the most common applications using 12-gauge. Pre-galvanized or hot dipped galvanized w-beams are used in most installations. Prior to 2000, most guardrail systems were mounted with the splices at the post and the top of the guardrail was approximately 27 in. (0.70 m) above grade. Vehicles, especially in the United States, have continued to get heavier with higher center of gravities. This has led to design modifications in the w-beam system. Top mounting height is currently recommended to be 31 in. (0.79 m) and the splices are moved to midspan to achieve a slightly greater redirective capacity.

Concrete barriers are the most expensive and rigid of all RSBs. The NCHRP Synthesis 244 Report states, “Although it is not clear exactly when or where the first concrete median barriers were used, concrete median barriers were used in the mid-1940s on US-99 on the descent from the Tehachapi Mountains to the central valley south of Bakersfield, California (Ray and McGinnis, 1997). This first generation of concrete barriers was developed to (a) minimize the number of out of control trucks penetrating the barrier, and (b) eliminate the need for costly and dangerous median barrier maintenance in high-accident locations with narrow medians” Clearly Caltrans was involved in this early version of what came to be known as the K-rail. In the 1950s concrete barriers were developed by the Stevens Institute of Technology, New Jersey, under the direction of the New Jersey Highway Department. Hence, the name New Jersey barrier was adopted. The first installed New Jersey Barrier was only 460 mm tall in 1955 and was subsequently increased in height to 810 mm in 1959 for improved performance.

In the 1970s research was undertaken at Southwest Research Institute to further improve the performance of the concrete safety shape that ultimately leads to the development of the F-shape. The name was derived from the research iterations conducted on variations to the New Jersey barrier. It was the sixth alternative identified in the research project, corresponding to the sixth letter of the alphabet, “F.” While similar in shape to the New Jersey barrier, the height of the lower slope was reduced by 75 mm, reducing the amount of climb that occurred when compared to the New Jersey barrier, improving vehicle stability. The New Jersey barrier and the F-shape are shown in **Fig. 5**.

Another variant of the New Jersey barrier was developed by the Ministry of Transportation of Ontario, the Ontario Tall Wall. The Ontario Tall Wall was 1070 mm tall with the bottom 75 mm embedded in the adjacent pavement. It was unreinforced and successfully crash tested with a 36,000 kg tractor trailer impacting at 85.3 km/h.

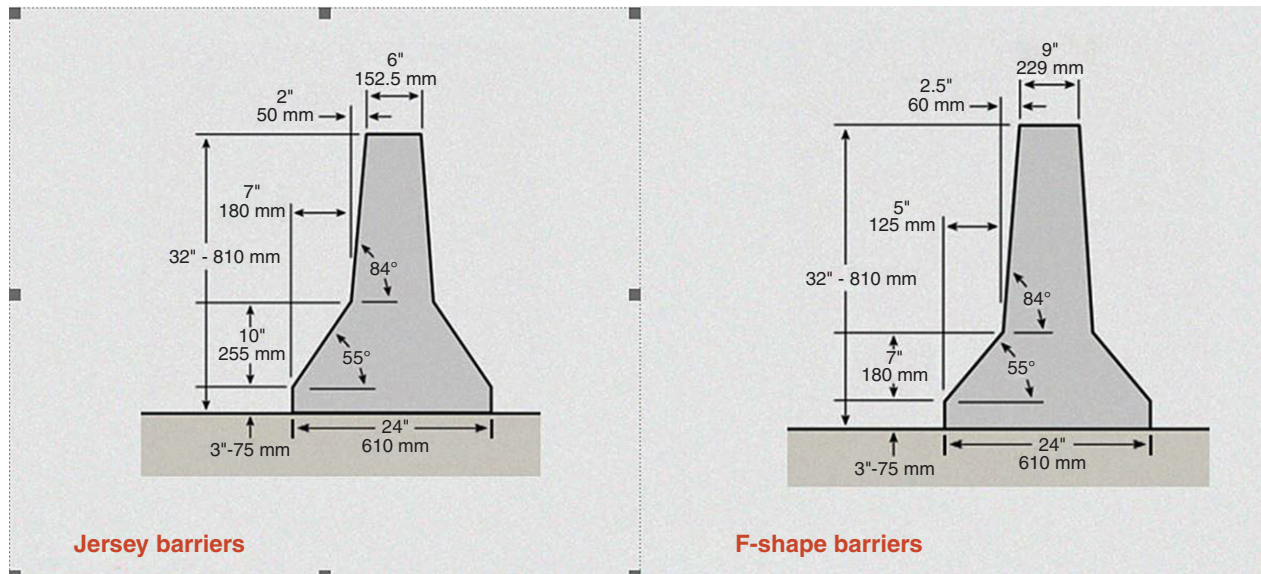


Figure 5 Concrete barrier shapes. Source: JJ Hooks used with permission.

Caltrans and the Texas Transportation Institute both developed Single Slope concrete barriers in the next evolution of concrete RSBs. Both of the systems have slopes of approximately 10 degrees in contrast to the upper slopes on the New Jersey and F-shape barriers of approximately 6 degrees. Caltrans design was slightly taller with a little less slope on the face creating a wider top on the barrier. These designs were selected to accommodate future pavement overlays without changing vehicular interaction with the face of the barriers. The removal of the lower, flatter slope improved vehicle stability during crashes.

Concrete RSBs are installed in both temporary and permanent installations. Temporary installations are often used in construction areas on the roadway, otherwise known as work zones. Temporary concrete barriers, also known as portable concrete barriers, are usually connected to one another but are only pinned or fixed to the roadway when deflections must be limited. Concrete RSBs in permanent locations are often cast in place and integral with the roadway providing anchorage to the system and minimal displacement occurs when impacted.

If temporary systems are unanchored, deflection is controlled by connection type and size of barriers connected. Longer and taller barriers weigh more and provide more resistance during impact. If the connection between barriers is loose or highly flexible, deflections increase. Conversely, tight or stiff connections tend to transfer more load to adjacent barriers, thus reducing deflections. As discussed previously, deflections tend to reduce accelerations to occupants of errant vehicles but, if workers are present behind a temporary barrier, deflections may cause injuries to those workers. There are a number of both proprietary and nonproprietary connection types used in the field today. Loose or flexible connections are easier to install, allowing for more rapid deployment in most cases.

Acknowledgment

This work was completed with financial support from Safe Roads Engineering, Stouffville, Ontario. Editorial review by Mr. Ben Powell is greatly appreciated.

Biography



Dr. Dean C. Alberson has served most of his career at the Texas A&M Transportation Institute (TTI). When he retired from TTI, he was a Senior Research Engineer and Program Manager of the Crashworthy Structures Program in the Roadside Safety Division. He is a registered professional engineer in the State of Texas. He obtained his PhD in Civil Engineering Structures at Texas A&M University.

See Also

The Value of Life and Health; Crash Not Accident; Attenuators; Automobile Accidents and Passive Prevention Systems; Side Area Safety and Side Slopes

References

- American Association of State Highway Transportation Officials, 1967. Highway Design and Operational Practices Related to Highway Safety. American Association of State Highway Transportation Officials.
- Bratland, D., 2019. U.S. Motor Vehicle Deaths. Available from: <https://commons.wikimedia.org/w/index.php?curid=66179446>
- Division of Design Bureau of Public Roads, 1936. A Preliminary Analysis and Discussion of Highway Guardrails. U.S. Department of Agriculture.
- Ray, M.H., McGinnis, R.G., 1997. Synthesis of Highway Practice 244, Guardrail and Median Barrier Crashworthiness. National Cooperative Highway Research Program, National Academy Press, Washington, DC, pp. 69–70.
- Smallwood, K., 2013. The First Car Accident. Today I Found Out. Available from: <http://www.todayifoundout.com/index.php/2013/07/the-first-car-accident/>.

Further Reading

- American Association of State Highway and Transportation Officials, 2011. Roadside Design Guide, fourth ed. American Association of State Highway and Transportation Officials, Washington, DC.
- McDevitt, C.F., 2000. Basics of concrete barriers, Public Roads, Vol. 63, No. 5, March/April 2000. Federal Highway Administration, Washington, DC.

Road Safety Management in Selected Countries

Paul Boase*, Brian Jonah†, *Transport Canada, Ottawa, ON, Canada; †Road Safety Canada Consulting, Ottawa, ON, Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction	519
Australia	519
Canada	521
The Netherlands	522
Sweden	523
United Kingdom	524
United States	525
Conclusion	526
Biographies	527
Related Websites	528
References	528
Further Reading	528

Introduction

The management of the safety of road transportation varies by country in western economies. Some countries are federations of provinces, states, and/or territories which have roles and responsibilities for the federal government and for other levels of government. Typically, federal governments address road safety issues which are common across the country such as motor vehicle regulations and standards for new and imported vehicles, motor carrier transportation, and criminal laws related to road safety (e.g., alcohol and drug impairment, dangerous driving, etc.). Lower levels of government usually manage driver licensing, vehicle registration, and road infrastructure. However, some road safety issues are handled by both the federal governments and the state/provincial governments such as road safety research, including collision data gathering, and policy and program development. In addition, there are oversight agencies which investigate collisions involving road vehicles, and make recommendations for improvements regarding safety (e.g., US National Transportation Safety Board).

Some countries have road transportation research institutes which conduct and/or foster road safety research. Some of these are established at the national level (e.g., Transport Research Laboratory in the United Kingdom, which is now privatized, SWOV in the Netherlands). In other countries, research is conducted within governmental agencies (e.g., US Department of Transportation, Transport Canada) or supported by funding such as the Transportation Research Board in the United States. A considerable amount of research is conducted in university-based institutes such as the University of Michigan Transportation Research Institute. There are also industry-based institutes such as the Insurance Institute for Highway Safety in the United States which is supported by the vehicle insurance industry. While there are a number of governmental, quasi-governmental, and nongovernmental organizations (NGOs) involved in various aspects of road safety management, it is debatable how coordinated these activities are. With so many separate organizations, there is a risk of duplication of activities or some issues may fall between the fractured responsibilities.

The purpose of this chapter is to describe how road safety is managed in selected countries including Australia, Canada, the Netherlands, Sweden, the United Kingdom, and the United States. **Table 1** shows the country profile for each of the subject countries ([International Transport Forum: Road Safety Annual Report, 2019](#)).

It should be noted that vehicle and road regulations are not always within the sole authority of the government. An example would be Sweden, and the Netherlands, which has regulations that all members are required to comply with. Another example would be Canada, which has limited influence with respect to vehicle safety standards given the relative size of the United States versus Canada market. These outside influences must be considered when examining the safety records of individual nations.

Australia

Australia is a federal state consisting of a federal government and seven state and territorial governments. The Australian Government is responsible for regulating safety standards for new vehicles, and for allocating infrastructure resources, including those for safety, across the national highway, and local road networks. State and territorial governments have primary responsibility for funding, planning, designing, and operating the road network, managing vehicle registration and driver licensing systems, and enforcing road user behavior.

The federal Department of Infrastructure, Regional Development, and Cities has a range of functions that support the Australian Government's role in road safety. These include: administering vehicle safety standards for new vehicles, administering the National

Table 1 Country profiles for target countries

	<i>Australia</i>	<i>Canada</i>	<i>Netherlands</i>	<i>Sweden</i>	<i>UK</i>	<i>USA</i>
Population (millions)	24.6	36.3	17.1	9.9	66.6	323.1
Fatalities ^a	1,296	1,898	533	270	1,860	37,806
GDP/capita (USD)	50,811	42,190	45,775	52,224	39,825	57,638
Cost of Crashes/GDP	1.8	2	2	1.3	1.7	1.6–6%
Road Network (km)	873,561	1,304,100	139,124	140,880	422,638	6,700,000
Registered Vehicles (m)	18.8	24.3	10.4	6.1	38.4	288
Cars %	0.75	0.92	79	78	0.83	0.93
Goods Vehicles %	0.2	0.05	10	10	0.12	0.04
Motorcycles %	0.05	0.03	6	7	0.03	0.03
Posted Speed Limits		Vary by Prov.				Vary by State
Urban Roads	50	40–70	30–50	30–50	48	32–128
Rural Roads	60–80	70–90	60–80	60–100	96	40–128
Motorways	100–110	100–110	100–130	110–120	112	40–128
Blood Alcohol (BAC) General Drivers g/dL	0.05	0.04–0.08	0.05	0.02	0.08	0.08
BAC Novice/Young Drivers g/dL	0	0–0.08	0.02	0.02	0.08	0–0.08
WHO Rate/100,000 popn ^b	5.6	5.8	3.8	2.8	3.1	12.4
Fatalities/billion Vehicle Kilometers Traveled (VKT)	5.2	5.1	4.7	3.3	3.4	7.3

^aIRTAD Database, deaths within 30 days of collision^b2016 data from the 2018 Global Status Report on Road Safety

Black Spot Program and other road funding, administering the keys2drive program for new drivers, producing national road safety statistics, and coordinating the *National Road Safety Strategy (NRSS) 2011–20* ([Australian National Road Safety Strategy, 2011–2020](#)). The NRSS, now overseen by the Transport and Infrastructure Council, represents the commitment of federal, state, and territorial governments to an agreed set of national road safety goals, objectives, and actions. It has the specific target of reducing Australia's annual number of road deaths and serious injuries by at least 30% by 2020.

The number of road fatalities in 2016 was 9% lower than the 2008–10 baseline and the fatality rate per 100 million vehicle kilometers traveled was 18% lower. Significant reductions were observed in single vehicle fatalities (–12%), fatal crashes involving young drivers/riders (–25%), fatally injured drivers with a blood alcohol concentration (BAC) over the 0.05 limit (–40%) and crashes involving heavy vehicles (–17%).

A recently published NRSS 2011–20 implementation status report identified areas where progress was being made as well as areas needing improvements.

All jurisdictions have integrated Safe System principles, as seen in [Table 2](#), into road safety project planning, and road authorities are continuing their efforts to increase understanding and acceptance of this accountability throughout their organizations and with local commitment to Safe System principles. All levels of government have assessed risks on their road network and refocused their road investment program on higher risk road sections while balancing safety and mobility objectives.

All jurisdictions are continuing to work to implement safer speeds in rural and urban environments, particularly on roads with a high-crash risk. For example, the state of Victoria has introduced lower speed limits as part of several Safer System Roads Infrastructure Programs (SSRIP), which include high-speed rural roads and local streets. Compliance with speed limits is also being improved. For example, Queensland and Western Australia (WA) are piloting and implementing point-to-point (average speed) camera enforcement. WA's trial has been evaluated and the site is continuing as an enforcement zone. A number of Australian Design Rules for vehicles have been promulgated by the federal government including pole side impact occupant protection standards for

Table 2 Four Principles of the Safe Systems Approach (SSA)

<i>Principle</i>	<i>Description</i>
Ethics	Human life and health are paramount and take priority over mobility and other objectives of the road traffic system (i.e., life and health can never be exchanged for other benefits within the society)
Responsibility	Providers and regulators of the road traffic system share responsibility with users
Safety	Road traffic systems should take account of human fallibility and minimize both the opportunities for errors and the harm done when they occur
Mechanisms for change	Providers and regulators must do their utmost to guarantee the safety of all citizens; they must cooperate with road users; and all three must be ready to change to achieve safety. It is recognized that Canadian jurisdictions will implement the SSA in a manner that is appropriate to their traffic environment

new vehicles (i.e., a test of a vehicle's ability to protect the occupant's head in the event of a side impact with a narrower object), electronic stability control for new heavy vehicles, and anti-lock brake systems for new motorcycles.

Austrroads is an association of road authorities whose purpose is to support its member organizations to deliver an improved Australasian road transport network that meets the future needs of the community, industry, and economy. To achieve this purpose, Austrroads undertakes leading-edge road and transport research underpinning policy development and publishes guidance on the design, construction, and management of the road network and its associated infrastructure. There are several universities that conduct road safety research such as the Monash University Accident Research Centre.

The state of Victoria has a government-owned insurer called the Traffic Accident Commission (TAC), the role of which is to promote road safety, improve the state's trauma system and support those who have been injured in collisions. It is funded through a TAC charge which is a component of the payment made by Victorian motorists when they register their vehicles each year. TAC has established a Toward Zero vision and adopted the Safe System Approach in order to achieve this vision. It conducts awareness campaigns on alcohol and drug-impaired driving, driver distraction by cell phones, speeding, fatigued driving, among others.

Canada

Canada is a federation consisting of a national government and 13 provincial and territorial governments. At the federal level, the Multi-Modal and Road Safety Programs Directorate is located within Transport Canada (TC). Under the authority of the *Motor Vehicle Safety Act*, TC is primarily responsible for the researching and promulgating of motor vehicle standards and regulations, as well as enforcing them through detailed compliance testing, and through defect and collision investigations. These vehicle standards and regulations are administered at the national level and the vehicles are labeled by the manufacturers with the National Safety Mark.

TC also gathers collision-related data from the provincial and territorial governments in order to update the National Collision Database (NCDB) annually and conducts research on road user behavior in order to assist jurisdictions to identify road safety-related issues and their magnitude and nature (e.g., distracted driving, drug-impaired driving, vulnerable road users' safety, etc.).

TC is also responsible for the safety of inter-jurisdictional motor carrier transportation (i.e., movements of heavy vehicles and buses across international and provincial/territorial borders) under the authority of the *Motor Vehicle Transport Act*. TC develops national regulations and standards to promote vehicle safety (e.g., electronic stability control), driver safety (e.g., hours of service regulations), and the management of carrier operations (e.g., electronic logging devices).

It should be noted that the provincial and territorial jurisdictions are responsible for the intra-jurisdictional movement of carriers. Efforts are made to harmonize these regulations with those of the national government but they are not always the same creating difficulty for some carriers operating between jurisdictions. There is also a National Safety Code (NSC) which consists of a number of standards which motor carriers are expected to follow. The NSC was developed jointly by the federal and provincial/territorial governments through the Canadian Council of Motor Transport Administrators (CCMTA), a not-for-profit national body which promotes cooperative road safety activity across the country. The NSC includes 16 standards for maintenance and periodic inspections, as well as driver training and licensing, drivers' hours of service, driver medical fitness, load securement, among others.

The 13 provinces and territories also play a major role in promoting road safety in Canada. They are responsible for the registration of all vehicles operating on public roads, the licensing of all classes of drivers, driver records, and driver improvement, traffic law enforcement on highways and the construction, maintenance, and operation of highways, as well as road safety-related legislation, policy, and regulations.

The jurisdictional governments are organized in different ways to manage road safety. In four provinces (British Columbia, Saskatchewan, Manitoba, and Quebec), there are Crown corporations that provide comprehensive compulsory auto insurance, driver licensing, and vehicle registration services. They also conduct road safety and loss management programs to reduce traffic-related deaths, injuries, and crashes. Road safety-related legislation is usually handled by government departments which works in partnership with law enforcement agencies, service providers, professional organizations, other government agencies in the development of legislation, regulations, policy, and programs.

In other jurisdictions, all aspects of road safety are managed by departments of transportation but in some cases, driver and vehicle licensing are handled by government service agencies or public safety offices. Vehicle insurance should be dealt with by private with private companies in these jurisdictions. In all 13 Canadian jurisdictions, the construction, operation, and maintenance of the highways are the responsibility of the jurisdictional departments of transportation.

Most large municipalities have transportation departments which are responsible for constructing, operating, and maintaining local streets. Typically, highways/expressways which run through municipalities are managed by the province. Some municipalities have their own road safety offices (e.g., Ottawa, Toronto, Edmonton).

Traffic law enforcement also varies by jurisdiction. Larger municipalities have their own police services. Two jurisdictions have provincial police forces (Ontario and Quebec) which perform traffic enforcement on highways and provide policing in smaller municipalities. The other provinces and territories have contracts with the Royal Canadian Mounted Police, the national police force, to provide enforcement on highways and in smaller municipalities.

The Council of Ministers responsible for Transportation and Highway Safety meets annually to address a variety of road safety issues (e.g., National Highway System, driver safety issues and motor carrier safety etc.). The members of this Council are federal, provincial, and territorial ministers who discuss common interests and goals.

Table 3 Road safety strategy 2025 safety matrix of risk groups and contributing factors

		Contributing factors								
		Alcohol impaired driving	Drug impaired driving	Distracted driving	Fatigue impaired drivers	Speed and aggressive drivers	Unrestrained occupants	Environmental factors	Road infrastructure	Vehicle factor
Risk group	Young/Novice driver									
	Medically at risk drivers									
	Vulnerable Road Users									
	Commercial drivers									
	High risk drivers									
	General population									

There are two not-for-profit quasi-governmental organizations which work to coordinate road safety activities across Canada, namely the Canadian Council of Motor Transport Administrators (CCMTA) and the Transportation Association of Canada.

CCMTA has a Board of Directors including senior-level representatives of all federal, provincial, and territorial governments. CCMTA is the custodian of Canada's latest strategy Road Safety Strategy (RSS) 2025, the vision of which is Toward Zero: The Safest Roads in the World. This vision is not a target to be achieved by a certain date; it is aspirational.

Under RSS 2025, each jurisdiction is responsible for identifying their specific road safety issues and developing programs and other interventions to improve road safety. To assist jurisdictions, an inventory of proven and promising countermeasures has been developed by CCMTA based on previously conducted evaluations of these measures in Canada or elsewhere. Jurisdictions are encouraged to implement these measures.

Table 3 shows the safety matrix of risk groups and contributing factors.

The Transportation Association of Canada is a not-for-profit, national technical association dealing with road and highway infrastructure and urban transportation. Membership includes all levels of governments, private sector companies, academic institutions, and other associations. The goal is to provide a neutral, nonpartisan forum to share ideas and information, build knowledge, and pool resources in addressing transportation issues and challenges. Transportation Association of Canada does not set standards but publishes guidelines for planning, design, construction, management, operation, and maintenance of road, highway, and urban transportation infrastructure systems and services.

While there are no road safety institutes in Canada, there are universities which conduct research on road safety (e.g., University of Montreal, University of Toronto, Western University, University of British Columbia). In addition, there are also research institutes which are not-for-profit (e.g., the Traffic Injury Research Foundation in Canada) which conduct contract-based road user research. There are also a number of NGO's which promote road safety such as Mothers Against Drunk Drivers Canada, Arrive Alive Drive Sober, and the Canadian Automobile Association.

The Netherlands

The Netherlands has a central government as well as provincial and local governments. The creation and monitoring of National Road Safety Strategy, the setting of targets, and the development of the RS programs involve the Ministry of Infrastructure and the Environment (MIE), the provinces, urban regions, Rijkswaterstaat (i.e., water boards) and municipalities, Safe Traffic Netherlands (VVN) and the Institute for Road Safety Research (SWOV). Improvements in road infrastructure are managed by the MIE, the Rijkswaterstaat and SWOV. Vehicle improvement is administered by the MIE. The improvement of road user education is led by the MIE but each province has a Regional Road Safety Body which provides information and education to the public. Publicity campaigns are conducted by the MIE while enforcement of road traffic laws is dealt with by the Ministry of Security and Justice, the National Traffic Prosecution Team, and the local police. Much of the road safety research in the Netherlands is carried out by SWOV.

The Netherlands follows a bottom-up approach such that ambitious but realistic targets are defined and updated on the basis of expected trends in casualties ([European Road Safety Observatory: Road Safety Country Overview-Netherlands, n.d.](#)). The prevailing policy plan (Road Safety Strategic Plan 2008–20, *From, for, and by everyone*) has separate targets for fatalities and serious injuries. Every four years, it is reviewed to determine whether the targets for fatalities and serious injuries are still achievable and whether the current policy strategy should be adjusted. This process starts with forecasts for the numbers of fatalities and serious injuries based on a two-step approach. First, extrapolation of past casualty rate trends for different road user categories are combined with forecasts on distances traveled (using a lowest and highest forecast scenario). Secondly, this extrapolation is corrected for changes in road safety policies based on the assumption that known changes in policies do not allow for extrapolation of past trends.

Table 4 Five principles of sustainable development

<i>Principle</i>	<i>Description</i>
Functionality	Related to roads: ideally, road sections and intersections have only one function for all modes of transport (mono-functionality): a traffic flow function or an exchange function
(Bio)mechanics	Ideally, traffic flows and transport modes are compatible with respect to speed, direction, mass, size, and degree of protection supported by the design of the road, the road environment, the vehicle, and, where necessary, additional protective devices
Psychologics	The design of the traffic system is well-aligned with the general competencies and expectations of road users, particularly older road users. The information from the traffic system is perceivable, understandable ("self-explaining"), credible, relevant and feasible. Road users are capable of carrying out their traffic task and adjust their behavior according to the task demands for participating in traffic safely under the prevailing circumstances
Effectively allocating responsibility	Responsibilities are allocated and institutionally embedded in such a way that they guarantee a maximum road safety result for each road user and optimally integrate with the inherent roles and motives of road users (i.e., road users follow the rules and set a good example for children and teenagers). Forgiving traffic systems ensure that road users will not be punished for their errors and weaknesses by crashing and sustaining serious injuries
Learning and innovating in the traffic system	Responsibilities are allocated and institutionally embedded in such a way that they guarantee a maximum road safety result for each road user and optimally integrate with the inherent roles and motives of road users (i.e., road users follow the rules and set a good example for children and teenagers). Forgiving traffic systems ensure that road users will not be punished for their errors and weaknesses by crashing and sustaining serious injuries

The Netherlands has had a Sustainable Road Safety approach since 1992 and it has been found to have a benefit-cost ratio of 3.6: 1. The vision of Sustainable Safety is to have an optimal approach to improve road safety in the Netherlands. A sustainably safe road traffic system prevents road deaths, serious road injuries, and permanent injuries by systematically reducing the underlying risks of the entire traffic system (i.e., Safe System Approach). Human factors are the primary focus: by starting from the demands, competencies, limitations, and vulnerabilities of road users, the traffic system can be realistically adapted to achieve maximum safety.

The current version of the Sustainable Safety Vision is characterized by five road safety principles adapted from the previous principles and strengthened with new insights, thereby providing a basis for specific solutions. The principles are explained in [Table 4](#).

Concerning the road design principles, vulnerable road users (pedestrians and cyclists in particular) and the competence of older road users are the more explicit foci now. The latest edition of the Sustainable Safety vision pays greater attention to cyclist crashes not involving motorized vehicles; responsibility is emphasized with respect to the role and potential actions of stakeholders in realizing an inherently safe road traffic system; in-depth analysis of all fatal road crashes to learn from the things that still go wrong; a pro-active and risk-based approach; and using both crash statistics and road safety performance indicators or surrogate safety measures as safety indicators as a basis for action.

For a number of years, road safety in the Netherlands has been in decline: while road deaths are no longer decreasing, serious road injuries continue to increase. The safety of cyclists is a specific problem, not only in terms of road deaths, but also in terms of serious road injuries. Four developments could perhaps explain why the Dutch performance has deteriorated: a lack of political priority for road safety, reduced government budgets for road safety, a decentralization of "power" from national to regional and local governments, and a lack of priority in the police forces for traffic safety, combined with a substantial increase in traffic, especially by cyclists, and an increased use of mobile phones resulting in distraction.

However, recently, a manifesto supported by about 30 organizations was published in 2017 and a new federal government has included road safety in its program. In addition, the Dutch Parliament has put more pressure on the government to address road safety and the new Transport Minister has supported this.

Sweden

Sweden has a national government as well as several provincial governments ([European Road Safety Observatory: Road Safety Country Overview-Sweden, n.d.](#)). The federal Swedish Transport Administration (Trafikverket) is the government agency responsible for the long-term planning of the transport system. Trafikverket is also in charge of the state road network and road safety policies. Other agencies involved include the Ministry of Enterprise and Innovation, which is responsible for the road infrastructure and the monitoring of road safety development and the Swedish Transport Agency, which is responsible for vehicle safety improvement and road user education and publicity.

Sweden was the first country to adopt the Vision Zero philosophy as a government policy in 1997 putting people first and focusing mainly on reducing collisions that can lead to fatalities or lifelong injuries. This vision's long-term objective is that no one should be killed or seriously injured in traffic and that the design, function, and use of the road transport system shall be adapted to the standards this vision requires. Responsibility for road safety is shared between individual transport system users and "system designers" (i.e., automotive industry, lawmakers, and infrastructure owners). This vision is being pursued by following the Safe System Approach whereby a road system should be designed so as to minimize the harm of potential human errors. Municipalities, companies, organizations, and authorities are collaborating to reach this ambitious goal. If road transport system users do not follow

the rules for reasons such as lack of respect, knowledge, acceptance or capacity or if personal injuries occur in a crash, the system designers must take further measures to prevent deaths and serious injuries.

A methodology was developed in Sweden called OLA: Objective data, List of solutions/actions, and addressed Action Plans. The OLA methodology was used when an analysis of collision statistics revealed an important road safety issue (e.g., too many moped drivers being killed). The aim of OLA was to have different stakeholders work together to solve a road safety issue. The use of this methodology has had an important impact as it committed stakeholders to implement solutions and not just recognizing a problem. Today, OLA is used as a complement to the Management by Objectives framework.

Some of the measures that have been adopted in Sweden include 2 + 1 roads which separate vehicles on highways with a cable on the centerline and replacing signalized intersections with roundabouts. Other measures focus on changing the behavior of drivers (e.g., installing speed cameras, building narrower roads and wider sidewalks in cities so that drivers are forced to keep a slower pace in busy areas, a .02 Blood Alcohol Concentration (BAC), random alcohol breath tests, encouraging car manufacturers to build safer vehicles with built-in features like loud signals that are only switched off when passengers buckle up).

Sweden has a national police force but most of the traffic law enforcement is conducted by local police services. This enforcement activity is overseen by the Swedish National Police Board. Sweden has recently renewed its commitment to Vision Zero.

Sweden has adopted the following target: the number of fatalities in road traffic shall be halved and the number of serious injuries shall decrease by one quarter between 2007 and 2020. In 2007, a total of 471 people died in traffic crashes in Sweden but that figure is now down to 252 in 2017, a reduction of 46.5%.

United Kingdom

The United Kingdom is a unitary state without states or provinces, although it does have local authorities and Scotland and Northern Ireland have some road-related authorities ([European Road Safety Observatory: Road Safety Country Overview-United Kingdom, n.d.](#)). The main central government departments and agencies with road safety responsibilities are transport, highways, health, justice, policing, and health and safety. There is a strong interdepartmental coordination to achieve agreed upon targets, orchestrated by the lead agency, the Department for Transport (DfT). The Secretary of State for Transport has responsibility on behalf of the government for the safety of the road traffic system in England and Wales. In Scotland, it is the Transport Minister and in Northern Ireland, it is the Minister of the Environment.

The DfT includes the Roads, Devolution, and Motoring Group (RDM) that consists of the Road Safety Standards and Services Directorate (RSSS) which is the lead on road safety policy and the coordination of road safety activity. Within the RSSS, the Road User Licensing Insurance and Safety Division (RULIS) is responsible for the development of road safety strategy. However, other divisions of RDM have responsibilities that are relevant to road safety policy such as International Vehicle Standards, Freight Operator Licensing, Active Accessible Travel, the latter being where cycling and pedestrian safety policy sits, Local Transport Infrastructure and national road investment. The RDM agencies responsible for Driver and Vehicle Standards (DVSA), Driver and Vehicle Licensing (DVLA), and Vehicle Certification (VCA) play key roles with regard to driver training and testing and vehicle roadworthiness standards. Responsibility for safety on the strategic roads network rests with Highways England.

The Department of Health, supported by a range of agencies including National Health England and Public Health England, has the mission of “helping people to live better for longer.” It has a core, strategic responsibility for road injury surveillance in the health sector, emergency medical response, major trauma care, the rehabilitation of road crash victims, and road injury prevention.

Local authorities have a general duty under the Road Traffic Act, 1988 to carry out road collision studies and a range of ensuing, appropriate preventative actions. Traffic law enforcement is carried by local police services with the Traffic Commissioners for Great Britain serving as an oversight body.

The United Kingdom has adopted the Safe System Approach in its national strategy to reach its vision of zero fatalities and serious injuries. In 2011, an independent evaluation of the Delivery of Local Road Safety commissioned by the Department for Transport found that “the existence of national targets had provided a useful stimulus to local partnership working.” However, post-2010, there was a change in focus away from national casualty targets. The review found widespread concern among most stakeholders including several key agencies about the absence of explicit national goals and interim targets and the safety performance monitoring associated with these targets that was having a negative impact on the focus on results and levels of activity, both nationally and locally. The lack of a national target had an impact on the priority given to road safety locally and its ability to compete for scarce funds.

In 2018, the DfT commissioned a Road Safety Management Capacity Review which examined institutional management functions, road safety coordination among the various partners, legislation, funding, and resource allocation, road safety promotion, monitoring, and evaluation and research. This review identified the strengths of the UK’s Road Safety Management to be a well established information sharing structure at the national level bringing together key road safety partners and mature, local road safety partnerships. New regional road safety coordination for the strategic road network (SRN) was being developed by Highways England.

Weaknesses in road safety management included the absence of a national road safety performance framework for the short and long-term resulting in a lack of focus and cohesion in coordination efforts; interdepartmental coordination was insufficient to ensure road safety objectives and the Safe System Approach were embedded in the policies of responsible agencies; little engagement in road safety delivery by key national Departments; vertical coordination between central and local government was present but

insufficient when compared with identified good practice; multisectoral involvement was reported to be falling away in local road safety partnerships.

There were many recommendations made in this review including a call for stronger leadership by the central government and a new national road safety coordination hierarchy to strengthen joint activity across the central government and between this government and local governments. This leadership would comprise a Minister-led, high-level Road Safety Strategic Partnership Group (RSSPG) with senior representatives from central and local government, police and other key road safety partners which would be focused on agreed upon priorities within a new road safety strategy and steering and overseeing the delivery of a Safe System Approach and quantified objectives.

The review also recommended that a new national road safety performance framework should be created which would set out the long-term Safe System/Toward Zero goal of working toward the ultimate prevention of deaths and serious injuries, set interim quantitative targets to 2030 to reduce the numbers of deaths and serious injuries, and set measurable, intermediate outcome objectives for activities to 2030 which are directly related to the prevention of death and serious injury. In order to achieve these targets, the strategy should focus on increasing compliance with speed limits on different road types, reducing average speeds on different road types, increasing the level of seat belt and child restraint use, increasing the level of helmet use for two-wheeled vehicle users, reducing driving while impaired by alcohol and drugs, increasing compliance with in-car telephone use rules, increasing the safety of the SRN and main road network to the highest International Road Assessment Programme (iRAP) star rating, and increasing the safety quality of the new car fleet to the highest Euro NCAP rating.

Subsequent to this review, a British Road Safety Statement (BRSS) was published by the national government with a commitment to invest further in continuing road safety activity but endorsing devolution and local decision-making rather than centralized national targets for the United Kingdom. The government's commitment to embedding a Safe System Approach nationally is evident in a safety performance framework for long-term goals and interim targets which are set for the SRN and in the setting up of a Safer Roads Fund. The objectives of the framework are that by 2040, the number of people killed or seriously injured on the SRN should approach zero; by 2020, the aim is for a 40% reduction in fatalities and serious injuries compared to a 2005–09 baseline, and by 2020 more than 90% of travel on the SRN should be on roads with an iRAP rating of 3.

The major public awareness campaign on road safety-related issues is the Think! Campaign, which is managed by the DfT. The campaigns address, alcohol and drug impaired driving, distracted driving, speeding, among others.

Some of the road safety interventions are managed at the local level with support from the DfT. There are also NGOs such as Brake and the Royal Society for the Prevention of Accidents, both of which promote road safety awareness and knowledge.

It is unclear how Brexit and the changing relationship with Europe may impact road safety in the United Kingdom.

United States

The United States is a federated country with a national government and 50 states. The national government has a Department of Transportation (US DoT) that includes the National Highway Traffic Safety Administration (NHTSA). This organization is responsible for the promulgation of motor vehicle safety regulations and their enforcement, as well as research related to vehicle safety and road user behavior and the development of road safety education programs which are supported by regional offices. It maintains a number of national databases including among others, the Fatality Analysis Reporting System (FARS) which is a census of all fatal collisions in the United States and the National Automotive Sampling System (NASS) General Estimates System (GES) which contains information on a representative sample of collisions that occur in the country. NHTSA also supports the New Car Assessment Program (NCAP) which provides consumers with safety ratings of new vehicles based on existing standards. The more stars given a vehicle, the safer it was found to be in testing. The Volpe National Transportation Systems Center conducts vehicle safety research at its facility. NHTSA also publishes the report Countermeasures that Really Work which provides guidance regarding which safety measures could be adopted (Richard et al., 2018).

The US DoT's Federal Motor Carrier Safety Administration (FMCSA) regulates the commercial vehicle (trucks and buses) industry in the United States by ensuring safety in motor carrier operations through enforcement of safety regulations, targeting inspections of high-risk carriers and commercial motor vehicle drivers, improving safety information systems and commercial motor vehicle technologies, strengthening commercial motor vehicle equipment and operating standards, and increasing safety awareness.

The Federal Highways Administration (FHWA) in US DoT supports state and local governments in the design, construction, and maintenance of the nation's highway system. Through financial and technical assistance to state and local governments, the FHWA ensures that America's roads and highways continue to be safe. The FHWA includes an Office of Safety which is responsible for highway designs and technologies that improve safety performance. Major program areas and initiatives include roadway departure, roadside hardware, retro-reflectivity, roadside safety, pavement safety, roadway systems design, intersection safety, Road Safety Audits, speed management, pedestrian/bicyclist safety, local and rural roads safety, and safety countermeasure analysis. Also, included are the Highway Safety Improvement Program (HSIP), Strategic Highway Safety Plans (SHSPs), High-Risk Rural Roads, driver penalty programs and monitoring, data analysis and tools, and coordination with external and internal safety stakeholders, partners, and advocates.

The National Transportation Safety Board is independent from the government and investigates major collisions in all modes including those involving commercial motor vehicles and occasionally light-duty vehicles. Their investigations often include recommendations to the federal government regarding how to prevent such collisions in the future.

In some of the states, road safety is managed by departments of transportation which license drivers and register vehicles, construct, operate, and maintain highways, and conduct research and policy and program development on road safety. In other states, the functions are split between the department of transportation and the department of motor vehicles. Each state has its own set of traffic safety laws. Traffic law enforcement is typically managed by municipal or state police services. Smaller communities are policed by Sheriff's Departments.

The American Association of Motor Vehicle Administrators (AAMVA) is a not-for-profit, nongovernmental organization, which develops model programs in motor vehicle administration, law enforcement and highway safety. AAMVA represents the state and jurisdictional officials in the United States and Canada who administer and enforce motor vehicle laws. AAMVA's programs encourage uniformity and reciprocity among the states and provinces. The association also serves as an information clearinghouse in these areas and acts as the international spokesperson for these interests. The association also serves as a liaison with other levels of government and the private sector. Its research and development activities provide guidelines for more effective public service. AAMVA's membership also includes associations, organizations and businesses that share an interest in the association's goals.

The Governors' Highway Safety Association (GHSA), whose board consists of state Governor appointed directors of traffic safety commissions, is organized to assist its members in implementing the highway safety programs of the Governors of several States, the District of Columbia, and territories to aid them in the development of policies consistent with the needs and goals of these states (e. g., study all problems connected with highway safety; to develop technical, administrative, and educational highway safety standards, to cooperate with other agencies in the consideration and solution of highway safety problems, etc.

The Insurance Institute for Highway Safety, which is funded by the insurance industry, also conducts a vehicle safety rating program based on common crash configurations and it also supports research on various aspects of road safety through special projects.

There are numerous university-based research institutes in the United States such as the University of Michigan Transportation Research Institute, the Virginia Technology and Transportation Institute, and the University of North Carolina Highway Safety Research Center, to name a few which conduct road safety-related research.

Although there is no formal road safety strategy in the United States, in recent years, NHTSA and many state and local transportation agencies have begun to embrace various versions of Safe System/Vision Zero strategies. NHTSA for example, has documented progress toward zero deaths on a state-by-state basis. US DoT, NHTSA, FHWA, FMCSA are all part of a Road to Zero coalition focused on strategies to end-vehicle fatalities by 2050.

Several federal programs mandated at the state level provide an organizational context for the deployment of proven countermeasures and for supporting the principles of a Safe Systems/Toward Zero Deaths approach. A major example is the Strategic Highway Safety Plan (SHSP). Each state is required to develop a coordinated plan for reducing fatalities and serious injuries on public roads. The SHSP for each state convenes the major stakeholders for that state and develops a plan for coordination of efforts in the areas deemed most critical for that state. The SHSP is a requirement of the Highway Safety Improvement Program (HSIP) (23 USC § 148) which provides funding to states.

Conclusion

In the 2018 Annual Report of the International Transport Forum and the Organisation for Economic Cooperation and Development, the change in the number of road deaths is reported for member countries from 2010 to 2016. According to this report, the changes for the six countries included in this review of road safety management can be seen in [Table 5](#). Clearly, despite the similarities in road safety management approaches, they have not resulted in the same outcomes for all countries.

Andras Varhelyi, a professor in Transport and Roads at Lund University in Sweden, has reviewed a number of articles on road safety management and identified several steps that should be followed in order to effectively manage road safety:

1. Define the burden and nature of road casualties;
2. Gain commitment and support from decision makers including politicians;
3. Establish a Road Safety Policy;
4. Define institutional roles and responsibilities;
5. Identify road safety problems;

Table 5 Change in road deaths, 2010–16

Australia	–4.1%
Canada	–15.2%
Sweden	+ 1.5%
The Netherlands	– 1.7%
United Kingdom	– 2.4%
United States	+13.5%

6. Set Road Safety Targets;
7. Formulate Strategy and Action Plan;
8. Allocate responsibility for measures;
9. Ensure funding;
10. Apply measures with known effectiveness;
11. Monitor performance;
12. Stimulate research and capacity building.

A number of international resources are available to assist in the management of road safety. The International Organization for Standardization (ISO 39001) endorses the Plan-Do-Check-Act road safety strategy. The United Nations has a number of working parties dealing with road safety and vehicle regulations. The Global Forum for Road Traffic Safety (WP.1) deals with harmonization of traffic rules by overseeing a number of UN conventions to guide member nations and the World Forum for Harmonization of Vehicle Regulations (WP.29) is tasked with developing a uniform system of vehicle safety regulations.

The United Nations Road Safety Collaboration is an informal forum to discuss and share road safety issues to facilitate international cooperation and to strengthen global and regional coordination among UN agencies. In addition, the World Health Organization is responsible for the production of the Global Status Report on Road Safety, the latest addition being published in 2018. It also coordinates the Decade of Action on Road Safety which is an international effort to improve road safety in all countries.

Six countries of varying sizes and political systems have been presented to identify different safety systems and how they are applied in these countries. These systems are very much a part of the political, fiscal, and social fabric of the country but a number of key points emerge; the need for good data to define the problem and track progress, leadership, the setting of specific short and long-term targets, cooperation among different key stakeholders in road safety and that financing road safety has significant social, health, and fiscal benefits.

Road safety is an important social, financial, and economic issue and as such, improvements in fatality rates may differ across time in developing countries based on changing rules of the roads, enforcement levels, vehicle safety, medical treatment and public attitudes respecting safety. All of these areas must be considered to continually improve the fatality rate in a given jurisdiction.

Biographies



Paul Boase graduated from York University in Toronto with BA in Sociology/Psychology in 1979. In 1982, he graduated from the University of Toronto with Masters Degree in Psychology.

In 1987, he joined the Ministry of Transportation and Communications Ontario as Assistant Research Officer, and in 1990, was promoted to Senior Research Analyst. In this capacity, he worked on the annual collision statistics as well as a number of safety-related projects such as graduated licensing, administrative license suspension and photo radar. In 1999, Paul joined Transport Canada as Chief, Road Users where he is responsible for research related to road user behavior.

Current affiliations include: Board of Directors of the Canadian Association of Road Safety Professionals (CARSP). Member of the Road Safety Research and Policy Committee of the Canadian Council of Motor Transport Administrators (CCMTA).



Brian Jonah is an independent road safety consultant working in Ottawa. He retired from the Road Safety Directorate of Transport Canada in 2008 where he had been the Director, Road Safety Programs responsible for collision data collection and analysis, road user and road infrastructure research, the development of road safety related policy and programs, and communication with the public.

He has a Ph.D. in Social Psychology from Western University in London, Ontario. He has worked over the past 40 years on road safety research, vehicle safety regulation, and policy and program evaluation, with particular emphasis on alcohol and drug impaired driving, seat belt use, risky driving, distracted driving, and young drivers.

He is a member of the Board of the Canadian Association of Road Safety Professionals and is a Past President of the association.

Related Websites

https://roadsafety.gov.au/performance/files/NRSS_Implementation_report_Nov2017.pdf
<https://www.tac-atc.ca/>
https://www.government.se/4a800b/contentassets/b38a99b2571e4116b81d6a5eb2aea71e/trafiksakerhet_160927_webny.pdf
<http://roadsafetystrategy.ca/en/strategy>
https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/erso-country-overview-2017-netherlands_en.pdf
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/717062/road-safety-management-capacity-review.pdf
<https://www.dot.gov/>
<https://www.irap.org/>
<https://www.iso.org/standard/44958.html>
<https://www.who.int/roadsafety/about/en/>
https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/
www.sciencedirect.com/science/article/pii/S038611121500014X?via%3Dihub
<https://benthamopen.com/FULLTEXT/TOTJ-10-137>

References

Australian National Road Safety Strategy 2011-2020: Implementation Status Report. Available from: <https://www.roadsafety.gov.au/nrss>.
European Road Safety Observatory: Road Safety Country Overview-Netherlands. Available from: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/erso-country-overview-2017-netherlands_en.pdf.
European Road Safety Observatory: Road Safety Country Overview-Sweden. Available from: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/erso-country-overview-2016-sweden_en.pdf.
European Road Safety Observatory: Road Safety Country Overview-United Kingdom. Available from: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/erso-country-overview-2016-uk_en.pdf.
International Transport Forum: Road Safety Annual Report, 2019. Available from: <https://www.itf-oecd.org/sites/default/files/docs/irtad-road-safety-annual-report-2019.pdf>.
Richard, C.M., Magee, K., Bacon-Abdelmoteleb, P., Brown, J.L., 2018. Countermeasures that work: a highway safety countermeasure guide for State Highway Safety Offices, 9th ed., Report No. DOT HS 812 478, National Highway Traffic Safety Administration, Washington, DC.

Further Reading

van Schagen, I.N.L.G., Aarts, L.T., 2018. *Sustainable Safety 3rd edition: The advanced vision for 2018-2030*. The Hague, SWOV Institute for Road Safety Research.

Roadway Pavement Conditions and Various Summer and Winter Maintenance Strategies

Kamal Hossain, PhD P. Eng, Assistant Professor, Department of Civil Engineering, Memorial University of Newfoundland, St. John's, Canada

© 2021 Elsevier Ltd. All rights reserved.

Overview of Various Road Conditions	529
Spring and Summer Road Conditions—Surface Distresses and Road Safety	530
Pavement Surface Distresses	530
Rutting	530
Potholes and Delamination	530
Cracking and Roughness	531
Patching	531
Low Surface Friction	532
Road Safety Improvement Through Summer Road Maintenance	532
Rut Maintenance	532
Pothole Maintenance	532
Pavement Crack Management	532
Surface Friction Management	533
Winter Road Surface Conditions and Road Safety	533
Winter Road Surface Conditions	533
Loose Snow-Covered Surface	534
Packed Snow-Covered Surface	534
Icy Road Surface	534
Wet Pavement	534
Winter Road Safety	534
Winter Road Maintenance Strategies, Materials, Guidelines for Improving Road Safety	535
Winter Road Maintenance Strategies	535
Plowing	535
Deicing	536
Anti-Icing	536
Abrasive Application	536
Combined Maintenance Strategies	536
Overview of Snow and Ice Control Materials	537
Winter Road Maintenance Guidelines	537
Determining the Level of Service Need	537
Determining Snow and Ice Control Treatment	537
References	538
Further Reading	538

Overview of Various Road Conditions

Generally speaking, road pavements are designed and built to provide comfortable, smooth, and safe travel for its service life, which is typically about 20–30 years. During this service time, various physical conditions and structural changes occur in roads. The main reason for this change is that, a roadway experiences complex environment throughout its life, in addition to bearing load-induced stresses from repeated traffic loading. Some major chemical compounds or environmental factors interacting with road materials are atmospheric oxygen, dissolved oxygen, UV radiation, and hundreds of other known and unknown solid, liquid, and gaseous compounds. As a consequence of the interactions between road materials and surrounding environmental factors, the road pavement loses its vital engineering properties over time. Often the deteriorated road conditions affect the travel and safety performance of roads. Under the effect of adverse winter events, snow and ice can also significantly affect road conditions. Transportation performance and road safety can be improved by improving roadway conditions. The first step toward this is the understanding on how various road conditions and distresses evolve.



Figure 1 Rutting distress in a roadway. *Source: Author.*

Spring and Summer Road Conditions—Surface Distresses and Road Safety

Pavement Surface Distresses

When designed and built properly, a road pavement has a smooth, continuous, and distress-free surface. Because of the negative effects of various factors including traffic loading and environmental factors, a pavement deteriorates which can result in a distorted, disintegrated, or cracked pavement surface (Garber and Hoel, 2015). The Distress Identification Manual for the Long-Term Pavement Performance Program developed by the Federal Highway Administration of the United States has summarized these distresses for asphalt and Portland-cement concrete roadways and provides detailed explanations of their occurrences (Miller et al., 2014). For example, for asphalt roadways (worldwide, over 90% of our roads are surfaced with asphalt pavements), they include rutting, cracking, patching and potholes, polished slippery surfaces, and miscellaneous distresses. Numerous studies have indicated that these distresses have moderately to strong relation with road safety.

Rutting

Rutting is a major form of road distress, which refers to the vertical depression on the road surface along the wheel paths (Fig. 1). Rutting can occur as a result of pavement being plastic and depressed by heavy loads, or by the grinding effect of studded tires. The former occurs due to the excessive deformation occurring in the pavement structure stemming from issues with the pavement mix design. For example, when excessive air void content is present in the paving mixture, one-dimensional densified rutting occurs. When aggregate mix selected does not provide adequate shearing resistance through the internal frictional action (which can be attained from a selection of properly graded aggregate skeleton), rutting related to lateral displacement (plastic movement) can be induced along the wheel path in the pavement. Rutting from the loss of mastic phase or fine grains of aggregate in surface layer due to the abrasive action of steel-studded tires can also occur. This latter phenomenon, in particular, is heavily observed in the transportation jurisdictions where studded tires are still allowed, specifically in northern states in the United States, Canada, and northern Europe. Not only does rutting significantly affect travel performance, but numerous evidences also exist that it affects ride quality and road safety.

Safety implications: Many studies in the literature show that rutting plays an important role in crash occurrence. Rutted wheel paths can easily be filled with water from rain and snowmelt. On days with heavy rain, a road that has poor drainage design, precipitated water flow through the rutted strip on the road surface like a surface drain—which is, of course, an unfortunate and undesirable aspect for maintaining safety on roads. Rutting can significantly enhance the probability of the loss of vehicle control from hydroplaning. Splashed-water by a vehicle in front or a vehicle in a side-lane can suddenly cover windshields with muddy water, and thus it can impair a driver's vision for safe driving. Based on a study in Wisconsin in the United States, a Transportation Research Board article reports that the crash rate dramatically increases when the rut depth on wheel paths exceed approximately 7.5 mm (Start et al., 1998). Additionally, a deeper rutting channel on the wheel path can lead to plunging oscillation for the vehicle; this vehicle oscillation affects a driver's ability to control the vehicle or forces the driver to change lanes abruptly. The latter can cause a rear end, sideswipe, or roadway departure collision.

Potholes and Delamination

Potholes are localized distresses in the form of bowl-shaped holes of varying sizes—created by the loosening of asphalt film, mastic, and aggregates due to adhesive and cohesive failure in the road materials through water and traffic action. Therefore, potholes often



Figure 2 Potholes distress on a roadway. *Source: Photo used with permission of A. Hawthorn.*

form in areas that have poor drainage, high traffic volumes, or frequent braking especially for heavy vehicles. Also, potholes are one of the prominent distresses in the region where road pavements experience a relatively high amount of freeze-thaw cycles during the winter months. Basically, during the freeze-thaw cycle, expansion and contraction occur in pavement materials. Thus, these expansions and contractions induce stresses in pavement and nucleate micro cracking. With the traffic loading, moisture and water from the atmosphere, rain, and snow, the small-crack turns into bigger holes. Besides, from the thawing of ices, voids, or holes are created in the pavement structure. **Fig. 2** shows a series of pothole in a service lane. Delamination is the gradual removal of an area of asphalt surface due to poor bonding with the underlying paving layer. This type of distress reduces the serviceability and road safety of the pavement as they can create water ponding in the road, which ultimately affects driver's driving behavior.

Safety implications: Physical defects, such as potholes and delamination can influence the driver to change the driving lane abruptly or swerve around the pothole. These types of driving actions are associated with high collision rates, especially on multi-lane roads. However, just making a road pavement smoother without improving its friction and without improving roadside features can lead to higher speeds and higher crash rates as observed on two-lane rural roads in New York State. So, in order to maximize safety, smooth pavement should be strived for and at the same time speed should be controlled and enforced through other means than by having people slow down by navigating around potholes. And even if drivers try to navigate around potholes, they often miss one and the damage to the vehicles wheels, suspension and shock absorbers can have serious implications on their safety at a later time.

Cracking and Roughness

Various types of pavement cracking occur in the life of a road pavement. They include fatigue, thermal, moisture, and aging/stiffness-related cracking. The cyclic loading from vehicular traffic, pavement temperature, frequency, and magnitude of the change of air and pavement temperatures can contribute to fatigue-related cracking. Pavements in regions with frequent freeze and thaw cycles, and/or extremely cold climate experience relatively more thermally induced stresses (from repeated contraction and expansion in pavement materials); this leads to thermal-related cracking. When an asphalt pavement ages, it becomes stiffer and exhibit more brittle behavior. With the loading from vehicular traffic, a stiffened pavement exhibits age-related cracking. If these pavement cracks are not treated in a timely manner, as indicated above, they can lead to bigger holes "potholes" which can have severe consequences on traffic safety.

Additionally, pavement cracks combined with other factors including pavement surface properties, extent, and severity of pavement distresses contribute to pavement roughness ([Hall et al., 2009](#)). Pavement roughness is generally defined as an expression of irregularities in the pavement surface that adversely affect the ride quality and safety. Degradation of the pavement roughness has a significant impact on safety. Roughness is measured from the vertical movement of a vehicle traveling at a speed of 80 km/h and is defined as the ratio of accumulated suspension motion of a car to the distance traveled ([Sayers, 1995](#)). Roughness is expressed in meters/kilometers or inches/mile. It commonly ranges from a value of about 1–5 m/km on a paved highway.

Safety implications: Low roughness values normally indicate a smooth surface. According to the US Federal Highway Administration, high-speed pavements with roughness value greater than 2.7 m/km are classified to be in "poor" condition ([Guerre et al., 2012](#)).

Patching

Generally, pavement patching is conducted to repair local distresses temporarily. A thin layer from the faulty area of the pavement is removed and replaced with a new paving material. Patching contributes to address road safety and ride quality issues. In general, less preparation and care are taken to conduct a patching project. Thus, patching areas are often found as an elevated area from the true planar road surface.

Safety implications: Patching can influence drivers to change lanes suddenly and can lead to rear-end, sideswipe or run-of-road crashes.

Low Surface Friction

From a roadway perspective, pavement surface friction mainly depends on the surface texture of the pavement, surface temperature, and moisture availability on the road surface. The surface texture refers to the deviations of a pavement surface from a given true planar surface, and it varies with the physical characteristics and gradation of aggregates used in the road construction and design of paving mixture (Hall et al., 2009). The sizes, angularity, shape factor, and durability of the aggregates used in pavement construction will dominate the pavement surface texture. Typically, the surface friction decreases over the pavement life as aggregates become polished by traffic action. And, frictional resistance (skid resistance) between the tire of the vehicle and the pavement surface has a profound effect on road safety. Therefore, during highway geometric design, the pavement friction value for the given pavement type is needed to determine the design values of stopping sight distance, highway curve radius and length, and superelevation.

Safety implications: A higher pavement surface friction value would mean that the driver would have more control of their vehicle when the pavement is wet or has a film of water on it.

Road Safety Improvement Through Summer Road Maintenance

Numerous studies present that poor road conditions significantly influence traffic safety and travel performance. Therefore, transportation agencies across the world strive to improve road safety by conducting proper road maintenance through delivering various road programs. When road conditions are evaluated along with road safety in terms of the number of collisions and their level of severity, obtaining a complete picture of various engineering and functional properties including the pavement's physical conditions, materials and mix designs, geometric designs, and roadway safety features, is essential.

Highway safety can be improved through active and passive safety programs. An active safety program is delivered through preventive changes to roadway features (e.g., installing high friction surface courses, guard rail) to prevent or reduce the severity of accidents. Pavement maintenance programs can be delivered as an active safety strategy by reducing distresses or deterring the deterioration rate of pavement surface conditions with rutting, potholes, cracking, and roughness.

Rut Maintenance

Minor surface ruts or a small strip of the rut can be filled with asphalt patching materials, and deeper ruts can be covered by applying an overlay in the rutted lane. If the surface asphalt layer is unstable, recycling and repaving can be performed to get rid of rutting. In the case of ruts related to subgrade inconsistencies, reconstruction is the best option.

Pothole Maintenance

Potholes can be repaired by asphalt patching work or by rebuilding a road section that has a relatively large number of potholes. Patching is a very common method used to cover up the potholes. Within patching methods, "throw-and-go" is the most widely used patching method (Simita et al., 2016). In this method, asphalt materials are placed into the pothole and then they are compacted with the maintenance vehicle tires. This is often done hastily and thus it leaves an elevated crown above the road surface plane. A semi-permanent pothole repair method can also be performed, and the performance of this method is better than the "throw-and-go" method. This method requires that water, disintegrated debris and aggregates are removed as much as possible before placing the asphalt mixture and compacting it with a proper compaction system. Edges of the patching area are leveled to obtain a smooth continuous pavement surface. Spray injection is another efficient pothole repair method that can be done to cover potholes. This process combines hot asphalt emulsion and crushed aggregate and uses forced air. Cleaning of debris and spraying tack coats are performed to further enhance the performance of this pothole repair method. If a road section is full of deep potholes, an overlay or reconstruction can be a suitable option, dependent on the severity of potholes.

To identify subsurface delamination effectively, non-destructive test methods, such as Ground Penetrating Radar or strain gages can be used. Delaminated areas can be treated by milling off the surface layer, replacing the wearing course, or placing a thin asphalt overlay.

Pavement Crack Management

To measure the extent of cracks, such as fatigue cracks or block cracking, the linear distance of the affected wheel path or square meters of the pavement area is considered. The affected zone can be classified according to the level of severity that can help rank the roads need in maintenance for allocating scarce maintenance dollars. A full depth patch or spray patching can be performed for temporary repair for fatigue cracking areas. A thin wearing course can be constructed to repair a low severity block cracking. For more severe situations, overlay with or without recycling can be performed. For base problems, reclamation and reconstruction of the pavement can be done. For longitudinal and traverse cracks, the number and length of cracks at each severity level are measured. Crack sealing can be performed for low severity cracks and an overlay can be placed for high severity cracks.



Figure 3 High friction surface treatment for enhancing pavement friction—treated (A) and untreated (B). *Source: Photo used with permission of The Transtec Group, Inc.*

Surface Friction Management

Maintaining safe friction on pavement surfaces is indispensable for road safety. Various high-speed and low-speed or stationary equipment including Locked-wheel skid trailer, Mu-meter, Fixed slip grip-tester, British pendulum tester, Sand patch system can be used to measure friction on pavements. Highway agencies aim to maintain an appropriate level of surface friction for all road sections within the network, based on each section's friction demand which depends on traffic volume levels, highway class, climatic zone, and crash history. The road sections that have friction values less than the friction demand value can undergo more rigorous testing in terms of micro and macro textures of the pavement surface and crash pattern. This process can help design immediate intervention with appropriate surface maintenance project to restore friction demand value. **Fig. 3** shows a treated and untreated pavement section, where an epoxy-based aluminum oxide additive is used to improve friction on the road surface.

The US National Cooperative Highway Program Report No. 108 indicates that aggregates' mineralogy, hardness, shape, angularity, and polish and abrasive resistance influence on pavement friction. Therefore, during pavement surface maintenance or reconstruction of pavement, a comprehensive material and mixture testing system can help identify the right recipe that can generate an appropriate combination of micro and macro textures in the pavement surface. The literature shows that using an appropriate surface friction course opposed to conventional asphalt mixtures can lead to millions of dollars in savings from reductions of crash numbers. **Table 1** summarizes the major roadway distresses, their road safety threats, and treatment options for improving road safety.

Winter Road Surface Conditions and Road Safety

Winter Road Surface Conditions

Winter weather can cause pavement surfaces to be contaminated with snow and ice, resulting in reduced surface friction. Pavement surface contaminants are the eventualities or products of snow after some external forces and actions. For example, compacted snow after plowing by truck or manual power, slushy snow due to positive thermal changes and traffic actions, and bonded ice due to the

Table 1 Pavement distress, road safety risks and treatment options

<i>Pavement distress</i>	<i>Road safety risks</i>	<i>Treatment options</i>
Ruts	Moderate to high	Asphalt patching Overlay Reconstruction
Potholes	High	Asphalt patching—throw and go, semi-permanent method Spray injection Reconstruction
Longitudinal and traverse cracks	Low	Crack sealing—rout and seal Seal coat—chip seal, slurry seal
Low friction pavement (Polished pavement surface)	High	Thin overlay High friction surface course Open-graded asphalt course

temperature falling below the freezing point are all types of pavement contaminants. Winter pavement surface conditions could vary significantly by contaminant type, coverage, and depth; however, in practice, they are commonly classified into four types, namely, surface with loose snow, packed snow-covered surface (leftover from a plowing operation), icy pavement surface, and wet pavement. These contaminants reduce the serviceability of a road, increasing travel times, or pose a threat to traffic safety.

Loose Snow-Covered Surface

Road surfaces become contaminated with newly fallen snow. The type of snow can vary from very loose to very packed, and thickness can vary from very thin to very thick amounts. A bond can start to form under the top layer of the snow when the pavement temperature falls below the freezing point. Road safety characteristic, that is, the surface friction of this type of pavement largely depends on the amount and physical properties of the snow (e.g., crust/packed snow vs. loose snow) rather than whether the snow is bonded with the pavement surface underneath or not.

Packed Snow-Covered Surface

This type of road surface can be observed when snow is present either in compacted form or in patches of snow after the snow has been plowed either by truck or manual power—which in North America is used primarily in small parking areas or walkways. If the initial snow is wet or the temperature is above the freezing point, a bond does not form and the snow can be mostly cleared by only plowing, leaving patches of snow; however, in the case of dry snow, plowing operation can leave the pavement with compacted snow. Thus, the friction of this type of pavement could become quite low (e.g., 0.20–0.50), either for initial snow type (e.g., dry snow vs. wet snow) and/or plowing method (manual plow vs. truck plow). Minor highways in some northern US states, and all roads in, for example, northern Sweden and all of Finland—including city streets and freeways—are always left “white” during all of the winters. They are plowed to be travelable but never salted and typically not sanded either except for in steep hills and at stop signs. Also, sidewalks are sanded but not salted.

Icy Road Surface

The pavement can be contaminated with ice when the temperature tumbles below the freezing point and no appropriate snow-melting chemicals (e.g., salt) have been applied to control the snow contamination. Icy pavements can also be caused by freezing rain, which happens when the ground and roadway temperatures are below the freezing point but the atmosphere above the roadway is warm. In addition to the regular ice, another type of ice is also observed, referred to as “black ice” in Canada or in northern states. This is a special type of ice that is formed when a very thin film of water on the pavement surface is frozen due to freezing temperature. “Black ice” is less visible than regular ice and has a much smoother surface. It can easily be overlooked and thus cause a higher risk of traffic accidents than a snow-covered road. Icy road surfaces pose significant threats to both vehicle and pedestrian safety during the winter season. The friction coefficient for an untreated road covered by ice may be as low as 0.1 (Hossain et al., 2015).

Wet Pavement

If the pavement is clear of contaminant types mentioned earlier, but is damp with a thin film of water from snow melted solution or winter precipitation, it is considered to be a wet pavement. Note that, in the cases where the puddles of water were present, pavement friction can certainly fall down below a safe friction value (approximately 0.40) and pose a significant threat to traffic safety. Over 12% of the total number of accidents that occurred between 2000 and 2009 in the United States occurred on wet pavements with reduced pavement friction (McGovern et al., 2011). **Table 2** summarizes the various winter road conditions, their threat levels to road safety, and treatment options to mitigate road safety issue.

Winter Road Safety

The accumulation of snow and ice contaminants on a pavement affects the mobility and safety of pedestrians and vehicular traffic. For example, every year there are thousands of pedestrians in Sweden who are injured because of slippery pavements and roadways and studies in that country shows that foot slippage caused 43% of all falls and 16% of all accidents at work, in the home and during leisure activities in the Nordic countries (Lund, 1984). Two-thirds of these slips occurred on sidewalks or other surfaces covered by ice and snow. Studies in other countries have also shown that surfaces covered with snow and ice increase a pedestrian’s risk of slipping and falling by reducing the average level of traction (friction) of the pavement surface. Fundamentally, slips and falls are caused by a loss of traction or friction between the pedestrian’s shoes and the surface being walked on. Pedestrian injuries on slippery surfaces are more severe than those on non-slippery surfaces. Therefore, maintaining safe traction or friction on these surfaces is critical to the safety of pedestrian and vehicular traffic during the winter months. To improve the surface conditions (i.e., the friction level), both public and private organizations put large efforts and resources into making sure the public is safe under winter weather conditions. However, in contrast to highways, parking lots and sidewalks are commonly maintained by small contractors and/or individual property owners, who may have varying or loose guidelines to follow for controlling snow and ice on these areas. This

Table 2 Winter road conditions, road safety risks, and treatment options

<i>Winter road conditions and weather events</i>	<i>Road safety risks</i>	<i>Treatment options</i>
Heavy snowfall forecast	Medium to high	Anti-icing—dry, pre-wet salt, or liquid salt application Do nothing and wait for forecast to be true
Road surface covered with loose snow	Low to high	Plowing only Deicing with dry or pre-wetted salt Plowing and deicing Abrasive use
Packed snow-covered road surface	High	Deicing with dry or pre-wetted salt Abrasive use
Icy pavement surface	High	Deicing with pre-wet solid salt or direct liquid Abrasive use
Wet pavement	Medium to high	Abrasive use Anti-icing, if temperature continues to drop

lack of uniform standards and guidelines is considered as one of the main factors contributing to significant legal challenges of slip-and-fall cases and over application of salt in fear of the former. From a sustainability perspective, a majority of the contractors applies excess salt to avoid slips and falls, in order to avoid litigations and increases in insurance premiums.

Reduced pavement friction due to snow and ice contamination contributes to increased collision risk for vehicular traffic. Therefore, vehicular traffic is significantly more likely to be involved in accidents in the winter than in the summer for a given distance of travel. With respect to fatal crashes, there is a reduction in the number of fatalities during the winter months in northern states of the United States and in the Nordic Countries. For example, in Sweden, the number of fatalities from traffic accidents for January through April is roughly half of that for May through August. And, in Maine, February has had the fewest fatalities of any month for each year during at least the last 10 years in spite of February typically is the month with the most snowfall. The primary cause that the snowiest months have the fewest fatalities is that people avoid traveling in winter weather. But, heavy snowfall and white roads also make people drive slower so that when they have accidents; they are less likely to be fatal. However, black ice is often missed by drivers and as long as we do not have good information systems informing drivers—or their cars—of that situation, black ice can and does increase fatality rates. Winter snow and ice-related traffic crash costs including both personal injury and property damages are in the range of billions of dollars each year globally. Various factors influence road safety in the winter season. The major factors affecting winter road safety are snow, ice, rain, and visibility from a weather perspective; vehicle speed, traffic volume, and vehicle safety features from a traffic perspective; and operational strategies, treatment types, degree, and efforts from winter road operation perspective. Suitable winter road maintenance can improve road safety in the winter months but studies in Maine indicate that injury rates are only marginally improved by improved winter maintenance. The big gain from improved winter maintenance is keeping travel times close to normal summertime values saving people large amounts of time. Various winter maintenance methods are employed which are based on the existing surface contaminants, expected snow event characteristics, traffic level, and level of service (LOS) requirements for the road.

Winter Road Maintenance Strategies, Materials, Guidelines for Improving Road Safety

Winter Road Maintenance Strategies

Transportation agencies that experience any amount of snowfall during winter months have adapted various methods of snow and ice control. These methods can sometimes vary according to winter behavior for that specific geographic area, and may include methods, such as plowing, salting, sanding, etc. Winter maintenance methods can be classified into three distinct categories: chemical, mechanical, and thermal categories. Applying a snow control chemical (e.g., sodium chloride) as a freezing-point depressant on pavement or integrating the freezing-point depressant into the pavement is an example of a chemical method. Methods such as plowing, scraping, or using high-velocity air to blow snow are all classified as mechanical. Finally, thermal methods include those that control or prevent the formation of snow through the application of heat, either from above or below the pavement surface. Recent technological and logistical developments have allowed a higher level of diversification in snow and ice control. Many different types of winter operations are currently in use, including deicing (e.g., post-salting only, plowing, and salting after), anti-icing (e.g., pre-salting only, pre-salting, and plowing after), and application of pre-wetted solid salts, liquid salts, or even combinations of these methods (Hossain et al., 2018). The following provides an overview of common snow and ice control methods.

Plowing

Plowing is the most basic form of snow and ice control that exists today. The exclusive use of plowing is often the most economical way to treat snow at some establishments, especially those with low LOS requirements. Despite this, a plowing only

method can still be effective even when a higher LOS is needed, if snow does not bond to the pavement, or if a white road is an accepted outcome as in most regions of the Nordic countries. For places where the exclusive use of plowing is not sufficient, it can be made more effective if it is conducted in conjunction with anti-icing. Indeed, a large number of studies indicate that crash rates decrease significantly after plowing.

Deicing

Deicing is a method of snow and ice control in which chemical agents are applied to melt the snow and ice already accumulated on a pavement surface. Chemical agents work by lowering the freezing point of water or by breaking the previously formed bond between snow and pavement. The most common type of salt for this application is rock salt; and, depending on the types and amounts of salts being used, these deicing methods can be effective at temperatures ranging from -7°C to -12°C .

Deicing treatment can also be conducted with other alternative materials, some of which contain less or no chlorides and thus have lower environmental effects. Salt can be pre-wetted using brine or other liquids for improved performance. Pre-wetting has been shown to be an effective method to provide a higher LOS for the following two reasons. First, wet salt can better adhere to the pavement surface, resulting in less scattering, and less material usage. Second, salt requires moisture to activate the deicing process.

The deicing operation has several unavoidable issues. First, the most common deicing chemical used is sodium chloride, and its application has the potential to cause substantial problems including contamination of drinking water and sub-soil layers and corrosion of concrete pavement and motor vehicles. Second, because of poor roadway conditions, both prior to and during maintenance, the potential for accidents may be increased when deicing operations are conducted. Additionally, it is important to note that deicing may also consume large amounts of snow control materials and labor to achieve a desired LOS.

Anti-Icing

Anti-icing is a strategy, which applies snow and ice control materials before or immediately after a snow event starts. The objective of anti-icing is to prevent the bonding of snow and ice to a pavement surface. If an anti-icing treatment is conducted at an appropriate time of the snow event, and not during severe storm conditions or on an extremely cold pavement surface, that is, colder than -8°C (23°F), it can be very effective. The anti-icing operations can contribute to cost savings for both highway agencies and motorists by reducing the use of materials and by reducing the number of accidents, respectively. In addition to its benefits against regular winter weather, anti-icing is particularly effective in dealing with heavy frosts.

In applying an anti-icing strategy, both solid and liquid chemicals can be used. Solid chemicals specifically have been in use for many years in anti-icing operations. If dry solid anti-icing agents are applied, the moisture in the air must be sufficient, as moisture reduces the ability of an anti-icing agent to be blown off the road by causing it to stick to the pavement. In addition to this, moisture has also been shown to improve the performance of applied chemicals. For roads, solid salt with water or another liquid chemical is recommended in order to minimize the bounce and scatter tendencies of salts and reduce the loss of particles from wind-blow caused by vehicular traffic.

Abrasive Application

The main objective of using abrasives is to provide improved traction on ice-covered roadways, especially when it is too cold for other chemicals to work effectively. Transportation and environmental agencies also prefer using abrasives when roads are too close to a water body, such as a lake, river, and pond, to reduce the impact of deicing salts in water quality. Abrasives such as sand and sand-salt mix are commonly used to provide improved traction on ice-covered roadways. Several kinds of abrasives can be used for snow and ice control, including natural sands, finely crushed rocks or gravels, bottom ashes, slags, ore tailings, and cinders. The application rate for abrasives varies among the different winter maintenance agencies due to the diverse weather conditions. It is important to note, however, that since most of the abrasives used are inert substances (e.g., sand), they will not help melt snow and ice. Moreover, abrasives can be expensive when the total costs of handling, cleaning, drainage, and environmental costs are considered. Also, pure sand tends to be swept away by trucks traveling at high speeds, so the treatment may lose most of its effect already after 10–20 trucks have passed by whereas salt sticks to the surface and the effect last much longer.

Combined Maintenance Strategies

Because winter weather events in different areas present a variety of weather and pavement conditions, a combination of strategies is almost always used (Blackburn et al., 2004). Different LOS requirements, scope of services, and maintenance agency's ability to provide various combinations of maintenance methods can influence to design combined maintenance strategies, such as using abrasive and chemical mixes, using mechanical and anti-icing methods at the same time, or applying mechanical and abrasives.

Table 3 Properties of some common snow and ice control materials

Material	Chemical composition	Forms used in snow and ice control operations	Eutectic temp and Eutectic conc. °C (°F)
Sodium chloride	NaCl	Primarily solid, but increasing use of liquid	–21 (–5.8) @23.3%
Calcium chloride	CaCl ₂	Mostly liquid brine, some solid flake	–51 (–60) @29.8%
Magnesium chloride	MgCl ₂	Mostly liquid brine, some solid flake	–33 (–28) @21.6%
Calcium magnesium acetate	CaMgAc	Mostly liquid with some solid	–27.5 (–17.5) @32.5%
Potassium acetate	KAc	Liquid only	–60 (–76) @49%
Agricultural by-products	NA	Liquid only	Usually blended with chloride-based products
Other organic materials	NA	Liquid only	Varies with product

Overview of Snow and Ice Control Materials

Snow and ice control materials include various chemical products as well as abrasives. Deicing materials can generally be classified into two types, namely chloride-based and nonchloride-based. The most commonly used chloride-based materials include sodium chloride (NaCl), calcium chloride (CaCl₂), and magnesium chloride (MgCl₂). These materials are generally produced from the mining of surface or underground deposits, extracting and fractionating brine from wells, industrial by-products, or through solarizing saltwater.

Nonchloride-based products, also called organic products, are mostly manufactured. Some are wholly synthesized (e.g., CMA and KA) while others are refined from agricultural sources (e.g., by-products from grain processing, brewing, winemaking, and similar sources). These materials are not as popular as chloride-based products, but are used as either stand-alone liquids, blended with inorganic liquids, or as stockpile treatments. In general, due to their high costs, they tend to be used for special situations (e.g., for low-corrosion applications, such as bridge decks). Most of these products are proprietary with little information about their actual manufacturing and refining process. These products are usually used in conjunction with chloride-based products, though stand-alone products have also been marketed. Many of these products have been claimed to have the benefits of being less corrosive and more effective in snow melting. Table 3 presents the chemical properties of common snow and ice control chemicals used.

Winter Road Maintenance Guidelines

This section gives an overview of winter road maintenance guidelines that can help the maintenance industry responsible for snow and ice control of highways, parking lots, and sidewalks to optimize maintenance operations and minimize salt usage.

Determining the Level of Service Need

Any maintenance program for a transportation facility starts with establishing the desired LOS that should be delivered during the winter event or the winter season. The desired LOS must be realistic and cost-effective due to the random nature of winter events. As a result, it is preferable that each LOS standard includes a probability quantifier. Some highway agencies designate that all Class I highways (high priority) must reach bare pavement within 8 h for 90% of the events over a season, that is, the LOS requirement for Class I highway is 8 h. Standards of this type also make sense from the users' point of view, as it would understandably be too cost prohibitive to maintain a facility in a bare pavement condition at all times and under all types of events. Another consideration in setting LOS targets is the need to strike a balance between costs and benefits. An ideal LOS policy for a parking lot should take into account the types of snow events that are to be expected in the area where the site is located as well as the service demand of the parking lot/establishment type (e.g., shopping plaza, restaurants, emergency buildings, etc.). Also, the public could reasonably expected to be able to travel at high speed at all times on Class I highways but should not have the same expectation for a residential, urban street. To salt urban streets and keep them with bare pavement in the winter should probably not be a goal in northern states. Rather, if safety rather than automobile mobility is a primary objective, it should be a priority always to keep sidewalks and crosswalks bare and they should be treated before the automobile travel lanes are plowed or salted. That should be in the municipal plan as it typically is in the Nordic countries.

Determining Snow and Ice Control Treatment

In responding to any upcoming weather events, maintenance operators must choose the right winter maintenance treatment, such as anti-icing or deicing and the application rate for each treatment. If an anti-icing operation is to be implemented, solid or liquid salt should be selected on the basis of LOS need and the expected event conditions. The recommended application rate should then be applied before the event begins.

For deicing operations, a number of weather conditions should be determined at the time of treatment. First, the average representative pavement surface temperature should be either measured or estimated on the basis of air temperature and site

conditions. Second, the type of snow and the total amount of the snow (depth) that is expected to accumulate during the event should be determined. The snow accumulation should include snow already present on the ground in addition to the forecasted precipitation. Finally, the desired LOS should be determined in terms of bare pavement regain time, based on the LOS requirements of the facility.

The Snow and Ice Control Guidelines for Materials and Methods developed by National Cooperative Highway Research can be consulted for highway salt application rates, whereas *Deicing Performance of Road Salt: Modeling and Applications*—an article published by the Transportation Research Board can be considered for determining optimal application rate for parking lots, sidewalks, and low volume roads (Blackburn et al., 2004; Hossain et al., 2014).

References

- Blackburn, R.R., Bauer, K.M., Amsler, D.E. Sr., Boselly, S.E., Mcelroy, A.D., 2004. Snow and ice control: guidelines for materials and methods. National Cooperative Highway Res. Prog. Transp. Res. Board. Report No 526.
- Colleen McGovern, Peter Rusch, David A.N., 2011. State Practices To Reduce Wet Weather Skidding Crashes. Federal Highway Administration.
- Garber, G.J., Hoel, L.A., 2015. Traffic and Highway Engineering. Fifth Ed. Cengage Learning Stamford, CT.
- Hall, J.W., Smith, K.L., Titus-Glover, L., Wambold, J.C., Yager, T.J., Rado, Z., 2009. NCHRP web-only document 108: guide for pavement friction. National Cooperative Highway Res. Prog. Transp. Res. Board.
- Hossain, S.M.K., Fu, L., Lu, C.Y., 2014. Deicing performance of road salt: modeling and applications. Transp. Res. Rec. 2440 (1), 76–84.
- Hossain, S.M.K., Fu, L., Law, B., 2015. Winter contaminants of parking lots and sidewalks: friction characteristics and slipping risks. J. Cold Reg. Eng. 29 (4), 4014–4031.
- Hossain, S.M.K., Muresan, M., Fu, L., 2018. Application guidelines for optimal de-icing and anti-icing. In: Shi, X., Fu, L., (Eds.), Sustainable Winter Road Operations. Wiley Online Library, pp. 443–471.
- Guerre, J., Groeger, J., Van Hecke, S., Simpson, A., Rada, G., Visintine, B., 2012. Improving FHWA's Ability to Assess Highway Infrastructure Health Pilot Study Report. Federal Highway Administration.
- Lund, J., 1984. Accidental falls at work, in the home and during leisure activities. Journal of Occupational Accidents 6, 181–193.
- Miller, J.S., Bellinger W.Y., 2014. Distress Identification Manual for the Long-Term Pavement Performance Program. Federal Highway Administration, Report No. FHWA-HRT-13-092.
- Sayers, M.W., 1995. On the calculation of international roughness index from longitudinal road profile. Transp. Res. Rec. 1501, Transp. Res. Board, 1995, pp. 1–12.
- Simita, B., Hashemian, L., Hasanuzzaman, M., Bayat, A., 2016. A study on pothole repair in Canada through questionnaire survey and laboratory evaluation of patching materials. Can. J. Civil Eng. 43 (5), 443–450.
- Start, M.R., Kim, J., Berg, W.D., 1998. Potential safety cost-effectiveness of treating rutted pavements. Transp. Res. Rec. 1629 (1), 208–213, doi: 10.3141/1629-23.

Further Reading

- Hauer, E., Terry, D., Griffith, M., 1994. Effect of resurfacing on safety of two-lane rural roads in New York State. Transp. Res. Rec. 1467, 30–37.
- Johnson, A., 2000. Best Practices Handbook on Asphalt Pavement Maintenance. Minnesota Technology Transfer Center, LTAP Program, Center for Transportation Studies, Minnesota Department of Transportation, MN, USA.
- Ketcham, S., Minsk, L.D., Blackburn, R.R., Fleege, E.J., 1996. Manual of Practice for an Effective Anti-icing Program: A Guide for Highway Winter Maintenance Personnel. Federal Highway Administration, Report No. FHWA-RD-95-202.
- Shi, X., Fu, L., 2018. Sustainable Winter Road Operations, Wiley-Blackwell. Available from: <https://www.wiley.com/en-us/Sustainable+Winter+Road+Operations-p-9781119185062>.
- Tighe, S., Li, N., Falls, L.C., Haas, R., 2000. Incorporating road safety into pavement management. Transp. Res. Rec. 1699 (1), 1–10.

Safety of Roundabouts

Khaled Shaaban, Associate Professor, Utah Valley University, Orem, UT, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	539
Modern Roundabouts	539
Flower Roundabout	541
Turbo Roundabout	543
Pedestrian Safety	544
Bicyclist Safety	545
Conclusion	547
References	547

Introduction

At-grade intersections may cause major bottlenecks in any roadway network system. Poorly designed intersection not only creates crashes but also leads to vehicular emissions and congestions. Crashes at intersections differ depending on the intersections' design and type. The design of an intersection should provide a safe and convenient environment for vehicles, bicycles, and pedestrians traveling through the intersection. Moreover, the design should aim to reduce the severity of potential conflicts between the users. There are two main types of intersections: grade-separated and at-grade intersections. In the case of at-grade intersections, two or more roadways cross at the same elevation. A roundabout is one type of at-grade intersection. For the grade-separated intersections, typically two roadways cross each other at different levels, but unless both roads are motorways (freeways), there are usually at-grade intersections between the minor road and ramps from the major road and those intersections can be yield-controlled, stop-controlled, signalized, or roundabouts.

Roundabouts rely on a physical barrier in the form of a central island to manage traffic compared to using only a regulatory sign or a traffic signal. Intercepting roads feed traffic into a one-way circulatory roadway that surrounds the central island. In this case, traffic is forced to move in one direction in the circulatory roadway, which typically is safer and more efficient than other types of intersections where a driver can run a stop sign or a traffic signal. All movements that drivers make when entering, while inside, and when leaving the roundabout are right turns, which is safer, and cause less delay than left turns. To make a U-turn, drivers have to circle all the way around the central island and exit the roundabout on the same street they entered it from.

As a treatment for the traditional intersection, roundabouts offer a safe, efficient, and economical solution to traffic signals in most situations if adequately designed (Elvik, 2003; Hydén and Várhelyi, 2000). Often, they are also less expensive to construct and maintain than signalized intersections. The maintenance cost is mainly related to standard roadway maintenance in addition to lighting and landscaping. The central island also provides space for landscaping, which is more esthetically appealing than a traditional intersection design. However, traditional modern roundabouts often require more right-of-way at the intersection itself than other types of control, which can increase the construction cost substantially if not available at an existing intersection. On the other hand, roundabouts keep traffic moving all the time, so sections between the intersections can be made narrower, saving right-of-way costs and construction costs, especially where several roundabouts are built in series along a so-called roundabout corridor.

After having been implemented and proven successful at numerous locations across the world, roundabouts have been incorporated as a standard design in many countries as a solution to congestion, high vehicle and pedestrian crash rates/counts, and a low-maintenance cost solution to the traditional signalized intersections. Today, after many years since the development phase, there are different types of roundabouts used around the world. Some types are already in frequent use in many countries; some of them are recent and used in specific countries, and others are still in the development phase. Therefore, there are still different points of view regarding the ideal roundabout from the perspective of safety. In this article, the description, advantages, and disadvantages of the most common types of roundabouts will be discussed.

Modern Roundabouts

Roundabouts have existed since before there were any automobiles, and they were typically built to act as landmarks. The first roundabout built with an intent to make automobile traffic flow smoother and safer is probably Sollershoth Circus in Letchworth Garden City, the United Kingdom, which was built in 1909. The inscribed circle diameter is 33 m (108 ft). When first built, there was no centerpiece beyond a grassy circular island, so this was not a case of locating a civic ornament, but just traffic engineering. Today, lighting and bushes have been added to the central island. However, with no one-way rule and no traffic priority, it was not exactly a modern roundabout at first but functioned well since traffic volumes were low, and the United Kingdom had a national speed limit of 20 mph (32 km/h) at the time.

The modern roundabouts are said to have been invented in the United Kingdom in order to eliminate the problems associated with larger traffic circles as traffic volumes increased and traffic circles experienced congestion because of traffic in the circle yielding to entering traffic and a high number of crashes because of high speeds. In the United Kingdom, they standardized yield at entry in November 1966. Sweden followed with a universal yield at entry in September 1967, and several other countries changed priority around that time as well. But in the United States, different states had conflicting rules, and there are still circular intersections in the United States where entering traffic has priority. In the United States, “standard” roundabouts are since the early 1990s called modern roundabouts to avoid confusion with the larger traffic circles and rotaries.

Modern roundabouts, like older traffic circles, are designed as one-way circulating roadways with channelized approaches and yield control on all entries. The main difference is that modern roundabouts have geometric curvature that creates a low-speed environment. Modern roundabouts can be classified under two categories: one-lane and multilane roundabouts. For single-lane roundabouts, the design is fairly easy, but to create a low-speed environment for multilane roundabouts has proven to be a challenge. Over 90% of roundabouts in France, Germany, the Netherlands, and Scandinavia are single-lane roundabouts and, in most of these countries, there are now more roundabouts than signalized intersections. In the United Kingdom, they often widen single-lane approaches to accommodate parallel vehicles in the roundabout itself, and allow “double” circulating lanes without any striping.

This is not the case for the rest of Europe or most of the United States where striping is used when allowing parallel vehicles in the circulating area. In some countries, concentric striping is still used, meaning a driver will have to change from the inside to the outside lane before exiting a roundabout from the left lane. Nowadays, typically exit striping is used instead. In this case, a driver picks a lane at entry and keeps that lane until they exit, except for if they want to make a U-turn, then a lane change is needed. This exit striping gives the modern roundabout somewhat similar characteristics to the turbo roundabout discussed later, except that many drivers do not stay within their lanes if they are separated only with striping.

As drivers enter a roundabout, they should slow down to get ready for crosswalks at the approach and then for the “intersection” itself. They should be able to see circulating traffic before they reach the yield line. When possible, vehicles merge with traffic inside the roundabout without stopping if a sufficient gap is available within the circulating traffic. Once in the roundabout, vehicles travel with the circulating traffic around the central island.

Operational speeds within a roundabout should be no more than 25 km/h (15 mph) on average, and the 90-percentile should not exceed 32 km/h (20 mph). However, many multilane so-called modern roundabouts have top speeds around 50 km/h (30 mph). The low speed should be achieved without signage indicating the speed limit within the roundabout, but an advisory speed limit sign recommending 25 km/h (15 mph) can be useful in areas where roundabouts are unusual. The geometric curvatures should ensure safe circulatory travel speeds. If bicycles share travel lanes in a roundabout, it is extra important to ensure that all cars travel at low speeds and speeds across crosswalks should never exceed 30 km/h (18 mph). When a vehicle approaches the required exit, in most jurisdictions, the drivers use their right-turn signal then turn right in order to exit the roundabout.

Modern roundabouts in high-speed environments in Scandinavia were often offset to the left already in the 1980s as seen in **Fig. 1**. The sharp deflection at entry ensures low entry speeds, but a potential safety issue is that the lack of deflection at the exit gives too high speed across crosswalks at exits. In the United States, such offsets have been introduced in the last decade, but the standard design in most countries is to have approaches centered toward the middle of the center island. Still, there are variants in the design of modern roundabouts.

Speeds within the roundabout are typically slower compared to a traditional signalized intersection, which gives drivers more time to handle potential conflicts. Moreover, in the case of a crash, the severity of damages and injuries is much lower than in high-

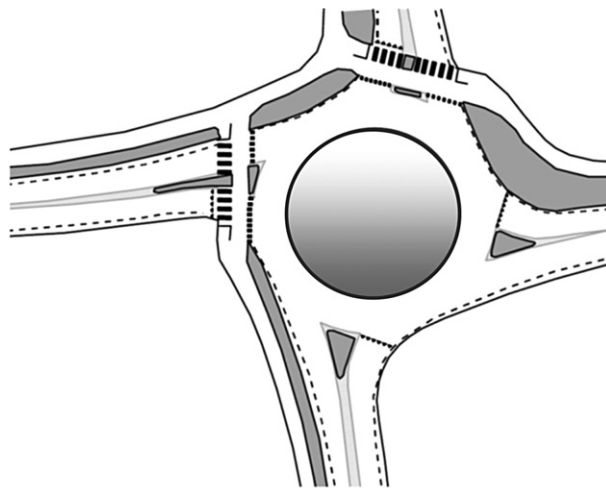


Figure 1 Typical Scandinavian roundabout.

speed crashes. Furthermore, since vehicles travel in the same direction through a roundabout, some of the most serious crashes that occur at signalized intersections (right-angle, left-turn, and head-on) are virtually eliminated.

Well-designed modern roundabouts have a good safety record. A study showed that the construction of roundabouts resulted in a reduction in the number of injury crashes in Australia (45%–87%), France (57%–78%), the United Kingdom (25%–39%), and the United States (51%) (Robinson et al., 2000). Another study indicated that converting intersection to roundabouts resulted in a 30%–50% reduction in the number of injury crashes and a 50%–70% in fatal crashes (Elvik, 2003). A similar study showed an 80% reduction for all injury crashes and about 90% in the numbers of fatal and incapacitating injury crashes (Persaud et al., 2001). In general, roundabouts are proven to reduce the high percentage of crashes where people are seriously injured or killed.

Most of the recorded crashes at roundabouts are glancing blows at low angles of impact, reducing crash severity, and an occupant of a car or truck is seldom seriously injured at speeds below 50 km/h (30 mph). However, a typical pedestrian has a 20% chance of dying if hit at 50 km/h (30 mph), and new US data show that a 70-year old pedestrian has a 37% chance of dying if hit at that speed by the typical vehicle, and an even higher risk of dying if hit by a pickup truck or SUV where the pedestrians' center of gravity is below the hood of the vehicle. Furthermore, bicyclists and motorcyclists are prone to get injured if hit at moderate speeds.

A summary that was made on May 12, 2019, shows there so far (from 1990 to 2019) have been 81 fatal crashes at modern roundabouts in the United States, killing 89 people. One pedestrian and two bicyclists have been killed, but at least 30 motorcyclists are victims. A number of the victims are shown as unknown type, but motorcycles make up over a third of all fatalities. Motorcyclists can get killed at low speeds, and motorcyclists also sometimes enter roundabouts at speeds way above the intended speed, losing control. In other words, roundabouts typically improve safety performance compared to other types of control, but in order to be safe, it is essential to ensure that everybody travels slowly. Because of their low-operating speeds, single-lane roundabouts are used as a traffic-calming device in many countries.

The conflict points between vehicles for a traditional four-leg intersection of two-lane roads are shown in Fig. 2. Traditional intersections have 32 vehicle/vehicle conflict points, including 8 merging points, 8 diverging points, and 16 crossing points. Modern roundabouts only have eight total conflict points, including four merging points and four diverging points. The crossing movement is eliminated as a potential conflict due to the central island. This is a 75% reduction from traditional four-way intersections. In the case of a three-leg intersection, the number of conflict points decreased from nine to six in the case of a roundabout, which is a 33% reduction. The reduction in the number of conflict points is an indication of few possibilities for crashes. This is one of the main reasons that a one-lane roundabout could offer safety conditions than a stop-controlled intersection under the same conditions. Furthermore, having few automobile conflict points gives the driver more time to look for pedestrians, especially if the driver is also slowed down to a lower speed than at a traditional intersection.

In the case of roundabouts with two circulating lanes, there are more conflict points compared to one-lane roundabouts, especially if lane changing occurs inside the roundabout, and dual entry and exit lanes may also cause confusion. In this case, there are 24 vehicle/vehicle conflict points: 8 crossing points, 8 diverging points, and 8 merging points. By replacing a traditional intersection by a two-lane roundabout, the number of conflict points is reduced from 46 to 24, 48% (Fig. 2). This may be a reason why safety appears to be better in one-lane roundabouts compared to two-lane ones. However, a more important reason may be that parallel lanes allow drivers to go much faster, especially if a driver enters in the right-hand lane, cuts into the left lane at the central island, and then again exits in the right lane. That is possible if the lanes are separated by only striping, and there is no vehicle in the other lane.

Flower Roundabout

In spite of the good performance levels of modern roundabouts, especially multilane ones have shown safety problems mostly associated with the behavior of the drivers at the entries, inside the roundabout, and at the exits, such as disregarding lane markings, improper lane changing, and speeding. Such behaviors can cause safety problems and increase the risk of crashes. Researchers have resolved the problems of the modern two-lane roundabouts by adopting new types of roundabouts. One such type is the flower roundabout (Fig. 3), which only has one circulating lane (Tollazzi et al., 2011). This is more or less the only design used in Denmark when traffic volumes require more than one lane.

This type of roundabouts can have two entry and exit lanes outside the roundabout, but it has only one circulating lane. There is an additional right-turn lane that directs right-turning vehicles to turn right separate from the circulating traffic. The capacity is lower than for a double-lane roundabout unless a high percentage of traffic turns right. A related design that increases the capacity is to put separate right-turn slip lanes on a roundabout with two circulating lanes. The safety benefit of that is questionable, especially if there are high pedestrian volumes.

An advantage of the flower design is the possibility of implementation on any existing roundabout. If the before design was a single-lane roundabout, and it has long delays, adding slip lanes can help. If the before design had two circulating lanes, the paved circulatory area would have to be narrowed, and the outer lane is changed to a right-turn lane only. These modifications will reduce capacity but will increase safety.

In the case of a roundabout with one circulating lane, the central island and intercepting roads remain in the same positions, and a right-turn lane is created by adding a new lane or separating an existing outside lane, from the inside lane by a splitter island. In this case, the inside circulating lane is used by vehicles desiring to go straight, turn left, or make a U-turn. If a vehicle incorrectly stays in the inside lane when entering the roundabout, it can still turn right at the first exit as long as the slip lane has its own exit lane. This

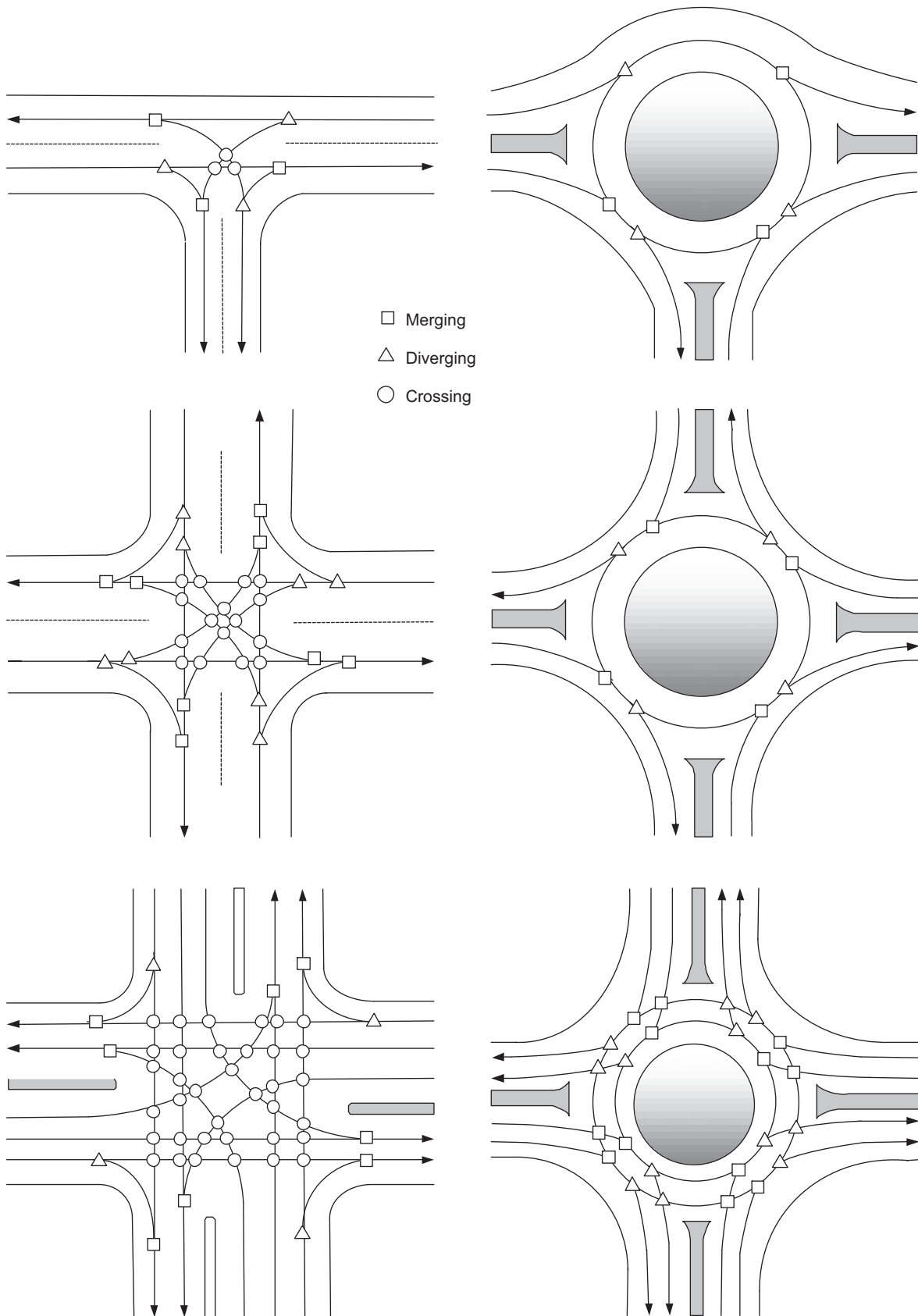


Figure 2 Vehicle/vehicle conflict points for different types of intersections versus modern roundabouts.

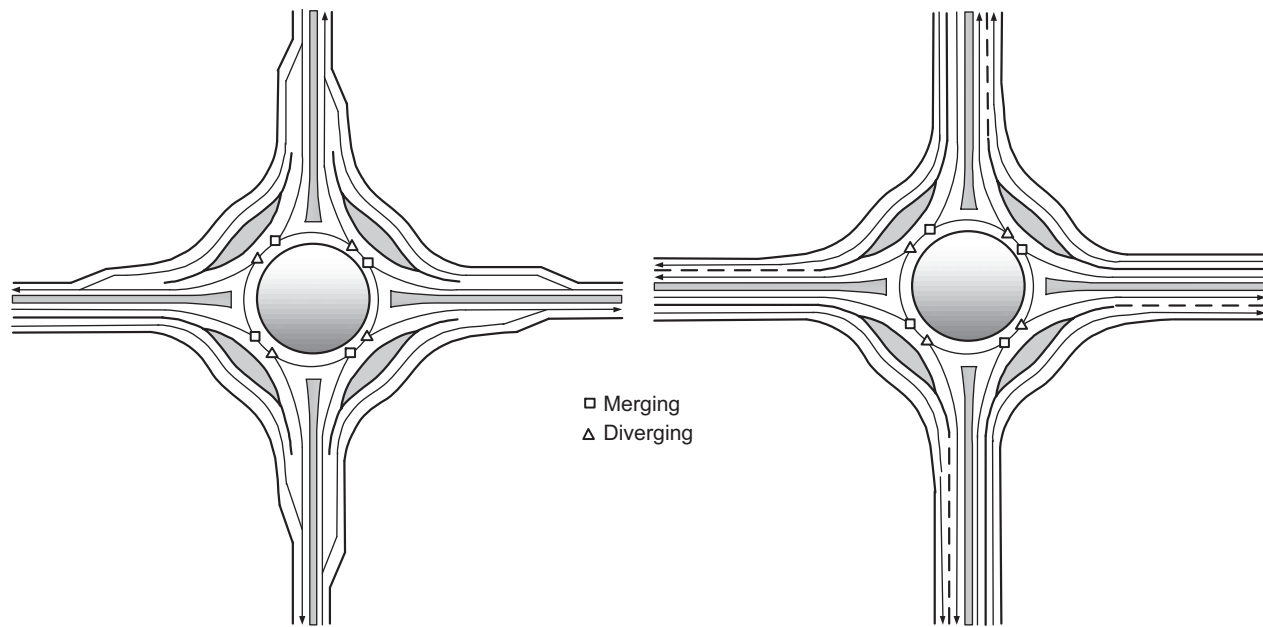


Figure 3 Vehicle/vehicle conflict points for the flower roundabout for one- and two-lane roads.

should not be allowed if the design follows the left graph of [Fig. 3](#) because drivers in the slip lane look for traffic already in the roundabout and do not plan to yield to right-turning traffic entering from the lane that was parallel to them on the approach, resulting in sideswipe crashes at the merge area.

When separating the right-turning movement from the other movements of what used to be a roundabout with two circulating lanes, the two-lane roundabout becomes a one-lane roundabout. All crossing conflicts and weaving conflicts are eliminated. Therefore, there are only eight vehicle/vehicle conflict points: four diverging points and four merging points similar to a traditional one-lane roundabout. However, the diverging conflict points concerning the right-turning movement and the merging points with the flow from the roundabout result in an additional four diverging and four merging conflict points that are introduced before and after the roundabout. The total number of conflict points for this case is 16, which is still a better design compared to the traditional roundabout with 24 conflict points.

Turbo Roundabout

One of the new types of roundabouts adopted recently in many countries is the turbo roundabout. In 1997, this type of roundabout was developed in the Netherlands in order to solve the weaving problems of the two-lane roundabout ([Fortuijn and Harte, 1997](#); [Fortuijn, 2009a, 2009b](#)). The concept is based on physically separating the vehicles before entering the roundabout while circulating within the roundabout, and while exiting from the roundabout. The separation of the traffic lanes disturbed only the entry point to the inner circulating lane. The separation is achieved by the installation of raised splitters (delineators) to restrict lane changing within the roundabout and eliminate the weaving conflict points.

These geometric features of the turbo roundabouts force the drivers to select the proper lane before approaching the roundabout. They also prevent any lane changing at the entry, exit, or within the circulating lanes. These restrictions lower the number of potential conflict points during the roundabout crossing path, lower driving speed near and through the roundabout because of the raised lane dividers, prevent weaving maneuvers, and lower the risk for sideswipe crashes. All these factors cause the safety level to increase significantly.

In the case of a traditional two-lane roundabout, drivers can ignore lane markings, drive in an almost straight path, and sustain their approach speed. A turbo roundabout can increase the safety level of multilane roundabouts by eliminating some of these problems. The raised lane dividers force drivers to remain in the correct lane and to follow paths with smaller radius at reduced speeds. These restrictions result in a lower number of conflict points. An intersection between a major road with two-lane entries and exits and a minor road with two-lane entries and one-lane exits on the minor road will result in 14 vehicle/vehicle conflict points: 4 crossing points, 6 merging points, and 4 diverging points ([Fig. 4](#)) compared to 24 conflict points for a two-lane roundabout. Although turbo roundabouts lead to few traffic conflicts, the traffic conflicts that do occur are more severe ([Vasconcelos et al., 2014](#)).

One negative feature of the turbo roundabout is that a driver failing to yield when entering in the inner lane can be hit on the left side of their car almost perpendicularly, or they may hit the right-hand side of the circulating vehicle. However, since the entry

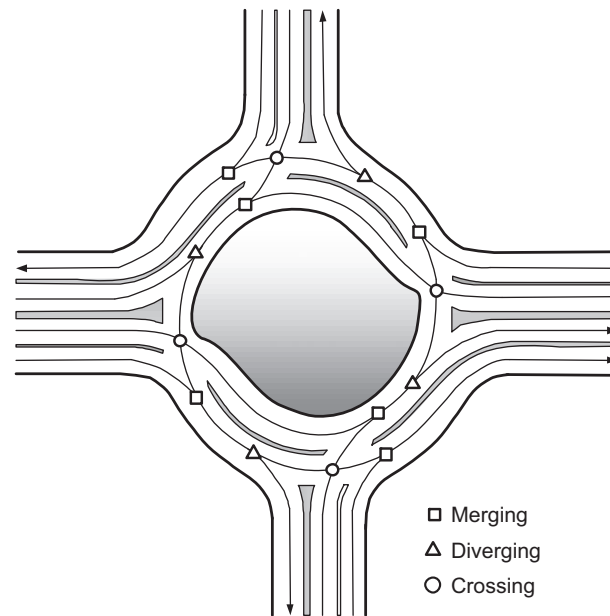


Figure 4 Vehicle/vehicle conflict points for the turbo roundabout.

geometry is very tight with a curvature guaranteeing low speed, there should be minimal risk of injury as long as one of the vehicles is not a motorcyclist or bicyclist. In the Netherlands, they, therefore, avoid this design if bicycles do not have their own separate lanes.

Pedestrian Safety

Pedestrian safety at roundabouts is considered high. The reduced speeds of the vehicles at roundabouts provide better crossing opportunities for pedestrians. In addition, the availability of a splitter island helps to provide a refuge area for pedestrians in order to focus on and cross one side of the road at a time instead of crossing the whole approach as in the case of un-signalized intersections. In addition, pedestrians can cross immediately in countries where pedestrians have the right-of-way in crosswalks, or as soon as a proper gap is available in countries where drivers have or take priority. Compared to waiting at a signalized intersection, this reduces the delay for the pedestrians. However, it also increases the delay for drivers, especially if the crosswalks at the roundabouts are signalized. Furthermore, the location of the conflicts between the vehicles and pedestrians occurs generally one car-length before the roundabout, which means the pedestrian does not have to walk in front of a driver stopped at the yield line, looking for traffic from their left. This results in few places for pedestrians to check for conflicting vehicles at the entry, whereas crosswalks at the exit are more similar to those at conventional intersections, but with traffic from only one direction and at a lower speed.

In a good design, a mid-block crosswalk is added to reduce the impact of pedestrian traffic on the functionality of the roundabout itself and increasing the traffic capacity when compared to their signalized counterparts. This type of design can improve pedestrian safety since pedestrian traffic interferes less with the operation of the intersection. In some cases, an island is employed, allowing pedestrians to cross one lane at a time. In general, pedestrians see a significant reduction in crashes at roundabouts. Studies in different countries have yielded similar results. On average, converting a traditional stop sign controlled intersection or a signalized intersection to a modern roundabout results in a 30%–50% reduction in pedestrian crashes. A Swedish study of “all” modern roundabouts in the country around the year 2000 showed that single-lane roundabouts are about 80% safer than traditional intersections, whereas multilane roundabouts have similar safety level as signalized intersections.

However, there are also some disadvantages to pedestrians. Drivers approaching roundabouts, which are yield-controlled, may slow down but not come to a full stop, and if a driver does not see a pedestrian, they may not stop for them. Therefore, pedestrians cannot be sure that vehicles will not pass over the crosswalk at a certain time. Additionally, the total walking path for pedestrians in the case of a multilane roundabout may be longer than at a signalized intersection since lanes often are 4.25 m wide rather than 3.5–3.75 as common at signalized intersections with bulb-outs. However, many signalized intersections have three or more approach lanes and sometimes even a parking lane extending all the way to the crosswalk. Moreover, visually impaired pedestrians may find it challenging to find the location of the crosswalks due to being generally located far from the usual location at a traditional intersection. Measures such as pedestrian signalization or a raised crossing may be needed in some cases. In general, if approach speeds for the fastest motor vehicles exceed 30 km/h (20 mph) the crosswalks should be raised, but that is the case for traditional non-signalized intersections as well.

Roundabouts reduce the number of potential conflict points for pedestrians for certain movements compared to other types of intersections. These conflicts include conflicts of pedestrians with high-speed vehicles, conflicts of pedestrians with right-turning

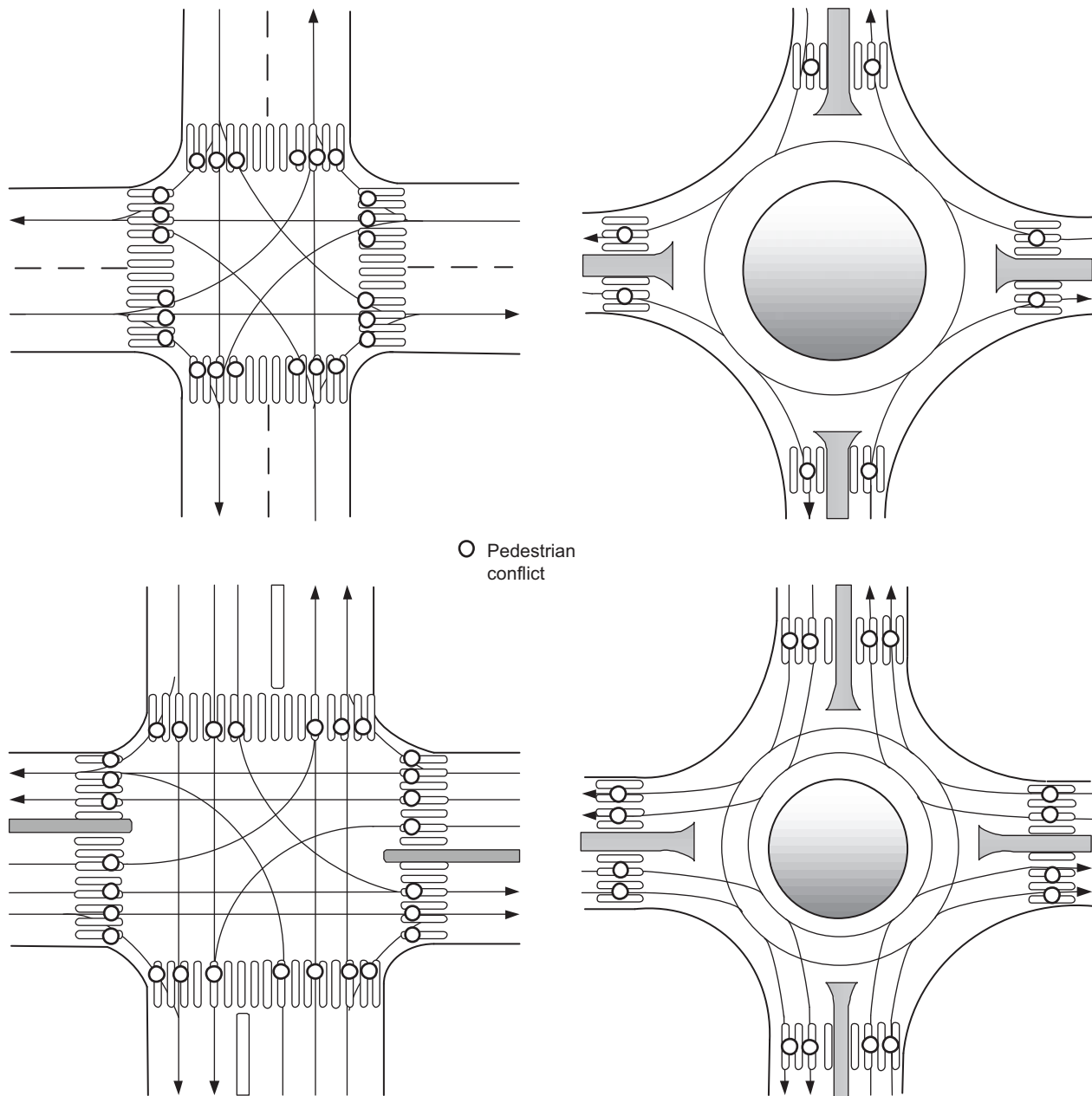


Figure 5 Vehicle/pedestrian conflict points for four-leg intersection with one- and two-lane approaches.

vehicles, and conflicts of pedestrians with left-turning vehicles. One-lane roundabouts produce 8 vehicle/pedestrian conflict points compared to 24 for a traditional intersection, and two-lane roundabouts produce 16 vehicle/pedestrian conflict points compared to 28 for a traditional intersection (Fig. 5).

Bicyclist Safety

Safety improvements at roundabouts are most evident in the case of vehicles and pedestrians. However, for bicyclists, it depends on the bicycle design treatment at the roundabout. In general, available studies do not indicate that roundabouts provide improved bicyclist safety compared to intersections that employ traditional traffic control, especially when considering multilane roundabouts. In the United Kingdom, the involvement of bicyclists in crashes at roundabouts was found to be 10–15 times higher than the involvement of car occupants after considering the exposure rates (Brown, 1995).

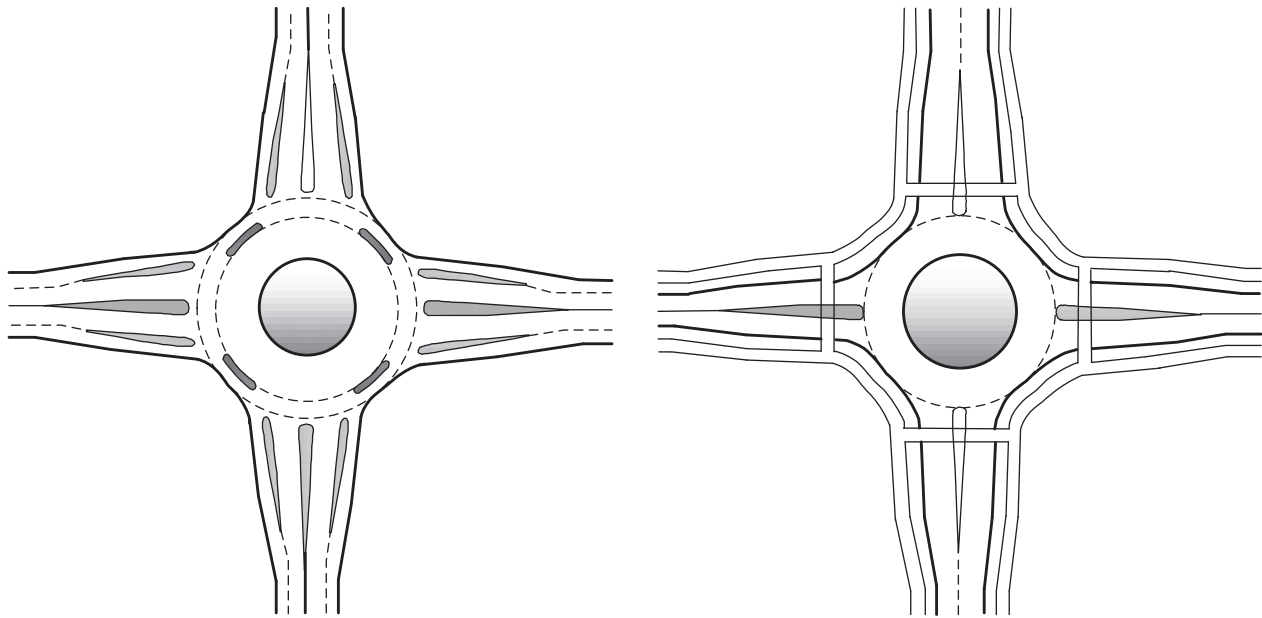


Figure 6 Roundabout with unprotected bike lanes/bike track.

There are different alternatives to deal with bicyclists at roundabouts. One alternative handling a bicyclist is to have it be treated like a vehicle. In this case, no bike lanes or tracks are provided, and bicyclists use the same lanes as other vehicles through the roundabout. As a result, they have to deal with the same potential conflicts as other vehicles. This option can be used in the United States since bicycles in most states are considered to be vehicles. In this case, bicyclists ride through the roundabout the way other drivers do. They should use the full lane, just like any other vehicle. Accordingly, drivers of motor vehicles have to yield to bicycles when entering, as if bicycles were any other vehicle on the road.

In this scenario, crashes are often due to a bicyclist misusing the roundabout and staying at the right edge of the roundabout and riding alongside vehicles. As a result, they risk having a collision when they circulate by an exit lane. Another crash scenario occurred when a vehicle is trying to pass bicyclists as if they were in a bike lane. This alternative is more suitable for one-lane roundabouts in low-volume environments, as speeds are reduced and the number of conflict points is low. In the case of multilane roundabouts, bicyclists may need another treatment.

For this alternative, if the bicyclists are not comfortable to go through the roundabout on their bikes as a vehicle, they should be allowed to walk or ride their bicycles on the sidewalk. To accommodate this type of movement, a ramp can be installed when approaching and leaving the roundabout to give bicyclists easy access to the sidewalk. In this case, in most jurisdictions, a bicyclist is required to get off his or her bicycle and to cross the travel lanes as a pedestrian. As a result, they have to deal with the potential conflicts for pedestrians.

A second alternative is to have an unprotected bike lane. This type of bike lane is provided in the roundabout adjacent to the travel lanes. In this case, bicyclists and vehicles have separate lanes without a physical barrier except at the four corners (**Fig. 6, left**). This method causes high-risk conflicts between bicyclists inside the roundabout and vehicles turning right to exit. Regarding the effects on all injury crashes, this alternative is considered the worst ([Daniels et al., 2008](#)). In the case of unprotected bike lanes, bicyclists have to give way to vehicles entering and exiting the roundabout. Therefore, bicyclists have minimal effect on the capacities and delays at the roundabout.

This alternative, riding on the edge of a roundabout, is potentially dangerous for bicyclists because it introduces a scenario where vehicles, especially large trucks, can pass them and crowd them off the lane. Another common type of cyclist-car crash occurs when a vehicle either enters or exits from the roundabout without yielding to a bicycle already in the roundabout.

A third alternative (**Fig. 6, right**) is to construct a separate bike track, which is a path that runs separately from the roundabout. The bike track is built outside the intersection, forcing bicycles to cross the path of vehicles outside the roundabout. It can be separated from the travel lane and sidewalks using curbs. In the case of a low pedestrian volume, striping may be enough to separate the bike lane from the pedestrians. Separating a bike track from vehicles using striping only is still used in some places, especially in the United States. In such cases, bike lane delineators (plastic poles with reflectors) are usually added to prevent drivers from driving in the bike lane and to improve visibility in the case of poor lighting or bad weather.

To accommodate crossing, ramps are installed to allow bicyclists to cross the different approaches when circulating the roundabout. Consequently, the number of conflicts between bicyclists and vehicles should decrease as both bicyclists and vehicle drivers are better aware of the presence of each other at the roundabout. In some countries, bicyclists crossing an approach have the priority over the drivers entering and exiting the roundabout, which may affect the capacities and delays at the roundabout.

A fourth alternative is to provide bicyclists with grade-separated bike paths in the form of bike tunnels or bridges. This method is considered the most expensive solution. However, it ensures the elimination of any conflicts between bicyclists and vehicles. Furthermore, bicyclists have no effect on the capacities and delays at the roundabout. This is the standard solution outside urban areas in Scandinavia.

In summary, if providing grade separation is not feasible, the safest approach for accommodating bicycle travel at a roundabout is a bike track, especially in countries where bicycle volumes are high such as Sweden and the Netherlands. In this case, clear yield lines should be provided not only at the roundabout entry but also just before the bike track, both on the entry and on the exit lanes to remind drivers to yield. However, due to the right-of-way needed, this is not a possible option at all roundabout locations.

In countries with low bicycle traffic volumes such as the United States, the primary recommendation is not to have bike lanes at a roundabout. In this case, the bike lanes are terminated some distance prior to the roundabout, and bicyclists have the option to use the travel lane as a vehicle or use the sidewalk if not comfortable to use the travel lane.

Conclusion

Roundabouts are characterized with significant design features to reduce vehicle speed such as the use of a central island, the right angle connection between roadways and circular roadways, and the right of way traffic arrangement, which serves to reduce crash rates and to reduce speed significantly in the area of the intersection. Compared to other types of intersection control, roundabouts eliminate some conflict types, reduce conflict points, and oblige drivers to decrease their speed through the roundabout.

Converting a signalized intersection into a roundabout typically results in a considerable decrease in the number of crashes and specifically crashes with injuries. This is due to the reduction in the number of conflict points, reduction in speeds, which are typically higher at signalized intersections, and the safer environment for pedestrians since they have few directions to check for vehicles than they do at signalized and non-signalized traditional intersections. Similarly, one-lane roundabouts produce greater safety benefits than multilane roundabouts due to the few potential conflicts between road users, shorter crossing distances for pedestrians, and on average lower speeds.

The modern roundabouts are the most commonly used type of roundabout. In recent years, different studies pointed out that there is a need to improve the traffic safety characteristics and to increase the capacity of modern roundabouts. For this reason, there was a need for different solutions to improve these existing roundabouts. Nowadays, many additional types of roundabouts are widely used at different levels within the road network. The flower and turbo roundabouts are two other types of roundabouts proposed to overcome some of the shortcomings of the modern roundabouts.

From the design perspective, the most straightforward design is the modern roundabout followed by the flower roundabout then the turbo roundabout, which requires a more sophisticated design. The main advantage of a flower roundabout is the possibility of building it within an existing modern roundabout without the need to modify the central island or splitter islands, unlike the turbo roundabout. From the capacity perspective, modern two-lane roundabouts can represent an appropriate design solution when high capacity is needed. The flower roundabout can be beneficial when the right-turning volume is high. The turbo roundabout has the advantage of accommodating high traffic volumes for specific directions using different designs of the circular island.

From the traffic-safety perspective, the safety levels of the multilane modern roundabout are considered the lowest. The flower roundabout has some advantages over the turbo roundabout. When converting a modern roundabout to a flower roundabout by separating the right-turning movement from other movements, a two-lane roundabout is transformed into a one-lane roundabout. Consequently, there are less merging and diverging conflict points, no weaving conflict points as in modern roundabouts, and no crossing conflicts as in turbo roundabouts. The turbo roundabout also offers high levels of safety by improving deflection and reducing the number of conflicts.

There are indications that pedestrians are safer at roundabouts compared to other types of intersections if pedestrian crosswalks are provided around the perimeter of the roundabout. The number of conflict points is less, splitter islands allow pedestrians to cross one direction at a time, and the low vehicle speeds through a roundabout allow more time for drivers and pedestrians to react to one another. As a result, few pedestrian crashes are usually found at roundabouts. For bicyclists, there is no clear evidence of the safety benefits for roundabouts over other types of intersections. Furthermore, there is no final evidence regarding the safety benefits of the different types of bicycle treatment. However, there are indications that roundabouts with bike tracks are safer than roundabouts with mixed traffic or unprotected bike lanes.

References

- Brown, M., 1995. The design of roundabouts. Transport Research Laboratory State-of-the-Art-Review. Her Majesty's Stationery Office, London, UK.
- Daniels, S., Brijs, T., Nuyts, E., Wets, G., 2008. Roundabouts and safety for bicyclists: empirical results and influence of different cycle facility designs. Presented at the 2008 TRB National Roundabout Conference, Kansas City, MO.
- Elvik, R., 2003. Effects on road safety of converting intersections to roundabouts: review of evidence from non-US studies. *Transp. Res. Rec.* 1847 (1), 1–10.
- Fortuijn, L., Harte, V., 1997. Multi-lane roundabouts: exploring new models. Traffic Engineering Working Days. CROW, The Netherlands.

- Fortuijn, L.G., 2009a. Turbo roundabouts: design principles and safety performance. *Transp. Res. Rec.* 2096 (1), 16–24.
- Fortuijn, L.G., 2009b. Turbo roundabouts: estimation of capacity. *Transp. Res. Rec.* 2130 (1), 83–92.
- Hydén, C., Várhelyi, A., 2000. The effects on safety, time consumption and environment of large scale use of roundabouts in an urban area: a case study. *Acc. Anal. Prev.* 32 (1), 11–23.
- Persaud, B.N., Retting, R.A., Garder, P.E., Lord, D., 2001. Safety effect of roundabout conversions in the United States: Empirical Bayes observational before-after study. *Transp. Res. Rec.* 1751 (1), 1–8.
- Robinson, B.W., Rodegerdts, L., Scarborough, W., Kittelson, W., Troutbeck, R., Brilon, W., Mason, J., 2000. Roundabouts: an informational guide. Report FHWA-RD-00-067. FHWA, U.S. Department of Transportation.
- Tollazzi, T., Renčelj, M., Turnšek, S., 2011. New type of roundabout: roundabout with “depressed” lanes for right turning—“flower roundabout”. *Promet-Traffic Transp.* 23 (5), 353–358.
- Vasconcelos, L., Silva, A.B., Seco, Á.M., Fernandes, P., Coelho, M.C., 2014. Turboroundabouts: multicriterion assessment of intersection capacity, safety, and emissions. *Transp. Res. Rec.* 2402 (1), 28–37.

Rumble Strips, Continuous Shoulder, and Centerline

Anna Anund^{*,†}, Anna Vadeby^{*,†}, ^{*}Swedish National Road and Transport Research Institute, Linköping, Sweden; [†]Rehabilitation Medicine, Linköping University, Linköping, Sweden

© 2021 Elsevier Ltd. All rights reserved.

The Intention With Rumble Strips	549
Different Designs and Placements of Rumble Strip	549
Rumble Strips Effect on Driver Behavior and Driver Performance	550
The Impact of Rumble Strips	551
Traffic Safety	551
Negative Effects of Rumble Strips	552
External Noise	552
Surface	552
Effect on Safety and Comfort for Vulnerable Road Users	552
Maintenance	553
Conclusions	553
References	553

The Intention With Rumble Strips

Driver fatigue, including both sleepiness-related and task-related underload and overload situations, is considered as a contributing factor in approximately 20% of all vehicle crashes, the view is the same for inattention. Rumble strips in the center of two-lane rural roads and on the shoulders of motorways as well as two-lane roads are a countermeasure aimed to help drivers who unintentionally are leaving the lane, for example, due to sleepiness or inattention.

The thump of the rumble strip is aimed to give the driver feedback in time to be able to react and take action to avoid running off the road or collide with on-coming vehicles. The rumble strip should be designed to support the drivers, without threatening them. In addition, they should give a clear feedback and not causing dangerous erratic maneuvers. They also need to be detectable for all kind of drivers like that of passenger cars, trucks, and buses; and at the same time not causing problems for riders of motorbikes, cyclists, or pedestrians.

In this chapter, we aim to summarize the knowledge about rumble strips on rural roads and their design, their effect on road users, and their general impact.

Different Designs and Placements of Rumble Strip

Rumble strips across the traveled roadway ahead of crosswalks, stop signs, and toll plazas have been used—in limited quantities—for at least 50 years. The intent of those is to get a driver's attention and signal that something unexpected is coming up. However, this article is limited to looking at more-or-less continuous rumble strips along a roadway. Experiments with such rumble strips started in 1950s and they have now been used widely in several countries for decades, especially in the United States and Canada. Sometimes they are installed in raised profile and sometimes as in-ground (milled or pressed). Variation is not only between raised versus in-ground rumble strips but also in their width, depth, length and design.

Rumble strips installed in raised profile typically has the aim to increase the visibility of the road marking during wet conditions at nighttime. The secondary effect of such installation, in terms of internal noise and vibrations, disappear after a short time and is not even possible to detect in heavy vehicles, such as trucks and buses.

Rumble strips are normally used on the shoulder or in the center of the road. The design of them might differ dependent on if they will be driven on frequently or not, since external noise is an important factor and more noise-friendly solutions are required if drivers are expected to drive across them frequently, especially if they are located near residential developments. Designs of in-ground rumbles strip causing less noise have been developed during the past few years.

In Fig. 1 two different examples of in-ground centerline rumbles strips are shown. One is the conventional one and other, with the same depth, has a sinusoidal design in order to reduce the level of external noise. They both are using a design that in Sweden is called Målilla, that is 0.35 m wide, 0.50 m center to center, and with a depth of approximately 1.0 cm.

In some countries, for example, Norway and Finland, the center of the road is equipped with two road markings with one or even two in-ground rumble strips in between them.



Figure 1 (A) Swedish Målilla design in road center. (B) Swedish sinusoidal Målilla design in road center.

The sinusoidal in-ground rumble strips are also used on shoulders. In Norway, the recommendation is to use those and apply the road marking in the rumble strip. Such a solution protects the road marking, which normally suffers from maintenance work during wintertime. The Norwegian recommendation of width of the rumble strip is 40 cm with a road marking of 10 cm or 55 cm, if the width of road marking is 15 cm (Høye, 2015).

The in-ground rumble strip generates both an internal noise and a sensation of vibrations. Most studies on internal noise have focused on passenger cars. The results show an increase in internal noise when driving on intermittent rumble strips that varies between 13 and 17 dB(A). For the sinus rumble strip, the corresponding values are 1–6 dB(A). Whether the level of signal obtained for the sinus rumble strip design is enough in order to attract the driver's attention is as far as we know not evaluated. However, results from simulator studies show that even low levels of internal noise are helpful for drivers who are about to leave the lane due to sleepiness. The sinus rumble strips provide not only noise but also vibrations. What the most important component to attract attention is not known, but it is most likely that the vibrations create an important contributing effect. There are almost no studies available on the effects on drivers of trucks and buses. In the studies done 20 years ago, it was highlighted that internal noise driving on rumble strips varies a lot depending on the type of truck or bus and for trucks it was highly dependent on the vehicle type and cargo. It was also shown that truck cabs often are so noisy that the vibration is a necessary part in waking dozing drivers, and that the rumble strip indentation must be significantly wider than the truck tire to get the full effect of vibration.

An alternative placement of rumble strips on rural roads is in the center of the lane. This has been tested on narrow roads in Sweden and also in the United States. The intention was to avoid rumble strips both on the shoulder and in the center of the road (Fig. 2). The evaluation showed that the solution itself might be good, but that the road included in the tests were too narrow (6.5 m) with major complains from users, especially motorcycle riders and truck drivers that argued the solution limited their right to use the lane and that in case of sleepiness there were too little time left to counteract the lane departure.

On motorways, and where drivers are not expected to cross them, it is common to use a deeper in-ground rumble strip. One type often seen, both in the United States and in Europe, is the so-called Pennsylvania rumble strip. This is a rumble strip with a depth of 1.2 cm, a width of 30–50 cm, a length on 13 cm with a center-to-center distance of 30 cm. They could be used both continuously or intermittent as in Fig. 3. This is the far most common design and the recommendation in, for example, the Swedish road construction directives for motorways.

Rumble Strips Effect on Driver Behavior and Driver Performance

Drivers are influenced by the rumble strips not only when passing over them but also when they are driving. In general, they seem to reduce their speed with 2–5 km/h, also increase the distance to them with 10–15 cm, and also contribute to a variability of the lateral position (Vadeby and Anund, 2017). This means that they might have an impact also when not used for alerting drivers. The increased variability on lateral position might have a negative impact on maintenance in terms of increased rutting and wear but also causing closer interactions with pedestrians and cyclists. In Norway, this has led to a recommendation to not use centerline rumble strips on roads that are narrower than 7.5 m. Additionally, in Sweden, the Road Construction Directives recommends that two-lane



Figure 2 Swedish conventional Målilla design in lane center.



Figure 3 (A) Pennsylvania design—intermittent on motorways shoulder. (B) Pennsylvania design—continuously on motorways shoulder.

roads with a width of 7 m or more should have in-ground rumble strips. If this is fulfilled, the speed limit is set to 80 km/h. Hence, the prevalence of rumble strips is highly related to the speed of the road.

A sleepy driver that hits a rumble strip will be alerted even though the rumble strip itself is not very deep. However, it must be kept in mind that the effect of the rumble strip in time is very short. Studies in simulators show that already 5 min before a rumble-strip hit on motorway, the drivers show signs of severe sleepiness. After the hit they are alerted to the same level as in the beginning of the drive, but after another 4 min of driving, they are no longer significantly alerted (Anund et al., 2008). This means that it is important to understand that hitting a rumble strip in a sleepy condition is a very late warning that require a stop for rest as soon as possible in order to avoid risky driving. In addition, there is support for any adverse effect in terms of crash migration from locations with to those without rumble strips. In other words, if motorways have shoulder rumble strips and secondary roads do not, crashes may migrate from the motorway—that typically has safe roadside areas—to the secondary road that may have trees or other obstacles near it (Smith and Ivan, 2005).

The Impact of Rumble Strips

Traffic Safety

Centerline rumble strips have been shown to reduce the number of single-vehicle crashes and crashes with oncoming vehicles, thereby reducing the number of crashes, fatalities, and injuries. The traffic safety effect may vary between different countries depending on the configurations used. A common effect is estimated to about 10% reduction of all injury crashes and 37% reduction of target crashes (head-on, single-vehicle crashes to the left etc.). The estimated effectiveness of centerline rumble strips on two-lane carriageways differs dependent on crash severity and crash type, within a range of approximately 10%–50%. In [Table 1](#), an estimate of the traffic safety effect can be found. The results are based on almost 20 studies and reported in *The handbook of road safety measures* (Elvik et al., 2009; Gårder and Davies, 2006).

Studies from the United States show a reduction in the total number of crashes by, on average, 9% and that fatal and injury crashes are reduced by about 12%. Looking at the target crashes only, a 30% reduction was seen in head-on and opposing direction

Table 1 Traffic safety effects of milled rumble strips

<i>Centre-line milled rumble strips</i>		
Crash type and injury level	Effect (%)	Confidence interval (95%)
All injury crashes	−10	(−14; −5)
Target crashes ^a , all injury	−37	(−42; −31)
<i>Centre-line and edge-line milled rumble strips</i>		
Crash type and injury level	Effect (%)	Confidence interval (95%)
All injury crashes	−14	(−23; −3)
Target crashes ^b , all injury	−32	(−36; −29)

^aHead-on, run-off-the-road to the left, side-collisions in opposite direction.

^bHead-on, run-off-the-road to the left/right, side-collisions in opposite direction.

sideswipe crashes, 25%–30% reduction in injury crashes while fatal crashes were reduced by 44% (Russels and Rys, 2005). In Western Australia, a quasi-experimental study resulted in a reduction in the all-severity crash rate of 58% and in casualty crashes of 80%. In Sweden, milled centerline rumble strips have been shown to reduce the number of severely injured in all crash types by 15% and by 24% in single-vehicle crashes. No significant changes were noted for “all injury crashes.” In Norway, installing milled centerline rumble strips resulted in a reduction in injury crashes. Head-on crashes were reduced by 32% and single-vehicle run-off road to the left by 54%.

Most of these studies consider a mixture of raised and in-ground rumble strips and include variations in the type of in-ground rumble strips as well as a variation of the environment where the road is located. These heterogeneities are confounding factors that cause limitations especially in estimating the optimal effect. A study from Norway including large-scale reports of real-world driver experiences supports the evidence that rumble strips reduce crash numbers and reduce the severity of consequences, specifically of fatigue-related driving; and they found a significant reduction in sleep-behind-the-wheel incidents resulting in road departure crashes when rumble strips were present.

Negative Effects of Rumble Strips

External Noise

In Denmark, Norway, and Sweden, measurements were made of the maximum external noise in the vicinity of rumble strips. There were major variations in results for both intermittent rumble strips and sinus rumble strips. The results show that the intermittent rumble strips provide an increase of external noise of around 2–8 dB(A). The corresponding figure for the sinus rumble strip was 0.0–4 dB(A). Further, it was found that the sinus rumble strips provide more low frequency noise (30–40 Hz) compared to the intermittent rumble strips (60–160 Hz). Maximum noise from intermittent rumble strips are obtained at around 80–90 km/h, and at 90 km/h the threshold for noise for those living close to the road is 90–140 m. It is not known at what speed the sinus rumble strip provide the maximum noise. There is reason to believe that what is perceived as disturbing is not only related to the maximum noise but rather to the fact that the sound deviates from the normal monotonous traffic noise. The rumble-strip sound is more low frequency and not continuous.

Surface

The introduction of centerline rumble strips can result in traffic confinement and cause a reduction in the amount vehicle wander laterally. This reduction is likely to increase the rate of rutting. A Swedish study investigated how annual rut development rates and rut area measurements on two-lane rural roads were affected by the introduction of milled centerline rumble strips. Comparisons between the test sections with milled rumble strips and control sections without them, showed that there were no major differences in rut development rates. The conclusions to be drawn from the results are that centerline rumble strips do not have a noticeable confining effect on traffic and have no adverse effect on the rate of rutting.

Effect on Safety and Comfort for Vulnerable Road Users

Motorcycle riders are normally sensitive to countermeasures on roads that might limit their use of the full road. However, several tests have been carried out where riders were involved and as long as rumble strips are not too deep and used as transverse strips across the road direction, there are no major complains, except if they are in the center of the lane. Studies from New Zealand and the United States has shown that the rumble strips have not been a contributing factor in crashes with motorcycles and that the rumble strips have very small impact on the stability of the motorcycle. Studies also shows that motorcyclists, in general, do not experience any significant negative effects of the rumble strips, as long as they are aware of that the road has rumble strips.

Centerline milled rumble strips can indirectly affect the safety of bicyclists due to that vehicle drivers tend to change their lateral position toward the right when there is a rumble strip in the center, leading to less available space for cyclists. This is a problem that has been raised especially in the United States where separate bicycle facilities seldom are provided, and bicyclists typically use the edge of the travel way or narrow shoulders (Gårder and Davies, 2006).

Maintenance

When rumble strips were first installed, there were worries that if they were not in-ground, the effect of them in terms of vibrations and sound would disappear after a winter of plowing snow. This was a correct reflection and an argument for the use of in-ground rumble strips.

There was also criticism that in-ground rumble strips would create a problem in snow plowing in winter when it would be difficult to clear snow in an optimal way using the wing of the plow. However, this has not been proved to be a problem. In order to extend the lifespan of rumble strips and to avoid cracking formation in the roadbed, it is usually required that the in-ground rumble strips should be sealed in connection with milling. It is difficult to find studies on cost-benefit analysis of the rumble strip taking all positive and negative impacts into consideration. However, milling or pressing in-ground rumble strips do not require any specific equipment besides the machine and is therefore a much cheaper solution as compared to, for example, guardrails or wire barriers, etc. An additional advantage is that rumble strips can also be used on more narrow roads where guardrails and wires will not be possible to fit.

Conclusions

Rumble strips save lives at a rather small cost in relation to other infrastructure-based countermeasures, they can be used on most roads including narrow ones, and they are effective for all drivers regardless of type of vehicle. The aim that rumble strips should help drivers who unintentionally are about to leave their lane, for example, due to fatigue or inattention are proven to be achieved. Despite, variations in design evaluations, they are typically shown to reduce all injury crashes by about 10% and target crashes (head-on, single-vehicle crashes to the left, etc.) by about 37%. Rumble strips contribute to a speed reduction of 2–5 km/h and an increased distance to them of 10–15 cm, something that is good for vehicle-to-vehicle interactions, but less good for pedestrians and cyclist using the shoulder. The in-ground strips do not cause problems for maintenance during wintertime or to the road construction itself and as long as they are sealed, they will remain in good shape for a long time.

References

- Anund, A., Kecklund, G., Vadeby, A., Hjalmdahl, M., Akerstedt, T., 2008. The alerting effect of hitting a rumble strip—a simulator study with sleepy drivers. *Accid. Anal. Prev.* 40 (6), 1970–1976.
- Elvik, R., Høy, A., Vaa, T., Sørensen, M. (Eds.), 2009. *The Handbook of Road Safety Measures*. Bingley Emerald Publishing, Oslo.
- Gårder, P., Davies, M., 2006. Safety effects of continuous shoulder rumble strips on rural interstates in Maine. *J. Transp. Res. Board* 1953, 156–162.
- Høy, A., 2015. *The Handbook of Road Safety Measures*, Norwegian (online) version. Retrieved from: <https://tsh.toi.no/index.html?147175>.
- Russels, E., Rys, M., 2005. *Centerline Rumble Strips—A Synthesis of Highway Practice*. Transportation Research Board, Washington DC.
- Smith, E., Ivan, J., 2005. Evaluation of safety benefits and potential crash migration due to shoulder rumble strips installation on Connecticut freeways. *J. Transp. Res. Board* 1908, 104–113.
- Vadeby, A., Anund, A., 2017. Effectiveness and acceptability of milled rumble strips on rural two-lane roads in Sweden. *Eur. Transp. Res. Rev.* 9(2), 29.

Safety Culture

Tor-Olav Nævestad, Institute of Transport Economics, Oslo, Norway

© 2021 Elsevier Ltd. All rights reserved.

Introduction	554
What is Organizational Safety Culture?	555
How Can Safety Culture Be Measured?	555
Organizational Safety Climate among Drivers at Work	556
Safety Culture among Nonprofessional Road Users	556
How Can Safety Culture Inform Preventive Measures in the Road Sector?	557
Approaches to Influencing Safety Culture among Drivers at Work	557
Approaches to Influencing Safety Culture among Nonprofessional Road Users	558
See Also	559
References	559

Introduction

Transport is fundamental to our society, but it unfortunately has several negative side effects. The annual number of road traffic fatalities is now 1.35 million people, while between 20 and 50 million people sustain nonfatal injuries (WHO, 2018). Thanks to safety strategies targeting general road user safety behaviors, as well as improvements in technology and infrastructure, the number of road fatalities has steadily decreased. For example, in the United States, fatality numbers in roadway traffic peaked in 1972 with 54,589 people killed and was reduced to 37,133 in 2017 while the population grew from 209.9 million to 325.7 million, bringing the fatality rate from 260.1 to 114.0 fatalities per million inhabitants, a reduction in the rate of 56%. Even more impressive, in Sweden, the total number of people killed in roadway crashes has gone from a high of 1313 in 1965 (and stayed the same in 1966) to 253 in 2017 in spite of the population growing from 7.77 to 9.95 million; it means that the fatality rate per million people has decreased from 168.9 to 25.4 in this period, a rate decrease of 84.9%. To reduce road crash risks for all road users further, it has been argued that new perspectives are required, and international research has pointed to safety culture as a promising new perspective (AAA, 2007; Ward et al., 2010, 2019). Safety culture generally refers to shared and safety relevant ways of thinking or acting that are (re)created through the joint negotiation of people in social settings.

Since its first use in the wake of the Chernobyl nuclear power accident in 1986, the organizational safety culture concept has become an established part of safety research. Although it has especially been applied in what is considered high-risk settings like the nuclear industry and in aviation, the relationship between organizational safety culture/climate and safety outcomes is robustly documented in studies reporting experiences across organizations, industries, and countries (Christian et al., 2009; Zohar, 2010). The crucial importance of safety culture is also documented in a range of accident investigations. Additionally, high-quality studies of safety culture interventions, with pre- and post-measurements, test, and control groups, have indicated up to 60% decrease in crash risk in the road sector (Nævestad et al., 2018a).

In spite of this, the organizational safety culture perspective has been applied to a limited extent among drivers at work in general and professional drivers in specific. Potential reason for this is the lack of rules requiring safety management systems (SMSs) in the sector. SMS typically includes formal routines and measures enabling the organization to work systematically with safety, by identifying and correcting risks, for example, appointment of key safety personnel, risk assessments, safety training, safety procedures, and safety performance monitoring. SMSs aiming to foster positive safety culture are legally required in aviation, rail, and the maritime sector. In contrast to the other transport sectors, formal SMSs for companies in the road sector are so far voluntary (e.g., ISO:39001). While SMS denotes the formal aspects of safety management in organizations ("what the organization is supposed to do"), safety culture denotes the informal aspects ("what the organization actually does") (Antonsen, 2009). Other factors that may explain the presumably lower focus on organizational safety culture in the road sector are the prevalence of small road transport companies that often have few resources and lower competence on organizational safety management, and that drivers at work often are treated in the same ways as private car drivers, by their companies and authorities. The latter means that drivers at work are given the primary responsibility for the safety of their driving, often neglecting the organizational influence on, and responsibility for their safety performance. This contrasts with Vision Zero, which fits well with a safety culture perspective (refer to Article The Swedish Vision Zero – A Policy Innovation).

When employing the safety culture concept to road transport, it is important to note that the organizational safety culture concept only applies to drivers who are part of work organizations. Given the potential importance of the safety culture perspective for road safety, and the fact that the majority of drivers on the road are not at work, several scholars have also argued that the safety culture perspective should be employed to private road users, linking it to other social units than organizations, for example, nations, communities, peer groups, and families (Edwards et al., 2014; Luria et al., 2014; Nævestad and Bjørnskau 2012; Rakauskas et al., 2009). We will return to this issue later.

What is Organizational Safety Culture?

The concept of (safety) culture lends itself to a wide range of different definitions and operationalizations, as it is abstract and general. A 1963 study conducted by Kroeber and Kluckhohn referred to more than 160 definitions of culture. As a consequence, safety culture is often criticized for being a fuzzy concept that refers to everything and nothing. Researchers have pointed to a fragmented literature and terminological confusion within safety culture research (Antonsen, 2009; Glendon, 2008; Guldenmund, 2000). Nevertheless, most researchers agree that safety culture is a sub-concept of the more general concept of organizational culture, and thus that safety culture can be described and understood by referring to definitions and conceptualizations of organizational culture (Edwards et al., 2014). The highly influential organizational culture scholar and social psychologist Edgar Schein defines organizational safety culture as:

“(. . .) a pattern of shared basic assumptions that was learned by a group as it solved its problems of external adaptation and internal integration, that has worked well enough to be considered valid and, therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems.” (Schein, 2004, p. 17)

Drawing on a sociological/social anthropological understanding of culture, Stian Antonsen refers to culture as:

“(. . .) frames of reference through which information, symbols and behavior are interpreted, and the conventions for behavior, interaction and communication are generated.” (Antonsen, 2009, p. 4).

Bearing these definitions in mind, organizational safety culture can be defined as “safety relevant aspects of culture in organizations.” The important thing to remember about these definitions is that safety culture provides frames of reference that guide individuals’ interpretations of actions, hazards, and their identities, and which motivate and legitimize behaviors that have an impact on safety, and that such shared frames of reference are created through interaction within groups.

It should also be mentioned that several scholars on culture in organizations discern between different levels of culture. Schein discerns between three different levels. Culture at the deepest level refers to underlying assumptions (beliefs, perceptions, thoughts, and feelings) influencing what we pay attention to, what things mean, how we react emotionally, and how we act. Organizational culture at the middle level refers to espoused beliefs and values, for example, explicit strategies, goals, and philosophies, while organizational culture at the most superficial level refers to artifacts, for example, visible organizational structure and processes. Accident investigations often find discrepancies between cultural levels, for example, between espoused values (e.g., safety first) and underlying assumptions (e.g., timetables must be met).

Due to the safety performance and acknowledged good safety culture of aviation, James Reason’s description of five key aspects of safety culture (based on aviation) is often used as a point of departure for describing the most important aspects of good safety culture (Reason, 1997). First, Reason refers to a good safety culture as an informed culture, which means that the organization collects information about both accidents and incidents and carries out proactive countermeasures. The second aspect is a reporting culture, which means that all employees report their errors or near misses, and take part in initiatives to improve safety. The third aspect of safety culture is a just culture, which means that there is an atmosphere of trust within an organization that encourages and rewards its employees for providing information on errors and incidents, with the confidence of knowing that they will receive fair and just treatment for any mistake they make. A just culture is a key premise of a reporting culture. The fourth aspect is a flexible culture, which involves that the organization and its members are capable of adapting effectively to changing demands. The final aspect of safety culture is a learning culture, which means that the organization learns from incident reports, safety audits, and so forth, resulting in improved safety. These aspects of good safety culture are by and large taken for granted as the key aspects of good safety culture in all sectors, including transport, for example, in rail, in maritime sector, and in road.

How Can Safety Culture Be Measured?

Safety culture, exemplified by Reason’s key aspects, can be studied using a qualitative or a quantitative approach, or ideally through a combination of the two. These methods facilitate examinations of different aspects and layers of safety culture. Quantitative studies are often referred to as studies of safety climate, which can be conceived of as “snapshots,” or manifestations of safety culture (Flin et al., 2000). Safety climate is also defined as perceptions of the value and importance of safety in a given context. The concepts of culture and climate are, however, often used interchangeably. Based on Schein’s depiction of different layers of culture, we may state that safety climate only gives access to the most superficial levels of safety culture: perceptions of managers’ and colleagues’ commitment to safety, perceptions of incident reporting, perceptions of procedures, safety training, etc (Guldenmund, 2007). Quantitative measurements of safety culture can provide leading indicators of safety and consequently offer predictive assessments that enable safety improvements without having to wait for crashes or incidents to happen. Quantitative measurements of safety culture are necessary to compare scores over time between organizations and to quantify the relationship between safety culture and safety outcomes.

The most commonly measured aspect in safety climate studies, independent of sector, is senior management commitment to safety. This is the most studied and best-documented characteristic of a good safety climate. It tends to influence all other safety-related

aspects of organizations (Flin et al., 2000). The other most commonly measured aspects of safety climate, independent of sector, are employee perceptions of SMSs (e.g., work permit systems and safety philosophies) and perceptions of, and attitudes to, risk.

Risking oversimplification, we may suggest that while safety climate often is studied quantitatively by social psychologists, safety culture is often studied qualitatively by social anthropologists and sociologists. The qualitative studies involve research interviews and/or time-consuming fieldworks, where researchers interact with people over long periods of time to learn how they see the world, how they think, how they (inter)act, etc. Qualitative studies of safety culture focus on how it guides individuals' interpretations of actions, hazards, and their identities, and motivates and legitimizes behaviors that have an impact on safety. These studies may give us access to the "deeper" levels of safety culture: the more implicit and taken-for-granted basic assumptions and "tacit knowledge." Qualitative studies focus on how safety culture provides a frame of reference that guides individuals' interpretations of actions, hazards, and their identities, and which motivates and legitimizes behaviors that have an impact on safety.

Organizational Safety Climate among Drivers at Work

There are few studies focusing explicitly on organizational safety climate among professional drivers, or drivers at work in road transport. The studies that do exist often combine organizational safety climate questionnaires with questionnaires measuring safety outcomes such as self-reported driving behaviors (e.g., the Driving Behavior Questionnaire [DBQ]) (Reason et al., 1990) and self-reported crashes (Davey et al., 2006; Öz et al., 2014).

Apart from also focusing on management commitment to safety, like most of the studies of safety climate independent of sector do, the studies of safety climate in road transport organizations also measure safety climate aspects such as, for example, the supervisor's role (adequacy of), fleet safety rules, support from managers and supervisors during the trips, communication and support, work pressure, driver training, colleagues' influence, competence, etc (Davey et al., 2006; Huang et al., 2013; Wills et al., 2005). The studies also often include questions on perceived time and work pressure, and other work-related factors related to efficiency, for example, commission pay, bonus, and reward systems. The studies also include questions measuring attitudes to various traffic safety interventions targeting risky behaviors, perceptions of risky behaviors, and so forth. Although management commitment, perceptions of SMS aspects, and work pressure are recurring safety climate aspects measured in the studies, this indicates the abundance of possible safety climate aspects. As in studies of safety climate in other sectors, it may be difficult to replicate factor structures in different organizational settings.

The existing studies generally find that positive organizational safety culture/climate scores are related to lower incidences of aberrant driving behaviors, measured by means of the DBQ. The DBQ originally distinguishes between three types of aberrant behaviors, based on Reason et al. (1990): lapses, errors, and violations. Lapses typically involve problems with attention and memory. Errors typically involve observation failures and misjudgments. Violations involve deliberate deviations from safe driving practices. The aberrant driving behaviors that have been found to be related to organizational safety culture/climate are violations and errors.

Studies of safety climate/culture in road transport also find a relationship between organizational safety culture/climate scores and crash involvement among professional drivers and drivers at work. As noted, high-quality studies of safety culture interventions, with pre- and post-measurements, test, and control groups, have indicated up to 60% decrease in crash risk (property damage accidents) in the road sector. Retrospective data indicate that an annual average of about 1.500 people in Norway are injured in accidents involving drivers at work annually, and most (81%) of the people injured in these accidents are other road users. At the same time, Norwegian and international research indicates that companies employing drivers at work often focus (too) little on organizational factors such as safety culture and SMS. Estimates including only road goods transport companies in Norway indicate that between 7 and 56 fatalities/severely injured potentially could have been avoided annually, if those companies had focused systematically on safety culture and SMS (Nævestad et al., 2018b). The estimates take into account prevalence, expected effects, etc. These estimates are conservative, and they only apply to about 40% of the accidents involving drivers at work in Norway. Thus, there is a large and largely untapped safety potential in focusing on safety culture and SMS in transport.

The relationship between safety climate/culture and crashes often seems to be mediated by driving behaviors, for example, measured by means of the DBQ. Studying the relationship between organizational safety culture and crashes is, however, not a straightforward issue, as it may be difficult to determine the causal relationship between safety culture and crashes. First, reporting of errors, injuries, crashes, etc., is one of the key characteristics of good safety culture. Thus, organizations with a good safety culture will also have a higher probability of reporting a high number of crashes, injuries, and incidents. (Still, police reported injury crashes would be relatively unbiased.) In organizations with poorer safety culture, on the contrary, we may expect a certain level of underreporting of incidents. Second, cross-sectional studies of safety culture and crashes may also find a negative relationship, e. g. if organizations have learned from crashes and developed a positive safety culture. Nevertheless, high-quality studies of organizational safety culture in road transport also indicate that good safety culture/climate is related to fewer crashes.

Safety Culture among Nonprofessional Road Users

It seems even more important to employ the safety culture perspective to nonprofessional road users, as these include high-risk groups, like those who are too young or too old to be employed by work organizations, and which also include drivers who drive in

high-risk contexts, for example, in weekend, at evening/nights, and with friends (Luria et al., 2014). Few studies have, however, so far been devoted to this issue. Employing the safety culture concept to these groups involves, however, shifting the focus to other social units than organizations, for example, nations, communities, peer groups, and families (Edwards et al., 2014; Luria et al., 2014; Nævestad and Bjørnskau, 2012; Rakauskas et al., 2009). There are several challenges related to this, as some sociocultural units may be less well defined than, for example, work organizations (Nævestad and Bjørnskau, 2012).

Although scholars seem to agree that safety culture is also important for nonprofessional road users, there are still no definitions of traffic safety culture that are commonly accepted by road safety researchers (Edwards et al., 2014; Girasek 2012). Previous studies of traffic safety culture among private drivers define it as, for example, the “beliefs, norms and values and things people use that guide their social interactions in everyday life,” “implicit shared values and beliefs,” and “common practices, expectations and informal rules that drivers learn by observation from others in their communities (Edwards et al., 2014).”

There seems to be agreement that traffic safety culture among nonprofessional road users can be viewed as a different application of the same foundational concept as organizational safety culture (Edwards et al., 2014). Thus, we may, for example, define safety culture among nonprofessional road users as, for example, shared norms, beliefs, and assumptions that provide frames of reference that guide individuals’ interpretations of actions, hazards, and their identities, and which motivate and legitimize behaviors that have an impact on safety, and which are created through interaction within groups. We may hypothesize that this definition also can be applied to the national level, the community level, peer groups, and families.

Safety culture is generally (re)created through interaction processes in groups, and, to explain the existence of traffic safety culture at different levels (e.g., national, communities, peer group, and family), we may hypothesize that this interaction is influenced by factors located at different analytical levels. At the more general level, we may hypothesize that national road safety culture is created through processes involving drivers in the same country. Several factors that could influence a traffic safety culture are national (e.g., traffic rules, the police enforcing the rules, and infrastructure). Based on these factors, we may expect the existence of different national traffic safety cultures. Studies comparing road safety in countries find considerable variations between national road safety records (the fatality risk varies with a factor of up to 10 between EU countries), and between self-reported road safety behaviors, measured by means of the DBQ. In accordance with this, studies find national differences between national traffic safety cultures. A recent study comparing national traffic safety culture among Norwegian and Greek private and professional drivers found a relationship between national traffic safety culture and driver behaviors, which in turn were related to self-reported accident involvement (Nævestad et al., 2019). The authors found that the differences were in accordance with the national road safety records in the two countries. This study defines road safety culture as shared patterns of behavior, shared norms prescribing certain road safety behaviors, and shared expectations regarding the behaviors of others. The study measures national traffic safety culture as descriptive norms, hypothesizing that the mechanisms between national traffic safety culture and road safety behavior are drivers’ perception of what is “normal” and expected from drivers within their country, generating a subtle social pressure to behave in certain ways. This hypothesized mechanism may explain the influence of road safety culture in several different groups (national, community, peer groups, and family) on behaviors, and we may hypothesize that the perceived social pressure is stronger, contingent on the importance of the group (e.g., that family and peers are most important).

It also seems well founded to hypothesize the existence of traffic safety culture at the community level, as road user interaction to a large extent occurs between a given number of actors within relatively geographical limited areas (Luria et al., 2014). This especially applies to islands, which to some extent are closed communities. The mentioned study of traffic safety culture in Norway and Greece found such an “island effect”, and a previous study has also found differences between traffic safety culture in rural and urban areas.

Some studies also apply the traffic safety culture perspective to peer groups, which can be defined as a group of people sharing some key characteristics, often a combination of age, sex, and/or crucial interests. This type of sociocultural group membership is hypothesized to be especially important, as its members are assumed to view each other as equal, significant others, who share identities and who often exercise a considerable level of normative pressure to comply with certain norms for behavior. In accordance with this, studies have found descriptive norms at the peer level to influence road user behavior.

Finally, it should be noted that the application of the safety culture concept to nonprofessional road users is less developed than the organizational safety culture concept. The successful application of the safety culture concept to private drivers seems to be contingent on developing more research-based knowledge on: (1) the mechanisms generating traffic safety culture in different sociocultural units (nations, communities, peer groups, and families), (2) a theoretical explanation of the influence of traffic safety culture on traffic safety behaviors in these sociocultural units, and (3) how this knowledge can be utilized to develop successful interventions to improve road safety (Nævestad et al., 2019). We will return to these issues later.

How Can Safety Culture Inform Preventive Measures in the Road Sector?

Approaches to Influencing Safety Culture among Drivers at Work

Although there are few studies of explicit safety culture interventions in the road sector, the research on organizational safety culture stands on the shoulders of the more well-developed organizational culture research. It is therefore easier to describe how this perspective can inform measures to prevent road crashes, than it is to describe how the research on nonprofessional road users can inform preventive measures (as this scarcely has been studied).

Based on existing research, we may conceive of at least three ways of influencing organizational safety culture in organizations employing drivers at work. First, managers may influence safety culture through their daily management of safety culture. Schein (2004) refers to “six primary embedding mechanisms” that managers can use to shape culture:

- 1 What managers pay attention to, measure, and control on a regular basis
- 2 How managers react to critical incidents and organizational crises
- 3 How managers allocate resources
- 4 Deliberate role modeling, teaching, and coaching
- 5 How managers allocate rewards and status
- 6 How managers recruit, select, promote, and excommunicate

Although this probably is the most important way of influencing safety culture, due to its regularity and comprehensiveness (ideally applying to all aspects of organizational life), these mechanisms have, to a little extent, been subjected to robust evaluations in transport. The reason is probably that this mode of influence is constant, indefinite, and hard to delineate and define. It must, however, be noted that the importance of these mechanisms is indicated by the fact that management commitment to safety is the most important aspects in studies of organizational safety climate. The mechanisms specify and operationalize how management may indicate commitment to safety. Several scholars criticize, however, the assertion that leaders initiate and shape integrated organizational cultures in organizations. They hold that culture emerges bottom-up, as cultures are created and recreated through group members’ interaction and negotiation over meaning (Nævestad et al., 2018a). This means that organizations comprise several subcultures that may respond negatively to managerial efforts to manage culture, and that changing and managing safety culture is a very demanding task.

Second, safety culture in organizations may also be influenced by means of interventions. A systematic literature review of safety culture interventions in transport organizations, published in 2018, identifies eight relevant studies of interventions that may be defined as safety culture interventions in road transport organizations (Nævestad et al., 2018a). Generally, the interventions have a broad safety focus, which means that safety culture is not the primary target of the interventions. Effects on safety culture are only measured (or assumed to happen) as an additional effect of a broader set of organizational safety measures. The interventions also vary widely in resource intensiveness. Nevertheless, the author concludes that safety culture interventions seem to be effective, as the most robust of the reviewed studies (with pre- and post-measurements, test, and control groups) indicates that safety culture interventions improve safety culture and safety behaviors, and reduce crashes. However, as the studied interventions often are very different, the author identifies four key elements that seem to be common in all the reviewed safety culture interventions:

- 1 Appointing a key person (generally a manager) to be responsible for implementing the intervention.
- 2 Institutionalizing joint discussions and risk assessments of workplace hazards, involving managers and employees.
- 3 Implementing and monitoring measures based on these discussions and joint risk assessments, for example, reporting systems and training.
- 4 Maintaining effective communication about safety issues in the organization, in line with Reason’s depiction of an informed safety culture.

The authors concludes that the basic requirements of safety culture change seem to institutionalize joint discussions of workplace hazards facilitated by manager commitment and employee involvement.

A third possible approach to influencing organizational safety culture is by implementing SMS. As noted, this is a common strategy in other transport sectors, in which SMS fostering positive safety cultures are required. The 2018 review referenced earlier mentions at least one high-quality study describing how voluntary SMS implementation (ISO:39001) in road transport organization fosters positive safety culture.

Approaches to Influencing Safety Culture among Nonprofessional Road Users

Describing how the safety culture perspective can be used to inform preventive measures among nonprofessional road users is, as noted, a relatively unresearched issue compared to the research on organizational safety culture interventions among drivers at work. One possible approach is to take the knowledge about the factors influencing traffic safety culture in the different sociocultural groups as the point of departure, and to identify how to effectively influence these. In a recent study comparing national traffic safety culture in Norway and Greece, the authors hypothesizes that national traffic safety culture is (re)created through road user interaction, which is influenced by infrastructure, enforcement, training, the composition of road users, and economic conditions (Nævestad et al., 2019). The authors also suggest that national traffic safety culture could be influenced by targeting the five factors influencing the interaction among road users. They argue that relevant questions in this respect are, for example, which factors are possible to influence? (How) do the factors interact? What are the expected outcomes? Which is the most important factor? They also state that a crucial question is the extent to which traffic safety culture is an independent social force, and the extent to which it merely reflects other influencing factors (e.g., infrastructure and enforcement). An important question related to this is the extent to which collectively shared norms for behavior remain unchanged, although the factors through which they originally were influenced by change, or whether shared norms “follow” from infrastructure, police enforcement, and education. We may hypothesize that the interaction between road users creates an independent dynamic where new norms are developed, which not necessarily reflect the infrastructure. Certain social groups may develop norms prescribing unsafe road behaviors, despite good training, infrastructure,

and police enforcement. This discussion and several of the discussed factors (e.g., interaction, enforcement, and infrastructure) are also relevant in the discussion of how to influence traffic safety culture at the community level.

A second and relatively widespread approach that is discussed by the mentioned study, and which already is applied in several countries, is to change traffic safety culture through campaigns aiming to change road users' attitudes, norms, and thus to motivate safe(r) road transport behaviors (Nævestad et al., 2019). Relevant questions in this respect are, for example, how do social norms for road safety behavior come about? To what extent is it possible to change such norms through other ways than the interaction processes in which (we assume) that they are created? How should it be done? Research on road safety campaigns often shows that they are more efficient when they are combined with police enforcement. Thus, future research should also shed more light on such relationships and the combinations of mechanisms generating changes in traffic safety cultures.

A third approach to influencing traffic safety culture among nonprofessional road users, that also is discussed in the aforementioned study of Norway and Greece, is to employ a social norms approach, targeting descriptive norms, that is, perceptions of what other people in a given group do (Nævestad et al., 2019). As noted, the study of national traffic safety culture in Norway and Greece hypothesized that the observed relationship between national traffic safety culture and road safety behaviors could be explained by pointing to descriptive norms, creating a subtle social pressure to behave in certain ways. The underlying idea behind the social norms approach to interventions is that people who behave in undesirable ways (e.g. driving under the influence of alcohol, failing to use seat-belt) tend to overestimate the prevalence of risky behavior among other people (e.g., in their country and in peer groups) to justify their own undesirable behavior (e.g. "everybody drink and drive"). This effect is referred to as false consensus. Thus, the social norms approach hypothesizes that by informing people about the actual prevalence of risky behavior in given groups, the false consensus effect will be reduced or removed, and mild social pressure (i.e., the descriptive norms mechanism) will make people behave in safer ways. This approach has successfully been employed in traffic safety interventions. The social norms approach may, however, not work if road users at risk have fairly correct perceptions of their peers' (un)safe behavior. This could be the case in high-risk subcultures, which may be based on, and defined by, risky behaviors.

See Also

The Swedish vision zero – a policy innovation

References

- AAA, 2007. Improving traffic safety culture in The United States – the journey forward. AAA (Ed).
- Antonsen, S., 2009. Safety Culture: Theory, Method and Improvement. Ashgate, Farnham.
- Christian, M.S., Bradley, J.C., Wallace, J.C., Burke, M.J., 2009. Workplace safety: a meta-analysis of the role of person and situation factors. *J. Appl. Psychol.* 94, 1103–1127.
- Davey, J., Freeman, J., Wishart, D., 2006. A study predicting crashes among a sample of fleet drivers. In: Proceedings of the Road Safety Research, Policing and Education Conference, 25–27 October 2006, Gold Coast, Australia.
- Edwards, J., Freeman, J., Soole, D., Watson, B., 2014. A framework for conceptualising traffic safety culture. *Transp. Res. F Psychol. Behav.* 26, 293–302.
- Flin, R., Mearns, K., O'Connor, P., Bryden, R., 2000. Measuring safety climate: identifying the common features. *Saf. Sci.* 34, 177–192.
- Girasek, D.C., 2012. Towards operationalising and measuring the Traffic Safety Culture construct. *Int. J. Inj. Contr. Saf. Promot.* 19 (1), 37–46.
- Glendon, I., 2008. Safety culture: snapshot of a developing concept. *J. Occup. Health Saf. Aust. N.Z.* 24 (3), 179–189.
- Guldenmund, F.W., 2000. The nature of safety culture: a review of theory and research. *Saf. Sci.* 34, 215–257.
- Guldenmund, F.W., 2007. The use of questionnaires in safety culture research – an evaluation. *Saf. Sci.* 45, 723–743.
- Huang, Y., Zohar, D., Robertson, M.M., Garabet, A., Lee, J., Murphy, L.A., 2013. Development and validation of safety climate scales for lone workers using truck drivers as exemplar. *Transp. Res. Part F* 17, 5–19.
- Luria, G., Boehm, A., Mazor, T., 2014. Conceptualizing and measuring community road-safety climate. *Saf. Sci.* 70, 288–294.
- Nævestad, T.-O., Bjørnskau, T., 2012. How can the safety culture perspective be applied to road traffic? *Transp. Res.* 32, 139–154.
- Nævestad, T.-O., Phillips, R.O., Hesjevoll, I.S., 2018a. How can we improve safety culture in transport organizations? A review of interventions, effects and influencing factors. *Transp. Res. Part F Psychol. Behav.* 54, 28–46.
- Nævestad, T.O., Phillips, R.O., Hovi, I.B., Jordbakke, G.N., Elvik, R., 2018b. Miniscenario: Sikkerhetsstigen. Innføre tiltak for sikkerhetsstyring i godstransportbedrifter, TØI Rapport 1629/2018, Oslo, Transportøkonomisk institutt.
- Nævestad, T.-O., Laiou, A., Phillips, R.O., Bjørnskau, T., Yannis, G., 2019. Safety culture among private and professional drivers in Norway and Greece: examining the influence of National Road Safety Culture. *Safety* 5 (2), 20, Special Issue: Social Safety and Security.
- Öz, B., Ozkan, T., Lajunen, T., 2014. Trip-focused organizational safety climate: investigating the relationships with errors, violations and positive driver behaviours in professional driving. *Transp. Res. Part F Traffic Psychol. Behav.* 26, 361–369.
- Rakauskas, M., Ward, N., Gerberich, S., 2009. Identification of differences between rural and urban safety cultures. *Accid. Anal. Prev.* 41 (5), 931–937.
- Reason, J., 1997. Managing the Risks of Organizational Accidents. Ashgate, Aldershot.
- Reason, J.T., Manstead, A.S.R., Stradling, S.G., Baxter, J.S., Campbell, K., 1990. Errors and violations on the road: a real distinction? *Ergonomics* 33, 1315–1332.
- Schein, E.H., 2004. Organizational Culture and Leadership, third ed. Jossey-Bass, San Francisco, CA.
- Ward, N.J., Linkenbach, J., Keller, S.N., Otto, J., 2010. White paper on traffic safety culture. White Papers for: "Toward Zero Deaths: A National Strategy for Highway Safety" Series—White Paper No. 2. Montana State University, Bozeman, MT, USA.
- Ward, N.J., Watson, B., Fleming-Vogl, K., 2019. Traffic Safety Culture: Definition, Foundation, and Application. Emerald Publishing Ltd., Bingley.
- WHO, 2018. Available from: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>.
- Wills, A.R., Biggs, H.C., Watson, B., 2005. Analysis of a safety climate measure for occupational vehicle drivers and implications for safer workplaces. *Aust. J. Rehabil. Counsel.* 11, 8–21.
- Zohar, D., 2010. Thirty years of safety climate research: reflections and future directions. *Accid. Anal. Prev.* 42 (5), 1517–1522.

Safety Data Quality Management

Robert A. Scopatz, VHB, Inver Grove Heights, MN, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction and Background	560
Surface Transportation Safety Data	560
Data Quality Principles	561
Data Quality Attributes	561
Data Standards, Data Governance, and Formal Data Quality Management	561
Statistical Methods of Data Quality Control	561
Establishing Statistically Valid Expectations	562
Calculating Data Quality Scores for Each Record	562
Formal Data Quality Management Processes	562
Summary and Benefits of Formal Data Quality Management	563
References	563
Further Reading	563

Introduction and Background

This article describes how and why transportation safety professionals manage data quality. It starts by defining the domain of transportation safety data (also known as traffic records) and presents the guiding philosophies and techniques for managing the core datasets. This article ends with a discussion of why data quality management should be considered a formal process that goes beyond measurement to include actions designed to improve data quality.

Surface Transportation Safety Data

The domain of transportation safety typically refers to surface transportation—the safety of people, vehicles, and commodities on public roadways. The domain excludes other forms of transportation (air, rail, water, and pipelines) that have their own data standards. The list of datasets used in transportation safety varies substantially by country. In North America, Canada, Europe, Australia, and parts of Asia, South America, and Africa, the list of available data includes six core system groups: crash, roadway, driver, vehicle, citation and court records, and injury surveillance data ([National Highway Traffic Safety Administration, 2011](#)). Every country has access to a subset of the following datasets:

- **Crash data:** Records of crash events including information on the locations, people, and vehicles involved, and multiple contributing factors. These records may come from law enforcement officers and/or from the involved parties (operator reports). The formal law enforcement reports may be limited to public roadways—crashes on private roads might result in a different report type and might not be included in traffic safety analyses. Insurance companies also keep records of crashes and associated costs, but this information is seldom accessible for analysis outside of the companies' actuaries or industry-based research groups.
- **Roadway data:** There are several relevant roadway data sources. The most important, for safety analysis, are the roadway inventory data (a list of attributes for each defined location) and the traffic volume data (counts of traffic at specific locations). Other relevant roadway data include the spatial definitions of locations (e.g., the start and end points of routes, segments, intersections, etc.), asset management data (inventories of the location and condition of signs, markings, roadside appurtenances, etc.), intersection inventory data (a listing of intersections and the features specific to places where two or more roadways meet), structures data (e.g., bridges, tunnels, overpasses, sign gantries, etc.), and more.
- **Driver data:** Records of each driver's licensing, training, and history of convictions and license sanctions. Driver history data relies on court records and initial citations if there is no court adjudication (e.g., for minor violations in some jurisdictions). Training data is often lacking except for first-time licensees and special endorsements (e.g., motorcycle or commercial vehicle) where completion of training may be recorded.
- **Vehicle data:** Records of vehicle titling and registration along with coded descriptions of the vehicle's type, features, and date and place of manufacture. In some places, the vehicle data may also include indications of the vehicle's condition and mileage.
- **Citation and adjudication data:** Records of original violations issued by law enforcement officers and (if any) the court process outcomes for each violation. If citations are disposed without a court process (i.e., fines for moving violations are payable to the licensing authority directly), the record of original citation and payment may complete the record. These records form the basis for the driver history records referenced earlier in the list. Court-imposed license sanctions, training, and treatment requirements are also recorded in this data source and are, typically, transferred to the driver licensing authority as well.

- **Injury surveillance data:** There are several relevant data sources describing injuries, treatments, and outcomes for crash-involved persons. Some of these databases are available publicly (in redacted form). The first dataset describes emergency medical services provided at the scene as well as any transportation to a medical facility for treatment. If the injured person is treated in an emergency department, records of their treatment and diagnostic codes describing the injury will be recorded. The same visit may result in a record of the injuries and treatments in a trauma registry, spinal cord injury registry, or traumatic brain injury registry. For persons admitted to a hospital, the hospital discharge record will contain a record of treatments, diagnoses, outcomes/prognosis, and, in some systems, the billing records for treatments. Physical and occupational therapy centers and long-term care facilities may also keep records on those persons treated in their facilities. Finally, for those who died as a result of a motor vehicle crash, coroner's reports and death certificate data record the causes and injuries sustained as well as the date of the original event and the date of death—for motor vehicle crash-related deaths this source may be the only official record of a relationship between the event and the cause of death if the person survived the crash for a long period. Efforts to combine injury surveillance and crash data are now common throughout the world. In the United States, the original Crash Outcomes Data Evaluation System (CODES) is no longer funded by the federal government, but the project continues in several states and under different names and methodologies. Sweden has the Swedish Traffic Accident Data Acquisition (STRADA) program that similarly links crash and injury data to provide a more robust dataset for analyzing crash-related injuries.

Data Quality Principles

Data Quality Attributes

Data quality is a quantifiable aspect of data. The list of data quality attributes includes the following four core items:

- **Timeliness:** Timeliness measures the interval between a recordable event and the appearance of the data in a system ready to be used for analysis and decision-making. Depending on the dataset and the needs of managers and users, timeliness can be measured for each step in a sequence; for example, we might measure the time from a crash to submission of a crash report separately from the time from submission to posting on the database.
- **Accuracy, precision, and resolution:** These are often described jointly as accuracy—a term that describes how close the match is between the recorded data and the ground truth of whatever is being measured. In spatial data, the terms accuracy, precision, and resolution take on more specific meanings. Accuracy is the difference between the actual location and the measured location. Precision is the difference between multiple measures of a specific location (how close the measurements are to one another even if their average is inaccurate). Resolution is the variability inherent in a spatial measurement based on the capability of the measurement method or tool (each estimated location is expressed $\pm$ a specific tolerance).
- **Completeness:** Completeness measures what portion of the expected data record or full data file is complete. For example, we may wish to know the width of every roadway segment, and the completeness measure would tell us what percentage of roadways are lacking pavement width information.
- **Uniformity:** Uniformity measures how well data collection adheres to the established data requirements or agrees with any established guidance.

There are many other measurable data quality attributes. Different sources list integrity, validity, reliability, accessibility, and integration ([Australian Bureau of Statistics, 2015](#); [ECCMA, 2019](#); [Statistics Canada, 2019](#)). This longer list does present some concerns for the ability to accurately measure the attributes—much of it would require a subjective judgments by selected users. Some of these, particularly accessibility and integration, are not directly related to how the data are collected or stored, but they can have importance for users and managers. In general, it is a good idea to use a formal data governance process as a way to determine which data quality attributes are most important, and how best to measure those attributes.

Data Standards, Data Governance, and Formal Data Quality Management

Data quality management is closely tied to the concepts of data standards—including the quality requirements for specific data elements—and data governance—the processes by which all data management, not just data quality management, is achieved. A data standard provides data collectors, managers, and users with requirements the data must meet in order to be accepted into the official government database. A data standard is anything that defines the inclusion and exclusion cases for a data element, the method of collection (an operational definition), and the measurable data quality levels that must be achieved. Data governance is a formal process for establishing and enforcing data standards. It typically involves executive-level decision-makers in setting the documentation and policy requirements for every data system and a practitioner-level data governance group responsible for implementing and enforcing those requirements.

Statistical Methods of Data Quality Control

Statistical methods of data quality control are numerous and varied; however, some common features of the methods are worth exploring here. As the name implies, these methods use statistical modeling as a way to measure data quality and, importantly, set

reasonable expected values for those measurements. Using statistical methods can help agencies save time and effort by highlighting those data quality measurement results that fall outside the expected range and thus would be worth spending time and resources to address. Statistical methods also allow for a refinement in data management by calculating a data quality score for each record in a database. This score reflects how well the particular record (the collection of all data elements within the report of a single location or event) meets the established data quality standards. Examples of these two concepts follow.

Establishing Statistically Valid Expectations

Data quality measurements, and associated goals, are properly considered as a mean (or, less frequently, a median or modal) value rather than a specific point value. The true value of the measurements should fall, predictably, in a distribution around that estimate of the central value. Statistical methods of quality control set an expected range around the mean value so that data managers can ignore changes that fall within the allowed range and act whenever the obtained value(s) fall outside the allowed range. One common technique is to establish a criterion range such as $\pm$ two standard deviations from the mean. That would coincide with an expectation that data managers would only research and respond to obtained values of the data quality performance measures that are so extreme as to be expected by chance only 5% of the time. This type of criterion can avoid wasted effort responding to random fluctuations that just happen to appear worrisome from the direction of movement in the raw numbers alone. Of course, the criterion range can be adjusted to suit the needs of the data managers and key users.

Calculating Data Quality Scores for Each Record

Within the traffic records datasets, records about an event, person, or location contain multiple data elements. Each element can be weighted for importance to decision-making, and each element can have one or more data quality measurements associated with it as part of data checking and validation. A record-level data quality score is a composite of all the quality scores for the data elements in a record, weighted by the importance of each data element. This composite quality score is then appended to the record as a new data element. It can be used by data managers and analysts to judge the overall quality of the entire dataset by taking the average quality rating of all records. It can also be used to filter the database when analysts may wish to use only those records that meet a minimum quality level. If data managers establish a minimum quality level, those records that fail to meet criterion could be rejected and returned to the data collectors for correction. It is also possible to provide different quality scores for different users and uses. For example, practitioners responsible for driver sanctions might weight data elements in a crash and citation database differently to how practitioners responsible for highway engineering would weight the same data elements. Thus, it is possible to create multiple quality scores, or business-need-specific component scores, to meet the needs of different data users.

Formal Data Quality Management Processes

Data standards and data governance are part of a formal data quality management process. Agencies may use other means of managing data; however, the best practice involves the following:

- System documentation including a complete data dictionary defining each data element and describing the necessary metadata and validation rules for each data element.
- A data governance committee that develops, communicates, and enforces data standards (including data quality standards).
- Validation rules (also known as edit checks) to verify that each data element in a submitted record is of the proper type, size, value range, and format, and that it meets the established rules for logical agreement with other data elements.
- Data quality measurements that quantify (at a minimum) the timeliness, accuracy, completeness, and uniformity levels achieved in the data. The selection of data quality attributes and measurements should meet the needs of users and data managers.
- Numeric goals for each data quality measurement so that actual performance can be compared to the desired levels of quality.
- Analytic reviews of database contents to check for unexpected changes period to period, in proportions of specific codes used based on prior years' data and trends, and for differences among reporting jurisdictions. This may be combined with statistical methods of measurement.
- Feedback and correction mechanisms that inform data collectors of errors and require timely corrections by the original data source (where possible).
- Data errors and data change logging to track the frequency of each type of error in each data element, and to log the original and final data values in each field, along with who made the change and when the change was made. Systems that allow filtering by the type and source of the change allow for flexibility in meeting different users' needs.
- Linking error reports and change logs to communications content provided to data collectors. This may include changes to data collection instruction manuals, as well as help files embedded in data collection software, training, and special communications.
- Reporting data quality measurements at a level of detail that can identify the need for improvement for individual data collection sources as well as system-wide summaries.
- Including data quality improvement in major plans by the custodial agencies responsible for each database. The planning efforts should be informed by the quantitative data quality measurements and goals for data quality should be addressed in the various plans as well. In particular, there should be a comprehensive strategic plan for traffic records improvement and that plan should be referenced when related plans (such as highway safety, injury surveillance, and others) are created.

Summary and Benefits of Formal Data Quality Management

Data quality management is a process that extends beyond merely measuring data quality. It is closely related to data governance. Agencies with mature, formal data governance practices will also meet the definition of formal data quality management described here. Likewise, formal data quality management involves many of the same processes that an agency would put into practice if they implement formal data governance. Doing both is the best practice defined here.

Data quality management is not without costs. Staff must be trained in and devote time to quality assessment and improvement efforts. Rejecting data after it is collected because it does not meet the enforced standards means that there may be costs associated with getting the data up to the required quality levels. Over time, and with practice and training, the cost of maintaining data quality should level off and perhaps even drop as ways are found to automate data validation before it is submitted. The value of high-quality data must be worth the initial (higher) costs and the long-term costs to maintain the desired quality levels.

It can be difficult to estimate the value of data. It is possible to, theoretically at least, make decisions using methods that do not use data and compare the outcomes to decisions made when the data are available and of sufficiently high quality. As a thought experiment, one would expect that data-driven decisions will result in fewer costly errors and less waste—for example, by directing resources only to locations and problems that are *real* and *treatable*. In reality, however, the comparison is between making decisions with data of one quality level versus making decisions with data of a higher quality level. Some of the obvious benefits are subjective or intangible: decision-makers feel more confident using recent and complete data; more accurate data are judged as more reliable. Efforts to quantifiably prove the benefits of better data, however, are difficult to find. We do know the following:

- Advanced decision-making methods, such as those described in the *Highway Safety Manual*, require more complete and often more accurate data that many departments have maintained prior to implementing those advanced methods. Thus, poor-quality data can inhibit use of the best practice analytic methods and tools.
- When an agency decides to integrate across multiple core datasets (such as CODES projects in the United States and STRADA in Sweden), it is common for the integration effort to fail due to poor data quality in one or more of the source files. Cleaning up the errors increases the number and quality of records that are added to the new, merged data resource. The merged datasets support new insights and new analyses that were not available to decision-makers before the data integration effort.
- Poor-quality data tend to lose users over time, whereas high-quality data gain users. The evidence for this conclusion comes, especially, from years of experience helping agencies prioritize data improvement projects. Priorities are based on business needs for the information, and ultimately determine which data systems get funding and resources. If a system loses users because of poor quality, that system will get less support from key users. Decision-makers will simply look elsewhere for the information they need.

In summary, formal data quality management is the way to gain users, promote data integration, and provide the best possible support for advanced analysis. Decision-makers are the natural allies of data collectors and data managers in making decisions about where to put efforts into improving data quality. Ultimately, poor data quality can lead to costly mistakes, but more importantly high data quality can better support the best practices in decision-making.

References

- Australian Bureau of Statistics, 2015. The ABS Data Quality Framework. Canberra, Australia. Available from: <https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Quality:+The+ABS+Data+Quality+Framework>.
- ECCMA, 2019. ISO 800 Quality Data Principles. International Standards Organization. Available from: https://eccma.org/private/download_library.php?mm_id=22&out_req=SVNPIDgwMDAgUXVhbGI0eSBFYXRhIFB5aW5jaXBsZXNM=.
- National Highway Traffic Safety Administration, 2011. Model performance measures for state traffic records systems. Report No. DOT HS 811 441. Washington, DC. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811441h>.
- Statistics Canada, 2019. Data Quality Toolkit. Ottawa, Canada. Available from: <https://www.statcan.gc.ca/eng/data-quality-toolkit>.

Further Reading

- Crash Outcomes Data Evaluation System, 2013. Available from: <https://www.nhtsa.gov/research-data/state-data-programs>.
- DAMA International, 2017. Data management body of knowledge 2. Available from: <https://dama.org/content/body-knowledge>.
- National Academies of Sciences, Engineering, and Medicine, 2010. Target-setting methods and data management to support performance-based resource allocation by transportation agencies—volume I: research report, and volume II: guide for target setting and data management. The National Academies Press, Washington, DC. Available from: <https://doi.org/10.17226/14429>.
- National Academies of Sciences, Engineering, and Medicine, 2017. Data management and governance practices. The National Academies Press, Washington, DC. Available from: <https://doi.org/10.17226/24777>.
- National Highway Traffic Safety Administration, 2012. Traffic records program assessment advisory. Report No. DOT HS 811 644. Washington, DC. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811644>.
- Swedish Traffic Accident Data Acquisition, 2015. Available from: <https://www.trafikverket.se/en/startpage/operations/Operations-road/vision-zero-academy/Vision-Zero-and-ways-to-work/strada/>.

School Bus Safety

Yousif A. Abulhassan, Department of Occupational Safety and Health, Murray State University, Murray, KY, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	564
School Bus Loading Zone	564
School Bus Safety Standards	565
School Bus Emergency Exits	565
School Bus Seat Belts	566
School Bus Security Concerns	566
Closing Remarks	567
References	567
Further Reading	567

Introduction

School buses are one of the most common modes of transportation for millions of students in the United States, Canada, Australia, and some countries in Europe and Asia. The prevalence of school bus transportation is dependent on several factors including the availability of different modes of transportation such as public transportation and social norms such as parents driving their children to school. Other factors such as the climate conditions and economic status also have a significant impact on the popularity of school buses. For instance, school transportation by private drivers is common in countries such as Kuwait and Saudi Arabia. In other countries, school bus transportation is primarily used only by private schools, and the availability of bicycle lanes leads to more students transporting themselves to school. The American School Bus Council estimates that over half the students who attend school in the United States (26 million) use school buses as the primary mode of transportation to and from school (McCray and Brewer, 2005). Taking into consideration the number of accidents and fatalities per mile driven, school buses are considered to be safer than any other mode of transportation for getting to and from school. The National Highway Traffic Safety Administration (NHTSA) estimates that students are 70 times more likely to make it to school safely traveling in a school bus compared to students traveling in a car.

School bus safety is an essential component of the pupil transportation system that primarily serves as a means of transportation between a child's home and school. As shown in Fig. 1, school bus transportation safety can be broken down into two levels. The role of the school bus transportation system starts from the point where a student leaves their home to the point of arrival at a school and vice versa. The safety concerns at each of these levels as outlined in Fig. 1 are controlled by traffic laws pertaining to school buses, Federal Motor Vehicle Safety Standards (FMVSS), and engineering controls such as rollover protection, fire protection, and emergency exits.

School buses are categorized to be one of the following types: Types A, B, C, or D dependent on their gross vehicle weight rating (GVWR), engine location, and passenger capacity. Types A and B school buses are usually smaller than Types C and D school buses and are either van conversions or constructed from the chassis of a van or a truck. Types C and D school buses are the most common type of school bus. The main noticeable difference between Types C and D school buses is the location of the entrance door and location of the engine. Type C school buses have the entrance door located behind the front wheels, whereas a Type D school bus door is located ahead of the front wheel. The front end of Type C school buses has a hood, whereas Type D school buses have a flat front face similar to commuter buses. The maximum seating capacity of Type D school buses can be up to 90 passengers, but these buses often transport 60–70 passengers comfortably. This paper provides an overview of school bus safety standards and design guidelines in addition to some of the research that is aimed to improve the safety of one of the safest modes of transportation.

School Bus Loading Zone

The American School Bus Council estimates that nearly two-thirds of fatalities involving school buses actually occur outside the school bus, whereas one-third of school bus-related fatalities occur inside the school bus. Factors such as environmental conditions and road traffic can have a significant impact on school bus passenger visibility when entering and exiting a school bus. Distraction from texting and use of electronic devices by other vehicle drivers are also a prevalent contributor to the increase in school bus stop-related fatalities and other vehicle contact with school bus passengers. Younger children especially in kindergarten through second grade are more susceptible to these types of accidents due to their small physical size and lower cognitive abilities to assess road hazards. Stop-arm extensions along with flashing red lights on school buses are the most common devices used to alert other drivers on the road of the presence of children near the loading zone of a school bus. Latest technology such as stop-arm cameras are also

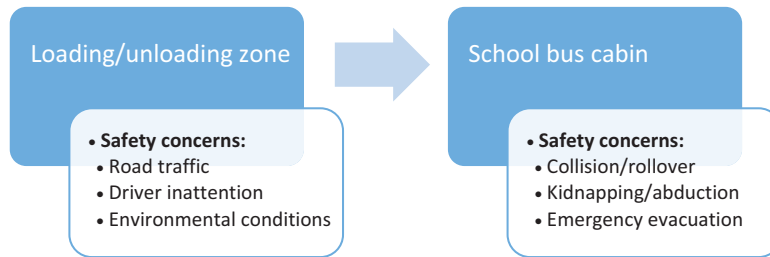


Figure 1 Primary components of school bus safety.

being used in some states to obtain license plate information of vehicles that do not abide with school bus stop laws. Each state has slightly different laws regarding stopping and passing around school buses. In general, all states require that vehicles traveling on the same road as a school bus, regardless of travel direction, to come to a complete stop when stop-arms are extended or red lights are flashing on a school bus.

School Bus Safety Standards

Motorized school buses used to transport children can be dated to the 1900s when vehicles were first used to transport groups of children between home and school. Columbia University's Teachers College established the first school bus safety standards in 1939 (NASDPTS, 2000). It was not until 1977 that the FMVSS (49 CFR Part 571) for modern school bus design requirements were promulgated (NHTSA, 2011). Currently, school buses are the most regulated mode of surface transportation. Some of the prominent school bus safety standards are outlined later.

Stop-arms used to protect children entering and exiting a school bus are regulated by FMVSS No. 131. The standard regulates the design of the stop sign including its shape, color, and reflectivity properties.

Emergency exits on school buses are regulated by FMVSS No. 217—Bus Emergency Exits and Window Retention and Release. This standard provides detailed guidelines on the following:

1. Number and location of emergency exits required on school buses.
2. Requirements to open a school bus emergency exit, including the force requirements to operate latching mechanics.
3. Size requirements of emergency exits.
4. Testing procedures to test for compliance of emergency exits.

The structural integrity of the school bus is primarily regulated by FMVSS Nos. 220 and 221. These standards protect school bus passengers in accidents resulting from rollover collisions and other impact collisions. FMVSS No. 222 provides the standards for school bus seating and crash protection controls that include maximum passenger capacity and passenger protection requirements in the event of a collision. According to the standard, maximum seating capacity of school buses can be calculated given by dividing the width of a school bus bench in millimeters by 380, and multiplying the rounded whole number by the number of benches on a school bus.

School Bus Emergency Exits

The number and location of emergency exits required on a school bus is dependent on the type of school bus, passenger capacity, and the location of the engine. School buses with a maximum seating capacity of up to 57 passengers are not required to have emergency exits. However, depending on the maximum seating capacity of larger school buses, they are required to have a combination of emergency exits that include roof hatches, sidewall emergency windows, and emergency exit doors. The usability and effectiveness of emergency exits are highly dependent on the orientation of a school bus after an accident. For instance, roof hatches are ideal if the school bus is rolled over on either its left or right side; however, they provide little to no utility if the school bus was in an upright orientation or rolled over on its roof. Similarly, emergency exit windows and side doors are mainly used in evacuations where a school bus is in an upright orientation as it would be difficult to reach these exits if a school bus is rolled over to either its left or right side. For these reasons, school buses typically have multiple exits to allow for egress in multiple scenarios.

While school buses are the safest mode of transportation, further improvements can be made to provide a greater level of protection for its occupants. Limited research has been conducted to evaluate the effectiveness of safety devices such as emergency exits. With children being the primary occupants of a school bus, the design of safety devices must take into consideration the limited physical and cognitive capabilities of the most vulnerable population riding on school buses. Research studies conducted by Abulhassan et al. (2016, 2018a, b) have identified discrepancies between current designs of roof hatches and rear emergency door to the physical and cognitive capabilities of young children. The size of young children is of primary concern when evaluating the effectiveness of emergency exits. For instance, as shown in Fig. 2, the significant difference between the size of an average-sized adult

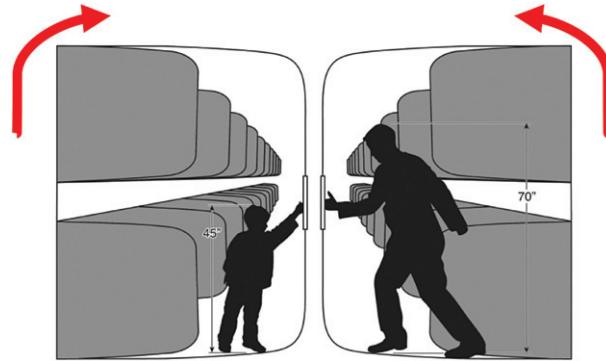


Figure 2 Illustration of a rolled-over school bus (Abulhassan et al., 2018a) (permission obtained from Elsevier RightsLink on January 28, 2019).

and an average-sized child hinders the usability of emergency exits, as some children may not be able to operate or reach these exits. Furthermore, research on the effectiveness of school bus emergency exits suggests that young children lack the physical and cognitive capabilities to meet the operating requirements of school bus emergency exits as set by FMVSS No. 217.

School bus evacuation time is an important safety measure that needs to be considered in scenarios where immediate evacuation is required such as an onboard fire. The size of emergency exits plays an important role in determining the flow rate or the number of passengers that can evacuate through the exit per minute. As the prevalence of higher obesity rates among children has increased over the last 20 years, it is crucial to evaluate the minimum sizes of emergency exits to determine if they provide an adequate level of utility in the event of an emergency.

School Bus Seat Belts

Currently in the United States, only school buses weighing less than 10,000 pounds are required to have seat belts. The increase of accidents resulting in child fatalities such as the school bus crash in Chattanooga, TN, on November 6, 2016 which resulted in the death of six children and the collision of a school bus with a dump truck in New Jersey on May 17, 2018 has led to eight states (Arkansas, California, Florida, Louisiana, Nevada, New Jersey, New York, and Texas) to require school buses with a GVWR of greater than 10,000 pounds to have seat belts. Each of the eight states has approached seat belt laws to various extents. Some states require only new buses to have seat belts, while others require seat belts on all school buses, but leave it up to the local jurisdiction to enforce wearing seat belts. Seat belts voluntarily fitted on school buses must still meet performance standards for lap/shoulder belts as outlined in FMVSS Nos. 222 and 209. Some of the seat belt performance standards include the following:

1. Seat belts ability to withstand forces in the event of a crash.
2. Buckle design including the size of the buckle and maximum push force to disengage the latching mechanism in a seat belt buckle.

School buses with a GVWR greater than 10,000 pounds use a process known as “compartmentalization” where the school bus seats are designed to act as a compartment for occupants in the event of a collision. Compartmentalization is shown to be effective in the event of a front- or rear-end collision; however, it provides little to no protection in the event of a side collision or a school bus rollover. While seat belts can provide a greater level of protection for passengers, there are some unique challenges for implementing seat belts on school buses. Enforcing seat belt usage can present a challenge due to the young age of school bus passengers. The only adult on most school bus routes is the school bus driver, and it would be time consuming for the school bus driver to ensure proper seat belt usage every time a passenger enters the school bus. Strategies such as additional training of young children may be required to properly use seat belts.

School Bus Security Concerns

School bus passenger safety from kidnapping, hijacking, and other criminal activity is an emerging concern in the wake of active school shootings and school bus kidnappings. School buses transporting young children are particularly vulnerable to these types of events due to the lack of physical measures that would prevent illicit access. The National Center for Missing and Exploited Children reports that 38% of attempted abductions occur while students are being transported to and from school in school buses and other modes of transportation including walking (Trump, 2017). The 2013 kidnapping of a 5-year-old child from a school bus after killing the school bus driver in Midland City, AL, has led to increased security measures to prevent unauthorized entry on school buses. As a result of this tragic event, the State of Alabama signed into law the Charles “Chuck” Poland Jr. Act, which makes unauthorized school bus entry and trespassing a Class A misdemeanor. Other states also have similar laws regarding unauthorized school bus entry. The

limited adult authority on school buses has led to increased training of school bus drivers to dealing with school bus security concerns, including identifying questionable bags and other items brought onboard, and diffusing altercations on board a school bus. Controls such as surveillance camera and GPS location systems are also widely used by many school districts to assist in improving security concerns onboard school buses.

Closing Remarks

A multidisciplinary approach is required to undertake safety concerns regarding the unique challenges of transporting vulnerable populations with limited capabilities compared to adults. Even though school buses are among the safest mode of transportation, continued development in the standards and design of these systems can further improve overall safety. Emerging safety and security concerns also need to be addressed to ensure an adequate level of protection for the millions of children that ride on school buses.

References

- Abulhassan, Y., Davis, J., Sesek, R., Callender, A., Schall, M., Gallagher, S., 2018a. Physical and cognitive capabilities of children during operation and evacuation of a school bus emergency roof hatch. *Saf. Sci.* 110, 265–272.
- Abulhassan, Y., Davis, J., Sesek, R., Schall, M., Gallagher, S., 2018b. Evacuating a rolled-over school bus: considerations for young evacuees. *Saf. Sci.* 108, 203–208.
- Abulhassan, Y., Davis, J., Sesek, R., Gallagher, S., Schall, M., 2016. Establishing school bus baseline emergency evacuation times for elementary school students. *Saf. Sci.* 89, 249–255.
- McCray, L.B., Brewer, J., 2005. Child safety research in school buses. In: *Proceedings of the International Technical Conference on the Enhanced Safety of Vehicles*, Vol. 2005. National Highway Traffic Safety Administration, Washington, DC, United States, pp. 1–5.
- NASDPTS, 2000. History of school bus safety—why are school buses built as they are? National Association of State Directors of Pupil Transportation. Available from: <http://nasdpts.org/Documents/Paper-SchoolBusHistory.pdf>
- NHTSA, 2011. CFR. Part 571, Federal Motor Vehicle Safety Standards (FMVSS), Subsection 571.217, FMVSS No. 217, Bus Emergency Exits and Window Retention and Release. US DOT. Office of the Federal Register, National Archives and Records Administration, Washington, DC.
- Trump, K., 2017. School bus transportation security. National School Safety and Security Services. Available from: <https://www.schoolsecurity.org/resource/school-bus-security/>

Further Reading

- American School Bus Council, 1840. Keeping children safe in and around the school bus. Western Avenue, Albany, New York. Available from: <http://schoolbusfacts.com/safety/>
- NHTSA, 2018. School transportation related crashes. National Highway Traffic Safety Administration. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812476.pdf>
- NHTSA, 2015. My choice their ride, school bus facts. National Highway Traffic Safety Administration, Washington, DC, United States.
- NHTSA, 1993. CFR. Part 571, Federal Motor Vehicle Safety Standards (FMVSS), Subsection 571.222, FMVSS No. 222, School Bus Passenger Seating and Crash Protection. US DOT. Office of the Federal Register, National Archives and Records Administration, Washington, DC.
- Purswell, J., Dorris, A., Stephens, R., 1970. Escapeworthiness of vehicles and occupant survival. Report No. 1729-FR-1-1. National Transportation Safety Board, Washington, DC, United States.
- School Transportation News, 2015. Understanding the different school bus types. Torrance, CA. Available from: <https://stnonline.com/news/school-bus-type-faqs/>
- Sliepcevich, C., Steen, W., Purswell, J., Krenek, R., Welker, J., 1972. Escape worthiness of vehicles for occupancy survivals and crashes. First part: research program. Report No. DOT HS-800 736. National Highway Traffic Safety Administration, Washington, DC, United States.

School Campus Traffic Circulation

Dimitrios Nalmpantis, School of Civil Engineering, Faculty of Engineering, Aristotle University of Thessaloniki, Thessaloniki, Central Macedonia, Greece

© 2021 Elsevier Ltd. All rights reserved.

Introduction	568
Pretertiary School Campus Traffic Circulation	568
Private Car	568
Carpooling	569
Active Mobility (Walking and Cycling)	569
Walking School Bus	570
School Bus	571
Motives	572
Interface Management	573
Conclusion	573
Tertiary School Campus Traffic Circulation	573
Segregated Traffic	573
No Traffic	573
Mixed Traffic	573
Conclusion	574
Biography	574
See Also	574
References	575
Further Reading	575

Introduction

School campus traffic circulation is a multidimensional issue that should be addressed on many levels. There are different educational stages, transportation modes, geographical regions, external and internal traffic, and so on. Moreover, there are different impacts, such as on road safety, congestion, air and noise pollution, greenhouse gas emissions, unhealthy mobility habits, etc. All these dimensions are intertwined and increase the complexity of addressing the issue.

First of all, there are many educational stages—apart from preschool and tertiary education, there are primary and secondary schools (in 2-stage systems) or elementary, middle, and high schools (in 3-stage systems). Many transportation modes are used for school commuting, for example, school buses, private cars, public transport, motorcycles, scooters, bicycles, on foot, etc. There are many differences per geographical region; the most studied ones being in the United States and Europe. External traffic is much more significant in pretertiary education, whereas in the case of tertiary education, where students are adults and campuses are usually much larger, internal traffic circulation is usually more important. This is the reason that these two different cases of school traffic circulation will be examined separately—(1) pretertiary school campus traffic circulation, focusing on external traffic circulation and (2) tertiary college and university campus circulation, focusing on internal traffic circulation.

Pretertiary School Campus Traffic Circulation

Going to and from school is a daily trip that millions of children undertake in the United States, Europe, and the rest of the world. Although many children walk or cycle to and from school, the major modes of transport for school commuting in most industrialized countries are school buses and cars (Anund et al., 2011). In low-income countries, walking may be the most common mode. In most cases, the elementary school population and work hours are smaller compared to middle and high schools. Therefore, middle and high schools have more complex traffic problems, such as teen drivers and, in many cases, after-school activities (Tsai et al., 2004).

Private Car

Although the main mode of transport for school commuting in the United States remains the school bus, it has become common that parents drive their children to elementary school through high school. Additionally, in the United States, where in most states it is allowed to drive from the age of 16 years or even earlier, many adolescent students drive to and from school with their own cars.

More specifically, in a hypothetical diagram of automobile use for school commuting a U-shape would be observed—the use of automobile is increased in elementary and high school and decreased in middle school, whereas the opposite is true for school buses. This is probably due to the fact that in elementary school parents prefer to drive their children to school, because they think that it is safer than the school bus, and in high school adolescent students prefer to driver to school rather than taking the school bus, probably as a sign of adulthood (Kelly and Fu, 2014).

School-campus traffic congestion is a big problem for urban areas and a source of frustration on a daily basis. This kind of congestion is short lived, usually most situations do not last more than 15 minutes, but the potential of pedestrian-to-vehicle and vehicle-to-vehicle conflicts is significant. Obviously, there are serious safety issues while mixing automobiles with walking and biking children and buses bringing children to school (Tsai et al., 2004).

In order to improve road safety around school campuses, and according to the legislative framework of each state or country, several measures can be taken, such as (1) speed zones, usually around 15–25 mh^{-1} ($\sim 25\text{--}40 \text{ kmh}^{-1}$); (2) use of traffic control devices; (3) routing and traffic layouts that separate vehicle and pedestrian conflicts; (4) use of alternative modes, such as bicycles, carpooling, and on foot; (5) better information provision with Variable Message Signs (VMS), marked crossings, flashers, etc.; (6) establishing positive interaction with stakeholders, such as parents, transport engineers, teachers, municipal and local authorities, etc. (Institute of Transportation Engineers, 1998).

Carpooling

One way to reduce school campus traffic congestion is carpooling. Carpooling can be combined both with parent driven private car drop-offs and pick-ups and with student driven private car commuting. Nevertheless, queuing and congestion problems around school campuses are created that have to be faced.

In a carpool operation, vehicle–passenger match is the single most prominent procedure. The objective is to minimize the time vehicles spend in the loading area. Ideally, students should be ready to board as soon as the vehicle arrives at the loading bay. For this to happen, the correct students should be notified in time and the vehicle must also be referenced to the respective students. This information is usually communicated through school staff members. Moreover, many schools have staff members, or upper-level students, as safety assistants. These assistants open vehicle doors, assist students' boarding, and then close vehicle doors. This procedure helps to significantly decrease the pick-up time. Other measures, such as enforcement, restrictions, and queue lane management are also used to keep queue length under control and prevent road accidents (Tsai et al., 2004).

Other recommendations regarding school campus traffic congestion may include the following (1) parents should be discouraged arriving before the afternoon school bell time, in order not to sit and idle in the carpool lane; (2) the actual loading and unloading time of students should take less than 10 sveh^{-1} and the whole process of a single vehicle entering and exiting the loading area should be completed in less than 45 s; (3) loading bays should be clearly identified; (4) schools should deploy a passenger identification system assisted with school staff members; (5) parent drivers should be guided and have limited options and restricted access to prevent vehicle-to-vehicle and, especially, vehicle-to-pedestrian conflicts; (6) schools with large walker and cyclist populations should separate them from carpooling traffic, releasing walkers and cyclists first; (7) schools should also consider implementing walking and cycling programs as a means to mitigate carpool traffic; and (8) if necessary, schools should consider implementing double queuing lane (Tsai et al., 2004).

In any case, carpooling is preferable to simple parent driven private car drop-offs and pick-ups, and student driven private car commuting, and it should be encouraged in order to reduce the total number of cars needed to serve the same number of trips. Indeed, studies have shown that carpooling can reduce traffic congestion by half, having a major positive impact on road safety and the environment.

Active Mobility (Walking and Cycling)

One of the countries that has a strong tradition of children going themselves to school, without being accompanied by parents, is Sweden. Back in 1981, 94% of all children ages 6 to 17 years walked or rode their bikes to school by themselves or with friends. Nationwide surveys show that percentage had dropped to 77% by 2000, 67% by 2003, 47% by 2009, and 45% by 2014. In 2014, an additional 34% used public transportation (or in some areas before age 8, or where public transportation is nonexistent, taxi, or buses dedicated to school transport), so only 18% were taken to school by automobiles and no one in Sweden would drive themselves by car since the license age is 18 years. The percentage that walk or ride bicycles vary between rural and urban areas and with the size of a city and with the distance to school. High schools are often located so far away that walking is not practical (Björklid and Gummesson, 2013).

Although over the last three decades the proportion of children walking to school has declined significantly, not only in Sweden but also in other European and western countries, with accompanying evidence of a sharp increase in driving children to school, lately there is a trend in Europe to prohibit car drop-offs in school campuses altogether! Some schools in Edinburgh have prohibited car drop-offs, the London neighborhood of Hackney has shut down the roads outside some schools, as has Vienna, Austria, where cars are banned from the roads leading up to a school between 7:45 and 8:15 a.m., in order to make more space for safer and cleaner modes of transport, such as bicycles and walking. Moreover, it seems that this is only the beginning as many cities are expected to follow, in the general frame of promotion of sustainable urban mobility in Europe (Schmitt, 2018).

When people walk or bike to school, the creation of woonerven zones/shared spaces around school campuses should also be considered. Studies show that although there might be some increase in traffic accidents, though with minor damages, at the



Figure 1 Bike to school day, Gastineau Elementary School, Juneau, AK. Source: Alaska Department of Transportation & Public Facilities, 2015.

beginning of their application, mid and long-term road safety increases, as users understand how they should behave and drive in a woonerf zone. Nevertheless, there should be no through traffic in a woonerf zone and, therefore, no car drop-offs should be allowed, or these should take place outside of them.

Woonerven zones, shared spaces, bicycle tracks, sidewalks, and every kind of urban regeneration that promotes active mobility (i. e., walking and cycling) should be promoted as much as possible (**Fig. 1**). Although there are road safety considerations, active mobility is necessary for students in order to reduce obesity and promote a healthy lifestyle. Child obesity has become a real problem in the developed world, especially in the United States, and the same goes for unhealthy/sedentary lifestyles (e.g., time spend on watching television and streaming video, on social media and Internet surfing, playing video games, etc.). This is one of the reasons that new policies that promote active mobility in the general frame of urban sustainable mobility are introduced, especially in the European Union (EU), expressed through Green and White Papers.

It should be noted that it is not enough that the area in close vicinity of the school is designed as a safe woonerf zone; the entire walking distance should be safe. This means that elementary schools should be located so that children do not have to cross arterials on their way to the school. Moreover, all street crossings should occur in low-speed environments. As pointed out in the article Bicycle Collision Avoidance Systems of this Encyclopedia, [Johansson and Leden \(2010\)](#) concluded that pre-school children should not encounter any cars in their play areas or where they walk; and where children ages 7–12 years are crossing a street, by bicycle or on foot, the actual vehicle speeds should be below $15\text{--}20\text{ kmh}^{-1}$, especially if they are walking unaccompanied by an adult.

Apart from the perceived risk, from traffic as well as risk of assault, there are also objective reasons that active mobility cannot fully replace motorized mobility for school commuting. Several studies show that the distance to travel is the most significant on-off determinant of mode choice. Two (2) km is a guiding “splitting line,” or threshold, between the active modes of mobility and using transit or other motorized modes. Above 2 km, students rarely walk to school and this threshold offers an evidence-based recommendation for specific school bus services or other motorized approaches. Below 2 km, a “walking school bus” can serve as a functional substitute to private car drop-offs, for certain areas given the student densities and school locations ([Kelly and Fu, 2014](#)).

Walking School Bus

“Walking school bus” is an interesting concept (**Fig. 2**). Two adults, a “driver” that leads and a “conductor” that follows, walk students to school along a set route, in much the same way a school bus would drive students to school ([Moening et al., 2016](#)). Walking school buses, like traditional buses, have a fixed route with designated “bus stops” and “pick up times.” The benefits of walking school buses can be summarized as follows: they (1) encourage physical activity by teaching the skills to walk safely, how to identify safe routes, and the benefits of walking; (2) raise awareness of a community’s walkability and also where improvements can be made; (3) raise environmental concern; (4) reduce crime and take back neighborhoods for walkers; and (5) reduce traffic congestion, pollution, and speed near schools ([Campos, 2011](#)).



Figure 2 Walking school bus program in Hermosa Beach, CA. Credit: Beach Cities Health District. Source: *The Beach Reporter*, 2013.

School Bus

Nevertheless, for distances longer than 2 km, school buses are, and should remain, the main transport mode, in at least the United States and countries in which the main alternative is car. Despite the fact that there are exceptions, especially in safer countries like Sweden in which children up to age 10 years are expected to walk or bike no more than 2 km but at age 11–13 years that distance is extended to 3 km and above age 13 years to 4 km, and regardless of what parents believe, school bus is the safest mode of transport for children going between home and school (see the article School bus safety in this Encyclopedia).

Buses are also an effective means of reducing traffic both on roads and near schools, promoting, thus, environmental protection and road safety. There are several schemes of school bus ownership and operation: although many are owned and operated by school districts, others are controlled by third-party contractors who run them on behalf of schools (Schroth and Helfer, 2014).

It should be mentioned, though, that there is a clear difference between school buses in the United States and Europe, as well as the rest of the world. In the United States school bus vehicles and operation were evolved after a long tradition of standards' development. We should remember "the father of the yellow school bus" Frank W. Cyr and his efforts, that led to a set of 44 standards for school bus construction, which were adopted by all manufacturers. These 44 standards led to consistent interior dimensions, seating configurations, and other features that permitted new consistency in industrial production for manufacturers, leading to decreased complexity, which subsequently reduced the price per unit for purchasers (Fig. 3). The "National School Bus Glossy Yellow" was selected as the easiest color to see at both dawn and dusk and, although this was never officially adopted by the federal government of the United States, it has become the de facto choice for school buses both in the United States and the rest of the world. Moreover, regarding the accessibility of People with Disability (PwD), small school buses are often equipped with special devices to assist PwD, such as automated lifts (Schroth and Helfer, 2014).

The system of school transport in the United States differs from the school transport system in Europe and the rest of the world. In the United States, the school bus itself is an icon and functions as a "mobile traffic STOP sign," as it requires other vehicles to stop and not overtake it while children are boarding or alighting. In the United States, on a daily basis, more than 22 million children are transported to and from school in specially designed school buses. School bus is the safest transportation mode available for students to go to school in the United States; a student on the way to school is eight times more likely to be injured while traveling to and from school in a vehicle other than a school bus. Nevertheless, illegal overtaking of school buses remains problematic as, statistically, a child is three times as likely to be killed as a pedestrian in the "loading zone" around the school bus when the school

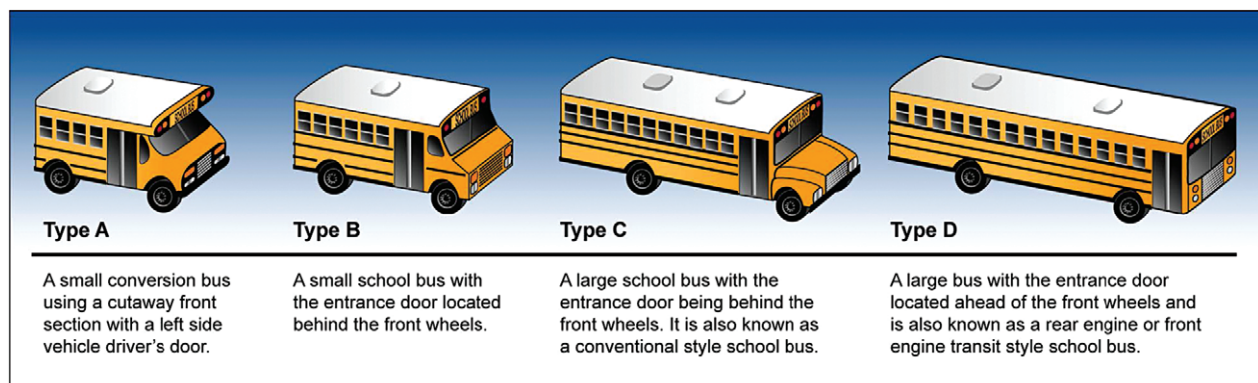


Figure 3 Types of school buses in the United States. Source: *United States Government Accountability Office*, 2017.

bus is present than to be killed as a passenger in the school bus. An interesting fact is that, apart from the illegal overtaking vehicles, many students are hit by the actual bus they intended to ride or rode with (Anund et al., 2011).

On the contrary, many of the features of the school transport system in the United States do not apply to most EU countries and the rest of the world. Instead, it is more common in EU to have a mixed system of school transport. Some students are transported by vehicles provided for school children, but they do not have any special features or traffic rules applying to their presence on the road, such as in the case of the United States, apart from specific EU regulations related to seat belts and the display of a school bus sign. Moreover, many students in EU use public transport for school commuting. Because of this mixed system, school-related EU injury statistics are more difficult to obtain. However, findings show that: (1) the majority of the children on school transport who are fatally injured are not vehicle occupants; (2) they are in the area outside the bus on their way to or from buses; (3) boys seem to be overrepresented; and (4) the afternoon journey is particularly vulnerable. Running out behind the bus in the afternoon was a common conflict scenario, and perhaps EU should adopt the practice of school buses in the United States that function as mobile traffic STOP signs (Anund et al., 2011). Parenthetically, on the other hand, official statistics of children ages 0–17 years killed in traffic from 2004 to 2008 was lower per capita in every single EU country than in the United States and the rate in the United States was 3.35 times that of Sweden's. Furthermore, Sweden, as opposed to the United States, has drastically improved its safety in the last decade, reducing the number of children killed in traffic from an average of 30.6 actual fatalities per year in the time period 2004–08 to 12.6 in 2013–17 in spite of a growing population. In addition, it should be noted that most children that are killed in traffic are 15 or older. In Sweden, only about one child per year is killed in each of the age groups 0–2, 3–5, 6–8, 9–11, and 12–14 years so about two-thirds of all child fatalities occur among 15–17 year olds. The fact that Europe has better traffic safety for children than the United States does not mean that Europe should not invest in safer school buses. It means that the United States needs to have more children use school buses and improve the walking and bicycle environment for children, as well as having them travel less by car, since the fatality rates above include all activities, not just going to school.

School bus occupants have a good safety record all around the world. There are several regional differences though. In the United States, there are three times as many children killed around school buses compared to those killed inside a school bus. In New Zealand, Australia, the United States, and Canada most school bus incidents happen while pedestrians are alighting from the bus in the afternoon. There is a notable difference, though—in the United States and Canada the majority of student pedestrian casualties are caused by their own school bus, whereas in Australia and New Zealand the students injured after alighting were more likely to be injured by another vehicle. Probably, this reflects the different approaches in traffic education—in the United States and Canada children are encouraged to cross the road in front of their bus, whereas in the United Kingdom, Australia, and New Zealand the norm is to teach students to cross the road after their bus has left the bus stop (Anund et al., 2011).

Studies show that children do not see school commuting as a part of their school day. Therefore, children when alighting feel that they enter “play time” and the result is a sort of cognitive distraction. Another explanation for the casualties when running out behind (or in front of) the school bus could be that children cannot perform beyond their degree of maturity. The frontal lobe, which is the brain region that is essential to foresee the consequences of behavior, is one of the latest parts in the brain that is developed. Consequently, children do not have fully developed the capability to foresee a possible critical situation when alighting. This fact raises the need for effective risk assessment and management of the location of school bus stops, to ensure they are not in complex traffic environments which children are inherently unable to cope with. The consistency of these casualty patterns in several regions emphasizes the need for effective road safety education, ensuring that school commuting is regarded as an integral part of the school day and, thus, under the school's responsibility (Anund et al., 2011).

Despite all the aforementioned problems with school bus commuting, it is clear that school bus is the safest mode of transport to and from school. It is important to keep in mind that most children are injured or killed as passengers in private cars or as vulnerable road users (i.e., while on foot or cycling). Actually, data from the National Highway Traffic Safety Administration (NHTSA) indicates that a child is 50 times more likely to be killed on the way to school if traveling in an automobile driven by themselves or their friends than if they take a school bus (Schroth and Helfer, 2014)!

Motives

Here comes the need to persuade both the parents and the adolescent student drivers that school bus is the safest mode of transport for school commuting. In this attempt, the role of motives is crucial. The perceived risk plays a very important role in modal choice for school commuting. Many parents believe driving their children to school is safer than letting their children ride the school bus, despite the contrary statistics.

Studies show that for each parent who believes that “school bus is very safe” two believe that “private car is very safe,” and, of course, parents who believe that private cars are more safe are more likely to ride in an automobile for school trips, whereas students whose parents believe the school bus or non-motorized modes to be most safe are more likely to ride the school bus. Parental perception of modal safety is paramount. If there is to be systemic change resulting in modal share shifts and in decreased traffic congestion around schools, addressing parents' subjective perceptions about the relative safety of the school bus and private car is critical. Most parents believe the automobile to be a “most safe” means of transport, although injury and fatality statistics suggest otherwise. In order to shift students from private cars to school buses, this bias has to change (Rhoulac, 2005)!

Some findings indicate that, in some cases, socioeconomic status, expressed in average median household income, is usually inversely proportional to the probability that a student will be driven in a private car for school commuting (Rhoulac, 2005). Although this could be explained by the probable longer home-to-school distances of wealthy suburbs, another reason could be that, due to better education, parents of higher socioeconomic status understand that school buses are safer than private cars.

Interface Management

Another very important aspect of school commuting road safety is the management of interfaces. Most accidents occur at interfaces in space and time. Examples of interfaces in space are carpooling loading bays, bus stops, pedestrian crossings, etc. Examples of interfaces in time are cases in which pupils feel that school time has finished and play or free time begins after alighting, etc. All these interfaces should be managed accordingly and one way to address this issue is to apply the road safety “travel chain perspective,” which means that every trip from and to school involves all necessary steps from “door to door.” For example, in the case of school bus, this means that going to and waiting at the bus stop, as well as boarding and alighting, are integral parts of the trip. The concept of road safety “travel chain” is derived from the concept of “accessible travel chain” and there are research projects that promote this approach, such as SAFEWAY2SCHOOL (Anund et al., 2012). By the way, obviously commuting travel chain should be accessible for all.

Conclusion

In summary, there is a need of an integrated approach to school commuting—(1) all private car drop-offs should be shifted to carpooling or school buses; (2) gradually even carpooling drop-offs should be shifted to school buses; (3) an exception to the previous guidelines is the trips realized with active mobility modes, such as walking and cycling, which have to be promoted; (4) even in the case of distances greater than 2 km safe active mobility measures should be promoted, such as alighting students at some distance before schools in order to be forced to walk or cycle to the school in woonerven zones or safe pedestrian and bicycle tracks; and (5) all the previous measures should be taken in an integrated frame of an accessible and safe “travel chain perspective.”

Tertiary School Campus Traffic Circulation

In the case of colleges and universities, almost all students are adults and their commuting does not differ so much compared to typical commuting of employees. Aspects closer to pre-tertiary education commuting have already been covered in the previous part of this article. Obviously, in the case of tertiary school campus traffic circulation, the most important part is internal traffic, that is, the traffic circulation inside the campuses, which in most cases resemble to small villages. However, many campuses have parking lots located on the outskirts of the campus, and most serious accidents occur not in the center of campus but when students cross arterial or collector roads walking from parking lots to classrooms. So, any improvement plans need to include such streets.

There are several approaches to college and university campus traffic circulation. Most of them can be grouped in the following three categories: (1) segregated traffic, (2) no traffic (i.e., full pedestrianization), and (3) mixed traffic (i.e., woonerf zone).

Segregated Traffic

The proponents of segregated traffic in campuses usually are transportation engineers with a conventional traffic engineering approach. According to this approach, creating segregated traffic lanes for cars, bicycle tracks, pedestrian sidewalks, and, in general, segregating traffic increases road safety and creates a viable campus for every type of user. Although this approach seems to be the dream of every transport engineer, who sometimes consider campuses as playgrounds in which they can make their dream come true, it has some drawbacks. More specifically, a lot of space is needed to segregate traffic, and especially in the case of most European campuses simply there is not enough. Moreover, segregated traffic lanes for cars have as a result speeding inside campuses; remember that most drivers in campuses are novice young drivers, not to mention that half of them are men. Therefore, although the approach of segregated traffic seems to be ideal at first glance, in real life this is not always the case.

No Traffic

The opposite approach is that of no traffic at all or, else, full pedestrianization. Usually, the proponents of this approach are environmentally aware members of the academic community and, indeed, in case there are enough parking spaces around the campus or adequate public transport service, this is the best solution. There are beautiful campuses all around the world that allow no traffic at all. Given that there are solutions for PwD and the elderly (Demir-Mishchenko et al., 2010), perhaps this should be the ultimate goal of every college or university campus. To achieve such a goal, undergrounding the necessary parking facilities should be considered but parking spaces should not be so many that would generate additional trips to campus. In the case of larger campuses, maintenance traffic as well as bus traffic may be needed, but such traffic can be restricted to occur only during class times and banned during the 10 or 15 minute periods when students walk between classes.

Mixed Traffic

Nevertheless, in case there are no parking lots around the campus for park-and-walk or adequate public transport service, the pressure to allow internal traffic in campuses is very strong and, sometimes, justified. In such campuses, especially if they are located



Figure 4 Woonerf zone in Downtown Raleigh, NC. Source: *Institute for Transportation Research and Education, 2013.*

in urban areas where public space is limited, there is an ongoing battle between drivers and pedestrians. In these cases, the woonerf zone approach seems to be a feasible and viable solution, a “via media” between the proponents and opponents of private cars, compared to infeasible maximalist proposals of full pedestrianization (Nalmpantis et al., 2017; Vasileiadis and Nalmpantis, 2019). Moreover, inside a woonerf zone human interaction is not disturbed by traffic (Fig. 4).

The beauty of the true woonerf zone, as it was defined by law in the Netherlands in 1976, is that the maximum speed allowed is qualitatively defined: “Article 88b KVV: Drivers within a ‘Woonerf’ may not drive faster than at a walking pace” (Ben-Joseph, 1995). As the average walking pace is very slow, there were interpretations of that being the pace of a walking horse, that is, around 10 kmh^{-1} ($\sim 7 \text{ mh}^{-1}$), and currently even 15 kmh^{-1} ($\sim 9 \text{ mh}^{-1}$), that is close to the speed of a racewalker, is considered to be acceptable. In the United States, shared spaces typically allow an upper speed limit of only 20 mh^{-1} ($\sim 30 \text{ kmh}^{-1}$), which is the maximum theoretical running speed of human beings; according to evolutionary psychologists our psychology and physiology have evolved based around that while hunting, therefore an individual’s ability to interact and retain eye contact with other human beings diminishes rapidly at speeds greater than 30 kmh^{-1} . Obviously, it is not a coincidence that safety analysts have known for several decades that the maximum speed at which pedestrians can escape severe injury upon impact in most cases is just under 30 kmh^{-1} ; that or less should be the upper speed limit in order to have an even lower possible impact speed (as recent data from the United States show that if hit by a motor vehicle at 30 kmh^{-1} young adults and young children with the elderly have a 2% and 15% fatality rate, respectively).

Conclusion

In summary, the management of internal college or university campus traffic circulation depends on the peculiarities of every institution. Preferably, the solution of full pedestrianization should be pursued but only in case there are parking lots around the campus for park-and-walk, adequate public transport service, and provisions for PwD and the elderly. In case these conditions are not met, the second best solution is a woonerf zone.

Biography

Dr. Dimitrios Nalmpantis is a member of the Laboratory Teaching Faculty of the School of Civil Engineering of the Aristotle University of Thessaloniki (AUTH) and a member of the Adjunct Teaching Faculty of the Hellenic Open University (HOU). He holds a PhD in Transportation and Road Safety, an MA, an MBA, and an MEng. He has participated in more than 25 research projects and he has published more than 60 peer-reviewed papers in scientific journals, conferences, and edited books as chapters.

See Also

School bus safety; Bicycle collision avoidance systems; Pedestrian safety, children; Age and Gender as factors in road safety; Safety culture; Transport modes and commuters; Shared space; Speed limits on urban streets

References

- Alaska Department of Transportation & Public Facilities, 2015. Gastineau Elementary Bike to School Day [Image]. Available from: <https://flic.kr/p/sbMSXk>.
- Anund, A., Chalkia, E., Nilsson, L., Diederichs, F., Ferrarini, C., Montanari, R., Strand, L., Aigner-Breuss, E., Jankowska, D., Wacowska-Slezak, J., 2012. SAFEWAY2SCHOOL (Final Report). Available from: http://www.safeway2school-eu.org/assets/docs/deliverables/SW2S_D10.5.pdf.
- Anund, A., Dukic, T., Thornthwaite, S., Falkmer, T., 2011. Is European school transport safe?—The need for a “door-to-door” perspective. *Eur. Transp. Res. Rev.* 3 (2), 75–83, doi:10.1007/s12544-011-0052-7.
- Ben-Joseph, E., 1995. Changing the residential street scene: adapting the shared street (Woonerf) concept to the suburban environment. *J. Am. Plann. Assoc.* 61 (4), 504–515, doi:10.1080/01944369508975661.
- Björklid, P., Gummesson, M., 2013. Children's independent mobility in Sweden. The Swedish Transport Administration, Borlänge, Sweden. Available from: https://trafikverket.ineko.se/Files/sv-SE/11521/RelatedFiles/2013_113_childrens_independent_mobility_in_sweden.pdf.
- Campos, D., 2011. *Jump start health!: practical ideas to promote wellness in kids of all ages* Teachers College Press, New York, NY.
- Demir-Mishchenko, E., Zorlu, F., Tsalis, P., Naniopoulos, A., Nalmpantis, D., 2010. ACTUS Guidebook: Accessibility of University Campus for People with Disabilities. Mersin University Press, Mersin, Turkey.
- Institute for Transportation Research and Education, 2013. ITRE Downtown Raleigh Bicyclists Woonerf Pedestrian Plaza [Image]. Available from: <https://flic.kr/p/y5raq5>.
- Institute of Transportation Engineers, 1998. Survey of traffic circulation and safety at school sites. Institute of Transportation Engineers, Washington, DC. Available from: https://safety.fhwa.dot.gov/ped_bike/docs/schoolsurvey.pdf.
- Johansson, C., Leden, L., 2010. Child pedestrians' quality needs and how these needs relate to interventions. 11th International Walk21 Conference and 23rd International Workshop of the International Co-operation on Theories and Concepts in Traffic safety, The Hague, the Netherlands, 17-19 November 2010. Available from: http://www.ictct.org/migrated_2014/ictct_document_nr_718_404A%20Lars%20Leden%20Child%20Pedestrians.pdf.
- Kelly, J.A., Fu, M., 2014. Sustainable school commuting – understanding choices and identifying opportunities: a case study in Dublin, Ireland. *J. Transp. Geogr.* 34, 221–230, doi:10.1016/j.jtrangeo.2013.12.010.
- Moening, K., Lieberman, M., Zimmerman, S., 2016. Step by Step: How to Start a Walking School Bus. California Department of Public Health, Sacramento, CA. Available from: https://www.saferoutespartnership.org/sites/default/files/resource_files/step-by-step-walking-school-bus-2017.pdf.
- Nalmpantis, D., Lampou, S.-C., Naniopoulos, A., 2017. The concept of woonerf zone applied in university campuses: the case of the campus of the Aristotle University of Thessaloniki. *Transp. Res. Proc.* 24, 450–458, doi:10.1016/j.trpro.2017.05.071.
- Rhoulac, T.D., 2005. Bus or car?: the classic choice in school transportation. *Transp. Res. Rec.* 1922 (1), 98–104, doi:10.1177/0361198105192200113.
- Schmitt, A., 2018. The European answer to school-drop-off chaos. Available from: <https://usa.streetsblog.org/2018/11/27/the-european-answer-to-school-drop-off-chaos/>.
- Schroth, S.T., Helfer, J.A., 2014. School Bus Safety. In: Garrett, M. (Ed.), *Encyclopedia of Transportation: Social Science and Policy*. SAGE, Thousand Oaks, CA, pp. 1202–1206.
- The Beach Reporter, 2013. 'Walking School Bus' finishing year after making strides in beach cities. Available from: http://tbrnews.com/news/redondo_beach/walking-school-bus-finishing-year-after-making-strides-in-beach/article_32cf9efc-d2b5-11e2-8732-0019bb2963f4.html.
- Tsai, J., Cranford, J., Lee, J.-J., 2004. Best practices in managing school campus traffic circulation. *Transp. Res. Rec.* 1865 (1), 41–47, doi:10.3141/1865-07.
- United States Government Accountability Office, 2017. School bus safety: Crash data trends and federal and state requirements. United States Government Accountability Office, Washington, DC. Available from: <https://www.gao.gov/assets/690/682077.pdf>.
- Vasileiadis, I., Nalmpantis, D., 2019. Development of a methodology, using Multi-Criteria Decision Analysis (MCDA), to choose between full pedestrianization and traffic calming area (woonerf zone type). In: Nathanail, E., Karakikes, I. (Eds.), *Data Analytics: Paving the Way to Sustainable Urban Mobility. CSUM 2018. Advances in Intelligent Systems and Computing* (Vol. 879). Springer, Cham, Switzerland, pp. 315–322, doi: 10.1007/978-3-030-02305-8_38.

Further Reading

- Falkmer, T., Horlin, C., Dahlman, J., Dukic, T., Barnett, T., Anund, A., 2014. Usability of the SAFEWAY2SCHOOL system in children with cognitive disabilities. *Eur. Transp. Res. Rev.* 6 (2), 127–137, doi:10.1007/s12544-013-0117-x.
- Kte'pi, B., 2014. School Transportation. In: Garrett, M. (Ed.), *Encyclopedia of Transportation: Social Science and Policy*. SAGE, Thousand Oaks, CA, pp. 1209–1211.
- Pitsiava-Latinopoulou, M., Basbas, S., Gavanis, N., 2013. Implementation of alternative transport networks in university campuses: the case of the Aristotle University of Thessaloniki, Greece. *Int. J. Sustain. High. Educ.* 14 (3), 310–323, doi:10.1108/IJSHE-12-2011-0084.
- Thornthwaite, S., 2009. *School Transport: Policy and Practice*. Local Transport Today, London, United Kingdom.
- Toor, W., Havlick, S., 2004. *Transportation and Sustainable Campus Communities: Issues, Examples, Solutions*. Island Press, Washington, DC.
- Transport, Innovation and Systems, 2004. Road safety in school transport (Final Report). Available from: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/projects_sources/rsst_final_report_v1.3.pdf.
- Transportation Research Board, 2002. The relative risks of school travel (Special Report 269). Transportation Research Board, Washington, DC. Available from: <http://onlinepubs.trb.org/onlinepubs/sr/sr269.pdf>.

Sexual Violence in Public Transportation

Vania Ceccato, Department of Urban Planning and Built Environment, KTH Royal Institute of Technology, Stockholm, Sweden

© 2021 Elsevier Ltd. All rights reserved.

The Nature of Sexual Violence in Public Transportation	576
Why Care About Sexual Violence in Public Transportation?	576
The People Most Affected by Sexual Violence in Transit Environments	577
Gender	577
LGBTQI Status	578
Age and Disability	579
Previously Victimized	579
Ethnicity	579
Socioeconomic Status	580
Underreporting	580
Situational Risky Conditions for Sexual Violence	580
Selection of International Evidence	580
The Impact of Sexual Victimization	581
Public Transport, Sexual Violence, and Prevention	581
Emerging Issues and Future Debate	581
Biography	582
Relevant Website	582
References	582
Further Reading	583

The Nature of Sexual Violence in Public Transportation

Sexual violence can be a vast array of sexual behaviors that vary from sexual harassment to sexual assault. The boundaries between these types of acts are blurred. In the literature, sexual harassment is said to be unwanted sexual attention, while sexual assault takes place when a person is threatened, coerced, or forced into a sexual act against her/his will. Cohan and Shakeshaft (1995) already in the 1990s distinguished between what they called noncontact and contact sexual violence. In the noncontact category, they identify nonverbal sexual abuse (such as exhibitionism, showing sexually explicit pictures, or making gestures) and verbal sexual abuse (such as sexual comments, jeering or taunting, and asking questions about sexual activity). In the contact category, they include sexual abuse such as touching, kissing, and rape. Fig. 1 illustrates a list of acts, which include a vast array of sexual behaviors under the denomination “sexual violence.”

Although not well understood, the risk of victimization of sexual violence is expected to vary across types of individuals (age, gender, (dis)ability, LGBTQI status, IT literacy, previous victimization, frequency of use of public transportation, length of trip, etc.), over time (rush hours, off-peak hours, days of the week, weekends, weekdays, and seasonal differences), and by different public transportation modes and settings along the trip (Gekoski et al., 2015, p. 6). Since some transit environments are often places of convergence, crime opportunities, including sexual violence, are inherent parts of the daily rhythmic flow of people passing by in these transit environments. Public transport or public transportation is the term used here to capture what North American readers often call “public transit,” “mass transit,” or “rapid transit” systems (Newton, 2014, p. 709). These systems compose forms of transport that are available to the public, charge set fares, and run on fixed routes, such as trains, buses, and trams. When safety is associated to public transportation, the journey not only involves the bus ride or the trip by train; it includes a whole journey perspective; in other words, this perspective considers safety throughout the whole trip, from door to door (Natarajan et al., 2017). For instance, safety in the environments to which an individual is exposed when walking to the bus stop, waiting for the bus to come, inside the carriage on the trip, changing transportation modes, arriving at a destination, and returning back home. In order to obtain a better understanding of risk of sexual victimization in public transportation, it is necessary to untangle the intersection between riders’ individual characteristics and the characteristics of the multiplicity of environments they are exposed to from door to door, at a particular time, and all possible times they would make such a trip.

Why Care About Sexual Violence in Public Transportation?

Firstly, sexual violence is highly gendered. Worldwide, women more often are exposed to unwanted sexual behavior than men are while in transit. This is partially because women travel by public transport more often than men do and, in some countries, they



Figure 1 Types of sexual violence in public places. *Source: Adapted from RedDot Foundation SafeCity initiative.*

constitute the so-called “transit captives,” which means they have relatively less access to nonpublic forms of transportation and are therefore overly reliant on it. For example, in Sweden, previous research has shown that women are less likely to have access to a car (Lundkvist, 1998), and tend to take more, shorter and more varied trips (trip chains) at more varied times (although they tend to travel less at night) (Kunieda and Gauthier, 2007; Ceccato, 2013, 2014b). In the United States, a study showed that 64% of the people riding Philadelphia’s subways and buses were women (Goodyear, 2015). That share is 62% in Chicago’s MTA and 60% in Washington, DC, New York, and Boston. Being a victim of sexual violence can have serious psychological and behavioral effects and may result in women and members of vulnerable groups (e.g., disabled people, LGBTQI) being too afraid to use buses or trains. If public transportation is not safe, or rather perceived not to be, the mobility of half of the population of the globe may be impaired, which directly affects individuals’ life opportunities (Junger, 1987).

Secondly, sexual violence has become more and more part of the public transportation agenda. The #MeToo! Movement has shown that sexual violence includes more than sexual harassment in the workplace, even including other unwanted types of sexual behavior that were in the past normalized by social and cultural norms in society. An example of this growing interest is a recent special issue that brought together 10 articles that characterize women’s victimization and safety in transit environments Ceccato, (2017b).

Thirdly, cities are to be planned having the UN 2030’s sustainability goals in mind (UN, 2019). If one wants to promote the use of more sustainable transportation (such as buses and trains but also walking and bicycling), personal safety must become a priority. This is particularly true for women and other vulnerable groups. Recent research shows that in order to foster transit use, personal safety has to come first and include an increasing public awareness of transportation-related environmental issues (Hsu et al., 2019). A non-neutral gender agenda for transit safety does not mean the implementation of radical solutions such as “women-only cars” or “women-only waiting rooms.” Although women-only transport may be an effective means of reducing unwanted sexual behavior on public transport, such measures are, at best, “short-term fixes” to the problem that at the end may penalize women by limiting their choices of transportation (limited supply of cars in particular time-frames). Moreover, there is an underlying assumption that these solutions may not be appropriate for countries where current gender inequalities do not result in extreme, unsafe transit conditions for women. Thus, there is an urgent need for a better understanding of the societal, economic, and institutional barriers that hamper the incorporation of women’s voices and views from vulnerable groups into transport service policies.

The People Most Affected by Sexual Violence in Transit Environments

Gender

Studies have shown that while men are more often crime victims on public transport (Morgan and Smith, 2006) women declare being more fearful than men (Dymén and Ceccato, 2012; Ceccato et al., 2013; Loukaitou-Sideris, 2016). That may most likely be because women are more exposed to sexual violence, and particularly in transit environments (Ceccato, 2013, 2014b; Ceccato and Loukaitou-Sideris, 2019; Moreira and Ceccato, 2019). In Britain, research has found that around 15% of women and girls have been subjected to unwanted sexual behavior on the London transport network, the vast majority of which goes unreported (Transport for London [TfL], 2013). A recent international survey performed among college students by Ceccato and Loukaitou-Sideris shows that victimization can reach more than half of the consulted population and gender is the most powerful predictor of sexual victimization (Fig. 2A, B). Although the sample may not be representative for the general youth population in some of the cities studied, the survey provides good indications of how sexual victimization affects those most targeted by this crime, specifically the youth

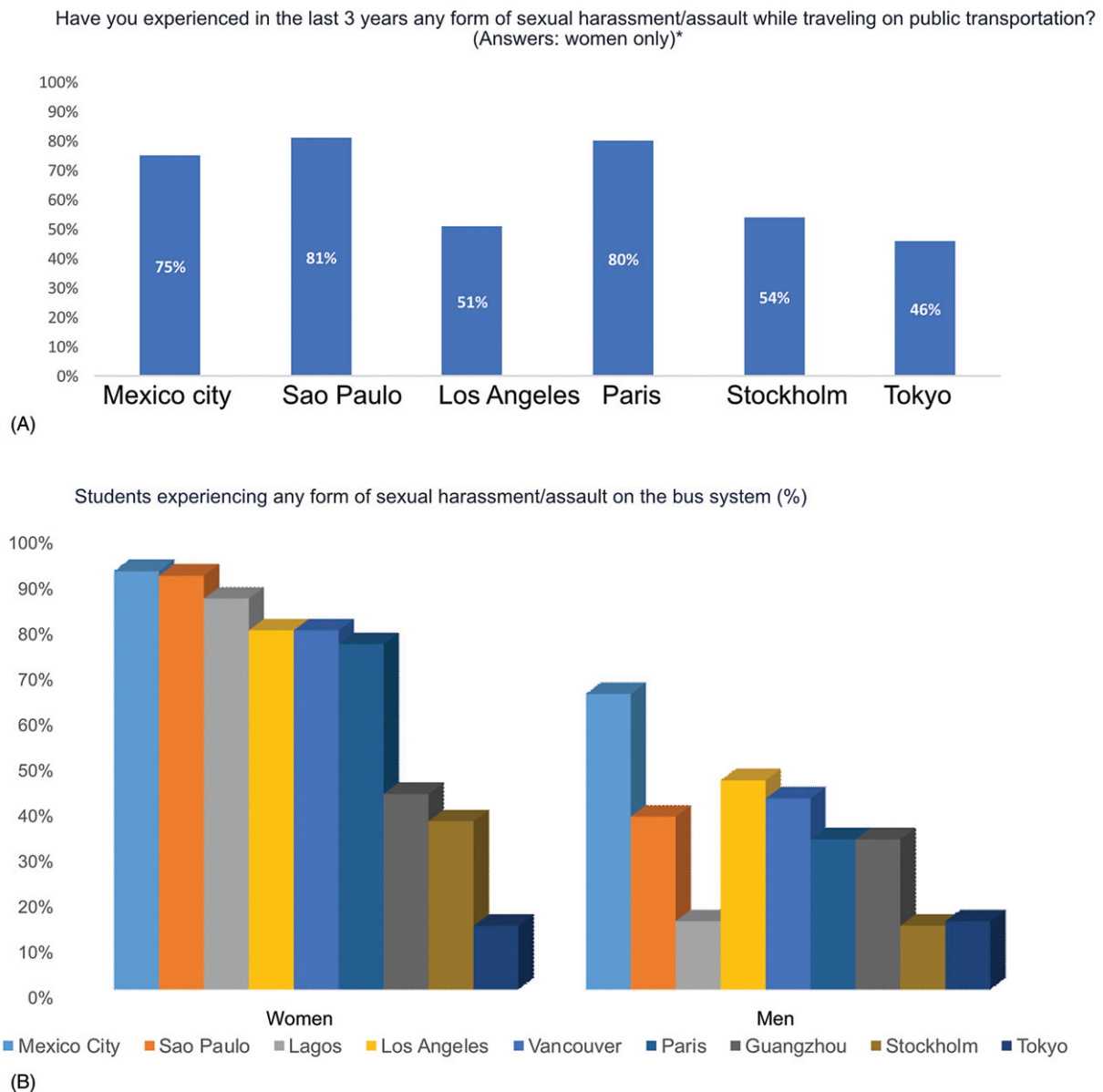


Figure 2 (A) Sexual violence victimization among university students in the transport system (on trains/buses and on the way to them, women only) in selected cities (% of respondents) and (B) sexual victimization on buses in selected cities (% of respondents). Note that samples vary from 350 to 1,200 students and that the cities differ in their public transportation system. Source: *Ceccato and Loukaitou-Sideris (2019)*.

population. In Stockholm, for instance, sexual violence on trains reaches 43% of those surveyed (62% women, 12% men) and for women only in public transportation, sexual victimization reaches 54%, but, as [Fig. 3](#) shows, it varies by type (verbal, nonverbal, or physical sexual violence) and whether the young rider is on the bus or on the train. Some of these events, such as rapes, cause a long-lasting trauma in the victims. In Stockholm municipality, reports of rape (including attempts) reached around 600 cases in 2016, about 30% of these cases happening outdoors committed by a person unknown to the victim. In a sample of 67 of these cases (all women) 2008–9, Ceccato in 2014 showed that 76% of these cases took place less than 500 meters from a bus stop and 53% of them less than 500 meters from a metro/train station ([Ceccato, 2014b](#)).

LGBTQI Status

In a study in the United States, [Lubitow et al. \(2017\)](#) found that transgender and gender nonconforming individuals often experience harassment, which undermines their access to safe public transportation. Similarly, evidence from the Global South study shows that sexual violence does differ among members of the LGBTQI community (gay women being more exposed to sexual harassment

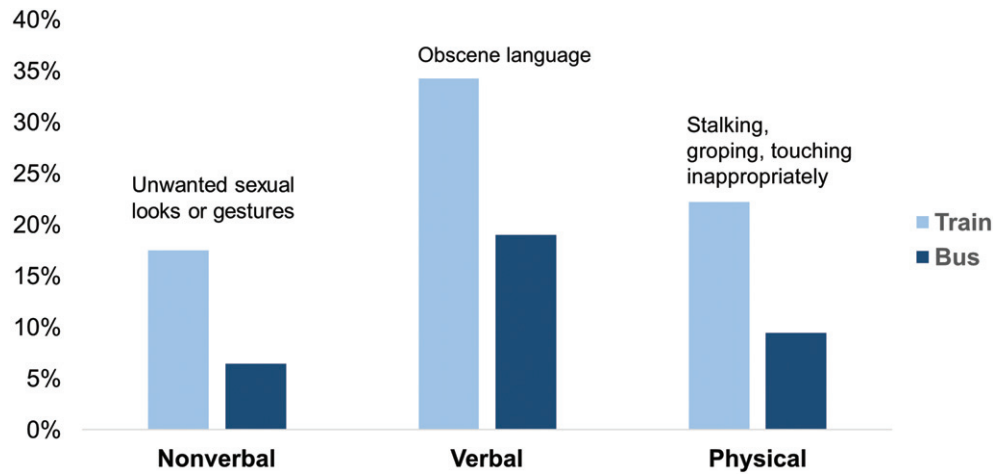


Figure 3 The nature of sexual violence by type and public transportation mode in Stockholm, Sweden. “In the last three years, have you experienced any of the following while traveling on, heading to, or waiting for the bus (please mark all that apply)?” Sample of university students only. $N = 991$ university students in rail bound public transportation and $N = 853$ university students in buses, Stockholm, Sweden, 2018.

than cisgender women) (UOL, 2014), but the main overall message is that LGBTQI as a group composes a vulnerable target for harassment, abuse, and potential street crime in public places (Doan, 2007). The interactions of LGBTQI status with ethnicity and socioeconomic status are also reported in the current international literature.

Age and Disability

Research shows that young women are more likely to be victimized by sexual violence in public places than are older women, including in public transportation (Whitzman, 2007; Madan and Nalla, 2015). As suggested by Gray (2018), young women in particular have to do their “safety work” before leaving their homes, or commuting to college, work, and recreation. In Sweden, the 2018 national crime victims’ survey showed that the vulnerability for sexual crimes differs greatly between different age groups. The proportion is highest in the age group 20–24 years for both women and men. In this age group, 36% of women state that they were subjected to sexual offenses while for men the figure is close to 5% only. Note that in the oldest age group for both men and women (75–84 years), only 0.4% state that they were exposed to sexual victimization in that country (BRÅ, 2019). For perceived safety, other studies found that older women are more likely to be afraid of victimization in transit environments (e.g. Levine & Wachs 1986).

With age, we may face a number of disabilities (Sochor, 2015), increasing the risk for unwanted sexual violence. As Iudici and colleagues indicate, disability is associated with a higher incidence of victimization, and scholars find that women with disabilities are among those who suffer the most from sexual assault in transit environments (Iudici et al., 2017). Evidence from Sweden by Ceccato et al. (2013) indicates the geography of fear among older adults follows a distance decay from their own residence; this means that older adults become more dependent on their own familiar surroundings to feel safe, although victimization may not necessarily follow the same pattern, happening often at home (e.g., residential burglary).

Previously Victimized

The chances of one being sexually victimized in transit are positively correlated with previous victimization of sexual violence (Ceccato, 2014a; Ceccato et al., 2017; Ceccato et al., 2018). This increased risk of sexual victimization is linked to an individual’s personal profile, including lifestyle choices and/or the environmental contexts she or he is exposed to.

Ethnicity

An individual’s ethnic background (here loosely associated with one’s “race” and “ethnicity”) may also be a key factor explaining one’s experiences of sexual violence (Sommers et al. 2006). The recent study across six continents of sexual violence among college students shows evidence of how ethnicity contributes to increased risk of sexual victimization and how such risk varies by transportation mode, along the trip, and by city and country contexts. This increased risk of victimization affects one’s perceived safety and may impair daily mobility. A study in the United States by Loukaitou-Sideris showed, for instance, that nonwhite women often experience higher levels of fear than white women experience, and tend to be fearful of white men, while white women tend to fear nonwhite men, which may be the result of both racial prejudice and persistent racial hierarchies (Loukaitou-Sideris, 2004; Ceccato, 2016; Newton, 2018).

Socioeconomic Status

The #MeToo! Movement has shown that sexual violence is an issue that transcends socioeconomic status. However, being often transit captives, poorer women worldwide may be more exposed to sexual violence than any other group. Evidence about the individual impact of income and/or socioeconomic status on sexual violence is relatively rare. In the United States, for instance, research has found that low-income women were more fearful of crime, assaults in particular, in public spaces and transit environments than higher-income women because they live in high-crime and unsafe neighborhoods (Loukaitou-Sideris 2005).

Underreporting

Unwanted sexual behavior is highly underreported (Gekoski et al., 2015). It is believed that around 15% of Londoners have experienced unwanted sexual behavior on the public transport network, but 90% of those affected do not go on to report it to the police. Most people report to family and friends but are reluctant to report to security guards or go to the police. Ball and Wesson in 2017 showed that passenger density in a transit setting influences the incident seriousness and also the likelihood of reporting an event (Ball and Wesson, 2017). The reasons for not reporting varied, but often riders thought the crime was not serious, it was a “waste of time” (as they think no one would be able to catch the abuser), or they did not want to create “more problems” or be reminded of the event. The embarrassment felt by victims of sexual harassment due to cultural pressure may also result in underreporting (Gray, 2018). Yet, despite the common occurrence, the majority of cases of sexual violence go underreported, as these unwanted sexual behaviors become normalized through individuals’ actions and experiences, becoming an invisible problem (Mellgren et al., 2018). High rates of underreporting mean a lack of reliable information about sexual violence in public transportation, constituting a key barrier to its prevention (Solymosi et al., 2018).

Situational Risky Conditions for Sexual Violence

Many of us would associate a crowded bus with groping and an empty station with the necessary conditions for indecent exposure. Research shows that some sexual behaviors actually “demand a crowd” while others “require anonymity” to happen. Less known is whether and how transit environments make sexual violence possible and, in other situations, they can prevent it. The internal design of carriages and buses, the physical and social environment of transport nodes (such as stations and/or bus stops), as well as the socioeconomic and land use contexts of transport nodes impact on the levels and types of crime that happen in these transit environments (Madan and Nalla, 2015) Newton (2018). Sexual violence is no exception. Studies based on traditional environmental criminology theories that associate the characteristics of the environment (in which transportation nodes are located) to poor social control and crime (e.g., broken windows theory, spiral of decay, crime prevention through environmental design, routine activity, social disorganization theory, situational crime prevention, etc.) provide a framework for the understanding of the situational conditions of sexual violence in transit environments as illustrated later.

Selection of International Evidence

High-density environments (in particular, overcrowded carriages) are a facilitator of crimes against women—a condition that may also trigger bystanders’ interventions. In an Indian study that explored actual and witnessed victimization as well as risk perceptions of a sample of female students, Tripathi et al. (2017) showed that sexual harassment appears to be most prevalent on buses, and increases with the frequency of public transport use. More interestingly, they found these events are not widely noticed by other passengers. There is evidence in the literature that most people pretend not to witness what is happening, or they declare not being ready to intervene. Other studies have investigated which strategies are the best for women when they are sexually assaulted in buses, for instance in Lea et al. (2017). In this particular study, women’s strategy of confronting incidents by “making a scene” and “engaging the crowd” plays in their favor in the closed, shared-space setting, such as crowded buses.

In the United States, previous research shows that commercial areas near transit stops, especially liquor stores, bars, and check-cashing establishments, as well as residential environments, particularly multifamily housing, are associated with higher crime rates at transit stops as compared with office and industrial environments, and such differences may be explained by higher density in the former (Loukaitou-Sideris, 2012). In the Sao Paulo metro, in Brazil, sexual violence constitutes 12% of total reported crime—crimes that are highly underreported. Sexual violence there is linked to the characteristics of these transport nodes (the presence of employees in the station, the presence of CCTV, type of station, etc.) and the different types of criminogenic land uses and socioeconomic demographics of the neighborhoods (poor areas, located close to a parking lot) in which public transportation operates. (See the recent study by Ceccato and Paz (2017) and Moreira and Ceccato (2019). Both studies indicated concentrations of sexual violence in space and time (the busiest central stations, often during morning and afternoon rush hours) for women.)

In a series of studies on rape in Stockholm, Sweden, Ceccato and colleagues found that subway stations increase the risk of outdoor rape in an area by over 50% with 95% credible interval (Ceccato et al., 2018). Note that rapes did not happen at the stations per se but on the way to or from these transport nodes. Using another data source, they showed that rape was significantly more likely to occur at a time when women were passing by a forested area, in an interstitial area (between buildings, desolated area), or in

private transportation (taxi, other private mode) than when they were in an indoor public setting (Ceccato, V., Wiebe, D.J., Eshraghi, B., Vrotsou, K., 2017). The study also showed significant variations over weekday and weekend, as well as seasonal patterns using police recorded data. This analysis, focusing on innercity areas, showed that rape typically happens in secluded inner courts between buildings, walkways, hilly areas where visibility from the street is difficult (stairs), ditches, tunnels that link to public transportation, and small parks, including cemeteries. In the periphery, the typical locations for rape included interstitial spaces between roads and buildings, often paths or streets with poor surveillance and close to transport nodes (Ceccato, 2014a). They also showed that the situational conditions 12 hours before a rape affected the risk (odds) of the rape taking place—findings that point out the importance of adopting a whole journey approach to ensure women’s safety. Informed by a “whole journey” approach to safety that includes walking to and from bus/subway stops as well as waiting for and riding the bus or subway Natarajan and colleagues assessed young women’s experiences of sexual victimization during the commute to college in New York, United States. Their findings showed the extensive patterns of victimization during all stages of the trip to college (Natarajan et al., 2017).

The Impact of Sexual Victimization

Transit users respond to the risk of transit crime by engaging in defensive behaviors. Common precautionary measures include changing travel time/mode/route, traveling in groups, choosing other times of the day, staying home, selecting defensible positions within a carriage, standing near “safe” people (typically other women), or using objects to shield themselves. According to Kash (2019), these behaviors allow some victims to regain a sense of safety while using public transportation while others may struggle with hypervigilance, anxiety, and other responses to sexual trauma. A recent study in France (d’Arbois de Jubainville and Vanier, 2017) assessed whether and how female passengers change their routines when feeling unsafe in the transit environment in the Paris region. They suggested that women’s education, previous victimization, and declared perceived safety are consistently associated with time-based and space-based avoidance.

Public Transport, Sexual Violence, and Prevention

A wide range of interventions have been introduced through particular programs that often encompass a bundle of measures to tackle sexual violence. Common safety interventions may include design strategies (e.g., improving lighting, decreasing disruptive objects, increasing visibility, good maintenance of transport facilities, implementing real-time scheduling information), passenger separation strategies (e.g., Women-only Cars/taxis, splitting passenger flows), improvement of surveillance (e.g., CCTV, the presence of security officers, training of personnel and passengers to be alert and intervene, hotlines, implementation of emergency buttons in carriages, SMS services), targeting routes (hot spots known to have higher crime rates), raising awareness (campaigns to motivate individuals to report, organizing public workshops, actions through social media in collaboration with existing campaigns used in raising awareness about sexual violence in public spaces). There is limited evidence about the effectiveness of these interventions, in particular campaigns aiming at public education and awareness raising.

An example of a campaign in Sao Paulo showed limited incentives for sexual abuse detection, and by doing so it is neglecting victims’ needs and support. Another problem has been that campaigns of this type rely heavily on users to report these offences, but very little is suggested about the transport operators’ role after the event has already happened. Some claim that these campaigns take away the transportation company’s responsibility and move it to victims and passengers, to deal with the problem (Ceccato and Paz, 2017).

Emerging Issues and Future Debate

Assuming that sexual victimization is never justified, it is important in future research to improve our knowledge about women’s behavior (over which women are free to have total autonomy and carry out in the absence of the threat of sexual victimization) in situational contexts in which women become a target in public transportation. This is of relevance since the identification of these conditions is crucial to create opportunities for prevention. Fundamental in this process is to incorporate women’s voices and other vulnerable group’s views into public transport services and policies.

In practice, adopting a “whole journey approach” to transit safety means that we must take into account aspects of the transit journey for female passengers and other vulnerable groups that are specific: some demanding a combined attention from transportation providers, local governmental authorities (including police departments), policy-makers, and other stakeholders.

Finally, increasing the safety of passengers has, for various reasons, not been high on the agenda of transport agencies and governments, often adopting gender-neutral policies, rarely sensitive to the needs of those who are often the target of sexual violence in transit environments (Loukaitou-Sideris, 2014). A non-neutral gender policy does not need to lead to “Women-only cars” or other extreme solutions that are found in some transportation systems around the world. Aligned to the UN’s 2030 sustainability goals, a more progressive agenda in transportation policy is needed. As previous research suggests, transit agencies often have a general interest in providing adequate services to all customers, but unless a specific effort is made to ask those riders that are more often the target (of sexual violence), what they need and want and how services can be improved, suboptimal solutions persist. Solutions may

require structural changes through, for instance, recruitment of female personnel in transportation systems, often a male dominant environment (Rivera, 2007). It is expected that as more women and members of vulnerable groups are recruited to work in transport agencies in every area (from design and maintenance of transportation nodes and carriages, to customer services and bus drivers), they have a chance to reflect upon their needs and use them to transform environments and services to make everyone safer.

Biography



Vania Ceccato is a Professor at the Department of Urban Planning and Environment, School of Architecture and the Built Environment, KTH Royal Institute of Technology, Stockholm, Sweden. She coordinates the national network Safeplaces (Säkraplatser nätverk) funded by The Swedish National Crime Prevention Council (BRÅ). Ceccato is interested in the relationship between the built environment and crime and perceived safety, in particular the space–time dynamics of crime and people's routine activity. Gendered safety and the intersectionality of victimization are essential components in her research and teaching. She is the author of numerous articles and books in the area of the situational conditions of crime and fear in urban and rural environments.

Relevant Website

Can architecture and planning ensure safety for women? TEDxKTHWomen. Available from: <https://youtu.be/L8oEZ16vFzk>.

References

- Ball, K.S., Wesson, C.J., 2017. Perceptions of unwanted sexual behaviour on public transport: exploring transport density and behaviour severity. *Crim. Prev. Community Safety* 19, 199–210.
- BRÅ, 2019. The Swedish National Council for Crime Prevention (Swedish: Brottsförebyggande rådet). The Swedish national Crime Victims Survey 2018. BRÅ, Stockholm. Available from: <https://www.bra.se/statistik/statistik-utifran-brottstyper/valdtakt-och-sexualbrott.html>.
- Ceccato, V., 2013. Moving Safely: Crime and Perceived Safety in Stockholm's Subway Stations. Lexington, Plymouth.
- Ceccato, V., 2014a. The nature of rape places. *J. Environ. Psychol.* 40, 97–107.
- Ceccato, V., 2014b. Safety on the move: crime and perceived safety in transit environments. *Secur. J.* 27, 127–131.
- Ceccato, V., 2016. Public space and the situational conditions of crime and fear. *Int. Crim. Justice Rev.* 26, 69–79.
- Ceccato, V., 2017a. Women's transit safety: making connections and defining future directions in research and practice. *Crim. Prev. Community Safety* 19, 276–287.
- Ceccato, V., 2017b. Women's victimisation and safety in transit environments. *Crim. Prev. Community Safety* 19, 163–167.
- Ceccato, V., Loukaitou-Sideris, A., 2019. *Transit Crime and Sexual Violence in Cities: International Evidence and Prevention*. Routledge, London.
- Ceccato, V., Newton, A. (Eds.), 2015. *Safety and Security in Transit Environments: An Interdisciplinary Approach*. Palgrave, New York.
- Ceccato, V., Paz, Y., 2017. Crime in São Paulo's metro system: sexual crimes against women. *Crim. Prev. Community Safety* 19, 211–226.
- Ceccato, V., Uittenbogaard, A.C., Bamzar, R., 2013. Safety in Stockholm's underground stations: the importance of environmental attributes and context. *Secur. J.* 26, 33–59.
- Ceccato, V., Wiebe, D.J., Eshraghi, B., Vrotsou, K., 2017. Women's mobility and the situational conditions of rape: cases reported to hospitals. *J. Interpers. Violence*. doi:10.1177/0886260517699950.
- Ceccato, V., Li, G., Haining, R., 2018. The ecology of outdoor rape: the case of Stockholm, Sweden. *Eur. J. Criminol.* 16 (2), 210–236.
- Cohan, A., Shakeshaft, C., 1995. Sexual abuse of students by school personnel. *Phi Delta Kappan* 76, 513–520.
- d'Arbois de Jubainville, H., Vanier, C., 2017. Women's avoidance behaviours in public transport in the Ile-de-France region. *Crim. Prev. Community Safety* 19, 183–198.
- Doan, P.L., 2007. Queers in the American City: transgendered perceptions of urban space. *Gend. Place Cult.* 14, 57–74.
- Dymén, C., Ceccato, V., 2012. An international perspective of the gender dimension in planning for urban safety. *The Urban Fabric of Crime and Fear*. Springer, The Netherlands.
- Gekoski, A., Jacqueline, M., Gray, M.H., Horvath, S.E., Aliye, E., Adler, J., 2015. 'What works' in reducing sexual harassment and sexual offences on public transport nationally and internationally: a rapid evidence assessment. Project Report. Middlesex University, British Transport Police and Department for Transport, London, UK.
- Goodyear, S., 2015. More women ride mass transit than men. Shouldn't Transit Agencies Be Catering to Them? CityLab. Available from: <https://www.citylab.com/transportation/2015/01/more-women-ride-mass-transit-than-men-shouldnt-transit-agencies-be-catering-to-them/385012/>.
- Gray, F.V., 2018. *The Right Amount of Panic: How Women Trade Freedom for Safety*. Policy Press, Bristol.
- Hsu, H.-P., Boarnet, M.G., Houston, D., 2019. Gender and rail transit use: influence of environmental beliefs and safety concerns. *Transp. Res. Rec.* 2673, 327–338.
- Iudici, A., Bertoli, L., Faccio, E., 2017. The 'invisible' needs of women with disabilities in transportation systems. *Crim. Prev. Community Safety* 19, 264–275.
- Junger, M., 1987. Women's experiences of sexual harassment: some implications for their fear of crime. *Br. J. Criminol.* 27, 358–383.
- Kash, G., 2019. Always on the defensive: the effects of transit sexual assault on travel behavior and experience in Colombia and Bolivia. *J. Transp. Health* 13, 234–246.
- Kunieda, M., Gauthier, A., 2007. Gender and urban transport: fashionable and affordable. In: *Sustainable Transport: A Sourcebook for Police Makers in Developing Cities*, GTZ, Eschborn.
- Lea, S.G., D'Silva, E., Asok, A., 2017. Women's strategies addressing sexual harassment and assault on public buses: an analysis of crowdsourced data. *Crim. Prev. Community Safety* 19, 227–239.
- Levine, N., Wachs, M., 1986. Bus crime in Los Angeles: II—victims and public impact. *Trans. Res.* 20 (4), 285–293.

- Loukaitou-Sideris, A., 2004. Is it safe to walk here? Research on women's issues in transportation. Conference Proceedings. Transportation Research Board, Washington, DC.
- Loukaitou-Sideris, A., 2012. Safe on the move: the importance of the built environment. *The Urban Fabric of Crime and Fear*. Springer, The Netherlands.
- Loukaitou-Sideris, A., 2014. Fear and safety in transit environments from the women's perspective. *Secur. J.* 27, 242–256.
- Loukaitou-Sideris, A., 2016. A gendered view of mobility and transport: next steps and future directions. *Town Plan. Rev.* 87, 547–565.
- Lubitow, A., Carathers, J., Kelly, M., Abelson, M., 2017. Transmobilities: mobility, harassment, and violence experienced by transgender and gender nonconforming public transit riders in Portland, Oregon. *Gen. Place Cult.* 24, 1398–1418.
- Lundkvist, H., 1998. *Ojämsställdhetens miljöer*. Svenska kommunförbundet och Kommentus förlag, Stockholm.
- Madan, M., Nalla, M., 2015. Sexual harassment in public spaces: examining gender differences in perceived seriousness and victimization. *Int. Crim. Justice Rev.* 26, 80–97.
- Mellgren, C., Andersson, M., Ivert, A.-K., 2018. It happens all the time: women's experiences and normalization of sexual harassment in public space. *Women Crim. Justice* 28, 262–281.
- Moreira, G., Ceccato, V., 2019. Gendered mobility and violence in the São Paulo metro, Brazil. *Urban Studies* (submitted).
- Morgan, R., Smith, M.J., 2006. Crimes against passengers: theft, robbery, assault and indecent assault. In: Smith, M.J., Cornish, D.B.C. (Eds.), *Secure and Tranquil Travel: Preventing Crime and Disorder on Public Transport*. Jill Dando Institute of Crime Science, London.
- Natarajan, M., Schmulh, M., Sudula, S., Mandala, M., 2017. Sexual victimization of college students in public transport environments: a whole journey approach. *Crim. Prev. Community Safety* 19, 168–182.
- Newton, A., 2014. Crime on public transport. In: *Encyclopedia of Criminology and Criminal Justice*. Springer, London.
- Newton, A., 2018. Crime at the intersection of rail and retail. In: Ceccato, V., Armitage, R. (Eds.), *Retail Crime: International Evidence and Prevention*. Palgrave, Cham.
- Rivera, R.L.K.C., 2007. Culture, gender, transport: Contentious planning issues. *Transp. Commun. Bull.* 76.
- Sochor, J., 2015. Enhancing Mobility and Perceived Safety via ICT: The Case of a Navigation System for Visually Impaired Users. In: V. Ceccato and A. Newton (Eds.), *Safety and Security in Transit Environments: An Interdisciplinary Approach*. Palgrave Macmillan, London, UK, pp. 344–361.
- Solymsi, R., Cella, K., Newton, A., 2018. Did they report it to stop it? A realist evaluation of the effect of an advertising campaign on victims' willingness to report unwanted sexual behaviour. *Secur. J.* 31, 570–590.
- Sommers, M.S., Zink, T., Baker, R.B., Fargo, J.D., Porter, J., Weybright, D., Schafer, J.C., 2006. The effects of age and ethnicity on physical injury from rape. *J. Obs. Gynecol. Neonatal Nursing* 35 (2), 199–207, doi:<https://doi.org/10.1111/j.1552-6909.2006.00026.x>.
- Transport for London (TfL), (2013). Safety and security annual report. Transport for London, London.
- Transport for London (TfL), (2013). Safety and security annual report. Transport for London, London.
- Tripathi, K., Borrión, H., Belur, J., 2017. Sexual harassment of students on public transport: an exploratory study in Lucknow, India. *Crim. Prev. Community Safety* 19, 240–250.
- UN, 2019. The sustainable development goals. New York. Available from: <https://www.un.org/sustainabledevelopment/sustainable-development-goals/>.
- UOL, 2014. Casal homossexual é espancado por 15 homens no metrô de São Paulo. UOL, São Paulo. Available from: <http://noticias.uol.com.br/cotidiano/ultimas-noticias/2014/11/12/casal-homossexual-e-espancado-por-15-homens-no-metro-de-sao-paulo.htm?cmpid=copiaecola>.
- Whitzman, C., 2007. Stuck at the front door: gender, fear of crime and the challenge of creating safer space. *Environ. Plan. A* 39, 2715–2732.

Further Reading

- Kelly, L., Radford, J., 1990. 'Nothing really happened': the invalidation of women's experiences of sexual violence. *Crit. Soc. Policy* 10, 39–53.
- Natarajan, M., 2016. Rapid assessment of "eve teasing" (sexual harassment) of young women during the commute to college in India. *Crim. Sci.* 5, 1–11.
- Whitzman, C., 2009. *The Handbook of Community Safety Gender and Violence Prevention: Practical Planning Tools*. EarthScan/Routledge, New York.

Shared Space

Allison B. Duncan, Coventry University, Coventry, Warwickshire, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	584
What is Shared Space?	584
History of Shared Space	585
Conceptual Principles of Shared Space	586
Urban Design	586
Traffic Calming	586
Risk	587
Elements of Shared Space	587
Benefits of Shared Space	588
Examples of Shared Space	588
Poynton, United Kingdom	588
Ashford, United Kingdom	588
Coventry, United Kingdom	589
Road Safety and Crashes	589
Shared Space Debates	590
Discussion and Conclusions	591
Relevant Websites	591
References	591
Further Reading	592

Introduction

Communities are increasingly grappling with the impact motor vehicles have on road safety and quality of life due to concerns surrounding speeding, congestion, and air pollution. This can be seen in the growing number of places enacting congestion charges, lower emission zones, and so on. There is also increasing interest in redesigning cities to minimize driving and keep vulnerable road users safer. Fatality rates for motor vehicle occupants has been decreasing for many years but the fatality rates of vulnerable users such as bicycle riders and pedestrians continues to increase throughout most of the world. There are many possible reasons for this including distracted driving, lax speed enforcement, and a growing number of SUVs on the road with their higher profiles. Other reasons for creating shared space areas include climate change and air pollution, the obesity epidemic, increasing traffic congestion, and vehicular injuries to pedestrians and bicycle riders. The desire for safer streets has spawned multiple movements including Livable Streets, Complete Streets, Safe Routes to School, and Shared Space (also known as Naked Streets or Curbless Streets).

This chapter discusses shared space as a concept and design solution. The chapter defines the concept and its foundations and explores the history of the movement. Then the chapter will delve into the features of shared space designs and provide a few examples of built projects. Because shared space is a controversial technique, this chapter will also touch on some of the current debates.

What is Shared Space?

The overarching idea behind the shared space concept is that the streets are for people and that no single transportation mode should dominate public space and/or roadways. This technique is used to modify road straightaway and intersections into more accessible spaces for all people including pedestrians and cyclists and users of all ages and abilities. Therefore, shared space can be best considered both a form of urban design as well as a form of traffic calming.

The primary goal of a shared space project is the reduction of motor vehicle speeds in order to improve the safety of pedestrians, bicyclists, and motorists without unduly interfering with traffic flow. The language of design and uncertainty are used to communicate to all users that no one mode has priority over any other. The appeal of shared space projects has become increasingly popular worldwide over the last several years. For instance, the United Kingdom has at least 20-shared space projects on the ground as of 2018. There are also several in the United States in cities ranging from Philadelphia to Salt Lake City as well as in Switzerland and Sweden. **Fig. 1** shows an example of a shared space project in Sweden.

Because of the design modifications, which will be discussed further, a shared space project may involve a shift in management and communication within a city for issues ranging from emergency vehicle access to street cleaning. For instance, in Portland,



Figure 1 Skvallertorget shared space project in Sweden. *Source: Photo courtesy of Norrköpings kommun.*

Oregon, one stumbling block for a proposed curbless road downtown was confusion over which agency was responsible for street cleaning once the street and sidewalk merged and became one surface.

There is not a large body of research on shared space. This may be for multiple reasons including the relative scarcity of actual projects on the ground. These projects are not monolithic and can be created on a degree of gradation from a “pure” shared space to one with fewer elements. There are many projects that are inspired by shared space and its concepts but far fewer of pure shared space projects themselves.

History of Shared Space

While similar to historic road design before motor vehicles were introduced in the early 1900s, shared space is said to have begun with Hans Monderman, a road safety investigator in the Netherlands. The popular history says that Hans Monderman was facing budget cuts while simultaneously needing to calm vehicular traffic in multiple high-crash areas. Given the lack of budget, he decided instead to remove street signs. The resulting slower speeds and lower crash rates convinced him to pursue this technique and develop it further. He worked on more than 100 shared space projects in the European Union before his death in 2008.

Hans Monderman found his initial inspiration in the woonerven in the Netherlands. A woonerf, also called a Living Street, a Neighborhood Unit, a Play Street, or a Home Zone is a small-scale individual street or neighborhood traffic-calming project which slows motor vehicle traffic down and enables residents to enjoy their community. A woonerf creates a safer “island” in the urban environment allowing children to play and neighbors to socialize. As opposed to share-space streets, woonerven are typically low-volume residential streets and different modes do not have the same priority, pedestrians are favored. For example, if a child is playing with his/her toy car in the middle of the street, a motorist will have to wait or get out of their car and negotiate with the child before they legally can continue (Article 88a RVV of 1976), and then proceed only at walking speed—though that is the speed with which a horse “walks” (Article 88b RVV). But woonerven have a lot in common with shared space, for example, both lack curbs and use physical traffic-calming tools.

The changes wrought in the standard road environment make it unusual and requires more attention on the parts of all users. The ambiguity requires all modes to actively negotiate with each other when navigating the roadway. The change in the more familiar road design accompanied by the potential for users to use the space in unpredictable manners, encourages users to pay more attention and communicate with other road users.

Conceptual Principles of Shared Space

Because shared space incorporates urban design and traffic calming, the following sections will define these two subjects as the foundational material for understanding shared space.

Urban Design

Urban design is not easily defined. Its definitions are wide-ranging and subject to debate in some circles. They can range from historic definitions looking at design precedents to qualitative definitions focusing on goals and objectives. The design literature lists multiple elements authors consider essential to good urban design and these range from large scale and general, such as a building or parcel, to smaller scale including elements like benches or street trees. Urban design elements also may be classified with respect to movement (roads, sidewalk) or barriers (walls, edges). For example, the American urban planner Kevin Lynch, in his most famous work, the *Image of the City* (1960), discusses how people take in information of streets, sidewalks, and other channels in which people travel by observing edges (perceived boundaries such as walls, buildings, and shorelines); districts (relatively large sections of the city distinguished by some identity or character); nodes (focal points, intersections or loci); and landmarks (readily identifiable objects which serve as external reference points).

In this chapter, urban design is defined as the creation of spaces and opportunities in the public realm by the function, sensitive, and flexible apportionment of space and structure with regard to users, environment, and infrastructure while encouraging the use and interaction of the public (Duncan, 2016). The appropriate scale for shared space is site-specific and primarily focused on human-scaled elements such as benches, street trees, and paving. It must be noted, though, that the larger scale aspect of urban design plays a role in the siting of shared space projects themselves.

Traffic Calming

The human body is not designed for impact at high speeds, and the death rate of vulnerable users goes up significantly with vehicle speed. While the literature varies on the specific vehicle speeds involved, in general, a pedestrian's chance of surviving a vehicular impact drops with increasing vehicle speed. Depending on multiple variables (Jurewicz et al., 2016; Kröyer, 2015), a pedestrian's chances of surviving may range from an approximately 90% survival rate at 30 km/h (20 mph) vehicle speed to a less than 50% survival rate at 50 km/h (31 mph). Pedestrians hit by motor vehicles moving at 40 mph have approximately an 85% chance of dying. Survival rates also vary depending on type of crash (head-on versus side impact) (Jurewicz et al., 2016), vehicle type (SUVs and large trucks with higher profiles), and age of person hit (Kröyer, 2015). The very young and old will have higher fatality rates at lower speeds with people 75 and older having the lowest survival rates. It is essential to recognize that the fatality rate for vulnerable users increases by almost 3% as vehicle speed increases by 1 mph. There is a clear link between survivability and speed with faster vehicle speeds being more deadly.

Traffic calming slows road users down and can be physical or psychological (Elvik, 2001; Kennedy et al., 2005). Effective traffic calming is the application of various calming elements used together to slow road users down. This works best when the elements are used in an integrated system rather than in isolation. Speed humps and bumps are the elements most commonly associated with standard traffic calming by slowing drivers with a physical element in the road. Other forms of physical traffic calming include bulb outs, chicanes, and traffic circles, which all deviate the path of the road users. For more detail, see article Speed-Reducing Measures. These physical forms often are used in shared space designs in conjunction with the entry monuments (Fig. 2).



Figure 2 Example of entry monument. Source: Photo by author.

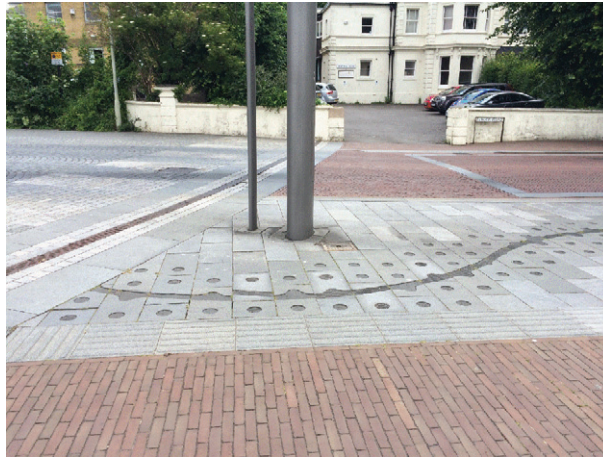


Figure 3 Example of textured and paving patterns. *Source: Photo by author.*

Psychological traffic calming includes techniques such as visually narrowing a roadway by introduction of elements such as trees or street furniture as well as painting the road (see The City Repair Project) or paving it with alternative materials (Fig. 3). It also includes designing ambiguity into the space removing road elements as well. Psychological treatments work best when used in conjunction with physical calming methods. It is interesting to note that sometimes lack of paint can also make the road seem narrower. If you drive down a street with a centerline and it consists of two 3.25 m lanes, the lane you are driving in will look slightly wider than the oncoming lane because of the perspective you see it from. And, the 3.25 m lane will seem wide enough to allow fairly high speeds. If we remove the centerline, the road will seem slightly narrower than 6.5 m, and you are likely to slow down when there is oncoming traffic.

Risk

This is perhaps the most controversial aspect of shared space. Engineers are trained to design risk out of roads by designing for passive safety but shared space projects intentionally bring perceived risk back. Some research literature finds that the lack of risk in the environment may make people engage in riskier behavior (risk homeostasis or at least risk compensation) (Wilde, 1998), and making roads feel less safe will actually encourage less risky behavior. This concept is not without its critics but there are multiple studies, which support it. Though, optimal may be to make it seem unsafe by using certain design elements but actually have redundant safety in the system if possible.

An example of a controversial design element introducing risk can be found in the original woonerven designed in Delft. These included new elements—such as planters and boxes—placed so that a child could hide behind them, and then suddenly jump out in front of a motor vehicle. The intent was that drivers would go “dead slow” rather than drive at walking speed at those places. That can be considered part of the “unpredictable” design principle mentioned earlier. When woonerf principles were adopted for use in other countries in the late 1970s, such as opholds- og legegader in Denmark, these sight obstructive elements were eliminated. The opholds- og legegader also allowed slightly higher speeds than woonerven, up to 15 km/h (9 mph). And, Denmark at that time also introduced shared-space streets called Stilleveje for use on busier streets with allowed speed of 30 km/h (19 mph). Both these street types seemed revolutionary for Danish cities where the default speed limit—applied to most urban streets—was kept at 60 km/h (37 mph) until April 1985, longer than in other Scandinavian countries, which had been using 50 km/h as the default speed limit since much earlier.

Elements of Shared Space

As discussed, shared space encompasses both urban design and traffic calming. There is overlap between them, and a successful project integrates both. There is a general catalog of “expected” elements associated with shared space; a complete, or “pure,” shared space project typically will have the following elements in the design. First, the curbs will be removed from the roadway blurring the segregation between vehicle drivers and people walking. All lane markings, including the centerline, will be removed. All street signs will be removed as well. Removing the curbs flattens the road and adjoining sidewalks to give the project the impression of a plaza or public square. The paved surface of a shared space design will have the paving typically go from building face to building face with no distinction made between road and pedestrian space. The resultant plaza will be paved in geometric designs with textured pavers. These designs will help direct road users while the textured pavers help less able users navigate the space as well as further indicate a change from standard road space to shared space. The shared space also will include street furniture such as street trees, seat walls, benches, and light fixtures to humanize the space and create areas where people may sit and socialize. Finally, a “pure” shared space

will have entry monuments. These serve to alert road users that they are entering a space that differs from the standard roadway and they need to stay alert. All of the above contribute to creating spatial ambiguity for drivers as well as making it feel more welcoming to nonmotorized users.

Benefits of Shared Space

The concept of shared space lists multiple potential benefits to these projects. These include:

Place-making and improved sense of community: By giving the roadways back to the people through calming and design, the community may be more willing to gather and use the space for more than just travel. By creating a place where people want to be, where they can have pride of place, place-making can enhance a community. (Karndacharuk et al., 2014)

Quality of life: Beyond improving the quality of life through improved community, shared space projects can improve quality of life through safer roads, lower vehicle speeds, reduced congestion, and improved accessibility options for some user groups.

Economic growth: Shared space projects have been linked to economic revitalization of the areas around the projects. Research has found increased property values and improved economic development opportunities.

Safety: As discussed previously, the decrease of vehicle speeds contributes to safer streets for all users. There may be a decrease in crime as well due to what Jane Jacobs called “ . . . eyes on the street.”

Traffic flow: Shared space has been shown to improve throughput in multiple projects such as Poynton and Coventry (discussed below). Despite slowing traffic down, the delay is decreased due to traffic not being stop and go.

Examples of Shared Space

Initially, in the Netherlands, shared spaces were only installed in quiet, low traffic areas. Over time, it was found that these designs could work in areas with higher motor vehicle traffic volumes than initially thought. Shared space, however, is not appropriate for all roads. Instead it is best implemented in areas in need of traffic calming, especially in sites with heavy (or desired increased) nonmotorized traffic. This works best in commercial and residential areas.

It can be difficult to accurately quantify the number of shared space projects as many people count all pedestrianized spaces toward this number of shared space projects. For instance, the United Kingdom has many shared space projects; however, the reported number of projects on the ground ranges from 20 to more than 100. Examination of some of the supposed shared space sites shows that some road projects classified as shared space may be wrongly classified as shared space. For instance, some were included due to being a pedestrianized zone. (It should be noted, shared space differs from shared-use paths. Shared-use paths are a form of nonmotorized infrastructure typically shared by pedestrians and bicycle riders. They remain physically separated from motor vehicle traffic by either a barrier or open space.)

As it has been discussed, a shared space design can take many forms. It will be site specific. The following section will discuss a few projects to illustrate the range of possible designs.

Poynton, United Kingdom

Poynton’s shared space project is one of the United Kingdom’s showcase projects. This project has received significant publicity due to both its innovative design as well as a good publicity scheme.

The village of Poynton’s central intersection had serious congestion issues with a significant amount of commercial truck traffic. Poynton chose a shared space retrofit because the road running through the center of town, north to south was dominating the village and decreasing quality of life. The residents were subjected to long waits while trying to cross the main road in their own village, pollution, noise, etc. The British urban designer, Ben Hamilton-Baillie, removed the entire existing road infrastructure, including traffic signals, and converted the large intersection to shared space. This complex intersection was designed with intricate and textured paving patterns forming a double roundabout.

Ashford, United Kingdom

The Elwick Square plaza is a good example of a large shared space intersection. It is unusual in the United Kingdom for its street furniture as well as its size. Elwick Square has many of the standard shared space elements. There is textured paving in geometric patterns and a seating area with vegetation within the plaza as well as a seat wall at a spot in the site where bicycle riders often stop to wait to cross (Fig. 4).

Note that the marked crosswalk (Fig. 5), also present in other shared space projects, is not considered a “pure” shared space element. However, its inclusion represents the flexibility of the shared space concept and the needed responsiveness of listening to the public when designing these spaces. For further understanding of this project, Moody and Melia (2013) researched pedestrian movements through this intersection, and Duncan (2016) analyzed bicyclist behavior.



Figure 4 The Elwick Square project illustrating the use of street furniture, landscaping, and textured paving. *Source: Photo by author.*



Figure 5 A portion of Elwick Square shared space intersection with marked zebra crosswalk. *Source: Photo by author.*

Coventry, United Kingdom

The Coventry University shared space intersection is not considered a “pure” version. The modifications on the scheme in this intersection include multiple large, stone bollards to protect the corners from vehicle drivers. The space also utilizes colored paving instead of textured paving—an affordable and lower maintenance option. There are three zebra crossings near this intersection. One can be seen with its yellow lights farther back in [Fig. 6](#). However, it can be seen that the space is level and lane markings and road signs have been removed. The space is still ambiguous and receptive to all modes. This is an illustration of the range-shared space projects can take.

It is also more common to put shared space projects in on straightaway instead of intersections. It is a simpler/less complicated task but leaves the intersection unmodified so that it remains the most dangerous section for vulnerable road users.

Road Safety and Crashes

Comparing crash rates of shared space projects can be difficult. Many projects are newer and data are often not available or not enough time has passed since project installation to gather crash data on the new design. Another potentially confounding issue is that minor crashes may not always be reported to the appropriate reporting agencies.

In 2017 a serious car crash occurred on Exhibition Road, a shared space project in London. A driver injured 11 people after driving into a crowd along Exhibition Road leading to increased calls by the media to eliminate shared spaces in the United Kingdom. Already a contentious subject in the United Kingdom, the Chartered Institute of Highways and Transport (CHIT) responded and performed a study on the crash rates of existing UK shared space projects. The Institute found crash rates either



Figure 6 A Coventry University shared space project on campus illustrating the use of colored asphalt and spherical bollards. *Source: Photo by author.*



Figure 7 Skvallertorget before conversion to shared space. *Source: Photo courtesy of Norrköpings kommun.*

decreased or stayed unchanged after project completion. Other studies have found similar results, or that in some cases the crash rate increased but the crash severity decreased. In other words, there were fewer serious crashes and more minor collisions. However, crashes are rare and most of the projects on the ground have not shown any changes in crash rates (Karndacharuk et al., 2014). The crash rates that have been measured have shown that the crashes that do happen are more minor due to lower speeds.

Another example of changing crash rates is the Skvallertorget space in Norrköping, Sweden. A higher traffic intersection with approximately 13,000 vehicles/day (Fig. 7), the shared space site is at an important intersection in central Norrköping.

The intersection was changed from a standard intersection to a shared space designing in uncertainty as well as incorporating elements such as textured and patterned paving. See Fig. 1 for Skvallertorget after conversion to shared space. The new plaza showed a decrease in the number of crashes and vehicle speeds after project completion.

The primary goal of shared space is to slow vehicle traffic down and make spaces more livable. Livable in this case translates into a safer, more welcoming environment for nonmotorized road users. The research on shared space has shown that the ambiguity of these spaces has the effect of slowing drivers down. This also leads to safer roads for all users. An added benefit of slower speeds is improved throughput, too which actually leads to less delay to motorists.

Shared Space Debates

Concerns regarding these spaces are frequently expressed by elderly, disabled, and/or visually impaired road users. These user groups may feel less protected while crossing these spaces than in more standard road layouts. One ongoing debate in the United Kingdom concerns the potential difficulties and concerns shared space causes people with visual impairments (Parkin and Smithies, 2012). There have been research articles [focus groups] looking at the impact these road designs have on this user group. One concern from this group is the confusion it may cause guide dogs (Norgate, 2012). This has led to a recent decision in the United Kingdom to put a pause on shared space projects and build no more of them pending further investigation.

There is also an increased interest in more segregation of modes on roads. This can be seen in the 8–80 movement where projects are designed to be safe and comfortable for users of all ages. The increasing number of cycle tracks, protected bike lanes, and protected intersections are indicative of this. Some argue that this means that shared space projects are no longer appropriate. However, segregation of modes is not always the best use of a site due to limitations such as available space or a project's goals and objectives.

The use of design to influence behavior relies on the belief that all modes will respond to the road treatment similarly. Research has shown, however, that shared spaces may not always perform as designed. In the United Kingdom, for instance, a large percentage of nonmotorized users, pedestrian, and cyclists, do not cross freely, communicating with drivers. Instead, many of these users skirt the edges minimizing their interactions with motor vehicles. Culture, education, and time are all factors in how the road is navigated by different modes. In some cases, nonmotorized users may adapt and grow more comfortable with the space and use it more as desired.

Concerns about these projects may vary from country to country. In the United States, designers and DOTs are reluctant to put in any project that varies from transportation guides like the AASHTO. The fear is of litigation. In the United Kingdom, the strongest anti-shared space group is the visually impaired group.

Discussion and Conclusions

As there is growing recognition of the need to make transportation safer and more welcoming to nonmotorized users, there will be more ideas on how to design and build safer streets. Shared space is one of those options—like mode separation, is one of many possible tools in a community's toolbox.

This technique's history and its continuing controversies show how important it is to include the public in the planning and design phases. Listening and addressing concerns raised can help the community feel more comfortable about the new space and mitigate future issues. Some elements, which may help concerned users, include more tactile paving, street furniture, and inclusion of refuge areas, if possible. An example of this is the Ashford project where a pocket of trees, benches, and small wall are placed along the path where vulnerable road users most typically cross.

Current evidence indicates these spaces are safer despite feeling riskier to road users. However, it is still not clear that shared space and its riskier feeling provides a more welcoming environment for nonmotorized users—especially for the visually impaired road user. The questions will remain—how can these more democratic spaces be welcoming to all users? Further research, including analysis of existing projects and modeling of new ones (VR and temporary) will be needed to improve the concept for more countries.

Designers and engineers have to design for the space and circumstances they have. Each situation is unique. Because shared space is both design and engineering, there are no set answers to a transportation problem. Instead, designers need to approach the problem holistically by looking at the issues, the users, the goals of the municipality, the adjacent land uses, and so on. The elements of shared space alone do not make a successful, or unsuccessful project—the integration of the elements best suited for the roadway and the community are essential to the success of the design.

The concept of shared space is a controversial idea to many planning agencies and departments of transportation. The idea of making something less segregated and seemingly less safe is uncomfortable for public officials and most road users. Public agencies often water down a shared space design due to concerns about safety by installing bollards, paint in bicycle lanes, and so on. What matters is that the design is site sensitive and addresses the concerns of the community and road users of all modes. The combination of urban design and traffic calming can allow for creative options to address road sections in need to improvement.

Relevant Websites

Cassini, M., 2013. Poynton regenerated. [online] Available from: <https://www.youtube.com/watch?v=vzDDMzq7d0>.

References

- Duncan, A., Cyclist Path Choices Through Shared Space Intersections in England. Google Scholar. Portland State University, Portland, 2016. Available from: https://pdxscholar.library.pdx.edu/open_access_etds/2704/.
- Elvik, R., 2001. Area-wide urban traffic calming schemes: a meta-analysis of safety effects. *Accid. Anal. Prev.* 33 (3), 327–336.
- Jurewicz, C., Sobhani, A., Woolley, J., Dutschke, J., Corben, B., 2016. Exploration of vehicle impact speed–injury severity relationships for application in safer road design. *Transp. Res. Procedia* 14, 4247–4256, doi:10.1016/j.trpro.2016.05.396.
- Karndacharuk, A., Wilson, D.J., Tse, M., 2014. Shared space performance evaluation: quantitative analysis of pre-implementation data. IPENZ Transportation Group Conference, 2011, Auckland.
- Kennedy, J., Gorell, R., Crinson, L., Wheeler, A., Elliott, M., 2005. Psychological traffic calming in TRL Report TRL641. Available from: <https://trl.co.uk/sites/default/files/TRL641.pdf>.
- Kröyer, H.R.G., 2015. The relation between speed environment, age and injury outcome for bicyclists struck by a motorized vehicle—A comparison with pedestrians. *Accid. Anal. Prev.* 76, 57–63, doi:10.1016/j.aap.2014.12.023.
- Moody, S., Melia, S., 2013. Shared space: research, policy, and problems. *Proc. Inst. Civil Eng.* 167 (6), 384–392.

- Norgate, S.H., 2012. Accessibility of urban spaces for visually impaired pedestrians. *Proc. Inst. Civil Eng.* 165 (4), 231–237, doi:10.1680/muen.12.
- Parkin, J., Smithies, N., 2012. Accounting for the needs of blind and visually impaired people in public realm design. *J. Urban Des.* 17 (1), 135–149, doi:10.1080/13574809.123456789.
- Wilde, G.J.S., 1998. Risk homeostasis theory: an overview. *Inj. Prev.* 4 (2), 89–91, doi:10.1136/ip.4.2.89.

Further Reading

- Duncan, A., 2017. A comparative analysis of cyclists' paths through shared space and non-shared intersections in Coventry, England. *J. Urban Des.* 22 (6), 833–844, doi:10.1080/13574809.2017.1336062.
- Biddulph, M., 2012. Street design and street use: comparing traffic calmed and home zone streets. *J. Urban Des.* 17 (2), 213–232, doi:10.1080/13574809.2012.666206.
- Edquist, J., Corben, B., 2012. Potential application of shared space principles in urban road design: effects on safety and amenity, NRMA-ACT Road Safety Trust, Australia.
- Hamilton-Baillie, B., 2008. Shared spaces: reconciling people, places and traffic. *Built Environ.* 34 (2), 161–181.
- Jones, P., Oakely, C., Higgs, G., Kirby, T., 2018. Creating better streets: Inclusive and accessible places: Reviewing shared space. Available from: https://www.ciht.org.uk/media/4463/ciht_shared_streets_a4_v6_all_combined_1.pdf.
- Kaparias, I., Bell, M.G., Singh, A., Dong, W., Sastrawinata, A., Wang, X., Mount, B., 2013. Analyzing the perceptions and behaviors of cyclists in street environments with elements of shared space. 45th Annual Conference of the Universities' Transport Study Group (USTG), Oxford, UK.
- Luca, O., Gaman, F., Singureanu, O.-G., 2012. Coping with congestion: shared spaces. *Theor. Empirical Res. Urban Manage.* 8 (3), 53–62.

Side Area Safety and Side Slopes

José María Pardillo-Mayora, Department of Transport Engineering, Urban and Regional Planning, Technical University of Madrid, Madrid, Spain

© 2021 Elsevier Ltd. All rights reserved.

Glossary	593
Introduction	594
Roadside Hazards	594
Obstacles	594
Side Slopes	595
Safety Zones	596
Vehicle Restraint Systems	597
Safety Barriers and Bridge Parapets	597
Barrier Terminals and Transitions	598
Motorcyclists and Cyclists Protection Devices	598
Crash Cushions and Arrester Beds	600
Testing and Approval of Road Restraint Systems	601
Biography	601
Relevant Websites	601
References	601
Further Reading	601

Glossary

abutment End support of a bridge deck or tunnel.

arrester bed Sand- or gravel-filled lane adjacent to a steep downhill grade section intended to allow vehicles that are having braking problems to safely stop.

breakaway support Roadside support designed to yield or break when struck by a vehicle.

bridge parapet Safety barrier mounted on bridges, culverts, or retaining walls.

crash cushion Energy absorbing device that prevents an errant vehicle from impacting fixed objects by gradually decelerating the vehicle to a safe stop or by redirecting the vehicle away from the obstacle.

curb Raised edge at the side of the roadway or outer side of shoulder.

cut slope Earth embankment sloping upward from the roadway.

ditch Drainage channel that runs parallel to the road.

fill slope Earth embankment sloping downwards from the roadway.

median Area that separates opposing traveled ways in a divided highway.

roadside Area beyond the edge line of the traveled way.

roadway Part of a road designed to be used by vehicles.

run-off-road crash Type of crash where an errant vehicle encroaches into the roadside and either overturns or collides with a non-traversable obstacle or with the terrain.

safety barrier Longitudinal vehicle restraint system.

safety zone or clear zone Strip of land adjacent to the traveled way that is cleared of roadside hazards to afford the driver of an errant vehicle the possibility of regaining control or stopping the vehicle.

shoulder Part of the roadway contiguous with the traveled way available for the accommodation of stopped vehicles and for emergency use.

shoulder rumble strip Longitudinal warning device made of a series of indented or raised elements installed on the shoulder near the outside edge of the travel lane to provide an audible and or tactile warning to roadside encroaching vehicles.

sidewalk Area reserved for pedestrian traffic that is separated from the carriageway by a curb.

slope Ground that forms a natural or an artificial incline.

slope rate The relative steepness of a slope expressed as a ratio of its vertical height to its horizontal length.

traveled way Part of the roadway reserved for vehicular travel.

vehicle restraint system Roadside device designed to contain or to redirect an errant vehicle.

Introduction

Roads are designed in principle to accommodate vehicle travel on the traveled way. Nevertheless, vehicles occasionally encroach beyond the edge of the traveled way either intentionally to make emergency maneuvers around objects blocking the roadway or to make emergency stops or unintentionally as a result of the loss of control of the vehicle by the driver. Consequently, the conditions of the area that lies beyond the edge line of the traveled way are important both for road operation and for traffic safety. Properly conditioned and maintained roadsides complement the functions of the roadway and contribute to integrate the infrastructure into its environment. Additionally, roadside conditions have a major effect on traffic safety as a significant part of road traffic fatalities result from uncontrolled road departures.

There are multiple factors that may contribute to the loss of control of a vehicle, including abrupt maneuvers performed to avoid an imminent collision, speeding, distracted driving or drowsiness, driving under the influence of alcohol or drugs, etc. Preventive measures are taken to prevent unsafe driving patterns that increase the risk of vehicle control losses, mainly through the implementation of targeted education and enforcement schemes.

Connected-vehicle and automated-vehicle technologies also have the potential to enhance highway safety by simplifying driving tasks and facilitating that the drivers perceive hazardous situations in time to adopt the adequate decisions and to perform the necessary driving maneuvers while maintaining control of the vehicle.

The risk of drivers losing control of their vehicles can also be reduced by adapting the geometric design of the roads to drivers' expectancies, and by providing ample sight distances at critical road sections, implementing effective road signing, marking and lighting, installing shoulder rumble strips (**Fig. 1**) ([Griffith, 1999](#)), and maintaining pavement regularity and skid resistance at adequate levels. Horizontal curvature and longitudinal grade also have an influence on vehicle encroachment frequencies and on the angle and speed of road departures, which in turn are important factors influencing the severity of roadside crashes.

Early research by Stonex in 1960 concluded that at that time 35% of traffic fatalities in the United States occurred as a consequence of roadside encroachments and identified a number of factors that contribute to increase ROR crash severity, including the presence of non-traversable obstacles in the vicinity of the roadway, steep roadside slopes, deep ditches, and inadequate guardrail terminal configuration.



Figure 1 Shoulder rumble strips.

Roadside Hazards

Once an errant vehicle leaves the roadway and encroaches on the roadside, both the likelihood of it crashing with an obstacle or rolling over and the severity of crashes that cannot be prevented need to be minimized. To this respect, the existence of unprotected non-traversable obstacles on the roadside, their offset from the roadway edge, and the gradient of the side slopes are the main risk factors ([Pardillo et al., 2010](#)).

Obstacles

The existence of rigid non-traversable objects in the vicinity of the road is a major factor of roadside risk. In particular, motorcyclists lack the protection of the vehicle bodywork in the event of a loss of control or of stability and are much more likely to get injured if they hit fixed objects or uneven surfaces.



Figure 2 Traversable roadside ditch.

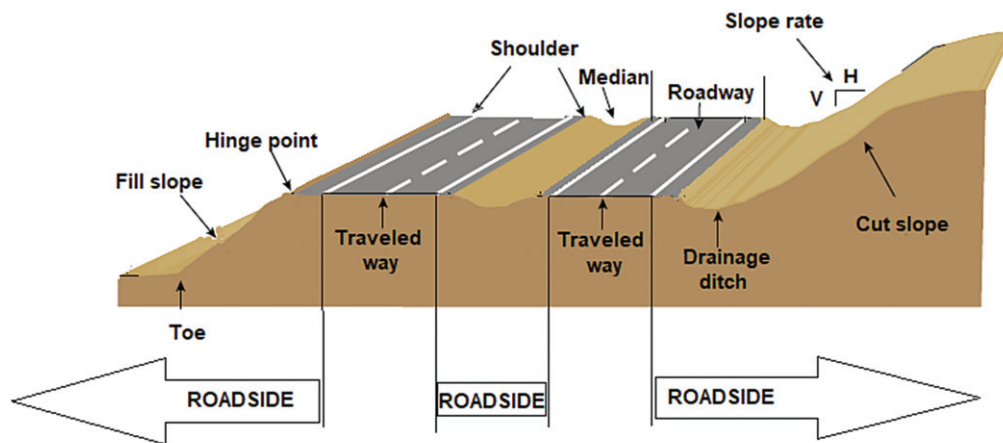


Figure 3 Road and roadside profile.

Roadside obstacles include trees, utility and lighting poles, signposts, rocks and boulders, bridge and overpasses pillars and abutments, culverts and culvert ends, headwalls and drainage pipes, and blunt or ramped guardrail terminals that do not bend toward the roadside. Other obstacles, such as embankments, slopes, ditches, rock face cuttings, retaining walls, and closely spaced trees, extend along a length of the roadside constituting continuous hazards. Vertical falls associated to terrain configuration or to traversing bridges or viaducts pose also a severe roadside risk.

Drainage channels and ditches perform the vital function of collecting and conveying surface water from the roadway. Deep ditches with steep sides can cause errant vehicles to overturn or collide against their back slope especially where rock cuts form the back of the ditch (Viner, 1995). Design guides recommend designs of roadside ditches with fore slope and back slope rates of 1V:6H to prevent rollovers and collisions. These configurations are primarily seen on motorways and other higher-order roads (Fig. 2).

Curbs serve multiple purposes including drainage control, pavement edge delineation, and pedestrian walkways delineation. Vehicles can readily cross-mountable curbs provided with flat sloping faces and with a height that does not exceed 0.1 m. Non-mountable curbs can destabilize vehicles and even cause rollovers when hit.

Side Slopes

Side slopes are constructed to ensure the stability of the roadway. Height, slope rate, and stability of the side slopes, particularly fill slopes, are decisive for rollover prevention and control recovery in the event of uncontrolled roadway departures. Three elements of the slope, the top or hinge point, the fore slope, and the bottom or toe of the slope, need to be considered (Fig. 3).

When an errant vehicle crosses a slope at excessive speed, it may become airborne in crossing its hinge point causing the loss of steering control. Once the vehicle has passed the hinge point and is traversing the fore slope, the driver can attempt a recovery maneuver or reduce speed if the gradient is not too steep and the soil is stable. Nevertheless, in many cases the toe of the slope is reached and consequently a smooth transition at this point is required.

Unobstructed fill slope rates of 1V:6H or lower provide the best conditions for vehicle control recovery, but transverse slopes of 1V:4H or flatter are also classified as recoverable. If such slopes are relatively smooth and traversable, motorists who encroach on them can generally stop their vehicles or slow them enough to return to the roadway safely. Nevertheless, unstable soils may cause

rollovers on side slopes that are otherwise safe. It is also important for control recovery that the hinges and toes of fill and cut slopes be round off (Zegeer et al., 1988).

Embankment slopes between 1:3 and 1:4 may be considered traversable, but non-recoverable, if they are smooth and free of fixed objects. A vehicle will normally not roll over on them, but the driver will not be able to recover the control of the vehicle. The vehicle will therefore end up at the foot of the bank if driving off the carriageway onto such slopes. In addition, back slopes of 1:3 or flatter make it easier for motorized equipment to be used in maintenance operations.

The possibilities of recovering vehicle control with side slope rates over 1:3 are severely compromised. At the same time the probability that the vehicle overturns increases substantially.

Unprotected embankments with gradients steeper than 1:2 pose a high risk of vehicle rollover and significant consequential injury if a vehicle veers off the carriageway. The combination of embankment height and side slope rate also has an effect on the severity of the crashes. Based on studies of the relative severity of encroachments on embankments versus impact with roadside barrier, the installation of a roadside barrier is recommended for embankment heights starting at less than 1 m.

On cut slopes, the slope rate and the shape of the transition between the ditch and the back slope are decisive for how a vehicle will behave when it drives off the road. On rising slopes flatter than 1:2 with a soft transition from the roadway to the slope and no other roadside hazards the driver has good possibilities of slowing down and regaining control of the vehicle. Protruding large stones and rocks located less than 2.0 m above the roadway must be blasted away or treated as hazardous roadside obstacles. Steep cut slopes with rising gradients of more than 1:2 are considered continuous non-traversable obstacles.

Safety Zones

To achieve forgiving roadsides, safety zones, also referred to as clear zones, are established by conditioning a strip of terrain adjacent to the roadway to minimize the likelihood of rollover occurrence and to provide the driver with an opportunity to regain control of his vehicle and come to a gradual stop or return to the carriageway in a controlled manner, without being a hazard to other vehicles (Thompson et al., 2005).

A safety or clear zone may consist of a shoulder, a recoverable slope, a non-recoverable slope, and/or a clear run-out area. Within the safety zone, hazards should be removed or relocated beyond its limits, substituted by passive safe types, or protected with vehicle restraint systems.

Shoulder width is an important safety feature as it contributes to increase the chances of correcting the trajectory of errant vehicles. Shoulder widths on freeways usually range from 2.5 to 3.0 m. For single-carriageway roads there is much more variability. Usual widths are 1.5–2.0 m for main rural roads in flat or rolling terrain, 0.75–1.5 m for mountain roads, and 0.5–1.0 m for local roads. Paved shoulders allow for a better control of an errant vehicle.

Although the risk of collision with an obstacle drops substantially after the first few meters, there still is a possibility that encroaching vehicles may reach farther distances from the roadway edge, and hence safety zones should be as wide as technically and economically feasible. As extending the safety zone implies increasing costs, the recommended safety zone widths are determined on the basis of cost-efficiency analyses depending on the class of road, the expected traffic volumes and operation speeds, the slope gradients, and the road geometry. Recommended clear zone widths range from 2.0 to 3.0 m for low-volume, low-speed roads up to 11.5–14.0 m for high-volume, high-speed roads.

Footpaths and cycle tracks that run along roadways with speed limits above 90 km/h should be aligned outside the safety zone of the roadway. This is also recommended for lower speed roads when pedestrian volumes are high. Otherwise, when footpaths and cycle tracks lie inside the safety zone along roadways with a speed limit of 70 or 80 km/h there ought to be a traffic divide of at least 3.0 m between the traveled way edge and footpath/cycle track, and of at least 1.5 m limited by deflecting curbstones when the speed limit is 60 km/h (Fig. 4). If the traffic divides are narrower, safety barriers should be installed to protect the pedestrians.



Figure 4 Traffic divide between the traveled way and the cycle track.

Within the safety zone, the alternatives for treating the existing hazards in order of priority are as follows:

1. Removing the hazards or moving them beyond the safety zone.
2. Modifying the hazards to render them safe. This can be achieved depending on the cases by flattening the slopes and rounding off its tops and bottoms, blasting road cuttings to obtain an even surface, and using closed ditches, collision-safe noise barriers, and earth mounds or catchment ditches as an alternative to the installation of safety barriers.
3. Replacing hazardous roadside supports with passive safe structures.

Passive safe supports are designed to break or deform in a controlled way when struck by a vehicle thus lessening the impact energy that is transmitted to the vehicle and minimizing the damage both to the vehicle and to its occupants. This can be achieved in the following two ways:

- a. Using poles made of materials that yield and crush at point of impact and absorb energy in the process.
- b. Installing breakaway devices at the base of posts and columns. The bottom flange of the pole is bolted to a frangible base plate that allows the post to slip and break away on impact. The release mechanism may be a slip plane commonly made of heat-treated cast aluminum, plastic hinges, fracture elements, or a combination of these.

Lighting columns, utility poles, and signposts are not considered a hazard if replaced with equivalent passive safe types that have been positively tested in accordance with the applicable roadside hardware standards.

4. Shielding the hazards with vehicle restraint systems.

Vehicle Restraint Systems

Vehicle restraint systems are installed on the roadsides to prevent vehicular penetration and to redirect errant vehicles away from a roadside or median hazard. There are different types of vehicle restraint systems including safety barriers, bridge parapets, crash cushions, and truck arrester beds. Their general purpose is to reduce as much as possible the extent of damage and injuries when uncontrolled vehicles leave the road. In addition, in certain situations there is also a need to protect road users near the road against errant vehicles, to shield critical installations located near the road such as railways, fuel tanks, etc., to prevent uncontrolled vehicles from falling down onto railroad tracks or roads, or into rivers passing under the road (Gardner: Bridge safety), and from damaging road structures that could cause serious consequential damage if impacted.

Hitting vehicle restraint systems may cause damage to the vehicle and its occupants. Therefore, they should be installed only if it is more hazardous to run into the hazard that they shield than to drive into it.

Safety Barriers and Bridge Parapets

Safety barriers are longitudinal restraint systems. They are used to prevent errant vehicles from hitting non-traversable obstacles, overturning on steep side slopes, or falling from a height. Depending on the amount of barrier deflection that takes place when it is struck, safety barriers are classified into three categories: flexible, semirigid, and rigid.

Bridge parapets or railings are safety barriers specially designed to be mounted on the verge of the road on bridges, viaducts, culverts, or retaining walls to prevent vehicles, pedestrians, or cyclists from falling off the edge. Bridge railings are a structural extension of the bridge, differing from other barriers that usually are set in, at, or on soil (Fig. 5).



Figure 5 Bridge parapet.



Figure 6 Semirigid safety barrier.

Flexible systems generally consist of steel wire ropes mounted on weak posts. The flexibility of the system absorbs impact energy and dissipates it laterally, which reduces the forces transmitted to the vehicle occupants.

Semirigid systems consist of galvanized steel rails and steel or hardwood posts (**Fig. 6**). When impacted by a vehicle, the rail deflects and the vehicle is led along the safety fence until it stops, or it is redirected back onto the carriageway by effect of the combined tensile strength of the rail and flexural strength of the posts. Lateral restraint is provided by the posts. The impact energy transmitted to the vehicle and its occupants is greater than for a flexible system but lower than for a rigid system.

Rigid barriers such as concrete barriers do not suffer large permanent deformation on impact. Concrete barriers were developed in the United States of America in the 1960s for use on narrow medians. Over the years, different profiles have been developed, each one having a vertical base section, a flatter lower section and then a taller and steeper upper section. The most widely used nowadays are the New Jersey barrier (**Fig. 7**) and the F-shaped barrier.

Rigid barriers are designed to work at limited impact angles, usually not greater than 20 degrees. An errant vehicle that strikes a rigid barrier at a sufficiently low angle will tend to ride up the lower slope. The lifting of the vehicle and the deformation of the vehicle suspension absorb vehicle energy, slowing it down and pivoting it away from oncoming vehicles and back into traffic heading in its original direction. Kinetic energy is dissipated by the friction of the vehicle side against the almost vertical upper face of the barrier and by the deformation of the bodywork on collision.

The selection of a barrier system for a particular location depends on the required containment level and on the space available for accommodating lateral deflection upon impact without reaching the hazard. When hit at the specified angles and speeds, all safety barriers are capable of redirecting and/or containing an errant vehicle without imposing intolerable deceleration forces on the vehicle occupants. Nevertheless, more rigid systems may cause more severe damages if impacted at sharp angles. In spite of this, they are required when the offset to the object is low and when it is necessary to contain heavy vehicles.

Safety barriers installation requirements define the length of need or total length of barrier needed to shield an area of concern, the lateral offset to the hazard, the barrier height, and the barrier terminal conditions.

Barrier Terminals and Transitions

A safety barrier terminal is a special structure that is installed at the beginning or end of a safety barrier to anchor the leading end without introducing additional hazards for vehicle occupants. They are required to have specified impact performances to ensure that a vehicle that strikes them either drives through the termination and gradually stops or is redirected into the carriageway without significant injury to the driver or the passengers.

Terminals may be flared or tangent. The alignment of flared terminals diverges from the alignment of the roadway edge to redirect vehicles back onto the carriageway, while the alignment of tangent terminals is parallel to the roadway edge. Tangent terminals need to be energy-absorbing devices that have to be tested in accordance.

Safety barrier transitions connect two safety barriers of different geometry and/or containment levels and/or lateral deformation. Common locations for transitions are connections between bridge parapets and roadside safety barriers. Transitions need to be configured to provide a gradual change in the stiffness between the two connected systems thus ensuring a safe performance if a vehicle strikes the transition area.

Motorcyclists and Cyclists Protection Devices

Safety barriers may pose a risk of injury to motorcyclists and cyclists. The most serious incidents that can cause serious injuries, sometimes fatal, occur when they fall and hit a sharp edge or a protruding part or when a direct impact of the motorcyclist against the

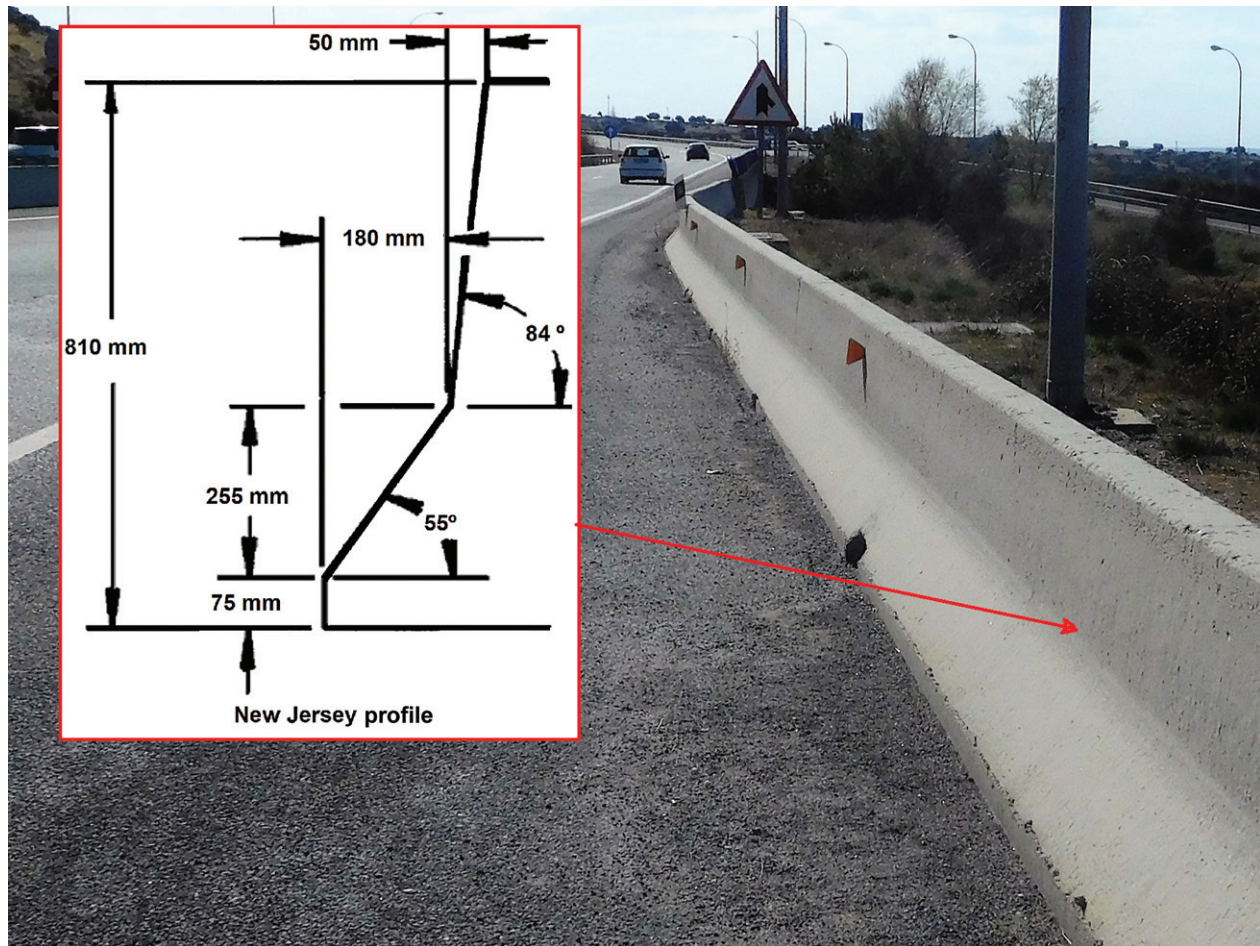


Figure 7 New Jersey concrete barrier.



Figure 8 Safety barrier with secondary rail for motorcyclist protection.

post barrier occurs. In consequence, safety barriers with sharp edges or corners with a radius of less than 0.009 m or protruding parts that can be struck should not be used unless they are effectively protected.

Protection systems designed to reduce the severity a motorcyclists' impact with a safety barrier fall into one of three categories: individual post protector, a secondary rail, or a barrier designed with motorcyclist safety incorporated. Safety barriers with a motorcyclist protection device should be erected in locations where there is a significant risk of motorcyclists rolling over and impacting the barrier such as the outer bends of roads with a significant level of motorcycle traffic (Fig. 8).



Figure 9 Crash cushion.

Crash Cushions and Arrester Beds

Crash cushions are energy absorbing devices used to prevent errant vehicles from head-on impacts with fixed objects at locations where the obstacle cannot be removed, relocated, or made breakaway, and cannot be adequately shielded by a longitudinal barrier.

Crash cushions (sometimes called energy absorbers or impact attenuators) are designed for protecting a point hazard. In contrast with safety barriers, crash cushions are designed for impacts at the end points of their structures, as well as impacts along their sides. They decelerate a vehicle over a short distance in front or side collisions, or which redirects it past a hazard. Commonly used crash cushions absorb the kinetic energy of the impacting vehicle by means of compression of plastically deformable components that are supported by a rigid structure (**Fig. 9**) or by transfer of the momentum of the vehicle to an expandable mass of material, usually sand contained in plastic material barrels.

Depending on their performance for oblique impacts crash cushions are classified into two categories:

- Redirect: Designed for oblique impacts and can behave like a safety barrier for short sections.
- Non-redirect: With no capacity for oblique impacts.

Arrester beds are a special type of vehicle restraining system typically a long, sand- or gravel-filled lane that allows a moving vehicle's kinetic energy to be dissipated gradually and is connected to the main roadway by an escape ramp. These devices are generally located on long steep descending gradients and designed to accommodate large trucks or buses. They are provided to enable vehicles that are having braking problems to decelerate and stop safely without crashing. When terrain conditions allow it, the escape ramp is angled away from the downhill road so that the arrester bed lies on an ascending gradient thus taking advantage of gravity in dissipating the kinetic energy of the vehicle (**Fig. 10**).



Figure 10 Arrester bed.

Testing and Approval of Road Restraint Systems

Prior to be approved for use in public roads, these systems are subject to exhaustive safety testing.

Tests are specified at different containment or test levels defined in terms of vehicle type and mass, and impact speed and impact. Approval criteria for each test or containment level impose conditions in relation with the structural adequacy of the system, the limitation of occupant risk, and the postimpact trajectory of the vehicle. The tested systems impacted by a vehicle of the specified mass at the specified speed and impact angle must contain and redirect the vehicle in a controlled and predictable manner with no overriding, overriding, or penetration. Furthermore, fragments of the system cannot penetrate the passenger compartment, the vehicle must remain upright during and after the collision, and the passengers must not undergo excessive impact or deceleration.

In the United States, testing conditions and requirements for roadside hardware are specified in the Manual for Assessing Safety Hardware (MASH), while in Europe the European Road Restraint Systems standard EN 1317 is applied.

Within MASH six test levels are defined. Test levels 1 and 2 are designated to qualify the features for work zones, lower service level roadways, and local roads. Test level 3 covers both passenger cars and pickups. The impact speed is 60 mph (97 km/h) with an impact angle of 25 degrees. It is applied to qualify a wide range of higher service level roadways and high-speed highways. Test levels 4–6 are designed to test the ability of restraint systems to contain heavy vehicles. Tests are conducted with single unit trucks at 56 mph (90 km/h), tractor trailers at 50 mph (80.5 km/h), and tractor-tank trailers at 50 mph (80.5 km/h), respectively. The impact angle is 15 degrees in the three cases.

EN 1317 defines three containment classes (normal, high, and very high) with several vehicle impact test descriptions for each class ranging from a light passenger car of 900 kg of mass at 100 km/h with an impact angle of 20 degrees to a 38,000 kg articulated truck at 65 km/h with an impact angle of 20 degrees.

Biography

After completing his PhD in Transportation Engineering at Madrid Technical University in 1995, Dr. Pardillo-Mayora collaborated with the Federal Highway Administration Office of Technology Applications in Washington, DC from May 1996 to September 1997. Since 1998 he is Associate Professor at the Department of Transport Engineering, Urban and Regional Planning of the Technical University of Madrid (UPM) and works as consulting engineer specializing in traffic engineering and road safety studies. He has participated as principal investigator in several national and European road safety research projects including DISCAM (Tools for Designing Safe Roads and Roadsides). He has collaborated on a continuous basis with the Spanish National General Directorate of Roads as senior road safety consultant in the planning and implementation of the National Road System Safety Improvement Programs since 1992. He has collaborated as independent expert with the European Commission's Innovation and Networks Executive Agency (INEA). He is member of the Conference of European Director of Roads (CEDR) Technical Group Road Safety since 2009. He also served as a member of the PIARC International Road Safety Committee (1992–96) and of several US Transportation Research Board Standing Committees since 1999 until 2013. In 2002 he was appointed to the Task Force for the Development of the US Highway Safety Manual (ANB25T) and participated in the preparation of the Manual until its publication by AASHTO in 2010.

Relevant Websites

Federal Highway Administration—Roadway Departure Safety. Available from: https://safety.fhwa.dot.gov/roadway_dept/

Roadside Safety Pooled Fund. Available from: <https://www.roadsidepooledfund.org/>

European Commission TRIMIS—Improving Roadside Design to Forgive Human Errors (IRDES) Project. Available from: <https://trimis.ec.europa.eu/project/improving-roadside-design-forgive-human-errors#tab-outline>

New Zealand Transport Agency—Safe Roads and Roadsides Programme. Available from: <https://www.nzta.govt.nz/safety/our-vision-of-a-safe-road-system/safe-roads/safety-solutions/>

References

- Griffith, M., 1999. Safety evaluation of rolled-in continuous shoulder rumble strips installed on freeways. *Transp. Res. Rec.* 1665, 28–34.
- Pardillo, J.M., Jurado, R., Domínguez, C.A., 2010. Empirical calibration of a roadside hazardousness index for Spanish two-lane rural roads. *Accid. Anal. Prev.* 42, 2018–2023.
- Stonex, K., 1960. Roadside design for safety. Highway Research Board proceedings. Vol. 39. Transportation Research Board. Washington, D.C.
- Thomson, R., Fagerlind, H., Martínez, A., et al., 2005. Roadside Infrastructure for Safer European Roads: D06 European Best Practice for Roadside Design: Guidelines for Roadside Infrastructure on New and Existing Roads. RISER Consortium, Brussels, Belgium.
- Viner, J., 1995. Rollovers on sideslopes and ditches. *Accid. Anal. Prev.* 27 (4), 483–491.
- Zeeger, C., Reinfurt, D., Hunter, W., et al., 1988. Accident effects of side slope and other roadside features on two-lane roads. *Transp. Res. Rec.* 1195, 33–47.

Further Reading

- AASHTO, 2011. Roadside Design Guide, 4th ed. American Association of State Highway and Transportation Officials, Washington, DC.
- AASHTO, 2016. Manual for Assessing Safety Hardware, 2nd ed. American Association of State Highway and Transportation Officials, Washington, DC.
- CEN, 1998. EN 1317—Road Restraint Systems. European Committee for Standardization, Brussels, Belgium.
- Saleh, P., La Torre, F., Domenichini, L., et al., 2013. Forgiving Roadsides Design Guide. Conference of European Directors of Roads, Paris, France.
- Veith, G., 2010. Guide to Road Design Part 6. Roadside Design, Safety and Barriers. Austroads, Sydney, Australia.

Simulators

Nichole L. Morris, Curtis M. Craig, Jacob D. Achtemeier, Peter A. Easterlund, Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	602
Simulation Characteristics, Fidelity, and Validity	602
Considerations in Fidelity	603
Measuring Fidelity	603
Scenario Design	603
Scenario Validation	604
Population and Participants	605
Simulator Sickness	606
Performance Metrics	607
Simulator Measures	607
Psychophysiology, Eye Tracking, and Cognitive Load in Driving Simulation Research	608
Eye Tracking Measures	608
Pupillometry in Task Performance and Cognitive Load	608
Psychophysiology Measures	608
Subjective Measures	609
Conclusions	609
Acknowledgment	609
References	609
Further Reading	610

Introduction

Simulation in transportation provides the promise of exploring, testing, and improving human performance in hypothetical worlds that would be either too expensive or too risky to conduct in the real world. The primary uses of simulation is to either conduct research or provide training. This chapter focuses primarily on the research use of simulation, particularly in driving. Readers should have the tools to critically examine research in transportation simulation and take the first steps to conduct their own research programs. While the domain of interest here is research in driving, the issues raised should be broadly applicable across other transportation domains, along with training methods.

Simulators have been primarily used in both flight transportation and ground transportation. In aviation, simulators are primarily used for pilot training, and both the US Federal Aviation Administration (FAA) and the European Aviation Safety Agency (EASA) have different standards and levels for flight simulators, ranging from training systems for procedures to full flight simulators. In driving, simulators are primarily used for training and evaluation of human behavior in a research context. Driving simulators have been employed for civilian and commercial vehicles (cars, trucks), as well as emergency vehicles and buses.

For effective use and design of a simulator scenario in both aviation and driving, the two predominant concepts are fidelity and validity, which overlap and yet are subtly distinct and are sometimes used interchangeably. Both concepts are covered in detail, but to summarize, fidelity could be considered the realism of the simulated world or the correspondence between the elements of the simulator and the real world, while validity generally refers to whether the simulation has achieved its research or training purpose. Fidelity and validity typically but do not necessarily correspond (Dahlstrom et al., 2009).

Simulation Characteristics, Fidelity, and Validity

Fidelity is the degree to which the simulation is realistic. In theory, if the simulation is high fidelity, performance in the simulator should correspond to performance in the real world. However, there are two major problems with simulation fidelity (Meister, 1990). First, increasing fidelity is expensive, and those building simulators and simulation worlds often find themselves determining how far they can diverge from the real world, while still maintaining value and generalizability in the simulation. Second, there is a subtle problem with the definition of simulation fidelity, in that it can represent at least two different subtypes of fidelity: physical and psychological-cognitive (Liu et al., 2009). Physical fidelity indicates the degree to which the simulation replicates the real environment, and can be broken up into subtypes: Equipment fidelity, motion fidelity, and visual-audio fidelity, see Fig. 1 for examples of low and high physical fidelity driving simulators. Psychological-cognitive fidelity represents the degree to which the simulation requires the use of the underlying psychological and cognitive processes necessary in the real-world setting.

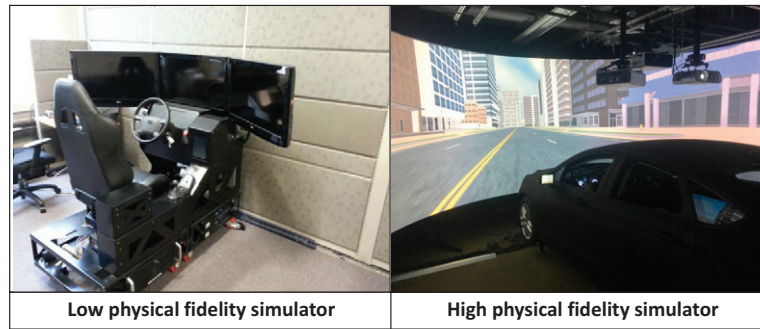


Figure 1 Images of University of Minnesota HumanFIRST Laboratory open cockpit (left) and full chassis (right) driving simulators.

Psychological-cognitive fidelity can be partially decomposed into task fidelity and functional fidelity. Determining and verifying psychological fidelity in simulation falls to methods such as task analysis (Hays, 1980). Hierarchical task analysis and task decomposition methods can guide the development of the simulation world and ensure some cognitive and perceptual components needed to perform the scenario under the research question are adequately implemented. For complex task environments, a survey of psychological and cognitive components in a simulated environment may require a cognitive task analysis, which is then integrated into the simulation with effective scenario design.

Considerations in Fidelity

In training, the focus on physical fidelity primarily arises from identical elements theory (Woodworth and Thorndike, 1901), which argues that practice leads to transfer of learning, if the stimulus elements in both the practice and testing environment are the same. Even outside of the training context, some have argued for the centrality of physical fidelity in simulation, in that perception and action are fully specified by the environment and the higher order relationships (i.e., affordances) between the participant and the environment (Stoffregen et al., 2006). Those emphasizing psychological-cognitive fidelity argue that more indirect cognitive processes moderate these participant-environment relationships and must be considered in both the research and training context. Besides the prohibitive issue of cost in designing high physical fidelity simulations, if the research question or training emphasis focuses on complex decision-making in safety critical domains, or environments with significant communication and team dynamics, then a focus on physical fidelity and face validity may be less effective than that of psychological-cognitive fidelity. This issue has been particularly relevant in aviation simulation and may emerge more often in other areas of transportation as automation and team processes become more predominant (e.g., automated vehicles) (Dahlstrom et al., 2009).

Measuring Fidelity

Verifying fidelity remains a challenge. For equipment fidelity, there are known design standards (e.g., comparison with real aircraft or motor vehicles). For visual-audio fidelity, there are metrics such as resolution, visual angle, and decibels. However, for most other types of fidelity, the metrics are less objective. The primary approach to determine the degree of fidelity has been to utilize subjective measurements (Liu et al., 2009). A typical method is to use experts or known users to rate simulation conditions on either a binary (0 or 1) or Likert (1–5) scale on what degree the simulation condition duplicated the real-world. These ratings can then be averaged to assess overall fidelity. Mathematical models have attempted to be more rigorous in their approach by building a referent or formal representation of reality, (e.g., objects, attitudes, and behaviors), derived from functional analysis and task analysis, and then use this referent and its components as the benchmark to determine fidelity in the simulation.

Scenario Design

Simulator scenario design for driving consists of specifying a road environment, a set of scenario events, and the data to be output. A temptation might exist to focus on the primary experimental event, but the entire scenario must be considered to ensure a successful design and avoid unwanted effects.

The road environment consists of the roadway itself, pavement markings, signs, and landscape surrounding the road. The entire road environment is significant to the scenario design. It is clear that the roadway and regulatory signs are important, but even the landscaping surrounding the road can influence the behavior of the driver in the simulator (Scallan and Carmody, 1999). The holistic design of the global experimental environment consisting of major roadway regions and adjacent environmental features must be considered. Most scenarios consist of several events linked by a period of driving not directly related to an event. The limitations of the simulator are a factor in scenario design. Low-workload driving on straight roads causes more fatigue than driving curved roads (Matthews and Desmond, 2002), which will impact subsequent events. A desktop-style simulator with a limited visual display is not suited for scenarios that require tracking objects in the periphery, such as crossing an intersection. Merging into traffic requires a visual system that incorporates the vehicle mirrors. Motion systems have a significant impact on lane-keeping when the participant is engaged in a secondary task (Greenberg et al., 2003). Each simulator is unique, and the characteristics of the device used must be a design consideration when creating scenarios.

Scenario events are the interactions between the driver and elements within the simulator world. The elements include other vehicles, pedestrians, regulatory and advisory signage, and environmental features. In developing a scenario, it is important to focus on conditions that create the event rather than the desired outcome, as the driver only has control over their speed, following distance, and headway (Papelis et al., 2003). Researchers must be cautious about making assumptions on how the driver will behave and design scenario events accordingly. Scenario testing and “pilot testing” are critical steps in the design of experiments to identify the range of behaviors produced from a set of event conditions and provides data on unexpected driver outcomes or challenges and problems in research protocols, in turn saving resources and time.

Data collected from the simulator is the entire purpose of running an experiment and the desired data fields should be clearly specified in the design prior to protocol creation and during the development and refinement of research questions. Research questions need to be in appropriate scope and assessed through experimental design to ensure accurate generalizability and replicability. When specifying data to be collected, fundamental measures can be used to later derive more complex information but the inverse is not true. The position of vehicles and time can be later used to derive a variety of measures including velocity, acceleration, and time headway. Direct output of higher measures (such as time headway) can be useful, but are of limited usefulness in deriving other data.

Scenario Validation

A key advantage of simulation is the ability to place participants in rare situations without harm that are capable of inducing the same sorts of emotional responses that real-world situations create (Caird and Horrey, 2011). However, balancing the ability to test extreme scenarios with the need for real-world generalization of the observed behavior is the major challenge of transportation simulation, and is partly accomplished through validation of the scenarios. Validity of driving simulation can roughly be defined as the extent to which the simulation matches the intended real-world driving target of interest, typically through comparison of simulated driving to on-road driving, naturalistic driving, self-report measures, or expert assessments. There are multiple subtypes of validity and determining what is most important in the research question or training focus helps to determine which type of validity is most important to achieve (Reimer et al., 2006). Six basic kinds of validity, shown in Fig. 2, should be considered to guide the iterative creation of a driving simulation scenario: face validity, construct validity, content validity, predictive validity, concurrent validity, and external validity.

Face validity of a simulation scenario relates to the appropriateness and relevance perceived by the observer. Achieving face validity often requires detailed design specifications and computer programming to present a high level of realism to subjects and while it is not necessarily a barrier to successful measurement, it can help create the right level of immersion to properly motivate drivers in the task. Beyond establishing visual details, achieving face validity might include the introduction of time. Notably, the types of real-world driving scenarios that result in a fatality are extremely rare per vehicle mile traveled, so recreating them perfectly is a challenge and repeated exposures in such a short exposure period diminishes the ability to elicit the intended

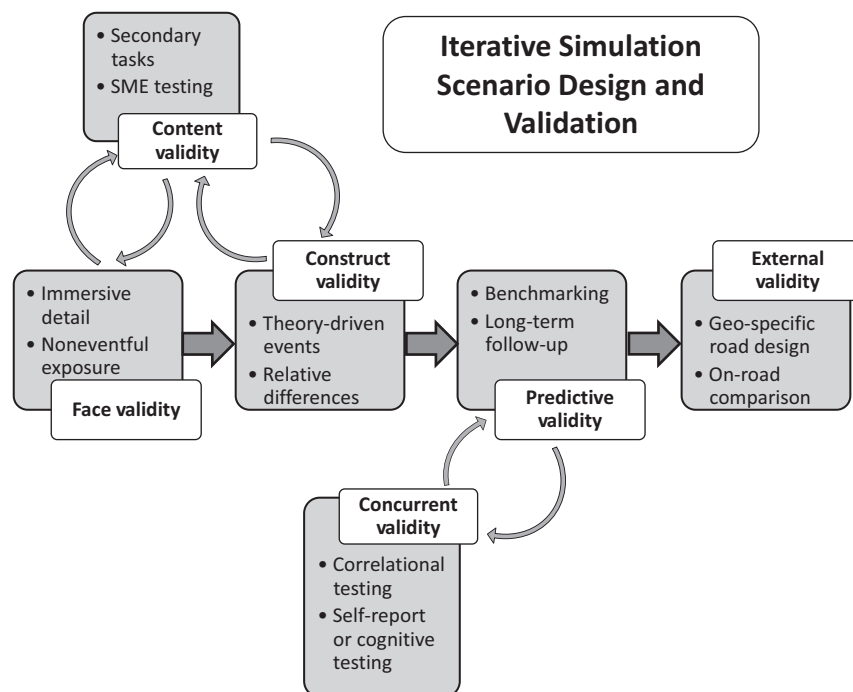


Figure 2 Schematic of validation steps in the iterative design life cycle of driving simulation scenarios.

response (Caird and Horrey, 2011). Countering this effect often requires introducing a surplus of mundane and uneventful road/scenario exposure.

Examining *construct validity* involves broadly determining if the performance concept (e.g., driving skills, risk taking) is being accounted for in the variance. For example, designing a simulator scenario to measure the construct of “driving skill” would likely include a series of complex maneuvers and difficult-to-avoid collision trajectories with other roadway units. Validating driving skill could involve measuring performance across two driving groups with known relative differences to ensure the expected variability is achieved. Such validation can occur within the experimental design or a priori.

Deeper within construct validity falls *content validity* which is the measure of how well the driving scenario captures all of the facets of the construct under investigation (e.g., driving performance). For example, real-world driving performance includes lateral and longitudinal control tasks to maintain position and speed (MacAdam, 2003) while avoiding hazards that are cognitively controlled with both strategic and automated processes (Ranney, 1994). Validating that a simulated scenario contains all of the requisite content (e.g., lateral + longitudinal control, hazard detection) often involves employing subject matter experts (SME). Generally, the majority of SMEs should agree that the scenario or task contains the necessary content of the task (Lawshe, 1975), although the percentage of agreement should be even higher, if the sample of SMEs is small (i.e., fewer than 10).

Predictive validity helps to serve the intention of measuring simulated driving performance to determine the likelihood of future driving outcomes in the real-world, namely the likelihood of crash involvement. Predictive validity studies that examine future crash risk often involve large sample sizes with long-term and costly examinations of self-reported or government recorded crash outcomes. Predictive validity is essential in instances where decision-making using simulated performance may result in discriminatory practices, or implementations with significant real-world impact.

An economical alternative to predictive validity to consider is *concurrent validity* in which simulation performance is tested to determine how well it correlates with intermediate outcomes or measures already validated with other metrics such as crash involvement. Simulated driving performance can be correlated with driving test performance and on-road test performance (Mayhew et al., 2011). Beyond evaluation, simulation performance may be validated against self-report measures (e.g., tendencies, driving history, etc.) and neuropsychological tests that predict driving outcomes. Alternatively, these processes can be used to identify *convergent* and *divergent validity* of which metrics are and are not correlated with driving performance.

External validity is often one of the foremost concerns of practitioners in determining the extent to which driving simulation findings should influence the world. There are a number of challenges to external validity including factors that can be controlled through experimental design, such as road choice, sample size or bias, participant fatigue, and motivation (Kaptein et al., 1996). Creating a simulated driving environment, using geo-specific databases and topographical data, based on an existing road and driving environment affords the opportunity to externally validate simulated driving performance to on-road driving performance.

The following is an example of the proposed procedure for designing and evaluating validity, as outlined in Fig. 2. In a simulated driving task across rural intersections in a driving simulator, the authors attempted to test the validity of the experimental design. Both face validity and external validity were constructed by way of high audio-visual and equipment fidelity by using a full-chassis on a motion base and a 210-degree forward arc screen, and a 7.1 surround sound stereo system, as well as the modeling of a known rural intersection (e.g., geo-specific road design). The content validity was assessed with local county engineers with experience with rural intersections who provided subject matter expert testing with representativeness scores on the simulated world, that is, infrastructure design, environmental elements, and sight distances, as well as the presentation of cross traffic speeds. Construct validity was considered by reviewing and referencing a theoretical framework of collision judgments that would be applicable in the context of intersections, in which both speed, size, and task type affect a person’s perception of collision. Because the limited proposed sample size did not allow for sufficient predictive validity, concurrent validity was instead provided. This was done by the utilization of self-report mental workload and risk perception questionnaires that were expected to covary along with the more difficult experimental trials. By considering these different forms of validity rigorously in the development process, the authors were able to have significantly higher confidence in the relevance of their results to real-world intersections.

Population and Participants

One of the greatest tenants within the field of human factors and ergonomics is “know thy user.” This principle naturally extends into the domain of transportation simulation in two significant ways. First, it is important to create your simulation design and test methods with the actual or target user population in mind. Second, it is critical to properly sample and classify your test population to ensure that your results are valid and properly interpreted. Driving performance and culture are not universal across any substantial fleet and the nuance of expected differences in simulation outcomes across sub-populations requires thoughtful consideration. Careful design of the simulated world, scenario, and tasks should account for the demographics, and the cognitive and physical characteristics of the target population. For cognitive characteristics, this includes experience and expertise level, attitudes, and individual differences within the specific domain (e.g., driving). For physical characteristics, this commonly is captured by height, reach, and physical strength. Gathering the data to define the target population may be challenging but will significantly strengthen the generalizability of the simulation to the real-world environment. Some techniques to gather this population data include surveys, interviews, and focus groups (Etches and Phetteplace, 2013).

Population demographics are most often segmented by gender, age, and experience. These three classification strategies allow blunt but useful parameters for extracting differences in metrics relating to performance, risk-taking, and subjective reporting. Additional sub-groups should be considered based on the task or research question. For driving simulation, there are a number of driver traits that have been associated with driving performance that are more difficult to ascertain through simple screening or eligibility protocols. Personality, skill level, and recent experiences may all contribute to driver performance or response to training in varying ways and must be assessed through structured surveys.

Gender is often a primary consideration for population sampling and generally should include a balanced sample of male and female drivers. In the United States, men are slightly overrepresented in early years of driving (e.g., approximately age 24 years and younger) and women are slightly overrepresented in later years of driving (e.g., age 70 years and older), but these distributions vary across regions and cultures. However, some examinations of occupational driving may not require parity in gender sampling, rather should include a representative sample of the population of focus. Stratified sampling may be required to obtain statistical power, if some subgroups are small.

Age and experience are two distinguishing factors for driving populations that are often interchangeable, that is, young drivers tend to be inexperienced and older drivers tend to be experienced, but there are exceptions to this rule. Migratory patterns may influence the disconnect between age and experience as immigrants with origins in countries where language, culture, or roadways are significantly different from their new location may result in a pronounced increase in crash risk. These changing licensing trends and evolving demographics of the driving fleet should be considered and controlled for in driving simulation population samples.

Performance in transportation simulation tends to vary based on age. Young drivers (i.e., often defined as 15–20 years old) and older drivers (i.e., defined as 65 years and older) are often the age groups that receive the greatest examinations due to their increased crash risks. These gross age classifications, however, may be too broad given known variability of performance and risks within them. Notably, young-old drivers (i.e., age 65–75 years) have been shown to have similar performance as middle-aged drivers (i.e., age 35–55 years) so it is likely inappropriate to group with middle-old (i.e., age 75–85 years) or old-old (i.e., age 85+) drivers because age-related differences would be obscured.

Simulator Sickness

Simulator sickness issues (the experience of feeling ill with their use) is a persistent challenge across multiple types of driving simulators (Stoner et al., 2011). This challenge places limitations on the population used in simulation. Although less severe and less frequent, simulation sickness is similar to the experience of motion sickness including symptoms of fatigue, headache, increased salivation, sweating, stomach awareness, and burping (Kennedy et al., 1993). The expected percentage of the population that will experience simulation sickness is reduced through necessary screening since several known factors are linked to its likelihood. In general, temporary conditions such as illness (particularly within the ear), fatigue, pregnancy, or hangover should delay simulation exposure until conditions are resolved. Permanent precluding factors include any history of motion sickness, epilepsy, or migraine.

Screening for a history of motion sickness is necessary but insufficient to eliminate all drivers who may be likely to experience simulation sickness. Potentially useful is an additional screening protocol of a brief introductory drive that is expected to mildly induce early symptoms of simulation sickness, but not so severe to unduly expose participants to triggering events who might otherwise complete the full simulation (Gable and Walker, 2013). Such measures may serve a dual purpose to acclimate individuals to the simulated environment and reduce sickness incidence rates. Assessments such as the simulator sickness questionnaire (SSQ; Kennedy et al., 1993) or the Motion Sickness Assessment Questionnaire (MSAQ; Gianaros et al., 2001) should be administered before, in between drive segments, and after completion to ensure individuals are well enough to continue or be released. Viewing the participant, directly or through a video stream, is critical to allow the experimenter to look for visual cues of simulation sickness, such as rapid swallowing, burping, mouth movements, sweating, and looking at a fixed object, since individual differences among participants may lead some to be less likely to disclose symptoms or attempt to mask symptoms.

Beyond screening, considering the simulation sickness in simulation scenario design is critical in helping to prevent its incidence rates. Considerations should include visual and temporal characteristics, optic flow, and exposure. Ensuring that the point of view for the driver matches their motion path and expectations is critical, as misalignment will induce discomfort and simulation sickness. The optic flow visually perceived by drivers will not coincide with the vestibular activation of their inner ear since true motion is not occurring. This is related to one of the several theories about the root cause of simulation sickness such as sensory or cue conflict theory (Brooks et al., 2010). Introducing stops, starts, turns, and large objects near the roadway create greater optic flow for drivers and add a greater potential for simulation sickness. Such elements may be instrumental for face validity and the research design but should be used sparingly. The detriment of the optic flow cues may be partially offset through the use of motion systems to stimulate the driver's vestibular system. These systems vary in cost and complexity, ranging from vibration or slight actuation of the cab by a few inches to greater range of motion through hexapod motion bases and track based bases where the entire vehicle moves around a large space. Finally, general guidance suggests that simulation exposure should be broken up with 5–10 minute breaks, with drives lasting no longer than 1 hour (shorter for more intense drives) and total exposure lasting no longer than 2 hours (Stoner et al., 2011).

Performance Metrics

Using a simulator for either research or training purposes requires the selection of useful metrics to determine whether a hypothesis is disproved, what patterns of behavior emerge in hypothetical scenarios, or whether training was effective. Measurements can be derived from the simulator program itself, such as the data output from the primary driving task in a driving simulator (e.g., average speed). Measurements can also be taken from other sources, such as secondary task measures (e.g., cognitive or physical tasks performed concurrently with the primary simulator task), subjective measures (e.g., attitude and usability questionnaires, mental workload scales), and physiological or neurological measures (e.g., pupillometry, eye-tracking, heart rate, functional near-infrared imaging).

Simulator Measures

For measurements derived from the simulator program, one useful organizing framework is to take a time-scale or levels of control approach. In aviation, this distinction is often made in the context of aviation and navigating tasks, in that control of the aircraft's altitude, lateral deviation, and longitudinal position is governed by direct pilot management of pitch, bank, and thrust (Wickens, 2007). In driving, the time-scale or levels of control framework can be delineated into strategic, tactical/maneuvering, and operational/control levels (Michon, 1985; Ranney, 1994). The strategic level involves control of the vehicle guided by long-term plans over the course of minutes, hours, and even days. Examples include route planning and navigation and may be assessed by posing decision-making scenarios within a driving simulation (e.g., best route choice based on some metric, goal planning). The tactical or maneuvering level of control typically involves explicit maneuvers made with the simulated vehicle and usually takes place in a span of seconds. This includes turning maneuvers, lane changes, and overtaking another vehicle. In a safety context, these maneuvers may include gap acceptance, braking and other forms of event monitoring, such as hazard detection and response (e.g., sudden braking of a lead vehicle, intruding vehicles, pedestrians). At the operational or control level, which is typically measured at millisecond time scales, the simulation scenarios are concerned with lateral and longitudinal control of the vehicle (Young et al., 2009). Examples of this include lane keeping, speed control, following behavior, and headway maintenance. See Fig. 3 for an overview of this time-scale or levels of the control framework for direct simulator measurements.

Even if distinguished in this fashion, there is still the question of exactly what measurement is used for each level and assessment task. Strategic level measurements usually comprise of choices and decisions made by the human participants, and the quality of these choices is determined by some prior referent or normative standard. For example, in a driving simulation, participants could decide which route to take upon arriving at a simulated intersection, given prior information such as travel time and complexity (Tian et al., 2011).

Tactical or maneuvering level measurements can be partially determined by the maneuver or event detection task scenario, given that the metric is also generally understood in a real-world recreation of the same task scenario. In the context of vehicle maneuvers: Turning can be measured by the degree of steering wheel turn and acceleration, although expert scoring of turning errors by participants has also been employed in this context (Shechtman et al., 2009). Lane changes may be measured by whether a lane change occurred (i.e., two or four wheels crossed over the lane line during the duration of the lane change event), and the average time to cross over the lane edge (e.g., time from two wheels to four wheels over). The lane change test (LCT) has also been used to measure the efficacy of changing lanes in the context of distraction, and primarily relies on reaction time as a metric (Mattes and Hallén, 2009). Overtaking maneuvers have been assessed looking at a count of when drivers crossed the roadway when it was not

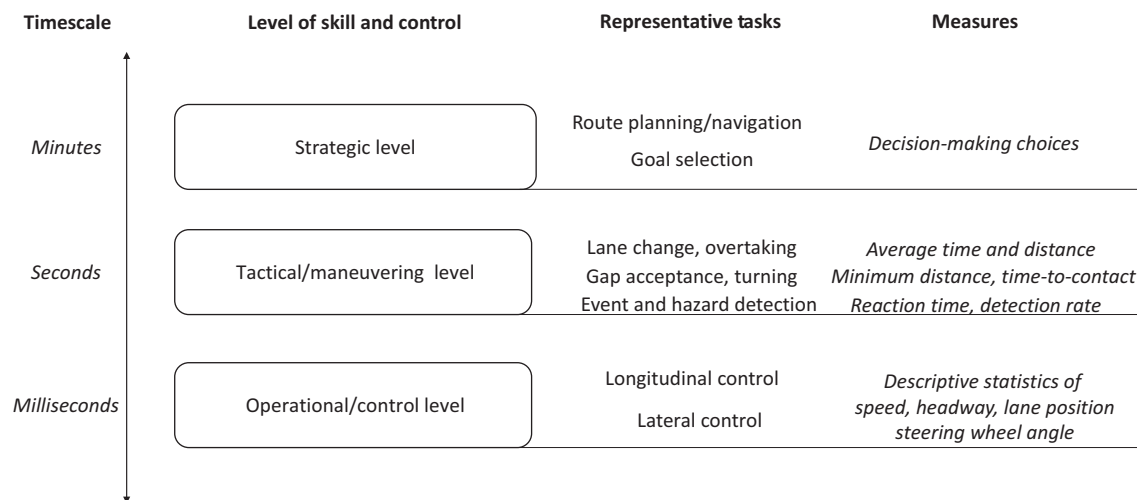


Figure 3 Overview of timescale or levels of control framework for simulator metrics.

permitted (e.g., double lines on the simulated roadway indicating no overtaking permitted), the lateral minimum distance between the simulated vehicle and the overtaken vehicle, and lane position and speed.

In the safety context of event detection and monitoring, this is often done in the context of distracting secondary tasks to determine whether participants in a simulator respond to events while dual-tasking (e.g., reaction time to traffic lights by acceleration or foot off the brake, [Young et al., 2009](#)). Also, event detection and monitoring are relevant to hazard detection, such as measuring the braking time from event onset, such as a hidden pedestrian or stop sign ([Lee et al., 2008](#)). Measures for event detection are frequently reaction or response time based, given that the speed of response generally indicates the moment a person has spotted a hazard or a critical event. However, one could consider distance metrics from the hazardous simulated object, or utilize time-to-contact metrics ([Hancock and Manser, 1997](#)). Time-to-contact is generally the calculation of the distance between two vehicles divided by their speed, which provides an estimate of when the two simulated vehicles will come into contact with each other. For somewhat more complex safety-critical scenarios such as gap acceptance, metrics typically focus on the gap in seconds between the participant's vehicle and the oncoming vehicle along with the reaction time of the decision to cross (usually via accelerator pedal press). Also, in the context of turning maneuvers and gap acceptance, one could measure the rotation angle of the steering wheel.

Operational or control level measurements are typically consistent no matter the overall driving scenario, as it is assumed that significant variation in these measures reflects poor vehicle control, whatever the proposed mechanism behind the poor control may be, such as driver inexperience, stress, fatigue, distraction, or high cognitive load. These measures are typically concerned with lateral and longitudinal control of the vehicle ([Young et al., 2009](#)). Longitudinal control has normally been measured in terms of speed (average speed, maximum speed, variability or standard deviation of speed, and 85th percentile speed) and vehicle following (time and distance based average headway, standard deviation or variability of headway, and minimum headway). Meanwhile, lateral control of the vehicle has been assessed via lane keeping or lateral position in relation to the center of the lane (e.g., average lane position, standard deviation of lane position, and count of deviations from the lane), along with steering wheel measurements (e.g., standard deviation of steering angle, reversal rate, and steering entropy).

Psychophysiology, Eye Tracking, and Cognitive Load in Driving Simulation Research

Psychophysiology, or the study of measuring interrelationships between physiological biomarkers and cognitive activity, afford researchers with opportunities to compliment driver performance in driving simulator scenarios to help validate and discern relative psychological changes based on physiological biomarkers. Heart rate variability, electroencephalography, and electrodermal activity are common psychophysiological methods used to infer driver psychological states, mental workload, and engagement.

Eye Tracking Measures

Saccadic eye movement patterns and fixations on areas of interest are two commonly used eye-tracking measures in driving simulation studies. Saccades are defined as the movement of the eyes between brief moments of pause when fixating, or examining a set location when the eyes are stationary. Fixations and saccades are determined by equipment sampling resolution and algorithms. Changing settings for fixation duration (ms) criteria and saccadic filtering algorithms using the same data set can result in markedly different output results in analyses ([Shic et al., 2008](#)). Ensuring fixation and saccade threshold criteria fit experimental interventions (e.g., target presence, collision event) is critical for conducting a robust driving simulator experiment with eye tracking measures. Most commercial eye tracking systems provide a minimal level of control in setting dependent measures criteria, and system algorithms' underlying calculations are not readily modifiable.

Pupillometry in Task Performance and Cognitive Load

Pupillometry is a derivative measure of general eye tracking methods considering at average pupil size and variability. While driving performance measures provide useful insights into the general state of the driver, assessing dynamic pupillary changes while drivers perform simulated driving tasks can indicate task-based, time-sensitive changes in such cognitive load ([Beatty, 1982](#)). Popular methods to capture cognitive load using pupillometry are based on the mean of the pupil and its rate of diameter change, and the number of measurable changes in diameter over a period of time ([Palinko et al., 2010](#)). Rapid changes in the mean pupil diameter change rate is a benchmark method that is easily accomplished using commercial eye tracking systems.

Psychophysiology Measures

Electroencephalography (EEG) typically involves a 56–256 electrode array on a wearable cap, and provides general insight into which cortical brain regions are activated or inhibited during task performance. In studies where specific events are triggered based on participant position (e.g., lane departure), driving behavior (e.g., collision event), or instructions to engage in secondary tasks, EEG measurement techniques such as event-related potentials can be implemented to discriminate cortical brain activity between task and non-task conditions. Data can be used to provide insights into the task demands via neural activity patterns. Electrodermal activity (EDA), historically referred to as galvanic skin response or skin conductance, involves the detection of resistance via skin

Table 1 Recognized criteria for subjective mental workload measurement

<i>Criterion</i>	<i>Description</i>
Sensitivity	Measure varies with changing task demand level
Diagnosticity	Assesses qualitative differences in task demand and their causes
Selectivity or validity	Measures construct in question (i.e., mental workload)
Intrusiveness	Extent to which mental workload measurement affects primary task performance
Reliability	Measurement method consistency in repeated administrations
Implementation requirements	Resources required to use measurements (time/materials)
Subject acceptability	Participant perceived validity and usefulness of measurement

sweat as secreted across the palm or bottom of the foot. While this psychophysiology measurement generally uses only two electrodes, these placement locations may prove problematic in some apparatus or experiment designs due to the involved use of hands and feet during driving tasks. EDA may be used for indicating cognitive demand and stress in a simulation (Collet et al., 2014; Reimer and Mehler, 2011). Electrocardiogram (EKG) affords many opportunities to capture autonomous nervous system activity during driving simulation research and is achieved by using electrode arrangements. One use of EKG includes measurement of heart rate variability (HRV), defined as the variation in heart rate (beats per minute) over time interval (s). HRV is a reliable and relatively non-invasive psychophysiological secondary measurement for assessing psychological states such as anxiety or stress levels, focused attention, boredom, and drowsiness during simulation tasks (Fujiwara et al., 2018).

Subjective Measures

Traditionally, subjective measures have been used to gather insight into mental workload and relative stress levels experienced by participants during simulation studies (Brookhuis and de Waard, 2010). The most commonly used examples include the NASA Task Load Index, the Subjective Workload Assessment Technique, and the Rating Scale Mental Effort. A useful review of mental workload assessment criteria for subjective measures appears in Table 1, which may be employed for other constructs besides mental workload (Paxion et al., 2014).

Conclusions

- Many long-standing issues with simulation research remain unresolved, such as adequacy of fidelity and validity, participant selection, selection of measures, and simulator sickness.
- Fidelity measures the degree to which the simulation is realistic. Validity measures the extent to which the simulation and scenario assess and measure what they were intended to assess and measure. Both concepts are multifaceted with several subtypes, and each subtype should be carefully considered in turn.
- The identification of the target population and careful selection of participants is important for behavioral research in simulators. Gender, age, and other individual differences such as experience, personality, and ethnic group (s) should all be considered.
- Simulator sickness remains a major constraint on those conducting simulation research. Best practices utilize careful screening measures to limit the impact of simulation sickness, including questionnaires for similar conditions (e.g., motion sickness), as well as short trial runs in the simulator to test proclivities toward simulation sickness in the longer experimental simulations.
- Effective selection of measurement within the simulated scenario requires consideration of both the research question and validity concerns. Measurement can comprise of performance within the simulator itself (e.g., choices, response time, vehicle control metrics), as well as secondary measures, including but not limited to subjective measures, eye-tracking measures, and psychophysiological measures.

Acknowledgment

The authors would like to thank HumanFIRST research assistants McKenzie Shepard, and Savindie Liyanagamage.

References

- Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* 91, 276–292, doi:10.1037/0033-2909.91.2.276.
- Brookhuis, K.A., de Waard, D., 2010. Monitoring drivers' mental workload in driving simulators using physiological measures. *Accid. Anal. Prev.* 42, 898–903.
- Brooks, J.O., Goodenough, R.R., Crisler, M.C., Klein, N.D., Alley, R.L., Koon, B.L., Logan Jr., W.C., Ogle, J.H., Tyrrell, R.A., Wills, R.F., 2010. Simulator sickness during driving simulation studies. *Accid. Anal. Prev.* 42, 788–796.
- Caird, J.K., Horrey, W.J., 2011. Twelve practical and useful questions about driving simulation. In: Fisher, D.L., Rizzo, M., Caird, J., Lee, J.D. (Eds.), *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. CRC Press, USA, pp. 5–1–5–18.

- Collet, C., Salvia, E., Petit-Boulanger, C., 2014. Measuring workload with electrodermal activity during common braking actions. *Ergon.* 57 (6), 886–896, doi:10.1080/00140139.2014.899627.
- Dahlstrom, N., Dekker, S., Van Winsen, R., Nyce, J., 2009. Fidelity and validity of simulator training. *Theor. Issues Ergon. Sci.* 10, 305–314.
- Etches, A., Phetteplace, E., 2013. Know thy users: user research techniques to build empathy and improve decision-making. *Ref. User Serv. Q.* 53, 13–17.
- Fujiwara, K., Abe, E., Kamata, K., Nakayama, C., Suzuki, Y., Yamakawa, T., Hiraoka, T., Kano, M., Sumi, Y., Masuda, F., Matsuo, M., Kadowaki, H., 2018. Heart rate variability-based driver drowsiness detection and its validation with EEG. *IEEE Trans. Biomed. Eng.* 66, 1769–1778, doi:10.1109/TBME.2018.2879346.
- Gable, T.M., Walker, B.N., 2013. Georgia Tech Simulator Sickness Screening Protocol. Georgia Tech School of Psychology Tech Report GT-PSYC-TR-2013-01. Georgia Tech, Atlanta, GA.
- Gianaros, P.J., Muth, E.R., Mordkoff, J.T., Levine, M.E., Stern, R.M., 2001. A questionnaire for the assessment of the multiple dimensions of motion sickness. *Aviat. Space Environ. Med.* 72 (2), 115.
- Greenberg, J., Artz, B., Cathey, L., 2003. The effect of lateral motion cues during simulated driving. *Proceedings of the Driving Simulation Conference North America*, Dearborn, Michigan.
- Hancock, P.A., Manser, M.P., 1997. Time-to-contact: more than tau alone. *Ecol. Psychol.* 9, 265–297.
- Hays, R.T., 1980. Simulator Fidelity: A Concept Paper. (Report 490). U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Va.
- Kaptein, N.A., Theeuwes, J., Van Der Horst, R., 1996. Driving simulator validity: some considerations. *Transp. Res. Rec.* 1550, 30–36.
- Kennedy, R.S., Lane, N.E., Berbaum, K.S., Lilienthal, M.G., 1993. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *Int. J. Aviat. Psychol.* 3, 203–220, doi:10.1207/s15327108ijap0303_3.
- Lawshe, C.H., 1975. A quantitative approach to content validity. *Personnel Psychol.* 28, 563–575.
- Lee, S.E., Klauer, S.G., Olsen, E.C., Simons-Morton, B.G., Dingus, T.A., Ramsey, D.J., Ouimet, M.C., 2008. Detection of road hazards by novice teen and experienced adult drivers. *Transp. Res. Rec.* 2078, 26–32, doi:10.3141/2078-04.
- Liu, D., Macchiarella, N.D., Vincenzi, D.A., 2009. Simulation fidelity. In: Vincenzi, D.A., Wise, J.A., Mouloua, M., Hancock, P.A. (Eds.), *Human Factors in Simulation and Training*. CRC Press, USA, pp. 61–73.
- MacAdam, C.C., 2003. Understanding and modeling the human driver. *Vehicle Syst. Dyn.* 40 (1–3), 101–134, doi:10.1076/vesd.40.1.101.15875.
- Mattes, S., Hallén, A., 2009. Surrogate distraction measurement techniques: The lane change test. In: Regan, M.A., Lee, J.D., Young, K.L. (Eds.), *Driver Distraction: Theory, Effects, and Mitigation*. CRC Press, Boca Raton, FL, pp. 107–121.
- Matthews, G., Desmond, P., 2002. Task-induced fatigue states and simulated driving performance. *Q. J. Exp. Psychol. Sec. A: Human Exp. Psychol.* 55, 659–686, doi:10.1080/02724980143000505.
- Mayhew, D.R., Simpson, H.M., Wood, K.M., Lonero, L., Clinton, K.M., Johnson, A.G., 2011. On-road and simulated driving: concurrent and discriminant validation. *J. Saf. Res.* 42, 267–275.
- Meister, D., 1990. Simulation and modeling. In: Wilson, J.R., Corlett, E.N. (Eds.), *Evaluation of Human Work: A Practical Ergonomics Methodology*. Taylor and Francis, UK, pp. 202–228.
- Michon, J.A., 1985. A critical view of driver behavior models: what do we know, what should we do? In: Evans, L., Schwing, R.C. (Eds.), *Human Behavior and Traffic Safety*. Plenum Press, USA, pp. 485–520.
- Palinko, O., Kun, A.L., Shyrovok, A., Heeman, P., 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-tracking Research & Applications*. ACM, Austin, TX, pp. 141–144.
- Papelis, Y., Ahmad, O., Watson, G., 2003. Developing scenarios to determine effects of driver performance: techniques for authoring and lessons learned. *Proceeding of the Driving Simulation Conference North America*, Dearborn, Michigan.
- Paxion, J., Galy, E., Berthelon, C., 2014. Mental workload and driving. *Front. Psychol.* 5, 1–11.
- Ranney, T.A., 1994. Models of driving behavior: a review of their evolution. *Accid. Anal. Prev.* 26, 733–750.
- Reimer, B., D'Ambrosio, L.A., Coughlin, J.F., Kafrissen, M.E., Biederman, J., 2006. Using self-reported data to assess the validity of driving simulation data. *Behav. Res. Methods* 38, 314–324.
- Reimer, B., Mehler, B., 2011. The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergon.* 54, 932–942, doi:10.1080/00140139.2011.604431.
- Scallan, S., Carmody, J., 1999. Investigating the effects of roadway design on driver behavior: Applications for Minnesota highway design. University of Minnesota and the Minnesota Department of Transportation, Saint Paul, MN.
- Shechtman, O., Classen, S., Awadzi, K., Mann, W., 2009. Comparison of driving errors between on-the-road and simulated driving assessment: a validation study. *Traffic Inj. Prev.* 10, 379–385, doi:10.1080/15389580902894989.
- Shic, F., Scassellati, B., Chawarska, K., 2008. The incomplete fixation measure. In *Proceedings of the 2008 symposium on Eye tracking research & applications*. Savannah, GA: ACM, Savannah, GA, pp. 111–114.
- Stoffregen, T.A., Bardy, B.G., Mantel, B., 2006. Affordances in the design of enactive systems. *Virtual Reality* 10, 4–10.
- Stoner, A.H., Fischer, L.D., Mollenhauer Jr., M., 2011. Simulator and scenario factors influencing simulator sickness. In: Fisher, D.L., Rizzo, M., Caird, J., Lee, J.D. (Eds.), *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. CRC Press, USA, pp. 1–14.
- Tian, H., Gao, S., Fisher, D.L., Post, B., 2011. Route choice behavior in a driving simulator with real-time information. In *Proceedings of the Transportation Research Board 90th Annual Meeting*. Washington D.C., USA.
- Wickens, C.D., 2007. Aviation. In: Durso, F.T. (Ed.), *Handbook of Applied Cognition*. Wiley, New York, NY, pp. 361–390.
- Woodworth, R.S., Thorndike, E.L., 1901. The influence of improvement in one mental function upon the efficiency of other functions (I). *Psychol. Rev.* 8, 247–261.
- Young, K.L., Regan, M.A., Lee, J.D., 2009. Measuring the effects of driver distraction: Direct driving performance methods and measures. In: Regan, M.A., Lee, J.D., Young, K. (Eds.), *Driver Distraction: Theory, Effects, and Mitigation*. CRC Press, USA, pp. 85–105.

Further Reading

- Fisher, D.L., Rizzo, M., Caird, J., Lee, J.D., 2011. *Handbook of Driving Simulation for Engineering, Medicine, and Psychology*. CRC Press, USA.
- Hancock, P.A., Vincenzi, D.A., Wise, J.A., Mouloua, M., 2008. *Human Factors in Simulation and Training*. CRC Press, Boca Raton, FL, USA.
- Lee, A.T., 2005. *Flight Simulation: Virtual Environments in Aviation*. Ashgate Publishing Ltd, Hampshire, UK.

Sleep-Related Issues and Fatigue

Prof Marion Sinclair, Estelle Swart, Department of Civil Engineering, Stellenbosch University, South Africa

© 2021 Elsevier Ltd. All rights reserved.

Fatigue and Crash Likelihood	611
Defining Fatigue	612
Causes of Fatigue	612
Drivers Most Susceptible to Fatigue	613
Diagnosing Fatigue	613
Preventing Fatigue	614
Enforcement	614
The United States	614
European Agreement (AETR)	614
United Kingdom	615
Australia	615
Educational Initiatives	615
Engineering Responses	615
See Also	616
References	616

Fatigue and Crash Likelihood

Driving is a complex task that requires constant attention and a high level of focus. Fatigue impairs cognitive and physical performance and so greatly reduces the ability to drive safely and efficiently (Schmidt et al., 2009). In fact, fatigue is reportedly one of the leading causes of crashes worldwide. In the United Kingdom, fatigue is a contributory factor in some 20% of crashes and in North America the rate is reportedly around 40%. In spite of the magnitude of the problem, driver fatigue is scarcely mentioned in many road safety policies across the world. Limited countermeasures are in use and the effectiveness of these methods is still largely unknown.

Of course fatigue is also a problem in transport modes other than road transport. In aviation, pilot error is believed to be responsible for around 80% of plane crashes, and fatigue itself is identified as a factor in between 15% and 20% of these. Likewise, investigations into rail crashes routinely identify driver fatigue as a problem—studies in Europe, for example, show that between 20% and 25% of serious train crashes have been associated with drivers being drowsy, asleep or generally fatigued. Fatigue and sleep-related issues are clearly a matter of concern across many modes of transportation. The vast majority of the research into sleep and fatigue has emerged from studies of motor vehicle drivers; this is natural as this is the most commonly experienced context of fatigue and one whose consequences are seen on a daily basis.

Although there is widespread recognition that fatigue and sleep-related issues pose serious problems in transportation systems, it is immensely difficult to implement effective solutions, though experts in all sectors of the transport industry are working to achieve this. There are two main problems with fatigue. The first is that there is no internationally recognized or accepted definition of fatigue. Fatigue is a concept far broader than sleepiness—at essence it is about performance, and the competence of a driver to respond appropriately to the demands of the drive (Philip et al., 2005; Liu and Wu, 2009). Performance itself is a function of physiological impairment and psychological state. While this may sound straightforward, neither of these dimensions is simple to identify or measure.

This links to the second problem with fatigue, which is that the physical manifestations of fatigue (generally identified as sleepiness) are difficult to detect without the use of invasive, high-tech, and resource-intensive measures, such as electroencephalography (EEG) or pupillometry. Identifying fatigue as a contributory factor after a crash is also difficult. Fatigue is seldom recorded as a contributory factor in road crash reports, unless it is very clearly evidenced in the crash forensics, or strongly suspected by the investigating officer. Such cases are rare and while there may be localized reports produced from police crash investigations, there are no published articles on sleep-at-the wheel crashes from which we can accurately determine the incidence of crashes caused by drivers actually falling asleep.

Falling asleep is, of course the most extreme outcome of fatigue and presents the highest risk in terms of crash and injury. Crashes caused by sleeping drivers tend to involve single vehicles drifting either off the road, or into oncoming vehicles; both likely to end in head-on crashes with high severity injuries. But the physical consequences of fatigue that lead up to sleep are also a concern. Even low levels of drowsiness cause drivers to pay less attention, to make errors in their judgment (for example of other vehicles position and speeds); and to miss key information and hazards.

Given the range of physical responses to fatigue, and the difficulty in identifying fatigue scientifically—or in identifying a critical point at which normal tiredness becomes something more dangerous—it is unlikely that fatigue as an element of fitness to drive (or to pilot an airplane for that matter) can easily be operationalized. One legal proxy currently utilized for fatigue is number of hours driven (or flown). This is finding its way into some traffic legislation across the world, and there are already restrictions on the number of hours that pilots can fly (in Europe this is currently 11 hours in one shift). In road transport however, it is difficult for traffic enforcement officials to enforce the number of hours a driver has been driving and it is often impossible to use it as a basis for prosecution. “Maggie’s law” (New Jersey, US) was the first piece of legislation to be passed to actively facilitate the prosecution of fatigued drivers involved in crashes. Under this law, a sleep-deprived driver can be certified as “reckless” and can subsequently be convicted of vehicular homicide. Maggie’s law was based on scientific evidence that sleep deprivation has direct and measurable effects of a driver’s competence (Arnedt et al., 2001; Rupp et al., 2007). This research had demonstrated that being awake for 18 hours produces impairment equivalent to a blood alcohol concentration (BAC) of 0.05%. At 24 hours the equivalence is around 0.1%, meaning that severely fatigued drivers who remain awake are potentially even less competent on the road than drivers who are legally drunk. Those drivers who fall asleep at the wheel face the highest risks of all.

The relationship between hours awake and safety performance is not as simple as it may seem, however, and the effect of fatigue on the performance of drivers is not identical in all drivers. In fact some research suggests that our brains have an extraordinary ability (when fatigued) to compensate for diminished cognitive functioning, allowing us to maintain adequate levels of performance under non-challenging driving environments (Philip et al., 2005). As a result, most drivers feel confident when driving tired; their brains have found a way to maintain a degree of cognitive competence. We don’t fully know how brains achieve this—we do, however, know that this competence is temporary and unreliable, as evidenced in high numbers of fatigue related crashes.

The scale of the actual problem of fatigue is hard to ascertain. Crashes resulting from fatigue and sleeping drivers are the tip of the iceberg, as many more near crashes occur which are never documented. Possibly the closest approximation of the extent of the problem in driving comes from research into drivers’ past experiences. Surveys of drivers from across multiple countries confirm that around half of drivers will admit to having driven while drowsy, and that frighteningly high levels (between 13% and 37%, depending on source) admit to having fallen asleep at the wheel. These numbers tend to be higher for males, and highest in the population of heavy goods vehicle drivers.

The significance of Maggie’s law was not that it conclusively defined fatigue, but that it was the first legislation of its kind to enable the prosecution of a driver for reckless driving, specifically for driving while knowingly fatigued. The fact that only one other state in the United States has adopted this legislation is possibly indicative of the difficulties that remain with proving fatigue; that is, proving the link between lack of sleep and performance impairment. This illustrates the importance of finding a measurable aspect of fatigue to eliminate fatigued drivers from the roads and therefore reduce crashes.

Defining Fatigue

Various researchers have defined fatigue; however, there is no consensus view and many definitions highlight the multidimensionality of fatigue as a construct. Elements commonly addressed include an overwhelming exhaustion, a lack of energy, a feeling of imminent sleepiness, and poorer physical and cognitive functioning. These concepts are helpful in conveying a shared understanding of the term but less useful when it comes to developing a watertight definition for prosecution purposes.

Causes of Fatigue

Complex in definition, fatigue is also complex in origin. It has many underlying potential causes and the interaction between contributory causes is also complex. The main causes are:

Sleep deprivation: This can be summarized succinctly as an insufficient amount of quality sleep over a sustained period. Different individuals require different amounts of sleep, but under this definition sleep deprivation emerges when a person has not had seven hours of sleep for two consecutive nights, or when that sleep has been fragmented, restricted or interrupted. The longer a person goes with less than average sleep per night, the faster sleep deprivation accumulates. Sleep deprivation has been shown to have a significant impact on cognitive functions of the brain including alertness, perceptual skills, reaction times, psychomotor coordination, judgment, decision making and risk propensity (Curcio et al., 2001; Lim and Dinges, 2008).

Circadian rhythm: All living organisms are subject to daily routines, which affect the amount of energy they are able to expend on tasks. This is known as circadian rhythm, and in humans the hormone responsible for circadian rhythm is melatonin, which is manufactured in the pineal gland. As melatonin levels increase, alertness and information-processing functions of the brain drop. There are typically two periods each 24 hours in which melatonin levels peak—for the average person these are between midnight and 6:00 a.m., and then again between 2:00 and 4:00 p.m. Of course age, gender, and sleeping routine can influence this (so, e.g., people working night shifts will experience different lows). These periods of lower cognitive functioning coincide with feelings of fatigue, and with reduced ability to operate machinery and to cope with the demands of complex tasks, such as driving. It is not a coincidence that fatal crash numbers are disproportionately high during the period of midnight to 6:00 a.m. (when normalized against traffic volumes).

Chronic sleep disorders: These include medical conditions, such as obstructive sleep apnea, restless leg syndrome (RLS), narcolepsy and chronic fatigue syndrome. All of these conditions undermine levels of alertness and physical and cognitive function, though to different degrees. While many people across the world suffer from such conditions, a key problem associated with chronic sleep disorders is that they can often go undiagnosed; with drivers being largely unaware that their sleep patterns may result in them driving fatigued. For example, meta-analyses of studies in the United States and Europe suggest that around 20% of adults suffer some degree of sleep apnea, with only 5% being formally diagnosed and undergoing treatment. Significantly fewer people (e.g., estimated at 0.05% of the US citizens) suffer from narcolepsy, with only a fraction of these receiving the treatment. Though a less common condition, sufferers of narcolepsy are particularly at risk of crashing as they are subject to periods of overwhelming sleepiness throughout the day.

Length of drive, and high demand conditions: Research into the relationship between work effort and fatigue has concluded that the degree of fatigue experienced is directly proportionate to the length of a task performed. Sustained task performance is related to performance decrements. Essentially, the longer a task is performed, the worse it is carried out, a fact that has been confirmed in both laboratory studies and also among drivers in real driving conditions.

When a task involves continuous cognitive engagement—that is, when the task is difficult or requires especially high levels of attention, fatigue will emerge at a faster rate than under low demand conditions. On the other hand, monotonous driving environments pose their own risks for drivers as they can induce a feeling of sleepiness simply because the driver's brain is not sufficiently engaged with the task. A road with a low flow density and a monotonous road environment can cause depreciation in alertness and driver vigilance (Thiffault and Bergeron, 2003).

Medication and drugs: The use of some antibiotics, antidepressants, and antihistamines can cause drowsiness and attention deficits in drivers. While these side effects are clearly stated on the medication inserts, not all drivers may be aware of these before they take them. Illegal drugs and alcohol also affect cognitive function. Alcohol in particular has depressive attributes, which affect awareness as well as reaction times.

Shift schedule: Professional drivers are one of the groups of people who are more likely to be affected by fatigue; especially when their employment requires shift driving at different times of the day. A science of shift schedules has emerged in recent years, which minimizes the impact of shift changes on circadian rhythm, and allows a driver's body time to adjust to new time demands. Shift schedules are applied in many industries such as the mining and medical industries, as well as (increasingly) in the transportation sector.

Drivers Most Susceptible to Fatigue

A few groups of drivers are more susceptible to fatigue than others. They include:

- Professional drivers—not only during their driving shifts, but also driving home at the end of the shift.
- Drivers who work varying shift patterns.
- People who work more than 72 hours a week.
- Drivers who drive at times when their circadian rhythm suggests they should be resting
- Long distance drivers, especially where insufficient rest stops have been taken.

Diagnosing Fatigue

Post-crash: Fatigue is often designated a contributory factor in crashes only when other factors have been eliminated. In fatal crashes, for example, fatigued drivers are commonly identified only by a lack of evidence of other causal factors, as well as no indication that evasive action had been taken by the driver. The underlying reason here is that fatigued drivers lose the ability to respond appropriately to unexpected events.

Pre-crash: In international legislation (where it exists) fatigue is most often defined simply on the basis of time without sleep, or the time elapsed since the last rest period. This is a pragmatic approach to some of the difficulties associated with physical identification of fatigue on the basis of physiological symptoms (e.g., changes in eye movement, heart rate, and so on). There are a number of scientific techniques available for conducting such analysis but they tend to be reliant on technology and on expensive equipment. Also, some measures available to accurately detect fatigue depend on prior knowledge of a driver's physiology in a non-fatigued state (e.g., heart rate changes requires an understanding of the baseline heart rate for that person before fatigue). This is clearly impossible in most roadside investigations.

Fatigue measuring techniques that have been developed for laboratory conditions include:

- **Percentage of eye closure (PERCLOS):** The PERCLOS technique calculates the percentage of time in a minute that eyelids are closed. According to the US Federal Highway Administration (FHWA) and US Highway Traffic Safety Administration (NHTSA), this is regarded as the most promising driver fatigue detecting technique (Singh et al., 2011).
- **Ocular parameters:** Measuring the characteristics of eye movement and pupil constriction.
- **Heart rate variability (HRV):** The use of HRV can serve as an early fatigue indicator. The HRV increases exponentially with increasing fatigue levels.

- Neurological parameters: Changes in oscillatory brain activity occur with fatigue, and can be used to identify mental fatigue.
- Head movements: Head nodding is the last cue before the occurrence of micro-sleeps and the onset of sleep. This measurable aspect is an efficient approach to determine fatigue. Unfortunately, it is the last cue providing an indication that the driver is falling asleep and might have been unfit to drive.

Assessing fatigue in real driving environments is more challenging, given the constraints of time, space, specialist staff, and budget limitations. There are numerous devices that have been developed for experimental use but these are not yet commercially available. One such device is the psychomotor vigilance test (PVT), which uses the reaction time of a driver to determine his/her level of alertness. Reaction times are arguably the most important aspect of driver performance when it comes to crash avoidance. They can be measured on palmtop machines with little in the way of supporting computer backup.

A second method, which avoids technology altogether, is a self-assessment test carried out by drivers. Many instruments have been developed over the years to allow drivers to record their experiences—of them the Swedish Occupational Fatigue Index (SOFI) stands out as a sound way of measuring fatigue among professional drivers. This technique incorporates multiple dimensions of fatigue, including physical, cognitive, and emotional fatigue, grounded in a test that can be customized to the demands to the task—in this case the driving environment.

A third method of assessing fatigue is the tachograph—a control device that records journey data of a vehicle. This technique, unlike the two mentioned above, does not attempt to measure the impact of the task on the driver, but makes a fairly simplistic assumption that the performance of the driver is directly related to the amount of time spent driving. It disregards any other external factors that could have an influence on the driver. The tachograph presents driving time as a proxy for fatigue. While this is a crude measure, it has the advantage of being consistent and for allowing legislation to be developed which similarly defines fatigue simply in terms of hours spent driving.

Preventing Fatigue

Enforcement

Fatigue in international legislation:

The manner in which traffic officials manage fatigued drivers is dependent on whether fatigue is defined in the law, the clarity of that definition, and the punishment that is consequently indicated. Where any of these elements is unclear, successful prosecution of drivers for fatigued driving is unlikely, and seldom attempted. Of course many countries have legislation that enables a wide range of generic forms of dangerous driving to be prosecuted, but because of the many difficulties associated with identifying fatigue as a definite cause, fatigue itself is seldom prosecuted under such laws.

Fatigue related legislation does exist in a small number of countries, and examples from the United States, Europe, and Australia are summarized below.

The United States

Maggie's law, New Jersey, which was passed in 2003, expanded the state's prior definition of a reckless driver to include fatigue. In this case, fatigue is defined simply in terms of time without sleep (defined as having had no sleep for 24 hours). Under this law, a driver who causes a fatal crash due to fatigue can be prosecuted for vehicular homicide. No performance-related impairments need to be proven; no physical indicators of fatigue are examined. This is a positive step forward in that fatigue is recognized as being detrimental to driving and a significant contributor toward recklessness on the roads. However, the take-up of the model has been low. Other states within the United States generally believe that there is sufficient provision in their legislation for fatigued drivers to be prosecuted (e.g., under their definitions of reckless driving and vehicle homicide laws). The fact that fatigued driving in the United States does not appear to be declining suggests that the legislation that does exist is not effective.

European Agreement (AETR)

The European Union has a number of restrictions on driving hours for drivers of goods or public transport vehicles, encapsulated in Articles 6 and 7 of the Regulation of the European Parliament and of the Council No 561/2006. The basics include the following requirements: A daily maximum driving time of 9 hours and weekly maximum of 56 hours is mandated. In addition, cumulative driving time over 2 weeks should not exceed 90 hours. Breaks should be taken after 4.5 hours, and should last no less than 45 minutes.

Driving crews crossing international borders are subject to a further Agreement, namely the European Agreement Concerning the Work of Crews of Vehicles Engaged in International Road Transport. The definitions for driving time, breaks, and rest periods as stipulated by the Agreement are similar to those defined earlier, but an additional stipulation regarding record keeping is included. This says:

. . . 6. This record (of driving) shall be entered either manually on a record sheet or printout or by use of the manual input facilities of the recording equipment. (Regulation (EC) No 561/2006 of AETR Art 6 & 7)

When records or driving time are required by law, a recording (control) device becomes necessary. Most countries in Europe use tachographs for this purpose.

United Kingdom

Drivers of public or goods vehicles in the United Kingdom are subject to the conditions laid out in the EU regulations. However, and additional requirement in the UK is the recording of time driven, even when driving within the boundaries of the UK itself. Here the primary fatigue management approach is to require Passenger Carrying and Goods Vehicles to have a control device installed. The relevant extract from UK legislation is Extract V—Part II Art 97 Installation and operation of recording equipment in vehicles, which reads as follows:

Subject to the provisions of this section, no driver shall drive a vehicle to which this Part of this Act applies unless-

There is installed in the vehicle in the prescribed place and manner equipment for recording information as to the use of the vehicle, being equipment of such type or design as may be prescribed or approved by the Minister for the purposes of this section; and

That equipment is in working order. (Transport Act 1988 Art 97)

The driver is required to be able to produce a work and rest record for the previous 28 days.

Australia

Australia is one of the more advanced countries regarding fatigue management for heavy goods vehicles. In February 2014, national legislation changed to facilitate a single regulation for the whole of Australia, namely the National Heavy Vehicle Regulator (NHVR). With this legislation, the responsibility for fatigue management of these now falls under the national authority and not the regional authority as in the preceding years. NHVR is valid for all vehicles with gross vehicle mass (GVM) over 12 tons and buses over 4.5 tons carrying more than 12 adults.

The NHVR addresses prescribed work and rest schedules, vehicle standards, maximum permissible mass and dimensions, and restraining of loads on heavy vehicles.

Fatigue is monitored by officials checking if drivers or employers are adhering to the work and rest requirements as prescribed by the legislation, depending on the type of accreditation of the driver. This is done through a requirement that each driver keep a work diary of previous shifts worked, recorded in a logbook.

A policy of fatigue management is managed nationally through the application of a policy of advanced fatigue management that applies a risk classification system (RCS) to drivers based drive/rest profiles of different drivers. This is applied to fleet drivers and heavy goods vehicle drivers only. This system works on identifying high-risk schedules by analyzing the combination of work and rest times and allocating countermeasures to mitigate the proposed risk. Each driver is required to evaluate the risk of his/her work and rest schedule by completing a standard RSC matrix.

Drivers and or employers can receive penalties and demerit points for not adhering to the requirements, failing to provide correct documentation or falsifying work and rest records. The police or enforcement officers determine the severity of the offence and the magnitude of fine; however, a maximum value is defined in the NHVR legislation itself. Drivers can receive driver's license demerits points when breaching work and rest schedules requirements.

Educational Initiatives

As with most risky driving behavior, prevention appears to be better than cure. With fatigue, however, few efforts internationally have been assessed in terms of educating drivers about the need to avoid fatigued driving; for example, publicizing the risks of driving during the circadian low periods, or even the need to take rest breaks while driving. A review of road safety campaigns across a number of countries showed very few interventions of this nature (Hashemi Nazari et al., 2017). More common are interventions aimed at addressing fatigue once it has already emerged. Interventions with fatigued drivers in Australia and Israel, for example, involved encouraging drivers to talk to passengers, to listen to the radio, open a window or wash their face. Drivers in an intervention in the Netherlands were encouraged to drink a caffeinated drink, or to stop their car and sleep for a short while. In fact, only sleep itself has proven to be a reliable antidote for fatigue, and it is unclear what effects other interventions have had—a systematic review carried out in 2017 or myriad interventions found that precise improvements in the degree of sleepiness were uncalculated and indeed uncertain.

Engineering Responses

The problem of fatigue cannot be dealt with by enforcement and education alone. Traffic engineers are important role-players in this regard and have multiple contributions to make, not least in developing new and effective ways of increasing driver alertness (surface treatments, such as rumble strips, interactive road signage, etc.), providing rest areas for fatigued drivers and implementing a policy of protective (e.g., center barriers on 2-lane roads to prevent head-on collisions) or forgiving road design (so that when crashes do occur, damage to the vehicle occupants is minimized). Vehicle designers possibly play the biggest role in both identifying fatigue in a driver and mitigating it—new vehicle technologies are emerging to prevent crashes by waking a driver who is falling asleep and even engaging emergency braking features before a crash occurs.

The biggest problem with fatigue lies in the difficulty we have to identify and operationalize it, and the consequent lack of public discourse there is about fatigue as a factor in road safety. More work is needed urgently on fatigue to increase public awareness as a first priority. While fatigue can and must be reduced, it is likely never going to be removed completely as a threat.

See Also

Roadside Safety Barriers; HUMAN FACTORS IN TRANSPORTATION; Rumble Strips, Continuous Shoulder and Centerline; Side Area Safety and Side Slopes

References

- Arnedt, J.T. et al., 2001. How do prolonged wakefulness and alcohol compare in the decrements they produce on a simulated driving task? *Accid. Anal. Prev.* 33(3), 337–344. doi: 10.1016/S0001-4575(00)00047-6.
- Curcio, G., Casagrande, M., Bertini, M., 2001. Sleepiness: evaluating and quantifying methods, in *Int. J. Psychophysiol.* 41, 251–263. doi: 10.1016/S0167-8760(01)00138-6.
- Hashemi Nazari, S.S., Moradi, A., Rahmani, K., 2017. A systematic review of the effect of various interventions on reducing fatigue and sleepiness while driving. *Chin. J. Traumatol.* 20(5), 249–258. doi: 10.1016/j.cjtee.2017.03.005.
- Lim, J., Dinges, D.F., 2008. Sleep deprivation and vigilant attention. *Ann. NY Acad. Sci.* 1129, 305–322. doi: 10.1196/annals.1417.002.
- Liu, Y.C., Wu, T.J., 2009. Fatigued drivers driving behavior and cognitive task performance: effects of road environments and road environment changes. *Saf. Sci.* 47(8), 1083–1089. doi: 10.1016/j.ssci.2008.11.009.
- Philip, P. et al., 2005. Fatigue, sleep restriction and driving performance. *Accid. Anal. Prev.* 37(3), 473–478. doi: 10.1016/j.aap.2004.07.007.
- Rupp, T.L. et al., 2007. Effects of a moderate evening alcohol dose. II: Performance. *Alcohol. Clin. Exp. Res.* 31(8), 1365–1371. doi: 10.1111/j.1530-0277.2007.00434.x.
- Schmidt, E.A. et al., 2009. Drivers misjudgement of vigilance state during prolonged monotonous daytime driving. *Accid. Anal. Prev.* 41(5), 1087–1093. doi: 10.1016/j.aap.2009.06.007.
- Singh, H., Bhatia, J.S., Kaur, J., 2011. Eye tracking based driver fatigue monitoring and warning system. *India International Conference on Power Electronics, IICPE 2010*. doi: 10.1109/IICPE.2011.5728062.
- Thiffault, P., Bergeron, J., 2003. Monotony of road environment and driver fatigue: a simulator study. *Accid. Anal. Prev.* 35(3), 381–391. doi: 10.1016/S0001-4575(02)00014-3.

Speed Governors and Limiters

Christer Hydén, Emeritus University of Lund, Lund, Sweden; Nye Sandviksveien, Bergen, Norway

© 2021 Elsevier Ltd. All rights reserved.

Speed Enforcement, Particularly Camera Enforcement	618
Speed Governor	618
Biography	621
References	621
Further Reading	621

Excessive speed is one of the main road safety problems according to many experts; and authorities in most parts of the world claim that excessive speed, that is, speed above the prevailing speed limit, is a strongly contributing factor to many crashes and—particularly—to fatal crashes and crashes with severe injuries. In the United States, the Federal Highway Administration (FHWA) states that “Speeding, traveling too fast for conditions or in excess of the posted speed limits—is a factor in almost one-third of all fatal crashes.” They continue, “Although much of the public concern about speeding has been focused on high-speed Interstates, nearly half of speeding-related fatalities occur on lower speed collector and local roads” (FHWA, 2019). In the EU, the problem is of the same magnitude: “500 people die every week on EU roads, a figure that has refused to budge for several years. And driving too fast is still the number one killer” (ETSC, 2019). And, the World Health Organization (WHO) has concluded that, in high-income countries, speed contributes to about 30% of road deaths, while in some low- and middle-income countries speed is the main factor in about half of road deaths (WHO, 2019).

The role of speed as a killer of people has been noted for almost as long as there have been vehicles on our roads. The way the problem has been tackled has been almost in vain, probably because speed is at the same time one of the most attractive characteristics of automobiles. Mobility in terms of “good speed performance” has obviously been valued so highly that controlling speed has not had any high priority at all. Therefore, even if there have been many efforts to reduce speeds, almost all of them have been without any real success. It is rather so that these efforts have the aim of demonstrating the will to do something without taking any political risks. Billions of US dollars are spent every year in order to try and convince drivers to drive slower/keep to the speed limit. This kind of information and education does not produce any disharmony to drivers. When they see the message, they can just think “yes, that sounds good, and it is good that authorities bother,” without being forced to follow it themselves. The result is discouraging regarding speed: speed campaigns have a very small effect. Meta-analysis has shown that while drinking and driving campaigns have the effect of an 18% reduction in all accidents, speed campaigns only have an effect of 4%. Besides, it seems as if the effect of the campaign goes down over time. In the 1980s, the overall effect was a 16% reduction in accidents, but it came down to 5% in 2000–10 (Høye et al., 2014). So, the conclusion regarding campaigns is that they are conditional on other speed reduction measures, such as speed enforcement and the acceptance of legal changes. On its own, the effect of campaigns in changing actual speed behavior is likely to be limited. It seems generally as if efforts have been quite small, and the problem of speeding does not seem to be solved. In Sweden, one of the countries leading the safety rankings, the conclusion is that speed development has not been positive during at least 15 years, in spite of very straightforward goals regarding speeds. The same goes for the whole of EU with its 28 countries. It is concluded that speeding remains a significant problem in most if not all European countries. On urban roads, where 37% of all EU road deaths occur, researchers found that between 35% and 75% of vehicle speed observations were higher than the legal speed. On rural non-motorway roads, where 55% of all road deaths in the EU occur, between 9% and 63% of vehicle speed observations were higher than the speed limit, while on motorways, where 8% of all road deaths occur, between 23% and 59% of observed vehicle speeds were higher than the speed limit (ETSC, 2019).

Global results indicate a similar problem. Of the 175 countries participating in the Global Status Report, 123 have road traffic laws that meet best practice for one or more key risk factors. During this review period, 10 additional countries (45 in total) have aligned with best practice on drink-driving legislation, 5 additional countries (49 in total) on motorcycle helmet use, 4 additional countries (33 in total) have aligned with best practice on the use of child restraint systems, and 3 additional countries (105 in total) on the use of seat belts. Less progress has been made on adopting best practice on speed limits, despite the importance of speed as a major cause of death and serious injury (WHO, 2018).

Speed is simply a major factor in overall road safety performance. It becomes obvious that authorities need to be more proactive and decisive regarding what types of measures are warranted. Soft measures such as education and campaigns are not at all sufficient. What is needed is hard measures, measures that are decided by decision-makers, passed into law, forcing drivers to comply. There are primarily two candidates that fulfill the criterion. They have documented effects on safety and a great potential of solving safety problems in a more general way. The problem is that they are still not implemented in a significant way, most probably because there is hesitation among politicians and other decision-makers in agencies. The two candidates are given in sections “Speed Enforcement, Particularly Camera Enforcement” and “Speed Governor.”

Speed Enforcement, Particularly Camera Enforcement

Drivers respond to a measure that will lead to fines and even harder consequences if you are not keeping the speed limit. The problem, however, is that there is still a lot of speeding, for many reasons: drivers know that there is enforcement, but those who really want to drive faster can inform themselves thanks to different apps that can provide them with information on where the enforcement is taking place. This is particularly true for camera enforcement, where the cameras are stationary. Besides, in most countries, the number of cameras is too small to be able to cover a large road network. There is an opening that should change the scope considerably and make camera enforcement much more effective. The measure is area control, that is, two cameras are linked and the average speed is calculated. This is now used in Norway, particularly in tunnels where the speed can be quite high. The longest is 22 km and forcing drivers to keep the speed limit all through the tunnel. Besides, when it becomes more common to connect adjacent cameras that are located a few kilometers or more apart, drivers will feel less safe that there is no enforcement on place wherever they are driving at any given time. An alternative to using cameras would be to use GPS and fine people, for example, \$1 per kilometer for each kilometer per hour they speed. Rather than having a low probability of getting a high fine, there would be a certainty of getting fined when speeding. An habitually speeding driver, driving 10,000 km per year with an average speed 10 km/h above the limit, would then have to pay \$100,000 in fines per year, whereas someone who was speeding by 10 km/h just once while passing a vehicle over a distance of 200 m would only have to pay \$2, and maybe the first \$100 per year would be forgiven. One drawback with such a system would be that it would seem costly to low-income individuals but affordable to upper-income people unless the fine is dependent on income. That certainly could be the case since it already is for criminal speeding in Scandinavian countries.

Speed Governor

A *speed limiter* is a governor used to limit the top speed of a vehicle. Today, modern motor vehicles most often have a speed performance capability that means a maximum speed of at least 200–250 km/h and an acceleration of 0–100 km/h in about 5–10 s. Even though, with higher demands on environmental aspects, car manufacturers are promoting their cars with arguments as “the driving pleasure” and speed and acceleration performance. They even claim that overtaking maneuvers are safer thanks to better performance. The impact of this promotion is quite big. As of 2006, total advertisement spending by the automotive sector amounted to over 20 billion dollars, 40% of which benefited the printed press (Beattie et al., 2017).

The problem is that there is no verification empirically regarding this development without any demands to assess the consequences regarding effects on speeds and safety.

In view of the slow progress regarding speed, a system that sees to it that the driver cannot drive faster than the prevailing speed limit should of course have a great theoretical potential. Figures from different countries and regions state that if everybody always kept the speed limit, then the number of fatalities would drop by anywhere between 21% and 55%. Thus, the potential is extremely high, but still the interest among drivers, authorities, and industry has been very limited. Almost all actors claim that there is no need for this “dramatic” and “non-popular” measure, because there is an ever-going development of new and more effective measures. They then refer to new both passive and active safety measures like crash-worthiness and new IT systems like autonomous emergency braking (AEB), automatic driving. These systems may have quite good effects individually; however, from a speed point of view, it does not solve the general problem with speeding. There is even the obvious risk that this kind of measure makes the driver more comfortable relying on these new systems. But, again, nobody has warranted an assessment of the present situation. How do all these new systems work from a speed point of view? Generally, research around the relation between vehicle performance—speed and acceleration—has not been given any priority neither by industry nor by governments. There are only a few studies. Those few examples that exist indicate primarily negative safety effects. In an analysis by Høyen et al. (2014) only five relevant studies were identified. Four of them indicate significant increase of accidents with increasing power.

Speed governors in private vehicles do exist already, but only in a very special setting. In Germany, the car industries themselves have agreed on a Speed Governor in their cars, set to 250 km/h. However, since Italian-made cars can go faster, Mercedes-Benz and BMW now also allow customers to get their cars without this limiter. Neither the German government nor the European Union seems to have been involved, and the outcome has not been assessed either. It is hard to believe that it would have any implications on driving and driving speeds away from German Autobahns. In March 2019, Volvo announced that it would be limiting the top speed on all of its vehicles, worldwide, to 180 km/h in a bid to reduce traffic fatalities. The new speed limit will be implemented on all model year 2021 cars. Also, European Citroën, BMW, Mercedes-Benz, Peugeot, Renault, Tesla, as well as some Ford and Nissan cars have driver-controlled speed limiters fitted or available as an optional accessory that can be set by the driver to any desired speed. The limiter can be overridden if required by pressing hard on the accelerator. So, they help drivers keep the posted speed but certainly do not guarantee that they do.

What is more relevant from a safety point of view is that true speed governors already exist in many countries in many vehicles. Authorities have introduced governors in many jurisdictions controlling the maximum speed of heavy vehicles and buses. It is not introduced in far from all countries, and the interest is not always great. Like in India, where the Transport Minister Nitin Gadkari has said that he will “soon demolish the governor system meant for commercial vehicles as limiting speed had no role in tackling the rising number of fatalities and accidents on the roads” (Pankaj Doval/TNN/September 6, 2018). Besides, these Governors that exist are often set on plus 10 km/h in relation to the highest prevailing speed limit for those vehicles. This means that it is still possible to

exceed the speed limit on all roads, especially on roads with lower speed limit than the maximum. The result is, therefore, very limited effects in total (Vaa et al., 2012) even if all vehicles are so equipped. In North America, Ontario and Quebec, both in Canada, are the only two jurisdictions to mandate speed limiter technology. It is set at 105 km/h on all trucks operating in those provinces even though the highest speed limit allowed in either province is 100 km/h for trucks as well as for cars. In most of the EU, heavy trucks have speed limiters allowing up to 90 km/h speed even on secondary roads where speed limits are much lower.

So far, every attempt to control the speed of the vehicle has stranded on the question of mandatory use. Therefore, the terms Speed Governor or Speed Limiter are abandoned. Instead, there is now one concept only, called intelligent speed adaptation (ISA), which includes all sorts of speed adaptation concepts. ISA is the general term for advanced systems in which the vehicle knows the speed limit at any given location and is capable of using that information to give feedback to the driver. ISA is a measure where the driver is assisted in keeping the speed limit by some kind of signal, auditory, visually, or haptic (resistance in the gas pedal). ISA can be operated in many different ways. The main point is that the use so far has been voluntary; you can always override the system, in one way or another. Today, ISA has been introduced in many car models, but again the driver always decides whether the system should be in operation or not.

An ISA that is a Speed Governor, that is, limiting speed to the prevailing speed limit, has never been tried by any of the car industries, as far as we know. In the 1980s and onward, there have, however, been a number of authority-driven, empirical trials of ISA with various degrees of demands regarding speed compliance, from Speed Governor type to various other solutions with different degree of voluntary use.

In France, the earliest study was that of Malaterre and Saad (1984) who tested vehicles equipped with two different ISA systems. The first system (System A) comprised a control panel placed near the steering wheel with fixed speed limit controls. Once the vehicle reached the selected speed limit, the accelerator pedal became stiffer, but this hard point could be overridden if necessary. This was the equivalent of a "mandatory" system described earlier. The second system (System B) involved a level, which the driver used to set a given driving speed, beyond which the accelerator pedal had no effect. A kickdown system enabled the vehicle to override the chosen speed for as long as the pedal stayed depressed. Releasing the pedal brought the vehicle back to the chosen speed. This was the equivalent of a voluntary system of speed control. Studies showed how subjects used the two systems, according to the posted speed limit and the chosen driving speed. A comparison between the two systems—system use was imposed and system use on a voluntary basis—showed that the correct speed was much higher in the first case, 72% versus 11% (speed limit: 60 km/h) and 96% versus 42% (speed limit: 90 km/h). The difference is striking.

In Tilburg, the Netherlands, the Dutch Ministry of Transport ran a mandatory ISA field test from 1999 to 2000. A total of 20 passenger cars and a bus were used (information collected from Loon and Duynstee, 2001). The test zone contained 30, 50, and 80 km/h speed limits.

The main conclusions of the Tilburg trial were:

- The ISA system as tested in the Netherlands is technologically feasible; minor improvements are needed.
- As expected, the ISA test in Tilburg has had a positive effect on driving behavior and speed patterns. 95 percentile speeds went from well over speed limit on 30 and 50 km/h roads without the ISA to somewhat under with the ISA.
- The ISA Tilburg test shows that great deal of public support can be gained; adequate communication, however, is essential.

In Eslöv, Sweden, 25 private people driving their own cars got them equipped with speed limiters, automatically activated when the car was inside the built-up area of the city. The trial went on for half a year, in the mid 1990s (Almqvist, 2006). The main results were that average speed for the whole test route for all drivers decreased from 43.5 km/h to 41.5 km/h. The total effect is relatively small (−4.6%), as the speeds were rather moderate in the city even before. Still it represents a theoretical reduction of fatalities by 15%–20%. More importantly, speeding was (of course) gone almost entirely for these 25 people (small technical problems led to some—but minor excessive speeds). The effect of the speed limiter can be clearly seen on major streets where speeds were considerably higher than the allowed 50 km/h, and the longitudinal speeds were strongly harmonized and kept under the speed limit in the after situation (with the speed limiter). Many drivers not having speed limiters also had to slow down.

Behavioral observations showed an improvement in behavior in interactions with other road users, for example, drivers more frequently giving the right of way to pedestrians, cyclists, and other vehicles. The test drivers stated that the Speed Limiter Function should be mandatory. Moreover, it was possible to imagine a mandatory speed assisting system (Speed Governor) for built-up areas.

In 2002, the French Ministry of Transport launched a significant program of experimentation and evaluation in order to better appreciate the effects of the ISA system in terms of acceptance by the drivers and influence on their driving behavior. The two French automotive manufacturers, PSA and Renault, participated with cars in the project (Driscoll et al., 2007). The potential safety benefits of the different versions, based on observed driving speeds, distributions of crash severity, and crash injury risk. Results are given for car frontal and side impacts that together represent 80% of all serious and fatal injuries in France. Of the three system modes tested (advisory, driver select, and mandatory), the results suggest that driver select would most significantly reduce serious injuries and death. The estimate is that 100% utilization of cars equipped with this type of speed adaptation system would decrease injury rates by 6%–16% over existing conditions depending on the type of crash (frontal or side) and road environment considered.

In Leeds, the United Kingdom, an on-road study was carried out, as part of an External Vehicle Speed Control (EVSC) project. A vehicle was equipped with two versions of EVSC, Driver Select and Mandatory. The main findings were:

- The Mandatory system trial in this experiment successfully reduced speeds, particularly in areas where drivers are renowned to being poor at adapting their speed.

- There were no negative behavioral compensation effects. Although drivers were initially unfavorable toward the mandatory system, they reported that driving with the system was safer due to enhanced awareness of potential hazards.
- The effectiveness of the Driver Select system is roughly half that of the mandatory one (Carsten and Fowkes, 2000).
- Prediction based on actual speed changes observed in an ISA project in the United Kingdom is that an ISA system targeted purely at compliance with existing speed limits could, in its strongest variant (i.e., a non-overrideable version), deliver a 29% reduction in injury accidents (Carsten, 2012). Applying the power model of Elvik (2014) translates into a 50% reduction in fatal accidents.

The UK ISA project produced a rich database with high-resolution data on driver behavior covering a comprehensive range of road environment. The field trials provided vital information on driver behavior in the presence of ISA. ISA is predicted to save up to 33% of accidents on urban roads, and to reduce CO₂ emissions by up to 5.8% on 70 mph (112 km/h) roads. In order to investigate the long-term impacts of ISA, two hypothetical deployment scenarios were envisaged covering a 60-year appraisal period. The results indicate that ISA could deliver a very healthy benefit-to-cost ratio, ranging from 3.4 to 7.4, depending on the deployment scenarios. Under both deployment scenarios, ISA has recovered its implementation costs in less than 15 years. It can be concluded that implementation of ISA is clearly justified from a social cost and benefit perspective. Of the two deployment scenarios, the market-driven one is substantially outperformed by the authority-driven one. The benefits of ISA on fuel saving and emission reduction are real but not substantial, in comparison with the benefits on accident reduction; up to 98% of benefits are attributable to accident savings. Indeed, ISA is predicted to lead to a savings of 30% in fatal crashes and 25% in serious crashes over the 60-year period modeled (Carsten, 2012).

The results of all the studies presented clearly indicate that there is a great potential of mandating speed limiters from a behavioral and safety point of view. All trials but the French Lavia project have clearly demonstrated the benefits of an “authority select” mode. And the predicted effect of a reduction of 50% of fatalities—as in the UK project—is verified by the EU-funded and SRA coordinated project PROSPER. The PROSPER project calculated crash reductions for six countries. Reductions in fatalities between 19% and 28%, depending on the country, were predicted in a market-driven scenario. Even higher reductions were predicted for a regulated scenario—between 26% and 50%. Benefits are generally larger on urban roads and are also larger if more intervening forms of ISA are applied (EC.Europa.eu, 2019).

The PROSPER also concluded that benefit-to-cost ratios ranging from 2.0 to 3.5 and from 3.5 to 4.8 were calculated for the two scenarios: market-driven and regulation-driven.

So, the general conclusion is that ISA is demonstrating quite positive results regarding behavior changes, safety changes, benefit-to-cost ratios, and even on climate aspects, even though small in the latter case. The effect of the “authority select” mode (= Speed Governor) is considerably larger in all cases but one. Besides, any possible negative aspects seem to be to a large degree nonexistent in the reported trials.

The main conclusion of these reported trials is that they were ended long ago, and no logical continuation has been following. There has been little effort to follow up the results. There are two lines to follow:

1. Research on how to proceed in order to obtain the theoretical effects of a voluntary—market-driven—concept.
2. How will drivers react once they realize that they cannot go above the speed limit? Behavioral effects, safety outcome? Other effects?

The first line will work as a preparation for line (2). What effort is demanded in order to reach behavioral changes without a mandatory use of ISA?

A breakthrough to get to this is a proposal from the ETSC to mandate speed limiters on all new vehicles sold in Europe has passed a vote by Members of European Parliament in February 2019. If approved for implementation, the technologies would be mandatory for new (not yet designed) models as of May 2022. Existing models (existing cars in production with no significant changes since before 2020) would have to comply by 2024 (NewsGlobe, 2019).

It will use recognition software to read speed limit signs, and, where none are present, GPS data. Drivers will reportedly be able to temporarily override the limit, which will continue to allow highway overtaking. Drivers will receive an audiovisual warning regarding their speed. “If the driver continues to drive above the speed limit for several seconds, the system should sound a warning for a few seconds and display a visual warning until the vehicle is operating at or below the speed limit again,” states the ETSC. The system may allow overriding it too easily to be effective but it certainly is a big step toward mandatory speed governors assuming lobbying does not stop implementation of this.

So far, research progress regarding ISA has been very slow. There are still no major breakthroughs, in spite of 30 years of research on ISA. And in spite of the fact that the speeds can be reduced significantly, and thereby fatalities/injured. Besides, users have been found to have generally positive attitudes and some sections of the public have been shown to be willing to purchase ISA systems (Carsten, 2012). From a survey of a representative sample of British drivers, we learn that “While a non-negligible part of the sample population has such strong opposition to ISA that no reasonable discounts or incentives would lead to them buying or accepting such a system, there is also a large part of the population that, if given the right incentives, would be willing or even keen to equip their vehicle with an ISA device” (Chorlton et al., 2012).

It is important to add to this picture that the car industry does a lot for safety. They have been, and are, extremely successful in improving the crashworthiness of cars. Much of this work has been enhanced thanks to the NCAP program where the industry has been encouraged—and, in practice, forced—to improve the performance. Lately Euro NCAP has introduced speed assistance that could be one important accelerating factor for the implementation of ISA. However, Euro NCAP only promotes installation of

voluntary systems. They assess different functions of Speed Assist systems, informing, warning, or actively preventing. Besides, Speed Assist is one of five systems under the heading Safety Assist. The other four are “Electronic Stability Control,” “Seat-Belt Reminders,” “AEB Interurban,” and “Lane Support.” This means that Speed Assistance only contributes to one fifth of the Safety Assist score. The Euro NCAP approach is of course encouraging. However, the assessment does not guarantee any actual change of speeds. The other included safety systems: AEB, electronic stability control, etc., are automatic systems that work once they are installed in the car. Speed Assist does not work automatically, but presuppose an active step by the driver, that is, to give high priority to a safety device. However, this is what normally not happens. In line with all other systems that are assessed by NCAP, with crash tests, etc., Speed Assist should be assessed based on the actual speed effect, not a fictive one.

So, the conclusion regarding ISA is that there still seems to be great unwillingness among industry and authorities to include Speed Governor on their agenda. It is obvious that the car industry has managed to set the scene, so as to exclude all discussions—and research—on the most promising measure that could prevent all speeding. But, as technical progress is made almost every day, it is going further toward concepts more and more similar to Speed Governor concepts. This is of great importance. A governor like this may change the world of driving entirely.

Biography

Christer Hydén has been active in the traffic safety field for around 40 years. His first achievement was the development of a Traffic Conflict Technique, a way of assessing risk on the roads with the help of near accidents. He was the President of ICTCT (at first standing for “International Co-operation in Traffic Conflict Techniques” and later for “International Co-operation on Theories and Concepts in Traffic Safety”) from 1977 to 2011. Another field of expertise is the impact of speed, and the effect on speed as well as safety of different measures, like traffic calming and speed limiters in vehicles. He is also a member of the Independent Council for Road Safety International (ICORSI).

References

- Almqvist, S., 2006. Loyal Speed Adaptation Speed Limitation by Means of an Active Accelerator and its Possible Impacts in Built-Up Areas. Lund University, Lund.
- Beattie, G., Durante, R., Knight, B., Sen, A., 2017. Advertising Spending and Media Bias: Evidence from News Coverage of Car Safety Recalls. National Bureau of Economic Research, Cambridge, MA.
- Carsten, O., 2012. Is intelligent speed adaptation ready for deployment? *Acc. Anal. Prev.* 48, 1–3.
- Carsten, O., Fowkes, M., 2000. External Vehicle Speed Control: Executive Summary of Project Results. Institute for Transport Studies, University of Leeds, Leeds.
- Chorlton, K., Hess, S., Jamson, S., Wardman, M., 2012. Deal or no deal: can incentives encourage widespread adoption of intelligent speed adaptation devices? *Acc. Anal. Prev.* 50, 282–288.
- Driscoll, R., Page, Y., Lassarre, S., Ehrlich, J., 2007. Lavia—an evaluation of the potential safety benefits of the French Intelligent Speed Adaptation project. *Annu. Proc. Assoc. Adv. Automot. Med.* 51, 485–505.
- EC.Europa.eu, 2019. Available from: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/projects/prosper.pdf.
- Elvik, R., 2014. Speed and Road Safety—New Models. Institute of Transport Economics, Oslo.
- ETSC, 2019. In-vehicle technology vital to tackling speeding in Europe. European Transport Safety Council. Available from: <https://etsc.eu/in-vehicle-technology-vital-to-tackling-speeding-in-europe/>.
- FHWA, 2019. Speed management safety. Available from: <https://safety.fhwa.dot.gov/speedmgt/>.
- Høye, A., Elvik, R., Vaa, T., Sørensen, M.W.J., Amundsen, A.H., Akhtar, J., et al., 2014. Trafikksikkerhetshåndboken (Traffic Safety Handbook) Transport Economics Institute, Oslo.
- Loon, A.V., Duynstee, L., 2001. Intelligent Speed Adaptation (ISA): A Successful Test in The Netherlands. Transport Research Centre (AVV), Ministry of Transport, The Hague.
- Malaterre, G., Saad, F., 1984. Contribution à l'analyse du contrôle de la vitesse par le conducteur: Evaluation de deux limiteurs. Cahier d'Etude n° 1984; 62 ONSER [Google Scholar].
- NewsGlobe, 2019. Available from: <https://thenewsglobe.net/eu-wants-speed-governors-data-recorders-in-cars-for-2022/>.
- Vaa, T., Assum, T., Elvik, R., 2012. Driver Support Systems: Estimating Road Safety Effects at Varying Levels of Implementation (No. 8248013340). Institute of Transport Economics (TØI), Oslo.
- WHO, 2018. Global status report. Global Status Report on Road Safety 2018. World Health Organization, Geneva.
- WHO, 2019. FACT SHEET #1 road safety: Basic facts. Available from: https://www.who.int/violence_injury_prevention/publications/road_traffic/Road_safety_media_brief_full_document.pdf.

Further Reading

- Ehrlich, J., Marchi, M., Jarri, P., Salesse, L., 2003. LAVIA, the French ISA project: main issues and first results on technical tests. Proceedings of ITS World Congress 2003. Madrid, Spain.
- ETSC, 2013. Intelligent Speed Assistance—Frequently Asked Questions. European Transport Safety Council, Brussels.

Speed Limits on Rural Highways

Peter Tarmo Savolainen, Timothy Jordan Gates, Michigan State University, East Lansing, MI, United States

© 2021 Elsevier Ltd. All rights reserved.

Glossary	622
Introduction	622
The Relationship Between Speed and Safety	623
History of Speed Limit Policy in the United States	625
Differential versus Uniform Limits for Trucks and Buses	627
Effects of Speed Limits on Actual Speeds	628
Two-Lane Rural Highways and Speed Limits	629
Current Practices in Setting Speed Limits	630
References	630

Glossary

85th percentile speed The speed at or below which 85 of vehicles travel. Speed limits are often established such that they are within 5 mph or 8 km/h of the 85th-percentile speed of free-flowing traffic.

advisory speed Advisory speeds warn drivers to proceed at a speed lower than the speed limit due to geometric, weather, or other conditions.

design speed The design speed is a selected speed used to determine the various geometric design features of the roadway.

differential speed limit A system that prescribes different maximum speed limits for different vehicle types or user groups. This is usually applied as one maximum speed limit for light passenger vehicles, and a lower maximum speed limit for trucks and heavy commercial vehicles.

free-flow conditions Free-flow conditions are characterized by drivers being free to drive at their desired speed and not constrained by the presence of other vehicles or downstream traffic control devices.

operating speed The speed at which vehicles are observed operating during free flow conditions. The 85th percentile of the distribution of observed speeds is the most frequently used measure of the operating speed.

speed limit, posted The posted speed limit is the maximum lawful vehicle speed for a particular location as conveyed to the motorist on a regulatory sign. This is a black-on-white rectangular sign in the United States and a black-on-white or yellow round sign with a red border in countries that have adopted the Vienna Convention. Standard engineering practice is to post speed limits where speed zoning has altered the limit from the statutory value.

speed limits, statutory Numerical speed limits specifically provided for under a country's or state's traffic codes that apply to various classes or categories of roads (e.g., interstates, rural highways, residential streets, etc.) in the absence of posted speed limits.

speed zoning Speed zoning is the process of performing an engineering study and establishing a reasonable and safe speed limit for a section of roadway where the statutory speed limits given in the motor vehicle laws do not fit the road or traffic conditions at a specific location.

speeding The legal definition of speeding is exceeding the posted speed limit, or a statutory speed limit. In the road safety community, speeding is defined as exceeding either of these limits or driving at a speed too fast for conditions.

Introduction

Speed limits inform drivers of the highest speed that is considered safe and reasonable for ideal traffic, road, and weather conditions. As they are established by legislative action, speed limits also provide a basis for the enforcement of unreasonably high travel speeds. Given the political nature of the subject, speed limits have been a longstanding concern of transportation agencies and the subject of numerous research studies.

The extant literature has generally shown that speed limit increases are associated with significant increases in traffic fatalities. The fact that fatalities increase with speed limits is unsurprising given the physics involved, which show the kinetic energy involved in a crash increases with speed. Despite this fact, 25 US states increased their maximum statutory speed limits on rural interstates between 2001 and 2018 as shown in Fig. 1. As of October 2019, there are 18 US states that have a maximum statutory speed limit of 75 mph (120 km/h) or above, including eight states with limits of 80 mph (129 km/h) or above. These and other historical increases have been supported by the fact that existing operating speeds tend to exceed the posted speed limit, often by significant amounts, on

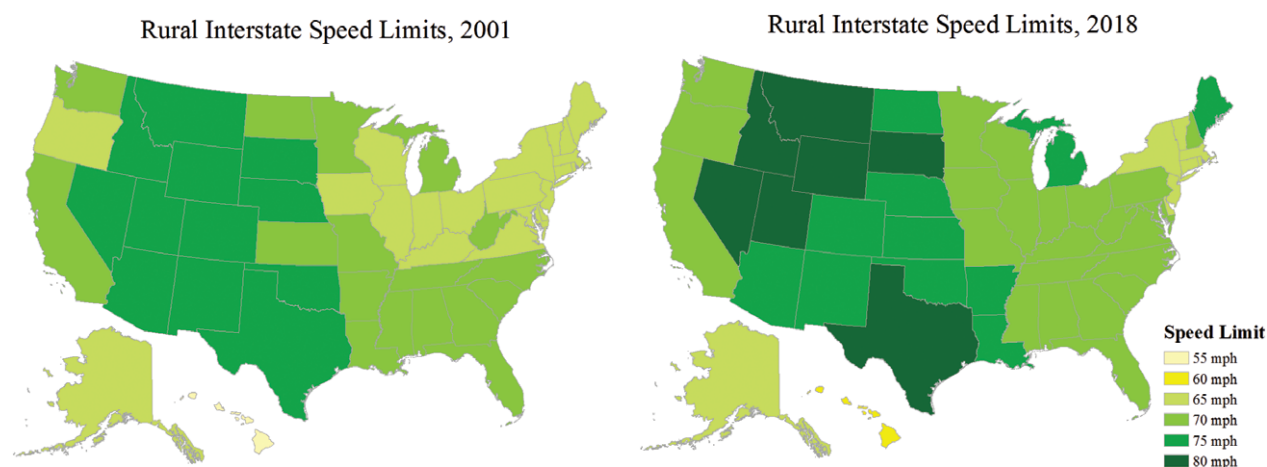


Figure 1 Maximum rural interstate speed limits by state in 2001 and 2018.

high-speed roadways (Fitzpatrick et al., 2003). The maximum speed limit on freeways in most other industrialized countries varies between 110 km/h and 130 km/h (SafetyNet, 2009).

Statutory limits are generally applicable for a particular class of highways with specified design, functional, jurisdictional, and/or location characteristics (FHWA, 2012). These limits are generally higher in driving contexts that are less complex and challenging to navigate, as well as where there is an expectation of higher travel speeds based on the contextual environment, such as rural highways.

Speed limits are typically established in consideration of the design speed of the road, which influences various geometric design features such as minimum stopping sight distance, minimum horizontal curve radius, and maximum grade (AASHTO, 2018). The American Association of State Highway and Transportation Officials (AASHTO) recommends using above-minimum criteria where practical (AASHTO, 2018) and, ideally, the statutory speed limit should be set at or below the highway's prevailing design speed. In most countries, speed limits are set using different criteria, including mobility and environmental considerations, though safety is typically the overarching criterion. In some countries, such as Finland (Kallberg and Toivanen, 1998) and Sweden (Lynam et al., 2005), there is an effort to set speed limits such that road-user or societal costs are minimized. Higher speeds result in lower cost for travel time, but higher costs for safety, gasoline consumption, vehicle maintenance, etc. As such, the speed limit should be set close to the minimum total cost. Environmental costs can be part of such a cost analysis, and from January 2020, the top speed of 130 km/h (81 mph) on Dutch freeways was reduced to 100 km/h (62 mph) between 6 a.m. and 7 p.m. in an attempt to cut emissions (Spiegel, 2020). Similar efforts have been discussed in other countries, including Germany (The Local, 2019) where much of the Autobahn operates without a federally mandated speed limit.

Some sections have geometric features with design speeds that are lower than the statutory speed limit, with a typical example being horizontal curves with below-minimum radii. Such curves introduce the need for advisory speeds, informing drivers that the inferred design speed of these features is less than the posted speed limit. It is important to note that these speeds, as the name implies, are advisory in nature and non-compliance with these advisories is typically not enforceable. In contrast, on extended sections of highways with speeds below the statutory limit, referred to as speed zones or altered speed zones (FHWA, 2012), a lower posted speed limit can be introduced.

In addition to design speed, the operating speed of the roadway has historically been a predominant factor in establishing the posted speed limit, and still is in at least parts of the United States. Consideration of operating speed assumes that most drivers are capable of selecting appropriate speeds according to traffic, road, and weather conditions and, also, that these drivers will operate at reasonable and prudent speeds. To this end, the Manual on Uniform Traffic Control Devices (MUTCD) states that speed zones shall only be established on the basis of an engineering study, which shall include an analysis of the current speed distribution of free-flowing vehicles. The MUTCD also states that "when a speed limit within a speed zone is posted, it should be within 5 mph of the 85th percentile speed of free-flowing traffic" (FHWA, 2012). The 85th percentile speed is that speed at which 85 percent of all drivers travel at or below under ideal, free-flow traffic conditions.

The Relationship Between Speed and Safety

The 85th percentile speed has been referenced directly in the MUTCD since the 1971 edition (FHWA, 1971) while its use as a basis for speed limit determination was used as early as the 1940s (TRB, 1998). The 85th percentile speed emerged as a determinant in speed limit selection based upon one of the earliest and most frequently referenced studies in the area of speed-safety research, which compared the estimated speeds of crash-involved vehicles with field-measured speeds from control vehicles on the same general day-of-week and time-of-day (Solomon, 1964). The results, illustrated in Fig. 2, present the crash involvement rate (per 100 million

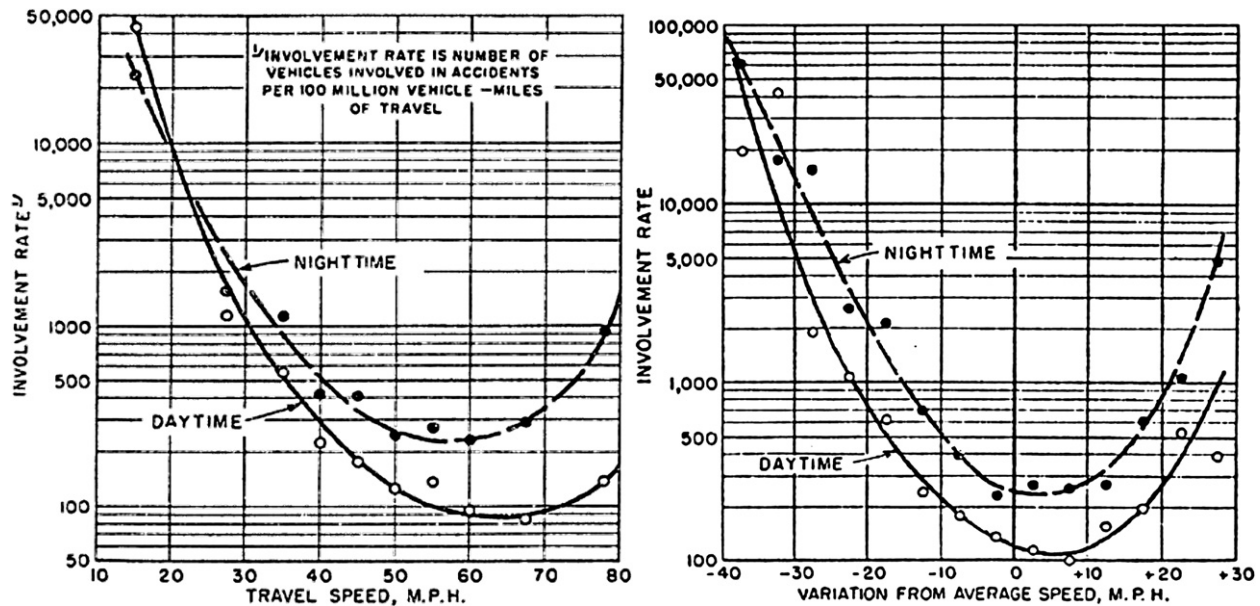


Figure 2 Crash rates by travel speed and variation from average speed (Solomon, 1964).

vehicle-miles of travel) with respect to travel speed (Fig. 2A) and with respect to variation from the average speed of traffic under similar conditions (Fig. 2B).

Collectively, these figures suggest that crash risk (i.e., possibility of being in a crash) is greatest at very low speeds and very high speeds, with vehicles traveling approximately 6 mph (10 kph) above the average speed exhibiting the lowest crash rates. However, subsequent research has shown that the crash risks at lower speeds are significantly overstated as 44% of these crashes involved low-speed maneuvers (e.g., vehicles turning into or out of traffic) and an analysis of the data excluding these maneuvers demonstrated crash risks were much less pronounced at low speeds (West and Dunn, 1971). A series of additional studies in Australia (Fildes et al., 1991), Switzerland (Finch et al., 1994), and the United Kingdom (Maycock et al., 1999; Quimby et al., 1999; Taylor et al., 2000) showed individual crash risk to consistently increase with increases in speed.

From a speed limit policy perspective, these trends suggest that drivers traveling at very high speeds are responsible for a disproportionate share of traffic crashes. This explains, in part, why setting the speed limit near the 85th percentile speed has become common practice. However, reliance on the 85th percentile speed as the primary factor in speed limit determination has come under criticism, with a National Transportation Safety Board (NTSB) report concluding that there is not strong evidence that the 85th percentile speed equates to the speed with the lowest crash involvement rate on all road types (NTSB, 2017). Furthermore, increasing speed limits to align with the 85th percentile speed may lead to higher operating speeds, resulting in a cycle of speeds escalation and reduce safety (Donnell et al., 2009).

The same NTSB report concludes that speed increases both the likelihood of being involved in a crash, as well as the severity of crash-involved occupants (NTSB, 2017). The relationship between speed and severity is generally straightforward, with increases in speeds resulting in more severe injuries by crash-involved occupants (ITF, 2018). However, the risk of crash involvement is very complex and influenced by various characteristics of both the driver and roadway environment.

A Swiss study compared changes in crash rates with respect to changes in average speeds using data from Denmark, Finland, Switzerland, and the United States (Finch et al., 1994) and showed fatal crashes to decrease by 12% when speed limits were lowered from 130 km/h (81 mph) to 120 km/h (75 mph). In aggregate, data from all four countries showed crash rates to increase consistently with speed, as illustrated in Fig. 3.

To this end, several meta-analyses have been conducted to provide insights as to the general relationship between mean speed and crash risk. A common form for this relationship is a power model, which relates the number of crashes after a speed limit change to the number of crashes before the change as a function of the ratio of mean speeds between these periods (Nilsson, 2004). The model takes the following general form:

$$\text{Crashes}_{\text{after}} = \text{Crashes}_{\text{before}} \cdot \left(\frac{\text{Speed}_{\text{after}}}{\text{Speed}_{\text{before}}} \right)^{\text{exponent}}$$

In this basic form, the exponent reflects the rate at which crashes change with respect to a relative change in speed. Data from 115 studies, which included 526 estimates of the relationship between changes in mean speed and the number of crashes or crash-involved injuries, show crash risk to consistently increase as speed increases (Elvik, 2013). Empirical results utilizing the power

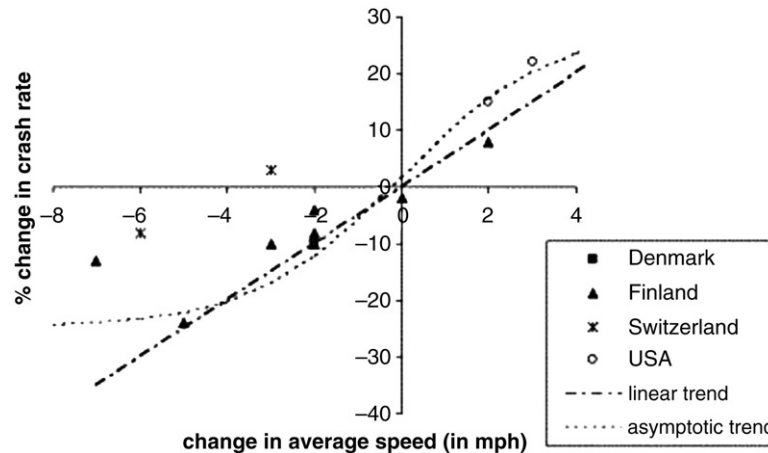


Figure 3 Change in crash rate with respect to average speed (Finch et al., 1994).

model suggest that a 1% increase in average speed results in increases in the average frequencies of injury, severe injury, and fatal injury crashes of 2%, 3%, and 4%, respectively (ITF, 2018).

Subsequent research has supported this general relationship (Castillo-Manzano et al., 2019), though an exponential function has been shown to provide improved fit as compared to the power model as shown in Fig. 4. This result implies that the impacts of the speed differential are greater at high initial speeds as compared to low initial speeds, which is particularly important, as these higher speeds are generally reflective of travel on rural highways. This summary is based upon aggregate-level data from various studies, though a subsequent meta-analysis has shown a similar relationship between speed changes and crash risk when considering individual drivers (Elvik et al., 2019).

A subset of the data from Fig. 3 were used as a part of the meta-analysis that investigated the impacts of speed limit changes on both mean speed and crash risk (Musicant et al., 2016). When speed limits were increased, the mean speeds increased, but to a lesser degree than the actual increase in limits. When speed limits were reduced, the reduction in mean speeds also tended to be inelastic as compared to the change in limits. However, the average impacts on traffic crashes and injuries was shown to be nearly proportional to the change in speed limits.

History of Speed Limit Policy in the United States

The preceding discussion demonstrates that the frequency and, particularly, the severity of crashes are both impacted by changes in travel speeds. These findings are further reinforced by research that was conducted in response to three major legislative actions that have influenced speed limit policies across the United States:

1. The Emergency Highway Energy Conservation Act of 1974 mandated a National Maximum Speed Law (NMSL) of 55 mph. The primary motivation of this legislative action was to reduce fuel consumption in response to the embargo put in place by the Organization of Petroleum Exporting Countries (OPEC). However, the NMSL was extended, in part, due to a reduction in traffic fatalities that occurred during this same period. An analysis of crash data from 1952 to 1979 estimated that traffic fatalities decreased by 7466 per year as a result of the speed limit reduction (Forester et al., 1984).
2. One issue that arose with the introduction of the NMSL was that observed driving speeds did not necessarily reflect the new lower speed limits. This was particularly true on rural interstate highways where posted speed limits were significantly below the design speeds of these roadways. In 1987, the Surface Transportation and Uniform Relocation Assistance Act (STURAA) permitted states to increase speed limits from 55 to as high as 65 mph on interstate highways in areas with populations of less than 50,000. Following the enactment of the STURAA, a series of evaluation studies showed increases in traffic crashes and/or fatalities in states where the speed limit had been increased. The average increase in fatality risk during this period was estimated at 19% when controlling for other pertinent factors (Baum et al., 1991).
3. In 1995, the National Highway System Designation Act (NHSDA) completely repealed the NMSL and gave states full autonomy to establish interstate speed limits at their discretion. As a result of this legislation, many states raised interstate speed limits to 70 mph or more, with the state of Montana introducing a “reasonable and prudent” speed limit law from 1996 to 1999. The repeal of the NMSL led to a series of additional studies, which estimate an average increase of 17 percent in traffic fatalities following these increases (Farmer et al., 1999). A follow-up study concluded that the increase in speed limits accounted for approximately 12,545 additional fatalities from 1990 to 2005 (Friedman et al., 2009).

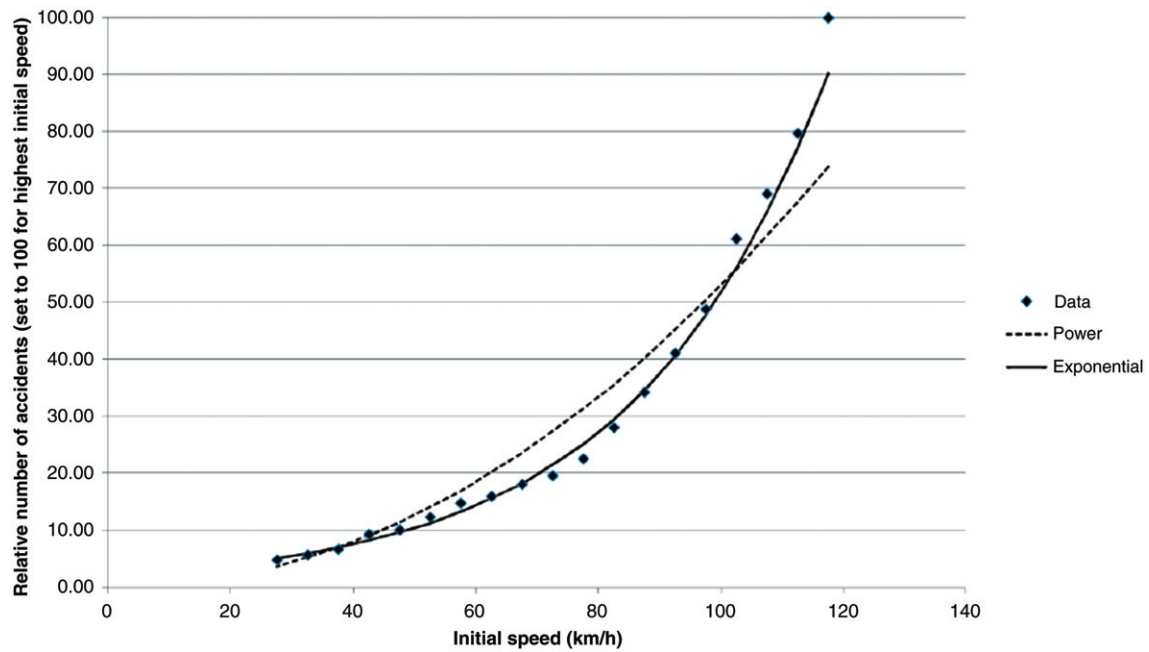


Figure 4 Comparison of power model and exponential function for injury crashes as a function of mean speed (Elvik, 2013).



Figure 5 Annual rural interstate fatality rates by maximum speed limit (Davis et al., 2015).

Collectively, the research literature shows that fatalities on rural interstates are consistently higher among those states with higher maximum statutory speed limits. However, research has also shown that the safety performance of these facilities to have improved

over time regardless of the maximum limit. Fig. 5 illustrates the differences in fatality rates with respect to the maximum statutory limit from 1999 through 2011 (Davis et al., 2015). At the aggregate level, these data show that fatality rates were markedly higher in states with 70 mph rural interstate speed limits and, particularly, in those states with speeds limits of 75 or 80 mph, as compared to states with 65-mph maximum limits. Over this 13-year period, states in each category experienced similar trends, with fatality rates remaining relatively stable through 2005, followed by a period of significant decreases, which is reflective of overall fatality trends in the United States during this same period.

Differential versus Uniform Limits for Trucks and Buses

Historically, the speed limits for trucks and buses have been a particular concern given their large size and weight, which results in restricted maneuverability, longer stopping distances, and higher impact forces in a collision. These concerns have led to the introduction of differential speed limits (DSLs), where the maximum speed limits for trucks and buses is set lower than the limit for passenger cars. In the early 1960s, the majority of states posted lower speed limits for trucks than for cars, especially on high-speed facilities. Most of the truck limits were 5–10 mph (8.0 to 16 km/h) lower than passenger car limits; however, in some states the differential was 15 mph (24 km/h). In most European countries, medium and heavy trucks are limited to 80 or 90 km/h (50 or 56 mph) on freeways whereas passenger cars typically are allowed to travel at 120 km/h (75 mph) or faster on freeways in almost all European countries (SafetyNet, 2009).

While lower speed limits for larger vehicles help to mitigate concerns with respect to high impact forces in truck- and bus-involved collisions, these differential limits potentially increase the variability in travel speeds and may increase the potential for truck-involved crashes. In light of this concern, numerous states subsequently transitioned to a uniform speed limit (USL), which establishes one maximum speed limit for all vehicles (English and Levin, 1978). On the other hand, Germany has no speed limit for passenger cars on a substantial portion of its freeways—and many drivers travel well above 130 km/h (81 mph)—whereas trucks are limited to 80 km/h (50 mph). In spite of this, German freeways have a lower overall fatality rate (0.25 per hundred million vehicle miles (HMVM) in 2017) than US freeways (0.79 for rural and 0.48 for urban segments in 2017). So, limiting the speeds of heavy vehicles may have net safety effects even if it leads to higher variation in speeds. It should also be noted that freeways in Germany often are congested, so average speeds are much lower than top speeds. And, European road safety in general is higher than that in the United States. So, the fact that Germany has no speed limit on some of its freeways should not be seen as an endorsement of that if we want safe roads. The United Kingdom, which has among the lowest top speed limits of any European country (70 mph or 110 km/h), had a fatality rate on freeways that was half of the German, one or 0.80 per billion vehicle km traveled, just below 0.13 per HMVM (Bundesanstalt für Straßenwesen, 2019).

Following enactment of the STURAA, 12 of the 40 US states that raised the maximum speed limit retained lower limits for trucks than for cars. A study by the National Highway Traffic Safety Administration (NHTSA) (1988) study in found few fatal crashes on rural interstates involving passenger cars and tractor-trailers and an effect of differential limits could not be determined. While there

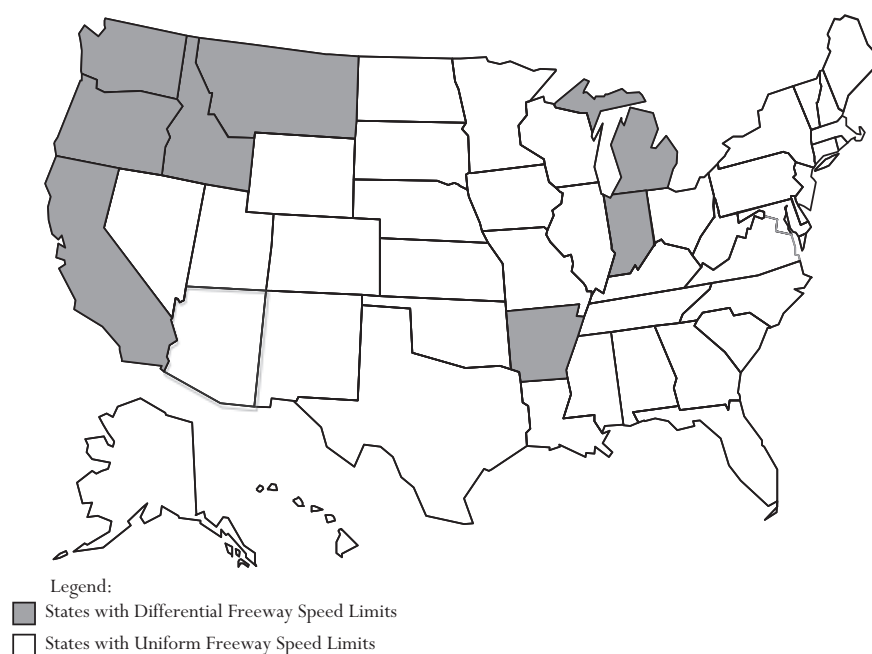


Figure 6 States with differential and nondifferential freeway speed limits.

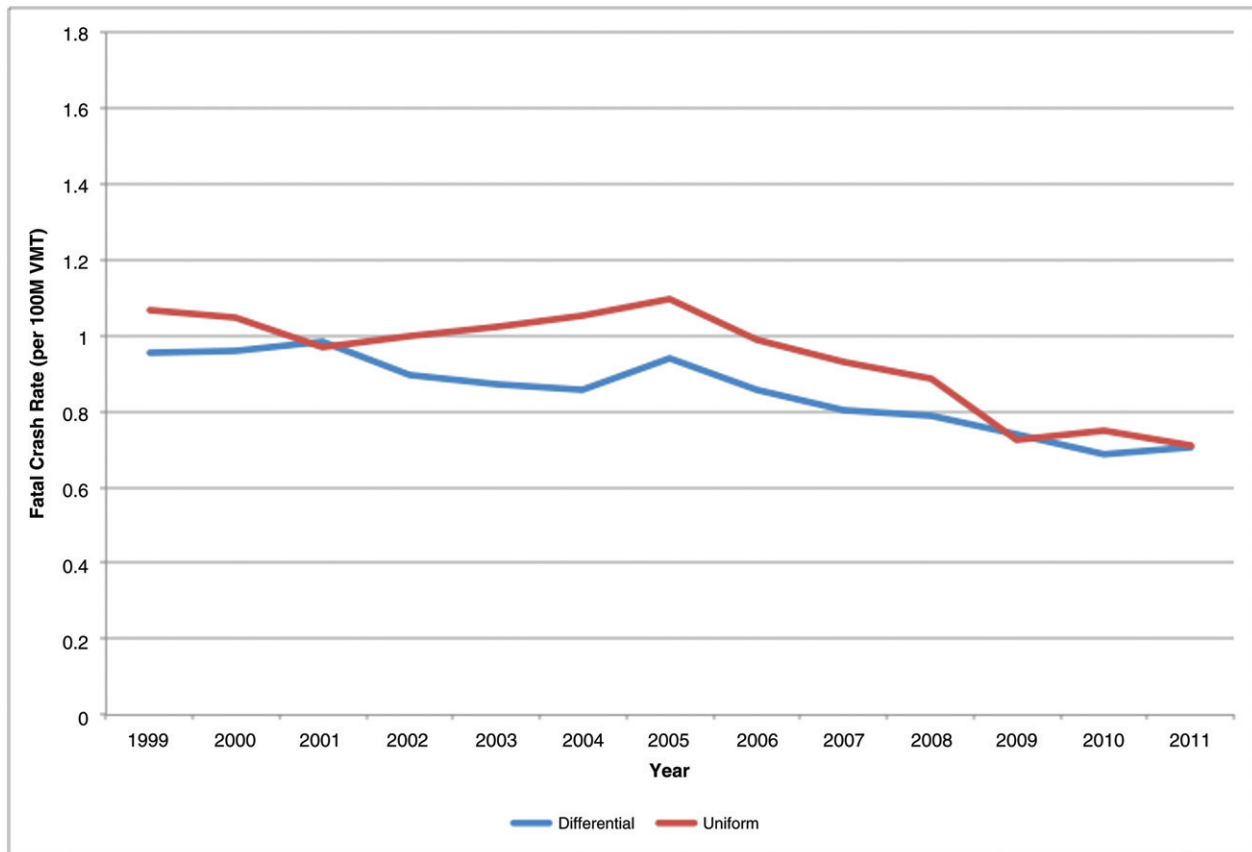


Figure 7 Annual rural interstate fatality rates by truck/bus speed limit policy (Davis et al., 2015).

has been extensive research comparing DSLs and USLs, a strong consensus has not emerged. A summary of the speed differences and crash impacts of differential speed limits for trucks on freeways was provided in *TRB Special Report 254*, which concluded, “a strong case cannot be made on empirical ground in support of or opposition to differential speed limits” (TRB, 1998). Subsequently, additional states have transitioned from differential to uniform laws and, as of October 2019, there were eight states with DSLs in place as shown in Fig. 6. On the other hand, all 28 European Union countries currently (January 2020) have DSLs.

Fig. 7 presents summary data, which compares the fatality rates between US states with uniform and differential freeway speed limits for trucks and buses from 1999 through 2011. States with differential limits exhibited slightly lower fatality rates, particularly from 2002 to 2008. However, the most recent data do not indicate a clear difference between these groups.

Effects of Speed Limits on Actual Speeds

In addition to the safety impacts of speed limits, another area of substantive debate is how speed limits influence the actual speed selection behavior of drivers. According to AASHTO (2018), driving speeds are affected by the physical characteristics of the road, weather, other vehicles, and the speed limit. Among these, road design is a principal determinant of driving speeds. Geometric factors tend to have particularly pronounced impacts on crashes. Ultimately, many factors affect speed selection beyond just road geometry and posted limit and research has generally shown that speed limit changes result in changes in the observed mean and 85th percentile speeds that are less pronounced than the actual speed limit changes (Freedman and Esterlitz, 1990).

This has been true for cases where speed limits were increased or decreased. One of the most extensive studies in this area investigated the impacts that raising or lowering posted speed limits on non-limited access highways had on driver behavior (Parker, 1997). Speeds limits were adjusted by as much as 20 mph (32.2 km/h) and, interestingly, the difference in speed after these changes was less than 1.5 mph (2.4 km/h) on average. This is one of several studies demonstrating that drivers select their speeds on non-limited access highways primarily on the basis of roadway geometry and traffic characteristics rather than the posted speed limits per se. A meta-analysis of research from the United States and various European countries showed similar results (Wilmot and Khanal, 1999).

Feedback from drivers reinforces this point. A national survey conducted in the United States in 2003 gathered information regarding driver attitudes and behaviors related to violating the speed limit and other unsafe driving behaviors (Royal, 2003). Results

showed that most drivers believe they can drive 7 to 8 mph (11.3–12.9 km/h) above the posted speed limit before being pulled over. While 51% of drivers admitted to driving 10 mph (16.1 km/h) over the posted speed limit, 68% felt that other drivers violating the speed limit were a danger to their own personal safety. Drivers reported that the most influential factors dictating their speed selection were weather, their perception of what speeds were “safe,” the posted speed limit, traffic volume levels, and the amount of personal driving experience they had on a particular road.

Two-Lane Rural Highways and Speed Limits

The discussion earlier has focused on speed limits on limited-access freeways, also called motorways. These are the types of roads with the highest speed limits, and also the highest volumes of traffic. For example, at least 27 US states allow for higher posted speed limits on freeways as compared to non-freeways while the remaining states provide the same maximum speed limit for both roadway types (Gates et al., 2015). These higher speeds are desirable since freeways are also the safest when considering crash risks per distance traveled. The reason for this is that freeways are designed to higher standards. They are divided by a median and typically include grade-separated interchanges as opposed to stop-controlled intersections. These differences virtually eliminate certain types of crashes, including head-on and angle collisions. In the United States, freeways carry 25% of all vehicle miles traveled (VMT), and 26% of rural VMT (FHWA, 2020).

However, the most recently available traffic fatality data (NHTSA, 2020) show that rural interstates experienced 1,685 fatalities whereas rural non-interstate facilities saw 13,053 fatalities. The latter group includes rural two-lane highways where safety is obviously also of great importance as such roads experience a disproportionate number of head-on collisions, as well as run-off-road crashes into roadside areas with hazardous fixed objects such as trees, utility poles, etc. Speeds, and speed limits, on such roads are obviously equally important. The majority of US states operate with 55 mph (89 km/h) or 65 mph (105 km/h) maximum speed limits on two-lane highways, though several states post limits as high as 70 or 75 mph (113 or 121 km/h) (Gates et al., 2015).

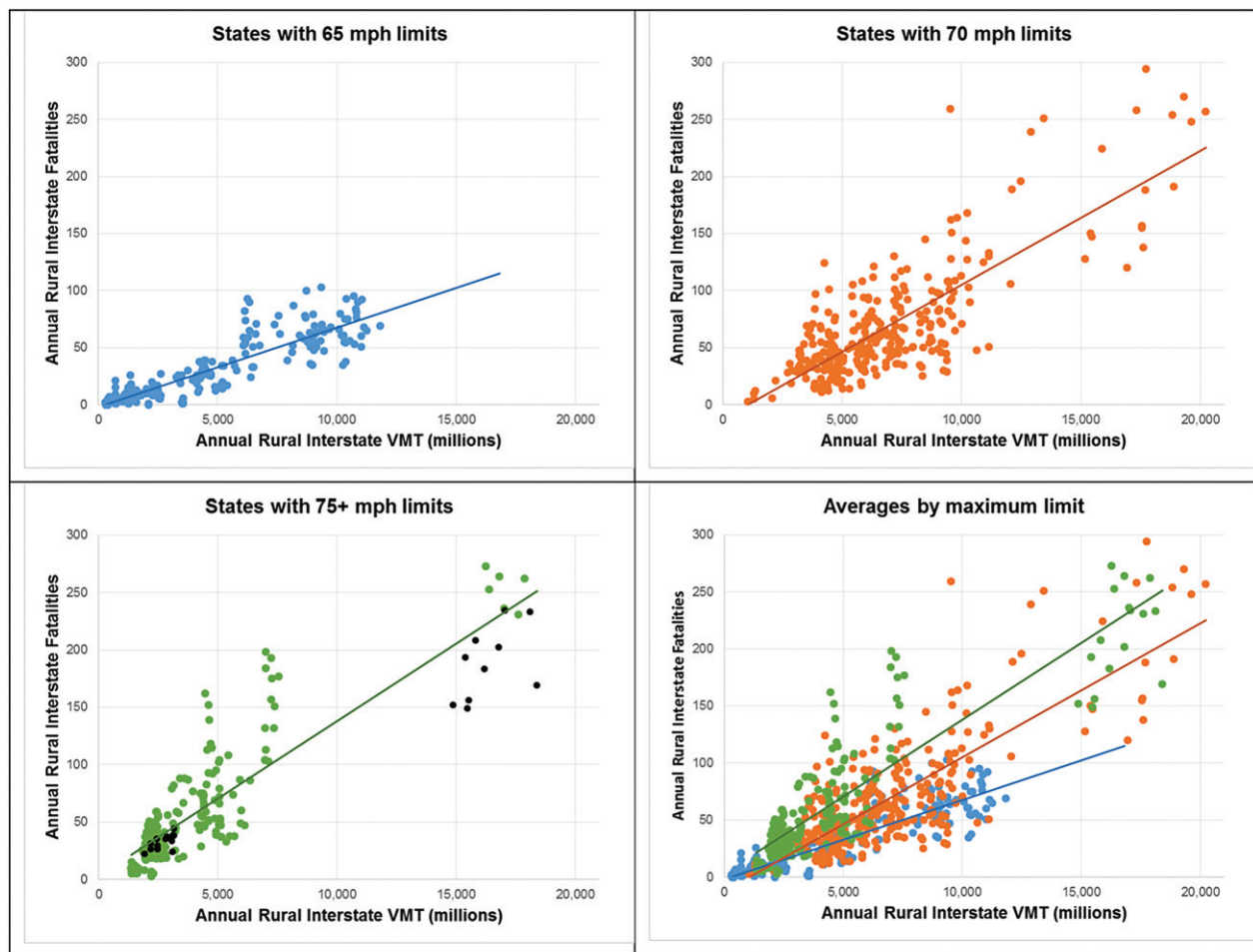


Figure 8 Rural interstate fatalities by maximum speed limit. (Sources: NHTSA Fatality Analysis Reporting System and FHWA Highway Statistics Series).

In general, the extant research literature has focused predominantly on limited access freeways. This is due, in part, to the fact that there is significant variability in the design characteristics of two-lane highways, where speeds can range from 25 to 75 mph (40–121 km/h). Continuing on this point, one concern that is unique to rural two-lane highways is the presence of speed reduction zones when such highways pass through incorporated cities or towns. Guidance is provided for local agencies in National Cooperative Highway Research Program (NCHRP) Report 737 for the implementation of such speed transition zones (Torbic et al., 2012). Several potential factors have been shown to affect drivers' selection of operating speeds as they enter speed reduction zones (Cruzado and Donnell, 2009, 2010). The magnitude of the speed reduction and other characteristics, such as lane width, paved shoulder width, and lateral clearance tend to reduce operating speeds entering a speed reduction zone.

Overall, the safety literature generally suggests that increasing speed limits on non-freeways will result in an increase in the overall crash rate and will also result in a higher proportion of more severe crashes due to the increase in the energy dissipated during crashes due to vehicles traveling at higher speeds. An NCHRP study estimated that increasing the speed limit from 55 to 65 mph (89–105 km/h) would increase the total crash rate by 3.3%, and the probability of a fatality (assuming a crash had occurred) would increase by 24% (Kockelman, 2006).

It should be noted that Sweden, which is one of the countries with the lowest fatality rates in the world, lowered the speed limits on most of its two-lane roads and other roads where head-on collisions are possible from 90 km/h (56 mph) or 100 km/h (62 mph) to 80 km/h (50 mph) in 2019 (ITF, 2019). Prior to the year 2000, many of those roads had a speed limit of 110 km/h (68 mph). However, some of the two-lane roads have had continuous centerline barriers installed, and the speed limit on these have been kept at, or increased to, 100 km/h (62 mph). As of 2019, there are around 1800 speed-enforcement cameras on rural roads in Sweden to enforce these speed limits. The reduction of speed limits and speeds are often credited with Sweden having seen an 83% reduction in the number of traffic fatalities since they peaked in 1966 at the same time as its population has increased by 32%. Sweden in 2019 had 2.16 fatalities per hundred thousand inhabitants whereas the United States has a population rate over 5 times that (WHO, 2020). The difference can be partly explained by higher travel speeds in the United States but also because of more automobile travel in general in the United States. But even if we account for the latter, the fatality rate per HMVMT is around 0.42 in Sweden and well over 1 in the United States.

Current Practices in Setting Speed Limits

As noted previously, 18 states have increased their maximum statutory speed limits to 75 mph or above. The preliminary evidence, as demonstrated in Fig. 8, shows that states with higher limits tend to experience consistently more fatal crashes than states with lower limits. Furthermore, one important contrast as compared to prior changes, such as those following the three speed limit law changes discussed previously, is that the more recent increases have generally been done selectively based upon traffic engineering, speed, and safety studies conducted by the state departments of transportation.

This is an important distinction as not all segments of a particular roadway class are likely to be acceptable candidates for speed limit increases. In particular, segments with extensive horizontal or vertical curvature, sight distance limitations, or other features that may not comply with current design standards (e.g., design exceptions) may not be suitable for speed limit increases. Similarly, locations at which prevailing speeds are generally aligned with the existing speed limit or locations where there is a history of crashes may not be suitable candidates for increases.

Increasing speed limits also introduces important economic considerations, including the costs required for geometric modifications (e.g., vertical or horizontal realignment) to satisfy minimum design criteria at higher design speeds. Consequently, increases to the higher ranges of speed limits should be considered only if the critical geometric elements can maintain design speed compliance without major modification. However, even for roadways on which design speed compliance is maintained, careful, site-specific consideration must be given to the potential safety impacts, particularly with respect to fatal and injury crashes, which may result after the speed limit is increased.

A 2018 survey conducted by the National Committee on Uniform Traffic Control Devices (NCUTCD) Task Force shows that those professionals who perform posted speed limit studies rarely use only the 85th percentile speed and, instead, consider the context of the roadway where the speed limit change is being considered (Fitzpatrick et al., 2019). This includes the factors explicitly noted in the MUTCD that may be considered when establishing or reevaluating speed limits, which include road characteristics, roadside development, and reported crash experiences.

Moving forward, the establishment of maximum speed limits in rural areas will likely continue to be an area of substantive interest to researchers, legislatures, and the traveling public. While additional research is warranted to fully understand the nature of these relationships, agencies should exercise caution whenever speed limit increases are under consideration. The impacts of emerging concerns, such as distracted driving and the integration of connective and autonomous vehicles, are among the issues that prompt the need for such scrutiny.

References

- American Association of State Highway and Transportation Officials (AASHTO), 2018. Policy on Geometric Design of Highways and Streets. Washington, D.C.
- Baum, H.M., Wells, J.K., Lund, A.K., 1991. The fatality consequences of the 65 MPH speed limits. *J. Safety Res.* 22 (4), 171–177.

- Bundesanstalt für Straßenwesen, 2019. International Traffic and Accident Data: Selected Risk Values for the Year 2012 (PDF). Bundesanstalt für Straßenwesen (Federal Highway Research Institute). Available from: http://www.bast.de/EN/Publications/Media/Unfallkarten-international-englisch.pdf?__blob=publicationFile.
- Castillo-Manzanao, J.I., Castro-Núñez, M., López-Valpuestaa, L., Vassallo, F.V., 2019. The complex relationship between increases to speed limits and traffic fatalities: evidence from a meta-analysis. *Safety Sci.* 111, 287–297.
- Cruzado, I., Donnell, E.T., 2009. Evaluating effectiveness of dynamic speed display signs in transition zones of two-lane rural highways in Pennsylvania. *Transp. Res. Rec.: J. Transp. Res. Board* 2112, 1–8.
- Cruzado, I., Donnell, E.T., 2010. Factors affecting driver speed choice along two-lane rural highway transition zones. *J. Transp. Eng.* 136 (8).
- Davis, A., Hacker, E., Savolainen, P.T., Gates, T.J., 2015. Longitudinal analysis of rural interstate fatalities in relation to speed limit policies. *Transp. Res. Rec.* 2514, 21–31.
- Donnell, E.T., Hines, S.C., Mahoney, K.M., Porter, R.J., McGee, H., 2009. *Speed Concepts: Informational Guide*, FHWA-SA-10-001. US Department of Transportation, FHWA, Washington, DC.
- Elvik, R., 2013. A re-parameterisation of the power model of the relationship between the speed of traffic and the number of accidents and accident victims. *Accid. Anal. Preven.* 50, 854–860.
- Elvik, R., Vadeby, A., Hels, T., van Schagen, I., 2019. Updated estimates of the relationship between speed and road safety at the aggregate and individual levels. *Accid. Anal. Preven.* 123, 114–122.
- English, J.W., Levin, S.H., 1978. *Traffic Speed Limit Laws in the United States*. National Highway Traffic Safety Administration.
- Farmer, C.M., Retting, R.A., Lund, A.K., 1999. Changes in motor vehicle occupant fatalities after repeal of the national maximum speed limit. *Accid. Anal. Preven.* 31 (5), 537–543.
- Federal Highway Administration (FHWA), 1971. *Manual on Uniform Traffic Control Devices for Streets and Highways*. US Department of Transportation, FHWA, Washington, DC.
- Federal Highway Administration (FHWA), 2012. *Manual on Uniform Traffic Control Devices for Streets and Highways: 2009 Edition Including Revision 1 dated May 2012 and Revision 2*. US Department of Transportation, FHWA, Washington, DC.
- Federal Highway Administration, Highway Statistics Series, 2020. Available from: <http://www.fhwa.dot.gov/policyinformation/statistics.cfm>.
- Fildes, B., Rumbold, G., Leeming, A., 1991. *Speed Behaviour and Drivers' Attitude to Speeding*. Monash University Accident Research Centre.
- Finch, D., Kompfner, P., Lockwood, C.R., Mayock, G., 1994. *Speed, Speed Limits and Crashes*. Transport Research Laboratory Report No. 58, Crowthorne.
- Fitzpatrick, K., Carlson, P., Brewer, M.A., Wooldridge, M.A., Miaou, S.-P., 2003. NCHRP Report 504: Design speed, operating speed, and posted speed practices. Washington, D.C., pp. 103.
- Fitzpatrick, K., McCourt, R., Das, S., 2019. Current attitudes among transportation professionals with respect to the setting of posted speed limits. *Transp. Res. Rec.: J. Transp. Res. Board* 2673, 778–788.
- Forester, T.H., McNow, R.F., Singell, L.D., 1984. A cost-benefit analysis of the 55 mph speed limit. *South. Econ. J.* 50 (3), 631–641.
- Freedman, M., Esterlitz, J.R., 1990. Effect of the 65 mph speed limit on speeds in three states. *Transp. Res. Rec.: J. Transp. Res. Board* 1281, 52–61.
- Friedman, L.S., Hedeker, D., Richter, E.D., 2009. Long-term effects of repealing the national maximum speed limit in the United States. *Am. J. Public Health* 99 (9), 1626–1631.
- Gates, T., Savolainen, P., Kay, J., Finkelman, J., Davis, A., 2015. *Evaluating Outcomes of Raising Speed Limits on High Speed Non-Freeways*, Michigan Department of Transportation, Report No. RC-1609B.
- International Transport Forum (ITF), 2018. *Speed and Crash Risk*. OECD/ITF, Paris.
- International Transport Forum, 2019. *Road Safety Annual Report 2019: Sweden*, Organization for Economic Cooperation and Development.
- Kallberg, V.-P., Toivanen, S., 1998. *Framework for Assessing the Impact of Speed in Road Transport*. MASTER Project Report. Technical Research Centre of Finland VTT, Espoo, Finland.
- Kockelman, K., 2006. *CRA International, Inc., Safety Impacts and Other Implications of Raised Speed Limits on High-Speed Roads*. Transportation Research Board.
- Lynam, D., Nilsson, G., Morsink, P., Sexton, B., Twisk, D., Goldenbeld, Ch., Wegman, F., 2005. *SUNflower + 6 – An Extended Study of the Development of Road Safety in Sweden, the United Kingdom and the Netherlands*. Transport Research Laboratory. Crowthorne.
- Maycock, G., Brocklebank, P., Hall, R., 1999. Road layout design standards and driver behavior. *Proc. ICE-Transp* 135 (3), 115–122.
- Musican, O., Bar-Gera, H., Schectman, E., 2016. Impact of speed limit change on driving speed and road safety on interurban roads. *Transp. Res. Rec.: J. Transp. Res. Board* 2601, 42–49.
- National Highway Traffic Safety Administration (NHTSA), 1988. *Preliminary Report to Congress on the 65 mph Speed Limit - Safety Effects in 1987*. Washington D.C.
- National Highway Traffic Safety Administration, 2020. *Fatality Analysis Reporting System (FARS)*. Available from: <http://www-fars.nhtsa.dot.gov/QueryTool/QuerySection/SelectYear.aspx>.
- Nilsson, G., 2004. *Traffic Safety Dimensions and the Power Model to Describe the Effect of Speed on Safety*. Lund University.
- National Transportation Safety Board, 2017. *Reducing Speeding-Related Crashes Involving Passenger Vehicles*. NTSB/SS-17/01 PB2017-102341. National Transportation Safety Board, Washington, D.C.
- Parker Jr, M., 1997. *Effects of Raising and Lowering Speed Limits on Selected Roadway Sections*. Federal Highway Administration.
- Quimby, A., Maycock, G., Palmer, C., Buttress, S., 1999. *The Factors that Influence a Driver's Choice of Speed: A Questionnaire Study*. Transport Research Laboratory Report No. 325.
- Royal, D., 2003. *National Survey of Speeding and Unsafe Driving Attitudes and Behaviors: 2002*. National Highway Traffic Safety Administration.
- SafetyNet, Speeding, 2009. Available from: https://ec.europa.eu/transport/road_safety/sites/roadsafety/files/specialist/knowledge/pdf/speeding.pdf.
- Solomon, D., 1964. *Accidents on Main Rural Highways Related to Speed, Driver, and Vehicle*. United States Bureau of Public Roads, Washington, D.C.
- Spiegel, C.H., 2020. *Niederlande: Klimaaktivisten siegen gegen eigene Regierung - DER SPIEGEL - Wissenschaft*. Available from: www.spiegel.de (in German).
- Taylor, M., Lynam, D.A., Baruya, A., 2000. *The Effects of Drivers' Speed on the Frequency of Road Accidents*. Transport Research Laboratory, Crowthorne.
- The Local, 2019. *Germany considers Autobahn speed limit to fight climate change*. Available from: <https://www.thelocal.de/20190121/german-weighing-speed-limit-on-autobahn>.
- Torbic, D.J., Gilmore, D.K., Bauer, K.M., Bokenroger, C.D., Harwood, D.W., Lucas, L.M., Frazier, R.J., Kinzel, C.S., Petree, D.L., and Forsberg, M.D., 2012. *Design Guidance for High-Speed to Low-Speed Transition Zones for Rural Highways*. NCHRP Report 737, Transportation Research Board, National Research Council, Washington, D.C.
- Transportation Research Board (TRB), 1998. *Managing Speed: Review of Current Practice for Setting and Enforcing Speed Limits*. Special Report 254. National Academy Press, Washington, DC.
- West, L.B., Dunn, J., 1971. *Accidents, Speed Deviation and Speed Limits*. Institute of Traffic Engineering.
- World Health Organization, 2020. *Road Traffic Deaths – Data by Country*, Available from: <https://apps.who.int/gho/data/node.main.A997>.
- Wilmot, C.G., Khanal, M., 1999. Effect of speed limits on speed and safety: a review. *Transp. Rev.* 19 (4), 315–329.

Speed Limits on Urban Streets

Anna Bray Sharpin, Claudia Adriazola-Steil, Ben Welle, Natalia Lleras, World Resources Institute, Washington, DC, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	632
History/Evolution of Speed Limits in Cities	632
The Safe System Approach and Speed Limits	633
Speed Limits and Operating Speeds	635
Safe Speed Limits for Urban Areas	635
Considering Road Function When Establishing Speed Limits	636
Benefits of Appropriate Speed Limits and Operating Speeds	639
Public Perceptions and the Politics of Speed Limits	639
Conclusion	640
Permissions	640
See Also	640
References	640
Further Reading	640

Introduction

Streets and roads form the foundation of travel and interaction in urban areas, as well as constituting most of the public space in cities. The concept of what streets are for and how they are used has changed over time, and continues to change, generating both opportunity and conflict that must be managed through an ever-evolving framework that includes political leadership, law, policy, regulation, design, monitoring and evaluation, and community participation. The interplay between all of these factors is relevant for setting safe speed limits.

The current situation in many urban areas around the world is streets that favor high speeds and a prioritization of space for moving and parked vehicles over other mobility and community needs. Allowing high speeds has resulted in urban areas that generate a high proportion of crashes that cause death and injury to people walking, biking, or using other human scale means of transport, act as a deterrent to these mobility modes for people who can afford other options, and degrade the urban environment with noise and air pollution. As the body of evidence demonstrating the direct link between vehicle speeds and crashes, deaths and serious injuries, the understanding of how and why speed limits are set in urban areas has responded accordingly, and efforts are now underway in many cities around the world to proactively determine and achieve safe speed limits and safe operating speeds. As well as the aforementioned safety benefits, but such speed limits also generate many co-benefits such as improved quality of life, accessibility, economic development, and more efficient fuel consumption, generating a positive cycle for urban areas.

Speeds are especially dangerous for vulnerable road users, pedestrians, cyclists and motorcyclists, as they do not have a protection of metal around their bodies as car drivers do. In case of a road crash, they receive the force of the impact directly into their bodies, and the higher the speed the higher the severity of the injuries. Lowering speeds results in more people walking or cycling as the risk of being involved in a road crash decreases, and these travel options feel safer and more comfortable. Low speeds are key to promoting active transport.

History/Evolution of Speed Limits in Cities

Historically, streets were used for walking or traveling at sedate speed by horse. In 1652, the American colony of New Amsterdam passed a law stating that "No wagons, carts, or sleighs shall be run, rode, or driven at a gallop." The punishment for breaking the law was "two pounds Flemish," the equivalent of US \$50 today. In the 1800s, bicycles and trams (streetcars) were added to many urban streets, but traffic still moved at human pace. The first numeric speed limit was the 10 mph (16 km/h) limit introduced in the United Kingdom in 1861. But that turned out to be too high for public comfort, so, by resident demand, it was reduced to 2 mph (3 km/h) in towns and 4 mph (6 km/h) in rural areas by the 1865 "Red Flag Act." In the United States, the first numerical speed limit for motor vehicles was set in 1901 in Connecticut with a maximum legal speed of 12 mph (19 km/h) in cities and 15 mph (24 km/h) on rural roads. In 1903, in the United Kingdom, the national speed limit was raised to 20 mph (30 km/h). In other words, streets were still low-speed environments where many users mingled and negotiated shared use of space on foot, bicycle, carriage, or public transport, such as trams. When cars became somewhat more common, this dynamic stayed much the same. But as the speed capacity and volume of cars grew, and the first fatalities caused by cars occurred, the discussion began to shift. Initially the expectation was on car companies and city regulators to limit the speed of cars and preserve the safe interaction and intermingling of many different types of

transport on the street. But influence from car companies and driver associations gradually shifted the expectation of what streets were for and what appropriate speed limits were, until other types of uses were pushed off streets and confined on sidewalks, and cars moving at much higher speeds became the default use of most streets. The type of road and urban areas that have been designed or evolved since private car use became widely accessible have further compounded the role of streets as thoroughfares for fast moving motor vehicle traffic.

Unfortunately, this shift towards a focus on streets as serving the sole purpose of moving (or parking) cars has resulted in a prevalence of urban environments that are hostile to people who are not in vehicles. The statistics on this are clear. Most victims of traffic fatalities and serious injuries in urban areas are so called “vulnerable users,” that is, people who are walking, bicycling, or otherwise traveling outside of a vehicle. For example, in New York City in 2018, 114 pedestrians, 10 bicyclists, and 39 motorcyclists were killed in comparison to a combined total of 37 car and truck occupants. That adds up to 200 deaths, the lowest in over a century, which certainly is an improvement over the 1360 deaths seen back in 1929, or the 701 traffic deaths as recently as 1990. However, more pedestrians died in New York in 2018 than in the year before. While safety for vehicle occupants, particularly in wealthy countries, has steadily improved, thanks to technological and legal advances regarding the use of seatbelts, airbags, and other safety devices to protect car occupants, design technology and legal requirements to protect people using the streets outside of cars have not advanced at the same pace or had the same impacts. As a result, in many places, even if traffic deaths and fatalities are declining overall, they are increasing among the most vulnerable groups. Furthermore, these hostile environments to people who are not using private vehicles have broader impacts because they affect people’s decision making about how to travel, generating significant financial, environmental and physical effects, and creating a negative feedback loop of increased car use and risk. For example, in wealthy countries, the proportion of children biking or walking to school is in dramatic decline, while childhood obesity is increasing. Meanwhile in many lower income countries women have been recorded as opting for substantially longer commutes than men in order to reduce their exposure to risk. This has a variety of negative carry-on effects on their health, access to jobs and services, and leisure time.

One traditional approach to setting or adjusting speed limits was to measure traffic operating speeds and then set the speed limit at **the 85th percentile**, that is to say, the speed that 85% of motor vehicles were traveling at or under. This method was first used in Europe. While this practice has waned there now, it is still often mentioned as a valid method in the United States and many other countries that follow the US transport planning and engineering guidance. However, this method should no longer be recommended for setting safe urban speed limits. This is because it is a reactive approach that allows the speed limits to be set based on driver perceptions and comfort with a certain speed. This is typically derived from the operating speed, which is heavily influenced by the design of the street. Unless the street has been proactively designed to generate a desired target speed, this is not likely to be the actual safe speed, both because the perception and reality of safety and comfort are very different between a vehicle occupant and a person walking, biking, or using another personal mobility device, and also because humans do not have a well-developed sense of speed and the risks associated with it, compared, for example, to our perception of the risks associated with height. Furthermore, many streets are designed in a way that is inappropriate for their current use or range of users (for example, traditionally many urban areas in developing countries, lacking their own street design guidance, have applied the US highway design guidance to streets in urban areas). In these cases, it is the design of the street that should be adapted to the needs of the location, rather than the speed that should be adapted to the design.

In other places, speed limits are set according to the type of road, in terms of its predominant purpose as a thoroughfare or local point of access. This is often referred to as “function” or “classification.” In this case, roads designated as arterials usually have higher default speed limits than roads designated as intermediate or local roads. However, most roads, especially long ones in larger cities, can vary from block to block in terms of adjacent land use and thus types of activity and road users present, or adjacent land uses can change over time, such as residences constructed on previously undeveloped or industrial land. Therefore, the default speed limit should not be rigid, but adjusted for particular needs. For example, a commercial center on an arterial road, that concentrates more pedestrians than other parts of the road. If the road classification does not take into account other variables as the context, road conditions, or quality of infrastructure for vulnerable road users in this way, then the speed limit may still not be appropriate, and the classification may also need to be reviewed. For more information, see the subsequent section, “Considering road function when establishing speed limits.”

The Safe System Approach and Speed Limits

The Safe System approach to road safety reframes the way streets and mobility systems in cities operate to place safety at the forefront, particularly the safety of the most sustainable, yet vulnerable, forms of transport—walking and cycling. This is in contrast to the status quo that has emerged in most regions of the world over the past 100 years that has placed private cars at the center of the mobility and safety planning system to the detriment of safety and accessibility of other types of transport. The Safe System approach, also referred to as Vision Zero in this Encyclopedia, has as its foundation the principle that **no serious injury or loss of life** is an acceptable trade-off for “efficiency” or “modernity.” People will always commit errors, and the mobility system should be designed to mitigate the possibility and severity of these, so that no one has to pay with their life or permanent disability (Welle et al., 2018).

Establishing appropriate speed limits, target speeds, and operating speeds is therefore a fundamental element of a safe mobility system, as speed is the key determinant of whether a crash happens at all, and if so, how severe it is. This is for several reasons. At

Higher vehicle speeds require longer stopping distances

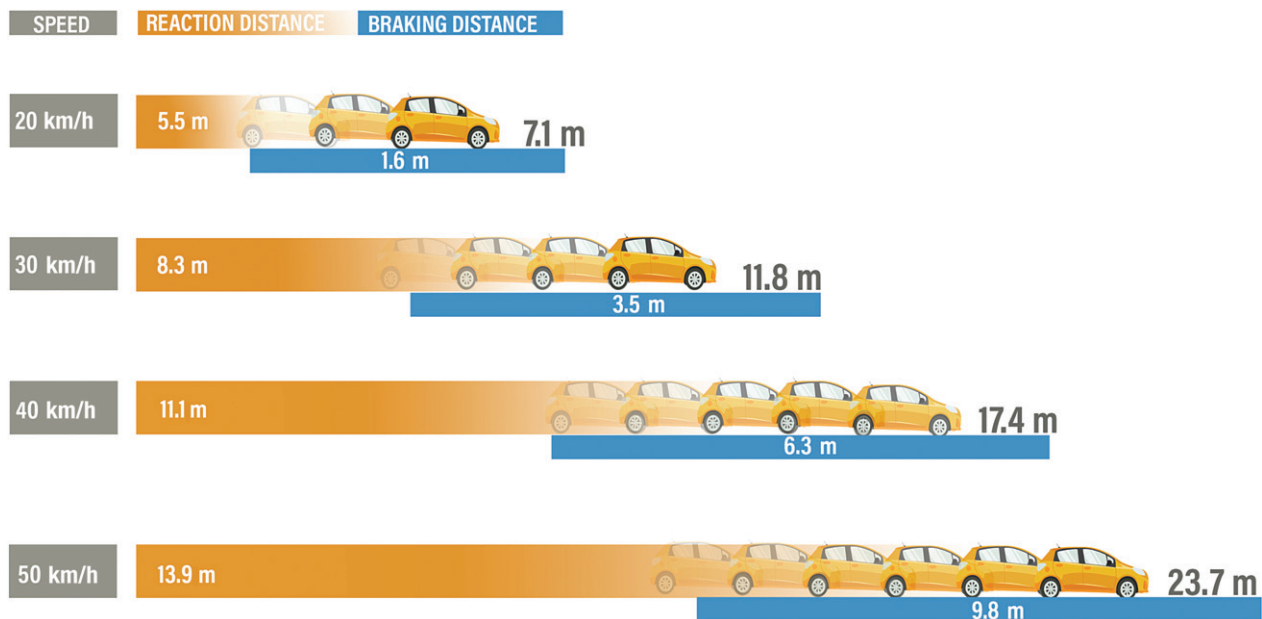


Figure 1 Combined reaction and braking distances required at various speeds. Note: The above distances are minimum distances. The total stopping distance also depends on the road surface, weather conditions, and age/condition of the vehicle. Source: Graphic prepared by World Resources Institute (WRI) with data and calculations supplied by Per Garder.

higher speeds, drivers have reduced peripheral vision. Driving at very high speeds can result in a combination of tunnel vision and decreased depth perception for the driver. At lower speeds, drivers have a wider field of vision and are more able to notice other road-users and any unexpected obstacles or events that they need to react to. Driving at lower speeds also enables drivers to stop within a shorter distance (Job and Sakashita, 2016). This is because the stopping distance of a vehicle is a combination of the distance traveled during the driver's reaction time and the distance it takes for the car to stop after the brakes are applied. At higher speeds, a car will travel further during this reaction time and due to momentum, the braking distance is also greater, and proportional to the speed squared. Another way to put this is that the impact of increased speeds is not linear but increases exponentially. The same rule applies to risk of death or serious injury.

For example, with good tires and dry pavement, and a typical reaction time of 1.0 s, a driver could stop in 11.8 m if the car is driven at 30 km/h and they brake as hard as possible (8.3 m in reaction distance and 3.5 m to brake). If the car instead is driven at 50 km/h, it would need 23.7 m to stop (13.9 m in reaction distance and 9.8 m to brake). So, if a driver experienced an unexpected event with a pedestrian appearing 12 m in front of them, and they were driving at 30 km/h, the driver would be able to stop in time to avoid hitting them, but if they were traveling at 50 km/h, they would not even have started to brake and would hit that pedestrian at 50 km/h. This would potentially cause serious injury or death. Even if they were traveling at a speed of 40 km/h, they would need 17.4 m to stop totally, so they would hit the pedestrian without managing to slow down by more than a couple of km/h (6.3 m reaction distance and 11.1 m braking distance). On the other hand, if the vehicle were traveling at 20 km/h, they would be able to stop in just over 7 m (5.5 m reaction distance and 1.6 m braking distance) giving them a margin of almost 5 m (Fig. 1) (Garber and Hoel, 2015). It is important to note that the distances given here are for a best-case scenario with good vehicle and road conditions. If their tires were worn or the pavement was wet, they would use up much of that 5-m margin for coming to a full stop. Speed thus clearly affects the possibility of a pedestrian's survival in the case of an incident on the road. Impact speed is particularly pertinent to consider when there are people walking and biking, whose bodies much more are exposed to the impacts of a crash than occupants of cars, trucks, and buses who are protected by the body of the vehicle.

Furthermore, a safe speed limit, which generates a safe operating speed thanks to infrastructure, regulation and police control, will reduce variation between the speeds of different vehicles on the street, and the related issue of risky mid-block acceleration.

Another key principle of the Safe System approach is that road safety should be tackled in a **proactive** manner, rather than a reactive one. In the context of speed limits, this means that safe limits should be established based on an analysis of multiple factors at hand, such as the types and volumes of different road users, the function of the road, adjacent land uses, any available data on near-misses or adapted travel patterns caused by perceived risk and discomfort, and the quality and adequacy of the infrastructure present, particularly at intersections, rather than as a reactive response based on the current operating speeds of vehicles, the original intent of the street (which may change over time), or to crashes only once they do happen.

Speed Limits and Operating Speeds

The **speed limit** is the legal speed permitted on a street under local, state, or national traffic laws or regulations. The purpose of a speed limit is to align the vehicle speeds on a street or road with the function, dynamics, context, and road users present, to create a predictable environment and avoid dangerous speeds. However, setting a speed limit is just the first step in achieving compliance with this limit. Ensuring that drivers follow speed limits requires a variety of enforcement, regulation and design, and awareness raising measures implemented in combination. One mechanism for this that is well proven in evidence is roads that are designed to be “self-enforcing,” that is, roads that have physical design features such that it is difficult or impossible for a vehicle to exceed the designated speed limit. Driver compliance with the speed limit can also be monitored and enforced by a specific authority, such as a police or traffic agent that is attributed with the power to issue a warning or fine directly to a driver, or via automated technology such as speed cameras. Speed limiting vehicle technology is also increasingly available and presents a significant opportunity to complement or reduce the need for enforcement. Street designs that relate directly to the speed limit and **desired operating speed**, also known as the **target speed**, should be complemented by other approaches such as integrating driver awareness and understanding of typical speed limits into driver education and licensing processes, and posting the speed limit in highly visible locations.

The actual combination of mechanisms applied to achieve compliance with the speed limit in a given location should be determined based on local conditions including: function and type of road, land-use, dynamics, road users present, street construction materials, design guidelines and policies, the culture of legality, law enforcement funding and priorities, access and cost of automated technology, the influence of national and state level laws on local enforcement attributions, and maintenance.

It is important to note that the speed limit alone does not actually determine the **operating speeds** in a street, neighborhood, or urban area. The operating speed is the actual speed that vehicles travel in a given street and has a direct impact on safety. Actual operating speeds can be significantly higher, or lower than the speed limit, or have major temporal variations. This is due to a variety of factors; mainly to what extent the mechanisms for compliance listed above are implemented. Limits may not be posted or enforced at all; speed regulation may include a degree of “tolerance” before enforcement is triggered; the physical design of the road may be out of line with the speed limit and allow drivers to feel comfortable moving at higher speeds (or the reverse). In addition, in many urban areas, traffic congestion, primarily caused by vehicle volume, and traffic conflicts, may mean that during peak times, vehicles move much more slowly than the posted limit.

For the same reasons, it is important that in addition to setting a safe speed limit, authorities ensure that the factors are in place to also generate a **safe operating speed**, that is, below or at the legal speed limit. This known as setting a **target speed**. Research has found that a regulation change such as lowering the posted speed limit typically only yields a reduction of a few kilometers/hour in average speeds if the change is made in isolation, indicating that proactive mechanisms for enforcement, either by enforcing agents or automated systems, as well as context-appropriate street design are necessary to achieve genuinely safe operating speeds that reflect the speed limit and desired target speed.

Traditionally, wide lanes, number of lanes and other street design factors such as long blocks, as well as a typical “tolerance” of 10 km/h or 10 mph before enforcement of speeding fines has meant that the de-facto speed limit is often substantially higher than the actual limit. The evidence shows that this can be the difference between life and death in a crash situation, so for a safe speed limit to truly be safe, it must be accompanied by a similarly safe operating speed. If there is concern about the limitations of enforcing speed limits at low speeds, or with no “tolerance” for excesses, then street design techniques may be applied to create a street environment that is “self-enforcing” by making it physically difficult or impossible for a driver to negotiate the street at a speed higher than the speed limit.

Safe Speed Limits for Urban Areas

Under the Safe System approach, a “safe” **speed limit** is considered one where most crashes can be avoided, and if one did occur, then those involved would have a 90% or more chance of survival and avoid permanent injury. This means that speed limits should be responsive to the specific types of users on that road. For example, an urban street with a commercial land use, and a high presence of people walking, biking, or using other human scale means of transport such as scooters, strollers, or wheel chairs, or where bicyclists share the road with drivers without any physical segregation of infrastructure—should have a speed limit of 30 km/h or less. This is because research has found that above approximately 30 km/h, the risk of a fatality in a crash with an unprotected human increases exponentially.

Data for pedestrians struck by a car or light truck in the United States from 2007 to 2009, shows that the average risk of severe injuries for a pedestrian struck by a vehicle reaches 10% at an impact speed of 26 km/h (16 mph), 25% at 37 km/h (23 mph), 50% at 50 km/h (31 mph), 75% at 63 km/h (39 mph), and 90% at 74 km/h (46 mph). The average risk of death for a pedestrian reaches 10% at an impact speed of 37 km/h (23 mph), 25% at 51 km/h (32 mph), 50% at 51 km/h (42 mph), 75% at 80 km/h (50 mph), and 90% at 74 km/h (58 mph) (Fig. 2). However, averages can be misleading, because risks vary significantly by both victim and contextual factors, such as age, gender, quality of emergency response and hospital care, and local vehicle fleet (Tefft, 2011).

It is important to note that most of this type of research has been undertaken in wealthy European countries and the United States where emergency response and hospital care is rapid and well-coordinated, which likely contributes to higher likely survival rates at these speeds than in countries with poorer access to medical care, such low- and middle-income countries where the majority of traffic fatalities are concentrated.

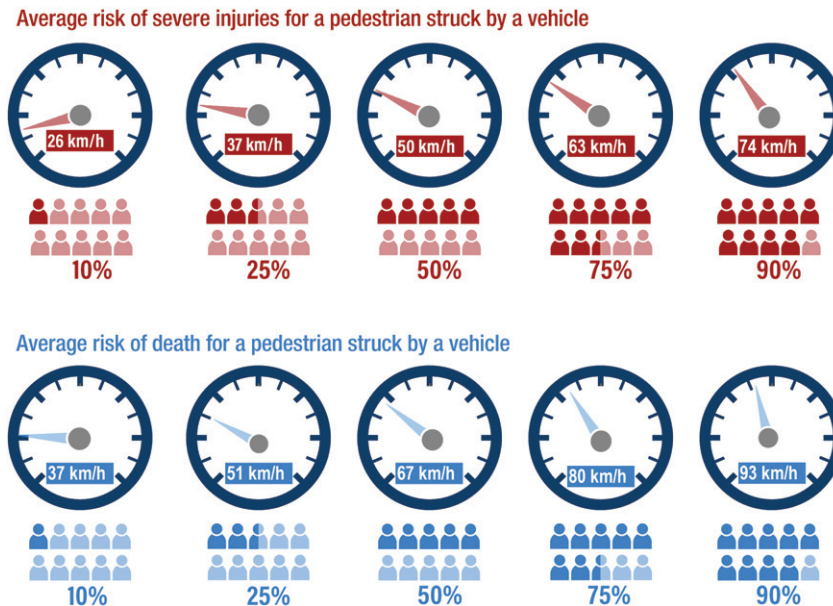


Figure 2 Average risk of pedestrian death or serious injury at various vehicle speeds from US data 2007–2009. Note: The above values are averages taken from one study based on the US data from 2007–09. The likelihood of survival will vary in specific cases depending on characteristics of the victim and vehicle and location involved, such as, age, gender, stature, vehicle type, speed, and quality of emergency response services. Source: Graphic prepared by World Resources Institute (WRI) with data from Tefft, 2011.

As an example of the impact of the local vehicle fleet on risk, the situation for pedestrian safety in the United States has declined over the last decade, since the above numbers were calculated, and one contributing factor is that the proportion of trucks and SUVs in the vehicle fleet has continually increased. Due to the height of the hood of such vehicles, even adults are knocked down in front of the vehicle if hit. In contrast, car design standards in Europe include design requirements to mitigate pedestrian injuries, a feature not required for cars in other regions that typically only require designs to protect the vehicle occupants.

With regard to age and gender and risk, the average risk of severe injury or death for a 70-year-old pedestrian struck by a car traveling at 40 km/h (25 mph) is similar to the risk for a 30-year-old pedestrian struck at 55 km/h (35 mph). Additionally, based on the history of both crash test dummy design based on average male measurements, and analysis of crash victims, which are predominantly male, it is likely that this research is skewed toward impact on the male body, so for women, the safe speed limit may be even lower. This is certainly the case when it comes to children, who are more vulnerable not just because their fragile bodies and less developed bones are less able to withstand a crash and their lower center of gravity and shorter stature means that they are more likely be knocked down in front of cars rather than land on the hood, and to receive upper body and head injuries, but also because they are more difficult for drivers to see, and do not yet have fully developed speed perception or impulse control. For this reason, appropriate speed limits around schools should be set even lower than other urban streets; ideally 20 km/h or lower. A lower limit (10–20 km/h) may also be appropriate in the case of street design types or functions where street space is shared between people driving vehicles, walking, playing, and bicycling.

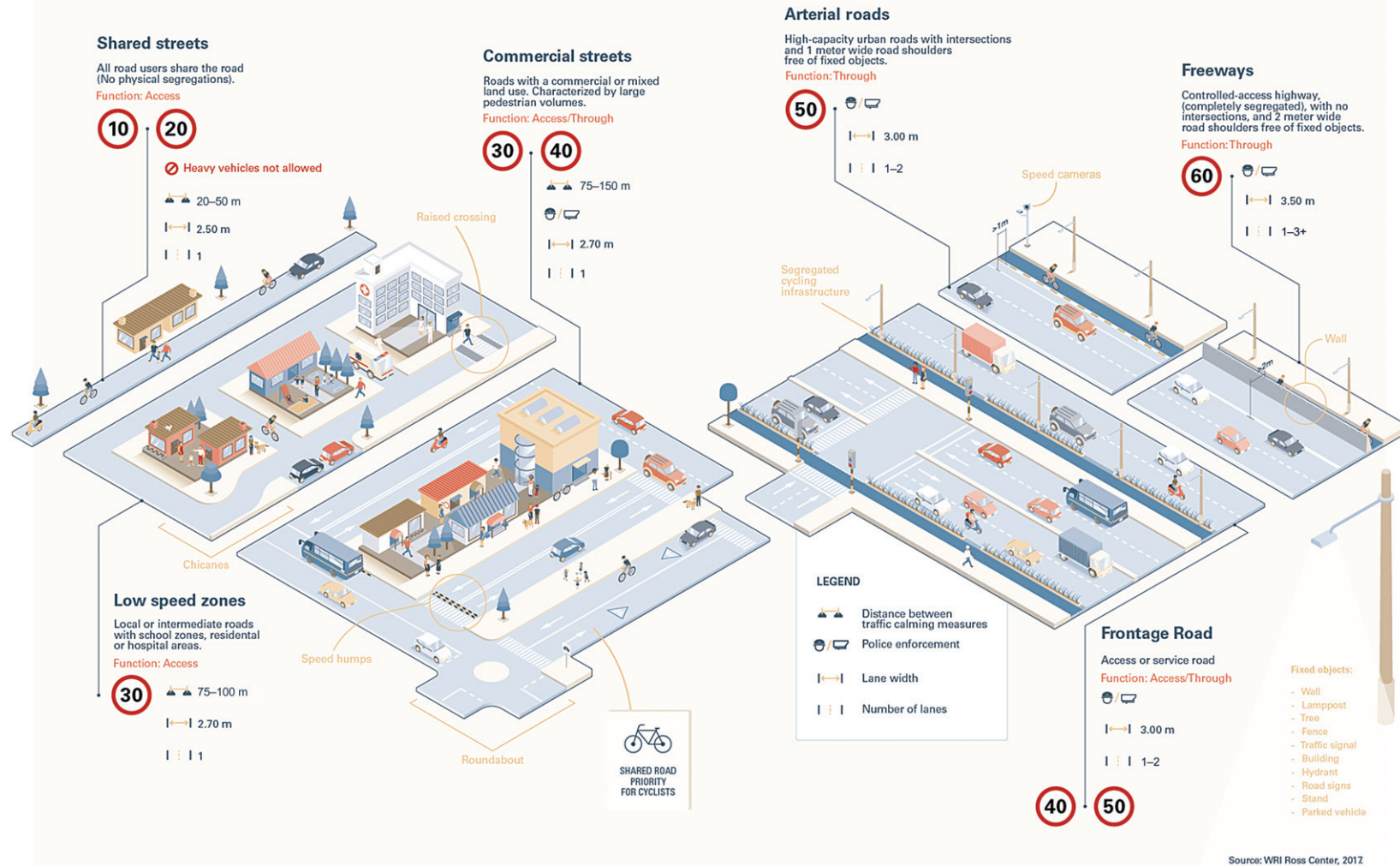
Many countries have also long set a **default maximum speed limit of 50 km/h** in urban areas, however research has found that this is too high to be a safe general speed limit in urban roads with a high prevalence of vulnerable users. This speed limit is best reserved for high traffic volume key arterials within cities, applied in conjunction with segregated infrastructure for people biking and walking and accessing public transport, and controlled intersections for safe crossing, so that the risks presented by higher speeds are counteracted by reduced exposure to risk.

In summary, the speed limit should be defined by both taking into account the function of the road and the context. Elements to consider include: the interaction or integration of vehicle traffic and vulnerable road users, the type of vehicles using the road and type of segregation between types of vehicle (e.g., the presence or absence of protected bicycle lanes), the volume and flow of traffic, the free-flowing traffic speed, the nature and intensity of surrounding land uses and infrastructure (e.g., the variety of activities/services, vehicle entryways, obstacles affecting visibility, etc.), the crash history of a location, and the design of intersections (Figs. 3 and 4) (Alcaldía Mayor de et al., 2019).

Considering Road Function When Establishing Speed Limits

One helpful mechanism to facilitate setting safe speed limits in urban areas is to establish a process of categorizing roads in terms of their function and context. Maximum speed limits can then be established for each category, with the option for local authorities to

Speed Management



Source: WRI Ross Center, 2017

Figure 3 Examples of road/speed classifications. Source: Graphic prepared by World Resources Institute for *Alcaldía Mayor de et al., 2019*. Translated from Spanish and adapted for the Elsevier Encyclopedia of Transportation.

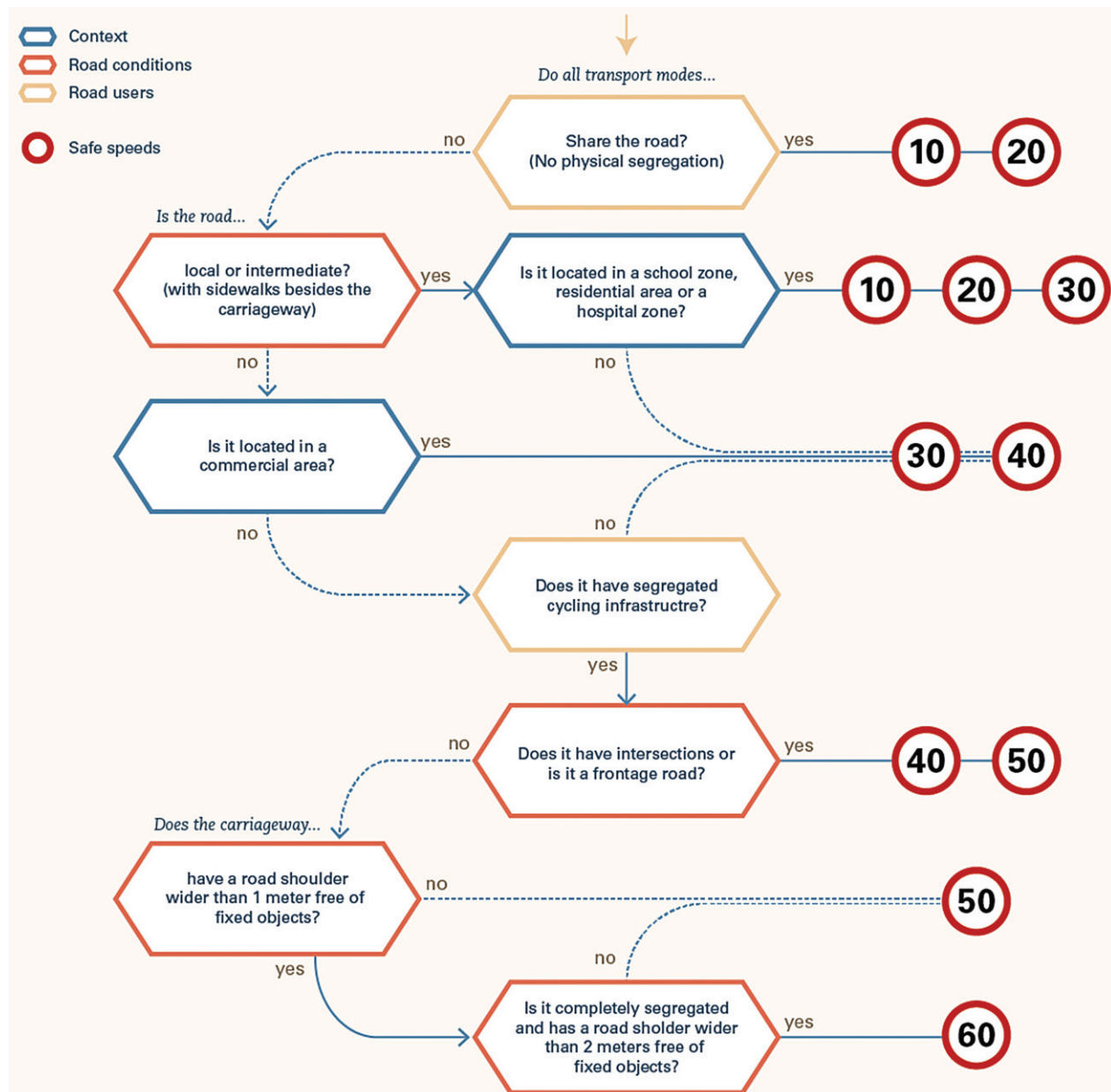


Figure 4 Flowchart for establishing speed limits based on local factors. Source: Graphic prepared by World Resources Institute for Alcaldía Mayor de et al., 2019. Translated from Spanish and adapted for the Elsevier Encyclopedia of Transportation.

lower the limit further depending on specific contextual details. If authorities are proactive about reviewing the default limits and adapting them to variations in the road features as necessary, a road function classification approach can facilitate the process of assigning appropriate speed limits and lowers the risk of inappropriate uses for roads. Because cities are dynamic, meaning that adjacent land uses and thus a road's purpose may change over time (e.g., urban expansion may result in a road that previous served a connecting purpose having housing built adjacent to it, adding an access purpose) road classifications should also follow a schedule of regular review and updating, in order to ensure that the classification and appropriate speed limit is appropriate to the function, or conversely, any issues with the road being used inappropriately (for example, commuters "rat running" through residential areas) are identified and addressed.

Road functions in urban areas can typically be divided into "through" functions and "access" functions. Through roads typically have a connector function, and are used to transport goods or people longer distances, a higher maximum speed limit is usually allocated to reflect this role. For example, in the Netherlands, a world leader in the Safe System approach, the maximum limit on such a road is 50 km/h. However, if the context of such a road on certain blocks is commercial, which attracts more pedestrians and generates crossing needs at different points of the road, then the speed limit in those areas should reflect that and be lowered to

40 km/h or less. The same applies for roads with an access function. These are locations where there is a concentration of residential, commercial, and/or institutional services. They should have maximum limits of 30–40 km/h. Access roads may be further broken down, either by formal classification or through other mechanisms such as design guidelines, into standard commercial or residential streets, shared streets, bicycle priority streets or play streets. Maximum or recommended speed limits may be set accordingly and be as low as 10–20 km/h in the case of streets with various types of shared status.

Benefits of Appropriate Speed Limits and Operating Speeds

Because speed is a contributing factor in the majority of crashes, setting safe speed limits and operating speeds is the fastest and most cost-effective way to reduce traffic fatalities and serious injuries, for all types of road users, but particularly people outside of cars. Moreover, the benefits of setting safe speed limits extend much more widely than this. Safe speed limits reduce both real and perceived risk for people biking and walking or using other human scale modes. The increased comfort generated reduces barriers to these active options, increasing accessibility, physical activity, and even play opportunities for children. In turn, this can generate a positive feedback loop whereby vehicle use is reduced, further reducing exposure to risk for people on the road, as well as reducing air and noise pollution, and advancing a positive feedback cycle of reduced private car use and increased use of more sustainable modes.

Reduced traffic fatalities and injuries and air pollution, and increased physical activity also have a beneficial economic impact as they reduce health costs. Furthermore, for cities battling major traffic issues, there is also emerging evidence that lowering speed limits reduces the conflicts and bottlenecks at intersection that contribute greatly to congestion. More consistent speeds and a reduction in acceleration and deceleration can both ease traffic flows and also reduce the emissions generated by cars.

Public Perceptions and the Politics of Speed Limits

The evidence of the safety benefits of taking a Safe System approach to road safety, and adjusting speed limits accordingly, is now well established. However, most countries are yet to build this knowledge into their default laws, regulations or planning policies, so there are few places where urban areas predominantly have a 30 km/h limit, and even fewer where it is the default.

This means that most places still require a review and implementation process to achieve safe speed limits in their urban areas. And, like most decisions involving public space, speed limits and use of streets are contentious and thus highly political matters, especially in urban and suburban areas that often face competing needs or demands between people commuting through them and living within them. For this reason, the local political economy must also be taken into account as it relates to setting, implementing and enforcing speed limits. Political leadership to prioritize road safety can play a major role in this process.

A common preoccupation when revising speed limits is the link between speed limits and travel times. The speed limit is frequently, but incorrectly seen as the key contributing factor to congestion in urban areas. Many people fear that lowering the speed limit in urban areas will increase journey times. However, average road speeds in cities are typically more determined by the volume of vehicles and the frequency of intersections than the speed limits. In many places, even if the speed limit of a road is 50 km/h, in a typical daytime or peak time speeds are in fact well below 30 km/h already, due to these factors. For the same reasons, if journey times are increased by the change in speed limit, it is typically by seconds rather than minutes. Conversely, emerging evidence from cities is demonstrating that lower speed limits can in fact ease congestion, by reducing traffic conflicts and bottlenecks at intersections.

After vehicle volume, distance from destination is the greatest determinant of travel time, not the speed limit. Due to a variety of factors such as the scale of urban areas, and vehicle-oriented planning or evolution into sprawl, people in all types of cities live at an increasing distance from their daily destinations, and as such are even more sensitive to anything that they perceive will increase their journey time. Applying a Safe System approach to road safety and mobility across the board can begin to address the preoccupation with journey times and congestion, because it also promotes more efficient land use arrangements that mix uses and encourage compact development, so that people live nearer to their frequent destinations. This way, safety can be improved by reduced exposure to vehicle travel or vehicles, alongside reduced speed limits.

One argument sometimes made against lowering speeds is that they can lead to higher fuel consumption. While it is true that traditional cars are the most fuel efficient at speeds of around 50 km/h (30 mph), constant low speeds of around 30 km/h (20 mph) are much more fuel efficient than the current typical urban situation of frequent acceleration and deceleration to stop at traffic signals or congested areas in order to reach 50 km/h between points. Furthermore, modern fleets are now transitioning towards hybrid and electric cars that typically get their best fuel consumption at speeds of around 30 km/h.

Cities that have successfully adapted their speed limits to safe levels for their current uses in recent years have avoided, reduced or addressed public backlash through a variety of means, including integrating speed limit changes with major physical and design upgrades of streets and public spaces, beginning with pilot projects such as “low speed zones” before scaling up the number of zones and extent of limits once the benefits are generated, developing integrated public outreach and enforcement campaigns so that drivers are more aware of the risks of speed and are exposed to verbal warnings from police for a “grace” period before new limits are enforced with fines, and conducting public consultation processes or community proposals on lowering speed limits.

Conclusion

As the concept of the Safe System for mobility is becoming more widely adopted, the concept of what safe speeds are in cities and who they should benefit has changed. There is now an increasing focus on protecting vulnerable users who typically make up most travel in cities and form the economic and social foundation of urban areas. This shift is generating measurable benefits in the places it is applied, both directly in terms of lives saved and injuries avoided, and indirectly in terms of other benefits such as increased physical activity, increased local economic activity, and improved quality of life.

As previously mentioned, establishing a safe speed limit in law at the national, state, or city level is the first step toward achieving safe speeds in urban areas, but it is not a sufficient measure on its own. Once a safe speed has been identified through a process of street classification and contextual analysis, and passed into law, regulation or policy, typically facilitated by a process of public outreach and awareness raising about the positive impacts of such a change, additional measures, specifically urban design and enforcement via police/automation or even vehicle features are necessary to ensure that driver behavior actually changes. Finally, while the changes necessary at policy, regulation, and street level are technical, they require a level of political leadership and commitment in order to be developed, introduced, and maintained, as well as to receive an adequate budget.

Permissions

Creative commons: Copyright 2018 World Resources Institute. This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of the license, visit <http://creativecommons.org/licenses/by/4.0/>.

See Also

Speed limits on rural highways; The Swedish Vision Zero – a policy innovation; Speed governors and limiters

References

- Alcaldía Mayor de Bogotá. Programa de Gestión de Velocidad. Documento Base/ City of Bogota Speed Management Program, Base Document (Spanish), 2019. Available from: <https://www.movilidadbogota.gov.co/web/sites/default/files/Paginas/2019-03-18/Programa%20de%20Gestión%20de%20la%20Velocidad%20para%20Bogotá.pdf>.
- Garber, N.J., Hoel, L.A., 2015. Traffic and Highway Engineering, fifth ed., Cengage Learning, Australia.
- Job, S., Sakashita, C., 2016. Management of speed: the low cost, rapidly implementable effective road safety action to deliver the 2020 road safety targets. J. Australas. Coll. Road Saf. 27 (2), 65–70.
- Tefft, B., 2011. Impact Speed and a Pedestrian's Risk of Severe Injury or Death, AAA Foundation for Traffic Safety. Washington, DC.
- Welle, B., Sharpin, A.B., Adiazola-Steil, C., Bhatt, A., Alveano, S., Obelheiro, M., Imamoglu, C.T., Job, S., Shotten, M., Bose, D., 2018. Sustainable and Safe: A Vision and Guidance for Zero Road Deaths. World Resources Institute, Washington, DC.

Further Reading

- Aarts, L., van Nes, N., Wegman, F., van Schagen, I., Louwerse, R., 2008. Safe speeds and credible speed limits (SaCredSpeed): a new vision for decision making on speed management. SWOV Institute for Road Safety Research, The Netherlands.
- Dumbaugh, E., Wenhao, L., 2011. Designing for the safety of pedestrians, cyclists, and motorists in urban environments, J. Am. Plan. Assoc. 77, 69–88.
- Greibe, P., 2005. Hastighedens Betydning for Trafiksikkerheden – Danske Og Udenlandske Studier. Dansk Vejtidskrift September. Available from: <http://asp.vejt看id.dk/Artikler/2005/09%5C4422.pdf>.
- Hydén, C., 1987. The Development of a Method for Traffic Safety Evaluation: The Swedish Traffic Conflicts Technique, Trafikteknik Tekniska Högskolan i Lund, Lund.
- Hydén, C., Linderholm, L., 1984. The Swedish traffic-conflicts technique. International Calibration Study of Traffic Conflict Techniques, 5, 133–139.
- Koorey, G., 2019. The mechanics and politics of changing a speed limit. Transportation Group 2019 Conference. Wellington, New Zealand.
- Sveriges Kommuner och Landsting, 2014. Traffic for an Attractive City—An Introduction to TRAST. Trafikverket, Borlänge.
- Tingvall, C., Haworth, N., 1999. Vision Zero: An ethical approach to safety and mobility, Papers of the 6th ITE International Conference – Road Safety and Traffic Enforcement, Victoria, Australia.
- Vejdirektoratet, 2013. Traffic regulations for urban areas in Denmark.
- WHO, 2008. Speed Management: A Road Safety Manual for Decision-Makers and Practitioners. WHO Press, Geneva.
- OECD Transport Research Center, 2006. Speed Management. Organization for Economic Cooperation and Development, Paris.

Speed-Reducing Measures

Vinod Vasudevan, Department of Civil Engineering, University of Alaska Anchorage, Anchorage, AK, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	641
Speed and Safety	641
Engineering-Based Measures	642
Traffic Calming Measures With Vertical Deflections	642
Speed Humps	643
Speed Bumps	643
Speed Cushions	643
Speed Tables	643
Raised Crosswalk	644
Raised Intersection	644
Traffic Calming Devices With Horizontal Deflections or Narrowing	644
Curb Extension	644
Chicanes	644
Lateral Shift	644
Curb Radius Reduction	645
Lane Narrowing	645
On-Street Parking	645
Raised Medians	645
Road Diet	645
Roundabouts	645
Other Infrastructure Improvements	645
Surface Treatment	646
Pavement Markings	646
Behavioral-Based Measures	646
Emerging Technologies and Measures	646
Active Vertical Deflections	646
GPS-Based Speed Governors	646
Mid-Block Traffic Signals With Speed Detection	647
Accommodating Nonmotorized Users	647
References	647

Introduction

The relation between speed and traffic safety is well established. Speed has been identified as a critical risk factor in road traffic injuries, influencing both the risk of a road crash and the severity of injuries. This knowledge influences speed limits for various functional classes of streets. As the accessibility of streets increases from freeways to local roads, possibilities for collisions also increase. Transportation agencies use lower speed limits to reduce the risk of severe injuries. Sometimes, setting speed limits alone would not help reducing vehicle speeds. In such cases, some interventions are required to reduce speeds. These could be engineering-based, or behavioral-based. The latter include enforcement and education. Combined, all these interventions are commonly known as speed-reducing measures.

This paper discusses the design, details of deployment, and effectiveness of some of the commonly used speed-reducing measures. This paper primarily references US-based design standards and documents.

Speed and Safety

Various studies have established that increasing the average speed by just 1.6 km/h (1 mph) typically increases the risk of crashes that result in injuries by about 4%. The same increase in speed adds 5% to the risk of fatal crashes. This happens because speed contributes to the severity of impact. For car occupants in a crash with an impact speed of 80 km/h (50 mph), the likelihood of death is 20 times what it would have been at an impact speed of 30 km/h (20 mph). The relationship between speed and injury severity is particularly critical for vulnerable road users, such as pedestrians and cyclists. For example, studies have shown that

pedestrians have a 90% chance of surviving when struck by a car traveling at 30 km/h (20 mph) or below, but that chance decreases to less than 50% at 50 km/h (30 mph). Pedestrians have almost no chance of surviving an impact at 80 km/h (50 mph).

Speed limits should be set such that the risk of fatalities from crashes is low, also see Article 10188, Speed Limits on Urban Streets; 10189, Speed Limits on Rural Highways. Typically, the risk of a fatality for the most critical crashes on that road category is set below 20%. For example, residential streets in the United States typically have speed limits of 40 km/h (25 mph). In Europe, urban street speed limits in the early 1950s were often 60 km/h (37 mph). Then they were reduced to 50 km/h (30 mph) and are now in many countries 40 km/h (25 mph) or 30 km/h (19 mph). In some cities, such as Berlin, many residential streets since 2017 have 10 km/h (6 mph) limits. In the United States, streets within school zones typically have a 25 km/h (15 mph) speed limit during times when children walk to and from them. The lower speeds intend to reduce the risk of pedestrian injuries. Similarly, most undivided urban collectors and arterials have speed limits of 50 km/h (30 mph) and sometimes 55 km/h (35 mph). Even 60 km/h (40 mph) and 70 km/h (45 mph) are used where there are no crosswalks and no pedestrians adjacent to the street. In Sweden, many rural undivided two-lane arterials used to have the speed limit 90 km/h (56 mph) but now many of those roads have been equipped with center cable barriers, and the speed limit increased to 100 km/h (62 mph) or, if they have not had barriers installed, the speed limit has been reduced to 80 km/h (50 mph) in order to reduce risks associated with head-on collisions. In other countries, 90 km/h (55 mph) speed limits are still allowed on undivided highways, and in some US states, even higher speeds are allowed.

However, speed limits alone will not always result in vehicle following posted speeds. Road design provides important cues to drivers about what constitutes a “safe” speed. And most drivers prioritize mobility and assume they will be safe at any speed, and reduce their speed more frequently to avoid tickets than to stay safe. The speed chosen by drivers is influenced by the road environment, including the road design and roadside infrastructure, and it is known as the perceived speed limit. If the posted speed limit and perceived speed limit do not match, drivers tend to deviate from the posted speed limit unless there is strict enforcement.

Driving at high speeds that are unsuitable for the prevailing road and traffic conditions has been responsible for many of the world’s vehicular injuries and fatalities. According to the World Health Organization (WHO, 2018), in some low-income and middle-income countries, speed is considered the primary contributing factor in about half of all road crashes. In high-income countries, in spite of availability to the resources to design safer roads and enforce traffic laws, speed still contributes to about 30% of deaths on the road according to some estimates. In the United States, for example, speeding has been at least partially responsible for about one-third of all vehicular fatalities over the last two decades. In Sweden, the authorities estimate that one-third of all fatalities would have been avoided in 2018 if no drivers had exceeded the existing speed limits, and more lives would have been saved if speed limits were reduced and enforced.

This goes to show that controlling speed can prevent crashes from happening and can reduce their severity when they do. However, many countries have introduced speed limits only to see them have a short-lived effect. Unless accompanied by sustained interventions, speed limits by themselves are often unable to curb speeding.

There are a few ways to reduce vehicle speeds, categorized mainly as engineering-based interventions and behavioral-based interventions. There are several resources available that discuss design, deployment strategies, and effectiveness of various speed-reducing measures. This paper refers to various resources by different agencies including the US Department of Transportation (USDOT, 2014, 2017, 2019), National Cooperative Highway Research Program (NCHRP, 2006, 2008, 2011, 2018), National Association of City Transportation Officials (NACTO, 2013), and the Institute of Transportation Engineers (ITE, 1999a, 1999b). The cost estimates mentioned for each of the speed-reducing measures (where available) are in 2017 US dollars.

Engineering-Based Measures

This category of speed-reducing measures introduces design changes to the roadway so as to enforce the desired speeds on it. One of the most popular engineering-based tools to curtail speeding is to implement traffic calming measures. Traffic calming measures are used either to reduce speeds or to reduce traffic volume. In this paper, only the measures used for reducing speeds are relevant. However, reducing the widths of streets from having two lanes in a direction to having only one lane is a fairly efficient speed-reducing measure in itself. As long as we have reasonably high volumes and a sizable proportion of drivers being law abiding, these drivers will reduce everybody’s speed to the limit. However, that strategy does not work when we have low traffic volumes or very few law-abiding drivers. Then other measures are needed, and such traffic calming measures primarily come in the form of two broadly defined types based on the design concept used to attain the desired speeds: (1) ones with vertical deflections, and (2) devices with horizontal deflections or narrowing. Due to their effectiveness and the relatively low cost of implementation and maintenance, traffic-calming measures are popular among road administrators and among residents living near them but they are often not very popular among commuters driving through areas with them. In addition to traffic calming, other engineering-based measures include active speed reduction techniques, such as speed governors (see Article 10517: Speed Governors and Limiters) and smart measures which are activated only when required.

Traffic Calming Measures With Vertical Deflections

Traffic calming devices with vertical deflections are the most common measures. Raising a small portion of a road surface creates discomfort for drivers traveling at high speeds. These devices reduce the speeds by inducing haptic vertical movement to the vehicles

passing over them and discomfort at speeds higher than the design speeds. The level of discomfort and optimum driving speeds depend primarily on geometry and type of vehicle traversing them. Examples of these measures include speed humps, speed bumps, speed cushions, speed tables, raised crosswalks, and raised intersections. Some of the commonly used ones are discussed as follows.

Speed Humps

A speed hump is a raised portion of the roadway surface extending transversely across the roadway. A common speed hump is Watts' hump, developed in England in 1973. It has a parabolic shape and a height of 10 cm (4 inches) and a length along the traveled way of 3.6 m (12 feet). The length of the hump is chosen so that a typical automobile has its front wheels move down at the same time as its rear wheels move up, increasing the feeling of vertical acceleration. Variations of this hump with slightly different dimensions, and only 75 mm height (3 inches), are sometimes used. In some cases, speed humps may raise the roadway surface to the height of the adjacent curb for a short distance. Humps can either extend the full width of the road, curb-to-curb, or be cut back at the sides to allow bicycles to pass and facilitate drainage. Notable speed reductions occur within 60 m (200 feet) of the speed hump, both downstream and upstream. A single Watts' hump reduces vehicle speeds to the range of 25–30 km/h (15–20 mph) at the hump. Vehicle speed increases at a rate of approximately 1 km/h (0.5–1 mph) for every 30 m (100 feet) outside of the 60-m (200-foot) approach and exit. This type of speed hump is considered appropriate if the posted speed limit is 50 km/h (30 mph) or lower (per ITE Guidelines for the Design and Application of Speed Humps), though some suggest a speed limit of 40 km/h (25 mph) as the maximum value. Speed humps in the United States are not considered appropriate if the 85th percentile speed prior to implementation is 45 mph or more. If the objective is a point speed control, then only a single hump is required. A series of humps are usually needed to reduce speeds along an extended section. Typically, speed humps are spaced at between 100 and 200 m (300 and 600 feet) apart. In many jurisdictions, speed humps are marked, and painted with white stripes to be noticed by drivers. In some countries, notably Norway, the preference is to have the speed hump blend with the pavement color and only advise drivers at the beginning of a street that there are speed humps. The idea is that drivers will not increase speeds as much in between the humps when they cannot easily see where the next hump is. Speed humps are low-cost measures with low maintenance. The typical cost of deployment of a speed hump ranges from \$2000 to \$4000.

Speed Bumps

A speed bump is similar to a speed hump but much shorter in direction of travel. A typical speed bump is 8–15 cm (3–6 inches) high and 0.3–1 (1–3 feet) long. Like speed humps, speed bumps extend transversely across the roadway. Due to their lower breadth, they cause higher discomfort to motorists than speed humps traversing them at speeds up to about 25 km/h (15 mph) and generally results in vehicles slowing to 8 km/h (5 mph) or less at each bump. However, if they are traversed at speeds above 50 km/h (30 mph), they do not cause any discomfort but may damage the suspension of a vehicle. Due to their lack of efficiency in reducing higher speeds, and extreme discomfort at modest speeds, they are not typically deployed on public roads but are often used on private property, such as parking lots, private streets, and within apartment complexes. The effect on driver speed is similar to the effect imposed by speed humps. Like speed humps, to reduce speeds along an extended section of the street, a series of bumps are usually needed, generally spaced between 100 and 200 m (300 and 600 feet) apart. Like a speed hump, a speed bump also is a low-cost measure with low maintenance. The typical cost of deployment per speed bump is very close to that of speed humps.

Speed Cushions

Speed cushions are modifications of speed humps. One of the significant drawbacks of speed humps is that they increase emergency response times. Speed cushions address this concern by allowing passages of larger vehicles without any vertical deflection. The primary difference is that a speed cushion has wheel cutout between the raised areas to enable a vehicle with a wide track, such as emergency response vehicles and buses to pass through without much discomfort. While most dimensions are similar to speed humps, it is common to design the top of speed cushion to be level, as opposed to parabolic or circular for speed humps. This is preferred on primary emergency routes or along transit routes with frequent service. The overall effectiveness of speed cushions is similar to that of speed humps. Therefore, spacing and other design specifications are the same as those recommended for speed humps. Speed cushion is a low-cost measure with low maintenance. The typical cost of deployment of a speed cushion ranges between \$2500 and \$6000, slightly more than the cost of a speed hump since it requires more material than a speed hump.

Speed Tables

Speed tables are another version of speed humps. The major difference is that the top of the speed table is flat instead of parabolic. Speed tables are typically used to slow drivers in the middle section of a city block or on approaches to multilane roundabouts. They are usually longer than speed humps, with a height of 7.5–10 cm (3–4 inches) and a length of about 7 m (22 feet). Slopes should not exceed 1:10 at entry and less steep than 1:25 at exit. Side slopes on tapers should be no steeper than 1:6. The most common speed table consists of a 3-m (10-foot) plateau with 2-m (6-foot) approaches on either side, which can be straight, parabolic, or sinusoidal in profile. Speed tables are not recommended on streets wider than 15 m (50 feet). Speed tables are typically paved using distinctive material, both in structure and in colors, such as brickwork, to attract driver and pedestrian attention. A single speed table can reduce 85th percentile speeds to a range of 40–55 km/h (25–35 mph), though the effect declines at a rate of approximately 1–1.5 km/h (0.5–1 mph) for every 30 m (100 feet) beyond the 200-foot approach and exit. Like speed humps, speed tables in the United States are not considered appropriate if the 85th percentile speed prior to implementation is 70 km/h (45 mph) or more. Compared to this practice, in northern Europe, crosswalks are not used if the speed limit is above 80 km/h (50 mph), and not recommended unless the

speed limit and actual speed are below 40 km/h (25 mph) and speed tables would typically not be used unless there are crosswalks. This means that if the 85th percentile was 70 km/h (45 mph), it would have to be reduced to below 50 km/h (30 mph) using speed tables or other devices. A speed table is considered as a medium-cost measure with a typical cost of deployment ranging between \$2500 and \$8000. The additional cost is due to the requirement for more materials.

Raised Crosswalk

Raised crosswalks are an extension of speed tables when placed near a mid-block pedestrian crossing. The breadth of the raised crosswalk remains the same as that of speed tables, but in most cases, they are slightly higher, with a height of 8–15 cm (3–6 inches) above street level, so that the top aligns with sidewalks. While all other design elements remain the same, raised crosswalks also require all of the standard crosswalk design elements, such as signs and markings. The effectiveness in speed reduction is very similar to that of speed tables. To make raised crosswalks pedestrian-friendly, they are often deployed in conjunction with curb extensions, which are discussed later in this paper. Raised crosswalk is considered as a medium-cost measure with a typical cost of deployment ranging between \$4000 and \$8000.

Raised Intersection

Raised intersections are another extension of speed tables. They are employed when speeds need to be reduced on two intersecting streets. Raised intersections have ramps on all approaches, leading to a flat, raised area that covers entire intersections. Like a raised crosswalk, a raised intersection typically rises to sidewalk level. They improve pedestrian safety at busy intersections and are often found in commercial business centers and residential areas. Like speed tables, raised intersections are typically paved with distinctive material. Raised intersections are appropriate on streets with a speed limit of 30 mph. The raised intersection is considered as a high-cost measure with a typical cost of deployment ranging between \$15,000 and \$60,000.

Traffic Calming Devices With Horizontal Deflections or Narrowing

Traffic calming devices with vertical deflections are effective at reducing speeds, but they are highly inconvenient to drivers and cause problematic discomfort for individuals with health concerns, such as elderly or disabled people. Horizontal deflections or narrowing can address these concerns. Horizontal deflections reduce speed by forcing vehicles to swerve slightly, while narrowing reduces the priority given to vehicles by limiting roadway area. This category of speed reduction techniques includes any technique that reduces the area of a street designated exclusively for motor vehicle travel. The additional reclaimed space is typically used to improve safety or attract other modes of transportation, primarily pedestrians. Commonly used techniques are curb extensions, chicanes, lateral shifts, curb radius reductions, lane narrowing, on-street parking, raised medians, and road diets. Some of the commonly used ones are discussed as follows.

Curb Extension

Curb extensions visually and physically narrow the roadway, creating safer and shorter crossings for pedestrians and increasing street-side space for public features, such as benches, shrubs, and trees. They are most often used in downtown neighborhoods and residential streets. Curb extension is an umbrella term that encompasses several different treatments and applications. Mid-block curb extensions, known as pinch points or chokers, may include a cut-through for bicyclists. Curb extensions used as gateways to minor streets are known as neckdowns. While design speed limit varies concerning the dimension and width of roadway, curb extensions in the United States are typically used on roads with speed limits of 55 km/h (35 mph) or less. Curb extension is considered as a medium-cost measure with a typical cost of deployment ranging between \$8,000 and \$12,000. However, sometimes they can cost over \$40,000 if drainage alteration is required. Curb extensions can be combined with raised medians (see later) and if the distance between the curbs is not more than 2.4 m (8 feet), they have the added benefit of slowing down traffic significantly. In London, England, spacing of 2.1 m (7 feet) giving speed around 5 km/h (3 mph) has been used at entries to intersections but then fire trucks and other wide vehicles need to use the exit side to get into the intersection which could cause head-on collisions.

Chicanes

Chicanes are speed-calming devices that sharply alter horizontal alignment over a small distance, forcing motorists to steer back and forth out of a straight travel path. Generally, chicanes are created of curb extensions that alternate from one side of the street to the other. Sometimes, on-street parking is used instead, but it may not be as effective when no vehicles are parked due to the lack of a physical imposition. A shift of one lane's width or a deflection angle of at least 45 degree is recommended to prevent drivers from maintaining their speed and alignment. For a given length of the chicane segment, as the deflection angle increases further, drivers are forced to reduce speeds further. Chicanes are appropriate for streets with a maximum speed limit of 55 km/h (35 mph). They are typically not necessary along primary access routes to commercial or industrial sites. Chicane is considered as a medium- to high-cost measure with a typical cost of deployment ranging between \$8,000 and \$25,000.

Lateral Shift

A lateral shift is the realignment of an otherwise straight street that causes travel lanes to shift in one direction. A typical lateral shift separates opposing traffic through the shift with the aid of a median. Without medians, drivers can cross centerlines to achieve the straightest possible path, thereby eliminating the need to reduce speeds. Medians also reduce the likelihood that motorists will veer

into oncoming traffic, further improving roadway safety. Typically, a lateral shift is deployed in conjunction with on-street parking. Similar to chicanes, lateral shifts are appropriate for streets with a maximum speed limit of 35 mph, but not along primary access routes to commercial or industrial sites. Lateral shift is considered as a medium- to high-cost measure with a typical cost of deployment ranging between \$8,000 and \$25,000.

Curb Radius Reduction

A wide curb radius at intersections typically results in rapid turning movements by motorists. By reducing the curb radius and causing tighter turns, drivers must reduce their right-hand turning speeds to below 20 km/h (12 mph). This also limits crossing distances for pedestrians. While standard curb radii are 3–4.5 m (10–15 feet), many cities use corner radii as small as 60 cm (2 feet) because of the high volume of vehicle and pedestrian traffic. A curb radius of 60 cm (2 feet) causes drivers to reduce their turning speed to less than 15 km/h (10 mph). However, this technique is not suitable in areas with a large number of turning trucks or buses, since raised aprons would be needed and pedestrians may get in conflict with trucks if standing on the apron. The cost of deployment of this measure varies significantly between \$2,000 and \$20,000 per site based on site condition.

Lane Narrowing

Wide streets encourage speeding, especially during off-peak hours, which makes lane narrowing an attractive technique. Studies have shown that a 30-cm (1-foot) lane reduction can reduce average speeds by as much as 1.6 km/h (1 mph), but it can also increase the number of collisions by 3%–5%. Therefore, lane narrowing should be implemented with caution and primarily only on streets with one lane in each direction and ideally only when there is a continuous raised median. This is one of the measures with a wide range of costs. The cost of deployment varies significantly, depending on various factors. For example, adding striped shoulders will cost only about \$1,000 per mile, whereas adding on-street parking, in addition to adding a striped bike lane, will cost between \$5,000 and \$10,000 per mile, depending on the number of lanes to be removed. If constructing physical infrastructure such as raised medians or improvement of sidewalks are included, the cost can go \$100,000 or more per mile.

On-Street Parking

On-street parking reduces available lane widths. This reduction in lane width reduces speeds by adding side friction to the traffic flow. On-street parking can be located on one or both sides of a roadway and can be located on alternating sides to create a chicane effect, as previously discussed. The result can be a speed reduction of 1.6–8 km/h (1–5 mph), with a reduction of 3–5 km/h (2–3 mph) being the most common. However, on-street parking poses serious safety concerns for pedestrians entering and exiting their vehicles. On-street parking is a low-cost measure. The cost of deployment of an on-street parking scheme varies depending on the length of applications and design specification. However, the typical cost of deployment is less than \$6000 for parallel parking. If diagonal parking is used, it should be of back-in, drive-out type.

Raised Medians

A raised median creates a barrier between opposing lanes of traffic. While raised medians reduce lane width and thereby reduce driver speed, they also provide other benefits, such as the lower risk of head-on collisions and the opportunity to create protected pedestrian refuges in the middle of crosswalks. The raised median is considered as a low- to medium-cost measure with a typical cost of deployment ranging between \$1,500 and \$10,000, depending on the length and width of islands.

Road Diet

Road diets have been described as “removing travel lanes from a roadway and utilizing the space for other uses and travel modes.” For example, the conversion of an undivided four-lane roadway into an undivided three-lane roadway made up of two unidirectional lanes, and one two-way left-turn lane would cause drivers to reduce their speed. The reduction of lanes can also be used for other purposes, such as bike lanes, pedestrian refuge islands, public transit, and on-street parking. A road diet is considered as a low-cost measure. Although the cost of deployment of a road diet scheme varies depending on the geometric changes and length of applications, the typical cost of deployment is less than \$6000.

Roundabouts

Single-lane roundabouts with good designs reduce 85th percentile speeds to 27 km/h (17 mph). Multilane roundabouts often need to be combined with speed humps or speed table to reduce top speeds to anything below 40 km/h (25 mph) during low-volume times when drivers can straddle lanes; but even that is a lower speed than typical at many traffic signals. Roundabouts can be used even if there is very little cross-traffic but motorists prefer them at busy locations where they not only slow down traffic, but also typically reduce average travel times significantly compared to signalization.

Other Infrastructure Improvements

In addition to traditional traffic calming measures such as vertical and horizontal deflections, there are other tools to reduce speeds. However, most have a limited effect and are, therefore, not discussed in detail. Two established tools are surface treatment and pavement markings.

Surface Treatment

Studies have shown that by changing the surface texture, speed can be reduced. Textured crosswalks and textured pavement are somewhat frequently adopted on urban streets with speed limits below 55 km/h (35 mph) with the aim of promoting nonmotorized activities. Transverse rumble strips that cause vehicles to vibrate are used on freeways and other high-speed facilities to warn drivers of impending speed limit reductions or toll plazas. Unfortunately, the implementation and maintenance of surface treatments can be costly and, like vertical deflections, may cause problematic discomfort for elderly and disabled people.

Pavement Markings

Pavement markings provide visual cues to reduce speed. Examples include converging chevrons, dragon teeth, full-lane transverse bars, on-road “signs,” and peripheral transverse bars. Studies have shown mixed effectiveness that is highly site-specific. Pavement markings are considered as a low-cost measure, although it requires regular maintenance. The cost of deployment of pavement markings varies depending on the type of application and design specification. However, the typical cost of deployment is less than \$6000.

Behavioral-Based Measures

As the name suggests, this category of interventions tries to achieve the desired speeds by influencing driving behavior. Examples include education and enforcement measures. Enforcement has shown a significant impact on speed reduction. These include manual enforcement or automated enforcement. Manual enforcement is costly, and it will lose its effects within a few days of completion unless regularly repeated. Therefore, automated enforcement is a better option, especially if it is area wide, for example, measuring the average speed between two adjacent cameras, see Article 10167, Photo/Video Traffic Enforcement. Education/awareness can also play a role in speed reduction at least if people’s attitudes are modified, see Article 10128, Education and Training of Drivers. In addition to media campaigns, measures such as speed trailers have shown to help reduce speeds, at least among drivers who were unaware that they were speeding, especially on low-speed facilities. Other than the traditional driver educational programs and enforcement initiatives, deployment of tools such as cameras automated enforcement and portable speed trailers to alter driver behavior have also shown short-term effectiveness in reducing speeds. The cost of educational and enforcement campaigns depends on the duration and details of campaigns. However, a typical speed trailer cost between \$8,000 and \$20,000 depending on the available messaging options. The speed cameras that are used extensively all over Sweden cost about \$45,000 each to install and then have operational costs as well.

Emerging Technologies and Measures

Recent technological advances, especially in sensors, are beginning to allow for traffic safety innovations. Some emerging technologies include active bumps, GPS-based speed governors, and mid-block traffic signals with speed detection.

Active Vertical Deflections

As discussed in the previous section, measures with vertical deflections such as speed humps are incredibly useful in reducing speeds. However, they are not popular due to the discomfort they cause even at low speeds. One way to address this issue is to develop active vertical deflections. These devices use sensors such as radars or on-road sensors to detect the speed of oncoming vehicles and activate the deflections only if the approach speed exceeds the posted speed limit. There are a few of these measures that are in conceptual or deployment stages. Some of the examples include Edeva, Smart Bumps, and Dynamic Rumble Strip System. Edeva is a trapdoor bump that raises for speeders (Edeva, 2019). Smart Bumps are angled vertical deflections that adjust based on driver speed (Smartbumps, 2019). Dynamic Rumble Strip System helps to alert the drivers when they pass this at high speeds using a set of closely placed trapdoor bumps with smaller widths (Paz, 2018). It is important to keep in mind that the capital costs of these devices would be high and would require electricity to operate. Also, since these tools use sensors, they require regular maintenance. In spite of these drawbacks, they offer a serious prospect to address some of the challenges that traditional traffic calming devices pose. While there is little field evidence to evaluate their effectiveness, active vertical deflections have the potential to offer results that are similar to their counterparts without actuation.

GPS-Based Speed Governors

Speed governors are physical devices used to measure and regulate vehicle speed. Governors have been in use for several decades, mainly on commercial vehicles. The devices prevent drivers from exceeding a set maximum, usually 105 km/h (65 mph). Such governors are mandated on trucks in much of the EU and in several Canadian provinces. However, most governors currently allow for only a one-speed setting. GPS-based speed governors, a recent innovation, adjust the maximum speed based upon the posted speed limit. Currently (May 2019), there is a proposed plan in Europe backed by Members of the European Parliament and the European Transport and Safety Council to introduce mandatory GPS speed limiters on new model vehicles starting May 2022 and on

new face-lifted models starting May 2024. It has been estimated that this system will reduce collisions by 30% and save 25,000 lives within 15 years of being introduced.

Mid-Block Traffic Signals With Speed Detection

This is an innovative signal that has been in use in the United Kingdom. Mid-block traffic signals remain green until a pedestrian activates them or until a vehicle approaches at speed 8 km/h (5 mph) or more above the posted limit, forcing drivers to stop completely. These devices, combined with red-light running cameras, have the potential to be highly effective in reducing vehicle speeds.

Since the measures discussed under this section are not yet available for commercial deployment, the costs of these measures are not readily available.

Accommodating Nonmotorized Users

As the primary objective of speed-reducing measures is enhanced safety of all road users, these are extensively used on residential roads, in school zones, and where high nonmotorized activities are expected. The level of haptic feedback and ideal operating speed depends on the type of traffic calming device deployed and its design. The human body can accept vibration energy up to a certain level above that repeated exposures pose a health hazard. Since most of the bicycles do not have proper suspension systems, a study by Patel and Vasudevan (2016) concludes that bicyclists experience higher levels of discomfort than those experienced by motorized vehicles while passing over speed-reducing measures with vertical deflections. The same would be true with altering surface types. Since traffic-calming measures with horizontal deflections do not induce vibrations, these would be appropriate in locations where high bicycle activities are expected so that these devices are not producing unintended consequences. Roundabouts are not nonmotorized user-friendly since the crossing facilities are placed away from the roundabout. Therefore, in areas where pedestrian presence is high, roundabouts are not recommended. Also, it is worth noting that the measures with horizontal deflections may not be as effective in reducing speeds when compared with the ones with vertical deflections. Therefore, choosing the right type of traffic calming measure requires a fine balancing of these two contradicting properties based on the site-specific requirements.

References

- Edeva, 2019. Edeva traffic systems for a livable city. Available from: <https://www.edeva.se/en/>.
- ITE, 1999a. Traffic Calming State of the Practice. Institution of Transportation Engineers, Washington, DC.
- ITE, 1999b. Traffic Safety Toolbox: A Primer on Traffic Safety. Institution of Transportation Engineers, Washington, DC.
- NACTO, 2013. Urban street design guide. National Association of City Transportation Officials, New York, USA. Available from: <https://nacto.org/publication/urban-street-design-guide/>.
- NCHRP, 2006. Improving pedestrian safety at unsignalized crossings. NCHRP Report 562. National Cooperative Highway Research Program, Washington, DC, USA.
- NCHRP, 2008. Guidelines for selection of speed reduction treatments at high-speed intersections. NCHRP Report 613. National Cooperative Highway Research Program, Washington, DC, USA.
- NCHRP, 2011. Speed reduction techniques for rural high-to-low speed transitions. NCHRP Synthesis 412. National Cooperative Highway Research Program, Washington, DC, USA.
- NCHRP, 2018. Design guide for low-speed multimodal roadways. NCHRP Report 880. National Cooperative Highway Research Program, Washington, DC, USA.
- Patel, T., Vasudevan, V., 2016. Impact of speed humps of bicyclists. *Saf. Sci.* 89 (2016), 138–146.
- Paz, A., 2018. Prototyping and field testing of a demand-responsive rumble strip mechanism. NDOT Final Research Report, Report No. 224-14-803 TO17. Carson City, Nevada, USA.
- Smartbumps, 2019. Smart bumps. Available from: <http://smartbumps.com/about/>.
- USDOT, 2014. A desktop reference of potential effectiveness in reducing speed. US Department of Transportation, Federal Highway Administration, Washington, DC, USA. Available from: https://safety.fhwa.dot.gov/speedmgt/ref_mats/eng_count/2014/reducing_speed.cfm.
- USDOT, 2017. Traffic calming ePrimer. US Department of Transportation, Federal Highway Administration, Washington, DC, USA. Available from: https://safety.fhwa.dot.gov/speedmgt/traffic_caln.cfm.
- USDOT, 2019. Reference materials for speed management. US Department of Transportation, Federal Highway Administration, Washington, DC, USA. Available from: https://safety.fhwa.dot.gov/speedmgt/ref_mats/.
- WHO, 2018. Global status report on road safety 2018. Management of Noncommunicable Diseases, Disability, Violence and Injury Prevention (NVI). World Health Organization, Geneva, Switzerland. Available from: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/.

Striping, Signs, and Other Forms of Information

Kasem Choocharukul, Kerkritt Sriroongvikrai, Infrastructure Management Research Unit, Department of Civil Engineering, Faculty of Engineering, Chulalongkorn University, Pathumwan, Bangkok, Thailand

© 2021 Elsevier Ltd. All rights reserved.

Overview	648
Applications	648
Safety Benefits	651
Issues to Consider	652
Retroreflectivity	652
Maintenance	653
Redundant Signs and Striping	653
Human Factors	653
Uniformity	653
Forgiving Devices and Self-Explaining Roads	654
Concluding Remarks	654
See Also	655
References	655
Further Reading	655

Overview

Traffic signs and striping are essential to highway and street designs and operations. They are commonly utilized to communicate with road users. Appropriate installation of striping and signs contribute to improved road safety and mobility along a roadway. This chapter describes typical applications of striping and signs, along with their safety benefits. Relevant issues from road safety perspective are also discussed.

At present, several countries have adopted the Vienna Convention on Road Signs and Signals, while traffic engineers and transportation planners in the United States (US) consider guidelines such as the Manual on Uniform Traffic Control Devices (MUTCD) in designing and selecting appropriate traffic signs and control devices for their road network. According to MUTCD, a traffic control device is defined as “a sign, signal, marking, or other device used to regulate, warn, or guide traffic, placed on, over, or adjacent to a street, highway, private road open to public travel, pedestrian facility, or shared-use path by authority of a public agency or official having jurisdiction, or, in the case of a private road open to public travel, by authority of the private owner or private official having jurisdiction (FHWA, 2009).”

In principle, traffic signs can be categorized into three groups serving three distinct functions, that is, regulatory signs, warning signs, and guide signs. Regulatory signs are used for statutory requirements according to traffic law and regulation, while warning signs alert drivers to hazardous or potentially hazardous situation that might not be easily apparent. The guide signs are typically aimed to provide relevant information on the directions or information specific to destinations or points of interest. Traffic signs are varied in dimension, symbol, color, and size. **Table 1** presents examples of traffic signs commonly found along roadways.

Road striping and lane marking is normally used to delineate a roadway path and separate traffic movements in either the same or opposing directions. A proper installation of road striping and lane marking can reduce several crash type risks, particularly sideswipe, head-on, and run-off road crashes. The effectiveness of striping is enhanced for people driving vehicles with lane departure alert systems. Striping can also be used to warn drivers of potential hazards ahead or conflict points along the roadway. Other applications include Chevron markings and word markings, such as STOP AHEAD. Rumble strips and audio tactile line markings differ from other signage and lane markings in way that they provide not only visual information but also vibratory and audible warnings. **Table 2** illustrates various types of road striping and pavement markings. All longitudinal, transverse, and other markings are common and are designed to provide drivers information necessary for efficient and safe driving.

Applications

The primary purpose of traffic signs is to regulate, warn, or guide road users along the roadway. Failure to comply with traffic control devices could lead road users to be fined as well as have crashes that may cause injuries and even death. In general, drivers are required to follow local traffic laws and regulations, which are frequently posted on regulatory traffic signs such as speed limit, travel direction, and restriction of certain vehicle types on the road. Likewise, warning signs are utilized to warn road users about approaching situations on the road such as changes in road geometry or risk conditions in the surrounding area. A proper warning sign should inform drivers that they can safely maneuver their vehicles to intended destination and avoid confusion at any conflict

Table 1 Types of traffic signs

Category	Examples	Typical sign sheets
Regulatory sign	Stop, yield/give way, speed limit, One way, No parking, Do not enter	
Warning sign	Intersection warning, School zone warning, Crossings, Slippery when wet, Speed bump	
Guide sign	City/Town name, Route/Highway number, Destination and distance	

Table 2 Types of road striping and pavement markings

Category	Examples
Striping/longitudinal line marking	Center lines, lane lines, edge lines
Pavement marking	Stop bars, crosswalks, arrows
Others	Chevron marking, Rumble strips, audio tactile line markings, word markings

points such as intersections or crossroads. Ideally, warning signs should be displayed only during the times when they are needed. For example, warning drivers of freezing roadways at bridges during summer-weather conditions makes little sense.

From the MUTCD guidelines, traffic control devices should fulfill a need, command attention, convey a clear, simple meaning, command respect from road users, and give adequate time for proper response. Therefore, traffic engineers should carefully select relevant types of traffic signs and information. However, in many jurisdictions that is not the case. For example, in many US states, stop signs are frequently used at intersections where fewer than 20% of drivers come to a full stop and a sizable percentage of drivers do not even slow down. Stop signs are still used because they supposedly send a message to drivers that they do not have the right-of-way at that location and need to yield to other traffic; and many drivers do not seem to understand that a yield sign could convey that information. But if stop signs are not respected at most locations, installing a stop sign where a full stop is needed will not have the intended safety effect. **Fig. 1** illustrates various road settings in some selected countries showing traffic signs and road striping.

The treatment of traffic signs on the road can be affected by several factors, including facility type, traffic volume, and characteristics, as well as economic benefit and cost consideration. A conventional octagonal stop sign, for example, is in the United States



(A) Australia



(B) Japan



(C) Russia



(D) Thailand

Figure 1 Various road settings showing traffic signs and road striping.

normally recommended on a minor road entering a designated through highway with traffic volume greater than 6000 vehicles per day. Alternatively, MUTCD states that the stop sign can also be considered when there exists a strong evidence of historical crash records particularly right-angle crashes from drivers on the minor-street approach. If this rule was followed, we may get more respect for stop signs than following the current prevailing practice—in most US states—to put stop signs at every intersection.

Placement of traffic signs also directly influences visibility of road users. Therefore, the minimum required visibility distance should be justified. Such a distance is a function of speed and depends on type of traffic sign. As an example, an intersection warning sign should be placed in advance of the approaching intersection, and the advance placement distance should, according to MUTCD, be computed based on the 85th percentile speed and suitable sign legibility distance. Placement of warning signs should also reflect the perception-response time (PRT) to ensure that drivers have adequate time to detect, comprehend, and respond to traffic signs safely.

The main purpose of road striping and pavement markings is to specify the designated maneuvering area of road users and to guide them in the direction of travel ahead. In addition, striping and markings provide guidance and warning to road users in terms of road rules such as lane changing, overtaking, and intersection crossing, as well as merging into a traffic stream. Road striping and pavement markings may also reduce crash risks and increase mobility associated with poor visibility. One hypothesis states that with appropriate road striping, drivers who maneuver during inclement weather or darkness could safely judge the alignment of the road and avoid being involved in crashes. In reality though, many studies show that nighttime crashes go up as a result of better lane guidance since it leads to people driving faster and they still do not see pedestrians, animals, and crossing traffic at any further distance than they would without that guidance.

Striping is commonly applied to designate travel lanes, showing desired path and direction of travel. A typical road has edge line striping that delineates the left or right edge of a traveled way. Edge line striping can also be considered for roadways with different usages such as pedestrian walkway and bicycle facility. However, it needs to be remembered that striping by itself gives very little protection in separating such road users from automobiles. For shared-use paths or undivided roads, centerline striping can be seen as a type of median. Again, it has to be stressed that a stripe by itself does not guarantee that drivers do not cross over it.

In addition to providing road users information for maneuvering on a road, striping and pavement signs provide potential benefits in terms of road safety, not only to drivers but also to vulnerable road users such as pedestrians and bicyclists if they have to share the road facility. Edge striping without the striping of centerlines can be used to slow down motorists if one makes the travel roadway very narrow, say around 6 m. This should only be used in urban areas with speeds of 40 km/h or less. Without the

centerline, motorists fear that they may have a head-on collision and will slow down significantly whenever there is oncoming traffic. However, this will only work if they cannot encroach onto the edge stripes, so the edge stripes need to be complemented with vertical delineators. Uncomfortable rumble strips can also be used to achieve this effect but have the disadvantage of creating noise, which may be unacceptable in residential areas.

Striping and pavement markings can be installed by several materials such as paint, thermoplastic, or pre-cut sheeting. Different patterns and colors of striping and pavement markings denote different communication messages to road users. For instance, a solid continuous white line striping indicates in the United States the edge of the roadway, while a broken white line striping separates travel lane in the same travel direction and that road users may cross this line when changing lanes. Similarly, double yellow lines in the United States prohibit passing in either direction unless it is the passing of a bicyclist or pedestrian whereas making a left turn across it into a driveway is permitted. On the other hand, a double solid white or yellow line in many European countries means that it is illegal to cross the line for passing or making a left turn. Application of striping and pavement markings today follow local guidelines and standards to ensure that road users can perceive and comprehend correctly and perform reasonably safe driving accordingly. It probably would be better if we had international standards so that striping gave the same message everywhere so that tourists and other visitors understand the meaning.

Safety Benefits

Information that drivers receive from properly installed and displayed striping and traffic signs help them make informed decisions on how to maneuver their vehicles safely under prevalent roadway and traffic conditions. Although there are only limited past studies showing tangible benefits in terms of crash risk reduction after striping or typical traffic signs are installed, it is widely accepted as being beneficial to road users and critical to a safe road design and standards. And, surveys of drivers indicate that improved striping belong among the most popular measures a road administration can undertake to please its drivers. Benefits of traffic signs are more profound for drivers approaching difficult driving tasks such as ahead of a complex curve, in bottleneck locations, or in an area where operating speed should be substantially reduced due to sudden change in road geometry or roadside environment. Under these circumstances, provision of warning traffic signs would potentially reduce severities and number of crashes and thus make the road safer. However, drivers often ignore such warning signs, and compound curves on highway ramps that start out with a gentle curve and then transitions into a sharp curve often see many crashes even after large chevron signs are erected. Changing the roadway geometry and making the curve gentler, or at least using a constant radius and having drivers see the curvature of the ramp before they enter the curve—so called self-explaining roads—may be more effective than signage.

For striping and pavement markings, centerline and edge line markings can be regarded as a low-to-moderate cost treatment that provides some crash reduction potentials, especially in situations with complex geometry. Additional benefit can be gained in certain road types. For instance, a high-speed, two-way undivided rural road with clear centerline markings may have a lower likelihood of head-on crashes than an unmarked roadway. However, installing centerline rumble strips cut fatalities from head-on collisions with opposing vehicles to roughly half of what is experienced with good striping alone, and installing cable barriers in the center of the road cuts most of the remaining ones.

A crash modification factor (CMF) is customarily utilized as a key effectiveness measure for selecting appropriate road safety treatments and countermeasures. By definition, the CMF is the ratio between the estimated number of crashes expected after implementing the road safety treatment and the estimated number of crashes without the treatment. A CMF value less than one indicates a sound treatment that is expected to reduce crash numbers. CMFs can be derived from a systematic before-after design, of which the safety performance of a group of locations is compared with the treatment of interest to similar locations without the treatment.

Table 3 lists potential crash effects for traffic sign installation based on the current Highway Safety Manual (AASHTO, 2014). Depending on type of setting and crash type, the CMF values for warning and speed sign treatments range from 0.54 to 0.87, signifying approximately 13%–46% reduction in estimated crash frequency after these measures are implemented.

It should be noted that the studies referenced for these effects were published in 1959 through 1971 and have been criticized for not accounting for regression-to-the-mean effects accurately. It would be useful if newer studies were conducted. Likewise, **Table 4** presents corresponding values for pavement markings, which range from 0.55 to 1.05. These CMF values may vary upon various circumstances and could only be used as a guideline for implementing road safety treatments in the study area. And it should also be

Table 3 Potential crash effects of treatments related to roadway signs

<i>Treatment</i>	<i>Setting</i>	<i>Crash type</i>	<i>CMF</i>	<i>s.d.</i>
Install combination horizontal alignment/advisory speed signs	Unspecified	All types (Injury)	0.87	0.09
Install changeable crash ahead warning signs	Urban (Freeways)	All types (Injury)	0.56	0.20
Install changeable “Queue Ahead” warning signs	Urban (Freeways)	Rear-end (Injury)	0.84	0.10
Install changeable speed warning signs for individual drivers	Unspecified	All types (All severities)	0.54	0.20

Source: AASHTO (2014).

Table 4 Potential crash effects of treatments related to pavement markings

<i>Treatment</i>	<i>Setting</i>	<i>Crash type</i>	<i>CMF</i>	<i>s.d.</i>
Install post-mounted delineators	Rural (Two-lane undivided)	All types (Injury)	1.04	0.10
Place standard edge line marking	Rural (Two-lane)	All types (Injury)	0.97	0.04
Place wide (8 inches) edge line markings	Rural (Two-lane)	All types (Injury)	1.05	0.08
Place center line markings	Rural (Two-lane)	All types (Injury)	0.99	0.06
Place edge line and center line markings	Rural (Two-lane/Multilane undivided)	All types (Injury)	0.76	0.10
Install edge lines, center lines, and post-mounted delineators	Rural (Two-lane/Multilane undivided)	All types (Injury)	0.55	0.10

Source: AASHTO, 2014.

noted that the only studies referenced as giving a clearly positive effect with a CMF of 0.55 were published in 1968 and 1970, respectively (Roth, 1970; Tamburri et al., 1968). And the authors of the 1968 study, Tamburri, Hammer, Glennon and Lew, themselves writes on page 36 of the 1968 Highway Research Record Number 257, with respect to delineation that, “the findings are based on a relatively small number of locations and, in some cases, a small accident experience; thus, the representativeness of the data is open to question. Additionally, data on items most relevant to the issues being investigated were sometimes unavailable. Consequently, the findings are of a provisional nature (Tamburri et al., 1968).” It should also be noted that the authors use a simple before and after methodology not accounting for regression-to-the-mean effects. And they themselves conclude after finding that most of the striping combinations did not improve safety, “The findings are somewhat discouraging. Naturally, the hope was to demonstrate much more improvement than what appears to have been accomplished. It may be that the worth of delineation measures lie not so much in terms of accident reduction but in terms of “near misses” which might have been averted and the psychological comfort they may provide the driver.” It seems like newer studies of the effectiveness are warranted. The only other study referenced for the 0.55 CMF is a 1970 study by Roth (1970) at Michigan Department of State Highways. That study pertains to a 40-mile (64 km) section of rural freeway. And, the before and after study actually showed an increase in crashes after post-mounted delineators had been installed but less so than for control sections. So, to found the effect of “Installing edge lines, center lines, and post-mounted delineators” and give an expected CMF of 0.55 based on these two studies certainly seems to lack precision and lack a basis for generalization, or even application to our current roadway environment 50 years later.

A deeper discussion of the safety effects of edge lines and whether they are beneficial to safety can be found in Ezra Hauer’s book *Observational before-After Studies in Road Safety* (Hauer, 1993). He writes, “The painting of edge lines is often thought to enhance safety. However, the studies on which this belief is based are often seriously flawed.” He concludes that it is clear that they increase the feeling of safety but do not actually reduce accident risk.

The FHWA crash modification factors clearinghouse provides a web-based data source of various CMF values with supporting documentation for road safety-related activities.

The Handbook of Road Safety Measures (Elvik and Vaa, 2004), give similar effects. That book states that striping normal edge lines reduce crashes as well as injuries by about 3% whereas widening them from 10 to 20 cm increase injuries by about 5%. Striping of centerlines increases crashes by 1% but reduce injuries by 1%. Lane striping between parallel lanes seems to reduce crash numbers but the confidence interval for the measure spans from a 51% reduction to a 36% increase in crashes.

Alternatively, the iRAP Road Safety Toolkit provides proven road safety treatments in terms of their effectiveness and implement issues.

Issues to Consider

This section discusses several safety issues associated with these treatments.

Retroreflectivity

Retroreflectivity refers to the property of how light is reflected from a surface and returned to its source. A retroreflective traffic sign is one of the key factors that affect nighttime visibility, which can be obscured due to reduced illuminance. Road safety could be enhanced and some nighttime crashes could be prevented by making traffic signs appear brighter and easier to be detected at night. And road-user orientation can be improved at night and in the daytime by making them larger so they can be seen from further away leading to less abrupt changes in speed. Thus, road striping and traffic signs should be retroreflective and should be highly visible both in day and at night.

The retroreflectivity of striping and traffic signs can be measured by handheld devices or mobile devices and is typically determined by the coefficient of retroreflection (R_A), expressed in units of candelas per lux per square meter ($\text{cd lux}^{-1} \text{ m}^{-2}$). It is necessary that the coefficient of retroreflection meet the minimum level appropriate for a corresponding type of striping and traffic signs. Standards and guidelines for minimum retroreflectivity levels of traffic signs are generally set based on sign color and size, position of sign, and sheeting types, whereas those for longitudinal pavement markings are justified based on road type and posted speed.

Various retroreflective sheeting materials used in traffic signs provide different level of retroreflection and service life. They could be made with glass beads or micro-prism and range from engineer grade (ASTM Type I), high intensity grade (ASTM Type III), to diamond grade (ASTM Type XI). Many traffic signs follow standard specifications such as the ASTM D 4956-17 (standard specification for retroreflective sheeting for traffic control). Similarly, longitudinal pavement markings such as centerlines, lane lines, and edge lines are generally required to be retroreflective with a minimum maintained retroreflectivity levels. A well-maintained retroreflective pavement marking significantly helps drivers maintain their maneuver along designated paths and may prevent run-off road and head-on crashes particularly during the nighttime. However, they may also lead to higher travel speeds and whether the net safety effect is positive or negative seems to depend on the type of location.

Maintenance

Both traffic signs and striping should be periodically inspected and evaluated as they degrade over time. A systematic traffic sign inspection and inventory of traffic control devices should be maintained so that a replacement program can be effectively implemented before the end of their expected service life. The current MUTCD classifies maintenance of traffic signs and striping into two types, that is, functional maintenance and physical maintenance. The former action is to ensure that traffic signs and striping meet current traffic conditions, while the latter is to warrant legibility, visibility, and proper functions.

Activities such as sign cleaning and vegetation control can be performed as part of preventive maintenance. Damaged and faded traffic signs should be repaired or replaced. In addition, sign supports should be constantly monitored so that the sign crashworthiness is maintained. Maintenance priority is usually given to mandatory and warning signs on a roadway with higher functional classification.

Periodic measurement and evaluation of striping and road markings is necessary to ensure that they are legible to road users. Maintenance of striping and pavement markings can be performed mainly from two perspectives, that is, replacement based on monitored markings and blanket replacement. In the former situation, when the retroreflectivity level of a representative sample of markings that are monitored falls close or below the minimum acceptable value, pavement markings of similar kind are replaced. Alternatively, in the blanket replacement, pavement markings of the entire area or corridor are replaced when they reach their expected service life.

Redundant Signs and Striping

Traffic signs that are redundant could potentially reduce traffic safety and obscure visibility. In some situations, these signs are unauthorized and arbitrarily installed by other agencies, resulting in a visual pollution and overwhelming information to drivers. In addition, installation of too many traffic signs, especially at intersections, could distract drivers and lead them to undesirable driving decisions and behaviors such as sudden lane changing and dangerous overtaking. Therefore, a regular inspection should be conducted in order to spot and remove redundant or unnecessary traffic signs.

The sign redundancy issue is also related to modern advertising billboards, which are permitted in some jurisdictions. Those are often illuminated with high luminance in the roadside of urban areas. These billboards obviously distract visual attention of road users and can possibly lead to increased crash risks. The situation could be worsened at nighttime for a large digital light emitting diodes video billboard screen that are commonly found along the roadsides of some countries.

Redundant striping and pavement markings could result from road construction or road rehabilitation when temporary markings are installed and are not properly removed afterwards. Similarly, it could remain after new road surface treatments are installed due to changes in traffic conditions or road geometry. It is thus recommended that any temporary striping and pavement markings be removed to prevent road users from potential mislead or misdirect.

Human Factors

Although a road crash can be attributed to a combination of human errors, vehicle failure, and road hazards, the majority of road crashes can be influenced by human factors. Human factors deal with how drivers perceive road system, how they process the information, and how a decision is made. Understanding human factors help traffic engineers design a safe road system that can accommodate driver performance and ensure higher road safety. It is thus challenging for traffic engineers to incorporate road users' needs and design traffic components that are functional and relevant to all road user groups.

When operating a vehicle, reaction time can be considered as the combination of time required to see visual traffic signs (perception), identify traffic signs (identification), reach the decision (emotion), and execute the action (volition). Road users of different ages may maneuver and react differently on the road. Older and impaired drivers generally require more time to process information and respond to traffic signs and related information. Cross-cultural differences may also impact driving behaviors in various aspects, including reaction time, risk perception, risk-taking behavior, and general driving performance.

Uniformity

Uniformity of traffic signs and striping can facilitate driving tasks of road users. Many countries have adopted international standards, that is, the Vienna Convention on Road Signs and Signals for traffic control devices in order to facilitate international traffic movements. According to the Convention, road signs are classified into 8 categories from A to H, namely, danger warning signs, priority signs, prohibitory or restrictive signs, mandatory signs, special regulation signs, information, facilities, or service signs, direction, position, or indication sign, and additional panels.

Sometimes these standards and guidelines are adapted to local needs. Nevertheless, certain traffic signs and striping treating similar situation are yet found different in their shape, color, or size across countries. In locations with foreign driving travelers, unfamiliar traffic signs could be misinterpreted, and local traffic law and regulations could be unknowingly violated. In the worst-case scenario, falsely perceiving and interpreting traffic signs could result in crashes.

Nonuniformity issues can also be observed on various local traffic signs with local text or messages, causing foreign drivers to be unable or having difficulties to interpret the true meaning. Using symbols rather than words is therefore typically preferred. As an alternative in some countries, an English or other preferred second language text is additionally augmented on the traffic signs to facilitate international drivers. It is therefore important that all traffic signs and information be consistent in order to ensure that road users can correctly comprehend and perceive traffic signs, which could in turn reduce cognitive workload in driving.

Forgiving Devices and Self-Explaining Roads

Forgiving devices and self-explaining roads are the key concepts in safe system approach, a systematic approach that considers the holistic view of transport system management, including road infrastructure and facilities, road users, vehicles, and travel speed. The safe system principle essentially incorporates the notions of human fallibility and human vulnerability in creating a road system that is safe and forgiving for all types of road users with the main goal to reduce fatalities and serious injuries on the road.

Forgiving traffic control devices aim to strengthen road safety and minimize negative consequences due to human errors. Traffic signs could be equipped with breakaway supports that fail in a relatively safe manner when hit by a vehicle. In the United States, the AASHTO Manual for Assessing Hardware (MASH) presents a uniform guideline for evaluating highway safety features in terms of crashworthiness (AASHTO, 2016). Breakaway supports that are formally tested based on the guideline can be safely used on the roadside. Nonetheless, they should be avoided in certain circumstances where there is the possibility of having a secondary accident involving the impacting vehicle.

Self-explaining roads are designed such that drivers can clearly perceive the intended design and function of roads, and thus they immediately and instinctively know how to perform a safe maneuver according to the design. In such cases, roads must have a clear design that meets driver expectations. There are several ways to make a road self-explained. For instance, striping and pavement marking can be used in speed management by modifying lengths and spacing of centerlines. Treatments such as rumble strip on select road section can be used to gain attention of drivers. Similarly, audio-tactile line marking (ATLM) provides an audible sound and vibration when traversed by a vehicle, thus reducing the risk of head-on and run-off road crashes.

Concluding Remarks

Road striping and traffic signs are among the most fundamental and indispensable components in a roadway design. Drivers need to be informed about speed limits, hazards and regulations, and directional sign are needed to find one's destination. Drivers also often request improved striping and good striping definitely improves the perceived safety of a road. To provide better visual cues may also play a role in road safety but the effect is at best very modest. Road striping and traffic signs can influence vehicle operations to keep vehicles further apart and have people slow down earlier but they may at time also increase speeds. To maximize the benefits of striping and sign treatments, installation configuration should be congruent with the prevailing roadway and traffic conditions. In addition, historical crash records, if available, can be analyzed for potential locations that might lack appropriate installation of striping and traffic signs.

Several issues related to road safety, road striping and traffic signs are discussed in this chapter. Road striping and traffic signs should be designed such that they are self-explained and in line with other traffic control devices. Forgiving devices such as sign posts with breakable supports should be considered when applicable. During the operations phase, road striping and traffic signs should always be in good working condition, that is, they should be functional, legible, and properly placed at designated locations. During nighttime or inclement weather conditions, striping and signs should be retroreflective, clearly delineate the road path, and safely direct drivers to desired direction. Redundant striping and traffic signs should be removed or replaced to prevent potential conflicts and uncertainty to road users.

Selection of striping and traffic signs should be consistent and complied with standards and guidelines. Traffic signs that might confuse drivers and complicate driving tasks should be removed. Similarly, non-standard striping and traffic signs may be used with caution. Driver familiarity and human factors are among essential issues when placing striping and traffic signs on a road. Although a road agency may have its own standards and guidelines, it is important that road users of all types clearly understand their meaning. When foreign drivers are present, applying international standards with additional context that is universally understandable could be beneficial in comprehending traffic signs and the road environment.

In terms of maintenance, road striping and traffic signs should be periodically examined. Inspection should reflect changes in functional requirement on a timely basis. Non-functional and obsolete traffic signs should be removed, and faded road striping and lane markings should be repainted. Retroreflectivity of striping and traffic signs should be constantly monitored. A systematic inventory and maintenance program of road striping and traffic signs are thereby recommended. In many cases, advance traffic signs and pavement markings, incorporated with the intelligent transport system (ITS), could further enhance road safety. With the advent of ITS technologies, it is hoped that better and safer road systems can be achieved, and road traffic injuries can be minimized.

See Also

Roadside Safety Barriers; HUMAN FACTORS IN TRANSPORTATION; Rumble strips, continuous shoulder and centerline; Road Safety Audits; Traffic signals and safety

References

- AASHTO, 2014. Highway Safety Manual. American Association of State Highway and Transportation Officials, Washington, DC.
- AASHTO, 2016. Manual for Assessing Safety Hardware. American Association of State Highway and Transportation Officials, Washington, DC.
- Elvik, R., Vaa, T., 2004. Handbook of Road Safety Measures. Elsevier, Netherlands.
- FHWA, 2009. Manual on uniform traffic control devices. Federal Highway Administration, U.S. Department of Transportation, Washington, DC.
- Hauer, E., 1993. Observational before-After Studies in Road Safety. University of Toronto, Toronto.
- Roth, W.J., 1970. Interchange ramp color delineation and marking study. Highway Res. Rec. 325, 36–50.
- Tamburri, T.N., Hammer, C.J., Glennon, J.C., Lew, A., 1968. Evaluation of minor improvements. Highway Res. Rec. 257, 34–79.

Further Reading

- Austroroads, 2018. Guide to Road Safety, Austroroads, Sydney, Australia.
- Crash Modification Factors Clearinghouse, 2019. Federal Highway Administration. Available from: <http://www.cmfclearinghouse.org/>.
- iRAP Road Safety Toolkit, 2019. International Road Assessment Programme. Available from: <http://toolkit.irap.org/>.
- United Nations Economic and Social Council, 1968. Vienna Convention on Road Signs and Signals. Technical Report, United Nations Economic and Social Council.

Suicides

Brendan Ryan, Human Factors Research Group, University of Nottingham, Nottingham, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	656
Railway Suicides	656
Statistics/Incidence	656
Types of Events	657
Influencing Factors	657
Prevention	658
Highway Suicides	658
Statistics/Incidence	658
Types of Events	659
Influencing Factors	659
Prevention	659
Comparing Railway and Road Suicides	659
Future Directions—Research, Practice, and Policy	660
Biography	660
Relevant Websites	661
References	661

Introduction

Transport offers mobility over increasingly greater distances, catering for differing priorities of travelers, from relaxation and comfort through to enjoyment of a sense of speed in others. Sadly, as a result of speed, high mass and resulting impacts, a number of people take steps to end their lives in suicides involving transport. Some of these are a result of deliberate choices, though the extent to which the people involved are fully in control of their decisions and actions is debatable. We know that a proportion of suicide events are impulsive and can be influenced by a wide range of individual and situational factors.

Suicide is not easy to define (Silverman, 2016) and there are many definitions in the literature. Roughly one million people take their lives each year worldwide, with 10 times as many attempts (7.9 deaths per 100,000 population in China, 11.9 in the United Kingdom, 17.4 Australia, and 21.1 in the United States). Suicides are evident in all forms of transport: rail, road, aviation, and marine. These range from single person incidents on train tracks, in cars, from road bridges, and from ships, through to the extremely rare, but serious incidents like the decision of a pilot to crash a plane carrying many passengers. Data are identifiable to some extent in the classes within the WHO International Statistical Classification of Diseases and Related Health Problems (ICD-10 Codes), for example: X80 Intentional self-harm by jumping from a high place; X81 Intentional self-harm by jumping or lying before moving object; X82 Intentional self-harm by crashing of motor vehicle.

In this chapter, the focus is restricted to railway and highway (road) suicides. In both situations, these incidents result in death of the person involved, but can also include other consequences: the injury or death of others, psychological trauma for witnesses, relatives of the person and those involved in dealing with events and recovering bodies or body parts, and the economic costs of delay in dealing with incidents and getting road and rail networks back to normal operations.

The chapter includes an overview of what is known about suicides on railways and highways. This considers the incidence of these events from some of the available statistics, including commentary on the factors that can affect our understanding on the extent and nature of the problem in each of these transport contexts. Commentary is also provided on the current research directions and the initiatives that are being implemented in industry. In reflecting on where progress has been made in each area and some of the factors that hinder such progress, suggestions are made for cross-sector learning and priorities for future research and practice.

Railway Suicides

Statistics/Incidence

It can be hard to determine how many people take their lives on railways internationally. Statistics from around the world show some variation. Data from the United States shows 219–328 annual railway suicide fatalities in recent years, less than 1% of total suicides in the United States. There have been between 235 and 307 suicides in each of the last ten years on the mainline in Great Britain and London; underground system, equating to 3.6%–4.7% of the total number of suicides in Great Britain. More widely in Europe, there have been between 2756 and 2982 suicides on the railway annually, accounting for 73% of fatalities on European Union railways. These range from 0.4% of all suicides in Latvia to 13.0% of all suicides in the Netherlands (Reynders et al., 2011).

Statistics are also reported and compared by traffic density to account for differences in exposure. As examples, lower rates (0.2–0.3 per-million train kilometer) have been observed in Norway, the Republic of Ireland, and Greece and higher rates (1.2–1.4 per-million train kilometer) in Portugal, Czech Republic, and the Netherlands.

These statistics are informative, but there can be issues of misclassification, for a number of reasons, potentially impacting on cross-country comparisons. These include difficulties in interpreting what happened from the available evidence (often requiring open or narrative verdicts), differences in legal and reporting systems employed across countries, or the potential influence of religious or cultural factors in decision-making. There can be a time lag in making a determination of whether an event is suicide, such as, while awaiting a coroner verdict. This can introduce uncertainty into records and statistics of events over the most recent eighteen months. As a means of making judgments as early as possible, the industry in Great Britain apply the Ovenstone criteria (Ovenstone, 1972) in determining whether an incident is likely to be suicide, using a balance of probability judgment.

There have been many efforts to analyze various demographic, situational, and environmental factors associated with these events. These studies can provide useful descriptive evidence and can help target some preventative measures (e.g., campaigns aimed at higher risk groups, such as male, aged 40–60), though have lesser value in helping to predict where and when an incident may occur. On the whole, relatively small numbers of suicides are spread across the vast railway network, posing challenges to those who are trying to improve safety and reduce fatalities on the railway. It has been possible to identify some locations where there is a thought to be higher risk (usually determined by several incidents) and prevention can be targeted in these areas. However, it is difficult to say to what extent there may be common factors that connect a series of events at a particular location.

Types of Events

There have been various studies of the ways in which people take their life on the railway. Railways in some countries have high security and restriction of access to most areas of track. Therefore, the main access points can be at stations and railway crossings, with smaller numbers of people accessing the track over or through less substantial areas of fencing in various locations. In many countries the access is easier to unfenced areas of the railway (especially outside towns and cities). It is often thought that people choose locations where there are faster trains (e.g., outside of cities on the open line, on fast lines through stations where some trains do not stop at a station). As such, people may target a specific platform at a station, jump from the early part of the platform where train speeds are higher, find a point where they can hide to wait for a train out on the open line, or may move around the track area in search of a fast train. There are still many incidents involving slower trains, including freight trains, especially at night where there may be fewer passenger trains or in countries where the proportion of freight trains is greater (e.g., United States).

There have been efforts to classify the locations of incidents at a national level. In Great Britain, 50% of fatalities occur on the running line, 40% at stations, and the remainder at crossings or other locations. Different proportions can be seen in data from other countries. These national data tend to mask the true situation at a local level. In some areas (e.g., mainlines close to major cities), there may be very few crossings and access to the line is commonly from some point at the station. Consequently, the proportion of incidents at the station or near to the station will be much larger. Similarly, in a rural area, especially away from the main inter-city lines, the numbers of incidents at or close to a crossing are likely to increase.

There are also classifications of the mode of railway suicide (jumping, lying, wandering along or around the track, and touching electrical traction supply) (Ladwig et al., 2009). Recent work is collecting more detailed descriptions of the behaviors of the people prior to an event (Ryan, 2018), with the intention of refining efforts for prevention measures to respond to different types of incidents and access behavior. While there are greater proportions of these commonly reported modes of railway suicide, there are also smaller numbers of somewhat anomalous events (including driving onto railway crossings or unusual ways of making contact with a passing train), that need some consideration in the development of preventative measures.

Influencing Factors

Recent work in Great Britain (Marzano et al., 2019) is considering *Why people choose the railway for suicide?* This is thought to be because of ease of access, a perception of certainty of death, and death that is quick and painless. These can be misplaced perceptions and studies of data on suicides and suicide attempts indicate that up to 60% on metro systems and up to 90% of mainline rail attempts are successful. It is likely that some events are not the result of a planned and rational decision and can be impulsive. While it is known that there are some high-risk areas that are close to psychiatric institutions, many people involved are not known to be at risk by health services. People can reach a low point in their life, such that suicide seems a plausible way of escaping from the position in which they find themselves. This does not mean that they will continue to be suicidal or that they will move to another location if they are prevented from taking their life. Longitudinal studies of survivors of bridge jumps (Seiden, 1978) give support to the transitory nature of some suicidal thoughts and the long-term survival of those involved.

It is likely that aspects of the location can encourage the selection of a railway site for suicide. Sometimes this may be a prominent location, where there are frequent fast trains, with the ability to hide and not be disturbed and have seclusion from others. The intensity of train traffic and ease of access to the track are thought to be influencing factors (van Houwelingen et al., 2013). It is possible that some people don't think too hard about choice, but pick what is familiar or look for opportunities to access trains. Locations tend to be quite close to where people live. Knowledge of events at a location, or high profile events in the media can influence copycat incidents (known as the Werther effect, imitating the suicide of a hero from the work of Goethe) (Niederkrötenhaler et al., 2010).

Prevention

There are many available methods in operation across the world in efforts to prevent these types of events. There is still a gap in knowledge of what prevention measures work and in what circumstances. This is improving, but there is still a lot to do.

The classification of five overarching types of prevention methods ([Rådbo et al., 2008](#)) is a useful framework that provides an effective way of thinking about how a preventative initiative might operate in practice. These includes the following:

Reducing the perceived attractiveness and availability of rail traffic as a means of suicide. This reduces ideation, such that people don't associate the railway with a means of suicide. In practical terms this can mean responsible media reporting after an event, to minimize the risk of copycat incidents.

Limiting access to the track area or means of suicide. This can be achieved through fencing, though many practical difficulties can be anticipated. It is hard to fence all parts of the network, with particular weaknesses at stations and crossings. Platform edge screens have been tried with success in various underground systems, but preventing access is harder to do on mainline railways, where different train types and door configurations pose problems for alignment with openings in the screens. Through design of the infrastructure (station configurations, removing locations for hiding, effective use of lighting or auditory cues) it may be possible to influence how people move around a station and keep them away from places of risk.

Early warning to enable a fast response. Better understanding of behavior can be used in improving surveillance, either by people or technology. Regardless of how surveillance is carried out, there is a need for consideration of what type of response is needed and what is possible when someone is suspected of being at risk. This relies on having more informed staff and members of the public, with the confidence and willingness to intervene. There may be limited time in many cases to react, but in others there are indications of ambivalence and time to intervene as someone may be deliberating about their course of action. An example is the "Small talk saves lives" campaign in Great Britain, encouraging people to approach and talk with people who they suspect may be considering taking their life. This can also be in situations away from the railways. There is recent discussion in the literature of the Papageno effect ([Niederkrotenthaler et al., 2010](#)), where the likelihood of suicide can be reduced by the supportive intervention of friends.

Measures to persuade people to leave a place of danger. This has some links to effective surveillance, but could include interventions using traditional warning signs. More recent innovations include lighting or audible messages, triggered by sensors, to encourage people to leave the track area or place of hiding on a station platform, or to dissuade access to a tunnel on the rail network.

Measures to minimize the consequences of a collision. These are mitigating measures and are challenging in the railway context. Train drivers are relatively helpless to respond and unable to steer away from someone on the track, in a train that can take several kilometers to stop from high speeds. Such measures could include more responsive train braking and better design of train frontages to protect the people struck by the train. These have been considered historically (e.g., use of airbags and nets on trains), but often discounted for practical reasons. If progress could be made in this area of train design there could also be accompanying psychological effects on people contemplating suicide (i.e., by reducing suicide ideation, on account of the reduced likelihood of success of a suicidal attempt).

There has been much progress in recent years in collaboration between many organizations and agencies to improve understanding and share best practice in this area (e.g., infrastructure managers, train operating companies, transport police, local police, local authorities, and health authorities). There are excellent examples of working together at national and international levels. One example is the major collaborative EU project RESTRAIL, which included efforts to identify and evaluate the likely success of wide ranging preventative measures ([Ryan et al., 2018](#)), also including a series of pilot tests of recommended preventative measures. Collaboration is on going, with widening worldwide participation in the GRASP network.

Highway Suicides

Statistics/Incidence

Approximately 1.25 million people die on roads worldwide each year. Studies reported in the literature ([Pompili et al., 2012](#)), covering data from wide ranging international sources, suggest that between 1% and 7.4% of motor vehicles fatalities could be a result of suicide, though reaching as high as 10% in some years of study. Many of the studies focus on motor vehicle collisions and do not take account of vehicle–pedestrian collisions.

There are indications of under-reporting and misclassification ([Harrison, 2017](#); [Tøllefsen et al., 2015](#)), and this may be more problematic in highways than in the rail context, for a number of reasons (e.g., changes in data quality over time and exclusion of suicides from some reporting systems). Various recommendations have been made: for the need for education of crash investigators; improved awareness of medical staff; advice to coroners and researchers; better communication between authorities; better access to information for decision-makers in certifying deaths; and improved training of coders of national databases. In Sweden, there are efforts to ensure more in-depth study of incidents. This includes collection of more details on the vehicle, road and user (details of life events and mental and physical health—psychological autopsy), and uses input from experts from forensic medicine, psychology, and traffic safety backgrounds to classify fatalities as suicide or accidents.

Types of Events

Like railway suicides, these events include collision between a vehicle and pedestrian. These can also involve a driver (usually alone) steering into the path of another vehicle (often large) or fixed object (e.g., tree or structure). Other event types include driving into rail crossings (potentially with injury to the train driver and on-board passengers) or driving into water and subsequent drowning, or jumps of a pedestrian from height (e.g., road bridge). The relative frequencies of these different event types are not consistent in different studies across the world.

These incidents can include deaths of others (e.g., when in collision with another vehicle). It is thought that deaths of others can occur in as many as 11% of driver suicides. These impacts clearly add to the tragic circumstances, though it is possible that the people don't necessarily understand what they are doing or the consequences of their actions.

Determining intent can be difficult, especially within in-vehicle incidents. There may be evidence at the site (e.g., lack of skid marks, indications of high, uncontrolled vehicles speed, or limited efforts to avert collision), but these can also be indications of fatigue related accidents. Additional evidence is therefore needed from witnesses, relatives, autopsy, and police and coroners for correct classification of the event.

Influencing Factors

While not reported specifically in literature, some of the influencing factors in rail (individual, location, and environmental cues) are also plausible in highway contexts, by reason of similarities in the events. Seclusion is a factor in some incidents (e.g., one pedestrian and single occupants in vehicles). People may choose this method of suicide as it can be hard to be sure that this was in fact suicide and may protect families from the psychological trauma associated with knowing that they had chosen to take their life. Studies have indicated the influence of the Werther effect, where reporting of events can influence future events of a similar nature. Excess alcohol, drugs, extreme stress, and life events are thought to be factors, potentially making it easier to act impulsively. However, some prescribed drugs could impact on driver behavior and lead to car accidents, becoming confounding variables in analyses.

Prevention

Understanding the type of event and influencing factors are important in selecting the types of countermeasures that are likely to be effective for suicide. Generally, there are limited ideas in literature for prevention, with suggestions largely drawn from general road safety measures (Routley et al., 2003). For people in vehicles, these include design of the infrastructure (e.g., effective road side barriers) and crash protection in vehicles (e.g., technologies for automatic application of brakes, interlock systems to prevent driving while influenced by alcohol, prevention of tampering of vehicle safety features such as airbags, automated emergency call, and location indicator systems to help survival in the event of a crash). For pedestrians, suggestions include the design of cars to minimize impacts on pedestrians from collisions, better braking, and night vision systems. Some general principles are also reported, notably relating to the reduction of crashes and the energies involved in these. Other suggestions for suicide reduction include reducing heavy transport, better separation of heavy and light vehicles, better design of vehicle fronts (especially heavy vehicles), stability control, and enhanced braking systems in vehicles. There are few if any detailed studies in the published literature to evaluate the effectiveness of suggested preventative measures for highway suicide.

Comparing Railway and Road Suicides

There are many similarities in events across the two transport contexts, but also a number of differences of interest. Some of the incidents are similar in nature, involving pedestrians in collisions with vehicles, which are often large/heavy and fast vehicles. Similarly, jumping from bridges can occur onto railway lines or roads. Isolation is a common factor in each of these incident types and is evident in many events where people hide in secluded areas of stations or the track area, or near to roads or bridges. People are also alone in single vehicle incidents or single occupant vehicle collisions.

A wide variety of locations are possible in each of the railway and highway contexts. These are often easy to access (potentially easier on some highways than many parts of railways), with small numbers of events in any location and difficulties with prediction of location specific risk. Intervention is possible to prevent a suicide in some situations (e.g., where people may be waiting and visible in a location for a period of time before an event). However, events can occur with very little notice on the railway and highway, as someone emerges from a place of hiding, from close proximity to the line or road, or from the impulsive action of the individual. There is the potential for psychological trauma for any witnesses (train drivers, car/lorry/bus drivers) and staff involved in recovery. Events in both railway and highway settings cause serious disruption and associated costs for the industry and society as a whole.

A notable difference is that many of the suicides on highways occur within vehicles. These events are more likely to cause physical injury or death to others on road settings and it can be difficult determining intent and classifying these as suicides. As such, under-reporting of suicides is likely to be a greater problem in the highways context. In terms of prevention, it is very hard to intervene in these types of vehicle-to-vehicle collisions, as the first indication may be very close to the point of impact. Fencing and restricting access to the road network may also pose more difficulties than in railways. In contrast, some vehicle design mitigations are likely to have more potential for prevention in highways settings than in railways.

Future Directions—Research, Practice, and Policy

The number of suicides on the railways and highways are relatively low and a modest proportion of all suicides. However, these are very violent deaths, with far reaching impacts and trauma (mental and physical) to others, plus huge costs to the industry and society. In addition, the scale of the problem is not fully appreciated, with some underreporting, misclassification and delay in classification (i.e., a time lag in official statistics).

Better data are needed to understand the true scale of the problem, along with the circumstances surrounding the incident and the most appropriate approaches to prevention that are aligned with these circumstances. While compiling data at national levels and benchmarking these internationally can be informative, looking in detail at events at a local level is necessary to fully understand and design solutions for future prevention. Clarity is needed in several areas. Some of this information is collected by various agencies, but not shared very well and accessible to researchers and practitioners involved in prevention.

Detail on incidents. This can include firstly, the locations of incidents. Multiple reference points are relevant, such as, the access point, movements to locations around the track or road infrastructure, the final location of the event (e.g., line or part of carriageway). Location based findings can be specific to the particular incident or more generic lessons with wider application. Another consideration is the pre-incident behaviors of people involved. This can include gestures, movements, and interactions with others, potentially betraying an emotional state or intention. There are thought to be many successful outcomes from interventions from bystander campaigns. Improved knowledge in this area can improve training for staff, information for public awareness, and the future development and application of visual analytic surveillance systems.

Background of the people involved and the things that influenced them. There is an increasing awareness within organizations (railways and highways) and positive efforts and campaigns to reduce incidents. It is recognized by people in the industry that these events are not inevitable and things can be done to prevent them. This has not always been the case and within general society it is still possible to hear people say that these events cannot be prevented. People should not be treated as a homogenous group. There may be a proportion of people who will be determined and very difficult to influence, but there will be many who can, for a variety of reasons, be at a transitory low point in their lives and small interventions to prevent suicide can be successful.

It is necessary to continue efforts to improve understanding of preventative measures, in particular, which countermeasures work and in what circumstances do they work. In doing so, it is also important to consider how countermeasures are implemented and the extent to which they are implemented in the way in which they were intended. There are often gaps between the intention and what is actually implemented in a prevention program.

Evidence on the evaluation of countermeasures in the scientific literature is growing in railways, but there is still a lack of good evaluation studies. In comparison, there is very limited evidence in literature of testing of countermeasures for suicide in the highway sector. Some of the knowledge of railway safety interventions is held within industry and there were efforts to access this and determine consensus on the evidence in the RESTRAIL program (Ryan et al., 2018). This was an exercise that focused on immediate priorities for implementation within the RESTRAIL project and there are arguments for repeating this, with new terms of reference (e.g., considering more ambitious and novel interventions—possibly using cross-sector groups to share learning). This may need to contemplate major system change. As discussed earlier, some of these events are impulsive and can happen with apparent ease, with minimal opportunities to intervene. Looking back from a point in time in the future, what will we think of our history of unfenced platforms and vehicles approaching one another on single carriageways with high combined speeds?

In practice, countermeasures are often implemented in industry without any evaluation. Even where there is an appetite to evaluate, the knowledge of how to do this in real world settings is often lacking. This type of evaluation is difficult to do and even harder to roll out in a program across an industry. However, this doesn't mean we shouldn't try to improve the organizational/inter-organizational practices and processes that will support this type of work.

On balance, there seems to have been greater progress so far in railways, in terms of compiling knowledge of events, implementation and evaluation of prevention opportunities and collaboration between agencies (highways authorities, emergency services, health authorities, and voluntary sectors). Suicides are difficult problems to resolve, but there is value in any lives that can be saved with continued efforts in these transport contexts.

Biography



Dr. Brendan Ryan is an Associate Professor in the Human Factors Research Group at the University of Nottingham. He has wide-ranging experience in safety, accident investigation and ergonomics/human factors roles, predominantly in transport settings. This has included research on rail fatality prevention and management (suicide and trespass), with particular areas of interest on the study of behaviors of people before incidents and the implementation and evaluation of prevention measures. This has included the development of evaluation programs and support tools for industry, taking account of the challenges of implementing solutions and research findings in industrial contexts. Recent work has also included how this work can be applied in highways contexts.

Relevant Websites

World Health Organization. *Classifications*. Available from: <http://www.who.int/classifications/icd/en/>.
 RESTRAIL—Available from: www.restrail.eu.
 GRASP—Available from: www.volpe.dot.gov/rail-suicide-prevention/grasp.

References

- Harrison, K., Suicides on UK roads—Lifting the lid, Parliamentary Advisory Council for Transport Safety, PACTS, UK, 2017.
- Ladwig, K.-H., Ruf, E., Baumert, J., Erazo, N., 2009. Prevention of metropolitan and railway suicide. In: Wasserman, D., Wasserman, C. (Eds.), *Oxford Textbook of Suicidology and Suicide Prevention. A Global Perspective*. Oxford University Press, Oxford, UK, pp. 589–594.
- Marzano, L., Mackenzie, J.M., Kruger, I., Borriall, J., Fields, B., 2019. Factors deterring and prompting the decision to attempt suicide on the railway networks: findings from 353 online surveys and 34 semi-structured interviews. *Br. J. Psychiatry* 6, 1–6, doi:10.1192/bjp.2018.303.
- Niederkrotenthaler, T., Voracek, M., Herberth, A., Till, B., Strauss, M., Etzersdorfer, E., Eisenwort, B., Sonneck, G., 2010. Role of media reports in completed and prevented suicide: Werther v. Papageno effects. *Br. J. Psychiatry* 197, 234–243, doi:10.1192/bjp.bp.109.074633.
- Ovenstone, I.M.K., 1972. A psychiatric approach to the diagnosis of suicide and its effect upon the Edinburgh statistics. *Br. J. Psychiatry* 123, 15–21.
- Pompili, M., Serafini, G., Innamorati, M., Monteboni, F., Palermo, M., Campi, S., Stefani, H., Giordano, G., Telesforo, L., Amore, M., Girardi, P., 2012. Car accidents as a method of suicide: a comprehensive overview. *Forensic Sci. Int.* 223, 1–9.
- Rådbo, H., Svedung, L., Andersson, R., 2008. Suicide prevention in a railway system: application of a barrier approach. *Safety Science* 46, 729–737, doi:10.1016/j.ssci.2006.12.003.
- Reynders, A., Scheerder, G., Van Audenhove, C., 2011. The reliability of suicide rates: an analysis of railway suicides from two sources in fifteen European countries. *J. Affect. Disord.* 131, 120–127.
- Routley V., Staines C., Brennan C., Haworth N., Ozanne-Smith J., 2003. Suicide and natural deaths in road traffic—review, Report no. 216, Monash University, Accident Research Centre, Melbourne, Australia.
- Ryan, B., 2018. Developing a framework of behaviours before suicides at railway locations. *Ergonomics* 61 (5), 605–626, doi:10.1080/00140139.2017.1401124.
- Ryan, B., Kallberg, V.P., Rådbo, H., Havårneanu, G.M., Silla, A., Lukascsek, K., Burkhardt, J.-M., Bruyelle, J.-L., El-Koursi, E.-M., Beurskens, E., Hedqvist, M., 2018. Collecting evidence from distributed sources to evaluate railway suicide and trespass prevention measures. *Ergonomics* 61 (11), 1433–1453, doi:10.1080/00140139.2018.1485970.
- Seiden, R.H., 1978. Where are they now? A follow-up study of suicide attempters from the Golden Gate Bridge. *Suicide Life Threat. Behav.* 8 (4), 203–216.
- Silverman, M.M., 2016. Challenges to defining and classifying suicide and suicidal behaviors. In: O'Connor, R.C., Pirkis, J. (Eds.), *The International Handbook of Suicide Prevention*. Wiley, Oxford, pp. 11–35.
- Tøllefsen, I.M., Helweg-Larsen, K., Thiblin I., et al., 2015. Are suicide deaths under-reported? Nationwide re-evaluations of 1800 deaths in Scandinavia. *BMJ Open* e009120. doi:10.1136/bmjopen-2015-009120.
- van Houwelingen, C., Baumert, J., Kerkhof, A., Beersma, D., Ladwig, K.-H., 2013. Train suicide mortality and availability of trains: a tale of two countries. *Psychiatry Res.* 209, 466–470.

Surrogate Measures of Safety

Nicolas Saunier*, Aliaksei Lareshyn[†], *Civil, Geological and Mining Engineering Department, Polytechnique Montréal, Montréal, QC, Canada; [†]Traffic and Roads, Department of Technology and Society, Faculty of Engineering, LTH, Lund University, Lund, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	662
How to Measure Safety?	662
A Short Historical Perspective	663
Definition and Properties of Surrogate Measures of Safety	663
The Concept of Severity	664
Operationalization of Surrogate Measures of Safety	664
Interaction Severity Indicators	664
The TTC Category	665
The PET Category	665
The Evasive-Action Category	665
Other Indicators	665
Data Collection Methods	666
How to Diagnose Safety	666
Perspectives	667
References	667

Introduction

How to Measure Safety?

The number of road crashes or injuries that occurred during a certain time period and within a certain area seems intuitively like a good measure of safety, with a clear and direct relation to the core of (un)safety—human suffering and property harm. From a statistical perspective, the crash count is a random variable governed by some underlying distribution. For example, the annual number of reported crashes in a given area does not remain the same from 1 year to another, even if there are grounds to believe the traffic system has not changed particularly. It has also been shown on many occasions that naive comparisons of the crash counts without taking into considerations their random nature may yield very misleading results and such practice has been banned by the safety experts long ago. Safety is better measured by the number of crashes *expected* to occur in a given area and time period.

It follows that, since crash counts may only be used as a raw material for safety estimates, they are not as *direct* a measure as it may seem. In fact, there are several issues that make the use of crash records for safety analysis problematic and motivates the search for alternative (surrogate) measures with the same aim (estimation of the expected number of crashes):

- Road safety improvements and the constant decrease in crash numbers in most Western countries has a side effect of leaving safety experts with less and less data to work with. Once disaggregated by type or location, the numbers become so low that no statistically meaningful analysis can be performed.
- Collection of “enough” crashes usually require long time periods (at least several years). It is very unlikely that all the conditions remain constant during that time.
- Crashes are heavily under-reported in all countries, depending on the crash severity level, types of road users involved, location, etc. Vulnerable road users and particularly fall-related injuries seem to suffer highest under-reporting rate.
- The information available in a standard crash reporting protocol is scarce and biased by the witnesses’ and police officer’s opinions.
- Crash reporting is always a postobservation of the crash scene. Very little information is available about the process of crash development and dynamic factors present at that moment (such as another road user involved who avoided the collision and left the crash scene).
- The original purpose of calling police to a crash scene was to determine *whom to blame*. While this attitude is gradually changing towards the impartial collection of facts that can explain why the traffic system failed, the legal aspect of police work still has a great effect on the prioritization and search for evidence.
- Finally, there is an ethical dilemma in the reactive nature of the crash-based safety analysis. Before a verdict can be given, one must wait for unsafety to manifest itself in form of crashes and injuries.

As a result, the need for alternative proactive measures of safety has been advocated by traffic safety experts for a long time. The goal of this chapter is to give an overview of the state of the art in surrogate measures of safety (SMoS). This first section provides the core definitions and concepts, as well as a short historical perspective. Section “Operationalization of Surrogate Measures of Safety”

covers the methods to collect and use SMOs. The last section highlights where research is still needed and promising areas for the application of SMOs.

A Short Historical Perspective

The first documented attempt to assess safety by observing critical events in traffic is a study performed by a research laboratory of the General Motors company in the 1960s. The conflict definition covered nearly any action taken by a driver while approaching an intersection (activation of the light for braking or changing lane was used as conflict indicator) and the first results were inconclusive. However, the idea rapidly gained popularity, leading to the development of numerous conflict observation techniques in the 1970s.

These traffic conflict techniques (TCTs) used different indicators and protocols and it was hard to compare results and methodologies. As a response, The International Committee on TCTs (ICTCT) was created in 1977, an organization that had a significant impact on the development of the field in the coming years. ICTCT organized several workshops and joint calibration studies and developed solid theoretical grounds for the collection and interpretation of traffic conflicts. These activities are well-documented and, despite their age, still can be considered a must-read introduction to the subject (a library of essential reading is available at <http://www.ictct.net>).

During the 1980–90s, the TCTs got a solid reputation as a research method, but still were hardly used by traffic practitioners in their daily work. Among the reasons were the high reliance on human observers as the main data collection tool (together with doubts in objectivity of such a tool) as well as the high costs related to that. Starting from 2000s, however, automated and semiautomated tools for data collection, such as video analysis programs, became available and had a significant impact on the field. While not perfect and completely problem-free, automated tools opened for much more extensive data collection under longer periods and various conditions, and, what is important, for empirical tests and validations of various objective measures of safety.

Definition and Properties of Surrogate Measures of Safety

The SAE standard on SMOs uses the following working definition (SAE Committee on Surrogate Measures of Safety, 2019) (In the remainder, *interaction* is used as a synonym for *non-crash event*, referring to any situation where two road users, or one road user and an obstacle [or the side of the road for a road departure], are within perception distance of each other.):

“An indicator derived from the observation and the safety evaluation of non-crash events in traffic with the goal to estimate the expected crash/injury frequency as well as to get a better understanding of the crash mechanisms and contributing factors.”

This definition meets the following three conditions for a good SMOs slightly adapted from (Tarko, 2018; Tarko et al., 2009):

1. It properly captures the effect of changes to the traffic system.
2. It is associated (statistically) with the crashes affected by these changes.
3. It is practical.

The *validity* of SMOs is the degree to which it measures what it is supposed to measure, that is, road safety or, in practical terms the expected frequency of crashes (Laureshyn et al., 2016). While the ultimate goal is to have a clear and stable relation to the expected number of crashes expressed in mathematical terms, as of today the documented attempts to establish such relations are few and not always conclusive.

This strong form of validity is called *absolute product validity* (Laureshyn et al., 2016). When used to estimate the expected crash frequency, the surrogate measure should provide better accuracy than what can be obtained from standard traffic data like flow counts. In other words, it must be a measure of safety and not simply a surrogate for exposure. Yet, it is often sufficient to have an indication of whether the safety is improving or deteriorating, for example, to get a quick feedback after a change in road design or regulation, without knowing the actual scale of the change. This is called *relative product validity*, a lower degree of validity than absolute product validity (Laureshyn et al., 2016).

The second part of the SMOs definition refers to *process validity*, or the level of similarity in the underlying mechanisms of how the studied interactions and actual crashes develop (Laureshyn et al., 2016). This ensures that the crash mechanisms can be studied by observing the safety-relevant interactions. For example, if certain types of interactions seem to occur in certain conditions (e.g., approach at a high speed or at an angle affecting the field of view), the same contributing factors should be found in crashes of this type. Similarly, if a counter-measure affects these conditions and decreases the severity or frequency of relevant interactions, one should expect a reduction in crashes of the same type explained by the same causal mechanisms.

From a practical perspective, it is also important that:

- SMOs are based on interactions that occur considerably more frequently than crashes, so that observation periods are shorter;
- The definition of a specific SMOs is expressed in objective terms, so that it can be measured repeatedly and in an unambiguous way.

SMOs should be distinguished from other noncrash-based indicators used in safety work practice, such as “safety performance indicators” (seat-belt use, alcohol use, etc.) or behavioral indicators (yielding, head or gaze movements, etc.) that are explanatory factors to safety, but do not measure safety as such.

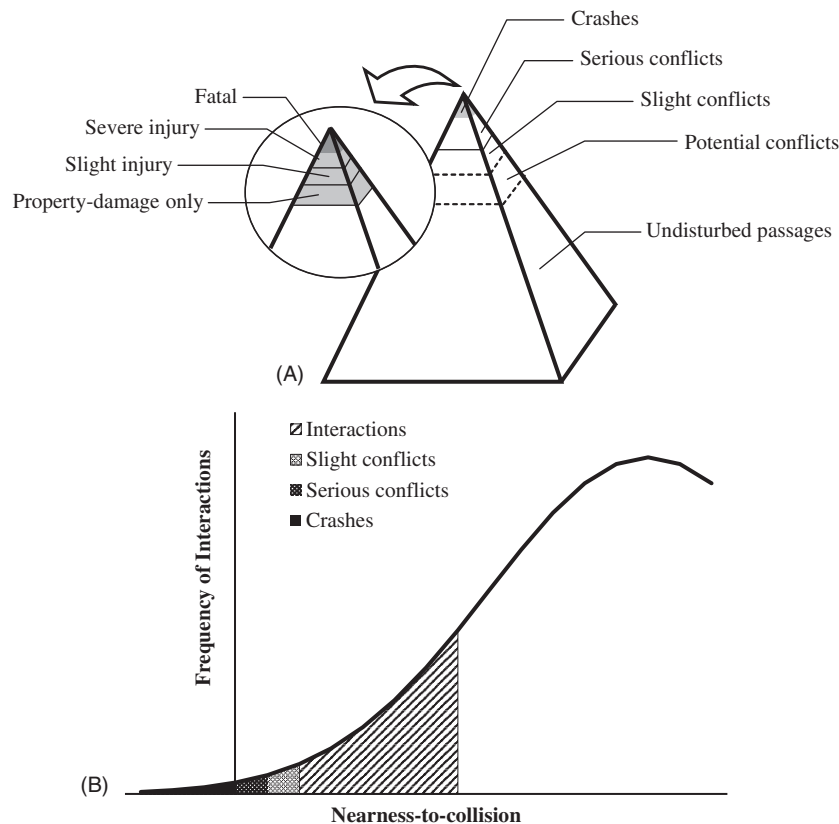


Figure 1 Examples of visual representations of the safety hierarchy as a pyramid (A) or a probability distribution (B) of interactions. *Source: Panel (A) adapted from Hydén, C. 1987. The Development of a Method for Traffic Safety Evaluation: The Swedish Traffic Conflicts Technique (PhD thesis). Lund University of Technology, Lund, Sweden (Bulletin 70); Panel (B) adapted from Glauz, W., Migletz, D. 1980. Application for Traffic Conflict Analysis at Intersections. NCHRP 219, Midwest Research Institute.*

The Concept of Severity

The idea of SMOs rests on the assumption of a hierarchy and continuity among the events taking place on the road (**Fig. 1**). While crashes are the most severe interactions in traffic, and fatal ones the most severe among crashes, they are also the least frequent. At the other end are the least severe events representing normal traffic. Frequency is assumed to decrease as the severity, or nearness to a crash, increases. The research and development of TCTs relies on the assumption that the most severe interactions such as serious traffic conflicts defined in TCTs are the basis for valid SMOs ([Tarko, 2018](#)). Yet these may not be frequent enough to gather enough observations during a reasonable period of time and there are attempts at deriving SMOs from less severe, more frequent interactions ([Svensson, 1998](#)).

The concept of severity is related to the safety hierarchy of an interaction, as the severity should reflect the proximity/nearness to a crash ([Tarko, 2018](#)), but also the potential consequences of a crash. Following the vision zero philosophy, this is reflected in the proposed SMOs definition with the goal of “estimating the expected crash injury frequency.” Different SMOs are related in different ways to these two dimensions of the severity and will result in different interaction distributions and safety hierarchies.

Operationalization of Surrogate Measures of Safety

The current state-of-the-art regarding SMOs is described in the extensive literature review made in 2016 for the European project “In-depth Understanding of Accident Causation for Vulnerable Road Users” (InDeV) ([Laureshyn et al., 2016](#)).

Interaction Severity Indicators

The severity indicators for individual interactions, referred to as “measures of crash proximity,” were grouped in four categories by [Laureshyn et al. \(2016\)](#): the most common category is time-to-collision (TTC), followed by the postencroachment time (PET) and evasive-actions (deceleration).

The TTC Category

The TTC is the time remaining until a collision of two road users (or a road user with a stationary obstacle) assuming they continue traveling as initially planned. TTC is a continuous indicator that can be calculated at any time instant as long as there is a collision course.

To determine the existence of a collision course, that is, what travel was “initially planned,” is a challenge. In the case of an evasive action in the form of changed path, speed, or both, the initial intentions are not revealed and alternative motion prediction methods have to be used. The most simplistic and straightforward method is to assume that the speed and the heading remain constant. Clearly, this assumption does not hold in case of a turning maneuver which requires both changes in the motion direction and speed adjustments to negotiate the turn. Various modifications to this basic method have been suggested, for example, constant yaw or/and acceleration rates: yet, all these methods do not take into account the actual context, composed of the spatial context (road geometry) and the surrounding traffic, and may therefore predict very unrealistic trajectories and speed profiles.

The state of the art in context-aware motion prediction lies rather in learning the typical motion patterns based on observation of large amount of traffic performing the same maneuvers, with or without interactions with other road users (Saunier et al., 2007). The advantage of this method is that instead of poorly motivated kinematic assumptions, the prediction is made based on the actual motion being repeatedly observed, and can work well even in case of unusual geometry, for example in a construction zone. The method can be easily extended within a probabilistic framework, in which each road user has a variety of options for future motion. Combining these multiple future trajectories for a pair of road users yields at each instant a distribution of possible collision points, TTC values and their related probabilities (Gomaa Mohamed and Saunier, 2013). Two road users are on a collision course at a given instant if there is at least one possible future collision point. Such calculations, however, require large amounts of trajectories of normal traffic which requires an automated tracking tool.

At least two TTC values during an interaction deserve attention. Time-to-accident is the TTC at the instant the first evasive action starts, indicating the evolution of the situation from *unaware* to at least one of the road users realizing their collision course and attempting to mitigate it. The second value of interest is the minimal TTC, TTC_{min} , during the interaction, which represents the instant when road users are closest to a collision. The TTC concept inspired a wide range of modifications and new indicators. Some to mention are time to pedestrian crossing (time to zebra), inverted TTC ($1/TTC$), Time Exposed and Time Integrated TTC (describing the relation of a TTC curve to a pre-defined threshold), and T_2 (time for the latest road user to arrive to a predicted crossing point in case there is no collision course and TTC cannot be calculated) (Laureshyn, 2010; Laureshyn et al., 2016).

The PET Category

If the trajectories of two road users cross (overlap), the PET is the duration between the instant the first road user leaves the crossing zone and the instant the second road user reaches the crossing zone, that is, “PET indicates the extent to which they missed each other” (Laureshyn et al., 2016). There is therefore only one measurement at most for an interaction, but it may be measured in some situations where there is no collision course and therefore no TTC measured.

Like TTC, there have been several variations around PET. In a car-following situation, the PET is the same as the gap and similar to time headway. When using motion prediction as for TTC, PET can be also predicted, under the name of time advantage or predicted PET. For each pair of predicted trajectories at t_0 , they either cross or not: if they cross, let t_1 and t_2 be the arrival instants at the crossing zone of the first and second road users, respectively. There are two cases:

- the road users may be predicted to arrive at the same time ($t_2 = t_1$) and a TTC can be calculated as $t_1 - t_0$; OR
- there is a predicted PET calculated as $t_2 - t_1$.

The Evasive-Action Category

Deceleration is the most common evasive action taken by a driver to avoid a collision (Hydén, 1987). Several indicators based on deceleration have therefore been proposed. The first subcategory is for indicators based on observed road user deceleration. The magnitude of the deceleration at the beginning of the evasive action is called the (initial) deceleration rate. The maximum deceleration rate during an interaction has also been used, but these indicators suffer from the varying behaviours, and how the same deceleration may be normal for some users and extreme for others. The derivative of acceleration with respect to time, the jerk, reflects the suddenness of braking and has been shown to better discriminate severe interactions than deceleration.

The second subcategory is not directly observed: the minimal necessary deceleration for a driver to avoid a collision, assuming constant deceleration, was called the deceleration-to-safety time (DST), or the deceleration rate to avoid a crash (DRAC) in car-following situations. Similarly, in car-following situations, the safe stopping distance is the minimum distance required for the following driver to safely reduce speed and avoid a collision if the leader braked at the maximum deceleration and stopped. The proportion of stopping distance is the ratio of the current distance to the minimal acceptable stopping distance (Laureshyn et al., 2016).

An obvious shortcoming of this category is that indicators based on the deceleration or other evasive actions cannot be calculated if there is no such evasive action, if the road users are completely unaware of the situation or apply a different evasive action.

Other Indicators

There have been some attempts at creating indicators that can also reflect the outcome of crashes. The best known are related to Delta-V, a common notation in physics to denote an object's change of velocity because of a collision with another object. Delta-V is

correlated with the severity of real crashes. Based on motion prediction, an expected Delta-V can be calculated. It was extended to take into account evasive actions (braking) to reduce the impact forces. It may reflect both the nearness to a crash and the severity of the potential crash, and is suitable to represent vulnerability of road users, for example, pedestrians with motorized vehicles.

The TCTs developed in the 1970s and after rely to varying degrees on the *subjective* evaluation of the severity of interactions by observers in the field or from video recordings. These may be subjective evaluations of objective indicators such as distance, speed, and TTC. It was shown that observers can be trained to estimate speed and distance with sufficient accuracy to calculate time-related indicators like TTC (Hydén, 1987). Observers can also be trained to provide a holistic evaluation of the “dangerousness” of interactions by watching videos of severe interactions. Laureshyn et al. (2016) argue that these human subjective perceptions may come closer to the “true” severity compared to other objective indicators.

Data Collection Methods

The two most common ways to collect SMOs are *roadside* and *in-vehicle* (Tarko, 2018). As described above, TCTs were the earliest methods for road-side studies. Since the beginning, employing trained human observers to detect interactions and evaluate their severity has raised questions. This led to the search for more reliable and objective methods. Video analysis in particular has generated considerable interest, with a recent boom in the 2000s thanks to the lower costs of recording video and improved computer vision algorithms. Video analysis is rarely completely automated, with important initial configuration of the camera and scene configuration and tuning of algorithm parameters often done manually, and often requires users to annotate, correct or even completely manually track road users, detect relevant interactions and measure safety indicators (a list of existing software for video analysis and SMOs calculation is maintained at <https://www.ictct.net/surrogate-measures-software/>). Progress is ongoing in this area, as new sensors like LIDAR and methods from artificial intelligence may help completely automate the collection of SMOs. A long-term benefit of automated data collection is to provide researchers and analysts with large amounts of traffic data with various levels of safety to better understand the processes that lead to crashes.

In-vehicle studies are called naturalistic studies, most notably naturalistic driving studies. They typically focus on individual drivers/users, their behavior and interactions with the vehicle, the other users and the road. They rely on vehicle instrumentation for data collection. Such instrumentation is similar to available driver assistance systems, for example, forward collision warning based on measuring TTC with the leading vehicle. Severity indicators based on evasive actions, in particular deceleration and jerk, were developed to detect severe interactions in which drivers/users get involved, especially when limited data is available for the surrounding road users. Road-side and in-vehicle studies have complementary advantages and disadvantages: the former provides continuous temporal coverage for a limited part of the road network with little personal information about the road users, while the latter provides coverage for large part of the road network for more limited periods of time and possibly personal information about the users (if doing a survey).

An important question related to data collection methods is the duration of the study. Since interactions, like crashes, occur randomly, the study duration should be long enough to provide enough data so that safety can be estimated reliably. There is little research and a large variability in practice as reported in Laureshyn et al., (2016). Previous recommendations about study durations, for example, from TCTs, may not be relevant given the changes in traffic and the overall improvement in safety (decreasing frequency of crashes and severe interactions) since then.

How to Diagnose Safety

A good SMOs should have a clear relationship to safety, that is, the expected number of crashes/injuries. This is done in two main ways, either by counting the number of severe interactions as in most TCTs or by studying the distribution of some severity indicators and aggregating them. An important problem of the presented severity indicators is that many have been theoretically formulated but hardly followed by any empirical validation. The majority of the validation studies available have focused on TTC and PET, but no clear guidance can be given on which one to prefer as a measure of an interaction severity (Laureshyn et al., 2016). There has been little research on testing the various statistics that can be obtained from continuous severity indicators like TTC, aggregated spatially and temporally, while other researchers have tried different indicator combinations to better reflect the severity of each interaction.

In line with most TCTs, Tarko (2018) strongly argues that the sound basis for SMOs are severe interactions, given their etiological consistency with crashes. He provides a comprehensive summary of the mathematical models proposed for the relationships between the observed number of severe interactions $N_{interaction}$ and the expected number of crashes $N_{crashes}$, the simplest and earliest being a proportional relationship $N_{crashes} = \pi N_{interaction}$ developed in Hydén (1987), where the coefficient π was supposed to be constant. Yet, efforts to estimate such a model have been mixed as interactions and crashes are generally collected over different periods of time in different conditions, leading to π depending on these conditions (Tarko, 2018). More complex models have been proposed to estimate safety, notably through the counter-factual technique (causal inference) and extreme value theory or statistical exceedances.

Instead of counting severe interactions, the severity indicators may be directly used as dependent variables in statistical models (Zangenehpour et al., 2016). The distributions of the severity indicators by interaction can also be used, although it has not been established how interactions of different severity levels are related to safety. A statistical model can identify a statistical association between a treatment and an indicator like TTC_{min} , but it may not mean the same thing if the TTC_{min} are very small, indicating severe interactions, or much larger. A more holistic safety diagnosis may be derived from interpreting the distributions of severity indicators

including less severe interactions and finding how shifts in these distributions translate into safety. Early work showed for example very different severity distributions for a signalized and a nonsignalized intersection in Svensson (1998).

Perspectives

Despite the considerable rise in interest and progress over the last two decades, some challenges remain for the widespread use of SMOs. First and foremost is to establish validity of the proposed SMOs. This may seem like a moving target as a large proportion of crashes are not reported and safety has improved in many countries to the point that crashes are too rare at specific locations for validation between SMOs and crash statistics. Moreover, the properties of the road transportation system are changing very rapidly, too. The evolution of urban mobility systems toward streets for people (rather than cars) yields more walkers and cyclists interacting with motor-vehicle traffic, which in turn calls for new infrastructure designs, speed regimes, and general attitude and behavioral changes. On the other hand, technical progress in driver assistance technologies and the automation of driving tasks will probably have a dramatic impact on the traffic interplay, the relation between interactions and crashes, and safety.

At the same time, the same technological evolution provides more and more data from road-side and in-vehicle sensors. These will provide larger amounts of data and more complete data about interaction and crash processes. Analysis of such big data will strongly benefit from artificial intelligence methods, in particular machine learning techniques. Sensors in probe vehicles and from other users may be used for example to detect evasive actions like harsh decelerations and accelerations that enable cost-effective network-wide screening.

Another venue for SMOs is the use of traffic simulation. Crashes and interactions can be simulated, but the main issue, as for other measures, remains validity. Traffic simulations rely on complex driver models that are difficult to calibrate to real world conditions, while most do not involve the actual mechanisms leading to driver errors. Currently, most driver models represent perfect drivers, or robots, which do not generate any crashes in simulations. The correlations observed between the number of simulated interactions and observed severe interactions and crashes may be introduced by exposure.

Transportation modes with high safety profiles (air transport, railway) have recognized early the value of collecting data on out-of-normal incidents in conditions when very few severe crashes take place. It seems that the safety of road transport, at least in the Western world, is developing in the same direction. Waiting for several years to collect crash records is not a feasible option with the rapid changes in transport systems, leaving the scene to SMOs as the main tool for rapid feedback on safety. With a solid base developed over the decades and with the technical data collection tools reaching maturity, SMOs are about to get an important and possibly irreplaceable role in the toolbox of all road safety practitioners.

References

- Glauz, W. and Migletz, D. (1980). Application for traffic conflict analysis at intersections. NCHRP 219, Midwest Research Institute.
- Gomaa Mohamed, M., Saunier, N., 2013. Motion prediction methods for surrogate safety analysis. *Transport. Res. Rec.* 2386, 168–178.
- Hydén, C. (1987). The development of a method for traffic safety evaluation: The Swedish Traffic Conflicts Technique. PhD thesis, Lund University of Technology, Lund, Sweden. Bulletin 70.
- Laureshyn, A. (2010). Application of automated video analysis to road user behaviour. PhD thesis, Lund University.
- Laureshyn, A., Jonsson, C., De Ceunynck, T., Svensson, Å., de Goede, M., Saunier, N., Włodarek, P., van der Horst, R., and Daniels, S. (2016). Review of current study methods for vru safety: Appendix 6 - scoping review: surrogate measures of safety in site-based road traffic observations. Technical report, In-Depth understanding of accident causation for Vulnerable road users (InDeV) Project.
- SAE Committee on Surrogate Measures of Safety (2019). Concepts, terms and definitions related to surrogate measures of safety. In revision.
- Saunier, N., Sayed, T., and Lim, C. (2007). Probabilistic Collision Prediction for Vision-Based Automated Road Safety Analysis. In *The 10th International IEEE Conference on Intelligent Transportation Systems*, pages 872–878, Seattle. IEEE.
- Svensson, Å. (1998). A Method for Analyzing the Traffic Process in a Safety Perspective. PhD thesis, University of Lund. Bulletin 166.
- Tarko, A.P., 2018. In: *Surrogate Measures of Safety*. Emerald Publishing, (Chapter 17), pp. 383–405.
- Tarko, A.P., Davis, G.A., Saunier, N., Sayed, T., and Washington, S. (2009). Surrogate measures of safety. White paper, ANB20(3) Subcommittee on Surrogate Measures of Safety.
- Zangenehpour, S., Strauss, J., Miranda-Moreno, L.F., Saunier, N., 2016. Are intersections with cycle tracks safer? Control case study based on automated surrogate safety analysis using video data. *Accid. Anal. Prev.* 86, 161–172.

Targeting Transit: the Terrorist Threat and the Challenges to Security

Brian Michael Jenkins, Mineta Transportation Institute, San Jose, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	668
Recent Terrorist Campaigns	668
Escalating Violence	669
Trends in Tactics and Targeting	671
Security Issues	672
References	673
Further Reading	674

Introduction

Public surface transportation has become a theater of operations for terrorists. Passenger trains and stations offer terrorists easy access to crowds of people, often in confined environments, which enhances the effects of explosives or chemical substances and enables the terrorists to spread fear and alarm. While some terrorists aim primarily at disruption and others involve attempts to derail speeding trains, terrorists determined to kill in quantity and willing to kill indiscriminately find the dense throngs of commuters that fill stations and trains during rush hours the most lucrative targets. (Some supporters refer to individuals who target innocent civilians as “freedom fighters,” but their basic mission in carrying out attacks is to cause terror, so we use the term “terrorists.”)

In 2017 a bomb hidden in a briefcase exploded on a metro in Saint Petersburg, Russia, killing 14 people and injuring 64. Later that year, a would-be suicide bomber, angered by US military operations in the Middle East, detonated an improvised explosive device in a subway station in New York City. His crudely made bomb only partially detonated, injuring him and four other people. Another bomber, who had trained with terrorists in Syria, detonated a crude bomb at the Parsons Green Underground station in London, England. His device was also defective, but 30 people were injured in the explosion. (The statistics in this chapter derive from the Mineta Transportation Institute’s database of attacks on public surface transportation.)

In 2018 two persons were stabbed by a man at the Central Railway Station in Amsterdam, the Netherlands. That same year, another man injured one person and briefly held another hostage at the train station in Cologne, Germany, in what police believe may have been a foiled terrorist attack. In 2019 an assailant stabbed three persons at a train station in Manchester, England. In a separate 2019 incident, small bombs were mailed to London airports and the Waterloo train station.

These more recent, and mostly crude, attacks were typical of the kinds of primitive terrorist incidents seen in the late 2010s, and they contrast with the far deadlier attacks on rail passengers seen earlier: A 2004 commuter train bombing in Madrid killed 192 persons; the 2005 bombings of London Transport killed 52; the 2006 Mumbai commuter train bombing killed 209; the 2010 bombings in Moscow’s metro killed 40; and the 2016 bombing of a metro station in Brussels killed 20.

Recent Terrorist Campaigns

Terrorists do not descend to earth like extraterrestrials. They arise from political causes, as well as personal circumstances. Much of the recent terrorist activity directed against transportation targets reflects specific terrorist campaigns. What this means for those charged with security is that the level of threat varies greatly according to location.

History is filled with accounts of bandits holding up trains and insurgents derailing them. Bombs in trains or train stations were less common until the late 20th century. The Provisional Wing of the Irish Republican Army (IRA) was one of the first groups to initiate a bombing campaign targeting British rail passengers. For the first few years of “The Troubles,” the IRA concentrated its violence in Northern Ireland, but in 1973, in an effort to increase pressure on the British government, the group began carrying out bombings in London, initially focusing on government buildings but later targeting train and Tube stations, shopping centers, restaurants, and tourist sites.

Many of these attacks caused injuries, and some caused fatalities, but the IRA was cautious. It had far more capability than the generation of jihadist terrorists that operated in the United Kingdom decades later—the IRA knew how to make bombs. Its goal, however, was attention and political leverage, not high body counts. It worried about provoking backlashes that could repel its supporters and reduce its financial contributions from sympathizers abroad. Above all, it wanted to impose a high psychological and economic burden on the British public and its government as a price for remaining in Northern Ireland.

The IRA often provided authorities with warnings of where and when bombs were about to go off, claiming that it did so in order to give them time to evacuate the area and avoid casualties. While it is true that the group generally operated within self-imposed constraints, warnings were not always given; the information was not always precise and was sometimes misleading; some warnings

were deliberate hoaxes; and in some cases, the terrorists planted secondary devices aimed at killing police and soldiers sent to deal with the primary bomb. Providing “warnings” offered the group cover when people were killed by IRA bombs, allowing it to shift culpability to the authorities. The warnings were also a way to increase disruption.

In one incident, the IRA set off a bomb at London’s Paddington Station at 4:20 a.m. on February 18, 1991. The explosion did little damage and injured no one, but it established credibility for an IRA call warning that bombs would explode at all 11 of the mainline train stations in London at 7:45 a.m. There had been no IRA bombings of London stations since 1976, and the authorities were reportedly slow to react. They tried to check for devices before halting all incoming trains and entirely evacuating every station, which would have put thousands of people into the streets. As a result, passengers were still present on a crowded platform when a bomb hidden in a trash can went off in Victoria Station at 7:40, 5 minutes before the detonation time given in the warning. The explosion killed one person and injured 38. Fearing further casualties, authorities shut down all of the other train stations for the first time in history.

The deadly cat-and-mouse game continued. In February 1992, the IRA gave just a 10-minute warning before a bomb exploded at the London Bridge Station. There had been enough real bombs to give the warnings credibility. The warnings were calculated to force the authorities to evacuate large facilities quickly, creating potential panics, giving them just enough time to respond but not enough time to search. To avoid catastrophic casualties, the IRA usually used small, although still lethal, devices, and the attacks escalated over time. The IRA campaign against British transport is described in detail in [Jenkins and Gersten \(2001\)](#).

The Basque separatist group Euskadi Ta Askatasuna (ETA) was a contemporary of the IRA. Beginning in the late 1960s, it carried out a campaign of assassinations, armed assaults, and bombings directed against Spanish government officials, Spain’s security forces, and civilian targets, which included rail systems. ETA claimed responsibility for or was directly linked to 25 attacks on transportation targets; another 77 unclaimed attacks were also attributed to its terrorist campaign. ETA was even more cautious than the IRA about civilian casualties. Its attacks on transportation targets were intended to demonstrate capabilities, cause fear and disruption, and discourage tourism as part of its campaign of economic warfare. Most of its attacks targeted empty trains and stations. When ETA planted bombs where civilian bystanders were likely to be killed, it—like the IRA—provided warnings. Only two of its many attacks on rail targets resulted in fatalities: Four people were killed by bombs detonated in the Atocha and Chamartín train stations in Madrid on July 29, 1979. Five other attacks resulted in injuries.

However, as is the case in many terrorist campaigns, there was pressure to escalate the violence. In December 2003, Spanish police thwarted an ETA plot to detonate two suitcases containing bombs with more than 50 kilos of explosives on a train in Madrid’s Chamartín train station on Christmas Eve. The arrest and interrogation of the individual attempting to plant one bomb alerted police to the second bomb, which was already on its way to Madrid. Authorities halted the Madrid-bound train, evacuated its passengers, and defused the device. The terrorists had also planned to detonate two smaller devices on the rail lines in Aragon. On February 29, 2004, police arrested two ETA members driving vans carrying 536 kilos of explosives on their way to Madrid. The terrorists planned to detonate a massive vehicle bomb there on the day of the national elections, scheduled for March 14.

It is understandable, therefore, that ETA would be at the top of the list of suspects when bombs went off on commuter trains arriving at Madrid’s Atocha rail station on March 11, 2004. However, police quickly realized that the design of the devices and the explosives used were different from devices fabricated by ETA. This, plus other evidence that was gathered, convinced police that the bombs were the work of jihadists, not Basque separatists. The police were therefore surprised when the Spanish Minister of Interior announced that ETA was responsible, a claim that Spanish political leaders persisted in making despite overwhelming evidence to the contrary. This infuriated crowds who considered the government’s claim as an attempt to cover up what people saw as a terrorist attack in retaliation for Spain joining the United States in the invasion of Iraq. A detailed analysis of the March 11 bombing by [Reinares \(2017\)](#) makes the point that both sides were wrong: ETA was not responsible for the attack, despite continuing conspiracy theories; and the decision by the jihadists to carry out a major terrorist attack in Spain was made in December 2001, the jihadist network in Spain that carried out the attack assembled in March 2002, and al Qaeda’s senior leadership approved the plan in October 2003, all long before Spain joined the American-led coalition to invade Iraq.

Transportation targets were not a feature in the terrorist campaigns of Germany’s Red Army Faction or Italy’s Red Brigades terrorists in the 1970s and 1980s. Neo-fascists in Italy, however, bombed an express train in Italy in 1974, killing 12 people and injuring 48, and they bombed the Bologna Central railway station in 1980, killing 80 people and injuring more than 200. The Sicilian Mafia was believed responsible for the bombing of a passenger train in 1984 that killed 16 and injured 256.

Escalating Violence

Terrorist attacks on transit targets in the 1970s and 1980s were seldom intended to cause mass casualties; separatists and ideologically motivated groups had constituencies, which they did not want to alienate. (Italy’s neo-fascists and Mafia were exceptions.) However, as terrorists in general escalated their violence in the 1980s and 1990s, attacks on transportation systems became deadlier. Transportation venues became killing fields.

On March 20, 1995, members of Aum Shinrikyo, an apocalyptic Japanese cult, released plastic bags filled with sarin, a lethal nerve agent, on five Tokyo subway trains as they reached the central part of the city, where government buildings were concentrated (see [Jenkins and Gersten, 2001](#)). Because the quality of the sarin manufactured by the group was poor and the method of dispersal was primitive, only 12 persons died in the attack, although more than 1000 people were directly exposed to the fumes, and more than 5500 were treated for a variety of medical conditions related to the attack. Had the sarin been more potent and the dispersal more effective, the number of deaths would have been much higher.

Weeks later, Aum Shinrikyo members still at large attempted another attack on Tokyo's subway system, this time using hydrogen cyanide. The dispersal scheme did not work as planned, although four people were injured. With these incidents, chemical weapons had become part of the threat matrix. However, old-fashioned bombings and derailments caused by sabotaged rails continued to account for the highest-casualty attacks.

In July 1995, a small group of Algerian extremists, angered by France's support of the Algerian government in its war against Islamists, detonated an explosive device aboard a French commuter train at the St. Michel station in Paris. Seven persons were killed and more than 80 were injured in the blast. This was the first of a series of attacks, several of them directed against transportation targets, that continued until mid-October. In August, the terrorists unsuccessfully attempted to derail a high-speed train between Lyon and Paris. In October, a bomb exploded on a commuter train in Paris, injuring 30. The campaign, which is described in [Jenkins et al. \(2010a, 2010b\)](#), appeared to end when police arrested most of the members of the terrorist network in November 1995, but another bomb detonated on a commuter train in Paris in December 1996, killing four persons and injuring more than 90.

Police uncovered several jihadist plots against French rail targets in the early 2000s, and a 2015 attempt to gun down passengers on the high-speed train between Amsterdam and Paris was foiled by passengers. Although a group of French and Belgian veterans of the Islamic State carried out devastating terrorist attacks in Paris in November 2015, transportation targets did not figure in their planning. However, the same network carried out a bombing on the Brussels metro in 2016, killing 20 people (see [Jenkins et al., 2016](#)).

Between 2003 and 2017, Chechen rebels and Islamists carried on a deadly bombing campaign targeting passenger trains and stations and urban metros in Moscow and other cities in Russia, killing more than 240 people and wounding many hundreds more (see [Jenkins and Butterworth, 2014a](#)). In 2017 Islamic State supporters from Tajikistan attempted to sabotage a high-speed train, causing it to derail in the path of another train traveling in the opposite direction. The trains did not crash and there were no casualties, but the coaches sustained extensive damage.

India has experienced more attacks on surface transportation than any other country and has suffered some of the deadliest incidents. Terrorist attacks on commuter and intercity passenger trains in India became increasingly common in the first years of the 21st century. In 2002 sabotage caused the derailment of a high-speed passenger train crossing a bridge, resulting in between 130 and 200 fatalities; and 209 people died in the 2006 bombing of a Mumbai commuter train. In 2007 bombs in the coaches of a passenger train killed 70. Another 148 persons died in the derailment of another passenger train in 2010, although the conclusion that it was sabotaged has been disputed.

The deadliest incident in a subway occurred in Daegu, South Korea, where a mentally unstable individual deliberately started a fire on the train, ultimately killing 198 persons. (The incident with the most total fatalities took place during Angola's civil war in 2001, when rebels derailed a passenger train and then shot fleeing survivors; about 100 people were killed in the initial attack and more than 150 were chased and shot down—in all, 256 people are believed to have died.)

In comparison to the high volume of attacks in South Asia and Russia, Canada and the United States have suffered few attacks. Between 1970 and September 2019, 12 attacks occurred in the United States, and four occurred in Canada. Only three of these—two of them by deranged individuals—resulted in fatalities.

Attacks by Islamist fanatics have been the source of greatest concern in the first decades of the 21st century. The deadliest attacks—most notably the 9/11 attacks on the United States—and many of the foiled plots during the first decade of the 21st century were part of an ongoing global campaign of terrorism directed against a variety of targets in countries branded by jihadists as enemies. This terrorist campaign was inspired by continuing exhortations from al Qaeda leader Osama bin Laden and other al Qaeda communicators, and it was waged by individuals who subscribed to al Qaeda's ideology, some of whom had trained in its training camps or had learned from information posted on various jihadist websites. This is not to say that the campaign was centrally directed.

Jihadists were responsible for the most-lethal surface-transit-system bombings, including the 2004 Madrid train bombing, the 2005 London Tube bombings, the 2006 Mumbai train bombing, and the train and metro bombings in Russia. An analysis of jihadist targeting preferences in Europe between 1996 and 2016 ([Hemmingby, 2017](#)) showed that transportation (aviation and surface) was the most frequent intended target in jihadist plots, ahead of public areas in general.

Jihadist attacks typically aim at death, not disruption. In Europe and North America, jihadists have been responsible for only 11 of the attacks on surface transportation, but approximately half of the total casualties. The jihadists have also made extensive use of suicide bombings.

The term "jihadist" gets slippery outside of Western countries. The Chechen insurgency in Russia started out as an independence movement, but over time it acquired a more radical jihadist ideology. Whether China's Uighur combatants, who have carried out deadly terrorist attacks on transportation targets, should be counted as separatists or jihadists depends on the perspective of the viewer. Their quarrel is largely local—they oppose rule by Han Chinese—although they have reportedly received support from al Qaeda, and some of their leaders describe their struggle as a religious jihad (see [Jenkins and Butterworth, 2014b](#)). And while some of the attacks on India's railways are inspired by jihadist ideology, Indian authorities believe there is considerable evidence that the jihadists are proxies of Pakistan's intelligence service in a continuing covert war between the two countries.

Police have thwarted numerous jihadist plots, in addition to the foiled attacks on transportation targets. British authorities uncovered plots to disperse poison gas and ricin on London trains in 2002 and 2003. Just two weeks after the July 7, 2005, bombings on London transport, a second jihadist group tried to replicate the attack, but their devices failed. In November 2005, Australian authorities uncovered a plot to carry out attacks in Sydney and Melbourne; targets included Melbourne's landmark Flinders Street Rail Station. In 2006 homegrown terrorists inspired by jihadist ideology planned to detonate two bombs hidden in suitcases on German commuter trains, but the devices failed to detonate. That same year, Moroccan police uncovered a jihadist plot to bomb Milan's metro. In 2008 police in Spain uncovered a jihadist plot to carry out bombings on Barcelona's metro.

Several jihadist plots have targeted subways and commuter trains in the United States. In 2003 jihadists planned to release cyanide gas in New York's subway system but reportedly were diverted to another project. Authorities thwarted a plot by homegrown jihadists to bomb a subway station in midtown Manhattan in 2004. In 2006 Lebanese authorities discovered another jihadist plot to bomb trains in the tunnel under the Hudson River. An American jihadist who traveled to Pakistan offered his experience as a former employee of the Long Island Railroad to help al Qaeda plan and execute a terrorist attack on that system. He was arrested by Pakistani authorities and returned to the United States in 2008. The following year, acting on a tip from British authorities, FBI agents uncovered the existence of a group of jihadists planning to carry out suicide bombings on New York's subways. Another individual was arrested the following year in an undercover operation. He intended to carry out a bombing on the Washington metro. The discovery of notes in Osama bin Laden's compound indicated that he also was contemplating attacking trains in the United States on the tenth anniversary of September 11, 2001.

The jihadist campaign reached a high point in the mid-2000s, with three major terrorist attacks on trains in Madrid, London, and Mumbai, and seven foiled plots between 2004 and 2006, after which the campaign declined.

The continuing revelation of jihadist plots underscores the reality of the threat—there is always someone out there planning to do something. And, given ongoing conflicts in the Middle East, it must be presumed that the threat will continue. Historical experience shows that if they are not thwarted, these attacks can have terrible consequences—fatalities in the scores or hundreds, potentially more if chemical weapons are used effectively.

There have been several plots involving crudely refined ricin possibly being spread in train stations or coaches, none of which would have caused mass casualties. And there is at least one case of deliberate radioactive contamination. In 1974 an individual with a history of mental illness spread nonlethal quantities of radioactive Iodine-131 on train coaches in Austria. The perpetrator intended his action, which received widespread publicity, as a protest against Austria's treatment of the mentally disturbed (DeLeon et al., 1978).

Nevertheless, terrorist attacks are statistically rare, and the overall threat to life is only a small fraction of the total casualties resulting from ordinary transportation accidents—trains colliding with vehicles at grade crossings, people walking on tracks or in tunnels, suicides by train.

Trends in Tactics and Targeting

Terrorist attacks on surface transportation targets have increased over the long run. Between 1970 and 1979, terrorists carried out a total of 23 surface transportation attacks that caused fatalities. (Only incidents with fatalities are included in this analysis to mitigate increases that may be due solely to better reporting.) The number grew to 44 attacks with fatalities in the 1980s, 271 in the 1990s, 446 in the decade between 2000 and 2009, and 814 in the decade between 2010 and (October 15) 2019. These numbers include attacks on both bus and passenger rail targets.

Many of the attacks on surface transportation involved few or no fatalities and did not make headline news, but 11 of them since 9/11 resulted in 50 or more deaths, and three of the attacks killed approximately 200 people each. The total number of fatalities in these 14 attacks would translate into the casualties from anywhere between 6 and 18 commercial airline crashes. The regulatory evaluation method used in 2001 to derive this estimate for airline crashes was based on an average of 78 passengers for a loaded Boeing 737. At the high end, one can instead look at the 1988 PanAm 103 and 2015 Russian Metrojet sabotage incidents to calculate an average of more than 200 fatalities per incident. The average number of fatalities in the four airline crashes caused by terrorist sabotage since 9/11 is 106.5, which would give a rough equivalent of 13 crashes. One can imagine the fear and furor that would have resulted if 13 commercial airliners had been brought down by terrorists after 9/11.

The total number of fatalities caused by attacks on surface transportation also increased, although both the number of incidents and the number of fatalities declined sharply after 2015. Lethality is more complicated. The Mineta Transportation Institute's database of attacks on surface transportation tracks lethality, measured as fatalities per attack (FPA). The annual average FPA (total fatalities divided by the total number of attacks) charted over the years shows whether attacks in general are becoming bloodier.

The record shows that overall lethality—the average number of FPA—has declined slightly over the years. This may reflect the growing volume of low-level attacks, which lowers the average lethality. For example, two attacks that caused 150 and 85 deaths occurred in 1980, when there were far fewer incidents overall, and thus drove up the average FPA for those years. This sharp peak, in turn, distorts the linear trend line. Excluding these two incidents tilts the trend line slightly upward. We also cannot dismiss the possibility that some portion of the increased volume of incidents recorded in recent years may be due simply to better reporting as more local news accounts become available via the Internet.

Bombings account for most of the attacks on surface transportation and most of the casualties. They are also the most lethal form of attack. Improvised explosive devices account for 51% of the total attacks. Explosives, including improvised explosive devices, grenades, mines, and dynamite, account for almost 60% of all attacks (see Jenkins and Butterworth, 2010). Shootings and armed assaults—incidents involving only firearms, including hijackings—account for 15%. Incendiary devices and other forms of arson account for nearly 13% of all attacks, and mechanical means of sabotage (mostly of rails) account for about 3%.

Not all attacks on rails themselves are intended to derail trains. In many cases, the saboteurs—anarchists, environmental extremists, opponents of nuclear power and weapons—seek only to cause disruption, thereby bringing attention to their cause and imposing an economic cost and inconvenience on society.

Recently, there has been an increase in more-primitive terrorist tactics—stabblings and vehicle-ramming attacks (see [Jenkins and Butterworth 2018a, 2018b; Jenkins et al., 2019](#)). There are also incidents in which people are pushed off platforms into oncoming trains. The trend reflects changes in how terrorists are recruited. In the past, terrorists usually joined small underground groups, which carefully vetted them to prevent infiltrators or individuals judged to be unreliable. In contrast, much of today's recruiting is done by exhortation. Jihadist groups, in particular, have exploited the Internet and social media to inspire followers abroad to carry out attacks wherever they are, using whatever means they have available. Often operating alone, the recruited individuals have limited resources and skills. Guns, where they are easily available, may be used. Knives and vehicles are readily accessible almost everywhere.

Remote recruiting emphasizes the violence of attacks as much as the cause, and it appeals to a broader audience of potential actors that includes not only fanatics, but individuals with histories of aggression and violence, as well as mentally disturbed persons. Intense media coverage adds to the attraction. The result is random attacks by individuals untethered to organizations.

It is low-yield ore—there are handfuls, not hundreds, of attacks in any country. The risk to the individual traveler is infinitesimal. Nonetheless, the very randomness of the violence creates an impression that no location is safe. The psychological effects of terrorism, which is above all about creating fear and alarm, create anxiety and psychological disorders that can last years. Media coverage creates a vast audience of vicarious victims. At the societal level, continuing fear fuels suspicion, generates demand for extreme measures, and exacerbates social divisions and other tensions.

Security Issues

Security for public surface transportation must be realistic. Protecting public places that by their very nature require easy access is difficult and costly. Prevention of all attacks is virtually impossible. (The challenges of protecting surface transportation against terrorist attacks are described in detail in [Jenkins, 2017](#); see also [Taylor et al., 2005](#).)

It can be argued that the extraordinary security measures deployed over the years at commercial airports have significantly reduced the number of airline hijacking and sabotage incidents, which is true, although other factors have also contributed to the decline. However, an aviation security model will not work for surface transportation. The number of passengers using urban rail in the United States daily is more than four times the number boarding planes.

The architecture of the two systems is also completely different. Airports are designed to funnel passengers onto airplanes. Transit stations are open systems designed to let as many people on and off trains as quickly as possible. Given the limits of current and even near-term technology, to replicate aviation security for surface transportation would require several hundred thousand screeners, add unacceptable delays to daily commutes, and impose costs that would likely destroy public transportation as it currently exists. Screening of 100% of passengers can be imposed only where governments exercise great social control and there are no alternate ways to travel. It is unlikely to work elsewhere.

The check-in and baggage pickup areas at airport terminals, which are outside of the checkpoints for passengers boarding planes, have also been venues for terrorist attacks. Security officials face the same problems there as do those charged with the security of transportation—crowds of people in public areas. Some airports have established outer security perimeters to surveil and screen passengers as they arrive at the airport or enter the terminal. This is a completely separate function from passenger screening to prevent weapons and explosives from being smuggled on airlines.

There is another important conceptual difference between protecting airplanes and protecting transit systems. Airline security is front-loaded, that is, it aims at deterrence and prevention. In contrast, surface transportation encompasses a broader spectrum of countermeasures, including deterrence, detection, prevention, mitigation of casualties through the design of coaches and stations, and reducing fatalities by prompt detection and rapid intervention when events do occur and by facilitating evacuation and rescue. In the case of armed assaults, rapid intervention can prevent further killing. The multitude of entry points and passenger volumes limit what can be done at the front end, but there are still opportunities to save lives, even during a terrorist event. Detection can be improved, and casualties can be further mitigated by station and coach design.

Aviation security is highly regulated. Rules must be standardized to ensure that passengers go through the same security procedures no matter where they enter the system. The same strict regulatory approach would not work for surface transportation, nor is it necessary. A train or bus passenger does not enter an international or even a national system. Local systems vary greatly in size, configuration, passenger volumes, and other attributes. Security may be provided by local police, specialized transport police, proprietary police departments, or private security. Instead of standardized rules, operators follow a best-practices approach.

Governments and other sources provide information about terrorist events and lessons learned. Individual transportation operators decide what is appropriate for their local threat situation and circumstances. Government may provide additional support in the form of intelligence about possible threats, research, and security-technology assessments and may further provide material assistance to assist operators in the development and deployment of new security measures.

Rail operators and transit systems have increased the presence of security personnel. They have added cameras to improve surveillance, assist in rapid diagnosis if an event occurs, and deter terrorists who want to avoid capture. Smart camera systems alert monitors to anomalies, for example, objects that do not move when they should or movement where there should be none, and can assist in tracking the past or current movements of persons arousing suspicion. Some systems have installed platform-edge doors.

Explosives-sniffing dogs are increasingly being used to search for bombs and detect vapor trails left by explosives. Remote explosives-detection technologies are being deployed on an experimental basis, but there is no near-term technological solution.

Some transportation systems have implemented random passenger screening, which introduces uncertainty for attackers. It is difficult to determine how much deterrent value such screening efforts may have, but they establish protocols, allow for training, and thereby provide a foundation for an orderly surge in security, should intelligence provide some warning of an increased threat level, or in the immediate aftermath of an event to discourage further planned or copycat attacks. These measures are also flexible and can easily be scaled back when an immediate threat is seen to diminish.

Most security measures enhance security, but local governments and transit operators, strapped for cash, can do only so much. They are often dealing with ordinary crime, aggressive behavior, homeless populations, and other challenges that upset passengers and discourage ridership. Whatever security measures are put into place must be sufficient to handle the huge volumes of passengers, must not trample civil liberties, must be economically sustainable, and must offer a net security benefit. Implementing such measures will require intelligence and imagination.

A realistic appraisal of the threat posed by terrorists must take into account the minuscule risk that they pose to passengers and the broader society. The number of persons killed by terrorists attacking rail passengers is a tiny fraction of the number killed in passenger railway accidents, mainly collisions with vehicles at level rail crossings or people being hit by trains while walking on the tracks. In 2019 The European Statistical System (Eurostat) found that in Europe, the number of people who die in suicides by train is greater than the number killed in accidents. To be sure, terrorist attacks attract more headlines and generate fear, and this causes people to exaggerate the threat.

An often overlooked cost of terrorism is the panoply of psychological effects—the terror—that can be immediately observable but may also persist over long periods of time and have a corrosive effect on society in general (see [Greenberg et al., 2009](#); [Sheppard, 2009](#)). Although the level of apprehension may seem to be a broader societal issue beyond the responsibility of transportation system operators, it may be directly affected by how effectively and efficiently government and transportation operators handle terrorist crises.

Often, this is a matter of communications—providing accurate information about what is going on, rapidly setting up centers where concerned people can obtain information about relatives who may be victims, arranging and informing the public about alternate means of transportation, rapidly countering rumors, and other messaging to reflect competence. Security measures rapidly put in place to reassure nervous passengers and deter copycats or malicious pranksters can acknowledge and respond to reports of suspicious objects or people from passengers who are understandably frightened and therefore hyperalert.

Employee and passenger awareness counts. Just as airline passengers have become the last line of defense against terrorists on airplanes, rail passengers can become the first line of defense in ensuring their own security. Daily riders know the environment intimately and are able to identify suspicious objects or behavior.

During the IRA bombing campaign, British authorities depended on alert passengers, already sensitized by IRA bombings, to report suspicious objects within minutes of the objects' being placed. Reports were followed up with visual diagnosis through cameras and personnel on scene. Bomb squads were summoned only when necessary, thereby minimizing disruption. This made London's Tube and trains a hostile environment for terrorist operatives. One can see the results over time, as the terrorists gradually edged away from their preferred targets in crowded central London to more-remote venues.

The Mineta Transportation Institute has examined the issue of citizen awareness in detail (see [Jenkins and Butterworth 2018a, 2018b](#)). Researchers there found that since 1970, on-site police, security personnel, transportation employees, and passengers have prevented 10.6% of all attacks on surface transportation by reporting suspicious objects. The overall detection rate was 14.2% in the higher-income countries—particularly in Europe and North America—where, in response to terrorism, "See Something, Say Something" campaigns have been emphasized. For trains and subways, the figure is 14.8%.

Current "See Something, Say Something" campaigns need to be further evaluated to see if the message is getting through and to determine how to better engage the public. Communications by members have to be facilitated—an easier task in the age of widespread mobile phones. Procedures have to be established to ensure rapid diagnosis and response. Callers need to be acknowledged for their efforts, even when an effort turns out to be a false alarm. Disruptions must be minimized.

Finally, it should be noted that police intelligence operations account for a majority of the failures of terrorist plots. Intelligence collection is a national effort that extends far beyond the security force charged with protecting a particular transportation system. However, transit police forces have developed creative ways to enlist staff, riders, vendors, and nearby merchants as additional eyes and ears. In today's terrorism, the tip that leads to uncovering a terrorist plot may just as easily come from a local source as from a foreign intercept.

References

- DeLeon, P., Jenkins, B., Kellen, K., Krofchek, J., 1978. *Attributes of Potential Criminal Adversaries to U.S. Nuclear Programs*, The RAND Corporation, Santa Monica, CA.
- Greenberg, N., Rubin, G.J., Wessely, S., 2009. The psychological consequences of the London bombings. Available from: <https://www.kcl.ac.uk/kcmhr/publications/assets/cbrn/Greenberg2009-psychologicalconsequencesLondonbombingsbookchapter.pdf>.
- Hemmingby, C., 2017. Exploring the continuum of lethality: militant Islamists' targeting preferences in Europe, *Perspectives on Terrorism*, No. 5. Available from: <http://www.terrorismanalysts.com/pt/index.php/pot/article/view/642/html>.
- Jenkins, B.M., 2017. *The Challenge of Protecting Transit and Passenger Rail: Understanding How Security Works Against Terrorism*, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/challenge-protecting-transit-and-passenger-rail-understanding-how-security-works-against>.
- Jenkins, B.M., Butterworth, B.R., 2010. *Explosives and Incendiaries Used in Terrorist Attacks on Public Surface Transportation: A Preliminary Empirical Analysis*, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/explosives-and-incendiaries-used-terrorist-attacks-public-surface-transportation>.

- Jenkins, B.M., Butterworth, B.R., 2014a. By the Numbers: Russia's Terrorists Increasingly Target Transportation. Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/Numbers-Russia%E2%80%99s-Terrorists-Increasingly-Target-Transportation>.
- Jenkins, B.M., Butterworth, B.R., 2014b. The Terrorist Attack in Kunming, China: Does It Indicate a Growing Threat Worldwide? Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/Terrorist-Attack-Kunming-China-Does-It-Indicate-Growing-Threat-Worldwide>.
- Jenkins, B.M., Butterworth, B.R., 2018a. An Analysis of Vehicle Ramming Attacks as a Terrorist Threat, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/Analysis-Vehicle-Ramming-Terrorist-Threat>.
- Jenkins, B.M., Butterworth, B.R., 2018b. Does "See Something, Say Something Work?" Mineta Transportation Institute, San Jose, CA. Available from: https://transweb.sjsu.edu/sites/default/files/SP-1118_SeeSomethingSaySomething.pdf.
- Jenkins, B.M., Butterworth, B.R., Clair, J.F., 2010a. The 1995 Attempted Derailing of the French TGV (High-Speed Train) and a Quantitative Analysis of 181 Rail Sabotage Events, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/1995-attempted-derailing-french-tgv-high-speed-train-and-quantitative-analysis-181-rail>.
- Jenkins, B.M., Butterworth, B.R., Clair, J.F., 2016. Trains, Concert Halls, Airports, and Restaurants—All Soft Targets: What the Terrorist Campaign in France and Belgium Tells us About the Future of Jihadist Terrorism in Europe, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/trains-concert-halls-airports-and-restaurants%E2%80%99all-soft-targets-what-terrorist-campaign>.
- Jenkins, B.M., Butterworth, B.R., Clair, J.-F., Trella, J., 2019. An Exploration of Transportation Terrorist Stabbing Attacks, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/SP0319-Terrorist-Stabbing-Attacks-Public-Transportation>.
- Jenkins, B.M., Butterworth, B.R., Gersten, L.N., 2010b. Supplement to MTI Study on Selective Passenger Screening in the Mass Transit Rail Environment, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/supplement-mti-study-selective-passenger-screening-mass-transit-rail-environment>.
- Jenkins, B.M., Gersten, L.N., 2001. Protecting Public Surface Transportation Against Terrorism and Serious Crime: Continuing Research and Best Security Practices, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/protecting-surface-transportation-systems-and-patrons-terrorist-activities>.
- Reinardes, F., 2017. Al-Qaeda's Revenge: The 2004 Madrid Train Bombings, The Woodrow Wilson Center Press/Columbia University Press, Washington, DC.
- Sheppard, B., 2009. The Psychology of Strategic Terrorism: Public and Government Responses to Attack, Routledge, Abingdon, UK.
- Taylor, B.D., et al., 2005. Designing and Operating Safe and Secure Transit Systems: Assessing Current Practices in the United States and Abroad, Mineta Transportation Institute, San Jose, CA.

Further Reading

- Goodman, A.I., 2003. Spain police find railway bomb, CNN.com. Available from: <https://www.cnn.com/2003/WORLD/europe/12/26/spain.bomb/index.html>.
- Jenkins, B.M., Butterworth, B.R., 2007. Selective Screening of Rail Passengers, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/selective-screening-rail-passengers>.
- Jenkins, B.M., Butterworth, B.R., 2016. Long-Term Trends in Attacks on Public Surface Transportation in Europe and North America, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/Long-Term-Trends-Attacks-Public-Surface-Transportation-Europe-and-North-America>.
- Jenkins, B.M., Trella, J., 2012. Carnage Interrupted: An Analysis of Fifteen Terrorist Plots Against Public Surface Transportation, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/carnage-interrupted-analysis-fifteen-terrorist-plots-against-public-surface-transportation>.
- Mineta Transportation Institute, 2007. The Fourth National Security Summit: Approaches to Passenger Screening, Mineta Transportation Institute, San Jose, CA. Available from: <https://transweb.sjsu.edu/research/Fourth-National-Security-Summit-Approaches-Passenger-Screening>.
- Múgica, F., 2004. La Extrana 'Caravana de la Muerte,' elmundo.es. <https://www.elmundo.es/elmundo/2006/05/19/espana/1148055732.html>
- Eurostat, 2019. Rail Accident Fatalities in the EU. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php/Rail_accident_fatalities_in_the_EU

The Swedish Vision Zero: A Policy Innovation

Matts-Åke Belin, KTH Royal Institute of Technology, Stockholm, Sweden; Swedish Transport Administration, Borlänge, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	675
Vision Zero as a Public Road Safety Policy Innovation	675
A Traditional Approach to Road Safety	676
The Vision Zero Approach	676
Vision Zero—To Make Things Happen	678
Vision Zero—Outputs and Outcomes	678
References	679
Further Reading	679

Introduction

Sweden has a long experience of a holistic, systematic, and strategical road safety work, introduced in the late 1960s in connection to and after the switch to right-hand driving in 1967. One important step was the establishment of a special government agency for road safety, the Swedish Traffic Safety Agency, in 1968. The number of people killed on the roads dropped from 16.7 killed per 100,000 inhabitants in 1966 to 9.1 killed per 100,000 inhabitants in 1982—a decrease of more than 40%. During the 1980s, the positive trend was broken and the road injury figures and the traffic growth began to follow each other like they had in the 1950s and early 1960s: the more car traffic, the more people were killed on the roads. In 1989, Sweden had 10.6 fatalities per 100,000 inhabitants and was, once again, approaching four-figure numbers of road deaths. A sense of lost control was spreading in society and this, together with political pressure, led to something more radically done about the problem. This eventually (in 1993) led to the closure of the Swedish Road Safety Agency and to an enhanced role for the Swedish Road Administration.

Parallel with this process to change the institutional prerequisites for the national road safety work, Sweden faced a severe economic recession in the first half of the 1990s. During the period 1990–93, the Swedish GDP dropped by almost 5% and the unemployment rate increased dramatically. From a road safety point of view, at least in the short run, we know that economic recessions might be good for safety and this was the case for Sweden in the 1990s. The number of fatalities dropped by more than 40% between 1989 and 1996 from 904 to 508.

However, with the recent negative trend in deaths and injuries on the roads in the 1980s freshly in mind, together with the knowledge that the recession gradually would change into a more normal pattern of economic development, road safety experts in Sweden shared a common insight that something different was needed. Strategies adopted in the late 1960s and business as usual was not a long-term sustainable strategy.

In 1994, a development work was established. This was led by Claes Tingvall, a former Road Safety Director of the Swedish Road Administration, and documented in a memorandum entitled “Vision Zero—An Idea for a Road Transport System Without Health Losses.” Vision Zero quickly attracted political interest, and the Minister of Transport and Communications at the time, Ms. Ines Uusmann, initiated a broad public policy process that culminated in October 1997 with the Swedish Parliament adoption of Vision Zero as a new national public road safety policy ([Swedish Government and Committee on Transport and Communications, 1997](#)).

It is now more than 20 years ago that the Swedish parliament adopted Vision Zero and this approach to safety has since then been spreading around the world and to other sectors of society. In this paper, the Swedish Vision Zero and its basic characteristics, its implementation, and results are described and explored.

Vision Zero as a Public Road Safety Policy Innovation

The Swedish Vision Zero is described and analyzed in many different studies and documented in scientific literature ([Tingvall et al., 2000](#); [Johansson, 2009](#); [Belin and Tillgren, 2012](#), [Kristianssen, 2018](#)), governmental reports, conference proceedings, and presentations. Vision Zero is a long-term goal to eliminate fatalities and serious injuries and it is based on an ethical imperative, which states that society can never accept that people are killed, or seriously injured when they travel in our road transport system. However, Vision Zero is more than this ethical imperative. It is a comprehensive theory, a belief system of different ideas, which connects to each other in a logical way. Broadly speaking, it contains ideas about the road safety problems and its causes, what future conditions the society should aim for to achieve “safety,” and the appropriate ways to reach these short- and long-term conditions. In order to explore the innovative characteristics of Vision Zero, it could be useful to compare it with a traditional approach to road safety ([Belin and Tillgren, 2012](#)). As follows, the ways that society defines its problems are paramount for what both people and organizations believe are appropriate strategies and goals. Let us start with looking at the predominant traditional approach and then move on to Vision Zero.

Population/ system perspective	People are victims in a complex technological system	People in general are influenced by wrong social norms
	Some individuals do not have the psychological and biological traits which are needed	We have subgroups of people who do not comply with social norms
Individual risk groups	Human traits “Error”	Human willingness “Violation”

Figure 1 Accident causations theories.

A Traditional Approach to Road Safety

The traditional road safety work focuses mainly on the problem with accidents or, as some researchers and policy-makers prefer to say, crashes. The definition of road safety problems is, to a significant extent, based on in-depth studies, which draw the conclusion that roughly 90% of all road traffic accidents can be explained by human factors, solely or in relation to its environment (Evans, 2004). Which type of human factors, human errors, or human violations that causes most safety problem is, however, an area of constant debate among scholars through the history. System-oriented human factor experts tend to focus on human errors, while social norms individual human behavior experts tend to focus on individual traits and subgroups (risk groups) (Fig. 1). In general, the individual perspective—and human error explanations—dominates among road safety experts around the world.

The statement that 90% of all accidents (crashes) are due to human factors is a powerful tool to direct interventions and to support society in their efforts to distribute responsibility for safety. The principal challenge, according to a traditional approach, is to prevent conscious and subconscious faulty individual human actions and its ultimate aim is to create a perfect human being, who always does the right thing, or to prevent faulty humans to get access to the system. The main strategy in order to achieve this has been to regulate individual road users' access to the system and its individual road user behavior (e.g., capability and willingness to comply) (Friedland et al., 1990). According to a traditional point of view, the individual road users are responsible for controlling and managing their own risks that may occur when traveling on the road transport system; and the individual road users are therefore ultimately responsible for their own and other's safety. If something goes wrong in the road transport system and an accident occurs, it is, in almost all cases, possible to find an individual to blame. The Vienna Convention of 1949 and the Vienna Convention of 1968 are excellent examples of how this universal principal of responsibility is formalized, and this is also reflected in the insurance systems surrounding road transport. Traditional road safety work is therefore largely focused on the so-called three E strategies—engineering, enforcement (regulation), and education—directed toward individual road users in order to modify their behavior with the intention of reducing the number and severity of accidents.

The ultimate goal, with a traditional approach to road safety, is to gradually reduce risks within the system but there is a limit for how far we can go. According to this traditional approach, it is unrealistic to eliminate the human factor and reduce the number of accidents to zero. People are willing to take risks and neither individuals nor society wants absolute safety and the perception is that it would be too costly to eliminate all accidents. Society therefore has to accept losses of human life and long-term health. The perception is that fatalities and serious injuries in the road transport system is a price that we have to pay for our mobility and our individual freedom. As a result of this, the long-term goal is to reach an optimal level of fatalities and serious injuries and strike a balance between costs and benefits to achieve this (Elvik, 2009).

The Vision Zero Approach

Vision Zero, as a road safety approach, does not deny the fact that human factors play a crucial role in the occurrence of accidents, but instead of defining this as the basic problem which needs to be solved, Vision Zero is based on the idea that human beings always will make conscious and subconscious mistakes. There is no perfect human (or machines for that part), there never will be, and we have to accept that accidents will occur. Instead of focusing on prevention of accidents, per se, Vision Zero rather focuses on severe injuries as the basic problem that needs to be solved. According to Vision Zero, the principal cause as to why people die and are seriously injured is that the energy to which people are exposed in a traffic accident is too excessive in relation to the energy that humans can withstand. This is not a new idea. These ideas are to a large degree based on the thoughts that Dr. William Haddon developed and elaborated upon already in the middle of the 1960s (Haddon, 1968, 1970). However, Haddon's utmost important contribution to

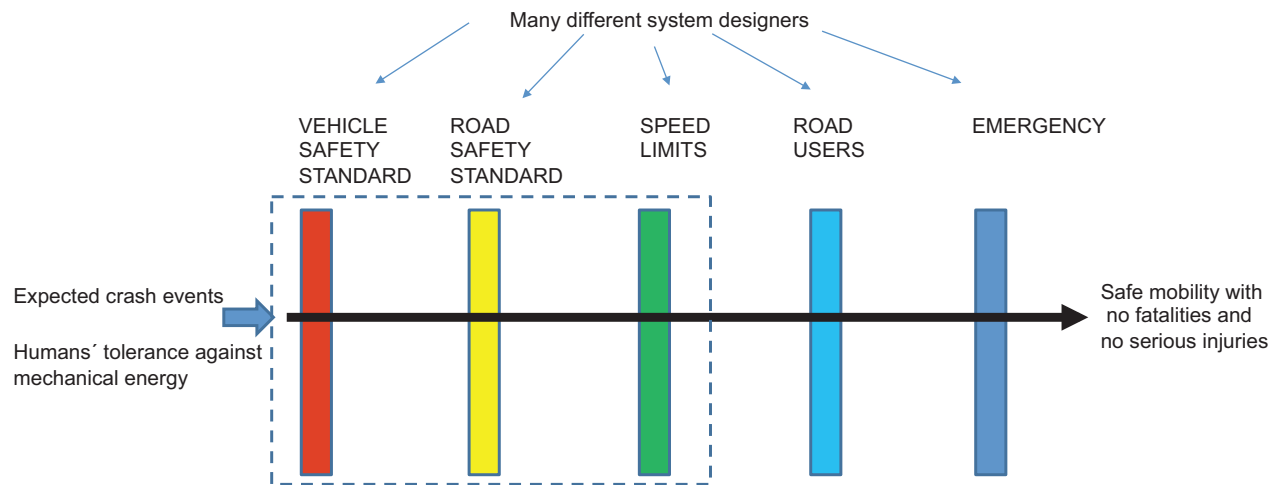


Figure 2 The Reason's cheese model applied on road traffic safety.

the road safety community has not been frequently adapted in the design of our road transport system except for in parts of the car design (e.g., crashworthiness) and individual protection equipment (e.g., helmets, seat belts, child restraint systems, etc.).

The basic strategy that underpins Vision Zero is therefore injury prevention rather than accident control. Vision Zero is a holistic policy that aims to create a safe system in the long run. The components, that is, roads, vehicles, and traffic environment and behavior, need to be integrated into a system that delivers safe mobility for all road users. All these aspects need to be designed as barriers and thereby control potential harmful energy to reach individual road users. We therefore need to add a prevention model to the famous injury causation model: precrash, crash, and postcrash. Based on Reason's "cheese model" (Reason, 1997), the road transport system and its different individual components could therefore be arranged into different protection layers (Fig. 2). According to this model, trauma care is the last protective layer just after individual behavior. The road environment is the first and most general and comprehensive protection layer that, to a large degree, defines what kind of accident types that could happen in what energy level. For example, if the roads were designed to eliminate head-on collisions (e.g., motorways, 2 + 1 roads) then the severity of factors, such as drink and driving, fatigue, distraction, among others, would also be eliminated. In a Vision Zero approach, the design of the road environment, both rural and urban, is the most important protection layer. Vehicles is the next protection layer and both the factors, how the vehicle can be used and its protection characteristics, play an important role. For example, cars, both top and dynamic speed performance, are important to support drivers to choose an appropriate speed. Nowadays, there are technologies such as intelligent speed adaptation systems that prevent drivers from speeding. Based on the design of the environment and the vehicle safety performance, safe speed limits could be set. If the road environment and the vehicles were well designed, from a safety perspective, a higher speed limit and mobility could be accepted. Poor safety standards will therefore instantly influence the mobility in the road system. In a Vision Zero approach the rural and the urban road environment, vehicle safety performance, and speed limits are therefore the most important protection layers and these need to be designed and operated according to safe system principles.

The road transport system is a complex system and there are several stakeholders at different levels in our society that influence the design and the operation of the system. Instead of putting the ultimate responsibility for safety on us as individual road users, Vision Zero puts the final responsibility for safety on those stakeholders who design and operate the system. The responsibility for safety is thus split between the road users and the system designers (i.e., infrastructure builders and administrators, the vehicle industry, the haulage sector, taxi companies, and all the organizations that use the road transport system professionally), on the basis of the principles that:

- the system designers have ultimate responsibility for the design, upkeep, and use of the road transport system and are thus responsible for the safety level of the entire system;
- as before, the road users are still responsible for showing consideration, judgment, and responsibility in traffic and for following the traffic regulations; and
- if the road users do not take their share of the responsibility (e.g., due to a lack of knowledge or competence) and personal injuries or other risky situations occur, the system designers must take further measures to prevent people from being killed or seriously injured.

In Vision Zero, the responsibility for safety is a chain that both begins and ends with the system designers.

Vision Zero is also based on the idea that individuals and society want and demand safety. The fact that people sometimes act as though they do not need safety rather seems to do with inability, ignorance, and a lack of social support than a lack of will. Not even suicide, in most cases, is a deliberate act based on the willingness to die. It is rather a desperate expression that the life is too difficult to tolerate and there are no other options to be seen. The long-term objective of Vision Zero is to create safe (e.g., no fatalities and no

serious injuries) mobility. Vision Zero as a public policy will therefore dramatically change our perception on mobility versus safety in our road transport system. Mobility becomes a function of safety rather than the other way around. This is nothing new; it is the way things work in other parts of our society and other transport systems. For example, if we discover safety problems in our aviation system that might threaten people's lives we stop the air traffic until we have fixed the problem. Luckily, we do not have to stop the road traffic in order to achieve safe mobility, since it, in most cases, is enough to reduce speeds.

Vision Zero—To Make Things Happen

Although we still need more road safety research, over the years a comprehensive knowledge base has been developed about traffic safety and what factors in the road transport system (e.g., road, vehicle, and behavioral factors) that need to be manipulated in order to reduce the number of fatalities and serious injuries. Our main challenge is therefore to ensure that all this knowledge is transformed into an action that eventually contributes to a safe road transport system. This moves us from knowledge to policy/innovations to implementation and result. There are various actors who individually and jointly act in the road transport sector, which in the end will affect the operation and the design of the road transport system. However, this implementation and innovation process is not well understood and there is a lack of understanding of the dynamic process that aims to formulate and implement road safety policy/innovations, as well as how road safety interventions are effectively disseminated. Therefore, we need more systematic studies of implementation and innovation processes. Such knowledge would contribute to how safety issues can be solved most effectively and appropriately and thus be of practical benefit to both the individual and the society (Belin et al., 2017).

The implementation of Vision Zero has been largely a trial-and-error process. The state (parliament, government, and authorities, particularly the Swedish Road Administration) played a central role, both directly and indirectly, in the implementation of Vision Zero. With regard to its role as an influencer of different stakeholders, the state must decide whether to intervene or whether to let the market and society take care of itself. Historically, traffic safety work has been focused on management by rules. In the work to realize Vision Zero, management by rules still exists, but occupies less space than in the past. New measures and new approaches were introduced, in which networking and management by objectives are fundamental approaches.

The work to make roads safer to achieve Vision Zero has been primarily characterized by four forms of governance, which have been shown to be effective and have driven the work forward:

- **Management by objectives and results:** In order to create long-term sustainability and a systematic way of working, the Swedish road safety efforts are based on management by objectives. The work includes setting objectives for various indicators judged to have the greatest impact on road safety, for example, adherence to the speed limit, safe roads, safe vehicles, and the use of helmets. Changes to the various indicators are measured, monitored, and discussed at annual results conferences.
- **Network governance:** Collaboration between various parties has been and is a key factor in the work toward achieving Vision Zero. This has led to different parties working together in the same direction through coordination and harmonization. Especially important is the collaboration between the vehicle industry and road authorities.
- **Benchmarking:** The market's demand is the mechanism that governs development and results. This way of working has had good results in terms of safer cars such as Euro new car assessment program.
- **Evidence-based work to improve road safety:** Implementing new road safety measures requires fundamental information about the problems that need to be solved. There are two sources of statistics and knowledge about road deaths in Sweden: Swedish Traffic Accident Data Acquisition (STRADA) and the Swedish Transport Administration's in-depth studies of fatal accidents. STRADA is an information system containing data about injuries (collected through hospitals) and accidents that have occurred in the road transport system. All these methods are examples of processes that are vital to make things happen.

Vision Zero—Outputs and Outcomes

In Fig. 3, we see how the number of fatalities per million inhabitants has developed since the adoption of Vision Zero. In 1999, the Swedish government launched an 11-point program for Vision Zero and based on this several interventions have been implemented: traffic calming in urban area, 2 + 1 roads, new speed limit system, and a large-scale traffic safety camera program, and safer vehicles among other things (Strandroth, 2015). The number of fatalities per 100,000 inhabitants has been reduced from 6.7 fatalities in year 2000 to 2.8 fatalities in 2010, a more than 50% reduction. Vision Zero is a policy and innovations and interventions based on this have contributed to the reduction of fatalities by more than 50% during the 10-year period. This is nothing short of an outstanding result, especially for a country that already was performing well in an international perspective. This is promising because it shows that continuous improvements of both strategies and interventions can help us to get better performance. However, in recent years there are some signs that Sweden has reached a plateau and during 2018 it was even an increased number of fatalities.

During 2009 and 2010 the government made major changes to the governmental structure of the transport sector. Among other things, the Swedish Road Administration and the Swedish Railway Administration were merged together into the Swedish Transport Administration. A new regulatory government body was also set up—the Swedish Transport Agency. However, the government did

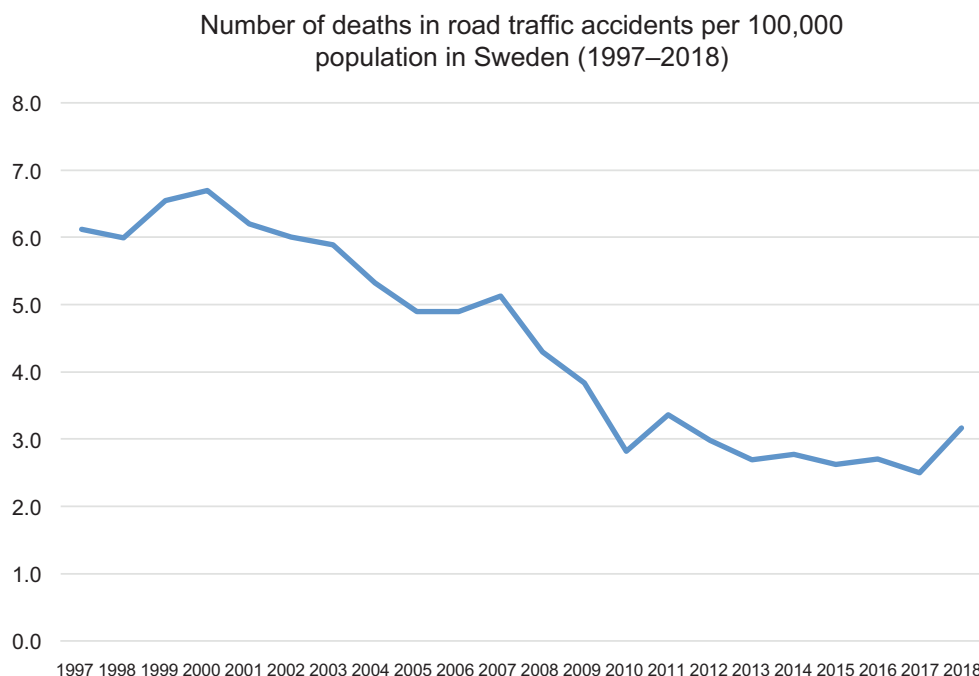


Figure 3 Number of deaths in road traffic accidents per 100,000 population in Sweden (1997–2018).

not decide which authority should be the leading agency and responsible for maintaining and developing Vision Zero. In 2016, the Swedish government decided to renew its commitment to Vision Zero and they also decided that the Swedish Transport Administration has the overall responsibility for managing and coordinating the road safety work in Sweden. Hopefully, after a few years of stagnation, through the government initiative to renew Vision Zero, Sweden will enter a new era in the work with traffic safety. Despite a continuously traffic and population growth, Sweden has a down-going trend of fatalities since late 1960s, which is impressive. However, there is still a long way to go until we have reached a safe road transport system where we have eliminated all fatalities and serious injuries and even then the remaining challenge will be to maintain a safe system.

References

- Belin, M.-Å., Tillgren, P., 2012. Vision Zero—how a policy innovation is dashed by interest conflicts, but may prevail in the end. *Scand. J. Public Adm.* 16 (3), 83–102.
- Belin, M.-Å., Andersson, R., Nilsen, P., 2017. Vision Zeros—from idea to implementation—a programme for implementation research within the transport sector. Publication 2017:032. The Swedish Transport Administration, Sweden.
- Elvik, R., 2009. Can injury prevention efforts go too far? Reflections on some possible implications of Vision Zero for road accident fatalities. *Accid. Anal. Prev.* 31 (3), 265–286.
- Evans, L., 2004. *Traffic Safety*. Science Serving Society, Bloomfield, MI.
- Friedland, M.L., Trebilcock, M.J., Roach, K., 1990. *Regulating Traffic Safety*. University of Toronto Press, Toronto.
- Haddon Jr., W., 1968. The changing approach to the epidemiology, prevention, and amelioration of trauma: the transition to approaches etiologically rather than descriptively based. *Am. J. Public Health* 58, 1431–1438.
- Haddon Jr., W., 1970. On the escape of tigers: an ecologic note. *Technol. Rev.* 72 (7), 44–52.
- Johansson, R., 2009. Vision Zero: implementing a policy for traffic safety. *Saf. Sci.* 47 (6), 826–831.
- Kristiansen, A.-C., Andersson, R., Belin, M.-Å., Nilsen, P., 2018. Swedish Vision Zero policies for safety—a comparative policy content analysis. *Saf. Sci.* 103, 260–269.
- Reason, J., 1997. *Managing the Risk of Organizational Accidents*. Ashgate Publishing Company, Farnham, Surrey.
- Strandroth, J., 2015. Validation of a method to evaluate future impact of road safety interventions, a comparison between fatal passenger car crashes in Sweden 2000 and 2010. *Accid. Anal. Prev.* 76, 133–140.
- Swedish Government and Committee on Transport and Communications, 1997. Bill 1996/97:137 Nollvisionen och det trafiksäkra samhället (Vision Zero and the traffic safety society) and Committee report 1997/98:TU4.
- Tingvall, C., Lie, A., Johansson, R., 2000. Traffic safety in planning: a multidimensional model for the zero vision. In: von Holst, H., Nygren, A., Andersson, A.E. (Eds.), *Transportation, Traffic Safety and Health—Man and Machine*. Springer, Berlin, Heidelberg, pp. 61–69.

Further Reading

- Belin, M.-Å., Johansson, R., Lindberg, J., Tingvall, C., 1997. The Vision Zero and its consequences. In: Hakkert, A.S. (Ed.), *Proceeding of the Fourth International Conference on Safety and the Environment in the 21st Century*, 23–27 November 1997, Tel Aviv, Israel.
- Belin, M.-Å., Tillgren, P., Vedung, E., Cameron, M., Tingvall, C., 2010. Speed cameras in Sweden and Victoria, Australia: a case study. *Accid. Anal. Prev.* 42 (6), 2165–2170.
- Belin, M.-Å., Tillgren, P., Vedung, E., 2012. Vision Zero: a road safety policy innovation. *Int. J. Inj. Contr. Saf. Promot.* 19 (2), 171–179.

- Belin, M.-Å., Höök, H., Tingvall, C., Hartzell, P., 2013. Advancing road safety management: the new standard ISO 39001. *Routes/Roads* 359, 30–35.
- Belin, M.-Å., Vedung, E., Meleckidzedeck, K., Tingvall, C., 2014. Carrots, sticks and sermons: state policy tools for influencing adoption and acceptance of new vehicle safety systems. In: Regan, M.A., Horberry, T., Stevens, A. (Eds.), *Driver Acceptance of New Technology: Theory, Measurement and Optimisation*. Ashgate, Farnham, pp. 241–252.
- Fahlquist, J.N., 2006. Responsibility ascriptions and Vision Zero. *Accid. Anal. Prev.* 38 (6), 1113–1118.
- Larsson, P., Dekker, S.W.A., Tingvall, C., 2010. The need for a systems theory approach to road safety. *Saf. Sci.* 48, 1167–1174.
- Rosencrantz, H., Edvardsson, K., Hansson, S.O., 2007. Vision Zero: is it irrational? *Transp. Res. Part A Policy Pract.* 41, 559–567.
- Swedish Transport Administration, 2018. Analysis of road safety trends 2017: management by objectives for road safety work towards the 2020 interim targets.
- Tingvall, C., 1997. The Zero Vision: a road transport system free from serious health losses. In: von Holst, H., Nygren, A., Thord, R. (Eds.), *Transportation Traffic Safety and Health*. Springer, Berlin, pp. 37–57.
- Tingvall, C., Haworth, N., 2000. Vision Zero: an ethical approach to safety and mobility. In: Jraiw, K. (Ed.), *Proceeding of the Road Safety & Traffic Enforcement: Beyond 2000*, 6–7 September 1999, Melbourne, Australia.

Tire Safety

Saied Taheri, Center for Tire Research (CenTiRe), Mechanical Engineering Department, Virginia Tech, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	681
Tire Parameters Affecting Force and Moment Generation Mechanism	682
Studded, Runflat and Non-Pneumatic Tires	683
Studded Tires	683
Run-Flat Tires	683
Non-Pneumatic Tires	684
References	684

Introduction

Although development of new tire materials, improvement of structural integrity, and tread design over the past decade have helped with improving tire and hence vehicle safety, the number of accidents per million vehicle miles traveled, has not changed drastically. National Highway Transportation Safety Administration's (NHTSA) Crash Causation Survey found that there was an issue with a tire before the crash occurred in 1 of 11 crashes (9%). Issues included tread separations, blowouts, bald tires, and under-inflation (Choi, 2012). They found that:

- Of the tires that were underinflated by more than 25% of the recommended pressure, approximately 10% were in vehicles that experienced tire problems in the pre-crash phase. In contrast, among the correctly inflated tires, a much smaller percentage (3.4%) belongs to vehicles that experienced tire problems. Thus, under-inflation is not the only cause of tire problems; however, when tires are underinflated by 25% or more, tires are three times as likely to be cited as critical events in the pre-crash phase.
- With at least one or more tires with lower tread depths (between 0 and 4/32"), vehicles experienced tire problems during crash occurrence significantly more than chance. Of tires with tread depth in the range 0 to 2/32", about 26% were in vehicles that experienced tire problems in the pre-crash phase while only 8% of tires with tread depth in the range 3/32–4/32" were in such vehicles.
- The percentage of vehicles experiencing tire problems is significantly higher among vehicles that rolled over as compared to vehicles that did not roll over for all vehicle body types: passenger cars, pickups, SUVs, and vans. Of all SUVs experiencing tire problems in the pre-crash phase, 45% rolled over. For the other body types (passenger cars, pickups, and vans), fewer than 25% of the vehicles experiencing tire problems rolled over. Thus, tire problems experienced in the pre-crash phase were more likely to result in a rollover in SUVs than in other vehicle types.
- A significantly higher percentage (11.2%) of vehicles were observed to experience tire problems when one or more roadway-related factors (e.g., wet road, road under water, slick surface) were present in the pre-crash phase as compared to when no roadway-related factors were cited (3.9%). Thus, the vehicles running under adverse roadway conditions may become more vulnerable to tire problems.

Aside from durability issues related to the findings in (Choi, 2012), for all vehicles leading to a crash, lack of tire force and moment generation capability, for the given condition, was the cause of all accidents. For example, pressure difference between the tires located in front of the car, as compared to those in the rear, could initiate and oversteering condition, which leads to instability of the vehicle and potential accidents.

In the current chapter, tire safety is viewed and discussed as force and moment generation capability of the tire and its effects on vehicle stability for studded, runflat (extended mobility), and non-pneumatic tires. Interested readers are directed to [Gent and Walter \(2006\)](#), Chapter 15, where, tire safety, durability, and failure have been discussed and analyzed in details. The analysis provided in [Gent and Walter \(2006\)](#) is mostly concerned with tire failure and its effects on safety. As stated by the authors, performing a failure analysis on a tire is a scientific process that begins with an understanding of the product, including its materials, mechanics, and operational characteristics. Therefore, elements of tire safety, basic mechanics, and factors pertinent to tire durability, including common tire failure modes and analysis are discussed.

Since tires are the only means of contact between the vehicle and the road, they are considered as one of the safety critical components of the vehicle. Aside from tire durability/failure, which can be due to various internal or external issues, forces and moments generated by the tires play the most important role in vehicle stability and safety. Tread separation and blowout are direct durability issues, under-inflation affects both durability and force and moment generation capability of the tire, and bald, studded, and runflat or non-pneumatic tires affect force and moment generation capability, and hence safety of the vehicle.

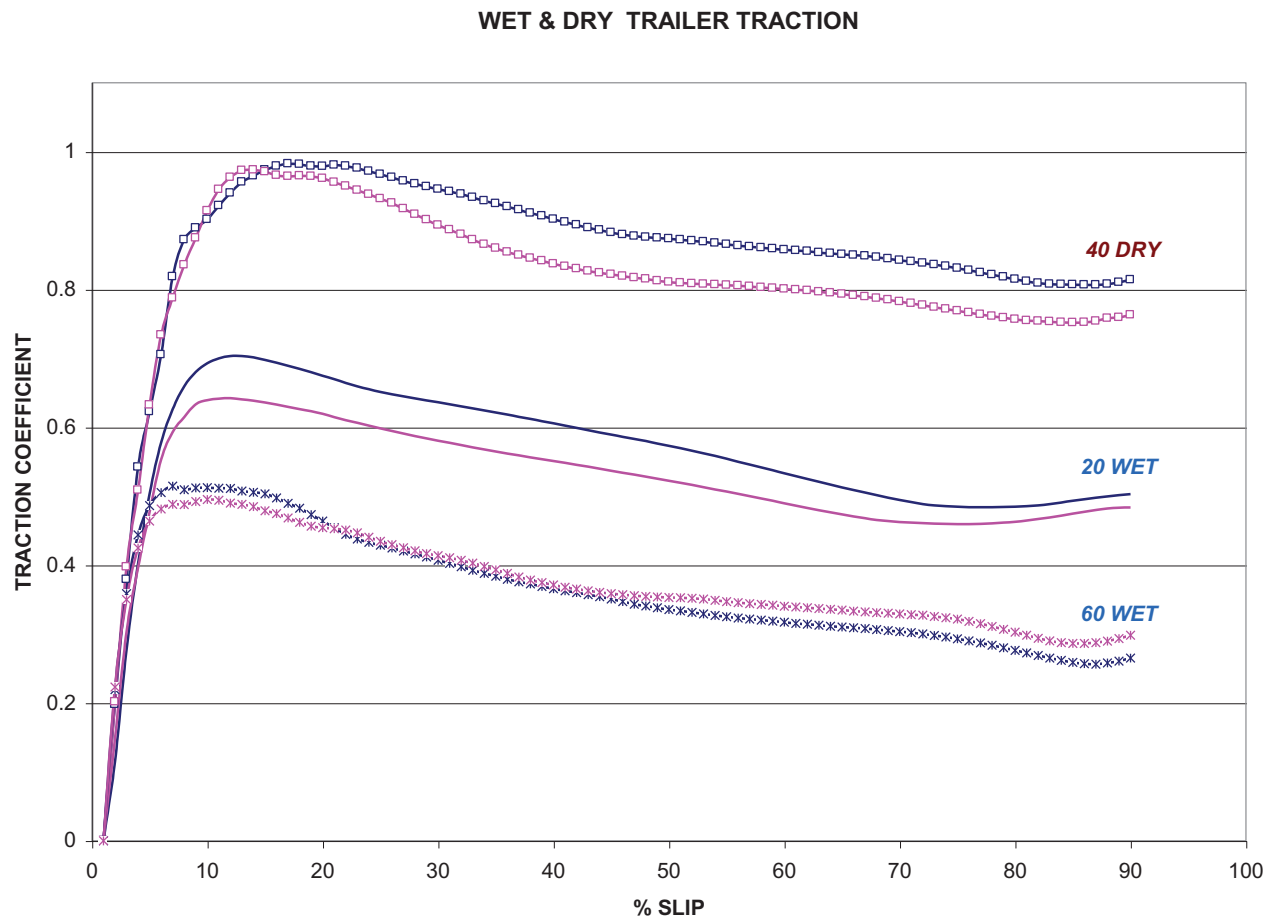


Figure 1 Longitudinal force/friction vs. slip.

Tire Parameters Affecting Force and Moment Generation Mechanism

During vehicle motion, aside from the ride characteristics of the tire, which are mostly determined by vertical force (dynamic load), tire longitudinal, and lateral forces are the only means for the vehicle to accelerate, decelerate, and negotiate curves. This determines vehicle acceleration potential, stopping distance, and turning accuracy. If these forces are not generated properly, the driver's chance of controlling the vehicle is reduced drastically. For a properly designed and selected tire for a given vehicle platform, tire inflation pressure and tread depth are among the most important factors effecting force and moment generation and vehicle safety.

As can be seen from **Fig. 1**, vehicle speed and road surface have paramount effects on tire longitudinal force generation, which is essential for vehicle stopping and accelerating. As vehicle speed increases from 40 to 60 km/h and road surfaces changes from dry to wet, the coefficient of friction drops nearly by 50%. Therefore, road accidents are more likely to occur on wet and slippery roads than dry road surfaces. This mainly can be defined as skid resistance resulting in friction force that is required for stability and control of the vehicle travelling on the road. Federal Highway Administration estimates that there are, on average, over 5,748,000 vehicle crashes each year. Approximately 22% of these crashes—nearly 1,259,000—are weather-related. The vast majority of the weather-related crashes happen on wet roads—either wet pavement or during rainfall—73% on wet pavement and 46% during rainfall. According to the FHWA report, over 900,000 crashes occur on wet pavement and nearly 600,000 car crashes occur when it is raining. Over 350,000 persons are injured and nearly 4500 persons are killed in the wet pavement crashes. Over 225,000 persons are injured and nearly 2800 are killed in the car accidents which occur during rain (Karzan and Noorance, 2017). Although **Fig. 1** shows that friction reduction between the tire and road on wet surfaces could be viewed as the main cause of accidents, however, this depends on the tire tread design, surface texture, depth of water, and human error.

The data shown in **Fig. 1**, is collected under proper tire inflation pressure. If the pressure is reduced below 25% of recommended value or drastically increased, the force and moment generation capability of the tire changes, leading to vehicle instability.

One of the most influential tire safety parameters is tire inflation pressure. The air inside the tire determines its performance as well as safe operations. In addition to driving with low pressure tires, which can be dangerous and even fatal, tire pressure changes the tire stiffness which in turn effects force and moment generation. When a tire is underinflated, the sidewalls will flex out farther than usual. This will change the effective rolling radius of the tire and can generate excessive heat which may cause tire failure. Additionally, studies have demonstrated the relationship between tire pressure and slip ratio. Slip ratio is defined as the relative

velocity between the tire and the road and is one of the main variables in longitudinal tire force generation. Antilock braking system (ABS), is a control system designed to prevent wheels from losing tractive contact with the road surface by modifying driver inputs during the braking process to prevent wheel lock up (Li et al., 2015). There are many different variations of control methods in ABS. Several involve the comparison of current wheel speed against some constant threshold value. This value is targeted to keep the slip ratio close to some optimal value where longitudinal braking force is maximized. It has been shown that decreased tire pressure increases the peak longitudinal force, but at a lower optimal slip ratio (Besselink et al., 2010). Additionally, variations in tire pressure can change the effective wheel radius by as much as two centimeters over a normal range of operating conditions (Schmeitz et al., 2005). A study by the National Transportation Safety Board (https://www.nts.gov/safety/safety-alerts/Documents/SA_044.pdf) found that underinflated tires lead to over deflection, that is, a change in the tire's radius that is beyond the intended operation limits, which lead to tread separation and ultimately in tire failure that resulted in fatalities. Further, it is also known that underinflated tires experience severe flexing which results in elevated temperatures particularly at high speeds and during hot weather which also results in tread separation and then sudden blowout (<https://www.nts.gov/safety/safety-studies/Pages/SIR1502.aspx>; <https://www.aacar.com/library/tirefail.htm>).

Looking at the vehicle stability from understeering viewpoint (understeering is a stability term used in vehicle dynamics, which refers to the difference between the turning potential of the vehicle front end as compared to the rear end of the vehicle), having less pressure in the rear tires can initiate an oversteering effect, which will initiate vehicle instability during turning at higher speeds. Therefore, tire pressure has multiple effects on vehicle performance and stability. Overinflated tires will result in a smaller contact patch which may adversely affect the force and moment generation of the tires leading to potential stability issues.

In order to eliminate some of these safety related issues, non-pneumatic and runflat or extended mobility tires were introduced. Although currently there are not any non-pneumatic (Tweel: <https://michelinmedia.com/tweel/>, for example) tires in use under passenger cars, they are being developed as a viable option to replace the pneumatic tire.

Studded, Runflat and Non-Pneumatic Tires

Studded Tires

Regarding studded tires vs. non-studded, there are several issues which have resulted in several states banning the use of these tires. A study conducted by Washington State Transportation Commission (Scheibe, 2002) concluded that:

1. Studded tires produce their best traction on snow or ice near the freezing mark and lose proportionately more of their tractive ability at lower temperatures than do stud-less or all-season tires.
2. The traction of studded tires is slightly superior to stud-less tires only under an ever-narrowing set of circumstances. With less aggressive (lightweight) studs being mandated, and with the advent of the new "stud-less" tire, such as the Blizzak, since the early 1990s, the traction benefit for studded tires is primarily evident on clear ice near the freezing mark, a condition whose occurrence is limited. For the majority of test results reviewed for snow, and for ice at lower temperatures, studded tires performed as well as or worse than the Blizzak tire. For those conditions in which studded tires provided better traction than stud-less tires, the increment usually was small.
3. Traction performance can be characterized in many ways, including braking, acceleration, cornering, controllability, and grade climbing. Though all factors are important, the single best indicator of tire performance is braking distance and deceleration. Studded tires reduce the difference in friction factor between optimum-slip and locked-wheel braking in comparison to non-studded tires. This may reduce the risk of drivers misjudging the necessary braking distance and may improve the braking potential for anti-lock brakes. In one set of stopping distance tests in Alaska, studded, stud-less, and all-season tires performed nearly equally on snow, when averaged across several vehicles. On ice, stopping distances for studded tires were 15% shorter than for Blizzaks, which in turn were 8% shorter than for all-season tires.
4. On bare pavement, studded tires tend to have poorer traction performance than other tire types. This is especially true for concrete.

Although all of the conclusions drawn from this study are related to longitudinal performance of the tire/vehicle, the lateral performance on dry surfaces can be degraded due to the nature of contact and the friction coefficient associated with material variations and less apparent area of rubber contact with the road. Fig. 2 shows the comparison between the longitudinal force of studded and non-studded tires (Ivanov et al., 2017).

Run-Flat Tires

A run-flat tire is a relatively new concept that has begun to integrate itself into modern vehicles. Rather than a standard tire, a run-flat tire has a polyurethane ring that allows the vehicle to remain driving after the tire loses all its pressure. The polyurethane ring is offered as a security blanket to drivers in case they run into issues with the tires. Run-flat tires have two basic designs: one with a sturdy sidewall and one with a stiff inner ring. With a standard flat tire, a driver must pull over immediately. This is because the weight of the vehicle on the rims of the wheel can cause serious, costly damage and make driving conditions unsafe. Drivers will also experience unpredictable steering. Run-flat tires are designed to keep drivers on the road safely for longer periods of time in the event of a blowout. The tires rely on strong reinforced sidewalls to support the weight of the vehicle if the tire is punctured or loses air pressure. The sides of the tire are thick enough to carry the vehicle for miles without damaging the wheels or suspension. Another type of run-flats has a tough and rigid ring inside of the tire, hugging the wheel. When these tires lose air pressure, the ring spins on

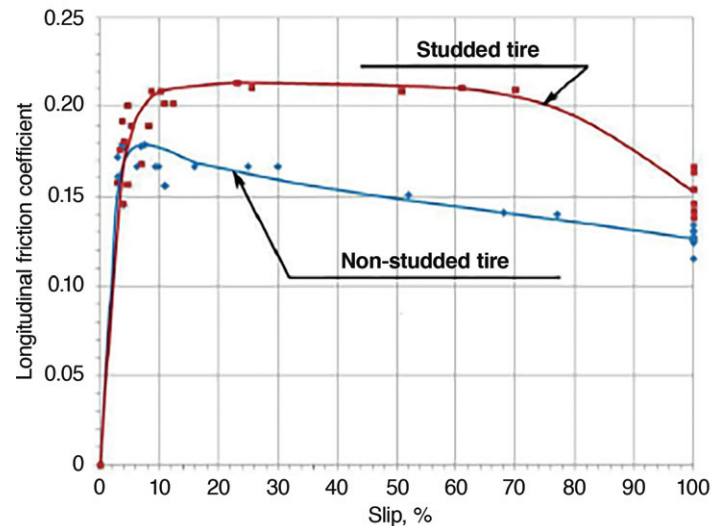


Figure 2 Longitudinal friction coefficient for studded and non-studded tires (Ivanov et al., 2017).

the road rather than the wheel rim. The miles of allowed driving distance let drivers find a safe spot to pull over to the side of the road and change the flat tire. However, many cars that are equipped with run-flat tires do not have spare tires, which means that a mechanic with a new tire must be called to the location.

Though runflat tires provide a great advantage over a completely flat or significantly underinflated tire, their characteristics do vary significantly from an inflated tire in normal conditions. For example, a runflat tire compared to a conventional tire each at a vertical load of 9 kN demonstrates a 45%–50% greater vertical stiffness. The vertical load is more comparable for loads under 3 kN. This will lead to a rougher ride as compared to regular tires. Additionally, runflat tires will exhibit degraded handling performance at higher speeds and could be a safety concern if transient steering inputs are commanded by the driver.

Non-Pneumatic Tires

In addition to runflat tires, which behave very similar to regular tires while fully inflated, the non-pneumatic tires have been studied for the past two decades. Tweel, a Michelin Tire Company product, was introduced as a concept in 2005 and was offered as a product for use on loaders in 2012. Once offered for passenger cars, these tires have the advantage of not requiring to be inflated and hence tires will never be flat or fail due to puncture. Researchers have been trying as long back as 1920s to build a non-pneumatic tire (NPT) that has sufficient resilience (Cozatt, 1924). With advancement in material science and manufacturing technologies, researchers were able to create NPTs having sufficient compliance (Rhyne and Steven, 2005). Modern NPT designs integrate wheel and tire into a single component. The NPT consists of a rigid hub, flexible spokes, shear band and tread which is made up of rubber.

References

- Besselink, I.J.M., Schmeitz, A.J.C., Pacejka, H.B., 2010. An improved magic formula/swift tyre model that can handle inflation pressure changes. *Veh. Syst. Dyn.* 48, 337–352.
- Choi, E.-H., 2012. Tire-Related Factors in the Pre-Crash Phase, Report No. DOT HS 811 617. National Highway Traffic Safety Administration, Washington, DC.
- Cozatt, C.P., 1924. Spring wheel. US patent-2502,908.
- Gent, A.N., Walter, J.D. (Eds.), (2006). The pneumatic tire. 'Introduction to Tire Safety, Durability and Failure Analysis. J.D. Gardner and B.J. Queiser, pp. 613–640 (Chapter 15).
- Ivanov, A.M., Gaevskiy, V.V., Kristal'niy, S.R., Popov, N.V., Shadrin, Sergey, Fomichev, V.A., 2017. Adhesion properties of studded tires study. *J. Ind. Pollut. Control* 33., 988–993.
- Karzan, I., Noorance, A., 2017. Traffic accidents analysis on dry and wet road bends surface in Greater Manchester-UK, 2. 10.24017/science, 2017, 3, 51.
- Li, G., Wang, T., Zhang, R., Gu, F., Shen, J., 2015. An Improved Optimal Slip Ratio Prediction considering Tyre Inflation Pressure Changes. [Online]. Available: <https://www.hindawi.com/journals/jcse/2015/512024/>.
- Rhyne, T.B., Steven, S.M., 2005. Development of a nonpneumatic wheel. *Tire Sci. Technol.* 34, 150–169.
- Scheibe, R.R., 2002. An Overview of Studded And Studless Tire Traction And Safety, WA-RD 551.1. Washington State Transportation Commission, 10/2002.
- Schmeitz, A.J. C., Besselink, I., de Hoogh, J., Nijmeijer, H., 2005. Extending the Magic Formula and SWIFT tyre models for inflation pressure changes. Available: https://www.researchgate.net/publication/252060616_Extending_the_Magic_Formula_and_SWIFT_tyre_models_for_inflation_pressure_changes.

Town Gates: Section on Transport Safety and Security

Charles Tijus, University Paris 8, Saint-Denis, France

© 2021 Elsevier Ltd. All rights reserved.

What are Town Gates?	685
The Logic of Things Inside Town Gates	686
Who	686
Where	686
When	686
Why	687
How	687
The Logic of Action Constraints Inside Town Gates	687
The Overall Logic of Action Constraints in Town	687
Externalization and Internalization of Transportation Information	688
How to Favor Correct Decision-Making in Towns: Recommendations	689
Conclusion: Future Uses of Town Gates	690
Biography	691
Relevant Websites	691
References	691
Further Reading	691

What are Town Gates?

Historically, town gates were used for protection of cities, towns, and villages against invaders. They were also used as a location for collecting tolls related to people or products. With the enlargement of cities, old town walls and gates can nowadays often be found toward the center of large cities. For transportation, modern town gates are of importance to prevent injury or accidents by informing people with the use of signs or other information at the entry to the town so they behave safely for themselves and for others. The information aims to engage mainly drivers that they are entering or leaving a place where other people live; thus a place with specific risks of conflicts and dangers and rules for avoiding these dangers. The importance of information from town and village gates increases with the degree of the transition from outside to inside the town or village. If a high-speed road is going through a village, seeing, paying attention to, making smart decisions, and reducing speed according to mandatory rules of traffic regulations are important in order to avoid accidents and reduce their severity if some accidents still happen, with a goal of especially avoiding fatal injuries.

To behave properly, people must know when they are entering or leaving a town or a village. The context, such as denser housing, might be of help but it does not guarantee that all drivers note that they have entered an urban area. Legal and safe behavior can be obtained to a great degree by using and getting compliance to traffic signs and road markings. When entering and leaving towns, there are specific international road signs indicating the perimeter of the compact area, so called town gates, although there may be no gates at all, just traffic signs for entrance and exit (Fig. 1). Such signs are sometimes labeled "Traffic Calming Signs" if they have the aim of slowing the flow of traffic entering the built-up area. Some communities also use non-conforming gates, which may resemble medieval gates or modern "portals" with structures spanning across the roadway at a height allowing vehicles to travel under them.

As for any road and traffic sign, Traffic Calming Signs are external representations using symbols and or text that should be matching all local and national regulations that drivers are supposed to know, to remember and to apply. Thus compliance to in-town and village traffic regulations is not solely a matter of perceiving the sign that indicate gates, but a matter of knowledge, and of interpretation based on a less or more precise cognitive model of situation about being inside this specific town or village. It follows that foreign visitors might be less aware of local rules and comply only according to their own habits. For instance, a foreign driver from a country where school starts at 9:00 a.m. might pay less attention to warning signs for schools when it is 5:00 a.m. in the morning although in the actual country school may start at 5:30 a.m. Thus, departments responsible for city infrastructure should be aware of how to inform visitors, and people just passing through, about road safety and security and make sure people understand the meaning of local road signs.

To communicate with drivers and obtain good traffic safety in towns and villages, motorists must be made to understand what is going on inside the town gates. Such an understanding is based on the "where," "who and what," "when," "why," and "how" dimensions of events and actions made by people who live there. Such knowledge activation or provision of new information is useful and can be provided through tourist's brochures, new traffic regulations, installation of road regulatory signs and road markings, and of the improvement of roads and streets by using traffic signals, roundabouts, and geometric features such as humps and pedestrian pathways (de Lavalette et al., 2009).



Figure 1 Traffic signs indicating (A) for any kind of Town gates with entry on top and exit on bottom panel, (B) for a village named St Vincent de Tyrosse with entry on top and exit on bottom panel, and (C) for entrance of a village named Elkington with a 30 km/h speed limit.

By providing guidance on logical and clear transportation information and regulations before, at and behind town gates, we expect this article to be of help for the design of regulatory procedures, rules and information signage and finally of help to improve safety for drivers, bicyclists, and pedestrians in general and in particular for specific users such as tourists or children, as well as people with disabilities including accessibility for the elderly, for example, when crossing streets. To do so, we first provide the logic of activities and other things inside town gates, the logic of action and behavior that are based on the nature of things inside towns, and how the regulatory authorities of transportation inside towns are delivering the information about obligations, prohibitions and permission by externalizing them in signs that everyone can understand.

The Logic of Things Inside Town Gates

When traveling from outside town gates to inside a town, there is a number of informational items dedicated to transportation through the town, including announcement of the town entrance (Fig. 1). There are roads and streets, some with and others without sidewalks, some that bypass the city, and some of those may be limited-access highways, others that go through the town and may be intentionally narrowed in width, with speed humps, a number of roundabouts or traffic signals, or both. In addition, one will find much information for drivers of passenger cars and of trucks, mostly on traffic signs. These use physical entities to encourage or force the desired driving behavior or remind road users of the driving rules. But most important are what is not displayed but is supposed to be known. For instance, inside town gates, in many countries, the speed limit is set at 50 km/h, unless there is another more restricted limit (such as shown in Fig. 1C).

In addition to purveying knowledge attached to town gates, there is rationality to how people should behave inside the gates. Transport safety and security depends on the understanding of this logic of functioning and of utilization of the city while traveling.

Who

To obtain safety in transportation, motorists should be able to easily imagine the kind of people who live in a place: unpredictable school children, slow walking aged people, joggers that avoid stopping, bicyclists are some groups that can be expected. To do so, drivers need to pay attention to the traffic signs that provide information about people who live or spend time at a location.

Where

To obtain safety in transportation, motorists should have a good representation of the structure of places inside the town. The city can be with or without an increasingly populated city center in which going downtown will provide places with more and more people, markets, tourist destinations, shops, and traffic congestion.

Visitors should know where to find useful locations (e.g., restaurants, parking lots, toilets, etc.) as well as specific locations such as hospitals, schools, and train stations and so on. Along streets, one should be able to find indication of what exists in the town (i.e., a hospital sign) but also in which direction to go to find WHERE is a special place (i.e., a sign with an arrow indicating where to go to find a hospital and distance to it). For instance, Fig. 2 indicates buildings that are located where the road signs are placed, here or further away.

When

Transportation information is mostly permanent information. For instance, information about how to behave (e.g., *do not bypass; speed limit, etc.*) is delivered with road signs that indicate that at any time drivers may behave that way. There is some information about WHEN. And if so, this is information about which days in the month (*No Parking from the 1st to the 15th of the month*), which day in the week (*Friday to Saturday*), at what times (*11 p.m. – 5 a.m.*), and sometimes how long (*2 hours limit*).

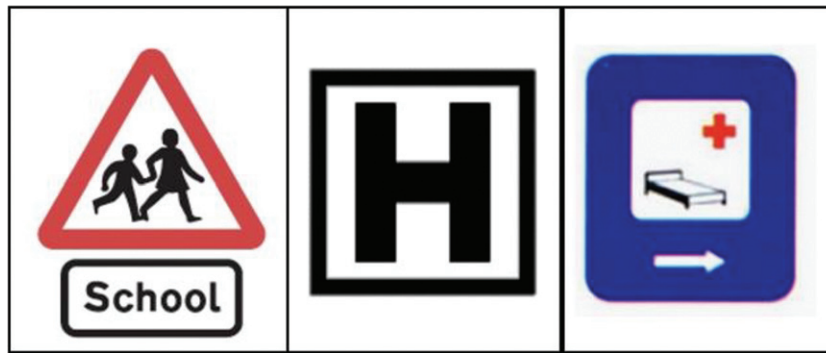


Figure 2 Road signs indicating the kind of building that exists there (left panel: a school, middle panel: a hospital) or which direction to go to find a kind of building (right panel: a hospital).

Why

WHY is about the understanding of the meanings of things inside town gates. For instance, Fig. 2 indicates buildings inside the town: a school here, a hospital here, and a hospital further away. The related action, —the reason why these road signs are placed there—, is

- for the school sign (Fig. 2, left panel) about increasing drivers' attention and precaution because of the danger (red triangle) that corresponds to a place frequented by children going to and from school, an increasing risk according to the school schedule,
- for the hospital here (Fig. 2, middle panel), the location of a hospital is indicated but it should be known that this is for non-emergency services, or a clinic. In case of an emergency, going there by mistake can be fatal in case of a serious injury,
- for the Hospital further away (Fig. 2, right panel), this is indicating a hospital providing emergency services (indicated by the red cross) and it is mandatory (blue background of the sign) to go in that direction to find emergency medical help.

It appears that transport safety depends not only on the pertinent and efficient implementation of regulatory traffic signs, but on the knowledge of road signs by people using the streets in the city, on their interpretation of what to do according to their understanding and misunderstanding of the whole transportation system implemented in the city and finally implications of this on their decision making.

How

Safe and secure transportation within a city depends on HOW people behave while interacting with others. This is influenced by their knowledge, expertise and familiarity with things inside the town gates, and the implementation of support of displayed information. Let's look at one of the worst scenarios based on incorrect decision-making: A child is running to school because s/he is late. A driver who does not respect the danger sign indicating "place with children going to or coming from school" (Fig. 2, left panel) runs into the child and takes her/him to the hospital located here (as indicated in Fig. 2, middle panel) instead of taking her/him to the hospital for emergencies (Fig. 2, right panel); a hospital which he can then have difficulty finding because he does not know which direction to take; not being reminded where the direction indicator sign was, although it was seen.

Note that in such cases, expertise can help; by heuristics: to find a sign that we have seen, take the same path that allowed us to see this sign earlier; but better to call an ambulance by using a mobile phone or from the nearest emergency telephone.

In summary, HOW people behave depends on their decision-making according to their understanding of WHY things in town function in a certain way according to WHO, WHERE and WHEN. These dimensions are those that are constraining actions: what one can do, and what one cannot do.

The Logic of Action Constraints Inside Town Gates

In a majority of countries, most people live in towns and villages. Contrary to motorways for which much of information and mandatory rules are focused on automobile regulations, inside town gates, much is about interactions between people in cars and people outside cars. Thus, the logic of expected behavior relates mainly to people in or out of cars. For people inside cars, there will be information about how to drive; and also, where to park, locations of services, tourist places to visit, and so on. What should be known is that this information inside town gates increases both in informational and procedural levels of complexity. Knowing this improves safe and secure transportation by anticipating next steps of interaction with the inside town gate system of delivering services and regulation information.

The Overall Logic of Action Constraints in Town

There has to be a logic to transportation information. Knowing this logic improves safety. The logic helps understand where signs should be placed in towns. The purpose of signage is to indicate, in iconic or linguistic form, places, directions, actions, or

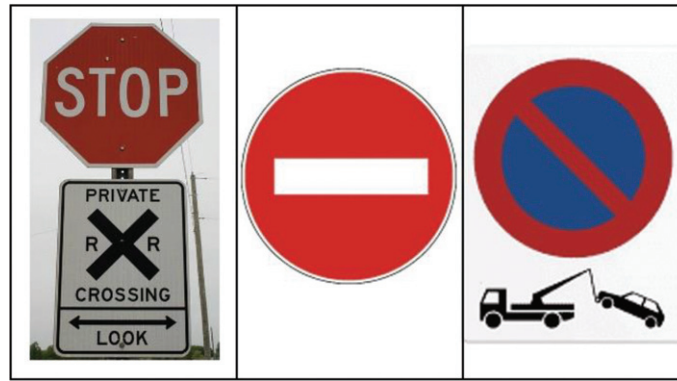


Figure 3 Road signs indicating action prohibition for different kinds of objects, respectively in left panel: not crossing without stopping at an at-grade railroad crossing, middle panel: not entering a street in that direction, right panel: do not park.

prohibitions of actions in the city. In fact, there are no other uses of signage than indications of objects (O), properties of objects (P of O), actions (A on O to change OP), prohibitions of action (Don't do A on O), or cessation of action (now you can do A on O) on objects or categories of objects. Following are the successive levels of difficulty:

1. the indication of an object ("this is O") or of a category of objects ("this is a kind of O") such as a panel on which is written "North Station" or "station," or as another example "pedestrian crossing,"
2. the indication of object properties ("O has this property P") such as the information according to which "this station is the bus station" or that "this pedestrian crossing is taken by children,"
3. the indication of action to be undertaken on an object property ("Action A allows you to transform this property of O"), such as the sign informing that going in that direction will reduce the distance between where you are and the bus station, or indicating that slowing down the speed of the car will increase the level of safety at the pedestrian crossing,
4. the indication of procedural sub-goals ("How to use O"). For instance, such as how to use the train at-grade crossing: "Always approach the at-grade crossing carefully and be prepared to stop. You must obey the traffic signals. When the amber or red lights show, you must stop behind the white line across the road. If you have already crossed the white line when the amber light comes on, keep going,"
5. the indication of prohibition of actions on an object ("Action A on O to change P is prohibited), such as signs indicating not crossing without stopping at an at-grade crossing, not entering a street in that direction, do not park (three prohibition corresponding respectively to the left, middle and right panels of Fig. 3), and
6. the indication of cessation of prohibition of actions concerning the ownership of an object ("Action A on O to change P is no longer prohibited), such as signs indicating that there is no more parking prohibition or indicating that there is no more prohibition of any kind.

Note that any given level implies the preceding. For instance "now you can park (f)," implies that "before it was not possible to park – e" (and maybe it will not be possible again to park later), that there are "subgoals to park – d" (entry, parking ticket, find a place, keep parking ticket with you, etc.), implying in turn "you can park your car on one of the free parking place, reducing the number of free parking places – c" implying the information about how much free parking places – b", implying the indication of this is the parking- a".

Finally, note that in addition to prohibition of actions and of cessation of prohibition of actions, you might get the reverse such as indications, permissions and obligations of action.

Externalization and Internalization of Transportation Information

Road signs in town are externalization of rules in the real world for safety and security. The transportation regulatory authorities expect that when people in town (drivers, passengers, pedestrians, etc.) see the content of a sign in its external environment, by processing this content (seeing, categorizing, reasoning, and interpreting), they will internalize its meaning and will take the correct decision. Thus, when a driver sees a pedestrian crossing sign, s/he will pay attention, reduce speed while the pedestrian will respect the corresponding rules by using the pedestrian crossing and doing so in a safe way. Driver-pedestrian interactions are to be regulated that way. However, signage is not solely indication of action to do or not to do, signage also has semantics. It is symbolic in nature so that the meaning of a road sign, or of a pictogram is not unequivocal. It therefore requires an interpretation largely based on the context, and here the context of the town. For this reason, in line with the work on ergonomics on the understanding of technical notices, work on semiotics and applied linguistics, the semantic analysis of signage is essential.

An action that is to undertake or not on an object because this object belongs to a specific category of objects according to this action. For a safe and secure transportation in town, it is of importance that people get the proper meaning of signage in town (Bazire and Tijus, 2009). For instance, most people are often mistaken what a road sign means because it is not explicit but implicit. In the



Figure 4 The possible misinterpretation of signs in town, when the sign is displayed on the surface of the street, when trucks are passenger cars, when cigars and pipes are cigarettes, when don't kiss means that "the departure of the train should not be delayed."

left panel of Fig. 4, a road sign indicating a speed limit (50 km/h) is directly drawn on the surface of the street. A driver might say that this is indicating that he has to drive his car at 50 km/h, which is not true. The proper meaning is that vehicles must keep a speed below 50 km/h at all times, and, most importantly, because we are inside town gates, for any street not posted, the speed limit may be 40 km/h in this country. Thus, the next street will have the same speed limit although there could be no sign displayed to remind this rule.

In towns, there are often narrow streets, and there is sometimes a sign about prohibition to overtake other vehicles. In countries following the Vienna Convention of signage, this is shown by displaying a red pictogram of a passenger car to indicate that it cannot overtake another passenger car or a red pictogram of a truck to indicate that trucks cannot overtake cars (Fig. 4, middle panel). First, these road signs may be used because roads are narrow or there are other reasons we do not want drivers to pass other vehicles. Second, the pictogram of a passenger car does not apply solely to passenger cars but also to trucks. And the truck that cannot overtake a passenger car also cannot overtake another truck. All drivers should know that the pictogram of a passenger car designates passenger cars as well as trucks while the pictogram of a truck designates only trucks but not passenger cars. However, drivers from for example the United States may not understand this since they are not used to Vienna Convention signs.

Outside of town centers, for example in a forest, a sign can be showing the pictogram of a cigarette with a slash across it (indicating a ban, as in the right panel of Fig. 4). The meaning of the sign is that forest fires should not be started by smokers. When the same signage is used in a town, its meaning is that people should have their clean air unpolluted. More important for understanding is that although the pictogram on the sign is a cigarette, in both cases, smoking a pipe or cigar is also not allowed. Misunderstanding could be more pronounced when the pictogram is used in a metaphorical way. For instance, the prohibition sign of "kisses" (the drawing of a man and a woman kissing, all crossed out by a slash), which was found for a while on the platform of a train station in UK could be mistaken as meaning that a man and a woman are not allowed to be kissing each other, whereas two women or two men could. The meaning was that, "The departure of a train should not be delayed by people taking long good-byes," and among all the possible situations that could delay the train departure, the administration of regulatory signs chose this example.

Note that for blind or deaf pedestrians, providing specific information, for example at pedestrian crossings, are often provided: marking on the ground sensible for the blind man's cane with synthetic voice; or flashing lights to compensate for loss of hearing (Tournier et al., 2016).

How to Favor Correct Decision-Making in Towns: Recommendations

To make people properly follow rules that arise from entering a city, the regulator must have two principles of transport ergonomics in mind. The first principle is based on a general distinction between system and user-centered design (Norman, 2005). The second principle is the distinction between metaphorical display categories and specific target categories (Tijus et al., 2007).

The first principle is that legal categories which define what concerns transportation and road safety may not correspond to the natural categories of people. As an example, we can consider, as in Fig. 4-middle panel, the truck that is a passenger car because all of the regulations that apply to the passenger car also apply to the truck (the reverse being not true). If this implication "trucks are cars" works in the case of prohibition, it should be noted that in case of permissions, or even worse in case of obligation, people can misunderstand the intent and reason that "if it is allowed for drivers of passenger cars, then it should be allowed for trucks too!" Thus, one is to be advised that the categories of behavior to regulate might not match the categories of the users' behaviors that need to be regulated. The legal categories may concern, for example, the use of traffic lanes and sharing them between different types of users while the users concern may be the things to be transported using vehicles in the traffic lane.

The second principle can be illustrated with an example such as the "two people kissing" category, as instantiated in Fig. 4-right panel, not to indicate that kissing is forbidden but to signify that "the departure of the train should not be delayed." However, people might have a literal interpretation: not kissing while delaying the train departure!

Signs, traffic signal displays, pavement markings, refuge islands, and demarcations the kind of information that might be affected by user-centered recommendations: on the one hand, to avoid using regulation categories instead of what is natural categories to people, and, on the other hand, to avoid using misleading analogies or pictorial metaphors that might not be matching the categories of the regulation. Note that the balance is to be provided by the goal of what legal behavior is to be attained.

The overall recommendation is to be careful with respect to the where, who, when, why, and how.

Being careful of where. For town gates, the main distinction is between the “urban built-up area” and the “outside urban built-up area.” The built-up area is defined as “the space where buildings are grouped closely together.” Its limits bounded by entrance signs and exit signs are usually defined by decision of the city council or mayor. In Europe, all permanent signs must comply with a certification using the CE marking regulation (European Conformity: European standard). The cluster of entrance and exit signs is placed along the lanes. In France, for example, these are EB panels: EB10 for the entrance panel and EB20 for the exit panel. They are designed for best visibility, especially at night, to signal to drivers that they are entering or leaving an area of concern. The road sign panels are important because they indicate the urban built-up area is the area within which the rules of driving, and police as well as town planning rules apply. However, care must be taken to avoid confusion because the location of the sign does not correspond to the real limits of the municipal territory. The road signs for entering the municipality is also of importance: it designates a transition zones from a high-speed area to a low-speed area with an increased risk of accidents, injuries and deaths (Lantieri et al., 2015).

Being careful of who. The city’s transportation regulator designs the way people will be using the town and will be using the displayed information that are provided about some of the city’s parts. In a user-centered regulation of inside town transportation, the regulator should be aware of the most usual mental representation, schemas and scenarios people use to drive themselves to enter the town and to travel inside, mainly for those that are unacknowledged about the city, such as tourists or those that are just passing by. For instance, people do have often the following mental models about a town: (1) shortest way to cross it is to go straight on, (2) main important monuments, markets, shops, parking lots and toilets will be located in the city center, (3) the city center should be crowded, and so on. Thus inquiry about who is using the town is to be the first step of an efficient transportation regulation for taking into consideration the opposite usages of the transportation means (e.g., pupils vs. drivers in hurry). Thus, traffic signs should provide information about locations and about people who live there.

Being careful of when. Inside towns, people’s typical activities are time related. Consequently, a number of permanent sign panels deliver information about when and for how long a transportation action can be executed (allowed to park here for a maximum of one hour from 10:00 to 12:00). Note that the “when/where” dimension, except for variable message signs used in “smart cities,” is a difficult regulation due to a combination of space and time constraints. A permanent sign panel is location-based. It is fixed all the time in a specific location and need a point of view and a maximum distance from where it to be seen by a viewer. If the displayed information is relevant all the time for the whole area of the town (e.g., speed limit), a single well-situated localization may be sufficient, and it should be properly seen every time it is passed. In opposite, time constrained regulations displayed on permanent sign panels are useless at times they do not apply. In addition, time constrained regulations are usually location constrained to a specific area, thus should not be displayed elsewhere. Consequently, the permanent sign panel has to be displayed in the right target place (e.g., to a square is free from parked cars when it is to function as a vegetable market place every Monday from 6:00 to 16:00).

Being careful of why and how. People follow a rule better when they understand its rationale in terms of a simple mental model. Very often, the reason for a rule is either known or implied and the purpose of sign panel is primarily to serve as a reminder. For example, Fig. 4-left displays the 50 km/h speed limit, which applies to the whole town. This sign is located where it can remind drivers, for example because a forest environment could negate the suggestion of being in a town. In opposite to the “why” dimension that relates to the understanding of the situation, the “how” dimension is to be developed and apply here and now as real actions in the current situation. Not knowing how to act is a potential dangerous situation. And because transportation regulation is in the form of “DO NOT DO,” a user that is confronted with the “DO NOT” constraint although s/he was going “TO DO IT” is facing a problem. The city’s transportation officials should be aware of providing a possible solution. For instance, if the constraint is “not to park here in order to free up a square to function as a vegetable market place this Monday from 6:00 to 16:00,” a sign indicating where else to park would be of benefit.

Conclusion: Future Uses of Town Gates

Instead of delivering information by signs, city ergonomics might be the futures where a city does not use written or pictorial instructions. Instead, the design of a smart city may be to go to providing a contextual and connected user.

First of all, following human-centered design concepts of transportation regulation inside the town gates, is designing the “what You See Is What You get” (WYSIWIG) ergonomic environment of the town. For instance, rather than using speed limits sign panels at the current town gates, at the entrance to the city, at the location which is the dangerous transition zone from high-speed areas to low-speed areas, physical measures can be used. The gateway to enter the city can instead be made of chicane deflections, bulb outs, and of wide central islands that narrow the lanes and indicate by sight a safe speed limit to drive. Vertical elements and portals can also be used. Thus, the speed limit is not delivered by signage to the user’s observance but in the physical situation.

Moreover, rumble strips and speed-activated humps in the road can cause the car and thereby the driver to tremble when the car is moving too fast or when it is not well positioned laterally. The driver’s body thus directly participates in the regulation of transport. These are promising solutions.

The next step is using smart town gates related to a connected transportation mode. This Internet of Town Gates can provide digital sign panels inside the vehicle instead of having the driver look outside the transportation mode. This can be done with connected or autonomous cars, buses or trains and can provide their users the information they need when they need it. More useful transportation regulations can thereby be providing the how-to-do-it and why, with consideration of time constraints, if any, to the target person, about what is where.

Biography



Charles Tijus heads the Cognitions Humaine et Artificielle Laboratory (CHArt), a Cognitive Science laboratory and the "Laboratoire des Usages en Technologies d'Information Numériques (FED 4246 - LUTIN), a Living UserLab located at the "Cité des Sciences et de l'Industrie" in Paris, France. Charles Tijus' PhD was on "visual memory span of 3D objects." As a Research Assistant, he has worked on visual perception with Adam Reeves (Vision Lab, Northeastern University, Boston, United States). He currently develops a contextual categorization-based approach to model the cognitive processes of understanding, decision-making, and learning. In the transportation research field, this is about road signs taxonomy, understanding pictograms as well as drivers and pedestrians' decision-making according to the contextual environment.

Relevant Websites

wikipedia: https://en.wikipedia.org/wiki/Vienna_Convention_on_Road_Signs_and_Signals#Road_markings

UK Government: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/782725/traffic-signs-manual-chapter-07.pdf

US Government: <https://mutcd.fhwa.dot.gov/index.htm>

wikipedia: https://fr.wikipedia.org/wiki/Panneau_d'entrée_ou_de_sortie_d'agglomération_en_France

References

- Bazire, M., Tijus, C., 2009. Understanding road signs. *Saf. Sci.* 47 (9), 1232–1240.
- de Lavalette, B.C., Tijus, C., Poitrenaud, S., Leproux, C., Bergeron, J., Thouez, J.P., 2009. Pedestrian crossing decision-making: A situational and behavioral approach. *Saf. Sci.* 47 (9), 1248–1253.
- Norman, D.A., 2005. Human-centered design considered harmful. *Interactions* 12 (4), 14–19.
- Tijus, C., Barcenilla, J., De Lavalette, B.C., Meunier, J.G., 2007. The design, understanding and usage of pictograms. In: *Written Documents in the Workplace*. Brill, Leiden, Nederland, pp. 17–31.
- Tournier, I., Dommes, A., Cavallo, V., 2016. Review of safety and mobility issues among older pedestrians. *Accid. Ana. Prevent.* 91, 24–35.
- Lantieri, C., Lamperti, R., Simone, A., Costa, M., Vignali, V., Sangiorgi, C., Dondi, G., 2015. Gateway design assessment in the transition from high to low speed areas. *Transp. Res. Traffic Psychol. Behav.* 34, 41–53.

Further Reading

- Castro, C., Horberry, T., 2004. *The Human Factors of Transport Signs*. CRC Press, Boca Raton, FL.
- Global Designing Cities Initiative, & National Association of City Transportation Officials, 2016. *Global street design guide*. Island Press.
- Great Britain Department of Transport, 2007. *Know your Traffic Signs*, Official Edition. TSO, London.
- Jackman, W.T., 2019. *The development of transportation in modern England*. Routledge, London.
- Sadik-Khan, J., 2012. *Urban Street Design Guide*. NACTO, New York.

Traffic Flow Volume and Safety

Athanasios Theofilatos*, **Apostolos Ziakopoulos†**, *Loughborough University, School of Architecture, Building and Civil Engineering, Loughborough, United Kingdom; †National Technical University of Athens, Department of Transportation Planning and Engineering, Athens, Greece

© 2021 Elsevier Ltd. All rights reserved.

Introduction/Overview of the Problem	692
Data Sources	693
Traditional Approach	693
Automatic Vehicle Identification (AVI)	693
Roadside Sensors	693
Video Vehicle Detection	693
Emerging Methods	693
Data Processing Using Real-time Traffic Data	694
Real-time Traffic Data Potential	694
Sum and Average of Traffic Flow	694
Variance of Traffic Flow	694
Less Frequent Traffic Flow Parameters	694
Additional Issues	694
Road Segmentation	694
Precise Crash Time Estimation and Correction	695
Safety Indicators	695
Real-time Crash Risk	695
Crash Frequency (Data and Methods)	696
AADT and Crash Frequency	696
Real-time Traffic and Crash Frequency	696
Crash Severity	697
AADT and Crash Severity	697
Real-time Traffic and Crash Severity	697
Conclusions	697
References	697
Further Reading	698

Introduction/Overview of the Problem

Modern road transport systems have improved societies by offering increased mobility and connectivity worldwide, especially since the 1950s. Despite these benefits, road safety is an ever-present issue, with roughly 1.35 million deaths occurring every year worldwide from road crashes, which constitute the leading cause of death for people of 5–29 years of age. It has been established that about 95% of all road crashes are caused primarily by human error (Singh, 2015), with the remaining percentage split between infrastructure and vehicle defects.

Factors affecting the risk of human error, such as drivers consuming alcohol or using cell phones or interacting with passengers, have been examined extensively by researchers. It can be argued, however, that traffic flow is a parameter that is tightly knit with human behavior and human error. Not only drivers are prone to changing their behavior based on fluctuations of the number and position of cars around them, ranging from becoming more aggressive or defensive, they also become more exposed to the behavior and the respective variations of others as well. If the traditional features of traffic flow that are utilized in highway engineering, transport system planning and traffic management are taken into account, it can be then assumed that traffic flow is always a relevant and interesting parameter to examine when considering road safety.

Traffic flow is defined as the number of vehicles passing a point per unit of time; it is also often called traffic volume, when the time unit is 1 h. The more straightforward approach is employing the parameter known as annual average daily traffic (AADT), which, as the name implies, is the average number of vehicles passing a point during 24 h measured over a year. This attribute is indicative of the intensity of use of an examined road segment, and can be further refined by lane or by vehicle type. AADT also has the potential of providing a hint for the macroscopic evolution of the state of the road if examined over the course of several years, and can be used to infer road safety impacts of changes in traffic volume, perhaps by interventions from road safety measures.

Other traffic flow parameters, such as hourly traffic flow, volume to capacity (V/C) ratio also known as degree of saturation and similar proxies of congestion, have also been considered in the past. All the aforementioned parameters, however, express aggregated quantities that have the drawback of being somewhat static and not reflecting the precise flow at any given time, especially when

compared to the alternatives offered in recent years of real-time data, big data, and intelligent transport systems (ITS). A disaggregated parameter with higher temporal resolution, average daily traffic (ADT) is a step between AADT and microscopic data. ADT has been implemented by researchers to gain more precise background than AADT in road safety studies, for instance as described in Wang et al. (2018).

Undeniably, the present and future trend is the use of real-time methods to obtain snapshots of the state of the road environment in a dynamic manner. These methods frequently include data collection through roadside video analyses, vehicle instrumentation, or smart phone sensors.

The aim of this chapter is the examination of the way traffic flow is treated by road safety researchers and also its overall impacts on road safety. The aim of this chapter is not to provide an exhaustive list of studies, methods, and results for real-time traffic safety. For more details about real-time traffic data and safety evaluation, the reader is encouraged to refer to the further reading list of the present chapter. Furthermore, the ever-important parameters of speed and occupancy (occupancy is defined as the time fraction during which a detector embedded in the roadway is occupied, i.e., whether there is a vehicle on top of the detector) warrant separate investigations and as such fall beyond the scope of this chapter. Similarly, parameters such as V/C ratios and hourly flows were not examined here, because when referring to the traditional approach, the most commonly used entity is AADT.

Data Sources

Traditional Approach

Until around 50 years ago, most traffic volume data were collected by roadside observers. These could be researchers or unspecialized employees that would remain by a dedicated road segment and count passing vehicles for predetermined time duration. Considerable care is required during the design of such studies on several fronts; the positions of measurement were important in order to ensure correct vehicle counting. Junctions should be adequately covered in all available entry and exit points. Similarly, the presence of complex road environments, such as ramps, tunnel exits, and junction arms should be taken into account to ensure the correct area is measured every time. The time and date of measurement is critical and tied to the objective of the research, as fluctuations and spikes are the norm: for instance, urban traffic flows are lowered during the holiday season or during nighttime.

Automatic Vehicle Identification (AVI)

Roadside Sensors

As an alternative to manual counting, pneumatic tubes placed across the roadway, typically fastened by hook that is nailed into the asphalt pavement at each end, and connected to a traffic counter, have been used since the 1930s. These tubes do not count vehicles but axles and the results are either presented as axle pairs or as estimates of vehicle numbers based on manual observations concluding that vehicles on average have, say, 2.1 axles at a given location.

An alternative to pneumatic tube that actually can distinguish actual vehicles is using automated loop detectors installed just below the top of the pavement, as those described in Yannis et al. (2014). Though even those may mistake a truck with a trailer for being two separate vehicles. Loop detectors can be temporary but are typically permanent devices installed in a lane or across the entire roadway and function by detecting electro-magnetic pulses when a metal object, such as a vehicle pass over the loop. Similar devices that are installed in proximity to the road surface, rather than on or under it, include radar, video, photocell/infrared sensors, and ultrasound sensors.

Traffic flow measurements were traditionally conducted over 24 h or more, with the counters registering the count numbers by 15 min periods but more modern equipment can record numbers for much shorter time intervals, typically of 30, 60, or 90 s; and more advanced equipment can distinguish between heavy vehicles, cars, motor cycles, and bicycles, thus providing separate counts. Data from roadside sensors can be used in crash analysis to provide insights on the traffic conditions at the time of a crash. One approach includes analyzing the values from sensors exclusively on the crash segment, while alternatively the upstream and downstream values are used in conjunction (e.g., as mentioned in Roshandel et al., 2015).

Video Vehicle Detection

Another method is the real-time analysis of video from traffic cameras to determine vehicle counts. This method is considered cost-effective as it can handle many lanes and directions simultaneously, and furthermore more precise information is available in cases where there is ambiguity or uncertainty in the measurements. Without proper design for minimized computational requirements, however, this method can lead to large and possibly unwieldy amounts of data.

Emerging Methods

Recently, new methods have been emerging in the transport sector for a wide array of applications, including traffic flow counts. These methods involve the use of affordable bluetooth and wi-fi readers that can be installed at road segments with little to no calibration. Alternative systems include utilizing smart phone data from a fraction of participant users and inferring the concurrent traffic flows from these counts.

Data Processing Using Real-time Traffic Data

Real-time Traffic Data Potential

As mentioned earlier, the introduction and rapid growth of ITS have enabled the access to modal disaggregated traffic flow data through the constant monitoring and recording of traffic at the installed locations or specific segments of freeways. Usually these kind of data are provided in very short time intervals, such as 30 s, 1 min, 2 min, or 5 min. Consequently, it has given a unique opportunity for traffic safety engineers and researchers to shift to more disaggregate types of data in order to ultimately explore the traffic conditions prior to a crash. As a result, traffic flow (volume), density, occupancy, speed, and heavy vehicle proportions could be measured. One of the most important research implications is the fact that researchers could not only calculate the total number of vehicles like AADT, but also aggregate the data to reduce noise and capture traffic fluctuations and temporal variability as well. Moreover, the analyses could also be made per lane. By that way, variability among adjacent lanes could also be captured. An illustration of typical traffic flow parameters used in research in different countries is provided in the following.

Sum and Average of Traffic Flow

The sum of traffic is a frequently used traffic-related variable in past literature. More specifically, the traffic volume is usually the sum of vehicles passing a reference point per unit of time, for example, vehicles per 2 min. However, the average value is also a common variable considered in real-time traffic analyses. Let us suppose, for example, that the raw traffic data are provided by the traffic management center of a freeway in 30 s intervals. Then, for instance, when the raw traffic data are further aggregated into 2 min time slices to reduce noise, the average traffic flow can be calculated simply by dividing the total sum of flows by the total number of time slices.

Variance of Traffic Flow

Another phenomenon of interest is the fluctuation of flow, which can be estimated in a spatial or temporal manner. In the former case for a specific time slice (e.g., 5–10 min prior to a crash), the difference in traffic flow values between upstream and downstream loop detectors is usually calculated. In some cases, the traffic flow difference between the crash segment and the upstream or downstream loop detector is also utilized as well. It is noted that if the researcher has access to disaggregate traffic data per lane, then the variability between adjacent lanes is another way to estimate spatial variability in traffic.

Overall, the spatial traffic flow difference expresses the variability in traffic conditions from a specific point to another specific point of a freeway. For example, there can be abrupt transitions from free-flow conditions at one location to more congested conditions at an adjacent location or vice versa. In international literature, temporal and spatial variations are expressed simply as continuous variables. In the spatial approach the produced variable could simply be the mathematical difference of the average flows in a specific time slice.

In the temporal approach in specific, the entity that is estimated is simply the variability in traffic conditions of a specific time slice at a specific point near a crash, that is, in the downstream loop detector. Let's assume for instance, that the raw 20 s data have to be aggregated to 2 min intervals and we wish to calculate the variability of traffic. This could be carried out by calculating the variance and the standard deviation of flow. Lastly, another proxy of traffic variability that researchers frequently resort to is the simple coefficient of variation of flow.

Overall, the aforementioned measures express the variability in traffic. Higher values of flow fluctuations can be translated, as unstable conditions and heterogeneity in traffic flow are often associated with more crashes and also more severe crashes. The variation of traffic flow and the respective impacts in traffic safety have also been examined under varying weather conditions; traffic flow was found to be reduced under fog conditions, indicating increased driver caution (Wu et al., 2018).

Less Frequent Traffic Flow Parameters

An alternative traffic flow related variable that sometimes is used is the median value of flow. Once again, data have to be aggregated in larger time intervals to obtain the median. It is straightforward to estimate the median. For instance, if the number of traffic observations is even, the median flow is calculated as the mean of the two middlemost numbers after the numbers have been ordered from the lowest to the highest. Otherwise, for traffic flow, in the say 20 s intervals, the median will be the middlemost value.

This section ends with the probably least observed traffic flow related variable, namely the chaos index, which expresses the impact of traffic chaos on traffic safety by multiplying the flow with a dimensionless coefficient which is obtained by dividing the speed variance with the average speed. When all cars travel at the same speed, they “take up less space” than the same number of cars traveling at very different speeds near each other.

Additional Issues

Road Segmentation

Another essential step in data preparation is road segmentation that is to divide a roadway into segments. One reason to resort to road segmentation is for assigning crashes to road segments (see also section of Crash Frequency). Another reason is the need to

account for unobserved heterogeneity (see also sections of Crash Frequency and Crash Severity). Generally, two main choices are available: fixed segmentation and homogeneous segmentation.

In the former case, the road network is divided into fixed-length segments, for example, 500 m. The problem with this approach is the fact that the variation in road geometry (e.g., straight segments, curves, tunnels, etc.), is not taken into account. Consequently, the latter approach is preferred in order to achieve the homogeneity in roadway alignment. In general, both horizontal and vertical alignments should be scrutinized. It is also noted that a minimum-length criterion should be generally set so as to avoid low exposure problem and low crash numbers for very short segments. It is therefore suggested that very short segments (e.g., shorter than 100–150 m) should be combined if possible with adjacent segments having similar geometrical characteristics.

Precise Crash Time Estimation and Correction

In real-time safety evaluation research, it is very important to know the precise time of a crash, because all data analyses will rely on very microscopic values of traffic parameters prior to crashes. As a result, if there is a divergence between the real time of a crash and the reported time of a crash, then the statistical models will be unreliable. For that reason, in almost all relevant research in the field, a correction at the precise time of the crash is conducted before matching the crash and collecting the required traffic information, due to errors and inaccuracies of the crash time reported by the police.

The most effective way of correcting this is by conducting a thorough visual examination of the traffic time series prior to a crash and then by subtracting a few minutes such as, 1–5 min (time lag), from the reported time of the crash to adjust it. However, larger time lags are also possible, if strong crash time inaccuracies are expected.

This correction is carried out in order to avoid the impact of the crash on the traffic itself, for example, traffic jam. More specifically, when a crash occurs on a freeway, a traffic jam is highly likely to occur near the crash location and consequently traffic speeds are facing a sudden drop. Therefore, the researcher can spot the precise moment of the crash and correct the reported crash time, which is provided by police authorities.

Safety Indicators

Real-time Crash Risk

The concept of crash likelihood or real-time crash risk evaluation has been introduced in the last decade or more. Under this approach, a sample of “non-crash” cases (i.e., time periods when no crash occurred) is selected together with the traditional crash cases. Usually a matched-case control method is applied, where a sample of non-crash cases are matched to crashes that occurred at the same location, day of week, and time of day but at different time periods. The advantage is that by using this particular method, researchers are able to isolate the impact of traffic flow related variables on real-time crash risk, while at the same time be able to control for the road segment geometry and other factors.

Determining the crash/non-crash ratio is not simple. For instance, if a crash occurred at a specific point of a freeway, then a typical approach could be to gather traffic data at the same point, same time of day, one week before, and one week after. In this example, there will be two controls; non-crash cases and one crash case. Other crash-non-crash ratios could range from 1:4 to 1:10. However, a problem is that the actual number of non-crash cases are many more and crashes could be considered as “rare events”. Although being time consuming and costly, researchers could gather all non-crash cases from a freeway (i.e., all time slices in all road segments). In any case, the researcher could choose any crash-non-crash ratio and modify the analysis methodology accordingly, since the selected data collection strategies may affect model development. A remark is that, crash risk cannot only be expressed as crash-non-crash (0 = non-crash, 1 = crash), but also be crash with severity/no-crash or crash with type/no-crash.

Regarding the methods of analysis, various different methods have been employed. Because of the fact that the dependent variable is binary, it would be expected that binary logistic regression based models were frequently utilized (fixed or random effects). After the construction of the well-known utility function, the respective crash risk probability (or crash likelihood) for any specific set of parameters can be estimated. Bayesian models are flexible and can handle problems with multiple configurations; it is noted that Bayesian inference is one of the most applied approaches to date.

Moreover, other types of the binary logistic model include the conditional logistic model, the sequential logistic model etc. Recently, the advances in computational methods stimulated a large part of research to utilize data mining, machine learning (ML), Bayesian networks as well as neural networks. Traffic time series classification is also an option that has not been extensively explored. By that method, ML classifiers are tested on the original raw time series data, which are not further aggregated (Theofilatos et al., 2018).

Since it is often pursued to explore the impact traffic flow for classification purposes, modeling estimation includes also goodness of fit measures such as, overall accuracy, recall, specificity, and area under curve (AUC). In order to calculate sensitivity and specificity one needs to estimate:

- True positive rate (correctly classified crash case as a crash case)
- True negative rate (correctly classified non-crash case as a non-crash case)
- False positive rate (a non-crash case incorrectly classified as a crash case)
- False negative rate (a crash case incorrectly classified as a non-crash case)

Then the classification performance metrics can be estimated as:

$$\text{Overall accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{Total crashes}} \quad (1)$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \quad (2)$$

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}} \quad (3)$$

Another well-known classification performance is the AUC value of the receiver operating characteristic (ROC) curve. Akaike information criterion (AIC) values, deviance information criterion for Bayesian models (DIC) and pseudo-R squares could be used when performing model comparisons.

Lastly, some general trends regarding the traffic risk factors per se, it is observed that usually higher traffic volumes lead to higher crash likelihood, however, the results are not always consistent. On the other hand, high variations in traffic temporal or spatial aspects are often strongly correlated with higher crash risk.

Crash Frequency (Data and Methods)

AADT and Crash Frequency

Crash numbers are positive integers; therefore Poisson or negative binomial regression models are the main statistical models that are employed. It is often meaningful to normalize crash numbers per road segment length, per unit of time, or more advanced parameters of exposure, such as vehicle-kilometers. This is achieved by dividing crash numbers by the respective unit, thus transitioning to crash rates, which are positive real numbers. When the impact of AADT on crash rates is examined, simple linear regression models or censored regression models can be applied. It should be noted that in both cases considerably more advanced statistical methodologies have been developed, which are beyond the scope of this chapter.

AADT has also been employed in a spatial analysis context. Spatial correlation, indicating that neighboring areas have correlated crash counts, has been discovered for intersections along a given corridor due to similar patterns in traffic flow. Spatial analyses have shown that higher AADTs directly contribute to increased crash rate, measured as crashes per million entering vehicles. But, the crash rate for intersections depends more on what percentage of traffic is carried by the minor road than on the total entering volume. So, if we have, say, 10% of traffic on the minor road and 90% on the major road, increasing the flows gives a higher crash rate. For example, a Swedish model based on its entire state-road network shows that doubling the total entering volume at a non-signalized T-intersection from 10,000 to 20,000 vehicles per day (vpd) increases the crash rate by about 37% if we keep the percentage of traffic on the minor road at 10%. But if we instead add most of the “new” traffic to the minor road, so that they have close to equal traffic volumes and still an entering volume of 20,000 vpd, the crash rate is increased by over 200%. And, even lowering the total entering volume to 5,000 vpd with a 50/50 split (having 2,500 vpd on each road) gives a 64% higher crash rate than the double entering volume with 9,000 vpd on the major road and 1,000 vpd on the minor road. Traffic flows have been used in spatial studies extensively as exposure parameters in order to establish a common basis for crash risk comparison.

Real-time Traffic and Crash Frequency

A major difference between real-time crash risk and frequency models is that the latter are more aggregate. More specifically, when analyzing crash frequency with real-time traffic data, the road network should be divided in road segments in order to analyze the impact on the number of crashes per segment for a given time period (day, month, etc.). Aside from the flow variables discussed earlier, traffic volume can also be multiplied by the segment length in order to express a proxy of exposure.

Overall, a few issues arise in crash frequency analyses utilizing real-time data. First, when a road segment has experienced zero crashes, it is not clear which values of traffic flow should be used. One solution could be to rely on an even more aggregate but representative variable, such as a sensible average value of traffic flow (e.g., seasonal, average flow of the adjacent segments, etc.). Another solution is to use different road segmentation. Also, when zero crashes have occurred, it is still obvious that the expected number of crashes for that segment and that time period is greater than zero, even if only by a small fraction of one crash. When we predict crash numbers for a segment, we can give estimates with several decimals, but when we construct our models, we cannot easily determine the expected number of crashes for a segment with zero crashes recorded.

Another remark is the application of a robust methodology, when there are segments with more than one crash. In this case, the researcher has to take the mean value of the traffic variables from different time periods. The problem with this issue is that much information is lost and the estimates of traffic flow might not be reliable. Consequently, the use of disaggregate data in such cases does not have the expected benefits. To the best of our knowledge, this issue has yet to be solved.

Another issue is unobserved heterogeneity, as the impact of traffic flow may vary among road segments, thus not being fixed. For that reason, the models that are preferred are hierarchical or random effects models. It is noted, that the problem of excessive

zeros (i.e., a large number of segments with no reported crashes) can be solved by applying zero-inflated Poisson or negative binomial models.

In general, increased flow values are found to be associated with an increased number of crashes, but at a more aggregate level (i.e., daily counts). Please note that traffic occupancy (the fraction of time during which a detector embedded in the roadway is occupied) or degree of saturation values have been found to have the same effect and as such, the same (indirect) effect of real-time traffic flow can be considered due to the correlation of flow and occupancy. Overall, this topic needs further research as there is a lot of open issues that need to be tackled.

Crash Severity

AADT and Crash Severity

Crash injury severity is a parameter which requires a higher degree of data collection (number of injured people and separate degrees of injuries), which may also exceed some of the aforementioned data collection processes since the consideration of single vehicles is often not enough (for instance pedestrian casualties). Separate data collection methods are required instead, that can involve third parties (hospitals, insurance companies etc.). This reality, combined with the (fortunately) low overall numbers of crashes, results in crash injury severity being not as investigated as crash frequency.

When injury severity is considered, usually an approach of binary nature is adopted, depending on whether property damage only (PDO) crashes are recorded as well: For instance, 1 is assigned to crashes with injuries or fatalities, while 0 to PDO crashes. Alternatively, if PDO crashes are not recorded, researchers can examine mortality: 1 is assigned to crashes with severe injuries or fatalities while 0 to non-severe crashes. Finally, there are studies dedicated to injury severity that exploit rich and detailed datasets and employ methods like multinomial logit models in order to determine the characteristics of parameters leading to the crash. In all the aforementioned cases, AADT or ADT is used either as an exposure parameter or as one of the explanatory independent variables.

Real-time Traffic and Crash Severity

The impact of real-time traffic flow parameters on crash severity has received less attention than crash risk. This could be attributed to the fact that real-time safety evaluation has more focus on preventing a crash from occurring, rather than explore how a severe a crash could be. Similar to previous studies in the field examining the impact of AADT on crash severity, severity of a crash is mainly binary in nature, namely severe/fatal versus non-severe, such as in [Christoforou et al. \(2010\)](#).

A similar issue arising when studying crash severity with real-time traffic data is the potential unobserved heterogeneity. Hierarchical models (i.e., with random effects) or other models (i.e., latent class models) which address this issue could be established on the basis of either segment level or crash individual level, because these are the two main sources of heterogeneity. These models allow a number of variables to vary across different roadway segments or crash observations and in this way they account for the unobserved effects (roadway characteristics, environmental factors, and driver behavior). Other models should be used when the correct prediction is sought, for example, machine learning models, such as support vector machines. The problem is that machine-learning models are basically black box methods with limited interpretability of results; therefore a sensitivity analysis is needed as well to complement the analysis.

Regarding some core findings so far, it could be suggested that congested traffic leads to less severe crashes because of the lower driving speeds, but variations in traffic flow increase the crash severity, especially when being close to but just below congestion when traffic alternates between high speeds and almost stand still. Overall, research on the influence of real-time traffic on crash severity is relatively limited and more future studies should focus on this topic.

Conclusions

Research regarding the impact of traffic flow on road safety analysis is of critical importance and is gaining increasing attention. In the era of big data, social media, and other technological advances, there is still a lot of room for further research. For instance, if black boxes are installed in all vehicles, information on the trajectories, speeds, headways, and maneuvers of all vehicles could be acquired and analyzed. The exploitation of the large amounts of data through appropriate advanced real-time analyses could result in highly effective pro-active traffic safety management systems with great potential for safety improvement. Additionally, there is still research needed regarding the impact of traffic flow on motorcyclist safety. Lastly, the upcoming era of connected and autonomous vehicles (CAVs) could bring radical changes to our knowledge of the impact of traffic characteristics on road safety and more research in that direction is needed.

References

- Christoforou, Z., Cohen, S., Karlaftis, M., 2010. Vehicle occupant injury severity on highways: an empirical investigation. *Accid. Anal. Prev.* 42, 1606–1620.
- Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: systematic review and meta-analysis. *Accid. Anal. Prev.* 79, 198–211.

- Singh, S., 2015. Critical reasons for crashes investigated in the national motor vehicle crash causation survey (No. DOT HS 812 115). Washington, DC.
- Theofilatos, A., Yannis, G., Antoniou, C., Chaziris, A., Sermpis, D., 2018. Time series and support vector machines to predict powered-two-wheeler accident risk and accident type propensity: a combined approach. *J. Transp. Safe. Secur.* 10 (5), 471–490.
- Wang, L., Abdel-Aty, M., Wang, X., Yu, R., 2018. Analysis and comparison of safety models using average daily, average hourly, and microscopic traffic. *Accid. Anal. Prev.* 111, 271–279.
- Wu, Y., Abdel-Aty, M., Lee, J., 2018. Crash risk analysis during fog conditions using real-time traffic data. *Accid. Anal. Prev.* 114, 4–11.
- Yannis, G., Theofilatos, A., Ziakopoulos, A., Chaziris, A., 2014. Investigation of road accident severity and likelihood in urban areas with real-time traffic data. *Traffic Eng. Control* 55 (1), 31–35.

Further Reading

- Abdel-Aty, M., Shi, Q., Pande, A., Yu, R., 2018. Real-time traffic safety and operation. In: Dominique, L., Simon, W. (Eds.), *Safe Mobility: Challenges Methodology and Solutions*. Emerald Publishing Limited, Bingley, pp. 175–204.
- Ahmed, M., Huang, H., Abdel-Aty, M., Guevara, B., 2011. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accid. Anal. Prev.* 43, 1581–1589.
- Hossain, M., Abdel-Aty, M., Quddus, M.A., Muromachi, Y., Sadeek, S.N., 2019. Real-time crash prediction models: state-of-the-art, design pathways and ubiquitous requirements. *Accid. Anal. Prev.* 124, 66–84.
- Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road section: Poisson versus negative binomial regression. *Accid. Anal. Prev.* 26, 471–482.
- Noland, R., Quddus, M., 2005. Congestion and safety: a spatial analysis of London. *Transp. Res. Part A* 39 (7–9), 737–754.
- Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. *Accid. Anal. Prev.* 72, 244–256.
- Theofilatos, A., Yannis, G., Kopelias, P., Papadimitriou, F., in press. Impact of real-time traffic characteristics on crash occurrence: preliminary results of the case of rare events. *Accid. Anal. Prev.* Available from: <https://doi.org/10.1016/j.aap.2017.12.018>.
- Wang, J., Huang, H., 2016. Road network safety evaluation using Bayesian hierarchical joint model. *Accid. Anal. Prev.* 90, 152–158.
- Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accid. Anal. Prev.* 50, 371–376.
- Yu, R., Abdel-Aty, M., 2014. Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. *Accid. Anal. Prev.* 62, 161–167.
- Zheng Z., 2012. Empirical analysis on relationship between traffic conditions and crash occurrences. *Soc. Behav. Sci.* 43, 302–312. Presented at the Eighth International Conference on Traffic and Transportation Studies. August 1–3, 2012, Changsha, China,

Traffic Safety and Security of Taxis and Ride-Hailing Vehicles

Zhe Wang, Helai Huang, Ye Li, School of Traffic and Transportation Engineering, Central South University, Changsha, China

© 2021 Elsevier Ltd. All rights reserved.

Background	699
Features	700
Comparison of Taxi Drivers, Ride-Hailing Vehicle Drivers, and Private Car Drivers	700
Comparison of Ride-Hailing Vehicle Drivers and Taxi Drivers	700
Facts	700
Taxi and Ride-Hailing Vehicle Accidents Analysis of Factors	700
Drivers	700
Age	700
Gender	701
Fatigue	701
Income	701
Speed and Speeding Violation	701
Vehicle	702
Vehicle Condition	702
Vehicle Color	702
Road	702
Road Section	702
Roadway Type	702
Road Surface	703
Horizontal Element	703
Travel Time	703
Day of Week	703
Time of Day	703
Vehicle Occupancy	703
Safety Belt Use	703
Speed Limit	704
Collision Type	704
Recommendations	704
Safer Drivers	704
Safer Vehicles	704
Ensure the Safety of Passengers	705
Relevant Websites	705
References	705
Further Reading	705

Background

Taxis play a significant role in the public transit system all over the world. In New York City, where the private car ownership rate is relatively low, the taxi service has a high utilization. In Europe, although private car ownership is relatively high, the taxi industry still employs more than 1 million people, accounting for 8% of transport employment (Bidasca and Townsend, 2016). With the rapid development of the sharing economy and the advent of mobile payment, ride-hailing service becomes a new choice of travelers. The ride-hailing service is characterized by the network information transmission and application ability and is different from the traditional taxi service. Globally, ride-hailing companies, such as Uber, Lyft, Ola, and DiDi, have jointly built a huge mobile travel network that serves more than 80% of the world's population and covers more than 1000 cities. In the context of traditional taxis and emerging ride-hailing service, traffic safety and security of these two modes become a vital issue that cannot be ignored. The objective of this study is to analyze and summarize the traffic safety and security of taxi and ride-hailing vehicles. We focus on the major risks and accident factors faced by taxi drivers and ride-hailing vehicle companies and provide suggestions for taxi and ride-hailing vehicle drivers as well as passengers to reduce risks.

Features

Comparison of Taxi Drivers, Ride-Hailing Vehicle Drivers, and Private Car Drivers

Taxi and ride-hailing drivers are more likely to be involved in accidents and accident-related injuries due to their occupational exposure to hazardous road conditions and particular activities, such as seeking locations, spotting potential passengers, and stopping to pick up or drop off passengers (Baker et al., 1976; Johnson et al., 1999; Chin and Huang, 2009). There are some factors that will influence driver behavior in taxi operations compared with private drivers, such as drivers, vehicles, roads, and so on (Chin and Huang, 2009). The corresponding factors and related reasons will be detailed introduced in the following text. The nature of the job is shift work, and the process includes cruising the streets, waiting at a fixed point, and a mix of both (Tseng, 2013).

Comparison of Ride-Hailing Vehicle Drivers and Taxi Drivers

Ride-hailing drivers may be less familiar with the road network compared with taxi drivers and they usually drive with navigation. While driving, drivers will check their mobile phones or car navigation screens from time to time, which may lead to potential safety risks. Besides, ride-hailing drivers have additional mobile phone operations, because customers request vehicles through their smartphones and ride-hailing drivers need to contact with the passenger and locate the passenger.

Taxi companies regularly or irregularly train taxi drivers on policies and regulations, etiquette, professional ethics, foreign languages, and other aspects. Traffic police and other departments also train taxi company managers and drivers in relevant aspects. In some countries, like Sweden, a driver needs a special national license to drive a taxi, which includes passing criminal background checks, medical tests, theoretical tests, and driving tests that go far beyond the standard testing for driving a passenger vehicle. Before March 2018, ride-hailing drivers in Sweden only needed a regular driver's license but, since that time, these drivers are also required to have the special taxi license. The management of ride-hailing vehicle drivers is relatively simple. The professional ability of ride-hailing drivers may in many countries be insufficient, and moral quality is uneven. The security problem of ride-hailing vehicles is also very prominent.

Facts

The traffic accident statistics in Shenzhen city in 2017 show that a total of 3879 crashes occurred that involved ride-hailing vehicles. In terms of the proportion of the number of vehicles involved and the total number of platform registrations, the proportion of ride-hailing vehicle accident accounts for 7.15% of all reported accidents, while the taxi traffic accident proportion is 1.78%. Private car accident accounts for 0.28%. The difference in traffic accident rates among traditional taxis, ride-hailing vehicles, and private cars is mainly related to the amount of travel mileage. In the same period, the liability rate of traffic accidents for taxis was 55.6%, very similar to the 54.5% of ride-hailing vehicle drivers being deemed at fault.

China judicial big data service platform has released a list of crimes committed during online ride-hailing and traditional taxi services in 2017. The incidence rate of 10,000 traditional taxi drivers is 0.627, while that of ride-hailing drivers is 0.048. In the process of providing services, the nighttime is the peak time for a crime. About 50% of the cases for ride-hailing drivers and nearly 30% of the cases for traditional taxis take place in this period. As reported by CNN on January 21, 2019, Uber has 1200 incidents reported every week. This includes rapes, other sexual assaults, serious traffic accidents, and even deaths. They also speculated that many incidents are not reported so the true numbers may be even higher even if some of the included reports are false reports.

Taxi and Ride-Hailing Vehicle Accidents Analysis of Factors

Since very few studies have been conducted specifically for the safety of ride-hailing vehicles, this paper majorly summarizes the accident factors obtained for generalized taxis, which may also be useful for understanding the ride-hailing vehicles given their similar operational features.

Drivers

Drivers of taxi and ride-share vehicles can be assumed to behave similarly to drivers of other vehicles. However, it is possible that drivers of these vehicles differ with respect to age, gender, or personality type. In some countries, it may be possible to get a ride-sharing job even if you have a personality type that makes it difficult to hold employment in many other sectors. The average education level may also differ.

Age

As an important risk group in traffic, the contribution of young drivers to motor vehicle deaths is confirmed to be disproportionate. The number of fatal crashes per 100,000 licensed drivers between the ages of 15 and 24 is much higher than those from 25 to 45-year-old people (Karpf and Williams, 1983). Young drivers drive more recklessly compared with older drivers. Besides, young drivers have less ability to perceive the environmental risks involved in the operations of motor vehicles compared with more experienced drivers

(Matthews and Moran, 1986; Finn and Bragg, 1986). Thus, youth participation in motor vehicle accidents in the United States and around the world has long been considered a major public health problem.

Moreover, there are, at least in Singapore, 50- to 59-year-old and above 59-year-old drivers who have a higher probability of being involved in accidents than those in their 40s. Causes have been found for the 50- to 59-year-old group and above 59-year-old group, respectively. The common cause is failing to give way. Drivers who are above 59 years old have a little bit more risks than those aged 50–59 because of causes including changing lanes without due care, failing to have proper control, and turning without due care. It is clear that the older group has more difficulties in making good judgments associated with traffic conditions, even though older people have more propensity to obey traffic rules. Failing to give way is another common problem among older drivers, which shows that older people have reduced capability in judging road rights. The most significant cause for above 59-year-old drivers having crashes is changing lanes without due care, which may be associated with a retrogressive ability to judge gaps in traffic streams. For the more senior group, failing to have proper control and turning without due care are also significant, suggesting that the ability to interact with other vehicles may also deteriorate with age when an interaction is required. These problems could be because the elderly have worse issues with physical, perception, and reaction capabilities derived from declining eyesight, weaker muscle strength, and reduced alertness.

However, in the United States, where elderly drivers have been driving since they were young, the driver fatality rate per mile driven stay at a low level and constant for age groups from 30 to 64 and then increase slowly for people up to 70, and by age 75 reach the same level as that of 20–24-year olds, and by age 80–85 is roughly as high as for teenage drivers.

Gender

Studies in some countries have shown that female taxi drivers have a higher risk of accident-related mortality and injury than male taxi drivers. This result accords with the data reported in a study in Australia that the rate of hospitalization of female car drivers is higher than that of men (Lam, 2004). This may be because accidents involving female drivers are more frequently caused by panic, especially for novice female drivers. On the other hand, male drivers are more prone to road rage and aggressive driving. And, overall, US data indicate that women on average are much safer drivers than men. FARS database for 2015 shows that there were 32,082 fatal crashes in the United States involving 48,613 vehicles. Those vehicles had male drivers in 35,472 cases and female drivers in 12,220 cases and unknown or unreported gender (or no driver) in 921 cases. This gives a male to female ratio of 2.90 (or 74.3% males and 25.7% women out of known gender). The Federal Highway Administration (FHWA) stated in July 2016 that the average miles driven by male drivers were 16,550 miles per year and by female drivers 10,142 miles per year. If we had equal number of male and female drivers, that would mean that 62.0% of all miles were driven by men and 38.0% by women. But the numbers of male and female drivers in the United States are not identical. Since 2005, there are slightly more female drivers and therefore the percent of all miles driven by men in 2015 was slightly below 62.0%, or about 61.8%. So men should be the driver in 61.8% of all fatal crashes if they were equally safe, but men were involved in 74.3%, so men are 20.2% worse than the average driver $[(74.3 - 61.8)/61.8 = 0.202]$. Women should be the driver in 38.2% but are involved in only 25.7%, meaning they are 32.7% safer than the average driver. This means that if the average driver was involved in 100 fatal crashes, male drivers, driving average distance, would be involved in 120.2 and female drivers would be involved in 67.3. In other words, men have 78.6% more fatal crashes than women per mile driven.

Fatigue

Driving for more than 8 h a day significantly reduces the response time of drivers. This finding is supported by a study showing that more than 2 million crashes in the United States in 2008 were attributed to driver fatigue.

In accident causal investigations, taxi drivers do not usually attribute accidents to fatigue, because the drivers may not realize they are tired, or if they realize it, they may not admit it (Corfittsen, 1993). If there is no emergency during the driving, a driver could drive without incidents. However, once there is an emergency, a driver may react too slowly to avoid the accident because of fatigue or sleepiness. Drivers may attribute accidents to the emergency instead of fatigue. Thus, fatigue is a potential boost to accidents through slow operations and awareness of danger. Fatigue generally comes from lack of sleep and circadian rhythm disruption (Dalziel and Job, 1997). For taxi drivers and ride-hailing drivers, picking up passengers is a career to make a living. Taxi drivers and ride-hailing drivers may choose to reduce the rest time as much as possible and constantly look for potential passengers. Moreover, taxi drivers may work in shifts. For those who do not work in regular shifts, the fatigue caused by the chaotic biological clock is a big safety risk. In many countries, ride-hailing services do not have limits to the hours a driver can work and even in a country like Sweden, a driver can operate up to 13 h per 24-h period, since the rest period needs to be only 11 h per day. Furthermore, not all vehicles are equipped with electronic logging that shows who has been operating a specific vehicle.

Income

US studies show that there is a strong relationship between a driver's crash rate and income. At least one study shows that the higher the driver's income, the lower the crash rate. Also, when the United States hit a deep recession in 2009, which lowered the incomes of taxi drivers, their crash rates significantly increased.

Speed and Speeding Violation

Literature shows that speed is an important factor in road safety and has a direct causal relationship with traffic accidents. The risk of a fatal crash is roughly proportional to speed to the power of four, meaning that if speed is doubled, the expected risk of a fatal crashes increases by a factor of 16. And, taxi and ride-share drivers may speed more than other drivers. One main harm of driving at higher

speed is that the line of sight becomes narrower and the definition decreases. When traveling faster, the braking distance also becomes longer. When speeding, people's psychology may also be stressed, since they fear being caught by police, especially under complex road conditions, in which drivers may panic and make operational errors. As for taxi drivers, it has been confirmed that they commit speeding violations associated with operating styles, working time, work experience, daily driving distance, and sensation-seeking personality (Tseng, 2013). Novice taxi drivers are more likely to commit speeding violations than senior taxi drivers because of insufficient knowledge of local geography and unfamiliarity with the roads and traffic conditions. Drivers who cruise the streets are less likely to commit speeding violations than drivers on route to pick up or deliver passengers. Taxi drivers cruising down a street often need to find a fare for the next ride and have more opportunities to turn around or stop suddenly to pick up passengers so they might drive at a slower speed but may also brake or turn suddenly. However, drivers waiting at fixed points usually take passengers directly to their destinations, and they may drive faster to shorten the travel time and increase the frequency of workload. People who drive late at night have a higher chance to commit speeding violations than those who drive at other times. One possible reason is that the traffic situation at night often is less congested than that during the day, so that drivers may drive at higher speeds. Besides, the more daily driving miles, the more speeding violations. This conclusion is following that driving mileage is the most important predictor of driver violations and accident involvement. Moreover, it has been suggested that people with sensation-seeking personalities are more inclined to have speeding violations, and they may also be more likely to be taxi drivers than people with other personality types (Burns and Wilde, 1995).

Vehicle

Vehicle Condition

In the United States and Europe, having a vehicle factor being the main contributor to a crash occurs in less than 3% of crashes, but vehicle factors contribute to approximately 13% of all crashes. Problems of vehicles, especially in poorer countries, often include lighting problems, brake failure, and a flat tire, among which the most important factor is the problem of tire failure, which is also the main cause of high-speed traffic accidents in China. The cause of flat tires is generally that the owners pay little attention to maintenance, including high or insufficient air pressure, and tires being severely worn. This may be more common for taxis than other vehicles since the owner often is not the operator. Road conditions and excessive and severe tire temperatures are potential dangers causing flat tires. And, the tire may get punctured by sharp objects, followed by air leakage which causes a flat tire. Also, if the ambient temperature is too high, or driving at higher speeds than the tire is rated for over long time periods, can make the tire temperature too high, and may cause the tire to burst, thus prone to traffic accidents. Besides that, the taxi may not necessarily be a taxi driver's car, where they do not pay attention to the maintenance of the car, taxis, and ride-hailing vehicles are driven longer distances than most vehicles. Hence, the vehicle condition may degrade quicker, meaning that annual safety inspections—which are not even mandatory in all jurisdictions—are not enough to maintain the vehicle in a safe condition, which may be extra important for taxi and ride-hailing drivers.

Vehicle Color

Studies show that yellow taxis have fewer accidents per month than blue taxis (Ho et al., 2017). The reason is that yellow taxis are more visible than blue ones, especially in front of another car. Other drivers are therefore better able to avoid hitting them, especially at night, and the accident rate is therefore reduced. Also, when driving in rain or fog, drivers have a worse vision, and the vehicle color may play a vital role in sideswipe and rear-end crashes. For head-on collisions, vehicle color has very little effect since little of the frontal area is painted and operating with daytime running lights is a much more important factor in being visible to others (Ho et al., 2017).

Road

Road Section

Per distance driven, intersections constitute a higher risk than other road sections. And, there is a higher probability of taxi drivers being involved in accidents when driving through intersections than elsewhere. At intersections, drivers need to adjust speed due to the signal controls or other junction controls. Besides, it is indicated that taxi drivers have higher risks at intersections compared with general car drivers. Taxi operations with an increased workload are responsible for the higher risk. The additional workload includes picking up or dropping off passengers near intersections as well as unfamiliarity of route and traffic conditions. Significant causes of accidents at intersections are disobeying traffic signs and signals, failing to give way, and turning without due care. These three causes may be because of reckless or careless driver actions resulting from high workload as mentioned earlier or from these drivers being more aggressive than drivers in general.

Roadway Type

As for taxis, like other drivers, limited-access highways (expressways) are safer than other roadway types. Taxi drivers have a higher risk of being responsible for accidents on non-expressway than the expressway. The reasons for the relatively high tendency are as follows: failing to give way, disobeying traffic signs and signals, and turning without due care. In non-expressway conditions, taxi drivers have more pressure resulting from the requirement to stop for boarding and alighting, look for potential passengers, and make quick decisions to navigate. In the process, taxi drivers may be at fault derived from indiscriminate lane changing, turning,

stopping, and possible lack of concentration. Besides, taxi drivers may have little awareness of street and traffic conditions. Taxi drivers' errors of judgment and actions affected by these possible factors will increase the risk of accidents.

Road Surface

Wet road surfaces make driving riskier. Taxi drivers being at fault have a higher risk on wet road surfaces compared with dry road surfaces. Two causes that have been found to have a higher propensity are failing to keep a proper lookout and failing to have proper control. When driving on wet roads, it is easy to slip, and driving in the rain, fog, and snow will be affected by not only wet roads but also reduced visibility, increasing the workload and driving tension of taxi drivers and ride-hailing vehicle drivers. Taxi and ride-hailing vehicle drivers, by nature of their profession, are subject to longer periods of stress in rainy conditions and bad visibility, which may reduce performance level and increase the likelihood of incorrect judgments. Under this circumstance, taxi driving will be more hazardous.

Horizontal Element

Taxi drivers have a higher risk on straight roads than in horizontal curves, compared with private car drivers. This factor is significant because of the way taxi drivers operate. There is a clear distinction between taxi drivers and private drivers driving on straight roads. More passenger-related activities exist on straight roads, including stopping for boarding and alighting for passengers. Because of this demand for more stops of taxi drivers while private car drivers do not have to do, and the longer period of stopping at the request of passengers, potential risks in the process of waiting make taxi driving more hazardous.

Travel Time

Day of Week

Taxi driving on weekdays has a higher probability to involve accidents than on weekends. Compared with ordinary car drivers, taxi drivers are more likely to get involved in accidents when driving on weekdays. Several causes are responsible for the higher risk. Driving during workdays is more stressful because the traffic is heavier and tends to be concentrated during peak periods due to commuting and the driver workload is higher due to increased interaction with traffic in a variety of operations. Besides, in modern cities such as Singapore, there are traffic controls that are only implemented on weekdays, such as an exclusive bus lane system and the Electronic Road Pricing (ERP) system, which increase the workload of drivers' in-vehicle control and navigation and make driving more dangerous. Thus, when taxi drivers drive on a weekday, a higher level of mental alertness is needed and this affects driving abilities.

Time of Day

Studies have found that nighttime driving has higher risks compared with other driving time for all drivers. Taxi drivers' nighttime driving has an even higher chance of involvement in accidents than daytime driving (Corfitsen, 1993). Although there are fewer pedestrians and lower traffic flow at night than during the day, visibility at night decreases, which negatively affects reading traffic signs and interpreting the street environment. Furthermore, travel speeds may increase during nighttime. And it has been suggested that poor vision and visual problems are associated with an increased risk of motor vehicle accidents among many taxi drivers.

Moreover, compared with general car drivers, taxi drivers have a higher probability to be involved in accidents when driving at night than in the day. Because of the nature of taxi drivers' work, they may need to drive in unfamiliar areas, and they may not be proficient in local geography and traffic conditions. Besides, taxi drivers still have the greater need to seek locations, spot potential passengers, stop for boarding and alighting under poorer night-vision conditions, and taxi drivers may suffer from fatigue. It appears that more attention should be paid to night driving, especially for taxi drivers' nighttime driving which has a considerable risk to drivers as well as passengers.

Vehicle Occupancy

The risk is increased by over 20% for taxi drivers who are not carrying any passengers. It is reasonable that taxi drivers will be more cautious when taking passengers (Lam, 2004). Once passengers are injured in a taxi, taxi drivers may have financial liability. Furthermore, taxi drivers need to pick up passengers, and when there are no passengers in cars, they will try their best to find possible passengers or answer the phone for a reservation. To maximize the passenger-carriage time, taxi drivers may rush to the waiting passengers for picking up. Under such circumstances, reckless driving behaviors may happen, such as speeding, changing lanes without due care, and turning without due care, and hence increase the chance of accidents.

Safety Belt Use

Safety belt use has been shown to reduce motor vehicle fatalities by about 45% and serious injuries by more than 50% (Fielding et al., 2001). However, there is in many jurisdictions an exemption for taxi drivers that they do not need to wear seat belts because of the need to frequently enter and exit a vehicle. The ratio of taxi drivers' seat-belt use in Boston is less than 10%, significantly lower than other drivers in Massachusetts (Fernandez, 2005). The safety impact of not using seat belts is very serious, so seat-belt legislation should be strengthened. Even in Sweden, which has mandated seat-belt use in taxis since 1999, it is still permitted that children travel short distances at low speed in the back seat without proper child seats. But seat-belt use in taxis in Sweden has been above 90% since

2006 and reached 96% in the latest comprehensive state-organized observation study in 2013, which included over 1200 taxis. That is almost as high a usage rate as in other passenger vehicles.

Even with safety restraints, taxi passengers have higher injury rates compared to car occupants. Experts believe that this trend is linked to the presence of partition in most taxis.

Speed Limit

The speed limit has been identified as a significant factor in risk. Many countries use 40 km/h, 70 km/h, and 90 km/h as standard speed limits and they are confirmed to be significant among categories of speed limits. On low-speed roads with a 40 km/h limit, taxi drivers have a higher risk of involvement in accidents than on 50 km/h streets in the empirical studies conducted in Singapore. The causes are failing to give way, changing lanes without due care, and turning without due care. The reason for accidents on the low-speed road is associated with passenger-related activities, including seeking for potential passengers and stopping and waiting for picking up or dropping off passengers. In this case, the taxi drivers have to undertake more searches and navigation workload. Besides, taxi drivers may have some reckless behaviors due to unfamiliarity with the road environment and the low-speed roads. It could be suggested that the speed limit on urban local streets should be lowered from 40 to 30 km/h, as in much of Europe, or even to 20 or 10 km/h as is the case in most of downtown Berlin since 2017.

The 70 km/h roads are typically semi-expressway routes, and taxi drivers are more likely to be involved in accidents there too. Driver errors are associated with navigation and other distracted attention at high speed rather than passenger activities. The causes of accidents are failing to have proper control and following a car too closely. It is more associated with driver behaviors than taxi operations. At a 90-km/h speed limit, there is an increased probability of accident responsibility when compared with the reference category. Several other causes are dominant on the expressway. These are failing to keep a proper lookout, changing lanes without due care, and failing to have proper control. Contrary to earlier findings, the expressway is less risky than other roads. The finding could be related to some high-risk behaviors by taxi drivers, such as speeding, which could indicate that taxi drivers are more likely to make mistakes in such situations. Although expressways are safer environments than urban streets, it suggests that expressway driving still has certain risk associations. Since the speed limit is not one of the important factors related to taxi operation, these specific expressway driving risks are related to poor driver behaviors.

Collision Type

Taxi drivers are nearly twice as likely to be involved in multi-vehicle collisions as in single-vehicle collisions when compared with general car drivers. This factor is related to taxi operations rather than drivers themselves. The significant causes include failing to give way and changing lanes without proper care. It may be because of drivers' making errors of judgment or action when responding to passenger calls rather than poor driving skills.

Recommendations

Safer Drivers

- ETSCs PRAISE (Preventing Road Accidents and Injuries for the Safety of Employees) project provides some corresponding countermeasures about how to make taxis safer ([Bidasca and Townsend, 2016](#)).
- Mandatory driving training programs and education and information need to be developed to ensure that taxi drivers are better prepared to conduct their business with considerations of the safety of passengers and other road users. Ride-share drivers should be mandated to have the same licensing requirements as taxi drivers if they pick up passengers for a fee.
- Countermeasures need to be taken to prevent fatigue among taxi drivers. This can be achieved through legislation to limit working hours and mandating electronic log systems where drivers use their personal state-issued taxi certificate to ensure that drivers do not drive extra hours for more than one company, or drive under two identities. Education, information, and training will be needed to get acceptance for this and other measures.
- Taxi drivers and all passengers including children should be required to wear seat belts or be seated in child seats/on fixed cushions if shorter than 135 cm, with no exception for short trips.
- Strengthen the law enforcement of drunk driving for improving the safety of services.
- Taxi companies formulate reasonable safety management policies to ensure that shift patterns, travel plans, employment contracts, and work schedules do not create time management stress for employees, leading to driver fatigue and stress.
- The ride-hailing vehicle platform and management department need to strengthen the function of reviewing and ensuring the safety of ride-hailing vehicle drivers and include the ride-hailing service into the taxi-service assessment system.

Safer Vehicles

- Develop vehicle management policies and procedures.
- The taxi vehicle safety technical requirements, including vehicle occupant protection, vehicle safety technical equipment, etc., shall be formulated strictly following the legal requirements.

- Those responsible for the procurement of taxis within the organization should communicate closely with the safety department and relevant supervisors and management.
- Taxi companies communicate vehicle safety technology to their employees to keep them safer and train them to use the equipment properly.
- Maintenance personnel shall periodically assess the appropriate high standards and ensure that approved replacement parts are used in the vehicle, particularly for safety-critical components such as brakes and tires.

Ensure the Safety of Passengers

- Draw up a list to be noticed and advertised when booking and entering a taxi.
- To equip taxis with child safety restraints where practicable.
- Allow passengers to provide meaningful feedback to drivers on key risk factors (speed, seat belts, pickup locations, etc.).
- Taxi drivers carrying designated vehicles for people with mobility problems should be trained and equipped with wheelchairs at no additional charge.
- Taxi passengers themselves can take measures to ensure their risks are minimized. When using the services of the taxi or the ride-hailing platform, passengers should have appropriate safety awareness.
- Passengers are often embarrassed to tell drivers to slow down. And, some passengers even egg on drivers because they like traveling fast or are short on time to get to airports, etc. Therefore, within a reasonable number of years, all taxis should be equipped with speed governors that use GPS or automatic sign reading, making it impossible for drivers to exceed the speed limit by even 1 km/h except for very short distances, similar to what is proposed to be an EU rule for all motor vehicles starting with new models from May 2022 on. In taxis, such systems should be retrofitted into older models too at a later date.

Relevant Websites

Southern Metropolis Daily. Available from: http://epaper.oeeee.com/epaper/H/html/2017-06/02/content_33360.htm.
 Prospective Industry Research Institute. Available from: <http://www.qianzhan.com/analyst/detail/220/180628-b51ef43b.html>.
 The Supreme People's Court of The People's Republic of China. Available from: <http://www.court.gov.cn/zixun-xiangqing-120431.html>.
 Oriental News. Available from: <http://mini.eastday.com/a/170829225941038.html>.

References

- Baker, S.P., Wong, J., Baron, R.D., 1976. Professional drivers: protection needed for a high-risk occupation. *Am. J. Public Health* 66 (7), 649–654.
- Bidasca, L., Townsend, E., 2016. Making taxis safer: managing road risks for taxi drivers, their passengers and other road users. Available from: http://etsc.eu/wp-content/uploads/TAXI_report_final.pdf.
- Burns, P.C., Wilde, G.J.S., 1995. Risk taking in male taxi drivers: relationships among personality, observational data and driver records. *Pers. Individ. Differ.* 18 (2), 267–278.
- Chin, H.C., Huang, H., 2009. Safety assessment of taxi drivers in Singapore. *Transp. Res. Rec.* 2114, 47–56.
- Corfittsen, M.T., 1993. Tiredness and visual reaction time among nighttime cab drivers: a roadside survey. *Acc. Anal. Prev.* 25 (6), 667.
- Dalziel, J.R., Job, R.F.S., 1997. Motor vehicle accidents, fatigue and optimism bias in taxi drivers. *Acc. Anal. Prev.* 29 (4), 489–494.
- Fernandez, W.G., Park, J.L., Olshaker, J., 2005. An observational study of safety belt use among taxi drivers in Boston. *Ann. Emerg. Med.* 45 (6), 626–629.
- Fielding, J., Mullen, P., Brownson, R., Fullilove, M., Guerra, F., Hinman, A., et al., 2001. Recommendations to reduce injuries to motor vehicle occupants: increasing child safety seat use, increasing safety belt use, and reducing alcohol-impaired driving. *Am. J. Prev. Med.* 21 (4), 16–22.
- Finn, P., Bragg, B.W.E., 1986. Perception of the risk of an accident by young and older drivers. *Acc. Anal. Prev.* 18 (4), 289–298.
- Ho, T.H., Chong, J.K., Xia, X., 2017. Yellow taxis have fewer accidents than blue taxis because yellow is more visible than blue. *Proc. Natl. Acad. Sci. USA* 114 (12), 3074–3078.
- Johnson, N.J., Sorlie, P.D., Backlund, E., 1999. The impact of specific occupation on mortality in the US National Longitudinal Mortality Study. *Demography* 36, 355–367.
- Karpf, R.S., Williams, A.F., 1983. Teenage drivers and motor vehicle deaths. *Acc. Anal. Prev.* 15 (1), 55–63.
- Lam, L.T., 2004. Environmental factors associated with crash-related mortality and injury among taxi drivers in New South Wales, Australia. *Acc. Anal. Prev.* 36 (5), 905–908.
- Matthews, M.L., Moran, A.R., 1986. Age differences in male drivers perception of accident risk: the role of perceived driving ability. *Acc. Anal. Prev.* 18 (4), 299–313.
- Tseng, C.M., 2013. Operating styles, working time and daily driving distance in relation to a taxi driver's speeding offenses in Taiwan. *Acc. Anal. Prev.* 52, 1–8.

Further Reading

- De Blasio, B., 2016. For-hire vehicle transportation study. Available from: <https://www1.nyc.gov/assets/operations/downloads/pdf/For-Hire-Vehicle-Transportation-Study.pdf>.
- Öz, B., Özkan, T., Lajunen, T., 2010. Professional and non-professional drivers stress reactions and risky driving. *Transp. Res. Part F Traffic Psychol. Behav.* 13 (1), 32–40.
- Thompson, J.P., Baldock, M.R.J., Dutschke, J.K., 2018. Trends in the crash involvement of older drivers in Australia. *Acc. Anal. Prev.* 117, 262–269.

Traffic Signals and Safety

Andrew P. Tarko, Purdue University, Lyles School of Civil Engineering, West Lafayette, IN, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	706
Red Signal Violation	706
Queue Presence	707
Pavement Friction	707
Permissive Left Turns	707
Permissive Right Turn on Red	708
Signal Design and Settings	709
New Signal Installations	709
Signals Visibility and Readability	709
Pedestrians at Signalized Intersections	710
Bicyclists at Signalized Intersections	711
Signalized Crosswalks	711
Signal Preemption	711
Signals During Low-Volume Periods	712
Eliminating Traffic Signals	712
Further Reading	712

Introduction

Traffic signals are installed at road intersections to improve their capacity or safety or both. Conditions for which traffic signals are justified as potentially beneficial are known as traffic signal warrants. In the United States, there are currently nine signal warrants, four of them relate to vehicle and pedestrian traffic volume, and another four consider signal coordination, road network, and the presence of schools and railway crossings. The remaining safety-focused warrant allows signalization if:

- other means of safety improvement are exhausted;
- at least five police-reportable crashes susceptible to correction by traffic signal control occurred within a 12-month period; and
- the threshold of at least one traffic volume warrant is met by at least 80%.

It is apparent that the use of traffic signals as a means of safety improvement is warranted only under quite restrictive conditions.

Traffic signals allow certain movements across an intersection while holding other movements at a stop. Each of the alternating periods contains a set of allowed movements, and is called the signal phase (Fig. 1). In the United States, signals go from green to yellow to red and back to green. In many other countries, there is also a period showing red plus yellow simultaneously to indicate that the green light will soon be displayed. The signal's safety benefit is affected by the absence of serious conflicts among movements allowed in each phase and by separating consecutive phases with sufficiently long clearance periods. A clearance period consists of a yellow signal that warns about an upcoming red signal and an all-red period when no traffic from any direction may enter the intersection (Fig. 1). The yellow signal must be sufficiently long to avoid what is known as a dilemma zone, where an approaching driver cannot stop before the intersection and cannot enter the intersection before the yellow signal turns red. Fig. 2 presents this zone as an overlap of two sections marked with gray bars: (1) the section where the vehicle is too far to reach the intersection during the yellow signal, and (2) the section where the vehicle is too close to stop before the intersection. Although traffic signals tend to reduce the number of right-angle collisions, traffic interruptions introduced by red signals expose vehicles to rear-end collisions. Yellow signals that are too short may cause both additional rear-end collisions due to rapid braking and right-angle collisions due to voluntary or unavoidable violations of red signals.

An intriguing dilemma zone control is being used in Sweden called LHOVRA. It uses a sequence of detectors placed along the approach to an intersection. The detectors determine vehicle type and speed and the change period is adjusted to minimize the dilemma zone occurrence. It may not be applied in countries where the signal standards require fixed yellow and all-red intervals.

Red Signal Violation

Drivers may violate red signals intentionally to avoid stopping and being delayed. Enforcement by police or red-light-running cameras may reduce both red signal violations and associated right-angle collisions. A flashing advance warning signal is installed at some intersections. Its flashing light and message, "Prepare to Stop When Flashing," reduces unintentional red-light running by

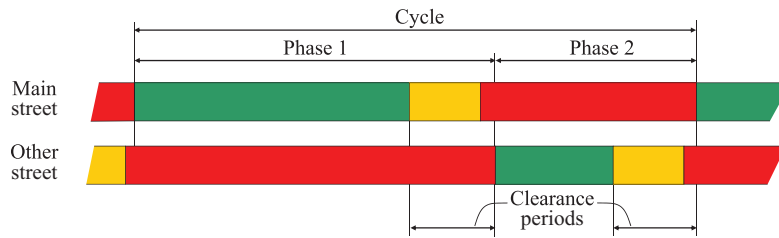


Figure 1 Signal phases and clearance periods.

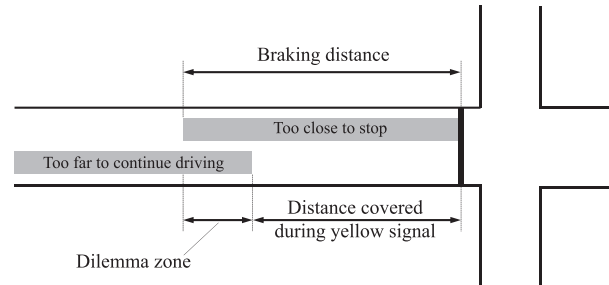


Figure 2 Dilemma zone on an approach to a signalized intersection.

warning that a yellow signal is coming soon, and will be followed by a red signal. The device also helps protect against the dilemma zone and alerts motorists where the sight distance is limited.

Long red signals may cause uncertainty, among drivers already stopped and waiting, about the proper functioning of traffic signals, particularly if no traffic movements occur from other approaches. This uncertainty may lead to a driver entering the intersection before the red signal ends. Long red signals may also increase signal violation by regular users of the roadway, who may elect to move through the intersection on a yellow signal to avoid what they anticipate to be a long waiting time.

Queue Presence

Research indicates that rear-end collisions may be associated with long queues on arterial streets with signal coordination. Clearing a queue takes time during the green signal, and during that time stopped vehicles block lanes to groups of vehicles arriving from upstream intersections. The presence of stopped vehicles may be difficult for these approaching drivers to see if they are following other vehicles closely. Forced to brake rapidly, they cause the additional risk of rear-end collision with vehicles behind them. Signal coordination reduces the risk of rear-end collision if it schedules the arrival of upstream vehicles after queues are cleared at the intersection they are approaching.

Missing or short turning bays are another potential cause of rear-end collisions because vehicles waiting to turn block the through traffic with whom they share the lane. Although designing longer turning bays or extending existing bays are effective countermeasures, shortage of space or excessive costs may prohibit these solutions. Where capacity is not a problem and recorded collisions confirm a short turning bay problem, a viable solution may be to control queue length in the turning bay by recalling a green signal. The green signal is recalled if a detector placed at the end of the turning bay indicates the continuous presence of vehicles.

Pavement Friction

The ability to effectively brake is a fundamental condition of successfully avoiding a rear-end collision. Approaches to signalized intersections may experience drivers' frequent braking maneuvers that involve locking brakes, causing tires to slide on the pavement. Over time, the mineral seeds in the pavement, particularly bituminous pavement, are exposed and polished. The coefficient of friction can be quite low, which considerably reduces driver acceleration rate. Proper treatment of the damaged pavement increases the friction coefficient even above the original friction of new pavement.

Permissive Left Turns

Left-turn collisions occur between vehicles turning left and vehicles moving through from the opposite approach. These collisions can be eliminated or reduced by letting left-turn vehicles move while displaying red signal to through vehicles (the protected left

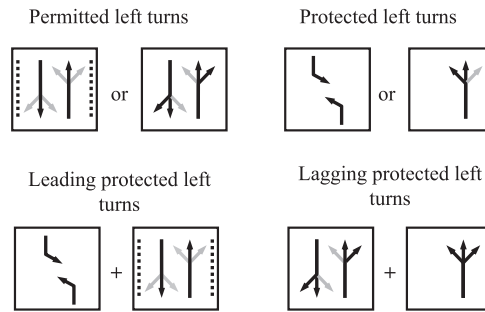


Figure 3 Permissive (or permitted) and protected left turns.

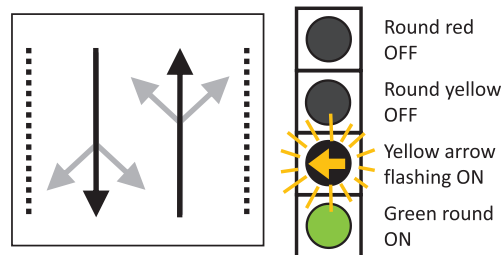


Figure 4 Flashing left arrow to emphasize that the left turn is permitted but not protected.

turns case in [Fig. 3](#)). This solution is not always applied, however, and instead, opposing traffic movements are presented with circular green signals simultaneously. In such cases, drivers turning left must wait for sufficiently long gaps in the oncoming traffic. The gray left-turn arrows in [Fig. 3](#) indicate this situation. Left-turn collisions are caused by drivers turning left who misperceive a gap in opposing traffic when road speed is high, sight distance of oncoming traffic obscured by vehicles moving in multiple lanes, and/or by a vertical curve on the opposite approach. These conditions, which are exacerbated when left-turning vehicles are permitted to turn left from two or more lanes, motivate protection of left-turn movements. An additional potential problem is drivers misunderstanding the green circular signal and assuming that they may safely turn left. An improvement related particularly to that misunderstanding warns drivers about permissive left turns by implementation of a flashing yellow left arrow presented when left-turning vehicles are not protected and must yield to the opposite through traffic ([Fig. 4](#)).

Restricting left-turn movements are considered when an intersection experiences an excessive frequency of left-turn collisions, or a shortage of capacity with its consequential risk of rear-end crashes. Restricting movement is viable as long as alternative routes for left-turning vehicles are available. A more expensive solution is to remove the left-turn through movement conflict from the intersection by adding signalized crossovers before an approach to the intersection that would permit travel on the opposite side of the road, and then divert it back to the original side once traffic has passed the intersection. Various forms of this solution include continuous flow intersections, signalized upstream crossovers, and diverging diamond interchanges ([Fig. 5](#)). The capacity benefits of such treatment have been confirmed, but the safety effect is uncertain. The addition of traffic signals at crossovers and potentially confusing movement paths are among the safety-related drawbacks.

Permissive Right Turn on Red

The permissive right turn on red (RTOR) signal was introduced in 1925 at signalized intersections in Los Angeles, California. In the 1970s, RTOR spread across the United States at signalized intersections, and then later to Canada, as a means of saving time and fuel. The maneuver was considered safe for its simplicity. After passing the crosswalk on approaching an intersection, a driver could proceed if there were no vehicle in the right-most lane of the intersecting road. The RTOR provision was common at signalized intersections until safety studies indicated a heightened risk of collision with pedestrians on the two crosswalks a driver had to pass. Motorists focused on vehicle traffic approaching on their left and did not pay sufficient attention to pedestrians on their right. In addition, attaining a position convenient for examining gaps on a major road before accepting them required the driver to pull forward, thus blocking the crosswalk on the current approach if the crosswalk did not offset a sufficient distance from the curb line. Research studies confirmed occurrence of collisions caused by the permissive RTOR.



Figure 5 Diverging diamond interchange.

Signal Design and Settings

Traffic signal design includes selecting type and sequence of signal phases and designing change periods. The length of the green signal must be determined in advance if the signal is to be controlled by clock. If signals are to be actuated by a signal controller, signal settings need to be decided. Additional design decisions include coordination with adjacent signals, in which more signal settings are added to establish a coordination plan. A coordination plan imposes constraints on local signal actuation to maintain smooth movements of vehicles along a sequence of coordinated signalized intersections.

Faulty signal design may affect safety at signalized intersections. Faulty design includes too short yellow signals, too long red signals, permissive left-turn phases in nonconductive conditions, and poor signal coordination that sends vehicles into the back of a queue still present at the downstream intersection. Poor coordination resulting in long queues and long waiting times may increase the violation of red signals and lead to additional collisions. A similar effect may result from premature termination of green signals by incorrect settings. The resultant phase failures force drivers to wait through another red signal, increasing their propensity to violate the signal. Another potential safety problem may be produced by inconsistent use of leading and lagging left-turn phases. In such cases, the through traffic on opposing approaches to an intersection does not receive green signals at the same time. Local drivers, who become accustomed to the order at which each approach receives a green signal, may become confused if that order is changed at different intersections in the same area, or even worse, at an intersection during different times of the day. Confused left-turning drivers may inadvertently violate the right way of through traffic from the opposite approach. Consistency in signal and geometric design of intersections increases drivers' expectancy, which promotes compliant behavior and enhanced performance.

New Signal Installations

New traffic signal installations should be supplemented with signs posting warnings about the new signals. Otherwise, local drivers may be surprised by the signals and the presence of stopped vehicles along their travel route. Where possible, stops signs should be installed on all approaches of the intersection in the period preceding traffic signals in order to prepare local drivers for the upcoming traffic interruption. Permanent warning signs may be warranted where the traffic signals' visibility is reduced by curves or other obstructions.

Signals Visibility and Readability

Drivers cannot follow traffic signals if these signals are not seen or not understood. The importance of clear and non-confusing signals for left-turning drivers in the United States has already been discussed. But, just crossing the border into Canada, drivers are faced with flashing green lights. In Canada this means the same as a green arrow, that they have an exclusive left turn. However, in many other countries, a flashing circular green signal means that the green will soon turn yellow. A general requirement for signals is their visibility and clarity to road users approaching an intersection. The first moment of importance is when the intersection first comes into view. Signal heads and their indications should be seen in advance to help road users notice the presence of traffic signals

and to approach the intersection in the proper lane. Information redundancy is important, and can be accomplished with properly aligned pavement edges and markings, road signs, and traffic signals. Enhancement of traffic signals' visibility and clarity is accomplished by placing the signals on the right-hand side and repeating them if needed on the left-hand side in medians, divisional islands, or elongated separators. The conditions that prompt repeating signals on the left-hand side include multiple lanes, the presence of tall vehicles in traffic, and the need to display different signals for different movements. An overhead position of signal heads is recommended under certain conditions, which include high speed, multiple lanes, the presence of tall vehicles in traffic, and lane-based traffic control.

Traffic signal visibility can be improved by (1) replacing 8-inch (200 mm) lenses with 12-inch (300 mm), (2) installing shades to eliminate direct sunlight on signal lenses, and/or (3) adding backplates with retroreflective borders. It should be obvious to approaching road users which signals apply to them. This communication is accomplished in most countries by (1) using turn signals (arrow silhouettes or cutoffs placed on circular lenses) to indicate the traffic movements being controlled by individual signal heads, (2) reserving solid round signal lenses for all turns including solid green for permissive left and right turns, (3) emphasizing the permissive left turn by adding a flashing left-turn signal to the round green signal, (4) reserving solid arrow signals for individual movements including protected left and right turns, (5) installing signal heads for individual lanes above the corresponding lanes, and (6) adding text or symbols for clarification, if needed.

Drivers seeing signals addressed to other road users may add confusion or uncertainty, so should be avoided. Reducing the visibility of signals addressed to others can be accomplished by (1) properly directing signal heads toward locations with drivers who are intended to see the signal, (2) using protruding shades (screens, covers, etc.) that hide the signal from certain angles, and (3) using polarized lenses that allow viewing the signal light only from a sufficiently close position. A separate precaution is proper maintenance of signal heads and trimming vegetation in their vicinity. Overgrown trees may partly or completely block the view of signal heads.

Pedestrians at Signalized Intersections

Pedestrians are present at most signalized intersections in urban areas. Pedestrians are vulnerable road users and their safety at signalized intersections must be carefully considered when designing and operating traffic signals. Due to the quite different character of vehicles and pedestrians, these two groups are treated differently. Pedestrian are not subject to formal instructions and training as drivers are, so their signals must be easy to interpret. Pedestrian *Walk* signals shaped as walking human silhouette and *Do not walk* signals shaped as upraised hand or standing silhouette convey a clear meaning and help distinguish them from vehicle signals. *Do not walk* signal flashes during the pedestrian change period and is solid in the vehicle phase. Unlike vehicles, pedestrians require neither considerable distance nor time to slow down and stop. For this reason, they do not require an additional signal, equivalent to the yellow signal for vehicles, to warn them about the shortly appearing *Do not walk* signal.

Similarly to vehicles, pedestrians require time to clear the roadway after their walk signal ends. The clearance time should allow pedestrians to cross the entire length of the crosswalk to an area where they are safely separated from moving vehicles. The time is determined by assuming a slightly higher walking speed than the recommended 4 ft/s (1.2 m/s) pedestrian speed for crosswalk design; this speed may be reduced where elderly, children, or strollers are expected at a higher rate than usual. Although recommended standards may allow a clearance time sufficient to bring pedestrians from a curb to a pedestrian refuge island, prudence encourages making this time long enough to let pedestrians walk completely to the other side of the roadway. Crossing a roadway in two phases is justified only if the dividing median is wide enough to discourage pedestrians from attempting to continue past the end of the walk signal.

Violation of the *Do not walk* signal, particularly first few seconds, by pedestrians is quite frequent. One method of reducing this behavior is displaying a countdown of time remaining to the end of the pedestrian change period when green signals for conflicting vehicles begin. Such information may discourage pedestrians from starting crossing too late to finish safely.

Where pedestrian safety is of concern, such as in downtown areas with high pedestrian traffic and on crosswalks of high-speed arterial roads, the walk signal should start as soon as possible after a pedestrian calls for a *Walk* signal if the phase for conflicting vehicles is not needed and may be safely terminated.

At many intersections, vehicles turning right are permitted to pass over an adjacent crosswalk during the walk signal after yielding to pedestrians. This quite common solution provides a green signal to right-turning vehicles at the same time the walk signal is displayed to pedestrians. The solution is acceptable if right-turn vehicles do not experience long queues caused by pedestrians blocking their movement. Otherwise, aggressive driver behavior resulting from long delays may expose pedestrians to an excessive risk of collision with a turning vehicle. The intersection corner negotiated by right-turning drivers must also be designed to be sufficiently sharp to slow drivers down and protect walking pedestrians. This measure is particularly important if a crosswalk is offset from the intersection and vehicles have a distance to accelerate before reaching it. On the other hand, if a crosswalk is immediately adjacent to the intersection, a turning vehicle may arrive at the crosswalk at the same time that pedestrians take their first steps onto the pavement. To avoid this potentially dangerous situation, the walk signal for pedestrians should start 2–3 s ahead of the vehicles' green signal to establish pedestrian presence on the crosswalk before arrival of the first vehicle. The visibility of pedestrians to drivers is of paramount importance. The corner should be clear of any sight obstructions such as bushes, kiosks, and advertising columns. Street lighting should provide sufficient illumination of any waiting and crossing pedestrians during dark hours. Separating a pedestrian signal phase from a signal phase for turning vehicles is a

recommendable solution where the vehicle–pedestrian conflict is considerable. The increased wait time for both parties makes signal violation more probable though.

Bicyclists at Signalized Intersections

Bicyclists are another group of vulnerable road users. One of the successful improvements for these road users are bicycle boxes and two-phase crossings for bicycles' left-turn movements. They typically require push buttons at actuated signals to call their signal if no motorized vehicles moving in the same direction are present. Otherwise, violating a red signal may be the only way for cyclists to avoid long waits or walking with the bicycle across the crosswalk. Also, if intersections do not have separate signal displays for bicyclists, the clearance phase for vehicle traffic should be designed for bicycle traffic and speeds around 12 mph (20 km/h) or lower where justified.

Signalized Crosswalks

Pedestrian and bicycle crossings on arterial roads may require traffic signals to mitigate the conflict between vulnerable road users and vehicular flows. Crosswalk and bicycle crossing signal design deserves the same, or, even higher, level of attention as vehicular signal design. To minimize signal violations by pedestrians, the best method is to provide walk signals as frequent and for as long as practical. The best signal coordination plan, from the point of view of pedestrians and bicycles, promotes vehicle groups from opposite directions to arrive at the crosswalk simultaneously in order to open as large a time window as possible for the pedestrians and cyclists to cross the road. If push buttons are provided to recall a signal, waiting for the walk signal should not be longer than the time needed for vehicles to clear the crosswalks, unless the call was placed in the part of the cycle reserved for the coordinated vehicular phase.

At signalized crosswalks without signal coordination, a fully actuated signal with push buttons for pedestrians and bicycles, and detectors for vehicles, allows signals to be set adequate to road user needs and priorities, as prompted by traffic volume and crash records. Particularly important for pedestrian and bicycle safety on high-speed arterials is prevention of the signalized crossing dilemma zone sometimes faced by arriving motorists. Protecting the safety of children crossing roads in school zone areas is of particular concern. It has been demonstrated that signalization of crosswalks in school zone areas may be an ineffective solution as children tend to violate signals while drivers may pay less attention at signalized than at unsignalized crosswalks.

Pedestrian Light CONtrolled (PELICON, then PELICAN) crossing was introduced in the United Kingdom, in 1969. It was the first pedestrian crossing located between intersections and protected with traffic signals. A similar solution, High-intensity Activated crossWALK (HAWK) is a type of crossing introduced in the United States. The HAWK beacons consist of two red lights on top and a yellow light on bottom for vehicles. The signals for vehicles are inactivate until a person presses a button on the signal pole and the yellow signal begins to flash. The flashing yellow turns to solid to prepare drivers to stop. The driver has to stop once both top lights turn solid red and pedestrians receive *Walk* signal. Pedestrians see standard *Walk* and *Do Not Walk* with optional time countdown. Drivers may pass the intersection when the two solid red signals turn to flashing and pedestrians have ended crossing for improving the safety of pedestrian crossings at locations where pedestrian volumes do not meet Manual on Uniform Traffic Control Devices (MUTCD) warrants for conventional traffic signals.

Protecting pedestrians and bicyclists from serious injuries and death on signalized crossing by reducing vehicle speeds of vehicles include raising the crossing area (Sweden) or painting zigzag lines on the approaches to the crosswalks (Great Britain). Other precautions to protect pedestrians and bicyclists include street lighting focused on the crosswalk, special signage to bring driver attention to the crosswalk, separation islands for pedestrians and bicyclists, speed limits, and display to pedestrians and bicyclists of the time countdown to the next walk signal.

Signal Preemption

Traffic signal preemption allows emergency vehicles, public transit vehicles, trains, and other vehicle types pass a signalized intersection or crossing without stopping or with minimum delay. Although signal preemption may be seen as a control intervention that disrupts regular operations, it is needed to provide priority to emergency and transit vehicles and to reduce the risk of collision.

Signal controllers are equipped with special devices that can receive a request from authorized users to activate preemption. Upon receiving a preemption request from an emergency vehicle, the preemption system overrides regular operations of the signal controller, forcing termination of the current signal phase and, after required clearance time, calling a signal phase that protects conflict-free passing through the intersection by the emergency vehicle. A similar operation is executed at traffic signals located near a railway crossing. To permit a safe train's passage, green signals that help clear the area and red signals that prevent motion of vehicles toward the railway crossing are called or extended. These signals are maintained from the period preceding the train arrival through the train passing the railway crossing. A preemption request is frequently sent automatically when emergency vehicles or trains are at predetermined distances from the required location in order to allow sufficient time for effective preemptive action.

Traffic signal preemption may include confirmation of the preemption to the requesting driver and a warning signal to other drivers. The warning signal is usually a single light placed near the intersection that is visible to all road users around the intersection, or it might be several directional signals aimed toward approaching traffic streams. The signals may flash or be continuous. A variety of solutions have been tested to inform local drivers where the emergency unit is coming from and which direction may need to be cleared. Since emergency preemptions at many locations are infrequent, it has been difficult to develop and disseminate among road users a general awareness of the notifying signals and knowledge of their meaning. Current and emerging in-vehicle technologies, however, bring an opportunity for transmitting emergency notifications directly to vehicles potentially affected by the preemption.

Preemption for prioritizing public transit is less intrusive as its main purpose is to reduce delays rather than improve safety.

Signals During Low-Volume Periods

Traffic control signals in the flashing mode may be a beneficial form of traffic control that increases driver awareness of an intersection's presence without interrupting traffic nor causing additional delays. For these reasons, traffic signals are often operated in flashing mode, with yellow flash on the major road and red flash on the minor road, during periods of low traffic volume during late-night and early-morning hours at intersections with adequate sight distance. Unfortunately, before/after research studies indicate a near doubling of crash rates when flashing signals of this type are implemented during these periods, as compared to the ordinary three-color operation. An alternative strategy is to make the intersection all-way stop during low-volume times with red flash on all approaches.

Eliminating Traffic Signals

Unwarranted installation of traffic signals at intersections may increase delays and reduce safety. A similar effect is also expected at signalized intersections when traffic signals are no longer justified after traffic volume has dropped below the level that would warrant the signals. This situation is possible if an alternative road has been opened that offers a more convenient connection with a major traffic attraction such as a nearby shopping mall, or when an attraction is closed. Elimination of signalization often requires redesigning the roadway geometry. This is a good opportunity to build a roundabout—one of the safest solutions for moderate or low-volume traffic.

Further Reading

- Antonucci, N.D., Hardy, K.K., Slack, K.L., Pfefer, R., Neuman, T.R., 2004. Guidance for implementation of the AASHTO Strategic Highway Safety Plan—volume 12: a guide for reducing collisions at signalized intersections. Transportation Research Board, Washington, DC. Available from: <http://nap.edu/23423>.
- Chandler, B.E., Myers, M.C., Atkinson, J.E., Bryer, T.E., Retting, R., Smithline, J., Trim, J., Wojtkiewicz, P., Thomas, G.B., Venglar, S.P., Sunkari, S., Malone, B.J., Izadpanah, P., 2013. Signalized intersections informational guide, second ed. FHWA-SA-13-027. Federal Highway Administration, Office of Safety, Washington, DC. Available from: <https://trid.trb.org/view/1267799>.
- Federal Highway Administration, 2009. Manual of uniform traffic control devices for streets and highways. U.S. Department of Transportation. Available from: <https://mutcd.fhwa.dot.gov/index.htm>.
- FHWA Office of Safety, 2009. Low-cost safety enhancements for stop-controlled and signalized intersections. FHWA-SA-09-020. U.S. Department of Transportation, Federal Highway Administration. Available from: https://safety.fhwa.dot.gov/intersection/other_topics/fhwasa09020/.
- FHWA Office of Safety, 2010. Removal of signal flashing mode during late-night/early-morning operation. FHWA-SA-09-012. U.S. Department of Transportation, Federal Highway Administration. Available from: https://safety.fhwa.dot.gov/intersection/conventional/signalized/case_studies/fhwasa09012/.
- Nevers, B. et al., 2011. Assessment of auxiliary through lanes at signalized intersections. National Academies of Sciences, Engineering, and Medicine, The National Academies Press, Washington, DC. Available from: <https://doi.org/10.17226/22830>.
- Noyce, D.A., Bill, A.R., Knodler Jr., M.A., 2014. Evaluation of the flashing yellow arrow (FYA) permissive left-turn in shared yellow signal sections. National Academies of Sciences, Engineering, and Medicine, The National Academies Press, Washington, DC. Available from: <http://nap.edu/22246>.
- Rodegerdts, L.A., Nevers, B., Robinson, B., Ringert, J., Koonce, P., Bansen, J., Nguyen, T., McGill, J., Stewart, D., Suggett, J., Neuman, T., Antonucci, N., Hardy, K., Courage, K., 2004. Signalized intersections: informational guide. FHWA-HRT-04-091. Federal Highway Administration, Turner-Fairbank Highway Research Center, McLean, VA. Available from: <https://www.fhwa.dot.gov/publications/research/safety/04091/04091.pdf>.
- Srinivasan, R., Baek, J., Smith, S., Sundstrom, C., Carter, D., Lyon, C., Persaud, B., Gross, F., Eccles, K., Hamidi, A., Lefler, N., 2011. Evaluation of safety strategies at signalized intersections. NCHRP Report 705. National Cooperative Highway Research Program. Available from: <http://nap.edu/14573>.

Tunnels, Safety, and Security Issues—Risk Assessment for Road Tunnels: State-of-the-Art Practices and Challenges

Konstantinos Kirytopoulos^{*,†}, Panagiotis Ntzeremes[†], Konstantinos Kazaras[†], ^{*}School of Natural & Built Environments, University of South Australia, Adelaide, SA, Australia; [†]National Technical University of Athens, Athens, Greece

© 2021 Elsevier Ltd. All rights reserved.

Introduction	713
Setting the Scene: Accidents and Fires in Tunnels	713
Fire Safety in Road Tunnels: Fundamentals of the Risk-Based Approach	715
System Description and Theoretical Framework	715
The Scenario-Based Approach	716
The Case of Vehicles Carrying Dangerous Goods	716
Challenges in Current Practice	716
The Human Factor	716
Organization and System Degradation	717
Deviating Provisions	717
Recommendations	717
Relevant Websites	718
References	718

Introduction

Tunnels constitute a significant building block for supporting an adequate and intelligent road network. There are different types of tunnels, such as road, rail, pedestrian, or bicycle tunnels. The latter are used to increase safety of pedestrians and cyclists who are trying to cross busy roads. However, such tunnels can create risks if they are poorly designed, not well lit or used improperly. On the contrary, rail tunnels also pose critical risks but, since there is no other “traffic” and train drivers do not interact with one another, the safety of rail tunnels is probably higher than that of road tunnels. This paper focuses on road tunnels and specifically examines the current best practices and challenges for risk assessment.

In general, road tunnels are divided into urban and rural depending on their different operational and construction characteristics. Urban tunnels are developed to improve transportation flow within cities. This includes the improvement of environmental quality of densely populated areas by easing traffic congestion and pollution. Thus, they are usually formed as underground constructions. On the contrary, rural tunnels are developed in order to reduce travel distances and increase safety by creating shortcuts in mountainous areas. Their importance is related to the fact that they minimize the environmental impact while reducing travel time and cost (Kirytopoulos et al., 2017). As a result of the significant improvement of underground construction-related technology in recent years, tunnels are nowadays a cost-effective engineering solution within road infrastructure. As a consequence, their numbers are increasing as more and more tunnels are developed, specifically in Asian countries due to the economic growth but also in Europe as part of the infrastructure enhancement, that, in turn, increases the number of people and the volume of goods transported through them (Ntzeremes and Kirytopoulos, 2019).

Despite the numerous advantages, tunnels bear an internal issue, which is the extreme impact of potential accidents. Undoubtedly, fire accidents are the biggest threat for tunnel systems since they can result in enormously adverse impacts regarding human losses and destruction of the tunnel’s equipment and infrastructure (Li and Ingason, 2018). The devastating experience from fire accidents in the past, such as Mont Blanc fire in France/Italy (1999; 39 fatalities) or the fire in Yanhou in China (2014; 31 fatalities), is indicative of the severity of such accidents. On top, such accidents will also bring significant economic losses and costs relevant to the rehabilitation of the infrastructure. The rapid growth of tunnels implies that more fire accidents may occur in the future, thus the development of efficient safety strategies is of utmost importance. Because of that, authorities and safety analysts around the world work on sophisticated, more effective safety strategies based on a risk-based approach in the tunnel area (Li and Ingason, 2018; NCHRP, 2011; PIARC, 2013).

Setting the Scene: Accidents and Fires in Tunnels

Tunnels, together with bridges, are the most advanced components of the road network. A quick look at tunnel facilities and equipment justifies this description. Nowadays, most long tunnels are continuously monitored through a control center. Complex surveillance systems for monitoring tunnel areas include systems using motion pattern sensors for detecting potential fire accidents

Table 1 Indicative road accident rates

Country	Accident rate in road tunnels	Accident rate on open road
Austria (data years 1999–2003)	10	14
France (data years <1998)	9	7
Norway (data years 2001–06)	12	>12
Greece (data years 2011–13)	4	6
Italy (data years 2006–09)	15	9

Data: Synthesis of multiple sources—values are approximate.

Accident rates per 10^8 vehicle-kilometres.

Note: The data are comparable only between road tunnels and open road per country and *not* across countries as the measurements are not always compatible since they may follow different protocols and different years.

Sources: Amundsen, F.H., 2009. *Studies on Norwegian Road Tunnels II. An Analysis on Traffic Accidents in Road Tunnels 2001–2006*. Roads and Traffic Department, Traffic Safety Section, Oslo.

Caliendo, C., De Guglielmo, M.L., 2012. Accident rates in road tunnels and social cost evaluation. *Procedia Soc. Behav. Sci.* 53, 166–177.

Nussbaumer, C., 2007. Comparative analysis of safety in tunnels. Young Researchers Seminar. Austrian Road Safety Board, Brno, pp. 1–9.

Tsantanosoglou, A., 2013. Egnatia Odos tunnel safety indices, Report. Available from: Egnatia Odos observatory.

and systems, such as the last generation devices, for smoke and temperature detection. These systems provide direct information on the location of an incident or fire and its development in order to enable assessment and appropriate execution of response protocols. Furthermore, advanced systems for controlling the development and consequences of a fire event, such as mechanical ventilation and fixed firefighting systems, enable tunnel staff to manage the system remotely from a control center and direct emergency services to rescue trapped users successfully and prevent damage to tunnel structure and facilities.

In general, tunnels are regarded as safe road facilities (Nvestad and Meyer, 2014). There are plenty of reasons for the increased safety in tunnels including drivers being more vigilant but also the standardized environment. For instance, apart from extreme cases, tunnels are not affected by weather conditions, they usually lack junctions, walkers, parked vehicles, bicyclists, or distracting advertisement stands, which are common causes of accidents or factors that affect an accident's evolution. Additionally, studies have indicated that users drive more carefully when passing through tunnels. In view of that, tunnels are perceived to be safer than open roads (Kirytopoulos et al., 2017); however, official statistics are not absolutely conclusive on that. For instance, as depicted in Table 1, Austria, Greece, and Norway exhibit a lower accident rate in road tunnels than on open roads, France has about equal rates in the two systems, while Italy has higher accident rates in tunnels than on open roads. Also, there are studies mentioning that although the accident rate in tunnels is lower, the severity of the accidents is higher. In translating these data though, one should keep in mind two important issues. The first is that safety, specifically in road tunnels, is constantly improving as new technologies are embedded into the tunnel systems. The second is that the actual monitoring and recording of accidents in tunnels is such that every single incident is captured while this might not always be the case for open roads.

In particular, fire accident rates are much lower in tunnels than on open roads. Indicatively, statistics from France show that there are fewer than two vehicle fires per tunnel kilometer traveled per 100 million vehicles ($<2/10^8$ vehicle-kilometres) (Beard and Carvel, 2012). When it comes to heavy goods vehicles, the index is increased to about $8/10^8$ vehicle-kilometres. However, only $1/10^8$ vehicle-kilometres is serious enough to cause any damage to the tunnel itself. In absolute numbers, for Norway, a country with more than 1000 tunnels, and with a combined length over 1000 km, the average number of tunnel fires sparking from vehicles is approximately 20 per year (Nvestad and Meyer, 2014). Only some of those will result in any harm to people, but it is still a serious concern for users. China features highway tunnels in more than 15,000 locations with a total length above 14,000 km. A sum of 161 fire accidents has been recorded over the last 15 years in China (Ren et al., 2019). The results indicate that over half of them were due to vehicles' technical problems.

In-depth studies have revealed some general conclusions regarding accidents in tunnels (Beard and Carvel, 2012; Kirytopoulos et al., 2017; Nvestad and Meyer, 2014):

- It is evident that bidirectional tunnels have higher accident rates than unidirectional ones. Moreover, the response to a fire in a bidirectional tunnel is much harder since vehicles typically get trapped both upstream and downstream of the fire, thus the typical ventilation systems cannot be used (pushing the smoke to either side will put some users in danger).
- The terminal zones are more prone to accidents than the central zones. This is related mostly to the changing conditions/geometry/environment, the black hole effect at the entry portal and the glaring at the exit portal.
- Higher accident rates are observed in sections that affect the traffic flow (e.g., speed changes, variations in alignment, etc.).
- Most accidents are caused by rear-end collisions, poor maintenance of vehicles, and disobedience by drivers in keeping a safe distance to vehicles in front.
- The prevailing causes for fires in road tunnels are collisions. That includes both collisions between vehicles and between vehicles and the tunnel structure. But fires are also caused by electromechanical issues such as overheated brakes.
- Users' behavior after fire ignition is decisive for the outcome in terms of injuries or losses. Quick self-evacuation is regarded to be the most important factor in mitigating the consequences of an accident as far as users' survival is concerned.

The potential severity of tunnel fires is related to some special attributes of the tunnel. The particular environment in tunnels exhibits: (1) the absence of natural light, in which drivers need to adjust, (2) directed and almost constant air movement because of the different pressure between tunnel portals that complicates the operation of ventilation, (3) difficulties for emergency services and rescue crews to approach the crash site in case of an accident, and (4) irregularities in fire combustion (Beard and Carvel, 2012; Nvestad and Meyer, 2014).

The last attribute is closely related to the challenges of properly modeling the fire and thus deciding on the right equipment to be installed in tunnels. It is not easy to model the fire behavior in a single way as fire is influenced by the different characteristics of the burning materials as well as the ventilation (Li and Ingason, 2018). Additionally, tunnels are different to one another. Tunnels can differ by being unidirectional or bidirectional, and have different geometrical characteristics (e.g., length, width, etc.), different structure, or traffic conditions (e.g., average volume, whether dangerous goods transportation is allowed, etc.). These differences influence fire development and evolution and consequently the appropriate safety strategy (Ntzeremes and Kirytopoulos, 2019). The enclosed environment affects both the fire and the smoke development owing to the mutual interactions between physical and chemical processes. Due to these interactions, tunnel fires are characterized by turbulence, combustion irregularity, and high radiation (Beard and Carvel, 2012). There is also the possibility of a quick fire spear between stopped vehicles. A fire spear occurs when fire is transmitted quickly from one car to the other when the cars have not kept the appropriate safety distance while stopping upstream the fire one behind the other. Furthermore, the oxygen required for combustion is not always available in tunnels as in the open environment. As a result, large amounts of toxic fumes and products of incomplete combustion are emitted inside the tunnel. These conditions create a highly unsafe environment for trapped users who are affected by the smoke even in cases of unidirectional tunnels, especially during the initial stage of the fire (before the ventilation system manages to push the smoke downstream of the fire) as hot gases and smoke flow upstream of the fire location create the well-known backlayering phenomenon (Li and Ingason, 2018; NCHRP, 2011).

Fire Safety in Road Tunnels: Fundamentals of the Risk-Based Approach

System Description and Theoretical Framework

The disastrous fire events that led policy-makers and safety experts to reconsider road tunnel safety were the serious Alpine accidents of Mont Blanc—France/Italy, 1999 (39 losses), Tauern—Austria, 1999 (12 losses), and St. Gotthard—Switzerland, 2001 (11 losses). Those tragedies remind us that tunnels should be considered complex socio-technical systems and that their safety is a very complex issue (PIARC, 2008, 2013).

A system consisting of technical, organizational, and social elements is a socio-technical one. Indeed, road tunnels involve human, organizational, and technical elements, and hence they can be understood as socio-technical systems (Kazaras and Kirytopoulos, 2013). They are complex systems since they can be influenced by the setting (environment) of operations and they also influence this setting in return. Road tunnels are open to transport and influence the area in which they are located by the transportation of people and goods. If a major tunnel accident occurs, this may significantly affect the nearby region. In this sense, tunnels can be regarded as complex socio-technical systems. The main issue with such systems, especially complex ones, is that their overall safety is dependent on many characteristics coming from the interaction of their components and not only by the isolated safety level of an individual component. In road tunnels, it is the complexity of the nonlinear interactions among the different elements of the system that characterize the overall system's safety. Modeling the complex tunnel system can, however, be simplified through grouping various system parameters into users, facilities, infrastructure, vehicles, and traffic characteristics (Ntzeremes and Kirytopoulos, 2019; PIARC, 2008).

Normative tunnel safety, especially in Europe, has changed drastically once the European Directive 2004/54/EC came into force (EC, 2004). Till then, tunnel systems were regarded as safe only if they were designed in line with the prescriptive regulatory requirements that each country had imposed. Given the deficiencies of prescriptive regulatory requirements to manage a complex socio-technical system, the new safety approach introduced in 2004 uses risk assessment along with adherence to normative requirements. This is the so-called risk-based approach, and it enables tunnel managers and safety analysts to better investigate the system and understand, assess, and manage any potential risks (Aven and Zio, 2014; PIARC, 2008). Hence, a proper safety approach toward tunnels does not count only on *learning from events* (which is the main source of prescriptive requirements), but also on *assessing the system proactively*, similar to roadway audits, as should be the case for any complex socio-technical system (Aven and Zio, 2014; Ntzeremes and Kirytopoulos, 2019). Current trends see the risk-based approach as a systemic approach including an in-depth examination of potential accidents as well as the observation of possible residual or concealed risks identified from past accidents.

Various risk assessment methods have been developed over the years. Many countries have developed their own methods. A brief report of such methods can be found in the World Road Association (formerly named PIARC) reports. There is also some work in methods developed from academia (INERIS, 2005; Kazaras and Kirytopoulos, 2013; Ntzeremes and Kirytopoulos, 2018). Despite which specific method is used, the general risk assessment framework consists of three discrete layers. After a top-to-bottom sequence, these layers are follows (Ntzeremes and Kirytopoulos, 2019; PIARC, 2008, 2013):

- Risk analysis, which forms the first layer of the framework. This layer consists of three sub-layers. The first sub-layer is the *system definition*. The better the tunnel system is defined, the better the goals of the analysis would be achieved. Subsequently, the next

sub-layer is the *hazard identification*. Hazard is defined as a *potential source of harm*. A combination of hazards can lead to the critical event such as fire, which can cause direct or indirect harm to tunnel users, damage to tunnel infrastructure and facilities, and paralyze the rest road network widely. The last sub-layer is the *risk estimation* that entails the measurement of risk according to probability of occurrence and relevant consequences.

- Risk evaluation, which focuses on determining whether estimated risks are acceptable or not. To do so, estimated risks are compared with certain predefined risk criteria.
- Risk treatment, which aims at mitigating or eliminating, if possible, nonacceptable risks. Through this process, a decision to treat risk is made by selecting additional to standard safety measures. In the end, risk analysis is repeated in order to reevaluate the new risk level of the system and examine whether the residual risk is acceptable.

The Scenario-Based Approach

Regardless of the particular features of the method being applied, each one enables safety analysts to mitigate both the relevant probabilities of occurrence and/or the expected consequences. However, the lack of a sufficient amount of accidents prevents the extraction of sufficient statistical information and thus the analysis tends to be consequences oriented (Ntzeremes et al., n.d.). Such an analysis is the scenario-based approach, in which the tunnel system is investigated under predetermined conditions. To do so, safety analysts rely on standardized fire scenarios. Although probabilities should play a role in determining the scenarios to be examined, the analysis itself of the scenarios considers only the impact, that is, expected losses associated with the occurrence of a possible critical event. Therefore, mitigating the fire risk is translated in mitigating primarily potential losses among tunnel users in case of fire. This approach considers the technical parameters and the fire scenarios and is completed with an evacuation model for the trapped users. Computational fluid dynamics is usually used to examine the evolution of the fire and the evacuation model reveals potential losses (Ronchi et al., 2012). Depending on the results, the analyst comes up with a number of potential additional measures in order to reduce the estimated risks.

The Case of Vehicles Carrying Dangerous Goods

Based on the Agreement of Dangerous goods in Roads (ADR), transported goods are separated between dangerous and non-dangerous goods. Following this division, risk assessment adopts a totally different approach depending on the kind of transported goods. Fire accidents with dangerous goods involvement are of significantly lower-frequency yet lead to much higher-consequence critical events that can cause multiple fatalities in a single event. Thus, risk here is recognized as societal, unlike risk deriving from fire accidents without dangerous goods involvement, which is recognized at an individual scale. Societal risks are the ones who can affect the general public and may result in multiple casualties (PIARC, 2013).

The most widely accepted risk assessment method used for dealing with accidents involving dangerous goods is the OECD/PIARC quantitative risk assessment model (QRAM) (INERIS, 2005). This model examines 13 predetermined scenarios concerning potential types/groupings of dangerous goods but in this case the probability of having such an accident is taken into consideration. Contrary to the scenario-based approach without the involvement of dangerous goods, this analysis includes historical data of relevant accident rates and potential consequences of such accidents. The outcome provides the relevant evaluation of their respective consequences in terms of fatalities (and/or injuries) and frequencies in the form of frequency/number of fatalities (F/N curves). The developed F/N curves are compared with predefined criteria and the level of the tunnel safety is estimated. Predefined criteria are either in the form of overall expected values or the As Low as Reasonable Practicable (ALARP) principle. The latter also considers the cost/benefit analysis by comparing the resources spent to reduce risk to the increase of the level of safety, making sure that the cost involved in reducing the risk further would not be grossly disproportionate to the benefit gained (Ntzeremes and Kiriopoulou, 2019).

Challenges in Current Practice

The Human Factor

Looking closer to the existing risk assessment methods, the quantitative approach is used, almost exclusively, in order to predict the fire events, examine the associated consequences, and estimate the overall level of safety of the system. This is mainly because the quantitative approach is easier to be measured against specific thresholds imposed by norms or legislation. However, the quantitative approach includes several challenges, the main of which is the challenge to accurately capture human behavior and specifically the modeling of either the trapped-users behavior or the performance of tunnel personnel in applying the response protocols (INERIS, 2005). Empirical studies of actual or intended users' behavior have revealed that the users ignore important safety rules for driving through or copying with emergency situations in tunnels (Fridolph et al., 2013; Kiriopoulou et al., 2017). Despite the efforts of tunnels' managers in educating the users, there is, still, plenty of work to be done.

Knowledge about users' behavior would enable the analysts to better estimate and improve road tunnels' level of safety. In general, the analysts need to estimate and take into account the actions that people take and the time it needs for these actions to be performed (including realization and implementation time) during evacuation (Ronchi et al., 2012). Although conducted studies

and full-scale experiments do provide some information regarding users' behavior, the available data about evacuation behavior and movement are still limited and relevant assumptions often raise criticism (INERIS, 2005). Frequently, oversimplified assumptions about the different stages of the evacuation process, the behavior of trapped users, and their speed in a smoke-filled tunnel are made during the use of relevant models. Therefore, the effectiveness of both risk analysis and risk evaluation can be challenged and the need for further research endeavors in this space is evident.

Because of these reasons, the idea of using qualitative methods, supplementary to the aforementioned quantitative ones, in a systemic way for analyzing road tunnels safety is gaining space.

Organization and System Degradation

A crucial part in the systemic analysis is the organizational aspects at the tunnel manager end. Based on the idea that road tunnels are complex socio-technical systems, it is essential to highlight that the organizational perspective (systems, culture, and procedures) significantly affects the overall safety of the tunnel system, since it sets the context in which the human element (both user and operator) takes decisions and the technical system operates.

The way tunnel facilities are organized and operating is different among countries. Nevertheless, there are some common aspects that have an effect on the level of safety. Such aspects include: (1) maintenance and inspection, (2) recruitment of operators, (3) training procedures, (4) preparation of emergency plans, (5) cooperation with the emergency services, and (6) analysis of past incidents and learning from events. In addition, the interface design of the tunnel control center and the systems available for detecting and controlling an event greatly affect the outcome of an emergency situation. All the aforementioned aspects are undeniably critical and the fact that they are seldom included in the analysis may hinder an accurate safety assessment process.

By following the system's theoretic perspective, the adaptation of the tunnel system over time needs to be considered as well (INERIS, 2005). Currently, during risk analysis it seems that there is no provision for differences that may occur between what was originally designed and the actual conditions after some time of operation. Lack of such a provision implies that systems are regarded to be fully functional and procedures followed always without any signs of degradation over time. This is probably a narrow perspective as variation or adaptation is an intrinsic characteristic of any system, especially systems including human and organizational components. The point is that degradation of systems does not happen accidentally but is attributed to specific factors that should be identified and then managed appropriately.

Deviating Provisions

Traditionally, safety analysts, after the evaluation of the estimated risk and if the risk is above the relevant thresholds, define a list of alternatives and make a decision that best meets the existing risk criteria. The use of absolute risk criteria means that the evaluation strategy requires well-established thresholds or risk targets being in line with the regulative requirements against which estimated risks are compared. Although the application of such criteria can be simple, the determination of such thresholds in order to define the acceptable level of risk has many difficulties. Also, thresholds need to be supported by reliable and up-to-date accident databases, otherwise they can lead to misleading results (Ntzeremes and Kirytopoulos, 2019; Ronchi et al., 2012).

One issue though is that not all risk assessment methods or regulations include the same reference conditions, which may lead to different results regarding the level of safety. That is, the same tunnel under different reference conditions and thresholds would appear as having a different level of safety and thus would require different additional measures. Because of that, a common approach through policies is suggested. It is to be acknowledged though that, in cases where cultural or other special characteristics (e.g., familiarity to tunnels or general road safety rules adherence in the country) dictate a different approach, this needs to be seriously taken into consideration and may allow for deviations from a common standard or policy.

Recommendations

Fire accidents are the biggest threat for tunnel systems since they may lead to severe impacts regarding human losses and destruction of tunnel's equipment and infrastructure. The devastating experience from fire accidents in the past is indicative of the severity of such accidents. Because tunnels are critical infrastructures of the road transportation system, potential tunnel fire accidents can be catastrophic. Thus, fire safety is a key issue for any tunnel project. Based on statistics, tunnels are not less safe than the rest of the road network, and fire accidents frequency particularly is much lower (Li and Ingason, 2018). However, the rapid growth of tunnels worldwide implies that more fire accidents may occur in the future. Thus, the development of efficient safety strategies is definitely needed but also a challenge. Despite the fact that a fire is probably the most impactful risk in a tunnel, it is far from being the only aspect of consideration and efforts should be put in reducing accident numbers and their severities in general.

Currently, legislation in most developed countries includes a periodical risk assessment of road tunnels (PIARC, 2008, 2013). Such assessment is usually done using a scenario-based approach or/and a probabilistic risk assessment when, for example, transportation of dangerous goods is examined. Both approaches are quantitative in nature and certain benchmarks are used in order to judge the safety level of a tunnel. Despite the benefits of such an approach, there are still some areas, crucial to safety of a socio-technical system such as a road tunnel, that a quantitative method cannot fully incorporate. Such areas include: the human factor, which relates to the behavior of the users and the tunnel personnel; the organization of tunnel safety, including training,

emergency plans, analysis of past incidents, and lessons learned; and the potential degradation of the system, that is the adverse changes that happen to the system reducing the overall level of safety over time (INERIS, 2005).

Another important issue is the current inconsistency between assumptions and thresholds mandated in different risk assessment methods. Different countries provide different assumptions for, say, the heat release rate or the speed of evacuation and, obviously, this results in inconsistent conclusions regarding the safety of the same tunnel, dependent on which method the analysts use. Worldwide organizations working with or responsible for road safety, such as PIARC, should (and do) significantly contribute toward the harmonization of relevant standards (Ntzeremes and Kirytopoulos, 2019).

Combining the benefits of the quantitative and qualitative analysis, creating common standards and adopting a holistic risk assessment view, will definitely contribute to safer tunnels and serve better the purposes of Vision Zero for safety in road infrastructure.

Relevant Websites

<https://tunnels.piarc.org/en/introduction/homepage>

References

- Aven, T., Zio, E., 2014. Foundational issues in risk analysis. *Risk Anal.* 34 (7), 1164–1172.
- Beard, A., Carvel, R., 2012. *Road Tunnel Fire Safety*. 2nd ed. Thomas Telford, London.
- EC, 2004. Minimum safety requirements for tunnels in the trans-European road network. Directive 2004/54/EC of the European Parliament and of the Council. European Union, Brussels.
- Fridolph, K., Nilsson, D., Frantzich, H., 2013. Fire evacuation in underground transportation systems: a review of accidents and empirical research. *Fire Technol.* 49 (2), 451–475.
- INERIS, 2005. *Transport of Dangerous Goods through Road Tunnels: Quantitative Risk Assessment Model (v. 3.61)*. Verneuil-en-Halatte, France.
- Kazaras, K., Kirytopoulos, K., 2013. Challenges for current quantitative risk assessment (QRA) models to describe explicitly the road tunnel safety level. *J. Risk Res.* 17 (8), 953–968.
- Kirytopoulos, K., Kazaras, K., Papapavlou, P., Ntzeremes, P., Tatsiopoulou, I., 2017. Exploring driving habits and safety critical behavioural intentions among road tunnel users: a questionnaire survey in Greece. *Tunn. Undergr. Space Technol.* 63, 244–251.
- Li, Y.Z., Ingason, H., 2018. Overview of research on fire safety in underground road and railway tunnels. *Tunn. Undergr. Space Technol.* 81, 568–589.
- Nævestad, T.O., Meyer, S., 2014. A survey of vehicle fires in Norwegian road tunnels 2008–2011. *Tunn. Undergr. Space Technol.* 41 (3), 104–112.
- NCHRP, 2011. *National Cooperative Highway Research Program: Design Fires in Road Tunnels*. Transportation Research Board, New York.
- Ntzeremes, P., Kirytopoulos, K., 2018. Applying a stochastic-based approach for developing a quantitative risk assessment method on the fire safety of underground road tunnels. *Tunn. Undergr. Space Technol.* 81, 619–631.
- Ntzeremes, P., Kirytopoulos, K., 2019. Evaluating the role of risk assessment for road tunnel fire safety: a comparative review within the EU. *J. Traffic Transp. Eng.* 6 (3), 282–296, [English Edition].
- Ntzeremes, P., Kirytopoulos, K., Filiou, G., n.d. Supporting quantitative risk assessment on road tunnel fire safety: an improved evacuation simulation model. *J. Risk Uncertainty Eng. Syst. Part A Civil Eng.*
- PIARC, 2008. *Risk Analysis for Road Tunnels*. World Road Association, Paris.
- PIARC, 2013. *Current Practice for Risk Evaluation for Road Tunnels*. World Road Association, Paris.
- Ren, R., Zhou, H., Hu, Z., He, S., Wang, X., 2019. Statistical analysis of fire accidents in Chinese highway tunnels 2000–2016. *Tunn. Undergr. Space Technol.* 83, 452–460.
- Ronchi, E., Colonna, P., Capote, J., Alvear, D., Berloco, N., Cuesta, A., 2012. The evaluation of different evacuation models for assessing road tunnel safety analysis. *Tunn. Undergr. Space Technol.* 30, 74–84.

Understanding, Managing, and Learning from Disruption

Karl Kim, Professor of Urban and Regional Planning, University of Hawaii, Honolulu, HI, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction: Disrupting Transportation Disruptions	719
Characterizing Disruption	719
Coping with and Mitigating Disruption	721
Individual Strategies	721
Agency and Organizational Strategies	721
Community Strategies	722
Key Lessons and Conclusions	723
Biography	724
References	724
Further Reading	725

Introduction: Disrupting Transportation Disruptions

There are different events and hazards that can disrupt transportation systems (Kim et al., 2018b). There are many different components of the transportation system—users, vehicles, facilities, roadways, equipment, networks, and technologies vulnerable to disruption by natural, human, technological, and combined forces. Efforts to cope with, minimize and mitigate the harm from disruptions can improve the resilience of transportation systems. Transportation resilience is the capacity to absorb shocks, to respond quickly and effectively to threats, and to recover functionality so that services and the movements of people and goods are not interrupted. A harm-causing event may result in the reduction or temporary shutdown of services, but a resilient transportation system rebounds and recovers quickly.

The extent of disruption can be measured in terms of the downtime, loss of service, reductions in capacity, travel volumes, speed, convenience, and increased costs of travel. The impacts of disruption include the number of travelers, trips, origins, and destinations affected, along with trip purposes as well as other metrics related to travel time. It is useful to assess the vulnerabilities of the transportation network (Murray-Tuite and Mahmassani, 2004). Disruptions affect both mobility and the ability to travel from one point to another and accessibility or the ability to use a particular mode of travel (walk, bike, auto, transit, paratransit, marine, air, etc.). Disruptions can especially impact those with limited access to transportation or special physical or health needs. Disruptions need to be seen in the larger social and economic context of travel behavior (National Cooperative Research Program, 2012).

Characterizing Disruption

The most common disruptions to the transportation system occur from weather events (heavy rainfall, flooding, landslides, storms, etc.), accidents, and large special events (sporting, political, conventions, etc.). Other events such as power blackouts, terrorism, and infrastructure failures (bridge collapses, derailments, airline crashes, structural defects, etc.) can also lead to traffic diversions, closed roadways, cancellation of services, and other impacts to the transportation system. Disruptions can also be caused by heavy traffic, congestion, and delay caused by insufficient capacity or labor disputes. Disruptions can result from internal factors related to the management and use of resources, factors of production, and operations or from external factors related to regulatory, political, or environmental changes that affect the availability and price of energy, movement and shipping of goods and people across borders and jurisdictions, or other factors such as conflict, terrorism, or changes in trade policy.

Hazards can be characterized by various spatial and temporal scales (see Fig. 1). Spatial and Temporal Dimension of Disruptive Hazards with Implications for Transportation Agencies shows common potential disruptions to transportation systems according to micro-, meso-, synoptic and planetary spatial and temporal scales. Hazards depicted in red correspond to more frequent, immediate, localized hazards, while those in orange refer to middle range threats in spatial and temporal scales, while those in green refer to longer term threats in time and cover greater geographic distances. While tornadoes and thunderstorms are microscale events with both rapid onset and concentrated potential for harm, other hazards and threats such as climate change, intra-seasonal oscillations and ENSO (El Nino Southern Oscillation) cover larger spans of time and distance in terms of their onset and location of effects. It is more difficult to predict with precision, microscale events. These hazards often threaten life and property and require effective alert and warning systems. While with some weather-related hazards such as wildfire, riverine flooding, and mid-latitude cyclones, there may be more time to prepare and preposition assets for response and recovery and evacuate communities, these hazards may cover a much larger physical area, making it difficult to focus attention and increase levels of preparedness. Synoptic and planetary hazards suffer from the “out of sight, out of mind” perspective, so individuals, agencies and organizations and communities may lack the

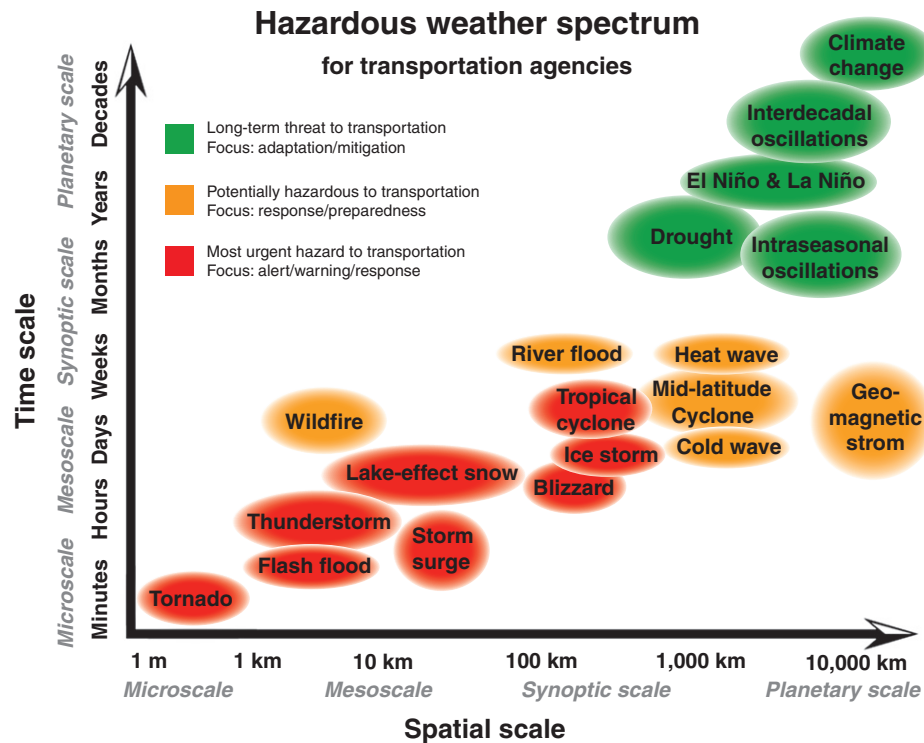


Figure 1 Spatial and temporal dimension of disruptive hazards with implications for transportation agencies.

sense of urgency and commitment to address these hazards and threats. More immediate, localized and understandable threats may divert attention away from longer-range threats and hazards.

While there may be time to prepare for and develop mitigation and adaptation strategies for climate change, much of the attention of the disaster and emergency management profession focuses on response and recovery from more immediate harm causing events. Moreover, threats associated with man-made hazards such as terrorism and industrial accidents have gained prominence since 9/11 and the Fukushima nuclear disaster. While the loss of life and economic losses associated with natural hazards is far greater, there is more attention to man-made threats as evidenced by the investments in the creation of the Department of Homeland Security and the Transportation Safety Administration ([Transportation Security Administration, 2018](#)).

Risk is the product of the likelihood and the consequences of hazards. There are frequently occurring, low consequence events and also rare, high consequence events. The management of risk involves understanding of both frequency and consequence of hazards.

Weather hazards are the most common type of disruption because weather varies greatly. Heavy rainfall, flooding, snow, and ice, and hazards such as landslides, avalanches, or downed trees and powerlines, or high winds can cause loss of vehicle control and accidents.

A seemingly minor weather event can cause major disruption such as in January 2014 when a few inches of snow fell in Atlanta ([Kim et al., 2018b](#)). While the National Weather Service had issued storm warnings, effort to plow and salt the roadways were hampered by lack of equipment, experience, and congestion caused by the early release of government employees and accidents, causing massive delays in a region which experiences severe peak traffic congestion even on days with good weather. Because of stalled vehicles, blocked roadways, and interrupted school bus transportation, students and teachers were forced to shelter in schools and other locations. A crisis situation was created because the disruption affected the travel of school children.

The largest disruptions to the transportation system result from big tropical storms and hurricanes, which may impact vast areas such as when Hurricane Katrina or Sandy struck. Hurricanes produce a combination of hazards and threats. In addition to damaging winds, storm surge, heavy rainfall, flooding, and the failure of power, water, wastewater, and other infrastructure systems can harm people and lead to the loss of property, roads, and other infrastructure. With Sandy, tunnels were flooded and closed, transit system equipment was damaged, and transportation services throughout the region were curtailed. Throughout the region hundreds of thousands of vehicles were damaged or destroyed. The towing, storage, repair, and disposal of damaged vehicles also present challenges.

While weather is the most common occurring hazard, there are also geophysical events and processes including volcanoes ([Ulfarsson and Unger, 2014](#)), earthquakes ([Sherrouse and Hester, 2008](#)), landslides, and other events which can significantly disrupt travel behavior. In some cases, such as when the Kilauea volcano erupted in Hawaii, the disruption is localized, affecting commuters, residents, and businesses because of damage to roads or because of evacuation orders requiring people to leave their homes or business or preventing students, workers, customers, tourists, or travelers engaged in recreational or social activities from

completing their trips (Kim et al., 2018a). These damaging events could lead to the rerouting of traffic or repositioning of schools, solid waste transfer sites, emergency services, and other activity centers.

With some volcanic events, the hazard is long-lived, lasting for weeks, months, or years. In addition to the lava threats, there are also concerns associated with volcanic gas, ash and other hazards such as acid rain which can harm human health, livestock, and agriculture. This can lead to the disruption of social and economic activities resulting in changes in travel behavior. The Icelandic volcanic eruption in 2010 led to more than 100,000 flight cancellations, millions of stranded airline passengers, and over \$1.7 billion in losses to the airlines. Geologic hazards are difficult to predict because the active forces can be located below ground and are difficult to detect, monitor, and map with precision. By the time an earthquake or ground deformation occurs, there may be only seconds to react and take protective action. Another challenge with earthquakes is that they can lead to cascading effects such as when the Great Japan Earthquake in 2011 led to a tsunami in which thousands of people drowned and flooded the Daichi Nuclear Power Plant, causing overheating of reactors, explosions and the release of radioactive material into the environment.

Different hazards and threats impact different transportation modes, facilities, and systems in different ways. Accidents involving motor vehicles and collisions with pedestrians, bicyclists and other roadway users are common events, which injure and harm people and cause disruptions, delay, and curtailment of transportation services. Coupled with fire or a hazardous materials spill, an accident can further harm individuals and the environment. Collisions with bridges, utility poles, roadside equipment or people lead to the need for accident investigation, repair work, environmental remediation, and construction work which could further interfere with the movement of people, goods, and vehicles. Accidents involve human, vehicle, roadway, and environmental factors. While most accidents are unintentional, there are examples of vehicles such as trucks or airplanes being weaponized to cause harm and governments have developed comprehensive strategies for reducing the threats from terrorism (Transportation Security Administration, 2018).

Accidents can result from the collapse of infrastructure such as when the Interstate 35 West Bridge connecting Minneapolis and St. Paul had a structural failure, causing 1000 feet of the eight-lane bridge to plummet into the river below (Kim et al., 2018b). This disaster occurred during the peak evening rush hour at 6:07 p.m. and resulted in 13 fatalities and more than 140 injuries. The National Transportation Safety Board determined the cause was the result of both design defects and improper inspection and maintenance practices (Hao, 2010). The disruption impacted not just motorists but also pedestrians, bicyclists, rail, and river traffic. Improper maintenance and repair of aging infrastructure is another hazard, which can result in significant disruption.

Special events can cause disruption of transportation systems. These are planned large sporting events, conventions and meetings, celebrations, parades, and special occasions. Often, they are regulated, permitted events such as the Boston Marathon which was impacted by terrorism when two homemade bombs exploded on April 15, 2013. Three people were killed and hundreds were injured. In addition to disrupting the marathon and impacting participants and spectators, the massive manhunt for the terrorists launched a few days later, led to the closure of the transit system, Amtrak service, as well as schools and businesses were closed as a city-wide, shelter-in-place order was implemented. There were significant economic impacts associated with the shutdown of transportation services affecting thousands of workers, customers, and residents of the region.

Coping with and Mitigating Disruption

Individual Strategies

When travelers encounter disruption, they can implement coping strategies. If possible, they change their route to avoid impassable areas or delay causing congestion. If the disruptions are known or planned events, they can change departure times, leaving either earlier or later to avoid slowdowns, road closures, or system shutdowns. Postponement or cancellation of trips can be facilitated by flextime, work from home, telecommuting, and improvements in communications and other technologies that support adjustments in travel behavior. Yet for some work, educational, and social activities, it may not be possible to avoid travel and face-to-face interactions. This may require the traveler to change destinations or change the travel mode, switching from surface to air transport or from ground based to marine travel. In some cases, travelers may need to switch from driving private automobiles to transit, shared ride, or combined modalities (walk, rideshare, bus, rail, etc.) to get to their destinations. In some instances, nonessential travel can be cancelled or deferred to a later time. Trips such as shopping can be accommodated with delivery services or other arrangements with suppliers as well as alternative providers of goods and services. Depending on the type and duration of disruption, the impacts on cost, convenience, consumers, and livelihoods can vary widely.

Of special concern is the effect of transportation disruption on evacuation. Disrupted or damaged transportation systems may strand evacuees or require rescue or assistance to move people, pets, and possessions to safety. Closed roads, cancellation of transit services, and decreased accessibility can also impair the ability of transit workers and service providers to get to work and operate vehicles, equipment, and facilities.

Agency and Organizational Strategies

Agencies and service providers need contingency plans, emergency operations procedures, and protocol for detecting and managing disruption. It is useful to consider both frequently occurring, low impact events such as heavy rainfall, power outages, accidents as well as rare, high consequence events including terroristic acts, earthquakes, flooding, and major disruptions leading to full or partial

shutdown of transportation systems. The experiences from managing smaller events are relevant to management of larger, more disruptive events. Often, the same personnel involved in operations and responding to hazards, threats, and system disruptions will need to be called upon for situational awareness, damage assessment, search and rescue, crowd management, and effective incident management.

In order to prepare for transportation disruptions, agencies and organizations must understand how to manage incidents, communicate and coordinate with other emergency response agencies (law enforcement, fire, emergency medical services, and others involved in disaster response). They must learn to optimize resources to reduce the harm from disruptions (Grayson and Noonan, 2010). For large-scale incidents, it may be necessary to call upon other agencies, jurisdictions, and assets from state and national agencies including the military to provide assistance, mutual aid, human and technical resources to support incident management. The I-35W Bridge collapse brought first responders from multiple local jurisdictions and required coordination between local, state, and federal authorities involved in search and rescue as well as the restoration and recovery of transportation services. A key to the effectiveness of the response to this disaster was training of personnel on Incident Command System, National Incident Management System (NIMS) and operating procedures, policies, and guidance for coordinated command and communications for emergencies and disasters. Implemented after the 9/11 terrorist attacks, NIMS is a method for managing response operations across all types of hazards and threats so that different agencies and jurisdictions can work together effectively. NIMS provides a framework for operations, planning, logistics, and financial administration, that is flexible, scalable, and makes the use of common definitions and terminology so that diverse personnel coming from different disciplines, agencies, jurisdictions, and levels of government can function together (U.S. Department of Homeland Security, 2019a, 2019b).

To reduce the impacts of disruptions, it is important for agencies and organizations to invest in drills, exercises, training, education, and capacity building in order to detect, communicate, manage, respond to and recover from disruptions. The training needs to be linked to risk and vulnerability assessments and the identification of potential hazards and threats and assets for response and recovery. In addition to building on the experiences from accidents and minor incidents, agencies and organizations conduct joint evaluations, after-action reports (Federal Emergency Management Agency, 2017), and “hot washes” to identify and categorize lessons learned as well as understanding of the materials, equipment, personnel, procedures, and policies needed to increase effectiveness and the capacity to manage disruptions and improve coordination across agencies (DeBlasio et al., 2004). The experiences of similar agencies experiencing common threats should be shared more widely across the transportation industry. Such evaluations should be conducted after disruptive events and following simulations or exercises designed to improve response and recovery capabilities. Transportation agencies and organizations need to create a culture of preparedness and to support investment in training and education for leaders, managers, and operational personnel.

Community Strategies

Effective response to disruptions involves communication and coordination with many different stakeholders. In addition to managers and field supervisors and system operators, it is important to work with and establish relationships with law enforcement, fire and emergency medical services and other responders (hazmat teams, urban search and rescue, utilities, etc.) which may be providing aid and assistance in the event of disruptive incidents. It is also important to focus on vulnerable, special needs populations which may require assistance in evacuation or transport to medical and care facilities.

Large incidents with multiple injuries and fatalities may require triage and additional transportation assets and personnel to manage care and the needs of victims, survivors, and affected people. In addition to physical and medical conditions, there may also be other social, cultural, language, psychological, and human needs of those affected by disruption. People may need to communicate with family members, friends, or employers. Travelers involved in accidents or incidents leading to damaged vehicles or equipment may need to find alternative modes to complete their journey or to manage the removal or towing of personal vehicles and may need to complete accident investigation forms working with law enforcement, insurers, and others involved in damage assessment and determination of the cause or “fault” associated with the incident.

For terroristic and intentionally caused incidents, the location may be treated as a crime scene in which evidence may need to be collected, analyzed, and preserved for further investigative or judicial action, such as immediately following the Boston Marathon bombing (Kim et al., 2018b). In these incidents, transportation officials will need to work closely with incident commanders and on-scene investigators who need to collect data and information from witnesses as well as physical evidence which can degrade quickly over time.

Large incidents may require the provision of on-scene medical as well as mass care (food, water, clothing, shelter, etc.) for victims, bystanders, and others (businesses, workers, etc.) impacted by a harmful event. In addition to government, various humanitarian assistance, nongovernmental organizations such as the Red Cross and faith-based organizations and spontaneous volunteer groups may materialize to render aid and assistance. With a large disaster, there may be volunteer rescuers such as the “Cajun Navy” or the evacuation of lower Manhattan following 9/11. For long duration disasters and disruptions such as volcanic eruption (Kim et al., 2018a) or events which lead to the longer-term displacement of residents and businesses, there may need to additional assistance and support during the restoration and recovery process (Kim et al., 2019).

Large incidents which damage infrastructure, buildings, and homes may create long-term disruptions to not just travel behavior, but also to the underlying social and economic activities for which the transportation system is designed to support. There may need to be adjustments or temporary services to accommodate transportation during the reconstruction phase. With disasters such as the

release of radioactive material following the tsunami in Japan, the cleanup and environmental remediation may take decades forcing long-term displacement of residents and businesses, prolonging the recovery for affected communities.

While individual travelers and transportation agencies can take steps to increase their preparedness and resilience to harm causing events, there is also a need for collective, community action to minimize the potential for disruption. This includes the planning for and mitigation and adaptation to hazards and threats such as climate change, urbanization, and stressors which increase the potential for harm to communities. In addition to identifying the risks of global warming, sea level rise, intensification of weather events, there is need for research and development of new materials and approaches to making transportation systems more durable, resistant, and stronger but also able to withstand and recover from flooding, wind, and seismic events (Asam et al., 2015). There is also need to invest in the development of robust supply chains to manage the flow of goods and services during and after disruptive events so that needed supplies and personnel can move quickly and efficiently across networks which may also be affected by disruption (U.S. Department of Homeland Security, 2019a). While these are large challenges related to social, economic, and geo-political forces, it is necessary to develop understanding of resource flows and effective strategies and actions for managing both the impacts as well as the systemic causes of disruption.

Small disruptions result from routine, more manageable incidents. Larger disruptions are caused by exogenous and global forces and require more extensive collective, collaborative efforts, agreements, and actions across regions, nations, and international partners. International scientific organizations such as the World Meteorological Association or the accords and agreements for sharing data on seismic hazards, volcanoes, tsunamis, etc. provide not just a basis for scientific advancement, but also opportunities to improve detect and warning systems as well as reduce stressors and social conflict arising from inequities and unbalanced concentration of wealth, resources, and economic opportunity. Transportation systems are major consumers of fossil fuels, contribute significantly to greenhouse gas emissions and global warming, exacerbating the potential for large-scale disruption of the climate system. There is, however, time and opportunity to reduce carbon footprints and the potential harm to transportation systems (Asam et al., 2015) and the communities they serve.

Key Lessons and Conclusions

Five key lessons emerge from this discussion on the effects and responses to transportation disruptions. They include the following:

1. anticipate the unanticipated,
2. focus on vulnerabilities and vulnerable populations,
3. agility = resilience,
4. learn from disruptive events, and
5. invest in research, training, and education.

Transportation agencies can avoid the unmanageable if they can learn to anticipate disruptive forces and manage unavoidable impacts to vulnerable populations. Those facing mobility and societal challenges are likely to be most severely impacted by disruptions. The physical components (vehicles, equipment, hardware) of the transportation system also need routine maintenance and upgrading. Transportation service providers need to avoid “unmanageable” incidents with complex, cascading effects such as the Fukushima disaster, which overwhelm local capacity and require outside assistance. They also need to manage “unavoidable” routine events such as accidents and flooding by anticipating and learning from disruptive events to address system vulnerabilities. Investments in research and capacity building promote greater understanding, agility and resilience.

In addition to plausible worst-case scenarios, transportation agencies should also anticipate the unanticipated. This includes studying catastrophic events that have happened in other jurisdictions and researching cases and best practices (American Association of State Highway and Transportation Officials, 2015). Prior to 9/11, there was little awareness of the potential for weaponization of transportation assets. The potential for disruption to the airline industry from the volcanic eruptions in Iceland was also not anticipated. The effects of other “creeping” disasters such as wildfires or sea level rise or increased storm intensities also need to be better anticipated and addressed in the planning, design, engineering, financing, construction, and maintenance of transportation systems. This may include planning, design, and construction of new seawalls or other infrastructure to mitigate harm from hazards. See for example, the work done in Virginia on the vulnerabilities of roads to storm surge and sea level rise (Kleinosky et al., 2006). Increased threats associated with cybercrime and the potential for disruption to transportation systems also need to be mitigated. Development of new technologies included connected and automated vehicles create new vulnerabilities. In addition to aging systems, the risks and vulnerabilities associated with new hardware and communications platforms (mobile, satellite, 5G, etc.) will need to be integrated into risk management. Such investments will require larger collective action across diverse constituencies and stakeholders who need to be better informed as to the benefits and returns on investment in resilient transportation systems.

A key element to minimizing the impacts of disruption is to focus on vulnerable populations and weaknesses in vehicles, facilities, and transportation systems. This involves strengthening facilities and structures to withstand earthquakes and tornadoes and other hazards. Components prone to breakdown require more maintenance as well as the stockpiling of spare parts. There is not enough attention to special needs populations as well as those with limited access and barriers to transportation assets due to medical, social, and economic conditions. There is need for reserve capacity and support services during incidents and disruptions (Cats and Jenelius, 2015). More attention to paratransit operations during disruptions as well as accommodation of the needs of

persons with disabilities, medical conditions, mobility limitations, and social barriers because of race, gender, age, or other factors needs to be a priority in the planning for and management of transportation disruptions. This requires closer coordination with social service providers and other partners handling special needs transportation services during and after disruptive events. With ridesharing, ride hailing services, the landscape of transportation service providers has changed greatly (Goodall and Lee, 2019). More attention to vehicle design, context sensitive design, accessibility, and transportation services for all will also have payback for the management of transportation disruption.

An important aspect of transportation resilience is the agility of travelers, agencies, and communities in responding to and recovering from disruption. Agility involves flexibility but also surplus capacity and resources which can be called upon when needed. It requires not just robust people and systems with effective technologies for seeing, sensing, detecting and assessing disruptions, but also effective communications between system operators, passengers, and first responders and others engaged in evacuation, sheltering, search and rescue and recovery. Agility involves not just speed, but also effective coordination and timely action to control vehicles, facilities, as well as manage crowds and those involved in the provision of emergency services.

Learning from disruptive events (Donahue and Tuohy, 2006) provides opportunity to study the causes and remedies associated with large and small disruptions. It is useful to look across a wide variety of different hazards and threats and consider how best to prepare, respond to, and mitigate the harm and impacts associated with natural and human caused incidents. Smaller, more routine events such as accidents or criminal activity demonstrate response capabilities and the technologies and systems for detection, alert, and warning and protocol for emergency response and system recovery (Goodall and Lee, 2019). The same personnel will likely be involved in the response and recovery from larger more disruptive events. Disruptions can be measured (Murray-Tuite and Mahmassani, 2004) in terms of the extent to which service is delayed, cancelled or closed, or degraded in terms of the number of travelers, travel times, mode changes, trip completions, inconvenience, and other measurable indicators of level of service. This is the basis for identifying vulnerable locations, links, nodes, and networks. Running plausible, worst case disruption scenarios due to weather, intentional acts of sabotage, accidents, or equipment and infrastructure failure due to poor design or inadequate maintenance is part of the effort to learn from disruptions and promote a culture of preparedness. Such efforts should include not just operators and managers, but the whole community of responders, caregivers, and volunteers who mobilize during disasters and disruptive events.

More attention to research, education and training is needed to promote understanding of disruption and the development of capabilities to reduce the effects on travelers and communities. In addition to studying and documenting the impacts and best practices with regard to different disruptions, there is need for interdisciplinary, international exchange of knowledge, technologies, and tools for detecting and alerting system users as well as for those involved in the planning, design, and operation of transportation services. In addition to better integration between routine operations and understanding of the management of congestion, incidents, and disruptions, it is necessary to encourage more testing and evaluation of procedures and actions in the event of disruptions. An emphasis on training and education of those in the transportation industry as well as more broadly in the communities at-risk from hazards and threats will serve to promote greater agility and resilience.

Biography

Karl Kim is Professor of Urban and Regional Planning at the University of Hawaii at Manoa, and Director of the graduate program in Disaster Management and Humanitarian Assistance. He studies transportation, cities, and resilience. He served as the Chief Academic Officer (Vice Chancellor for Academic Affairs) overseeing tenure and promotion, strategic planning, international programs and program review for the University of Hawaii. He is Executive Director of the National Disaster Preparedness Training Center (ndptc.hawaii.edu), authorized by the US Congress to develop and deliver FEMA certified training courses for underserved, at-risk communities on natural hazards, mitigation, and urban planning. The Center has trained more than 50,000 first responders, emergency managers and leaders in over 400 different communities across the world. Dr. Kim has served as the Chairman of the National Domestic Preparedness Consortium (ndpc.us). He has published many papers on disaster management, transportation, and urban planning. He is Editor-in-Chief of *Transportation Research Interdisciplinary Perspectives* (Elsevier) and is editor of a 10-volume series on disaster risk reduction and resilience (Routledge). He has been a Fulbright Scholar to Korea and the Russian Far East. Dr. Kim was educated at Brown University and the Massachusetts Institute of Technology.

References

- American Association of State Highway and Transportation Officials, 2015. *Managing Catastrophic Transportation Emergencies: A Guide for Transportation Executives*. ASHTO, Washington, DC ISBN: 978-1-56051-639-2.
- Asam, S., Bhat, C., Dix, B., Bauer, J., Gopalakrishna, D., 2015. *Climate Change Adaptation Guide for Transportation Systems Management, Operations and Maintenance*. Federal Highway Administration, Washington, DC.
- Cats, O., Jenelius, E., 2015. The value of reserve capacity for public transportation network robustness. *Trans. Res. A* 81, 47–61.
- DeBlasio, A., Regan, T., Zirker, M., Fichter, K., Lovejoy, K., 2004. *Effects of Catastrophic Events on Transportation System Management and Operations: Comparative Analysis*. DOT-VTSC-FHWA-04-03. Federal Highway Administration, U.S. Department of Transportation, Washington, DC.
- Donahue, A., Tuohy, R., 2006. *Lessons we don't learn: a study of the lessons of disasters, why we repeat them, and how we can learn from them*. Homeland Secur. Affair. II (2), 1–28.
- Federal Emergency Management Agency, 2017. *2017 Hurricane Season FEMA After-action Report*. Federal Emergency Management Agency, Washington, DC.
- Goodall, N., Lee, E., 2019. Comparison of Waze crash and disabled vehicle records with video ground truth. *Trans. Res. Interdiscip. Perspect.* 1., 100019. <https://doi.org/10.1016/j.trip.2019.100019>.
- Grayson, M., Noonan, T., 2010. *Optimization of Resources for Mitigating Infrastructure Disruptions Study*. National Infrastructure Advisory Council, Washington, DC.
- Hao, S., 2010. I-35W bridge collapse. *J. Bridge Eng.* 15 (5.), 608–614.

- Kim, K., Pant, P., Yamashita, E., 2018a. Managing uncertainty: lessons from volcanic lava disruption of transportation infrastructure in Puna, Hawaii. *J. Emer. Manag.* 16 (1), 29–40.
- Kim, K., Francis, O., Yamashita, E., 2018b. Learning to build resilience into transportation systems. *Trans. Res. Rec.* 2672, 1–13, doi:10.1177/0361198118786622.
- Kim, K., Pant, P., Yamashita, E., Ghimire, J., 2019. Analysis of transportation disruptions from recent flooding and volcanic disasters in Hawai'i. *Trans. Res. Rec.* 2673, 1–14, doi:10.1177/0361198118825460.
- Kleinosky, L., Yarnal, B., Fisher, A., 2006. Vulnerability of Hampton roads. Virginia to storm surge flooding and sea level rise. *Nat. Hazar.* 40 (1), 43–70.
- Murray-Tuite, P., Mahmassani, H., 2004. Methodology for determining vulnerable links in a transportation network. *Trans. Res. Rec.* 1882, 88–96.
- National Cooperative Research Program, 2012. Methodologies to Estimate the Economic Impacts of Disruptions to the Goods Movement System. Report 732. Transportation Research Board, Washington, DC.
- Sherrouse, B., Hester, D., 2008. Potential effects of a scenario earthquake on the economy of southern California: Intraregional commuter, worker, and earnings flow analysis. Open File Report 2008-1254. U.S. Geological Survey, U.S. Department of the Interior, Reston, VA.
- Transportation Security Administration, 2018. 2018 Biennial national strategy for transportation security. Report to Congress. U.S. Department of Homeland Security, Washington, DC.
- U.S. Department of Homeland Security, 2019a. Supply Chain Resilience Guide. U.S. Department of Homeland Security, Washington, DC.
- U.S. Department of Homeland Security, 2019b. National Response Framework, fourth ed. U.S. Department of Homeland Security, Washington, DC.
- Ulfarsson, G., Unger, E., 2014. Impacts and responses of Icelandic aviation to the 2010 Eyjafjallajökull volcanic eruption. *Trans. Res. Rec.* 2214, 144–151, doi:10.3141/2214-18.

Further Reading

U.S. Department of Transportation, nd, Road weather management best practices. Version 3.0. U.S. Department of Transportation, Washington, DC.

Use/Analysis of Crash Data and Underreporting of Crashes

Mohammadali Shirazi*, Dominique Lord†, Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States; †Zachry Department of Civil and Environmental Engineering, Texas A&M University, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Crash Data Modeling, Methodological Issues, and Data Limitations	726
Underreporting of Crash Data	727
Estimation of Underreporting Rate	727
Underreporting and Modeling Issues	728
Summary	729
References	729

Crash Data Modeling, Methodological Issues, and Data Limitations

Various statistical models have been proposed to model the frequency and severity of crash data (Lord and Mannering, 2010; Lord et al., 2021; Mannering and Bhat, 2014; Savolainen et al., 2011). Crash data are often over dispersed, which means that the variance of crash data is greater than its mean. Therefore, the Poisson distribution which is often used to model the data characterized by an equal mean and variance is not a suitable choice for crash data modeling. The Negative Binomial (NB)—also known as Poisson-gamma—is the most common model used to analyze over-dispersed (i.e., variance > mean) crash data. The Negative Binomial model is a mixture of Poisson and gamma distributions. The gamma distribution gives additional flexibility to the Poisson distribution to accommodate the over dispersed data.

Although the NB model is the most common model used to model crash frequency, it is not without limitations. In fact, there are several methodological issues with the NB model. As such, the NB model is not adequate for modeling data with many zero observations (Geedipally et al., 2012; Lord and Geedipally, 2011; Lord and Mannering, 2010; Shirazi et al., 2016, 2017), or accounting unobserved heterogeneity (Mannering et al., 2016). On rare occasions, also, the mean of crash data might be greater than its variance, or in other words, the crash data are under-dispersed, which usually occurs due to the data generating process (Lord and Mannering, 2010; Lord et al., 2021). In this scenario, the NB model is no longer appropriate and other models that can accommodate data with a variance less than the mean such as Conway–Maxwell–Poisson model (Lord et al. 2008; Lord et al., 2010) are recommended to be used. In addition, data may be characterized by the bias caused by a low sample mean and a small sample size. Lord and Mannering (2010) listed three major reasons that leads to a dataset with a low sample mean and a small sample size. First, there are data collection limitations. Collecting safety data is expensive and requires significant time and resources. Hence, collecting all data may not be economically possible. Second, this issue can occur when analyzing unique and rare facility types. In fact, some facility types are so unique and rare in the jurisdiction under study that the analyst does not have access to enough number of sites or locations, to collect adequate data to develop a reliable model. Third, datasets can be characterized by excessive zero observations. In addition to the methodological issues caused by data characterized by many zero observations (as noted above), excessive zero observations in data can also result in a low sample mean and a long tail which adversely affect the performance of the model. Data characterized by a low sample-mean and a small sample size result in inaccurate estimation of the NB dispersion parameter (Lord, 2006; Lord and Miranda-Moreno, 2008).

Before describing data limitations and modeling issues with underreported crash data, it is important to discuss one more important modeling issue that should not be confused or mixed with the underreporting of crash data. This issue is referred to as the omitted variable bias in statistical analysis. Omitted variable bias refers to instances when some important independent variables are not included in the model, which can cause significant bias or inaccuracy in the modeling results. A positive bias and over estimation of parameters happen when both of the independent and dependent variables are positively (or negatively) correlated with the omitted variable. On the other hand, a negative bias occurs when the dependent variable is positively (or negatively) and independent variables are negatively (or positively) correlated with the omitted variable. Omitted variable bias result in inaccuracy in parameter estimations, and consequently leads to incorrect crash predictions, and erroneous safety evaluations and analyses.

As a closing note to this section, it is worth pointing out that, regardless of the type of the model or the method used for the safety analysis, modeling itself is not a trivial task and involves different steps and nuances, from data collection to cleaning the data to estimating, evaluating and selecting the best model. Various questions should be answered by the analyst when he or she works with the data during the modeling phase. For example, should he or she use a panel or cross sectional data? What should be the functional form? What variables should be used in the model? What is the best model? All these questions should be answered one by one during the modeling process. A final model is usually selected based on comparing several alternative models and considering multiple modeling objectives such as goodness of fit, the intuitive relationships between the response and variables, goodness of logic (as explained by Lord et al., 2021; Miaou and Lord, 2003; Shirazi et al., 2017; Shirazi and Lord, 2019) and prediction accuracy.

Underreporting of Crash Data

Crash data recorded by law enforcement are considered as the primary source of information in analyzing crash data and evaluating safety. However, not all crashes are reported in police records. As such, in most jurisdictions, a minimum threshold needs to be met to include a crash in official statistics. For example, a crash must have caused \$1200 or more in damages to be reported. In fact, incomplete reporting of crashes in official statistics has been a critical issue in analyzing crash data for decades (Elvik and Mysen, 1999). The severity of crashes is typically classified in ordinal scales using five categories: fatal injury (K), incapacitating injury (A), nonincapacitating injury (B), possible injury (C) and property damage only (PDO). While in most developed countries, fatal crashes have been properly defined and carefully reported, the same cannot be said for nonfatal crashes. Reviewing 18 different studies, Hauer and Hakkert (1988) estimated an underreporting rate of less than 5% for fatal crashes, but they estimated much higher rates for the less severe crashes. Hauer and Hakkert (1988) pointed out that approximately 20% of severe injuries, and 50% of all injuries are often not reported. In another study, Alsop and Langley (2001) noted that only about two-thirds of hospitalized crashes were reported by law enforcement in New Zealand in 1995. The level of underreporting of accident data can induce uncertainty about the modeling and analysis of crash data and result in significant bias in estimation of model coefficients.

Wood et al. (2016) documented several factors that influence underreporting of crash data. First, and probably by far the most critical criteria, the extent of the damage to the vehicle. As noted above, each jurisdiction has a certain threshold to report crashes. If it is a minor crash, and the damage is below the reporting threshold, the crash will not be reported in official statistics. Second, drivers' willingness to report the crash. For instance, a driver involved in a PDO collision (above the minimum reportable threshold) may choose to pay the damage him or herself and not report the accident to authorities (e.g., law enforcement or his or her insurance company), in fear of penalty for committed violations or that his or her insurance premium may go up. Finally, yet importantly, the officer may deem the crash minor and does not file the report.

In addition to the above factors, the jurisdiction or the area where the crash occurs can also influence the underreporting rate. In fact, a case study in France (Amoros et al., 2006) showed that the underreporting rate varies based on the crash area (e.g., urban vs. rural or remote locations). In addition, even the road type and the police availability in the area considered as a major factor influencing the underreporting rate. The research studies also analyzed the relationship between various demographic factors with underreporting rates. While not all demographic factors are correlated with the level of underreporting, there are strong correlations between some of these factors and underreporting rate. For example, Alsop and Langley (2001) noted that the underreporting rate varies by the age of participants in New Zealand crashes. On the other hand, they concluded that the gender or race of participants are not significant factors. Last but not least, the time of crash and the number of vehicles involved in a crash are also considered as two additional factors, apart from severity levels (e.g., minor crashes are not completely reported) and geographical locations (urban vs rural), that influence the underreporting of crash data (Alsop and Langley, 2001). For example, crashes are more likely to remain unreported at nights or some months of the year, or the crashes that involve two or more vehicles are more likely to be reported than those that involve only one vehicle.

As a closing note to this section, it is also worth noting that in the United States, crash records collected or managed at the National Highway Traffic Safety Administration (NHTSA) and all departments of transportation (DOTs) include only crashes that involve motor vehicles and only crashes that occur on public roads (or in some states, like Maine, on public roads or in private areas where public traffic is anticipated, such as in Walmart parking lots). So, it may not be surprising to notice that traffic injuries involving bicyclists but not motor vehicles are not collected or compiled by governmental agencies. Even motor vehicle–bicycle crashes may not be reported to (or by) the police since the motorist may drive the bicyclist to the emergency room (ER) without involving the law enforcement.

Estimation of Underreporting Rate

The previous section described the issue with the underreporting of crashes, and how underreporting seems inevitable due to various factors. The next question is whether or not we can estimate the underreporting rate, and to what extent we can achieve this goal. This section reviews the methods and data sources used to estimate the underreporting rate.

Earlier studies mostly estimated or determined the underreporting rate by comparing the statistics recorded in hospitals or emergency rooms with those reported by law enforcement. For example, Aptel et al. (1999) found that there was a 62.7% discrepancy in nonfatal injury crashes between hospital data and police records. However, although initially considered a promising approach, comparing the hospital records and law enforcement data may not be sufficient to provide an accurate estimation of the underreporting rate. Several reasons were noted in safety literature to justify this claim (Aptel et al., 1999; Elvik and Mysen, 1999). First, hospital records are not without limitations themselves and are also likely to be incomplete, similar to the police records. Second, the source of injury in patient files or hospital records may not be clear or could contain mistakes. Also, privacy rules may mean that the crash location and other pertinent data cannot be easily obtained. Third, there might be a significant regional discrepancy between the regions where the crash occurs and where an injured person is hospitalized which will add substantial level of difficulty to determine the bias caused by underreporting, or estimate the underreporting rate. Last but not least, the definitions used in hospitals and police records may not be consistent and can be incompatible with each other. One advantage with hospital records is that E-codes, N-codes and AIS-scales are identically defined around the world whereas the definition of evident injury may vary from country to country and even within countries.

Despite these limitations, however, still hospital or emergency room records are used as a major source of information to estimate the underreporting rate. In fact, hospital or emergency room records might provide satisfactory results to estimate the underreporting rate for some types of crashes. For instance, [Janstrup et al. \(2016\)](#) used data reported in police records, and those collected from the emergency rooms to find out—given a set of variables—what type of crashes are more likely to appear on both police and hospital records. The study found that crashes associated with the use of helmet or seat belt, or those related to the alcohol usage or speeding are more likely to appear on both police and hospital records. As another example, In Sweden, the Swedish Traffic Accident Data Acquisition (STRADA) system ([OECD, 2019](#)) deploys a systematic connection between the law enforcement records (crash information) and hospital data (medical information) to provide greater accuracy in reporting severe crashes (fatal and serious injury crashes). However, although STRADA system is often reliable for serious injury crashes, the less severe crashes (e.g., minor injuries that only need primary care) are likely to remain underreported in the system.

Research studies resorted to other sources of data such as those collected by insurance companies to determine the underreporting rate as well. However, using other supplementary data such as insurance records can also be inadequate to determine the level of underreporting, as these sources are also likely to be incomplete or inconsistent with law enforcement records. For example, as noted above, it is likely that a driver decides not to report a minor crash to both his or her insurance company and law enforcement in fear to see a raise in insurance premiums or to be considered guilty for committed violations. In order to obtain more robust results, several studies have attempted to estimate the underreporting rate or alleviate the bias caused by underreporting using data from multiple sources. For example, [Watson et al. \(2015\)](#) merged and combined several data sources collected in Queensland, Australia to estimate the degree of underreporting in crash injuries reported by the police department. Data used for this purpose include the Queensland road crash database, admitted patients in Queensland hospitals, data from the emergency department information system, and data provided by the Queensland injury surveillance units. Using all these data, merged and combined, the researchers found that around two-third of injury crashes were not reported in statistics reported by law enforcement. Further analysis revealed that the extent of underreporting is larger in crashes involved bicycles or motorcycles, as well as those occurring in remote locations.

Crowdsourcing is another approach that was recently suggested by researchers. Crowdsourcing (e.g., Web-based mapping or GPS-enabled mobile devices) brings new avenues for data collection compared to traditional datasets that can be used in addition to the archived data to explain, analyze or estimate a phenomenon ([Lord et al., 2021](#)). [Medury et al. \(2019\)](#) used crowdsourcing to estimate underreporting in pedestrian and bicycle crashes around three university campuses in California. The crowdsource data used in this research included two sources of information: self-reported crashes as well as the perceived hazardous locations. Similar to the previous research studies, the results of the study indicated that the police records significantly underreported non-motorized crashes. The next section will discuss the impact of underreporting of crashes on crash data modeling and analysis.

Underreporting and Modeling Issues

Neglecting the underreporting issue will result in a bias sample; this bias sample induce inaccuracy in modeling, create inconsistency in classification outcomes, and result in erroneous parameter estimations. Research studies reported that both crash-frequency and crash-severity models are affected when underreporting is ignored and not considered in modeling. The underreported crash data create a biased sample that include unclear or unknown portions of the injury severities that do not represent the population ([Yamamoto et al., 2008](#)). Hence, the sample is over represented by higher severities, which can skew the estimation of parameters for some key safety variables. Underreporting will cause the threshold to distinguish injury and possible injury and the threshold to distinguish possible injury and PDO to be estimated with bias; a biased threshold consequently results in erroneous estimation of parameters or prediction of crash severities ([Hauer 2006](#)).

Crash data are often treated as outcome-based or choice-based samples instead of random samples from the population. [Ye and Lord \(2011\)](#) examined the impact of underreporting on estimation of parameters in three commonly used traffic crash severity models: multinomial logit, mixed logit, and ordered probit. They concluded that all three models are sensitive to the extent of the underreporting rate. They provided a few recommendations to minimize the bias. First, for the multinomial logit and mixed logit models, the fatal crashes are recommended to be set as the baseline severity level. Second, for the ordered probit model, the rank of the crash severity is recommended to be set from descending order from fatal to PDO. Other studies reported similar observations. For instance, research studies noted that the effect of explanatory variables could be estimated with substantial bias in severity outcome models if underreporting is not considered or ignored in modeling. A research study by [Yamamoto et al. \(2008\)](#) analyzed the impact of underreporting on ordered probit and sequential binary probit models, and concluded significant bias for both models. They also found that the bias in estimation of parameters is substantially larger when crash injuries are underreported compared to PDO crashes.

The underreporting issue was mostly studied for crash severity models; however, although it may not seem obvious initially, there is a clear need to address this issue for crash frequency models as well. Imagine a location with many minor crashes that was not reported in official statistics. This location then may not be identified as a hazardous location in hotspot identification, although it has experienced many crashes and is not considered safe. In fact, minor changes in some variables such as traffic volume or weather conditions can convert many of these minor crashes to more severe crash consequences ([Mannering and Bhat, 2014](#)). This is particularly important as in many countries the emphasis is on reduction of fatal and injury crashes (vision zero). However, if the link between minor and serious crashes is not properly defined, unsafe locations cannot properly be selected in hotspot

identification. Underreporting can also influence the number of zero observations in the dataset. As noted earlier in this article, excessive zero observations make the commonly used models such as the NB model inadequate for modeling. In addition, underreporting can lead to the low sample mean issue. When it is combined with the small sample size, this can adversely affect the parameter estimations by crash-frequency models (especially the NB model).

In short, underreporting of crash data induces significant modeling inaccuracy and should not be ignored or down played by safety researchers or practitioners. Estimating the underreporting rate might seem to be the most reliable way to handle the modeling issue with the underreported crash data. In other words, once the underreporting rate is determined or estimated, it can be applied to adjust the number of crashes. However, recall that although the sample can be adjusted when the underreporting level is properly defined, it is not easy or straightforward to assess or estimate the underreporting information in many instances—as discussed in detail in Section 3. To overcome this limitation, researchers explored other methods or techniques such as the application of adjustment estimators to alleviate the limitations imposed by underreporting of crash data. For example, Patil et al. (2012) used the weighted conditional maximum likelihood estimator to overcome the limitations imposed by an unknown underreporting rate, when statistical models such as logit or nested logit are used to analyze severity of crashes. Yet, still further research is needed to address the issue of underreporting of crash data directly or indirectly in statistical modeling of crashes.

Summary

Crash data recorded by law enforcement are considered as a crucial source of data in various stages of modeling, evaluations and analyses of highway safety. There are specific characteristics and limitations with crash data that make the data unique and the statistical analysis of crash data more challenging than a typical statistical analysis. For example, crash data may include a lot of zero observations, the sample size and the sample mean of the dataset might be small, or there might be instances that not all critical variables associated with crashes can be observed or collected. In all these cases, the typical models such as the NB model may not be adequate for modeling and the analyst may need to switch to a more advanced model to obtain better results. Most importantly, even the recording process of the crash data itself is not without limitations. In fact, one of the most critical issues with crash data modeling is related to the issue of the underreporting of crashes. While in most developed countries underreporting is rare for fatal records documented in official statistics, injury and PDO crashes are frequently underreported. Traditionally, research studies resorted to hospital data to estimate the level of underreporting in crash data. Hospital data, however, are inadequate to estimate underreporting in most instances, as hospital data itself may not be complete or can be inconsistent or may involve discrepancy with the geographical location of where the crash occurred. Recently, researchers resorted to using multiple sources of data or crowdsourcing to alleviate the inaccuracy in estimation of the underreporting rate. Yet, the accurate estimation of underreporting is a difficult task and demands significant time and resources. Safety researchers show that if not properly addressed, underreporting could result in substantial bias in estimation of parameters; consequently, the effect of independent variables might not be captured properly and can be significantly different than the reality, which will result in inaccurate safety evaluations or analyses, and incorrect investigations of the source of safety problems. Further research is needed to investigate new and modern data sources to estimate the underreporting rate, as well as exploring new methods or strategies to directly or indirectly overcome the modeling limitations associated with the underreporting of crash data.

References

- Amoros, E., Martin, J.L., Laumon, B., 2006. Under-reporting of road crash casualties in France. *Accid. Anal. Prevent.* 38 (4), 627–635.
- Aptel, I., Salmi, L.R., Masson, F., Bourd , A., Henrion, G., Erny, P., 1999. Road accident statistics: discrepancies between police and hospital data in a French island. *Accid. Anal. Prevent.* 31 (1–2), 101–108.
- Alsop, J., Langley, J., 2001. Under-reporting of motor vehicle traffic crash victims in New Zealand. *Accid. Anal. Prevent.* 33 (3), 353–359.
- Elvik, R., Mysen, A., 1999. Incomplete accident reporting: meta-analysis of studies made in 13 countries. *Transport. Res. Rec.* 1665 (1), 133–140.
- Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-Lindley generalized linear model: characteristics and application using crash data. *Accid. Anal. Prevent.* 45, 258–265.
- Hauer, E., Hakkert, A.S., 1988. Extent and some implications of incomplete accident reporting. *Transport. Res. Rec.* 1185 (1–10), 17.
- Hauer, E., 2006. The frequency-severity indeterminacy. *Accid. Anal. Prevent.* 38 (1), 78–83.
- Janstrup, K.H., Kaplan, S., Hels, T., Lauritsen, J., Prato, C.G., 2016. Understanding traffic crash under-reporting: linking police and medical records to individual and crash characteristics. *Traffic Injury Prevent.* 17 (6), 580–584.
- Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accid. Anal. Prevent.* 38 (4), 751–766.
- Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transport. Res. Part A: Policy Pract.* 44 (5), 291–305.
- Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective. *Safety Sci.* 46 (5), 751–770.
- Lord, D., Guikema, S.D., Geedipally, S.R., 2008. Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accid. Anal. Prevent.* 40 (3), 1123–1134.
- Lord, D., Geedipally, S.R., Guikema, S.D., 2010. Extension of the application of Conway–Maxwell–Poisson models: Analyzing traffic crash data exhibiting underdispersion. *Risk Anal.: An International Journal* 30 (8), 1268–1276.

- Lord, D., Geedipally, S.R., 2011. The negative binomial–Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. *Accid. Anal. Prevent.* 43 (5), 1738–1742.
- Lord, D., Qin, X., Geedipally, S.R., 2021. *Highway Safety Analytics and Modeling*. Netherlands, Elsevier Publish Ltd, Amsterdam.
- Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. *Analyt. Methods Accid. Res.* 1, 1–22.
- Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analyt. Methods Accid. Res.* 11, 1–16.
- Miaou, S.P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. *Transport. Res. Rec.* 1840 (1), 31–40.
- Medury, A., Grembek, O., Loukaitou-Sideris, A., Shafizadeh, K., 2019. Investigating the underreporting of pedestrian and bicycle crashes in and around university campuses— a crowdsourcing approach. *Accid. Anal. Prevent.* 130, 99–107.
- OECD, 2019. Road safety annual report, Sweden.
- Patil, S., Geedipally, S.R., Lord, D., 2012. Analysis of crash severities using nested logit model—accounting for the underreporting of crashes. *Accid. Anal. Prevent.* 45, 646–653.
- Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. *Accid. Anal. Prevent.* 43 (5), 1666–1676.
- Shirazi, M., Lord, D., Dhavala, S.S., Geedipally, S.R., 2016. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: characteristics and applications to crash data. *Accid. Anal. Prevent.* 91, 10–18.
- Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017. A methodology to design heuristics for model selection based on the characteristics of data: application to investigate when the Negative Binomial Lindley (NB-L) is preferred over the Negative Binomial (NB). *Accid. Anal. Prevent.* 107, 186–194.
- Shirazi, M., Lord, D., 2019. Characteristics-based heuristics to select a logical distribution between the Poisson-gamma and the Poisson-lognormal for crash data modelling. *Transportmetr. A: Transport Sci.* 15 (2), 1791–1803.
- Yamamoto, T., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. *Accid. Anal. Prevent.* 40 (4), 1320–1329.
- Ye, F., Lord, D., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit, and mixed logit. *Transport. Res. Rec.* 2241 (1), 51–58.
- Watson, A., Watson, B., Vallmuur, K., 2015. Estimating under-reporting of road crash injuries to police using multiple linked data collections. *Accid. Anal. Prevent.* 83, 18–25.
- Wood, J.S., Donnell, E.T., Fariss, C.J., 2016. A method to account for and estimate underreporting in crash frequency research. *Accid. Anal. Prevent.* 95, 57–66.

Utility Poles

Lai Zheng^{*,†}, Tarek Sayed[†], ^{*}School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin, Heilongjiang, China;
[†]Department of Civil Engineering, University of British Columbia, Vancouver, BC, Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction	731
Utility Pole Crashes	731
Contributing Factors and Statistical Modeling	732
Importance of the Clear Zone	733
Safety Treatments	733
Countermeasures	734
Keep Vehicles on the Roadway	734
Place Utility Lines Underground	734
Relocate Utility Poles to Locations Less Likely to be Struck	734
Increase Lateral Offset of Utility Poles	734
Decrease the Number of Utility Poles	734
Use Breakaway or Energy Absorbing Poles	734
Shield Utility Poles With Safety Devices	735
Delineate Utility Poles in Hazard Locations	735
Countermeasure Selection	736
Summary	736
See Also	736
References	736
Further Reading	736

Introduction

Utility poles are fixed objects that are intentionally placed on the roadside to carry electricity or communication wires. These objects are substantial both in sheer number and structure strength and can negatively affect the safe operation of roads. It is estimated that there are over 88 million utility poles on highway right-of-way in the United States, over 5 million utility poles in Australia, and more than 500 million utility poles in China. Most of these utility poles are made of wood, steel, or reinforced concrete. The rigid material allows the pole to have a longer service life and survive high winds of a storm and other environmental extremes, but unfortunately makes the pole unforgiving in a traffic crash. More importantly, in most countries utility poles are usually located within the highway right-of-way to avoid inference with trees and to ensure the reliability of the services carried on poles, and in some cases they are too close to the traveled lane, which makes them too dangerous to errant vehicles that run off the road and frequently leads to serious or fatal injuries.

The traffic safety problem related to utility poles, especially those not of forgiving (e.g., breakaway upon impacts) nature, has been recognized by many countries. For example, as pointed out in Transportation Research Board's *State of the Art Report 9: Utilities and Roadside Safety* (TRB, 2004), utility pole improvements have been made in the United States as early as highway engineers' recognition of the need for forgiving roadsides since the 1960s. Since then many utility poles have been removed, relocated to safer locations, or made safer in some other manner. However, there are still many more utility poles that remain in place as a significant roadside hazard in the United States. Moreover, some other countries have not paid enough attention to the utility pole crash problem. Far too many people are being killed or seriously injured each year in utility pole crashes all over the world, and many of these deaths and injuries could be avoided with appropriate safety treatments to utility poles.

This article is organized as follows. First, road safety problem related to utility poles is discussed emphasizing the fact that utility pole crashes are one of the most common fixed-object crashes and have a higher frequency of severe to fatal crashes. Contributing factors to utility pole crashes and some widely used utility pole crash prediction models are then presented. Subsequently, the importance of clear zones in reducing utility pole crashes is discussed. Last of all, safety treatments to utility poles are summarized with a discussion on their pros and cons and selection methods.

Utility Pole Crashes

A utility pole crash is the run-off-road/fixed-object crash that involves a vehicle leaving the traveled lane, encroaching the roadside, and striking a utility pole. Many people are killed and injured each year in utility pole crashes. In the United States, the proportion of fatalities involving crashes with fixed objects has remained between 21% and 23% since 2005. Among all the fixed objects, utility poles are one of the most common fixed objects to be struck, usually ranked the second after trees. The proportion of fatalities

Table 1 Fatalities of fixed-object crashes, tree crashes and utility pole crashes in the USA

Year	Fixed-object crashes		Tree crashes		Utility pole crashes	
	Number	Proportion to all crashes	Number	Proportion to fixed-object crashes	Number	Proportion to fixed-object crashes
2005	9062	21	4573	50	1223	14
2006	9181	21	4552	50	1142	12
2007	9168	22	4341	48	1191	13
2008	8661	23	4182	48	1029	12
2009	7813	23	3817	49	980	13
2010	7290	22	3614	50	1015	14
2011	7069	22	3563	50	910	13
2012	7697	23	3672	50	1004	14
2013	7553	23	3604	50	913	13
2014	7518	23	3514	47	953	13
2015	7700	22	3611	47	926	12
2016	8039	21	3801	48	897	11
2017	7833	21	3691	47	914	12

Note: The statistics are based on analysis of data from the US Department of Transportation's Fatality Analysis Reporting System (FARS).

involving utility pole crashes is between 11% and 14% of fixed-object crash fatalities and approximate 3% of total fatalities, as shown in [Table 1](#). In British Columbia and Canada, over a period of 12 years (1999–2011), the province experienced a total of 2314 crashes that involved utility poles, resulting in 68 fatalities, 1203 personal injuries, and 1043 with property damages only.

In general, utility pole crashes are not as frequent as other types of crashes. As such, highway agencies and utility companies seem to give the problem a low priority. Nonetheless, because of the high structural strength of utility poles as well as the small contact area of a collision, utility pole crashes tend to be severe. Moreover, the negative effects of utility poles go beyond the run-off-road crash type. Crashes are usually categorized by “first harmful event.” Some crashes, in which striking a utility pole is a secondary event that may be more severe than the first harmful event, are not classified as run-off-road/fixed-object crashes. A common example is that motorcycle crashes with utility poles often involve more than one impact event, and the collisions with utility poles usually have a higher fatality risk than collisions with either the ground or another motor vehicle.

Contributing Factors and Statistical Modeling

The occurrence of utility pole crashes, similar to the occurrence of other types of crashes, is influenced by traffic conditions (e.g., traffic volume, speed), road conditions (e.g., geometric alignments, pavement surface), driver characteristics (e.g., risk driving behavior, fatigue), and environmental factors (e.g., weather, lighting). Among these factors, the most important ones appear to be traffic volume, lateral offset, and pole density. Traffic volume is the exposure factor. Offset is measured laterally from the edge of the traveled lane to the utility pole. Both the crash frequency and crash severity will decrease when utility poles are placed farther away from the traveled lane. The frequency of utility pole crashes will also decrease when the number of poles along the roadway decreases, intuitively because less poles are exposed to errant vehicles. Other than the traffic volume, offset, and density, factors such as roadway class, shoulder width, horizontal curvature, intersection type, divided/undivided road, rural/urban, and speed limit are also important contributing factors.

To quantitatively measure the effects of contributing factors on utility pole crashes, several statistical models (e.g., safety performance functions) have been developed. These mathematical models are derived from crash data and are used to acquire inference on different factors that contribute to crash frequency.

The most-referenced utility pole crash model was a multiplicative model developed by ([Zegeer and Parker, 1984](#)). The model predicts the utility pole crash frequency as a function of average daily traffic (ADT, vehicle day⁻¹), pole density (Density, km⁻¹), and pole offset (Offset, m) using the linear and non-linear regression. The model is as follows:

$$\text{Crash} = \frac{(3 \times 10^{-5} \cdot \text{ADT}) + (1.735 \times 10^{-2} \cdot \text{Density})}{\text{Offset}^{0.6}} - 0.025 \quad (1)$$

The model is developed based on the data collected from four states in the United States: Michigan, North Carolina, Washington, and Colorado. Based on the available data, it is found that the model is a reasonably good predictor within the specific range of input values: ADT from 1000 to 60,000 vehicles per day, pole offsets from 0.18 to 2.77 m, and pole densities of 6–56 poles per kilometer.

In the current practice, most safety performance functions are developed with Poisson and negative binomial regression. The approach has the following theoretical basis. Let Y be the random variable that represents collision frequency at a given location during a specific time period, and let y be a certain realization of Y . The mean of Y , denoted by Λ , is itself a random

variable. For $\Lambda = \lambda$, Y is Poisson distributed with parameter λ .

$$P(Y = y | \Lambda = \lambda) = \frac{\lambda^y \cdot e^{-\lambda}}{y!} \quad (2)$$

$$E(Y | \Lambda = \lambda) = \lambda; \text{Var}(Y | \Lambda = \lambda) = \lambda \quad (3)$$

It is typical to assume that the distribution of Λ can be described by a gamma probability function with a shape parameter κ and a scale parameter κ/μ , and then the distribution of Y around $E(\Lambda) = \mu$ is a negative binomial with an expected value and variance as follows:

$$P(Y = y | \mu, \kappa) = \frac{\Gamma(y + \kappa)}{y! \Gamma(\kappa)} \cdot \left(\frac{\kappa}{\kappa + \mu} \right)^\kappa \cdot \left(\frac{\mu}{\kappa + \mu} \right)^y \quad (4)$$

$$E(Y) = \mu; \text{Var}(Y) = \frac{\mu^2}{\kappa} \quad (5)$$

Following the Poisson and negative binomial regression approach, a utility pole crash prediction model is developed by [Elesawey and Sayed \(2012\)](#), for the rural-undivided highway segments in British Columbia, Canada. The developed model includes four explanatory variables: segment length (L), traffic volume (ADT), poles density (Density), and poles offset (Offset). The final model is found to be the Poisson structure and has the following form.

$$\lambda = 0.1095 \cdot (L \cdot \text{ADT})^{0.1247} \cdot \text{Density}^{0.1162} \cdot e^{-0.905 \text{Offset}} \quad (6)$$

By setting segment length to 1.0 km, the expected utility pole crash frequency as crashes/km/year is calculated as follows:

$$\lambda = 0.1095 \cdot (\text{ADT})^{0.1247} \cdot \text{Density}^{0.1162} \cdot e^{-0.905 \text{Offset}} \quad (7)$$

Using the developed model, the percentage change of crash frequency associated with a change in the explanatory variables is estimated. It is found that the impact of density and offset on the crash frequency is more evident than the impact of the volume, and moreover increasing the pole offset has a more significant effect on reducing the frequency of utility pole crashes compared to increasing poles spacing.

Importance of the Clear Zone

The significant effect of offset on the utility pole crash frequency implies that utility poles should be placed as far as practicable from the traveled lane to provide a safe environment for traffic operation. One of the most important principles is to place the utility poles outside the clear zone. A clear zone is an unobstructed, traversable roadside area that allows an errant driver to stop safely or regain control of a vehicle that has left the roadway. Utility poles located outside the clear zone are more likely not to be struck by errant vehicles and thus collisions would be avoided.

The clear zone concept stresses the provision of wide enough roadsides free from hazards like utility poles and other ground-mounted structures. Usually, variable clear zone widths are suggested based on types of roads, speeds of vehicles, traffic volumes, horizontal curves, steepness of roadside slopes, and some other factors. For example, the *Roadside Design Guide* of the United States ([AASHTO, 2011](#)) suggests the clear zone widths of rural highways may vary from 2.0 m (for front slopes 1:4 or flatter and back slopes 1:3 or flatter, design speed ≤ 60 km h⁻¹, and annual average daily traffic (AADT) < 750 vehicle day⁻¹) to 14.0 m (for front slopes 1:4–1:5, design speed of 110 km h⁻¹, and AADT > 6000 vehicle day⁻¹) based on traffic volumes, speeds, and roadside geometry. The suggested width may be modified further with adjustment factors to account for horizontal curvature. For urban streets, although the same clear zone widths as rural highways are preferred, they are sometimes not feasible because of physical constraints. In these instances, a minimum width of 0.5 m from the face of the curb is required. The *Geometric Design Guide for Canadian Roads* of Canada ([TAC, 2017](#)) suggests the minimum clear zone widths ranging from 2.0 m (for front slopes 1:4 or flatter & back slopes 1:3 or flatter, design speed ≤ 60 km h⁻¹, and AADT < 750 vehicle day⁻¹) to 8.5 m (for front slopes 1:4 or flatter and back slopes 1:3 or flatter, design speed ≥ 110 km h⁻¹, and AADT > 6000 vehicle day⁻¹). For instances where the application of the “full” clear zone width requirements is not appropriate or cost-effective, a “reduced” clear zone by as much as 40% of the “full” width with a minimum width of 2.0 m is proposed.

Ideally no utility poles should be located within the clear zone. In practice, however, it has been difficult to satisfy this requirement due to many constraints. For road sections where it is not cost-effective to place the utility pole outside the clear zone, safety treatments as detailed in the following section may be necessary for utility poles.

Safety Treatments

Lists of countermeasures to utility poles have been provided in AASHTO’s *Roadside Design Guide* ([AASHTO, 2011](#)), and general guidelines on utility pole crash mitigation are also reiterated in Transportation Research Board’s *State of the Art Report 9: Utilities and*

Roadside Safety (TRB, 2004). This section provides a summary on these countermeasures and discusses the methods to select proper countermeasures.

Countermeasures

Keep Vehicles on the Roadway

The most obvious way to prevent utility pole crashes is to keep the vehicles on the roadway and from encroaching on roadsides. This can be realized by some low-cost, quickly implementable treatments, such as reducing speeds, installing shoulder rumble strips, installing edgeline profile marking, installing mid-lane rumble stripes, providing enhanced shoulder or in-lane delineation and marking for sharp curves, providing enhanced pavement marking, providing skid-resistant pavement surfaces, widening and/or paving shoulders, and eliminating shoulder drop-offs.

Place Utility Lines Underground

Placing utility lines underground will greatly reduce the utility pole crash potential because no poles exist unless there are streetlights along the road. In addition, this countermeasure saves the costs of removing and replacing a pole damaged in a crash and of repairing the utility lines after a crash. However, the initial cost of placing lines underground is usually very high. The full safety benefit cannot be achieved if other existing above ground structures such as transformers and switching cabinets pose a hazard to vehicles. There are also other disadvantages related to the countermeasure. For example, the underground line is vulnerable to excavation damage, the carrying capacity on underground lines is limited, and the underground line is more difficult to patrol in the case of an outage.

Relocate Utility Poles to Locations Less Likely to be Struck

Some locations are vulnerable to utility pole crashes, and an effective countermeasure is to alter the pole position to avoid these utility pole crash prone locations. A typical example of crash prone locations is the outside of horizontal curves as there are many more run-off-road crashes on the outside of horizontal curves than on the inside. Other utility pole crash prone locations as presented in *Utilities and Roadside Safety* include, the gore of highway diverging areas; downstream of a lane drop or the area where the roadway narrows; traffic islands or in the median of divided roadways; and the critical quadrants of an intersection that is intersected by two roads with significantly different speeds (TRB, 2004). An illustration of the crash prone locations to utility poles is shown in Fig. 1.

Increase Lateral Offset of Utility Poles

This countermeasure is to address the “Offset” factor in utility pole crash problems. Increasing the lateral offset of utility poles is an attractive countermeasure, since both the frequency and severity of utility pole crashes will decrease when poles are moved farther away from the roadway. Ideally, utility poles should be placed at the right-of-way line and outside the clear zone. This means the lateral offset should be at least 2.0 m from the edge of the traveled lane of rural highways and 0.5 m from the face of curb of urban streets.

Practically, the right-of-way is usually expensive to acquire and not always available, which makes this countermeasure heavily site dependent. As well, the full safety benefit of the countermeasure cannot be achieved if other fixed objects remain in the clear zone.

Decrease the Number of Utility Poles

The countermeasure is to address the “Density” factor in utility pole crash problems. Decreasing the number of poles along a road section indicates fewer obstructions that are less frequently being struck and larger openings that allow a vehicle to pass through. If there is a wide enough recovery area behind the poles, the larger openings allow a vehicle more space to recover.

There are several methods to decrease the number of utility poles along a roadway. For example, encouraging the joint use of existing poles, with one pole carrying electric power, telephone, and other utility lines; placing poles on only one side of the road; increasing pole spacing by extending span lengths for utility poles. Increasing the pole spacing or placing poles on one side of the road may require using larger poles. These larger poles, however, may cause an increase in the severity of crashes because of their larger size.

Use Breakaway or Energy Absorbing Poles

To use breakaway or energy absorbing poles is an alternative treatment for utility poles that have to remain in place. Both types of poles are directed at reducing the severity of utility pole crashes. However, they are different in operating principles. Breakaway poles are designed to break away upon impact and swing out of the path of the vehicle. Energy absorbing poles are designed to “capture” the vehicle and stop it gently enough so that velocity change and deceleration do not exceed requirements established for the safety of a vehicle’s occupants.

The breakaway design has shown its effectiveness in reducing crash severity in sign supports and luminaire supports, and it has also been applied to utility poles and proved to be effective in some countries. However, unlike the signs and luminaires, a utility pole is not an isolated element, and there is concern regarding the “domino effect” of a single weak pole causing many poles to fall. Considering that the final position of the pole and conductors (wires) may create hazard for pedestrians, the breakaway pole is also not suggested in an area of significant pedestrian activity such as an urban area. Instead, the energy absorbing poles that can

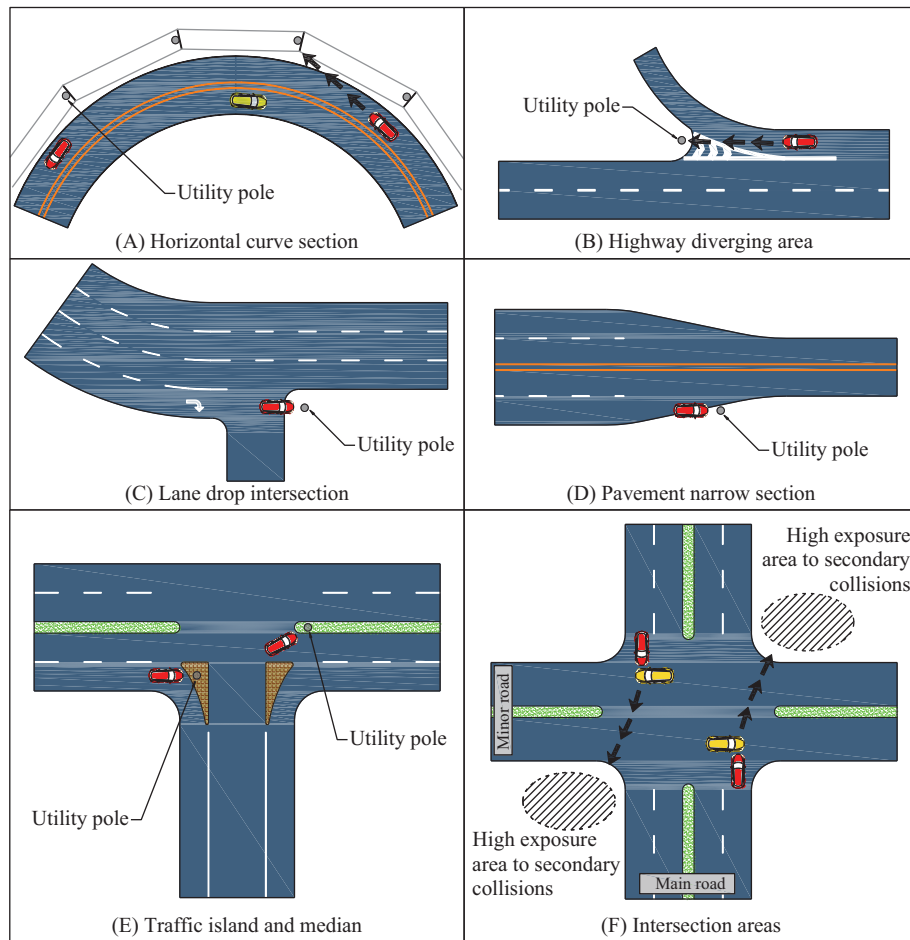


Figure 1 Utility pole crash prone locations.

“capture” the vehicle and stop it from encroaching the pedestrian area or hitting other road hazards are preferred. The problem with the energy absorbing poles is that they can entrap the vehicle at high speeds causing yaw, which overturns the vehicle and results in serious bodily harm.

Shield Utility Poles With Safety Devices

For utility poles that cannot be removed or placed underground, a potential countermeasure is to shield them with safety devices, such as roadside barriers and crash cushions. A roadside barrier is a longitudinal system used to shield vehicles from natural or man-made hazards along the roadway, and widely used barriers include low profile barrier and W-beam guardrails. A crash cushion functions by collapsing upon impact and slowing the vehicle at a controlled rate, and examples include sand or water filled containers.

Using safety devices to shield utility poles can only reduce the severity of crashes, and it is usually considered after removal or relocation has been ruled out. An important reason is that the installation of these devices increases the probability of a crash occurring, by exposing a larger area to the errant vehicle, and being closer to the roadway than the protected utility poles. In addition, there are some instances in which installing safety devices are not appropriate. A common example is that there is not enough room between the utility pole and the traveled lane to allow the safety device working functionally.

Delineate Utility Poles in Hazard Locations

Another countermeasure for utility poles in hazardous locations is to make them more visible, and thus errant drivers are alerted to the presence of utility poles and may take necessary evasive actions if the vehicles are under some level of control. However, if the vehicle is out of control, the safety benefits of pole delineation would be minimal. The delineation can be done with reflective paint, sheeting, object markers placed on utility poles, and roadway lighting.

A major problem with this countermeasure is that its low cost may make it appear attractive. However, the effectiveness is questionable, and there may not be any real safety improvement after the treatment. Therefore, this countermeasure should be only considered when all other options are exhausted.

Countermeasure Selection

The aforementioned countermeasures have their pros and cons, and there is not such a one-fits-all countermeasure. Countermeasures should be selected depending on local conditions, size of the program, funds available, and other factors. The selection may be made through the judgment of an informed individual or a group of individuals, and the best way for selecting countermeasures is to perform a cost-effectiveness (or cost-benefit) analysis.

The cost-effectiveness evaluation involves comparing the costs of various countermeasures to determine the most effective use of limited funding. Costs include potential future crashes, initial construction costs, ongoing maintenance costs, and similar items. Benefits include a reduction in the number of crashes, reduced maintenance costs, and the salvage value of the facility at the end of useful service life. Most importantly, the reduction in the number of crashes is estimated based on safety performance functions of utility poles, as shown in Eqs. (1) and (7). Benefits and costs are compared to determine whether an improvement is cost-effective and to set priorities among different countermeasures or their combinations competing for limited funds. It is noted that the comparison should consider the period of time in which each countermeasure provides a benefit and thus costs and benefits should be annualized using discount rates. Examples of detailed cost-effectiveness analysis procedures can be found in the *Roadside Design Guide* (AASHTO, 2011).

Summary

Utility poles, which are placed near the road substantially affect the safe operation of roads because of their large number and high structural strength. Vehicles that run off the road and strike a utility pole tend to suffer large deformation, which frequently leads to serious or fatal injuries. Statistics have shown that utility poles are one of the most common fixed object struck among all the fixed objects.

The occurrence of utility pole crashes are influenced by many contributing factors, and among them traffic volume, speed, lateral offset, and pole density are the most prevalent ones. Different statistical models are developed to relate utility pole crashes to these significant contributing factors. This article summarized the most widely used multiplicative model developed in 1984 and the safety performance function developed with the state-of-the-practice Poisson and negative binomial regression methods. The latter model also shows that the impact of density and offset on the crash frequency is more evident than the impact of the volume, and moreover increasing the pole offset has a more significant effect on reducing the frequency of utility pole crashes compared to increasing pole spacing. This finding suggests the importance of placing the utility poles outside the clear zone.

A comprehensive group of safety treatments for utility poles is summarized at last. These include countermeasures as following: keeping vehicles on the roadway, placing utility lines underground, relocating utility poles to locations less likely to be struck, increasing lateral offset of utility poles, decreasing the number of utility poles, using breakaway or energy absorbing poles, shielding utility poles with safety devices, and delineating utility poles in hazard locations. These countermeasures have their pros and cons, and they should be applied after a thorough consideration on local conditions, size of the program, funds available, and other factors. The cost-effectiveness analysis is suggested to aid the treatment selection.

See Also

Roadside Safety Barriers

References

- American Association of State Highway and Transportation Officials (AASHTO), 2011. *Roadside Design Guide*, Washington, D.C.
- Elesawey, E., Sayed, T., 2012. Evaluating safety risk of locating above ground utility structures in the highway right-of-way. *Accid. Anal. Prev.* 49, 419–428.
- Transportation Association of Canada (TAC), 2017. *Geometric Design Guide for Canadian Roads*, Ottawa.
- Transportation Research Board (TRB), 2004. *Utilities and roadside safety, State-of-the-Art Report 9*. Prepared by the Utility Safety Task Force, Committee on Utilities (A2A07), Transportation Research Board, National Research Council, Washington, D.C.
- Zegeer, C.V., Parker Jr., M.R., 1984. Effect of Traffic and roadway features on utility pole accidents. In: *Transportation Research Record 970*. Transportation Research Board, National Research Council, Washington, D.C, pp. 65–76.

Further Reading

- Fox, J.C., Good, M.C., Joubert, P.M., 1979. *Collisions With Utility Poles-A Summary Report*. University of Melbourne. Commonwealth Department of Transport, Victoria, Australia.
- Gabler, H.C., Gabauer, D.J., Riddell, W.T., 2007. *Breakaway utility poles. Feasibility of energy absorbing utility pole installations in New Jersey*. Report No. FHWA-NJ-2007-018, Department of Transportation, New Jersey.
- Lacy, K., Seinivasan, R., Zegeer, C.V., Pfefer, R., Neuman, T.R., Slack, K.L., Hardy, K.K., 2004. *Guidance for implementation of the AASHTO strategic highway safety plan-Volume 8: A guide for reducing collisions involving utility poles*. NCHRP Report 500, Transportation Research Board, National Research Council, Washington, D.C.
- Thome, J., Turner, D., Lindly, J., 1993. *Highway/utility guide*. Report No. FHWA-SA-93-049, Federal Highway Administration, Washington, D.C.

Value of Life and Injuries

David A. Hensher, Institute of Transport and Logistics Studies, The University of Sydney Business School, Sydney, NSW, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	737
The Value of Fatal Risk Reductions in a Road Safety Context	737
Making the Model Operational	738
Aggregating Individual WTP to the Population	739
Altruistic Versus Individual Values	740
Concluding Comments	741
Acknowledgments	741
References	741
Further Reading	741

Introduction

One important conceptual advance in the state of practice of road safety valuation was achieved in the 1980s by valuing road safety according to subjective preferences rather than by using the heavily criticized *human capital* (HC) *approach*. The HC rests on accounting principles: the benefit of avoiding a premature death is given by the present value of the income flow the economy could lose in that case. More appropriately, the *value of risk reductions* (VRR)—initially known as the *value of a statistical life* (VSL)—is based on subjective preferences and given by the amount of money that individuals are willing to pay for reducing the risk of their premature death while performing a certain risky activity. The focus on the VRR in contrast to the HC value yields higher benefits for risk avoidance, and hence the social net benefit of safety policy measures has increased in recent years, prompting many road safety interventions, otherwise not socially profitable, in the developed world (Viscusi et al., 1991).

The VRR for road contexts was estimated originally using contingent valuation (CV), standard gamble, or the chain method (Schwab and Soguel, 1995; Hausman, 1993), but the approach, in general, was heavily criticized by specialists in human behavior and economics. In original studies, people were confronted with situations expressing risk as tiny probabilities, and needing a trade-off between risk and money to arrive at a monetary value (Some of these studies posed a risk–risk trade-off. However, in order to arrive to a monetary value a risk–money trade-off is necessary, sooner or later). This kind of simulated context may not bear upon actual choices on route selection or travel behavior in general, where individuals have to consider a bundle of attributes describing each alternative (i.e., travel time, toll, and safety in a route choice context).

Some authors have proposed a different approach based on stated choice (SC) techniques (Rizzi and Ortúzar, 2003; Hensher et al., 2009; Iragüen and Ortúzar, 2004). In an SC survey, individuals are asked to choose among different alternatives, the attribute levels of which vary according to a statistical design aimed at maximizing the precision of the estimates; as such, SC allows the analyst to mimic actual choices with a relatively high degree of realism, and for this reason most experts believe that it is an appropriate elicitation method for the valuation of intangibles (Louviere et al., 2000).

The Value of Fatal Risk Reductions in a Road Safety Context

To understand the interpretation of the willingness to pay (WTP) for the safety attributes, it is useful to provide some theoretical background to the meaning of the evidence. We focus on a road setting. Assume a route, used by N users. If person n travels more than once in a reference period, say m_n times, this gives rise to m_n pseudo-members with a total population of $N = \sum_{n=1}^N m_n$, observations, that is, the individuals of a population. This population exactly amounts to the flow on a route in a given period (say a year). Bear in mind though that a population is a stock variable whereas a flow is not. A route is defined as a path connecting one origin–destination (O–D) pair. A trip on a route provides a level of dissatisfaction (or disutility) given by a deterministic indirect utility function $V = V(r, c, t)$, where r stands for risk of a fatal crash or class of injury, c for the cost of travel, and t for the travel time (mean and variability) on a route; there could be more attributes, of course.

Focusing only on fatality, we can formally define VRR as the value of avoiding one expected death, and this corresponds to the population (or sample) average of the marginal rate of substitution (also known as WTP) between income and risk of death for person n (MRS_n) plus a covariance term that accounts for possible correlation between MRS and reduced risk (δr_n):

$$MRS_n = \frac{\partial V_n / \partial r}{\partial V_n / \partial c |_{V=\bar{V}}} \quad (1)$$

$$VRR = \frac{1}{N} \sum_{n=1}^N MRS_n + N \text{cov}(MRS_n, |\delta r_n|) \quad (2)$$

where $\text{cov}(MRS_n, \delta r_n) = \sum_n MRS_n \delta r_n - \frac{1}{N} \sum_n MRS_n \frac{1}{N} \sum_n \delta r_n$

In empirical work, it is typically assumed that there is no correlation between WTP and δr in the population. Then, Eq. (2) simplifies to Eq. (3), and to estimate the VRR, it is sufficient to have a good estimate of the MRS (or WTP). This assumption would be correct, for example, if δr is the same for every individual.

$$VRR = \frac{1}{N} \sum_{n=1}^N MRS_n \quad (3)$$

The MRS can be interpreted as an implicit value for the traveler's own life, and averaging it over all individuals traveling on the route yields the VRR. The MRS clearly depends on personal risk perceptions according to the functional form of V_n . The same analysis can be carried out in terms of fatal crashes, f , (or injuries), instead of risks, r . However, in this case, the VRR is derived differently (but yields the same value):

$$VRR = \frac{1}{e} \sum_{n=1}^N \frac{\partial V_n / \partial f}{\partial V_n / \partial c |_{V=\bar{V}}} = \frac{1}{e} \sum_{n=1}^N SVCR_n \quad (4)$$

where e represents the number of fatalities or injuries (by class) per crash (effectively the information that relates to exposure to risk) and SVCR stands for the subjective value of fatal crash injury (by class) reductions, and is a Lindahl price. Eq. (4) embodies the definition of community WTP for a public good, road safety in this case, as the sum of individual marginal rates of substitution between income and number of fatalities and injuries (by class).

As a context example, if it is thought of in terms of a hypothetical tolled route, whose operators were able to extract the full consumer's surplus, the SVCR would be the maximum toll increase due to a safety improvement for individual n , such that they are as well-off as before the improvement. If the VRR is higher than the cost of reducing one fatality or one injury (by class), the safety project should be desirable from the community standpoint; in what follows it will be assumed that e is equal to one.

It can be seen that one advantage of dealing with the variable number of fatal or injury crashes, rather than risk, in empirical work. From Eqs. (2) and (4), it follows that

$$VRR = \sum_{n=1}^N SVCR_n = \frac{1}{N} \sum_{n=1}^N MRS_n + N \text{cov}(MRS_n, |\delta r_n|) \quad (5)$$

In other words, estimating the SVCR and aggregating across individuals will yield the correct VRR irrespective of the value of the covariance and this follows from the definition of our public good; one statistical death (or injury class reduction) reduction (per unit of time) on a particular route. (A statistical death reduction means saving one life, on average, per unit of time (obviously whose life is saved is unknown). This suggests that to elicit the VRR, rather than asking people to place a value on risk reductions, they should be asked to value a reduction in fatal or injury class crashes; it is believed this task is far easier from the respondents' standpoint. The two approaches are mutually consistent only when respondents have the correct aggregate flow in mind (i.e., they would value an extra fatal or serious injury crash per year different if they were to make the only trip on that road that year, than when millions of trips would be made on that road). In this sense, although a formulation in terms of number of crashes may sound more natural and easy-to-understand than a formulation in terms of probabilities to most respondents, the cognitive burden may not become any lighter. This is the basis of the popular WTP approach and reflected in the outputs derived from a discrete choice model.

Making the Model Operational

The model can be made operational within a (binary) discrete choice framework where the indirect deterministic utility of each available alternative j is

$$V_j = \alpha f_j + \eta \text{MajI}_j + \theta \text{MinI}_j + \phi \text{IncapI}_j + \beta c_j + \gamma \text{meant}_j + \lambda \text{stdev}_j + \text{other influences } (j = 1, 2) \quad (6)$$

where f is the number of fatalities, MajI is the number of major injuries, IncapI is number of incapacitating injuries, MinI is the number of minor injuries, t is trip time, and c is trip cost. The SVCR is equal to α/β for fatalities for every individual, η/β for major injuries, θ/β for minor injuries, and ϕ/β for incapacitated injuries for every individual. The definition of injury and its classification tends to vary by the geographical jurisdiction and is often in disagreement.

Also note that by computing γ/β , the behavioral value of travel time savings (VTTS) is obtained and λ/β is the value of reliability (VoR), two other useful outputs that relate to trading of the features of a trip setting. This model form is couched in the context of a specific trip context where travelers face travel times, travel costs, and safety of the road environment (defined by a history of deaths and injuries). By varying the levels associated with different travel contexts, a rich array of data can be gathered that is the basis of identifying traveler preferences for each of attributes being investigated. Simply put, the variability in an attribute level within a package of trips provides the necessary data to reveal, through choice model estimation, the contribution of each attribute (i.e., its marginal utility or disutility) to overall utility which is the representation of preferences. By comparing the marginal (dis)utility of an attribute of interest (e.g., travel time) to the marginal (dis)utility of cost an estimate of a WTP for a unit of the attribute of interest in dollars can be obtained.

For safety, the estimation of the choice model (Hensher et al., 2015) provides parameters for the safety attributes that are used to obtain the WTP to avoid a fatality or a class of injury in a road environment. To allow for preference heterogeneity, the number of deaths and injuries by category can be estimated with random coefficients. As the modeler does not possess all the relevant information, he must assume randomness in the utility function. Random utility, U , is simply expressed as the sum of two terms: deterministic utility, V , and a random component ε :

$$U_i = V_i + \varepsilon_i \quad (7)$$

It is assumed that each alternative i has a probability of being chosen, given by the probability that U_i is the highest random utility. A useful extension of Eq. (7) is to parameterize indirect utility according to socio-economic (SE) variables in order to attain what can be called as deterministic taste variations:

$$V_{ij} = \left(\alpha + \sum_l \alpha_l s_{lj} \right) f_{ij} + \beta c_{ij} + \lambda t_{ij} \quad (i = 1, 2) \quad (8)$$

where the binary variable s_{lj} represent the SE characteristic l of individual j , and i is the choice alternative (All level-of-service parameters can be tested for systematic variations in this form.). This is an interesting way of incorporating SE variables; with the advantage that if the information is used to estimate the SVCR, say, Eq. (8) states that there may be different coefficients for each attribute depending on the characteristics of the individual.

If the random terms are identically and independently Gumbel distributed (iid) between alternatives and across individuals, the popular multinomial logit (MNL) model is obtained and the probability of choosing alternative i is given by:

$$Prob(i) = \frac{\exp(\lambda V_i)}{\sum_{m=1,2} \exp(\lambda V_m)} \quad (9)$$

where λ is a scale factor inversely related to the unknown standard deviation of ε , and in practice has to be normalized to one as it is not identifiable (Hensher et al., 2015).

The MNL is not capable of addressing the issue that many responses may be provided by each individual in the SC case. We need a more powerful model that allows correlation among the choices made by the same person. The mixed logit model allows not only to do that, but also to consider random parameters. The mixed logit (ML) choice probabilities consist of a logit kernel that has to be integrated over the distribution of taste parameters:

$$Prob(i) = \int \frac{\exp(V_i)}{\sum_{m=1,2} \exp(V_m)} f(\alpha, \beta, \gamma | \Xi) d\alpha d\beta d\gamma \quad (10)$$

where $f(\bullet)$ represents the joint distribution of the parameters α , β and γ , and Ξ , the (usually unknown) population parameters that characterize the distribution.

Aggregating Individual WTP to the Population

The overall value for safety improvements (public goods) is a summation of individual WTP (WTP_i). The resultant figure is a reflection of what the safety improvement is “worth” to the affected group, relative to the alternative ways in which each individual might have spent his or her limited income. Research supports the WTP approach to the valuation of safety to determine the maximum amounts that those *affected* would individually be willing to pay for (typically small) improvements in their own and others’ safety.

For car drivers, the WTP is the vehicle driver’s marginal rate of substitution between income and number of annual road fatalities or number of serious injuries; and VRR is the summation of WTP (*separately for fatalities and serious injuries on a specific route*) over all drivers that traverse a specific road in a given year. Summing WTP values over all drivers (annual flow) on each route, we obtain two values—the societal value of fatality risk reductions (VF) and the societal value of serious injury risk reductions, that is, value of a severe injury (VSI). Note that the sources of societal aggregate WTP are many and varied. These include users’ WTP (including altruism if it exists) and nonusers WTP; with respect to the latter we would only count altruism (in the sense of taking care of others’ road safety). The WTP of nonusers is unlikely to be the same as users’ WTP (which includes the self-interest value). You cannot scale to the whole population from a subpopulation. To capture WTP of non-road users, one should consider a survey on nonusers to infer their WTP. Thus, multiplying drivers’ WTP by the total population rather than by the drivers’ population will overestimate the WTP for his segment.

With a focus on specific car trips, we have to convert the individual WTP to a *driver population exposure risk measure*. The traditional method based on the HC approach simply took the aggregate cost associated with all fatalities and serious injuries and paid no attention to the risk spectrum in which the cost of HC was linked. This link is critical to the validity of the societal WTP and cannot be disassociated from the specific level of risk associated with each trip.

The *exposure* of interest is reflected in the number of trips and associated kilometers undertaken by each driver in the population. The trip kilometers associated with driving have to be expanded up to the relevant population based on the number of times an individual in a subpopulation is exposed to risk. Identifying the actual amount of trip activity is crucial in aggregating up the average WTP per trip. The formulae for inputting the calculations for each risk class are:

$$\begin{aligned}\text{Societal } VRR_{rnf} (= VSL_{rnf}) &= \text{Societal } VF_r; r = \text{regions } 1, 2, \dots, R. \\ \text{Societal } VRR_{rvi} (= VSL_{rvi}) &= \text{Societal } VSI_r; r = \text{regions } 1, 2, \dots, R.\end{aligned}$$

The components of the calculation of the Societal VRR can be defined simply as WTP/chance , where *chance* is defined by the relationship between the *risk* as measured by the number of fatalities per annum or number of serious injuries per annum and *exposure* is defined by the annual number of vehicle kilometers.

$$\text{Societal } VF = \frac{\overline{WTP \text{ per trip}_r}}{\overline{\text{Trip kilometers}_r}} \times \frac{AAVKM_r \times 365}{\# \text{ fatalities}_r} \quad (11)$$

$$\text{Societal } VSI = \frac{\overline{WTP \text{ per trip}_r}}{\overline{\text{Trip kilometers}_r}} \times \frac{AAVKM_r \times 365}{SI_r} \quad (12)$$

WTP is an average or median WTP per person per trip; the average number of fatalities or severe injuries is an average over a period, such as the last three 3–5 years; and the average annual daily vehicle kilometers (AAVKM), is also over the same period. The reason for 3–5 years is because accidents are very random in nature, so it is good to have averages over such a period of years. That figure would be more stable than the number in one specific year.

To illustrate how this formula will work, let us assume a representative driver WTP (averaged over all drivers in the sample in the r th region) of \$0.20 per vehicle kilometer. Let us assume from our population data in the r th region that the chance of death per annum associated with driving a car is three fatalities divided by 6,000,000-vehicle kilometers per annum. The VSL is the WTP of the representative car-driving trip divided by the chance of death. This is $\$2 \div [3/6,000,000] = \$400,000$. This is the sum that society would be willing to pay to reduce the risk by one statistical death.

One major challenge is in how to work with a road network where the risks vary quite substantially across the network, within each region. That is, where the ratio of accidents to flow or exposure per route differs. There are two main options: (1) to work with each route safety record, or (2) to group routes according to some characteristic and work with their aggregated road safety record (for example, route A: 1 death per year; flow, 2,000,000; route B: 2 death per year; flow 3,500,000; and aggregated: 3 death per 5,500,000 flow). In the latter situation, if routes are of different lengths, it may be necessary to standardize in terms of deaths and serious injuries per million vehicle-kilometers.

We suggest that a way forward for car drivers is to sample representative routes and obtain data on annual trip kilometers (based on number of trips and average distance if trip kilometers are not readily available) and accident record (road vehicle fatalities and serious injuries). As long as we have information on the amount of traffic on these roads, compared to the entire eligible regional network, we can use this data on exposure (annual vehicle kilometers) and risk (aggregated fatalities and serious injuries) to obtain the chance indicator for Eqs. (11) and (12).

Where we have different routes for different O–D trips, everything will work fine as per the text. But if you have a link as part of a route for different O–D trips, you will also have to know the proportion of trips for each O–D, because vehicle kilometers will differ. For example you have Route 5 joining Taree–Sydney–Wollongong. Someone going from Taree to Wollongong will traverse the link Taree–Sydney, so on that stretch of the road all drivers (whether they go to Sydney, Wollongong, or further south) will be exposed the same travel time and road safety conditions, but the individual when choosing her preferred route will see what each alternative route offers to accomplish the trip. So to expand, you need know the proportion of trips corresponding to each O–D, to be able to establish a WTP measure per kilometer. As long as you know that proportion, the methodology will work fine. An O–D matrix with its corresponding traffic assignment available from any past study would be useful.

Altruistic Versus Individual Values

Altruism is only valid to be counted if an individual's utility is a function of other people's safety $U = U(\dots; \text{others' safety})$, not if it is a function of other people's welfare $U = U(\dots; \text{other's } U)$. To the extent that people display "altruistic" concern (people's WTP for others' safety as well as their own) then *prima facie* one would expect that it would be appropriate to augment the WTP-based VSL and VSI to reflect the amounts that people would be willing to pay for an improvement in others' safety.

However, under some simple assumptions regarding the nature of people's altruistic concern for others' safety on the one hand and their material well-being on the other (the latter being reflected by their wealth or consumption), augmenting the VSL and/or VSI to reflect WTP for other's safety would involve a *form of double-counting* and "would therefore ultimately be unwarranted."

In Chilean surveys, it was found that if someone travels with someone else, his or her WTP is higher than that of someone traveling alone. It was also found that WTP was affected by whether the driver has children (independently of whether they are

traveling with children or not). Those two factors are important to control for because otherwise one could think of altruism when actually it is not present. Of course, for many drivers other passengers could be their children.

Concluding Comments

There is growing support for the WTP approach with an emphasis on ex ante identification of VRR in contrast to the ex post HC measure. In addition, there is support for assessing road safety benefits in the context of real travel settings such that we can identify the trade-offs being made between all of the attributes influencing specific trip-related decisions. The VRR studies find that the VRR associated with a fatality in western economies are in the range of \$US8–\$15 m. The evidence on VRR for various human injury classes (e.g., minor, serious, incapacitated) vary a great deal and appear to be typically in a range of \$US200,000–\$US1.5 m. There remains controversy on the definitions of injury classes and measurement of exposure (defined in terms of kilometers of travel and history of actual injury levels in various geographical jurisdictions).

Acknowledgments

I acknowledge the contributions of Luis Rizzi, Juan de Dios Ortuzar, and John Rose who contributed to the research synthesized in this article.

References

- Hausman, J.A. (Ed.), 1993. *Contingent Valuation: A Critical Assessment*. North-Holland, Amsterdam.
- Hensher, D.A., Rose, J., Greene, W.H., 2015. *Applied Choice Analysis*, second ed.. Cambridge University Press, Cambridge.
- Hensher, D.A., Rose, J.M., Ortuzar, J., deDios, Rizzi, L., 2009. Estimating the willingness to pay and value of risk reduction for car occupants in the road environment, (ARC DP 0770633; 07–09). *Transp. Res.* 43 (7), 692–707.
- Iragüen, P., Ortúzar, J., de, D., 2004. Willingness-to-pay for reducing fatal accident risk in urban areas: an internet-based web page stated preference survey. *Accid. Anal. Prev.* 36, 513–524.
- Louviere, J.J., Hensher, D.A., Swait, J.D., 2000. *Stated Choice Methods: Analysis and Application*. Cambridge University Press, Cambridge.
- Rizzi, L.I., Ortúzar, J., de, D., 2003. Stated preference in the valuation of interurban road safety. *Accid. Anal. Prev.* 35, 9–22.
- Schwab, C.N., Soguel, N. (Eds.), 1995. *Contingent Valuation, Transport Safety and the Value of Life*. Kluwer Academic Publishers, London.
- Viscusi, W.K., Magat, W.A., Hube, J., 1991. Pricing environmental health risks: surveys assessments of risk-risk and risk-dollar trade offs for chronic bronchitis. *J. Environ. Econ. Manage.* 21, 32–51.

Further Reading

- Beattie, J., Covey, J., Dolan, P., Hopkins, L., Jones-Lee, M., Loomes, G., Pidgeon, N., Robinson, A., Spencer, A., 1998. On the contingent valuation of safety and the safety of contingent valuation: part 1—caveat investigator. *J. Risk Uncertain.* 17, 5–25.
- Fischhoff, B., 1997. What do psychologists want? Contingent valuation as a special case of asking questions. In: Kopp, R.J., Pommerehne, W.W., Schwarz, N. (Eds.), *Determining the Value of Nonmarketed Goods*. Plenum, New York, pp. 189–217.
- Hojman, P., Ortúzar, J. de D., Rizzi, L.I., 2005. On the joint valuation of averting fatal victims and serious injuries in highway accidents. *J. Safety Res.* 36(4), 377–386.
- Jones-Lee, M., 1994. Safety and the savings of life. In: Layard, R., Glaister, S. (Eds.), *Cost-Benefit Analysis*. Cambridge University Press, Cambridge.
- Jones-Lee, M., Hammerton, M., Philips, P., 1985. The value of safety: results of a national sample survey. *Econ. J.* 95, 49–72.
- McFadden, D., 1998. Measuring willingness-to-pay for transportation improvements. In: Gärling, T., Laitila, T., Westin, K. (Eds.), *Theoretical Foundations of Travel Choice Modelling*. Elsevier Science, Amsterdam, pp. 339–364.

Effects of Weather

Maria Pregnolato*, **Amirhassan Kermanshah†**, **Wisinee Wisetjindawat‡**, *Department of Civil Engineering, University of Bristol, Bristol, United Kingdom; †Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, United States; ‡Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	742
Chapter Scope	742
Key Definitions	743
Method	743
Impact of Adverse Weather Events to Roads	744
Modeling Methods	746
International Case Studies	746
Flash Flood Impact to Roads in Europe (Newcastle, United Kingdom)	746
Flood Impact to Roads in Japan (Nagoya)	746
Flash Flood Impact to Roads in the United States (Various Cities)	747
Discussion and Research Challenges	748
Conclusion	749
References	750

Introduction

Urban Transportation Systems (UTSs) are the lifeblood of cities, supporting mobility of millions of people and providing the means for exchange of goods and connecting businesses. As they underpin the global supply chain, the efficient and sustainable performance of these systems is crucial in all urban, rural, and in-between spaces, particularly for emergency management (e.g., evacuation, rescue, and recovery).

UTSs are under increasing pressure from two fronts. First, due to mass in migration into cities, UTSs are under increasing overcrowding and use-related deterioration. Second, the impacts from weather-related hazards (like heavy rainfalls, flash floods, etc.) are likely to increase due to climate and socioeconomic changes. These hazards are especially disruptive for transportation systems, both at a financial and a societal level; for instance, Hurricane Sandy caused \$7.5 billion of damages to New York City's transportation system of New York City, while immobilizing access for millions of users.

Considering the magnitude of societal debilitation that these events can cause to cities, understanding the resilience of UTSs is one of the most important missions of local, national, and international governing bodies (Schweighofer et al., 2014). A resilient and safe transportation system aims to efficiently respond to gradual and sudden environmental changes, as well as adapt to emerging technologies in the fields of transportation, data science, climatology, and meteorology.

Transport is a challenging sector in the context of climate impact modeling because its components have unique vulnerabilities to changes in climate system, which may cause their failure or impairment (Faturechi and Miller-Hooks, 2014). Such components can respond to weather differently according to the spatial and temporal scale (Pregnolato et al., 2019). In addition, the answer of the components is not independent since UTSs are related by many interdependencies. In other words, a failure in one component can cause cascading failures that could be inter-sectoral. The interlinked nature of UTSs means that impacts in one location can affect other areas in various period of time.

Chapter Scope

The focus of this chapter is on safety of UTSs. In particular, the content was tailored to provide insights into the state-of-the-art regarding road transport performance in adverse weather conditions. After a general overview of all major impacts to UTSs due to natural and man-made hazards, the chapter emphasizes how roadway conditions are influenced by a variety of adverse weather events (e.g., snow, rain, heat, and windstorm) in terms of safety, service delivery, and accessibility. This chapter also provides the thresholds of adverse weather events that can impact roadway conditions in different countries around the world to give the readers a better understanding of geographical location importance in the context of road safety.

The chapter sections are: (1) an overview of the latest understanding on weather impact on transportation networks, specifically road systems; (2) overview of the modeling technique used in road transport sector and technological gains that improve how these models impact the study of these networks; (3) case studies of technological use in practical application. This chapter will also introduce new approaches and identifying challenges within transportation networks, putting them in the wider context of traditional approaches to study these systems. The topic is discussed by underlining multi-disciplinary cross-cutting issues to invite the reader to personal reflection and creative thinking.

Key Definitions

In recent years, transport resilience has turned into a rapidly growing topic with extensive specialized literature. The key definitions detailed below are used in the realm of transport resilience study (and beyond). It is worth noting that some of these terms have overlaps and interchangeable meanings depending on the projects and preferences of research scholars. Authors may differ in the use of terms; therefore, the definitions below represent the specific use of terminology in this chapter.

Risk. Combination of hazard, exposure, and vulnerability; it measures the likelihood of occurrence of an event, and the consequences on the object at risk.

Hazard. Natural or man-made event that threatens objects of interest, causing perturbation or disruption of its physicality and/or functionality. It is measured by specific parameters (intensity measures) and probability of occurrence (frequency). Usually its location, impact area, and severity cannot be predicted with certainty.

Exposure. Inventory of the objects of interest, including information, such as location, material, age, etc. It represents objective information, independent from the hazard considered. Such information is usually stored in (geo-)databases.

Vulnerability. It measures the object susceptibility to a certain hazard, given the exposure characteristics. It is usually expressed by damage (vulnerability or fragility) curves that express the negative consequences due to the impact.

Resilience. It is defined as the system's ability to resist and absorb the negative consequence due to a hazard. This includes potential adaptation or mitigation measures to support the persistence of the performance in non-favorable conditions.

Recovery. Post-event actions to restore functionality at (at least) pre-event levels. The cost and time of recovery are important parameters when planning and prioritizing those actions.

Serviceability. State or condition under which an asset is considered useful. The level of serviceability refers to the performance of the asset, from full service (100%) to non-serviceability (0%).

Method

UTSs consist of a range of different modes, namely water, road network, railway, and air (Fig. 1). All of these networks are vulnerable to a range of natural and man-made hazards that can be categorized into four main groups (Faturechi and Miller-Hooks, 2014):

1. Natural weather events (e.g., rain, flood, wind gust, heat, and snowfall)
2. Natural geological events (e.g., earthquakes, volcano ashes, tsunamis, and hurricanes)

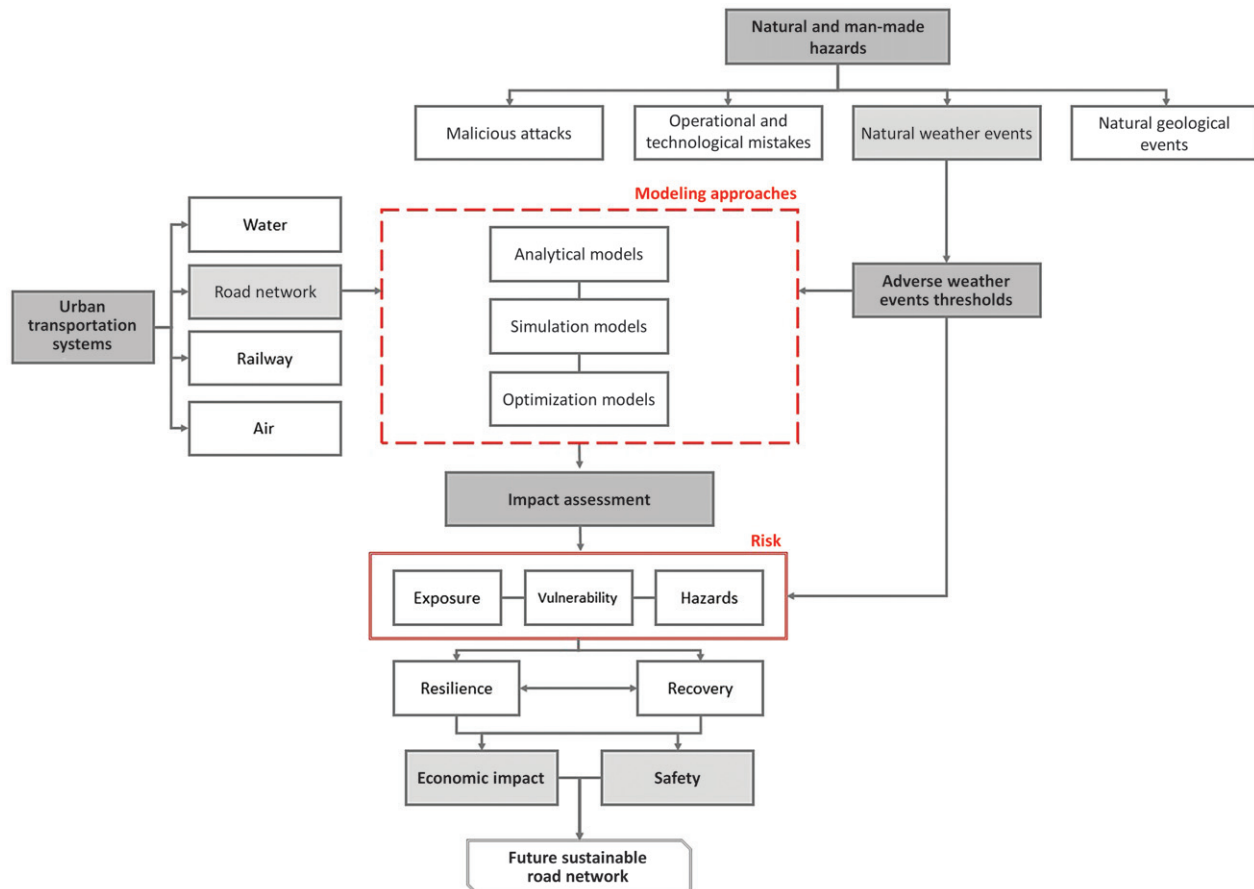


Figure 1 The range of different modes of the urban transportation system; this chapter focuses on road systems impacted by adverse weather events.

3. Operational and technological mistakes (e.g., hardware breakdown)
4. Malicious attacks (e.g., terrorism)

The risk from natural hazards encompasses the hazard characteristics (in both intensity and frequency), the asset exposure to that special hazard, and its vulnerability to that specific event (Grossi et al., 2005).

The vulnerability of the infrastructure asset depends on a series of factors (Pregnotato et al., 2016; Kermanshah et al., 2017): (1) the role of the links in the network (measurable by e.g., graph theory); (2) the exposure to the hazard; and (3) the number of users who rely on the asset. Identifying the areas with higher likelihood of hazard occurrence and higher consequences is fundamental for risk management. Such “hotspot” identification provides strategic information regarding urban dynamics and urban planning, which can be used to select critical links and target adaptation options. When dealing with infrastructure adaptation work, resources are limited, and prioritization is crucial. Within the network system, the degree to which the reduced performance of a link or node will impact the overall functioning of the network is called network criticality. This helps identify which network sections are likely to have large-scale impacts for society and economy, and thus are of critical importance to inform cost-effective adaptation strategies.

Within the context of road safety, adverse weather conditions are to be seen from the operational perspective as “atmospheric conditions unfavorable to optimal traffic conditions” (Schweighofer et al., 2014). Adverse weather events can cause direct and indirect damages to networks. Direct damages are those related to physical losses (e.g., damage to road surface); indirect damages relate to the loss of functionality, such as business interruption, service disruption, and lack of accessibility. The latter have a wider spatial and temporal scale, and are potentially more costly than the direct losses. Functional measures assess the effectiveness of a network (i.e., performance of the provided service) when it serves the community. More specifically, functional measures focus on serviceability of the transportation system, including travel time/distance, flow, and accessibility.

The journey time is usually considered the key output measure to assess the performance of a transport network, and transport modeling enables the estimation of travel times across a network (Pregnotato et al., 2016). Specifically, the (indirect) economic cost of the weather impact consists of estimating the delays in the traveling time of users, converted into monetary terms using the average value of time (VoT). The users’ VoT depends on many factors, such as the trip type (e.g., leisure/working), traveling conditions (e.g., mean, class), and the location (e.g., country, urban/rural). This appraisal could also include other metrics and quantify additional impacts, such as the increase in air pollution due to vehicle emission, or social impacts in terms of driver health and well-being. However, the appraisal of these metrics is difficult to be captured in economic terms. Certainly, more inclusive and refined calculations would imply extra complexities in the modeling. The complex nature of transport (and wider infrastructure networks) is typical for most developed countries, and implies unique demands for modeling and analysis.

Impact of Adverse Weather Events to Roads

Precipitation, rain and snow, and ice represent the biggest risks of disruption due to slipperiness, loss of visibility, or vehicles maneuverability reduction (Schweighofer et al., 2014). High winds can cause significant disruption by blowing over vulnerable vehicles (e.g., high-sided vehicles, motorbikes, and caravans) and trees. Flooding is another major cause of disruption, resulting in partial or total closure of lanes. Fog is hazardous to traffic since it reduces visibility and increases the risk of vehicle accidents and congestion. High temperature does not usually cause disruption to roads; however extended hot periods can lead to uneven road surface (due to thermal expansion), which require repairs, and may lead to indirect damages due to roadworks. In the period 2010–13 in the United Kingdom, approximately 60% of highway disruptions were related to high wind speeds, and approximately 40% were related to flooding; at local level, flooding is the most widespread problem, since disruption can last several days. Between 1996 and 2010, six US states, experienced increased rates of car crashes and injuries at a rate of 10% and 8%, respectively during rainfall days. These numbers increased to 51% and 38% during days that had over 50 mm of rainfall.

As a practical approach, precautionary road closures can be implemented to avoid incidents. Various agencies have different regulations, but their decisions are usually based on expert judgment and thresholds. Thresholds define the upper limit of the level of service a network can guarantee while preserving safety. Through review of available literature (e.g., Schweighofer et al., 2014), a series of thresholds for adverse weather events impacting roads are proposed in Table 1, for the European Union, Japan, and the United States of America.

These thresholds closely relate to the geographical location and are sensitive to individual contexts; therefore, they are not fixed values and might differ from place to place. As a result, different countries developed their own standard. For example, considering the same amount of rainfall, Japan tends to set the speed reduction to be higher than other countries, since the country regards safety as the highest priority. We can see also that Japan has no concern on the ice temperature. According to the geographic location of the country, most areas in Japan rarely face a winter as bitter as -20°C . This is different in the United States where each state has totally different climate standards, due to the size of the country; hence determining common temperature thresholds is not possible. Generally, many US cities are susceptible to other natural hazards, such as hurricanes, storm surges from hurricanes, and sea level rise. Each of these natural hazards can have different impacts (e.g., sea level rise does not have instantaneous impacts) in different cities. For example, Miami (Florida) is likely to be the worst-affected city in North America, according to the latest climate change projections (more than 3°C temperature increase by 2100) happen, which will cause the city to be totally submerged; moreover, the barrier islands and geology is such that seawalls cannot be built in an effective way (e.g., as for New York). Hence, the country focuses more on adaptive policies and how to set standards on when to evacuate.

Table 1 Thresholds for adverse weather events for several areas (EU, Japan, and United States)

Hazard	EU		Japan		USA	
	Threshold	Impact	Threshold	Impact	Threshold	Impact
Rainfall (RF)	$0.16 \leq \text{RF} < 6.4 \text{ mm h}^{-1}$	Light rain, speed reduction range 4%–10%	$\text{RF} < 15 \text{ mm h}^{-1}$	Light rain, recommended speed 80 km h^{-1} .	$0.25 \leq \text{RF} < 6.35 \text{ mm h}^{-1}$	Light rain, speed reduction range 2%–14%, reductions in capacity 0%–15%
	$\text{RF} \geq \text{mm h}^{-1}$	Heavy rain, speed reduction range 10%–30%	$15 \leq \text{RF} \leq 20 \text{ mm h}^{-1}$	Heavy rain, recommended speed $60\text{--}70 \text{ km h}^{-1}$ Very heavy rain, recommended speed 60 km h^{-1} .	$\text{RF} \geq 6.35 \text{ mm h}^{-1}$	Heavy rain, speed reduction range 5%–17%, reductions in capacity 0%–15%
Snowfall (SN)	$10 \leq \text{SF} < 100 \text{ mm in 24 h}$	Slipperiness, increased car accident rate	$\text{RF} > 20 \text{ mm h}^{-1}$		$1.5 \leq \text{SF} < 12.7 \text{ mm h}^{-1}$	Light/moderate snow, speed reduction range 3%–10%, reductions in capacity 5%–10%
	$100 \leq \text{SF} < 200 \text{ mm in 24 h}$	Reduced friction, double car accident rate and delays	$\text{SF} \geq 10 \text{ mm}$	Slipperiness and increased accident rate.	$\text{SF} \geq 12.7 \text{ mm h}^{-1}$	Heavy snow, speed reduction range 20%–35%, reductions in capacity 25%–35%
	$\text{SF} \geq 200 \text{ mm in 24 h}$	Snow accumulation, road closed				
Ice (T as mean Temperature)	$-7^{\circ}\text{C} < \text{T} \leq 0^{\circ}\text{C}$	Ice formation, increased accident risk	na	na	na	na
	$-20^{\circ}\text{C} < \text{T} \leq -7^{\circ}\text{C}$	Reduced effect of salting, delays, increased car accidents				
Flooding (FD as Flood Depth)	$\text{T} \leq -20^{\circ}\text{C}$	Fuel problems, slipperiness				
	$0 \leq \text{FD} < 50 \text{ mm}$	Reduced speed	$0 < \text{FD} \leq 100 \text{ mm}$	Normal condition	na	na
	$5 \leq \text{FD} < 300 \text{ mm}$	Reduced speed, delays, congestion	$100 < \text{FD} \leq 300 \text{ mm}$	Reduced braking performance		
	$\text{FD} \geq 300 \text{ mm}$	Unsafe driving conditions, roads closed	$300 < \text{FD} \leq 500 \text{ mm}$	Engines stop, closed roads		
Wind Gust (WG)	$17 \geq \text{WG} > 25 \text{ m s}^{-1}$	Trees can fall onto roads	$\text{FD} \geq 500 \text{ mm}$	High danger	Wind gusts $\geq 20 \text{ m s}^{-1}$ in 36 h	Wind advisory: Unsecured objects may be blown around and reduced visibility
	$25 \geq \text{WG} > 32 \text{ m s}^{-1}$	Reduced visibility, fallen trees	$15\text{--}25 \text{ m s}^{-1}$	Reduced speed	Wind gusts $\geq 26 \text{ m s}^{-1}$ in 12–48 h	High wind watch: Vehicle instability, damage to power lines and small structures due to falling trees
	$\text{WG} \geq 32 \text{ m s}^{-1}$	Road closed, high number of fallen trees	$> 25 \text{ m s}^{-1}$	Road closed	Wind gusts $\geq 26 \text{ m s}^{-1}$ in 36 h	High wind warning: Vehicle instability and loss of control, difficult driving conditions

Note: For the United States, the values were converted from the US unit to International unit (na means not available).

Table 2 Three different approaches to model natural hazards and their consequences, with arguments for (pro) and against (cons)

<i>Model</i>	<i>Description</i>	<i>Pros</i>	<i>Cons</i>
Analytical	Identification of risk/failure states through logical structures (e.g., risk matrix)	Easy to apply	Not suitable for large-scale investigation
Simulation	Representation of potential post-disaster consequences given a range of starting conditions	Random large sample of scenarios	Offer separated decision strategies, sub-optimal for decision-making to due uncertainty
Optimization	Determination of the optimal network according to the maximization of a range of parameters	Capable of providing decisions to address multiple possible future damage events	Input-data demanding, high-level of sophistication

Modeling Methods

Since adverse weather events are unpredictable in their frequency, location, and intensity, risk-based modeling techniques have been developed in order to plan for likely consequences. Each community/sector faces differing threats and need strategies fitted to their context. For example, models can support carriers to anticipate and plan for supply chain resilience in the face of broad disruptions. Also, models are suitable for investigating the need of relocating highly vulnerable structures (e.g., bridges). Models are particularly useful for testing adaptation options (i.e., strategic interventions) to understand if investments for improving resilience will save money in the long term.

Different approaches exist to model natural hazards and their consequences (Table 2). These models can be categorized as: (1) analytical, (2) simulation, or (3) optimization models (Faturechi and Miller-Hooks, 2014). Depending on data availability, they can be applied to different hazards.

Analytical methods investigate potential failure and risk states using disaster probabilities and consequences through different forms of logical structures, for example, risk matrix. Simulation methods are scenarios-based methods in which a range of potential situations is generated to evaluate the degradation of the network (by comparing pre-and post-performance level of service). Monte Carlo simulation considers random damage states and probability of occurrence to generate a large sample of scenarios. Optimization models optimize one or more network performance function (e.g., traffic flow), whether they are deterministic (ignoring events probability) or stochastic (considering events probability). They are mostly employed for pre-disaster planning and post-disaster resource allocation. They are suitable to capture the non-recurrent and uncertain nature of disaster impacts, supporting decisions to address multiple possible future damage events.

A model can be obtained by combining different heterogeneous models; in this case, the single models are organized in modules with distinctive processes and input data. For exchanging data and working together, input and output data must be compatible between the models. User-friendly interfaces can be created in order to facilitate the adoption of the final model. All models should be validated in order to be a good representation of reality; however, this is usually a challenging task due to data availability and rarity of extreme events. Data from sensors or large database are a promising resource for this purpose.

International Case Studies

Three case studies are presented to illustrate the application of simulation models to assess adverse-weather impact to roads. Each study shows a combination of models to integrate hazard, exposure, and vulnerability aiming at identifying road links at risk and impact at network-level.

Flash Flood Impact to Roads in Europe (Newcastle, United Kingdom)

This case study shows the impact of heavy rainfall in terms of flood to the urban road network by means of a simulation model of a middle-size city in the United Kingdom, Newcastle-upon-Tyne (Pregolato et al., 2016).

In this study, a risk-based framework for assessing the disruption to the urban road network from flood events was presented. This framework coupled a flood and a transport model to assess the impacts of extreme rainfall events, and to quantify the resilience value of different adaptation options. This study showed that extreme flooding events can cause the increase of travel times by up to 50%, while adaptation options can reduce transport disruption from flooding by reducing delays in travel times up to 22% (Fig. 2).

The proposed method can be used to optimize investment and target limited resources at critical locations, providing means of prioritizing limited financial resources to improve resilience. This is particularly important, as flood management investments must typically exceed a far higher benefit-cost threshold than transport infrastructure investments. By capturing the value to the transport network from flood management interventions, it is possible to create new business models that provide benefits to, and enhance the resilience of, both transport and flood risk management infrastructures.

Flood Impact to Roads in Japan (Nagoya)

This case study shows the impact of torrential rain (typhoon) and the consequent flash flood on the urban road network in Nagoya, Japan (Wisetjindawat et al., 2018). The study uses a stochastic modeling approach that assigns a failure probability to each road segment based on climate-model outputs for the region, which is later used to determine the travel time reliability between any given

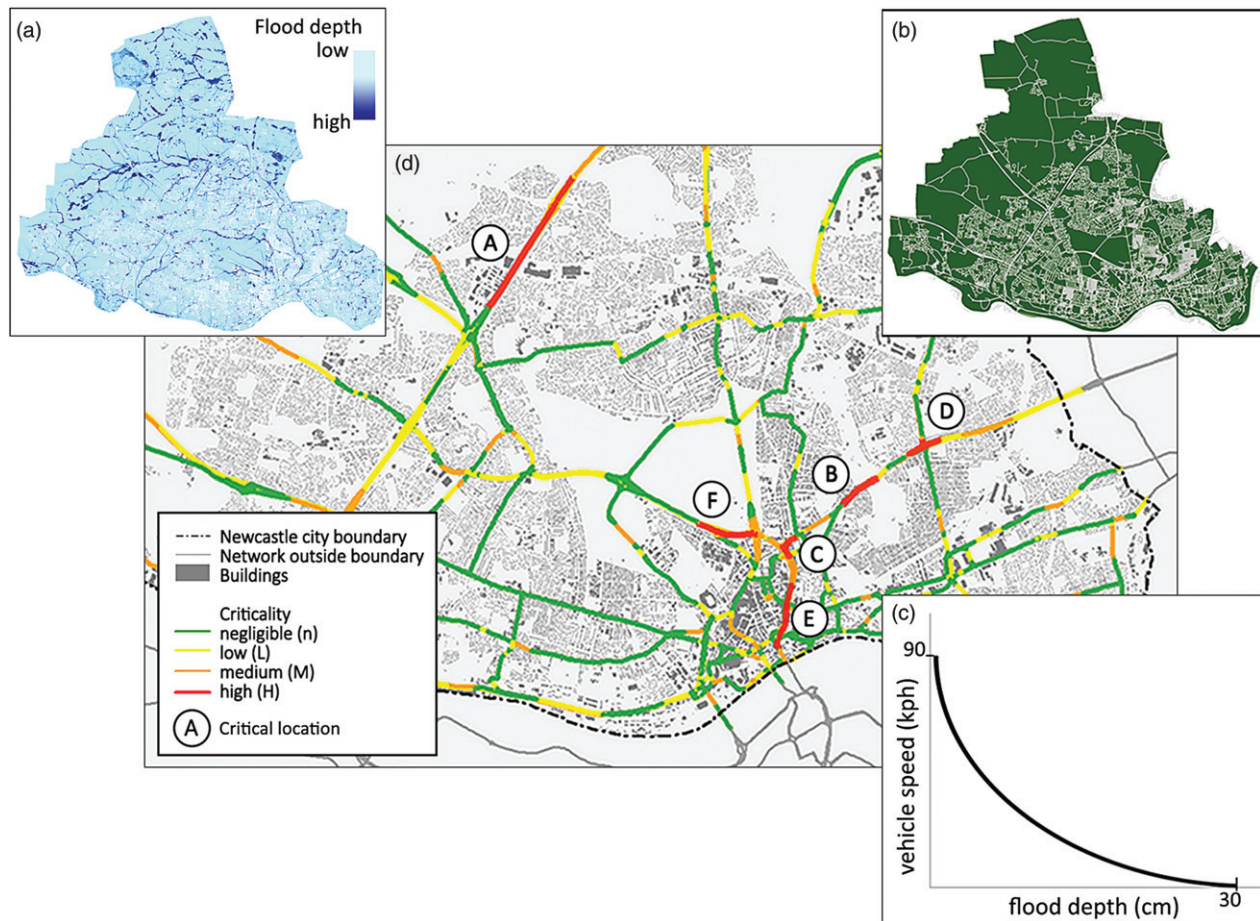


Figure 2 The impact of flooding on the road network of Newcastle upon Tyne (UK) (Pregnotato et al., 2016). (a) the flood hazard map for the city council area; (b) the road network; (c) the flood-transport curve that relates flood depth and vehicle speed during flooding; (d) flooding impact on the network as a result.

origin-destination pair. Travel time reliability represents the probability that a trip between a given origin-destination pair can be completed within a given time duration. For example, the simulation result estimates that there is a 70% probability that a trip between Nagoya city center and Tajimi can be made within a travel time 4-times higher than usual during the typhoon (Fig. 3).

The method can be used to evaluate the effectiveness of multiple policy schemes to reduce the impacts of an extreme weather event. For example, the model was used to evaluate two types of policy measures: (1) infrastructure management schemes (e.g., strengthening heavily used roads, thus reducing the failure probability of major roads) and (2) demand management schemes (e.g., changing departure time or adopting work-from-home schemes). Although it was found that strengthening important roads had a bigger impact in relieving congestion than delaying departure time, combining both strategies can better alleviate the congestion.

Flash Flood Impact to Roads in the United States (Various Cities)

This case study shows the impact of flash flood events on an urban road network considering changes in precipitation patterns under two different climate scenarios (high and low climate change scenarios) by simulating flash floods in five different cities in United States, namely: Boston (Massachusetts), Houston (Texas), Miami (Florida), Oklahoma City (Oklahoma), and Philadelphia (Pennsylvania) (Kermanshah et al., 2017). In this study, a vulnerability assessment framework for urban road networks against extreme rainfall-induced flash floods was presented. The framework coupled climate models, network modeling, GIS, and stochastic modeling to compile a "vulnerability surface" for each city in each climate scenario. More specifically, the impact of simulated extreme flash floods was assessed using percentage drops in static (i.e., overall properties and robustness topological indicators) and dynamic (i.e., accessibility and travel demand metrics) network properties and combined these metrics for each climate scenario onto one vulnerability surface. A vulnerability surface is the polygon created by a spider diagram that each axis is the percentage drop in one metric.

This study showed that flash floods and climate change are bound to seriously impact cities across the United States (Fig. 4). Generally, flash floods can have a significant impact on road networks and the magnitude of these impacts are mainly related to the geographic location of the cities, size of the road networks, and also the projection of climate scenarios. Moreover, different metrics can capture different resilience aspects of a road network. For example, in Houston resulted the most vulnerable city to flash floods in

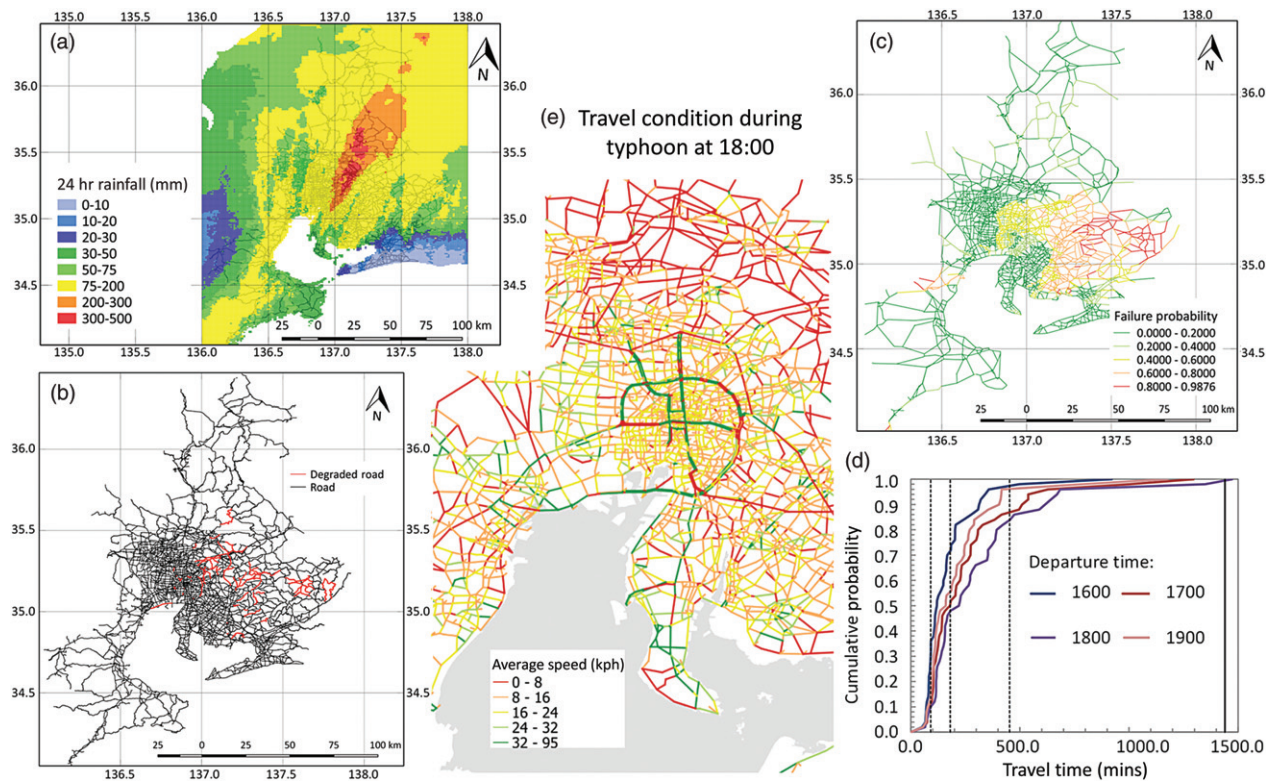


Figure 3 The impact of the simulated typhoon on roads and travel condition in Tokai Region (Nagoya, Japan) (Wisetjindawat et al., 2018). (a) map of 24-hr-typhoon rainfall; (b) observed degraded road network; (c) predicted road failure probability; (d) travel time reliability from Central Nagoya city to Tajimi city at different departure times; (e) average travel speed during the typhoon at 18:00.

high climate change scenario where on average the 8% of road networks were impacted causing more than 30% loss in long-distance accessibility in the city.

The proposed approach can be reproduced in different cities around the world, and researchers can use the results as guidelines for infrastructure design and planning purposes. From a policy perspective, governments can prioritize adaptation policies (e.g., infrastructure maintenance) in regions with a higher risk of vulnerability and allocate more resources to them. This can also reduce possible economic losses in those regions in future. Also, they can alter their urban infrastructure design criteria to adapt to climate change. For example, by changing the design criteria (by simulating green infrastructures such as street planters and permeable pavement) stormwater management strategies can be tested to reduce the catastrophic impacts of extreme flash floods.

Discussion and Research Challenges

The resilience of transport systems is highly dependent on geography. In the previous sections, we have learned how different countries assess impacts through thresholds and scenarios simulation. Each country develops its own standard, which is adapted to suit its specific geography and needs. Some countries, such as Japan, are prone to large-scale disasters, thus tend to focus more on evacuation plans. In the United Kingdom, floods happen more regularly but will less catastrophic aftermath; therefore, the focus is more on alleviating travel interruptions and severe traffic congestions. In the United States, because of the size of the country, each state has special weather condition and therefore, the weather hazards are totally different in each state and some regions experience natural hazards like hurricanes that will never happen in other countries.

In terms of modeling approaches, in addition to functional measures, topological measures from graph theory (e.g., connectivity and betweenness centrality) can be adopted to consider the network as a system of links and nodes. This method gives an insight into the weaknesses and is able to identify the important and vulnerable locations. In the third case study (United States), the authors adopted a graph theory approach to identify vulnerable locations in the system. More specifically, they calculated the changes in betweenness centralities (which can be a representative of the traffic volume), alongside with five other metrics like accessibilities, after the simulated flash floods to find out the vulnerable regions of the road networks.

Validation is a tricky stage of modeling, since data are not always available in the needed type, quality, or format. We recommend standardizing the data and collect it for a larger set of disasters, preferably in a GIS format to enable onward use.

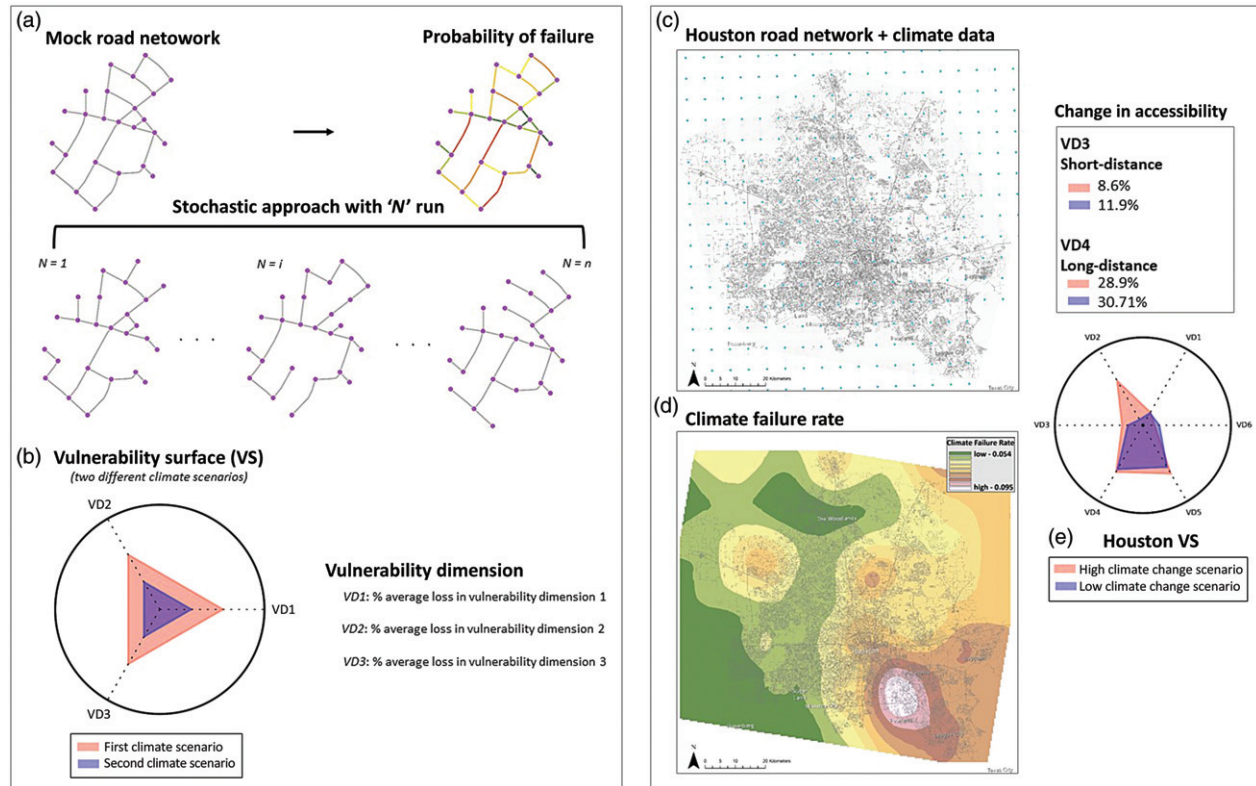


Figure 4 The impact of flash floods on the road networks in United States (Kermanshah et al., 2017). The left panel explains the framework used for the vulnerability assessment of road networks using (a) stochastic methodology for different climate scenarios; (b) the vulnerability surface method that combines different metrics in one single number. The right panel presents a case study on road network of Houston, Texas: (c) extracting precipitation data from climate models; (d) climate failure rate for the area; (e) resulting vulnerability surface.

Despite efforts in increasing infrastructure resilience, achieving a complete no-fail infrastructure is an impossible task. Infrastructure management is not isolated and limited to the physical assets, but imbedded in the complexity of the built environment and strongly linked to the users' demand and behavior. For example, most commuters tend to underestimate the risk of travel during adverse weather and continue their usual travel habits. Hence, along with improving the infrastructure, establishing an efficient communication and education on the risk of travel at different level of hazards with commuters is important.

Although this chapter focuses on the road system, transport sector widely embraces all the other means (such as rail, aviation, and navigation). Deep analysis should incorporate complexities of the real world, including the interdependencies underlying urban environment focusing on cascading effects and interconnected failures. For the same reason, studies are moving from a "single-hazard" to a "multi-hazard" approach, looking at the effects of compounding events (e.g., flooding that triggers landslides).

In the near future, autonomous vehicles are expected to share the transport infrastructure with other conventional vehicles and pedestrians; these vehicles are highly dependent on sensors and wireless communication between vehicles-to-vehicles and vehicles-to-infrastructures. Further studies should address how adverse weather might increase failures in vehicles' perception of obstacles or induce the likelihood of errors in the communication to design a robust system less vulnerable to environmental (cascading) impacts.

Conclusion

Up-to-date well-functioning transportation systems are fundamental for modern societies, who rely on infrastructure for everyday activities. Transportation systems are vulnerable to adverse weather events and the impact of such phenomena in cities can potentially cripple the flow of life for millions of people. Considering the magnitude of societal debilitation that these events can cause to cities, increasing efficiency and resiliency in urban transportation system is pivotal to protect our cities and surrounding areas.

This chapter provided an overview about transport resilience, illustrating the impact of adverse weather events to roads. Analytical, simulation, and optimization models were explained, as tools to study transport resilience problems. Three international case studies were presented as examples of application of simulation models looking at the impact of flooding on road networks in

different areas of the world. Such case studies could be extended by considering the interdependence with another infrastructure type, or the effect of technology change (e.g., autonomous vehicles) or multiple hazards.

In addition to flooding, winter storms can have a crippling effect on any transportation systems. Air and sea traffic, as well as railways and highways, especially in remote rural areas, can be closed down for hours or even days during times of heavy snowfall or ice events. The Chapter 10176: *Roadway Pavement Conditions and Various Summer and Winter Maintenance Strategies* discusses the safety effect of different winter-maintenance strategies for highways.

References

- Faturechi, R., Miller-Hooks, E., 2014. Measuring the performance of transportation infrastructure systems in disasters: a comprehensive review. *J. Infrastruct. Syst.* 21 (1), 1–15.
- Grossi, P., Kunreuther, H., Windeler, D., 2005. An introduction to catastrophe models and insurance. In: Grossi, P., Kunreuther, H. (Eds.), *Catastrophe Modeling: A New Approach to Managing Risk*. Springer, USA, pp. 23–42.
- Kermanshah, A., Derrible, S., Berkelhammer, M., 2017. Using climate models to estimate urban vulnerability to flash floods. *J. Appl. Meteorol. Climatol.* 56 (9), 2637–2650.
- Pregolato, M., Ford, A., Robson, C., Glenis, V., Barr, S., Dawson, R., 2016. Assessing urban strategies for reducing the impacts of extreme weather on infrastructure networks. *R. Soc. Open Sci.* 3 (5), 1–15.
- Pregolato, M., Jaroszewski, D., Ford, A., Dawson, R.J., 2019. Climate extremes and their implications for impact modelling in transport. In: Sillmann, J., Sippel, S., Russo, S. (Eds.), *Climate Extremes and Their Implications for Impact and Risk Assessment*. Elsevier, London.
- Schweighofer, J., 2014. The impact of extreme weather and climate change on inland waterway transport. *Nat. Hazards* 72 (1), 23–40.
- Wisetjindawat, W., Derrible, S., Kermanshah, A., 2018. Modeling the effectiveness of infrastructure and travel demand management measures to improve traffic congestion during typhoons. *Transp. Res. Rec.* 2672 (1), 43–53.

Wrong-Way Driving on Motorways

Huaguo Zhou, Md Atiquzzaman, Department of Civil Engineering, Auburn University, Auburn, AL, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	751
Contributing Factors	751
Safety Countermeasures	752
Signs, Pavement Markings, and Traffic Signals	752
Geometric Design Features	754
Intelligent Transportation Systems (ITS)	755
Systemic Safety Approach for Reducing WWD Crashes	757
Discussions	758
References	759
Further Reading	760

Introduction

Wrong-way driving (WWD) is the act of driving against the legal direction of traffic flow. This article covers only WWD on high-speed limited access facilities (e.g., freeways, expressways, and multilane divided highways), and that is what WWD stands for in the text. Although this type of event is rare, it is a serious traffic safety issue due to the severe outcomes of the crashes resulting from WWD. Typically, WWD crashes are head-on or opposite-direction sideswipe collisions at high speeds, which result in multiple fatalities and/or severe injuries. However, WWD crashes occur at low speed too, typically at cloverleaf ramps where only a yellow stripe separates traffic in the two directions. These crashes occur on what is part of the limited-access facility but of a different character—more similar to head-on collisions on rural two-lane highways—than what is the focus of this article. According to the WWD crash data, an average of 265 such fatal crashes resulting in 355 deaths occurred annually in the United States from 2004 to 2013 (Baratian-Ghorghi et al., 2014a). In terms of magnitude, WWD fatal crashes and fatalities are comparable to those occurring at highway-rail grade crossings (HRGX). However, unlike the HRGX experience, WWD has not been declining over the years. Fig. 1 compares the overall trends of all traffic fatalities to those that are specifically WWD-related in the United States over the 10-year study period. This figure illustrates that overall traffic fatalities declined substantially from 2004 to 2013, while WWD fatalities remained almost unchanged. A comprehensive national analysis has not been done for the years after 2013, but preliminary data indicate that the situation has not improved. There has been little sustained, coordinated national campaign to address WWD, which may somewhat explain this difference. It also suggests that such an effort could make a positive difference by helping to align WWD with overall national traffic safety trends. In other countries, WWD is also a serious issue of similar magnitude, and, in some countries, intentional WWD—the so-called ghost driving—seems to be more common than in the United States.

Fig. 2 depicts the annual average frequency of WWD fatalities across all 50 US states. In these illustrations, Group 1 includes 16 states with more than 2% of WWD fatalities, Group 2 represents states with less than 2% but more than 1%, and Group 3 consists of the states with less than 1% WWD fatalities (Zhou and Pour-Rouholamin, 2014a). Texas is ranked first with more than 15% of WWD fatalities occurring, followed by California and Florida. The three states with the highest frequency comprise approximately one-third of total WWD fatalities in the United States. Furthermore, these three states have put a considerable amount of efforts and investment in solving the WWD issue, given the significance of this problem in their jurisdictions.

Much of the focus of WWD research has been dedicated to their occurrences on freeways since this type of roadway has the highest speed and thus the outcomes of WWD crashes are most severe. As such, the discussion in the remaining part of this article primarily focuses on WWD crashes on freeways. Specifically, the discussion covers the following areas: (1) contributing factors to WWD crashes; (2) current traditional and advanced safety countermeasures to mitigate WWD crashes; and (3) future systemic safety approach to combat WWD.

Contributing Factors

The special investigation report of National Transportation Safety Board (NTSB) reported that impaired drivers constitute an alarming 60% of fatal WWD crashes (NTSB, 2012). This number might be even higher as some WWD drivers either refused to take sobriety test or had no test results recorded in the crash reports. NTSB also reported that the drivers who are older than 70 years have greater odds of being involved in WWD crashes than non-WWD crashes. Other studies found that most WWD events occur during the early morning hours (12–5 a.m.) of weekend days; this is possibly related to high alcohol consumptions at those times. In addition, poor lighting condition and severe weather may contribute to the occurrences of WWD crashes. The findings of the studies

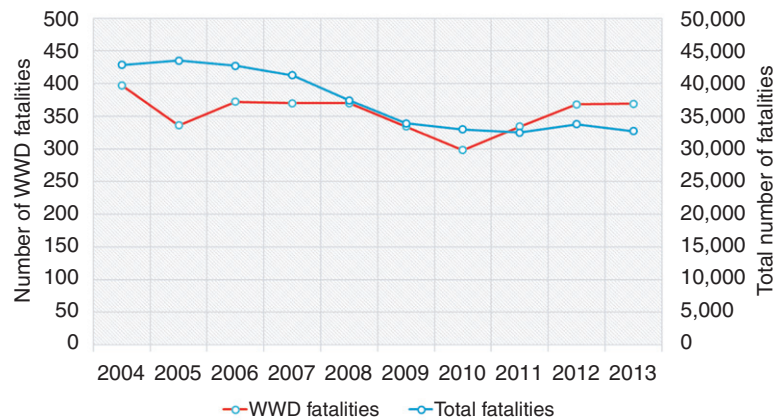


Figure 1 Overall fatalities versus WWD fatalities in the United States (2004–13).

conducted in California, Texas, Florida, North Carolina, Illinois, Alabama, Arizona, and other states are consistent to NTSB special investigation report (Zhou et al., 2015a, 2017a, 2017b; Ponnaluri, 2016; Lathrop et al., 2010; Braam, 2006). Most of the past studies on contributing factors to WWD crashes were based on the crash reports that provide little information on roadway conditions. Therefore, the results might be biased and focused more on the drivers' errors. There are a number of other factors that may increase the probability of WWD occurrence on a roadway facility. For instance, studies also found that inconsistency in location, angle, and size of wrong-way-related traffic signs, lack of pavement markings, and improper geometric design plays a significant role (Zhou and Pour-Rouholamin, 2014a). Typically, wrong-way drivers enter controlled access highways through their exit ramps. Particularly, the partial cloverleaf ramp terminals with parallel two-way ramps increase driver confusions and contribute to a major portion of the WWD events (Morena and Leix, 2012; Cooner et al., 2004). Although WWD movements sometimes originate from making U-turns on the mainline or at median crossover, they account for only a small portion of all WWD events. A study in Illinois showed that only 6.5% of 217 confirmed WWD crashes occurred from wrong-way drivers making U-turns on the mainline, while the other 93.5% likely occurred from drivers entering the freeway through exit ramps (Zhou and Pour-Rouholamin, 2015; Zhou et al., 2012). Therefore, the geometric design characteristics and usage of traffic control devices (TCDs) at the exit ramp terminals as well as along the ramps are critical in reducing the occurrence of WWD crashes on freeways. Although properly designed geometric features can physically obstruct the drivers from entering wrong way, proper use of wrong-way-related TCDs can help the drivers to take corrective measures if they mistakenly enter wrong way.

Safety Countermeasures

Over the last few decades, different engineering countermeasures have been proposed, implemented, and tested by various state and local transportation agencies to mitigate WWD activities, including (1) changing the size, location, and angle of wrong-way-related traffic signs, (2) proper use of conventional and innovative pavement markings, (3) implementing proper geometric elements (raised medians, channelizing islands, and control radii at intersections), and (4) application of intelligent transportation systems (ITS). The following sections briefly discuss these countermeasures and their potential application in combating WWD crashes.

Signs, Pavement Markings, and Traffic Signals

Signs are among the more traditional and least expensive WWD countermeasures. They are meant to guide, warn, or regulate drivers, and, in the WWD context, deter them from making wrong-way maneuvers at the intersections of exit ramps and crossroads. Proper signage can notify the wrong-way driver that she/he is traveling in the wrong direction should it occur. Commonly used wrong-way-related signs used in the United States include DO NOT ENTER, WRONG WAY, ONE WAY, keep right, and turn prohibition signs. The effectiveness of these signs greatly depends on the location, orientation, size, and nighttime visibility. The following general considerations are recommended to increase the effectiveness of WW-related signs:

- Ensure that signs are positioned to face, and are clearly visible to the driver for whom signs are placed.
- Place the first set of WRONG WAY signs within 200 feet from the exit ramp terminal, so that these signs are clearly visible to the drivers on the crossroad when they are approaching the exit ramp terminals.
- Consider optional use of oversized sign and/or supplemental signs to enhance their visibility and conspicuity.
- Evaluate the use of lower mounting height of DO NOT ENTER and WRONG WAY signs where appropriate.
- Consider various optional enhancements to increase conspicuity of signs (particularly at night), including red retroreflective strips on sign supports, flashing LED borders, or flashing beacon.

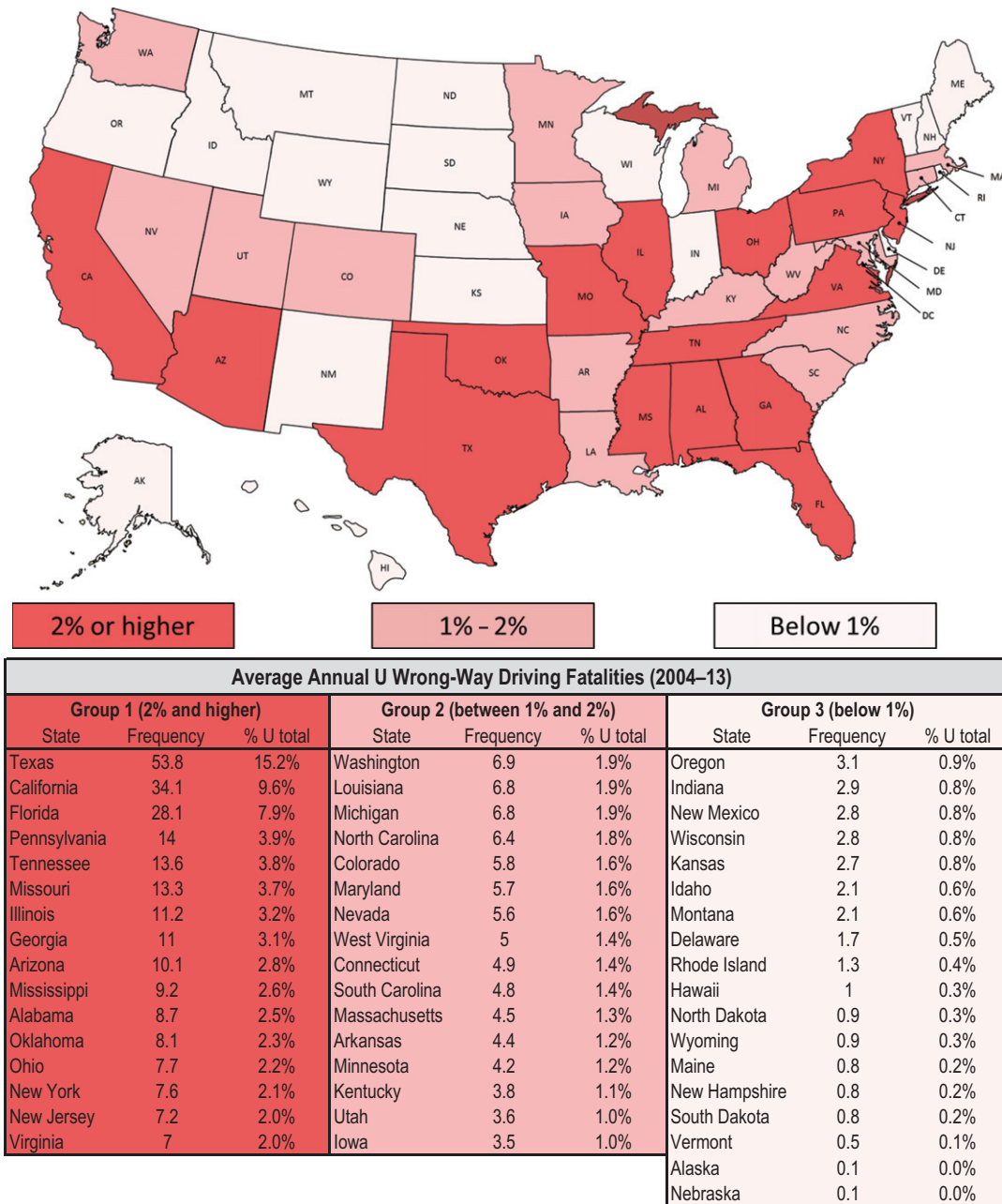


Figure 2 Percentage of WWD fatalities in each state (2004-13).

Pavement markings, like other TCDs, are a way of communicating with drivers and they can convey important messages about proper and safe road use. Types of pavement markings used to address WWD include longitudinal lines, lane line extensions, lane-use arrows, wrong-way arrows, and delineation. Detailed information about designing these arrows and their dimensions can be found in *Standard Highway Signs and Markings* by Federal Highway Administration (FHWA, 2004). Some key considerations for effective use of pavement markings to reduce WWD are as follows:

- Ensure that pavement markings complement the geometry of the location and provide positive guidance to drivers about proper direction and movement.
- Ensure that pavement markings and signs are consistent with one another so they reinforce the information conveyed to drivers.
- Consider various optional enhancements to increase conspicuity of pavement markings.

Most exit ramp terminals, if signalized, currently use circular green signals supplemented with No Right/Left Turn signs. Some locations use green arrow signals. Field observations show that green arrow signals send a more obvious message to drivers and can

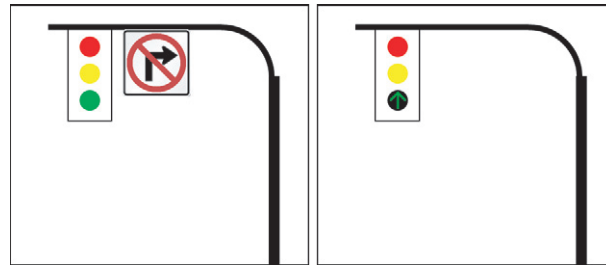


Figure 3 Analogous traffic signals for exclusive through lanes.

replace a circular green signal with a No Right/Left Turn sign. As shown in [Fig. 3](#), the combination of a circular *green* indication and a turn prohibition sign (No Right Turn shown in the figure) mounted on the mast arm traffic signal for an exclusive through-lane has the same meaning as a *green* arrow indication (through only).

The placement of WW-related signs/markings can vary considerably depending on the interchange types. For instance, the appropriate placements are different between diamond and partial cloverleaf interchange due to their difference in geometric configurations at the exit ramp terminals. Therefore, signs and pavement markings should be placed where it complements the geometry of interchange and provide positive guidance about proper direction and movement. A detailed guideline for appropriate application of WW-related signs and markings at different types of interchanges can be found in *Guidelines for Reducing Wrong-Way Driving Crashes on Freeways* by the Illinois Department of Transportation (IDOT) ([Zhou and Pour-Rouholamin, 2014a](#)) and *Manual on Uniform Traffic Control Devices (MUTCD) 2009* by FHWA (2009). A historical evolution on wrong-way-related TCDs in MUTCD was summarized by [Baratian-Ghorghi and Zhou \(2017\)](#).

Geometric Design Features

Past studies have indicated that certain interchange configurations and geometric design elements may be more susceptible to WWD and that minor geometric changes to ramps can reduce wrong-way movements onto freeways. In this section, geometric elements that are capable of discouraging wrong-way maneuvers are identified. Those elements should be given special considerations during the design stage of interchanges and ramp intersections.

Exit ramp terminals are highly critical points because they can be the entry locations from which wrong-way maneuvers originate. Their geometric characteristics (arrangement, angle with crossroad, cross-section, etc.) are of high importance. For instance, adjacent exit and entrance ramps (parallel, side by side) may be more prone to wrong-way maneuvers when there is a nearby side street. Therefore, care should be taken to consider how each geometric characteristic can make a location more susceptible or less susceptible to WWD. In general, use of acute angles to connect the exit ramps to one-way crossroads and obtuse angles to connect with two-way crossroads is recommended to better convey direction of travel. Use of sweeping connections to crossroads is reported to be less susceptible to WWD because they form acute angles with crossroads where turning movements in either direction are not allowed. Reducing the width of exit ramp throats by placing channelizing islands and increasing the width of entrance ramp throats by removing islands can be helpful in reducing WWD. Non-traversable raised median on the crossroad may help reducing left-turning wrong-way drivers from the crossroads. On the other hand, to reduce wrong-way right turns (particularly at diamond interchanges), a short-radius curve or angular break at the intersection of the left-edge of exit ramps and the right edge of crossroads is recommended. An open sight distance should be provided for drivers from crossroads to see the entire length of ramps. To ensure the visibility at night, uniform lighting levels should be provided for both entrance and exit ramps. At the intersection of two-way ramps and crossroads (partial cloverleaf interchanges), the stop lines for left turns from crossroads should be moved forward so the motorists have a better view of the entrance ramp and a shorter left-turning radius.

There are certain geometric features that are more susceptible to WWD, which should be avoided when possible to reduce the chance of WWD. Some of the examples may include: adjacent entrance and exit ramps intersecting a crossroad ([Fig. 4](#)), isolated exit ramps, left-side exit ramps ([Fig. 5](#)), one-way exit ramps connected as unchannelized T-intersections, exit ramps intersecting with two-way frontage roads, less common arrangement of exit ramps (e.g., buttonhook or J-shaped ramp, as shown in [Fig. 6](#), connected to a parallel or diagonal street of frontage road), temporary ramp terminals at work zones, freeway feeders (where exit ramps transition into local roads), side streets adjacent to exit ramps ([Fig. 7](#)), and scissors channelization ([Fig. 8](#)).

Interested readers are referred to the following publications for a more detailed discussion on designing geometric features for reducing WWD:

- *Guidelines for Reducing Wrong-Way Driving Crashes on Freeways* by the IDOT ([Zhou and Pour-Rouholamin, 2014a](#))
- *A Policy on Geometric Design of Highways and Streets—7th Edition* by the American Association of State Highways and Transportation Officials ([AASHTO, 2018](#))

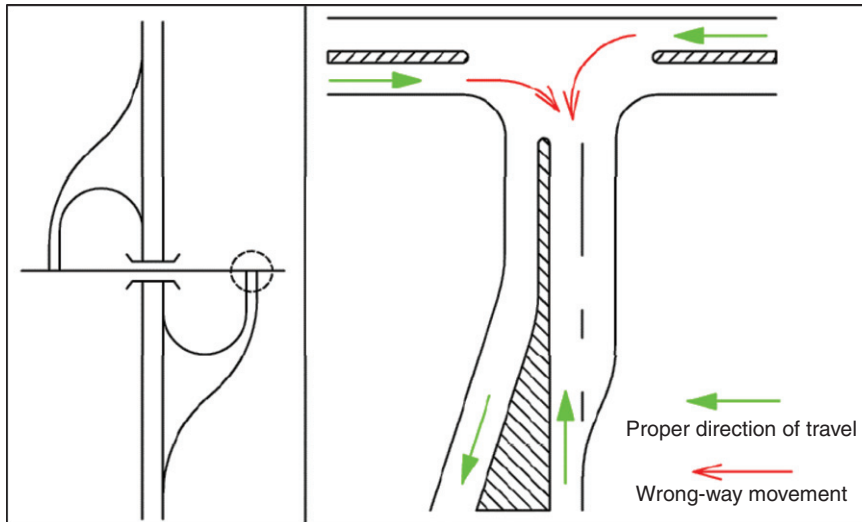


Figure 4 Partial cloverleaf interchange with adjacent exit and entrance ramps.

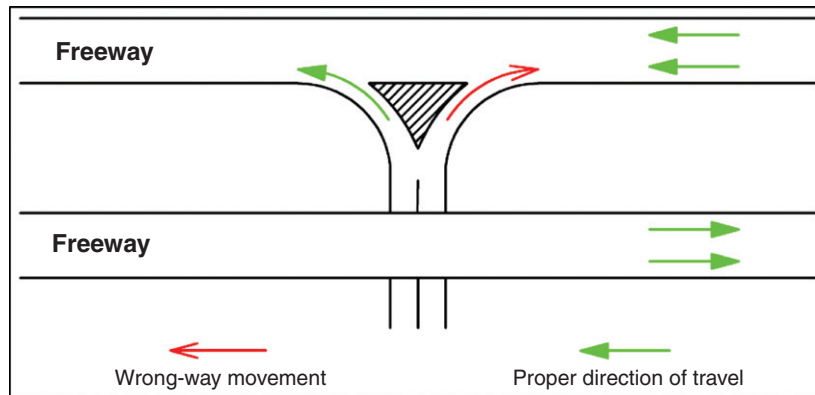


Figure 5 Left-side exit ramps.

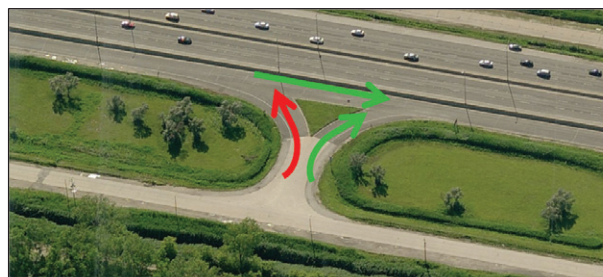


Figure 6 Buttonhook ramp connected to parallel to frontage road.

Intelligent Transportation Systems (ITS)

ITS technologies have been recently used by many transportation agencies to develop WWD countermeasures. The application of ITS in detecting and deterring WWD incidents consists of three steps: (1) detection, (2) warning, and (3) action. Detection step can be accomplished using a variety of different detectors such as inductive loop detectors (ILD), video image processing (VIP) systems, microwave radar-based traffic detection systems, infrared detection systems, and others. Regarding the warning step, we note that different methods can be employed to warn both wrong-way and right-way drivers. In-pavement warning lights, flashing wrong-way signs, warning lights, and dynamic message signs (DMS) are some examples of these warning systems. Action can be taken by patrol

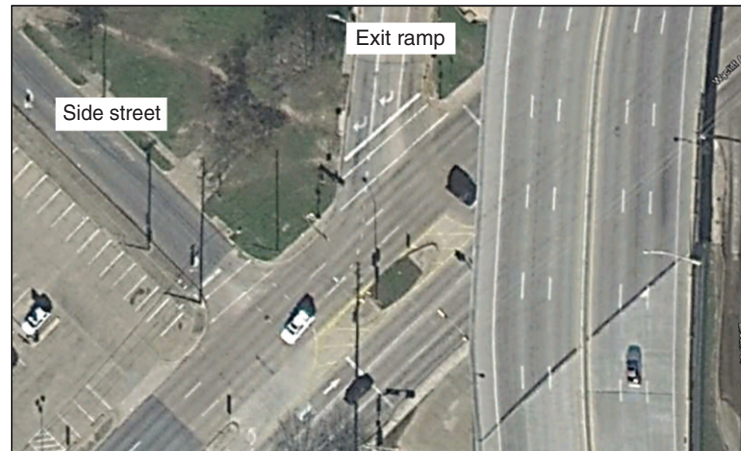


Figure 7 Side streets adjacent to exit ramps.

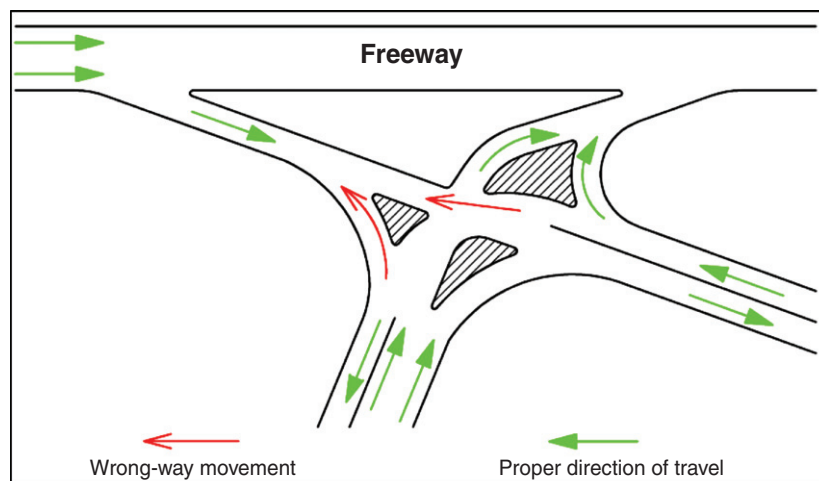


Figure 8 Scissors channelization with possible movements.

units or other responsible parties after receiving an alert from the traffic management center (TMC) or some centralized dispatches to intercept wrong-way drivers. A typical scheme of ITS detection and warning system for WWD can be found in [Fig. 9](#).

As shown in [Fig. 9](#), a WW driver is first detected using sensor technologies, and she/he is immediately notified of the mistake by means of warning devices such as LED WW signs. These work well in rural areas where modern TMCs and quick responses by police are not available. In some large metropolitan areas, TMCs can receive the signals from field detectors and sensors and further verify WWD using video cameras. After the incident is verified, law enforcement officers will be provided with the location and direction of the wrong-way driver through dispatch centers. Law enforcement officers and other incident responders can either help at-fault vehicles correct their directions or stop them. Meanwhile, pertinent messages can be posted on changeable message signs (CMSs) to notify right-way drivers of the presence of errant drivers.

Given that the vast majority of WWD crashes are caused by driving under the influence (DUI) of alcohol or drugs, the traditional WWD countermeasures might not be capable to reduce the crashes by drunk drivers. Therefore, WWD detection and warning systems may be more required. According to the past studies, only a small portion of transportation agencies exploited ITS technologies to identify WW drivers and to take prompt and proactive actions. Radar detectors, closed-circuit television (CCTV) cameras, and ILDs were used as popular detectors. After detection, flashing WW signs, warning signs, and DMS are available means of warning other drivers of an imminent danger ahead. Various messages may appear on DMS such as "Wrong-Way Driver Ahead" and "All Traffic Move to Shoulder and Stop." After detection and verification of the at-fault driver, patrol units may step in and position ahead of the WW driver to either help the driver pull over or to correct his or her direction. If it is not possible to position, responding units may attempt to intercept the vehicle by deploying tire deflation devices (e.g., portable spike) to slow or stop the WW driver or by using extra force to stop the vehicle.

In the United States, there are several existing systems that apply ITS technologies for WWD detection and warning. The Harris County Toll Road Authority ([HCTRA, 2012](#)) in Houston, Texas, implemented a radar-based WWD detection system. This system was

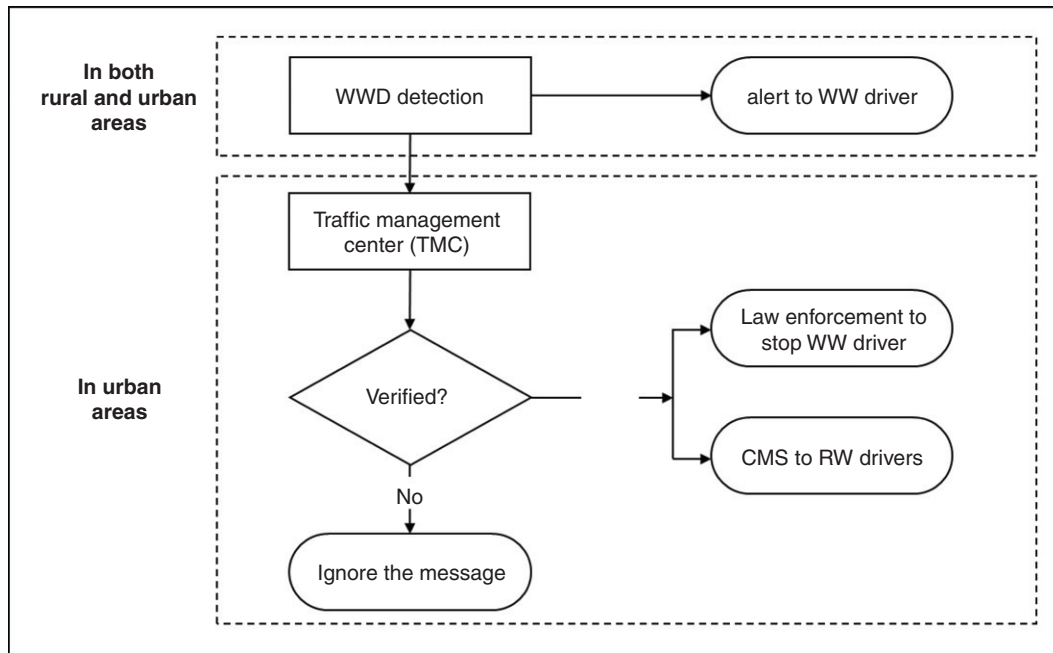


Figure 9 Typical scheme of WWD detection and warning system.

designed for 12 sites at exit ramps and along the mainline, all connected to the HCTRA TMC, at an overall cost of \$337,000. The New York State Thruway Authority (Thruway) implemented an ITS-based, wrong-way entry warning system at Porter Avenue (Exit 9) along I-190 in Buffalo, New York. The system consists of two major components: Doppler radar detection and programmable CMSs. Arizona Department of Transportation (ADOT) recently implemented a thermal-detection technology, the first-of-its-kind camera detection system, to mitigate WWD crashes (AZCentral, 2018). This system consists of 90 thermal-detection cameras placed on exit ramps and the mainline of I-17 in Phoenix, Arizona, at an overall cost of \$4 million. It takes a three-phase approach in case a WW vehicle is detected including alerting WW drivers thus they can self-correct themselves, warning right-way drivers, and informing law enforcement.

In addition to the traditional ITS application, several studies explored the application of connected and autonomous vehicle technologies, which uses vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communication, for developing wrong-way driver detection and warning system. In 2016, the Texas A&M Transportation Institute (TTI) developed an initial concept for a connected vehicle WWD detection and management system, which would detect WWD vehicles, notify the Texas Department of Transportation (TxDOT) and law enforcement personnel, and alert affected travelers (Finley et al., 2016). In 2007, a driver assistance program was developed in Germany, which uses the in-vehicle navigation system to automatically recognize when a driver enters the wrong direction in a roadway and provides a series of audible and visual warnings in the heads-up display (HUD) (Brigl, 2007). Using V2V communication, the program was also capable of warning other motorists within 2000 feet of the wrong-way driver. The V2I technology was used for communicating information to vehicles in a wider area, in which the wrong-way vehicle sends its position to a service center, and the service center disseminates this information to other vehicles.

Although not common in the United States, WWD originated from ghost driving has been critical safety issue in some European countries including Germany and France. By definition, ghost driving is the act of driving in the wrong direction to travel from one interchange to another interchange. Several detection and alert systems have been introduced to reduce crashes resulting from this type of WWD events. In France, when a ghost driver alert is sent out, roadway operators attempt to confine the problem promptly by activating electronic signs at roadway entry points and closing access to the road using physical barriers. In Germany, an electronic system is being experimented that will reactively warn drivers through visual and acoustic means of imminent threat. For further details on this topic, interested readers are referred to *Preventing and Managing Ghost-Driver Incidents: The French Experience* (Vicedo, 2007) and *Ghost Drivers: A Direct Experience of Toll Road Operators* (ASECAP, 2017).

Systemic Safety Approach for Reducing WWD Crashes

Previous literature shows some evidence that the WWD risks can be attributed to some common risk factors, which indicate that a proactive systemic safety approach can be developed to address the WWD issue on freeways. A systemic approach involves making improvements at locations with high-risk roadway features or high-predicted crash risk correlated with particular severe crash types, regardless of the actual crash history. It is a more proactive approach than a spot safety or a corridor retrofit approach that focuses

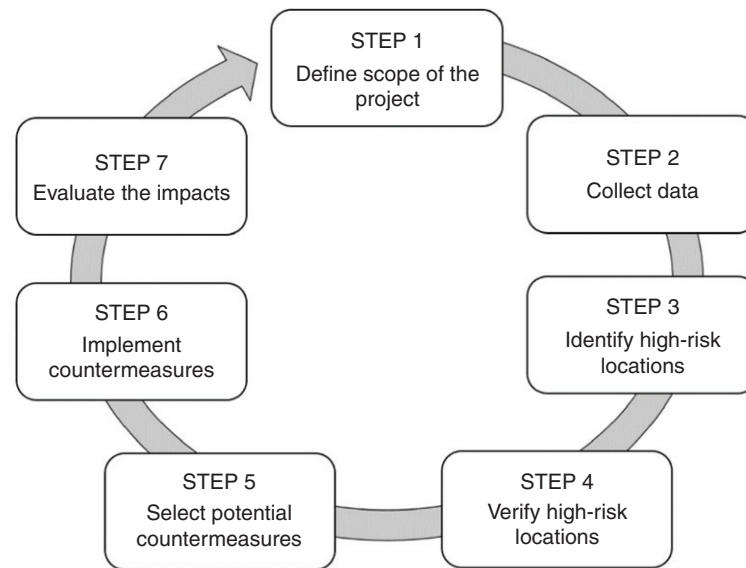


Figure 10 Systemic safety approach for reducing WWD crashes.

only on treating specific locations with a crash history and less costly than systematic approach that makes improvements at all sites in an area, regardless of predicted crash risk or crash history. Since WWD crashes are rare events, it is challenging to select hot spots for WWD countermeasures based on crash history alone. In addition, the selection of locations for WWD countermeasures based on engineering judgments alone may not be sufficient in most cases since the occurrences of WWD events are often resulted from the combined effect of a wide range of contributing factors. The complex relationship among all these contributing factors cannot be accounted by engineering judgments alone. Therefore, a systemic approach could be an ideal tool for transportation agencies to mitigate WWD activities within their jurisdiction by acting proactively and ensuring an efficient use of their limited funding.

A systemic approach to address the WWD issue on limited access motorways would involve the seven steps shown in [Fig. 10](#). Step 1 involves defining the project scope. Depending on the interest of a particular transportation entity, the project scope can be the exit ramp terminals along a specific segment of a freeway or all the exit ramp terminals along the freeways within their jurisdiction. After defining the scope of the project, steps 2–4 help to identify and verify the locations (i.e., the exit ramp terminals) having highest risk of WWD. Steps 5–7 involve selecting potential countermeasures to reduce WWD risk at those locations, implementing countermeasures, and evaluating their impacts on reducing the risk of WWD. The most challenging part of developing a systemic safety approach for WWD crashes is the identification of locations with high-predicted crash risk. As mentioned earlier, WWD crashes are rare events on our roadways, which make it difficult to develop predictive crash-risk models for these kinds of crashes. However, in recent years, a number of researchers made significant efforts to develop WWD crash-risk prediction models. The works of [Atiquzzaman \(2019\)](#), [Atiquzzaman and Zhou \(2018a\)](#), [Sandt et al. \(2017\)](#), [Rogers et al. \(2016\)](#), [Pour-Rouholamin \(2016\)](#), and [Baratian-Ghorghi et al. \(2014b\)](#) are some examples of efforts to develop WWD crash-risk prediction models. These studies explored the idea of predicting WWD risk at an individual exit ramp terminals or along a specific roadway segment based on the geometric design, usage of TCDs, traffic characteristics, WWD incidents (i.e., citations and/or 911 calls), etc. Transportation agencies can take advantage of the probabilistic/linear models developed in these studies to identify high-risk locations, which will enable them to adopt a systemic approach to reduce WWD crashes within their jurisdictions.

Discussions

Historically, it has been difficult to study the characteristics of WWD crashes due to the rareness of these types of crashes. Additionally, it is difficult to identify the true WWD crashes from a crash database alone. The identification of true WWD crashes may involve reviewing the crash narratives in the crash reports, which is time-consuming and laborious. Having said that, looking at the crash data alone may not be sufficient to perceive the actual gravity of WWD problems on motorways. Looking at the WWD incidents along with crash data is probably the best way to understand the underlying characteristics of wrong-way drivers on a certain location/facility. Based on their occurrences, WWD incidents can be categorized into three broad groups—recurring, intentional, and random incidents.

Recurring incidents happens regularly mainly due to the geometric design deficiencies. For instance, an average of 17 WWD incidents during a typical weekend was reported at the I-65 Exit 284 southbound ramp terminal in Alabama, United States. The potential design deficiencies that might have contributed to the recurring WWD movements at this location include a traversable median on crossroad, narrow median width between exit and entrance ramp, faded pavement markings, lack of proper signage, and

inadequate sight distance for drivers turning left from crossroad to the entrance ramp. Intentional WWD incidents are caused by a combination of roadway design deficiency and undesirable driving behaviors. For instance, there were 15 WWD movements over a 48-hours video monitoring at an intersection on a rural divided highway near a gas station in Alabama, United States. WWD movements at this location were due to the fact that the drivers exiting from the gas station were inclined to drive wrong way to make a shortcut instead of driving downstream until the next median opening to make a U-turn. A combination of access control and education strategies is needed to reduce this type of incident. However, intentional WWD incidents also include suicidal attempts and ghost driving, which may only be addressed through public awareness, stricter enforcement, and application of ITS. Random WWD incidents are primarily caused by impaired driving. WWD with a suicidal attempt also falls into this category. ITS technologies and quick incident responses can work together to reduce this type of incident and prevent related WWD crashes.

Prevention of recurring WWD incidents greatly depends on the geometric design and TCDs at the exit ramp terminals, where most WWD drivers enter wrong way initially. Therefore, careful considerations should be taken at the design stage of an exit ramp terminal to make it less susceptible to WWD. Signs and pavement markings should be oriented and placed carefully to complement the geometric design so that it is a positive source of information to the motorists and helps them make corrective measures quickly when someone enters wrong way or to prevent them from going wrong way in the first place.

Given the current rapid technological advancement, the available ITS technologies are becoming less costly, which indicates that many transportation agencies will adopt ITS technologies to mitigate WWD crashes within their jurisdiction. In addition, if the connected and autonomous vehicles become open to public use, it is likely that these vehicles will be equipped with special safety features that warn the motorists should it go wrong way and also disseminate information to the nearby right-way drivers to inform them about the imminent wrong-way driver.

Over the years, in the context of WWD, many transportation agencies have struggled to select locations for safety improvements because these crashes are rare and thus difficult to find hot spots based on the historical crash data. To overcome this issue, the transportation agencies may consider adopting a systemic safety approach, taking advantage of the available WWD crash-risk prediction models, to address WWD proactively.

Having said that, engineering countermeasures alone may not be enough to combat WWD crashes by impaired driving or intentional WWD activities. For instance, a study conducted by TTI reported that lowering the height of the white-on-red signs, making the signs larger (i.e., oversized), adding a red retroreflective sheeting on the sign support, or adding red flashing LEDs around the border of the sign did not improve the ability of alcohol impaired drivers to locate WRONG WAY signs. Therefore, to combat the WWD resulting from impaired drivers or undesirable driving behaviors, the best options for the transportation agencies are education and enforcement, possibly combined with making it impossible to drive a vehicle with alcohol on one's breath, but that is a topic for a different article. Transportation agencies are recommended to organize regular educational campaigns and media coverages, and adopt stricter enforcements to combat WWD resulting from impaired driving within their jurisdictions.

References

- American Association of State Highway and Transportation Officials (AASHTO), 2018. *A Policy on Geometric Design of Highways and Streets*, seventh ed. AASHTO, Washington, DC, USA.
- ASECAP, 2017. *Ghost Drivers: A Direct Experience of Toll Road Operators*. Permanent Committee on Safety, Security, Environment and Sustainability (COPER II), Association Européenne des Concessionnaires d'Autoroutes et d'Ouvrages à Péage, Paris, France.
- Atiquzaman, M., 2019. *Modeling the Risk of Wrong-Way Driving at Freeway Exit Ramp Terminals*. Doctoral Dissertation. Department of Civil Engineering, Auburn University, Auburn, AL, USA.
- Atiquzaman, M., Zhou, H., 2018a. Modeling the risk of wrong-way driving entry at the exit ramp terminals of full diamond interchanges. *Transp. Res. Rec. J. Transp. Res. Board* 2672 (17), 35–47, doi:10.1177/0361198118783152.
- AZCentral, 2018. ADOT: freeway technology has detected more than dozen wrong-way drivers. Available from: <https://www.azcentral.com/story/news/local/arizona-traffic/2018/06/12/adot-thermal-detection-technology-detects-15-wrong-way-drivers/693172002/>.
- Baratian-Ghorgchi, F., Zhou, H., 2017. Traffic control devices for deterring wrong-way driving: historical evolution and current practice. *J. Traffic Transp. Eng.* 4 (3), 280–289, doi:10.1016/j.jtte.2016.07.004.
- Baratian-Ghorgchi, F., Zhou, H., Jalayer, M., Pour-Rouholamin, M., 2014a. Prediction of potential wrong-way entries at exit ramp terminals of signalized partial cloverleaf interchanges. *Traffic Inj. Prev.* 16 (6), 599–604, doi:10.1080/15389588.2014.981651.
- Baratian-Ghorgchi, F., Zhou, H., Shaw, J., 2014b. Overview of wrong-way driving fatal crashes in the United States. *ITE J.* 84, 41–47.
- Braam, A.C., 2006. *Wrong-Way Crashes: Statewide Study of Wrong-Way Crashes on Freeways in North Carolina*. Traffic Engineering and Safety System Branch, North Carolina Department of Transportation, Raleigh, NC, USA.
- Brigl, S., 2007. Advance Warning of Drivers Heading in the Wrong Direction—The “Wrong-Way Driver” Information. BMW Group Press Release, September 7, 2007.
- Cooner, S.A., Cothron, A.S., Ranft, S.E., 2004. Countermeasures for wrong-way movement on freeway: overview of project activities and findings. FHWA/TX-04/4128-1. Texas Transportation Institute, TX, USA.
- Federal Highway Administration (FHWA), 2004. *Standard Highway Signs and Markings (2004 Edition with 2012 Supplement)*. US Department of Transportation, Washington, DC, USA.
- Federal Highway Administration (FHWA), 2009. *Manual on Uniform Traffic Control Devices for Streets and Highways*. Sections 2A.21, 2B.37, 2B.38, and 2B.40. US Department of Transportation, Washington, DC, USA.
- Finley, M.D., Balke, K.N., Rajbhandari, R., Chrysler, S.T., Dobrovolsky, C.S., Trout, N.D., Avery, P., Vickers, D., Mott, C., 2016. Conceptual design of a connected vehicle wrong-way driving detection and mitigation system. FHWA/TX-16/0-6867-1. Texas A&M Transportation Institute, College Station, TX, USA.
- HCTRA, 2012. *Incident management annual report 2012*. Harris County Toll Road Authority (HCTRA), Houston, TX, USA.
- Lathrop, S.L., Dick, T.B., Nolte, K.B., 2010. Fatal wrong-way collisions on New Mexico's interstate highways, 1990–2004. *J. Forensic Sci.* 55 (2), 432–437.
- Morena, D.A., Leix, T.J., 2012. Where these drivers went wrong. *Public Roads* 75 (6), 33–41.
- National Transportation Safety Board (NTSB), 2012. *Highway special investigation report: wrong-way driving*. NTSB/SIR-12/01. Washington, DC.
- Ponnaluri, R.V., 2016. The odds of wrong-way crashes and resulting fatalities: a comprehensive analysis. *Acc. Anal. Prev.* 88, 105–116, doi:10.1016/j.aap.2015.12.012.

- Pour-Rouholamin, M., 2016. Wrong-Way Driving: Crash Data Analysis and Safety Countermeasures. Doctoral Dissertation. Department of Civil Engineering, Auburn University, Auburn, AL, USA.
- Rogers Jr., J.H., Al-Deek, H., Alomari, A.H., Gordin, E., Carrick, G., 2016. Modeling the risk of wrong-way driving on freeways and toll roads. *Transp. Res. Rec. J. Transp. Res. Board* 2554, 166–176, doi:10.3141/2554-18.
- Sandt, A., Al-Deek, H., Rogers Jr., J.H., 2017. Identifying wrong-way driving hotspots by modeling crash risk and analyzing traffic management center response times. Presented at the Transportation Research Board 96th Annual Meeting. Washington, DC, USA.
- Vicedo, P., 2007. Preventing and managing ghost-driver incidents: the French experience. *Tollways* 4 (3), 43–49.
- Zhou, H., Pour-Rouholamin, M., 2014. Guidelines for reducing wrong-way crashes on freeways. FHWA-ICT-14-010. Illinois Center for Transportation, Illinois Department of Transportation, Springfield, IL, USA.
- Zhou, H., Pour-Rouholamin, M., 2015. Investigation of contributing factors regarding wrong-way driving on freeways, Phase II. FHWA-ICT-15-016. Illinois Center for Transportation, Illinois Department of Transportation, Springfield, IL, USA.
- Zhou, H., Pour-Rouholamin, M., Zhang, B., Wang, J., Turochy, R., 2017. A study of wrong-way driving crashes in Alabama, volume 1: freeways. ALDOT-Belt-SP07(906)-Wrong Way. Alabama Department of Transportation, Montgomery, AL, USA.
- Zhou, H., Pour-Rouholamin, M., Zhang, B., Wang, J., Turochy, R., 2017. A study of wrong-way driving crashes in Alabama, volume 2: divided highways. ALDOT-Belt-SP07(906)-Wrong Way. Alabama Department of Transportation, Montgomery, AL, USA.
- Zhou, H., Zhao, J., Fries, R., Gahrooei, M.R., Wang, L., Vaughn, B., Bahaaldin, K., Ayyalasomayajula, B., 2012. Investigation of contributing factors regarding wrong-way driving on freeways. FHWA-ICT-12-010. Illinois Center for Transportation, Illinois Department of Transportation, Springfield, IL, USA.
- Zhou, H., Zhao, J., Pour-Rouholamin, M., Tobias, P., 2015a. Statistical characteristics of wrong-way driving crashes on Illinois freeways. *Traffic Inj. Prev.* 16 (8), 760–767, doi:10.1080/15389588.2015.1020421.

Further Reading

- Arizona Department of Transportation, 2013. Wrong-way vehicle detection: proof of concept. FHWA-AZ-13-697. Phoenix, AZ.
- Arizona Department of Transportation, 2015. Detection and warning systems for wrong-way driving. FHWA-AZ-15-741. Phoenix, AZ.
- Atiquzzaman, M., Zhou, H., 2018. A comparative analysis of exit ramp terminal design practices to deter wrong-way driving in different states. Proceedings of the 97th Annual Meeting of Transportation Research Board. Washington, DC, USA.
- Chang, Q., Atiquzzaman, M., Zhou, H., 2019. Evaluation of low-cost countermeasures for preventing wrong-way driving incidents based on two before-and-after case studies. Proceedings of the 98th Annual Meeting of Transportation Research Board. Washington, DC, USA.
- ENTERPRISE, 2016. Countermeasures for wrong-way driving on freeways—project summary report. Transportation Pooled Fund Study TPR-5 (231). College Station, TX.
- Florida Department of Transportation (FDOT), 2015. Statewide wrong way crash study final report. Project ID: 190258-1-32-37, FDOT Central Office, FL, USA.
- Finley, M.D., Venglar, S.P., Iravarapu, V., Miles, J.D., Cooner, S.A., Ranft, S.E., 2014. Assessment of the effectiveness of wrong way driving countermeasures and mitigation methods. FHWA/TX-15/0-6769-1. Texas A&M Transportation Institute, TX, USA.
- Jalayer, M., Pour-Rouholamin, M., Zhou, H., 2018. Wrong-way driving crashes: a multiple correspondence approach to identify contributing factors. *Traffic Inj. Prev.* 19 (1), 35–41, doi:10.1080/15389588.2017.
- Jalayer, M., Zhou, H., Zhang, B., 2016. Evaluation of navigation performances of GPS devices near interchange area pertaining to wrong-way driving. *J. Traffic Transp. Eng.* 3 (6), 593–601, doi:10.1016/j.jtte.2016.07.003.
- Moler, S., 2002. Stop. You are going the wrong way! *Public Roads* 66 (2), 110.
- Pour-Rouholamin, M., Zhou, H., 2016. Analysis of driver injury severity in wrong-way driving crashes on controlled-access highways. *Acc. Anal. Prev.* 94, 80–88, doi:10.1016/j.aap.2016.05.022.
- Pour-Rouholamin, M., Zhou, H., Zhang, B., Turochy, R., 2016. Comprehensive analysis of wrong-way driving crashes on Alabama interstates. *Transp. Res. Rec. J. Transp. Res. Board* (2601), 50–58, doi:10.3141/2601-07.
- Wang, J., Zhou, H., 2017. In-depth investigation of wrong-way crashes on divided highways in Alabama using Haddon matrix and field observations. *Transp. Res. Rec. J. Transp. Res. Board* 2618, 8–26, doi:10.3141/2618-02.
- Wang, J., Zhou, H., 2018. Using naturalistic driving study data to evaluate the effects of intersection balance on driver behavior at partial cloverleaf interchange terminals. *Transp. Res. Rec. J. Transp. Res. Board* 2672, 255–265, doi:10.1177/0361198118774670.
- Wang, J., Zhou, H., Zhang, Y., 2017. Improve sight distance at signalized ramp terminals of partial cloverleaf interchanges to deter wrong-way entries. *ASCE J. Transp. Eng. Part A Syst.* 143 (6), doi:10.1061/JTEPBS.0000039.
- Yang, L., Zhou, H., Zhu, L., 2018. New concept design of directional rumble strips for deterring wrong-way freeway entries. *J. Transp. Eng. Part A Syst.* 144 (5), doi:10.1061/JTEPBS.0000131.
- Zhou, H., Pour-Rouholamin, M., 2014. Proceedings of the 1st national wrong-way driving summit. FHWA-ICT-14-009. Illinois Center for Transportation, Illinois Department of Transportation, Springfield, IL, USA.
- Zhou, H., Pour-Rouholamin, M., Jalayer, M., 2014. Emerging Safety Countermeasures for Wrong-Way Driving. American Traffic Safety Services Association (ATSSA), Fredericksburg, VA.
- Zhou, H., Zhao, J., Reisi-Gahrooei, M., Tobias, P., 2015b. Identification of contributing factors to wrong-way crashes on freeways in Illinois. *J. Transp. Safety Security* 8 (2), 97–112, doi:10.1080/19439962.2014.991814.

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

Page left intentionally blank

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

EDITOR-IN-CHIEF

Roger Vickerman

*School of Economics, University of Kent, Canterbury, United Kingdom
and*

Transport Strategy Centre, Imperial College, London, United Kingdom

VOLUME 3

Freight Transport and Logistics

SECTION EDITORS

Prof. Kevin P.B. Cullinane

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden



Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2021 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-08-102671-7

For information on all publications visit our
website at <http://store.elsevier.com>



Publisher: Oliver Walter
Acquisitions Editor: Oliver Walter
Content Project Manager: Natalie Lovell
Associate Content Project Manager: Manisha K and Ramalakshmi Boobalan
Designer: Matthew Limbert

EDITORIAL BOARD

Editor in Chief

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Section Editors

Maria Börjesson

Professor of Economics, VTI Swedish National Road and Transport Research Institute; Affiliated Professor at Linköping University, Sweden

Per Gårder

Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States

Prof. Kevin P.B. Cullinane

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

Prof. Edward C.S. Chung

Department of Electrical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

Chandra R. Bhat

Center for Transportation Research (CTR), The University of Texas at Austin, TX, United States

Edoardo Marcucci

Department of Political Sciences, University of Roma Tre, Rome, Italy; Department of Logistics, Molde University College, Molde, Norway

Prof. Maria Attard

Department of Geography, Faculty of Arts, University of Malta, Msida, Malta

Prof. Carlo G. Prato

School of Civil Engineering, The University of Queensland, Brisbane, Australia

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Page left intentionally blank

INTRODUCTION



Roger Vickerman

In an increasingly globalized world, despite reductions in costs and time, transportation has become even more important as a facilitator of economic and human interaction; this is reflected in technical advances in transportation systems, increasing interest in how transportation interacts with society and the need to provide novel approaches to understand its impacts. This has become particularly acute with the impact that Covid-19 has had on transportation across the world, at local, national, and international levels. This Encyclopedia brings a cross-cutting and integrated approach to all aspects of transportation from work in many disciplinary fields, engineering, operations research, economics, geography, and sociology to understand the changes taking place.

Transportation is both influenced by, and an influencer of, changes in the economy and society. Increasing speeds have reduced journey times and made the world a smaller place as globalization has affected both where people live and work and from where they source their goods and materials. Increasing volumes of traffic, often on old and life-expired infrastructure, lead to congestion and delays. Constraints on public budgets have led to increasing pressure on the private sector to fund improvements requiring new and innovative financial solutions. While there are clear differences in the nature of the pressures felt in the developed and less developed economies, there is an increasing recognition that in all societies, there is an accessibility problem such that certain groups become disadvantaged by the lack of access to reliable and cost-effective transport. While the problems are clearly multidimensional, research on transportation is often constrained by single disciplinary approaches and this carries over into the practice of transport planning and policy. The Encyclopedia cuts across these artificial boundaries by taking an approach that emphasizes the interaction between the different aspects of research and aims to offer new solutions to understand these problems. Each of the nine sections is based around a familiar dimension of work on transportation, but brings together the views of experts from different disciplinary perspectives. Each section is edited by an expert in the relevant field who has sought chapters from a range of authors representing different disciplines, different parts of the world, and different social perspectives. In this way, the work is not just a reflection of the state of the art that serves as a starting point for researchers and practitioners, but also a pointer toward new approaches, new ways of thinking, and novel solutions to problems.

The nine sections are structured around the following themes: Transport Modes; Freight Transport and Logistics; Transport Safety and Security; Transport Economics; Traffic Management; Transport Modeling and Data Management; Transport Policy and Planning; Transport Psychology; Sustainability and Health Issues in Transportation. Some of the chapters provide a technical introduction to a topic while others provide a bridge between topics or a more future-oriented view of new research areas or challenges. While there is guidance to cross-referencing between chapters, readers are encouraged to explore the tables of contents of all the sections to get a full understanding of the issues. Much of the Encyclopedia was completed before the Covid-19 pandemic and clearly this will have changed the situation in many areas covered by this work; the advantage of this type of reference work is that relevant updates will be possible in future editions.

The Encyclopedia has only been possible because of the cooperation of a large number of people. Robert Noland, Georgina Santos, Xiaowen Fu, and Dick Ettema served as an Editorial Advisory Board identifying possible editors of sections and advising on the overall structure. The Section Editors, Edoardo Marcucci, Kevin Cullinane, Per Garder, Maria Börjesson, Edward Chung, Chandra Bhat, Maria Attard, and Carlo Prato,

carried out the work of identifying potential authors of individual chapters, commissioning these, encouraging authors and reviewing drafts. More than 600 authors and co-authors of chapters are, however, ultimately responsible for the success of this venture. Thanks are also due to the key people at Elsevier, particularly the Publishers, Andre Wolff and Oliver Walter, and the Project Managers, Sophie Harrison and Natalie Bentahar; they have shown exemplary patience over more than 3 years in bringing this work to fruition.

LIST OF CONTRIBUTORS TO VOLUME 3

Sharon Cullinane

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

Kevin Cullinane

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

David J. Closs

Department of Supply Chain Management, Eli Broad College of Business, Michigan State University, East Lansing, MI, United States

David B. Grant

Supply Chain Management and Social Responsibility Group, Department of Marketing, Hanken School of Economics, Helsinki, Finland

Sarah Shaw

Logistics and Management Systems Subject Group, Hull University Business School, University of Hull, Hull, Yorkshire, United Kingdom

Wayne K. Talley

Old Dominion University, Norfolk, VA, United States

Luca Zamparini

Department of Law, University of Salento, Lecce, Italy

Aura Reggiani

Department of Economics, University of Bologna, Bologna, Italy

Nicolas Raimbault

Institute of Geography and Regional Planning, University of Nantes, CNRS UMR 6590 "Espaces et SOciétés" (ESO), Nantes, France; Urban Development and Mobility Department, Luxembourg Institute of Socio-Economic Research, Esch-sur-Alzette, Luxembourg

Richard Wilding

Centre for Logistics, Procurement and Supply Chain Management, Cranfield University, Cranfield, Bedfordshire, United Kingdom

Maria G. Burns

College of Technology, Houston, TX, United States

Zhuohua Qu

Liverpool Business School, Liverpool John Moores University, Liverpool, United Kingdom

Chengpeng Wan

National Engineering Research Center for Water Transport Safety, Wuhan, China; Intelligent Transport Systems Research Center, Wuhan University of Technology, Wuhan, China; Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, United Kingdom

Zaili Yang

Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, United Kingdom

Lisa M. Ellram

Miami University-Farmer School of Business, Department of Management, Oxford, OH, United States

Maria Björklund

Linköping University, Department of Management and Engineering, Logistics and Quality Management, Linköping, Sweden

Maja Piecyk-Ouellet

School of Architecture and Cities, University of Westminster, London, United Kingdom

Prof Usha Ramanathan

Nottingham Business School, Nottingham Trent University, Nottingham, England

Prof Ramakrishnan Ramanathan

University of Bedfordshire Business School, University Square, Luton, Bedfordshire, England

Petri Helo

Networked Value Systems, School of Technology and Innovations, University of Vaasa, Vaasa, Finland

Javad Rouzafzoon
*Networked Value Systems, School of Technology and
Innovations, University of Vaasa, Vaasa, Finland*

Aicha AGUEZZOUL
University of Lorraine, Metz, France

Kee-hung Lai
*The Hong Kong Polytechnic University, Hung Hom,
Kowloon, Hong Kong, China*

Jinan Shao
*Zhejiang University, Hangzhou City, Zhejiang Province,
China*

Yongyi Shou
*Zhejiang University, Hangzhou City, Zhejiang Province,
China*

Christina K. Wiederer cwiederer
*Global Trade and Regional Integration Unit, The World
Bank, Washington, DC, United States*

Jean-François Arvis
*Operations & Supply Chain Management, University of
Turku, Turku, Finland*

Lauri M. Ojala
*Operations & Supply Chain Management, University of
Turku, Turku, Finland*

Tuomas M. M. Kiiski
*Operations & Supply Chain Management, University of
Turku, Turku, Finland*

Evi Hartmann
Lange Gasse, Nuremberg, Germany

Hendrik Birkel
Lange Gasse, Nuremberg, Germany

Matthias Kopyto
Lange Gasse, Nuremberg, Germany

James H. Bookbinder
*Department of Management Sciences, University of
Waterloo, Waterloo, ON, Canada*

M. Ali Aœelkü
*Rowe School of Business, Dalhousie University, Halifax,
NS, Canada*

Erik Hofmann
*Institute of Supply Chain Management, University of St.
Gallen, St. Gallen, Switzerland*

Jesús García-Arca
*Industrial Engineering School, University of Vigo, Campus
Lagoas-Marconsende, Vigo, Spain*

Alicia Trinidad González-Portela Garrido
*Industrial Engineering School, University of Vigo, Campus
Lagoas-Marconsende, Vigo, Spain*

J. Carlos Prado-Prado
*Industrial Engineering School, University of Vigo, Campus
Lagoas-Marconsende, Vigo, Spain*

Jan C. Fransoo
Kuehne Logistics University, Hamburg, Germany

Maximiliano Udenio
*Research Center for Operations Management, KU Leuven,
Leuven, Belgium*

Yingli Wang
*Cardiff Business School, Cardiff University, Cardiff,
United Kingdom*

Shong-lee Ivan Su
*Department of Business Administration, School of
Business, Soochow University, Taipei, Taiwan*

Charles Kunaka
World Bank, Singapore, Singapore

Lóránt Tavasszy
Delft University of Technology, Delft, The Netherlands

Yousef Maknoon
Delft University of Technology, Delft, The Netherlands

Gerard de Jong
Significance, The Hague, The Netherlands

*Institute for Transport Studies, University of Leeds, Leeds,
United Kingdom*

Genevieve Giuliano
*Department of Urban Planning and Spatial Science,
University of Southern California, Los Angeles, CA,
United States*

Michael Browne
*School of Business, Economics and Law, University of
Gothenburg, Gothenburg, Sweden*

José Holguin-Veras
*Department of Civil and Environmental Engineering,
Rensselaer Polytechnic Institute, Troy, NY,
United States*

Julian Allen
*School of Architecture and Cities, University of
Westminster, London, United Kingdom*

Tom Van Woensel
*Eindhoven University of Technology, Eindhoven, The
Netherlands*

Gyöngyi Kovács
*HUMLOG Institute, Hanken School of Economics,
Helsinki, Finland*

Diego Vega
*HUMLOG Institute, Hanken School of Economics,
Helsinki, Finland*

M. Teresa Ortuño
*Facultad de Ciencias Matemáticas, Complutense
University of Madrid, Spain*

Jose M. Ferrer
*Facultad de Ciencias Económicas y
Empresariales, Complutense University of
Madrid, Spain*

Inmaculada Flores
*Facultad de Ciencias Matemáticas, Complutense
University of Madrid, Spain*

Gregorio Tirado
*Facultad de Ciencias Económicas
y Empresariales, Complutense University of Madrid,
Spain*

Rev. Ruth Dowson
*UK Centre for Events Management, Leeds Beckett
University, Leeds, United Kingdom*

Dan Lomax
*Macaulay Hall, Leeds Beckett University, Leeds, United
Kingdom*

Dale S. Rogers
*Arizona State University, W. P. Carey School of Business,
Tempe, AZ, United States*

Ronald S. Lembke
University of Nevada, Reno, NV, United States

Tolga Bektaş
*University of Liverpool Management School, Liverpool,
United Kingdom*

Hyun Chan Kim
*Waikato Institute of Technology, Hamilton, Waikato,
New Zealand*

Alan Nicholson
*University of Canterbury, Christchurch, Canterbury,
New Zealand*

María Feo-Valero
*Department of Applied Economics II, University of
Valencia, Valencia, Spain*

Amaya Vega
IEI, Valencia, Spain

Bárbara Vázquez-Paja
GMIT Galway Campus, Galway, Ireland

Edoardo Marcucci
*Faculty of Logistics, Molde University College, Molde,
Norway*

Valerio Gatta
*Department of Political Science, University of Roma Tre,
Roma, Italy*

Michela Le Pira
*Department of Civil Engineering and Architecture,
University of Catania, Catania, Italy*

Theo Notteboom
*Center for Eurasian Maritime and Inland Logistics
(CEMIL), China Institute of FTZ Supply Chain, Shanghai
Maritime University, Shanghai, China; Maritime
Institute, Ghent University, Ghent, Belgium; University of
Antwerp, Antwerp, Belgium; Antwerp Maritime Academy,
Antwerp, Belgium*

Manolis G. Kavussanos
*Athens University of Economics and Business, Athens,
Greece*

Stella A. Moysiadou
*Athens University of Economics and Business, Athens,
Greece*

Lourdes Trujillo
*Department of Applied Economics, University of Las
Palmas de Gran Canaria, Faculty of economics, Las
Palmas de Gran Canaria, Spain*

Alba Martínez-López
*Department of Mechanical Engineering, University of Las
Palmas de Gran Canaria, Industrial and Civil
Engineering School, Las Palmas de Gran Canaria, Spain*

Karin Andersson
*Mechanics and Maritime Sciences, Chalmers University of
Technology, Gothenburg, Stockholm, Sweden*

Selma Brynolf
*Mechanics and Maritime Sciences, Chalmers University of
Technology, Gothenburg, Stockholm, Sweden*

Lena Granhag
*Mechanics and Maritime Sciences, Chalmers University of
Technology, Gothenburg, Stockholm, Sweden*

J. Fredrik Lindgren
*Mechanics and Maritime Sciences, Chalmers University of
Technology, Gothenburg, Stockholm, Sweden*

Harilaos N. Psaraftis
*Technical University of Denmark, Kongens Lyngby,
Denmark*

Mary R. Brooks
*Rowe School of Business, Dalhousie University, Halifax,
NS, Canada*

Geraldine Knatz
*USC Sol Price School of Public Policy, USC Viterbi School
of Engineering, Los Angeles, CA, United States*

Francesco Parola
*Department of Economics and Business Studies,
University of Genoa, Genoa, Italy*

Giovanni Satta

Italian Centre of Excellence on Logistics, Transport and Infrastructures, University of Genoa, Genoa, Italy

Francesco Vitellaro

Italian Centre of Excellence on Logistics, Transport and Infrastructures, University of Genoa, Genoa, Italy

Peter W. de Langen

Department of Strategy and Innovation, Copenhagen Business School, Frederiksberg, Copenhagen, Denmark

María Manuela González

Department of Applied Economics, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Casiano Manrique-De-Lara-Peñate

Department of Applied Economics, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Ivone Pérez

Department of Applied Economics, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Michael G.H. Bell

ITLS, Business School, University of Sydney, Australia

Yuquan Du

National Centre for Ports and Shipping, Australian Maritime College, University of Tasmania, Launceston, TAS, Australia

Qiang Meng

Department of Civil and Environmental Engineering, National University of Singapore, Singapore

Gordon Wilmsmeier

Kühne Professorial Chair in Logistics, School of Management, Universidad de los Andes, Bogota, Colombia

Jason Monios

Associate Professor in Maritime Logistics, Kedge Business School, Marseille, France

Yufeng Lin

Department of Supply Chain Management, Asper School of Business, University of Manitoba, Winnipeg, MB, Canada

David G. Babb

International Forum on Climate Change Adaptation for Port, Transportation Infrastructure, and the Arctic (CCAPPTIA)

Adolf K.Y. Ng

Centre for Earth Observation Science, University of Manitoba, Winnipeg, MB, Canada; St. John's College, University of Manitoba, Winnipeg, MB, Canada

Kenneth Button

Schar School of Policy and Government, George Mason University, Arlington, VA, United States

Keith Debbage

Department of Geography, Environment and Sustainability, University of North Carolina at Greensboro, Greensboro, NC, United States

Neil Debbage

Department of Political Science and Geography, University of Texas at San Antonio, San Antonio, TX, United States

Lucy Budd

Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

Stephen Ison

Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

Oliver Kunze

HNU Neu-Ulm University of Applied Sciences, Neu-Ulm, Germany

Thomas M. Corsi

Robert H. Smith School of Business, University of Maryland, College Park, MD, United States

Diñçer Konur

Department of Computer Information Systems and Quantitative Methods, McCoy College of Business Administration, Texas State University, San Marcos, TX, United States

Gonca Yildirim

Department of Industrial Engineering, Cankaya University, Ankara, Turkey

Bahriye Cesaret

Faculty of Business, Ozyegin University, Istanbul, Turkey

Heikki Liimatainen

Tampere University, Tampere, Finland

Tooraj Jamasb

Copenhagen School of Energy Infrastructure (CSEI), Copenhagen Business School, Frederiksberg, Denmark

Manuel Llorca

Copenhagen School of Energy Infrastructure (CSEI), Copenhagen Business School, Frederiksberg, Denmark

Heike Flømig

TUHH, Am Schwarzenberg-Campus 3, Hamburg, Germany

Dr Allan Woodburn

University of Westminster, London, United Kingdom

Maksym Spiryagin
*Central Queensland University, Centre for Railway
 Engineering, Rockhampton, QLD, Australia*

Qing Wu
Railway Consultant, Rostov-on-Don, Russian Federation

Peter Wolfs
*Central Queensland University, Centre for Railway
 Engineering, Rockhampton, QLD, Australia*

Colin Cole
*Central Queensland University, Centre for Railway
 Engineering, Rockhampton, QLD, Australia*

Valentyn Spiryagin
*Central Queensland University, Centre for Railway
 Engineering, Rockhampton, QLD, Australia*

Tim McSweeney
*Central Queensland University, Centre for Railway
 Engineering, Rockhampton, QLD, Australia*

Hendrik Rodemann
Unter den Eichen, Wolfsburg, Germany

Simon Templar
*Cranfield School of Management, Cranfield University,
 Cranfield, Bedfordshire, United Kingdom*

Tomas Ambra
*Vrije Universiteit Brussel, MOBI research Center,
 Brussels, Belgium, Europe*

Koen Mommens
*Vrije Universiteit Brussel, MOBI research Center,
 Brussels, Belgium, Europe*

Cathy Macharis
*Vrije Universiteit Brussel, MOBI research Center,
 Brussels, Belgium, Europe*

Matthew E. Oliver
*Georgia Institute of Technology, School of Economics,
 Atlanta, GA, United States*

Wouter P.C. Boon
*Copernicus Institute of Sustainable Development, Faculty
 of Geosciences, Utrecht University, Utrecht, The
 Netherlands*

Bert van Wee
*Transport and Logistics Group, Faculty Technology, Policy
 and Management, Delft University of Technology, Delft,
 The Netherlands*

Eric Ballot
*MINES ParisTech, PSL University, Centre de gestion
 scientifique (CGS), i3 UMR CNRS, Paris, France*

Shenle PAN
*MINES ParisTech, PSL University, Centre de gestion
 scientifique (CGS), i3 UMR CNRS, Paris, France*

Paul Tae-Woo Lee
*Maritime Logistics and Free Trade Islands Research
 Centre at Ocean College, Zhejiang University, Zhoushan,
 China*

Barbara Lenz
German Aerospace Center, Berlin

Johannes Gruber
German Aerospace Center, Berlin

Page left intentionally blank

CONTENTS OF ALL VOLUMES

<i>Editorial Board</i>	<i>v</i>
<i>Introduction</i>	<i>vii</i>
<i>Contributors to Volume 3</i>	<i>ix</i>

VOLUME 1

Introduction to Transportation Economics	1
Market Failures and Public Decision Making in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	2
Demand for Freight Transport <i>José Holguín-Veras and Diana G. Ramírez-Ríos</i>	7
Cost Functions for Road Transport <i>José Manuel Vassallo</i>	13
Future of Urban Freight <i>Russell G. Thompson</i>	20
Operation Costs for Public Transport <i>Marco Batarce</i>	26
Natural Monopoly in Transport <i>André de Palma and Julien Monardo</i>	30
Freight Costs: Air and Sea <i>Yulai Wan and Dong Yang</i>	36
Transport Production and Cost Structure <i>Ricardo Giesen and Darío Farren</i>	46
The Concept of External Cost: Marginal versus Total Cost and Internalization <i>Sofia F. Franco</i>	56
Value of Time <i>John J. Bates</i>	67
Valuation of Carbon Emissions <i>Svante Mandell</i>	72
Valuation of Travel Time Variability Using Scheduling Models <i>Katrine Hjorth</i>	76

Value of Crowding <i>Daniel Hörcher</i>	84
What Drives Transport and Mobility Trends? The Chicken-and-Egg Problem <i>Nathalie Picard</i>	89
Pricing Principles in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	95
Long-Run Versus Short-Run Valuations <i>Stefanie Peer</i>	102
Value of Noise <i>Jon P. Nelson</i>	106
The Value of Life and Health <i>Henrik Andersson</i>	114
The Value of Security, Access Time, Waiting Time, and Transfers in Public Transport <i>Raquel Espino, Juan de Dios Ortúzar, and Luis I. Rizzi</i>	122
Demand for Passenger Transportation <i>Kenneth A Small and Robin Lindsey</i>	127
Real-World Experiences of Congestion Pricing <i>Charles Raux</i>	134
Distributional Effects of Congestion Charges and Fuel Taxes <i>Jonas Eliasson</i>	139
The Bottleneck Model <i>Dereje Abegaz and Yili Tang</i>	146
Dynamic Congestion Pricing and User Heterogeneity <i>Kathrin Goldmann and Gernot Sieg</i>	150
Economics of Parking <i>Daniel Albalade and Albert Gragera</i>	159
Loss Aversion and Size and Sign Effects in Value of Time Studies <i>Andrew Daly</i>	165
Intertemporal Variation of Valuations <i>James Fox</i>	170
The Rebound Effect for Car Transport <i>Bruno De Borger, Ismir Mulalic and Jan Rouwendal</i>	174
Elasticities for Travel Demand: Recent Evidence <i>Fay Dunkerley, Charlene Rohr and Mark Wardman</i>	179
Parking Price Elasticities <i>Stephan Lehner</i>	185
The Pareto Criterion and the Kaldor Hicks Criterion <i>Harald Minken</i>	190
Social Discount Rates <i>Lars Hultkrantz</i>	195
Cross-Elasticities between Modes <i>Mark Wardman and Jeremy Toner</i>	201
First-Best Congestion Pricing <i>Moez Kilani</i>	209

Ethical Aspects-Can We Value Life, Health, and Environment in Money Terms? <i>Jose M. Grisolia and Ken Willis</i>	216
Car tolls, Transit Subsidies for Commuting, and Distortions on the Labor Market <i>Ioannis Tikoudis¹ and Kurt Van Dender¹</i>	221
Demand Management and Capacity Planning of Airports <i>Stephen Ison and Lucy Budd</i>	227
Dealing With Negative Externalities: Low Emission Zones Versus Congestion Tolls <i>Valeria Bernardo, Xavier Fageda, and Ricardo Flores-Fillol</i>	231
The Rule-of-a-Half and Interpreting the Consumer Surplus as Accessibility <i>Mogens Fosgerau and Ninette Pilegaard</i>	237
Producer Surplus <i>Chau Man Fung</i>	242
The Robustness of Cost-Benefit Analyses <i>Morten Welde and James Odeck</i>	249
The GDP Effects of Transport Investments: The Macroeconomic Approach <i>James Laird and Daniel Johnson</i>	256
The Mohring Effect <i>Hugo E. Silva</i>	263
Public Transport Fare and Subsidy Optimization <i>Qianwen Guo and Zhongfei Li</i>	267
Transportation Equity <i>Rafael H. M. Pereira, and Alex Karner</i>	271
Impact of Transport Cost-Benefit Analysis on Public Decision-Making <i>Niek Mouter</i>	278
Causal Inference for Ex Post Evaluation of Transport Interventions <i>Daniel J. Graham</i>	283
Transport Cost and Location of Firms <i>Adelheid Holl</i>	291
Commuting, the Labor Market, and Wages <i>Jan Rouwendal, and Ismir Mulalic</i>	297
How to Buy Transport Infrastructure <i>Johan Nyström</i>	302
Procurement of Public Transport: Contractual Regimes <i>Andrew Smith, and Chris Nash</i>	308
The Mono-Centric City Model and Commuting Cost <i>Zhi-Chun Li, and Ya-Juan Chen</i>	315
Value of Time in Freight Transport <i>Gerard de Jong</i>	321
The Economics and Planning of Urban Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	326
Incentives in Public Transport Contracts <i>Andreas Vigren</i>	332
Transportation Improvements and Property Prices <i>Alex Anas</i>	337

Transport Infrastructure Effects on Economic Output: The Microeconomic Approach <i>Patricia C. Melo</i>	347
Wider Economic Impacts of Transport Investments <i>Anthony J. Venables</i>	355
Employment Effects of Transport Infrastructure <i>Anna Matas, and Javier Asensio</i>	360
Regulatory Reforms and Competition in Public Transport <i>Didier van de Velde</i>	365
How to Finance Transport Infrastructure? <i>Tiziana D'Alfonso, and Giuseppe Catalano</i>	371
Congestion, Allocation and Competition on the Railway Tracks <i>John Armstrong, and John Preston</i>	378
Public Private Partnership <i>David Meunier, and Emile Quinet</i>	385
Airline Economics <i>Anming Zhang, Yahua Zhang, and Zhibin Huang</i>	392
Privatization and Deregulation of the Airline Industry <i>Achim I. Czerny, and Hao Lang</i>	397
Price Discrimination and Yield Management in the Airline Industry <i>Guoquan Zhang, Colin C.H. Law, Yahua Zhang, and Hangjun Yang</i>	404
Regulation and Competition in Railways <i>Chris Nash, and Andrew Smith</i>	409
Cycling Economics <i>Bert van Wee</i>	414
The Economic Rationale for High-Speed Rail <i>Ginés de Rus</i>	419
Rail Cost Functions <i>Kristofer Odolinski, and Phill Wheat</i>	425
Transport and International Trade <i>Siri Pettersen Strandenes</i>	431
Maritime Economics: Organizational Structures <i>Kevin P.B. Cullinane</i>	436
Port Planning and Investment <i>Jasmine Siu Lee Lam</i>	443
Estimating the Capital Stock of Transport Infrastructure <i>Heike Link</i>	449
The Economics of Reducing Carbon Emissions From Air and Road Transport <i>Olga Ivanova</i>	457
Regulation and Financing of Toll Roads <i>Marco Ponti</i>	464
Are Megaprojects too Transformational for Cost-Benefit Analysis? <i>Tom Worsley</i>	470
Economics of Transportation Safety <i>Ian Savage</i>	476

Cost Overruns of Transportation Infrastructure Projects <i>James Odeck, and Morten Welde</i>	483
The Downs-Thomson Paradox <i>Joel P. Franklin</i>	490
Policy Instruments for Plug-In Electric Vehicles: An Overview and Discussion <i>Jake Whitehead, Patrick Plötz, Patrick Jochem, Frances Sprei, and Elisabeth Dütschke</i>	496
Vertical and Horizontal Separation in the European Railway Sector and Its Effects on Productivity <i>Pedro Cantos-Sánchez</i>	503
How will Autonomous Vehicles Impact Car Ownership and Travel Behavior <i>Patrick M. Bösch, Felix Becker, Henrik Becker, and Kay W. Axhausen</i>	508
Policy Instruments to Reduce Carbon Emissions from Road Transport Computable General Equilibrium Analysis in Transportation Economics <i>Johannes Bröcker</i>	520
Estimation of Value of Time <i>Stefan Flügel, and Askill H. Halse</i>	527
The Taxation of Car Use in the Future <i>Griet De Ceuster, and Inge Mayeres</i>	534
Cost-Benefit Analysis and Other Assessment Techniques: Contrasts and Synergies <i>Paolo Beria</i>	540
Demand for Air Travel and Income Elasticity <i>Jing Lu, Yucan Meng, Changmin Jiang, and Cheng Lv</i>	547
Generalized Cost for Transport <i>Jepppe Rich</i>	555
The Impact of Electric Vehicles on Energy Systems <i>Patrick E.P. Jochem, Jake Whitehead, and Elisabeth Dütschke</i>	560
Uber versus Taxis <i>Georgina Santos</i>	566
Contract Efficiency in Public Transport Services <i>Philippe Gagnepain, and Marc Ivaldi</i>	572
Company Cars <i>Stefan Gössling</i>	580
Transportation Network Companies (TNCs) and the Future of Public Transportation <i>Susan Shaheen, and Adam Cohen</i>	584
Public Transport in Low Density Areas <i>Jani-Pekka Jokinen, Leif Sörensen, and Jan Schlüter</i>	589
Market Failures in Transport: Direct and Indirect Public Intervention <i>Federico Boffa, and Alberto Iozzi</i>	596
The Braess Paradox <i>Anna Nagurney, and Ladimer S. Nagurney</i>	601
VOLUME 2	
Introduction to Transportation Safety and Security <i>Per Gårder</i>	1

The Concept of “Acceptable Risk” Applied to Road Safety Risk Level <i>Claes Tingvall</i>	2
Crash Not Accident <i>Robert A. Scopatz</i>	6
Age and Gender as Factors in Road Safety <i>Marion Sinclair</i>	11
Aggressive Driving and Road Rage <i>James E.W. Roseborough, Christine M. Wickens, and David L. Wiesenthal</i>	17
Aircraft Maintenance and Inspection <i>Alan Hobbs</i>	25
Airport Security <i>Richard W. Bloom</i>	34
Transport Safety and Security: Alcohol <i>James C. Fell</i>	40
Animal Crashes <i>Michal Bil</i>	53
Attenuators <i>Simonetta Boria</i>	63
ATV, Snowmobile, and Terrain Vehicle Safety <i>David P. Gilkey, and William Brazile</i>	77
Automobile Safety Inspection <i>Subasish Das</i>	85
Aviation Safety: Commercial Airlines <i>Clarence C. Rodrigues</i>	90
Aviation Safety, Freight, and Dangerous Goods Transport by Air <i>Glenn S. Baxter and Graham Wild</i>	98
Bicycle Collision Avoidance Systems: Can Cyclist Safety be Improved with Intelligent Transport Systems? <i>Lars Leden</i>	108
Bicycle Infrastructure <i>Rock E. Miller</i>	115
Bicycles, E-Bikes and Micromobility, A Traffic Safety Overview <i>Höskuldur Kröyer</i>	125
Bicycles: The Safety of Shared Systems Versus Traditional Ownership <i>Mercedes Castro-Nuño and José I. Castillo-Manzano</i>	139
Bridge Safety <i>Per Erik Garder</i>	144
Carsharing Safety and Insurance <i>Elliot Martin and Susan Shaheen</i>	150
Carjacking <i>Terance D. Mieth, Christopher Forepaugh, Tanya Dudinskaya</i>	157
Collision Avoidance Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	164
Collision Avoidance Systems, Automobiles <i>Erick J. Rodríguez-Seda</i>	173

Connected Automated Vehicles: Technologies, Developments, and Trends <i>Azra Habibovic and Lei Chen</i>	180
Construction Zones <i>Jalil Kianfar</i>	189
Costs of Accidents <i>Ulf Persson</i>	196
Critical Issues for Large Truck Safety <i>Matthew C. Camden, Jeffrey S. Hickman, Richard J. Hanowski, and Martin Walker</i>	200
Demerit Points and Similar Sanction Programs <i>Matúš ťucha and Kristýna Josrová</i>	210
Driver State and Mental Workload <i>Dick de Waard and Nicole van Nes</i>	216
Drugs, Illicit, and Prescription <i>Rune Elvik</i>	221
Education, Training, and Licensing <i>Matúš ťucha and Kristýna Josrová</i>	228
Elderly Driver Safety Issues <i>Mark J King</i>	233
Emergency Response Systems <i>Frances L. Edwards</i>	240
Emergency Vehicles and Traffic Safety <i>Shamsunnahar Yasmin, Sabreena Anowar, and Richard Tay</i>	247
Encouragement: Awards and Incentives <i>Fred Wegman</i>	255
Enforcement and Fines <i>Matúš ťucha and Ralf Risser</i>	263
Epidemiology of Road Traffic Crashes <i>Sherrie-Anne Kaye, Judy Fleiter, and Md Mazharul Haque</i>	269
Evacuation Planning and Transportation Resilience <i>Karl Kim</i>	276
Exposure: A Critical Factor in Risk Analysis <i>Frank Gross, PhD, PE</i>	282
Passenger Ferry Vessels and Cruise Ships: Safety and Security <i>Wayne K. Talley</i>	290
Fuel Economy Standards: Impacts on Safety <i>Kenneth T. Gillingham and Stephanie M. Weber</i>	296
Hazardous Materials Transport <i>Dr. Arjan Vincent van der Vlies</i>	304
Head-on Crashes <i>John N. Ivan</i>	311
Helicopters in Emergency Medical Response <i>Stephen J.M. Sollid, M.D., PhD</i>	316
Horizontal and Vertical Geometry <i>Victoria Gitelman</i>	322

Human Factors in Transportation <i>Alison Smiley, Christina (Missy) Rudin-Brown</i>	331
In-Depth Crash Analysis and Accident Investigation <i>Yong Peng, Helai Huang, and Xinghua Wang</i>	346
Incident Detection Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	351
Inequality and Traffic Safety <i>Miles Tight</i>	358
Lighting <i>John D. Bullough</i>	361
Macroscopic Safety Analysis <i>Mohamed Abdel-Aty and Jaeyoung Lee</i>	367
Motor Vehicle Crash Reportability <i>John J. McDonough</i>	380
Nominal Safety <i>Per Erik Garder</i>	386
Parking Lots <i>Maxim A. Dulebenets</i>	392
Passenger Van Safety <i>Saksith Chalermpong and Apiwat Ratanawaraha</i>	401
Passive Prevention Systems in Automobile Safety <i>B. Serpil Acar</i>	406
Pedestrian Safety, Children <i>Mette Møller</i>	415
Pedestrian Safety, General <i>Muhammad Z. Shah, Mehdi Moeinaddini, and Mahdi Aghaabbasi</i>	420
Pedestrian Safety, Older People <i>Carlo Lui</i>	429
Visually Impaired Pedestrian Safety <i>Robert S. Wall Emerson</i>	435
Photo/Video Traffic Enforcement <i>Charles M. Farmer</i>	439
Powered Two- and Three-Wheeler Safety <i>Fangrong Chang, Helai Huang, and Md. Mazharul Haque</i>	443
Railroad Safety <i>Xiang Liu and Zhipeng Zhang</i>	451
Railroad Safety: Grade Crossings and Trespassing <i>Rahim F. Benekohal and Jacob Mathew</i>	466
Recreational Boating Safety: Usage, Risk Factors, and the Prevention of Injury and Death <i>Amy E. Peden, Stacey Willcox-Pidgeon, and Kyra Hamilton</i>	477
Refuge Islands <i>Christer Hydén</i>	487
Risk Perception and Risk Behavior in the Context of Transportation <i>Martina Raue and Eva Lerner</i>	494

Road Diets <i>Robert B. Noland</i>	500
Road Safety Audits <i>Xiao Qin</i>	508
Roadside Safety Barriers <i>Dean C. Alberson</i>	513
Road Safety Management in Selected Countries <i>Paul Boase and Brian Jonah</i>	519
Roadway Pavement Conditions and Various Summer and Winter Maintenance Strategies <i>Kamal Hossain, PhD P. Eng</i>	529
Safety of Roundabouts <i>Khaled Shaaban</i>	539
Rumble Strips, Continuous Shoulder, and Centerline <i>AnnaAnund and AnnaVadeby</i>	549
Safety Culture <i>Tor-Olav Nvestad</i>	554
Safety Data Quality Management <i>Robert A. Scopatz</i>	560
School Bus Safety <i>Yousif A. Abulhassan</i>	564
School Campus Traffic Circulation <i>Dimitrios Nalmpantis</i>	568
Sexual Violence in Public Transportation <i>Vania Ceccato</i>	576
Shared Space <i>Allison B. Duncan</i>	584
Side Area Safety and Side Slopes <i>José María Pardillo-Mayora</i>	593
Simulators <i>Nichole L. Morris, Curtis M. Craig, Jacob D. Achtemeier, and Peter A. Easterlund</i>	602
Sleep-Related Issues and Fatigue <i>Prof Marion Sinclair and Estelle Swart</i>	611
Speed Governors and Limiters <i>Christer Hydén</i>	617
Speed Limits on Rural Highways <i>Peter Tarmo Savolainen and Timothy Jordan Gates</i>	622
Speed Limits on Urban Streets <i>Anna Bray Sharpin, Claudia Adriazola-Steil, Ben Welle, and Natalia Lleras</i>	632
Speed-Reducing Measures <i>Vinod Vasudevan</i>	641
Striping, Signs, and Other Forms of Information <i>Kasem Choocharukul and Kerkritt Sriroongvikrai</i>	648
Suicides <i>Brendan Ryan</i>	656

Surrogate Measures of Safety <i>Nicolas Saunier and Aliaksei Laureshyn</i>	662
Targeting Transit: the Terrorist Threat and the Challenges to Security <i>Brian Michael Jenkins</i>	668
The Swedish Vision Zero: A Policy Innovation <i>Matts-Åke Belin</i>	675
Tire Safety <i>Saied Taheri</i>	681
Town Gates: Section on Transport Safety and Security <i>Charles Tijus</i>	685
Traffic Flow Volume and Safety <i>Athanasios Theofilatos and Apostolos Ziakopoulos</i>	692
Traffic Safety and Security of Taxis and Ride-Hailing Vehicles <i>Zhe Wang, Helai Huang, and Ye Li</i>	699
Traffic Signals and Safety <i>Andrew P. Tarko</i>	706
Tunnels, Safety and Security Issues-Risk Assessment for Road Tunnels: State-of-the-Art Practices and Challenges <i>Konstantinos Kirytopoulos, Panagiotis Ntzeremes, and Konstantinos Kazaras</i>	713
Understanding, Managing, and Learning from Disruption <i>Karl Kim</i>	719
Use/Analysis of Crash Data and Underreporting of Crashes <i>Mohammadali Shirazi and Dominique Lord</i>	726
Utility Poles <i>Lai Zheng and Tarek Sayed</i>	731
Value of Life and Injuries <i>David A. Hensher</i>	737
Effects of Weather <i>Maria Pregnolato, Amirhassan Kermanshah, and Wisinee Wisetjindawat</i>	742
Wrong-Way Driving on Motorways <i>Huaguo Zhou and Md Atiquzzaman</i>	751

VOLUME 3

Freight Transport and Logistics <i>Sharon Cullinane and Kevin Cullinane</i>	1
Expanding the Perspective of Logistics and Supply Chain Management <i>David J. Closs</i>	8
Logistics and Supply Chain Management Performance Measures <i>David B. Grant and Sarah Shaw</i>	16
Economic Regulation/Deregulation and Nationalization/Privatization in Freight Transportation <i>Wayne K. Talley</i>	24
Freight Transport Policy <i>Luca Zamparini and Aura Reggiani</i>	29

Planning and Financing Logistics Spaces <i>Nicolas Raimbault</i>	35
Supply Chain Risk Management: Creating the Resilient Supply Chain <i>Richard Wilding</i>	41
Transportation Safety and Security <i>Maria G. Burns</i>	47
Resilience in Freight Transport Networks <i>Zhuohua Qu, Chengpeng Wan, and Zaili Yang</i>	53
Environmental Sustainability in Freight Transportation <i>Lisa M. Ellram</i>	58
Sustainable Logistics, CSR in Logistics, and Sustainable Supply Chain Management <i>Maria Björklund and Maja Piecyk-Ouellet</i>	64
Information Sharing and Business Analytics in Global Supply Chains <i>Prof Usha Ramanathan and Prof Ramakrishnan Ramanathan</i>	71
Logistics Information Systems <i>Petri Helo and Javad Rouzafzoon</i>	76
Factors Affecting the Selection of Logistics Service Providers <i>Aicha AGUEZZOUL</i>	85
Logistics Service Performance <i>Kee-hung Lai, Jinan Shao, and Yongyi Shou</i>	89
The World Bank's Logistics Performance Index <i>Christina K. Wiederer cwiederer, Jean-François Arvis, Lauri M. Ojala, and Tuomas M. M. Kiiski</i>	94
Outsourcing Logistics Functions <i>Evi Hartmann, Hendrik Birkel, and Matthias Kopyto</i>	102
Freight Transport and Logistics in JIT Systems <i>James H. Bookbinder and M. Ali Aælkü</i>	107
Supply Chain Finance <i>Erik Hofmann</i>	113
Packaging Logistics <i>Jesús García-Arca, Alicia Trinidad González-Portela Garrido, and J. Carlos Prado-Prado</i>	119
The Bullwhip Effect <i>Jan C. Fransoo and Maximiliano Udenio</i>	130
Blockchain Applications in Logistics <i>Yingli Wang</i>	136
Logistics in Asia <i>Shong-lee Ivan Su</i>	143
Logistics in the Developing World <i>Charles Kunaka</i>	150
Freight Network Modeling <i>Lóránt Tavasszy and Yousef Maknoon</i>	157
National Freight Transport Models <i>Gerard de Jong</i>	162
Freight Flows in Cities <i>Genevieve Giuliano</i>	168

Urban Logistics and Freight Transport <i>Michael Browne, José Holguín-Veras, and Julian Allen</i>	178
Omni-Channel Logistics <i>Tom Van Woensel</i>	184
Humanitarian Logistics <i>Gyöngyi Kovács and Diego Vega</i>	190
Optimization of Humanitarian Logistics <i>M. Teresa Ortuño, Jose M. Ferrer, Inmaculada Flores, and Gregorio Tirado</i>	195
Event Logistics <i>Rev. Ruth Dowson and Dan Lomax</i>	201
Reverse Logistics <i>Dale S. Rogers and Ronald S. Lembke</i>	208
E-Tailing and Reverse Logistics <i>Sharon Cullinane</i>	219
Green Routing of Freight Vehicles <i>Tolga Bektaş</i>	224
Freight Mode Choice <i>Hyun Chan Kim and Alan Nicholson</i>	231
The Value of Time in Freight Transport <i>María Feo-Valero, Amaya Vega, and Bárbara Vázquez-Paja</i>	236
Behavioral Research in Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	242
Container (Liner) Shipping <i>Theo Notteboom</i>	247
Bulk Shipping Markets: An Overview of Market Structure and Dynamics <i>Manolis G. Kavussanos and Stella A. Moysiadou</i>	257
Ferries and Short Sea Shipping <i>Lourdes Trujillo and Alba Martínez-López</i>	280
Shipping and the Environment <i>Karin Andersson, Selma Brynolf, Lena Granhag, and J. Fredrik Lindgren</i>	286
Energy Efficiency of Ships <i>Harilaos N. Psaraftis</i>	294
Seaports <i>Mary R. Brooks and Geraldine Knatz</i>	299
Port Hinterlands <i>Francesco Parola, Giovanni Satta, and Francesco Vitellaro</i>	305
Seaports as Clusters of Economic Activities <i>Peter W. de Langen</i>	310
Port Efficiency and Effectiveness <i>Lourdes Trujillo, María Manuela González, Casiano Manrique-De-Lara-Peñate, and Ivone Pérez</i>	316
Container Port Automation <i>Michael G.H. Bell</i>	323
Optimizing Crane Operations in Ports Scheduling of Liner Container Shipping Services	335

<i>Yuquan Du and Qiang Meng</i>	
Dry Ports	344
<i>Gordon Wilmsmeier and Jason Monios</i>	
Arctic Shipping	349
<i>Yufeng Lin, David G. Babb, and Adolf K.Y. Ng</i>	
Airfreight and Economic Development	355
<i>Kenneth Button</i>	
Air Freight Logistics	361
<i>Keith Debbage and Neil Debbage</i>	
Air Freight Marketing	369
<i>Lucy Budd and Stephen Ison</i>	
Drones in Freight Transport	374
<i>Oliver Kunze</i>	
Duty of Care in the Selection of Motor Carriers	382
<i>Thomas M. Corsi</i>	
Carrier Selection for Less-Than-Truckload (LTL) Shipments	388
<i>Dinçer Konur, Gonca Yildirim, and Bahriye Cesaret</i>	
Decarbonizing Road Freight Transport	395
<i>Heikki Liimatainen</i>	
The Rebound Effect in Road Freight Transport	402
<i>Tooraj Jamasb and Manuel Llorca</i>	
Autonomous Goods Transport	407
<i>Heike Flømig</i>	
Rail Freight	413
<i>Dr Allan Woodburn</i>	
Rail Freight Vehicles	423
<i>Maksym Spiryagin, Qing Wu, Peter Wolfs, Colin Cole, Valentyn Spiryagin, and Tim McSweeney</i>	
Eurasia Rail Freight: Enablers and Inhibitors of Future Growth	436
<i>Hendrik Rodemann and Simon Templar</i>	
Intermodal and Synchromodal Freight Transport	456
<i>Tomas Ambra, Koen Mommens, and Cathy Macharis</i>	
Pipelines	463
<i>Matthew E. Oliver</i>	
3D Printers and Transport	471
<i>Wouter P.C. Boon and Bert van Wee</i>	
The Physical Internet and Logistics	479
<i>Eric Ballot and Shenle PAN</i>	
Bicycles for Urban Freight	488
<i>Barbara Lenz and Johannes Gruber</i>	
China's Belt and Road Initiative	495
<i>Paul Tae-Woo Lee</i>	

VOLUME 4

Introduction to Traffic Management <i>Edward C.S. Chung</i>	1
Urban Motorway Management <i>John Gaffney and Hendrik Zurlinden</i>	2
Ramp Metering Application <i>John Gaffney and Hendrik Zurlinden</i>	10
City Wide Coordinated Ramp Meters <i>John Gaffney and Hendrik Zurlinden</i>	21
Variable Speed Limits for Traffic Efficiency Improvement <i>José Ramón D. Frejo and Bart De Schutter</i>	33
Hard Shoulder Running <i>Justin Geistefeldt</i>	41
High-Occupancy Vehicle (HOV) and High-Occupancy Toll (HOT) Lanes <i>Roxana J. Javid, Jiani Xie, Lijiao Wang, Wenruifan Yang, Ramina Jahanbakhsh Javid, and Mahmoud Salari</i>	45
Reversible Lanes: Guidelines, Operation and Control, Research Directions <i>Gowri Asaithambi, Venkatesan Kanagaraj, and Madhuri Kashyap</i>	52
Electronic Toll Collection <i>Azusa Toriumi</i>	60
Road Pricing-Theory and Applications <i>Kian Keong Chin</i>	68
Road Pricing 1: The Theory of Congestion Pricing <i>Timothy D. Hau</i>	74
Road Pricing 2: Short- and Long-Run Equilibrium of Road Transportation <i>Timothy D. Hau</i>	83
Road Pricing 3: The Implications for Pricing Public Transportation <i>Timothy D. Hau</i>	90
Road Pricing 4: Case Study-The Implementation of Electronic Road Pricing in Hong Kong <i>Timothy D.</i>	103
Advanced Travelers Information Systems (ATIS) <i>Chintan Advani and Ashish Bhaskar</i>	106
Travel Time Reliability <i>Sharmili Banik, Anil Kumar, and Lelitha Vanajakshi</i>	109
Traffic Incident Management <i>Ruimin Li</i>	122
Traffic Incident Detection <i>Shuyan Chen and Yingjiu Pan</i>	128
Bottleneck <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	134
Flow Breakdown <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	143
Recurrent Congestion <i>Takahiro Tsubota</i>	154

Nonrecurrent Congestion <i>Takahiro Tsubota</i>	158
Freeway to Arterial Interfaces <i>Abolfazl Karimpour and Yao-Jan Wu</i>	162
Arterial Road Management <i>Jiaqi Ma, Yi Guo and Adekunle Adebisi</i>	169
Signalized Intersections <i>Chaitrali Shirke</i>	178
Protected Phase <i>Jiarong Yao, Yumin Cao, and Keshuang Tang</i>	185
Permitted Phase <i>Yumin Cao, Jiarong Yao, and Keshuang Tang</i>	196
Hook Turns: Implementation, Benefits, and Limitations <i>Sara Moridpour and Amir Falamarzi</i>	203
Traffic Signal Coordination <i>Rahim F. Benekohal</i>	213
Emergency Vehicle Priority (Preemption): Concept and Advancements <i>Chaitrali Shirke</i>	221
Signalized Roundabouts <i>Yetis Sazi Murat and Rui-jun Guo</i>	227
Turbo Roundabouts: Design, Capacity and Comparison With Alternative Types of Roundabouts <i>Marco Guerrieri and Raffaele Mauro</i>	238
Capacity of an Intersection <i>Ashish Verma and Milan Mathew Thomas</i>	247
Delay <i>Shinji Tanaka</i>	258
Queue Length <i>Shinji Tanaka</i>	263
Local Area Traffic Management <i>Michael A.P. Taylor</i>	268
On-Street Parking <i>Jun Chen and Guang Yang</i>	278
Off-Street Parking <i>Jun Chen and Guang Yang</i>	285
Parking Information Systems <i>Behrang Assemi and Douglas Baker</i>	289
Performance-Based Parking Management <i>Douglas Baker and Behrang Assemi</i>	299
Transit Priority <i>Wanjing Ma, Qiheng Lin, and Ling Wang</i>	307
Adaptive Bus Control <i>Monica Menendez</i>	315
Transit Fare Collection <i>Mahmoud Mesbah and Kamal Khanali</i>	325

Transit Information Systems <i>Ankit Kumar Yadav and Nagendra R Velaga</i>	331
Tram Lane Configurations and Driving Rules <i>Farhana Naznin</i>	338
Railway Crossing <i>Chunliang Wu and Inhi Kim</i>	341
Pedestrian Crossing (Crosswalk) <i>Miho Iryo-Asano, Wael K.M. Alhajyaseen, and Koji Suzuki</i>	346
Bike-Sharing System: Uncovering the “1Success Factors” <i>S.K. Jason Chang and Amanda Fernandes Ferreira</i>	355
Airspace Systems Technologies-Overview and Opportunities <i>Banavar Sridhar and Gano B. Chatterji</i>	363
Air Traffic Flow and Capacity Management <i>Li Weigang and Cristiano P Garcia</i>	380
Port Management <i>Maria G. Burns</i>	390
Port Performance Measurement from a Multistakeholder Perspective <i>Min-Ho Ha, Zaili Yang, and Young-Joon Seo</i>	396
Transport Modeling and Data Management <i>Chandra Bhat</i>	407
Computational Methods and Data Analytics <i>Bilal Farooq and David López</i>	408
Activity-Based Models <i>Renato Guadamuz and Rajesh Paleti</i>	414
Advanced Traveler Information Systems <i>Stephen D. Boyles</i>	418
Demand-Responsive Transit, Evaluation Studies <i>Sebastián Raveau</i>	423
Location Choice Models <i>Adam Wilkinson Davis</i>	428
Full Feedback and Equilibrium Modeling in Urban Travel Forecasting <i>Yu (Marco) Nie, Jun Xie, and David Boyce</i>	432
The Use and Value of Geographic Information Systems in Transportation Modeling <i>Ming Zhang</i>	440
Latent Demand and Induced Travel <i>Charisma F. Choudhury</i>	448
ICT, Virtual and In-Person Activity Participation, and Travel Choice Analysis <i>Jacek Pawlak and Giovanni Circella</i>	452
Public Transit Ridership Forecasting Models <i>Ipsita Banerjee, Deepa L, and Abdul Rawoof Pinjari</i>	459
Spatial Mismatch, Job Access, and Reverse Commuting <i>Gian-Claudia Sciarra</i>	468

Microsimulation and Agent-Based Models in Transportation <i>Milos Balac</i>	473
Choice Models in Transportation <i>Naveen Chandra Iraganaboina and Naveen Eluru</i>	477
Multi-Criteria Decision Analysis <i>Zhanmin Zhang and Srijith Balakrishnan</i>	485
The National Household Travel Survey Data Series (NPTS/NHTS) <i>Nancy McGuckin</i>	493
Route Choice and Network Modeling <i>Emma Frejinger and Maëlle Zimmermann</i>	496
Departure Time Choice Modeling <i>Khandker Nurul Habib</i>	504
Traveler Responses to Congestion <i>David T. Ory and Gayathri Shivaraman</i>	509
Origin-Destination Demand Estimation Models <i>William H.K. Lam, Hu Shao, Shuhan Cao, and Hai Yang</i>	515
Parking Demand Models <i>S.C. Wong, Zhi-Chun Li, and William H.K. Lam</i>	519
Pavement Management Systems <i>Senthilmurugan Thyagarajan</i>	524
Residential Location Choice Models <i>Shlomo Bekhor and Sigal Kaplan</i>	531
Transport Demand Management <i>Feiyang Zhang and Becky P.Y. Loo</i>	537
Traffic Flow Analysis <i>H. Michael Zhang and Jia Li</i>	544
Autonomous Vehicles and Transportation Modeling <i>Annesha Enam, Felipe de Souza, Omer Verbas, Monique Stinson, and Joshua Auld</i>	557
Ride-Hailing and Travel Demand Implications <i>Felipe F. Dias</i>	564
Transportation Modeling and Planning Software <i>Joel Freedman</i>	569
Transportation Statistics and Databases <i>Taha Hossein Rashidi</i>	574
Travel Surveys <i>Stacey G. Bricka</i>	587
Travel Demand Forecasting: Where Are We and What Are the Emerging Issues <i>Thomas F. Rossi</i>	590
Travel Model Calibration and Validation <i>Ram M. Pendyala</i>	596
Trip Chaining Analysis <i>Cynthia Chen and Yusak Susilo</i>	606
Vehicle Ownership Models <i>Dr. Sören Groth and Prof. Dr. Dirk Wittowsky</i>	612

Bicycle Sharing/Bikesharing <i>Catherine Morency and Jean-Simon Bourdeau</i>	617
Carsharing <i>Shiva Habibi and Frances Sprei</i>	623
Urban Recreational Travel <i>Long Cheng and Frank Witlox</i>	629
Place Perception and Travel Behavior <i>Kathleen Deutsch-Burgner and Konstadinos G. Goulias</i>	635

VOLUME 5

Transport Modes <i>Edoardo Marcucci</i>	1
Infrastructure Transport Investments, Economic Growth and Regional Convergence <i>Xavier Fageda and Cecilia Olivieri</i>	2
Transport Modes and an Aging Society <i>Charles B.A. Musselwhite and Theresa Scott</i>	6
Sustainable Mobility Paths <i>Erling Holden and Geoffrey Gilpin</i>	13
Vehicles that Drive Themselves: What to Expect with Autonomous Vehicles <i>Michele D. Simoni and Kara M. Kockelman</i>	19
Transport Modes and Tourism <i>Ila Maltese and Luca Zamparini</i>	26
Transport Modes and Accessibility <i>Bert van Wee</i>	32
Transport Modes and Globalization <i>Jean-Paul Rodrigue</i>	38
Transport Modes and Cities <i>Erick Guerra and Gilles Duranton</i>	45
Transport Modes and Remote Areas	51
Modeling Mode Choice in Freight Transport <i>Lóránt Tavasszy and Gerard de Jongb</i>	57
Travel Mode Choice as Reasoned Action <i>Sebastian Bamberg, Icek Ajzen, and Peter Schmidt</i>	63
Energy Consumption of Transport Modes <i>Zissis Samaras and Ilias Vouitsis</i>	71
Transport Modes and People With Limited Mobility <i>Roger L Mackett</i>	85
Transport Modes and Commuters <i>Colin G. Pooley</i>	92
Shopping and Transport Modes <i>Antonio Comi</i>	98
Transport Modes and Health <i>Jennifer S. Mindell and Sandra Mandic</i>	106

Multimodality in Transportation <i>Sören Groth and Tobias Kuhnimhof</i>	118
Transport Modes and Disasters <i>Brian Wolshon</i>	127
Big Data for Public Transport Planning <i>Jan-Dirk Schmöcker</i>	134
Active Transport: Heterogeneous Street Users Serving Movement and Place Functions <i>Regine Gerike, Stefan Hubrich, Caroline Koszowski, Bettina Schröter, and Rico Wittwer</i>	140
Electric Vehicles <i>Christine Eisenmann, Daniel Görges, and Thomas Franke</i>	147
Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service <i>Susan Shaheen, PhD and Adam Cohen</i>	155
Adoption of new travel information platforms <i>Sigal Kaplan</i>	160
ICT and Transport Modes <i>Galit Cohen-Blankshtain</i>	165
Mode Choice and Life Events <i>Joachim Scheiner</i>	171
Introduction to Air Transport <i>Milan Janić</i>	178
The History of Air Transportation <i>Richard P. Hallion</i>	192
The Geography of Air Transport <i>Lucy Budd and Stephen Ison</i>	198
The Future of Air Transport <i>Rico Merkert and James Bushell</i>	203
Air Transport and Its Territorial Implications <i>Lanfranco Senn</i>	208
Next Generation Travel: Young Adults' Travel Patterns <i>Tobias Kuhnimhof and Scott Le Vine</i>	215
Airport Network Planning and Its Integration with the HSR System <i>Francesca Pagliara, Juan Carlos Martín, and Concepción Román</i>	222
Airport Management <i>Peter Forsyth and Hans-Martin Niemeier</i>	229
Airport Regulation <i>Achim I. Czerny</i>	234
Air Route Planning and Development <i>Renan Peres de Oliveira and Gui Lohmann</i>	241
Airline Management <i>Sveinn Vidar Gudmundsson</i>	249
Air Cargo <i>Volodymyr Bilotkach</i>	258

Airline Regulation <i>Andrew R. Goetz</i>	263
Air Vehicles Classification <i>Vincenzo Torre</i>	269
Aircraft Manufacturing <i>Antonio Sollo</i>	290
A Geography of Road Transport in Cities <i>Cristian Domarchi and Juan de Dios Ortúzar</i>	300
The Future of Road Transport <i>Preston L. Schiller</i>	306
Road Transportation and Territorial Scale <i>Ana M. Condeço-Melhorado</i>	315
Road Modes: Walking <i>Kevin Manaugh, PhD Associate Professor</i>	320
Bus Public Transport Planning and Operations <i>Ehab Diab and Ahmed El-Geneidy</i>	326
Street Design for Active Travel <i>Bruce Appleyard</i>	333
Transit Planning and Management <i>Zakhary Mallett and Marlon G Boarnet</i>	349
Road Infrastructure: Planning, Impact and Management <i>Jos Arts, Wim Leendertse, and Taede Tillema</i>	360
Road Transport Planning at the Urban Scale <i>David A. King and Kevin J. Krizek</i>	373
Road Traffic Regulation: Road Pricing and Environmental Quality <i>Marco Percoco</i>	378
Car Ownership and Car Use: A Psychological Perspective <i>J.L. Veldstra, A.B. Ünal, E.M. Steg</i>	384
Road Transport: E-Scooters <i>Gysele Lima Ricci and Klaus Bogenberger</i>	391
Railway Station and Network Planning <i>Ingo Arne Hansen</i>	399
Introduction to Rail Transport <i>Chris Nash and Tony Fowkes</i>	406
The History of Rail Transport <i>Carlo Ciccarelli, Andrea Giuntini, and Peter Groote</i>	413
The Geography of Rail Transport <i>Frédéric Dobruszkes and Amparo Moyano</i>	427
Rail Transport and Territorial Scale <i>Prof. Andrés Monzón and Dr. Elena López</i>	437
Railway Management <i>Vassilios A. Profillidis</i>	444
Railway Terminal Regulation <i>Nacima Baron</i>	454

Service Network Design for Freight Railroads <i>Teodor Gabriel Crainic</i>	464
Subway Systems <i>Guillaume Monchambert, Daniel Hörcher, Alejandro Tirachini, and Nicolas Coulombel</i>	471
Railway Company Management <i>Vilius Nikitinas, Skaistė Miliauskaitė</i>	479
Regulation of Rail Infrastructure and Services <i>Javier Campos</i>	485
Rail Vehicle Classification <i>Christos Pyrgidis and Alexandros Dolianitis</i>	490
Introduction to Maritime Shipping <i>Christa Sys and Thierry Vanelslander</i>	508
The Geography of Maritime Transport <i>César Ducruet and Justin Berli</i>	517
The Future of Maritime Transport <i>Harilaos N. Psaraftis</i>	535
Maritime Transport and Territorial Scale <i>Brian Slack</i>	540
Containerization and the Port Industry <i>Hercules Haralambides</i>	545
Port Management <i>Lourdes Trujillo, Daniel Castillo Hidalgo, and Manuel Herrera</i>	557
Economic and Environmental Regulation in the Port Sector <i>Beatriz Tovar and Alan Wall</i>	563
Maritime Route Planning <i>Johan Woxenius</i>	570
Inland Waterway Transport and Inland Ports: An Overview of Synchromodal Concepts, Drivers, and Success Cases in the IWW Sector <i>Behzad Behdani, Bart Wiegman, and Yun Fan</i>	577
Maritime Company Passenger Management/Liner Industry <i>Claudio Ferrari and Alessio Tei</i>	587
Cruise Industry <i>Athanasios A. Pallis and Aimilia A. Papachristou</i>	593
International Maritime Regulation: Closing the Gaps Between Successful Achievements and Persistent Insufficiencies <i>Laurent Fedi</i>	600
Ship Classification <i>Gareth C. Burton and Mimosa T. Miller</i>	607
The Shipbuilding Industry and its Interactions With Shipping <i>Paul William Stott</i>	617
Methods for Designing Public Transport Networks <i>Zain Ul Abedin and Avishai (Avi) Ceder</i>	625
Space Transportation <i>Mark Hempsell</i>	638

Pipelines	646
<i>Franco Cotana and Mattia Manni</i>	
Women and Transport Modes	656
<i>Priya Uteng, PhD and Yusak Susilo, PhD</i>	
Transport Modes and Big Data	665
<i>Hannah D Budnitz, Emmanouil Tranos, and Lee Chapman</i>	
Transport Modes and Inequalities	671
<i>Caroline Mullen</i>	
Railway Traffic Management	678
<i>Francesco Corman</i>	
Indoor Transportation	684
<i>Lutfi Al-Sharif</i>	
Bicycle as a Transportation Mode	697
<i>Raktim Mitra and Paul M. Hess</i>	
Urban Air Mobility: Opportunities and Obstacles	702
<i>Adam Cohen and Susan Shaheen, PhD</i>	
Transport Modes and Sustainability	710
<i>Long Cheng, Jonas De Vos, and Frank Witlox</i>	

VOLUME 6

Introduction to Transport Policy and Planning	1
<i>Maria Attard</i>	
Workplace Parking Levy	2
<i>Stephen Ison and Lucy Budd</i>	
Air Transport	7
<i>Lucy Budd and Stephen Ison</i>	
Mobility as a Service	12
<i>Milo N. Mladenović</i>	
Bicycle Sharing	19
<i>Cyrille Médard de Chardon</i>	
Light Rail	31
<i>Fiona Ferbrache</i>	
Planning Tourism Travel	39
<i>Luca Zamparini</i>	
Transport Planning and Management and its Implications in Chinese Cities	44
<i>Mengqiu Cao</i>	
Road Safety	51
<i>George Yannis and Eleonora Papadimitriou</i>	
Mobility Planning and Policies for Older People	59
<i>Charles B A Musselwhite</i>	
Electric Mobility	64
<i>Graham Parkhurst</i>	
Transport Policy and Governance	73
<i>Lisa Hansson</i>	

Transferability of Urban Policy Measures <i>Paul Martin Timms</i>	77
Accessibility Tools for Transport Policy and Planning <i>Benjamin Büttner</i>	83
Demand Responsive Transport <i>Marcus Enoch</i>	87
Transport and Climate Change <i>Robin Hickman and Christine Hannigan</i>	94
Taxicabs and Microtransit <i>David A. King</i>	101
Land-Use and Transport Planning <i>Luis A. Guzman</i>	107
Parking <i>Stephen Ison and Lucy Budd</i>	113
Transport Planning in the Global South <i>Daniel Oviedo and Mariajose Nieto-Combariza</i>	118
Planning for Children's Independent Mobility <i>E. Owen D. Waygood and Raktim Mitra</i>	125
The Politics of Mobility Policy <i>Geoff Vigar</i>	131
Technology Enabled Data for Sustainable Transport Policy <i>Susan M. Grant-Muller, Mahmoud Abdelrazek, Hannah Budnitz, Caitlin D. Cottrill, Fiona Crawford, Charisma F. Choudhury, Teddy Cunningham, Gillian Harrison, Frances C. Hodgson, Jinhyun Hong, Adam Martin, Oliver O'Brien, Claire Papaix, and Panagiotis Tsoleridis</i>	135
Car Sharing <i>Cyriac George and Tanu Priya Uteng</i>	142
Toward a More Holistic Understanding of Mega Transport Project (MTP) Success <i>John Ward</i>	147
Equity Considerations in Transport Planning <i>Karel Martens</i>	154
Planning for Rail Transport <i>Simon P. Blainey</i>	161
Connected and Autonomous Vehicles: Priorities for Policy and Planning <i>Dr. Alexandros Nikitas</i>	167
Gendered Mobility <i>Sheila Mitra-Sarkar</i>	173
Modeling and Simulation for Transport Planning <i>Michela Le Pira, Giuseppe Inturri, and Matteo Ignaccolo</i>	184
Externalities and External Costs in Transport Planning <i>Silvio Nocera</i>	191
Planning for Public Transport with Automated Vehicles <i>Gonçalo Homem de Almeida Rodriguez Correia</i>	198
Urban Congestion Charging in Transport Planning Practice <i>Ida Kristoffersson and Maria Börjesson</i>	206

Energy and Transport Planning <i>Debbie Hopkins and Christian Brand</i>	214
Customer Satisfaction as a Measure of Service Quality in Public Transport Planning <i>Laura Eboli and Gabriella Mazzulla</i>	220
Car Sharing and the Impact on New Car Registration <i>Mario Intini and Marco Percoco</i>	225
Evaluation Methods in Transport Policy and Planning <i>Niek Mouter</i>	230
Transport and Air Quality Planning and Policy <i>Dr Fabio Galatioto</i>	236
Cycling Policies <i>Esther Anaya-Boig</i>	241
Community Severance <i>Paulo Anciaes and Jennifer S. Mindell</i>	246
Planning for Bus Priority <i>Claus H. Sørensen, Fredrik Pettersson, and Joel Hansson</i>	254
High-Speed Rail and the City <i>Marie Delaplace</i>	261
Public Engagement in Transport Planning <i>Miriam Ricci</i>	266
Long-Distance Travel <i>Giulio Mattioli and Muhammad Adeel</i>	272
Urban Freight Policy <i>Laetitia Dablanç</i>	278
Regional Transport Planning <i>Chia-Lin Chen</i>	286
Transport Project Financing <i>Romeo Danielis and Lucia Rotaris</i>	292
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	298
Transitions and Disruptive Technologies in Transport Planning <i>Kate Pangbourne and Maria Attard</i>	309
Community Transport: Filling the Gaps for Those in Need of Mobility <i>Ian Shergold</i>	314
Fundamental Emerging Concepts and Trends for Environmental Friendly Urban Goods Distribution Systems <i>Sandra Melo</i>	320
Plateau Car <i>David Metz</i>	324
Emerging Trends in Transport Demand Modeling in the Transition Toward Shared Mobility and Autonomy <i>Patrizia Franco</i>	331
Policy and Planning for Walkability <i>Carlos Cañas Sanz and Maria Attard</i>	340

Public Transport Subsidy and Regulation <i>Jonathan Cowie</i>	349
Urban Regeneration and Transportation Planning <i>Thomas Vanoutrive</i>	356
Social and Distributional Impact Assessment in Transport Policy <i>Laura Walker and Angela Curl</i>	361
Mobility Planning for Healthy Cities <i>Ersilia Verlinghieri</i>	368
Transport Demand Management <i>Begoña Guirao</i>	374
Planning and the Global Movement of Goods and Commodities <i>Christopher Clott and Chris Petrocelli</i>	380
Public Transport Network Planning <i>Corinne Mulley and John D. Nelson</i>	388
A Timely Perspective on Planning for Ageing Infrastructure <i>Anthony Perl</i>	395
The role of media in transport planning and the transport policy process <i>Özgül Ardiç and J.A. Annema</i>	400
Travel Plans <i>Stephen Potter and Marcus Enoch</i>	408
Home Deliveries and their Impact on Planning and Policy <i>Rosário Macário</i>	413
Planning for Safe and Secure Transport Infrastructure <i>Per Erik Gårder</i>	418
Sensors and Data Driven Approaches in Transport <i>Mohammad Sadrani and Constantinos Antoniou</i>	426
Planning Active Travel and School Transport <i>Fahimeh Khalaj, Dorina Pojani, and Sara Alidoust</i>	432
VOLUME 7	
Introduction to Transport Psychology <i>Carlo Prato</i>	1
From Self-reports to Auto-Tech-Detect (ATD)-based Self-reports in Traffic Research <i>Türker Özkan and Timo Lajunen</i>	2
Observational Field Studies in Traffic Psychology <i>Tova Rosenbloom and Hodaya Levy</i>	8
Driving Simulators <i>Karel A. Brookhuis</i>	14
Naturalistic Driving Studies: An Overview and International Perspective <i>Johnathon P. Ehsani, Joanne L. Harbluk, Jonas Bärghman, Ann Williamson, Jeffrey P. Michael, Raphael Grzebieta, Jake Olivier, Jan Eusebio, Judith Charlton, Sjaanie Koppel, Kristie Young, Mike Lenné, Narelle Haworth, Andry Rakotonirainy, Mohammed Elhenawy, Gregoire Larue, Teresa Senserrick, Jeremy Woolley, Mario Mongiardini, Christopher Stokes, Paul Boase, John Pearson, and Feng Guo</i>	20

A Detailed Approach to Qualitative Research Methods <i>Sonja Forward and Lena Levin</i>	39
Behavioral Change <i>Sonja Haustein</i>	46
Habitual Behavior <i>Carlo G. Prato</i>	54
Driving Behavior and Skills <i>Timo Lajunen and Türker Özkan</i>	59
The Multidimensional Driving Style Inventory <i>Orit Taubman - Ben-Ari</i>	65
Risk Perception in Transport: A Review of the State of the Art <i>Trond Nordfjrn, An-Magritt Kummeneje, Mohsen F. Zavareh, Milad Mehdizadeh, and Torbjørn Rundmo</i>	74
Social-Symbolic and Affective Aspects of Car Ownership and Use <i>Birgitta Gatersleben</i>	81
Pitfalls of Statistical Methods in Traffic Psychology <i>J.C.F. de Winter and D. Dodou</i>	87
Data Analysis: Structural Equation Models <i>Marco Diana</i>	96
Data Analysis: Integrated Choice and Latent Variable Models <i>Carlo G Prato</i>	102
Explaining Data Analysis Using Qualitative Methods <i>Lena Levin and Sonja Forward</i>	107
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	113
Driver Aggression and Anger <i>Mark J.M. Sullman and Amanda N. Stephens</i>	121
Speeding: A "Tragedy of the Commons" Behavior <i>Bryan E. Porter, Thomas D. Berry, and Kristie L. Johnson</i>	130
Cycling as a Mode Choice: Motivational Psychology <i>Sigal Kaplan</i>	136
Motorcyclists <i>Narelle Haworth</i>	144
Drivers' Hazard Perception Skill <i>Mark S. Horswill and Andrew Hill</i>	151
Driver Education and Training for New Drivers: Moving beyond Current 'Wisdom' to New Directions <i>Teresa Senserrick, Oscar Oviedo-Trespalcacios, David Rodwell, and Sherrie-Anne Kaye</i>	158
Road Safety Advertising: What We Currently Know and Where to From Here <i>Ioni Lewis, Barry Watson, Katherine M. White, and Sonali Nandavar</i>	165
Traffic Law Enforcement Theories and Models <i>Richard Tay</i>	171
Satisfaction with Travel and the Relationship to Well-Being <i>Tommy Gørling and Filip Fors Connolly</i>	177
	182

Electromobility: History, Definitions and an Overview of Psychological Research on a Sustainable Mobility System <i>Josef F. Krems and Isabel KreiÃŸig</i>	
Sharing: Attitudes and Perceptions <i>Yi Wen and Christopher R Cherry</i>	187
Women's Travel Patterns, Attitudes, and Constraints Around the World <i>Sandra Rosenbloom</i>	193
Driver Stress and Driving Performance <i>Lisa Dorn</i>	203
Introduction to Sustainability and Health in Transportation <i>Roger Vickerman</i>	225
Sustainable Development Goals and Health <i>Rosa Surinach</i>	226
Human Ecology <i>Roderick J. Lawrence</i>	234
Car- Free Cities <i>Haneen Khreis and Mark J. Nieuwenhuijsen</i>	240
Superblocks Base of a New Model of Mobility and Public Space. Barcelona as an Example <i>Salvador Rueda Palenzuela</i>	249
Resilience of Transport Systems <i>ErikJenelius and Lars-GöranMattsson</i>	258
Impact of Shipping to Atmospheric Pollutants: State-of-the-Art and Perspectives <i>Daniele Contini and Eva Merico</i>	268
Noise Pollution From Transport <i>Marianna Jacyna, Emilian Szczepański, Konrad Lewczuk, Mariusz Izdebski, Ilona Jacyna-Gotda, Michał, Kłodawski, Paweł, Gołda, Piotr Goł, and Ebiowski</i>	277
Visual Impacts From Transport <i>Paulo Rui Anciaes</i>	285
Light Pollution <i>John D. Bullough</i>	292
Wildlife Crossings and Barriers <i>Scott D. Jackson</i>	297
Environmental Justice, Transport Justice, and Mobility Justice <i>Devajyoti Deka</i>	305
Transport Noise and Health <i>Elisabete F. Freitas, Emanuel A. Sousa, and Carlos C. Silva</i>	311
Climate Change and Health, Related to Transport <i>Ersilia Verlinghieri</i>	320
Urban Greenspace, Transportation, and Health <i>Payam Dadvand and Mark J. Nieuwenhuijsen</i>	327
Transport Access and Health <i>Alireza Ermagun</i>	335
Social Exclusion and Health, Related to Transport <i>Roger L. Mackett</i>	341

Burden of Disease Assessment <i>David Rojas-Rueda</i>	347
Achieving a Near-Zero CO <i>Lewis M. Fulton</i>	353
Disabled Travelers <i>Bryan Matthews</i>	359
Health Impacts of Connected and Autonomous Vehicles <i>Soheil Sohrabi</i>	364
Electric Vehicles and Health <i>Kanok Boriboonsomsin</i>	372
Shared Mobility Opportunities and their Computational Challenges for Improving Health-Related Quality of Life <i>Cristiano Martins Monteiro, Cláudia Aparecida Soares Machado, Adelaide Cassia Nardocci, Fernando Tobal Berssaneti, José Alberto Quintanilha, and Clodoveu Augusto Davis</i>	376
Bike Sharing and Health <i>David Rojas-Rueda and Mark J. Nieuwenhuijsen</i>	384
E-Bikes and Health <i>Aslak Fyhri and Hanne Beate Sundfør</i>	393
<i>Index</i>	399

Freight Transport and Logistics

Freight Transport and Logistics

Sharon Cullinane, Kevin Cullinane, School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Defining Logistics	1
Freight Transportation	3
Logistics Decisions	5
Biographies	6
References	7

Defining Logistics

So, what is logistics and how is it different from supply chain management? Logistics can be defined as the following:

"Logistics management is that part of supply chain management that plans, implements, and controls the efficient, effective forward and reverse flow and storage of goods, services and related information between the point of origin and the point of consumption in order to meet customers' requirements."
(Council of Supply Chain Management Professionals, 2019).

In contrast, the same organization defines supply chain management as:

"Supply chain management encompasses the planning and management of all activities involved in sourcing and procurement, conversion, and all logistics management activities. Importantly, it also includes coordination and collaboration with channel partners, which can be suppliers, intermediaries, third party service providers, and customers. In essence, supply chain management integrates supply and demand management within and across companies."
(Council of Supply Chain Management Professionals, 2019).

These definitions indicate that logistics management should be considered as a subset of supply chain management. Where the precise boundary between the two is actually drawn, however, is rather fluid. The definition of logistics has become wider over time to include new functions, and it continues to expand today.

Logistics is an integrating and coordinating function, which encompasses the following elements:

- demand forecasting,
- supply forecasting,
- inventory management,
- communications,
- warehousing and storage,
- transportation,
- customer service,
- reverse logistics,
- materials handling,
- order processing,
- procurement,
- order fulfillment,
- logistics network design,
- packaging and assembly, and
- staffing.

Logistics is about delivering the "7 Rights": The Right product in the Right place at the Right time in the Right quantity to the Right customer in the Right condition at the Right cost (Christopher, 2016). A successful logistics function within any organization will integrate seamlessly with other company functions such as marketing, production, finance, IT, and human resource management. In the absence of such seamless integration between the logistics function and other organizational functions, a company's profitability will be undermined so that it is not reaching its full potential. As a consequence of this or any other form of underperformance of the

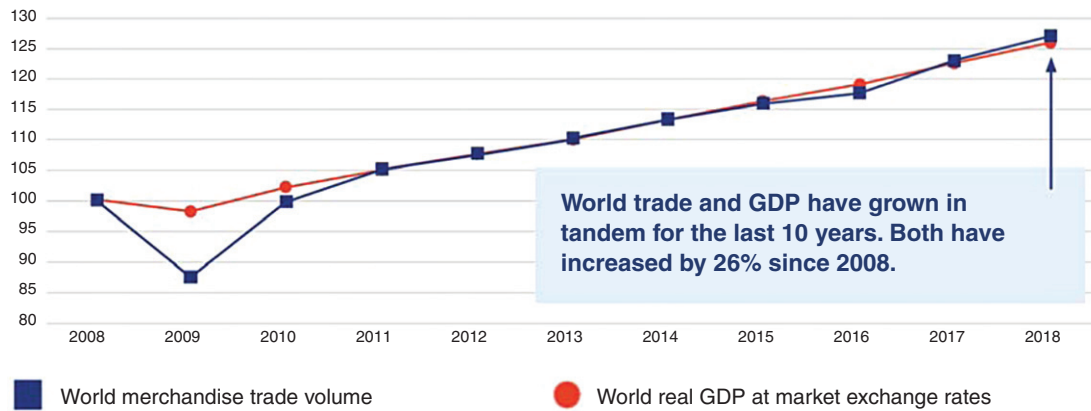


Figure 1 World merchandise trade volume and real GDP at market exchange rates, 2008–2018 (indices, 2008 = 100). Source: WTO (2019a)



Figure 2 World merchandise trade volume and real GDP growth, 2011–2018 (annual percentage change). Source: WTO (2019a)

logistics function within an organization, when aggregating across all underperforming companies within a nation, it becomes clear that for any country's economy as a whole, GDP will not be maximized (Civelek et al., 2015).

In recent decades, increasing industrialization, the liberalization of national economies and the globalization of production and consumption have all served to fuel greater free trade and a growing demand for consumer products. Thus the trade in physical products (i.e., "merchandise trade") that is facilitated by the logistics industry has grown consistently over time (Fig. 1). In seeking to meet this greatly increased derived demand for its services, the logistics industry has responded by not only expanding to meet the demand, but also by adopting advances in technology that have enhanced efficiency and the quality and range of the services offered. As graphically illustrated in Fig. 1, because of the close historical correlation that exists between the volume of trade and world real GDP, it is actually vitally important that the global logistics industry continues to successfully respond to what appears to be continuous growth in merchandise trade volume.

To emphasize and quantify the precise nature and closeness of the relationship between the volume of merchandise trade and the value of real GDP, Fig. 2 focuses on the growth in both variables over time, as measured in terms of annual percentage change, to reveal how closely they have moved together over recent years.

In order to gain a better insight into the nature of the required response from the logistics industry in facilitating merchandise trade, Fig. 3 shows the most important product groups (in value terms) that are exported around the world, each of which gives rise to different generic forms of logistics solutions. Given the trends in merchandise trade volumes revealed in Figs. 1 and 2, but having a focus on export values rather than volumes, Fig. 3 also implies that export price inflation has not kept pace with the growth in export volumes. Potentially, this may be at least partially attributable to the increasing efficiency of the logistics industry, especially in the light of the economies of scale that are likely to be reaped from the handling of increasing export volumes.

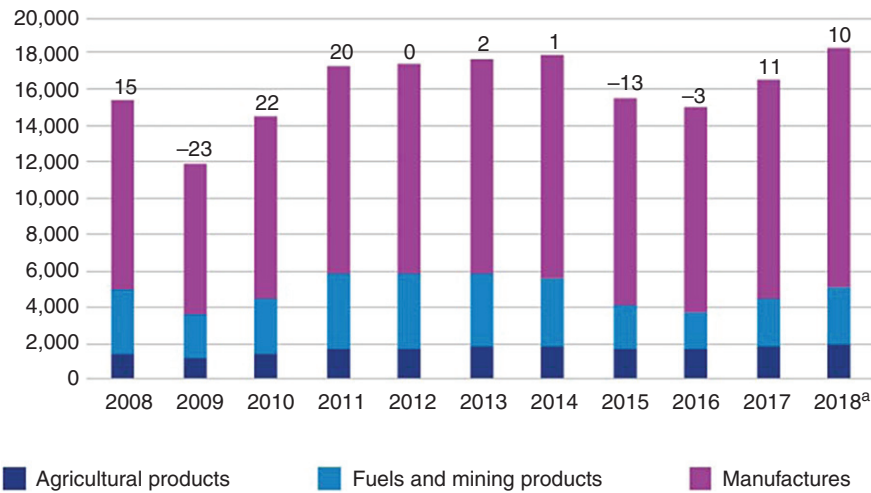


Figure 3 World merchandise exports by product group and annual growth, 2008–2018. ^a US\$ billion and average annual percentage change. Source: WTO (2019a)

In order to complete the macroeconomic picture of the demands that are placed upon the global logistics industry, Fig. 4 gives some insight into the major nations exporting merchandise trade (in value terms) and, by implication, the geographical scope of coverage the international logistics industry is required to respond to.

Freight Transportation

Logistics activities obviously revolve around the movement of freight that potentially can be carried out through the use of several modes of transport. Sometimes, the available modes of freight transport are used individually, on a *unimodal* basis. For example, the

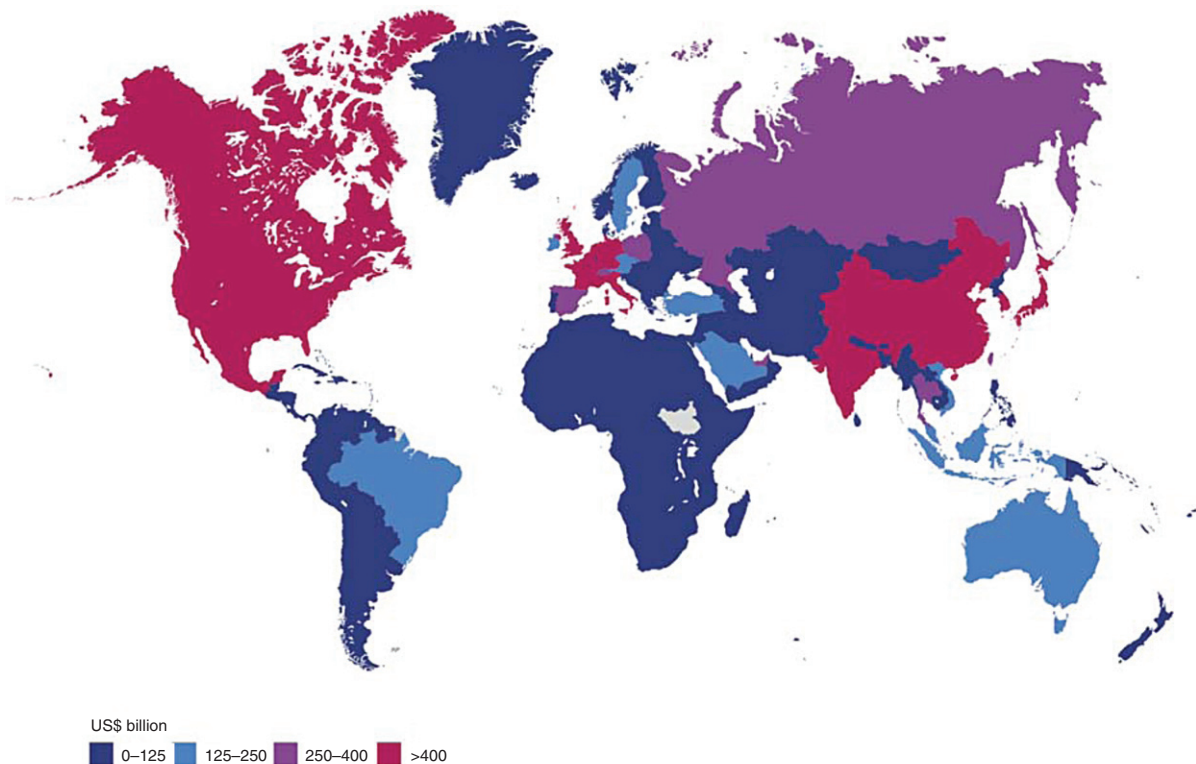


Figure 4 Economies by size of Merchandise Trade, 2018. Note: Includes significant reexports or imports for reexport. Source: WTO (2019b)

movement of a consignment by truck from the factory where it is produced direct to the premises of the end user. On other occasions, however, the available modes of freight transport may be used in combination, on a *multimodal* basis. For example, virtually all international movements of freight by sea or air will require a movement to and from port or airport via another mode, most usually by road but sometimes rail or other modes. The term “intermodal” refers exclusively to a specific subset of multimodal transport where a cargo is moved from origin to destination via two or more modes of transport, but without any transfer between different types of loading unit and, strictly speaking, under a single contract of carriage.

The amount of freight transportation can, of course, be measured in terms of simply the weight of the consignments/shipments/cargoes carried. However, most importantly, this fails to take account of the distance that the freight is carried. Conventionally, therefore, the amount of freight transportation taking place is typically measured in terms of weight and distance combined, most typically in the form of ton-miles or ton-kms. Typically, these are also the basic units used for enumerating the levels of supply and demand for freight transportation.

The main modes of freight transport are sea (using ships of various sizes and types), pipeline, inland waterway, rail, road (mostly using trucks of various shapes and sizes, but also including vans, bicycles etc.), and air (mostly using planes, but may in the future also include an element of drone transport). In addition to these mainstream modes, but often excluded, are cars—since people also carry freight in their cars. Each of these modes of transport has its own characteristics, advantages, and disadvantages, although their precise advantages and disadvantages vary geographically.

By implication, it should be apparent that the geographical scope of freight transport movements range from intercontinental, through international, national (sometimes referred to as domestic), regional, and local. As already suggested, different modes are likely to be used for different geographical levels. Thus, for intercontinental trade (e.g., from China to Europe), the main alternatives will be to move goods by either sea or air, depending on the quantity and the value of the goods being transported. At the other extreme, for local distribution, the main alternatives will be small truck, van, and in urban areas, perhaps moped, bicycle, etc.

In general terms, shipping is used for carrying large consignments of cargo and, almost invariably, represents the cheapest mode of freight transport per unit of weight carried. Because of this and the distances covered by cargoes moved by the shipping mode, it is often said that shipping carries over 90% of world trade, as measured in ton-miles (IMO, 2019) and that “without shipping, half the world would starve and the other half would freeze” (IMO, 2007). Over the past 50 years, estimates of total seaborne trade have grown by approximately a factor of 6, from just over 10 thousand billion ton-miles in 1970 (UNCTAD, 1977) to over 60 thousand billion ton-miles in 2019 (UNCTAD, 2019). The growth in seaborne trade by cargo type over the past 10 years is shown in Fig. 5. What is not really shown, however, is the phenomenal growth in container cargoes which has occurred over the past 50 years; as illustrated in Fig. 5, they now make up a significant proportion of seaborne trade carried, but back in 1970 containerized cargoes formed a mere component of what were classified as “other cargoes,” outside the main cargo types.

In contrast to shipping, freight transport by air is generally used for smaller, more valuable cargoes and, in the majority of cases, is the most expensive mode of freight transport, with the cost of other modes generally falling in-between. Shipping is, however, the slowest mode and air is the fastest and this means that the inventory costs of using shipping are higher than if an airplane is being used.

As mentioned earlier, different countries use different modes. Fig. 6 shows the modes of freight transport used within the European Union.

In most countries, freight transport will consist of a variety of types; movements of goods within the country, exports out of the country, imports into the country, hub movements through a country, and transit movements through a country. For some countries in Europe, such as Belgium, the Netherlands, and Germany, transit movements through the country are very important. In these

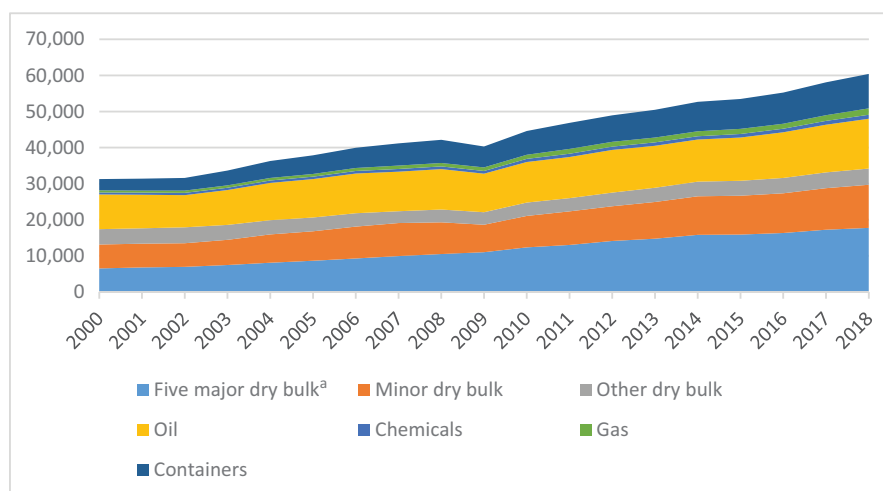


Figure 5 International maritime trade in cargo ton-miles, 2000-2019 (estimated billions of ton-miles). ^aIron ore, grain, coal, bauxite/alumina, and phosphate. Source: The figure is own compilation using data published in UNCTAD (2018), which was originally sourced from Clarkson's Research (2018)

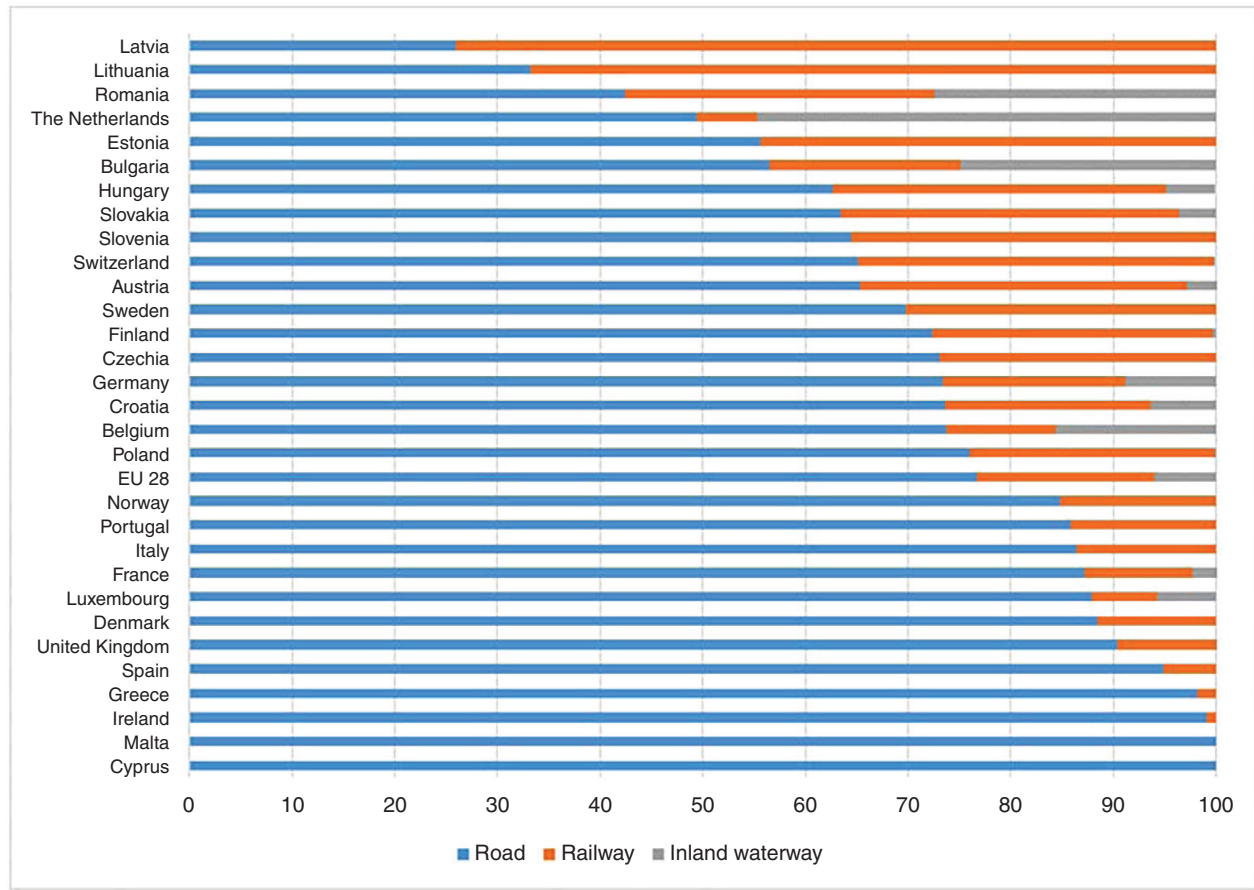


Figure 6 Modal split of inland freight transport within the EU, 2017 (% of total ton-kms). Source: The figure is own compilation, based on data from Eurostat (2019a)

cases, the country is merely a bridge between the cargo origin and destination and the vehicle carrying the products may not even stop. Fig. 7 shows the importance of this element of freight transport in the EU.

Logistics Decisions

Thus far, logistics and freight transport have been considered exclusively at a macro level. At a more microlevel, every organization will need to decide on what mode (or combination of modes) of transport it uses, but will also need to decide whether it owns and organizes its own logistics or whether it subcontracts some or all of its logistics to a specialist logistics provider (third party provider, or 3PL). Many large manufacturers use several (up to about 70) 3PLs. Some manufacturers will organize some of their logistics function, but subcontract other elements to specialists.

Over the past decade, there has been a movement toward the use of global logistics companies, such as Schenker, DHL, UPS, and Fedex. This is largely because production, consumption and trade has become more globalized. Global trade has increased the need for such mega-logistics companies but, in turn, the development of such companies has facilitated and enabled the globalization of trade. They basically promise delivery of any goods anywhere and achieve this through partnerships with local companies, as well as through the ownership of global logistics facilities.

Goods to be transported can be categorized into raw materials, semifinished products, finished products, and consumer products, or sometimes referred to as business-to-business (B2B) goods and business-to-consumer (B2C) goods. At a company level, logistics can be further categorized by dividing into “upstream” and “downstream,” where upstream refers to the logistics involved in getting goods to the company (which may involve the organization of many different inputs and companies) and “downstream” involves the logistics of getting goods from the company to the consumer (which could be another producer). Equivalently, these may also be referred to as “primary” and “secondary” logistics or “supply side,” and “demand side” logistics.

In the past, manufacturers and retailers worked on a “push” basis, where demand was forecasted and goods were produced/supplied based on these forecasts. Nowadays, most manufacturing/retailers work on a “pull” basis, where consumers buy products

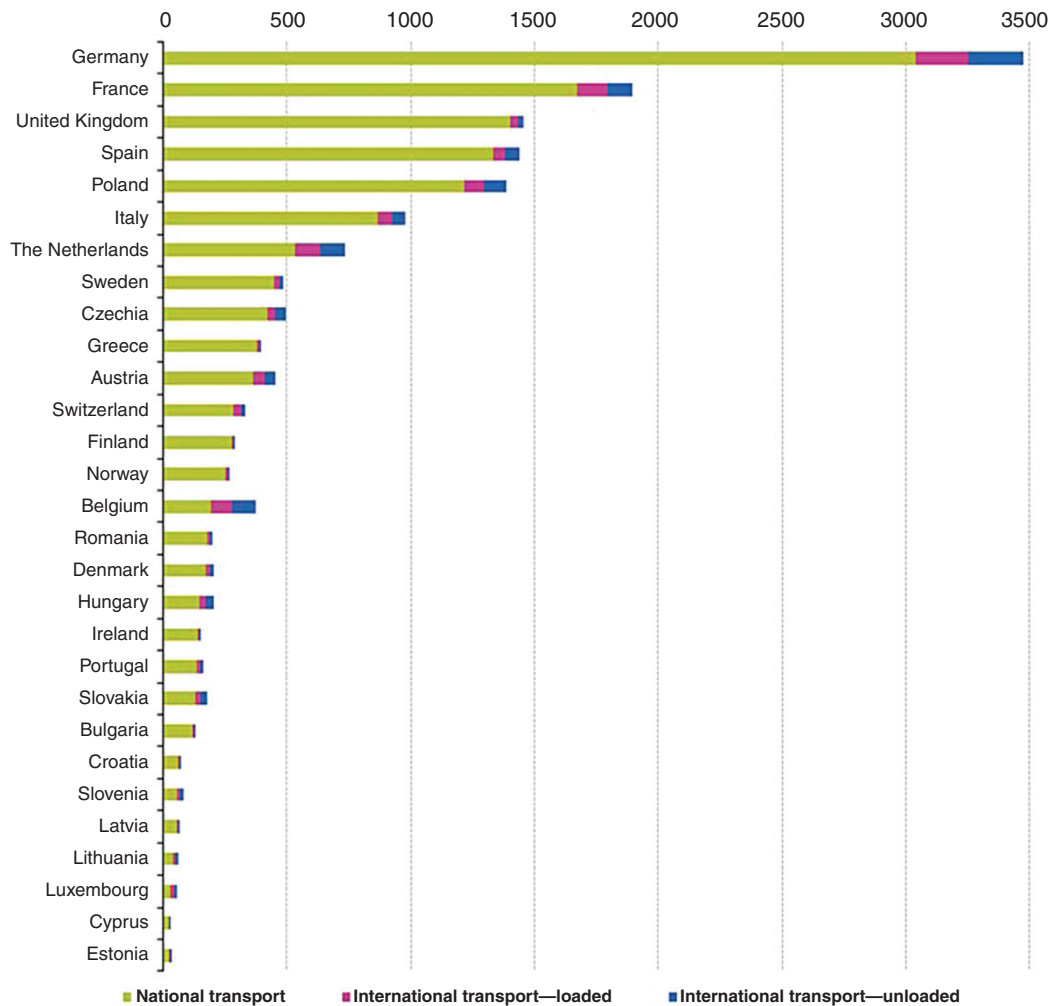


Figure 7 Transport of goods on countries' territory by type of transport in 2017 (million tons). *Source: From Eurostat (2019b)*

and this purchase triggers a “pull” for more of the product to be supplied/manufactured. The use of IT and digitalization has greatly facilitated this trend. The outcome is that inventory has been reduced and less waste is produced. “Just-in-time” and “lean” logistics have evolved to deal with this change. It remains to be seen how future systems and innovations, such as block chain technology for example, will impact worldwide logistics practices and freight transportation in particular.

Biographies

Sharon Cullinane is Professor of Sustainable Logistics at the University of Gothenburg Business School. Having gained her PhD in logistics from Plymouth University in 1987, she has lectured, researched, and published in the field of transport policy and the environment around the world. She has been employed at Heriot-Watt University in Edinburgh, the University of Hong Kong, Oxford University, the Egyptian National Institute of Transport, the Ecole Supérieur de Rennes, and Plymouth University. For many years she has been interested in e-commerce and its impact on the environment. Her current research interest lies in the area of returns in the e-commerce sector and its impacts on the environment, focusing particularly on the clothing sector in Sweden. She is also Program Coordinator of the Masters degree in Logistics and Transport Management at Gothenburg University.

Kevin Cullinane is Professor of International Logistics and Transport Economics at the University of Gothenburg. Kevin has been a logistics adviser to the World Bank and transport adviser to the governments of Scotland, Ireland, Hong Kong, Egypt, Chile, and the United Kingdom. He holds an Honorary Professorship at the University of Hong Kong and is a Visiting Professor at the VTI (the Swedish National Road and Transport Research Institute), Liverpool John Moores University, Dalian Maritime University, and Shanghai Maritime University. Prior to joining the University of Gothenburg on a permanent basis, he was appointed to the lifetime position of Honorary Visiting Professor at the University. Kevin has personally won research and consultancy projects to the value of \$US 6.5 million, many of which have informed the policies of national and/or regional governments. He has held other positions as the Director of the Transport Research Institute (TRI) at Edinburgh Napier University, Chair in Marine Transport & Management at Newcastle University, Professor and Head of the Department of Shipping & Transport Logistics at Hong Kong Polytechnic University, Head of the Centre for International Shipping & Transport at Plymouth University, Postdoctoral Research Fellow at the University of Oxford Transport Studies Unit and as Senior Partner in his own transport consultancy

company, based in Cairo, Rennes, Hong Kong, Edinburgh and now Gothenburg. Kevin has published 14 books and over 300 journal papers and is an Associate Editor of *Transportation Research D*, *Maritime Economics & Logistics* and the *International Journal of Applied Logistics*. He is also a member of the Editorial Boards of *Transportation Research A*, *Transport Reviews*, the *Annals of Maritime Studies*, the *Journal of Shipping & Logistics* and, formerly, of the *Proceedings of IMechE Part M: Journal of Engineering for the Maritime Environment* and the *International Journal of Logistics Management*. Kevin has been a Chartered Fellow of the Chartered Institute of Logistics & Transport since 1997. His expertise has been recognized at UK national level with his appointment to REF 2014 and REF 2021; the United Kingdom's national research evaluation exercise.

References

- Christopher, M., 2016. Logistics and supply chain management, fifth ed. Prentice-Hall, Harlow.
- Civelek, M.E., Uca, N., Çemberci, M., 2015. The mediator effect of logistics performance index on the relation between global competitiveness index and gross domestic product. *Eur. Sci. J.* 11 (13). Available from: <https://ssrn.com/abstract=3338312>.
- Clarksons Research, 2018. Shipping Review and Outlook, Spring, Clarksons Research, London.
- Council of Supply Chain Management Professionals, 2019. CSCMP supply chain management definitions and glossary. https://cscmp.org/CSCMP/Educa/SCM_Definitions_and_Glossary_of_Terms.aspx (accessed 06.11.19).
- Eurostat, 2019a. Modal split of freight transport. https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=tran_hv_frmod&lang=en. (accessed 14.11.19).
- Eurostat, 2019b. File:Transport of goods on countries' territory by type of transport, 2017 (million tonnes)-up.png. [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Transport_of_goods_on_countries%27_territory_by_type_of_transport,_2017_\(million_tonnes\)-up.png](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Transport_of_goods_on_countries%27_territory_by_type_of_transport,_2017_(million_tonnes)-up.png) (accessed 14.11.19).
- IMO, 2007. World Maritime Day 2007: the IMO's response to current environmental challenges. IMO News, No. 2, 14–27.
- IMO, 2019. IMO Profile. <https://business.un.org/en/entities/13> (accessed 13.11.19).
- UNCTAD, 1977. Review of Maritime Transport 1975, TD/B/C.4/149/Rev.I, United Nations, New York.
- UNCTAD, 2018. Review of Maritime Transport 2018, UNCTAD/RMT/2018, United Nations, New York.
- UNCTAD, 2019. Review of Maritime Transport 2019, UNCTAD/RMT/2019, United Nations, New York.
- WTO, 2019a. World trade statistical review 2019, World Trade Organisation, Geneva.
- WTO, 2019b. Highlights of world trade, World Trade Organisation, Geneva. https://www.wto.org/english/res_e/statis_e/wts2019_e/wts2019chapter02_e.pdf (accessed 21.11.19).

Expanding the Perspective of Logistics and Supply Chain Management

David J. Closs, Department of Supply Chain Management, Eli Broad College of Business, Michigan State University, East Lansing, MI, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	8
Supply Definition and Activities	8
Extended Logistics and Supply Chain Applications	9
Product Supply Chain	10
Promotional Supply Chain	10
Bulk Material Supply Chain	12
Talent Supply Chain	12
Business to Consumer (B2C) Supply Chain	12
Recycling Supply Chain	13
Resource Supply Chain	13
Construction Supply Chain	13
Recovery Supply Chain	13
Humanitarian Supply Chain	13
Global Supply Chain	13
Durables Supply Chain	14
Agricultural Commodity Supply Chain	14
Innovative Supply Chains	14
Military Supply Chain	14
Clinical Trials Supply Chain	15
Conclusion	15
References	15

Introduction

Logistics and supply chain management have traditionally focused on the acquisition, manufacture and delivery of goods to consumers. These capabilities traditionally evolved from military applications but expanded significantly for use in more broad commercial settings. Historically, the capabilities have been grouped into the functions of purchasing, production, and logistics. The traditional focus has been on optimizing the performance of these functions. For example, purchasing, production and logistics decisions were made to minimize the expenses within their own operations.

More recently, firms have shifted their focus from functional to integrated process. The integrated process emphasizes the combination of functions and activities that must be coordinated to deliver value to customers and consumers. **Table 1** lists and describes the processes. The concept behind a process is a group of activities that use the functional resources (inventory, production equipment, transportation equipment, technology, etc.) to provide value to customers. In most cases, the processes closely match with organizational components of the firm, including product development/launch, purchasing, production, logistics, and post-sales support.

Table 1 Supply chain processes

Process	Description
Demand planning responsiveness	The assessment of demand and strategic design to achieve maximum responsiveness to customer requirements.
Customer relationship collaboration	The development and administration of relationships with customers to facilitate strategic information sharing, joint planning, and integrated operations.
Order fulfillment/service delivery	The ability to execute superior and sustainable order to delivery performance and related essential services.
Product/Service development launch	The participation in product service development and lean launch.
Manufacturing customization	The support of manufacturing strategy and facilitation of postponement throughout the supply chain.
Supplier relationship collaboration	The development and administration of relationships with suppliers to facilitate strategic information sharing, joint planning, and integrated operations.
Life cycle support	The repair and support of products during their life cycle includes warranty, maintenance, and repair.
Reverse logistics	The return and disposition of inventories in a cost effective, secure, and responsible manner.

Source: Bowersox et al. (2019, p. 9).

Supply Definition and Activities

Due to the many views and perspectives regarding supply chain, there are varying definitions that include different institutions, processes, functions, and activities. As such, there is not one common definition. There is not even a common set of processes or activities that most experts would agree to include. However, it is important that this article establishes a foundation by providing a definition and describes the range of applications for logistics and supply chain management. According to Bowersox et al. (2019), supply chain management is a set of processes to effectively and efficiently integrate suppliers, manufacturers, distribution centers, distributors, and retailers so that products are produced and distributed at the right quantities, to the right locations, and at the right time to minimize system-side costs, while achieving the consumer's desired value proposition.

What began during the last decade of the 20th century and will continue to unfold well into the future can be increasingly characterized as the information-based or digital supply chain. The information or digital enterprise will connect collaborating business organizations to drive a new order of relationships called supply chain management. Managers are increasingly improving and integrating traditional marketing, manufacturing, purchasing, and logistics practices. This information-based evolution has facilitated the development of the supply chain discipline, which can be defined as a coordinated, cross-functional strategy, involving both internal and external partners that utilize process and information to improve operating efficiency and leverage strategic positioning. Supply chain strategy applies the functions and processes to effectively and efficiently integrate suppliers, manufacturers, distribution networks and channels, as well as final consumers, all of which ensures that the firm's value proposition is achieved while minimizing total system cost.

While there are other definitions that come from different perspectives (Institute of Supply Management, Association for Supply Chain Management and Council of Supply Chain Management Professionals), there are common themes that suggest a common framework. These definitions emphasize key concepts including effective and efficient flow, crossfunctional collaboration, collaborative institutional partners, achieving the consumer's value proposition, and minimizing system-wide cost. Efficient and effective flow emphasizes the need for a firm to work collaboratively with other supply chain partners to deliver the product to the consumer at a minimum cost. The crossfunctional collaboration requires that the firm's internal functions, particularly those involved in the supply chain, work together to minimize waste and the duplication of time and resources. Achieving the consumer's value proposition means that the supply chain can deliver the product or solution in a form that can meet the unique consumer requirements. Finally, "deliver at minimum cost" means that the firm and its supply chain partners try to deliver the product or solution to the consumer while minimizing the total end-to-end cost for all activities occurring in the supply chain.

With this process perspective as background, my experience as the chairperson of the Department of Supply Chain Management at Michigan State University suggested a broader range of applications for supply chain principles. This perspective was developed through many conversations with recruiters at MSU. These discussions suggested a wide range of other applications for supply chain principles beyond the traditional product focus. Table 2 lists and describes some of these nontraditional applications.

Why is it important to take this broader perspective? While there are certainly many firms that are primarily concerned with the production and movement of traditional products, there are many more that are interested in these broader supply chain solutions. These other applications suggest many opportunities for supply chain professionals and extend the role that supply chain can have in firm competitiveness. While the typical supply chain is focused on a manufacturer with the support of suppliers, distributors, retailers, and supply chain service providers, there are many nontraditional environments where supply chain principles can be effectively applied.

Although many believe that supply chain principles and practices are primarily relevant for major manufacturing firms, Table 2 demonstrates that the principles can be applied in many other scenarios and environments. It is increasingly common for nonsupply chain executives to reach out to supply chain professionals when they are challenged with problems that have process and resource management characteristics. It is important for supply chain professionals to understand the environments that supply chain principles can be applied to. This insight can also result in a wide range of job opportunities for supply chain professionals and increased influence in their firm.

Extended Logistics and Supply Chain Applications

Whereas supply chain principles can be applied in these broader contexts, there needs to be some context refinements to define specific requirements. These refinements include: (1) service focus; (2) financial focus; (3) procurement focus; (4) manufacturing focus; (5) logistics focus; (6) opportunities for performance enhancement; and (7) product types. Service focus refers to the dimensions of the service requirements desired by customers. Financial focus refers to the benefits desired by the firm, such as profit, return on assets or satisfying humanitarian needs. In the case of procurement, manufacturing, and logistics, the focus would relate to whether the firm should emphasize economies of scale or flexibility. Opportunities for performance enhancement refers to the firm's ability to create competitive advantage using a low cost, service differentiation, or hybrid strategy. Product type refers to the product categories that would be appropriate for each application. Tables 3A–D lists these extended applications and some of the ways that they differ from typical supply chain applications. The follow sections discuss some of these differences.

Table 2 Supply chain applications

<i>Supply chain applications</i>	<i>Description</i>
Product supply chain	Traditional supply chain involving suppliers, manufacturers, distributors, and retailers for consumer products. This is the primary focus of many supply chain operations and most commonly covered in classes and texts.
Promotional supply chain	Supply chain for items that are being heavily promoted such as end-aisle tasting displays in wholesale clubs. The major challenge is that all items related to the promotion (product, utensils, cooking materials, display materials, and demonstrator must be assembled with the cart and delivered to the store to meet the weekend display schedule.
Bulk material supply chain	Supply chains designed to move bulk products such as grains, metals, and chemicals. In many cases, these materials are relatively low value so all movement and handling in the supply chain must take advantage of significant economies of scale and often specialized vehicles.
Talent supply chain	Applying supply chain principles to talent management where individual talent represents the products that are moved through the supply chain with the value-added process being training and education.
Business to consumer (B2C) supply chain	Business to consumer supply chains represents the increasing volume of product that is sold online from manufacturers or distributors directly to consumers.
Recycling supply chain	Supply chains responsible to handle returns for recycling of products, components, reprocessing, and packaging.
Resource supply chain	Resource supply chains are designed to provide facility resources for information-based supply chains such as server farms for cloud or social media applications. This includes the purchasing and sequencing of land, regulatory approvals, utilities, and equipment to provide the technology services.
Construction supply chain	Construction supply chains provide and sequence the equipment, building supplies, and specialized labor for construction.
Recovery supply chain	Recovery supply chains are employed to recover material that has reached its useful life in the field. A recovery supply chain is useful following military, construction, mining, or drilling operations.
Humanitarian supply chain	Humanitarian supply chains provide postevent support for disaster recovery. This includes bring in equipment for recovery, food and medical care items, and commodities to support reconstruction.
Global supply chain	Global supply chains source and deliver from multiple regions around the world. While most supply chains include global aspects, it is important to consider specific global characteristics such as demand variation, distance, and documentation.
Durables supply chain	Durables supply chains are designed to facilitate the handling and delivery of heavy equipment such as agricultural, construction, or military equipment. The major differentiator for durables supply chains is specialized transportation equipment due to infrastructure restrictions.
Agricultural commodity supply chain	Agricultural commodity supply chains move agricultural product from the farm to the commodity elevator or the processing plant. In most cases, the challenge is to move this bulk product economically so the farmer can still make money even when the buyer sets the price. In other words, if the farmer is too far from the buyer, there will have no market for his products.
Innovative supply chain	An innovative supply chain is one that must rapidly introduce new product to the market. This is typically a responsive supply chain that is defined to bring new product variations to market or to have souvenirs such as for movies, athletic events, or customized product introductions available when the event is taking place (e.g., concerts, movies, openings, etc.).
Military supply chain	Military supply chains are designed to support military operations. Specialized requirements include the ability to provide supply chains for a range of products (food, medical, and equipment) in demanding environments (dessert, jungle, and supporting combat operations) under varying requirements for responsiveness.
Clinical trials supply chain	Clinical trials supply chains are designed to support the very precise demands for completing pharmaceutical clinical trials. Clinical trials are very demanding due to the need for precise controls of dosages, ingredient combinations, and drug combinations.

Source: Bowersox et al. (2019, p. 7).

Product Supply Chain

The standard product supply chain demonstrates the flow of product from the raw material provider to the consumer. This is typical of most organizations that manufacture product on a relatively consistent basis without significant seasonal spikes. The classical example is “every day low price.” The service focus is to meet the availability and access requirements desired by the consumer. In general, the financial focus is to grow through increased revenue or market scope. In terms of procurement, manufacturing, and logistics, the primary focus is to take advantage of economies of scale in overall supply chain operations. Typically, the firm’s performance will likely be enhanced through increased volume. The product supply chain is typical for fast moving food and general merchandise.

Promotional Supply Chain

The promotional supply chain demonstrates the flow of products that are pushed at different times during the year through price discounts or advertising. In this case, the product flow, from raw material acquisition through delivery to customer, is seasonal or very spiky. The implication of a promotional supply chain is that it must be flexible because it needs to accommodate significant swings in volume. The primary focus of a promotional supply chain is to meet the sales and revenue objectives of the promotion. In general, a promotion is designed to increase market share by introducing a product to new consumers or to get existing consumers to stock up so they will not purchase competitive products. In the case of purchasing, manufacturing, and logistics, the focus is to be able to acquire, produce, and deliver the product in an environment where there are significant swings in material, production, and

Table 3A Supply chain applications

	<i>Product supply chain</i>	<i>Promotional supply chain</i>	<i>Bulk material supply chain</i>	<i>Talent supply chain</i>
Service focus	Meeting customer service objectives	Meet sales objectives of promotion	Provide service with maximum asset utilization	Provided appropriately trained and experienced personnel
Financial focus	Focus on volume	Increased market share	Take advantage of raw material capacity	Maximize utilization of trained and experienced expertise
Procurement focus	High volume reliable suppliers	Obtain raw materials in volumes to meet specific promotion	High volume reliable suppliers	Shift from fixed to variable cost with high degree of specialization
Manufacturing focus	High volume manufacturing	Customized agile production	Minimize changeovers	Increased application of specialized talent under time constraints
Logistics focus	Economies of scale	Obtain transport capacity to meet volume requirements	Require bulk transport (water, rail, and truck)	Finding talent that can be employed in desired location (language and permits)
Opportunities for performance enhancement	Increased revenue	Increased market share	Move from commodity to specialized	Maximum utilization of specialty talent
Product types	Food and general merchandise	Food and general merchandise. Event related promotions	Petroleum, chemicals, and construction commodities	Research, information technology, medical, and supply chain talent

Table 3B Supply chain applications

	<i>B2C supply chain</i>	<i>Recycling supply chain</i>	<i>Resource supply chain</i>	<i>Construction supply chain</i>
Service focus	Responsive home delivery for significant volume variations	Maximize quantity recycled	Meet solution requirements of consumers	Meet building availability requirements
Financial focus	Increased access to customer base	Minimize cost of recovery	Minimize assets required	Minimize cost of materials and labor
Procurement focus	Increased range of products	Gain access to volume of recyclables	Schedule asset arrival in a timely manner	Trade-off acquisition cost versus storage requirements
Manufacturing focus	Increased customization and breadth of line	Effective process of recyclables to create reusable materials	Parallel processing	Parallel processing and site logistics
Logistics focus	Responsive delivery and low-cost returns processing	Minimize transport cost in noneconomy of scale environments	Arrival sequencing and inventory minimization	Arrival sequencing, inventory minimization, and licensing
Opportunities for performance enhancement	Increased volume over last mile. Processing time and capacity considerations	Easy to disassemble and sort	Take advantage of opportunities for parallel processing	Take advantage of opportunities for parallel processing
Product types	Increasing range of products	Metals, electronics, and durables	Project management and financial documents	Facilities construction

Table 3C Supply chain applications

	<i>Recovery supply chain</i>	<i>Humanitarian supply chain</i>	<i>Global material supply chain</i>	<i>Durables supply chain</i>
Service focus	Return goods for reuse	Search, recovery, and meet critical necessities	Meet unique regional requirements	Meet customer delivery requirements
Financial focus	Reuse assets and minimize cost	While cost is important, perception may be more important	Low total cost including duties and taxes	While it is important, cost is often secondary with more emphasis on solution provided
Procurement focus	Organize items to be recovered	Preposition requirements	Meeting regional and regulatory requirements	Standardize components
Manufacturing focus	Remarketing or remanufacturing	Adapt products to environment	Standardize as much as possible	Extensive variation and complexity
Logistics focus	Adapting to low scale economies	Infrastructure is often limiting	Work within local infrastructure and service providers	Specialized equipment
Opportunities for performance enhancement	Consolidation	Product prepositioning	Apply appropriate global strategy	Collaboration with other firms serving similar customers
Product types	Military, construction, and electronics	Floods, earthquakes, hurricanes, and political uprisings	Durables and furniture	Farm and construction equipment

Table 3D Supply chain applications

	<i>Agricultural commodity supply chain</i>	<i>Innovative supply chain</i>	<i>Military supply chain</i>	<i>Talent supply chain</i>
Service focus	Meet production requirements of production facilities	Make sure product is available when required in desired markets	Meet requirements of war fighters	Very precise delivery to trial patients
Financial focus	Maximum use of commodity priced items	Cost is definitely secondary	Increasingly important	Cost is important but not nearly as important as delivery precision
Procurement focus	Commodity items	Work with suppliers to develop innovative components	Meet Defense Acquisition Regulations System (DFARS)	Meet quality and certification requirements
Manufacturing focus	Maximize productivity for commodity conversion	Pilot plant capability	Make versus buy	Apply quality pilot production capabilities
Logistics focus	Transport cost results in reduced price received by farmer	Precise delivery requirements with security	Non-standard and hazardous goods transportation	Precise delivery is critical
Opportunities for performance enhancement	Reduce transport distance while maximizing scale of production	Specialty innovative supply chain	Apply lessons from commercial sector	Collaboration with competitors involved in clinical trials
Product types	Agricultural commodities	New products	Military consumables and capital goods	Pharmaceuticals involved in trials

delivery capacity requirements. This can typically only be done effectively when capacity requirements can be smoothed out across multiple products. The typical result of promotional supply chains has increased market share and is often used to increase the volume of food and general merchandise products and for special event items.

Bulk Material Supply Chain

The bulk materials supply chain demonstrates the flow of commodity products, such as petroleum, chemicals, and coal, typically from a mine or well to the processing facility. In this case, large volumes are being moved requiring bulk transportation equipment such as pipelines, ships, barges, or rail cars. Due to the product bulk, a significant percentage of the product value is related to transportation, so transport mode is a critical consideration. In terms of purchasing, manufacturing and logistics, product volume and changeovers need to be maximized and minimized, respectively. While these bulk materials/commodities still require economies of scale in processing and transportation, their customers are increasingly looking smaller and more responsive shipments, which is driving them to smaller shipment sizes and more less than truckload transport modes. While bulk materials supply chains are very important, their unique requirements for bulk movement are sometimes not well understood.

Talent Supply Chain

The talent supply chain refers to the application of supply chain principles to talent management and development. In effect, the management and operational talent becomes the product that is moved through talent development stages, such as training and job experiences. Increasingly, human resource management firms are using supply chain principles to manage the individuals, so that their specialized skills can be enhanced and applied. As an example, with the increasing shortage of talent, it is critical that the individuals be provided with the experiences that they need. In an environment where this is a shortage of supply chain managerial and operational talent, the effective development and use of existing talent is critical. Applying supply chain principles to talent management is particularly relevant when there is a talent resource shortage. Supply chain principles are specifically being used to manage specialized talent in the fields of research, information technology, medicine, and supply chain management. The other fields that these principles could be applied in include disciplines, where there is a high degree of specialization, significant benefits of experience, and a shortage of talent. [Kelly OCG \(2017\)](#), a division of Kelly Services, has used supply chain management principles to manage the talent pipeline for many of its clients.

Business to Consumer (B2C) Supply Chain

The B2C supply chain refers to the application of supply chain principles for firms delivering directly to homes. The obvious application is Amazon and similar firms who offer a wide range of items within a delivery time of a few days. Consumers are increasingly requesting this type of agile responsiveness to deliver everything from books to specialized electronics to general merchandise directly to their homes. While some of the items are reasonably expensive and can justify premium transportation, most of them are not, so it is difficult to provide quick delivery without having distribution centers in close proximity. The result is that B2C supply chains typically require many distribution centers with the capability to customize the product. While such last-mile delivery is very expensive, firms are using their last mile delivery capabilities to offer a wider range of products to their consumers,

which will spread the delivery cost over a wider range of products. The result is a supply chain that is agile and responsive focused rather than economies of scale focused.

Recycling Supply Chain

The recycling supply chain refers to a firm's ability to collect and process recyclable material. The major focus is to recover material for remarketing, remanufacturing, or recycling. Remarketing, remanufacturing, and recycling are becoming increasingly important to satisfy a firm's need to meet environmental and sustainability requirements for consumers and regulators. The remarketing application includes the recovery of products such as electronics or office furniture that can be renewed and resold. The remanufacturing application includes the recovery and reprocessing of products such as electronics, office furniture, mechanical components, and packaging. In these cases, the product may be disassembled and then rebuilt using new or renewed components for resale to consumers. The recycling application collects used products and packaging to reprocess them into materials such as metal or plastic. These recycled materials can then be reused to manufacture new product. A recycling supply chain is particularly relevant for materials such as plastics and aluminum. The supply chain challenge for recycling is effectively collecting used materials from consumers and transporting it to a processing facility while achieving some level of scale economies. There are many articles that discuss the nuances of recycling (Shahparvari et al., 2018).

Resource Supply Chain

The resource supply chain focuses on processes that handle products that would not traditionally be considered supply chain products. Examples might include financial or mortgage documents. While these documents may be processed electronically, there is still often a need to move the documents physically through a process with capacity constraints. These constraints might be in the form of equipment or individuals with specialized expertise. As such, this process is not unlike the processing of products through a job shop production facility. The challenge is to design the document flow to take advantage of batch or parallel processing opportunities to optimize the use of constraining resources.

Construction Supply Chain

The construction supply chain focuses on the building of facilities such as houses, commercial buildings, and manufacturing complexes. While construction has not typically been viewed as a supply chain application, a review of the process makes it obvious that any construction initiative requires process and resource management. The resources include not only the construction materials, but also the approvals and regulatory authorizations. The processes include the construction activities and the fact that many of them must be completed in sequence. The combination of constraints regarding site logistics and construction expertise emphasize the importance of effective process and resource management.

Recovery Supply Chain

The recovery supply chain focuses on returning goods for reuse. The typical application of these principles occurs when products are used for an event (construction or military initiative) with the desire to return for reuse. Examples might include construction equipment like earth movers, cranes, material for oil wells, and military equipment, like personnel transporters or tanks. At the completion of such an initiative, there may be a desire to return these goods for reuse or to provide them to another organization. While the goods can sometimes be remarketed or remanufactured, there are often restrictions on how and where the returned goods need to be handled. For example, there are often restrictions where military equipment can be returned and how it must be handled due to the environment it was used in. While the focus should be to take advantage of consolidation opportunities when possible, the challenge is that there are many complexities that must be considered.

Humanitarian Supply Chain

The humanitarian supply chain focuses on products and relief supplies to those who do not have regular access to items like medical supplies or life support materials. The access may be on a regular or emergency basis, such as when medical supplies are not available due to a regular or emergency lack of infrastructure. As an example, it is often the case that it is not possible to get medical, pharmaceutical, water, or food supplies to populations outside of the developed or available infrastructure. The challenge in these situations is to either preposition commodities when possible or utilize alternative modes to overcome the modal limitations. The recent hurricanes and floods have demonstrated the need for these capabilities and the role that they can play in enhancing firm perception.

Global Supply Chain

The global supply chain reflects the need for recognizing the challenges inherent in global operations. While most larger firms operate and understand the implications of operating global supply chains, most smaller firms also operate globally whether they

know it or not. They often have international suppliers and customers that are not visible to them. The result is that such firms do not understand the supply chain differences characteristic of global operations. These differences include distances and time delays, documentation and compliance requirements, technical differences, and population diversity. The overall result is that the global supply chain is substantially more complex. The challenge, then, is that all firms must increasingly become more proactive in understanding the implications and requirements to support global supply chain operations.

Durables Supply Chain

The durables supply chain refers to the challenges related to moving large manufacturing, construction, marine, or agricultural equipment. These durable delivery supply chains are generally unique because they require specialized equipment and, often, infrastructure monitoring to move the finished product. Infrastructure monitoring refers to the need to have tracking vehicles to protect traffic and infrastructure, such as bridges that can accommodate large durable equipment. The products are obviously highly specialized and complex. While cost is a consideration, the primary focus is to complete delivery of the product. Unlike most supply chain applications, durables supply chains are different in that each of them has unique characteristics.

Agricultural Commodity Supply Chain

The agricultural commodity supply chain focuses on the movement of bulk agricultural commodities such as lumber, grains, or soybeans. While agricultural commodity supply chains are similar to bulk material supply chains in that they are typically moved in large loads, there are significant differences in that agricultural commodities must be collected from farms across the geography. The implication is that these bulk shipments are often made on secondary roads, which results in seasonal congestion and road damage. Another issue related to commodities in general, and agricultural commodities specifically, is that the effective selling price of the product to the processor must be reduced by the transportation charges. This limits the distance that the farms can be from the processors and still be competitive. The challenge in agricultural supply chains is to understand the trade-offs related to the location for the demand of processed commodities, the production economies for the processing facility and the source for enough commodity volume to support production. Topics of specific interest for agricultural commodities include food security and flow optimization. Due to the complexity of the food supply chain, there has been an increasing use of optimization and simulation tools applied to identify an integrated solution (Ge et al., 2015).

Innovative Supply Chains

The innovative supply chain focuses on the processes to get relatively low quantities for new product introductions to trial markets. While some firms believe that their supply chain for innovative products is a subset of the supply chain for their standard products, the requirements are actually quite different. As discussed previously, the typical product supply chain has the primary focus on economies of scale, whereas the innovative supply chain should be primarily focused on agility and responsiveness. The innovative supply chain must be able to purchase the raw material in enough quantity to facilitate production of the new product, produce it in what is typically a pilot plant setting, deliver the product in agile fashion to test markets, and provide confidentiality through the entire process to protect competitive advantage.

Military Supply Chain

The military supply chain focuses on one of the oldest applications in the world. Two examples illustrate the role of logistics and supply chain professionals in the military. The first describes the role of logistician in the military: "Logisticians are a sad and embittered race of men who are very much in demand in war, and who sink resentfully into obscurity in peace. They deal only in facts, but must work for men who merchant in theories. They emerge during war because war is very much a fact. They disappear in peace because peace is mostly theory. The people who merchant in theories, and who employ logisticians in war and ignore them in peace are generals. Generals are a happily blessed race, who radiate confidence and power. They feed only on ambrosia and drink only nectar. In peace, they stride confidently and can invade a work simply sweeping their hands grandly over a map, pointing their fingers decisively up terrain corridors, and blocking defiles and obstacles with the sides of their hands. In war, they must stride more slowly because each general has a logistician riding on his back. He knows that, at any moment, the logistician may lean forward and whisper: 'No, you can't do that.' Generals fear logisticians in war and try to forget them in peace. Romping along beside generals are strategists and tacticians. Logisticians despise strategists and tacticians. Strategists and tacticians do not know about logisticians until they grow up to be generals – which they usually do. Sometimes a logistician becomes a general. If he does, he must associate with generals whom he hates; he has a retinue of strategists and tacticians whom he despises; and on his back, is a logistician whom he fears. This is why logisticians who become generals always have ulcers and cannot eat their ambrosia" (Anon, cited in Bowersox et al., 2002, p. 596).

President Dwight D. Eisenhower put it more succinctly when he wrote: "You will not find it difficult to prove that battles, campaigns and even wars have been won or lost primarily because of logistics" (Bowersox et al., 2019, p. 383).

One of the major challenges related to military supply chains is the shift to increased consideration of cost. While kings and generals had traditionally thought that "cost was no object" when providing resources for the military (because the implications of

losing a war were so severe), today's civilian leader knows that military spending does not generate a substantial number of votes. As such, today's military supply chain is becoming more like a commercial supply chain, where there are strong cost considerations. While logistics has historically been the focus of the military, there has been increasing emphasis on a more holistic supply chain consideration, where the military objective is to minimize total supply chain cost (Haraburda, 2016).

Clinical Trials Supply Chain

The clinical trials supply chain focuses on the processes and flows to support the regulatory trials of prescription pharmaceuticals. These extensive trials must be conducted with a controlled sample of patients prior to making the drugs available to the general population. The supply chain challenge for clinical trials is that the sample products (both those with the active ingredients to be tested and those that are placebos (no active ingredients) must be manufactured and distributed in a very controlled manner to the test patients. The unique characteristic of clinical trials is that the different types of products must be tracked and recorded very accurately to ensure that the result statistics are accurate. Any accuracy or delivery failures could invalidate the trials and require the pharmaceutical firm to begin the trials over again. There are very few other applications, which require such a high level of supply chain precision and consistency.

Conclusion

Opportunities for the application of supply chain principles are increasing dramatically. However, many supply chain professionals tend to limit their scope when considering where their expertise and skill can be applied. Based on years of interactions with students and recruiters, it became clear that there was a need to identify and describe the broader range of applications so that nonsupply chain professionals can identify the potential benefits and supply chain professionals can identify the opportunities.

References

- Bowersox, D.J., Closs, D.J., Bixby Cooper, M., 2002. *Supply Chain Logistics Management*, first ed. McGraw Hill, New York.
- Bowersox, D.J., Closs, D.J., Bixby Cooper, M., Bowersox, J.C., 2019. *Supply Chain Logistics Management*, fifth ed. McGraw-Hill, New York.
- Ge, H., Gray, R., Nolan, J., 2015. Agricultural supply chain optimization and complexity: a comparison of analytic vs simulated solutions and policies. *Int. J. Prod. Econ.* 159, 208–220.
- Haraburda, S.S., 2016. Transforming military support processes from logistics to supply chain management. *Army Sustainment* 48 (2), 12–15.
- Kelly OCG, 2017. Talent Supply Chain Management. Available from: <https://www.kellyocg.com/expertise/talent-supply-chain-management>.
- Shahparvari, S., Chhetri, P., Chan, C., Asefi, H., 2018. Modular recycling supply chain under uncertainty: a robust optimization approach. *Int. J. Adv. Manuf. Technol.* 96 (1–4), 915–934.

Logistics and Supply Chain Management Performance Measures

David B. Grant*, **Sarah Shaw†**, *Supply Chain Management and Social Responsibility Group, Department of Marketing, Hanken School of Economics, Helsinki, Finland; †Logistics and Management Systems Subject Group, Hull University Business School, University of Hull, Hull, Yorkshire, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	16
Performance Measurement and Measures	16
Performance Measurement Frameworks	17
The Balanced Scorecard (BSC)	18
The Deming Plan-Do-Study-Act (PDSA) Cycle	18
Special Cases	18
Sustainable or Environmental Performance Measures	18
Sustainable or Environmental Performance Measurement Frameworks	19
Sustainability or Environmental Management Schemes for Logistics and SCM	20
Humanitarian Logistics and SCM	21
Conclusion	22
Biographies	23
References	23
Further Reading	23

Introduction

The phrase that “you can’t manage what you can’t measure” is attributed to various sources and has become an important business aphorism. There are two additional, related quotes that highlight today’s preoccupation with performance measurement and key performance indicators (KPIs). The management guru Peter Drucker argued that “if you can’t measure it, you can’t improve it,” while the statistician and quality control expert W. Edwards Deming suggested that “in God we trust, all others bring data.”

These three concepts are very applicable in the logistics and supply chain management (SCM) domains where organizations are concerned with what have been called the “7 rights:” the right product in the right place at the right time to the right customer in the right amount or quantity in the right condition and at the right cost. These “rights” are predicated on being able to measure what goes into a logistics and SCM process and, more importantly, tracking and tracing where goods, products, or services are in that process.

This chapter provides an overview of logistics and SCM performance measurement in three sections. First, performance measurement and measures or metrics in general are discussed regarding current practice and provide suggestions for managers on selecting appropriate measures for the organizations. Then, different types of performance measurement frameworks are introduced regarding management of performance measurement and measures. Finally, two special contexts, sustainable or environmental and humanitarian logistics and SCM, are presented to consider nuances on usual organizational performance measurement and include the use of two frameworks, the Balanced Scorecard (BSC) and the Plan-Do-Study-Act (PDSA) cycle, to illustrate these nuances.

Performance Measurement and Measures

Performance measures (PMs) are a set of metrics or measures used to quantify the efficiency and/or effectiveness of an action. PMs enable an organization’s mission or strategy to be translated into reality by generating “actionable” knowledge that is essential for managing and navigating organizations through turbulent and competitive global markets. They allow organizations to track progress against their strategy, identify areas of improvement and act as a good benchmark against competitors or industry leaders. The information provided by PMs allows managers to make the right decisions at the right times.

One of the most prevalent issues associated with performance measurement is having too many metrics. Some organizations are using hundreds of metrics, which are often not aligned to the organization’s strategy. One review of logistics and supply chain performance measures by identified almost 90 metrics, of which 40% were financial and 60% were functional (Shaw, 2013). Such metric proliferation can lead to confusion, often results in “paralysis by analysis” and presents difficulties in conducting benchmarking exercises. Thus, organizations need to determine a meaningful but parsimonious set of PMs, so they can be effectively implemented and monitored.

Table 1 Common logistics and SCM performance and service measures

Accurate invoices
Accurate orders
After-sales support
Appropriate order cycle time
Availability of products
Complete orders
Consistent order cycle time
Consistent product quality
Customized services
Delivery time
Helpful customer service representatives
On-time delivery
Products arrive undamaged
Products arrive to specification
Supplier commitment
Supplier integrity
Supplier trust

Source: Compiled from [Grant, 2012](#)

Table 2 Eight criteria for judging performance measures

Criteria	Description
Validity	PM accurately captures activities being measured and controls for any external factor
Robustness	PM can be interpreted similarly by users, is comparable across time, location and organizations, and is repeatable
Usefulness	PM is readily understandable by managers and provides a guide for action
Integration	PM includes all relevant process aspects and promotes coordination across functions and organizational units
Economy	PM benefits are greater than the costs of PM data collection, analysis and reporting
Compatibility	PM is compatible with existing information, material and cash flow systems in the organization
Level of detail	PM provides enough degree of granularity or aggregation for the user
Behavioral soundness	PM minimizes incentives for counter-productive acts or game-playing and is presented in a useful form

Source: Compiled from [Caplice and Sheffi, 1994](#)

Traditionally, logistics and supply chain PMs have been quantitative and orientated around measuring cost, time, and accuracy. For example, order lead-times, delivery performance, customer query time, and total cash flow time within a framework of strategic, tactical, and operational performance levels. [Table 1](#) lists, in alphabetical order, almost 20 PMs related to logistics and SCM functions and service previously documented by [Grant \(2012\)](#). While some may appear to be the same, there are nuances as to their meanings for individual or focal organizations.

For example, appropriate order cycle (or lead time) may be what a supplier can offer in terms of delivery dates, that is, 3 weeks or 3 days. The product type, whether it is perishable, and final demand from the focal organization's customers will be important factors in determining the appropriate order cycle. Consistent order cycle time is the variance around the appropriate order cycle time, for example, 3 weeks $\pm$ 4 days. Focal organizations are also concerned with this element, as too great a variance can affect their own scheduling and planning.

Thus, a key challenge for organizations is to determine and select the most appropriate and effective PMs. Managers should continually review and evaluate these PMs to make sense of the large number of metrics available and to ensure they reflect the ever-evolving business environment. [Table 2](#) presents eight criteria from [Caplice and Sheffi \(1994\)](#) for judging the quality of PMs or metrics, as they relate to an organization's strategic objectives and strategies and considering its external environment. Managers can use any or all of these criteria to assess any PMs they would consider choosing. However, once the appropriate PMs are chosen, the next step is to embed them into some form of performance measurement framework or system. The further section discusses this topic.

Performance Measurement Frameworks

Over the last 40 years, researchers have proposed various performance frameworks to manage firm performance, including a performance measurement matrix, performance pyramid, performance prism, the BSC, Deming's PDCA cycle, and the supply chain operations reference (SCOR) model in a logistics and SCM context. This has also led to confusion for managers over which framework may work best for their organization and how it should be developed and deployed. Due to space limitations in this chapter we focus on two that are demonstrated in the following section.

The Balanced Scorecard (BSC)

The BSC is a strategic planning and management system that organizations can use to communicate what they are trying to accomplish; align the day-to-day work that everyone is doing with strategy; prioritize projects, products and services, and measure and monitor progress toward strategic targets. The BSC suggests the organization takes a view from four perspectives and develops objectives, measures (KPIs), target, and initiatives (actions) relative to each of these points of view, which are:

- *Financial*, often renamed stewardship or another more appropriate name in the public sectors, this perspective views organizational financial performance and the use of financial resources;
- *Customer/Stakeholder*, this perspective views organizational performance from the point of view of the customer or other key stakeholders that the organization is designed to serve;
- *Internal process*, this perspective views organizational performance through the lenses of quality and efficiency related to its products, services, or other key business processes; and
- *Organizational capacity (formerly called Learning and Growth)*, this perspective views organizational performance through the lenses of human capital, infrastructure, technology, culture, and other capacities that are key to breakthrough performance.

The BSC has been adapted for use by different organizations in different contexts, including logistics and SCM. Although pioneering and popular, the BSC was originally devised by Kaplan and Norton (1992) and thus it is over a quarter of a century old. Criticisms of the BSC and its applications are that it excludes the ability to manage sustainable or environmental performance within an overall business strategy, as well as the people and suppliers who are key stakeholders in the sustainable or environmental management process. The BSC is also a static tool and does not have a dynamic “cause and effect” evaluation loop process and, thus, has no ability to guide businesses through change, which is of central importance to PM. Further, there is a lack of evidence to suggest that the application of BSC improves performance. Finally, extensions to the BSC to incorporate sustainable or environmental PMs have been advocated but not extensively studied, and there are some arguments suggesting that adding sustainable or environmental elements to the BSC might overcomplicate it.

The Deming Plan-Do-Study-Act (PDSA) Cycle

The PDSA cycle is a systematic process for gaining valuable learning and knowledge for the continual improvement of a product, process, or service. Also known as the Deming cycle, or Deming Wheel (Deming, 1950), this integrated learning–improvement model was first introduced to Deming by his mentor, Walter Shewhart of Bell Laboratories in New York (Shewhart, 1939).

The cycle begins with the Plan step. This involves identifying a goal or purpose, formulating a theory, defining success metrics and putting a plan into action. These activities are followed by the Do step, in which the plan components are implemented, such as making a product. Third is the Study step where outcomes are monitored to test the validity of the plan for signs of progress and success, or problems and areas for improvement. The Act step closes the cycle, integrating the learning generated by the entire process, which can be used to adjust the goal, change methods, reformulate a theory altogether, or broaden the learning–improvement cycle from a small-scale experiment to a larger implementation Plan. These four steps can be repeated over and over as part of a cycle of continual learning and improvement.

Another variant is the Plan-Do-Check-Act (PDCA) cycle. Although PDSA and PDCA are often confused and used interchangeably, there are important differences. The main difference is that PDSA attempts to study the consequences of the applied changes more closely and in more detail compared to PDCA. The meaning of the Check step often entails running some basic tests to ensure that the new state of the system corresponds to some baseline measurements. However, that may be insufficient to understand the entire situation and ensure that changes will lead to long-run improvement. Deming emphasized the PDSA cycle, not the PDCA cycle, as he found the focus on the Check step is more about implementation of a change, with success or failure followed by needed corrections to the Plan in the event of failure. His focus was on predicting results of an improvement effort, studying these results, and comparing them to possibly revise theory. He stressed that the need to develop new knowledge from learning is always guided by a theory.

Some authors note that PMs tend to be retrospective, inward looking, not aligned to business strategy, and focused too much on cost as a primary measure. As a result, organizations need to continually review and renew their PMs to ensure they meet the organization’s needs and changing business environment. Further, the development and evolution of PM research has reached a critical point and is entering a new phase or direction, categorized by context, theme, and challenge. The following section considers two special cases of sustainable and humanitarian logistics and SCM.

Special Cases

Sustainable or Environmental Performance Measures

Organizations are facing increased pressure from the government, customers, and competition on their environmental and social performance. Since the beginning of this Millennium, organizations have been looking to quantify and mitigate their impact on

Table 3 Sustainable logistics and SCM performance measures

Electricity consumption measures
Driver behavior for telematics
Carbon dioxide emissions of an activity (route/product)
Carbon dioxide emissions per item/case/pallet delivered
Overall company carbon footprint measures
Vehicle mileage measures
Packaging consumption/reduction measures
Fuel consumption measures
Number of pallet movements or touches per delivery
Utilization/consolidation measures (warehouse/back haul/pallet occupancy)
Fuel consumed per item/case/pallet delivered
Vehicle running costs
Waste recycling measures/reduction/percentage to landfill
Warehouse efficiency measures
Water consumption measures
Gas consumption measures
Overall supply chain carbon footprint measures
Vehicle fill/utilization measures to reduce empty running
Energy used per item/case/pallet delivered
Greenhouse gas emissions (carbon dioxide, nitrous oxide, sulfur, methane, etc.)
Employee environmental or sustainability training
Container size/fill/movements (number of TEUs)
Employee travel (e.g., company car mileage, car share, and cycle to work schemes)
Resource efficiency (raw materials, asset utilization)

Source: Compiled from Shaw, 2013

natural environment ecosystems that contain, inter alia, air, water, and land without which these ecosystems would not exist. Society and humans are both a participant in, and exploiter of, the natural environment and society's use of the natural environment's resources to produce economic goods, products, and services can upset a fragile balance.

Logistics and SCM present a major challenge to the sustainability of these ecosystems in the way that goods and products are transported, handled, stored, manufactured, and supplied throughout the world. For example, fresh air and excess packaging are shipped, with empty return loads, products sit idle and become obsolete in warehouses, and unnecessary products move backward and forward. Supply chain networks are not robust, and innovation is thwarted as every organization works to achieve targets within their own corporate silos, all of which has an impact on natural environment ecosystems. It is important, therefore, for organizations to be able to measure this environmental impact and reduce, mitigate or eliminate it.

As discussed previously, logistics and SCM performance measurement is difficult due to the numerous tiers and echelons found within supply chains and networks and is more difficult in a sustainable or environmental context, due to its increasing importance. A major barrier to the adoption of sustainable PMs is financial, due to the large upfront investment required that results in low adoption rates. Organizations are also confused over what to measure and how to do so. There is a challenge in convincing organizations to invest in sustainability and determining if it is worthwhile. There is also a challenge in assisting organizations to find the most appropriate sustainability measures, especially in a logistics and SCM context. Shaw (2013) developed a list of over 20 logistics and SCM sustainable PMs (Table 3), through research with members of the UK's Chartered Institute of Logistics & Transport (CILT).

Given the nature of the research sample, many of these PMs are transport- and storage-related and, thus, the list in Table 3 is not exhaustive. However, it provides organizations with possible sustainability or environmental PMs in these functions and includes aggregate (e.g., carbon dioxide emissions of an activity) and granular (e.g., carbon dioxide emissions per item/case/pallet delivered) measures. An organization can examine their own logistics and SCM activities to develop a similar list of possible measures that can be reduced to those that directly affect the organization's performance. The next step is to consider what performance measurement framework would best suit that parsimonious list.

Sustainable or Environmental Performance Measurement Frameworks

Some authors have identified that the types of sustainable PMs used are reflected by an organization's evolutionary stage in the sustainable management process. One of the key issues related to the number of current sustainable metrics, which range from air emissions through to water usage. Accordingly, there is a requirement to conceptualize a sustainable and environmental performance measurement system and its requirements, in order to address both this complexity and the volume of metrics and to provide a way forward for the organization.

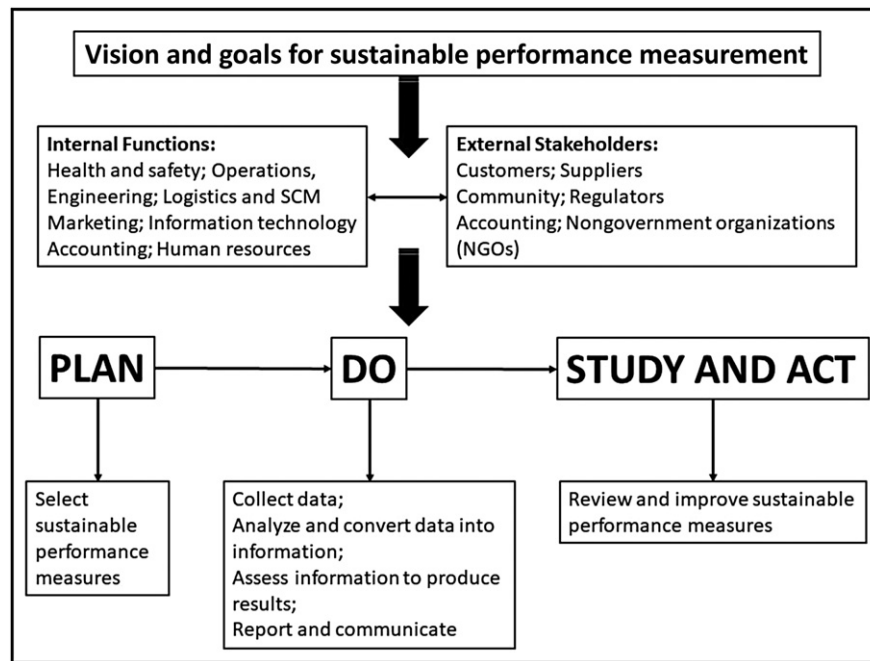


Figure 1 A Deming PDSA cycle framework for sustainable performance measurement. *Source: Adapted from Hervani et al., 2005.*

Fig. 1 presents an adapted PDSA cycle framework proposed by Hervani et al. (2005) for determining PMs based on the ISO 14001 family of environmental performance management systems to help design, evaluate, and adapt sustainable PMs for logistics and supply chain applications.

Interorganizational performance management systems play an important role in this framework, since organizations are required to explicitly consider the environment in their strategic and operational planning and execution, particularly across their logistics and supply chain activities. This framework can be managed through actions of vertical integration of governance and horizontal integration of internal and external stakeholders.

These actions are required to ensure the structure and functioning of a natural system is protected and maintained, while at the same time providing the goods, products, services, and benefits required by society. Such frameworks have yet to fully exist and operate across many organizations. However, their development and introduction is inevitable as further integration and pressures cause organizations to seriously consider them for their long-term well-being. One way to enhance or enable a framework can be to align it with an environmental or sustainability management certification or accreditation scheme.

Sustainability or Environmental Management Schemes for Logistics and SCM

Sustainable PMs in the logistics and SCM domain have largely focused on greenhouse gas emissions, due to their importance in the fight against climate change. For example, the 2015 COP 21 Paris agreement legally bound industrialized nations to limit temperature rises to below 2°C. However, organizations are also starting to more seriously consider broader sustainable management issues, not only from a mitigation and legislative perspective but also from an adaptation perspective. Consequently, several sustainability and environmental certification or accreditation schemes have been created and developed to assist organizations in establishing and managing their sustainable performance initiatives.

Such schemes range in scope from generic business certifications that assess environmental performance, such as the International Standards Organization's ISO 14000, which specifies a process for controlling and improving an organization's environmental performance, and the European Commission's Eco Audit and Management Scheme or EMAS, developed by the European Commission for organizations to evaluate, report and improve their environmental performance. There are also more industry-specific certifications such as the UK's Building Research Establishment Environmental Assessment Method (BREEAM) which measures sustainability in the building and construction sector, or the Electronic Product Environmental Assessment Tool (EPEAT), an environmental product ratings service that makes it easy to select high-performance electronics that support organizations' information technology (IT) and sustainability goals.

However, this diversity of certifications presents additional challenges for organizations when selecting the most appropriate certification due to the differing scopes and levels each certification may have; such as country, industrial sector, type of organization, type of supply chain, and/or global presence. Most sustainability certifications and standards currently available are voluntary, except for country-specific government acts such as the UK Climate Change Act. However, a significant number of

Table 4 Ten criteria for the sustainable or environmental management of natural ecosystems

Socially desirable/tolerable	Sustainable management PMs are required or understood by society as being required.
Ecologically sustainable	Sustainable management PMs will ensure the safeguarding of fundamental ecosystem features, functions and final ecosystem services.
Economically viable	A cost-benefit assessment of the sustainable management PM indicates economic or financial viability and sustainability.
Technologically feasible	Methods, techniques and equipment for ecosystem and society/infrastructure protection are available.
Legally permissible	Regional, national or international agreements and/or statutes are in place to enable and/or force sustainable management PMs.
Administratively achievable	Statutory bodies such as governmental departments and environmental protection/conservation bodies are in place and functioning to enable success.
Politically expedient	Management approaches and philosophies are consistent with prevailing political climates and have the support of political leaders.
Ethically defensible	Costs to act are determinable and calculable for current and future generations.
Culturally inclusive	Cultural considerations are included in required sustainable management PMs.
Effectively communicable	Communication is effective all stakeholders to achieve the vertical and horizontal integration encompassed in the foregoing criteria.

Source: Compiled from Grant and Elliott, 2018

voluntary schemes are audited to ensure organizations that use and advertise the standards comply with them, for example, the ISO series of standards.

Integrated sustainable management requires combining many aspects into a holistic natural ecosystem and problems caused by materials (e.g., pollution) or infrastructure added to or removed from the ecosystem (e.g., aggregates) require a risk assessment framework to be considered in conjunction with sustainable PMs. Consideration of interactive sustainable or environmental relationships, as discussed previously, and their inherent risks gives rise to assessing whether the strategy or strategic option fulfils various criteria related to sustainable or environmental management.

Table 4 presents ten criteria for sustainable or environmental management discussed by Grant and Elliott (2018) which should be largely fulfilled to ensure that the management of, and solutions for, an environmental problem will be sustainable, successful, and acceptable to society. As such, they should fall within what is realistically possible by encompassing socioeconomic and governance aspects noted in Fig. 1. Fulfilling them would potentially be seen by wider society as achieving sustainability and, in turn, would be more likely to be accepted, encouraged, and successful.

Humanitarian Logistics and SCM

Fast-onset humanitarian crises, such as natural disasters or conflicts, have their own impact regarding appropriate PMs to use and inherent trade-offs affecting usage. In a humanitarian crisis, a very important logistics and SCM consideration is to establish a transport network to provide aid tailored to fit that crisis. The network may be considered as a conduit running from a donor country's port of origin all the way to warehouses or distribution centers (DCs) in or near the affected areas. The network will often pass across an ocean arriving at a discharge port in the destination country or a country near to the destination, but may also pass enroute through warehouses or DCs on land, and be served by rail, road, air, and even barge.

In a crisis situation, humanitarian organizations look to ensure this transport network is up and running quickly, to establish it as robust as circumstances dictate so that supply interruptions do not occur and, while satisfying these objectives, attempt to keep network operating costs to a minimum. Hence, PMs for this situation may not be as determinable or efficient as they would be for a normal business organization, due to the immediacy of need and the life or death consequences for beneficiaries or recipients of aid. Humanitarian organizations will thus need to evaluate several shipping parameters or PMs to distinguish between transport carriers, which will vary by country and type of humanitarian crisis, as part of their carrier selection process. **Table 5** lists a number of these parameters in no particular order (Banomyong and Grant, 2016).

While many of these parameters appear "normal" in the context of business situations, they will be affected by the temporal and fulfilment requirements of ensuring beneficiaries achieve aid goods, products, and services, and hence cost may be less important. However, whatever is saved on transportation cost can be spent on vital aid supplies. The decision as to whether to charter a ship or to place cargo on a liner ship is one example of this trade-off. The World Food Program (WFP) has estimated that transporting aid on liners typically costs between two and twenty-one times as much as on ships chartered by the WFP. Consequently, to keep expensive liner shipments to the minimum, WFP has intensified their efforts to consolidate consignments into larger quantities, suitable for charter vessel carriage.

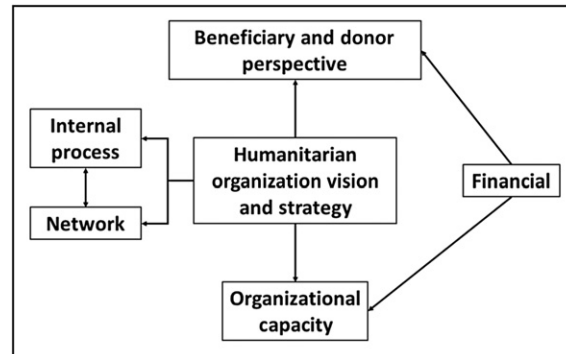
Another issue is what type of PM framework to use given various interrelationships between several humanitarian organizations and, of course, aid recipients. One attempt has been made to use a modified BSC by Widera and Hellingrath (2016) as shown in Fig. 2.

The BSC has been modified for specific humanitarian operations to reflect the daily tasks of various organizational units involved. Each operation is unique, has different challenges and set-ups, the importance of PMs or KPIs is not the same and several

Table 5 Transport carrier selection criteria for humanitarian logistics and SCM

Geographic coverage of the transport carrier, i.e., local or world-wide
Volume, weight, value and type of aid goods the carrier can handle
Any consignment, load and dimension limits
Transit time from door-to-door
Carrier reliability versus risk
Cost for throughput, distance, time per aid goods “unit” moved
Service frequency and schedule flexibility
Carrier’s service range and choice including the use of technology
Any intermediate handling and/or alternative routings

Source: Compiled from *Banomyong and Grant, 2016*

**Figure 2** A humanitarian logistics and SCM balanced scorecard. Source: Adapted from *Widera and Hellingrath, 2016*

may be conflicting, the organizational stages involved may have different objectives and operate under varying circumstances, and the functioning of the BSC requires understanding and support by all personnel involved. *Banomyong and Grant (2016)* identified ten generic objectives or PMs/KPIs unique to humanitarian operations: responsiveness and speed, cost efficiency, standardization, innovation, co-operation, satisfaction of donors and beneficiaries, reliability, flexibility, inventory performance, and bottleneck management.

Three of the BSC elements, financial, process, and organizational capacity, are well known from the classical BSC. However, the introduction of the network element emphasizes the performance of collaboration with external stakeholders in the humanitarian supply chain. Also, an adjustment for the humanitarian context is the replacement of the customer by the “beneficiary and donor” as it is difficult to transfer the concept of a customer to a humanitarian context. The authors addressed this issue by integrating all beneficiary and donor-related KPIs, partly conflicting and partly conforming, into one perspective to adjust the balance between the respective interests.

The advantage of this humanitarian BSC framework is that the different perspectives allow role-based monitoring in terms of both functions and data aggregation levels that consider a decentralized organization structure that is usual practice, i.e. with headquarters and field offices. Thus, performance levels achieved by the different perspectives can be balanced with regards to organization and operation-specific objectives.

Conclusion

Performance measurement is important for organizations to determine how they are meeting their strategic objectives, particularly about their logistics and SCM activities. This chapter could only provide an overview of this topic to guide readers, especially organizations.

Two key elements were presented in this chapter. First, organizations must determine which PMs are applicable and appropriate for monitoring and control. They should naturally devolve from the organization’s strategic objectives. However, as discussed in the special case of sustainable or environmental measures, some reflection and research may be required to ensure they are correct and meet the eight criteria for an appropriate measure.

Second, the selection of an appropriate performance measurement system or framework is important, in order to best use the selected PMs. Several were presented and details on two of the most popular frameworks, the BSC and Deming’s PDCA cycle, were discussed in more detail and their use demonstrated in the two special cases of sustainable or environmental and humanitarian logistics and SCM, which were provided to illustrate issues outside normal business performance measurement.

Biographies

David B. Grant is Professor of Supply Chain Management and Social Responsibility at Hanken School of Economics in Helsinki, Finland. His doctoral research at the University of Edinburgh investigated customer service, satisfaction, and service quality in UK food processing logistics and received the *James Cooper Memorial Cup PhD Award* from the Chartered Institute of Logistics and Transport (UK) in 2003. David's research interests include logistics customer service, satisfaction, and service quality; in-store and online retail logistics; reverse, closed-loop and sustainable logistics; and humanitarian and developmental logistics. He has over 250 publications in various refereed journals, books, and conference proceedings and is on the editorial board of seven international journals. In 2019 David was ranked 5th in *Economics, Business and Management* and 1st in *Industrial Economics and Logistics* in an academic study determining the "top ten professors in Finland" for research impact and productivity and awarded the *Bualuang ASEAN Chair Professorship* for 2019–21 at Thammasat University in Bangkok, Thailand.

Dr Sarah Shaw is Senior Lecturer and Program Director in Logistics and Supply Chain Management at the Hull University Business School, UK. Her PhD thesis investigated environmental supply chain performance measures. Her research interests include green and sustainable logistics, closed-loop supply chains and performance measurement and reporting. She won "Leader of the Year" in the Women in Logistics Awards, 2017 UK for her work with the Chartered Institute of Logistics & Transport UK encouraging young people to work in the logistics sector. She leads various multi-disciplinary research projects and has published in a variety of scientific journals. One of her publications was selected by Emerald Group Publishing to appear in *A Focus on Sustainable Supply Chains and Green Logistics*, part of "Emerald Gems," which brings together some of the most highly cited, read, and innovative research in its field.

References

- Banomyong, R., Grant, D.B., 2016. Transport in Humanitarian Supply Chains. In: Kovács, G., Spens, K., Haavisto, I. (Eds.), *Supply Chain Management for Humanitarians: Tools for Practice*. Kogan Page, London, pp. 191–208.
- Caplice, C., Sheffi, Y., 1994. A review and evaluation of logistics metrics. *Int. J. Logist. Manag.* 5, 11–28.
- Deming, W.E., 1950. *Elementary Principles of the Statistical Control of Quality*. Japanese Union of Scientists and Engineers Conference, Tokyo.
- Grant, D.B., 2012. *Logistics Management*. Pearson Education, Harlow UK.
- Grant, D.B., Elliott, M., 2018. A proposed interdisciplinary framework for the environmental management of water and air-borne emissions in maritime logistics. *Ocean Coast. Manag.* 163, 162–172.
- Hervani, A.A., Helms, M.M., Sarkis, J., 2005. Performance measurement for green supply chain management. *Benchmark.: Int. J.* 12, 330–353.
- Kaplan, R.S., Norton, D.P., 1992. The balanced scorecard - measures that drive performance. *Harvard Bus. Rev.* 70, 71–79.
- Shaw, S., 2013. *Developing and testing green performance measures for the supply chain*. PhD thesis University of Hull, UK. Available from: <https://hydra.hull.ac.uk/resources/hull:8108>.
- Shewhart, W.A., 1939. *Statistical Method from the Viewpoint of Quality Control*. Department of Agriculture, Dover.
- Widera, A., Hellingrath, B., 2016. Making performance measurement work in humanitarian logistics: the case of an IT-supported balanced scorecard. In: Kovács, G., Spens, K., Haavisto, I. (Eds.), *Supply Chain Management for Humanitarians: Tools for Practice*. Kogan Page, London, pp. pp339–352.

Further Reading

- Association for Supply Chain Management, 2019. Supply Chain Operations Reference (SCOR) Model. Available from: <https://www.apics.org/apics-for-business/frameworks/scor>.
- Balanced Scorecard Institute, 2019. Available from: <https://www.balancedscorecard.org/>.
- Beske-Janssen, P., Johnson, P.M., Schaltegger, S., 2015. 20 years of performance measurement in sustainable supply chain management – what has been achieved? *Supply Chain Manag.:Int. J.* 20, 664–680.
- Grant, D.B., Shaw, S., 2019. Environmental or sustainable supply chain performance measurement standards and certifications. In: Sarkis, J. (Ed.), *Handbook on the Sustainable Supply Chain*. Edward Elgar Publishers, Northampton MA, pp. 357–376.
- Grant, D.B., Trautrim, A., Wong, C.Y., 2017. *Sustainable Logistics and Supply Chain Management*, second ed. Kogan Page, London.
- Gunasekaran, A., Kobu, B., 2007. Performance measures and metrics in logistics and supply chain management: a review of recent literature (1995–2004) for research and applications. *Int. J. Prod. Res.* 45, 2819–2840.
- Kasilingam, R.G., 1998. *Logistics and Transportation: Design and Planning*. Springer, Boston MA.
- Neely, A.D., 2005. The evolution of performance measurement, developments in the last decade and a research agenda for the next. *Int. J. Operat. Prod. Manag.* 25, 1264–1277.
- Shaw, S., Grant, D.B., Mangan, J., 2010. Developing environmental supply chain performance measures. *Benchmark.: Int. J.* 17, 320–339.
- The W. Edwards Deming Institute, 2019. Available from: <https://deming.org/>.

Economic Regulation/Deregulation and Nationalization/Privatization in Freight Transportation

Wayne K. Talley, Old Dominion University, Norfolk, VA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	24
Economic Regulation	24
Rate Regulation	24
Entry Regulation	25
Service Regulation	25
Financial Regulation	25
Economic Regulation of US Freight Railroads	25
Economic Regulation of US Truck Carriers	25
Economic Deregulation	26
US Freight Railroads	26
Railroad Revitalization and Regulatory Reform Act (4-R Act)	26
Staggers Rail Act of 1980	26
Economic Deregulation of US Truck Carriers	26
Motor Carrier Act of 1980	26
Household Goods Transportation Act of 1980	26
Nationalization	27
Nationalization of Railroads	27
Privatization	27
Relevant Websites	28
References	28

Introduction

Freight transportation is a service that provides for the movement of goods from one place to another place. Freight transportation decision-makers include: (1) providers of freight transportation service, (2) users of freight transportation service, and (3) government (Talley, 1983). If the unreliability of the transportation of goods within a country by privately owned freight transportation firms is negatively affecting the country's economic growth, for example, the country's government may seek to address this slow-growth-economic problem by economically regulating its privately owned freight transportation firms, that is, by regulating certain economic behaviors of the firms as found in Section "Economic Regulation." US laws that economically regulate US freight railroads and US truck carriers are discussed in Sections "Economic Regulation of US Freight Railroads" and "Economic Regulation of US Truck Carriers," respectively.

The economic deregulation by government of a country's economically regulated transportation firms is discussed in Section "Economic Deregulation," followed by discussions of the economic deregulation of economically regulated US freight railroads and US truck carriers in Sections "Economic Deregulation of US Truck Carriers" and "Nationalization," respectively. The privatization of a country's nationalized freight transportation firms by government is discussed in Section "Privatization."

Economic Regulation

Economic regulation is the imposition of rules by government to modify economic behavior.

The economic regulation of privately owned freight transportation firms by a country's government includes rate, entry, service, and financial regulations (or rules) to modify the economic behavior of the firms (Talley, 1983, p. 41).

Rate Regulation

A rate is a price charged by a freight transportation firm in the provision of a freight transportation service. Rate regulation of a freight transportation firm encompasses both the firm's *level of rates* and its *rate structure*. The *level of rates* controls to some extent the firm's earnings, that is, "rates should be established for a regulated firm so that such rates yield no more than a fair return on a fair value of the property committed to public use . . . or earnings high enough to compensate for risk" (Davis et al., 1975, p. 42). The rate structure addresses the specific rates to be charged, that is, the rates should be nondiscriminatory rates where

discrimination is defined as the “use of different rates under similar circumstances or similar rates under different circumstances” (Davis et al., 1975, p. 117).

Entry Regulation

Regulating the entry of freight transportation firms into geographical areas (or on given routes) for the provision of freight transportation services is concerned with how many freight transportation firms should be allowed to compete in given areas (or on given routes). Specifically, the concern is that the entry of new freight transportation firms in given areas (or on given routes) might have negative impacts on the operations of existing transportation firms that are competing in the areas (or on the routes).

Service Regulation

Once an economically regulated transportation firm has been given the authority to operate exclusively in a given area, the quality of the transportation service that it provides in the area may deteriorate over time from the lack of freight transportation competition. This possibility is addressed by government economically regulating the transportation services provided by the freight transportation firm, that is, by government imposing “three general service requirements on the regulated freight transportation firm: 1) to serve, 2) to deliver and 3) to avoid discrimination” (Talley, 1983, p. 52). The requirement to serve requires the freight transportation firm to serve all who request its services. The requirement to deliver requires the freight transportation firm to deliver freight to the consignee in the same physical condition as received from the shipper and with dispatch. The requirement to avoid discrimination requires that both preferential and prejudicial discrimination by a freight transportation firm be avoided between origin-destination points and among classes of freight.

Financial Regulation

The financial regulation of a freight transportation firm consists of government regulating various financial aspects of the firm such as mergers and insurance. The rationale for controlling mergers and requiring that certain types of insurance (e.g., liability insurance) be held by the economically regulated freight transportation firm is that the firm might endanger itself financially, for example, from mergers and not having certain types of insurance, such as liability insurance, when service accidents occur.

Economic Regulation of US Freight Railroads

The industrialization of the US economy grew rapidly following the end of the US Civil War (in 1865) with the growth of the US railroad industry being the driving force. In 1887, the US Congress responded to the US railroad industry having indulged in business malpractices by passing the Interstate Commerce Act of 1887 that established the economic regulation of US freight railroads engaged in interstate commerce. The Act also sought to protect the public against railroad discrimination and other monopoly-related abuses.

The Interstate Commerce Act of 1887 also established the federal regulator agency, the Interstate Commerce Commission (ICC), to monitor US freight railroads to ensure that they complied with ICC regulations. The ICC had the authority to: (1) establish maximum, minimum, and actual rates to be charged by ICC-regulated freight railroads and (2) control the entry of new freight railroads and changes in the operating structures of existing US freight railroads. The ICC-regulated railroad service is to: (1) ensure that adequate rail service is provided to the public, (2) prevent unjust discrimination in service provision by requiring railroads to have adequate facilities and equipment, and (3) control the abandonment of rail tracks. Financial regulation of US railroads by the ICC included: (1) prescribing the accounting systems to be used by regulated railroads and (2) controlling whether a railroad could issue bonds and capital stocks and whether railroads can merge.

Economic Regulation of US Truck Carriers

The US Motor Carrier Act of 1935, an amendment to the US Interstate Commerce Act of 1887, gave the ICC the authority to economically regulate US truck transportation carriers involved in interstate commerce. The 1935 Act classified ICC truck carriers into two categories: (1) common truck carriers that provided transportation service to the general public and (2) contract truck carriers that held contract agreements with a limited number of customers for the provision of truck transportation service.

ICC truck carriers are required by the Interstate Commerce Act to adhere to a “public convenience and necessity” clause, that is, serve a useful public purpose, be responsive to public demand, and operate the service without endangering or impairing the operations of other existing companies. The 1935 Act also required that rates charged by ICC truck carriers be “just and reasonable” and “not discriminate between customers of similar circumstances” (<https://object.cato.org/sites/cato.org/files/pubs/pdf/pa012.pdf>).

Economic Deregulation

Economic deregulation (also known as economic regulatory reform) of an economically regulated industry is the process by which government removes the economic regulation of the industry by removing (in part or in total) the rules that government had placed on the industry in the past in its economic regulation of the industry.

US Freight Railroads

During the post-World War II period, US ICC economically regulated freight railroads experienced financial difficulties attributed to: (1) payment of property taxes on their rail rights of way (unlike US truck carriers that use public rights of way, i.e., government highways) and (2) costly railroad union work rules. The suppression of railroad intermodal and intramodal price competition from minimum-rate regulation made it difficult for railroads to lower their rates. In response, the US Congress passed the railroad economic deregulation acts—the Railroad Revitalization and Regulatory Reform Act of 1976 (commonly referred to as the 4-R Act of 1976) and the Staggers Rail Act of 1980.

Railroad Revitalization and Regulatory Reform Act (4-R Act)

Under the 4-R Act, railroad rates could be suspended by the ICC on the basis of being unreasonable. “Rates equal to or exceeding variable costs were not to be found unreasonable or unjust on the basis of being too low. A rate is no longer to be declared unreasonable on the basis of being too high unless the ICC first determines that a carrier has market dominance over the traffic involved. The ICC is required to develop market dominance standards” (Talley, 1983, p. 178).

The 4-R Act exempted US railroads from ICC economic regulation in the transportation of certain types of cargo, for example, fresh fruits and vegetables. US railroads responded to the growing list of cargo exemptions by reducing (increasing) rail prices when the freight transportation demand for exempt cargo was low (high) and the availability of rail equipment was high (low).

Staggers Rail Act of 1980

Although the 4-R Act reduced the ICC economic regulation of US railroads, the railroads continued to have financial problems. In response, the US Congress passed the Staggers Rail Act of 1980 to further reduce ICC economic regulation of US railroads, that is, “a railroad may establish any rate it desires so long as the rate contributes to going-concern value, a rate that equals or exceeds variable cost” (Talley, 1983, p. 178). The Act also established a revenue-to-variable-cost schedule for determining whether a railroad has market dominance over transportation service to which a challenged rate applies.

Economic Deregulation of US Truck Carriers

An investigation by the [Committee on the Judiciary of the US Senate \(1980\)](#) on the impact of US economic regulation of the US trucking industry found that economic regulation had negatively affected the industry by contributing to waste and inefficiency, absence of competition, and high rates in the industry. Under economic regulation, ICC truck carriers were restricted to hauling only those commodities so specified in their ICC certificates—thereby possibly resulting in trucks traveling without cargo (or traveling empty) on return trips.

In the early 1980s, there were “approximately 18,000 truck carriers under ICC regulation” and “generally only a few carriers served particular city-pairs. Hence, . . . far less competition among carriers in providing truck service than indicated by the size of the trucking industry” (Talley, 1983, p. 199). Under ICC regulation, truck-carrier rates were “based on value of service and not necessarily on the actual costs of providing the service, i.e., rates were set according to what the traffic will bear” (Talley, 1983, p. 199).

Motor Carrier Act of 1980

In July 1980, the US Congress passed the Motor Carrier Act of 1980 that provided for economic deregulation of the US trucking industry. The Act sought to promote a competitive and efficient trucking industry by “increasing opportunities for new carriers to get into the trucking business and for existing carriers to expand their services” (Talley, 1983, p. 200). With respect to pricing, the Act established a zone of rate freedom that permitted regulated truck carriers to raise or lower their rates by as much as 10% a year without ICC approval. Truck carriers were also permitted to carry deregulated and exempt commodities in the same vehicle at the same time.

Household Goods Transportation Act of 1980

Household-goods truck carriers transport personal effects of individuals as opposed to shipper-cargo truck carriers that transport shipper cargo. Since truck transport of household goods occurs relatively seldom in comparison to the truck transport of shipper cargo, the degree to which individuals understand the use of household-goods truck carriers in the transport of their personal effects is expected to be less than the degree to which shippers understand the use of shipper-good truck carriers in the transport of their cargo.

Given the unique nature of household-goods truck carriers, the US Congress's economic deregulation of household-goods truck carriers was considered separately from the economic deregulation of shipper-good truck carriers (as under the Motor Carrier Act of 1980) by passing the Household Goods Transportation Act of 1980 that sought to reduce unnecessary government regulation of household-goods truck carriers by establishing "new remedies and protections for their customers" (Talley, 1983, p. 201).

Nationalization

Nationalization "is the process of transforming private assets into public assets by bringing them under the public ownership of a national government or state" (https://en.wikipedia.org/wiki/Railway_nationalization). A nationalized freight transportation firm is owned by government versus an economically regulated freight transportation firm that is privately owned but economically regulated by government.

The objectives of a nationalized freight transportation firm in the provision of freight transportation service are those of the government owner, whereas the objectives of an economically regulated privately owned transportation firm in the provision of freight transportation service are those of the private owner of the firm subject to the economic regulations established for the firm by government. The management of a nationalized firm is less effective when its public managers pursue actions that are not in the interests of the government owners of the nationalized firm.

Nationalization of Railroads

Numerous countries have nationalized their privately owned railroads (Table 1). The United Kingdom and the United States nationalized privately owned railroads during World War I.

Privatization

Privatization is the transfer of the production (provision) of goods (services) from the public sector to the private sector. This transfer can occur from: (1) the sale of public assets to private owners to produce goods and provide services and (2) contracting-out of services formerly provided by the public sector to the private sector for provision.

The rationale for the privatization of goods formerly produced by government is to reduce the production costs of the goods. The rationale for privatization of services formerly provided by government is to reduce the provision costs of the services and to improve the quality of the services. The efficiency of privatization in the production (provision) of goods (services) increases when private managers of privatized firms find it in their interests to serve the public interest.

British Railways was nationalized in 1948 when placed under state ownership and its operations were placed under the control of the British Railways Board (BRB). British Railways was privatized in 1993, following passage of the Railways Act of 1993 (<https://en.wikipedia.org/wiki/Impact-of-the-privatisation-of-British-Rail>).

Table 1 Nationalization of railroads

Country	Year nationalized	Railroads nationalized
Argentina	1948	Argentine railways
Canada	1918	Canadian National Railway
France	1982	France's five main railways
India	1951	Indian Railways
Ireland	1953	Great Northern Railway
Japan	1906	Majority of Japan's private railway lines
Spain	1941	Broad-gauge railways
United Kingdom	1914	United Kingdom's railways (returned to their original owners in 1921)
United Kingdom	1948	The four major British railways—Great Western Railway; Southern Railway; London and North Eastern Railway; and London, Midland, and Scottish Railway
United States	1917	Following the entrance of the United States into World War I in 1917, President Wilson on December 26, 1917 nationalized most US railways under the Federal Possession and Control Act, creating the United States Railroad Administration (USRA) that took control of the railways on December 28, 1917. On March 1, 1920, following the end of the war, the railways were returned to their original owners.

Source: The information on the nationalization of railroads found in the table was taken from https://en.wikipedia.org/wiki/Railway_nationalization.

Under privatization, the operations of the BRB were broken up and sold and its various regulatory functions were transferred to the newly created office of the Rail Regulator. The ownership of infrastructure (including larger stations) was transferred to Railtrack, while track maintenance was to be provided by 13 companies across the network (<https://en.wikipedia.org/wiki/Railtrack>).

The ownership and operations of freight trains became the responsibility of two companies—English, Welsh & Scottish (EWS) and Freightliner. A significant change occurred in 2001 with the collapse of Railtrack and the passage of its assets to the state-owned Network Rail.

Relevant Websites

<https://en.wikipedia.org/wiki/Impact-of-the-privatisation-of-British-Rail> - Impact of the privatisation of British Rail.

<https://en.wikipedia.org/wiki/Railtrack> - Railtrack.

https://en.wikipedia.org/wiki/Railway_nationalization - Railway nationalization.

References

Committee on the Judiciary of the US Senate, 1980. Federal Restraints on Competition in the Trucking Industry: Antitrust Immunity and Economic Regulation. US Government Printing Office, Washington, DC.

Davis, G.M., Farris, M.T., Holder Jr., J.J., 1975. Management of Transportation Carriers. Praeger Publishers, New York.

Talley, W.K., 1983. Introduction to Transportation. South-Western Publishing Company, Cincinnati, OH.

Freight Transport Policy

Luca Zamparini*, Aura Reggiani†, *Department of Law, University of Salento, Lecce, Italy; †Department of Economics, University of Bologna, Bologna, Italy

© 2021 Elsevier Ltd. All rights reserved.

The Evolution of Freight Transport Policy	29
Freight Transport Policies in the European Union, United States and Asia	30
European Union Freight Transport Policy	30
United States Freight Transport Policy	31
Asian Freight Transport Policies	32
Conclusions	33
References	33
Further Reading	33

The Evolution of Freight Transport Policy

According to the definition by Slack et al. (2017), transport policy is related to the development of a set of propositions and constructs in order to achieve specific goals pertaining to economic, social, and environmental conditions jointly with the seamless and efficient functioning of the transport system. Transport policies have a paramount importance given that transport activities are at the core of virtually all social, political, and economic activities. Among the main factors influencing the development of freight transport policies in the last decades, it is important to underline the trend toward deregulation and privatization and the increases in world economic growth and in globalization. Both sets of factors have led to the rising importance of policies aimed at regulating, improving and making more efficient freight transport. Moreover, the 1980s and 1990s have witnessed the diffusion of multimodal transportation and of greater concentration on the supply sides of the markets for freight transportation (Reggiani et al., 2000). The move toward combined transport was caused by the concentration of transportation activities in ever larger firms that started to subsume transportation activities in integrated logistics chains. Another phenomenon that was at the base of this shift in freight transportation was the diffusion of third party logistics providers, which serve a large number of clients and take advantage of the emerging economies of scale and of scope in freight transportation. It must also be highlighted that some country policies (i.e., the Austrian system of ecopoints) constituted an important factor in the development of multimodal transportation.

In the last decade, the role of smartness in freight mobility policies has gained importance. This phenomenon has been clearly summarized in a report by Caltrans (2010), which has discussed the principles that should be at the base of smart transport initiatives. Among them, those related to freight transport mobility and policies are the following: (1) location efficiency; (2) reliable mobility; (3) robust economy; (4) health and safety; and (5) environmental stewardship. Location efficiency pertains to the integration of mobility and land use policies. Reliable mobility should be based on transportation network management and on multimodal mobility options. The robust economy principle should consider network performance optimization, the congestion effects on productivity, the efficient use of system resources, and the return on investment, with the aim of enhancing competitiveness. Health and safety should foster management and design practices that lessen exposure to pollution, minimize injuries and fatalities, and encourage active living. This principle is clearly linked to environmental stewardship, based on the reduction of the emission of greenhouse gases and of pollutants. The two latter principles lead to the issue of sustainability; a core topic of transport policies since the late 1990s.

One of the concepts that have been widely considered in order to minimize the impact of freight transport on the environment is the decoupling of transport growth from overall economic growth. A contribution by Schleicher-Tappeser and Hey (1998) has highlighted three basic strategies that may be pursued to accomplish decoupling: (1) the dematerialization of the economy via the replacement of material products with services, miniaturization and the increased durability of products; (2) a reduction of the spatial range of material flows by the use of incentives to increase local production for the local market; and (3) the optimization of transport organization. On the basis of a review of their review of the literature, Tight et al. (2004) proposed a slightly different taxonomy whereby the measures aiming at decoupling may be based on the moderation of demand growth, on modal shift, on increasing transport system efficiency and on improving vehicles and fuels.

On a more general sustainability level, a review paper by Ruamsook and Thomchick (2012) has proposed a conceptual framework of policies which partitions them into four categories:

- *Private sector operational strategies* are based on leveraging technology and implementing Information and Communication Technology (ICT) innovations in order to save fuel and reduce CO₂ emissions, on designing new packaging methods to minimize packaging size and to eliminate unnecessary packaging layers, and on collaborative (shipper-carrier, multishipper, and shipper-costumer) transportation to share capacity, to create round trips and to align delivery schedules.

- *Private sector strategic actions* include the restructuring of physical logistics systems, the training of personnel and compensation schemes and the alignment of incentives and performance measures. The first class of strategies concerns the location of warehousing and manufacturing sites in the proximity of large clusters of customers and/or supply facilities. Personnel training is aimed at optimizing the behavior of drivers in all phases of transport activities. The alignment of incentives and performance measure requires the involvement of all suppliers, carriers and third party logistics providers in changing and updating performance measures, altering contract specifications and modifying payment schemes. Moreover, performance metrics related to cost, quality, and time are required to take into account sustainability measures, such as reducing empty miles, freight comingling, the use of alternative fuels, route optimization and delivery efficiency and fuel reduction. Finally, differential price schemes should be implemented in case customers purchase a less than full shipment or less fuel-efficient transport modes.
- *Public sector operational actions* deal with the improvement of traffic capacity in the extant transport systems. Notable examples in this respect are represented by the application of intelligent transportation systems (leading to the reduction of emissions and congestion) and the restrictions on truck delivery hours in cities.
- *Public sector strategic actions* aim at changing substantially the transport systems and comprise incentives, facilitation, and direct regulation. The incentives can either take the form of pricing and taxation in the case of negative externalities (i.e., carbon pricing and cap and trade systems in the United States) or grants and subsidies if positive externalities are generated (i.e., the Diesel emission reduction act, the Air quality standard attainment program). The facilitation can be represented by grants for research and development (see, among others, Horizon 2020, which is the biggest European Union (EU) research program), as well as by new patents that constitute the instrument for sustainable transport options or by the pursuit of investments aimed at facilitating behavioral changes and modal shift in the operational, tactical, and strategic management of freight activities. Examples of possible regulations include emission and engine standards for all transport modes and the promotion of renewable fuel standards.

Freight Transport Policies in the European Union, United States and Asia

The previous section has highlighted the evolution of freight transport policies in the last decades and the general shift toward smartness and sustainability. The current section is devoted to the discussion of the main principles that have guided the freight transport policies that have been carried out in the EU, in the United States and in Asia.

European Union Freight Transport Policy

Before the beginning of the 1990s, freight transport policies in the EU were mainly based on local and national initiatives. Only the Single Act of 1986 and the consequent liberalization of the movement of goods (among other liberties) paved the way for unified EU freight transport policies. As a result, many of the interventions that were proposed in the 1990s aimed at establishing a common market. In this context, the strengthening of the Trans European Transport Networks constituted the main basis of the EU policies in this field. Moreover, many of the proposals carried out at the EU level were based on previous experiences in the EU countries. In 2011 the White Paper ([European Commission, 2011](#)) recommended a 20% reduction in transport emissions (excluding international maritime transport) between 2008 and 2030, and a reduction of at least 60% between 1990 and 2050. It also sought a 40% reduction in emissions from international maritime transport between 2005 and 2050. Consequently, as also indicated in the first section, a new emphasis was placed on sustainability issues and on the need to decouple economic growth from transport growth.

An interesting analysis carried out by [Aditjandra \(2018\)](#) by means of a literature review has highlighted that the EU freight transport policy has apparently been based on the following seven main pillars: (1) general freight issues, (2) rail, (3) port/maritime, (4) intermodal, (5) road freight and intelligent transport systems; (6) urban freight mobility, and (7) TransEuropean Networks (TENs).

With respect to general freight issues, the main themes that have emerged are transport planning, supply chain and green logistics. Rail freight concerns are mainly related to open access, material, and technology improvements and to modal shift initiatives. The port/maritime policy has targeted both the nodes in terms of port policies and the arcs with the possibility to establish the, so-called, motorways of the sea that were aimed at achieving the largest possible degree of modal shift from other (especially motorized) transport modes. The diffusion of intermodal transport was to be pursued mainly by the transfer of technology and terminal comodality. [Zografos et al. \(2012\)](#) proposed a taxonomy of the major policy impact areas:

- *Competition*—obtained through the improvement of market access rules, a reduction in the differences between transport related taxes, use based charging, a strategy for a common European maritime space, a freight oriented rail network, and a framework for fair competition among ports.
- *Efficiency*—based on the promotion of innovative technologies, the use of corridors in order to take advantage of the economies of scale, the training of freight personnel, the minimization of the infrastructure and technical limits to rail transport, and the diffusion of ITS and ICT systems.
- *Linkage of modes*—by fostering short sea shipping and the related land connections, simplifying the administrative burdens of freight transport chains, integrating waterborne transport modes, improving the quality of logistics services, and enlarging the capacities of ports and terminals

- *Social and environmental safety and security*—by the implementation of common security rules, the promotion of green transport and of energy efficiency in all modes, the introduction of new technologies and by facilitating cooperation among private and public stakeholders in different countries.

The main themes pertaining to road freight and ITS were related to the consideration of externalities and to the possible tax systems that may be put in place in order to internalize them. On the other hand, intelligent freight systems should be used to mitigate security risks, to improve the efficiency of freight transport activities and to coordinate road transport with other transport modes. At the local level, urban freight mobility should have the main aim to minimize the impact on air quality, to improve sustainability and to be based on proper planning in terms of land use in order to carefully implement an appropriate quantity of nodal infrastructure, such as warehouses and logistic platforms. In this context, the issues of value of time savings and reliability in freight transport, jointly with the related freight transport forecasts by means of innovative techniques and models, play a significant role in the strategic decisions of the policy makers (de Bok et al., 2018; de Jong, 2008). However, urban and regional freight transport policies should also be aimed at moving from traditional quantitative measurements of flows to qualitative considerations that would consider the economic impact at the local level and the enhancement of local industrialization and development processes (Zamparini, 2017). Moreover, efficient freight policies should manage to reconcile the heterogeneous interests of the various urban freight transport stakeholders (supply chain actors, resource supply actors, public authorities, residents, and visitors).

The main issues that have characterized the policies related to TENs for Transport (TEN-T) are the modeling of freight scenarios, the possibility to deploy green corridors and the regional impacts emerging from their implementation. The main aims of the TEN-T policy have been attained through the development of interoperability and of interconnectivity (Janic and Reggiani, 2001), in the light of the emerging era of globalization. Interconnectivity requires the coordination of various transport modes to deliver door-to-door services, while interoperability deals with the possibility to use compatible and standardized equipment, technology, facilities, and infrastructure with compatible vehicles.

It should also be noted that the 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, will certainly move current freight transport policies toward more innovative solutions, with the aim of fulfilling the 17 Sustainability Goals. Concerning the EU, the related EU documents approved in 2016 and 2017 (European Commission, 2019) show the EU initiatives within this framework, such as the multistakeholder platform on SDGs, set up on May 22, 2017, which aims to provide a forum for the exchange of experience and best practice on the implementation of the SDGs across sectors and at local, regional, national and EU level. The efforts of the EU toward achieving an integrated sustainable policy at all spatial levels (EU, national, regional, urban and local) should be emphasized in this respect.

United States Freight Transport Policy

A document published in 1997 by the Department of Transportation (DoT, 1997) has set the eight main principles that should guide the United States freight transport policy.

- The first principle was related to the provision of funding and of a planning framework that would establish priorities for the allocation of Federal resources to cost-effective infrastructure investments to support broad national goals. In this respect, the Federal administration would intervene either to facilitate the resolution of a freight transportation problem or to fund projects with a national significance. Under the same principle, rigorous cost benefit analyses were deemed necessary for all proposed investment initiatives.
- The second principle aimed at promoting economic growth by removing unwise or unnecessary regulation and through the efficient pricing of publicly financed transportation infrastructure.
- The third principle stated the necessity to ensure a safe transportation system. This would require both public regulation and enforcement and private voluntary compliance efforts that should be supported.
- The fourth principle endorsed the protection of the environment and the conservation of energy. The cooperation of all levels of government and the private sector was considered an important requirement to fulfill this goal. In this respect, the federal government was called to reduce the social costs of transportation and to devise taxes and subsidies that would consider the negative and positive externalities generated by transportation activities. Research and technology development were considered as the two most important drivers of the minimization of environmental impact and energy conservation.
- Coherently, the fifth principle advocated the use of advances in transportation technology to promote efficiency, safety, and speed.
- The sixth principle aimed at fostering policies that would effectively meet defense and emergency transportation requirements. The main focus of this principle in 1997 was related to facilitating an efficient response to natural disasters. It is evident that, after the events of September 11 2001, the focus drastically shifted toward the enhancement of security against terrorist acts, with a clear reorganization of the responsible Federal agencies.
- The seventh principle aimed at facilitating international trade and commerce by upgrading the existing facilities that connected the country to global supply chain systems.
- The last principle fostered the promotion of effective and equitable joint use of transportation infrastructure for freight and for passenger services, with a particular emphasis on the consideration and minimization of the safety issues that may arise from this conjoint use.

Table 1 Policies on the infrastructure systems that impact on freight transport

	<i>Federal</i>	<i>State</i>
Operation and maintenance policies	Truck size and weight rules Corps of Engineers maintenance dredging Corps lock and dam maintenance Corps decisions on water levels and flows on rivers	Highway operations and maintenance decisions Enforcement of size and weight rules for trucks Seasonal load limits on highways Truck routing restrictions
Investment policies	Level of highway funding Highway design standards Aid for railroad infrastructure Inland waterway investment Modal split of funding	Level of highway funding Project selection Design and build highway projects Modal split of funding
Finance policies	Fuel taxes Approval for tolls and other user charges Airport peak pricing policies	Fuel taxes Tolls and other user charges Privatization of roads Port fees

Source: Authors' elaboration based on *NASEM (2011)*.

A report by the National Academies of Sciences, Engineering, and Medicine (*NASEM, 2011*) has clarified that the Federal administration has deployed in the last decades two major roles: funding and regulation. The report also provides a taxonomy of the policy categories that may affect the various modes of freight transport and constitutes, therefore, an update of the report issued in 1997 (*DoT, 1997*): (1) safety, (2) security, (3) land use, (4) environment, (5) energy and climate change, (6) trade and economic regulation, (7) infrastructure operations and management, (8) infrastructure investment, and (9) infrastructure finance.

Safety policies that have been implemented at the federal level include truck, railroad, and aviation hours-of-service rules, interstate speed limits, truck speed governor rules, truck electronic onboard recorder rules, rules for the inspection of truck and vehicles, rules for aircraft design, hazmat rules, and Coast Guard rules for barges and barge operations. At the state level, highway speed limits, the enforcement of federal truck rules and restrictions on locomotive horns. Federal security policies are the transportation worker identification credential, truck driver background checks, US exit fingertips rules, maritime administration foreign crew ID requirements, airport security protocols, chemical facility antiterrorism standards, and screening of cargo on passenger aircraft and of import containers, customs rules, and programs. At the state level, it is worthwhile mentioning routing and infrastructure access restrictions. Federal and state land use policies affecting freight transport are respectively represented by brownfield programs and by land use planning requirements. Federal environmental policies are constituted through emission and fuel standards, air quality standards and planning requirements, the management of dredging spoils, water pollutant discharge rules for vessels, and oil spill prevention rules. With respect to energy and climate change policies, it is possible to highlight at the federal level the requirements or subsidies for renewable fuels, the greenhouse gases emissions regulations and the programs and financing to improve fuel efficiency and to introduce alternative fuels. Most of these policies are also pursued at the state level.

Table 1 provides a list of examples of policies related to the infrastructure system that have an impact on freight transport.

Asian Freight Transport Policies

A report by the Asian Infrastructure Investment Bank (*AIIB, 2018*) provides a clear picture of the situation of the transport system in Asia and of the related needs in terms of policies and investments. Despite the importance of transport for trade and economic development, the provision of transport infrastructure and services is highly uneven across Asia and, for most countries, it lags behind that of developed economies. Policies aimed at upgrading and making more efficient the provision of transport infrastructure and systems are particularly important in a region that is generally marked by rising degrees of affluence and freight transport demand. Consequently, freight transport policies should provide the necessary connectivity among countries. At the time of writing, there are five main initiatives in this respect:

- The ASEAN connectivity initiative focuses on sustainable infrastructure and seamless logistics within ASEAN countries.
- The Central Asia Regional Economic Cooperation program defines six transport corridors among the 11 countries participating in the initiative with a particular emphasis on rail, roads, and dry ports.
- The Greater Mekong Subregion Cooperation program involves Cambodia, China, Lao People's Democratic Republic, Myanmar, Thailand and Vietnam, with the aim of facilitating crossborder freight transport with a mixture of infrastructure and of immaterial investments.
- The South Asia Subregional Economic Cooperation program aims at promoting crossborder transport networks and the reduction of freight transport costs among Bangladesh, Bhutan, India, Maldives, Myanmar, Nepal and Sri Lanka.
- Last, the "One Belt, One Road" initiative or "Belt and Road Initiative" is mainly backed by the Chinese government and is not just related to transportation. It involves Asian, African and European countries and its stated purpose is to increase economic cooperation and prosperity among countries. In this context, freight transport is expected to play a very important role. The

expected outcomes are a rebalancing of modal shift with a lower dependency on road transport, a higher diffusion of multimodal transportation chains, an enhanced rail–road connectivity in Central Asian Countries and the development of several logistics hubs in Eastern Europe, Central Asia, and the Middle East.

As for India, a report by the [National Transport Development Policy Committee \(2014\)](#) has stated that the main objective of Indian freight transport policies lies with the creation of an interconnected, hierarchical, and integrated transport network. In this context, a rebalancing of modal shift toward railway transport is considered as an important achievement, both in terms of sustainability and energy efficiency. Moreover, the creation and completion of a series of dedicated freight transport corridors is envisaged jointly with the provision of new local ports for the diffusion of coastal shipping, dedicated facilities for air cargo in airports and various logistics parks close to the main metropolitan areas. In terms of governance, the proposals re aimed at creating a stable Transport Ministry and an Office of Transportation Strategy, at providing a comprehensive regulatory environment and at decentralizing policy decisions, where appropriate, to urban and metropolitan governments. Urban freight management is likely to become a pressing issue in all cities in Asia (and in the rest of the world). Appropriate policies should reduce the need to move goods, to shift local freight transport to more sustainable modes and to improve the operations, technologies, and energy efficiency of current road freight transport ([GIZ, 2013](#)).

Conclusions

This review of freight transport policies has underlined the fact that, in this current era of increasing connectivity, networking, and globalization, the urgent objective of achieving sustainable transport, as established by the SD-2030 Agenda, will most probably raise the issue of how to develop current freight transport policies in the EU, the United States and Asian Countries toward common policy strategies at a very global level.

References

- Aditjandra, P.T., 2018. Europe's freight transport policy: analysis, synthesis and evaluation. In: Shiftan, Y., Kamargianni, M. (Eds.), *Preparing for the New Era of Transport Policies: Learning from Experience*. Elsevier, Amsterdam, pp. 197–243.
- AIIB, 2018. Transport sector study, Asian infrastructure investment bank. https://www.aiib.org/en/policies-strategies/_download/transport/2018_May_AIIB-Transport-Sector-Study.pdf (accessed 23.10.19).
- Caltrans, 2010. *Smart Mobility 2010: A Call to Action for the New Decade*. Caltrans, Los Angeles, CA.
- de Bok, M., Wesseling, B., Kiel, J., Francke, J., Miete, O., 2018. A sensitivity analysis of freight transport forecasts for the Netherlands. *Inter. J. Trans. Econ.* 45 (4), 567–584.
- de Jong, G., 2008. Value of freight travel-time savings. In: Hensher, D.A., Button, K.J. (Eds.), *Handbooks in Transport, Volume 1: Handbook of Transport Modelling*. Elsevier, Amsterdam, pp. 649–663.
- DoT, 1997. National Freight Transportation Policy Statement, Department of Transportation, Washington, DC. https://rosap.nhtl.bts.gov/view/dot/15739/dot_15739_DS1.pdf? (accessed 13.11.19).
- European Commission, 2011. White Paper: Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system. <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0144:FIN:en:PDF> (accessed 14.11.19).
- European Commission, 2019. Sustainable development goals. https://ec.europa.eu/info/strategy/international-strategies/sustainable-development-goals_en (accessed 14.11.19).
- GIZ, 2013. *Sustainable Urban Freight in Asian Cities*. GIZ, Bonn.
- Janic, M., Reggiani, A., 2001. Integrated transport systems in the European Union: an overview of some recent developments. *Trans. Rev.* 21 (4), 469–497.
- NASEM, 2011. *Impacts of Public Policy on the Freight Transportation System*. National Academies of Sciences, Engineering, and Medicine, The National Academies Press, Washington, DC.
- National Transport Development Policy Committee, 2014. *India Transport Report. Moving India to 2032*. Routledge, New Delhi.
- Reggiani, A., Cattaneo, S., Janic, M., Nijkamp, P., 2000. Freight transport in Europe. Policy issues and future scenarios on Trans-Border Alpine connections. *IATSS Res.* 24 (1), 48–59.
- Ruamsook, K., Thomchick, E.A., 2012. Sustainable Freight transportation: A Review of Strategies, Transportation Research Forum Annual Forum Conference Proceedings, pp. 20. <https://ageconsearch.umn.edu/record/207083/> (accessed 02.10.19).
- Schleicher-Tappeser, R., Hey, C., 1998. Policy Approaches for Decoupling Freight Transport from Economic Growth, Paper Presented at the 8th World Conference on Transport Research. https://www.researchgate.net/publication/228716486_Policy_approaches_for_decoupling_freight_transport_from_economic_growth (accessed 12.11.19).
- Slack, B., Notteboom, T., Rodrigue, J.P., 2017. Transport planning and policy. In: Rodrigue, J.P., Comtois, C., Slack, B. (Eds.), *The Geography of Transport Systems*. Routledge, New York. https://transportgeography.org/?page_id=147 (accessed 02.10.19).
- Tight, M.R., Delle Site, P., Meyer Rülhe, O., 2004. Decoupling transport from economic growth: towards transport sustainability in Europe. *Eur. J. Trans. Infrastruct. Res.* 4 (4), 381–404.
- Zamparini, L., 2017. Urban mobility in a smart development perspective. In: Antonelli, G., Cappiello, G. (Eds.), *Smart Development in Smart Communities*. Routledge, Oxon, pp. 294–304.
- Zografos, K.G., Sedlacek, N., Bozuwa, J., 2012. A comparative assessment of freight transport and logistics policies in Europe. *Procedia Soc. Behav. Sci.* 48, 2523–2532.

Further Reading

- Banister, D., 2002. *Transport Policy and the Environment*. Spon, London, New York.
- Banister, D., 2018. *Inequality in Transport*. Alexandrine Press, Abingdon, UK.
- Button, K., 2010. Transportation economics: some developments over the past 30 years. *J. Transp. Res. Forum* 45 (2), 7–30.
- Danielis, R., Rotaris, L., Marcucci, E., 2010. Urban freight policies and distribution channels: a discussion based on evidence from Italian cities. *Eur. Trans.* 46, 114–146.
- Divall, C., Hine, J., Pooley, C., 2016. *Transport Policy: Learning Lessons from History*. London, Routledge.
- Gatta, V., Marcucci, E., 2014. Urban freight transport and policy changes: improving decision makers' awareness via an agent-specific approach. *Trans. Policy* 36, 248–252.
- Givoni, M., Banister, D. (Eds.), 2010. *Integrated Transport: From Policy to Practice*. Routledge, Abingdon, pp. 75–96.

- Gwilliam, K.M., Mackie, P.J., 1975. *Economics and Transport Policy*. Routledge, Abingdon [reprinted in 2017].
- Konings, R., Priemus, H., Nijkamp, P. (Eds.), 2008. *The Future of Intermodal Freight Transport. Operations, Design and Policy*. Edward Elgar, Cheltenham.
- Leinbach, T., Capineri, C. (Eds.), 2007. *Globalised Freight Transport. Intermodality, E-Commerce, Logistics and Sustainability*. Edward Elgar, Cheltenham.
- Macharis, C., Nocera, S., 2019. The future of freight transport. *Eur. Trans. Res. Rev.* 11 (21), 1–2.
- Stanley, J., Hensher, D. (Eds.), 2019. *A Research Agenda for Transport Policy*. Edward Elgar, Cheltenham.
- Stopher, P., Stanley, J., 2014. *Introduction to Transport Policy. A Public Policy View*. Edward Elgar, Cheltenham.
- van Geenhuizen, M., Reggiani, A., Rietveld, P. (Eds.), 2007. *Policy Analysis of Transport Networks*. Ashgate, Aldershot.
- van Wee, B., Annema, J.A., Banister, D. (Eds.), 2013. *The Transport System and Transport Policy. An Introduction*. Edward Elgar, Cheltenham.

Planning and Financing Logistics Spaces

Nicolas Raimbault, Institute of Geography and Regional Planning, University of Nantes, CNRS UMR 6590 “Espaces et SOciétés” (ESO), Nantes, France; and Urban Development and Mobility Department, Luxembourg Institute of Socio-Economic Research, Esch-sur-Alzette, Luxembourg

© 2021 Elsevier Ltd. All rights reserved.

Introduction	35
Logistics Zones: The Silent Privatization of Ordinary Logistics Spaces	36
Warehouses, Distribution Centers, and Logistics Zones	36
From Industrial Zones to Logistics Zones	36
Local Policies of Logistics Zones Development	36
Financialization of Logistics Real Estate and Emergence of Private Logistics Parks	37
International Gateways: Planning and Financing Public Logistics Infrastructures	37
The Logistics Facilities of International Gateways	38
International Gateways Governance: The Production of Logistics Spaces and Services	38
Planning Policies for Logistics Facilities: Between Regional Planning and Local Urban Logistics Innovations	39
Regional and Metropolitan Spatial Planning Policies	39
Local Urban Logistics Innovations	39
Conclusions	40
Biography	40
References	40

Introduction

Logistics encompasses a diversity of activities concerned with the management and the operation of goods flows, including transportation, handling and packing, warehousing, transshipment, pre- and post-manufacturing, and supply chain management. Within this large economic sector, three main logistics functions can be identified: international flows, regional distribution, and city logistics. The management of international flows is a crucial task in a context of global production and distribution networks (Coe, 2014). International flows mostly rely on maritime and air transport modes—key infrastructures are thus seaports and airports. However, most goods flows are national or regional and are managed by regional distribution systems. They depend on, mainly, road and, secondary, rail and river transport modes. Warehouses, often called distribution centers (DCs) (section, Logistics Zones: The Silent Privatization of Ordinary Logistics Spaces), represent crucial nodes to (re)organize shipments within regional distribution systems (Hesse, 2008). Intermodal terminals can be needed for transshipments from the different transport modes. Eventually, city logistics, also called last mile logistics, corresponds to the final step of the delivery process. It is mainly performed by light trucks even if greener vehicles (electric vans and cargo cycles) are now being experimented with. These logistics services can need specific facilities in dense parts of metropolitan areas (Browne et al., 2019). Most logistics flows articulate these three functions, which are required to connect the production of goods to their consumption and their recycling. Thus a large range of logistics spaces—seaports, intermodal terminals, logistics parks, DCs, urban consolidation centers to name a few—have been built during the last decades in order to perform these different logistics activities.

The diversity of logistics spaces, however, does not only reflect this diversity of logistics functions. In this perspective, the way these spaces are planned, produced, regulated, governed, financed, maintained, and accepted by local communities matters a great deal. In other words, logistics spaces display different configurations corresponding to their institutional, political, economic, and social environments (Hall and Hesse, 2013).

This chapter proposes an exploration of the way current logistics spaces are planned and financed by the different public and private actors involved. It will highlight the impacts of planning policies and financial circuits on their spatial and institutional configurations. In this perspective, along with their logistics roles, the chapter underlines three main dimensions of the current systems of production of logistics spaces: the modalities of the financialization of logistics properties, the scope and effectiveness of planning policies for logistics facilities, and the scope of traditional public infrastructure regulations and management regimes. These three main dimensions are indeed the three central variables of the existing modes of governance of logistics spaces.

The outline of this paper is as follows. The next section is dedicated to logistics zones, where most warehouses and distributions centers are concentrated. Current logistics zones are embedded in historical industrial urban areas or correspond to new private business parks. In the second next section, the specific mode of governance of international gateways (airports, seaports, and inland ports) is analyzed as a part of traditional public infrastructure policies. Against these two main configurations of logistics spaces production, the role of current spatial planning policies for logistics facilities are investigated in the third next section.

Logistics Zones: The Silent Privatization of Ordinary Logistics Spaces

The development of logistics activities and flows entails the construction of thousands of warehouses and DCs that are mainly concentrated in logistics zones in urban regions. This section presents a typology of three kinds of logistics zones based on the comparison of different European and North American case studies in the urban regions of Paris, Atlanta, and Frankfurt (Raimbault et al., 2019; Barbier et al., 2019). This typology indicates the existence of different modes of logistics zone governance, corresponding to different phases of logistics development, planning, and financing regimes. A comparison with case studies in other regions could reveal the existence of other modes of logistics zone governance.

Warehouses, Distribution Centers, and Logistics Zones

Warehouses and DCs are industrial sites dedicated to logistics operations. Shippers and logistics providers both operate warehouses and DCs within their distribution networks. In these sites, four basic operations are performed: the pickup of goods, consolidation/deconsolidation of shipments, change from one means of transportation to another, and storage. In contrast to traditional warehouses, DCs are designed in order to minimize the storage of goods and to keep goods as mobile as possible in a context of just-in-time flows (Cidell, 2015): the goods are collected from different origins, grouped in DCs, and then redistributed to their respective destinations. DCs are the contemporary configuration of warehousing.

Yet the establishment of DCs is not solely the result of corporate logistics strategies. These buildings of thousands of square meters are subject to several spatial planning regulations, as well as land development and real estate investment strategies. In most developed countries, municipalities and local communities design and implement local zoning plans that authorize the use of land for establishing DCs. They also issue the required building permits. Through this process, the real estate industry (land developers, developers and investors) come under the influence of local public and even political power that is expressed in planning regulations or social movements. Moreover, some local governments initiate the development of new business zones dedicated to logistics facilities.

The development of logistics activities during the last decades has led to the emergence of different kinds of business zones concentrating logistics facilities. Some logistics zones are former industrial zones, while others are planned by local government or produced by property investors.

From Industrial Zones to Logistics Zones

During a first period of logistics development (1970s–1990s), logistics providers and shippers, looking for sites in urban regions to build warehouses, found spaces in the large existing industrial zones of European and North American cities. Warehouses have replaced former factories or have been built on plots that became available, when the demand for new manufacturing sites declined. This process has led to the incremental conversion of industrial zones into logistics zones.

The establishment of these logistics sites is not based on complex political negotiations, or specific real estate or land development operations. The land, generally developed by public land developers, is available for any type of industrial purpose, whether manufacturing or logistics. Municipal authorities are only asked to give their formal agreement by signing the building permits. The shift to logistics of these former industrial sites is therefore almost invisible, without explicit public discussion or negotiation between public and private stakeholders.

The silent conversion of industrial zones into logistics zones corresponds in this way to a first mode of governance of logistics zones, which is limited to the enforcement of generic rules of land use, and does not tilt local political agendas toward other issues of logistics industry development. It mainly takes place in the former manufacturing belt in metropolitan areas (Raimbault et al., 2019).

Local Policies of Logistics Zones Development

The increasing demand for logistics spaces has led to a second generation of logistics zones corresponding to another mode of governance. Many local governments or authorities implement economic development policies based on attracting logistics facilities in order to increase their tax revenues and the number of local jobs. The implementation of this agenda consists of publicly developing new business zones dedicated to logistics activities.

Thus local governments plan new logistics zones in their territory and invest in the corresponding land development operations and in the subsequent road (and sometimes rail or river) infrastructure. Then they sell the plots directly to shippers or logistics providers looking for new logistics sites or, much more often, to property developers building DCs in order to rent them. Some municipalities also plan and finance rail or river container terminals in these zones, in order to strengthen its attractiveness or to favor more ecological transport modes.

In this way, the development of this second generation of logistics zones is the result of voluntarist local spatial planning policies. This mode of logistics zone governance focuses on land development issues, at the expense of other issues such as employment or housing and public transport for the workers of the zone. Moreover these logistics zones are only planned at the municipal scale (Cidell, 2011). In most urban regions, the issue of planning logistics zones is not in the scope of regional or metropolitan planning policies (section, Planning Policies for Logistics Facilities: Between Regional Planning and Local Urban Logistics Innovations). Thus,

in the absence of strong planning policies at this scale, these local policies result in logistics zones spreading toward suburban and outer-suburban areas, participating in the process of “logistics sprawl” (Dablanc and Ross, 2012). However, the development of suburban zones can meet with objections from environmentalists, which can hinder the sprawl (Barbier et al., 2019).

Financialization of Logistics Real Estate and Emergence of Private Logistics Parks

Since the 1990s, logistics firms have largely opted for flexible real estate solutions and thus have looked for warehouses to rent, rather than building and managing their own facilities. This has contributed to the emergence of a development and investment market in logistics real estate (Hesse, 2008), which refers to the general process of the financialization of real estate. The financialization of logistics real estate is connected to a third mode of governance that is specific to the private logistics parks.

The logistics real estate market is dominated by international firms, which specialize in logistics and manage global investment funds. These companies take direct charge of the development of the warehouses they buy, as investment fund managers. In order to reduce their dependence on negotiations with local public authorities, they also tend to be the developers of the logistics zones in which they invest. In other words, instead of building warehouses scattered around different business zones, the industry leaders develop private logistics zones containing several warehouses. These “logistics parks” are entirely owned and operated by the same investment fund manager responsible for property management. They are fenced and protected by private security.

This business model leads to the privatization of a number of local policies. Logistics real estate firms privatize land development policies, compared to the situation of business zones directly developed by local governments. To the extent that logistics parks are entirely private, real estate firms become the *de facto* owners and managers of the streets and green spaces that constitute the public spaces in the business parks. Moreover the model also enables real estate companies to decide on local economic development issues, insofar as they select the firms that settle in the municipality, which considerably affects the latter’s economic specialization and prospects.

However, logistics parks must be authorized and supported by local governments, which are responsible for issuing spatial planning documents and building permits. The production of logistics parks implies that the local authorities accept this dynamic of privatization. Different mechanisms explain this political choice. First, some local authorities in the outer suburbs lack the financial, technical and even political resources needed for developing business zones. In order to get the capacity to attract logistics sites in their territory, they look for private investors able to establish private business zones. Second, some outer suburban municipalities argue that the private logistics park model is superior to traditional publicly developed business zones: the general design of the park and the fact that it is fenced and secure, the fact that private development and management of parks makes no demands on the public purse, eventually the fact that the property manager is solely responsible for the entire park and can be contacted directly by local governments. These different aspects of private logistics parks give local governments a greater sense of control over their territory.

This last mode of governance, dominated by the logistics real estate industry, becomes dominant in large urban regions prone to strong political fragmentations. The consequences are twofold. At the local scale, local governments negotiate only with property developers and investors. They rarely meet the users of the warehouses, the workers, or even the logistics firms themselves. Managing the relations with the firms that rent the warehouses becomes the task of the property manager alone. In consequence, logistics issues are seen as a question of real estate, disconnected from matters relating to logistics activities and employment, such as employee transport or transfer of goods flows from road to rail or river modes. At the regional scale, private logistics parks directly challenge planning policies (section, Planning Policies for Logistics Facilities: Between Regional Planning and Local Urban Logistics Innovations). As these real estate products are particularly attractive for outer-suburban areas, where local authorities do not have the resources or the desire to develop logistics zones alone, the financialization of logistics real estate largely contributes to logistics sprawl.

The evolution of the planning and financing regimes dedicated to logistics zones, from traditional industrial zones to private logistics parks, is due to the emergence of a global financialized development and investment logistics real estate market. In a double context of weakness of planning policies for logistics facilities at the metropolitan scale and of the lack of financial, technical and even political resources in many local authorities in the outer suburbs, the financialization of logistics real estate leads to the privatization of the production of logistics spaces in many metropolitan areas. Besides this fragmented mode of governance of logistics zones, the management of international gateways (airports, seaports, and inland ports) corresponds to a specific mode of governance tied to traditional public infrastructure policies.

International Gateways: Planning and Financing Public Logistics Infrastructures

Maritime transportation represents between 70% and 80% of the total of international flows, with the rest mainly completed by air transport (Rodrigue et al., 2017). Efficient transshipment facilities from maritime and air transport to land transport modes are thus needed for international flows. Airports, seaports and also different kinds of inland ports, that complement maritime gateways, constitute the systems of international gateways. As they are framed as being of strategic importance by local and national governments, international gateways are mainly organized and managed by public or semipublic authorities, which constitutes specific modes of governance of logistics spaces. Most airports and seaports and some inland ports thus constitute emblematic cases of logistics spaces planned and financed as public infrastructure.

The Logistics Facilities of International Gateways

International gateways are based on two main kinds of facilities. First of all, intermodal terminals are used for the transshipment of goods from maritime or air transport to land transport modes. Seaports are organized around container terminals and different types of bulk terminals (dry and liquid bulk, general cargo, and roll on/roll off). These terminals are usually operated by private terminal operators, which can be subsidiaries of the main maritime shipping companies. In airports, cargo stations organize the links between airport runways and trucking. These facilities are directly rented by air, courier, and express delivery companies.

Airports and seaports also concentrate numerous other logistics facilities needed for the organization of hinterland flows. On the one hand, shippers' DCs and carriers' parcel service facilities are located in the domain of seaports and airports in order to organize the different inland shipments at the local, regional, national, and continental scales. On the other hand, some river and rail (containers and bulk) terminals are usually localized in seaports, as rail and river transport modes can be efficient hinterland transport solutions.

Several inland ports complement the major seaports. Inland ports are "inland facilities with or without an intermodal terminal and logistics companies, which [are] directly connected to seaport(s) with high capacity transport mean(s) either via rail, road or inland waterways, where customers can leave/pick up their standardized units as if directly to a seaport" (Witte et al., 2019, 54). They offer the services of an extended gateway within the context of seaports becoming integral parts of extensive hinterland networks, intermodal transport corridors and inland ports (Notteboom and Rodrigue, 2005). Inland ports are located in different cities of different size, from small cities located on major transport corridors to large urban regions.

International gateways are thus complex logistics spaces, corresponding to different scales of flows, different modes of transport, and different logistics services. A large variety of logistics firms are involved in international gateways, which supposes specific institutional arrangements in terms of infrastructure, land and operations management.

International Gateways Governance: The Production of Logistics Spaces and Services

The way most seaport spaces are produced and governed is emblematic of the specific public infrastructure mode of governance. Even if there is a diversity of governance configurations, the governance of airports and inland ports generally corresponds to this infrastructure mode of governance (Rodrigue et al., 2010).

The ownership, management, and operation of the facilities of seaports and airports are rarely entirely public or entirely private. Concerning seaports, the most common configuration is known as *Landlord port* (Brooks and Pallis, 2012; World Bank, 2007). This mode of governance is based on the separation of public and private spheres of intervention. A public port authority (PA) is responsible for the production and the maintenance of the port spaces, while the private operators are responsible for the production of the different logistics services and the private facilities needed for them. In some seaports, especially in England (Baird, 2006), the ownership, the regulation and the operation can be the responsibility of the private sector. This mode of governance, known as *Private Service port*, remains much less common than the Landlord port. The ownership of airports is often public, while the management of the infrastructure can be entrusted to a private operator under a concession contract. Some inland ports are governed by PAs according to the principle of a Landlord port. The next sections develop the way logistics spaces are planned and financed in the case of a Landlord port. This configuration corresponds to the most common mode of governance for international gateways.

PAs are organized as public corporations that are accountable to a local (municipal or regional) government (for instance: the case of Antwerp in Belgium or Los-Angeles and Long Beach in United States) or a national government (the case of Marseille and Le Havre in France). They produce and maintain port spaces with an overarching goal of financial self-sustainability. Besides public subsidies, most of their revenues come from port dues and land leases. Thus, PAs aim to increase port traffic, in order to collect port dues, which requires financing and providing plots for efficient terminals, and land uses, in order to increase their land revenues, which requires financing and providing attractive plots to rent for industrial sites.

In a Landlord port, the PA owns the land and the collective infrastructure (docks, roads, railways and waterways). It manages nautical issues, land use planning and the promotion of the port in general. The role of the PA is thus mainly focused on the production of different kinds of port spaces. It finances, produces and maintains the infrastructures mentioned above. Moreover, it acts as a public land developer in order to attract two kinds of companies in the port domain. On the one hand, it produces terminal land (docks, wharfs and quay walls). Private terminal operators rent the land through long-term administrative leases and finance the construction of the terminals they operate. On the other hand, in the rest of the port domain, it also produces the land plots that enable logistics or manufacturing firms to establish industrial sites in the port domain. These firms contract administrative leases for the lots where they construct and own their DCs, parcel service facilities or factories. In the context of the financialization of logistics real estate, numerous property investors lease land in seaport domains in order to build DCs to rent. In this way, even with the growing importance of property investors, the land remains under public ownership and control.

The planning of the different facilities and plots are negotiated with local or national governments according to the local and national institutional configurations. The autonomy and the financial responsibility of PAs are important in this domain. Besides, the major strategic projects, such as port extensions or the construction of new transport ways toward the port (railways, waterways or highways) are planned and financed with local or national governments that exert control over PAs, in the framework of transport infrastructure planning policies.

PAs are not only concerned, therefore, with the production of spaces, but also with the flows and logistics services issues. The negotiations on the leases are a tool enabling port authorities to choose companies that will generate traffic or, from a sustainability perspective, greener practices (e.g. river and rail traffic instead of road traffic, slow steaming). PAs usually develop strong relationships with the companies established in the port domain. They can then provide advice and sell business research services. With the companies forming port communities, PAs act as managers of “port clusters”, which are characterized by “various forms of coordination and resource sharing as a consequence” of the cluster of firms in the port (de Langen and Haezendonck, 2012, 638).

The management of international gateways corresponds in this way to a specific mode of governance tied to public infrastructure policies. The public control on the production of logistics spaces and the involvement of PAs in flows and logistics services issues clearly contrast with the privatized production of logistics parks and the indifference of local governments over logistics issues. These two parallel modes of governance of the logistics spaces are thus often at work in cities and regions. The next section analyses the potential articulation of these two modes of governance within emerging planning policies for logistics facilities at the scale of metropolitan areas and regions.

Planning Policies for Logistics Facilities: Between Regional Planning and Local Urban Logistics Innovations

Logistics is now included in most regional and metropolitan agendas, framed mainly as an issue of sustainability. It leads to spatial planning policies for planning facilities and to policies favoring greener city logistics practices.

Regional and Metropolitan Spatial Planning Policies

Regional and metropolitan spatial planning policies have increasingly included logistics spaces issues, often framed as an issue of sustainability. For instance, in the Paris region, regional planning documents aim to limit the logistics sprawl in order to reduce the urbanization of new spaces. The logistics sprawl is also framed as being at odds with the goal of modal shift of freight flows from road to river and rail modes (Raimbault et al., 2019). However, as developed in section (Logistics Zones: The Silent Privatization of Ordinary Logistics Spaces), the governance of logistics zones development remain essentially local in most urban regions. Indeed, despite these ambitions, regional or metropolitan planning policies are rarely legally binding for municipal land use plans. Comparison between US metropolitan areas, the Paris region and the Frankfurt metropolitan area shows that, Paris region planning policies include logistics issues more clearly than in the US metropolitan areas (Raimbault et al., 2019) and that land-use legal restrictions are weaker in the Paris region than in the Frankfurt case. As a consequence, Frankfurt’s planning regulations restrict the space available for logistics, reduce logistics sprawl and the development of huge private logistics parks compared to the Paris region and US metropolitan areas (Barbier et al., 2019).

In this context, seaports, airports and inland ports can be considered as public tools for implementing logistics regional planning. Against the fragmented governance of logistics zones, a planning solution consists of working with the already existing public authorities dealing with logistics spaces such as (air)port authorities. This situation explains the role of river port spaces and institutions in the regional planning policies of the Paris region (Raimbault, 2019). Indeed, the Paris regional master plan aims to concentrate the development of logistics sites around the main rail and river terminals, which corresponds to spaces managed by the river PA. In this way, the public infrastructure mode of governance meets some goals of the regional planning policies.

Eventually, spatial planning policies also aim to preserve spaces for urban logistics sites in dense urban areas close to city-centres.

Local Urban Logistics Innovations

Besides spatial planning, in order to limit “logistics sprawl,” and thus to comply with sustainability goals, some municipalities facilitate the supply of logistics facilities in inner areas. The literature provides several examples, mostly in Paris and Tokyo (Diziain et al., 2012; Raimbault et al., 2019; Dablanc, 2019), of policies developing new formats of urban warehouses in city centers and dense urban places.

In this domain, flagship projects consist of multistory urban buildings, where logistics activities (consolidation centers, rail terminals, small storage facilities) coexist with several types of activities such as housing, offices, sport, and leisure facilities or even datacenters. The purpose of cohabitation is that these activities bear a portion of the cost of building and land, since they are more profitable. This financial equalization strategy would allow logistics to return to dense urban areas at an acceptable cost.

While private real estate investors are involved in the development of suburban logistics zones, they are just emerging in the development of dense urban logistics buildings. The latter are generally developed by public stakeholders backed by municipalities, as the profitability of such real estate operations is still hazardous. The support of public authorities is crucial in these projects, which can slow down the diffusion of this type of logistics format, as it can make investors risk averse. Urban logistics infrastructure has, to this day, relied mostly on the public initiatives of municipalities, which indicates a third mode of governance of logistics spaces.

Conclusions

The chapter underlines three main dimensions of the current systems of production of logistics spaces: the modalities of the financialization of logistics properties, the scope and effectiveness of planning policies for logistics facilities, and the scope of traditional public infrastructure regulations and management regimes. The different combinations between these three main dimensions explain the coexistence of different modes of planning and financing logistics zones, international gateways, and urban logistics infrastructure.

At the metropolitan scale, the result is a “dualization” of the governance of logistics spaces (Heitz, 2019). On the one hand, international gateways and urban logistics facilities are governed thanks to specific public planning and financial tools. On the other hand, suburban logistics zones are prone to strong privatization dynamics and largely remain out of the scope of regional planning policies and regulations.

These conclusions indicate two main research perspectives. First, the analysis of the coexistence of different modes of governance should be developed through new systemic approaches to the coevolution of the different spaces and modes of governance of logistics dynamics, for instance at the scale of urban regions (Raimbault, 2019). Second, as this chapter is mainly built on European and North American case studies, these findings should be challenged by future international comparisons with Asian, Latin American, Middle Eastern, and African case studies.

Biography

Nicolas Raimbault is an Assistant Professor at the University of Nantes, researcher at ESO (Espaces et Sociétés, UMR CNRS 6590, France) and research fellow at Luxembourg Institute of Socio-Economic Research (LISER, Luxembourg). He teaches urban geography, urban planning, and urban policies and his research deals with the urban and social impacts of logistics activities development. His current work investigates the transformations of blue-collar places and their governance in the dual context of the rise of logistics blue-collar jobs and the fall of manufacturing jobs in European and North American urban regions. He completed his PhD in urban studies in 2014 at Paris-Est University. His PhD thesis studies the governance of logistics development in the Greater Paris Region and in the inland corridor of the port of Rotterdam in the Netherlands. He also received a Master's degree in regional and urban affairs from Sciences-Po Paris.

References

- Baird, A.J., 2006. Port Privatisation in the United Kingdom. In: Brooks, M.R., Cullinane, K.P.B. (Eds.), *Devolution Port Governance and Port Performance*, vol. 17. Research in Transportation Economics, pp. 55–84.
- Barbier, C., Cuny, C., Raimbault, N., 2019. The production of logistics places in France and Germany: a comparison between Paris, Frankfurt-am-Main and Kassel. *WOLG* 13 (1), 30–46.
- Brooks, M.R., Pallis, A.A., 2012. Port governance. In: Talley, W. K. (Ed.), *The Blackwell Companion to Maritime Economics*, Wiley-Blackwell, Hoboken, NJ, pp. 491–516.
- Browne, M., Behrends, S., Woxenius, J., Giuliano, G., Holguín-Veras, J. (Eds.), 2019. *Urban Logistics*, Kogan Page, London.
- Cidell, J., 2011. Distribution centers among the rooftops: the global logistics network meets the suburban spatial imaginary. *IJURR* 35 (4), 832–851.
- Cidell, J., 2015. Distribution centers as distributed places. In: Birtchell, T., Savitzky, S., Urry, J. (Eds.), *Cargomobilities: Moving Materials in a Global Age*, Routledge, New York, pp. 17–34.
- Coe, N.M., 2014. Missing links: logistics, governance and upgrading in a shifting global economy. *Rev. Internat. Polit. Econ.* 21 (1), 224–256.
- Dablanc, L., 2019. City logistics. In: Richardson, D., Castree, N., Goodchild, M.F., Kobayashi, A., Liu, W., Marston, R.A. (Eds.), *International Encyclopedia of Geography*, Wiley-Blackwell, New-York, NY. doi: 10.1002/9781118786352.wbieg0137.pub2.
- Dablanc, L., Ross, C., 2012. Atlanta: a mega logistics center in the Piedmont Atlantic Megaregion (PAM). *J. Trans. Geo.* 24, 432–442.
- de Langen, P.W., Haezendonck, E., 2012. Ports as clusters of economic activity. In: Talley, W. K. (Ed.), *The Blackwell Companion to Maritime Economics*, Wiley-Blackwell, Hoboken, NJ, pp. 638–655.
- Dizain, D., Ripert, C., Dablanc, L., 2012. How can we bring logistics back into cities? The case of Paris metropolitan area. *Proc.-Soc. Behav. Sci.* 39, 267–281.
- Hall, P.V., Hesse, M. (Eds.), 2013. *Cities Regions and Flows*, Routledge, Abingdon.
- Heitz, A., 2019. Planning urban freight and logistics: duality in the logistics real estate market, the case of the Paris Metropolitan area. In: 15th World Conference on Transportation Research, 26–31 May, Mumbai, India.
- Hesse, M., 2008. *The City as a Terminal. Logistics and Freight Distribution in an Urban Context*. Ashgate Publishing, Farnham.
- Notteboom, T.E., Rodrigue, J.-P., 2005. Port regionalization: towards a new phase in port development. *Marit. Policy Manag.* 32 (3), 297–313.
- Rodrigue, J.-P., Comtois, C., Slack, B., 2017. *The Geography of Transport Systems*. Routledge, Abingdon.
- Rodrigue, J.-P., Debie, J., Frémont, A., Gouvenal, E., 2010. Functions and actors of inland ports: European and North American dynamics. *J. Trans. Geo.* 18 (4), 519–529.
- Raimbault, N., 2019. From regional planning to port regionalization and urban logistics. The inland port and the governance of logistics development in the Paris region. *J. Trans. Geo.* 78, 205–213.
- Raimbault, N., Heitz, A., Dablanc, L., 2019. Urban planning policies for logistics facilities. a comparison between US metropolitan areas and the Paris region. In: Browne, M., Behrends, S., Woxenius, J., Giuliano, G., Holguín-Veras, J. (Eds.), *Urban Logistics*, Kogan Page, London, pp. 82–108.
- Witte, P., Wiegman, B., Ng, A.K., 2019. A critical review on the evolution and development of inland port research. *J. Trans. Geo.* 74, 53–61.
- World Bank, 2007. Port Reform Toolkit, Module 3: Alternative Port Management Structures and Ownership Models. Available from: http://rru.worldbank.org/Documents/Toolkits/ports_fulltoolkit.pdf.

Supply Chain Risk Management: Creating the Resilient Supply Chain

Richard Wilding, Centre for Logistics, Procurement and Supply Chain Management, Cranfield University, Cranfield, Bedfordshire, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	41
Categorizing Supply Chain Risk	42
Constructing Resilient Supply Chains	42
Supply Chain Design	43
Engendering Supply Chain Agility	43
Establishing Collaborative Supply Chain Relationships	44
Fostering a Culture of Supply Chain Risk Reduction	44
Conclusion	45
Acknowledgment	45
Appendix A: Case Study	45
Further Reading	46

Introduction

Supply chains involve supply chain risk: every supply chain professional knows that. And as global cross-border trade continues to increase, and supply chains lengthen, that risk mounts. Observers point regularly to the threats posed by currency fluctuations, supplier failures, geopolitical uncertainty, infrastructure failings and natural disasters. Time and again, events demonstrate how diverse those risks can be—and how interconnected are the world's factories and supply chains.

Consider the earthquake and ensuing tsunami, which hit North-eastern region of Japan in 2011. Within a matter of hours, large parts of Japan's automotive industry had ground to a halt. Honda, Toyota, Nissan, and Subaru all had plants in or close to the affected zone, and were forced to close them. And soon, other Japanese automotive plants—in many cases far less directly affected by either the earthquake or the tsunami—were also forced to stop production, because of parts shortages caused by postdisaster logistics difficulties, or by earthquake or tsunami damage to tier-1, tier-2, or tier-3 suppliers' plants.

Toyota, Suzuki, and Nissan ceased production completely. Even Mazda, located far to the south near Hiroshima, closed down—unable to keep its plants running without critical parts sourced from suppliers closer to the damage zone. Take into account the 10 days of unaffected production prior to March 11, when the earthquake struck, and postearthquake volumes for the rest of March were close to zero.

Inevitably, the disruption soon spread offshore, affecting manufacturers as diverse as Ford, Volvo, GM, Renault, Chrysler, and PSA Peugeot–Citroën. Across Asia, North America and Europe, crippled assembly plants either shut, or scrambled to build whatever vehicles they could, with the components, materials, and paint colors that were available.

Simply put, the world's automotive industry found that huge numbers of components turned out to be sourced from Japan and its intricate supply chains. What made matters worse was that in many cases, many of these components were single-sourced—and single-sourced not just to a single company, but to a single manufacturing plant within that company. A widely used metallic paint pigment, called Xirallic, had been produced at only one factory in the world, located in a town close to Fukushima.

It might be thought easy to dismiss this supply chain disruption as a “one off;” an unfortunate outcome from a natural disaster that is extreme by any standard. Except, of course, for the fact that a little thought reveals plenty more examples from the last few years.

The 2010 eruption of Iceland's Eyjafjallajökull volcano, for instance, filled European air space with high-altitude ash clouds, and led to the closure of much of Europe's air space, and resulted in the cancellation of some 95,000 flights, causing significant disruption to air freight schedules. Again, the automotive industry's lean supply chains were quickly affected, with production halted at two Ford plants in Michigan as well as at BMW's plant at Spartanburg, South Carolina.

Or the flooding in Thailand in October 2011, which submerged seven of the country's largest industrial zones, causing national manufacturing output to fall by 36%. In the country's Khlong Luang industrial zone, factory floors lay under up to 2 m of polluted water for several weeks and, of the 227 factories operating in the zone, a mere 15% had restarted production six months later. Two of the world's largest manufacturers of computer hard drives—Seagate and Western Digital—had factories in Thailand, with Western Digital reporting that it manufactured 60% of its hard drives in the country.

In short, in each case, events on the other side of the world had supply chain impacts on businesses and end-consumers thousands of miles away. Contingency plans had been of little use, the events in question could not have been predicted, and in many cases the affected businesses did not know the extent of their vulnerability until it was too late.

What can businesses do to improve their resilience to such supply chain disruption? How can they reduce their risk profile? What changes do they need to make to their organizations, and their supply chains? This chapter explores the answers to these questions.

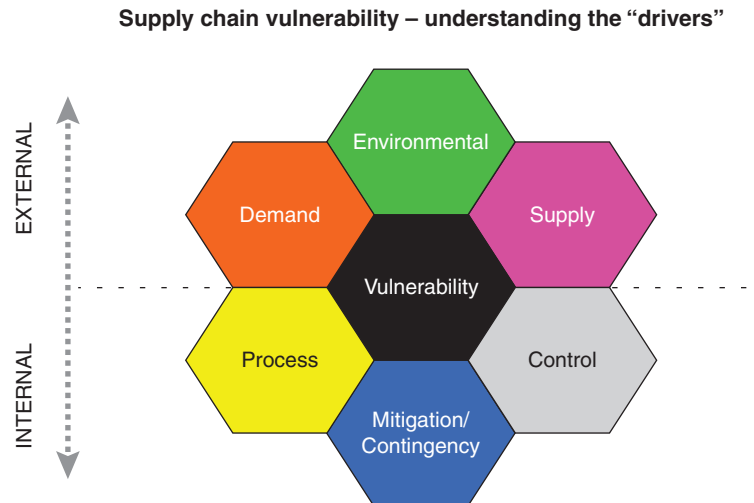


Figure 1 The drivers of supply chain risk.

Categorizing Supply Chain Risk

It is possible to think of supply chain vulnerability to risks as being split between external and internal “drivers,” as shown in **Fig. 1**. Importantly, risks will always be context-specific and because every supply chain is different, then risks and their relative weightings will always be different.

External risks can be summarized as follows:

- **Demand risk**, which relates to potential or actual disturbances to the flow of products, information and cash, between the focal firm and its market. Cash resources are rarely considered, but in economic downturns can prove to have significant impact on the operating capability of organizations.
- **Supply risk** is the upstream equivalent of demand risk; it relates to potential or actual disturbances to the flow of product or information upstream of the focal firm. In a similar way to demand risk, the disruption of key resources coming into the organization can have a significant impact on the organization’s ability to perform.
- **Environmental risk** is the risk associated with external and, from the firm’s perspective, uncontrollable events. The risks can impact the firm directly or through its suppliers and customers. Environmental risk is broader than just natural events like earthquakes or storms. It also includes, for example, regulatory changes such as alterations to legislation or customs procedures.

Internal risks relate to both how the firm addresses the external risks that it faces, and its competence in planning and executing its own business:

- **Processes** are the sequences of value-adding and managerial activities undertaken by the firm. Process risk relates to disruptions to key business processes that enable the organization to operate. Some processes are key to maintaining the organization’s competitive advantage, while others may underpin the organization’s activities. It is important to recognize and classify the importance of the various processes to enable effective risk management strategies to be implemented.
- **Controls** are the assumptions, rules, systems, and procedures that govern how an organization exerts control over the processes and resources. In terms of the supply chain these can include such things as order quantities, batch sizes, and safety stock policies, plus the policies and procedures that govern asset and transportation management. Control risk is therefore the risk arising from the application or misapplication of these rules.
- **Mitigation** is a hedge against risk built into the operations themselves and, therefore, a lack of mitigation is a risk in itself. It is important to consider mitigation during the supply chain design process, as if this is not done, then the risk profile can be increased. Allied to mitigation is the notion of *contingency*, which is the existence of a prepared plan for dealing with risk, and the identification of resources that can be mobilized in the event of a risk materializing. This requires every stakeholder in the supply chain to understand which resources should be mobilized, and the correct procedures for doing this.

Constructing Resilient Supply Chains

When building a resilient supply chain, four key things need to be considered:

- The supply chain design
- How supply chain agility can be engendered

- How collaborative supply chain relationships can be established
- How a culture of supply chain risk reduction can be fostered

Let us now address each of these areas in turn.

Supply Chain Design

Fairly obviously, supply chains do not operate in a vacuum. In particular, when considering the design of the supply chain, it is important to maintain a clear linkage to the corporate and competitive strategies of the business. In other words, if the overarching strategy of the business is to compete on, say, price, or service, or the ability to nimbly respond to changing market tastes, then in each case the supply chain must be configured to suit—in terms of its processes, in terms of its organization, in terms of its infrastructure, and in terms of the systems that operate it. Fig. 2 depicts this.

Therefore, the first consideration in creating resilient supply chains is ensuring a clear strategic alignment between the organization's competitive strategy, and the supply chain. Not doing so risks either having a supply chain that is not synchronized with the goals of the business, or—importantly from a vulnerability point of view—a supply chain that is riskier than it needs to be, in order to support the goals of the business. An example of the latter would be a long supply chain stretching back to factories in China, incurring transportation-related risks, on the mistaken assumption that a low product cost was of overarching importance. Dell, Apple, and fashion firm Zara all stand out as businesses with supply chains perfectly aligned to their own—very different—business strategies.

Engendering Supply Chain Agility

To reduce the overall risk of a supply chain, supply chain agility is critical. Again, supply chain agility needs to be reflected in the supply chain's processes, infrastructure, information systems, and organization. Agile supply chains have a number of distinguishing characteristics. They not only need to be network-based, but they also need to be market-sensitive, with highly integrated virtual and critical processes. Agile supply chains need to synchronize both supply and demand, in order to respond quickly to both volume and variety changes. In addition, agile supply chains need to be able to adjust output levels quickly to match market demand, and to be able to switch rapidly from one variant to another.

Central to the achievement of agility is an understanding of what has been termed the “three Ts of highly effective supply chains”—*time, transparency, and trust*. Each is interrelated with, and dependent upon, the others. Understanding the **time** dimension of the supply chain enables organizations to gain transparency into what is happening within the supply chain; when supply chain participants know what is happening, confidence builds due to the resulting **transparency** and **trust** that then develops between all the players in the supply chain. Moreover, a virtual circle develops; trust engenders a greater understanding of the time dimension, and hence greater levels of transparency and then trust arise.

Where to start? It is often easier to begin by simply mapping the supply chain with respect to time. Once the time dimension is understood, then it is relatively straightforward to plot relationships such as demand with respect to time, processes with respect to time, or inventory holdings with respect to time. This then helps to both generate transparency and build an understanding of the key issues involved in the supply chain, further cementing trust.

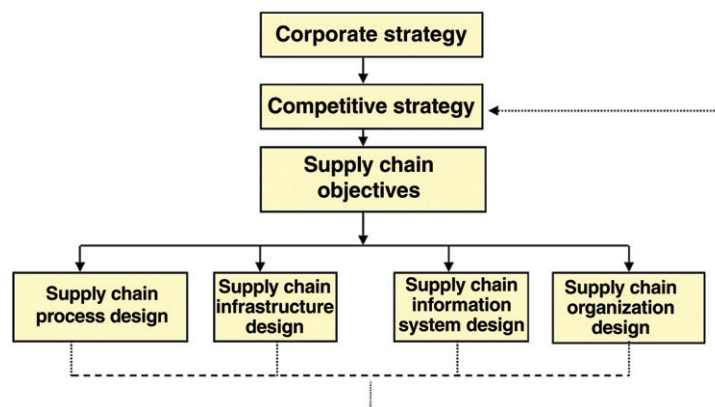


Figure 2 Effective supply chain design.

Establishing Collaborative Supply Chain Relationships

To create a truly resilient supply chain structure, relationships are key. As Martin Christopher has noted, supply chains—and hence supply chain management—can be defined in terms of relationships: *“Supply chain management can be defined as the management of upstream and downstream relationships with suppliers, distributors, and customers to achieve greater customer value-added at less total cost.”*

Defined in this way, it becomes abundantly clear that these relationships must be actively managed, if the overall supply chain is to operate effectively. Organizations often underestimate the importance of dedicating resources to the management of relationships, especially in the upstream supply chain. Nevertheless, in real-life case studies of supply chain resilience, the quality of supply chain relationships, and companies’ investment in the requisite resources, often stand out as important contributory factors.

That said, businesses cannot build close collaborative relationships with *every* link in the upstream and downstream supply chain—and certainly not all at once. Some degree of prioritization is necessary. While individual circumstances will differ, logic suggests that the closest links should be forged with a combination of the largest suppliers, customers and partners, as well as those suggested by appropriate risk and resilience audits, even if the trading relationship is not especially significant in monetary terms. A case in point: supply chain risk buried within its tier-2, tier-3, and tier-n components.

After 2011 Japanese earthquake and tsunami, for instance, car manufacturer BMW’s first reaction to the disaster was to assume that it would be unaffected, as it had no suppliers in Japan. But it soon discovered that some critical tier-4 semiconductors came from North-eastern region of Japan. And on probing more deeply, the company then discovered several more instances of supply chains terminating in North-eastern Japan—in some cases, to factories that were the sole source of the components and materials in question.

Renesas Electronics, for example, the world’s largest manufacturer of automotive microcontrollers, was forced to cut production by 70%, with production shifted from its badly-damaged Naka wafer fabrication factory in the Japanese coastal city of Hitachinaka to a number of other sites around the company. Altogether, output was suspended at seven of the company’s 22 plants, with the worst affected remaining off-line until July—four months after the disaster.

Just as businesses cannot build close collaborative relationships with every link in the upstream and downstream supply chain, neither can businesses build equally deep collaborative relationships with every supplier. Inevitably, some relationships will be—and will need to be—deeper than others. In a spectrum of possible relationship types stretching from “arm’s length partnerships” through to “virtual vertical integration,” some selectivity is required.

Again, relative exposure to supply chain risk acts as a useful lodestar. Often, when organizations fully recognize the resource implications of managing highly collaborative relationships, it is not uncommon for them to take a more pragmatic view of relationship building. A natural priority: begin by identifying those relationships that could put the supply chain at risk, and invest scarce resources there first.

Finally, it can be useful to consider relationships with competitors. Collaboration between competitors is not so bizarre a notion as one might expect: clearly, those businesses with supply chain challenges and requirements that will be closest to a given business’s *own* supply chain challenges and requirements will generally be its competitors.

Often, the goal is cost reduction—or, increasingly, benefits allied to cost reduction, such as carbon reduction or other societal benefits. But there can be a supply chain risk dimension, as well. Automotive manufacturers PSA Peugeot Citroën and Toyota, for instance, shared component development across their Peugeot 107, Toyota Aygo and Citroën C1 models, potentially creating dual-sourcing possibilities.

Logistics possibilities exist, too. In the motor industry, for instance, tire, battery, and exhaust system manufacturers and distributors must deliver to the same dealerships and aftermarket retail outlets. In the grocery industry, grocery manufacturers must deliver to the same supermarket regional distribution centers, wholesalers, and retail outlets. In such circumstances, collaboration—sometimes dubbed “coopetition”—can make a lot of sense. To the extent that such arrangements provide access to a “backup” delivery option, or a larger pool of delivery resources, or make it possible to leverage the increased volumes and select a higher-quality and more resilient transport provider, then an element of supply chain de-risking clearly exists.

Fostering a Culture of Supply Chain Risk Reduction

Culture has been defined as “what we do or how we act, in the absence of instructions.” For businesses to develop maximally resilient supply chains, it is important that an effective culture of supply chain risk is fostered. Only in that way will supply chain risk be front-of-mind each time a decision relating to the supply chain is made. In other words, rather than supply chain risk—and supply chain risk reduction—being considered separately from other supply chain-specific considerations, the supply chain risk dimension is considered alongside, and as a part of, any ongoing supply chain deliberations.

This is not necessarily easy or straightforward. Supply chain strategies are generally couched in terms of issues such as cost-to-serve, price, responsiveness, service levels and inventory holdings. When managers are targeted and incentivized upon such things, their goals are very explicit, and an implicit impact on supply chain risk may not always be welcome. Nevertheless, if managements are serious about supply chain risk reduction, it is necessary to constantly ask questions such as:

- How will this action impact on the risk profile of the supply chain?
- Will this action make us more vulnerable to disruption?
- Will this action improve or detract from our resilience?

In moving to this way of thinking, it is a blunt fact that risk—and risk reduction—are often very intangible notions when set against such things as cost, price, and inventory levels. In terms of cost, for instance, the impact of sourcing a particular component from (say) China can be very clear. Far less clear will be the impact on supply chain risk. Nevertheless, a serious commitment to supply chain risk reduction will require managements to give greater priority to risk reduction than they may do at the moment.

A serious commitment to supply chain risk reduction will also involve the organization at times transcending long-standing organizational boundaries. Proper consideration of supply chain risk rarely takes place within tight organizational limits: some element of cross-functional collaboration is usually required, albeit on an *ad hoc* basis. This is particularly true of the “supply chain continuity” or “supply chain crisis management” teams that can be created, so as to be quickly brought to bear on issues and problems as and when they arise.

During the UK’s September 2000 fuel crisis, for instance, which was brought about by striking tanker drivers, it quickly became apparent that food supplies were threatened and the military was partly mobilized to help in the distribution of fuel, so as to alleviate this. Quite simply, supermarkets were unable to deliver to their stores, either because products had not reached their distribution depots, or because they lacked the fuel to undertake the outbound distribution leg to the store. But some businesses fared better than others: supermarket chain Sainsbury, for instance, was able to activate a team of supply chain contingency planning experts, helping it to make day-to-day decisions that delivered a competitor-beating performance.

Finally, building a culture of supply chain resilience may involve overturning long-standing traditions and existing cultural norms. The supply chain disruption caused by the Japanese earthquake and tsunami of 2011, for instance, resulted in considerable introspection among automotive manufacturers, both within Japan and overseas. Long an advocate of single-sourcing, Nissan began mandating that suppliers’ business continuity plans should include an alternate sourcing capability—in other words, the ability to seamlessly switch output to an alternative facility in the event of disruption at the primary facility. Toyota and BMW, for their part, reckoned to have learned important lessons about supplier transparency and openness in terms of supply chain mapping and sourcing visibility—including, in Toyota’s case, lessons regarding the advisability of modifying its product development processes so as to make it easier to use alternative parts.

Conclusion

For the most part, supply chain disruption happens with little warning. Even extreme weather—think 2017 Hurricane Harvey, for instance, which was the second-costliest tropical cyclone on record, inflicting a reported \$ 125 billion in damage—provided only a few days’ notice of potential disruption. So, for a resilient supply chain, forward planning is a key. While planning and preparation for a specific event is rarely possible, planning for disruption in general is certainly possible, as is planning for specific categories of events. Virtually every large retailer, for instance, will have a “playbook” for how to deal with product recalls or the collapse of a major supplier. In short, the time spent by organizations thinking through the options and understanding the risk profile of the supply chain is an important investment that will protect the organization’s performance.

That said, it is important to realize that such contingency planning is planning in the context of today’s organization: in other words, how should *today’s* organization respond to disruption, should it occur? But detailed planning—however extensive—has natural limits, and can only take an organization so far. “*No plan survives contact with the enemy*,” is how nineteenth-century Prussian military commander Helmuth van Moltke described these limits. “*Everyone has a plan—until they get punched in the face*,” said boxing champion Mike Tyson, more prosaically. Real resilience demands something else.

In short, as this chapter has shown, it is better to create resilience by reshaping the organization’s response to disruption, so that it is inherently better able to react and recover. With an organization that is equipped with the right culture, the right chain relationships, the right level of agility, and the right supply chain design, that is eminently possible.

Acknowledgment

I would like to thank my colleague Dr. Malcolm Wheatley, Visiting Fellow, Centre for Logistics, Procurement and Supply Chain Management for researching a number of the examples and cases utilized in this chapter.

Appendix A: Case Study

On Monday 3 February 1997, Toyota announced that all its Japanese assembly lines had been brought to a halt. The problem: a devastating fire the previous Saturday, at a plant belonging to one of its suppliers, Aisin Seiki. The company was Toyota’s only supplier of brake fluid proportioning valves (P-valves), which were used right across the manufacturer’s model range. With just half a day’s stock in hand, it was just a matter of hours before Toyota’s assembly lines ground to a halt—closely followed by those of the rest of its supplier base. Gloomy prognostications predicted that production would be halted for weeks; consumed within the blaze were the transfer lines used to manufacture the P-valves, along with other special-purpose machinery and drills.

Instead, P-valve deliveries to Toyota resumed just three days later, rolling off temporary lines hastily set up by a tiny second-tier Aisin supplier, Koritsu Sangyo. By the following Friday, production was back up to 90% of normal levels, reaching 100% on Monday

10 February, exactly a week after the fire. In all, some 62 firms found themselves manufacturing P-valves: 22 of Aisin Seiki's own suppliers, Toyota itself, 36 Toyota suppliers, and a number of independent businesses—including a sewing machine manufacturer, Brother Industries, which had never previously manufactured vehicle components. Also involved were around 150 other businesses, including 70 machine tool manufacturers, drill manufacturers and various suppliers of fixtures and gauges.

Impressively, the whole thing was coordinated by a four-team "emergency response unit" communicating with suppliers by fax machine and phone, distributing engineering drawings and specifications and liaising with materials suppliers. After the event, recognizing their achievement, Aisin Seiki published a booklet containing the lessons learned, making it available to its supplier network and all the businesses that had participated in the recovery effort.

Further Reading

CERES Cranfield, 2003a. Creating Resilient Supply Chains: A Practical Guide. Available from: <http://hdl.handle.net/1826/4374>.

CERES Cranfield, 2003b. Understanding Supply Chain Risk: A Self-Assessment Workbook. Available from: <http://hdl.handle.net/1826/4373>.

Wilding, R. 2013. Supply chain temple of resilience. *Logist. Transp. Focus*. 15 (11), 54-59. ISSN: 1466-836X. Available from: <http://dspace.lib.cranfield.ac.uk/handle/1826/8308>.

Transportation Safety and Security

Maria G. Burns, College of Technology, Houston, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Defining Safety and Security	47
The Human Factor in Security and Safety	48
Risks and Vulnerabilities	48
Trends in the 21st Century	50
Conclusions	51
Acknowledgment	52
Biography	52
References	52
Further Reading	52

Defining Safety and Security

Transportation professionals measure their performance according to: (1) business development, that is, commercial success, and (2) ensuring safe and secure operations, that is, the ability to prevent, eliminate, and recover from safety and security risks. While these terms are used interchangeably, in fact safety differs from security.

Transportation safety risks pertain to a company's ability to protect its operations from nonintentional, yet harmful elements, such as:

1. natural disasters or weather conditions, for example, hurricanes, flooding, earthquakes;
2. machinery breakdown, operational malfunction, technological failures, loss of structural integrity, construction vulnerabilities, etc.
3. human factors, that is, human error, fatigue, negligence, lack of situational awareness, unintentional damage, etc.

Transportation security involves risks of deliberate attacks aiming to compromise transport networks. Security risks are driven either by terrorist motives (i.e., nationalist, religious, or sociopolitical) or by financial motives. Security threats include terrorist attacks, vehicle hijacking or sea piracy, cargo theft, cyber warfare, money laundering, contraband, human or narcotic trafficking, and others (Burns, 2015).

While security attacks entail intentional damage, they are further categorized based on the nature of the motive:

1. Terrorism-based attacks are triggered by ideology based on religious, political, or social differences. Terrorism aims toward the destruction of life, property and the environment, aiming to instigate fear and social turmoil.
2. On the other hand, criminal-based attacks are driven by financial motives and the goal of profiting through illegitimate activities. Hence, criminals do not destroy property, but steal, hijack, or use transport for illegitimate activities such as contraband, money laundering, human trafficking, drugs and weapons trafficking, among others (Fig. 1).

Over the past years security investigations have associated terrorists with criminal networks, as part of larger, more intricate supply chains. This means that the funds generated from criminal activities, often end up being used by terrorist groups. For example, connections have been verified between religious fanatic groups and sea piracy, gun smuggling, and money laundry activities. Cyber security incidents have also revealed associations between hackers requesting ransom fees and the financing of terrorist groups.

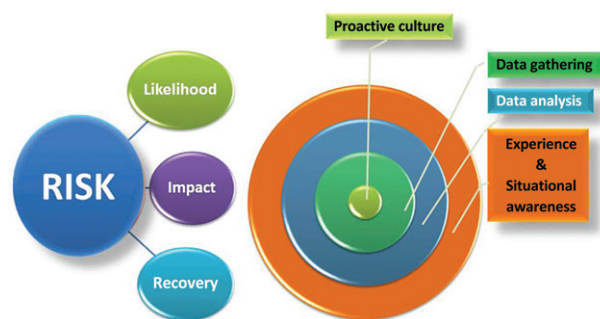


Figure 1 Risk mitigation. Source: Burns. M.G.

The Human Factor in Security and Safety

We live in an era of unprecedented technological innovations, many of which were designed to strengthen corporate safety and security systems. Despite the high-tech achievements, it is still the human factor which plays a pivotal role in attaining a successful safety and security strategy (Burns, 2014).

The human factor plays a significant role in security and safety: Namely, there are two distinctions in security risks: (1) humans are motivated by ideology, hence commit acts of terrorism, or (2) humans are motivated by financial or commercial gains, and commit criminal acts such as money laundering, contraband, human trafficking, smuggling of weapons or narcotics, and others.

In the case of safety, while unintentional, the human factor still plays a critical role: human-driven factors are typically focused on negligence (including poor operational handling or inadequate maintenance leading to wear and tear); lack of situational awareness (such as failure to identify or act upon signs of an existing or potential safety threat); but also fatigue, lack of training, lack of familiarization with the facility or equipment, and so on.

The degree of familiarization with the premises, the technologies, the processes is another important factor. Surveillance cameras may be installed or cargo scanners may be used. However, it is the human element, that is, the operators of the respective equipment which can make a difference. The more familiarized company employees are, the stronger their defense is against safety and security threats. This is the reason that pertinent regulations request employees' familiarization as a prerequisite.

Risks and Vulnerabilities

Although incidents may occur unintentionally in the case of safety, and intentionally in the case of security, what is common is the fact that external risks manifest themselves in the presence of internal systemic vulnerabilities. In the case of safety, a risk will coincidentally be triggered due to a weakness or exposure. In the case of security, the perpetrators will intentionally and meticulously search for a systemic vulnerability and will plan their attack with the particular vulnerabilities in mind. What makes security risks more impactful, is the intentional attack of a systemic vulnerability (Burns, 2019b). Consequently, both safety and security risks cover the same physical, cyber, and supply chain realm, with the only distinction being the intentional damage in security versus the unintentional damage in safety.

Supply chain activities and operations are categorized according to their risk factor, and the domain they encompass is distinguished into:

1. Physical, that is, focusing on the tangible assets, commodities, and aspects of transport.
2. Virtual, cyber, that is, focusing on the intangible assets, that is, data available on IT, cyber and cloud-based platforms.
3. Holistic, including the entire supply chain with all its transportation networks, and various companies.

For modern transportation companies to safeguard their systems and enjoy professional success, they must find the perfect balance between operational efficiency, and an impeccable safety and security track record.

The likelihood of increased safety and security risks depends on different factors. Safety risks increase as companies' resources diminish, and limited budget is available for repairs, maintenance and investment in the physical and cyber protection of their supply chains. On the other hand, security risks pertain to targeted attacks by entities with financial motives (such as cargo theft) or ideological motives (aiming to destroy a particular target). Hence, security threats are initiated by entities targeting healthy, prosperous, and critical segments of a nation's infrastructure (Fig. 2).

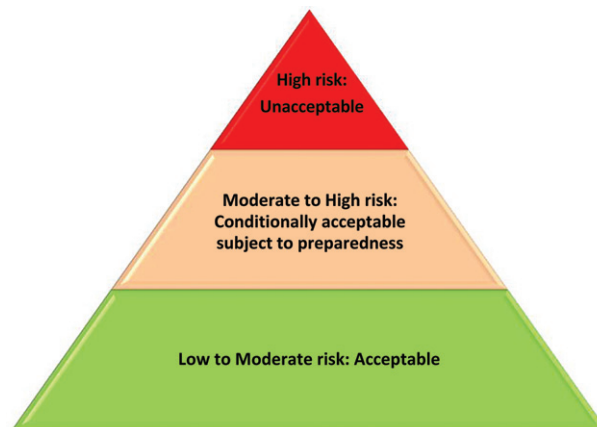


Figure 2 The levels of risk tolerance.



Figure 3 The stages of supply chain.

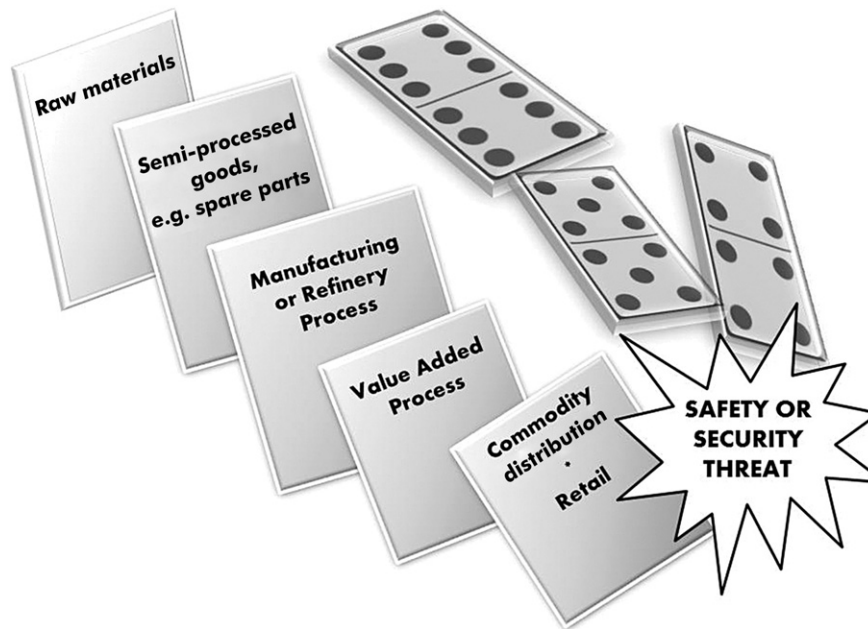


Figure 4 The Domino effect, or how a single threat can impact the entire supply chain. *Source: Burns. M.G.*

Transportation is driven by the laws of derived demand, which suggest that as passengers and cargoes increase, so does the demand for specific transport modes and geographical trade routes. Also, the growth of global population and the economy will urge the transportation industry to grow accordingly.

Transportation networks are components of intricate supply chains, where transport is the glue, or catalyst for commodities to move from one stage of the supply chain to the next (Burns, 2018a, 2019a). Modern transportation networks are highly complex and interactive, which suggests that when a single company takes risks, its supply chain partners are also influenced. This is a characteristic of the trade and transport industry: it is considered as the norm when complex networks function in an effective and efficient manner. However, any delay or malfunction becomes instantly noticed.

As seen in Fig. 3, each stage of the supply chain is interrelated. Hence, when a safety or security threat impacts a single stage of the network, all others are affected through the domino effect. To eliminate safety and security risks, transportation stakeholders and their supply chain partners must ensure supply chain visibility. The interrelation between the supply chain and the domino effect are presented in Figs. 3 and 4, respectively.

The transportation industry is a high-risk, high-profit realm, suggesting that companies often striving for higher risk endeavors, are likely to maximize business development, commercial and financial gains. This is a highly competitive, ever evolving industry, where companies establish supply chain networks with the purpose of taking risks that are not only acceptable, but also rewarding. Fig. 2 demonstrates the levels of risk tolerance, with the high risk area reflecting safety and security challenges, among other types of risks.

As a transportation network grows, risks may grow accordingly. As global population grows, the demand for trade and transport also grows, production grows, and both physical and cyber networks grow. On an annual basis, the volume of global trade exceeds 10.3 billion tons of cargo, with its value exceeding \$17 trillion. By 2030, the global trade volume is anticipated to reach 20 billion tons of cargo. In the United States alone, cargo exceeds 2 million metric tons, among which, there are 15 million TEUs (twenty feet equivalent units) of containers. According to the FBI, annual cargo theft statistics become alarming, ranging from \$15 to \$30 billion each year.

As defined by the FBI, cargo theft pertains to “the criminal taking of any cargo including, but not limited to, goods, chattels, money, or baggage that constitutes, in whole or in part, a commercial shipment of freight moving in commerce, from any pipeline system, railroad car, motor truck, or other vehicle, or from any tank or storage facility, station house, platform, or depot, or from any vessel or wharf, or from any aircraft, air terminal, airport, aircraft terminal, or air navigation facility, or from any intermodal

container, intermodal chassis, trailer, container freight station, warehouse, freight distribution facility, or freight consolidation facility” (FBI, 2015).

To mitigate this threat, the FBI’s Uniform Crime Reporting (UCR) Program was established with the purpose of gathering and analyzing cargo theft information to notify law enforcement and other federal and state agencies, as well as legislators, universities and the general public. The data is gathered and duly analyzed in order to promote general awareness of the growing size and consequences of cargo theft upon our national security, economy, and other risks (FBI, 2016; Burns, 2018b).

When it comes to the transportation mode selection, it is worth noting that maritime transportation is responsible for about 90% of global trade. This determines the transportation risks, considering that each transportation mode has its own distinct safety and security risks. In general, geography and route selection will determine weather conditions, but also the risks of piracy or terrorist attack. On top of the transport mode selection, the selection of supply chain partners will determine the type, likelihood and impact of safety and security risks.

An optimum transportation network is characterized by the uninterrupted flows of passengers and cargoes. When delays and bottlenecks are present, safety and security risks are likely to be present as well. For example, safety issues are manifested through mechanical or operational failures. Most important, in the case of security attacks, perpetrators intentionally select systems with high traffic, bottlenecks, and systemic delays. There is a close relation among delays and border security: as global volumes of trade and transport grow each year, the incidents of safety and security also grow.

Transportation is the glue which connects all the components of global and domestic supply chains. Therefore, any safety and security disruptions trigger the domino effect or ripple effect spreading throughout the entire supply chain. Hence, a single vulnerability point has the potential to damage the operational and financial performance of several partners within a supply chain. Cargo theft, contamination, damage or simply delayed delivery may lead to losses, product recalls, or cancellations and reorders. This in turn triggers losses caused by reverse logistics processes, insurance claims, and long-term disruptions in partnerships.

Modern companies adopt a proactive safety and security culture, aiming to control as many factors as possible, and thus eliminate risk. As a principal rule, the closer the cargo is to a company’s premises, the greater the control and protection from risk. While the company’s assets and cargoes are in motion, especially in long-term haulage, the higher the exposure and the lower the control.

In order to assess the true risks of safety and security, it is worth measuring the cost of an incident. In the aftermath of an incident, the company’s brand promise or reputation are tarnished (Burns, 2017). This can be measured in the stock market where the value of shares can drop dramatically. In medium or high-impact incidents, the stock value diminishes by 30%–60% within only hours or days after the incident.

From a safety and security perspective, each supply chain nodal point is the convergent node where illegitimate activities affect legitimate trade and transport, and economic development. Based on the assumption that knowledge is power, visibility helps notify transport networks of a particular threat occurring in a specific nodal point and hence protects the entire supply chain through the transfer of resources and the creation of contingency plans.

While a company’s safety and security managers focus on strengthening the vulnerabilities within their premises, they often neglect to conduct meticulous safety and security background checks on their network partners, contractors and employees. Hence, a frequent outcome suggests that companies are involved in safety and security incidents, not due to their own omissions or weaknesses, but due to their partners or contractors’ shortcomings. After all, it has been established that about half of the intentional systemic damages are caused by insiders. As a conclusion, companies should be aware that we are as weak as our weakest links.

Trends in the 21st Century

Over the past decades, there has been a shift in the safety and security risks from the purely physical realm toward cyber and broader, supply chain risks. In response, companies and governments are increasingly adapting their preparedness and defense strategies accordingly.

The multilayered approach is based on the Swiss cheese theory, demonstrated in Fig. 5, suggesting that an incident occurs when multiple vulnerabilities are aligned, enabling the safety or security risk to manifest. For this reason, modern companies and governments adopt the multilayered approach. As several complimentary measures are taken to safeguard a company’s safety or security, the risks are eliminated. For physical security, the multilayered defense includes surveillance cameras, uniform/armed guards, fences, walls, vehicle barriers, and other protective measures.

The 21st century holds tremendous opportunities for the transportation industry, based on the new technological paths: Artificial Intelligence (AI), Big Data, 5G networks, Blockchain, Internet of Things (IoT), Bullet trains, and the list goes on... It is worth considering that new technologies do not only serve as instruments for safeguarding our systems and leading to prosperity. The benefits that these new technologies offer to legitimate entities, can also be offered to security violators, such as criminal entities and terrorists, to support their own goals. Hence, the job of safety and security professionals is to explore the capabilities of each innovative technology used in their supply chain, and consider how these tools can be used in the hands of security violators.

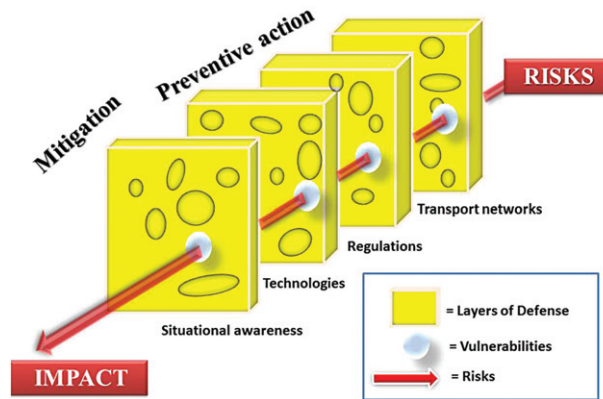


Figure 5 The Swiss cheese model applied in transportation or how safety and security threat penetrate layers of our defense.

One of the many challenges, and opportunities of the transportation industries is to be able to quantify and evaluate the benefits of a proactive safety and security culture. One significant shift in the industry entails the customer-driven business development approaches.

The US Government has developed MSRAM, a security risk analysis tool used to assist in the prioritization and protection of Critical Infrastructure and Key Resources (CIKR). It is aligned with the goals of the US Department of Homeland Security, in alleviating terrorist attacks within the United States; minimizing our nation's vulnerability to terrorism, but also the pertinent consequences, including damages; facilitating the fast recovery, and promoting socioeconomic security and sustainability (Burns, 2014).

Safeguarding the global transportation network from safety and security threats, is not only beneficial to the company, but also to the client: it promotes loyalty, long-term partnerships, and builds the brand name. While this notion was becoming popular in the dawn of the 21st century, the 2008 global economic meltdown led some companies to place low cost as a key driver. Financial restrictions forced many companies to make strategic transport decisions driven more by cost-efficiency factors and less by safety and security considerations.

Every organization, be it public or private, is asked to allocate budget across two main objectives: First, business development for organic and financial growth. Second, risk management including safety and security risks. Funds and resources are allocated at the managerial discretion. The advantage of requesting funds for financial and commercial development purposes is that their performance metrics are easily defined, measured and therefore better justified. On the other hand, safety and security managers may find it hard to negotiate the increased allocation of funds to avert threats which may, or may not occur, as it may not always be feasible to measure unmaterialized losses, or the benefits of risk aversion. It may be hard to measure and quantify the extent of the risk that was averted or eliminated, or prove that the company's preventive action has totally deterred an incident.

Once the brand promise is compromised, business loyalty is the second intangible commodity which suffers tangible financial damage. Order cancellations, loss of clients, credit and payment terms, as well as insurance premium levels are some of the anticipated losses following a major safety and security incident. Companies in general are reluctant to publicize their incidents, even in cases where the safety or security problem is inevitable. Small and medium-sized companies with limited resources and recovery alternatives, are the ones being mostly impacted by safety and security threats. Many of these companies collapse every year, unable to cover the unexpected expenditure.

Companies can reap tremendous benefits by showcasing an excellent track record and their safety and security performance over a number of years. An impeccable record provides tremendous advantages in building a company's reputation in the present tense. However, companies can only benefit from an excellent record of compliance, until the moment a serious accident happens (Burns, 2017). A company may have an impeccable safety and security record for decades, yet this does not presuppose a more favorable public opinion should an incident occur. Companies should remember that fine reputation takes decades to build, hours to destroy, and forever to recover.

Conclusions

This article has established a theoretical and practical framework that transportation professionals can use when working on safety and security risk factors. The process of safety and security is multi-dimensional and proper risk management requires contingency planning with the consideration of selecting alternative trade routes, partners, technologies, markets, and overall contingent plans. Promoting a culture of safety and security requires ongoing, dedicated efforts and a sincere alignment with the company's safety and security mission and strategic goals. The company's leadership must be dedicated to building and nourishing this culture. All company employees and partners must be responsible, alert, trained, and accountable when implementing the company's safety and security strategy.

Acknowledgment

The author wishes to acknowledge the support and collaboration of the editors and publishers of the Encyclopedia, for making this work possible.

Biography



Prof. Maria G. Burns serves as the Director for the Logistics and Transportation Policy Program, University of Houston and as SCLT Faculty and Researcher for the US Department of Homeland Security. She is a National Academies scholar, serving as Chair in the Supply Chain Security Subcommittee, in Washington DC. An acclaimed author of transportation security books used as textbooks in Universities worldwide: “Port Management & Operations” (2014); “Logistics & Transportation Security” (2015); “Managing Energy Security & Critical Infrastructure” (2019), and “Maritime Security” (in print, 2020). She is an Honorary Member of the US Coast Guard Auxiliary. In 2016, she received the “Excellence in Emergency Management Award” by the Emergency Management Association of Texas (EMAT). She has developed a series of Training Manuals approved by the US Coast Guard NMC. She is a Certified Auditor for Security (ISPS), Safety (ISM), Quality (ISO9001) and the Environment (ISO14001).

References

- Burns, M.G., 2014. Port Management and Operations, first ed. CRC Press, Florida, United States.
- Burns, M.G., 2015. Logistics and Transportation Security: A Strategic, Tactical, and Operational Guide to Resilience, first ed. CRC Press, Florida, United States.
- Burns, M.G., 2017. Energy Security: Supply Chain and the Paradox of Change. Transportation Research Board of the National Academies, Washington DC (Conference proceedings and presentation). Available from: <https://ctssr.transportation.org/wp-content/uploads/sites/54/2017/11/Energy-and-the-Supply-Chain.pdf>.
- Burns, M.G., 2018a. Participatory Operational & Security Assessment on homeland security risks: an empirical research method for improving security beyond the borders through public/private partnerships. J. Transp. Secur. 11 (3–4), 85–100.
- Burns, M.G., 2018. Cargo Theft and Logistics Security. Texas Trucking Association (TXTA), Houston.
- Burns, M.G., 2019a. Containerized Cargo security at the US–Mexico border: how supply chain vulnerabilities impact processing times at land ports of entry. J. Transp. Secur. 12 (1–2), 57–71.
- Burns, M.G., 2019b. Managing Energy Security. An All Hazards Approach to Critical Infrastructure, first ed. Routledge, Abingdon, United Kingdom.
- FBI, 2015. Crime in the United States, Cargo Theft, 2015. An UCR FBI Publication. Available from: https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/additional-reports/cargo-theft/cargo-theft-report_-2015-_final.
- FBI, 2016. Crime in the United States, Cargo Theft, 2016. An UCR FBI Publication. Available from: <https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/additional-publications/cargo-theft/cargo-theft.pdf>.

Further Reading

- Burns, M.G., 2014. Hazardous Materials (HAZMAT); Technical Reports: US Coast Guard-Approved Manuals.
- Burns, M.G., 2014. Advanced Chemical Tankers’ Training; Technical Reports: US Coast Guard-Approved Manuals.
- Burns, M.G., 2014. Tank Ship Familiarization & Dangerous Liquids; Technical Reports: US Coast Guard-Approved Manuals.
- Burns, M.G., 2014. Train the Trainer - STCW Part B; Technical Reports: US Coast Guard-Approved Manuals.
- Burns, M.G., 2014. Crisis Management & Human Behavior; Technical Reports: US Coast Guard-Approved Manuals.
- Burns, M.G., 2014. Crowd Management & Passenger Safety; Technical Reports: US Coast Guard-Approved Manuals.
- Burns, M.G., 2014. Marine Environmental Awareness, STCW 2012; Technical Reports: US Coast Guard-Approved Manuals.
- Burns, M.G., 2014. Personal Safety/Social Responsibilities (PSSR); Technical Reports: US Coast Guard-Approved Manuals.
- Burns, M.G., 2014. First Aid, CPR & AED; Technical Reports: US Coast Guard-Approved Manuals.

Resilience in Freight Transport Networks

Zhuohua Qu*, **Chengpeng Wan†**, **Zaili Yang‡**, ^{*}Liverpool Business School, Liverpool John Moores University, Liverpool, United Kingdom; [†]National Engineering Research Center for Water Transport Safety, Wuhan, China; Intelligent Transport Systems Research Center, Wuhan University of Technology, Wuhan, China; Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, United Kingdom; [‡]Liverpool Logistics, Offshore and Marine (LOOM) Research Institute, Liverpool John Moores University, Liverpool, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Definitions of Transportation Resilience	53
Characteristics of Freight Transport Networks	53
Challenges in Research Methods and New Techniques Used in FTN Resilience Studies	54
An Advanced Method to Measure the Risk-based resilience of FTN Components Facing High Uncertainty in Failure Data	55
Quantitative FTN Resilience Analysis from a Global Network Perspective	56
Conclusions	57
References	57

Definitions of Transportation Resilience

Resilience is commonly used to describe the ability of an entity or a system to bounce back to a normal condition after its original state has been affected by a disruptive event (Henry and Emmanuel Ramirez-Marquez, 2012). Since resilience was first introduced in the context of ecological systems by Holling (1973), there are a number of different opinions and definitions of resilience in various application domains. From the perspective of transportation, various types of research on resilience have also been conducted, aiming to figure out what transport resilience is, what kind of features a resilient transport system has and what capabilities it should have. As a result, there are a variety of definitions proposed for the notion of resilience, though some of them are similar, having overlaps with other relevant concepts such as reliability, vulnerability, robustness, and survivability. The resilience definitions applied by previous transportation-related studies are summarized in Wan et al. (2018). Even though the research foci of these studies are transport systems, they are conducted from different perspectives. A careful review of definitions of resilience shows that there is no universal description as to what transport resilience is or what the standard definition of it should be. However, most similarities and differences can be observed across these resilience definitions. The highlights of resilience definitions from previous transportation-related studies are analyzed as follows (Wan et al., 2018):

- The majority of research defines resilience as a kind of ability (or capability) of a system/network, belonging to a system/network's inherent nature.
- Almost all these definitions are given with a consideration of abnormal conditions such as shocks, disturbances, disruptions, or even disasters. This reveals that one of the core intentions of resilience is the performance of a system in the face of disruptive events.
- Different verbs (such as resist, absorb, maintain, withstand) are used to describe the resilience of a system when a disruptive event occurs. Among all the actions, "recovery" is considered as the most critical one.
- It normally requires a system to return back to a predisaster state or at least be close to it.

Taking into account such highlights from the definitions in the literature, it is proposed that the resilience of a freight transport network (FTN) is the ability to recover its function of delivering freight cargoes from origin to destination within an acceptable time and cost, after the occurrence of disruptions. In the meantime, it is noteworthy that transport resilience is a contextual based definition, requiring further analysis of its characteristics to find the best way to describe and measure it with reference to the stages of occurrence of disruptions.

Characteristics of Freight Transport Networks

Different terms have been used to describe resilience and its characteristics, including but not limited to vulnerability, adaptability, robustness, preparedness, redundancy, response, and recovery. Currently, there are only a few studies analyzing the similarity and differences in the application of such terms in the transportation arena. Here, the most commonly used terms when describing the features and connotations of resilience have been extracted from the literature, as summarized in Table 1.

As a crossdisciplinary concept, resilience has been studied in different research fields from various aspects, with emphases on one or several of its particular properties. Sometimes it is not sufficient to describe resilience by only using mathematical equations, especially in a more general situation. It will increase the difficulty for decision makers to understand and apply it

Table 1 Interpretations and analysis of terms related to resilience.

Term	Interpretation/analysis
Vulnerability	It is defined as the susceptibility to damage or perturbation—especially where small damage or perturbation leads to disproportionate consequences. It is also regarded as the property of a transportation system, which may weaken or constrain its ability to endure, handle and survive threats and disruptive events that originated both within and outside the system boundaries.
Adaptability (or adaptive capacity)	It is defined as one of the functions of a resilient system, reflecting its flexible ability to response to new pressures. Its main features lie in its response to changes reflecting the dynamic nature of complex systems.
Robustness	It is the property of being strong, healthy, and hardy. Thus it is generally defined as the ability to withstand or absorb disturbances and remain intact when exposed to disruptions.
Flexibility	It's the ability of a system to respond to shocks and adjust itself to changes through contingency planning after disruptions. It is also referred to as an ability to reconfigure resources as well as to cope with uncertainties. As such, connotations of flexibility are opposite to that of robustness, which emphasizes the ability to endure these changes rather than to adapt to them.
Reliability	It is generally defined as the probability that a network remains operative given the occurrence of a disruption event. It can be either a pre-disruption or postdisruption metric for measuring system performance.
Recoverability (or the ability to recover)	It has been discussed the most in terms of the research of transportation resilience. It is defined as the ability of a network to recover functionality in a timely manner. It is regarded as an important feature of secure and highly functioning transport networks.
Redundancy	It indicates the ability of certain components of a system to take over the functions of failed components without adversely affecting the performance of the system itself. In the context of transportation, redundancy is generally viewed as the existence of optional routes between origins and destinations. It is commonly accepted that the more redundancy a system has, the more resilient it will be.
Survivability	It is generally defined as the ability to withstand sudden disturbances while meeting original demands. Survivability techniques have been considered as an access to mitigating the vulnerability of a network or system.
Preparedness	It refers to “ <i>prepare certain measures before disruption happens</i> ”, and it enhances the resilience of a system by lessening potential negative impacts from disruptive events. It can be subdivided as emergency preparedness and response preparedness.
Resourcefulness	Resourcefulness is defined as the availability of materials, supplies, and crews to restore functionality in the study of transportation resilience. Resourcefulness was treated as one of stabilizing measures in resilience. It indicates the level of preparedness in effectively resisting an adverse event.
Responsiveness	It is regarded as an important factor to the resilience of transportation networks. Similar to redundancy, responsiveness factors of a system may also increase the costs although it is able to improve the service level of a system.
Rapidity	It is a well-studied concept in the “ <i>resilience triangle</i> ”, a framework that has been applied in civil infrastructure for decades. It contains a hidden meaning of recovery, but with more emphases on the speed to recover. It affects the duration of reduced performance of a system.

Source: Reproduced with permission from Taylor & Francis Ltd (www.tandfonline.com). The original source of publication is from Wan et al. (2018).

in practice, and inevitably result in the neglect of parts of its properties in theoretical research. Thus Wan et al. (2018) concluded and expounded the key characteristics of resilience by using graphic representation. The hypothetical system performance of a resilient system under normal conditions and in the face of disruptive events are shown as Fig. 1, which attempts to incorporate as many characteristics of resilience mentioned in the literature as possible, and provides a general overview of performance of a time-dependent system.

For a transport system, performance can be understood as the service function it offers, and it is usually measured with operational metrics such as a component's capacity, traffic flow, and throughput. Overall, the performance with respect to the time of occurrence of a disruptive event can be divided into three stages: predisruption (t_0, t_e), disruption (t_e, t_r), and postdisruption (t, t_r) periods. From Fig. 1, it is observed that the resilience of a FTN is largely determined by reliability, redundancy, robustness, and recoverability (the 4Rs), as they measure the overall performance of an FTN in terms of how long it can perform without failing, what actions it will take in the face of a disruptive event, how much function will remain after being disrupted and how it reaches a new equilibrium.

Challenges in Research Methods and New Techniques Used in FTN Resilience Studies

In previous transport resilience studies, the most used methods are quantitative and include mathematical modeling, simulation, and their combination. It is often the case that simulation is used to validate the analysis of mathematical modeling. Qualitative studies involve surveys, case studies and conceptual work. Reviewing the previous methods used in transport resilience reveals that (1) there are a lack of empirical data when conducting transportation resilience studies; (2) natural hazards such as flooding and earthquakes are identified as predominant sources of external disturbances, particularly in qualitative analysis; and (3) resilience is often measured at a microlevel of transport nodes or links, with the support of a real case study, thus lacking a systematic analysis from a whole transport network viewpoint. For instance, in terms of the applied context, ports attract the most qualitative studies, largely because of their crucial roles in FTNs. However, the issues as to how the resilience of a port affects the resilience of the wider

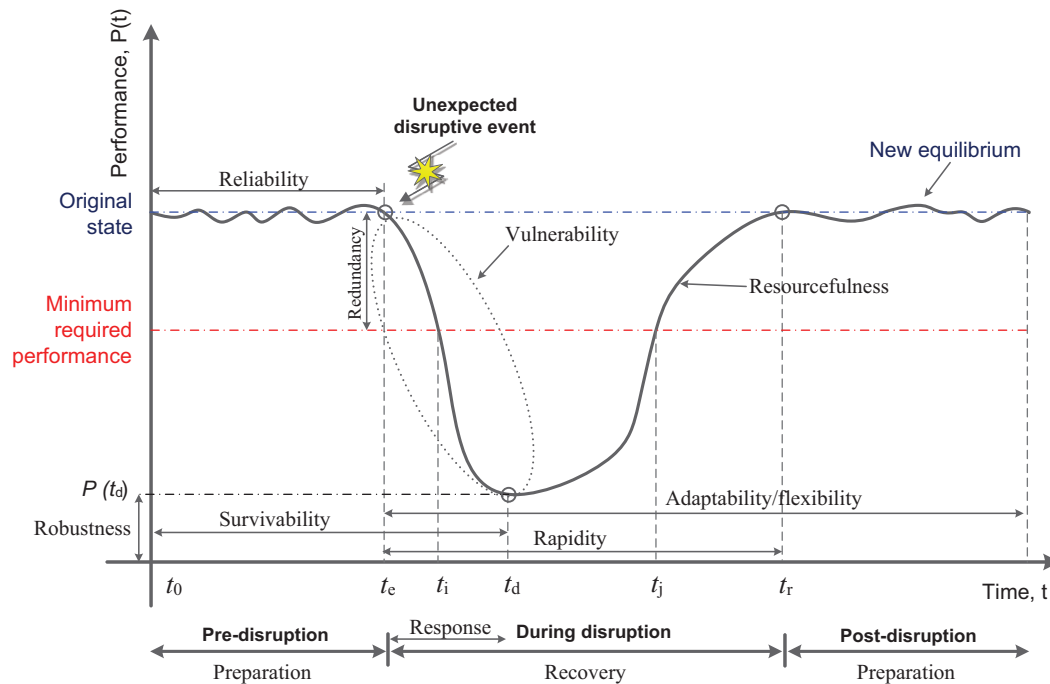


Figure 1 Schematic of performance of a resilient system. Source: Reproduced with permission from Taylor & Francis Ltd (www.tandfonline.com). The original source of publication is from Wan et al. (2018).

FTN has yet to be answered. In light of the above, analyzing the resilience of an FTN requires addressing the following challenges with some urgency:

- How to quantify the resilience of freight transport with uncertainty in data? Quantifying the resilience of FTNs will be very beneficial in justifying the investment in new infrastructure through cost benefit analysis. However, traditional resilience analysis techniques that are capable of dealing with disruptions by classical risks will probably be insufficient to tackle emerging ones associated with climate risks, against which objective failure data are often not available to aid the relevant analysis. In a recent study, Wu et al. (2019) quantified the resilience of maritime shipping networks under natural hazards based on the concept of resilience triangle. In the study, a resilience index was proposed according to Fig. 1, which was defined as the ratio between the area of the changed/influenced performance and the initial status with respect to the total recovery time needed. However, the recovery cost was not considered in their research, which still remained as an important issue to be dealt with.
- How to measure the resilience of transport nodes and links from a global FTN perspective? In the design and management of transport networks, it is crucial to understand which components are the most important to the resilience performance of the whole network, and are thus vulnerable when facing disturbances. However, previous relevant studies focus more on the risk-based resilience analysis of individual segments (e.g., road, port, and shipping) of large multimode FTNs (e.g., international container supply chains). Few studies have been found to measure the vulnerability of components considering the resilience of the whole FTN (e.g., Liu et al., 2018). It is often the case that marginalized nodes (e.g., ports) in a large FTN are exposed to high risks but have poor resilience due to economic constraints compared to more centralized nodes. However, their impact to the whole network resilience is relatively low. It is therefore essential to take into account both local internal risk and global external safety impact when measuring FTN resilience.

An Advanced Method to Measure the Risk-based resilience of FTN Components Facing High Uncertainty in Failure Data

Conventional risk assessment approaches [e.g., quantitative risk assessment (QRA)] which have been widely used to carry out risk analysis in many sectors are not well-suited to dealing with emerging risks (e.g., climate risks) in which, as mentioned earlier, a high level of uncertainty in data exists. Due to the serious scarcity of historical/statistical data, the assessment needs to be conducted using subjective judgments for the purpose of prioritizing climate risks and rationalizing the allocation of adaptation resources. This may comprise various types of uncertainties including, but not limited to, “fuzziness” and “incompleteness.” An effective way to cope with such imprecision is to use linguistic assessment. However, such linguistic descriptions define risk parameters to a discrete extent

so that they can at times be inadequate. Fuzzy set theory has the ability to model such subjective linguistic variables and deal with the discrete problem. In Yang et al. (2018), new risk parameters such as *timeframe* (T), *likelihood* (L), and *severity of consequences* (S) are modeled by fuzzy linguistic terms to calculate the climate risk index. The risk parameters are tested through a survey, with the possibility of adding or eliminating parameters. Through the survey the linguistic variables/grades used to describe each identified risk parameter are defined and used to evaluate the risk levels of climate risk. A novel climate risk assessment model is developed by combining fuzzy logic and Bayesian network approaches in a complementary way, in which a fuzzy rule based method is employed to construct the collected information (i.e., climate risk parameters and their associated grades) through fuzzy belief IF-THEN rules, while a Bayesian reasoning mechanism is used to aggregate all relevant rules for facilitating risk calculation and realizing the climate risk evaluation. The theoretical foundation of the new climate risk model shares the same principle as with a fuzzy rule based Bayesian reasoning (FuRBaR) approach, which was originally presented in Yang et al. (2008).

The new climate risk model has been reinforced by expanding the risk parameters to a more detailed level [e.g., the use of economic and environmental damage to model *severity of consequences* (S)] in which they can be better evaluated by using both objective data and subjective judgments. It strikes a new thought on the solutions to this issue in future. Climate risk parameters can be further broken down to a level where data from other climate models (e.g., sea level rise) can be used as the input and integrated into the proposed risk model. In the meantime, the risk parameters will be examined to find their correlations with climate change risk indexes (CCRIs) (e.g., temperature), which can be measured using historical or quantitative objective data. By doing so, the objective measures on the CCRIs can complement the expert evaluations based on the defined linguistics variables to improve the analysis accuracy. Finally, sensitivity analysis can be conducted to investigate the resilience of the investigated FTN component(s) to different levels of disruptions of the prioritized climate risks.

Quantitative FTN Resilience Analysis from a Global Network Perspective

Analyzing the resilience of any FTN requires the consideration of both internal risk levels of FTN components (e.g., shipping and ports) and their external safety impacts on the global performance of the network. This is because the high risk of a transport node or link does not necessarily indicate a high impact on the resilience of the whole FTN, as its impact also depends on the specific role it plays in the whole network. Conventional resilience analysis of a transportation network often compares the variation of certain indices related to network performance, before and after the removal of node(s) and link(s). In the context of FTN, the removal of a node implies the shutdown of a port/dry port or an inland depot, such as a regional distribution center, while the removal of a link refers to the changing of freight transport service configurations. Here, the concept of network vulnerability is often used to describe the degradation of network performance due to a random or selected removal of nodes and links (Liu et al., 2018). Random removals imply the disruptions that may occur on any node or link in a random manner. They are caused by natural disasters or unexpected accidents. Selected removals (e.g., deliberate attacks) will prioritize certain target nodes and links following a strategy of degrading the network performance in the fastest way possible. Such priority is obtained by studying network topology features, only without the incorporation of the selected nodes/links' resistance ability to the attacks, which are often denoted by its internal risk levels. In other words, in traditional transport network resilience studies, there was a strong assumption that the effort to remove the selected nodes/links when facing deliberate attack is identical. Such an assumption becomes error-prone in the context of an FTN, in which ports and shipping routes often have different resistance ability to external attacks/disruptions. In this study, we use the internal risk level of the nodes and links to express their resistance ability. The higher the risk level is, the lower the resistance ability. While the internal risk estimates of FTN components are obtained from traditional QRA or the advanced method presented in the fourth section, the focus of its resilience analysis is shifted onto: (1) how to expand the network topology measures (e.g., centrality analysis) to describe the external impact of individual components on the performance of an FTN and (2) how to combine the internal risk estimate of a component with its external impact.

To measure the impact of individual nodes in a FTN, a model containing degree centrality, node strength centrality, betweenness centrality, and closeness centrality is developed based on the Borda Count method (Liu et al., 2018). The impact of individual nodes based on Borda Count is normalized and then used to calculate the impact of any link by the multiplication of the impact values of the two nodes it links. The impact of a node/link is then used to express their importance/weights. To combine the internal risk estimates of a node/link and its external impact, an evidential reasoning (ER) approach (Yang et al., 2018) is used because of its advantages of accommodating both quantitative and qualitative data and the relative ease of conducting sensitivity analysis by assigning a zero weight to any eliminated nodes and links of the FTN. The quantitative data refer to the risk estimates of certain nodes/links by QRA, while the qualitative data mean the risk results by FuRBaR in a form of belief based linguistic terms. The quantitative data, if any, can be converted into qualitative data using the rule base in the ER approach. As a result, the performance of any investigated FTN can be obtained, which reflects the resilience of the FTN during normal operational environments. Next, the internal risk level of any component can be adjusted to test its impact on the performance of the FTN through sensitivity analysis. Compared to previous studies of eliminating nodes/links, the new model allows the analysis of the impact of the partial functioning of a node/link (due to a particular disruption) on the resilience of the FTN. The difference between the reduced performance after the sensitivity tests and the original performance is used to express the resilience of the network to the impact of a particular disruption on a node/link. Currently, the external impact is based on static centrality analysis without the incorporation of the freight throughput of any node/link. This prompts a new research direction in the future to investigate how different freight throughput affects the resilience of an FTN.

Conclusions

This article answers the questions on what TFN resilience is, what characteristics it should have, and how to build and measure TFN resilience. Furthermore, the research challenges and useful remarks on resilience measures in TFN systems are analyzed and discussed. The relevant findings and implications are concluded as follows:

- *Definition of contextual resilience.* There is, as yet, no universal and widely accepted definition of transport resilience. It is left to be argued if, as an interdisciplinary concept, TFN resilience requires a universally accepted definition, and if resilience should be utilized in different ways depending on specific applications at hand. Nevertheless, it is essential and significant to propose a specific definition of resilience to define its study scope, research methods, and required data before applying it within certain domains, such as disaster resilience, and climate change resilience.
- *Quantitative resilience analysis.* The majority of the resilience studies define it as an ability of a system that can absorb disruptions, while some describe it as a function or metric to measure the system performance against disruptions. In theory, both ability and metrics can be quantified. In fact, from a risk perspective, if resilience measures cannot be assessed quantitatively, the resilience management system does not motivate industrial companies/local authorities to adopt them, possibly because their effects are not visible in a state-of-the-art risk assessment. Risk-based resilience of TFN components with respect to particular emerging risks (e.g., climate change) can be coped with by incorporating subjective data based on advanced uncertainty models such as FurBar. TFN system resilience can be measured by incorporating both internal risk estimates of individual nodes/links with their external impact based on an ER approach. The recommended evaluation of the external impact should take into account dynamic cargo throughput in future.

In general, attention needs to be given to the application of resilience in the early design of the associated infrastructure by the introduction of a “design for resilience” concept. It is necessary to develop and standardize a framework for quantitative resilience analysis that can incorporate the features of transportation resilience, consider the different phases of a disturbance striking the system and evaluate resilience from a global systematic perspective, connect resilience with safety management, and properly cope with the uncertainty in both qualitative and quantitative data. It is required to enable this framework, not only to assess the resilience status of existing transportation systems to identify vulnerable parts and prepare for unpredictable disasters, but also in the system design process, to provide a reference for optimal decision making for the development of transport infrastructure, on issues such as route planning and key infrastructure maintenance and renewal.

References

- Henry, D., Emmanuel Ramirez-Marquez, J., 2012. Generic metrics and quantitative approaches for system resilience as a function of time. *Reliab. Eng. Syst. Safe.* 99, 114–122.
- Holling, C.S., 1973. Resilience and stability of ecological systems. *Annu. Rev. Ecol. Systemat.* 4, 1–23.
- Liu, H.L., Tian, Z.H., Huang, A.Q., Yang, Z., 2018. Analysis of vulnerabilities in maritime supply chains. *Reliab. Eng. Syst. Safe.* 169, 475–484.
- Wan, C., Yang, Z., Yan, X.P., Zhang, D., 2018. Resilience in transportation systems: a systematic review and future directions. *Transp. Rev.* 38 (4), 479–498.
- Wu, J., Zhang, D., Wan, C., 2019. Resilience assessment of maritime container shipping networks—A case of the Maritime Silk Road. In: 2019 5th International Conference on Transportation Information and Safety (ICTIS), IEEE, pp. 252–259.
- Yang, Z., Bonsall, S., Wang, J., 2008. Fuzzy rule-based Bayesian reasoning approach for prioritization of failures in FMEA. *IEEE Trans. Reliab.* 57 (3), 517–528.
- Yang, Z., Ng, A., Lee, P.T.W., Wang, T., Qu, Z., Rodrigues, V.S., et al., 2018. Risk and cost evaluation of port adaptation measures to climate change impacts. *Transp. Res. D* 61, 444–458.

Environmental Sustainability in Freight Transportation

Lisa M. Ellram, Farmer School of Business, Department of Management, Miami University, Oxford, OH, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	58
The Impact of Freight Transportation on the Environment	58
The Contribution of Freight Transportation to GHG Emissions	58
Transportation Impact by Mode	59
Reducing Freight Impacts	61
Moving Forward	62
References	63
Further Reading	63

Introduction

Emissions from freight transportation is one of the largest sources of industrial emissions, and it is growing. Yet, it receives very little attention from consumers or in the popular press. As a result, we are missing opportunities to make significant progress that can simultaneously reduce emissions, reduce fossil fuel use, and reduce costs and slow climate change. A fundamental question in logistics and transportation is “what transportation mode should I choose?” The modal choice depends on many factors, such as when the item is needed, accessibility to that mode of transportation, cost of shipping, item characteristics such as weight, any special handling required, value, fragility, and so forth. Another important criterion that is often overlooked is that of the environmental impact of each mode, and the carrier selected within the mode. The purpose of this paper is to explore the environmental impact of freight transportation in general to understand the environmental and other performance characteristics of each mode, explore the progress that has been made in environmental transportation, the barriers to further progress, and how to move forward in research and in practice.

The Impact of Freight Transportation on the Environment

The greenhouse gas (GHG) emitted by burning fossil fuels has a significant, negative impact on the natural environment and on human health. GHG from transportation emissions includes carbon dioxide (CO₂), methane (CH₄), and nitrous oxide (N₂O). It also includes sulfur and particulate matter that are harmful to human health and the natural environment. These are well-known negative impacts documented by the World Health Organization, and other highly credible international organizations ([World Health Organization, 2018](#)). The implications are staggering: 9 out of 10 people globally live with air pollution, which takes approximately 7 million lives annually. Ninety percent of children breathe toxic air, creating lasting damage to their developing lungs, increasing rates of asthma, emphysema, and other health-related breathing problems, increasing health-care costs, and reducing quality of life.

In addition, air pollution is influencing a faster rate of climate change. This contributes to extreme heat, more natural disasters, food insecurity, and more premature deaths, to the elderly and children in particular. Freight emissions is one important and significant source of GHG that can be reduced, with resulting reductions in the negative impact of emissions ([Environmental Protection Agency, 2017](#)).

The Contribution of Freight Transportation to GHG Emissions

Understanding the magnitude and impact of freight emissions is extremely complex. Various reports break out emissions in different ways, some including all passenger and freight transportation, some including only domestic freight, and some including only international shipping. In addition, they include different assumptions about the vehicle types, weights, and use kilometers or miles, so comparisons are challenging. In spite of these complexities, several things are clear from the data. First, freight transportation relies heavily on the burning of fossil fuels, which generates a high level of air pollution and associated environmental damage. Second, while each country includes its domestic transportation impacts in its total GHG emissions, international aviation and marine emissions are reported separately, as no country is willing to claim these. These are not included in the Paris Climate agreement, but are managed separately by the International Civil Aviation Organization and the International Maritime Organization (IMO). Thus, country-level emissions are understated. Third, with growth in international trade and travel, international transportation emissions of all types, including freight, are large and increasing. A recent study by the International Council on

Table 1 Estimated freight emissions included in US total transportation emissions^a

Category	Percentage
Heavy duty trucks	21%
Pipelines	2%
Rail	1.9%
Aviation/aircraft	3.3%
Domestic ships/boats	2.1

^aAll estimates are from the EPA Fast Facts 1990–2017, except aviation is estimated from the EDF Green Freight Handbook.

Clean Transportation (ICCT) indicates that if treated as a country, the impact of international freight shipping would be the sixth largest emitter of energy-related CO₂ in 2015 (Olmer et al., 2017).

Most people do not spend much time thinking about the environmental impact of transportation, or the differences in environmental impact between modes. Companies do not provide insight into the differential environmental impact of overnight shipping (probably air) versus shipping that takes a couple of days (probably truck), more than a week (rail and truck combination), or over a month (intermodal—ocean freight plus some domestic mode such as truck or rail). Because consumers do not demand this visibility, companies do not offer the visibility. Most companies do not spend a lot of time trying to minimize the environmental impact of freight. They focus on getting a low cost while meeting their customers' delivery expectations.

Good data are available for the impact of freight transportation for the United States from the environmental protection agency. This is used for illustrative purposes here to gain an understanding of implications of freight emissions (Table 1).

This means that freight transportation is around 8.8% of total GHG emissions in the United States, excluding international ocean and airfreight shipping. It is a significant amount from one source. Ocean shipping, the impact of which is not acknowledged by individual countries, adds an additional 2%–3%.

Given what we know about the science of climate change, it is important to work on reducing the impacts of freight transportation. It is possible to reduce the amount of transportation by working smarter, and to reduce the impact of transportation by shifting to less polluting modes, and working on reducing the impact of existing transportation. These ideas are presented in further detail later.

Transportation Impact by Mode

While the earlier detail provides information on the impact of modes based on their usage, it does not give insight into the impact of each mode per ton-mile. A few things to keep in mind when considering transportation impacts are as follows:

- In advanced economies, virtually all consumer goods travel via truck at some point on their path to the ultimate consumer.
- All ocean freight involves very long distances. In addition, while the CO₂ emissions of ocean freight may look relatively low per ton-mile, ocean freight has additional negative impacts not captured within GHG accounting.
- The figures are averages, and the choice of a specific carrier within a mode, as well as how fully utilized the capacity is on a given trip, results in significantly different impacts.

Table 2 summarizes the relative performance of the major transportation modes according to their key performance characteristics.

From reviewing Table 2, several things should be clarified:

1. Air and motor carriers provide the shortest and most reliable transit times.
2. The faster the shipping mode, the worse the environmental footprint per ton-mile. This makes sense, as these modes are more expensive because they burn more fossil fuel, and the fossil fuel consumption directly impacts the emissions.

Table 2 Relative modal performance

Performance criteria*	Motor (truck)	Ocean shipping	Railroad	Air	Pipeline**
Cost per ton-mile	4	2	3	5	1
Speed/transit time	2	4	3	1	5
GHG***	4	2	3	5	1

* Ranking is based on 1 as the best (fastest, cheapest, lowest emissions) and 5 as the worst.

** Pipeline considers the impact of transmission.

*** Cost and GHG rankings are based on performance per ton-mile.

Table 3 Estimated average emissions performance by mode in gram per ton-mile^{a,b}

Mode	CO ₂	NO _x	PM ₁₀	PM _{2.5}	SoX
Truck	210	1.145	0.047	0.045	Not available
Ocean ^c	11	0.158	0.025	Not available	Very high
Rail	22.94	0.427	0.012	0.012	Not available
Air, short haul ^d	4,236	38.2342	0.250797	Not available	Not available
Air, long haul	1,461	13.1497	0.086482	Not available	Not available
Intermodal	77.19	0.635	0.022	0.022	Not available
Barge ^e	17.48	0.4691	0.0117	0.0111	Not available

^aCalculations are in grams emitted per short ton (2000/lbs) of freight moved. No data on pipeline were available.

^bRail, truck air, and intermodal are from 2018 SmartWay Logistics Company Partner Tool.

^cBased on EDF data.

^dShort haul is 3000 miles or less.

^eFrom US Maritime Administration and the National Waterways Foundation (US MARAD), amended January 2017. *A Modal Comparison of Domestic Freight Transportation Effects on the General Public*. Prepared by Center for Ports & Waterways, Texas Transportation Institute (Table 10). Available from: <http://www.portsofindiana.com/wp-content/uploads/2017/06/Final-TTI-Report-2001-2014-Approved.pdf>; provided in 2018 SmartWay Logistics Company Partner Tool report.

3. Faster shipping is directly more expensive.
4. An increasing amount of freight is being moved by personal and light utility vehicles to support the last mile of e-commerce logistics for home delivery. The impact of this is not included in the transportation figures reported here.

The earlier detail provides the relative differences. What about the absolute impacts of the various modes? **Table 3**, derived from numerous sources, provides some initial insights.

These numbers can be a bit overwhelming and confusing. Most of us are not familiar with these units of measure. The figures are comparable to each other, because each reports the emissions impact of that mode for 1 ton of freight shipped 1 mile. To provide some insights, **Table 4** scales these numbers so that ocean shipping, which is the lowest in terms of grams per ton-mile for CO₂ is a base of 1 for each type of emission. In addition, keep in mind that they are estimates based on many assumptions and averages related to the exact technology used, the capacity utilized and more, and an individual carrier within a mode may perform quite differently: either better or worse.

Truck is the most common mode of domestic transportation in most cases. The impact of ground freight transportation is increasingly understated as more personal and passenger vehicles are being used for the last mile of home deliveries to end customers. In addition to the various types of emissions to air of trucks, they also negatively contribute to noise pollution, road congestion, traffic accidents, and deaths (Environmental Protection Agency, 2019).

Ocean shipping is the most important form of transportation for global trade. Not only is it essential for the movement of goods between continents, it causes significantly less impact in most emissions per ton-mile than does airfreight. Important exceptions to this are black carbon and sulfur dioxide. Inexpensive transportation is essential for efficient and competitive world trade. While transportation by ocean can be long, 20–40 days depending on the ports and seasonal congestion, it is also fairly consistent. Importantly, it is extremely inexpensive, costing around \$0.05–\$0.10 to get a piece of clothing from Asia to North America, less than the transportation of the item within North America. While all of this is good, ocean shipping does have some devastating, and little-understood environmental impacts that are extremely well summarized in the work by Walker et al. (2019).

The impacts of ocean shipping are difficult to study and difficult to regulate, as they are international in nature, not “owned” by any one government. Ocean transportation is regulated by the IMO. First, ocean transport of all types, including shipping, historically uses a very high-sulfur, low-quality type of fuel known as bunker fuel, accounting for about 13% of global sulfur-

Table 4 Relative estimated emissions performance by mode per ton-mile, ocean shipping as base^a

Mode	CO ₂	NO _x	PM ₁₀	PM _{2.5}	SoX ^b
Truck	19.09	7.24	1.88	0.045	0.0004
Ocean	1	1	1	Not available	1
Rail	2.09	2.70	0.48	0.012	Not Available
Air, short haul	385.09	241.99	10.03	Not available	Not available
Air, long haul	132.8	88.23	3.46	Not available	Not available
Intermodal	7.02	4.01	0.88	0.022	Not available
Barge	1.59	2.96	0.468	0.0111	Not available

^aOcean is set to 1 in each relevant category to make relative comparison.

^bShipping bunker fuel is 27,000 ppm sulfur versus about 10 ppm for vehicles, per Walker et al. (2019).

dioxide emissions. This fuel causes extreme outputs of sulfur-dioxide emissions, which is harmful to human and marine life. As a result, the IMO has mandated that as of 2020, ships must switch to marine fuel that is 0.5% sulfur, away from the 3.5% sulfur bunker fuel. This will raise shipping costs, which will ultimately be passed on to consumers. However, a report from Yale indicates that it will also save thousands of lives, as 70% of emissions from ships are said to occur within 250 miles of land, exposing millions of people to dangerous emissions. It is estimated that the move to cleaner fuel will reduce ocean shipping-related premature mortality by 34% (150,000 deaths that are primarily heart and lung-cancer related) and illness by 54%, (a reduction of 7.6 million cases of childhood asthma). Another alternative way that companies can reduce ocean shipping sulfur emissions is to switch to scrubbers. The problem with scrubbers is that they pull the pollution out of the air but must be cleaned and the waste from the cleaning must be disposed of. This is often done at sea, which has other negative implications. Switching to liquid natural gas (LNG) would reduce all sources of shipping pollution even more (IMO, 2019).

While it appears that the fuel emissions issue is finally being addressed in some measure, with more sensitive areas enforcing a 0.1% sulfur content in marine fuel, there are numerous other negative environmental implications from ocean shipping. Two-thirds of oil spilled during transportation happens on the ocean, and has ravaging, often long-term effects on many plants, animals, and birds. There are also spills of other hazardous materials, which are increasingly being transported by ocean. Garbage waste from ships is also a problem, though freight ships have few people compared to cruise ships. They are still allowed to discharge food waste and raw sewage at sea, which is certainly harmful to the marine environment. While there are laws about proper discharge of ballast water, we know that invasive aquatic species are transferred through the release of such water. This is very difficult to police. European green crab, zebra mussel, and Chinese mitten crab are some examples of invasive species that have been spread by ballast water.

While we think of noise pollution when we think of land-based travel and airplanes, we think of it less with ocean travel because of the distance ships travel from shore. Unfortunately, the noise pollution that ships create can be very destructive to marine species, even over long distances. Sound is used by most marine animals for everything from communication and navigation to reproduction and avoidance of predators and other hazards. Noise can create stress, behavioral change, and even death. Thus, while one aspect of ocean shipping, sulfur emissions, is being addressed, there are many more.

Rail is excellent for large, bulk shipments, usually traveling 500 miles or more. While it tends to be slower than truck and does not go door-to-door unless you have a rail spur, it has significantly less emissions and is cheaper. It is also less reliable. It is often used *intermodally*, combined so that trucking is used for final delivery. Combining it in this way can be cost-effective and better for the environment when shipments are not very cost sensitive. One of the biggest issues with rail is the noise that it makes, and the space that it takes. No one wants railroad tracks in their backyard. In the United States, the miles of usable rail track have dropped dramatically in the last 50 years. It is difficult to return rail track to use once it is out of service (Ellram et al., 2019).

Airfreight has the worst environmental emissions profile of any mode. It uses so much fuel on takeoff and landing that it is very expensive and very bad for the environment. In addition, not reflected in these figures, emissions released at higher altitude have a worse impact because the air is thinner. Thus, the true CO₂ impact of air is even worse than depicted in the table. Short haul is much worse than long haul due to the impact of the takeoff and landing, with fewer miles in the air to average out those impacts. One thing is certain: you do not want to live near an airport due to the high level of emissions that planes release on takeoff and landing, and also due to the noise pollution (McKinnon, 2018).

The bottom line is that companies need to be smart about using airfreight. When is it really essential in terms of retaining customer business, or in some cases shipping medical supplies or organs to save lives? From a cost and an environmental standpoint, it should be avoided except when it is essential. For example, some fashion companies such as Victoria's Secret ship their fashions by ocean in the pre-season. During peak season for an item, stock is replenished by air in order to meet customer demand before the season ends or tastes change. Dell computer, which used to rely heavily on air to ship its final products, has also reevaluated what it can forecast versus ship-to-order, and have dramatically shifted its freight shipments to ocean away from air, saving both money and the environment.

Barge can be very effective for the same types of freight as rail, but you do need a waterway to make it work. Barge is common in some areas of Europe, and for bulk commodity items in the United States, especially on the Mississippi River.

Reducing Freight Impacts

New technologies are emerging which could dramatically reduce the impact of freight emissions. Electrification of freight transportation could also drastically reduce noise pollution. Electric semitrucks for freight haulage are under development by numerous companies, with Tesla, Nikola Motors, Volvo, Kenworth–Toyota partnership, Daimler, and Thor all having models which are either being tested or underway. These still suffer from somewhat limited range and a lack of a good charging network, so these will not take over quickly. However, as they do come on board, their positive impact on noise pollution and emissions will be dramatic. Hydrogen trucks are also underway, but suffer similar problems. Still, these provide very quiet, potentially zero emissions offerings if they are charged by alternative energy.

Hydrogen and wind offerings are also under development for ocean shipping, with renewable energy and biodiesel alternatives also being tested for air. Solar and electrification experiments are also underway for railroads. These provide great options for reducing the damage from using fossil fuel. But these are evolving and will not be widely developed and implemented overnight. In

Table 5 Reducing transportation emissions**Technology**

- Purchase the most efficient and cost-effective technology for your transportation requirements.
- Install auxiliary power units so you can turn off the main engine when you are not moving.
- Install governors to limit the maximum speed to one that is efficient.
- Install GPS units to reroute out of bad traffic areas.

Company priorities

- Elevate the awareness of freight transportation impacts. Set goals for reduction and reward and publicize the achievement of goals.
- Use SmartWay certified or other environmentally certified carriers who use current equipment and know how to reduce their environmental footprint.

Educate and increase awareness

- Train your drivers of any mode to drive efficiently, and consider rewarding them for their efforts.
 - Turning off the engine (no idling) when stopped for more than a few minutes.
 - No sudden acceleration, which burns up a disproportionate amount of gas.
 - Driving at a steady and moderate speed.
- Join a green transportation initiative such as SmartWay (UA, Canada) or Green Freight (globally) to learn more about freight emission and techniques to improve freight efficiency.

Improve equipment planning and utilization

- Use intermodal transportation where possible.
- If you own your own fleet, utilize it where you can be efficient, and outsource your transportation to more efficient carriers in areas where you have limited volume and are no backhaul. Empty miles are not only expensive, they add to your environmental footprint by moving air.
- Work on improving your utilization of loads as much as possible. There are many ways to do this, including:
 - Look at your product configuration to make sure you are shipping in the most efficient way in terms of weight, space, palletization, and so on.
 - Redesign and reduce your packaging weight and size so that you can fit more in each shipment while also reducing packaging waste.
 - Plan your loads and routes so that they can be a fully utilized as possible.
- Continually revisit your logistics network design, loads, customer needs, and improve upon your transportation processes.
- Plan routes to avoid traffic at peak hours in congested areas.

Collaborate with others in your supply chain to reduce transportation emissions

- Consider your customer's true needs, and use less damaging, slower shipping by planning better. These means saving air shipments for emergencies only. Switch from truck to rail wherever possible.
- Talk to your carriers and customers about ways to improve your load utilization and get their ideas, including ways to better utilize backhaul capacity.
- Encourage your suppliers to reduce their freight transportation emissions and join organizations such as SmartWay.

the meantime, there are certain behaviors that firms who want to reduce their fossil fuel use and associated cost and environmental impact can implement. These are presented in [Table 5](#).

Following the approaches mentioned earlier can have a tremendous impact. For example, Walmart was able to more than double the efficiency of its freight fleet between 2005 and 2015, saving almost \$1 billion and avoiding 65,000 metric tons of CO₂. It is now focusing on the electrification of fleets and encouraging customers to combine trips and orders to reduce the impact of the consumer last mile pickup and delivery ([Walmart, 2018](#)).

Moving Forward

How urgent is the need to act and reduce emissions and other environmental impacts associated with decreased quality of human life? The fact of the matter is we cannot say with certainty. Climate changes over time due to natural shifts as well as man-made causes, and it is difficult to separate the effects completely. The vast majority of scientific evidence points to the fact that dramatic action is needed immediately if we are to maintain a good quality of human life for future generations. A 2019 Harvard University study indicated we have 5 years to take significant action, or it will be too late to slow the affects of climate change ([Vandette, 2019](#)).

The good news is that we can make positive change. It begins with education that this problem exists, the impacts, and what companies and individuals can do to reduce the impact. For consumers, this means combining trips to the store—or online shopping purchases to minimize their carbon impact. Choosing the slower shipping mode allows the retailer to combine your shipments even if it means waiting. It means buying items in person that are likely to be returned if you cannot see them or try them on. This will reduce the impacts of the returns process, which affects freight emissions as the item is picked up from a consumer's home and shipped to another location, then often shipped again to a distribution point where it may be reshipped to a consumer, a

store, or landfill. Clothing and shoes purchased online are returned about 30%–50% of the time, versus about 5% for in-store purchases. Consumers should think twice before hitting the purchase button for these types of buys.

Businesses can ship smarter and instill a low-carbon shipping culture into their company. Many of the techniques in **Table 5** will save money. This money can be reinvested in technology, bonuses for reaching GHG reduction targets, or more education.

References

- ElIram, L., Fawcett, S., Goldsby, T., Hofer, C., Dale, R., 2019. Logistics Management: Enhancing Competitiveness and Customer Value. MyEducator. Available from: www.myeducator.com.
- Environmental Protection Agency, 2017. EPA's Endangerment Finding: Health Effects. Available from: <https://archive.epa.gov/>.
- Environmental Protection Agency, 2019. Fast Facts U.S. Transportation Sector Greenhouse Gas Emissions 1990-2017. Available from: <https://www.epa.gov/>.
- IMO, 2019. Introduction to the IMO, International Maritime Organization. Available from: www.imo.org.
- McKinnon, A., 2018. Decarbonizing Logistics: Distributing Goods in a Low-Carbon World. Kogan Page.
- Olmer, N., Comer, B., Roy, B., Mao, X., Rutherford, D., 2017. Greenhouse Gas Emissions from Global Shipping, 2013–2015, ICCT.
- Vandette, K., 2019. Harvard Scientist: We have 5 years to Mitigate Climate Change. Earth.com. Available from: <https://www.earth.com/news/5-years-mitigate-climate-change/>.
- Walker, T.R., Adebambo, O., Del Auila Feijoo, M.C., Elhaimer, E., Hossain, T., Edwards, S.J., Morrison, C.E., Romo, J., Sharma, N., Taylor, S., Zomorodi, S., 2019. Environmental effects of marine transportation. In: *World Seas: An Environmental Evaluation*. Academic Press, pp. 505–530. Available from: <https://doi.org/10.1016/B978-0-12-805052-1.00030-9>.
- Walmart, 2018. 2018 Global responsibility report. Available from: <https://corporate.walmart.com>.
- World Health Organization, 2018. How air pollution is destroying our health. Available from: <https://www.who.int/>.

Further Reading

- ElIram, L.M., Murfield, M., 2017. Environmental sustainability in freight transportation: a systematic literature review & agenda for future research. *Transp. J.* 56 (3), 263–298.
- Environmental Protection Agency, 2019. SmartWay Logistics Company partner tool. Available from: <https://www.epa.gov/smartway/smartway-truck-carrier-partner-resources>.
- Gallucci, M., 2018. At last, the shipping industry begins cleaning up its dirty fuels. *Yale 360*. Available from: <https://e360.yale.edu>.
- Mather, J., Craft, E., Norsworthy, M., Wolfe, C., 2015. The green freight handbook. Environmental Defense Fund. Available from: <http://www.cadenadesuministro.es/wp-content/uploads/2015/07/The-Green-Freight-Handbook-EDF.pdf>.
- OECD/ITF, 2017. ITF transport outlook 2017. OECD Publishing, Paris. Available from: <http://dx.doi.org/10.1787/9789282108000-en>.
- van Loon, P., Deketele, L., Dewaele, J., McKinnon, A., Rutherford, C., 2015. A comparative analysis of carbon emissions from online retailing of fast moving consumer goods. *J. Clean. Prod.* 106, 478–486.
- World Health Organization, 2018. Climate change and health. Available from: <https://www.who.int/>.

Sustainable Logistics, CSR in Logistics, and Sustainable Supply Chain Management

Maria Björklund*, Maja Piecyk-Ouellet[†], *Linköping University, Department of Management and Engineering, Logistics and Quality Management, Linköping, Sweden; [†]School of Architecture and Cities, University of Westminster, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	64
The Characteristics of CSR in Logistics and SCM	64
CSR Covers a Myriad of Aspects	65
Three Interrelated and Interdependent Sustainability Dimensions	65
Benefits From Applying CSR in Logistics	65
On a Voluntary Basis	66
Consideration of Numerous Stakeholders	66
Collaboration and Coordination	66
Long-Term Perspective	66
An Add-On or an Integrated Part	67
Corporate Social Responsibility in Logistics Activities	67
Purchasing	67
Freight Transport	68
Warehousing and Terminals	68
From Sustainable Logistics Toward Logistics for Sustainability	69
Acknowledgments	70
Biographies	70
References	70
Further Reading	70

Introduction

As a response to increased environmental and social challenges, new concepts and constructs have entered the logistics arena, such as the consideration of corporate social responsibility (CSR) and sustainability in supply chain management. What do these terms mean? This very much depends on whom you ask and when you ask, as the concepts may mean different things to different people at different times. In the 1980s and 1990s, when logistics managers referred to sustainability they often only meant the economic sustainability of the company. Logistics researchers in the late 1990s used the term “environment” solely for the business environment in terms of, for example, competitors and markets (and “green” for what is now more commonly referred to as the environment). In 1987, the United Nations-funded report “Our common future” proposed the still most-cited definition of sustainable development: “sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (Brundtland et al., 1987). Sustainability, therefore, comprises the three interdependent pillars of economy, environment, and society, commonly termed as the triple bottom line (3BL). Today, the United Nation’s Agenda for Sustainable Development 2030 presented in the year 2015 is commonly referred to when describing sustainability.

Commercial organizations, together with other stakeholders (e.g., politicians, consumers, and researchers), share the responsibility to work toward sustainable development. CSR can be described as a corporation’s way to address sustainability. CSR has developed from being a theoretical discussion about the role of business in society to become an integral part of logistics management and something seen as a business opportunity. Increasing recognition of CSR as a core business activity was also demonstrated by the launch of ISO 26000 in 2010. Thus, the traditional logistical mission of creating business/economic advantages is now starting to be replaced, or at least complemented, by a broader concept driven by social responsibility.

The Characteristics of CSR in Logistics and SCM

Despite the growing importance of CSR considerations in logistics and freight transport, there is still a lack of comprehensive approaches to the CSR construct in academic literature and organizational practice. Concepts such as logistics social responsibility and sustainable supply chain management are discussed by some (Carter and Jennings, 2002). Taken as a point of departure in the wide range of different definitions and descriptions of sustainable supply chain management, sustainable logistics, social responsible logistics, etc., a number of more commonly mentioned characteristics can be found. These are further elaborated on later.

CSR Covers a Myriad of Aspects

Sustainability for a corporation encompasses the consideration of its impact on the environment and society, while maintaining economic viability. The environment, society, and economy are the three sustainability dimensions. The 3BL, often referred to as “people, planet, and profit,” is a terminology commonly applied by corporations when describing their sustainability performance. However, when the CSR concept needs to be operationalized and sustainability plans put into action, it quickly becomes complex and a myriad of aspects and areas to consider arise.

One framework commonly used by corporations for accounting and reporting of their economic, environmental, and social performance is the Global Reporting Initiative (GRI) framework, developed by the GRI (Milne and Gray, 2013). The framework divides the three dimensions into close to 30 areas to report on, covering social aspects such as labor practices (e.g., employment, health and safety, labor/management relations, diversity, etc.), product responsibility (e.g., labeling and marketing communication), environmental (e.g., the use of energy, materials, and emissions), and economic aspects (such as income and market presence).

At the same time, one of the broadest and most universal policy agendas that world leaders have pledged common action on is the United Nation’s Agenda for Sustainable Development 2030 (United Nations, 2015). This agenda put forward 17 sustainable development goals (e.g., climate action, no poverty, zero hunger, gender equality, etc.) and 169 associated targets. However, most authors and practitioners within the field tend to focus on a selected sustainability area, for instance impact on climate change, air pollution, waste generation, diversity of workforce, safety records, etc. This is because the scope of CSR is so broad and complex that it is difficult for one single corporation or research study to cover it all. The development of CSR policies and the management, monitoring, measurement, and reporting of CSR performance are resource-intensive activities, and only selected companies are able to dedicate the time and money necessary to provide a holistic picture of their CSR involvement.

Three Interrelated and Interdependent Sustainability Dimensions

These three dimensions of sustainability are interrelated and interdependent, which means that synergies/win-wins, as well as trade-offs, exist between them. This emphasizes the importance of taking a holistic approach when analyzing the consequences of, for example, different logistics decisions. There could be trade-offs both within one dimension (e.g., individual vs. collective interests within the social dimension) and among the dimensions (Piecyk and Björklund, 2015a). The interdependence can also be described as “pop-up” effects that can arise from actions taken in order to eliminate or decrease a negative environmental or social impact. Addressing one problem (press down) makes new problems arise. One example is, as a response to the climate challenge, to change from traditional fossil fuels to batteries, raising new challenges such as the labor situation in the mining industry and the type of new unfamiliar substances/metals with unknown long-term health and/or environmental impact. Despite this, both researchers and practice commonly address one dimension or one aspect within a dimension separately and in isolation.

Environmental and social performance is sometimes described as an “add on” to traditional (economically focused) business practice: to carry on “business as usual,” but to add in the consideration of the impact on the environment and/or society. However, more and more researchers and practitioners are starting to have another approach toward this, presenting the sustainability work as an integrated part of their performance, thus something that cannot be separated from other activities or decisions. For example, is the person responsible for sustainability in a logistics firm seen as a resource for guidance or an important member of the company board? Can the purchasing manager order from suppliers not approved by the sustainability department or not? Some researchers within logistics might even go so far as to suggest the need for a more “eco/sociocentric” approach where all logistics decisions/changes are furthestmost taken on the basis of their social and/or environmental impact.

Benefits From Applying CSR in Logistics

Despite the fact that CSR derives from the altruistic desire to do good and act in a morally correct way, current thinking suggests that CSR initiatives can also generate competitive advantage for the firm (Porter and Kramer, 2006; Kucht Campos et al., 2018). Several studies have focused on the trade-offs between economic and environmental/social performance. Environmental and social actions have traditionally been associated with additional cost for the corporation. However, a paradigm shift has started to take place linking sustainability initiatives to improved economic performance and several of the more recent studies suggest that the link between CSR and a firm’s profitability is becoming closer. If undertaken in the correct manner, CSR initiatives can provide several corporate benefits such as: improved reputation; facilitating employee recruitment, commitment, and motivation; increased operational efficiency; improved relations with customers, suppliers, investors, and local society; improved risk management and a readiness for new legal demands/requirements; improved product quality and increased product value; a potential to target new markets and consumer groups; as well as knowledge development. CSR activities can thereby be vital to a company’s profitability, pointing at the critical role CSR can play with regard to a corporation’s overall strategy and success (Piecyk and Björklund, 2015b).

It is a challenge to understand the complexity and possible trade-offs in the specific situation. Initiatives might not influence the 3BL in a straightforward way, as these initiatives could strengthen company reputation and thereby indirectly increase sales, while simultaneously increasing manufacturing costs. Furthermore, the results are also dependent on the measurements and system limitations applied. Different results can be found, for example, if restricting economic performance to unit manufacturing cost compared to total costs or sales. Furthermore, these win-win situations seem to have as much to do with the type of actions taken as

the situation and specific context of the corporation taking these actions, as well as the classical “it is not about what you do, but how you do it.”

Closely related to the benefits of applying CSR are the reasons why companies engage in the CSR programs:

- Regulations—where sustainability actions are undertaken in order to comply with existing regulations (i.e., a passive/responsive approach). Proactive organizations may also develop strategies to comply with anticipated future regulatory changes (e.g., logistics operators investing in alternatively fueled vehicles in the expectation of potential carbon taxes or low emission zones, etc.).
- Reduced cost—where environmentally or socially sustainable behaviors can reduce the operating cost of an organization (such as avoiding payouts or reduced insurance premiums as a result of few employees involved in workplace accidents, reduced fuel bills as a result of energy saving programs), or help it to avoid penalties for noncompliance.
- Increased profits—a company can gain a competitive advantage from its sustainability-focused differentiation strategy and the development of new sustainable products or services. Improved corporate image and green marketing can attract new and existing customers who are willing to pay a premium for green and/or ethical offerings.

Organizational culture and top management support are key antecedents to successful CSR programs in organizations. Other factors include the regulatory environment within which a firm operates, stakeholder pressure, and the personal values and beliefs of managers and staff.

On a Voluntary Basis

CSR is commonly described as “a concept whereby companies decide voluntarily to contribute to a better society and cleaner environment” (CEC, 2001, p. 2). It is notable that CSR is commonly described as something corporations do voluntarily, as a response to, for example, ethical/moral and/or economic incentives. The sustainability responsibility of corporations has a basis in economic responsibility: thus, a responsibility to generate revenue for the owners and to provide products and services in a market economy (Carroll, 1979). This is followed by the legal responsibilities to obey laws and regulations. Climbing on the social responsibility ladder, or pyramid as commonly referred to, there are two more voluntary steps: ethical responsibility, to manage business in a fair and responsible way, and philanthropic responsibility, to contribute to the development of society.

Consideration of Numerous Stakeholders

To meet stakeholder sustainability requirements is the most important driver for companies to improve sustainability. This has also been singled out as an important driver within the logistics area. Stakeholders could be local communities, employees, investors, as well as members of the supply chain, including suppliers and customers. In particular, the social dimension involves a wide range of stakeholders with different goals, demands, and opinions.

Governmental regulations and increasingly strong market signals from environmentally and socially conscious consumers are also important drivers for CSR considerations in logistics practices. For example, research studies show that purchasing companies place a high value on the environmental performance of logistics service providers (LSPs).

Collaboration and Coordination

In response to shifting business climates (e.g., as experiences with globalization and an increased focus on core competences) and thereby growing supply chains, sustainability has expanded from the domain of individual companies to that of their supply chains. A product or service is never more sustainable than the weakest link in its supply chain. Corporations have started to recognize the need to address sustainability in a way that encompasses not only their own operations, but also those of other supply chain actors. This calls for greater consideration of sustainability not only within the focal company, but also encompassing the entire supply chain. Sustainable supply chains call for interorganizational interactions in order not to move sustainability problems between supply chain actors; practices performed in firm-internal isolation are also said to be less successful. The importance of collaboration and coordination between actors in the supply chain (described as, e.g., cooperation among companies along the supply chain to collaborate with its supply chain partners and collaboratively manage intra- and interorganization processes) are put forward in the descriptions of sustainable logistics/supply chain management. The interorganizational interactions necessary can range from strategic, based on collaboration and/or partnership, to more operational, arms-length and control-based forms. More close and strategic forms of interorganizational interaction is often put forward as necessary for the development toward sustainable supply chain management, described in terms of, for example, partnerships, mutual investments and adjustments, knowledge exchange, and standardization of/development of mutual goals and norms.

Long-Term Perspective

Sustainable development put forward “future generations” as its time perspective. This is far beyond the range of quarterly reports and even the strategic focus of corporations that seldom span beyond a 5–10-year time horizon. Even if several definitions of sustainable supply chain management use wordings such as “to improve in a long term perspective,” few (none) provide a clear

description of how long this could be. Selected time frames when calculating, for example, a return on investment, can decide if an investment in a more environmental or social logistics solution is economically beneficial or not.

An Add-On or an Integrated Part

Some definitions/researchers describe sustainability as an aspect that could be added to traditional approaches to logistics management and decisions with regard to the design of logistics systems. Thus, adding sustainability to the agenda, while keep on working more or less the same way as traditionally. One example is where the sustainability actions are not influencing normal practice, but instead could be seen as adding additional KPIs or taking actions that do not result in changes in the way the logistics system is managed, such as donating unsold product to charity or using more modern and less polluting trucks.

Another approach to sustainability within logistics and supply chain management involves viewing it as an integrated element of logistics management. Thus, sustainability performance should not be seen in isolation, but instead as something that influences the traditional way of managing logistics systems. One example is if the top management decides that all suppliers must be approved by the sustainability department. Another example is if it is decided that sea transport should be used to a larger extent for transporting the company's goods. These types of actions also influence the management of other parts of the logistics system redefining, for example, the potential supplier basis or the time management of the logistics flow.

Corporate Social Responsibility in Logistics Activities

Approaches and responses to logistics sustainability pressures will vary depending on the type of organization. A manufacturing company with a mainly outsourced logistics function will engage in the sustainability agenda in a different way than a manufacturer with in-house logistics operations, a retailer, or an LSP. As the core value-adding activities will vary among these organizations, so will the magnitude and nature of their environmental, social, and economic impacts and thus their CSR agendas. The size of an organization will also determine the range and the level of sophistication of its sustainability initiatives (Piecyk and Björklund, 2015a). This may be due to the available resources, workforce knowledge and skills, legislative and regulatory pressures, and external support offered to the company in question. The experience or CSR maturity of a company will also play an important role in developing and implementing sustainability programs. Organizations that are only just starting on the CSR journey might focus more on output-oriented actions, such as a change of technology to achieve an instant drop in greenhouse gas (GHG) emissions. More mature organizations are more likely to engage in a broader scope of initiatives, often with a more strategic focus. Actions implemented may be more closely related to the redesign of processes, organization, management, etc., that is, a focus on improving the ways of working, rather than immediate results. These types of initiatives tend to produce consistent but slow-paced gains in the future, rather than quick wins often desired by new CSR players willing to quickly demonstrate their dedication to the cause.

This section takes its point of departure in three central logistics activities (purchasing, warehousing, and freight transport), and elaborates on how different aspects of sustainability can be addressed within these activities.

Purchasing

Purchasing social responsibility (PSR) can be defined as the inclusion of CSR issues advocated by a company's stakeholders in purchasing decisions. The purchasing function can be viewed as a key mechanism an organization can use to extend its CSR standards to suppliers, sub-suppliers, and contractors, thus improving the social and/or environmental performance of the whole supply chain.

The key environmental practices in PSR are related to purchasing goods with better environmental credentials, using environmentally sound suppliers, cooperating with suppliers to ensure their processes are environmentally sustainable, participating in the design of products and packaging for reuse or recycling, with the aim of minimizing the use of energy and materials at the production, and using an end-of-life stages in the product life cycle. Other actions include sharing cost, risk, and responsibility for investments in new technologies in places where they can generate the most difference, that is, usually the poorer countries where manufacturing operations take place but the cost of new technologies is prohibitive for local companies.

With respect to the social aspects of PSR, managers should ensure that fair and transparent protocols are followed in the purchasing process, whereby the offer and any confidential information obtained from a supplier is not shared with their competitors in order to obtain better quotes, contracts do not use unclear or misleading terms to gain advantage over suppliers, and common terms are stated in, for example, codes of conduct that are signed by the focal firm and its suppliers. Not accepting bribes, gifts, meals, etc., from potential suppliers are also key components of purchasing ethics. Staff involved in the purchasing process should be made aware that potential suppliers should be offered an equal playing field independent of personal preferences, likes and dislikes, gender, political affiliations, religious beliefs, or ethnic status.

While managers have full control over the internal processes, due to progressing globalization and a trend toward the outsourcing of noncore activities, over the last decades supply networks have become increasingly complex, often involving multiple tiers of sub-suppliers located in different countries. It is, therefore, very challenging but crucial for businesses to monitor the human rights performance of their suppliers in order to avoid forced, sweatshop, or child labor, and ensure that their suppliers pay living wages

and ensure safe working conditions for their employees. More proactive companies may get directly involved in educating workers with respect to their rights at suppliers' premises. Collaborations with third party organizations such as, for example, the Better Cotton Initiative are also a way to ensure sustainable supply sources in global supply chain networks.

Unfortunately, while the purchasing function is a powerful tool to ensure that a company's sustainability principles penetrate deeply into the supply chain, it is also sometimes used to simply transfer the efforts and investments required to make the product or service more sustainable to the least powerful partner in the chain. While purchasing from a greener supplier is often an important first step for companies starting on the environmental sustainability agenda, as their experience, knowledge, and skills develop, attention should be shifted to internal efforts to improve their own operations and strengthening external links to actively participate in the joint development of greener products or services.

As the trend toward the outsourcing of logistics services continues to grow in importance, companies increasingly incorporate CSR-related requirements into their purchasing process. As a result, demonstrating dedication to environmentally and socially responsible ways of operating becomes a significant differentiation factor for LSPs. The subsections later present an overview of key CSR considerations in freight transport and warehousing.

Freight Transport

Companies often tend to focus their CSR considerations in relation to freight transport on the environmental aspects of sustainability. As moving cargo is a significant source of anthropogenic GHG emissions and air pollutants including nitrogen oxides, sulfur oxides, and particulate matter, these are often reported and acted upon as key performance areas. This prioritization is also based on the green-gold approach, where actions aimed at reducing energy use and resulting emissions also result in lower operating costs. Mode-specific environmental impacts are also important: for example, vessel discharges and invasive species transfer for shipping, noise for airfreight, or chemicals used for weed control on tracks in the case of rail freight operators.

In terms of the social aspects of CSR, the emphasis is usually put on the safety aspects. The CSR actions focus on continuous vehicle operator training, drug and alcohol testing, appropriate equipment maintenance, abiding by hours of service, and weight restrictions requirements. Also, logistics is a very male-dominated industry. The need for some physical strength, time spent away from home, and not well equipped and potentially unsafe "resting places" could partially explain the reasons for that. Actions to improve the gender balance and diversity of the workforce are important social aspects of CSR agendas. Quality of life for long-distance drivers is a vital CSR concern in logistics, particularly if the type of operation requires frequent and lengthy periods away from home. Paying adequate wages and the diversity of workforce are also commonly mentioned as important social issues companies need to address. Human rights, for instance complying with child labor laws, also need attention from companies involved in freight transport, especially where the supply chain involves operations in less developed countries with regulatory regimes that leave poorer members of society vulnerable to the danger of exploitation and human rights abuse.

The sustainability of last-mile transport operations has recently attracted significant attention. From availability and fair pricing of deliveries to far-flung places, to the contribution to congestion in densely populated urban areas, there is a whole range of sustainability impacts that should be considered with reference to the final leg in the distribution chain. A significant social sustainability issue related to the last-mile delivery that has been attracting perhaps most attention in recent years relates to the emergence of the so-called gig economy. By the gig economy, we mean a system in which, rather than using own staff on traditional employment contracts, companies use short-term contracts with independent, self-employed workers to carry out the work. Such practices significantly cut costs for the organizations in question. However, while some workers enjoy the flexibility of their self-employed status, there is a growing concern about the lack of employment and income protection for those on zero-hours contracts and the long-term social consequences (e.g., the lack of workplace pension support) of such practices.

Autonomous freight vehicles, currently being tested in a number of countries around the world, may also have some significant CSR consequences in the freight transport sector. On one hand, they are likely to alleviate the economic pressures many companies currently face as a result of increasing driver shortages. On the other hand, at least in the short term, they may exaggerate the problem, as young people are less likely to invest in the training necessary to enter the profession (e.g., an HGV license) knowing the jobs may not even exist anymore in 10 years time.

Warehousing and Terminals

Logistics sustainability initiatives should also be incorporated into the warehouse, distribution center, terminal or logistics hub design, and operation. The first key decision in design for sustainability is the number, location, and capacity of distribution facilities, as this will largely determine the transport intensity of the overall logistics network. The decisions taken at this stage will have consequences for transport distances, potential to use more efficient equipment, access to terminals and opportunities to use less energy-intensive transport modes, as well as the ability to consolidate transport flows for a better utilization of transport resources. Locating distribution facilities in the proximity of environmentally sensitive areas may have profound impacts on biodiversity and/or soil quality (e.g., due to leaks or spills). 24/7 operations are increasingly common in the distribution industry and this too may have a detrimental effect on nearby residential areas, with noise and visual intrusion being of particular concern to local communities.

In warehousing there is a strong focus on the social aspects of CSR, with particular attention dedicated to measures aimed at improving workers' safety and accident prevention. Typical actions include training provided to staff to increase the safety of

handling equipment use, maintenance of a clean, safe, and healthy work environment, and the ability to quickly respond to any accidents in order to save lives and to minimize damage and disruptions. Policies to ensure workforce diversity, equal pay, and ensuring the availability of healthy meal options in workplace eateries are also keys to socially responsible warehouse operation. Another key dimension of social responsibility is paying warehouse workers living wages rather than minimum wages. A living wage can be defined as the minimum income that allows a worker to meet their basic needs without relying on extra support from the state. It is usually higher than the minimum wage, that is, the lowest wage an employer can legally pay their staff.

Environmental aspects of warehouse operation focus on reductions of energy use in lighting, temperature control and handling equipment use, prevention of refrigerants leakage and associated GHG emissions, as well as proper storage, labeling, and packaging of hazardous materials. Appropriate inventory management is also important to prevent economic and environmental losses associated with the obsolescence of stock, and so are policies to maximize the level of reuse and recycling of packaging material and the safe disposal of waste.

Actions to prevent theft and shrinkage are important for the economic sustainability of warehouse operations. Careful inventory management to prevent product obsolescence is another example of economically sustainable action in a warehouse. By tight control of stock rotation, companies can not only improve their profitability, but also reduce the risk of products going out-of-date and being wasted, thus also improving their environmental results.

From Sustainable Logistics Toward Logistics for Sustainability

Traditionally, the key objective of logistics as a core business function was to contribute to gaining and maintaining the competitive advantage of a company or a supply chain, by maximizing the level of customer service while simultaneously minimizing the cost of delivering that service. The focus was purely on profit generation, that is, ensuring the economic sustainability of a business. As the concept of sustainable development has risen to the top of political and corporate agendas, social and environmental concerns put constraints on untamed economic development in order to prevent social exploitation and ensure that enough natural capital is preserved to meet the needs of future generations.

However, what is the future role of logistics systems? The perception and understanding of logistics ought to look beyond its traditional business function, and focus on the true social values that well-functioning and sustainable logistics systems bring to modern economies. Logistics in itself is a core to a sustainable society.

First of all, without logistics there would be no modern societies. Without everyday transport of chemicals for water treatment/purification plants, we would not have clean water in our homes. Without transport links, housing areas would be dependent on the proximity of food supply. Waste and sewage removal is a highly complex logistics operation. Modern cities only exist thanks to highly complex, efficient networks ensuring that the goods and services we consume in everyday life are delivered at the right time and at the right cost to the place where they are needed.

Recently, increasing attention is given to the development of the logistics systems that are created with the aim of contributing to a better environment and/or society. Examples of such systems include:

- Humanitarian logistics, for example, the design, management, and operation of a logistics system, developed to support people in need due to, for example, natural disasters by providing them with emergency or continuous supplies of food, shelter, or equipment needed to alleviate poverty.
- Food banks and the redistribution of surplus food from private homes, supermarkets, food producers and retailers, cafes, and restaurants to people in need. Apart from formal initiatives introduced, for instance, by retailers as part of their CSR programs, the widespread reach of social media facilitates the development of less formal civil society networks that want to do good by sharing functional food that otherwise would go to waste.
- Secondhand, freecycle, and upcycle networks where people sell or give away unwanted items in order to reduce waste and support other members of society.
- City logistics initiatives spurred from the will to improve the quality of life in densely populated urban areas. In order to combat the increased level of noise, congestion, and pollution, various initiatives aiming, for example, to consolidate transport, use greener vehicles, or introduce off-peak deliveries are initiated.

The aforementioned logistics systems are designed primarily with nonfinancial objectives in mind (i.e., the urge to do right, to help people in need, to improve societies, and/or to decrease the environmental impact) that take priority over the traditional criteria of profit and competitiveness. Ad hoc distribution activities, planned and carried out by non-logisticians, and a short required response time, often mean that the nature of transport movements is inefficient. For instance, leftover food from restaurants is redistributed in poorly utilized passenger cars. As a result, the positive social impact might come at the price of increased environmental pollution. While the negative environmental impact of logistics activities is often an obvious sacrifice to make in order to alleviate the suffering of vulnerable people (e.g., when airfreight is used in humanitarian logistics to reach people affected by a natural disaster as quickly as possible), in the longer term the lack of economic, environmental, and social balance in logistics activities may put the whole initiative in jeopardy. Evidence shows that several well-intended sustainability initiatives in city or humanitarian logistics, for example, have been abandoned due to the inadequate economic performance or lack of basic logistics skills and knowledge. Logistics systems, even when designed with a grand sustainability objective in mind, need to be economically viable to survive and a healthy profit provides a basis for further development and upscaling potential. The highly skilled, well-

trained, flexible, and resilient logistics managers of the future are needed not only in businesses, but also in municipalities, hospitals, and nonprofit organizations such as charitable or humanitarian agencies.

The way we see logistics and the knowledge researchers develop within the field is constantly evolving in response to demographic and social pressures, the changing nature of the demand and the type of logistics services required, advances in data collection and processing, and social media and communication protocols. However, as environmental and social sustainability are definitely going to stay at the top of the social and political agendas, more and more attention needs to be given to CSR considerations in logistics and supply chain management operations.

Acknowledgments

The authors thank “the Kamprad family foundation for Entrepreneurship, Research & Charity,” for financing the project “Innovative sustainability performance in logistics supply chains,” in which this paper is a part.

Biographies

Maria Björklund is an Associate Professor of Logistics Management at the Department of Management and Engineering, Linköping University, Sweden. She received her PhD within the area of green logistics in 2005 from Lund University, Sweden. She researches, publishes, and lectures within the fields of sustainable supply chain management and logistics, with a focus on environmental and social considerations in logistics and purchasing, city logistics, logistics for sustainability, sustainable logistics-based business models, interorganizational interaction for sustainability, and green performance measurement. She is currently leading two research projects financed by the Swedish Energy Agency and the Kamprad family foundation for Entrepreneurship, Research & Charity.

Maja Piecyk-Ouellet is a Reader in Logistics at the University of Westminster. She is a former Deputy Director of the Centre for Sustainable Road Freight, an EPSRC-funded research center between Westminster, Heriot-Watt, and Cambridge Universities. Her research interests focus on the environmental performance and sustainability of freight transport operations. Much of her current work centers on city logistics, CSR, GHG auditing of businesses, and the forecasting of long-term trends in energy demand and environmental impacts of logistics. She is a Fellow of the Higher Education Academy. She is currently leading the University of Westminster’s input to the EPSRC-funded Freight Traffic Control 2050 project (EP/N02222X/1), the Volvo Centre of Excellence in Sustainable Urban Freight Systems (CoE-SUFS) and is carrying out research for the Centre for Sustainable Road Freight.

References

- Brundtland, G.H., Khalid, M., Agnelli, S., Al-Athel, S., Chidzero, B., 1987. *Our Common Future*. United Nations, New York.
- Carroll, A.B., 1979. A three-dimensional conceptual model of corporate performance. *Acad. Manag. Rev.* 4 (4), 497–505.
- Carter, C.R., Jennings, M.M., 2002. Logistics social responsibility: an integrative framework. *J. Bus. Logist.* 23 (1), 145–180.
- CEC, 2001. Green Paper: Promoting a European Framework for Corporate Social Responsibility. DOC 01/09. Commission of the European Communities, Brussels. Available from: https://europa.eu/rapid/press-release_DOC-01-9_en.htm.
- Kucht Campos, J., Alcântara Cardoso, P., Cunha Callado, A.A., Piecyk, M.I., 2018. The potential strategic role of logistics service providers in extending sustainability to the supply chain. In: McIntyre, J.R., Ivanaj, S., Ivanaj, V. (Eds.), *CSR & Climate Change Implications for Multinational Enterprises*. Edward Elgar Publishing, Cheltenham, UK.
- Milne, M.J., Gray, R., 2013. W(h)ither ecology? The triple bottom line, the global reporting initiative, and corporate sustainability reporting. *J. Bus. Ethics* 118 (1), 13–29.
- Piecyk, M.I., Björklund, M., 2015a. Logistics service providers and corporate social responsibility: sustainability reporting in the third party logistics industry. *Int. J. Phys. Distrib. Logist. Manag.* 45 (5), 459–485.
- Piecyk, M.I., Björklund, M., 2015b. Green logistics, sustainable development and corporate social responsibility. In: McKinnon, A., Browne, M., Piecyk, M., Whiteing, A. (Eds.), *Green Logistics, Improving the Environmental Sustainability of Logistics*. Kogan Page, London, pp. 3–26.
- Porter, M.E., Kramer, M.R., 2006. Strategy and society: the link between competitive advantage and corporate social responsibility. *Harv. Bus. Rev.* 84 (12), 78–92.
- United Nations, 2015. Transforming our world: the 2030 Agenda for sustainable development. Available from: <https://sustainabledevelopment.un.org/post2015/transformingourworld>.

Further Reading

- Perry, P., Towers, N., 2013. Conceptual framework development: CSR implementation in fashion supply chains. *Int. J. Phys. Distrib. Logist. Manag.* 43 (5/6), 478–501.
- Tyagi, M., Kumar, D., Kumar, P., 2015. Assessing CSR practices for supply chain performance system using fuzzy DEMATEL approach. *Int. J. Logist. Syst. Manag.* 22 (1), 77–102.

Information Sharing and Business Analytics in Global Supply Chains

Usha Ramanathan*, Ramakrishnan Ramanathan[†], *Nottingham Business School, Nottingham Trent University, Nottingham, United Kingdom; [†]University of Bedfordshire Business School, Luton, Bedfordshire, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	71
Collaboration for Supply Chains and Logistics	72
Role of Information and Business Analytics	72
Level of Collaboration Among Supply Chain Partners	72
The Ordering Pattern—Traditional Approach	74
Environmental Sustainability and Logistical Constraints	74
Conclusions	75
References	75

Introduction

In recent years, many global businesses have been involved in collaboration to improve overall performance through collaborative mandates. The initiatives of Walmart in collaborative planning forecasting and replenishment (CPFR) have attracted many businesses around the globe. Subsequently, several retailers and manufacturers, such as Procter and Gamble, Sara Lee, Nabisco, etc., have adopted CPFR (Ireland and Crum, 2005; Seifert, 2003). However, there is still no clear evidence on the benefits of collaboration, information exchange, or business analytics for the participating companies in specific contexts. One of the main reasons behind this ambiguity is the paucity of case studies documenting experiences and success stories.

Supply chain research on information exchange (Barbosa et al., 2018) focuses mainly on two areas. One is production and replenishment, and the other is forecasting and planning. Either cost reduction or inventory control is the primary motive of these supply chain collaborations. Initially at the inception of CPFR, an understanding of the collaboration process and the framework to collaborate were considered the two basic requirements for a collaborative supply chain. At a later stage, information sharing was recognized as one of the key elements of success for collaborative relationships (Ramanathan, 2012) and it was asserted that the benefit of information sharing is dependent on two factors: one is content and the other is proper use of information. Distorted information and inefficient use of available data will lead to excess inventory in each level of the supply chain. Hence, it may be important for the companies under collaboration to decide on what information is to share to reduce the cost of logistics and inventory, to create more accurate demand forecasts, to make production flexible, and to achieve timely replenishments.

In the 21st century, information in collaborative supply chains is being used to make joint business decisions. Business analytics involving in-depth data analysis are becoming integral parts of leading businesses in order to plan production and distribution. In recent times, particularly after the introduction of 3G and 4G technology, many logistical services namely DHL, FedEx, Amazon, Ocado, and many others around the globe are connected with warehouses, production plants, and service centers for inventory data, order data production, completion schedules, delivery schedule plan, etc. This connection is made available through technology, such as Internet of Things (IoT) sensors, Radio Frequency Identifications (RFID), bar codes, social media platforms, cloud computing, and Big data analysis.

From the literature, it is evident that the importance of sharing demand information with all supply chain members (both upstream and downstream) can help reduce the manufacturer's supply chain cost, specifically reducing the inventory costs of both supplier and customer (Ramanathan, 2012). The manufacturer could reduce the variance in demand forecast, if readily available historical order data and current demand data are used efficiently for decision-making. Walmart has set an example for retailers to practice transparent information sharing. Especially, the sharing of demand information with manufacturers or suppliers has been widely encouraged by Walmart under CPFR adoption in the USA. Point of sale (POS) data and market-data were found influential in achieving forecast accuracy in the "augmented CPFR" model (Chang et al., 2007). Recognizing the type of information to be shared among supply chain members to build visibility is a big challenge in achieving collaboration. In some businesses, under high demand variability, long-term forecasts performed better than short-term forecasts. Under low demand variability, short-term forecasts performed better than long-term forecasts. Previous research using store level stock keeping unit (SKU) data, found that simple time series forecasting will be appropriate for normal sales without promotions, but we need advanced techniques to improve forecast accuracy of promotional sales and logistical planning. Several examples on the benefits of sharing supply chain information and business analytics in managerial decision-making have been reported in professional journals and magazines, such as Focus from CILT and Professional Manager from CMI.

The main objective of this chapter is to discuss the role of supply chain information and business analytics in collaborative logistics planning, inventory management, and distribution. The rest of this chapter is organized as follows: Section, "Collaboration for Supply Chains and Logistics" describes collaboration among supply chain partners both upstream and downstream with respect

to the role of information sharing, business analytics, order generation and replenishment. This section also touches upon sustainability aspects in logistical collaborations. Section, “Conclusions” concludes with possible future research notes.

Collaboration for Supply Chains and Logistics

Role of Information and Business Analytics

While we appreciate the role of data analytics in business decision-making process, we also understand some negative effects of business analytics on the decision making of leading companies. For example, Tesco has used customer data extensively by means of their customized software and has used club card data for customer analysis to distribute gift coupons and discount vouchers. This attracted many customers for nearly half a decade and then this promotional approach slowly lost its power of retaining loyal customers. This sheds a light on the decision making of leading companies who used the power of data; conquering the competition with data analysis may not be successful for companies if they compromise collaborative decision-making and innovations.

In a collaborative framework such as CPFR, the point of sale (POS) information among organization is flowing mainly to forecast the demand. All partners will normally discuss various factors behind the sales figures, especially peaks and dips. This will help them to arrive at a single collaborative forecast figure that will be used to plan and control the production, warehousing, and inventory in line with the sales data, preferably the POS data. In Vendor Managed Inventory, solely suppliers will carry out the same exercise (Barratt and Oliveira, 2001). However, today’s competitive world requires better data storage and management in order to use business analytics to ensure a comprehensive business solution. Real time data analysis and business intelligence are an integral part of businesses to fulfill logistical requirements to gain customer satisfaction while fighting to survive the competition (Fig. 1).

In collaborative supply chains, all players will come forward to see the usefulness of the available information through multi-level discussions and formal agreements. For example, what information is required to maintain service level, inventory, production planning, and demand? Can in-depth data analysis lead to operational planning and forecasting? Will further managerial decision-making based on the analysis inform the business to improve their services? These details will help the businesses to decide the level of collaboration among partners for various purposes such as warehouse management, logistical services, and distribution planning (Ramanathan, 2011). The CPFR framework supports part of these processes, namely collaborative planning, forecasting, and replenishment, by sharing demand information from downstream partners. Fig. 1 represents the current business process enhancements in the presence of collaborative information exchange and the introduction of Business Analytics. This approach helps to inform collaborative decisions in fast moving supply chains. Global businesses such as Amazon and Alibaba manage their logistics, services (both delivery and warehouses) effectively through business analytics. Table 1 provides a list of important business analytics capabilities for modern logistics operations as compiled by the 2019 3PL Logistics study (3PL Study, 2019) (Table 1).

The inclusion of block chain in the list is an interesting addition to logistical capabilities and a true reflection of the inclusion of technology innovation in logistics. It has been reported that the IoT sensors, along with block chain capabilities, have the potential to revolutionize the way logistics firms operate (Porter and Heppelmann, 2014). For example, smart roads (using IoT sensors) and specially devised SatNavs with information on road quality, fuel efficiency, and local traffic are fitted in many trucks in the UK and Europe to make a quick decision on routes, considering the live traffic situation. The capabilities of IoT sensors and block chain technology are further enhanced using artificial intelligence techniques (e.g., smart contracts) and other decision support tools.

Level of Collaboration Among Supply Chain Partners

In order to have a higher level of support from supply chain partners, one of the essential factors is establishing a close relationship with partners. In the literature, it is argued that not all supply chain partners will have same level of collaboration (Ramanathan, 2011). Companies operating under high competition will need to have open information sharing and strong collaborative relationships with trusted allies in order to thrive and succeed. This is named as “Futuristic Collaboration.” The other two levels of collaboration namely “progressive collaboration” and “preparatory collaboration” are suggested by Ramanathan (2011). Here,

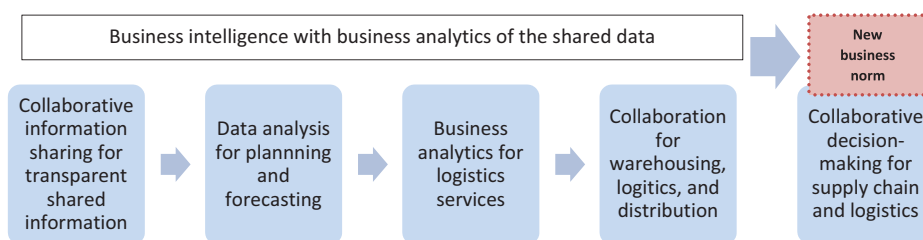
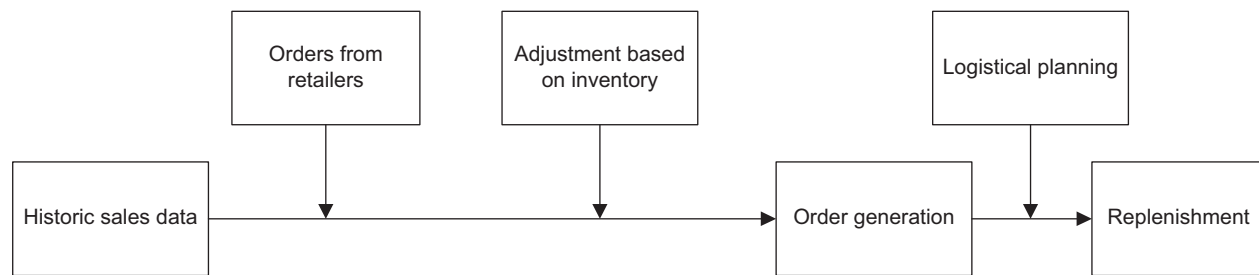


Figure 1 Role of collaborative information in supply chain logistics and distribution.

Table 1 IT-based analytics capabilities for logistics

Transportation management (planning)
Warehouse/distribution center management
Visibility (order, shipment, inventory, etc.)
EDI data interchange—orders, advanced shipment notices, updates, invoicing
Transportation management (scheduling)
Transportation sourcing
Global trade management tools (e.g., customs processing and document management)
Network modeling and optimization
Bar coding
Supply chain planning
Web portals for booking, order tracking, inventory management and billing
Customer order management
Cloud-based systems
CRM (customer relationship management)
Advanced analytics and data mining tools
RFID
Distributed order management
Yard management
Block chain

Source: 3PL Study, 2019

**Figure 2** Order generation—traditional approach.

the choice as to the level of collaboration is highly dependent on type of business and partner requirements. For example, functional products with no fluctuation in their demand patterns will have a basic preparatory level of collaboration. However, in businesses with high variability in customers' demands due to the fast moving nature of the product, the level of collaboration needs to be intense, such as futuristic collaboration. This will allow companies to have open and transparent information exchange. The level of collaboration and order generations are closely related to each other (Baratt and Olivera, 2001; Ireland and Crum, 2005). For example, a vendor managed inventory (VMI) approach will make ordering straightforward with vendors in place and a CPFR will manage automated orders. However, "no collaboration" or "no information" will trigger orders only when the buyer is making efforts to place orders either through historic sales figures or on checking the inventory level. Fig. 2 represents the order generation in traditional supply chains. Here historic sales data was in use to make production, inventory, and replenishment plans. This required very basic collaboration such as "preparatory level" collaboration (Fig. 2).

In contrast, Fig. 3 represents supply chains of the 21st century in the presence of live data. Global businesses normally value all customers and attempt to serve them to the highest possible level. As the market becomes demand driven, the companies need to go for business analytics to see customers' preferences and market trends in order to survive the competition. This requires the involvement of live data in planning, forecasting, delivery, and decision-making; hence, business analytics has become a new norm for businesses (Fig. 3).

Many global companies that operate from various countries, such as DHL, FedEx and Amazon, have a wide range of customers from big national chain customers, such as Tesco, to small local shops and convenience stores. It is crucial to maintain a close and healthy relationship with all these customers. From the inception of the internet, the companies around the globe are using a variety of Windows-based solutions such as an "EPOS ordering system" or something similar to Walmart's "retail-link" to communicate with business and supply chain partners. This new system assists retailers to benefit from a full product file, pre-installed system, so that orders can be placed electronically to the suppliers, and encourages their customers to use Windows-based systems by valuing customers' participation in profit earning and timely replenishment.

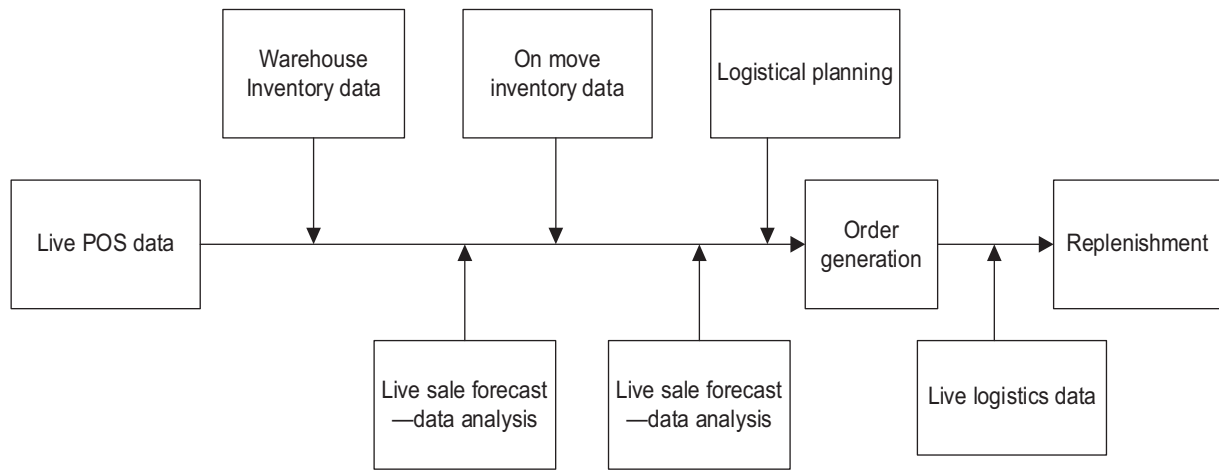


Figure 3 Replenishment using live data and business analytics.

The Ordering Pattern—Traditional Approach

Some companies still follow order generation through the “traditional method,” which is based on their respective order from the customers (Fig. 2). In order to avoid stock-outs, some companies maintain 5 days of safety stock for functional products, such as coffees and beverages. Initially, demand analysts make their order forecasts by looking at historic sales data and adjusting this based upon order input from the retailers. Based on order forecasts and current inventory levels, retailers, or wholesalers place orders to suppliers. On receiving orders from downstream partners, the supplier plans its replenishment. This process usually takes place for a total of 3 days, where Day 1–Day 3 delivery signifies that:

- Day 1—placing order, deciding site from where order has to be replenished and deciding plan for dispatch.
- Day 2—moving necessary stock in between the warehouses to gain cost benefits.
- Day 3—delivery of order.

This ensures timely replenishment at retail outlets, as long as the outlets are located within 100 miles. The above stated procedure adopted by the supplier for order generation is a simple and straightforward method, but does not consider potential complications. The sales team of the company will normally use historic sales data to place orders with suppliers. Orders from retailers and inventory status are incorporated in the process of order generation. The supplier replenishes orders in 3 days. Some of the existing traditional operational problems can potentially be improved on the back of the mutual understanding of partners and on streamlining the current system using data analytics and collaborative decision-making.

Environmental Sustainability and Logistical Constraints

Some global companies have started encouraging bulk orders of one full truckload in order to maintain sustainability objectives. This quantity discount on bulk orders is hoping to provide customers with some extra time to order-up-to fill level. Obviously, customers need to wait to order until their order comes closer to the fill level of one truckload, if the company wants to enjoy a quantity discount. Initially, this approach has complicated the global company’s delivery to their customers; slowly, the companies have started learning how to make a profit on ordering a full container or full truckload.

In the case of ordering less than a full container or full truckload, companies try the other option of using its own vehicles to obtain delivery from the supplier. Some other companies with no investment capability have also started using 3PLs. A few companies have experimented with multipurpose/multi compartmental vehicles carrying a variety of products. Unfortunately, this has not proved successful as these vehicles have fixed space and hence it is difficult to accommodate unplanned products and deliveries. Later this idea was dropped in the international market due to its operational difficulties.

The introduction of online sales has added some extra logistical issues for retailers and for producers around the globe. However, e-tailing has shown its success in a variety of ways by increased customer satisfaction through logistical and service excellence despite the difficulties.

Any broken link in the supply chain increases lead-time and service performance. However, it can be concluded that the sharing of demand forecasts with suppliers will reduce uncertainty and, hence, could improve inventory positions. Information sharing and collaboration for data analytics has proved to have led to a higher level of precision in decision-making, as shown in the case of Walmart.

Conclusions

This chapter has highlighted, with some practical examples, how the power of business analytics can be combined with transparent information sharing in collaborative supply chains. If this study is extended for many global companies operating in different industrial sectors, such as pharmaceuticals, automobiles etc., a general opinion on the type of information and Business analytics can be suggested. It will also be beneficial to look at other possibilities for information exchange among wholesale companies and upstream partners.

References

- [3PL] Study, 2019. 2019 Third-party logistics study—The State of Logistics Outsourcing: Results and Findings of the 23rd Annual Study, Multiple authors. Available from: www.3plstudy.com.
- Barbosa, M.W., Vicente, A.D.L.C., Ladeira, M.B., Oliveira, M.P.V.D., 2018. Managing supply chain resources with Big Data Analytics: a systematic review. *Int. J. Logistics Res. Appl.* 21 (3), 177–200.
- Barratt, M., Oliveira, A., 2001. Exploring the experiences of collaborative planning initiatives. *Int. J. Phys. Distribution Logistics Manage.* 31 (4), 266–289.
- Chang, T., Fu, H., Lee, W., Lin, Y., Hsueh, H., 2007. A study of an augmented CPFR model for the 3C retail industry. *Suppl. Chain Manage. Int. J.* 12, 200–209.
- Ireland, R.K., Crum, C., 2005. *Supply Chain Collaboration: How to Implement CPFR and Other Best Collaborative Practices*. J. Ross Publishing Inc, Florida.
- Porter, M.E., Heppelmann, J.E., 2014. How smart, connected products are transforming competition. *Harvard Bus. Rev.* 92 (11), 64–88.
- Seifert, D., 2003. *Collaborative Planning Forecasting and Replenishment: How to Create a Supply Chain Advantage*. AMACOM, Saranac Lake, NY, USA.
- Ramanathan, U., 2011. *Role of Information Exchange in Collaborative Supply Chains: Analysing the Role of Information Exchange for Demand Forecasting in Collaborative Supply Chains*. LAP Lambert Academic Publishing, Saarbrücken, Germany.
- Ramanathan, U., 2012. Supply chain collaboration for improved forecast accuracy of promotional sales. *Int. J. Oper. Prod. Manage.* 32 (6), 676–695.

Logistics Information Systems

Petri Helo, Javad Rouzafzoon, Networked Value Systems, School of Technology and Innovations, University of Vaasa, Vaasa, Finland

© 2021 Elsevier Ltd. All rights reserved.

Introduction	76
Enterprise Planning Systems	76
Transport Management Systems	77
Warehouse Management Systems	78
Tracking and Tracing Systems	80
Advanced Supply Chain Planning Systems (APS)	81
Modeling and Simulation	82
Conclusions	82
Relevant Websites	83
References	83

Introduction

Logistics information systems deal with monitoring and control of information flows, material flows, and financial transaction-related flows within companies and supply chains. The purpose of the logistics information systems is to provide decision-making support, automated transactions, and data storage services for organizations to deliver the right quantities of goods at the right place and right time (Helo and Szekely, 2005).

Information systems in business logistics are gaining increasing importance as competitive advantage of supply chains is very often based on reliable, fast-response operations. Handling material flows within and between companies takes place across several functions, for example, (1) intra-logistics: procurement, production, shipping, and inventory handling functions, and (2) logistics occurring within the network: transportation, e-business, and supply chains.

Logistics information systems very often consist of various independent, but interconnected, information systems (Fig. 1). A central system for any company is enterprise resource planning (ERP) software that keeps records of the key business and material handling transactions of the company in a central location. Advanced planning and scheduling (APS) software packages operate on a network level that often consists of several enterprises and aims to optimize supply chain operations. Warehouse management systems (WMS) are handling information at the warehouse level and are often connected to operational automation systems. Transport management systems (TMS) are used for the planning and control of transportation events, route planning, and scheduling. Tracking and tracing software systems provide a real-time or close to real-time visibility for supply chain actors, in order to answer the questions of the origin of the items and where these are currently located. Business intelligence and big data analytics are bespoke systems built to collect information from various sources and to provide up-to-date performance indicators for each part of the operations (Addo-Tenkorang and Helo, 2016).

Enterprise Planning Systems

ERP systems have evolved from material requirements planning software to cover the entire integrated operations of business transactions in the company, including purchasing orders, sales orders, production planning, inventory management, finance, and human resources. ERP systems are based on a centralized database and client server architecture, which allows users to access current information in real-time (Kurbel, 2016). This integrated and centralized approach helps companies to handle automated processes between functions. For example, automated purchase handling can generate purchase orders for a vendor based on inventory level and triggering rules; goods received can be linked to purchase order number and the stock updated at the same time approving payment to the vendor.

ERP software functionality could include the following areas:

- Sales and distribution: offers, orders, pricing.
- Purchasing: quotations, vendor management, purchase orders.
- Production: production planning, master production scheduling, forecasting, material requirements planning, bill of materials, capacity management, quality control.
- Financial accounting: cost accounting, general ledgers, cash management, invoicing.
- Warehousing: goods receiving, shipping.
- Human resources: payroll, training records, recruiting.
- Project management: cost management, scheduling.

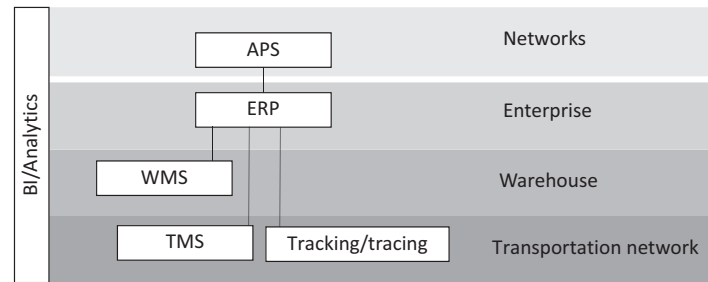


Figure 1 Logistics information systems in planning hierarchy.

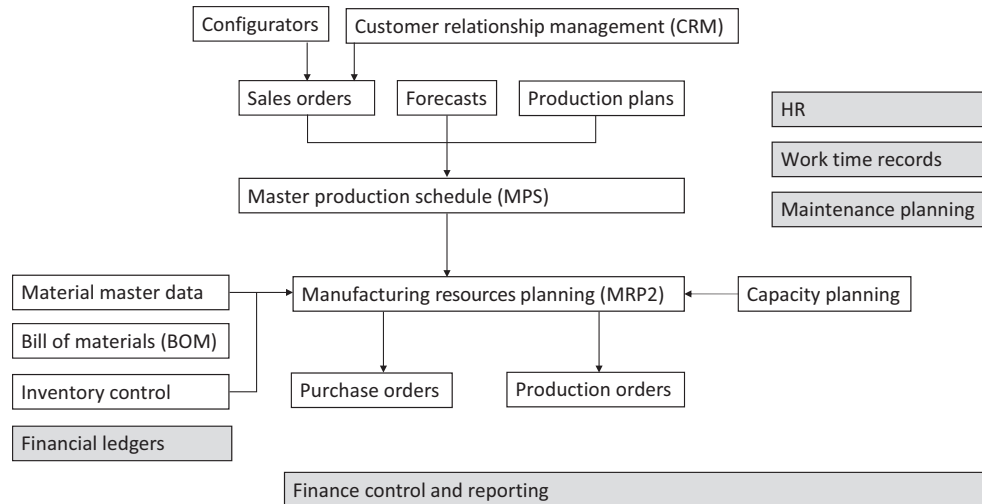


Figure 2 Enterprise resource planning (ERP) system functionality is linked with business processes.

These functionalities are interconnected and provide the master data for important logistics transactions—including product data, customers, vendors, warehouse data—that is stored in a single up-to-date location available for different needs (Fig. 2).

Many ERP systems are based on industry best practices and standardized configurable business processes. Despite the differences, ERP systems share similarities as to how software modules are organized and how processes are built. Fig. 3 is a screenshot capture of Odoo ERP, an open source software. This figure shows possible applications (modules) available for the use of organizations.

SAP, Microsoft, and Oracle are some of the leading software ERP providers, but a great share of the ERP market is controlled by several smaller national vendors, which have industry-specific solutions or good local support for localized needs. Currently, ERP solutions are moving from on-premises hosted solutions to new cloud-based systems. This allows the building of better connectivity in supply chains. Typical use cases are portal solutions for the customers and vendors. Extending the scope of ERP beyond the enterprise is one of the leading themes in this area.

Transport Management Systems

TMS maintain an efficient delivery of materials from source to destination. The primary purpose of TMS is to schedule and optimize the transportation routes, track the fleet in real-time, minimize the transportation cost and delays, and improve logistics reliability (Barreto et al., 2017).

TMS is a component of supply chain management (SCM) that enables collaboration with an order management system (OSM) and distribution center (DC) or warehouse. Companies of all sizes have been employing TMS, particularly following the recent development and growth in TMS, to handle the high freight cost, coordinate with other supply chain technologies such as WMS and Global Trade Management System and to manage the electronic communication with customers, trade partners, and carriers (Barreto et al., 2017).

TMS take care of the daily planning and execution of transportation tasks. The TMS system typically received orders from an ERP system and is integrated with WMS as well. The functions of a TMS system are often tied with certain types of freight modes and can include the following capabilities:

- Transportation vendor management
- Freight invoicing

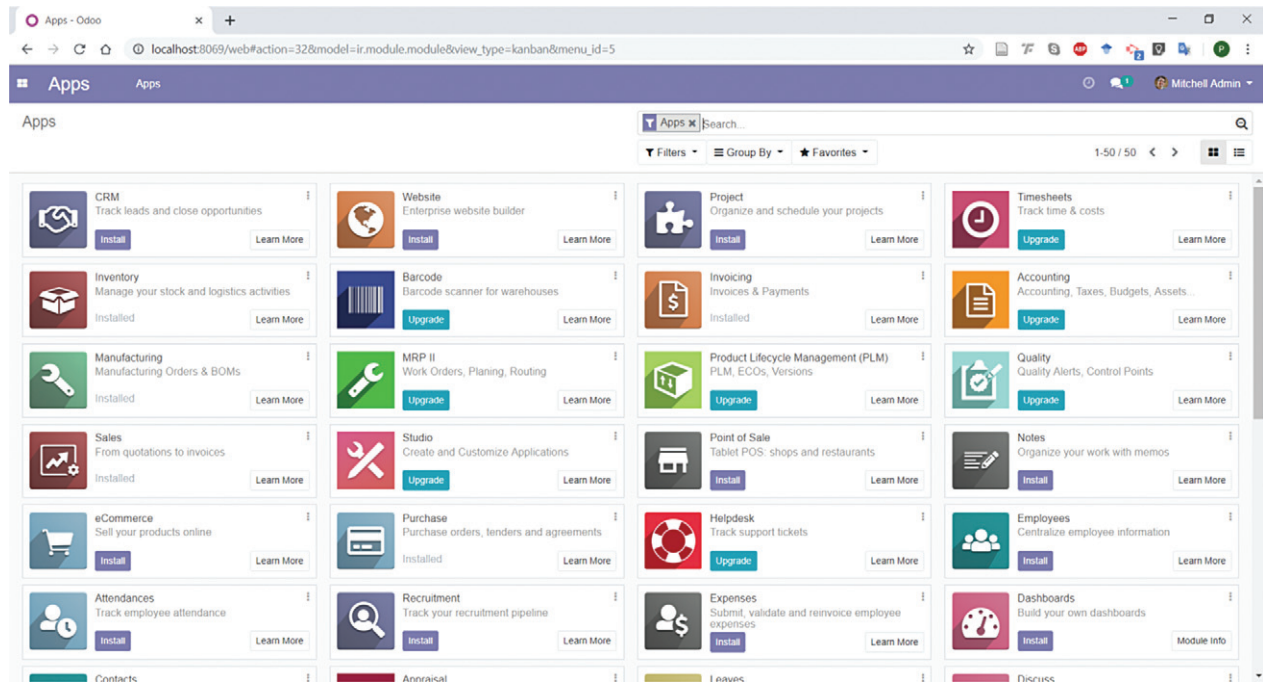


Figure 3 Odoo ERP application menu.

- Route planning
- Load and unload planning
- Asset management of vehicles, trailers, and containers
- Yard management systems
- Cost control and reporting
- Transportation tracking
- Quality management and transportation claims processing

A TMS system enables companies to accurately locate their fleets while in operation through GPS technology, oversee freight movements, negotiate with carriers, consolidate freights, and collaborate with intelligent transportation systems (ITS). On the planning side, TMS systems often combine a geographical information system (GIS) with an optimization system such as vehicle routing problem solvers which determine the daily planning based on cost minimization or according to a similar system-wide objective (Braekers et al., 2016). Fig. 4 illustrates Open Door Logistics, an open source software used to plan daily delivery routes and schedules. The optimization algorithm provides a solution for the relatively complex problem in less than a minute.

The Internet of Things (IoT) and TMS play a critical role in the transportation and logistics industries. Since more and more physical objects are equipped with sensors, barcodes and RFID tags, logistics companies can perform real-time monitoring of physical objects from an origin to a destination throughout the entire supply chain. On the other hand, vehicles are empowered these days with sensing, networking, communication, and data processing competencies, and IoT technologies can be utilized to improve logistics operations and fleet utilization (Atzori et al., 2010).

TMS are important sources of competitive advantage for companies operating in transportation. Large logistics companies such as DHL and FedEx have invested a lot of resources to optimize their transportation network by using advanced planning systems (APS) that aim to improve capacity utilization, minimize total costs, and reduce emissions. Industry-specific solutions have been built for air-based logistics (Feng et al., 2015), for maritime logistics (Helo et al., 2018), as well as for road transportation (Dibbelt et al., 2017). The possibilities to use various optimization algorithms have been increased as cloud-based solutions enables low-cost computing. The interest in autonomous vehicles and drones (Wang et al., 2017) is expected to drive this sector with artificial intelligence-related investments.

Warehouse Management Systems

A WMS is a database-driven computer application to enhance warehouse efficiency and retain accurate inventory through the recording of warehouse transactions (Ramaa et al., 2012). WMS are employed by various companies to obtain transportation and production economies, fulfill changing market conditions and uncertainties, achieve the minimum logistics cost corresponding to the desired level of customer service, and offer a combination of products instead of a single product per order.

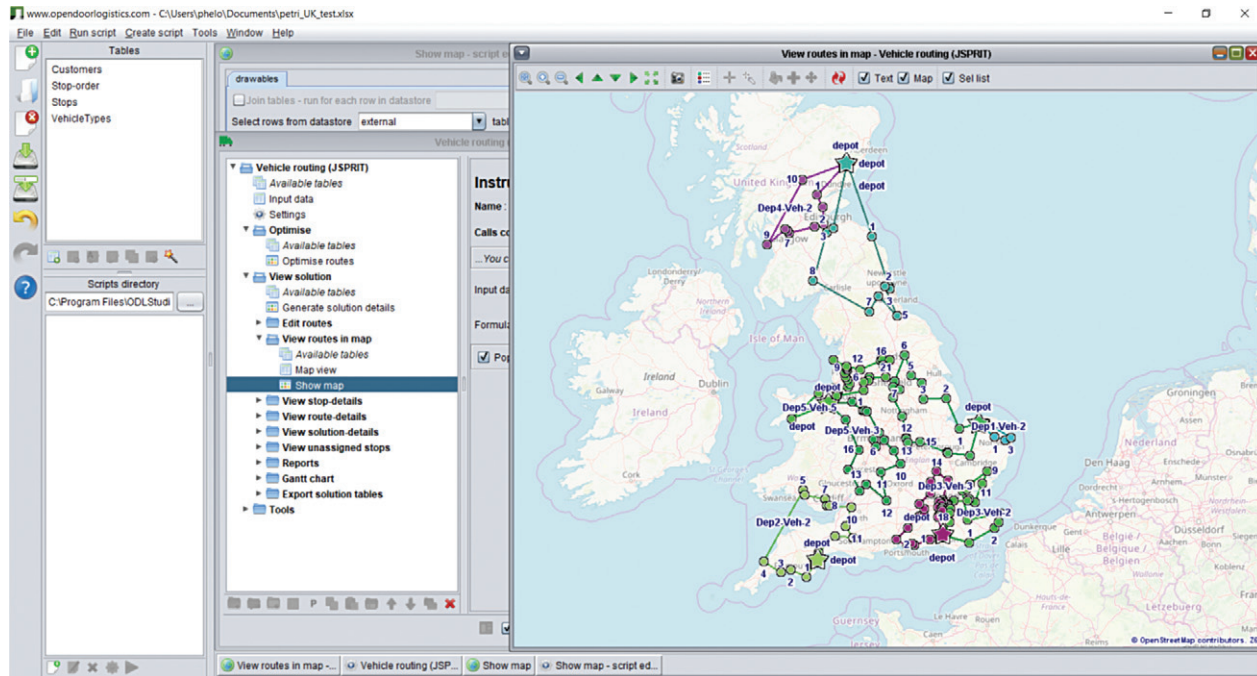


Figure 4 Open Door Logistics software used for daily vehicle route planning.

WMS are software solutions used to conduct warehouse and distribution-related planning and execution tasks. These systems are often linked to certain types of warehousing system and increasingly linked with automation. The systems take care of planning, resourcing, and managing tasks to receive, move, and ship materials to and from warehouses. The controlled tasks can include:

- Goods receiving
- Order picking
- Packing
- Goods issue and dispatch
- Inventory management, storage locations, detailed stock keeping units (SKU)

In addition, a WMS can determine a temporary storage for disposable or recyclable materials, provide an intermediary location for transshipments, and maintain the just-in-time programs of suppliers and customers. WMS are a crucial component of a company's logistics system and contains four procedures: receiving products, storing the materials until they are requested, picking products from inventory, and shipping the products in response to customer orders. The order picking process is the primary function of a warehouse management system and lies at the interface between production and distribution systems. Order picking policy comprises basic order picking, zone picking, batch picking, and wave picking. In a basic order picking policy, pickers conduct a tour through the warehouse to collect all SKUs for a single order, while under a zone picking policy, a warehouse is divided into zones and pickers gather SKUs from a single zone. In batch picking, an order picker is restricted to a specific zone of a warehouse and simultaneously collects orders in order to reduce the mean trip time per pick. However, orders need to be sorted subsequently. In wave picking compared to zone picking, several workers pick different items of the same order at the same time from all zones.

The possible automation systems are often built around standardized material handling units, for example, plastic crates or pallets. The system layout and configuration may be a combination of the following automation building blocks:

- Automatic storage and retrieval systems (AS/RS), high-bay warehouses
- Conveyor belt-based systems
- Pick-to-light, voice picking, or other computer-guided human-based warehouse system
- Automated-guided vehicles (AGV)
- Robotic bin-picking systems (Huang et al., 2015)

Fig. 5 illustrates a screenshot capture of Open WMS systems example layout, which combines a conveyor belt-based automation system with multiple aisle collection points and manual operations. In this example, the WMS controls the procedures and guides each station operator.

The order picking systems are classified based on utilizing humans or automated machines to three system types, picker-to-parts, parts-to-picker, and a put system. In the picker-to-parts systems, which is the most common method in order picking, order pickers

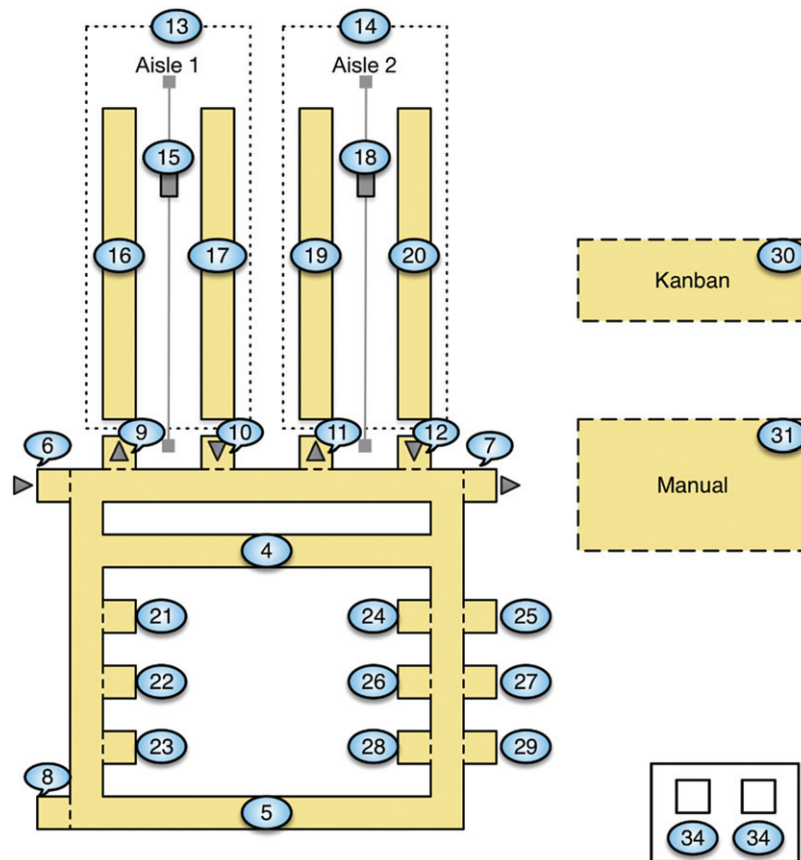


Figure 5 Open WMS user layout with conveyor belts and collection points.

collect the items by walking or driving through the aisles. The parts-to-picker system employs AS/RS, which is majorly aisle-bound cranes retrieving one or more orders and delivering them to the pick location. The put system comprises both retrieval and distribution processes, where items are collected according to a parts-to-picker or picker-to-parts type in the retrieval process of the put system. An order picker afterward dispenses the carrier with the pre-picked items over customer orders.

Large logistics operators have invested a lot in advanced warehouse automation. Companies such as Amazon, which acquired Kiva Robots, and Alibaba are well-known examples of supply chain operators who are driving the technological developments (Li and Liu, 2016). Efficient and flexible automation is expected to deliver value for the entire supply chain.

Tracking and Tracing Systems

The identification, real-time tracking, and traceability of materials in supply chains have all been a demanding challenge for companies. Constant monitoring and handling rising shipment chains is vital in complex supply chains and logistics management. Various tracking and tracing solutions have been introduced to visualize where products are (tracking) and where they are originating (tracing).

In their simplest form, tracking systems are based on identification systems such as bar codes and processes supporting reading transactions into a centralized database. RFID is technology that has improved the automatic identification and data capture (AIDC) technologies (Angeles, 2005). RFID systems comprise three components: the tag or transponder, the reader, and the data processing subsystem containing both middle-ware and internal databases. RFID tags are active or passive. Active tags have internal batteries and can be utilized for longer distances since these tags are not dependent on a nearby field to transfer or obtain signals. Passive tags do not contain a battery or power source. RFID utilizes radio frequency waves to transmit the information between a reader and an item required to be identified, tracked, or located. RFID technology can identify objects without line of sight and with a greater reading distance compared to barcode technology. In addition, in contrast to barcode technology, RFIDs can store more data, maintain larger sets of unique IDs, and detect many tags at once in a single area, without human assistance. RFID reading systems can detect, scan, trace, and monitor items attached with RFID tags within the defined proximity of the reader. In addition, the IoT is an evolution in computer technology and communication with the purpose of linking items through the internet and RFID is one of the essential components of the IoT (Sarma et al., 2002; Thornton et al., 2006).

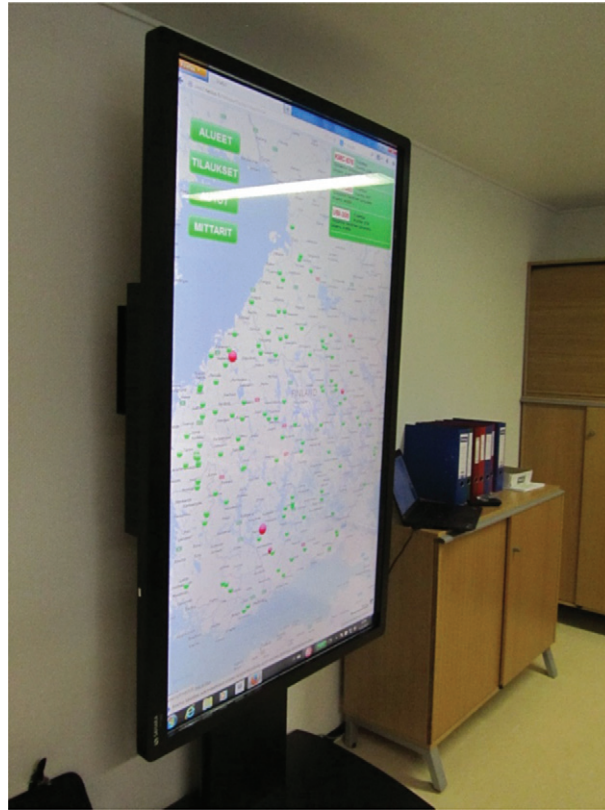


Figure 6 Tracking view for fleet of delivery maps on GIS.

The authenticity of the product and the absence of any tampering with supply chain transactions are essential for products in the food industry and healthcare. Blockchain is an emerging technology which is solving this problem by immutable distributed ledgers which are verified by a large number of server computers simultaneously (Helo and Hao, 2019).

Transportation vehicles are also increasingly being tracked. The automatic identification system (AIS) is an imperative collision avoidance system required by the international maritime organization (IMO) to be installed on ships. AIS devices transfer data to stations such as harbor authorities, equipment such as buoys, or search and rescue airplanes. Every individual vessel utilizing AIS devices is identified by a unique Maritime Mobile Service Identify (MMSI) number and IMO registry number. The data transition intervals can be 3 s to a few minutes and two VHF channels are employed to transfer the AIS information. The initial use of AIS in maritime safety has been extended to support logistics needs (Chen et al., 2016; Last et al., 2017). Similar solutions are available for commercial airfreight. Also, many road transportation companies have built full-scale fleet tracking of trucks and trailers within their own assets. Real-time tracking and tracing allow fast response on adjusting plans and updating estimates for supply chain actors. Fig. 6 is a tracking monitor used for the planning and control of a truck fleet. Real-time truck location is combined with a planned delivery schedule and deviations between planned and actual are constantly monitored.

Advanced Supply Chain Planning Systems (APS)

APS are information systems that utilize advanced solution approaches such as operations research, metaheuristics such as genetic algorithms or computer simulations to resolve supply chain planning problems. An APS must interact with enterprise systems (ES) to accomplish a synchronized workflow. APS is not a replacement for human planners, however, as the users of APS can always approve, adjust, or deny the outcome generated by an advance planning system. An APS is a computer planning system that delivers various function of SCM such as procurement, production, distribution, and sales at the strategic, tactical, and operational planning level (Stadtler, 2005; Vidoni and Vecchiotti, 2015). Fleischmann et al. (2005) described the characteristics of an APS as follows:

- *Integral planning:* APS plans for the entire supply chain and integrates from suppliers to customers of a single company or even a network of enterprises.
- *True optimization:* APS improves solutions by setting alternatives, objectives, and constraints for various planning problems and utilizes optimization planning approaches, such as deterministic and heuristics methods.

- *Hierarchical planning systems*: APS employs hierarchical planning systems which is a concession between practicability and consideration of interdependencies between the planning tasks. The APS contains different planning levels or modules incorporating decisions at long-term, mid-term, and short-term decision levels (Fleischmann and Meyr, 2003).
- *Strategic network planning (SNP)*: It comprises the quantitative elements of strategic planning and aspects of network design, such as facility location, stock size, or production capacities.
- *Demand planning (DP)*: It forecasts the future demand in a make-to-stock environment on mid-term aggregate and short-term basis.
- *Master planning (MP)*: It integrates the material flow of the supply chain as a whole for a mid-term planning horizon.
- *Production planning and scheduling (PP&S)*: It encompasses lot-sizing, machine assignment, scheduling, and sequencing in the short-term planning horizon.
- *Distribution Planning (DtP)*: It deals with mid-term tasks within distribution systems, such as regular transport links, delivery areas of warehouses, and resource allocation.
- *Transport planning (TP)*: It covers the short-term dispatching of shipments in the distribution and procurement side.
- *Demand fulfillment and available to promise (ATP)*: It concerns the arriving of customer orders and comprises the tasks of order promising, checking the availability of materials, due date setting and actions in case of shortage.

Modeling and Simulation

Utilizing simulation models for complicated systems has become more popular and practical. The simulation approach captures the behavior of all units, the interactions between entities, and the uncertainties associated with systems. Simulation methods can be classified into three types: (1) system dynamics, (2) agent-based modeling, and (3) discrete-event simulation (DES).

The system dynamics methodology includes developing the causal diagrams and policy-oriented simulation models that are unique to each problem setting. The main principle of system dynamics is that the complex behavior of organizational and social systems are derived from the balancing of feedbacks, reinforcing feedbacks, and the constant accumulation of people, materials, financial assets, information, and even psychological or biological states. A completed model may contain scores or equations alongside numerical inputs. Modeling is an iterative process of scope selection, hypothesis generation, causal diagramming, quantification, reliability testing, and policy analysis (Homer and Hirsch, 2006).

Agent-based modeling employs a bottom-up modeling approach and is considered a valuable method for decision support in SCM. The most distinguishing aspect of agent-based modeling simulation (ABMS) is the agent perspective, derived from a view that every system consists of agents. In ABMS, agents are individual entities and have diverse characteristics. In addition, agents have internal behaviors, allowing them to act autonomously, sense any conditions, which occur within the model at any time, and react with the appropriate response. Autonomous agents are also able to interact with other agents and the environment. In addition, the adaptive ABMS is a modeling method in which agents can change their behavior during the simulation as they learn (Macal, 2016; Rouzafzoon and Helo, 2018).

DES has been the mainstay of the operational research (OR) simulation community. DES models are process-oriented and employ a top-down modeling approach. The focus is on modeling the system in detail and is not on the entities. Queues are considered the core components of modeling and entities flow through a system. In addition, the attributes of entities are changed while they transfer through a system. DES was originally employed to resolve the uncertainties inherent in real systems and to deal with those complicated scenarios that are too complex to be resolved by mathematical methods. DES has been vastly implemented in the manufacturing domain. DES is an operational research technique that allows the users to evaluate the efficiency of the current system, assess various scenarios, and devise a new system. In addition, a DES can be employed to analyze resource utilization, to forecast the impact of modification on the flows of entities, and to examine the complex connections between different variables of a model (Jun et al., 1999; Siebers et al., 2010; Wang et al., 2018).

Fig. 7 shows an agent-based model of the transportation schedules of a fleet of collection trucks taking orders from four factories under dynamic demand. Various complex scenarios related to capacity planning and extreme conditions can be modeled in a relatively small development time.

Conclusions

Digital technologies are the salient drivers of supply chain development. Logistics information systems connect the top layers of SCM to execute daily tasks and interconnect various actors in the networks to access the same real-time information. Currently, the main technological drivers are:

- Cloud-based software applications taking place from on-time premises software by introducing software as a service (SaaS).
- Increasing use of artificial intelligence techniques for solving problems related to mathematical optimization, such as warehouse optimization, APS systems, and the vehicle route planning used in TMS.
- Use of big data and analytics is enabling organizations to make automated logistics decisions in real time (Waller and Fawcett, 2013).

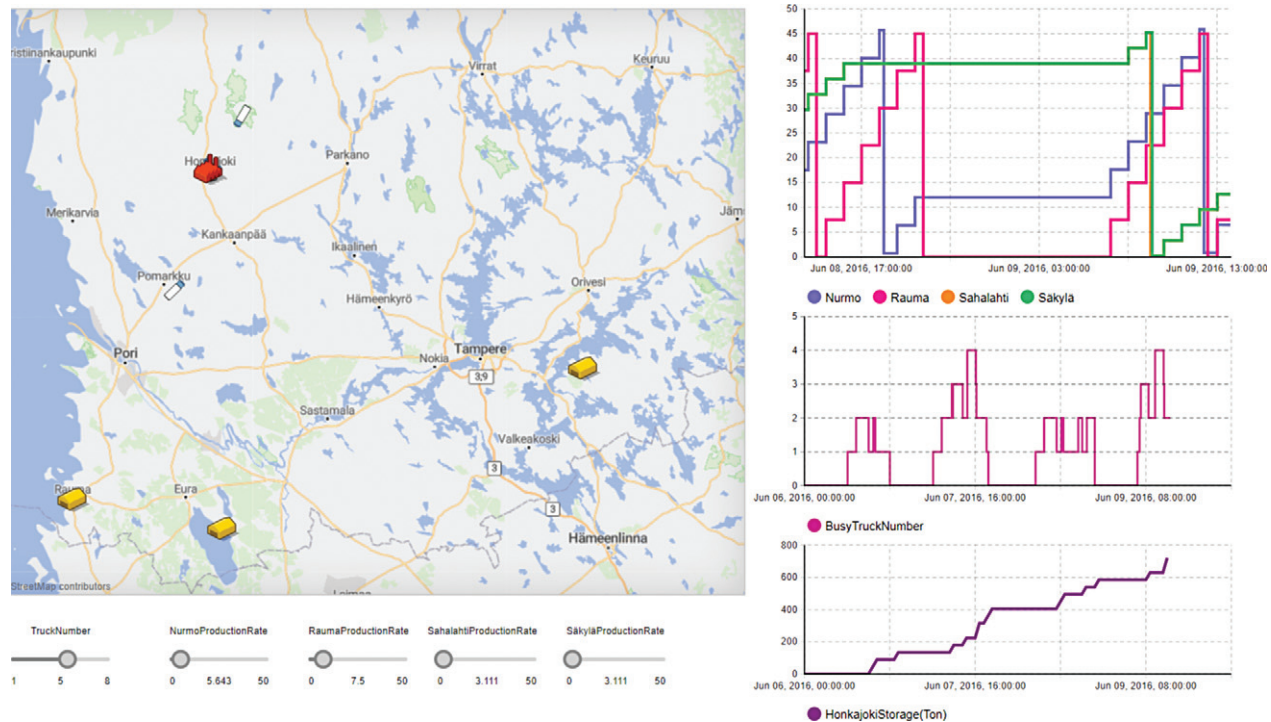


Figure 7 Agent-based simulation of transportation schedules of a fleet of collection trucks (AnyLogic software).

- Tracking and tracing is becoming a requirement of logistics applications: blockchains are used to guarantee the authenticity, with various IoT systems providing information from the entire transportation network.
- Robotics and automation are interlinking information management systems with the execution of the tasks.

Relevant Websites

Odoo: Open Source ERP and CRM. Available from: <https://www.odoo.com/>

Open Warehouse Management System—An Open Source Software to Build Intralogistic Projects With. Available from: <https://openwms.github.io/org.openwms/>

Open Door Logistics—Intelligent software for vehicle routing and territory mapping/management. Available from: <https://www.opendoorlogistics.com/>

References

- Addo-Tenkorang, R., Helo, P.T., 2016. Big data applications in operations/supply-chain management: a literature review. *Comput. Ind. Eng.* 101, 528–543.
- Angeles, R., 2005. RFID technologies: supply-chain applications and implementation issues. *Inf. Syst. Manag.* 22 (1), 51–65.
- Atzori, L., Iera, A., Morabito, G., 2010. The internet of things: a survey. *Comput. Netw.* 54 (15), 2787–2805.
- Barreto, L., Amaral, A., Pereira, T., 2017. Industry 4.0 implications in logistics: an overview. *Procedia Manuf.* 13, 1245–1252.
- Braekers, K., Ramaekers, K., Van Nieuwenhuysse, I., 2016. The vehicle routing problem: state of the art classification and review. *Comput. Ind. Eng.* 99, 300–313.
- Chen, D., Zhao, Y., Nelson, P., Li, Y., Wang, X., Zhou, Y., Lang, J., Guo, X., 2016. Estimating ship emissions based on AIS data for port of Tianjin, China. *Atmos. Environ.* 145, 10–18.
- Dibbelt, J., Kehagias, D., Pantziou, G., Gavalas, D., Konstantopoulos, C., Wagner, D., Giannakopoulou, K., Kontogiannis, S., Zaroliagis, C., 2017. Eco-aware vehicle routing in urban environments. 2017 IEEE Symposium on Computers and Communications (ISCC). IEEE, Heraklion, Greece, pp. 208–213.
- Feng, B., Li, Y., Shen, Z.J.M., 2015. Air cargo operations: literature review and comparison with practices. *Transp. Res. Part C Emerging Technol.* 56, 263–280.
- Fleischmann, B., Meyr, H., 2003. Planning hierarchy, modeling and advanced planning systems. *Handbooks Oper. Res. Manag. Sci.* 11, 455–523.
- Fleischmann, B., Meyr, H., Wagner, M., 2005. Advanced planning. In: Stadtler, H., Kilger, C. (Eds.), *Supply Chain Management and Advanced Planning*. Springer, Berlin, pp. 81–106.
- Helo, P., Hao, Y., 2019. Blockchains in operations and supply chains: a model and reference implementation. *Comput. Ind. Eng.* 136, 242–251.
- Helo, P., Szekely, B., 2005. Logistics information systems: an analysis of software solutions for supply chain co-ordination. *Ind. Manag. Data Syst.* 105 (1), 5–18.
- Helo, P., Paukku, H., Sairanen, T., 2018. Containership cargo profiles, cargo systems, and stowage capacity: key performance indicators. *Mar. Econ. Logist.* 1–21, <https://doi.org/10.1057/s41278-018-0106-z>.
- Homer, J.B., Hirsch, G.B., 2006. System dynamics modeling for public health: background and opportunities. *Am. J. Public Health* 96 (3), 452–458, doi:10.2105/ajph.2005.062059.
- Huang, G.Q., Chen, M.Z., Pan, J., 2015. Robotics in ecommerce logistics. *HKIE Trans.* 22 (2), 68–77.
- Jun, J.B., Jacobson, S.H., Swisher, J.R., 1999. Application of discrete-event simulation in health care clinics: a survey. *J. Oper. Res. Society* 50 (2), 109–123, doi:10.1057/palgrave.jors.2600669.

- Kurbel, K.E., 2016. Enterprise Resource Planning and Supply Chain Management. Springer-Verlag, Berlin.
- Last, P., Kroker, M., Linsen, L., 2017. Generating real-time objects for a bridge ship-handling simulator based on automatic identification system data. *Simul. Model. Pract. Theory* 72, 69–87.
- Li, J.T., Liu, H.J., 2016. Design optimization of Amazon robotics. *Automat. Control Intel. Syst.* 4 (2), 48–52.
- Macal, C.M., 2016. Everything you need to know about agent-based modelling and simulation. *J. Simul.* 10 (2), 144–156, doi:10.1057/jos.2016.7.
- Ramaa, A., Subramanya, K.N., Rangaswamy, T.M., 2012. Impact of warehouse management system in a supply chain. *Int. J. Comput. Appl.* 54 (1), 14–20.
- Rouzafoon, J., Helo, P., 2018. Developing logistics and supply chain management by using agent-based simulation. Paper presented at the 2018 First International Conference on Artificial Intelligence for Industries (AI4I). Laguna Hills, CA, USA.
- Sarma, S.E., Weis, S.A., Engels, D.W., 2002. RFID systems and security and privacy implications. Paper presented at the International Workshop on Cryptographic Hardware and Embedded Systems. Springer, Berlin.
- Siebers, P.O., Macal, C.M., Garnett, J., Buxton, D., Pidd, M., 2010. Discrete-event simulation is dead, long live agent-based simulation! *J. Simul.* 4 (3), 204–210, doi:10.1057/jos.2010.14.
- Stadtler, H., 2005. Supply chain management and advanced planning—basics, overview and challenges. *Eur. J. Oper. Res.* 163 (3), 575–588.
- Thornton, F.B., Das, A., Bhargava, H., Campbell, A., Kleinschmidt, J., 2006. RFID Security. Syngress Publishing, Inc., Rockland.
- Vidoni, M.C., Vecchiotti, A.R., 2015. A systemic approach to define and characterize Advanced Planning Systems (APS). *Comput. Ind. Eng.* 90, 326–338.
- Waller, M.A., Fawcett, S.E., 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *J. Bus. Logist.* 34 (2), 77–84.
- Wang, X., Poikonen, S., Golden, B., 2017. The vehicle routing problem with drones: several worst-case results. *Optim. Lett.* 11 (4), 679–697.
- Wang, Z., Hu, H., Gong, J., 2018. Simulation based multiple disturbances evaluation in the precast supply chain for improved disturbance prevention. *J. Clean. Prod.* 177, 232–244, <https://doi.org/10.1016/j.jclepro.2017.12.188>.

Factors Affecting the Selection of Logistics Service Providers

Aicha Aguezzoul, University of Lorraine, Metz, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	85
Characteristics of 3PLs	85
The 3PL Selection Process	86
Selection Criteria	86
Importance of Criteria	87
Conclusions	88
References	88

Introduction

In a global economic environment, which is increasingly competitive and uncertain, any company could remain competitive and efficient without working closely with external partners. The concept of supply chain management has emerged in this direction and aims to manage in an efficient way, physical, informational, and financial flows between all its actors, for permanent intra and interorganizational coordination and organizational agility. In addition, in the current context of sustainable development, environmental, social, and societal factors must be taken into consideration in the management of logistics processes.

This coordination in the supply chain involves relationships between its various partners—customers, suppliers of goods, third-party logistics service providers (3PLs). It is strategic since in most industries, 60%–80% of value-added services come from suppliers. In the case of 3PLs, Lieb and Bentz (2005) report that about 60% of fortune 500 companies in the United States have signed at least one contract with a 3PL. Similarly, the number of European companies engaging 3PLs has nearly doubled since 2000 and the European volume of 3PLs and SCM business in 2016 was around USD 173 billion, while total global revenues were over USD 802 billion (Langley, 2018).

In order to set up this interorganizational cooperation, however, the company must select a well-performing set of partners to meet its needs in terms of profitability and risk. This selection is a complex process because, on the one hand, it is multifunctional and involves purchasing actors as well as other skills and resources of society, and on the other hand, it is multicriteria because it takes into account various tangible and intangible criteria such as price, quality, flexibility, reputation, delivery, expertise, etc. This process is obviously related to the product or service complexity to be purchased. To solve this complex problem, several techniques and methods are proposed in the literature, mainly in the case of the selection of product suppliers.

In this article, we focus on the 3PL selection process according to two aspects: the first one aims to identify the main criteria proposed in this process, while the second one aims to understand and explain the evolution over time of the importance of these criteria. Thus, to understand these aspects, the first part of this article deals with 3PL characteristics, and presents the various services they offer. The second part presents, from a recent study on this subject, an analysis of the different criteria as well as the evolution of their importance. The last part summarizes the implications, limitations, and potential perspectives of future research.

Characteristics of 3PLs

A substantial volume of research has been carried out on the outsourcing of logistics, the objective of which is to entrust to a 3PL, all or part of the realization and sometimes of the design, of a customer's logistics activities, possibly involving the physical transfer of materials or personnel. This strategic and complex decision allows the customer to focus on its core business.

Although logistics outsourcing is associated with a loss of control of outsourced activities, a reduction in direct customer contact and the potential for 3PL opportunism, the growing interest in this area is motivated by several factors (Razzaque and Sheng, 1998; Selviaridis and Spring, 2007), such as:

- reduction of costs, thanks to the economies of scale generated by a 3PL or by a release of immobilized capital,
- transformation of fixed costs into variable costs, thus allowing a company to focus on its own resources (financial, human, etc.), production and know-how,
- better control of logistics costs,
- increased flexibility of a 3PL, derived from its market knowledge,
- privileged access to 3PL skills and expertise, and
- in the context of globalization, logistics outsourcing is a way to get closer to customers all over the world.

These 3PLs may perform, in whole or in part, the logistics activities of their customers, whether or not they own warehouses and transportation means; knowing, however, that many of them claim a growing global presence, such as DHL, Kuehne-Nagel, DB Schenker, and Geodis.

Warehousing - Storage - Warehouse - Bonded warehouse - Storage at three temperatures (+, -, 0) - Automation, etc.	Packaging - Packing - Custom packaging - Copacking - Kitting - Stuffing-stripping - Palletization, etc.	Sustainable logistics - Ecofriendly packaging - Returns management - Waste management - Electric vehicles, - Mega-truck, etc.	Customer services - Call centers - Logistics advice - After sales service - Billing for account - Direct marketing - Logistics in situ, etc.
Order management - Cross-docking - Comanufacturing - Merchandizing - Postponement - Pick by voice - Pick by vision, - Pick to light - Put to light, etc.	Information & communication technology - E-commerce - EDI, ERP, WMS, TMS - Control Tower - Front office solutions - RFID - Bar code - Digital solutions - Geolocation etc.	Transport-distribution - Modes of transport - National, international, and multimodal transport - Last mile delivery (LMD) - Transport at controlled temperature - Urban distribution - Express distribution - Customs engineering, etc.	

Figure 1 3PL services by logistics activity.

To achieve effective integration and coordination in supply chains, most 3PLs have expanded their offered services from traditional or basic services (transport and warehousing) to innovative and value-added services (Fulconis et al., 2016; Aguezzoul, 2019). Fig. 1 provides a categorization by logistics activity, of the main services offered by 3PLs. This evolution, diversification, and innovation in their service offers have allowed 3PLs to access increasingly important roles in their customers' supply chains, from simple operators of logistics operations to becoming integrators of logistics solutions, sometimes at the worldwide level.

In order to set up the outsourcing contract with a 3PL, however, the customer must first make a rigorous selection from a set of potential performants of the 3PL function; those that have met their initial requirements. The following section details the 3PL selection process in terms of criteria, as well as the evolution of their importance over recent years.

The 3PL Selection Process

Selection Criteria

Once the customer has decided to outsource its logistics processes, the next step is to determine a set of potential performants from which to choose a 3PL. This strategic decision is influenced by several factors such as price, quality, services offered, technology mobilized, reputation, and the professionalism of the 3PL. Moreover, in the context of trade globalization, new factors must be taken into account, such as cultural adequacy to the target markets, sustainable aspects, political stability, and security. These factors represent the customer's sources of added value, since the expertise of the 3PL, for example, allows the customer to distinguish itself from its competitors.

A detailed analysis of 67 articles dealing with the 3PL selection problem, and published during the period 1994–2013 in reputable journals in the fields of logistics, SCM, and purchasing, has allowed the identification of 11 main criteria (Aguezzoul, 2014). An extension of this study to 17 articles published during the period 2014–17 (Aguezzoul, 2019), allowed these criteria to be extended to 12 and for them to be classified in this order of importance—relationship, cost, quality, services, flexibility, ICT, delivery, professionalism, financial situation, physical equipment and infrastructure, location, reputation, and sustainable logistics. Each of these main criteria is defined by a set of sub-criteria (Table 1).

It should be noted that the majority of the studies analyzed (80%) are empirical, conducted mainly in Asian countries, particularly China and Taiwan. Furthermore, international transport and warehousing remain the most outsourced logistics activities. The use of international 3PLs offering international transport services is due to the globalization of markets; 3PLs are increasingly integrating into global and complex logistics chains.

In addition, the multicriteria approach is the most used in the 3PL selection process and continues to be enriched by other criteria, such as ICT and sustainable logistics (Aguezzoul and Pires, 2016; Aguezzoul and Paché, 2018). The list of criteria to be considered differs depending on the outsourced services. For example, in the case of international transport, criteria such as the relationship with the 3PL, delivery reliability, location, ICT, and reputation are important (Yayla et al., 2015; Hwang et al., 2016). Similarly, in the case of returns management, equipment, and storage capacity of the 3PL, ICT, environmental aspects, as well as the existing relationship with the customer, are the most used criteria (Guarnieri et al., 2015). Finally, the main criteria that affect the selection of national operators are ICT, financial stability, quality, and range of services offered (Solakivi and Ojala, 2017).

Table 1 Summary of evaluating criteria

Criteria	Definition and attributes
Relationship	It includes shared risks and rewards, ensure cooperation between the user and the 3PL. It also helps in controlling the 3PL opportunistic behavior. Reliability, truth, dependence, alliance, compatibility, reciprocity are among its attributes.
Cost	It refers to the total cost of logistics outsourcing. Its related attributes include price, cost reduction, low-cost distribution, expected leasing cost, operations cost, warehousing cost, and cost savings.
Quality	Quality of the 3PL includes many aspects such as commitment to continuous improvement, SQAS/ISO standards environment issues, and risk management.
Services	It is related to attributes such as breadth of services, characterization, and specialization of services, variety of available services, presale/postsale customer services, and value-added services.
Flexibility	Ability to adapt to changing customers' requirements and circumstances. Its attributes include ability to meet future requirement, capacity to accommodate and grow the client's business, flexibility index, responsiveness to target market or service requests, capability to handle specific business requirements, and time response capability.
ICT	ICT capacity is an important factor in improving internal and external capabilities of the 3PL. It facilitates the execution of logistics operations and the rapid exchange with its customers. Its related attributes include EDI, WMS, ERP, and CRM.
Delivery	It's represented by attributes such as time, on-time performance, on-time shipment and deliveries, delivery speed, accuracy of transit/delivery time, shipment delivery, and on-time delivery rate.
Professionalism	3PL exhibit sound knowledge of services in the industry and display punctuality and courtesy in the way they interact and present to the customers. It is characterized by attributes such as expertise, competence, and experience.
Financial position	A sound financial performance of the 3PL ensures continuity of service and regular upgrading of the equipment and services, which are used in logistics operations.
Physical equipment and infrastructure	It corresponds to physical equipment and infrastructure that the 3PL uses to facilitate execution of logistics operations of its customers. It's related to attributes such as hub, warehouse, materials handling equipment, and means of transport.
Location	It's related to attributes such as distribution coverage, geographical specialization and coverage, international scope, market coverage, shipment destinations, and distance.
Reputation	It refers to the opinion of the customers about how good is the 3PL in satisfying their needs. This is more relevant in the initial screening of 3PL.
Sustainable logistics	Its activities are such as pallet flow management, recycling, reuse, repair, and return management. Its attributes include energy consumption, CO2 emissions, congestion rate, and number or volume of returns.

Importance of Criteria

Table 2 compares the two studies discussed above and conducted on 3PL selection during different time periods, in order to illustrate the evolution of the importance of these various criteria. The rank of each criterion is deduced from the number of times it is cited in the articles analyzed.

From these results, several trends are observable. Over time, it seems that the relative importance of ICT has increased significantly, while that of cost is relatively lower than in the past. This overall increase in the relative importance of ICT is due to the change in customer/3PL relationships. Indeed, ICT has facilitated the rapid exchange of information, to build cooperative relationships between a 3PL and its customers, which are increasingly distant geographically, and finally, to expand their (sustainable) supply chains to other partners around the world. Moreover, and although largely underestimated so far in the literature, the social

Table 2 Evolution of the importance of the selection criteria for 3PL

Criteria	2014–17 (<i>Aguezzoul, 2019</i>)		1994–2013 (<i>Aguezzoul, 2014</i>)	
	Nb. papers	Rank	Nb. papers	Rank
ICT	14	1	23	8
Relationship	14	1	44	2
Quality	11	2	40	4
Cost	10	3	46	1
Physical equipment and infrastructure	9	4	15	11
Services	8	5	42	3
Financial position	8	5	19	9
Delivery	7	6	28	6
Professionalism	7	6	24	7
Flexibility	6	7	35	5
Reputation	5	8	11	12
Location	4	9	16	10
Sustainable logistics	4	9	6	13

Table 3 Overall rank of selection criteria for 3PL

<i>Criteria</i>	<i>Total</i>	<i>%</i>	<i>% Cumulated</i>
Relationship	58	12.72	12.72
Cost	56	12.28	25.00
Quality	51	11.18	36.18
Services	50	10.96	47.15
Flexibility	41	8.99	56.14
ITC	37	8.11	64.25
Delivery	35	7.68	71.93
Professionalism	31	6.80	78.73
Financial position	27	5.92	84.65
Physical equipment and infrastructure	24	5.26	89.91
Location	20	4.39	94.30
Reputation	16	3.51	97.81
Sustainable logistics	10	2.19	100
Total	456		

relationships between 3PL and their customers constitute an important dimension in generating trust and reducing both opportunism and informational asymmetries.

The criteria related to sustainable logistics, especially reverse logistics, have also become increasingly important, due to the ongoing development of sustainable supply chains. As for the proximity of geographic markets, it does not seem to be a determining factor. Finally, using the Pareto method, we deduce the result as shown in **Table 3**, which aggregate the number of citations of each criterion as well as its respective rank.

The criteria most commonly used during the period 1994–2018 are found to be in this order of importance: relationships, cost, quality, services, flexibility, ICT, delivery, and professionalism. These criteria represent a little less than 80%, while criteria related to financial position, equipment and infrastructure, location, reputation, and sustainable logistics represent just over 20%. As noted earlier and following the emergence of new ICT, relational performance is shown to be a stronger driver of customer loyalty than are costs. This could also be explained by the fact that customer/3PL collaboration could reduce logistics costs by, for example, setting up a pooling system which also reduces CO₂ emissions and improves the loading rate of vehicles.

Conclusions

In the work presented in this article, the main decision criteria used in practice in the 3PL selection process are reviewed. An analysis of the various studies published up to 2017 helps to identify the relative importance that companies place on the criteria they use in this process. Over the 2014–17 period, this relative importance has changed considerably as a result of the emergence of ICT, which has encouraged close collaboration between 3PLs and their customers, even globally. Moreover, with the development of sustainable supply chains, other criteria particularly related to returns activities and waste processing, are becoming increasingly important and prevalent in this selection. From a managerial point of view, the results of this study could serve as a guide, on the one hand, for companies in the planning of logistics outsourcing actions and, on the other hand, for new purchasing managers or logisticians in identifying criteria favored when negotiating with 3PLs.

References

- Aguezoul, A., 2014. Third-party logistics selection problem: a literature review on criteria and methods. *Omega: Int. J. Manag. Sci.* 49, 69–78.
- Aguezoul, A., Pires, S., 2016. 3PL performance evaluation and selection: a MCDM method. *Supply Chain Forum: Int. J.* 17 (2), 87–94.
- Aguezoul, A., Pache, G., 2018. A decision making tool for selection of LSPs. *Int. J. Bus. Manag.* 13 (5), 95–104.
- Aguezoul, A., 2019. Prestataires de services logistiques: évolution de l'importance relative des critères de sélection. *Logist. Manag.* 1–8.
- Fulconis, F., Nollet, J., Paché, G., 2016. Purchasing of logistical services: a new view of LSPs' proactive strategies. *Eur. Bus. Rev.* 28 (4), 449–466.
- Guarnieri, P., Sobreiro, V., Nagano, M., Serrano, A., 2015. The challenge of selecting and evaluating third-party reverse logistics providers in a multi-criteria perspective: a Brazilian case. *J. Clean. Prod.* 96, 209–219.
- Hwang, B.-N., Chen, T.-T., Lin, J., 2016. 3PL selection criteria in integrated circuit manufacturing industry in Taiwan. *Supply Chain Manag.: Int. J.* 21 (1), 103–124.
- Langley, C.J., 2018. Infosys 22nd annual third-party logistics study. Available from: https://mymaritimeblog.files.wordpress.com/2017/10/3pl_2018_study.pdf.
- Lieb, R., Bentz, B., 2005. The use of third-party logistics services by large American manufacturers: the 2004 survey. *Transp. J.* 44 (2), 5–15.
- Razzaque, M.A., Sheng, C.C., 1998. Outsourcing of logistics functions: a literature survey. *Int. J. Phys. Distrib. Logist. Manag.* 28 (2), 89–107.
- Selviaridis, K., Spring, M., 2007. Third party logistics: a literature review and research agenda. *Int. J. Logist. Manag.* 18 (1), 125–150.
- Solakivi, T., Ojala, L., 2017. Determinants of carrier selection: updating the survey methodology into the 21st century. *Transp. Res. Proc.* 25, 511–530.
- Yayla, A.Y., Oztekin, A., Gumus, A.T., Gunasekaran, A., 2015. A hybrid data analytic methodology for 3PL transportation provider evaluation using fuzzy multi-criteria decision making. *Int. J. Prod. Res.* 53 (20), 6097–6113.

Logistics Service Performance

Kee-hung Lai*, Jinan Shao*,†, Yongyi Shou†, *The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China; †Zhejiang University, Hangzhou City, Zhejiang Province, China

© 2021 Elsevier Ltd. All rights reserved.

Evaluation of Logistics Service Performance	89
Customer Segments for Logistical Services	90
Improvement of Logistics Service Performance	91
Summary	93
References	93
Further Reading	93

Evaluation of Logistics Service Performance

Due to its influence on downstream customer experience and the cost implications for executing these logistical activities in the process, logistics service is an important element in logistics management in order for firms to competently manage their cargo movement activities. Providing quality services with the aim of meeting customer requirements at reasonable cost is essential for a firm to create customer satisfaction and loyalty, increasing its ability to compete in the market. Toward achieving this end, firms need to understand customer requirements with respect to the seven “Rights” in logistics management and to satisfy them fully. Logistics service performance evaluates organizational ability in logistics management to deliver the right product, in the right amount and the right condition, to the right location at the right time for the right customer at the right price (and cost). This seven “Rights” principle highlights the importance of handling cargo movements to satisfy customer requirements in an efficient, timely, and reliable manner (Rushton et al., 2014). In a broad sense, achieving the seven “Rights” in logistics services reflects the “Availability” of goods for customers, which is fundamental in logistics management for firms in any industry (Lai and Cheng, 2009).

There is no universal definition of logistics service performance and it depends on the contextual situations for application. Several studies have conceptualized the attributes used for evaluating a firm’s logistics service performance. For example, Stank et al. (1999) examined the logistics service performance of fast food restaurants and identified two critical aspects: operational performance and relational performance. While operational performance focuses on the characteristics of logistical services (e.g., timeliness, reliability, and cost), relational performance focuses on activities that can enhance a firm’s closeness to customers (e.g., the courtesy of employees and responsiveness to customer requests). Stank et al. (2003) operationalized the logistics service performance of third-party logistics firms in three performance dimensions: operational performance, relational performance, and cost performance. Murfield et al. (2017) examined logistics service performance for the online retail industry. They suggested that relational performance is less important in the online context as buyers and sellers are physically apart and the services provided are aimed at goods and not at people. In this context, logistics service performance is conceptualized as comprising three performance dimensions: consumers’ perceptions of availability, delivery timeliness, and product condition. In the airline industry, Farooq et al. (2018) emphasized that airline tangibles, terminal tangibles, personnel services, empathy, and airline image are five crucial dimensions of logistics service performance. Overall, it is important for managers to consider the contextual situations in determining the appropriate logistics service performance attributes for applications in their organizations.

There are studies highlighting the importance of examining logistics service performance from customers’ viewpoints because their satisfaction level is dependent on how they perceive the service quality. The focus is on assessing the gap in logistics service quality regarding the services desired by customers relative to their perceived service delivery. Furthermore, customers experience logistics service in three different stages, starting from their pre-purchase inquiry (e.g., written policy for services, delivery schedule) and in-purchase interaction (e.g., ease of payment, stock availability) through to post-purchase follow-up (e.g., product returns, alterations). It is vital that firms fulfill their customer requirements in these different service stages to improve their experience and satisfaction. Examples for pre-transaction stage logistics activities include a written statement of customer service policy, methods of ordering, accessibility of order personnel, and system flexibility, such as handling last minute or minimum quantity orders. For the transaction stage, example activities include stocking products in good condition, inventory availability, delivery time, and delivery reliability. The post-transaction stage includes activities such as product tracing and warranties, product substitution and returns procedures, as well as customer claims and complaints handling. Many firms have recognized the importance of balancing and improving their logistics activities in these different stages without a lopsided focus. For example, JD.com, one of the largest online retailers in China, emphasizes quality logistics services for its customers. In the pre-transaction stage, JD.com provides its customers with service policy statements in written form (e.g., to deliver customer orders within 24 hours upon order receipt). JD.com also operates a call center to answer customer enquiries before their purchase. In the transaction stage, JD.com can responsively handle the delivery of goods from its warehouses to fulfill customer orders. JD.com offers “same-day delivery” and “next-day delivery” services to its customers, such that their ordered goods can be delivered in a timely manner. In the post-transaction stage, customers

can trace their purchase orders online in real-time, while JD.com is able to promptly respond to customer complaints and efficiently handle returned products.

Customer Segments for Logistical Services

Logistics management can create value for customers in terms of achieving the seven “Rights,” reflecting an organizational objective to meet customer requirements at reasonable cost with product availability. However, the value creation process comes with costs. Is it always advisable for a firm to provide a 100% service level for its customers and meet their requirements fully? This 100% service level means that customers come to purchase a hundred times and get exactly what they want without discrepancy. If they can only get what they desired ninety times out of their hundred visits, the service level achieved is only 90%. Whether or not a firm provides a high service level depends on the importance of the product items and customers for the firm’s profitability. For “cash cow” items (or important customers such as VIPs), a firm needs to provide a 100% service level because they are core businesses (e.g., representing 80% of its total sales) and can bring in a steady flow of income. But for trivial items (e.g., representing less than 1% of its total sales), providing a 100% service level is unnecessary and costly, compromising the firm’s profitability due to excessively high logistics costs for handling and storage, but with low financial returns. This situation is particularly salient for retail fronts such as convenience stores, where physical retail space is limited and it is expensive to maintain a higher service level for trivial items. In this situation, an occasional stock-out can be a better logistics option.

As another example to illustrate, customers can demand as many as twenty service items for liner shipping services (e.g., on-line booking, tracking and tracing, customs clearance, high frequency of sailing, door-to-door service, consolidation service and temporary storage). Yet, it is not necessary to offer all these services as doing so can be very costly and uneconomical. It is crucial for firms to identify the critical ones (e.g., based on the 20/80 rule) and make their best efforts to ensure that these critical items are fully achieved. The 20/80 rule suggests an operations strategy that coverage of 20% of service items (e.g., offering and excelling in the four most desired shipping service items out of the twenty items in the above example) can satisfy 80% of the services required by customers (the remaining sixteen items are either not offered or offered at a low level of service due to high costs with low marginal economic returns). It is essential that firms balance and improve their logistics service with the incurred cost considered, so as to avoid providing a high service level with low economic returns. In this regard, they can be more selective in providing logistics services for their customers. One way to do so is to customize their logistics services to specifically cater for the desires of different customer segments based on their characteristics, such as demographic (e.g., age, sex, and occupation), geographic (e.g., domestic, international), behavioral (e.g., use frequency, customer loyalty), and product characteristics (product value, product durability).

The rationale is that customer segments vary in their requirements for logistics services. For example, timeliness is a crucial factor for delivery of fresh foods because such product items are perishable. Fresh foods need to be refrigerated at the right time and delivered to the right location in order to keep them in the right condition. For products characterized with high value, security and timeliness are important service aspects to emphasize in logistics operations. Firms are encouraged to create value for customers by achieving the seven “Rights” in logistics, so long as the activities are profitable. Doing so will require firms to better understand the requirements of their target customer segments and fulfill them at reasonable costs. It can be a waste in logistics management that firms provide the wrong service items that are not desired by the market and/or their targeted customer segments or, alternatively, even the right service items but to an excessively high service level with low marginal economic returns.

In view of the varying requirements of different customer segments for logistics services, a classification framework can be useful for managers to design, plan, and execute logistics operations in their organizations. In general, there are two broad product characteristics, namely product durability and product value, which can be used by firms to segment their customer groups characterized with different logistics service requirements. As nondurable products (e.g., fruits and fashion) have relatively shorter product life cycles, timeliness is a prioritized factor to emphasize in their logistics arrangements. In considering the liability aspect of high value products (e.g., electronics and jewelry), security and damage free handling are vital for processing these items in logistics operations. These two dimensions of product characteristics provide useful insights for firms to understand logistics service requirements and develop logistics operations strategy for handling different product categories with the classifications shown in Fig. 1.

In segment 1, the products are characterized by high durability and low value. Examples of such kinds of products include textiles, books, papers, and plastics. For this segment, firms can choose to offer simply standardized logistics services, such as warehousing, distribution and pick and pack to meet customer requirements at the right cost.

In segment 2, the products are characterized by low durability and low value. For such kinds of products, firms are advised to put special emphasis on the timeliness of their logistical services. Customers in this segment care more about whether orders can arrive at their locations as promised. For example, staple food (e.g., vegetables and fruit) need to arrive at the right time in order to keep the food healthy, fresh, and in the right condition.

In segment 3, the products are characterized by high durability and high value. As the financial value of the products is high, secure and reliable logistical handling for the product movement are especially emphasized by customers of this segment. Firms need to pay the utmost attention to ensure security in handling these product items (e.g., electrical appliances and automotive products) at the right quantity with the right condition throughout the logistics management process.

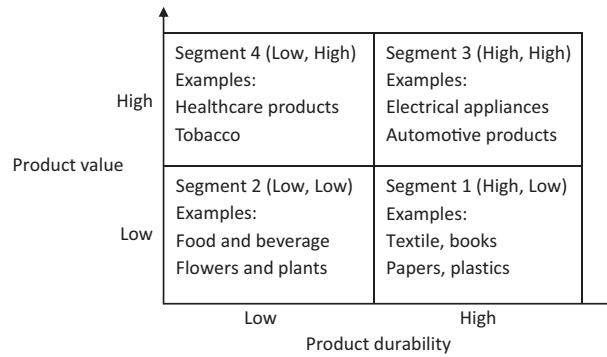


Figure 1 Product classifications based on their value and durability.

In segment 4, the products are characterized by low durability and high value. The storage and delivery of such products (e.g., healthcare products and tobacco) require an emphasis on delivery at the right time with the right quantity and condition.

In sum, product categories servicing different customer segments vary in their logistics service requirements. Managers are advised to understand the cost and service implications in designing and delivering the types and levels of logistics services catering to the needs of their targeted customer segments. While it is desirable for a firm to strive for excellence in their logistics services, the wastes caused by providing excessive services, unwanted by their customer segments, must be avoided.

Improvement of Logistics Service Performance

For logistics service improvements, many firms have taken initiatives to enhance their service capability and performance. In the process, they need to pay attention to both relational and operational aspects of logistics service in planning and executing their logistics activities to attain the seven “Rights.”

For the relational aspect, this is concerned with the effectiveness of servicing customer needs, that is, doing the right thing. Building strong relationships with customers is essential for firms to enhance their logistics service performance. Firms can organize training programs to improve their boundary-spanning employees’ ability to develop long-term relationships with customers. These employees are frontline staff who directly interact with customers and identify customer needs first. Training programs can develop their knowledge on how to interact with customers and satisfy their logistics needs well. In addition, it is critical for firms to establish customer relationship management (CRM) systems, which can help facilitate the collection, access, transfer, and analysis of customer information for planning operations strategy to improve their customer experience. Such CRM tools are helpful for firms to communicate with their customers and to respond promptly to their requests, including claims and complaints. Utilizing CRM can also make it easier for firms to generate and disseminate customers’ current and future logistics requirements throughout their organizational functions and along the supply chain for responsive actions. For example, SF Express, the logistics giant in China, has developed an “E-learning” mobile platform, which provides its employees with plentiful training materials to facilitate their knowledge acquisition. This learning platform is helpful for employees to keep themselves updated with the latest trends in the logistics industry and to cultivate their customer service skills (e.g., communication skills, problem-solving skills, and logistics operations skills). Moreover, SF Express has collaborated with SAP to set up their CRM system. This system compiles customer data across different communication channels (e.g., e-mail, telephone, and social media) operating in SF Express. It collects and stores information about customers’ contact information, purchase history, personal preferences for logistical services and feedback. The system can also help the company analyze the demands of different customer groups and suggest targeted marketing campaigns to improve customer relationships.

The operational aspect is concerned with the efficiency in providing quality services to meet customers’ requirements, that is, doing the things right. Performing logistics activities wrongly (e.g., delivery delay due to lack of planning) can be costly for firms to rectify, such as arranging product returns and rehandling. To improve logistics operations, firms need a mechanism to manage and control the logistics activities for continuous improvements (Lai et al., 2004). The six-sigma approach in quality management can be a viable way for firms to identify areas for continuous improvement actions to enhance the efficiency of their logistics management activities. This approach involving a continuous improvement cycle in the five DMAIC stages (Define-Measure-Analyze-Improve-Control), together with appropriate quality management tools, will be useful for firms to identify and eliminate inefficiency in their logistics operations. For the application of the six sigma approach, there are seven basic tools of quality management including process maps, check sheets, histograms, scatter plots, control charts, cause and effect diagrams, and Pareto analysis, as well as the seven new tools including the affinity diagram, interrelationship graph, tree diagrams, prioritization grid, matrix diagram, process decision program chart and activity network diagram. Firms will also find other quality management tools for performance measurement, such as the balanced scorecard and dashboards, useful for monitoring and controlling their logistics service performance.

On another note, technological development is a trend for logistics service performance improvement. For instance, technological tools such as electronic data interchange, warehouse management systems, and global positioning systems can help enhance the operational efficiency and service performance of firms. In recent years, firms have adopted the new "ABCD" technologies in response to the increasing customer expectations with the aim of enhancing their logistics efficiency. The "ABCD" refers to artificial intelligence (AI) (A), blockchain (B), cloud computing (C), and data analytics (D). These emerging technological developments are taking increasingly central roles in shaping logistics operations in many enterprises.

Artificial intelligence is a branch of information technology that copes with the automation of intelligent behavior. AI aims to utilize computers or machines that can process problems independently, in a manner similar to the way a human with appropriate training would do. Advances in AI applications are crucial to the service performance improvement of logistics activities. The use of AI-enabled robotics and autonomous vehicles is the physical embodiment of AI in logistics operations. Intelligent robotics can provide firms with effective high-speed sorting of letters and parcels. For example, JD.com has started to use robotics to move and sort parcels. The robots can automatically drive and "see" different items, and perform the tasks in a more efficient way than manual labor. The robots can sort 9000 parcels per hour in a warehouse center, which minimizes the amount of man-hour time it takes to complete the sorting tasks. In addition, JD.com has tried to use robots to deliver items. It has designed a four-wheeled drone, which can carry five packages at once and travel 20 km, if fully charged. SF Express has also tested the commercial unmanned aerial vehicle with a planning schedule to use it to deliver goods in rural areas, mountainous landscapes, and to islands.

Blockchain refers to the use of distributed ledger technology that can record transactions between multiple parties in a secure and permanent way. Blockchain technology can be used for any exchanges, contracts, tracing, and payments. It can enable data transparency and access among different stakeholders involved in the supply chain, because every transaction is recorded on a block and across multiple copies of the ledger that are distributed over many nodes. Due to the improved data transparency, blockchain technology can help build trust between different transaction parties along the supply chain. Firms can incorporate blockchain into their operations to improve supply chain transparency and traceability for their transactions. Operational data concerning how goods are made, where they come from, and how they are delivered can be stored in a blockchain-based system accessible to multiple companies in the supply chain. For instance, SF Express has used blockchain technology to build a traceable supply chain for improving its logistics operations. It has launched a project called "Fenghui GO" and incorporated a blockchain source code with the ID "Haito Goods." Consumers are required to scan traceable QR codes to reveal the product descriptions and the logistics information of the whole process. This technology helps increase the speed of flow of goods, as well as the transparency and traceability of the transaction process.

Cloud computing technology has begun to play a major role in logistics operations. It allows firms to access different types of data in real-time from anywhere via the internet and to deploy resources immediately in case responses are needed. For example, Cainiao, a technology company in China, has used the Alibaba cloud platform to build a "logistics cloud" that tracks packages at every stage of the supply chain. This platform provides real-time access to logistics information for both buyers and sellers. Cainiao also deploys logistics resources (e.g., vehicles) and performs analyses to optimize transportation routes through the Alibaba cloud. Cloud technology also allows real-time inventory management and data flow from the cloud helps firms more precisely control their inventory levels in response to demand fluctuations. For instance, Tencent, the technology giant in China, provides cloud services for numerous companies to help them better manage inventories. Through the Tencent cloud platform, firms can keep track of the stock movement and inventory levels in real-time and ensure that inventory information recorded in the platform is consistent with the actual inventory quantity stocked in the warehouse. This cloud-based technology provides a more accurate control of inventory levels for firms and helps reduce the incidences of stock-out. In addition, cloud computing technology can enable firms to more effectively monitor how equipment (e.g., vehicles and trays) is utilized and discover utilization patterns to better exploit abundance or to eliminate wasteful excess.

Data analytics technology brings an incredible potential for improving logistics operations and service performance. In contemporary logistics management, firms manage a massive flow of goods and meanwhile create huge amounts of data sets for their logistics planning and operations. Many companies have implemented a "data-driven" strategy to derive value from large-scale quantities of data (Singh et al., 2019). For example, Cainiao has applied big data techniques to optimize routing by analyzing comprehensive historical utilization of data collected from transit points and transportation routes. Through route optimization, the data analytics helps Cainiao save 260 million hours a day in delivery time. In addition, Hema Fresh, a new grocery retailer in China, has leveraged the power of big data to choose the locations for its supermarkets. It has applied big data techniques to analyze the number of active users of Alipay and estimate the purchasing power of users in a region. The data analytics helps Hema Fresh determine the supermarket locations where sufficient target customers may exist to support their operations. Hema Fresh can deliver online orders to customers within 30 minutes in a 3 km radius from each of its supermarkets. Learning about customer demand and satisfaction is also a promising application for big data techniques. For instance, JD.com has developed a secure database that includes massive customer shopping and feedback data. Using data analytics, JD.com can better understand customer preferences and satisfaction, creating a personalized shopping experience for them. In sum, all these emerging "ABCD" technological applications represent contemporary directions for firms to improve logistics service performance in their operations.

Further to technological adoption, environmental management is an increasingly important area for firms to seek improvements in their logistics service performance due to stakeholder pressure for environmental management practices (Vejvar et al., 2018). The rapid development in e-commerce and online retail sales in recent years has boosted the demand for logistics handling and delivery. The cargo movement processes can cause serious pollution and environmental damage, urging a need for greener service (Chan et al., 2016). For instance, there are increased greenhouse gas emissions by vehicles leading to poor air quality. The cargo movement

processes can also generate large amounts of waste, including wrapping materials and pallets. As such, it is essential that firms adopt green practices to lessen environmental damage and conserve resources in handling cargo movements, enhancing their logistics service performance (Lai et al., 2010). The increasing government pressures for environmental preservation have also urged firms to mitigate the operational damage they cause to the natural environment. In response, many enterprises in China have begun to embrace green practices as a part of their logistics service improvements. For instance, JD.com is proactive in implementing environmentally friendly initiatives. Some of its exemplary practices include the use of energy-efficient materials and handling equipment, automated loading, recyclable packaging and dissolvable materials, alternative fuels, electric vehicles and consolidated delivery. Overall, the environmental and sustainability management aspects are growing as important areas for firms to improve their logistics service performance.

Summary

In this chapter, we discuss the principles of logistics service performance with respect to the attainment of the seven “Rights” and the need for different treatments in different customer segments, as well as the emerging trends on technological adoption for logistics service performance improvements. Firms need to pay attention to different contextual situations which vary in the performance requirements for planning actions to improve their logistics services. There are also three stages of logistics services for customers that firms need to balance and improve for performance enhancement. The emerging “ABCD” technological applications and environmental management are the trending directions for firms to seek logistics service performance in the years ahead.

References

- Chan, T.Y., Wong, C.W.Y., Lai, K.H., Lun, V.Y.H., Ng, C.T., Cheng, T.C.E., 2016. Green service: construct development and measurement validation. *Prod. Operat. Manag.* 25 (3), 432–457.
- Farooq, M.S., Salam, M., Fayolle, A., Jaafar, N., Ayupp, K., 2018. Impact of service quality on customer satisfaction in Malaysia airlines: a PLS-SEM approach. *J. Air Transp. Manag.* 67, 169–180.
- Lai, K.H., Cheng, T.C.E., 2009. *Just-in-Time Logistics*, 1st ed. Routledge, London.
- Lai, K.H., Lau, G., Cheng, T.C.E., 2004. Quality management in the logistics industry: an examination and a ten-step approach for quality implementation. *Total Quality Manag. Bus. Excell.* 15 (2), 147–159.
- Lai, K.H., Cheng, T.C.E., Tang, A.K.Y., 2010. Green retailing: factors for success. *California Manag. Rev.* 52 (2), 6–31.
- Murfield, M., Boone, C.A., Rutner, P., Thomas, R., 2017. Investigating logistics service quality in omni-channel retailing. *Int. J. Phys. Distrib. Logist. Manag.* 47 (4), 263–296.
- Rushton, A., Croucher, P., Baker, P., 2014. *The Handbook of Logistics and Distribution Management: Understanding the Supply Chain*, 5th ed. Kogan Page, London, United Kingdom.
- Singh, N., Lai, K.H., Vejvar, M., Cheng, T.C.E., 2019. Big data technology: challenges, prospects and realities. *IEEE Eng. Manag. Rev.* 47 (1), 58–66.
- Stank, T.P., Goldsby, T.J., Vickery, S.K., 1999. Effect of service supplier performance on satisfaction and loyalty of store managers in the fast food industry. *J. Operat. Manag.* 17 (4), 429–447.
- Stank, T.P., Goldsby, T.J., Vickery, S.K., Savitskie, K., 2003. Logistics service performance: estimating its influence on market share. *J. Bus. Logist.* 24 (1), 27–55.
- Vejvar, M., Lai, K.H., Lo, C.K.Y., Fürst, E.W.M., 2018. Strategic responses to institutional forces pressuring sustainability practice adoption: case-based evidence from inland port operations. *Transp. Res. Part D: Trans. Environ.* 61, 274–288.

Further Reading

- Mentzer, J.T., Flint, D.J., Hult, G.T.M., 2001. Logistics service quality as a segment-customized process. *J. Mark.* 65 (4), 82–104.
- Mentzer, J.T., Myers, M.B., Cheung, M.S., 2004. Global market segmentation for logistics services. *Ind. Mark. Manag.* 33 (1), 15–20.
- Rao, S., Rabinovich, E., Raju, D., 2014. The role of physical distribution services as determinants of product returns in Internet retailing. *J. Operat. Manag.* 32 (6), 295–312.

The World Bank's Logistics Performance Index

Christina K. Wiederer*, Jean-François Arvis*, Lauri M. Ojala†, Tuomas M. M. Kiiski†, *Global Trade and Regional Integration Unit, The World Bank, Washington, DC, United States; †Operations & Supply Chain Management, University of Turku, Turku, Finland

© 2021 Elsevier Ltd. All rights reserved.

Introduction to the LPI	94
The LPI Methodology	94
Guidelines on How to Use the LPI and How to Interpret it	95
Logistics Performance Matters	95
Gaps in Logistics Performance Persist	95
Supply Chain Reliability and Service Quality are Strongly Associated with Logistics Performance	98
Delivering Good Quality Services is Key to Successful Operations, and its Importance is Growing	99
Supply Chain Resilience and Sustainability are Emerging Concerns	99
Pushing the Envelope of Implementation	100
The Policymaking and Business Impact of the Logistics Performance Index	100
References	100
Further Reading	101

Introduction to the LPI

Logistics is understood as a network of services that support the physical movement of goods, trade across borders, and commerce within borders. It comprises an array of activities beyond transportation, including warehousing, brokerage, express delivery, terminal operations, and related data and information management. The global turnover generated by these networks exceeds US\$ 4.3 trillion (2018). Although logistics is global in nature, the quality and efficiency of logistics delivery actually available to traders or manufacturers in a given country, or part of a country, is contextual and depends on a series of local factors. The concept of logistics performance, proposed by the World Bank in 2007, captures these outcomes and helps understand the magnitude and consequences of differences in performance across the world.

Logistics performance is key to economic growth and competitiveness. It determines how effectively a country connects to global networks and accesses world markets. Inefficient logistics raises the cost of doing business and reduces the potential for both international and domestic integration. Research shows that logistics performance and freight connectivity have a higher impact on the costs of trade than geographical distance. The Logistics Performance Index (LPI) has been conceived as a synthetic metric of how well countries are connected.

On July 24, 2018 the World Bank published the sixth edition of Connecting to Compete, the biennial Logistics Performance Index report. The first LPI report was published in 2007 using an identical approach (Arvis et al., 2007), which allows for comparison over time, too. All LPI reports and their aggregated data are freely available in the public domain in an interactive format at: <http://lpi.worldbank.org/>.

The LPI report puts together views of logistics professionals in more than 160 countries about how easy or difficult it is in these countries to transport general merchandise. The LPI 2018 index is compiled based on almost 6000 country assessments made by logistics professionals, which are also highly relevant in a global sourcing context.

The LPI captures a broad range of factors affecting trade logistics performance. Its results show clear benefits, particularly for developing countries, in moving forward on a broad range of fronts to improve logistics. Of particular importance is the role of public policies in shaping logistics performance. Logistics service delivery and the efficiency of supply chains depend on public sector provisions and interventions in a number of domains. Logistics uses publicly funded or regulated infrastructure. International trade is processed by border agencies. Services and logistics activities are regulated with fiscal, environmental, safety, land use, or competition objectives. The World Bank's LPI reports have emphasized the contribution of consistent policy interventions in the areas of infrastructure, the regulation of services, skills or trade facilitation. Since 2007, the LPI has stimulated national initiatives in many countries around the world.

This paper uses the materials published in the LPI 2018 report (Arvis et al., 2018), and it aims to highlight the key elements and policy-making messages of the LPI 2018 report, which the reader is encouraged to explore in depth afterward by reading the initial report.

The LPI Methodology

The LPI survey data provide numerical evidence on how easy or difficult it is in these countries to transport general merchandise—typically manufactured products in unitized form. The 2018 LPI survey employed the same methodology as the previous five

editions of Connecting to Compete: a standardized questionnaire with two parts, the international and the domestic part, respectively (Arvis et al., 2007; Arvis et al., 2010, 2012; Arvis et al., 2014; Arvis et al., 2016).

In the international part of the questionnaire, respondents evaluate six indicators of logistics performance in up to eight of their main overseas partner countries. The six main indicators of the international part of the LPI summarize on a five-point scale the assessments of logistics professionals worldwide trading with the country. The six indicators used to construct the international LPI are:

1. The efficiency of customs and border management clearance,
2. The quality of trade- and transport-related infrastructure,
3. The ease of arranging competitively priced international shipments,
4. The competence and quality of logistics services,
5. The ability to track and trace consignments, and
6. The frequency with which shipments reach consignees within the scheduled or expected delivery time.

In the domestic questionnaire, respondents are asked to provide qualitative and quantitative data for the logistics environment in the country where they work. The domestic part of the LPI indicates the quality and availability of key logistics services within a country, but due to the small number of responses, these data are more informative in comparisons by region or income group.

The next LPI edition—scheduled for 2020/2021—will feature a methodological update. While previous LPI editions relied on perception-based data from surveys of logistics professionals, the new edition will be based on micro-performance data (vessel and shipment tracking) as the main input.

This is relevant for an endeavor such as the LPI, given the network structure of logistics. In addition to micro-level data, the revised methodology aims to incorporate geospatial data as well as data from online professional sources and social networks. The goal is to improve the current LPI by addressing its challenges, such as large year-on-year fluctuations in some countries' scores and ranks, and by building on the LPI's strengths as a comprehensive supply chain performance indicator.

Guidelines on How to Use the LPI and How to Interpret it

Since 2007, LPI findings have become standard reference material in numerous studies and policy papers on trade logistics. The LPI has been adopted by several countries and international organizations as a key performance indicator in their national transport or logistics strategies. This makes it important to highlight how best to use the LPI and its indicators to avoid possible misinterpretations.

In short, individual country data—especially rank positions tracked from one LPI report to the next—should preferably not be used as the sole indicator, but should be considered in combination with scores, while also keeping the size of the confidence intervals of these scores in mind. Using the weighted aggregate score and rank data that rely on the four latest LPI ratings (e.g., for years 2012–18) is also a good idea, as they provide a more balanced picture.

Logistics Performance Matters

An effective logistics sector is now recognized almost everywhere as one of the core enablers of development. Previous editions of Connecting to Compete have highlighted how implementing better policies leads to better logistics performance. Such policies cover, for example, regulating services; providing transportation infrastructure; implementing controls, especially for international goods; and raising the quality of public-private partnerships (PPPs).

The policy focus has evolved since 2007, when the first LPI report was published. Initially, logistics policies tended to concentrate on facilitating trade and removing border bottlenecks. Today, international logistics is increasingly intertwined with domestic logistics. Policy makers and stakeholders deal with a wide range of policies. Growing concerns include spatial planning; skills and resources for training; the environmental, social, and economic sustainability of the supply chain; and the resilience of the supply chain to disruption or disaster (physical or digital).

Gaps in Logistics Performance Persist

Overall, the score profile of the entire set of more than 160 countries has remained similar since the 2007 edition, which is an indication of the robust nature of the underlying data. (Table 1).

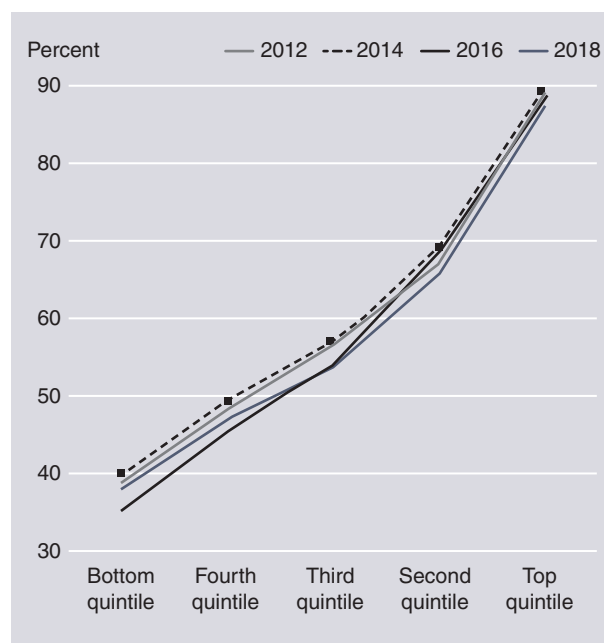
In 2018, the gap between top and bottom performers narrowed somewhat again, and the average score for the top 10 countries dropped to 4.03, whereas the bottom 10 countries scored an all-time high score of 2.08 (Fig. 1).

High-income countries occupied the top 10 rankings in 2018, eight in Europe plus Japan and Singapore—countries that have traditionally dominated the supply chain industry (Table 2). Germany is at the top, scoring 4.20. The scores of the following nine countries are in a tight interval, with Sweden in 2nd with a score of 4.05 and Finland in 10th with a score of 3.97. The composition of

Table 1 Top 10 average and lowest 10 average LPI scores, 2007–18

1=lowest; 5=highest	2007	2010	2012	2014	2016	2018
Top 10 average	4.06	4.01	4.01	3.99	4.13	4.03
Lowest 10 average	1.84	2.06	2.00	2.06	1.91	2.08

Source: Arvis et al., 2018

**Figure 1** LPI score as a percentage of the highest score by quintile average, 2012, 2014, 2016, and 2018. Source: Arvis et al., 2018.**Table 2** Top 10 LPI economies, 2018

Economy	2018		2016		2014		2012	
	Rank	Score	Rank	Score	Rank	Score	Rank	Score
Germany	1	4.20	1	4.23	1	4.12	4	4.03
Sweden	2	4.05	3	4.20	6	3.96	13	3.85
Belgium	3	4.04	6	4.11	3	4.04	7	3.98
Austria	4	4.03	7	4.10	22	3.65	11	3.89
Japan	5	4.03	12	3.97	10	3.91	8	3.93
Netherlands	6	4.02	4	4.19	2	4.05	5	4.02
Singapore	7	4.00	5	4.14	5	4.00	1	4.13
Denmark	8	3.99	17	3.82	17	3.78	6	4.02
United Kingdom	9	3.99	8	4.07	4	4.01	10	3.90
Finland	10	3.97	15	3.92	24	3.62	3	4.05

Source: Arvis et al., 2018.

the 15 best-performing countries has not significantly changed either. But it is worth highlighting major improvements in the LPI scores of Japan, Denmark, the United Arab Emirates, and New Zealand since 2012.

The bottom 10 countries are mostly low income and lower-middle-income countries in Africa or isolated areas (Table 3). Some are fragile economies affected by armed conflict, natural disasters and political unrest. Others are landlocked countries naturally challenged by geography or economies of scale in connecting to global supply chains. Afghanistan ranks 160th with a score 1.95, preceded by Angola (2.05), Burundi (2.06), and Niger (2.07).

The composition of the top-performing upper-middle-income economies has changed marginally, with China (26th with a score of 3.61), Thailand (32nd with a score of 3.41), and South Africa (33rd with a score of 3.38) leading the group (Table 4). Romania,

Table 3 Bottom 10 LPI economies, 2018

Economy	2018		2016		2014		2012	
	Rank	Score	Rank	Score	Rank	Score	Rank	Score
Afghanistan	160	1.95	150	2.14	158	2.07	135	2.30
Angola	159	2.05	139	2.24	112	2.54	138	2.28
Burundi	158	2.06	107	2.51	107	2.57	155	1.61
Niger	157	2.07	100	2.56	130	2.39	87	2.69
Sierra Leone	156	2.08	155	2.03	na	na	150	2.08
Eritrea	155	2.09	144	2.17	156	2.08	147	2.11
Libya	154	2.11	137	2.26	118	2.50	137	2.28
Haiti	153	2.11	159	1.72	144	2.27	153	2.03
Zimbabwe	152	2.12	151	2.08	137	2.34	103	2.55
Central African Republic	151	2.15	na	na	134	2.36	98	2.57

Source: *Arvis et al., 2018*.**Table 4** Top-performing upper-middle-income economies, 2018

Economy	2018		2016		2014		2012	
	Rank	Score	Rank	Score	Rank	Score	Rank	Score
China	26	3.61	27	3.66	28	3.53	26	3.52
Thailand	32	3.41	45	3.26	35	3.43	38	3.18
South Africa	33	3.38	20	3.78	34	3.43	23	3.67
Panama	38	3.28	40	3.34	45	3.19	61	2.93
Malaysia	41	3.22	32	3.43	25	3.59	29	3.49
Turkey	47	3.15	34	3.42	30	3.50	27	3.51
Romania	48	3.12	60	2.99	40	3.26	54	3.00
Croatia	49	3.10	51	3.16	55	3.05	42	3.16
Mexico	51	3.05	54	3.11	50	3.13	47	3.06
Bulgaria	52	3.03	72	2.81	47	3.16	36	3.21

Source: *Arvis et al., 2018*.**Table 5** Top-performing lower-middle-income economies, 2018

Economy	2018		2016		2014		2012	
	Rank	Score	Rank	Score	Rank	Score	Rank	Score
Vietnam	39	3.27	64	2.98	48	3.15	53	3.00
India	44	3.18	35	3.42	54	3.08	46	3.08
Indonesia	46	3.15	63	2.98	53	3.08	59	2.94
Côte d'Ivoire	50	3.08	95	2.60	79	2.76	83	2.73
Philippines	60	2.90	71	2.86	57	3.00	52	3.02
Ukraine	66	2.83	80	2.74	61	2.98	66	2.85
Egypt, Arab Republic	67	2.82	49	3.18	62	2.97	57	2.98
Kenya	68	2.81	42	3.33	74	2.81	122	2.43
Lao PDR	82	2.70	152	2.07	131	2.39	109	2.50
Jordan	84	2.69	67	2.96	68	2.87	102	2.56

Source: *Arvis et al., 2018*.

Croatia, and Bulgaria also improved their rankings. Among low-income countries, those in East and West Africa lead in this year's edition.

Among the lower-middle-income countries, large economies such as India (44th with a score of 3.18) and Indonesia (46th with a score of 3.15) and emerging economies such as Vietnam (39th with a score of 3.27) and Côte d'Ivoire (50th with a score of 3.08) stand out as top performers (**Table 5**). Most of these countries either have access to the sea or are located close to major transportation hubs.

Among low-income countries, countries in East and West Africa are leading performers in the 2018 report (**Table 6**).

Table 6 Top-performing low-income economies, 2018

Economy	2018		2016		2014		2012	
	Rank	Score	Rank	Score	Rank	Score	Rank	Score
Rwanda	57	2.97	62	2.99	80	2.76	139	2.27
Benin	76	2.75	115	2.43	109	2.56	67	2.85
Burkina Faso	91	2.62	81	2.73	98	2.64	134	2.32
Mali	96	2.59	109	2.50	119	2.50	na	na
Malawi	97	2.59	na	na	73	2.81	73	2.81
Uganda	102	2.58	58	3.04	na	na	na	na
Comoros	107	2.56	98	2.58	128	2.40	146	2.14
Nepal	114	2.51	124	2.38	105	2.59	151	2.04
Togo	118	2.45	92	2.62	139	2.32	97	2.58
Congo, Dem. Rep.	120	2.43	127	2.38	159	1.88	143	2.21

Source: Arvis et al., 2018.

Supply Chain Reliability and Service Quality are Strongly Associated with Logistics Performance

Supply chain reliability is key to logistics performance. In a global environment, consignees require a high degree of certainty as to when and how deliveries will take place. Reliability is typically much more important than speed, and many shippers are willing to pay a premium. In other words, supply chain predictability is a matter not just of time and cost, but also a component of shipment quality (Fig. 2).

In the top LPI quintile, just 13% of shipments fail to meet company quality criteria—the same proportion as in 2014 and 2016. By comparison, two to three times as many shipments in the two bottom quintiles fail to meet these criteria, and quality criteria in low-performing countries tend to be less rigid than in high-performing ones. This finding illustrates the persistence of the logistics gap from an overall perspective of supply chain efficiency and reliability.

The differing pace of progress is also seen in the ratings of domestic trade and transport infrastructure, where respondents were asked to assess how much these have improved since 2015. As in previous surveys, satisfaction with infrastructure quality varies by infrastructure type. For the first time, however, the perceived improvement is higher in the bottom quintile than in the top, although the difference is weaker in the middle of the distribution.

Respondents in all LPI quintiles are highly satisfied with information and communications technology (ICT) infrastructure. The infrastructure gap continues to narrow, particularly between the top and the bottom, where the rate of improvement seems noticeably faster. Improvement in the middle quintiles is on a par with what has been observed previously. In contrast to ICT, rail infrastructure continues to elicit general dissatisfaction. Similar patterns emerge when the domestic LPI data on infrastructure are disaggregated by World Bank region, excluding high-income countries. ICT is rated at the top or very close to the top in all regions.

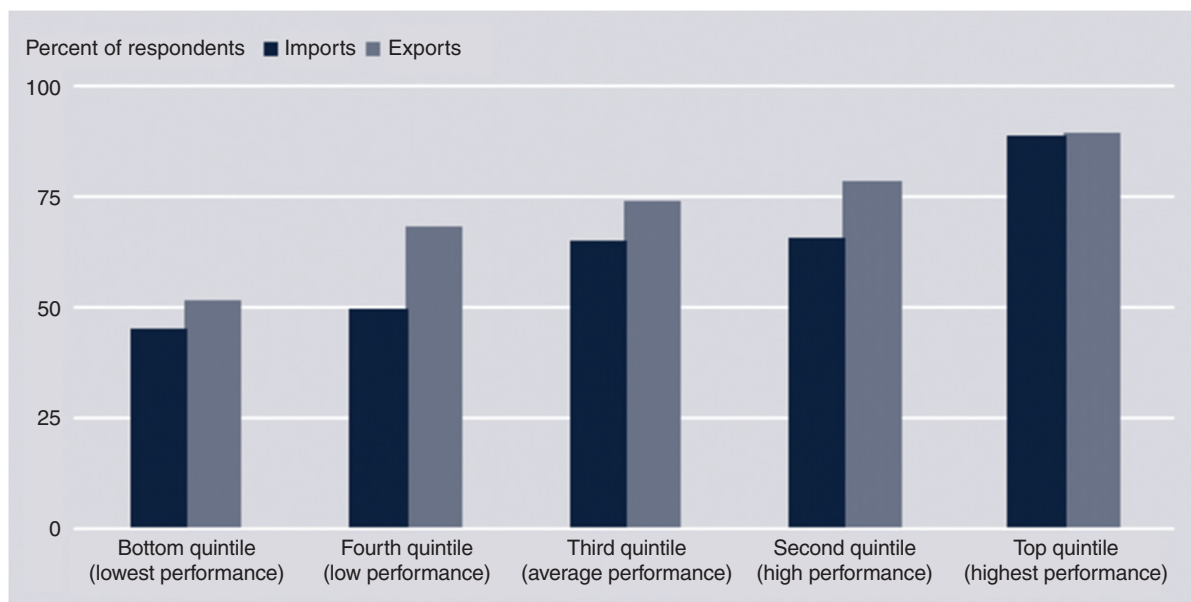


Figure 2 Respondents reporting that shipments are "often" or "nearly always" cleared and delivered as scheduled, by LPI quintile. Source: Arvis et al., 2018.

Delivering Good Quality Services is Key to Successful Operations, and its Importance is Growing

The LPI has shown that service quality drives logistics performance in practically all economies. Yet developing advanced services, such as third-party or fourth-party logistics, requires following a complex policy agenda, partly because such services cannot be created from scratch or developed purely domestically. In logistics-friendly countries, manufacturers and traders already outsource much of their basic logistics operations to third-party providers and focus on pursuing their core business while managing more complex supply chain issues.

This handoff is reciprocal: the more that such advanced services are available at a reasonable cost, the more shippers will outsource their logistics. But the less that reliable and comprehensive service are available, the more shippers will handle logistics in house.

Logistics services are provided under very different operational environments globally. In a pattern recurring across years, the quality of the services that logistics firms provide is often perceived as better than the quality of the corresponding infrastructure that they operate. This may be explained partly by who the respondents are—freight forwarders and logistics firms rating their own services.

In a pattern seen across LPI editions, operations that support international trade, such as air and maritime transport and supporting services, tend to receive high scores even when infrastructure bottlenecks exist. Railroads, on the other hand, have low ratings almost everywhere. Low-income countries score poorly on road freight and warehousing.

Service quality can differ substantially at similar levels of perceived infrastructure quality. Even high-quality “hard” infrastructure cannot substitute for operational excellence, based on “soft” infrastructure such as professional skills and smooth business and administrative processes.

Supply Chain Resilience and Sustainability are Emerging Concerns

The resilience of international and domestic supply chains has emerged as a growing policy concern worldwide. Resilience is understood as the ability of an organization (or a country) to recover from severe disruptions, whether caused by humans or natural. For 2018, the LPI survey included a question on cyber security resilience. The perceived magnitude of cyber threats and preparedness to mitigate their effects go hand in hand (Fig. 3). Developing countries lag far behind high-income countries in both.

The 2018 LPI survey confirms that demand for sustainable supply chain management goes hand in hand with logistics performance. This is especially true for environmentally sustainable services (green logistics). In the top quintile of LPI performers, 28% of respondents indicated that shippers often or nearly always ask for environmentally friendly options. In the second-highest quintile, the share drops to 14%, and it falls steadily in the third (9%), fourth (7%), and fifth (5%) quintiles.

This trend is in line with the increasing number of global and national commitments to reduce freight- and logistics-related greenhouse gases, particulate matter and other harmful emissions. Regulatory changes have been implemented in all transport modes and the international targets for 2030 and 2050, for example, are ever more challenging.

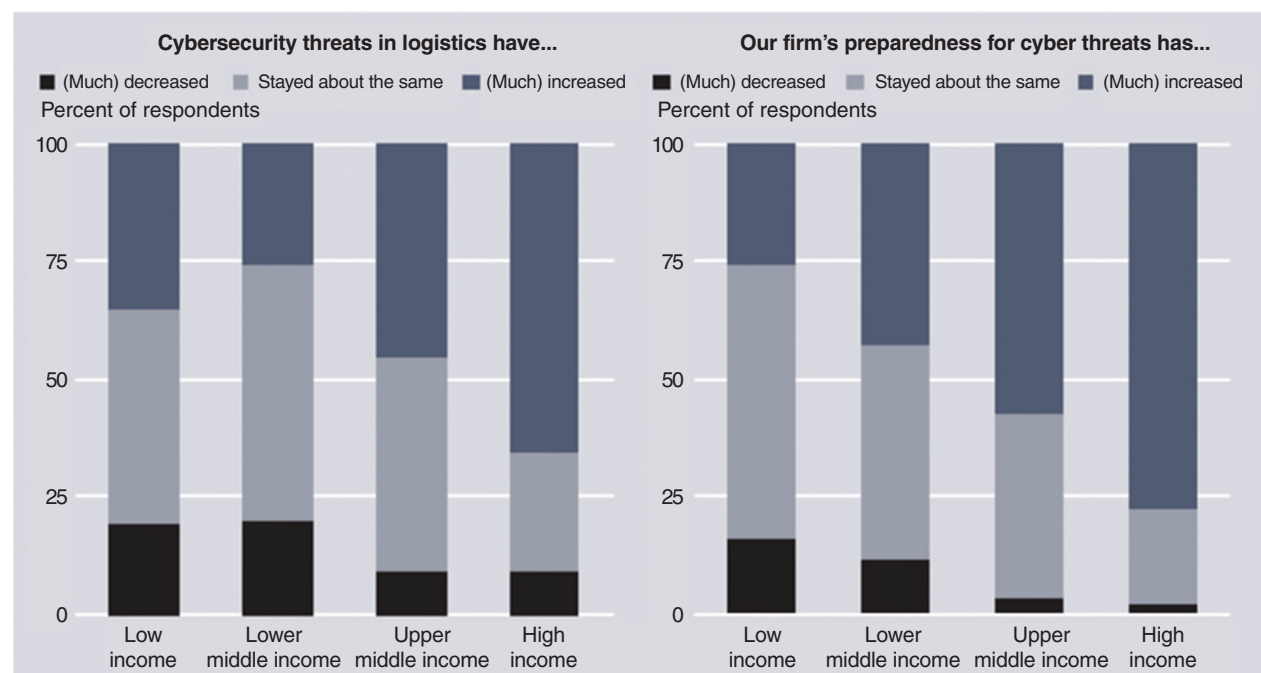


Figure 3 Cyber security threats and preparedness by income. Source: *Arvis et al., 2018*.

Pushing the Envelope of Implementation

The implementation of effective policies to improve logistics performance is at least as challenging today as in 2007—for two reasons. First, the scope of implementation has widened from the traditional focus on infrastructure and trade facilitation. Sustainability and resilience receive more attention, and not only by developed countries, as do skills development and training, spatial dimensions of logistics, and the specificity of the regulatory and legal framework.

In addition to these emerging fields, regulatory reforms of the logistics services sectors are critical but remain challenging to implement in many developing countries. Regulatory improvements aim to improve the quality of service delivery, building on market mechanisms and private sector participation, in the sectors that constitute the core of logistics activities, such as trucking, brokerage, and terminal or warehousing operations. The broad and crosscutting logistics agenda challenge policy makers to make sense of which policy measures are needed, when, and using what resources.

Second, most reforms involve more than one agency and many stakeholders, slowing implementation, or even reversing it, if cooperative mechanisms are not sustainable. This problem is well known in developing countries for transport (for example, transport corridors) and trade facilitation (e.g., single-window trade facilitation).

For consistent and broad reforms and improvements, countries must deal with this complexity. But countries in the middle and lower tiers of performance are more deterred by weaker coordination mechanisms and private sector constituencies than countries with modern and innovative logistics sectors. Even though logistics services are provided overwhelmingly by the private sector, public sector actors, and institutions play an essential role, without which logistics competitiveness is unlikely to improve.

Administrative reforms can be rapid when countries with strong political reforms will align their efforts. In some cases, soft reforms in the facilitation of trade and transport were implemented with considerable impact even before hard infrastructure projects were completed. The soft reforms provided a higher and quicker return on investment than hard infrastructure. Examples can be found in low- and middle-income countries such as India, Lao PDR, Southern African countries, and Vietnam, as well as in high-income countries such as Oman.

Unfortunately, performance may be degraded by governance weaknesses and economic and social turmoil, as for some Arab countries in the 2016 and 2018 LPI reports. Low-performing countries with serious governance challenges (conflict-ridden or post-conflict countries and fragile states) are the most in need of attention from their neighbors and the international community.

Ultimately, countries that introduce far-reaching changes appear to be those that treat logistics as integral to the economy. They tend to combine policy perspectives, such as regulatory reform, trade facilitation, and trade and investment planning. Seamless interagency coordination and, above all, strong public–private dialogue characterizes the top performers. They offer very positive examples of coordinating and facilitating logistics bodies, some of them public–private institutions, such as the most famous one, the Dutch Dinalog.

The Policymaking and Business Impact of the Logistics Performance Index

Since its inception in 2007, the Connecting to Compete report has moved trade logistics firmly onto the policy agenda, even for countries that had not previously considered it. LPI results have also been used in many policy reports and documents prepared by multilateral organizations or the consultants they have engaged. The findings provide a worldwide general benchmark for the logistics industry and for logistics users.

LPI results have been embraced also by the academic community, as evidenced by the widespread use of LPI data in research reports, journal articles, and textbooks. The results have also been used in teaching, and thousands of theses at all levels have cited the LPI.

Logistics performance is based largely on reliable supply chains and predictable service delivery for traders. Global supply chains are becoming more and more complex. Ever more demanding regulatory requirements for traders and operators are motivated by safety, social, environmental, and other reasons. Efficient management and information technology solutions in both the private and public sectors are tools for high-quality logistics. National competitiveness depends on the ability to manage logistics in today's global business environment. More than ever, comprehensive reforms and long-term commitments are needed from policy makers and private stakeholders. The current LPI data provide a unique and updated reference for better understanding the impediments to trade logistics worldwide and for informing policymaking and business decisions.

References

- Arvis, J.F., Mustra, M., Panzer, J., Ojala, L., Naula, T., 2007. Connecting to Compete 2007: Trade logistics in the global economy. The logistics performance index and its indicators, The World Bank, Washington, DC.
- Arvis, J.F., Mustra, M., Ojala, L., Shepherd, B., Saslavsky, D., 2010. Connecting to Compete 2010: Trade logistics in the global economy. The logistics performance index and its indicators, The World Bank, Washington, DC.
- Arvis, J.F., Mustra, M., Ojala, L., Shepherd, B., Saslavsky, D., 2012. Connecting to Compete 2012. Trade logistics in the global economy. The logistics performance index and its indicators, The World Bank, Washington, DC.
- Arvis, J.F., Saslavsky, D., Ojala, L., Shepherd, B., Busch, C., Raj, A., 2014. Connecting to Compete 2014. Trade logistics in the global economy. The logistics performance index and its indicators, The World Bank, Washington, DC.

- Arvis, J.F., Saslavsky, D., Ojala, L., Shepherd, B., Busch, C., Raj, A., Naula, T., 2016. Connecting to Compete 2016. Trade logistics in the global economy. The logistics performance index and its indicators, The World Bank, Washington, DC.
- Arvis, J.F., Ojala, L., Wiederer, C., Shepherd, B., Raj, A., Dairabayeva, K., Kiiski, T., 2018. Connecting to Compete 2018. Trade logistics in the global economy. The logistics performance index and its indicators, The World Bank, Washington, DC.

Further Reading

- Arvis, J.F., Shepherd, B., 2013. Trade cost around the World (2013). Policy Research Paper. The World Bank, Washington, DC.
- Memedovic, O., Ojala, L., Rodrigue, J.P., Naula, T., 2008. Fuelling the global value chains: what role for logistics capabilities? *Int. J. Technol. Learn. Innov. Dev.* 1 (3), 353–374.
- Shepherd, B., Archanskaia, L., 2014. Evaluation of value chain connectedness in the APEC region, APEC Policy Support Unit, Singapore.
- Shepherd, B., 2016. Trade Facilitation and Global Value Chains: Opportunities for Sustainable Development. International Center for Trade and Sustainable Development (ICTSD), Geneva.
- Solakivi, T., Ojala, L., Laari, S., Lorentz, H., Toyli, J., Malmsten, J., Lehtinen, N., Ojala, E., 2017. Finland State of Logistics 2016. Turku School of Economics Publications, Turku.
- World Bank, 2010. Trade and Transport facilitation assessment: a practical toolkit for country implementation. Available from: http://siteresources.worldbank.org/EXTTLF/Resources/Trade&Transport_Facilitation_Assessment_Practical_Toolkit.pdf.

Outsourcing Logistics Functions

Evi Hartmann, Hendrik Birkel, Matthias Kopyto, Lange Gasse, Nuremberg, Germany

© 2021 Elsevier Ltd. All rights reserved.

Overview of Logistics Outsourcing	102
Logistics Service Providers	102
The Outsourcing Process and Determinants	103
Strategy Phase	103
Selection Phase	103
Design Phase	104
Operation Phase	104
Potential Advantages and Risks of Logistics Outsourcing	104
Advantages of Logistics Outsourcing	104
Risks of Logistics Outsourcing	105
Issues and Influences	106
References	106
Further Reading	106

Overview of Logistics Outsourcing

The term outsourcing consists of the words *outside*, *resource*, and *using*, respectively, and evolved in the 1970s and 1980s when companies found themselves facing increasing global competition. To resolve the prevailing lack of agility, many large companies decided to sell noncore activities and focus on their core business. At the same time, the deregulation of transport in the Western world facilitated companies to externalize their road freight transport. However, the growth of contracting out did not emerge until the change in managerial attitudes in the late 1980s. Due to rising cost pressures in the 1990s, the proportion of outsourced functions increased further, especially in the logistics area, but also in other fields such as information systems, human resources, or accounting (Pickles et al., 2016).

Outsourcing can be described as the execution and management of activities or functions by external specialized service providers and includes both material goods and services. Notable are the lack of a uniform definition and the existence of a variety of descriptions, as well as definitional inconsistencies, within the concept of outsourcing. It can be divided into partial and total outsourcing. While in partial outsourcing only certain tasks are transferred, in total outsourcing, an entire functional area is externalized and the corresponding functions are completely eliminated internally. In general, the outsourcing of functions aims at establishing long-term and lasting partnerships. However, the intensity of the cooperation can vary and, for example, be increased in order to overcome short-term bottlenecks. The opposite of outsourcing is represented by insourcing and is defined as the reintegration of processes back into the focal company.

Logistics constitutes one of the classic corporate functions and is one of the most frequently outsourced operations. In most industries, the majority of logistics costs are attributable to external service providers and a further increase in share is expected (Lemoine and Dagns, 2003). The reasons which prompt and favor the outsourcing of logistical functions are manifold, such as the increasing specialization of companies, the reduction of vertical integration and associated decreasing complexity, or the trend toward the lowering of high asset intensity of logistic functions, such as infrastructure for transportation, handling, or storage. In addition to the high resource intensity, the complexity of processes continues to increase due to the higher level of globalization, shorter lead-time expectations, and higher requirements for value-added processes. Further crucial aspects are the shift of production to low-cost economies, growing global brands and retailers, as well as the greater availability of information and communication technologies.

At the same time, logistics offers a wide range of specialized elements and, therefore, a large number of tasks that can be outsourced. Logistics services can be subdivided into four categories: transport services (e.g., internal transport, collection and distribution tours, route planning, and route optimization), handling services (e.g., loading and unloading, sorting, packing, and stowage optimization), warehousing services (e.g., storage and retrieval, order picking, quality inspection, and storage location management), and value-added services (e.g., filling, customs clearance, collection, and assembly work). The most frequently outsourced activities are domestic and international transport, warehousing, freight forwarding, and customs brokerage, which are all characterized by transactional, operational, and repetitive characteristics.

Logistics Service Providers

Logistics service providers form the core of outsourcing. In the literature, there are several possibilities for classification. A general differentiation is the classification into individual service providers, joint service providers, and system service providers. An

individual service provider offers a specialized range of services, while a joint service provider integrates several individual services and offers them to a broad customer field. In contrast, the system service provider is fully responsible for operating an integrated logistics system for one or more customers in long-term business relationships. Its special feature is the comprehensive range of services and the particularly high proportion of non-logistics applications.

The most commonly and intensively discussed concept to categorize service providers was developed by the management consultancy Accenture and classifies the basic types of logistics service providers according to their increasing integration into a customer's logistics chain. The classification encompasses first-party logistics (1PL) to fourth-party logistics (4PL) and lead logistics providers (LLPs). The 1PL (carrier) is a small entrepreneur or transport operator with core competence as a freight specialist or warehouse operator, high-handling competence, and a high level of service. The 1PL relies mainly on its own resources and is often used as a subcontractor. The second-party logistics (2PL) (operator) is a classic freight service provider for transport orders and, as a specialist for order logistics, can be compared with the joint service provider. The operator partly uses its own resources and has disposition centers and settlement systems. Its core competence is in order logistics with a focus on one mode of transport. By purchasing transport services on its own, the 2PL enables complex combined transports. This is achieved through its expertise in IT-supported processing systems and the processing competence across several carrier systems. The third-party logistics (3PL) (solution provider) offers comprehensive solutions and focuses on contract logistics in selected industries. It has its own resources in logistics, including warehouses and disposition areas, and its core competences are the provision of customer-specific solutions, the integration of the process chain regarding warehouse management, and the fulfillment of customer requirements. The 4PL (solution integrator), which emerged from the 3PL, is also specialized in individual solutions. The 4PL has a comprehensive expertise in the design, integration, coordination, and management of networks in individual industries but, in contrast, does not own assets in the field of classic logistics, such as trucks, warehouses, or distribution centers. This also applies to the LLP (solution designer). However, in contrast to the 4PL, the LLP does not act as an operator of the customer-specific solution itself. Its core competence lies in consulting.

The distinction between 3PL, 4PL, and LLP in the literature is often ambiguous due to overlapping tasks and responsibilities. This leads to diverse information about the share of the individual providers. In practice, several mixed forms of 3PL, 4PL, and LLP exist.

The Outsourcing Process and Determinants

The outsourcing process consists of a large number of complex, parallel activities. The implementation involves a variety of factors, circumstances, challenges, and potential effects, which have to be considered in a comprehensive context to be able to reach a reliable decision and ensure the success of the outsourcing procedure. Depending on the task to be outsourced, different priorities have to be defined, including strategic and operational aspects. Overall, regarding the single outsourcing phases, a high degree of transparency and rationality is required, which facilitates discussions and thereby consensus-building processes. Similar to the definition of providers, there is also a multitude of different procedure models for the process of outsourcing. In general, the procedure can be categorized into four phases: strategy, selection, design, and operation phase.

Strategy Phase

The strategy phase includes the description of the initial situation and the objectives, as well as the definition of the outsourcing strategy. The competitive situation, core competencies, and potential transaction costs are analyzed for an additional assessment of the strategic orientation. This also includes the definition of requirements for a suitable logistics service provider based on the previously defined outsourcing goals. The preparation of requirements, evaluation standards, control parameters, and further objectives is of particular importance, because the defined strategic aspects have a considerable influence on outsourcing success and can hardly be adapted subsequently. Furthermore, the definition and distribution of responsibilities, tasks, and decision-making powers have to be determined in order to exclude irritations and conflicts of competence.

Selection Phase

The selection phase deals with the choice of partners. First, a bidder preselection is conducted based on the previously developed requirement specifications. This is followed by the call for tenders, an evaluation, and bidder negotiations. For this purpose, the in-house production and external procurement costs are evaluated, which serve as the basis for the make-or-buy decision. Both the strategy and selection phase should be concluded with a comprehensive screening of the logistics service providers. The entire process is iterative in order to enable continuous improvement of the individual determinants and thus ensure the highest possible fit between the corporate cultures of the cooperation partners. In addition to the soft factors, the fit of hard factors, such as information systems or service level agreements, is guaranteed. Another success factor, which should be included, is the logistics service provider's focus on innovation and optimization, as these efforts will reflect the contractual partner's long-term aptitude and commitment. However, sustainable success can only be ensured through the pursuit of innovation and a fair sharing of the improvement potential.

Design Phase

The design phase includes all steps for the successful initiation of the outsourcing project. For this purpose, a consistent target vision of all parties must be created, which is followed by the definition of the structure and standardization of the interfaces and business processes. In addition, the respective resources, the transfer of personnel, and the cost structure have to be planned and recorded in a contract. With special regard to the design phase, decisions on competences and prices of outsourcing are important. However, the determination of competences from the customer's point of view is fraught with uncertainty. In order to ensure the success of outsourcing, reputation and references from previous project experience and perceived expertise of the service provider can be considered. However, this should not only be taken into account in the design but also in the preceding phases. The outsourcing of logistic processes plays a special role in this context, since in general a high degree of complex tasks is outsourced in comparison to other business areas, which at the same time require error-free integration. Despite the uniqueness of projects, the company's knowledge in outsourcing has a crucial impact, as the experience of already completed projects in logistics can be transferred.

Operation Phase

The operation phase comprises operational aspects for the successful maintenance of the outsourced tasks. During the ongoing operation, continuous management and control with a regulated exchange of information is essential and, to a large extent, determines the quality of the cooperation. In addition to the selection and design process, the implementation of the outsourcing project plays an important role. In contrast to the selection phase, mainly qualitative factors determine the success or failure of the implementation. The shift of company boundaries and the associated interventions in the process structures and resource distribution increase the cross-company interfaces and, thereby, the effort expended in coordination. This issue and the influence of outsourcing on many company processes require the implementation of multidisciplinary teams. Not only employees within the company and the logistics service provider must be involved, but also external stakeholders, as well as the top management.

Continuous and comprehensive communication within the business relationship reduces the risk of conflicts, strengthens trust, and prevents conflicts and opportunism (Swann, 2009). Communication can be divided into formal and informal. In contrast to informal communication, formal communication is based on predetermined communication channels and contents. Which information content and information channels are most important for the outsourcing partnership depends on the specific case, but both communication channels must be developed and enabled due to their complementary character. A further central element of the exchange of information is the regular controlling of the interfaces and the services to be provided within the outsourcing partnership. In this context, controlling does not simply rely on the control of the service provided, but the coordination and provision of information for optimization measures. Therefore, the open and cooperative exchange of relevant information is one of the most important determinants. In a successful partnership, long-term cooperative development with common goals should evolve as part of the outsourcing process. However, the partnership may also fail, resulting in a termination of the project and the reintegration of the tasks.

Potential Advantages and Risks of Logistics Outsourcing

The transfer of a company's own business processes to external partners is associated with both positive and negative consequences, which have to be analyzed extensively before a decision should be made. While some aspects apply to outsourcing in general, the outsourcing of logistic functions requires special consideration.

Advantages of Logistics Outsourcing

The decision to outsource logistics activities is primarily motivated by strategic and operational reasons, although a distinct assignment of the benefits to one of the two groups is difficult. The following description highlights the multitude of benefits of both groups and their interdependencies and illustrates the close link to a further important dimension: the cost-related dimension.

From a strategic point of view, outsourcing enables a generous variabilization of fixed costs. Transferring logistics activities reduces capital commitment and investment risks, as outsourcing companies have to allocate few resources to capital-intensive assets such as vehicles, warehouses, or information technology. These investments are carried by the logistics service provider itself, enabling outsourcing companies to focus their limited capital to a greater extent on their own core business areas and increase efficiency. Furthermore, obtaining the services from a third-party provider for a contractually agreed period guarantees strategic flexibility. This, in turn, allows the outsourcing company to continuously adapt the contracts to current market conditions or to change the service provider.

In addition, the transfer of nonbusiness-related logistics activities helps to reduce vertical integration, as well as internal complexity, enabling the companies to concentrate the released resources even more on their core competencies and strategically relevant operations. Most particularly, the coordinating support provided by a 4PL is gaining importance for a large number of suppliers and customers.

The operational side also benefits from a number of advantages from outsourcing. For example, by reducing the number of processes and the complexity, the degree of specialization of the remaining processes can be increased. In addition, outsourcing

allows access to the know-how of other external providers and reduces the danger of operational blindness. The operational execution of logistics processes by external service providers leads to higher productivity, flexibility, and quality of logistics services. This results in a multitude of advantages, such as faster lead times, improved on-time deliveries, less damage, and higher responsiveness to market changes. Logistics service providers are able to offer all the services at competitive prices, which cannot be realized internally by the outsourcing firm. On the one hand, this benefit is achieved through the advantage of volume-dependent cost depressions (economies of scale) due to the providers' high volumes and market shares. On the other hand, specialized logistics service providers are able to operate with lower personnel costs, mainly due to lower wage levels and working time regulations in the logistics sector compared to other industries, as well as the more rational use of personnel and higher personnel productivity.

The benefit of increased flexibility not only applies to deliveries to the end customer, but also within the company itself, since outsourcing usually allows for a decrease in workforce and enables the transfer of risks and intricacies regarding labor regulations or unions. Other risks that can be transferred are, for example, the rising costs of input factors (e.g., fuel), difficulties with fluctuations in transportation volumes, or customer changes to delivery requests.

With growing pressure to differentiate, service providers increasingly offer innovative solutions, which further improve the services and benefits for the outsourcing company. This includes, for example, not only value-added services, such as tracking and tracing or insurance services, but also the supply of concrete value-adding processes such as preassembly or quality control.

Finally, companies that engage in outsourcing also profit from two beneficial side effects. By conducting a comprehensive analysis of strengths and weaknesses, the basis is created for testing their own internal business processes against changed external conditions, and adapting them if necessary. Additionally, the input from independent third parties reduces the risk of a company's shortsightedness.

Risks of Logistics Outsourcing

The potential benefits of outsourcing are contrasted by risks, which could minimize or even negate the positive effects of outsourcing. Because of this, they have to be included in both the initial outsourcing decision and the subsequent evaluation of the partnership. One of the key benefits, cost reduction, is also a crucial potential risk at the same time. A common threat lies in the erroneous calculation of one's own cost structure, which can lead to a positive outsourcing decision even though the in-house costs might be lower. The reason for this is the insufficient cost accounting basis of companies, which often means costs cannot be calculated in detail and prevents sufficient differentiation between the individual logistics services.

In addition, the risk of being unable to achieve the calculated cost-saving potentials exists, which can be attributed to several aspects. First, companies may fail to reduce fixed costs to the necessary extent. This is the case if the released capacities cannot be reduced by the same amount or used for other purposes. Second, the total costs of outsourcing may be higher than planned due to hidden or underestimated costs. For example, the coordination effort between the partners and the corresponding transaction costs is often estimated at too low a level beforehand. Such additional costs during the project represent a large burden for the company, but may have a negative impact on the relationship with the partner as well. The higher coordination effort can also be seen as a form of complexity that counteracts the reduction in complexity previously mentioned in the advantages section.

Furthermore, companies increase their dependency through the transfer of formerly internal company processes and the disclosure of information. This disclosure leads to a vulnerability for opportunistic behavior from logistics service providers, for example, in the form of unexpected, subsequent price increases. Thereby, the dependency toward the contracted companies rises with an increasing specificity of the outsourced process or potential lost knowledge.

Despite careful preparation and selection, outsourcing partners can prove to be inefficient, which manifests itself, for example, in the inadequate quality of logistics services or noncompliance with delivery times. Since logistics service providers have direct contact with customers, but customers usually associate service quality with the focal company, the inadequate performance of the providers translates into an immediate damage of reputation for the client company. A contracting company can further be exposed to major reputational damage if the service provider becomes conspicuous through poor working conditions, exploitation of employees, or illegal practices. These aspects necessitate continuous monitoring and follow-up evaluation of the relationship with contracted logistics service providers.

A further risk, which could lead to the failure of the outsourcing project, is poor data exchange and poor communication between the partners. This can occur at any level of personnel and is another bottleneck for top-level performance. Mutual trust is an essential aspect, because the customer in particular must show a willingness to grant access to internal data, even if this leads to a disclosure and exposes the focal company to the risk of this data being misused or shared illicitly. Generally, the sovereignty over the controlling of logistics performance is lost, as the data are reported by the logistics service providers and they could provide embellished data. Preventing these issues would require further control mechanisms, which again would entail higher transaction costs.

Lastly, the advantage of risk shifting must also be considered in detail, as if the logistics service provider bears the risk on its own and gets into financial difficulties, the outsourcing company must provide financial support in order to further guarantee operational capability. This, in turn, puts the risks on the contracting company. This risk is particularly relevant for small logistics service providers, as their financial resources are limited. Empirical evidence also shows that many outsourcing projects only deliver suboptimal results, or even fail completely, because too little attention is paid to the initial strategic goal definition, cost analysis, and process design.

Issues and Influences

The characteristic of logistics to link companies, supply chains, and individuals globally exposes the field to economic, political, social, and technological developments and, therefore, leads to a changing environment for logistics outsourcing as well. Changes in the past were driven by a multitude of circumstances, such as constantly rising customer demands, the growing linkage between companies, and the evolution of information technology. Nowadays, the industry is influenced by even more factors and developments, which will change the requirements and design of logistics outsourcing operations further.

A fundamental change is the demand for sustainability, which will continue to rise in importance and significantly influence the logistics market. The increasing awareness in society of the need to reduce CO₂ emissions affects companies, the production of goods, and, as a consequence, transportation chains and decisions on outsourcing (Lieb and Lieb, 2010). In the future, companies will urge logistics service providers to carry out their freight transport using alternative means of transport with environmentally friendly power trains. Aggravating trends, such as the growing density of cities, severe traffic congestion, or possible daytime driving restrictions for deliveries, require a special focus on the last mile, as well as general alternatives for transportation (Zijm and Klumpp, 2017). Outsourcing companies will shift the development of solutions to these questions to logistics service providers. Ecological sustainability and corporate social responsibility will become established standards and lose their function as a differentiating feature for logistics service providers. The service providers, therefore, need to address new fields in order to distinguish themselves from the competition. Building on the earlier, outsourcing companies will call for further integrated solutions, instead of individual services, from additional value-added services to proposed solutions for traffic reduction, stockless warehouse operations close to high-density centers, or the management of reverse-logistics flows.

Another factor that will have a particular impact on future logistics outsourcing activities is the high volatility in markets, competition, and politics. Triggered by market changes, supply chain vulnerability, and optimization efforts, outsourcing companies will increasingly dynamically review and change the elements of their value chains. This requires logistics service providers to make their offerings more flexible in order to respond to the highly dynamic context. Therefore, both partners will have to intensify the networking and integration of their IT systems. Advances in information technology and the use of new technologies will create new types of collaboration, foster the reduction of risks associated with the exchange of internal information, and ensure increased security expectations.

Political developments that will influence the overall outsourcing strategies of companies refer to uncertainties regarding the openness of markets. For example, the European Union follows the strategy to facilitate the exchange of goods between all its member states, including the harmonization of taxes and legislation between their member countries. At the same time, the calls for protectionism intensify in many countries around the world. The country-specific developments particularly affect the cross-border movement of goods, but indirect consequences also impact the operations of regionally active service providers. Furthermore, legal issues such as new laws, increasing security regulations, and bureaucratization will influence outsourcing activities continuously.

References

- Lemoine, W., Dagnæs, L., 2003. Globalisation strategies and business organisation of a network of logistics service providers. *Int. J. Phys. Distr. Logist. Manage.* 33 (3), 209–228. Available from: <https://doi.org/10.1108/09600030310471961>.
- Lieb, K., Lieb, R., 2010. Environmental sustainability in the third-party logistics (3PL) industry. *Int. J. Phys. Distr. Logist. Manage.* 40 (7), 524–533. Available from: <https://doi.org/10.1108/09600031011071984>.
- Pickles, J., Smith, A., Begg, R., et al., 2016. *Articulations of Capital: Global Production Networks and Regional Transformations*. Wiley-Blackwell, Oxford.
- Swann, G.M.P., 2009. *The economics of innovation: an introduction*. Edward Elgar Publishing, New York.
- Zijm, H., Klumpp, M., 2017. Future logistics: what to expect, how to adapt. In: Freitag, M., Kotzab, H., Pannek, J. (Eds.), *Dynamics in Logistics*. Lecture Notes in Logistics. Springer, Cham.

Further Reading

- Deepen, J.M., 2007. *Logistics Outsourcing Relationships: Measurement, Antecedents, and Effects of Logistics Outsourcing Performance (Contributions to Management Science)*. Springer, Heidelberg.
- Garrett, M., 2014. *Encyclopedia of Transportation: Social Science and Policy*. Sage Publications, Thousand Oaks, CA.
- Grossman, G.M., Helpman, E., 2005. Outsourcing in a global economy. *Rev. Econ. Studies* 72 (1), 135–159.
- Huiskonen, J., Pirttilä, T., 2002. Lateral coordination in a logistics outsourcing relationship. *Int. J. Prod. Econ.* 78 (2), 177–185.
- Knemeyer, A.M., Corsi, T.M., Murphy, P.R., 2003. Logistics outsourcing relationships: customer perspectives. *J. Bus. Logist.* 24 (1), 77–109.
- Kress, M., 2016. *Operational logistics: the art and science of sustaining military operations*. In: *Management for Professionals*, second ed. Springer, New York.
- Mattfeld, D.C., 2016. *Logistics Management: Contributions of the Section Logistics of the German Academic Association for Business Research*, 2015, Braunschweig, Germany. Lecture Notes in Logistics. Springer, Switzerland.
- Win, A., 2008. The value a 4PL provider can contribute to an organisation. *Int. J. Phys. Distr. Logist. Manage.* 38 (9), 674–684.

Freight Transport and Logistics in JIT Systems

James H. Bookbinder*, M. Ali Ülkü*, *Department of Management Sciences, University of Waterloo, Waterloo, ON, Canada; †Rowe School of Business, Dalhousie University, Halifax, NS, Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction	107
JIT Success, Pros and Cons	107
Transportation in a JIT System	108
A Possible JIT Contradiction	108
Toyota Mixed-Load System	108
Use of a JIT Warehouse	108
Mode of Transportation	109
Outbound Versus Inbound Transportation	109
Inventory-Transportation Tradeoffs in JIT	109
Continuous Improvement and JIT Logistics	110
JIT-L and Sustainability	111
Concluding Remarks	111
Biographies	112
References	112

Introduction

Just-in-Time (JIT), sometimes referred to as a “lean” system, is a management philosophy based on continuous search for improvement in ways to make processes more efficient. The goal is to produce goods and services without incurring any waste. JIT originated in Japan in the 1950s, helping Japanese companies, notably Toyota Motor Corporation, the world’s largest automaker, to enhance their productivities and cost efficiencies. Industry-wide diffusion of JIT was largely due to the 1973 oil embargo, and increased shortages of natural resources in Japan.

The Toyota Production System (TPS) is famous for elimination of “waste,” which was defined by Fujio Cho of Toyota as “ . . . anything other than the minimum amount of equipment, material, parts, and working time which are absolutely essential in production.” And Shigeo Shingo, the father of TPS, drew attention to the broad applicability of JIT by saying (Shingo, 1988) “ . . . the goal of improvement now is not simply to reduce the magnitude of problems, but to eliminate them entirely. Attention is focused on the causes of evils previously accepted as unavoidable, and steady progress is being made.”

JIT can thus trace its origins to *manufacturing*, TPS, and related approaches such as zero-inventory systems and zero-defect systems. Although there are subtle differences between the preceding, their similarities are what concern us here. Let us take as “given,” that the JIT system is well-suited to the appropriate manufacturing contexts.

The emphasis of this chapter, then, is on two questions:

- How can the methods and procedures of logistics and freight transport best enable a JIT production system to function well?
- And in designing those procedures, can we apply the JIT principles to logistics?

Our third and fourth sections, “Transportation in a JIT System” and “Inventory-Transportation Tradeoffs in JIT,” deal with the first question; the second is discussed in the sections that follow, on “Continuous Improvement and JIT Logistics” and “JIT-L and Sustainability.” Waste, in a logistics sense, includes excess inventory, inefficient routing of parts, and unnecessary movement or waiting. For example, delayed delivery of an order, due to a backorder at suppliers, is a waste. The extra time needed to deliver the order is of no value to the shipper.

The JIT approach is seen to involve potential cost-saving and is thereby a competitive strategy. JIT applications in business logistics concern, for example, the move-store activities within the supply chain. We will stress applications in the auto industry, both for historical reasons and because that industry has had excellent results in its utilization of JIT.

JIT Success, Pros and Cons

Following its stunning triumph at Toyota, the JIT approach reached North America a few decades ago. In the mid 1980s, General Motors decided to partner with Toyota to resuscitate its previously solely-owned auto-manufacturing plant in Fremont, California. That plant had needed to close after suffering low productivity indicators. This joint venture, naturally ingraining the JIT approach in its operations, showed tremendous success: Most of the laid-off plant workers were rehired. They were trained to improve

performance by eliminating waste. The morale of the workers increased, resulting in almost no absenteeism. The productivity and quality of the products drastically improved so much, that Fremont surpassed all other GM plants!

Such JIT success stories abound. This renewed definition of waste, and its accompanying goals of eliminating disruptions (e.g., quality problems during manufacturing) and of making the system flexible (e.g., faster setups), made JIT an eye-opener. JIT has become popular among other industries, including services and healthcare. Of course, the freight logistics industry is no exception in benefiting from the broad applicability of the JIT philosophy. Even more so in recent decades, firms realize that driving excessive logistics costs and wastes out of their supply chains emerges as a most viable competitive weapon for an enduring position in the marketplace.

Consider, for example, how Walmart became the world's largest retailer, via improving its logistics tactics by employing cross-docking. With the advent of emerging technologies in production (e.g., full automation) and transportation (such as increased use of drones in local deliveries and of driverless trucks for long-hauls), the JIT approach may be even more applicable in freight transportation and logistics systems.

As previously noted, JIT system benefits include decreased waste, lower costs, increased quality, cycle time reduction via eliminating non-value added operations, enhanced flexibility (from quick changeovers and smaller lot sizes), and improved productivity. Yet, JIT brings some drawbacks. For example, greater responsibility of workers may put extra stress on their shoulders; when unseen problems surface, the fewer available resources (time, people, inventory) may be insufficient to overcome supply chain disruptions, perhaps halting production. These pros and cons of JIT should be kept in mind. We also note that JIT involves cross-functional activities within an organization that extend to its supply chain. Within the firm, support from various functional groups (mainly operations, marketing, and finance), and between the supply chain agents, is essential if JIT is to achieve its full potential.

Transportation in a JIT System

A Possible JIT Contradiction

Here is an important issue, which a Just-In-Time system needs to address. Such a system famously delivers small quantities, repeatedly, rather than sending fewer, larger, "lot-size" quantities. Those JIT-type, frequent shipments, inbound to a manufacturing plant, likely worsen traffic congestion (a societal problem). More frequent deliveries of shipments result in extra vehicles on the roads, increasing emissions and creating noise pollution (environmental problems). The JIT approach does lower inventory cost by reducing necessary safety stocks, but JIT then increases transportation costs. Higher inventory costs associated with greater stock levels may be partly offset by savings in using less expensive transport (e.g., rail). However, in light of the considerable emissions due to (often unnecessary) frequent deliveries, the value of JIT-L needs to be justified.

Resolution of the preceding issue might again lie with suitable integration of an SCL (shipment consolidation) policy (Ülkü, 2009) and JIT-L. We provide several references, drawing attention to JIT's impact on traffic congestion (Rao et al., 1991), and the fact that SCL can reduce both costs and environmental damages (Ülkü, 2012). SCL and JIT-L can be considered side by side; both move largely less-than-truckload (LTL) type shipments. A parallel analysis of these two powerful logistics strategies is important.

Toyota Mixed-Load System

Can there be a system of *daily* deliveries to a plant, and still be consistent with zero waste (low emissions, fewer vehicles on the road)? The Toyota mixed-load system facilitates that. A group of five suppliers, say, would take turns on particular days of the week, in delivering its own components for that day plus having picked up the respective components of the other four suppliers. The truck of supplier A furnishes the transport on Monday, say, and supplier B plays that role on Tuesday, etc. This is analogous, of course, to a car pool taking children to school.

This mixed-load system enables, each day, deliveries of small amounts of each component (quantities required for that day's production). Perhaps equally important, a low-inventory system can now function in a "green" manner. That system permits daily deliveries of small quantities of the various items, without extra vehicles on the road or at the plant's loading dock.

Use of a JIT Warehouse

In Japan, the distances between suppliers and the (Toyota) plant are much smaller than in North America or Europe. The mixed-load system would be difficult to implement when the suppliers are that much further apart. Daily deliveries of the small quantities that the plant requires each day would be impractical (and also wasteful). While JIT is famous for its emphasis on minimal inventories, additional stocks may be required for a non-Japanese JIT system.

The solution is a "JIT warehouse." This would be operated by a 3PL, a third-party logistics service provider, and located as close (say a few kilometers away) to the assembly plant as practicable. Large (approximately TL) quantities of each component are shipped, for example, weekly, to the JIT Warehouse. The 3PL then makes daily, or even twice-daily, deliveries of the respective components to the proper loading docks at the plant.

It should be noted that expenses for warehousing and inventory are greater in this case than for the archetypal JIT system. Nevertheless, the total costs of inventory plus transportation in the above JIT system, as modified for application outside of Japan, is considerably less than it would be, if a “traditional” JIT approach were to be attempted.

Mode of Transportation

We tacitly assumed, above, that JIT transportation is by truck. And why not? Each shipper can dispatch goods by truck; any consignee can receive those loads. Not every supply-chain node has a rail siding, but those that do should not automatically assume that only shipments by truck can succeed in a JIT system. Indeed, the railways may be guilty of allowing the trucking industry to “eat their lunch.” A number of present or recent rail programs have been customized for JIT (Higginson and Bookbinder, 1990).

JIT transportation needs to be time-definite. Average speeds are slower for rail than for truck; when freight is hauled within a JIT system, the railway in question must give a *priority* to those dispatches. Shipment plans run by particular railroads were conceived for the JIT market. The French National Railway (SNCF) ran programs for JIT in which a commitment was made that the goods would arrive at particular destinations by a promised time. Dispatching decisions at intermediate nodes were made with this commitment in mind (Dejax and Bookbinder, 1991). The shipper paid a premium to receive this transportation service; a rebate was paid by the railway in the relatively rare cases that this commitment could not be honored.

In general, the application by a railroad of a shipment consolidation (SCL) program needs to be tempered in the case of JIT. SCL does typically lead to economies of scale in the cost per unit weight that is moved, or in cost per load, or cost per unit time. However, with those economies come some variability in the lengths of a consolidation cycle. The railway will receive a premium for a JIT shipment whose time commitment is kept; that premium is the benefit to the railway to roll back its SCL program, segregating JIT loads from the other (ordinary) shipments tendered to it.

When rail freight movements are part of a JIT system, the road in question must give priority to regularly-scheduled JIT trains. Those trains should, of course, bypass the classification yards. Within one of the railway intermodal plans, the pickup of goods from shippers, and delivery to consignees, should be prearranged, ideally in trucks owned by the railway or a carrier affiliate.

Outbound Versus Inbound Transportation

The above discussion on transportation aspects of JIT systems considered solely the inbound side, that is, daily deliveries of raw materials and components to, say, an assembly plant. Let us now discuss *outbound* transportation from a JIT manufacturing facility. That is, a distribution system, from that plant to warehouses and retailers. Can that outbound system also operate on a JIT basis? Is there any advantage in doing so?

Bookbinder and Locke (1986) conducted a simulation study of JIT distribution, whose main results are now summarized here. Two systems were compared, each consisting of a factory that served two warehouses that then shipped to three retailers apiece. Both systems operated on a JIT basis, with daily deliveries to restore the inventory at a given echelon to its base-stock value. In System 1, inventory was held at each echelon. System 2, however, carried inventories only at the factory and retailers; each warehouse in that system played the role of a cross-dock.

At each stock-keeping location, the performance metrics concerned the service measures P_1 and P_2 . These are respectively the probability of no stockout during a replenishment lead time (the cycle service level), and the fraction of demand that is immediately satisfied from on-hand stock (the fill rate). By the careful use of a simulation methodology, Systems 1 and 2 were compared in a statistical sense.

Naturally, System 2 had a greater annual inventory turnover, since it carried stock at one fewer echelon. This resulted in System 2 having a statistically significant lower cost than System 1. Both systems, however, could achieve high retailer service levels (whether P_1 or P_2), with only modest safety stock at the factory. A main conclusion is that, although the flows outbound from a JIT production system need not function in a JIT manner themselves, there is no reduction in retailer service levels in doing so. And there is a demonstrable cost advantage in operating that way. Of course, certain technical changes in the loading and unloading of JIT vehicles would enhance JIT's approach to distribution, rendering it smoother and more feasible.

Inventory-Transportation Tradeoffs in JIT

It is important to be aware that, in the analysis and operation of a JIT system, inventory decisions have tended to take priority over transportation. Two factors may explain this. First, we have already mentioned the geography of Japan, in which distances are “small,” and Tier-1 suppliers are located near the assembly plant. Short travel distances, especially when combined with the Toyota mixed-load system, make transportation costs seem less of an issue. Second, JIT became especially prominent in North America in the early 1980s. Interest rates went through the roof, generally greater than 20% annually. In that environment, it was no wonder that inventory was considered not an asset, but “waste.” Firms worried more about bank loans to finance their cycle stocks, than about the extra vehicles on the road, and the additional miles they were driving.

Today, companies and individuals are concerned about social responsibility, and the degree to which a supply chain is “green” (environmentally friendly). At present, interest rates are so low that, whether from a societal or a cost point of view, inventories look much better than transportation. Can this conclusion be supported by a modeling approach?

One way to study the issue of inventory-transportation tradeoffs is to consider the multiple-objective problem, $[\text{Min (Inventory Costs)}, \text{Min (Transportation Costs)}] = [\text{Min } Z_1, \text{Min } Z_2]$. Each objective could first be considered as a constraint on the other: $\text{Min } Z_1$, subject to the constraint $Z_2 \geq Z_{2 \min}$. The optimal values of the inventory costs would be found for good, feasible values of the transportation costs. Then, a reverse series of problems would be solved: $\text{Min } Z_2$, subject to $Z_1 \geq Z_{1 \min}$.

Here is why it may be necessary to consider a sequential method, similar to the preceding, but with a more detailed approach to the constraint on the secondary objective. In a firm with *decentralized* inventory and transportation departments, where we wish to minimize the total logistics costs, a simultaneous model to optimize those functions may not be implementable.

Bookbinder and Coulombe (2013) develop a sequential procedure, one that is applicable to any transportation-inventory model with separable costs. In that technique, there is first a full optimization by the “primary” department of its costs. The “secondary” department next minimizes its own costs, subject to a bounded allowable increase in the costs of the primary department. It was found that, when each department’s relative deviation from its optimal cost is equal, a near-optimal solution is attained for overall total logistics cost.

Continuous Improvement and JIT Logistics

The focus of JIT is on business processes, rather than on products and services. Therefore, tactics for Continuous Improvement, or CI (*kaizen* in Japanese), proven to be successful in manufacturing settings, can also be applied in business logistics (Coimbra, 2013). Let us call the application of JIT philosophy in the logistics context, *JIT-L*.

Increased competition, globalization, and a shift from manufacturing to more service-oriented industries especially in developed countries, necessitate finding innovative ways to cut costs by eliminating waste. Continuous improvement helps reveal nonvalue adding activities. One eliminates them if possible, or simply enhances the pertinent business processes, so that system performance metrics are constantly on an upswing.

How can we apply CI to *JIT-L*? We should start by understanding the core elements of logistics. These include customer service, order processing, inventory management, and of course, transportation. Customer service is a function of the quality with which the flow of goods is managed. Customer value-creation implies that a *JIT-L* system should observe the seven R’s of distribution: deliver in the *right* conditions the *right* product at the *right* quantity and *right* (least) costs to the *right* customer at the *right* place, and at the *right* time. With a closer look at customer needs, and after surveying feedback on customer satisfaction, the logistics service can benefit from informed decisions. For instance, even if a company’s on-time delivery performance is impressive, it may be nullified in terms of customer satisfaction and costs, if the delivered item is the *wrong* one to begin with.

Paperwork is still the “gum in the gear” in transport operations, especially in border-crossing operations. Such paperwork may cause delay in information flow between sellers and buyers, resulting in “waste” due to administrative procedures. For example, to reduce such unnecessary delays and therefore enhance on-time delivery performance, a 3PL could adopt electronic data interchange, Internet-based solutions, and Logistics 4.0 systems which are based on the Internet of things (IoT). The IoT refers to the data-communication technology and sensors that are built into physical objects, enabling those objects to be tracked and controlled over the internet. The IoT could provide valuable information and business insights, enabling logistics managers to reduce costs and upgrade service. For instance, an IoT-compatible forklift could alert a warehouse manager to potential mechanical or safety issues prior to their occurrence. The forklift could also provide increased visibility about inventory in the warehouse. If the related infrastructure were in place, it is apparent that the enabling technology may align with CI, and encourage progress in *JIT-L*. Venkatadri et al. (2016) investigate the use of the IoT and SCL in the context of integrated inventory-transportation management.

Due to its “pull system” nature, JIT systems are reactive; that is, production is initiated once the actual orders are received. Naturally, this reduces the amount of inventory (raw materials, WIP, or finished goods) to the level that is just needed. However, from a cost perspective, a JIT-embracing manufacturer may be still be lured to carry large volumes of stock. This could be due to price discounts offered by upstream suppliers, or simply to mitigate the risks of possible shortages. In such situations, the focal JIT firm can apply CI to magnify the precision in estimating demand, and also to forge stronger relations with its vendors. The “pricing” of inbound materials thus becomes a coordination issue, rather than a mechanism for the supplier to sell in volumes. Alternatively, as mentioned earlier, a JIT warehouse, especially in North America, may be a good “win-win” solution for both seller and buyer.

JIT-L should also strive to eliminate wasted capacity in transportation. For instance, because of inefficient truck scheduling, many vehicles could be waiting at the loading and unloading docks. Such unnecessary truck delays imply underutilized transport capacity, incurring obvious cost. Using advanced Logistics Information Systems (LIS; Bookbinder and Diltz, 1989), both the inbound and the outbound flows from the focal JIT company might be synchronized, similar to cross-docking. Bookbinder and Barkhouse (1993) propose such a system in terms of SCL, and discusses implications for JIT.

It should be noticed as well that, because of fewer handling activities (loading, unloading, sorting), the quality (condition) of goods delivered via SCL is better than that of separate shipments. This perspective could argue in favor of the use of SCL in a JIT setting: A pillar of JIT is quality; zero defects, until products reach the hands of the consignee.

Implementing JIT is not straightforward; many parts of the system must move together swimmingly. That necessitates the creation of dependable, long-lasting relations with supply chain partners. Therefore, JIT companies would tend to work with fewer

carriers, and more closely. As opposed to traditional firms, the JIT philosophy would dictate refraining from adversarial buyer-supplier negotiations. Rather, the focus should be on information sharing and developing “trust” between partners, to jointly make the most of mutual benefits in moving toward achieving the seven R’s. That is, CI needs to be embraced as a part of the organizational *culture* by all parties involved in the supply chain.

JIT-L and Sustainability

Daugherty and Spencer (1990) study the application of JIT principles to logistics and transportation. Let us now discuss the additional concept, “sustainability.” JIT logistics, whose aim is zero logistical waste, positively impacts three dimensions of sustainability: JIT-L yields economic benefits, because of reduced waste in motion and time (economic dimension). This is also reflected in diminished emissions of carbon and water, due to the elimination of the previously-mentioned unnecessary move-store activities (environmental dimension). Finally, JIT-L, by its management principles, requires close and prolonged relationships with relevant partners in the supply chain. JIT-L thus needs to be socially responsible in terms of how the involved organizations treat their employees, provide safe working environments, and support the communities involved (societal dimension).

The implementation of JIT-L can especially affect *omni-channel* marketing strategies. The consumer has the option of buying online, where, unlike in bricks-and-mortar retail stores, s/he is not able to physically examine the product and make a better informed purchase decision. That is, a holistic JIT-L system should consider the ramifications of product returns. The reverse flow of products is a significant issue, both in terms of cost and environmental impact. We need to factor in the impact of such reverse flows, with the recapture of value of the returned products.

Current models of closed-loop supply chains do not often embrace JIT-L. This might be because returned products may further benefit from SCL through designated collection depots. Although JIT “promises” zero-defects, we note that a large part of the items returned may not be due to product malfunction, but rather due to the consumers’ abuse of returns policies or simply regretting an impulsive purchase. (Ülkü et al., 2013; Ülkü and Gürler, 2018) analyze such consumer-return behavior and its impact on logistics performance indicators. Considering the fact that intense competition in the retailing world forces some companies to allow customers to return items for no reason at all, there is merit in exploring the impact of reverse logistics on JIT-L and its environmental footprint.

The JIT viewpoint implies complete elimination of waste from a productive system, whether it be a manufacturing, service, or logistics system. For CI to be achieved, the people in the system must contribute to it. Since, JIT requires constant feedback for improvement opportunities, and because employees are at the front line in observing what is going on in the business processes, it is of the utmost importance that the JIT-embracing company be respectful to its employees. Those personnel must be encouraged, so they are motivated to propose process improvement changes that may eliminate waste. Therefore, workers are essential assets of the JIT system, in which they are given additional authority to make decisions compared to more traditional (non-JIT) systems.

Concluding Remarks

The JIT approach originated in manufacturing, but is no longer confined there. As emphasized in this chapter, logistics and other business processes can benefit from the elimination of waste. The term “waste” should be understood as comprising any non-value added activities along the supply chain, such as unnecessary waiting times in loading and unloading, or underutilized conveyance capacity.

The emphasis of JIT-L is on rendering more “lean,” the move-store activities within and between supply chain partners. JIT-L obliges an integrated approach and dependable relationships between those partners.

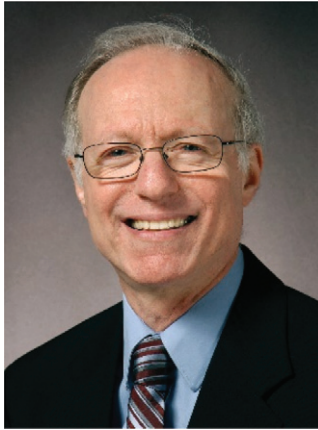
Unlike their Japanese counterparts, North American firms are dispersed over larger areas, impeding the small distances that enable JIT success. Use of a “JIT warehouse” permits the JIT approach to function when suppliers to an assembly plant are not nearby.

JIT-L requires a real “cultural shift” in management to operationalize the concept of CI. But with that shift, JIT-L may enable new practices, especially with the advent of technology in the logistics sector. Similar to successful tactics such as cross-docking, JIT-L may show the way in novel areas such as the IoT (the “physical internet”; Venkatadri et al., 2016).

The JIT philosophy aids a renewed approach to supply chain sustainability. Logistical activities of course have an economic dimension (e.g., reduced waste in motion and time), but also environmental and social impacts of respective supply chains. Examples of the latter two effects include diminished carbon emissions by eliminating unnecessary move-store activities, and social responsibilities to the organization’s employees, respectively.

We remark in closing that JIT-L can easily be the subject of further research. Areas of interest include the reverse logistics of returned products, whose forward flows were managed as a JIT system, and the contractual mechanisms between partners in a JIT-operated supply chain.

Biographies



James H. Bookbinder is a Professor of Management Sciences at the University of Waterloo. For 20-plus years, he has worked with manufacturing firms, third parties, and transportation carriers on the modeling and analysis of logistics strategies and operations. He is the Director of WATMIMS, the University of Waterloo's Center for Research in Logistics and Manufacturing. He is a past-president of the Canadian Operational Research Society and past-chair of the transportation science and logistics section of INFORMS. Jim holds an MBA from the University of Toronto and a PhD from the University of California, San Diego.

His current research involves optimization of joint inventory-transportation decisions, and models for supply chains in the post-NAFTA, post-TPP (Trans Pacific Partnership) eras. For 10 years, Dr. Bookbinder was an Associate Editor (AE) of *Operations Research* (OR Practice), and still serves as an AE for the *Journal of Business Logistics* and for *Naval Research Logistics*. Jim is co-author of a 2009 INFORMS tutorial chapter on "Global Supply Chains." He is the editor of *Handbook of Global Logistics: Transportation in International Supply Chains* (Springer, 2013).



M. Ali Ülkü is a Professor and the Director of the Center for Research in Sustainable Supply Chain Analytics, in the Rowe School of Business at Dalhousie University, Canada. He received his PhD in Management Sciences from the University of Waterloo, MSc in Operations Research from Çukurova University, and BSc in Industrial Engineering from Bilkent University. Prior to his academic career, he worked as a productivity consultant at Anadolu Efes, the largest multinational brewery company in Turkey. Dr. Ülkü's writings include the theoretical modeling of sustainable supply chain and logistics systems, operations-marketing interface, and mathematical modeling of consumer behavior and societal problems. He is an avid supporter interdisciplinary research at the confluence of theory and practice. He published in such journals as *Annals of Operations Research*, *European Journal of Operational Research*, *International Journal of Production Economics*, *Journal of Business Research*, *Journal of Cleaner Production*, and *Service Science*. He served as the Program Chair for the 2018 Conference of the Canadian Operational Research Society.

References

- Bookbinder, J.H., Barkhouse, C.I., 1993. An information system for simultaneous consolidation of inbound and outbound shipments. *Transp. J.* 32 (4), 5–20.
- Bookbinder, J.H., Coulombe, M., 2013. Sequential decision-making schemes in inventory and transportation environments. *Proceedings of European Operations Management Assoc.*, Dublin, pp. 1–10.
- Bookbinder, J.H., Dilts, D.M., 1989. Logistics information systems in a just-in-time environment. *J. Bus. Logist.* 10 (1), 50–67.
- Bookbinder, J.H., Locke, T.D., 1986. Simulation analysis of just-in-time distribution. *Int. J. Phys. Distrib. Mater. Manag.* 16 (7), 31–45.
- Coimbra, E.A., 2013. *Kaizen in Logistics and Supply Chains*. McGraw-Hill Education, New York.
- Daugherty, P.J., Spencer, M.S., 1990. Just-in-time concepts: applicability to logistics/transportation. *Int. J. Phys. Distrib. Logist. Manag.* 20 (7), 12–18.
- Dejax, P., Bookbinder, J.H., 1991. Goods transportation by the French national railway (SNCF): the measurement and marketing of reliability. *Transp. Res. Part A: General* 25 (4), 219–225.
- Higginson, J.K., Bookbinder, J.H., 1990. Implications of just-in-time production on rail freight systems. *Transp. J.* 29 (3), 29–35.
- Rao, K., Grenoble, W., Young, R., 1991. Traffic congestion and JIT. *J. Bus. Logist.* 12 (1), 105–121.
- Shingo, S., 1988. *Non-stock Production: The Shingo System of Continuous Improvement*. CRC Press, Florida, United States.
- Ülkü, M.A., 2009. Comparison of typical shipment consolidation programs: structural results. *Manag. Sci. Eng.* 3 (4), 27–33.
- Ülkü, M.A., 2012. Dare to care: shipment consolidation reduces not only costs, but also environmental damage. *Int. J. Prod. Econ.* 139 (2), 438–446.
- Ülkü, M.A., Dailey, L.C., Yayla-Küllü, H.M., 2013. Serving fraudulent consumers? the impact of return policies on retailer's profitability. *Serv. Sci.* 5 (4), 296–309.
- Ülkü, M.A., Gürler, Ü., 2018. The impact of abusing return policies: a newsvendor model with opportunistic consumers. *Int. J. Prod. Econ.* 203, 124–133.
- Venkatadri, U., Krishna, K.S., Ülkü, M.A., 2016. On physical internet logistics: modeling the impact of consolidation on transportation and inventory costs. *IEEE Trans. Autom. Sci. Eng.* 13 (4), 1517–1527.

Supply Chain Finance

Erik Hofmann, Institute of Supply Chain Management, University of St. Gallen, St. Gallen, Switzerland

© 2021 Elsevier Ltd. All rights reserved.

Introduction	113
Historical Background of SCF	113
Relevance of SCF	114
Definition of SCF	114
Measuring the Performance of SCF	115
Value Proposition of SCF	115
SCF Solutions	116
SCF Stakeholders	117
Concluding Remarks	117
Biography	118
References	118

Introduction

In the past, and even today, many companies had, and still have, high levels of capital bound up in the form of working capital in accounts payable, accounts receivable, or inventories. This capital cannot be used in a profit-yielding way. Thus, in the middle of the 20th century, many companies, supported by academic research, started to optimize their working capital positions on their balance sheets. The aim of the optimization focused on the company's benefit, and did not take into account the effects on other related parties in the supply chain. Moreover, at that time the main focus of companies was the flow of goods in the supply chain (thus, the physical supply chain), and not optimization of the flow of cash and affiliated information (thus, the financial supply chain). In the 1990s, however, supply chains became more and more globalized. With the development of direct sourcing, the complexity of supply chains increased, while the globalization of the economy lengthened the supply chains, by increasing tremendously the number of suppliers, and therefore the number of physical and financial transactions.

Since the global economic and financial crisis (starting in 2008/09), companies have had to rethink their supply chain optimization efforts, especially for financials. This crisis had two major negative effects on the world economy. First, the global demand for every kind of good and service collapsed, severely reducing the order inflow and backlog. Many companies have had to introduce short-time working or make staff redundant, or even cease their business activities. Second, liquidity was no longer available on the market, as financial institutions tightened their criteria for business financing, and increased the costs of financing. Since this credit crunch, the distance between investment grade and non-investment-grade financing rates has widened even more. Companies, which did not have sufficient liquidity, had to negotiate unfavorable conditions, or they went bankrupt. Both effects of the crisis destabilized supply chains as, for example, the timely supply of necessary product inputs was no longer guaranteed. Finally, this crisis clearly underscored the fact that a supply chain is only as financially strong as its weakest link.

In this context, the concept of supply chain finance (SCF) has been seen by companies as a suitable solution to reduce counterparty risk and sustainably stabilize the different links in their value networks, to prevent the disruption of whole production lines resulting from financial problems of one important party involved, and to optimize the total costs within the supply chain. SCF can be understood as a business practice whose principles were firstly examined by Koch (1948), who discusses the *economic aspects of inventory and receivables financing*, or Petersen and Rajan (1994), who emphasize *the benefits of lending relationships* from a small business perspective. SCF, in its contemporary manifestation, has come out of hibernation. In the meantime, SCF has become an established business practice, with an enormous range of techniques and solutions. Banks, fintechs, IT providers, advisors, logistics service providers, and others are forming an ecosystem for SCF.

This paper aims to provide an overview of the background, the relevance, and the definition of SCF. This paper also introduces the cash-to-cash (C2C) cycle as a key performance measurement of SCF. After the value proposition of SCF is elaborated, the solutions of SCF (especially the reverse factoring approach) are presented. Before the paper ends with concluding remarks, relevant internal and external stakeholders within the SCF ecosystem are discussed.

Historical Background of SCF

From a historical viewpoint, the industry of trade finance has been around for hundreds of years, with sample letters of credit from the 16th century on display at the museum of Banca Monte dei Paschi Di Siena in Italy (Malaket, 2014). Trade finance is a commercial discipline that divides into two opposing facets: on one side are letters of credit (L/C), standby credits, documentary collection, and bid bonds, all representing the traditional subdivision of trade finance. Typically, such traditional products of trade

finance address single transactions, and establish bilateral financing arrangements between a buyer and a seller involving other parties only if, and when, directly concerned. On the other side, the modern form of trade finance promotes financing programs occasioning the emergence of entire ecosystems of relationships. These contemporary interorganizational financing solutions signify an evolutionary step as they disassociate themselves from atomistic product-transaction thinking to a holistic solution-relationships rationale in supply chains. Thus, the modern method of trade finance formed the basis of SCF (Hofmann, 2005; Hofmann and Belin, 2011). In the not too distant past, businesses exchanging goods and services conceived of one another as autonomous entities, rather than constituents of the self-same supply chain. In need of liquidity and working capital, interdependent companies not only sought financing based on their own terms but also adopted ill-considered measures that got their trading partners in hot water (Dyckman, 2011). Many large US and European corporate buyers striving to optimize working capital, for instance, extended payment terms to the point that small- and medium-sized suppliers situated in states with less developed capital markets faced a sharp upsurge in production cost and financial risk (de Meijer and de Bruijn, 2013). Although lack of liquidity and financing is a menace ever-present with smaller companies anyway, multilayered developments induced a refrain from egocentrism and a turn toward cooperation in what may be considered a paradigmatic shift resulting in SCF, among others (Malaket, 2014).

The presumably most prominent calamity propelling the growth of SCF in recent years was the 2008/09 financial crisis in whose aftermath the implementation of Basel III impelled banks to hold sufficient high-quality liquid assets and intensified capital requirements, to name but a few amendments. While the Basel Committee on Banking Supervision's regulatory framework discourages banks from financing (typically) less creditworthy small- and medium-sized enterprises (a withdrawal of liquidity that halted the operation of quite a few small suppliers, and disrupted the supply chains of larger corporations in much the same way as the ill-advised extension of payment terms), and emphasizes SCF as a last resort, other developments deserve acknowledgment, too. The formation of SCF harks back to globalization and an increase in cross-border commerce, the contraction of traditional trade finance's letters of credit in favor of open account (O/A) trading, and enhanced technology facilitating automation and supply chain transparency. In point of fact, exchanging goods and services globally stretched supply chains, considerably incentivizing businesses to reduce capital exposure to supply chain risks by optimizing working capital. Although open account trading is one alternative to optimizing importers' working capital with their cost and cash flows under control, open account trading may jeopardize the health of exporters if employed without proper risk management measures. In light of this, the first movement by international banks toward SCF assuredly was not altruistic, but aspired to rejuvenate declining trade finance revenues. Although the industry of SCF continues to experience changes, with, for example, nonbank players entering the market, a further breakdown of SCF requires to emphasize a common understanding of its relevance at the minimum.

Relevance of SCF

Often, non-investment-grade companies and small-to-medium enterprises find it difficult to finance their working capital requirements, as buying firms have tightened their supply chains. This is particularly true for smaller suppliers or customers (buyers) in emerging economies. The result is that there is a significant credit arbitrage between large firms in established markets and their supply chain partners in emerging economies.

Referring to an interorganizational scale, single supply chain members often act against the supply chain surplus when achieving the same goal of improvement (Hofmann and Kotzab, 2010). Financial factors, such as credit risk and capital costs, are often transferred to the upstream supply chain (to suppliers), or, in fewer cases, to the downstream supply chain (to customers). Although extended payment terms on the supply side (or shorter payment terms on the demand side) constitute a lower risk to the buyer (vendor), self-centered payment terms can destabilize the entire supply chain, as a result of a higher-risk supplier (customer) base, and their restricted access to short-term financing (Seifert et al., 2013).

SCF aims to minimize the cost of tied-up capital, including its related risks, while maximizing the gains of cash received across the participating supply chain members. From an interorganizational perspective, firms with lower refinancing rates and greater financial strength should carry more working capital compared to firms with higher financing costs. Grosse-Ruyken et al. (2011) agreed that working capital management (WCM) decisions should consider up- and downstream partners in a collaborative way, and that an adequate level of working capital depends on the business model, specific design configurations, as well as risk aspects within the supply chain.

Definition of SCF

The term "supply chain finance" often refers to a branch of WCM, and is connected to financial supply chain management (FSCM). FSCM comprises the financial processes (e.g., customs clearance or invoicing) and payments within a supply chain, with the purpose of introducing measures to optimize planning, managing, and controlling cash flows, to facilitate efficient material flows in interorganizational settings. There are different understandings of SCF (Zhao and Huchzermeier, 2018):

- *Supply chain finance in the narrow sense:* Within this definition, SCF is understood as a financing solution that enables a customer to prefinance payables to suppliers. This form of supplier financing—also known as buyer-driven payables solutions—is the most common SCF practice. Such a common used instrument is "reverse factoring."

- *Supply chain finance in a wider sense:* However, SCF encompasses much more. The concept covers the financial flows of the entire supply chain and extends far beyond supplier financing. It aims to optimize cross-company working capital and the integration of financial processes between suppliers, customers, and external service providers (e.g., banks, fintechs, and logistics service providers). Thus, SCF can be understood as an approach to creating value by planning, monitoring, and steering the flow of financial resources for two or more organizations in a supply chain, including external service providers.
- *Supply chain finance in the broadest sense:* In the broadest definition, SCF covers not only the financing of all the transactions being done in supply chains (current assets), but also the financing of all resources and capacities required for operating activities (fixed assets). In particular, this includes moveable assets (e.g., vehicles), equipment (e.g., tools and machinery), and real estate (e.g., warehouses). Furthermore, this perspective goes beyond the point-in-time view, focusing, on the contrary, on the whole lifetime of the supply chain-relevant infrastructure.

In support of standardization efforts by the European Banking Association's Supply Chain Working Group, SCF refers to (EBA, 2014, p. 44): "The use of financial instruments, practices and technologies to optimise the management of the working capital and liquidity tied up in supply chain processes for collaborating business partners. SCF is largely 'event-driven'. Each intervention (finance, risk mitigation or payment) in the financial supply chain is driven by an event in the physical supply chain. The development of advanced technologies to track and control events in the physical supply chain creates opportunities to automate the initiation of SCF interventions."

Measuring the Performance of SCF

On the face of it, net working capital is defined as current assets minus current liabilities. In somewhat greater detail, working capital is calculated as working capital = accounts receivable + inventory – accounts payable + cash (EBA, 2014). Although holding more cash and inventory (i.e., raw material, work in progress, and finished goods), and accounts receivable (i.e., the money customers owe a business) increases working capital, adding accounts payable (i.e., the payments due to suppliers in exchange for their goods and services) decreases working capital. Although research suggests that a low level of positive working capital performs best, each company has to find the right balance between positive and negative working capital. That is, extraordinarily broadly speaking, positive working capital comes with less risk, but greater cost of capital, as expensive liquid assets are relatively abundant; negative working capital is accompanied by greater risk, but less cost of capital, as expensive liquid assets are relatively scarce. Nonetheless, optimizing working capital necessitates ways and means to draw conclusions pertaining to the commercial efficiency and effectiveness of WCM. Appealing to the time-based ratios of days sales outstanding (DSO), days inventory held (DIH), and days payable outstanding (DPO) together is one such approach to evaluating WCM. The C2C cycle, measured as $C2C\ cycle = DSO + DIH - DPO$, links the three time-based ratios, with the intent of demonstrating the period of time between disbursing cash for purchasing inventory and collecting cash for selling inventory. Considered in isolation (Hofmann and Belin, 2011):

- DSO portrays the average number of days required to collect outstanding receivables. Accordingly, DSO illustrates how fast customers are invoiced and payments received.
- DIH portrays the average number of days, raw materials, work in progress, and finished goods are held until sold. Accordingly, DIH illustrates how long funds remain tied up in inventory.
- DPO portrays the average number of days required to pay suppliers. Accordingly, DPO illustrates a trade-off between sufficient working capital and sustainable business relations.

As a local improvement can have negative effects on the C2C cycle of the previous and/or following supply chain partners, a holistic approach must be taken, creating a win-win situation for all involved parties. This requires that companies share their financial information concerning the weighted average costs of capital (WACC) and inventory carrying costs (ICC), to leverage financial strengths throughout the whole supply chain. The C2C cycle would be optimized so that it is inversely related to the companies' WACCs and ICCs, as the firm with lower costs of capital should carry more C2C cycle days. Degradation of one company's C2C numbers must be accepted, as it brings overall gains for the whole supply chain. These gains, for example, cost savings, have to be spread equitably, thus also compensating companies that reduce their profits. Further benefits include improved overall cash flows, and, thus, increased profitability and higher company value.

Value Proposition of SCF

Starting from the premise that SCF is set to optimize working capital, the following insights may eventually be pinpointed. First, the C2C cycle is said to be negatively correlated to the enterprise value. Changes in enterprise value can be described with the return on net assets (RONA). In line with this train of thought, optimizing working capital, too, adds economic value: optimizing the working capital intends to shorten the C2C cycle, which again increases enterprise value, an added value measured by RONA. Put differently, a chain of causation links working capital optimization, C2C cycle, enterprise value, and RONA. As is well known, SCF and its solutions are envisioned to optimize working capital, too.

Consequently, the value proposition of SCF incorporates at least two kinds of benefits for the parties involved. Qualitative benefits include an increase in payment transparency, an improved flow of information, and enhanced compliance and

relationships between buyers and suppliers (Templar et al., 2016). Quantitative benefits incorporate improvements in working capital due to reduced C2C cycle times, a reduction of operational and administrative costs, and increased liquidity (Caniato et al., 2016). However, from an interorganizational perspective, the shifts in the payment terms or program structure may result in a different allocation of benefits between the actors (Liebl et al., 2016). As an example, an improvement in the DPO from the buyer's perspective does not necessarily lead to a beneficial situation for the supplier, or the additional administrative effort needed for an SCF program may lead to resistance from within a company. Therefore, it is crucial to consider the actors' interests and motivations, to support the win-win concept of SCF.

SCF Solutions

SCF is different from traditional trade finance in that it elicits ecosystems of relationships. For such ecosystems to materialize, two elements are needed: an SCF solution and stakeholders of SCF. In point of fact, this combination motivates the market players to think in terms of a larger, generic ecosystem in which all sorts of stakeholders interrelate, irrespective of the specific solution.

The following SCF solutions are available:

- Accounts payable or buyer-centric solutions, comprising approved payables finance or reverse factoring, dynamic discounting or purchase order financing. Of the earlier, this category includes payables-based credit arbitrage or supplier finance and import financing.
- Accounts receivable or supplier-centric solutions, comprising receivables purchase, invoice discounting, factoring, and forfaiting. Of the earlier, this category includes key accounts receivable sales, sales finance, export financing, and the matching accounts receivable solutions.
- Inventory-centric or pre-shipment-centric solutions, comprising off- and on-balance inventory financing. Off-balance inventory financing refers to the financing of inventories where operational goods logistics and ownership are transferred to an external (logistics) service provider. In contrast, within the on-balance solution, inventories remain on the company's balance sheet, and serve as a guarantee for a credit agreement (the asset-based lending approach).
- Other SCF-related solutions, comprising asset financing and asset-based lending, longer-term export and project finance, bank payment obligations, hedging and payment instruments, and products of traditional trade finance.

Example: Approved payables financing is a branch of SCF, and focuses on the accounts payable aspect of WCM (Wuttke et al., 2013). From this branch, a commonly used program is called "reverse factoring." Reverse factoring (RF) is a buyer-driven payables solution (Klapper, 2005), and is commonly defined as an arrangement of a buyer and its supplier collaborating with a financier or bank to facilitate the optimization of financial flows along their respective supply chains (Liebl et al., 2016). RF first emerged as a financing solution for larger corporations, with the goal of diminishing the financing issues of their suppliers (Lekkakos and Serrano, 2016). Unlike factoring, RF is not the supplier asking the bank for anticipated credit, but the buyer initiating a reverse factoring program by offering the accounts receivable of its suppliers to a bank, once the product and the invoice have been sent by the supplier and received by the buyer (Zhao and Huchzermeier, 2018).

Fig. 1 illustrates the concept of reverse factoring, and the main parties involved in the process. The following explanation of the concept and its components is based on depictions in literature (Hofmann and Belin, 2011; Templar et al., 2016; Hofmann et al., 2018). The four parties are the affiliated suppliers or vendors, the initiating buyer, the financier, and the SCF platform

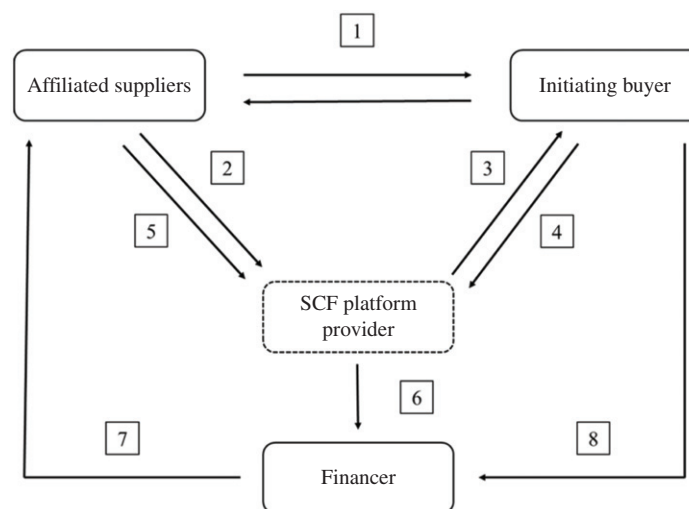


Figure 1 The concept of reverse factoring.

provider serving as the intermediary in the process, which can either be the financier directly or an independent service provider. Further, from the buyer's perspective, it is possible to interact with multiple suppliers and financiers at once, to establish a fully efficient reverse factoring cycle. The whole process starts with the buyer placing the order and the supplier sending the goods (1). Then, the supplier submits the invoice for the transaction to the buyer through the SCF intermediary, or any other form of invoicing favored by the buyer (2). This allows the buyer to receive the invoice in their enterprise resource planning system (3). Once the buyer approves the payable, the approval is conveyed through the SCF platform to the supplier (4). At this point, the supplier has the choice of waiting for the payment to reach its maturity or requesting an early payment executed by the financier (5). If the supplier chooses the financing method, the financier approves the early payment request through the platform (6), pays the receivable amount of the invoice (7), and earns the premium by withholding a small portion of the invoice payment as a discount (8). In the final step, when the invoice reaches maturity, the buyer pays the full amount of the invoice to the financier (9).

It becomes evident when examining the reverse factoring concept that the complexity and number of interactions between the actors may lead to complications and issues (Liebl et al., 2016). Therefore, it is crucial to better understand which individuals and groups within these parties are involved in the process, and what role they play in an SCF program.

SCF Stakeholders

A successful implementation of SCF requires a profound collaboration between all relevant parties. The working relationship has to evolve into a truly intensive partnership, in which all parties pay careful attention to the impact of their decisions on others in the supply chain. In view of the cross-functional nature of SCF, close coordination between individual internal and external stakeholders is required.

Internal stakeholders are responsible for successfully implementing SCF solutions. Be it sales dealing with DSO, treasury managing cash, procurement addressing DPO, or supply chain managers, together with manufacturing, planning inventory, these and other stakeholders internal to businesses are part and parcel of SCF ecosystems. With purposes and objectives contrasting in large part, the stakeholders within a company effectively constitute risk sources, just as organizations as a whole do. Staff from procurement and logistics must work closely with colleagues from the finance department to ensure a united front in dialogue with suppliers. The traditional job profiles and skill sets of the individual stakeholders must be improved in a targeted way, using SCF expertise, and a common understanding of SCF must be established.

External stakeholders have to be addressed, too: corporate customers, banks, nonbank financial providers, IT vendors, business-to-business networks and e-invoicing service providers, marketplaces and hubs, consultancies, logistics services providers (LSPs), industry associations and governments, along with the public sector. Apparently, this is a rich landscape not yet accounted for; for instance, a supply chain's upstream suppliers and downstream buyers in international contexts are often referred to as exporters and importers, of first, second, and higher tiers. To not lose a grip on feasibility, however, it is imperative to pare the external stakeholders down to the following vital ones:

- The focal company is usually an international enterprise sufficient in sales and trade volume to cater to the cost and sophistication of SCF. As the availability of buyer- and supplier-centric solutions indicate, the focal company is interchangeably a buyer or a supplier driving the adoption of SCF solutions in its supply chain.
- The trading partners, that is, first-tier suppliers and buyers, exchange goods and services with the focal company. Usually smaller than the focal company, the trading partners benefit from the adoption of SCF in ways that are subject to the solution in question.
- The service providers, that is, financial services providers (FSPs), LSPs, and insurance companies (ICs). FSPs are banks and nonbank providers of financial services. Banks, both international and national, facilitate the adoption of SCF on a stand-alone basis or in partnership with other banks and nonbank FSPs. Nonbank FSPs are financial arms of corporations, among others. LSPs offer a range of services, giving them insight into the physical supply chain. ICs assume the risk of incurring damage from certain negative events, for an insurance premium.

In practice, it is not always easy to convince internal departments of an SCF solution, while also taking account of the concerns of external stakeholders, such as suppliers and auditors. Auditors, in particular, should be involved at an early stage (i.e., when the SCF program is being structured), to obtain their evaluation of the compliance on accounting treatments and of the impact on financial reporting. The persuasive efforts are worthwhile, as customers who proactively offer their suppliers earlier payments through an SCF program enjoy clear competitive advantages.

Concluding Remarks

With the current advances in technological developments (e.g., automation, platforms, blockchain technology solutions [e.g., tokenization], and machine learning), the supply chain has become a cheap source of cash in many organizations (Hofmann et al., 2018). SCF enables managers to improve the company's balance sheet and income statement. There is a symbiotic effect between the combination of supply chain management and finance that makes the whole greater than the sum of the parts.

SCF can bring structure and discipline to the financial portion of the supply chain, which can, in turn, improve the physical supply chain. SCF can reduce the variability of payments in the supply chain and, therefore, can reduce the need for additional cash to alleviate uncertainty. In general, SCF has the strong potential to be a multiple-win for the ecosystem partners, with numerous additional benefits, such as providing significant liquidity, enforcing discipline in the approval of invoices, reducing risks, and taking the variability out of the timing of payments.

In a more comprehensive view, SCF addresses all the assets to be financed within the supply chain. Beyond the (net) working capital, SCF should contain all other necessary material resources (e.g., warehouses, trucks, and transshipment facilities), enabling infrastructures (e.g., information and communication systems, platforms, and software) or even human resources (e.g., workforce, truck drivers, IT specialists, and analysts).

Biography

Prof. Erik Hofmann (Dr. rer. pol., University of Technology, Darmstadt, Germany) is director of the Institute of Supply Chain Management and a senior lecturer at the University of St. Gallen, Switzerland. He is also head of the Supply Chain Finance Lab (SCF Lab). His primary research focuses on supply management and the intersections of operations management and finance issues. He has published in several operations management journals. He is coauthor of several award-winning books, including *Supply Chain Finance Solutions*, *Ways Out of the Working Capital Trap*, *Performance Measurement and Incentive Systems in Purchasing* (all Gabler Springer), and *Financing the End-to-End Supply Chain* (Kogan Page).

References

- Caniato, F., Gelsomino, L.M., Perego, A., Ronchi, S., 2016. Does finance solve the supply chain financing problem? *Suppl. Chain Manag. Int. J.* 21, 534–549.
- de Meijer, C.R., de Bruijn, M., 2013. Cross-border supply-chain finance: an important offering in transaction banking. *J. Payments Strat. Syst.* 7 (4), 304–318.
- Dyckman, B., 2011. Supply chain finance: risk mitigation and revenue growth. *J. Corp. Treas. Manag.* 4 (2), 168–173.
- EBA, 2014. Supply chain finance: EBA European market guide. European Banking Association. Available from: <https://www.abe-eba.eu/N=EBA-Market-Guide-on-SCF.aspx>.
- Grosse-Ruyken, P.T., Wagner, S.M., Jönke, R., 2011. What is the right cash conversion cycle for your supply chain? *Int. J. Services Oper. Manag.* 10 (1), 13–29.
- Hofmann, E., 2005. Supply chain finance: some conceptual insights. In: Lasch, R., Janker, C.G. (Eds.), *Logistik Management—Innovative Logistikkonzepte*. Deutscher Universitätsverlag, Wiesbaden, pp. 203–214.
- Hofmann, E., Belin, O., 2011. *Supply Chain Finance Solutions: Relevance—Propositions—Market Value*. Springer, Heidelberg.
- Hofmann, E., Kotzab, H., 2010. A supply chain-oriented approach of working capital management. *J. Bus. Logist.* 31 (2), 305–330.
- Hofmann, E., Strewe, U.M., Bosia, N., 2018. *Supply Chain Finance and Blockchain Technology—The Case of Reverse Securitisation*. Springer, Cham.
- Klapper, L., 2005. *The Role of Factoring for Financing Small and Medium Enterprises*. The World Bank, Washington, DC.
- Koch, A.R., 1948. Economic Aspects of Inventory and Receivables Financing. *Law Contemp. Probl.* 13 (4), 566–578.
- Lekkakos, S.D., Serrano, A., 2016. Supply chain finance for small and medium sized enterprises: the case of reverse factoring. *Int. J. Phys. Distrib. Logist. Manag.* 46 (4), 367–392.
- Liebl, J., Hartmann, E., Feisel, E., 2016. Reverse factoring in the supply chain: objectives, antecedents and implementation barriers. *Int. J. Phys. Distrib. Logist. Manag.* 46 (4), 393–413.
- Malaket, A.R., 2014. *Financing Trade and International Supply Chains: Commerce Across Borders, Finance Across Frontiers*. Gower, Surrey.
- Petersen, M.A., Rajan, R.G., 1994. The benefits of lending relationships: evidence from small business data. *J. Finance* 49 (1), 3–37.
- Seifert, D., Seifert, R.W., Protopappa-Sieke, M., 2013. A review of trade credit literature: opportunities for research in operations. *Eur. J. Oper. Res.* 231 (2), 245–256.
- Templar, S., Hofmann, E., Findlay, C., 2016. *Financing the End-to-End Supply Chain—A Reference Guide to Supply Chain Finance*. Kogan Page Limited, London.
- Wuttke, D.A., Blome, C., Henke, M., 2013. Focusing the financial flow of supply chains: an empirical investigation of financial supply chain management. *Int. J. Prod. Econ.* 145, 773–789.
- Zhao, L., Huchzermeier, A., 2018. *Supply Chain Finance—Integrating Operations and Finance in Global Supply Chains*. Springer, Cham.

Packaging Logistics

Jesús García-Arca, Alicia Trinidad González-Portela Garrido, J. Carlos Prado-Prado, Industrial Engineering School, University of Vigo, Campus Lagoas-Marconsende, Vigo, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	119
Design Requirements and the Packaging System	119
Systems for Evaluating Design Alternatives	122
Organizational Considerations in Packaging Design	123
Conceptualizing the “Packaging Logistics” Approach	124
Conclusions	128
Biographies	128
References	128
Further Reading	129

Introduction

In increasingly turbulent and volatile markets, where competition is not so much between firms as between supply chains, organizations must design and implement initiatives to increase their competitiveness. Recent years have witnessed the appearance of four situations which, just like a perfect storm, are significantly affecting the efficiency and sustainability of supply chains: the globalization of purchases and sales; a constant increase in raw materials prices, particularly oil; the sudden arrival and implementation of new materials and production processes; and finally, the advent of new technologies and technological paradigms such as e-commerce, additive manufacturing, the Internet of Things (IoT) and radio frequency identification (RFID).

The combination of these four developments points to an urgent need to undertake actions that seek maximum performance in logistics activities taking place along the whole supply chain (production, transport, handling, storage, etc.) not only to eliminate waste or activities that do not add value (for example, by applying Kaizen or Lean Manufacturing approaches) but also to develop innovations in processes, products, and services.

At the same time, various stakeholders are showing increasing interest in developing sustainability (economic, environmental, and social) applied to supply chains, in line with the promotion of activities that foster the “Circular Economy.” Although a growing number of firms are backing the development of sustainable supply chains, most are still limiting themselves individually to meet legal commitments and sidestepping any “extended” responsibility they may have upstream or downstream along their supply chain. Nevertheless, different sources highlight an interest in developing sustainable supply chains given that they make it possible to save resources, reduce waste, and pollution levels, while at the same time providing competitive advantages.

In this context, packaging design is one of the key elements that can actively contribute toward improving the efficiency and sustainability of supply chains. Suitable packaging design can help eliminate “waste” such as unused storage, transport and sales points, difficulties in product handling, excessive use of materials, rejected products, and the loss of productivity in processes for packaging production, or product breakages and deterioration. What is more, packaging itself is a source of innovation in products, processes, and materials, which further strengthens its strategic contribution.

The packaging design process demands a holistic vision that allows efficient and sustainable alternatives to be sought. Well-known international firms, such as Ikea, Walmart, Procter & Gamble, Amazon, or Inditex, have turned their packaging into the cornerstone of their commercial and logistics strategies. However, the statement is not true just for firms operating in the consumer sector but also for those in the industrial sector. Proper integration of packaging design, product, and supply chain is the ultimate objective of “Packaging Logistics,” which this chapter deals with.

The chapter is structured into six sections, the first of which is this introduction. The following section details the different functional requirements associated with packaging design and how they are structured as a system. The third section covers the different assessment methods for packaging design alternatives, while the fourth deals with the different organizational connotations associated with the design process. The fifth section defines and reveals the “Packaging Logistics” approach and the chapter ends with a section giving some conclusions.

Design Requirements and the Packaging System

Before explaining the “Packaging Logistics” concept, it is necessary to know the different design functions or requirements that the packaging system must satisfy. Thus, beyond the traditional– and vital–function of protecting the products, the packaging must be designed, so that it not only has the capacity to differentiate the product commercially but also to provide production or logistics efficiency. That must all be done in a context in which sustainability is developed in its economic, environmental, and social dimensions.

Thus, the economic dimension could be analyzed from the different, and yet complementary, perspectives of income and expenditure. On the one hand, good packaging design can act as a “silent salesperson,” highlighting certain (tangible and intangible) features of products that increase their capacity to be different from others and surprise the market positively. On the other hand, suitable packaging can help reduce overall costs in the supply chain (particularly those for production and logistics), which is a critical aspect when it comes to conceptualizing and understanding the impact of certain design decisions. That impact, in terms of costs, will be dealt with in greater detail in the next section.

Regarding the environmental dimension of sustainability, packaging design also affects the amount of resources consumed. That consumption not only includes the materials used for the packaging, but also any other resources used along the supply chain, such as the fuel needed to transport the goods. Likewise, packaging design also affects the amount of waste and pollution produced. Logically, spoilt or damaged products caused by unsuitable packaging design have an equally important environmental impact.

Awareness and understanding of the environmental impact, particularly concerning waste and pollution, has led to the development of specific legislation, including European Directive 94/62/EC and its updates, which in practical terms has meant the promotion of the green dot for recycling and the adoption of specific mechanisms for managing packaging and its waste products (European Commission, 1994). That legislation, in line with growing public demands in environmental issues, establishes the prevention of waste generation as a conceptual priority, that is, the development of noncontaminating products and techniques in order to reduce any impact on the environment from the materials and substances used in packaging and present in its waste products during the production, commercialization, distribution, use, and elimination processes.

When this priority objective of impact prevention is not feasible, a second one seeks the use of any waste, starting with reuse of the packaging. If reuse is not an option, then recycling or adding value to the waste is the next ranked objective. Value is added by making the most of the resources contained in the waste packaging, including burning it for the energy it contains, so as to avoid any environmental impact resulting from the final treatment associated with its disposal, such as storing or dumping it in a controlled way or destroying it totally or partially by incineration or other methods that do not involve energy recovery.

In order to understand the impact that packaging waste has on the environment, it should be noted that it accounts for a third of all municipal solid waste. The European Commission estimated that in 2016, for each European citizen, 170 kg of packaging waste was generated (Eurostat, 2019), all of which not only needs managing at the end of its life cycle, but also means that there have been previous costs for the manufacturing firms in terms of raw materials, supplies, and so on.

Finally, packaging design also has an impact on sustainability's social dimension. This includes issues such as facilitating recycling schemes, supplying honest, understandable, and truthful information, guaranteeing safety when products are consumed, and adaptation of the product's use, ergonomics, and dose to the needs of different customers, such as senior citizens or people with disabilities.

On the basis of the above points on sustainability, the multi-dimensional impact and character of the design requirements that packaging must satisfy can be appreciated, along with the very different levels they exist on and needs they meet (Lindh et al., 2016). The different design requirements are summarized in Fig. 1. However, even though it is important to be aware of, and understand,

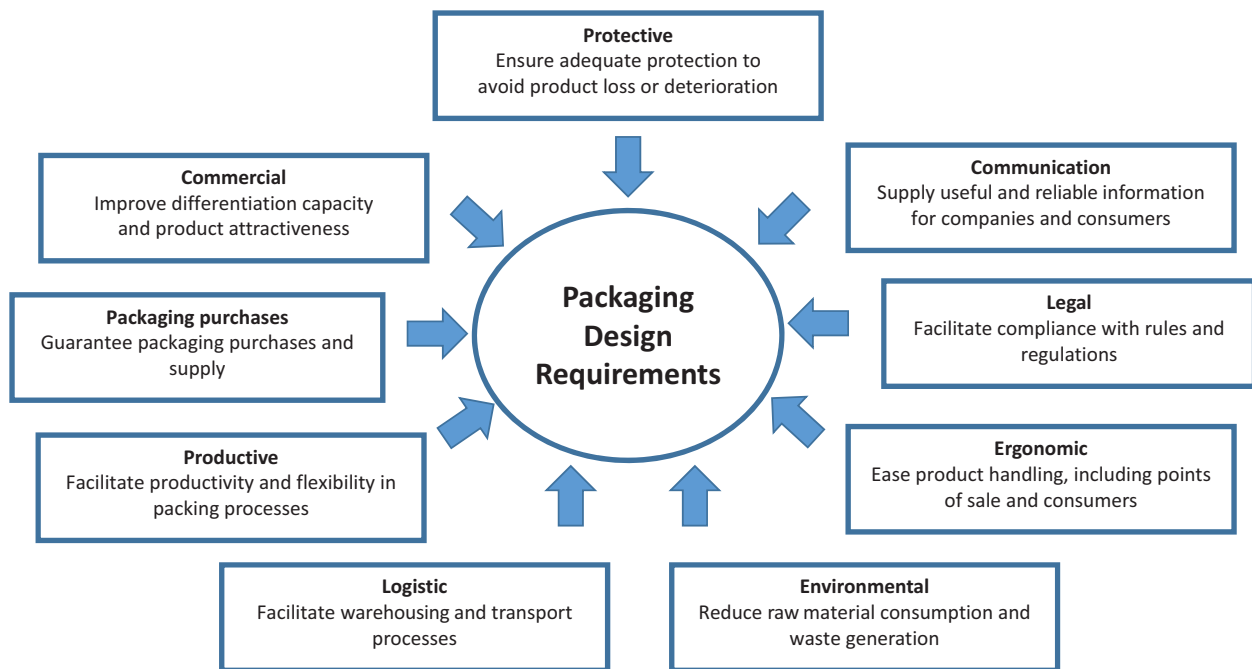


Figure 1 Packaging design requirements. Source: Own compilation and development

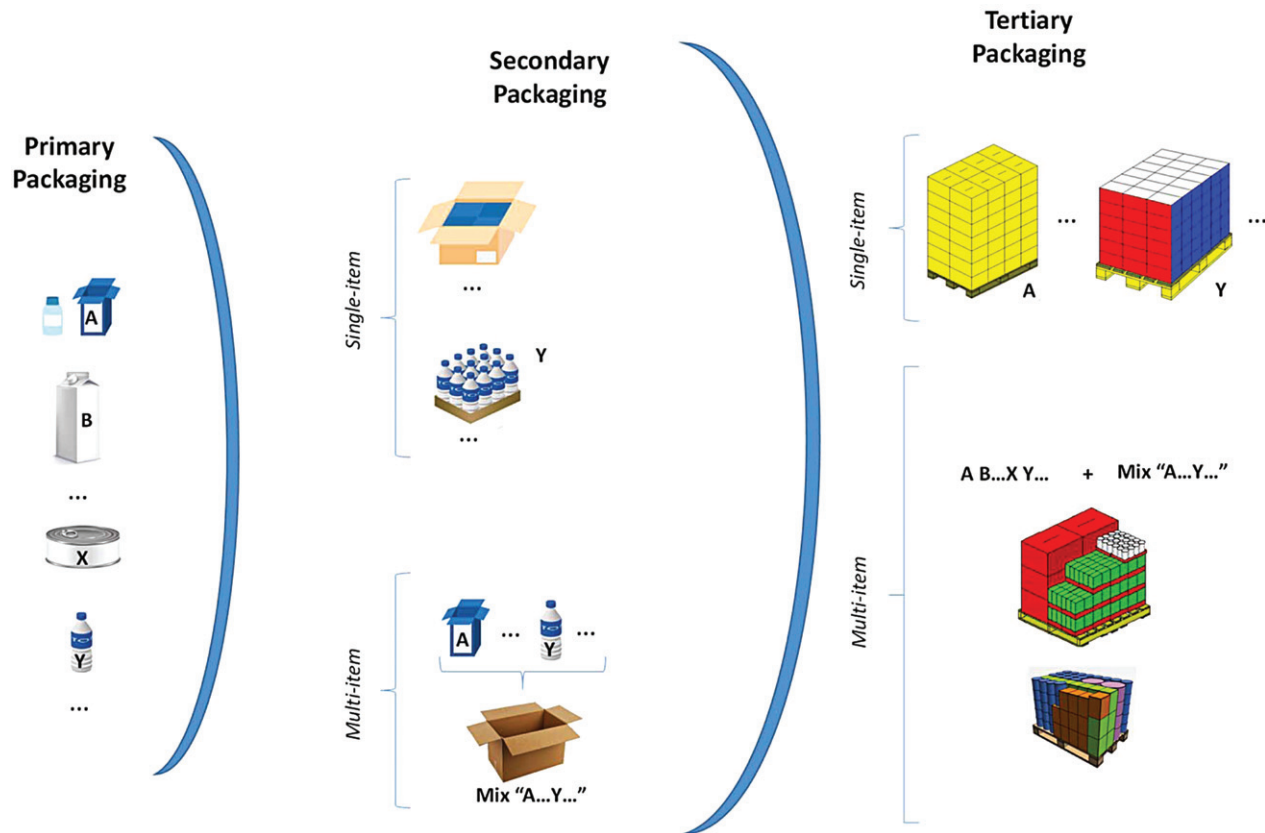


Figure 2 Structure of the packaging system. Source: Own compilation and development

the many varied design requirements that packaging is subject to, it is also necessary to know the structure of the packaging system itself, which typically has three levels: primary, secondary, and tertiary (Fig. 2).

"Primary" packaging is the product's first protector and is typically in contact with it; it is also known as the inner container, first container or consumer package. This level can simultaneously have various formats, materials and/or layers (e.g., a bottle in a cardboard box). "Secondary" packaging is designed to group several primary units of the same product or different products to provide protection and make handling easier during commercial distribution and sale. It is usually removed when the product is used or stored in the home. The common generic term is outer packaging or retail packaging. This level can also have several formats, materials, and/or layers (e.g., a cardboard tray and plastic film to group bottles of water or cartons of milk). Finally, "Tertiary" packaging is typically associated with several secondary units (of the same product or different products) which are grouped, for example, on a pallet or roll container to facilitate handling, storing, and transport tasks. This level is also called a unit load or transport packaging.

The fact that secondary and tertiary packaging can include one or more products is linked to the need to meet the needs of orders coming from firms, stores, or end customers (e.g., in e-commerce). Logically, given that these multi-product groupings usually diverge from the initial configuration established by the packaging company, additional handling activities are needed to prepare them and this will condition the efficiency and sustainability of the whole supply chain. Likewise, variables such as a product's size, its fragility, the number of units sold, and delivery frequency will also condition the design.

The three-level packaging system is applicable, conceptually, to all types of product, both in the consumer sector, where packaged products go to end customers, and in the industrial sector, where they go to be consumed and transformed by other firms. Clearly, if the supply chain changes, so will the design requirements. The interrelation between the three different levels of the packaging system (primary, secondary, and tertiary) should be coherent with the design requirements they are subject to. So, as there is a considerable dependency between the three levels in order to ensure that design requirements are met, adaptation to suit each level should not be evaluated separately, but rather in a joint and integrated way with the other two levels (Pålsson and Hellström, 2016). Furthermore, that integrated vision should be linked to the requirements and features of the product design and its associated supply chain.

So, for example, coordinated decisions about product, supply chain, and packaging can find a natural meeting point in the implementation of packaging postponement strategies. Such strategies pursue a reduction in the amount of stock of a certain product by delaying how it is specifically prepared for each customer. This aim is met by using a modular design for a product's

components and promoting flexibility in packaging processes. The more valuable a product is and the wider the range available, the greater the impact of packaging postponement will be.

Systems for Evaluating Design Alternatives

The business and academic literature highlights the lack of specific metrics for globally measuring and comparing the efficiency and sustainability of different packaging design alternatives. However, having a suitable evaluation system available for these alternatives is a critical factor when determining the most suitable design. It would thus seem reasonable to catalog the best packaging as that which simultaneously combines the most attractive and useful one for the customer (end or industrial) with the smallest overall cost, while at the same time generating the smallest impact on the environment (including the reduction in product losses and deterioration). As has already been mentioned, however, packaging is subject to a variety of design requirements. Logically, different metrics may exist that are linked to the application of each of those requirements. So, a suitable packaging design can highlight certain product features to increase its capacity to differentiate itself from others and, as a result, increase sales. In all events, sales are an after-the-fact indicator, as they measure suitability once new formats have already been implemented, which means that they are not the most reasonable way to approach the initial problem of measuring the pros and cons of each design alternative. One way of measuring this impact before a new format launch is based on techniques such as product, visibility, or stimulation tests “in the lab.” Likewise, those techniques can be complemented by using comparisons with competitors’ packaging (benchmarking), although the priority in this case is generally to study commercial aspects regarding the primary packaging.

At the same time, suitable packaging design can help reduce overall costs in the supply chain (**Table 1**). Cost reduction is an area where many firms find enough incentive to promote changes in their packaging. Regrettably, the relationship between those costs and any packaging design decisions is both direct—they show up directly in the firm’s accounts—and indirect, which means that the overall impact of certain designs does not always have “visibility.” This indirect relationship is what stops many firms from understanding the impact certain design packaging decisions have on the related costs. Furthermore, many firms do not even have reliable data and/or information to make the costs totally or partially visible or clear.

Nevertheless, when we attempt to measure the suitability of a certain packaging alternative from perspectives other than cost, the assessment or measurement system is not so clear and in many cases it is very difficult to fully convert them into economic

Table 1 List of potential costs affected either directly or indirectly by packaging design decisions in a retail supply chain

<i>Type of cost</i>	<i>Typical cost unit</i>	<i>Likely impact on the supply chain</i>	<i>Visibility of impact</i>
Raw materials supply	Purchase cost (including logistics costs linked to management, handling, storage, and transport)	Supplier and packer	Indirect
Purchase of packaging	Purchase costs (may or may not include supply cost)	Packaging manufacturer; Packer	Direct
Management, handling and storage of packaging	Cost per package (and/or grouping)	Packer	Indirect
Packing	Production cost (including costs for production management, machine preparation, packing process and rejects)	Packer	Indirect
Internal transport of finished product	Cost/pallet	Packer (between factories and warehouses; or to customers, if paid by packer)	Indirect
Internal handling, storage, and picking linked to finished product	Cost/pallet and/or cost/box (picking)	Packer’s warehouse	Indirect
External transport of finished product	Cost/pallet	Distributor (between packer and distributor’s warehouse when paid by distributor)	Indirect
External handling, storage, and picking linked to finished product	Cost/pallet and/or cost/box (picking)	Distributor’s warehouse	Indirect
Delivery transport	Cost/pallet	Distributor (from distributor’s warehouses to stores or sales points)	Indirect
Handling and storage at point of sale	Cost per pallet and/or cost per box	Stores or points of sale	Indirect
Environmental management of packaging waste	Cost per weight/volume depending on the material (may be linked to the green Dot cost)	Packer and distributor	Direct
Product losses due to inefficient packet and/or packing.	Cost of deteriorated returned products (including loss of product’s commercial value)	Any part of the supply chain (including stores and households)	Indirect

Source: Own compilation and development

terms. For example, whether packaging fulfills the protective function could be partially measured in economic terms by the cost of losses, deterioration, or returns. However, it would prove difficult to measure the dissatisfaction that these problems create in each section of the chain, particularly for the end user, as it is not necessarily reflected in sales (at least not immediately).

A similar difficulty emerges when trying to measure the environmental impact of the different packaging alternatives, because even though some of the effects can partially be measured in cost terms (e.g., waste management costs paid for by the Green Dot license fee), others are measured using scales that are not economic. For instance, the most widely-used system to measure environmental impact, "Life Cycle Assessment" (LCA), as embodied within ISO 14040:2006, is a technique that measures the carbon footprint generated not only by the packaging but also, more comprehensively, by the products throughout their life cycle (ISO, 2006).

With the rise of "ecodesign," the use of this technique has included different criteria and parameters for assessing environmental impact. A web-based platform used by firms such as Procter & Gamble, Johnson and Johnson and UPS, the COMPASS (Comparative Packaging Assessment) tool stands out among the more relevant initiatives which measure and compare packaging design from an environmental perspective and have the LCA technique in mind.

In all events, in the face of all the difficulties when objectively assessing each packaging alternative from a multifunctional perspective (i.e., not just environmental impact or cost) and under a global chain approach, different assessment models have been developed that combine quantitative and qualitative scales. By doing so, however, a certain degree of subjectivity is introduced into the results of any analysis. The most widely-used of these models is the Packaging Scorecard method (Olsmats and Dominic, 2003), popularized by firms such as IKEA. By analyzing cases of supply chains or specific packaging, they establish an assessment system for alternatives that combines the needs of each stage of the supply chain with specific qualitative and quantitative criteria.

Likewise, the literature includes other techniques and tools to deal with more specific aspects in the search for the best packaging design alternatives. Thus, measurement systems have been proposed that seek size optimization for boxes and/or volume optimization for packaging or loads. For the latter, there are IT design tools that prioritize the search for improvements in logistics efficiency (mainly to improve the occupied space in boxes, pallets, trucks and/or intermodal containers) by using different heuristic, metaheuristic, or integer programming techniques.

Finally, it is worth noting that over the last few years in the consumer sector, a set of IT tools has appeared aimed at managing more efficiently the space products take up on point-of-sale shelves. These shelf management programs not only look for the best physical use of the space, but also try to connect brand image and product categories (facings) with their economic sales performance by defining "planograms." Logically, these considerations on the rationalization of space at the point of sale are of little interest on purely electronic sales platforms.

Organizational Considerations in Packaging Design

After describing the design requirements, how the packaging system operates and the methods for measuring and assessing different alternatives, it is necessary to look at the structure of the design process itself, that is, delve into the organizational considerations. Logically, the variety of requirements demands a certain level of external and internal coordination in the design process. The importance placed upon each design requirement will depend on which area, department or firm within the supply chain is making the assessment. Furthermore, these design requirements are not evenly spread out over the various levels of the packaging system (primary, secondary, and tertiary), and neither are they to be found at each stage, process, or task within the chain.

At the same time, the final decisions affecting the three different levels of the packaging system could also be related to the part of the supply chain that is able to exercise greater influence or leadership. For example, if the distribution companies "dominate" over the packing firms (which is common place in the consumer sector, for example), the final decisions relating to packaging could be dependent on what the distributor considers a priority. Logically, this approach will not necessarily lead toward the best packaging solution from the perspective of the supply chain as a whole. Therefore, the organizational structure adopted should have the job of taking decisions about packaging design in a way that coordinates and seeks consensus among the affected parties, that is, the different areas or departments within each firm and the different firms within the supply chain (packaging manufacturers, packers, distributors, logistics operators, etc.). Likewise, it would be desirable for this organizational structure to be coordinated and integrated with those responsible for taking decisions about product design and about the supply chain. Conceptually, there are four ways to address this organization in order to face the structure of the design process as a whole (Olander-Roese and Nilsson, 2009):

- The basic or "traditional" approach. Packaging design is inefficient, starting during the last stages of development for a new product (and its associated supply chain).
- The "simultaneous" approach. Product design (and its associated supply chain) is developed at the same time as packaging design.
- The "static" approach. Packaging design is undertaken in a way that is integrated into the first stages of product design, but only the first versions. It is not then updated dynamically when the product and/or supply chain need changes in order to adapt to new design requirements and/or threats and opportunities in the environment.

- The “dynamic” approach. Packaging design leads and promotes the design of the product itself (and its associated supply chain), that is, the packaging provides the basis for innovative design of new products, updating its design in the face of changes in the requirements and/or threats and opportunities in the environment.

When the “dynamic” approach is used, there is not only a greater number of efficient and sustainable alternatives, but also feedback of coordinated criteria for selecting between them, creating a true driver for change and continuous improvement in the processes taking place throughout the supply chain. This is the approach put forward by “Packaging Logistics” which is explained in the next section. The design decisions that this organizational structure should be taking at a packaging level will be centered on five complementary and closely interconnected aspects:

- Selection of the materials to be used, including their type and quality.
- Selection of the technologies related to the processes in which the packaging will be involved. These technologies include, for example, those adopted in packaging processes or in handling, storing and transport, and will contemplate what level of automation to use. Likewise, identification technologies (such as RFID or barcodes) that enable tracing of the physical flow of products, in line with an approach linked to the IoT.
- The packaging’s esthetic design (shape, colors, text, symbols, images, etc.).
- Details of the structure in terms of levels and interconnections in the packaging system (primary, secondary, and tertiary).
- Selection of the dimensions used in the packaging.

Although all the decisions to be taken are relevant, the choice of dimensions is given here to illustrate the impact of such a selection in the areas of production, logistics, and the environment. In recent years there has been some standardization in dimensions, particularly for secondary packaging, where the use of 600 mm × 400 mm modules and its submultiples has been promoted in the ISO 3394:2012 standard (ISO, 2012). Implementing this standard facilitates the efficient use of some of the most commonly used pallets in the world (Fig. 3): the EUR (1200 mm × 800 mm) and the American (1200 mm × 1000 mm) pallets. Such standardization in packaging dimensions (for both the base and the height) also facilitates dimensional standardization to a large extent in resources used for handling, storing, and transport, which in turn facilitates the automation of some processes and, therefore, has an effect on logistics efficiency, and sustainability.

If standard format packaging is chosen, which has high efficiency in terms of logistics (palletizing) and production (reduced number of changes and adjustments in packaging processes), lower costs can be obtained. However, this approach may mean the sacrifice of some possible ways to differentiate the product. There are also questions to be asked about how product protection is ensured or how waste is minimized. Furthermore, in increasingly complex global supply chains, standardization (and automation) based on palletizing criteria can differ from those employed in courier systems or 20- or 40-foot intermodal containers. Couriers, often used by e-commerce, seek minimized size and weight and reduced risk of breakage, whereas intermodal containers, typically used in industrial logistics between manufacturers, do not efficiently fit pallets into the available space. Answers are also needed to the question of what dimensions to adopt when products with different packaging are mixed and efficient, sustainable deliveries must be still achieved along with the whole supply chain.

Looking in further depth at this issue, there are over 50 different cardboard boxes (typically used for secondary packaging) on the market. Fig. 4 shows one of the most widely-used types, which has overlapping lid flaps. For each target volume (V) there are an infinite number of alternatives combining different lengths, widths, and heights (shown as feasible alternatives on the graph). However, there is only one alternative for each volume V that optimizes the amount of cardboard needed (S) and the waste generated. The *blue line* on the graph shows this set of alternatives.

On a different note, throughout the course of structuring the process for product design, packaging, and supply chain, it can be useful to develop standards used for different management fields. The best known among these are quality management (e.g., the ISO 9000 standards) or environmental management (e.g., the ISO 14,000 standards). The literature particularly mentions the positive effect of these in terms of structuring and systemizing certain areas of management such as defining tasks, responsibilities, or deadlines. In packaging design, this positive effect could be associated, for example, with the documentation for the design process itself (for tasks, leaders, and deadlines) or recording technical, production, or logistics documentation associated with the process (e.g., technical sheets for purchasing and supply or palletizing forms).

Conceptualizing the “Packaging Logistics” Approach

Within the context described above of requirements, assessment systems and organizational considerations, the need arises to conceptualize the close relationship between the packaging design process, product design and supply chain (logistics) system design. This is the perfect scenario for developing the “packaging logistics” approach. A very early vision of the approach, still unnamed, dates back to academic circles in the 1990s, although it was very limited as it basically highlighted the relationship between physical distribution processes (handling, storage, and transport). The term used today was coined at the end of the 1990s in northern Europe, and was initially proposed as a structured relationship between the logistics system and the packaging system that would enable firms to improve profits (Björnemo et al., 2000). That notion was later developed and qualified as a more academic concept based on contributions from different authors, particularly from Lund University in

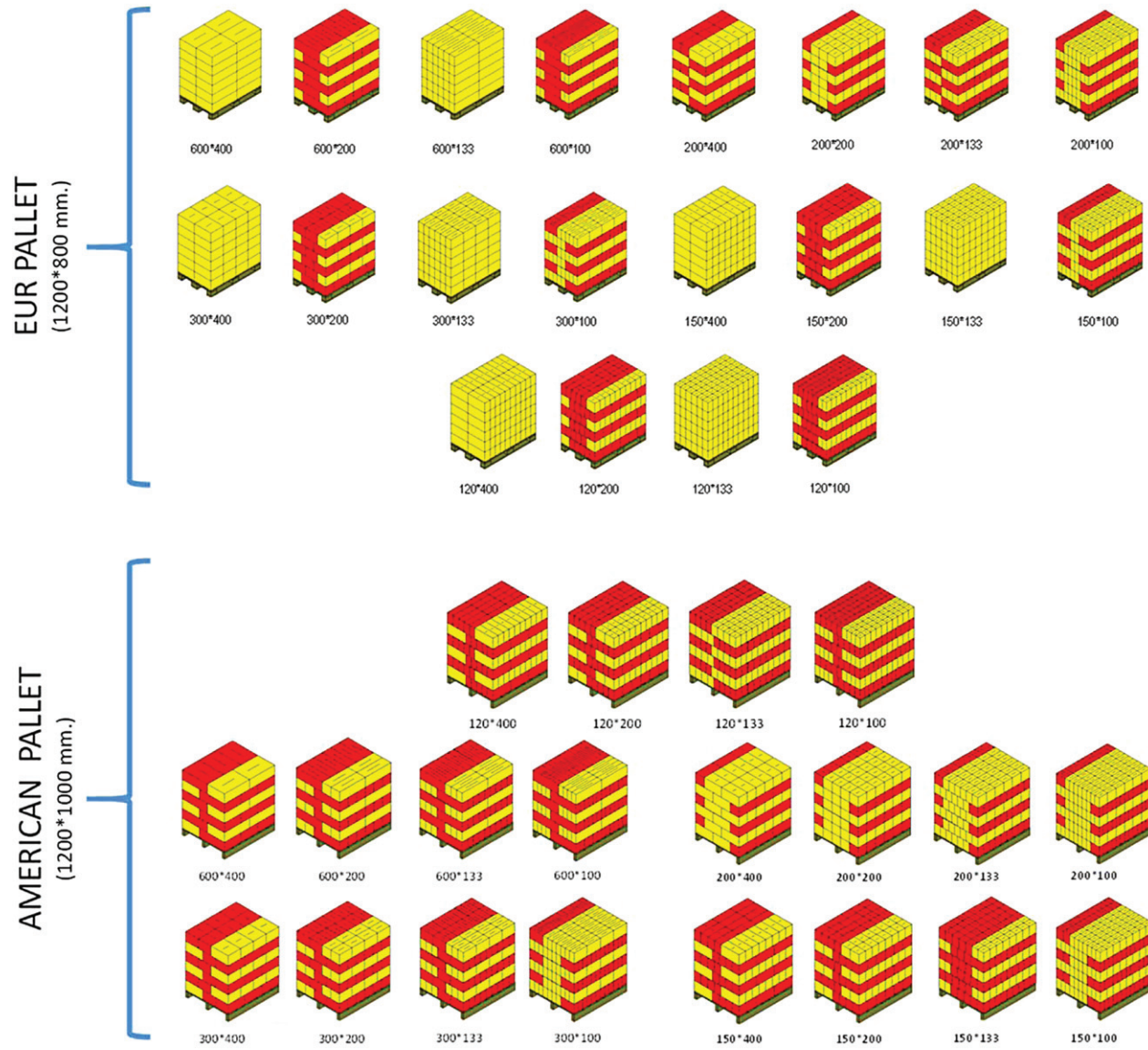


Figure 3 Efficient box dimensions for use on standard pallets under the 600*400 modular system. Source: Own compilation and development

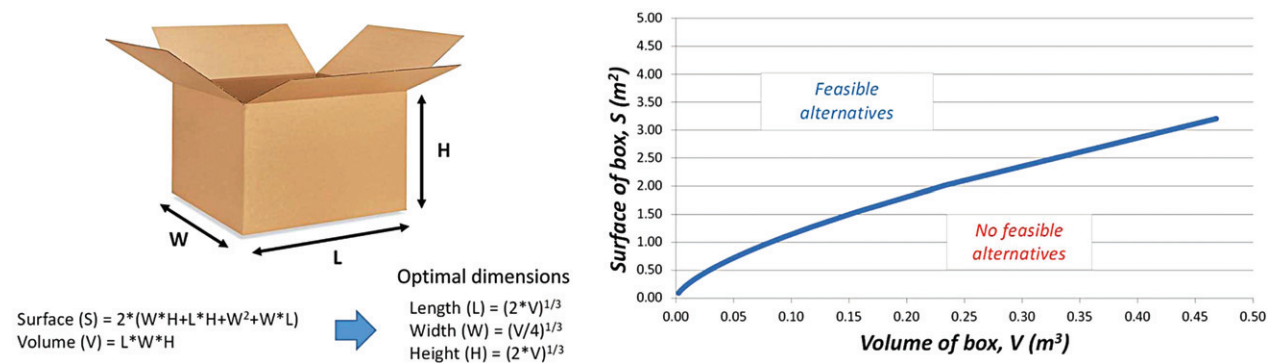


Figure 4 Example of the search for dimension alternatives in box design Source: Own compilation and development

Sweden. However, this description should be considered as a first approach at a more complex definition, because even though it already highlighted the interaction between the logistics system and the packaging system, it oversimplified integration in the overall concept of the firm (rather than the supply chain) and sidestepped any relationship with product design.

Saghir (2002, p. 45) proposed a more integrated definition of the concept, “The process of planning, implementing, and controlling the coordinated packaging system of preparing goods for safe, secure, efficient and effective handling, transport, distribution, storage, retailing, consumption and recovery, reuse or disposal and related information combined with maximizing consumer value, sales and hence profit.” This definition now emphasized the active role of the packaging system in order to integrate different viewpoints at commercial, logistics and environmental levels and to affect the product system and the supply chain itself—with everything oriented toward improvement in the competitiveness of firms.

Later, García-Arca et al. (2014, p. 330), would adjust, qualify, and enlarge on the “packaging logistics” approach put forward by Saghir (2002) to include the context of sustainability, coining the term “Sustainable Packaging Logistics” (SPL), which they defined as: “The process of designing, implementing, and controlling the integrated packaging, product, and supply chain systems in order to prepare goods for safe, secure, efficient and effective handling, transport, distribution, storage, retailing, consumption, recovery, reuse or disposal, and related information, with a view to maximizing social and consumer value, sales, and profit from a sustainable perspective, and on a continuous adaptation basis.”

This new definition expressly mentions the processes involved, the sustainable sphere (social, economic, and environmental) and the dynamic vision of the five design decisions associated with packaging. As mentioned previously, this dynamic vision is critical, if the design requirements are to be adapted or fine-tuned to the threats and opportunities of their surroundings. Likewise, this definition reinforces the idea of integration of the three systems involved: packaging, product and the supply chain itself. Finally, it strengthens the idea of connecting packaging design with a firm’s competitive improvement, that is, obtaining better business results. In this context, there are three pillars supporting deployment of the “packaging logistics” approach in its broadest sense, that is, its contribution to deployment of the three areas of sustainability (economic, social, and environmental) (García-Arca et al., 2017):

- Definition of the design requirements, based on identification from a sustainable perspective of the overall needs of the different stages of the supply chain. The current level of competition in most markets would suggest integrating all these requirements without disregarding any of them in an attempt to achieve objective consensus that will satisfy all the priorities of the areas and firms along the whole supply chain.
- Definition of a system enabling measurement of the impact of certain design decisions (materials, technologies, levels structure, dimensions, and esthetics) on the efficiency and sustainability of the supply chain. This pillar is linked to the need to measure and compare costs, environmental impact and any other associated variable mentioned earlier in this chapter.
- Definition of an organizational structure to coordinate all the areas affected by packaging design along the whole supply chain, both internally—within each firm—and externally—between firms in the chain. This would mean integrating packaging design into the initial stages of product and logistics system design, and using a “dynamic” approach in order to adapt constantly not only to the different views and perceptions that each stage of the chain has of the design requirements, but also to the changes that can occur in the design requirements themselves and/or the environment. In practice, this would mean firms and supply chains learning and adapting efficiently and sustainably in packaging issues, that is, acting as “learning organizations.”

At the same time, the deployment of these three pillars would promote a culture of packaging-based change and continuous improvement in firms, which could be approached from different and yet complementary perspectives: on the one hand by adopting “management best practices” in the design process; and on the other by promoting changes and innovations in the packaging. The scope of best practice could include all those actions that pursue an efficient and integrated design process. This could include, for example, providing other parts of the chain with knowledge of logistics processes, documenting the design aspects (for instance, generating design procedures, technical instructions for materials, or palletization sheets), or using IT design tools to simplify decision-making.

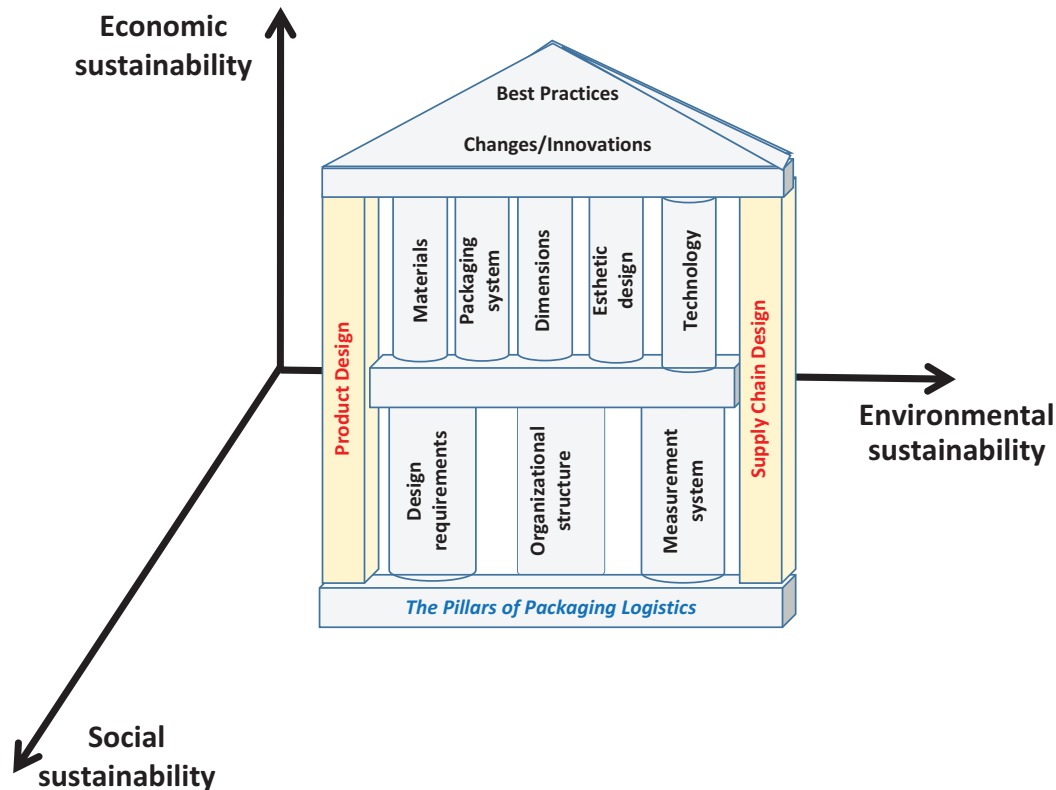
The scope of change and innovation could include improvement guidelines that have already proved successful in other firms but which cannot be applied indiscriminately without prior knowledge of factors that condition the product, the other design requirements, and the supply chain itself. Logically, the field of innovations would include those provided by the dynamic packaging industry (e.g., in terms of materials, qualities, and packaging equipment or processes). Table 2 shows these possible changes or innovations and indicates the potential impact on overall costs.

Finally, in Fig. 5, there is a diagram of the authors’ management model for “packaging logistics” deployment. This diagram reflects the conceptual aspects mentioned throughout this section and uses a building as an analogy, in that, the more suitable the pillars of its structure become, the stronger it is able to function and propose and improve packaging alternatives, by taking the five associated decisions and growing and developing the three areas of sustainability (economic, social, and environmental). That is all done as an integrated part of the product and supply chain design.

Table 2 List of potential actions to improve packaging

Action	Main costs affected
Changes in packaging dimensions (primary, secondary and/or tertiary packaging)	Packaging purchasing costs Logistics costs (handling, storage and transport) Waste management costs
Change in the materials used in packaging (material type, quality, etc.) (including changes to improve the capacity to recycle or take advantage of the materials)	Packaging purchasing costs Waste management costs
Variation in the amount of product per primary packaging unit or the shape/size of the product	Packaging purchasing cost (per unit of product) Logistics cost (handling, storage and transport) (per unit of product) Waste management costs (per unit of product)
Change in the packing process (new equipment, new tools, etc.)	Packaging production cost Deterioration or reject costs
Change in the number of primary packaging units per secondary packaging unit	Packaging purchase cost (per unit of product) Logistics cost (handling, storage, and transport) (per unit of product) (includes handling at point of sale) Waste management costs (per unit of product)
Standardization in dimensions and formats for packaging	Packaging purchasing cost (economies of scale) Production cost (reduction in setup times)
Standardization in the qualities of the materials used in packaging	Packaging purchasing cost (economies of scale) Production cost (reduction in preparation times)
Elimination of "excess packaging"	Packaging purchasing cost Waste management costs Packaging purchasing cost
Change in the graphic art design or esthetics of the packaging (colors, texts, labels, shapes, etc.)	Packaging purchasing cost
Implementation of returnable packaging	Packaging purchasing cost Logistics cost (handling, storage, and transport; includes reverse logistics) Waste management costs
Implementation of SRP formats ("Shelf Ready Packaging")	Packaging purchasing cost (per unit of product) Packaging production cost Logistics cost (handling, storage, and transport) (per unit of product) (includes handling at point of sale) Waste management costs (per unit of product)
Rationalization of packaging and pallets for raw materials and components	Logistics cost associated with supplies (handling, storage and transport)

Source: Own compilation and development

**Figure 5** Deployment model for "Packaging Logistics" from a sustainable perspective. Source: Own compilation and development

Conclusions

The implementation of the “packaging logistics” approach mentioned in this chapter should actively contribute to the competitive improvement of a supply chain and, by extension, to each and every firm within it. It is an implementation that will have been developed from a sustainable perspective.

Researchers can also find new lines for research in this chapter related to the role of packaging as a promoter of competitiveness for firms and supply chains within the contexts of sustainability (less consumption, waste, pollution, and losses), market globalization (including developing countries), commercialization channel diversity (including e-commerce and logistics flows between manufacturers) and new technologies in materials, processes, equipment, and traceability (IoT). Likewise, the way a design process is structured, in the pursuit of internal and external coordination in order to integrate design into a scope that encompasses product-packaging-supply chain, is another fascinating line for research that will require future study.

At the same time, practitioners can find in this chapter a structured methodological proposal for developing “packaging logistics” in their firms, based on the pillars of decisions and improvements that aim to obtain competitive advantages. The chapter not only explains how it is possible to align the “packaging logistics” approach with the strategy of firms, but also with that of their supply chains. After approaching the issue from the two standpoints of academic research and business practice, some final points are offered for reflection to arouse or encourage interest among professionals in businesses and research institutions when it comes to paying attention to packaging from a “packaging logistics” perspective:

- Suitable packaging design facilitates the capacity to make savings, improve sales, and operate in a more sustainable way.
- To obtain those results, it is necessary to identify the design requirements, measure, and compare alternatives from an all-encompassing perspective and strengthen collaboration and coordination between all the areas, departments, and firms in the supply chain.
- The more integrated the packaging, product, and design processes are, the more efficient and sustainable any adopted solutions will be.
- The selected alternatives for packaging should not be considered from a static standpoint, but from a dynamic one which is subject to potential changes and improvement.
- The final point for reflection is to point out that good packaging may not turn a bad product into a good one but, in almost all certainty, bad packaging will turn a good product into a bad one.

Biographies

Jesús García-Arca, PhD. He graduated in Industrial Engineering and is a Tenured Associate Professor at the Engineering School of the University of Vigo (Spain).

A. Trinidad González-Portela Garrido, PhD. She graduated in Business and Economic Sciences and Law and is a lecturer at the Engineering School of the University of Vigo.

J. Carlos Prado-Prado, PhD. He graduated in Industrial Engineering and is a Full Professor at the Engineering School of the University of Vigo.

They all have extensive experience in teaching, research and consultancy in the field of supply chain design, management and improvement, including efficient and sustainable packaging design.

References

- Björnemo, R., Jönson, G., Johnsson, M., 2000. Packaging Logistics in Product Development, Proceedings of the 5th International Conference: Computer Integrated Manufacturing Technologies for New Millennium Manufacturing, Singapore, pp. 135–146.
- European Commission, 1994. Directive 94/62/EC on Packaging and Packaging Waste, European Union, Brussels.
- Eurostat, 2019. Packaging Waste Statistics. Available from: https://ec.europa.eu/eurostat/statistics-explained/index.php/Packaging_waste_statistics.
- García-Arca, J., Prado-Prado, J.C., González-Portela Garrido, A.T., 2014. Packaging logistics: promoting sustainable efficiency in supply chains. *Int. J. Phys. Distrib. Logist. Manag.* 44, 325–346.
- García-Arca, J., González-Portela Garrido, A.T., Prado-Prado, J.C., 2017. Sustainable packaging logistics: the link between sustainability and competitiveness in supply chains. *Sustainability* 9 (7), 1098.
- ISO, 2006. ISO 14040: 2006. Environmental management. Life cycle assessment. Principles and framework, Geneva, International Organization for Standardization.
- ISO, 2012. ISO 3394: 2012. Packaging. Complete, filled transport packages and unit loads. Dimensions of rigid rectangular packages, Geneva, International Organization for Standardization.
- Lindh, H., Williams, H., Olsson, A., Wikström, F., 2016. Elucidating the indirect contributions of packaging to sustainable development: a terminology of packaging functions and features. *Packag. Technol. Sci.* 29 (4–5), 225–246.
- Olsmats, C., Dominic, C., 2003. Packaging scorecard. A packaging performance evaluation method. *Packag. Technol. Sci.* 16, 9–14.
- Olander-Roese, M., Nilsson, F., 2009. Competitive advantages through packaging design- prepositions for supply chain effectiveness and efficiency. *Proc. Int. Conf. Eng. Des. (ICED 2009)*, USA, Stanford University, pp. 279–290.
- Pålsson, H., Hellström, D., 2016. Packaging logistics in supply chain practice – current state, trade-offs and improvement potential. *Int. J. Logist. Res. Appl.* 19 (5), 351–368.
- Saghir, M., 2002. Packaging logistics in the Swedish retail supply chain. Lund University, Sweden.

Further Reading

- Azzi, A., Battini, D., Persona, A., Sgarbossa, F., 2012. Packaging design: general framework and research agenda. *Packag. Technol. Sci.* 25 (8), 435–456.
- García-Arca, J., Prado-Prado, J.C., García-Lorenzo, A., 2006. Logistics improvement through packaging rationalization: a practical experience. *Packag. Technol. Sci.* 19 (6), 303–308.
- García-Arca, J., Prado-Prado, J.C., 2008. Packaging design model from a supply chain approach. *Supply Chain Manag.: Int. J.* 13 (5), 375–380.
- Hellström, D., Nilsson, F., 2011. Logistics-driven packaging innovation: a case study at IKEA. *Int. J. Retail Distrib. Manag.* 39 (9), 638–657.
- Hellström, D., Saghir, M., 2006. Packaging and logistics interactions in retail supply chain. *Packag. Technol. Sci.* 20 (3), 197–216.
- Kye, D., Lee, J., Lee, K., 2013. The perceived impact of packaging logistics on the efficiency of freight transportation (EOT). *Int. J. Phys. Distrib. Logist. Manag.* 43 (8), 707–720.
- Pålsson, H., 2019. *Packaging Logistics. Understanding and Managing the Economic and Environmental Impacts of Packaging in Supply Chains*. Kogan Page, London.
- Regattieri, A., Santarelli, G., Piana, F., 2019. *Packaging Logistics in Operations, Logistics and Supply Chain Management*. Springer International Publishing, Luxemburg.
- Saghir, M., Jönson, G., 2001. Packaging handling evaluation methods in the grocery retail industry. *Packag. Technol. Sci.* 14 (1), 21–29.
- Sohrabpour, V., Hellström, D., Jahre, M., 2012. Packaging in developing countries: identifying supply chain needs. *J. Humanit. Logist. Supply Chain Manag.* 2 (2), 183–205.
- Svanes, E., Vold, M., Møller, H., Pettersen, M.K., Larsen, H., Hanssen, O.J., 2010. Sustainable packaging design: a holistic methodology for packaging design. *Packag. Technol. Sci.* 23 (3), 161–175.
- Vernuccio, M., Cozzolino, A., Michelini, L., 2010. An exploratory study of marketing, logistics, and ethics in packaging innovation. *Eur. J. Innov. Manag.* 13 (3), 333–354.
- Wikström, F., Verghese, K., Auras, R., Olsson, A., Williams, H., Wever, R., Grönman, K., Pettersen, M.K., Møller, H., Soukka, R., 2019. Packaging strategies that save food: a research agenda for 2030. *J. Ind. Ecol.* 23 (3), 532–540.

The Bullwhip Effect

Jan C. Fransoo*, Maximiliano Udenio[†], *Tilburg School of Economics and Management, Tilburg University, Tilburg, The Netherlands;
[†]Research Center for Operations Management, KU Leuven, Leuven, Belgium

© 2021 Elsevier Ltd. All rights reserved.

Definition	130
Causes of the Bullwhip Effect	131
Delay Structure	131
Behavioral Causes of the Bullwhip	132
Rational Decision-Making Causes	132
Measuring the Bullwhip Effect	133
The Bullwhip in Real Life—Anecdotal Evidence	133
The Bullwhip in Real Life—Empirical Evidence	133
Evidence from Macroeconomic Shocks	133
Taming the Bullwhip	134
References	134

Definition

The bullwhip effect is a supply chain phenomenon comprising two information distortion mechanisms. The first is that, for a given firm, the orders it places to its suppliers tend to be more variable than the demand it observes from its customers. This is called “demand distortion”. The second mechanism is the observation that demand distortion increases the further upstream a firm is in its supply chain (i.e., the further away it is from the final consumer). This is called “variance amplification.”

Due to the combined effects of demand distortion and variance amplification, a demand shock downstream generates demands that oscillate with increasing amplitude at each successive stage of the supply chain. This is said to resemble the visual effect of a bullwhip, with which a cattle farmer can use a flick of the wrist to break the sound barrier; hence the name. **Fig. 1** illustrates this resemblance, showing the evolution of demand on a stylized supply chain following a sudden, 2%, change in the end market.

Even though the term itself dates from the late 1990s (see [Lee et al., 1997a, 1997b](#), for the earliest appearances of the term), the mechanisms behind the bullwhip effect were first described in the late 1950s by Jay Forrester, a professor at Massachusetts Institute of Technology. Thus the bullwhip effect is also frequently referred to as the “Forrester effect,” named after his seminal initial work ([Forrester, 1961](#)). The bullwhip effect is often described in research through stylized models, typically using a nonstationary downstream demand stream (e.g., a sudden shock) as a trigger for instability. However, the bullwhip effect can also be observed under stationary demand.

One of the most significant contributions of the work of [Lee et al. \(1997a\)](#) was to show, mathematically, that the bullwhip effect could appear even under conditions of full rationality. Thus, while irrational behavior can increase its impact ([Croson and Donohue, 2006](#); [Sternan, 1989](#)), its appearance is not limited to improperly managed firms.

The bullwhip effect is undesirable because it generally increases operational costs and may cause further management problems in the supply chain, such as poor service levels. Demand distortion leads to inefficiencies in production and inventory management: highly variable production and/or inventories are required to serve highly variable (often oscillatory) customer demands. Therefore the bullwhip effect is directly responsible for costs associated with starting and stopping production equipment, successive idling and overtime periods, and/or workforce changes. Moreover, the bullwhip effect can also have a negative impact in ways that are difficult to quantify, such as difficulties in forecasting demand, and customer relations problems such as stockouts or delays in supply ([Disney et al., 2013](#)).

Because of variance amplification, the bullwhip effect is specially problematic for firms that are upstream in the supply chain. The compounding effect of variance amplification increases the impact of the bullwhip effect on firms located upstream within a supply chain; a chemical company, for instance, will typically suffer a larger bullwhip than a retailer. In fact, the inefficiencies associated with the progressive increase in upstream variability have a measurable effect on the relative production and inventory costs for upstream firms. Some studies estimate that these inefficiencies account for excess costs between 10% and 25% ([Lee et al., 1997a](#)).

From an information perspective, the importance of the bullwhip effect lies in the fact that it makes it increasingly difficult for firms to extract meaningful information from demand data. As one travels upstream in a supply chain, demand signals progressively diverge from the original consumer end-market demand. It is possible that after a few stages, very little resemblance remains ([Udenio et al., 2015](#)). Therefore, it is difficult for upstream firms to use their own demand signals to estimate consumer demand and to account for changes in the consumer market. In fact, most common examples and case studies based on the bullwhip effect illustrate this particular aspect; that is, how the production of diapers is more variable than what would be expected from the (relatively stable) usage by babies and toddlers, or how the DRAM market faces a much higher volatility than the computer market.

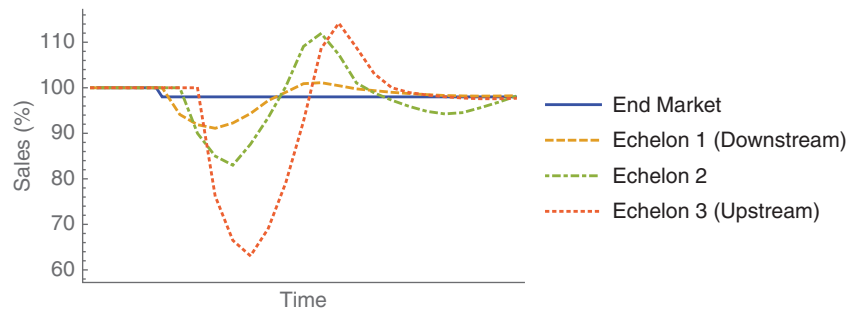


Figure 1 Illustration of variance amplification of demand on successive supply chain echelons.

This is of particular importance when consumer demand experiences a shock and when no information is shared among firms. In fact, minor consumer demand shocks result in upstream production levels that oscillate around consumer demand levels; making it impossible for upstream firms to link their demand observations with downstream market behavior merely based on the demand signal. This lack of visibility makes accounting for the bullwhip effect (i.e., taming the bullwhip) a difficult problem. Depending on the structure of the supply chain, downstream shocks can propagate and the related oscillations can persist for years.

In the academic literature, the bullwhip effect is classified in numerous ways. In this article, we analyze the bullwhip effect through three different theoretical causes. Namely, the bullwhip effect caused by delays within the supply chain structure (Forrester, 1961), the bullwhip effect caused by behavioral biases (Sterman, 1989), and the bullwhip effect caused by rational decision-making policies (Lee et al., 1997a). Whereas any one of the above is able to explain the appearance of the bullwhip effect, it is typical—in real life—to find more than one working concurrently.

The bullwhip effect is commonly introduced to students and executives in the classroom, using the “beer distribution game” developed by John Sterman. In the beer distribution game (see Sterman, 1989, for a detailed overview), four players (or eventually, groups of players) need to coordinate a supply chain, consisting of four echelons, without any interaction among the echelons being allowed. Commonly, this leads to excessive variance amplification over the course of the game. The bullwhip effect observed during a typical session of the beer game is caused by a combination of all the causes we discuss in this entry: the delay structure of the supply chain makes the oscillations all but inevitable, whereas the behavioral biases of the players, as well as a number of the structural causes, exacerbate the effect.

It is important to note that the bullwhip effect is at odds with the economic theory of production smoothing, whereupon production is kept constant in time and demand fluctuations are absorbed entirely through inventory (Abramovitz, 1950). The theoretical justification behind production smoothing is the assumption that changes in production output are generally more expensive than the necessary inventory buffers, thus it would be rational for firms to maintain production as stable as possible; allowing inventory levels to increase when sales decrease and vice versa. Successive steps of such countercyclical behavior implies that in a supply chain the variability of upstream demand would be lower than the variability of the end market.

Production smoothing was a common assumption in macroeconomic modeling until the 1980s. However, this modeling paradigm is difficult to justify empirically (Blinder and Maccini, 1991). At a macroeconomic level, inventory investment is strongly procyclical and, as discussed, production tends to be more variable than sales. To accommodate these observations, a number of economists have expanded the production smoothing model, noting that the addition of cost and/or productivity shocks can reconcile the model to observations. A different view in macroeconomic modeling eschews the production smoothing model altogether, arguing that avoiding stockouts is the more important metric for a firm, and that procyclical behavior and a production variability that is larger than demand variability (i.e., the bullwhip effect) follow naturally from this assumption (Kahn, 1992).

Causes of the Bullwhip Effect

The bullwhip effect was first analyzed in the late 1950s as a product of inventory and ordering policies and inherent delays in the supply chain structure. In the late 1980s, classroom experiments using the beer distribution game were conducted and human biases were introduced to extend the understanding and analysis. In the late 1990s, a more mathematically rigorous analysis of the bullwhip effect showed that supply chain decision-making, without the impact of delays or human biases, is enough to trigger the bullwhip. This triggered extensive research work encompassing more formal modeling, more empirical work at firm, supply chain, and industry sector levels, as well as more laboratory experiments detailing out earlier understanding.

Delay Structure

In the late 1950s, Jay Forrester introduced “industrial dynamics” (Forrester, 1961), and with it a new modeling paradigm for production and inventory systems. In this view, the defining factor causing the oscillatory dynamics typical of the bullwhip effect is the interaction of inventory policies and delays.

Delays induce the bullwhip effect because they disconnect decisions from the observation of their results. When there is a delay between making a decision and the observation of its result, then every subsequent decision must consider this delay and the associated uncertainty. Operationally, this implies keeping track of decisions taken but whose results are not yet observed and basing these decisions on expectations of future realizations. A typical example of a delay in a supply chain is the replenishment lead time; if the replenishment lead time is longer than the ordering frequency, then this delay forces the decision maker to keep track of multiple orders and to predict demand multiple periods ahead. Such delays complicate the analysis, noting that not all information in the supply chain can be observed and future demand is unknown.

Assuming there is a desired “end state” for a system, all decisions must be made as a function of this result. In practice, this implies that decision makers must decide, not only on desired inventory and production levels, but also on the path toward them. This is modeled through adjustment times; the desired amount of time that a decision maker wishes to spend chasing the desired state. Implicitly, this assumes a policy maker, and policy-related decisions. Such policies, the underlying behavior, and their effect on the bullwhip effect are explicitly modeled in the behavioral operations literature.

Behavioral Causes of the Bullwhip

Behavioral models of the bullwhip effect are commonly based upon the dynamic models introduced in [Forrester \(1961\)](#). Behavior is explicitly modeled through adjustment times—short adjustment times, implying “nervous” behavior, where quickly reaching the end state is prioritized at the expense of variability in orders/production; and long adjustment times as a way to model “relaxed” behavior, prioritizing stability over time to reach the desired state.

Behavioral models of the bullwhip effect are commonly investigated in experimental research, using the beer distribution game ([Croson and Donohue, 2006](#); [Sternan, 1989](#)). In this game four players assume different roles in a supply chain, from retailer to beer producer. In its basic form, no communication is allowed between players and an unknown consumer demand must be filled by the retailer, who can place orders to its supplier who, in turn, can place orders to its supplier, and so on, until the beer producer places manufacturing orders. The beer distribution game introduces information and material delays into the system, that is, it takes time for order information to travel upstream and it takes time for material (“beer”) to travel downstream. Analysis of data from the beer game shows that, in addition to the delays, behavioral attributes increase the magnitude of the bullwhip effect.

The main behavioral insight to come out from the beer game is that humans are not good at accounting for the pipeline (i.e., everything that has been ordered but not yet received). Compounded with long delays, this causes players to overreact to stockouts and order every period without taking into account outstanding orders. The effect of ignoring the pipeline is often seen as enormous oscillations; deep stockouts followed by equally large excess inventory. This effect is very robust and has been replicated innumerable times, in research as well as in the classroom ([Croson and Donohue, 2003](#); [Croson et al., 2014](#); [Narayanan and Moritz, 2015](#); [Moritz et al., 2019](#)).

Being based on the same paradigm as Forrester models, the underlying cause of the bullwhip effect in the beer game is the delay structure of the system. The insight from experimental research lies in identifying the behavioral biases present and how this irrational behavior further amplifies the bullwhip effect.

Rational Decision-Making Causes

Since the publication of the work by [Lee et al. \(1997a, 1997b\)](#), there has been significant interest in the theoretical mechanisms causing the bullwhip effect under fully rational decision makers. Under a periodic review inventory policy, for a firm calculating the optimal order quantity every period, the bullwhip effect appears due to four fundamental causes: demand signal processing, order batching, price fluctuations, and shortage games. These insights have been obtained using stylized analytical models.

- Demand signal processing: when a firm uses past demand information to forecast future demand, a nonstationary change in demand (e.g., a one-period surge or drop in demand) will trigger the bullwhip effect. In the absence of communication among firms, a supplier uses observed demand to update its forecasts. In such situations, nonstationary customer demands trigger changes in supplier forecasts that lead to increased variability due to a constant update of the optimal order quantity. The amplification of demand variability increases with the lead time, but it has been shown in an analytical model that amplification exists in this setting even when lead times are zero ([Lee et al., 1997a](#)).
- Order batching: if there is a nonzero ordering cost associated with placing an order, then order batching occurs naturally. If a supplier has a single customer, order batching implies that it will observe less frequent, larger orders, that is, increased variability. If a supplier has multiple customers, the impact of order batching depends on the correlation and timing of the individual demand streams. Under “perfectly synchronized” retailer ordering, that is, when a constant number of customers order every period, then the contribution of order batching to the bullwhip effect is avoided. In all other cases, order batching generates a bullwhip effect. Positively correlated ordering results in higher amplification than uncorrelated (i.e., random) ordering, which in turn results in higher amplification than balanced demands. Balanced demands are those, whereupon groups of customers place orders within a designated period every review cycle without any overlap (unlike perfectly balanced demands, the number of retailers placing orders in a given period is not constant).
- Price fluctuations: when a firm experiences variations in the purchase price, then it is optimal to adopt an ordering policy such that there exist different order-up-to levels, decreasing the purchase price. All else being equal, this results in increased variance of

orders and hence a bullwhip effect, compared to a policy with constant prices. As an example, consider a product with two prices, the regular price and a sale price. Unless the buying firm has full advance information on the sales periods, it follows logically that the optimal order-up-to level will increase whenever there is a sale and decrease when the price returns to normal.

- **Shortage gaming:** shortage gaming appears when a firm orders from a supplier that regularly suffers from stockouts. Facing a shortage, a supplier must allocate the available inventory among its customers. If such an allocation is (or is thought to be) proportional to the order size, customers have an incentive to inflate orders in an effort to obtain a larger share of the allocation. This sends a false demand signal that, interpreted by the supplier as a real increase in the demand, compounds the amplification problem (see demand signal processing). In addition, customers that inflate orders will typically cancel superfluous orders once the supplier catches up with demand; leading to a further increase in demand variability.

Measuring the Bullwhip Effect

The Bullwhip in Real Life—Anecdotal Evidence

Anecdotal evidence of the bullwhip effect abounds. In addition to the examples mentioned earlier, other famous cases are often used as case studies in business schools and the practitioner literature. Canned soup, dried pasta, and printer ink are some of the well-known examples of evidence of the bullwhip effect—products with stable consumer demand and highly variable upstream production. In recent decades, the bullwhip effect has become shorthand for the “inevitable” variability that upstream firms are required to contend with. Nevertheless, for all the usefulness of anecdotal evidence and case studies, rigorous research requires rigorous evidence of the existence of the bullwhip effect as a widespread phenomenon.

The Bullwhip in Real Life—Empirical Evidence

For all our understanding of the theoretical and behavioral causes of the bullwhip effect, extensive empirical research has been and is still conducted to show whether the bullwhip effect is an industry-wide problem, or if it only affects a limited number of individual firms. To this end, numerous researchers have attempted to rigorously measure it, using large panels of data. The evidence for the bullwhip effect in empirical data is not clear-cut (Bray and Mendelson, 2012; Cachon et al., 2007). There are a number of reasons that explain the nuances observed in the research.

First, empirical research, whether it uses primary or secondary data, typically uses aggregated data. It has been shown mathematically that measuring the bullwhip effect in aggregate data underestimates its impact (Chen and Lee, 2012). The particular degree of aggregation varies from study to study. Most empirical bullwhip effect studies use secondary data aggregated at the firm level with regards to products (i.e., all products are aggregated into a single stream), to inter-firm links (i.e., all customers/suppliers are aggregated into a single stream), as well as at the temporal level (observations are aggregated by month, quarter or even year) (Fransoo and Wouters, 2000).

Second, empirical research uses proxies to measure certain quantities. For example, sales data are used as a proxy of demand data and are combined with inventory data to estimate production data. In addition to the inherent errors that such estimations can introduce, the use of such data is conceptual at odds with the theory of the bullwhip effect. This is because the theoretical development of the bullwhip effect relies on the use of information flows in addition to material flows. As an example, the implicit assumption behind the modeling of the delay structure of a supply chain used in the beer game is that lead times are the result of a combination of the delay in information transmission (e.g., the time between placing an order and a supplier executing the necessary steps) and the material delay (e.g., the time it takes an order to be physically transported from a firm to its customer). Chen et al. (2016) show that using material flow data to estimate the bullwhip effect can over- or underestimate the real bullwhip effect, that is, the bullwhip effect as measured using information flow.

Another nuance that appears when measuring the bullwhip effect on empirical data is the stabilizing effect of demand seasonality. In real life, a significant number of products exhibit demand seasonality to a certain extent. For researchers estimating the bullwhip effect observed in the real data, the question is, then, whether this should be measured using the original demand stream or a deseasonalized demand stream (Cachon et al., 2007; Chen and Lee, 2012). From a theoretical perspective, including seasonality in the measurements can dampen the observed bullwhip; particularly if the variability of seasonality dominates the variability of the demand itself and, more so, if there are capacity constraints.

Of course, another source of ambiguity in the measured data is the fact that many firms actively seek to limit the bullwhip effect. Thus, in reality, there is a mix of bullwhip effect and production smoothing, and disentangling the one from the other is nontrivial.

Evidence from Macroeconomic Shocks

A number of studies took advantage of the 2008 credit crisis as a quasi-natural experiment, where the demand collapse in multiple industries approximated a nonstationary shock across entire industries. The evidence of an inventory shock is particularly strong. Such studies support the predictions regarding the transmission of the variability. In general upstream firms have observed more severe drops in demand than downstream firms. Moreover, recent studies claim that the inventory shocks by themselves,—that is, independent as a response to changing demand, but as a consequence of general financial or economic conditions—further amplify the bullwhip effect (Udenio et al., 2015).

Taming the Bullwhip

It is impossible to eliminate all structural and behavioral causes of the bullwhip effect from real life supply chains. Thus decision-makers can attempt to, at best, “tame” the bullwhip as much as possible. Ample research exists prescribing strategies to “tame” the bullwhip. Each of these strategies tackles one (or more) of the causes of the bullwhip effect highlighted earlier.

Delays in the supply chain structure amplify variability, thus minimizing delays is an obvious strategy to reduce amplification. Clearly, firms can only reduce delays by altering the structure of the supply chain in which they operate. All else being equal, a supply chain with short lead times will experience a smaller bullwhip effect than a supply chain with very long lead times. Firms willing to limit the appearance of the bullwhip are recommended, to the extent possible, to participate in agile supply chains; source (sell) from (to) suppliers (customers) with short lead-times. In practice, local supply chains are typically more agile than global supply chains. Note that the mode of transport used for deliveries affects lead times and thus the bullwhip effect. Thus such a strategic decision must consider the impact on variability amplification in addition to trade off between logistics costs, lead times, and environmental impact.

From a behavioral standpoint, the main recommendation is to avoid under-weighting the pipeline. While this recommendation directly addresses the main behavioral bias exhibited by human players of the beer game, research has shown that firms operating with policies that track “desired” inventory and pipeline targets (e.g., order-up-to policies) are also prone to under-weighting the pipeline (be it by design or through arbitrary adjustments). Furthermore, adopting fractional, rather than full, adjustments decreases the bullwhip effect generated by such policies. However, such fractional policies reduce the bullwhip at the expense of responsiveness to changes in demand (Disney et al., 2013).

Information sharing is often prescribed as a tool to counter the bullwhip effect because it tackles a number of structural issues. Under full information sharing, all firms in a supply chain can observe up-to-date inventory and/or demand information for all other firms in the supply chain. The visibility afforded by information sharing breaks the main mechanism behind the demand forecast updating; in this setting, firms can generate more reliable forecasts, for example by using downstream demand and inventory information to estimate future demand. This has been shown to be a particularly powerful action when combined with knowledge about the inventory policies that other firms employ. Directly sharing demand (and future order) forecasts among firms can similarly reduce the bullwhip effect.

In addition, information sharing can also mitigate the effect of shortage gaming, on the one hand by allowing suppliers to understand the “real” demand from customers (and thus avoid reacting to artificially inflated orders) and on the other hand by providing customers with information regarding the shortages that a supplier is facing. If a customer knows about the severity of a particular shortage, as well as the allocation mechanism and the expected recovery time, the incentive to artificially inflate orders to increase the allocation is greatly reduced. Full information sharing across a supply chain was previously considered all but unimplementable due to operational and technological barriers. Today, information sharing is starting to become technologically feasible. However, a strong barrier to its adoption is that, oftentimes, firms see demand and inventory information as a source of competitive advantage and thus classify it as sensitive and are not willing to share. Furthermore, the immediate benefits of information sharing are not symmetric. In fact, taming the bullwhip through information sharing requires downstream firms (the least affected by the bullwhip) to share information so that upstream firms reap the benefits. Downstream firms do not get comparable benefits from obtaining upstream demand/inventory information. Thus many downstream firms do not see this as a fair trade-off. However, this reasoning negates the second order benefits that downstream firms are able to obtain; if upstream companies achieve bullwhip reduction, then the decrease in inefficiencies and associated costs will sooner or later reach downstream firms in the form of lower costs, increased reliability and shorter lead times. Hence, information sharing typically needs to be associated with incentive alignment to be effective.

Other operational policies are designed to tame the bullwhip by attacking the root of each of the structural causes. For example, adopting “every day low prices” policies, or pricing contracts, to avoid the variability generated by price fluctuations; encouraging less-than-full truckload ordering via discounts and arrangements with 3PLs to discourage order batching; and adopting allocation rules that are independent of current orders (e.g., allocating based on past sales) to prevent rationing games.

Finally, there is the case for vertical integration. Information sharing is widely viewed as a way forward in regards to synchronizing the management of multiple firms and thus reducing the bullwhip effect. As discussed earlier, such implementations are relatively rare to find at large scale due to the view of information as competitive advantage. The vertical integration of firms enables what in practice is full information sharing of different stages of a supply chain, while at the same time avoiding competitive issues.

References

- Abramovitz, M., 1950. *Inventories and Business Cycles, with Special Reference to Manufacturer's Inventories*. National Bureau of Economic Research.
- Blinder, A., Maccini, L., 1991. Taking stock: a critical assessment of recent research on inventories. *J. Econ. Perspect.* 5 (1), 73–96.
- Bray, R.L., Mendelson, H., 2012. Information transmission and the bullwhip effect: An empirical investigation. *Manag. Sci.* 58 (5), 860–875.
- Cachon, P., Randall, T., Schmidt, G., 2007. In search of the bullwhip effect. *Manufact. Ser. Operat. Manag.* 9 (4), 457.
- Chen, L., Lee, H.L., 2012. Bullwhip effect measurement and its implications. *Operat. Res.* 60 (4), 771–784.
- Chen, L., Luo, W., Shang, K., 2016. Measuring the bullwhip effect: discrepancy and alignment between information and material flows. *Manufact. Ser. Operat. Manag.* 19 (1), 36–51.
- Croson, R., Donohue, K., 2003. Impact of pos data sharing on supply chain management: an experimental study. *Product. Operat. Manag.* 12 (1), 1–11.
- Croson, R., Donohue, K., 2006. Behavioral causes of the bullwhip effect and the observed value of inventory information. *Manag. Sci.* 52 (3), 323–336.

- Croson, R., Donohue, K., Katok, E., Sterman, J., 2014. Order stability in supply chains: coordination risk and the role of coordination stock. *Product. Operat. Manag.* 23 (2), 176–196.
- Disney, S.M., Hoshiko, L., Polley, L., Weigel, C., 2013. Removing bullwhip from Lexmark's toner operations. In: *Production and Operations Management Society Annual Conference*, pp. 3–6.
- Forrester, J.W., 1961. *Industrial dynamics*. MIT Press, Cambridge, Mass.
- Fransoo, J., Wouters, M., 2000. Measuring the bullwhip effect in the supply chain. *Supply Chain Manag.* 5 (2), 78–89.
- Kahn, J.A., 1992. Why is production more volatile than sales? theory and evidence on the stockout-avoidance motive for inventory-holding. *Quarterly J. Econ.* 107 (2), 481–510.
- Lee, H.L., Padmanabhan, V., Whang, S., 1997a. Information distortion in a supply chain: the bullwhip effect. *Manage. Sci.* 43 (4), 546–558.
- Lee, H.L., Padmanabhan, V., Whang, S., 1997b. The bullwhip effect in supply chains. *Sloan Manag. Rev.* 38, 93–102.
- Moritz, B., Narayanan, A., Parker, C., 2019. Unraveling behavioral ordering: relative costs and the bullwhip effect. Available from: SSRN <https://ssrn.com/abstract=3195282> or <http://dx.doi.org/10.2139/ssrn.3195282>.
- Narayanan, A., Moritz, B.B., 2015. Decision making and cognition in multi-echelon supply chains: an experimental study. *Product. Operat. Manag.* 24 (8), 1216–1234.
- Sterman, J., 1989. Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Manag. Sci.* 35 (3), 321–339.
- Udenio, M., Fransoo, J., Peels, R., 2015. Destocking, the bullwhip effect, and the credit crisis: empirical modeling of supply chain dynamics. *Int. J. Prod. Econ.* 160, 34–46.

Blockchain Applications in Logistics

Yingli Wang, Cardiff Business School, Cardiff University, Cardiff, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Glossary	136
Introduction	136
How Does It Work?	137
Types of Blockchain	138
Why Blockchain in Logistics	138
Blockchain Use Cases in Logistics	139
Product Provenance and Traceability	139
Process and Operations Improvement	140
Automation and Smart Contract	141
Trade Finance and Settlement	141
Anticorruption and Humanitarian Logistics	141
Conclusions	142
Further Reading	142

Glossary

A **block** is comprised of a group of transactions from the same time period, like a page from a record book.

A **blockchain** uses a special append-only data structure that are composed of transactions batched into blocks, which are cryptographically linked to each other to form a sequential, tamper-evident chain that determines the ordering of transactions in the system.

Consensus is the collaborative process that members of a distributed blockchain network use to agree that a transaction is valid and to keep the ledger consistently synchronized.

Merkle tree hashing is a computational function that ensures data blocks received from other peers in a distributed network are undamaged and unaltered.

A **miner** is a blockchain node that gets rewarded by some form of cryptocurrency or other incentives for creating blocks of validate transactions and including them in the blockchain.

A **hash** is a unique string of letters and numbers created from text using a mathematical formula. A hashing function turns any input (e.g., a password, or jpeg file) into a string of characters (i.e., a hash) that serves as a virtually unforgeable, unique, and encrypted digital fingerprint of the data.

A **node** is simply a computer on the blockchain network that stores the ledger.

A **timestamp** in a blockchain marks the time for each transaction on the blockchain proves when and what has happened on the blockchain and is tamper-proof.

Introduction

Widely recognized as one of the most disruptive technologies, blockchain technology has gained increasing popularity in recent years. Its core innovation lies in its ability to enable parties who are geographically distant or have no particular trust in each other to record, trace, monitor, and transact assets (tangible, intangible, or digital) without the need for third party verification. It therefore powers a range of innovative products, tools, and concepts such as cryptocurrency (e.g., Bitcoin or Ether), digital identity, and smart contacts and leads to new business and economic models.

Blockchain is a shared, distributed electronic ledger technology that can record transactions as they occur between parties in a secure and tamper-resistant way. Transactions in a blockchain are typically confirmed by all participants via a consensus mechanism and are not subject to any form of central control. Once validated and recorded in a blockchain, a transaction becomes permanent. An identical copy of the ledger is thus held by all users (known as nodes) on the network. No single party is able to delete or change a transaction unilaterally. Hence it is resistant to unauthorized change or malicious tampering, as participants in a blockchain network will immediately spot a change to one part of the ledger.

Blockchain and distributed ledger technology (DLT) are often used interchangeably. But strictly speaking, blockchain is an architectural subset of DLT. A critical difference between a distributed ledger and blockchain is that a distributed ledger does not have to consist of blocks of transactions to keep the ledger growing. The blockchain technology was invented by someone using the name

Satoshi Nakamoto in 2008 and is an innovative combination of a few existing technologies such as distributed ledgers, cryptographic protocols, and consensus mechanisms.

How Does It Work?

Let us assume that a transaction between two parties (from A to B) needs to be initiated. A transaction in this exchange could represent money, contracts, deeds, medical records, or any other asset that can be described in digital form. As shown in Fig. 1, Party A starts the transaction by first creating and then digitally signing it with its private key. The transaction is then propagated to peers (i.e., other nodes) of the blockchain network, which validate the transaction based on preset criteria. Once the transaction is validated by peers, it is combined together with transactions to create a new block of data for the ledger. The new block is then added to the existing blockchain and becomes part of the ledger. The transaction is complete and becomes permanent. In a blockchain, the blocks that contain batches of transactions are mathematically “chained” together in a chronological order (a term known as “time stamping”) using cryptographic techniques, hence the name “blockchain.”

The network participants in a blockchain use a consensus mechanism to eliminate the need to rely on central or third parties to validate transactions, and to ensure the single version of truth for all. Consensus mechanisms vary based on the type of blockchains

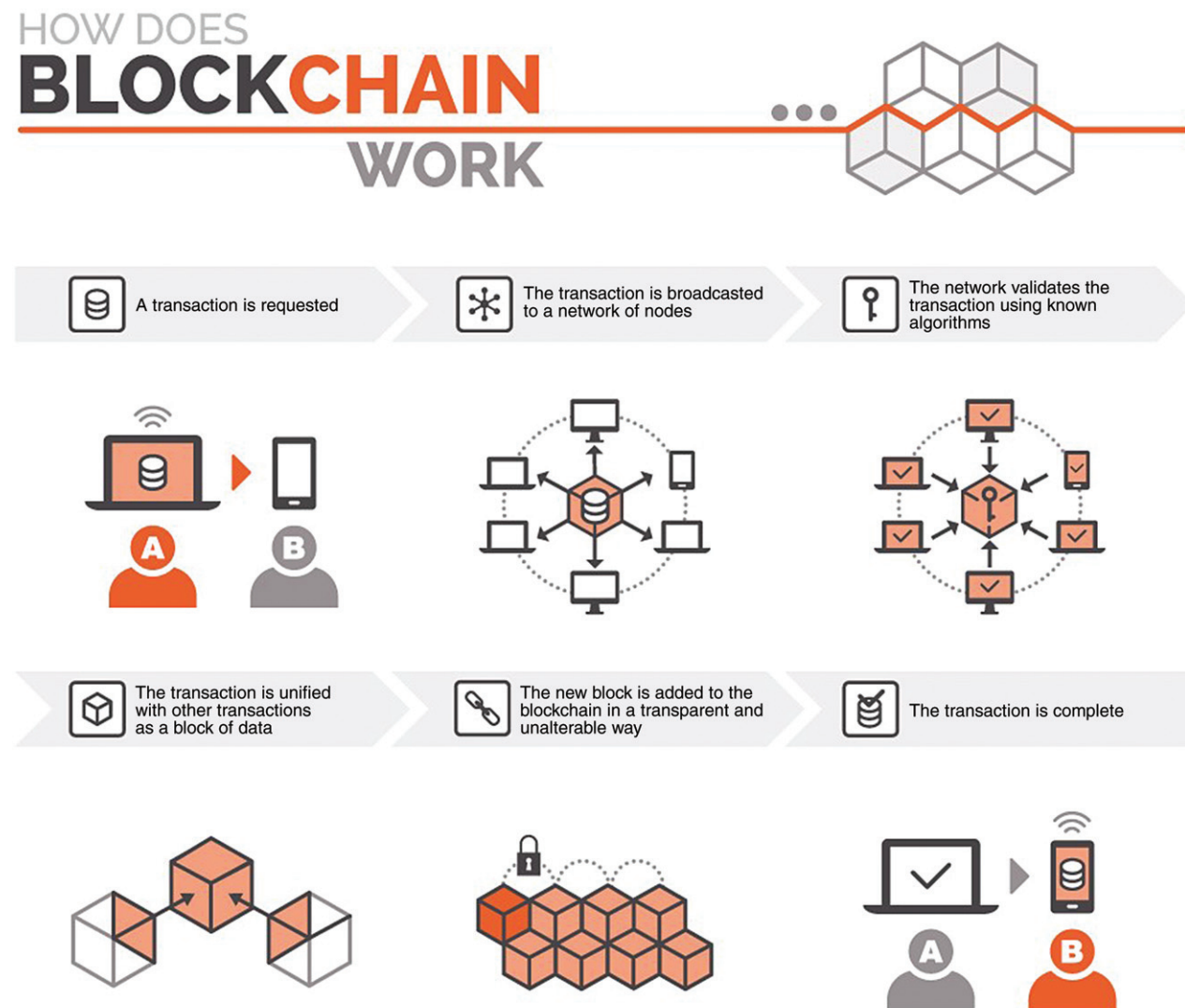


Figure 1 How blockchain works. Source: https://kilroyblockchain.com/skin/frontend/kilroy/default/images/how_does_blockchain_work.png, free to modify, share and use).

discussed earlier. For example, Proof of Work (PoW) is the most popular mechanism used in public blockchains, where miners must solve complex mathematical problems to verify transactions (a process known as mining). A reward is given to the first miner who finds the solution. Another popular mechanism is Proof of Stake (PoS), which tends to be deployed in a permissioned blockchain. PoS replace miners with validators where they will lock up some of their deposit (e.g., in terms of cryptocurrency) as a stake, and a group of validators takes turns proposing and voting on the next block and the weight of each validator's vote depends on the size of its stake. The validator collects network fees as the reward. Both PoW and PoS have their pros and cons. The former scales well, but is often criticized for its heavy time and computing power consumption, while the latter is more energy efficient but does not scale well. Currently, there are active efforts from industry and academia trying to find the right balance between scalability and performance and as a result a plethora of consensus mechanisms have been proposed, for instance Delegated Proof of Stake, Proof of Elapsed Time, Proof of Importance, and Proof of Activity, to name only a few.

Types of Blockchain

Due to its embryotic state, there is currently no agreement on blockchain categorization. A general approach is to distinguish the type of blockchain based on its access control, that is, who can read and submit transactions to a blockchain and participate within the consensus process.

If a blockchain is open to all, whereby transactions can be validated by anyone, then it is considered as a permissionless (sometimes also known as public) blockchain. Earlier versions of blockchain are permissionless; transactions are broadcast to every single participant (node); and every node thus keeps a complete record of the entire transaction history. Economic incentives are built in to encourage honest behavior, for example, rewarding miners with tokens. Bitcoin is a typical example of a permissionless blockchain.

Later, organizations realized that due to the sensitive nature of their data and regulatory concerns, sometimes a public blockchain was not a feasible option. As a result, permissioned (also known as private) blockchains (where only authorized participants can join) emerged to adapt to the needs of those organizations. Most blockchain technology networks observed so far in logistics, and supply chains are permissioned-based—where transactions and transaction-related data are only broadcast to selected parties (those involved in a specific trade to which these transactions relate). As a result, each node only holds records of which it is involved with. A few consortia-based initiatives in industry, such as the one led by Walmart and IBM on product provenance can be considered as permissioned blockchains because only related supply chain actors will be participating in this blockchain.

In practice, there is a broad spectrum of distributed ledger models, with different degrees of centralization and different types of access control to suit different purposes. Other terms such as hybrid blockchain (which allows participants to use permissioned networks that limit access to some data while also interacting with any public blockchain when needed) have also emerged. Over time, more blockchain variations have been added, such as public permissioned blockchain, federated, or consortium blockchains. The blockchain landscape is far from being settled.

Why Blockchain in Logistics

Therefore does blockchain matter to logistics and supply chains? Before offering a quick answer, we need to examine the unique attributes of blockchain technology and the core innovations that it entails.

- **Disintermediation:** blockchain's peer-to-peer network reduces reliance on any third party. Network participants can independently verify the integrity of the shared database and have a shared control over the entire database. Thus the integrity of data is guaranteed by a whole network, not by an intermediary. This opens up a number of opportunities for supply chain actors. For example, with blockchain, interfirm payments or digital assets transfer can occur without the authentication of a third party. Disintermediation will likely take place if the cost of existing supply chain intermediaries exceeds the value they add. New intermediaries that offer blockchain related products and service may emerge to facilitate blockchain's further diffusion along the supply chain.
- **Transparency with pseudonymity:** the information within a blockchain is visible to all participants and cannot be altered by a single entity, thus this improves auditability, creates trust, and reduces fraudulent behavior. Users can remain anonymous or reveal their identity to others. The extended transparency and traceability is seen as being of great value for industries that are sensitive to products and materials provenance. For instance, the blockchain could be used to trace the origin and production process of food ingredients, to detect fake products in a pharmaceutical supply chain, to assist the management of the whole product lifecycle in the automotive sector and support agile responses in the process of product recalls. Blockchain can be useful in supply chains where there is a need to minimize the degree of trust required between participants, or reliance on an intermediary service provider.

Given that the trust has been "programmed" into a blockchain through cryptography, relational investment is not as essential as in traditional supply chains (a costly, time consuming process). This may support the provision of a more agile type of logistics and supply chain operation, which is only configured temporarily to fulfil a particular demand and later could evolve and dissolve when conditions change.

- **Immutability:** once data are appended to a block and these blocks are collected in a chain, they cannot be changed or deleted by a single actor. Instead, they are verified and managed using governance protocols. This set up can effectively secure the data in blockchain ledgers against unauthorized access or manipulation. This is an important advancement, as any falsification of the information has to be done in real time, making it much harder than simply substituting new data. Therefore blockchain could be used as a registry and inventory system for recording, tracing, monitoring, and transacting tangible or intangible or digital assets. Caution though as absolute immutability does not exist. Reverse transaction, in theory, can be implemented if enough nodes (i.e., 51% or more) decide to collude.
- **Security:** cryptographic techniques, including cryptographic one-way hash functions, Merkle trees, and digital signatures, are deployed to safeguard the records in the database. Logistics and supply chain actors can encrypt and protect sensitive and commercial information and use business rules to control access by related parties, such as customs. Unlike a centralized system, if a system is hacked or there is some technical malfunction, the whole system may be brought to a halt, since the blockchain, by design, has no single point of attack and failure and can be more resilient to cyber hacks.
- **Automation:** blockchains can be programmed to automate a large number of business processes across different entities (e.g., make a payment) once certain conditions are met. Part of blockchain's novelty lies in its ability to set rules about a transaction (business logic) that are tied to the transaction itself. This contrasts with conventional databases in which rules are often set across the entire database level or in the application, but not in the transaction. This function supports the much discussed concept of a smart contract, whereby a smart contract is a computerized transaction protocol that automatically executes the terms of a contract upon a blockchain. Smart contracts are believed to be able to satisfy common contractual conditions, while reducing the costs and delays associated with traditional contracts. For example, a smart contract might send a payment to a supplier as soon as a shipment is received by the buyer. A smart contract, thus, eliminates payment withheld issues and improves efficiency by eliminating contract registration, monitoring, and updating efforts and time.

Blockchain Use Cases in Logistics

As blockchain technology matures, there are active efforts from almost every industry, as well as public and third sectors, in exploring how to capture the new opportunities afforded by blockchain. The following outlines some of the most popular use cases observed in the area of logistics and supply chain management.

Product Provenance and Traceability

A blockchain system offers extended visibility among multiple supply chain actors of a (digitized) product's digital footprint, from manufacturing to distribution and sale, thus creating so-called "see-through" supply chains. The immutability afforded by a blockchain system enhances data integrity and leads to increased confidence from customers of a product's legitimacy. Moreover, the use of timestamping (the process of providing a temporal order among sets of events) in the blockchain can prove the existence of certain data at a point of time, avoiding potential conflicts and disputes between parties over time-sensitive issues. Information completeness can be enhanced as well, as blockchain can accommodate a wide range of data, including ownership, location, product specification, and cost.

There is currently a strong industrial appetite in using blockchain for product provenance, in order to respond to the increasing demand from customers for the proof of legitimacy and authenticity of the products they purchase. Products range from luxury items (e.g., diamonds, watches), pharmaceutical products, aerospace, and automotive parts to organic and fair-trade food products. Therefore it is not surprising to observe that crucial supply chains are currently paving the way of blockchain deployment, where offering extended visibility is seen as a strategic value adding activity to multiple stakeholders.

Product provenance and enhanced traceability do not only deliver commercial benefits but are also of critical importance for product safety and quality assurance. It allows companies a quick response to product recalls and other safety-related issues. For instance, Walmart announced in September 2018 that it required all of its leafy green vegetable suppliers to submit and upload data to its blockchain tracking system by end of September 2019. This initiative is a direct response to the two Romaine Lettuce E. Coli Outbreaks in the United States in 2018, which caused over 200 people to fall sick and left 5 dead. Blockchains can also play an important role in stolen merchandise recovery, and in detecting counterfeit products and fraudulent transactions.

There are also emerging efforts in the blockchain's convergence with other technologies, such as Internet of Things (IoT), for real-time product tracking and whole life cycle asset management. This has important implications for asset-intensive industries, such as power, oil and gas, telecommunications, automotive, transportation, and construction, where asset management is critical to risk management and operational efficiency. Life cycle asset management also powers the concept of the circular economy and enables the reuse and recycling of materials (e.g., steel, batteries) back to the supply chain. The integrative use of artificial intelligence with blockchain for autonomous analysis and decision making is another area that attracts increasing attention.

Process and Operations Improvement

Implementing blockchains could improve efficiency in logistics and supply chains because the technology streamlines the data transfer between parties and thus reduces the time products spend in the transit process, improves inventory management, and ultimately reduces waste and cost. In a supply chain ecosystem, there are numerous bilateral transactions between organizations, creating many data silos and multiple versions of “truth.” Processes tend to be fragmented and, in some cases, largely manual and complex.

A blockchain platform can help to ease the current heavy workload on information transfer and processing, and bring multiple stakeholders together by the digitalization of document transfers and the acceleration of the flow of data. Fig. 2 shows the current state (A) and future (B) blockchain enabled information exchange process in a cross-border logistics and supply chain context. In a blockchain enabled cross-border system, data and digitized critical documents, such as a Bill of Lading, are appended by related parties into the blockchain along the progress. Customs clearance has been seen as a major bottleneck in a cross-border supply chain. The system will provide real-time visibility into the cargo status, therefore significantly improving the information available for customs authorities to inspect cargo, make decisions, and conduct risk analysis. The real-time end-to-end visibility also affords shippers, consignees, freight forwarders, port operators, and carriers greater predictability, earlier notification of issues and full transparency to validate fees and surcharges.

International payment requires a series of intermediaries for both clearing and settlement, and is complicated and expensive. A blockchain system can facilitate the business-to-business settlement at a fraction of the cost and time of traditional correspondent banking. A Deloitte report in 2016 has claimed that cross-border payment on blockchain has resulted in a reduction of transaction time from 2 to 3 days to a few seconds, and a 40%–80% reduction in transaction cost. Therefore cost savings can be substantial for organizations that make frequent international transactions. The added transparency and security can eliminate the risk of discrepancies in record keeping and safeguard the cash flows in a global supply chain. For cross-border settlement between two

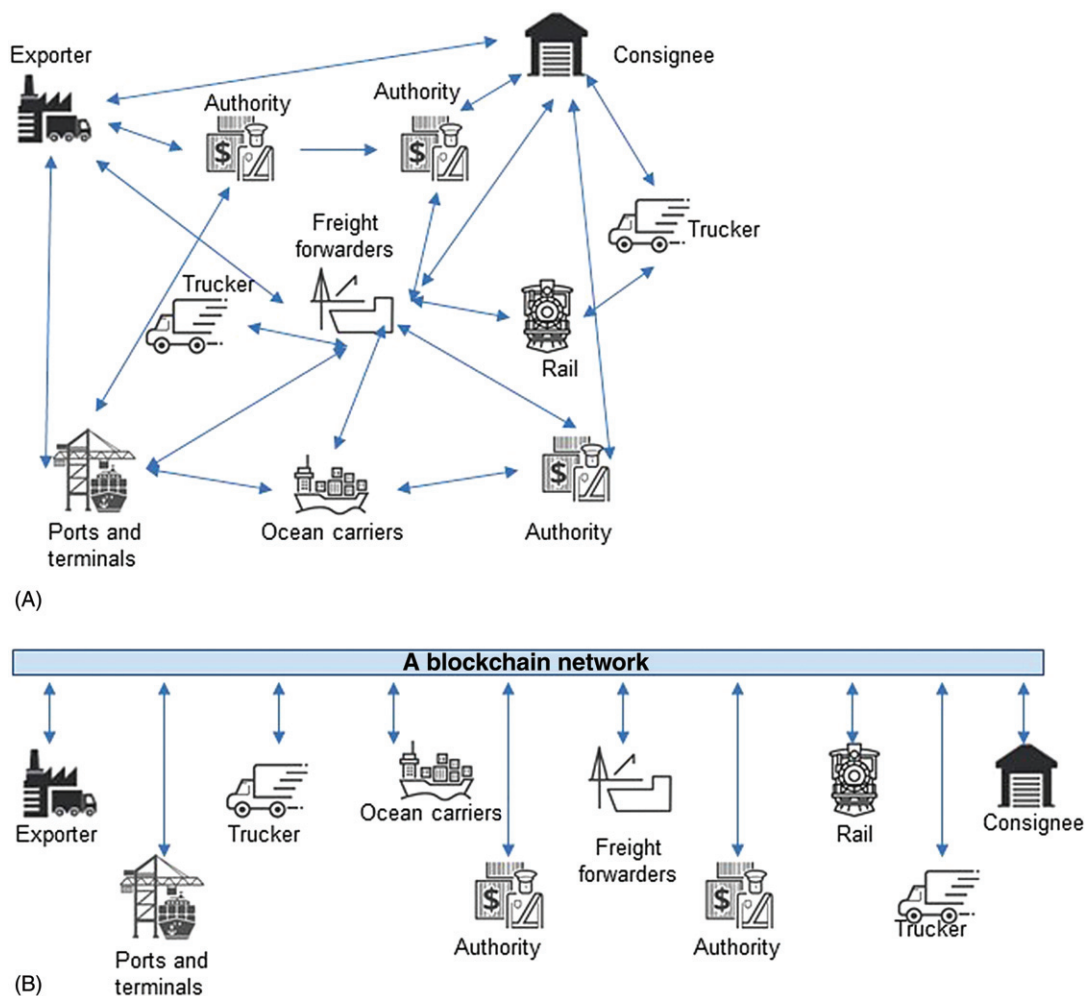


Figure 2 (A) Current state: cross border logistics and supply chain information exchange. (B) Future state: a blockchain enabled cross border logistics and supply chain information exchange. Source: Yingli Wang.

financial institutions, it would normally require that two parties agree together to use a stable coin (a cryptocurrency which is pegged to an asset with a stable value, such as gold or fiat money), central bank digital currency or other digital asset as the bridge asset between any two fiat currencies. The digital asset facilitates the trade and supplies of important settlement instructions. Cross-border payment is an important area where blockchain solutions have been developed and deployed commercially.

Automation and Smart Contract

Current operations, processes, and data exchange in logistics and supply chains are often manual, slow, and error-prone. With smart contracts, blockchain technology allows for increased automation and efficiency through avoiding the rekeying of data, speeding up of transactions and reduction of errors. In the blockchain context, a smart contract is a computer code running on top of a blockchain, containing a set of rules under which the parties to that smart contract agree to interact with each other. If and when the predefined rules are met, the agreement is automatically enforced. The smart contract code facilitates, verifies, and enforces the negotiation or performance of an agreement or transaction.

Smart contracts can potentially help with the cascading of purchasing orders, invoices, change orders, receipts, ship notifications, other trade-related documents, and inventory data across a supply chain. The automation of contractual terms in traditional database architectures is known as “stored procedures”. The key differences of running a smart contract in a blockchain context lies in the fact that smart contracts are guaranteed by system rules and the outcome is verifiable and auditable by all network participants. For example, a smart contract might send a payment to a supplier as soon as a shipment is received by the buyer. A GPS-tracked product return could log its location in real time and trigger a signal within the blockchain for immediate refunds.

Blockchain may also be utilized to support machine-to-machine payment in the IoT context. For instance, IOTA is currently developing an alternative DLT application named Tangle that is designed specifically for micro transactions between IoTs. A potential use scenario is that autonomous vehicles—such as forklifts, trucks, or any industrial machine—could pay for their own energy or fuel, maintenance, and insurance. Taking an example of a smart charging station for an electric vehicle, an electric vehicle plugs itself into a charging station. Both the vehicle and the charging station are equipped with their own digital wallet. A smart meter tracks usage for calculating payment. Payments (e.g., using a cryptocurrency) between the two wallets take place automatically after the vehicle is fully charged. The same concept could be applied to toll and parking payments. A smart contract, thus, eliminates payment withheld issues and improves efficiency by eliminating contract registration, monitoring, and updating efforts and time.

Trade Finance and Settlement

Blockchain used in trade finance mainly focuses on removing inefficiencies from existing processes. For example, blockchain can be used for faster credit risk assessment, minimized human errors in documentation checks, instant verification and reconciliation of records, automatic execution of workflow steps via smart contract and instant and secure exchange of data.

Crowdsourcing is another area of application. A blockchain-enabled finance platform could help to widen access to capital. This will be of particular benefit to micro-, small-, and medium-sized businesses (SMEs), which tend to have more difficulties getting loans from banks, due to their limited credit ratings. A blockchain platform would allow a micro or SME organization to source from a wider group of stakeholders, for example from a large number of investors with very small amounts of capital. Therefore, access and entry requirements are democratized.

Another application is the tokenization of existing assets (e.g., physical items) in supply chains. Tokenization refers to the process of digitally representing an existing, off-chain asset on a distributed ledger. In a blockchain network, cryptotokens can be created to denote custody or ownership of the bearer. By tokenizing the trade asset, its transfer between transaction participants on the blockchain network mirrors the movement of the physical asset, tracking the change of ownership and supporting product the traceability and provenance discussed earlier.

Finally, a token could be a native currency in a permissioned, blockchain-enabled supply chain system to facilitate fast and secure financial transactions between community members. This could reduce the high transaction cost associated with traditional finance and may help to reduce currency exchange volatility for cross-border transactions (see previous discussion on international payment in the Process and Operations Improvement section). In this case, tokens tend to link to national currencies. As a result, a price can be established for these tokens, and they become digital assets that have a certain value.

Anticorruption and Humanitarian Logistics

In a blockchain, unethical or opportunistic behaviors are made visible to all participants. This level of transparency can be of great value to traditional supply chains, such as in pharmaceuticals, where dominant supply chain actors may manipulate the market to inflate product prices; or in coffee supply chains where a fairer payment to farmers can be made visible to relevant stakeholders. Similarly, a blockchain system could help to expose and eliminate the corruption that is witnessed in certain public–private interactions. In a similar vein, blockchain has been deployed in humanitarian supply chains to ensure that financial or other emergency aid reaches the target beneficiaries.

For example, the World Food Programme (WFP)’s Building Blocks pilot project uses an ethereum-based blockchain technology to make the delivery of funds more efficient and secure to help refugees of the Syrian Civil War. In the Azraq refugee camp in Jordan,

10,000 people receive food from entitlements recorded on a blockchain-based computing platform. Refugees purchase food from local supermarkets in the camp by using an eye scan instead of cash, vouchers, or ecards. The blockchain allows the WFP to account for balances without the need for a third-party financial service provider to act as an intermediary. The major benefits include the large reduction of payment to financial service firms, increased privacy for beneficiaries and quicker reconciliation of accounts.

Conclusions

Despite the great potential of blockchain for logistics and supply chain applications, its deployment is not without challenges. Blockchain deployment implies significant technological (e.g., security), legal (enforcement mechanism when disputes arise), and societal (e.g., removal of intermediaries and potential job losses) implications. The adoption of blockchain in logistics and supply chains imposes some fundamental changes in relational, structural, process, and technological configurations between multiple parties in the ecosystem. There are a number of issues that have been raised in practice concerning its further diffusion:

- Confidence and related necessity issues—block chain's infancy state, lack of demonstrable benefits at scale, lack of expertise and knowledge in organizations;
- Cultural, governance, and collaboration issues—traditional mind-set, complex multiple-stakeholder involvement, conflicts in interests and objectives, and lack of clarity on roles and responsibilities;
- Technical issues—data integrity, interoperability and standards, data ownership, information sharing and scalability, and performance trade-offs; and
- Cost, privacy, legal, and security issues—relatively high deployment cost, commercially sensitive information and privacy, regulatory uncertainties.

Despite the aforementioned challenges, it can be observed that blockchain developments have gradually shifted from the pilot and proof of concept stage to larger scale deployments. Most efforts in logistics and supply chain currently concentrate on the building of permissioned, purpose-built platforms. Lessons and best practices learnt from those early efforts indicate that for any blockchain projects to succeed, the starting point is to identify areas where there are pertinent problems in an existing supply chain that blockchain technology would help address. This requires a clear purpose and business case to be established. The core value of a blockchain system is in its shared, append-only, distributed ledger across a network of organizations. If the problem is organizational rather than interorganizational, the blockchain may not be the right technological solution. However, that said, blockchain can be used as an enterprise solution for large organizations that have multiple divisions across the globe, where it can facilitate interdivision transactions within the organization.

Once a business case is identified, the next stage is to build a blockchain ecosystem by deciding on participating organizations. In practice, a small group of consortia tends to be set up and act as the founding members of a blockchain ecosystem. This small community tends to be led by an influential supply chain actor (e.g., a large retailer, a shipping line or a port operator), which has sufficient power to pull and push other community members toward the execution of a project goal. The blockchain platform value needs to be clearly articulated, which will then dictate its core activities and processes. The value proposition also largely influences the choice of application solutions, for instance whether to build in house or purchase propriety solutions or deploy blockchain as a service model. It is critical to set up a proper governance and legal framework within the community that defines the ways by which the participants access or exit a blockchain network; the types of roles each member will play; and the ways to resolve potential disputes among members. Once the pilot is completed, the organizations can then start to explore how to scale up.

There will not be a one-size-fits-all approach when it comes to blockchain deployment. Blockchain technology can be utilized along, or better still, combined with other digital technologies to offer innovative products and services that otherwise would not be available in the marketplace. Given the rapid changing landscape of blockchain and DLT, both industry and academia need to keep abreast of its development, understand the potential areas where it brings value, and prepare for its disruptive effect.

Further Reading

- Hewett, N., Lehmacher, W., Wang, Y., 2019. Inclusive Deployment of Blockchain for Supply Chains: Part 1 – Introduction, World Economic Forum. Available from: <https://www.weforum.org/whitepapers/inclusive-deployment-of-blockchain-for-supply-chains-part-1-introduction>.
- Nascimento, S., Polvora, A., Lourenco, J.S., 2018. Blockchain4EU: Blockchain for Industrial Transformations, European Commission, Brussels. Available from: <https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/blockchain4eu-blockchain-industrial-transformations>.
- UK Government Office for Science, 2016. Distributed Ledger Technology: Beyond Block Chain, UK Government Office for Science, London. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/492972/gs-16-1-distributed-ledger-technology.pdf.
- Wang, Y., Han, J.H., Beynon-Davies, P., 2019a. Understanding blockchain technology for future supply chains: a systematic literature review and research agenda. *Supply Chain Manag. Int. J.* 24 (1), 62–84., doi:10.1108/SCM-03-2018-0148.
- Wang, Y., Singgih, M., Wang, J., Rit, M., 2019b. Making sense of blockchain technology: How will it transform supply chains? *Int. J. Prod. Econ.* 211, 221–236, doi:10.1016/j.jipe.2019.02.002.

Logistics in Asia

Shong-lee Ivan Su, Department of Business Administration, School of Business, Soochow University, Taipei, Taiwan

© 2021 Elsevier Ltd. All rights reserved.

Introduction	143
Asian Logistics Within the Global Economy	144
Evolution	144
Research Findings	145
Asian Logistics Performance	145
Asian Logistics Industry	146
Third-Party Logistics Development	147
Research Findings	147
The Future	148
References	148
Further Reading	149

Introduction

Asia is the Earth's largest and most populous continent, covering an area of 44,579,000 square kilometers (17,212,000 square miles), about 30% of Earth's total land area and constituting 4.5 billion people (as of September 2018), roughly 60% of the world's population. There are more than 50 political entities in Asia, most of which are members of the United Nations.

Asia has the largest economy of any continent, both by nominal gross domestic product (GDP) and PPP, and is the fastest growing economic region. As in all regions of the world, the wealth of Asia differs widely between, and within, states. This is due to its vast size, meaning a huge range of different cultures, environments, historical ties, and government systems are represented. The largest economies in Asia in terms of nominal GDP are China, Japan, India, South Korea, Indonesia, Philippines, Thailand, Taiwan, United Arab Emirates, Iran, Malaysia, and Singapore.

As is shown in **Table 1**, there exists a disproportionate relationship among GDP, population size, and area coverage for subregions in Asia. The top 16 economies in terms of GDP, representing about one-third of Asia's economies (16/48), create more than 90% of Asia's GDP, revealing that only one-third of Asian economies dominate the wealth in Asia. Though not the largest subregion in terms of population and area, East Asia, with China and Japan as the representative economic powerhouses, is adjacent to the Pacific Ocean and has a major trade interaction with North America. East Asia is also the largest economic power in Asia, holding over 65% of the total GDP in Asia. In contrast, Central Asia falls within a landlocked area with the least population size and the poorest economic performance in Asia, with less than a 1% share in Asia's GDP. Thus, in the last 20 years, East Asia has been the most vibrant and innovative Asian subregion involved in global economic and logistics development, while Central Asia has seen the opposite.

In recent years, Southeast Asia and South Asia (mainly India) have experienced strong economic growth due to a more open trade environment, low cost production factors, and the trade war between the United States and China. These regions have attracted a great deal of foreign direct investment (FDI) from global firms wishing to build their new manufacturing base. Furthermore, a burgeoning middle class in these regions has created greater demand for consumer goods of all kinds which, in turn, creates greater demand for not only trade logistics but also domestic logistics and distribution within the region. This trend has also altered the inter- and intra-regional trade logistics landscape in recent years.

There have been wars in different parts of Asia due to religious conflicts and the fight against terrorism. In particular, there have been civil wars in West Asia and South Asia. These wars have increased in intensity since the US 911 terrorist event in 2001. Therefore, the ongoing military actions in these two Asian subregions have raised a great demand for military logistics and weapon shipments. However, the wars and military actions have also created many refugees that are fleeing away from the war zones and looking for shelter in neighboring countries or even distant promised lands without war zones. Thus, there has been a great need for humanitarian logistics implemented, in particular, by the United Nations and international relief organizations. Increasingly devastating natural disasters, largely linked to climate change, such as the 311 earthquake and the tsunami in Japan causing the Fukushima Nuclear Plant Disaster in 2011 have brought great damage to life and economies, as well as tremendous challenges for disaster relief operations and logistics.

In brief, it is a highly challenging task to write an article about logistics in Asia since Asia is such a diversified and fragmented region with so many different geopolitical, economic, and natural influences on the demand for different logistics solutions for many varied stakeholders, with sometimes similar, but sometimes conflicting, interests. Thus, this article will focus solely on providing an overview of the economic and trade aspects of Asian logistics development.

Table 1 A regional view of Asia's scale in the global economy

Region	Economy count	GDP nominal (2018)		Population (2016)		Area coverage	
		\$ M	%	Size	%	km ²	%
Asia ^a	48	29,734,672	100.00%	4,535,326,831	100.00%	48,261,992	100.00%
Top 16	16	27,601,551	92.83%	NA	NA	NA	NA
Central Asia	5	279,659	0.94%	69,787,760	1.54%	4,003,451	8.30%
East Asia	8	19,364,810	65.13%	1,642,093,255	36.21%	11,797,156	24.44%
North Asia	1	1,469,341	4.94%	143,964,513	3.17%	17,098,242	35.43%
South Asia	8	3,321,964	11.17%	1,783,888,730	39.33%	5,215,729	10.81%
Southeast Asia	11	2,722,560	9.16%	641,775,797	14.15%	4,537,945	9.40%
West Asia	15	2,576,338	8.66%	253,816,776	5.60%	5,609,469	11.62%

^a48 economies excluding Cyprus, Egypt, Georgia, Palestine, Turkey, North Korea, and Syria which do not have GDP statistics.

Source: Wikipedia/economy of Asia, 2019/2/9 downloaded with author's compilation.

Asian Logistics Within the Global Economy

Asia differs very widely among and within its subregions with regard to ethnic groups, cultures, environments, economics, historical ties, and government systems. With approximately 4.5 billion people, Asia hosts 60% of the world's population and continues to experience high growth rates in both population and economy. Over the last 3 decades, Asia's economic development and integration into the global economy has, arguably, been the most intriguing commercial story on the planet. However, due to its diversity and associated complexity, Asia continues to be the most disconnected region in the world. Therefore, it is fascinating as well as challenging to uncover the critical issues and theories relating to Asian logistics development.

Evolution

The integration of national economies into a global economic system has been one of the most important developments of the last century. This process of integration, often called "globalization," has materialized in the form of a remarkable growth in trade between countries. Exports today are more than 4000 times larger than in 1913. Up to 1870, the sum of worldwide exports accounted for less than 10% of global output. Today, the value of exported goods around the world is more than 25%. This shows that over the last hundred years of economic growth, there has been an exponential growth in global trade. This development has also created huge demands on both goods and service logistics flows either globally, regionally, or domestically.

Asia first emerged as a manufacturing power in the 1960s, when Japan began exporting electronics and consumer goods. Taiwan and South Korea followed its lead. By the 1980s, Japanese firms were building plants across Southeast Asia. But China's opening up was the game changer. In 1990, Asia accounted for 26.5% of global manufacturing output. By 2018, this had surpassed 50%. China accounts for half of Asia's output today. The region's share of the global trade in intermediate inputs—the goods that are eventually pieced together into final products—rose from 14% in 2000 to more than 50% in 2018.

Trade transactions include both goods (tangible products that are physically shipped) and services (intangible commodities, such as tourism and financial services). The production chains for these goods and services are becoming increasingly complex and global. In today's global economic system, countries exchange not only final products, but also intermediate inputs. This creates an intricate network of economic interactions that cover the whole world. According to recent estimates, about 30% of the value of global exports come from foreign inputs. Under this trend, intra-industry trade, that is, the exchange of similar products belonging to the same industry, has grown substantially to alter the traditional export–import logistics flow between two countries under inter-industry trade to an intricately intertwined global supply chain network. Recent years have seen a trend in international trade called splitting up the value chain. The value chain describes how a good is produced in stages in multiple countries. The production and distribution of the iPhone, for example, involves the design and engineering of the phone in the United States, parts supplied from Korea, Japan, Taiwan, and many others, the assembly of the parts in China, and the advertising and marketing done in the United States and many consumption countries.

The success of Factory Asia over the last 2 decades is a sign that Asian countries have been able to put business ties above political disputes. Commerce has brought them closer together. ASEAN nations have a free-trade agreement, with China. Japan, China, and South Korea negotiate a deal among themselves. There are also talks, still early, about a broader pact that would tie all countries in the region together, including India. Countries with lots of cheap workers should produce labor-intensive goods; rich countries should focus on those requiring plenty of capital. Richard Baldwin, an economist, argues that simply comparing national advantages is outdated. As supply chains spread across borders, regional comparative advantage matters even more. With its bounty of both labor and capital, Asia has already built up a huge lead in manufacturing. It only stands to grow.

Since 2018, world trade has run into a strong headwind due to trade imbalances and the resulting economic power shift between the West and the East causing the trade war between the world's two largest economies, the United States and China. This will

certainly affect how global firms adjust their supply chains. However, it will not deter the development of already highly connected and mature global value chain networks.

Research Findings

A 2007 study identified an emerging global logistics model that has now become a mature model for high value, lightweight, and non-dangerous global products (Su, 2007). Due to the heat-up of competition among major notebook computer brands in the global consumer market, these brand name companies such as HP, Apple, Lenovo, and Acer have started to tighten up their control of their supply chains and raised the production and service delivery requirements to their original design manufacturers (ODMs) and third-party logistics (3PL) service providers. To cope with these stricter requirements, ODMs started to work with their brand name customers and global logistics firms on developing a Just-in-Time (JIT) supply chain that would not only meet the global distribution requirements, but also surpass the expectations of the end customers or consumers. This JIT supply chain is defined as the Global Direct Distribution Business Model (abbreviated as GDD), which combines and synchronizes the fast production cycle and the fast delivery cycle of global products similar to that of notebook computers. With GDD, it is possible for global marketers to centralize production or warehousing in the global supply chain and distribute products to many destinations in the world, as if they are distributing their products domestically without using any trade middleman. This model has now become a critical facilitator for the fast growing cross-border e-commerce businesses in Asia.

In 2015, the *International Journal of Physical Distribution & Logistics Management* published a special issue on "Contemporary Strategic Supply Chain Management and Logistics Issues in Asia" (Su, 2015). A total of 43 papers were submitted. The most prevalent topics are supply chain collaboration/integration and green supply chain strategic issues in Asia. Asian supply chain risk and 3PL are also popular topics. The following topics were also submitted: sustainability, e-commerce, service orientation, and logistics innovation. One intriguing paper examined the very unique topic of logistics corruption in Asia. For research methods, nearly half the papers submitted employed survey-based empirical research methods such as regression analysis, structural equation modeling (SEM), and factor analysis. Case-based and optimization-based methods were also both well represented. This special issue showcases the diverse strategic supply chain management (SCM) and logistics research interests in Asia, as well as the highly diverse nature of Asian economies. Two representative research findings will be presented next.

Ke et al. (2015) used annual US trade statistics and manufacturing industry data for the years 2002–9 between the United States and its top 12 Asian trading partners to examine key factors associated with the international transport modal decision. The study findings indicate that manufacturing industries use more airfreight and less ocean freight when facing positive sales surprises, high monthly demand variation, a high contribution margin ratio, a high cost of capital, and increased competition. While transportation cost remains an important concern, a logistics manager must also consider non-cost factors such as competition, working capital, and demand uncertainties in their modal decisions.

Pujawan et al. (2015) presented a new method for integrating operational and strategic decision parameters, namely, shipment planning and storage capacity decisions under uncertainty for the maritime logistics of bulk shipments of commodity items. The results suggest that the number of ships deployed, silo capacity, working hours of ports, and the dispatching rules of ships significantly affect both total costs and service levels. Interestingly, the study findings indicate that operating few ships enables firms to achieve almost the same service levels while gaining substantial cost savings if constraints on other parts of the system are alleviated, that is, storage capacities and working hours of ports are extended.

A longitudinal single case supply chain study recently published (Cui et al., 2019) explores a global heavy vehicle manufacturer's servitization process in the Chinese trucking user market with many 3PL firms as its customers. The key research finding reveals that the decision-makers adjust their decision-making logic depending on the stage of the servitization process and associated risk patterns. As the servitization process evolves into a more sophisticated stage, decision-makers will change their decision-making logic from a causation dominant logic to an effectuation dominant logic, in order to cope with the increased risks. Effectuation theory originally developed from entrepreneurship research and is found to be a valid theory for the explanation of the risk and uncertainty control behaviors in the servitization transformation process of manufacturing firms.

Asian Logistics Performance

Logistics disparity among nations or economies has caused both hard and soft connectivity issues for the intra- and inter-Asia trade routes and subsequent time delay and additional logistics costs in global supply chains. The World Bank has recognized the critical role of national logistics performance in the effectiveness of world trade and also the gaps of logistics performance among nations. Thus, since 2007, the World Bank has initiated a biannual global survey on national logistics performance for most countries in the world and has developed a comprehensive *logistics performance index* (WB LPI) database, examining six logistics performance indices: *customs, infrastructure, international shipments, service quality, tracking and tracing, and timeliness*.

National logistics is not a stand-alone system, rather it is a complex interorganizational and, oftentimes, cross-country set of processes prone to degenerate if not managed on a continuous improvement basis. Even until now, for many countries, it is still a highly challenging job to measure and manage their national logistics performance due to its cross-ministry administrative nature and many other complexities involved in moving international goods across national borders. The WB LPI database is currently the only source that provides the national logistics performance measures for most countries in the world. Previous studies have found

Table 2 Asian subregional LPI comparison

Asia subregion	LPI average			LPI standard deviation		
	Score	Relative scale	Rank	Score	Relative scale	Rank
Central Asia (5)*	2.54	1.00	108	0.18	1.00	25
East Asia (6)*	3.52	1.39	38	0.59	2.87	46
East Asia-M (5)*	3.75	1.48	19	0.20	5.66	10
North Asia (1)*	2.76	1.09	75	0.00	1.43	0
South Asia (8)*	2.53	1.00	105	0.35	1.02	36
Southeast Asia (9)*	3.06	1.21	60	0.51	1.79	39
West Asia (14)*	2.88	1.14	74	0.49	1.45	42

(n)*, n is the number of the countries in the subregion; East Asia-M, East Asia excluding Mongolia; relative scale, the relative LPI scale of a country to the worst performer.

Source: Extracted and compiled from 2018 full LPI dataset from the World Bank (<https://lpi.worldbank.org/>)

Table 3 Top and bottom LPI performers in Asia subregion

Asia subregion	Top LPI performer		Bottom LPI performer		Gap
	Country	LPI rank	Country	LPI rank	
Central Asia (5)	Kazakhstan	71	Tajikistan	134	–63
East Asia (6)	Japan	5	Taiwan	27	–22
East Asia-M (5)	Japan	5	Mongolia	130	–125
North Asia (1)	Russia ^a	75	Russian	75	0
South Asia (8)	India	44	Afghanistan	160	–116
Southeast Asia (9)	Singapore	7	Myanmar	137	–130
West Asia (14)	UAE ^b	11	Iraq	147	–136

^aRussian Federation

^bUnited Arab Emirates

Source: Extracted and compiled from 2018 full LPI dataset from the World Bank (<https://lpi.worldbank.org/>)

that better LPI scores are associated with more trade and lower logistics costs, while a few studies have indicated that the relationships are nonlinear. One study even developed a process-based benchmarking method to facilitate the management and continuous improvement of the national logistics performance of a country or economy (Su and Ke, 2017).

Asia can be divided into seven subregions. Only Russia is a single-country region, while all other subregions are multiple-country regions ranging from 5 to 14 countries/economies in the WB LPI database. Table 2 clearly shows East Asia (without Mongolia) leads in Asia in terms of average LPI score and rank, with large relative gaps surpassing the lowest performers—Central Asia and South Asia, that is, 48% versus 5.66 times higher, respectively. Furthermore, other than North Asia with only a single-country Russia, Asian subregions internally also demonstrate logistics disparity, with different degrees of LPI score and rank standard deviations. Central Asia has both the lowest LPI averages and standard deviation, which implies that all countries in the subregion might suffer from poor logistics connectivity and stagnating economic development. Table 3 provides a view of the top and bottom LPI rank performers in each subregion and the gap between top and bottom performers to elaborate the findings from Table 2.

The extended LPI analysis results in Tables 2 and 3 reinforce the World Bank's claim that logistics disparity exists among nations/economies and has become a hurdle to global economic and trade development. The Asia Development Bank estimates that between 2016 and 2030, an investment of US \$8.4 trillion in transport infrastructure will be needed to address the shortfall of transportation connectivity in Asia. Countries have responded by upgrading aging and/or developing new transport infrastructure, as well as more ambitious initiatives like China's "One Belt, One Road" initiative. Relative to major hubs elsewhere in the world, Asia is under-provided with modern logistics space; only Shanghai, Tokyo, Hong Kong, and Singapore match the supply elsewhere—and even then, only the smaller global hubs. Asia—developing Asia, in particular—lags the industrialized world in terms of logistics facility supply and sophistication, presenting a significant opportunity for real-estate investors and developers.

Asian Logistics Industry

In the last 20 years, accompanying the rapid economic growth globally, particularly in Asia, the 3PL industry (i.e., the logistics outsourcing industry) has evolved in a revolutionary manner into a global service industry booming in both developed and also emerging economic blocks. There are several industry survey reports for the global status of 3PL development. One that has a long

Table 4 Global 3PL revenues

Region	2014 global 3PL revenues	2015 global 3PL revenues	2016 global 3PL revenues	2017 global 3PL revenues	Percent change 2015–16	Percent change 2016–17	Percent change 2017–18	CAGR 2010–17
Africa	29.6	27.1	25.5	26.1	−8.4	−5.9	2.4	1.30%
Asia Pacific	289.3	292.7	306.1	329.3	1.2	4.6	7.6	6.10%
CIS/Russia	33.7	23.5	21.7	25.5	−30.3	−7.7	17.5	−0.50%
Europe	196.4	172.6	173.4	184.1	−12.1	0.5	6.2	−0.40%
Middle East	45.3	40.3	40.5	42.2	−11	0.5	4.2	2.70%
North America	195.9	195.7	200.3	220	−0.1	2.4	9.8	4.60%
South America	45	37.9	36.7	41.8	−15.8	−3.2	13.9	1.10%
Grand total	835.2	789.8	804.2	869	−5.4	1.8	8.1	3.50%

Note: Revenue in US \$ billions.

Source: Langley and Infosys (2019).

history and recognition is the Annual Third-Party Logistics Study (Langley and Infosys, 2019). The most recent release of the survey analysis report is the 23rd Annual Report in 2019. It is worth extracting some of its findings for this article, particularly relating to Asian logistics.

Third-Party Logistics Development

Table 4 shows global 3PL revenues by region from 2014 to 2017, and the year-over-year (YOY) percentage increases or decreases. Global 3PL revenues increased to \$869 billion in 2017 from \$804.2 billion in 2016. This reflects a strong global growth in the 3PL market of +8.1% from 2016 to 2017. The last column in the table indicates the CAGR of global 3PL revenues was calculated at +3.5% from 2010 to 2017. This is a clear indication that the 3PL industry has been a growing industry over the last decade. Globally, the Asia Pacific region is the largest logistics outsourcing market, accounting for 38% of total global 3PL revenues; North America with 25% share comes next; and Europe with 21% is in third place. The three regions together control more than 80% of the world 3PL market.

To keep up with the increasing level of complexity within the supply chain, companies must be agile enough to meet rapidly changing environments. Shippers are continuing to leverage what 3PLs offer, allowing them to optimize the supply chain, minimize costs, and create value, as well as to align expectations as a key to achieving success for both parties. Logistics is being transformed through the power of data-driven insights, and current technology is enabling unprecedented amounts of data to be captured from various sources along the supply chain. The use of technology is exploding within every area of the supply chain, which is driving increased agility. Shippers are increasingly aware that if they do not have the technological capabilities to accomplish their goals, they should partner with those that do. As the amount of available data increases, shippers and their logistics partners will need to be able to take the information and make it relevant, as many 3PLs are already making significant investments in technology that allows them to analyze shippers' operations.

A report released by Armstrong & Associates, Inc. (2017) shows e-commerce and outsourcing are fueling demand for 3PL services, driving 3PL revenues to over US\$1.1 trillion by 2022. Recent projections by Armstrong & Associates indicate optimism for the growth of global 3PL revenues into the next 5 years. Among the factors contributing to the likelihood of continued growth are data-driven technology, increased capacity and demand, and overall rising prices.

Research Findings

Since the early 1990s, 3PL services have evolved into a highly innovative and competitive global service industry and, as such, have attracted great research interest from supply chain and logistics researchers. Few Asia-related 3PL studies can be found, even though a number of Asian 3PL studies are growing. Two highly recognized international journals have published two special issues regarding logistics development in the Greater China area: an immense and diverse geographic area that includes Mainland China, Hong Kong, Macau, and Taiwan. Over the last 3 decades, there has been a tremendous influx of financial and human capital that has contributed to the development of a complex and dynamic mix of domestic and global supply chains. Moreover, rising labor and rent costs, stricter environmental regulations, escalating trade conflicts, and regional political instabilities have dramatically impacted and changed supply chains in the region. Rapid changes make collecting empirical data for strategic SCM research studies in Greater China somewhat challenging for multiple reasons.

In 2009, the *Transportation Journal* (Vol. 48, No. 3) published a special issue to better understand the transportation and logistics-operating environment in the Greater China region. Five papers were selected for the special issue with only one relating to 3PL development. Cui et al. (2009) conducted a pioneering 3PL innovation study and compared three mid-size regional 3PL firms from Hong Kong, Sweden, and Taiwan. How the 3PL firms innovate illustrates the characteristics of different regions. Many firms enter China for sourcing and outsourcing. They have the need for cross-border operations and require customs-related services and international transport from innovative and highly motivated regional 3PL firms. Hong Kong is the logistics center in

East Asia and companies tend to choose Hong Kong as a distribution center for their global pipelines and regional 3PLs constantly develop new services related to distribution and operation. In contrast, Sweden is in a more developed region. Firms focus more on sophisticated end consumers in this region and 3PL providers concentrate on providing new services to their clients in order to meet end consumers' needs. This pioneering 3PL innovation study laid out a foundation for a 3PL innovation competence assessment study that developed a 3PL innovation competence assessment methodology based on the 3PL and innovation literature (Su et al., 2014).

In 2017, an *International Journal of Physical Distribution & Logistics Management* special issue (Vol. 47, Issue 9) on the strategic supply chain and logistics management in Greater China provides several valuable research findings on 3PL firms (Su, 2017). In general, the special issue recognized that the 3PL industry in Greater China has experienced rapid growth in recent years and demand for logistics services has grown tremendously. Meanwhile, 3PL firms in the region are facing intense competition in the logistics market, and such competition has forced many 3PLs to reconsider their competitive advantages from a resource-and-capability point of view. Nevertheless, logistics users' satisfaction with 3PLs in mainland China is lower than the level of satisfaction found in more developed countries and the range of logistics service provisions seems to be much narrower.

The four 3PL papers published in this special issue reveal the crucial status regarding the contemporary strategic development of 3PLs in the Greater China region. Interestingly, the four papers all use structural equation modeling as the key research method for their data analysis, showing the popularity of more advanced statistical methods among Greater China logistics and SCM scholars who conduct survey-based research. Three papers examine 3PL topics from the resource-based view (RBV) of the firm and the factors affecting horizontal alliances and firm performance, while the other paper analyses the views of 3PL customers regarding the relationships between dependence and integration, with trust as the mediating factor. Two papers in the special issue are selected to present their research findings in this article.

Shou et al. (2017) build on extant research to examine the effects of relational resources on firm performance in the Chinese logistics industry, surveying randomly selected 3PL providers across different regions in China. The results of this research confirm that relational resources have a positive effect on firm performance. However, such an effect is not direct; instead, it is realized through the mediation of the innovation capability. This finding corresponds to the core idea of the RBV, which argues that firms deploy strategic resources through capabilities in order to achieve competitive advantages and superior performance.

Gao et al. (2017) explores the impact of partner similarity on the success of horizontal alliances of 3PLs in China. Primary data were collected through questionnaire to 380 Chief Executive Officers and Managing Directors in 262 small- and medium-sized logistics enterprises in China. This research verifies that partner similarity and logistics alliance management capability are positively correlated with alliance stability and performance in horizontal alliances among Chinese 3PLs, especially where competence and cultural similarities are present. Moreover, alliance stability mediates the impacts of partner similarity and logistics alliance management capability on alliance performance.

The Future

As the Asian economy moves toward a new level of uncertainty and complexity in the global environment, the logistics services and support to Asian businesses and customers will definitely continue to transform to a more agile form, in order to cope with the changing and increasingly demanding service requirements from the markets. For logistics practitioners, ample growth and business renewal opportunities certainly exist in Asia and can be captured by those who are prepared. For logistics/supply chain scholars, many challenges need to be addressed regarding developing better educational programs to train the next generation of Asian logistics managers and analysts. The opportunity also exists for researching critical emerging topics to better understand Asian logistics development that is becoming more dynamic and complex, in order that organizations can formulate better logistics policies and strategies.

References

- Armstrong & Associates, Inc., 2017. Latest top 3PL lists. Available from: https://www.3plogistics.com/armstrong__associates_top_3pl_lists/.
- Cui, L., Su, S.I., Hertz, S., 2009. How do regional third party logistics firms innovate? A cross-regional study. *Transp. J.* 48 (3), 44–50.
- Cui, L., Su, S.I., Feng, Y., Hertz, S., 2019. Causal or effectual? Dynamics of decision making logics in servitization. *Industrial Marketing Management*. Available from: <https://doi.org/10.1016/j.indmarman.2019.03.013>.
- Gao, H.J., Yang, J.H., Yin, H.W., Ma, Z.C., 2017. The impact of partner similarity on alliance management capability, stability and performance: empirical evidence of horizontal logistics alliance in China. *Int. J. Phys. Distribution Logistics Manag.* 47 (9), 906–926.
- Ke, J.Y., Windle, R.J., Han, C.D., Britto, R., 2015. Aligning supply chain transportation strategy with industry characteristics: evidence from the US-Asia supply chain. *Int. J. Phys. Distribution Logistics Manag.* 45 (9/10), 837–860.
- Langley Jr., C.J., Infosys, 2019. 23rd annual third-party logistics study: the state of logistics outsourcing. Available from: https://dsqapj1lkrkc.cloudfront.net/media/sidebar_downloads/2019-3PL-Study.pdf.
- Pujawan, N., Arief, M.M., Tjahjono, B., Kritchanchai, D., 2015. An integrated shipment planning and storage capacity decision under uncertainty: a simulation study. *Int. J. Phys. Distribution Logistics Manag.* 45 (9/10), 913–937.
- Shou, Y.Y., Shao, J.N., Chen, A.L., 2017. Relational resources and performance of Chinese third-party logistics providers: the mediating role of innovation capability. *Int. J. Phys. Distribution Logistics Manag.* 47 (9), 864–883.
- Su, S.I., 2007. The emerging global direct distribution business model: industry and research opportunities. *Transp. J.* 46 (4), 58–65.

- Su, S.I., 2015. Guest editorial: contemporary strategic supply chain management (SCM) and logistics issues in Asia. *Int. J. Phys. Distribution Logistics Manag.* 45 (9/10), doi:10.1108/IJPDLM-05-2015-0135.
- Su, S.I., 2017. Guest editorial: strategic supply chain and logistics management in Greater China: evolution, innovation and future challenges. *Int. J. Phys. Distribution Logistics Manag.* 47 (9), 766–771.
- Su, S.I., Ke, J.F., 2017. National logistics performance benchmarking: a process-based approach using World Bank logistics performance index database. *J. Supply Chain Oper. Manag.* 15 (1), 55–78.
- Su, S.I., Ke, J.F., Cui, L., 2014. Assessing the innovation competence of a third-party logistics service provider: a survey approach. *J. Manag. Policy Pract.* 15 (4), 64–79.

Further Reading

- Arvis, J.F., Ojala, F., Wiederer, C., Shepherd, B., Raj, A., Dairabayeva, K., Kiiski, T., 2018. Connecting to Compete 2018 Trade Logistics in the Global Economy: The Logistics Performance Index and Its Indicators. The International Bank for Reconstruction and Development/The World Bank, Washington, DC.
- Su, S.I., Ke, J.F., Lim, P., 2011. The development of transportation and logistics in Asia—an overview. *Transp. J.* 50 (1), 124–136.

Logistics in the Developing World

Charles Kunaka, World Bank, Singapore, Singapore

© 2021 Elsevier Ltd. All rights reserved.

Introduction	150
Measuring the Performance Efficiency of Logistics Systems	150
Key Bottlenecks in Logistics	151
Improving Logistics Performance in Developing Countries	153
Understand the Demand and Benchmark Logistics Performance	153
Define Comprehensive Logistics Strategies	154
Develop and Maintain Infrastructure	154
Improve the Regulatory Environment	154
Trade Facilitation	154
Skills Development	155
Exploit Disruptive Technologies	155
References	156

Introduction

Logistics performance has emerged as one of the important ingredients for trade competitiveness for firms and countries. Logistics systems help firms and producers to reach markets at the domestic, regional and international levels. The systems also play an important role in a country's ability to participate in global value chains. Trade logistics refers to the process of planning, implementing, and controlling the efficient flow and storage of internationally traded goods, services, and related information from point of origin to point of consumption. The logistics system comprises a wide variety of activities that include transportation, distribution, packaging, warehousing, and supply chain management and consulting services, among others (Fig. 1). Logistics systems are a network where the quality of services is a function of the level and type of demand that has to be served, the availability and efficiency of core infrastructure, the sophistication of the services, and the effectiveness and appropriateness of the regulatory regime in a particular country. The services have to continuously evolve with changes in any one of these variables, but especially demand.

The flow of goods in trade logistics is structured in supply chains. A supply chain is a system of people, technology, activities, information, and resources organized into activities for moving a product or service from supplier to customer. Given the variety of elements, the efficiency of a logistics system is determined by how well the constituent components function together, and their ability to offer seamless services to shippers. A supply chain is designed for specific types of goods or services moving on a specific route. The logistics services used in the chain can be modified to meet a specific objective with regard to the time, cost, and reliability. In developing economies, the main flows are primary products, linked especially to the agriculture and mining sectors and a few manufacturing or distribution firms that together generate the core demand for logistics services, as they have to move inputs needed for production and products that are needed elsewhere or to supply the market.

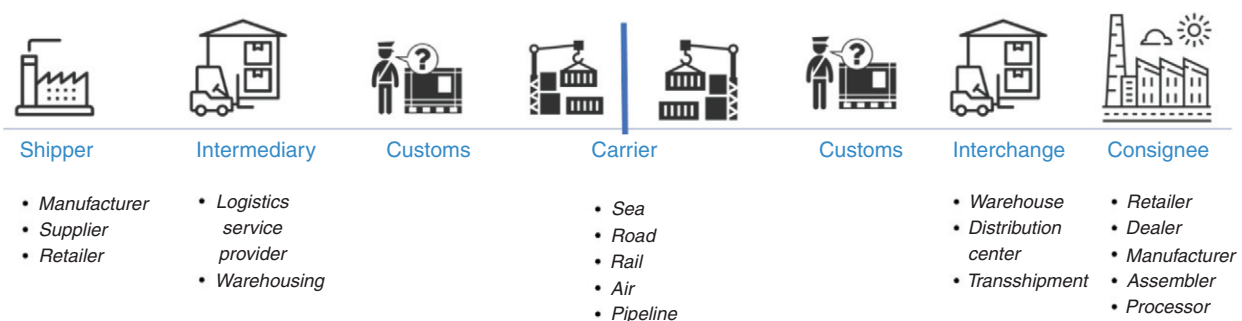


Figure 1 Core elements of an International Logistics System.

Measuring the Performance Efficiency of Logistics Systems

There are a few core metrics that are used to measure the performance of a supply chain. The most common measures are time and its variance to reach target markets and the cost of shipping goods.

The time it takes shipments to reach their destinations is an important consideration in logistics. One of the predicaments in developing countries is that their dominant economic activities, typically agriculture, generate time-sensitive products and yet, at the same time, their logistics system performance is not always suited to speedy deliveries. As a result, the incidences of postharvest losses are very high, as products deteriorate on the farm, during transfer to market or at the market. The small scale of production can result in lengthy delays before products are collected, or if the consignments are collected individually, then the unit costs of logistics become very high. A possible solution that is often used in advanced economies is to have scheduled services. However, individual farmers are typically not able to sustain regular supply. Consequently, the market depends on on-demand services and spot pricing of services. There is limited use of long-term contracts, except in the case of larger firms or cargo consolidators. Spot contracting generally comes with high prices and often discourages service providers from upgrading their services and can result in frequent mismatches between demand and supply of logistics services. In contrast, long-term contracts, typical of economies with thick markets, allow for a better distribution of risk, by assigning it to the party in the best position to manage associated risks. In addition, long-term contracts also provide a formal mechanism for coordination between logistics service providers, as typified by multimodal transport contracts.

The reliability of logistics systems can often be more important to traders than the actual time. A trader needs to be sure that the goods will arrive as expected, so even if the transit time is a few days longer this can be allowed for in the production and delivery schedule, whereas uncertainties (especially in the transit time) require an allowance that is greater than the uncertainty. Where uncertainties in the time required for supply chain activities occur, they will propagate down a supply chain, often becoming more magnified as sequential schedules are missed. In order to avoid these problems, shippers either store larger volumes in warehouses and storage facilities or introduce slack time into their schedules. Logistics in developing economies is therefore characterized by large volumes of products in storage or frequent stockouts as a consequence of unreliable systems. Storage and inventory costs can therefore be very high—or alternatively, shippers are excluded from some markets due to an inability to guarantee reliable supply.

Ultimately the direct cost of logistics, usually monetary, is the performance measure that producers in developing economies pay the most attention to. Individually, the producers are price takers and therefore are not able to influence price trends in the market. They therefore pay attention to cost in order to retain market competitiveness. In many instances, freight rates are the largest component of direct costs. The rates typically reflect the balance or rather, imbalance in traffic flows that are observed in most developing economies. There is usually one dominant direction of goods flow, often associated with agricultural seasons. In many cases, prices during the peak season are set above the long-run marginal cost and the rates for backhaul cargo during slack seasons set at the marginal operating cost. Even then, when averaged over all shipments, prices tend to be set above the average cost, to cover for unreliable demand. It is also quite common in many developing countries for costs to include charges that are informal, to expedite shipment or processing of products through regulatory controls. However, the administration of the informal payments is often very opaque, which makes it difficult to ascertain the ultimate beneficiaries.

There are significant seasonal variations in logistics flows and prices in many developing economies, in concert with agricultural cycles. Often, periods of peak demand lead to rates that are significantly higher than during off-peak periods. For instance, in road transport there is usually a high level of competition but access to certain trades, such as transit cargoes may be restricted, while other markets such as port imports or commodities requiring special vehicles, may be vulnerable to cartels. Rail freight rates are often determined through regulation, but increasingly are set relative to road transport rates.

Often, there are trade-offs between the various performance measures of logistics. The most common trade-off is between time and cost. Generally higher value products give greater importance to time because of the impact on the costs of inventory in transit, as well as the higher value the consumers give to immediate availability. The opposite applies for lower value products. However, for most products there is a differentiation between the value of time for large orders and for restocking. Where demand is volatile, a third category or rapid restocking places additional emphasis on time.

Key Bottlenecks in Logistics

There are a variety of bottlenecks that affect the performance of supply chains by increasing costs, delays, and the unpredictability of arrival. These may be physical constraints associated with a lack of capacity or diminished throughput. The effect of the bottlenecks is magnified in sequential activities, as uncertainties become more pronounced. They may result from delays in the processing of documents and payments or other procedures that temporarily interrupt the flow of cargo.

Global indices such as the World Bank's Logistics Performance Index (LPI) show that the logistics performance of low-income economies as a group lags behind that of middle- and high-income countries (Fig. 2). For instance, the 2018 edition of the LPI shows an average score of 2.35 out of 5 for low-income economies compared to 2.57 and 3.14 for middle- and high-income economies, respectively. On average, low-income economies have scores that are 48% below the average performance of high-income countries. Weaknesses in the logistics performance of low-income countries in particular are across all the dimensions of the LPI, covering infrastructure, logistics competence, quality of services, customs, ability to track and trace shipments, and ability to manage

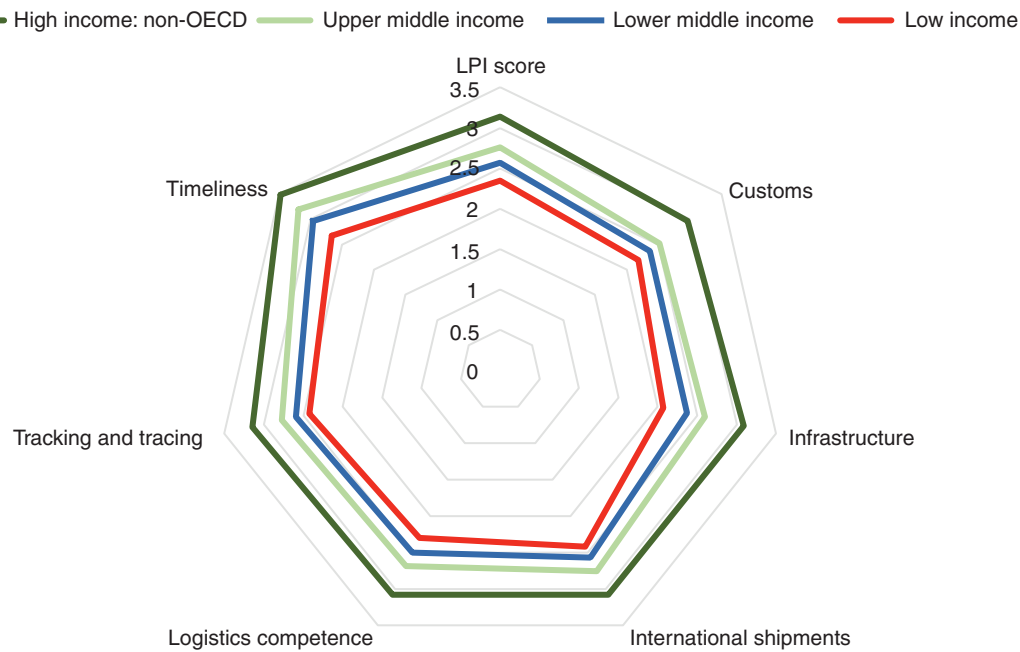


Figure 2 LPI performance by Income Group, 2018. *Source:* Own estimates based on World Bank, 2018a. Connecting to Compete: Connecting to Compete 2018, Trade Logistics in the Global Economy, The Logistics Performance Index and Its Indicators. World Bank, Washington, DC.

international shipments. The causal factors behind these recorded system attributes are many, reposed in infrastructure, regulatory regimes and institutional mechanisms.

Various factors affect the efficiency of transport services. In addition to the condition and congestion on the right-of-way, which determine the average velocity, there are delays at the various intermodal and intramodal nodes. These delays arise due to difficulties with problems of productivity and congestion at the nodes as well as synchronization of the connecting services. These become more problematic during periods of peak demand, when there is a loss of capacity or where quotas are introduced on the services that receive the shipments.

A major source of high costs of logistics in developing countries is the thin and fragmented demand for logistics services. Freight rates are affected both by the industry-wide factors and the characteristics of individual shipments. Two critical factors are the size of a shipment and the potential for finding a backhaul cargo. The latter, which can reduce freight rates by as much as 40%, is especially problematic from cross-border movements where the return cargo is limited to cargoes destined for the same country from which the cargo originated. Significant savings can also be obtained by shipping full container loads, truck loads, rail wagon loads, and airfreight shipments of 500 kg or more. For rail transport there are significant savings for longer trips due to the time spent loading and unloading trains. The same applies for road transport because of the cost for mobilization and return to the dispatch point. For shorter distances, the cost is dependent on the number of trips that can be made in a day.

There are significant economies of scale in logistics services. Wide bodied aircraft, Post-Panamax container vessels, larger trucks, and rail wagons, all offer significant savings because of lower fixed costs per ton of capacity and lower fuel consumption per ton-kilometer of capacity size and capacity of transport units. Investments in the extension of runways and deepening of port entrances are meant to reduce the cost of transport by attracting larger, more efficient aircraft and vessels. The conversion to broad gauge and strengthening of rails that is taking place in South Asia and East Africa is partly meant to accomplish the same for rail transport. A reduction in road transport can be accomplished through strengthening pavement and increasing the weight capacity of bridges. However, a challenge in developing economies is that the average shipper generates only a relatively small amount of cargo and it is usually only possible to reduce costs in the short run by increasing the average load factors and annual utilization for existing transport units.

On services, one major variable is the level of competition for services and differences between peak and off-peak demand. The degree of competition within the logistics sector, for both transport and nontransport services, is important for both the efficiency and variety of the services provided. For the transport sector, there are obvious issues of public sector restrictions through quotas, standards, and financial requirements, but also indirect constraints related to taxes on equipment, access to finance and lack of security and safety on the road network. For nontransport logistics, there are various opportunities for public sector restrictions through licensing and certification, but these are rarely a problem. More common are limitations on the participation of foreign service providers. The principal constraint for local service providers is the preference of foreign importers and exporters for using international logistics service providers, though this does not prevent local service providers from entering into various forms of partnership with their international counterparts.

Competition within and between modes is another important factor in reducing the cost of transport and can also play a role in increasing the quality of service. The extent of competition can be determined for each mode, but applying different measures in each case. For road transport, the size of large trucking firms and the number of such firms provides such a measure. In most cases, the large companies will hire independents, but this form of subcontracting does not affect the level of competition. While rail transport is usually provided by a single company, the competition from road transport is usually intense. Information on the market share for freight transport usually indicates a small market share and most of this is dominated by shipments of petroleum and minerals. The exception is for container movements where there are large inland container depots located far from the port. While international air freight and ocean shipping are relatively competitive, the former is often limited by bilateral agreements and restrictions on the carriage of cargo. A more significant constraint on competition is the cargo handling operations at the airports and seaports. Increasingly, these services are being provided by private operators having concessions obtained through competitive bidding. However, national airlines continue to have a strong presence in cargo handling operations and the port authorities often participate in the regulation of cargo handling charges while collecting royalties from the concessionaires. The results have only a modest impact on the efficiency of operations but can result in excessive charges for these services.

Market failures can also occur because of lack of knowledge on the part of the participants or because of external events. The most common of these is excess supply and capacity, especially in transport services, which leads to a sharp decline in rates and degradation in the quality of service. However, there is also the failure of shippers to plan strategically and treat logistics as part of their value chain rather than as a cost. Although there is widespread recognition of the need to diversify exports and move up the value chain, there is frequently a failure to appreciate the importance of logistics in achieving these goals. The same applies for imports, albeit in a less competitive environment.

In terms of international logistics, the main weaknesses in developing countries are typically the trade facilitation environment and, especially, the performance of main trade gateways, particularly seaports, airports, and border crossings. These dimensions are part of the trade facilitation agenda which refers to the ease of clearance of goods, including the movement of cargo through international gateways. Efforts to improve the efficiency and transparency of trade facilitation core functions have to be an ongoing endeavor. The basic problems in developing countries are delays, informal payments to expedite clearance, lack of transparency, and limited compliance with regulatory requirements. Uncertainty in the time required to clear cargo, especially imports, limits the ability to introduce more efficient distribution systems and reduce inventory in the supply chain. Because of the importance of exports to a country's economies, most countries have largely eliminated the delays for clearing export cargo, thereby reducing delivery times and allowing for scheduled movements from the factory to the final destination.

Even in the case of imports, there have been significant improvements over the last decade. For a significant proportion of companies, imports with proper documentation can be cleared within a couple of days versus 1 day or less for exports. In a growing number of countries, the percentage of import shipments subject to physical inspection has decreased to 20% or less, while the number of shipments covered by various programs for expedited clearance is increasing from lows of 20% to highs of 80%–90%. These figures apply to the total trade, most of which moves through seaports. The clearance time at land borders and airports is much less, but so are the requirements. The delays in clearing cargo at ports only affect the cargo whereas at the land borders they also affect the movement of vehicles and therefore the cost of transport. Therefore, delays for more than 4–8 h increasingly are unacceptable. Airfreight is more demanding because the impact of delays at the airport have a greater impact on total delivery time. Airfreight is used where the value of time is extremely high. Any delay in clearing import cargo beyond 4 h is considered excessive.

The productivity of the gateways, as measured in terms of the cargo dwell time, is of paramount importance. Many of the factors affecting gateways apply in general to intermodal and intramodal transfer points. The efficiency of these transfer points is improving as a result of greater coordination between the transport operators delivering and receiving the cargo. Increasingly, these are the same company which either control the transport services on both sides or control one and subcontract the other. In some cases, these movements are controlled by a third party which contracts both services. The capacity and efficiency of these hubs are also being improved as the technology used in the gateways is transferred. Inland container depots, logistics hubs, and urban truck terminals have all been upgraded to provide faster transfer of cargo and better tracking of cargo movements.

Improving Logistics Performance in Developing Countries

Understand the Demand and Benchmark Logistics Performance

What an economy trades matters a lot in logistics. The logistics system has to be able to meet the needs of the specific products that are already traded or those that a country aspires to trade. A first step in improving logistics performance is to understand the nature of the demand. There are a variety of national and international sources of trade, transport and logistics data that can be tapped. The data should allow a review of the country's trade to identify those that are important, growing, or otherwise of strategic importance. They should also provide information on both volume and value for the trade in specific commodities. Information on traffic volumes is available from international sources, but greater detail is available from annual statistical reports prepared at the national level or by specific organizations. The latter are increasingly available on the web.

Based on the data collected, benchmarking can be implemented to assess how well supply chains or an economy is performing compared to others with which it competes. Various indices assessing the quality of logistics services have been developed in recent

years. However, most indices measure overall performance which is determined by a combination of private service providers and public regulators. They all provide an aggregate measure. Ideally, the data should be at a disaggregate level to offer granular insights about the performance of specific supply chain components.

Define Comprehensive Logistics Strategies

Logistics in many countries is characterized by fragmented service provision, regulations, institutions, and modal imbalances. Many countries have more than one mode of transport but most traffic is carried by road. This results in severe congestion on the roads, underutilized railways and other systems that may exist, and increased negative impacts on climate. In fact, each system would have evolved as a discrete layer in the overall network with limited integration and coordination of investments and operations across the different networks. The separation of the networks is reinforced by the institutional arrangements for the transport sector and the regulation of logistics services which are similarly fragmented. Emphasis should be placed on enhancing planning and strategic development that are optimized to minimize system costs, from a development and user perspective. Emphasis should be on multimodal systems of logistics. Consideration should be given to rationalizing or at least coordinating the agencies that are involved in providing logistics infrastructure and services. In addition, any strategy should be codeveloped together with the private sector, shippers, and services providers, as major stakeholders (Oxera, 2017).

Develop and Maintain Infrastructure

There are many studies that point to large infrastructure backlogs across much of the developing world. The symptoms of poor infrastructure are congestion, poor service quality, and poor connectivity. Filling the infrastructure gaps is important, but how to finance it has become one of the major issues. Increasingly, governments are turning to the private sector for infrastructure financing. There is great potential for the private sector to finance network improvements, as well as having responsibility for the management of transport and logistics systems. Most success has been registered in parts of the systems with guaranteed revenue streams, such as ports and airports, but also railways and more and more highways, especially in countries with high economic densities. However, care should be taken to connect lagging regions, as often they are also the poorest ones. There is also a need to move toward the standardization of infrastructure design to interconnect with neighboring countries. Ultimately, regional integration offers great potential to enhance the economic outcomes of investments in logistics infrastructure and associated services.

Improve the Regulatory Environment

A modern logistics framework should be clear on what needs to be regulated and what need not to be regulated. Obtaining the input of all relevant stakeholders is critical for a sound and well-understood legal and regulatory framework that will help modernize and develop transport and logistics services. It is also good practice when formulating legislation to carry out some regulatory impact assessment to balance costs and benefits of such new legal instruments. Overall regulation should be kept minimal and limited to only a few issues relating to safety or market failures; regulations should be accessible; and bilateral, regional, and international agreements should be taken into account. More importantly, it is critical to enforce any regulations in a consistent and nondiscriminatory manner.

A conducive regulatory regime should nurture the evolution of logistics services in concert with changes in demand (Fig. 3). Over time, a professional class of service providers, some with no physical assets, will emerge to offer sophisticated bundles of services. Since the 1980s, firms in developed economies have developed customer-oriented production and supply chain management that has contributed to increasing market segmentation. This has caused vertical disintegration as a result of companies focusing on their core business and utilizing distributed processing. Starting in the 1990s, logistics structures were reorganized to improve performance. This included reducing inventories, outsourcing transport, and establishing horizontal collaborative networks to share the costs of logistics facilities. This has resulted in an emphasis on improved scheduling of product flow—lead time management, lean and agile logistics, process mapping, and efficient consumer response. Many low-income economies are still to go through this evolution, which is very much part of modern dynamic supply chains.

Trade Facilitation

In many countries, the focus of trade facilitation is on the simplification and harmonization of procedures and documentation, with special attention given to the release and clearance of goods, including goods in transit. While most of the attention has been given to customs procedures, there is growing interest in the other regulatory procedures performed at the border. A number of standard reforms have been introduced in most countries, corresponding to the various conventions and standard practices introduced by the WTO and WCO, more recently through the WTO Trade Facilitation Agreement (Kunaka et al., 2013). The various international instruments have promoted the classification and valuation of goods, the format, submission and processing of customs declarations, the procedures for selecting shipments to be inspected and the treatment of courier services. Most of the improvements have been achieved with the use of improved IT systems.

Efforts to reduce the level of inspection have focused on the introduction of IT-based risk management systems. Traditionally, risk management is implemented through subjective judgments of customs officials. IT systems have introduced selectivity modules that

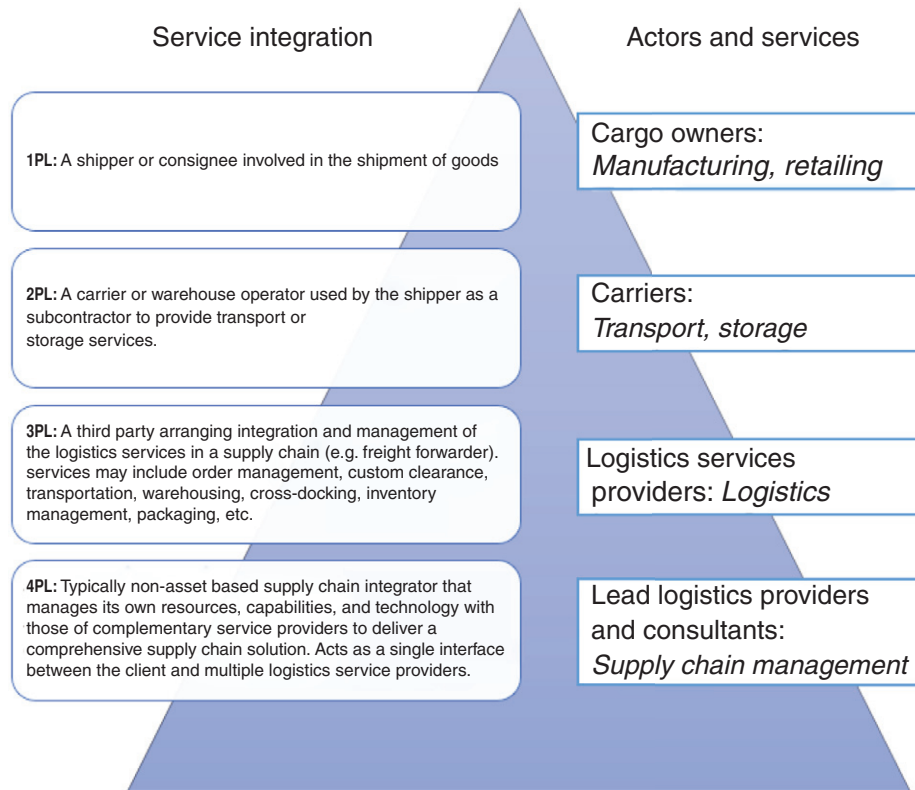


Figure 3 Hierarchy of Logistics Services. Source: Own construction after Rodrigue, J.P., 2017. *Developing the Logistics Sector: The Role of Public Policy. Technical Report*.

apply objective criteria related to the type of commodity and the source in terms of origin country and shipper. These criteria are developed through the preparation of risk profiles developed by designated customs officials. Efforts to apply statistical analysis and business analytics have begun but face significant institutional challenges as customs authorities replace their cadre of officers with personnel having modern technical skills.

Skills Development

The low skills level of logistics service providers increases logistics costs and the costs of externalities. For instance, in many countries, even in advanced economies, logistics services are labor-intensive; dominated by truck drivers, warehouse operators, and administrative personnel. The difference in low-income countries is that many of these workers are low skilled and illiterate. Besides lowering service quality, poor driving skills also contribute to road crashes. It is therefore important to invest in skills development across the public and private sector, so the various role players can continue to improve performance. Fortunately, there are increasingly many schools and universities that offer courses on logistics (McKinnon et al., 2017).

Exploit Disruptive Technologies

Disruptive technologies offer the promise of improved and even different forms of connectivity, when compared to what exists now. As such, the long-term planning of infrastructure and services needs to take into account the changes that modern technologies are feeding. The World Bank (2018b) defines disruptive technologies as technologies that result in a step change in the cost of, or access to, products or services, or that dramatically change how we gather information, make products, or interact. In logistics, they are impacting how and where products are manufactured and delivered, the speed of delivery and the management of inventory. All countries, regardless of level of income, have to take into account some the emerging patterns. This is despite that fact that past performance is no longer a good predictor of future patterns.

References

- Kunaka, C., Mustra, A., Saez, S., 2013. Trade Dimensions of Logistics Services: A Proposal for Trade Agreements. Policy Research Working Paper Series 6332. World Bank, Washington, DC.
- McKinnon, A., Flöthmann, C., Hoberg, K., Busch, C., 2017. Logistics Competencies, Skills, and Training: A Global Overview. World Bank Studies. World Bank, Washington, DC.
- Oxera, 2017. Policy Approaches for Improving Logistics Unpublished Paper Prepared for the World Bank, Washington, DC.
- Rodrigue, J.P., 2017. Developing the Logistics Sector: The Role of Public Policy Technical Report.
- World Bank, 2018a. Connecting to Compete: Connecting to Compete 2018, Trade Logistics in the Global Economy, The Logistics Performance Index and Its Indicators. World Bank, Washington, DC.
- World Bank, 2018b. Disruptive Technologies and the World Bank Group – Creating Opportunities – Mitigating Risks. Joint Ministerial Committee of the Boards of Governors of the Bank and the Fund on the Transfer of Real Resources to Developing Countries. World Bank, Washington, DC.

Freight Network Modeling

Lóránt Tavasszy, Yousef Maknoon, Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	157
The Network Flow Model	157
Infrastructure Networks and Their Operation	158
Service Network Design	159
Vehicle Routing	160
New Frontiers in Research	160
References	161

Introduction

In this chapter we review the different approaches and methodologies for network modeling and their main applications for freight transport. Network modeling in transportation science is one of the main areas of application of network theory. It is based on graph-theoretic principles—mathematically speaking, networks are graphs that represent relations (edges) between discrete objects (nodes), where the edges and/or nodes have properties (e.g., travel time, capacity). These conventions are the cornerstones of models of the real world of transportation networks, services and flows. In freight transportation, networks are visible in physical transport infrastructure, transport services, and even trading activities. Network models do not necessarily only represent tangible aspects of the transport system, but may also represent the availability of a service offering as potential movement, or a trading relationship as abstract connection.

The basic techniques for freight network modeling evolved from operations research (Ahuja et al., 1995). Our description of models largely follows this tradition. Applied freight network modeling for civil engineering purposes developed in the 1960s, during the times of rapid expansion of national infrastructure networks, and was extended with microeconomic models to study the regulation of network industries, like the US rail freight industry. During the early 1980s, the emphasis in passenger network modeling in transportation shifted to congested networks and traffic management (Sheffi, 1985), while in freight transport, descriptive models of broader logistics and trading processes gained momentum (Harker and Friesz, 1985).

Network models require a representation of the demand and the supply side of the market of transport services. Network design models are often applied to calculate the optimal structure of the network, or its services. In such a setting, demand and supply models can be framed normatively as interdependent optimization models. The optimal design of a network structure or its services has to follow the expected usage of the network. At the same time, the usage of the network will depend on the available network and, based on the principle of shortest or cheapest paths, optimizing the flow over the network given the available services. Users can either work together to minimize the total transport costs in the network (system optimization) or choose paths separately (user optimum). Seen from a game theoretic perspective, the framing is one of a bilevel optimization problem. Here the design of the network or services is the main problem, and determination of network flows is the underlying problem; these two have to be solved together. Economic theory adds the idea that user behavior can only be described to a limited degree using optimization; therefore, descriptive models can be used that build on ideas of optimization but also explain how users deviate from a supposedly optimal situation. These descriptive models can still be applied in a bilevel framework, but will result in relatively sophisticated network flow calculation that allows for typical behavioral phenomena like imperfect information and bounded rationality. Network flow models can also be applied in isolation, to predict the expected use of a given network, in order to evaluate impacts of transport policies or traffic management measures.

The remainder of this chapter is built up as follows. Below we first discuss the way that flows are portrayed on the network (Section “The Network Flow Model”). Section “Infrastructure Networks and Their Operation” discusses approaches to modeling infrastructure networks and their operations. In Section “Service Network Design”, we introduce the models for service network design. Section “Vehicle Routing” follows up with the related vehicle routing problems. Final section suggests some directions for new research in freight network modeling.

The Network Flow Model

In a network flow model, a network with nodes (N) and arcs (A) is given, and we try to determine the flows over this network, given a table of origin-destination flows between nodes. We will generally assume that flows will try to follow paths through this network such that supply and demand in each node is satisfied and transport costs through the network are minimized. To this end, we also

assume unit costs through the network for each arc (noted by c_{ij}). The general mathematical formulation of the minimum cost flow problem is as follows:

$$\text{Minimize } \sum_{(i,j) \in A} c_{ij}x_{ij} \quad (1a)$$

$$\text{subject to :} \\ \sum_{\{j:(j,i) \in A\}} x_{ij} - \sum_{\{j:(i,j) \in A\}} x_{ij} = b(i) \quad \forall i \in N, \quad (1b)$$

$$l_{ij} \leq x_{ij} \leq u_{ij} \quad \forall (i,j) \in A. \quad (1c)$$

The objective is to find an arrangement of flows x on arcs (i,j) such that the network total of the products of arc flow volumes and unit costs are minimized. The first constraint relates to the node flow balance—outflow minus inflow has to equal the supply (positive balance) or demand (negative balance) of each node. The second relates to the bounds of the arc capacities (mostly zero as lower bound and a highest flow value due to service or physical limitations to flow).

Starting from this general form of the minimum cost flow problem, many extensions are possible. Extensions can be formulated by a broader interpretation of the term network than relating only to the single mode, single user, single good, fixed unit cost case.

- The assumption about the dependence of the link unit costs c upon the flow x , or $C(x)$, is crucial for applications (indicating congestion or economies of scale). Convex cost functions, where C increases with x , allow the calculation of a global optimum, or equilibrium solution, for a network. Nonconvex functions, where C decreases with increasing x , do not exhibit one global optimum. There will be multiple solutions of such networks, only distinguishable through the starting conditions.
- Single commodity flow problems relate to one origin and one destination. Multicommodity problems include several origins and destination and will have a unique, flow-dependent constraint (1b) for each origin destination pair. Second, commodity attributes may differ across the network. If flows from multiple O/D pairs use the same network, they will compete for scarce capacity on that network. In some multicommodity applications, the optimization will not involve the minimization of the costs of all flows together, but of groups of flows individually, each trying to minimize their own costs. This is typically the case in a multicarrier network where cooperation does not exist, but flows use one and the same network.
- Multimodal networks include links representing different modes of transport, including transshipment between these modes. Beyond the individual links, services that connect links or even modes of transport can also be defined. Links may represent transshipment movements, and paths will then indicate intermodal transport chains, where goods are passed from one mode to another.
- Network models may also include choices that relate to nonphysical aspects such as departure time, or volumes of demand and supply. The first problem will involve multiple networks for different periods of time, where the user can choose which network to enter. The minimum cost solution can be calculated in the extended network as before, but now the solution will implicitly include the choice of the time of travel. In the second problem, demand and supply is assumed variable and dependent on network costs—responding positively to cost reductions and negatively to cost increases. This also creates an additional layer in network models, which can be mathematically included in the network optimization formulation. Solving the extended optimization problem will also result in an equilibrium state of the demand and supply of each node. These applications are known as extended freight network equilibrium models or spatial price equilibrium models.

Applications of network flow models are essentially predictive in nature, that is, meant to describe the expected use of the network under different future network and policy scenarios. Because of that it is essential to validate these models against observed behavior of network users. Users will often not follow a simple cost minimization logic, in many respects. In the simple form as formulated above, the network cost is an abstract term and may be related to any cost measure such as distance, service rates or time. In practice, many dimensions of costs will matter and will lead to a multidimensional formulation of link costs, including all important components. Standard practice is to use a weighted linear, so-called generalized cost function. In addition, recognizing variations between users and various uncertainties that determine individual behavior, probabilistic approaches become important. These phenomena have led to the development of sophisticated network choice models that depart from the assumption that a flow can only choose one link between two nodes. If there are alternatives, choice will be probabilistic. At the individual level, this leads to choice probabilities. At the aggregate level, this leads to a spread of flows of multiple users over multiple alternatives. Discrete choice, random utility models constitute the state of practice here today. Effectively, they add another constraint to the basic formulation, forcing flows between origin and destination nodes to follow a path dictated by probabilities of choice that conform to observed choice behavior of users.

Infrastructure Networks and Their Operation

Network design focuses on the strategic design of infrastructure. It is a joint decision about the design of infrastructure and its operation. The infrastructure design deals with the decision on establishing facilities as well as their connection topology. The infrastructure operations decisions focus on the assignment of the customers to the facilities and the routing of customers' demand

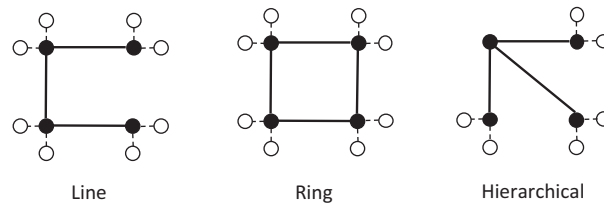


Figure 1 Network topologies.

in the designed network (Contreras and Fernandez, 2012). The problem looks to optimize an economic (e.g., total transportation cost), an environmental (e.g., congestion, CO2 emissions) or a social (e.g., service accessibility) function.

The network design problems are generally presented on a multigraph $G(V, A)$. The set of vertices represents a potential facility location and/or customers. Two types of arc exist in this graph. The first type is used to present the transportation of commodities in the network. Commodities could represent a specific origin-destination pair or specific material type. The second types of arcs/edges are used to model the network topology and the allocation of customers to the facilities (Fig. 1). In location planning, facilities may need to be connected with certain rules. For example, for the case of the railway network, facilities may need to be connected as a ring or line. Another example will be the design of the distribution network for postal service in which facilities (local and regional hubs) may need to be connected hierarchically (Laporte et al., 2015).

In the general network design problem, facilities have different functions and roles. They can either act as the sender/receiver of the commodities or as a transshipment point for consolidation. In the literature, the well-known facility location problem can be seen as a specific case of the general network design problem, in which facilities operate as a sender or a receiver of commodities. In the classical facility location problems, it is assumed that facilities are disconnected. The design decision is then limited to the selection of facilities. Similarly, the operational decision is limited to the allocation of customers to the facilities. In the p -median problem, the decisions are to select p facilities and allocating each customer to the selected facilities to minimize total transportation cost. By relaxing the number of selected facilities, the fixed-charge problem looks to simultaneously select facilities and allocate customers to minimize the total weighted cost (cost of establishing facilities and serving customer demand). The p -center problem is another variant of the p -median problem which aims to minimize the maximum cost. In recent years, studies have shifted toward the general problem in which the network topology plays a key role in the service. These problems have application in the multiechelon production-transportation system.

In some logistics systems, facilities act as a transshipment point, where flows between users are consolidated at the facilities and then rerouted to the final destination. This problem is the subject of studies in the network design problem. Early studies develop models based on topology in which all facilities are connected. The design decision is then limited to the selection of facilities and operational decisions that focus on the allocation of customers to the facilities, and routing of commodities in the network. Also, it is assumed that facilities have an unlimited capacity to process customer demand. This assumption facilitates the formulation approach, as one can prove that at most two facilities have to be visited between every pair of customers' demand. Similar to facility location problems, the p -median problem aims to locate p facilities to minimize the total transportation cost. The p -center problem aims to select p facilities to minimize the maximum transportation cost. In recent years, research studies have been performed toward the problems in which the connection of facilities have to follow certain network topology (e.g., ring, line, hierarchy).

Most network design problems are NP-Hard. For moderate size problems, they can be solved by using exact methods such as enumeration, branch and bound. For larger problems, various heuristic and meta-heuristics methods are developed to solve the problem. Interested readers are referred to Laporte et al. (2015), Mladenović et al. (2007), and Alumur and Kara (2008) for a review of the resolution approach.

Service Network Design

The service network design refers to planning freight movement over the relatively long distance between terminals with a given, or combinations of, transportation mode (e.g., truck, rail, ship, plane). For a given network, service network design looks to answer three main questions: (1) selection and scheduling of services, (2) traffic distribution, and (3) relocation of empty services. Service network design problems are modeled as the generalization of the multicommodity flow problem. Most of the exact approaches were developed based on the Dantzig Wolfe decomposition approach. Interested readers are referred to Wieberneit (2008) and Andersen et al. (2009) for an overview of the resolution approach for service network design problems.

Freight traffic assignment is a basic problem underlying service network design. This problem is an equivalent of the minimum cost network flow problem. The goal is to find the minimum cost of routing products over an existing network. Special cases of the assignment problem are the well-known shortest path problem and transportation problem. The traffic assignment problem has been extended in the fixed-charge network design problem. In this problem, products can be transported between nodes if the associated service is established. The problem can be formulated as a multigraph in which nodes represent terminals and arcs either denote product movement between nodes or a service that has been established between two nodes. Important issues to consider include (1) usually the transported demand is assumed to be constant over time. Almost all practical problems have some demand

variation. If they are not considered in the planning problem, it causes inefficiency in the capacity usage of transporters. (2) The routing of services is not always considered. In these cases, the cost associated with transport resources can be over- or underestimated. For example, in air transportation, aircraft routing is the main contributor to transportation costs. Thus, its detailed itinerary has a great impact on total transportation cost. To consistently solve the above-mentioned issues, new definitions of the service network design problem were introduced. Here, service network design is generally defined as a time-space network. The planning horizon is described as a repetitive pattern (e.g., week, month). The problem is then defined as a fixed-charge network design problem over the time-space graph. This representation enables the planners to address the rotation of services and demand fluctuation explicitly.

Vehicle Routing

Last-mile delivery problems deal with the transportation of products in which origin-destinations are in a relatively small-sized geographical area (e.g., city, country). The pickup and delivery problem is the general problem dealing with last-mile deliveries. When the transportation plan has a single origin or destination, then the problem is the well-known vehicle routing problem (Toth and Vigo, 2002). The vehicle routing problem deals with finding a minimum cost route to satisfy the demand. This problem has been extensively studied in the literature. The problem can be formulated as a network flow problem. With this approach, branch and cut algorithms are a common exact method of solution. Most of these approaches focused on introducing valid inequalities to improve computational performance. The vehicle routing problem can also be formulated as a set-partitioning model and be solved by using branch and price algorithms. Interested readers are referred to Baldacci et al. (2012) for an overview of the exact approach to solve vehicle routing problems. For very large problems, a solution to vehicle routing problems can be found by using a heuristics approach (Laporte et al., 2000).

Real-life routing problems may require to satisfy various restrictions (e.g., pickup and delivery regulations, maximum ride time, time windows). These restrictions are generalized as a rich vehicle routing problem. The term “rich” refers to various features associated with the vehicle tour (Lahyani et al., 2015). In recent years, the last-mile delivery problem is studied as integrated operational planning. The inventory-routing problem (Coelho et al., 2014) deals with inventory decisions and routing of products. Similarly, the production routing problem looks to optimize production, inventory and distribution jointly (Adulyasak et al., 2015). The difficulty of solving inventory and production routing problems are inherited from the travel-salesman problem. In general, these problems are NP-Hard and an exact method fails to solve real-life problems with a reasonable time. Readers are referred to Coelho et al. (2014) and Adulyasak et al. (2015) for an overview of the models and resolution approaches to solve these problems.

New Frontiers in Research

Advancement of information technology, innovations in transport technology and green logistics are three main trends that will shape the future of freight transport systems.

- Freight transport is one of the main contributors to air pollution. In recent years, innovations in transport technology has revolutionized transportation systems. With the introduction of electric vehicles and drones, freight transport systems have the possibility of offering emission-free services. On the other hand, they have limited autonomy. Electrical vehicles need to be recharged and drones cannot be used to carry large items. These limitations are needed to be addressed at strategic, tactical, and operational planning levels. For example, last-mile delivery can be executed by trucks and drones. With this approach, one can possibility reduce the cost while improving the response time. At the same time, the solution is required to overcome the challenge of synchronizing the operation of trucks with drones (Agatz et al., 2018).
- The growth of online retail has shifted the traditional logistics planning from offering a common low-cost solution toward personalized and swift services. In this environment, the companies need to have an insight into the customers' behavior and balance the need of customers with the operational restrictions. This is not limited to the end-user but also service providers. For example, in crowdsourced delivery systems, companies can outsource their logistics tasks (Arslan et al., 2019). However, they need to predict the reaction of each individual while a task is assigned to them. Similarly, in time-slot management systems, companies offer delivery time to the end-customers (Agatz et al., 2011). In this case, they need to know how they can distribute customers between delivery timeslots to reduce their transportation costs while making the customers satisfied.
- Sustainability has entered into network modeling as a new goal, and adds environmental and social objectives to the classical criterion of maximum efficiency. These objectives individually will generally lead to different network designs. Measures needed to reduce environmental impacts of transportation (such as collaborative routing, higher inventory levels, internalization of external costs) may negatively affect the performance of services for individual shipments. New technologies like electric vehicles, the tracking of shipments' individual trajectories and advanced emissions measurements further add to the complexity of designing sustainable networks. New approaches are needed, therefore, to reconcile the different objectives and to consider more advanced network designs (Psaraftis, 2015).

References

- Adulyasak, Y., Cordeau, J.F., Jans, R., 2015. The production routing problem: a review of formulations and solution algorithms. *Comput. Oper. Res.* 55, 141–152.
- Agatz, N., Bouman, P., Schmidt, M., 2018. Optimization approaches for the traveling salesman problem with drone. *Transport. Sci.* 52 (4), 965–981.
- Agatz, N., Campbell, A., Fleischmann, M., Savelsbergh, M., 2011. Time slot management in attended home delivery. *Transport. Sci.* 45 (3), 435–449.
- Ahuja, R.K., Magnanti, T.L., Orlin, J.B., Reddy, M.R., 1995. Applications of network optimization. In: Ball, M.O., Magnanti, T.L., Monma, C.L., Nemhauser, G.L. (Eds.), *Handbooks in Operations Research and Management Science*, vol. 7. Elsevier, pp. 1–83.
- Alumur, S., Kara, B.Y., 2008. Network hub location problems: the state of the art. *Eur. J. Oper. Res.* 190, 1–21.
- Andersen, J., Crainic, T.G., Christiansen, M., 2009. Service network design with asset management: Formulations and comparative analyses. *Transport. Res. C* 17 (2), 197–207.
- Arslan, A.M., Agatz, N., Kroon, L., Zuidwijk, R., 2019. Crowdsourced delivery—a dynamic pickup and delivery problem with ad hoc drivers. *Transport. Sci.* 53 (1), 222–235.
- Baldacci, R., Mingozzi, A., Roberti, R., 2012. Recent exact algorithms for solving the vehicle routing problem under capacity and time window constraints. *Eur. J. Oper. Res.* 218 (1), 1–6.
- Coelho, L.C., Cordeau, J.F., Laporte, G., 2014. Thirty years of inventory routing. *Transport. Sci.* 48 (1), 1–19.
- Contreras, I., Fernandez, E., 2012. General network design: a unified view of combined location and network design problems. *Eur. J. Oper. Res.* 219 (3), 680–697.
- Harker, P.T., Friesz, T.L., 1985. The use of equilibrium network models in logistics management. *Transport. Res. B* 19 (5), 457–470.
- Lahyani, R., Khemakhem, M., Semet, F., 2015. Rich vehicle routing problems: from a taxonomy to a definition. *Eur. J. Oper. Res.* 241 (1), 1–14.
- Laporte, G., Gendreau, M., Potvin, J.Y., Semet, F., 2000. Classical and modern heuristics for the vehicle routing problem. *Int. Trans. Oper. Res.* 7 (4-5), 285–300.
- Laporte, G., Nickel, S., da Gama, F.S. (Eds.), 2015. *Location Science*, vol. 528. Springer, Berlin (Chapters 2 and 3).
- Mladenović, N., Brimberg, J., Hansen, P., Moreno-Pérez, J.A., 2007. The p-median problem: a survey of metaheuristic approaches. *Eur. J. Oper. Res.* 179 (3), 927–939.
- Psaraffis, H.N. (Ed.), 2015. *Green Transportation Logistics: The Quest for Win-Win Solutions*, vol. 226. Springer, Berlin.
- Sheffi, Y., 1985. *Urban Transportation Networks*, vol. 6. Prentice-Hall, Englewood Cliffs, NJ.
- Toth, P., Vigo, D. (Eds.), 2002. *The Vehicle Routing Problem*. Society for Industrial and Applied Mathematics.
- Wieberneit, N., 2008. Service network design for freight transportation: a review. *OR Spectrum* 30 (1), 77–112.

National Freight Transport Models

Gerard de Jong, Significance, The Hague, The Netherlands; Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	162
The Four-Step Model and Beyond	162
Modeling Freight Generation and Distribution	165
Modeling Mode Choice and Route Choice in Freight Transport	166
Short Overview of National Freight Transport Models	166
Conclusions	167
References	167
Further Reading	167

Introduction

National passenger transport models have been in use at least since the 1980s, and national freight transport models followed shortly after. Nevertheless, much more effort has been spent on passenger transport modeling than on freight, also within the subcategory of models at the national scale. National freight transport models deserve more attention because, with increasing globalization, freight transport has been growing fast, and a large part of the benefits and internal and external costs of transport (and of investments in transport infrastructure) are related to freight transport.

Important differences between the freight and passenger transport markets are the diversity of decision-makers in freight (shippers, carriers, intermediaries, drivers, operators, etc.), the diversity of the items being transported (from parcel deliveries with many stops to single bulk shipments of hundreds of thousands of tons), and the limited availability of data (especially disaggregate data, partly due to confidentiality reasons).

In this paper, we review models that are used for national transport authorities to forecast commodity flows or freight vehicle flows and the impacts of transport policy and infrastructure projects on these flows.

National freight transport models are used for different purposes:

- to assess the impacts of different types of autonomous developments (e.g., in GDP, population, and oil prices); and
- to (ex ante) evaluate the impact of policy measures, such as changes in national regulations and taxes, or infrastructure investments in specific links, nodes, and corridors.

A wide range of freight transport models and model systems [see [Tavasszy and de Jong \(2014\)](#) for a reference book] are applied by public agencies, either developed by these agencies themselves or contracted by these. Furthermore, a considerable amount of freight transport modeling takes place at universities and at the individual firm level. Models to optimize transport and logistics within a specific firm or supply chain are not discussed in this paper. However, lessons that modeling for public agencies can learn from modeling in the private sector will be discussed where relevant.

Since we restrict ourselves to national freight transport models, the focus will be on transport between regions (more or less intercity flows), not on intraregional flows. The latter are handled in urban freight models. Of course, what is being studied in a national model depends on the size of a country and its freight transport, which can vary enormously between countries. For the largest countries, a national model would mainly have to cover domestic flows, but, for other countries, a national freight transport model by definition will be an international model, containing domestic, export, import, and possibly also transit flows. Also, in interregional transport in, to, from, and through a nation, often several modes are important (road, rail, sometimes also inland waterways, and/or sea transport; possibly also air transport). Many urban models can ignore other modal competition and devote themselves to road transport only.

The Four-Step Model and Beyond

The four-step modeling structure from passenger transport has been adopted in freight transport modeling ([de Jong et al., 2004](#)) with some success ([Fig. 1](#)):

- Generation models for the production of freight transport in each of the zones and the attraction to each zone, per sector (e.g., mining) or commodity group (e.g., petroleum). The output dimension is tons of goods (transport flows) or monetary units (trade flows).
- Distribution models for the flows between pairs of zones, sometimes with a dependence of the distribution on the transport resistance between the zones from the modal split model; the dimension is tons.

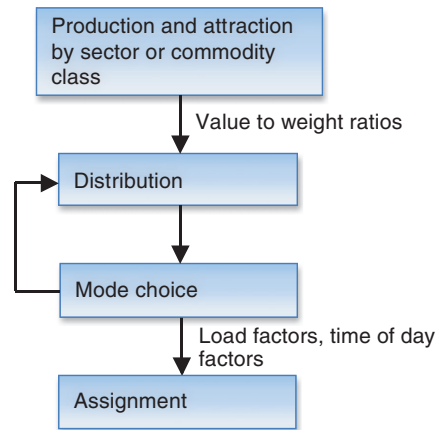


Figure 1 The conventional four-step model in freight transport.

- Modal split models, in which the allocation of the commodity flows to transport modes (e.g., road, train, combined transport, inland waterways, etc.) is carried out.
- Network assignment. After conversion of the flows in tons to flows in vehicle-units, they can be assigned to networks (in some model systems the truck flows are assigned to the road networks together with the passenger cars).

Nevertheless, a comprehensive freight transport model system often requires additional steps. Some of these can be regarded as transformation modules, as in freight transport modeling the units (of measurement) may change from step to step. Freight transport model systems often start in money units, then do distribution and modal choice in tons and finally a network assignment in vehicles.

If the models for the production and attraction step use trade flows in money units, a transformation converting flows in money units into physical flows (tons) is required, since the central part of the freight model will be in tons. The conversion can be carried out using value/weight ratios by commodity group. These ratios could have a large impact on the final model predictions (in terms of tons per mode or number of vehicles per link). For this reason, it is very important to assemble good data on this conversion. Possibly, value/weight ratios can be made an endogenous, policy sensitive choice within the model system.

A second transformation that is used in many freight transport models is that for going from flows in tons to flows in vehicle units (e.g., number of heavy goods vehicles [HGVs]). This transformation is required when mode choice is in tons and the network assignment in vehicles (sometimes assignment is done directly for tons). The factors used for this are called “load factors,” usually consisting of a fixed value in tons for each mode or vehicle type. However, in reality the vehicle load is influenced by a number of decisions that are usually not modeled in freight transport models (notably decisions on shipment frequency, shipment size, return loads, and vehicle utilization rates). Alternatively, these decisions could be modeled explicitly in additional logistics modules. Also, other logistics aspects that are related to the trade-off between transport and inventory costs are usually not included in freight transport models, even though the logistics solutions of firms influence the mode split. Some models also need to produce the number of trucks for specific periods of the day, such as the morning peak and the afternoon/evening peak. Once again, this is usually done using fixed factors, not through explicit departure time choice modeling.

This brings us to the following list of missing components in the conventional four-step setup:

- Several logistics choices that individual firms are making: shipment size, use of consolidation and distribution centers; integration with production logistics;
- Scheduling: the choice of departure time (e.g., choice between peak and off-peak, day and night), as it is nowadays sometimes also added in national passenger transport models;
- Location choice for the firms and also for consolidation centers, distribution centers, and warehouses, as can be included in integrated land use-transport interaction models;
- Formation of multiproducer collection tours and multiconsumer distribution tours; however, this is much more relevant for urban freight models than for national freight models; and
- Treatment of interaction (e.g., between senders, carriers, and receivers; or even with the consumers in the household/passenger transport).

In some of these directions, considerable progress has been made in recent years; in particular, the integration of more logistics choices in freight transport modeling has been a thriving area (de Jong et al., 2012; Tavasszy et al., 2012).

A better representation of the freight transport model system in this context might be the following.

This structure has fewer steps: only three models—economic activity choices, logistics choices (including the choice of transport chain with a mode for each chain), and route choice—but each of the first two steps includes several related choices.

This model representation is consistent with the distinction between Producer-Consumer (PC) flows and Origin-Destination (OD) flows:

- PC flows are flows of goods from the location of production (P) all the way to the location of consumption. Between P and C, there could be a sequence of modes (transport chain) and vehicle types within modes, with transshipments, consolidation, distribution, and keeping inventories.
- OD flows are usually defined as flows of goods using the same mode. So every change of mode within a PC flow leads to a new OD flow.

This implies that the number of OD flows that corresponds to a single PC flow can be one or more. For instance, the PC flow with the transport chain “road-sea-road” has three OD flows. Trade data (e.g., statistics on export and import) usually are at the PC level, and transport statistics are usually collected for each mode separately and are, therefore, at the OD level. For most commodity flows, the natural decision level will be the PC flows, so that there is an optimization over the whole transport chain and not at the level of its components. However, there are also goods that are transported from the production location to an inventory and/or a port without knowing who the client at the consumption end will be. For such flows the optimization of the transport to and from the warehouse/port can be done separately, for example, by regarding them as separate PC flows (each of which could use a multimodal transport chain).

The ADA (aggregate-disaggregate-aggregate) model (de Jong and Ben-Akiva, 2007; Ben-Akiva and de Jong, 2013) follows the structure depicted in Fig. 2, and as such constitutes an alternative to the conventional four-step model, that allows for including more logistics choices. The basic representation of the ADA model is given in Fig. 3.

The ADA model consists of three submodels, two of which are at the aggregate (zonal) level, but one submodel (where disaggregation matters most) represents decisions at the level of flows between individual firms (disaggregate):

- The first aggregate submodel predicts PC matrices (this model is similar in structure to the generation and distribution models of the four-step model).

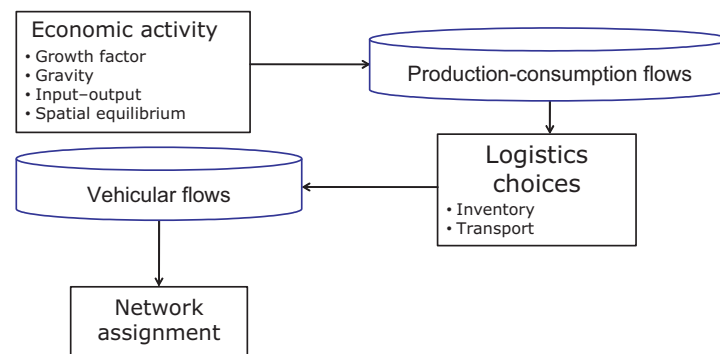


Figure 2 A revised representation of the structure of the freight model system.

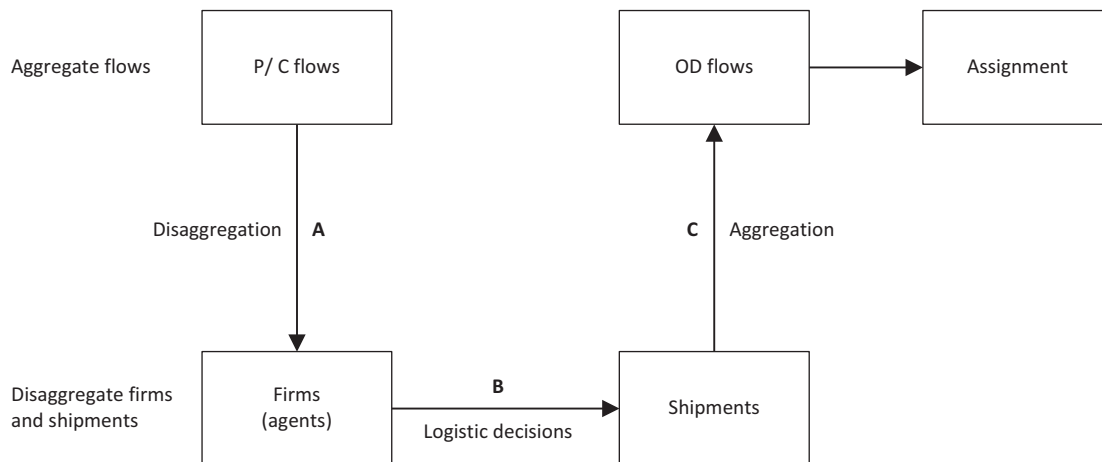


Figure 3 The ADA model.

- Then comes the disaggregate submodel 2, the logistics model:
 - This reads in the base matrices of commodity flows from producers to consumers: PC flows;
 - This represents the logistics decisions (especially transport chain and shipment size choice) at the level of annual firm-to-firm flows; and
 - This delivers OD matrices to the network model (assignment);
- The second aggregate submodel carries out assignment to networks (this model is similar to the assignment models in the four-step model).

The outputs of the first submodel, the PC flows, are at the level of annual flows between two zones, within a certain commodity group. In order to serve as input in the logistics model, these flows are allocated to firm-to-firm flows in step A (Fig. 3). Step B then is the actual logistics model, where the choice of transport chain from P to C and that of shipment size are made for each firm-to-firm flow by minimizing the logistics costs. Because the assignment to the networks needs to take place at the aggregate level for OD flows, the transport chains from step B are decomposed into OD movements and aggregated back to flows between zone pairs (step C), also taking into account empty return flows.

The logistics costs comprise:

- Order cost
- Distance-based transport cost
- Time-based transport cost
- Loading at sender; unloading at receiver
- Transfer at terminal
- Deterioration and damage
- Capital costs in transit (time)
- Inventory and capital cost at receiver
- Safety stock cost

Other solutions than the ADA model to introduce more aspects of logistics into the four-step model in national freight transport planning have been developed. These include:

- The aggregate, but very detailed in terms of sectors and commodities, model SMILE+ (Tavasszy et al., 1998) that was developed as a national freight transport model for The Netherlands in the late 1990s (the current Dutch national freight transport model BasGoed took a step back to the four-step structure, because the more sophisticated model became hard to maintain).
- The aggregate EUNET model that was developed for the United Kingdom that includes choice of distribution structure and multimodal transport chains.
- The disaggregate FAME model (Samimi et al., 2010) developed for the United States, and is rather similar in structure to the ADA model but includes a disaggregate model for the choice of supplier by the receiver of the goods (to predict the PC flows), so moving toward a disaggregate-disaggregate-aggregate (DDA) model.

Modeling Freight Generation and Distribution

Model types at the zone-to-zone level that can be used for the generation and distribution steps, or for the PC step of ADA-type model systems, are:

- gravity models;
- input–output (I/O) models; and
- spatial computable general equilibrium (SCGE) models.

A simpler approach specifically for generation is the use of sector-specific trip rate tables (or for tons of freight). Zonal trip rates for production and attraction are usually derived from classifying cross-sectional data on transport volumes to/from each zone in the area under investigation (or another similar area) into a number of homogeneous zone types.

Gravity-based models predict the flow of goods between two zones as a function of production and attraction of these two zones (e.g., represented by employment in specific sectors) and transport resistance between the two zones. For the estimation of these models, trade statistics (PC level) or transport statistics (OD level) are needed. The PC level is most appropriate for a gravity model, since this relates to the zones where these goods were actually produced and consumed (including further processing). A gravity model on transport statistics is less appropriate, since, for multimodal transport chains, the starting and ending points from the transport data may not be the actual locations that caused the goods flow, but just transshipment locations.

I/O tables describe, in money units, what each sector of the economy (e.g., steel manufacturing) delivers to the other sectors of the economy (e.g., car manufacturing). The final demand by consumers, imports, and exports is also included in these I/O tables. National I/O tables have been developed for many countries, usually by the central statistical office of the country (e.g., Statistics Sweden). I/O models are macroeconomic models that use the production structure links between sectors as described by the I/O

tables to predict how a change in demand or production or consumption in some sector influences the other sectors. The multiregional or spatial input–output (MRIO) table is a special form of I/O table, which not only includes deliveries between sectors, but also between sectors in different regions (interregional trade flows). Not all countries have up-to-date national I/O tables, and recent multiregional I/O data are only available for a few countries. An I/O model that takes account of the spatial structure of the links between sectors (MRIO model) can be used to predict trade flows between zones. In doing this, some I/O models also take transport distance, time, and/or costs between the zones into account.

Another option for production and attraction is the computable general equilibrium (CGE) model that establishes equilibrium in several related markets (not only transport, but also goods markets, labor markets, and land markets). They require similar data as I/O models: I/O tables or Make and Use tables from the national accounts data, preferably multiregional (otherwise, one might carry out a rationalization). CGE models in economics (not focusing on transport) often include economic issues that are not handled in transport models, such as type of competition and economies of scale. Spatial CGE (SCGE) models have been developed more recently, and could become operational models for practical (inter)national and regional forecasting and evaluation in years to come.

In order to make distribution models dependent on transport costs and time between zones, one needs to include time and distance data from transport network skims (possibly combined with cost functions), or accessibility indicators (such as logsums) from mode choice models.

There are hardly any disaggregate models for generation and distribution in freight transport.

Modeling Mode Choice and Route Choice in Freight Transport

Aggregate mode choice models can be estimated on transport statistics data by mode, which are available in many countries. For the estimation of disaggregate mode choice models, one needs shipper surveys, stated preference surveys with mode choice experiments or consignment bill data. Only few countries have such data sources available to researchers. Shipper surveys can also be used for models of shipment size or for joint models of mode and shipment size.

In the ADA models and other model systems where generation and distribution are carried out at the PC level, the corresponding mode choice component is a transport chain choice model. The transport chains in this model can be unimodal or multimodal. The limited direct observations that we have of transport chains come from disaggregate shipper surveys (including commodity flow surveys).

Both for a mode choice and for a transport chain choice model, one also needs network data (time and cost by mode) and cost functions. For a transport chain choice model, data on the locations of terminals for transshipment is also required.

Network assignment models typically do not use information on observed choices or market shares, but use a deterministic rule to assign vehicles to a shortest path. The same modeling philosophy that is routinely used in network assignment can also be used for other choices where information on observed choices is missing. In the absence of information about transport chains, one might use a deterministic model that predicts the transport chain choice on the basis of the minimization of the full logistics costs. However, the outcomes of this are normative and not necessarily realistic. The latter problem can be reduced by a calibration of the model to other data, such as mode shares at the OD level from transport statistics.

A number of freight transport models at the national or international level use a simultaneous assignment to modes and routes: multimodal assignment. This requires that the networks of the various modes are connected to each other to represent the multimodal terminals: the nodes where transshipments between different modes can take place. The existing multimodal assignment models in freight all work at the aggregate level.

Short Overview of National Freight Transport Models

In **Table 1**, we list the currently used model approaches in national freight transport modeling and provide some examples from practice for each approach (de Jong et al., 2012). The same approaches are not only used for national models, but also for international models, such as transport models systems for Europe as a whole (Transtools1, 2, 3, TRIMODE, LOGIS) or world models (Worldnet) and for states/provinces and metropolitan regions within larger countries (e.g., the MetroScan model for the

Table 1 Practical examples of different approaches in national freight transport modeling

<i>Model approach</i>	<i>Model</i>
Aggregate four-step	Dutch national model (BasGoed), Flanders strategic model, France (MODEV), Great Britain freight model (GBFM)
Aggregate multimodal assignment	Belgian model for the Walloon region (NODUS)
Aggregate model with logistics	Previous Dutch national model (SMILE+), EUNET model for the United Kingdom
Partly disaggregate, optimization	National models of Sweden, Norway, and Denmark, Model for Mobility Masterplan Flanders (= the ADA family)
Partly disaggregate, statistically estimated on observed data	German Federal Model, US national model FAME, US national freight model for FHWA (prototype under development), Danish model for Fehmarn Belt corridor, prototype stochastic national model for Sweden, and planned national model for Austria

Sydney area or the models for Alberta or Toronto in Canada, or Oregon or Chicago in the United States). The category that is used most in practice is the aggregate four-step model. Within the partly disaggregate model systems, there are models that use a deterministic minimization of logistics costs to find the transport chains and shipment sizes, and there are models that are statistically estimated on data from a disaggregate data source such as a commodity flow survey or a stated preference survey.

Conclusions

National freight transport models have been used for practical transport planning by the national transport authorities at least since the late 1980s/early 1990s, for scanning the future under various exogenous scenarios and for providing the transport inputs in the appraisal of transport projects. Since these early models, several new developments have taken place in freight transport modeling at the national and related spatial scales. Many of these new developments were related to attempts to include more logistics choices in the modeling framework. In spite of these new developments, the approaches that were used in the earliest national models (aggregate four-step freight transport models) are still the most common methods in practice. The main reasons for this seem to be a lack of data (especially on transport chains and, more generally, disaggregate data) and high costs of new data collection. It is hoped that new electronic data collection technologies, in combination with older survey methods, can lead to improvements in this area.

References

- Ben-Akiva, M.E., de Jong, G.C., 2013. The aggregate-disaggregate-aggregate (ADA) freight model system. In: Ben-Akiva, M.E., Meersman, H., van de Voorde, E. (Eds.), *Recent Developments in Transport Modelling: Lessons for the Freight Sector*. Emerald, Bingley.
- de Jong, G.C., Ben-Akiva, M.E., 2007. A micro-simulation model of shipment size and transport chain choice. *Transp. Res. B Methodol.* 41, 950–965.
- de Jong, G.C., Gunn, H.F., Walker, W., 2004. National and international freight transport models: overview and ideas for further development. *Transp. Rev.* 24 (1), 103–124.
- de Jong, G.C., Vierth, I., Tavasszy, L., Ben-Akiva, M., 2012. Recent developments in national and international freight transport models within Europe. *Transportation* 40 (2), 347–371.
- Samimi, A., Mohammadian, A., Kawamura, K., 2010. Freight Demand Microsimulation in the US. *World Conference on Transport Research (WCTR)*, Lisbon.
- Tavasszy, L.A., de Jong, G.C., 2014. *Modelling freight transport*. Elsevier Insights Series. Elsevier, London/Waltham.
- Tavasszy, L.A., Ruijgrok, C.J., Davydenko, I., 2012. Incorporating logistics in freight transportation models: state of the art and research opportunities. *Transp. Rev.* 32 (2), 203–219.
- Tavasszy, L.A., Smeenk, B., Ruijgrok, C.J., 1998. A DSS for modelling logistics chains in freight transport systems analysis. *Int. Trans. Opera. Res.* 5 (6), 447–459.

Further Reading

- Beuthe, M.B.J., Geerts, J.-F., Koul à Ndjang Ha, C., 2001. Freight transport demand elasticities: a geographic multimodal transportation network analysis. *Transp. Res. E* 37, 253–266.
- Bröcker, J., 1998. Operational spatial computable general equilibrium models. *Ann. Reg. Sci.* 32, 367–387.
- Holguín-Veras, J., 2008. Necessary conditions for off-hour deliveries and the effectiveness of urban freight road pricing and alternative financial policies in competitive markets. *Transp. Res. Part A* 42 (2), 392–413.
- Holguín-Veras, J., Aros-Vera, F., Browne, M., 2015. Agent interaction and the response of supply chains to pricing and incentives. *Econ. Transp.* 4 (3), 147–155.
- de Jong, G.C., Tavasszy, L., Bates, J., Grønland, S.E., Huber, S., Kleven, O., Lange, P., Ottemöller, O., Schmorak, N., 2016. Issues in modelling freight transport at the national level. *Case Studies Transp. Policy* 4, 13–21.
- Liedtke, G., 2009. Principles of micro-behavior commodity transport modelling. *Transp. Res. E* 45 (5), 795–809.
- Roorda, M.J., Cavalcante, R., McCabe, S., Kwan, H., 2010. A conceptual framework for agent-based modelling of logistics services. *Transp. Res. Part E* 46 (1), 18–31.

Freight Flows in Cities

Genevieve Giuliano, Department of Urban Planning and Spatial Science, University of Southern California, Los Angeles, CA, United States

© 2019 Elsevier Ltd. All rights reserved.

Introduction	168
The Urban Freight Problem	168
Air Pollution and GHG Emissions	168
Congestion	169
Conflicts and Safety	169
Two Types of Urban Freight Flows	170
Local Freight Demand	170
Trade-Related Freight Demand	170
Two Major Trends	170
Spatial Organization of Warehousing and Distribution	171
The Rise of E-Commerce	172
The Problem of Free Shipping	173
Strategies for Managing Urban Freight	174
Managing Trade-Related Freight	174
Managing Local Freight Demand	175
Conclusion	176
Global Trade: Might Ever Greater Integration Be Coming to an End?	176
Might the Consumption of Material Goods Decline?	176
References	177
Further Reading	177

Introduction

There was a time when freight, whether moving in cities or outside, was the purview of logistics and transport economics. No more. Freight is a major component of urban transport flows. It impacts and interacts with passenger transport, generates many externalities, and influences the changing spatial organization of cities. Freight demand is largely an urban demand: producers and consumers are concentrated in metropolitan areas, the economic engines of the global economy. As the global economy has become more liberalized, supply chains have expanded, and regional networks of production, tied together via transport, have emerged. Large cities are the major nodes in these regional networks.

At the same time, the growth in consumption associated with rising per capita income generates more freight flows in getting goods to and from local markets. Online shopping and instant deliveries only add to this growth. As a result of these trends, urban freight demand has generated interest within urban planning, geography and civil engineering, in addition to economics and logistics. This new interdisciplinary field seeks to understand the dynamics of urban freight, its relationship with urban spatial structure, and to generate effective strategies for managing urban freight and increasing its sustainability.

This chapter provides an introduction and overview of urban freight. First, the major impacts of freight are discussed, including congestion, air pollution, and safety. Second, urban freight is categorized into two types: trade related and local. Third, two trends affecting urban freight are described: the changing spatial organization of warehousing and distribution, and the rise of e-commerce. Fourth, an overview of freight mitigation is presented. The chapter closes with some considerations on longer term trends that may bring further changes to urban freight.

The Urban Freight Problem

Freight generates many problems for metropolitan areas, yet the efficient movement of freight is essential for economic and social well-being.

Air Pollution and GHG Emissions

Freight (primarily trucks) accounts for a significant share of air pollution and GHG emissions. Specific contributions vary by country, as they depend on the age and technology of the fleet, as well as the total volume of emissions. In the United States, the transport sector accounts for nearly 60% of NOX and 22% of VOC, the precursors to smog and ozone. Trucks account for about one-

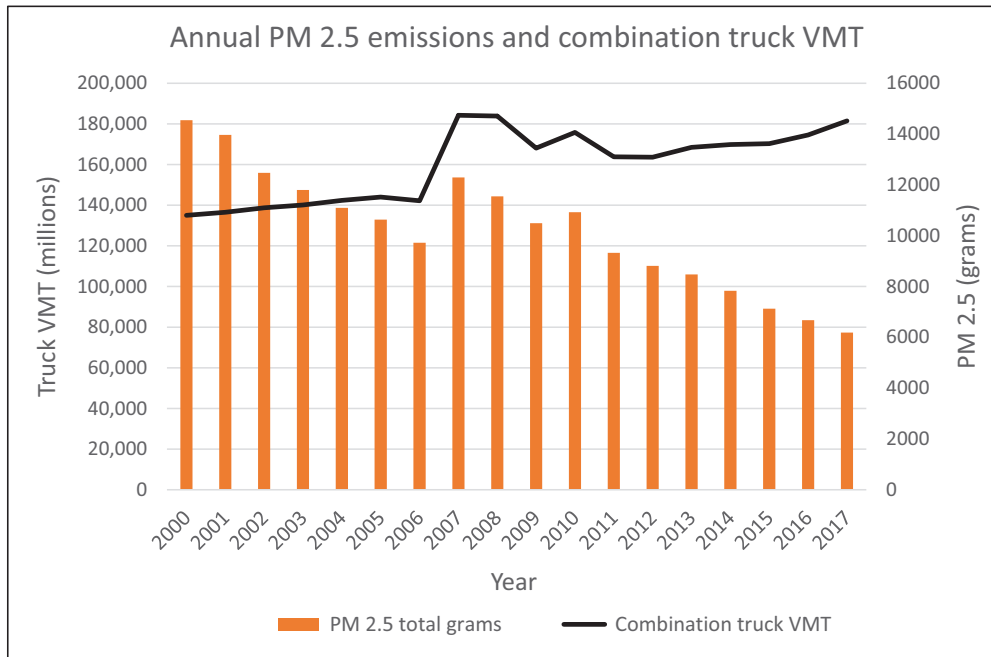


Figure 1 Annual PM 2.5 emissions and combination truck VMT. Source: Calculated by author from USDOT Bureau of Transportation Statistics data.

third of the transport share. The regulation of engines and fuels has resulted in lower emission rates. To date, declining emissions rates have offset the growth of truck VMT (vehicle miles traveled). For example, [Fig. 1](#) shows the annual estimated PM_{2.5} emissions and combination truck VMT from 2000 to 2017. The increase from 2006 to 2007 is due to a change in the method used to calculate VMT.

The US transport sector accounts for 29% of GHG emissions. Within the on-road transport sector, freight accounts for 23% of all GHGs. As other sectors reduce GHGs, the transport share is increasing. The contribution of trucks is increasing more than that of other transport modes. From 1990 to 2017, the total GHG transport sector emissions increased by 22%, but truck emissions increased almost 90% ([US EPA, 2019](#)). The truck share is increasing due to the growth in overall freight volumes, the continuing trend toward using faster modes, rising congestion, and the technology challenges of developing cleaner medium and heavy duty vehicles.

Congestion

Freight contributes disproportionately to congestion. Large trucks are more difficult to maneuver and are slow to accelerate and decelerate, adding delays to the surrounding auto and bus traffic. In city cores, limited parking and loading facilities result in illegal parking (in bicycle lanes or at bus stops, or double parking in the roadway).

It is difficult to identify the contribution of freight activity to the overall level of congestion due to data limitations. [Giuliano et al. \(2018\)](#) developed the concept of “freight impact area” to identify road segments with high levels of freight-related congestion. Applying the concept to the Los Angeles region, they identified and ranked freight impact areas based on total peak period congestion and the share of trucks in the traffic flow. The top 15 freight impact areas were located along major corridors connecting intermodal facilities, industrial, and warehouse zones. The top 15 together account for 53,615 h of vehicle delay each PM peak period, or about 14 million hours per year.

Conflicts and Safety

Freight conflicts with passenger transport for limited road and sidewalk capacity, creating safety hazards. As cities become more dense and pedestrianized, safety hazards increase. Freight is known to get delivered no matter the circumstances, whether it requires parking on a sidewalk, double parking, or blocking a bike lane or transit stop. Local safety data is not widely available. [Table 1](#) gives the number and share of deaths of motorcyclists, bicyclists, and pedestrians from crashes involving large trucks from 2000 to 2018 in the United States. The drop in fatalities in 2010 is explained by the Great Recession. The total number of deaths has decreased since 2000, but deaths of motorcyclists, bicyclists, and pedestrians have increased. Consequently, these mode users represent an increasing share of all fatalities. This trend is an indicator of more urban truck traffic, where conflicts with bicyclists and pedestrians are more likely. A recent study examined crashes by location—urban/nonurban, on or off interstate highway—and found that the share of freight-related crashes with nonfatal injuries taking place in urban areas off interstates is increasing, another indicator of growing freight safety problems in cities ([McDonald et al., 2019](#)).

Table 1 Deaths of motorcyclists, bicyclists and pedestrians from crashes involving large trucks, 2000–18

<i>Year</i>	<i>No. of deaths, motorcyclist, bicyclist, pedestrian</i>	<i>No. of deaths, all</i>	<i>Percent motorcyclist, bicyclist, pedestrian</i>
2000	490	5173	9.5
2005	595	5049	11.8
2010	445	3418	13.0
2015	568	3876	14.6
2018	619	4136	15.0

Source: Calculated from US DOT/NHTSA Fatality Analysis Reporting System, <https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars>.

Two Types of Urban Freight Flows

Metropolitan freight activity may be described in terms of two main types: freight related to local supply or demand, and freight related to national or international trade.

Local Freight Demand

Freight related to local supply and demand is often termed “last-mile” freight: local supply and demand is typically the first or last move in complex regional and global supply chains. Local freight activity includes delivery or pickup of imports/exports, intra-metropolitan trade of commodities (local production and consumption), construction, waste removal, and shipments between warehouse and distribution facilities, retail facilities, and homes and business.

Local freight demand is a function of population and employment; as metropolitan areas grow, so does local freight demand. It is also a function of per capita income. Thus, as metropolitan areas become richer, the demand for local goods and services increases. Other trends are accelerating local freight activity, notably the rise of e-commerce and higher velocity supply chains. E-commerce, or online shopping, spreads freight demand throughout the city, leading to more freight deliveries. Higher velocity supply chains mean less inventory and more frequent deliveries. In city cores, restaurants, retail shops, offices, and even apartment dwellers minimize storage space in response to high rents.

Trade-Related Freight Demand

Freight related to national or international trade is a function of the integration of the global economy, itself a result of advances in information and telecommunication technology and trade liberalization policies (Dicken, 2007). Production supply chains have become more complex as producers seek out comparative advantage opportunities around the world. This process has resulted in consistent growth in cross-border trade for the last several decades. For example, total imports grew by 92% from 2000 to 2017 while total exports grew by 98% in the United States (US Census, 2018).

Cities that serve as trade nodes have an additional layer of freight and logistics demand, because they serve both local and nonlocal supply and demand. Table 1 gives an example from California. Los Angeles and San Francisco are well-known international trade centers, while Sacramento and San Diego are not. Table 2 gives domestic and international commodity flows in US dollars per capita. Domestic flows have a narrow range, from about 35,000 to 41,000. In contrast, international flows are much higher for Los Angeles and San Francisco. Although international flow is about half as much as domestic flows for these cities, the volume of trade adds significantly to overall freight demand. This added demand implies more congestion, air pollution, and goods movement facilities.

Two Major Trends

Urban freight is dynamic; supply chains change constantly in response to prices and consumer demand. In this section, two changes are described that are affecting both the volume of urban freight flows and their spatial organization.

Table 2 Commodity flow per capita by Freight Analysis Framework area

<i>Metro area</i>	<i>Domestic flow/capita (US\$)</i>	<i>International flow/capita (US\$)</i>
Los Angeles	35,331	17,036
San Francisco	37,725	18,196
San Diego	37,564	4992
Sacramento	41,013	3450

Source: Calculated from 2007 Commodity Flow Survey.

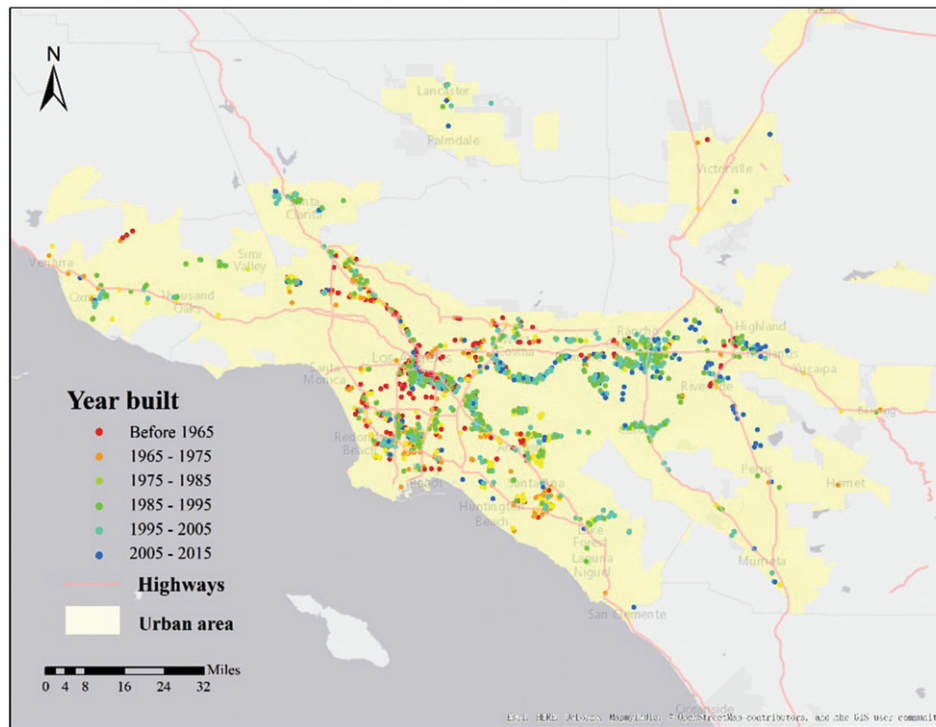


Figure 2 Warehouse location by year built, Los Angeles region. Source: *Giuliano and Yuan (2017)*.

Spatial Organization of Warehousing and Distribution

Significant changes have taken place in the warehousing industry during the last decade. First, the growth of warehousing and distribution activity has been greater than that of the general economy (*Giuliano and Kang, 2018*). Second, warehousing facilities are increasingly specialized, for example, in transshipping, packaging, or inventory management (*Akman and Baynal, 2014*). They serve a wide range of customers, including 3PL logistics providers, trucking and freight forwarders. Third, warehousing facilities serve more geographically dispersed markets (*Hess and Rodrigue, 2004*) as economies of scale intensify within the industry. Fourth, warehouses provide more frequent deliveries as retail businesses seek to reduce inventory and save on rent costs. Finally, the scale of warehousing and distribution has increased due to automation and high-volume product flows. Although increases in e-commerce and consumer demand for instant deliveries are creating demand for close-in distribution or fulfillment centers, these trends have not reduced demand for largescale warehousing. Rather, distribution chains are changing.

The result of increased warehousing activity, scale, and demand has been the decentralization of warehouse and distribution facilities, as documented by studies of several large metropolitan areas in the United States, Europe, and Japan. *Fig. 2* is an example from the Los Angeles region. Warehouses are color coded by year built. It can be seen that the newest facilities are generally located far to the east of the urban core.

One US study of warehouse location from 2003 to 2013 in the 48 largest metropolitan areas revealed the following: (1) warehousing decentralized by about 1 mile, measured as the average distance of all warehouses to the CBD; (2) large warehouses decentralized more than smaller warehouses, and (3) large warehouses decentralized more in the largest metropolitan areas. Decentralization is largely a function of land values. The study used the population density gradient as a proxy for land value, and found that decentralization is negatively associated with the steepness of the gradient and positively associated with peak density (*Kang, 2017*).

The shift of warehousing and distribution activity to larger facilities far from the urban core suggests more concentrated freight traffic as the scale of distribution increases and distribution networks become larger. Spatial clusters of distribution facilities add to this concentration.

The emergence of clusters of large warehouses (over 1 million square feet) has led to questions of environmental justice. Warehouses are major truck trip generators and, hence, major sources of air pollution. Truck traffic generates noise and pavement damage as well as traffic safety threats. An extreme example of warehouse development is the Highland Fairview Logistics Park in Moreno Valley, California. It is the home of the Sketchers warehouse, a 2 million square foot facility that is over ½ mile in length (see *Fig. 3*). It is part of a larger World Logistics Center that is proposing 40 million square feet of warehouses.

The environmental justice question is whether warehouses are disproportionately located near low income or minority neighborhoods. There are three arguments for why this may be the case. First, warehouse developers seek places with low land prices and access to low wage labor, and such places are often where low income or minority households reside. Second, disadvantaged



Figure 3 Sketchers warehouse, Moreno Valley, California. *Source: Laetitia Dablanc photo collection, by permission.*

populations have less political power, and hence less capacity to resist locally undesirable land uses. Third, discrimination in housing markets constrains location choice, especially for poor minority households.

A study of four metro areas in California reveals a more complex relationship: warehousing location is associated with medium and low income minority neighborhoods in Los Angeles, San Francisco, and Sacramento, and only with medium income minority neighborhoods in San Diego (Giuliano and Yuan, 2017). A possible explanation for these findings is that low income neighborhoods are not necessarily located in areas with low land prices and adequate transport access required by large facilities. This work was extended to test causality: do households follow warehouses as they search for cheaper housing, or do warehouses follow households as they search for cheaper land? Results showed the latter, suggesting that warehouse location is an environmental justice problem in the four California metro areas (Yuan, 2018).

The Rise of E-Commerce

E-commerce is defined as any type of consumption that takes place via an online platform. The most common type of e-commerce is online shopping; consumers order and pay online, and the goods are delivered to the consumer's residence. There are variations: the consumer may order online and pick up at a retail establishment, or at a local locker facility. E-commerce includes all types of consumption, from meal deliveries to cat litter to furniture. The emergence of online shopping has transformed where and how goods are produced, distributed and sold, and how consumers make shopping, as well as shopping travel, decisions (Mokhtarian, 2004).

The rise in e-commerce has been rapid. In the United States, the market share of online shopping has increased from about 3.7% in 2008 to 9.5% in 2018. The annual rate of growth of online shopping sales has been around 11%–13% since 2011, much greater than the rates for total retail sales, 3%–4% (US Census, 2019). Global online sales as of 2018 are estimated to be about \$2.5 trillion and account for about 9% of all retail sales. Some countries have much higher shares of e-shopping: 18% for the United Kingdom and 16.6% for China as of 2018 (Investp, 2020). E-shopping is expected to continue to grow, with great uncertainty with regard to when the rate of growth will slow.

E-commerce is also changing rapidly. The variety of goods available continues to grow, and some new products have emerged, such as ingredients and instructions for home prepared meals (e.g., Blue Apron). Speed of delivery is also increasing. Some firms now offer "instant deliveries" (within 2 h), and one day delivery is now routine in many metropolitan areas. The growth of e-commerce is illustrated in Fig. 4, which shows trends in mail volume and package volume in billions of units. Mail volume has declined consistently from 2007, while parcel volume has increased. For the same period, the total number of parcels increased from 7.2 to 17.8 billion, a 140% increase.

E-shopping has many impacts on cities. First, e-shopping affects travel behavior. If shoppers simply substitute an online purchase for an in-store purchase, travel is reduced. However, shoppers may behave in many different ways. Shoppers may inspect merchandise in stores and then purchase online, or the reverse. Shoppers may shop more overall because of the convenience of online purchases. Even if shoppers make fewer shopping trips, they may use the extra time to travel for other purposes. The mode of travel is also an important consideration. If the trip to the store is made by walking or biking, there is no environmental benefit, even if it is

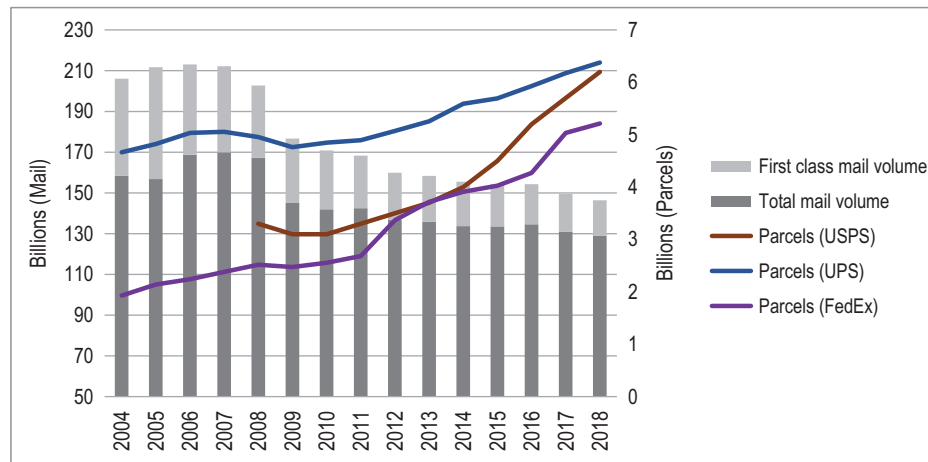


Figure 4 USPS mail, parcels carried by major carriers, US, 2004–18. Source: Rodrigue, J.-P., 2020. *The Geography of Transport Systems*, fifth ed. Routledge, London. Available from: https://transportgeography.org/?page_id=1698.

fully replaced by online shopping. At this time there is insufficient evidence to draw any conclusions on travel behavior impacts (Cao, 2009).

Second, e-shopping affects goods distribution. E-commerce requires delivery to a residence or local common pickup point. Small-scale deliveries (small packages in small trucks) are less efficient than the large scale deliveries made to retail establishments. Although e-shopping eliminates at least some large-scale deliveries (due to the loss of in-store business), these losses will be more than offset by the added truck travel generated by small scale deliveries. Efficiency of freight deliveries is further reduced by the rate of failed local deliveries and the higher rate of returns associated with online shopping. The outcome is increased truck VMT (Rotem-Mindali and Weltevreden, 2013).

One day and instant deliveries also affect the distribution supply chain. In order to offer one day or less delivery, retailers must have enough inventory and be close enough to consumers to be able to fulfill, package, and deliver an order in a matter of hours. Technology and predictive analytics help to stock fulfillment centers with goods in anticipation of orders, yet the entire system depends on fast transport, from warehouse to fulfillment center to customer. In highly congested metropolitan areas, these facilities must locate closer to customers.

This leads to the third impact on cities: the changing demand for retail and industrial space in metropolitan areas. One of the most visible impacts of changing shopping behavior in the United States is the failure of shopping malls. From 2012 to 2017 nearly 20% of regional malls were closed down (ICSC, 2013, 2017). E-shopping is a contributing factor. The traditional mall is organized around anchor stores—the major department store chains (e.g., Macy’s, Sears)—with smaller stores between. The anchor stores provide the customers for the smaller specialty stores. As department stores consolidated and closed stores, the traditional structure was no longer viable, and the weaker malls failed. The trend continues, and it is expected that 20%–25% of existing malls will close by 2022 (Credit Suisse, 2018). In other countries, the growth in online shopping is associated with a decline in in-store shopping; hence generating declining demand for retail space.

At the same time, the continued growth of online shopping and instant deliveries is generating increased demand for in-city warehouse and distribution facilities, despite higher rent and operating costs. Amazon and other online retailers are engaging in adaptive re-use to develop local instant delivery networks. Examples include the conversion of a printing facility in Barcelona and a parking structure in Paris (see Fig. 5). In the United States, failed malls have become distribution centers, and retailers such as Target are using their extra space as local micro-distribution centers and pick up points. The transportation implications of these shifts are significant: more truck traffic in local residential and commercial areas, truck traffic replacing auto traffic, and more complex distribution chains.

E-shopping is having additional impacts on the built environment. For example, the traditional design of an apartment building includes a reception area and tenant mailboxes. There are no provisions for large numbers of package deliveries, which results in reception areas becoming de facto parcel storage areas. New apartments are being built with parcel storage areas and communications systems to inform tenants of deliveries. In an effort to reduce the cost of home deliveries, Amazon and others are establishing “pickup points” and lockers to consolidate home deliveries. Pickup points are common in many countries, including France and Korea. They are often located at transit stations so that shoppers may pick up parcels on the way home from work or other activities, but also may be located in convenience stores, banks, or other places.

The Problem of Free Shipping

One of the motivators of the growth in e-shopping is free delivery. Although consumers value the convenience of online shopping, they are also price sensitive. Paying for shipping makes online purchase less competitive with the in-store option. Online retailers



Figure 5 Amazon fulfillment center, Barcelona. *Source: Laetitia Dablanc photo collection, with permission.*

have responded by offering free shipping, first for standard shipping (3 days or more), then for one or two day shipping, and now often for instant shipping. Rather than competing on product price, online retailers are competing on free shipping.

Free shipping is of course not free. It is a subsidy to consumers; the delivered product is priced below cost. As with any other good that is priced below cost, more will be consumed than economically optimal. Without free shipping, at least some online sales would be diverted to in-store purchases. In addition, free shipping is an incentive for consumers to be wasteful. Under most circumstances, a consumer would not travel to a store for one small item, for example, a pair of socks. Rather, the trip would be made when the total purchase is worth the trip. With free shipping, there is no penalty for single item purchases. The consumer has no reason to bundle purchases, or to delay their satisfaction in obtaining the item. The fragmentation of purchases adds more inefficiency to freight deliveries, generating more truck traffic and pollution.

Strategies for Managing Urban Freight

Cities around the world are employing a vast array of strategies to mitigate negative impacts and better manage freight flows. It is not possible to provide a comprehensive discussion of the many strategies considered or employed around the world. This section provides a brief overview, along with a few specific examples.

Managing Trade-Related Freight

Strategies for managing trade-related freight are summarized in [Table 3](#). Strategies are organized into four categories, and examples of each are presented. Infrastructure strategies tend to be effective, because they add some form of freight capacity to the system. However, high costs, the lack of available space in urban areas and local environmental impacts often preclude such strategies. For example, truck-only lanes would require highway expansion and a design that would prevent conflicts between trucks and passenger vehicles. Corridors most in need of truck lanes are located in dense urban areas, where the taking of additional land is most

Table 3 Strategies for trade-related freight

Type of strategy	Infrastructure	Efficiency	Emissions	Pricing
Examples	Truck only lanes Truck bypass facilities Railroad grade separations Added highway capacity On-dock rail Inland ports	Advanced freight traffic management systems Integrated freight information systems Freight priority traffic management Cargo matching services Smart truck parking Terminal appointment systems Truck platoons	Low emission fuel standards Zero and near zero emission trucks Low carbon fuel standard Mode shifts	Truck and passenger VMT tax Congestion tolls Truck lane tolls Port cargo or gate tolls Carbon tax Carbon cap and trade system

problematic. In addition, truck only lanes would attract more truck traffic to the corridor, increasing the environmental burden on the nearby population.

Efficiency improvements seek to take advantage of advancements in information and communication technologies. Connected vehicle systems make it possible to manage traffic in real time, increasing safety and system efficiency. Efficiency improvements are attractive because many have relatively low cost and increase the productivity of the existing infrastructure. Innovation is taking place in both public and private sectors. Governments are developing various types of freight traffic management strategies such as route guidance or dynamic traffic signal operations. The private sector is developing platform-based services that seek to efficiently match buyers and sellers. As connected and automated vehicle technology continues to develop, trucking is expected to be among the early adopters. As of 2020, semi-automated trucks are being operated in pilot studies along the I-10 corridor from Arizona to Texas, and truck platoons have been demonstrated in several US states.

By far, the most effective emissions reduction strategy to date has been regulation based on emissions and fuel consumption standards. In the United States, the first national emissions standards were established in 1973. Emissions of four criteria pollutants declined by half or more by 2015, despite vehicle travel tripling over the same period (US EPA, 2018). Zero emission technology is currently limited to battery electric trucks. Hydrogen fuel cell technology has promise, but remains experimental. Battery technology is improving, making possible some penetration in the medium duty market. Heavy duty trucks remain a challenge due to the limited range and long charging times for batteries large enough to haul a heavy load.

Pricing is the most effective strategy for solving externality problems of all types, yet it remains politically difficult to implement. Pricing provides clear incentives for behavioral change and provides revenue for infrastructure investment. For example, a truck VMT tax could be used to fund highway rehabilitation, or to subsidize the purchase of electric trucks.

Text box reference: Zhao, Y., Ioannou, P., 2019. A co-simulation, optimization, control approach for traffic light control with truck priority. Annu. Rev. Control 48, 283–291.

Intelligent Traffic Priority for Trucks

Could we reduce traffic delay for all vehicles if we give traffic priority to trucks under some circumstances? It is well known that heavy trucks have disproportionate impacts on traffic due to their large size, slow braking and acceleration, and wide turning radius. Traffic signal timing and control is based on the performance characteristics of passenger vehicles. In a dynamic system, signals are timed to minimize delay based on passenger vehicle parameters. However, every time a truck stops at a signal, the following vehicles are slowed down by the slow acceleration of the truck when the light turns green, and it takes more time for all the waiting traffic to clear the intersection. If the truck were able to move through the intersection without stopping, the delays behind the truck would not occur.

Researchers have developed methods to manage traffic signals in real-time, giving priority to trucks when doing so leads to overall reductions in delay. Simply setting a principle of giving trucks signal priority is not sufficient, as a net reduction in delay depends on traffic patterns and composition. These methods were tested via simulation modeling using traffic data in an area around the ports of Los Angeles and Long Beach. Results show that truck priority results in reduced delay for all travelers. The reduction in delay also reduces truck emissions and energy consumption. Vehicle-infrastructure technology will allow for such customized versions of traffic management.

Managing Local Freight Demand

Strategies for managing local freight have an added category, regulation, as regulation plays such a critical role in urban freight. Infrastructure strategies are much smaller in scale, reflecting the more local focus of last-mile pickups and deliveries (see Table 4). Truck lanes may be demarcated to clearly identify major truck routes through the city and prevent conflicts with other traffic. Adequate parking and loading facilities reduce parking violations. Street and sidewalk design can be used to accommodate truck traffic where needed, while preserving space for nonmotorized transport. In-city consolidation centers facilitate more full truck movements.

There are many options for efficiency improvements beyond advanced traffic management. Deliveries to retail establishments or warehouses can be shifted to evening hours, moving truck traffic out of crowded daytime hours. Pickup points reduce the number of residential deliveries.

Table 4 Strategies for managing local freight demand

Type of strategy	Infrastructure	Efficiency	Emissions	Pricing	Regulation
Examples	Truck lanes Truck parking and loading zones Flush curbs Sidewalk cutouts Consolidation centers	Advanced traffic management systems Freight priority traffic management Off hour deliveries Lockers and pickup points	Zero and near zero emission trucks Cargo bikes Low emission zones	Cordon tolls Loading zone and parking fees Licenses	Truck size and weight restrictions Parking enforcement On-site loading docks

Text box references: Holguín-Veras et al., 2018; Sanchez-Díaz et al., 2016.

Off-Hour Deliveries

A major source of urban congestion is commercial deliveries to restaurants, stores, hotels, etc. Traditionally, these deliveries take place during regular business hours (during the day) so that labor is available to receive the shipments. Shifting these deliveries to night time hours, termed off-hours deliveries (OHD), would generate many benefits. Reduced daytime truck traffic would reduce congestion, truck emissions, and energy consumption. Vehicles and delivery infrastructure would be more efficiently used, and delivery times would be less uncertain. Given the potential benefits, it is logical to ask why OHD is not widely practiced. The main reason is cost: receivers must employ a crew for an additional shift in order to accept the deliveries. Another option is to deliver to a secure location, with the truck driver completing the delivery. However, receivers (cargo owners) may be reticent to trust others to safely access and egress facilities and to forego checking the order before acceptance. Other challenges include noise impacts on neighbors, and desire of drivers to work the daytime shift.

The New York State Department of Transportation funded a large, multiyear OHD pilot project in New York City in order to determine how a successful OHD program could be implemented. Years of research revealed that with the right incentives (subsidies to offset added labor costs, shipping discounts, and public recognition) businesses were willing to participate. By 2016 there were 400 participating businesses. The New York City pilot demonstrated that OHD benefits are widespread. Reduced daytime truck traffic and associated reduced congestion generates benefits for both passenger and freight traffic.

The New York City pilot has led to implementation of OHD programs around the world. Sao Paulo has the largest program, with over 700 participants. London, Copenhagen, Stockholm, as well as Bogota, Medellin, and Cartagena (Columbia) have operating OHD programs as of 2020.

Pollution impacts are of particular concern in cities, as human exposure is concentrated. Thus, extensive experimentation is taking place with zero emission vehicles. In very dense urban areas, cargo bikes (electric powered bicycles with a cargo hold) may deliver the last few meters. Low-emission zones have been established in several cities. For example, Paris requires heavy duty trucks to be registered as meeting Euro V standards in order to gain access between 8 AM and 8 PM on weekdays. In California, several zero and near zero emission heavy duty truck demonstration projects are taking place to promote the use of ZE trucks in short haul service. California is also developing a “ZEV mandate” for trucks that will require truck manufacturers to sell a minimum percent of ZE trucks as part of new truck sales.

Pricing is also relevant at the local level. Cordon tolls began in Singapore in 1975 in order to manage traffic in the central core. London launched cordon tolls in 2003 in response to debilitating traffic congestion, and has since expanded the system to incentivize use of low emission trucks in the central area. In contrast to the Paris example, London uses pricing to promote cleaner trucks. The London low emission zone (LEZ) was introduced in 2015 and encompasses all of Greater London. Trucks must meet a specified Euro standard in order to enter the zone without paying a fee. The standard is graduated. Currently, the standard is Euro IV; in October 2020 trucks will have to meet Euro VI.

Local governments have regulatory authority that can be used to manage freight. Building permits can require construction of on-site loading docks so that the use of curb space is reduced. Codes can require ample parcel or inventory storage space.

Conclusion

This chapter has provided an overview of a relatively new field of transport research. It is a field that is changing rapidly as global economics and politics affect trade patterns, as technology changes transportation and as consumer behavior evolves to take advantage of more convenient travel and consumption alternatives. The topic is concluded with some observations on current emerging trends that may bring additional changes to urban freight.

Global Trade: Might Ever Greater Integration Be Coming to an End?

Global trade has expanded as a result of technology and trade policy. It is not inevitable that this expansion will continue. First, some countries are moving away from multilateral trade as they adopt nationalist political agendas. The United States and Great Britain are notable examples of this trend. Second, aggressive efforts to manage climate change will require major changes in the transport sector. Through a carbon tax, cap and trade, or strict regulation, transport costs will rise, and the advantage of off-shore production will decline. Public perceptions may also be changing. Indicators include the emergence of the “flight shame” movement in Europe and the Locavore movement in the United States, which urges people to eat locally sourced foods. These factors may shrink the geographic reach of supply chains, possibly reducing trade-related urban freight activity.

Might the Consumption of Material Goods Decline?

As noted earlier in the chapter, freight demand is largely a function of population and per capita income. However, if the environmental cost of disposable or lightly used goods becomes more widely understood, it is possible that consumption per capita would decrease. The early stages of reducing waste are evident in the widespread bans on plastic bags, disposable cups, etc. Less evident but growing are movements away from “fast fashion”—buying garments for a few uses and then replacing them with the

newest style. Fast fashion consumes a lot of energy in producing and transporting the garments and then generates more energy consumption as the disposed garments join the waste stream. There is an emerging backlash against fast fashion, especially from Generation Z, which is particularly attuned to global climate change. In addition, the sharing economy continues to grow, with more apps available every year for trading or sharing clothing, tools, furniture, and much more. These apps increase the efficiency of using material goods and hence can reduce overall demand. Reduced demand may or may not reduce freight flows, as sharing or swapping implies transport.

While these potential changes would affect patterns of urban freight, it seems unlikely that freight volumes would be significantly reduced. Peer-to-peer sharing further fragments shipments. Shorter supply chains do not imply less impacts on cities. Urban freight will continue to be in need of innovative solutions to increase the sustainability of the world's metropolitan areas.

References

- Akman, G., Baynal, K., 2014. Logistics service provider selection through an integrated fuzzy multicriteria decision making approach. *J. Ind. Eng.* 2014., doi:10.1155/2014/794918, Article ID 794918.
- Cao, X., 2009. E-shopping, spatial attributes, and personal travel: a review of empirical studies. *Transp. Res. Rec.* 2135 (1), 160–169.
- Credit Suisse, 2018. CMBS Retail Outlook: Transformation Accelerating. January 4, 2018. Available from: https://research-doc.credit-suisse.com/docView?language=ENG&format=PDF&sourceid=csplusresearchcp&document_id=1080091741&serialid=pXHSk4mvE%2FPg%2BYIsoOUUnVyVpX6htl%2By2jIDMyUDG0r1%3D (Accessed 20 July 2018).
- Dicken, P., 2007. *Global Shift: Mapping the Changing Contours of the World Economy*, 5th ed. SAGE Publications, London.
- Giuliano, G., Kang, S., 2018. Spatial dynamics of the logistics industry: evidence from California. *J. Transp. Geogr.* 66, 248–258.
- Giuliano, G., Showalter, C., Yuan, Q., Zhang, R., 2018. Managing the Impacts of Freight in California. National Center for Sustainable Transportation Research Report, University of California, Davis. Available from: <https://ncst.ucdavis.edu/project/managing-impacts-freight-california>.
- Giuliano, G., Yuan, Q., 2017. Location of Warehouse and Environmental Justice: Evidence from Four Metros in California. Final Report, Project 16-1.1g. MetroFreight Center of Excellence, University of Southern California. Available from: www.metrofreight.org.
- Hess, M., Rodrigue, J.-P., 2004. The transport geography of logistics and freight distribution. *J. Transp. Geogr.* 12 (3), 171–184.
- ICS, 2013. Economic Impact of Shopping Centers. Available from: <https://www.icsc.org/uploads/default/2013-Economic-Impact.pdf> (Accessed 18 December 2018).
- ICSC, 2017. U.S. Shopping Center Classification and Characteristics. Available from: https://www.icsc.org/uploads/research/general/US_CENTER_CLASSIFICATION.pdf (Accessed 18 December 2018).
- Investp 2020 <https://www.invespro.com/blog/global-online-retail-spending-statistics-and-trends/> (Accessed 1 January 2020).
- Kang, S., 2017. Unraveling Decentralization of Warehousing and Distribution Centers: Three Essays. University of Southern California Doctoral dissertation.
- McDonald, N., Yuan, Q., Naumann, R., 2019. Urban freight and road safety in the era of e-commerce. *Traffic Inj. Prev.* 20 (7), 764–770, doi:10.1080/15389588.2019.1651930.
- Mokhtarian, P., 2004. A conceptual analysis of the transportation impacts of B2C e-commerce. *Transportation* 31 (3), 257–284.
- Rotem-Mindali, O., Weltevreden, J.W.J., 2013. Transport effects of e-commerce: what can be learned after years of research? *Transportation* 40 (5), 867–885.
- US Census, 2018. Foreign Trade Statistics 2018. <https://www.census.gov/foreign-trade/data/index.html>.
- US Census, 2019. Sales for U.S. Retail Trade Sector at \$5,046.9 Billion [Internet]. The United States Census Bureau. Available from: <https://www.census.gov/newsroom/press-releases/2019/retail-trade.html>.
- US EPA, 2018. Fast Facts US Transportation Sector Greenhouse Gas Emissions 1990–2016. US Environmental Protection Agency, Office of Transportation and Air Quality Report EPA-420-F-18-013.
- US EPA, 2019. Fast Facts: U.S. Transportation Sector Greenhouse Gas Emissions, 1990–2017. US Environmental Protection Agency, Office of Transportation and Air Quality. EPA-420-F-19-047; Available at <https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P100WUHR.pdf>.
- Yuan, Q., 2018. Mega freight generators in my backyard: a longitudinal study of environmental justice in warehousing location. *Land Use Policy* 76, 130–143.

Further Reading

- Browne, M., Behrends, S., Woxenius, J., Giuliano, G., Holguin-Veras, J., 2019. *Urban Logistics: Management, Policy and Innovation in a Rapidly Changing Environment*. Kogan Page, London.
- Conway, A., Williamson, J., Gjorgjievska, M., Chen, Q., Prasad, L., Xing, C., Peters, D., Hodge, S., Mammes, N., Klatsky, M., 2019. Complete Streets Considerations for Freight and Emergency Vehicle Operations. New York State Energy Research and Development Authority, Albany, NY. Available from: <https://www.metrofreight-publication-a-guidebook-for-considering-freight-in-complete-street-design->.
- Dabanc, L., Rodrigue, J.-P., 2017. The geography of urban freight. In: Giuliano, G., Hanson, S. (Eds.), *The Geography of Urban Transportation*. 4th ed. Guilford Press, New York, Chapter 2.
- Giuliano, G., Kang, S., Yuan, Q., 2018. Using proxies to describe the metropolitan freight landscape. *Urban Stud.* 55 (6), 1346–1363.
- Giuliano, G., O'Brien, T., Dabanc, L., Holliday, K., 2013. Synthesis of Freight Research in Urban Transportation Planning NCFRP Report 23.
- Hesse, M., 2016. *The City as a Terminal: The Urban Context of Logistics and Freight Transport*. Routledge, London.
- Holguin-Veras, J., Hodge, S., Wojtowicz, J., Singh, C., Wang, C., Jaller, M., Aros-Vera, F., Ozbay, K., Weeks, A., Replogle, M., Ukegbu, C., Ban, J., Brom, M., Campbell, S., Sanchez-Diaz, I., González-Calderón, C., Kornhauser, A., Simon, M., McSherry, S., Rahman, A., Encarnación, T., Yang, X., Ramírez-Ríos, D., Kalahashti, L., Amaya, J., Silas, M., Allen, B., Cruz, B., 2018. The New York city off-hour delivery program: a business and community-friendly sustainability program. *Interfaces* 48 (1), 70–86, doi:10.1287/inte.2017.0929.
- Sánchez-Díaz, I., 2017. Modeling urban freight generation: a study of commercial establishments' freight needs. *Transp. Res. A: Policy Pract.* 102, 3–17.
- Sánchez-Díaz, I., Georen, P., Brolinson, M., 2016. Shifting urban freight deliveries to the off-peak hours: a review of theory and practice. *Transp. Rev.* 37 (4), 521–543, doi:10.1080/01441647.2016.1254691.
- Taniguchi, E., Thompson, R. (Eds.), 2008. *Innovations in City Logistics*. Nova Science Publishers, New York.
- Zhao, Y., Ioannou, P., 2019. A co-simulation, optimization, control approach for traffic light control with truck priority. *Annu. Rev. Control* 48, 283–291.

Urban Logistics and Freight Transport

Michael Browne*, **José Holguín-Veras†**, **Julian Allen‡**, *School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden; †Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, NY, United States; ‡School of Architecture and Cities, University of Westminster, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	178
The Complexity Challenge	179
Urban Logistics Initiatives	180
Supply-Side Initiatives	180
Demand-Side Initiatives	181
Discussion and Conclusions	182
References	183

Introduction

Cities around the world are changing rapidly. Many are growing and this trend seems set to continue with over half the world's population already living in urban areas. Much attention has focused on the growth of some of the largest cities—those with a population over 10 million—referred to as mega-cities. However, many smaller cities are also growing. The growth and the scale of cities mean that urban logistics has become more important and relevant to public authority policy making. Urban logistics has in many cases responded very effectively to the requirements of modern urban economies. However, it is a major contributor to environmental impacts, particularly to local air pollution and noise and, as a result, has an important impact on the health of the most vulnerable residents of cities. This rise in importance and the initiatives that can be taken in urban logistics provides the focus for this chapter.

Urban freight transport and logistics involves the delivery and collection of goods and provision of services in towns and cities. It also includes activities such as goods storage and inventory management, waste handling, office and household removals and home delivery services (Allen et al., 2015). Scope and definitions can be important because they have implications for the policies developed by those responsible for transport and related activities in cities and metropolitan areas. A key change has been the growing tendency to include the last-mile deliveries to final consumers associated with e-commerce within the scope of urban freight transport and urban logistics. A report by the RPA (Barone and Roach, 2016) noted that different geographical scales could be considered in urban freight and logistics activities. At the wider or regional scale, the commodities travelling into and out of a metropolitan area need to be addressed. At the more local level, there are those flows that remain within a metropolitan area or perhaps a smaller part of such an area.

Hicks (1977, p.100) noted that: "any urban area depends for its existence on a massive flow of commodities into, out of, and within its boundaries. Yet the transport of goods remains a forgotten aspect of urban transportation study."

At the time, few would have disagreed with Hicks that this important activity had received relatively limited attention from researchers. However, there has been a major growth in research into urban freight, encouraged by the increased attention that has been focused on wider logistics activities. The extra attention has been strongly influenced by the growing awareness and concern about the local and global environmental impact of transport and the implications for the economic vitality of cities caused by congestion problems.

Over the past 40 years, the range of commodities, the scale and the complexity of the flows has increased. For example, during the past 10 years the growth in e-commerce means that many products now go direct from a seller or supplier to the end-customer. These changes in the routing of products has led to a significant rise in the use of small vehicles, such as vans and cargo cycles, in many cities. Although there have been major changes in the volume and types of commodity moving within cities it has, in most cases, been difficult to change the infrastructure to accommodate these new patterns.

A major challenge for improving urban logistics lies in the nature of the activities. Decisions about freight movements and truck operations lie almost entirely in the hands of the private sector. However, the regulatory and legal questions about freight movements in cities are addressed by public authorities—for example, city councils or metropolitan authorities. Many cities have a powerful political figure, such as a mayor, who takes overall responsibility for transport and other strategies in the development of the city. The political level typically relies on professional policy makers and technical managers to develop policy and manage transport and traffic in the city. Long-term land-use planning is also controlled by public authorities with considerable variation in the detailed legal powers governing these actions in different countries and indeed sometimes between cities in the same country. In addition, urban areas vary substantially in terms of factors including their geographical and physical size, population density, economic composition, land-use patterns, transport infrastructure, and street layouts. All the above makes for a complicated

commercial, technical, social, and political landscape within which urban logistics activities take place. This in turn has made it difficult to find simple solutions to urban logistics and freight problems that can be readily scaled up and transferred.

Urban logistics plays an important economic role in cities and its efficiency or otherwise has implications including:

- the effect of logistics and supply chain costs on the cost of commodities consumed in a city or metropolitan region;
- the fundamental importance of urban logistics in sustaining urban lifestyles;
- the role urban logistics plays in the service, retail, and trading activities which are essential for wealth generation in cities.

At the same time there has been a major focus on the problems caused by urban freight movements—the majority of which are by road. Road freight vehicles operating in an urban environment generally emit a greater proportion of certain pollutants per kilometer travelled than other motor vehicles, such as cars and motorcycles. This is due to their higher fuel consumption per unit of distance travelled and the fact that many of them use diesel as a fuel. However, when considered in terms of the quantity of goods carried, a well-loaded truck will produce fewer pollutants per ton or cubic meter of goods carried than say a van.

As well as making a journey within an urban area to collect or deliver goods or provide a service, a freight vehicle also has to be parked while loading, unloading or servicing activity is carried out by the driver away from the vehicle. This requires the parking of the vehicle either at the kerbside on-street, or in an off-street facility, either at the building being served or somewhere else sufficiently nearby. Many dense, central urban and residential facilities requiring goods or services comprise little off-street parking space, and in these locations freight vehicles compete for kerbside space with many other road users. Freight companies and their drivers are, therefore, subject to various challenges associated with stopping/parking their vehicles as well as the journey itself. These include:

- Parking and loading/unloading problems including complicated regulations concerning time and location, fines, lack of unloading space, and handling problems resulting from the kerbside design and paving material, for example, as well as the distance that goods may need to be carried or moved by hand in order to make a delivery.
- Customer/receiver-related problems—for example, identifying the exact address can be difficult and for some types of delivery there will be a need to obtain signatures as proof of delivery and perhaps the most common problem—narrow time windows during which receivers are willing to accept deliveries—which leads to complicated scheduling requirements.

As discussed above, while it is generally possible to agree that urban freight transport is an essential part of the economic and social well-being of a city, there is also agreement that further improvements are needed. Urban logistics also suffers from the impacts of the rise in traffic congestion, as this reduces the scope for efficient, reliable deliveries to be made. Logistics strategies and improved technology, especially related to information systems, has opened up possibilities for major gains in delivery efficiency. Yet much of this has been lost because of the rise in general congestion levels in cities and because of the inflexibility of other parts of the urban logistics systems—for example, the options to deliver outside peak times. Urban congestion and an inertia in decision-making by both the public and private sector actors has imposed constraints on further improvements.

The Complexity Challenge

Urban logistics exhibits a number of complexities including:

- Heterogeneous commodities, resulting in complicated freight flows and a variety of transport sectors;
- Overlapping regulatory regimes;
- The trend toward “logistics sprawl”;
- The range of organizations involved in decision-making processes.

These points are discussed in more detail below.

The very varied range of commodities being transported in the urban logistics system leads to complexity. Different products flow through different channels from supplier to end user and they are carried by different types of transport operator using a variety of vehicles (types and sizes). Some food products, for example, require strict adherence to specific temperature limits and will be carried in specialist refrigerated vehicles. In construction, there are also vehicles dedicated to a single type of freight movement—for example, cement mixers. Small packages and parcels may be delivered by van or by bicycle or when combined with postal services—on foot. All of this results in a very complicated pattern of vehicle movements where the main influencing factors vary between sectors.

Regulations applying to freight movements come from different sources. For example, emission standards in Europe have been established at a European Union wide level with increasingly stringent limits set on emissions by new engines for trucks. On the other hand, many regulations relating to urban logistics may be made and enforced by the city itself—for instance, there is no national standard for a low emission zone in the UK and certainly no Europe-wide standard, although there are broad similarities in some cases. At the street level then, there is considerable divergence in the way kerbside parking regulations and loading spaces are designed, operated and enforced. In some cases, such spaces will be controlled by a district within a city, leading to the situation where the rules at one end of a street could be different to the rules at the other end.

Increased urbanization has in turn resulted in generally higher values of land, especially in major urban centers. The rent or purchase price of such land has been too high to justify its use as distribution property. As a result, land classified for industrial and distribution activities has become increasingly scarce and much of this activity is now to be found only outside the city. Longer access

trips by vehicles delivering in the city may impose limitations on the type of vehicle being used—for example, it becomes less feasible to use a small electric powered vehicle if the city delivery trip is longer. City authorities have noted this development and some have now sought to create regulation to safeguard future land availability for urban freight use. For example, in the soon to be published new London Plan, the spatial development strategy for the city, the Mayor of London is proposing to include new and improved policies to retain, enhance, and provide additional capacity of logistics land, prioritizing these efforts in locations that are suitable for last-mile distribution services and support access to supply chains that extend beyond the city ([Mayor of London, 2019](#); [Transport for London, 2019](#)).

Different types of stakeholders can implement change and influence urban logistics in different ways. Public authorities may take the lead by changing regulations or by shaping policies in such a way as to contribute to broader urban sustainability goals ([Giuliano et al., 2013](#); [Holguin-Veras et al., 2020a,b](#)). In the private sector, two categories of stakeholders have received considerable attention: (i) freight transport companies or carriers and (ii) the receivers of products, for example, retail companies, offices and in some cases industrial companies. Most attention has probably been focused on the carriers, with many initiatives aimed at requiring or encouraging them to make their operations more sustainable, primarily from environmental initiatives. However, the scope for carriers to change transport operations may be limited because they frequently have to respond to their customers, who may be the senders or receivers of the goods.

Take as an example the scope to change the time of delivery in order to allow deliveries to take place outside the busiest hours on the road network (the peak hours). Many initiatives have been aimed at finding ways to encourage carriers to deliver in the off-peak hours ([Browne et al., 2014](#); [Department for Transport, 2011](#); [Holguin-Veras et al., 2011, 2014](#); [Transport for London, 2018a](#)). Despite many operational successes, carriers frequently encounter resistance from receivers, particularly from smaller stores and businesses where there are few employees and where the owner may already work long hours. Engaging with the decision-making stakeholders has been recognized as important and more attention is now focused on property owners and managers and business-led initiatives, such as Business Improvement Districts that represent local firms.

Urban Logistics Initiatives

A number of frameworks have been developed to categorize the many initiatives that can be seen in urban logistics. By comparing pilot tests, projects and initiatives around the world it has been possible for researchers to provide valuable insights for policy makers who are concerned with finding ways to improve urban logistics performance and sustainability in their cities. An extensive exercise to review urban logistics initiatives was carried out in 2015 with the leadership of Rensselaer Polytechnic University, in a project on behalf of the Transportation Research Board in the United States ([Holguin-Veras et al., 2015](#)). The project devised a classification system that allowed projects and initiatives to be grouped into seven categories divided between supply and demand initiatives.

Supply side initiatives include: (1) infrastructure management; (2) parking/loading areas management; (3) vehicle-related strategies; (4) traffic management. Demand side initiatives were categorized as: (5) incentives and taxation; (6) logistical management; (7) freight demand/land use management. The report was subsequently developed using further original research ([Holguin-Veras et al., 2020a,b](#)). The initiatives within these seven categories are discussed in more detail below.

Supply-Side Initiatives

1. *Infrastructure management*: The focus here is on changes to the built infrastructure—because most urban freight involves a road movement, then most of these initiatives are concerned with building new road infrastructure or making modifications to existing roads. Some of these can be considered as major—for example, a new road to bypass a congested area or one where there is high housing density. Others may be more minor, such as reengineering the kerb to allow deliveries to be made more efficiently or changing a ramp to enable loading and unloading to take place. In practice, minor changes to infrastructure can have important implications for freight efficiency.
2. *Parking/loading zone management*: Much attention in urban freight movement focuses on moving vehicles. However, in some sectors (e.g., package or parcel delivery), it is common for the vehicle to spend up to 60% of the time parked for unloading and delivery operations ([Allen et al., 2018a](#)). Therefore, ensuring that loading zones are managed efficiently can provide important benefits for carriers and, in turn, can improve the way the city center operates from a public authority perspective. Changing regulations about the time of access and the length of time that can be spent to load or unload has important efficiency impacts. In addition, designating areas for loading or unloading can be a major contributor to better urban freight operations.
3. *Vehicle-related initiatives*: As noted in the opening sections of the chapter, vehicle emissions have a major impact on urban air quality. As a result, a lot of attention has focused on improving existing fossil fuel emissions standards and also on the scope to change to alternatives such as electric vehicles. Furthermore, programs that have been designed to support off-peak hours delivery have also led to the search for standards relating to noise of both vehicles and handling equipment ([Sanchez-Díaz et al., 2017](#)). An important way to encourage a higher proportion of low emission vehicles is by creating low emission zones within which vehicles are only allowed to operate if they meet certain standards. Low emission zones—also known as clear zones or environmental zones—have been introduced by a range of cities, especially in Europe. One of the largest zones of this type is to be found in London which has also now introduced the ultra low emission zone ([Transport for London, 2020](#)). One issue for

transport operators has been that there can be argued to be some inconsistency in terms of regulations between cities in different countries. This can make for complications in some cross-border operations. Low emission zones may specifically restrict certain types of vehicle that do not meet specific standards.

1. *Traffic Management:* Vehicle access restrictions based on weight, size, or time of day are widespread initiatives taken by city regulators to limit what may be seen as negative environmental impacts from urban truck operations. Restrictions may make it more difficult for freight operators to combine deliveries and, in some cases, can lead to an increase in distance travelled or an increase in the number of vehicles (or both). A better understanding of urban logistics by public authorities has helped to reduce the unintended consequences of restrictions. Consultation with industry has resulted in changes to restrictions that have a positive impact—for example, reconsidering vehicle length limits may allow freight operators to continue to use the most efficient truck for the deliveries that are required.

Many cities have identified that some roads and parts of the network are not suitable for all freight vehicles. This may be because of the weight and size of the vehicle or in some cases it could be the products being carried (e.g., dangerous or hazardous cargo). In these cases, cities have defined specific routes for trucks to follow. Enforcement is an important issue when defining and maintaining such a route or network. Increasing interest now exists in the scope to use geo-fencing techniques to monitor and enforce such initiatives. Geo-fencing would require vehicles to be able to demonstrate their whereabouts in the network and the access to some parts of the network would be limited to certain trucks and types of operations.

Allocating road space can also be an important decision for public authorities. Initiatives such as “Complete Streets” in the United States have led to discussions about how freight vehicles should be accommodated in urban traffic plans (McCann and Rynne, 2010). Traditionally, vehicle weight and size limits have been widely used in cities to control vehicle movements in order to avoid large trucks being in sensitive and historic areas. However, restricting trucks according to size may result in the increased use of smaller vehicles which can have negative consequences for congestion. In the same way, access time restrictions may lead to more trucks being present in the shorter times that are available for delivery.

Demand-Side Initiatives

1. *Pricing, incentives, and taxation:* Road user charging or congestion pricing has been introduced in a number of cities around the world. In most cases, freight vehicles also pay these charges. However, research has highlighted that freight movements do not change much as a result of charges being introduced. This is mainly because carriers have limited power over the time of day of the delivery because that power lies with the receivers (or possibly the senders of the products). Incentives have been discussed as a means to encourage higher standards and better operating practices related to urban freight. However, there are relatively few examples of incentive schemes around the world. Certification projects whereby carriers can apply to receive a certification based on the quality and sustainability of their operations have been introduced with some measures of success—see, for example, the FORS scheme in the UK (FORS, 2018).
2. *Logistical management:* Managing urban logistics in new ways underpins many initiatives. These range from approaches such as urban consolidation centers and the use of lockers and collection points, as well as initiatives related to operations management, such as the routing of the vehicles and special training for urban freight drivers. The concept of consolidating products to achieve a reduction in the number of truck trips and distance run in urban areas has been promoted for many years. Consolidating products can improve the load utilization and ensure better loaded vehicle trips. However, it is sometimes argued that the consolidation operation adds time and costs to the activity. As a result, there has been commercial reluctance to adopt this approach unless there are strong regulatory or other reasons. Despite the reluctance many urban consolidation center schemes can be identified (Allen et al., 2012; Gonzalez-Feliu and Morana, 2010; Greening et al., 2015; Kin et al., 2015). The goal of using electric vehicles has also led to some support for the consolidation center approach. In many cases the lower range that can be accomplished by electric vehicles means that the operations are simpler when there is a consolidation center available at, or close to, the edge of the urban area.

Locker boxes, collection points, and other ways to support the rise in home delivery and e-commerce all have an important role in urban logistics (Morganti et al., 2014; Yuen et al., 2018). Delivery direct to the consumer can work very efficiently, but there are also many studies that have highlighted the problem of failed deliveries caused by the person receiving the products being absent at the time of delivery. Using parcel lockers or a collection point (such as a retail store) can enable a transport operator to make efficient deliveries and, at the same time, allow a consumer to receive the product at a time that is most convenient to them. As these systems become more widespread, the proximity of locker boxes and collection points to the home of urban residents has also increased, leading to greater satisfaction with such systems.

Finding the right size vehicle for urban freight operations can be challenging. Larger vehicles are required for the movement of bulky products or where there are many goods destined to one delivery point. However, for smaller packages and diverse loads, there may be scope to use smaller vehicles such as cargo cycles or light electrical vehicles. Other fuel sources may also be an option in the search for a lower environmental impact—for example, the use of biofuels. Cargo cycle use has increased significantly in recent years (Allen et al., 2018b; Rudolph and Gruber, 2017; Transport for London, 2018b) and such vehicles can now be seen providing package delivery and other services in many cities around the world. Despite the growth, the total scale of such operations remains modest at present, but there are signs of growth and such vehicles can be expected to represent a larger share of movements in inner city logistics in the future.

3. *Freight demand/land use management*: This can be considered as an important category in making fundamental changes to the way urban logistics is carried out. During recent years it has become apparent that many logistics activities are strongly connected to decisions about land use. As previously discussed, warehouse and distribution center activity that was formerly within city boundaries is often now found only on the edge of cities or even at some distance from the urban area. As a result, vehicles have longer access trips and the vehicle kilometers in relation to freight movements have increased.

The timing of deliveries can also be altered; changing to off-peak hours deliveries can have major impacts on the efficiency and sustainability of urban logistics, as well as reducing the freight vehicle contribution to emissions (Browne et al., 2014; Holguin-Veras et al., 2011, 2014; Sanchez-Díaz et al., 2017). Major gains have been achieved when such initiatives have been introduced in cities around the world. Changing the time of deliveries to move them outside the peak hours when traffic volumes are greatest has been shown to produce benefits (Browne et al., 2014; Department for Transport, 2011; Holguin-Veras et al., 2011, 2014). However, there has been resistance to implementing such initiatives, based on concerns about vehicle noise and disturbance to residents and also because the change could lead to additional costs for some organizations in the supply chain. Early discussions on this subject tended to lead to fears about delivery being made in the middle of the night. In practice, it is the scope to shift to a slightly later time in the evening or a slightly earlier time in the morning that seems to have had many benefits. When this change is combined with the opportunity to make what are called unattended deliveries, then the gains seem to be especially valued (Sanchez-Díaz et al., 2017). Unattended deliveries mean that there does not need to be a person present at the receiving address at the time the delivery takes place. Instead, drivers are either granted secure access to the delivery point or technologies that can ensure security in terms of delivery can be used.

A further important possibility is for receivers to find ways to group together or bundle their purchases and thereby reduce the number of vehicles required to make deliveries—such receiver-led consolidation approaches have received considerable attention (Allen et al., 2017; MacLean et al., 2019).

Mode choice is a very important topic in freight transport and logistics. However, in the case of urban logistics it has been difficult for companies to find viable alternatives to road deliveries using vans and trucks. Despite the problems, there are alternative modes used for urban deliveries and these include switching from road to the use of tram, rail or waterway. Over the past few years an increasing number of alternatives have become established in some cities. The scope to increase the use of modes such as waterways is especially attractive when considering the transport of bulky products such as construction materials and waste.

Discussion and Conclusions

In order to achieve successful implementation of the initiatives outlined above, a constructive process of engagement between the relevant public and private sector actors is required. This multistakeholder approach is necessary because there are many different actors involved in urban logistics and freight transport operations, and also because no single stakeholder is usually capable of solving the problem on its own. Urban logistics depends on decisions made by the public sector and by those in the private sector. However, the goals of organizations from the public and private sector differ and as a result it can be difficult to achieve a harmonious working relationship. Strategies for engagement include Freight Partnerships and Freight Networks (Lindholm and Browne, 2013; Quak et al., 2016).

As Holguin-Veras et al. (2020b) argue, public authorities typically lack detailed knowledge of how and why urban freight activities and operations take place—formal training is scarce and the level of information and data on urban freight movements is far lower than in the case of public transport and personal mobility. Almost all the operational and commercial decisions about urban freight and logistics are in the hands of the private sector, raising complicated questions about confidentiality and the scope to share data and insights. When addressing freight and logistics, urban public authorities have tended to rely on approaches based on regulation and traffic engineering.

This chapter has outlined the wide range of initiatives that can be implemented by public authorities to improve the performance of the urban freight and logistics system. However, as researchers have, there remains the need to (1) assess the effectiveness of initiatives and consider the advantages and disadvantages in different situations; (2) provide material that transportation professions can use to determine the efficacy of one initiative compared with another; and (3) identify the best ways to help public authorities take the necessary steps in improving urban freight and logistics in terms of management, policy, and planning (Holguin-Veras et al., 2020a,b).

Furthermore, the results of the research noted above also demonstrated that, in most cases, the initiatives discussed reduce congestion, either significantly or slightly, and initiatives in the group “Freight Demand/Land Use Management” were identified as having the most profound impacts. Moreover, those initiatives related to “Logistical Management” and “Freight Demand/Land Use Management” were not perceived to have significant negative effects. By contrast, some negative impacts were argued to exist for initiatives relating to financial approaches, as a result of increasing costs. Overall, the results of the work by Holguin-Veras et al. (2020a,b) show the great potential of initiatives in “Freight Demand Management” to reduce congestion with minimal negative impacts.

Research into urban logistics and freight transport has certainly grown significantly over the past 20 years. However, this is from a small base of researchers. At the end of the 1990s, there were already research conferences focused on urban freight and logistics, but the number of researchers attending was modest. Since then there has been a major increase in interest from cities and a parallel

growth in the number of researchers and the range of countries interested in this field (Browne et al., 2019). The research community has grown and the practitioners concerned with urban freight, and in particular those involved in the public sector and city authorities, have also increased. However, in comparison to practitioners and authorities involved in personal travel, the number remains very low. It is of considerable importance that the research that has been produced in the past 20 years is made more widely available to practitioners and policy makers and that guidance on implementation continues to be developed. If this can be achieved, then the promising progress outlined in this chapter will become far more widespread.

References

- Allen, J., Bektas, T., Cherrett, T., Friday, A., McLeod, F., Piecyk, M., Piotrowska, M., Zaltz Austwick, M., 2017. Enabling the freight traffic controller for collaborative multi-drop urban logistics: practical and theoretical challenges. *Transport. Res. Record* 2609, 77–84.
- Allen, J., Browne, M., Holguin-Veras, J., 2015. Sustainability strategies for city logistics. In: McKinnon, A., Browne, M., Whiteing, A., Piecyk, M. (Eds.), *Green Logistics: Improving the Environmental Sustainability of Logistics*. third ed. Kogan Page, London, pp. 293–306.
- Allen, J., Browne, M., Woodburn, A., Leonardi, J., 2012. The role of urban consolidation centres in sustainable freight transport. *Transp. Rev.* 32 (4), 473–490.
- Allen, J., Piecyk, M., Piotrowska, M., McLeod, F., Cherrett, T., Nguyen, T., Bektas, T., Bates, O., Friday, A., Wise, S., Austwick, M., 2018a. Understanding the impact of E-commerce on last-mile light goods vehicle activity in urban areas: the case of London. *Transport. Res. D* 61 (B), 325–338.
- Allen, J., Piecyk, M., Piotrowska, M., 2018b. Same-day delivery market and operations in the UK, report as part of the Freight Traffic Control 2050 project. University of Westminster. Available from: http://www.ftc2050.com/reports/ftc2050-2018-Same-day_delivery_market_UK.pdf.
- Barone, R., Roach, E., 2016. Why Goods Movement Matters: Strategies for Moving Goods in Metropolitan Areas. RPA. Available from: <http://library.rpa.org/pdf/Why-Goods-Movement-Matters-ENG.pdf>.
- Browne, M., Allen, J., Wainwright, I., Palmer, A., Williams, I., 2014. London 2012: changing delivery patterns in response to the impact of the Games on traffic flows. *Int. J. Urban Sci.* 18 (2), 244–261.
- Browne, M., Behrends, S., Woxenius, J., 2019. Introduction to urban logistics. In: Browne, M., Behrends, S., Woxenius, J., Giuliano, G., Holguin-Veras, J. (Eds.), *Urban Logistics: Management, policy and innovation in a rapidly changing environment*. Kogan Page, London, ISBN 97809749478711, pp. 3–18.
- Department for Transport, 2011. Quiet Deliveries: Good Practice, Principles and Processes. Department for Transport. Available from: <https://www.gov.uk/government/publications/quiet-deliveries-demonstration-scheme>.
- FORS, 2018. Fleet Operator Recognition Scheme Standard. FORS. Available from: https://www.fors-online.org.uk/cms/wp-content/uploads/2019/11/FORSStandard_v5.pdf.
- Gonzalez-Feliu, J., Morana, J., 2010. Are city logistics solutions sustainable? The Cityporto case. *Territorio Mobilità e Ambiente* 3 (2), 55–64.
- Greening, P., Piecyk, M., Palmer, A., McKinnon, A., 2015. An Assessment of the Potential for Demand-Side Fuel Savings in the Heavy Goods Vehicle (HGV) sector. The Centre for Sustainable Road Freight. Available from: <https://www.theccc.org.uk/wp-content/uploads/2015/11/CFSRF-An-assessment-of-the-potential-for-demand-side-fuel-savings-in-the-HGV-sector.pdf>.
- Giuliano, G., O'Brien, T., Dabian, L., Holliday, K., 2013. Synthesis of Freight Research in Urban Transportation Planning. Transportation Research Board NCFRP Report 23.
- Hicks, S., 1977. Urban freight. In: Hensher, D. (Ed.), *Urban Transport Economics*. Cambridge University Press, Cambridge.
- Holguin-Veras, J., Amaya Leal, J., Sanchez-Diaz, I., Browne, M., Wojtowicz, J., 2018a. State of the Art and Practice of Urban Freight Management Part I: Infrastructure, Vehicle-Related, and Traffic Operations. *Transport. Res. A* 137, 360–382.
- Holguin-Veras, J., Amaya Leal, J., Sanchez-Diaz, I., Browne, M., Wojtowicz, J., 2020b. State of the art and practice of urban freight management Part II: Financial approaches, logistics, and demand management. *Transport. Res. A* 137, 383–410.
- Holguin-Veras, J., Amaya-Leal, J., Wojtowicz, J., Jaller, M., González-Calderón, C., Sanchez-Díaz, I., Wang, X., Haake, D., Rhodes, S., Hodge, S., Frazier, R., Nick, M., Dack, J., Casinelli, L., Browne, M., 2015. Improving Freight System Performance in Metropolitan Areas, NCFRP 33. Transportation Research Board.
- Holguin-Veras, J., Ozbay, K., Kornhauser, A., Ukkusuri, S., Brom, M., Iyer, S., Yushimoto, W., Allen, B., Silas, M., 2011. Overall impacts of off-hour delivery programs in the New York City Metropolitan Area. *Transport. Res. Rec.* 2238, 68–76.
- Holguin-Veras, J., Wang, C., Browne, M., Hodge, S., Wojtowicz, J., 2014. The New York City off-hour delivery project: lessons for city logistics, eighth international conference on city logistics. *Procedia* 125, 36–48.
- Kin, B., Verlinde, S., van Lier, T., Macharis, C., 2015. Is there life after subsidy for an Urban Consolidation Centre? In: *Conference Proceedings, 9th International Conference on City Logistics, Tenerife*, 17–19 June, pp. 422–437.
- Lindholm, M., Browne, M., 2013. Local authority cooperation with urban freight stakeholders – a comparison of partnership approaches. *Eur. J. Transport Infrastruct. Res.* 13 (1), 20–38.
- MacLean, F., Bates, O., McLeod, F., 2019. Parcel Carrier Collaboration – Can Big Cities Learn from Small Communities? CILT Focus, pp. 43–45. Available from: http://www.ftc2050.com/papers/ParcelColab_Focus_April_19.pdf.
- Mayor of London, 2019. The London Plan: Spatial Development Strategy for Greater London, Intend to Publish. Greater London Authority. Available from: https://www.london.gov.uk/sites/default/files/intend_to_publish_-_clean.pdf.
- McCann, B., Rynne, S., 2010. Complete Streets: Best Policy and Implementation Practices, PAS Report 559. American Planning Association.
- Morganti, E., Dabian, L., Fortin, F., 2014. Final deliveries for online shopping: The deployment of pickup point networks in urban and suburban areas. *Res. Transport. Bus. Manag.* 11, 23–31.
- Quak, H., Lindholm, M., Tavasszy, L., Browne, M., 2016. From freight partnerships to city logistics living labs – giving meaning to the elusive concept of living labs. *Transport. Res. Proc.* 12, 461–473.
- Rudolph, C., Gruber, J., 2017. Cargo cycles in commercial transport: Potentials, constraints, and recommendations. *Res. Transport. Bus. Manag.* 24, 26–36.
- Sanchez-Diaz, I., Georén, P., Brolinson, M., 2017. Shifting urban freight deliveries to the off-peak hours: a review of theory and practice. *Transp. Rev.* 37, 521–543.
- Transport for London (2018a) Transport for London's code of practice for quieter out-of-hours Deliveries, Transport for London. Available at: <http://content.tfl.gov.uk/codeofpractice.pdf>.
- Transport for London (2018b) Strategies to increase uptake of cycling freight in London, report prepared by Element Energy and WSP for TfL, Transport for London. Available at: <http://content.tfl.gov.uk/cycle-freight-study.pdf>.
- Transport for London (2019) Freight Action Plan, Transport for London. Available at: <http://content.tfl.gov.uk/freight-servicing-action-plan.pdf>.
- Transport for London (2020) Ultra Low Emission Zone, Transport for London. Available at: <https://tfl.gov.uk/modes/driving/ultra-low-emission-zone>.
- Yuen, K., Wang, X., Ng, W., Wong, Y., 2018. An investigation of customers' intention to use self-collection services for last-mile delivery. *Transport Policy* 66, 1–8.

Omni-Channel Logistics

Tom Van Woensel, Eindhoven University of Technology, Eindhoven, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Omni-Channel Retailing and Omni-Channel Logistics	184
Omni-Channel Logistics: Toward a Definition	184
Decision Problems	186
Network Design	186
Challenges	187
Warehousing	187
Challenges	187
Inventory Management	187
Challenges	188
Last-Mile Delivery	188
Challenges	188
Return Logistics	188
Challenges	189
Closing Remarks	189
References	189

Omni-Channel Retailing and Omni-Channel Logistics

Omni-channel *retailing* refers to a business model that increases the customer experience by allowing the use of a variety of channels in a customer's shopping journey (Fig. 1). The increase in usage of digital technologies, social media and mobile devices has dramatically changed the shopping behavior of the customer and hence, the retail environment. This change necessitates retailers to rethink their marketing, supply chain, and logistics strategies. The customer journey not only involves the physical store, but also alternative sources of information (e.g., via mobile phones, internet, etc.), and potentially, buying online. In the marketing literature (Gallino and Moreno, 2014), this is referred to as the *ROPO* effect (Research online, purchase offline) versus show rooming (researching in a physical store and buying online [Gensler et al., 2017]).

Many initiatives aim to revitalize the offline retailers. Clearly, these retailers face challenges to compete with online retail, and are more and more in need of combining the traditional, offline channels with the growing (in size and importance to customers) online channels.

Omni-channel retailing thus involves the organization and integration of the different multiple sales channels (think of stores, website, etc.), mainly within the B2C (business-to-consumer) context. The customer is always central in this respect regardless of the actual *marketing* channel used. From a customer perspective, his or her demand is thus fulfilled seamlessly across all the channels. The customer experiences the brand rather than channels, hence omni-channel marketing. In summary, omni-channel marketing involves integrating multiple sales channels into a single-integrated sales channel, planned and controlled accordingly. Clearly, this buying process approach is a marketing oriented focus.

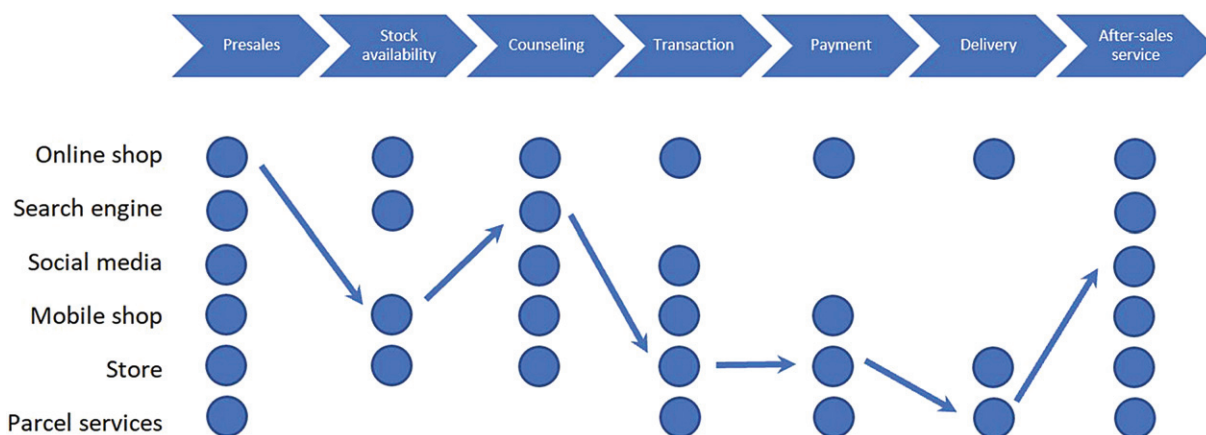


Figure 1 The shopping journey.

Obviously, there is a clear difference between the buying process of the customer and the order-fulfillment process, after buying. In the latter, delivery reliability and logistics costs are important drivers, also affecting the performance and experience of the sales process and the ability to sell. Specifically, the consumer not only asks for detailed product information, but also wants information about (real-time) availability, accurate delivery times, and a broad range of return options during the buying process. These decision and organization problems are typically in the omni-channel *logistics*' domain (Savelsbergh and Woensel, 2016).

Omni-Channel Logistics: Toward a Definition

We define omni-channel logistics as *all necessary activities involved with the functioning of the supply chain considering an omni-channel retail strategy* (Van Woensel and Broft, 2016). In its essence (Simchi-Levi et al., 2019), logistics and supply chain management aims at providing good service (the right product, at the right time, and at the right place), for a low cost. This becomes more and more complex in this omni-channel context, considering different channels to connect to the customers. In our definition, omni-channel logistics is mainly involved with the fulfillment process of the orders, in connection to these customers. More narrow, *e-fulfillment* deals with the fulfillment activities for online buying.

Fig. 2 gives an overview of the omni-channel logistics activities related to the supply chain. We simplified the supply chain in the sequence sourcing–production–distribution–retail/detail/e-tail–Consumer. Contrary to the most graphical depiction, the consumer is not the last element in this sequence, but more and more connects directly to all echelons. Think of the following examples. Consumers dictate that producers (e.g., fast moving consumer goods or fashion companies) to source their raw products in a clean, sustainable, and ethical way. Production is increasingly being done in a *batch-size-equal-to-one* way: NIKEiD allows customers to design and personalize their own Nike merchandise, being produced and shipped directly to their customers. Similarly, customers expect full visibility of the inventory in warehouses and can order their products directly from these sites, rather than going to the stores (e.g., DIY chains). Finally, the customer is also more and more in the lead on how and where their orders need to be delivered (the “logsumer” takes an active role on time, price, quality, and sustainability in decisions of logistics services [DHL, 2013]).

As in any supply chain, physical products move back and forth, next to information, and money flows. Adequately managing these flows is a key driver for successful omni-channel logistics. Over the past years, managing the reverse logistics flows became an important distinguishing success factor. Clearly, reverse logistics also offers a wide range of different possible return channels (aligned with the Retail, Detail, and E-tail channels).

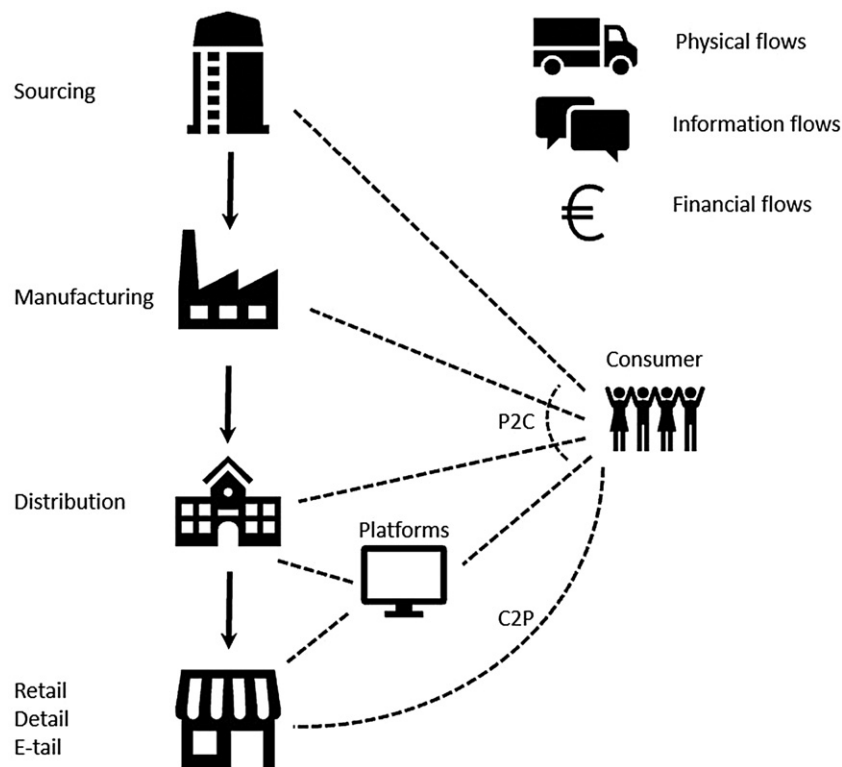


Figure 2 Omni-channel logistics.

We summarize the different channels for reaching customers into the following three broad categories:

- **Retail:** The retail channel is primarily focusing on larger retail chains. Here, we see large to mega-stores with a large and broad assortment (not necessarily deep or specialized). These chains typically have large buying power and operate their own network of warehouses and, in many cases, their own vehicle fleet.
- **Detail:** The detail channel is more oriented toward smaller stores and shops (“mom and pop stores”). The assortment is smaller, but more deep. More experienced and specialized personnel work in the store. As they are usually not part of a larger chain, they rely on wholesalers. Sometimes, these detail stores are part of a large retail chain, for example, think of a city-center convenience store (express stores), offering a fast-moving assortment.
- **E-tail:** Consumer online orders in web stores are the primary focus of the E-tail channel. We see that many E-tail stores are part of a retail or detail chain. Pure online players are getting more and more rare, as the need for a physical network is increasing, due to the higher demand of the customers (e.g., same hour delivery).

As an example of the above categories think of flowers. Usually, in a retail setting, only few flower cultivars are sold (typically as ready-made bouquets). However, in a detail store, the assortment is much larger, also more variety in bouquets is available. In the e-tail setting, again, the same broadness might be reached.

Note that the e-tail channel penetrates throughout the complete supply chain and potentially affects the sourcing, production, etc. This makes it possible for manufacturers to reach out directly toward their customers, rather than via retailers, wholesalers, etc. Shortening the supply chain toward the customer can be very beneficial in terms of costs and in terms of short delivery lead times. On the other hand, the other channels still need to be handled as well. This again calls for efficient and effective omni-channel logistics.

Clearly, the retail and detail channels are typically bricks-and-mortar stores where customers need to go physically and buy their product. This is denoted in [Fig. 2](#) as *C2P*, short for customer-to-product. The e-tail channel is of course part of the online stores, web shops, in-store ordering with home delivery, etc. This is denoted in [Fig. 2](#) as *P2C*, short for product-to-customer. For both settings, the order fulfillment processes can be very different. Considering the omni-channel environment creates additional challenges, as we need to manage a more complex demand portfolio and multiple inventories with cross-replenishment over channels. All should be planned and controlled with total visibility of the supply chain, implying that integrated business planning is a prerequisite.

Over the past years, the role of sales platforms is increasing. Platforms like Amazon, Alibaba, Ebay, or Bol.com are growing in importance. Platforms originally were carrying and owning the inventory themselves. However, these platforms are moving toward a partner model setup, connecting multiple retailers. As such, the platform is not carrying inventory anymore, but only acts as the sales channel. The inventory holding is executed by the participating retail partners. In some cases, the participating retailers are making use of the logistics infrastructure of the platform (e.g., delivery, invoicing, etc.).

The latter system of organizing platforms is also observed to move toward a dropshipping execution model. Here, the retailer does not carry inventory, but uses the wholesaler to fulfill their customer’s order. The latter ships directly to the customer and the retailer earns the profit based on the difference between the wholesale price and the customer sales price. Similarly, within the omni-channel context, end-consumer orders are directly fulfilled from the various upstream suppliers ([Peinkofer et al., 2019](#)). The platform then serves as a collection of small to large retailers, as such broadening and deepening the assortment offered by that platform.

Decision Problems

In order to create a seamless customer experience, coordination between all logistics processes is crucial ([Hubner et al., 2015](#)). To enable demand fulfillment from anywhere to anywhere, decisions on network design, inventory management, warehousing, and transportation need to be aligned with the fulfillment options and promised lead times. In network design, it is important to consider that as well as pursuing economies of scale and cost benefits, a local presence is required. Without local presence, it is either unfeasible or too expensive to be able to offer the required short lead times to customers. In addition, real-time inventory visibility across all locations is required such that fulfillment options can be offered in real-time. The warehouse management systems and transportation systems need to be designed in such a way that picking, packing, and transport are feasible within the promised lead time to the customer.

Network Design

Network design is a typical supply chain process of strategic planning directly impacting its performance. Network design decisions are greatly affected by the channel strategy chosen. Omni-channel network design needs to consider more key performance indicators (KPIs) than the traditional network design methodologies.

Network design is about the physical configuration of the supply chain ([Ballou, 2001](#)). As such, network design involves organizing the supply chain network extending from raw material sources to the final consumption destination. Think of determining the number, size, and location of terminals, ports, plants, warehouses, stores, lockers, cross docks, etc. Specifically, it deals with how the network is as depicted in a very abstract manner in [Fig. 2](#), needs to look like^a.

^a As a side-note, note that network design is usually a redesign exercise, that is, improving the existing current network.

Traditionally, we evaluate the performance of the network along with two dimensions: customer needs versus logistics costs (Chopra, 2003). Various KPIs can be considered for both dimensions: response time, product variety, product availability, customer experience, order visibility, and returnability (customer needs) and inventory, facility, and (inbound and outbound) transportation costs (logistics costs).

Challenges

An adequate omni-channel network design requires considerations of the two dimensions presented above. However, since this omni-channel network fulfills both online and offline customer orders from anywhere in the network, an additional complexity arises. Moreover, since the customer connects to more echelons in the supply chain (Fig. 2), the proximity to the customer gains in importance, in terms of response time, and costs. Clearly, the integrated network design for an omni-channel strategy leads to high complexity and important challenges.

Clearly, C2P and P2C processes are difficult to separate, necessitating a network design that is able to handle both processes. Network integration to fulfill customer orders from all retail, detail, and e-tail channels is a key for successful omni-channel order fulfillment. Combining this integration with the increasing need for speed, calls for high-performance, and more dense networks.

Infrastructure and transportation costs need to be optimized by integrating the networks and adequately sharing the usage of warehouses and trucks. (Hubner et al., 2015) listed a number of order fulfillment networks, that is, either from the stores, from a warehouse, or directly from the manufacturer. These authors identified three large archetypes of networks for omni-channel logistic: (1) a fully integrated network (online and offline are organized from the same location), (2) a parallel network (online and offline are separately organized), and (3) a sequential network (online is replenished via the offline warehouse). These networks are however too limited, as due to the short response times, e-fulfillment centers need to be much closer to customers. Hence, fulfillment options directly from the offline stores are needed as well, creating one extra layer of complexity in the network organization.

Warehousing

An omni-channel strategy forces retailers to organize their warehouses not only as a traditional distribution center but also manage the online sales from their web-shops. Summarizing, two types of order flows are then received: (1) retail and detail store orders (large size, less frequent) versus (2) e-tail orders (small size, but highly frequent and irregular). This evolution toward more and irregular orders, and decreasing order sizes is a direct consequence of the omni-channel strategy (Hamberg and Verriet, 2012).

The material handling is the core activity in the warehouses. From a cost-perspective, this constitutes the largest part of the operational costs: "Most of the expense in a typical warehouse is in labor; most of that is in order-picking; and most of that is in travel" (Bartholdi and Hackman, 2017). Clearly, these processes that deal with the organization of the order flow and the storage and picking operations are most affected when implementing an omni-channel supply chain.

Challenges

The impact on the warehouse management and operations following an omni-channel strategy mainly impact the picking, sorting, and packing processes. A critical challenge is the ability to handle different transaction sizes within one warehouse setting, without losing efficiency in the key processes (Niels et al., 2006). Handling the different order flows (online and offline) puts pressure on the material handling flexibility. Picking of individual items versus pallet or case-pack picking is significantly different in terms of picking equipment, planning, and organization (i.e., real-time information availability and real-time communication requirements) for the warehouse management system (WMS), enterprise resource planning (ERP), and/or distributed order management (DOM) system.

Following the above discussion, separate material-handling systems seem to be required, as a fully integrated system, handling both online and offline flows, seems difficult. This could be either in two physically separate warehouses or in one building, controlled by independent systems. Moreover, in a B2C context with increasing online orders, picking quality is important. Advanced picking technologies (e.g., radio frequency terminals, wireless speech technology, and pick/put-to-light systems) are interesting design decisions, but their high investments require high order volumes for a viable business case. Next to the required high picking quality, online order picking involves small pick quantities for many items. In general, these activities are more labor-consuming than case-pack or pallet picking (Niels et al., 2006).

Adding the online order flow to the traditional flow gives rise to extra complexity in the warehouse operations. Outsourcing to a 3PL (third-party logistics) service partner is then an interesting solution. These 3PL service providers can leverage on their economies of scale.

Inventory Management

Customers expect high service levels, in terms of availability, delivery times, and reliability. On the other hand, operating the supply chain leads to costs, which needs to be kept under control. Inventory management helps to reduce these costs. Inventory models aim to balance the different cost factors (ordering, holding, shortage, salvage costs, and discount rates) in such a way that the supply chain performance is optimized.

Over time, inventory management has become increasingly complex in its effort to adequately consider, as much as possible, the underlying real-life situation. Although less representative of real-life, single-echelon inventory models are relatively easy to analyze.

More representative, but more complex, multi-echelon models are also available, depicting serial models, distribution systems and assembly systems (Ton, 2018).

Challenges

Putting the customer at the center of an omni-channel supply chain, leads to the expectation that the inventory should be available at any place (both physical and channel), at any time and in the right quantity. This leads to many more inventory locations in the supply chain. Another consequence is that inventory is more distributed in the supply chain, as such, losing the inventory pooling advantages, potentially leading to higher inventory costs.

Moreover, a careful characterization of out-of-stocks is needed within the omni-channel setting, as the product can be available in the company, but not available to the customer (e.g., wrong place or wrong time or dedicated to another channel). This last point also relates again strongly to the network design (e.g., proximity to the customer).

Many inventory models in the literature assume that consecutive orders cannot overtake each other (Nahmias and Olsen, 2015). This means that, once an order is allocated to inventory ("promised"), this inventory cannot be used anymore for later orders, but with an earlier due date than any previous order. For example, replenishment of the inventory happens on Wednesday, an online order comes in on Monday with a delivery due date on Friday, depleting the on-hand inventory. An order coming in on Tuesday with a due date on Wednesday cannot be fulfilled, as the inventory is promised for the Monday order, resulting in lost sales for the Tuesday order. Swapping or allowing order overtaking, would resolve this situation.

Advanced algorithms are needed, adjusting the retail inventory rules to better exploit these opportunities and order fulfillment challenges provided by an omni-channel strategy. Moreover, richer performance measures are needed. Product availability for a certain service level based on a single channel demand fulfillment strategy might not be sufficient anymore and does not cover all the discussed challenges in a sufficient way. Short-term seasonality effects on the demand and the connected inventory rules are needed as well (Ehrental et al., 2014).

Last-Mile Delivery

Last-mile delivery refers to the last transportation echelon in the supply chain. Traditionally, brick-and-mortar retailers are the last stock point, and the customer comes there (Quak and de Koster, 2009). As such, the last mile is toward the store. However, following an omni-channel strategy, click-and-mortar stores and their last-mile delivery toward the customer's home is exploding (Niels et al., 2006). Whereas the brick-and-mortar option gives rise to consolidated flows from warehouses to the store, the click-and-mortar sales leads to a magnitude of fragmented delivery locations with small drop sizes.

Note that the last-mile delivery is highly related to other discussed processes in the supply chain. To a large extent, the network structure and the availability of inventory determine the delivery origin and destination. As this last-mile distribution happens mostly in a city context, all relevant considerations and challenges originating from city logistics are highly relevant to mention here as well (Bektas et al., 2017).

Challenges

Several important factors for store distribution (the C2P situation) are discussed in (Quak and de Koster, 2009). Specifically, these authors mention three large categories, that is, urban policies, logistical concepts, and retailer-specific characteristics. Each influences the retail store delivery process and its performance. Following the evolution of larger assortments (i.e., high product mix), brick-and-mortar retailers are moving into more pull versus push supply chains. This leads to more fluctuating demand and requires faster and more frequent delivery toward the stores.

Also note that time-window policy and vehicle restrictions, as imposed by the urban policy makers, also create additional challenges for the performance of the last-mile delivery to stores. These policy measures lead to under-utilization of vehicles and more vehicles in the routing and scheduling operations (Quak and de Koster, 2009). Ghilas et al. (2016) aim to improve this drop in efficiency in searching for good combinations of people and freight within a *cargo hitching* context.

Clearly, the rise of the online channel (click-and-mortar), has given broader and deeper assortments to the customer, but also increased the pressure on last-mile delivery, following this e-commerce boom. In this case, the products go to the customers (P2C). Savelsbergh and Woensel (2016) discuss various opportunities to reach the customers in their preferred locations (trunk delivery, drones, etc.). Dabia et al. (2017) discuss a variant of the pollution-routing problem where, next to the cost parameters, sustainability in terms of greenhouse gas emissions are also considered (Tolga Bektas and Laporte, 2011).

Several new initiatives are observed in practice: crowd shipping (Le et al., 2019), urban consolidation centers (Dellaert et al., 2019), integrated routing and inventory (Song et al., 2019), horizontal collaboration (Hezarkhani et al., 2019), etc.

Return Logistics

Reverse logistics refers to the product flow going back into the supply chain, rather than out of the supply chain in the direction of the customer (i.e., forward distribution) (Fleischmann et al., 1997). Serving the customer from everywhere means that for product returns, the customer should have more options than merely bringing the product back to the store. These options range from physical store returns, parcel shops, locker walls or being picked-up at home. Return requests also happen via a wide range of possibilities: internet, mail, phone, social media, etc.

Table 1 Buy and return options

<i>Return</i>	<i>Buy offline</i>	<i>Buy online</i>
Collection point	(1)	(2)
Home pickup	(3)	(4)

Challenges

Providing a generous return policy reduces the customer's reluctance to buy online and creates confidence. Moreover, it has positive effects on sales volumes and revenue. On the other hand, this increases overhead costs and complexity in the return process (Mukhopadhyay and Setoputro, 2004).

Combining the forward and return flows gives four possible combinations (Table 1). Each combination in Table 1 already gives rise to additional complexity in the supply chain. However, within an omni-channel strategy, *all* combinations would ideally be offered to give full flexibility to the customers. Combination (1) is the most standard one, existing for a long time within a brick-and-mortar setting. The other combinations (2), (3), and (4) exist following the growth of the online market. The collection point returns for combinations (1) and (2) lead to extra movements of the customer. In this case, retailers aim to use their physical network (e.g., stores) or the network of a service provider (post office, collection points) as effectively as possible. Home pickup of the returns is an expensive solution, but the least effort for the customer. Adequate routing and scheduling, combining pickup and delivery along the same routes, are much needed here (Peng et al., 2018).

Closing Remarks

An overview is presented of a selection of important logistics challenges in adopting an omni-channel strategy. This retail strategy leads to new models and insights which might not be there to the full extent. Especially, this integration of offline and online sales channels has many implications for current logistics and operations models. We hope that this chapter delivers some insight into these challenges and implications.

References

- Ballou, R.H., 2001. Unresolved issues in supply chain network design. *Inform. Sys. Front.* 3, 417–426.
- Bartholdi, J., Hackman, S., 2017. *Warehouse and Distribution Science*. Georgia Institute of Technology, Atlanta, GA.
- Bektas, T., Crainic, T.G., van Woensel, T., 2017. From managing urban freight to smart city logistics networks. In: *Series on Computers and Operations Research*, World Scientific, United States, pp. 143–188.
- Chopra, S., 2003. Designing the distribution network in a supply chain. *Transp. Res. Part E: Logist. Transp. Rev.* 39 (2), 123–140.
- Dabia, S., Demir, E., van Woensel, T., 2017. An exact approach for a variant of the pollution-routing problem. *Transp. Sci.* 51 (2), 607–628.
- Dellaert, N., Dashty, S.F., van Woensel, T., Crainic, T.G., 2019. Branch-and-price-based algorithms for the two-echelon vehicle routing problem with time windows. *Transp. Sci.* 53 (2), 463–479.
- DHL, 2013. Logistics Trend Radar. DHL Customer Solutions & Innovation, Germany. Available from: https://www.dhl.com/content/dam/Campaigns/InnovationDay_2013/90310673_HI-RES.PDF
- Ehrental, J.C.F., Honhon, D., van Woensel, T., 2014. Demand seasonality in retail inventory management. *Eur. J. Operat. Res.* 238 (2), 527–539.
- Fleischmann, M., Bloemhof-Ruwaard, J.M., Dekker, R., van der Laan, E., Jo, A.E.E., van, N., Wassenhove, L.N.V., 1997. Quantitative models for reverse logistics: a review. *Eur. J. Operat. Res.* 103 (1), 1–17.
- Gallino, S., Moreno, A., 2014. Integration of online and offline channels in retail: the impact of sharing reliable inventory availability information. *Manag. Sci.* 60 (6), 1434–1451.
- Gensler, S., Neslin, S.A., Verhoef, P.C., 2017. The showrooming phenomenon: it's more than just about price. *J. Interact. Market.* 38, 29–43.
- Ghilas, V., Demir, E., van Woensel, T., 2016. An adaptive large neighborhood search heuristic for the pickup and delivery problem with time windows and scheduled lines. *Comp. Operat. Res.* 72, 12–30.
- Hamberg, R., Verriet, J., 2012. *Automation in Warehouse Development*. Springer, Verlag London.
- Hezarkhani, B., Slikker, M., van Woensel, T., 2019. Gain-sharing in urban consolidation centers. *Eur. J. Operat. Res.* 279 (2), 380–392.
- Hubner, A., Holzapfel, A., Kuhn, H., 2015. Operations management in multi-channel retailing: an exploratory study. *Operat. Manag. Res.* 8 (3–4), 84–100.
- Le, T.V., Stathopoulos, A., van Woensel, T., Ukkusuri, S.V., 2019. Supply, demand, operations, and management of crowd-shipping services: a review and empirical evidence. *Transp. Res. Part C: Emerg. Technol.* 103, 83–103.
- Mukhopadhyay, S.K., Setoputro, R., 2004. Reverse logistics in e-business: optimal price and return policy. *Int. J. Phys. Distrib. Logist. Manag.* 34 (1), 70–89.
- Nahmias, S., Olsen, T.L., 2015. *Production and Operations Analysis*. Waveland Press, Long Grove, IL, USA.
- Niels, A., Agatz, H., Fleischmann, M., Jo van, Nunen, 2006. E-fulfillment and multi-channel distribution – a review. *Eur. J. Operat. Res.* 187, 339–356.
- Peinkofer, S.T., Esper, T.L., Smith, R.J., Williams, B.D., 2019. Assessing the impact of drop-shipping fulfillment operations on the upstream supply chain. *Int. J. Prod. Res.* 57 (11), 3598–3621.
- Peng, S., Veelenturf, L.P., Hewitt, M., van Woensel, T., 2018. The time-dependent pickup and delivery problem with time windows. *Transp. Res. Part B: Method.* 116, 1–24.
- Quak, H.J. (Hans), de Koster, B. M.(René), 2009. Delivering goods in urban areas: how to deal with urban policy restrictions and the environment. *Transp. Sci.* 43 (2), 211–227.
- Savelsbergh, M., van Woensel, T., 2016. 50th anniversary invited article-city logistics: challenges and opportunities. *Transp. Sci.* 50 (2), 579–590.
- Simchi-Levi, D., Kaminsky, P., E., Simchi-Levi., 2019. *Designing and managing the supply chain: concepts*. In: *Strategies and Case Studies*, Irwin Publishers, New York.
- Song, R., Zhao, L., van Woensel, T., Fransoo, J.C., 2019. Coordinated delivery in urban retail. *Transp. Res. Part E: Logist. Transp. Rev.* 126, 122–148.
- Tolga Bekta, A. Y., Laporte, G., 2011. The pollution-routing problem. *Transp. Res. Part B: Method.* 45 (8), 1232–1250.
- Ton, D.K., 2018. Inventory management: modeling real-life supply chains and empirical validity. *Found. Trends Technol. Inform. Operat. Manag.* 11 (4), 343–437.
- Van Woensel, T., Broft, A.D., 2016. *Omni-Channel Logistics: State of the Art*. Technische Universiteit Eindhoven, Eindhoven.

Humanitarian Logistics

Gyöngyi Kovács, Diego Vega, HUMLOG Institute, Hanken School of Economics, Helsinki, Finland

© 2019 Elsevier Ltd. All rights reserved.

Introduction	190
Special Features of Humanitarian Transportation	190
The Context Impacts on Modal Choice	191
You Never Walk Alone	191
The Humanitarian Transportation Chain	192
New Developments in Humanitarian Transportation	192
Conclusions	193
Biographies	193
Relevant Websites	194
References	194
Further Reading	194

Introduction

Disasters, conflicts, crises, and pandemics have one thing in common: they impact on livelihoods and require that people assist one another. It is the discipline of humanitarian logistics that has taken it on itself to focus on the delivery of aid to people in need. Humanitarian logistics has come a long way from a reactive, ad hoc way of collecting and delivering materials, to a coordinated, complex, global, and professional effort of delivering aid. Notwithstanding rapid developments, challenges in this area will always prevail, and come with the next natural hazard actually impacting on human lives, or the outbreak of the next conflict or next pandemic.

Given that the core of the discipline is to deliver aid, freight transportation is an essential part of the overall logistical effort. This goes so far that many humanitarian organizations regard logistics as synonymous with transportation (Kovács, 2018). This may be a narrow view, not just with respect to logistics, but also on the humanitarian supply chain. However, it illustrates the point about the importance of transportation in this area. What is more, humanitarian organizations have themselves been likened to logistics service providers, 3PLs, and 4PLs (Abidi et al., 2015; Jensen, 2012; Vega and Roussat, 2015). After all, they extremely rarely manufacture any of the goods they deliver, while they purchase, pre-position, prepackage, and deliver all sorts of materials and finished products for their teams and for other humanitarian organizations (Vega and Roussat, 2019). Many are specialized in some product groups that relate to the thematic clusters they operate in, such as food, shelter, health, water and sanitation, etc.

Overall, of course, the same logistical principles apply to humanitarian logistics just as well as to any other application area of logistics and transportation. Yet, humanitarian logistics does also have some special features and faces particular challenges. These will be in the limelight of this article.

Special Features of Humanitarian Transportation

The very aim of humanitarian logistics being to assist people in need makes it different from any other application area of logistics and transportation. There are at least two main issues that stand out from this aim:

- This being a not-for-profit environment; and
- That it is needs-driven.

Humanitarian logistics has these features in common with emergency services or public health and, in fact, many of the logistical models and solutions developed for those areas are easily adaptable to the humanitarian context. Other models and solutions require more adaptation, as one cannot just not deliver to areas or groups of people that would not result in profit, thereby cutting off remote areas, difficult-to-navigate informal settlements, or dangerous environments, as often it is exactly in these spots where people are found who are most in need. To some extent, the problem is not just one of access to people in need, who may even be behind dynamically morphing lines of conflict, but also to identify and find people in need who may be actively hiding or are constantly moving, either due to a conflict, their position in society, or a stigma related to their health condition. The very aspect of mobility of vehicles used in freight transportation can be used to address this problem, converting trucks to distribution points or even donkeys to mobile clinics, or vessels to warehouses or hospitals (Tatham et al., 2016). In other areas of humanitarian logistics, containers have been used to set up different types of facilities, whether warehouses (Şahin et al., 2014), offices, or actual shelter.

At the same time, being needs-driven is quite different to responding to the indefinite nuances of demand. Again, it is from healthcare or emergency services that one can apply the concept of triage, in that cases are addressed in terms of their degree of urgency. There are some quite advanced models that have been developed to assess the needs of a population, and to further also assess the vulnerabilities versus capacities of individuals to determine who is in greatest need. Minimum standards of a response have also been set in the SPHERE standards, which outline details from the liters of water per person per day, to the square meters of shelter per household to the maximum weight of package sizes so they can be carried by individuals in the absence of any equipment. Apart from that, however, opinions vary as to the service level that meets needs, and whether the aim is to restore what people had, or to build back better, or to contribute to individual and community resilience. The target is constantly shifting.

The Context Impacts on Modal Choice

Other special features of humanitarian logistics stem from the impact of the actual “event” a so-called humanitarian program is responding to. Natural hazards and conflicts tend to impact greatly on the infrastructure around them, including on the transport and energy infrastructure. Destabilized, if not destroyed, ports and airports also impact on what is left in terms of modal choice, or at least vehicle selection. Railway lines often need to be rehabilitated and many roads may be unpassable, and bridges down, or are only capable of operating with limited capacity. Countless optimization models have been developed for vehicle routing in disasters, facing the common problem of not knowing which routes are *de facto* still available, and for which types of vehicles. Aerial photography (whether from satellites or drones) can help, but is not yet there to establish not just that a bridge is still standing, but that it can take a particular axle load.

Modal choice is not only restricted, however. Humanitarian transportation has come up with fascinating solutions, from the use of “quake jumpers” (i.e., people who jumped out of helicopters to assess the needs of populations trapped in the mountains of Kashmir after an earthquake) to riverboats in the Congo, dhows from Djibouti to Yemen, and all kinds of pack animals. Rare to be found in the typical logistics textbook, it is quite useful to know, for the humanitarian logistician, which kinds of loads different pack animals can take, for how long, and with what speed. As a side note, who would have known that lowland mules could have vertigo and refuse to ascend a mountain trail?

Perhaps the most misleading yet dominant picture of disaster relief is that of airdrops. Far from ideal, these are in fact applied as last resort. After all, without people who can oversee the distribution of aid in the last mile, it is the strongest and fittest, rather than the people in the most dire need, who are reached with such deliveries. Yet, air cargo is essential and much used in humanitarian transportation, especially in the very first moments of disaster relief. Sadly, while other transportation modes could at times be faster (Tatham et al., 2016), it is the image of air cargo being faster than any other mode of delivery that has led to the persistence of air cargo as the preferred mode of transportation in inbound logistics. This is despite the fact that air cargoes are more likely to get stuck in customs—leading of course to higher overall lead times after all.

When it comes to larger volumes and steady flows on the inbound, it is the combination of sea and road transportation that comes to the rescue. Banomyong and Grant (2016) have developed a humanitarian aid transport reference model that revisits all these constraints and requirements along a time line after a disaster. The guiding questions of the humanitarian aid transport reference model are as follows:

- What are the needs and requirements of the affected area?
- Can aid be diverted by land or sea within 24 h?
- Can it be diverted or bought locally after 24 h?
- Is the affected country landlocked?

As the last of these points indicates, responding to disasters in landlocked countries comes with further challenges. The disaster region may have its own views on treating incoming cargo for disaster relief, but they are not always shared across transport corridors when it comes to transit transport (Beresford and Pettit, 2019), as exemplified also in Choi et al. (2010) and Beresford et al. (2016). At the same time, movement can be restricted back out from the disaster area, either out of the fear of displaced people crossing the border as refugees, or for wanting to restrict the spread of pandemics.

You Never Walk Alone

Disaster relief is not a one-man show. It requires the combined effort of many organizations. Thus, many endeavors have focused on the coordination of aid deliveries, whether in the form of regular meetings to keep one another informed, to more actively establishing who delivers what where and in which time frame. Structured approaches to such coordination include alliances (e.g., the UNGM platform shared across UN agencies, the IASC Consolidated Appeal Process, and the ACT Alliance for Protestant and Orthodox organizations) to more those with an explicit logistical focus, for example, the Logistics Cluster, the Fleet Forum, the Logistics Emergency Teams, or the Humanitarian Logistics Association. Of these, the Logistics Cluster has both a global and a very operational disaster-specific coordination role in case of large-scale disasters (with a UN presence). Both they and the Fleet Forum also set standards, develop templates, and are used to exchange experiences. The latter is at the very heart of the Humanitarian Logistics Association, which has a strong focus on the professionalization of the discipline.

There is also a considerable commercial interest in the humanitarian sector. Apart from the sheer magnitude of this “industry” (OCHA’s Global Humanitarian Overview 2018 indicates 135.3 million people in need, 97.9 million people to receive aid, and the requirements of 25.2 billion USD to be able to deliver this aid), the reasons for companies and especially logistics service providers to get involved are many (Bealt et al., 2016; Tomasini, 2011; Vega and Roussat, 2015): apart from the very commercial interest in this sector, this involvement is used as a corporate social responsibility initiative, to attract and retain a motivated workforce, to the commercial reason of helping a disaster area to get back to business (i.e., the LSP’s core business) sooner, to using this as a market entry to a particular region, and to wanting to learn from the agility of humanitarian organizations. However, the interests of humanitarian organizations and LSPs do not always align, especially not when an LSP is required to be involved (rather than wanting to be), or when it comes to the *modus operandi* (Bealt et al., 2016; Falagara Sigala and Wakolbinger, 2019).

The Humanitarian Transportation Chain

Broadly speaking, there are three main ways in which a humanitarian delivery can take place:

- With the same humanitarian organization handling everything from procurement to delivery in the last mile;
- Through a myriad of organizations and implementing partners; and
- Through convoys.

The first option signifies organizations that value their neutrality and independence even from other humanitarian organizations, but also the newbies that have not yet heard of all the existing coordination platforms. As for the professional ones, they tend to work in conflict zones and derive their legitimacy from working alone. Only in rare cases would they turn to the help of other humanitarian organizations, though they may in fact share some logistical information (stock levels, scheduled arrivals, etc.) through coordination mechanisms. That said, even these may engage commercial logistics service providers especially for inbound flows: with the main restriction to that being the lack of geographical reach of such LSPs in conflict zones.

The second most common option is that of a larger international humanitarian organization orchestrating the delivery of aid to a specific destination and people of concern. This is what is meant by humanitarian organizations, not to speak of their coordination platforms, acting as 4PLs (Abidi et al., 2015; Jensen, 2012; Vega and Roussat, 2015, 2019). In this vein, humanitarian organizations procure items for persons of concern and see to it that these items reach those people. In the extreme, this could not require any material flow through this focal humanitarian organization, as procured items can be delivered directly to their downstream implementing partners, with everything else being outsourced. Few humanitarian organizations have their own fleet for freight; what is covered under “fleet management” also in the Fleet Forum refers mostly to the 4×4 s humanitarians use for their own access to particular locations. That said, the management of this fleet has developed into an art of itself, with highly interesting solutions being used for the leasing and maintenance of the global fleet of international humanitarian organizations (Kunz and van Wassenhove, 2019).

Most likely, however, the focal organization will still be the consignee of the shipment in the country of operation, and then involve several LSPs and (humanitarian) implementing partners in the last mile. What is more, many countries require their own freight forwarders to be used within their territories, leading to cross-docking at the border from one truck to another (Kachali et al., 2017). On the other hand, the focal humanitarian organization can very well also be involved in the value-adding logistics activities: in the storage of blank goods till their destination is known, in their packaging, labeling, kitting, customization and assembly (Vega and Roussat, 2019), and in consolidating the transportation of different shipments (Vaillancourt, 2016) or, alternatively, outsource some of these activities to LSPs (Falagara Sigala and Wakolbinger, 2019). At the same time, specialized humanitarian activities may also be carried out, including (re-)construction, repair and maintenance, assembly of water and sanitation infrastructure, etc. (Vega and Roussat, 2019). To add to this complexity, humanitarian organizations also offer both asset-based and network-based logistical services to one another (Heaslip et al., 2018). Not surprisingly, notwithstanding the illusion that items are tracked to their final use, commodity tracking systems focus on the delivery of a shipment to the next implementing partner only. Instead, whether they have or have not reached their target population is assessed by independent consultants visiting this population when carrying out monitoring and evaluation surveys.

The third option of using convoys is again a last resort. It is mostly used in conflict situations where safe passage needs to be negotiated through particular transport corridors (Choi et al., 2010; Kachali et al., 2017; Tatham and Kovács, 2017). Convoys may still include the leasing of trucks for the specific operation, but are closely supervised and operated by humanitarian organizations themselves. They can be operated by one single organization (especially if that is the only one that has been granted access through a line of conflict) or involve the consolidation of cargo across several organizations.

New Developments in Humanitarian Transportation

Ever since its beginnings, logistics has constantly adapted to the evolution of the humanitarian context, responding to the different changes that the field has witnessed. After a pick on the number of natural disasters worldwide at the beginning of the 21st century, the following years have seen an increasing number of outbreaks of infectious diseases, conflicts and displacements, which have a

tremendous impact on the type of aid delivered. On the other hand, the global economic downturn has pushed humanitarian organizations to adopt a more cost-conscious approach (Heaslip et al., 2018), which in turn has an impact on how that aid is delivered. Finally, many humanitarian organizations have developed a particular expertise in certain areas, including the use of technologies, which substantially reshapes the humanitarian network and the relationships between humanitarian organizations and commercial actors. New developments have seen the light in response to these changes.

As the world moves toward more slow-onset disasters such as droughts, pandemics, and conflict-related crises, so does humanitarian aid. In these particular crises, which share the same characteristics as other disasters, humanitarian organizations have to deal with an extra layer of complexity, such as following “people on the move” even in the sea and sometimes finding people “hiding on the move” in the case of population displacement, ensuring cold chains for vaccines (which need to be maintained between 2°C and 8°C, or in the extreme case of the Ebola vaccine, require an ultra-cold chain of up to –60°C to –80°C) in places with no electricity, or following and developing safety regulations when responding to an infectious disease outbreak.

With the concern for cost reduction comes also the need to rethink the way humanitarian aid is provided. Cash-based assistance has come as a solution not only to shorten the supply chain, simplify procurement, and remove the need for many logistics activities, but also to empower beneficiaries by providing them with purchasing power that will be used locally and so contribute to strengthening the local economy. Thus, in addition to setting up and managing cash transfers instead of product delivery, humanitarian organizations move toward the localization of the response, mapping in-country capabilities and, when needed, building logistics capacity. This represents an important shift in the humanitarian logistician’s skill set: from being a jack-of-all-trades, to a logistics and cash office, but more important it shows the evolution that the context is having from product-based to more service-oriented.

Service provision across humanitarian organizations is a well-known fact. Whether it is simple logistical support from one organization to another, sharing services, or exchanging services, the many activities performed by international and local NGOs are at the heart of the servitization trend in humanitarian logistics (Heaslip et al., 2018). Many of the major and mid-size humanitarian organizations provide all kinds of logistics services to other organizations, but also to grassroots movements, health partners, public institutions, and governments, either ad hoc or as part of a larger agreement and both on a profit and not-for-profit basis. Also, these organizations engage in the sharing or exchange of logistics services (Vega and Roussat, 2019). One example of this is UNICEF managing procurement services during the Ebola crisis on behalf of 100 partners including GAVI, who in turn used UPS aircraft based in Cologne, Germany, to transport UNHCR and UNICEF equipment to the countries worst hit by the epidemic.

All these changes have, to some extent, been made possible by new technologies and their adoption for humanitarian relief. Information technology plays an important role in coordinating physical flows between the different actors of the humanitarian network. New technologies like blockchain and the Internet of Things help organizations to manage transfers and ensure security in the case of cash-based assistance, improve visibility through the tracking and tracing of goods, and in some situations support emergency response through enhanced geographical information and urgent deliveries by the use of drones (Tatham et al., 2018).

Conclusions

By and large, humanitarian logistics is of course just logistics in another context, whether considering strategic questions of transportation planning (Gralla and Goentzel, 2018) or operational ones of vehicle routing (Fikar et al., 2018). This context may though impact not only on the aim and outreach of a program, but in rather operational terms for any delivery, on:

- The members of the transportation chain, including the availability (or not) of logistics service providers, the intricacies of coordination and service tiering across humanitarian organizations, and the implementing partners distributing the aid in the last mile;
- Access to an area and/or to a particular group of people, including the routes and transportation modes that are still usable, and the way of gaining safe passage and the choice of delivery method; and
- Product-specific requirements of materials handling, for example, the requirement of medical personnel to handle health products; temperature control for various food and health items; and the requirement of being able to deliver cash and/or goods.

Yet, this context is constantly changing. Both new challenges and new solutions will emerge over time. Watch this space!

Biographies

Gyöngyi Kovács is the Erkkö Professor in Humanitarian Logistics, and the Head of Subject of Supply Chain Management and Social Responsibility at the Hanken School of Economics, Helsinki, Finland. She is also the co-editor-in-chief of the *Journal of Humanitarian Logistics and Supply Chain Management*. Her research interests include humanitarian logistics as well as sustainable supply chain management.

Diego Vega is Assistant Professor of Supply Chain Management and Social Responsibility at Hanken School of Economics, Finland, and Deputy Director of the Humanitarian Logistics and Supply Chain Research Institute (HUMLOG Institute). His interests include humanitarian logistics, logistics service providers, competence-based strategic management, and temporary organizations.

Relevant Websites

Fleet Forum: Available from: <https://www.fleetforum.org/>

Humanitarian Logistics Association: Available from: <https://www.humanitarianlogistics.org/>

Logistics Cluster: Available from: <https://logcluster.org/>

OCHA's Global Humanitarian Overview: Available from: <https://interactive.unocha.org/publication/globalhumanitarianoverview/>

SPHERE standards: Available from: <https://www.spherestandards.org/>

References

- Abidi, H., De Leeuw, S., Klumpp, M., 2015. The value of fourth-party logistics services in the humanitarian supply chain. *J. Humanitarian Logist. Suppl. Chain Manag.* 5 (1), 35–60.
- Banomyong, R., Grant, D.B., 2016. Transport in humanitarian supply chains. In: Haavisto, I., Kovács, G., Spens, K. (Eds.), *Supply Chain Management for Humanitarians*. Kogan Page, Philadelphia, PA (Chapter 6.1), pp. 191–208.
- Bealt, J., Fernández Barrera, J., Mansouri, S., 2016. Collaborative relationships between logistics service providers and humanitarian organizations during disaster relief operations. *J. Humanitarian Logist. Suppl. Chain Manag.* 6 (2), 118–144.
- Beresford, A., Pettit, S., 2019. Humanitarian aid supply chain management. In: Wells, P. (Ed.), *Contemporary Operations and Logistics*. Palgrave Macmillan, Cham, pp. 341–364.
- Beresford, A., Pettit, S., al Hashimi, Z., 2016. Humanitarian aid supply corridors: Europe—Iraq. In: Haavisto, I., Kovács, G., Spens, K. (Eds.), *Supply Chain Management for Humanitarians*. Kogan Page, Philadelphia, PA (Chapter 6.2), pp. 209–222.
- Choi, A.K., Beresford, A.K., Pettit, S.J., Bayusuf, F., 2010. Humanitarian aid distribution in East Africa: a study in supply chain volatility and fragility. *Suppl. Chain Forum Int. J.* 11 (3), 20–31.
- Falagara Sigala, I., Wakolbinger, T., 2019. Outsourcing of humanitarian logistics to commercial logistics service providers: an empirical investigation. *J. Humanitarian Logist. Suppl. Chain Manag.* 9 (1), 47–69.
- Fikar, C., Hirsch, P., Nolz, P.C., 2018. Agent-based simulation optimization for dynamic disaster relief distribution. *Central Eur. J. Oper. Res.* 26 (2), 423–442.
- Gralla, E., Goentzel, J., 2018. Humanitarian transportation planning: evaluation of practice-based heuristics and recommendations for improvement. *Eur. J. Oper. Res.* 269 (2), 436–450.
- Heaslip, G., Kovács, G., Grant, D., 2018. Servitization as a competitive difference in humanitarian logistics. *J. Humanitarian Logist. Suppl. Chain Manag.* 8 (4), 497–517.
- Jensen, L.M., 2012. Humanitarian cluster leads: lessons from 4PLs. *J. Humanitarian Logist. Suppl. Chain Manag.* 2 (2), 148–160.
- Kachali, H., Kovács, G., Grant, D., 2017. Determining tracking and tracing user requirements in conflict zones. *EUROMA 2017 Conference Proceedings*, paper 42899, Edinburgh, UK.
- Kovács, G., 2018. Where next? A glimpse of the future of humanitarian logistics. In: Tatham, P., Christopher, M. (Eds.), *Humanitarian Logistics: Meeting the Challenge of Preparing for and Responding to Disasters*. third ed. Kogan Page, London, UK, pp. 316–330.
- Kunz, N., van Wassenhove, L., 2019. Fleet sizing for UNHCR country offices. *J. Oper. Manag.* 65, 282–307.
- Şahin, A., Alp Ertem, M., Emür, E., 2014. Using containers as storage facilities in humanitarian logistics. *J. Humanitarian Logist. Suppl. Chain Manag.* 4 (2), 286–307.
- Tatham, P., Kovács, G., 2017. Overcoming international freight transport challenges in a disaster response context. In: Beresford, A., Pettit, S. (Eds.), *International Freight Transport: Cases, Structures and Prospects*. Kogan Page, London, pp. 299–314.
- Tatham, P., Ball, C.M., Wu, Y., Diplas, P., 2018. Using long endurance remotely piloted aircraft systems to support humanitarian logistic operations: a case study of Cyclone Winston. In: *Smart Technologies for Emergency Response and Disaster Management*. IGI Global, Hershey, PA, pp. 264–278.
- Tatham, P., Kovács, G., Vaillancourt, A., 2016. Evaluating the applicability of sea-basing to support the preparation for, and response to, rapid onset disasters. *IEEE Trans. Eng. Manag.* 63 (1), 67–77.
- Tomasini, R.M., 2011. Helping to learn? Learning opportunities for seconded corporate managers. *J. Glob. Responsibility* 2 (1), 46–59.
- Vaillancourt, A., 2016. A theoretical framework for consolidation in humanitarian logistics. *J. Humanitarian Logist. Suppl. Chain Manag.* 6 (1), 2–23.
- Vega, D., Roussat, C., 2015. Humanitarian logistics: the role of logistics service providers. *Int. J. Phys. Distr. Logist. Manag.* 45 (4), 352–375.
- Vega, D., Roussat, C., 2019. Towards a conceptualization of humanitarian service providers. *Int. J. Logist. Manag.* doi: 10.1108/IJLM-04-2018-0091.

Further Reading

- Haavisto, I., Kovács, G., Spens, K. (Eds.), 2016. *Supply Chain Management for Humanitarians*. Kogan Page, London, UK.
- Kovács, G., Spens, K., Moshtari, M. (Eds.), 2018. *The Palgrave Handbook of Humanitarian Logistics and Supply Chain Management*. Palgrave Macmillan/Springer Nature, London, UK.
- Kovács, G., Spens, K.M. (Eds.), 2012. *Relief Supply Chain Management for Disasters: Humanitarian, Aid, and Emergency Logistics*. IGI Global, Hershey, PA, USA.
- Kovács, G., Tatham, P., Larson, P.D., 2012. What skills are needed to be a humanitarian logistician? *J. Bus. Logist.* 33 (3), 245–258.
- Matopoulos, A., Kovács, G., Hayes, O., 2014. Examining the use of local resources and procurement practices in humanitarian supply chains: an empirical examination of large scale house reconstruction projects. *Decis. Sci.* 45 (4), 621–646.
- Tatham, P., Christopher, M. (Eds.), 2018. *Humanitarian Logistics: Meeting the Challenge of Preparing for and Responding to Disasters*. third ed. Kogan Page, London, UK.
- Tomasini, R., van Wassenhove, L., 2009. *Humanitarian Logistics*. Palgrave MacMillan, Basingstoke, UK.

Optimization of Humanitarian Logistics

M. Teresa Ortuño*, Jose M. Ferrer†, Inmaculada Flores*, Gregorio Tirado*, *Faculty of Mathematics, Complutense University of Madrid, Madrid, Spain; †Faculty of Economics and Business, Complutense University of Madrid, Madrid, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	195
Humanitarian Logistics and Disaster Management	195
Commercial Logistics Versus Humanitarian Logistics	196
The Disaster Management Cycle: Tasks and Decisions	196
Evacuation in Humanitarian Logistics	197
Last Mile Distribution in Humanitarian Logistics	198
Exact and Heuristic Methods in Humanitarian Logistics	198
Multicriteria Optimization for Humanitarian Logistics	199
Acknowledgment	199
References	200
Further Reading	200

Introduction

Throughout the history of humanity, disasters, both natural and human-made, have hit the population and devastated infrastructure. In the last decades, the number and impact of disasters seem to be increasing, predominantly due to weather-related events, and trying to minimize their impact and consequences is an aspiration of Governments and organizations that involves many decision-making processes that can be optimized.

The word *Optimization* refers to the action of choosing “the best” (in some sense to be defined) among different options—what can be called “the feasible.” The Cambridge Dictionary states that optimization is the act of making something as good as possible. In the same way, it states that *logistics* is the careful organization of a complicated military, business, or other activity so that it happens in a successful and effective way. The terms *Military logistics* and *Business logistics* have a long tradition in operational research. In fact, logistics is an extension of something that armies throughout history have taken care of: the supply of goods, movement, and maintenance of armed forces. In a more general way, logistics can be defined as the process of planning and organizing to make sure that resources are in the places where they are needed, when they are needed, so that an activity happens effectively.

In recent decades, a new term has been coined: *Humanitarian logistics*. This kind of logistics is focused on alleviating the suffering of vulnerable people. In fact, humanitarian logistics has been defined in the Humanitarian Logistics Conference, 2004 (Fritz Institute), as the process of planning, implementing and controlling the efficient, cost-effective flow and storage of goods and materials, as well as related information, from the point of origin to the point of consumption for the purpose of meeting the end beneficiary’s requirements and alleviate the suffering of vulnerable people.

Humanitarian Logistics and Disaster Management

Humanitarian logistics can appear in different contexts, but it is in the management of disasters where its specificity is greater. Some definitions related to the field of disaster management (Ortuño et al., 2012) are next introduced.

- A hazard is a threatening event or probability of occurrence of a potentially damaging phenomenon within a given time period and area. It can be both natural and human-made.
 - Natural: naturally occurring physical phenomena caused either by rapid or slow onset events, which can be geophysical, hydrological, climatological, meteorological, or biological.
 - Technological or human-made: events caused by humans and which occur in (or close to) human settlements, such as complex emergencies/conflicts, famine, displaced populations, industrial accidents, catastrophic transport accidents, etc.
- Vulnerability: This is the degree to which people, property, resources, systems, and cultural, economic, environmental, and social activity are susceptible to harm, degradation, or destruction on being exposed to the consequences of a hazard.
- An emergency is a situation that poses an immediate risk to health, life, property, or environment.
- A disaster is the disruption of the normal functioning of a system or community, which causes a strong impact on people, structures and environment, and goes beyond the local capacity of response. It is a combination of hazard and vulnerability.
- A catastrophe is considered an extremely large-scale disaster.

Table 1 Differences between commercial and humanitarian logistics.

<i>Characteristics</i>	<i>Commercial logistics</i>	<i>Humanitarian logistics</i>
Objective pursued	Minimize the logistic cost (inventory and transportation costs, among others)	Alleviate suffering
Demand	Known or estimated	Unpredictable in terms of timing, geographic location, type and quantity
Decision makers	Few decision makers	Multiple decision makers sometimes difficult to identify
Periodicity	Usually repetitive	Sudden apparition of unexpected demand for large amounts
Initial resources	Controlled	Lacking
Transportation network	Stable and functional	Can be damaged Dynamically changing
Uncertainty	Mainly due to demand, price or resources	Additional uncertainty due to damaged structures, dependence on donations or safety concerns

From the beginning of history, disasters have struck human beings and they have had to face their impact and consequences. In addition, far from diminishing, disasters, natural or human-made, every year affect more people around the world. The use of optimization approaches to tackle disaster management problems dates back to the 1950s, where they were used to determine the optimal location of fire-fighting resources. In the 1970s the first models on emergency logistics were developed, facing some maritime disaster situations. The decision-making processes in disaster management are extremely difficult, due to the multiple actors which are involved and the complexity of the tasks addressed. Among those tasks, humanitarian logistics stands out as a field where optimization techniques can be of great help. From the 1980s up to now, a growing literature on the optimization of humanitarian logistics has been developed.

Note that based on its definition, humanitarian logistics also appears in contexts different from disaster management. However, it is in the latter where the application of humanitarian logistics is more complex and difficult and where more differences with commercial logistics appear.

Commercial Logistics Versus Humanitarian Logistics

Optimizing logistics processes has been at the core of operational research since the beginning. The business and commercial sector has contributed to the development of such methods and benefited from them, and the question about whether such methods can guide decision making in the humanitarian sector has interested both researchers and practitioners.

The main characteristics that differentiate humanitarian supply chains in the context of disaster management from commercial supply chains can be summarized in [Table 1](#).

These differences show the need for new optimization models that can be used in the humanitarian sector which address its peculiar characteristics.

The Disaster Management Cycle: Tasks and Decisions

In the decision-making processes needed in humanitarian logistics for disaster management, the context and the nature of the decisions to be made change over time, especially as we move from before to after the disaster event. Deciding about preventive actions to mitigate the effects of a possible future hurricane is not the same as deciding about the precise actions to undertake just after it strikes, or two weeks later. Time pressure varies substantially from one situation to the other, being at a maximum just after the disaster occurs, while uncertainty is greater when only hurricane forecasts are available. The same differences can be found in the nature of the decisions and the criteria considered by the involved actors. This has led to distinguishing four successive phases in the management of emergencies and disasters according to the main nature of the tasks to be performed and their temporal location with respect to the disaster event ([Altay and Green, 2006](#)), as illustrated in [Fig. 1](#).

The four phases may be defined as follows:

Mitigation: This phase encompasses the actions taken to either prevent the onset of a disaster or reduce the impact in case it occurs. Typical optimization models arising in this phase are the selection of infrastructures to reinforce, the location of permanent shelters or warehouses, or the allocation of resources for them, among others.

Preparedness: All the actions taken to prepare the community to respond when a disaster occurs. Typical optimization models arising are inventory prepositioning problems, location of temporary shelters and evacuation plans, among others.

Response: This phase includes the tasks and decisions taken just after the disaster strikes, and it is characterized by a short duration with high time pressure and high uncertainty. Typical optimization models arising are transport, evacuation, and distribution problems, among others.



Figure 1 Disaster management cycle.

Recovery: This refers to all long-term actions and decisions taken in order to stabilize and recuperate the normal functioning of the affected population. It is characterized by its long duration with low uncertainty and low time pressure. Typical optimization models refer to infrastructure recovery, among others.

These phases are not independent from each other, they could even overlap, and the process of disaster management has to be understood as a nonstop cycle (**Fig. 1**). In order to help in the decision-making processes arising in each of the interrelated previously mentioned phases, different optimization models can be developed, and in recent decades an explosion of literature regarding this topic has taken place. In particular, very good surveys on models and optimization techniques used in humanitarian logistics can be found in the literature.

Regarding transportation problems, they present different characteristics. Vehicles can transport people, goods or a combination of both. They can visit the same node more than once or not, and can traverse the same arc more than once or not. Subtours can be allowed or not. Problems can be multimodal (including different types of vehicles, trucks, helicopters, ambulances, etc.), multidepot (including multiple supply nodes and/or multiple vehicle initial locations), and multiloading (each vehicle can load multiple times, even in the same node). Models can split delivery or carry out transshipments in any node. Also, data can be static (single-period models) or dynamic (multi-period models), deterministic or stochastic.

We will focus next on two of the most important and complex transportation problems arising in the humanitarian context: evacuation problems (both in the preparedness and in the response phase) and last mile distribution problems (arising in the response phase), both aimed at saving the lives of vulnerable people and alleviating suffering.

Evacuation in Humanitarian Logistics

According to the United Nations Office for Disaster Risk Reduction, an evacuation “consists of moving people and assets temporarily to safer places, before, during or after the occurrence of a hazardous event in order to protect them. Moreover, evacuation plans refer to the arrangements established in advance to evacuate properly, as well as plans for return of evacuees and options to habilitate safe shelters in the affected area.”

Following the previous definition, evacuation problems appear mainly in the preparedness and response phases. In the context of the preparedness phase, it is necessary to design the evacuation plans, prelocate supplies and detect possible locations that will be used as temporary shelters. As part of the response plan, the actions to be taken in order to evacuate the affected population include the rescue and secure allocation of affected people and the distribution of basic needs. Sometimes the evacuation could be part of the recovery phase, following the reevaluation of the possible infrastructure failures that may occur during the response phase.

According to the [London Resilience Mass Evacuation Framework \(2018\)](#), the following forms of evacuation are possible:

- **Self-evacuation:** individuals will make their own arrangements to move from a place of danger or threat of danger to a safe place using existing public transport or their own mode of transportation. They require no assistance from the community beyond general warning and informing communications.
- **Assisted evacuation:** individuals are capable of transporting themselves but require some type of assistance, as indications regarding where to evacuate or by which means.
- **Supported evacuation:** individuals cannot evacuate by their own means and require support from the emergency services and public authorities.

In addition, evacuation problems may arise in buildings, cities, regions or collective means of transport, such as trains, ships or airplanes. Due to the increasing height of buildings and the expansion of travel, there is a large collection of works related with building, train, airplane, and ship evacuation.

In general, self-evacuation is encouraged, if circumstances allow. For this reason, self-evacuation has been vastly studied in the literature facing the problems of traffic congestion, bottlenecks, density waves, training lines, oscillations, patterns, and panic. These characteristics are traditionally modeled via nonlinear programming models and queuing theory. Besides that, in the field of Humanitarian logistics it is necessary to open research lines with methods related to supported evacuation. These include routing,

location and distribution problems. A characteristic to take into account concerning these models is data inaccuracy, since most censuses have information about where people live, but there is not accurate information about the population present in the area at the time an evacuation has to be carried out. Furthermore, dynamic models prevail over static models due to the variability of the characteristics, and the fact that uncertainty has to be captured via stochastic models. The usual constraints appearing in these models are flow constraints, to ensure proper movement of vehicles and people; capacity constraints in vehicles and shelters, and time-related constraints.

The primary objective of evacuation models is to maximize the number of evacuees, combined with minimizing response time or travel distance from the affected area to the safe one. Besides that, other objectives may arise, such as cost, coverage, reliability, security or equity, so evacuation models are in general multicriteria. The size and complexity of the resulting models imply, in many instances, the use of metaheuristic approaches.

Last Mile Distribution in Humanitarian Logistics

The distribution of humanitarian aid to the population affected by the occurrence of a disaster is one of the most important tasks to be carried out in the context of humanitarian logistics and disaster management. The Logistics Operational Guide of the Logistics Cluster states that “the distribution chain or channel represents the movement of a product or service from the point of purchase to the time it is handed over to the final user/consumer. Some of the distribution activities embrace materials handling, storage and warehousing, packaging, transportation, etc.” In particular, last mile distribution refers to the last leg of the distribution chain, dealing with the transportation of products or people from a certain hub or depot to their final recipient or destination. These definitions hold both for commercial and humanitarian logistics. However, if we focus on the latter, the products involved are usually humanitarian aid that must reach the regions in need. If, in addition, the situation is a disaster relief context, issues such as the lack of reliable information, the urgency to reach the affected people or the possible damages to the local infrastructure all serve to complicate significantly the decision process and the implementation of the distribution plans. According to these elements and other relevant information related to aid and vehicle availability, budget, demand, etc., the involved organizations need to define the target of the operation and the plans to be developed. Last mile distribution operations usually begin at hubs or regional warehouses and end at field warehouses, delivery points managed by other local organizations or even at the beneficiaries directly. The transportation of humanitarian aid may involve the use of heterogeneous fleets that could include, among others, vans (or minivans), trucks of different sizes, 4x4 vehicles, helicopters, etc. Their availability depends on the local resources and the characteristics of the existing infrastructure.

In last mile distribution there is a variety of elements to be considered when evaluating the performance of a certain distribution plan, or in other words, several attributes must be taken into account when deciding how the operations should be carried out. First, the cost of the operation is always a key issue in any real situation, no matter the circumstances, and the time required to perform the whole distribution is usually an important factor as well, not only to provide a good service to customers receiving purchased products but especially relevant when the recipients are affected people and there are lives at stake. However, in the context of disaster relief, there are additional performance indicators that arise naturally: for example, when the available aid is not sufficient to meet the requirements of all potential beneficiaries, it is crucial to ensure that the distribution is carried out in an equitable way, trying not to leave out certain population groups because they are difficult or expensive to reach (or for any other reasons). In addition, designing reliable and secure distribution itineraries, in order to avoid areas which may have been significantly damaged by the disaster or that could be dangerous due to likely assaults, is also desirable. All these elements should be taken into account jointly by the decision maker, when planning the operation, and yield multicriteria optimization models, as will be discussed later (Ferrer et al., 2018).

Most of the underlying optimization problems arising in last mile distribution are vehicle routing problems (VRPs), which can take many different forms: capacitated VRPs, pickup and delivery problems, VRPs with time windows, split delivery, heterogeneous fleets, loading conditions, precedence constraints... and even combinations of these or other variants with certain modifications. In some cases, the problem can be formulated as a single or multiple flow problem, which are usually easier to solve. Furthermore, most real problems have a dynamic (data change over time) and stochastic (there is uncertainty regarding certain elements of the problem) nature, which are important to consider but complicate the resulting mathematical models to a great extent. These problems usually have a combinatorial number of solutions, making them very hard to solve to optimality and requiring long computational times to obtain proven optimal solutions. For this reason, and also taking into account the urgency associated with disaster relief intervention, heuristics and other approximation methods are frequently used to be able to provide the decision maker with reasonable solutions within short running times. This is crucial, so that these models and methods can be useful in practice as decision making aids or tools.

Exact and Heuristic Methods in Humanitarian Logistics

In the humanitarian logistics literature, mixed integer linear and non-linear models have been developed and solved to optimality, but in real case applications some of them fail to give a solution in reasonable time, due to the complexity of the models. In the context of optimization, a heuristic is a method that solves a problem in a short time without guaranteeing to reach

the optimal solution. Heuristic algorithms are appropriate to complex optimization problems and/or problems that must be solved with high urgency. Both characteristics—complexity and urgency—are usually present in humanitarian logistics, especially in the response phase, so heuristics are frequently applied in this field. A metaheuristic is a special higher-level heuristic, which iteratively guides a subordinate procedure, mostly a heuristic as well, to find near-optimal solutions. Metaheuristics generally include learning strategies, in order to efficiently explore the feasible set and may incorporate mechanisms to avoid getting trapped in a local minimum.

Metaheuristics can be classified according to several criteria. One of these possible classifications deals with the way in which solutions are obtained. In the first group we can include metaheuristics in which the solutions are obtained by performing alterations to previously obtained solutions. Local search-based heuristics belong to this group. Local search algorithms start from an initial solution and iteratively replace the current solution by other solutions belonging to a neighborhood of the current solution. Some local search-based metaheuristics applied to humanitarian logistics problems are Iterated Local Search (vehicle routing), Tabu Search (vehicle routing, large scale distribution) or Variable Neighborhood Search (water distribution, infrastructure recovery). In evolutionary computation, that also fits into this first group, the solutions form populations in evolution. In each iteration, several modifications are applied to the solutions of the current population to generate solutions of the population of the next generation. Successful evolutionary computation algorithms applied to humanitarian logistics are Genetic algorithms (infrastructure modeling, inventory planning, vehicle routing, evacuation, etc.), but other evolutionary computation methods have been used as well (evacuation, last mile distribution, etc.).

The second group is formed by constructive-based metaheuristics. Constructive algorithms generate solutions by iteratively adding elements to an initially empty solution until a feasible solution is complete. Some learning techniques may be integrated to improve the construction of future solutions during the process. In some humanitarian logistics problems arising in an insecure environment, vehicles are forced to travel together forming convoys, so they can be escorted. In such problems, the strong links between the elements of the solutions make it difficult to apply metaheuristics of the first group, since any small alteration of a solution can make it infeasible. Instead, metaheuristics of this second group fit much better. Two of these metaheuristics that had been applied to humanitarian logistics are Greedy Randomized Adaptive Search Procedure (GRASP) and Ant Colony optimization.

Multicriteria Optimization for Humanitarian Logistics

An especially rapidly growing line of research in humanitarian logistics uses multicriteria optimization methods, which are necessary due to the multiple actors involved and the multiple objectives present in these types of operations. There are extensive reviews on multicriteria optimization in humanitarian logistics (Gutjahr and Nolz, 2015). Usually, criteria are classified in three categories: efficiency (i.e., cost), effectiveness (i.e., time, coverage, travel distance, reliability, and security) and equity (fairness), and these criteria appear in different combinations and with different priorities in predisaster and postdisaster models.

Some of these criteria have commonly been considered in the commercial logistics literature, but others are specifically concerned with humanitarian operations. In this way, poverty and armed conflicts increase security issues, and may affect not only the objectives, but also the very nature of the transportation models, forcing the vehicles to travel in convoys to increase security or the possibility of being escorted. The same applies to the equity criterion, especially important in the humanitarian field. It is usually modeled as fairness in distribution, allowing transport to reach far and costly demand points, or fairness in the delivery times.

Time and cost are the most common criteria in transportation models. Regarding time, papers propose the total waiting time, the sum of travel times, the sum of delivery times, or the maximal time of operation. Cost criterion usually aggregates fixed and variable costs. The coverage criterion refers to the total met demand, and finally, reliability is related to the probability that the network arcs will be available for transportation and with the probability of being able to follow the routing plan designed.

Regarding solution approaches to address multicriteria models, they have been classified in scalarization (aggregation of single objectives), Pareto optimization, goal programming, and compromise programming. When many different and conflicting criteria are present, the last two approaches are the most commonly used in humanitarian logistics. Goal programming gives a solution that satisfies the goals of decision makers, or if this is not possible, the closest solution to such goals, even if the solution is not always efficient. On the contrary, under some regularity conditions, compromise programming ensures that no other solution has better or equal values for all the criteria. Both methods are related through the mathematical concept of distance to a given reference point. Both can include weights for the criteria, to capture the importance given by the decision makers to each of them. When the importance of different criteria cannot be compared, as could be the case between saving human lives and cost of operation, it is possible to define priority levels and use lexicographical optimization to prevent any trade-offs between them.

Acknowledgment

This work has been supported by the Government of Spain, grant MTM2015- 65803-R.

References

- Altay, N., Green, W., 2006. OR/MS research in disaster operations management. *Eur. J. Operat. Res.* 175 (1.), 475–493.
- Ferrer, J.M., Martín-Campo, F.J., Ortuño, M.T., Pedraza-Martínez, A.J., Tirado, G., Vitoriano, B., 2018. Multi-criteria optimization for last mile distribution of disaster relief aid: test cases and applications. *Eur. J. Operat. Res.* 269, 501–515.
- Gutjahr, W., Nolz, P., 2015. Multicriteria optimization in humanitarian aid. *Eur. J. Operat. Res.* 252, 351–366.
- London Resilience Group, 2018. London Resilience Mass Evacuation Framework. Available from: https://www.london.gov.uk/sites/default/files/mass_evacuation_framework_2018_v3.0_0.pdf.
- Ortuño, M.T., Cristóbal P., Ferrer J.M., Martín-Campo, F.J., Muñoz, S., Tirado, G., Vitoriano, B., 2012. Decision aid models and systems for humanitarian logistics. A survey. In: Vitoriano, B., Montero, J., Ruan, D. (Eds.), *Decision Aid Models for Disaster Management and Emergencies*. Atlantis Computational Intelligence Systems Series 7:vii-x. Atlantis Press, Paris, France, 17–44.

Further Reading

- Habib, M.S., Lee, Y.H., Memon, M.S., 2016. Mathematical models in humanitarian supply chain management: A systematic literature review. *Mathemat. Prob. Eng.* 2016, 1–20.
- Hoyos, M.C., Morales, R.S., Akhavan-Tabatabaei, R., 2015. OR models with stochastic components in disaster operations management: A literature survey. *Comput. Indust. Eng.* 8, 183–197.
- Liberatore, F., Pizarro, C., Simón, C., Ortuño, M.T., Vitoriano, B., 2012. Uncertainty in humanitarian logistics for disaster management. A review. In: Vitoriano, B., Montero, J., Ruan, D. (Eds.), *Decision Aid Models for Disaster Management and Emergencies*. Atlantis Computational Intelligence Systems Series 7:vii-x. Atlantis Press, Paris, France, pp 47–74.
- Ozdamar, L., Ertem, M., 2015. Models, solutions and enabling technologies in humanitarian logistics. *Eur. J. Operat. Res.* 244, 55–65.
- Van Wassenhove, L.N., 2006. Humanitarian aid logistics: supply chain management in high gear. *J. Operat. Res. Soc.* 57, 475–489.

Event Logistics

Rev. Ruth Dowson, Dan Lomax, UK Centre for Events Management, Leeds Beckett University, Leeds, United Kingdom; and Macaulay Hall, Leeds Beckett University, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	201
Traffic Management	202
Transporting People	203
Types of Vehicles and Their Uses	204
Transport Logistics Considerations in New Build Permanent Event Venues	204
On Tour Perspectives	204
Transport Economics and Events	205
Sustainability	205
Illegal Events	205
Car Parking	206
Public Transport	206
Biographies	206
See Also	207
References	207

Introduction

This article discusses the role that transport plays in the context of events and event logistics. In the 21st century, events dominate our cultural landscapes and inhabit the lives of people of all ages, backgrounds, religions, nationalities, and ethnicities. For the last 2 decades, the study of events has entered the academic curriculum, with researchers and practitioners developing a growing range of literature that instructs and analyses, questions and resolves, written by, for, and with those who deliver events, and, increasingly, those who wish to study events management, as the field professionalizes around the world.

What we now know as “events” bring people together in a specific place, for a specific period of time, to engage in experiences and activities. This meeting together is with a purpose. Such gatherings may comprise people that are known to each other, such as family or close friends, or they may bring strangers together who have never before encountered each other and who may never connect again. Events range from small, intimate affairs to complex global mega-events engaging with hundreds of thousands, or even millions of people in one place. A major component of events is the logistics of bringing those people together to one location, ensuring that all the resources needed for the event are present in a timely fashion, whether the event is held in a permanent purpose-built or adapted event venue or in one that has been temporarily created for the specific event. In observing the world around us, we can identify events that contain cultural meaning, events for business purposes, events with political and state functions, personal celebrations and rites of passage, leisure activities and entertainment, education, and sport (Bowdin et al., 2011; Getz, 2005; Raj et al., 2009).

Transport and event logistics are vital constituent parts of the event planning process, as shown in the model in Fig. 1.

In telling the story of the development of an event, there are multiple transport implications and impacts on the event logistics that are necessary to enable the event to proceed. Travel is a vital element, as events bring people together, and much of the planning process involves managing logistics, which inevitably includes transport, whilst pre-event logistical considerations also include organizational aspects such as site visits and enabling face-to-face communications through meetings for planning and coordinating with suppliers to outsource support services. Post-event activities include transport logistics through face-to-face debrief meetings and evaluation reporting.

A key element of event logistics that requires transport is the onsite delivery of goods and services to and from the chosen event venue: these will take place pre-event, during the event, and post-event. The range of deliveries will vary from one event to another, and the type of venue will influence the requirement for deliveries onto the site. There are many different types of venues, and, increasingly, there are “hybrid” venues that have a primary function that may or may not relate to events (Hassanien and Dale, 2011). The venue plays a key role in defining and creating the event experience. Financial pressures and concerns about making the most of a space, combined with the burgeoning requirement for events venues, have led many organizations to open their buildings and spaces up for hire for the purpose of hosting events, in return for a financial contribution to enable these organizations to survive. Whether operating on a commercial basis or not, event venues range from purpose-built convention centers, hotels, and sports stadia to fields, parks, and churches, and from historic houses to schools and universities. Meanwhile, many sports organizations responsible for managing football, cricket, or rugby grounds are keen to increase the financial contribution by hiring their facilities out for other non-sports events (John et al., 2013). All of these venues require access for deliveries, as well as access for people working at and attending

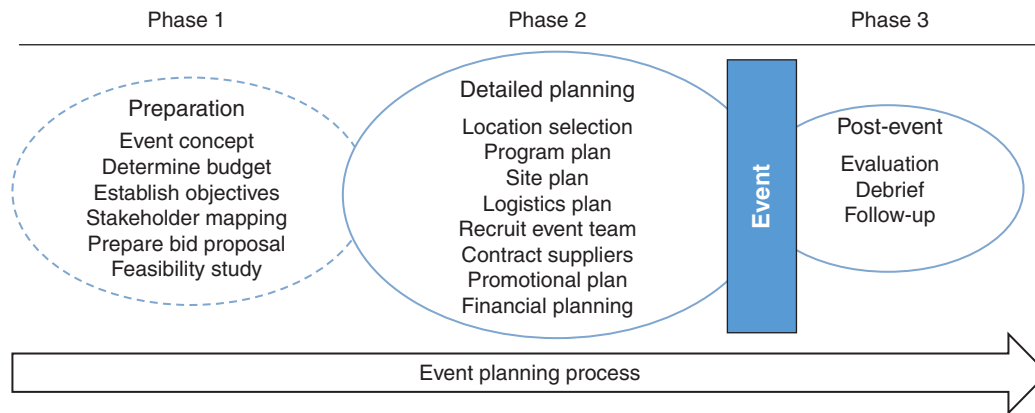


Figure 1 The event planning process (Dowson and Bassett, 2018, p. 24).

the event, before the event, during the event, and even after the event. However, many event creators choose to use greenfield sites for their event locations. This means that literally everything has to be brought onto the site, to enable the building of a temporary event venue, and returning the site to its original state afterward. The adage “leave no footprint” may be more popular with those who affirm sustainable values, but, whether it’s a festival, a weekly park run, or even an illegal rave, the carrying away of event detritus and removal of the physical resources needed to deliver the event require managed logistics.

Traffic Management

For event participants, arriving at the venue and leaving the venue are important contributors to the overall event experience, as the initial and final components, and all too often include enduring long queues for toilets, refreshments, and cloakrooms inside the venue, and queues for traffic when leaving the venue. The impact on the local community in the area surrounding the event should be taken into consideration, whether roads are signposted, closed completely, rerouted as one-way, or simply clogged up with slow-moving vehicles trying to enter the event car park or drop off attendees or supplies; meanwhile, local residents are unable to go about their daily lives. Event timing is an important factor, for example, combining an event start with the weekday evening rush hour is more likely to lead to traffic queues as tired commuters, anxious to get home, clash with excited event goers, eager to see their favorite band playing. Weekend traffic may have different peaks and troughs, or be equally busy, depending on the area, and so traffic management plans need to include local knowledge and experience of road conditions for travelers at all times. For example, in Yorkshire, United Kingdom, travelers hoping for a smooth journey to the English East Coast over the August Bank Holiday weekend are very likely to sit in traffic for hours as floods of young people make their way north of the city for the popular annual Leeds Festival, and have sympathy for parents dropping off and collecting their offspring and assorted friends, as they queue to get onto the site, crawl within the site boundaries to park, and then queue again to make their way out, a process that can take hours. In 2016, technical and production staff arriving to set up stages at the Glastonbury Festival spent up to 13 hours sitting in their cars and trucks queuing to get onto the site because some festival attendees had decided to arrive early, blocking the narrow country lanes with their parked vehicles when they were rightly denied early access by festival security staff.

The legal and regulatory requirements for managing transport elements of an event can be found in health and safety guides promoted by industry associations to develop good practice. For the topic of Transport Management, such advice is available in an online publication, “The Purple Guide,” which contains a specific chapter that addresses transport logistics onsite at events held in the United Kingdom (The Event Industries Forum, 2014). The chapter provides a useful health and safety resource for transport logistics at events, even in wider international contexts.

The Purple Guide details the following counsel:

- Once on an event site, the movement of site vehicles and traffic are found to be a major cause of serious and fatal accidents, so organizers are urged during the planning stages of the event to assess the risks of vehicle movement at all times, from pre-event setup and post-event breakdown, to imposing considerable restrictions on movements during the event, when the site is likely to be full of people.
- A traffic management plan is required as part of the event license, to control not only internal site traffic but also to consider the implications for external traffic and public highways leading to the event location and venue. Planning takes account of the requirement that pedestrians and vehicles should be segregated onsite at all times, which leads to fenced-off internal traffic routes within an outdoor festival site, for example, where possible, incorporating one-way systems and minimizing the need for reversing vehicles.
- Appropriate training and authorization processes are required for drivers and traffic marshals within the traffic management plans, with protocols in place and communicated to all.

Appropriate official signage should be used. In the United Kingdom, the Purple Guide advises that event organizers should agree detailed traffic sign plans and schedules with police and local authorities, and consider using a sign contractor for events where the majority of people will be arriving by cars or coaches ([The Event Industries Forum, 2014](#)). On-street signs should be placed and erected following guidelines in the Purple Guide, and by accredited personnel only.

In the United Kingdom, only the police can legally undertake traffic regulation on the public highway, although they may accredit some trained and approved stewards to work under their direction under the PATO (Police Accredited Traffic Officers) scheme. On a private event site, stewards can direct traffic. Logistical requirements will determine how many stewards will be required (directing traffic and managing closed or access-only roads), and all stewards should receive correct training and PPE (Personal Protective Equipment), such as high-viz clothing and appropriate weather protection.

The UK Construction Industry Research and Information Association (CIRIA) issued Guidance on Designing for Crowds in 2008 ([Rowe and Ancliffe, 2008](#)). This guidance considers the operational requirements of venues, and includes the following transport and logistics aspects: routes and way finding including signage, security (gates, search areas, CCTV, and communication), disabled users, deliveries, access for queuing and ticketing. The CIRIA Guidance is intended for designing venues for crowds but is useful for all aspects of venue design and encourages event organizers to engage all stakeholders actively in the design from the start of the planning process. Transportation and access considerations form a vital component of the guidance, which encourages event development teams to visit similar, existing environments and meet with operational staff to identify what works well and what does not work well in the venue and location. This information can be effectively used to develop the transport planning documentation, with special consideration for people with disabilities and impaired mobility, and for transporting goods and services onto the event site.

In a post-9/11 world, in which vehicles can be weaponized, events have been targeted by suicide bombers in terrorist acts. Two such attacks on events took place on November 13, 2015, in Paris, France, where the entrance to the Stade de France (football stadium) was struck by three suicide bombers during a football match, and a mass shooting took place during a concert by the Eagles of Death Metal in the Bataclan Theatre, killing 130 people and injuring another 413. On September 22, 2017, at the end of a concert by the singer Ariana Grande at the Manchester Arena in Manchester, England, a suicide bomber killed 22 people and injured 139. Post-Manchester there has been a focused response by the security services (including police) and event planners, taking on board the security and safety impacts in event logistics and planning, which impacts on transport arrangements for people and deliveries of goods to and from event venues. Following attacks in Nice, France, on July 14, 2016, when a truck was deliberately driven into crowds celebrating Bastille Day, governments and police forces have reviewed event safety measures. Security technologies, surveillance, and policies to guide best practice are readily available, along with experts to implement and advise ([Safe Events, 2019](#)).

Transporting People

A key component of event logistics involves the transport of people to and from the event location. The range of people involved in an event can be assessed through identification of the different stakeholders. Every event has multiple users who access the site, and different events and different sites will have different logistical requirements. These users might include the following groups:

- Clients
- Staff (e.g., technical staff, security, marshalls, cleaners, venue staff, caterers, event managers, event production staff, etc.)
- Performers and artists
- Promoters
- VIPs
- Event attendees
- Users with disabilities and mobility issues
- Emergency services

Each of these groupings of people has different logistical requirements, some of which are potentially in conflict with others. Staff working onsite at an event may be in-house or outsourced, but they will need access to the site prior to the event (sometimes for days, weeks, or even months), throughout the duration of the event, and after the event (again, sometimes for days or weeks, as the site is returned to its former state). This access may involve traveling in on a daily basis, or less frequently, perhaps from long distances, as well as traveling around the site itself. Staff will need to familiarize themselves with the site layout, and especially the back-of-house areas. For a festival site, this includes special access routes that are not open to event attendees, but which are specially required for deliveries, emergency vehicular access, transportation of event production resources, and waste removal (including emptying toilets). Event performers and some VIPs may need specialist access, with more direct routes than those available to event attendees. For example, the site may need to be accessed via helicopter, to avoid road congestion and ensure timely arrivals onsite. Access will be restricted onto and within the event site where groups and individuals will be identified onsite by their accreditation, usually in the form of passes, stickers, or wristbands, which may be timebound, as well as enabling or restricting access to specific zones, areas, or buildings/structures, defining back-of-house and front-of-house accessibility. Increasingly, such labeling identifies the named person allowed onsite, through the use of scanned barcodes or QR codes. Users of the site with disabilities and special requirements may need to access different routes within an event site, and music festivals will often have specially reserved camping areas available near to facilities and event activities. Wheelchair and other disabled access and mobility considerations are an important part of

event logistical planning, with expert advice best taken from those with specific experience of the requirements. The placement and size of blue badge parking sections for people with disabilities and mobility issues should reflect the numbers attending an event, but often assumptions are made rather than plans for known participants and event staff. Some festivals provide shuttle buses from accessible campsites to the event stage areas, and most include a free pass for a carer. Electric wheelchairs for use on an event site may require electric charging facilities to last through the whole event.

Types of Vehicles and Their Uses

Event transport can be provided by specialist companies, providing quality transportation in a range of vehicles, from cars, MPVs, mini coaches, and large coaches, with professional experienced drivers and chauffeurs with local knowledge. Such companies provide quality transportation to the events industry, whether for meetings, conferences, sports events, and festivals, supplying services such as ground support, liaison with event organizers and venues. Recognition of sustainability issues includes the development of carbon offset programs, and quality attention to detail.

Examples of transport use for events include:

- Transport from hotel to dinner venue for 200 conference delegates, using three coaches out and two coaches providing a shuttle service back, every 15 minutes, from 10.30 pm to midnight.
- Passenger transfers, with luggage distribution, for an international conference, with 12,000 transfers over 3 days: 750 conference delegates arriving from all over the world, staying in 12 different hotels in and around a city.
- Bespoke chauffeur-driven services, for airport transfers, client show rounds, local tours and transfers, long-distance meetings, corporate entertaining, ceremonies and dinners, using executive cars or people-carriers.
- For disabled or infirm event participants, providing assistance getting in and out of vehicles, escorting to and from buildings, with or without walking aids or folding wheelchairs accommodated within the vehicle. Specialist support for visually impaired passengers, with or without guide dogs. Transporting multiple guide dogs.
- Corporate and special event vehicles for national and international coach tours, sports teams, and music tours.

Transport Logistics Considerations in New Build Permanent Event Venues

Considering the transport logistics of a new build permanent event venue requires significant planning, with documentation developed in collaboration between multiple stakeholders, including local authorities, site owners, site management (if different), and specialist consultants, in a Design and Access Statement. This statement contributes to the supporting documentation submitted within the planning application process. New venues and their surrounding space are explicitly designed to promote a sense of arrival, enhancing attendees' expectations, whilst meeting requirements for access and egress for pedestrians and vehicles throughout normal operations and in emergencies. Pedestrianized space is prioritized, for people to gather as they wait to enter the building for a performance, and providing safe capacity for departing crowds.

New access roads and enhanced pedestrian footpaths and cycle routes may be needed to improve existing access to the site of a new venue. The aim is to minimize vehicular access and encourage event attendees to travel by public transport and on foot. Access to the site may be designed to enable deliveries and other vehicular access and departure with minimal disruption to the existing traffic flow. Environmental sustainability concerns and building on existing public transport links discourage event participants arriving by car, using existing car parks. Disabled parking is prioritized, with disabled parking bays adjacent to the venue, with additional disabled parking in nearby public car parks.

Significant attention is paid to enabling connectivity, including arrival routes by foot and bicycle, developing new access points, with well-lit footpaths and cycleways to promote the safety and security of pedestrians and cyclists in the area. Environmental sustainability considerations closely combine with social and economic concerns to improve surrounding areas.

On Tour Perspectives

Touring events are common, as musicians, sports teams, and even conferences travel nationally and internationally, performing, playing, or delivering their messages. The logistical structure of a tour can be complex and is subject to multiple factors (which sometimes appear illogical to the observer). The road show is a convenient method of engaging with multiple audiences in different geographic areas, especially in the era of experimental marketing and product activation events. The events team managing such a road show might also be involved in the delivery of other events, as well as coordinating suppliers, from audiovisual technical support to trained facilitators and industry specialist speakers, which may be consistent, or vary from one event to another, as the event moves around the country to the next location and venue. Delivering two, 36-hour business events a week, should be easy enough, with travel within the weekday period being as limited as possible. However, other factors may conspire to force contiguous events to be held geographically further apart than seems logical, such as the availability of suitable (and appropriately priced) venues that meet diverse needs of participants and organizers.

In the music sector, in which live events have become the major source of income, the requirement for bands and artists to go on tour to generate income has seen a significant growth in the scale and frequency of tours. This is true of small and emerging acts working to build a following, just as much as it is of mega-artists. At the largest scale, acts will have multiple stage set ups (U2's 2017 tour had three touring stage sets each requiring 16 trucks), enabling complex stage production involving stage build times of several days, whereby one stage is moving, one is building, and the other is being used for the show. In addition, 30 trucks worth of lighting, sound, and stage production traveled from stage to stage. This level of complexity requires highly sophisticated logistical control and large numbers of personnel (again U2 in 2017 had a touring crew of 196 traveling on nine buses) and a significant industry of expert tour managers and operators has developed as a result.

In addition, specialist suppliers will regularly "parachute" in to an event location on another continent, highlighting the global nature of the events industry and the impact that such transport logistics has on sustainability issues arising from events, especially in terms of air travel.

Transport Economics and Events

In addition to the important logistical considerations relating to the various transport-related aspects of delivering events, it is important to consider the impact that this will have on both the business model of event organizers and the impact on the wider economy. Transport economics aims to provide a means of understanding individual, business, and wider decisions about transport decisions and in the case of events this will most clearly be seen in terms of understanding the relationship between travel distance and cost incurred by attendees and the costs of staging in terms of venues and travel costs incurred in event staging. A successful event business model will seek to manage the ability of a chosen venue to deliver an excellent event experience alongside the willingness of attendees to travel the required distance to access the show and the cost of transporting equipment and personnel to the site.

Sustainability

Environmental concerns are increasingly recognized in the events industry; as events bring people together, they are heavily dependent on resources, impacting on the environment (Jones, 2017; Powerful Thinking, 2015; Raj and Musgrave, 2009). Events can generate significant waste, and put pressure on local resources such as transport and energy. As an example, it has been estimated that audience travel alone accounted for 80% of the carbon footprint of UK music festivals. The quality benchmark, Sustainable Events ISO 20121, published in 2012, was developed by the events industry to provide an international quality standard for events management, and all event supplier organizations can be accredited (ISO, 2012). Obvious solutions for event transport would focus on the use of local public transport. However, many events, such as music festivals, are often held in less accessible places with no public transport links. A UK music festival, Greenbelt, undertook research into how people arrived at the festival, observing the type of vehicle and the number of occupants, as they arrived for the start of the festival, using the format in Table 1.

Subsequently, festival organizers were able to encourage car-sharing as one method of improving overall sustainability. Other festivals included coach transport in the price of the festival ticket, with access to the site limited to official coaches, whilst other festivals developed specialist online programs for attendees to share cars and offer lifts. Some event organizations develop carbon-offset programs to support sustainability aims.

Illegal Events

Whilst the events industry and event management studies are mostly concerned with legal events, there are lessons to be learned from illegal events such as raves or freeparties, where the logistical planning process remains hidden and secret, until the night of the event, when there may be as little as a 1-hour window for event communication to take place (Dowson et al., 2015). In planning a rave, organizers select a venue for its hidden nature, using out-of-the-way outdoor countryside locations or urban abandoned buildings (and sometimes, not-so-abandoned buildings). The continuous search for new venues is an activity that runs alongside daily life, interspersed with regular forays that build complete familiarity with a chosen site, often undertaken under cover of darkness, and, almost inevitably, involve some form of subterfuge to ensure that the location remains undetected.

Table 1 Format for counting arrivals at Greenbelt Festival

Observation #	Minibus	Type of vehicle	Truck	SUV	Passenger car	Van	Motor home
# of occupants	1	2	3	4	5	6	6+

until the last moment. The secretive arrangement of logistics is paramount to the success of such an event. Transporting sound equipment to the depths of a forest or into the bowels of an abandoned warehouse without arousing suspicion of neighboring residents or the authorities is quite a feat. Facilitating the transport of thousands of people to arrive at the same time in a location that may be far from anywhere or in the middle of a city, with an hour's notice, is an accomplishment that many conventional event organizers would envy. Having the foresight to hide the trucks that delivered the sound systems a mile from the event location underneath a motorway flyover, to avoid confiscation by police, demonstrates skilful and well-considered planning processes.

Car Parking

Car parking for events offers a choice between self-regulated and controlled parking. A city-center event venue may rely on external car parks in the vicinity. Self-regulated parking in an unstructured and unmarked space such as a field requires minimal event staff but is less efficient than planned parking, so more space is required and problems arise if everyone arrives or leaves at the same time. A controlled car park is more efficient but requires large numbers of staff at entry and exit times. Car park requirements include marshals to direct traffic, adequate signage and signposting, and a numbered reference system for drivers and passengers to find their cars at the end of the event. For events that take place at night, adequate lighting is important along with considerations around safety of pedestrians (who may or may not be intoxicated). Specialist businesses exist to manage car-parking arrangements for large events, whether in permanent locations (such as sports stadia) or open fields (SEP, 2019; TRACSIS, 2019). Links to road networks are essential, and for larger urban events it is worth using several separate car parks that link to different parts of the road networks to avoid or restrict queues and congestion. Event car parks should include a separate entrance and exit, and a managed holding area with space for vehicles to turn in for drop-offs.

Public Transport

Public transport links aid the accessibility of event venues and enable a wider attendance. Some events specifically choose a venue due to its proximity to bus, coach, train, and underground stations. Factors to consider within public transport hubs include: effective signage (e.g., saying which bus is going where), space for safe queuing, boarding, and alighting of passengers, addressing capacity issues (e.g., are additional buses or trains needed), accessibility, station and platform capacities, stewarding and signage at bus or train stations, and the impact on existing services. The Purple Guide advises event organizers to consult with transport authorities about introducing new or enhancing existing services to service an event. Negotiations with train operating companies might include the need for bigger trains, extra, or later-running services.

In the United Kingdom, event traffic management comes under the requirements of The Road Traffic Act 1984, as inserted by the Road Traffic Regulations (Special Events) 1994. For large-scale events, traffic planning and implementation should take a multi-agency approach, working with traffic authorities such as the highways agency (for trunk roads) and the local authority (for other roads), the emergency services (police, fire, ambulance, etc.), local bus companies and businesses on routes taken by event participants, local car park operators (if separate from the venue), local residents, hospitals, travel assistance organizations (e.g., Automobile Association, Green Flag), and the media. Advance warning should be given of events and advance warning notices displayed.

Biographies



Rev. Ruth Dowson has been a Senior Lecturer at the UK Centre for Events Management at Leeds Beckett University, United Kingdom, since 2007, and was previously an events professional with 25 years' experience in the events industry. She holds an MBA from Bradford University School of Management, United Kingdom (1985), and an MA in Theology and Ministry from the University of Sheffield, United Kingdom (2013). Ordained in the Church of England in 2012, her research combines her passion for church and events. Her published research interests focus on events and church, the eventization of faith, venuefication, and the use of religious buildings for events. She is also co-author with David Bassett, of "Event Planning & Management," now in its second edition, and is a Fellow of the Higher Education Academy.



Dan Lomax has been a Senior Lecturer at the UK Centre for Events Management at Leeds Beckett University, United Kingdom, since 2010 following a varied 20-year career in music event production and management. He holds an MSc in Events Management from Leeds Metropolitan University (2003) and an MA in Transport Economics from the Institute of Transport Studies at the University of Leeds (1995). His research focuses on entrepreneurialism, business, and the economics of the Live Music Industry, and published work includes studies of both music festivals and music events. He is a Fellow of the Higher Education Academy.

See Also

Introduction to Transport Economics; Emergency Vehicles and Traffic Safety; Parking Lots; Safety Culture

References

- Bowdin, G., Allen, J., O'Toole, W., Harris, R., McDonnell, I., 2011. *Events Management*, third ed. Butterworth Heinemann, London.
- Dowson, R., Bassett, D., 2018. *Event Planning and Management*, second ed. Kogan Page, London.
- Dowson, R., Lomax, D., Theodore, B., 2015. Rave culture: freeparty or protest? In: Lamond, I.R., Spracklen, K. (Eds.), *Protests as Events: Politics, Activism and Leisure*. Rowman & Littlefield, London, 215.
- Getz, D., 2005. *Event Management and Event Tourism*, second ed. Cognizant, New York.
- Hassanien, A., Dale, C., 2011. Toward a typology of events venues. *Int. J. Event Festival Manag.* 2 (2), 106–116, doi:10.1108/17582951111136540.
- ISO, 2012. Sustainable events with ISO 20121. Available from: <https://www.iso.org/publication/PUB100302.html>.
- John, G., Sheard, R., Vickery, B., 2013. *Stadia: The Design and Development Guide*, fifth ed. Routledge, New York.
- Jones, M., 2017. *Sustainable Event Management: A Practical Guide*, third ed. Routledge, Abingdon.
- Powerful Thinking, 2015. The show must go on: environmental impact report and vision for the UK festival industry. Available from: <http://www.powerful-thinking.org.uk/site/wp-content/uploads/TheShowMustGoOnReport18.3.16.pdf>.
- Raj, R., Musgrave, J., 2009. *Event Management and Sustainability*. CAB International, Wallingford.
- Raj, R., Walters, P., Rashid, T., 2009. *Events Management: An Integrated and Practical Approach*. Sage Publications, London.
- Rowe, I., Ancliffe, S., 2008. *Guidance on Designing for Crowds*. UK Construction Industry Research and Information Association (CIRIA), London.
- Safe Events, 2019. Available from: <https://www.safeevents.ie/>.
- SEP, 2019. SEP events. Available from: <https://sepevents.co.uk/>.
- The Event Industries Forum, 2014. The purple guide: to health, safety and welfare and music and other events. Available from: <https://www.thepurpleguide.co.uk/>.
- TRACSIS, 2019. Traffic and data services. Available from: <https://tracsistraffic.com/event-traffic-management/>.

Reverse Logistics

Dale S. Rogers*, **Ronald S. Lembke†**, *Arizona State University, W. P. Carey School of Business, Tempe, AZ, United States; †University of Nevada, Reno, NV, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	208
Definition of Reverse Logistics	209
Strategic Importance of Drains	209
Reverse is Different	209
Overview of Reverse Logistics Activities	209
Strategic Use of Returns	210
Retail Return Rates	210
Lowering Transaction Risk	210
Changing Policies	210
Omnichannel Synergies	211
Types of Returns	211
Centralized Returns Centers	211
Disposition Decision-Making	212
Disposition Options	212
As New	212
Open Box—New	212
Open Box—Returned	212
Refurbished to New Standards	212
Refurbished, Cosmetic Imperfections	212
As Is	212
Sell to Secondary Market Broker	212
Component Harvesting	213
Recycle/Recover Primary Materials	213
Landfill	213
Recalls	213
Food Recalls	213
Unsaleables	214
Secondary Markets	214
Description of the Secondary Market	214
Financial Issues	214
Disposition Decision-Making	215
Secondary Market Size	215
Secondary Market Flow	215
Total US Secondary Market	216
Secondary Market Relative To GDP	216
Reverse Logistics Metrics	216
Value of Returned Product	216
Conclusions	217
Biographies	217
References	218
Further Reading	218

Introduction

Reverse Logistics (RL) capabilities within a firm and across the supply chain act as “drains” that purge waste and allow fresh product to take the place of old and bad product. RL systems bring product back from forward positions in the supply chain and move those items to locations where they may be resold or disposed of properly at a minimal impact on the environment. These “drains” are tasked with cleaning the unwanted product out of the network while contributing as much as possible to profitability (Rogers et al., 2013a). In order for these drains to function as effectively and efficiently as possible, the RL system must make strategic use of the “secondary market,” the network of buyers and sellers who specialize in merchandise that is not suitable for sale in the primary retail market.

Over 100 years ago, both Sears & Roebuck and JCPenney became successful retailers when they introduced “no questions asked” returns. They realized that this was the best way to keep loyal customers, and offered them cash. Both evolved from very simple roots to become retail powerhouses. Sears started in a small train station in Minnesota, while JCPenney began as a tiny general store in Kemmerer, Wyoming. Both of these retailers utilized a unique business philosophy, which included open, liberal return policies where consumers could bring products back with no penalty (Rogers et al., 2013a).

This innovation, which was designed to increase customer satisfaction, formed a risk-free relationship for their customers. This meant that their stores were likely to keep long-term customers because the risk associated with shopping there was reduced. It was safe to buy something from Sears or JCPenney because the consumer knew that if it did not function well, or it was the wrong size, or it just did not work, they could return the product and get their money back. For these two pioneering retailers, taking back products from consumers was a smart marketing move that over the years consistently translated to business success. This realization was the beginning of the investment in developing RL processes to handle the returned material.

Definition of Reverse Logistics

RL has been defined as: “The process of planning, implementing, and controlling the efficient, cost-effective flow of raw materials, in-process inventory, finished goods and related information from the point of consumption to the point of origin for the purpose of recapturing value or proper disposal” (Rogers and Tibben-Lembke, 1999a, 1999b). Forward logistics is concerned with the flow of product from the point of production toward the point of consumption. RL is concerned with product flowing in the opposite direction, which leads to many important differences. Materials in the reverse flow are moving toward another use, or disposal. Although retailers have long understood the importance of strong forward logistics networks, retailers were much slower to appreciate the importance and competitive benefits of a strong RL network (Carter and Ellram, 1998; Dowlatshahi, 2000; Rogers, 2011; Wang et al., 2017).

Strategic Importance of Drains

One important function of RL networks is to serve as a drain for the firm, a place to get rid of unneeded or unwanted products. Biological systems have processes for ridding themselves of waste material and retail systems need the same. By removing unwanted products from the network, RL processes allow a retailer to ensure that the product selection remains fresh and current, full of product that customers want to buy.

Many categories of customer products including clothing, beauty products, and housewares, have significant annual seasonality. At the end of a selling season, the firm must have a way to remove unsold product from its shelves, to make way for the next season’s products. With an RL network, the firm can place the unsold products into the network to quickly and easily remove all unsold products from their network. Without an RL network, the only way to get rid of unsold product would be to mark it down repeatedly, which would have bad consequences. Companies may over-estimate demand for a product, or market conditions may change, or other factors may lead to an over-abundance of an item. Drains also provide the retailer with the opportunity to dispose of these unwanted items. Retailers invest heavily to create a brand image, and stores filled with shelves or racks full of clearance-priced merchandise would not make a positive contribution to their brand. Retailers already offer a series of increasing markdowns toward the end of a selling system. If the product stayed behind after the end of the season, it would only further train customers to wait for better prices. Repositioning those items to an alternative distribution channel can be a better strategy so that consumers buy fresh product from their original destination.

Reverse is Different

As mentioned in the introduction, RL processes are different than typical forward logistics processes and systems. The processes required to manage RL are more complex than those needed to manage the forward flow of fresh product, while at the same time the amount of product in the reverse flow is smaller. In Table 1, some of those differences are identified.

Overview of Reverse Logistics Activities

RL networks perform a variety of important tasks. For a retailer, a product may enter an RL network for different reasons, and from different positions. From a retail location, a product may be returned by a customer, or put into the network by the retailer directly. For a customer return, where the consumer brings an item back to the retailer or ships it through the post to the ecommerce retailer, an assessment must be made of the condition of the item, and the disposition destination must be selected. In fact, RL systems that include predetermined dispositions are almost always better than taking the decision about where to send the item after it arrives at the place where it is being sorted.

Table 1 Differences between forward and reverse

<i>Forward</i>	<i>Reverse</i>
Product quality uniform	Product quality not uniform
Disposition options clear	Disposition options unclear
Routing of product unambiguous	Routing of product ambiguous
Distribution costs easily understandable	Distribution costs less understandable
Pricing of product uniform	Pricing of product not uniform
Inventory management consistent	Inventory management not consistent
Product life cycle manageable	Product life cycle less manageable
Financial management issues clear	Financial Management issues unclear
Negotiation between parties straightforward	Negotiation less straightforward
Type of customer easy to identify and market to	Type of customer difficult to identify and market to
Visibility of process transparent	Visibility of process less transparent

Source: Modified from *Tibben-Lembke and Rogers (2002)*

Product may be placed into the RL network directly from the retailer. This is especially true with ecommerce returns that are sent by the consumer directly to a processing facility. A retailer may have too many units of an item at one location, and use the RL network to reposition the items from a place of low demand to one of high demand. A retail item typically enters the RL network via the service desk at the store. At the service desk, the items are scanned and a disposition decision is typically made, particularly regarding whether the item will be sold at the current location, or sent elsewhere.

For a manufacturer, RL activities primarily consist of selling off unwanted new product. Because of economies of scale, a manufacturer may produce a larger batch of product than they have orders for. If enough additional orders are not received, the firm will have product that they may want to sell off. A similar situation arises when a large customer order is cancelled.

Strategic Use of Returns

For most retailers, the largest source of unwanted items is customer returns, but there are many reasons products enter the RL network.

Retail Return Rates

Return rates for general merchandise retailers in the United States are typically around 6%. This number has stayed fairly constant over recent decades. In fact, books about retailing in the 1950s also give this number as being typical. The rates are typically much higher for clothing, easily reaching 20% or higher. For ecommerce retailers, the typical return rate is much higher and for some products can range between 10% and 30%.

Lowering Transaction Risk

Returns may seem like an undesirable condition that should be avoided as much as possible. However, returns are actually very important and can become strategic. By making it easy for customers to return items, retailers lower the transaction risk of doing business with the retailer, which makes it easier for the customer to purchase from the retailer. As previously described regarding Sears and Roebuck and JCPenney, lowering consumer risk is a key benefit of a liberal returns policy.

Changing Policies

Companies would like to reduce the volume of returned product they must deal with, but they are well aware of their strategic benefits. However, companies have experimented in recent years to see what types of restrictions customers would accept. In the 1990s, Nintendo created a technological solution that allowed them to capture the serial number of each gaming console as it was sold. In addition to the serial number, the system also recorded the date and the retailer selling the item. If a customer tried to return the item to a different retailer, or outside of the allowable returns window, the system would prevent it. Nintendo spun the business off into a separate company called Siras, which is now part of InComm.

Their technology allowed Nintendo to avoid a very serious form of returns abuse. When a new gaming console system came out, customers would return consoles of the previous generation, and falsely claim to have purchased the unit recently, within the returns window, and Nintendo had no way to dispute that claim. Because the new system merely provided a way for Nintendo to enforce an existing policy, there was little customer push back.

Table 2 Types of returns

<i>Return types</i>	<i>Attributes</i>
Consumer returns	Returns to the retailer from consumers due to buyers' remorse or defects. This category is generally the largest returns grouping.
Marketing returns	Product returned from a position forward in the supply chain often due to slow sales, loading the trade, or the need to reposition inventory.
Damage returns	Returns of product to the manufacturer driven by the retailer's choice but not by the consumer. These are different from marketing returns in that the items are not first quality product because they were not checked carefully prior to shipment, or more likely, they were damaged in transit. These items may be returned to a reclamation center and then the cost is deducted from the manufacturer's invoice.
Asset returns	Recapture and repositioning of an asset such as reusable pallets, racks, or totes.
Product recalls	Usually initiated because of a safety or quality issue. Requires planning security. May be voluntary or mandated by a government agency.
Environmental returns	The disposal of hazardous materials or abiding by environmental regulations.

From its founding in 1912, LL Bean had a lifetime, no-questions-asked returns policy. In 2018 it changed the policy to one year, unless the item is defective, because too many people were abusing it. People were buying damaged old items at garage sales, and returning them for new products, which they then sold. Although a majority of customers accepted the policy as reasonable, there have been unsuccessful lawsuits challenging the policy.

Omnichannel Synergies

Although online shopping typically has a higher returns rate than in-person shopping, retailers have simplified the returns process by allowing customers to return the item directly to a store, instead of having to mail the product back in. This is often easier for the customer, and it brings the customer into the retail store, where the retailer hopes the customer will purchase a replacement item or additional items. This is one more way that returns can offer a competitive advantage. In 2017 Amazon purchased Whole Foods, and one of the services they began to offer was the ability to return items at Whole Foods stores, no shipping required.

In July, 2019 Kohls department stores announced a partnership with its online competitor Amazon, in which Kohls allows customers to return Amazon items at Kohls stores. In its limited trial before rolling out the program nationwide, Kohls said stores accepting Amazon returns saw sales growth of 8%, compared to 2% growth for the rest of their stores.

Types of Returns

There are several types of returns. It is not enough to merely know that returns are being generated. It is important that managers be aware of the source and type of return they are receiving. They should also understand the reason for the return. A firm needs to develop different strategies and systems to deal with each type of return (Table 2).

Centralized Returns Centers

If a retailer determines that an item is not to be sold in the store and needs to be sent elsewhere, it is frequently sent to a Centralized Returns Center, or CRC. At a CRC, employees have training and experience in assessing the quality of the items. They also have access to information systems about the number of units of different quality already available.

CRCs are used because companies have learned that it is unwise to process returns at a forward distribution center. Trying to do both in one facility creates the problem of "serving two masters." The primary focus of a forward distribution center is delivering new product to stores, in large volumes. Forward logistics supplies product that provides the firm with revenue. RL represents a cost. When things get busy and labor is constrained, as it always is, DC managers are likely to allow returns management to be ignored.

Also, the processes for forward and reverse goods are entirely different. An arriving pallet of new product is all in identical boxes, clearly labeled, and the products are interchangeable. RL is, by definition, exception-driven. When a pallet of returns arrives, it will contain many different products, some with packaging, some without. Some of the advantages of a central location to manage returned product are the following:

- An important factor for success in RL is the speed of processing. CRCs can help companies increase the speed of processing returns. As one industry professional observed, returned goods are not like fine wine. They do not get better with age.
- Because the items often lack original packaging, the more times they are moved around a facility, the more likely they are to become damaged. Quickly deciding what to do with an item and getting it out of the facility reduces the opportunities for damage.
- If an item has been discontinued or in some other kind of shelf pull, the sooner the item is disposed of in the secondary market, the better the likelihood of receiving a higher price for the units.

Disposition Decision-Making

A key part of RL is deciding which disposition to choose for an item, what the destination of the material should be. Every item must be carefully evaluated, to find the most valuable destination for it. Without a strong RL network, a retailer would be forced to landfill all unwanted product. Such a retailer would lose out on millions of dollars of potential revenues, which could easily make the difference between profitability and loss. Selecting the proper disposition option has important financial implications, therefore, and knowing the amounts that could be received for each potential option is a critical part of this selection.

In addition to different potential amounts of financial return, different disposition options also provide the seller with differing abilities to control where the product is resold, and the prices at which it is resold. Maximizing financial payments are an important consideration, but often these are secondary to being able to protect the value of the brand.

Disposition Options

The possible destinations for a product that has entered the RL network include the following, generally listed in order of decreasing revenues:

As New

If the item is unopened and can be sold as new, the retailer may be able to still make a profit on the item, but not as great as before. The original retail profit margin takes into account the cost of buying, transporting, displaying and selling the item. Accepting the item as a return and placing it back on the shelf does add labor cost, but most items still generate a positive profit.

Open Box—New

An open box item has been opened, but is still in new condition. However, most customers would expect to receive a discount off the original retail price.

Open Box—Returned

If an item has been opened and returned, it cannot be sold as new. Some retailers may be able to inspect an item to ensure it still performs to specifications and sell it, but at a greater discount off the original retail price. The value received will be determined by the condition of the item.

Refurbished to New Standards

Some items may require some repair or refurbishing before being resold. Refurbishing can be completed by the original manufacturer, by third parties contracted by the manufacturer, or by outside companies who purchase the items in the secondary market.

Refurbished, Cosmetic Imperfections

Sell as refurbished, but perhaps cosmetically inferior to new products. The value will be significantly influenced by the level of cosmetic imperfections.

As Is

An item sold “as is” may be fully functional, or perhaps not functional. Anyone buying such a product is accepting all of the risk of the transaction. Because most buyers do not have the knowledge or materials to repair most items, they would expect to receive a very large discount in compensation for accepting this risk. As with the related category of open box, the value is heavily influenced by the condition and degree of functionality of the item.

Sell to Secondary Market Broker

As described further, the secondary market is the term for all of the brokers and retailers who buy and sell product outside of the primary sales channel. When a retailer sells to a secondary market broker, the product may change hands multiple times before being sold to a retail customer. As the products pass between brokers, more detailed diagnostics will be performed on the item, and where economically feasible, repairs will be made and the item sold as refurbished.

Table 3 The 10 biggest product recalls

<i>Rank</i>	<i>Recall</i>	<i>Cost (as of March 2018)</i>
10	Tylenol	\$0.1B
9	Peanut Corp. of America	\$1.0B
8	Toyota Floor Mats	\$3.2B
7	Pfizer's Bextra	\$3.3B
6	General Motors Ignition Switches	\$4.1B
5	Samsung Galaxy Note 7	\$5.3B
4	Firestone Tires and Ford	\$5.6B
3	Merck's Vioxx	\$8.9B
2	Volkswagen Diesel Engines	\$18.3B
1	Takata Air Bags	\$24.0B (and counting)

Component Harvesting

If a product cannot be refurbished profitably, instead of disposing of it in a landfill or recycling any recyclable components, there may be major components such as computer boards, semiconductors, motors, etc., that can be removed and used or sold as replacements.

Recycle/Recover Primary Materials

If there are no components that can be harvested, or after any have been harvested, any metals, plastics, or glass materials can be recycled. The value of plastic or glass recyclables is generally quite low and depends on market conditions, which often depend on global trade realities. Metals, however, can be significantly more valuable. In some cases, then, the firm can receive a payment for the recyclable materials, or pay a lower cost for disposal than for mixed waste when landfilling.

Landfill

The very last option for any product is the landfill. All other options offer at least the possibility of some payment. If a product is sent to the landfill, the firm not only loses all of the investment it made in purchasing and processing the product, it now must pay someone to take the product away and bury it in a landfill.

Recalls

There are a lot of reasons for a product to be recalled, but typically they involve a liability issue. Some are mandated by a consumer product agency or the Food and Drug Administration. Faulty parts generally occur in two places along the supply chain: Either the product's original equipment manufacturer or at the facility that assembles the product (**Table 3**).

Traditional recalls often involve items at the consumer or the last step before the consumer, but it can be anywhere. For this reason, recalls are a major concern for supply chain managers. Recalls are split about evenly between voluntary and mandatory. When a product is recalled, the company has a responsibility to ensure that all potentially harmful products are removed from circulation. This means that all of the items should be collected and either destroyed, or retrofitted in such a way as to prevent anyone from being injured. This collection, destruction and refurbishing is an important RL function. Unlike typical customer returns, which come in gradually, however, recalls come in a short period of time, and may require that dedicated teams be created to process them.

Food Recalls

Many recalls represent significant risks to the public. In 2018, according to the US Department of Agriculture, 20.6 million pounds of food (10,000 tons) were recalled in 2018. Salmonella accounted for nearly 13 million pounds alone. Companies are working hard to stop recalls from becoming necessary, but given the complexity of today's global food networks, this will be difficult. This is why companies are also working hard to reduce the scope of any required recalls.

Current recalls involve such large volumes of product because companies have very limited visibility of exactly where products from a given production lot end up. Companies are working hard on technological fixes, such as blockchain, to provide better reporting. Despite the numerous headlines about blockchain, it is, in the end, merely a secure method for storing databases. The real obstacle to more limited recalls is the tracking of where products come from and where they have gone to. Technologies for tracing products throughout the distribution channel must be improved before recall volumes will be reduced.

Unsaleables

Crushed cans, rumpled boxes and expired cartons—“unsaleable” products continue to create challenges for the food and consumer product industry up and down the supply chain. A new GMA report on unsaleables costs and business practices shows that unsaleables are still on the rise, costing the grocery industry \$4.2 billion in 1998—up from \$4 billion in 1997. Pharmacy chain retailers such as Boots or Walgreens experience a much higher return rate than other channels, while mass merchants and club stores, such as Costco, generally experience much lower unsaleables rates than average.

When Walmart moved to open code dating and most retailers followed their lead, unsaleables increased dramatically. Every food or pharmaceutical product typically has an expiration date. In 1997 Walmart began to ask their suppliers to place expiration dates on the consumer package as “open codes” that consumers would be able to easily read. Before 1997 most expiration dates on consumer packaging were in a format that was unclear to consumers. Because of these changes, manufacturers went to reduce shelf life, so that consumers would not know how long the shelf life was. Some reduced shelf life by 90%. An example of how these changes impacted brand owners is Campbell’s Soup. They manufactured cans of condensed soup with three years shelf life. When Walmart asked them to go to open date coding they had 2 years of shelf life before the expiration date occurred. After open code dating, Campbell’s decided to move to 90 days of shelf life for marketing reasons. These changes have caused large problems for the industry.

Secondary Markets

Secondary markets effectively act like efficient drains that are supplied through RL systems. Having “drains” that can carry unwanted product to places where it can be resold is an important method for reclaiming value where it could otherwise be lost. The product sold into to the secondary market is generally being sold for less than its original price, so it might seem surprising to speak about RL systems contributing to profits (Tibben-Lembke, 2004).

As described further, to monitor progress in its RL processes, a company should measure the financial impact of returns on the firm and on other members of the supply chain. This can include measuring aspects of RL such as material scrapped and products resold through secondary markets that can reclaim value.

Description of the Secondary Market

Secondary markets have become a significant portion of domestic economic activity in the United States. Many companies recover, refurbish, remanufacture, and recycle product for additional use elsewhere instead of throwing the product away. Secondary markets act as system “drains” where excess inventory or assets can be recovered and resold.

The secondary market consists of retail store returns, surplus inventory, bankruptcy sale inventory and products that did not sell the first time around. It includes firms that specialize in buying close-outs, surplus, and salvage items. The secondary market firms then sell the product through their own stores or to other markdown retailers, such as dollar stores. The secondary market is fueled by the RL processes in a supply chain. eBay is the biggest player in the secondary market and many of the items for sale there are company returns.

As described above, RL flow is very different from the forward flow. Reverse flows are more reactive, with much less visibility. Because any supply chain may have product flowing in the reverse channel, the range of RL and secondary market development issues facing a particular supply chain will be as diverse as the number of supply chains.

Financial Issues

Another issue in examining RL practices is that the accounting methods that most firms use to manage their business are designed to manage the business in forward distribution. The costs and processes related to managing product that is moving backwards through the supply chain and being dispositioned to a secondary market typically do not show up neatly in an easily understood package in a firm’s balance sheet or income statement. In forward logistics, costs are usually well defined and well known.

Accounting systems are built to handle the comprehensive cost development for a product as it moves through the forward channel. The systems used for accounting for RL are more specialized, even if straightforward. Firms specializing in forward logistics are usually not very effective at managing the amount of detail derived from a product’s non-standardized journey backwards through the firm and the channel. For example, transportation costs on a firm’s income statement may be derived from both forward distribution and RL.

Disposition Decision-Making

Additionally, when deciding how to dispose of product or select the appropriate secondary market, there are a variety of options. Unlike forward distribution, RL managers may need to spend significant time determining where a particular item will be shipped because the condition of the item must be considered, as well as any agreements with the manufacturer regarding disposition. Manufacturers often have strict requirements about where product can be sold into the secondary market. They want to ensure that

the product is not advertised or sold any place that could potentially damage the brand equity. Therefore they often place significant restrictions on firms' secondary market activities.

Given the range in returned product quality, variations in pricing are to be expected. While the prices of products in the forward channel can vary due to factors such as value-added services or total quantity purchased, the pricing variability of returned products can be even more complex. Making quick disposition decisions into a particular secondary market can improve the profitability of selling into that secondary market.

Secondary Market Size

In 2018 the size of the secondary market in the United States was estimated to be approximately \$618 billion. This number was about \$100 million larger than the total amount of ecommerce revenues in the United States. (Rogers et al., 2010). The secondary market reduces waste and provides a stable income for a growing sector of the US economy, providing jobs for millions of workers around the world. Salvage products are not only an important part of the domestic economy, but they also make up a significant portion of US exports.

US consumer returns are estimated to exceed \$100 billion per year. This estimate is larger than two-thirds of all countries' total Gross Domestic Product (GDP). To illustrate the size of the market, Walmart is estimated to have processed over \$12 billion in consumer returns in the United States during 2009.

Secondary Market Flow

In this section, the reverse flows that feed various secondary markets are depicted. Fig. 1 includes the major secondary markets through which consumer electronics flow. The appropriate secondary market and disposition route depends on the type of product, its desirability in the marketplace, condition, and demand levels. As depicted in Fig. 1, there are several definable secondary markets for returned materials.

Fig. 1 reflects the complexity of the secondary market. The journey from asset recovery to its final destination is often filled with many twists and turns. The route to the secondary market consumer is not necessarily an efficient one.

Researchers have divided the secondary market into meaningful segments. These segments are as follows:

- Retailer,
- Manufacturer,
- International disposition,
- Auctions,
- Salvage dealers,

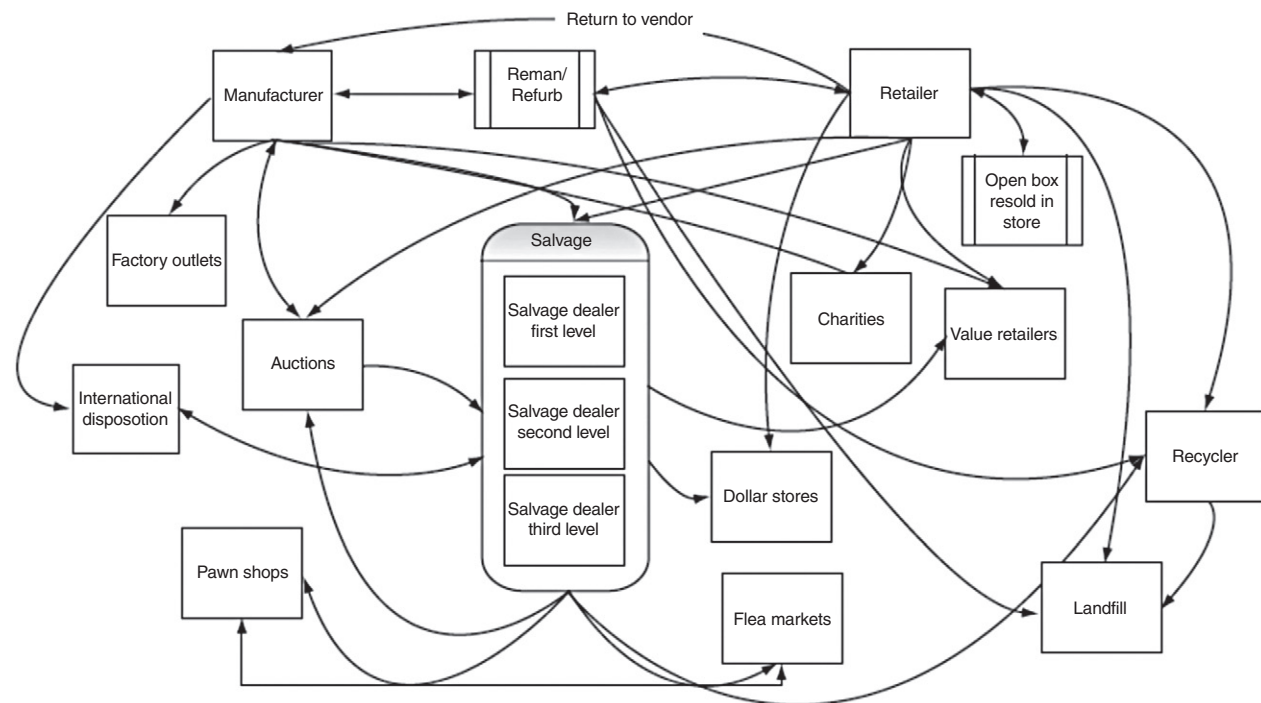


Figure 1 Structure of the secondary market.

- Factory outlets,
- Pawn shops,
- Dollar stores,
- Flea markets
- Charities, and
- Value retailers.

As the product flows from one level of the secondary market to the next, the dealers incur shipping and handling costs at each level of roughly 8%, and receive a profit margin of roughly 10%. The dealers purchasing directly from retailers typically purchase at a cost of approximately 15% of the original wholesale cost of the item. The price increases to approximately 18% for the Level 2 Salvage Dealers and 21% for Level 3 Salvage Dealers.

Total US Secondary Market

Summing the totals from each secondary market segment gave a conservative estimate of over \$618 billion in the United States for 2018. Without the secondary market to efficiently wick away product that cannot be sold through a primary market, many of these goods would be thrown away. It is useful to understand the size and structure of this market because of its environmental, social and economic impact. The secondary market provides environmental and social benefit, and is a significant portion of the US economy. In [Table 4](#), the various segments are compared.

Secondary Market Relative To GDP

Clearly, the secondary market is a large portion of the US economy. The methodology to estimate the size of the secondary market is deliberately conservative. The secondary market in the United States was estimated at \$618 billion in 2018, which represents more than 3% of US 2018 GDP. The concept of GDP has been defined to be “The market value of final goods and services newly produced within a country’s borders within a fixed time period.” This means that much of the value of the secondary market, as defined here, does not show up as a final good or service and is not included in GDP. Items that are not part of retail sales or the final step of a supply chain in an alternate channel are ordinarily not included in US GDP calculations.

Reverse Logistics Metrics

As with all other supply chain activities, firms are increasingly finding ways to measure and track the performance of their RL activities. Typical RL metrics are included in [Table 5](#).

Value of Returned Product

There can be several moderating factors on the pricing of product. Inventory levels, whether the product is current or at the end of life and seasonality all impact secondary market pricing. For example, in October a higher price can typically be gained than in March when retailers or manufacturers are more likely to be dumping product. If the market has a glut of a specific type of product, or a product is near the end of its lifecycle, the return from selling a product in the secondary market is likely to be lower. Also, product quality can be a factor in pricing. If the item is perceived to be a low-quality product or it is damaged in some way, then the likely return will also be lower.

Table 4 Size of the secondary market

<i>Sector</i>	<i>2018</i>
Factory outlets	\$48,439,708,714
Value retailers	\$59,192,930,000
Online auction houses	\$159,300,000,000
Dollar stores	\$38,769,174,400
Salvage dealer	\$230,861,990,000
Pawn shops	\$25,943,730,000
Flea markets	\$38,201,625,000
Charities	\$17,912,235,000
Total US secondary market	\$ 618,621,393,114

Table 5 Typical reverse logistics metrics

Dispositions
Amount of product to be reclaimed and resold
Percentage of material recycled
Freight costs and fuel utilized in transportation of returns
Handling costs and energy used in handling returns
Waste
Recycled materials
Total cost of ownership
Credit reconciliation
Transformation and repair decisions
Zero returns
Channel conflict
Declining product values
Time to cash
Store efficiency
Exploitation of vendor agreements
Maximize recovery

The type of consumer electronic product can impact pricing. For example, large flat screen televisions are likely to have a different depreciation curve than cell phones. Different brands of the same product can depreciate at diverse rates.

Conclusions

For firms selling product around the world, solving the problems of managing returned items and RL processes is critical to their ongoing success in an environment where consumers have become more demanding around the world. Understanding how to implement best RL practices is something that all supply chain managers need. It is important that managers can both efficiently ship new products and manage returned products that have entered the reverse flow.

Biographies



Dale Rogers is the ON Semiconductor Professor of Business at the Supply Chain Management Department at Arizona State University. He is also the Director of the Frontier Economies Logistics Lab and the Codirector of the Internet edge Supply Chain Lab ASU. Dale is the Leader in Supply Chain Finance, Sustainability, and Reverse Logistics Practices for ILOS—Instituto de Logística e Supply Chain in Rio de Janeiro, Brazil. In 2012 he became the first academic to receive the International Warehouse and Logistics Association Distinguished Service Award in its 130-year history. He is a Board Advisor to Flexe, Enterra Solutions and Droneventory and is a founding board member of the Global Supply Chain Resiliency Council, Reverse Logistics and Sustainability Council and serves on the board of directors for the Organización Mundial de Ciudades y Plataformas Logísticas.

Dale is a Leading Researcher in the fields of reverse logistics, sustainable supply chain management, supply chain finance, and secondary markets and has published in the leading journals of the supply chain and logistics fields. He has been Principal Investigator on research grants from numerous organizations. He is a Senior Editor at the *Rutgers Business Journal*, Area Editor at *Annals of Management Science*, and Associate Editor of the *Journal of Business Logistics* and the *Journal of Supply Chain Management*.

He has made more than 300 presentations to professional organizations and has been a faculty member in numerous executive education programs at universities in the United States, China, Europe, and South America as well as at major corporations and professional organizations. Dr. Rogers has been a Consultant to several companies and a principal investigator on research grants from numerous organizations and is the author of a new book on the subject of Supply Chain Financing along with Rudi Leuschner and Tom Choi.



Ron Lembke is an Associate Professor and the Chair of the Managerial Sciences department at the University of Nevada. He is the Chair of the Standards Committee of the Reverse Logistics Association where in 2016 they established a new ANSI approved standard for including multiple pieces of product information in a single 2D barcode, which has been assigned the 12 N designator, and is also known as Standardized QR for Logistics. Ron has been at the University of Nevada since 1995.

References

- Carter, C.R., Ellram, L.M., 1998. Reverse logistics: a review of the literature and framework for future investigation. *J. Bus. Log.* 19 (1), 85.
- Dowlatshahi, S., 2000. Developing a theory of reverse logistics. *Interfaces* 30 (3), 143–155.
- Rogers, D.S., 2011. Sustainability is free—the case for doing the right thing. *Suppl. Chain Manag. Rev.* 15 (6), 10–17.
- Rogers, D.S., Lembke, R., Bernardino, J., 2013a. Reverse logistics: a new core competency. *Suppl. Chain Manag. Rev.* 17 (3), 40–47.
- Rogers, D.S., Rogers, Z.S., Lembke, R., 2010. Creating value through product stewardship and take-back. *SAMPJ* 1 (2), 133–160.
- Rogers, D.S., Tibben-Lembke, R.S., 1999a. Going Backwards: Reverse Logistics Trends and Practices. Reverse Logistics Executive Council, Pittsburgh.
- Rogers, D.S., Tibben-Lembke, R.S., 1999b. Reverse Logistics: Strategies et techniques. *Logist. Manag.* 7 (2), 15–25.
- Tibben-Lembke, R.S., 2004. Strategic use of the secondary market for retail consumer goods. *CMR* 46 (2), 90–104.
- Tibben-Lembke, R.S., Rogers, D.S., 2002. Differences between forward and reverse logistics in a retail environment. *Suppl Chain Manag. Int. J.* 7 (5), 271–282.
- Wang, J.J., Chen, H., Rogers, D.S., Ellram, L.M., Grawe, S.J., 2017. A bibliometric analysis of reverse logistics research (1992-2015) and opportunities for future research. *Int. J. Phys. Distr. Logist. Manag.* 47 (8), 666–687, <https://doi.org/10.1108/IJPDLM-10-2016-0299>.

Further Reading

- Carter, C.R., Rogers, D.S., 2008. A framework of sustainable supply chain management: moving to new theory. *Inter. J. Phys. Distr. Logist. Manag.* 38 (5), 360–387.
- Rogers, D.S., 2015. Reversing your supply chain. *Insid. Suppl. Manag.* 28 (8), 24.
- Rogers, D.S., Lambert, D.M., Croxton, K., Garcia-Dastugue, S., 2002. The returns management process. *Int. J. Logist. Manag.* 13 (2), 1–18.
- Rogers, D.S., Melamed, B., Lembke, R.S., 2012. Modeling and analysis of reverse logistics. *J. Bus. Log.* 33 (2), 106–116.
- Rogers, D.S., Lembke, R.S., Bernardino, J., 2013b. Cover story: reverse logistics - a new core competency. *ASSET 2.0*, 5, 1,8,10,12. <http://www.invrecovery.org>.
- Rogers, D.S., Tibben-Lembke, R.S., 2006. Returns management and reverse logistics for competitive advantage. *CSCMP Explores* 3 (1), 1–16.

E-Tailing and Reverse Logistics

Sharon Cullinane, Professor of Sustainable Logistics, University of Gothenburg Business School, Gothenburg, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	219
Operational Issues	220
Delivery Windows	220
Delivery Options	220
E-Fulfillment	220
Cross-Border	221
Environmental Impacts	221
Reverse Logistics	221
Background	221
Reverse Logistics Costs	221
Stock Visibility and Availability	222
Environmental Issues	222
Capturing Value	222
Digitalization	223
Future Developments in E-Tailing and Reverse Logistics	223
References	223

Introduction

E-tailing is short for electronic retailing and basically refers to the sale of goods online. E-tailing has increased substantially over the past 10 years and now accounts for over 10% of sales of most categories of product (and substantially more for some categories). In 2017, Statista.com states that global e-tail sales amounted to US\$2.3 trillion, up from US\$1.3 trillion in 2014, and that 1.7 billion people worldwide shop online. When we talk about e-tailing, the discussion is often confined to the sale of consumer goods, that is, business-to-consumer (B2C) sales. It is important, however, to also include business-to-business (B2B), and consumer-to-consumer (C2C) sales in the case where the purchaser is the final consumer. In China, e-tailing is sometimes referred to as O2O (online-to-offline). The term e-tailing is sometimes used synonymously with e-commerce. Strictly, however, e-commerce is a more general term that refers to all electronic communications, so includes things like e-mail and electronic document transfers. When looking at statistics, it is important to pay attention to the precise definition used by the data source (Cullinane and Cullinane, 2018).

E-tailers can be divided broadly into those, which only sell products online, termed “pure e-tailers” and those which have both an online and a physical store presence, termed “mixed retailers.” Both “pure” and “mixed” e-tailers own the stock of products, which they sell. A third and increasingly important category of e-tailers are the so-called “market places” of which e-bay and Alibaba are examples. Such companies advertise and sell the products of other companies and generally do not own the stock that they sell. Some companies, such as Amazon and Zalando are hybrids: Amazon sells some of its own products; it sells products, which it has bought from other retailers and acts as a marketplace for other retailers. Zalando sells its own products and acts as a marketplace for other retailers. The e-tail sector is certainly dynamic and difficult to categorize as new formats and channels appear and disappear.

E-tailing is often referred to as being multi-channel or omni-channel. The difference between the two is that multi-channel refers to the fact that an item is available through more than one channel (e.g., online, in store, in a catalogue or maybe through a television channel), whereas omni-channel refers to the case where a customer uses more than one channel to purchase a product (Saghiri et al., 2018). An example of an omni-channel purchase is one where a customer has looked at a product in a shop and then purchases it online.

The logistics associated with e-tailers is sometimes referred to as e-logistics and an e-supply chain can be very complex, involving many links, for example, B2B2B2B2C. In some respects, the logistics associated with e-tailing is much the same as that required for traditional retailing; the major difference is that rather than servicing physical retail outlets (stores), the logistics operations are geared towards servicing the final, individual customer. This so-called “last-mile” delivery (i.e., delivery from a hub to the final consumer) is the problematic element of e-logistics as instead of servicing a fixed number of often quite large stores with well-organized and efficient goods-receiving procedures, companies are dealing with millions of individual people, each with their own delivery requirements and expectations and with little appreciation of the logistics involved. The last-mile delivery element costs e-tailers a disproportionate amount of money and is responsible for considerable environmental problems.

Operational Issues

Delivery Windows

The efficiency of last-mile operations is affected by the width of the “delivery windows.” A delivery window is the time-slot in which the parcel delivery company can deliver an ordered item to the customer. This is often chosen by the customer at the time of purchase and can range from a 2-hour slot (say between 10:00 and midday, or between 16:00 and 18:00) to half a day or even a full day. The customer usually wants the window to be as narrow as possible to increase the convenience to them, but for the delivery company, the wider the window the more convenient and the cheaper the delivery. Many delivery companies now notify the customer just before they arrive to maximize the convenience to the customer. Some delivery companies allow the customer to track the progress of their parcel and thus have a more precise knowledge of when their parcel will arrive.

Delivery Options

There are various delivery options used by customers. Delivery options can be divided into “attended,” where a person must be present to receive the item, and “unattended” where there is no requirement for the presence of a person. Unattended deliveries are in many ways preferable as it eliminates the “not at home” problem; this can mean increased convenience for the customer and means that deliveries can be made at a time suitable to the delivery company (including overnight for instance). Attended delivery options include delivery to the home or other location chosen by the customer (such as their place of work); delivery to a collection point of some description (such as a local shop or petrol station). Unattended delivery options include delivery to a doorstep or safe location around a person’s residence, to a locker in a locker bank (such as that used by Amazon), to a purpose-built reception box located outside a person’s residence, or even to the back of car (a solution which was trialed by Volvo). While in the United Kingdom, delivery to the home is the most popular; in other countries (such as Sweden) delivery to a collection point is more popular.

The usual means of transportation for the last-mile delivery is a van. The number of vans in circulation over the past 10 years has increased substantially as e-tailing has increased (although it cannot be stated categorically that one is a direct result of the other). There are many parcel delivery companies, all competing with each other and each with a slightly different method of operating and dealing with the customer. In addition to vans, and often in a bid to be more “environmentally friendly,” other delivery solutions have appeared, including motorcycles, pedal and electric bikes of various configurations, mini trains, taxis, and even trams. Drones are being piloted by several companies as are small ground-based autonomous pods. In more rural areas, the final mile is sometimes made by private individuals who pick up parcels from a hub and deliver them to customers in their private cars. This can save parcel delivery companies money and can also sometimes increase the proportion of successful deliveries. While all these alternative delivery methods can be useful, they also disguise the count of distance traveled in the course of parcel deliveries, once again casting doubts over the official statistics.

An increasingly popular option is “click-and-collect,” whereby the customer picks up their online purchase from a branch of the retailer or other collection point, or “reserve-and-collect,” which is the same as the previous option except the item still needs to be paid for when the customer collects the item ordered online. The attraction of these two options is that they are usually free for the customer (as opposed to home delivery which some companies charge for) and it means that the customer can collect the item at their own convenience rather than waiting around for a delivery at home. For the e-tailer it may also provide an opportunity for them to sell additional items when the customer collects their ordered items. It is a much cheaper option than delivering to the home because it eliminates the last mile element of the trip, including the redeliveries resulting from the unsuccessful “not at home” trips.

E-Fulfillment

From the e-tailer’s perspective there is a range of possibilities that can be used to fulfill customers’ orders (termed e-fulfillment). For mixed e-tailers, customer orders might be picked from the shelves of a nearby physical store. Supermarkets often fulfill at least some of their orders this way. In locations close to centers of high population, however, this method is not very efficient: it reduces the stock available for physical shoppers; the range of in-store stock might be substantially less than that available to the customer on the website and; the pickers (i.e., those employees who actually go around the shop making up the customers’ online orders) might interfere with the customers in the store. In such cases, companies might set up dedicated e-fulfillment centers (sometimes termed “dark stores”).

For products other than groceries, e-fulfillment normally takes place from a warehouse or distribution center. In the case of a mixed retailer, this might be in a dedicated e-tailing area within a warehouse containing the stock destined for the physical stores, or in some cases, it might be a dedicated e-tail warehouse. Where a company needs to dispatch a large volume of goods very quickly, the e-tailer might cut out the warehouse altogether and engage in what is termed “drop-shipping.” In this case, the e-tailer never touches the products, but the manufacturer or wholesaler delivers the goods direct to the customer. This is very prevalent in B2B e-tailing but is increasingly important in B2C e-tailing, particularly in the case of products which are required by the customer very quickly, where the manufacturer has logistics expertise which a parcel company does not have or where the goods are heavy and using a drop-shipper can reduce double handling.

As e-tailing has increased, customers’ expectations of delivery times have changed. There is an increasing desire for almost immediate delivery of products ordered online. In a bid to be competitive, companies are offering faster delivery times. Next-day delivery is an option in many cases and an increasing number of companies are even offering same-day delivery. In some cases,

customers are expected to pay for these super-fast delivery options, but not in all cases. Amazon Prime is an example where for a relatively small annual fee, customers can request next-day delivery as a “free” option for many products. In larger countries, such as Canada, Amazon offers two-day delivery. It can generally be said that the faster the delivery request, the greater the demands on the logistics operation. Fast delivery means that load consolidation is less possible, so vehicles are more likely to be running less-than-full. This increases the logistics costs, as well as the costs to the environment.

Cross-Border

The online market is truly global and one of the major developments over the past decade has been the increase in cross-border online trade. For many countries, cross-border trade accounts for more than 50% of online trade, with China and the United States being two of the main beneficiaries of this trade. This obviously impacts on e-tail logistics, making the logistics chains much longer and complex and increasing the propensity for goods to be transported at least in part by air.

Environmental Impacts

There is some debate over the environmental impacts of e-tailing. Some research has suggested that buying things online is less environmentally damaging than buying them in-store, whereas other research has shown the opposite. In order to properly calculate the environmental impacts of e-tailing, both individual travel behavior and freight transport must be taken into account. Some studies do not take into account the former. A problem arises in where to draw the boundaries of the calculations. For example, if a person does not drive to the shop to buy a product, then it might seem that this saves a car trip. But what happens to the car when it is not being used for the shopping trip? Will it be used by another member of the family for some other purpose or perhaps it will be used for another purpose by the person who would have made the shopping trip. And does the fact that shopping trips are not so frequent mean that there will be an overall impact on car ownership? On the freight side, the extent of environmental damage will depend on the drop-density of the load (i.e., the number of parcel deliveries per unit of distance traveled by the van). The higher the drop-density the lower the environmental impact per parcel. It will also depend on the traffic conditions in which the van is driving, the type of fuel used, the level of urbanization, the driving behavior of the van driver and many more things. How often when you order a set of 4 printer cartridges, for example, do they arrive in 4 separate packages from 4 different companies maybe at 4 separate times and with 4 lots of packaging? Several studies have shown that the level of packaging associated with e-tailing is much greater than that associated with shopping in store. It has also been shown, however, that the level of waste is greater for physical shops than it is for online shops, as physical shops often over-stock on some products which then need to be disposed of in some way. Finally, there is the issue of reverse logistics and the negative environmental impacts of this activity in e-tailing.

Reverse Logistics

Background

Reverse logistics refers to the flow of goods back from the consumer to the retailer or the manufacturer. It involves most of the logistics operations covered above, but in reverse. Consumers have always returned products back to stores but, in the case of e-tailing, the number and proportion of returns is much greater. There are many reasons why a customer might return a product: it might be damaged or faulty or the wrong product might have been delivered; it might not match the expectations the customer had from looking at it on the website; the customer might have ordered without much thought and subsequently changed their mind. However, many of the returns occur because of consumer behavior, which is almost specific to the process of buying online. Over-ordering is a key issue in this context. The consumer cannot try things on or see the product that they are buying, so they over-order and then routinely send back what they do not want. For example, a person may order a shirt in 4 different colors and 3 different sizes, try them on at home, perhaps keep one or two and send back the rest. Their home becomes the changing room. In addition to this there are levels of “wardrobing,” that is, purchasing a product, using it once or twice and then returning it and “renting” where a customer uses a product for a specific purpose (e.g., a camera for a birthday party) and then sends it back. There is also a level of totally fraudulent returns which might include purchasing a high value item and sending back a fake; falsely claiming that something is faulty in the hope of getting a refund and customers sending back random items hoping that a refund for their original purchased item will be made. This all adds up to the facts that return rates are often very high. In the clothing sector, which has the highest level of online purchasing, the returns levels are also the highest. Typically, the returns rate for clothing is around 22%, but it can increase to around 60% for high-fashion goods and even 90% for “occasion wear” such as evening wear. Consumer behavior is reinforced by the policy of e-tailers to offer free returns.

Reverse Logistics Costs

The logistics costs to e-tailers of dealing with their returns is very high, in fact it is around 3 times that of the outward logistics costs. Each return must be collected, transported, inspected, sorted, and repackaged before it can be put back into outbound stock. An item of returned clothing will need to be visually inspected and sometimes sniffed; it may need to be cleaned or hairs and fluff might need

removing. It might need re-ironing and then repackaging. Manpower and space is required for all these processes. Generally, it takes three times more manpower to deal with a returned item than to deal with the initial dispatch of that item.

Outward logistics is often extremely well-planned and efficient and often involves the delivery of large quantities of goods to a relatively small number of often well-organized companies. With reverse logistics however, the returns are being made from a very dispersed, large number of individuals who have little knowledge and often little interest in the logistics requirements of getting their product back to the retailer. These individuals may keep their item for 1 day before they return it or for 63 or 364 days. Obviously, the longer they keep the product, the less likely it is that the e-tailer can re-sell it, particularly at full price ([Ferne and Sparks, 2018](#)).

The e-tail market is very competitive and keeping costs as low as possible is crucial for e-tailers. High wages are a major cost factor in dealing with returns, so some e-tailers actually process their returns in low-wage countries. In Europe, for example, some clothing e-tailers process their returns in Estonia or Poland and then transport them back to warehouses in western Europe to re-dispatch the goods to new customers. This can result in returns being transported for many hundreds of kilometers. In environmental terms, this is not very sustainable. E-tailers are sometimes loathe to deter customers from making returns, as there has been research to show that those customers who return the most also spend the most; so by deterring returns, profitability might be adversely affected.

Stock Visibility and Availability

Two key issues in reverse logistics are stock visibility and stock availability. Stock visibility refers to the e-tailer's ability to see the stock electronically wherever it is in the system, whether this be in a warehouse, en-route to a customer, or being returned by a customer. With so much trade being international by nature, it is difficult to maintain stock visibility in the system as a whole. Stock availability is linked to its visibility and refers to the ability of the e-tailer to know how much stock is available in the system as a whole—that is, having a picture of their *virtual* stock. Thus, if a customer A in location x wants to order a particular item, the e-tailer needs to know that even if it is not in stock at their warehouse, there is a customer B in location y who is in the process of returning that particular item and that it could be available for customer A in say, 2 days. Unfortunately, the ICT (information and communications technology) required to have full stock visibility and availability is very complex and expensive and so very few e-tailers have it.

E-tailers should take returns into account when ordering stock. For instance, if they know that 30% of their products sold are likely to be returned, they should order 25%–30% fewer products to start with. That is, returns should be taken into account in their inventory management systems. However, this does not always occur. As stated above, the timing and condition of returns is often unknown to the e-tailer, so there is little stock visibility—making it difficult to operate accurate inventory management systems which should take into account the overall availability of a particular SKU (stock keeping unit).

Environmental Issues

The increase in cross-border trade has also impacted on reverse logistics trends. Although not always the case, returns are frequently sent back to the country from which the products originated; so if a British customer receives a product from Germany, for example, the product may also be returned to Germany. This is not always the case, as some e-tailers have taken the step of opening a “returns” center in some countries, particularly those in which the market for their products is quite large.

For mixed retailers, that is, those with both an online and a physical store presence, the potential exists for customers to return unwanted goods to the physical store. Once an item is returned to the store, it usually stays at the store for resale, thereby reducing the whole reverse logistics process. Many companies, even the largest ones, however, have not yet developed their computer software sufficiently to be able to cope with this and to enable them to deal with customer refunds for goods bought through a different channel. Currently, the omni-channel possibilities have not yet been extended to returns for many companies. This is an area where a great deal more work needs to be done.

Because of the scale of the returns problem, specialist reverse logistics companies have started to appear. Such companies, such as Clipper and ReBound logistics, not only deal with the processing of returns for other companies, but also specialize in providing intelligent digital algorithms to reduce a company's returns in the first place (for a review of this issue, see [Mangiaracina et al., 2015](#)).

Capturing Value

When products are returned by customers, companies will seek to maximize the value they can receive from these goods. Where possible, goods will be re-sold as new with little or no loss of value. This will not be possible for all items however. Some will be damaged; some will be counted as “old stock” and some will be past their sell-by date. There is a so-called hierarchy of reverse logistics. The precise hierarchy varies slightly between actors but, in general, from best option to worst option, it will include the following: Resell, repair, recondition, refurbish, remanufacture, cannibalize, recycle, and dispose. This hierarchy nowadays might also contain words such as “re-purpose” or “up-cycle.” The aim is to contribute to the circular economy with minimal waste and maximum profitability. Some items might be sold in end-of-season sales, in outlet centers, discount stores or secondary shops (either online, offline, or both) and some will be donated to charity. There have been horror stories about some e-tailers burning returned items, particularly in the case of “luxury brands” which do not wish to reduce a product's exclusivity by selling it in a secondary market, but this action adds little to profitability, is unpopular with the general public and is unlikely to persist in the long run.

The reverse logistics of certain products can be covered by legislation. In Europe, for instance, the Waste Electrical and Electronic Equipment Directive (the WEEE Directive, or EU Directive 2012/19/EU) was introduced in 2003 and has had several subsequent amendments, covers all electrical and electronic goods such as fridges, computers, toasters, lighting, cookers, etc. The Directive covers the collection, treatment, and recycling of all these goods and compels producers who sell their products in the EU to deal with them according to legislation. The aim is to reduce the quantity of such waste, which is sent to landfill in order to reduce the environmental and health effects of some of their hazardous content.

Reverse logistics also includes dealing with packaging and other items necessary for the safe and secure delivery of items ordered online, including logistics handling units such as pallets, trolleys, and crates.

Digitalization

Key to efficient e-logistics is digitalization. In reverse logistics, digitalization can help in reducing the number of returns that are made to start with and can aid the efficiency of the returns process. Included in the former category are developments such as digital changing rooms and AI (Artificial Intelligence) which can improve the alignment between expectations and reality; social media which can alert e-tailers to problems with the products through instant feedback pages and can help in the customization and tailoring of products to specific customers or groups of customers, as well as providing peer review opportunities. The quicker an e-tailer knows that a product is being returned, the greater the value that can be captured from the return. Digital tools such as EDI (electronic data interchange) that speed up communications between different parts of the logistics process and RFID (radio frequency identification) which enables product to be tracked and traced through the supply chain, can aid in increasing the visibility of a product to an e-tailer. Problems with operationalization of digital systems and platforms however are many. The issue of integration of different digital systems both within the company and between the company and its logistics partners is a difficult one. Most companies cannot afford to have new, bespoke systems and so have to incorporate additional elements or modules into their existing systems. This action is invariably second best and causes considerable integration problems. Another problematic area is making a choice decision between the huge arrays of digital systems available on the market. Many companies can identify multiple data and information requirements needed to improve the efficiency of their logistics processes. Prioritizing these requirements is difficult and trading off benefits against costs poses additional problems. Introducing new forms of digitalization can cost millions of dollars and this money can be completely wasted, if the system is not appropriate for the company (for a review, see Pettit and Wang, 2016).

Future Developments in E-Tailing and Reverse Logistics

Advances in this field are rapid and diverse. What seems futuristic one day can become quite normal within a few months. There is a great deal of competition in e-tailing and this is driving future developments, as companies strive to gain a slice of the e-tail pie. Such developments include the relatively low-key possibilities that might come with unattended collection and delivery points such as lockers, through innovative post boxes that allow companies to access them to deliver items and pick up returns. In the middle of the spectrum are drones and delivery pods. At the other end of the spectrum, developments in Artificial Intelligence, block-chaining and digitalization generally will improve the efficiency of e-logistics in ways that are currently unimaginable. It could, however, be changes in customer behavior or in fields seemingly unlinked to logistics (such as has happened with the development of social media) that dictates future developments in e-tail logistics.

References

- Cullinane, S.L., Cullinane, K.P.B., 2018. The reverse logistics of cross-border e-tailing in Europe: developing a research agenda to assess the environmental impacts. *Int. J. Appl. Logist.* 8 (1), 1–19.
- Fernie, J., Sparks, L. (Eds.), 2018. *Logistics & Retail Management: Emerging Issues and New Challenges in the Retail Supply Chain*. fifth ed. Kogan Page, London.
- Mangiaracina, R., Marchet, G., Perotti, S., Tumino, A., 2015. A review of the environmental implications of B2C e-commerce: a logistics perspective. *Int. J. Phys. Distrib. Logist. Manage.* 45 (6), 565–591.
- Pettit, S., Wang, Y., 2016. *E-logistics: Managing Your Digital Supply Chains for Competitive Advantage*. Kogan Page, London.
- Saghiri, S., Bernon, M., Bourlakis, M., Wilding, R. (Eds.), 2018. Omni-channel logistics Special Issue. *Int. J. Phys. Distrib. Logist. Manage.* 48, 4.

Green Routing of Freight Vehicles

Tolga Bektaş, University of Liverpool Management School, Liverpool, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Green Routing	224
Emission Models	225
Activity-Based Models	225
Macroscopic Models	225
Microscopic Models	226
Planning for Green Routes	226
Generic Formulations	227
Route Optimization	227
Route and Speed Optimization	228
Speed Optimization	228
Optimal Control	229
Outlook	229
See Also	229
References	229

Green Routing

Freight distribution entails the movement of goods from their origin to destination for which freight vehicles provide the basic means of transportation. While the term vehicle may have a more general connotation to include those of different modes, for example, a barge in water transport or rail cars in railways, the term is especially understood to be relevant to land transportation, and the vehicles that are used therein, such as a truck, a lorry, or a van.

Distribution of freight has undesirable, yet inevitable, impacts on the environment, also known as externalities, most of which arise in the process of physically moving the goods (Forkenbrock, 1999). Vehicles emit greenhouse gases, such as carbon dioxide (CO₂) or nitrogen oxide, and pollute the air through the release of others, such as methane, and small particles of dust. Emissions and air pollution result in phenomena such as acidification, summer smog, and depletion of the ozone layer. Other impacts include noise and vibration that cause damage to buildings and are detrimental to human well-being, use of land and natural resources for building and maintaining the necessary infrastructure, and accidents that cause injuries and death.

Freight can either be transported in full truckloads when goods are to be sent directly from their origin to destination, or in less-than-truckload shipments. The latter is relevant when goods are to be delivered to several destinations by a single vehicle, and are consolidated at the point of origin and loaded onto a vehicle prior to it being dispatched. In this case, the vehicle departs from the point of origin, often called the depot, visits the destinations (also called customers) in successive order, and returns to the depot. The sequence of visits is called a route, which may also entail the pickup of goods, in addition to delivery. Planning the routes of such vehicles is the aim of the class of problems known as vehicle routing, the basic version of which entails finding the best sequence of customers for a vehicle to optimize a given objective. The problem arises at the operational level of decision-making, typically executed on a daily basis. The vehicle routing problems that arise in practice, however, are often compounded with additional complexities, including the planning of routes for a fleet of vehicles, either all identical or with varying features, and other practical restrictions such as the limits on the amount of commodity transported due to vehicle capacity, or the need to deliver the goods within time slots (or time windows) dictated by the customers. A solution to the basic vehicle routing problem is a sequence of customers to be visited, one for each vehicle, where each customer is visited only once. The objective itself is often the minimization of economic cost, which may be measured in terms of the total distance traversed, the total time spent, or a similar measure of performance.

A green routing of freight vehicles is of concern when the problem entails an explicit consideration of environmental impacts of the distribution itself, with an aim to reduce the externalities as much as possible (Piecyk et al., 2015). Incorporating a particular externality into the planning of vehicle routes requires, however, an explicit representation of the said externality in some quantified form, such as a mathematical function of the various parameters that govern the externality. This is necessary to be able to measure the environmental impact of a given routing plan. Not all externalities can easily be represented using a functional form, however. Accidents or vibration are difficult to quantify, for example. Others are more amenable to such a representation because they are better understood. Emissions is a primary example of such an externality, particularly that of CO₂, which is known to be proportional to the amount of fuel consumed by a vehicle. Although the amount of carbon content in the fuel varies, the average value used by the European Environment Agency, for example, is 3.169 kg of CO₂ per kilogram of fuel (Ntziachristos and Samaras, 2018). The amount of fuel consumed, in turn, is a function of the energy required to move the vehicle, which depends on factors such as vehicle

load, speed, acceleration, and engine efficiency, and can therefore be better approximated. This is one reason why green routing of freight vehicles is predominantly concerned with reducing carbon emissions, which will also be the focus of this paper.

Emission Models

Emission models quantify the amount of greenhouse gases released by freight vehicles, and can either be in the form of a simple estimation factor expressed per unit of transport activity, or represented using a more detailed algebraic expression that includes various vehicle- and environment-related parameters. From a planning perspective, these factors can be classified into three, namely, activity-based, macroscopic, and microscopic, as detailed later (Demir et al., 2014; Bektaş et al., 2019).

Activity-Based Models

Also known as emission factors, these models calculate the emissions from a transport activity, and come in two types. The first type is used when the total amount of energy used or the fuel consumed by the transport activity is known, in which case it is simply multiplied by an emissions factor (e.g., in grams per kWh). If the total amount of energy used is not available, then factor models provide pre-calculated rate emissions from a transport activity based on the type (including capacity) of vehicle used, typically expressed in amount of carbon emitted per vehicle kilometer. The typical emission factors reported by the European Environment Agency, for example, range from 25.5 g of CO₂ per kilometer for a light commercial vehicle using conventional engine technology to 1.3 g of CO₂ per kilometer for the same vehicle using a Euro VI standard technology (Ntziachristos and Samaras, 2018).

Macroscopic Models

Macroscopic models are used for wide-area emissions assessment, where emission rates are measured for a given trip characterized by the type of vehicle used and the average speed v at which the vehicle travels. One of such models is described in the Air Pollutant Emission Inventory Guidebook by the European Commission (Ntziachristos and Samaras, 2018) that is an EU standard used to calculate vehicle emissions. The model is of the form $E(v) = b(a_1v^2 + a_2v + a_3 + a_4/v)/(a_5v^2 + a_6v + a_7)$, where $a_1 - a_7$ are parameters that are predefined on the basis of the fuel, class of vehicle, and engine technology used, and b is a correction factor applied, if necessary, to account for different types of road (i.e., urban, rural, and highway). The function $E(v)$ computes the hot emissions expressed in terms of grams per kilometer for a given average speed v . Figs. 1 and 2 show the resulting emissions produced using the data described by Ntziachristos and Samaras (2018) for light and heavy goods vehicles, respectively, that use different engine technologies.

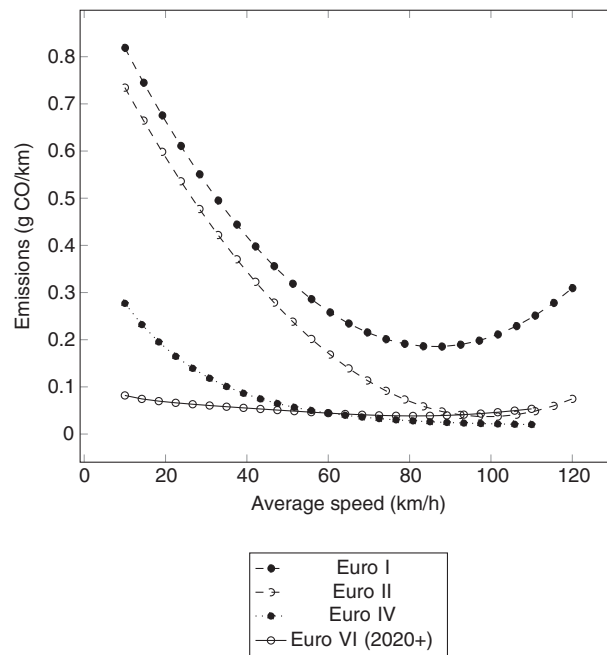


Figure 1 CO₂ emissions calculated with the average speed model using the data described by Ntziachristos and Samaras (2018) for diesel light goods vehicles (maximum weight not exceeding 3.5 tons) and for different engine technologies.

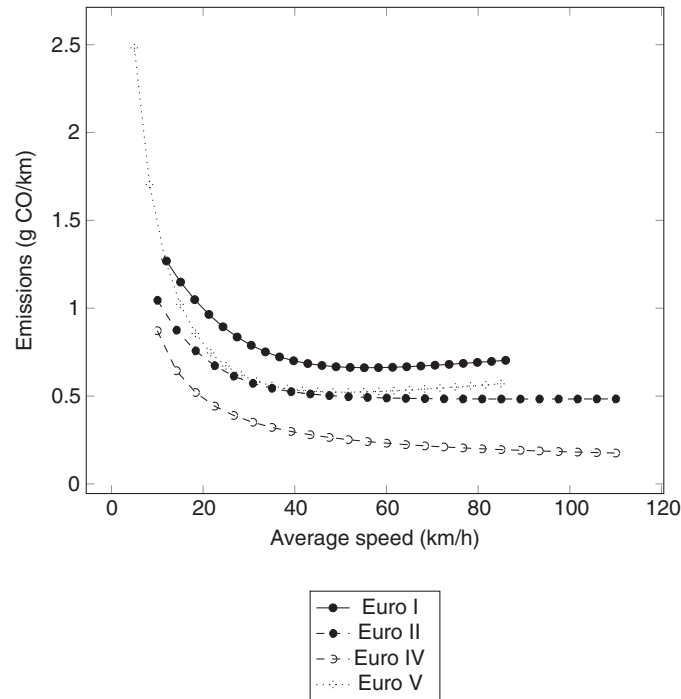


Figure 2 CO₂ emissions calculated with the average speed model using the data described by Ntziachristos and Samaras (2018) for diesel heavy goods vehicles (maximum weight between 3.5 and 7.5 tons) and for different engine technologies.

Microscopic Models

Microscopic models provide for a much finer way to estimate emissions from a vehicle in an instantaneous manner (e.g., on a second-by-second basis). These models involve highly detailed expressions derived from mechanical physics of automobile engines (Ross, 1997) to calculate the tractive power required by the engine to move the vehicle over a given length of road. They also incorporate other demands for power including accessories (such as air conditioning). The tractive power itself is a function of a number of parameters that can broadly be classified into two, the first of which includes those that relate to the vehicle, such as vehicle mass (including empty weight and any load carried), instantaneous speed and acceleration, and the frontal surface area. The second class includes environment-related parameters such as road gradient, rolling resistance, drag, gravitational force, and air density. Assuming that all parameters, except for vehicle mass w and speed v , remain constant, a simplified form of a microscopic model to calculate the energy requirement $F(w, v)$ (in kWh) for the engine to traverse a road segment of length d can be expressed as follows:

$$F(w, v) = g(w \cdot d) + h(v^2 \cdot d), \quad (1)$$

where g and h are constants that can be calculated using the vehicle- and environment-related parameters. In calculating the total energy requirement, one would also have to factor in engine efficiency, which can then be converted into total fuel consumption by taking engine-specific parameters into account (Barth et al., 2005).

Planning for Green Routes

Vehicle routing problems are generally represented by optimization models, where the latter are solved using mathematical programming techniques. Green vehicle routing entails incorporating one or more externalities into the planning, which can be done by embedding the functions that represent the externalities into the corresponding optimization model. In the case of emissions, this can be done by adopting a particular emissions model and treating it as a minimizing objective function, or as a component thereof (Figliozzi, 2010; Qian and Eglese, 2016). The way in which the emissions model is incorporated into the optimization depends on the variables that can be controlled along a route, which in turn gives rise to several types of green vehicle routing problems, some of which are described as follows:

1. Fleet composition entails choosing, from a given set of vehicles with varying characteristics, such as capacity and weight, those to be dispatched, and the route that each vehicle is to follow. Fleet composition is particularly relevant to green vehicle routing as

the choice of vehicles will impact on emissions, which is due to the way in which vehicle-specific parameters affect fuel consumption. The type of vehicle used and the route chosen are interdependent decisions that should be made jointly (Koç et al., 2014).

2. Route optimization forms the main set of decisions to be made in the standard vehicle routing problem, which is also the case in green vehicle routing. One difference between the two problems is the need to factor in environment-related parameters into the planning models which are known to affect fuel consumption and emissions, such as road gradient and drag. Another difference arises due to the change of vehicle load from one delivery location to the next on a given route. As load affects fuel consumption, it would need to be considered jointly with the distance traversed. This combined effect in green routing is sometimes modeled by a function where the distance between successive delivery locations is weighted by the load carried over that distance.
3. If the traffic conditions allow for the (average) vehicle speed to be set freely, within legal limits, and implemented accordingly, then speed selection arises as another set of decisions that can be made to reduce emissions. If there are no time window restrictions present in the problem, then the choice of speeds is trivial. In this case, slowing vehicles down to their optimal speed at which fuel consumption is minimized can reduce emissions (McKinnon, 2016). Where there are time window restrictions, however, then speed choice becomes a nontrivial problem, especially given that speeds are inherently linked to route feasibility with respect to time. In this case, problems of route and speed selection need to be solved jointly (Bektaş and Laporte, 2011). Given a fixed route, speed selection gives rise to what is known as the speed optimization problem, which entails computing the average speed to be used when traveling between successive customers in order to minimize the resulting emissions (Kramer et al., 2015).
4. Acceleration is one micro-level factor that affects fuel consumption and emissions, and is inherently linked to speed choice on a given road segment. Choosing the speed and acceleration profile of a vehicle on a given road gives rise to an optimal control problem to maximize fuel economy (Hooker et al., 1983).

Generic Formulations

This section presents formulations to help conceptualize some of the problems described in the previous section. For simplicity, the models will assume the use of a homogeneous fleet of vehicles. Consequently, decisions concerning the selection of the number and type of vehicles will not be part of the decision problems described later, but models involving fleet composition decisions can be written in a similar fashion.

Route Optimization

The standard vehicle routing problem can be modeled on a graph described by a node set C of customers, with node 0 as the depot, and a set of arcs in between (possibly, but not necessarily, all) pairs of customers. The set R includes all routes that are feasible with respect to the problem restrictions, such as those on vehicle capacity and time. A constant cost c_r is defined for each route $r \in R$, which may be the total distance on the route, the total time to traverse the route, or a linear function of either.

The problem can be formulated using a binary variable x_r that is equal to 1 if route $r \in R$ is chosen, and is equal to 0 otherwise, and a binary parameter ξ_{ir} that equals 1 if the customer $i \in C$ is included in the route r , and 0 otherwise. Using this notation, one possible formulation for the standard vehicle routing problem can be described as follows.

$$\text{Minimize } \sum_{r \in R} c_r x_r \quad (2)$$

subject to

$$\sum_{r \in R} \xi_{ir} x_r = 1 \quad i \in C \quad (3)$$

$$x_r \in \{0, 1\} \quad r \in R, \quad (4)$$

which will select a number of routes from within the set R so as to minimize the total overall cost, and in such a way that each customer is included in only one route as dictated by constraints (3).

If one is to use activity-based models, then the cost c_r of a route $r \in R$ can simply be changed to reflect the unit energy or fuel consumption, or emissions, such as $\hat{c}_r = u \cdot d_r$, where u is the rate of emissions per vehicle per distance traveled (e.g., kilometers), and where d_r is the length of route r (e.g., in kilometers). However, the use of such a model assumes that parameters that affect fuel consumption or emissions along a given route, such as vehicle load or speed, remain constant, which may not necessarily hold in practice. This is particularly the case when a vehicle makes several deliveries on a given route, which means that the load carried on the vehicle will change from one stop to another.

Route and Speed Optimization

If the intention is to plan for vehicle routes at a more detailed level by controlling variables such as load and speed, then the use of microscopic models that incorporate such variables as input in calculating emissions may allow for a finer level of detail in the modeling. In vehicle routing, the load of the vehicle changes from one customer to another as deliveries are made along a route. Vehicle speed, on the other hand, is continuous and can change instantaneously along an arc. For planning purposes, however, speed could be expressed as the set $S = \{s_1, s_2, \dots\}$ of discretized average speeds. For a given route characterized by the sequence of customers visited, let $V_r \subseteq S$ be the set of speed vectors for a given route $r \in R$, where each vector $v_r \in V_r$ represents the feasible speeds that can be assigned to each pair of successive customers on the route. Under time window constraints, not all discretized speeds may be achievable. In this case, the planning involves solving a route and speed optimization problem, which concerns jointly determining the routes and the speeds on each leg of each route. One formulation for this problem, shown later, uses a binary variable x_{rv_r} equal to 1 if a route $r \in R$ is selected that uses the speed vector v_r , and equal to 0 otherwise.

$$\text{Minimize} \quad \sum_{r \in R} \sum_{v_r \in V_r} \bar{F}_r(w, v_r) x_{rv_r} \quad (5)$$

subject to

$$\sum_{r \in R} \sum_{v_r \in V_r} \xi_{ir} x_{rv_r} = 1 \quad i \in C \quad (6)$$

$$\sum_{v_r \in V_r} x_{rv_r} = 1 \quad r \in R \quad (7)$$

$$x_{rv_r} \in \{0, 1\} \quad r \in R, v_r \in V_r, \quad (8)$$

where constraints (6) choose one route for each customer and constraints (7) ensure the selection of one speed vector for each route. In the objective function, $\bar{F}_r(w, v) = \delta \cdot F_r(w, v)$ the total amount of emissions on route r , derived using (1) and δ is a conversion factor from energy to emissions.

In most instances, fuel consumption and total time spent en route are conflicting objectives, as the optimal speed that minimizes fuel consumption or emissions is not always the fastest possible speed (Figs. 1 and 2). The latter can be expressed in terms of value of time or money paid to the drivers. If τ_r denotes the time spent on each route, minimizing both objectives would give rise to the following bi-objective formulation:

$$\text{Minimize} \quad \sum_{r \in R} \sum_{v_r \in V_r} F_r(w, v_r) x_{rv_r} \quad (9)$$

$$\text{Minimize} \quad \sum_{r \in R} \sum_{v_r \in V_r} \tau_r x_{rv_r} \quad (10)$$

subject to (6)–(8). It is possible that the two objectives can be converted into a single monetary objective by using unit fuel cost (e.g., cost per liter) for the former and driver wages (e.g., cost per hour) for the latter.

Speed Optimization

Speed optimization is a scheduling problem that determines the arrival times to each of the k customers visited in a fixed sequence $(0, 1, \dots, k)$ by adjusting the speed $v_{i,i+1}$ in between every successive pair $(i, i+1)$ of locations, where $0 \leq i \leq k-1$, in order to optimize a given function. The problem is nontrivial when each customer $i = 1, \dots, k$ requires a visit within a prescribed time interval denoted by $[a_i, b_i]$. In addition, speeds must be chosen within an interval $[v_l, v_u]$, dictated by legal restrictions and traffic conditions. Within green routing, the objective would be to minimize a fuel consumption function such as (1), specified for each leg $(i, i+1)$ of the route as $F_{i,i+1}(w, v)$. A model for the speed optimization problem is given as follows.

$$\text{Minimize} \quad \sum_{i=0}^{k-1} F_{i,i+1}(w, v) \quad (11)$$

subject to

$$t_{i+1} - t_i - d_{i,i+1}/v_{i,i+1} \geq 0 \quad i = 0, \dots, k-1 \quad (12)$$

$$a_i \leq t_i \leq b_i \quad i = 1, \dots, k \quad (13)$$

$$v_l \leq v_{i,i+1} \leq v_u \quad i = 1, \dots, k \quad (14)$$

The variables t_i that appear in the formulation indicate the service start time for each customer $i = 1, \dots, k$, calculated using constraints (12), which are constrained to be within the required service by (13). Constraints (14) simply ensure that the chosen speeds are within the allowable limits.

Optimal Control

This planning problem pertains to finding an optimal driving profile for a vehicle characterized by the speed $v(t)$ and acceleration $a(t)$ of the vehicle, expressed as a function of time t , as it traverses a road segment of a given length d over a duration of T time units. The vehicle is assumed to have an initial speed equal to v_0 and the road has a gradient equal to $\theta(l(t))$. The length $l(t)$ is the distance traversed by the vehicle at time t from the starting point. A formulation for the optimal control problem is given as follows.

$$\text{Minimize}_{v,T} \int_0^T f(v(t), a(t)) dt \quad (15)$$

subject to:

$$\dot{l}(t) = v(t) \quad (16)$$

$$a(t) = \dot{v}(t) + g \sin \theta(l(t)) \quad (17)$$

$$a(t) \leq \bar{a}(v(t)) \quad (18)$$

$$v(0) = v_0 \quad (19)$$

$$s(0) = 0 \quad (20)$$

$$s(T) = d. \quad (21)$$

The expression $f(v(t), a(t))$ that appears in (15) is the rate of fuel consumption that is a function of both speed and acceleration. The acceleration itself is subject to a maximum value of $\bar{a}(v(t))$.

Outlook

Various algorithmic techniques have been proposed to solve the types of mathematical models of green routing problems described earlier, ranging from exact techniques to heuristics. Applications have been reported on practically relevant problems that differ with respect to the model that is used to calculate emissions, the variables used in the planning (e.g., load and speed), practical restrictions, type of environment (long vs. short-haul, urban vs. intercity), and so on. While the reported results are case-specific, they collectively indicate that savings in emissions are possible by incorporating a finer level of detail in the route planning, which can be significant in some cases. In some cases, the savings in emissions can be reaped without a significant compromise in operational cost, suggesting that win-win solutions may well be possible in green vehicle routing in which environmental and economic criteria are simultaneously optimized.

See Also

The Concept of External Cost: Marginal Versus Total Cost and Internalization; Valuation of Carbon Emissions; Freight Vehicle Routing & Scheduling; Road Freight Vehicles; Decarbonizing Road Freight Transport; Energy Consumption of Transport Modes

References

- Barth, M., Younglove, T., Scora, G., 2005. Development of a heavy-duty diesel modal emissions and fuel consumption model. Technical Report UCB-ITS-PRR-2005-1. Institute of Transportation Studies, University of California at Berkeley.
- Bektaş, T., Ehmke, J.F., Psaraffis, H.N., Puchinger, J., 2019. The role of operational research in green freight transportation. *Eur. J. Oper. Res.* 274 (3), 807–823.
- Bektaş, T., Laporte, G., 2011. The pollution-routing problem. *Transp. Res. Part B Methodol.* 45 (8), 1232–1250.
- Demir, E., Bektaş, T., Laporte, G., 2014. A review of recent research on green road freight transportation. *Eur. J. Oper. Res.* 237 (3), 775–793.
- Figliozzi, M., 2010. Vehicle routing problem for emissions minimization. *Transp. Res. Rec. J. Transp. Res. Board* 2197, 1–7.
- Forkenbrock, D., 1999. External costs of intercity truck freight transportation. *Transp. Res. Part A Policy Pract.* 33 (7), 505–526.
- Hooker, J., Rose, A., Roberts, G., 1983. Optimal control of automobiles for fuel economy. *Transp. Sci.* 17 (2), 146–167.
- Koç, Ç., Bektaş, T., Jabali, O., Laporte, G., 2014. The fleet size and mix pollution-routing problem. *Transp. Res. Part B Methodol.* 70, 239–254.
- Kramer, R., Maculan, N., Subramanian, A., Vidal, T., 2015. A speed and departure time optimization algorithm for the pollution-routing problem. *Eur. J. Oper. Res.* 247, 782–787.

- McKinnon, A.C., 2016. Freight transport deceleration: its possible contribution to the decarbonisation of logistics. *Transp. Rev.* 36 (4), 418–436.
- Ntziachristos, L., Samaras, Z., 2018. EMEP/EEA air pollutant emission inventory guidebook 2016—Update Jul. 2018: Section 1.A.3.b.i-iv Road transport 2018. Publications Office of the European Union, Luxembourg.
- Piecyk, M., Browne, M., Whiteing, A., McKinnon, A. (Eds.), 2015. *Green Logistics: Improving the Environmental Sustainability of Logistics*. Kogan Page Publishers, London.
- Qian, J., Eglese, R., 2016. Fuel emissions optimization in vehicle routing problems with time-varying speeds. *Eur. J. Oper. Res.* 248 (3), 840–848.
- Ross, M., 1997. Fuel efficiency and the physics of automobiles. *Contemp. Phys.* 38 (6), 381–394.

Freight Mode Choice

Hyun Chan Kim*, Alan Nicholson*, *Waikato Institute of Technology, Hamilton, Waikato, New Zealand; †University of Canterbury, Christchurch, Canterbury, New Zealand

© 2021 Elsevier Ltd. All rights reserved.

Introduction	231
Freight Mode Choice Decision Factors	231
Freight Transport Decision-Makers	232
Road Transport: Owned-Fleet or For-Hire Vehicles	232
Modeling Freight Demand and Mode Choice	233
Behavioral Demand Model	234
Inventory-Based Model	234
Revealed Preference and Stated Preference Methods	234
References	235
Further Reading	235

Introduction

In a firm's logistics activity, the mode choice decision process would be considered only at one stage by shippers, who must make choices on "How" to move the consignments taking into account destination and shipment sizes. However, it is at this stage the interests of three major institutions associated with freight transport (i.e., the government, the carrier, and the shipper) mainly coincide.

Due to growing concerns about urban congestion, environmental impact, public health, and safety considerations, freight transportation has increasingly become of vital importance in logistics decision making in particular, and in the industrial process in general. In order to manage the issues, firms consolidate their freight flows and outsource the transport and logistics activities. Among the shippers and logistics providers, the effective use of varying transport modes is not yet widely accepted as an alternative when addressing transport problems. Although the choice of an appropriate transportation mode is vital for the logistics process in a global industrial process, shippers, and firms feel constrained by the logistics trade-off, such as the trade-off between the level of inventory versus the level of customer service.

Market globalization and the development of service economies have increased the demand for reliable, flexible, timely, and cost-effective freight services to shippers. The Organisation for Economic Cooperation and Development (OECD) predicts that global freight transport demand (in ton-km) is expected to grow by about 294% between 2015 and 2050 (OECD/ITF, 2017). The OECD expects the steady growth of freight movements to continue and increase by almost 100% by 2040. Concurrently, the modal share of road transport has increased significantly and is expected to increase further in the coming years. In addition, with rising fuel prices and growing awareness about the challenge of global climate change, innovative policies and technologies to reduce the negative impacts of the dependency on road transport are being introduced. Despite the dynamic nature of the business environment, few studies have attempted to systematically establish a relationship between the shipper's mode choice perception and their logistics characteristics in the freight transportation market.

Freight Mode Choice Decision Factors

The allocation of freight among transport modes, often known as mode choice, is one of the most controversial topics in the field of transport logistics. This is because many mode choice decisions are not always based upon a full and rational appraisal of options available, and commercial decisions generally do not consider the total cost of each mode or modal service, especially external costs related to safety and environmental impacts.

The choice of transport mode has a direct impact on the efficiency of logistics channels and systems. Each transport mode possesses different characteristics and different strengths and weaknesses. Depending on the mode chosen, the overall performance of the logistics system will be affected. Transport decision-makers choose the transport mode within a logistics system and depending on the decision maker's requirements, unimodal, multimodal, or integrated transport logistics will be utilized. It is important to recognize the impact of the decision maker's perceptions on the selection decision.

The perceptual approach assumes that the explanatory variables influencing transport mode choice are determined by the transport user's perception of subjective quality measurements, rather than by objective measurements. This approach treats transport as a product purchased like any other product. A range of empirical studies performed from the 1970s to the 2000s shows that freight shippers have varying perceptions of alternative transportation modes such as road, rail, and intermodal. Research

indicated that shippers generally consider freight logistics characteristics and modal characteristics as important in the mode choice decision process. More recent research shows that logistics attributes such as frequency of shipments and flexibility of mode use, are important factors, as the globalization of business increases the need to have effective and efficient transport.

The shipper's freight logistics characteristics, such as the attributes of the shipper, the attributes of the commodities to be transported, and the spatial attributes of shipments such as the origin and destination of shipments and location of warehouses or distribution centers, strongly influence the choice of mode. These can be categorized as follows:

- the operational characteristics of the firms (i.e., business type, size of the firm, and type of transport fleets);
- the physical characteristics of shipments (i.e., type and value of goods, size of shipment, packaging of shipment, and frequency of shipment);
- information about the linkages and itinerary of the shipments (i.e., location of customers, location of warehouses, and distribution centers); and
- the choice between alternative carriers.

The shipper's decision to use a particular transportation mode is generally based on several factors. Several studies have been conducted to identify the specific service attributes often considered important in the shipper decision process. The decision-maker's own perception is a significant input to the decision-making process in mode selection. Also, the overall perception is typically driven largely by at least six perception factors, such as transport rates (including cost and charges), transport time (including transit time), transport reliability (i.e., on-time delivery), loss and damage to goods, modal availability (i.e., service frequency), and the value of transport service provided by carriers or transport service providers.

The rankings of factors differed between different studies of carrier selection in freight mode choice studies from the 1970s to the 2000s. The importance of freight rates increased in the 1980s, but reliability was always ranked on top. Transit time was the second most important factor in the 1970–80, but has steadily declined the importance in the 2000s. In some market segments, though, freight rates were more important than all other service factors. According to the study, the US shippers' overall perceptions are more affected by timeliness and availability than rates, which is often the last criterion for selecting a transport service provider. More recent studies have shown that shippers value service and reliability higher than cost or any other factors.

Freight Transport Decision-Makers

In freight transport, shipping decision-makers are generally classified into three categories: own-mode shippers, carriers, and shipping brokers. Own-mode shippers (or consignors) are shippers that deliver shipments with their own transport. Carriers, also called for-hire carriers, are the agents (e.g., trucking company, railway company) that move the shipment from the shipper to the consignee or customer, on behalf of the owners of the goods.

In the for-hire freight transport market, the prices of freight services are usually difficult for an analyst to get, as both transport firms and their clients try to keep such data confidential. [Ortuzar and Willumsen \(2011\)](#) classify what are thought to be the important factors affecting charges or cost implications:

- the length of the contract;
- the volume of shipments;
- the availability of alternative modal facilities;
- the availability of own vehicle; and
- the hierarchy of the transport system and network.

Freight agents or brokers act as intermediaries and handle the booking of the trucking company or other modes of transport for shippers without their own transport. These categories are not necessarily mutually exclusive. For example, it is possible for third-party logistics (3PL) companies to own their own vehicles and deliver a shipper's goods. 3PL providers typically specialize in the integrated operation of warehousing and transportation services. This can be scaled and customized to meet a customer's needs based on market conditions and the demands and delivery service requirements for the customer's products and materials. These days, most 3PL companies handle all aspects of transportation including warehousing, intermodal transfer, rail, container, cold storage, and air freight transport.

Road Transport: Owned-Fleet or For-Hire Vehicles

The firm or shipper has a choice whether a carrier will be contracted to move goods or the firm will own and use its own trucks. An owned-road fleet, also called private carriers (the United States), private trucking (Canada), or own account carriers (EU), are operated by firms whose principal occupation is not trucking, but who transport their own goods. For-hire (the United States and Canada) or hire-or-reward (EU) operations are transport services provided by transport service providers.

Owned-fleet trucking accounts for a significant amount of the trucking industry's activity in the United States and Canada. According to the study commissioned by the American Trucking Association in 2006, it has been estimated that private carriers

handled about 48% of the freight volume by tons hauled by all trucks in the United States, while owned-fleet trucking is dominated by a large number of small fleets operating in and around urban areas in Canada, where it holds an 85% share of the urban trucking market (Arthurs, 2006). The Private Motor Truck Council in Canada indicates that for urban goods movements, owned-fleets account for 85% of the trips within 160 km distance, for 25% of the trips within 500 km distance and for 10% of the trips of 500–1000 km.

Nowadays, not many firms set up their own private transportation operation, unless they have unique service requirements that cannot be met by for-hire carriers. The disadvantage of having one's own fleet is the investment in equipment, facilities, and people to operate and manage it. According to the Road Transport Vademecum (European Commission, 2011), about one-sixth of ton-km of road freight in the EU is transported by owned-fleet. In 2010, the share of owned-fleets was only 15.2%, compared with 15.8% in 2008 and 16.9% in 2009. Owned-fleet transport is relatively strong in Germany where it has a share of slightly more than 20% of total activity.

In other EU member states, such as France, the United Kingdom, Finland, and Sweden, the proportion of goods moved by for-hire carriers rose during the last ten years. The Vademecum also found that the proportion of empty runs is higher for owned-fleet transport than for for-hire transport. In 2010, it was 30.6% for owned-fleets but just 21.4% for for-hire carriers. Reflecting the situation with empty runs, for-hire transport is more efficient than owned-fleet transport in terms of the average load factor, with 8.5 ton-km for owned-fleets and 14.5 ton-km for for-hire carriers. The EU study reveals that around three-quarters of goods are transported 150 km or less. This is similar to the situation in the United States and Canada.

Modeling Freight Demand and Mode Choice

In freight demand studies, the discussion of shipper behavior modeling is quite extensive and approached in many different ways. Regan and Garrido (2002) classified the mode choice research on shipper behavior models as:

- the private versus for-hire carrier selection process;
- shipper-carrier relationships and carrier selection criteria, including the development of shipper-carrier alliances; and
- 3PL service provider.

Freight demand models have been developed since the 1960s. The published studies fall into three main categories: econometric models, spatial price equilibrium models, and network equilibrium models. Based on the nature of the data used for estimation, Winston (1993) also classified freight demand models as either aggregate or disaggregate models.

Aggregate models are defined as models that use an aggregate share of freight mode as the basic unit of observation for the model at a specified geographical level. The majority of freight models applied in practice have been of the aggregate kind, which applications follow the four-stage model as used in the personal transport model. In comparison to aggregate demand models, disaggregate models reflect in more detail the behavioral realities of freight transport decision-making, including the question of which “actor” (i.e., the owner of goods, carriers, or brokers) actually chooses the mode of transport. Disaggregate demand models are based on theories of individual behavior. Therefore, disaggregate demand models are probabilistic and attempt to explain individual behavior using individual data. The explanatory variables included in such models can have explicitly estimated coefficients. In principle, the utility function allows any number and specification of explanatory variables, as opposed to the case of the generalized cost function in conventional models, which is generally limited to only a few fixed parameters, such as the monetary and nonmonetary costs of a journey.

Disaggregate demand models can be classified into two classes: the so-called “inventory” and “behavioral” models. Inventory-based models analyze freight demand from the perspective of an inventory manager who deals with several production decisions, while the behavioral models deal with only one decision, the choice of mode.

In disaggregate freight mode choice models; mode choice is modeled at the level of individual shipments. The models use data from surveys of shippers, commodity surveys and so on. A combination of revealed preference (RP) and stated preference (SP) data have been widely used in the field of travel behavior studies. In general, the RP data captures individuals' actual choice responses to actual alternatives in current options. Using SP data allows the impact of policy variables on mode choice to be incorporated in the modeling. In particular, the relative importance of variables other than cost and time (e.g., reliability) can be identified. The models tend to be multinomial or nested logistic regression models, which can be based on the economic theory of individual utility maximization. The theory and method of discrete choice models were developed by Daniel McFadden, who won the Nobel Prize for economics. Kenneth Train extended the framework for modeling choice behavior by introducing simulation methods and software. Multinomial logit (MNL) and mixed logit (ML) are probably the most widely used methods for predicting freight mode choice.

During the last two decades, the development of discrete choice models has mainly been concerned with modeling the heterogeneity of the unobserved effects at the individual level. This is referred to as scale heterogeneity. The ML models account for scale heterogeneity in individual preference and error variance, but they cannot be directly explained by the inclusion of socio-demographic or behavioral variables through the use of random parameters. To control for scale heterogeneity, the generalized MNL model and generalized ML model have been introduced recently as an extension to the MNL models.

Behavioral Demand Model

The behavioral models attempt to explain freight transport demand as a process of utility maximization with respect to the choice of mode. For generating discrete choice models, “random utility theory” assumed that a freight shipper has possible multi-attribute transport modes from which to choose. The shipper will choose the mode with the largest associated utility among the available modes. Therefore, the mode choice probabilities depend on the random utility differences and their distribution (e.g., the receiver’s attitude toward risk), and the expected value of the unobserved modal, commodity and firm attributes, as well as measurement error.

Winston (1981) developed a model of freight demand based on the random utility model and used disaggregate data for a broad set of markets. In practice, the choice of mode may be made by either the shipping or receiving firm. When a firm is paying the transportation charges, the regional physical distribution or logistics manager at a shipping point generally has control over the manner of shipping. It is often the case, however, that purchase orders issued by the receiving firm include, among other things, the choice of mode as requested by the receiver’s logistics department. In this case, the receiving firm makes the mode choice, and hence, pays the transportation costs. Where the shipping firm makes the mode choice, it pays the transportation costs.

Different transportation modes are distinguished by their service attributes. The transportation cost is affected by such attributes as the availability of equipment, travel time, price, the flexibility of the service, reliability, insurance cost, loading facilities, etc. In addition, the component of the level of service of each mode introduces risk (e.g., higher risk of damage to goods and low on-time reliability) into the shipper’s decision regarding mode and destination.

Inventory-Based Model

Inventory-based models are theoretical in nature. Such a demand model analyses the transport mode decision made by shippers and the total demand for transportation services. The shipping cost per unit, the mean shipping time, the variance of shipping time, and the in-transit cost are typically assumed to determine how a shipper chooses between modes. The inventory theory investigates the trade-off between attributes for firms maximizing profit. From the total profit equation, the optimal demand for transportation in the firm’s supply chain (i.e., the location of logistics facilities, level of inventory, and vehicle routing) can be calculated using nonlinear estimation techniques.

The optimal demand for transportation, explicitly defined as an annual tonnage shipped, is calculated using nonlinear estimation techniques. This approach contributes analytical power to the inventory-based theoretical model and enables estimation of what would happen to demand given a change in any of the attributes. However, the limitations of the approach include its inapplicability to situations involving examining anything more than mode choice.

Revealed Preference and Stated Preference Methods

To develop a choice model, it is necessary to have data on an individual’s or firm’s choice response and operational characteristics. Parameters of behavioral models could be estimated using both surveys of actual travel behavior in a real context and surveys of hypothetical travel behavior in nonexistent scenarios.

In general, the RP survey captures individuals’ actual choice responses to actual alternatives. This method assumes that the preference of respondents can be revealed by their choice behavior. The advantage of the RP survey is that it captures actual behavior, which gives high validity to capturing freight shippers’ actual mode choices and ranking of several predetermined attributes that may influence the choices. However, because data from revealed preferences considers only current options, RP surveys are constrained to making predictions only for observed options. RP surveys are also restricted in their use due to the difficulty in measuring when a higher number of alternatives exist, and there is a need to allow for collinearity of attributes, as they can be closely correlated in real life.

The SP survey uses a specially constructed questionnaire to elicit estimates of the willingness to pay (WTP) for or the willingness to accept (WTA) a particular choice. WTP is the maximum amount of money an individual is willing to give to receive a good or service. WTA is the minimum amount of money they would need to abandon a good or service.

Different SP methods are available under a wide variety of names, for including contingent valuation (CV), choice experiment (CE), conjoint, functional measurement, trade-off analysis, and the transfer price method. The best-known are the CV and CE methods. CV is a technique for the valuation of nonmarket resources and involves constructing and presenting hypothetical situations to the survey respondents. The term “CV” was originally used for a survey to estimate the economic value of a public good or service. The CV conveys three main elements: (1) information related to preferences is obtained using an SP survey, (2) the study’s purpose is placing an economic value on one or more goods, and (3) the good (s) being valued are public ones. A detailed description of a good or service, how it will be provided, and the method and frequency of payment, are usually highlighted in the CV questionnaire. Following this, questions are posed in order to infer a respondent’s WTP or WTA. CV questionnaires also generally contain additional questions to gain information on a respondent’s socioeconomic and demographic characteristics, their attitudes toward the goods, and the reasons behind their stated valuations. The key outcome of the analysis of the responses is an estimate of the average WTP across the sample of people surveyed. If the sample is representative of the target population, then this estimate can be expanded to obtain an estimate of the total value of the outcome or good.

The CE method focuses on a good's or service's attributes and their values. There are three essential elements to the CE method: (1) a respondent is asked to make a choice between two or more alternative hypothetical options (choice set), (2) the alternatives are described by its characteristics, also referred to as attributes, and (3) the attributes are specified by several levels depending upon the nature of these attributes. To develop valuation estimates, CE questionnaires present respondents with a series of alternative goods or services. The alternative descriptions are constructed by varying the levels of the attributes. Depending on the specific CE adopted, respondents are either then asked to rank, choose, or rate the descriptions presented.

The SP method is more flexible than the RP method and can avoid multicollinearity problems, such as similarities in attribute levels and the shared-relationships between attributes, because all the relevant information for making choices between alternatives is provided to the respondent. However, there has been debate regarding whether stated intentions reflect current conditions, as SP data does not always accurately predict overall market share. Despite this weakness, the SP approach is widely used for the valuation of nonmarket resources, such as environmental preservation or the impact of contamination; while those resources give people "utility," they do or may not have a market price and are not sold directly. In addition, SP studies are used to investigate respondents' preferences toward alternative (s) not yet available in the market (e.g., improvement of the public transport system or introducing a new service), or to investigate alternative (s) outside the current technological frontier.

Utility is a broad concept that may refer to a person's welfare, well-being, or happiness. For example, freight shippers or firms receive non-price benefits from choosing optimal transport modes to distribute goods, but the decision-making process would be difficult to value using a price-based model such as the RP method. The SP method is a good technique for estimating the non-price aspects.

Traditionally, RP data were commonly used in transport analysis for estimating ton-km transported by a given mode. However, in the last two decades, SP techniques have been gradually introduced in passenger transport demand analysis and, less frequently, in freight transport demand analysis. The advantages and disadvantages of the two techniques compensate for each other and can be used jointly.

References

- Arthurs H.W., 2006. Fairness at work: federal labour standards for the 21st century, Fed. Labour Stand. Rev., Human Resources and Skills Development Canada, Ottawa.
- European Commission, 2011. Road freight transport vademecum: market trends and structure of the road haulage sector in the EU in 2010. DG for Mobility and Transport, European Commission. Available from: <http://ec.europa.eu/transport/modes/road/doc/2010-roadfreight-vademecum.pdf>
- OECD/ITF, 2017. ITF (International Transport Forum) Transport Outlook 2017. Available from: <https://www.itf-oecd.org>.
- Ortuzar, J. de D., Willumsen, L.G., 2011. Modelling Transport, fourth ed. Wiley, Chichester, UK.
- Regan, A.C., Garrido, R.A., 2002. Modeling freight demand and shipper behavior: state of the art, future directions (No. UCI-ITS-LI-WP-02-2), The Institute of Transportation Studies.
- Winston, C., 1981. A disaggregate model of the demand for intercity freight transportation. *Econometrica: J. Economet. Soc.* 49 (4), 981–1006.
- Winston, C., 1993. Economic deregulation: days of reckoning for microeconomists. *J. Econ. Lit.* 31 (3), 1263–1289.

Further Reading

- Ballou, R.A., 2003. Business Logistics Management, fifth ed. Prentice-Hall, Englewood Cliffs, NJ.
- Ben-Akiva, M., Lerman, S.R., 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambridge, Massachusetts.
- Greene, W.H., 2009. Discrete choice modeling. In: Mills, T., Patterson, K. (Eds.), *The Handbook of Econometrics. Applied Econometrics*, Vol. 2. Palgrave, London.
- Hensher, D.A., Rose, J., Greene, W.H., 2015. *Applied Choice Analysis*, second ed. Cambridge University Press, Cambridge.
- Louviere, J.J., Hensher, D.A., Swait, J.D., 2000. *Stated Choice Methods: Analysis and Application*. Cambridge University Press, Cambridge.
- Train, K., 2009. *Discrete Choice Methods with Simulation*, second ed. Cambridge University Press, Cambridge.

The Value of Time in Freight Transport

María Feo-Valero^{*,†}, Amaya Vega[‡], Bárbara Vázquez-Paja[†], ^{*} Department of Applied Economics II, University of Valencia, Valencia, Spain; [†]IEI, Valencia, Spain; [‡]GMIT Galway Campus, Galway, Ireland

© 2021 Elsevier Ltd. All rights reserved.

Introduction	236
Estimating the Value of Time in Freight Transport: Key Challenges	236
The Decision Maker	236
Difficulty in Obtaining Data	237
Heterogeneity	237
Value of Freight Transport Reliability	239
Conclusions and the Future of the VOT in Freight Transport	239
See Also	240
References	240
Further Reading	240

Introduction

The value of time (VOT) is a critical aspect in decision-making in transportation. One of the key objectives of any transport investment is the reduction in travel times. Governments employ cost-benefit analysis (CBA) as a framework for a transparent evaluation of the economic and social effects of projects and to ensure that their actions and investments in transportation infrastructure use resources in the most efficient manner. The process of evaluating the benefits of particular transport infrastructure investments often requires assigning money values to factors that lack observable market prices. One of the most important of these factors is travel time, which has been the subject of research for decades, with significant advances both from a theoretical and empirical perspective.

The theoretical underpinnings of the VOT are inherently related to the economic theory of time allocation introduced in the 1960s (Becker, 1965). Based on the idea that time is a scarce resource, pioneering theories were formulated in the context of people—commuters—and framed around a utility maximization problem, subject to constraints on income and the minimum amount of time required by the various activities that individuals undertake.

In the case of freight transport, while the problem still remains a maximization problem, it is generally defined as a profit maximization problem, subject to constraints on costs of production and transportation. The main difficulty with freight transport is to identify the economic agent whose profit function needs to be maximized. While the decision about the transport mode and route may be made by a transport hauler, a freight forwarder or other transport operator, the benefits obtained from a reduction in transit times may be part of the profit function of the shipping firm, as well as of the firm or consumer that receives the goods.

In general, the VOT is defined as the monetary value that users are willing to pay—or to accept—for a reduction—or an increase—in travel times along a particular origin-destination pair. In this chapter, the specific characteristics of the VOT in freight transport are outlined, highlighting how the various methodologies have dealt with the idiosyncrasies of this key indicator in infrastructure-related CBA.

The monetary value of travel time in freight transport goes beyond the valuation of the willingness to pay for transit time savings. While this is indeed a key aspect to take into account, the final delivery time—total time since the customer places the order until he receives it—is determined not only by how fast goods are transported from origin to destination, but it is also determined by the reliability of these transit times and the frequency of the transport service. Monetary values for the reliability of travel times relate to reductions in travel time variability on a determined route. This time variability has been identified as a key determinant in the valuation of time in freight transport.

Overall, the relationship between transit time, reliability, and frequency becomes more relevant for transport modes with fixed schedules and comparatively lower levels of frequency, such as maritime and rail transport. This relationship is particularly important for multimodal transport chains typically characterized by having a more limited scope for transit time improvements. In modal switching initiatives, where governments are to make policy decisions about freight transport by road and the promotion of alternative modes to induce modal shift, the VOT should be assessed in conjunction with the value of reliability (VOR) and the value of frequency.

Estimating the Value of Time in Freight Transport: Key Challenges

The Decision Maker

Unlike passenger transport, where the decision maker is the traveler, there are multiple agents involved in the freight transport decision-making process, with no clearly identifiable pattern of responsibility and involvement by each decision maker in the

various logistics decisions. Agents involved in freight transport decision-making are normally aggregated into two groups: carriers and shippers. Carriers are those agents whose primary activity is directly related to the transportation of the goods. These agents are involved in the provision of transport services both directly—road haulers, shipping lines, rail operators, and air carriers—or indirectly—freight forwarders and logistics operators. Alternatively, shippers are those agents whose primary activity is related to the production and distribution of goods. Shippers can either be the shipping firm or the receiving firm, and the agreed International Commercial Terms or INCOTERMS ultimately determines this role. In some cases, shippers act both as shippers as well as carriers—own account shippers—carrying out their own transport operations.

The correct identification of the decision maker is one of the key challenges in modeling freight transport demand and, therefore, in estimating the VOT in freight transport. This is important from a data collection perspective as well as from a modeling perspective. First, it is crucial to ensure that the agent providing the information is the one responsible for the final decision under study. Here the difficulty is not limited to identifying the company responsible for the decision, but to identifying within that company, the person in charge of these decisions. From a modeling perspective, while the benefits obtained by carriers from transit time savings relate to savings in operational costs, such as labor costs, fuel, and vehicles, the benefits obtained by shippers relate to savings in the associated production or distribution process. This has motivated a distinction in the literature between two approaches for estimating values in freight transport: the factor-cost approach, which focuses on the transport cost component of the VOT, and the modeling approach, which centers on the cargo component of the VOT, with the sum of both representing the overall VOT (De Jong et al., 2014).

Under the factor cost approach, the debate focuses on the type of costs that should be included in the estimation process, which in turn depends on the time horizon of reference. In the short run, firms may have difficulties in taking full advantage of the benefits associated with transit time savings. However, in the long run, with a greater perception of the steady benefits from transit time savings, firms are likely to make a more effective use of resources that may result in a further cost reduction and therefore a higher VOT.

The modeling approach is the most widely used method in the estimation of the VOT in freight transport. Based on the implied time-cost trade-offs made by the decision maker, this method is based on the use of discrete choice models, where the VOT corresponds to the time-cost marginal rate of substitution. These models are in line with those used in passenger transport studies, but with added specific challenges specifically associated with the freight transport context, such as the greater heterogeneity in transport flows and difficulty in obtaining data.

Difficulty in Obtaining Data

The data collection process in freight transport studies can be arduous and expensive. At the aggregate level, where the basic unit of observation is the market share at the regional or national level, freight transport datasets are relatively scarce, lacking in detail and too broad to carry out any comprehensive demand side analysis. Therefore, there is often the need to complement these datasets with information from other secondary sources.

Most of the empirical applications that use the modeling approach require the collection of their own disaggregated data through surveys and questionnaires. In these cases, the basic unit of observation is the transport or logistics decision made by the decision maker. Two main approaches can be distinguished: the inventory approach (Baumol and Vinod, 1970), where the transport decision is integrated in the production decision-making process, and the behavioral approach, where the transport decision is based on a consumption decision based on maximizing utility (Winston, 1983). In any case, the data collection process in freight transport is generally more complex than in the case of passenger transport. This is mainly due to the reduced availability of firms willing to participate in the studies, as well as the greater reluctance to provide sensitive commercial information.

Under the behavioral approach and depending on the type of data used, studies can be classified into revealed preference studies—data from choices actually made by decision makers—and stated preference (SP) studies—data from choices based on hypothetical situations, with the latter being the most widely used in freight transport applications. In SP studies, the firm is asked to declare their choice under a hypothetical situation put forward by the researcher and, therefore, does not have to provide information about their actual behavior, which contributes to reducing the respondent's understandable reluctance to provide sensitive commercial information. However, the main challenge in SP studies is in ensuring that these hypothetical situations are perceived by the interviewee as realistic, as well as in avoiding lexicographic behaviors due to an inadequate design of the experiment, that is, negligible trade-offs between the levels of service of the attributes displayed. The high levels of heterogeneity of freight transport flows make the identification of the boundary values leading to changes between alternatives particularly complex. In this sense, the use of SP experiments pivoted around the reference alternative provided by the interviewee, and the definition of the variations in the levels of the attributes in relative terms allows the experiment to be adjusted to the particularities of the shipment. Moreover, recent methodological advances in the last decade have increased the efficiency of SP studies, that is, efficient designs (Rose et al., 2008), and have allowed these methods to partially offset the lower sample size often found in freight transport applications.

Heterogeneity

There is a considerable level of heterogeneity in freight transport flows. This is due to the physical characteristics of the goods transported and the transport chains involved, as well as to the logistics requirements of individual consignments. In contrast to

passenger transport, where the unit of reference is straightforward, the traveler, there is no single unit of reference in freight transport. With a wide range of units used—tonnes, shipments, TEUs, lorry, wagon, barge, etc.—the choice of the reference unit ultimately depends on the decision maker, the data availability, and the research objective. In any case, even with identical units of reference, results from different studies may not be directly comparable. The high level of heterogeneity that characterizes the goods transported means that if using the shipment as the reference unit, it may differ by size to a great extent from, for example, a shoebox to a grain barge. This also applies to some extent to the tonne, with great heterogeneity in terms of volumetric weight across goods.

Some of this heterogeneity can be captured by simply taking into account the mode of transport and type of cargo—bulk or unitized and, within the latter group, full container load (FCL) or less container load (LCL). For example, dry bulk shipments, which are unpacked and homogeneous, show certain specific characteristics from a logistics perspective—larger size and reduced unit value—and therefore, attributes, such as the cost of transport, become a key determinant of their relative competitiveness. This means that generalizing the various estimated transport attribute values, such as the VOT, may be more challenging. In the case of unitized cargo and in particular the transport chains associated with LCL, which imply additional loading/unloading operations, there are often significant differences in the relative competitiveness for the various transport modes, being the VOT per unit of product usually higher for smaller shipments—LCL—than for FCL. In any case and from a CBA perspective, while taking into account the distinction between FCL and LCL may result in higher accuracy in the valuation of time, this may be partially hampered by the relatively reduced quality of data from transport national statistics, which often ignore this distinction by equating shipments with trips.

There is a general consensus across the literature that freight transport valuation models should incorporate to some extent basic systematic sources of heterogeneity, such as those linked to the decision maker, the type of mode of transport, the format of cargo, size, distance, or the value of the goods. This is often carried out by either segmenting the sample or through interactions with other transport attributes, such as cost or transit time. The challenges associated with modeling freight transport demand—large heterogeneity, small population in relative terms and the difficulty in obtaining freight data—can make it difficult to simultaneously analyze all these factors, even though technological and methodological advances have made it possible to better address the issue of heterogeneity in the valuation of freight transport attributes.

From a methodological perspective, the use of mixed logit models (McFadden and Train, 2000)—where coefficients can be specified as random variables following a continuous distribution—and latent class models (Greene and Hensher, 2003)—where the population is split into Q latent classes with heterogeneous preferences across classes and homogeneous preferences within the class—have allowed for significant advances in the way heterogeneity is now dealt with in freight transport studies.

Advances are also being made in the way in which decision-making processes are incorporated into the models. In this sense, while model specifications with symmetrical preferences—implying equal willingness to pay (WtP) and willingness to accept (WtA) measures for variations in transit times—are more common in the case of freight transport, there is an increasing body of literature that introduce elements of prospect theory (Kahneman and Tversky, 1979) to assess if the loss aversion hypothesis holds in the case of freight transport. These empirical applications are focused on determining whether the impact of deterioration in the level of service is higher than an improvement in the same magnitude, which would imply that the WtA would be greater than the WtP for a particular level of service. This line of research is especially relevant in the context of reducing the environmental impact of transport systems as a policy objective. From this perspective, the development of more environmentally efficient transport chains might imply reconsidering the levels of service provided in terms of transit time, in which case the validation of the loss aversion hypothesis becomes crucial.

Until now, most of the empirical evidence available on this issue in the area of freight corroborates the loss aversion hypothesis (Masiero and Hensher, 2010; Masiero and Rose, 2013). In any case, differences in WtP and WtA will ultimately depend on the relative level of service provided across the various transport alternatives considered and the extent of the deterioration or improvement considered. In this sense, more research is needed to analyze the existence or not of diminishing sensitivity, as there is hardly any empirical evidence on this issue in the field of goods transport.

Noncompensatory behaviors are also being considered through the specification of piecewise linear functions (Swait, 2001). In these specifications, a penalty component linked to the magnitude of the cutoff violation is added to the traditional linear-in-the-parameters nonpenalized utility function, where low levels of service of one attribute can always be offset by high levels of service of other attributes. This makes it possible to obtain different VOTs, depending on whether the corresponding cutoff is disregarded. When cutoffs are imposed, the VOT is expected to be higher when cutoffs are breached. While the empirical evidence on the imposition of cutoffs in the evaluation of transit time is not unique, in the case of transit time reliability it is possible to identify a larger number of investigations showing that these cutoffs apply (Danielis and Marcucci, 2007; Feo-Valero et al., 2016). Indeed, in some of those cases, the willingness to pay for improvements in the reliability within the “acceptable level of service” is zero. This result has significant implications in terms of the valuation of the benefits derived from specific actions.

Heterogeneity is not such a major challenge in valuation studies under the factor cost approach. The unit of reference is the unit of transport and the impact of variations in transit time over the operating cost do not usually depend on the size and the type of cargo. Under the factor cost approach, the difficulty comes from the conversion to a single unit of reference of the transport and cargo components of the VOT. Moreover, since empty trips may represent a significant proportion of the overall trips, a correction factor should be applied to the good component of the VOT when calculating total cost benefits linked to a specific transport project. Here again, accurate and up-to-date statistics on the proportion of empty trips would be needed.

Value of Freight Transport Reliability

The benefits from reductions in travel time variability were not taken into account in the valuation of travel time in freight transport until recent times. Measuring these benefits is important, as well as including them in the CBA of transport investment projects. From the shippers' perspective, these benefits may allow them to cut down on inventory costs and develop just-in-time strategies. From the carriers' perspective, benefits from reductions in travel time variability may result in greater optimization of their resources through more efficient planning. With the recent developments in production and distribution processes based on fast and reliable transport chains, minimizing vulnerability and the risk of disruption from unexpected delays have become more important. This has motivated an increasing interest in the measurement of the VOR in freight transport.

Most of the empirical studies regarding the value of freight time have historically focused on the analysis of transit time, with less empirical evidence on the role that delays had on the logistics decision-making process and the WtP for improvements in reliability. One of the main reasons for this limited empirical evidence is the difficulty in defining and measuring reliability. While other typical freight transport attributes, such as cost and frequency, are generally perceived by the decision maker as static for a given shipment, the key aspect to consider by the decision maker in the case of reliability is their experience over multiple shipments, which requires a greater level of mental effort (Tseng et al., 2009).

In freight transportation, reliability in the terms of delivery refers to the extent to which transit times for particular shipments correspond with the transit times initially agreed and expected by the decision maker. In general, reliability refers to the level of unexpected variability in travel times. There is no single method to measure this variability, which implies that a range of different reference units can be used. Therefore, it becomes increasingly hard to make comparisons across estimates from different empirical studies.

One of the most widely used methodologies to measure reliability in the terms of delivery is the percentage of shipments arriving within the scheduled time or alternatively, the percentage of shipments suffering significant delays. This is often completed with additional information on the average extent of the delays when these actually take place. Reliability can also be measured by calculating either the variance or the standard deviation with respect to the agreed transit time along the logistics chain under study. However, the need for the decision maker to assess a relatively complex concept, such as the variance, has motivated that recent studies substitute this variable by a range of possible transit times with an equal chance of occurring.

From the shipper's perspective, an additional aspect to consider is the impact of reliability on the preferred departure and arrival times (Fowkes, 2007). Taking into account these issues leads to the discussion between nonscheduled and scheduled shipments, depending on how important is the timing of the shipment for the decision maker. Under the scheduled-approach timing preferences are explicitly introduced in the model, usually through the magnitude of early and late schedule delays. The importance attached to the timing is determined by the production strategy adopted by the shipper/receiver, the type of transport chain—unimodal or multimodal—and the level of service in terms of frequency. While the importance of departure and arrival times for full truck loads may be relatively low, as the transport provider can adapt to the loading time window set by the shipper, it becomes critical in the case of multimodal transport chains where the main transport mode—maritime, IWT, rail, or air—displays relatively low levels of frequency and fixed schedules. In this sense, there are only a few studies that specifically analyze if the relative value of delays is affected by the point at which the delays take place along the door-to-door transport chain, by taking into account the interrelation between the different nodes and agents involved.

The reliability ratio (RR), defined as the ratio between the VOR and the VOT, is a useful indicator in empirical studies that include both estimates. Most studies have reported RRs greater than one, which implies that there is a relatively higher weight given to reliability relative to transit time (Shams et al. 2017). This suggests that while transit times matter in the case of freight transport, it is the reduction in the variability in transit times that matters the most to decision makers.

Conclusions and the Future of the VOT in Freight Transport

One of the key aspects of the evaluation of any transport project involves the monetary valuation of travel time. The intrinsic characteristics of freight transport make the estimation of the value of freight travel time somehow more challenging than the corresponding value for passenger transport. These challenges involve the identification of the decision maker—carriers or shippers—and the heterogeneity of freight transport flows—transported goods, transport chains, units of reference—as well as an additional difficulty in data collection due to the commercial nature of the activities under study.

The high level of casuistry inherent in logistic flows makes it very complicated to provide a unique VOT reference value. Thus the relative weight given to time in a short distance shipment will probably have little to do with that of an interoceanic shipment whose estimated transit time is more than two weeks. In the latter case, it is likely that the relevant aspect for the decision maker is the reliability and frequency of service, rather than the transit time. In fact, the latest restructuring movements of liner shipping services undertaken by some of the major shipping lines are partially aimed at improving the reliability of their schedules. Special mention should also be paid to the estimation of the value of freight transport attributes in the last mile logistics context, where limited and inflexible delivery windows together with the increase of transport flows linked to ecommerce operations make the analysis of the VOT and its reliability even more relevant.

The importance of recognizing the relationship between transit time, and the reliability of transit time has been highlighted in recent empirical studies. The importance of estimating the VOR as the level of unexpected variability in travel times also becomes

particularly relevant in the case of multimodal transport chains and the promotion of environmentally sustainable transport chains.

Taking into account the aforementioned challenges in freight transport, special attention should be paid to the effect that future technological innovations and ICT developments may have on the valuation of transit time and reliability. Digitization and automatization are two of the main disruptive innovations in the transport and logistics industry. New digital developments built on big data, cloud and connected platform technologies, blockchain, or AI will provide ease of access and transparency for the entire logistics chain, with a swift, near real-time integrated transport service. This implies greater flexibility in dealing with potential delays in delivery time and an increased ability to reduce variability in transit times. While having real-time information on unexpected delays is still negatively affecting passenger's utility, technological advances that provide this information for freight transport operators and users can have unprecedented effects in the level of industry competitiveness in the long term.

See Also

Transport Economics

Demand for Freight Transport; Value of Time; Valuing Travel Time Variability: Scheduling and Reduced form Models; Loss Aversion and Size and Sign Effects in Value of Time Studies; Value of Time for Freight; Estimating the Value of Time

Freight Transport and Logistics

Logistics and Supply Chain Management; Logistics Performance and Supply Chain Management; Freight Transport Policy; Logistics Costs; National Freight Transport Models; Freight Vehicle Routing and Scheduling; Carrier Selection; Freight Mode Choice; Behavioral Research in Freight Transport; Intermodal Freight Transport

Transport Modeling and Data Management

Mode Share/Choice Models; Departure Time Choice Modeling; Transportation Statistics and Databases; Transportation Surveys and Diaries; Stated Preference Surveys: Experimental Design and Modeling

Transport Modes

Data Sources and Transport Modes; Modeling Freight Transport Mode Choice

References

- Baumol, W., Vinod, H.D., 1970. An inventory theoretic model of freight transport demand. *Manag. Sci.* 16 (7), 413–421.
- Becker, G.S., 1965. A theory of the allocation of time. *Econ. J.* 75, 493–517.
- Danielis, R., Marcucci, E., 2007. Attribute cut-offs in freight service selection. *Transp. Res. E Logist. Transp. Rev.* 43, 506–515.
- De Jong, G.C., Kouwenhoven, M., Bates, J., Koster, P., Verhoef, E., Tavasszy, L., Warffemius, P., 2014. New SP-values of time and reliability for freight transport in the Netherlands. *Transp. Res. E Logist. Transp. Rev.* 64, 71–87.
- Feo-Valero, M., García-Menéndez, L., del Saz-Salazar, S., 2016. Rail freight transport and demand requirements: an analysis of attribute cut-offs through a stated preference experiment. *Transportation* 43, 101–122.
- Fowkes, T., 2007. The design and interpretation of freight stated preference experiments seeking to elicit behavioural valuations of journey attributes. *Transp. Res. B Methodol.* 41, 966–980.
- Greene, W.H., Hensher, D.A., 2003. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transp. Res. B Methodol.* 37 (8), 681–698.
- Kahneman, D., Tversky, A., 1979. An analysis of decision under risk. *Econometrica* 47 (2), 263–292.
- Masiero, L., Hensher, D.A., 2010. Analyzing loss aversion and diminishing sensitivity in a freight transport stated choice experiment. *Transport. Res. A Policy Pract.* 44 (5), 349–358.
- Masiero, L., Rose, J.M., 2013. The role of the reference alternative in the specification of asymmetric discrete choice models. *Transp. Res. E Logist. Transp. Rev.* 53 (1), 83–92.
- McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. *J. Appl. Econ.* 15 (5), 447–470.
- Rose, J.M., Bliemer, M.C., Hensher, D.A., Collins, A.T., 2008. Designing efficient stated choice experiments in the presence of reference alternatives. *Transp. Res. B Methodol.* 42 (4), 395–406.
- Shams, K., Asgari, H., Jin, X., 2017. Valuation of travel time reliability in freight transportation: a review and meta-analysis of stated preference studies. *Transp. Res. A Policy Pract.* 102, 228–243.
- Swait, J., 2001. A non-compensatory choice model incorporating attribute cutoffs. *Transp. Res. B Methodol.* 35 (10), 903–928.
- Tseng, Y., Verhoef, E., de Jong, G., Kouwenhoven, M., van der Hoorn, T., 2009. A pilot study into the perception of unreliability of travel times using in-depth interviews. *J. Choice Model.* 2 (1), 8–28.
- Winston, C., 1983. The demand for freight transportation: models and applications. *Transp. Res. A* 17 (6), 419–427.

Further Reading

- De Jong, G.C., 2000. Value of freight travel-time savings. In: Hensher, D., Button, K. (Eds.), *Handbook of Transport Modelling*. Pergamon Press, Oxford, pp. 553–564.
- Feo-Valero, M., García-Menéndez, L., Garrido-Hidalgo, R., 2011. Valuing freight transport time using transport demand modelling: a bibliographical review. *Transp. Res.* 31 (5), 625–651.
- Goenaga, B., Cantillo, V., 2018. Willingness to pay for freight travel time savings: contrasting random utility versus random valuation. *Transportation*, doi:10.1007/s11116-018-9912-5.

- Halse, A., Samstad, H., Killi, M., Flügel, S., Ramjerdi, F., 2010. Valuation of transport time and reliability in freight transport. Summary of TOI-report 1083/2010, Oslo.
- Ortuzar, J., de, D., Willumsen, L.G., 2001. Modelling Transport. John Wiley & Sons, Ltd., Chichester, UK.
- Tavasszy, L., de Jong, G., 2014. Modelling Freight Transport. Elsevier, London, UK.
- Vierth, I., 2013. Value of freight time savings. In: Ben-Akiva, M., Meersman, H., Van de Voorde, E. (Eds.), Freight Transport Modelling. Emerald Group, Bingley, UK, pp. 299–317.
- Zamparini, L., Reggiani, A., 2007. Freight transport and the value of travel time savings: a meta-analysis of empirical studies. *Transp. Rev.* 27 (5), 621–636.
- Zamparini, L., Reggiani, A., 2010. The value of reliability and its relevance in transport networks. In: Givoni, M., Banister, D. (Eds.), Integrated Transport: From Policy to Practice. first ed. Routledge, pp. 97–116.

Behavioral Research in Freight Transport

Edoardo Marcucci^{*,†}, Valerio Gatta[†], Michela Le Pira[‡], ^{*}Faculty of Logistics, Molde University College, Molde, Norway; [†]Department of Political Science, Roma Tre University, Rome, Italy; [‡]Department of Civil Engineering and Architecture, University of Catania, Catania, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	242
Behavior Change and Stakeholder Engagement in UFT Policy-Making	242
Stakeholder Behavioral Analysis	243
Discrete Choice Models	243
Agent-Based Models	243
Stakeholder Behavioral Analysis	244
Conclusion	245
Acknowledgment	245
Biographies	245
References	246
Further Reading	246

Introduction

Freight transport is a fundamental component of modern life and the world economy, increasingly influenced by globalization and e-commerce. Since cities are expected to host the majority of the world population in the next years, most of the goods will have urban settlements as final destination, with an increase of transport activities and space sharing among different components. This implies greater stress on city sustainability, in terms of negative externalities like congestion, accidents, noise, reduced quality of life, air pollution, and climate change. Besides, urban freight transport (UFT) is a complex world characterized by numerous actors with heterogeneous interests. Cities need appropriate solutions and policies to increase their sustainability and efficiency. However, policies are effective only if the impacted agents adopt them and change their behavior. As pointed out by Holguín-Veras et al. (2017a), “the lack of research on how best to influence the behaviour of the freight system and supply chains remains a formidable obstacle to the implementation of comprehensive policies to improve their overall performance.” A strong knowledge of behavioral underpinnings characterizing stakeholder choices is thus required.

Based on this premise, this paper reports on the international debate on freight behavior research by focusing on the importance of: (1) behavior change and stakeholder engagement for UFT policy-making; and (2) *ex-ante* behavioral assessments of UFT policies to guarantee their uptake and spread. New behavioral analyses are needed to complement technical analyses, by considering both solution feasibility and policy acceptability. We illustrate a procedure to perform a stakeholder behavioral analysis, based on integrated models to adequately tackle this issue. Before introducing the procedure, a literature review on behavior change and stakeholder engagement in UFT policy-making is presented.

Behavior Change and Stakeholder Engagement in UFT Policy-Making

The complexity of UFT issues implies that there are no perfect solutions, forcing policy-makers to accept compromises based on the trade-offs involved. These should be clearly identified when evaluating and selecting alternatives, representing an important and crucial phase which should not be underestimated. In this respect, policy-makers are interested in the *ex-ante* assessment of UFT solutions, which can be difficult due to the multiple interests and variety of stakeholders involved. This is even truer for innovative solutions in freight transport based on stakeholder cooperation, for example, off-hour deliveries (Gatta et al., 2019; Holguín-Veras, 2008; Marcucci and Gatta, 2017), or when promoting the sharing economy paradigm, for example, “crowdshipping” (Marcucci et al., 2017a; Punel and Stathopoulos, 2017; Simoni et al., 2019). In addition, it should be noted that policy-planning activities do not take place in a vacuum and that the only successful way to foster change is to constructively engage all stakeholders to develop consensus-based strategies. From a modeling standpoint, the individual interests and the ensuing behavioral changes that a modification of the *status quo* would most likely induce must be included in the policy-making process (Gatta et al., 2017).

Behavior change is the end goal policy-makers aim for. A recent trend to engage and promote sustainable behaviors foresees the use of game dynamics. This is usually referred to as “gamification,” that is, the use of game design elements in nongame contexts. This instrument is also gaining success in the mobility domain. In this respect, Marcucci et al. (2018) proposed an advanced user-centered approach to reliably build a user-centered gamification process applied to UFT that accounts for players’ heterogeneous preferences.

Behavioral analysis is fundamental to elicit stakeholder preferences and investigate their utility maximizing behavior. This can be performed through disaggregate behavioral freight models like discrete choice models (DCMs) (Gatta and Marcucci, 2015; Holguín-Veras and Wang, 2013; Marcucci and Gatta, 2016). Typically, preferences about hypothetical policy scenarios are elicited through stated choice experiments (SCEs), while DCMs are used to estimate the willingness to pay related to single policy components. Besides, one should not underestimate the possible policy effects arising from the dynamic interactions among agents often characterized by conflicting objectives.

Stakeholder engagement in the planning processes and a deep understanding of the behavioral aspects, intertwined with their decision-making process, are therefore crucial to define and implement effective policies, given the variety of heterogeneous UFT stakeholders with conflicting interests. However, the traditional planning approach transport planners and practitioners adopt does not consider freight a priority, focuses mainly on passengers, and is implemented within a well-established framework.

In this respect, transport planning has undergone a continuous evolution over the years, from “planning” intended as a “plan document” to a “plan process,” and from a “strongly rational” to a “cognitive rationality” approach, where the concept of choice optimization has been overcome with that of satisfying the community needs in an evolving process (Cascetta et al., 2015). However, while quantitative methods for *ex-ante* technical/economic assessments of transport policies are well established, there is still a general deficiency of methods aimed at assessing policy acceptability from a stakeholder’s point of view (Le Pira, 2018). New behavioral analyses are needed to complement the technical ones, so to identify and evaluate effective and shared policies, by considering both their feasibility and acceptability. In what follows, this paper illustrates a procedure aimed at a stakeholder behavioral analysis based on integrated models.

Stakeholder Behavioral Analysis

The important role of a behaviorally consistent and agent-specific approach in modeling UFT policy impacts on stakeholder behavior is widely recognized, even if only a limited number of papers have adopted this perspective, both for data acquisition and for model estimation (Gatta and Marcucci, 2016). The proposed stakeholder behavioral analysis is based on a well-thought-out integration of DCMs and agent-based models (ABMs). The rationale for these two approaches will be presented in the following.

Discrete Choice Models

DCMs can be quite useful in helping policy-makers devise the possible policies stakeholders accept, given their preferences for alternative configurations. One can also estimate the willingness to pay/accept for each mix of measures, providing a monetary evaluation of alternative policy components.

Data acquisition is based on SCEs (Louvière et al., 2000). Sampled stakeholders respond to a sequence of choice tasks where they can choose one policy alternative out of a numerable and finite choice set. Stakeholders’ choices are supposedly based on the perceived utility that alternative policy components have. Microeconomics assumes rational agents maximize utility, which is modeled as a random variable reflecting the assumption that the decision-maker has perfect discriminative capability while the analyst has incomplete information. Utility is composed of a deterministic and a stochastic term. Different assumptions on the distribution of the latter give rise to different DCM specifications. DCMs represent valuable instruments in: (1) addressing stakeholder preferences with respect to alternative policy configurations; (2) determining the respective substitution rates between the different, and possibly contrasting, policy components; and (3) predicting the behavior and acceptability through scenario analysis. While all these features are useful and desirable with respect to an empirically relevant and robust *ex-ante* policy evaluation, when dealing with simulations DCMs do not prove versatile in mimicking interaction effects among different stakeholders which, in reality, permeate the UFT environment. In this respect, knowing how single stakeholders would choose is not sufficient when the system outcome is highly influenced by the interactions taking place among agents. A dynamic simulation approach seems more appropriate to deal with interaction effects.

Agent-Based Models

ABMs aim at simulating a system whose main components are single agents, defined as autonomous entities capable of acting under certain behavioral rules (Macal and North, 2010). The structure of an ABM consists of: (1) a set of agents, with certain properties and behaviors; (2) a set of relationships and methods of interaction, with a topology that defines how and with whom agents interact; and (3) the agents’ environment, where agents interact with each other and with the environment itself. Agents are provided with simple behaviors linked to local information. However, when many agents are simulated as a group, emergent phenomena often arise. ABMs are widely used to represent real-world complex systems and to reproduce distributed decision-making processes as the outcome of a set of autonomous entities interacting in a common environment. Several ABMs have been implemented lately to reproduce and investigate UFT problems when explicitly considering the interaction among UFT actors (Holmgren et al., 2012; Roorda et al., 2010; van Duin et al., 2012).

However, despite the convenience and adaptability of these models, a lack of data to calibrate the models is a common problem encountered in many ABMs. In general, this induces assigning random values to the attributes, so as to study the emergence of collective phenomena in a bootstrapping fashion. In this regard, data characterizing single agents’ preferences, such as utility functions derived from SCEs, can improve the predictive capacity of the model and its forecasting reliability.

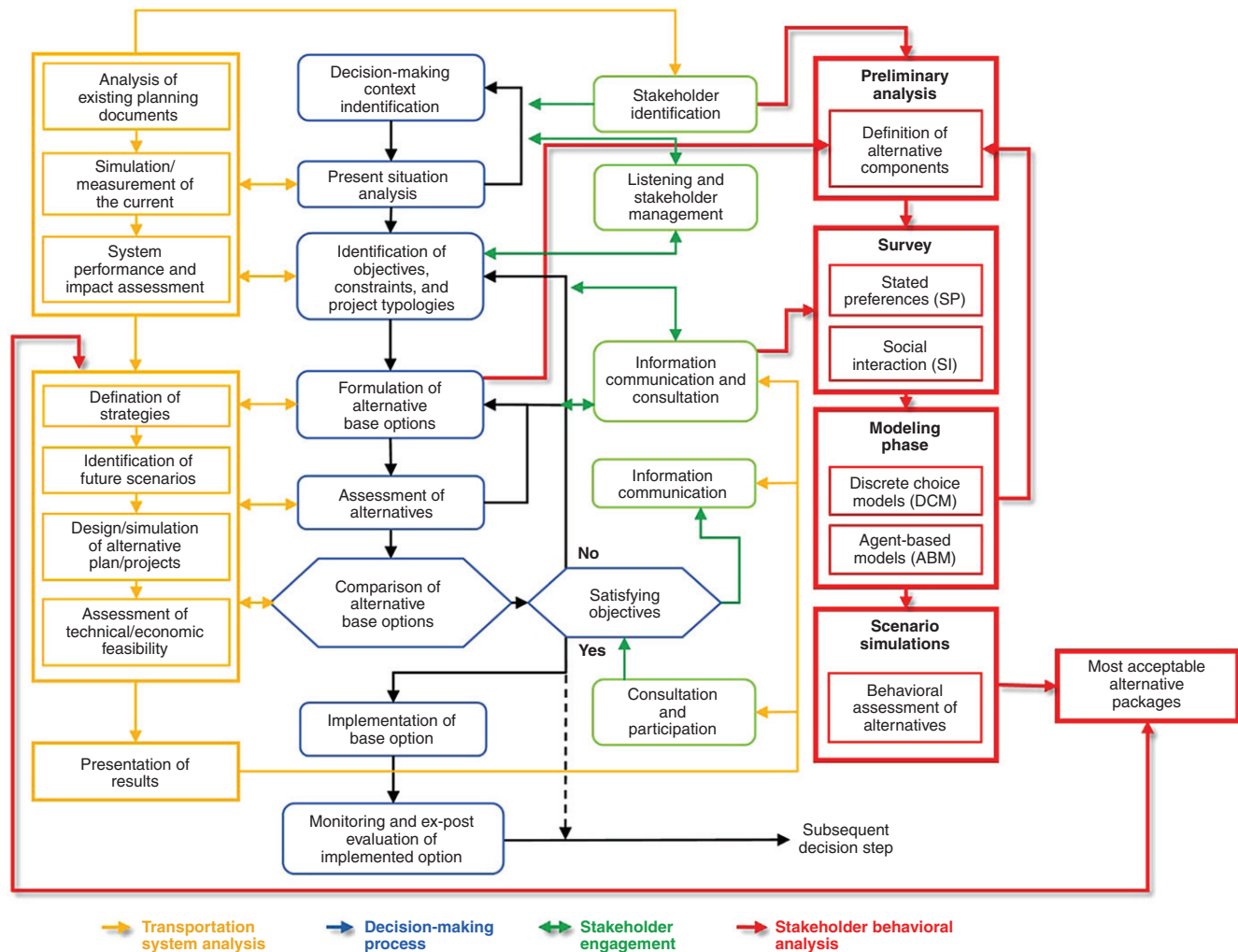


Figure 1 Decision-making model for UFT planning including stakeholder behavioral analysis. Source: *Le Pira et al., 2017*.

Stakeholder Behavioral Analysis

A combination of DCMs with ABMs has proved to be suitable for modeling stakeholder involvement in policy-making, allowing their intrinsic limits to be overcome. DCMs can provide sophisticated agent-specific data accounting for personal heterogeneity in preferences, while ABMs simulate dynamic processes and, therefore, interaction. *Marcucci et al. (2017b)* apply this integrated approach in the field of participatory UFT policy-making, by implementing an ABM based on DCM utility functions. The integrated modeling approach produces an added value for UFT policy-making. In fact, along with technical and economic analyses, the stakeholder behavioral analysis contributes both to the *ex-ante* policy assessment needed to support an effective participatory decision-making process and to help policy-makers in taking well-thought-out decisions.

Le Pira et al. (2017) adopt an integrated extension of the commonly used decision-making modeling approach, which combines: (1) a cognitive rational approach to organize the decision-making process, (2) a five-level stakeholder engagement process, and (3) quantitative analyses and methods. The stakeholder behavioral analysis can and should be considered as an additional set of activities representing the “fourth leg” of the UFT planning process (the red blocks in *Fig. 1*). The “fourth leg” introduced consists of a:

1. preliminary analysis aimed at identifying relevant UFT stakeholders and define the policy package components and elements to be used in the experimental design;
2. survey aimed at eliciting stakeholder preferences through SCEs and understanding the nature of their existing social interactions;
3. modeling and simulation phase, where DCMs are estimated and ABMs implemented accordingly, with all the acquired data and the agent-specific utility functions from DCMs; and
4. set of scenario simulations of interaction processes among stakeholders with the aim of unveiling acceptable alternative packages, using opinion dynamics mechanisms.

The technical/economic feasibility of the “most acceptable policy package” derived from the simulations has to be evaluated by quantitative methods and tools, capable of mimicking the effects of alternative transport system configurations (i.e. transportation system analysis). This should guarantee decisions that are consistent with previously identified objectives, while taking into account the heterogeneous preferences of stakeholders and behavioral aspects linked to their decision-making. This decision-making model ensures stakeholder engagement and behavioral evaluations in designing and assessing UFT solutions toward sustainability and efficiency.

Conclusion

A deep understanding of the behavioral aspects linked to stakeholder decision-making is crucial in order to define and implement effective policies, given the variety of heterogeneous UFT stakeholders with conflicting interests. Besides, behavior change forms the basis of any solution of urban freight problems. Shared and long-standing policies must account for and accommodate both private and public objectives. Therefore, urban freight policies/solutions should emerge from a collaborative/participatory approach, compatible with larger societal sustainability goals. This paper focused on the importance of stakeholder engagement and behavior change for UFT policy-making, and of an *ex-ante* behavioral assessment of UFT policies to guarantee the spreading of solutions toward efficiency and sustainability. A stakeholder behavioral analysis, based on integrated simulation/econometric models, has been presented, with the overall aim to include the heterogeneous preferences of stakeholders and their interactive behavior in the decision-making process. Any planning process oriented to sustainability should consider stakeholder-led behaviorally based evaluations as an important part of the decision-making process.

Acknowledgment

We would like to thank all those who have collaborated in the activities performed by the Transport Research Lab at Roma Tre University (TRELab—www.trelab.it) that supported most of the research that forms the basis of the present paper.

Biographies



Edoardo Marcucci is Full Professor of Transport Economics at both Molde University College and Roma Tre University, where he is codirector of *Transport Research Lab*. He is currently coeditor in Chief of Research in Transportation Economics. Being an author of several articles published in international top journals, his research interests mainly focus on urban freight distribution, stated preference, discrete choice modeling, and interaction effects in group decision-making. He has been involved in several international, national research projects, and evaluation committees.



Valerio Gatta is a Researcher of Transport Economics and codirector of *Transport Research Lab* at Roma Tre University. He has an extensive research experience in innovative transport solutions, especially in urban freight, linked to decision-making processes and sustainability, with particular reference to stated preference survey designs and discrete choice modeling techniques. Advanced methods and models for policy acceptability and behavior change analysis are at the core of his academic path.



Michela Le Pira has a PhD in transport engineering and is currently a Research fellow/lecturer specialized in Transport Planning at the University of Catania (Italy). Her PhD research focused on public participation in transport planning supported by agent-based modeling and multi-criteria decision methods. Postdoc at University of Roma Tre and at the University of Catania, she worked on participatory decision-support methods for sustainable urban freight transport policy-making. As a research fellow at the University of Catania and guest researcher at TU Delft (Netherlands), she is currently investigating new mobility services for both passengers and freight.

References

- Cascetta, E., Carteni, A., Pagliara, F., Montanino, M., 2015. A new look at planning and designing transportation systems: a decision-making model based on cognitive rationality, stakeholder engagement and quantitative methods. *Transp. Policy* 38, 27–39.
- Gatta, V., Marcucci, E., 2015. Behavioural implications of non-linear effects on urban freight transport policies: the case of retailers and transport providers in Rome. *Case Study Transp. Policy* 4 (1), 22–28.
- Gatta, V., Marcucci, E., 2016. Stakeholder-specific data acquisition and urban freight policy evaluation: evidence, implications and new suggestions. *Transp. Res. Part A* 36 (5), 585–609.
- Gatta, V., Marcucci, E., Le Pira, M., 2017. Smart urban freight planning process: integrating desk, living lab and modelling approaches in decision-making. *Eur. Transp. Res. Rev.* 9 (3), 32.
- Gatta, V., Marcucci, E., Delle Site, P., Le Pira, M., Carrocci, C.S., 2019. Planning with stakeholders: analysing alternative off-hour delivery solutions via an interactive multi-criteria approach. *Res. Transp. Econ.* 73, 53–62, doi: 10.1016/j.retrec.2018.12.004.
- Holguín-Veras, J., 2008. Necessary conditions for off-hour deliveries and the effectiveness of urban freight road pricing and alternative financial policies in competitive markets. *Transp. Res. Part A Policy Pract.* 42 (2), 392–413.
- Holguín-Veras, J., Marcucci, E., Wang, X.C., 2017a. Editorial for the special issue freight behaviour research. *Transp. Res. Part A Policy Pract.* 102, 1–2.
- Holguín-Veras, J., Wang, Q., 2013. Behavioral investigation on the factors that determine adoption of an electronic toll collection system: freight carriers. *Transp. Res. Part C Emerg. Technol.* 19 (4), 593–605.
- Holmgren, J., Davidsson, P., Persson, J.A., Ramstedt, L., 2012. TAPAS: a multi-agent-based model for simulation of transport chains. *Simul. Model. Pract. Theory* 23, 1–18.
- Le Pira, M., 2018. Transport planning with stakeholders: an agent-based modeling approach. *Int. J. Transp. Econ.* 45 (1), 15–32.
- Le Pira, M., Marcucci, E., Gatta, V., Ignaccolo, M., Inturri, G., Pluchino, A., 2017. Towards a decision-support procedure to foster stakeholder involvement and acceptability of urban freight transport policies. *Eur. Transp. Res. Rev.* 9 (4), 54.
- Louvière, J.J., Hensher, D., Swait, J., 2000. *Stated Choice Methods: Analysis and Applications*. Cambridge University Press, Cambridge, MA.
- Macal, C.M., North, M.J., 2010. Tutorial on agent-based modelling and simulation. *J. Simul.* 4, 151–162.
- Marcucci, E., Gatta, V., 2016. How good are retailers in predicting transport providers' preferences for urban freight policies? ... And vice versa?. *Transp. Res. Procedia* 12, 193–202, doi:10.1016/j.trpro.2016.02.058.
- Marcucci, E., Gatta, V., 2017. Investigating the potential for off-hour deliveries in the city of Rome: retailers' perceptions and stated reactions. *Transp. Res. Part A Policy Pract.* 102, 142–156.
- Marcucci, E., Gatta, V., Le Pira, M., 2018. Gamification design to foster stakeholder engagement and behavior change: an application to urban freight transport. *Transp. Res. Part A Policy Pract.* 118, 119–132.
- Marcucci, E., Le Pira, M., Carrocci, C.S., Gatta, V., Pieralice, E., 2017. Connected shared mobility for passengers and freight: investigating the potential of crowdshipping in urban areas. In: *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)* (pp. 839–843). IEEE, Naples, Italy.
- Marcucci, E., Le Pira, M., Gatta, V., Inturri, G., Ignaccolo, M., Pluchino, A., 2017b. Simulating participatory urban freight transport policy-making: accounting for heterogeneous stakeholders' preferences and interaction effects. *Transp. Res. Part E Logist. Transp. Rev.* 103, 69–86.
- Punel, A., Stathopoulos, A., 2017. Modeling the acceptability of crowdsourced goods deliveries: role of context and experience effects. *Transp. Res. Part E Logist. Transp. Rev.* 105, 18–38.
- Roorda, M.J., Cavalcante, R., McCabe, S., Kwan, H., 2010. A conceptual framework for agent-based modelling of logistics services. *Transp. Res. Part E Logist. Transp. Rev.* 46, 18–31.
- Simoni, M.D., Marcucci, E., Gatta, V., 2019. Potential last-mile impacts of crowdshipping services: a simulation-based evaluation. *Transportation*, doi: 10.1007/s11116-019-10028-4.
- van Duin, J.H.R., van Kolck, A., Anand, N., Tavasszy, L.A., 2012. Towards an agent-based modelling approach for the evaluation of dynamic usage of urban distribution centres. *Procedia Soc. Behav. Sci.* 39, 333–348.

Further Reading

- Gatta, V., Marcucci, E., 2014. Urban freight transport and policy changes: improving decision makers' awareness via an agent-specific approach. *Transp. Policy* 36, 248–252.
- Holguín-Veras, J., Amaya-Lea, J., Wojtowicz, J., Jaller, M., González-Calderón, C., Sánchez-Díaz, I., Wang, X., Haake, D.G., Rhodes, S.S., Frazier, R.J., Nick, M.K., Dack, J., Casinelli, L., Browne, M., 2015. Improving freight system performance in metropolitan areas: a planning guide. NCFRP Report no. 33. Washington, DC.
- Holguín-Veras, J., Amaya-Lea, J., Seruya, B., 2017b. Urban freight policymaking: the role of qualitative and quantitative research. *Transp. Policy* 56, 75–85.
- Le Pira, M., Marcucci, E., Gatta, V., Ignaccolo, M., Pluchino, A., 2017a. Integrating discrete choice models and agent-based models for ex-ante evaluation of stakeholder policy acceptability in urban freight transport. *Res. Transp. Econ.* 64, 13–25.
- Marcucci, E., Gatta, V., Scaccia, L., 2015. Urban freight, parking and pricing policies: an evaluation from a transport providers' perspective. *Transp. Res. Part A Policy Pract.* 74, 239–249.
- Taniguchi, E., Thompson, R.G., Yamada, T., 2008. Modelling the behaviour of stakeholders in city logistics. In: Taniguchi, E., Thompson, R.G. (Eds.), *Innovations in City Logistics*. Nova Science, New York, pp. 1–16.

Container (Liner) Shipping

Theo Notteboom, Center for Eurasian Maritime and Inland Logistics (CEMIL), China Institute of FTZ Supply Chain, Shanghai Maritime University, Shanghai, China; Maritime Institute, Ghent University, Ghent, Belgium; University of Antwerp, Antwerp, Belgium; and Antwerp Maritime Academy, Antwerp, Belgium

© 2021 Elsevier Ltd. All rights reserved.

Container Transport by Sea: From Infancy to a Mature Business	247
Demand-Supply Dynamics and the Financial Performance of Container Shipping Lines	247
Freight Rates and Surcharges	248
Liner Service Networks	249
The Importance of Asset Management in Container (liner) Shipping	250
Scale Increases in Vessel Size	251
Operational Strategies of Container Shipping Companies	251
Horizontal Integration in the Container Shipping Industry	252
Vertical Integration in the Container Shipping Industry	254
The Challenges Ahead	255
Biography	256
References	256
Further Reading	256

Container Transport by Sea: From Infancy to a Mature Business

The container liner shipping industry consists of shipping companies involved in the transportation of containerized goods over sea. Scheduled liner services are offered between ports using a fleet of owned and or chartered container ships. The ships carry containerized cargo in standardized unit loads of 20 foot long (i.e., 20 Foot Equivalent Unit or TEU) or in forty foot dry cargo containers (i.e., FEU). Occasionally, slightly diverging container units are also loaded on container vessels such as high cube containers, tank and open top containers and 45-foot containers. The launching of the first containership *Ideal X* by Malcolm McLean in 1956 is often considered as the beginning of containerization (Levinson, 2006). In the early years of container shipping, vessel capacity remained very limited in scale and geographical deployment, and the ships used were simply converted general cargo ships or tankers. Shipping companies and other logistics players hesitated to embrace the new technology, as it required large capital investments in ships, terminals, and inland transport. The first transatlantic container service between the US East Coast and Northern Europe in 1966 marked the start of long-distance containerized trade. Soon the containerization process expanded over maritime and inland freight transport systems (Rodrigue and Notteboom, 2009).

Container shipping developed rapidly due to the adoption of standard container sizes in the late 1960s and the awareness of industry players about the advantages and cost savings resulting from faster vessel turnaround times in ports, the reduction in the level of damage and associated insurance fees, and the integration with inland transport modes such as trucks, barges, and trains. The container and the associated maritime and inland transport systems proved to be very instrumental to the consecutive waves of globalization. Container shipping also became an essential driver in reshaping global supply chain practices allowing global sourcing strategies of multinational enterprises and the development of global production networks. New supply chain practices in turn increased the requirements on container shipping in terms of frequency, schedule reliability/integrity, global coverage of services, rate setting, and environmental performance.

The large-scale adoption of the container, in combination with the globalization process, drove the worldwide container port throughput from 36 million TEU in 1980 to 237 million TEU in 2000, 545 million TEU in 2010, and 771 million TEU in 2018 (figures Drewry). This figure includes full and empty containers. The world container traffic, that is the absolute number of loaded containers being carried by sea (excluding empties), has grown from 28.7 million TEU in 1990 to 152 million TEU in 2018 (figures UNCTAD). The ratio of containerized trade over container port throughput shows that a container on average is loaded or discharged multiple times between the first port of loading and the last port of discharge.

The center of gravity of the container business is shifting to Asia. During the last 20 years, the transatlantic container trade gradually lost its dominance to the transpacific and Europe-Far East trades. About 423 million TEU or 55% of the world's total port throughput in 2018 was handled by Asian container ports. The dominance of Asia is also reflected in the world container port rankings. In 2018 16 of the 20 busiest container ports came from Asia, mainly from China (Lloyd's List, 2019). In the mid-1980s there were only six Asian ports in the top 20, mainly Japanese load centers.

Demand-Supply Dynamics and the Financial Performance of Container Shipping Lines

Despite sustained growth brought by the containerization process, container carriers have always somewhat underperformed financially compared to other players in the logistics industries. Fig. 1 shows that since 2009, the container shipping industry has been struggling to achieve positive operating margins.

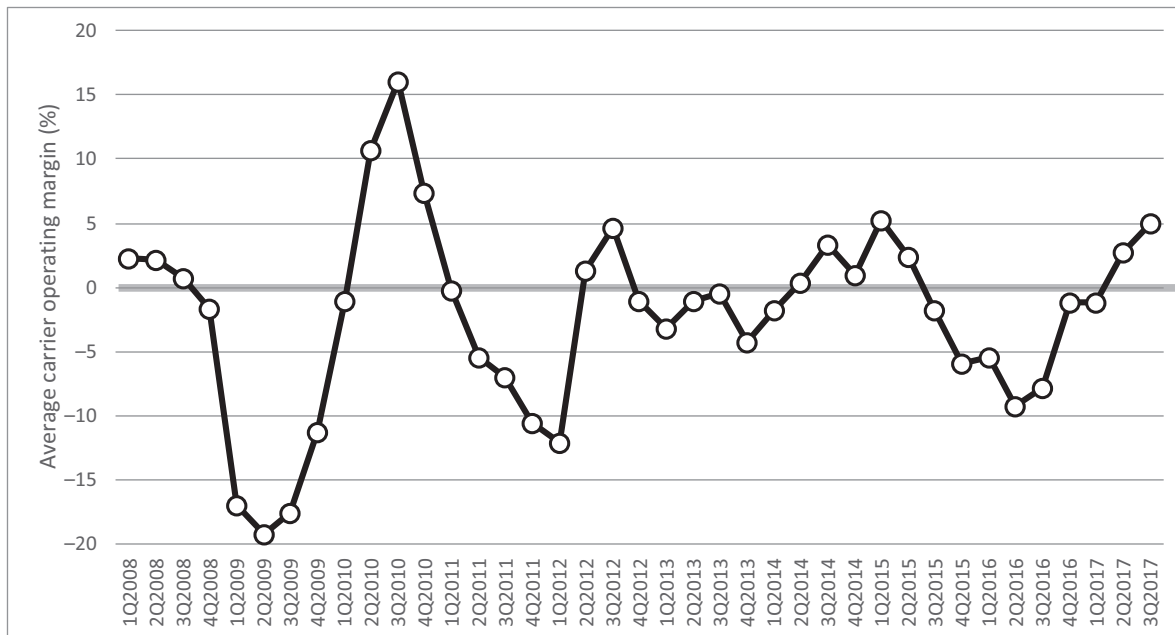


Figure 1 Average operating margin of main carriers by quarter, 2008–2017. Average of CMA CGM (including APL to 2Q2016), China Shipping (to 1Q2016), EMC, Hanjin (to 3Q 2017), Hapag-Lloyd (including CSAV to 2014), HMM, K-Line, Maersk Line, MOL, NYK, Wan Hai Lines, Yang Ming and Zim. *Source: Author's compilation based on data Alphaliner*

The weaker performance is linked to the combination of capital-intensive operations and high risks associated with the revenues. For most shipments, the freight rate only accounts for a very small portion of the shipment's total value, but as carriers cannot influence the size of the final market, they will try to increase their short run market share by reducing prices (Fusillo, 2006). As such, shipping lines may reduce freight rates without substantially affecting the underlying demand for container freight. The fairly inelastic nature of demand for shipping services constitutes the core problem behind the financial performance of container shipping lines. When the market is characterized by vessel oversupply, processes of rate erosion unfold as shipping lines try to increase their respective market shares by lowering the freight rates without substantially impacting the total demand. Such price competition continues till the freight rates stabilize at a low level, just above the "refusal rate," that is the lowest rate at which the shipping line prefers to operate its vessel, rather than having it in lay-up. If the freight rate on a specific route would fall below the refusal rate, then many vessels would be laid up. The resulting reduction in vessel capacity would push rates back up to a level above the refusal rate. When demand starts to pick up, ships are taken out of lay-up. The freight rates move away from the refusal rate once all vessels are out of lay-up and all deployable capacity is operational in the market. It is only then that rates start to increase significantly. Rates will reach their highest level when the utilization of the fleet reaches its upper limits and not enough new capacity is added to the market. At that moment, container shipping lines massively order new vessel capacity, leading to a shockwave of new capacity being brought into the market 1.5 to 2 years later (i.e., the time lag between vessel order and delivery), typically pushing the shipping market back into a period of overcapacity and lower rates. These dynamics in the shipping market, combined with the rather inelastic demand, force shipping lines into an intense concentration on costs and to seek negotiated long-term contracts with large shippers, with a view to securing cargo. Even though lower rates may allow carriers to take on extra cargo, in most cases where there remains capacity, it also reduces their profitability.

The combination of poorly differentiated rates (too many customers to negotiate a rate for every cargo) and inflexible capacity causes a pricing problem and explains existing freight rate volatility in the market. Shipping lines are not able to achieve rate stability as they cannot adjust vessel capacity to meet short run demand fluctuations (see later). High commercial and operational risks are associated with the deployment of a fixed fleet capacity (at least in the short run) within a fixed schedule between a set of ports of call at both ends of a trade route. Unused capacity cannot be stored and used later. Once the large and expensive container liner services are set up, the pressure is on to fill the ships with freight.

Freight Rates and Surcharges

The revenue base of container shipping lines consists of freight rates collected from the shippers or their representatives for the maritime transport of containerized cargo, and mostly complemented by a set of surcharges. All-in ocean freight rates have fluctuated significantly since 2009, as exemplified by the Shanghai Containerized Freight Index (Fig. 2) and other similar indices

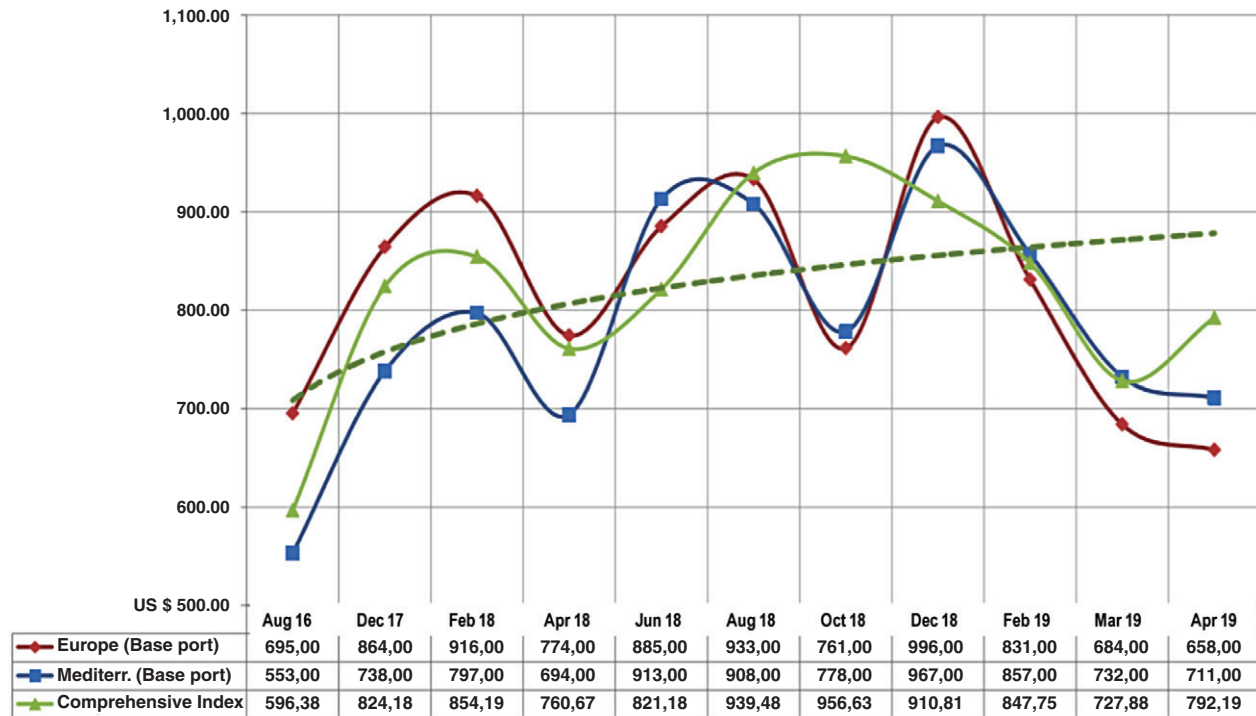


Figure 2 Shanghai Containerized Freight Index—Comprehensive Index and indices for Northern Europe and the Mediterranean, August 2016–April 2019. *Source:* Author's compilation based on data www1.chineseshipping.com.cn

such as the World Container Index or the China Containerized Freight Index). Also, contract container freight rates are influenced by the balance of power between shippers and shipping lines in rate negotiations.

The freight rates can vary greatly depending on the economic characteristics (e.g., cargo availability, imbalances, competitive situation among shipping lines) and technological characteristics (e.g., maximum allowable vessel size) of the trade route concerned. The existence of large cargo imbalances on a number of trade routes has a significant impact on the pricing problem. Trade imbalances also affect shippers in their ability to access equipment. In order to guarantee space, shippers may double book their container loads, leading to missed bookings for shipping lines. Container liner companies can react by imposing surcharges in the form of a “no show” fee.

Base freight rates or freight all kinds rates are applicable in most trades. These freight rates are lump sum rates for a container on a specific origin-destination relation irrespective of its contents and irrespective of the quantity of cargo stuffed into the box by the shipper himself. On top of these base freight rates liner companies charge separately for additional items through a range of surcharges. Still, some (larger) customers receive all-in prices. The most common surcharges include fuel surcharges (Bunker Adjustment Factor or BAF, but also other terms are used), surcharges related to the exchange rate risk (Currency Adjustment Factor or CAF), port congestion surcharges, terminal handling charges (THC) and various container-equipment related surcharges (e.g. demurrage charges, detention charges, equipment handover charges, equipment imbalance surcharge, special equipment additional for open top container, heavy lift, etc.). THC are a tariff, charged by the shipping line to the shipper and which (should) cover (part or all of) the terminal handling costs, which the shipping line pays to the terminal operator (Dynamar, 2003). The THC vary per shipping line and per country and are a negotiable item for large customers. Fuel surcharges are aimed at passing (part of) the fuel costs on to the customer through variable charges (Notteboom and Cariou, 2013). In the second half of 2018, container shipping lines started to adjust fuel surcharges on a trade-by-trade basis ahead of the deadline for the introduction of the low sulfur fuel rules of the IMO (i.e., a global sulfur cap of 0.5% on ship fuel from January 2020). For example, MSC implemented a new global fuel surcharge in January 2019 to replace the earlier bunker surcharge mechanism.

Liner Service Networks

Container shipping lines have designed liner services connecting global ports on each side of the trade routes. The networks are based on traffic circulation through a network of specific hubs. However, shipping lines do not necessarily opt for the same hubs. When designing their networks, shipping lines implicitly have to make a trade-off between the requirements of the customers and operational cost considerations. Shippers demand direct services between their preferred ports of loading and discharge. Shipping lines, however, have to design their liner services and networks in order to optimize ship utilization and benefit the most from scale

economies in vessel size. Their objective is to optimize their shipping networks by rationalizing coverage of ports, shipping routes, and transit time. Shipping lines may direct flows along paths that are optimal for the system, with the lowest cost for the entire network being achieved by indirect routing via hubs and the amalgamation of flows.

Bundling is one of the key drivers of container service network dynamics. The objective of bundling within an individual liner service is to collect container cargo by calling at various ports along the route instead of focusing on an end-to-end service. Such a line bundling service is conceived as a set of x roundtrips of y vessels each with a similar calling pattern in terms of the order of port calls and time intervals (i.e., frequency) between two consecutive port calls. By the overlay of these x roundtrips, shipping lines can offer a desired calling frequency in each of the ports of call of the loop.

Most liner services are line bundling itineraries connecting between two and five ports of call scheduled in each of the main markets. The Europe–Far East trade provides a good example. Most mainline operators and alliances running services from the Far East to North Europe stick to line bundling itineraries with direct calls in 3 to 5 ports in each of the main markets.

Liner services are combined to form complex multi-layer networks (Ducruet and Notteboom, 2015). The advantages of complex bundling are higher load factors and/or the use of larger vessels in terms of TEU capacity and/or higher frequencies and/or more destinations served. For example, shipping lines bundle container cargo by combining/linking two or more liner services in specific sea–sea transshipment hubs. These transshipment hubs have a range of common characteristics in terms of nautical accessibility, proximity to main shipping lanes and ownership, in whole or in part, by carriers or multinational terminal operators. Most of these hubs are located along the global beltway or equatorial round-the-world route (i.e. the Caribbean, Southeast and East Asia, the Middle East and the Mediterranean). These nodes multiply shipping options and improve connectivity within the network. In channeling gateway/hinterland cargo and sea–sea transshipment flows through their shipping networks, container carriers aim for control over key terminals in the network.

Existing liner shipping networks feature a great diversity in types of liner services and a great complexity in the way liner services are connected to form extensive shipping networks. For example, Maersk Line, MSC, and CMA-CGM operate truly global liner service networks, with a strong presence also on secondary routes. Notwithstanding the demand pull for global services, a number of individual carriers remain regionally based. Historically, Asian carriers mainly focused on intra-Asian trade, transpacific trade and the Europe–Far East route, partly because of their huge dependence on export flows generated by the respective Asian home bases. Thus while some carriers have clearly opted for a true global coverage, others are somewhat stuck in a triad-based service network forcing them to develop a strong focus on cost bases. Alliance structures (see later) provide its members with easy access to more liner services with relatively low-cost implications and allow them to share terminals.

The Importance of Asset Management in Container (liner) Shipping

Container shipping is a very capital-intensive industry where some assets are owned and others leased/chartered (Brooks, 2000). The total slot capacity of the cellular container fleet stood at 23.06 million TEU in August 2019 (covering 5330 fully cellular ships), 9.44 million TEU in 2007, 3.09 million TEU in 1997 and 1.22 million TEU in 1987 (figures Alphaliner). Container shipping lines on average charter about half of the vessels from third ship owners. Ship chartering is a particularly common practice for mid-size ships (i.e., 1000 to 3000 TEU). Container shipping lines also face large investments in their containers. The number of containers needed to run a regular container service is about twice the joint TEU capacity of the vessels deployed in that liner service. Container shipping lines and other transport operators typically own 55% to 60% of the total global container equipment, while leasing the remainder from specialized leasing companies.

Asset management is a key component of the operational and commercial success of container shipping lines, since they are primarily asset-based. Common asset management decisions for shipping lines include the management of the equipment (e.g., ships and boxes) to reduce downtime and operating costs, to increase the useful service life and residual value of vessels, to increase equipment safety and reduce potential liabilities and to reduce costs through better capacity management (Table 1).

Table 1 Asset management domains in container shipping

Lifecycle management of ships and boxes	Purchase/ordering	Cost management	Total cost of ownership and operation
	Deployment		Ship/box finance
	Performance measurement		Depreciation/amortization Decision on flag state (ship registry)
	Maintenance		Operational, financial and fiscal accounting
	Disposal (selling on second-hand market, scrapping)		Location and modalities of ship management
Contract management	Leases/charter parties(C/P)	Risk management	Safety/security (assets, people, IT)
	Warranties		Regulatory compliance
	Service-level agreements		
	Outsourcing of services		

Source: Author's compilation

Container shipping lines are particularly challenged to develop an effective asset management program for the fleet they own or operate. Vessel lifecycle management includes the procurement, acquisition, deployment, and disposal of container vessels. Fleet capacity management is complex given the inflexible nature of vessel capacity in the short run due to fixed timetables, the seasonality effects in the shipping business and cargo imbalances on trade routes. Lines vie for market share and capacity tends to be added as additional loops (i.e., in large chunks) to already existing services. Lines incur high fixed costs in this process. For example, 10 to 11 ships are needed to operate one regular liner service on the Europe-Far East trade with each of the postPanamax container vessels having a typical newbuilding price ranging from USD 100 to 150 million depending on the unit capacity of the ship and the market situation in the shipbuilding industry at the time the vessel order is placed.

Scale Increases in Vessel Size

The growing demand for maritime container transport has been met via vessel upscaling. Since the 1990s a great deal of attention is devoted to larger and more fuel-efficient vessels (Cullinane and Khanna, 1999). The mid-1970s brought the first ships of over 2000 TEU capacity. The Panamax vessel of 4000 to 5000 TEU (i.e. maximum dimensions for transit through the old Panama Canal locks) was introduced in the early 1990s. In 1989 APL was the first shipping line to deploy a post-Panamax vessel. Maersk Line introduced the Regina Maersk (nominal capacity of 6418 TEU but “stretchable” to about 8000 TEU) in 1996. Consecutive rounds of scale increases led to the introduction of the “Emma Maersk” in 2006, a container ship which can hold more than 15,000 TEU and measures 397 m length overall, a beam of 56 m, and a commercial draft of 15.5 m. The introduction of the Triple E class (about 18,000 TEU) by Maersk Line in 2013 was followed by even bigger vessels of competing lines such as COSCO Shipping (ships of up to 21,237 TEU), CMA CGM (up to 20,954 TEU), and OOCL (units of up to 21,413 TEU). The first ULCs of 23,656 TEU have been delivered to MSC in 2019, while evergreen placed an order for six ships of almost 24,000 TEU in October 2019. According to Clarkson’s database, there were 73 ULCs of over 18,500 TEU in operation on the Asia-Europe route in early 2019. In October 2018 the average container vessel on the Asia—North Europe trade measured 14,193 TEU compared to 11,711 TEU in 2015, 9,444 TEU in 2012, 6,164 TEU in 2006 and 4,250 TEU in 2002 (data Blue Water Reporting). The introduction of ever larger container vessels has resulted in an overall upscaling across the main east-west trade routes, with big vessels also cascading to north-south routes.

However, the focus of container carriers on larger vessels did not lead to a more stable market environment. Consecutive rounds of scale enlargements in vessel size have reduced the slot costs in container trades, but carriers have not reaped the full benefits of economies of scale at sea (Lim, 1998). The large container vessels can be deployed efficiently on the major trade lanes, provided they are full. Unpredictable business cycles and the seasonality on some of the major trade lanes (e.g., the demand peak just before Chinese New Year) have more than once resulted in unstable cargo guarantees to shipping lines. Adding postPanamax capacity gave a short-term competitive edge to the early mover, putting pressure on the followers in the market to upgrade their container fleet and to avert a serious unit cost disadvantage. A boomerang effect eventually also hurt the carrier who started the vessel scaling up round.

Operational Strategies of Container Shipping Companies

The container shipping industry goes through consecutive cycles of vessel oversupply and capacity shortages. However, the periods with vessel overcapacity are generally much longer. Overcapacity situations are often sustained as a result of a vicious circle, which takes many years to dissolve. Overall, crises and overcapacity situations in the container shipping industry are partly the result of exogenous factors such as a demand drop and partly the result of endogenous factors such as wrong investment decisions by shipping lines. Much of the overcapacity problems are linked to the failure by the key stakeholders, in the first place the ship owners and providers of ship finance to correctly anticipate the future markets for different ship types and sizes. The liner shipping community creates the cyclicalities and adds significantly to the volatility of the business environment through investment and allocation decisions.

One way for shipping lines to fight overcapacity is by postponing or canceling newbuilding orders. Shipping lines can also decide to absorb overcapacity by laying up vessels. For example, in the Autumn of 2009 during the financial-economic crisis, the worldwide laid-up fleet totaled an elevated 12% of the world container fleet (figures Alphaliner). Another measure to absorb overcapacity involves the suspension of liner services. This can involve (1) a full and indefinite termination of a liner service; (2) a temporary suspension of a liner service for a couple of months; or (2) the announcement of some blank sailings to cope with a short-lived seasonal drop in demand. The latter practice is very common around the Chinese New Year period each year, given the significant drop in cargo availability to/from China around that time.

Vessel lay-ups, order cancellations and service suspensions are not the only tools used by shipping lines in an attempt to absorb overcapacity when it occurs. Many vessels are now slow steaming at around 14 to 18 knots to save on bunker costs (Ferrari et al., 2015; Notteboom and Vernimmen, 2009). The resulting longer roundtrip times help to absorb surplus capacity in the market (i.e., more vessels needed per liner service). A service speed of 14 knots or less is considered by some as the future norm for containerships, in contrast to speeds of up to 22 to 23 knots before slow steaming was introduced. Such slow steaming practices also help to reduce emissions to meet stringent MARPOL and other regulations of the International Maritime Organization (IMO). In a reaction to slow steaming and super slow steaming, shippers might have to redesign their supply chains to meet the new reality of longer transit times.

Horizontal Integration in the Container Shipping Industry

Shipping lines are viewing market mass as one of the most effective ways in coping with a trade environment that is characterized by intense pricing pressure. Trade agreements in the form of liner conferences were very common till this type of cooperation was outlawed by the European Commission in October 2008. At present, the horizontal integration dynamics in the container shipping industry are based on mergers and acquisitions (M&A) and operational cooperation in many forms ranging from slot-chartering and vessel-sharing agreements to strategic alliances. A slot chartering agreement is a contract between partners who buy and sell a defined allocation (space, weight) on a vessel in general on a "used" or "unused" basis at an agreed price and for a minimum defined period of time. A vessel sharing agreement (VSA) is established between a limited numbers of shipping companies who agree to operate a liner service along a specified route using a specified number of vessels. A VSA is usually dedicated to a certain trade route with terms and conditions specific to that route, whereas an alliance is more global in nature and could include many different trade routes usually under the same terms.

The first strategic alliances among shipping lines date back to the mid-1990s, a period that coincided with the introduction of the first 6000+ TEU vessels on the Europe-Far East trade. In 1997 about 70% of the services on the main East-West trades were supplied by the four main strategic alliances. At the time of writing, three alliances were operational in the market: 2M, Ocean Alliance and THE Alliance. The landscape looked very different compared to 2015 when four alliances were still active: 2M, Ocean Three, CKYHE and G6 (Table 2). The alliance partnerships evolved as a result of M&A and the market entry or exit of liner shipping companies (e.g., the bankruptcy of Hanjin in 2016). In recent years, even the largest shipping companies have resorted to joining alliances for their survival and in the hope of increasing margins.

The main incentives to join an alliance relate to achieving critical mass in the scale of operation, exploring new markets, enhancing global reach, improving fleet deployment, and spreading risks associated with investments in large container vessels (Ryoo and Thanopoulou, 1999; Slack et al., 2002). Strategic alliances provide its members with easy access to more loops or services with relative low cost implications and allow them to share terminals, to cooperate in many areas at sea and ashore, thereby achieving costs savings in the end.

Table 2 The changing alliance structures in container liner shipping

<i>Q2 1996</i>	<i>Q1 1998</i>	<i>Q4 2005</i>	<i>Q 2009</i>	<i>Q2 2015</i>	<i>Q 2019</i>
<i>GLOBAL</i>	<i>NWA</i>	<i>NWA</i>	<i>NWA</i>	<i>G6</i>	<i>THE ALLIANCE</i>
<i>ALLIANCE APL</i>	<i>APL/NOL</i>	<i>APL/NOL</i>	<i>APL/NOL</i>	<i>ALLIANCE</i>	<i>ONE</i>
<i>MOL</i>	<i>MOL</i>	<i>MOL</i>	<i>MOL</i>	<i>APL/NOL</i>	<i>Yang Ming</i>
<i>Nedlloyd</i>	<i>HMM</i>	<i>HMM</i>	<i>HMM</i>	<i>MOL</i>	<i>Hapag-</i>
<i>OOCL</i>				<i>HMM</i>	<i>Lloyd/UASC</i>
<i>MISC</i>				<i>Hapag-</i>	<i>HMM</i>
				<i>Lloyd</i>	
				<i>NYK Line</i>	
				<i>OOCL</i>	
<i>GRAND</i>	<i>GRAND</i>	<i>GRAND</i>	<i>GRAND</i>		
<i>ALLIANCE</i>	<i>ALLIANCE II</i>	<i>ALLIANCE III</i>	<i>ALLIANCE IV</i>		
<i>Hapag-Lloyd</i>	<i>Hapag-Lloyd</i>	<i>Hapag-Lloyd</i>	<i>Hapag-Lloyd</i>		
<i>NYK Line</i>	<i>NYK Line</i>	<i>NYK Line</i>	<i>NYK Line</i>		
<i>NOL</i>	<i>P&O Nedlloyd</i>	<i>OOCL</i>	<i>OOCL</i>		
<i>P&OCL</i>	<i>OOCL</i>	<i>MISC</i>			
	<i>MISC</i>				
	<i>UNITED</i>	<i>CKYH</i>	<i>CKYH</i>	<i>CYKHE</i>	<i>OCEAN</i>
	<i>ALLIANCE Hanjin</i>	<i>Hanjin</i>	<i>Hanjin</i>	<i>Hanjin</i>	<i>ALLIANCE CMA</i>
	<i>Cho</i>	<i>K-Line</i>	<i>K-Line</i>	<i>K-Line</i>	<i>CGM</i>
	<i>Yang</i>	<i>Yang Ming</i>	<i>Yang Ming</i>	<i>Yang Ming</i>	<i>COSCOCS/OOCL</i>
	<i>UASC</i>	<i>COSCO</i>	<i>COSCO</i>	<i>COSCO</i>	<i>Evergreen</i>
				<i>Evergreen</i>	
	<i>KYK ALLIANCE</i>			<i>2M</i>	<i>2M</i>
	<i>K-Line</i>			<i>MSC</i>	<i>MSC</i>
	<i>Yang Ming</i>			<i>Maersk Line</i>	<i>Maersk Line (ircl</i>
	<i>COSCO</i>				<i>Hamburg Sud)</i>
<i>Maersk</i>	<i>Maersk</i>			<i>Ocean Three</i>	
<i>Sea-Land</i>	<i>Sea-Land</i>			<i>CMA CGM</i>	
				<i>China</i>	
				<i>Shipping</i>	
				<i>UASC</i>	

Source: Author's compilation

Table 3 Slot capacities of the fleets operated by the top 20 container lines (in TEU)

	January 1980		September 1995		January 2000		November 2005	
1	Sea-Land	70,000	Sea-Land	196,708	Maersk Sealand	620,324	Maersk Line	1,620,587
2	Hapag-Lloyd	41,000	Maersk	186,040	Evergreen	317,292	MSC	733,471
3	OCCL	31,400	Evergreen	181,982	P&O Nedlloyd	280,794	CMA/CGM Group	485,250
4	Maersk	25,600	COSCO	169,795	Hanjin/DSR Senator	244,636	Evergreen Group	458,490
5	NYK Line	24,000	NYK Line	137,018	MSC	224,620	Hapag Lloyd/CP Ships	413,281
6	Evergreen	23,600	Nedlloyd	119,599	NOL/APL	207,992	China Shipping	334,337
7	OOCL	22,800	Mitsui OSK Lines	118,208	COSCO	198,841	NOL/APL	331,639
8	Zim	21,100	P&OCL	98,893	NYK Line	166,208	Hanjin/Senator	315,153
9	US Line	20,900	Hanjin Shipping	92,332	CP Ships/ Americana	141,419	COSCO	311,644
10	APL	20,000	MSC	88,955	Zim	136,075	NYK Line	303,799
11	Mitsui OSK Lines	19,800	APL	81,547	Mitsui OSK Lines	132,618	OOCL	236,789
12	Farrell Lines	16,400	Zim	79,738	CMA/CGM	122,848	CSAV Group	230,699
13	NOL	14,800	K-Line	75,528	K-Line	112,884	K Line	228,612
14	Trans Freight Line	13,900	DSR-Senator	75,497	Hapag-Lloyd	102,769	Mitsui OSK Lines	220,122
15	CGM	12,700	Hapag-Lloyd	71,688	Hyundai	102,314	Zim	201,263
16	Yang Ming	12,700	NOL	63,469	OOCL	101,044	Yang Ming	185,639
17	Nedlloyd	11,700	Yang Ming	60,034	Yang Ming	93,348	Hamburg-sud	185,355
18	Columbas Line	11,200	Hyundai	59,195	China Shipping	86,335	Hyundai	148,681
19	Safmarine	11,100	OOCL	55,811	UASC	74,989	Pacific Int'l Lines	134,292
20	Ben Line	10,300	CMA	46,026	Wan Hai	70,755	Wan Hai Lines	106,505
Slop capacity top 20		435,000		2,058,063		3,538,103		7,185,608
C4-index		38.6%		35.7%		41.4%		45.9%
Share top 5 in top 20		38.6%		42.3%		47.7%		56.3%
Share top 10 in top 20		44.1%		67.5%		71.7%		73.9%
		69.1%						
	March 2007		February 2010		April 2015		April 2019	
1	Maersk Line	1,758,857	Maersk Line	2,061,607	Maersk Line	2,966,765	Maersk Line	41,115,597
2	MSC	1,081,005	MSC	1,536,244	MSC	2,541,347	MSC	3,386,293
3	CMA CGM Group	746,185	CMA CGM Group	1,042,308	CMA CGM	1,696,332	COSCO group	2,822,551
4	Evergreen Group	566,271	Evergreen Group	554,316	Hapag-Lloyd	974,024	CMA CGM Group	2,661,799
5	Hapag-Lloyd	467,030	APL	548,788	Evergreen	963,599	Hapag-Lloyd	1,700,337
6	CSCCL	417,337	Hapag-Lloyd	495,894	Cosco Container Lines	814,240	ONE	1,552,639
7	COSCO	391,527	COSCO	453,922	China Shipping (CSCCL)	703,331	Evergreen	1,248,375
8	NYK	353,832	CSCCL	438,176	Hanjin Shipping	630,215	Yang Ming	674,914
9	Hanjin/ Senator	345,037	Hanjin Shipping	428,436	MOL	595,872	HMM	436,768
10	APL	342,899	NYK	407,300	APL	546,100	PIL	405,870
11	OOCL	303,864	CSAV Group	348,746	Hamburg Sued	544,675	ZIM	305,247
12	K-Line	283,076	OOCL	342,512	OOCL	527,343	Wan Hai Lines	250,396
13	MOL	281,447	MOL	336,971	NYK Line	503,852	KMTC	167,460
14	Yang Ming Line	253,104	K Line	325,071	Yang Ming	450,468	I RI SL Group	154,415
15	CSAV Group	250,436	Zim	310,568	UASC	412,149	Antong Holdings (QASC)	148,264
16	ZIM	248,922	Yang Ming Line	308,664	K-Line	387,806	Zhonggu Logistics	137,513
17	Hamburg-Sud Group	222,907	Hamburg Sud Group	302,056	HMM	382,812	X-Press Feeders Group	127,844
18	HMM	168,966	Hyundai M.M.	283,550	PIL	370,848	SITC	108,864
19	PIL	146,174	UASC	202,099	ZIM	323,794	TS Lines	86,538
20	Wn Hai Lines	116,439	PIL	189,281	Wan Hai Lines	208,341	SM Line Corp.	76,852

(continued)

Table 3 Slot capacities of the fleets operated by the top 20 container lines (in TEU) (*cont.*)

	January 1980	September 1995	January 2000	November 2005
Slot capacity of top 20	8,745,315	10,916,509	16,543,913	20,568,536
C4-index	47.5%	47.6%	49.4%	63.1%
Share top 5 in top 20	57.6%	57.2%	60.2%	79.0%
Share top 10 in top 20	74.0%	73.3%	75.1%	92.4%

Source: Author's compilation based on data compiled from Alphaliner, ASX Alphaliner and Containerisation International.

The shipping business has also been subject to several waves of M&A. The number of acquisitions rose from three cases in 1993 to thirteen in 1998 before peaking at eighteen in 2006. The main M&A events included the merger between P&O Container Line and Nedlloyd in 1997, the merger between CMA and CGM in 1999 and the take-over by Maersk of Sea-Land in 1999 and P&O Nedlloyd in 2005. The economic crisis of late 2008 had an impact on the market structure. There was no major M&A activity in liner shipping between October 2008 and early 2014, but a new wave of acquisitions and mergers appeared inevitable in the medium term. The most recent wave in carrier consolidation started in 2014 (Crotti et al., 2019). The main recent take-overs and mergers include: the take-over of CCNI by Hamburg-Süd (2014); the merger between Hapag-Lloyd and CSAV (2014); the sale of APL container division to CMA CGM by NOL (2015); the merger between China Shipping and COSCO (2016); the merger between Hapag-Lloyd and UASC (2016); the merger of NYK line, MOL and K-Line in ONE (Ocean Network Express; 2016); the take-over of Hamburg-Süd by Maersk Line (2017) and; the take-over of OOCL by COSCO (2017).

Shipping lines opt for M&A to obtain a larger size, to secure growth, to benefit from scale advantages, and to gain instant access to markets and distribution networks. Through a series of major acquisitions (Sea-Land, P&O Nedlloyd, Safmarine, Hamburg-Süd, etc.), Maersk Line was able to maintain its position as the world's largest container shipping line and to make strategic adjustments to secure its competitive advantage on key trade routes. Fusillo (2002) argued that a large fleet capacity enabled Maersk Line to use excess capacity as a form of entry deterrence by saturating the market and reducing profit opportunities for competing carriers. In contrast to Maersk Line, MSC reached the number two position in the world ranking of container lines based on organic or internal growth. MSC was only involved in two minor take-overs such as Kenya National in 1997.

The liner shipping industry has witnessed a concentration trend in slot capacity control, mainly as a result of M&A activity. The top 20 carriers controlled 89.7% of the world's container vessel capacity in April 2019. This figure amounted to about 83% in late 2009, 56% in 1990 and only 26% in 1980 (Table 3). The top 4 companies controlled 79% of the total fleet capacity of the top 20 shipping lines in 2019, while this share was 60% in 2015 and 42% in 1995. The consolidation trend has raised concerns with respect to an overly concentrated market and the potential oligopolistic behavior of the large carriers and alliances between them. Therefore M&A activity and alliance formation in liner shipping are under scrutiny by competition authorities around the world. Alliances and carrier consolidation are having their full impact on inter-port competition given the large container volumes involved and the shift in the associated bargaining power.

Vertical Integration in the Container Shipping Industry

In response to the low margins in shipping and the demand of customers for door-to-door and one-stop shopping logistics services, shipping lines may extend the reach of their activities to other parts of the supply chain (Notteboom, 2004). Over the past decades, the largest container lines have shown a keen interest in developing dedicated terminal capacity in an effort to better control costs and operational performance, improve profitability (Cariou, 2001; Haralambides et al., 2002; Notteboom et al., 2017) and as a measure to remedy poor vessel schedule integrity (see Notteboom (2006) for a discussion on schedule unreliability). For example, Maersk Line's parent company, AP Moller-Maersk, operates a large number of container terminals through its subsidiary APM Terminals. CMA CGM, MSC, Evergreen, and Cosco are among the shipping lines fully or partly controlling terminal capacity around the world. Independent global terminal operators such as Hutchison Ports, PSA, and DP World are increasingly hedging the risks by setting up dedicated terminal joint ventures in cooperation with shipping lines and strategic alliances. Terminal operators also seek long-term contracts with shipping lines using gain sharing clauses. The above developments gave rise to a growing complexity in terminal ownership structures and partnership arrangements (Notteboom and Rodrigue, 2012).

The scope extension of a number of shipping lines goes beyond terminal operations to include inland transport and logistics. A number of shipping lines are developing door-to-door services based on the principle of carrier haulage in an attempt to get a stronger grip on the routing of inland container flows (Van Den Berg and De Langen, 2015). A number of shipping lines try to enhance network integration through structural or ad hoc coordination with independent inland transport operators and logistics service providers. They do not own inland transport equipment. Instead they attempt to use trustworthy independent inland operators' services on a (long term) contract base. Other shipping lines combine a strategy of selective investments in key supporting activities (e.g., agency services or distribution centers) with subcontracting of less critical services. With only a few exceptions (e.g., CEVA Logistics as part of CMA CGM, and Damco now fully integrated in Maersk Line), the management of pure logistics services is done by subsidiaries that share the same mother company as the shipping line, but operate independently of liner shipping

operations. A last group of shipping lines is increasingly active in the management of hinterland flows. The focus is now on the efficient synchronization of inland distribution capacities with port capacities.

Shipping lines face important challenges to further optimize inland logistics. Competition with the merchant haulage option remains fierce. The logistics requirements of customers (e.g., late bookings, peaks in equipment demand) typically lead to money-wasting peaks in inland logistics costs. Given the mounting challenges in inland logistics, shipping lines that do succeed in achieving a better management of inland logistics can secure an important cost advantage compared to rivals.

Many shipping lines are also heavily focusing on digital transformation. Through investments and initiatives in digital infrastructure and services, shipping lines are aiming for the creation of value-adding activities in the following areas:

- The optimization of operations by using data for real-time network optimization in terms of bunker savings; optimized automated vessel stowage planning; planning of container repair; and empty container repositioning; predictive maintenance of vessels and containers, etc.;
- The development of advanced commercial decision-making instruments by using data to target customers; to optimize the cargo mix within restrictions (e.g., dangerous goods); to generate transparency in the supply chains of the customers, and to develop intelligent pricing engines;
- The development of new services that can generate new revenue streams. Examples include consultancy/advisory services related to logistics chains, the aggregation and selling of trade data or the commercialization of weather data collected by vessels navigating the open seas.

Shipping lines are setting up cooperation schemes to support digital transformation processes. For example, Maersk, MSC, Hapag-Lloyd, and ONE launched a digital container platform in 2019. This Digital Container Shipping Association (DCSA) has been established with the intention of setting standards for the digitalization of container shipping to overcome the lack of a common foundation for technical interfaces and data. The neutral and nonprofit association is open to all ocean carriers who wish to join. DCSA has no intention of developing or operating any digital platform of its own. DCSA will interact with other associations and organization such as the United Nations (i.e., Rules for Electronic Data Interchange for Administration, Commerce and Transport or UN/EDIFACT), the International Organization for Standardization, the Blockchain in Transport Alliance, and OpenShipping.org, which offers an open source standard for global shipping.

The Challenges Ahead

The establishment of world-embracing container shipping networks has facilitated globalization processes and associated global production and logistics practices. Still, the financial performance of many container shipping lines remains poor. As discussed earlier, the core of the problem lies in a combination of the capital-intensive operations and high risks associated with the revenues, due to a combination of volatile markets and inflexible capacity in the short run. Shipping lines have developed a very strong focus on market shares and on the lowering of the cost base by ordering ultra large vessels and by focusing on horizontal integration. The resulting low rates and often negative margins did not make shipping lines fundamentally revise their business strategy. The increasing "commoditization" of the liner shipping market is urging shipping lines to consider other growth strategies and revenue models by focusing on supply chain integration/vertical integration and digital transformation. Thus the liner shipping industry is challenged to walk the path of breakthrough innovation in order to reinvent itself. This is a difficult task for an industry where business model changes are not drastic and tend to be implemented slowly.

Still, there are clear signs that the container shipping industry is (slowly) walking a different path. For example, Maersk Line, one of the companies that have always embodied the race for market shares and big ships is no longer pushing for bigger vessels and wants to focus on profits through a stronger focus on service quality and integrated logistics services supported by data. Also other shipping lines are developing a much stronger supply chain focus and embrace digitalization. The increased digitalization of the container market through platforms and apps will create a far greater transparency in terms of freight rates, vessel positions and port capacities, and will facilitate demand forecasting exercises. Data analytics will give customers and other actors in the supply chains much more possibilities to compare, benchmark, and play-off one shipping line against the other. These developments challenge shipping companies to become an integral part of these digital transformation processes by making their own processes open to standardization and transparency (instead of being subject to what other players impose) and to extract new business models and revenue streams from the digital environment. These transformation processes within container shipping lines are not easy, as the weak financial base of most companies does not leave a lot of room for large-scale investment programs in initiatives with a long or highly uncertain payback period.

The same applies to the enormous pressure on container lines to prepare for zero emission vessels in the foreseeable future. At present, container ships carry about 80% of global trade and contribute about 3% of the world's emissions. The IMO has developed regulations on energy efficiency to support the demand for ever greener and cleaner shipping (Cariou et al., 2019). Since 1 January 2020, the limit for sulfur in fuel oil used on board ships operating outside designated emission control areas (ECAs) is 0.5% m/m (mass by mass). There is an even stricter limit of 0.1% m/m in effect in ECAs, which have been established by the IMO. Shipping lines had to take action by implementing a range of measures, including the switch to other low-sulfur fuels (e.g., LNG, low-sulfur Marine Diesel Oil, or biofuels), by retrofitting vessels with so-called scrubbers designed to remove sulfur oxides from the ship's engine and

boiler exhaust gases, by reducing vessel speed (i.e., slow steaming), by designing fuel-efficient and low-emissions vessels, etc. The longer-term challenges are far greater with a push for zero-emission ships by 2050. It is hopeful to observe that, despite the current market conditions and poor margins, some of the world's largest container shipping companies are pledging to cut net carbon emissions to zero by 2050. This will only be feasible if the entire chain from engine makers and shipbuilders to new technology providers come up with carbon-free ships by as early as 2030 and if the final customers share the cost burden to invest in new technologies related to biofuels, hydrogen, electricity or even wind or solar power. Shipping companies are aware that not just governments and countries, but also companies and industries need to make a change.

Biography

Theo Notteboom is Professor in Port and Maritime Economics and Management. He is Director and Research Professor at Centre for Eurasian Maritime and Inland Logistics (CEMIL) of China Institute of FTZ Supply Chain of Shanghai Maritime University in China. He is Chair Professor "North Sea Port" at the Maritime Institute of Ghent University. He also is part-time Professor at University of Antwerp and the Antwerp Maritime Academy in Belgium. He is cofounder and codirector of www.porteconomics.eu, an online platform on port studies. He is past President (2010–2014) and Council Member of International Association of Maritime Economists (IAME). He is Editor, Associate Editor, or Editorial Board Member of a dozen leading academic journals in the area of maritime economics and logistics. His work is widely cited.

References

- Brooks, M., 2000. Sea change in liner shipping: regulation and managerial decision-making in a global industry. Pergamon, Oxford, UK.
- Cariou, P., 2001. Vertical integration within the logistic chain: does regulation play rational? The case for dedicated container terminals. *Trasporti Europei* 7, 37–41.
- Cariou, P., Parola, F., Notteboom, T., 2019. Towards low carbon global supply chains: a multi-trade analysis of CO₂ emission reductions in container shipping. *Int. J. Prod. Econ.* 208, 17–28.
- Crotti, D., Ferrari, C., Tei, A., 2019. Merger waves and alliance stability in container shipping. *Marit. Econ. Logist.* 1–27.
- Cullinane, K.P.B., Khanna, M., 1999. Economies of scale in large container ships. *J. Trans. Econ. Policy* 33 (2), 185–208.
- Ducruet, C., Notteboom, T., 2015. Developing liner service networks in container shipping. In: Song, D.W., Panayides, P. (Eds.), *Maritime Logistics: A Guide to Contemporary Shipping and Port Management*. Kogan Page, London, pp. 125–146.
- Dynamar, 2003. Terminal Handling Charges – A Bone of Contention, Dynamar Special Report, Alkmaar. Available from: https://www.dynamar.com/system/table_of_contents/24/original/THC%20Contents.pdf?
- Ferrari, C., Parola, F., Tei, A., 2015. Determinants of slow steaming and implications on service patterns. *Marit. Policy Manag.* 42 (7), 636–652.
- Fusillo, M., 2002. Excess capacity and entry deterrence: the case of ocean liner shipping markets. *Marit. Econ. Logist.* 5 (2), 100–115.
- Fusillo, M., 2006. Some notes on structure and stability in liner shipping. *Marit. Policy Manag.* 33 (5), 463–475.
- Haralambides, H., Cariou, P., Benacchio, M., 2002. Costs, benefits and pricing of dedicated container terminals. *Int. J. Marit. Econ.* 4 (1), 21–34.
- Levinson, M., 2006. The Box: How the Shipping Container Made the World Smaller and the World Economy Bigger. Princeton University Press, Princeton.
- Lim, S.-M., 1998. Economies of scale in container shipping. *Marit. Policy Manag.* 25, 361–373.
- Lloyd's List, 2019. One Hundred Ports 2018: The Definite Ranking of the World's Largest Container Ports. Available from: <https://lloydslist.maritimeintelligence.informa.com/one-hundred-container-ports-2018>.
- Notteboom, T., 2004. Container shipping and ports: an overview. *Rev. Net. Econ.* 3 (2), 86–106.
- Notteboom, T., 2006. The time factor in liner shipping services. *Marit. Econ. Logist.* 8 (1), 19–39.
- Notteboom, T., Cariou, P., 2013. Slow steaming in container liner shipping: is there any impact on fuel surcharge practices? *Int. J. Logist. Manag.* 24 (1), 73–86.
- Notteboom, T., Vernimmen, B., 2009. The effect of high fuel costs on liner service configuration in container shipping. *J. Trans. Geogr.* 17 (5), 325–337.
- Notteboom, T., Rodrigue, J.P., 2012. The corporate geography of global container terminal operators. *Marit. Policy Manag.* 39 (3), 249–279.
- Notteboom, T.E., Parola, F., Satta, G., Pallis, A.A., 2017. The relationship between port choice and terminal involvement of alliance members in container shipping. *J. Trans. Geogr.* 64, 158–173.
- Rodrigue, J.P., Notteboom, T., 2009. The geography of containerization: half a century of revolution, adaptation and diffusion. *GeoJournal* 74 (1), 1–5.
- Ryoo, D.K., Thanopoulou, H.A., 1999. Liner alliances in the globalization era: a strategic tool for Asian container carriers. *Marit. Policy Manag.* 26 (4), 349–367.
- Slack, B., Comtois, C., McCalla, R., 2002. Strategic alliances in the container shipping industry: a global perspective. *Marit. Policy Manag.* 29 (1), 65–76.
- Van den Berg, R., De Langen, P.W., 2015. Towards an 'inland terminal centred' value proposition. *Marit. Policy Manag.* 42 (5), 499–515.

Further Reading

- Graham, M.G., 1998. Stability and competition in intermodal container shipping: finding a balance. *Marit. Policy Manag.* 25, 129–147.
- Heaver, T., 2002. The evolving roles of shipping lines in international logistics. *Int. J. Marit. Econ.* 4, 210–230.
- Lam, J.S.L., Wong, H.N., 2018. Analysing business models of liner shipping companies. *Int. J. Ship. Transp. Logist.* 10 (2), 237–256.
- Meng, Q., Zhao, H., Wang, Y., 2019. Revenue management for container liner shipping services: critical review and future research directions. *Transport. Res. E Logist. Transport. Rev.* 128, 280–292.
- Rodrigue, J.-P., Notteboom, T., 2009. The future of containerization: perspectives from maritime and inland freight distribution. *GeoJournal* 74, 7–22.
- Slack, B., Frémont, A., 2009. Fifty years of organisational change in container shipping: regional shift and the role of family firms. *GeoJournal* 74, 23–34.
- Sys, C., 2010. Is the container liner shipping industry an oligopoly? *Trans. Policy* 16, 259–270.
- UNCTAD, 2019. Review of Maritime Transport. UNCTAD, Geneva.
- Yang, C.S., 2018. An analysis of institutional pressures, green supply chain management, and green performance in the container shipping context. *Transport. Res. D Trans. Environ.* 61, 246–260.
- Yuen, K.F., Wang, X., Ma, F., Lee, G., Li, X., 2019. Critical success factors of supply chain integration in container shipping: an application of resource-based view theory. *Marit. Policy Manag.* 1–16.

Bulk Shipping Markets: An Overview of Market Structure and Dynamics

Manolis G. Kavussanos, Stella A. Moysiadou, Athens University of Economics and Business, Athens, Greece

© 2021 Elsevier Ltd. All rights reserved.

Introduction to the Shipping Industry	257
The Main Economic Participants	257
The Main Sectors of the Shipping Industry	257
Sector Segmentation Per Vessel Size	258
The Bulk Freight Market	261
Supply of Sea Transportation	261
Demand for Sea Transportation	262
Market Equilibrium and Shipping Cycles	262
Dry Bulk Sector of the Shipping Industry	263
The Derived Demand for Dry Bulk Sea Transportation—Dry Bulk Commodity Trades	263
Iron Ore Seaborne Trade	263
Coal Seaborne Trade	264
Grains Seaborne Trade	266
Supply of Dry Bulk Sea Transportation—Fleet, Freight Rates, and Market Integration	268
The Wet Bulk (Tanker) Sector	270
The Derived Demand for Wet Bulk Sea Transportation—Commodity Trades	270
Crude Oil Seaborne Trade	270
Oil Products Seaborne Trade	274
Supply of Wet Bulk Sea Transportation—The Tanker Fleet Throughout Market Cycles	275
Summary	278
Relevant Websites	278
References	278
Further Reading	279

Introduction to the Shipping Industry

The Main Economic Participants

More than 90% of world trade is carried by sea. Vessels are required for this purpose, for offering the sea freight transportation service to those lacking it. Thus the shipping industry provides a service, not a product; that is, the transportation of commodities (or people) from point A to B in the world. This service is measured in ton-miles, that is, the quantity of tonnage transported times the distance.

A large number of economic participants, maritime stakeholders, are involved in the sea transportation service, and they collectively constitute the maritime industry. They include shipowners, charterers, ship managers, freight forwarders, cargo owners, brokers, classification societies, seafarers, ports, shipyards, bunker suppliers, scrapyards, repair and maintenance yards, equipment manufacturers, marine insurers, maritime flag states, regulators, lawyers specializing in maritime law, financiers providing capital for the needs of the industry, and others.

The degree of integration between the functions of economic participants differs between the various shipping sectors, which are presented in section, *The Main Sectors of the Shipping Industry*. Shipping companies' economic activities are structured into entities, with their legal nature ranging from single purpose companies (or single purpose vehicles, SPVs)—owning a single vessel, to family businesses—owning a number of ships, to fully integrated shipping divisions or large corporate structures, which may be private or listed in one of the major stock exchanges around the world. Their economic activity is subject to the various shipping markets—the freight, the newbuilding, the sales and purchase, the demolition markets, the bunker, the money and capital, and the risk management markets, amongst others.

The Main Sectors of the Shipping Industry

Merchant shipping refers to the transportation of commodities, which are carried as bulk, break-bulk/general or containerized cargoes.

Bulk shipping involves large volumes of homogeneous commodities transported around the world in bulk, mainly on a “one ship, one cargo” basis. The cargo transported can be dry or liquid. Bulk carriers are used for the transportation of dry commodities in bulk, and tankers for the transportation of liquid ones, while combined carriers can carry either dry or liquid cargoes. There are also

specialist vessels, particularly designed to accommodate the transportation of commodities with special handling or storage requirements, such as the transportation of gas under pressure, refrigerated products, etc.

There are also cargoes which do not justify the use of the whole ship because of their consignment size. These are called “general cargoes” and include machinery, heavy cargoes, sling cargoes, pallets, drums, etc. Two broad groups are distinguished: (1) conventional or break-bulk cargoes, involving transportation of individual items (e.g., pieces of machinery, bags, cases, bales, drums, barrels, etc.; and (2) unitized cargoes, which involve containers, pallets, slings (e.g., planks of wood lashed together), etc. Multipurpose and specialized vessels, as well as containerships, are used for the transportation of these cargoes. Thus containerships carry intermodal containers, which are shipping containers of standardized dimensions that can be carried across different modes of transport (rail, truck, and ship) and handled via automated transport system technologies. Cargoes that are unitized or impossible/ineffective to be transported in bulk are carried in intermodal containers.

Thus the decision of what type of vessel will be used for the transportation of a commodity depends (among other factors) on the particular characteristics of the commodity, as well as its consignment size. For any commodity, there are typical parcel (consignment) sizes for transportation, which appear over a range—the so called “Parcel Size Distribution.” This range is determined by the economics of the industry demanding the commodity as well as its physical and economic characteristics^a.

The different commodities and vessels used for their transportation define the different sectors of merchant shipping, which can be summarized as follows:

- The dry bulk sector, which involves bulk carriers for the transportation of dry cargo in bulk form, such as iron ore, coal, grains, bauxite and alumina, and other minor bulk commodities.
- The sector of general cargo vessels, which carry other dry cargo that is not shipped in bulk or in not large enough sizes to justify the use of bulk cargo transportation.
- The oil and product tanker sectors, where crude oil and oil product tankers are used for the transportation of crude oil and refined products.
- The container sector, which involves the transportation of finished products, such as manufactured commodities, electronic goods etc. in containerships.
- The offshore sector, with vessels involved in the supply of oil extracting platforms.
- Several other types of vessels serve the different sectors of merchant shipping, like combined carriers, specialized vessels, tugs, crew/workboat fleet, etc.

As at the start of July 2019, the world fleet comprised 97,142 vessels with a total carrying capacity of 2,020.71 million deadweight (DWT)^{b,c}. As can be observed in **Fig. 1**, bulk carriers constituted the largest part of this and amounted to 43% of the world total carrying capacity (857.78 million dwt, with 11,643 vessels), tankers comprised the second biggest component of the world fleet with 30% share (609.83 million dwt, 6,873 vessels) while containerships came third with a 13% share (269.92 million dwt, 5,311 vessels). LNG (Liquefied Natural Gas) and LPG (Liquefied Petroleum Gas) carriers amounted to 3% of the world total fleet, while other types of vessels—like combined carriers, offshore vessels, anchor handling tugs, cruise ships, crew/workboat fleet, etc.—comprised 11% of the world total carrying capacity.

Fig. 2 shows the development of the world fleet in terms of DWT capacity for the years 1970–2018. The increase in carrying capacity, from 57.7 to 817.5 million DWT for bulk carriers, and from 138.8 to 582 million DWT for tankers, is impressive, and is a consequence of the globalization process in production and services and the consequent need for sea-transportation to serve its needs. This continuously growing need for transportation of raw industrial materials and fuels over time includes commodities such as iron ore, coal, crude oil, and its products, which are indispensable in industrial production. As observed, the world total carrying capacity has more than doubled from 2005 (921.7 million dwt) to 2019 (1976.5 million dwt), with carrying capacity increasing in all segments.

Up to 2003, tankers had the greatest share in the world fleet, followed by bulk carriers. This has been reversed since 2004, partially due to switches from long-haul to short-haul routes in the transportation of oil, which led to a decrease in the average tanker size (see section *Sector Segmentation Per Vessel Size*), but also due to the over-ordering of dry bulk vessels during the booming years of 2003–07.

Sector Segmentation Per Vessel Size

The shipping industry aims to provide fast and cheap transportation service around the globe. In comparison to all alternative modes of cargo transportation, particularly over large consignments and/or longer distances, waterborne transportation is the most economical mode on a per-unit of cargo basis. For bulk vessels, the unit cost of transporting a ton of cargo is calculated as the sum of the capital, operating and voyage expenses, divided by the tonnage the ship can carry. Since capital, operating and voyage costs do not increase proportionally with tonnage capacity, the unit cost generally falls as the size of the ship increases. In real terms,

^a For instance, iron ore is transported in large parcels of around 180,000 tons, as the steel industry which is using it as a raw material requires such large consignments, in order to make it cheaper for them to be transported. Also, it does not require special storage facilities—open land next to the steel mill is sufficient for storage as weather does not affect it, while the commodity can be handled in bulk. As a consequence, it will be transported on a one ship one cargo basis in relatively large consignments, as this will reduce the unit (per ton) transportation cost. PSD's vary over time as the world economy, trade patterns, technological developments in shipping and commodity economics change.

^b Source: Clarksons SIN (Shipping Intelligence Network).

^c Deadweight is a measure of the carrying capacity of a vessel, calculated as the sum of the weights of cargo, fuel, fresh water, stores, and crew that it can carry.

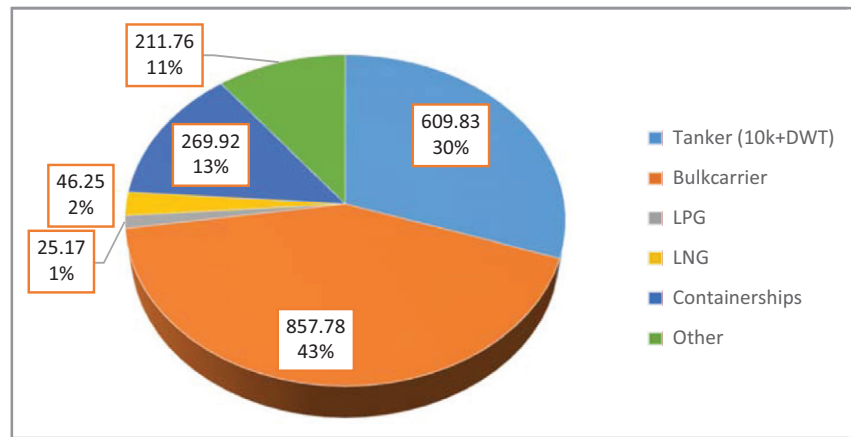


Figure 1 World fleet per sector (July 2019, measured in million dwt and as percentage of total world fleet). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

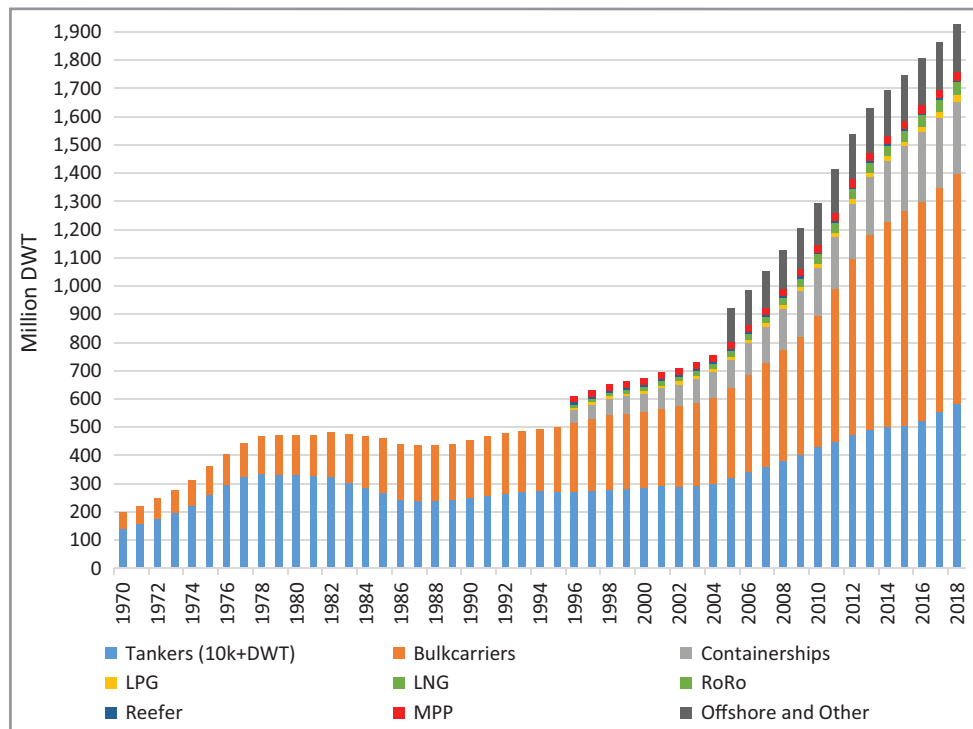


Figure 2 World fleet development (in million DWT, 1970–2018). Note: For the years 1970–95, Tanker and Bulk Carrier fleets only are reported due to data unavailability for the rest of the sectors. Similarly, for all subsequent years, data are depicted depending on its availability. *Source: Figure created by the authors. Data derived from Clarksons SIN.*

the cost per ton of transportation has been reducing continuously since the Second World War, due to increases in the size of the vessels carrying the cargoes and thus economies of scale setting in. As shown in Fig. 3, from January 1970 to January 2019, the average bulk carrier increased in size from 28,095 dwt to 73,427 dwt.

However, since demand for transportation is measured in ton-miles, there may be cases where changes in trading routes (for example from long-haul to short-haul ones) may result in a decrease in vessel sizes. As can be seen in Fig. 3, even though the average tanker vessel doubled in size from January 1970 to January 2019, with the greatest increase observed during the 1970s (from 43,714 dwt in January 1970, to 100,434 dwt in January 1980), thereafter there has been a decrease due to switches from long-haul to short-haul routes in the transportation of oil.

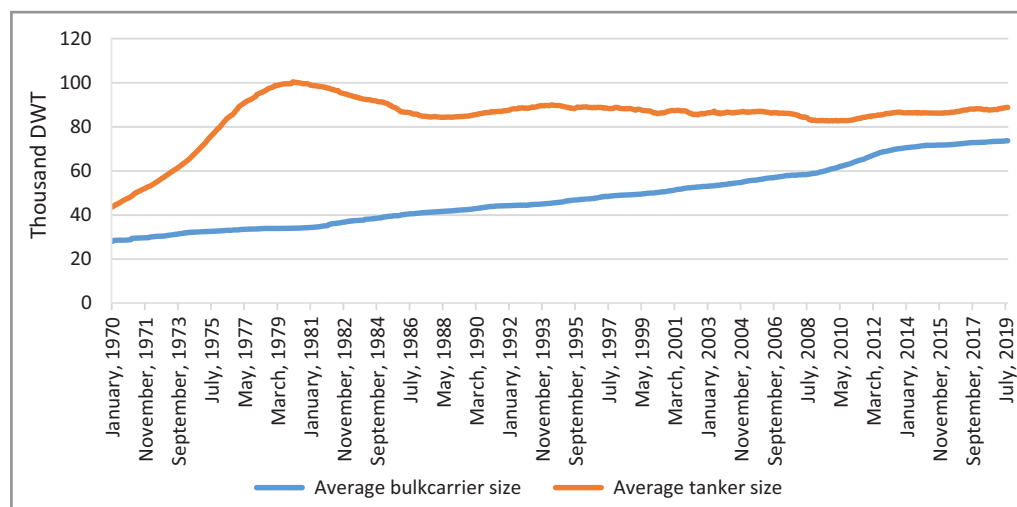


Figure 3 Development of the average bulkcarrier and tanker sizes overtime (1970–2019, in thousand dwt). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

Table 1 Dry bulk sector segmentation per vessel deadweight capacity.

Segment/Vessel type	Vessel DWT capacity
Very Large Ore/Bulk Carrier (VLOC/VLBC)	200,000–400,000 dwt
Valemax ^a	400,000 dwt
Capesize	120,000–200,000 dwt
Newcastlemax ^b	Approx. 185,000 dwt
New Panamax	100,000–120,000 dwt
Panamax	65,000–100,000 dwt
Kamsarmax ^c	Approx. 82,000 dwt
Supramax	50,000–61,000 dwt
Handymax	40,000–65,000 dwt
Handysize	10,000–40,000 dwt

^aSubcategory named according to the cargo owner/vessel charterer: Valemax vessels comprise a fleet of very large ore carriers -VLOC-owned or chartered by the Brazilian mining company Vale S.A. to carry iron ore from Brazil to European and Asian ports.

^bSubcategory named according to infrastructural particularity: Newcastlemax is the largest vessel able to enter the port of Newcastle in Australia.

^cSubcategory named according to infrastructural particularity: Kamsarmax got its name by meeting the 229-meter limit on ship length at the Port of Kamsar, a major bauxite shipping port in the Republic of Guinea on the west coast of Africa.

Source: Created by the authors using various sources.

Another feature of water transportation economics and particularly those of bulk shipping freight markets is that shipping freight rates represent a small proportion of the final value of the transported product. As a consequence, a large change in freight rates will not result in a substantial move in the price of the commodity being transported, and as a result on the final demand for the commodity, rendering demand for sea transportation inelastic.

The dry bulk shipping sector comprises the vessel categories per deadweight capacity presented in Table 1.

An important differentiation between vessel sizes involves the availability of cargo-handling equipment on the vessel. Typically, vessels below Panamax size are geared, and thus are self-sufficient in terms of being able to approach even ports of the world with less developed cargo-handling facilities.

In the tanker sector, one can distinguish between the following on the basis of the type of cargo they carry:

- crude oil or dirty tanker carriers,
- crude oil-product or clean tanker carriers, and
- gas carriers.

Table 2 Tanker sector segmentation per vessel deadweight capacity.

	Segment/Vessel type	Vessel DWT capacity	
Crude oil	Ultra large crude carrier (ULCC)	> 320,000 dwt	
	Very large crude carrier (VLCC)	200,000–300,000 dwt	
	Suezmax	115,000–200,000 dwt	
	Aframax	70,000–115,000 dwt	
	Panamax	50,000–70,000 dwt	
	Handysize	10,000–50,000 dwt	
Crude oil products	Long range 2 (LR2)	80,000–160,000 dwt	
	Long range 1 (LR1)	55,000–80,000 dwt	
	Mid range (MR)	25,000–55,000 dwt	
Gas	Liquefied natural gas (LNG)	Differentiation by volume (or cabin) capacity:	
	Liquefied petroleum gas (LPG)	Q-Max	250,000–300,000 m ³
		Q-Flex	200,000–250,000 m ³
		Standard type	100,000–200,000 m ³
		Small	<100,000 m ³
	Ethylene and other gas carriers	Differentiated by boiling point of the gas	
			Differentiation by tank (or cabin) type: Thin film cabin type Spherical Cabin type

Source: Created by the authors using various sources.

The various subsegments of the tanker sector are presented in **Table 2**. As can be observed, the loading capacity (deadweight) of subsegments may be overlapping.

The Bulk Freight Market

Freight rates in tramp bulk shipping are determined by the equilibrium (intersection) of supply and demand for freight services. In the maritime economics literature, the bulk shipping freight market is usually considered perfectly competitive (Norman, 1981), with freight rates constantly changing, according to the balance of demand with supply. There are thousands of vessels and hundreds of shipowners, offering the “homogeneous” service of sea transportation. Each shipping company only holds a relatively small market share and is thus considered a “price taker”—that is, no individual firm has a substantial proportion of the market, thus enabling it to affect freight rates. On the demand side, there are many buyers of the freight service, competing to hire the most economical vessel for cargo transportation. Shipbrokers match supply and demand and create a transparent market where information is efficiently disseminated. Even though large amounts of capital are required to enter the business, there are no other entry/exit barriers for new/already existing market participants.

In order to understand equilibrium freight rates, one needs to understand the supply and demand functions for freight services, their relative positions and how these evolve over time. The answer to these questions depends on the factors that affect supply and demand.

Supply of Sea Transportation

The supply component of sea transportation corresponds to the cargo carrying capacity of the fleet. In contrast to demand, supply is determined by the investment decisions of the independent market agents (the shipowners) involved in the maritime sector and therefore, is endogenous to the industry. Shipping supply increases through the ordering of new vessels, and decreases through demolition of existing ones.^d The delivery of a newbuilding order requires a time-to-build that can vary from 18 months upwards and depends on the prevailing market conditions, as well as shipyard capacity. Hence, in the short-term, shipping supply can be inelastic and adjusts sluggishly to changes in demand (Greenwood and Hanson, 2015; Kalouptsi, 2014). In this case, freight rates may increase rapidly in response to demand increases. In the short-run, supply can only increase either by taking ships out of lay-up, or by increasing the fleet productivity; the latter can be achieved by increasing the speed of vessels in operation, or shortening the time spent in ports. However, if demand is too strong, supply cannot adjust fully in the short run and freight rates skyrocket. In the long run, supply is not sticky and can adjust to the level of demand through the delivery of newly built vessels.

The response of freight rates to changes in demand depends significantly on the state of the freight market. In bear markets, when freight rates are low and expectations are dismal, part of the vessels in the fleet which are typically older, cannot cover their operating expenses (OPEX) and as a consequence are laid up. At the same time the (younger) fleet at sea is slow steaming to save on (voyage) costs. Thus an increase in demand is absorbed by vessels being brought back from lay-up and into service and by some of the slow-steaming ships gradually increasing their speeds. As a consequence, the impact of the increased demand on freight rates in this elastic segment of the supply curve is not significant, as opposed to its significant impact on the inelastic part of the supply curve described earlier.

^d Since supply (and demand) for freight services is measured in ton-miles, changes in the average speed of vessels will also affect supply in the market.

Owning and operating a fleet entails high operating and financial risks. Shipowners invest large amounts of capital in mobile assets (the vessels) which travel on the Earth's oceans, under high levels of operational risk, which includes adverse weather conditions, ship accidents, piracy, war risks, etc. To finance their investments, they usually rely on foreign capital, which include bank loans, bond issues, leasing, etc. Excess volatility of freight rates and vessel prices will increase their default risk; that is, the probability to be unable to honor the promised payments emanating from contractual agreements (repayment of bank loan, fulfillment of charter rates, bunkering, etc.). A ship can continue its operation only if freight earnings can cover OPEX and voyage costs. Otherwise, operating losses can drive the shipowner to put the ship into lay-up or even scrap it during prolonged periods of depressed freight markets, particularly for older, less efficient vessels.

Demand for Sea Transportation

Demand for sea transportation derives from demand for the goods being carried. For instance, demand for transportation of grains derives from demand for grains itself, which in turn depends on factors such as demographics, economic growth, changes in dietary and consumption patterns, etc. Equivalently, demand for iron ore comes from steel mills, which produce steel that is utilized in the construction of infrastructure projects. The latter, in turn, depends on industrial production and the growth of mainly the developed economies and the BRICS^e countries of the World.

Overall, world macroeconomic factors should be considered in order to understand demand for international seaborne trade. They include the business cycle of the world economy, which in turn is affected by factors such as the multiplier, time lags in decisions, stock-building, mass psychology, and random shocks. Trade elasticity of the world economy, that is, the percentage change in sea trade over the percentage change in industrial production, is a significant factor. This changes over time and depends also on the changing trade patterns in the world economy.

Commodity trade patterns are influenced by developments in relevant industries, for example, oil, gas, coal, agriculture, etc. Such developments may include changes in production technologies, the relocation of processing units or refineries^f, discoveries of new resources and/or depletion of others^g, changes in relative prices between commodities or (for the same commodity) between regions, concerns about security of the cargo transported^h, etc.

Seaborne trade patterns are also affected by seasonality patterns in relevant industries (Kavussanos and Alizadeh, 2001, 2002). For instance, more oil shipments are typically demanded in autumn and early winter, due to the need for heating. In metallurgical industries, where iron ore, coking coal and bauxite/alumina are transported, there are seasonal increases in demand in February and March, mostly in Japan, due to the financial year change. In agricultural industries, grain cargoes increase during the harvest seasons (spring and autumn).

In bulk shipping, demand is considered to be inelastic in the short term. The underlying reason is that the freight rate payable represents only a small part of the final price of the good transported, so only if a substantial change in the freight rate occurs, will demand for sea transportation possibly be affected. In addition, for most commodities transported in bulk, there is no economically viable alternative to sea transportation.

Market Equilibrium and Shipping Cycles

In bulk shipping, three types of equilibrium can be distinguished:

- Long-run equilibrium, which emanates from the long-run relationship between supply and demand, driven by shipbuilding and scrapping (demolition) and the changing long term trends in commodity trade patterns.
- Short run equilibrium, resulting from the interaction between global supply and demand. The time frame can be between a month and several months, a time frame that enables ships to move from one part of the world to another according to where the demand is emanating from. Over this time frame, factors such as lay-up rates, speed, bunker prices, seasonal trades, and fleet productivity are important.
- Momentary Equilibrium, which comes from the interaction between regional supply and demand. The time-frame in this case is less than a week, and concerns cargoes and vessels in the vicinity of one-week's travel to where the cargoes appear.

Three types of shipping cycles in freight rates and vessel values can be identified (Stopford, 2009): (1) the long-term cycle, which lasts for long periods of time (e.g., 60 yearsⁱ) and is driven by major changes in the economy, technological developments or regional factors; (2) the short-term cycle (also referred to as "business cycle"), which lasts from 5 to 10 years; (3) the seasonal cycle, which is observed within the year, prompted by the seasonality in demand for the sea transportation of goods (Kavussanos and Alizadeh, 2001, 2002).

^e BRICS stands for Brazil, Russia, India, China and South Africa.

^f For instance, oil refineries built near the market place rather than in oil producing countries are switching fleet requirements from product carriers to large crude carriers.

^g For instance, in the 1960s, crude oil was imported from the Middle East to the West. The discovery of Alaskan and North Sea oil resources reduced average haul; these areas being closer to Unites States and Europe, where demand comes from.

^h Disturbances to strategic sea-lanes impact average haul. For example, the closure of the Suez Canal increases distance between the Middle East and Europe from 6,000 to 11,000 miles, leading to a freight market boom. Other important passages include the Panama Canal, the Straits of Hormuz, the Bosphorus, the Straits of Malacca, the Straits of Gibraltar and the British–French Channel.

ⁱ Today, with the fast changing pace of technology, this would be shorter.

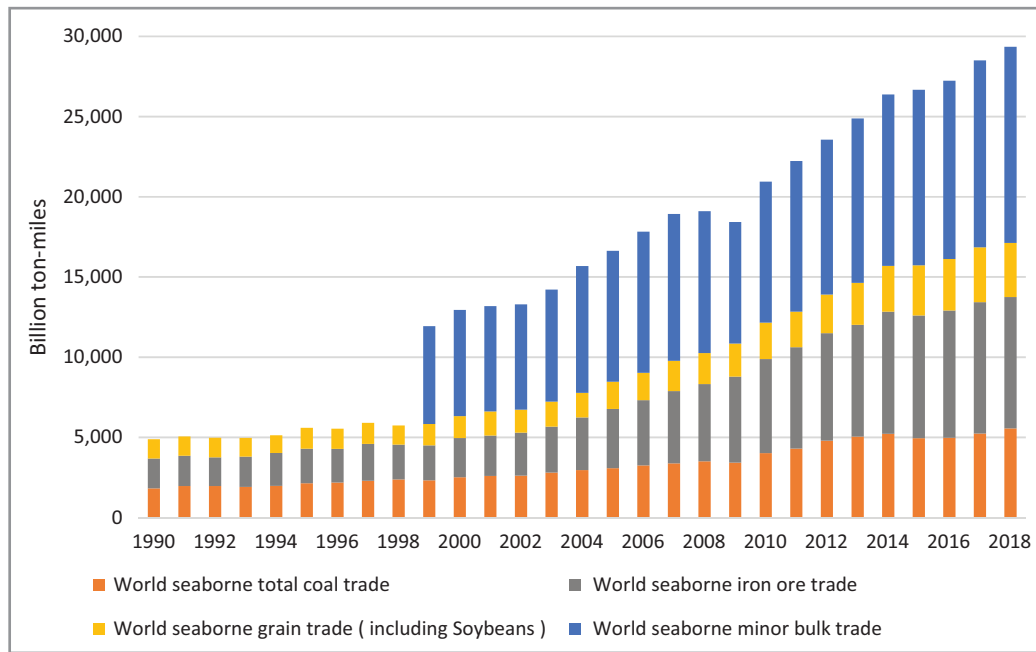


Figure 4 World fleet seaborne major and minor dry bulk trade development (billion ton-miles, 1990–2018). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

Shipping is a highly cyclical business, with shipping cycles lagging behind the wider economic cycles. The short-term cycle is the one most commonly referred to as the “shipping cycle.” The stages of a shipping cycle include the market trough, the recovery, the peak and the collapse. During a cycle, integration between the shipping markets—newbuilding, sales and purchase, demolition and freight—is what reinstates equilibrium in the industry.

The sections that follow provide an insight into the specifics of supply and demand of the bulk sectors of the shipping industry, as well as their interaction through the various shipping markets.

Dry Bulk Sector of the Shipping Industry

The Derived Demand for Dry Bulk Sea Transportation—Dry Bulk Commodity Trades

The main commodities transported by ocean going dry bulk vessels can be classified into:

- the major bulks, which are iron ore, coal, and grains;
- the minor bulks, which include: bauxite, alumina, phosphates, steel, steel products, cement, sugar, rice, gypsum, nonferrous metal ores, salt, sulfur, forest products, wood chips, cotton, chemicals, fertilizers, cement, tapioca, jute, wine, coffee, and others;
- Specialist bulk cargoes, which have specific handling or storage requirements and include heavy lift, cars, timber, refrigerated cargo, etc.

The iron ore and coal trades are the two most important drivers of demand for dry bulk sea transportation. The coal trades include coking and thermal coal. Coking coal is used in metallurgical processes (e.g., in iron smelting to make steel), while thermal (or steam) coal is used for energy production (e.g., in generating electricity via steam). Iron ore and coking coal are the main components in the production of steel, which is massively used in construction, manufacturing and many other industries worldwide.

The grains trade is mostly regional, thus not driving major changes in the dry bulk sector. It depends on world population, feedstock, and human consumption. Minor bulk commodities are mostly used in construction and agriculture; thus their trade is mainly driven by industrial growth and world GDP growth. **Fig. 4** shows the World Seaborne Major and Minor Dry Bulk Trade development over the time period 1990–2018.

Iron Ore Seaborne Trade

Iron ore is mainly exported from Australia and Brazil (see **Fig. 5**). Their exports over the years 2000–10 reached approximately 70% of the world total seaborne iron ore trade. After 2010, given the increased exports from Australia, the two countries’ export shares jointly surpassed 80% of the world total seaborne iron ore trade. India had been the third main exporter until 2010, but a year later

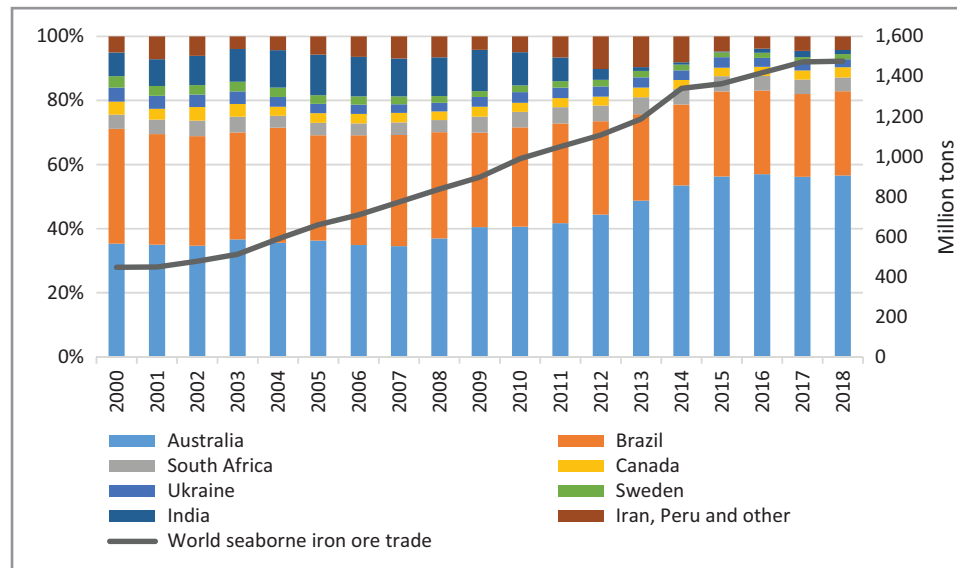


Figure 5 World seaborne iron ore exports (million tons, 2000–18). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

the Indian government started implementing a series of mining restrictions and iron ore export tariffs, in an attempt to create a framework to protect India's natural resources. From 2012 onwards, the third biggest iron ore exporter has been South Africa, followed by Canada.

The world seaborne iron ore trade increased by 230% during the period 2000–18. The key driver of this growth has been the increased demand for iron ore imports into China for the implementation of vast infrastructure and urbanization projects, as part of the Chinese policy for industrial growth. Chinese demand for high quality iron ore imports from Brazil, Australia, South Africa and Canada boosted demand, not only in terms of tons of iron ore transported, but also in terms of ton-miles, strengthening total demand in the iron-ore trades of the dry bulk sector—more iron ore cargo had to be transported over longer distances. From 2008 onwards, several countries implemented financial packages to stimulate infrastructure, one of them being China's CNY 4 trillion package. As can be seen in [Fig. 6](#), from 2009 onwards, around 90% of the world seaborne iron ore trade was transported into Asia, while approximately 70% of it specifically into China.

Coal Seaborne Trade

Two types of coal may be distinguished, based on their use: (1) metallurgical coal or coking coal and (2) thermal coal. They account respectively for approximately 25% and 75% of total world coal seaborne trade.

Coking coal is used in the process of creating coke. Coke, in turn, is used together with iron ore to produce iron and steel. Coke is a porous, hard black rock of concentrated carbon that is created by heating bituminous coal without air to extremely high temperatures.

Even though world seaborne coking coal trade has increased from 2000 to 2018, its increase is spikier and more volatile than the respective iron ore trade. China, the industrial growth of which has driven dry bulk trades during the last 15 years, has the third largest coal natural reserves in the world and is able to satisfy demand domestically. Only in 2009 did it start importing considerable amounts of coking coal, as seen in [Fig. 7](#), affecting the global coal trade balance and becoming the fourth largest coking coal importer after Japan, Europe and India. Moreover, China's coking coal imports derive mainly from price arbitrage practices, making them harder to predict. The major exporters of coking coal are Australia, Canada and the United States. As seen in [Fig. 8](#), Australia has been exporting more than 50% of the world seaborne coking coal over most of the years from 2000 to 2018, followed by the United States and Canada.

The second type of coal transported by sea is thermal coal or steam coal. This is ground to become powder, then fired into a boiler and burned for steam to run turbines which themselves generate electricity. Electricity is used either by electricity producing companies or directly by industry requiring electrical power. Examples of the latter include chemical industries, paper manufacturers, the cement industry, and brickworks manufacturing. Steam coal seaborne trade almost tripled during the years 2000–18, from 336 to 996 million tons. China and India increased considerably their imports from 2009 onwards, driving a surge in demand for steam coal and Indonesian exports. The main exporters are Australia and Indonesia. While South Africa used to be the third

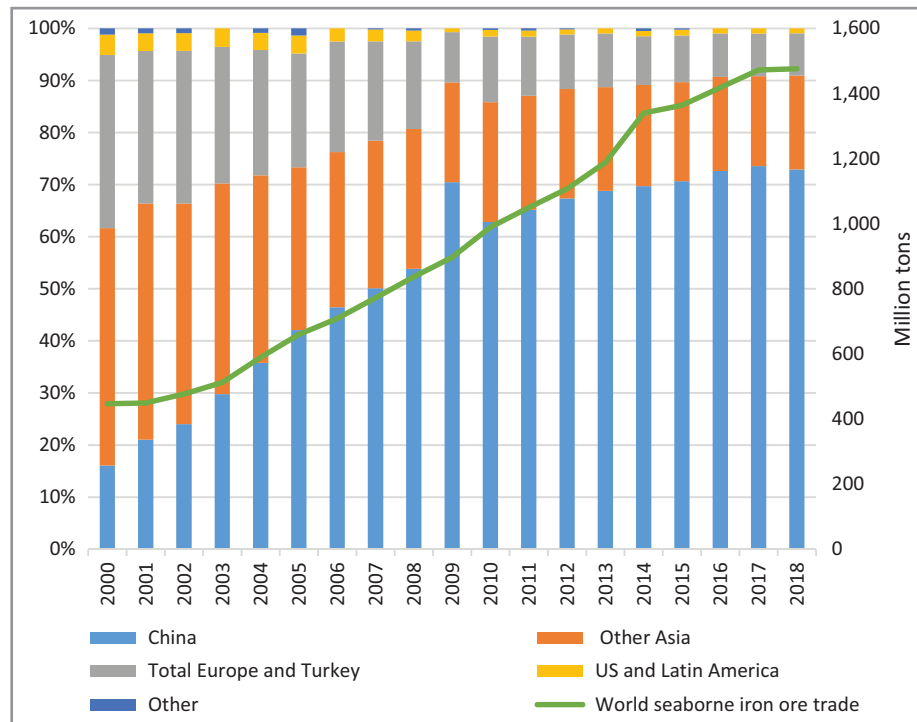


Figure 6 World seaborne iron ore imports (million tons, 2000–18). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

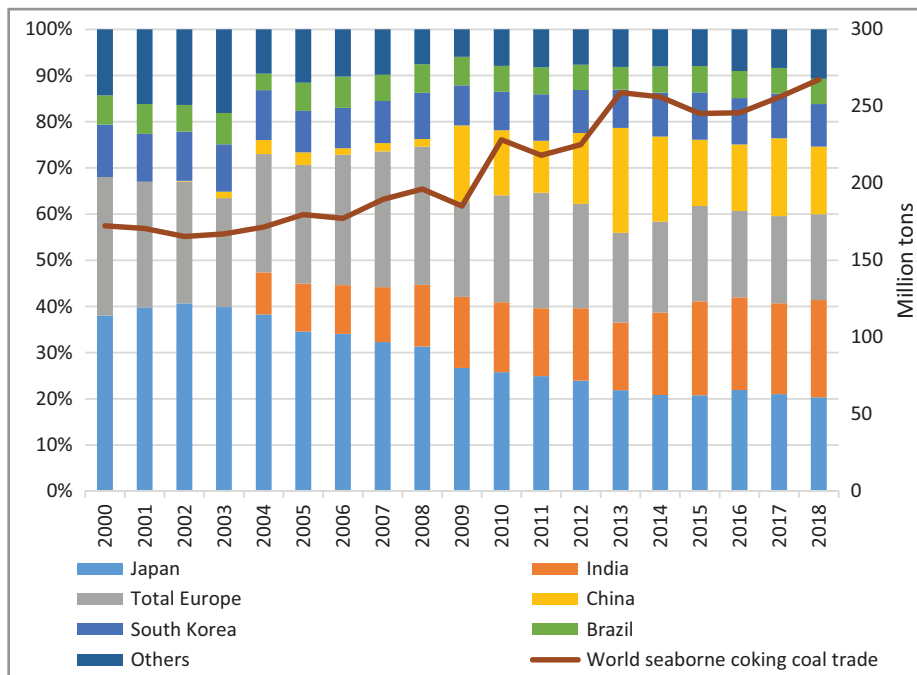


Figure 7 World seaborne coking coal imports (million tons, 2000–18). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

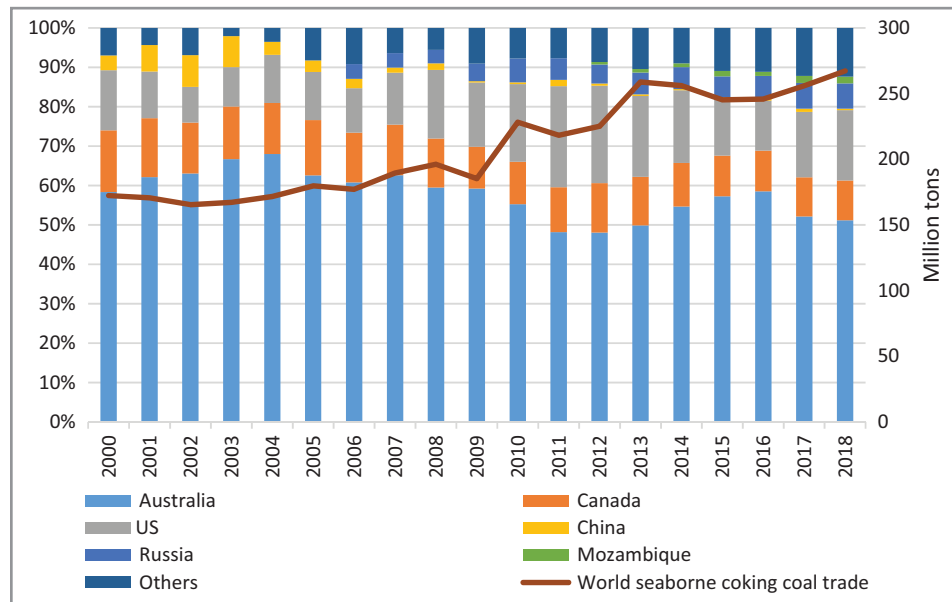


Figure 8 World seaborne coking coal exports (million tons, 2000–18). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

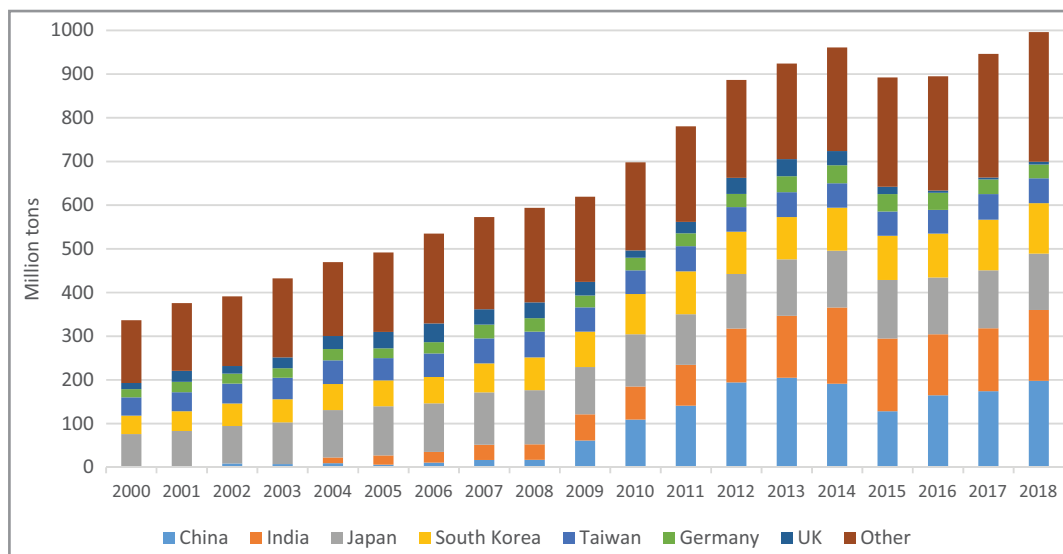


Figure 9 World seaborne steam coal imports (million tons, 2000–18). *Note: Category “Other” includes Malaysia, Thailand, Philippines, Pakistan, Vietnam, France, Italy, Netherlands, Spain, Portugal, Turkey, United States, Israel, Chile and Morocco, among others. Source: Figure created by the authors. Data derived from Clarksons SIN.*

biggest thermal coal exporter until 2013, Russia took over the third position ever since, increasing its exports from 13 million tons in 2000 to 118 million tons in 2018 (Figs. 9 and 10).

Grains Seaborne Trade

World seaborne grains trade has accounted on average for 11% of the world total dry bulk trade during the period from 1999 to 2018. It involves the transportation of wheat, coarse grains and soybean. Wheat is a cereal grain and a staple food worldwide. Its seaborne trade constitutes 36% of the world total seaborne grain trade. Coarse grains generally include cereal grains other than wheat and rice, such as maize, oats, sorghum, barley and various millets. Despite being used primarily for animal feed or brewing, varieties of coarse grains are also consumed by humans given their high nutritional value. The world seaborne coarse grain trade amounts to

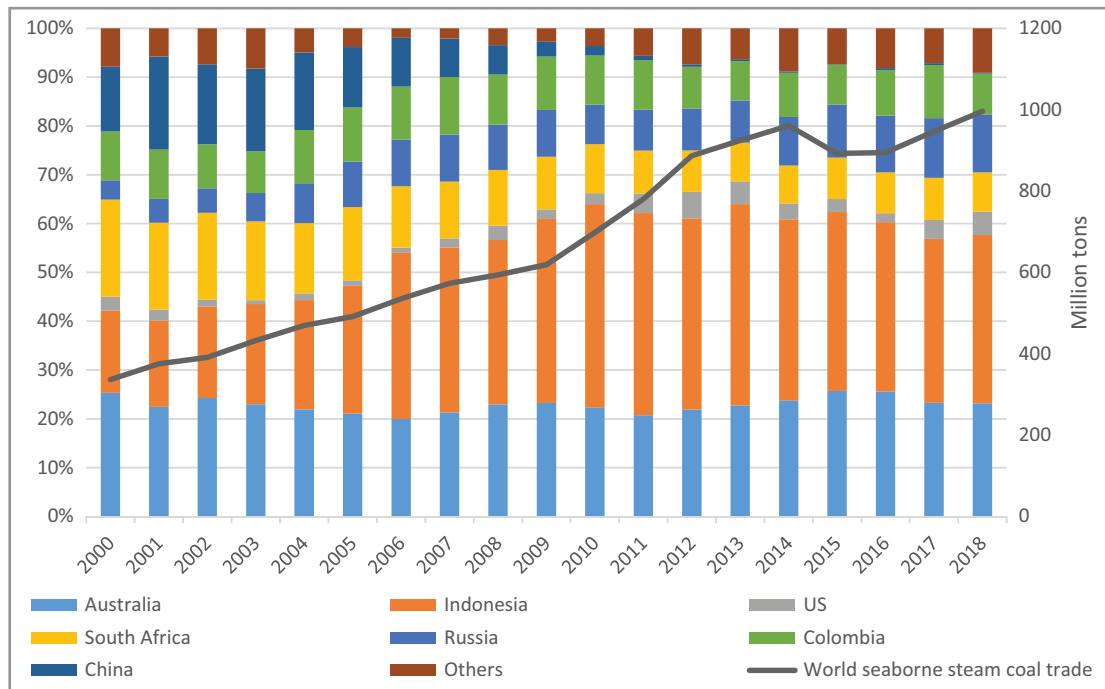


Figure 10 World seaborne steam coal exports (million tons, 2000–18). Source: Figure created by the authors. Data derived from Clarksons SIN.

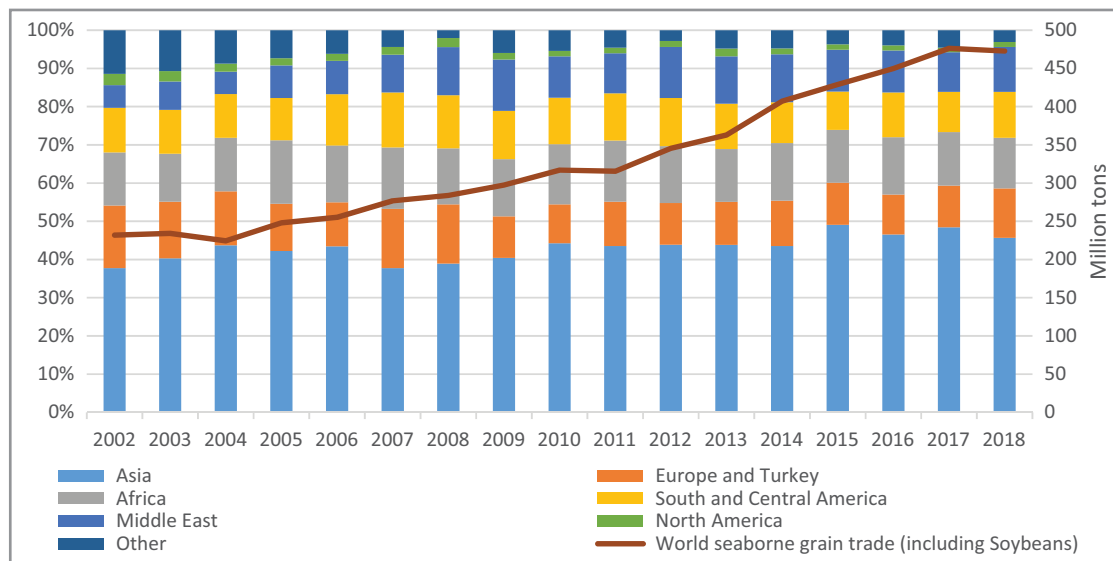


Figure 11 World seaborne grain imports (million tons, 2002–18). Source: Figure created by the authors. Data derived from Clarksons SIN.

38% of the world seaborne grains trade. Soybean is an annual legume consumed in vast quantities worldwide and also an ingredient for hundreds of chemical products. It accounts for 26% of the world seaborne grain trade.

Over 50% of the world's grain cargoes transported by sea are imported by Asia and Europe, while the remaining 50% is mainly imported by Africa, South and Central America, and the Middle East (see Fig. 11). The main grains exporter is the Americas, having exported on average 69% of the world seaborne grain trade flows over the years 2000–18. Other important export regions include Europe and Turkey, Russia and Ukraine, and finally Australia (see Fig. 12).

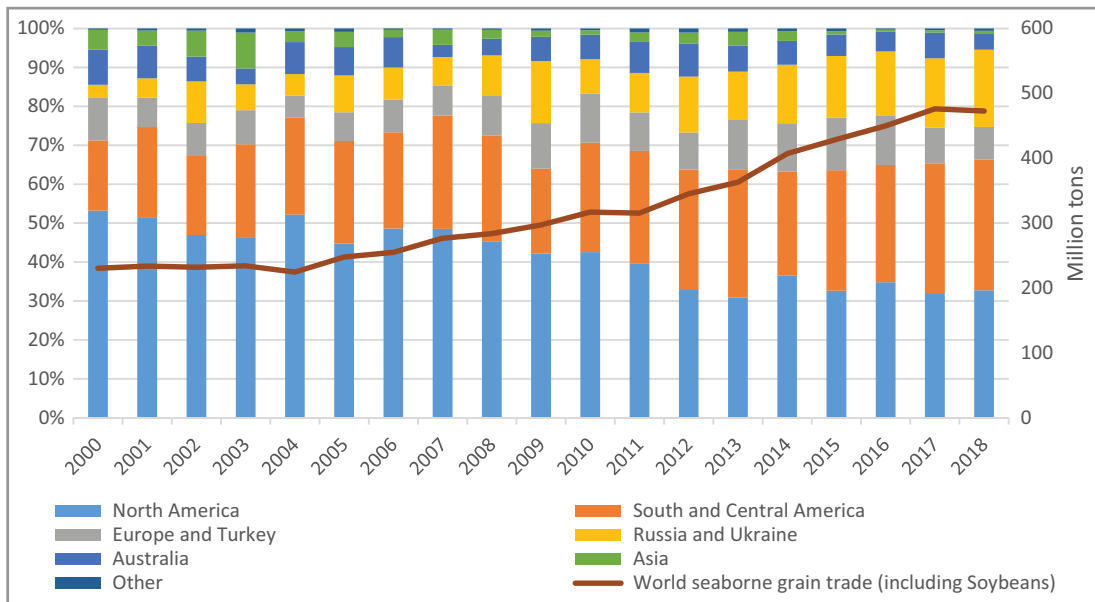


Figure 12 World seaborne grain exports (million tons, 2000–18). Source: Figure created by the authors. Data derived from Clarksons SIN.

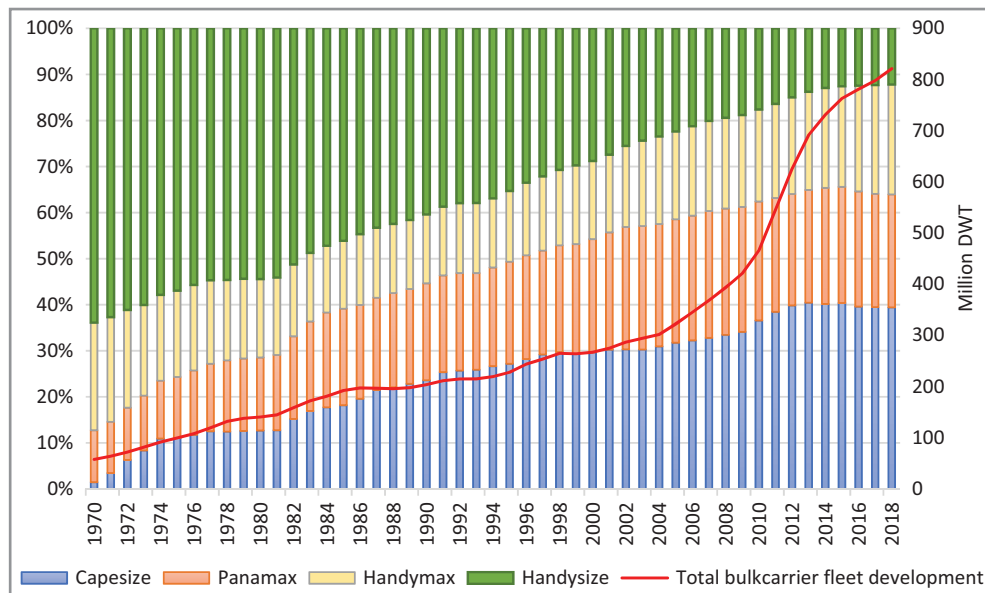


Figure 13 World bulkcarrier fleet development per vessel type (in million DWT, 1970–2018). Source: Figure created by the authors. Data derived from Clarksons SIN.

Supply of Dry Bulk Sea Transportation—Fleet, Freight Rates, and Market Integration

Figs. 13 and 14 present the dry bulk fleet development per vessel type during the years 1970–2018, in terms of deadweight capacity and number of vessels, respectively. The increase in fleet carrying capacity is impressive, from 64 million DWT in 1970 to 821 million in 2018—that is, almost 13 times in terms of deadweight capacity, and 4.5 times in number of vessels, in less than 40 years. It is obvious that the structure of the bulk carrier fleet has changed considerably over the years, with Capesize and Panamax vessels gaining share over Handysize ships, as a percentage of the total bulk carrier fleet capacity.

The fleet growth is greater during periods of increased demand for sea transportation and flourishing freight markets. As can be seen in Fig. 13, after 2002 the bulk carrier fleet capacity rose from 287 to 822 million DWT. The high demand for sea transportation of iron ore to Asia drove freight rates to historically high levels, with the Baltic Dry Index (BDI) skyrocketing from 1137.5 basis points in 2002 to 4510 points in 2004, and 7071.2 points in 2007 (see Fig. 15 for the historical evolution of the components of the BDI per

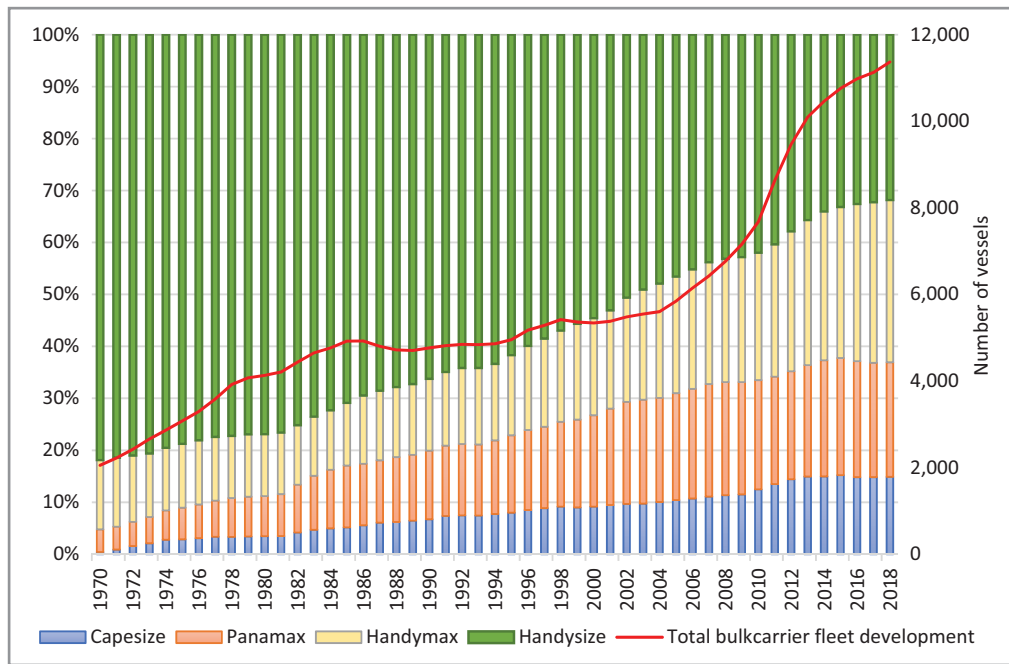


Figure 14 World bulkcarrier fleet development per vessel type (number of vessels, 1970–2018). Source: Figure created by the authors. Data derived from Clarksons SIN.

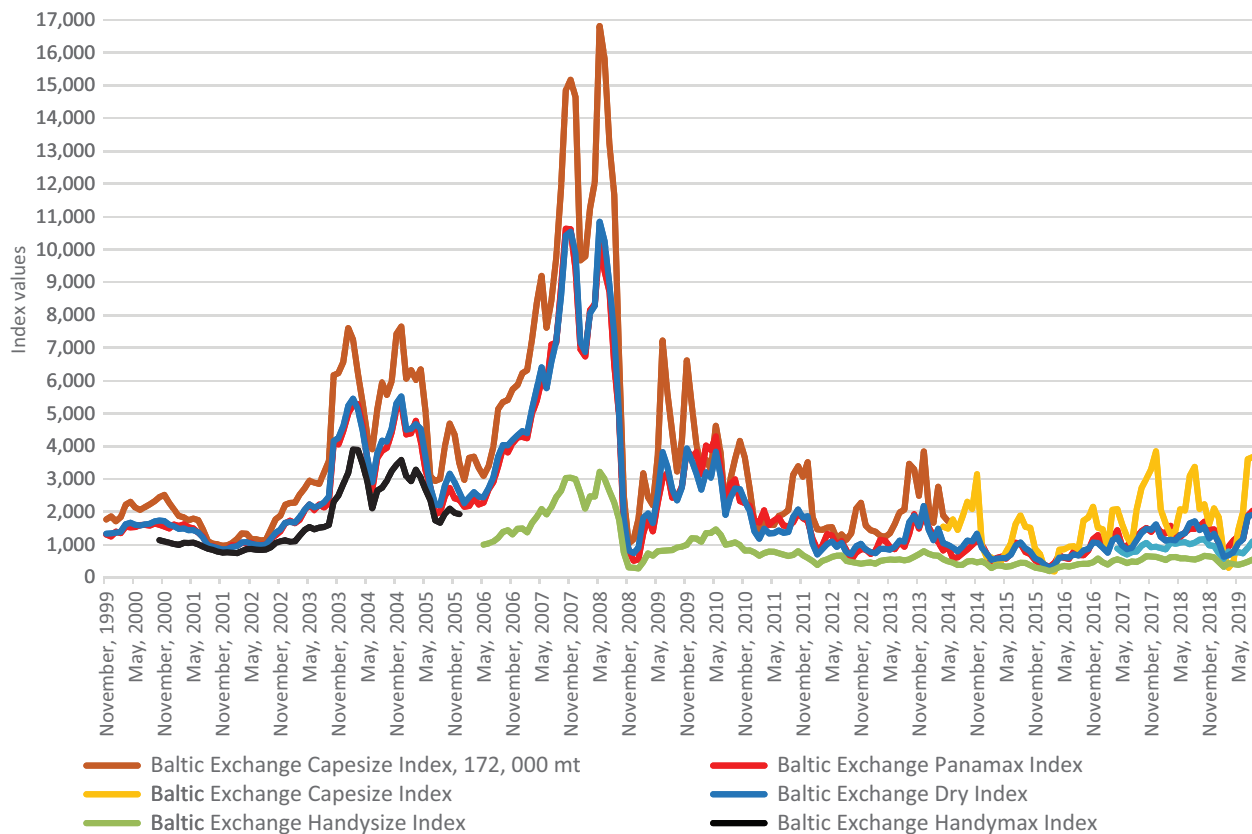


Figure 15 Baltic exchange dry bulk freight indices (in index values, 1999–2019). Source: Figure created by the authors. Data derived from Clarksons SIN.

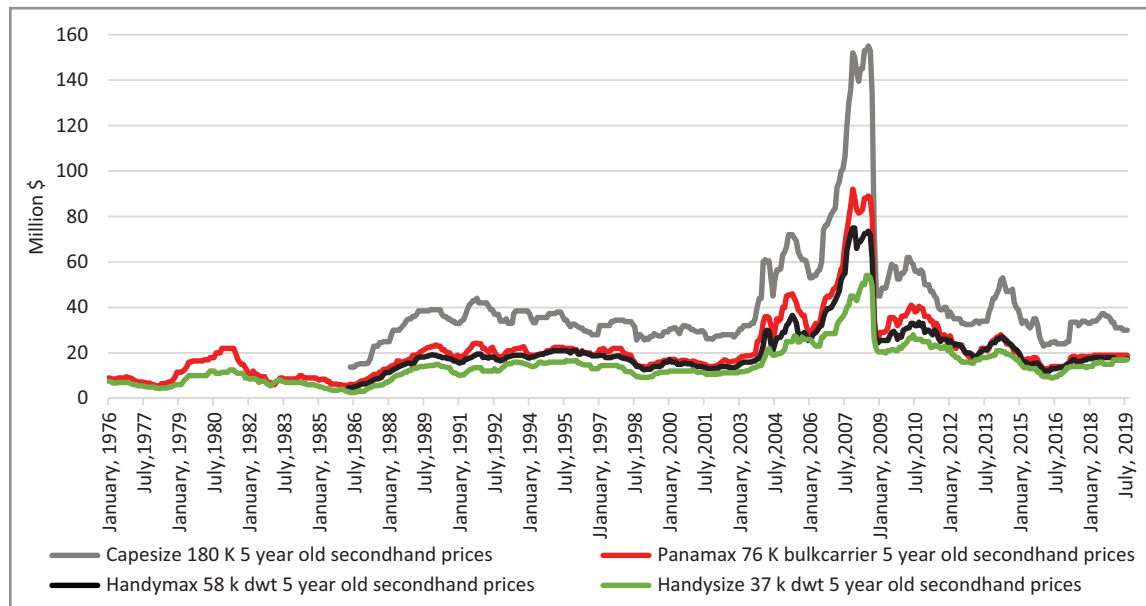


Figure 16 Dry bulk vessels, 5 year old secondhand prices (in million \$, 1976–2019). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

vessel category). Bulk carrier 5-year-old secondhand prices for all types of vessels followed suit, and also reached their highest levels in 2007, with Capesize vessels selling for more than \$150 million, Panamax reaching \$90 million, and Handymax and Handysize prices also peaking during this period (Fig. 16).

Expecting that China's economic growth would continue at the same or higher pace for the years to come, shipping investors kept ordering new vessels, thereby increasing world vessel tonnage. The bulk carrier 5-year-old secondhand price index reached the highest level of 422.8 points in 2007, from its level of 89.2 points in 2002 (see Fig. 16 for the historical evolution of 5-year old second hand prices per segment). The increase in the bulk carrier orderbook was impressive, and increased from 24 million deadweight tons in 2002, to 327 million in 2009 (Fig. 17). Since new orders usually have a delivery time of 18 months to 3 years, there is obviously a considerable time-lag between the peaks of demand and supply. In 2009, the bulk carrier orderbook was 78% of the total bulk carrier fleet, with the Capesize sector accounting for the largest part of it. This peak was reflected in newbuild deliveries in 2011 and 2012, their level surpassing the historical value of 100 million DWT (see Fig. 18).

Accordingly, during the period of booming markets, the demolition market was almost inactive. It was only in 2008 and 2009, during the global economic crisis, that demolitions recovered, reaching their highest level since 1970 in 2012, with a value of 8.3 million DWT (see Fig. 19). At that time, while freight and sales and purchase markets were at their lowest levels, with oversupply of dry-bulk capacity having already been created (newbuild deliveries were at their highest level), the demolition of old capacity could be seen as a potential relief to the markets. The inevitable collapse came in the latter part of 2009, and was a result of the choking off of world demand following the global economic crisis, coupled with the oversupply of vessels in the shipping markets. Ever since, freight rates and ship values fluctuate well below the levels seen prior to 2008.

The Wet Bulk (Tanker) Sector

The Derived Demand for Wet Bulk Sea Transportation—Commodity Trades

In the tanker sector, the main commodities transported are crude oil, oil products, liquid chemicals (such as caustic soda), liquefied gases (e.g., Liquefied Natural Gas/LNG or Liquefied Petroleum Gas/LPG), vegetable oils, and wine. As can be seen in Fig. 20, crude oil trades dominate the sector, followed by the trade of oil products, LNG, and Chemicals.

Crude Oil Seaborne Trade

Approximately 85-90% of world seaborne crude oil exports come from the Middle East, Africa, the Former Soviet Union (FSU) countries and Latin America (see Fig. 21), with the Middle East being by far the largest exporter, holding a 45% world export share. The main importers are South and Southeast Asia, Europe and the United States (see Fig. 22). Chinese seaborne crude oil imports from 2000 to 2018 quintupled, from 69 to 414 million tons, due to oil demand rising faster than domestic production, given the country's strong economic and refinery-capacity growth. China became an oil importer only in the 1990s, when its domestic oil production was not sufficient to cover demand.

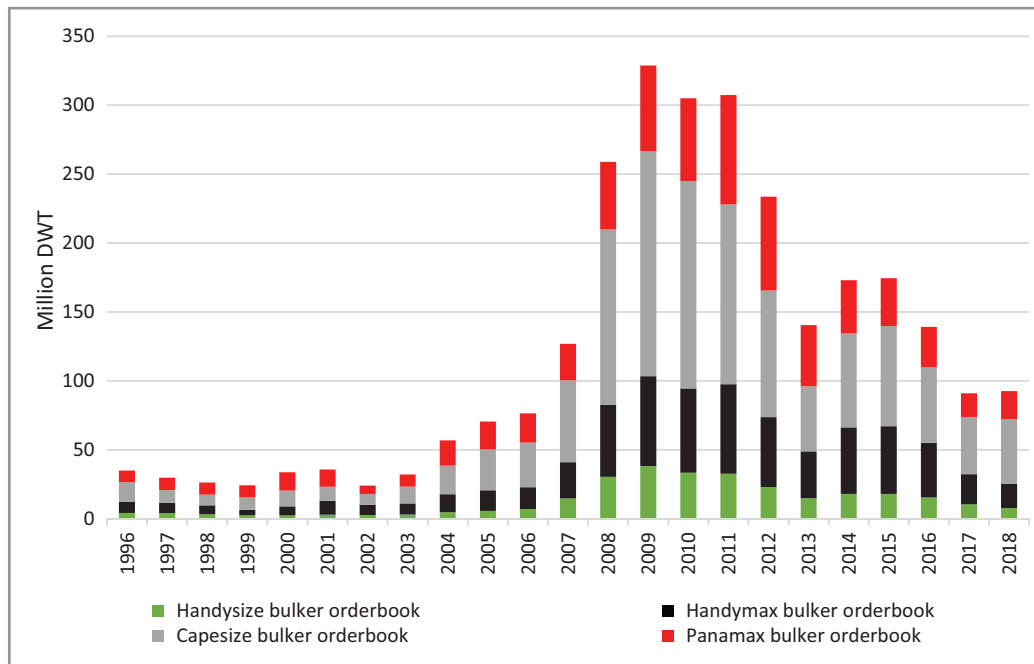


Figure 17 World bulkcarrier fleet orderbook per vessel type (in million DWT, 1996–2018). Source: Figure created by the authors. Data derived from Clarksons SIN.

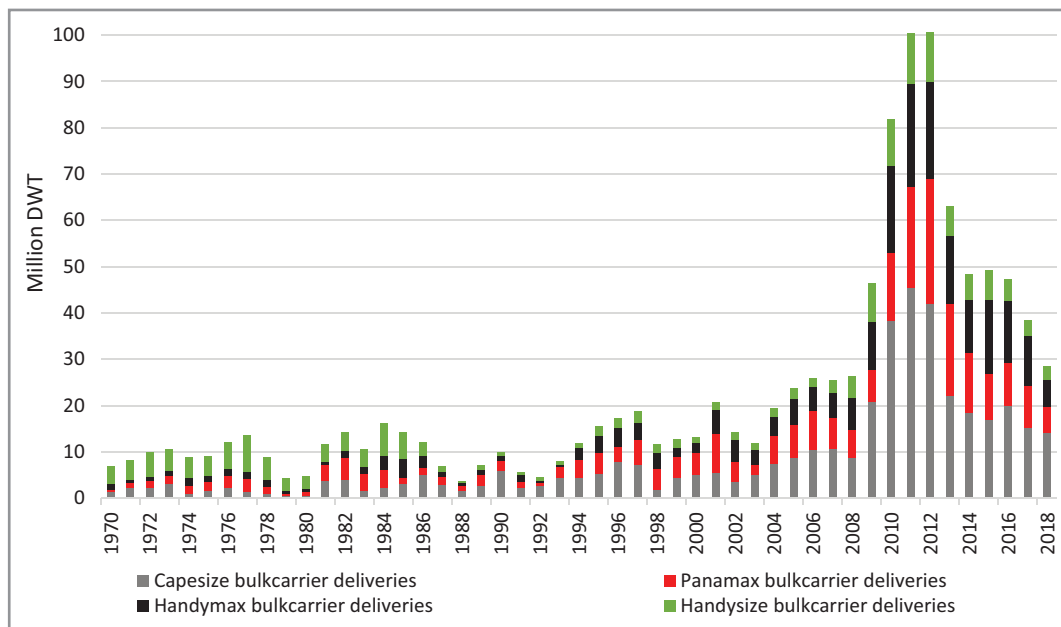


Figure 18 World bulkcarrier fleet deliveries per vessel type (in million DWT, 1970–2018). Source: Figure created by the authors. Data derived from Clarksons SIN.

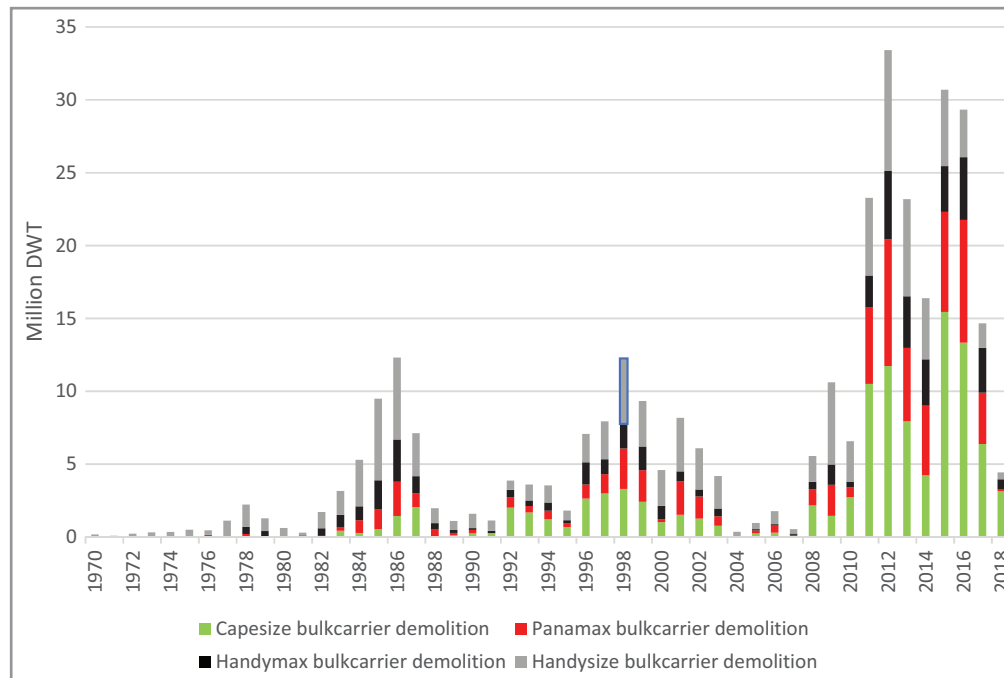


Figure 19 World bulkcarrier fleet demolition per vessel type (in million DWT, 1970–2018). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

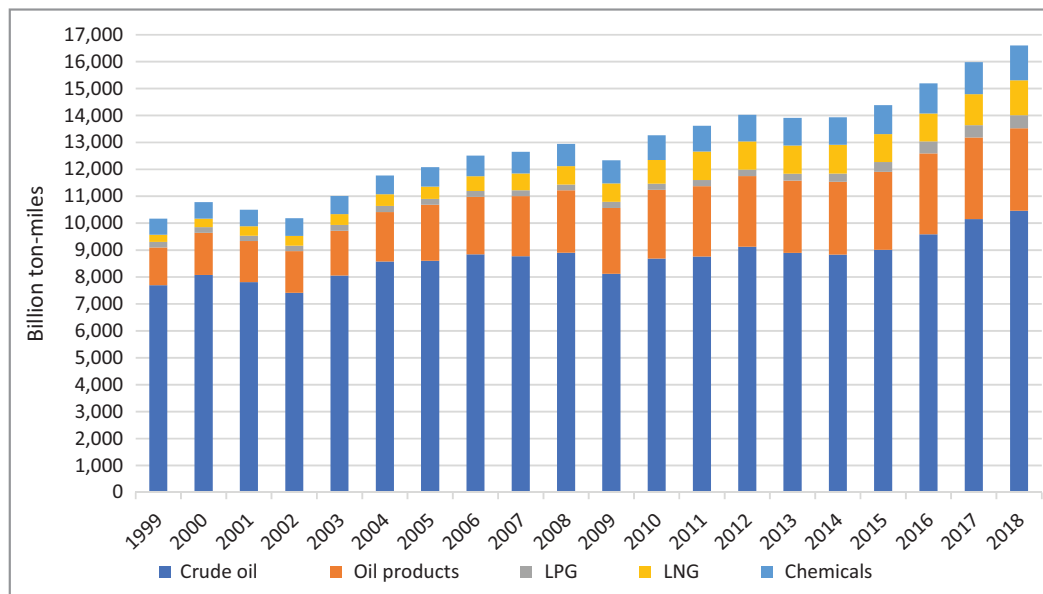


Figure 20 Tanker sector major trades (in billion ton-miles, 1999–2018). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

The world seaborne oil trade is driven by developments in the oil market, which is characterized by high volatility. Observing **Figs. 21 and 22**, one can distinguish three major downturns in the total world seaborne oil trade: one in 2002, a second in 2009 and a third in 2014. The first drop is attributed to the 9/11 terrorist attack which caused a severe decrease in oil prices. From 2002 to 2008,

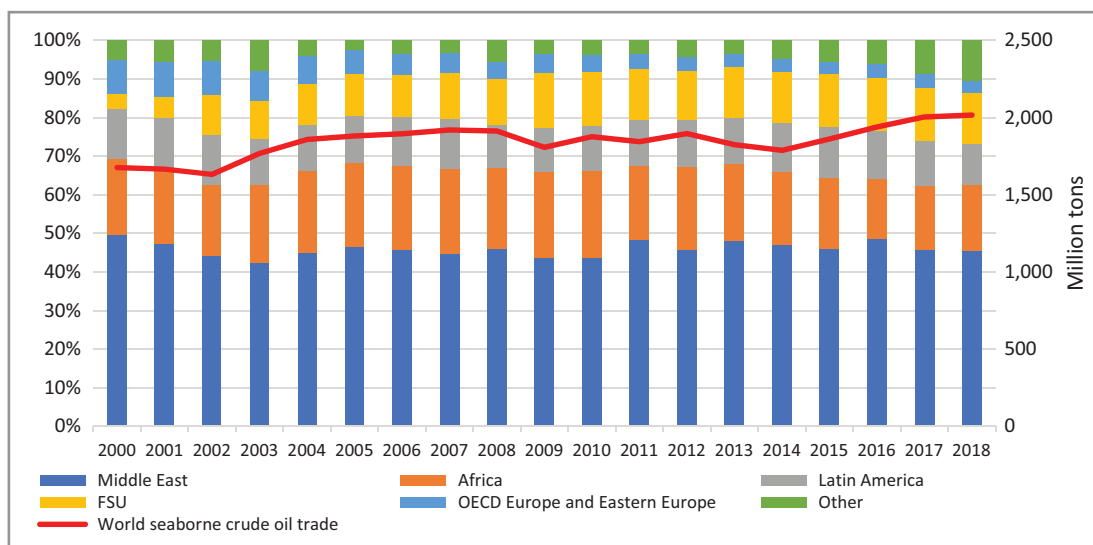


Figure 21 World seaborne crude oil exports (in million tons, 2000–18). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

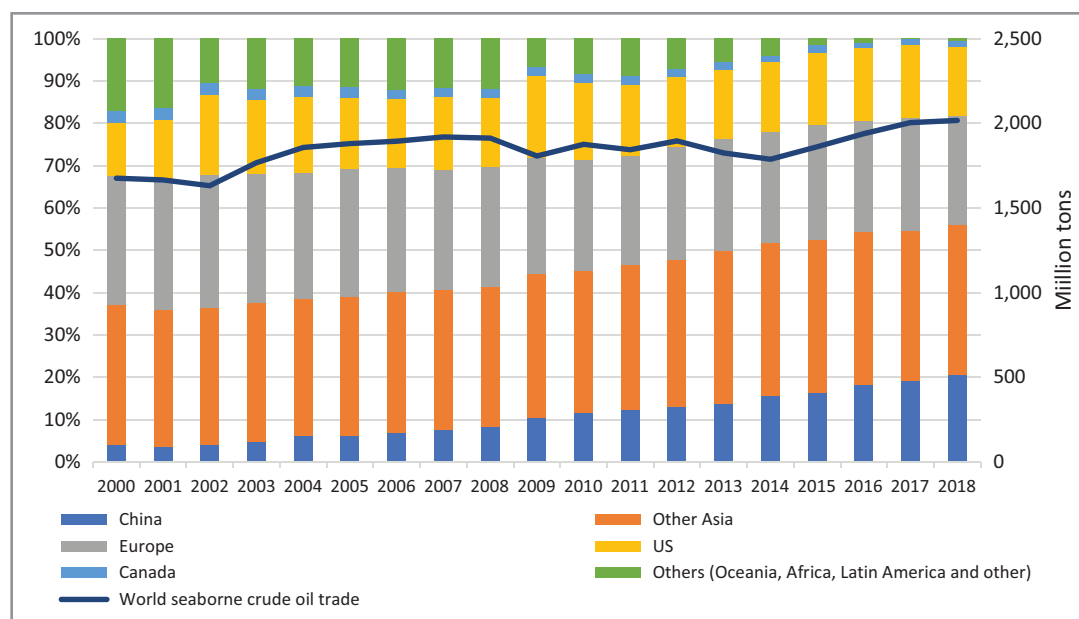


Figure 22 World seaborne crude oil imports (in million tons, 2000–18). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

the oil market experienced a boom, with oil prices increasing year after year, until the 2009 spectacular collapse in the aftermath of the Jeddah Energy Meeting in June 2008^j and the deepening global economic recession. In subsequent years, given the boom in US shale oil production and the weaker economic growth in China and Europe, the oil market has been highly volatile. In the interest of restoring oil market equilibrium, OPEC 2014 Conference Meeting in Vienna^k decided to maintain oil production levels at 30 million barrels per day, causing another sharp drop in oil prices.

^j Given their concerns about sharp increases in oil prices and their volatility, which threaten global oil market stability and sustainable economic growth, Ministers and representatives from many oil producing and consuming countries met in Jeddah, Saudi Arabia, on June 22, 2008. A Joint Statement by the Kingdom of Saudi Arabia and the Secretariats of the International Energy Agency, the International Energy Forum and the Organization of Petroleum Exporting Countries (OPEC) was the result of this meeting. In this Statement, participants agreed on the necessity of specific measures regarding the timeliness and adequacy of oil supply in the market, regulation and transparency in financial markets, quality, completeness and timeliness of oil data, etc. See OPEC Press release on www.opec.org.

^k See OPEC Press release on www.opec.org.

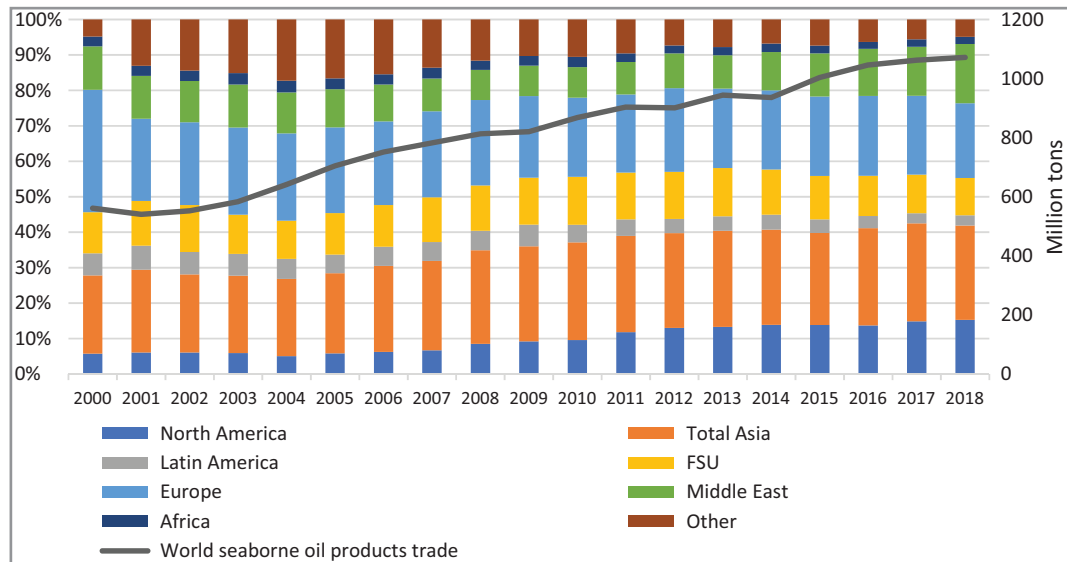


Figure 23 World seaborne oil product exports (in million tons, 2000–18). Source: Figure created by the authors. Data derived from Clarksons SIN.

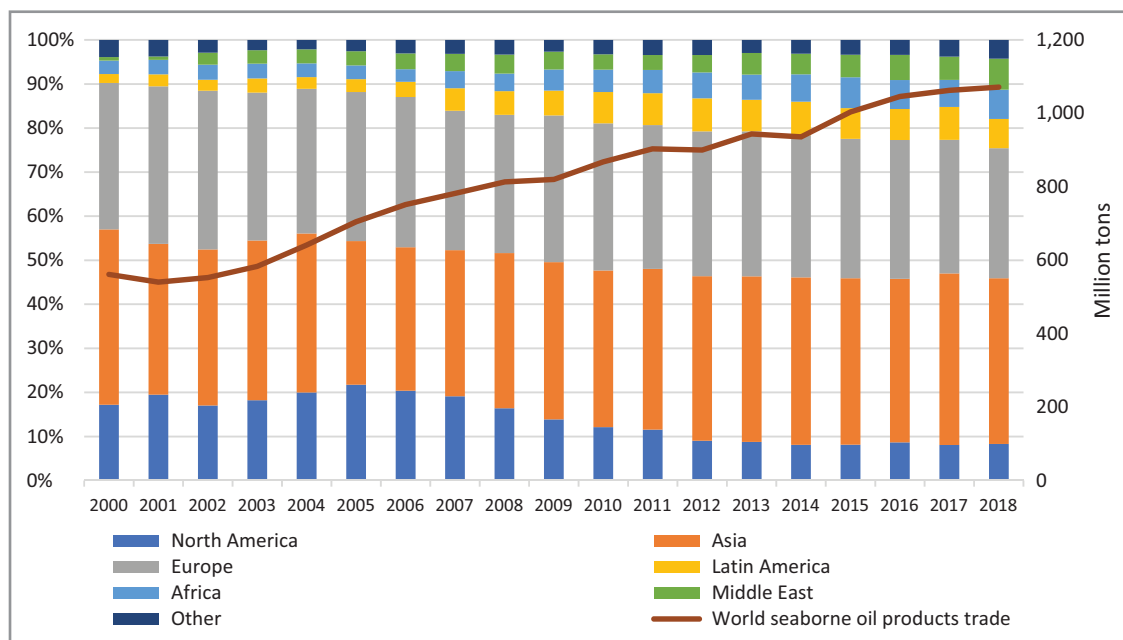


Figure 24 World seaborne oil product imports (in million tons, 2000–18). Source: Figure created by the authors. Data derived from Clarksons SIN.

Oil Products Seaborne Trade

Seaborne exports of oil products mainly originate from Asia, Europe, the Middle East, North America and the FSU (see Fig. 23), while the main importers are Asia and Europe (see Fig. 24). Developments in the oil refining industry affect both crude and products trade. During the past decade, China and the United States have been increasing their refinery capacities, growing also their intra-regional products exports—from China to Southeast Asia and from the United States to North and Latin America. China's rapidly growing refinery capacity is associated with the growth in Chinese seaborne crude oil imports, while the same effect is less strong in the case of the United States, due to increased domestic crude oil production and low oil demand levels in recent years. The European products trade has experienced several shifts, prompted by significant changes in oil prices and inventories. The intra-European products trade accounts for over 70% of European imports.

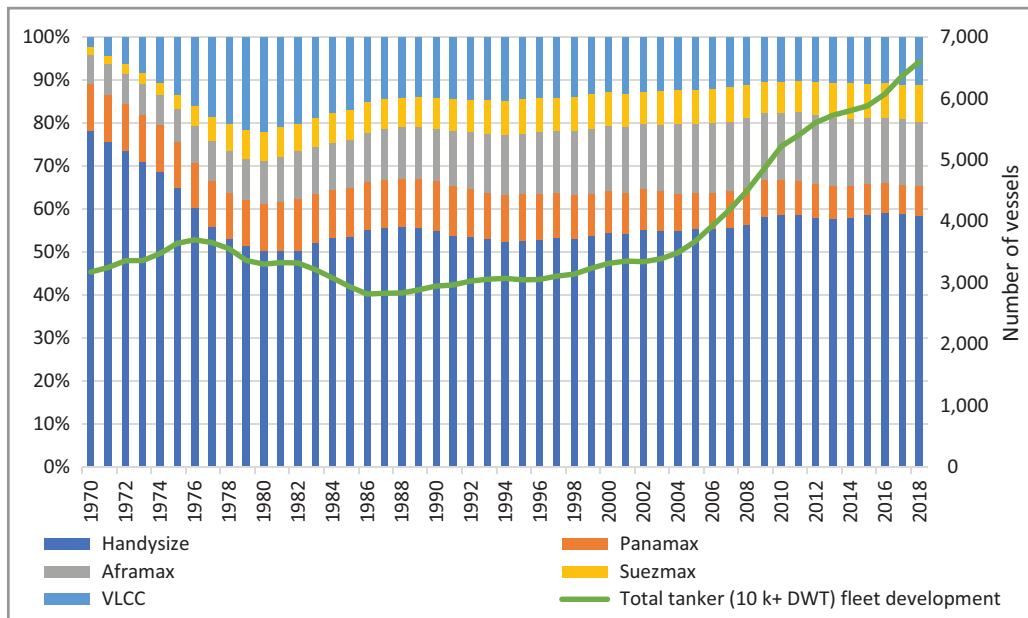


Figure 25 World tanker fleet development per vessel type (number of vessels, 1970–2018). Source: Figure created by the authors. Data derived from Clarksons SIN.

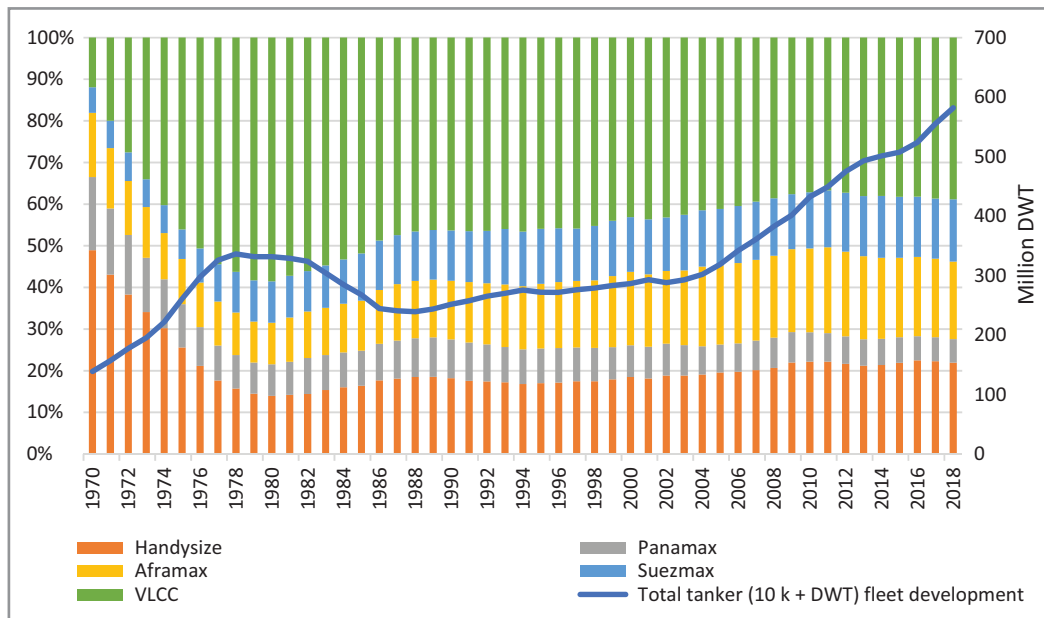


Figure 26 World tanker fleet development per vessel type (in million DWT, 1970–2018). Source: Figure created by the authors. Data derived from Clarksons SIN.

Supply of Wet Bulk Sea Transportation—The Tanker Fleet Throughout Market Cycles

From 1970 to 2018, the tanker fleet doubled in number of vessels (see Fig. 25) and tripled in deadweight capacity (see Fig. 26), reflecting the increasing need for transportation capacity to match the demand for wet bulk trades. The firm tanker market in the early 1970s resulted in speculative investments in newbuild tanker vessels, which were delivered in the following years. In 1975, new deliveries reached their second historically highest level (for the period from the 1970s onwards - see Fig. 27), while the demolition market was almost inactive during the period 1970–74 (see Fig. 28). The oversupply of deadweight capacity and rises in oil prices in 1973 and 1979, brought the tanker freight market to a trough that lasted until 1987. During that period of low markets, there was a

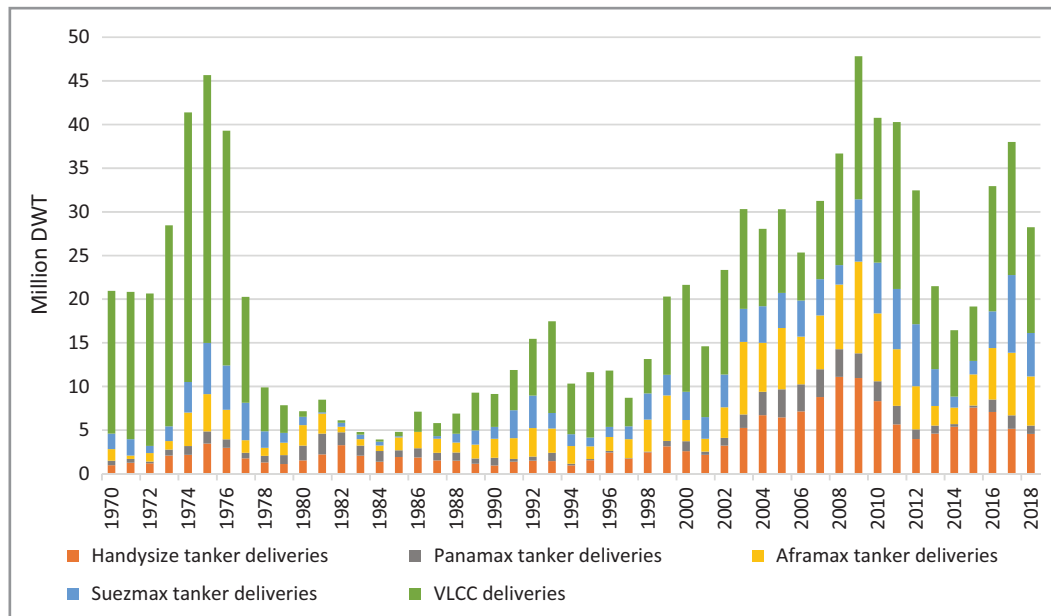


Figure 27 World tanker fleet deliveries per vessel type (in million DWT, 1970–2018). Source: Figure created by the authors. Data derived from Clarksons SIN.

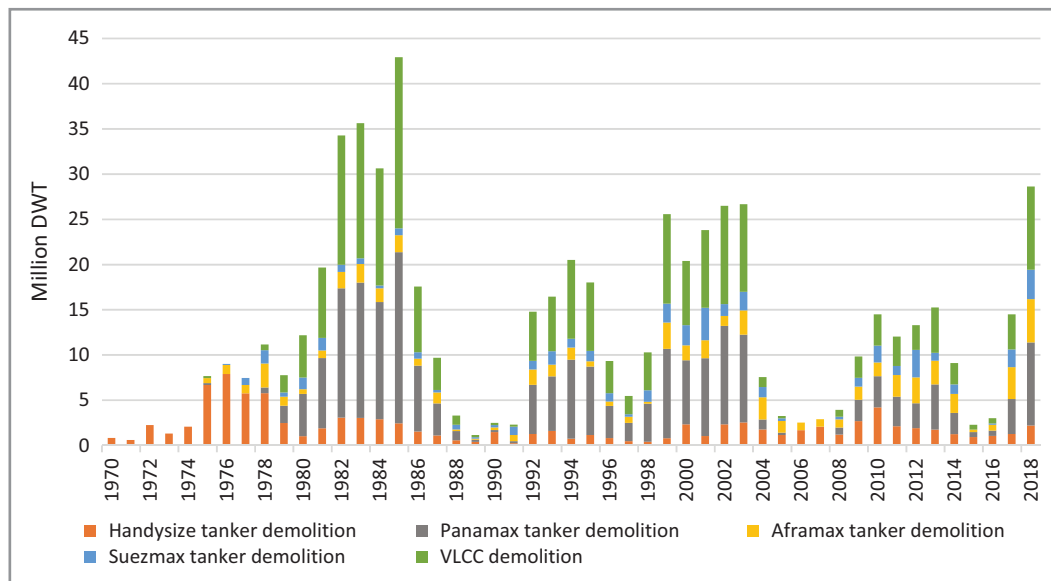


Figure 28 World tanker fleet demolition per vessel type (in million DWT, 1970–2018). Source: Figure created by the authors. Data derived from Clarksons SIN.

decline in total tanker fleet capacity as demolition skyrocketed to levels not since realized, and the newbuilding and sales and purchase market prices (see Fig. 29) collapsed.

The tanker market started recovering after excess capacity had been scrapped, rising to a new freight peak in 1989. Expectations for growing oil demand and high future demolition of vessels built in the 1970s led investors to place new orders. However, expectations were not realized and when new orders were delivered in 1992, they pushed the market to a new recession. During the 2000s, the tanker market flourished, driven by China's increased demand for oil imports to "fuel" its rapidly growing economy. The market boom lasted until the 2009 collapse (see Figs. 30 and 31), due to the deep, global economic recession and the 2008 Jeddah Energy Meeting's decisions to put a ceiling on oil supply, in an attempt to restore equilibrium in the market for crude oil.

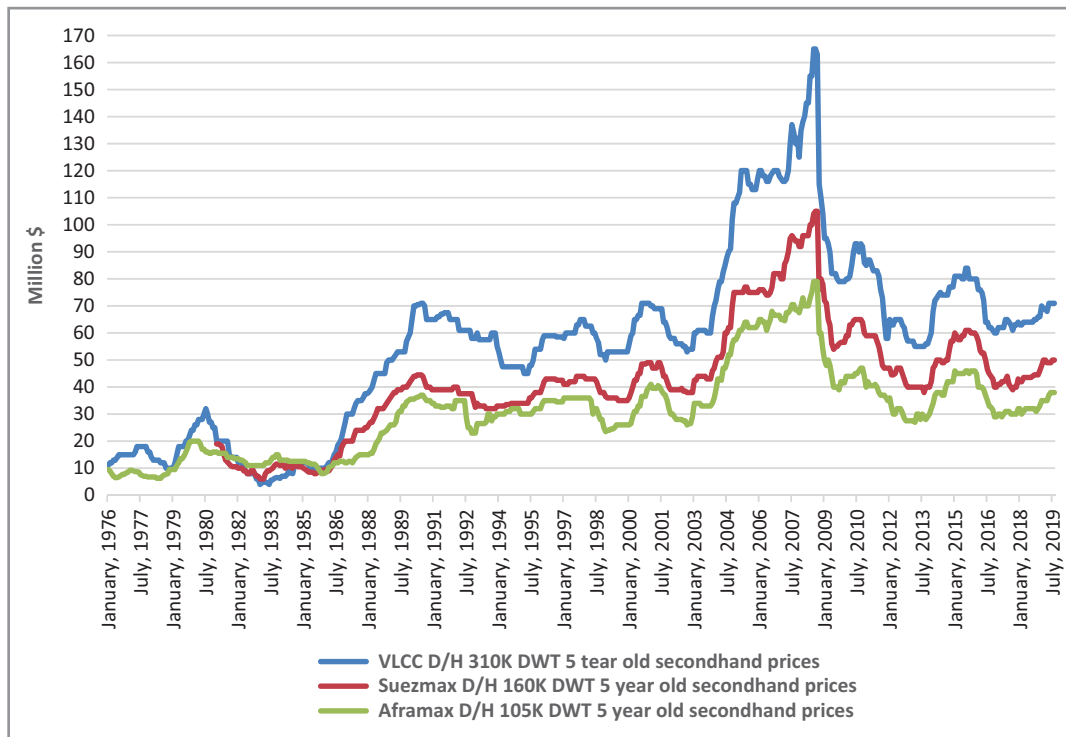


Figure 29 Tanker vessels, 5 year old secondhand prices (in million \$, 1976–2019). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

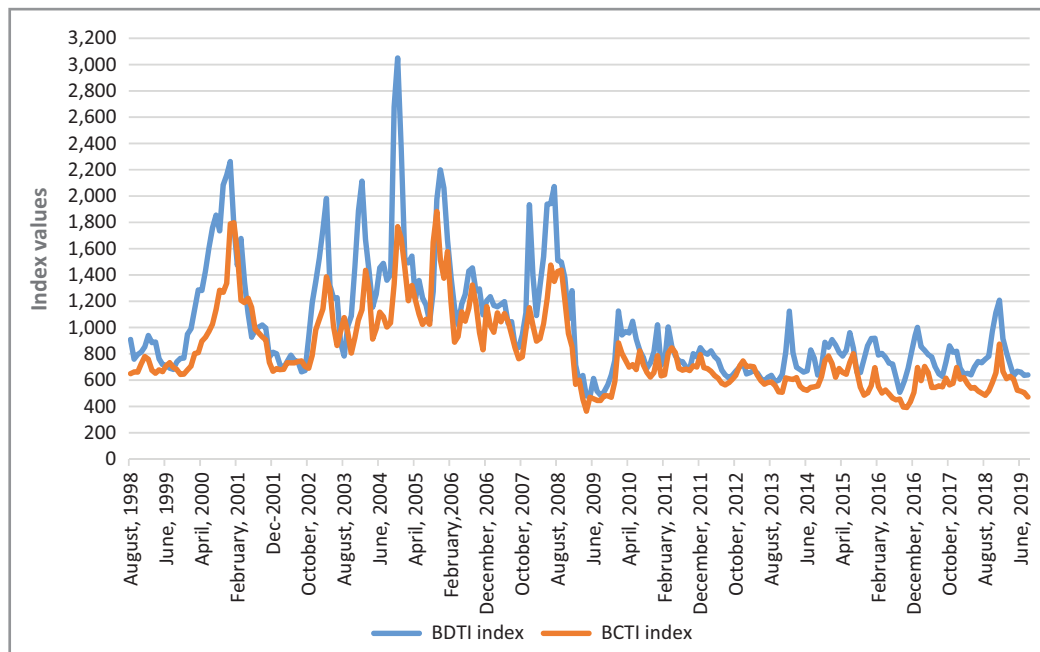


Figure 30 Baltic exchange dirty and clean tanker indices (index values, 1998–2019). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

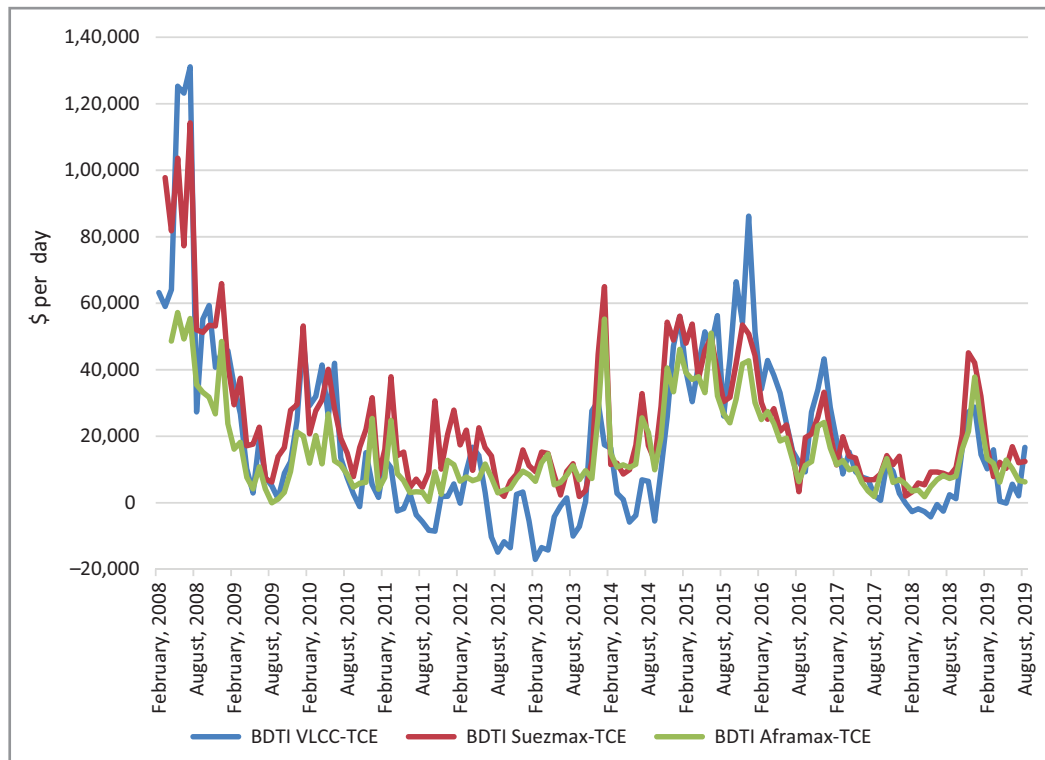


Figure 31 Baltic exchange dirty tanker time-charter equivalents (\$ per day, 2008–19). *Source: Figure created by the authors. Data derived from Clarksons SIN.*

After 2009, the market has been highly volatile during its course to recovery, affected by the weaker economic growth in China and Europe, the US shale oil production boom and OPEC's 2014 decision to prolong the ceiling on oil production levels. In 2018, tanker demolition hit the highest level since 1985, driven by the depressed freight market, the prolonged limits in OPEC oil cargoes and environmental regulation requirements regarding tanker vessels, which are costly to comply with.

Summary

The dry and wet bulk sectors of the maritime industry account for the largest carrying capacity of ocean going vessels, with containers coming third. It is thought that conditions of perfect competition prevail in bulk freight markets, with freight rates determined by demand and supply in each market. There are distinct submarkets within dry bulk and within tanker sectors, respectively. Albeit driven by common factors, such as the world economy and its economic activity, there are distinct factors in each submarket that should be analyzed separately in order to understand the dynamics of each market. This article has considered the main drivers of demand in each of the markets, including the commodities being transported and their characteristics. It also clearly identifies the types of vessels involved in each market and considers how the constituents of demand, supply, freight rates, and ship prices have evolved over time in each of the segments. In sum, it provides a brief and concise description of the dry and wet bulk sectors of the maritime industry.

Relevant Websites

OPEC Press Releases: www.opec.org.

References

- Greenwood, R., Hanson, S.G., 2015. Waves in ship prices and investment. *Quart. J. Econ.* 130 (1), 55–109.
- Kalouptisidi, M., 2014. Time to build and fluctuations in bulk shipping. *Am. Econ. Rev.* 104 (2), 564–608.
- Kavussanos, M.G., Alizadeh, A.H., 2001. Seasonality patterns in dry bulk shipping spot and time-charter freight rates. *Trans. Res. E* 37 (6), 443–467.
- Kavussanos, M.G., Alizadeh, A.H., 2002. Seasonality patterns in tanker spot freight rate markets. *Econ. Model.* 19 (5), 747–782.

Kavussanos, M.G., Visvikis, I.D. (Eds.), 2016. *The International Handbook of Shipping Finance, Theory and Practice*. Macmillan Publishers Ltd, London.

Norman, G., 1981. Spatial competition and spatial price discrimination. *Rev. Econ. Stud.* 48 (1), 97–111.

Stopford, M., 2009. *Maritime Economics*. Routledge, New York pp. 94–98.

Further Reading

Cullinane, K.P.B. (Ed.), 2011. *International Handbook of Maritime Economics*. Edward Elgar, Cheltenham.

Grammenos, C.T. (Ed.), 2010. *The Handbook of Maritime Economics and Business*. Lloyd's List, London.

Ferries and Short Sea Shipping

Lourdes Trujillo, Alba Martínez-López, Department of Applied Economics, University of Las Palmas de Gran Canaria, Faculty of Economics, Las Palmas de Gran Canaria, Spain; and Department of Mechanical Engineering, University of Las Palmas de Gran Canaria, Industrial and Civil Engineering School, Las Palmas de Gran Canaria, Spain

© 2021 Elsevier Ltd. All rights reserved.

Glossary	280
Introduction	280
The Motorways of the Sea (MoS)	282
The MoS Concept and Main Milestones in Its Evolution	282
Current Situation of the MoS and Policy Trends	283
Environmental Sustainability and Short Sea Shipping	284
References	285

Glossary

CEF Connecting Europe Facility program
CNCs core network corridors
DIP detailed implementation plan
ECA Emission Control Area
LNG liquefied natural gas
MGO marine gas oil
MoS Motorways of the Sea
RO-PAX vessel: ferry vessels rolled cargo and passengers for the vessel
RO-RO vessel roll-on roll-off cargo for the vessel
SCR selective catalytic reduction
SECA Sulfur Emission Control Area
SMEs small and medium enterprises
SSS short sea shipping
TEN-T Trans-European Transport Network
TEN-T EA Trans-European Transport Network Executive Agency

Introduction

Despite *short sea shipping* (SSS) being widespread, consensus has not yet been reached about the limits of its nature. Although SSS is an international concept, one of its most cited definitions in the literature is provided by the European Commission (COM(95) 317): *The movement of cargo and passengers by sea between ports situated in geographical Europe or between those ports situated in non-European countries having a coastline on the enclosed seas bordering Europe*. In order to facilitate comprehension, this article will assume SSS to be short-range maritime transport for the movement of cargo and passengers, since this definition is widely accepted by transport stakeholders as distinguishing SSS from deep-sea transport.

Even though SSS is a *traditional concept* as it has always existed, it has gained special attention in the last 2 decades, mainly due to explicit support obtained from policy-makers in different regions, especially from the European Union (EU). The absolute predominance of road haulage over all other transport alternatives in Europe had already been identified as a significant problem in the first *European White Paper on Transport in 1992*. With the intention of finding alternatives to the congestion problems of European roads, the EU promoted its *Transport Policy* and the idea of *intermodality*, as a feasible solution for *door-to-door* transport (Medda and Trujillo, 2010). Hence, since the 1990s, through successive *White Papers on Transport*, European Transport Policy has supported the development of intermodal chains through SSS as a strategic measure to mitigate the negative effects for citizens (e.g., accidents, pollution, congestion, noise, bottlenecks, etc.) of the intra-community transport of commodities.

Far from being a regional concern, the previously mentioned negative consequences of land transport are recognized by the international community, and this reality has driven SSS to be of global interest. Likewise, increasing environmental concerns have reinforced the preference for SSS, since maritime transport has been shown to be a sustainable solution to meet transportation

needs (Medda and Trujillo, 2010). Thus, whereas in the EU, SSS has evolved toward Motorways of the Sea (MoS) (*White Paper: European Transport Policy for 2010*), Canada and the United States signed a *Memorandum of Cooperation on Sharing Short Sea Shipping Information and Experience* in 2003. The competitiveness and feasibility of SSS as the “trunk haul” of intermodal chains has been largely attended in the last decade by transport researchers in several geographical contexts beyond the EU: in Asia, America, and Australia, among others.

This body of research has frequently demonstrated that SSS is a convenient trunk haul for intermodal transport because it enables not just the social costs derived from high traffic concentration on land to be reduced, but also the private costs for transport stakeholders (i.e., shorter traveling distances). Thus, when SSS is a *feasible stretch* for the intermodal chain, this is often the leading option for policy-makers mainly due to: its low requirements in terms of infrastructure investment, its capacity to reach ultra-peripheral regions, and because it can take advantage of economies of scale by using vessels (this also affects the environmental costs). All these motives are very convenient regardless of the location of transport services; however, they are especially significant in regions where transport by land is particularly difficult, the maintenance costs of corridors are high in relation to the size of the economies of the countries, crossing borders by road involves high costs, and there are incomplete connections in the road network. Therefore, despite scarce attention paid in the literature to intermodality in emerging markets, SSS is also put forward as a convenient solution for freight transport in those regions.

As a result of the promotion of SSS since the 1990s, numerous studies and projects have been financed by the EU and by countries from other regions, with the intention of identifying and addressing the weak points of intermodality through SSS. One of the most significant conclusions of these studies was the identification of the weakest attribute of the intermodal chains: the time invested in the transport (Martínez-López et al., 2015a; Suárez-Alemán et al., 2014). This finding was reached through analysis from very diverse points of view. However, the modal choice studies did not just identify this improvement point, but also revealed the negative perception of SSS compared to trucking on the part of shippers: that is, SSS is an outdated, slower, rigid, and complex alternative from an administrative point of view, in addition to being less reliable (Medda and Trujillo, 2010).

As a consequence of these results, enormous efforts were made by policy-makers to reverse this perception. In this regard, the simplification and standardization of bureaucratic processes associated with loading cargoes in European ports was especially successful. In parallel, considerable attention was paid to research studies focused on minimizing the time invested in intermodality through SSS. Among them, it is worth noting the studies related to the feasibility of operating *High Speed Crafts* (although this possibility was finally ruled out due to the high associated costs), the port efficiency impact on intermodality (Suárez-Alemán et al., 2014, 2015), and the identification of the most advantageous routes to operate under SSS conditions (Martínez-López et al., 2015a).

Regarding port performance, the insights drawn from these analyses are highly significant: SSS vessels in the North and Baltic seas were found to spend up to 40% of their time in ports and half of this time was unproductive. As a consequence, specific recommendations were made: to improve cargo-handling efficiency in ports and increase energy efficiency for the SSS fleet by reducing waiting times in port (Suárez-Alemán et al., 2015). In such a way, the port's role in intermodality was developed over time, from the initial nodes for the transport networks to the current situation where they may be regarded as consolidation centers for cargo shipments.

Early in the development of SSS in Europe, considerable efforts were made to identify the most promising potential corridors in which to establish SSS services. It has been widely demonstrated that the relative competitiveness of intermodal transport against road is determined by a combination of the operating and technical features of the transport modes, customer requirements, and geographic characteristics (Medda and Trujillo, 2010). Despite this, the EU established as a priority the geographical definition of the maritime corridors to operate under SSS conditions. Thus, the research studies firstly concluded distance thresholds (general recommendations: maritime distances ranging between 834 and 1400 km and a minimum threshold of 1385 km of land distance between origin and destination points) or particular routes that ensure the “intermodality success” of SSS against road haulage (Martínez-López et al., 2015a).

To reach these findings, load patterns, the features of SSS services, and the technical features of the vessels at the time of the study were all assumed as fixed parameters. As a consequence, optimum route selection was based on their fit with the operating features of the existing fleets in Europe (Martínez-López et al., 2015b). This practice was commonly accepted.

As a consequence of this, most of the initial projects investigating the competitiveness of intermodal chains involving SSS considered solely rolled cargo (i.e., trucks) as cargo units. By so doing, only the roll-on roll-off traffic conditions for the vessels (RO-RO and RO-PAX vessels/ferries) were included in these studies. This choice of the cargo unit was also supported by the idea that RO/RO vessels were the most suitable for intermodal transport through SSS (the lowest times for the cargo-handling operations). In addition, this assumption also involves a political intention to reinforce the idea of complementing, instead of substituting for truck transport through intermodality. Nevertheless, this argument was not sufficiently convincing to rule out other compatible cargo units under intermodality. For that reason, subsequent studies included containers as cargo units and, therefore, feeder vessel activity was included among the intermodal possibilities (Martínez-López et al., 2015b).

In 2003, as a consequence of the *Van Miert Report*, the European Commission reviewed the extension of the *Trans-European Transport Network* (TEN-T) (revision of TEN-T guidelines 2004, Decision No 884/2004), by including four maritime corridors under a new concept: *The Motorways of the Sea*: that is the *Motorway of the Baltic Sea*, *Motorway of the Sea of Western Europe*, *Motorway of the Sea of South-East Europe* and *Motorway of the Sea of South-West Europe*.

Despite the efforts made by various administrations and political institutions with transport responsibilities, it is commonly accepted that the development of SSS has not currently achieved the success expected (Suárez-Alemán et al., 2015). Beyond the

wrong assumptions about the expected behavior of transport decision-makers (based on the value added to the freight, the size of the cargo unit, the type of shippers, etc.), there are several operative reasons for this. The following are notably cited in the literature: the low frequency of transport services, slow speeds, and the lack of efficiency of freight transitions between transportation modes. In addition, numerous authors have also pointed out the policy-makers' responsibility for the imbalance in (Gesé and Baird, 2013):

- the internalization of external costs between SSS and the road (considerable public funds have been uniquely channeled into the development of roads and railway infrastructure);
- the progressive deregulation of road freight; and
- the environmental requirements demanded by the environmental regulations for each transport mode (Cullinane and Cullinane, 2013; Hjelle, 2014).

The latter involves an environmental regulation that is less demanding and has had a slower evolution in the maritime sector than the land normative. Consequently, the technological development in the reduction of pollutant emissions has clearly been slower in the maritime sector (Cullinane and Cullinane, 2013). This fact, along with the characteristics of SSS vessels (small and fast vessels; Martínez-López et al., 2015b), has motivated that the environmentally friendly label traditionally given to SSS in comparison to other transport systems has recently been under discussion (Hjelle, 2014).

Hence, there is a perception that the regions, which have politically supported the development of maritime transport, have, however, transferred the responsibility for the success of intermodal transport (a matter of general interest for citizens), to private initiatives, which not only must take decisions about the operative and technical characteristics of the fleets and facilities (with considerable influence on intermodal performance), but also cope with an initially disadvantaged position with regard to road transport (Martínez-López et al., 2015b, 2018).

The following sections of this article will analyze more closely some of the issues addressed in this introduction. This introduction is firstly followed, however, by a review of the state of the MoS and their implications. Afterwards, there is a brief discussion about the current sustainability of the intermodal chains that involve SSS, by considering the necessary technology to improve the environmental performance of seaborne transport. This section also includes the main challenges for SSS due to its current technological limitations.

The Motorways of the Sea (MoS)

The MoS Concept and Main Milestones in Its Evolution

The MoS concept did not formally appear in European Transport Policy until 2001 (*White Paper: European Transport Policy for 2010: Time to Decide*, 2001). However, although the concept is quite clear, no formal definition was provided. The MoS are introduced in the European Transport Policy as a wider concept than SSS, which attempts to respond with high effectiveness to the initial aims of seaborne transport in the EU: alleviating congestion and bottlenecks in the European road and rail networks through modal shift. Thus, the MoS not only involve the maritime corridors of the chains, but also imply their integration into "door-to-door" transport by including all necessary services into the modal shift in ports. Consequently, the MoS concept represents greater influence on the intermodal chain, with wider implications for the overall service (Medda and Trujillo, 2010). This new reality forces improvements on modal integration, administrative procedures among maritime regions, environmental performance, modal shift operations, etc.

Since then, the MoS concept has evolved within European Transport Policy through various promotional programs and legislative events. One of the most significant milestones in the development of the MoS was its inclusion in the TEN-T by the European Commission (revision of TEN-T guidelines 2004, Decision No 884/2004). The revision of the TEN-T guidelines 2004 was carried out according to the *Van Miert Report* (2003), which recommended as a priority the inclusion of the MoS in the TEN-T guidelines and its implementation in the *Marco Polo program* (the European program to reduce transport pollution through alternatives: in this case to support new SSS services).

The Van Miert Report defined four MoS corridors (the *Baltic Sea*—linking the Baltic Sea region with Central and Western Europe, the *Sea of Western Europe*—the Atlantic arc from the Iberic Peninsula toward North and Irish seas, the *Sea of South-East Europe*—connecting the Adriatic Sea with the Eastern Mediterranean, and the *Sea of South-West Europe*—linking the western Mediterranean with the corridor of the *Sea of South-East Europe* and the *Black sea*) to be integrated into the TEN-T, and that member states should support links between them, in accordance with their interests.

Thus, through the inclusion of the MoS into the TEN-T in 2004, a legal and financial context was provided for MoS activities (European Commission, 2017). In the call for proposals for TEN-T 2005 a number of MoS studies were funded, most aimed at identifying the most promising maritime routes and forecasting the cargo flow attracted by the proposed service. Even though the Marco Polo program began in 2003 with the aim of improving the environmental performance of the freight transport system through the reduction of road congestion, this did not include specific actions for MoS until the Marco Polo II (Regulation 1692/2006, updated by Regulation 923/2009) program (2007–13). Despite this retardation, this has been the main support program to develop the MoS concept.

In 2007, the MoS attained a new role in EU policy by being considered as one of the five axes to promote international exchange, trade, and traffic between the EU and its neighboring countries (COM(2007) 32: *Extension of the major trans-European transport axes to the neighbouring countries*). In the same year, the European commission published a report about the state of the MoS (COM(2007)

606: *The EU's freight transport agenda: Boosting the efficiency, integration and sustainability of freight transport in Europe. Report on the Motorways of the Sea: State of play and consultation*). Among other conclusions, the report proposed the creation of a one-stop help desk (the MoS Help Desk launched by the Trans-European Transport Network Executive Agency—TEN-T EA—and Executive Agency for Competitiveness and Innovation—EACI—in 2010) that could provide information about the financing of MoS projects and the establishment of specialized commission executive agencies to follow the TEN-T and Marco Polo project proposals.

In 2010, the *Europe 2020 strategy for smart, sustainable and inclusive growth* (COM(2010) 2020) included energy and climate targets that affected the MoS projects in the TEN-T framework. The 2013 work program: *Annual Report of the European Coordinator for the Motorways of the Sea* (Valente de Oliveira, 2013) also identifies sustainability as part of the general aim of the TEN-T MoS network: including the use of liquefied natural gas (LNG) as an environmental trend. In addition, as part of the *Europe 2020 strategy*, the communication launched on 2009 (COM(2009) 10) about the establishment of a European maritime transport space without barriers was developed. The communication proposed harmonizing and simplifying administrative procedures in SSS through the following actions (among others): the simplification of customs formalities for transporting goods between EU ports; the simplification of port reporting formalities (Directive 2010/65/EU), creating an “e-maritime” system for electronic data transmission; and the establishment of the *national administrative single windows* for port formalities.

In 2013, the TEN-T guidelines were revised again (EU Regulation No.1315/2013). This revision established the *Connecting Europe Facility* (CEF) as one of the funding instruments for TEN-T. The CEF Regulation (1316/2013) defines the Core Network of transport infrastructure: nine *core network corridors* (CNCs) start or finish in a port. In such a way, the MoS are the seaborne stretches of these corridors.

Since the end of the Marco Polo II program (2007–13), the CEF is the only instrument that can finance MoS projects. From 2014 to 2018, 46 MoS projects were funded under the CEF program. Thus, the CEF and the *European Fund for Strategic Investment* (EFSI) are the sources of EU financing for the transport infrastructure projects (2014–20).

Current Situation of the MoS and Policy Trends

According to the *detailed implementation plan* (DIP) (Simpson, 2018) provided by the European Coordinator for the MoS, the MoS in 2015 moved nearly 59% of all maritime transport of goods to and from European ports. This shows the key role that the MoS connections have to ensure the competitiveness of the European intermodal chains. In 2017, more than 800 regular RO-RO and container services were operating in the European transport network, with more than 400 calls from different EU ports, and connections with hundreds of ports around the world.

Even though a significant amount of tramp services are operating in the SSS regime in Europe, the feeder and RO-RO shipping (RO-PAX/ferries as well) show the greatest relevance due to their suitability for securing a modal shift. Consequently, they have essentially been the focus of the MoS investments. Since 2001–18, accumulated EU funding of EUR 857.5 million was targeted at boosting MoS projects (essentially *TEN-T* and *Marco Polo* programs, but also through others like *Interreg*, *Horizon 2020*, etc.).

The most important corridors for container traffic in the EU are as follows (European Commission, 2017):

- In the North of the EU: between the Baltic Sea and the North Sea, the services between continental Europe in the North Sea area, and the transport between the Republic of Ireland and the United Kingdom.
- In the South of the EU: intra-Mediterranean services and between the Black Sea and the Mediterranean.

Since 2002, the number of companies involved in carrying RO-RO cargoes has been decreasing, whereas the size of vessels has been increasing in all regions of the EU (European Commission, 2017). In 2013, the Mediterranean offered the highest number of SSS services for RO-RO cargo, followed by the Baltic Sea and the North Sea.

Despite the recognized relevance of maritime transport in Europe, the European Commission report: “Motorways of the Sea: An ex-post evaluation on the development of the concept from 2001 and possible ways forward. Final Report” (European Commission, 2017) assessed the effectiveness of the MoS projects for the period 2001–13 as “mixed,” by making a reference to the scarce contribution of the MoS to road congestion reduction to date. The industry’s evaluation of the MoS policy with respect to the promotion of modal shift was even more categorical. It was suggested that a disadvantage exists for MoS in relation to their competitors (moving cargo by road or rail) due to the additional regulatory and administrative obstacles (European Commission, 2017).

The limited contribution of MoS to the modal shift from road to sea has led to not only an exhaustive revision report about MoS results (European Commission, 2017), but also to a DIP provided by the European Coordinator for MoS in 2018 (Simpson, 2018). Among other points for improvement, the reports mention that:

- scarce attention was paid to modal shift targets by the MoS projects (predominantly these were simply related to market conditions and administrative obstacles); and
- the absence of road transport operators and small and medium enterprises (SMEs) in most MoS projects. This is particularly noteworthy given the objectives inherent for the MoS in European Transport Policy.

These circumstances, along with the proven fact that the EU subsidies by themselves are insufficient to maintain the interest of shippers in the MoS option, leads observers to consider that the intermodal chains through MoS have not reached the necessary level of effectiveness to be a clear alternative to the road mode for freight transport in the EU.

In light of the earlier, the MoS reports (European Commission, 2017; Simpson, 2018) concluded that there exists a need for a more holistic approach, by covering the whole freight logistics chain, in order to enhance intermodality. In that sense, it is remarkable how few MoS links are currently integrated into the *TEN-T CNCs*. The current corridors are land-based corridors that simply start or finish in the ports. The role of the ports in the *TEN-T* network necessarily must be modified in order to adequately integrate seaborne transport into the nine *CNCs*. The new role of ports must be as a center of consolidation of the load (intermodal node). This point is one of the most recognized improvement points in the development of the MoS, since this involves a wider discussion about the current classification of the ports in the *TEN-T* network (core and comprehensive ports—Regulation (EU) No 1315/2013) and the implications of this for the funding of MoS projects.

The current system of port classification has determined the funding priorities for the MoS projects (CEF program); this being one of the reasons why many peripheral regions and islands have not only been excluded from the network planning and transport maps of the *TEN-T*, but also why despite the importance of their maritime routes for local trade (the routes are connected through outside ports of the core/comprehensive list), they were largely absent from the projects funded from MoS funding programs (Simpson, 2018). Consequently, there is a consensus about the need to extend the current *TEN-T* corridors by reconsidering the current port eligibility criteria (with considerable difficulties applying across all ports, with great contextual variety).

Finally, further trends for the SSS and ports in the EU were provided by the European coordinator for MoS (Simpson, 2018), who highlighted three development pillars as key priorities for SSS and ports in the following years:

- Environment
- Integration of maritime transport in the logistics chain
- Safety, traffic management, and the human element

Environmental Sustainability and Short Sea Shipping

Traditionally, maritime transport was considered a means of sustainable transport. This assumption was mainly supported by the positive effects on costs of the economies of scale to be reaped in the utilization of vessels and because the evaluated pollutants were solely the greenhouse gases (mainly just CO₂ emissions) in most cases. However, the assessment is not so favorable when acidifying substances or ozone precursors are analyzed (Cullinane and Cullinane, 2013; Hjelle, 2014). This performance is even more disadvantageous when the analysis is carried out for SSS conditions, due to the technical and operational characteristics of SSS vessels (fast and small ships, along with the modal shift necessary for door-to-door transport) that generate higher emissions per ton and kilometer (Martínez-López et al., 2018).

As a consequence of the results obtained in recent evaluation reports, numerous authors are demanding special attention be paid to SSS sustainability and rejecting the *self-evident idea* of SSS as an environmentally friendly means of transport. In 2010, in a revision of exhaust emissions for Norwegian domestic maritime transport, the relevance of the NO_x, CH₄, black carbon emissions, and particulates in the SSS was confirmed (Nielsen and Stenersen, 2010). Along the same lines, the European Environment Agency (EEA) concluded in 2013 (Technical Report No. 4/2013) that the NO_x emissions for European maritime transport will reach the same level as land transport NO_x emissions by 2020. Taking into account the local and regional impact of this pollutant, this finding is especially concerning for the SSS in Europe, due to the virtually nonexistent regulations (Bengtsson et al., 2014).

Even though certain EU zones are classified as “Sulfur Emission Control Areas (SECAs)” (the Baltic Sea, the North Sea, and the English Channel, with the Mediterranean shortly expected to join) by the *International Convention for the Prevention of Pollution from ships* (MARPOL 73/78; Annex VI: Regulations for the Prevention of Air Pollution from Ships), this classification simply involves limitations of SO₂ emissions (SECA zones); while no NO_x is restricted (Bengtsson et al., 2014). In this sense, despite the reaction of the EU to these environmental results, no specific European regulation has limited the NO_x emissions in maritime transport to date. This issue has contributed to the increasing criticism about the unbalanced treatment of sustainability among transport modes by policy-makers.

Evidence of this inequality is the different evolution of the European restrictions for environmental sustainability in road transport in relation to seaborne transport (Cullinane and Cullinane, 2013; Brynolf et al., 2014). The *White Paper on Transport 2011* drew up targets for maritime transport greenhouse emissions that were specified in 2013 through a “step” strategy (COM (2013) 479). Since then, only the first step has materialized through the setting up of *Monitoring, Reporting and Verification* (MRV) systems for CO₂ emissions in vessels (the first reporting period scheduled was for 2018). In addition, in 2016, the EU, through the Directive (EU) 2016/802 of the European Parliament, limited sulfur emissions in all European territorial seas and Exclusive Economic Zones from 2020, but the restrictions simply met the International Maritime Organization (IMO) requirements for the SECA.

This regulation for seaborne transport contrasts with the evolution of standards for land-based industries. Since the 1990s, Euro standards (from Directive 91/542/EEC, amending Directive 88/77/EEC) have driven the environmental requirements for road transport in Europe. With an increasing level of restriction (from Euro I to Euro VI) for all main pollutants (including NO_x and particle pollutants targets) the technological development of trucks has adapted over time to meet these standards.

Even though the most restrictive standard of the *International Convention for the Prevention of Pollution from Ships* (MARPOL 73/78; Annex VI: Regulations for the Prevention of Air Pollution from Ships) was imposed on the European seas (Emission Control Area—ECA zones—which restricts SO₂ and NO_x emissions at the same time), the disadvantage in terms of NO_x and SO₂ emissions for

seaborne transport compared to trucking is notable. Moreover, suitable technological alternatives used in combination with marine bunker fuels and different marine engines to meet the ECA requirements involve significant disadvantages (Brynolf et al., 2014; Martínez-López et al., 2018). Specifically, the abatement of NO_x emissions currently requires: the use of Selective Catalytic Reduction (SCR) systems, which are compatible with petroleum fuel engines, the use of expensive fuels (marine gas oil—MGO), or the use of gas engines. The first option has to cope with the ammonia slip problem or, where gas engines are fed by LNG, methane slip then becomes an issue. This is especially significant for SSS and, thereby, arises as a particular problem (Martínez-López et al., 2019).

Despite international opinion supporting LNG as a feasible solution for SSS environmental problems (the EU has committed itself to establishing infrastructure that will allow LNG to be available to ships in all of its 144 maritime and inland ports by 2025), it is worthwhile noting the current difficulties associated with the use of LNG in ships. On the one hand, this gas by itself remains far from providing emission levels close to Euro VI standards (trucking), and, in addition, recent publications highlight the fact that expected maintenance savings, along with the reduced LNG price, are insufficient to balance the capital costs associated with retrofitting. Finally, the lack of specific regulations and technical solutions for the methane slip from LNG engines has emerged as an additional problem (Brynolf et al., 2014; Martínez-López et al., 2019).

References

- Bengtsson, S., Fridell, E., Andersson, K., 2014. Fuels for short sea shipping: a comparative assessment with focus on environment impact. *Proc. Inst. Mech. Eng. Part M J. Eng. Mar. Environ.* 228 (1), 44–54.
- Brynolf, S., Magnusson, M., Fridell, E., Andersson, K., 2014. Compliance possibilities for the future ECA regulations through the use of abatement technologies or change of fuels. *Transp. Res. Part D* 28, 6–18.
- Cullinane, K.P.B., Cullinane, S.L., 2013. Atmospheric emissions from shipping: the need for regulation and approaches to compliance. *Transp. Rev.* 33 (4), 377–401.
- European Commission, 2017. Motorways of the Sea: an ex-post evaluation on the development of the concept from 2001 and possible ways forward. Directorate-General for Mobility and Transport (DG MOVE). Available from: <https://ec.europa.eu/transport/sites/transport/files/2017-ex-post-evaluation-mos.pdf>.
- Gesé, X., Baird, A., 2013. Motorways of the sea policy in Europe. *Mar. Policy Manag.* 40 (1), 10–26.
- Hjelle, M.H., 2014. Atmospheric emissions of short sea shipping compared to road transport through the peaks and troughs of short-term market cycles. *Transp. Rev.* 34 (3), 379–395.
- Martínez-López, A., Munin, A., Alonso, L., 2015a. A multi-criteria decision method for the analysis of the Motorways of the Sea: the application to the case of France and Spain on the Atlantic Coast. *Mar. Policy Manag.* 42 (6), 608–631.
- Martínez-López, A., Caamaño, P., Castro, L., 2015b. Definition of optimal fleets for sea motorways: the case of France and Spain on the Atlantic coast. *Int. J. Shipping Transp. Logist.* 7 (1), 89–113.
- Martínez-López, A., Caamaño, P., Chica, M., Trujillo, L., 2018. Optimization of a container vessel fleet and its propulsion plant to articulate sustainable intermodal chains versus road transport. *Transp. Res. Part D* 59, 134–147.
- Martínez-López, A., Caamaño, P., Chica, M., Trujillo, L., 2019. Choice of propulsion plants for container vessels operating under short sea shipping conditions in the European Union: an assessment focused on the environmental impact on the intermodal chains. *Proc. Inst. Mech. Eng. Part M J. Eng. Mar. Environ.* 233 (2), 44–54.
- Medda, F., Trujillo, L., 2010. Short sea shipping: an analysis of its determinants. *Mar. Policy Manag.* 37 (3), 285–303.
- Nielsen, J., Stenersen, D., 2010. Emission factors for CH₄, NO_x, particulates and black carbon for domestic shipping in Norway. Revision 1. Norwegian Pollution Agency. Marintek, Klima og forurensningsdirektoratet.
- Simpson, B., 2018. Motorways of the sea detailed implementation plan of the European coordinator. European Commission. Mobility and Transport. Available from: https://ec.europa.eu/transport/sites/transport/files/101_web_final_i_i_mos_dip_2018.pdf.
- Suárez-Alemán, A., Cullinane, K.P.B., Trujillo, L., 2014. Time at ports in short sea shipping: when timing is crucial. *Mar. Econ. Logist.* 16 (4), 399–417.
- Suárez-Alemán, A., Trujillo, L., Medda, F., 2015. Short sea shipping and intermodal competition. *Mar. Policy Manag.* 42 (4), 317–334.
- Valente de Oliveira, L., 2013. Priority Project 21 Motorways of the Sea. Annual Report of the Coordinator. European Commission. Available from: https://www.onthemosway.eu/wp-content/uploads/2016/06/Valente-report-2012_2013-pp21_en.pdf.

Shipping and the Environment

Karin Andersson, Selma Brynolf, Lena Granhag, J. Fredrik Lindgren, Mechanics and Maritime Sciences, Chalmers University of Technology, Gothenburg, Stockholm, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Shipping Today—The Relation Between Shipping and the Natural Environment	286
Emissions to Air and Impacts—The Role of Fuel	286
Greenhouse Gases and Growing Shipping	287
Sulfur Emissions	287
NO _x	287
Particles	287
Fuel Alternatives	288
Life Cycle Impact of Fuel	288
Propulsion Alternatives	289
Electricity/Batteries	289
Wind Assisted	289
Nuclear	289
Fuel Cells	290
Emissions to Water and Impacts	290
Oil Spills	290
Environmental Effects of Oil Spills	290
Biofouling and Use of Hull Coatings	290
Ballast Water	291
Waste Generated on Board	291
Noise	291
Particularly Sensitive Areas, Baltic Sea, Arctic, and Others	291
Sustainable Shipping in the Future	292
Biographies	292
See Also	292
Relevant Websites	292
References	292
Further Reading	293

Shipping Today—The Relation Between Shipping and the Natural Environment

The sea provides the infrastructure for shipping, but it is also a very important part of the natural environment, providing all kinds of ecosystem services to man. More than 70% of the surface of the earth is covered by oceans, and life on earth is completely dependent on the function of oceans. The sea plays an important role in the hydrological cycle but also in the geochemical cycles of carbon, nitrogen, sulfur, and other elements. The sea is used by man for many different purposes other than shipping, like fishing, food production, energy production (wind, waves) as well as oil/gas and mineral extraction. Since “what is below the surface is not visible” has been treated as a truth for a long time, the sea has traditionally been used for waste dumping and objects lost into the sea have not been recovered. When technology allows more sophisticated exploration below the surface and the environmental impacts of dumped or lost materials like plastic objects or fishing nets, becomes obvious, this has become an increasing problem. The competition between the different interests of using the sea can also be a source of conflicts.

Shipping is a large sector globally, more than 90% of goods transport is performed by sea, and in spite of the fact that shipping is the most energy efficient means of transport in terms of energy used per ton km (or ton nm), it also causes impacts. These include impacts not only on the natural environment but also on health, crops, and the built environment. Emissions from shipping are regulated in various ways, locally in ports and fairways, as well as nationally with the framework for international shipping regulations by the International Maritime Organization (the IMO). The use of marine spatial planning (MSP) has been developed as a tool to be used in order to avoid problems that can arise from multiple activities in a marine area.

Emissions to Air and Impacts—The Role of Fuel

The main part of emissions to air from shipping is related to the fuel. Traditionally, combustion engines using heavy fuel oil (HFO) or marine gas oil/marine diesel oil (MGO/MDO) in diesel engines, have been dominating. Also dual fuel engines with the possibility

to run on different fuels [diesel/liquefied natural gas (LNG) or diesel/methanol], or diesel electric engines are used. The use of fossil fuel is associated with emissions of carbon dioxide, but there are also other combustion-related emissions that have environmental impacts.

There are international regulations for shipping emissions to air regarding sulfur (in terms of the fuel content) and nitrogen (in terms of emission of nitrogen oxides related to engine speed). For carbon dioxide (CO₂), the reduction has been handled in terms of energy efficiency (EEDI, energy efficiency design index) for new built ships.

Greenhouse Gases and Growing Shipping

Greenhouse gases (GHG) are gases in the atmosphere that absorb infrared (heat) radiation, keeping the temperature on earth more even through the hours of the day and by this making the earth habitable. The most abundant GHGs are water vapor and CO₂ but there are many more heat-absorbing gases, both naturally occurring and anthropogenic.

The CO₂ is part of the natural carbon cycle between soil and water. However, increased use of fossil fuels after industrialization, combined with a growing population has increased the amount of carbon available in the cycle and thus of CO₂ in the atmosphere. Emissions of more strongly absorbing gases like methane add to the effect. The most discussed result is global temperature increase but the carbon dioxide also leads to increased carbonate contents and increased acidity of the oceans.

In the *Third IMO GHG Study 2014*, it was estimated that GHG emissions from international shipping in 2012 accounted for about 2.2% of anthropogenic CO₂ emissions and that the emissions may grow by between 50% and 250% by 2050 (Smith et al., 2014). The Paris agreement with a very tight goal for CO₂ reduction is not applied to international shipping, since the IMO has the responsibility for the shipping sector. The IMO has attempted to decrease the CO₂ emissions by regulations in terms of demands on energy efficiency for new ships (Energy Efficiency Design Index, EEDI). This has been shown not to be sufficient to obtain a decrease, especially given expected future increases in sea transport, and the IMO in 2018 set a goal of a CO₂ decrease by 50% by 2050 and of zero emissions “within this century” (IMO, 2018).

Sulfur Emissions

Sulfur oxide emissions from shipping are mainly due to the use of fuel with high sulfur content. HFO, that has for a long time been a dominating fuel for larger ships, contains various, and sometimes high, amounts of sulfur—up to 4%, while MGO usually has low-sulfur contents and fuels like LNG or methanol are sulfur-free. The sulfur is a contaminant in the fuel and fulfills no function for the ship engine and the sulfur oxides formed in the combustion are forming a strong acid when dissolved in water. High sulfur content also promotes the formation of particles in the exhausts. The permitted sulfur content has been lowered in steps both globally and in special “Emission Control Areas” (ECA). From 2020, the global limit is 0.5%. In ECAs in the Baltic Sea and North Sea and in North America, the limit is 0.1% (Smith et al., 2014).

The sulfur regulations may be met by using low (or no) sulfur fuel or by continuing with high sulfur fuel and combining this with emission abatement on board. Emission abatement can be performed by washing the sulfur into liquid, by the use of a “scrubber.” This can be done using seawater that is discharged into the sea after use, “seawater scrubbing,” or by using a closed-water/chemicals circuit on board, “closed-loop scrubbing.” In addition to sulfur these can be various hydrocarbons and metals. Seawater scrubbing thus moves the contaminants from air to the ocean in a concentrated form. The impact of this is not fully understood, and the method has been questioned. For both scrubber alternatives, a sludge that has to be treated or disposed of on land is resulting, setting demands on port reception facilities. The scrubber comes with an investment cost as well as a need for space and increased weight on board. The economics of the alternatives varies depending on the fuel price at a certain time. The fuel prices in 2018–19 have been low, making scrubbers attractive, and a large number have been installed. Use of sulfur-free fuel like liquid natural gas, LNG, is another alternative that is increasing.

NO_x

Nitrogen oxides are formed in the combustion, mainly due to the nitrogen (N) and oxygen contents in the combustion air (the N₂ contents in air is around 79%). At the high-combustion temperatures, the nitrogen forms oxides that are both acidic and contributing to the formation of secondary air contaminants like ozone. NO_x formation in engines is thus dependent on the combustion conditions as well as the operational mode of the ship. Low-NO_x emissions are observed using methane (LNG or LBG) or methanol, but still the combustion technology and operation of the engine are important to achieve this. Exhaust gas recirculation (EGR) is one technology that reduces the emissions. NO_x can also be removed after combustion by a catalytic technology, SCR, where the NO_x is reduced back to nitrogen gas. A limitation of the SCR technology is that it can only be used when the exhaust gas is warm enough, (> 200°C) and thus the NO_x emissions will be high when the engine is started up, and often also for some time after leaving port.

Particles

Particles are formed in the combustion and also in the exhaust gas (“secondary particles”). There is a strong correlation between sulfur contents of the fuel and particle formation. The impact of particles varies somewhat with the particle size.

Large particles are deposited close to the source and are visible, “soot.” Small particles (up to 10 μm , PM_{10}) are not as visible, but are spread over larger areas and may constitute a health hazard. Particles are absorbing heat and “black carbon” particles are a contributor to global warming. Particles can also act as catalysts for the formation of secondary air pollutants.

Measurements of particle emissions are mainly performed in terms of mass (g/kWh). However, this is a poor indicator for the small particles, where the number of particles of low weight may be large, but with a small total mass.

Fuel Alternatives

In order to decrease emissions, various types of fuels have been suggested and tested as alternatives to HFO. Fossil-based alternatives are MDO and MGO with low-sulfur content, LNG, and methanol. Renewable alternatives can be biodiesel but also liquefied biogas (LBG), methanol made from renewable material and hydrogen (Brynolf, 2014; Balcombe et al., 2019). Recently, also ammonia (NH_3) has been proposed as a shipping fuel. The production of synthetic fuels from (renewable) electricity and CO_2 , “electrofuels” is also being investigated. Electricity is increasing as an energy carrier for ships, especially in the short sea-shipping segment in the form of hybrid, but also as fully electric, vessels.

The outcome of a change of fuel in terms of emissions varies depending on the fuel and the propulsion technology. Converted traditional engines can be used, but engines developed and optimized for the specific fuel may give better performance. There are two principal kinds of fuel in terms of storage, liquid and gas. A change from liquid to a gaseous fuel sets demands on fuel handling systems and also on regulations.

The fuel alternatives are in many cases the same as discussed for other transport sectors. However, for shipping some additional focus has been on methanol and recently also on ammonia, two chemicals that are widely used in industry and having a well-developed global production and distribution. The production of both is today based on natural gas. However, the production may be using renewable electricity, “electrofuel.” In this production hydrogen is an intermediate step. Methanol and ammonia can also be used as a “hydrogen carrier” from which hydrogen can be released for use in fuel cells or engines. Also “biomethanol” can be produced from all kinds of raw material, including waste.

Methanol, (CH_3OH), is liquid at ambient temperature, which makes it easy to store and handle. Methanol can be used as a marine fuel in diesel engines as well as in Otto engines. In the diesel engine, a small amount of other fuel like diesel, “pilot fuel,” is needed for ignition. Dual-fuel diesel engines running on methanol/diesel are in use in some ships, for example, a RoPax Ferry between Sweden and Germany. Methanol is toxic to man, setting demands on handling. However, it mixes with water and is biodegradable, making spills into the environment of low and local impact.

Hydrogen has long been considered as a future fuel and is possible to use in both combustion engines and fuel cells. The only direct emission is water vapor when used in fuel cells. However, for the use of hydrogen to reduce climate impacts it is important that hydrogen is produced from renewable resources or from nonrenewable resources in combination with carbon capture and storage. The fuel cell technology has matured and reduced in cost, but there are still challenges, such as the cost of the fuel and with hydrogen storage and infrastructure. Hydrogen gas has a very low volumetric energy density and needs to be compressed, liquefied or bound to other molecules or substances to increase energy content per volume. One way to store hydrogen is to bind the H_2 molecule to nitrogen and form ammonia.

Ammonia (NH_3) is a compound consisting of only nitrogen and hydrogen. Due to the lack of carbon in the molecule, ammonia has recently (2019) been proposed as a future energy carrier for many different purposes, including shipping. Ammonia is a gas at room temperature. (Not to be mixed up with the aqueous solution sold for household purposes). Ammonia as a fuel is not yet tested in modern engines, although engine development has been started. Ammonia has been used as a refrigeration liquid, but is often avoided due to safety reasons. A release of ammonia to air is toxic to humans and has an unpleasant smell in low concentrations, which will be an issue in handling. However, ammonia is quickly decomposed, so the effect is local.

Life Cycle Impact of Fuel

Although the emission regulations only apply to the operation of the ship, the environmental impact is also dependent on what happens upstream, in the production and distribution of the fuel to the ship. A fuel that has low emissions in use, may need a large effort in production, including phases like harvesting (biofuels), transport, reforming, and purification processes. The production of renewable fuels may also include use of fossil energy, making the gain in GHG savings smaller.

The widespread use of HFO in shipping has been the result of a competitive price of fuel. Also the total energy used for production of the HFO is low, adding to the competitiveness. Production of renewable fuels requires more energy input and processing, resulting in a higher price compared to HFO, often 2–3 times higher. A challenge for future renewable fuels is to develop production at low energy input and cost.

Examples of environmental impact and energy use in a well to propeller perspective for some fuels as compared to HFO are given in Fig. 1.

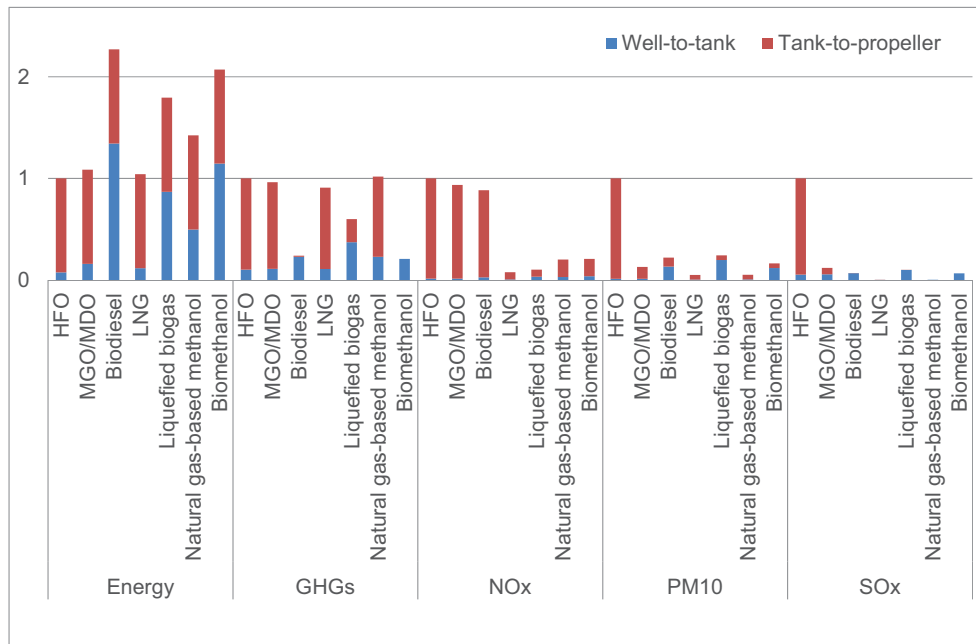


Figure 1 Properties of some alternative fuels from “well-to-propeller” (production and use of fuel) in comparison to HFO (= 1) in terms of total energy input and emissions. Source: Compiled from Brynolf (2014).

Propulsion Alternatives

Electricity/Batteries

The development of electrical propulsion for vehicles has made electricity an interesting energy carrier also for ships. There has been a use of electric propulsion on board ships for a long time as “diesel-electric” propulsion, where the diesel engines generate electricity that is used for propulsion. For short distances, passenger ferries, etc., all-electric propulsion with batteries is coming into use. The use of batteries for onboard recovery and storage of energy is used.

Battery-electric propulsion has a very high-energy efficiency and roughly half the energy is needed for the same propulsive power compared to the use of diesel in internal combustion engines. Despite the higher efficiency, the battery-electric propulsion system is only viable for short distances unless considerable improvements in the energy density of batteries are achieved. Another way to reduce the climate intensity of the propulsion system is to harvest energy directly from the wind and sun (Brynolf et al., 2018).

Wind Assisted

Although shipping has been powered by wind throughout history, the use of wind has been very limited during the last 100 years. Today, however, with the demands for decreased emissions and fossil-free propulsion and with higher fuel costs, various concepts of wind assistance in sailing have been tested. There is very little information on the outcome of the tests.

The use of rigid sails has been tested in various projects since the 1980s, but has not (in 2019) received widespread acceptance as yet. A 5-month test on a 7000 DWT bulk carrier pointed at a fuel saving of up to 30%, which is in agreement with model calculations for use in open sea. There are also “hybrid sail” concepts with combinations of soft and rigid sails.

Wind-assisted sailing using kite sails has also been tested and used. Calculations have shown that there is a theoretical possibility to save between 10% and 50% of the fuel, depending on wind speed, by using a kite on a 50,000 dwt tanker.

Flettner rotors are a way of creating additional thrust by wind. It is a high cylinder, standing vertically and rotating. It is driven by an electric motor. Tests on board have shown that it can deliver more thrust than the main engine under some conditions. Fuel savings of 7–10% have been recorded in on board tests. Theoretical calculations point at a potential fuel saving of up to 20%. Model calculations have shown that the use of Flettner rotors has the potential to provide cost-efficient propulsion, however, the payoff time is strongly dependent on the price of fuel (Tillig et al., 2018).

Nuclear

An alternative source of energy that can be used for propulsion of ships is the nuclear reactor. This has a long history for propulsion of military ships, mainly submarines, but also ice-breakers have been built. In 2016, around 166 naval reactors were in operation worldwide. The nuclear reactor has advantages in that it is emission-free during operation. It also needs refueling

only at very long intervals. The use of nuclear propulsion for merchant ships has been tested in three ships in the 1960s, in Germany, Japan, and the United States. The conclusion from the tests was that the cost was far too high for commercial application. The ships were rebuilt to conventional propulsion or decommissioned. Security, as well as the opinion from some states and ports against calls by nuclear powered commercial ships, constitutes the main limitations on use. One argument for this is that the fuel in these reactors is high enriched and can easily be used in nuclear weapons, which makes the ships a security issue (Schoyen and Steger-Jensen, 2017). In recent years, there have been some initiatives to take up commercial development again, and the regulatory framework has been updated by class societies. At present (March, 2019) there is a tender from China for building a 30,000 tons ice-breaker.

Fuel Cells

Fuel cells are used to convert chemical energy in a fuel into electrical energy and have been tested for use on board ships, especially for generating electricity for on board use. Dependent on the fuel cell technology different fuels can be used. Hydrogen is the most common but it is also possible to use LNG, methanol, and ammonia, for example.

Emissions to Water and Impacts

Oil Spills

Crude oil and various forms of petroleum products, for example, heavy fuel oil and diesel commonly end up in the natural environment. Calculations show that on average, 1,250,000 tons of oil annually end up in the marine environment from sea-based sources. Sources of oil to the marine environment that originate from shipping-based activities are accidents, groundings, operational discharges, and shipwrecks. These sources are responsible for approximately 34% of the total annual input. With better regulations, vessel design, and on-board safety installations accidents leading to large oil spills have decreased over time, however as shipping activities are increasing globally the volumes from operational discharges, for example, leaking-propeller shaft bearings, bilge water, and illegal discharges is increasing (Farrington, 2013).

Environmental Effects of Oil Spills

Oil spills can have two different effects in the aquatic environment, acute toxic or nonlethal effects. The volume of oil, type, location of the spill, time of the year, and weather conditions determines the effect of the spill in the environment. The normal biota at the site of the oil spill and the possibility of these organisms for recolonization also affects the total impact of oil spill.

Effects from single, large oil spills often have lethal effects in organisms. Through narcotic effects the organisms' cellular functions are highly degraded, as cell membranes lose the ability to uphold its normal homeostasis, which have lethal consequences. Organisms with feathers and fur are vulnerable to oil spills, especially seabirds. Even a small amount of oil on the feathering causes a loss of insulation capabilities, leading to hypothermia. Larger amounts cause seabirds to lose buoyancy and their ability for flight. Furthermore, effects from a large oil spill include the physically hindering of oxygen and sunlight transfer into the water caused by the slick. This leads to anoxia in shallow waters and lowered capability for photosynthesis, which entails reduced densities in marine plants and benthic fauna.

Small and continuous oil spills also cause negative effects in the marine environment, and is primarily caused by the most toxic part of oil, Polycyclic Aromatic Hydrocarbons (PAHs). This group of compounds exists in crude oil and petroleum products in various concentrations. PAHs have mutagenic, carcinogenic, dioxin-like, and endocrine-disrupting potential, consequences from exposure to PAHs can be considerable to a wide range of organisms. Exposure can, for example, result in the induction of cancer, decreased ecological and taxonomical community diversity, lowered fecundity and growth, and lowered tolerance to other stresses. Over time effects from a small, continuous oil spill can inflict equal negative effects on a community or population level as a single large oil spill (Lindgren et al., 2017).

Biofouling and Use of Hull Coatings

Hull coatings with antifouling properties is commonly used on ships to deter the accumulation of biological growth (biofouling), for example, microorganisms, algae, and animals on the hull and other surfaces of the ship immersed in seawater. Processes leading up to this growth begins immediately after a surface is lowered into the sea. In the final stage of this process, larvae of macrofoulers, such as barnacles, bryozoans, molluscs, and polychaetes will be under favorable conditions adhere (settle) to the surface and begin to grow. Finally, the surface will be covered with a thick layer of biological growth. The amount of biofouling on the immersed surface will depend on several factors, for example, water temperature, salinity, solar radiation, water depth, and interactions within the fouling community. If the fouling is left uncombated on the hull this will significantly increase the ship's hull roughness and drag, which will result in increased weight, reduced speed, and decreased maneuverability, effects which must be compensated for by increased fuel consumption. Another consequence is the possibility of alien species getting attached to the hull and thereby transferred to new areas.

Throughout history, seafarers have attempted to protect their hulls from fouling organisms. Indications exist that people used antifouling systems to protect ships even before 1000 B.C. As antifouling, in general some sort of toxic substance, biocide, is used to prevent organisms from attaching to the hull. In the 1960s Tri-Butyl-Tin (TBT) became the dominating biocide in antifouling paints and was so until 2008, when the substance was prohibited due to its negative effects also to nontarget organisms.

In modern hull coatings copper in various chemical compounds is used to deter fouling organisms. Copper is a metal that commonly occurs in the environment and living organisms need it in small amounts and use it for growth. However, copper is toxic to organisms at high concentrations and can occur in several different chemical states. The most toxic form of copper to living organisms is in its ionic form, Cu^{2+} . This easily passes through cell membranes and causes damage. The usage of copper in antifouling paints is often in the form of cuprous oxide, copper thiocyanate, or metallic copper, all leaking copper in ionic form. This leakage of copper ions then has an antifouling effect on for example barnacles, tube worms, and some species of algae. The leakage rate of copper from the coating is governed by several factors, for example, paint formula, age of the paint on the hull, the water temperature, and the pH. Due to both leisure boating and commercial shipping, high concentrations of dissolved copper are commonly found in shallow, near-coastal marine areas. Copper also adsorbs to particulate matter present in the water column, with the consequence of an accumulation in sediments. These shallow coastal areas are at the same time used as spawning, nursery, and feeding grounds by numerous marine organisms. The potential impact of copper on nontarget species is therefore substantial.

Ballast Water

Ballast water taken up to stabilize the ship when operating with less-cargo load, will include the marine organisms at the place of intake. This can be both larval stages of larger organisms as well as smaller planktonic animals, which will get transferred between ports in different geographic areas. When a new species is introduced in a sea area the consequence will be dependent on several factors. For example, if the introduced species is outcompeting native species by being fast in nutrient uptake and growth and if it is tolerant to physical factors like temperature and salinity, it might become invasive. The introduction can on the contrary be limited, for example due to predation by present species or by mismatch in reproduction requirements. The problem of invasive species transfer with ballast water was highlighted by researchers at the end of the 1990s and in 2004 the IMO Ballast Water Management Convention (BWMC) was adopted with the aim to limit the issue. In the BWMC the requirement of treatment to remove marine organisms from the water before discharge is described. The BWMC has been in force since September 2018.

Examples of invasive species transferred by shipping with great ecological impact and high mitigation costs is, for example, the *zebra mussel* that clog locks and water intakes, *Chinese mitten crab* that erode sediment banks and spread parasites, microalgae that can form toxic algal blooms and cholera bacteria that can start outbreaks.

Waste Generated on Board

Various kinds of waste are generated on board ships. This includes operational waste like oil/water mixes as well as waste generated by crew and passengers like sewage water and food waste. The tradition of dumping waste into the sea is old, and still in the 1960s there were recommendations of how to sink waste from leisure boats in a safe way. In general, there are IMO regulations prohibiting release to the sea, with some exceptions in MARPOL annex V. The exceptions include release of food waste outside special and arctic areas, if ground to smaller fractions at a certain distance (usually 12 nm) from the shore. The regulations put demands on the port reception facilities for waste.

Noise

Noise is also present underwater. Fish, invertebrates, and mammals are using sounds for communication, sometimes at frequencies that are similar to those generated by ship propellers and engines. Different effects can be the result of anthropogenic noise. First, permanent and temporary hearing loss can be a consequence. Then modifications in the amplitude and frequency that are needed to detect sound have been the negative result from the noise. Secondly, masking can occur. Then alien sounds cover a desired signal, rendering detection more difficult, which can hinder communication, mating calls, and the search for prey. Finally, behavioral changes in response to a sound, which can include the abandonment of an important activity, for example, feeding, nursing, or the use of an area, can be the response to a sound (Andre, 2018).

Particularly Sensitive Areas, Baltic Sea, Arctic, and Others

Some marine areas are classified by the IMO as Particularly Sensitive Sea Areas (PSSAs) for ecological, socioeconomic, or scientific reasons. These areas are considered as being in risk of damage from maritime activities. Actions like stricter environmental regulations (The Baltic Sea) or routing systems (The Galapagos Islands) are taken. There are currently 17 areas classified as PSSAs. The IMO MARPOL regulation does also include Special Areas with stricter discharge limits under its six Annexes for different pollutants.

Ships trading in the polar regions are since 2017 regulated under the IMO Polar Code. The Polar Code include safety and environmental provisions ranging from design rules to navigational advice like for example prohibitions on the release of oily mixtures and prevention of pollution from garbage and noxious liquid substances. Areas with stricter national regulation can be exemplified by Norway that targets zero emissions in the fjords, US-Alaska with strict wastewater treatment requirements and New Zealand and Australia with special biosafety regulations to avoid transfer of invasive species with hull fouling.

Sustainable Shipping in the Future

After a couple of decades with stricter regulations of emissions to sea and air, there are still more challenges to address to achieve an environmentally sustainable shipping sector. Shipping has a large challenge, which it shares with other transport and energy sectors, in becoming fossil-free in the coming decades. In addition, there is a need for a general decrease of emissions, and a “zero vision” for shipping emissions is discussed as a goal. Although regulations and technical improvements are making the environmental impact from shipping less detrimental to the environment, geographical areas and/or toxic chemicals are at risk of becoming new or increased problems in the future. Future shipping will most probably use many different technologies and fuels in order to strive toward the zero-emission vision.

Biographies

Karin Andersson is a professor emerita in Maritime Environmental Sciences at Chalmers University of Technology. Starting with M.Sc. in Chemical Engineering and a Ph.D. in Nuclear Chemistry, Karin's research is focusing on environmental and energy assessments of technical systems. In the last 10 years the main focus has been on the relation between shipping and the natural environment, providing decision support for industry, and policy makers.

Selma Brynolf, Ph.D., is a researcher at the Department of Mechanics and Maritime Sciences at Chalmers University of Technology. Selma's research is focused on assessing future fuels and propulsion technologies for transport with focus on shipping using tools such as life cycle assessment, multicriteria decision analysis and global energy system modeling.

Lena Granhag is an associate professor in Maritime Environmental Science at Chalmers University of Technology. Lena's background is in marine ecology and she works with questions related to the impact of shipping on the marine environment. Lena's main research interest is in ship hull biofouling, transfer of invasive species with shipping, and in finding sustainable solutions to these issues.

J. Fredrik Lindgren, Ph.D. is a senior adviser at the Swedish Agency for Marine and Water Management and researcher at the Department of Mechanics and Maritime Sciences at Chalmers University of Technology. Fredrik's research has been focused on ecotoxicological effects of the shipping industry, for example, operational oil spills, biocides from antifouling technologies, and dumped-chemical warfare agents.

See Also

Energy Efficiency of Ships; Arctic Shipping

Relevant Websites

The International Maritime Organisation, International convention for prevention of pollution from ships (MARPOL). Available from: [http://www.imo.org/en/About/Conventions/ListOfConventions/Pages/International-Convention-for-the-Prevention-of-Pollution-from-Ships-\(MARPOL\).aspx](http://www.imo.org/en/About/Conventions/ListOfConventions/Pages/International-Convention-for-the-Prevention-of-Pollution-from-Ships-(MARPOL).aspx)

References

- Andre, M., 2018. Ocean noise: making sense of sounds. *Soc. Sci. Info. Sur Les Sciences Sociales* 57 (3), 483–493.
- Balcombe, P., Brierley, J., Lewis, C., Skatvedt, L., Speirs, J., Hawkes, A., Staffell, I., 2019. How to decarbonise international shipping: options for fuels, technologies and policies. *Ener. Conv. Manag.* 182, 72–88.
- Brynolf, S., 2014. Environmental assessment of present and future marine fuels, unpublished thesis Chalmers University of Technology.
- Brynolf, S., Taljegard, M., Grahn, M., Hansson, J., 2018. Electrofuels for the transport sector: a review of production costs. *Renew. Sust. Ener. Rev.* 81, 1887–1905.
- Farrington, J., 2013. Oil pollution in the marine environment inputs, big spills, small spills, and dribbles. *Environ.* 55 (6), 3–13.
- IMO., 2018. Initial IMO strategy on reduction of GHG emissions from ships. RESOLUTION MEPC.304 (72).
- Lindgren, J.F., Hassellöv, I.M., Nyholm, J.R., Östin, A., Dahllöf, I., 2017. Induced tolerance in situ to chronically PAH exposed ammonium oxidizers. *Mar. Pollut. Bull.* 120 (1–2), 333–339.
- Schoyen, H., Steger-Jensen, K., 2017. Nuclear propulsion in ocean merchant shipping: the role of historical experiments to gain insight into possible future applications. *J. Clean. Prod.* 169, 152–160.
- Smith, T.W.P., Jalkanen, J.P., Anderson, B.A., Corbett, J.J., Faber, J., Hanayama, S., O'Keeffe, E., Parker, S., Johansson, L., Aldous, L., Raucci, C., Traut, M., Ettinger, S., Nelissen, D., Lee, D.S., Ng, S., Agrawal, A., Winebrake, J.J., Hoen, M., Chesworth, S., Pandey, A., 2014. Third IMO GHG Study. International Maritime Organisation (IMO), London.
- Tillig, F., Ringsberg, J., Mao, W., Ramne, B., 2018. Analysis of uncertainties in the prediction of ships' fuel consumption - from early design to operation conditions. *Ships Offshore Struct.* 13, 13–24.

Further Reading

- Andersson, K., Brynolf, S., Lindgren, J.F., Wilewska-Bien, M., 2016. *Shipping and the Environment*. Springer, Berlin.
- Andre, M., 2018. Ocean noise: making sense of sounds. *Soc. Sci. Info. Sur Les Sciences Sociales* 57 (3), 483–493.
- Corbett, J.J., Lack, D.A., Winebrake, J.J., Harder, S., Silberman, J.A., Gold, M., 2010. Arctic shipping emissions inventories and future scenarios. *Atmos. Chem. Phys.* 10 (19), 9689–9704.
- Corbett, J.J., Winebrake, J.J., Green, E.H., Kasibhatla, P., Eyring, V., Lauer, A., 2007. Mortality from ship emissions: a global assessment. *Environ. Sci. Technol.* 41 (24), 8512–8518.
- DNV-GL, 2018. Assessment of selected alternative fuels and technologies. Hamburg.
- GESAMP, 2007. Reports and studies 75: Estimates of oil entering the marine environment from sea-based activities. London.
- Gilgen, A., Huang, W.T.K., Ickes, L., Neubauer, D., Lohmann, U., 2018. How important are future marine and shipping aerosol emissions in a warming Arctic summer and autumn? *Atmos. Chem. Phys.* 18 (14), 10521–10555.
- Larkin, A., Smith, T., Wrobel, P., 2017. Shipping in changing climates. *Mar. Policy* 75, 188–190.
- Lloyds Register and UMAS, 2017. *Zero-emission Vessels 2030. How do we get there?* Lloyds Register, London.
- Noor, C., Noor, M., Mamat, R., 2018. Biodiesel as alternative fuel for marine diesel engine applications: a review. *Renew. Sust. Ener. Rev.* 94, 127–142.
- Sofiev, M., Winebrake, J.J., Johansson, L., Carr, E.W., Prank, M., Soares, J., Corbett, J.J., 2018. Cleaner fuels for ships provide public health benefits with climate tradeoffs. *Nat. Commun.* 9 (1), 406.
- Traut, M., Larkin, A., Anderson, K., McGlade, C., Sharmina, M., Smith, T., 2018. CO₂ abatement goals for international shipping. *Clim. Policy* 18 (8), 1066–1075.
- Yebra, D.M., Kiil, S., Dam-Johansen, K., 2004. Antifouling technology - past, present and future steps towards efficient and environmentally friendly antifouling coatings. *Prog. Org. Coat.* 50 (2), 75–104.

Energy Efficiency of Ships

Harilaos N. Psaraftis, Technical University of Denmark, Kongens Lyngby, Denmark

© 2021 Elsevier Ltd. All rights reserved.

Introduction	294
The Energy Efficiency Design Index (EEDI)	294
The MBM Track	296
The Way Ahead	297
Acknowledgments	298
References	298

Introduction

International shipping is currently at a crossroads. The decision of the 72nd session of the Marine Environment Protection Committee (MEPC 72) of the International Maritime Organization (IMO) in April 2018 to achieve by 2050 a reduction of at least 50% in maritime greenhouse gas (GHG) emissions vis-à-vis 2008 levels epitomizes the last among a series of recent developments as regards sustainable shipping. It also sets the scene on what may happen in the future. The so-called Initial IMO Strategy (IMO, 2018) is in the form of Resolution MEPC.304(72), and includes, among others, the following elements: (1) the vision, (2) the levels of ambition, (3) the guiding principles, (4) a list of short-term, medium-term, and long-term candidate measures with a time line, and (5) miscellaneous other elements, such as follow-up actions and others.

Against this background, the purpose of this paper is to present some basics as regards the energy efficiency of ships, including related regulatory activity at the IMO and elsewhere. This paper mainly draws from prior work by this author and his colleagues.

Before we proceed, a clarification is in order, and this concerns the definition of the term “energy efficiency.” In economics, efficiency typically means optimal allocation of resources to achieve a certain (multi-objective) goal. A solution is called efficient if it is impossible to improve one of the objectives without deteriorating another objective. Such a solution is also called Pareto-optimal. In that sense, the term “efficient” has a binary meaning, which means that a solution is either efficient or not efficient, and there is no sense of saying that a solution is more efficient than another solution.

In transportation, including shipping, “energy efficiency” is not binary, but can be measured. Energy efficiency is typically defined as the ratio of the amount of energy required to perform a certain amount of transport work, appropriately defined, divided by that transport work. In that sense, a solution can be said to be more energy efficient than another solution if it achieves a lower such ratio, that is, uses less energy to perform the same transport work. In shipping, energy efficiency is often associated with various indices, such as the Energy Efficiency Design Index (EEDI) or the Energy Efficiency Operational Indicator (EEOI), among others (definitions will follow).

For a specific type of fuel, since reducing the energy required to propel a ship is tantamount to reducing the amount of fuel consumed and since the latter is proportional to GHG emissions produced, achieving energy efficiency in shipping typically means reducing GHG emissions. In that sense, we shall not restrict ourselves to discussing EEDI and EEOI, but will also cover other measures to reduce GHG emissions, such as market-based measures (MBMs), and comment on the Initial IMO Strategy. Reducing GHG emissions can also be achieved through alternative (low carbon or zero carbon) fuels. For instance, nuclear propulsion will yield zero GHG emissions. However, in this paper we shall focus on instruments that can reduce GHG emissions assuming a specific carbon footprint for the fuel used. For a more detailed discussion of the issues that pertain to the decarbonization of maritime transport, see Psaraftis (2018).

The rest of this paper is organized as follows: section “The Energy Efficiency Design Index (EEDI)” introduces EEDI and also talks about EEOI. Section “The MBM Track” discusses MBMs. Finally, section “The Way Ahead” comments on the way ahead, including the Initial IMO Strategy.

The Energy Efficiency Design Index (EEDI)

The so-called EEDI, adopted by the IMO in 2011 as an amendment of MARPOL Annex VI, is thus far the most important (and in fact thus far the only mandatory) regulatory instrument for maritime GHG emissions reduction. EEDI basically aims to induce changes at the technological/design level that would bring about GHG reduction in the world fleet. Technological measures to reduce GHGs include optimized hulls, more efficient engines and propellers, and a variety of possible solutions such as wind-assisted propulsion, air lubrication, waste heat recovery, and others.

For a specific ship of 400 GRT and above, and built as of January 1, 2013, its EEDI is computed by the following formula (IMO, 2012):

$$\frac{\left(\prod_{j=1}^n f_j \right) \left(\sum_{i=1}^{nME} P_{ME(i)} \cdot C_{FME(i)} \cdot SFC_{ME(i)} \right) + (P_{AE} \cdot C_{FAE} \cdot SFC_{AE*}) + \left(\left(\prod_{j=1}^n f_j \cdot \sum_{i=1}^{nPTI} P_{PTI(i)} - \sum_{i=1}^{neff} f_{eff(i)} \cdot P_{AEff(i)} \right) C_{FME} \cdot SFC_{AE} \right) - \left(\sum_{i=1}^{neff} f_{eff(i)} \cdot P_{eff(i)} \cdot C_{FME} \cdot SFC_{ME**} \right)}{f_i \cdot f_c \cdot Capacity \cdot f_w \cdot V_{ref}} \quad (1)$$

The numerator in Eq. (1) is the total CO₂ emissions produced by the ship and is a function of all power generated by the ship (main engine and auxiliaries). The denominator is a measure of transport work and is a product of the ship's capacity (usually deadweight) and its "reference speed," defined as the speed corresponding to 75% of maximum continuous rating (MCR), the maximum power of the ship's main engine. The units of EEDI are grams of CO₂ per ton mile.

The EEDI of a specific ship, also known as attained EEDI, as computed earlier, is to be compared with the so-called EEDI (reference line), which is only a function of ship type and DWT (deadweight) and is defined as follows:

$$EEDI(\text{reference line}) = aDWT^{-c} \quad (2)$$

In Eq. (2), a and c are known positive coefficients that have been determined by regression from the world fleet database, and have been finalized for each major ship type after a long debate within the IMO. For a given ship, the requirement for EEDI compliance is that the attained EEDI value should be equal to, or less than, the so-called *required EEDI* value. The required EEDI value is proportional to the value of EEDI (reference line), as defined in Eq. (2), and the requirement can be written as follows:

$$\text{Attained EEDI} \leq \text{Required EEDI} = (1 - X/100) aDWT^{-c} \quad (3)$$

where X is a reduction factor ranging from $X = 0\%$ for ships built from 2013 to 2015, $X = 10\%$ for ships built from 2016 to 2020, $X = 20\%$ for ships built from 2020 to 2025 and $X = 30\%$ for ships built from 2025 to 2030. The year of build is defined as the year the ship's keel is laid. The rationale for factor X is the wish of IMO policy-makers to see newer ships becoming more energy efficient in the future, the latter being defined as having a lower EEDI than ships of similar type and size built earlier. In that sense, IMO policy-makers assume that EEDI compliance is a proxy for achieving GHG emissions reductions. However, things are not that simple and it turns out that several factors sometimes make the earlier assumption hard to justify.

First of all, commercial ships do not necessarily trade at the speed corresponding to 75% of MCR, or at any other predetermined speed. Whoever pays for the fuel (owner or charterer) will select an appropriate speed, which is a function of basically two factors: fuel price and freight rate. High fuel prices and/or low freight rates will induce slower speeds and hence lower fuel consumption. Conversely, low fuel prices and/or high freight rates will induce ships to speed up. Slow steaming, a much prevalent practice these days, may involve speeds drastically lower than the 75% MCR speed, and the corresponding reduction in fuel consumption will be even higher.

Second, fuel consumption is very much a function of prevailing weather conditions, and ship A that has a lower EEDI than ship B under calm weather conditions may have a far poorer fuel consumption than ship B in actual weather conditions. The behavior in other than calm weather conditions cannot be captured by EEDI, as defined earlier. An IMO delegation reported that even identical sister ships built by the same yard in the same period as part of a series program have EEDIs varying between 8% and 10%, the sole reason being different entries for the design speed recorded in commercial fleet databases. Such information is reported by ship owners or other sources, and is not independently verified.

But by far the most important deficiency of EEDI is what we call the "horsepower limit" deficiency. In Eq. (1), the traditional assumption, which comes from marine hydrodynamics, is that the ship engine's MCR, which is in the numerator of Eq. (1), grows to the cube of speed, V^3 , where V is the reference speed that appears in the denominator of Eq. (1). This means that EEDI grows like V^2 . In Eq. (3), speed does not enter the formula at all. It is straightforward to check that this combination is tantamount to imposing an upper bound on speed, corresponding to 75% MCR, and this would translate to an upper bound on MCR itself. Thus, in the quest of EEDI compliance, one would run the risk to see the construction of underpowered ships, which would be less safe to navigate and which, in their attempt to go faster or just maintain speed in bad weather, would emit more CO₂. Perhaps more important, this might also shift the focus of action from designing the best possible hull forms, engines, or propellers, which is the intended aim of EEDI, to the easy solution: just reduce MCR, or equivalently, speed at the design level. This means that any totally inefficient design could be made acceptable with the easy way out: a reduction in horsepower, or design speed. The existence of the earlier easy way out can hardly serve as an incentive for more efficient future ship designs.

The horsepower deficiency has been recognized at the IMO. Since 2011, a serious discussion has ensued, both within MEPC and MSC (IMO's Maritime Safety Committee), on how to reconcile EEDI compliance with the minimum safe power requirement. But the approach for such a reconciliation does not involve modifying the EEDI formulation. Recent documents submitted to the IMO suggest that an impasse cannot be ruled out. The impasse revolves around the dilemma between (1) seriously downgrading the requirements for ships so as to be able to safely navigate in adverse weather conditions, and (2) admitting that the whole EEDI formulation needs to be radically modified. The latter is already the case for Ro-Ro vessels. For these, the EEDI formulation is more complex, as some coefficients in the EEDI formula are not constant. The EEDI for this type of vessel was thus reconsidered by the IMO, as some industry stakeholders reported problems with the current formulation. In a recent study on CO₂ reduction targets and

related pathways (Smith et al., 2016) it was shown, among other things, that the existence of EEDI versus a scenario in which there is no EEDI as we move to 2050 amounts to a GHG emissions difference of about 3%.

EEDI being only a *design* index, its operational equivalent is the so-called EEOI, which is a *voluntary* index with a formula very similar to that of EEDI but, instead of DWT in the denominator, the actual amount of cargo carried is used. Also, total CO₂ emitted at the actual voyage speed (and not at 75% MCR) is estimated.

It turns out that there are several problems with EEOI as well, rendering it an unreliable indicator of operational efficiency. First, the effect of bad weather where a ship increases fuel consumption to keep a certain speed penalizes the EEOI value. In addition, a voyage with less than maximum cargo, or with zero cargo (on ballast), heavily worsens the EEOI value. Also, bad weather and ballast voyages are mostly out of the control of the operator and their effect on EEOI may be much larger than any best practices applied by the ship owner (e.g., course optimization, frequent hull and propeller cleaning, etc.) An energy efficient ship with the most prudent operator may have a good EEOI one year and a bad EEOI next year, thanks to the whims of nature and the markets.

Based on all of the aforementioned, the concept of EEDI (and by extension the EEOI), which is at this point in time the only mandatory instrument to reduce GHG emissions from ships, even though formulated and implemented with the noblest of intentions, suffers from some basic deficiencies. These will have to be resolved if one is to see a credible dent in GHG emissions in the future. However, how or if these deficiencies will be resolved seems pretty much open at this point in time. For a discussion of EEDI/EEOI including associated problems see Polakis et al. (2019) and for an alternative EEDI formulation taking into account weather conditions see Lindstad et al. (2019).

The MBM Track

The second most important instrument that the IMO has considered in order to curb GHG emissions has been the class of MBMs. By and large, and following the “compartmentalization” process that is prevalent in many of these discussions at the IMO and other fora, the MBM track has run in parallel to (and has been independent of) the EEDI track, even though there have been cross-linkages between the two, as some MBM proposals embedded EEDI into their formulation.

MBMs are instruments that apply the “polluter pays” principle and internalize the external costs of GHG emissions. In the short run an MBM (for instance, a bunker levy) can induce speed reduction and other logistical adjustments and in the long run it can incentivize technological advances (such as better hulls, engines, propellers and other technologies, and/or low carbon fuels). In both cases, GHG emissions would be reduced.

In 2010, an IMO Expert Group, appointed by the IMO Secretary General, and chaired by none other than the (then) Chairman of the MEPC, was tasked to evaluate as many as 11 separate MBM proposals, submitted by various member states and other organizations. These included a levy scheme, several separate Emissions Trading Systems (ETS), and various other proposals. All MBM proposals described schemes that would target GHG reductions through either in-sector emissions reductions from shipping or out-of-sector emissions reductions through the collection of funds and the spending of such funds to reduce emissions outside the maritime sector (for instance, building a wind farm in New Zealand or a solar farm in Indonesia). By making shipowners pay for their ships’ CO₂ emissions, an MBM is an instrument that can implement the “polluter pays” principle. In that sense, it may help internalize the external costs of these emissions. The IMO formulated a list of criteria for the evaluation of the MBM proposals, including environmental effectiveness, practical feasibility, administrative burden, compatibility with existing legal frameworks, and others.

The MBM Expert Group met on several sessions and produced a 300-page report with a detailed analysis of the 11 MBM proposals, including a discussion of alternative scenarios as regards fuel prices, projected emissions growth, and other parameters (IMO, 2010). The group’s modeling effort, which also involved the work of external consultants, was to develop and apply a model to make quantitative estimates of emissions reductions, revenues generated, costs, and other attributes of each MBM proposal. The group’s report contained no recommendation on any specific MBM proposal that could be chosen, and did not even contain a short list of MBMs, keeping all of them on the table. In fact, discussion on MBMs at the IMO level after 2010 was not very productive. The period to July 2011 focused on the adoption of EEDI, and not much was done on MBMs. The period immediately after the adoption of EEDI focused on practical matters involving its implementation and again there was little discussion on MBMs. A proposal by Greece in 2012 (who had submitted no MBM proposal of its own) for the IMO to decide on a short list of MBMs (essentially a bunker levy and ETS) was rejected. All MBMs continued to be on the table, except one that was withdrawn.

In addition to the almost complete lack of consensus among the MBM proposers (except for those promoting ETS who had a common position), the same group of developing countries (China, India, Brazil, Saudi Arabia, etc.) were as much against any MBM as they were against EEDI, particularly after they lost the EEDI vote in 2011. Their main objection was mainly on the grounds that MBMs were not compatible with the principle of *Common But Differentiated Responsibilities and Respective Capabilities (CBDR-RC)* (Stone, 2004). Then in May 2013, the MEPC decided to suspend discussions on MBMs. This was accompanied by a channeling of the discussion toward the subject of monitoring, reporting and verification (MRV) of CO₂ emissions. Today, some 6 years later, the MBM discussion at the IMO remains suspended, with no visible sign of reappearance any time soon.

A later twist in the MBM saga came with the February 2017 vote of the European Parliament (EP) to include shipping into the EU ETS as of 2023, in case no global agreement is reached by 2021. As mentioned earlier, this caused serious concern among industry stakeholders that such a *regional* MBM would create serious distortions, not to mention that it might not necessarily reduce maritime CO₂ emissions. In November 2017, and after some negotiations between the EP and the EU Council of Ministers, it was agreed to

align the EU with the IMO process, and essentially refrain from taking action on ETS before seeing what the IMO intends to do on GHGs. Industry circles, concerned with the effects of an early EU ETS, welcomed this development. However, the European Commission will closely monitor the IMO process, starting from what is agreed on the initial strategy in 2018 and all the way to 2023. Whether or not this latest agreement at the EU level might put some pressure on the IMO to resume the suspended discussion on MBMs and adopt a global MBM before the EU moves on ETS is unclear at this time. And even though the ETS looks like the default scenario for the EU if progress at the IMO is not deemed satisfactory, precisely what action the EU will take and when that action will be taken is equally unclear.

In April 2018, the IMO/MEPC reached the landmark decision to adopt an initial strategy for GHG reductions that sets (among other things) a target of GHG reductions of at least 50% by 2050, vis-à-vis 2008 levels (IMO, 2018). A broad list of potential measures has also been proposed, broken down as short-term (finalized and agreed to up to 2023), medium-term (2023–30), and long-term (2030 and beyond). MBMs are included in the set of *medium-term* measures, but only obliquely. To quote from the decision text, “new, innovative emission reduction mechanism(s), possibly including Market Based Measures (MBMs), to incentivize GHG reduction.” No measures have been specified to be adopted until 2023 at the earliest, the year when the strategy is supposed to be finalized. Other than the earlier wording, and as this paper was being finalized, there is nothing in the IMO process that would reopen the MBM discussion anytime soon.

How can MBMs help with the recently adopted ambitious IMO GHG reduction targets? Irrespective of the fact that MBMs seem not to be up for discussion in the foreseeable future, one idea that might be worth considering would be to impose a significant bunker levy on a global level. By significant we mean not 10 or 20 USD per ton of oil, as is being occasionally contemplated by industry, but at least one order of magnitude higher. This would induce both technological changes in the long run and logistical measures in the short run. In the long run, it would lead to changes in the global fleet toward vessels and technologies that are more energy efficient, more economically viable and less dependent on fossil fuels than those today.

How much CO₂ can be reduced by a substantial global bunker levy? Devanney (2010) estimated that with a base BFO price of USD 465/ton, a USD 50/ton bunker levy would achieve a 6% reduction in total Very Large Crude Carrier (VLCC) emissions over their life cycle and that for a USD 150/ton levy the reduction would be 11.5%. Some estimates of CO₂ reductions for tankers (Gkonis and Psaraftis, 2012) and handymax bulk carriers (Kapetanidis et al., 2014), and for several bunker levy scenarios, were also made. These estimates showed CO₂ reductions of more than 50% for a single VLCC if fuel price rises from 400 to 1000 USD/ton. However, the long-term fleet-level impacts of substantial levies are by and large unknown. For a discussion of the MBM concept and of the various MBMs proposed to the IMO, see Psaraftis (2012) and Psaraftis and Woodall (2019).

The Way Ahead

The reception of the April 2018 Initial IMO Strategy was almost universally laudatory. Not only industry associations but also the European Commission, the Organisation of Economic Cooperation and Development, and several nongovernmental organizations hailed the result as an important first step toward the eventual full decarbonization of shipping. There were very few expressions of dissatisfaction. At the same time, the following can be said.

First, the two stated principles that are centrally included in the Initial IMO Strategy, that is, (1) nondiscrimination/no more favorable treatment and (2) CBDR-RC, are in direct conflict with one another. The latter principle was included so as to please the group of developing countries who stood and continue to stand firmly behind CBDR-RC. Some would suggest, however, that one will need to find a way to circumvent or even eliminate the CBDR-RC principle altogether if any serious progress is to be made.

Second, that GHG emissions are to reach a peak as soon as possible is a laudable goal, but raises the obligatory question, how soon. According to the third IMO GHG study (IMO et al., 2014), in 2008, the baseline year as far as comparison to target values is concerned, CO₂ emissions of international shipping were estimated at 920.9 million tons, and declined to 795.9 million tons in 2012, even though they reached 849.5 million tons in 2011. As of yet, and pending the fourth IMO GHG study (which will be commissioned in 2019 and finalized in 2020), there is no consensus on GHG emissions figures after 2012. And even the earlier figures are based on the “bottom-up” (activity-based) method, whereas emissions figures based on the “top-down” (fuel sales-based) method are significantly lower (624.9 million tons in 2008 vs. 648.9 million tons in 2011—there was no top-down estimate for 2012). There should certainly be consensus on which method is used (“bottom-up” numbers are 30%–50% higher than “top down”), plus consensus on when the GHG peak is expected to occur. Barring any major world trade slowdown, it seems self-evident that for any GHG peak to be reached some measures will have to be implemented—no peak will happen by itself.

“Speed reduction” (or mandatory speed limits) is included as a potential short-term measure, even though the term “speed optimization” was added so as to make the measure more palatable to some South American countries who are concerned about their agricultural exports to Asia. At first glance, speed limits may seem like a reasonable measure. However, they are plagued by various deficiencies and would create distortions and other problems. For a comparison between GHG emissions reductions that can be achieved by a bunker levy and those that can be achieved by speed limits, see Psaraftis (2019).

Perhaps more important, there is not yet a sense of priority among the wide array of candidate measures, all of which are on the table. Indicative is the fact that, after a fierce debate at MEPC 73 (October 2018), the updated plan all the way to MEPC 80 in 2023 changed the wording from “prioritization” to “consideration” of the candidate proposals. Is this really the message that one would like to convey as a clear indication that one means business on the subject of energy efficiency of ships?

For a more thorough discussion on what may lie ahead as regards sustainable shipping and specifically the quest to improve the energy efficiency of ships, see [Psaraftis and Zachariadis \(2019\)](#).

Acknowledgments

The author would like to thank Poul Woodall of DFDS for his comments on a previous version of this paper.

References

- Devanney, J., 2010. The impact of EEDI on VLCC design and CO₂ emissions. Technical Report. Center for Tankship Excellence, USA.
- Gkonis, K.G., Psaraftis, H.N., 2012. Modelling tankers' optimal speed and emissions. Archival Paper. 2012 SNAME Trans. 120, 90–115.
- IMO, 2010. Full report of the work undertaken by the expert group on feasibility study and impact assessment of possible market-based measures. IMO doc. MEPC 61/INF.2. International Maritime Organization, London, UK.
- IMO, 2012. Resolution MEPC.212(63) (adopted on March 2, 2012). 2012 Guidelines on the method of calculation of the attained Energy Efficiency Design Index (EEDI) for new ships. International Maritime Organization, London, UK.
- IMO, 2018. Resolution MEPC.304(72) (adopted on 13 April 2018). Initial IMO Strategy on Reduction of GHG Emissions from Ships, IMO doc. MEPC 72/17/Add.1, Annex 11. International Maritime Organization, London, UK.
- IMO, Smith, T.W.P., Jalkanen, J.P., Anderson, B.A., Corbett, J.J., Faber, J., Hanayama, S., O'Keefe, E., Parker, S., Johansson, L., Aldous, L., Raucci, C., Traut, M., Ettinger, S., Nelissen, D., Lee, D.S., Ng, S., Agrawal, A., Winebrake, J., Hoen, M., Chesworth, S., Pandey, A., 2014. Third IMO GHG Study 2014. International Maritime Organization, London, UK.
- Kapetanios, G.N., Gkonis, K.G., Psaraftis, H.N., 2014. Estimating the operational effects of a bunker levy: the case of Handymax Bulk Carriers. TRA 2014 Conference. Paris, France, April 2014.
- Lindstad, E., Borgen, H., Eskeland, G., Paalsson, C., Psaraftis, H.N., Turan, O., 2019. The need to amend IMO's EEDI to include a threshold for performance in waves (realistic sea conditions) to achieve the desired GHG reductions. *Sustainability* 11, 3668, doi:10.3390/su11133668.
- Polakis, M., Zachariadis, P., de Kat, J., 2019. The energy efficiency design index. In: Psaraftis, H.N. (Ed.), *Sustainable Shipping: A Cross-Disciplinary View*. Springer, Berlin.
- Psaraftis, H.N., 2012. Market based measures for green house gas emissions from ships: a review. *WMU J. Mar. Affairs* 11, 211–232.
- Psaraftis, H.N., 2018. Decarbonization of maritime transport: to be or not to be? *Mar. Econ. Logistics* 21, 353–371, doi:10.1057/s41278-018-20180098-8.
- Psaraftis, H.N., 2019. Speed optimization vs speed reduction: the choice between speed limits and a bunker levy. *Sustainability* 11, 1–18, doi:10.3390/su11080000.
- Psaraftis, H.N., Woodall, P., 2019. Reducing GHGs: the MBM and MRV agendas. In: Psaraftis, H.N. (Ed.), *Sustainable Shipping: A Cross-Disciplinary View*. Springer, Berlin.
- Psaraftis, H.N., Zachariadis, P., 2019. The way ahead. In: Psaraftis, H.N. (Ed.), *Sustainable Shipping: A Cross-Disciplinary View*. Springer, Berlin.
- Smith, T., Raucci, C., Haji Hosseinloo, S., Rojon, I., Calleya, J., De la Fuente, S., Wu, P., Palmer, K., 2016. CO₂ emissions from international shipping. Possible reduction targets and their associated pathways. Report prepared by UMAS, London, UK.
- Stone, C.D., 2004. Common but differentiated responsibilities in international law. *Am. J. Int. Law* 98 (2), 276–301.

Seaports

Mary R. Brooks*, **Geraldine Knatz†**, *Rowe School of Business, Dalhousie University, Halifax, NS, Canada; †USC Sol Price School of Public Policy, USC Viterbi School of Engineering, Los Angeles, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	299
Seaport Typology and Governance	299
Port Business Models	300
Port Planning	301
Marketing and Competition Between Ports	301
Labor	302
Critical Issues for Ports	302
Port Security	302
Environmental and Sustainability Challenges Facing Ports	303
Climate Change and Adaptation to Sea Level Rise	303
Biographies	303
Relevant Websites	304
References	304
Further Reading	304

Introduction

Ports are an eclectic mix of public and private activities that are geographically situated on navigable waterways, generally for the purpose of moving goods in domestic or international trade. Ports may also move people, as cruise ports, to tourist destinations around the globe. This section of the encyclopedia provides an overview of the types of cargo-handling in ports, by commodity and by governance structure. Fundamental aspects of the role played by port authorities are explained in the sections devoted to port planning and port marketing. Current perspectives related to port labor, particularly in the wake of automation of many traditional port labor functions is provided, despite the fact that the hiring and supervision of port labor may not be the obligation of the port authority, particularly for ports that are landlords. Today ports face a myriad of challenges. Highlighted here are ones of primary importance—port security, environmental challenges, and climate change and the consequence of sea level rise.

Seaport Typology and Governance

There are several ways to characterize ports: by cargo, like container ports; by function, like feeder or hub; or by governance model, like private or government-owned.

Characterized by cargo type, seaports may be single purpose or multipurpose ports, and so may serve a varying combination of market segments depending on their business opportunities. Ports are often called by their predominant cargo type, such as a container port, or bulk port but it is not uncommon to find that a so-called container port may also handle other cargo types. Similarly, some smaller government-owned ports, typically non-container ports, are often referred to as “niche” ports. The market segment of niche ports tends to rely on traditional, non-containerized cargoes, such as automobiles or seasonal fruit moved as break-bulk cargo.

Generally, there are no clear-cut patterns to categorize port market segments by governance type. Successful ports tend to be opportunistic and strive to accommodate as many different types of cargo and customers as they have facilities for. As such, larger ports managed by a governmental authority tend to be multipurpose ports. Fully private ports tend to be single purpose ports, often a single terminal that specialize in one particular cargo type such as liquid bulk (crude oil, petroleum products, molasses), lumber, or dry bulk cargoes like coal or alumina. For every generality about port typology, one can find exceptions due to the complexity of port ownership, management, and operational control and the global variations in the market.

Within a particular region, ports may also be characterized by the function they play in relationship to other ports or trading patterns. Feeder ports serve to collect and transport goods, such as containers, from their local hinterland to a larger hub port for international shipment. This is common in a situation where not all the ports in a particular region have the market strength or facilities to handle larger ships, but rather serve to ferry goods to a larger hub port or gateway that has sufficient facilities and capacity for larger ships moving in the global trading lanes. These feeder ports can be part of a cargo relay system or one of the “spokes” in a “hub and spoke” system. For example, some Japanese ports serve as feeder ports to Busan, Korea. Similarly, transshipment ports (e.g., Shanghai), where cargoes are moved from one vessel to another, can also serve as collecting points for a hub port. A hub port might also be characterized as a “gateway” port, meaning it has extensive rail, highway, or

waterway access to serve a deep hinterland (e.g., Los Angeles). Governments under international regulatory regimes remain responsible for port security and port state control, but may have agreements for alternative compliance regimes to deal with these responsibilities.

Ports are often characterized by governance model because it is possible to identify a reasonable number of discrete patterns found among governance in our world ports (Baltazar and Brooks, 2007). Because seaport operations typically involve multiple actors charged with the management, regulation and/or operation of services that utilize a navigable waterway, port governance is a significant reflection of the tension between regulators, local and national governments, and the private sector. There are five typologies of port ownership, management and operations, which are useful in understanding the key relationships that direct the ability of the port to serve the various user types:

1. Central government-owned with central government management and operations (Examples: National Harbors Board ports in Canada pre-1982, Spain pre-1992, Italy pre-1994, Greece pre-1999);
2. Regional or local government-owned with management by the regional or local government body. Operations may be a combination of public terminals operated by the government (operating port) or with some or all facilities operated by the private sector via a concession or lease (Examples: Most US ports);
3. Government-owned (national, regional, or municipal) but managed by a corporatized entity (Examples: Melbourne in Australia pre-2016; Sydney in Australia pre-2013; Rotterdam in Europe);
4. Government-owned but managed by a private sector entity via a concession or lease arrangement with a terminal operator, or owned and managed via a public-private partnership agreement (Examples: Many ports in developing countries, Melbourne post-2016; Sydney post-2013); and
5. Fully privately owned, managed and operated (Examples: Many UK ports)

While private ports exist in most countries, few countries have contemplated full privatization of the port sector. Outside of the United Kingdom, most privately held ports serve single cargo facilities—mines, refineries, power plants, and the like. Multi-purpose ports are usually developed with public funds. The dominant model of port governance in the second decade of this century is the landlord model. A publicly-owned port authority acts as the landlord and private companies carry out port operations on behalf of those government owners, usually through a lease or concession contractual arrangement.

Port reform efforts by governments have often focused on trying to extract lease payments for positive cash flows to government or long-term lump sum concessions for government debt reduction. While the principles of good governance would encourage that there be a separation of regulator versus operator functions, this does not always happen. Across the globe, there are many models of port governance, and most research concludes that there are few best practice approaches and no one right way for governments to establish the governance structure and port policies for the ports in the country (Brooks, et al., 2017; Brooks and Pallis, 2012; World Bank, 2007).

Port Business Models

Seaports generate revenues through fees for services offered to port users. Seaports typically have published tariffs, which indicate their rates and charges for port users. Those fees generally consist of wharfage, the fee associated with moving a particular cargo across a wharf for loading or offloading a ship; dockage or berthage, the fee assessed against a vessel for berthing or making fast against a wharf, pier, bulkhead or other structure; demurrage, the fees for storing cargo left on the wharf; and, pilotage, the fees for the service of providing a harbor pilot to guide movement of a ship into, out of, or within a port. Fees can also be assessed by the port authority for the right to anchor a vessel within the jurisdiction of the port.

Landlord ports will enter into short- or long-term agreements (leases or concessions) where real estate is leased to a private company, like a stevedoring company, that will operate the facility or terminal. In some cases, the landlord port will build the terminal for its customer and lease the land with improvements. In other cases, the tenant will construct some of all of the improvements. Port customers that have leased a facility may have a compensation structure that differs from the published rates and charges. Seaports strive to generate sufficient revenue through compensation paid by its customer(s) to allow reinvestment for improving harbor facilities. Historically, a typical business model for a port authority was to lease a terminal to an ocean carrier or stevedore for a long-term, such as 20–30 years, with a guaranteed annual minimum payment to the port authority. This minimum was typically set to allow the port authority to realize a return on its investment, as well as allow for reinvestment in modernizing the terminal facilities over time.

The port's investment in the capital expense of terminal construction may be financed by issuing bonds, using taxing authority or via a public-private partnership (P3) with an investor or its customer.

Structural changes in the ocean shipping industry, including the consolidation of the seaport's multiple customer base into fewer but larger ocean carrier alliances, have challenged traditional port business models. Ocean carriers may have eliminated their commitment to call at specific terminals for long periods of time, jeopardizing a port's reliance on the long-term commitment of cargo. Investment in new terminals or terminal expansion projects are riskier if ports cannot guarantee that enough cargo will use the terminal long enough for the port authority to recover its investment. Thus, ports have developed various alternative financing strategies to their traditional approaches. From the port's perspective, risk reduction means strategies that transfer or share the risk. These strategies range from a private concession/equity approach working with private equity funds that focus on infrastructure

investments, to P3s that continue the port's traditional role as the landowner and landlord role but with private equity making the infrastructure investment. Having a commitment by ocean carriers to move cargo through the terminal is vital to making a P3 economically viable.

Port Planning

Port planning is a key activity of port managers. Capital investments and growth for new business should be built on measuring port performance and productivity numbers. The challenge here is that there is no consensus on how a port's efficiency or productivity should be measured. Without understanding the existing productivity of a port, it is difficult to plan for future infrastructure needs. In less developed countries, for example, containers may be lifted from a ship's deck by a mobile crane, using a crew of several workers and managing one container every several minutes, while in a developed country port, there may be multiple shore cranes working the ship, each crane able to offload one or two containers with efficiencies that have been recorded as high as less than one minute per lift, and using one operator per crane.

There are multiple plans produced by a port authority. These may include:

1. Comprehensive, master, or long-range plans, which detail the spatial organization of the port, often serving as the fundamental guide to port land use policies and the designation of specific uses for land and water areas. These plans require periodic updating, generally after five years.
2. Facility or project plans, which focus on the engineering, economics, and environmental impacts of the proposed development project. The outcome of this process should be a development plan sufficient to pursue project funding or financing.
3. Capital Improvement Plans, which prioritize multiple projects and are tied to the port's budgetary process. Because the development of a port project may take many years, ports typically have 5- or 10-year capital improvement plans.
4. Risk management plans, which dictate the policies for regulating the siting of port facilities that handle hazardous cargoes. A fundamental tenet of port planning is co-location and segregation; these plans co-locate similar land uses as much as feasible, and segregate incompatible land uses such as public use areas and hazardous cargo facilities.
5. Strategic plans, which include an analysis of the forces affecting the port's internal and external environment and those that will shape the port in the future. A plan can be both outward-facing—that is dealing with the port's relationship with its customers and stakeholders, and inward-facing—that is dealing with its own governance, internal operations, and employees. The strategic plan identifies the port's mission, vision and values. The desired outcome of a strategic planning process should yield an understanding of a port's market and competitive position in context with those ports against which it competes. These plans are typically updated annually or every few years.

A comprehensive master plan should facilitate project decision-making based on longer-term trends in trade growth and an analysis of the port's particular market. The plan should be based on an econometric forecast of trade volumes, examining long-range forecasts of GDP, bilateral trade movements and potential disruptors to established trading patterns, along with a determination of the throughput capacity of the existing port facilities to identify gaps that trigger new facility development. Examples of disruptors to trading patterns can include the expansion of the Panama Canal in 2016, climate change and the potential of year-round shipping through the Arctic. Research on Asian ports has found that natural disasters and labor strikes are the two most common disruptors that ports must prepare to address in the master plan (Lam and Su, 2015).

Another important feature of a comprehensive port plan can be the port-city interface, particularly for urban ports where there is a desire to enhance the bond or create buffers between the city and the port. Here, factors like preserving scenic views of the water and ensuring consistency between urban mobility plans and port connections could be important. Where a port has developed historically in the core of an urban center, planning for new facilities can be constrained by limited land area, creating a need to examine efficiency improvements within the port's existing jurisdiction as the strategy for growth or examining growth opportunities in the port hinterlands. A best practice in port planning is conducting outreach with stakeholders, including not only port users and regulators, but also members of the public and residents of surrounding communities.

With the unprecedented pace of changes in maritime-related technology however, port planning should be viewed as a continuous process and not an end result. This is particularly true in the areas of port strategic planning and facility and project planning.

Marketing and Competition Between Ports

While some bulk ports serve captive customers (mines, refineries, and the like), most ports must compete for the cargo they handle. A seaport is only as competitive as the origin-destination total supply chain (for the products it handles). A cargo owner seldom uses only one port, as mitigating route risk (due to disruption by natural forces, labor unrest, or terrorist actions) encourages some splitting of the business across more than one route (and port); a slower but more reliable route will share traffic with a faster but less reliable one. The intention of route diversification is to minimize the probability of an inventory stock-out at

destination. While this may result in higher inventory carrying costs in transit for the cargo owner, such a strategy may more than offset the cost of the additional buffer stock or the consequential losses arising from a production shutdown or a retail stock-out for the cargo owner.

Other factors affecting port competitiveness include the quality of the hinterland connections, and the origin-destination transit time. In North America, rail connections and efficiency can pit a port on one coast like Vancouver, Canada against another port, like New York, on a different coast. Ocean carrier reliability concerns spill over onto port competitiveness, just as port capacity concerns, strikes, and lockouts spill over onto a cargo interest's choice of ocean carrier. Transit time variability is examined closely as shippers seek performance data not just on ocean carrier arrival reliability but also on container yard dwell time, and trucking company delivery performance. Total logistics costs drive the route decision for those companies seeking greater control of their operations, better customer experience and shareholder value creation.

There is divided opinion on the best way to market a port. Some ports see their role as being focused on serving the ocean carrier as customer, delivering promised port efficiency in cargo handling. Others, particularly those without a "must call" population base, are more focused on the target market of cargo owners, convincing those importers and exporters that the port can meet their needs, with marketing to the shipping line secondary. Still others are port community-focused, knowing that if the region is not served with good transportation access to warehouses, and efficient hinterland connections, the social license they need to continue to serve port customers will evaporate. Most port managers today focus on a multi-pronged marketing approach, a combination of these three. There remain, however, ports whose management continues to see their primary "customer" as the government they act on behalf of, and their decision-making has a political focus. Many ports undertake economic impact studies, identifying the jobs created and financial benefits to a particular region, to assist with local political efforts.

Labor

Port labor is often a contentious issue in the management and operations of ports. There are multiple perspectives about what is the correct number of dock workers.

First, there is the view of the day-to-day terminal operator, whose efficiency may be measured in terms of tonnage or containers moved per employee hour, and therefore is seeking to improve the level of productivity of each employee and minimize the number of employees for the expected cargo volume. This may trigger plans by the operator to automate many port processes, substituting capital investment in high-technology equipment for port labor costs.

Second is the view of those employees. In many parts of the world, dock labor is highly unionized and has been effective in gaining high wages for port workers. In such cases, unions are reluctant to concede hard-won labor contract gains, even under the threat of automation of terminal activities.

Third is the view of government. In many countries, particularly those in the developing world, ports are seen as important employers in the community, and the government is keen to increase employment numbers rather than focus on efficiency, putting political pressure on ports where governance structures allow.

Efforts to automate ports may therefore come up against either governments or unions or both, while in high-labor-cost countries, decisions to automate can be dependent on the availability of capital for that investment or driven by the share labor costs represent in the port or terminal's total budget.

Critical Issues for Ports

Port Security

The modern port security regime is a product of global security initiatives galvanized by the terrorist attacks on the United States of 11 September 2001. Multilateral regulators implemented new security requirements under the auspices of the International Maritime Organization, which in 2002 developed the *International Ship and Port Facility Security Code (ISPS Code)* and added it as an amendment to the *Safety of Life at Sea (SOLAS) Convention, 1974*. As most countries have adopted SOLAS, the ISPS Code effected a global common approach to port security as the Code sets forth mandatory security requirements to be taken by governments, ports, shipping companies, and terminal operators. These requirements came into force on 1 July 2004. Each country's national authority certifies those ports that meet the Code.

The ISPS Code defines maritime security (MARSEC) levels. The local government sets the security level and informs the port and ships prior to entering the port, or when berthed in the port. Each of the security levels under the ISPS code describe the current scenario related to the security threat to the country and its coastal region, including the ships visiting.

- Level 1 is the normal level that the ship or port facility maintains on a daily basis, and is a minimum of appropriate security on a 24 hour a day, seven days a week basis.
- Level 2 is a heightened level for a time period during a security risk that has become visible to security personnel.
- Level 3 is an additional level of security with measures for an incident that is forthcoming or has already occurred.

Each MARSEC level has specific actions expected of the port and a vessel's personnel to maintain port security.

Environmental and Sustainability Challenges Facing Ports

Most ports face a myriad of environmental and sustainability concerns, including air pollution from ships, access road congestion, and the resulting air pollution in the vicinity of the port, water pollution, ballast water that carries invasive species, local habitat protection, and the disposal of dredged materials. Globally, ports are working to reduce their environmental impact in a multitude of ways ranging from reducing water and air pollution, reducing traffic congestion, minimizing the production of greenhouse gases, and shifting to cleaner energy sources. By virtue of the port being a node where multiple types of transportation sources come together, ships, trains, trucks, and terminal equipment, ports have served as laboratories for the testing of new types of energy efficient and green technologies. In many cases, the increase in environmental consciousness on the part of ports can be attributed to the environmental activism among near-port communities. While some ports have been forced into greener operations by regulators, all ports have realized the benefits of sharing best practices and adapting them to their own port's environment.

Since 1996, the European Sea Ports Organization has surveyed its members to identify the top 10 environmental issues facing their member ports. Not surprisingly, in 2018 air quality is number one and has been since at least 2013. Many regional and national port organizations have committees or areas of specialization devoted to the environment. Many of these port associations have created environmental or green port awards to encourage ports to share their best practices (for example, American Association of Port Authorities Environmental Awards, European Sea Ports Organization's Environmental Award, Asia Pacific Port Services Network Green Ports Award). Many global ports have become members of the World Ports Sustainability Program (WPSP), part of the International Association of Port Authorities. Here, individual ports will lead specific focus areas, encouraging information sharing and dialogue. One international program developed by the predecessor to WPSP, the World Port Climate Initiative, is the Environmental Ship Index, an incentive program ports can join to attract the cleanest ships to their harbor, similar to the Green Award program established by the Port of Rotterdam and the Dutch Ministry of Transport and Water Management to initiate market incentives that promote green shipping. The Clean Air Action Program developed by the Port of Los Angeles and Long Beach includes a technology advancement program to encourage "proof of concept" for port related technologies to reduce air emissions.

Climate Change and Adaptation to Sea Level Rise

Climate change and sea level rise are global challenges that necessitate actions at the global and local level for seaports. Global actions in the maritime realm fall within the purview of the International Maritime Organization, which is currently focused on reducing the greenhouse gas emissions of vessels. Ports sit on the frontlines of global climate change due to their coastal location and are thus positioned to take necessary actions with their region as part of their ongoing development functions. Mitigating climate change through the reduction of greenhouse gases is now a mainstream activity of ports. Most ports are focusing on "hard" engineering measures as their main response to address sea level rise (UNCTAD, 2018). Thus, many ports already address sea level rise in their design of new piers and infrastructure (Ng et al., 2015). Landlord ports have additional challenges where they have leased their waterfront to port terminal operators under long-term concessions. It may be hard to incentivize a concession holder to take actions now to adapt to sea level rise when the length of their concession may only be 20 or 30 years. Many concessionaries view this as making investments that would benefit future concession holders. Yet, the ports that will become the most resilient to climate change in the future are the ones where climate change policy is considered in concession agreements.

Sea level rise is only one of multiple challenges ports face due to climate change. The melting permafrost that supports railroad tracks in northern Canada, the extreme temperatures experienced by dockworkers in Australia, the lower water levels of the North American Great Lakes and Gatun Lake in the Panama Canal are all examples illustrating the wide range of impacts on port infrastructure, operations, and worker health and safety. The challenge for all ports is to figure out how to proactively address climate adaptations today rather than waiting until they are forced to react in a crisis situation. Remarkably, the efforts of many ports to "green" their operations have had the unintended consequence of making the ports less resilient. Reducing the production of greenhouse gases by replacing the use of diesel-powered equipment with electrically powered equipment is one example where the vulnerability of the port may be increased in the case of a severe weather event involving the loss of electrical power.

Biographies

Dr. Mary R. Brooks is Professor Emerita at Dalhousie University, Halifax, Canada, and a founding editor of *Research in Transportation Business & Management*. In 2016–18, she served as Chair of the Marine Board of the US National Academies. Her research focuses on competition policy in liner shipping, port strategic management and short sea shipping. She has authored and published more than 25 books and technical reports, more than 25 book chapters, and more than 80 articles in peer-reviewed scholarly journals. In September 2018, she received the Onassis Prize in Shipping 2018, in recognition of 'her outstanding contribution to the study of shipping.'

Dr. Geraldine Knatz is Professor of the Practice of Policy and Engineering, a joint appointment between the University of Southern California School of Public Policy and School of Engineering. She served as the CEO of the Port of Los Angeles from 2006 to January 2014 and prior to that was Managing Director at the Port of Long Beach. She is past president of the American Association of Port Authorities and the International Association of Ports and Harbors, and founding Chairman of the World Port Climate Initiative. In 2014, Knatz was elected to the National Academy of Engineering in recognition of her international leadership in the

development of environmentally clean urban seaports. That same year she was honored with Containerization & Intermodal Institute's Lifetime Achievement Award.

Relevant Websites

Association of Port Cities, 2015. Plan the city with the port, guide of good practices, Paris. Available from: <http://www.iuav.it/Ateneo1/docenti/architettura/docenti-st/Umberto-Tr/materiali-/09—Labor/Plan-the-City-with-the-Port.pdf>.

European Sea Port Organization, 2018. Top 10 environmental priorities. Available from: <https://www.espo.be/publications/top-10-environmental-priorities-2018>.

International Association of Ports and Harbors, 2019. World ports sustainability program (WPSP). Available from: <https://sustainableworldports.org>.

International Association of Ports and Harbors, 2010. Carbon footprinting guidance. Available from: http://wpci.iaphworldports.org/data/docs/carbon-footprinting/PV_DRAFT_WPCI_Carbon_Footprinting_Guidance_Doc-June-30-2010_scg.pdf.

International Maritime Organization, 2019. SOLAS XI-2 and the ISPS Code. Available from: http://www.imo.org/en/ourwork/security/guide_to_maritime_security/pages/solas-xi-2-%20isps%20code.aspx.

Maritime Administration. Port planning and investment toolkit, US Department of Transportation. Available from: <http://www.maritime.dot.gov/ports/strong-ports/port-planning-and-investment-toolkit>.

Maritime Administration and American Association of Port Authorities Port Planning and Investment Toolkit, Introduction and User's Guide, Available from: <http://aapa.files.cms-plus.com/PDFs/Toolkit/Introduction%20and%20Users%20Guide.pdf>.

Ports of Los Angeles and Long Beach, 2019. Technology advancement program of the clean air action plan. Available from: <http://www.cleanairactionplan.org/technology-advancement-program/>.

References

- Baltazar, R., Brooks, M.R., 2007. Port governance, devolution and the matching framework: a configuration theory approach. In: Brooks, M.R., Cullinane, K.P.B. (Eds.), *Devolution, Port Performance and Port Governance*. Res. Transp. Econ. 17, 379–404.
- Brooks, M.R., Pallis, A.A., 2012. Port Governance. In: Wayne, T. Talley (Ed.), *Maritime Economics – A Blackwell Companion*. Blackwell Publishing Ltd., pp. 491–516.
- Brooks, M.R., Cullinane, K.P.B., Pallis, A.A., 2017. Revisiting port governance and port reform: a multi-country examination. Res. Transp. Bus. Manag. 22, 1–10.
- Lam, J.S.L., Su, S., 2015. Disruption risks and mitigation strategies: an analysis of Asian ports. *Maritime Policy Manag.* 42 (5), 415–435.
- Ng, A.K.Y., Becker, A., Cahoon, S., Chen, S.L., Earl, P., Yang, Z., 2015. *Climate Change and Adaptation Planning for Ports*. Routledge, London.
- UNCTAD, 2018. Port Industry Survey on Climate Change Impacts and Adaptation, United Nations Conference on Trade and Development, Research Paper No. 18, /SER.RP/2017/18. Available from: https://unctad.org/en/PublicationsLibrary/ser-rp-2017d18_en.pdf.
- World Bank, 2007. *World Bank Port Reform Toolkit*, second ed. Available from: <https://ppp.worldbank.org/public-private-partnership/library/port-reform-toolkit-ppiaf-world-bank-2nd-edition>.

Further Reading

- Bottasso, A., Conti, M., Ferrari, C., Merk, O., Tei, A., 2013. The impact of port throughput on local employment: evidence from a panel of European regions. *Transport Policy* 27, 32–38.
- Brooks, M.R., Button, K.J., 2007. Maritime container security: a cargo interest perspective. In: Bichou, K., Bell, M., Evans, A. (Eds.), *Risk Management in Port Operations, Logistics and Supply Chain Security*. Informa, London, pp. 221–236.
- Knatz, G., 2017. How competition is driving change in port governance, strategic decision-making and government policy in the United States. Res. Transp. Bus. Manag. 22, 67–77.
- Turnbull, P., 2010. From social conflict to social dialogue: counter-mobilization on the European waterfront. *Eur. J. Indus. Relat.* 16 (4), 333–349.

Port Hinterlands

Francesco Parola^{*,†}, Giovanni Satta^{*,†}, Francesco Vitellaro[†], ^{*}Department of Economics and Business Studies, University of Genoa, Genoa, Italy; [†]Italian Centre of Excellence on Logistics, Transport and Infrastructures, University of Genoa, Genoa, Italy

© 2021 Elsevier Ltd. All rights reserved.

Ports and Their Hinterlands: The Impact of Containerization	305
Hinterlands Become More Contestable	305
Ports are Inserted into Broader Supply Chains	306
Shaping Port Hinterlands: A Mix of Positive and Negative Forces	306
The Active Role of Key Actors in Developing Port Hinterlands	307
Shipping Lines	307
Terminal Operators	307
Rail Transport Operators	308
Port Authorities	308
The Emergence of Port Regionalization	308
References	309
Further Reading	309

Ports and Their Hinterlands: The Impact of Containerization

The concept of a port hinterland has changed dramatically over the years, following the evolution of maritime transport and the logistics industry. A hinterland represents the inland catchment area of a port, from where it produces the majority of its business. Indeed, the points of origin/destination of cargo flows, which passes through the port, are located in the hinterland. In summary, the traditional notion of a hinterland refers to the economic area influenced by the ports. It is rather complex to identify the boundaries of a hinterland's extent, as this largely depends on the different commodity flows and transport modes. Furthermore, a hinterland's size is constantly changing over time due to economic cycles, seasonality effects, technological transformations, evolutionary changes of carriers the strategies of multimodal transport operators, capacity constraints of infrastructure, modifications in transport policy, etc. Hence the concept of a port hinterland is very dynamic and a static approach, considering port hinterlands as being everlasting may be misleading (Olivier and Slack, 2006).

Academics began in-depth investigations into port–hinterland relationships in the 1960s, even though some pioneering studies date back even earlier. Formerly, the concept of a hinterland was mainly founded on zonal models, paying scarce attention to the relevance of inter-port inland competition. Then, the containerization process and intermodality broke this approach and pushed the integration of ports as nodes of global transport chains. Therefore the selection of gateways and corridors by carriers has been placed at the center of the debate, reshaping profoundly the concept of the hinterland (Ferrari et al., 2011).

Nowadays, a port hinterland is viewed as a broad logistics area strictly linked to global supply chains and international transport players. The shape of the hinterland is no longer considered a function of the physical distances involved and it seems to be affected by numerous discontinuities along its boundaries. Such discontinuities are provoked by inefficiencies across the logistics chain and also by the contextual competition of distant ports. Although some scholars still persist in considering distance as a driver for defining port hinterlands, most academics assert that the competition among the top players is not necessarily grounded on inland distance.

Hinterlands Become More Contestable

The emergence of global networks affected the relationship between the nodes of the supply chain and their market areas. In this perspective, it is possible to identify some key factors that have facilitated the growth of gateway ports with the characteristics to compete for contestable hinterlands. Academics have extensively studied the relationships between gateway seaports and the hinterland, focusing on the drivers, which have affected port catchment areas. In particular, containerization and intermodality represent two pivotal factors for the extension of port hinterlands and the strong intensification of interport competition.

The invention of maritime containers triggered a significant revolution in the transport industry. Containers have become the common transport unit, enabling maritime transport to stretch the business toward land and to build innovative door-to-door systems. In the second revolution, the development of intermodality, most closely associated with rail and barge transport, further enlarged the land penetration of maritime containers through the creation of landbridges (Muller, 1999). This revolution, together with the introduction of logistics corridors, has largely widened port hinterlands, leading to a paradigm shift from “captive to contestable” hinterlands. Moreover, the perception of port markets has passed from being monopolistic or oligopolistic to being competitive, opening new possibilities for further exploring and developing the business.

These changes have deeply affected the relationships among ports located in the same range (e.g., Rhine-Scheldt delta, Southern California, etc.) and also, to a certain extent, in opponent ranges (e.g., USWC vs USEC, Northern vs Southern range ports in Europe, etc.), producing a much fiercer on-shore competition for capturing cargo. As a result, the competition led to the “natural selection” of some gateways, which emerged as leaders for their operational reliability, maritime strategic location, as well as their effectiveness in inland penetration. Notably, these gateways represent nodes of global maritime networks and, thus, allow long-haul transport flows to reach broad continental areas and vice versa. It is widely demonstrated that containerization and intermodality have progressively eroded traditional hinterland paradigms. Indeed, the growing hinterland contestability has altered hinterland size, as well as its shape. Hinterlands also became characterized by more discontinuous physical boundaries, even in the immediate vicinity of the port. [Notteboom and Rodrigue \(2005\)](#) argued that this process may lead to the formation of “islands” in the distant hinterland, where the load center port in question achieves a comparative cost and service advantage with respect to their competitors. Hence traditional perspectives based on the concept of distance decay are unsuitable to face this new scenario. In this regard, as mentioned earlier, the effectiveness of inland connections and the relationship with inland terminals are fundamental for port competitiveness.

Ports are Inserted into Broader Supply Chains

The rise of corridors has influenced the relationship between gateway ports and their hinterland. These logistics infrastructures enable maritime gateways to enhance their penetration toward the hinterland and increase traffic volumes. In the past, the competitiveness of a container port was grounded on its standalone “physical” features (e.g., asset endowment, nautical accessibility, ship-to-shore performance, container dwell time, etc.). Nowadays, however, port selection is strongly conditioned by hinterland accessibility and, thus, by the quality of gateway port-hinterland connections. In this vein, physical attributes cannot be used as the only parameters for assessing the competitiveness of a port, since they no longer reflect the complexity of global supply chains. This has been provoked by globalization, the delocalization of production plants as well as the physical dispersion of manufacturing inputs over a broader geographic space. Multinational manufacturers, indeed, have introduced a more flexible multifirm organizational structure, based on the fragmentation of production and logistics activities throughout globally outstretched supply chains ([Olivier and Slack, 2006](#)).

In this perspective, ports became only a stage in the entire logistics equation, and their competitiveness is mostly shaped by the efficiency of logistics operations beyond port boundaries and, especially, by their inland connectivity and reliability. Therefore ports have been increasingly competing not as individual maritime nodes, basically responsible for loading and unloading cargo, but as crucial interfaces within different logistics chains. For this reason, ports are expected to shift their strategy to an overarching supply-chain approach aiming at meeting the requirements of both carriers and shippers. Consequently, this might also suggest in terms of governance mechanisms, the creation of an independent public body for coordinating the activities along the supply chain, including port operations and hinterland transportation. This is crucial for the proper functioning of the network, as the lack of integration among public bodies (e.g., port authority, municipality, region, central government, etc.) and private companies can generate institutional problems and serious bottlenecks.

Shaping Port Hinterlands: A Mix of Positive and Negative Forces

The earlier-mentioned arguments suggest that the capacity of a port for attracting cargo is strictly correlated to the degree of inland penetration. In this vein, port and hinterland appear as a unique interconnected entity and the competitiveness of this broad logistics area is determined by the mutual attractiveness of port and hinterland together, as well as their attitude toward collaborating with each other. In particular, the quality of hinterland connections heavily affects port competition; port attractiveness is increasingly dependent on overall network costs and performance. Hence port selection criteria strongly take into account the role of the port as a pivotal node within the global supply chain ([Ferrari et al., 2011](#)).

Hinterland contestability can be seen as the contextual intervention of positive and negative “forces” (standalone and external factors) across the overall transport chain. These “gravitational forces” attract ports and hinterlands to each other and generate traffic flows between them. Among these drivers, the extant literature suggests that the main important ones are: the effectiveness of port handling operations, the location of distriparks and inland terminals within and nearby the port domain, the presence of an international business environment, a strong local cargo base (calling for transshipment operations as well as additional gateway cargo for long inland distances), fast vessel turnaround times, short container dwell times, efficient rail, road and barge networks, and effective port governance. These positive forces extend port hinterland coverage and foster its capacity to compete against faraway ports. Nonetheless, such positive forces are counterbalanced by negative ones, which determine “frictions” to the transit of cargo flows and limit hinterland contestability. In particular, the lack of capacity in port and inland infrastructures, high port costs, inefficient, expensive, and unreliable inland connections, all negatively affect port development toward the hinterland, deviating the traffic flows to other corridors and/or ports. When these negative forces occur, logistics players are pushed to find alternative routes with a lower “resistance” in terms of costs and reliability. Consequently, ports that are located on inefficient or low-capacity corridors are in a disadvantageous position.

In summary, gravitational forces and frictions can be assessed through the analysis of costs (considering the overall logistics costs), capacity (e.g., port handling, inland infrastructure, etc.), and the reliability of the port-hinterland system. Nevertheless, this system is affected by other drivers concerning the competitive position of a port in relation to a specific hinterland region, that is,

historical, psychological, political, cultural, institutional, legislative, and personal ones, which are particularly difficult to manage and can largely twist the “perfect” market-based division. In particular, bounded rationality, inertia, and opportunistic behavior are among the behavioral factors that could lead to a deviation from a “distance-decay” optimal solution.

The Active Role of Key Actors in Developing Port Hinterlands

The success of a seaport is based on its ability to cooperate with the maritime cluster and logistics operators in order to exploit all the synergies with the other nodes of the supply chain. In particular, the collaboration between seaports and inland terminals has led to the development of large logistics areas that have arisen both spontaneously, as the result of a slow market-driven process, and carried out by national, regional, and/or local authorities through financial and regulatory actions. These partnerships have proven to be not only of a competitive nature but also of a complementary nature, as they can enhance cargo flows and the competitiveness of the entire supply chain. As regards seaports, the geographical concentration of logistics companies in the hinterland and the development of frequent, well-structured connections lays the foundations to a superior network that positively influences port attractiveness and encourages shipping lines to call and invest in that specific port.

Therefore the development of a broader port hinterland is mainly due to the strategies of maritime and logistics actors. In particular, this section reports the impact of horizontal and vertical relations among shipping lines, terminal operators, rail transport operators, and port authorities (PAs) that aim at supporting the growth of logistics zones and supply chains.

Shipping Lines

Over the last years, shipping lines have implemented internal and external growth strategies aiming at reaching major economies of scale and lowering transport costs. As a result, a significant number of mergers and acquisitions have been carried out in the industry, giving rise to a gradual market concentration. Nowadays, the top 10 carriers control 82.8% of the world market share of the container shipping industry, with 64.5% share controlled by only the top 5 (Alphaliner, 2019). In addition, the formation of global alliances and the deployment of increasingly bigger ships has contributed to reducing the overall ship system costs, as well as to increasing the bargaining power of ocean carriers versus terminal operators and PAs. Nonetheless, shipping lines have to face more demanding shipper requirements that represent a serious challenge for their competitiveness, especially in terms of frequency, reliability and geographical coverage of the services demanded. Therefore due to the inefficiency of many ports of call, liners have gone beyond simple long-term contracts with terminal operators and have started to acquire shares in several maritime container terminals located in strategic positions, in order to directly control and enhance their capacity.

When it comes to the development of a hinterland network, scholars have demonstrated that the proportion of total transport costs represented by intermodal costs is constantly increasing, reaching a range somewhere between 40% and 80%. This supports the notion that inland logistics represents a fruitful area for shipping lines’ growth strategies, due to the possibility of managing the entire transport service and provides more competitive services (i.e., carrier haulage). In this vein, they can further reduce the overall costs related to transport activities (i.e., point-to-point), as well as foster and monitor logistics operational performance. Hence shipping lines have developed rail, barge, and truck services through their own companies (i.e., associated firms or controlled entities) or long-term strategic partnerships with major logistics players. Ultimately, the most ambitious companies have invested in inland terminals and depots aiming at accelerating inbound and outbound container flows.

Terminal Operators

Large global terminal operators have emerged in maritime logistics and inland transport to counterbalance the new dominant position of shipping lines. In this vein, an increasing consolidation has also occurred in this industry, fed by the growing presence of institutional investors (e.g., banks, private equity funds, insurance companies, pension funds, etc.) that have revealed a strong interest in the terminal business. As a result, four worldwide operating companies (i.e., Hutchison Port Holdings, PSA International, DP World, and APM Terminals) dominate the container handling industry, controlling more than 40% of the market share. In addition, for the last two consecutive years, Cosco Shipping Ports, formed by the merger of the two Chinese conglomerates—Cosco Group and China Shipping Group, has reached primacy in the industry (Lloyd’s List, 2018). CS Group achieved this result through the takeover of Greek transshipment hub Piraeus and the Spanish port operator Noatum, and this has considerably enhanced its volume of containers handled. Contextually, terminal operators have strengthened their position in relation to customers by establishing equity joint ventures with shipping companies aiming at co-managing strategic container terminals worldwide. In such a way, terminal operators can secure a more stable cargo base and reduce traffic volatility.

Given the earlier, leading terminal operators have been developing diverging strategies toward extending their business and network. In particular, they have increased the array of logistics services provided, including inland transport (i.e., terminal operator haulage), and their influence throughout the supply chain. Furthermore, they have incorporated several inland terminals in order to create extended gates of seaport terminals. In this vein, they can meet shippers’ requirements, moving closer to the “final customer”, as well as relieve the pressure and congestion of seaport terminals. Hence, terminal operators nowadays appear as key actors in the successful development of port hinterlands.

Rail Transport Operators

Rail transportation is at the heart of the logistics debate, especially in Europe, because it appears geographically, politically and economically fragmented and far away from the realization of a greater intermodal infrastructure for supporting seaport expansion. Nevertheless, the ongoing sectoral liberalisation and the increasing involvement in the industry of major shipping lines, terminal operators, and PAs is helping to bring about a new era in railway transport. Their initiative has proved to be essential in overcoming the limits of rail services that lay with the high entry costs and the fragmentation of cargo flows. In this vein, rail operators have designed more efficient networks by connecting the main nodes of the supply chain that includes seaport terminals and logistics centers located in the hinterland. According to the hub-and-spoke model, cargoes are bundled in the aforementioned logistics nodes and transported by frequent fast shuttle trains. This system has proved to be particularly effective in the European ports of the Northern range, thanks to the well-developed railway infrastructure designed by local and national governments.

The increasing role of rail transport operators in the growth of the port-hinterland is due to their new commercial attitude that fosters the improvement of the services provided (e.g., decreasing the lead time for booking, digitalization of documentation, track and tracing, integration of intermodal services, and door-to-door transport). In particular, the integration of effective railway connections by seaports allow enhancing supply chain performances and promoting intermodality that constitutes a prerequisite for the creation of a broader logistics areas.

Port Authorities

The main objective of PAs is to attract users by providing competitive services to shippers, carriers, and the other logistics stakeholders in the supply chain. In this regard, the current main challenge is represented by hinterland connections that allow PAs to access cargoes far from port boundaries, promote intermodality, and increase their revenues. As well known, ports are pivotal nodes of the logistics chain thanks to their accessibility to the sea and foreland connections. This gives PAs the power to shape hinterland networks and preserve their centrality in the system. Nevertheless, the worldwide transport network, considering both land and sea transport services, is becoming increasingly complex due to the growth in the number of intermodal logistics centers (on the land side) and sea ports, especially transshipment ports that offer valuable routing options for container flows. Indeed, most European container ports are facing a dramatic increase in the relative growth of transshipment traffic flows that requires a strengthening of both port infrastructures and inland long-distance infrastructure. Therefore PAs are expected to develop even closer collaboration with various logistics operators, especially inland and intermodal terminals, in order to foster port competitiveness and intensify demand flows.

Given this perspective, many PAs have launched several projects and initiatives in partnership with inland terminals aiming at creating new logistics zones along the main hinterland corridors. For example, Hamburg Port Authority have invested in rail terminals (Melnik, Budapest, etc.) to support its rail traffic via Potzug, Metrans, and HHCE, whereas Barcelona Port Authority entered into partnerships to expand the port hinterland through the setting up of dry ports and logistics zones (i.e., Toulouse, Zaragoza, and Madrid).

As reported, the success of a port lies with the ability of the PA to collaborate with the maritime cluster and logistics players in order to create synergies for a broader regional logistics network. The development of these logistics poles generates benefits for all the actors involved. In particular, [Rodrigue and Notteboom \(2009\)](#) draw attention to the advantages originating from the cooperation between PAs and inland terminals that include, amongst other things, stronger support for the cargo handling function (due to a better use of port space area); expansion toward the hinterland, and the possibility to fight with competing ports for their market shares; a stronger position for luring (foreign) investments and subsidies because of more attractive, reliable, and frequent integrated hinterland services; the enhancement of intermodal services and better modal split in favor of rail transport; and the simplification of customs procedures.

The Emergence of Port Regionalization

As reported in the previous section, inland distribution is becoming a fundamental dimension of the maritime transportation and freight distribution paradigm. The development of global supply chains has increased the pressure on the main actors of maritime logistics, including shipping lines, terminal operators, rail transport operators, PAs, and inland terminals. In particular, hinterland accessibility constitutes a cornerstone for port competitiveness and, thus, the ability of a PA to develop port systems commercially oriented toward the hinterland is essential for its success.

In this regard, [Notteboom and Rodrigue \(2005\)](#) coined a new concept, known as “port regionalization”, to explain this new phase of growth of port systems, which was traditionally based on the activities confined to the port perimeter. In the regionalization model, instead, inland distribution becomes an essential driver for port competition, as it promotes the emergence of transport corridors and logistics poles. The authors argue that PAs do not lead this process as “motivators” or “instigators”. The process, conversely, derives from the collaboration and strategic coordination between shippers and third party logistics providers. In this vein, PAs are expected to embrace the regionalization process in order to face current port-related challenges, including congestion, limited handling capacity, and growing costs associated with additional traffic. Hence, hinterland accessibility resolves these problems by exploiting logistics corridors and rail transport to accelerate cargo flows and increase the handling capacity of seaport

terminals. Furthermore, regionalization extends seaport terminals toward the hinterland, moving the port closer to its customers and enhancing freight distribution.

Finally, regionalization enables PAs to go much beyond the role of a traditional facilitator. Indeed, PAs undertake an active role in the development of a more efficient supply chain, grounded on inland freight distribution, information systems, and intermodality (Notteboom and Rodrigue, 2005). In particular, PA networking activities represent one of the most peculiar aspects of regionalization. Coordination and cooperation among the maritime logistics actors constitute the pillars of this model and guarantee a valuable competitive advantage to seaports.

References

- Alphaliner, 2019. Top 100. Available from: <https://alphaliner.axsmarine.com/PublicTop100/>.
- Ferrari, C., Parola, F., Gattorna, E., 2011. Measuring the quality of port hinterland accessibility: the Ligurian case. *Trans. Policy* 18 (2), 382–391.
- Lloyd's List, 2018. Surging Volumes Drive Up Cosco Shipping Group's Earnings. Available from: <https://lloydslist.maritimeintelligence.informa.com/LL1122397/Surging-volumes-drive-up-Cosco-Shipping-Ports-earnings>.
- Muller, G., 1999. Intermodal Freight Transportation. Eno Transportation Foundation Inc., Virginia.
- Notteboom, T.E., Rodrigue, J.P., 2005. Port regionalization: towards a new phase in port development. *Marit. Policy Manag.* 32 (3), 297–313.
- Olivier, D., Slack, B., 2006. Rethinking the port. *Environ. Plan. A* 38 (8), 1409–1427.
- Rodrigue, J.P., Notteboom, T., 2009. The terminalization of supply chains: reassessing the role of terminals in port/hinterland logistical relationships. *Marit. Policy Manag.* 36 (2), 165–183.

Further Reading

- Blauwens, G., Van de Voorde, E., 1988. The valuation of time savings in commodity transport. *Int. J. Trans. Econ.* 15 (1), 77–87.
- Carbone, V., & Gouvenal, E. (2017). Supply chain and supply chain management: appropriate concepts for maritime studies, In J. Wang, D. Olivier, T. Notteboom, and B. Slack (Eds.), *Ports, Cities, and Global Supply Chains* (pp. 27–42). Aldershot (UK): Routledge.
- Debie, J., Guerrero, D., 2008. (Re)-spatialiser la question portuaire: pour une lecture géographique des arrière-pays européens. *L'Espace Géographique* 37 (1), 45–56.
- Fleming, D.K., Hayuth, Y., 1994. Spatial characteristics of transportation hubs: centrality and intermediacy. *J. Trans. Geogr.* 2 (1), 3–18.
- Magala, M., Sammons, A., 2015. A new approach to port choice modelling. In: *Port Management*. Palgrave Macmillan, London, pp. 29–56.
- Notteboom, T.E., 2008. The relationship between seaports and the inter-modal hinterland in light of global supply chains. Discussion Paper 2008-10, OECD-ITF.
- Robinson, R., 2002. Ports as elements in value-driven chain systems: the new paradigm. *Marit. Policy Manag.* 29 (3), 241–255.
- Song, D.W., Panayides, P.M., 2008. Global supply chain and port/terminal: integration and competitiveness. *Marit. Policy Manag.* 35 (1), 73–87.
- Van Klink, H.A., van Den Berg, G.C., 1998. Gateways and intermodalism. *J. Trans. Geogr.* 6 (1), 1–9.
- Vigarié, A., 1979. Ports de commerce et vie littorale. Paris (France): Hachette.

Seaports as Clusters of Economic Activities

Peter W. de Langen, Department of Strategy and Innovation, Copenhagen Business School, Frederiksberg, Copenhagen, Denmark

© 2021 Elsevier Ltd. All rights reserved.

Introduction	310
Synergies From Clustering	310
Clustering as a Perspective	311
Components of the Port Cluster	311
Particularities of Port Clusters	312
The Particularities of Port Clusters; Relevant for Analyzing Port Clusters	313
The Relevance of Synergy Services and Implications for Cluster Governance	313
The Business Ecosystem Perspective: Cluster Development in the Strategic Management Literature	313
Ports in Proximity: One Port Cluster?	314
References	314

Introduction

The application of insights from cluster theories to ports started roughly 20 years ago. The port cluster concept has been applied in practice: the port of Valencia has embraced the port cluster concept, the Chinese government uses the port cluster concept in port planning, the Korean Maritime Institute analyzes the potential of Korea's ports to further develop as logistics clusters and United Nations publications promote the development of port and logistics clusters.

Central to this cluster perspective is the recognition that interdependent firms cluster together in port regions, with various forms of synergies as a result (Porter, 2000). Anyone who drives around in a seaport, takes a boat tour in a port or flies over a port will note that ports are collections of many different activities, such as terminals, warehouses, chemical plants, tank storage, container depots, rail terminals, and cargo processing activities. All these activities locate in the port because of the agglomeration advantages that ports offer. Some of the activities in a port complex are very closely related (e.g., a tank terminal and a refinery), while others are more "loosely coupled" to each other (e.g., a container terminal and a port related ICT company).

It may be hard to imagine in our digital and service driven economy, but the major world cities today have emerged largely because of port activities. London was the world's largest port in the 19th century, to be overtaken by New York, while currently Shanghai is the largest port in the world. All these—and many other world cities such as Hong Kong, Singapore, Tokyo, Buenos Aires, Barcelona, and Venice—flourished in their early age as ports. The arrival, storage, and trade of goods in these port cities led to an agglomeration effect that re-enforced itself: ports attracted other economic activities, which led to population growth, which led to more economic activities, and so on^a.

Synergies From Clustering

All the companies related directly or indirectly to ships and international commodity chains and located in the port area can jointly be considered a port cluster. Synergies arise from the colocation of interrelated companies in ports. Some of these synergies relate to the shared use of infrastructure and services, others to the availability of specialized knowledge and skills.

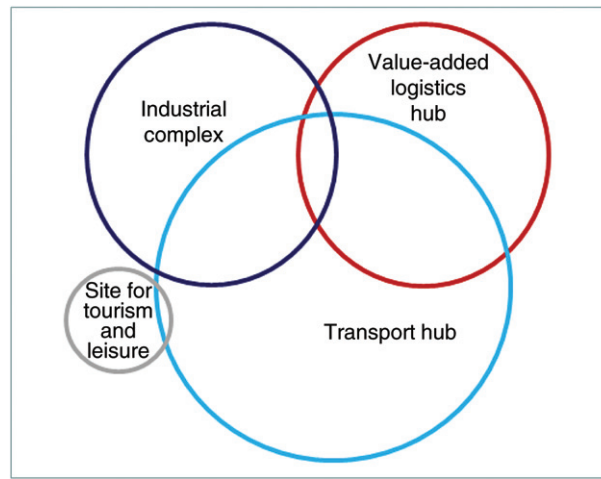
Some synergies arise "spontaneously," through transactions between the clustered firms. For example, the Siemens plant for rotor blades in the port of Hull will use third-party run terminal facilities for receiving parts and transporting the blades to the offshore wind locations. Siemens will also use energy from power plants in the port and contract transport operators that are established in the port. Engaging in transactions with suppliers and customers in the cluster is attractive since "trade costs" (that include transport costs) for transactions between firms in the same cluster are lower than those of transactions between a firm located in the cluster and another firm from outside the cluster. As another example of spontaneous, transaction-based synergies, because of the "constant market for skill," it is attractive for workers to invest in specific training and education. Thus, specialized training and education programs emerge in clusters and the overall quality of human capital availability is high.

Other colocation benefits arise as a result of *active cooperation* (beyond arms-length transactions) between firms in the cluster. For instance, firms may develop partnerships for innovation, jointly invest in the development of human capital and exchange data in a so-called "port community system."

^a In the words of Fujita, Krugman and Venables (leading scholars in economic geography, Krugman winning a Nobel Prize in Economics in 2009, p236): "Nearly all great cities are major ports (or were at one time), but many have long since ceased to be mainly port cities (. . .). There is nothing paradoxical in the fact that the Erie Canal made New York—and is now no more than a tourist attraction."

Table 1 Stylized characteristics of the transport node and cluster perspective on ports

	<i>Port as transport node</i>	<i>Port as economic cluster</i>
Definition	A gateway through which goods are transferred between ships and the shore.	An economic complex consisting of all firms related to the arrival of ships and cargo and located in one region.
Typical performance indicator	Throughput volume	Value added in the port (cluster)
Analytical models for analyzing the role of the government	Classification landlord, tool port and service port	Port authority as central organization in cluster governance.
Typical performance variables	Maritime accessibility Geographic location Hinterland connections	Intra-port competition Knowledge spill-over A qualified labor pool
Geographical focus	Specific terminals	Geographical and institutional proximity of actors in ports.

**Figure 1** Four components of a port cluster. Source: De Langen (2020).

Clustering as a Perspective

The port cluster concept has proven to be a useful “lens” for analyzing ports (de Langen, 2004). In port studies, the perspective that regards seaports as “transport nodes” is well established. The “cluster perspective” complements this “transport node perspective.” With the cluster “lens,” scholars address research issues that are complementary to the established lens of treating a port as a transport node. For example, the cluster perspective provides new insights into the determinants of port competitiveness, such as the importance of intracluster competition for the overall competitiveness of the port. In addition, the cluster perspective can be used to derive additional port performance indicators. Furthermore, the cluster perspective provides a lens for analyzing coordination and *collective action*, analyzed widely in cluster literature and very relevant in ports (de Langen and Visser, 2005). Finally, the cluster perspective yields additional insights regarding the appropriate governance of the port cluster. Table 1 shows stylized characteristics of the “cluster” and “transport node” perspectives.

Components of the Port Cluster

The “borders” of a cluster are somewhat vague in practice, and it may not be very fruitful to try to precisely define the port cluster^b. The port cluster does not just encompass freight operations, but also logistics, manufacturing, and maritime leisure activities (e.g., cruise and marina activities). Even though the borders are inevitably “blurred at the edges,” four components of a port cluster can be distinguished, as shown in Fig. 1.

The function of the port as a transport node is at the center of a port cluster: cargo-handling terminals are the *raison d’être* of the port infrastructure. Since cargo handling is a part of a transport chain, ports also attract transport companies, as well as related specialized activities, such as storage, ship repair, and ship management. All these activities can be regarded as elements of the function of the port as a transport hub.

^bThe cluster perspective has also been applied with a focus on “logistics clusters” (Sheffi, 2012). While there is a clear overlap with port clusters, the “delimitation” of a logistics cluster is different.

Table 2 Activities often included in each of the four components of a port cluster

<i>Transport hub</i>	<i>Logistics value added hub</i>	<i>Manufacturing cluster</i>	<i>Tourism and leisure</i>
Handling and transport of containers (deep sea and short sea), oil & oil products, chemical products, LNG, agribulks, biofuels, cars, iron ore, coal, construction materials, and other commodities.	LCL consolidation and deconsolidation Heavy/special equipment storage & stockholding Stock holding & value adding for slow moving products (B-to-B and to a lesser extend consumer goods). Stock holding & value adding for fast moving products (consumer & B-to-B goods). Fuel blending Refrigerated warehousing	Food processing Energy production Downstream chemicals Upstream metals industries Car/truck manufacturing Shipbuilding & ship repair (including maintenance—all ship types) Downstream metals industries Upstream chemicals	Cruise activities Large & mega yachts Marina Congress, hotel & leisure facilities.

Logistics activities, such as storage, repacking, and assembling are frequently located in seaports. Two reasons explain the presence of logistics activities in seaports. First, cargo is almost always temporally *stored* in ports, because of the scale differences between ships and inland modes (e.g., a 300,000 ton ship vs a 3,000 ton train and a 20,000 TEU ship vs a truck for 2 TEU). This necessity of storage makes ports natural (de) consolidation points and provides a rationale for locating logistics activities (such as blending and repacking) in seaports. Second, a port is an attractive location for logistics activities due to the high (international) connectivity. Thus, goods can be delivered to end markets, both over land and by sea, at lower costs than from most other locations. In addition, many ports are located in metropolitan areas and, thus, are well located for distribution facilities to these metropolitan regions. Included under the label logistics are also *trade activities*. Traditionally, trading and storage (in a port) are closely linked. All major ports were also commercial hubs. Although the strength of links between storage and trade has diminished, for some commodities (for instance fruit, crude oil, and grain), trading and related service activities are still located in the port—and part of the port cluster.

Examples of *manufacturing activities* frequently strongly linked to ports are energy production, steel production, shipbuilding and repair, oil refining, chemical production, and food production. Manufacturing activities in seaports are directly related to commodities handled in the port cluster. Manufacturing is not necessarily limited to large-scale plants; ports may also be attractive locations for small-scale manufacturing, as well as combinations between manufacturing and service activities.

Finally, ports increasingly also function as tourism sites. Ports accommodate cruises and (mega) yachts and may host tourism and leisure facilities such as hotels, restaurants, shops, and museums. In addition, a part of the (former) port land may be a public space. Because of the attractiveness of waterfronts, the potential for tourism-related developments in ports is huge. While such developments may be somewhat isolated from the freight-oriented activities, there are relations with the other port functions. For instance, RoRo vessels carry passengers, part of them tourists, as well as freight. **Table 2** shows activities often included in each of the four components of a port cluster.

Particularities of Port Clusters

Even though the cluster perspective is useful, ports are not the “typical clusters” often described in textbooks and on the mental map of policy-makers, such as Silicon Valley. **Table 3** summarizes the differences between “port clusters” and stylized “tech clusters,” these differences are described in further section.

First, the economic activities in a port are mainly *operational*, since ports, as well-connected pieces of infrastructure for freight flows, are specialized in handling and processing of cargo. This operational specialization implies limited scope for “spontaneous” interactions that may lead to path-breaking innovations. For instance, even the introduction of the maritime container was done by an industry outsider. This is radically different from the widely studied “high-tech clusters,” that are typically associated with spontaneous, noncontractual interactions, and nontradable interdependencies in pitching new ideas and raising venture capital for innovative exploration (Maskell, 2001), often depicted to take place in the coffee corners of Palo Alto or Tel Aviv.

Second, port clusters consist of a *diverse* set of potentially interrelated activities. This is because these port activities are linked to transport and port infrastructure, and include all kinds of activities (terminal operations, transport operations, warehousing, manufacturing, and trading), in all kinds of supply chains (including food, energy, automotive, steel, and so on).

Third, given the international nature of the activities in ports mentioned earlier (transport, logistics, manufacturing, and trade), multinational corporations (MNCs) are dominant players in port clusters. These MNCs have operations in various countries and are thus (more or less) embedded in a variety of port clusters. As a consequence, in most port clusters, only a small part of the clustered firms (or even none) is headquartered in the port. In addition, most MNCs manage operations in a large number of port clusters. This global operational presence in port clusters allows for the diffusion of operational knowledge through the regular transfer terminal managers, but hampers exploration.

Fourth, a “port cluster” developer may play a central role in the governance of the cluster. In most ports, a publicly owned and territorially focused port authority holds a key responsibility in developing the port cluster. Indeed, this organization may think of itself as a “cluster developer.” In many landlord ports, the port authority is run as an autonomous and commercially operating entity, it does not aim at maximum profits, but is more orientated toward creating “value for society.”

Table 3 The differences between “port clusters” and stylized “tech clusters”

	<i>Port cluster</i>	<i>Tech cluster</i>
Type of activities	Operational, capital intensive.	Human capital intensive.
Heterogeneity of clustered activities	High. The type of activities that cluster together in ports range from transport to manufacturing to advanced services. The core driver of clustering is the presence of port infrastructure and the resulting international connectivity for freight.	Low to medium. A technological specialization (medical, media, materials) is often the core of a tech cluster.
Dependence on extra-cluster corporate decisions.	Very high. Most activities located in ports are operational activities that are managed by local branches of MNCs with limited autonomy.	Low to medium. Substantial decision-making powers in the cluster, a relatively high share of the firms are headquartered in the cluster or have a high degree of autonomy.
Importance of “cluster governance” and cluster governance structures	Cluster governance is very important and relatively formalized. Given the need for investments (in tangibles and intangibles) for creating synergies between companies in the cluster (e.g., pipelines, transport infrastructure, a “port community system”) a formalized governance structure is required. In most ports, a territorially focused and often state-owned port development company plays a central role in the governance of the cluster.	Relatively unimportant. The clusters, as well as synergies in the cluster evolve dynamically, market-based initiatives to create synergies dominate (e.g., incubators). The most important collective interest is human capital formation, universities often play a leading role in this respect. Governance is mostly informal and temporal.

The Particularities of Port Clusters; Relevant for Analyzing Port Clusters

The differences between the port cluster and the tech cluster as discussed above have three important implications for the port clusters, as well as for the analysis of such port clusters. First, much more than the “tech clusters,” “port clusters” are vulnerable to technological change, especially given the clear (and sometimes disruptive) impact of new technologies on ports (for instance, think of digitalization, the energy transition, 3D printing, circularity of supply chains).

Second, contrary to most of the work on innovation in clusters, for ports, the core challenge is not so much improving the innovation performance (as measured by indicators such as patents, start-ups, and scale-ups), but enhancing the *absorptive capacity* (Giuliani, 2005) of the port. Because of the characteristics of port clusters (diverse, specialized in operations, and dominated by local branches of MNCs), R&D activities in port clusters are limited and the dominant innovation challenge is the (early) application of new knowledge (e.g., regarding digitalization, recycling technologies, bio-based chemicals, and smart grids). Thus, the vitality of the port cluster depends mainly on its absorptive capacity, which can be defined as the capacity to absorb, enhance, diffuse, and exploit knowledge from extra-cluster sources. The absorptive capacity is determined both by the formation of linkages with extra-cluster sources of knowledge and the intra-cluster knowledge system. For ports, the most relevant indicators are first the investments in the port cluster to apply new technologies in new or established operations and the number and growth of start-ups with “cluster specific” products and services based on new technologies.

Third, contrary to most “tech clusters,” in which governance is rather loose, chaotic, and decentralized, more structured governance mechanisms may be critical to enhancing the competitiveness of port clusters.

The Relevance of Synergy Services and Implications for Cluster Governance

One of the core challenges in developing the port cluster is creating synergies. As argued previously, these do often not emerge spontaneously. *Synergy services* can be defined as services provided to a group of geographically clustered companies that allow these companies to create synergies (Siskos and Van Wassenhove, 2017). Fig. 2 shows examples of such synergy services for each of the four components of the port cluster.

The core challenge of developing synergy services is creating value for users and capturing (a part of) the created value through charges. As the value of the synergy services often increases with the number of users, capturing value directly through pricing may not be attractive, as it reduces the number of users. In addition, users may be deterred by the risk of high costs associated with ending the use of synergy services (e.g., a pipeline), which may allow the synergy service provider to extract an economic rent. Finally, synergy services generally create “value for society” (e.g., reduction of emissions, truck movements, congestion, and the like). This justifies partial public funding for such services. For these reasons, a “for profit approach” for synergy services is often problematic. Consequently, a “not-for-profit approach” may be required, for instance through service provision by a “special purpose vehicle.”

The Business Ecosystem Perspective: Cluster Development in the Strategic Management Literature

In strategic management, the term business ecosystem is used for the network of firms, which collectively produce a system, service, or product that creates value for customers (Iansiti and Levien, 2004). A cluster can be regarded as a territorially focused business

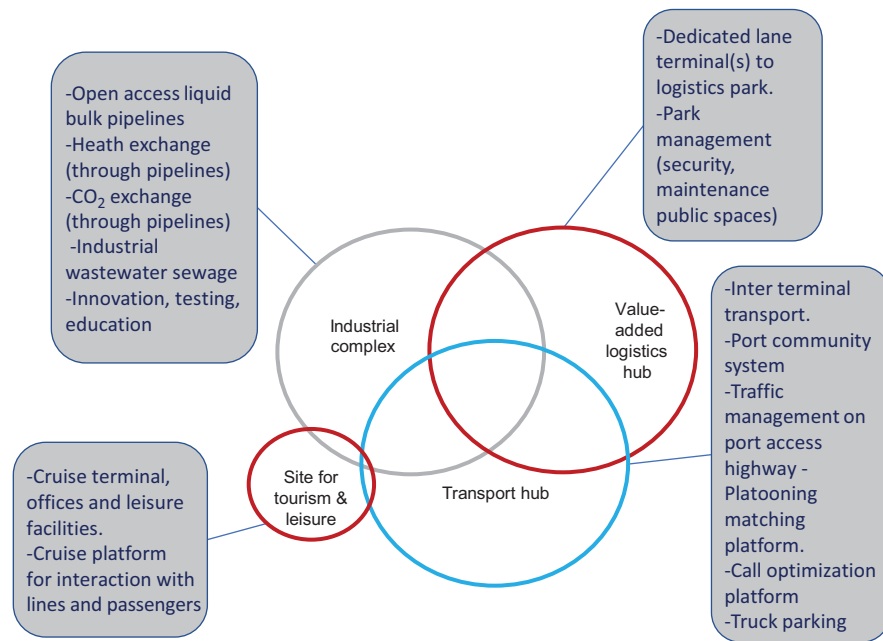


Figure 2 Examples of synergy services in port. Source: *De Langen (2020)*.

ecosystem. A first strategic management insight is that many companies do not compete on their own with other companies, but instead, business ecosystems compete with each other; the ecosystem of companies that constitute the windows, respectively iOS business ecosystems compete. Likewise, the European and North American business ecosystems for design and production of aircraft are competing. Applied to ports, on top of the obvious competition between individual terminal operators in different ports, the competition between the ports as business ecosystems is relevant.

A second insight from strategic management scholars is that some companies actively develop a business ecosystem. Such companies need a specific set of capabilities and have a specific business model and are labeled “keystones.” A keystone develops and regulates the overall function of the ecosystem and, as a consequence, its actions influence the success of all other members. In port clusters, port authorities can be considered as companies that develop a business ecosystem. Port governance structures that enable port authorities to effectively play this role are required. In line with the arguments provided above, an ideal type port governance structure is emerging in which a state owned, but autonomous and commercially operating, port development company develops the port as a business ecosystem and provides various types of “synergy services.”

Ports in Proximity: One Port Cluster?

One of the key questions concerning port clusters is: “in which cases can ports in proximity be regarded as one port cluster and thus are best developed in an integrated manner?” In many parts of the world, initiatives to develop port policies for groups of ports or gateways can be observed: Chinese policy makers regard the Pearl River delta ports as one port cluster and the ports in the Adriatic sea position themselves as one cluster. The state-owned port development companies of Copenhagen and Malmo and Ghent and Zeeland Seaports are even fully merged, even though they are located in different countries.

In the case of the United Kingdom, where port development is mainly done by private companies, mergers between port companies that develop ports in proximity have taken place; there is a natural tendency toward the appropriate geographical scope. Such a mechanism is absent in cases where port development companies are state-owned. Consequently, the geographical scope of publicly owned port development companies is often not optimal. The appropriate geographical scope is one publicly owned port development company for one port cluster. In this case, over investment in similar facilities is prevented, while the port development company remains focused on the development of one integrated port cluster.

References

- Giuliani, E., 2005. Cluster absorptive capacity: why do some clusters forge ahead and others lag behind? *Eur. Urban Reg. Stud.* 12 (3), 269–288.
 Iansiti, M., Levien, R., 2004. *The Keystone Advantage: What the New Dynamics of Business Ecosystems Mean for Strategy, Innovation, and Sustainability*. Harvard Business Press, Cambridge.

- de Langen, P.W., 2004. The performance of port clusters; a framework to analyze cluster performance and an application to the port clusters of Durban, Rotterdam and the Lower Mississippi, ERIM and TRAIL thesis series. (No. ERIM PhD Series; EPS-2004-034-LIS). Available from: hdl.handle.net/1765/1133.
- De Langen, P.W., 2020. Towards a Better Ports Industry: Principles of Port Management and Development. Forthcoming at Routledge, London.
- de Langen, P.W., Visser, E.J., 2005. Collective action regimes in port clusters: the case of the lower Mississippi port cluster. *J. Transp. Geog.* 13, 173–186.
- Maskell, P., 2001. Toward a knowledge-based theory of the geographical cluster. *Indus. Corp. Change* 10 (4), 921–943.
- Porter, M., 2000. Location, competition, and economic development: Local clusters in a global economy. *Econ. Dev. Quart.* 14 (1), 15–34.
- Sheffi, Y., 2012. *Logistics Clusters: Delivering Value and Driving Growth*. MIT press, Cambridge.
- Siskos, I., Van Wassenhove, L.N., 2017. Synergy management services companies: a new business model for industrial park operators. *J. Indus. Ecol.* 21 (4), 802–814.

Port Efficiency and Effectiveness

Lourdes Trujillo, María Manuela González, Casiano Manrique-De-Lara-Peñate, Ivone Pérez, Department of Applied Economics, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	316
Port Effectiveness Measurement	316
Port Efficiency Measurement	317
The Frontier Function	318
Empirical Analysis in Port Efficiency	320
Conclusion	321
References	321
Further Reading	322

Introduction

Port efficiency and effectiveness are normally two complementary components of the broader concept of port performance. Although the academic analysis of effectiveness is more recent than that dealing with efficiency, there is still much to be done, if the sector wants to adequately expand performance measurement programs.

Efficiency, effectiveness, and productivity are related concepts. While efficiency means producing as much as possible with a given quantity of inputs or producing a specific number of goods with a minimum amount of inputs, effectiveness tends to put its emphasis on the users. In the words of various authors, being efficient is about “doing things right,” whereas being effective is about “doing the right things.” The right things are considered those that are identified as such by the users. On the other hand, for a company to be productive, in addition to being efficient, it has to have an optimal scale, that is, an adequate size.

The distinct services offered by a port can be performed by a combination of public and private initiatives. There are a number of port models that show how private participation might be introduced. The most common model is the landlord model in which the port authority (PA), generally public, is the owner of the infrastructure and the remaining services are offered by private port operators. The provision of infrastructure by the PA and the cargo handling by private operators are the most relevant services in terms of port efficiency. Other services can be provided by private port operators working in more or less competitive conditions and, therefore, it is expected to be provided efficiently.

The port industry has been immersed in a process of permanent change for decades. During the last century in many ports worldwide, there was a change of governance toward this decentralized model of port management, along with a greater increase in the presence of private port terminals that would alleviate public budgets and assume part of the risk of port operations. More recently, the increase in ship sizes, the alliances between shipping operators, especially in container traffic, and economic globalization, have all increased competition and led to new services and port logistics. All these changes will be successful if they are accompanied by improvements in port efficiency.

Since the first studies on port efficiency in the 1990s, publications in this area have increased. However, one characteristic stands out in a review of the empirical literature on port efficiency: “diversity.” Studies on port efficiency seem to approach the subject from multiple perspectives. For example, different methodological approaches are used to study port efficiency. The same can be said about other factors in the reviewed studies, including: port services analyzed, inputs and outputs, variables influencing or explaining inefficiency, sample size, and the time period. This “diversity” or variety makes comparisons between research results difficult. However, it is important to note that most articles are dedicated to studying the provision of infrastructure and cargo handling services. In this article, we conducted a review of the methodologies used to estimate port efficiency and effectiveness in order to highlight their characteristics, scope, and limitations.

Port Effectiveness Measurement

The study of effectiveness tends to emphasize the users, who are the final recipients of the services and who tend to influence ports and government in order to improve port performance. Therefore, their perspective should be actively incorporated into the analysis. Efficiency mainly relates to physical aspects and researchers have at their disposal methodologies that allow it to be objectively measured, whereas effectiveness can only be measured against the goals that the users seek and is usually based on qualitative indicators (Brooks and Pallis, 2008).

Normally, any measure that improves efficiency would positively influence the level of effectiveness, but the opposite may not necessarily be the case. Users tend to highlight issues relating to the ability to respond to special requests (working overtime or at

weekends) and relating to cargo damage as important. Maintaining resources ready to respond to potential unexpected situations or security issues may not necessarily be efficient, as these demands may introduce higher costs while managing a similar level of throughput.

The first step involved in the measurement of port effectiveness consists in identifying the main stakeholders in a port. The main customer is obviously the carrier and anybody that buys maritime transportation services, whether in their own name or in that of others, but other agents should also be considered. In fact, many elements in the supply chain may also be clearly interested in the well-functioning of ports, even if they are located outside of their physical location. This can be the case for many intermodal distributors, forwarders, warehouses, and road and rail transport companies.

The second step consists in identifying the main aspects of interest for users. To prepare the list of performance criteria, the authors normally review previous academic studies and then employ surveys directed at users, either to obtain an evaluation of the identified criteria or to ask them directly for those they think fulfil their needs. Once the relevant criteria for users have been identified, the last step involves users, evaluating how a particular port performs in relation to those attributes.

The combination of these two valuations—what aspect is important for the user and how the port behaves in relation to that criterion—can help in identifying the level of performance of a specific port from the users' points of view. A similar procedure can be used to test the performance from the point of view of the PA. If the PA undertakes its own evaluation of the criteria, this internal evaluation can be compared with that of the users, and even with the evaluation of other ports. These comparisons could help in identifying potential performance gaps, especially if they are evaluated over time.

Importance–performance analysis is a technique that has been used by Brooks et al. (2011) for evaluating user satisfaction and service quality, combining what users identify as important drivers with how they appraised a specific port performance. Attributes can be considered determinant, if they help us understand how the user discriminates between various alternatives. The main problem in this procedure is the fact that both elements, importance and performance valuations, are highly correlated and, therefore, generate multicollinearity problems. Normalised pairwise estimation has been identified as a convenient procedure under these circumstances. After rating the importance of the considered criteria, users are asked to indicate the level of all defined attributes for each port. They are also asked to evaluate the general performance of each port using variables like satisfaction, competitiveness, and effectiveness of their services. The determinance importance of each attribute is obtained through a correlation analysis, whereas the general performance of the port is explained by the level of the attributes given by users to the port.

Recently, Ha et al. (2017) proposed the use of a hybrid multistakeholder framework for the modelling of port performance indicators (PPIs). This new port performance measurement (PPM) methodology allows for the integration of a multistakeholder dimension. These authors actually consider a wide group of dimensions ranging from core activities like output—as measured by throughput growth—through aspects like financial strength, user satisfaction, supply chain integration, and sustainable growth. In other words, they do not limit themselves to efficiency and effectiveness measures.

This hybrid PPM model has four steps. The first consists of identifying the needs and aims of the different stakeholders and presenting them as PPIs adapted to each group. The second employs methodologies that solve the problems generated by interdependency, uncertainty of the PPIs and the need to use multigroup multicriteria decision-making methods. The third assigns weights to the different PPIs evaluating their performance for each port. Finally, the PPIs are evaluated with their weights using “fuzzy evidential reasoning” and utility techniques.

Port Efficiency Measurement

Comparing ports in terms of their performance is relevant for economic analysis and crucial for the PA and port operators. The most widely used concept to establish this comparison is productive or technical efficiency. Efficiency arises from the comparison of observed values of a firm's outputs and inputs with optimum values from the optimum evidence provided by other firms. That means efficiency is a relative concept that can vary depending on the firms being compared and that it is necessary to have a standard to compare firms.

Economic efficiency has two dimensions: technical (or productive) efficiency and allocative efficiency. Firms that produce the maximum output from a given set of inputs or, alternatively, firms that use the minimum quantity of inputs to obtain a given output level represent technical efficiency. Firms that use inputs in the right proportions, given their prices and marginal productivity, are known as “allocative efficient.”

The frontier concept, introduced by Farrell (1957), allows us to measure technical and allocative efficiency as shown in Fig. 1A (input-oriented efficiency) and Fig. 1B (output-oriented efficiency). We use the input-oriented approach to introduce technical and allocative efficiency (Fig. 1A). The efficient production function is $Y = f(X_1, X_2)$, where Y is the output obtained with two inputs X_1 and X_2 . If constant returns to scale are assumed, the production function becomes the isoquant (SS') $1 = f(X_1/Y, X_2/Y)$.

Firms Q and P use the same input proportions, but only Q is on the isoquant. Q and Q' are technically efficient, and Q' is also allocative efficient (as represented by the point of tangency between the isoquant and isocost curves). Following an input-oriented measurement (Table 1), it is possible to produce the same quantity of P using the ratio OQ/OP inputs. This ratio ranges from 1 if the company is efficient to 0, if it is inefficient. Comparing Q (in the same isoquant as Q') and R (in the same isocost as Q'), allocative efficiency can be measured as the ratio OR/OQ . If this ratio is 1, the firm is allocative efficient, whereas values less than 1 indicate the degree of allocative efficiency achieved by the company. Finally, economic efficiency can be calculated by multiplying technical and allocative efficiency (Table 1). The degree of inefficiency is calculated as one minus the corresponding efficiency ratio.

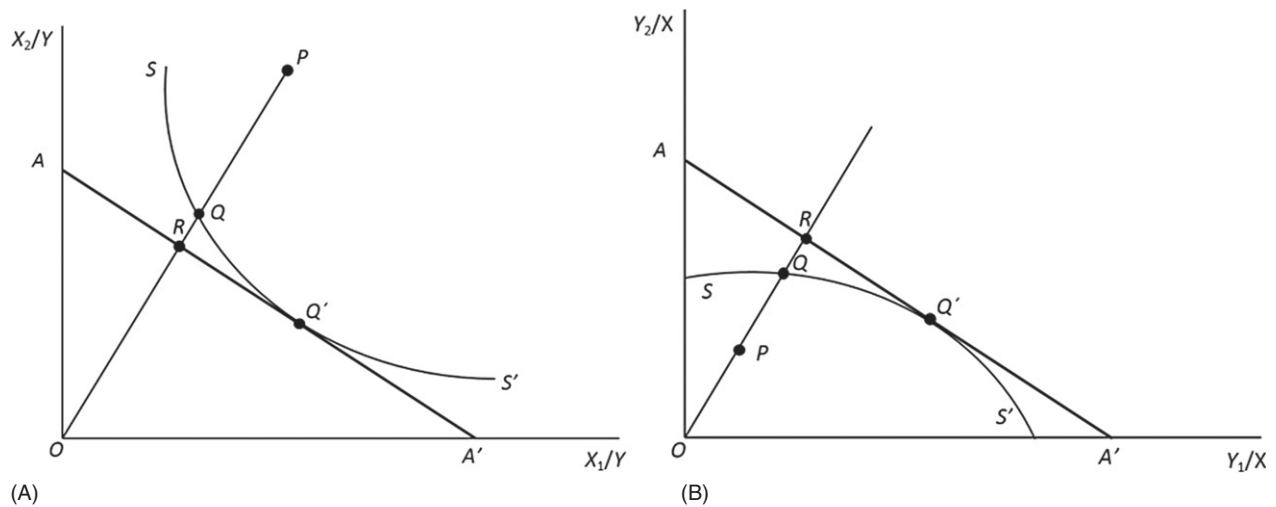


Figure 1 (A) Input-oriented efficiency. (B) Output-oriented efficiency.

Table 1 Technical, allocative, and economic efficiency measurement

<i>Input-oriented efficiency</i>	<i>Output-oriented efficiency</i>
Technical efficiency = OQ/OP	Technical efficiency = OP/OQ
Allocative efficiency = OR/OQ	Allocative efficiency = OQ/OR
Economic efficiency = OR/OP	Economic efficiency = OP/OR

In the assessment of port performance, the calculation of technical efficiency prevails. The main reason is that it is less demanding of data, it only requires information on quantities of inputs and outputs, whereas efficiency in costs and allocative efficiency require data on input price or cost.

The Frontier Function

Efficiency measurement is joined to the estimation of a frontier. Once the frontier has been estimated, the efficiency shows how the firm behaves in relation to the performance of the best companies in the industry, if they were faced with the same conditions as the firm analyzed.

Frontier models can be classified according to several criteria. Taking into account the type of data they use, there are models of cross-sectional and panel data. Suppose a port intends to invest in infrastructure and equipment to address a future demand. If the cross-sectional analysis is carried out before the port meets the demand, this port will be classified as inefficient. As the effect of the indivisibilities of the port infrastructure cannot be captured with cross-sectional models, it is advisable to use panel data.

Taking into account the type of variables (physical and/or prices) and the assumed assumptions about the optimization behavior of the companies, the frontier modalities shown in Table 2 can be distinguished. The most elementary assumption is that companies do not waste productive resources. In this case, a production frontier will be estimated, which allows us to obtain an evaluation of technical efficiency.

The choice of the frontier model depends on the hypothesis about the behavior of the producers most appropriate to the case study. Most of the empirical applications to the port industry estimate the production frontier and, therefore, technical efficiency, although there are a few studies that estimate the cost frontier. The main reason is that it is much more difficult to access input price data or financial data on cost and revenues than physical data related to outputs and inputs (e.g., cargo handled, berth length, yard area, cranes, etc.).

Table 2 Frontier and efficiency typology

<i>Optimizing behavior</i>	<i>Frontier</i>	<i>Efficiency</i>
Maximum production	Production	Technical efficiency
Minimum cost	Cost	Economic efficiency in cost
Maximum income	Income	Economic efficiency in income
Maximum benefit	Benefit	Economic efficiency in benefit

Table 3 Characteristics of DEA and SFA

DEA	SFA
• Nonparametric approach • Deterministic approach • Does not consider random noise • Does not allow statistical hypothesis to be contrasted • Does not carry out assumptions on the distribution of the inefficiency term • Does not include error term • Does not require specifying a functional form • Sensitive to the number of variables, measurement errors, and outliers • Estimation method: mathematical programming	• Parametric approach • Stochastic approach • Considers random noise • Allows statistical hypothesis to be contrasted • Carries out assumptions on the distribution of the inefficiency term • Includes a compound error term • Requires specifying a functional form • Can confuse inefficiency with a bad specification of the model • Estimation method: econometric

According to the estimation method, it is possible to differentiate two categories: estimated frontiers through linear programming and those that use econometric methods. The frontier can be defined by assuming a particular functional form for the technology or from a series of properties that it must satisfy, which allows the distinguishing between parametric and nonparametric methods, respectively. Considering the cause of the separation from the frontier, these are classified as deterministic and stochastic. Finally, within the econometric approach it is possible to establish a distinction based on the number of equations of the model (single equations, or systems of several equations).

As in other productive sectors, two approaches have been developed to measure efficiency in the port industry (Charnes et al., 1997; Fried et al., 1993; Kumbhakar and Lovell, 2003): the linear programming method, basically data envelopment analysis (DEA) and the econometric model, mainly stochastic frontier analysis (SFA), with the former being the most used. The main characteristics of these two approaches (Lovell, 1993) are presented in Table 3.

DEA assumes that there are no errors in data collection. For the port industry, this is a quite strong assumption to make. More often than would be desirable, port statistics contain errors in the measurement, especially in the input variables. Bootstrap resampling methods can ease this problem, although with its own limitations. Stochastic frontiers do not suffer from this potential flaw, so it appears to be more appropriate to evaluating efficiency in the port sector when there are suspicions of data collection errors.

Since the first DEA model, a number of improvements have been developed: from constant returns to scale to variable returns to scale, additive model, from radial efficiency measures to nonradial measures, super efficiency models, DEA window model, partial or metafrontiers, etc.

SFA requires the specification of a functional form and makes assumptions related to the distribution of the error term. The Translog production function and normal distribution for the random error and seminormal for the nonnegative random variable associated with technical inefficiency are the most used in empirical applications.

As the first SFA models specified for cross-sectional data, the stochastic frontier approach has evolved in an attempt to overcome the limitations of previous models. Thus, models were designed to accommodate panel data, individual efficiency, variant inefficiency over time, the factors that determine inefficiency, multiple outputs (in stochastic distance functions), etc.

As the interest of researchers moved beyond simply establishing a ranking of ports in relation to efficiency, the focus shifted to explaining the causes of port inefficiency. As a consequence, the models were changed to adjust to the new research objectives.

The DEA approach has been mainly completed with Tobit models, in which the efficiency indices are explained by efficiency explanatory variables and, more recently, with bootstrapped truncated regression. At the beginning, SFA models in two stages was widely criticized, since some of the assumptions established in the first stage are breached in the second. Subsequently, specifications were developed in which the effects of inefficiency are defined as a function of the specific factors of firms that affect efficiency, carrying out the estimation in a single stage.

In the empirical measurement of efficiency in ports, DEA is the approach that is most frequently applied. Ports are heterogeneous entities that include not only port services provided by port authorities, but also services provided by a range of agents and operators. Ports have a multioutput dimension (e.g., services to ships, several types of cargo and passengers). In part, the ease of accommodating multiple outputs explains the prevalence of DEA. The stochastic distance function solves this limitation of stochastic production frontiers without imposing additional restrictions. However, stochastic production functions predominate, perhaps because many empirical works refer to container terminals. Even in this case, it is likely that the terminals also move general cargo, with distance functions being the appropriate approach.

Regarding which approach is best suited to the port sector, some researchers argue that DEA has a business orientation, whereas the econometric approach is more oriented to policy decisions. So one recommendation might be to apply DEA when estimating the efficiency of individual cargo terminals and the stochastic frontier if the estimation relates to port authorities. The characteristics of the data, the port activity analyzed, and the assumptions that can be made will also guide the choice of the efficiency calculation method.

Empirical Analysis in Port Efficiency

In recent decades there has been a continuous increase in the number of academic papers focusing on port efficiency, along with an expansion of the geographic scope of these studies (Cullinane, 2013; González and Trujillo, 2009).

In the literature on port efficiency, two activities are the most studied: the provision of infrastructure service by port authorities and the cargo handling service provided by stevedoring companies and predominantly carried out in port terminals. Regarding cargo handling, most research has focused on container terminals and some has analyzed the aggregate of all container terminals in a port, as the separate output for each company is not available from main secondary data resources. This limits the scope of the studies, since it would be more useful to know the efficiency of each terminal and investigate the causes of the differences between them.

In addition to estimating efficiency levels, research on port efficiency seeks to shed light on the impact of specific events (e.g., port reforms, type of governance, port size, etc.) on efficiency, with the aim to provide evidence that assists policy or management decisions. The diversity of approaches and variables used makes it difficult to compare the works, while offering opportunities to continue analyzing such issues.

Most ports have been involved in regulatory changes that comprise, among other aspects, the decentralization and privatization of port operations. One of the most studied aspects is whether port reforms and the ownership structure influence port efficiency, with inconclusive results. These two aspects are often related to each other, and it is not easy to introduce them into the efficiency analysis. Most of the works that analyze the effects of the reforms do not introduce a specific variable; they simply undertake the analysis in a time period that encompasses before and after the reform, so that the observed variations in efficiency are attributed to the reform, which is a major assumption. Other studies use a naive dummy variable (1 if it is private and 0 otherwise). Classifying ports as public versus private property is very simplistic and hides the complexity of the structure of port ownership and governance, so it would be better to use a more multidimensional approach, such as the port function matrix (Baird, 1995), as has been done in a number of papers.

Under the hypothesis that large ports are more efficient, a number of authors have analyzed this issue. Port size has been approximated by the length of quays, the number of quays and by the output (TEU, tons, or dummy variable). The literature review does not provide conclusive results. Port size is closely related to other aspects such as level of complexity, competition, economies of scale, transshipment traffic, and location. The lack of agreement on the effect of port size on efficiency is perhaps due to the interrelation of all these aspects or can be related to a failure of the exogeneous regressors assumption.

In relation to the role of port competition as a driver of port efficiency, two hypotheses can be contrasted. Ports subject to high competition levels may have an incentive to overinvest in capital, which may lead them at some point to lower levels of efficiency. On the other hand, it can also be argued that the incentive in these competitive environments leads to innovation and, in turn, to an efficiency improvement. The empirical evidence is not yet conclusive, so this is a topic that can continue to be explored.

With respect to the data, the benefits of data panels are highlighted compared to the use of cross-sectional data. While the latter provides static information and reflects differences among units of analysis at particular points of time, the former allows us to observe both the differences between the units of study over time and their evolution. Panel data sets better capture the indivisibilities of the port infrastructure, providing richer information for policy decisions.

We offer a detailed and critical description of the main variables used to measure port output and inputs, as well as the variables that determine port efficiency.

- In spite of the fact that ports have a multioutput dimension, most authors focus on cargo as the port output, mainly measured in tons, whereas a few use a monetary approximation. Recent works also include the number of vessels, the berthing time, and considering the cruise boom, the number of passengers is also gaining special relevance.
- It is useful to distinguish several types of cargo: solid bulk, liquid bulk, break bulk and containers, each one requiring specific facilities, and services. However, not all the papers reflect these differences in cargo. Monoproduction is used in two cases. Papers that focus on a port as a whole use total throughput in tons. When the objective is to analyze container terminal efficiency, the output is containers (in tons, number or TEUs). However, as general cargo is also usually handled in container terminals, it might be useful to distinguish between gateway traffic and transshipment traffic as different outputs. However, the lack of disaggregated data have led to the use of global measures in many of the studies.

In relation to output, the recommendation is to try to reflect the multiproductive nature of port services; both DEA and the stochastic distance function allow for this.

Inputs can be measured through physical and monetary approaches. Most of the research use capital and labor to carry out the port activity, although there are differences in the way of measuring them.

- The capital factor is the the most controversial input. This factor includes the docks (length and depth), terminal area, and mechanical equipment. Quay length and depth indicates the ability of the port/terminal to accommodate ships. Larger quays will be able to serve larger vessels with higher loading levels. The port draft constitutes an entry restriction on certain types of ships. Likewise, the terminal area also represents the storage capacity of the port. The mechanical equipment refers both to the cranes needed to handle the cargo from/to the ships and the equipment involved in the organization and storage of the merchandise in the terminal. Most studies measure capital in physical terms, although some of them use monetary approximations, such as the net value of capital or capital depreciation.

- Depending on the type of cargo, the equipment differs. Containers are handled via cranes on terminal docks (ship-to-shore cranes) or on board the ship (ship cranes) and their storage is done by other equipment (e.g., tow truck, reachstacker, straddle carriers, forklift trucks). Rolling loads do not require cranes. Bulk cargo is handled using a system of suction pipes or grab cranes (both on dock and on ship). Liquid bulk can be found both in liquid and in gaseous state, and its discharge (or loading) is carried out by pipes.
- The inclusion of mechanical equipment in an efficiency study is a challenge. The existence of different types of cranes with disparate characteristics – such as outreach, speed or maximum lifting capacity – makes it difficult to compute them. Something similar happens with storage resources. One of the main drawbacks of the data set is that they do not reflect the technological improvements in cranes.
- Several measures are used to include mechanical equipment in efficiency studies. Although some use the number of operative hours, the most common practice is to include the number of cranes available, differentiating by typology in some cases (quay cranes, yard gantries, and straddle carriers). Other studies include the load handling capacity of the cranes, also differentiating by typology, which makes more sense.
- In summary, it is necessary to develop approaches that capture the technological characteristics of modern cranes. It is always preferable to use the load capacity of the cranes than the number of cranes. If data are available, it would be ideal to work at a level of disaggregation that reflects the lifting capacity of the different types of cranes. If it is not possible, it is necessary to develop an index that reflects all these characteristics and allows comparison between the different cranes.
- Regarding labor input, two types of workers are included: those who work for the port authorities and the stevedores, who work for the cargo terminals. The labor input can be measured as numbers of employees, hours worked or work shifts, as well as in terms of remuneration or labor costs, with the most common being the number of workers.
- Measuring the port workforce presents difficulties. While information concerning the employees of a PA is relatively easy to find in annual port reports, to obtain data regarding the stevedores is difficult, especially when comparing different countries, because it is often necessary to carry out questionnaires in ports and terminals, whose response rate is usually low. This has led, in some studies, to use the employees of the PA to define the service of handling cargo, which does not reflect the reality of the service and can lead to incorrect results. Other papers, to avoid the labor factor, argue that there is a constant relationship between the number of cranes and stevedores, so they use a ratio to the number of cranes as an approximation of the number of stevedores. However, this relationship may be affected by technological issues and the authors themselves warn that this assumption limits and conditions the results obtained.
- Some research also add other input variables. These are measured mainly in monetary terms to capture inputs such as energy, intermediate consumptions, etc.

Conclusion

Making comparisons between the results and conclusions of research that measure efficiency in the port industry is not an easy task. The diversity of methods and specification of the functions, the differences in the variables of inputs, outputs, determinants of efficiency, the different time periods and countries compared all add to the difficulty and make it highly problematic to achieve robust results.

It is necessary to define as sharply as possible the port activities under study, which requires arriving at a balance between the number of necessary variables and the number that the sample and the models are capable of supporting.

It is possible to continue advancing in empirical research to reach a degree of consensus in relation to the determinants of efficiency and even develop issues that the literature has not addressed sufficiently, such as questions about quality, user time (berthing time, waiting time and delay time), emissions of greenhouse gases, climate change, etc.

In terms of effectiveness, more evidence about this indicator is also welcome. In addition, effectiveness can be considered a complement to efficiency, so that taking into account the objectives of users would greatly help improve efficiency. It seems clear that if the port is not effective, it can hardly be efficient.

References

- Baird, A.J., 1995. Privatization of trust ports in the United Kingdom: review and analysis of the first sales. *Trans. Policy* 2 (2), 135–143.
- Brooks, M.R., Pallis, A.A., 2008. Assessing port governance models: process and performance components. *Marit. Pol. Manag.* 35 (4), 411–432, doi:10.1080/03088830802215060.
- Brooks, M.R., Schellinck, T., Pallis, A.A., 2011. A systematic approach for evaluating port effectiveness. *Marit. Pol. Manag.* 38 (3), 315–334, doi:10.1080/03088839.2011.572702.
- Charnes, A., Cooper, W., Lewin, A.Y., Seiford, L.M., 1997. Data envelopment analysis theory, methodology and applications. *J. Operat. Res. Soc.* 48 (3), 332–333.
- Cullinane, K.P.B., 2013. Revisiting the productivity and efficiency of ports and terminals: methods and applications. In: Grammenos, C. (Ed.), *The Handbook of Maritime Economics and Business*. Informa Law from Routledge, London, pp. 937–976.
- Farrell, M.J., 1957. The measurement of productive efficiency. *J. R. Stat. Soc. Vol.* 120, 253–290.
- Fried, H.O., Lovell, C.A.K., Schmidt, S.S., 1993. *The Measurement of Productive Efficiency: Techniques and Applications*. Oxford University Press, Oxford.
- González, M.M., Trujillo, L., 2009. Efficiency measurement in the port industry: a survey of the empirical evidence. *JTEP* 43 (2), 157–192.
- Ha, M.H., Yang, Z., Notteboom, T., Ng, A.K.Y., Heo, M.W., 2017. Revisiting port performance measurement: a hybrid multi-stakeholder framework for the modelling of port performance indicators. *Transp. Res.* 103, 1–16.

Kumbhakar, S.C., Lovell, C.A.K., 2003. *Stochastic Frontier Analysis*. Cambridge University Press, Cambridge.

Lovell, C.A.K., 1993. Production frontier and productive efficiency. In: Fried, H., Lovell, C.A.K., Schmidt, S.S. (Eds.), *The Measurement of Productive Efficiency: Techniques and Applications*. Oxford University Press, Oxford.

Further Reading

Banker, R.D., Charnes, A., Cooper, W.W., Swarts, J., Thomas, D., 1989. An introduction to data envelopment analysis with some of its models and their uses. *Res. Gov. Nonprofit Account.* 5 (5), 125–163.

Brooks, M.R., Pallis, A.A., 2013. Advances in port performance and strategy. *Res. Transport. Bus. Manag.* 8, 1–6.

Brooks, M.R., Schellinck, T., 2013. Measuring port effectiveness in user service delivery: what really determines users' evaluations of port service delivery? *Res. Transport. Bus. Manag.* 8, 87–96.

Coelli, T., Estache, A., Perelman, S., Trujillo, L., 2003. *A Primer on Efficiency Measurement for Utilities and Transport Regulators*. World Bank Institute Development Studies, Washington, DC.

Coelli, T.J., Rao, D.S.P., O'Donnell, C.J., Battese, G.E., 2005. *An Introduction to Efficiency and Productivity Analysis*. Springer Science & Business Media, New York.

Cooper, W.W., Seiford, L.M., Tone, K., 2000. Data envelopment analysis: history, models, and interpretations. In: Cooper, W.W., Seiford, L.M., Zhu, J. (Eds.), *Handbook on Data Envelopment Analysis*. Springer, New York, pp. 1–39.

Färe, R., Färe, R., Grosskopf, S., Lovell, C.K., 1994. *Production Frontiers*. Cambridge University Press, Cambridge.

Panayides, P.M., Maxoulis, C.N., Wang, T.F., Ng, A.K.Y., 2009. A critical analysis of DEA applications to seaport economic efficiency measurement. *Trans. Rev.* 29 (2), 183–206.

Seiford, L.M., Thrall, R.M., 1990. Recent developments in DEA: the mathematical programming approach to frontier analysis. *J. Econom.* 46 (1–2), 7–38.

Container Port Automation

Michael G.H. Bell, ITLS, Business School, University of Sydney, Sydney, NSW, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	323
Containerization	323
Container Terminal Automation Technologies	323
Container Terminal Automation Challenges	324
Slow Pace of Container Terminal Automation	324
Green Ports and Automation	325
Future for Container Terminal Automation	325
Biography	326
References	326

Introduction

Containerization transformed the way general cargo is transported in the 1970s and facilitated the globalization of supply chains thereafter. The standardization of container dimensions and twist locks for lifting enabled the mechanized transfer of containers between ships, trucks and trains and greatly reduced the workforce in ports from large numbers of dockers working in gangs to a much smaller number of drivers of cranes and vehicles. It was then a small step to the automation of container handling, replacing some drivers by control room operators remotely controlling equipment or managing exceptions for automated equipment. Experience suggests that automation approximately halves the workforce, but the main gains are in terms of reliability and safety.

To date, the investment required to automate container terminals is huge and the returns on investment have been modest. As automated equipment becomes more “off the shelf”, its cost will fall and the equipment control software will improve in reliability and efficiency. As a consequence, container terminal automation is expected to continue its slow spread, not just to more ports, but also to cargo handling upstream and downstream of the port, and in documentary transfer which finances and facilitates international trade. This chapter looks at container terminal automation in detail, describes the pros and cons of various forms of automation, and anticipates future developments for the container port of the future.

Containerization

The mechanized handling of general cargo that containerization facilitated also opened the possibility of automating the loading, unloading and stacking of containers. The containerization of general cargo revolutionized breakbulk shipping, bringing with it the consolidation of those shipping lines carrying containers and massively reducing the labor requirement of those ports handling containers. The result was a period of industrial unrest in the 1970s among those dock workers no longer required for loading and unloading ships, who were replaced by very much fewer drivers of cranes and vehicles. The sheds used to protect loose cargo from the elements were replaced by yards to store containers. The larger space requirement led to some container terminals relocating to sites where more land was available. From the seafarer perspective, life on board container carriers also lost some of its appeal as the ships spent hours rather than days in port, but the supply of seafarers from developing countries like the Philippines has so far sustained container shipping (see Phillips, 2014, for a short and readable account of the impact of containerization).

Container Terminal Automation Technologies

The potential to automate the handling of containers in ports was evident from early on, starting with stacking cranes in the yard. Initially attention was focused on rail-mounted gantry cranes, referred to as RMGs, because of the positional accuracy offered by rails. The Delta Terminal of Europe Container Terminals (ECT) in Rotterdam is believed to be the first terminal where Automated Stacking Cranes (ASCs) entered operation in 1993 (PEMA IP03, 2012). Despite a slow start, ASCs have gradually spread, to the extent that by 2016 there were over 1100 in operation around the world. The technology is now standard and as a consequence the cost has fallen and reliability improved (PEMA IP12, 2016).

There are two types of ASC in common use. In Europe, the end loading/unloading ASC working a block of containers 6–10 containers wide and between 28 and 48 TEU (20 foot equivalent units) long positioned perpendicular to the quay appears to be preferred (Kemme, 2012). There would generally be two ASCs per block, with one serving the quay and the other serving the landside. This layout works well when the proportion of transhipped containers is low. It has the advantage of separating the internal

vehicles serving the quay cranes from the external vehicles collecting and delivering containers on the landside, with safety benefits for drivers. In Asia, however, a block layout parallel to the quay worked by side loading ASCs is preferred. The block may be wider (10–12 containers). A cantilever on the ASC allows containers to be loaded or unloaded from a trailer alongside the crane. This layout is less sensitive to the proportion of transhipped containers. When there are cantilevers either side of the ASC the internal vehicles can be separated from the external vehicles, with safety benefits for vehicle drivers. The side loading ASCs are heavier than the end loading ASCs, but travel less, saving energy and track wear and tear. Although in both layouts containers are generally stacked 5 high, the yard density achieved with end loading ASCs is higher because of the absence of vehicle lanes between the blocks (PEMA IP03, 2012).

Automation of the horizontal transport between the quay and the blocks in the yard has not reached the same level of maturity. Initially Automated Guided Vehicles (AGVs) were deployed by ECT in Rotterdam in 1993 (PEMA IP12, 2016). After resolving initial problems encountered in Rotterdam, AGVs were deployed in HHLA's CTA terminal in Hamburg in 2002. However, AGVs cannot load or unload themselves, so must be located beneath the ASC or quay crane during loading or unloading. Significant gains in vehicle productivity are possible when crane and vehicle activity are decoupled. One possibility used in a number of terminals is to decouple the ASC and AGV by use of a rack. The ASC unloads an export container on to a rack, a lift-AGV slides under the container, lifts it off the rack from below, and transports it to the quay crane. The process can also be reversed for an import container, however racks cannot be used in conjunction with quay cranes. Lift-AGVs have been deployed at APMT's Maasvlakte II terminal, Rotterdam, in 2015 (PEMA IP12, 2016).

Another possibility is to use shuttle carriers, reduced format 1 over 1 straddle carriers, which can pick up a container left on the ground by an ASC or a quay crane. Shuttle carriers at present have drivers. An operational disadvantage of the shuttle carrier over the lift-AGV is that the former requires a wider lane. On the other hand, racks cannot be easily used in combination with quay cranes, because ships can change position.

New forms of automated horizontal transfer between the quay crane and yard storage block are being developed. After the implementation of Autostrads (driverless straddle carriers developed by Kalmar) in Patrick's Brisbane Terminal 2002, this technology has matured, and is in operation in the Patrick Terminal, Port Botany, Australia, since 2015. Having successfully automated the straddle carrier, it is a small step to the automation of the shuttle carrier. Since 2017, automated shuttle carriers have been in operation in Melbourne's VICT facility.

Elsewhere in the terminal, remote-controlled RMGs with rotating 'spreaders', the device that locks on to the container, have been used to load and unload trains. In APMT's Maasvlakte II terminal, lift-AGV have been used to transport containers to and from the railhead (PEMA IP12, 2016). The automation of quay cranes has only been partial so far, because of ship movement and the safe setting or removal of twistlocks on the ship. However, HHLA's CTA terminal in Hamburg pioneered an automated second trolley, which moves containers between an intermediate lashing platform on the quay crane and AGVs. This facilitates parallel working, since containers can be lifted from a ship to the lashing platform at the same time as containers are lifted from the lashing platform to an AGV. The quay cranes in APMT's new Maasvlakte II terminal are equipped with spreaders to lift containers to or from the ship up to four 20 foot containers at a time and spreaders capable of loading or unloading lift-AGVs at up to two 20 foot containers at a time.

Remote-control is revolutionizing the work of the quay crane driver. Rather than sitting in a cab high up in the crane and peering down between his/her legs at the spreader below, the driver can sit or stand in the comfort of the control room, operate one or more cranes simultaneously, switch to a colleague for a break, and consult when a problem arises.

Container Terminal Automation Challenges

Automation and remote crane control approximately halves the workforce. In 2014, the Patrick Terminal of Port Botany, Australia, employed 436, however after automation in 2016, when conventional straddle carriers were replaced by driverless Austostrads, the workforce fell to 213 (Hellenic Shipping News, 2019). Safer well-paid jobs remain in the control room, providing a basis for a negotiated settlement with labor unions. A study by Prism Economics and Analysis, commissioned by the International Longshore and Warehouse Union (ILWU) Canada and quoted in Hellenic Shipping News (2019), found that the automation and digitization of container terminals in British Columbia could lead to significant job losses, reduce tax revenues and impact the local economy. This was a major issue in contract negotiations between the ILWU and terminal operators, which however were resolved after a short strike.

Another major challenge in automation has been linking the Equipment Control Software (ECS), which is generally provided by the equipment suppliers, with the Terminal Operating System (TOS), which keeps track of container location and status and increasingly initiates jobs. The TOS is transitioning from a database to a workflow optimizer and an initiator of jobs for the ECS to execute. The TOS increasingly optimizes slot allocation and storage/retrieval schedules.

Slow Pace of Container Terminal Automation

Since the first example of port automation in 1993, the spread of the technology has been slow. Chu et al. (2018) estimate that by 2018 there were almost 40 fully or partially automated container ports around the world, with at least US\$10 billion invested in such projects. After the removal of initial niggles and software bugs, container terminals have proven to be safer and more reliable after

automation. However, [Chu et al. \(2018\)](#) reported that the return on invested capital and productivity has disappointed industry managers. This contrasts with the mining sector, which also involves heavy equipment working outdoors, where early adopters of automation have experienced productivity improvements of 20%–40%. Nonetheless, the survey of port terminal managers conducted by [Chu et al. \(2018\)](#) anticipates more automation in the future.

The automation of container terminals can be characterized by four stages ([Chu et al., 2018](#)):

1. Port 1.0: “Management by hero”, where managers make decisions supported by a TOS, which keeps track of containers in the yard.
2. Port 2.0: “Management by process”, where the TOS allocates jobs to cranes and vehicles, and managers supervise the process.
3. Port 3.0: “Management by exception”, where the ECS interacts with the TOS to generate, schedule, allocate and execute jobs carried out by automated cranes and vehicles, leaving the managers to deal with exceptions where automation cannot cope.
4. Port 4.0: “From manage to orchestrate”, where documentary exchanges are performed electronically and automatically, and where automation is extended beyond the port upstream and downstream in the supply chain.

While Port 3.0 is now technologically well-established and spreading slowly, Port 4.0 is now emerging.

Green Ports and Automation

Steps are currently being taken to reduce the environmental footprint of ports. In general, this involves the electrification of cranes, *cold ironing* (switching to landside electricity supply when at berth), and the use of hybrid straddle carriers. Electric AGVs are coming into use, for example in APMT’s Maasvlakte II terminal, but this involves a complicated automated battery swapping system. Battery powered straddle carriers pose a challenge because of the distance traveled and the lifting. However, battery powered shuttle carriers look promising (see [Leskinen and Hägglom, 2019](#), for a full discussion on the electrification of port equipment).

The “greenness” of electrification depends on the extent to which electricity is generated from renewable energy sources, in particular wind turbines and solar panels. Ports frequently offer large roof areas suitable for solar panels and piers or breakwaters suitable for wind turbines, so the load on the electricity grid can be reduced. The use of regenerative technology to reclaim electricity when lowering containers or when vehicles brake can also reduce the load on the electricity grid, but will require intermediate storage of electricity in batteries. Fuel cells are a promising technology for the future, but a supply chain for hydrogen suitable for vehicles in ports is not currently available.

Automation and electrification are synergetic. In combination they offer possibilities for optimizing the use of electricity. Machine acceleration and deceleration, top and cornering speed as well as loads can be adjusted to improve the overall efficiency of the electric powertrain. Automation and digitalization enables the scheduling of charge cycles as part of the routing of vehicles, and the entire operation can be optimized incrementally through machine learning ([Leskinen and Hägglom, 2019](#)).

Future for Container Terminal Automation

International trade relies on the exchange of documents between the buyer and seller, or their respective banks. The seller would like to be paid when shipping the cargo, while the buyer would not like to pay until the cargo (or more specifically, the entitlement to the cargo) is received. The classic and most reliable way to fund international trade is the *Letter of Credit* (L/C). The buyer goes to his/her bank, orders a L/C, and sends this to the seller’s bank. The L/C lists documents the buyer requires before paying. Key among these is the *Bill of Lading* (B/L), which is both the receipt issued by the shipping line to the exporter when loading the cargo and the title document that enables the importer to take possession of the cargo upon arrival. When the buyer’s bank receives the L/C, B/L and other documents (like certificates of origin, warranties, inspection certificates, etc), the buyer’s bank pays the seller’s bank.

There are simpler, and therefore cheaper, forms of documentary exchange, but they are less secure. For example, under the *sight draft* procedure, the buyer’s bank holds on to the B/L and other documents received from the seller (or seller’s bank) until payment has been made, but there is a risk that the buyer decides not to purchase the cargo after the seller has shipped it, leaving the seller with unsold cargo “at sea”.

Increasingly, documents are exchanged electronically. Consequently, distributed ledger and encryption technologies, in particular blockchain technology, is receiving attention as a way to ensure secure electronic document exchange and to automate document checking and payment ([DHL, 2018](#)). To date a number of platforms have been developed, but their use is spreading slowly. The best known is TradeLens, developed jointly by IBM and Maersk, but according to a news report ([Wee, 2018](#)) take-up among other shipping lines has been limited, perhaps due to a lack of trust in the impartiality of the platform given the involvement of Maersk.

An interesting example of the extension of automation beyond the port, also a feature of Port 4.0, is the use of hybrid AutoStrads (driverless straddle carriers developed by Kalmar) to connect the rail terminal with adjacent warehousing in Sydney’s new Moorebank intermodal terminal ([Kalmar, 2019](#)). When the facility is finished, a significant number of containers offloaded at Port Botany will be transported to Moorebank intermodal terminal by train and then transferred automatically to warehousing in the adjacent logistics park. The Port 4.0 vision will be automation “from ship to shop”, to include the documentary exchanges

involved. This vision will then match Australia's iron ore mining and exporting industry, where automation is being extended "from pit to port" (Issa, 2019).

Biography



Michael Bell is the professor of Ports and Maritime Logistics in the Institute of Transport and Logistics, at the University of Sydney Business School. Prior to this, he was for 10 years the Professor of Transport Operations at Imperial College London and for the final 5 years at Imperial the Founding Director of the Port Operations Research and Technology Centre (PORTeC). He graduated from Cambridge University with a BA in Economics and obtained an MSc in Transportation and a PhD on Freight Distribution from Leeds University. His research and teaching interests span ports and maritime logistics, transport network modeling, traffic engineering, and intelligent transport systems. He is the author of many papers, a number of books (including *Transportation Network Analysis*, published in 1997), was for 17 years an Associate Editor and now an Editorial Board Editor of *Transportation Research B*, the leading transport theory journal, was an Associate Editor of *Maritime Policy & Management* and is currently an Associate Editor of *Transportmetrica A*.

References

- Chu, F., Gailus, S., Liu, L., Ni, L., 2018. The Future of Automated Ports. McKinsey & Company, New York.
- DHL, 2018. Blockchain in Logistics <https://www.logistics.dhl/content/dam/dhl/global/core/documents/pdf/glo-core-blockchain-trend-report.pdf>, accessed 22/1/2020.
- Hellenic Shipping News, 2019. The Fallout From Container Port Automation <https://www.hellenicshippingnews.com/the-fallout-from-container-port-automation/>, published 16/09/2019, Accessed 07/01/2020.
- Issa, J., 2019. Automating from Pit to Port: Australia Aspires to Define 'Mining 4.0' <https://www.gbreports.com/article/automating-from-pit-to-port-australia-aspires-to-define-mining-40>, Accessed 23/1/20.
- Kalmar, 2019. OneTerminal powers Australia's largest logistics development at Moorebank Logistics Park <https://www.kalmarglobal.com/news-insights/2019/kalmar-oneterminal-powers-australias-largest-logistics-development-at-the-moorebank-logistics-park-mlp/>, accessed 22/1/2020.
- Kemme, N., 2012. Effects of storage block layout and automated yard crane systems on the performance of seaport container terminals. *OR Spektrum* 34, 563–591.
- Leskinen, J., Häggblom, H., 2019. Energy management and battery powered horizontal transportation at container terminals https://www.kalmar.com.au/491467/globalassets/media/216119/216119_FastCharge-WP-2019-WEB.pdf, Accessed 24/1/20.
- PEMA IP03, 2012. Container Terminal Automation: A PEMA Information Paper <http://www.pema.org/publications>, Accessed 08/01/2020.
- PEMA IP12, 2016. Container Terminal Automation: A PEMA Information Paper <http://www.pema.org/publications>, Accessed 08/01/2020.
- Phillips, K., 2014. How container shipping changed the world through globalisation <https://www.abc.net.au/radionational/programs/rearvision/the-big-metal-box-that-changed-the-world/5684586>, Accessed 23/1/20.
- Wee, V., 2018. Maersk, IBM launch TradeLens blockchain platform with 94 signed up <https://www.seatrade-maritime.com/asia/maersk-ibm-launch-tradelens-blockchain-platform-94-signed>, Accessed 22/1/2020.

Optimizing Crane Operations in Ports

Frank Meisel, Christian Bierwirth, School of Economics and Business, Kiel University, Kiel, Germany; School of Economics and Business, Martin-Luther-University Halle-Wittenberg, Halle, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	327
Crane Technology	327
Quay Cranes	327
Gantry Cranes	328
Optimizing Quay Crane Operations	330
Quay Crane Assignment	330
Quay Crane Scheduling	330
Optimizing Gantry Crane Operations	331
Yard Crane Deployment in Container Yards	331
Gantry Crane Scheduling in Container Yards	332
Gantry Crane Scheduling at the Hinterland Interface	332
Research Trends	333
Biographies	333
See Also	334
References	334
Further Reading	334

Introduction

The container shipping industry has grown tremendously since its emergence in the 1950s. Today, the largest container ports in the world handle millions of containers per year. The largest ocean-going container ships can carry more than 20,000 container units each. To handle the huge flows of containerized cargo, port facilities have become billion dollar investments that provide the infrastructure for quickly serving even the largest container ships, for intermediate storage of thousands of containers, and for handling dozens of trains and thousands of container-carrying trucks every day. The typical workflow of a container terminal is illustrated in Fig. 1. At the seaside, ships moor at berths for being served. The unloaded import containers are transported to the yard area where they stay (typically for a couple of days) until they are further transported into the hinterland by trucks, trains, or barges. Hinterland traffic based on road and rail transportation requires specialized handling facilities configuring a hinterland interface, while inland waterway transportation typically uses the facilities given at the seaside area. Generally, containers carrying export goods take the opposite way through the terminal, that is, they arrive at the port from the hinterland by truck, train, or barge, stay in the yard, and are then loaded onto their designated container ship.

Terminals require specialized, powerful crane equipment at the seaside area, in the yard, and at the hinterland interface in order to stack containers and lift them onto transport vehicles. A highly productive usage of this equipment calls for an optimization of crane operations in container terminals. The corresponding optimization problems have been investigated in great detail in recent years. With this article, we provide a fundamental description of the crane technology used in container ports and the operational optimization problems that occur when using this equipment.

This article is organized as follows. In Section “Crane Technology” we provide technical details of the most relevant types of cranes, namely, quay cranes (QCs) that are used at the seaside and gantry cranes that are used in the yard and at the hinterland interface. A discussion of the fundamental optimization problems in the operations management of QCs and gantry cranes is provided in Sections “Optimizing Quay Crane Operations” and “Optimizing Gantry Crane Operations.” Finally, we briefly outline topics for future research in Section “Research Trends.”

Crane Technology

Quay Cranes

QCs are deployed at the seaside of a container terminal for retrieving containers from ships and for loading containers onto ships (Fig. 2A). The largest container ships in service nowadays can carry more than 20 rows of containers next to each other in a bay and up to 24 containers on top of each other (e.g., ships from the Triple-E class). Hence, for serving such ships, QCs have to have an outreach of more than 60 meters and a lifting height of 40 meters. Unloading a container from a ship requires placing the crane's spreader on the container where it is fixed using twistlocks before being lifted by a hoist. A trolley then moves the container to the

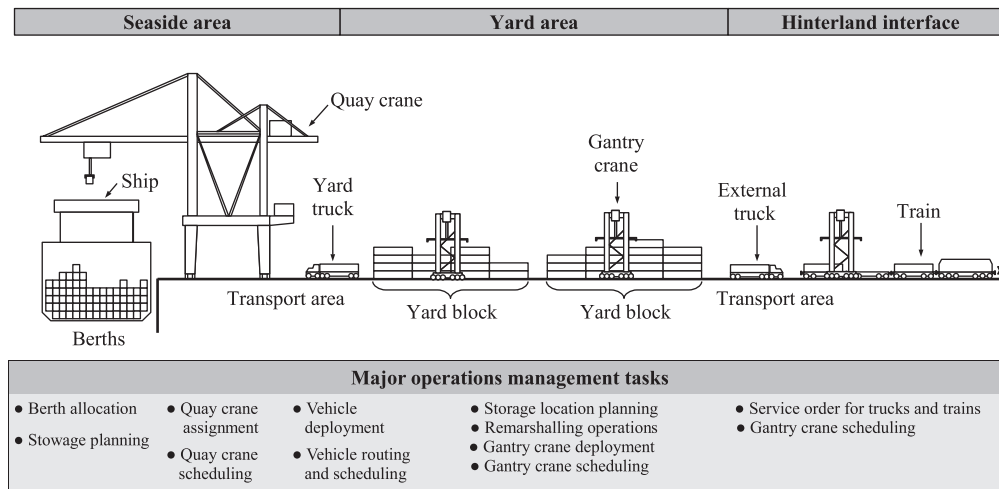


Figure 1 A schematic view of a container terminal.

shore where it is lowered to the ground or onto a transport vehicle and released by unlocking the twistlocks and lifting the spreader. The loading of a container is performed using the inverse procedure, starting at the quay and ending at the ship. One such transfer operation is referred to as a “move.” The number of moves per hour is the major productivity measure for QCs. Typical productivity rates in practice are 25–35 moves per hour. Up to now, QCs are still manually operated by crane drivers, but numerous technical innovations may lead to (partly) automated crane operations and a general increase of crane productivity in the future. Advanced technologies already existing are twin-spreaders for handling two container units in one move, dual trolley QCs where one trolley serves the ship and hands over containers to a landside trolley putting them onto transport vehicles, and remote control, which enables one human operator to operate multiple QCs in parallel. A general difficulty observed by crane operations control is that QCs are mounted on rails for being moved alongside the quay. This has the effect that QCs cannot overtake each other, but have to keep their relative position at the quay at all times. Furthermore, due to their geometry, a safety distance has to be kept between cranes, which means that no two cranes can operate at adjacent ship bays at the same time. Needless to say, technical crane operation rules are able to considerably reduce crane productivity. For this reason, research on optimization in crane scheduling puts a strong focus on these constraints.

Gantry Cranes

Gantry cranes are used for stacking and retrieval operations in the container yard and for unloading and loading containers from/to trains at the hinterland interface of a terminal (Fig. 2B and C). Gantry cranes used in the container yard of a terminal are also called yard cranes (YCs). The container yard is partitioned into yard blocks, where each block consists of a number of container rows with lengthwise-arranged storage positions (the so-called bays) (Fig. 2B). A gantry crane spans such a block, which means that the number of rows in a block is determined by the width of a crane. Furthermore, adjacent blocks need to be separated by traffic lanes. Gantry cranes can be rubber-tired (RTGCs) or rail-mounted (RMGCs). RTGCs can span up to eight rows of containers. RMGCs are somewhat larger and can span 10 or even more rows. Although RMGCs support a higher overall storage density for the yard, the advantage of RTGCs is that they are mobile and can be moved relatively freely among different blocks. If automated, the gantry crane is called an automated stacking crane (ASC). Gantry cranes can stack containers up to six tiers high where the topmost tier typically remains empty for moving lifted containers. In a train area, gantry cranes span multiple rail tracks, which allow them to serve a number of trains in parallel. Gantry cranes use the same technology (spreader, twistlocks, hoist, trolley, etc.) and similar work processes for moving containers as previously described for QCs. However, while QCs reposition containers only along one axis (ship-to-shore), gantry cranes reposition containers along two axes. They can move a container from one row to another using the trolley, while the crane itself remains at its position. They can also move containers lengthwise along the block by letting the crane travel with a lifted container. This allows gantry cranes to reposition containers among any stacks in a container block. The horizontal transport vehicles carrying containers to/from the yard block can be served either at the two ends of the block, or are served by the YC beneath a bay, provided an outer row of the container block is left empty to serve as a parking lot. The former layout allows the stacking of more containers in a bay, but requires additional movement of YCs, while the latter option has a lower storage capacity but faster crane operations.

The number of container moves per hour is also a relevant productivity measure for gantry cranes. However, this measure can strongly vary for gantry cranes, as the durations of individual container moves depend on the distance between the particular retrieval and stacking positions in the yard block. Several technological innovations have been developed to make gantry crane systems more productive. Different to QCs, gantry cranes can already be run on an automated or semiautomated basis today. In combination with remote-control automation, this enables large container yards to be run by only a few human operators. Another approach often found

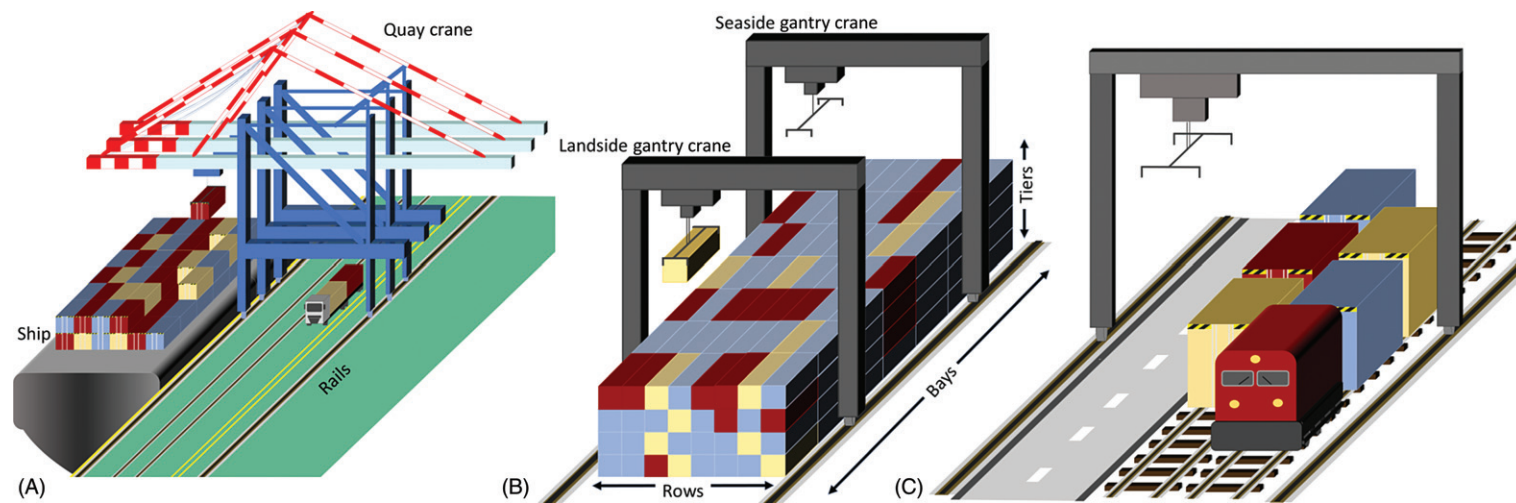


Figure 2 Quay cranes and gantry cranes. (A) Quay cranes at the seaside. (B) Twin gantry cranes in container yard. (C) Gantry crane at train interface.

in practice is to deploy two gantry cranes in a yard block instead of just one, in order to increase the number of containers handled in the block per hour. If the two cranes are of the same size (the so-called twin cranes), each of them operates at one end of the yard block, also called landside and seaside cranes (Fig. 2B). Obviously, twin cranes cannot pass each other and have to keep a safety distance at all times. In such a system, if a container has to be repositioned from one end of the block to the other, either the noninvolved crane has to move away (if possible), or the container is placed at an exchange point from which the other crane picks it up. If the gantry cranes are of different size (the so-called dual cranes), the smaller crane can pass beneath the larger crane. In this kind of system, both cranes can reach every position in the yard block, but they are not allowed to operate at the same yard bay at the same time.

It should be noted that a terminal might also employ the so-called straddle carriers that can lift, drop, and carry containers to any position in a terminal. Hence, straddle carriers can be used for horizontal transportation of containers as well as for stacking and retrieving operations in the yard and at the hinterland interface. Particularly for smaller terminals, straddle carriers are an interesting alternative to circumvent major investments in gantry technology, which in general is profitable only if highly utilized.

Optimizing Quay Crane Operations

The QC resource of a container terminal is the subject of several strategic, tactical, and operational decisions. For example, when building a new terminal or extending an existing terminal, an equipment selection problem is faced regarding the type and number of QCs that are to be procured (Meisel and Bierwirth, 2011a). Decisions on equipment are typically made in response to changing market conditions (e.g., growing vessel categories), sometimes even in accordance with the demands of shipping companies. They enable a terminal to gain competitive advantage against other terminals in the port or in the region. An important tactical problem is how to dimension the workforce of the terminal to meet the workforce demand in the daily operations. A crucial operational question is how to utilize the available cranes for highly productive transshipment processes and a fast service of ships. This challenge is related to two operational optimization problems, the quay crane assignment problem (QCAP) and the quay crane scheduling problem (QCSP), which are described here in more detail.

Quay Crane Assignment

The QCAP is to assign a number of QCs to each ship that is served at the terminal. Next to the number of cranes, the QCAP also has to decide on the specific cranes to be assigned to a ship. Smaller ships are typically served by one to three cranes, while larger ships can be served by up to six cranes simultaneously. Since QCs are a scarce resource of container terminals, the QCAP is especially relevant if multiple ships are served simultaneously. Clearly, the total number of cranes available has to be respected at all times in the QCAP. A further constraint can be that a ship operator and the terminal operator contracted a least number of cranes that must be assigned to a ship. The number of cranes assigned to a ship might be constant for the whole service process (the so-called time-invariant QC-to-vessel assignment) or it might change in the course of time (the so-called variable-in-time QC-to-vessel assignment). While time-invariant assignments are easier to plan and implement at a terminal, variable-in-time assignments give more flexibility in utilizing the QC resource. Eventually, all ships must receive sufficient QC hours for loading and unloading based on the total number of container moves to be executed. Additionally, it needs to be considered that QCs have to keep safety distances and cannot pass each other because they are rail-mounted. These restrictions account for the fact that cranes can interfere with each other when moving to dedicated bays. Consequently, assigning more and more cranes to a ship increases the probability of crane interference and reduces individual crane productivity. As the number of cranes assigned to a ship determines how fast a ship is served, the QCAP has significant impact on the departure times of ships and, thus, on the service quality achieved by the terminal. Therefore, crane assignment is a crucial decision, especially if a large number of ships are served in parallel. A typical objective in crane assignment is, therefore, to minimize tardiness in ship departures.

Assigning not just a number, but specific cranes, to a ship is of relevance, if not all available cranes are in use. Consider, for example, a terminal with eight QCs and two small ships to be served with three QCs each. The specific cranes to use might be 1–3/4–6, 1–3/5–7, 1–3/6–8, 2–4/5–7, 2–4/6–8, or 3–5/6–8 as the rail-mounted QCs have to keep their relative position at the quay. In such a situation, the cranes to use are selected such that the movement of cranes to the ship positions or the setup operations of cranes are minimized.

In the scientific literature, the QCAP is rarely investigated as a problem on its own. Due to the interdependency with berth allocation and quay crane scheduling, the QCAP is typically combined with one of these problems by merging the corresponding decisions, constraints, and objectives into an integrated optimization problem. In particular, the integrated berth allocation and crane assignment problem has been studied intensively. Especially for terminals with continuous quays (one long, straight quay wall) decisions about where to berth ships and which cranes to use are highly interrelated. An explicit QCAP is not relevant for terminals where quays are viewed as discrete resources. Here, every ship is assigned exclusively to one of a number of available berth segments, each equipped with a fixed set of QCs.

Quay Crane Scheduling

The QCSP is to determine starting times and the respective cranes for all loading and unloading operations of containers belonging to a certain ship. The predominant objective of this optimization problem is to minimize the latest completion time among all crane

tasks, also called the makespan of the schedule. This objective coincides with the minimization of a ship's handling time and, as such, strives for short port stay times and on-time departures of ships. Further objectives to be considered in the QCSP are, among other things, the minimization of the total finishing time of cranes and the minimization of unproductive crane movements. Such goals mainly aim at reducing crane operation cost, which, in practice, receives attention only in the case that an on-time departure is assured for a ship.

The basic constraints of the QCSP are quite similar to constraints met in general scheduling problems: every loading/unloading task must be assigned to exactly one QC, each QC can process only one task at a time, and task processing is non-preemptive. More specific constraints of the QCSP concern, for instance, the initial positions and ready times of the QCs. Even time windows might be given that restrict the period at which a crane is available for serving a ship. Ready times and time windows typically follow from the QCAP, which is solved prior to the QCSP.

The fact that QCs are rail-mounted leads to a set of mandatory interference constraints. Most fundamental is the non-crossing constraint, which ensures that cranes keep their relative positioning at the quay at all times in the crane schedule. Some QCSP models also consider a safety requirement, saying that a minimum distance (typically measured as a number of ship bays) is kept between any adjacent QCs at all times in the crane schedule. Non-crossing is often ensured by moving all cranes in parallel into the same direction along the ship (from bow to stern, or from stern to bow) throughout the service process. This leads to the so-called unidirectional crane schedules. More advanced models also care for the stability of the ship during the service process, which might be endangered in a unidirectional loading/unloading process.

The containers to load and unload for a ship are given in the ship's stowage plans. From this, a QCSP can be derived which determines a schedule for each single container move to be executed (Meisel and Wichmann, 2010). However, this becomes computationally intractable if a large number of containers (probably several thousands) have to be handled for a ship. To reduce complexity, practicable QCSP models typically cluster the huge set of containers into the so-called transshipment tasks, which aggregate subsets of operations to be scheduled. Transshipment tasks can refer to: stacks of containers (the so-called stack-level); several groups of containers in the same bay (group-level); all containers to load and unload at a certain bay (bay-level); or a subset of the bays (area-level). The most prevalent aggregations found in the literature are the group-level and the bay-level. These aggregations significantly affect smaller task sets for the QCSP without losing the flexibility to balance the total workload among the available cranes effectively. In the group-level aggregation, there are typically precedence constraints involved to ensure that groups of import containers are unloaded at a bay before groups of export containers are loaded (Kim and Park, 2004; Bierwirth and Meisel, 2009). The more detailed stack-level and container-level aggregations are used to model advanced loading techniques like crane double cycling or container reshuffling, where the scope is restricted to a single ship bay. Such models have to be assembled and solved for each bay of a ship individually. Then the detailed schedules obtained are inserted into the overall crane schedule that is built for the entire ship service process.

A further distinction of QCSP models is whether a crane schedule is determined for a single ship or for a set of ships served in parallel at the terminal. QCSP models of the latter type support the possibility to transfer cranes from one ship to another during service, which implicitly solves a QCAP. Here, the moving speed of QCs becomes a crucial parameter. Multi-ship models usually strive for a minimization of the total weighted completion time of ships. Other advanced models integrate quay crane scheduling with stowage planning or with horizontal transport management. Stowage planning is closely related with crane scheduling as it makes decisions on where to locate containers onboard a ship. Implementing the stowage plans for the bays obviously impacts the workload of the QCs as given in the QCSP. Horizontal transport management is also connected with the QCSP, as unloaded import containers have to be transported to the container yard and export containers have to be carried from the yard to the quay. To enable a high productivity of QCs, transport management is responsible for providing empty vehicles as soon as containers are unloaded and for delivering the right export containers from the yard such that they arrive at the quay on time.

Optimizing Gantry Crane Operations

Similar to QCs also, the gantry cranes are subjects of strategic and tactical management. Crucial decisions concern, for instance, the question of equipment selection, the intended degree of automation, and the yard layout design problem. Making these decisions is difficult for several reasons. The decisions to be made are highly interdependent and they are linked with high acquisition costs, and thus constitute a long-term basis for the operation of the container yard and the traffic facilities that build the hinterland interface. To utilize these infrastructures and equipment in the most profitable way is the subject of operations management. The operational problems met in the yard and the hinterland interface include crane deployment problems and scheduling problems for container storage and retrieval operations, which are described in the following subsections.

Yard Crane Deployment in Container Yards

The yard crane deployment problem (YCDP) is to decide which crane to work at which yard block in the course of time. The problem is most relevant for terminals where the number of yard blocks exceeds the number of available cranes. Such terminals often deploy RTGCs that can be moved flexibly from one-yard block to another. In contrast, RMGCs are bound by the rail infrastructure to one

particular block or to a line of blocks. In the latter case, crane deployment decisions can be relevant to some extent. The projected workload at each yard block serves as an input for crane deployment decisions. Assigning cranes to yard blocks typically has the objective of completing workload as fast as possible, or to shift as little uncompleted workload as possible into the next planning period. The planning horizon is usually set to a few hours. Therefore, the YCDP delivers a (dynamic) assignment of cranes to blocks, which is similar to the assignment of QCs to ships in the QCAP. A typical constraint in the YCDP is to assign at most one or two cranes to a container block at the same time in order to avoid increasing crane interference effects, which in turn reduce productivity. Furthermore, as the YCs available at a terminal can be heterogeneous, particular cranes might be unsuitable for operating in certain blocks. Like the QCAP, the YCDP is rarely investigated in the scientific literature as a problem on its own, but is often integrated with YC scheduling due to the strong interdependencies of both problems.

Gantry Crane Scheduling in Container Yards

The major task of operations management in the yard is to schedule stacking and retrieval operations of containers by YCs, which is referred to as the yard crane scheduling problem (YCSP). In this problem, a set of container operations (retrievals, stackings, repositionings) is given for a yard block, for which a service order needs to be determined together with a start time for each single operation. If multiple YCs are deployed to a block, it also needs to be decided which crane is to perform which operation. In contrast to the QCSP, there is no task aggregation applied in the YCSP, meaning that each container is treated as an individual operation to be scheduled. Relevant objectives are the minimization of a schedule's makespan and the minimization of earliness or lateness in completing operations (Wu et al., 2015). In the latter case, crane operations to be planned have a given due date which coordinates the YCs with the service processes at the seaside and the hinterland interface. Another objective of the YCSP aims at minimizing the effort of reshuffling containers. Reshuffling is necessary when stacked containers need to be removed in order to access a target container that is to be retrieved.

Standard constraints in a YCSP are to ensure that each operation is executed by a YC, where no YC can perform more than one container operation at a time. Scheduling is preemptive just like in the QCSP. Furthermore, precedence relations might be prescribed among operations to ensure that the YCs obey the loading sequences of the QCs and respect the stacking order of containers in the yard. Ready times might be given for stacking operations to respect the arrival times of transport vehicles at a block. Hard due dates can be given for time-critical retrieval operations. If two YCs operate simultaneously in a yard block, crane interference comes into play similar to the QCSP. For twin cranes, the non-crossing condition, as well as safety-distance requirements, has to be fulfilled (Gharehgozli et al., 2015). For dual cranes, it merely needs to be ensured that the cranes do not operate at the same yard bay at the same time. As a twin crane cannot reach every position in the yard block, a further constraint might be imposed for assigning container operations to specific cranes. In this case, an additional decision is to select a slot position where the container is handed over to the other crane, provided the container must be repositioned from one end of the block to the other.

It is evident that the YCSP is closely related to the YCDP. Hence, both problems can be solved together by a joint model. A related planning task results for an activity that is called "premarshalling." In this problem, containers are selected and repositioned in the yard in times of low workload. The idea is to provide better access (less reshuffling) when the service process of a ship starts, or when the time comes to move containers to/from trains or trucks. Repositioning operations are part of the job set in the YCSP. Further related planning tasks for the yard management are the allocation of storage slots to containers that are to be stacked, and the coordination of the crane operations with horizontal transport vehicles. Both of these decision fields impact YC operations, as they define parameters for storage/retrieval operations and impose operational constraints. Therefore, coupling storage slot allocation and horizontal vehicle scheduling with the YCSP improves the performance of the terminal subsystems. As horizontal transport operations are dynamic and volatile regarding vehicle arrival times, short-term planning and control, as well as online optimization of YC operations are important instruments too.

Gantry Crane Scheduling at the Hinterland Interface

Container terminals serve external trucks and trains that deliver export containers to the port and import containers to their destinations in the hinterland (Froyland et al., 2008). External trucks are often served directly at the yard block where the projected containers are located. Scheduling the service operation of external trucks is part of the YCSP in this case. Trains, however, are served in dedicated areas of a terminal that are equipped with their own gantry cranes. For these cranes, a scheduling problem has to be solved in order to sequence the loading and unloading process of containers to and from the trains. Similar to the YCSP, the objective is to achieve a fast handling of containers for a timely departure of trains.

The constraints are also similar to the YCSP: all operations have to be processed, due dates of containers or trains might be considered, and so on. As most rail systems do not allow for the stacking of containers on trains, direct access to every container is possible. Precedence relations can be of relevance if containers onboard a train have to be removed from their position before the loading of containers starts. As freight trains can be rather long, the shunting of trains might be required to reduce the travel effort of gantry cranes or to grant gantry cranes access to all parts of a train. This leads to a combined problem of gantry crane scheduling and train shunting, where a trade-off between train shunting and gantry crane movement must be sought.

Research Trends

In the paper, we have briefly discussed recent optimization approaches for operating the prevalent types of cranes met in container ports. There are, of course, many further topics and open questions in optimizing crane operations, which might create innovative and powerful planning models and methods in the future. New features coming up will be driven by technological progress (twin cranes, twin spreaders, indented berths), by experimental technologies (e.g., floating platforms for QCs at the waterside of a ship), or by advanced operational concepts (QC double cycling, direct ship-to-ship transshipment of containers, balancing of QC and YC workload). Even some issues of tremendous practical importance are not addressed sufficiently in the literature so far. One such example mentioned before is to ensure ship stability throughout the ship handling process, that is, when a QC schedule gets executed.

The optimization methods proposed to solve the planning problems described in this paper are quite similar to those in other scheduling problems like, for example, machine scheduling problems. Exact optimization based on mixed integer programming is, in the vast majority of cases, only applicable in relatively few instances. More advanced exact optimization techniques are not capable of substantially widening the range of solvability. Therefore, the development of tailored heuristics and, especially, of local-search-based metaheuristics has become a vital field of research. However, most studies propose methods for deterministic planning situations only, whereas few studies consider dynamics and uncertainty in crane operation processes. Such issues could be captured in terms of robust optimization or stochastic programming. As there are numerous sources of disturbances in the terminal environment (uncertain arrival times of ships and trains, breakdown of equipment, incomplete stowage plans or container data, uncertain handling times), it is of importance to support practice in coping with such challenges. More systematic comparisons are also required for the variety of models and methods proposed in these fields. This should help to consolidate knowledge and to identify those streams of research that are worth being further developed.

Finally, as was already mentioned in the previous sections, optimization problems for crane operations management have numerous interrelations among one another and also with other planning problems faced in container ports like, for example, berth planning. Although integration approaches have received some attention in the past, holistic concepts for a terminal-wide management of resources are still missing. An adjustment of a terminal's operations plans with external processes (e.g., arrival time planning for ships, trains, and trucks) can also be fruitful for achieving higher productivity and better service quality in the future.

Biographies



Frank Meisel is a Professor of Supply Chain Management at the University of Kiel, Germany. He holds a Diploma Degree in Transportation Engineering from the Technical University of Dresden and a PhD in Business Administration from Martin-Luther-University Halle-Wittenberg, Germany. His research interests include maritime logistics, sustainable transportation management, and supply chain management. He has published in *Transportation Science*, *Transportation Research Part B*, *European Journal of Operational Research*, and other peer-reviewed journals.



Christian Bierwirth is a Professor of Production and Logistics at Martin-Luther-University Halle-Wittenberg, Germany. He holds a Diploma Degree in Mathematics from the University of Heidelberg and a PhD in Business Administration from the University of Bremen. His research focuses on developing heuristics for complex scheduling problems from the fields of production and logistics. His work has been published in international journals such as the *Journal of Scheduling*, *OR Spectrum*, *Evolutionary Computation*, *European Journal of Operational Research*, *International Journal of Production Economics*, *Transportation Science*, and others.

See Also

Ports; Automation in ports; Port efficiency and effectiveness; Scheduling issues in container terminals

References

- Bierwirth, C., Meisel, F., 2009. A fast heuristic for quay crane scheduling with interference constraints. *J. Scheduling* 12, 345–360.
- Froyland, G., Koch, T., Megow, N., Duane, E., Wren, H., 2008. Optimizing the landside operation of a container terminal. *OR Spectrum* 30, 53–75.
- Gharehgozli, A.H., Laporte, G., Yu, Y., de Koster, R., 2015. Scheduling twin yard cranes in a container block. *Transp. Sci.* 49, 686–705.
- Kim, K.H., Park, Y.M., 2004. A crane scheduling method for port container terminals. *Eur. J. Oper. Res.* 156, 752–768.
- Meisel, F., Bierwirth, C., 2011a. A technique to determine the right crane capacity for a continuous quay. In: Boese, J.W. (Ed.), *Handbook of Terminal Planning*. Springer, Berlin, 155–178.
- Meisel, F., Wichmann, M., 2010. Container sequencing for quay cranes with internal reshuffles. *OR Spectrum* 32, 569–591.
- Wu, Y., Li, W., Petering, M.E.H., Goh, M., de Souza, R., 2015. Scheduling multiple yard cranes with crane interference and safety distance requirement. *Transp. Sci.* 49, 990–1005.

Further Reading

- Bierwirth, C., Meisel, F., 2010. A survey of berth allocation and quay crane scheduling problems in container terminals. *Eur. J. Oper. Res.* 202, 615–627.
- Bierwirth, C., Meisel, F., 2015. A follow-up survey of berth allocation and quay crane scheduling problems in container terminals. *Eur. J. Oper. Res.* 244, 675–689.
- Boysen, N., Briskorn, D., Meisel, F., 2017. A generalized classification scheme for crane scheduling with interference. *Eur. J. Oper. Res.* 258, 343–357.
- Carlo, H.J., Vis, I.F.A., Roodbergen, K.J., 2014. Storage yard operations in container terminals: literature overview, trends, and research directions. *Eur. J. Oper. Res.* 235, 412–430.
- Lehnfeld, J., Knust, S., 2014. Loading, unloading and premarshalling of stacks in storage areas: survey and classification. *Eur. J. Oper. Res.* 239, 297–312.
- Meisel, F., Bierwirth, C., 2011b. A unified approach for the evaluation of quay crane scheduling models and algorithms. *Comput. Oper. Res.* 38, 683–693.

Scheduling of Liner Container Shipping Services

Yuquan Du*, Qiang Meng†, *National Centre for Ports and Shipping, Australian Maritime College, University of Tasmania, Launceston, TAS, Australia; †Department of Civil and Environmental Engineering, National University of Singapore, Singapore

© 2021 Elsevier Ltd. All rights reserved.

Introduction	335
Schedule Design of Liner Shipping Services	336
Sailing Speed Decision and Bunker Fuel Cost	337
Number of Ships and Schedule Tightness	337
Container Routing Decision and Transit Time	339
A Base Mathematical Model for Schedule Design of Liner Shipping Services	340
Liner Shipping Service Scheduling Under Uncertainties	341
Uncertainties and Their Impacts on Liner Shipping Service Scheduling	341
Addressing Uncertainties With a Proactive-and-Reactive Framework	341
Main Solution Methodologies	341
Future Research Opportunities	342
Acknowledgments	343
Relevant Websites	343
References	343
Further Reading	343

Introduction

Liner shipping has been conveying cargoes to and from almost every place in the world and has become the major artery of world trade. In the process of economic globalization driven by containerization, liner shipping is the predominant service provider of containerized cargo transportation. Containerized trade volume in 2017 was 1.83 billion tons and accounts for 17% of global seaborne trade by weight, attaining up to 148 million twenty-foot equivalent units (TEUs). The liner shipping connectivity of a country measures its capacity to transport its containerized foreign trade by shipping companies and has a stronger impact on trade costs than the combined effects of logistics performance, air connectivity, costs of starting a business, and lower tariffs.

Compared to industrial shipping and tramp shipping, liner shipping is characterized by its fixed ship route and fixed schedule. The ships deployed on a liner service follow the same port rotation (itinerary). For instance, in service 1 illustrated in Fig. 1 (Wang and Meng, 2011), the deployed ships visit Hong Kong, Jakarta, and Singapore in sequence and then come back to Hong Kong again, which forms a loop. Thus, a service in liner shipping is also termed as *ship route*. The *service frequency* in liner shipping is usually weekly, with all the ships deployed on a service calling at a particular port on the same day of the week. This is the case especially for liner services connecting different countries, regions, and even continents. In the example illustrated in Fig. 1, two ships can be deployed on service 1, visiting Hong Kong every Friday, Jakarta every Friday, and Singapore every Monday. However, the feeder services serving a particular country or region generally adopt more diverse frequencies. For instance, the frequencies of the feeder services operated by Shanghai PanAsia Shipping Co. Ltd. vary from biweekly, through weekly to 2–7 visits per week to each port along a service route.

The costs of running a ship can be immense, consisting of (1) daily fixed operating costs (manning/crew costs, repair and maintenance, stores and lubricants, registration costs, management fees, etc.), (2) voyage costs (bunker fuel cost, port fees, canal dues, etc.), and (3) capital costs. From the perspective of the ship operations department of a shipping company, the operating costs of a ship only involve daily fixed operating costs and voyage costs, whereas capital costs are the concern of the financial department. Among these operating costs of the ship operations department, bunker fuel cost is the biggest single cost component: it typically represents more than 30% of the total operating cost; when the bunker fuel price is high, it can account for more than 50% of the total operating cost. The published balance sheet of Maersk Line show that a change in bunker fuel price (USD/ton) by several dollars would result in a tremendous change in its quarterly profit by several million dollars.

A liner container shipping company often operates a group of services over a number of ports to meet shipping demand and these services constitute a liner shipping *service network*. For instance, the service network shown in Fig. 1 is composed of three services. To give a more practical example, CMA CGM is operating around 500 container ships serving a global shipping network with more than 250 liner services. Without loss of generality, let us consider a service network that consists of a set of weekly services, denoted by $\mathcal{R}$, serving a group of ports, denoted by $\mathcal{P}$. For a particular service $r \in \mathcal{R}$, a port may be visited more than once by a container ship during a round trip of this service. Let m_r denote the number of ports of call on service $r \in \mathcal{R}$ and $p_r^i \in \mathcal{P}$ the corresponding i th port of call. The port rotation of service r can be expressed by:

$$p_r^1 \rightarrow p_r^2 \rightarrow \cdots \rightarrow p_r^{m_r} \rightarrow p_r^1 \quad (1)$$

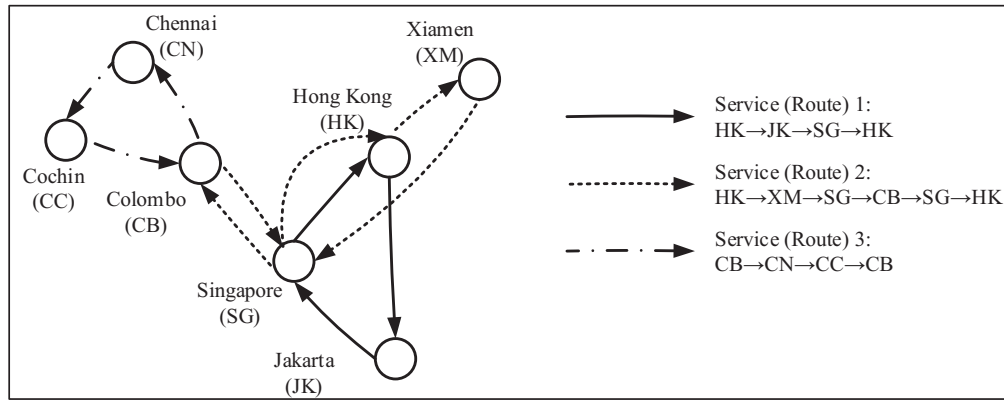


Figure 1 A network with three liner services (ship routes). Source: Wang and Meng (2011).

Let $\mathcal{I}_r = \{1, 2, \dots, m_r\}$ be the set of indices of all port calls. We define $p_r^{m_r+1} = p_r^1$ indicating that the ship comes back to the first port of call after a round trip. The voyage between two consecutive ports of call p_r^i and p_r^{i+1} is called *leg i* of the service r . For instance, the three services shown in Fig. 1 can be respectively expressed as follows:

$$r = 1, m_r = 3 : p_r^1(\text{HK}) \rightarrow p_r^2(\text{JK}) \rightarrow p_r^{m_r}(\text{SG}) \rightarrow p_r^1(\text{HK}) \quad (2)$$

$$r = 2, m_r = 5 : p_r^1(\text{HK}) \rightarrow p_r^2(\text{XM}) \rightarrow p_r^3(\text{SG}) \rightarrow p_r^4(\text{CB}) \rightarrow p_r^{m_r}(\text{SG}) \rightarrow p_r^1(\text{HK}) \quad (3)$$

$$r = 3, m_r = 3 : p_r^1(\text{CB}) \rightarrow p_r^2(\text{CN}) \rightarrow p_r^{m_r}(\text{CC}) \rightarrow p_r^1(\text{CB}) \quad (4)$$

For the service 2, $\mathcal{I}_2 = \{1, 2, 3, 4, 5\}$. Its leg 2 is from Xiamen to Singapore, and leg 4 is from Colombo to Singapore. The notation p_r^i can easily differentiate these two port calls at Singapore: p_2^3 denotes the call at Singapore after Xiamen, while p_2^5 denotes the call after Colombo.

Container transshipment plays a paramount role in a modern liner shipping network, which allows containers to be discharged from a ship deployed over one service, stay in the transshipment port for some time, and then be loaded onto another ship deployed over the other service bound for their destination. Around 85% of containers arriving at Singapore are for the purpose of transshipment. In 2017, the transshipment cargo of Hong Kong accounted for 50.9% of its total port throughput. Transshipment enables a liner container shipping company to consolidate containers at a major port (transshipment hub) to improve the capacity utilization of its mega ships over trunk services. In the meantime, transshipment also expands the geographical coverage of a shipping company's service network because much more origin-destination (O-D) port pairs can be served owing to transshipment. For instance, in the shipping network illustrated by Fig. 1, there is no direct service from Xiamen to Chennai. However, we can transport the containers from Xiamen to Chennai by utilizing services 2 and 3 with transshipment at Colombo:

$$p_2^2(\text{XM}) \xrightarrow{\text{Service 2}} p_2^3(\text{SG}) \xrightarrow{\text{Service 2}} p_2^4(\text{CB}) \mapsto p_3^1(\text{CB}) \xrightarrow{\text{Service 3}} p_3^2(\text{CN}) \quad (5)$$

The optimal design of a liner shipping network is important for a shipping company, and it has also attracted considerable research efforts from academia. To design a shipping network, we need to determine the optimal ship route (port rotation) of each service, fleet deployment (type/capacity and the number of ships for each service), sailing speed of the ships over each service, and container flows through the whole network (i.e., over each sailing leg in the network), in order to meet transport demand and generate more profit. The liner shipping network design problem is quite computationally challenging, and the solution of its medium- and large-sized instances still requires further collaborative efforts of academics and industrial professionals.

Liner shipping network design is a strategic-level decision problem confronting a shipping company. However, it does not answer how all the services are synchronized to ensure smooth container transshipment operations among services. This will be addressed in the schedule design of liner shipping services, which provides a clear answer on the arrival time of the ships at each port of call of each service.

Schedule Design of Liner Shipping Services

Simply speaking, the scheduling of a liner shipping service means the determination of the arrival time of the ships at each port of call in the service. The first point that complicates this scheduling problem is the need to consider all the services in a shipping network simultaneously, since we must ensure smooth container transshipment operations among different services. Fig. 2 (Wang and

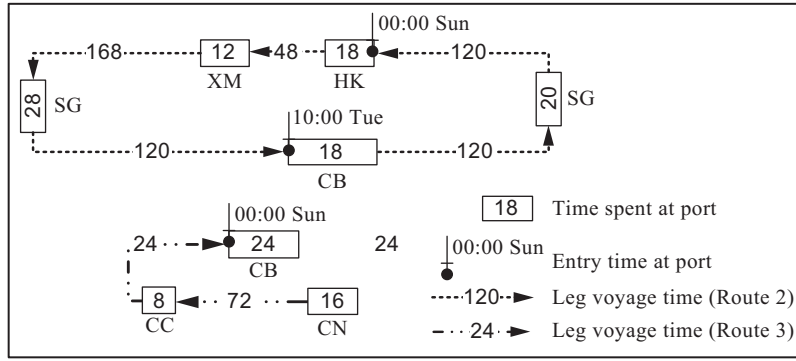


Figure 2 Schedules of service (ship route) 2 and service (ship route) 3. Source: Wang and Meng (2011).

Meng, 2011) illustrates a possible schedule design for services 2 and 3 in the example of Fig. 1. The ships over service 2 arrive at Colombo at 10:00 h on Tuesday, and the ships over service 3 arrive at Colombo at 00:00 h on Sunday. If we want to transfer some containers from service 2 to service 3 and fulfill the transshipment operations shown in Eq. (5), these containers have to stay at Colombo and wait for a ship over service 3 for around 4 days, which is undesirable. Thus, the first group of decision variables of this scheduling problem involves the arrival time of the ships at each port of call over each service of the whole network, denoted by $\{\phi_r^i | r \in \mathcal{R}, i \in \mathcal{I}_r\}$.

Sailing Speed Decision and Bunker Fuel Cost

The sailing speeds of the ships represent an essential factor in liner shipping schedule design, because sailing speed is the main determinant of a ship's fuel efficiency (i.e., daily bunker fuel consumption) and speed choice over each sailing leg will have an impact on the ship's bunker fuel consumption/cost. A *power law* is usually adopted to describe the impact of sailing speed v (knots) of a ship on its daily bunker fuel consumption Q (tons/day):

$$Q = a \times v^b \quad (6)$$

where the coefficients a and b can be calibrated based on historical data. When $b = 3$, the power law is also known as the *cubic law*. Fig. 3 (Wang and Meng, 2012a) provides some regression results over several container ships with the capacities varying from 3000 to 8000 TEUs, where the coefficient of determination R^2 is at least 0.96.

Assume that in each port of call i , a ship deployed on service r requires $\hat{t}_r^i$ units of time for queuing, pilotage in, mooring and unloading, and $\hat{t}_r^i$ units of time for loading and pilotage out of the port, then the sailing speed of the ships over leg i can be calculated later based on the schedule and its sailing distance l_r^i (nautical miles):

$$V_r^i = \frac{l_r^i}{\phi_r^{i+1} - (\phi_r^i + \hat{t}_r^i + \hat{t}_r^i)} \quad (7)$$

where V_r^i is in a technically feasible range $[\underline{v}_r, \bar{v}_r]$. Based on Eq. (6), the daily bunker fuel consumption of a ship over this leg is:

$$a \times (V_r^i)^b \quad (8)$$

Furthermore, the bunker fuel consumption of a ship over leg i is calculated as:

$$F_r^i = \frac{l_r^i}{V_r^i \times 24} \times a \times (V_r^i)^b = \frac{l_r^i \cdot a}{24} \times (V_r^i)^{b-1} \quad (9)$$

where "24" in the denominator converts the unit of sailing time from hours to days. We can see through Eqs. (7) to (9) that the schedule design of a liner service determines the ships' bunker fuel consumption. The bunker fuel cost can also be calculated given the bunker fuel price c^{fuel} (USD/ton).

Number of Ships and Schedule Tightness

The number of ships deployed in a service has a great impact on the ships' sailing speeds. Consider a liner service with three port calls. The time duration for a ship's dwelling at each port call is 24 h, and the total sea travel distance of a round trip is 6480 nautical miles. If we deploy three ships on this service, to maintain the weekly service frequency, the total trip time is 3 weeks $\times$ 7 days/week $\times$ 24 h/day = 504 h. The sea travel time, which is the difference between the total trip time and the port time, can be calculated as

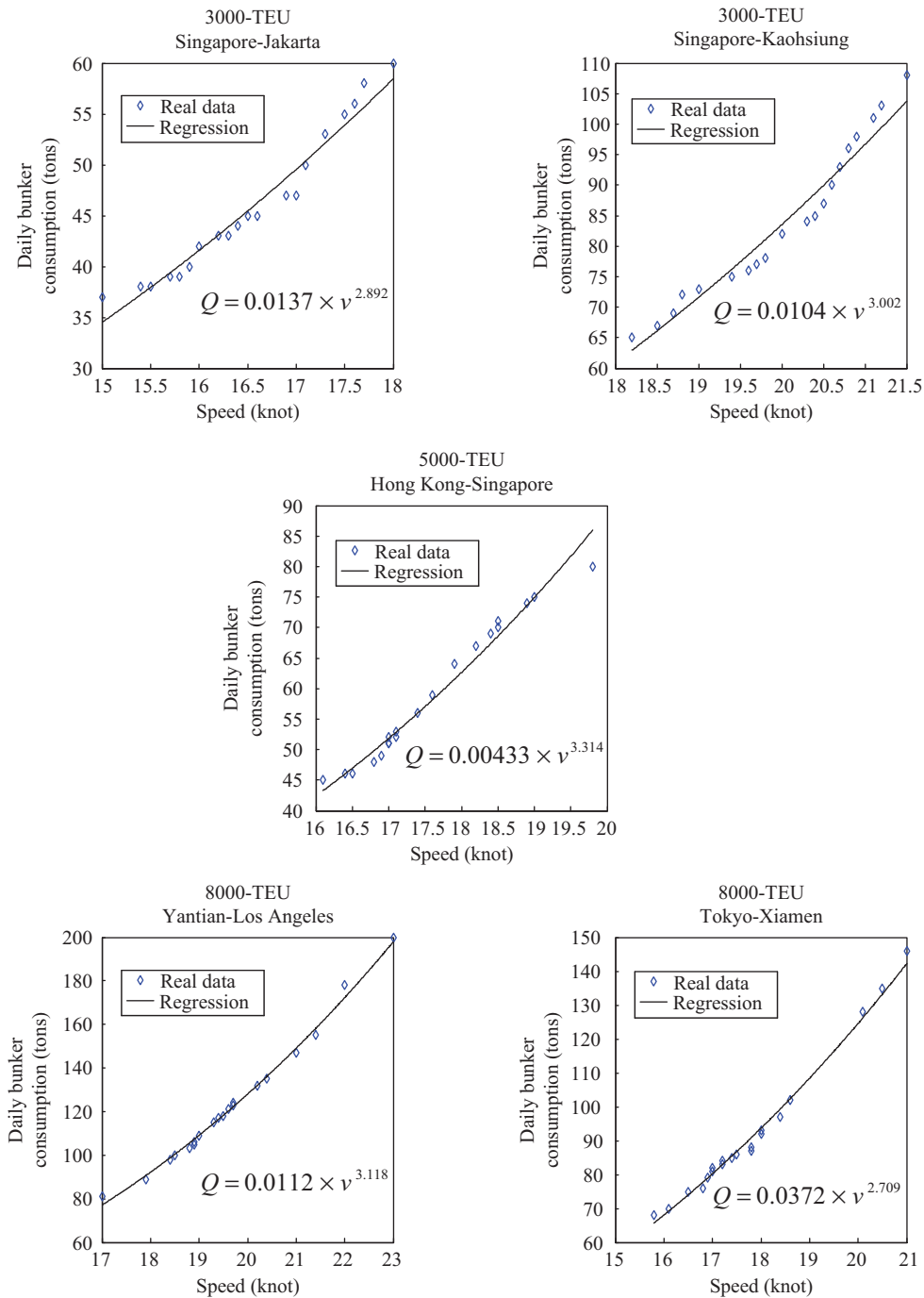


Figure 3 Bunker consumption—sailing speed relation. Source: Wang and Meng (2012a).

$504 - (24 + 24 + 24) = 432$ h. Thus, the average sailing speed of the ships is $6480/432 = 15$ knots. Alternatively, if we deploy two ships on this service, the sailing speed to maintain the weekly service frequency is:

$$\frac{6480}{2 \times 7 \times 24 - (24 + 24 + 24)} \approx 25 \text{ knots} \quad (10)$$

Compared to the slow speed of 15 knots, sailing at 25 knots reduces the number of ships required to maintain the weekly frequency and saves the fixed operating costs such as crew costs, but significantly increases the bunker fuel consumption and cost. That is why most of the shipping companies are not willing to sail their ships at high speeds when the bunker fuel price is high. Instead, they adopt a *slow-steaming* policy by deploying more ships over their services. Sailing at a slow speed also means more *buffer*

time is built into the schedule of a service. When we experience port delays or bad weather at sea, we can take advantage of the sufficient buffer time, speed up and make up the time loss due to uncertainties or disruptions. In this sense, deploying more ships over a service will reduce the tightness in a service and enhance the robustness of its schedule. Therefore, the number of ships deployed over a service concerns the trade-off between fixed operating costs and bunker fuel cost, as well as the tightness and robustness of its schedule. The number of ships deployed over service r is denoted by N_r throughout this chapter. Moreover, we denote by c_r^{fix} the fixed operating cost associated with service r .

Container Routing Decision and Transit Time

Liner container shipping schedule design is also often interwoven together with container routing decisions. We denote by $\mathcal{P}_r = \{p_r^1, \dots, p_r^{m_r}\}$ all the ports of call in service r . To transport containers from port $o \in \mathcal{P}_r$ to port $d \in \mathcal{P}_s$, a transshipment operation must be conducted from service r to service s at a connected port $p_r^i = p_s^j$. In other words, the containers traverse through some legs on service r , arrive at the connected port, switch onto service s , traverse further on some legs of service s , and finally arrive at their destination $d \in \mathcal{P}_s$, which actually forms a *container route*. The transshipment involved in this container route is denoted by $\langle r, s, i, j \rangle$. For instance, Eq. (5) shows a container route from Xiamen to Chennai with a transshipment denoted by $\langle 2, 3, 4, 1 \rangle$. If the origin and destination ports are in the same service, a container route between them exists that does not involve any transshipment. For instance:

$$p_2^2(\text{XM}) \rightarrow p_2^3(\text{SG}) \rightarrow p_2^4(\text{CB}) \quad (11)$$

is a container route that only takes advantage of service 2. Two or more transshipment operations might be involved in a container route in practice, even though this is unfavorable. All the possible transshipment operations between services of the network are assumed to be given and denoted by a set:

$$\mathcal{T} = \{\langle r, s, i, j \rangle \mid r \in \mathcal{R}, s \in \mathcal{R} \setminus \{r\}, i \in \mathcal{P}_r, j \in \mathcal{P}_s, p_r^i = p_s^j\} \quad (12)$$

For the container flow for a specific O-D pair, there may exist several container routes serving it. Let us denote by $\mathcal{H}_{od}$ the set of container routes serving the O-D pair $\langle o, d \rangle$, with the shipment demand n_{od} , and define $\mathcal{H} = \bigcup_{o \in \mathcal{P}, d \in \mathcal{P}} \mathcal{H}_{od}$. Theoretically, the total number of container routes in $\mathcal{H}$ could be exceedingly significant. However, if some business rules apply, a set $\mathcal{H}$ of practical container routes will become constructible. For instance, a container route involved in a very long journey or involving transshipment at too many ports is considered as practically infeasible. We thus assume that the set $\mathcal{H}$ is also given. Two binary parameters σ_h^r and δ_h^{rsij} are additionally adopted, with $\sigma_h^r = 1$ indicating that container route $h \in \mathcal{H}$ contains leg i of service r , and $\delta_h^{rsij} = 1$ indicating that container route $h \in \mathcal{H}$ involves the transshipment operation $\langle r, s, i, j \rangle \in \mathcal{T}$.

Obviously, we can split the shipment demand n_{od} to different container routes $h \in \mathcal{H}_{od}$, and make a decision on the container flow Y_h through each container route $h \in \mathcal{H}_{od}$ (or equivalently the container flow X_r^i over each leg $i \in \mathcal{I}_r$ of service $r \in \mathcal{R}$, taking into account the capacity of the ship cap_r in number of TEUs). This is termed the *container routing decision*. The container routing decision influences the total port charge of all the containers (a part of voyage costs), because there is a unique port handling cost c_h^{port} (USD/TEU) associated with transporting one TEU along a particular container route h . For instance, for the container route shown in Eq. (5) from Xiamen to Chennai, the total port handling cost for one TEU is the sum of the loading cost at Xiamen, the transshipment cost at Colombo, and the discharge cost at Chennai. In the meantime, the container routing decision matters because different container routes have different *transit times* and transit times reflect the level of service (LoS) provided by a shipping company. In general, for a specific O-D, a shorter transit time is more attractive for shippers/freight forwarders. A shipping company is also willing to provide diverse transit times between a specific origin and destination for the choices of shippers, because a shorter transit time usually means a more expensive freight rate and shippers can choose the appropriate freight rate—transit time combination(s) based on their preferences.

The transit time along a container route consists of three components: the sailing time at sea, the port time, and the *connection time* (waiting time) of containers for transshipment from one service to another. From the perspective of the whole shipping network, the total time at sea and in port can be calculated based on the schedules of services and container flows:

$$\sum_{r \in \mathcal{R}} \sum_{i \in \mathcal{I}_r} X_r^i \cdot (\phi_r^{i+1} - \phi_r^i - \bar{t}_r^i + \bar{t}_r^{i+1}) \quad (13)$$

Consider the connection time for transshipment $\langle r, s, i, j \rangle \in \mathcal{T}$. If $\phi_r^i \leq \phi_s^j$, the containers in service r can be directly transshipped to service s after experiencing a waiting (connection) time of $\phi_s^j - \phi_r^i$ at the transshipment port. This is represented by the case where the auxiliary decision variable $\Lambda_{rs}^{ij} = 1$ in the model of the next section. Otherwise, the containers from a ship call on service r at time ϕ_r^i will be transshipped to a ship on service s arriving at the transshipment port at time $\phi_s^j + \Phi_{rs}^{ij} \cdot 168$ (168 h every week) such that:

$$\phi_s^j + (\Phi_{rs}^{ij} - 1) \cdot 168 < \phi_r^i \leq \phi_s^j + \Phi_{rs}^{ij} \cdot 168 \quad (14)$$

where $\Phi_{rs}^{ij} \in \mathbb{Z}^+$. Solving Eq. (14) yields:

$$\Phi_{rs}^{ij} = \left\lceil \frac{\phi_r^i - \phi_s^j}{168} \right\rceil \quad (15)$$

Thus, the connection time for this case is calculated as $\phi_s^j + \Phi_{rs}^{ij} \cdot 168 - \phi_r^i$. This is represented as the case where the auxiliary decision variable $\prod_{rs}^{ij} = 1$ in the model of the next section.

In summary, the scheduling of liner container services determines the number of ships deployed N_r , the schedule $\langle \phi_r^1, \phi_r^2, \dots, \phi_r^{m_r}, \phi_r^{m_r+1} \rangle$ of each service $r \in R$ (which also implies a decision on sailing speeds), and the container flow Y_h through each container route $h \in \mathcal{H}$, in order to achieve some predefined objectives, such as the minimization of the fixed operating cost and bunker fuel cost and of the total transit time of containers through the whole network.

A Base Mathematical Model for Schedule Design of Liner Shipping Services

Based on the earlier discussion, the schedule design problem of liner shipping services can be formulated as the mixed-integer nonlinear programming model as follows:

$$\min_{N_r, V_r^i, Y_h} \sum_{r \in R} c_r^{\text{fix}} N_r + c^{\text{fuel}} \sum_{r \in R} \sum_{i \in \mathcal{I}_r} \frac{l_r^i \cdot a}{24} (V_r^i)^{b-1} + \sum_{h \in \mathcal{H}} c_h^{\text{port}} Y_h \quad (16)$$

$$\begin{aligned} \min_{X_r^i, Y_h, \phi_r^i, \phi_r^{i+1}, \phi_s^j, \Phi_{rs}^{ij}} & \sum_{r \in R} \sum_{i \in \mathcal{I}_r} X_r^i \cdot (\phi_r^{i+1} - \phi_r^i - \tilde{t}_r^i + \tilde{t}_r^{i+1}) \\ & + \sum_{o \in \mathcal{P}} \sum_{d \in \mathcal{P}} \sum_{h \in \mathcal{H}_{od}} \sum_{\langle r, s, i, j \rangle \in \mathcal{T}} \delta_h^{rsij} \cdot Y_h \cdot (168 \Phi_{rs}^{ij} + \phi_s^j - \phi_r^i) \end{aligned} \quad (17)$$

subject to:

$$\phi_r^i \leq \phi_s^j + M(1 - \Lambda_{rs}^{ij}), \quad \langle r, s, i, j \rangle \in \mathcal{T} \quad (18)$$

$$0 \leq \Phi_{rs}^{ij} \leq M(1 - \Lambda_{rs}^{ij}), \quad \langle r, s, i, j \rangle \in \mathcal{T} \quad (19)$$

$$\phi_r^i > \phi_s^j - M(1 - \Pi_{rs}^{ij}), \quad \langle r, s, i, j \rangle \in \mathcal{T} \quad (20)$$

$$\frac{(\phi_r^i - \phi_s^j)}{168} - M(1 - \Pi_{rs}^{ij}) \leq \Phi_{rs}^{ij} < \frac{(\phi_r^i - \phi_s^j)}{168} + 1 + M(1 - \Pi_{rs}^{ij}), \quad \langle r, s, i, j \rangle \in \mathcal{T} \quad (21)$$

$$\Lambda_{rs}^{ij} + \Pi_{rs}^{ij} = 1, \quad \langle r, s, i, j \rangle \in \mathcal{T} \quad (22)$$

$$X_r^i = \sum_{h \in \mathcal{H}} \sigma_h^{ri} \cdot Y_h, \quad r \in R, i \in \mathcal{I}_r \quad (23)$$

$$X_r^i \leq \text{cap}_r, \quad r \in R, i \in \mathcal{I}_r \quad (24)$$

$$V_r^i = \frac{l_r^i}{\phi_r^{i+1} - (\phi_r^i + \tilde{t}_r^i + \tilde{t}_r^{i+1})}, \quad r \in R, i \in \mathcal{I}_r \quad (25)$$

$$\underline{v}_r \leq V_r^i \leq \bar{v}_r, \quad r \in R, i \in \mathcal{I}_r \quad (26)$$

$$\sum_{h \in \mathcal{H}_{od}} Y_h = n_{od}, \quad o \in \mathcal{P}, d \in \mathcal{P} \quad (27)$$

$$\Lambda_{rs}^{ij}, \Pi_{rs}^{ij} \in \{0, 1\}, \Phi_{rs}^{ij} \in \mathbb{Z}^+ \cup \{0\}, \quad \langle r, s, i, j \rangle \in \mathcal{T} \quad (28)$$

$$0 \leq \phi_r^1 < 168, \quad r \in R \quad (29)$$

$$0 \leq \phi_r^{m_r+1} - \phi_r^1 \leq 168 N_r, \quad r \in R \quad (30)$$

$$N_r \in \mathbb{Z}^+, \quad r \in R \quad (31)$$

$$Y_h \geq 0, \quad h \in \mathcal{H} \quad (32)$$

$$\phi_1^1 = 0 \quad (33)$$

The objective (16) minimizes the sum of the fixed operating cost, bunker cost, and port handling cost, while objective (17) minimizes the total transit time composed of sailing time, port time, and connection time for transshipment. The constraints (18–22) assist the calculation of connection time in the second term of the objective (17). Constraint (23) represents the relationship between the container flow of each sailing leg and the container flows over the container routes. Constraint (24) ensures that the container flow of each sailing leg of a service should not exceed its ship capacity. Constraints (25) and (26) calculate the sailing speed over each leg and impose a technically feasible range. Constraint (27) requires that the shipment demand for each O-D pair should be exactly met. Constraints (28–32) define the domains of decision variables. Constraint (33) breaks the symmetry of the feasible domain.

Liner Shipping Service Scheduling Under Uncertainties

Uncertainties and Their Impacts on Liner Shipping Service Scheduling

The on-time arrival of ships at each port of call is a key performance indicator of a liner service's reliability. SeaIntelligence reported that the schedule reliability of the world's top 15 container shipping companies in 2018 are all below 76%, and 3 of them maintained schedule reliability of below 70% (shippingwatch.com). In January of 2019, the on-time performance of the eastbound trans-Pacific trade liner services fell below 40%.

There are two types of uncertainties in container shipping transportation that have a great impact on schedule reliability (Li et al., 2016). The first type involves the regular uncertainties that occur frequently, such as port congestion, fluctuation of port handling productivity, and unexpected waiting time before port/channel entrance. It is possible to use probabilistic models to capture the realization patterns of these types of uncertainties. The second type represents a one-off uncertainty event that occurs occasionally and irregularly, such as bad weather, labor strikes, and port maintenance. The first type of uncertainty can usually be hedged against through buffers embedded in tactical-level plans/schedules, whereas the second type of uncertainty, also known as disruption events, needs greater managerial effort in the form of real-time recovery plans. In reality, there are often combinations of these two different types of uncertainties.

Schedule unreliability, caused by the inappropriate treatment of uncertainties, has a wide influence on the partners along the whole container shipping supply chain. First, the late arrival of ships or missed deliveries at ports will increase the inventory cost of shippers, and even lead disastrously to the stock-out of their products and/or product parts. The 2019 Trans-Pacific Maritime Conference in Long Beach witnessed the disappointment of shippers about the schedule unreliability of ocean carriers. Second, schedule unreliability also causes operational cost increases to shipping companies and container ports, such as a bunker cost increase due to speedups to make up sea time loss due to uncertainties, and port costs additionally incurred to deal with the ships that missed their schedules.

Addressing Uncertainties With a Proactive-and-Reactive Framework

A proactive-and-reactive planning framework can be adopted to address the uncertainties in liner shipping service scheduling. In the proactive planning phase, a robust tactical-level shipping schedule is targeted, which is capable of absorbing regular uncertainties and provides a hedge against the possibility of bad weather at sea (sea contingency). The robustness of the schedule can be developed by introducing a certain amount of *buffer time* (slackness) that can be regarded as the additional time included in the timetable on top of the minimum required tour time. Most existing studies on liner shipping service scheduling under uncertainties have addressed how to build the appropriate amount of buffer time into a tactical-level shipping schedule of a service (Wang and Meng, 2012b). Robust liner shipping service scheduling usually considers the arrival time at each port and the number of ships deployed as the decision variables, because the number of ships together with sailing speed determines the tightness of the schedule (buffer time). Sailing speed adjustment is often considered as the recourse action for any sailing delays. The optimization objective includes the fixed operating costs, the expected bunker fuel cost under uncertainties, and the delay handling cost.

In the reactive phase, an operational-level recovery plan is targeted which provides the shipping company with optimal recourse actions after uncertainties occur. The uncertainties here can be either or both of regular and one-off uncertainty events. The recourse actions include changing sailing speed, skipping a port call, swapping two port calls, and even a cut-and-go action which stops the handling of containers at a port and allows the ship to depart immediately. The objective of an operational-level recovery plan is to enable the ships to recover from these uncertainties affecting defined schedules as soon as possible and at the least cost.

Some existing studies also take advantage of two phases to build a robust sailing schedule, where the first phase builds optimal buffer time and the second phase assesses the performance of the buffer time through recourse actions when uncertainties actually happen. Many practical considerations have also been addressed by existing studies, such as the impact of the dam lock system along a river, and the influence of the traffic convoy system and toll pricing policy of the Suez Canal.

Main Solution Methodologies

As the first resort, the state-of-the-art commercial optimization solvers, such as IBM ILOG CPLEX Optimizer and Gurobi, can be utilized to solve these scheduling models if the built models can be cast as a mixed-integer linear programming model, mixed-integer quadratic programming model (with linear and/or quadratic constraints), and even mixed-integer bilinear programming model. As

an example, the nonlinearity caused by the bunker fuel calculation in Eq. (9) can be addressed by second-order cone programming or outer approximation with linear constraints, and then CPLEX is used. Sometimes, it is also possible to design tailored cutting planes. When the impact of payload on the fuel consumption rate is additionally considered (as in the well-known *Admiralty coefficient*), nonconvexity will be created in models. In this case, the spatial branch-and-bound algorithm can be adopted which is able to produce epsilon-optimal solutions. Dynamic programming and time-space network models are also widely adopted to solve the liner shipping service scheduling problem.

As far as robust schedule design under uncertainties is concerned, stochastic programming models can be adopted to capture both here-and-now and wait-and-see decision variables. Sampling average approximation techniques are also helpful in estimating the expected values of random variables. Existing studies have also developed some stochastic dynamic programming and stochastic optimal control models for robust shipping schedules and operational recovery plans.

Future Research Opportunities

The schedule design for a liner shipping network with diverse service frequencies needs more research effort. Nowadays, global shipping routes basically adopt weekly liner services because of inertia, possible production and selling rhythms of shippers, and

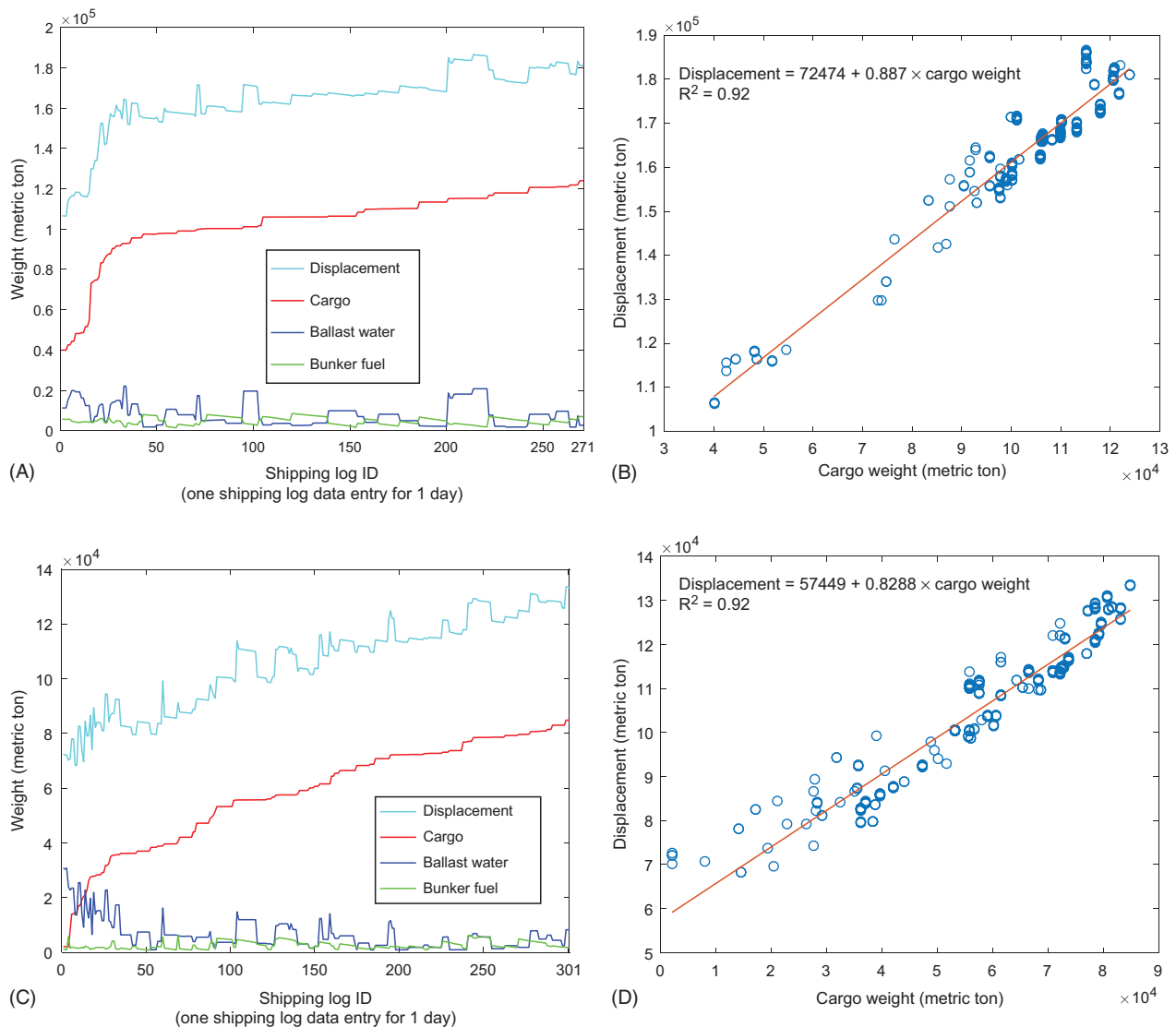


Figure 4 Relationship between ship displacement, cargo weight, and ballast water. (A) Displacement and weights of cargo, ballast water, and bunker fuel of a 14,000-TEU container ship (Ship S1) in service of 271 days. (B) Estimating the displacement of a 14,000-TEU container ship (Ship S1) with cargo weight. (C) Displacement and weights of cargo, ballast water, and bunker fuel of a 9,400-TEU container ship (Ship S2) in service of 301 days. (D) Estimating the displacement of a 9,400-TEU container ship (Ship S2) with cargo weight.

remarkable difficulty in reconfiguring the whole shipping network and synchronizing intercontinental and regional/feeder services. This motivates researchers to conduct their studies mainly in the context of weekly service frequencies. However, diverse frequencies should be further considered in future for two reasons. First, no theoretical fundamentals exist that prevent the shipping companies from introducing non-weekly services in the main trade routes in future. Second, the frequencies of feeder services are quite diverse in reality (for instance, biweekly, weekly, 5 days, 2.5 days, etc.). It would be interesting to study the schedule design problem for a shipping network with diverse frequencies and investigate the benefits of changing the frequencies (Du et al., 2017).

Ship payload is also widely ignored in existing studies, even though it has an impact on the daily bunker fuel consumption of a ship. Few studies consider this in a simple form, for example, the Admiralty coefficient, in ship routing and scheduling problems. However, in practice, the influence of weather and sea conditions on a ship's fuel efficiency might exceed that of its payload. In the meantime, directly adopting cargo load as the proxy of a ship's displacement may not be reliable, because ballast water is also a significant part of a container ship's displacement (Fig. 4). It would be intriguing to investigate, with tangible industry data support, how a ship's payload and cargo/container routing influence the bunker fuel cost in a shipping service network.

Designing the schedules for a liner shipping service network also requires more data-driven studies, through collaboration with industry partners. For instance, researchers are encouraged to conduct empirical studies on the buffer times involved in the service schedules of their industry partners based on real data, and further investigate their performance in enhancing the robustness of schedules. Historical data accumulated in shipping companies and other organizations, such as voyage report data and weather archive data, can also be adopted to model the synergetic influence of sailing speed, displacement, trim, and weather and sea conditions on ship fuel efficiency (Du et al., 2019). In the meantime, the state-of-the-art robust optimization theories developed in the past decade (Bertsimas et al., 2018; Wiesemann et al., 2014) provide advanced methodologies to deal with uncertainties in different ways, such as distributionally free robust optimization, which is promising in introducing greater insights into liner shipping schedule design with different types of uncertainties.

Acknowledgments

This paper is supported by the research project "Container haulage problems: model development, effective algorithm design and applications" funded by the NOL Fellowship Programme of Singapore.

Relevant Websites

Hong Kong SAR. Available from: www.censtatd.gov.hk.

JOC Magazine. Available from: www.joc.com.

Shanghai PanAsia Shipping. Available from: www.panasiashipping.com.

Singapore PSA. Available from: www.singaporepsa.com.

References

- Bertsimas, D., Sim, M., Zhang, M., 2018. Adaptive distributionally robust optimization. *Manag. Sci.* 65 (2), 604–618.
- Du, Y., Meng, Q., Wang, S., 2017. Mathematically calculating the transit time of cargo through a liner shipping network with various trans-shipment policies. *Marit. Policy Manag.* 44 (2), 248–270.
- Du, Y., Meng, Q., Wang, S., Kuang, H., 2019. Two-phase optimal solutions for ship speed and trim optimization over a voyage using voyage report data. *Transp. Res. Part B Methodol.* 122, 88–114.
- Li, C., Qi, X., Song, D., 2016. Real-time schedule recovery in liner shipping service with regular uncertainties and disruption events. *Transp. Res. Part B Methodol.* 93, 762–788.
- Wang, S., Meng, Q., 2011. Schedule design and container routing in liner shipping. *Transp. Res. Rec. J. Transp. Res. Board* 2222, 25–33.
- Wang, S., Meng, Q., 2012a. Sailing speed optimization for container ships in a liner shipping network. *Transp. Res. Part E Logistics Transp. Rev.* 48 (3), 701–714.
- Wang, S., Meng, Q., 2012b. Liner ship route schedule design with sea contingency time and port time uncertainty. *Transp. Res. Part B Methodol.* 46 (5), 615–633.
- Wiesemann, W., Kuhn, D., Sim, M., 2014. Distributionally robust convex optimization. *Oper. Res.* 62 (6), 1358–1376.

Further Reading

- Brouer, B.D., Dirksen, J., Pisinger, D., Plum, C.E., Vaaben, B., 2013. The vessel schedule recovery problem (VSRP)—a MIP model for handling disruptions in liner shipping. *Eur. J. Oper. Res.* 224 (2), 362–374.
- Du, Y., Meng, Q., Shi, W., 2018. Impact analysis of the traffic convoy system and toll pricing policy of the Suez Canal on the operations of a liner containership over a long-haul voyage. *Int. J. Ship. Transp. Logistics* 11, 1–26.
- Qi, X., Song, D.P., 2012. Minimizing fuel emissions by optimizing vessel schedules in liner shipping with uncertain port times. *Transp. Res. Part E Logistics Transp. Rev.* 48 (4), 863–880.
- Tan, Z., Wang, Y., Meng, Q., Liu, Z., 2018. Joint ship schedule design and sailing speed optimization for a single inland shipping service with uncertain dam transit time. *Transp. Sci.* 52 (6), 1570–1588.
- Wang, S., Meng, Q., 2012. Robust schedule design for liner shipping services. *Transp. Res. Part E Logistics Transp. Rev.* 48 (6), 1093–1106.
- Wang, Y., Meng, Q., Kuang, H., 2019. Intercontinental liner shipping service design. *Transp. Sci.* 53 (2), 344–364.
- Xia, J., Li, K.X., Ma, H., Xu, Z., 2015. Joint planning of fleet deployment, speed optimization, and cargo allocation for liner shipping. *Transp. Sci.* 49 (4), 922–938.

Dry Ports

Gordon Wilmsmeier*, **Jason Monios†**, *Kühne Professorial Chair in Logistics, School of Management, University of the Andes, Bogota, Colombia; †Associate Professor in Maritime Logistics, Kedge Business School, Marseille, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	344
Toward a Contemporary Definition of Dry Ports	344
Dry Ports in the Port Development Cycle	345
Emerging Key Questions	346
Conclusion and Research Agenda	347
References	347
Further Reading	348

Introduction

Dry ports fulfill a defined role in the development of port and hinterland systems and comply with specific functions within the multiplicity of types of inland facilities serving the logistics and transport industry. Thus it is relevant to understand their specific role from both spatio-temporal and institutional perspectives, particularly their position within the port life cycle. Given the multifarious features and concepts found in previous discussions of inland facilities in the literature, it seems necessary to clearly differentiate the dry port from other types of inland freight facilities and identify its role from a systems perspective.

Toward a Contemporary Definition of Dry Ports

The original dry port concept dates back to 1982 and emerged with the expansion of global trade and the advancement of containerization, allowing through bills of lading and directly connecting inland regions. This dry port concept was defined as “an inland terminal to which shipping companies issue their own import bills of lading for import cargoes, assuming full responsibility of costs and conditions and from which shipping companies issue their own bills of lading for export cargoes” (UNCTAD, 1982).

Beresford and Dubey (1991) argued that dry ports and inland clearance depots (ICDs) were functionally synonymous, as the focus of such facilities was to enable customs clearance at distant inland locations, particularly but not exclusively for the benefit of landlocked countries not possessing their own seaports. Nevertheless, despite the clear maritime role played by dry ports from their inception, early dry ports were generally developed by inland actors, a requirement emerging from being landlocked or otherwise suffering from poor port access. While access to rail or inland waterway transport was considered desirable, in these early discussions the transport mode was not prescribed, as the most important contribution of the dry port facility was to improve the flow of international trade for less accessible regions by saving time, delays and fees in the port, with the potential for other related services, such as intermediate storage, processing and repacking, enabling greater consolidation of shipments. While the focus in developed countries tends to be on the usage of dry ports for high volume container shuttles (see below), in developing countries where customs clearance at a seaport can take several days, customs reform is ongoing and this activity remains one of the key reasons for passing containers through a dry port, even if using road transport. For example, a study of 18 dry ports in China by Monios and Wang (2013) found that seven were rail-connected and three had rail connections nearby, while eight used road transport under customs bond to connect to seaports.

In short, a dry port is “a location in the interior of a country that fulfils original port functions” (Cullinane and Wilmsmeier, 2011). Considering that dry ports should fulfill such original port functions, they need to comply with specific infrastructure and services and must be established as a common user facility, accessible to all shippers, whether operated under a public, private or mixed model. Beresford and Dubey (1991) stressed in relation to the latter that any exclusivity in the access would defy the original idea of a dry port, a distinction that has received little attention by later authors.

Beyond the “original port functions”, the dry port concept by definition includes the type and characteristics of accessibility and interlinkage with the seaport. This high-level accessibility is related to the “main street” concept outlined by Taaffe et al. (1963), who argued that certain inland destinations will create *Bedeutungsüberschuss* due to high-priority linkages toward the seaports. In their conceptualization of port rationalization, Notteboom and Rodrigue (2005) characterize inland terminals (without defining dry ports specifically) as active nodes in shaping the logistics and transport chains in the context of linkages between ports and their hinterlands.

In the 37 years since the first definition of dry ports, the discussions and related research topics on the concept have clearly evolved (Wilmsmeier and Monios, 2020; Witte et al., 2019). Nevertheless, this evolution has been accompanied by a process

Table 1 Dry port development matrix

Phase	Development	Introduction	Operation	Extension strategy
Direction	Inside-Out	Inside-Out	Inside-Out	Inside-Out
	Outside-In	Outside-In	Outside-In	Outside-In
	Bi-directional	Bi-directional	Bi-directional	Bi-directional
	Inland only	Inland only	Inland only	Inland only

Source: The data is based on the intermodal terminal life cycle (Monios and Bergqvist, 2016) and the directional model (Wilmsmeier et al., 2011).

of vagueness in the definition of dry ports and related concepts of inland terminals or inland ports in the literature (Beresford, 2009; Cardebring and Warnecke, 1995; Jaržemskis and Vasiliauskas, 2007; Lévêque and Roso, 2002; Roso et al., 2009; UN ECE, 1998; UNESCAP, 2009). During this wave of dry port studies, where a large growth in containerized transport through ports in developed countries has led to modal shift and port congestion rising up the policy agenda, dry ports are more likely to be connected to seaports by high capacity transport modes such as rail and inland waterway. Thus the definition provided by Lévêque and Roso (2002), is that a dry port is “an inland intermodal terminal directly connected to seaport(s) with high capacity transport mean(s), where customers can leave/pick up their standardized units as if directly to a seaport.” Under this conceptualization, the role of the seaport is also foregrounded, whereby Roso et al. (2009) state that “for a fully developed dry port concept the seaport or shipping companies control the rail operations.” This emerging duality between inland and seaport actors controlling the dry port system was conceptualized in the directional model of Wilmsmeier et al. (2011) discussed in the following section.

Thus two primary axes of dry port concepts can be identified: on the one hand an ICD where the role of customs and an inland-specified bill of lading are dominant, and on the other an intermodal terminal providing access to intermodal transport for high-capacity port flows. While both definitions recommend the provision of supporting logistics activities at the dry port, this third axis of the dry port concept has become more relevant considering recent advances in logistics and the increasing complexity of the port-hinterland system. Such evolution raises the importance and desirability of colocating a dry port with some kind of logistics platform or freight village (Rodrigue et al., 2010; Monios, 2015). A modern or full-service dry port could be expected, therefore, to be a larger facility including supporting logistics activities either onsite or close nearby. Rodrigue et al. (2010) drew a useful distinction between the transport function (classifying them as satellite terminals, transmodal centers, and load centers) and the supply chain function (i. e., the level of logistics or value-added activities performed onsite or nearby). This functional approach is somewhat similar to the close, mid-range and distant dry port model presented by Roso et al. (2009) and the later sea port-based, city-based and border-based model proposed by Beresford et al. (2011). It could be argued that this kind of functional approach, based on the usage of each node, has more utility than overall terms such as “dry port” or “inland port.” It allows a research agenda to be developed along the lines of the purpose and usage of these nodes in the transport chains that they shape.

Dry Ports in the Port Development Cycle

Dry ports provide a means both to improve the accessibility of an inland location and to expand the influence of a seaport in its respective hinterland. Thus the dry port concept stands next to, and is complementary to, different types of inland terminals and any discussion cannot be separated from the evolution of ports, port hinterland systems, and the logistics sector.

Cullinane and Wilmsmeier (2011) and Wilmsmeier and Monios (2020) argue that the dry port concept is one of the possibilities in the extension of a port life cycle, when a port or port system has reached the maturity phase. A port in its maturity phase suffers from limitations in its physical expansion, due to its traditional location (often close to a city centers). These limitations are usually paired with negative externalities, such as congestion, air pollution or noise, affecting the proximity of the port and port city, a major driver being port-related freight transport. Cullinane and Wilmsmeier (2011) argued that rationalization efforts of port services, as well as process innovations and feasible investment primarily aimed at capacity effects (e.g., conversion to more effective storage technologies), will reach their limits. Consequently, a dry port development allows the possibility of “location splitting” (Schaezel, 1996), whether at short, medium or long distance to the seaport, and can create positive expansion effects. This development can enhance the flexibility and the adaptability of the port-hinterland system, release pressure on the traditional port setting and support the improvement of its environmental performance. However, it can also require adjustments in the regulatory and institutional environment, as the roles and activities of key actors in this system are transformed (van den Berg et al., 2012). The anticipated development should not only create new system capacities and alleviate negative external effects, but also needs to comply with the previously defined common user access. Further, dry port development may affect traditional hinterland competition patterns, as this development will significantly alter the hinterland of the respective port(s).

Wilmsmeier et al. (2011) argued, however, that dry port development is not always initiated as a result of a limitation or crisis in the seaport system, thus introducing the directional development perspective. This approach combines both first and second wave dry port definitions, according to which dry ports may be developed by inland or seaport actors, respectively. The directional model

helps to accentuate the challenges of port-hinterland integration and the potential conflicts between different actors. Inside out development strategies (land-driven, e.g., developed by rail operators or public bodies) can be contrasted with those that are pursued outside in (sea-driven, e.g., developed by port authorities or port terminal operators). In response to recent empirical research depicting different permutations around these two broad directions, such as bidirectional developments (Bask et al., 2014; Raimbault et al., 2016), as well as to better reflect the life cycle of both ports and inland facilities (Monios and Bergqvist, 2016), Wilmsmeier and Monios (2020) refined the directional model to account for a four-phase dry port development via-à-vis seaports: inside out, outside in, bidirectional, and no seaport involvement (Table 1). According to a recent review of publication trends on the topic, Witte et al. (2019) found that the implementation of the directional model defines what they refer to as the contextualization stage (2012–17) in inland/dry port research.

Examples of outside in development, driven from the seaward side, are Rotterdam sea port terminal operator ECT vertically integrating inland through the purchase of terminals at Venlo, Willebroek, and Duisburg (Veenstra et al., 2011), Virginia Inland Port developed by the port of Virginia, United States (Roso and Lumsden, 2009) and the Terminal Intermodal Logístico Hidalgo, located 50 km north of Mexico City and connected to the Port of Veracruz (Wilmsmeier et al., 2015). A public sector alternative is demonstrated in Spain where, for example, port authorities such as Barcelona are investing in inland terminals in the Spanish hinterland (e.g., Madrid/Coslada, Azuqueca de Henares, and Zaragoza) (Monios, 2011). The ECT example shows how a private sector company can expand its industry position through integration strategies, while the Spanish case demonstrates the use of joint ventures between public and private bodies to achieve a common goal, both being developed outside in.

Inside out developments can be quite diverse and do not always conform explicitly to one dry port definition. One example is freight villages developed by regional organizations in Italy, which can be characterized as colocated intermodal terminals and logistics platforms with customs facilities connected to ports (Monios, 2016; Monios and Wilmsmeier, 2013). Another is the Rickenbacker inland intermodal terminal in Columbus, Ohio, developed by rail operator Norfolk Southern and linked to the port of Virginia through the Heartland Corridor (Monios and Lambert, 2013). Smaller developments include several dry ports in Sweden connected to the Port of Gothenburg (Bergqvist and Woxenius, 2011). Each of these examples has been developed through different business models and leading actors range from cities to regions to rail operators to logistics providers (Bergqvist et al., 2010; Monios, 2015).

While not all site development strategies can be classified solely as one or the other, this broad conceptual distinction highlights conflicting strategies and the importance of port investment if a dry port development is to lead to a successful business, handling port container shuttles for one or more seaports (Monios and Wilmsmeier, 2012). Furthermore, there has been some debate concerning whether outside in or inside out development is the most likely model in developed or developing countries (Monios and Wilmsmeier, 2012; Monios and Wang, 2013; Ng and Cetin, 2012). It can be expected that the growing geographical coverage on port hinterland integration concepts in general and the analysis of dry ports in different geographical settings will give the opportunity to compare characteristics, factors of success or failure, as well as institutional settings.

This directional model highlights that the assumed levels of hinterland integration are accompanied by difficulties arising from the nature of creating intermodal transport corridors. The actual development and integration of such transport corridors, whether developed by ports (outside in) or by inland actors (inside out), faces significant challenges; since changing the participation of rail from a marginal share in hinterland transport requires the implementation of long-term strategies and sometimes modal shift policies. Such a major industrial shift can be difficult when the transport industry remains fragmented, large shippers refuse consolidation, and government subsidies to shift traffic to rail are fragile. At the same time, dry ports change the competition in the hinterland and the potential capture and control of inland transport networks require public sector regulation in order to prevent potential monopolization or market power (particularly given the continuing vertical integration in the shipping and port sectors).

Moreover, in many cases, the complexity of institutional design, conflicts of interest and collective action problems continue to constrain integration between maritime and inland transport systems. For example, the level of operational integration between an inland freight facility and a port (either solely through joint activities or, more formally, through ownership) can influence how daily operational issues and contractual problems between rail operators serving the dry port would be resolved (Bergqvist and Monios, 2014). Which party takes responsibility for such complexity is directly related to the institutional relationships between the various organizations (e.g., whether the port authority or port terminal operator owns the dry port, whether they own or have any integration with the rail shuttles or whether the relationship is handled through third-party contracts—Monios, 2015). Accordingly, development decisions need to be based on an analysis of whether implementing a dry port is simply a protectionist measure that would prop up a failing seaport or whether it will be planned as a node within an integrated seaport hinterland system (as argued by Cullinane and Wilmsmeier, 2011).

Emerging Key Questions

While many studies relate to the potential and development of dry ports and advance the theoretical and conceptual discussion, the question remains whether dry port development has a long-term positive effect on the performance and sustainability of logistics systems (Awad-Nunez et al., 2016). The majority of research argues for positive environmental effects of dry ports on hinterland logistics, originating from the shift toward the use of high-volume transport modes (rail or inland waterway) to connect the seaport and the dry port (Hanaoka and Regmi, 2011; Lättilä et al., 2013; Tsao and Linh, 2018). This originates from the fact that dry port development encourages investment in railway or inland waterway infrastructure to create the basis of high-volume flows (Bergqvist

et al., 2015). A shift of modes can bring significant reductions in externalities, such as alleviating road transport congestion and accidents in proximity to seaports (Fazi and Roodbergen, 2018; Gonzalez-Aregall and Bergqvist, 2019; Hanaoka and Regmi, 2011), and to reduce GHG emissions due to the better environmental performance of rail or inland shipping in comparison to road freight transport. However, the calculation of emissions reductions should not be limited to the seaport-dry port corridor, but should take the overall emissions of the whole logistics chain being served in the port hinterland production system.

Further, as derived from the port life cycle, research argues that the successful implementation of a dry port, beyond alleviating congestion, will reduce capacity limitations of storage space and operations in the port(s). A particularly relevant concept in this context is the extended gate, discussed by Rodrigue and Notteboom (2009), van Klink and van den Berg (1998), and Veenstra et al. (2011) that has been considered an example of outside-in integration as mentioned earlier. The concept of “terminal haulage” (as opposed to carrier or merchant haulage), represents a new stage of integration that could hold significant potential if technical and operational obstacles can be overcome. The extended gate terminal haulage concept can also be related to a move from push logistics toward pull logistics or even “hold logistics”, as outlined by Rodrigue and Notteboom (2009) in their concept of supply chain terminalization, whereby inland terminals are actively used to manage inventory flows.

This discussion of managing hinterland flows between the seaport and dry port emerges from the development proposal stage and remains key throughout the life of the facility. A dry port may be developed under certain assumptions about its business model and traffic sources, with particular expectations of the role it will play in the port’s ongoing strategic development (e.g., will the port guarantee traffic levels, will the port and dry port collaborate in organizing rail shuttles or will it be left to the decision of rail operators). But over the years of its operational life, the conditions may change and with that the role of the dry port (Monios and Bergqvist, 2016). The dry port may be sold or reconcessioned, it may gain or lose the business of various rail operators or even ports and it may require maintenance or upgrades and become involved in contractual disputes over who should fund these investments. Thus the governance arrangements and business model for the dry port are closely linked to its relationship with the seaport and their separate, but linked, life cycles (Table 1).

Conclusion and Research Agenda

In the evolution towards more sustainable transport systems, effective port hinterland integration is a key element and dry ports are one option. The aforementioned discussion reveals the system complexity when including the hinterland and its elements (especially the dry port) in the port life cycle perspective. The expansion and evolution of the dry port concept is intrinsically linked to capacity and structural changes, as well as to vertical integration, all of which increase both the complexity of systems and the potential price of errors or dysfunctionalities in the system. Thus key strategic questions remain, which provide an agenda for future research. Why, when and how should a seaport invest in a dry port? What kind of business model is most suitable for the dry port? How are governance and institutional issues managed between the various inland and seaport actors? Finally, regarding all three questions, how do these change over time?

References

- Awad-Nunez, S., Soler-Flores, F., González-Cancelas, N., Camarero-Orive, A., 2016. How should the sustainability of the location of dry ports be measured? *Trans. Res. Procedia* 14, 936–944.
- Bask, A., Roso, V., Andersson, D., Hämäläinen, E., 2014. Development of seaport-dry port dyads: two cases from Northern Europe. *J. Trans. Geogr.* 39, 85–95.
- Beresford, A., 2009. Dry ports: a comparative study of the UK and Nigeria. Paper presented at the International Forum on Shipping, Ports and Airports, Cardiff, UK.
- Beresford, A.K.C., Dubey, R.C., 1991. Handbook on the Management and Operation of Dry Ports. UNCTAD, Geneva, Switzerland RDP/LDC/7.
- Beresford, A., Pettit, S., Xu, Q., Williams, S., 2011. A study of dry port development in China. in: Cullinane, K.P.B., Bergqvist, R., Wilmsmeier, G. (Eds.), *Maritime Economics and Logistics*, vol. 26, pp. 73–98 (Special Issue on Dryports).
- Bergqvist, R., Falkemark, G., Woxenius, J., 2010. Establishing intermodal terminals. *WRTR* 3 (3), 285–302.
- Bergqvist, R., Macharis, C., Meers, D., Woxenius, J., 2015. Making hinterland transport more sustainable a multi actor multi criteria analysis. *Res. Transp. Bus. Manag.* 14, 80–89.
- Bergqvist, R., Monios, J., 2014. The role of contracts in achieving effective governance of intermodal terminals. *WRTR* 5 (1), 18–38.
- Bergqvist, R., Woxenius, J., 2011. The development of hinterland transport by rail – The story of Scandinavia and the Port of Gothenburg. *J. Interdis. Econ.* 23 (2), 161–177.
- Cardebring, P.W., Warnecke, C., 1995. Combi-terminal and Intermodal Freight Centre Development. KFB-Swedish Transport and Communication Research Board, Stockholm, Sweden.
- Cullinane, K.P.B., Wilmsmeier, G., 2011. The contribution of the dry port concept to the extension of port life cycles. in: Böse, J.W. (Ed.), *Handbook of Terminal Planning, Operations Research Computer Science Interfaces Series*, vol. 49. Springer, Heidelberg, Germany, pp. 359–380.
- Fazi, S., Roodbergen, K.J., 2018. Effects of demurrage and detention regimes on dry-port-based inland container transport. *Transp. Res. C Emer. Technol.* 89, 1–18.
- Gonzalez-Aregall, M., Bergqvist, R., 2019. The role of dry ports in solving seaport disruptions: a Swedish case study. *J. Trans. Geogr.* 80.
- Hanaoka, S., Regmi, M.B., 2011. Promoting intermodal freight transport through the development of dry ports in Asia: an environmental perspective. *IATSS Res.* 35 (1), 16–23.
- Jaržemskis, A., Vasiliauskas, A.V., 2007. Research on dry port concept as intermodal node. *Transport XXII* (3), 213.
- Lättilä, L., Henttu, V., Hilmola, O.-P., 2013. Hinterland operations of sea ports do matter: dry port usage effects on transportation costs and CO2 emissions. *Transp. Res. E Log. Transp. Rev.* 55, 23–42.
- Lévêque, P., Roso, V., 2002. Dry port concept for seaport inland access with intermodal solutions. Master’s Thesis, Chalmers University of Technology, Gothenburg.
- Monios, J., 2011. The role of inland terminal development in the hinterland access strategies of Spanish ports. *Res. Trans. Econ.* 33 (1), 59–66.
- Monios, J., 2015. Identifying governance relationships between intermodal terminals and logistics platforms. *Trans. Rev.* 35 (6), 767–791.
- Monios, J., 2016. Intermodal transport as a regional development strategy: the case of Italian freight villages. *Growth Chan.* 47 (3), 363–377.
- Monios, J., Bergqvist, R., 2016. Intermodal freight terminals: a life cycle governance framework. Abingdon, Routledge.
- Monios, J., Lambert, B., 2013. The heartland intermodal corridor: public-private partnerships and the transformation of institutional settings. *J. Trans. Geogr.* 27 (1), 36–45.

- Monios, J., Wang, Y., 2013. Spatial and institutional characteristics of inland port development in China. *Geo J.* 78 (5), 897–913.
- Monios, J., Wilmsmeier, G., 2012. Giving a direction to port regionalisation. *Transp. Res. A Policy Pract.* 46, 1551–1561.
- Monios, J., Wilmsmeier, G., 2013. The role of intermodal transport in port regionalisation. *Trans. Policy* 30, 161–172.
- Ng, A.K., Cetin, I.B., 2012. Locational characteristics of dry ports in developing economies: some lessons from Northern India. *Regional Stud.* 46 (6), 757–773.
- Notteboom, T.E., Rodrigue, J.-P., 2005. Port regionalization: towards a new phase in port development. *Marit. Policy Manag.* 32 (3), 297–313.
- Raimbault, N., Jacobs, W., van Dongen, F., 2016. Port regionalisation from a relational perspective: the rise of Venlo as Dutch International Logistics Hub. *Tijdschrift voor Economische en Sociale Geografie* 107 (1), 16–32.
- Rodrigue, J.-P., Debrie, J., Fremont, A., Gouvello, E., 2010. Functions and actors of inland ports: European and North American dynamics. *J. Trans. Geogr.* 18 (4), 519–529.
- Rodrigue, J.-P., Notteboom, T., 2009. The terminalization of supply chains: reassessing the role of terminals in port/hinterland logistical relationships. *Marit Policy Manag* 36 (2), 165–183.
- Roso, V., Lumsden, K., 2009. The dry port concept: moving seaport activities inland? *Trans. Commun. Bull. Asia Pacif.* 5 (78), 87–102.
- Roso, V., Woxenius, J., Lumsden, K., 2009. The dry port concept: connecting container seaports with the hinterland. *J. Trans. Geogr.* 17 (5), 338–345.
- Schaetzl, L., 1996. *Wirtschaftsgeographie 1 – Theorie*, sixth ed. UTB, Paderborn.
- Taaffe, E.J., Morrill, R.L., Gould, P.R., 1963. Transport expansion in under-developed countries: a comparative analysis. *Geograph. Rev.* 53, 503–529.
- Tsao, Y.-C., Linh, V.T., 2018. Seaport- dry port network design considering multimodal transport and carbon emissions. *J. Clean. Prod.* 199, 481–492.
- UNCTAD, 1982. Multimodal transport and containerisation. TD/B/C.4/238/Supplement 1, Part Five: Ports and Container Depots. United Nations Conference on Trade and Development, Geneva.
- UN ECE, 1998. UN/LOCODE – Code for Ports and Other Locations. Recommendation 16, Geneva.
- UNESCAP, 2009. Development of Dry Ports. UNESCAP Transport and Communications Bulletin for Asia and the Pacific, Bangkok.
- van den Berg, R., de Langen, P.W., Rua Costa, C., 2012. The role of port authorities in new intermodal service development; the case of Barcelona Port Authority. *Res. Transp. Bus. Manag.* 5, 78–84.
- van Klink, H.A., van den Berg, G., 1998. Gateways and intermodalism. *J. Trans. Geogr.* 6, 1–9.
- Veenstra, A., Zuidwijk, R., van Asperen, E., 2011. The extended gate concept for container terminals: expanding the notion of dry ports. in: Cullinane, K.P.B., Bergqvist, R., Wilmsmeier, G. (Eds.), *Maritime Economics and Logistics* vol. 19, pp. 14–32 (Special Issue on Dryports).
- Wilmsmeier, G., Monios, J., 2020. Port and dry port life cycles. in: Böse, J.W. (Ed.), *Handbook of Terminal Planning, Operations Research Computer Science Interfaces Series* (second ed.), vol. 49., Springer, Heidelberg, Germany (forthcoming).
- Wilmsmeier, G., Monios, J., Lambert, B., 2011. The directional development of intermodal freight corridors in relation to inland terminals. *J. Trans. Geogr.* 19 (6), 1379–1386.
- Wilmsmeier, G., Monios, J., Rodrigue, J.-P., 2015. Drivers for outside-in port hinterland integration in Latin America: the case of Veracruz, Mexico. *Res. Transp. Bus. Manag.* 14, 34–43.
- Witte, P., Wiegmann, B., Ng, A.K.Y., 2019. A critical review on the evolution and development of inland port research. *J. Trans. Geogr.* 74, 53–61.

Further Reading

- Bergqvist, R., Wilmsmeier, G., Cullinane, K.P.B. (Eds.), 2013. *Dry ports – A Global Perspective: Challenges and Developments in Serving Hinterlands*. Ashgate, Farnham.
- Chang, Z., Dong Yang, D., Wan, Y., Han, T., 2019. Analysis on the features of Chinese dry ports: Ownership, customs service, rail service and regional competition. *Trans. Policy* 82, 107–116.
- Cullinane, K.P.B., Bergqvist, R., Wilmsmeier, G., 2012. Special Issue: The dry port concept – theory and practice. *Marit. Econ. Logist.* 14 (1), 1–13.
- Do, N.H., Nam, K.C., Ngoc Le, Q.M., 2011. A consideration for developing a dry port system in Indochina area. *Marit. Policy Manag.* 38 (1), 1–9.
- Garnwa, P., Beresford, A., Pettit, S., 2009. Dry ports: a comparative study of the UK and Nigeria. *UNESCAP Trans. Commun. Bull. Asia Pacif.* 78, 40–56.
- Jeevan, J., Salleh, N.H.M., Loke, K.B., Saharuddin, A.H., 2017. Preparation of dry ports for a competitive environment in the container seaport system: a process benchmarking approach. *Inter. J. e-Navigat. Marit. Econ.* 7, 19–33.
- Jeevan, J., Shu-ling Chen, S.-L., Lee, E.-S., 2015. The challenges of Malaysian dry ports development. *Asian J. Shipp. Logis.* 31 (1), 109–134.
- Komchornrit, K., 2017. The selection of dry port location by a hybrid CFA-MACBETH-PROMETHEE method: a case study of Southern Thailand. *Asian J. Shipp. Logis.* 33 (3), 141–153.
- Ng, A.K.Y., Gujar, G.C., 2009. Government policies, efficiency and competitiveness: the case of dry ports in India. *Trans. Policy* 16 (5), 232–239.
- Ng, K.Y.A., Gujar, G.C., 2009. The spatial characteristics of inland transport hubs: evidences from Southern India. *J. Trans. Geogr.* 17 (5), 346–356.
- Ng, A.K.Y., Padilha, F., Pallis, A., 2013. Institutions, bureaucratic and logistical roles of dry ports: the Brazilian experiences. *J. Trans. Geogr.* 27, 46–55.
- Ng, A.K.Y., Tongzon, J.L., 2010. Transportation improvements as a catalyst for regional development in India: the role of dry ports. *Euras. Geogr. Econ.* 51 (5), 669–682.
- Nguyen, L.C., Notteboom, T., 2018. The relations between dry port characteristics and regional port-hinterland settings: findings for a global sample of dry ports. *Marit. Policy Manag.* 46 (1), 24–42.
- Padilha, P., Ng, A.K.Y., 2011. The spatial evolution of dry ports in developing economies: The Brazilian experience. in: Cullinane, K.P.B., Bergqvist, R., Wilmsmeier, G. (Eds.), *Maritime Economics and Logistics*, vol. 23, pp. 99–121 (Special Issue on Dry ports).
- Panova, Y., Hilmola, O.P., 2015. Justification and evaluation of dry port investments in Russia. *Res. Transp. Econ.* 51, 61–70.
- Roso, V., 2008. Factors influencing implementation of a dry port. *Inter. J. Phys. Distrib. Logis. Manag.* 38 (10), 782–798.
- Roso, V., Lumsden, K., 2010. A review of dry ports. *Marit. Econ. Logis.* 12 (2), 196–213.
- UNESCAP, 2006. Promoting Dry Ports as a Means of Sharing the Benefits of Globalization with Inland Locations. Bangkok: UNESCAP Committee on Managing Globalization, E/ESCAP/CMG (3/1)1.
- Wilmsmeier, G., Bergqvist, R., Cullinane, K.P.B., 2012. Ports and Hinterland: evaluating and managing location splitting. *Res. Transp. Econ.* 33 (1), 1–5.

Arctic Shipping

Yufeng Lin^{*,†}, David G. Babb[‡], Adolf K.Y. Ng^{†,§}, ^{*}Department of Supply Chain Management, Asper School of Business, University of Manitoba, Winnipeg, MB, Canada; [†]International Forum on Climate Change Adaptation for Port, Transportation Infrastructure, and the Arctic (CCAPPTIA); [‡]Centre for Earth Observation Science, University of Manitoba, Winnipeg, MB, Canada; [§]St. John's College, University of Manitoba, Winnipeg, MB, Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction	349
Actors in the Arctic	349
The Physical Aspects: The Arctic Marine Environment Under the Influence of Climate Change	350
Sea Ice	350
Arctic Weather	352
The Socioeconomic Aspects	352
The Opening of Shorter Shipping Routes	352
Infrastructural Limitations	352
The Legal Aspects	353
Regulations Regarding the Arctic Area in General	353
The Polar Code	353
United Nations Convention on the Law of the Sea (UNCLOS)	354
Regulations Regarding the Canadian Arctic	354
The Arctic Ice Regime Shipping System (AIRSS)	354
Guidelines for the Operation of Passenger Vessels in Canadian Arctic Waters (TP 13670E)	354
Pollution Prevention Guidelines for the Operation of Cruise Ships Under Canadian Jurisdiction (TP14202E)	354
Other Legislation	354
References	354
Further Reading	354

Introduction

Arctic shipping includes shipping activities occurring north of the Arctic Circle or in seasonally ice-covered regions south of the Circle for the entirety or part of their voyage. Historically, Arctic shipping has been confined to the summer season, when vessels could operate without the threat of sea ice or would only have to deal with remnant melting ice floes. This limited Arctic shipping activity to areas such as the Baltic, Bering, Barents, and southern Chukchi and Beaufort Seas, along with Hudson and Baffin Bay. However, with declining sea ice extent, trans-Arctic routes have attracted much attention and become a reality. The Northern Sea Route (NSR) has become well traveled during the open water season, connecting Europe and Asia via a trans-Arctic route through the marginal Sea of the Russian Arctic. On the North American side of the Arctic, the Northwest Passage (NWP) through the Canadian Arctic Archipelago has become increasingly open and, while several successful transits have occurred, sea ice does remain a threat from year to year along the narrow passages. Finally, a true Trans-Polar Route (TSR) connecting Europe and Asia has been forecast to become a viable shipping route as the areal extent of Arctic sea ice continues to decline into the coming decades. These routes are shorter than conventional routes via major canals and are becoming more viable as the number of ice-free days during summer continues to increase. Furthermore, the reduction in sea ice extent increases the accessibility to known resources in both the terrestrial and marine areas of the Arctic. Unsurprisingly, both Arctic and non-Arctic countries have shown substantial interest in how Arctic shipping should be developed. For example, calling itself a “near-Arctic state,” in 2018, China has included the Arctic into its Belt and Road Initiatives (BRI) (in the form of 2.0 BRI) through its “Polar Silk Road” policy. However, the feasibility and economic viability of these Arctic routes are still under intense debate, while issues pertaining to Arctic sovereignty and environmental protection are paramount. Key controversial points include the extent of genuine interest in Arctic shipping—whether shipping stakeholders treat Arctic shipping routes as serious alternatives, as well as the unclear environmental and socioeconomic impacts that Arctic shipping would pose, especially in places with a lack/shortage of infrastructural support or emergency response capacity.

Actors in the Arctic

Due to geography and different understandings of maritime boundaries, the governance of the Arctic marine environment is the subject of great debate. Within the Arctic, the Arctic Council is an intergovernmental forum that promotes cooperation, coordination, and interaction among Arctic countries and Arctic indigenous groups on economic, social, and environmental issues. It

Actors in Arctic geopolitics

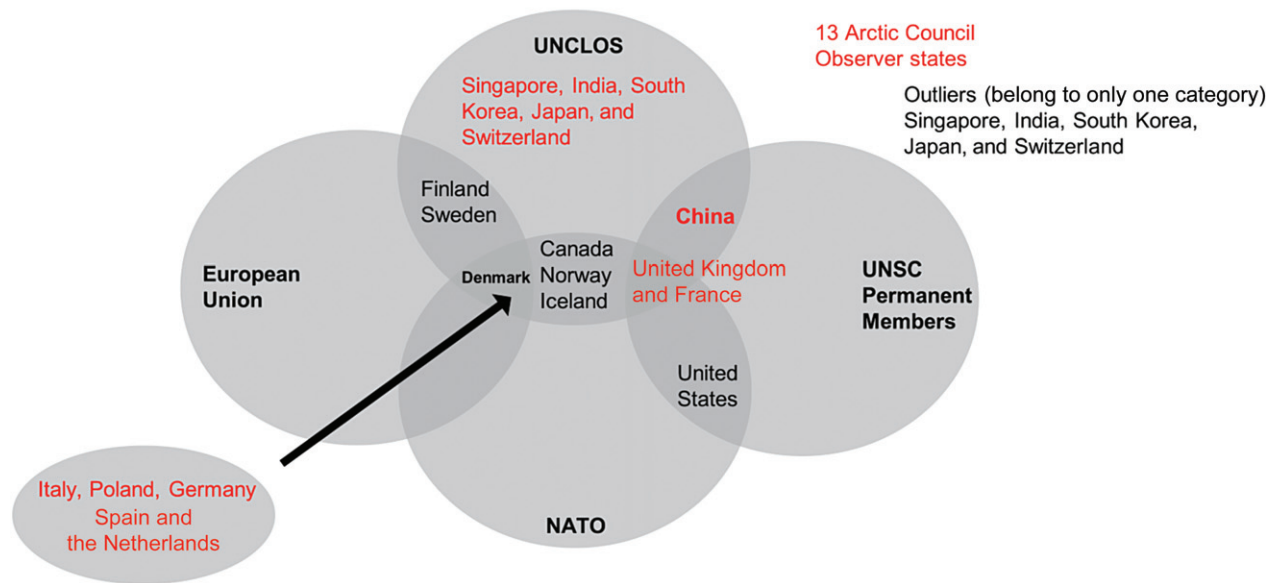


Figure 1 The major actors in Arctic geopolitics.

comprises eight member states: Canada, the Kingdom of Denmark, Finland, Iceland, Norway, the Russian Federation, Sweden, and the United States; six indigenous organizations: the Aleut International Association, the Arctic Athabaskan Council, Gwich'in Council International, the Inuit Circumpolar Council, Russian Association of Indigenous Peoples of the North, and the Saami Council; as well as other non-Arctic countries that are granted observer status such as China, India, Japan, and others (Fig. 1). Arctic Council observers primarily contribute through their engagement in the Council at the level of working groups: Arctic Contaminants Action Program (ACAP), Arctic Monitoring and Assessment Program (AMAP), Conservation of Arctic Flora and Fauna Working Group (CAFF), Emergency Prevention, Preparedness and Response Working Group (EPPR), Protection of the Arctic Marine Environment (PAME), and the Sustainable Development Working Group (SDWG). In this case, Arctic countries cooperate closely to address common challenges and solve territorial disputes by diplomatic approaches. However, the more attention the Arctic gets on the international agenda, the more the Arctic becomes an arena for strategic rivalry. Possible tension between Russia and NATO or United Nations Security Council (UNSC) Allies, as well as China's increasing engagement, could lead to a more complicated political environment regarding development in the Arctic. The United Nations Convention on the Law of the Sea (UNCLOS) is the international agreement with respect to the use of the world's oceans, establishing guidelines for businesses, the environment, and the management of marine natural resources. The UNSC is one of the six principal organs of the United Nations (UN), ensuring international peace and security. Permanent members include China, France, Russia, the United Kingdom, and the United States. NATO (the North Atlantic Treaty Organization) is an international alliance, established at the signing of the North Atlantic Treaty.

The Physical Aspects: The Arctic Marine Environment Under the Influence of Climate Change

In its Fifth Assessment Report (AR5), the Intergovernmental Panel on Climate Change (IPCC) estimated, with high confidence, that between 1880 and 2012 the global mean temperature increased roughly 0.85° C (IPCC, 2013). Moreover, temperatures in the Polar regions have increased at a significantly greater rate than the lower latitudes, due to a number of mechanisms collectively referred to as "Polar Amplification" (ACIA, 2005; Stocker, 2014). Warming temperatures in the Arctic have driven a rapid decline in sea ice over the last 50 years and have triggered changes in Arctic weather systems (IPCC, 2013), with considerable implications for marine transportation.

Sea Ice

Sea ice is a complex and dynamic medium that forms from freezing seawater and presents the greatest physical constraint on Arctic shipping. It is not only important to know where the ice is present and how much of it is there, but also what type of ice is present and how thick it is, if it is mobile and if there are hazards such as large ridges or icebergs present within the ice cover. Given the remoteness of the Arctic, information on the ice cover in a particular region is typically provided by satellite-based

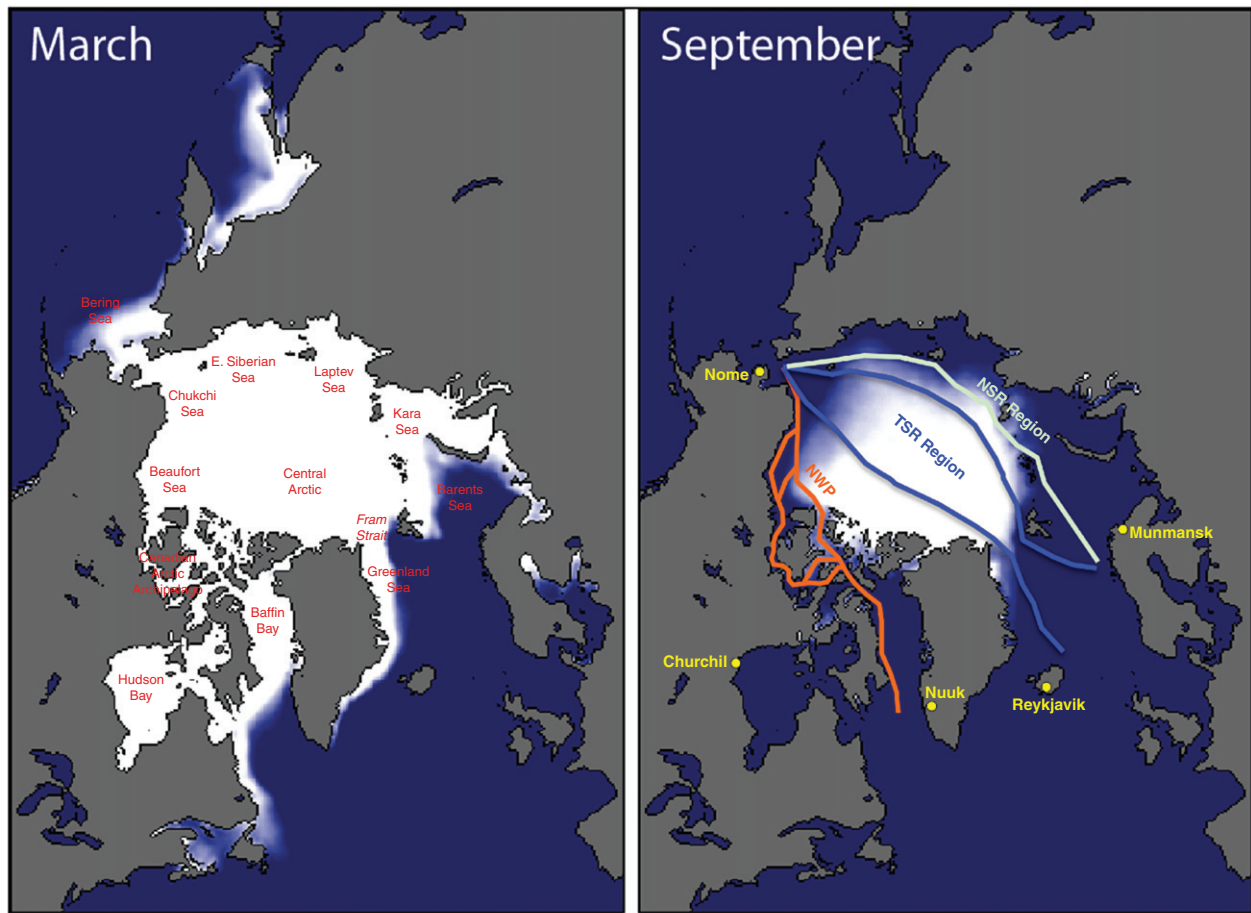


Figure 2 The 1981–2010 average maximum (March) and minimum (September) sea ice extent. Note: The three trans-Arctic shipping routes have been overlaid. Source: Data from the National Snow and Ice Data Center.

observations. National ice services present information on sea ice concentration and type in the form of regional ice charts that utilize the Egg Code of the World Meteorological Organizations (WMO). The Egg Code is a standardized method for representing sea ice concentration (in tenths), stage of development and motion, and indicates the presence of heavily deformed ice or icebergs for a particular region with standardized colors and numeric codes. Sea ice concentration describes the percent coverage of sea ice within an area with the ice edge typically being defined at the 15% (or one-tenth) contour. Sea ice type describes the stage of development of the ice cover, which can range from new thin seasonal ice to thick deformed multiyear ice. While observations of ice thickness are typically limited, ice thickness corresponds to ice type, which can therefore be used to infer the thickness and strength of ice cover in a particular area. Sea ice is characterized by a pronounced annual cycle from a maximum extent and thickness in late winter (March to April) to minimal extent and thickness at the end of the summer melt season (September) prior to the onset of fall freeze-up. At its maximum extent, there is approximately 14 million km² of sea ice in the northern hemisphere, with sea ice completely covering the Arctic Ocean and peripheral Seas like the Bering Sea and Sea of Okhotsk on the Pacific side, Hudson Bay in central Canada, and Baffin Bay and the Greenland and Barents Seas on the Atlantic side. Conversely, during the September sea ice minimum the ice cover retreats from the southern marginal seas toward the central Arctic (Fig. 2). Historically, the ice cover declined to around 7 million km², presenting open water conditions in a narrow coastal fringe around the central Arctic Ocean. However, since routine satellite observations of the ice cover began in 1978 there has been a significant decline in monthly sea ice extent during all months. This negative trend is most pronounced during summer when processes like the ice-albedo feedback loop amplify ice melt. Increasing temperatures combined with feedback cycles has reduced the average September ice extent to 4.54 million km² over the last decade, with a record minimum extent of 3.39 million km² in 2012. Reduced sea ice extent has occurred concomitantly with a decline in sea ice thickness and a transition from an older, thicker resilient ice cover to a younger, thinner, more mobile ice cover prone to enhanced ice melt. Global climate models project that this change will continue into the future, with the Arctic one day becoming ice-free. But while projecting the exact occurrence of an ice-free Arctic is inexact, the models do show continued trends toward progressively longer open water periods along both the NWP and the NSR and eventually the recurrence of open water along a transpolar sea route over the TSR. Underlying these trends, it is important to note that sea ice maintains a high degree of interannual variability, so routes that may

be ice-free one year may remain ice covered the next year. Such has been the case along the NWP during recent years as the ice cover within the Canadian Arctic remains a complex dynamic feature that still maintains thick multiyear sea ice that is able to persist through summer.

While the declining ice cover has attracted attention to Arctic sea routes, the fact is that sea ice continues to pose a threat to vessels operating in the Arctic. These vessels can either avoid sea ice by confining their activities to the seasonal open water periods in certain Arctic regions, or they can operate within the ice cover using an appropriately ice strengthened vessel. Operating during the open water period is feasible for southern areas that reliably become ice-free every summer, such as the Bering, Baltic and Barents Seas, Hudson Bay, and a majority of Baffin Bay. However, open water periods may be too short, or interannual variability in the ice cover may not reliably present ice-free conditions along certain routes. Hence, vessels operating in the Arctic typically have some measure of ice-strengthening or ice-breaking capability to contend with the risk of ice along their route. A vessel's degree of ice strengthening reflects its design (e.g., hull strength and thickness) and power. The world's shipping fleet contains a broad range of ice-strengthened vessels, varying from "Open Water" (OW) vessels capable of traveling in ice up to 15 cm thick to "Polar Class" vessels that are capable of breaking the ice and can therefore operate in ice up to several meters thick ([Transport Canada, 2010](#)). In recent decades, vessels are built according to numerous different ice-capacity classification systems, resulting in a patchwork of vessel designs and regulations ([Canadian Coast Guard, 2012](#); [Government of Canada, 1985](#)). However, with regard to ice-strengthened vessels, it is now recommended that new vessels be built according to international guidelines set out in the *Requirements Concerning Polar Class* by the International Association of Classification Societies (IACS) at the behest of the International Maritime Organization (IMO) ([International Association of Classification Societies, 2011](#); [Transport Canada, 2009](#)). Ice-strengthened vessels designed according to an earlier classification system may apply for an equivalency in the IACS system.

Arctic Weather

While sea ice is the main determinant of Arctic marine accessibility, the weather along shipping routes is a concern and has garnered far less research attention. Historically, in situ weather observations from the Arctic are limited, particularly within the marine environment, where observations are limited to a few ice camps and ship-based studies. However, the advent of autonomous drifting platforms and satellite remote sensing has improved the observational and forecasting capabilities over the Arctic marine environment. Events such as storms that increase winds and waves can be particularly dangerous in areas of the Arctic where seafloor mapping, infrastructure, and emergency response capacity can all be limited. Under a changing climate, the Arctic is expected to experience more storms, particularly during the summer when a majority of vessels operate in the Arctic. Additionally, the reduction in sea ice extent is increasing fetch and therefore promoting wavier conditions within Arctic waters ([Thomson and Rogers, 2014](#)). While vessels contend with waves in all waterways, increasing wave activity can make navigation through narrow poorly charted channels, particularly along the NWP, more hazardous. Perhaps the greater impact of a wavier Arctic Ocean will occur on coastal infrastructure that is both limited in presence and capacity and may be compromised by coastal erosion.

An issue that is unique to Arctic shipping is the potential for the accumulation of ice on the vessel from rain, sleet, snow, and fog or sea spray in near freezing temperatures. Ice accumulation impacts the balance and stability of a vessel and may create hazardous conditions for vessels operating late into the open-water shipping season or through winter when temperatures are hovering near, or well below the freezing point.

The Socioeconomic Aspects

The Opening of Shorter Shipping Routes

With climate change driving considerable changes in the Arctic, previously impeded routes are now opening, offering the potential to connect international markets across shorter polar routes ([Lasserre and Pelletier, 2011](#)). For instance, the Arctic route can cut the Rotterdam-Seattle route from 9,000 nautical miles (nm) through the Panama Canal, to 7,000 nm and the Rotterdam-Yokohama route from 11,200 nm through the Suez Canal, to 6,500 nm. Considering canal fees, fuel cost, and freight rates, Arctic shipping routes offer the potential to cut transportation costs substantially. For instance, the NSR can reduce travel distance by approximately 40%, saving up to 10% in fuel costs when compared to traditional transcontinental routes *via* major canals. However, there are cons to Arctic shipping. A major factor is related to seasonality. The lack of accessibility and reliability contributes to a much higher insurance premium. High icebreaking fees, constraints on sailing speed, and investments in dedicated ships all contribute to high costs and uncertainty, although recent developments suggest that icebreaking ships may not be needed for certain periods.

Infrastructural Limitations

Shipping in the Arctic needs to be supported by reliable, resilient infrastructure (e.g., ports, navigational aids, rescue systems, etc.) and know-how to ensure safety and security and to minimize potentially negative environmental and socioeconomic impacts. Military installations in the Russian Arctic and large port projects (e.g., Yamal) are now under construction. Within Canadian Arctic waters, the only deep-water Arctic port is located in Churchill and is returning to operation in 2018 (following its shutdown in

2016), while a deep water port at the Nanisivik Naval Facility near Arctic Bay on the northern end of Baffin Island is set to open soon, and the community of Iqaluit, on the southern end of Baffin Bay, is set to have a new modern port facility built in the coming years. Still, hitherto, ports and other infrastructures along the Arctic shipping routes are scarce and of low quality. While professional know-how in the construction, maintenance, and operation of such facilities is in serious short supply. All these require capital-intensive investments that, in turn, push up the costs of Arctic shipping. Inevitably, this raises the question as to whether Arctic shipping should be “general” or “specific” to particular type(s) of passengers and/or cargoes. For example, rather than positioning Arctic shipping as an alternative to intercontinental shipping routes based on shorter distance, there are views that shipping in the Canadian Arctic should be oriented toward resource extraction (e.g., oil and gas), food security (e.g., toward the northern communities), and other local activities (e.g., commercial fishing and cruise tourism).

The impacts of Arctic shipping on regional development cannot be ignored, as negative impacts posed by shipping are often generated by vessels. With the highly environmentally sensitive nature of the Arctic area, negative impacts can be substantial even without any serious accidents (e.g., oil spills). Hitherto, the academic and professional communities have not yet reached a consensus on the cost implications regarding insurance, escorts, and emergency response, all of which are far from being straightforward (Ostreng et al., 2013). In addition to environmental concern, there are socioeconomic aspects that need to be considered. In Canada, for example, there are at least 600 indigenous communities scattered across the northern territories and the Prairie provinces (e.g., Manitoba). Any pollution generated by shipping may affect migration routes and marine behaviors that can pose huge impacts on their livelihoods. However, vessel operators may not be aware of such impacts, especially those based in non-Arctic countries that have no direct connections with the Arctic itself. Facing such threats, *A Circumpolar Inuit Declaration on Sovereignty in the Arctic* states the rights of the Inuit (i.e., indigenous people that inhabit the Arctic regions speaking the Inuit language) in terms of natural wealth and resources (Inuit Circumpolar Council, 2015). Thus, Arctic shipping can trigger competing conflicts between global and local interests.

Of the major Arctic routes, the NSR has experienced the greatest increase in marine traffic. This is due to several reasons, including: (1) greater reductions in summer sea ice extent have occurred along the NSR compared to the NWP which is confounded by ice conditions in the Canadian Arctic Archipelago, (2) Russia, the country sovereign over this area, has invested much in terms of icebreakers, infrastructure, and other necessary facilities to make it navigable, thereby building confidence for shipowners and operators to travel along this route, and (3) the routes are better defined compared to those in the NWP (Ostreng et al., 2013). For an area that is estimated to contain much of the world’s hydrocarbon deposits, precious minerals, and other resources, it is a very strategic location for the global economy, wealth, and energy security.

Despite the heated ongoing discussion of sovereignty between Canada and the United States on the Arctic marine channels, the Canadian government has stepped up its work in this area to assure that the development of shipping can take off smoothly. A key benefit of doing such is the benefit that the northern communities are likely to receive. Unlike the Russian side that has a (relatively) high population, it is not the case for the Canadian side. The increased activities along the NWP will, therefore, open the northern communities up and make movement of goods easier. Due to the nature of the Canadian Arctic, communities are mostly accessible by air. This is normally expensive. In this case, although not technically part of the Canadian Arctic area, the town of Churchill (Manitoba, Canada), with its already-established rail links to the south, a port with decent facilities and proximity to the northern regions, has real potential to become an important maritime gateway to the Arctic area. Also, Arctic shipping is likely to bring wealth and opportunities to an otherwise remote community: people could be more willing to move to the communities to work, knowing they are assured of supplies. The cost of living is high in these areas and this discourages people from taking up employment. An improved and fast means of delivering cargo means that employees are assured of food and other essentials. The uncertainty about food and travel could be minimized.

Finally, another aspect of Arctic shipping is cruise tourism. Like many Arctic cities or towns, accessibility in terms of cruise ships could improve the tourist potential of these places. Many would like to know how the Arctic looks and the natural fauna and flora in this area, but the difficulty in traveling through ice has prevented such voyages. Arctic shipping may therefore present an unprecedented opportunity for tourism in the Arctic area.

The Legal Aspects

Regulations and guidelines for Arctic shipping vary from nation to nation. Examples of the regulations used within Canada are provided later.

Regulations Regarding the Arctic Area in General

The Polar Code

The Polar Code (IMO, 2015) stresses that ensuring stricter practice within Polar waters is warranted as an essential and positive development for Arctic shipping. The shipping and resource extraction impact on wildlife (e.g., caribou, seals, etc.) can, in turn, pose key challenges to the indigenous population whose (traditional) livelihoods depend on them. With intensified shipping in the Arctic, the risks to local autonomy and sovereignty may increase as non-Arctic countries, such as China, start to become actively involved. Such potential threats have led to the *Polar Inuit Declaration of Arctic Sovereignty* (2009) that addresses Inuit rights in terms of natural wealth and resources.

United Nations Convention on the Law of the Sea (UNCLOS)

The Law of the Sea, as included in the 1982 UNCLOS, sets out the legal framework for the regulation of shipping based on maritime zones of jurisdiction. The Law of the Sea ensures the power balance of coastal states, flag states, and port states to exercise jurisdiction and control over shipping. However, the jurisdiction status of some Arctic waters, especially internal waters and straits utilized (or potentially to be utilized) for international navigation, remains controversial and could give rise to disputes. Also, Article 234 of UNCLOS identifies coastal state powers to regulate foreign shipping in order to prevent, reduce, and control marine pollution in the Arctic.

Regulations Regarding the Canadian Arctic***The Arctic Ice Regime Shipping System (AIRSS)***

The Arctic Ice Regime Shipping System (AIRSS) is a numerical go/no-go breakdown for vessels facing ice that has been adopted heavily in shipping projections based on climate projections. The AIRSS is developed to manage the risk of pollution in the Arctic due to damage to vessels by ice, highlighting the responsibility of the shipowners for safety and providing a flexible framework for the decision-making process.

Guidelines for the Operation of Passenger Vessels in Canadian Arctic Waters (TP 13670E)

The Canadian Government has drafted the TP 13670E guidelines with the aim of improving passenger ship operations in Canada, including cruise ships. Although not a legally binding document, the guidelines strive to offer a solid framework for Arctic passenger ship operators to operate and manage their ships in a safe and pollution-free way. In addition to operational standards, the guidelines also clarify the division of responsibilities between different agents that have interests in Arctic shipping (e.g., Transport Canada, Canadian Coast Guard, Department of Fisheries and Oceans, etc.). Finally, the guidelines also consolidate the expectations and commitments of those who organize Arctic tourism and trips.

Pollution Prevention Guidelines for the Operation of Cruise Ships Under Canadian Jurisdiction (TP14202E)

The TP14202E guidelines were promulgated by the Canadian Government in 2013. It intends to help cruise ship operators develop better procedures to comply with Canadian legislation on Arctic shipping and pollution. Simultaneously, the guidelines suggest best practices that cruise ship operators are strongly encouraged to follow. The guidelines state clearly how cruise ship operators should undertake waste management (e.g., X-ray development fluid waste, contaminated materials, garbage, oily water residues, etc.). Finally, the guidelines also offer a framework on reporting, inspections, and the preparation of training and educational materials.

Other Legislation

In addition to the aforementioned, there are several legally binding laws in Canada that address Arctic shipping in general. These include the Arctic Waters Pollution Prevention Act, Canada Shipping Act 2001, Marine Liability Act, Marine Transportation Security Act, Coasting Trade Act, and the Canada Labor Code. Although these acts and legislation are not directly related to Arctic shipping, each of them plays an important role in ensuring safety, the quality of health, property, and the maritime environment within the Arctic region.

References

- ACIA, 2005. Arctic Climate Impact Assessment. Cambridge University Press, Cambridge.
- Canadian Coast Guard, 2012. Ice Navigation in Canadian Waters. Available from: <http://www.ccg-gcc.gc.ca/foi/00913/docs/ice-navigation-dans-les-galces-eng.pdf>.
- Government of Canada, 1985. Arctic Waters Pollution Prevention Regulations. Available from: http://laws-lois.justice.gc.ca/eng/regulations/C.R.C._c._354/.
- IMO, 2015. The Polar Code. International Maritime Organization, London. Available from: <http://www.imo.org/en/MediaCentre/HotTopics/polar/Pages/default.aspx>.
- International Association of Classification Societies, 2011. Requirements Concerning Polar Class. Available from: http://www.iacs.org.uk/document/public/Publications/Unified_requirements/PDF/UR_L_pdf410.pdf.
- Inuit Circumpolar Council, 2015. A Circumpolar Inuit Declaration on Sovereignty in the Arctic. Available from: <https://www.lawnow.org/circumpolar-inuit-declaration-sovereignty-arctic/>.
- IPCC, 2013. Summary for policymakers. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK/New York, NY, USA.
- Lasserre, F., Pelletier, S., 2011. Polar super seaways? Maritime transport in the Arctic: an analysis of shipowners' intentions. J. Transp. Geogr. 19, 1465–1473.
- Ostreng, W., Eger, K.M., Fløistad, B., Jørgensen-Dahl, A., Lothe, L., Mejlender-Larsen, M., Wergeland, T., 2013. Shipping in Arctic Waters: A Comparison of the Northeast, Northwest and Trans Polar Passages. Springer Science & Business Media, Berlin.
- Stocker, T. (Ed.), 2014. Climate Change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK/New York, NY, USA.
- Thomson, J., Rogers, W.E., 2014. Swell and sea in the emerging Arctic Ocean. Geophys. Res. Lett. 41 (9), 3136–3140.
- Transport Canada, 2009. Bulletin No.: 04/2009. Available from: <http://www.tc.gc.ca/eng/marinesafety/bulletins-2009-04-eng.htm>.
- Transport Canada, 2010. Arctic Ice Regime Shipping System Pictorial Guide—TP 14044. Available from: <https://www.tc.gc.ca/eng/marinesafety/debs-arctic-acts-regulations-airss-291.htm>.

Further Reading

- IPCC AR5, n.d. Available from: <https://www.ipcc.ch/report/ar5/syr/>.
- National Snow and Ice Data Center, n.d. Available from: www.nsidc.org.

Airfreight and Economic Development

Kenneth Button, Schar School of Policy and Government, George Mason University, Arlington, VA, United States

© 2021 Elsevier Ltd. All rights reserved.

Key Features of Air Cargo	355
History	356
The Air Cargo Supply Chain	356
Nature of the Economic Development Impacts	357
Empirical Findings	358
National Development	358
Regional Development	359
Sectoral Developments	359
References	359
Further Reading	360

Key Features of Air Cargo

We begin with the vocabulary. In the past “cargo” often referred to goods conveyed by ship, boat, or aircraft, whereas “freight” referred to goods conveyed through trucks and trains. With the subsequent recognition that virtually all supply chains are multimodal, this distinction has largely disappeared. Here, the terms cargo and freight are, therefore, used interchangeably. The term air-courier (or air package) services is often encountered and relates to the rapidly growing activity of moving small consignments by air, often as part of a door-to-door intermodal service. Economic development is treated generally to include increases in income or employment but is also seen to include the modernization of industrial structure in a region.

Now, let’s see some basic statistics. In 2017, according to the UN’s International Civil Aviation Organization, air transportation was globally responsible for 231,215 million ton-kilometers of cargo, including mail. Of this, 199,835 million was international. Air cargo revenue in 2018 was estimated at \$109.8 billion. Much of the cargo traffic moves through the world’s largest cargo airports. In 2018, the five largest were Hong Kong (5,120,811 tons), Memphis International Airport (4,470,196), Shanghai Pudong International Airport (3,768,573), Incheon International Airport (2,952,123), and Ted Stevens Anchorage International Airport (2,806,743). In terms of airlines, in 2017 Federal Express, which mainly offers air-courier services, was the largest carrier providing 16,851 million freight ton-kilometers, followed by Emirates (12,715), United Parcel Service (11,940), Qatar Airways (10,999), and Cathay Pacific Airways (10,722).

Air cargo services can alleviate surface infrastructure deficiencies by providing fast and reliable transportation for high-value and perishable products. It can also offer alternative distribution systems by providing a different purchasing mechanism, including e-commerce with next-day delivery instead of going to the physical store. Additionally, it provides suitable transportation for perishable luxury and exotic agricultural goods for affluent consumers, such as flying flowers and produce from Ecuador and Colombia to the United States and from Kenya to Europe. Further access by air cargo suppliers may permit firms to quickly take advantage of new market opportunities. This is because changes in freight links require minimal infrastructure investment. Inbound air cargo enables businesses to improve their physical capital stock by providing a reliable transportation mode for moving high-value equipment, machinery, and spare parts.

Access to air cargo can affect firms’ strategies. For example, being fast and reliable, air cargo can be used for emergency delivery of products allowing reduced inventory holding and just-in-time business practices. This permits disintegration of production of components and semi-processed parts as elements of an integrated supply chain. Availability of air cargo can allow firms to gain competitive advantage by using low-cost local labor and rapid access to main markets for manufactured goods, such as clothes and some electronics manufactured in, say, China and Korea for US consumers with shelf life that rapidly decay with fashion dictates.

Air transportation, therefore, has important attributes that can help facilitate economic development (Kasarda, 1991). By linking together geographically separated, but complementary, economic activities, it can generate various forms of economies of scale and allow those of specialization to be realized. Basically, it can generate economies of agglomeration (Venables, 2006). But whether this will occur, its extent and which areas will benefit is an empirical matter. It is also a matter of timing. For example, in terms of freight, while the economic growth of the likes of China, India, and Brazil has led to an overall increase in demand for the movement of heavy, bulky commodities that mainly make use of maritime modes, the rapid expansion of high-technology industries and the demand for more “exotic” products stimulated by higher incomes has seen consistent growth in air cargo.

History

The changed institutional environment from the late 1970s fostered growth in air transportation by removing the stringent regulations that had tended to stymie both air cargo and passenger aviation following the 1944 Chicago Conference. While this added some flexibility to the pre-1939 situation, rather than pursuing more liberal, and possibly multilateral policies, the tendency of countries to seek protection of their own carriers within bilateral air service agreements hampered the growth of the industry. Domestically, the geographically large nations suited to air cargo activities acted to control market entry and air cargo rates.

The first two decades of the 21st century has also seen the evolution of Internet technology and this has led to the explosive growth of e-commerce. Easier access to the global marketplace, accompanied by a rise of e-commerce, has radically changed business and consumer buying behavior. The shortening of product life spans in many industries (e.g., computers, pharmaceuticals, and designer clothes) has meant manufacturers and retailers putting a premium on reducing inventories. As a result, in 2018, e-commerce sales at \$2.8 trillion were more than double the \$1.1 trillion spent in 2012. Much of this has involved air movements, including direct from producer to consumer logistics. This had clear impacts on sectors such as retailing, with resultant shifts in location patterns and supply chains, with some companies adopting “virtual warehousing,” involving keeping goods in transit rather than using storage. Others, notably e-commerce retailers, rely on strategically located “fulfillment centers” that enable speedy and economical delivery of goods bought online.

The result overall is that, in quantitative terms, 2018 saw the annual cargo carried by air transportation to be only 1% of total world trade by tonnage, but over \$6 trillion, or 35%, in terms of value. Revenue-ton kilometers of air cargo grew annually by 2.6% between 2007 and 2017 despite the global economic recession. Boeing’s *World Air Cargo Forecast 2018–2037* (Boeing Commercial Aeroplanes, 2018) predicts this annual average rate will average 4.2% between 2018 and 2037. There is, however, quite a substantial range around this average, with growth within China’s domestic market growing by an average of 6.2% but intra-European by only 2.3% and Middle East–Europe by only 3.2%.

The traffic, however, is not evenly spread and nor is the way it is carried. Dedicated air freighter routes are, for example, concentrated on a few trade lanes, notably East Asia–North America and East Asia–Europe. In contrast, passenger networks are much broader and include destinations where cargo demand is minimal. These factors explain the load factor differential between belly space services and freighters, which average about 30% and about 70%, respectively, and the fact that freighters generate more than 90% of the air cargo industry’s revenue.

There has also been an increasing internationalization of the air cargo industry brought about through the outward expansion of former domestic carriers, the merging of two or more formerly domestic carriers, and the creation of alliances of airlines. This has been mirrored in the provision of air cargo-related services. Much of this has involved developing countries and new markets, most notably in Asia.

The range constraints on fully loaded passenger aircraft and their limited services to major cargo airports also contribute to the differences. Express carriers such as UPS and FedEx operate substantial freighter fleets, flying more than half of the wide-body freighters and generating 43% of air cargo revenues in 2017. Low-cost carriers have increased their share of air cargo traffic, particularly in Southeast Asia, but are still estimated to carry less than 2% of cargo traffic.

In terms of the implications of these trends on economic development, at the global and the mega-region level there is a clear correlation between economic development and air cargo provision; for example, annual revenue-ton kilometers fell between Europe and slowly developing Africa by 1.0% from 2007 to 2017, but rose by 4.2% between Europe and dynamic South Asia.

The Air Cargo Supply Chain

Cargo aviation is always one part of a longer supply chain involving production, storage, information systems, and a variety of feeder and distribution modes, as well as airlines. Because of this, and the numerous other factors, it is difficult to isolate the specific development effects associated with it. What is of more relevance, and more easily assessed, are the roles the air cargo supply chain can play in stimulating economic development.

Many of the earliest aviation services were for mail transportation. The first reliable scheduled airmail service began in May 1918 when a daily, round-trip service between New York and Washington, DC, was started, with a stop at Philadelphia. Airmail continued to be the main form of air cargo business during the interwar period. It was seen by the imperial powers as a means of tying empires together, by large countries, as a way to integrate and develop more remote areas, and by aspiring military powers to gain influence in lesser developed economies. As technology improved so did the range and capacity of the networks. Germany, for example, had services to South America from 1934 that, amongst other things, involved a catapult launched flying boat refueling at a permanently stationed tanker in the South Atlantic and dirigibles. By World War II planes such as the German Junkers Ju 52, which could carry up to 3 tons of cargo, were used for supplying mining and remote areas in many parts of the world.

From the late 1940s, when large numbers of former military aircraft became available for civil use, air cargo traffic grew considerably. Beside mail and package services, which were often shipped on passenger planes, converted large bombers began to be used as dedicated freighters. The advent of improved technology, most notably wide-bodied freighters and passenger aircraft with significant belly-hold capacity, together with increased efficiencies that have been built into the materials handling and air cargo systems, has significantly reduced the costs of air cargo.

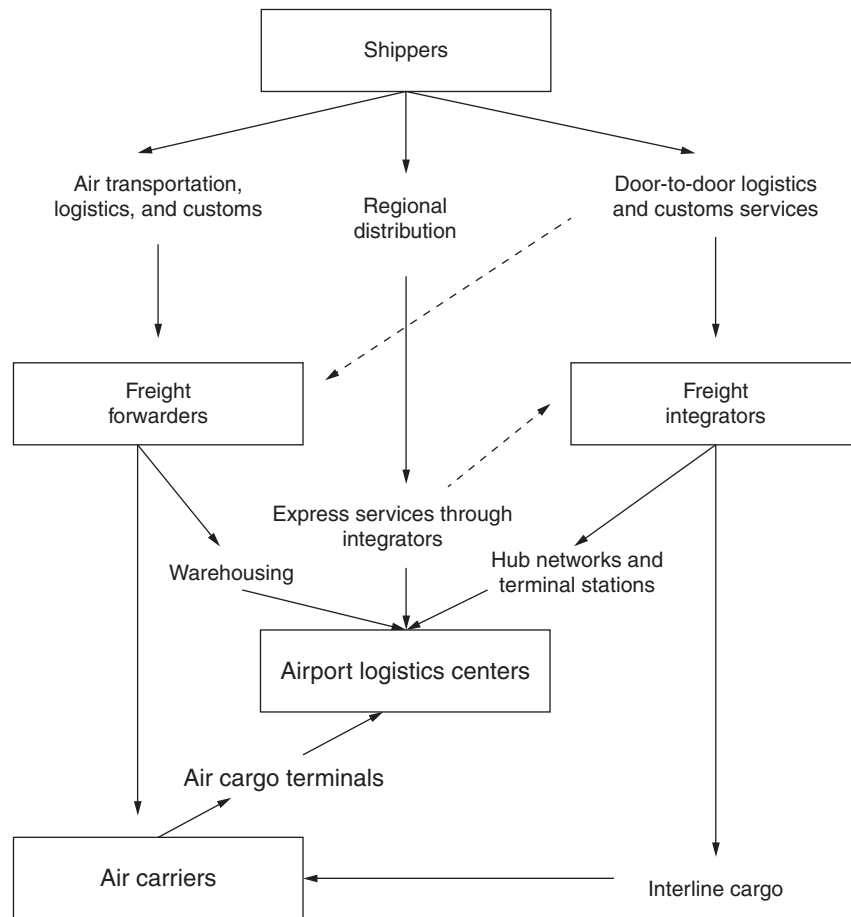


Figure 1 The air cargo logistics supply chain.

On the demand side, as industry has increased its production of high-value, light-weighted goods, such as microelectronics and pharmaceuticals, 80%–90% of their international movements are by air. The shortening of product life cycles and adoption of just-in-time manufacturing necessitates fast and reliable transportation to minimize inventories, warehousing, and packaging. As a consequence of these developments, the modern air cargo industry serves a multiplicity of functions.

Fig. 1 provides a stereotyped abstraction of the sorts of activities involved in providing air cargo logistics. It is not a simple matter of aircraft. Indeed, the strict air transportation part of the chain, highlighted in bold, is relatively limited but it does entail a significant variety of actors serving various supporting functions higher up the chain. These may be coordinated through the interactions of markets or by a coordinating agency, such as an integrated carrier, providing institutional integration and planning of the supply chain. Which is optimal is dependent on the cargo to be moved, the nature of the businesses involved, and the larger institutional environment in which the chain is set. The nature of the chain in turn affects the scale of any economic development effect and its location.

Focusing particularly on the trunk-haul air cargo part of the chain that has the most obvious development implications, this can pose serious supply issues. In particular, air cargo does not have the bilateral demands of return traffic as with passenger services. Individuals largely make return flights, but air cargo carriers have to find backhauls or fly half the time largely empty. In many cases involving developing countries, the bulk of traffic consists of exports, often natural, final products, and, in others imports, often of components and replacement parts. Use of belly-hold capacity in areas between regions with large, relatively high incomes can help spread the costs of a paucity of backhaul traffic. Complex route patterns can also help, but this may lengthen haul time and reduce service frequency. Thus, while aviation may have the potential in many cases to foster economic development, there can be threshold factors that can stymie this.

Nature of the Economic Development Impacts

The economic development effects of the supply chain can occur at different points in the chain. But even before this, air cargo can result in significant economic development as the consequence of the locations where the aircraft are constructed. The main

suppliers of both dedicated cargo carriers and the passenger aircraft that provide belly-hold capacity, Boeing and Airbus, have had major economic development impacts at their main production locations, Seattle (about 80,000 direct employees) and Toulouse (about 32,500 direct employees), respectively, and locations supplying components and undertaking aircraft maintenance and conversions. It is impossible, given the dual nature of most large aircraft and the numerous conversions from passenger to cargo planes, to isolate the number of jobs that can be attributed to cargo-plane manufacture. But most of these jobs are relatively highly paid, creating significant income in the area involved with subsequent multiplier effects.

It is around the nodes in the supply-chain, the airports and airfields, that economic effects are largely seen. There are the activities at the airport itself involving the handling and transshipment of cargo, its storage and sorting, its security, and associated data processing. The maintenance and handling of aircraft, and of the associated airport facilities, adds further jobs. These all, to varying degrees, through the incomes paid labor and fees paid companies, generate additional local fiscal multiplier effects in the local economy. The availability of air cargo services can further attract businesses that make use of air cargo to the hub's location allowing them to reduce surface transportation costs and also speed up their own logistics. This results in the creation of new industries and activities in the area; these are often called induced multiplier effects (Raguraman, 1997).

Empirical Findings

While the largest economic benefit of increased air cargo connectivity lies in its impact on the long-term performance of the wider economy through enhancement of the overall level of productivity, isolating the effects of any particular cargo mode, including air cargo, despite a mass of empirical studies, has not been easy. First, few movements are by a single mode. For example, Memphis, Tennessee, the main logistics center for FedEx Express—the world's largest airline in terms of freight tons flown and ninth largest in terms of fleet size—sees well over 1.5 million packages pass through it each night. But this is part of a larger FedEx structure that, in addition to its major cargo airport, also makes use of five Class 1 railroads and the I-40 and I-50 interstate highways that link the city to the national road network. The air cargo contribution to the city's economy is unquestionably large, but what would it be without the other transportation services?

Secondly, passenger and freight aviation are very often jointly supplied. At one level they generally share the same infrastructure, airports, and air traffic control, and at another considerable freight is moved in the belly-hold of planes that carry passengers as well. There are also complements in demand. Many tourist destinations, and notably island destinations, for example, not only rely on air transportation to bring their customers but also to bring in materials such as food, retail goods, and components necessary for a successful hospitality trade. Even with more traditional engineering industries where spare parts and components may be moved by air, these may be in the holds of planes carrying marketing, finance, and other personnel.

Causality is another matter often poorly dealt with in empirical analysis. Freight aviation may bring jobs and income to an area, but equally high-income, high employment areas can be catalysts for air services (Hakim and Merkert, 2016). Indeed, global air cargo transportation correlates 98% with world GDP growth, but does it cause it or result from it? There are statistical techniques that, with suitable data, can in some cases isolate the directions of causality, but these have not been widely used in studying the air transportation/economic development interface.

Fourthly, investment in air cargo facilities has an opportunity cost. Inevitably, while it might increase jobs and incomes where it is located, this may entail transfers of business from other areas with consequential reductions in jobs there. From an economic perspective, if there is a net gain in economic activity this may be seen as national development. But it can raise questions when it entails transfers of economic activity from a poorer to a wealthier region. The latter is not uncommon given the nature of the aviation sector and the types of things that its cargo industry moves.

But even allowing for all these caveats, there are clear cases where air cargo has been found as a necessary input for economic development. This work can be divided into that focusing on its areal implications, say the development of cities, or of its effects on specific industrial sectors, such as perishables.

National Development

Separating out the role of air cargo in global or even national economic development is impossible; there are just too many things that affect macro-level developments. At best it can be said that the emergence of freer trade in 1995 under the auspices of the World Trade Organization increased demands for international cargo services. The impacts of the resultant increase in the supply of air cargo services on national economies are difficult to isolate given other significant economic and political trends since then. What does seem important is the stage in an economy's development. Advanced air cargo services in particular have proved important in core economies and world cities in postindustrial economies, where the preponderance of knowledge-based functions is located. But at the other extreme, it has been important to less developed countries in developing labor-intensive, export-oriented industries, such as floriculture, in which they have a comparative advantage.

One thing that is clear is that air-based courier delivery services such as FedEx, UPS, DNL, etc., have been facilitators of e-commerce on a global scale and with this has come changes in the ways that business in general is conducted. Air-courier delivery services have, for example, changed how the retail sector functions, how sectors like apparel are sourced and function, and reduced inventory requirements in virtually all sectors. Whether this is development or a change in doing things is debated, but the effect has been substantial. We have earlier noted the large-scale, international network services of these courier service suppliers.

Regional Development

At the meso-regional level, the limited econometric analysis that has been conducted offers somewhat conflicting results. [Button and Yuan \(2013\)](#), for example, examined trends in employment and income for US metropolitan statistical areas with a focus on causality rather than simple correlation. Granger Causality testing based on panel data covering 35 airport and 32 metropolitan areas from 1990 to 2009 indicated that airfreight was a positive driver for local economic development at this level of spatial aggregation.

Another US study by [Green \(2007\)](#) looked at the impact of airports on the population and employment growth of 83 US metropolitan areas. After both controlling for a variety of other factors and considering causality, he found that while passenger airport activity is a powerful predictor of growth, cargo activity is not. The reason seems to be that the major cargo hubs, such as Memphis, tend to be highly automated and thus have limited direct employment effects and small local multiplier effects on their economies.

Sectoral Developments

Turning to some industries that make considerable use of air services, in 2008, about 15% of global air cargo traffic was made up of perishables and exotics where suitable transportation is defined in terms of speed, reliability, and security. In terms of trends, however, changes in maritime technology mean that computers and other goods that previously had been moved by air are increasingly being transported by ship. On the other hand, perishables, such as fruit and flowers, have experienced an annual growth in volume of over 7%, or put another way, as a share of new freight ton-kilometers they have been growing at about 12%. Much of this has been because of developments in some established production areas, but new areas are also being developed.

One example of the role of air cargo stimulating this type of economic development can be seen in its facilitating growth of the flower industry in the Victoria Lake region of East Africa, mainly Kenya, and around Addis Ababa, Ethiopia ([Button, 2020](#)). These two regions together account for about 20% of the world's cut flower exports. Without good surface access to Nairobi's and Addis's international airports, suitable refrigeration at the airports, and adequate, reliable airlines to take the blooms to European and Middle-eastern markets, the industry could not have grown. But the spread of the economic development impacts is quite small. For example, 70% of Kenya's flowers are grown at the rim of Lake Naivasha, northwest of Nairobi, and Ethiopian's floriculture industry involves a growing area of 1442 hectares. Both, however, do employ significant numbers of unskilled workers, and, notable from the economic development perspective, the vast majority are women.

But there has also been economic development in parts of South America, notably regions within Argentina, Peru, and Ecuador, that have depended on air cargo to move perishables, notably flowers, to the US market, and less so to Europe. Vega, for example, found that between 2000 and 2006 the value of perishables and exotics flown from South America to the United States grew by almost 50%, and specifically perishables from Peru by 20.1%, Argentina by 18.8%, and Ecuador by 11.3%. In terms of the economic development standpoint this, by providing jobs, improved the socioeconomic condition of impoverished rural communities. The Ecuador flower industry, for example, added an average of more than 5500 jobs a year, mainly in areas where floriculture is the single largest employer ([Vega, 2008](#)).

In terms of industrial components, [Bowen and Leinbach \(2003\)](#), looking at electronic manufacturers in Singapore, Malaysia, and the Philippines found that knowledge-intensive air cargo services played a significant role in the development of the industry. Findings of a similar general nature were found by [Aunurrofik \(2018\)](#) in a study looking across Indonesian regions. In 2011, the Chinese city Zhengzhou opened a 5 km², customs-free bonded zone on and adjacent to the airport for high-value, time-critical manufacturing and distribution. Foxconn located a manufacturing campus there that employs 240,000 workers assembling Apple's iPhones and other digital devices. Smartphone output from this doubled the value of the province's exports in 2012.

References

- Aunurrofik, A., 2018. The effect of air transportation on regional economic development: evidence from Indonesia. *Jurnal Ilmu Ekonomi* 7, 45–58.
- Boeing Commercial Aeroplanes, 2018. *World Air Cargo Forecast 2018–2037*, Boeing, Seattle, WA.
- Bowen, J., Leinbach, T.R., 2003. Air cargo services in Asian industrializing economies: electronics manufacturers and the strategic use of advanced services. *Pap. Reg. Sci.* 82, 309–332.
- Button, K.J., 2020. The economics of Africa's air cargo supply chains: the floriculture industry. *J. Transp. Geogr.*
- Button, K.J., Yuan, J., 2013. Airfreight transportation and economic development: an examination of causality. *Urban Stud.* 50, 329–340.
- Green, R.K., 2007. Airports and economic development. *Real Estate Econ.* 31, 91–112.
- Hakim, M.M., Merkert, R., 2016. The causal relationship between air transport and economic growth: empirical evidence from South Asia. *J. Transp. Geogr.* 56, 127.
- Kasarda, J.D., 1991. Global air cargo-industrial complexes as development tools. *Econ. Dev. Q.* 5, 187–196.
- Raguraman, K., 1997. International air cargo hubbing: the case of Singapore. *Asian Pac. Viewpoint* 38, 55–74.
- Vega, H., 2008. Air cargo, trade and transportation costs of perishables and exotics from South America. *J. Air Transp. Manag.* 14, 324–328.
- Venables, A.J., 2006. Incorporating wider economic impacts within cost-benefit appraisal. In: *Quantifying the Socio-Economic Benefits of Transport*, OECD, Paris, pp. 109–127.

Further Reading

- Kasarda, J.D., Green, J., 2005. Air cargo as an economic development engine: a note on opportunities and constraints. *J. Air Transp. Manag.* 11, 459–462.
- Kasarda, J.D., Green, J., Sullivan, D., 2004. *Air Cargo: Engine of Economic Development*. Center for Air Commerce, University of North Carolina at Chapel Hill, Chapel Hill, NC.
- Ying, Y.-H., Chang, C.-P., Hsieh, M.-C., 2008. Air cargo as an impetus for economic growth through the channel of openness: the case of OECD countries. *Int. J. Transp. Econ.* 35, 31–44.
- Zhang, A., Zhang, Y., 2002. Issues on liberalization of air cargo services in international aviation. *J. Air Transp. Manag.* 8, 275–287.

Air Freight Logistics

Keith Debbage*, **Neil Debbage†**, *Department of Geography, Environment and Sustainability, University of North Carolina at Greensboro, Greensboro, NC, United States; †Department of Political Science and Geography, University of Texas at San Antonio, San Antonio, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	361
Growth and Change in Air Freight Demand	361
The World's Largest Cargo Airports and the Aerotropolis Phenomenon	362
The Complex and Heterogeneous Air Freight Logistics Market	363
Major Trends and Future Challenges	366
Conclusions	366
Biographies	367
References	368

Introduction

The air freight logistics industry has become an increasingly important part of the modern world economy. In 2018, airlines transported just over 63 million metric tons of goods valued at USD 6.5 trillion, including products as diverse as cut flowers, cool chain pharmaceutical products, consumer electronics, perishable foodstuffs, and medical diagnostic devices (IATA, 2019). In many respects, the shipment of high value-to-weight commodities by air has become a crucial link in global trade, where a premium is increasingly placed on speed and timely deliveries. Although air freight shipments may account for less than 1% of global trade by volume, they now make up 35% of all global shipments by value (Shepherd et al., 2016).

The aim of this article is to better understand the changing role of air freight logistics in the contemporary global economy. This is achieved by first outlining the historical evolution of global air freight services and the geography of the major cargo airports. We then discuss Kasarda and Lindsay's (2011) theory of airport logistics-centered growth nodes, focusing on the aerotropolis concept. This is followed by a section that highlights the complex and heterogeneous nature of the air freight logistics market and concludes with an analysis of the major trends and future challenges facing the industry.

Growth and Change in Air Freight Demand

Although the 1919 Paris Convention and the 1925 Warsaw Convention essentially established the international regulatory framework that would facilitate the growth of air cargo, most developed nations initially utilized air space to primarily develop extensive air mail networks, both domestically and internationally (Budd and Ison, 2015). By the 1950s, demand for air freight services was growing, partly because longer-range, higher-capacity, and more fuel-efficient jet aircraft, such as the Boeing 707 and the Douglas DC-8 among others, had been introduced to the market, some of which are still in use today as re-engined freighters. These planes helped reduce flight costs, while new airline business models simultaneously emerged to maximize the revenue that could be obtained from transporting air freight.

Since the 1970s, demand for air freight has remained strong, with the only significant reductions occurring partly as a consequence of the first Gulf War in the early 1990s and the September 11th World Trade Center attack in 2001. Fig. 1 illustrates air freight traffic trends from 2004 through 2019, where the only substantive drop-off in demand was in response to the 2008/9 global economic recession. Although air freight traffic was relatively sluggish from 2010 until the early part of 2016, in response to the stagnant economy, traffic rebounded in 2017 with near-record growth driven by the global inventory restocking cycle and buoyant demand for high-value manufactured exports (Air Cargo World, 2018). More recently, the spread of nationalism, populism, protectionism, and trade tensions, particularly between the United States and China, have all acted to stymie growth rates in both 2018 and 2019. Others have also observed that additional factors, such as the increased competition in the air freight industry, fuel price volatility, modal shifts to cheaper sea shipping and overcapacity in the air freight sector, have all conspired to depress air cargo yields (Air Cargo World, 2019; Budd and Ison, 2015, 2017; Kupfer et al., 2017; Merkert et al., 2017).

Despite these recent trends, air freight has experienced a relatively steady long-term growth rate over the past fifty years, and in its recent World Air Cargo forecast, Boeing projected that from 2019 to 2038 air cargo operators will need more than 2800 additional freighters (Boeing, 2019). Although freighters (or all-cargo aircraft) comprise just 8% of the total commercial jet fleet, they carry more than 50% of all air cargo traffic. Boeing also expected global freight traffic to double with 4.2% growth annually largely due to the ongoing rise of e-commerce, with key markets including East Asia–North America, East Asia–Europe, and intra–East Asia.

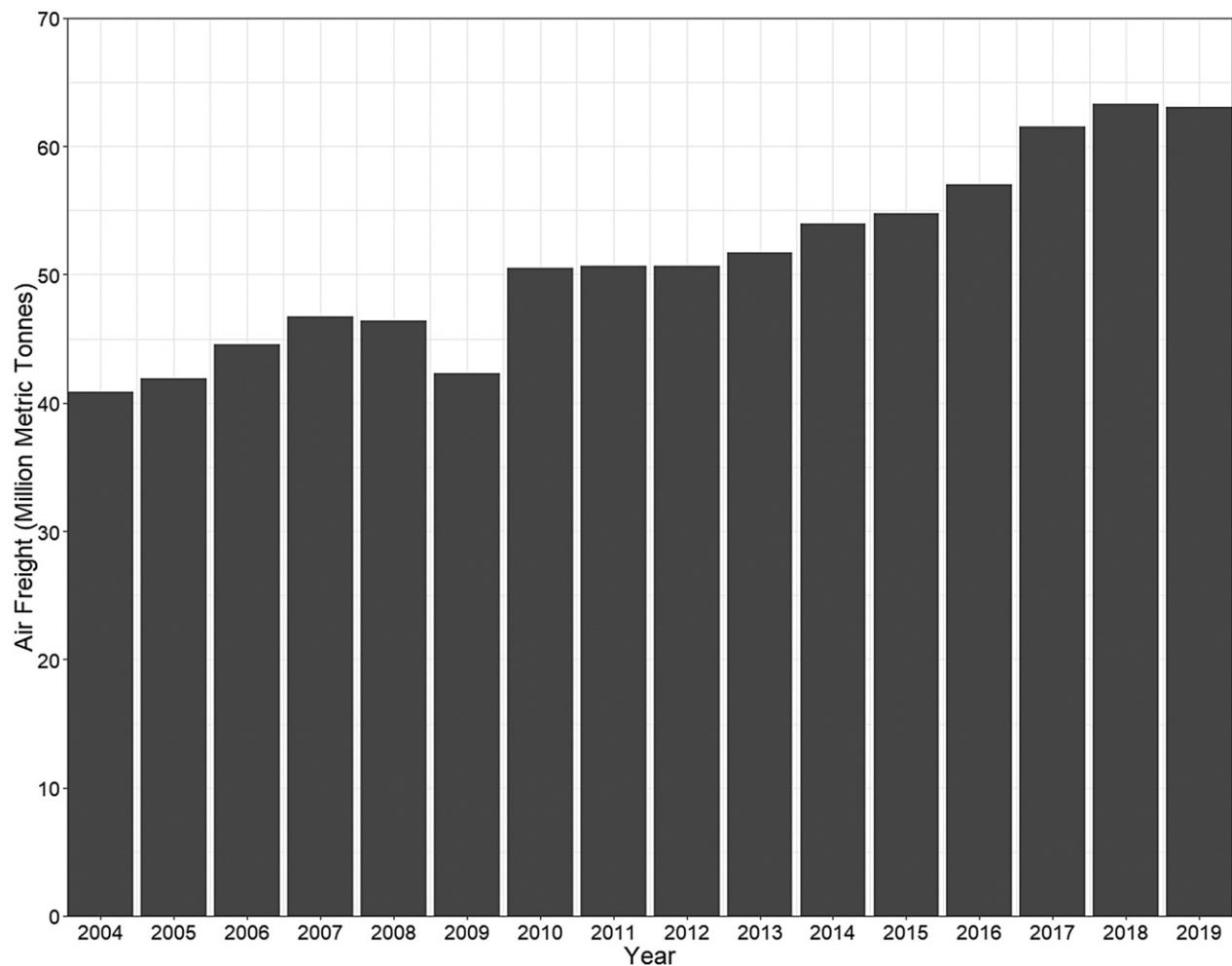


Figure 1 Worldwide air freight traffic: 2004–2019 (Million Metric Tons). Source: Data from IATA (2019)

The World's Largest Cargo Airports and the Aerotropolis Phenomenon

In 2018, the 20 largest cargo airports handled 51 million metric tons of cargo, accounting for 42% of global air cargo volume (ACI, 2019)—see Table 1. The major air cargo airports of the world play a vital role in air freight logistics, acting as the key interface between land and sky, although the spatial distribution of air freight demand is highly uneven by airport, with a small number of air freight nodes dominating the global network (Fig. 2). In 2018, the five largest cargo airports by weight included Hong Kong (5.1 million tons), Memphis (4.5 million tons), Shanghai (3.8 million tons), Incheon (2.9 million tons), and Anchorage (2.8 million tons).

Although air freight services operate on all seven continents, the majority of flights are focused on North America, Western Europe, the Middle East and East Asia. Most of the world's major cargo airports are located in heavily populated metropolitan areas and/or substantive manufacturing centers. Some exceptions to this rule include Anchorage (ranked fifth in the world), which functions largely as an important transit and refueling layover given its strategic location at the intersection of major world trade routes. Other major exceptions include the Federal Express and UPS hubs located in Memphis (ranked second) and Louisville (ranked seventh), respectively. Neither of these airports have major passenger facilities and both metropolitan areas are only medium-sized, each with less than 1.5 million in total population.

In recent years, the relative ranking of the major cargo airports has changed quite dramatically with a noticeable decline in the relative importance of Western European and North American hubs (e.g., London and Miami) and a fairly dramatic rise in the significance of both East Asian and Middle Eastern airport operations (e.g., Taipei and Doha). In 2000, 14 of the top 20 cargo airports in the world (by weight) were located in either the United States or Western Europe. By 2018, the relative mix had changed substantially, with 10 of the top 20 cargo airports now located within either East Asia or the Middle East, due to rapidly emerging markets in places like Shanghai, Dubai, Taipei, and Doha (Fig. 2).

Many observers have argued that this process of intense geographic concentration and regional specialization has fundamentally reshaped the air freight industry and related global value chains (Alkaabi and Debbage, 2011; Boonekamp and Burghouwt, 2017; Budd and Ison, 2015; Gardiner et al., 2005; Walcott and Fan, 2017). One of the most provocative conceptual frameworks for much of this research

Table 1 Top 20 air cargo airports, 2018

<i>Airport</i>	<i>Total cargo metric tons (000's)</i>
Hong Kong (HKG)	5,120
Memphis, US (MEM)	4,470
Shanghai, China (PVG)	3,768
Incheon, SK (ICN)	2,952
Anchorage, US (ANC)	2,806
Dubai, UAE (DXB)	2,641
Louisville, US (SDF)	2,623
Taipei, Taiwan (TPE)	2,322
Tokyo, Japan (NRT)	2,261
Los Angeles, US (LAX)	2,209
Doha, Qatar (DOH)	2,198
Singapore (SIN)	2,195
Frankfurt, Germany (FRA)	2,176
Paris, France (CDG)	2,156
Miami, US (MIA)	2,129
Beijing, China (PEK)	2,074
Guangzhou, China (CAN)	1,890
Chicago, US (ORD)	1,868
London, UK (LHR)	1,771
Amsterdam, Netherlands (AMS)	1,737

Source: Data from [ACI \(2019\)](#) (Preliminary Figures)

agenda is provided by [Kasarda and Lindsay \(2011\)](#). They argued that the rapid expansion of business-to-business (B2B) supply-chain transactions and the elevated emphasis on timely, fast, and reliable deliveries have played a fundamental role in generating new planned airport-related agglomerations in places like Amsterdam, Seoul, and Dubai that are fundamentally linked to the air freight logistics sector. The concept of the airport city, or what [Kasarda \(2013\)](#) has coined the “aerotropolis”, is based on the notion that some freight airports are now shaping business locations and urban development patterns, like highways, railroads, and seaports have in the past.

Critics have suggested that the aerotropolis concept is heavily contingent on relatively affordable oil prices and overlooks the substantial carbon footprint of many air freight logistics operations ([Charles et al., 2007](#)). Others have questioned the loss of natural lands, population displacement issues, problems linked to excessive aircraft noise and general lack of urban ambience associated with the aerotropolis model of development ([Bridger, 2015](#)). Despite these critiques, the aerotropolis concept is in some ways the explicit spatial manifestation of the agglomeration of air freight industries, given that it is functionally linked to time-sensitive manufacturing, e-commerce, telecommunications and third-party logistics firms. However, the air freight industry is both complex and highly heterogeneous ([Chu, 2014](#); [Kupfer et al., 2017](#); [Merkert et al., 2017](#))—a sector of the economy that defies simplistic classifications and sweeping generalizations.

The Complex and Heterogeneous Air Freight Logistics Market

Although air freight logistics plays a crucial role in the air transport chain and in the larger globalized economy, it is a highly heterogeneous industry with a wide variety of major actors and traffic flows. Airports and airlines obviously lie at the heart of most air freight logistics operations, but they are surrounded by a complex web of players including handling companies, customs brokers, and maintenance and fuel suppliers that can make any conceptualization of air freight logistics appear contrived ([Merkert et al., 2017](#)).

That said, the air freight supply chain consists of three major actors: shippers, forwarders, and carriers ([Boonekamp and Burghouwt, 2017](#); [Popescu et al., 2011](#)). The shipper can be thought of as the party that wants to have a good shipped from one place to another. The forwarder arranges the door-to-door transport of the shipment and handles all of the necessary documentation. Finally, the carrier is responsible for the airport-to-airport shipment.

With respect to the various carriers, [Kupfer et al. \(2017\)](#) have recently proposed an alternative business model for air cargo delivery and suggested that an important distinction should always be drawn between integrated and nonintegrated air cargo carriers in any discussion of air freight operations ([Fig. 3](#)). They pointed out that most of the companies in the air freight industry operate as nonintegrated service providers and they include forwarders, combination carriers (e.g., IAG, Lufthansa, Emirates) and all-cargo carriers (e.g., Kalitta Air, Southern Air, Cargolux). Combination carriers transport passengers along with cargo in the belly-hold while all-cargo carriers usually operate contract services or provide guaranteed space for third party operators and are not involved in the passenger business. However, the distinction drawn between combination carriers and all-cargo carriers is not always clear because many of the world's major passenger carriers also operate dedicated cargo-only aircraft alongside their passenger fleets (e.g., Lufthansa Cargo and Emirates Sky Cargo)

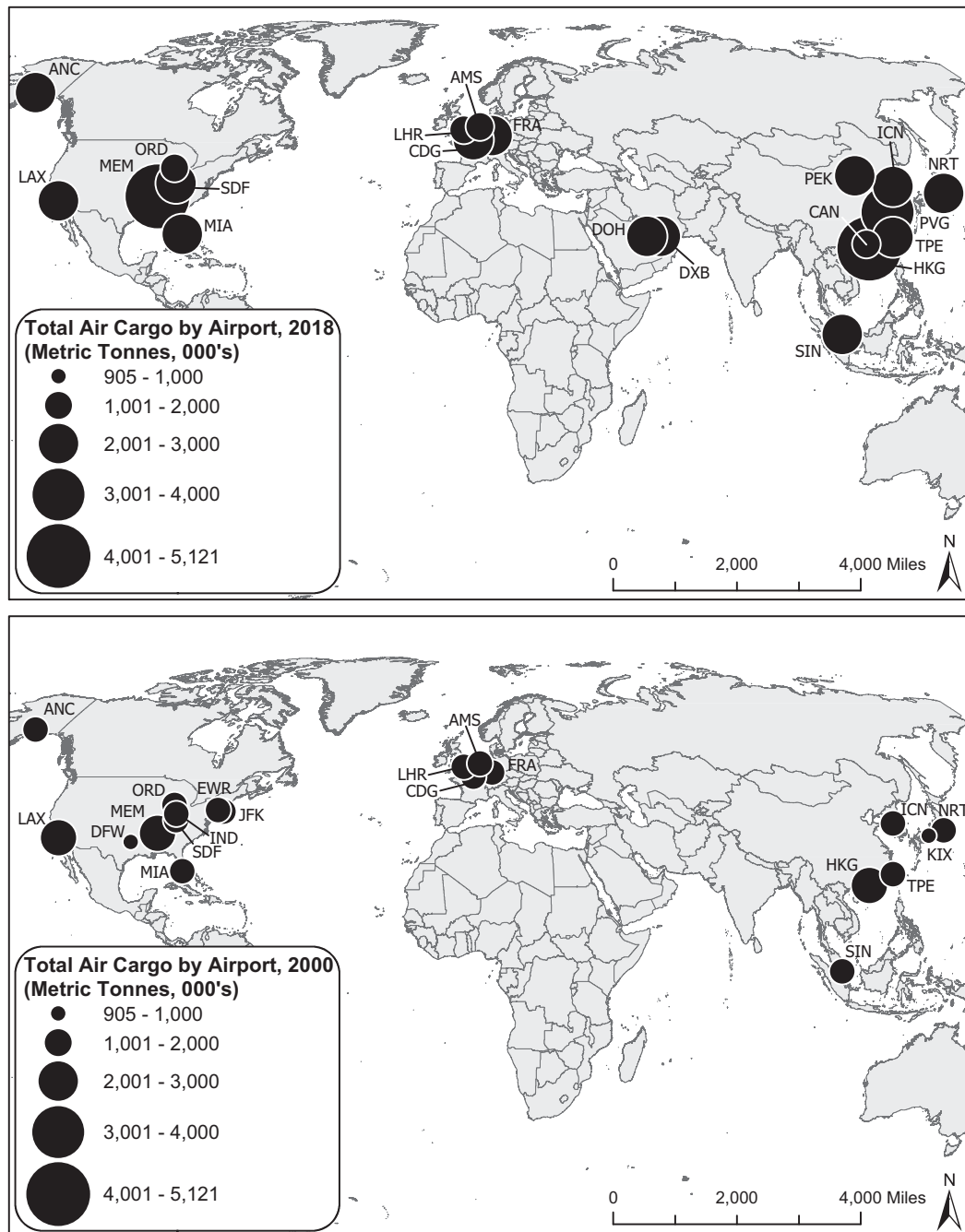


Figure 2 Spatial distribution of the top 20 largest cargo hubs in the World, 2018 (top) and 2000 (bottom). Source: Data from *ACI (2019)*

By contrast, integrated or express carriers (e.g., FedEx, UPS, and DHL) provide door-to-door collection and delivery services using a fleet of aircraft and road vehicles. In this sense, the integrators tend to own all the assets of production throughout the entire logistics value chain and, consequently, play a formidable role in shaping the air freight industry. Based on total scheduled freight ton-kilometers (FTKs) in 2017, the largest (FedEx), third largest (UPS), and fifth largest (DHL) cargo airlines worldwide were all integrated carriers (*Table 2*).

However, *Merkert et al. (2017)* remind us that any business model should also consider the vast network of service providers that act as critical external influences on the air freight logistics supply chain (*Fig. 3*). These factors can include the broader regulatory environment and labor pool as well as the practices and policies of the airport authority, among other things.

Although air freight can be transported in slightly different ways (e.g., integrated vs. nonintegrated carriers), the vast majority of approaches essentially offer the same basic service. However, *Budd and Ison (2017, p.36)* have pointed out that the decision to fly freight on a dedicated freighter (i.e., all-cargo carrier) as opposed to belly-freight on a combination carrier or via an express carrier “is

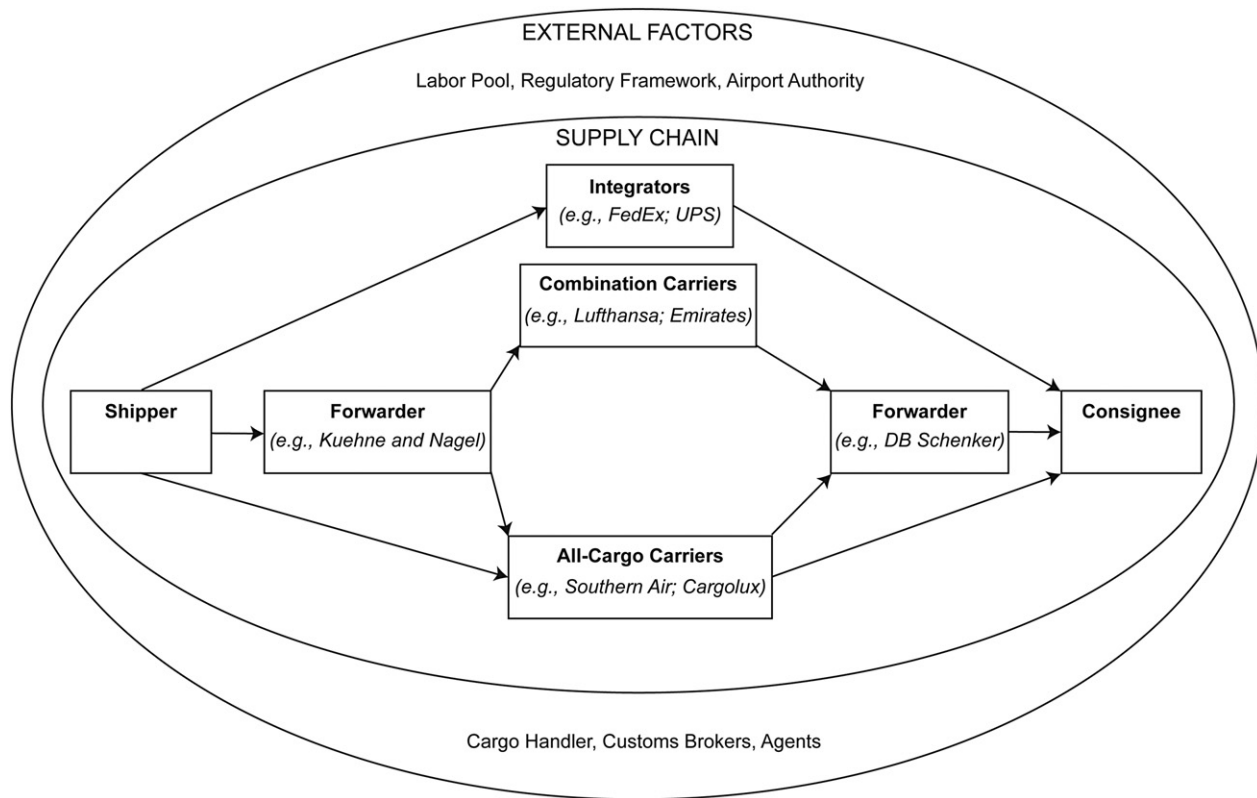


Figure 3 Business model of air freight delivery. Source: Modified from Kupfer et al. (2017) and Merkert et al. (2017)

Table 2 Top 10 cargo carriers, 2017

Airline (or airline group)	Total Cargo Traffic (FTK millions)
Federal Express	16,911
Emirates	12,979
UPS Airlines	11,940
Cathay Group	11,634
DHL Express Group	11,176
Qatar Airways	11,156
Lufthansa Group	10,692
China Southern Group	9,983
Air France – KLM Group	8,583
Korean Air	8,554

Source: Data from Air Cargo World (2018)

a commercial decision based on considerations of cost, convenience, reliability, the commercial value of the shipment and required levels of service, as well as variables including an airline's market share and market power." The end result is that air freight logistics is an industry that can often exhibit very different cost structures, operating characteristics and spatial patterns of supply and demand.

One major commercial challenge facing all air freight operators is the unique geography of many air transport routes (e.g., East Asia–Western Europe). Unlike the bi-directional nature of much air passenger transport, in air freight logistics an outbound flight is not necessarily complemented with a return or backhaul flight. The uneven spatial nature of demand is a consequence of moving goods from their site of manufacture (frequently East Asia) to their point of consumption (e.g., Western Europe and North America). To mitigate the so-called "backhaul problem", many carriers are forced to operate triangular, often times intercontinental, routes to avoid flying empty or with minimal freight loads. Although this may help to promote higher loads, it can also increase shipment times and, consequently, may not be suitable for certain types of perishable or time-critical commodities (Budd and Ison, 2017).

Although the dynamics of the air freight carrier industry are relatively well known, much less research has been conducted on the major air freight forwarders (Fig. 3). Freight forwarders act as the intermediary between shippers and carriers, as they typically

Table 3 Top 10 freight forwarders, 2018

<i>Provider</i>	<i>Country</i>	<i>Airfreight metric tons (000's)</i>
DHL Supply Chain and Global Forwarding	Germany	2,150
Kuehne and Nagel	Switzerland	1,743
DB Schenker	Germany	1,304
Panalpina ^a	Switzerland	1,038
Expeditors	United States	1,011
UPS Supply Chain Solutions	United States	935
Nippon Express	Japan	899
Bolloré Logistics	France	690
DSV ^a	Denmark	689

^aDSV acquired Panalpina in early 2019 making the combined company the third largest air freight forwarder in the world.

Source: Data from *Air Cargo World* (2019)

consolidate the merchandise of individual shippers into larger quantities of freight and then book cargo space with the major air freight carriers to ensure a more efficient and economical shipment (Chu, 2014). Based on airfreight tons shipped in 2018, the leading freight forwarders included German-based DHL Supply Chain and Global Forwarding (2.15 million tons), Kuehne and Nagel (1.7 million tons), and DB Schenker (1.3 million tons) (Table 3). Unlike the world's largest cargo airports, the leading freight forwarders are still largely based in Western Europe or the United States, although companies like China's Apex Logistics International and Hong Kong's Kerry Logistics are both rapidly moving up the rankings.

Major Trends and Future Challenges

Given the markedly heterogeneous nature of air freight logistics, it can be difficult to predict major trends and future challenges. Merkert et al. (2017, p.3) have argued that "there is no such thing as a single, unique, air freight model." However, it appears that at least two major drivers of growth will shape the broader market in the immediate future and these include the ongoing global increase in e-commerce and the cost of jet fuel.

The rapid rise of e-commerce has connected retailers and consumers in different geographic locations across both domestic and international markets, subsequently driving up the demand for faster, more reliable air freight delivery services. The growth in e-freight transactions, as well as information and communication technologies, are likely to lead to additional innovations and efficiencies in the air freight logistics supply chain. Some of these innovations have the potential to be disruptive, including the growth of artificial intelligence programs that can analyze supply chain "big data" and use machine learning to identify the hidden patterns of daily e-commerce. It may also include the implementation of block chain-based global trade digitalization platforms that grant trusted parties exclusive access to data in real-time, although the full potential of this technology for air freight logistics is still open to question.

Jet fuel prices will also substantially impact the demand for air freight transport. If oil prices (and by extension air freight rates) increase, then the demand for air freight logistics may fall or the services will become more expensive. Budd and Ison (2017, p.39) have pointed out how the period of high fuel prices in 2008/9 "had the effect of temporarily leading to a rise in slow shipping and a shift to slower but cheaper land and sea transport." Another round of modal substitution would likely occur if fuel prices rise in the near future.

The heavy reliance on fossil fuels and the related carbon emissions (Debbage and Debbage, 2019) also have the potential to radically reconfigure air freight supply chains. Recent research has suggested that air freight transportation can produce up to 40 times more carbon emissions than maritime transportation (Howitt et al., 2011). Consequently, the ongoing efforts to "decarbonize" the economy, such as placing CO₂ labels on airfreighted produce, may have the potential to substantially alter shipper and consumer behavior, particularly as the so-called "flight shaming" eco-movement gains momentum.

Despite these caveats, similar trends are expected in the near future regarding the continued expansion of air cargo services and infrastructure in East Asia and the Middle East. An elevated demand for express services for certain specialist products, particularly cool-chain pharmaceuticals, cut flowers, electronics and medical diagnostic devices, is also anticipated. Additionally, although air freight logistics has not followed the trend observed in the passenger market, where most carriers have joined one of three large strategic alliances (i.e., Star, OneWorld, or SkyTeam), expect the progression towards consolidation in the air shipment industry to continue unabated. Achieving economies of scale and scope at both airports, and through the value chain, allows shippers and freight forwarders to achieve efficiencies of volume as either belly freight or through full-freighter aircraft (Merkert et al., 2017).

Conclusions

In general, air freight logistics has experienced steady growth over the last 50 years. However, it is a complex form of transportation due to the heterogeneous nature of the industry, which has also contributed to a fairly uncertain future trajectory. Despite various

challenges facing the industry, air freight logistics will likely remain a crucial part of the global supply chain over the long-term, particularly regarding the rapid shipment of high-value, low-weight products. Overall, air freight is a key element of the globalized economy and much more than simply a “by-product” of passenger transport.

Despite its importance, however, this form of transportation remains under-examined from both theoretical and empirical research perspectives. Budd and Ison (2015) make a persuasive case for why there has been a general lack of scholarship focused on air freight logistics. They argue that the relative “invisibility” of air freight in much academic research is largely explained by the sector’s unique and distinctive operating characteristics. “To serve the demanding time-sensitive requirements of global logistics, just-in-time manufacturing and international commerce, air cargo operations typically occur during the hours of darkness, at anti-social hours and at secure inaccessible facilities, bonded warehouses and multimodal logistics centers that are often located at a distance from passenger terminals” (Budd and Ison, 2015, p. 164). The implication is that air freight logistics has tended to operate “in the shadows” with respect to academic research in favor of the more visible, “interesting” and culturally familiar network of air passenger airlines. The goal of this article was to shine more light on this crucially important form of transportation with the hope that the research imbalance regarding air freight logistics will be remedied in the near future.

Future research questions should focus on elevating our understanding of the key decision-making processes that determine the choice of an air freight carrier and airport relative to alternative shipping modes (by road, rail or sea). Additionally, what are the key drivers that shape the air freight logistics supply chain dynamic? How does the relatively high spatial concentration of the supply of air freight services to specific airports and particular carriers contour the geography of demand? In what way will the expanding regulations linked to carbon emissions impact supply and demand in air freight logistics? How will the recent development of the Amazon Air cargo fleet and the experimentation with drone delivery through Amazon Prime Air impact this form of transportation? Although not a complete or comprehensive list, these questions highlight the range and diversity of potential research questions that still remain to be fully answered. It is clear that much more research is needed if we are to more completely understand the dynamics that drive air freight logistics – a crucially important part of the contemporary global transport system.

Biographies



Keith Debbage is a Joint Professor of Geography in the Department of Geography, Environment and Sustainability and the Department of Marketing, Entrepreneurship, Hospitality, and Tourism at the University of North Carolina at Greensboro. His research interests include air transportation, the economic geography of the tourist industry, and urban development.



Neil Debbage is an Assistant Professor in the Department of Political Science and Geography at the University of Texas at San Antonio. His research interests include air transport carbon emissions, environmental sustainability, resilience, and climate change.

References

- Air Cargo World, 2018. The Freight 50: Turbulence Ahead for the Top Cargo Carriers? August, 18–22.
- Air Cargo World, 2019. The Power 25: Top Forwarders Face Turbulent Headwinds, June, 6–20.
- ACI, 2019. Preliminary World Airport Traffic Rankings Released, Airports Council International, March 13th. Available from: <https://aci.aero/news/2019/03/13/preliminary-world-airport-traffic-rankings-released/>.
- Alkaabi, K., Debbage, K.G., 2011. The geography of air freight: connections to U.S. metropolitan economies. *J. Trans. Geogr.* 19, 1517–1529.
- Boeing, 2019. Commercial Market Outlook: 2019–2038. Boeing Commercial Airplanes, Seattle.
- Boonekamp, T., Burghouwt, G., 2017. Measuring connectivity in the air freight industry. *J. Air Trans. Manag.* 61, 81–94.
- Bridger, R., 2015. Aerotropolis Alert! Airport Mega-Projects Driving Environmental Destruction Worldwide. *The Ecologist* 8 May. Available from: <https://theecologist.org/2015/may/08/aerotropolis-alert-airport-mega-projects-driving-environmental-destruction-worldwide>.
- Budd, L., Ison, S., 2015. Air cargo mobilities: past, present and future. In: Birtchell, T., Savitzky, S., Urry, J. (Eds.), *Cargo Mobilities: Moving Materials in a Global Age*. Routledge, New York, pp. 163–179.
- Budd, L., Ison, S., 2017. The role of dedicated freighter aircraft in the provision of global airfreight services. *J. Air Trans. Manag.* 61, 34–40.
- Charles, M.B., Barnes, P., Ryan, N., Clayton, J., 2007. Airport futures: towards a critique of the aerotropolis model. *Futures* 39, 1009–1028.
- Chu, H., 2014. Exploring preference heterogeneity of air freight forwarders in the choices of carriers and routes. *J. Air Trans. Manag.* 37, 45–52.
- Debbage, K., Debbage, N., 2019. Aviation carbon emissions, route choice and tourist destinations: Are non-stop routes a remedy? *Annal. Tourism Res.* 79, j.annals.2019.102765.
- Gardiner, J., Ison, S., Humphreys, L., 2005. Factors influencing cargo airlines' choice of airport: an international survey. *J. Air Trans. Manag.* 11, 393–399.
- Howitt, O.J.A., Carruthers, M.A., Smith, I.J., Rodger, C.J., 2011. Carbon dioxide emissions from international air freight. *Atmosp. Environ.* 45, 7036–7045.
- IATA, 2019. Air Cargo. International Air Transport Association, Montreal. Available from: <https://www.iata.org/whatwedo/cargo/Pages/index.aspx>.
- Kasarda, J.D., 2013. Aerotropolis: business mobility and urban competitiveness in the 21st century. In: Benesch, K. (Ed.), *Cultures and Mobility*. Universitätsverlag, Heidelberg, pp. 9–20.
- Kasarda, J., Lindsay, G., 2011. *Aerotropolis: The way we will live next*. Allen Lane, London.
- Kupfer, F., Meersman, H., Onghena, E., Van de Voorde, E., 2017. The underlying drivers and future development of air cargo. *J. Air Trans. Manag.* 61, 6–14.
- Merkert, R., Van de Voorde, E., de Witt, J., 2017. Making or breaking – key success factors in the air cargo market. *J. Air Trans. Manag.* 61, 1–5.
- Popescu, A., Keskinocak, P., Al Mutawaly, I., 2011. The air cargo industry. In: Hoel, L., Giuliano, G., Meyer, M. (Eds.), *Intermodal Transportation: Moving Freight in a Global Economy*. Eno Foundation for Transportation, pp. 209–237.
- Shepherd, B., Shingal, A., Raj, A., 2016. Value of air cargo: air transport and global value chains. *Internat. Air Trans. Assoc.*, Montreal.
- Walcott, S., Fan, Z., 2017. Comparison of major air freight network hubs in the U.S. and China. *J. Air Trans. Manag.* 61, 64–72.

Air Freight Marketing

Lucy Budd, Stephen Ison, Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	369
Air Freight Defined	369
The Evolution of Airfreight Marketing	370
The Scope of Contemporary Airfreight Services	370
Marketing Airline Airfreight Services	371
Marketing Airport Airfreight Services	371
Marketing Air Freight Services Using Sponsorship and Giveaways	372
Future Challenges	372
References	372

Introduction

The role of marketing in air transport has evolved to extend beyond mere travel promotion to incorporate the entire product, planning, and purchasing process. Airline services are perishable, heterogeneous, and intangible commodities, and many specialists and technologies are involved in their creation. Traditionally the four Ps—of product, place, promotion, and price—have been used in the study of airline passenger services marketing (Waguespack, 2018) but the characteristics of air freight marketing remain relatively underexplored. The purpose of this entry is to identify the principal ways in which air freight services are marketed and discuss possible future trends. The chapter begins by introducing and defining the scope and characteristics of the global airfreight sector before moving on to consider the ways in which both airline and airport airfreight services are marketed.

Air Freight Defined

Between the first dedicated air freight flight, which took off in the United States in November 1910, and today, the conveyance of time sensitive, high value, and perishable commodities by air has developed into a multitrillion dollar sector of the commercial air transport industry. In 2017, 62 million tons of freight with a combined value of US\$6 trillion (representing 35% of international trade by value) was transported around the world by air (ATAG, 2018). The ability to routinely transport large volumes of high value-to-weight and often time critical consumer goods, industrial components, express consignments, medical samples, pharmaceutical products, perishable commodities, and livestock around the world by air has become essential to the functioning of the modern world economy. Yet the sector is volatile, cyclical, and vulnerable to a range of external influences, which suppress demand. These include the state of the global economy, changes to international trade tariffs, the introduction of new technologies, new security interventions, and alterations in patterns of consumer demand. Almost from its inception, marketing has played a crucial role in stimulating and sustaining demand for air freight services. However, the particular form and function of air freight provision means that both the message and the media that are used to market freight services exhibit distinctive characteristics, some of which differ significantly from conventional air passenger services marketing.

Air freight describes the transport of goods and commodities (other than mail and passengers' luggage) by air. It is distinct from "air cargo," which comprises both the carriage of freight (excluding passengers' luggage) and mail (which includes letters, documents and parcels carried under the terms of a national postal convention or postal regulation). Depending on the nature of the consignment and the speed with which it needs to be delivered, air freight can be classified as being express (urgent), general (less urgent or routine), outsize (usually heavy and/or bulky loads that cannot fit in the holds of conventional passenger aircraft), specialist (such as temperature sensitive pharmaceutical products and live animals), or humanitarian freight (such as tents, food and water purification equipment that is required in the aftermath of an emergency).

Unlike passengers, who can only travel in dedicated aircraft specifically designed for the purpose of human conveyance, air freight can be transported in the holds of regular (often scheduled) passenger flights as "bellyhold" freight or as "pure freight" in dedicated freight-only aircraft that have been specially constructed, converted, and/or chartered for this purpose (Budd and Ison, 2017). Air freight can represent an important additional revenue stream for passenger airlines. Indeed, many of the world's major carriers accept freight on their passenger services, either as a by-product of their passenger provision or as a joint product in which both passengers and freight are equally important, and some operate dedicated freight-only aircraft alongside their passenger fleet. Alongside the passenger airlines are freight only specialists who operate dedicated freight aircraft. The final type of operation are integrators, such as DHL, FedEx, and UPS, who specialize in end-to-end journeys using dedicated fleets of road vehicles and aircraft.

In addition to the type of aircraft and nature of the operation, a key difference between passenger services and freight is that passenger demand is bidirectional and, generally speaking, equal numbers of passengers want to travel from A to B as from B to A. The same is not true of freight, which only moves from its point of production to its point of consumption. The unidirectional nature of freight demand (which is often referred to as the “back-haul” problem) has important implications both for airline operations and practices of marketing, as the return sectors are often not commercially viable. In response, air freight operators may perform triangular or multileg global itineraries to maximize their returns and minimize low loads. This schedule offers commercial advantages to the airline but it may also increase total trip duration and lower the number of direct point-to-point flights that are performed.

Owing to the diverse and demanding nature of the global market it serves, the provision of air freight services is complex and dynamic and the array of activities that are performed represent a challenging prospect for marketing, not least because the behavioral decisions of customers are based not on personal subjective factors, such as in-flight cabin comfort and customer service, but on price, punctuality, reliability (also known as “flown as booked” performance), flight frequency, and efficiency.

The Evolution of Airfreight Marketing

In the early decades of commercial flight in the 1920s and 1930s, high value freight, including diplomatic bags, bullion, and personal effects, was routinely transported on aircraft, but the volumes were small and air freight only accounted for a fraction of the world’s trade by volume. The introduction of larger passenger aircraft from the mid-1940s onwards not only transformed the range (i.e., nonstop distance) that could be flown, but also increased the airlines’ capacity for transporting freight. This presented a challenge. The airlines had newfound supply (in terms of the freight carrying capacity of the new aircraft), but there was relatively little demand, as airfreight services were often considered prohibitively expensive when compared to conventional forms of surface transport and general awareness of the availability of such services among business and the general public was low. The airlines’ marketing challenge was thus to raise awareness of the existence of air freight services and stimulate demand for them by emphasizing the commercial advantages of shipping long-distance freight by air rather than by road, rail, or sea.

One of the principal benefits air travel afforded was its superior speed and its ability to deliver freight consignments over medium and long distances far faster than was previously possible. The early air freight marketing campaigns that were run by the airlines thus emphasized the time savings that could be achieved by air and related this to first mover advantage by stressing that companies who used air freight could reach their consumers quicker than their competitors. There was, however, the issue of cost and door-to-door delivery. It was undeniable that air freight services were more expensive, mile-per-mile, than conventional modes. The airlines thus had to convince potential customers that the additional cost represented good value for money as air freighting enabled them to deliver their products to market sooner. This was deemed particularly attractive for producers who needed to ship time critical and/or perishable commodities. The challenge of providing door-to-door pick-up and delivery was quickly addressed through the provision of integrated courier services who would transport goods from their point of origin to the departure airport and then, on landing, to their ultimate delivery address.

The ability to transport freight around the world by air in this way has both been a driver of, and in turn been driven by, changes in global industrial strategy, the rise of overseas manufacturing facilities and practices of outsourcing and just in time production techniques that demand the rapid and routine international movement of goods and consumer products.

The Scope of Contemporary Airfreight Services

Today, access to safe, secure, and reliable airfreight underpins the international supply chains of many of the world’s largest companies operating in the biotechnology, pharmaceutical, aerospace, electronics, fashion, financial services, and entertainment, sports, livestock, humanitarian, and food production sectors. Much of this activity requires the time critical shipment of precious and/or perishable commodities across international borders and may require specific handling, tilt protection, and/or cool chain protection at all stages of the journey.

The types of commodities that are transported by air are many and varied and range from the routine transport of hazardous materials for the chemical industries to live vaccines and drugs for the pharmaceutical and medical sectors, and from live crustaceans and shellfish to racing cars, consumer electronics, and fast fashion items. Every consignment has particular needs and places exacting demands on air freight providers. The transshipment of valuable commodities, including works of art, for example, requires customized packing, handling, and security. The movement of live animals requires specialist handling and in-flight support to ensure their safe arrival. Vaccines and perishable food require temperature-controlled conditions to prevent them from being spoiled. In every case, the consignments need to be accompanied by accurate and valid airway bills and documentation. Yet in addition to routine shipments, air freight providers must also have the capacity and flexibility to respond to emerging contingencies and react to supply chain disruption that follows service interruptions in other transport modes. In order to examine the nature and complexities of air freight marketing, the fifth and sixth sections detail the service attributes and media that are used by airlines and then airports to market air freight services.

Marketing Airline Airfreight Services

It is apparent from the previous sections that air freight services exhibit distinctive characteristics compared to passenger services. One of the key features that differentiates air freight marketing from conventional air passenger services marketing is that air freight companies can only market their services to customers (i.e., people who take decisions about how freight travels) and not consumers (because they are not actually traveling on the flight that has been purchased). This has important implications for marketing practice, not least since there is no potential for consumers to generate or cocreate marketing content on blogs, review sites, and social media platforms. As a result, air freight marketing relies on a particular set of attributes which manifest themselves in the marketing copy employed by airlines.

One of the ways in which air freight marketing activities seek to create awareness of their products among the target market, influence purchasing behavior, and engender brand loyalty is through printed advertisements which are published in the specialist trade press. One of the challenges of using this form of print media is that the advertisements are seen in close proximity to competitor offerings and so the message must be clear and carry a corresponding appeal. [Budd and Ison's \(2015\)](#) content analysis of the air freight advertisements published in dedicated air freight trade publications revealed that air freight is marketed using five key attributes.

The first of these attributes is connectivity and seamless global flight with advertisements typically emphasizing the geographic extent of an airline's freight network. Such advertisements often use rhetorical devices, including images of the earth from space or abstract representations of the globe, to convey notions of a world that is seamlessly interconnected by freight services. Such images are often accompanied by text, which describes the number of cities, counties, and continents the airline (and, if relevant, its alliance partners) serve. An advertisement in 2018 for Emirates SkyCargo promotes the carrier's "intricate global network of over 250 seamless cargo connections," which "enables us to link your business to the world," while [Air Canada Cargo \(2019\)](#) suggested that, with them, the "entire world is within reach" courtesy of the 450 destinations on six continents that they serve. In some instances, advertisements focus on the connectivity that is provided by the airline's principal operating hub(s) and the speed with which shipments can be dispatched to multiple destinations. This focus on geographic connectivity has clear parallels with the marketing copy that is employed by passenger airlines, although the airports that are served by freight may be different from those used by passenger services.

In addition to promoting notions of seamless global connectivity, a second theme that is evident in air freight advertisements relates to the capacity and capability of the airline in question, as well as its ability to customize its service delivery to meet specific customer needs. In this context, capacity can be described both in terms of the payload weight or volumetric capacity (expressed in tons or number of unit load devices (ULDs) (ULD is a metal container that is loaded into the main deck or hold of an aircraft that contains multiple small freight consignments.) that can be accommodated) and also the airline's ability to accept irregular or outsized shipments. Images of railway locomotives, racing cars, endangered animals, and stage sets for major entertainment events may be depicted to stand proxy for the range and diversity of consignments that the airline can convey. For example, in an advertisement in 2019 announcing the inauguration of two new Boeing 777F aircraft into service, Turkish Airlines Cargo division stated that "Now, you can reach to more countries [sic] than any other airlines in the world with our long haul 2 new Boeing 777F planes with 102 tons capacity," while [SWISS World Cargo \(2019\)](#) promoted the fact that "This year we delivered a world-famous painting custom packaged in airtight conditions from Shanghai to Zurich and then onwards to a museum in Bern"

The third attribute is convenience, as it pertains both to freight consigners and end users. This attribute is typically expressed in terms of the frequency and intensity of freight flights, but also the possibility of employing door-to-door pick-up and delivery using fleets of dedicated road vehicles. "With our frequent flights and logistics know-how we've got your business" back ([Southwest Cargo, 2017](#)). Convenience also extends to the ability to track consignments in real time using a dedicated website or app to provide reassurance that the freight is being flown as booked: "Around the World in One Night. From Premium capacity allocation and service-level guarantee to GPS tracking, 24/7 monitoring and clearly visible pink wrapping, your parcels are delivered with care" ([Delta Cargo, 2018](#)).

In addition to promoting the connectivity, capacity, and convenience of air freight, a fourth set of advertisements emphasize the *care* that will be afforded to a customer's products, particularly those that demand specialist handling, enhanced security, or cool chain capabilities. Accompanying a photograph of a hand holding a delicate porcelain vase, [Qatar Airways Cargo \(2019\)](#) promise that "Whatever your cargo, you can trust us to carry it with as much care as you do" while [Thai Cargo \(2019\)](#) employed an image of a fresh fish and the caption "Keep it Ocean Fresh Across Continents. [Thai] employs an end-to-end cooling system to ensure freshness for temperature sensitive goods and perfect delivery to recipients at every destination." Arguably among the most demanding freight consignments are those for the medical and pharmaceutical industries. [Alaska Air Cargo \(2019\)](#) promoted the fact that "We ship medical items, including lab samples, supplies and live organs overnight to more than 100 destinations in North America."

Unlike the majority of passenger services which tend to operate during the day, air freight services are often scheduled to meet the demanding next-day delivery requirements of consumers and just in time manufacturing systems. The final attribute identified in the advertisements is speed and 24-hour service that stresses that, irrespective of the local time, the company will keep the freight moving. In every case, the aim is to convey the notion that the company is capable, professional, trustworthy, and reliable.

Marketing Airport Airfreight Services

In a highly competitive marketplace, it is not only the airlines that have to market their services. Airports too must compete for custom and the marketing they undertake aims to attract new airlines and routes by stressing the competitive advantages, both to

potential airline operators and air freight forwarders, of conducting business at the facility (Halpern, 2018). Freight airlines and freight forwarders have particular requirements with respect to the provision of ground handling services, the speed and availability of customs checks and border inspection posts, specialist freight handling, storage and processing equipment, and the availability of warehousing space adjacent to the airfield. The marketing of air freight services by airports usually seeks to provide information about existing capabilities and/or investment or expansion potential. Airports may utilize a range of media to market their services. This includes, but is not limited to, corporate event sponsorship, exhibiting, conference attendance, giveaways, and targeted print advertising. As with the air freight marketing copy that is produced by airlines, five key attributes can be identified in the printed material that is produced by airports.

The first of these is capability, with the advertisements stressing that the airport can handle a wide range of goods, from temperature controlled warehouses to e-commerce delivery. Accompanying a montage of images which included a panda, a helicopter, a bunch of fresh tulips, a framed painting, and a piece of fresh cheese, Changi Airport Group in Singapore asked "What makes us one of the world's most trusted cargo hubs: our capabilities."

The second attribute is capacity. In these adverts, the airport operator may seek to reassure potential airlines and freight forwarders that the facility can meet both current and future needs in terms of runway capacity, operating hours, stand space, warehousing, and surface access. A 2019 advert for Miami International Airport, for example, promoted the fact that the facility had no slots, curfews or delays that could inhibit the movement of freight.

The third theme that is evident in the printed advertisements is convenience, both in terms of proximity to major cities and local markets and also easy and reliable connections to the local and national road (and rail) network. The co-location of other freight firms on and around the airport was the fourth marketing point as the co-location of specialist freight handling providers and a local critical mass of expertise appears to be important in attracting new custom.

The final attribute concerns control or the security of goods at the airport. Such advertisements may use images of CCTV cameras and remote monitoring devices to emphasize the security protocols that are enforced and/or the range of secure landside and airside warehousing that is available to potential customers. These attributes are, of course, not mutually exclusive and advertisements may feature two or more of these themes.

Marketing Air Freight Services Using Sponsorship and Giveaways

A further way in which air freight services are marketed by both airlines and airports is through giveaways and corporate sponsorship. Giveaway items can be useful in promoting goodwill and raise customer awareness but they need to be carefully considered and controlled to reinforce the company brand. Ideally, they would be robust practical or fun items. Examples of giveaways include branded pens, lanyards, key fobs, rulers, USB sticks, memo blocks, and sticky notes or items such as branded model aircraft. In every case, the giveaway seeks to raise awareness of the brand in question and reinforce the message that the services that it supplies are useful, dependable, and reliable. There is also marketing value from supplying specialist services at or below cost price, particularly if this supports the company's CSR (corporate social responsibility) agenda. Examples of such practice might include transporting captive bred or rescued endangered animals to release sites in the wild and engaging in postactivity marketing to demonstrate the care, capacity, capability, and compassion that has been shown. Airlines and airports may also use event and/or venue sponsorship or naming rights to raise awareness of their brand and convey their corporate values. Typically, this may involve partnering with a particular charity or nongovernmental organization, sponsoring an event, conference or symposium, or paying for naming rights for sports stadia or major cultural events.

Future Challenges

Although it is difficult to speculate on the future for airfreight marketing, it is reasonable to assume that recent key trends will continue into the short to medium term. In a competitive operating environment, airlines and airports will continue to need to effectively market their services to existing and potential future customers. The key issues facing the sector include: overcapacity and low yields; the impact of volatile fuel prices; fluctuations in global economic activity; the introduction of more stringent environmental regulations and moves to "decarbonize" the economy, changing consumer priorities and the introduction of new aircraft technologies. The key challenges for air freight marketing are thus being able to anticipate, identify and meet the changing needs of shippers and consumers, while deriving a profit from the international transport of freight by air.

References

- Air Canada Cargo, 2019. <https://www.aircanada.com/cargo/en/> (accessed 24.11.19).
 Alaska Air Cargo, 2019. <https://www.alaskaair.com/content/cargo/cargo-home> (accessed 24.11.19).
 ATAG, 2018. Aviation benefits beyond borders. Air Transport Action Group, Geneva.
 Budd, L., Ison, S., 2015. Air cargo mobilities, past, present and future. In: Birtchnell, T., Savitzky, S., Urry, J. (Eds.), *Cargo Mobilities Moving Materials in a Global Age*. Routledge, London, pp. 163–179.

- Budd, L., Ison, S., 2017. The role of dedicated freighter aircraft in the provision of global airfreight services. *Journal of Air Transport Management* 61, 34–40.
- Delta Cargo, 2018. <https://www.deltacargo.com/Cargo/> (accessed 24.11.19).
- Halpern, N., 2018. Airport marketing. In: Halpern, N., Graham, A. (Eds.), *The Routledge Companion to Air Transport Management*. Routledge, Abingdon, pp. 220–237.
- Qatar Airways Cargo, 2019. <http://www.qrcargo.com/> (accessed 24.11.19).
- Southwest Cargo, 2017. <https://www.swacargo.com/> (accessed 24.11.19).
- Swiss World Cargo, 2019. <https://www.swissworldcargo.com/en/home> (accessed 24.11.19).
- Thai Cargo, 2019. <https://www.thaicargo.com/en/main> (accessed 24.11.19).
- Waguespack, B., 2018. Airline marketing. In: Halpern, N., Graham, A. (Eds.), *The Routledge Companion to Air Transport Management*. Routledge, Abingdon, pp. 206–219.

Drones in Freight Transport

Oliver Kunze, HNU Neu-Ulm University of Applied Sciences, Neu-Ulm, Germany

© 2019 Elsevier Ltd. All rights reserved.

Glossary	374
Introduction	374
Classification of Drones by Mode and Size	374
Logistics Use Cases	375
Loading and Unloading Infrastructure	376
Reach and Bimodal Approaches	376
Regulations and their Commercial Implications	376
Selected Safety and Security Issues	378
Noise Emissions	379
Critical Discussion of Air Versus Ground Drones	379
Open Issues	380
See Also	380
References	381
Further Reading	381

Glossary

AGV automated guided vehicle
ATC air traffic control
ATM air traffic management
BVLOS beyond visual line of sight
CEP courier—express—parcel
RPAS remotely piloted aircraft system
UAV unmanned aerial vehicle
VLOS visual line of sight

Introduction

“The hype over commercial drones is, so far, largely just that. [. . .] widespread deliveries by drone are not yet a reality, neither by Amazon nor by any other company. Regulatory thickets, technical complexity and the public’s skittishness have proven to be formidable hurdles. [. . .] But that doesn’t mean there’s likely to be a drone-free future” (Rosen, 2019). This assessment describes the perceived situation of drone use in logistics in early 2019.

A lot of speculation about the potential of drone use in freight transport is published by different stakeholders. While public media mostly speculates about drone use from a qualitative point of view, some consultancy companies even attempt quantitative forecasts. Pricewaterhouse Coopers LLP for instance stated in 2018 that by 2030 “cost reductions and efficiency improvements from drone usage will feed into an increase of 3.2% in ‘multi factor’ productivity across the UK economy, and contribute to considerable GDP uplifts in many industries . . . ” which include “ . . . £ 7.7 bn in wholesale, retail trade and food services . . . ” (PwC, 2018, p. 3). They also estimate that 11,008 drones will be used in the “transport & logistics” sector in the UK skies by 2030 (PwC, 2018, p. 16).

Shippers as, for example, Amazon or Walmart (FAZ, 2019), as well as transport companies like DHL or UPS have been and still are technically testing air and ground drone deliveries. Still, only a few—Amazon probably the most prominent of them—have actually announced drone delivery services.

Classification of Drones by Mode and Size

To discuss drones and their developing role in freight transport, it is necessary to first roughly categorize the multitude of currently developing drone types. From a logistics point of view, they can roughly be categorized by transport mode (air vs. ground) and by size (small, medium, and large). Table 1 shows a rough overview of typical representatives of this scheme.

Table 1 Types of freight drones

Mode→ Size ↓	Air drones or unmanned aerial vehicles (UAVs)	Ground drones or automated guided vehicles (AGVs)
Small (up to parcel size)	(a1) Small UAV  For example, by EmQopter GmbH (www.qopter.de)	(b1) Small AGV  For example, by Starship Technologies LTD (www.starship.xyz/kit/)
Medium (up to pallet size)	(a2) Medium UAV  For example, by DLR (Deutsches Zentrum für Luft- und Raumfahrt e. V. DLR)	(b2) Medium AGV  For example, by SSI Schäfer (SSI SCHÄFER)
Large (pallet size plus)	Unmanned airplanes	Unmanned trucks

This scheme is quite rough with respect to the size scale—still, one can assume that a *large* drone can transport items exceeding Euro-pallet size, a *medium* drone items up to Euro-pallet size, and *small* drone items up to parcel size. Small air drones can be subcategorized, for example, in accordance with the upcoming EU-drone-flight regulations which most probably will differentiate the total weight categories: –250 g/–900 kg/–4 kg/–25 kg. This article focuses on small and medium drones.

Logistics Use Cases

In general, one can assume that air drones are an interesting option in DDDD (dull, dirty, distant, and/or dangerous) transport environments.

In *intra-logistics*, ground drones (small, medium, and large) have successfully been used for decades, where they are known under the name of autonomous ground vehicles (AGVs). Whereas ground drones usually are the dominant technology in intra-logistics, air drones have started to be seen in niches—for example, if facilities are separated by public roads which need to be crossed.

In *extra-logistics* there are at least three distinct use-case clusters: emergency logistics, rural logistics, and urban logistics. It is difficult to predict how successfully drones will be used in these clusters, as not only technical aspects are to be considered, but political will (with respect to noise emissions, air traffic regulations, energy consumption, etc.) on the one side and lobbyism on the other side will also be important decisive factors. Focusing on technical top-level aspects (existence of infrastructure, energy consumption, and noise emissions) one can derive the following conclusions:

1. In *emergency logistics*, where road transport infrastructure is damaged (e.g., after flooding or after earthquakes) the use of air drones shows great potential, to get the goods to the needy in due time.
2. In *rural logistics*
 - a. the lack of dense road infrastructure together with

- b. the absence of dense population (which will oppose air-drone noise emissions), are two facts which support the speculation that the use of air drones is more promising than the use of ground drones. Here, the possibility of air drones to take shortcuts (e.g., to cross a river or a mountain range) can also compensate for their higher energy consumption.
- 3. In *urban logistics* the use of ground drones seems more promising than the use of air drones, because
 - a. sufficient road infrastructure is available
 - b. air drones need more energy than ground drones (because they need energy to be kept airborne throughout their journey), and
 - c. one cannot shield residents from noise from earlier.

For urban as well as for rural commercial freight transport, both ground and air drones currently still have to face the lack of payload standardization. Today, this lack of standardization leads to a multitude of incompatible infrastructures for loading and unloading of the drones. One may speculate that once the payload dimensions become standardized (as, e.g., today in the industry-standards within the KEP business) and a standardized infrastructure for loading and unloading of drones is established, the use of drones for freight transport may get a boost.

Another factor that promotes the use of drones is their potential to substitute for human drivers and/or couriers. Thus, an increasing use of drones may help to reduce costs for last mile transport while simultaneously low-income transport jobs in last mile delivery face the threat of being substituted by drones.

As of today, *drone use cases* are emerging in CEP-logistics (see, e.g., Amazon Prime Air, DHL, UPS, etc.), food home delivery (see, e.g., Domino's, UberEats, Wing, etc.), and in medical and disaster logistics (see, e.g., ApoAir, QuiQui, Wingcopter, Zipline, etc.).

Loading and Unloading Infrastructure

As all drones for freight transport need to be loaded and unloaded, the relevant loading and unloading infrastructure is relevant for the usability assessment of a drone. Whereas manual loading and unloading is almost always an option, this also is the most cumbersome, slowest and most expensive way to handle cargo. Thus, automated or semiautomated loading and unloading are important for the successful use of drones in transport logistics.

The loading and unloading infrastructure can either depend on handling devices mounted on the drone or on stationary devices at each loading and unloading point. The main options for loading and unloading of small drones are summarized in **Table 2**.

This multitude of handling options calls for standardization efforts. (Note that the extreme success of worldwide ISO-container transport was and is based on container size standardization and container handling standardization by means of corner castings and twist locks on spreaders.) As long as this standardization is not achieved for drones, this lack of standardization will reduce handling efficiency as well as the commercial potential to use freight drones.

Reach and Bimodal Approaches

Even though drones can be operated by combustion engines too, this article shall focus on electrically propelled drones, because on the one hand electric propulsion offers the option to use energy generated by solar, wind, or hydro power, and thus does not necessarily add to global CO₂ emissions, and on the other hand it usually produces less noise.

The reach of electrically propelled ground and air drones is limited by their battery capacities. However, the limited reach of ground and air drones becomes a less important factor, if they are used in a multimodal conceptual setting. In such a setting, the main transport leg is conducted by larger vehicles (e.g., vans), whereas only the last transport leg is carried out by means of a drone (**Table 3a1** and **b1**).

An alternative option to expand the reach of air drones is the concept of hybrid UAVs that have wings as well as copter engines (**Table 3a2**). Hybrid UAVs can operate in a copter or hover mode for takeoff, briefly go through a transition mode into a cruise mode for most of the mission, and return to copter mode for landing. In cruise mode the wings help to carry the drone weight, and thus can cover longer distances than sheer copter drones with identical batteries.

An alternative option to expand the reach of ground drones is the concept to co-use public transport infrastructure (**Table 3b2**). Especially at night, when trams and subways usually do not operate, they could be used to transport ground drones instead of passengers, and bring them to their last mile operation area without significant exhaustion of the drone batteries. The only requirement for this infrastructure co-use is a level access to the trains and the technical ability of the ground drones to use elevators and/or escalators.

Table 2 Selected types of drone loading and unloading devices

Technology	Active stationary devices			Passive drone mounted device	
(1) Horizontal handling	(1a) Active rolls, (1b) pushers, (1c) tilting ramp			Loading bay	
(2) 3-dimensional handling	(2a) Robot arm, (2b) spreader and twist locks			Loading bay	
(3) Manual handling	(3) Humans			Loading bay	
(1a)	(1b)	(1c)	(2a)	(2b)	Legend
Technology	Passive stationary device			Active drone mounted devices	
(I) Horizontal handling	Loading dock			(Ia) Active rolls, (Ib) pushers, (Ic) tilting floors	
(II) 3-dimensional handling	Loading dock			(IIa) Robot arm, (IIb) spreader and twist locks, (IIc) trap door	
(Ia)	(Ib)	(Ic)	(IIa)	(IIb)	(IIc)

Regulations and their Commercial Implications

The use of air drones in freight transport is subject to new regulations, which are currently developing. Worldwide regulations are significantly inhomogeneous. In the EU, a master plan for air traffic management (ATM) has been published and is being updated. Within this process, a “roadmap for the safe integration of drones into all classes of airspace” (Guillermet, 2018) has been released by the project group SESAR. A new drone regulation is scheduled to be issued in 2019. The key ideas of this regulation to come are:

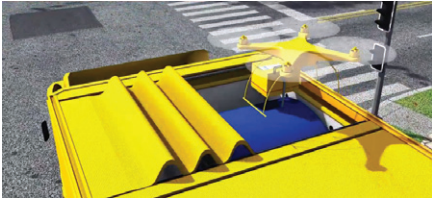
- Distinction of different risk categories C0-C4 which are mostly dependent on total drone weight (–250 g, –900 g, –4 kg, –25 kg), but also include other factors such as speed.
- Defined non-flight zones (especially over airports, industrial plants, public authorities, private ground, nature reserve areas, human crowds, and over emergency response efforts).

Depending on the EU drone risk category, different regulations apply with respect to the minimal distance to people during operations, drone identification, certified drone operator skills, maximal altitude, and the need for a certified authorization of use.

In the United States, the use of air drones for commercial reasons is regulated by the Small Unmanned Aircraft Rule (Part 107 of FAA regulations) (FAA, 2018), which also includes rules for maximum weight, maximum speed, maximum altitude, drone identification, operator skills, etc.

However, there are other aspects of the different worldwide regulations that have a significant influence on the commercial use of air drones, for example:

Table 3 Options to expand the reach of drones

Air drones	Ground drones
(a1) Vehicle-based air drone  Vehicle-based air drone concept of Dropney: truck with onboard rotating parcel magazines and drone magazine (http://dropney.com/showreel/)	(b1) Vehicle-based ground drones  Mothership concept of Daimler and Starship: van with goods magazine and ground drones (https://www.daimler.com/innovation/specials/future-transportation-vans/paketbote-2-0.html)
(a2) Hybrid drone with cruise and hover mode  Use of a hybrid drone with cruise and hover mode for transportation of medication (here in Vanuatu) (www.wingcopter.com)	(b2) Co-use of public transport  First ideas to co-use public transport for logistics purposes are explored. This generic concept can be combined with ground drones (www.logistiktram.de)

- Visual line-of-sight (VLOS)—that is, the operator needs to be able to directly see the air drone during operation. Where this rule applies, it prohibits remote air drone operations, as well as autonomous flight modes outside the vision range of the operator on site.
- Beyond visual line-of-sight (BVLOS)—that is, the operator does not have to be able to directly see the air drone during operation.
- 1:1 Operator control—that is, no person may act as a remote pilot in command for more than one unmanned aircraft operation at one time. Where this rule applies, each air drone needs a dedicated drone operator, and this significantly increases the costs of operating drones.
- Daylight operations only—that is, the drones must not operate in the dark. This regulation limits the use of air drones significantly, as air drone logistics service providers will either have to shut down their services at nightfall, or they need a non-air-drone-based fallback delivery concept.
- Must yield right of way to other aircraft—this generic regulation implies that there are rules for “right of way” and rules on how to yield correctly—these rules must be known to the drone operator and they must be supported by the relevant autonomous flight modes.

From today’s point of view, it seems likely that air drone regulations will undergo several adjustments in the short- and medium-term future, as lobbyists will ask lawmakers to lower restrictions, whereas voters may ask for additional restrictions for noise, privacy, and damage protection once air drones become more widely used.

Selected Safety and Security Issues

Safety and security are essential for the commercial use of drones in freight transport. Whereas safety issues focus on potential harm to humans (i.e., recipients as well as third party bystanders) in this context, security issues focus on potential harm to the transported goods. Safety and security issues can also be categorized by “on route” (while moving) and “on site” (during loading and unloading).

Table 4 Selected safety and security issues

Mode→	Air drones	Ground drones
Safety on route	⚡ <i>Crash</i> due to collision with moving objects, stationary objects, weather conditions, or due to energy exhaust ⇒ Regulations for flight over people ⇒ Air-drone parachute 🚨 ●●● High	⚡ <i>Stop</i> due to collision with moving objects, stationary objects, weather conditions, or due to energy exhaust ⇒ General traffic regulations ⇒ Active signaling (flashing warning lights) 🚨 ●○○ Low
Safety on site	⚡ <i>Rotating rotors</i> hurting persons ⇒ Caged rotors ⇒ Presence detectors ⇒ Cargo drop 🚨 ●●● High	(No similar issues due to small weight and slow speed of ground drone) 🚨 ○○○ None
Security on route	⚡ <i>Air piracy</i> by catcher drones or by redirecting software hacks ⇒ Permanent monitoring 🚨 ●○○ Low	⚡ <i>Road piracy</i> by manual theft or by redirecting software hacks ⇒ Permanent monitoring 🚨 ●○○ Low
Security on site	⚡ <i>Theft</i> ⇒ Identification check prior to unlocking 🚨 ●○○ Low	⚡ <i>Theft</i> ⇒ Identification check prior to unlocking 🚨 ●○○ Low

⚡: Hazard; ⇒, countermeasure; 🚨○○○, potential commercial damage in event.

These issues as well as possible countermeasures and a rough estimate of their potential commercial damage potential are shown in **Table 4**.

The qualitative assessment of potential commercial damage in the safety or security event in **Table 4** is based on the following ideas. First, if humans are harmed, this usually induces significant legal settlement costs, especially compensation payments for the victims. If transported goods are harmed, the *settlement cost dimensions* are smaller by magnitudes compared to harmed humans. Second, usually incidents where humans are harmed cause a significant *reputation problem*, whereas incidents where goods are damaged or lost often can be settled without great public attention. Third, ground drone transport problems can be considered to be *known problems* as there exists decades of experience from other ground transport executions. However, air drone transport problems are rather unknown, because air drone traffic is new and not yet widely used. Fourth, the magnitudes of piracy and theft issues are similar for air and ground drones.

Noise Emissions

Especially in urban logistics, noise emissions play a big role. Whereas huge sums of money are invested to shield citizens from road-and/or rail-noise emissions, there simply is no technical way to protect citizens outdoors (e.g., in their garden or on their balcony, or citizens who want to enjoy fresh air through an open window) from air noise. While some flats facing busy streets are already overexposed to street noise, often other flats in the same building provide the noise-wise rest and peace of an enclosed atrium. Thus, there are three options:

- Low-altitude corridors: Should air drones be limited to flight corridors over existing streets below average roof height, where they would not add to the noise pollution of currently non-noisy living quarters, but they would significantly add to the noise exposure of higher floor flats. At the same time, within this corridor option air drones would lose their potential to use beelines as route trajectories, and thus save on distance traveled.
- High altitude corridors: Should air drones be limited to flight corridors over existing streets above average roof height, where they would add to noise pollution of currently non-noisy living quarters—the higher they fly, the higher the noise emission radius (and the lower the emitted noise level per unit). In this option the beeline advantage would also be lost.
- No corridors: Should air drones be permitted to fly all possible beelines without corridors, such that their noise emissions would potentially cover all areas. In this case, the highest possible flight altitude (without infringement of current air traffic) would minimize noise emissions for citizens, whilst creating a permanent background hum noise from earlier (if many drones are airborne simultaneously).

Within all three options, air drones would create local noise during takeoff and landing, which freight recipients might accept, but neighbors may not. Ground drones on the other hand do not add to the noise problem, as on the one hand they produce almost no noise (compared to air drones) and on the other hand they share the infrastructure of today's noise-emitting traffic.

Critical Discussion of Air Versus Ground Drones

Both drone types share the potential that (if operated autonomously) they can operate *without a driver*, who is a significant cost factor in last mile logistics. Both lose this potential if regulations require a 1:1 full-time operator-drone surveillance. It is likely that ground drones will be exempt from this 1:1 regulation because they could “stop and wait” in critical situations (i.e., the default security mode in ground traffic) until a 1:many operator can take care of the situation. To achieve an exemption from 1:1 full-time operator surveillance for air drones seems more difficult from today’s point of view.

The one advantage of air drones is their potential average *cruise speed*. Whereas ground drones would most probably operate at pedestrian speed on sidewalks, at cyclist speed in bicycle lanes and at average traffic speed on roads, air drones could operate at higher speeds. This advantage is reduced, if, due to reach expansion requirements, a bimodal transport (e.g., truck and air drone, or subway and ground drone) is required.

The other advantage of air drones is their independence from road infrastructure, which renders air drones especially valuable for emergency freight transport.

For most other aspects, such as energy costs, noise emissions, safety, and sustainability, ground drones seem the superior option to execute the last mile freight transport—especially in urban environments.

Open Issues

There are a number of open issues that need further research and further experience from first operations. These issues are the following:

- Risk assessments are needed for different use cases. As a first step, these risk assessments will need to be based on assumptions and test cases, but they will need to undergo revisions as soon as significant experience from real-life operations is available on a larger scale.
- Drone traffic regulations are currently being developed by lawmakers. Similar to the risk assessments, this lawmaking process will constitute first regulations and later modified regulations based on the feedback from different stakeholders (shippers, drone operation service providers, customers, citizens, and lobbyists). One interesting aspect here is the willingness to open up existing public transport infrastructure for co-use by ground drones.
- Sustainability assessments for a wider use of drones are needed soon. Whereas technical innovations may reduce a few percent of energy consumption rates or noise emission rates, they do not change the big picture. Is it wise to propagate air drones for urban convenience services while humankind faces global warming and mass extinction of species? Therefore, further sustainability research is needed as well as political will to take actions.
- An interesting and, most probably, the decisive question in assessing the future of commercial freight drone operations is: will freight drone business models be profitable? This question is strongly dependent on the relevant use case, as well as on a multitude of cost factors (required investment, HR costs for operators, energy costs, etc.) and on regulations, which may influence the cost factors, and prohibit or allow business cases.

See Also

On urban logistics: The Future of Urban Freight; The Economics of Urban Freight; City Logistics

On political pricing: Value of Time for Freight; Value of Noise; The Value of Life and Health; Pricing Versus Regulation

On air traffic: Introduction to Air Transport; Airspace Management; Air Traffic Control; Air Traffic Flow and Capacity Management; Air Vehicles Classification; The Future of Air Transport

On autonomous vehicles: Autonomous Vehicles

On specific use cases: Humanitarian Logistics

References

- FAA (Federal Aviation Agency), 2018. Small Unmanned Aircraft Rule (Part 107 of FAA regulations). United States Department of Transportation, Federal Aviation Agency, Washington, DC.
- FAZ (Frankfurter Allgemeine Zeitung), 2019. Eine Kühlbox auf Rädern. Frankfurter Allgemeine Zeitung GmbH. January 25, 2019, Wirtschaft, Frankfurt, p. 21.
- Guillemet, F., 2018. European ATM Master Plan: roadmap for the safe integration of drones into all classes of airspace. SESAR—Single European Sky ATM Research Project, Brussels.
- PwC, 2018. Skies without limits drones—taking the UK’s economy to new heights. Pricewaterhouse Coopers LLP. Available from: www.pwc.co.uk/dronesreport.
- Rosen, E., 2019. Skies aren’t clogged with drones, yet, but don’t rule them out. The New York Times, New York, March 19, 2019.

Further Reading

Kunze, O., 2016. Replicators, ground drones and crowd logistics—a vision of urban logistics in the year 2030. *Transportation Research Procedia*, Elsevier, Amsterdam.

Duty of Care in the Selection of Motor Carriers

Thomas M. Corsi, Robert H. Smith School of Business, University of Maryland, College Park, MD, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	382
Dynamic Structure of the Trucking Industry in the United States	382
FMCSA Motor Carrier Safety Performance Data Initiatives	383
Assessing “Reasonableness” in the Carrier Selection Process	384
Defining Carriers as Independent Contractors or Agents; Assessing Negligence in Carrier Hiring	384
The Legal Battle for Assessing Carriers as Independent Contractors and Determining Negligence in Hiring	385
Shippers, Brokers, and Third-Party Logistics Providers Seek Relief From Reasonable Care Obligations	385
Safety Implications of Minimal Shipper, Broker, and Third-Party Logistics Provider Attention to Safety Performance	
Matters in Carrier Selection	386
Conclusion	386
References	387

Introduction

Transporting freight in large trucks is an inherently dangerous activity with the potential risk of a fatal or injury crash causing significant costs as a result of deaths, injuries, property damage, and cargo losses. In fact, throughout the period 2010–17, there has been a significant increase in the United States in the rate of large truck involvement in fatal crashes per 100 million vehicle miles from 1.2 in 2010 to 1.6 in 2017, an increase of 33% (Fig. 1) (FMCSA, 2017). There has also been an even more dramatic increase in the large trucks involved in injury crashes per 100 million miles from 20.3 in 2010 to 38.1 in 2016, an increase of 89% (Fig. 2) (FMCSA, 2019). These significant increases contrast sharply with a continuous decline in both trucks involved in fatal and injury crashes per 100 million miles between 1975 (the first year of recorded truck crash and injury data by the Federal Motor Carrier Safety Administration [FMCSA]) and 2009. Each of the 4237 fatal truck crashes in 2017 involved an economic cost of \$7.2 million or a total cost of \$30.5 billion.

Recognizing the inherent danger associated with large trucks operating on the nation’s highways, it is important to understand the obligations for shippers, brokers, and third-party logistics providers, the users of motor carrier transportation services, to ensure that the carriers they select can, in fact, conduct their operations in a safe and efficient manner. Over the last 15 years, the issue of shipper, broker, and third-party logistics provider responsibility has been litigated in civil cases based on truck crashes that resulted in fatalities and/or serious personal injuries. These cases litigated in state and federal courts have attempted to define the scope of responsibility for users who select individual motor carriers to transport freight on public highways (e.g., Schramm v. Foster et al., 2004; Jones v. C.H. Robinson Worldwide, Inc., 2008).

Prior to examining specific aspects of broker, shipper, and third-party logistics provider responsibility in the carrier selection process, it is necessary to examine the dynamics of the trucking industry structure as well as the role of the FMCSA in providing the user community with near real-time data regarding the safety performance of individual motor carriers. Both of these topics need to be explored in order to establish the context for assessing the user community responsibility in the carrier selection process.

Dynamic Structure of the Trucking Industry in the United States

Congress essentially deregulated the industry with the passage of the Motor Carrier Act of 1980. The Act created a process allowing for entities to obtain a DOT number quite easily, by completing a registration form where a declaration is made with respect to the type of operation requested (interstate or intrastate), as well as the commodity segments intended to operate in (general freight, refrigerated cargoes, etc.). Once the entity or carrier received a DOT number, the process of obtaining an operating certificate is also straightforward involving the payment of a \$300 fee, the completion of an information form and safety certification, proof of insurance, and the designation of a process agent. These minimal requirements to initiate a valid motor carrier operating business have resulted in an explosion in the total number of interstate motor carriers. In fact, the total number of carriers stood at over 500,000 in 2019. Recognize, however, that the majority of the interstate motor carriers with valid operating authority own and operate a small number of power units. In fact, the overwhelming majority of the 500,000 carriers own fewer than five power units, with the majority owning one or two power units (Cantor et al., 2013, 2017).

The industry structure has a profound impact on motor carrier safety as well as on the responsibility of the user community in the carrier selection process. Previous academic research indicates that a one-vehicle owner-operator who drives for many different employers is likely to have significantly worse safety performance (in terms of both regulatory compliance and crash rates) than would a carrier with larger vehicle fleets and established safety management programs and policies. This makes sense, because there is

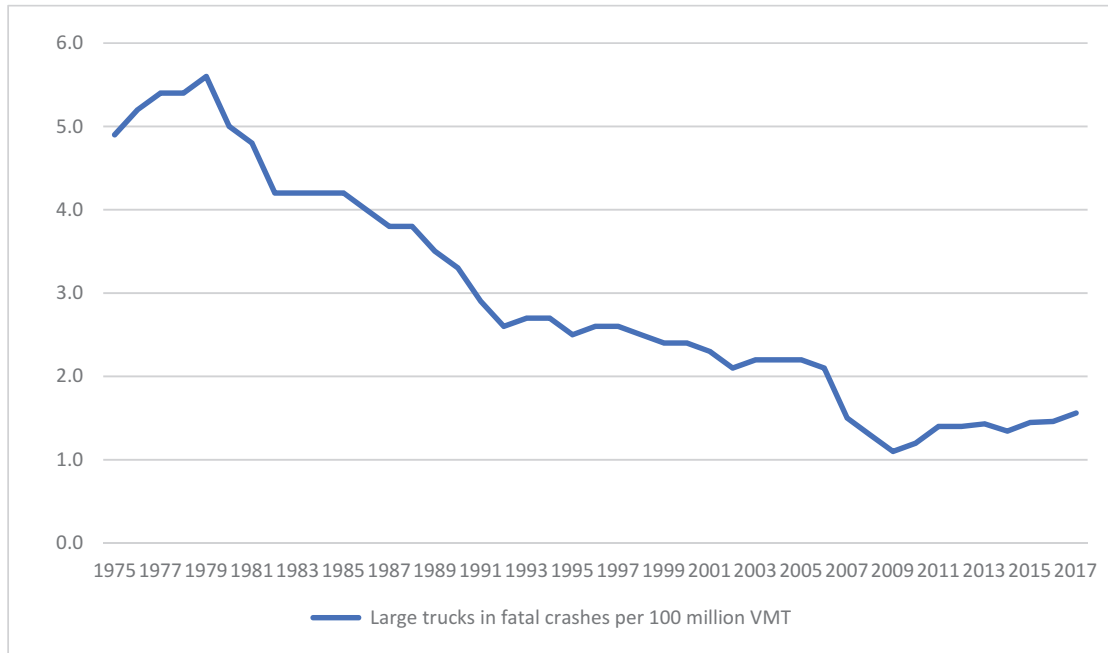


Figure 1 Large truck fatal crash rate: 1975–2017. *Source: Analysis Division, Federal Motor Carrier Safety Administration, Large Truck and Bus Crash Facts 2017, March 2019.*

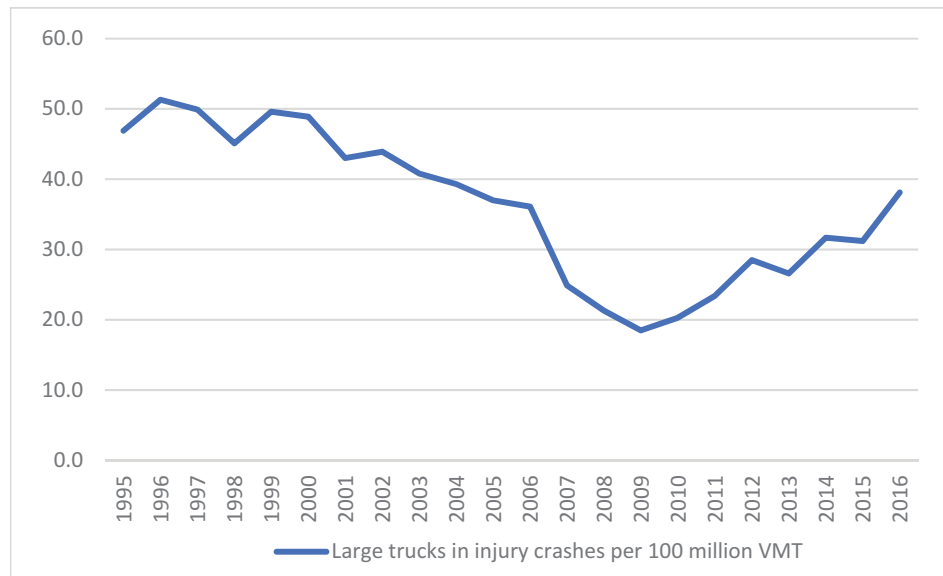


Figure 2 Large truck injury crash rate: 1995–2016. *Source: Analysis Division, Federal Motor Carrier Safety Administration, Large Truck and Bus Crash Facts 2017, March 2019.*

a good chance that a one-person owner-operator will not have the resources necessary to establish internal safety monitoring programs and processes, and to properly train and supervise himself. These findings have important implications for the carrier selection process and the responsibility of the user community in exercising reasonable care in the selection process (Cantor et al., 2017).

FMCSA Motor Carrier Safety Performance Data Initiatives

The FMCSA recognized that with the rapidly increasing number of motor carriers with valid interstate operating authority, it was critical to provide information to the transportation user community about the safety performance of the carriers. Indeed, the

objective of the FMCSA was to provide the user community with ongoing, real-time information regarding the safety performance of individual carriers so that users could incorporate this information in their selection process.

From December 1999 through November 2010, FMCSA provided information about carrier safety performance through the Analysis and Information section of its website. The SafeStat (Motor Carrier Safety Status Measurement System) evaluated carrier performance in four areas (safety evaluation areas or SEAs: driver, vehicle, safety management, and accidents) by scoring carriers on a percentile scale of 1–100, with higher scores indicating poorer safety performance. Carriers with scores above the 75th percentile on any of the SEAs were judged to be “above the threshold” level and subject, therefore, to additional inspections and reviews. In December 2010, the SafeStat system was replaced with the safety measurement system (SMS). The SMS was a finer grained assessment of carrier safety performance. It broke down the Driver SEA into its components (unsafe driving, hours-of-service violations, drug and alcohol abuse, and driver fitness). In addition, the SMS scored carriers on vehicle maintenance and provided a crash indicator. The FMCSA set threshold scores for each of the indicators, labeled in SMS as BASICS (behavioral analysis and safety improvement categories), based on the relationship between each BASIC to future crash likelihood. The threshold scores for unsafe driving, hours-of-service violations, and vehicle maintenance were set lower than the scores for the other BASICS, based on systematic empirical research demonstrating that higher scores on these three BASICS (thus indicating poorer safety performance) were more directly associated with a significantly higher future crash rate. Like SafeStat, SMS tagged carriers as being above the threshold if their percentile scores, in a particular BASIC, exceeded the designated threshold for that BASIC (FMCSA, 2019).

Both the SafeStat and the SMS converted carrier safety scores to percentiles so that carriers with like characteristics (in size or number of safety events) were grouped together for comparison purposes. The basic inputs into both SafeStat and SMS were the results of roadside inspections and “Compliance Reviews.” Each inspection or review included a number of items for assessment. Scores were computed based on the number of violations detected, along with the severity of each violation. Once each carrier was scored based on its accumulation of violations during inspections and reviews, the carrier was assigned to a specific size group. Within each group, carriers were arrayed based on their raw violation score and percentiles were assigned within each group, with the carrier having the fewest violations obtaining the lowest percentile score and the carrier with the most violations obtaining the highest percentile.

The underlying philosophy of FMCSA throughout this period was that with the large number of operating carriers in a deregulated environment, the best strategy for ensuring safe operations on the nation’s highways was to make real-time safety performance data available online so that the user community could incorporate this information as they selected individual carriers. Recognize that both SafeStat and SMS had data sufficiency tests such that carriers with insufficient data regarding inspections or reviews were not scored in the respective systems. Thus, many small carriers, in particular, do not have performance scores.

It is also important to note that separate from SafeStat and SMS, Congress mandated in 1984 that regulators determine the safety fitness of interstate motor carriers through an onsite comprehensive “Compliance Review.” As a result of this review, carriers receive a score of either “satisfactory,” “unsatisfactory,” or “conditional.” Carriers with an “unsatisfactory” result ultimately will lose their operating authority if they do not obtain approval of a corrective action plan. The results of Compliance Reviews continue at this time to be the sole basis for assigning an interstate carrier with a Safety Fitness Rating. FMCSA has resources to conduct a limited number of Compliance Reviews (some 15,000 annually). Thus, the overwhelming majority of carriers operating in interstate commerce have no Safety Fitness Rating, although many of these carriers do, in fact, have an array of scores in the BASICS that are displayed within the Analysis and Information section of FMCSA’s website.

Assessing “Reasonableness” in the Carrier Selection Process

With an understanding of the trucking industry structure and knowledge about the public availability of real-time motor carrier performance available as a consequence of FMCSA’s SMS program, it is important to examine the level of responsibility for shippers, brokers, and third-party logistics providers in the carrier selection process.

There is a consensus that there are three necessary conditions that users of motor carrier services must establish that their selected carriers possess prior to engaging their services. Users of motor carrier services have the obligation to ascertain that their selected carriers have valid motor carrier operating authority, have the required levels of insurance coverage, and do not have a Safety Fitness Rating of unsatisfactory. While these conditions are universally accepted as necessary conditions for selection, the discussion revolves around whether or not these three conditions constitute a sufficient level of evaluation in order to ensure that motor carrier operations on the nation’s highways are conducted in a safe and efficient manner (Corsi, 2018).

Defining Carriers as Independent Contractors or Agents; Assessing Negligence in Carrier Hiring

Many brokers, shippers, and third-party logistics providers have the opinion that the motor carriers whose services they engage are “independent contractors.” As a consequence, the entire burden of complying with FMCSA rules and regulations is up to the carriers themselves. The brokers, shippers, and third-party logistics providers often state in broker-carrier or shipper-carrier agreements that the carrier is obliged to comply with the regulations, but that responsibility, according to the signed agreements, is the sole responsibility of the carriers, themselves. Many brokers, shippers, and third-party logistics providers assert that they have absolutely no responsibility beyond ensuring that the earlier necessary conditions for carrier selection are met.

There are two important challenges to this interpretation put forward by many shippers, brokers, and third-party logistics providers. First, in many situations, the relationship between the shippers, brokers, and third-party logistics providers, and the carriers does not meet the requirements associated with the engagement of an independent contractor, since the users exercise a degree of control over the carriers activities that exceeds the actions normally associated with an entity engaging the services of an “independent contractor.” In many instances, the shipper, broker, or third-party logistics provider arranges the shipment, dispatches the carrier drivers, monitors the freight shipment from origin to destination, requires the driver to call the dispatch at multiple times over the course of the shipment, penalizes the driver for late deliveries, bills the shipper, pays the carrier, and prohibits the carrier from contact with the shipper/receiver. These actions are totally inconsistent with actions normally associated with an entity engaging the services of an “independent contractor.” In these situations, the motor carrier is far from “independent.” The relationship is more properly defined as a “principal-agent” relationship, with the principal directly responsible for the actions of its agent (*Tryg Insurance v. C.H. Robinson Worldwide, Inc.*, 2019).

Second, the viewpoint that shippers, brokers, and third-party logistics providers have no responsibility in motor carrier selection beyond the necessary conditions cited earlier is inconsistent with the obligation of a contractor to use reasonable care in selecting a subcontractor. Indeed, the shipper, broker, or third-party logistics providers can be found negligent if they fail to exercise caution, oversight, or reasonable care in the selection process.

The Legal Battle for Assessing Carriers as Independent Contractors and Determining Negligence in Hiring

Legal disputes have emerged over the last 15 years regarding the issue of whether carriers selected by shippers, brokers, and third-party logistics providers are, in fact, independent contractors or simply the agents of the party who hired them. Equally subject to significant dispute is the question of whether or not the users exercised reasonable care in the selection of carriers, if the carriers were, in fact, considered as independent contractors.

Two seminal cases in this area are the Schramm Case (2005, Maryland) and the Jones Case (2008, Virginia). In both cases, Courts ruled that the broker, C.H. Robinson, the nation’s largest broker, had an obligation to review available safety performance data as part of its carrier selection process. The respective Courts ruled that this requirement was not, in fact, a burden but a necessity. The cases were decided on the basis of negligent hiring of a carrier by the broker involved.

These seminal cases led to a series of additional cases in which plaintiffs sued shippers, brokers, and third-party logistics providers in the aftermath of fatal or serious crash injury cases on the basis either that the carriers were not, in fact, independent contractors but were subject to the control and oversight of the shippers, brokers, or third-party logistics providers, or that there was not reasonable care exercised in the selection process. In fact, the failure to use reasonable care constituted negligence in the hiring of an independent carrier by the shipper, broker, or third-party logistics provider. A number of these cases have resulted in out-of-court settlements between the parties. As a result, from a legal perspective, there is continuing uncertainty about the ultimate resolution of the issue.

Shippers, Brokers, and Third-Party Logistics Providers Seek Relief From Reasonable Care Obligations

In this environment, shippers, brokers, and third-party logistics providers have mounted a concerted effort to obtain relief from any obligation in hiring motor carriers beyond the three minimum standards discussed earlier. The user community launched an extended campaign against the percentile rankings associated with the BASIC scores in the SMS. As a result, Congress in December 2015 passed the Fixing America’s Surface Transportation Act (FAST). This legislation required the FMCSA to remove from the public website percentile scores for each carrier on each of the BASICs. The legislation allowed the FMCSA to continue publishing on its public website each carrier’s underlying scores (called measure scores). As noted, the underlying measure scores were converted in the methodology to percentile scores, after carriers were divided into size groups based on each carrier’s number of safety events (inspections and reviews). The user community had launched its attack with the claim that the SMS methodology was unreliable. In passing the FAST Act, Congress commissioned the National Academy of Sciences/Transportation Research Board to conduct a thorough examination of the efficacy of the SMS methodology. The report concluded that the methodology was sound in its ability to associate high percentile scores with increased likelihood of future crash involvement (*Volpe National Transportation Systems Center*, 2014). However, the report did, in fact, propose that the FMCSA evaluates some alternative methodologies, believed by the authors of the report to have the potential to be more effective than the SMS methodology. The report made no recommendation on whether the percentile scores should be made available once again (*National Academy of and Sciences*, 2017). As of 2019, the percentile scores and the associated threshold warnings are not part of the public website.

The user community as of 2019 were actively pursuing additional federal legislation to establish that the process of carrier selection be based on the necessary criteria discussed earlier with one important change. Rather than using the existing Safety Fitness Rating, based solely on Compliance Review, the user community advocate a far more generous method, resulting in almost all carriers being regarded as satisfactory, as the basis for the third component of the necessary selection criteria. The user community sought this legislation in the hopes of being provided with a legislative basis for relieving them of any obligation to review a carrier’s safety performance data at the level of the BASICs in making a selection decision. Indeed, the nation’s largest broker, C.H. Robinson, has repeated the assertion in numerous court depositions, that it considers absolutely no safety performance data for carriers in their selection process beyond establishing that the carrier has a valid operating authority, proper insurance levels, and does not have an

“unsatisfactory” safety rating (Craig, 2019). The intended legislation will codify a selection process with minimal attention to safety performance of the carriers with an emphasis on necessary, but not sufficient, considerations.

Safety Implications of Minimal Shipper, Broker, and Third-Party Logistics Provider Attention to Safety Performance Matters in Carrier Selection

There are a number of concerns regarding the proposed legislative cover for shippers, brokers, and third-party logistics providers to minimize their responsibility in evaluating and selecting motor carriers for transportation services. First, recognize that the inclusion of a carrier’s safety performance record as a key component of the selection process constitutes a powerful incentive for carriers to maintain an excellent safety record by avoiding violations of FMCSA regulations during inspections and reviews. Carriers have recognized, beginning with the introduction of online safety performance data with SafeStat, that their poor safety performance record will have a direct impact on future business opportunities, especially among the users who place a high value on securing safe carriers. Indeed, some shippers, brokers, and third-party logistics providers would not engage the services of carriers with multiple BASIC values exceeding FMCSA established thresholds. With the removal of the BASIC percentile scores and above-threshold indicators from the FMCSA website, it can certainly be argued that it is more challenging for users to incorporate a carrier’s safety performance record into its selection algorithm.

Second, recognize that, in the current environment, many truckload shipments are arranged through internet load boards. Shippers frequently contract with brokers/third-party logistics providers to transport specific loads on their behalf for a specific freight rate. Brokers/third-party logistics providers will fulfill the particular shipper request by engaging the services of a carrier with whom the broker/third-party logistics provider has an ongoing relationship (formalized in a broker–carrier agreement) or by posting the load on an internet load board. Posting the load elicits bids from carriers to provide the requested transportation services. The broker/third-party logistic provider’s business model is based on the margin between the amount they have contracted with the shipper for the designated transportation services and the amount bid by the carriers to transport the load in question. The lower the carrier bid for the transportation services, the higher the broker margin. This system creates an incentive for the brokers to select carriers offering the lowest bid. Previous academic research has demonstrated that there is a direct link between carriers under financial stress (likely carriers consistently submitting low bids in response to broker/third-party logistics postings) who have significantly higher rates of FMCSA violations and crashes (Miller and Saldanha, 2016). Minimizing the responsibility of brokers/third-party logistics providers to incorporate safety considerations in the carrier selection process would only intensify the use of low-bid carriers, likely under financial distress.

Third, the discussion of the dynamic trucking industry structure in a deregulated environment revealed that there are thousands of carriers with valid interstate operating authority who are no more than single-vehicle owner-operators—that is, one-person carriers. Previous academic research has shown that these one-vehicle carriers have significantly worse safety performance than do carriers with larger vehicle fleets and established safety management programs and policies. Indeed, the one-person carriers do not have resources necessary to establish internal safety monitoring programs and processes. Nor do they have the resources to properly train and supervise themselves. Yet, under the current environment, shippers, brokers, and third-party logistics providers want to minimize their oversight of these single-vehicle carriers with respect to safety performance (Cantor et al., 2013). Rather than hiring a carrier with an established vehicle fleet and safety program, engaging the single-vehicle carrier is directly comparable to hiring an individual driver, not a carrier. Allowing shippers, brokers, and third-party logistics providers to equate these single-vehicle owner-operators as carriers and independent carriers as opposed to requiring them to use reasonable care in examining the safety performance of these individuals will have negative safety consequences with respect to crash frequency.

Finally, recognize that if users are successful in obtaining legislation from Congress that restricts the requirements for carrier selection to the minimum levels stipulated earlier (with a new less restrictive determination of Safety Fitness, not tied to the results of a Compliance Review), users must consider that they will still be subject to legal challenges based on the control exercised over the carriers, defining the relationship as that between a principal and its agent, rather than as being defined as an independent contractor relationship. Second, there will remain the issue to determine whether the user exercised reasonable care in the selection process or whether the user was negligent in contracting with a carrier who represented a known risk to the traveling public through its record of safety performance violations and past crashes.

Requiring users to incorporate safety considerations in a carrier selection process, due to the inherent danger associated with motor carrier operations, has been incorporated into the broader concept of a “chain of responsibility,” shared by all parties involved in freight transportation. Thus, Australia has codified the shared responsibility concept that encompasses all parties involved in the transportation of freight from the shipper to the broker/third-party logistics provider to the carrier and, ultimately, to the driver. In the Australian legal environment, each of these entities has a shared duty to ensure that trucking operations are conducted in a safe manner (OECD, 2011).

Conclusion

The trend in higher rates of truck involvement in both fatal and injury crashes since 2010 is disturbing. What intensifies concerns about this trend is the concerted effort by many shippers, brokers, and third-party logistics providers to minimize the role of carrier safety performance in the selection process. There is a clear danger in breaking the “chain of responsibility,” covering shippers,

brokers, third-party logistics providers, the carriers themselves, and the individual drivers to ensure that motor carrier operations are conducted in a safe and efficient manner.

As noted, the increasing reliance on internet load boards for selecting carriers creates an environment that encourages brokers and third-party logistics providers to select low-cost carriers, often carriers who have reduced rates at the expense of safety management programs and practices. Requirements for users to exercise “reasonable care” in selecting carriers represent a powerful counter to ensure that safety receives priority in the selection process.

Furthermore, the changed industry structure has resulted in a large number of single-vehicle, owner-operator carriers. The shippers, brokers, and third-party logistics providers engaging the services of a single-vehicle, owner-operator carrier are, in reality, hiring an individual driver, who requires management oversight to ensure compliance with FMCSA regulations. Shielding shippers, brokers, and third-party logistics providers from the “chain of responsibility” in these cases will further erode truck safety on US highways.

Reversing the recent trend toward higher rates of truck involvement in both fatal and injury crashes requires a greater commitment to enforcing a “chain of responsibility” to ensure that the carriers selected have processes in place that will minimize the likelihood of future serious truck crashes. Efforts to minimize the responsibility of parties in the selection and oversight of carriers are counterproductive to the goal of reducing truck involvement in serious crashes.

The FMCSA should continue to support requirements for interstate carriers to adopt new technologies designed to enhance truck safety (Cantor et al., 2016). The FMCSA requirement for carriers to install electronic logbooks is an excellent example of a regulatory initiative to enhance safety. There is no question that there are rapid advancements in vehicle technology to make transportation operations safer by reducing crash risk. FMCSA should be diligent in examining new technologies that will enhance safety and requiring the industry to adopt those that have a documented impact on truck crashes. The FMCSA has the legislative mandate to reduce truck crashes and should continue in its efforts to be proactive in fulfilling its mission.

References

- Cantor, D.E., Celebi, H., Corsi, T.M., Grimm, C.M., 2013. Do owner-operators pose a safety risk on the nation's highways? *Transp. Res. Part E Logist. Transp. Rev.* 59, 34–47.
- Cantor, D.E., Corsi, T.M., Grimm, C.M., Singh, P., 2016. Technology, firm size, and safety: theory and empirical evidence from the US motor-carrier industry. *Transp. J.* 55 (2), 149–167.
- Cantor, D.E., Corsi, T.M., Grimm, C.M., 2017. The impact of new entrants and the new entrant program on motor carrier safety performance. *Transp. Res. Part E Logist. Transp. Rev.* 97, 217–227.
- Corsi, T., 2018. Broker/third party logistics provider and shipper responsibility in motor carrier selection: considering carrier safety performance. In: Peoples, J., Bitzan, J. (Eds.), *Transportation Policy and Economic Regulation: Essays in Honor of Theodore Keeler*. Elsevier, Amsterdam, pp. 311–330.
- Craig, J., 2019. Under pressure: the state of the trucking industry in America. Testimony of Jason Craig, Director of Government Affairs, C.H. Robinson Worldwide, Inc. before the House Subcommittee on Highways and Transit. Available from: <https://transportation.house.gov/imo/media/doc/Testimony-Craig.pdf>.
- FMCSA, 2017. Large Truck and Bus Crash Facts 2017. FMCSA-RRA-18-018. Federal Motor Carrier Safety Administration, Washington, DC.
- FMCSA, 2019. Safety Measurement System (SMS) Methodology Version 3.10. Federal Motor Carrier Safety Administration, U.S. Department of Transportation, Washington, DC.
- Jones v. C.H. Robinson Worldwide, Inc., 2008. United States District Court for the Western District of Virginia (Roanoke Division), Civil Action No. 7:06CV 00547, Roanoke, VA.
- Miller, J.W., Saldanha, J.P., 2016. A new look at the longitudinal relationship between motor carrier financial performance and safety. *J. Bus. Logist.* 37 (3), 284–306.
- National Academy of Sciences, 2017. Improving Motor Carrier Safety Measurement. National Academy Press, Transportation Research Board, Washington, DC.
- OECD, 2011. Moving Freight With Better Trucks: Improving Safety, Productivity, and Sustainability. Organization for Economic Co-Operation and Development Publishing, Paris.
- Schramm v. Foster et al., 2004. District Court, Maryland, 341 F Supp 2d 536, Ocean City, MD.
- Tryg Insurance v. C.H. Robinson Worldwide, Inc., 2019. U.S. Court of Appeals for the Third Circuit, April 19, 2019, No. 17-3768 (C.H. Robinson defined as a motor carrier as opposed to a freight broker), Camden, NJ.
- Volpe National Transportation Systems Center, 2014. The Carrier Safety Measurement System (CSMS) Effectiveness Test by Behavior Analysis and Safety Improvement Categories (BASICS), prepared for Federal Motor Carrier Safety Administration, Washington, DC.

Carrier Selection for Less-Than-Truckload (LTL) Shipments

Dinçer Konur*, **Gonca Yıldırım†**, **Bahriye Cesaret‡**, *Department of Computer Information Systems and Quantitative Methods, McCoy College of Business Administration, Texas State University, San Marcos, TX, United States; †Department of Industrial Engineering, Cankaya University, Ankara, Turkey; ‡Faculty of Business, Ozyegin University, Istanbul, Turkey

© 2021 Elsevier Ltd. All rights reserved.

Introduction	388
LTL Carrier Selection and Supply Chain Planning	389
LTL Carrier Selection and Strategic Level Planning	389
LTL Carrier Selection and Tactical Level Planning	389
LTL Carrier Selection and Operational Level Planning	389
LTL Carriers and Shippers	390
How Do the LTL Carriers Work?	390
What Do the LTL Shippers Want?	390
Inventory Control and the Impact of Key Factors	391
Impact of Transit Time	392
Impact of Transit Time Variability	392
Impacts of Freight Minimums and Maximums	393
Conclusion	393
References	394

Introduction

Logistics costs account for a substantial portion of the overall cost of a product delivered to the end consumers. In some product categories, logistics costs can make up more than 50% of the total cost. It is not surprising, therefore, that companies seek to manage their logistics operations as efficiently as possible. Unless a company operates its own private fleet (such as Walmart, PepsiCo), it must utilize carriers to transport its shipments. Carrier selection is one of the most important logistics decisions a company must make, as the carrier selection decision has a huge impact on a company's strategic, tactical, and operational planning phases.

One can categorize carriers into three classes based on the typical size of the shipments: courier service providers (parcel delivery), less-than-truckload (LTL) carriers, and full-truckload (FTL) carriers. Courier service providers, such as USPS, UPS, and FedEx, deliver parcels of relatively small size nonpalletized packages, which typically weigh less than 100 lbs (50 kg). Courier service providers are mostly used for residential deliveries. LTL carriers generally transport shipments consisting of multiple pallets that do not make up a full trailer, that is, a full truckload. On the other hand, FTL carriers, as the name suggests, dedicate trucks for the shipments they make. It is worthwhile to note that an FTL carrier does not necessarily ship a fully loaded truck because the shipper might prefer to partially fill a truck.

LTL transportation is crucial for the success of many businesses, especially considering the growth in ecommerce. In the United States, the LTL market is somewhere around \$35–\$40 billion (Ozkaya et al., 2010; Bokher, 2018; Robinson, 2018). An LTL shipper refers to a business using the services of an LTL carrier. Considering that there are many LTL carriers available with varying attributes, it is a challenging task to select a carrier for the shipments. Furthermore, the carrier selected will affect the performance of the other supply chain practices undertaken. Therefore, a business needs to be careful in selecting a carrier for its LTL shipments. Prior to discussing LTL carrier selection further, it is important to discuss when a shipper should prefer LTL over FTL shipments.

As noted earlier, the size of the shipment is an important factor to determine the type of the shipment; however it is not the only factor. The frequency of shipments is another key factor. For instance, instead of making weekly LTL shipments, a shipper can consolidate several of its shipments and make monthly FTL shipments. Another factor to consider is the number of different shipments made. Again, a shipper can consolidate several relatively small LTL shipments destined to different locations into one FTL shipment, which would require a truck to make several delivery stops. Such a consolidation of shipments would require coordinating the shipments to different destinations, as this consolidation necessitates common delivery windows (see, e.g., Konur and Geunes, 2019). Furthermore, most LTL and FTL carriers offer discounts based on the size and the frequency of the shipments. Therefore, the shipper should consider consolidation opportunities and carefully evaluate the implied tradeoffs between the frequency and the amount of individual and consolidated shipment options before deciding to use an LTL carrier over an FTL carrier. Moreover, the freight type, shipment budget, delivery time restrictions, and loading/unloading requirements should be considered when choosing between LTL and FTL shipments.

The focus in this chapter is on LTL carrier selection. In particular, we will first discuss the implications of carrier selection in strategic, tactical, and operational planning phases of a supply chain. Then, the factors shippers typically consider in carrier selection for LTL shipments are reviewed. Next, we quantify the impacts of four key factors on the logistics costs and service levels of a company through a stochastic inventory control model. The reader is referred to [Meixell and Norbis \(2008\)](#) for a review of transportation mode choice and the carrier selection literature for other various aspects of LTL carrier selection research.

LTL Carrier Selection and Supply Chain Planning

Effective supply chain management is key to the success of any business. Every planning decision made will have an impact on every stage of a supply chain. Typically, supply chain planning decisions are managed at three levels: strategic, tactical, and operational planning ([Nahmias, 2009](#)). The strategic planning phase consists of long-term decisions, which will be in effect for a relatively long period of time (e.g., 5 year, 1 year), such as supply chain network design, market entry choices, products offered, and manufacturing strategies. For instance, location decisions of production facilities and/or warehouses are strategic planning decisions, as a business will operate such a facility for a long time at the same location. The tactical planning phase consists of mid-term (e.g., quarterly, monthly) decisions such as seasonal promotions, supplier contracting, and inventory control decisions. The operational planning phase consists of short-term decisions, related to weekly or daily operations such as deliveries, routing and loading/unloading operations. LTL carrier selection can be regarded as a tactical or an operational planning decision; however it will have an impact on the performance of strategic planning decisions as well.

LTL Carrier Selection and Strategic Level Planning

Strategic level decisions are crucial for successful supply chain management, since such decisions constrain the tactical and operational planning decisions. LTL carrier selection will be directly affected by several strategic planning decisions. The supply chain network configuration and the market entry decisions of a business are the main determinants of the shipments to be made. Accordingly, a company might need to use LTL shipments and select carriers for each shipment. The performance of a strategic decision made will be affected by LTL carrier selection. In order to efficiently design a supply chain network, a business should account for the implied LTL shipment costs, which means that it should evaluate the available LTL carriers and their attributes for various network configurations. Similarly, in market entry decisions, a business should analyze the feasibility of such entry decisions considering the carriers available to make LTL shipments to various customers in the markets to be entered. While strategic level planning typically does not account for the detailed cost implications of a decision on the mid- and short-term practices, a business should not completely ignore the effects of LTL shipments in their strategic planning.

LTL Carrier Selection and Tactical Level Planning

At the tactical level, the decisions made are not binding for a very long time. However, it is not very easy to adjust the tactical level plans without significant operational changes. LTL carrier selection can be accepted as a tactical level planning decision, especially considering the contracting opportunities with an LTL carrier. Typically, a business can negotiate a relatively long-term contract with a carrier. Such contracts may include conditions on the annual amount of shipments needed to be made by the shipper. Inventory management decisions at the tactical level should be made in an integrated manner with LTL carrier selection. Particularly, inventory control is the key driver of the amount and frequency of the shipments to be made, which are the key factors in deciding which carriers to use. The inventory management literature has focused intensely on joint carrier selection and inventory control decisions. LTL carrier selection should also be accounted for in supplier selection decisions because the selected suppliers will characterize the origin/destination of the required shipments. In addition, seasonal promotions offered by a business might imply significant changes in the shipment requirements, therefore LTL carrier selection should be taken into account in designing seasonal promotions. To this end, while LTL carrier selection itself can be considered as a tactical decision, it should also be integrated with several other tactical planning decisions.

LTL Carrier Selection and Operational Level Planning

Short-term recurring operations are crucial for the seamless supply chain activities of a business. On-time deliveries, material handling operations and inventory replenishment activities have significant cost, as well as service level, implications. LTL carrier selection is, therefore, essential for efficient operational planning. One reason is that the carrier selected determines the transit time and, thereby, when the shipments are estimated to be delivered. Delivery times are critical for both inbound and outbound shipments. Inbound shipment times might affect the production at a manufacturing facility and inventory replenishments at a retail store. Outbound shipment times can make the shipper liable for delayed shipments, considering the just-in-time requirements of the receiving party. Furthermore, the shipment is not just about transportation. A shipment will be loaded and unloaded, and these operations might require materials handling equipment. LTL carriers might offer material handling services in addition to the transportation of the shipments. As such, LTL carrier selection has an impact on many short-term operations, and therefore, available carriers should be carefully evaluated considering various operational implications.

LTL Carriers and Shippers

The two main decision makers about LTL shipments are the carriers and the shippers. The LTL carriers decide on the rates for the services they provide and how they provide these services. The LTL shippers decide on their shipment sizes and which carrier(s) to use. It is important for each party to understand the dynamics of the other.

How Do the LTL Carriers Work?

The main advantage of LTL carriers for a shipper is the cost savings from a shipment that does not justify an FTL shipment. LTL carriers also offer flexibility through alternative schedules, and they can provide additional services such as loading/unloading support. In general, LTL carriers are able to offer such benefits because they utilize economies of scale due to the aggregation of individual demand for their services. Typically, LTL carriers operate on a hub-and-spoke network (Cerasis, 2016). Drivers with small trucks collect the regional demand and carry it to a regional distribution point; the cumulative regional demand is then shipped to a hub; and the hub serves as a cross-dock facility where shipments from different regions are consolidated and dispatched according to their destinations. This consolidation of various requests from different shippers enables the high utilization of the transportation capacity of the LTL carrier, which in return, reduces the costs. Therefore, LTL carriers can offer competitive shipment rates. Besides the base rate of a carrier in relation to its service quality and business background, the LTL rates are typically determined based on three main factors: origin-destination, shipment size, and freight class (Cerasis, 2016).

- Origin-destination will determine the distance of the shipment to be made, hence will affect the rate of the shipment. While it is expected that the rate would increase with the distance, the shipping rate is not always directly proportional to the distance of the shipment. Some LTL carriers might operate only in specific regions, hence a shipment out of the carrier's region would be subject to higher rates as the carrier might need to use other service providers. Moreover, depending on the carrier's network, the direct distance between the origin and the destination of the shipment will not necessarily define the distance the shipment travels.
- Shipment size is a key factor for an LTL carrier in pricing the shipment. The weight as well as the volume of the shipment are used by LTL carriers to determine the rates. The weight and the volume define the density of the shipment, which is used in determining the freight class. The freight class is then used in calculating the rate for the shipment. Furthermore, LTL carriers offer weight- and/or volume-based discounts to attract larger shipments.
- Freight classification helps describe a shipment's transportability based on its density, handling, stowability, and liability. These factors are all cost sources for a carrier, therefore the freight class will directly impact the rate an LTL carrier charges for a shipment.

Besides these factors, the rate for an LTL shipment will depend on the speed desired, additional services (such as loading/unloading support, residential pickup/delivery, shipments from/to limited-access locations), the carrier's freight minimum and the contract between the shipper and the carrier. Shippers can negotiate rates with carriers based on their shipment frequency and the cumulative amount of different shipments they make with the same carrier. An LTL shipper should understand that what drives a carrier's shipment rates and consider making adjustments accordingly in order to obtain better rates.

What Do the LTL Shippers Want?

As any customer for any service, LTL shippers seek the best service at the best rates from LTL carriers. Of course, just the cost of the service does not reveal the best LTL carrier and the service quality is not straightforward to measure, considering the various aspects of the service that are involved. Essentially, LTL shippers want many things, and thus LTL carrier selection is a multiattribute decision-making problem. We summarize common attributes typically used by LTL shippers for carrier selection in six categories as follows (Lambert et al., 1993; Premeaux, 2009; Williams et al., 2013; Vega et al., 2018). It should be noted that some attributes might be specific to a business type and a business might have other special attributes for evaluating LTL carriers.

- Cost and billing: shippers prefer competitive rates. As discussed earlier, a carrier's shipping rate depends on the origin/destination, shipment size, freight class as well as the speed desired. Furthermore, carriers can offer discounts and a shipper can negotiate rates. In addition, the costs of the other services provided are important for a shipper. Billing accuracy and transparency are also important attributes for LTL shippers. Billing should accurately reflect the shipping charges, how any damages are billed, costs of the additional services, and items included in the charge.
- Time performance: shipping time is a key performance measure of a carrier's service. LTL shippers regard the transit time, transit time reliability, on-time pickups, and on-time deliveries as crucial attributes for evaluating LTL carriers, because these attributes have a significant impact on a shipper's profitability and service quality.
- Shipment restrictions: typically, LTL carriers have restrictions on the shipment sizes and/or the cost of the shipment. These are generally referred to as freight minimums and freight maximums. Freight minimums define the minimum amount required for a carrier to accept a shipment and carriers set freight minimums in order to ensure a certain profit margin from a shipment. Freight maximums define the maximum amount a carrier can accept and carriers set freight maximums considering their fleet and capacity. Freight minimums and maximums should be taken into account by a shipper, as they will significantly affect the cost and transit time of the shipments. In addition, some carriers might have restrictions on what they can ship. For instance, some carriers will not have refrigerated trucks in their fleet, or some carriers will not have the required equipment and/or capability to

ship certain freight classes, such as hazardous materials. Furthermore, some carriers will have restrictions on where they can ship. A shipper should consider such restrictions in selecting a carrier.

- **Service quality:** here, we consider service quality as the set of attributes of a carrier's service other than shipping rates and times. One important service attribute LTL shippers consider is the damages incurred during transit. It is important to define who is liable and how the damages are to be billed. The carrier might offer insurance for the shipment, and LTL shippers should collect data about the damages to their shipments when they use an LTL carrier. Another crucial attribute is information sharing. It is essential for most LTL shippers to have real-time information on their shipment so that they can make necessary adjustments accordingly and/or update the receivers about the status of the shipment. Related to this, a carrier's flexibility and ability to make modifications are also important for an LTL shipper. Things might change, due to the changes in either the carrier's or the shipper's plans, and LTL shippers value the ability to make necessary changes before or during the shipment. Finally, LTL carriers can increase their service quality by offering additional services such as loading/unloading support, packaging services, redelivery and residential pickups and/or deliveries.
- **Customer service:** the level of customer service an LTL carrier provides can be a winning factor. LTL shippers enjoy prompt responses to their inquiries and claims. Furthermore, the relationship between the carrier personnel and the shippers can go a long way in prompting a shipper to work with a carrier. In addition, the professionalism and high working standards of sales personnel, drivers, and executives of a carrier are attractive attributes for LTL shippers.
- **Other attributes:** in addition to the aforementioned attributes, several other factors can attract LTL shippers. Other attributes include but are not limited to security of the shipment, sustainability of the carrier, carrier's brand image and financial stability, use of technology, fleet type and safety, data availability and ease of access to information, and other resources.

Some of the aforementioned attributes are quantifiable, whereas others are more qualitative in their nature. This makes LTL carrier selection a challenging task. LTL shippers can benefit from using third party logistics providers. In addition, transportation management systems and the integration of shipments with existing enterprise resource planning platforms can help LTL shippers manage their shipments better and achieve cost savings, as well as offer higher service levels to their customers. In the following section, we numerically discuss the effect of several key attributes on the logistics costs and service level of an LTL shipper, using an inventory control model.

Inventory Control and the Impact of Key Factors

Inventory control is the key driver of the amount and frequency of the shipments made by a business, and therefore it should be integrated with LTL carrier selection (for a review of transportation mode selection in inventory control models, see [Engbrethsen and Dauzere-Peres, 2019](#)). In this section, a numerical analysis of the impacts of four key LTL carrier attributes on a shipper's expected cost and service level using a continuous review inventory control model is undertaken. Particularly, the well-known (Q, R) model for a stochastic demand inventory control system is applied ([Hadley and Whitin, 1963](#)), where an order of Q units is placed when the on-hand inventory level drops to R . Here, Q defines the shipment size a shipper will be shipping from its supplier. The demand per unit time, here annual demand is used, is a random variable with mean λ and standard deviation ν . The shipper incurs inventory-holding cost (h per unit per year), order-setup cost (K per order), and penalty cost for each unit short (p per unit short). It is assumed that shortages are backordered. Note that shortages can occur only during the delivery of the shipment, that is, lead time, which is considered as the LTL carrier's transit time. Under the settings of the classical (Q, R) model, the shipper's annual expected cost as a function of the order-quantity Q and reorder-point R , denoted by $EC(Q, R)$, is given by:

$$EC(Q, R) = h \left(\frac{Q}{2} + R - \mu \right) + \frac{K\lambda}{Q} + \frac{p\lambda n(R)}{Q}$$

where the first term is the expected annual holding cost, the second term is the expected annual order setup cost, and the last term is the expected annual penalty cost, such that $n(R)$ denotes the expected number of shortages backordered per replenishment cycle and μ denotes the mean of the lead time demand. Assuming a constant purchase cost per unit and a constant unit shipping cost, the annual transportation and procurement costs are omitted from the expected cost function, as these will be constant. One can easily note that as the LTL carrier's unit shipping rate increases, the expected annual cost of the shipper will also increase.

In the following analyses, it is assumed that the annual demand is normally distributed. In addition to analyzing the effects of four key attributes of an LTL carrier on the shipper's expected annual costs, the shipper's service level is also considered. In particular, the changes in the expected portion of the shipper's demand that is not backordered (i.e., the shipper's service level, which is also known as the type-II service level in the (Q, R) model) are also analyzed. The type-II service level represents the shipper's fill rate and is measured as $\beta = 1 - \frac{n(R)}{Q}$. Furthermore, three classes of problem instances are considered based on the coefficient of variance of the shipper's annual demand, denoted by CV , such that $CV = \frac{\nu}{\lambda}$. The three classes of problem instances are defined such that $CV \in [0, 1, 0.4]$, that is, low variability, $CV \in [0.4, 0.7]$, that is, medium variability and $CV \in [0.7, 1]$, that is, high variability of the shipper's annual demand. For each problem class, 100 problem instances are randomly generated as follows. First, the cost parameters are randomly generated from uniform distributions such that $h \in [1, 10]$, $p \in [1, 10]$, and $K \in [150, 250]$. After that, λ is randomly generated such that $\lambda \in [2, 000, 10, 000]$ and CV is randomly generated depending on the problem class.

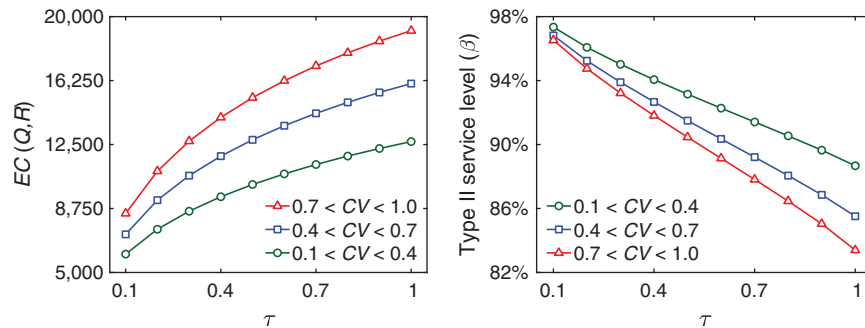


Figure 1 The impact of transit time on shipper's costs and fill rate.

Once λ and CV are known, ν is also known. Each problem instance from a problem class is solved with the same 10 varying values of the key LTL carrier attributes under investigation and the average results over 100 problem instances for each problem class are summarized.

Impact of Transit Time

As it is discussed earlier, the transit time of an LTL carrier for a shipment is one of the key attributes for the shipper. The transit time defines the lead time for the shipper, which directly influences the shipper's annual expected cost and fill rate. Here, the transit time is considered to be deterministic, and τ denotes the corresponding lead time. Fig. 1 summarizes the changes in $EC(Q, R)$ and β under the shipper's optimal (Q, R) policy as τ increases for different problem classes, respectively.

It can be observed from Fig. 1 that the expected annual cost increases, whereas the fill rate decreases, as the lead time increases. These are expected results because the longer the transit time is, the higher a shipper's shortage probability becomes, which, in turn, imply increased costs and decreased service levels. Furthermore, one can notice that when the variability of a shipper's annual demand is larger, that is, when CV is larger, the adverse effects of increased lead time duration are relatively more severe.

Impact of Transit Time Variability

The lead time variability measures the transit time reliability of the LTL carrier. In order to analyze the impact of lead time variability, we randomly generate mean lead time for a problem instance, and then solve the same problem instance with 10 different lead time standard deviation, w , values. Note that one can characterize the lead time demand distribution easily given the mean and the standard deviation of the lead time along with the details of the annual demand. Fig. 2 summarizes the changes in $EC(Q, R)$ and β under the shipper's optimal (Q, R) policy as lead time variability w increases for different problem classes.

Fig. 2 reveals that, as the lead time variability increases, the expected annual cost increases and the fill rate decreases. These are expected results because increasing lead time variability implies a higher probability of shortages which, in turn, implies higher penalty costs. To avoid these, the shipper might increase inventory levels which increases inventory holding costs. Furthermore, one can notice that when the variability of a shipper's annual demand is larger, that is, when CV is larger, the adverse effects of increased lead time variability are relatively less severe because, with larger CV, the shipper's demand during the lead time already has high variability, thus increasing the lead time variability adds up relatively less than it adds up for the cases with smaller CV values.

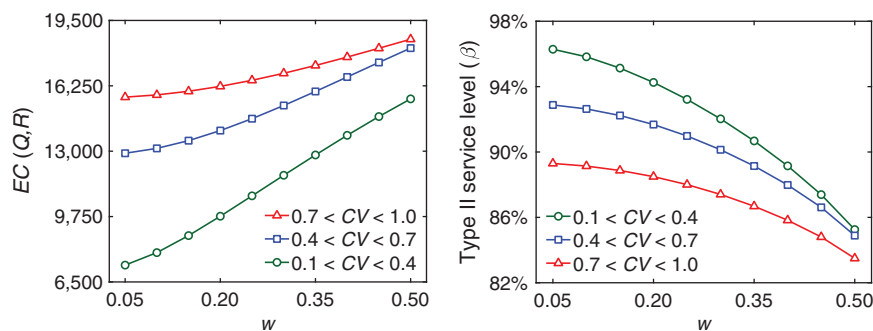


Figure 2 The impact of transit time variability on shipper's costs and fill rate.

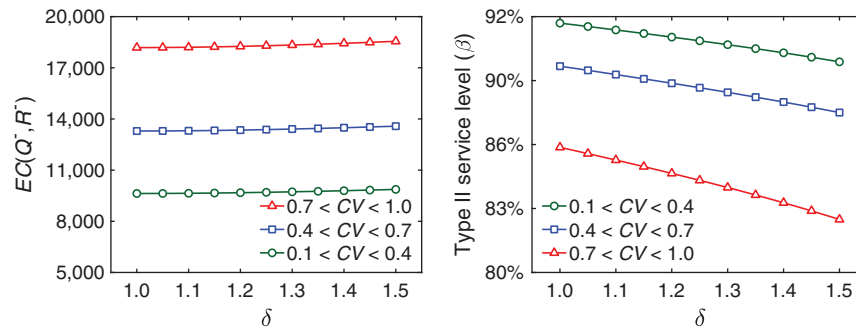


Figure 3 The impact of freight minimum on shipper's costs and fill rate.

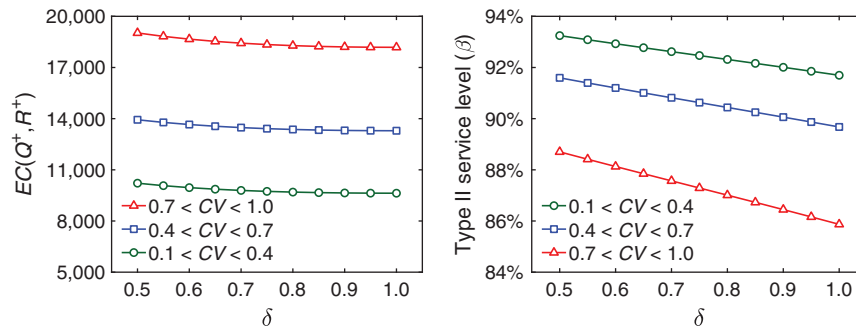


Figure 4 The impact of freight maximum on shipper's costs and fill rate.

Impacts of Freight Minimums and Maximums

To analyze the effects of an LTL carrier's freight minimums and maximums on a shipper's expected costs and service levels, the following definitions are applied. Given a problem instance, let Q^* be the shipper's optimal shipment size without any restriction on the shipment size, that is, $(Q^*, R^*) = \operatorname{argmin}\{EC(Q, R)\}$. To define binding freight minimums, the following problem instance is solved subject to $Q \geq \delta Q^*$, where δ increases from 1 to 1.5 in increments of 0.1, and let $(Q^-, R^-) = \operatorname{argmin}\{EC(Q, R) : Q \geq \delta Q^*\}$. Similarly, to define binding freight maximums, the following problem instance is solved subject to $Q \leq \delta Q^*$, where δ increases from 0.5 to 1 in increments of 0.1, and let $(Q^+, R^+) = \operatorname{argmin}\{EC(Q, R) : Q \leq \delta Q^*\}$. In order to determine (Q^-, R^-) and (Q^+, R^+) for a problem instance, the interior-point method is used. Furthermore, in this analysis, it is assumed that the carrier's transit time is deterministic and that lead time is randomly generated for each problem instance.

Fig. 3 illustrates the changes in the shipper's annual expected costs and service levels as the freight minimum increases from Q^* to $1.5Q^*$ for different problem classes. Similarly, Fig. 4 illustrates the changes in the shipper's annual expected costs and service levels as the freight maximum increases from $0.5Q^*$ to Q^* for different problem classes.

It can be observed from Figs. 3 and 4 that an LTL carrier's freight minimums and maximums hinder the shipper's efficiency and diminish the service quality. Indeed, freight minimums and maximums mean less flexibility for a shipper, hence if the shipper has to use the LTL carrier and obey the carrier's freight minimum and maximum, the shipper might need to deviate from its optimal operational plans which, in turn, reduces the shipper's efficiency and service quality. Finally, it can be noticed that the effects of freight minimums and maximums are similar for different characteristics of the shipper's annual demand (i.e., different problem classes).

Conclusion

This chapter has discussed the basics of carrier selection for LTL shipments. In particular, LTL carrier selection is crucial for the success of a business and LTL carrier selection affects the performance of strategic, tactical, and operational planning decisions in a supply chain. A shipper should pay attention to the working principles of LTL carriers and understand what drives LTL shipment rates, in order to effectively manage its LTL shipments. Nonetheless, shipping rates should not be the sole focus in selecting an LTL carrier. Typically, shippers regard multiple attributes simultaneously while evaluating potential LTL carriers. These attributes are related to the costs and billing, transit time, shipment restrictions, service quality, customer service, and other aspects of the services offered by LTL carriers. Of course, the importance of each specific attribute will depend on the shipper; therefore, the shipper should also understand the consequences of its LTL shipment decisions on the overall operations of its business.

While some of the LTL carrier attributes can be quantified, some of them are less tangible. Through numerical analyses using an inventory control model, the effects of several key LTL carrier attributes on a shipper's expected costs and service level have been demonstrated. Such inventory control models should be integrated with LTL carrier selection for cost efficiency and improved service levels. Furthermore, it is worth noting that there exist various multicriteria decision-making tools, such as the analytic hierarchy process, technique for order preference by similarity to ideal solution, and multiobjective optimization, that can be used for selecting LTL carriers. LTL shippers should carefully use such tools and try to include as much detail as possible in selecting carriers for their LTL shipments.

References

- Bokher, S., 2018. Segments of U.S. Trucking Industry. Available from: <https://medium.com/>.
- Cerasis, 2016. The Less-Than-Truckload Guide: From the Basics to Best Practices for Complete Mastery. Cerasis, Minnesota, United States. Available from: <https://cerasis.com/wp-content/uploads/2016/03/Less-Than-Truckload-E-Book.pdf>.
- Engelbrethsen, E., Dauzere-Peres, S., 2019. Transportation mode selection in inventory models: a literature review. *Eur. J. Oper. Res.* 279 (1), 1–25.
- Hadley, G., Whitin, T.M., 1963. *Analysis of Inventory Systems*. Englewood Cliffs NJ, Prentice-Hall.
- Konur, D., Geunes, J., 2019. Integrated districting, fleet composition, and inventory planning for a multi-retailer distribution system. *Ann. Oper. Res.* 273 (1–2), 527–559.
- Lambert, D.M., Lewis, M.C., Stock, J.R., 1993. How shippers select and evaluate general commodities LTL motor carriers. *J. Bus. Logist.* 14 (1), 131–143.
- Meixell, M.J., Norbis, M., 2008. A review of the transportation mode choice and carrier selection literature. *IJLM* 19 (2), 183–211.
- Nahmias, S., 2009. *Production and Operations Analysis*, sixth ed. McGraw-Hill Irwin, New York, NY.
- Ozkaya, E., Keskinocak, P., Joseph, V.R., Weight, R., 2010. Estimating and benchmarking less-than-truckload market rates. *Transport. Res. E* 46, 667–682.
- Premeaux, S.R., 2009. The similarity of motor carriers' and shippers' perceptions of the carrier choice decision improve. *JTRF* 48 (1), 39–47.
- Robinson, A., 2018. Less-Than-Truckload Market Analysis Insight Q4 2018 & 2019. Available from: <https://www.supplychain247.com/>.
- Vega, D.S.D.L., Vieira, J.G.V., Toso, E.A.V., de Faria, R.N., 2018. A decision on the truckload and less-than-truckload problem: an approach based on MCDA. *Int. J. Prod. Econ.* 195, 132–145.
- Williams, Z., Garver, M.S., Taylor, G.S., 2013. Carrier selection: understanding the needs for less-than-truckload shippers. *Transp. J.* 52 (2), 151–182.

Decarbonizing Road Freight Transport

Heikki Liimatainen, Tampere University, Tampere, Finland

© 2019 Elsevier Ltd. All rights reserved.

Decarbonization Framework	395
Framework Indicators	395
Framework Key Indicators	396
Limitations of the Decarbonization Framework	397
Affecting the Determinants of Road Freight Transport	397
Gross Domestic Product	397
Value Density	398
Modal Split	398
Average Length of Laden Trips	399
Average Load on Laden Trips	399
Empty Running	399
Average Fuel Consumption	399
Fuel CO ₂ Content	400
See Also	401
References	401
Further Reading	401

Decarbonization Framework

Frameworks for analyzing the relationships between the economy and road freight transport have been developed since the mid-1990s, and the frameworks have been expanded since to include the environmental effects and also monetary valuation of the environmental effects for determining the external costs of logistics operations. The basic structure of the frameworks has remained similar and included the components of economic development, freight transport performance, and environmental outputs. A detailed framework for illustrating the complexity of decarbonizing road freight transport is used in this article, as shown in Fig. 1.

Framework Indicators

The decarbonization framework disaggregates the link between the economy and the CO₂ emissions of road freight transport into eight indicators. The first indicator is Gross domestic product (GDP) or added value, usually presented using fixed prices to enable time series analysis. GDP is a widely used indicator for national economic output, and it has usually been categorized in the decarbonization framework as an aggregate or output value, but it is actually a variable, which needs to be forecast, unlike the aggregates of the framework, which are calculated as the results of forecast indicator values.

The value density is defined as the ratio of GDP to the total weight of goods transported within a country by all modes of transport. The value density is expressed as the unit €/t. Value density can also be defined as the ratio of GDP to the total weight of goods produced, but many countries lack the data of the weight of goods produced. If such data are available, an indicator called the handling factor, that is, the ratio weight of goods transported and produced, can be used.

The modal split is defined as the percentage of total weight of goods transported by road and other modes. The energy efficiency and CO₂ emissions of other modes than road freight can be studied using a similar framework as in figure for each mode. However, the focus of decarbonizing freight transport is usually on road freight transport because it is the most important mode of freight transport in most countries.

Average length of laden trips expresses the average distance, which trucks travel on one trip. It is calculated by dividing the mileage of laden trips by the number of laden trips. Average length can also be calculated by dividing the road haulage (tkm) with weight of goods transported by road, but this method is slightly misleading, as the payload and type of trip (long haul or pick up/distribution round) affect the road haulage.

The fifth indicator in the decarbonization framework is the average load on laden trips, which is expressed in tons. It is calculated by dividing the weight of goods transported by road with the number of laden trips. This actually gives the value for the average maximum load on laden trips, that is, the changes in the load during a pick up/distribution trip is not taken into account. Changes in the load during a trip can be taken into account if the average load is calculated by dividing the road haulage by the mileage of laden trips. The average load on laden trips can be disaggregated into vehicle utilization rate, or 'load factor', and the average maximum capacity of trucks. The vehicle utilization rate is the ratio of actual load to maximum load.

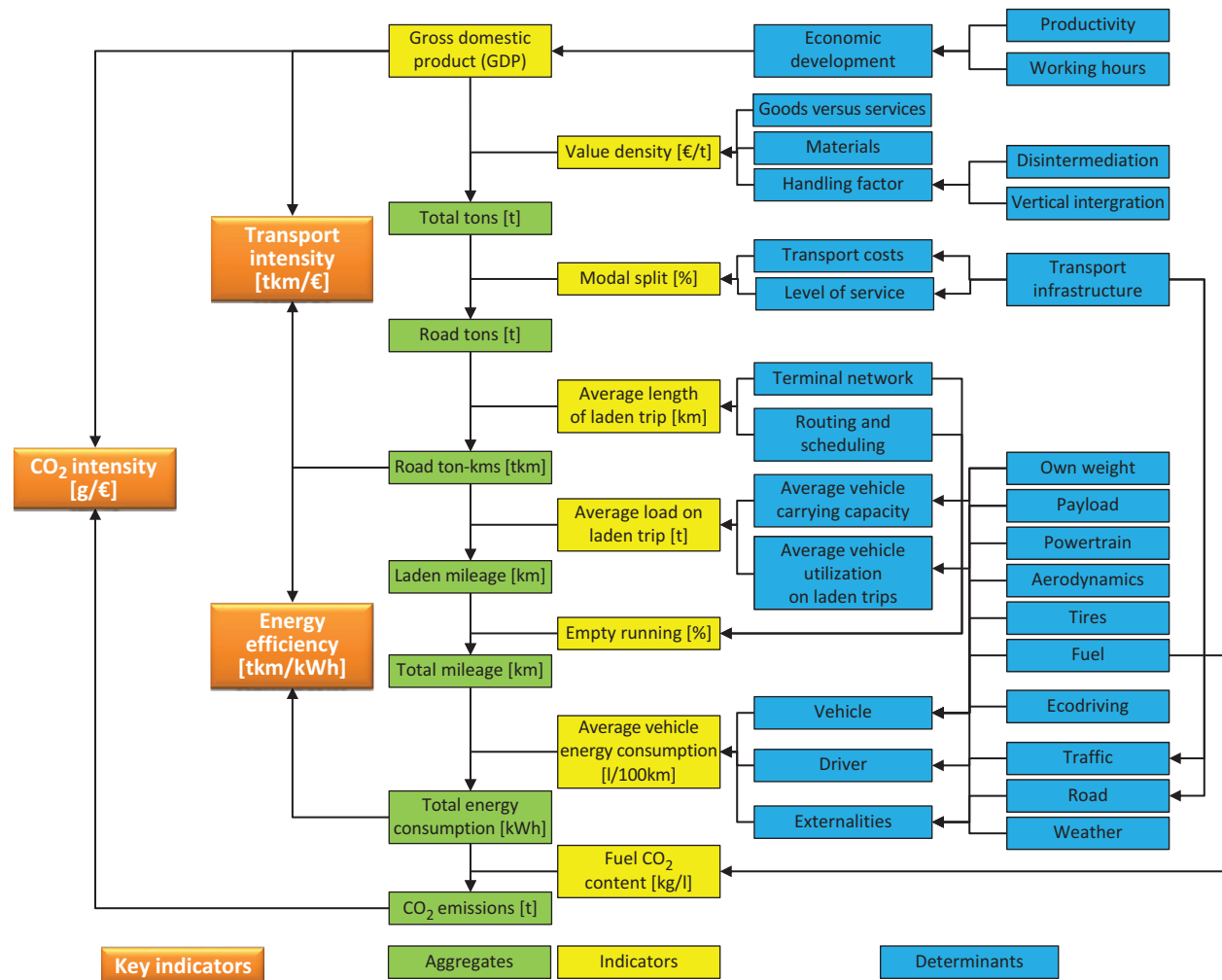


Figure 1 Road freight decarbonization framework.

Empty running is the percentage of total mileage run without a load. It is a characteristic feature of road freight transport as goods almost never return loaded to the point of origin. Average load and empty running are sometimes analyzed together as a single vehicle utilization indicator. One such indicator, is the “efficiency of vehicle usage” indicator, which is calculated by dividing ton-kilometers (tkm) by mass-kilometers, with mass meaning the sum of the vehicle’s own weight and the payload (Léonardi and Baumgartner, 2004). However, some valuable information may be lost by doing this, because even though empty running and average loading are interrelated, they are also determined by separate affecting factors.

Average fuel consumption is the amount of fuel needed for the trucks to travel a certain distance. The unit l/100 km is often used, but fuel economy can also be expressed using the inverse ratio of miles per gallon (mpg). Average fuel consumption is the result of a very complex system of, for example, engine, vehicle design, driving behavior, vehicle loading, and traffic conditions. There are usually no direct data available in road freight statistics on the fuel consumption, so it has to be estimated separately.

The last indicator in the decarbonization framework is the fuel CO₂ content, which expresses how much carbon dioxide is emitted when burning 1 L of fuel. In many countries, the fuel used in trucks is virtually solely diesel. Diesel has a fixed CO₂ content of approximately 2.66 kg/L. Biodiesel or some other alternative fuels may replace some or all of diesel and change the CO₂ content.

Framework Key Indicators

In addition to the eight indicators, three key indicators are defined. These key indicators enable analysis of the issue at a more aggregate level and can be used especially in decoupling analysis. On the most aggregate level, CO₂ intensity can be analyzed to find out whether decarbonization, that is, the decoupling of road freight transport CO₂ emissions from economic growth, has occurred. CO₂ intensity is the ratio of road freight CO₂ emissions to GDP (g/€), so decreasing CO₂ intensity means decarbonization has occurred. However, usually some additional information about the reasons for decarbonization is wanted and the simplest way of

doing this is by introducing the key indicators of transport intensity and energy (or CO₂) efficiency. Until recently, the changes of energy and CO₂ efficiency were the same, due to the fixed CO₂ content of diesel, but alternative energy is increasingly used and needs to be taken into account in the future.

Transport intensity is the ratio between road haulage (tkm) and economic output (GDP). It expresses the changes in the demand for road freight transport in the economy. The term immaterialization can be used of a situation, where decoupling of road haulage from economic growth occurs. Energy efficiency expresses the changes in the efficiency of the supply of road freight transport. Also the term dematerialization can be used to describe the decoupling of transport CO₂ emissions (or energy consumption) from road haulage. Energy efficiency is defined as a ratio between an output of performance, service, goods, or energy and an input of energy. The energy efficiency of road freight transport is thus generally the ratio between road haulage and energy consumption, indicated as ton-kilometers per kilowatt-hours. This can also be turned the other way around to yield energy intensity (kWh/tkm). Other possibilities for indicating the same subject include energy intensity (MJ/tkm), fuel efficiency (koe/tkm, where koe means kilograms of oil equivalent) and emissions efficiency (g CO₂/tkm), as well as CO₂ efficiency (tkm/kg CO₂). All these indicators are interdependent as the current major fuel of road freight vehicles, diesel, has fixed energy content (approximately 10 kWh/L, 36 MJ/L, or 0.87 koe/L) and produces a fixed amount of CO₂ (2.66 kg/L) when burned in the engine.

Limitations of the Decarbonization Framework

The decarbonizing framework disaggregates the relationship between the economy and CO₂ emissions into indicators, which can be analyzed to find out the causes of changes. However, by doing that, some complexity may be lost and one should be cautious not to lose sight of the various feedback loops between the indicators. For example, the value density affects the modal split, as high value goods are more often transported by road or air, and the average load on laden trips and share of empty running affects the average fuel consumption. While the framework includes the modal split, other modes of transport than road freight are omitted from the framework, but similar analysis can be made of other modes and changes in indicators in other modes can affect road freight. The geographical scope of the study is also an important issue. It can be seen from the various studies that analyses have mostly been made at a national level, as the data are best available for that scope. There are limited and scattered data available for freight analysis because of a variety of units for measuring freight movements, confidentiality issues, various decision makers involved in a shipment and different types of loads. The solution is to combine both quantitative and qualitative data from several sources.

The framework has mainly been utilized for studying the changes that have affected the road freight transport sector. It has also been used somewhat, and increasingly, for forecasting the future development of road freight transport and its environmental effects, mainly CO₂ emissions. Forecasts of the CO₂ emissions of road freight transport should be produced by combining various forecasting methods, such as focus group discussions, Delphi survey, scenarios, and spreadsheet modeling, with the framework. This is because the accuracy of forecasts that rely on trend extrapolation or linking freight transport with economic development is considered low. Extrapolation is useful only in stable conditions with continuous trends, but road freight transport has changed over the past decades and is likely to change in the future due to, for example, globalization and climate change mitigation. Hence, the elasticity of freight transport to GDP changes over time in relation to the maturity of commercial exchange between countries. This has to be taken into account when building long-term business-as-usual scenarios, but on short-term scenarios elasticity might have only a minor effect.

Affecting the Determinants of Road Freight Transport

There is a wide variety of measures for decarbonizing road freight transport, because each indicator of the decarbonization framework is affected by various determinants, which in turn may be affected in various ways to change the indicator toward a more sustainable direction.

Gross Domestic Product

The necessity of economic growth (as measured by GDP) has been the dominant paradigm in politics since the Second World War. GDP is widely identified with social welfare and hence the need for growth is taken for granted. However, there is a growing critique toward GDP for being inadequate in capturing human welfare and the environmental and social problems, which accrue from continuous growth. Several alternatives for GDP have been developed, including, for example, the index of sustainable economic welfare, sustainable national income, genuine savings, and the human development index. Economic and social concerns have led to the emerging idea of degrowth, which is defined as an equitable downscaling of production and consumption that increases human well-being and enhances ecological conditions at the local and global level, in the short and long term.

The downscaling of production and consumption of goods results in decreasing GDP and would decrease the demand for road freight transport. However, it seems highly unlikely that any national government would take downscaling of production and consumption as a policy measure for decarbonizing road freight transport.

Value Density

Decarbonizing road freight transport by increasing the value density is politically much more acceptable than by decreasing GDP. When value density is considered, it is essential to distinguish between changes at a national and a global scale. At both the national and global scale, decarbonization through increasing value density may occur if:

- lighter materials are used to produce same goods,
- smaller goods are produced to fulfill the same purpose,
- goods are substituted with services, and
- the effect of these changes is greater than the growth of consumption (GDP).

Offshoring of manufacturing to other countries while maintaining the headquarters of the company in another decreases the weight of goods produced and transported in the country of headquarters, but maintains the level of GDP generated in that country and maintains the weight of goods produced at a global scale. The effect of offshoring on the global total haulage (tkm) is yet another matter, as it depends on the locations of raw materials, production and final consumption, and thus the handling factor (ratio of weight of goods transported and produced) and the average length of trips. The handling factor can be reduced through:

- disintermediation, that is, eliminating links in supply chains;
- vertical integration, that is, locating interrelated processes on the same site; and
- decreasing intermodality, that is, eliminating road feeder movements by connecting or relocating the site to railroad, port, or airport.

Modal Split

Modal split affects the energy efficiency and CO₂ emissions of road transport, as it defines what kinds of goods are transported in each transport mode. Waterway and rail transport are generally considered to be more energy efficient than road transport, but this view is also questioned. The energy efficiency is highly dependent on the utilization of payload capacity in each mode. With average loading, rail transport has the lowest CO₂ emissions per ton-kilometers, closely followed by coastal shipping and inland waterways, whereas trucks have considerably higher emissions and vans far higher still (Fig. 2).

No mode of freight transport can be categorically designated as the most environmentally friendly. The potential for decarbonization through modal shift depends on:

- energy efficiency of the mode,
- carbon intensity of the energy source by mode,
- utilization of equipment, and
- energy use of modal interchange.

This has to be taken into account when the energy efficiency and CO₂ emissions of different modes are evaluated. It is desirable from the decarbonization perspective to promote co-modality, the efficient use of transport modes on their own and in combination. If sufficient decarbonization potential is available, modal shift can be promoted through infrastructure investments and direct support for rail and waterways equipment purchase and operations.

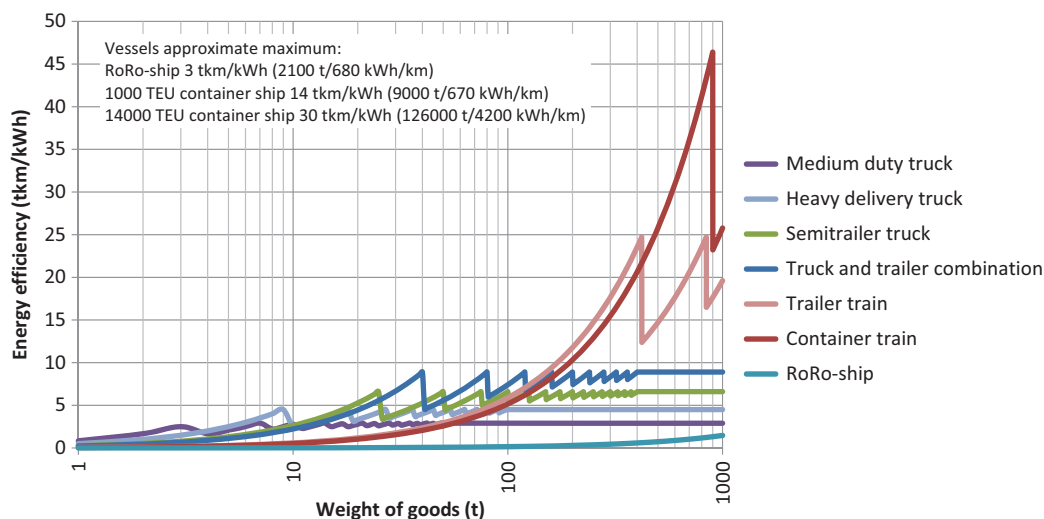


Figure 2 Energy efficiency of various freight transport modes. Source: Based on VTT (2019).

Average Length of Laden Trips

The average length of laden trips is mainly affected by the geographical scale of sourcing and market areas of production, as well as the location of production sites. The globalization of sourcing and markets, with centralization of production and warehousing, has been the major trend in the global economy in recent years. These changes may decrease or increase the average length of haul at a national scale, and it is very difficult to influence these changes. Furthermore, the changes in average length of haul influence the modal split, vehicle loading, and the level of empty running, so it is very difficult to implement decarbonizing policy measures directed at the average length of laden trips. At haulier level, decreasing the average length of laden trips is possible by improving the vehicle routing and scheduling, possibly with the help of optimization software (computerized vehicle routing and scheduling). Also relocating distribution centers, using a hub and spoke network, and participating in open cooperation networks are possible decarbonization measures, which also affect the vehicle loading and empty running.

Average Load on Laden Trips

Increasing the average load on laden trips increases the fuel consumption on the basis of l/100 km, but decreases the mileage required to perform the same haulage (tkm), and thus improves the energy efficiency (tkm/kWh) and decreases the total CO₂ emissions. Average load is mostly considered in terms of weight, although in many cases the area or volume of the cargo space is the limiting factor, rather than maximum weight. However, data on volumetric vehicle fill are rarely available at a sufficient level of detail for analysis. Average load may be disaggregated further into maximum payload weight and vehicle utilization rate. This disaggregation may be helpful, as the largest trucks are not suitable for every operation (e.g., in urban areas) and, hence, the changes in the average load may be caused by necessary changes in average maximum payload weight of the fleet.

Opportunities for decarbonizing road freight transport through increasing average load on laden trips are numerous, but mainly constrained by the interfunctional relationships between transport and other business activities. Increasing the transport cost through fuel taxation or road user charging may be used as policy measures to increase the importance of vehicle utilization. Other policy measures include allowing larger trucks and relaxing access restrictions, as well as promoting better vehicle loading through awareness campaigns and benchmarking information.

Policy measures can only guide companies to improve their vehicle utilization, but better results are gained if companies realize the potential cost savings and implement measures such as:

- nominated day delivery system, which decreases the inefficiency caused by demand fluctuations;
- more space efficient cargo handling equipment and packaging, which reduce the need for packaging without compromising the protection of goods from damage;
- double deck trailers, which allow stacking pallets or roll cages to better utilize the maximum payload weight;
- intra- and intercompany cooperation, which decreases the inefficiency caused by geographical imbalance of goods flows and lack of interfunctional coordination; and
- lightweight vehicles, which decrease the own weight of the vehicle and increase the maximum payload weight for transporting more high-density goods.

Empty Running

Most reasons for inefficiency and decarbonization measures mentioned for the average load on laden trips also apply for empty running, but empty running is even more dependent on the level of intercompany cooperation. The potential for increasing backloading and reducing empty running is not fully utilized, quite simply because there is a lack of knowledge of available loads. Even if there is such knowledge, backloading may not be realized because outbound delivery is prioritized and there is an increased risk for delays with backloading. Also, the vehicles or cargo handling equipment may not be suitable for the backloading products, or the driver's working hours limit the possibilities for backloading.

Intercompany cooperation, through outsourcing of freight operations to a logistics service provider or through establishing a network of hauliers, improves the sharing of knowledge on backloads. Also the wider use of web-based tendering of freight transport services improves the level of knowledge. Increased visibility of freight operations and cooperative efforts for avoiding delays can be used to develop confidence between the shipper and haulier and, thus, make backloading more acceptable. Empty running also decreases when reverse logistics, such as the recycling of packaging waste, reuse of handling equipment, or refurbishment of products, increases. Transport policy may affect empty running by increasing the cost of transport, relaxing the working time and cargo handling restrictions, as well as by standardizing cargo-handling equipment.

Average Fuel Consumption

In addition to the vehicle loading, the fuel consumption is determined by three main factors: traffic conditions, vehicle specifications, and driver behavior. Traffic conditions include road geometry and traffic flow. Road geometry affects fuel consumption mostly in hilly terrain, but winding roads may cause braking and acceleration, which increases fuel consumption. Traffic flow is affected by the number and behavior of other road users and by the regulation of traffic flow, that is, traffic lights, speed limits, etc. The effects of

Table 1 Fuel saving measures and approximate potential savings

Aerodynamics	Under-run air dam	<1%
	Cab roof fairing	4%
	Cab side edge turning vanes	<1%
	Body/trailer side panels	1%
	Body/trailer front fairing	3%
	Tipper sheeting systems	<1%
	Teardrop trailer	10%
	Sloped roof trailer	5%
	Reduce height of vehicle	3%
	Louvered spray suppression flaps	3%
Tires	Fuel efficient tires (all axles)	3%
	Super single tires	2%
	Tire pressure management	2%
	Regrooving tires	1%
	Wheel alignment	2%
Other	Synthetic engine oil	2%
	Anti-idling campaign	2%
	Speed limited to 84 instead of 90 km/h	<1%
	Reduce vehicle tare weight	1% per 1 t

Source: Based on *FTA (2012)*, *RICARDO (2009)*, and *RASTU (2009)*

road geometry are minor compared to the effects of irregular traffic flow. Hence transport policy may decarbonize road freight transport by investing in road infrastructure, improving road traffic management, introducing road user charges, and relaxing restrictions for night deliveries. Hauliers, on the other hand, may use dynamic vehicle routing to avoid congested roads and negotiate with their customers to reschedule the deliveries.

Vehicle specifications affect the fuel consumption in numerous ways. First, there are significant differences in the fuel consumption between the new trucks of different brands. There are several possibilities for improvements in truck engine and transmission, such as, downsizing, supercharging, and automated manual transmission, which when combined result in substantial fuel savings. The improvement with greatest fuel saving potential is the hybrid electric powertrain. Truck manufacturers are responsible for the development of truck powertrains, but the hauliers may decarbonize their operations by purchasing trucks with low consumption and the government can help hauliers to make the right decision by introducing fuel consumption standards. In addition to engine and transmission, there is a variety of measures which the haulier can implement to achieve considerable fuel savings. **Table 1** summarizes some of these measures and their approximate potential fuel savings.

Driving behavior has a great effect on fuel consumption. The difference between drivers can be up to 30%. Because of this, many companies have implemented ecodriving training and gained short-term fuel savings. However, the effects of ecodriving training often decrease in the long term if the driving behavior is not monitored and feedback is not given. Regular monitoring can also include an incentive system for drivers. However, there are challenges facing incentive systems that are operated within transportation companies, such as the complexity of implementation, lack of accurate consumption data, and lack of information on how to operate such a system. Monitoring driver performance fairly is a complex matter, because weather, traffic, road geometry, vehicle and load carried are all constantly changing, independent of a driver's actions. However, each can have a considerable effect on fuel consumption. Nevertheless, current onboard computers enable taking these issues into account to establish a fair incentive system. Policy makers can promote ecodriving through awareness campaigns, mandatory ecodriving as part of driving license training and fiscal incentives for the purchase of onboard monitoring equipment.

Fuel CO₂ Content

The final indicator in the framework is the carbon dioxide content of fuel. Road freight transport may be decarbonized by using alternative fuels such as renewable diesel, natural gas, or electricity. Electric vehicles depend on batteries, which are heavy and also quite large, but still have very limited range. These characteristics make them unsuitable for transporting heavy goods over long distances, but possible to exploit in urban distribution operations and the long haul of light goods. However, the decarbonizing effect of electric vehicles depends on the source of electricity.

Natural or biogas vehicles have a greater range than electric vehicles, but require large investments in distribution infrastructure. Natural gas has about the same CO₂ content as diesel and thus does not have a decarbonizing effect. Biogas, on the other hand, reduces methane, which would otherwise be emitted from the waste disposal process and thus reduces greenhouse gases, but the production of biogas is still very limited.

Renewable diesel is currently the most viable option for decarbonizing road freight transport through decreasing the CO₂ content of fuel. Renewable diesel may be produced from various sources, including vegetable oil, frying oil and animal fat, and the CO₂ reduction depends on the source. The well-to-wheel GHG emissions savings from renewable diesel are typically 30%–90%

compared to fossil diesel. However, the savings may be much smaller when the changes in land use are taken into account. Nevertheless, renewable diesel has advantages as an alternative to diesel. First, it is partly or completely compatible with the current diesel engines, and it uses the same distribution infrastructure. Second, it can be produced from waste or cellulosic crops, so it does not compete with food production or require land use change. Because of these issues, renewable diesel is already widely used blended with fossil diesel and its use is likely to increase. The political pressure for increasing the use of renewable diesel is certainly great due to its benefits in decreasing oil dependency and increasing the number of sources of energy and domestic self-sufficiency in energy.

See Also

Demand for freight transportation; Freight costs, road transport; Valuation of carbon emissions; Environmental sustainability in freight transportation; Green routing of freight vehicles; Freight mode choice; Road freight vehicles; The rebound effect in road freight transport; Transport modes and sustainability; Energy consumption of transport modes

References

- FTA, 2012. Carbon Intervention Modelling Tool. Freight Transport Association, United Kingdom.
- Léonardi, J., Baumgartner, M., 2004. CO₂ efficiency in road freight transportation: Status quo, measures and potential. *Trans. Res. D Trans. Environ.* 9 (6), 451–464.
- RASTU, 2009. Heavy-duty vehicles: Safety, environmental impacts and new technology. Final report, Kytö, M., Erkkilä, K., Nylund, N.-O. (Eds.) (in Finnish). Available from: http://www.motiva.fi/files/2278/RASTU-loppuraportti_2006-2008.pdf.
- RICARDO, 2009. Review of low carbon technologies for heavy goods vehicles. Prepared for Department for Transport. Available from: <http://www.lowcvp.org.uk/assets/reports/090715%20Review%20of%20low%20carbon%20technologies%20for%20heavy%20goods%20vehicles.pdf>.
- VTT, 2019. LIPASTO Unit Emissions Database. Available from: <http://lipasto.vtt.fi/yksikkopaastot/indexe.htm>.

Further Reading

- Alises, A., Vassallo, J., Guzman, A., 2014. Road freight transport decoupling: a comparative analysis between the United Kingdom and Spain. *Trans. Policy* 32, 186–193.
- Aronsson, H., Hüge Brodin, M., 2006. The environmental impact of changing logistics structures. *Int. J. Logist. Manag.* 17 (3), 394–415.
- Kamakate, F., Schipper, L., 2009. Trends in truck freight energy use and carbon emissions in selected OECD countries from 1973 to 2005. *Energ. Policy* 37 (19), 3743–3751.
- Liimatainen, H., Kallionpää, E., Pöllänen, M., Stenholm, P., Tapio, P., McKinnon, A., 2014. Decarbonising road freight in the future – detailed scenarios of the carbon emissions of road freight transport in 2030. *Technol. Forecast. Soc. Change* 81, 177–191.
- Liimatainen, H., van Vliet, O., Aply, D., 2019. The potential of electric trucks – an international commodity-level analysis. *Appl. Energ.* 236, 804–814.
- McKinnon, A., 2018. Decarbonising logistics: Distributing Goods in a Low Carbon World. Kogan Page, London.
- McKinnon, A., Cullinane, S., Browne, M., Whiteing, A. (Eds.), 2010. *Green Logistics: Improving the Environmental Sustainability of Logistics*, Kogan Page, London.
- McKinnon, A., Ge, Y., 2006. The potential for reducing empty running by trucks: a retrospective analysis. *Int. J. Phys. Distrib. Logist. Manag.* 36 (5), 391–410.
- McKinnon, A., Piecyk, M., 2009. Measurement of CO₂ emissions from road freight transport: a review of UK experience. *Energ. Policy* 37 (10), 3733–3742.
- Nahlik, M., 2016. Goods movement life cycle assessment for greenhouse gas reduction goals. *J. Indust. Ecol.* 20 (2), 317–328.
- OECD/ITF, 2017. *ITF Transport Outlook 2017*. OECD Publishing, Paris.
- Sorrell, S., Lehtonen, M., Stapleton, L., Pujol, J., Champion, T., 2012. Decoupling of road freight energy use from economic growth in the United Kingdom. *Energ. Policy* 41, 84–97.
- Tapio, P., Banister, D., Luukkanen, J., Vehmas, J., Willamo, R., 2007. Energy and transport in comparison: Immaterialisation, dematerialization and decarbonisation in the EU15 between 1970 and 2000. *Energ. Policy* 35 (1), 433–451.

The Rebound Effect in Road Freight Transport

Tooraj Jamasb, Manuel Llorca, Copenhagen School of Energy Infrastructure (CSEI), Copenhagen Business School, Frederiksberg, Denmark

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	402
Brief Historical Background	402
What is the Rebound Effect?	402
The Rebound Effect in Road Freight Transport	403
Policy Aspects	405
Acknowledgment	406
References	406

Nomenclature

EPA Environmental Protection Agency
ESI Energy Systems Integration
EU European Union
EV Electric Vehicle
GHG Greenhouse Gas
IPCC Intergovernmental Panel on Climate Change
NHTSA National Highway Traffic Safety Administration
US United States

Brief Historical Background

The rebound effect as a phenomenon was first presented by the English economist William Stanley Jevons in his book “The Coal Question” in 1865. He observed that technological improvements resulted in more efficient utilization of coal, but this also led to an increase in the consumption of this fuel in several industries during the 18th and 19th centuries (Jevons, 1865). This unforeseen effect, initially known as the “Jevons Paradox,” was revisited in the 1980s by both J. Daniel Khazzoom and Leonard G. Brookes, who discussed the issue independently. They stressed that, at a macroeconomic level and contrary to what common sense suggests, energy efficiency gains (or in the words of Leonard G. Brookes, “reductions in energy intensity of output”) that are not harmful to the economy, are linked with increases in energy consumption (Khazzoom, 1980; Brookes, 1990).

The validity of the Khazzoom-Brookes hypothesis was fiercely debated in a series of academic articles in the 1990s (Brookes, 1990,1992; Grubb, 1990,1992). In 1992, Harry D. Saunders demonstrated that under certain assumptions the postulate is consistent with results directly obtained from neoclassical growth theory. He identified two channels through which the rebound effect may arise (Saunders, 1992). First, energy efficiency improvements result in energy appearing cheaper than other inputs, which implies that energy will substitute other input factors such as labor. Second, increases in economic growth lead to an additional boost in energy consumption. Since then, despite these theoretical underpinnings, the scale and relevance of the rebound effect have remained contentious and spurred by the development of an extensive empirical literature.

What is the Rebound Effect?

From a theoretical point of view, the rebound effect describes the potential energy savings that are offset when there are energy efficiency enhancements. In other words, rebound effects imply that energy efficiency improvements may lead to less than proportional reductions in energy consumption. In a general and simplified way, an energy efficiency improvement can be viewed as a reduction in energy use while the demand for energy services is maintained and the comfort and the quality of life are not diminished. Full potential energy savings are usually mechanically derived by assuming that the demand for energy services remains unchanged after energy efficiency gains. However, energy efficiency enhancements implicitly imply a reduction in the marginal cost of providing a given level of energy services, which may lead to an increase in demand for those services and consequently to an increase in energy consumption. It is possible that this reaction may partially or fully offset the initial expected energy savings, often described as the rebound effect or take-back. If the magnitude of the rebound effect is non-negligible and the phenomenon is overlooked, the benefits from energy savings may be overstated and policies aimed at reducing energy consumption through promotion of energy efficiency may not be fully effective. This in turn may lead to decisions such as the (over) allocation of public funds to ineffective environmental and energy policies.

From consumer choice theory in microeconomics, the rebound effect can be understood as a combination of income and substitution effects (Gillingham et al., 2016). The substitution effect refers to a situation in which part of the demand for one good or service is replaced by the demand for another good or service simply due to a change in their relative prices. The income effect also implies a change in the consumption pattern of goods and services, but it arises from the increase in consumer purchasing power due to the price reduction. The literature on the rebound effect has often attempted to pin down the definition and implications of the effect through distinguishing at least three types of them, namely, direct, indirect, and economy-wide rebound effects (Sorrell and Dimitropoulos, 2008).

The most apparent effect is the so-called direct rebound effect. This type of rebound implies that an improvement in energy efficiency for a particular energy service decreases the effective price of that service and provides incentives to increase the demand for the service. Therefore, the reactions of individuals tend to offset the expected energy savings that could be attributed to energy efficiency improvement. The indirect rebound effect arises from energy savings due to energy efficiency enhancement when they are reverted into demand for other goods and services that require energy for their provision. The economy-wide rebound effect relates to the reduction in the price of intermediate and final goods in the economy due to the decrease in real price of energy services. This results in adjustment of the prices and quantities of goods and services consumed in the economy. These adjustments create a bias toward consumption of the goods in the most energy intensive sectors, which may then yield an increase in overall energy consumption.

The rebound effect is generally measured as the percentage of the potential energy savings that are not achieved. They normally take values between 0% (zero rebound effect) and 100% (full rebound effect). However, rebound effects can also take values larger than 100%, labeled as “backfire” in the literature (Saunders, 1992), which means that improvements in energy efficiency can lead to a higher level of energy consumption than before the efficiency improvement. The opposite case, that is, a negative rebound effect, which can seem counterintuitive, is labeled as super-conservation (Saunders, 2008). This rebound effect represents a reduction in energy consumption that exceeds the initial expected savings. The diverse definitions, set of methodologies applied, data, sectors and countries analyzed, and the resulting broad range of results obtained in the empirical literature have helped to keep the debate about the rebound effect open.

It is worth noting that some researchers consider the rebound effect as a natural adjustment to changing economic factors. In essence, this means that the rebound effect can be considered as a reoptimization process in response to price and income variations (Borenstein, 2015). From that perspective and using standard economic analysis, the rebound effect can be seen as a welfare improvement. However, it should be noted that in order to assess the net impact of the rebound effect on overall economic welfare, the external costs generated by this phenomenon (e.g., through GHG emissions) should also be incorporated in the analysis.

The Rebound Effect in Road Freight Transport

The literature on the rebound effect has focused both on the analytical definitions of the concept and its empirical measurement using alternative approaches generally based on elasticities. Elasticities are measures, often used in economics, to represent the proportional change of a specific variable in relation to the change of another variable. A basic description of the rebound effect can be proposed by using the elasticity of the demand for energy services with respect to energy efficiency. This elasticity represents how demand for energy services change (and hence, also the energy demand) when there is an increase in energy efficiency. The estimation of this type of elasticity is contentious because data on either energy services and energy efficiency are often unavailable or flawed in terms of accuracy. In road freight transport, measures such as vehicle or ton kilometers are frequently used for energy services. However, independent estimates of fuel efficiency are normally difficult to use due to lack of information or insufficient variation within the data.

Instead, own-price elasticities of the demand for energy have frequently been econometrically estimated and used as indirect measures or proxies for the rebound effect. A great majority of studies assume that the response to changes in energy price is equal to the response in changes to energy efficiency. However, it has been shown that, especially in the case of transport, this approach is likely to overestimate the rebound effect due to the endogeneity between energy prices and efficiency choice (e.g., through increasing load factor or purchasing more efficient vehicles), since these two factors are not independent from each other. Therefore, price-induced energy efficiency improvements should not be neglected when analyzing rebound effects to avoid biased estimates. Other likely sources of bias, when rebound effects are econometrically estimated, are the overlooking of trade-offs between energy and time costs, the analysis of rising energy price periods only (which are generally linked with high energy price elasticities) and the correlation between changes in energy efficiency and changes in input costs such as capital costs (Sorrell and Dimitropoulos, 2008).

The empirical research on the rebound effect in the transport sector has yielded a broad range of results both for the short and the long run. However, most of these results have been estimated to be between 10% and 30% (Sorrell, 2007; Hymel et al., 2010). The wide range of results has been a consequence of the differences in the definitions of the rebound effect, the type of elasticities and the approaches used, the countries analyzed and the level of aggregation in the data used. Nonetheless, as it has been remarked by some authors, there is still little research on the rebound effect in road freight, despite the need to assess the effectiveness of the policies adopted to reduce fuel consumption in the sector.

Recent surveys show that, on average, rebound effects in road freight transport are estimated to be between approximately 0% and 40% in the short run, while in the long run the values range between 10% and 85%. This is not unexpected in economic analysis. Elasticities are often higher in the long run than in the short run because changes and adjustments, especially in the stock of capital, are more likely to

happen over longer periods of time. However, these results reveal again that even when we restrict the rebound effect analysis to road freight transport, there is still an ample range of results arising from the diverse data, countries, and approaches used.

For instance, the use of a theoretical framework taken from thermodynamics and evolution, in conjunction with a traditional economic approach, has been proposed to explain the rebound effect (Ruzzenenti and Basosi, 2008). This methodology has identified an increase in the energy efficiency of road freight transport in Europe of 40% between the second half of the 1970s to the first half of the 1990s. The main argument is that, in more complex systems, there is a greater offset of the positive effect from higher energy efficiencies. Network theory has also been proposed to study the direct rebound effect in the European road freight transport sector (Ruzzenenti and Basosi, 2017). This approach provided negative rebound effects, that is, “superconservation,” indicating that energy efficiency improvements, along with energy prices, proved to be successful in reducing energy consumption in Europe in context of increasing demand for freight transport services.

European road freight transport has been analyzed using a stochastic frontier analysis approach taken from the efficiency and productivity analysis literature (Llorca and Jamasb, 2017). The results indicate that the achieved energy efficiencies are retained to a large extent. The average rebound effect is found to be 4%, but with large differences across the EU countries. Moreover, the rebound effect is found to be larger in countries with a higher share of trucks with respect to the total number of vehicles in the sector, as well as a better quality of transport infrastructure. In addition, the existence of a strong railway freight transport sector seems to be linked with a lower rebound effect. It is also shown that significant costs associated with environmental externalities can arise, even in countries with a small rebound effect, due to the large scale of their road freight transport activity.

Computable general equilibrium models have been frequently utilized in the literature to obtain economy-wide measures of the rebound effect. For instance, this approach has been used to analyze the oil rebound effect from energy efficiency improvements in the Scottish commercial transport industry (Anson and Turner, 2009). A rebound effect of 36.4% was found in the short run and 39.2% in the long run for the sector. At economy-wide level, the rebound effect was found to be 36.5% in the short run and 38.3% in the long run. The study also found a disinvestment effect, which implies that in most cases the long-run effect is constrained by a disinvestment that reduces the productive capacity in the energy sectors.

The case of demand for road freight transport in Portugal has been analyzed using a two-stage least squares model (Matos and Silva, 2011). Using aggregate time series data for the period 1987–2006, the study considers the changes in the energy cost of transport, while correcting for the endogeneity of the price variable. It shows that a large share of the operating costs in the road freight industry depends on energy consumption, and an increase in fuel efficiency that reduces these costs will in turn result in a significant increase in demand for energy services. The rebound effect is estimated to be approximately 24%.

The fuel used in the trucking industry of Denmark has been studied using aggregate time series data (De Borger and Mulalic, 2012). Using a simultaneous equations model, a rebound effect of 10% in the short run and 17% in the long run was found for the period 1980–2007. Moreover, the study found that higher fuel prices increased the average capacity of trucks. This feature also led the firms to invest in new and more efficient trucks. The joint influence of their findings suggests that the effect of an increase in fuel prices on fuel use is not very relevant. The results show the short-run and long-run price elasticities to be -0.13 and -0.22 , respectively. They also find a reduction in the realized energy efficiency of the trucking industry in Denmark. The study justifies this reduction as being partially due to rising congestion and “Just-in-Time” behavior of firms, which is also a consequence of the reduction in the utilization of vehicle capacity.

In the case of the United States, some recent studies estimate fuel price elasticities for single-unit truck operations and the combination trucking sector in the country (Winebrake et al., 2015a; 2015b). These analyses apply first-difference and error correction models to time series data for the periods 1980–2012 and 1970–2012, respectively. The authors state that their estimates may, under certain assumptions, be used as a proxy for the rebound effect. They find fuel price elasticities that are not statistically different from zero, although in the case of the combination trucking sector this has occurred since the deregulation of the sector in 1980. They offer possible explanations for their findings, such as the lack of incentives for firms to reduce travel or fuel consumption, adjustments in other modifiable operational costs, the “principal-agent” problem or the nature of freight transportation. However, another study of the United States using detailed truck-level microdata from over 100,000 medium- and heavy-duty vehicles estimated a rebound effect of 30% for tractor trailers and 10% for vocational vehicles (Leard et al., 2016). This study suggests that US regulatory agencies (EPA and NHTSA) might have overestimated long-term fuel and GHG savings from their cost benefit analysis of the truck standards applied.

The case of China has been studied using a double logarithmic regression equation and an error correction model to analyze the direct rebound effect in the short run and long run of different regions of the country from 1999 to 2011 (Wang and Lu, 2014). It found a partial rebound effect in the long run between 52% and 84% for different provinces of China. This implies that a large share of the expected reduction in energy demand could not be achieved and the policies intended to improve energy efficiency are not effective. Nevertheless, they find evidence of a slight super conservation effect in the short run.

The long-run direct rebound effect in the UK road freight transport has been analyzed using time series data for the period 1970–2014 (Sorrell and Stapleton, 2018). The analysis uses different model specifications, conducting several tests to check the robustness of the specifications and using diverse elasticities. It found a direct rebound effect of 61%. Using the most robust models, the direct rebound effect was 49%, while the individual estimates ranged between 21% and 137%. The study points to three possible explanations for these results. First, the large sensitivity of the operators to fuel cost, because it represents one-third of total operational costs. Second, the small variation of fuel efficiency and the fuel costs of the goods moved in the country could be causing unreliable results. Third, the increases in vehicle weight limits, which have encouraged more activity in the sector.

Policy Aspects

In recent decades, there has been an increasing interest in the adoption of energy and environmental policies that attempt to stimulate the reduction of energy consumption. The main goals of these policies are to minimize the dependence on fossil fuels and mitigate local air pollution and GHG emissions. This has been particularly relevant for energy-intensive sectors such as transportation. In 2010, freight transport was estimated to account for about 43% of global energy use in transport, which in turn represented 12% of total energy consumption and 10% of energy-related CO₂ emissions (IPCC, 2014). Moreover, in forthcoming years, the amount of fuel used by the trucking industry is expected to rise in the United States and the EU (Léonardi and Baumgartner, 2004; IEA, 2010; De Borger and Mulalic, 2012), where trucks and vans emit 67% of the GHG emissions associated with transport (McKinnon, 2015).

The demand for road fuel is a derived demand for energy services in road transport. This means that there is intermediate demand for transporting goods and people that is met through a combination of capital, labor, and fuel. Increases in the demand for freight transportation has globally driven road fuel consumption. As a counterbalance, the promotion of energy efficiency has become a key energy policy mechanism around the world to tackle the increased fuel consumption and emissions, and countries and regions have adopted different targets and policies to achieve energy and environmental objectives.

In the European Union, where transport accounts for 25% of energy-related GHG emissions (Walnum et al., 2014), the European Commission has passed specific directives for the sector. Since the introduction of the Council Directive 88/77/EEC in 1988 and followed by other legislation, the implementation of “Euro” standards since 1992, and the 2001 White Paper on transport, the energy consumption reduction objectives have been partially achieved. The goal has been to reduce GHG emissions by 80%–95% below the 1990 levels by 2050. However, the Commission has recognized that the specific targets for the transport sector needed to be adjusted downward to 60% due to the complexity of the sector (European Commission, 2011; Walnum et al. 2014).

In the case of the United States, trucks represent 6% of emissions and by 2030 these are expected to be greater than the emissions from light-duty passenger vehicles (Leard et al., 2016). The general approach to limit the emissions of the sector has been through the regulation of GHG emissions rates and fuel economy standards proposed by the EPA and the NHTSA. The implementation of the latest phase of standards (Phase 2) covers the years from 2018–27. These performance-based standards (i.e., they are technology-neutral) are estimated to yield GHG emissions reductions of up to 27% in tractor trucks, up to 24% in vocational vehicles, 16% in heavy-duty pickup trucks and vans, and up to 9% in commercial trailers by 2027 with respect to Phase 1 standards (2014–18).

In general, fuel costs represent approximately one-third of the operating costs in the sector. This implies that, regardless of policy targets, freight operators have incentives to minimize fuel consumption. It is clear that the development and deployment of new and more efficient technologies from vehicle manufacturers is, along with operations/logistics, the main channel to achieve these environmental and energy objectives. However, as it has been discussed in this chapter, energy efficiency improvements can lead to changes in the demand for energy services that offset some of the expected energy savings and consequently, forecasts of energy consumption reduction may be overstated. As evidenced by the empirical literature, rebound effects can be a non-negligible issue in road freight transport. Fuel efficiency improvement reduces the cost of freight transport, which in turn makes transportation cost-efficient for more goods, for longer distances and at greater frequency.

Ignoring rebound effect can overestimate the benefits of energy efficiency policy, which can in turn lead to decisions such as the (over) allocation of public funds to ineffective environmental and energy policies. Policy makers need to take rebound effects into account for air quality, energy security and climate change policy reasons. A rebound effect different from zero implies that the expected proportional reductions in emissions from fuel efficiency improvements might not be achieved. Therefore, the policy goals to reach specific levels of emissions through fuel efficiency enhancements might need to be adjusted accordingly in order to compensate this effect.

Fuel use per unit of energy service in the road freight sector has declined over recent decades and this may be further encouraged in the future through different means. For long-haul transportation, one possibility is the deployment of high capacity vehicles such as megatrucks or European Modular System combinations. Despite the potential for fuel consumption and CO₂ emissions reduction of this type of alternative, the expected savings might not be fully achieved, if the rebound effect generates unexpected modal shifts, in the form of a higher modal share for road transport. At local level, a transition of freight transport from trucks to light vehicles may also be possible, which has been observed for cities such as London in recent years. There are different reasons for this trend, such as the relative lack of regulation governing light vehicle use compared to HGVs, the popularity of e-commerce and Just-in-Time deliveries, the growth in demand for express and parcels services or the impact of congestion (Allen et al., 2016).

Regarding fuels, the main alternatives that may take over the future of long-haul heavy duty are hydrogen, e-fuels (i.e., synthetic fuels from renewable sources) and electrification. In particular, one of the most envisaged future scenarios for a decarbonized sector is one in which EVs or electrified road/transport systems take place. This is happening for the case of urban delivery. An electrified sector could imply a reduction in GHG emissions and local pollutants without any decrease in freight activity and avoiding concerns about rebound effects. However, this solution is likely to be affected by other considerations, such as the capital costs of EVs, investment in infrastructure (e.g., EV charging stations or electric roads) and the management of the electrical grid. Ultimately, energy-mix will be paramount in this paradigm, since for this solution to become effective, electricity should be generated from low-carbon energy sources.

There are other scenarios in which freight transport demand may be affected by the development of new systems or technologies. However, in most cases it is still unclear how likely it is that these developments will effectively happen and their chances of becoming “game changers.” For instance, despite the attention that 3D printing attracts, recent research is still inconclusive about its role and long-term impact on logistics. Another prospect is a setting of highly integrated energy systems, called ESI in the specialized literature. ESI is a relatively new concept that is still under discussion and will require policy and regulatory frameworks to be rethought across different

sectors of the economy. A paradigm, which seems easier to be implemented, is the development of integrated transportation demand management. This encompasses different transportation options and infrastructures to optimize the logistics of all modes in the transport system, which can become a relevant means to reduce and redistribute the transport demand.

Finally, it should be noted that due to the scarcity of empirical studies and limited understanding of the rebound effect in freight transport, further research is still needed. The evolution and magnitude of the rebound effect arising from higher energy efficiency in the sector is relevant from a policy perspective to assess the effect of energy efficiency improvements and national and international energy and climate policies. So far, the results from the literature have evidenced that rebound effects can be of sufficient magnitude to be explicitly taken into account in policy design. These results are of direct policy relevance given that fuel efficiency standards are among the main policy regulations to reduce environmental effects in the commercial transport sector and especially for heavy-duty trucks. In order to assess the effects of fuel economy regulation, information about the rebound effect should be based on analyzing data series of different countries. If efficiency standards fail as pollution control tools, then other measures such as fuel taxes, carbon taxes, feebates, or cap and trade programs could maintain a prominent role in climate policies, in addition to other measures.

Moreover, it should be noted that even when the “relative” magnitude of the rebound effect is low, the environmental impact of foregone reductions from efficiency improvement could be significant, due to the scale of the transport activity and the marginal cost of the externalities in some countries. The rebound effect is a potentially relevant factor and it is, therefore, important to consider specific policies that ideally can be combined with price signals in the sector, the promotion of intermodality and, where feasible, the provision of alternative and environment friendly means of transport.

Acknowledgment

This chapter relies on the content of Llorca, M. and Jamasb, T., Energy efficiency and rebound effect in European road freight transport, *Transp. Res. Part A: Policy Prac.* 101, 2017, 98–110. The authors acknowledge the support of the Centre for Sustainable Road Freight Transport (EP/R035199/1) funded by the Engineering and Physical Sciences Research Council, UK.

References

- Allen J., Piecyk M. and Piotrowska, M., 2016. An analysis of road freight in London and Britain: Traffic, activity and sustainability, Report from the project Freight Traffic Control 2050 (FTC2050): Transforming the energy demands of last-mile urban freight through collaborative logistics.
- Anson, S., Turner, K., 2009. Rebound and disinvestment effects in refined oil consumption and supply resulting from an increase in energy efficiency in the Scottish commercial transport sector. *Ener. Policy* 37, 3608–3620.
- Borenstein, S., 2015. A microeconomic framework for evaluating energy efficiency rebound and some implications. *Ener. J.* 36 (1), 1–21.
- Brookes, L.G., 1990. The greenhouse effect: the fallacies in the energy efficiency solution. *Ener. Policy* 18 (2), 199–201.
- Brookes, L.G., 1992. Energy efficiency and economic fallacies: a reply. *Ener. Policy* 20 (5), 390–392.
- De Borger, B., Mulalic, I., 2012. The determinants of fuel use in the trucking industry – volume, fleet characteristics and the rebound effect. *Trans. Policy* 24, 284–295.
- European Commission, 2011. White Paper on Transport: Roadmap to a Single European Transport Area - Towards a Competitive and Resource-Efficient Transport System. European Commission, Luxembourg.
- Gillingham, K., Rapson, D., Wagner, G., 2016. The rebound effect and energy efficiency policy. *Rev. Environ. Econ. Policy* 10 (1), 68–88.
- Grubb, M.J., 1990. Energy efficiency and economic fallacies. *Ener. Policy* 18 (8), 783–785.
- Grubb, M.J., 1992. Reply to Brookes. *Ener. Policy* 20 (5), 392–393.
- Hymel, K.M., Small, K.A., Van Dender, K., 2010. Induced demand and rebound effects in road transport. *Transp. Res. Part B: Methodol.* 44 (10), 1220–1241.
- IEA, 2010. Transport energy efficiency—Implementation of IEA recommendations since 2009 and next steps, Technical report, International Energy Agency.
- IPCC, 2014. Mitigation of climate change: Contribution of working group III to the fifth assessment report of the Intergovernmental Panel on Climate Change, Chapter 8, Cambridge University Press, Cambridge and New York.
- Jevons, W.S., 1865. *The Coal Question: Can Britain Survive?* Macmillan, London, UK 1906.
- Khazoom, J.D., 1980. Economic implications of mandated efficiency in standards for household appliances. *Ener. J.* 1 (4), 21–40.
- Leard B., Linn J., McConnell V. and Raich W., 2016. Fuel costs, economic activity, and the rebound effect for heavy-duty trucks. Resources for the Future, RFF DP 15-43-REV, September 2015, revised February 2016.
- Léonardi, J., Baumgartner, M., 2004. CO2 efficiency in road freight transportation: status quo, measures and potential. *Transp. Res. Part D: Trans. Environ.* 9, 451–464.
- Llorca, M., Jamasb, T., 2017. Energy efficiency and rebound effect in European road freight transport. *Transp. Res. Part A: Policy Prac.* 101, 98–110.
- Matos, F.J.F., Silva, F.J.F., 2011. The rebound effect on road freight transport: empirical evidence from Portugal. *Ener. Policy* 39, 2833–2841.
- McKinnon, A.C., 2015. Environmental sustainability: A new priority for logistics managers. In: McKinnon, A.C., Browne, M., Whiteing, A., Piecyk, M. (Eds.), *Green Logistics: Improving the Environmental Sustainability of Logistics*. Kogan Page, London, United Kingdom, pp. 3–31.
- Ruizenenti, F., Basosi, R., 2008. The rebound effect: an evolutionary perspective. *Ecol. Econ.* 67, 526–537.
- Ruizenenti, F., Basosi, R., 2017. Modelling the rebound effect with network theory: an insight into the European freight transport sector. *Energy* 118, 272–283.
- Saunders, H.D., 1992. The Khazoom–Brookes postulate and neoclassical growth. *Ener. J.* 13 (4), 130–148.
- Saunders, H.D., 2008. Fuel conserving (and using) production functions. *Ener. Econ.* 30, 2184–2235.
- Sorrell, S., 2007. The Rebound Effect: An Assessment of the Evidence for Economy-wide Energy Savings from Improved Energy Efficiency. UK Energy Research Centre, London, UK.
- Sorrell, S., Dimitropoulos, J., 2008. The rebound effect: microeconomic definitions, limitations and extensions. *Ecol. Econ.* 65 (3), 636–649.
- Sorrell, S., Stapleton, L., 2018. Rebound effects in UK road freight transport. *Transp. Res. Part D: Trans. Environ.* 63, 156–174.
- Walnum, H.J., Aall, C., Løkke, S., 2014. Can rebound effects explain why sustainable mobility has not been achieved? *Sustainability* 6, 9510–9537.
- Wang, Z., Lu, M., 2014. An empirical study of direct rebound effect for road freight transport in China. *Appl. Ener.* 133, 274–281.
- Winebrake, J.J., Green, E.H., Comer, B., Li, C., Froman, S., Shelby, M., 2015a. Fuel price elasticities for single-unit truck operations in the United States. *Transp. Res. Part D: Trans. Environ.* 38, 178–187.
- Winebrake, J.J., Green, E.H., Comer, B., Li, C., Froman, S., Shelby, M., 2015b. Fuel price elasticities in the US combination trucking sector. *Transp. Res. Part D: Trans. Environ.* 38, 166–177.

Autonomous Goods Transport

Heike Flämig, TUHH, Am Schwarzenberg-Campus 3, Hamburg, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	407
Autonomy in Freight Transport	407
Automated Transport Systems in Closed Environments	408
Automated Transport Systems in Open Environments	409
Future Scenarios of Autonomous Freight Transport	409
Process Changes in the Supply Chain through Automated Goods Transports	409
Conclusions and Outlook	410
Acknowledgment	411
See Also	411
References	411
Further Reading	412

Introduction

Freight transport's task is overcoming space as a function of production and consumption. The transport of goods becomes necessary because the places where goods and services are produced or provided and the places where they are consumed are rarely identical. The same applies in the case of people who are transported to their place of employment for the purpose of providing a service; craftsmen for example need to carry tools or components with them. Therefore, in contrast to the mobility of individuals, freight and commercial transport is not a question of mobility per se, but develops as a consequence of the spatial and functional division of labor in systems of production and ways of life. This brings about certain transport requirements for commodities, goods, auxiliary materials, and operating supplies as well as for personnel. For this reason, technological development in the area of automated freight transport has been distinct from that of passenger transport.

The trend toward greater automation has long been evident in all modes of transport. Remote control is being used in the case of driverless railways, and the autopilot has been the norm for steering aircraft for quite a while. The first Unmanned Aerial Vehicles (UAVs—commonly referred to as drones) are already flying in the civilian sector and being used for inspections, surveillance and surveying or for film productions or the transport of small consignments such as pharmaceuticals. Drone ships are being developed where, for example, a crew is only aboard when the ship is in port. Unmanned submarines have been in service for quite some time. All modes of transport have different degrees of freedom, which are partly dealt with in other chapters of this encyclopedia.

This contribution aims to create an understanding of autonomy in the realm of transport. For that purpose, automated solutions already implemented in the area of freight transport are outlined, and the effects in the area of the supply chain and road transport system are critically reflected. The article sums up with some conclusions and makes recommendations for further research.

Autonomy in Freight Transport

Autonomy is a state of freedom of decision and action, and it is usually equated with self-determination, autonomy and independence. While these concepts can be understood by everyone in the context of the mobility of individuals, the question arises as to what this means for the mobility of goods.

There are different approaches that can be distinguished by degree of automation of spatial mobility. First, it can be differentiated with respect to the degree of automation of the driving task itself, of the vehicle and as concerns the benefits of mobility. The driving task itself consists of steering, accelerating and braking to achieve the desired vehicle movement. For this purpose, decisions are made about speed and lane selection, that is, the longitudinal and lateral guidance of the vehicle. Decisions required to carry out the driving task are made on the basis of information about the environmental conditions and associated knowledge of how to act in response to these conditions.

Autonomy of the vehicle, therefore, means handing over decisions about one's own driving behavior to the vehicle itself. Here, the autonomy of the driving task and the vehicle are closely interrelated. Donges (1982) differentiates the classical driving task assumed by the driver on three levels, according to the required information and action knowledge about the behavior of the vehicle and the environment: At the stabilization level, information about the road surface is a prerequisite for a decision on possible corrective measures. The same applies to the traffic situation at guidance level. Without driver assistance systems, the driver must contribute to keeping the vehicle in a controlled condition by changing the speed, lane selection and lane position. At the navigation level, information on the condition of the road network, on the traffic and environmental situation and their possible effects is

processed, in order to allow for a permissible, safe (or also fast or ecological) route selection. Then related action strategies will be derived.

In the field of road traffic, internationally recognized classification systems have been introduced by the German Federal Highway Research Institute (BASt) (Gasser, 2012, p. 9), the American National Highway Traffic Safety Administration (NHTSA, 2013, p. 4–5) and the Society of Automotive Engineers (SAE International, 2016, p. 19–24). In the meantime, the DIN (German Institute for Standardization) and the SAE International have been working together to promote an internationally uniform understanding of the autonomy of road-bound vehicles. Most recent definitions, such as those of SAE International, assume an automated vehicle (AV) as a vehicle equipped with driving automation technology (DAT) on levels 1 to 5. At present, there are usually six levels of automation considered, including level 0 which describes the vehicle without any assistance. However, the differentiation of the various levels has not yet been finalized. In principle, the different steps toward fully automated driving differ primarily in their functional scope, that is, how many subtasks the system performs for a certain period of time and in specific situations. The driver is considered to represent either a fall-back level or a control organ.

The autonomy of vehicles starts at the level of stabilization and track guidance. The driver's use of the vehicle is facilitated by the installation of technical components that compare target values and actual status, then trigger independent corrective measures. In order to further increase the safety and comfort of vehicle use and the driving task, driver support features and automated driving features were developed, which can react in response to specific conditions or situations. At a lower automation level, the vehicle is thus able to detect, for example, the vehicle in front or stationary objects. Typical examples of features of the first automation level of traffic flows are the lane-keeping assistant and adaptive cruise control. When several of these features are installed, usually Level 2 is ascribed. Up to this level, reference is made only to the Data Sharing Framework. At Levels 3–5, automated driving features are installed. ADF not only support the driver in the correct execution of the driving task at the stabilization and lane guidance level, but also make decisions at the navigation level. AVs starting at Level 3 are able to recognize moving objects, which can also be subjects, and formulate decisions and actions depending on the road conditions (e.g., black ice, construction sites) or environmental conditions (e.g., heavy precipitation, snow). At lower speed ranges, these conditions will usually trigger an immediate braking action. The parking assistant for example, which is already available today, is able to carry out steering movements automatically in manageable environmental conditions.

At higher speed ranges, the decision-making process is usually much more complicated and often more complex. This is mainly due to the information required on environmental conditions, the complexity of which is influenced on the one hand by the performance of the technical systems required for environment recognition and interpretation range of the situation assessment interpretation, and on the other hand by possible feedback loops between the different subsystems. In particular, the decision algorithms for action preparation in complex environments must take into account not only the target variables of safety and comfort for the vehicle occupants, but also ethical aspects, in particular with regard to the evaluation of other road users, living creatures or objects.

Until now, no worldwide ethical standards apply. The first ethical rules on “automated and networked driving” were presented for the first time by an independent commission in Germany in July 2017. In the meantime, this initiative has been transferred to the European Union level, in order to develop a common standard. To date, there seems to be an international consensus that in the case of passenger transport, individual autonomy is valued so highly that the driving task must always involve human drivers, even in vehicles that have the autonomy level of automation.

Only from Level 4 onwards can the specific case of application involve the omission of components, which allow a human being to accelerate, brake and steer the vehicle. Level 5 stands for vehicles that can move without a driver in all situations and describe the term autonomous. As far as ethical values speak against vehicles without components for human driving in individual traffic, such solutions are only being used for the transport of goods in demarcated environments. In the field of freight transport, the main benefit lies in overcoming the distance between goods production and consumption. From a logistical perspective, greater vehicle autonomy aims above all at optimized economic and ecological efficiency. In addition to safety aspects, aspects such as time, reliability, predictability, etc. are also taken into account during system design and can be evaluated differently, depending on the application.

Since conditions and applications for an autonomous vehicle on public roads are not yet fully known, these are described using scenarios. A scenario consists of a sequence of scenes, which describe a current situation where at least one dynamic element, for example, steering, changes its condition because of an activity, for example, parking, takes place (Geyer et al., 2014). Selected scenarios for freight transport are outlined in the following sections.

Automated Transport Systems in Closed Environments

Even before industrialization or the invention of the steam engine, transport systems, which relieved people of the burden of transporting typically heavy bulk goods, were developed. With the emerging functional division of labor across different factories and localities, the need for transport between processing steps increased. At the same time, an increasing management focus on technology came to envisage a complete automation of repetitive labor. While economic situations varied, this vision was largely driven by a mixture of labor shortages, efficiency concerns and reliability aspects. More recently, in-house conveyor (belt) systems, initially adapted from bulk goods transport, were integrated into the production flow. They were placed between the production and assembly stages and in incoming and outgoing zones, in the picking zone and also in the warehouse area.

The increasing flexibility of the production systems led to a growing demand for a flexible and reliable material flow. The first automated guided transport (AGT) systems were introduced in the early 1950s. AGT systems consist of a control system specifically for order placement and route planning, a communication system and vehicles. The name AGT is derived from the fact that there is no driver on board, and the transport order processing is carried out via a control system. This control system handles the central management of transport orders, vehicle-scheduling and order-processing. The traffic control system is similar to that for rail traffic in block sections, where only one vehicle is allowed to move in a particular section. Newer approaches to decentralized control rely, for example, on cooperative, agent-based negotiation processes between vehicles, or use biomimetic approaches, such as swarm intelligence, particularly for route planning and order placement.

In the past, physical guidelines were used for vehicle positioning, such as magnetic tapes, active-inductive, or optical guidance tracks. Metallic or magnetic ground marks or transponders are used for grid navigation. Newer technologies combine laser scanners and camera systems with digital environmental maps and enable free navigation by means of environmental features. This means that existing reference points can be used without committing to major investments in additional infrastructure. The AGTs will be equipped with 3D lasers, laser scanners and cameras for 3D environmental detection. Security concepts also increasingly work with sensors and laser scanners. However, they are limited to a minimum functionality, since the operating speed is at low levels, and the scenario is controlled.

Automated Transport Systems in Open Environments

Typical areas of application for driverless vehicles outside of plant premises are, for example, AGT for heavy transports or internal shuttle transports. One of the oldest examples are automated guided vehicles (AGVs), which transport containers between the container gantry cranes and the container warehouse, for example at container terminals in the port of Hamburg, Germany. The AGV's position is determined by transponders, and route planning is carried out independently, as is battery replacement. The system is controlled via radio data transmission. Driverless truck systems operate according to similar principles on factory premises between production and logistics buildings, often combined with automatic loading and unloading devices. In outdoor areas, the safety concept usually consists of radar sensors, with an increasingly frequent use of laser scanners.

Automated trucks have been used for even longer to transport raw materials in mines. Navigation is now carried out using radar and laser technologies and waypoints for orientation. Control and intervention options are implemented via a control center and in smaller environments by wireless LAN. In bigger areas GPS and dead reckoning are used through ongoing positioning (localization) which measures course, speed and time, as with a ship or aircraft. In addition, so-called driving robots have been developed for use in dangerous or difficult-to-access situations. A German logistics service provider, Hermes, has already tested so-called starship-delivery robots.

Future Scenarios of Autonomous Freight Transport

As the aforementioned remarks have revealed, driverless freight transport already exists in many areas. When operating fully automated freight transport, including free navigation across the entire road system, a driver is certainly no longer present. As a consequence, personnel would be missing for fulfilling other activities in the supply chain, such as loading or unloading vehicles (see the following section). However, this scenario is rather unlikely to happen in the foreseeable future. Therefore the following solutions for road-bound traffic that are under development, or even already in use, are more likely to happen:

- The “autopilot” is already being tested as a form of highly automated driving on motorways with the driver available as fallback. The driving robot controls the vehicle between motorway access and exit, or in clear driving situations.
- The “platoon” on motorways consists of vehicles coupled with electronic drawbars that use their own (dedicated) electrified lane for safety reasons. In the first period, a fallback driver still steers the lead vehicle.
- The “follow-me” semi-autonomous vehicle serves the user as a rolling depot for goods or tools when stops are close by. If necessary, the user jumps on the vehicle and steers it. In the approved scenario—for example, a city center delivery at low speed—navigation is remote-controlled by the user.
- “Valet parking or delivery” is the use case in which the driving robot on the “last-last mile” takes over the parking of the vehicle, or the docking of a truck at the receiver's ramp or navigation without a driver from, for example, a waiting zone to a large construction site. Depending on the environmental conditions, this would also be already possible to implement without the need for a fallback driver.

Process Changes in the Supply Chain through Automated Goods Transports

Increasing automation in road freight transport also triggers process-related changes in the supply chain, which can be trip-related, vehicle-related, load-related and supply-chain-related. In those automation stages where a person remains in the vehicle as a fallback level, it is expected that the range of activities will expand further. In particular, an increase in planning, documenting and controlling activities is anticipated. These activities may require at least some advanced training, in particular in the area of

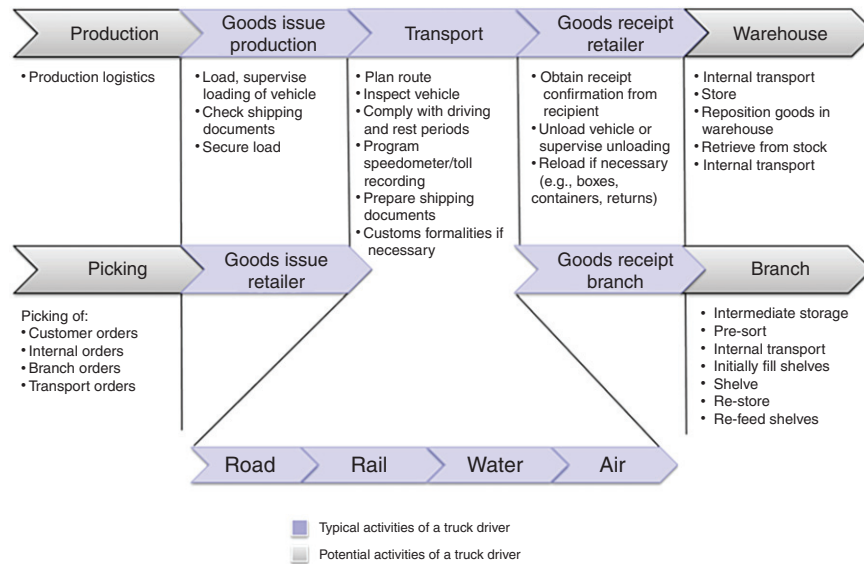


Figure 1 Typical fields of activities of truck drivers. *Source: Adapted from Flämig, 2016*

disposition and controlling. However, completely new occupational profiles can also emerge. In principle, it can be expected that tasks will shift further from physical to (co-) thinking activities, which usually require a higher level of training and expertise. At the same time, more demanding technical knowledge and physical activities, such as vehicle inspection, load securing, etc., are emerging. The profile of the professional driver is likely to expand accordingly.

In the case of highly automated, driverless, or autonomous transport, the frequent problem of delays owing to traffic and waiting time would become obsolete and itineraries could be planned with a higher degree of freedom. From the perspective of the driver's current job profile in production and retail distribution, the so-called added-value services in particular must be reorganized. However, it is conceivable that the emerging driverless supply chains could also create low-skilled jobs. In addition to driving and vehicle inspection, many drivers also perform activities in the area of handling incoming and outgoing goods or picking inventory. Tasks such as loading and unloading of the vehicles will continue to be performed—if not by the supplier or transporter, then by the recipient. Delivery robots, which have already been tested in last-mile distribution, may transfer the task of the removal of goods to the recipient. It is also conceivable, however, that a further business model could emerge for logistics service providers. This could lead to a revaluation of logistics work in general, and the creation of more localized employment, especially in urban areas.

In the area of production, one could imagine that transport contracts will be put out to tender and autonomous trucks could “bid” for them, winning certain contracts according to predefined criteria. This would neither require central control nor pathfinding. As freight locks are already existing in practice, no physical staff would be needed to be present, in order to organize the goods transfer or to minimize risk. In general, this means that the framework conditions in which companies design supply chains need to be redefined, and new degrees of freedom must be introduced within associated design practices (Fig. 1).

Conclusions and Outlook

Automated driverless vehicles have been used for the transport of goods in raw material extraction, production and in closed logistics systems for a long time. Tele-operated driving often takes place when an operator controls the vehicle from a remote location. The transport of goods outside of, or between, companies takes place in transport systems in open environments, and this usually includes a driver, conductor, or pilot. An increasing automation of vehicles can be observed in road traffic. In vehicles with lower levels of automation, assisted systems, when engaged, warn the driver if the vehicle is not running safely. In the case of semiautomated vehicles, the system takes over control for a certain period of time or in specific situations (or both), in order to ensure the longitudinal and lateral guidance of the vehicle. Fully automated vehicles, on the other hand, can navigate freely, regardless of situation and infrastructure.

The relationship between fully automatic driving robots and forms of traffic controlled by human beings will become more interwoven over time. For the use of autonomous vehicles alongside road-bound traffic, the introduction of a higher-level authority is therefore conceivable, which can intervene directly in vehicles' kinetics, as already occurs in the case of internal driverless transport. By comparison, in existing air, rail, and sea traffic systems, higher-level means of control extend only to regulating traffic flow. The digitalization of the road infrastructure can, therefore, be assumed to be in an early stage of development. The automation of vehicles is primarily a technical issue with direct implications for the prospects of automating driving or transport tasks. Initially, it served only to improve the safety and comfort of manual driving, while more recent developments aim at the complete automation of the

driving task. However, analysis of possible developments in supply chains with regard to driving and transport tasks should go beyond the transport system per se. Gradual automation of individual logistics processes which, in their totality, may lead to an autonomous supply chain and may also mean that consideration of micro- and macroeconomic cost-benefit factors become relevant. For example, the lower safety buffers and more homogeneous driving speeds which can be achieved through automation may help to reduce energy consumption and to increase the capacity of existing road infrastructure.

Meanwhile road traffic has become more diverse. In addition to cars and trucks of different sizes and degrees of automation, (cargo) bicycles, trailers, or scooters such as Segways with e-drive are widespread. At the same time, parts of goods transport have shifted to drones or underground transport systems. Intermodal vehicles, such as the RiverBus in Hamburg, must also be considered in any future automation of road-bound transport. More and more structures and purely mechanical components are being replaced or supplemented by electronic components. Sensor technology, lasers and laser scanners, cameras, radar, and transponder technology are becoming an integral part of material flows of goods and commodities, transport and handling technology, supra-structures, and infrastructures. Infra- and supra-structures increasingly provide real-time information about their (expected) conditions, for example about potential accident hazards and traffic jams. With digitalized, smart infrastructure emerging, individualized end-devices (such as smartphones) may be able to communicate with other smart vehicles or other end-devices.

The integration of information flows across the entire value chain not only changes the predictability and, thus, the accuracy of material requirements planning. Rather, networked, decentralized, real-time-capable and self-optimizing production and logistics systems will become possible, with which small units can be produced at a higher degree of individualization than before. These smaller units may have just as much influence on the autonomy of goods transport as the generative manufacturing processes made possible by the computerization of production.

In further discussion about the automated flow of goods and automated vehicles, embeddedness into the cyber-physical system of logistic systems for goods and commodities must be kept in mind. The simultaneous emergence of autonomous systems in other transport modes, for example drones, indicates that completely new business models in the field of logistics can be conceived through the emergence of autonomous systems. In these systems, the flows of goods tend to be more granular, more individualized. However, a proper conceptualization and evaluation of these flows is as yet pending. Therefore the current way of introducing these new systems incrementally seems appropriate and realistic. In the course of this associated networking process, new partners, partnerships, new cooperations, or new commercial enterprises need to be integrated into the system. Testing of these developments in real-world settings should therefore be accompanied by appropriate implementation research, in order to minimize negative feedback effects in particular. Digitalization in the form of networking, robotics, and automation is likely to change the physical networking of transport systems and the transition from the logistics system to the transport system. It also changes networking in relation to the economic and social framework conditions. These different dimensions must, therefore, be considered together.

Acknowledgment

This article is based on research that was undertaken in the following funded projects:

Daimler und Berta Benz Stiftung: Autonomes Fahren im Straßenverkehr der Zukunft: Villa Ladenburg (2012–14)

Bundesministerium für Verkehr und digitale Infrastruktur: ATLaS—Automatisiertes und vernetztes Fahren in der Logistik (2017–19)

See Also

Demand for freight transport; Cost functions for road transport; The Future of Urban Freight; ATV, snowmobile and terrain vehicle safety; Automatic vehicles and connected vehicles; Introduction: Trade, Freight Transport and Logistics; Logistics & supply chain management; Logistics information systems; Drones in freight transport; Road freight vehicles; ITS in road freight transport; 3D printers and logistics; The physical internet and logistics; Autonomous vehicles and transportation modeling; Data sources and transport modes; Vehicles that drive themselves: what to expect with autonomous vehicles; ICT and transport modes; The future of air transport; The future of road transport; The future of rail transport; The future of maritime transport; Transport modes and big data; Mobility as a service (MAAS)

References

- Donges, E., 1982. Aspekte der Aktiven Sicherheit bei der Führung von Personenkraftwagen. *Automobil-Industrie* 2, 183–190.
- Gasser, T.M., 2012. Rechtsfolgen zunehmender Fahrzeugautomatisierung. In: Bundesanstalt für Straßenwesen (Ed.), *Berichte der Bundesanstalt für Straßenwesen Fahrzeugtechnik Heft F 83*. Bremerhaven, Wirtschftsverlag NW.
- Geyer, S., Baltzer, M., Franz, B., et al., 2014. Concept and development of a unified ontology for generating test and use-case catalogues for assisted and automated vehicle guidance. *IEE Intelligent Transport Systems* 8 (3), 183–189.
- Flämig, H., 2016. Autonomous vehicles and autonomous driving in freight transport. In: Maurer, M., Gerdes, J.C., Lenz, B., Winner, H. (Eds.), *Autonomous Driving Technical, Legal and Social Aspects*. Springer-Verlag, Berlin, Heidelberg, pp. 365–385.
- NHTSA, 2013. Preliminary Statement of Policy Concerning Automated Vehicles, National Highway Traffic Safety Administration, Department of Transportation, Washington.

SAE International, 2016. Surface and vehicle recommended practice: taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Society of Automotive Engineers. www.sae.org/.

Further Reading

Maurer, M., Gerdes, J.C., Lenz, B., Winner, H. (Eds.), 2016. Autonomous Driving Technical, Legal and Social Aspects. Springer-Verlag, Berlin, Heidelberg.

National Highway Traffic Safety Administration, 2016. Federal Automated Vehicles Policy: Accelerating the Next Revolution In Roadway Safety. U.S. Department of Transportation, Washington.

Society of Automotive Engineers (SAE International), 2018. SAE international releases updated visual chart for its "Levels of Driving Automation" standard for self-driving vehicles. 2018-12-11 WARRENDALE, Pa. www.sae.org/.

Vahrenkamp, R., 2013. Von Taylor zu Toyota: Rationalisierungsdebatten im 20. Jahrhundert. 2. korrigierte und erw. Auflage. Josef Eul Verlag, Lohmar-Köln.

Rail Freight

Allan Woodburn, University of Westminster, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

An Introduction to Rail Freight	413
What is Rail Freight?	413
Rail's Share of the Freight Transport Market	413
Typical Rail Freight Commodities	415
Fundamentals of Rail Freight Operations	415
Relationship between Infrastructure and Service Provision	417
Key Opportunities and Challenges for Rail Freight in the 21st Century	417
The Changing Nature of Supply Chains	417
Information and Communication Technologies (ICT) for Rail Freight	418
Rail's Contribution to Making Freight Transport More Sustainable	418
Operational Efficiency Improvements	419
Competitive Markets Within Rail Freight	420
Innovation	420
Summary	420
Relevant Websites	421
Further Reading	421

An Introduction to Rail Freight

This article presents a critical review of rail freight, split into two sections: an introduction to rail freight and a discussion of key opportunities and challenges facing rail freight in the 21st century. Within this introduction, rail freight activity is first defined and its role in the freight market is set out. Key commodities carried by rail are then elaborated and the fundamentals of rail freight operations are examined. The section concludes with a discussion of the relationship between the rail infrastructure and the provision of freight trains over that infrastructure.

What is Rail Freight?

At its most basic, rail freight comprises the movement of goods (as opposed to people) on railway/railroad systems. This is typically distinct from passenger train operations and involves dedicated freight (or goods) trains comprising one or more locomotives hauling a set of freight wagons. There are alternatives for certain flows; however, such as freight multiple units (e.g., those used by the Royal Mail in the United Kingdom (UK) or, until 2015, the dedicated high-speed TGV trains carrying mail in France for La Poste), passenger trains with space devoted to carrying small quantities of freight, or even combined trains conveying both passenger carriages and freight wagons. Where they exist, however, these alternatives tend to be niche.

While data are incomplete and not fully comparable, Fig. 1 shows that global rail freight activity is highly concentrated in just a few countries. While data sources do not always concur, UIC data show that the Russian Federation, United States, and China each account for 25%–30% of global activity and in combination have more than 80% of the total market. While Germany, the largest rail freight market within the European Union, has just 1% of global rail freight activity, some European countries, notably Germany and Switzerland, rank highly in terms of rail freight activity per capita, reflecting the importance of rail freight to their economies.

Rail's Share of the Freight Transport Market

It is generally the case that rail's mode share of freight transport has declined since the middle of the 20th century, in both developed and developing countries, mainly because of the growing competition from road haulage. In the United Kingdom, for example, rail's share of freight moved declined from 54% in 1953 to just over 8% in 2016, while in India, rail's share more than halved from over 80% in 1950–51 to around 35% in 2014–15.

In the three biggest rail freight markets identified in Fig. 1, rail's share varies considerably: Russian Federation (17%), United States (32%), and China (8%). Eurostat data reveal considerable variations between European countries in rail's share of the inland freight market (Fig. 2). In conjunction with the type of rail freight activity (Fig. 3), it can be seen that those countries with a higher rail share tend to be more integrated into international rail systems, with a dominance of international and/or transit traffic and a smaller share for national flows. Country size, geographical location, and industrial structure are important factors influencing rail's share of the market.

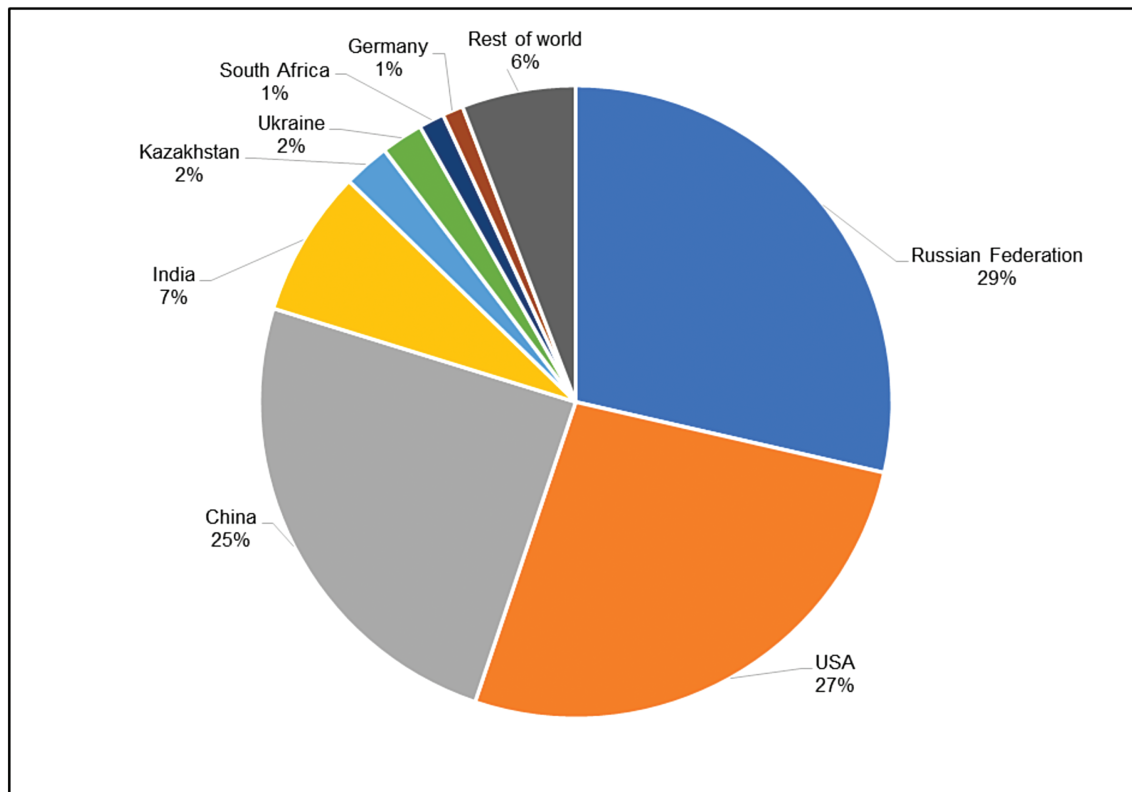


Figure 1 Worldwide rail freight activity (% of ton kilometers, 2017).

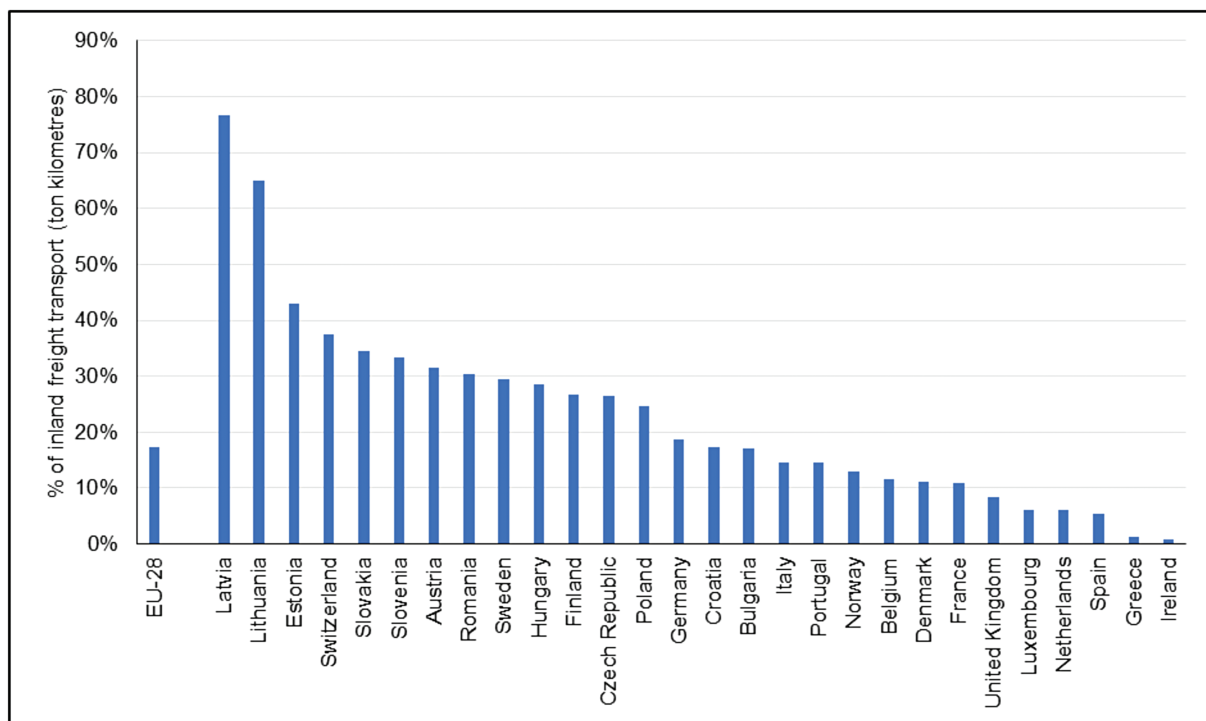


Figure 2 Rail share of inland freight transport (EU plus others, 2016).

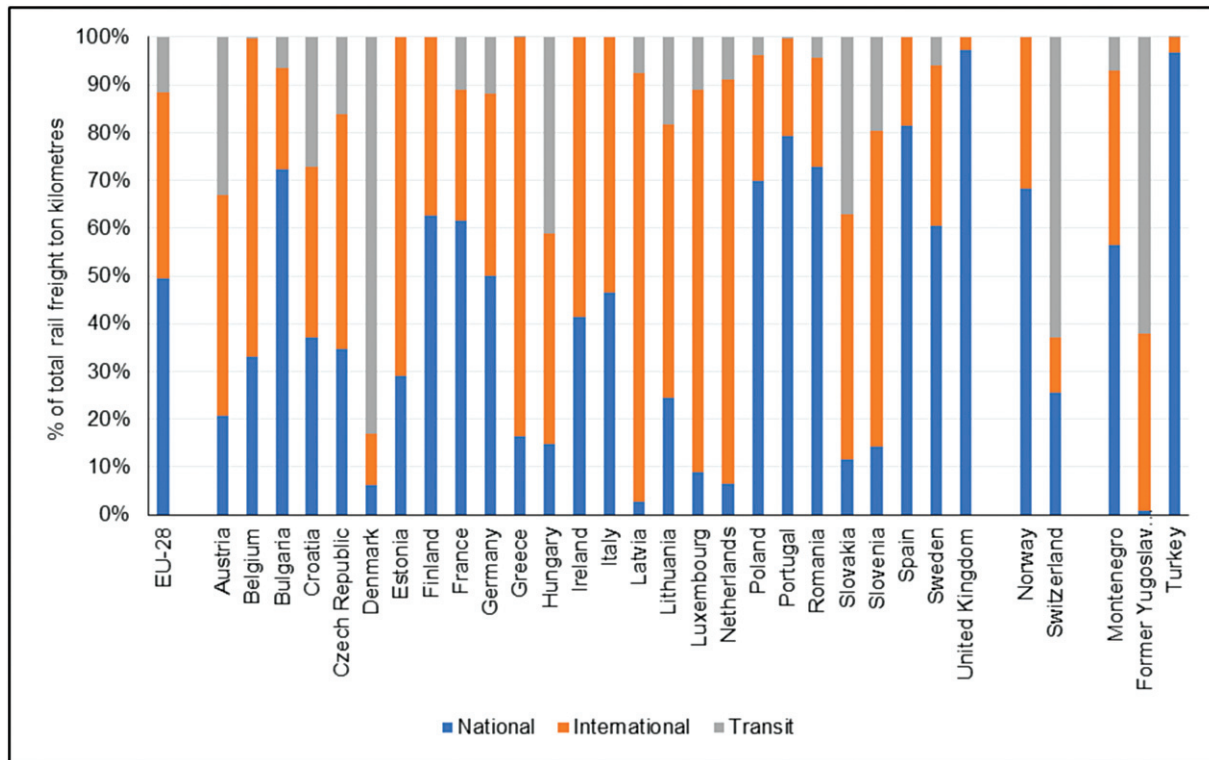


Figure 3 Rail freight activity by type of transport for main undertakings (EU plus others, 2017).

Even in countries where rail's overall share of the freight market is relatively low, rail often plays an important role in certain markets. For example, around 45% of the land-based freight volume to/from the Port of Hamburg is moved by rail, and 40% of construction materials used in London arrives there by rail.

Typical Rail Freight Commodities

A high proportion of rail freight's costs tend to be "fixed," certainly in comparison with those for road haulage. To allow freight trains to operate, considerable investments in track, terminals, locomotives, wagons, and skilled staff must be made. As a consequence, rail is best-suited to high volume flows over long distances, since this helps to reduce unit costs and provides rail with a competitive advantage over road haulage. The commodity composition of a country's rail freight market is dependent upon numerous factors including economic structure, availability of raw materials, industry activity, and geography. Activity is usually dominated by bulk flows of raw materials (e.g., coal, construction materials, crude oil, grain, iron ore, timber) or manufactured goods (e.g., cement, chemicals, petroleum, steel). In the Russian Federation, for example, coal and oil and petroleum products comprise almost half of the tonnage carried by rail, while coal alone is around one-third of rail freight tonnage in the United States. In certain countries, including Australia and South Africa, freight trains carrying more than 20,000 tons of raw materials are commonplace.

Over the last few decades, intermodal techniques (e.g., for unitized loads such as ISO containers) have become more common, though their uptake on rail has been variable. For example, **Fig. 4** shows how the importance of intermodal rail freight has changed across the European Union between 2007 and 2016. In most countries where data were available for both years, intermodal rail freight as a proportion of the total rail freight market has increased. While it remains a minority of total rail freight activity overall, in some European countries, it was approaching a 50% share in 2016 and had already exceeded this in Ireland. In general terms, intermodal activity is a major growth area for rail freight, with rail focusing on its strengths and road haulage satisfying the requirements of the door-to-door journey of dispersed customers' consignments.

Fundamentals of Rail Freight Operations

In most cases, to be able to operate freight trains, the following components of the system are required:

- Locomotives—to provide the power to move the goods
- Rolling stock—the wagons in which the goods will be carried
- Staff—to crew the trains, load, and unload goods and provide administrative services
- Line-of-route infrastructure—to allow trains to move from A to B

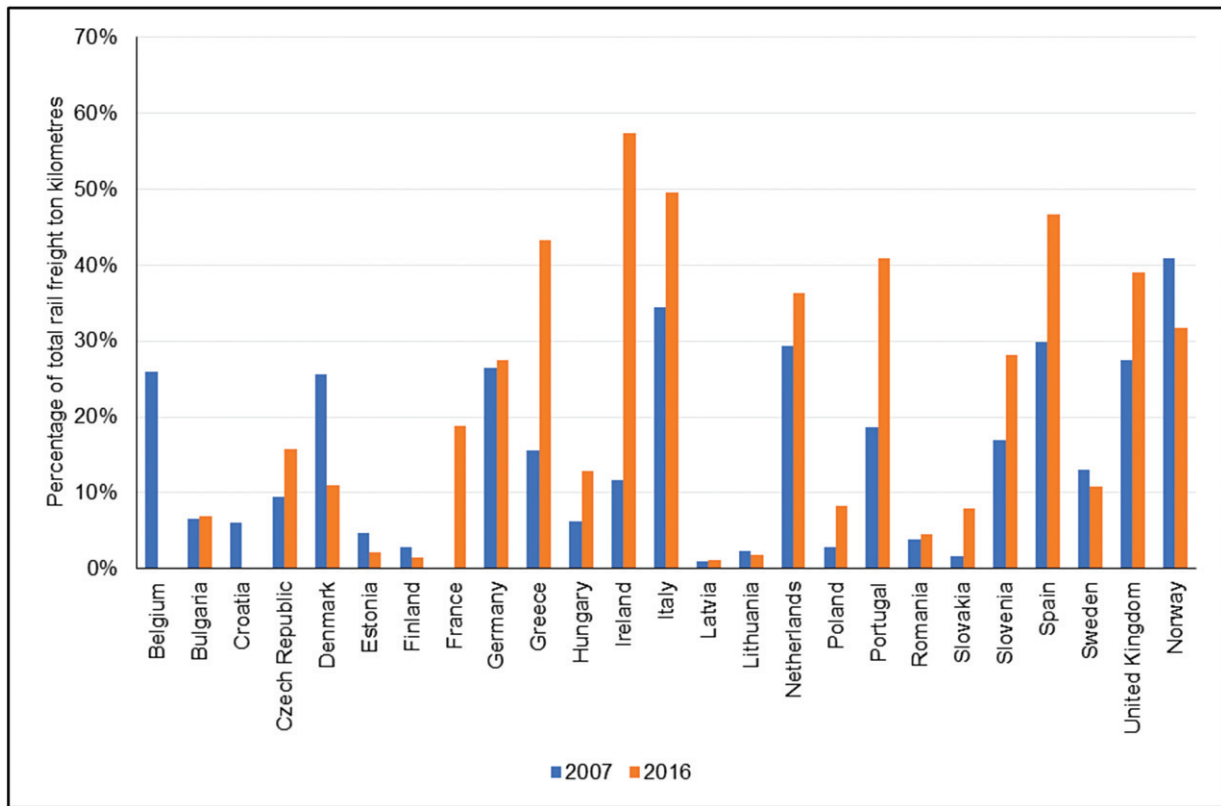


Figure 4 Share of intermodal transport units in rail goods transport (Eurostat data—2007 and 2016).

- Terminals—to provide the interface between the rail network and the “wider world” (either suppliers’ and customers’ premises or other transport modes)
- Handling equipment—the means by which the goods are loaded on and off the rolling stock

In a basic sense, freight train operations can be categorized as either trainload or less-than-trainload/wagonload in nature. Dedicated trainload provision is the most straightforward type of operation, with a block train moving from origin to destination without any intermediate marshalling activity, generally carrying a single commodity type for a single customer. Intermodal rail freight can also operate on a trainload basis, direct from one terminal to another, though conveying unit loads of different commodities for a range of customers. At the other extreme, a true wagonload operation will permit the movement of individual wagons through the network on shared-user services, but often this is not an economically viable option. Where viable, though, it does provide a rail freight service for customers with smaller quantities of goods to be moved. **Fig. 5** shows an example of a



Figure 5 Examples of wagonload (left) and trainload (right) rail freight operation.

wagonload service on the left, with a range of different wagon types clearly visible behind the locomotive, and a bulk trainload service carrying only coal in large hopper wagons on the right.

The ideal freight flow attributes for rail are high volume, long distance, and predictability/regularity, since a combination of these plays to the strengths of rail transportation. As with wagonload services, there may be potential to combine several small volume flows into a long-distance trunk movement. While wagonload operation generally has been in decline in recent years, there have been attempts to reinvigorate it, such as with the Lineas (Belgian Railways) Green Xpress initiative which combines wagonload and intermodal flows into viable trunk trains between European hubs. Wagonload remains an important form of rail freight operation in many countries, such as Germany and the Czech Republic, but has been all but eliminated elsewhere (e.g., in the United Kingdom).

Relationship between Infrastructure and Service Provision

Rail network density is considerably lower than that for the road network, so door-to-door rail freight is not an option for many flows. Terminals are a critical part of the overall rail freight system, given that freight flows need access to, and egress from, the rail network, connecting with other modes of transport such as road or shipping. They can also be used for transferring freight from one rail service to another, but this is more usually carried out in marshalling yards. On the rail network itself, route characteristics such as axle weight limits, loading gauge, signal spacing and gradients will influence the maximum length and weight of freight trains which are able to operate.

Some countries' rail networks are geared toward freight activity as the prime (or sole) user, while for others, freight must compete with passenger traffic to gain access to the network. The operation of freight trains on a mixed-use rail network, such as in the United Kingdom and most other European countries, poses particular problems and often results in less than ideal journey times. Rail network capacity is fixed and freight trains require paths, or slots, in the timetable. Freight trains often receive less priority than do passenger trains, and are perceived (sometimes rightly, sometimes not) as being slower. This need for scheduling can reduce rail's flexibility, particularly where there is little spare network capacity, but can also lead to robust and reliable service provision, since capacity must be available for a train to operate. This contrasts with road haulage, where network access is typically less controlled, offering greater service flexibility but also an increased risk of being delayed by congestion. Owing to capacity constraints, freight trains are sometimes limited to off-peak operation only, held in passing loops to allow passenger trains to pass, or diverted along more circuitous routes where spare capacity exists. Such treatment of freight traffic may not encourage new customers, whose requirements may not be compatible with the constraints imposed by such network access, or lack of it. Even on freight-only routes, line capacity may be limited because of low operating speeds and long single-track sections where trains cannot pass. In addition to line capacity, terminal capacity may be a constraining factor.

The respective involvement of governments and the private sector has varied over time and in different parts of the world. There is often a distinction between the provision (and maintenance) of the infrastructure and the operation of services over that infrastructure. The main models are vertical integration, where both infrastructure and service provision are the responsibility of one organization, and vertical separation, where infrastructure is managed by one entity and one or more operators provide services over the infrastructure. The USA has a vertically integrated system whereby private railroad companies own the infrastructure and provide services over it. In contrast, China has a vertically integrated system, which is state-owned and operated. European rail operations have moved toward vertical separation, with generally state-owned infrastructure with competitive service provision, often a mix of public and private sector. Competitive issues are considered further in Section Competitive Markets Within Rail Freight.

Key Opportunities and Challenges for Rail Freight in the 21st Century

As with any other freight transport mode, rail has both positive and negative attributes, which present opportunities and challenges when it comes to satisfying contemporary transport requirements and in responding to the wider economic, environmental and societal issues associated with freight flows. This section discusses some of the key issues affecting rail freight.

The Changing Nature of Supply Chains

Wider changes in supply chain activities influence both the total demand for freight transport and the modal split of this demand between the various transport modes, providing both opportunities and challenges for rail. Globalization has led to greater specialization of production, more supply chain complexity and an increasing average distance between the points of production and consumption. In some cases, where flows have become more frequent, time-sensitive and dispersed, rail is disadvantaged, since it is generally less flexible than road haulage and is perceived as being less customer-responsive. Where manufacturing activity has been transferred from the long-established industrialized nations to cheaper and less-developed countries, such as from western Europe to southern and eastern Asia, freight transport requirements in the industrialized nations have witnessed a decline in bulk flows of raw materials and semimanufactured goods, typically well-suited to rail, and a growth in dispersed flows of consumer products, for which road haulage is generally more capable.

However, there have also been changes which have worked in rail's favor. For example, the trends toward globalization and containerization have led to increased volumes needing to be moved between major ports and suppliers/customers located within their hinterland areas, feeding into and out of deep-sea maritime services (such as at the key European container ports of Antwerp,

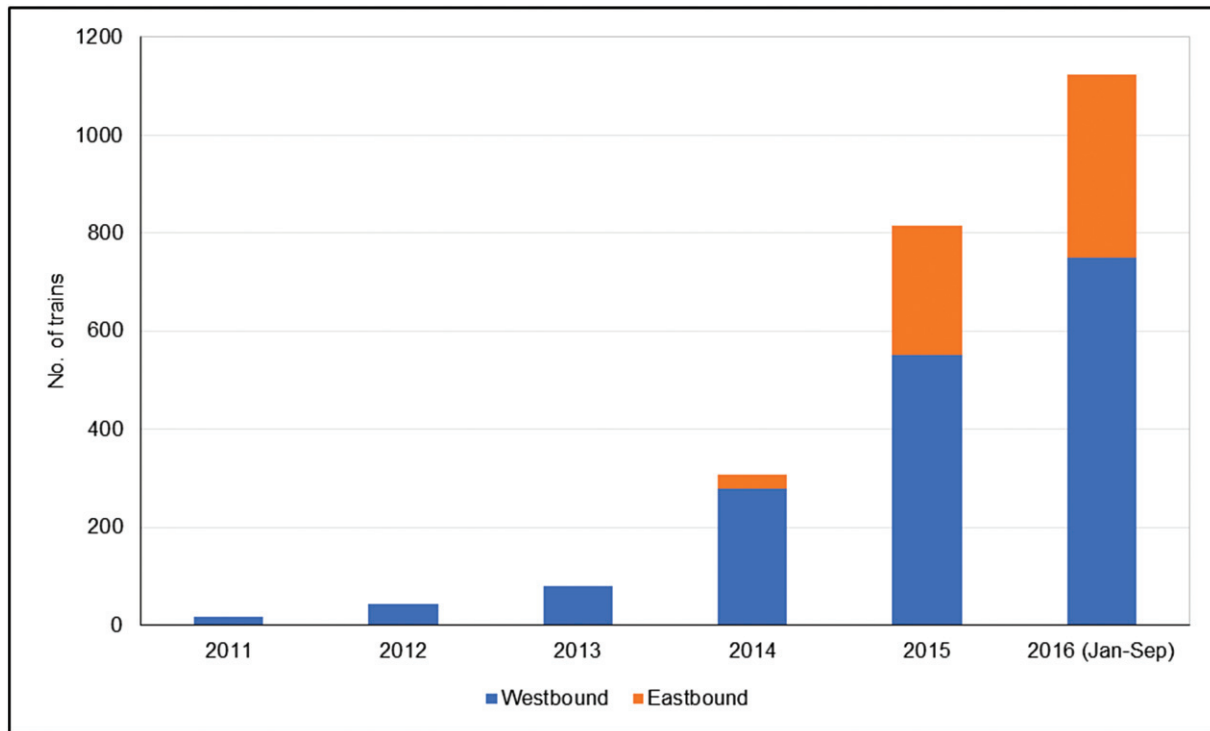


Figure 6 Number of China Railway Express (CRE) trains operated (2011–Sep 2016).

Hamburg and Rotterdam). In some circumstances, notably between China and the European Union, rail has begun to offer a direct alternative to sea and air; **Fig. 6** shows the rapid growth of such provision between 2011 and 2016.

Information and Communication Technologies (ICT) for Rail Freight

Rail has a long history of developing internal ICT systems to monitor and control the movement of freight trains since access to, and movement over, the network tends to be highly disciplined and signaling systems keep track of train movements. Key applications include establishing the position of trains on the network, identifying which locomotives and wagons are included in a train, recording the commodities carried on each train or wagon, and determining maintenance schedules and requirements for the rolling stock. While these are important uses, their benefits are often not shared with customers. This leads customers to complain frequently that their consignments disappear into a “black hole” when on the rail network. An increasing focus on customer service by at least some rail freight operators is seeing the diffusion of ICT to include the interface between internal systems and customers. Examples of operators introducing track and trace systems for their customers include CSX (USA), DB Cargo (Europe), KiwiRail (New Zealand), and Pacific National Rail (Australia).

ICT developments are assisting rail freight in other ways. Modern locomotives can be controlled more intelligently to provide more power at rail or to control other locomotives distributed through the train. In such ways, rail can play even better to its heavy haulage strengths. In a different application, remote-control shunting locomotives can improve the operations within a marshalling yard. More broadly, in-cab signaling offers potential to make better use of existing rail infrastructure, allowing trains to safely operate with closer spacings than is possible with traditional lineside block signaling. The European Train Control System (ETCS) is being developed as a means to better integrate the different rail networks across Europe with standardized technology and to contribute to operational efficiency improvements (see Section Perational Efficiency Improvements).

Rail's Contribution to Making Freight Transport More Sustainable

There is considerable evidence, which demonstrates that rail freight's impacts on the environment and society are lower than those of road haulage, its main competitor mode. On a normalized basis (i.e., per ton kilometer), rail freight performs relatively well when compared to the other freight transport modes, particularly road and air. As an example, **Fig. 7** presents the standardized UK government emissions of greenhouse gases for various freight transport modal options. Of course, these figures are averages, so it cannot be said that rail freight always outperforms road haulage; at the extreme, a diesel locomotive with a single wagon carrying 20 tons of goods would likely have a considerably worse impact per ton kilometer than would the use of a lorry for the same trip.

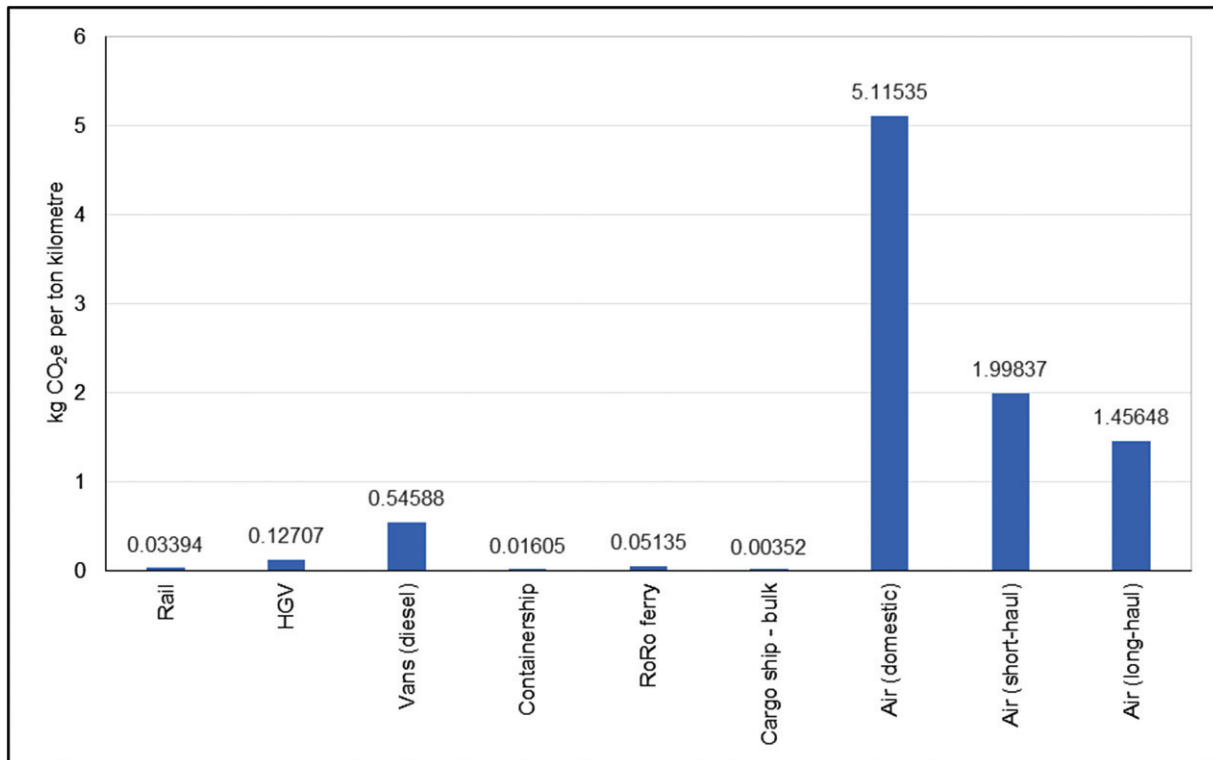


Figure 7 Freight greenhouse gas (CO₂e) emissions per ton kilometre, by mode (UK, 2017).

However, under normal circumstances, rail is considerably better than road per unit of transport activity when it comes to its contribution to climate change impacts and, as such, its use is often promoted in transport policies.

In addition to these existing environmental benefits, rail offers a faster route to freight transport decarbonization than is currently achievable for the other freight modes. In many countries, a high proportion of rail freight is electrically hauled already, increasingly using electricity generated from sustainable, carbon-free sources. In many cases where diesel traction is used at present, a switch to electric power can be achieved with relative ease.

Rail freight also outperforms road haulage when it comes to local air pollutants, with far lower emissions per ton kilometer of pollutants including particulates (PM10s), oxides of nitrogen (NO_x) and volatile organic compounds (VOCs). As the effects of air pollution have risen up the public health agenda, greater attention has been focused on the benefits of using rail freight in densely populated urban areas.

In an attempt to maximize the environmental and resource efficiency benefits, government freight transport policies tend to favor modal shift from road to rail. For example, the 2011 European White Paper on Transport set out a target for 30% of road freight travelling over 300 km to transfer to rail (and water) by 2030 and more than 50% by 2050.

Operational Efficiency Improvements

To make the most of rail's capabilities and to spread fixed costs more broadly, operators generally attempt to maximize the load carried on each train. **Fig. 8** shows that the freight load factor, essentially the tonnage carried per train, increased by an annual average of 1.2% across Europe from 2013 to 2017. Over a longer time period, from 2003–04 to 2017–18, the average freight train load increased by 63% in the United Kingdom. The increase by 2012–13 was actually 89%, but then dropped back as a consequence of the dramatic decline in heavy coal traffic. This demonstrates the importance of the commodity mix in influencing average freight train loads, since bulk trains of coal or aggregates tend to be heavier than intermodal trains.

Improving rail freight efficiency on a mixed traffic railway is a particular challenge, since both passenger and freight trains have to co-exist on the same infrastructure. Where the provision of passenger train services is intensified, freight activity can be marginalized. For example, slower freight trains may have to stop more frequently to allow faster passenger trains to overtake, lengthening journey times and reducing asset utilization for freight operators. To this end, the European Union is pursuing an initiative to develop priority corridors for rail freight, with greater international cooperation, enhanced train pathing and an overall aim to improve rail freight performance. In conjunction with the implementation of ETCS (see Section Information and Communication Technologies (ICT) for Rail Freight), it is hoped that this will make rail a more cost-effective and efficient option for customers and, as a result, lead to an increase in its share of the freight market.

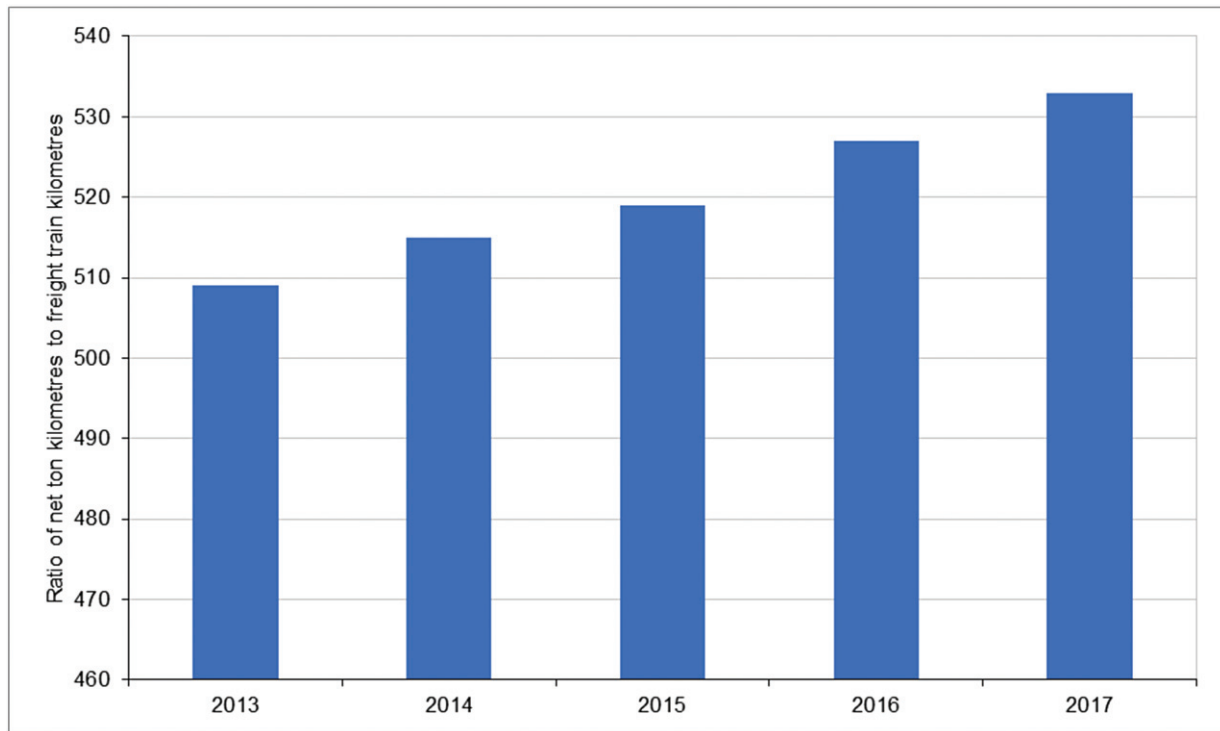


Figure 8 Freight load factor (IRG-Rail data for 26 European countries, 2013–17).

Competitive Markets Within Rail Freight

While rail predominantly competes with other modes of transport for a share of the freight market (see Section Rail's Share of the Freight Transport Market), a switch from monopoly state provision to on-rail competition has become more common in certain parts of the world, notably the European Union. Over the last two decades, the rail freight market in the European Union has been liberalized and on-rail competition has been encouraged. While the process of liberalization started at different times in different countries, and has developed at differing speeds, [Fig. 9](#) shows that the domestic incumbent share had decreased to around 40% by 2015 across the sample of IRG-Rail countries included. In just a few small rail freight markets was there no competition, while most countries had a range of operators competing with each other, comprising new operators (i.e., nonincumbents) and, in many cases foreign incumbents, notably Germany's DB Cargo which has expanded into many other European countries.

Analyses of rail freight liberalization in Europe, for example, the IBM Rail Liberalisation Index and a CERRE study of rail freight regulation, have provided evidence that those countries, which are most advanced in the liberalization process have greater on-rail competition and, typically, have seen rail fare better in competition with other freight transport modes.

Innovation

In addition to those already discussed, several other initiatives have been developed to try to increase rail's flexibility and reduce costs and time to make it a more competitive option for end-to-end journeys. For example, transferring goods to and from rail at terminals can be time-consuming and costly, so new loading and unloading techniques have been introduced: in intermodal rail freight, new wagon types offer the ability to quickly transfer road semi-trailers between modes without the need for expensive lifting equipment.

There is a movement in support of the development of unconventional rail freight, as a means of making rail more competitive for smaller volume flows in the final stages of the supply chain. For example, research has been conducted into the potential for freight multiple units (FMUs) to offer a rail solution for smaller volume flows, but the uptake to date has been extremely limited.

More broadly, the European Union has invested in innovation through its Shift2Rail initiative, aiming to introduce new technologies and working practices to improve rail's competitiveness. Examples include new composite materials for lighter weight and more energy-efficient rolling stock and seeking examples of best practice in life-cycle costs for rail infrastructure and rolling stock.

Summary

Rail plays an important role in meeting the freight transport requirements of many different markets, both geographically and commodity-wise. While it does not offer the flexibility of road haulage, rail is well-placed to benefit from the increasing distances

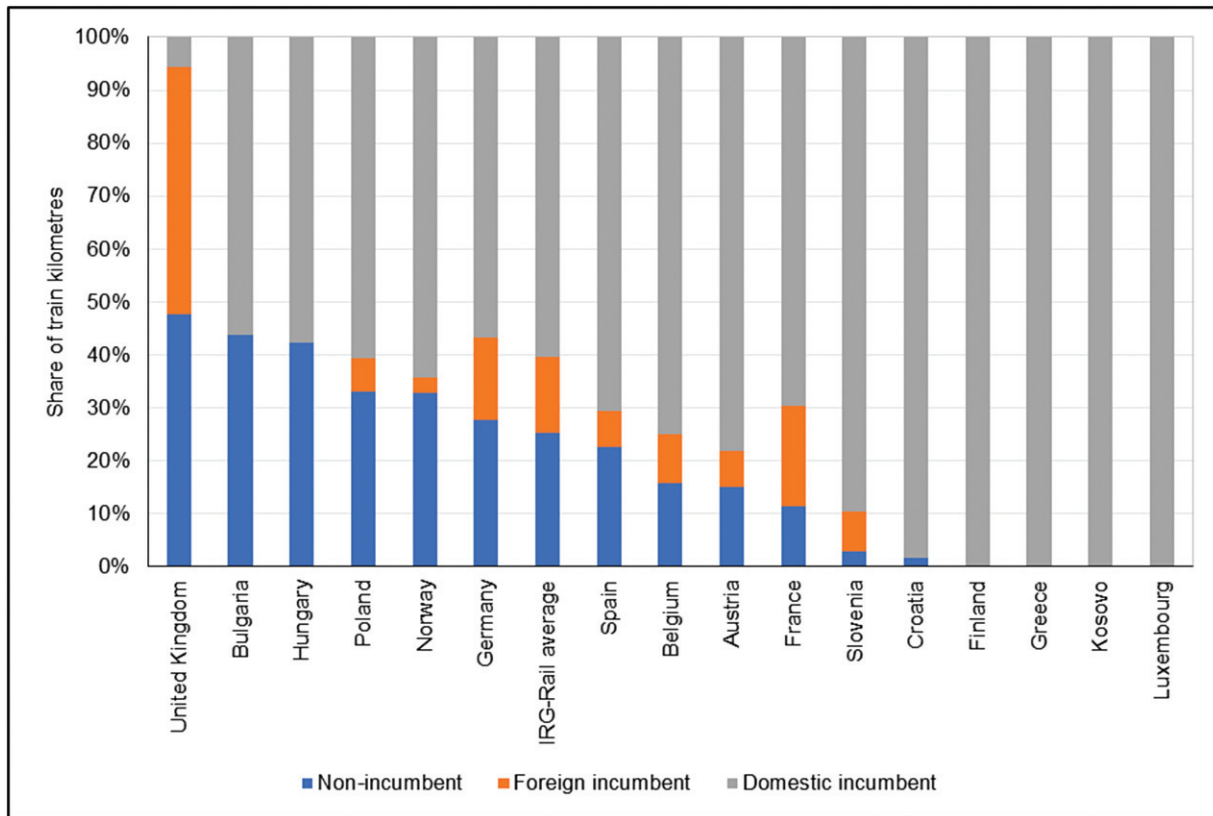


Figure 9 Market shares of domestic incumbents, foreign incumbents and nonincumbent rail freight undertakings (% of train km in IRG-Rail countries, 2015).

over which goods travel in a globalized economy and where concern for the environment is increasing. It is important that policy makers support rail, though, as it can be challenging to increase its use given the prevailing economic and regulatory circumstances.

Relevant Websites

Association of American Railroads (AAR) (<https://www.aar.org/>)
 Community of European Railway and Infrastructure Companies (CER) (<http://www.cer.be/>)
 European Commission (https://ec.europa.eu/transport/modes/rail_en)
 European Rail Freight Association (ERFA) (<http://erfarail.eu/>)
 European Union Agency for Railways (ERA) (<https://www.era.europa.eu/>)
 Federal Railroad Association (FRA), U.S. Department of Transportation (<https://railroads.dot.gov/>)
 International Railway Journal (IRJ) (<https://www.railjournal.com/>)
 Independent Regulators' Group – Rail (IRG–Rail) (<https://www.irg-rail.eu/irg/>)
 International Union of Railways (UIC) (<https://uic.org/>)
 Office of Rail and Road (ORR) (<https://orr.gov.uk/>)
 Rail Freight Group (RFG) (<http://www.rfg.org.uk/>)

Further Reading

Aritua, B., 2019. The Rail Freight Challenge for Emerging Economies: How to Regain Modal Share. World Bank Group, Washington.
 Boyer, K.D., 2014. Why is the rail share of US freight traffic so low? *Journal of Transport Economics and Policy* 48 (2), 333–344.
 CERRE, 2014. Development of rail freight in Europe: what regulation can and cannot do, CERRE Policy Paper, Brussels: Centre on Rail Regulation in Europe (CERRE). Available from: <https://www.cerre.eu/>.
 Eurostat, 2018. Railway freight transport statistics, Luxembourg: Publications Office of the European Union. Available from: <https://ec.europa.eu/eurostat/>.
 IBM, 2011. Rail Liberalisation Index 2011, IBM Global Business Services (IBM), Brussels.
 IRG–Rail, 2019. Seventh Annual Market Monitoring Report. Independent Regulators' Group – Rail (IRG–Rail), Paris.
 Laroche, F., Sys, C., Vanelslander, T., Van de Voorde, E., 2017. Imperfect competition in a network industry: The case of the European rail freight market, *Transport Policy*, 58 (August 2017), pp. 53–61, <https://doi-org.ezproxy.westminster.ac.uk/10.1016/j.tranpol.2017.04.014>.

- Liang, X.-H., Tan, K.-H., Whiteing, A., Nash, C., Johnson, D., 2016. Parcels and mail by high speed rail – A comparative analysis of Germany, France and China. *J. Rail Transp. Plan. Manag.* 6 (2), 77–88, doi:10.1016/j.jrtpm.2016.04.003.
- Morvant, C., 2015. Challenges raised by freight for the operations planning of a shared-use rail network: A French perspective. *Transport. Res. Part A: Policy Practice* 73, 70–79, doi:10.1016/j.tra.2014.12.003.
- Woodburn, A., Whiteing, A., 2015. Transferring freight to 'greener' transport modes. In: McKinnon, A., Browne, M., Piecyk, M., Whiteing, A. (Eds.), *Green Logistics: Improving the Environmental Sustainability of Logistics*. 3rd edition. Kogan Page, London.
- Zunder, T.H., Islam, D.M.Z., Mortimer, P.N., Aditjandra, P.T., 2013. How far has open access enabled the growth of cross border pan European rail freight? A case study. *Res. Transport. Business Manag.* 6, 71–80, doi:10.1016/j.rtbm.2012.12.005.

Rail Freight Vehicles

Maksym Spiryagin*, Qing Wu*, Peter Wolfs*, Colin Cole*, Valentyn Spiryagin[†], Tim McSweeney*, *Central Queensland University, Centre for Railway Engineering, Rockhampton, QLD, Australia; [†]Railway Consultant, Rostov-on-Don, Russian Federation

© 2021 Elsevier Ltd. All rights reserved.

Introduction	423
Locomotive Classification and Design	423
Locomotive Power Traction Systems	427
Tractive Effort and Dynamic Braking Characteristics	429
Freight Wagon Classification and Design	430
Freight Train Configuration and Distributed Power	432
Driverless Freight Train Operation	433
Conclusions	434
References	434
Further Reading	435

Introduction

On railway lines around the world, there are great numbers of various types of freight trains in operation. A freight train consists of multiple vehicles connected in a series that generally operates on a railway track and transports cargo or goods (Cole et al., 2013; Wu et al., 2017). The word “train” has been transformed from the Old French word “trahiner,” which can be translated as “to pull” or “to draw.” Freight trains usually consist of a mix of different types of rail vehicles. Most rail vehicles in freight train consists are unpowered and they are generally referred to as “wagons” in most countries in the world and as “cars” in the United States. In this article, all unpowered freight rail vehicles will be referred to as wagons. The term “wagon” can be defined as an unpowered rail vehicle that is used to transport various types of cargo and its design is significantly dependent on the type of cargo that it is designated to transport, the train consist that it will be hauled in, and the standard of rail infrastructure that it will operate over. Wagons in a train consist are generally hauled by one or more powered rail vehicles, called locomotives, that use motive power to haul cargo and goods to the destination point. The word “locomotive” originates from the two Latin words “loco” and “motivus,” which mean “from a place” and “causing motion,” respectively. The term “locomotive” has been strongly used to define this type of rail vehicle since the first locomotive building process that can be dated to the construction in Cornwall in 1801 by Richard Trevithick of his high-pressure steam locomotive engine “Puffing Devil” designed to run on roads. However, transport of railway cargo began with the application of the first practical freight locomotive in 1814 by another British inventor, George Stephenson, with his locomotive named Blücher hauling eight wagons loaded with coal, a total load of 30 tons, on a railway track up a gradient of 1 in 450 at a speed of 6.4 km/h. From that moment, rail cargo operations have strongly relied only on adhesion between rails and wheels.

The following sections provide an introduction to the classification and design of locomotives and wagons with a focus on existing modern designs, and then the principles of their operational application in train consists are discussed with details of the practical configurations in use for hauling the long and heavy modern freight trains.

Locomotive Classification and Design

The proper way to classify freight locomotives is to consider the energy sources used for their motion process. At the uppermost level, there are only two groups in use: nonautonomous and autonomous. A nonautonomous locomotive usually requires energy to be provided from a power source that is located outside the locomotive. A good example of this is an electric locomotive. An autonomous locomotive generates energy from its own power plant mounted directly inside of the locomotive and uses the generated energy for its motion process. Good examples of modern autonomous locomotives are diesel and gas turbine locomotives. Both groups have their own advantages and disadvantages. However, the main advantages of nonautonomous locomotives are a possibility to realize higher traction and to use energy in a more efficient way. The main advantages of the autonomous locomotive are that it does not need a very expensive electrical network infrastructure and it is very mobile in its application on the railway network, that is, it can be used on powered and non-powered railway lines.

It is possible to extend the classification further based on the type of power supply, and the modern freight locomotives are then divided into four groups (Spiryagin et al., 2014, 2017a; Steimel, 2008):

- Electric;
- Diesel;

- Gas turbine; and
- Hybrid locomotives.

An electric locomotive receives electricity from a power station to support its movement. The electricity generated by the power station is transmitted through high-voltage distribution lines to traction substations. The latter perform the transformation of the transmitted electricity to the parameters (voltage, current type, frequency, etc.) required to power traction and other equipment installed in electric locomotives. The transmission of the power from a traction substation to the locomotive pantographs is performed through overhead line equipment. An example of a general equipment layout of an electric locomotive is shown in **Fig. 1**.

A diesel-powered locomotive is commonly equipped with an internal combustion engine as a power plant, typically operating on diesel fuel or, in a limited number of countries, liquefied natural gas. The latter requires the installation of an additional tank for storing liquefied natural gas on the locomotive or it might be attached as a separate wagon adjacent to the locomotive (or shared between two locomotives). An example of a general layout scheme of a diesel-electric locomotive is shown in **Fig. 2**.

Gas turbine locomotives are equipped with a gas turbine as a power plant. This type of locomotive has not found wide application in freight operations yet due to the noise and heat produced during the motion process. However, their design scheme has potentially significant advantages in power density and the cost of fuel, and it is a particularly appropriate design for long-distance trips and extreme cold weather conditions.

Hybrid locomotives, which are similar in design to diesel locomotives, are equipped with an additional energy storage source(s) (electric batteries, supercapacitors, or flywheels) along with a diesel engine (Spiryagin et al., 2015, 2017c). The energy storage source is charged from the diesel generator during locomotive operation when some power is not required for traction purposes, and/or from the locomotive braking process where kinetic energy is transformed into electrical energy. It is a very efficient way to manage

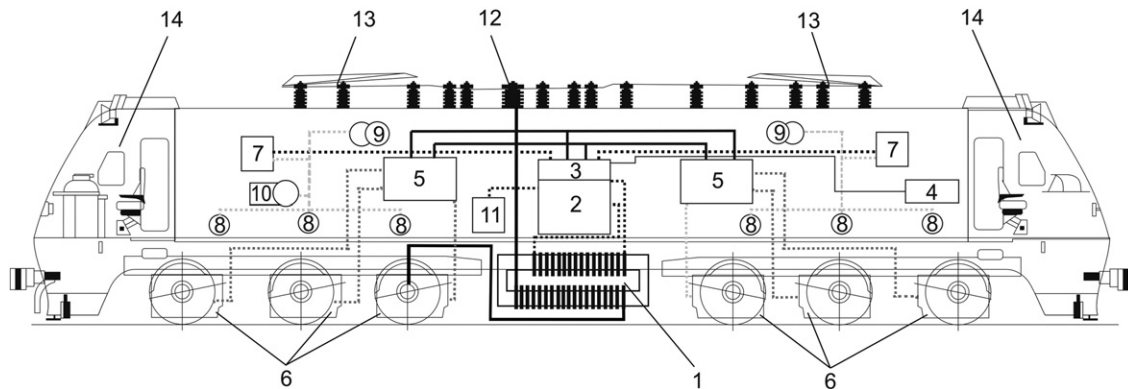


Figure 1 Overview of basic components of an AC electric locomotive. Note: 1—traction transformer; 2—main rectifier; 3—traction control; 4—train control and monitoring system; 5—traction converters; 6—traction motors; 7—auxiliary converters; 8—motor blowers; 9—cooling fans; 10—air compressor; 11—battery equipment; 12—circuit breaker; 13—pantographs; and 14—driver's compartments (cabs).

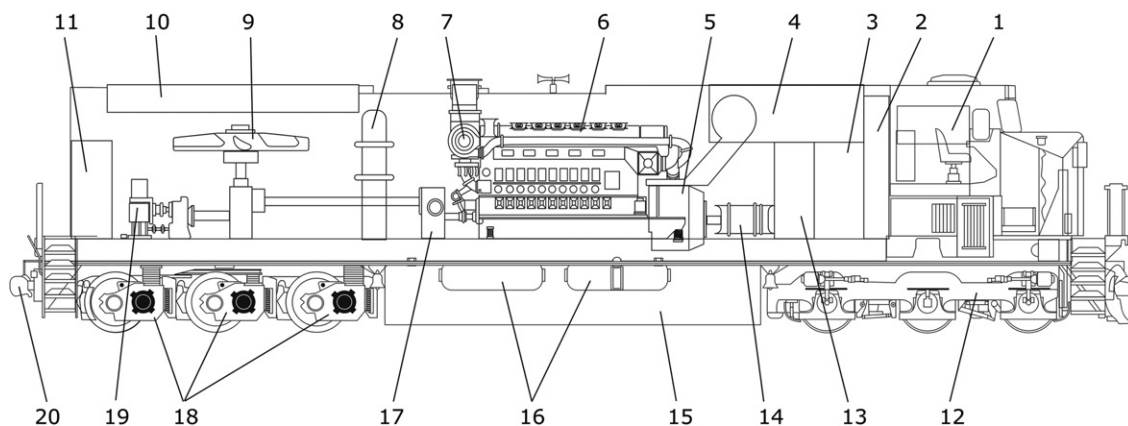


Figure 2 General equipment layout of a diesel-electric locomotive. Note: 1—driver's compartment (cab); 2—electronic control equipment; 3—high-voltage chamber; 4—centralized air intake system; 5—main alternator; 6—diesel engine; 7—turbocharger; 8—water-oil heat exchanger; 9—cooling fan; 10—cooling sections; 11—sand box; 12—bogie (truck); 13—motor-fans; 14—auxiliary generator; 15—fuel tank; 16—air reservoirs; 17—gear boxes; 18—traction motors; 19—brake compressor; and 20—coupler.

energy resources. However, the freight hybrid locomotive designs are still only at the prototyping stage due to the significant costs of hybrid power and energy storage equipment along with the associated additional operating costs.

A further breakdown of the classification of freight locomotives might be done by types of use. There are only two groups: freight service and heavy haul operation. The main difference between these is that heavy haul freight locomotives are designated to haul very long trains with extreme axle loads and these types of locomotives can be commonly found on iron ore, coal, or other mineral traffic railways.

The general design classification of freight locomotives can be systematized as follows (Spiryagin et al., 2014, 2017a):

- By gauge of the track: There are a lot of different gauges in the world, but the main gauges of the freight railways in most countries can be divided into three gauges: “narrow gauge,” “standard gauge,” and “broad gauge.” The standard gauge of 1435 mm (4 ft 8 1/2 in) represents approximately 60% of global railway track length at the present time.
- By speed: There are two common groups—regular freight and high-speed freight. The latter is still not common these days.
- By an envelope that represents a limiting rolling stock outline: In other words, this limits dimensional characteristics of rail vehicles in order to allow them to operate on designated railways.
- By maximum power of the locomotive: There are commonly two characteristics in use—power generated by the power plant (e.g., engine) and the traction power delivered to the wheels. The difference in these characteristics is that the power plant power shows the maximum power produced by an engine only and does not take into account any losses in transmission systems. As a result, it means that the total power at the wheels is undefined, but it is clear that the power at the wheels is slightly lower than that generated by the power plant. The traction power much more closely reflects the real world because it shows the available traction power or force applied tangentially from all traction wheels to the rails. The values of power are usually presented in horsepower or kilowatts units. The power of freight locomotives is commonly positioned in the range from 2000 hp to 6000 hp.
- By the car body design, which might be classified as the cab unit or the hood unit and, at the same time, by the number of driving cabs, that is, one-cab or two-cab designs as shown in Fig. 3(A).
- By the number of units in a locomotive set: Common designs for freight locomotives can have from one to four units. In multiunit locomotive designs, a unit without a driving cab can be present and such a unit is commonly called a booster unit. In the general terminology, a driving cab unit is defined as an A unit, as shown by the examples in Fig. 3(A), and a booster unit is defined as a B unit. An example of a locomotive set with three (A-B-A) units is shown in Fig. 3(B).
- By the wheelset/axle vertical loads: These characteristics are critically important because avoiding any track damage or derailments requires that the maximum loads allowed by the track design are not exceeded.
- By the set-up of driven axles/wheelsets: The International Union of Railways (UIC) classification of locomotives has traditionally been used to specify the arrangement of axles within the same bogie (truck) starting from the front end of the locomotive by alphabetical symbols for the number of consecutive motorized axles (A for one, B for two, C for three, etc.). The use of a lower case “o” as a suffix after the letter is also in use to indicate an individually driven wheelset, that is, a wheelset that has its own separate traction motor.

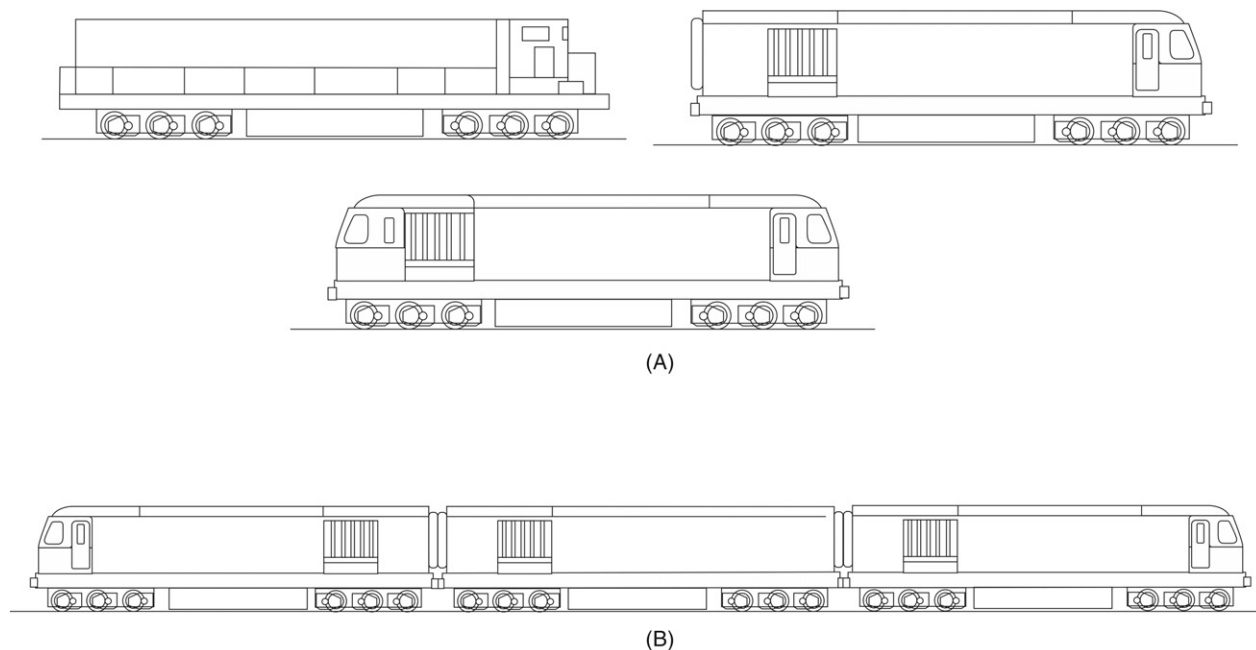


Figure 3 Types of locomotive sets. Note: (A) one- and two-cab locomotive designs; and (B) three unit locomotive set with a booster as the middle unit.

- Locomotives usually consist of the following basic systems for the energy source (power plant): mechanical, electrical, pneumatic, and hydraulic. However, in terms of general design, locomotives are commonly classified by their components as follows: locomotive car bodies and frames, bogies, traction drives, suspension systems, and brake systems.

The car body of the locomotive is designed to accommodate the driver/s and various items of system equipment and components, and it must be able to cope with the application of all external and internal loads. It is commonly designated in two variants: a mainframe design where the frame is considered as a load-bearing component and a monocoque design. The latter integrates the body shell and the mainframe into a single structural element designed to resist all loads acting on it.

The bogies are used in a rail vehicle design to improve the dynamic performance of a vehicle in curved track sections. Depending on the design parameters of locomotives (weight, length, tractive effort, etc.) and restrictions due to loading gauge and axle load, bogies can be designed in two-, three-, and four-axle design variants. At the present time, the four-axle design is not commonly in use for freight train operations. However, it is more applicable for locomotives that are required to haul a long train over track that has axle load restrictions. Examples of designs of two- and three-axle bogies are shown in **Figs. 4 and 5**, respectively. The main elements of a bogie are the bogie frame, the braking system equipment, elements of the locomotive sanding system, spring suspension, and wheelsets with associated assemblies and traction drives.

The spring suspension can be divided in two systems—secondary and primary suspension systems. A locomotive frame/car body and bogies are connected with the secondary suspension system that supports the car body on the bogie frames. This system is commonly not a part of the system that is designated to transmit traction and braking forces from the bogies to the car body. In most locomotive design cases, special connection elements such as traction rods or a center pivot are in use. The primary suspension connects the wheelset or its axle boxes with a pair of bogie frames and is designed to minimize the effects resulting from dynamic interaction at the wheel–rail interface. Traction drives are commonly considered as part of a bogie in the freight locomotive design. The main purpose of traction drives is to transfer kinematic power from the traction motors through the mechanical gear transmission to the wheelsets. Various designs of drives exist, and they are commonly classified based on the design of wheelsets/wheels and the mounting methods of the traction motor. The traction drive design for freight locomotives is commonly customized for that locomotive type and operational use.

The main task of the brake system is the creation of an artificially controlled resistance force by a locomotive or train in order to reduce the speed or bring its movement to a full stop. Brake systems are divided into two groups based on the method used to create the resistance force, these being either friction or dynamic. The friction brake requires energy to be absorbed by the friction (brake shoes or pads), while the dynamic brake of a freight locomotive works on principles of transformation of kinetic energy of the

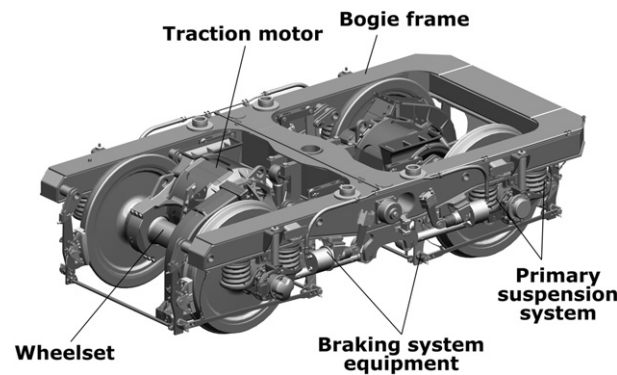


Figure 4 Two-axle locomotive bogie (manufactured by Ural Locomotives, Yekaterinburg, Russia).

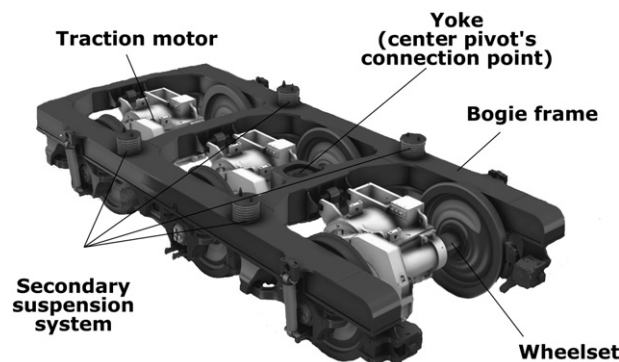


Figure 5 Three-axle locomotive bogie (manufactured by United Group Limited, Newcastle, Australia).

locomotive or train into other electrical energy that can be dissipated, transmitted, or stored. The freight locomotive is also equipped with a parking brake that is designed to prevent any movement of a locomotive or train when it is fully stopped or parked on inclined parts of the track.

Locomotive Power Traction Systems

Electric traction systems are widely applied in freight railway systems. The drives are highly efficient, precisely controllable, and complex mechanical transmissions are avoided. The traction system consists of the following elements: a source of electrical energy—either an electrical overhead power system or an on-board generator, the power conditioning equipment, and the traction machines (motors).

Modern diesel-electric locomotives (Spiryagin et al., 2017b) use a diesel motor to drive an alternator/rectifier combination that provides electrical power for the traction system as shown in Figs. 6 and 7. Electric freight locomotives (Spiryagin et al., 2017a) are

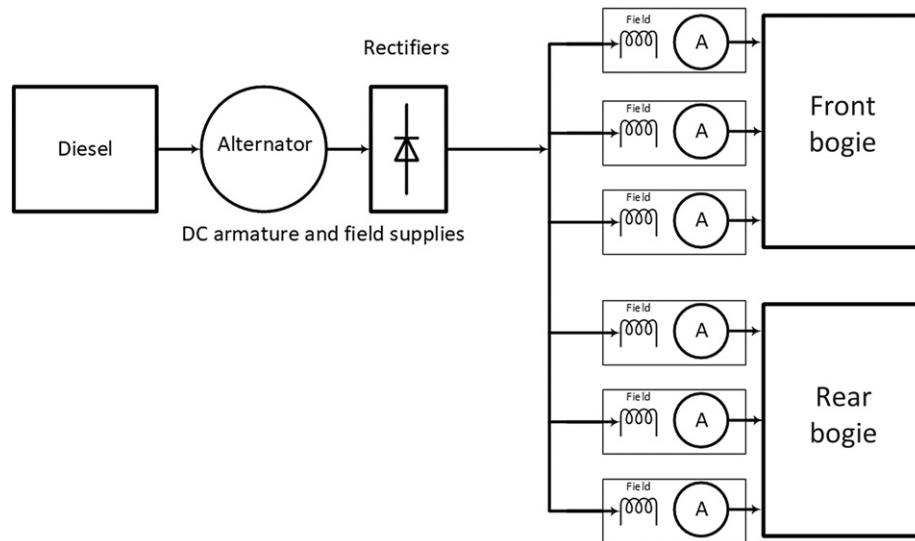


Figure 6 Example of an electric traction scheme for a diesel-electric locomotive with an AC-DC topology.

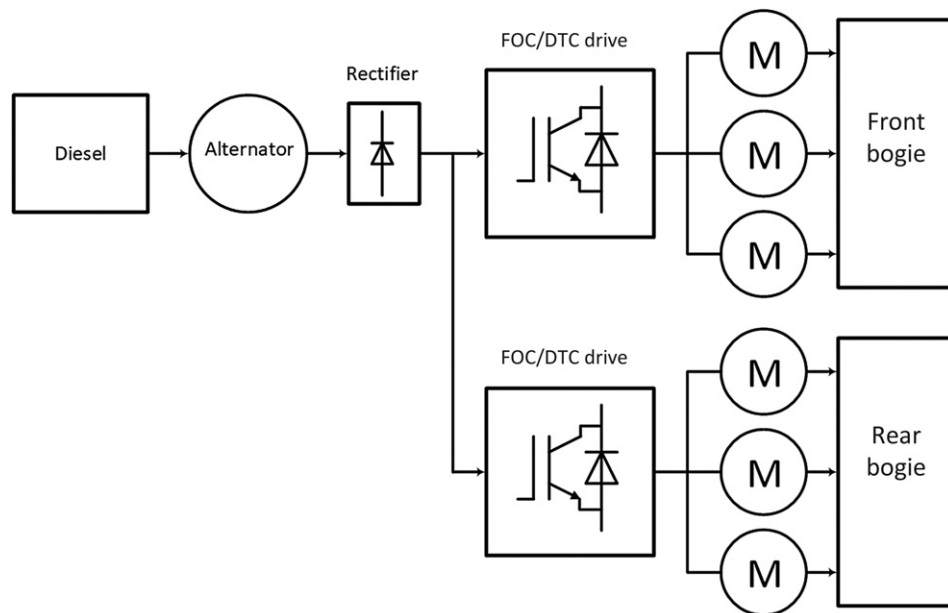


Figure 7 Example of an electric traction scheme for a diesel-electric locomotive with an AC-DC-AC topology.

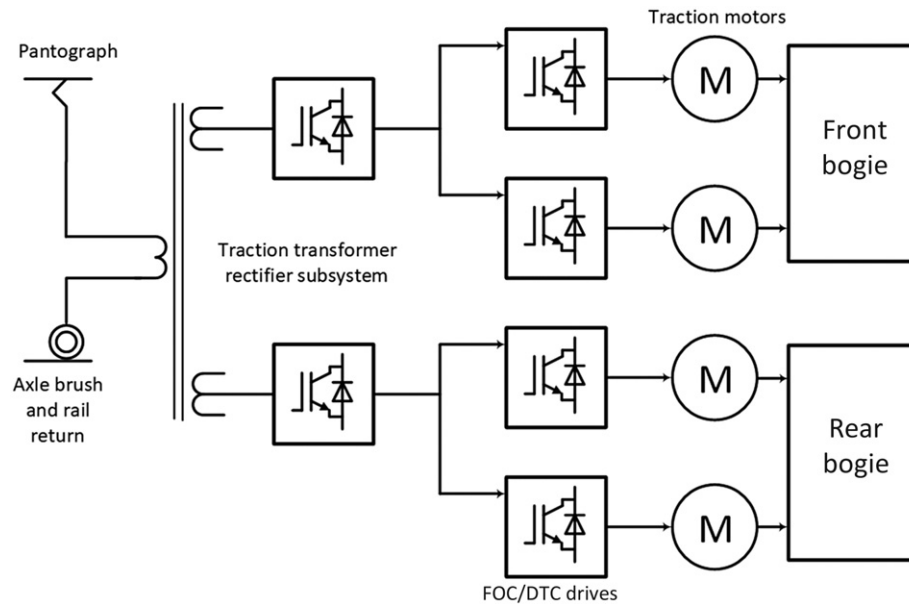


Figure 8 Example of an electric traction scheme for an electric locomotive with an AC-DC-AC topology.

normally supplied with AC power at 25 kV as shown in Fig. 8, but some 3 kV DC systems still exist. Both diesel-electric and electric locomotives can use either AC or DC motors.

AC motors, and generally squirrel cage induction machines, are the dominant technology choice for modern designs, but many older locomotives still operate with DC motors. The AC machine is much cheaper and more robust than the DC machine. The existence of affordable high-performance inverters and inverter control strategies allow an AC machine to be as easily controlled as a DC machine. The features of AC and DC machines are now discussed.

The DC traction machine best suited for heavy haul applications is the separately excited machine. These have independent control of the armature and field windings. There are two clearly identifiable operating regions. The low-speed or torque limited region applies from standstill to approximately one-third of the freight locomotive top speed. Above this, the high-speed operating region, which is also known as the field weakening or constant power region, occurs because of limitations on the motor voltages.

A freight locomotive will develop its maximum tractive effort in the low-speed or torque limited region. This is highly beneficial as the largest tractive forces are available from standstill to roughly one-third of the maximum locomotive speed. The motor field is maintained at its full rated current and the machine torque is limited by the maximum allowable armature current. In many locomotives, it is permissible to allow the motors to operate at higher currents for short periods. This allows even higher tractive forces for relatively short events such as climbing steep grades.

As the armature voltage is dependent upon the vehicle speed and the field winding magnetizing flux, there is a speed at which the motor voltage will equal the available source voltage. To operate at higher speeds, the field strength must reduce. This is the field-weakening region. The available motor torque, and hence locomotive tractive effort, reduces as the locomotive speed increases.

For DC motors, the torque and power are primarily controlled by adjusting the armature current. In a diesel-electric locomotive, the armature current is easily controlled by adjusting the on-board generator voltage through the control of the generator exciter winding. In an AC-supplied electric locomotive with DC motors, the armature circuit is supplied by a transformer and a thyristor phase-controlled rectifier. In both locomotive types, it is normal to have wheel slip control systems. If a wheel loses adhesion and wheel slippage occurs, the armature current will be rapidly reduced to regain adhesion. Adhesion control allows modern locomotives to produce the highest possible tractive efforts even in poor wheel-rail interface conditions.

A DC traction machine will have a rating up to several hundred kilowatts. Machines of this size have compensation windings or interpole windings to maintain the best operating conditions for the commutator. The commutator imposes a number of design constraints. The commutator diameter increases with voltage and the armature voltage will be limited to less than 1000V.

The AC traction machine most favored for modern heavy haul designs is the induction machine. The preferred machines are squirrel cage machines with four or six poles. These have no electrical connection to the rotor and no slip rings or commutators are present. AC traction motors will have power ratings in the range of a few hundred kilowatts to above 1 MW. The motor rated voltages will range up to 3000V and this allows lower currents.

The machines will be controlled by inverters that convert a DC power source into variable frequency AC power. Modern inverters generally use insulated gate bipolar transistors (IGBTs) but Gate Turn Off Thyristors (GTOs) are still found in some designs. The rotational speed of an induction machine is nearly directly proportional to frequency. Modern AC traction systems utilize

Field-Oriented Control (FOC) or Direct Torque Control (DTC) methods that provide precise control of the induction machine torque. An induction machine exhibits constant torque and constant power modes of operation that are very similar to those found with DC machines. At lower speeds, from standstill to up to about one-third of the locomotive top speed, the traction machines can produce their full torque capability. The machines operate with their rated magnetic flux and the torque is limited by the permissible stator and rotor currents. In the higher speed range or field-weakening range, the achievable magnetizing flux is limited by the available voltage. As the magnetizing flux falls, the achievable torque reduces.

The FOC and DTC control methods can provide very precise traction and adhesion control. Diesel-electric locomotives use a main generator to provide power to the inverters. It is normal practice to impose a torque rate limit upon the AC drive so that the diesel and generator combination is able to follow the changing demands in electrical power. For those locomotives that derive their power from an overhead system, rapid changes in tractive effort are possible.

Both DC and AC motors can provide dynamic braking. Both types of motors can be controlled to act as generators, and this provides a retarding force as the kinetic energy in the train is converted into electrical energy. An important benefit is the avoidance of wear in mechanical braking systems that rely on friction. This electrical energy must be either dissipated, stored, or returned to an overhead power system. The most common solution is to dissipate the electrical energy in braking resistors. These will often have ratings of a few megawatts. They can be mounted on the roof or within the locomotive in which case they will be force cooled using fans. In an electric locomotive, some power may be returned to the overhead power system. The amount of power that can be returned may be limited by voltage rises within the overhead power system. Electric locomotives will often blend the two approaches and regenerate some power into the overhead system and dissipate the remainder.

Hybrid traction systems use batteries or supercapacitors to capture some of the braking energy and store this for future use. Hybrid systems provide the most benefits in railway systems where braking is frequent. Braking occurs frequently in metropolitan passenger railways, but is less common in freight systems. In a freight system, the energy storage requirements can be several megawatt-hours, and this may require batteries with weights of tens of tons. Batteries can be integrated into the locomotive structure or an alternative solution is to use a battery tender that is a purpose designed battery wagon. Hybrid solutions are emerging but are not yet common.

Tractive Effort and Dynamic Braking Characteristics

The main performance descriptors for freight locomotives are generally based on their electro-traction transmission characteristics, which generally describe the tractive and braking capabilities of a locomotive and such characteristics are dependent on the train speed, adhesion conditions at the wheel–rail interface, and the load being hauled (Spiryagin et al., 2017a).

Locomotive tractive effort characteristics are usually presented by a graph of dependence between the tangential tractive effort and the speed at a given engine power (notch position of the throttle set by a driver). An example of characteristics for a diesel-electric locomotive is shown in Fig. 9. In this example case, the tractive effort characteristic also contains a restriction on the tractive effort due to the effect of an adhesion coefficient between wheels and rails, and the limit on the tractive effort set by the power plant.

Tractive effort characteristics of electric locomotives are different from those of diesel-electric locomotives in that the shape for the former is parabolic. An example of tractive effort characteristics of AC current electric locomotives is shown in Fig. 10 which displays

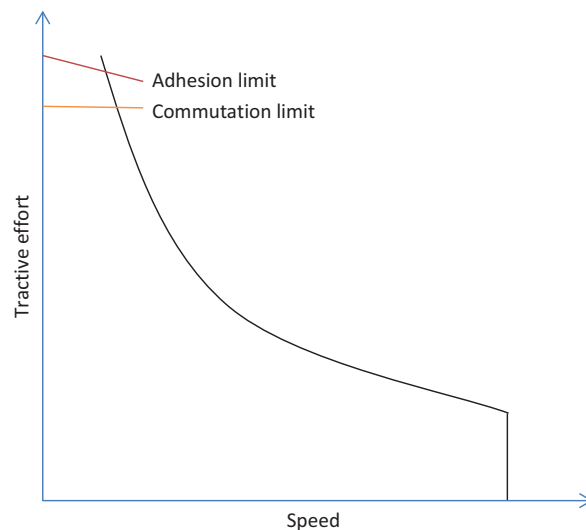


Figure 9 Typical tractive effort versus speed characteristics of a diesel-electric locomotive.

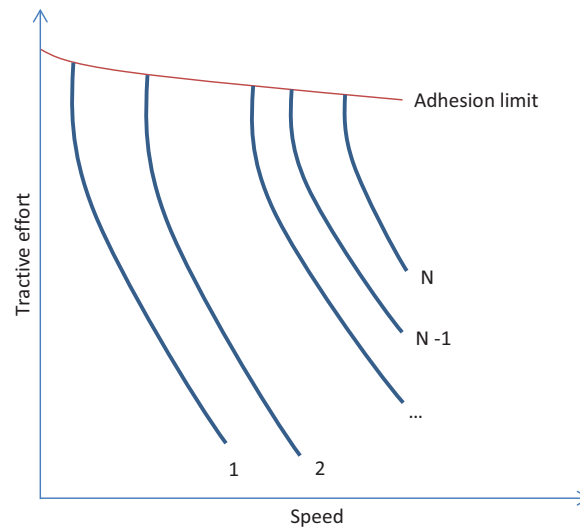


Figure 10 Typical tractive effort versus speed characteristics of an AC current electric locomotive for various throttle notch positions numbered 1 to N .

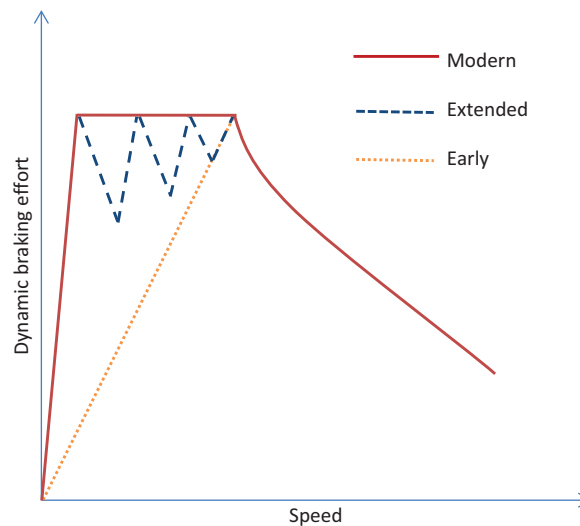


Figure 11 Typical dynamic braking effort versus speed characteristics of a diesel-electric locomotive.

different curves for different notch positions of the throttle (represented by the numbering 1 to N) which can be set by the locomotive driver.

Examples of dynamic braking characteristics are shown in **Fig. 11**. The figure contains three curves that represent different stages of the dynamic braking developments in locomotive designs. The modern curve is generally observed on modern AC freight locomotives, while the extended one is common for DC locomotives. At a low-speed range, less than 5–10 km/h, the traction motors cannot provide the required braking effort due to insufficient terminal voltages.

Freight Wagon Classification and Design

Large railway networks require a systematic freight wagon classification system to facilitate efficient rolling stock management. Freight wagon classification systems are commonly based on vehicle uses and mechanical designs. Again, as mentioned in the introductory section, different countries and regions can have different classifications under both methods, but common rationales can be seen within these different classification methods. A good understanding of any classification system can help to achieve an easier understanding of others. In this article, classifications used in Chinese Railways are taken as an example (Yan and Fu, 2008). Chinese Railways defines most of its wagons using 14 types, namely, box wagons (enclosed body), hopper wagons (open body), flat wagons, refrigerated wagons, tanks wagons, container wagons, ore wagons (bottom or side dump), oversize wagons (drop center

and Schnabel type), dangerous goods wagons, stock wagons (cattle), cement wagons, grain wagons, special wagons, and caboose wagons (becoming obsolete). Each type has its specific alpha registration code. For example, all tank wagons are registered with the initial “G” that is also the initial of the Chinese spelling for tank wagons. Individual wagons are then given unique numerical identifications.

Chinese Railways also categorizes different types of freight wagons into two categories according to their use as: (1) general-purpose wagons; or (2) dedicated-purpose wagons.

- General-purpose wagons include box wagons, hopper wagons, flat wagons, and refrigerated wagons. These wagons are designed to transport cargoes of a considerable number of different types. For example, a box wagon, as shown in [Fig. 12](#), has an enclosed cargo space and sliding side doors. It can transport many different types of cargoes such as bagged or boxed products, small machinery, and small items of merchandise: literally anything that can fit in the wagon.
- Dedicated-purpose wagons are those designed to transport a single or small number of types of cargoes. For example, a tank wagon, as shown in [Fig. 13](#), can transport only a small number of different types of liquid and liquefied gas cargoes such as acids, petroleum products, LNG, LPG, etc.

Bogie design is probably the most important part of the mechanical design of freight wagons. Discussions of freight wagon designs in this article are based on the different bogie designs. Currently, worldwide, there are mainly three types of freight bogie and running gear designs ([Iwnicki et al., 2015a](#)): (1) three-piece bogies, (2) Y-25 bogies, and (3) single axle running gear.

- A three-piece bogie ([Fig. 14](#)) mainly consists of a bolster, two side frames, and two wheelsets. The suspension system is a wedge-spring system in which the wedges generate friction forces to damp vibrations and impacts for the bogie ([Wu et al., 2018](#)).
- A Y-25 bogie is shown in [Fig. 15](#). It mainly consists of a rigid H-shape frame and two wheelsets. Flexibility of the bogie is provided by the suspension springs. A so-called Lenoir link is used for each quarter of the bogie to convert part of the vertical load into friction normal forces between a pusher and the frame. In this way, friction forces can be generated in suspensions to damp vibrations and impacts ([Vaghi et al., 2013](#)).



Figure 12 Box wagon—a general-purpose wagon.



Figure 13 Tank wagon—a dedicated-purpose wagon.

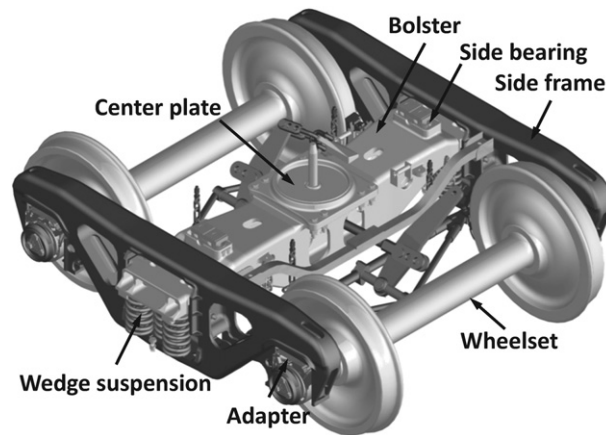


Figure 14 Three-piece freight wagon bogie. Source: From Wu et al. (2018).

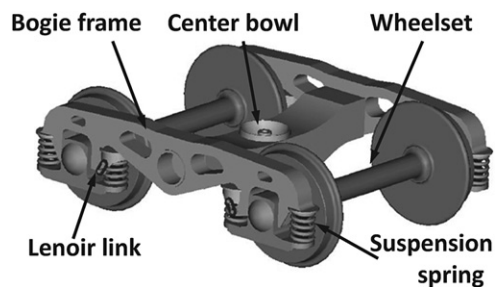


Figure 15 Y-25 freight wagon bogie (brake rigging not shown).

- A single axle running gear is the oldest running gear arrangement used when no bogie design is needed in the freight wagon design. The UIC-Link Suspension design provides a double-link between the leaf spring and the wagon body that allows yaw and lateral movement of the wheelset for better curving performance (Jonsson, 2007; Ahmad, 2018).

Freight Train Configuration and Distributed Power

The concept of distributed power freight trains (Rail Industry Safety and Standards Board, 2018) is to use multiple numbers of locomotives or locomotive groups (commonly two or three) and place the locomotives at different positions along the train (head, middle, or tail) instead of placing all locomotives at the front of the train. Example of various distributed power freight train configurations is shown in Fig. 16.

The distributed power train concept was first put into service in the 1960s in the United States. It has two obvious advantages: (1) decreasing in-train forces during traction as the in-train coupler forces (Cole et al., 2017; Wu et al., 2017) are distributed at different locations along the train, as shown in Fig. 17. And (2) decreasing in-train forces and braking distance when slowing or stopping, as shown in Fig. 18. During braking, locomotives at the middle or the tail of the train act as extra brake initiating points that can decrease brake delays throughout the train and, in turn, decrease braking distance and in-train forces.

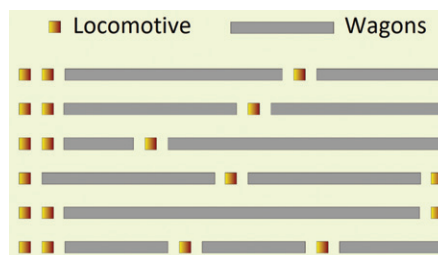


Figure 16 Example of distributed power freight train configurations.

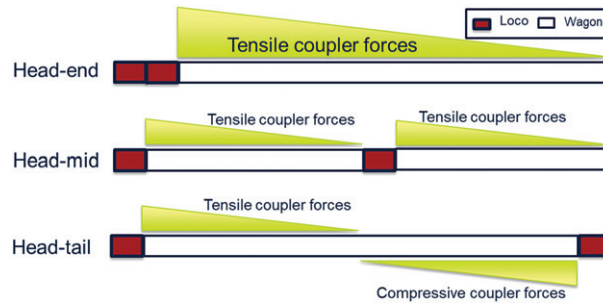


Figure 17 Simplified in-train force distribution during traction.

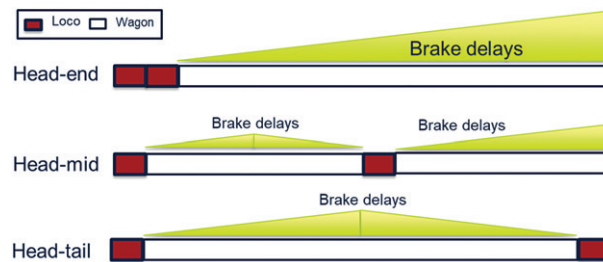


Figure 18 Simplified brake delay distribution during slowing/stopping.

Driverless Freight Train Operation

Despite the apparent significant advances in autopilot and automated control systems for ground vehicles, examples of driverless freight train operation are rare. There are slowly increasing numbers of automated metro and airport passenger trains, so one could conclude that sufficient technology is available. Perhaps the best-known driverless freight train system is that of the Karratha Hamersley railway in the Pilbara region of Western Australia. As reported in [Briginshaw \(2018\)](#), a train route of over 400 km is now operating in driverless mode. This impressive rail system operates an average of 34 trains per day with train payloads typically greater than 28,000 tons.

Full automation of freight trains is complicated by the train length. Trains are the longest ground vehicles; small freight trains might be 400 m long, while state-of-the-art heavy haul trains may be as long as 4 km. The challenge posed for the control, either by a human driver or by an automated control system, is that the vehicle can be subject to several different operational conditions at the same time. In particular, the train can be subject to different speed limits and different curves and grades. In long trains, these forces from grades (gravity forces) can be transferred through the train to add to locomotive traction forces and braking forces and increase risks of train coupling failure.

Large freight trains also have very low power to mass ratios. This means that traction and braking will effect changes only slowly. Trains will take several minutes to get up to speed as the traction forces give limited acceleration. Dynamic and regenerative braking is similarly limited. Both traction and dynamic braking are limited by the friction at the wheel–rail contact, and so performance is limited by locomotive power, mass, and the rail surface conditions. The mass of the locomotive plays a role in the maximum performance, as the typical wheel rail friction coefficient will rarely exceed 0.5. Traction and dynamic braking are typically limited to forces less than 35% of locomotive weight to reduce the risk of slipping. To ensure reasonable stopping distances, all freight wagons are also fitted with air-operated brakes, but trains still take a long time to stop. Deceleration for a loaded freight train under full air braking is typically of the order of 0.45 m/s^2 (or just 0.05 g), which gives a stopping distance of at least 525 m from a speed of 80 km/h. If older brake technology is being used, stopping distances can be more than double this figure due to inherent delays in the way the equipment works. Suffice to say that trains, although guided by rails, are not necessarily simple to drive. Good train control requires forward planning of power and braking. The difficulty of this task increases with train length and with the difficulty of the terrain. Compliance with speed is, however, only one issue. Trains store a large amount of kinetic energy due to their mass. Unnecessary stops or reducing too much speed is an expensive waste of energy. On a typical trip, 10%–30% of the total energy is expended in braking. One of the incentives to automate train control is to save some of the energy due to improved or more consistent control. Energy saving from improved control methods is presented in [Yee and Pudney \(2004\)](#). A third issue is the coupling forces. Couplings are generally manufactured to a specified strength. As longer trains are marshaled, the drawbar forces increase due to the double or triple stacking of locomotives and increased gravity forces on grades. Once the train reaches a certain size, the locomotives need to be distributed in the train to reduce drawbar forces as was described in the previous section. On difficult

terrain, that is, on certain combinations of hills and valleys, it is sometimes necessary to use different locomotive controls in the different locomotive groups to reduce coupling forces and prevent train partings.

As control is not just a matter of speed compliance, driverless control of locomotives requires several sources of data and algorithms to replace the many thought processes and judgments made by a human driver. As a minimum requirement, the control system needs to have a full track database describing the topography, to allow gravity forces on grades to be assessed. The system needs speed limit information to ensure speed compliance. The system needs full details of the train mass, rolling resistance, curve resistance, and braking characteristics. Positioning systems such as GPS or suitable track signaling equipment is required to ascertain train position and, therefore, the track topography to be applied. The technology to enable these developments now exists.

Prior to the emergence of large driverless train systems, such as in Briginshaw (2018), several developments that give information, assistance, and advice to human drivers were developed. Many products and systems have been developed during the last 2 decades, noting the following systems: driver information (Hawthorne and Smith, 1998; Yee and Pudney, 2004; Cole et al., 2007), driver advice (Yee and Pudney, 2004), and cruise control (Cole and McLeod, 2004; United States Patent, 2008; Cole et al., 2009). No doubt, these provided the technical steps along the way to achieving driverless freight trains.

Conclusions

The optimization of locomotive and wagon designs, as well as of train operating configurations, are very important aspects in improving the efficiency of the railway land transportation market. The designs of powered and unpowered rail freight vehicles can vary quite significantly, and can result in major modifications to the design information presented in this article. The authors' intent in writing this article is to provide a basic knowledge on common freight vehicle designs and their classification. Powered rail vehicles were also covered in terms of the design of their traction systems and how to interpret their traction and dynamic braking design characteristics. In addition, the authors have placed strong emphasis on the various train configurations that can be found in real train operations around the world. The contents of this article are structured in a manner that allows readers to become acquainted with the common terminology used in the rail vehicle and train operational world. We hope that readers have found the information presented in an easily readable manner for the general public and that, at the same time, it is sufficiently comprehensive for railway practitioners. The references provided at the end of this article will allow readers to further explore the world of rail freight vehicles from descriptions of more detailed design issues and exploring some of the complex train operational problems that exist in freight railway systems.

Readers with enquiries regarding rail freight vehicle design or train configurations can contact the Centre for Railway Engineering at Central Queensland University by email at cre@cqu.edu.au or visit the website at www.cqu.edu.au/cre.

References

- Ahmad, S.N.N., 2018. Analysis of a very low tare mass wagon concept for intermodal freight. Doctoral thesis. School of Engineering and Technology, CQ University, Rockhampton, Australia.
- Briginshaw, D., 2018. Rio Tinto completes automation of Pilbara rail network. *Int. Railway J.* Available from: <https://www.railjournal.com/freight/rio-tinto-completes-automation-of-pilbara-rail-network/>.
- Cole, C., McLeod, T., 2004. Optimising train operation using simulation, fuzzy logic cruise control and evolutionary algorithms. *Proceedings of the 5th Asia Pacific Industrial Engineering and Management Systems Conference*, December 12–15, 2004, Gold Coast, Australia, pp. 30.6.1–30.6.16.
- Cole, C., Scott, S., McClanachan, M., McLeod, T., Phelan, J., 2009. Train cruise control. *Proceedings of the 9th International Heavy Haul Conference*, June 22–24, 2009, Shanghai, China, pp. 639–645.
- Cole, C., Simson, S., Bosomworth, C., Hayman, M., McLeod, T., 2007. Advances in in-cabin systems for the management of long trains. *Proceedings of the 9th International Heavy Haul Conference*, June 11–13, 2007, Kiruna, Sweden, pp. 305–312.
- Cole, C., Spiriyagin, M., Sun, Y.Q., 2013. Assessing wagon stability in complex train systems. *Int. J. Rail Transp.* 1 (4), 193–217.
- Cole, C., Spiriyagin, M., Wu, Q., Sun, Y.Q., 2017. Modelling, simulation and applications of longitudinal train dynamics. *Vehicle Syst. Dyn.* 55 (10), 1498–1571.
- Hawthorne, M.J., Smith, N., 1998. On-board train management—LEADER proves it's worth. *Proceedings of Conference on Railway Engineering*, September 7–9, 1998. Railway Technical Society of Australasia, Rockhampton, Australia, pp. 265–272.
- Iwnicki, S.D., Stichel, S., Orlova, A., Hecht, M., 2015a. Dynamics of railway freight vehicles. *Vehicle Syst. Dyn.* 53 (7), 995–1033.
- Jonsson, P.-A., 2007. Dynamic vehicle-track interaction of European standard freight wagons with link suspension. Doctoral thesis in Railway Technology. Royal Institute of Technology, Sweden.
- Rail Industry Safety and Standards Board, 2018. *Distributed Power Freight Trains: Code of Practice*, Version 1. Spring Hill, Australia.
- Spiriyagin, M., Cole, C., Sun, Y.Q., McClanachan, M., Spiriyagin, V., McSweeney, T., 2014. *Design and Simulation of Rail Vehicles*. Ground Vehicle Engineering Series. CRC Press, Boca Raton, FL.
- Spiriyagin, M., Wolfs, P., Cole, C., Spiriyagin, V., Sun, Y.Q., McSweeney, T., 2017. *Design and Simulation of Heavy Haul Locomotives*. Ground Vehicle Engineering Series. CRC Press, Boca Raton, FL.
- Spiriyagin, M., Wolfs, P., Cole, C., Stichel, S., Berg, M., Plöchl, M., 2017b. Influence of AC system design on the realisation of tractive efforts by high adhesion locomotives. *Vehicle Syst. Dyn.* 55 (8), 1241–1264.
- Spiriyagin, M., Wolfs, P., Szanto, F., Sun, Y., Cole, C., Nielsen, D., 2015. Application of flywheel energy storage for heavy haul locomotives. *Appl. Energy* 157, 607–618.
- Spiriyagin, M., Wu, Q., Wolfs, P., Sun, Y., Cole, C., 2017c. Comparison of locomotive energy storage systems for heavy-haul operation. *Int. J. Rail Transp.* 6 (1), 1–15.
- Steimel, A., 2008. *Electric Traction—Motive Power and Energy Supply—Basics and Practical Experience*. Oldenbourg Industrieverlag GmbH, Munich, Germany.
- United States Patent, 2008. Control system for operating long vehicles. United States Patent No. 7,359,770, April 15, 2008.
- Vaghi, C., Österle, I., Milotti, A., 2013. Technical implementation, acceptance issues and the role of human factors in rail freight innovation. *Proceedings of XV Conference of the Italian Association of Transport Economics and Logistics (SIET)*, September 18–20, 2013, Venice, Italy, pp. 244–258.

- Wu, Q., Spiryagin, M., Cole, C., 2017. Longitudinal train dynamics: an overview. *Vehicle Syst. Dyn.* 55 (4), 1–27.
- Wu, Q., Sun, Y., Spiryagin, M., Cole, C., 2018. Methodology to optimize wedge suspensions of three-piece bogies of railway vehicles. *J. Vib. Control* 24 (3), 565–581.
- Yan, J., Fu, M., 2008. *Vehicle Engineering*, third ed. China Railway Press, Beijing, China [In Chinese].
- Yee, R., Pudney, P., 2004. Saving fuel on long-haul trains: FreightMiser initial trial results. *Proceedings of Conference on Railway Engineering*, June 20–23, 2004. Railway Technical Society of Australasia, Darwin, Australia, pp. 42.1–42.6.

Further Reading

- AAR, 2015. *Manual of Standards and Recommended Practices, Section C, Part II, Design, Fabrication, and Construction of Freight Cars, M-1001*. Association of American Railroads, Washington, DC.
- International Heavy Haul Association, 2015. In: Leeper, J., Allen, R. (Eds.), *Guidelines to Best Practices for Heavy Haul Railway Operations: Management of the Wheel and Rail Interface*. Simmons-Boardman Books, Omaha, NE.
- Iwnicki, S., 2006. *Handbook of Railway Vehicle Dynamics*. CRC Press, Boca Raton, FL.
- Iwnicki, S.D., Orlova, A., Jonsson, P.-A., Fartan, M., 2015. Design of the running gear for the SUSTRAIL freight vehicle. *Proceedings of IMechE Stephenson Conference: Research for Railways*, April 21–23, 2015, London, UK, pp. 299–306.
- SKF, 2012a. *Railway Technical Handbook, Volume 1—Bogie Designs*. Available from: https://www.skf.com/binary/tcm:12-62732/RTB-1-02-Bogie-designs_tcm_12-62732.pdf.
- SKF, 2012b. *Railway Technical Handbook, Volume 2—Drive Systems: Traction Motor and Gearbox Bearings, Sensors, Condition Monitoring and Services*. Available from: <https://www.skf.com/binary/21-96059/13085EN.pdf>.

Eurasia Rail Freight: Enablers and Inhibitors of Future Growth

Hendrik Rodemann*, **Simon Templar†**, *Unter den Eichen, Wolfsburg, Germany; †Cranfield School of Management, Cranfield University, Cranfield, Bedfordshire, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	436
Enablers and Inhibitors of Eurasia Rail Freight	436
Eurasia Rail Freight: Development and Current Market Position	437
Running Trains Across Eurasia	438
Infrastructure	438
Environmental and Political Framework Conditions	443
Train Length	443
Border Crossings and Gauge Changes	443
Line Speed Versus Travel Speed	443
Policies and Legislation	444
Cost and Subsidies	444
Political and Macroeconomic Conditions	445
Summary	446
Future Outlook	449
Acknowledgment	451
References	451
Further Reading	451
Appendices	452

Introduction

The aim of this paper is to revisit the research study undertaken by Cranfield University in 2013 into the enablers and inhibitors of future growth for rail freight between Asia and Europe. In the last 6 years since the original research was undertaken, the volumes have grown exponentially with an increase of 7868% (Berger, 2017; Pomfret, 2018; Xinhua Silk Road Information Service, 2019; Zhang, 2017) in train movements from 2013 to 2018. It is now time to take another look at Eurasian rail freight and to address two key questions:

- Which enablers and inhibitors still exist?
- Which enablers and inhibitors have changed/emerged?

The British economist John Maynard Keynes advocated that in certain circumstances state intervention into an economy to encourage growth was essential. There are many notable examples over the years where the state has intervened in an economy by investing in capital infrastructure projects, which have been the impetus for future economic growth. They include Roosevelt's New Deal in the 1930s, the Marshall Plan at the end of World War II, and now recently China's President Xi Jinping's Belt and Road Initiative (BRI), which seek to improve land and sea corridors with China and the rest of the world. A key element of the BRI is the investment in upgrading the rail infrastructure that forms the Eurasian land bridge that links China with Europe.

The correlation between the introduction of the railroad industry and economic development can be traced back to the Industrial Revolution in the United Kingdom, with the introduction of the first steam powered freight railway in 1825 that carried coal between Darlington and Stockton-on-Tees. The global expansion of the railroad industry in the 19th and early 20th centuries and its relationship with economic growth is undeniable: a significant example is the Transcontinental Railroad in the United States opened in 1869. The railroad provided an important land bridge that enabled freight and passenger traffic to flow between the east and west coasts of America and facilitated growth at either end and the intermediate railheads along its route. Similarly, the Trans-Siberian railway (TSR) completed in 1916 established a significant land bridge between Moscow and Vladivostok. The BRI is now being heralded as an agent of economic development and a renaissance for the rail freight industry.

Enablers and Inhibitors of Eurasia Rail Freight

An investigation based on exploratory case-study research undertaken by Cranfield University in 2013, when traffic across the Eurasia land bridge was in its infancy, identified a set of enablers and inhibitors that would have an impact on the growth of freight

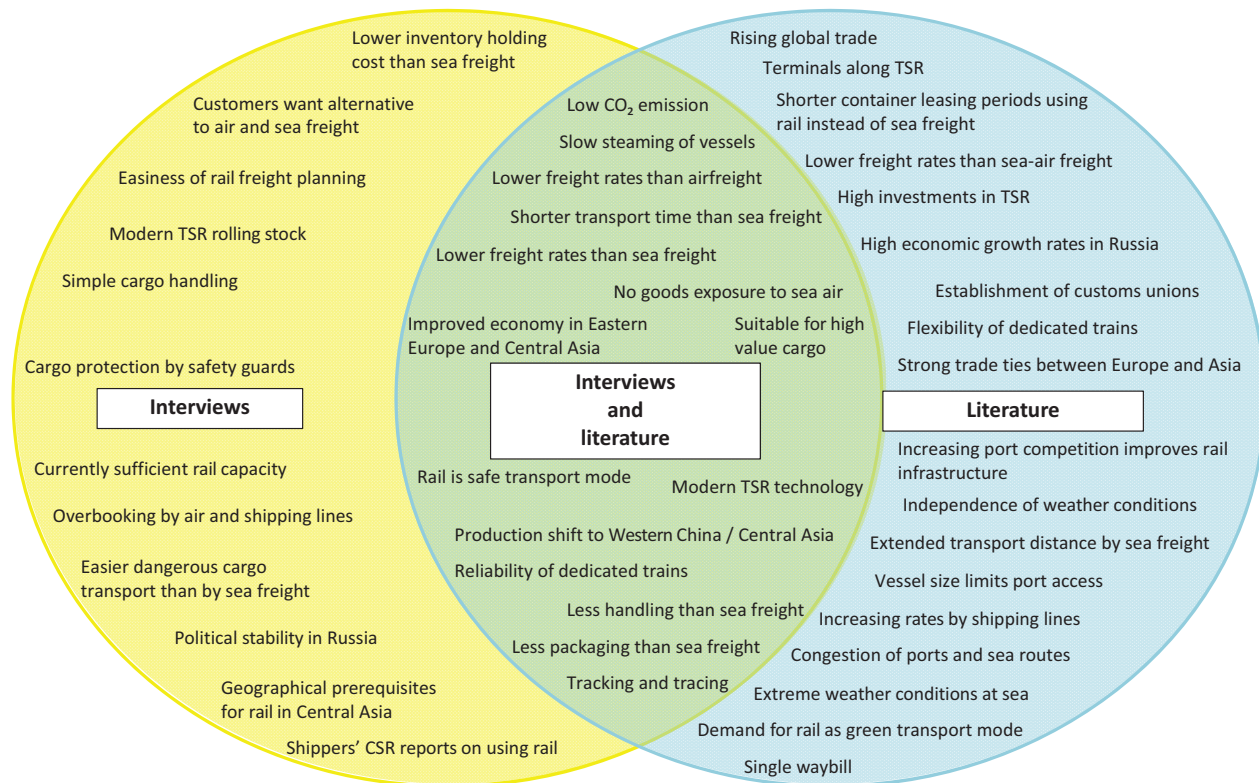


Figure 1 Venn diagram: enablers. *Source: Author.*

train movements over the land bridge. The inventory of enablers and inhibitors was compiled by the research, which was based on a comprehensive review of the relevant literature and was underpinned by primary data collected from 24 semi-structured stakeholder/expert interviews. The outputs of the research study are illustrated in **Fig. 1** (enablers) and **Fig. 2** (inhibitors), with both Venn diagrams illustrating the elements that were identified exclusively from the literature, by stakeholder interviews and common to both sources. Later research adopted similar methods.

The 2013 research identified 45 enablers and 54 inhibitors and these were further classified using PESTLE (political, economic, social, technological, legal, and environmental) analysis and the output is illustrated in **Figs. 3 and 4**. Again, similar methods were adopted in later research.

Eurasia Rail Freight: Development and Current Market Position

The basic idea of using rail freight across Eurasia remains unchanged. Given its characteristics of being faster than sea freight and more economic than airfreight, rail freight fits into a strategic market niche between the two established options of low cost/long lead time (sea freight) and high cost/short lead time (airfreight) (**Fig. 5**). Other alternatives, such as sea-airfreight or intercontinental road freight that might have cost/time ratios comparable to rail freight, are limited with regard to capacity. Rail freight is hence conceptually suitable for cargo that is too time-critical for sea freight (e.g., due to urgency or high capital tied-up), too cost-sensitive for airfreight, and too bulky or heavy for air, sea air, or road freight.

When comparing Eurasian rail freight in 2013 to today, a trend becomes clearly noticeable in terms of increasing transport volume and number of trains (**Figs. 6 and 7**).

The range of services offered by the Eurasia railroad for the years 2013 and 2019 are compared in **Table 1**. The key development in the marketplace is the introduction of a less than full container (LCL) offering.

Table 2 (east-west) and **Table 3** (west-east) compare the variety of goods being shipped by rail along the Eurasia rail bridge for 2013 and 2019.

From this, it can be concluded that the demand for Eurasian rail freight has increased which, in turn, indicates an expansion of the previously introduced market niche. Referring to **Fig. 5**, such expansion comes with a reduction of transport cost and/or transport time and leads to the following hypotheses on the recent development and current state of Eurasian rail freight:

- An increased number of routes and terminals enhances the geographical catchment area of rail freight and thus reduces transport distance to and from terminals;

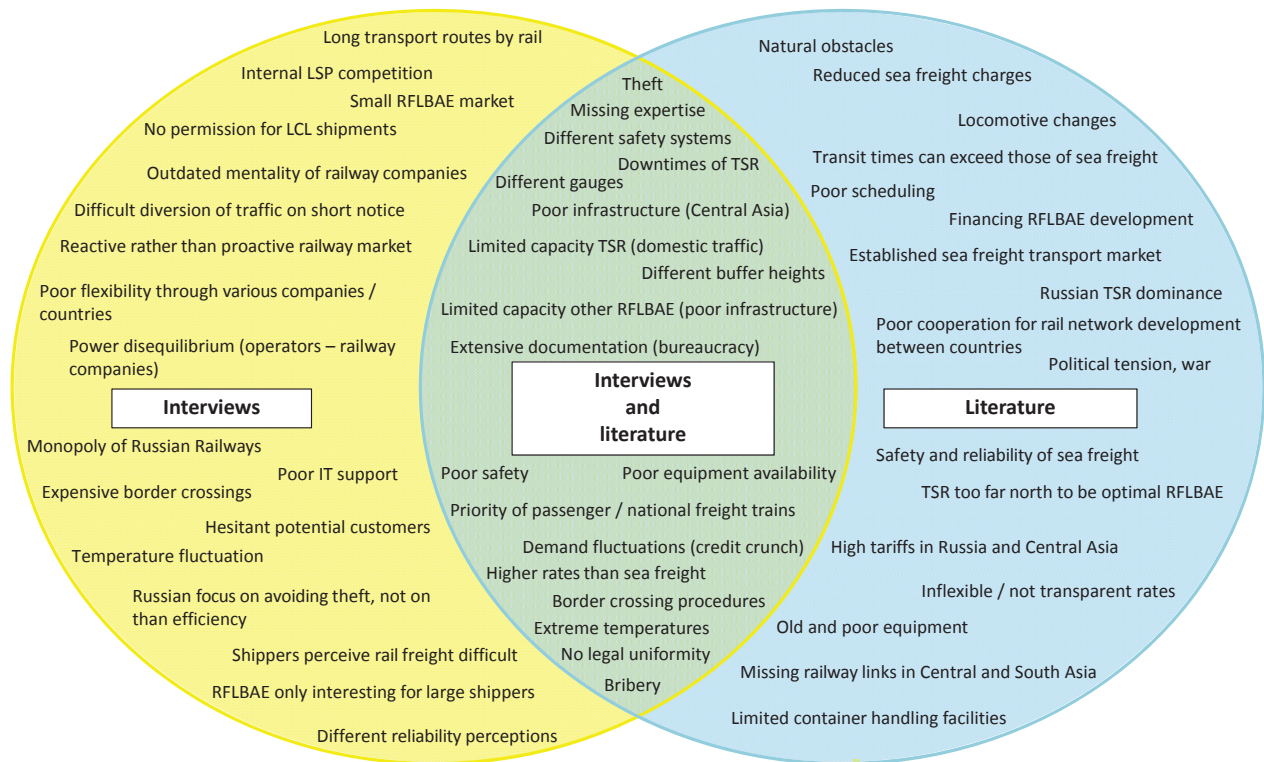


Figure 2 Venn diagram: inhibitors. *Source: Author.*

- An increased number of departures reduces waiting time for individual shipments;
- An increased number of departures reduces cost (e.g., overheads or transshipments); and
- An increased number of routes and departures enhances the importance of intercontinental rail freight for the Eurasian economy and may hence contribute to smoothing processes in order to reduce both cost and time.

The research conducted in 2013 grouped enablers and inhibitors using the PESTLE framework and found the majority of both being technological and economic aspects that are predominantly related to the category's infrastructure, processes, and transport cost. Hence, in the following the recent development of each category is outlined and assessed with regard to its impact on transport cost and time.

Running Trains Across Eurasia

Infrastructure

Fig. 8 shows the current Eurasian railway network and additional connections under construction.

Appendix A provides an overview of the technical characteristics (rail gauges, loading gauges, safety/signaling systems, power source, degree of electrification, and double tracks) of the individual national railway systems of the major Eurasian/Asian rail freight routes. It becomes obvious that the current Eurasian railway network is far from being integrated in terms of track and loading gauge, safety systems, propulsion method, and the degree of track electrification. As a change of these preconditions is extremely unlikely in the near future, infrastructure development predominantly focuses on the improvement and extension of certain (national or regional) networks and neglects the realization of infrastructural uniformity.

In terms of network improvement, tracks in many Central Asian (including Russia) and Eastern European countries have been modernized in order to allow higher speeds, increased weights, and extended lengths of trains. Likewise, many sections have been electrified and equipped with double tracks in order to increase capacity and reduce the number of required locomotive changes. In Central Asia, an increased use of alternative routing options can be witnessed that is related to a broader use (and modernization) of existing infrastructure. While great progress has been made, network modernization has not yet been completed.

Projects to facilitate the extension and further integration of existing networks are in progress, such as the construction of normal gauge routes from China to Turkey (via Central Asia), Pakistan or Singapore (Kunming–Singapore railway), as well as the elongation of the broad gauge network to Vienna, but none have been accomplished to date. In contrast, the number of access points to Eurasian rail freight (and hence the land bridge's geographical catchment area) has already been significantly increased by constructing new terminals or by including existing ones in (new) rail services.

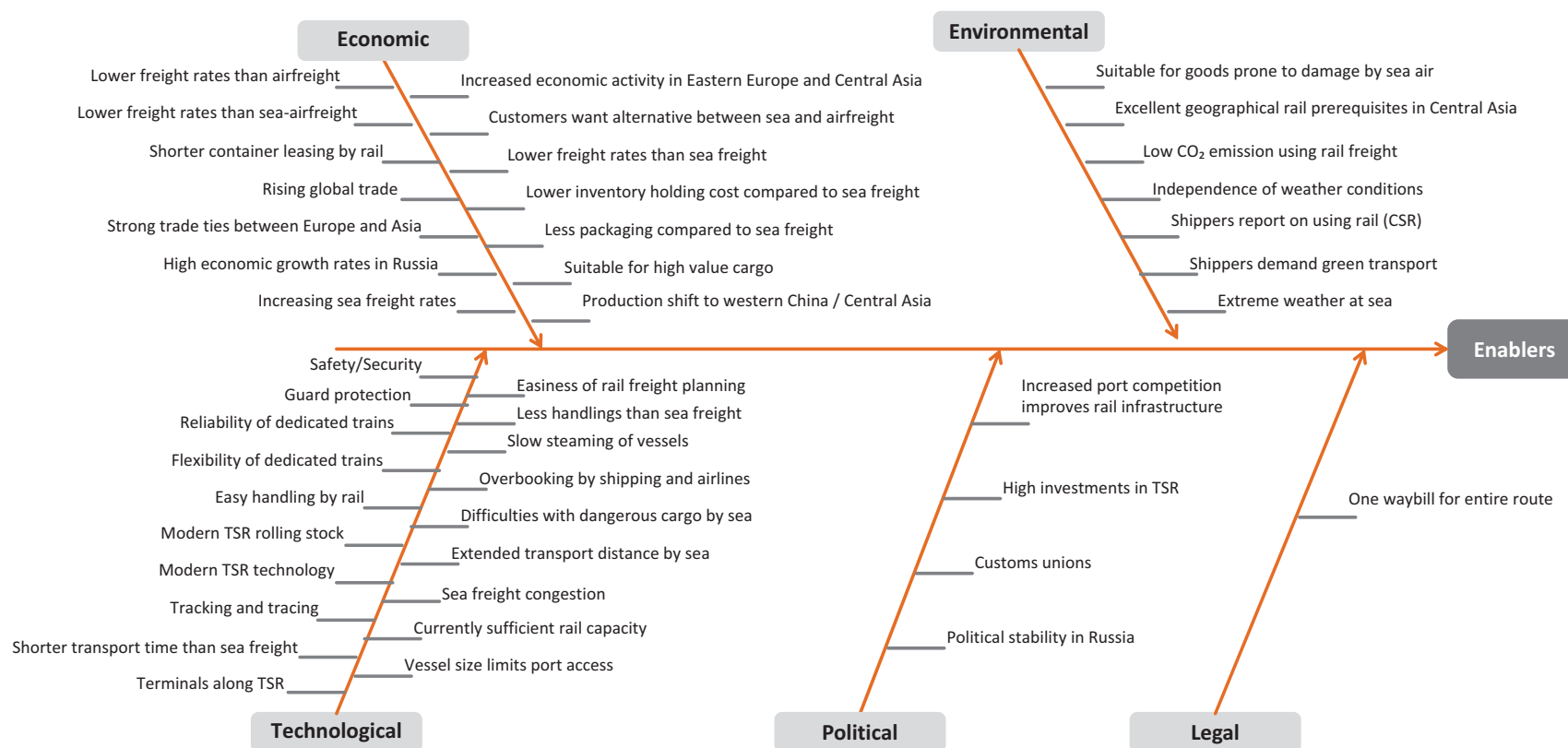


Figure 3 Ishikawa diagram: enablers. Source: Author.

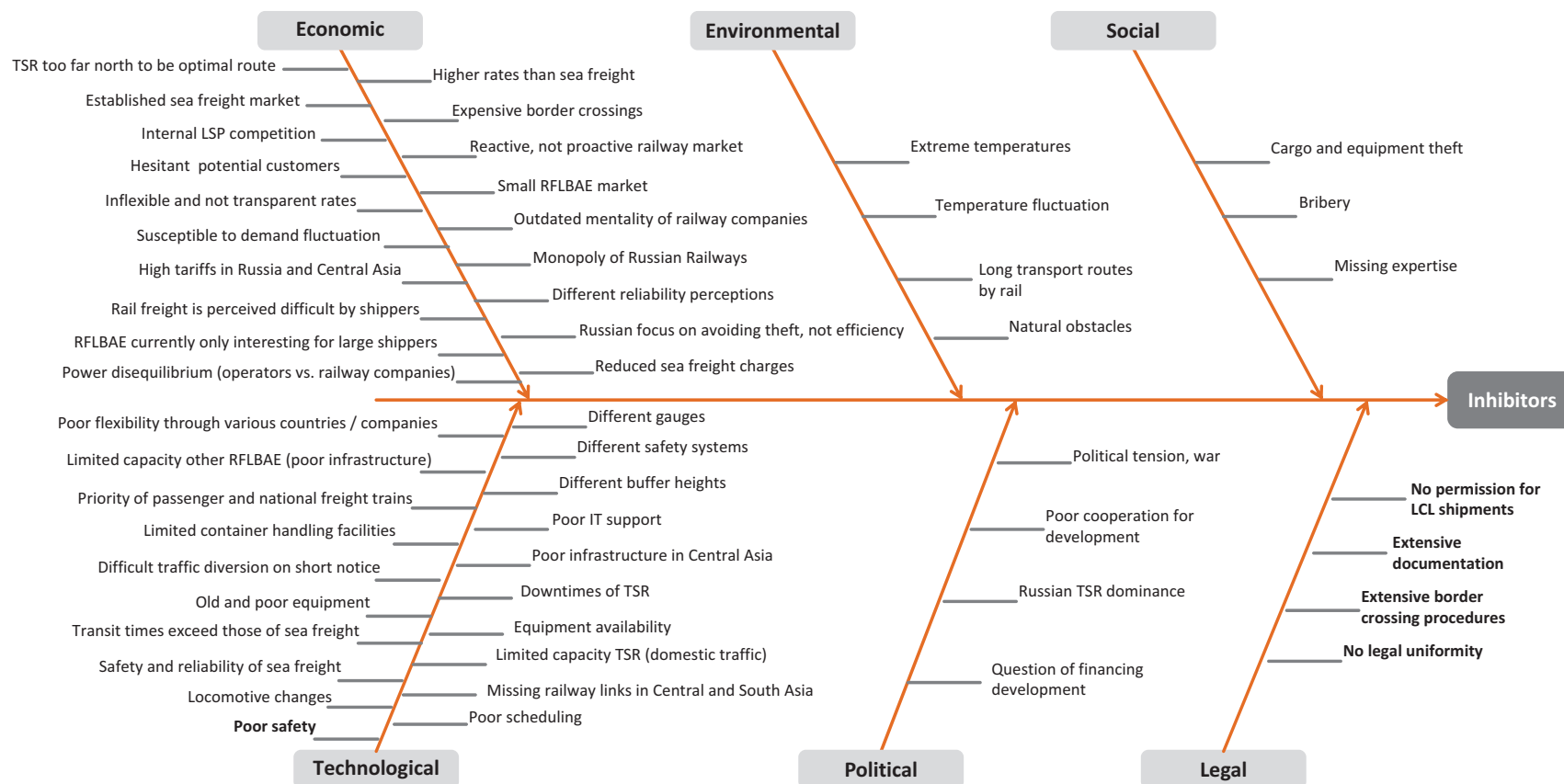


Figure 4 Ishikawa diagram: inhibitors. Source: Author.

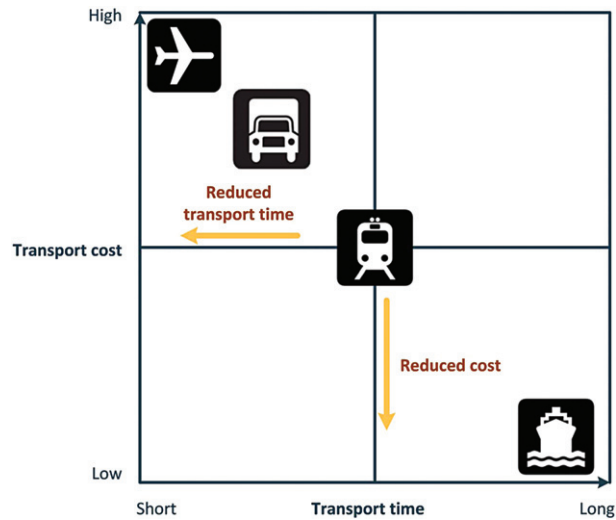


Figure 5 Transport modal choice characteristics. Source: Rodemann et al. (2018).

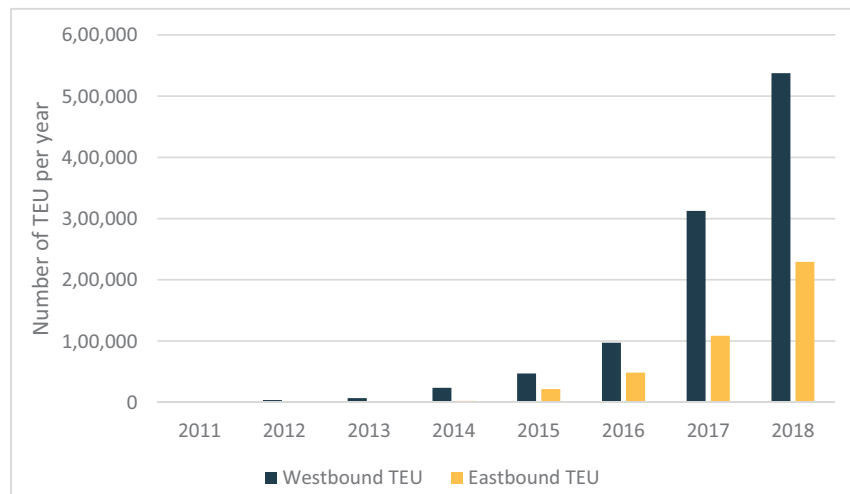


Figure 6 Number of TEU shipped across Eurasia by rail. Source: Adapted from Pomfret (2018), Xinhua Silk Road Information Service (2019), and Zhang (2017).

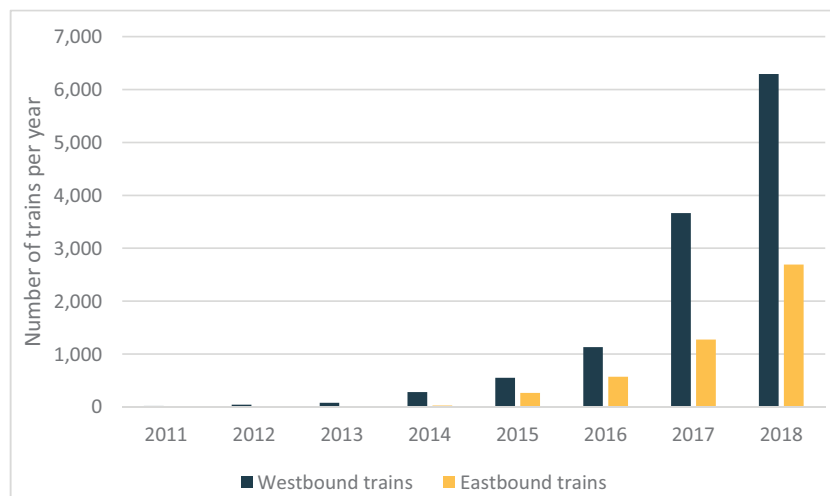


Figure 7 Number of container trains across Eurasia. Source: Adapted from Pomfret (2018), Xinhua Silk Road Information Service (2019), and Zhang (2017).

Table 1 Eurasia railroad offered services for the years 2013 and 2019

2013	2019
• Block trains • Multiple wagons • Multiple containers • Single wagons • Single container (full container load)	• Block trains • Multiple wagons • Multiple containers • Single wagons • Single container (full container load) • Less than container load (LCL)

Table 2 Eurasia railroad type of goods shipped east to west for 2013 and 2019

2013	2019
• Consumer electronics (laptops, tablets, computer screens)	• Consumer electronics (laptops, tablets, computer screens, smartphones, television sets, printers, electronic parts) • Machinery and equipment • Industrial goods (robots, high-tech, cooling fans) • Cars (to Central Asia), car parts, and spare parts • Consumer goods, e-commerce, retail, and merchandize (cosmetics, toothpaste, paper, ceramics, luggage, stationery, handicrafts, household appliances) • Garments/high-end fashion • (Industry) minerals • Flowers and trees

Table 3 Eurasia railroad type of goods shipped west to east for 2013 and 2019

2013	2019
• Car parts • Chemicals and hazardous material	• Cars, car parts, and spare parts • Chemicals and hazardous material • Food (cheese, chocolate, olive oil, pasta, wine, baby formula, spirits, beer, milk, meat) • Luxury goods, apparel, cosmetics, garments, and leather goods • Flowers • Pharmaceuticals • Machinery and equipment • Oil, coal, and gas (from Central Asia) • Grain and vegetables (from Central Asia) • Timber

For a detailed overview of current rail freight connections including route and service characteristics, as well as their start and end terminals, refer to [Appendix B](#). Please note that the indicated routes exclusively refer to direct connections. While Chinese terminals rather present immediate transshipment points to their respective hinterland, Europe has a more integrated intermodal rail freight network. In this, terminals that are connected with China via a trunk route collect or distribute containerized cargo via branch lines from start to end terminals (hub and spoke system). Needless to say, this amplifies the number of possible connections to a large extent.

Over recent years, there has been a considerable growth in cargo volumes ([Fig. 6](#)), train departures, and offered connections: from barely a scheduled service in 2011 to about 9000 trains in 2018 ([Fig. 7](#)). While these numbers are still small compared to sea freight, the market state cannot be considered as “infant” anymore, but rather as an established (although small) alternative to the incumbent modes of transport. The growth of volumes is expected to continue over the next years and hit 1 million TEU [Twenty Foot Equivalent Units] (equaling more than 10,000 trains) by 2020. Logically, the majority of trains currently run via the “established” routes, the TSR, or through Kazakhstan, and only a small share via the new connections through Turkey ([Fig. 8](#) and [Appendix B](#)). Shipments through Turkey, however, are likely to quickly gain popularity, especially for shipments between Western China and Southern Europe.

Despite the achieved and forecasted growth, Eurasian rail freight at large scale is still a new transport product. Hence, many shippers and logistics companies are inexperienced and thus insecure when it comes to aspects such as documentation, regulation, and, importantly, simply knowing “the right people” to facilitate and expedite solutions to operational issues. While anyone can send cargo across Eurasia, the picture changes when it comes to operating and steering trains. Whereas in many European countries

open market access applies to industries like logistics and rail transport, competition is impossible in the case of state-owned railways in China, Central Asia, and Russia. Moreover, in China, logistics companies and freight operators that manage international rail freight (block train companies) are required to possess special licensing.

As opposed to the logistical simplicity of maritime transport, running trains across Eurasia comes with an immense increase in complexity that affects shippers and logistics companies at strategic, tactical, and operational level. Strategic and tactical considerations range from understanding demographics via redesigning supply chains (e.g., production and storage facilities) to pursuing the adaptation or even construction of infrastructure and, most importantly, developing a strong pioneer mentality. At operational level, it is mainly technical, geographical, and legal aspects that are key to planning and decision-making.

Reliable information on available services and their departures have improved considerably. This can be seen in the overall market growth, as well as in the increase of involved logistics companies, forwarders, and freight operators that compete for cargo and hence pursue marketing activities and provide solid information. Incoterms are usually like those of sea and airfreight (e.g., EXW [ExWorks], FOB [Free On Board], or DDP [Delivered Duty Paid]). Due to the rising demand, certain services require reservations about 5 days before the train departure.

Environmental and Political Framework Conditions

Bridging Eurasia by rail means exposing both equipment and cargo to a range of challenging circumstances such as:

- Extreme temperatures (ranging from 40 °C in Chinese and Mongolian deserts to −40 °C in Siberian Tundra and Taiga during winter);
- Strongly fluctuating temperatures and related moisture (at yearly, daily, and also regional level);
- Geographical obstacles such as deserts, mountain ranges, earthquake-prone regions, forests, swamps, as well as rivers, lakes, and straits; and
- Unstable political situations.

While the first two put a high demand on running trains (e.g., regarding punctuality) and especially protecting cargo (e.g., food, cosmetics, or electronics), the latter two impact the construction and upkeep of infrastructure, as well as the trust of shippers.

Tracking and tracing of cargo is generally possible, so that shippers can follow their cargo online. With regard to security, containers can be equipped with electronic seals or light sensors to indicate when (and where) a container has been opened. This has immensely reduced the need for employing armed guards. Referring to tracking and tracing as well as equipment for cargo security, it must, however, be considered that many devices work similar to a SIM card in a mobile phone and hence do not necessarily work in each country passed. Integrated tracking and security systems are still to be introduced.

Train Length

As the maximum possible train length differs per country, there is no overall train length for the Eurasian land bridge. As a rule of thumb, however, a train is made up of about 43 railcars (that hold 86 TEU) reflecting the maximum lengths in China and Western Europe (750 m). Countries like Russia that allow much longer trains (up to 1 km) often combine two trains to cross the country and only separate them at the outbound border.

Border Crossings and Gauge Changes

Referring to the previous paragraph on missing infrastructural integration, running trains across Eurasia rather equals an “Olympic torch relay” (Hillman, 2017), where coordination and collaboration are essential to achieve a seamless outcome. Thus, interoperability between the individual network sections is currently ensured by changing equipment like locomotives and railcars at certain points along the route. For example, gauge changes typically take place at (or close to) national borders. While it seems to be best practice to transship liquid or dry bulk cargo with pipe systems, containers are moved by cranes between railcars of different gauge that are lined up side by side. Parallel to the transshipment of cargo that takes at best an hour per container train, the authorities usually manage import and export activities including documentation, and train operators perform quality checks on the rolling stock. Even though the individual changes of equipment only account for a relatively small share of the total transport time, they present a considerable risk of delays (e.g., in case of unavailable equipment for the upcoming section) and consequently require a high degree of asset planning and chain synchronization. In addition, regular waiting times (of many trains) at a certain point along the route are an attractive spot for vandalism or theft. While cargo within a container seems to be generally secured, diesel generators for refrigerated containers are easy targets, which in turn exacerbate the transport of temperature sensitive cargo.

Line Speed Versus Travel Speed

Considering the distances and transit times depicted in Appendix B, one can easily derive the average line speed per connection (28.3 km/h for the example Zhengzhou–Hamburg). Travel speeds are usually higher, ranging from 80 km/h up to 90 km/h or even 120 km/h on modernized sections. What causes relatively low average speeds is considerable waiting times, either within the network of a certain country or at borders between countries. The former is about priorities given to passenger or national freight

trains, extremely dense networks (e.g., in Western Europe, Moscow), or infrastructure sections that need improvement (e.g., in Mongolia, Uzbekistan, Azerbaijan, or Iran) resulting in speeds being down to 250 km per day (as opposed to speeds of more than 1000 km per day in Russia). The latter refers to procedures around borders that are still extremely inefficient. In this, it is not necessarily the pure transshipment of containers, but rather upstream and downstream processes that amplify the borders' characteristics as bottlenecks in the chain. Not only do these circumstances slow down planned travel times, they also present a major reason for delays that can easily amount to 2–6 days. As the previously outlined procedures seem to be alike across the different borders, it is difficult to identify "a route with best practice" in terms of interoperability. What rather impacts the quality of border crossings is the number of trains to be handled. In this, serving few trains results in smooth processes while many trains cause bottlenecks and delays. Hence, referring to Appendix B it becomes obvious that the border crossing Terespol–Brest at the Polish–Belarusian border is one of the weakest links of the entire network and shippers already seek to find alternatives. Other border crossings, however (e.g., between China and Kazakhstan respectively with Russia), already show similar tendencies, which is due to the overall increase in freight volume. Fig. 5 clearly shows that unpunctuality negatively impacts lead time and is thus a major threat for the competitiveness of Eurasian rail freight.

Policies and Legislation

Obviously, rail freight across Eurasia remains complex, which is due to the number of countries crossed, and thus different political, legislative, and organizational conditions are to be considered. In particular, China is to be regarded as complicated as each province has its own regulations that may exacerbate processes, especially with "nonstandardized" cargo such as dangerous goods. Nevertheless, several former pitfalls, such as slow and nonstandardized customs clearance, arbitrary procedures, as well as extensive documentation (and hence many opportunities for bribery) due to missing legal uniformity have been eased considerably by the creation of customs unions such as the EU or the Eurasian Economic Union (EAEU) that was founded by Belarus, Kazakhstan, and Russia in 2011 and extended to Armenia and Kyrgyzstan in 2014. While rail freight still requires more documentation compared to sea freight, procedures have been simplified and standardized (e.g., customs surveillance code 9610, the EAEU customs code in 2018, or the definition of general terms and conditions by the international rail transport committee in 2019) so that shippers and logistics companies do not have to arrange documentation with each country crossed any more. Together with the buildup of a solid base of experience, this has reduced formalities around borders to about 4 h. Hence, customs clearance and documentation with border crossings do not present a serious barrier to intercontinental rail freight any more. A further legal factor that has considerably contributed to growing volumes, broadening the base of potential (medium to small) shippers and realizing more stable demand patterns, is the possibility to send single containers or even LCL shipments, as opposed to a dedicated block train as in the past.

Cost and Subsidies

Obviously, a major modal choice criterion is cost. Cost for running trains (across Eurasia) among others comprises elements such as:

- Equipment (wagon and container rent);
- Infrastructure (access charges, traction, energy);
- Processes and procedures (customs, transshipment, security guards, border crossings, formerly bribery, change of rolling stock to due gauge change);
- Transport chain (pre- and post-haulage, terminal handling fees, storage fees at terminals, repositioning of empty wagons and containers); and
- Capital cost related to the shipped goods.

In 2013, transport rates were described as high, nontransparent and characterized by a high dependency on Russia and Central Asia. Comparing rates between 2013 and today show that for sea and airfreight there has been little change, while rail freight rates have been reduced by about one-third (Table 4).

Table 5 provides an example for current LCL freight rates. As LCL shipments (refer to Table 1) have only become possible in recent years, a comparison is obviously impossible.

Obviously, as rates are highly dependent on the origin and destination of the shipment, as well as on the characteristics of the route, the earlier values should serve as an indication of price ranges in comparison to other modes of transport. Both, lower and

Table 4 Freight rates 2013 and 2019

<i>Charge rates per container</i>	<i>2013</i>	<i>2019</i>
Sea freight (east-west)	\$2,000–\$5,200	\$2,000–\$6,200
Rail freight (east-west)	\$5,500–\$6,600	\$3,500–\$4,500
Airfreight (east-west)	\$18,000–\$24,000	\$18,000–\$28,000

Source: Adapted from Vinokurov et al. (2018).

Table 5 Freight rates 2013 and 2019

LCL freight rate	2019
Sea freight (east-west)	\$40 per m ³ or 1,000 kg
Rail freight (east-west)	\$115 per m ³ or 500 kg

Source: Adapted from Desormeaux (2019).

higher (rail freight) rates are possible. What is important for shippers in practice is a comparison of rates at individual route level and under consideration of needs and demands in the supply chain.

When comparing rates, it is moreover important to consider that rail freight across Eurasia is still heavily subsidized. The real cost for sending a container across Eurasia by rail is estimated at about \$9,000–\$10,000 (Wallis, 2018). Referring to the rates as indicated in Table 4, this results in subsidies of about \$5000 per container (Liang, 2018). Hence, in order to maintain current prices in the future, significant efficiency improvements are necessary after the ceasing of subsidies.

While the determination of rail freight rates is generally complex and hence prone to nontransparency (including charges for access, capacity, traction energy [electricity or diesel], as well as stations or depots), it is considerably exacerbated when it comes to the numerous activities related to rail transport across Eurasia. Research (Jakobowski et al., 2018) roughly divides revenues for terminal-to-terminal container transport between forwarding company (5%), intermodal operator (5%), and carrier as well as infrastructure provider (90%). However, a detailed distribution of costs incurred along the route (e.g., using a pie chart) could not be found in any of the literature considered and seems (currently) impossible. For that reason, it is unlikely that the transparency of freight rates has improved over the last years or that it will improve any time soon. At least with regard to the publishing of rates, as well as the debating of subsidies, in the literature, transparency can be considered as slightly improved.

Political and Macroeconomic Conditions

The dependency on Russia and Central Asia persists (refer to Fig. 8 and Appendix A). However, the use of various routes and construction of new connections make shippers less dependent on individual (less favorable) sections.

A macroeconomic phenomenon that characterizes trade between Asia and Europe is the imbalance of trade and hence cargo volumes. As is apparent in Figs. 6 and 7, there is a considerable imbalance of trains and shipped TEU in favor of east-west flows (more than two-thirds) compared to west-east flows (less than one-third). Certain references go a step beyond and also consider the loading degree of trains (and more importantly of containers on trains) showing an even more extreme imbalance, with loading

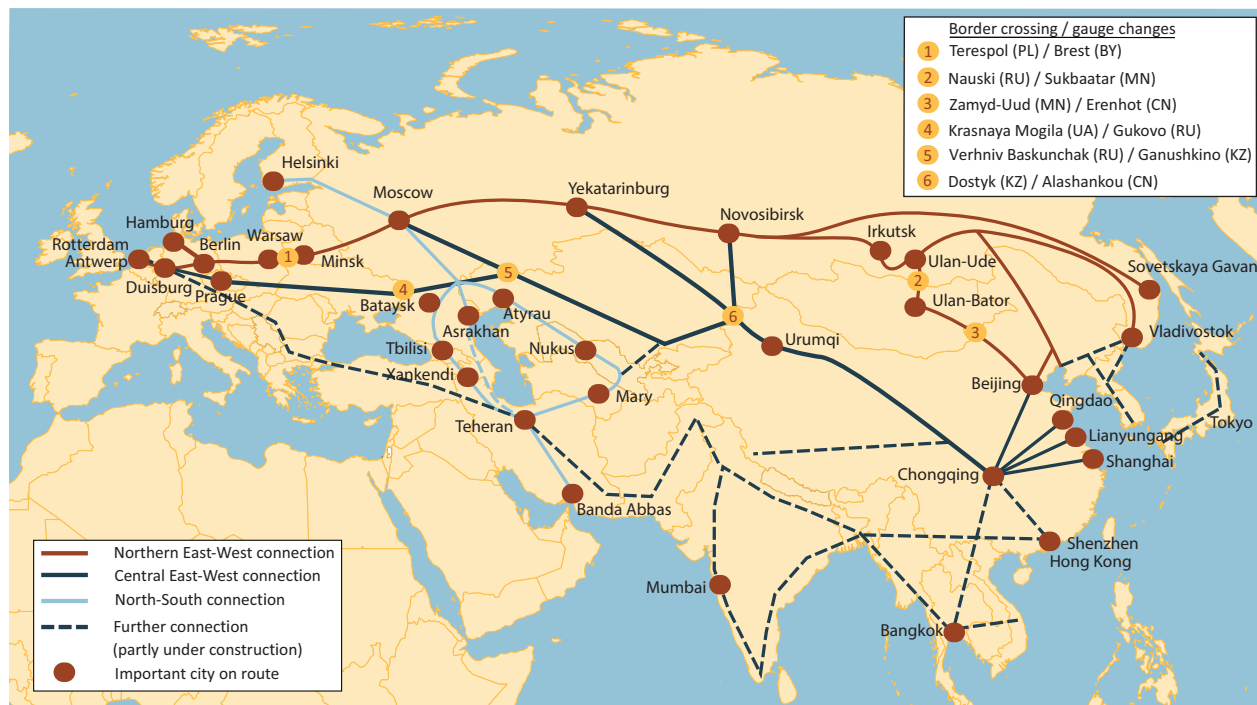


Figure 8 Eurasian railway network. Source: Rodemann et al. (2018).

factors of westbound trains being about 90% (new Chinese policies do not allow shipping empty containers to Europe by rail) compared to 20% for eastbound trains (it does happen that an empty pallet is put into a container to change its status to “loaded” [Szakonyi, 2018]). This represents a huge financial burden with regard to freight rates as well as cost for repositioning of empty containers and rolling stock, especially in the light of intercontinental rail freight being a relatively young product seeking competitiveness without subsidies.

Summary

The earlier description of the current state of Eurasian rail freight is now summarized with regard to the impact on transport cost and time in order to assess whether the former strategic niche of intercontinental rail freight is expanding or not. This is done with the help of the previously introduced hypotheses (Table 6). Subsequently, the mitigation strategies (Table 7) and inhibitors (Table 8) identified in 2013 are reviewed with regard to their continued presence or status in 2019, in order to answer the two key questions of this paper. Finally, a set of newly emerged enablers is presented (Table 9).

As mentioned earlier, the 2013 research collected possible strategies to mitigate inhibitors (and turn them into strengths). Table 7 lists these strategies and indicates the individual extent of realization in 2019. The order of the listed mitigation strategies reflects the frequency of findings in 2013 (starting with the most frequent strategy and descending subsequently).

Corresponding to the realization of mitigation strategies presented in Table 7, Table 8 shows the applicability (in 2019) of those inhibitors identified in 2013.

Based on the previous paragraph, Table 9 lists new enablers and inhibitors that have emerged over the recent years and will influence the future development and growth of the land bridge. To be in line with the 2013 research, Table 9 categorizes the enablers and inhibitors based on the PESTLE framework.

From the earlier, one can draw the conclusion that the strategic market niche of Eurasian rail freight is extending. In a virtuous circle, the increase of services has enhanced the number of shippers, which, in turn, has caused further connections to emerge. This

Table 6 Time and cost analysis

<i>Hypothesis</i>	<i>Applies to a</i>	<i>Reason</i>	<i>Impact on transport cost</i>	<i>Impact on transport time</i>	<i>Future development potential</i>
An increased number of routes and terminals enhance the geographical catchment area of rail freight and thus reduce transport distance to and from terminals.	High extent	• More terminals are linked to the land bridge which results in more routes • The network is extended which allows more route options	• Shorter total transport distance reduces cost • Shorter pre- and post-haulages reduce cost	• More routes link more cities which reduces the total transport distance for individual shippers/consignees to and from the closest terminal	• Additional terminals and routes • Further extension and integration of the railway network
An increased number of departures reduce waiting time for individual shipments.	High extent	• Especially highly frequented routes allow daily departures and hence reduce lead times for individual shipment	• No direct impact on transport cost • Less waiting time, however, reduces the capital tied-up	• No direct impact on transport time • Higher departure frequencies reduce total (average) lead time for individual shipments	• Further increase of departure frequency
An increased number of departures reduce cost (e.g., overhead or transshipments).	Low extent	• This is generally possible • Current rates, however, are to a huge extent achieved through subsidies	• Reduction of indirect cost per shipment • Reduction in capital tied-up	• See earlier line	• Volumes must considerably increase in order to generate sufficient economies of scale • Subsidies are currently more decisive than cost
An increased number of routes and departures enhance the importance of intercontinental rail freight for the Eurasian economy and may hence contribute to smoothing processes in order to reduce both cost and time.	High extent	• Border processes have been smoothed • Documentation has been eased • LCL shipments have been made possible	• Less administrative cost • Less capital tied-up	• Less waiting times • Less delays	• Further smoothing of processes • Standardization and harmonization of regulations • Eased market access

Table 7 Possible strategies to mitigate inhibitors

<i>Strategy</i>	<i>Main influencer</i>	<i>Extent realization until 2019</i>	<i>Evidence</i>
Joint infrastructure development (track modernization and new lines)	G	High	Modernization of infrastructure in Central Asian countries (e.g., Kazakhstan)
Simplification of border procedures/single-way bill	G	High	New lines from China to Turkey, Pakistan, and Singapore
IT infrastructure for real-time information (tracking and tracing/GPS use)	L&T	Medium	Customs/border crossing procedures have been eased (electronic processes)
Marketing	L&T	High	Demanded and already in use, but functionality along the entire route to be improved
Enhancing value propositions	L&T	Low	Major logistics companies offer Eurasian rail freight
Introduction of competitive rates	RC	High	BRI (Belt and Road Initiative) as a tag
Information sharing	AP	Medium	No evidence found
Extension of wide gauge network	G	Low	Subsidies have considerably reduced rates
New T (container handling and gauge changes)	G	Medium	Availability of information on services is improving. Still little information sharing between individual parties and still no transparency regarding cost and rates
Improving traffic management	RC	Low	No evidence found
Using equipment with adjustable bogies	RC	Low	Route to Vienna not yet finished
Discussion with authorities (improving/reducing customs procedures)	L&T	High	Several new services have been started and built/linked new terminal to the network
Tariff harmonization	G	Medium	No evidence found
System integration and collaboration to achieve standardized interoperability (technical, commercial, administrative)	G	High	Still considerable bottlenecks
Introduction of heatable containers and insulating packaging	S	High	No evidence found
China as major money supplier	G	High	Customs/border crossing procedures have been eased (electronic processes)
Speeding up loading and gauge changes	T	Medium	Rates have been reduced through subsidies
Improving economic relations between Russia and China	G	Medium	Subsidies still vary considerably between Chinese provinces
Creating/enlarging customs unions	G	High	Customs/border crossing procedures have been eased (electronic processes)
Running test trains	S	High	Refer containers are commonly used
Normal gauge line across Central Asia	G	High	China is to a great extent financing national and international infrastructure projects
Electrification and double tracking	G	High	Transshipment process lasts about an hour per train
Use of passenger locomotives	RC	Low	Up- and downstream processes still present bottlenecks
Rate standardization/single through rates	RC	Low	Increased number of services
Increasing global container trade	S	Medium	Still low number compared to China–Europe
Transit in 7 days program	RC	Medium	European Union (EU) and Eurasian Economic Union (EAEU)
China–Europe–Northeast America program	RC	Low	New routes and services are likely to use test trains (e.g., to Madrid or London)
Using dedicated trains	S	High	Normal gauge route built between China and Turkey
Minimizing empty wagons/containers	L&T	Low	Many tracks in Central Asia (e.g., Kazakhstan) have been modernized
Joint ventures	RC	Low	No evidence found
New/modernization of locomotives and wagons	RC	Medium	Rates are highly dependent on subsidies of Chinese provinces and hence differ a lot
Hybrid propulsion	RC	Low	Considerable growth of Eurasian rail freight
Speeding up trains	RC	Medium	Global trade is reaching saturation phase
Separation of freight and passenger trains	RC	Low	China's economic growth is slowing down
Reducing amount of network operators	G	Low	Trains in Russia achieve speeds of +1,000 km per day
Developing industrial production along RFLBAE	G	High	Bottlenecks at borders remain or increase due to increased traffic
			No evidence found
			Several Chinese provinces use dedicated trains to Russia and Europe
			There is still many empty containers / containers “loaded” with one pallet
			The service is still provided by individual train companies
			In China there is little collaboration between competing provinces
			Central Asian countries (e.g., Kazakhstan) have modernized their rolling stock
			No evidence found
			Modernization of certain sections allows higher speeds
			Lead time is still low due to bottlenecks and increased traffic
			No evidence found
			The service is still provided by individual train companies
			Chinese (mega) cities such as Yiwu, Wuhan, Chengdu, and Chongqing
			Logistics parks in Kazakhstan, Belarus, and Ukraine

AP, all parties; G, government; L&T, logistics and transport; RC, railway companies; S, suppliers; T, terminals.

Table 8 Eurasia inhibitors identified by the 2013 Cranfield study

<i>Pestle category</i>	<i>2013</i>	<i>2019</i>
Political	Political tension and war Poor cooperation for development	Political instability in Eastern Europe, Central Asia, Iran, and Turkey remains Despite heavy involvement of China and also Belgium and Germany, there is still little cooperation between countries
Economical	Russian TSR dominance Question of financing development Higher rates than sea freight Monopoly of Russian Railways Reactive, not proactive railway market Small RFLBAE (rail freight land bridge between Asia and Europe) market Outdated mentality of railway companies Russian focus on avoiding theft, not efficiency	Reduced due to alternative routes High involvement of China Subsidies have reduced the difference between rail and sea freight rates Reduced due to alternative routes No evidence for change Considerable growth in terms of cargo types, volume, trains, and connections
	Power disequilibrium (operators vs. railway companies) Hesitant potential customers RFLBAE currently only interesting for large shippers	No evidence for change Growth of market has increased the number of shippers Introduction of single wagon and LCL shipments also makes rail freight interesting for smaller shippers
	Susceptible to demand fluctuation (credit crunch) Rail freight is perceived difficult by shippers	Volume growth has made market less susceptible to demand fluctuation Rail freight is still complex, but an increase in market size and players also increases the experience of participants (including shippers)
Technological	Internal LSP (logistics service provider) competition Different reliability perceptions TSR too far North to be optimal route Establish sea freight market Inflexible and not transparent rates High tariffs in Russia and Central Asia Reduced sea freight charges Expensive border crossings	No evidence found (in theory it is still existing) No evidence for change No change; however, there is alternative routes No evidence for change No evidence for change No evidence for change No evidence for change Procedures have been standardized and simplified
	Different gauges Different safety systems Limited capacity on other RFLBAE (poor infrastructure)	No evidence for change No evidence for change The establishment of new infrastructure has increased the capacity; the increase in volumes, however, also requires high capacity
	Limited capacity on TSR (domestic traffic)	No evidence for change; a huge share of trains takes the TSR (at least for a certain part of the trip)
	Different buffer heights Priority of passenger and national freight trains Poor infrastructure in Central Asia	No evidence for change No evidence for change Investments in infrastructure have improved the network (especially in Central Asia)
	Poor IT support	IT support has improved; a comprehensive tracking and tracing along the entire route remains difficult
	Poor flexibility through various companies and countries involved Difficult traffic diversion on short notice Limited container handling facilities Old and poor equipment Transit times exceed those of sea freight	No evidence for change New routes are expected to increase the flexibility for shippers Heavy investments in infrastructure have increased the number of terminals Heavy investments have modernized or renewed equipment Rail freight is generally faster than sea freight; there can significant delays with rail freight
	Safety and reliability of sea freight Locomotive changes	No evidence for change No evidence for change; the uniform electrification of several sections, however, reduces the number of necessary locomotive changes
	Missing railway links in Central and South Asia	New links and connections between existing infrastructure have been established
	Poor scheduling Poor safety	Scheduling of border crossings and in dense networks remains a weak point There is less evidence and less armed guards; border crossings remain dangerous especially in terms of diesel and generator theft
	Downtimes of TSR Equipment availability	No evidence for change Sufficient equipment is available; short-term unavailability of equipment exists with transfer points (e.g., locomotive changes and border crossings)

(Continued)

Table 8 Eurasia inhibitors identified by the 2013 Cranfield study (*cont.*)

<i>Pestle category</i>	<i>2013</i>	<i>2019</i>
Social	Cargo and equipment theft	There is less evidence and less armed guards; border crossings remain dangerous especially in terms of diesel and generator theft
	Bribery	The standardization of procedures has reduced the possibility of bribery; it, however, still exists
	Missing expertise	The growth of market has contributed to the increase of expertise with various parties
Legal	No permission for LCL shipments	LCL shipments are now possible
	Extensive documentation	Documentation has been standardized and simplified
	Extensive border crossing procedures	Procedures have been standardized and simplified
	No legal uniformity	Still no complete legal uniformity, but EU and EAEU have standardized regulations within their territories
Environmental	Extreme temperatures	No evidence for change; the use of reefer containers is a possible mitigation strategy
	Long transport routes by rail	No evidence for change
	Natural obstacles	No evidence for change
	Temperature fluctuation	No evidence for change; the use of reefer containers is a possible mitigation strategy

Table 9 New enablers and inhibitors to Eurasian rail freight in 2019

<i>PESTLE category</i>	<i>2019 enablers</i>	<i>2019 inhibitors</i>
Political	- Subsidies by Chinese government - More and more countries are involved in BRI and are interested in investing	No new political inhibitors identified
Economical	Enhanced flexibility in the supply chain through competition between newly established routes	- Cannibalism through competition between terminals - Cannibalism through competition between routes
Social	- New growth potential, for example, through increased e-commerce - No new social enablers identified	- Imbalance of trade volume (east-west vs. west-east) - No new social inhibitors identified
Technological	- Improved tracking and tracing - Block chain and smart contracts	No new technological inhibitors identified
Legal	Implementation of new customs procedures	No new legal inhibitors identified
Environmental	Increased environmental awareness and low emissions of rail freight compared to airfreight	No new environmental inhibitors identified

growth of the market is mutually related to significant enhancements in terms of processes, procedures, and legal framework conditions (including subsidies for lower rates) that created more attractive circumstances for shippers and contributed to the earlier outlined reduction of transport cost and time.

Future Outlook

Having evaluated the current state of Eurasian rail freight, the focus is now on the future of the Eurasian land bridge between China and Europe, focusing on the original inhibitors that still need to be addressed (**Table 8**), together with recent enablers and inhibitors that will have a positive impact on the future development of the land bridge (**Table 9**). Finally, this paper focuses on the current plans that are being implemented and the future opportunities that are being considered that will influence the development of the Eurasian land bridge.

The future development and growth of the Eurasia railroad will be dependent on tackling the following issues:

- The need to build a financially sustainable railroad;
- Managing the future expectations associated with westward migration;
- Mitigating risk and achieving resilience;
- Solving the trade imbalance between freight moving from China to Europe compared to that from Europe to China;
- Managing internal demand for freight terminals;
- Future economic growth; and
- Extending the geographic reach of the railroad.

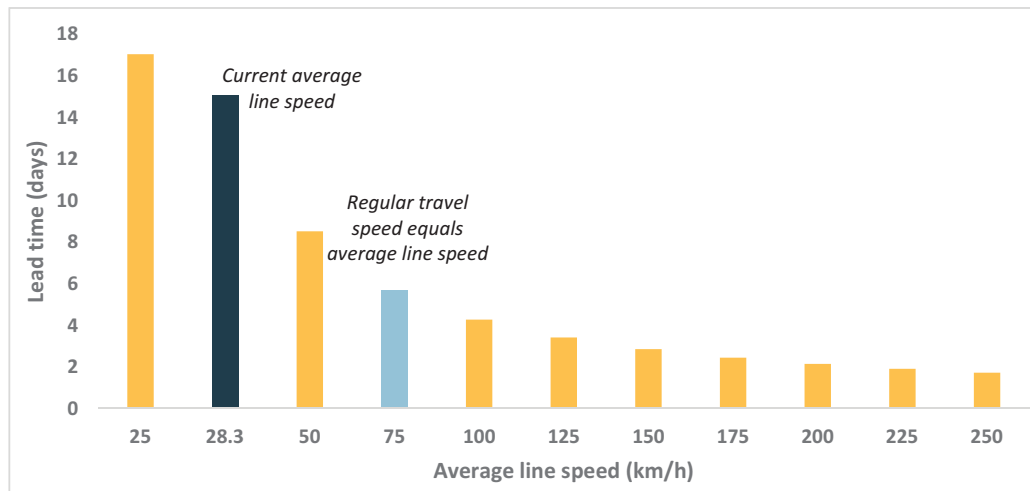


Figure 9 Scenario analysis. *Source: Author.*

The recent success of the railroad has been generated by the Chinese state investment in new infrastructure and subsidized freight rates. There are no guarantees that this state support will continue into the future, and the warning signs are already being signaled by the Chinese state that freight subsidies will be reduced and possibly phased out entirely in the near future. China wants to see additional investment in the railroad from their rail partners along the various routes and is not prepared to fund the project alone and let other states have a free ride on the back of Chinese investment.

Over recent years there has been a significant movement of people and businesses migrating westward away from the Chinese coastal cities inland to the central regions. This has resulted in greater demand for connectivity to facilitate the movement of people, resources, and goods for exporting to international markets. Migration has been one of the contributing factors that has stimulated the demand for, and interest in, improved and new investment in Eurasia rail links.

Traditionally, high value goods such as electronics, computers, and mobile phones have been exported to overseas markets from China by air. However, as manufacturing costs have decreased, airfreight costs are now becoming a greater percentage of total costs. Alternative transport modes such as rail freight are now being considered as an alternative to airfreight and the Eurasia railroad is now well placed to take advantage of this modal shift. This is because transit times from China to Europe are between 2 and 3 weeks, which is half the time of sea freight, and are also considerably cost-effective when compared to airfreight charges. A further reduction of rail freight lead time would obviously improve the competitive position of the Eurasia railroad compared to airfreight. **Fig. 9** presents a lead-time scenario analysis at different average line speeds using the previous example Zhengzhou–Hamburg. What is striking is that an increase of the average line speed to the (approximate) regular travel speed of freight trains would reduce the current lead time by about 60%. This could be achieved by a combination of advanced scheduling and comparably small improvements of certain infrastructure sections to remove current bottlenecks. The potential for further lead-time reduction exists with high-speed rail freight. However, this would require the extremely costly upgrade of practically the entire railway network and, moreover, the development and acquisition of new rolling stock.

The enhancement of the Eurasia railroad links also provides China with an alternative conduit for international trade, thus reducing strategic risk and enhancing resilience, just in case their traditional sea-lanes are disrupted in the future.

For the railroad to be financially sustainable, the trade imbalance needs to be addressed, as freight movements from China to Europe considerably outweigh the reverse flow. Empty running/low utilization is a significant issue for freight transport, whichever transport mode your business operates.

The prolific growth in the number of Chinese cities wishing to participate in the phenomenon of the Chinese railroad renaissance is clear to see by the increase in new routes. However, this growth has been supported and encouraged by government incentives and policy. Also, the international prestige that the rail links provide to a city and its hinterland is also a factor. The big issue that needs to be considered for the future development of these railheads is whether there is sufficient demand to support all of them. If not, the future could see existing cities and new entrants moving into the marketplace and cutting freight charges to attract customers, resulting in a “beggar my neighbor” culture, which may well benefit shippers in the short term, but could have a detrimental impact on the mid- to long-term future development of the railroad.

Economic growth will be a considerable factor that will determine the growth in demand for rail freight services along the Eurasia rail network, not only the demand for Chinese exports to the west, but also the demand for western imports into China. China’s economic growth is also essential in order to continue the capital investment in the BRI infrastructure projects if the Eurasia railroad is to achieve its aim and objectives. The impact of future trade deals will also influence the level of goods shipped via the Eurasia railroad.

Extending the geographical reach of the Eurasia railroad to link islands such as Japan, Taiwan, and the Philippines (by vessel to China), as well as mainland countries in Southeast Asia and the Korean peninsula, provides additional services for shippers. Furthermore, it enables railheads such as Chongqing to operate as a hub-and-spoke node or an aggregator. However, it requires additional investment in railway infrastructure to achieve the enhanced network.

To stabilize and further improve its position, Eurasian rail freight literally still has a long way to go. Having become more economically interesting, it is now important to provide sufficient infrastructural capacity, quality, and transport chain resilience to remove bottlenecks and to guarantee smooth and reliable operations. Moreover, in the medium to long run, further cost reductions will be inevitable in order to become independent of subsidies and ensure financial stability. Eventually, however, the decisive factors will be macroeconomic conditions that generate future growth, as well as balanced cargo flows between Asia and Europe and an appropriate use of terminals to realize (but not jeopardize) a comprehensive catchment area of the railroad across Eurasia.

Acknowledgment

We would like to sincerely thank Cranfield University, the sponsor of the initial research, for its financial support and the contribution of contacts. In addition, we owe special thanks to all interviewees for the engaging conversations and the provision of valuable information. We are extremely grateful to Fred Wan for his informative support, the validation of information, and most importantly his enthusiasm on the research. Last but not least, our gratitude goes out to Frans de Jong for his support with the map.

References

- Berger, R., 2017. Eurasian Rail Corridors: What Opportunities for Freight Stakeholders. International Union of Railways (UIC), Paris.
- Desormeaux, H., 2019. Overland gains traction on Asia-European trade. Available from: <https://www.americanshipper.com/news/overland-gains-traction-on-asia-europe-trade?autonumber=73289&origin=relatedarticles>.
- Hillman, J.E., 2017. Trains from China laden with hype and subsidies. Available from: <https://www.ft.com/content/dd6196f8-715e-11e7-aca6-c6bd07df1a3c>.
- Jakobowski, J., Poplawski, K., Kaczmarek, M., 2018. The Silk Railroad. Centre for Eastern Studies, Warsaw.
- Liang, L.Y., 2018. World focus: China on track for rail boom. Available from: <https://www.straitstimes.com/asia/china-on-track-for-rail-boom>.
- Lobyrev, V., Tikhomirov, A., Tsukarev, T., Vinokurov, E., 2018. Belt and road transport corridors: barriers and investments. Report No. 50. Eurasian Development Bank, Saint Petersburg.
- Pomfret, R., 2018. The Eurasian landbridge: linking regional value chains. Available from: <https://voxeu.org/article/eurasian-landbridge-linking-regional-value-chains>.
- Rodemann, H., de Jong, F., Ruijgrok, K., Verweij, K., 2018. The Vitality of Intermodal Transport. InterRoJo Publications, Barendrecht, The Netherlands.
- Szakonyi, M., 2018. China's Asia-Europe rail subsidies—a two edged sword. Available from: https://www.joc.com/rail-intermodal/international-rail/china/asia-europe-rail-subsidies-come-hidden-costs-network_20181004.html.
- Vinokurov, E., Lobyrev, V., Tikhomirov, A., Tsukarev, T., 2018. Silk road transport corridors: assessment of trans-EAEU freight traffic growth potential. Report 49. Eurasian Development Bank, Saint Petersburg.
- Wallis, K., 2018. Maersk aims to double its Asia-Europe rail volume. Available from: https://www.joc.com/rail-intermodal/international-rail/maersk-aims-double-its-asia-europe-rail-volume_20181120.html.
- Xinhua Silk Road Information Service, 2019. China-Europe freight trains see robust growth in trips 2018. Available from: <http://en.silkroad.news.cn/2019/0125/127398.shtml>.
- Zhang, X., 2017. Eurasian Rail Freight in the One Belt One Road Era, MSc thesis, Cranfield University.

Further Reading

- Mirchandani, V., 2012. HP—The Quest for a 10 Out of 10 Supply Chain. Courier Westford, Westford, MA.
- Rodemann, H., 2013. An Investigation into the Enablers and Inhibitors of Intermodal Rail Freight Land Bridges between Asia and Europe, MSc thesis, Cranfield University.
- Rodemann, H., Templar, S., 2014. The enablers and inhibitors of intermodal rail freight between Asia and Europe. J. Rail Transp. Plan. Manag. 4 (3), 70–86.
- Zhang, H., 2013. Eurasian Landbridge: Past, Present and Future, MSc thesis, Cranfield University.

Appendices

Appendix A Infrastructural characteristics of Eurasian railway network

<i>Direction</i>	<i>Origin—destination</i>	<i>Countries passed</i>	<i>Rail gauges (mm)</i>	<i>Infrastructure owner</i>	<i>Loading gauges (height in mm)</i>	<i>Loading gauges (width in mm)</i>	<i>Safety/signaling systems</i>	<i>Power source for traction</i>	<i>Degree of electrification</i>	<i>Degree of double tracks</i>
East-west	Russia/China to Western Europe	China (CN)	1,435	State/China Railways	4,800	3,400	CTCS-3/CBTC	25,000 V AC	68%	56%
		(Mongolia [MN])	1,520	State/Mongolyn Tömör	N/A	N/A	N/A	Diesel	Low	Low
		Russia (RU)	1,520	Osam	5,300	3,750	ALSN/KLUB-U	25,000 V AC ^a	40%	N/A ^f
		Belarus (BY)	1,520	Russian Railways	5,300	3,750	ALSN/KLUB-U	25,000 V AC	15%	N/A
		Poland (PL)	1,435	Belarusian Railways	4,650	3,150	SHP	3,000 V DC	65%	N/A
		Germany (DE)	1,435	State/Polish State Railways	4,650	3,150	INDUS/LZB	15,000 V AC	60%	N/A
East-west	Russia/China to Western Europe	China (CN)	1,435	State/China Railways	4,800	3,400	CTCS-3/CBTC	25,000 V AC	68%	56%
		Kazakhstan (KZ)	1,520	Qasagstan Temir	5,300	3,750	N/A	25,000 V AC ^c	26%	Low
		Russia (RU)	1,520	Scholy	5,300	3,750	KLUB-U	25,000 V AC ^a	40%	N/A ^f
		Belarus (BY)	1,520	Russian Railways	5,300	3,750	ALSN/KLUB-U	25,000 V AC	15%	N/A
		Ukraine (UA)	1,520	Belarusian Railways	5,300	3,750	ALSN/KLUB-U	25,000 V AC ^b	45%	N/A
		Poland (PL)	1,435	Ukrainian Railways	4,650	3,150	SHP	3,000 V DC	65%	N/A
		Germany (DE)	1,435	State/Polish State Railways	4,650	3,150	INDUS/LZB	15,000 V AC	60%	N/A
				DB Netze						
East-west	Russia/China to Western Europe	China (CN)	1,435	State/China Railways	4,800	3,400	CTCS-3/CBTC	25,000 V AC	68%	56%
		Kazakhstan (KZ)	1,520	Qasagstan Temir	5,300	3,750	N/A	25,000 V AC ^c	26%	Low
		Russia (RU)	1,520	Scholy	5,300	3,750	ALSN/KLUB-U	25,000 V AC ^a	40%	N/A ^f
		Georgia (GE)	1,520	Russian Railways	N/A	N/A	N/A	3,300 V DC	100%	N/A
		(Uzbekistan [UZ])	1,520	Georgian Railways	5,300	3,750	N/A	25,000 V AC	52%	N/A
		(Turkmenistan [TM])	1,520	Uzbekistan Railways	5,300	3,750	N/A	Diesel	Low	N/A
		(Azerbaijan [AZ])	1,520	Türkmen demir yollary	N/A	N/A	N/A	3,000 V DC	Low	Low
		(Iran [IR])	1,435	Azerbaijan Railways	N/A	N/A	N/A	25,000 V AC	20%	15%
		Turkey (TR)	1,435	Islamic Republic of Iran Railways	4,650	3,150	TDS	25,000 V AC		
				State						
North-south	Northern/Western Europe to Iran/Pakistan/India	Finland (FI)	1,524	Ratahallintokeskus	5,300	3,400	EBICAB 900	25,000 V AC	55%	8%
		Russia (RU)	1,520	Russian Railways	5,300	3,750	ALSN/KLUB-U	25,000 V AC ^a	40%	N/A ^f
		Georgia (GE)	1,520	Georgian Railways	N/A	N/A	N/A	3,300 V DC	100%	N/A
		Azerbaijan (AZ)	1,520	Azerbaijan Railways	N/A	N/A	N/A	3,000 V DC	43%	28%
		Iran (IR)	1,435	Islamic Republic of Iran Railways	N/A	N/A	N/A	25,000 V AC	Low	15%
		Pakistan (PK)	1,676	Pakistan Railways	7,000	3,250	N/A	25,000 V AC	Low	Low
		India (IN)	1,676	India Railways	7,000	3,250	AWS/TPWS	25,000 V AC	34%	33%

North-south	Russia/China to Pakistan/ India	Russia (RU)	1,520	Russian Railways	5,300	3,750	ALSN/KLUB-U	25,000 V AC ^a	40%	N/A ^f
		China (CN)	1,435	State/China Railways	4,800	3,400	CTCS-3/CBTC	25,000 V AC	68%	56%
		Pakistan (PK)	1,676	Pakistan Railways	7,000	3,250	N/A	25,000 V AC	Low	Low
		India (IN)	1,676	India Railways	7,000	3,250	AWS/TPWS	25,000 V AC	34%	33%
North-south	Russia/China to Iran	Russia (RU)	1,520	Russian Railways	5,300	3,750	ALSN/KLUB-U	25,000 V AC ^a	40%	N/A ^f
		China (CN)	1,435	State/China Railways	4,800	3,400	CTCS-3/CBTC	25,000 V AC	68%	56%
		(Kazakhstan [KZ])	1,520	Qasaqstan Temir	5,300	3,750	N/A	25,000 V AC ^c	26%	Low
		(Turkmenistan [TM])	1,520	Scholy	5,300	3,750	N/A	Diesel	Low	N/A
		Iran (IR)	1,435	Türkmen demirjollary	N/A	N/A	N/A	25,000 V AC	Low	15%
				Islamic Republic of Iran Railways						
North-south	Russia/China to Singapore	Russia (RU)	1,676	Russian Railways	5,300	3,750	ALSN/KLUB-U	25,000 V AC ^a	40%	N/A ^f
		China (CN)	1,435	State/China Railways	4,800	3,400	CTCS-3/CBTC	25,000 V AC	68%	56%
		Vietnam (VN)	1,435	Vietnam Railways	N/A	N/A	N/A	Diesel	Low	Low
		Cambodia (KH)	1,435	State/Toll Royal	N/A	N/A	N/A	Diesel	Low	Low
		Laos (LA)	1,000	Railways	N/A	N/A	N/A	Diesel	Low	Low
		Thailand (TH)	1,435	State	N/A	N/A	N/A	25,000 V AC	Low	Low
		Malaysia (MY)	1,435	State/State Railway of	N/A	N/A	N/A	25,000 V AC	42%	42%
		Singapore (SG)	1,435	Thailand State/Malaysian Railways State	N/A	N/A	N/A	1,500 V DC ^d	0% ^e	0% ^e

^aAlso 3,000 V DC; 6,000 V DC; 10,000 V AC.

^bAlso 3,000 V DC; 10,000 V AC.

^cAlso 10,000 V AC.

^dCurrently used for public transport—it is likely that a link to Malaysia will also be equipped with 25,000 V AC.

^eA link to Malaysia is planned.

^fThe Trans-Siberian railway is completely double tracked.

Connection	Distance (km)	Transit time (days)	Start	Frequency (departures per week)	Asian gauge change at	European gauge change at	Transit countries
Zhengzhou–Hamburg	10,214	15	2013	4	Alashankou/Dostyk Khorgos/Altynkol	Terespol/Brest ^a	KZ, RU, BY, PL
Zhengzhou–Wuhan–Hamburg	10,214	15	N/A	2	Zamyd-Uud/Erenhot	Terespol/Brest ^a	MN, RU, BY, PL
Chongqing–Duisburg	11,179	15	2011	18	Alashankou/Dostyk Khorgos/Altynkol Zamyd-Uud/Erenhot	Terespol/Brest ^a	MN/KZ, RU, BY, PL
Chongqing–Cherkessk	N/A	10	N/A	2	Zabaikalsk/Manzhouli ^b	None	None
Chengdu–Lodz–Nuremberg–Tilburg	9,826 (to Lodz)	12–15	2013	21	Alashankou/Dostyk Khorgos/Altynkol	Terespol/Brest ^a	KZ, RU, BY, (PL, DE)
Wuhan–Minsk–Hamburg	N/A	12–15	N/A	2	Zabaikalsk/Manzhouli ^b	Terespol/Brest ^a	RU, BY, PL
Wuhan–Pardubice–Lodz–Hamburg–Duisburg	10,863 (to Pardubice)	15	2012	4	Alashankou/Dostyk Khorgos/Altynkol	Terespol/Brest ^a	KZ, RU, BY, (PL, CZ, DE)
Yiwu–Madrid	13,052	18	2014	1	Alashankou/Dostyk Khorgos/Altynkol	Terespol/Brest ^a Irún/Hendaye	KZ, RU, BY, PL, DE, FR
Yiwu–Minsk	N/A	12	N/A	1	Zabaikalsk/Manzhouli ^b	None	RU
Yiwu–Istanbul	N/A	18	N/A	1	Alashankou/Dostyk Khorgos/Altynkol	None	KZ, AZ, AM, GE
Hefei–Hamburg	11,000	15	2014	1	Kartsakhi/Çıldır Alashankou/Dostyk Khorgos/Altynkol	Terespol/Brest ^a	KZ, RU, BY, PL, DE
Suzhou–Warsaw	11,200	12	2012	4	Zabaikalsk/Manzhouli ^b Zamyd-Uud/Erenhot	Terespol/Brest ^a	(MN), RU, BY, PL
Lianyungang–Istanbul	N/A	18	N/A	1	Alashankou/Dostyk Khorgos/Altynkol	None	KZ, AZ, AM, GE
Shenyang–Hamburg	N/A	13	N/A	10	Kartsakhi/Çıldır Zabaikalsk/Manzhouli ^b	Terespol/Brest ^a	RU, BY, PL
Changchun–Dresden	N/A	13	N/A	2	Zabaikalsk/Manzhouli ^b	Terespol/Brest ^a	RU, BY, PL
Shenyang–Moscow	N/A	12	N/A	1	Zamyd-Uud/Erenhot		MN
Changsha/Guizhou–Hamburg	N/A	15	N/A	1	Alashankou/Dostyk Khorgos/Altynkol	Terespol/Brest ^a	(KZ/MN), RU, BY, PL
Guangzhou–Moscow	N/A	12	N/A	3	Zamyd-Uud/Erenhot Zabaikalsk/Manzhouli ^b	None	None
Tianjin–Moscow	N/A	11	N/A	2	Zabaikalsk/Manzhouli ^b	None	None
Chifeng–Chelyabinsk/Kleshchikha	N/A	10	N/A	1	Zabaikalsk/Manzhouli ^b	None	None
Xiamen–Hamburg	N/A	16	N/A	1	Alashankou/Dostyk Khorgos/Altynkol	Terespol/Brest ^a	KZ, RU, BY, PL
Changsha–Duisburg	11,808	18	2014	1	Alashankou/Dostyk Khorgos/Altynkol	Terespol/Brest ^a	KZ, RU, BY, PL
Xiamen–Moscow	N/A	13	N/A	1	Zamyd-Uud/Erenhot	None	MN
Nantong–Mazar-i-Sharif	N/A	5	N/A	1	Alashankou/Dostyk Khorgos/Altynkol	None	KZ, UZ

Harbin–Moscow–Warsaw–Hamburg	9,820	10–15	2015	3	Zabaikalsk/Manzhouli ^b	Terespol/Brest ^a	(RU, BY, PL)
Jining–Moscow	N/A	5	N/A	1	Zamyd-Uud/Erenhot	None	MN
Jiaozhou–Hanoi	N/A	5	N/A	1	Pingxiang/Dong Dang	None	None
Hamburg–Zhengzhou	10,214	18	2013	3	Alashankou/Dostyk	Terespol/Brest ^a	PL, BY, RU, (KZ, MN)
					Khorgos/Altynkol		
Duisburg–Chongqing	11,179	18	2011	7	Zamyd-Uud/Erenhot	Terespol/Brest ^a	KZ, RU, BY, PL
					Alashankou/Dostyk		
					Khorgos/Altynkol		
Tilburg–Nuremberg–Lodz–Chengdu	9,826	18	2013	10	Zamyd-Uud/Erenhot	Terespol/Brest ^a	(DE, PL), BY, RU, KZ
	(from Lodz)				Alashankou/Dostyk		
Madrid–Yiwu	13,052	20	2014	1	Khorgos/Altynkol		FR, DE, PL, BY, RU, KZ
					Alashankou/Dostyk		
Hamburg–Wuhan	N/A	18	N/A	2	Khorgos/Altynkol	Terespol/Brest ^a	PL, BY, RU, KZ
					Alashankou/Dostyk		
Brest–Suzhou	N/A	15	N/A	1	Zabaikalsk/Manzhouli ^b	None	RU
Brest–Shenyang	N/A	15	N/A	1	Zabaikalsk/Manzhouli ^b	None	RU
Tomsk–Wuhan	N/A	15	N/A	2	Zabaikalsk/Manzhouli ^b	None	None
Hamburg–Zhengzhou	10,214	15	N/A	N/A	Alashankou/Dostyk	Terespol/Brest ^a	PL, BY, RU, KZ
					Khorgos/Altynkol		
Dresden–Tomsk–Harbin–Changchun	N/A	15–18	N/A	2	Zabaikalsk/Manzhouli ^b	Terespol/Brest ^a	PL, (RU)
Tomsk–Chongqing	N/A	16	N/A	2	Zabaikalsk/Manzhouli ^b	None	None
Vorsino–Jining	N/A	10	N/A	1	Zamyd-Uud/Erenhot	None	MN
Tangshan–Antwerp	11,000	16	2018	N/A	Alashankou/Dostyk	Terespol/Brest ^a	KZ, RU, BY, PL, DE
Yiwu–London ^c	12,000	12	2017	N/A	Alashankou/Dostyk	Terespol/Brest ^a	KZ, RU, BY, PL, DE, BE, FR
					Khorgos/Altynkol		

^aWhile the majority of rail freight to Europe crosses the Terespol/Brest border, there are further gauge changes such as Visoko-Litovsk/Czeremcha, Bruzgi/Kuznica, Svisloch/Siemianowka, and Dzerzhinskaya-Novaya intermodal terminal in Kaliningrad.

^bFurther gauge changes between Russia and China are Grodekovo/Suifenhe, Makhalino/Hunchun, and the new crossing Nizhneleninskoye/Tongjiang.

^cContainers that are destined for the United Kingdom are reloaded to special wagons for the Channel tunnel. Transshipment amongst others takes place in Duisburg, Germany.

Intermodal and Sychromodal Freight Transport

Tomas Ambra, Koen Mommens, Cathy Macharis, Vrije Universiteit Brussel, MOBI Research Center, Brussels, Belgium

© 2021 Elsevier Ltd. All rights reserved.

Introduction	456
Intermodal Transport—Its Costs and Evolution	456
Sychromodal Transport	458
Increasing the Capacity of Loading Units	459
Synchronization of Assets	459
Infrastructure and Networks	459
Digitalization and Data Sharing	460
Conclusion	461
Acknowledgment	461
Relevant Websites	461
References	462
Further Reading	462

Introduction

As international trade and cargo demands are growing, existing infrastructure capacities will be stretched to their limits, leading to increasing congestion, safety concerns, environmental issues, and diminishing reliability of services. Given the major challenge of decarbonizing the freight transport sector in the face of climate change, existing instruments will not suffice to sustainably deal with the expanding market (European Commission, 2011). By 2050, freight is projected to account for 60% of total transport greenhouse gas emissions, up from 42% in 2010 (OECD, 2015). Therefore, the European Commission aims to move 30% of road freight transport to more sustainable modes such as rail and inland waterways by 2030, increasing to 50% by 2050. These ambitions will require innovative solutions for a more seamless and cost-effective integration of the various transport modes. On the way to more environmental and socioeconomic sustainability, the usage of the existing assets constitutes a major challenge for the transport sector. Here, the core issue is the current user preferences and mode selection in favor of road transport. Given the dense road network, road transport can offer door-to-door transport solutions from almost anywhere to everywhere on the continent. Such unimodal door-to-transport chains are much less likely to be available via rail or inland waterway transport. Instead, they often need to be combined with other modes to form one integrated transport chain. Such transportation of goods by at least two differing modes is defined as multimodal transport. Further defining concepts exist: Intermodal transport is the movement of goods (which remain in the same vehicle or loading unit) by numerous modes of transport without the goods themselves being handled when switching modes (Eurostat, 2003). Combined transport applies more emphasis on reducing the road delivery element within the initial/final leg, so that the main leg by barge or train is as long as possible in order to reap economies of scale. Comodality focuses on balancing different modes and their combinations for optimum overall resource efficiency (Eurostat, 2003). Lastly, synchronomodality emphasizes the structured and synchronized combination of modes, including real time (re-) evaluations of routes and mode selections during transport (Steadie Seifi et al., 2014). The chapter firstly describes intermodal transport, its costs and evolution in the following sections. Secondly, the concept of synchronomodal transport is addressed in section, “Synchronomodal Transport” which extends the idea of intermodal transport by enabling real-time re-routing of goods along alternative routes in response to disruptions and/or shippers’ requirements. Thirdly, digitalization and data sharing are briefly discussed in section, “Digitalization and Data Sharing” followed by concluding remarks in section, “Conclusions.”

Intermodal Transport—Its Costs and Evolution

A distinctive characteristic of intermodal transport is the use of an intermodal loading unit (ILU). The success of intermodality depends upon the more general adoption of standards that has facilitated globalization, namely through the global use of containers. In 1961, the International Standards Organization (ISO) reached an agreement to set standard sizes of containers to 20-foot (1 TEU or Twenty-foot Equivalent Unit) and 40-foot (2 TEU) lengths. The introduction of containers contributed to lowering freight charges by improving port handling efficiency, but the real outcome of containerization was a boost in trade flows, with major effects that are even noticed by the average person. Every day at every major port, thousands of containers arrive and depart to carry goods that we are depending on. Containers can be moved seamlessly between ships, inland waterway vessels, rail wagons, and trucks through a system of intermodality. In order to establish a complete intermodal freight transport chain, a broader perspective of the logistical chain is considered: not just that the containers need to be standardized, but also the container ships, intermodal terminals,

trucks, and trains need to be adapted to carry the same container, to get it into the hinterland. The major disadvantage of the use of loading units is, however, the need to transport empty or half-empty loading units to satisfy the demand. As a consequence of imbalanced trade flows, empty container repositioning is common practice.

Intermodal transport systems can include a variety of transport modes like shortsea, inland waterway, rail, and air transport. Road transport is commonly used for the pre- and post-haulage, as its extensive infrastructure network enables it to access many locations. Given the growing volumes of maritime containers transhipped, the seaports represent an ideal starting point to stimulate intermodal transport. Establishing intermodal connections not only enables an extension of the hinterland potential of seaports but also contributes to improving the efficiency of the transport system. Intermodal terminals are the critical nodes that allow transhipments between different transport modes. For the transhipment of containers in these intermodal terminals, dedicated infrastructure is required.

In intermodal transport, different logistics chain configurations may involve different types of actors, or actors of the same type playing different roles (Vrenken et al., 2005). Shippers initiate the movement of cargo between locations directly or by means of contracts on their behalf. Intermodal transport services can be optimized by freight forwarders on behalf of shippers. Ocean shipping lines, can also be involved in the hinterland container shipment segment. On the supply side, we find: terminal, rail, inland navigation, shortsea and road transport operators. Operational organization places terminal operators at the core of the intermodal transport chain because of their role in transhipping ILUs between the main haul and drayage. Transport operators handle the movement of the loading units between terminals via rail, inland waterway or sea routes respectively, and road transport operators arrange the local transportation of cargo from origin and destinations. Offering door-to-door or terminal-to-terminal transport, intermodal transport operators procure transport and transhipment services. In addition to these purely commercial market players, the public sector can also be included on the supply side of intermodal transport. Infrastructure managers, port authorities, regional and national public authorities, and international institutions contribute to making the best possible use of infrastructure and provide an environment to encourage intermodal initiatives.

Fig. 1 represents an intermodal and a road-only cost function. For a port-to-door intermodal transport chain, the function shows the total intermodal transport costs, relative to distance, between a port and a final destination. Post-haulage requires an interchange from road transport to another transport mode in an intermodal terminal. From Fig. 1, it is possible to derive the importance of transhipments in an intermodal transport chain.

At the port, intermodal transport has larger handling costs compared to unimodal road transport. This is due to the operations needed for the transhipment of containers on barges or trains. The advantage of intermodal transport lies in the smaller variable costs during main haulage, as a result of the scale economies that are obtained by the large capacities that can be transported by a single movement. Scale economies, gained by the main haulage leg of an intermodal transport chain, can further be increased by the introduction of larger vessels or longer trains. As the variable costs of barge or rail transport is lower compared to road-only transport, the longer distances covered by the intermodal leg will make intermodal transport more efficient than road-only transport. At the end of the chain, this advantage is partly compensated by the extra handling cost that has to be paid for the terminal handling.

Once the different cost components are known, it is possible to make comparisons between intermodal and unimodal road transport. This relationship enables us to understand which transport alternative is costlier in a given situation based on the concept

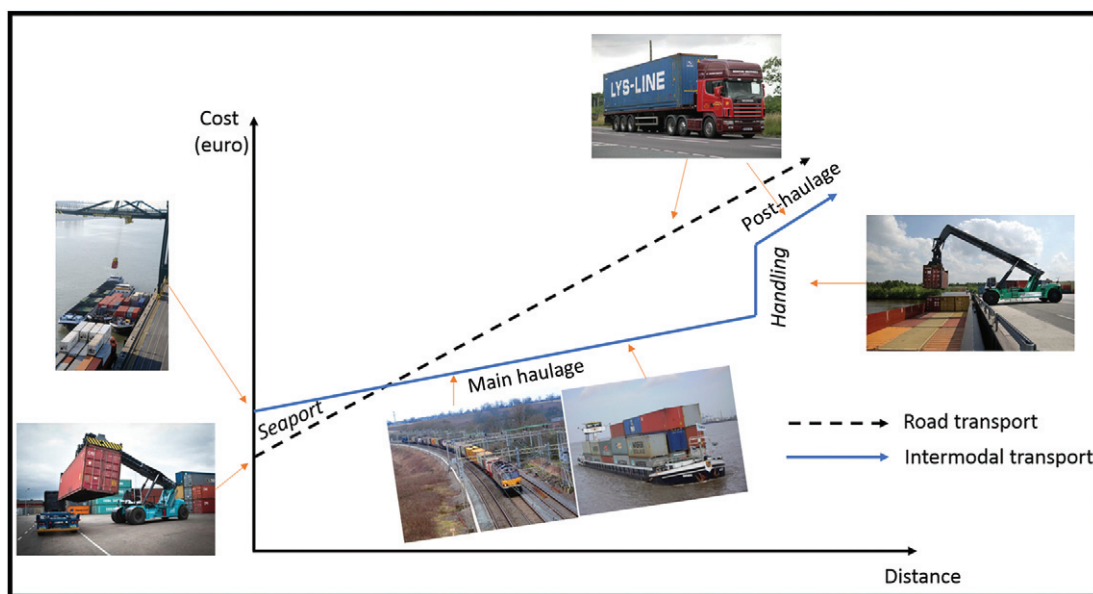


Figure 1 Intermodal cost function.

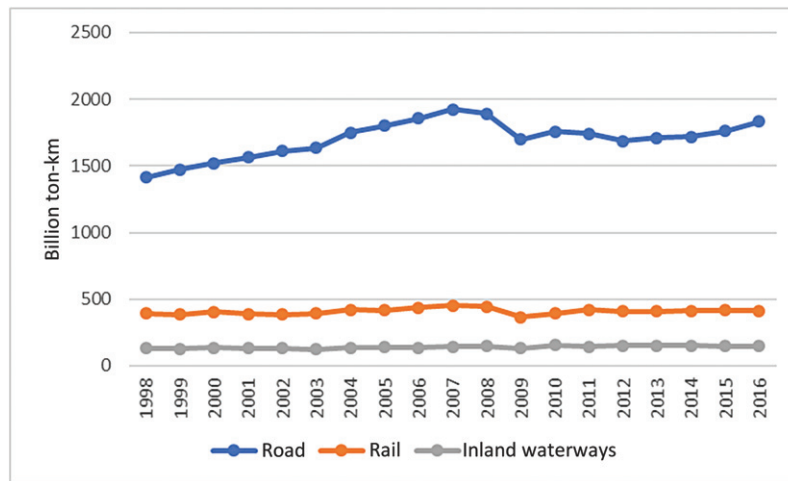


Figure 2 Evolution of transport performance in EU-28 between 1998 and 2016. Source: Eurostat (2003)

of break-even analysis. Fig. 1 shows that unimodal road transport performs better compared to intermodal transport for short distances. When a certain distance is reached, the costs of road and intermodal transport are equal. This is called the break-even distance. Above the critical distance point, intermodal transport costs are lower than those of unimodal road transport. The break-even distance is a theoretical concept, as in practice the transport costs are influenced by factors such as capacity utilization, the presence or lack of a full backhaul, etc., and the fact that actual transport distances can be impacted by detours. Other cost factors such as the empty container depot functioning as an intermodal terminal and the effects of congestion also have to be considered in calculating the break-even distances. The actual break-even distance will in any case depend on local market conditions.

Intermodal transport is supported and facilitated by the European Commission as well as by national and regional governments. For port authorities a modal shift can increase their accessibility—by decreasing congestion and extending or creating hinterlands (van Klink and van den Berg, 1998) and, therefore, increase their competitiveness (Vermeiren, 2013). Also, shippers can benefit from a modal shift if the new transport chain meets their preferences (cost, quality, reliability, security, etc.) more than the previous chain. Yet, we see that modal split has remained more or less constant over the last two decades in Europe (Fig. 2).

Synchromodal Transport

Intermodal transport has been seen as a solution to accommodate the shift from road to rail and inland waterways. However, incorporating all possible actors, modes, and trajectories into functioning decision support models for intermodal transport is challenging. The available data is limited and its static nature poses problems. Furthermore, it is perceived as slow, inflexible, and limited in its application from the user perspective of actors such as shippers and logistics service providers (Meers et al., 2017). This also explains the evolution in Fig. 2 in which the modal shift to inland waterways and railways has been modest at best. The target of achieving a modal shift as set out in the introduction is, therefore, still a challenge to attain. Here, the concept of synchromodal transport extends the idea of intermodal transport by enabling real-time re-routing of goods along alternative routes in response to disruptions and/or shipper requirements (Ambra et al., 2019a). It therefore differs from intermodal transport, where services are defined long in advance leading to a static service with low flexibility, and where disturbances and delays easily lead to loss of time or money. Furthermore, the intermodal setting is too inflexible to accommodate newly arriving orders in time. Within the synchromodal approach, routing, and modal choices are made as late as possible to account in real-time for operational and infrastructural developments (Reis, 2015; Tavasszy et al., 2017). Such dynamic re-routing and re-scheduling make it more competitive. This way, synchromodal systems may outperform intermodal transport in terms of reliability and flexibility, and lead to a better integration of different transport modes. Synchromodality is also seen as a solution to counteracting congestion problems and environmental concerns raised by the ever-increasing cargo demands. The concept has even become a component in the European Technology Platform called Alliance for Logistics Innovation through collaboration in Europe (ETP-ALICE) which develops research and innovation agendas for action on supply chain management and logistics at EU and national levels (Fig. 3).

To implement synchromodality, bookings should be made mode-free by shippers. This type of a-modal booking would allow logistics services providers (LSPs) to plan and combine orders more flexibly by making use of various modes, while adhering to customers' requirements. Next to enabling a-modal booking, synchromodal transport requires three further system changes: (1) a shift in approach and thinking from "predict and prepare" to "sense and respond," (2) the creation of standardized cooperation schemes, and (3) a governance system leading to improved operational alignment of the various modes (Tavasszy et al., 2017).

In the synchromodal setting, three aspects of freight transport require synchronization (Ambra et al., 2019b). These are (1) loading units, (2) assets, and (3) infrastructure, with each aspect addressing problems at a different level.

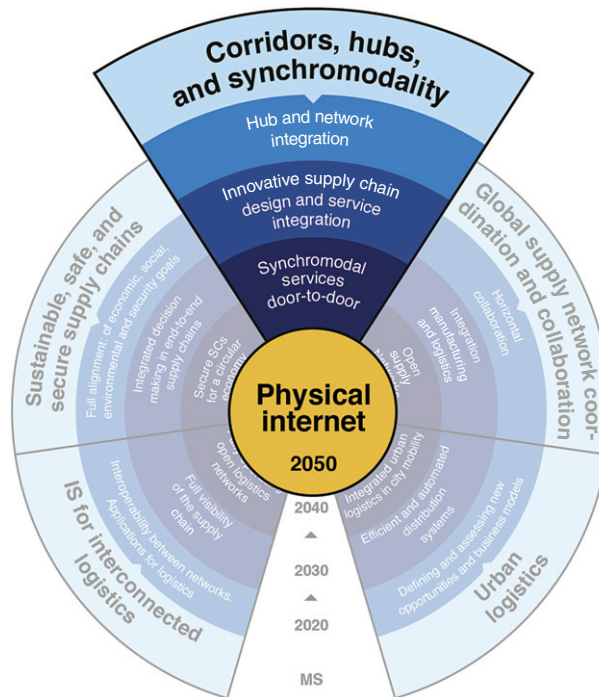


Figure 3 ETP-ALICE roadmap toward zero emissions.

Increasing the Capacity of Loading Units

At the basis of synchromodal transport stands the loading unit. Loading units such as intermodal freight containers are widely standardized to allow for easy handling and stacking at transport hubs and their capacity is usually indicated in TEUs. To achieve a higher fill rate of containers, the lower parcel and pallet levels of a loading unit should be accounted for via modularization, advanced encapsulation processes, and multifunctional load units (Sallez et al., 2015; Tran-Dang et al., 2017), leading to transparency and efficiency across levels. To maximize capacity utilization and revenue, smart containers supported by cloud-computing can achieve better grouping strategies in real time with less empty containers. This way, revenue management of the overall supply chain (Van Riessen et al., 2017) and of individual TEUs should be intertwined. Lastly, sensors and smart tags allow for the tracking of various orders enclosed in a container. Such visibility and location intelligence are paramount for the creation of a unified system that can reliably identify overlapping flows. Such inefficiencies could be assessed and visualized by the synchromodal system, to be then more efficiently bundled at the level of boxes, pallets, and containers (TEUs).

Synchronization of Assets

Another aspect to be synchronized is the asset such as barges, trains, and trucks that move between hubs and delivery locations. In addition, there are also cranes, tugs, and reach stackers which operate in the ports and terminals. Just like containers, these moving assets vary in terms of their size, geographical scale, and infrastructure they follow. Here, synchronization encompasses the optimization of train and barge schedules, the allocation of loading units to modes, and the handling of disruptions and their consequent cost impact (Ambra et al., 2019a; Behdani et al., 2016; Nabais et al., 2015). Nevertheless, intermodal and synchromodal applications concern mainly longer hauls; at the European interregional scale, for instance, while detailed local trucking tends to be detached from the overall solutions. Thus, the main European hubs and corridors should interact with local distributors to create mass for barges and trains, while trucks support synchromodal services by carrying out last-mile deliveries. Merging these two would offer more reliable door-to-door synchromodal coverage where, through integrated planning of modes, fill rates of all transport means can be increased once trucking companies and rail/barge operators optimize their schedules and routings together. Reinforced usage of assets is very important, yet it requires standardized coordination protocols to ease bundling or transitioning, if one intends to preempt closed collaboration contracts that would exclude some service providers.

Infrastructure and Networks

A network in freight models is made up of nodes as well as arcs connecting these nodes. In the case of the freight network, the nodes or hubs are ports, warehouses, and terminals, which are linked by arcs such as roads, rail, and inland waterways. At hubs, the

challenge arises to integrate and align intra-terminal operations with those of logistics service providers and road-inland waterway-rail network developments. This is facilitated by a combination of two types of algorithms: Route prediction algorithms, which are integral to the calculation of the estimated time of arrival and cargo peak prediction algorithms at hubs (Nabais et al., 2015). The coordination of internal-hub movements with external movement outside of the hubs can be facilitated by RFID, Bluetooth, and WIFI structures employed by manufacturers and terminal operators for higher visibility, and other “smart” object solutions for longer distances by means of GPS-enabled devices. For more resilient supply chain structures, synchromodal solutions should be able to react both to external events such as rail strikes or lower water levels, and to factory production disruptions. Besides these, opportunities such as vacant capacities and newly incoming orders can be accommodated in an agile manner.

Transparently tracking overlapping flows is crucial for efficiently exploiting all capacities. Through IoT-enabled infrastructure and RFID features, interregional synchromodal flows could be synced with higher density networks so that goods could be grouped effectively, while keeping loading and delivery times to a minimum. Similarly, traffic management systems of various network operators should be integrated toward using arcs more efficiently to reduce overall network congestion through better planning and route selection. Such data is necessary to enable real-time response modeling and further realistic and holistic models. With the growth in big data analytics, research into control-tower information technology needs, decentralized applications, as well as semantic data models will gain traction.

The dynamic shifting of modes often helps to improve the service level in the synchromodal context. Yet, it may not always be appropriate, as it can lead to deviations and longer lead-times given the fact that some route and mode switching could lead to hubs located further away from the order’s final destination (Ambra et al., 2019a). Thus, when considering such proactive solutions, spatial attributes and request urgency should also be taken into account, and weighed against the impact of deviations and switching on key performance indicators. Balancing these considerations is challenging due to many uncertainties and hidden consecutive processes, such as the objectives of other service providers and the location of their assets, working conditions, congestions levels, blockages, weather forecasts and similar contextual network information. The necessary transparency of the network and overview of assets may be attained with IoT technologies and geo-spatial network coverage (e.g., 5G), subsequently enabling remote control, information exchange and automation of assets. The next section touches upon this matter in more detail.

Digitalization and Data Sharing

The various modes (road, rail, and inland waterway) currently do not share integrated data platforms. This reduced interoperability, exacerbated by noncompatible software, hamper the creation of holistic synchromodal services at the national and international scale. For inland waterways, the European Rive Information Services (RIS) addresses the latter by acquiring and sharing dynamic information across borders. Since planning in a synchromodal setting is to take place as late as possible, planning and execution horizons almost align, making the reliance on real-time developments for adjustment vital. Real-time response modeling should be developed further, since switching and transporting disturbed or newly incoming flows comes with additional costs. To make the system financially viable, one may develop approaches that bundle disturbed flows with items that have a long lead time. Specifically, should a truck break down or other developments impede the transport of a given payload, the goods could be switched to a different mode at a terminal (loaded on a train/barge). The disturbed or newly incoming payload should, therefore, be combined with other shipments, potentially resulting in higher fill rates or in a new service.

The key to effective synchromodal transport is real-time visibility across levels and continuous track and trace beyond mere point-to-point monitoring. This includes the inventory and in-house processes of hubs, the state of infrastructure such as roads, rail, and inland waterways, and the status of all moving assets. Uncertainties can be further reduced through more data transparency and technologies increasing visibility. Synchromodality extends the more static intermodal approach by enabling dynamic planning and mode switching. Yet as data volumes increase, control towers as central information hubs, become necessary when providing a single reference point to terminals and assets for effective service integration and container allocation. Such control towers vary on the ground. There are port community systems (PCS) enabling exchange of information between stakeholders at a given port, as well as proprietary control towers run by a single logistics service provider (LSP) where the data is only exposed to their clients and contractors.

The biggest hurdle on the way to the envisioned holistic system is the currently fragmented landscape of applications ranging from various port community systems, field force automation tools, fleet and freight management tools, enterprise resource planning systems, and supply chain execution tools to terminal operating systems. A unified freight transport system (Fig. 4) will only be achievable when the various tools are made compatible and standardized and when the flows are monitored through control towers in combination with decentralized system solutions. In addition, the production and freight transport systems should be connected so that local flows can be linked to synchromodal corridors, to facilitate a positive modal shift by reaching a critical mass for barges and trains. Furthermore, synchromodal control towers should be supported by decentralized applications that take care of point-to-point tracking, freeing up capacities for the analysis of newly incoming orders in terms of their temporal and spatial attributes and transport needs.

In the EU, significant financial resources are being invested in the creation of the Trans-European Transport Network (TEN-T) consisting of nine main corridors. Conversely, these should also link up to local distribution and manufacturing systems and

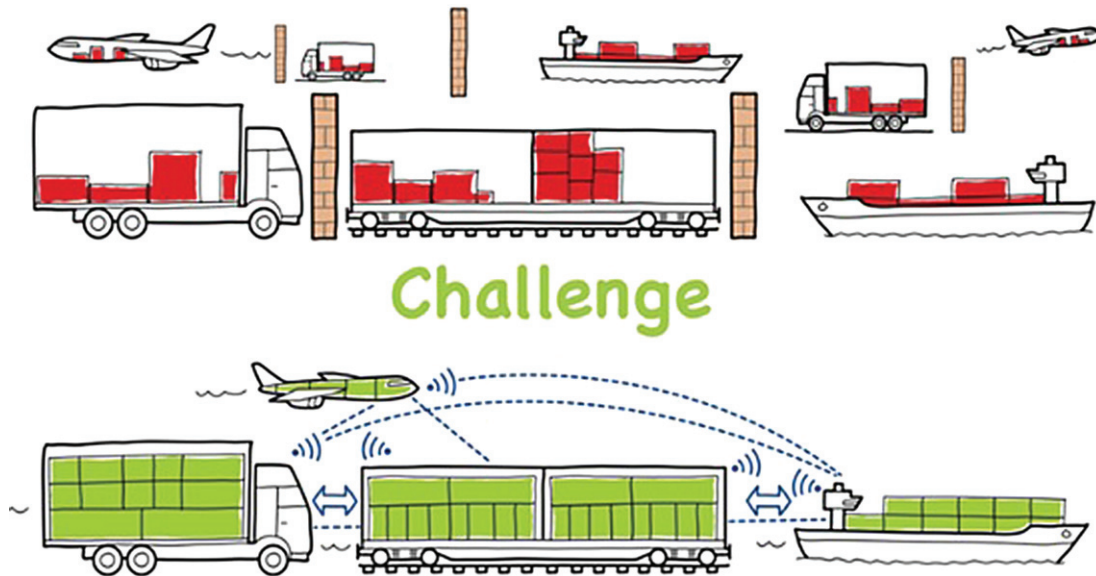


Figure 4 Conceptual barriers and underutilized assets (red) to be improved by interoperable communications and transparency (green). Source: Adapted from ETP-Alice

enable synchromodal planning. However, as flexibility and real-time planning become more prominent, flows that are currently static and predictable may become less predictable in a world of continuous updates and shifting of contexts. For example, low water levels could trigger everyone to shift to road or rail, potentially resulting in adverse effects in other infrastructure sections. Therefore, algorithms should be developed that could account for such unintended cascading effects in other parts of the freight system.

Conclusion

Multimodal/intermodal freight transport has undergone various changes. The breakthrough with the “box” has facilitated the physical handling of goods (Levinson, 2006). The next step is to induce a breakthrough with the digital dimension, which should facilitate information exchange about the “box.” Data availability and accessibility can improve synchromodal models and planning tools in order to provide more flexibility, higher speed, reliability, and a transparent overview of ongoing freight transport processes; these are the aspects that hinder most shippers from using intermodal services. In fact, models and decision-making tools without real-time situational data are generic. With real-time data feeds, they become unique and more realistic. Synchromodal transport is an extension of intermodal transport with a premise to make the freight transportation system more resilient, flexible, and responsive to newly incoming orders, disruptions, and breakdowns. But most importantly, it should be perceived as another word for modal shift, in which case fragmented orders are bundled in a transparent manner, with a goal to create critical mass for barges and trains. Such developments will contribute to increasing fill rates of environmentally friendlier transportation modes and reduce the share of road freight transport. Nevertheless, trucks will remain to play a crucial role in last-mile deliveries, as the final delivery locations are rarely located in ports and terminals. In this regard, the last-mile transport carried out by trucks also needs to be accounted for, to reduce empty and half-empty truckloads. To conclude, the most recent synchromodal stage is an exciting trend that will reveal many new opportunities, but also problems. Nonetheless, the scientific community, transport users, transport providers and policy makers can expect many new developments in terms of models and applications that will, hopefully, contribute positively to the roadmap toward the 30% reduction of road freight transport by 2030 and 50% by 2050.

Acknowledgment

This work was supported by the Research Foundation Flanders (FWO)—Doctoral grant for strategic basic research.

Relevant Websites

ETP-ALICE. Available from: www.etp-logistics.eu.

References

- Ambra, T., Caris, A., Macharis, C., 2019a. Should I stay or should I go? assessing intermodal and synchromodal resilience from a decentralized perspective. *Sustainability* 11 (6), 1765.
- Ambra, T., Caris, A., Macharis, C., 2019b. Towards freight transport system unification: reviewing and combining the advancements in the physical internet and synchromodal transport research. *Int. J. Prod. Res.* 57 (6), 1606–1623.
- Behdani, B., Fan, Y., Wiegman, B., Zuidwijk, R., 2016. Multimodal schedule design for synchromodal freight transport systems. *Eur. J. Transp. Infra. Res.* 16 (3), 424–444.
- European Commission, 2011. White Paper on Transport: Roadmap to a Single European Transport Area: Towards a Competitive and Resource-efficient Transport System. Publications Office of the European Union, Brussels.
- Eurostat, 2003. Glossary for Transport Statistics European Conference of Ministers of Transport (ECMT), United Nations Economic Commission for Europe (UNECE).
- Levinson, M., 2006. *The Box*. Princeton University Press, New Jersey.
- Meers, D., Macharis, C., Vermeiren, T., van Lier, T., 2017. Modal choice preferences in short-distance hinterland container transport. *Res. Transp. Bus. Manag.* 23, 46–53.
- Nabais, J.L., Negenborn, R.R., Carmona-Benítez, R., Botto, M.A., 2015. Cooperative relations among intermodal hubs and transport providers at freight networks using an MPC approach. In: *International Conference on Computational Logistics*, Springer, Cham, pp. 478–494.
- OECD, 2015. ITF Transport outlook 2015. Available from: <http://www.oecd.org/publications/itf-transport-outlook-2015-9789282107782-en.htm>.
- Reis, V., 2015. Should we keep on renaming a +35-year-old baby? *J. Transp. Geogr.* 46, 173–179.
- Sallez, Y., Montreuil, B., Ballot, E., 2015. On the Activeness of Physical Internet Containers. In: Borangiu, T., Thomas, A., Trentesaux, D. (Eds.), *Service Orientation in Holonic and Multi-Agent Manufacturing*. Springer International Publishing, Cham, pp. 259–269.
- Stadie Seifi, M., Dellaert, N.P., Nuijten, W., Van Woensel, T., Raoufi, R., 2014. Multimodal freight transportation planning: a literature review. *Eur. J. Operat. Res.* 233 (1), 1–15.
- Tavasszy, L., Behdani, B., Konings, R., 2017. Intermodality and synchromodality. In: Geerlings, H., Kaipers, B., Zuidwijk, R. (Eds.), *Ports and Networks*. Routledge, London, United Kingdom, pp. 251–266.
- Tran-Dang, H., Krommenacker, N., Charpentier, P., 2017. Containers monitoring through the physical internet: a spatial 3D model based on wireless sensor networks. *Int. J. Prod. Res.* 55 (9), 2650–2663. doi:10.1080/00207543.2016.1206220.
- van Klink, H.A., van den Berg, G.C., 1998. Gateways and intermodalism. *J. Transp. Geogr.* 6, 1–9.
- Van Riessen, B., Negenborn, R.R., Dekker, R., 2017. The cargo fare class mix problem for an intermodal corridor: revenue management in synchromodal container transportation. *Flex. Serv. Manu. J.* 29, 1–25.
- Vermeiren, T., 2013. *Intermodal Transport - The delta in the Delta*. Vrije Universiteit, Brussel.
- Vrenken, H., Macharis, C., Wolters, P., 2005. *Intermodal Transport in Europe*. EIA, Brussels 267 pp.

Further Reading

- Arancibia, A.I., Feo-Valero, M., García-Menéndez, L., Román, C., 2015. Modelling mode choice for freight transport using advanced choice experiments. *Transp. Res. Part A: Policy Prac.* 75, 252–267.
- Bontekoning, Y.M., Macharis, C., Trip, J.J., 2004. Is a new applied transportation research field emerging?—a review of intermodal rail–truck freight transport literature. *Transp. Res. Part A: Policy Prac.* 38, 1–34.
- Caris, A., Macharis, C., Janssens, G.K., 2013. Decision support in intermodal transport: a new research agenda. *Comp. Indus.* 64, 105–112.
- Crainic, T.G., Kim, K.H., 2007. Intermodal transportation. In: Barnhart, C., Laporte, G. (Eds.), *Handbook in OR & MS. Handbook in OR & MS: Transportation*, Vol. 14. North Holland Publishing Co, Amsterdam, doi:10.1016/S0927-0507(06)14008-6, pp. 467–537.
- European Commission, 2006. *Keep Europe moving. Sustainable mobility for our continent - Mid-term review of the European Commission's 2001 transport White Paper*.
- Kreutzberger, E., Macharis, C., Woxenius, J., 2006. Intermodal Versus Unimodal Road Freight Transport – A Review of Comparisons of the External Costs. In: Jourquin, B., Rietveld, P., Westin, L. (Eds.), *Towards Better Performing Transport Systems*. Taylor and Francis, London, pp. 17–42.
- Macharis, C., Bontekoning, Y.M., 2004. Opportunities for OR in intermodal freight transport research: a review. *Eur. J. Operat. Res.* 153, 400–416.
- Mathisen, T.A., Sandberg Hanssen, T.-E., 2014. The academic literature on intermodal freight transport. *Transportation Research Procedia*. 17th Meet. EURO Work. Gr. Transp. EWGT2014, 2–4 July, Sevilla, Spain 3, pp. 611–620.
- McGinnis, M.A., 1989. A comparative evaluation of freight transportation choice models. *Transp. J.* 29, 36–46.

Pipelines

Matthew E. Oliver, Georgia Institute of Technology, School of Economics, Atlanta, GA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	463
Basic Engineering	465
Basic Economics	467
Additional Considerations	469
Pipeline Safety and Integrity	469
Pipelines and Geopolitics	469
Acknowledgment	469
References	470

Introduction

Pipelines, or pipeline systems, generally refer to an interconnected network of pipes and related facilities used for transporting gas or liquids—typically fuels such as natural gas or hydrocarbon liquids such as crude oil—over long distances. Pipelines are a highly efficient means of inland fuel transport, connecting oil and/or gas producers at one end to refineries, retail gas distributors, or power plants at the other. Pipelines are often referred to as the oil and gas *midstream* because they connect upstream producers with downstream users (e.g., retail gas markets, oil refineries, etc.). This article reviews the basic engineering and economics of pipeline systems, giving particular focus to natural gas pipelines and oil pipelines, which make up the vast majority of existing pipeline systems worldwide. Also included are short sections reviewing basic pipeline safety considerations and the role of pipelines in geopolitics.

Pipelines are of central importance to the overall hydrocarbon fuel value chains. Fig. 1 depicts a simple schematic diagram of these value chains. The gas value chain begins at the point of production, or wellhead, from which gathering pipelines deliver raw gas to nearby processing plants. These processing plants separate out natural gas liquids, water, carbon dioxide and other inert gases, and various impurities, and then feed a homogeneous finished gas product consisting primarily of methane, ethane, and propane, into long-distance transmission pipelines, which transport the gas to various end users. Alternatively, liquefied natural gas (LNG) may be imported via specialized import terminals, which, after regasification, feed into the long-distance transmission system. At various points along the pipeline system, gas may be held in storage in natural underground caverns or pressurized tank farms. In the United States, storage facilities are regulated as part of the overall interstate pipeline transmission system. Large mainline industrial customers—primarily gas-fired electric power generators—receive gas directly from the long-distance transmission system via ancillary lines. Some gas may also be delivered to LNG export terminals. The largest users of gas are local distribution companies (LDCs) and gas-fired power plants. LDCs receive gas from the long-distance transmission system at specialized hubs called “city gates.” From there, gas is delivered to homes and other retail customers via smaller distribution pipelines.

The oil value chain begins with crude oil. Domestically produced crude oil may be refined locally or shipped via long-distance pipelines (or by truck or rail) either for export or for delivery to major refining centers. Imported crude oil may be shipped via pipeline or delivered directly to the major refineries. Refineries process the crude oil into various derivative petroleum products (e.g., various types of gasoline, diesel, heating oil, lubricants, and a range of petrochemical precursors), which are then shipped via petroleum products pipelines, truck, or rail to petroleum products markets.

The world’s first oil and gas pipelines were built in the United States during the second half of the 19th century (American Petroleum Institute, 2019). At that time, western Pennsylvania was the center of the early oil industry. In 1863, a group of oil producers built a 9-mile wooden pipeline to a nearby rail station in order to circumvent the regional monopoly position in local oil transportation held by the *Teamsters* (organized teams of men driving horse-drawn carriages who transported barrels of oil and other goods). The first successful natural gas pipeline—2 inches in diameter and just over 5 miles long—was built in 1872 to transport “associated” gas to the Titusville, PA town center from a nearby oil well. The first long-distance oil trunk line was the *Tidewater* line, built by independent oilmen in 1879 to compete with aggressive oil baron John D. Rockefeller’s stranglehold over oil transportation (soon thereafter Rockefeller purchased a half ownership stake in *Tidewater*). The first long-distance gas pipeline was built in 1891 to carry gas to Chicago from fields in Indiana. Today, most countries around the world have at least some type of pipeline system (or systems) for transporting oil, natural gas, refined petroleum products, or simply water. The most extensive oil and gas pipeline systems in the world are located in the United States and Russia (Table 1).

The world’s largest pipeline system is the US natural gas transmission and distribution network, which comprises nearly 2 million km of gathering, transmission, and distribution pipelines. The overall US natural gas market, however, is divided into three distinct sectors—production, transmission, and distribution. Fig. 2 shows maps of the US long-distance natural gas pipeline transmission network (gathering and distribution lines not shown). This network comprises roughly 500,000 km of high-capacity pipelines

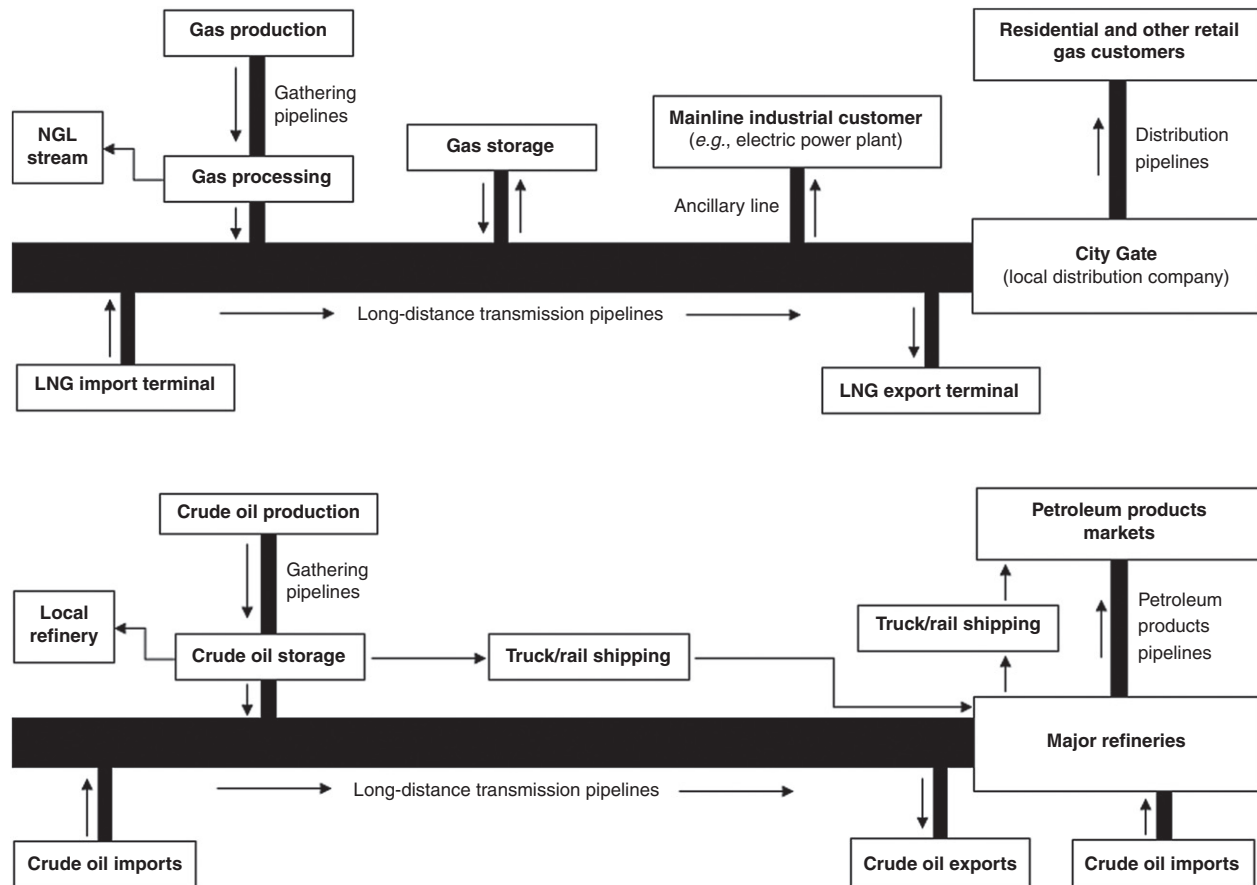


Figure 1 Hydrocarbon value chains: natural gas (top) and oil (bottom).

Table 1 Largest pipeline systems by country

<i>Natural gas</i>		
<i>Country</i>	<i>Total installed length (km)</i>	<i>As of (year)</i>
United States	1,984,321	2013
Russia	177,700	2013
Canada	74,980	2007
China	70,000	2015
Ukraine	36,720	2013
Australia	30,054	2013
Argentina	29,930	2013
United Kingdom	28,603	2013
Germany	26,985	2013
Iran	20,794	2013
<i>Petroleum and refined petroleum products</i>		
<i>Country</i>	<i>Total installed length (km)</i>	<i>As of (year)</i>
United States	240,711	2013
Russia	74,100	2013
China	48,400	2015
Canada	23,564	2007
India	20,012	2013
Iran	16,562	2013
Mexico	16,340	2013
Kazakhstan	12,408	2013
Venezuela	10,347	2013
Colombia	10,225	2013

Source: US CIA World Factbook 2017.

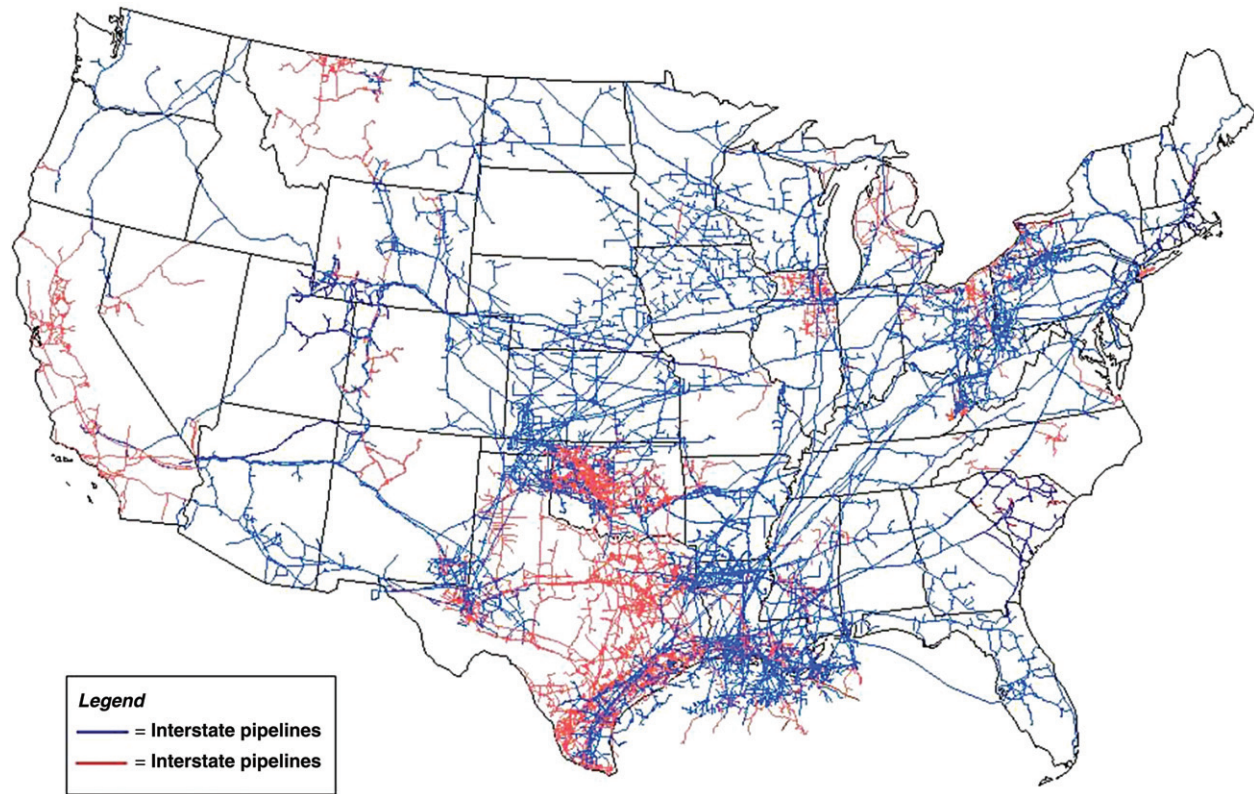


Figure 2 US natural gas pipeline network. Source: US Energy Information Administration.

operated by over 200 independent firms, including over 1,400 compressor stations (not shown) and around 400 underground storage facilities (Fig. 3). Nearly 70% of these long-distance gas pipelines are classified as interstate pipelines, and are therefore regulated federally (Oliver and Mason, 2018).

Basic Engineering

The engineering aspects of oil and gas pipelines are quite similar (Meisner and Leffler, 2006). Both oil and gas are fluids. A fundamental characteristic of a fluid is that it conforms to the shape of the vessel in which it is contained, be it a storage unit or a pipeline. The key differences between oil and gas are that (1) gas is considerably less dense and more compressible than oil; and (2) a liquid settles by force of gravity, whereas a gas fills a volume. This means the parameters that govern the flow of either fuel through a pipe are quite different, despite the underlying physical process being largely the same. The basic operational principle of a pipeline, in any case, is to pump fluid into one end of the pipe at pressure, which forces the fluid down the length of the pipe, from which it is extracted at the other end. In a gas pipeline, a compressor changes the pressure and gas density, generating flow. In a liquid pipeline, pumps increase the fluid's pressure while the density is not affected. Considerable energy may be required to generate the pressure needed to push a fluid through the length of a pipeline, and the longer the line, the greater the energy required, although pumping requires significantly less energy by volume than compression. Pressure is greatest near the pumping or compressor station, and fluid friction losses in the pipe cause pressure to decline as the fluid travels along the line—referred to as the *pressure drop* (Fig. 4). Particularly long pipelines often require multiple, strategically placed pumping or compressor stations to ensure adequate pressure along the full length of the line. Pressure—and the energy required to generate it—may also be affected by changes in temperature (for gases) or elevation (for liquids).

The most important feature of any pipeline is its *capacity*. A pipeline's capacity is defined as the maximum level of service, measured in energy or volumetric units (e.g., MMBtu of gas or barrels of oil) per unit time (typically expressed in daily increments), that it can maintain over an extended period, and is subject to various technological, safety, and regulatory constraints.^a The most important factors in determining a pipeline's capacity are its (inner) diameter and operating pressure. Diameter, of course, refers to the circular width of the steel tube used to form the pipe, which is fixed upon construction. Pressure, by contrast, is variable along the

^a From an economic or regulatory perspective, capacity typically does *not* refer to the maximum physical throughput capability of the system or segment on any given day, which can greatly exceed the pipeline's certificated capacity.

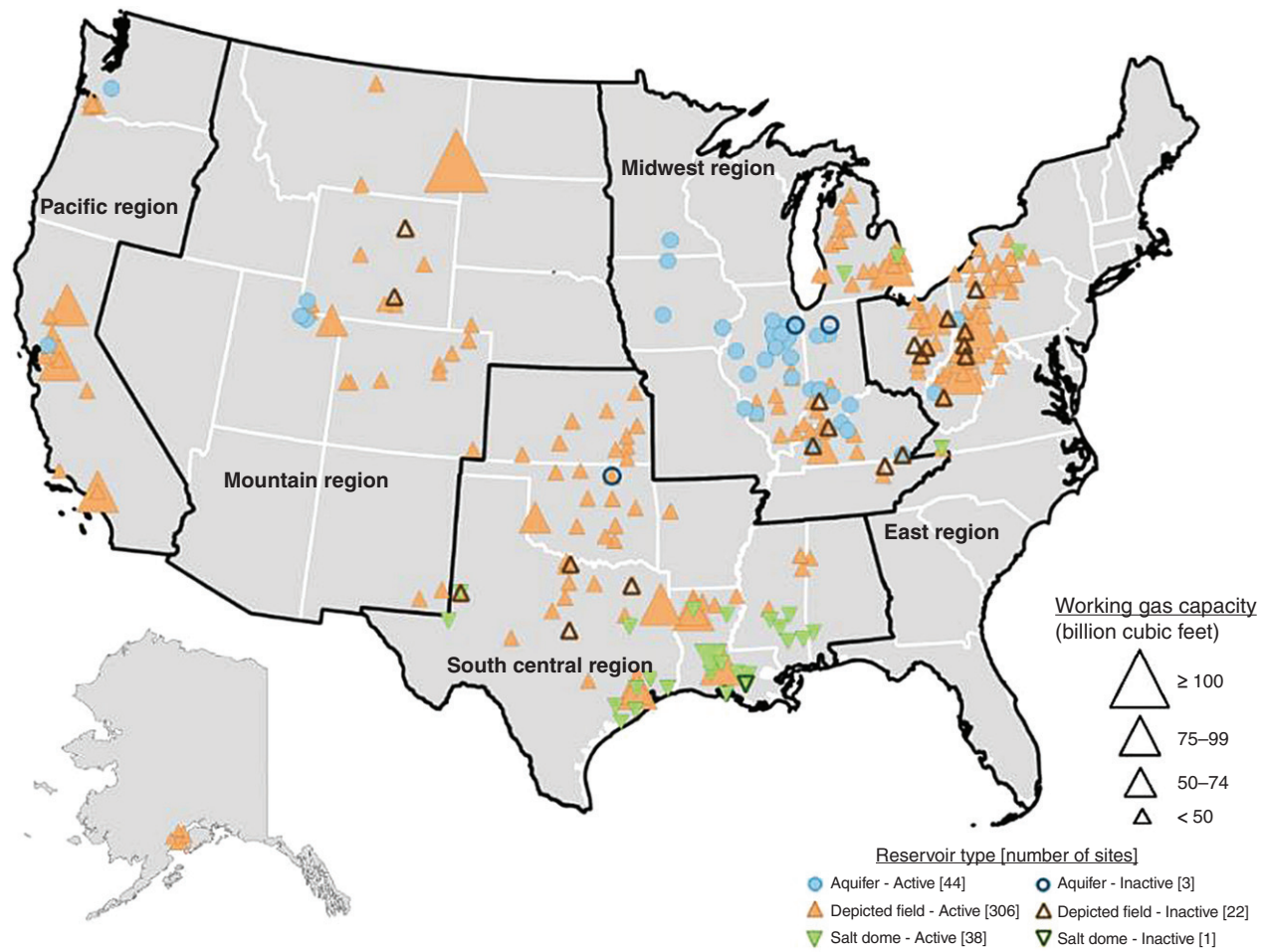


Figure 3 US underground gas storage facilities. Source: US Energy Information Administration.

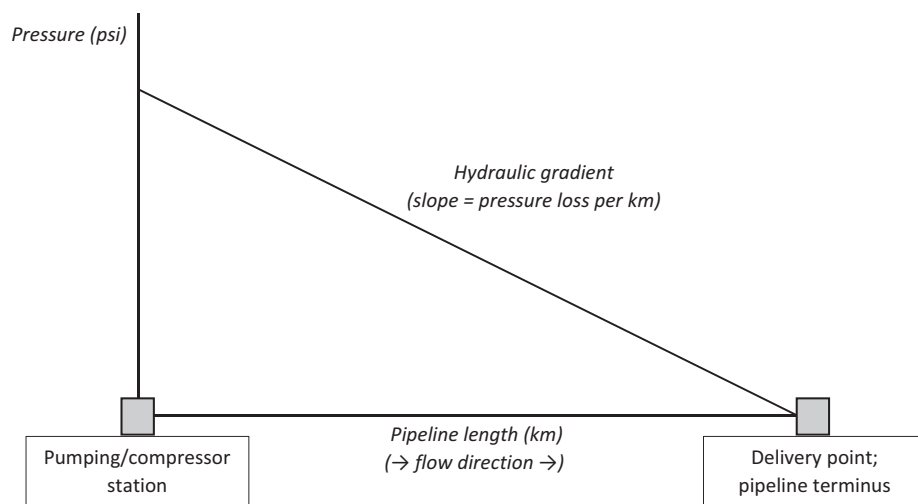


Figure 4 Stylized graph of the pressure drop along a pipeline segment.

length of the pipeline and depends on a number of factors such as pipe pressure rating, corrosion, and other safety considerations. Adding a pump or compressor increases operating pressure. A pipeline's *maximum allowable operating pressure* (MAOP), as the term implies, is the highest pressure at which the pipeline is allowed to operate, as determined by the jurisdictional regulator. The MAOP may change both along the pipeline and over time depending on its initial design characteristics (e.g., pipe wall thickness) and subsequent safety inspections. In the United States, for example, the MAOP for gas pipelines is determined by US Department of Transportation safety regulations, and is set considerably lower than its maximum design pressure. For safety reasons, the US Department of Transportation requires the MAOP to be lower in more densely populated areas (Oliver, 2015).

For a pipeline of a given diameter operating at a given pressure, the hydraulic properties of the fluid being transported also greatly affect flow rates. Such properties include density, viscosity, and vapor pressure (Meisner and Leffler, 2006). Oil, gas, and other hydrocarbon fuels vary significantly in these and other hydraulic properties. A fluid's density, viscosity, and other fluid properties can be calculated from an *equations-of-state* empirical model relating key fluid properties. *Density* refers to mass per unit volume, which depends on temperature, pressure, and fluid composition (or compressibility). A fluid's density is typically measured by pipeline operators in terms of *specific gravity*, defined as the ratio of the fluid's density to the density of a reference material (e.g., air for gas, water for liquid, etc.).^b *Viscosity* is defined as a fluid's internal resistance to flow (i.e., to pipe-wall friction); the higher a fluid's viscosity, the more energy is required to push it through a pipe. Viscosity is partly related to density, in that more dense fluids tend to be more viscous. The *vapor pressure* of a liquid is defined as the pressure exerted by a vapor in thermodynamic equilibrium with its condensed (i.e., liquid) form at a given temperature in a closed system. Vapor pressure and *dew point* are important for both gas and oil pipeline operations. For gas pipelines operating at a given temperature, when pressure and temperature drop below the dew point of (naturally occurring) heavier molecules such as ethane, butane, or water, they turn to liquids, which can block flows and damage compressors or other equipment. For oil and petroleum products pipelines, if the operating pressure drops below the vapor pressure of the particular fluid, this allows bubbles of vapor to form, which, similarly, can be damaging to pumping stations, meters, and other equipment.

The flow rate of a pipeline segment transporting a compressible gas can be calculated as a function of its diameter and length, the injection and withdrawal pressures, and several other parameters including the gas's specific gravity and compressibility factor. Mokhtahab et al. (2015) provide the following simplified form for the general flow equation over a pipeline segment:

$$Q = C \left(\frac{T_b}{P_b} \right) D^{2.5} \left(\frac{P_1^2 - P_2^2}{f \gamma_g T_a Z_a L} \right)^{0.5} E,$$

where Q is the standard flow rate measured in ft^3/day ; C is a constant that is adjusted for the volumetric unit used (77.54 for ft^3); T_b is base temperature in R ; D is the pipeline's internal diameter in inches; P_b is the pressure base in psia; P_1 is initial pressure in the pipe in psia; P_2 is terminal pressure in the pipe in psia; f is Moody's friction factor; γ_g is gas specific gravity; T_a is the mean flowing temperature of the pipeline in R ; Z_a is the average compressibility factor, L is the pipeline's length in miles; and E is a flow efficiency factor. In practice, specialized software tools are used to determine flow rates based on an advanced computational calculation of the complex pipeline system with all its boundary conditions.

The flow of an incompressible liquid such as oil or water is simpler, but still depends on a host of pipeline features and fluid characteristics, requiring sophisticated software tools to calculate local flow rates from complex pipeline models and the transient boundary conditions. First, the degree of *turbulence* of the flow within the pipe is determined by the *Reynolds number*:

$$\Re = \frac{\rho V D}{\mu},$$

where ρ is the fluid's density; μ is the fluid's dynamic viscosity; V is the cross-sectional mean velocity of flow inside the pipe; and D is the pipe's inner diameter. A more turbulent flow has a higher Reynolds number. Computing the flow rate requires first computing the average point velocity along the length of the pipe, $\bar{u}$, which depends on both pressure and the degree of turbulence. The flow rate can then be calculated as:

$$Q = 2\pi \int_0^a r \bar{u}(r) dr,$$

where a is the pipe's radius. In short, the flow volume rate equals velocity multiplied by the pipe's cross-sectional area. See Liu (2003) for a complete exposition.

Basic Economics

Pipelines exhibit *economies of scale*. This means that, all else being equal, a larger pipeline can transport fuel over a fixed geographic distance at a lower average cost than a smaller pipeline. These cost economies emerge over two margins (Oliver, 2015). The first is

^b For example, natural gas has a specific gravity ranging from 0.55 to 0.7, or 55%–70% of the mass of the same volume of air, depending on temperature. Crude oil at 60°F (15.56°C) has a specific gravity of roughly 0.9, or 90% of the mass of the same volume of water.

the fact that a pipeline's capacity—and therefore its throughput—can be increased for a less-than-proportional increase in *fixed cost*. Fixed costs are primarily those related to pipeline capital investment (e.g., construction costs), but also include other costs such as those related to regulatory compliance or obtaining right-of-way. Empirical research has confirmed, for example, that gas pipelines exhibit short-run economies of scale in production, as throughput can be increased at a less-than-proportional increase in pipe and compression capital costs. The second margin is that the *variable cost* of shipping a unit of fuel typically decreases as the capacity of the pipeline increases (Yépez, 2008). Economies of scale in production are important, as they generate productivity growth over time.

Pipelines also exhibit *economies of scope*. Economies of scope exist in a production process when the unit costs of producing two or more different goods or services in combination are lower than would occur under separate production processes. In the pipeline network setting, point-to-point transmission between different geographic locations on the network is analogous to the production of different goods or services in a traditional economies-of-scope model. Thus, scope economies exist when cost savings emerge as deliverability is expanded spatially to serve a greater number of geographically dispersed customers. This is easily illustrated by a simple but intuitive example of three network nodes—A, B, and C—spatially distributed along a roughly linear geographic arc, such that B lies between A and C. Transmissions from A to B, from A to C, and from B to C comprise three distinct services. Clearly, it is less costly to build a single transmission line from A to B to C that can provide all three services simultaneously than to build three separate lines connecting A to B, A to C, and B to C. In this sense, economies of scope exist in expanding a pipeline network spatially.

As a result of economies of scale (and, to a lesser extent, economies of scope), pipelines typically have significant market power, and in many (but not all) cases are considered to be local natural monopolies over their respective routes. A natural monopoly is defined in economics as a situation in which a single firm or supplier can satisfy the entirety of market demand at the lowest cost. This is true of many pipeline transmission routes. As a result, pipelines are typically either state-owned monopolies, as is the case for many European gas pipelines, or, if privately owned, are subject to price regulation, as is the case with US oil and gas pipelines. In many cases, pipelines are also subject to *common carriage* obligations, which prohibit discrimination against customers in terms of either prices or access. A notable exception is US gas pipelines, which are categorized by regulators as private carriers.

A fundamental aspect of the economic relationships pertaining to pipeline investment and transmission is that of *asset specificity* (Makholm, 2012). A pipeline is a capital investment whose economic value is specific to a single type of transaction—for example, transporting fuel from one geographic point to another—but which is otherwise essentially valueless. Asset specificity often leads to what economists have referred to as the “holdup” problem. In complex economic relationships with transaction-specific assets and incomplete contracts, a prior commitment made by one party may increase the rent-extraction capability of another party at subsequent stages. In the pipeline context, consider an example in which a gas producer sells gas to a pipeline, which then ships the gas and sells it to a single retail distributor at a bundled delivery price that covers both transmission and the value of the gas commodity itself.^c If the gas retailer has no other means of procuring gas, then in the absence of either regulation of, or a binding contractual agreement over, the price paid for transmission and resale of gas, the pipeline might be able to “hold up” the retailer by raising the bundled price of gas after the pipeline has been built.

Because of the asset specificity problem, the risk of holdup means the success of a pipeline project depends on cooperation between upstream suppliers (e.g., oil and gas producers), the pipeline, and downstream buyers (e.g., refineries, retail gas distributors, etc.). Once investment in the transaction-specific pipeline asset is sunk, investors face the risk that a party occupying a different position in the value chain might engage in “holdup” style rent-seeking behavior. This creates a considerable barrier to investment, the only two remedies for which are (1) long-term contracts between the pipeline and its upstream and/or downstream stakeholders regarding future revenue streams, or (2) vertical integration. Many pipeline systems around the world are vertically integrated joint ventures owned by oil and gas companies (as is the case for US crude oil pipelines, e.g.).

The US interstate gas pipeline network stands alone as the most economically competitive pipeline system in the world, largely due to the regulations enforced by the US Federal Energy Regulatory Commission (Makholm, 2012). Gas pipelines in the United States are investor-owned, privately operated facilities. Access to segments of point-to-point transmission capacity is allocated via long-term contracts between the pipeline companies and their “firm” customers. These contracts grant firm customers guaranteed access to a specified portion of a pipeline's capacity at a regulated price. Capacity rights have become a fully transferable, freely traded commodity in a largely competitive market for gas shipping services. This enables firm pipeline customers to either utilize their contracted capacity to ship gas or to sell the rights to its use to other shippers at an unregulated price in a secondary “capacity release” market.

Another key consideration in pipeline economics—particularly on a pipeline network that allows competitive point-to-point transmission like the US gas pipeline network—is the effect of congestion on market prices in the commodity market served by the pipeline infrastructure (Oliver et al., 2014). Importantly, at any given point in time a pipeline's capacity is fixed—that is, it cannot be increased incrementally in response to an increase in transmission demand. Congestion occurs when the demand for shipping services over a given pipeline route exceeds the capacity of the pipeline. When available space over the pipeline route is plentiful then, in the absence of any other idiosyncratic service interruptions, the difference in the commodity price between the points of receipt and delivery should equal the marginal cost of transmission. When available space becomes scarce and the pipeline is congested, competition for capacity drives the market price of transmission above the marginal cost, allowing owners of the rights to

^cThis was the standard industry practice for US gas pipelines throughout most of the 20th century, until the market was restructured by regulators in the 1990s.

that capacity to capture potentially large congestion rents. Such congestion rents depress the commodity price at the point of purchase (e.g., the wellhead) and drive up the commodity price at the point of sale (e.g., the city gate), which has salient consequences for economic surplus in the commodity market served by the pipeline.

Additional Considerations

Pipeline Safety and Integrity

Pipelines are statistically quite safe and reliable. Spills, leaks, and other accidents are comparatively rare relative to other means of fuel transport. However, they do occur, and major incidents can be extremely dangerous and/or environmentally damaging. Significant failures tend to make national or even global headlines, and the risk of such failures lead many to oppose pipeline construction through or near their communities.

There are several reasons a pipeline might fail. Remarkably, as [Kandiyoti \(2012\)](#) reports, over a third of all pipeline incidents are caused by *third-party* action, either negligently or intentionally. For example, a pipeline may be struck accidentally by an excavator. Or, in countries beset by violence and political instability, pipelines are routinely targeted by oppositional groups seeking to interrupt operations as a means of affecting some sociopolitical agenda. Operational reasons for pipeline failure include pipe corrosion, material failures (e.g., faulty welds), and mechanical failures. In yet other cases, damage may be caused by natural phenomena like earthquakes or hurricanes.

Both private industry groups and public regulatory agencies take part in determining operational safety rules and standards. Such industry groups include the Association of Oil Pipelines (AOPL) and the Interstate Natural Gas Association of America in the United States and the Conservation of Clean Air and Water in Europe (CONCAWE) in the European Union. Federal regulation of US pipeline operations is administered by the US Department of Transportation, Office of Pipeline Safety. No such body exists in the EU as a whole; operational safety regulations in Europe are administered at the national level. As a general rule, regulatory bodies require pipeline operators to maintain detailed *pipeline integrity management systems* (PIMS) and *pipeline safety management systems* (PSMS) that provide a comprehensive, integrated framework for pipeline asset management. Such operational standards minimize the risk of minor leaks and other, more significant failures, ensuring safe and reliable pipeline operations within the bounds of public tolerance ([Revie, 2015](#)).

Pipelines and Geopolitics

Because oil has been a lynchpin of global geopolitics since World War II, pipelines themselves have regularly been the focus of geopolitical tensions, particularly when they cross national boundaries. Nowhere was this more apparent than in the Middle East during the post-World War II era of the 20th century. Western developed economies—primarily the United States and Great Britain—were ravenous for oil supplies, and oil-rich Middle Eastern countries (e.g., Iraq, Iran, Saudi Arabia, etc.) were poised to deliver it. The geographic locations of many Middle East oil fields dictated that vast transport cost savings could be gained by building pipelines to deliver oil to tanker terminals along the Mediterranean. The alternative was to load tankers in the Persian Gulf, which were then forced either around the Arabian peninsula and through the Suez Canal or, even worse, around the southern cape of Africa. To reach the Mediterranean, however, such pipelines required passage through countries such as Syria, Jordan, Turkey, and Israel. Geopolitical tensions flared over a variety of issues including proposed pipeline routes, transit fees, and fuel prices, as well as over risks of service interruptions, sectarian violence, and even outright political blackmail ([Kandiyoti, 2012](#)).

More recently, oil and gas pipelines have taken center stage in Eurasian and south Asian geopolitics. After the dissolution of the Soviet Union in the 1990s, the Russian Federation found itself in a compromising position with respect to its own oil and gas exports to Europe, as major pipelines crossing former Soviet republics (e.g., Ukraine, Belarus, Kazakhstan, and the Baltic states) were claimed and nationalized by the newly independent nation-states. This has led to significant geopolitical wrangling between fuel-hungry European countries, Russia, and the intermediary nations over new pipeline projects and transit fees on existing lines. In South Asia, India has long sought to increase its gas imports by proposing pipelines from both the east (from Myanmar through Bangladesh) and the west (from Iran through Pakistan). Neither has yet been successfully implemented (or even agreed upon by the intermediary nations and other interested parties) due to seemingly intractable geopolitical disagreements ([Kandiyoti, 2012](#)). Anytime a major pipeline project requires a border crossing, it forces the international stakeholders to come to terms about a number of important factors—a difficult proposition when such neighboring countries already experience tension with regard to other, sometimes long standing, geopolitical issues.

Acknowledgment

I thank Klaus Brun of the *Elliot Group* for a number of helpful suggestions and clarifications regarding the section on pipeline engineering.

References

- American Petroleum Institute, 2019. Pipeline 101. Available from: <https://pipeline101.org/>.
- Kandiyoti, R., 2012. *Pipelines: Flowing Oil and Crude Politics*. I.B. Tauris, London.
- Liu, H., 2003. *Pipeline Engineering*. Lewis Publishers, Boca Raton, FL.
- Makholm, J.D., 2012. *The Political Economy of Pipelines*. The University of Chicago Press, Chicago, IL.
- Meisner, T.O., Leffler, W.L., 2006. *Oil & Gas Pipelines in Nontechnical Language*. PennWell Corp, Tulsa, OK.
- Mokhtab, S., Poe, W.A., Mak, J.Y., 2015. *Handbook of Natural Gas Transmission and Processing: Principles and Practices*, third ed. Gulf Professional Publishing, Oxford, UK.
- Oliver, M.E., 2015. Economies of scale and scope in expansion of the US natural gas pipeline network. *Energy Econ.* 52 (Part B), 265–276.
- Oliver, M.E., Mason, C.F., 2018. Natural gas pipeline regulation in the United States: past, present, and future. *Found. Trends Microeconomics* 11 (4), 227–288.
- Oliver, M.E., Mason, C.F., Finnoff, D., 2014. Pipeline congestion and basis differentials. *J. Regul. Econ.* 46 (3), 261–291.
- Revie, R.W., 2015. *Oil and Gas Pipelines: Safety and Integrity Handbook*. John Wiley & Sons Inc., Hoboken, NJ.
- Yépez, R.A., 2008. A cost function for the natural gas transmission industry. *Eng. Econ.* 53 (1), 68–83.

3D Printers and Transport

Wouter P.C. Boon*, **Bert van Wee†**, *Copernicus Institute of Sustainable Development, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands; †Transport and Logistics Group, Faculty Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

The Emergence of 3D Printing	471
The Concept of 3D Printing	471
Influence of 3D Printing on Transport and Logistics	472
Wants and Needs of Consumers and of Producing Companies	473
Location of Production, Clients and Distribution	474
Transport Resistance	474
Transport Volumes and Technologies	475
Societal Impact: Accessibility, Safety, and the Environment	475
Discussion	476
Concluding Remarks	476
Biographies	477
References	477

The Emergence of 3D Printing

3D printing (3DP) is a set of technologies that allows users to create products using digital models and an additive, layer-upon-layer manufacturing process. This new form of manufacturing is a marked shift away from shaping products from raw material as desired or from subtractive manufacturing methodologies, that is, constructing objects by cutting material away from a solid block of material. Trend watchers had already claimed that 3DP would lead to the next industrial revolution ([Economist, 2016](#); [Rifkin, 2011](#)). Initial growth figures (2014–2016) did not match these expectations, resulting in disappointment ([Deloitte, 2019](#)). The last couple of years saw an increase in growth because of more diverse 3D printable materials (moving from plastics to metals), higher printing speed, and a gain in printing volumes. Short-term expectations already envision that the worldwide 3D industry will grow steeply, from \$7.3 billion in 2017 to a forecasted \$15.8 billion in 2020, and \$35.6 billion by the year 2024. The growth is expected to come from more industrial 3DP systems, rather than desktop printers being sold ([Wohlers Report, 2019](#)).

3DP has been around for some time in the form of printed dental products, and has been adopted in other sectors that value production in small batches like hearing aids, and tailor-made jewelry and apparel ([Rogers et al., 2016](#)). In the context of product development in the automotive and aerospace industries, additive manufacturing has been used in the form of “rapid prototyping,” that is, to make prototypes in a relatively fast way ([Bradshaw et al. 2010](#)).

Two recent developments are changing the scene of 3DP. First, the quality of additive manufacturing is increasing rapidly. Traditionally, 3DP was associated with lower quality, related to issues like geometric repeatability (degree to which a digital image can be reproduced), surface quality, and fatigue resistance ([Huang et al., 2016](#)). Scientific and technological advances are currently overcoming these issues ([Gausemeier, 2014](#); [Ngo et al., 2018](#)). Second, the introduction of relatively cheap desktop 3D printers, such as RepRap, MakerBot, or Ultimaker, has made it possible for consumers to produce personalized home products such as jewelry and bicycle parts ([De Jong and De Bruijn, 2013](#)). With this, 3DP is moving from factory-based rapid prototyping to home fabrication ([Rayna and Striukova, 2016](#)).

3DP is thus associated with a more tailor-made, distributed way of manufacturing which, if successful, is bound to have repercussions on transport, supply chains, and logistics. But what is 3DP, how will it influence consumer wants and needs, location choices and transport resistance, and how will 3DP impact society? What are the repercussions for the transport and logistics sector? What is the state of knowledge, and which challenges are looming due to the emergence of 3DP? This article aims to answer these questions, of course taking into account the fact that the emergence of 3DP and its consequences are subject to a high degree of uncertainty.

The Concept of 3D Printing

The distinctive element of 3DP is building products layer by layer, rather than shaping from material (through forging, casting, stamping and molding) or subtracting from a block of material (through carving). 3DP should be understood as an aggregate of three elements: (1) the additive process technique, (2) the design and/or scanning software, and (3) the materials used ([Fig. 1](#)). The emergence of 3DP can be understood as an innovation pathway in which new developments are happening on all of these three elements, as well as in the combinations of these elements ([Robinson et al., 2019](#)).

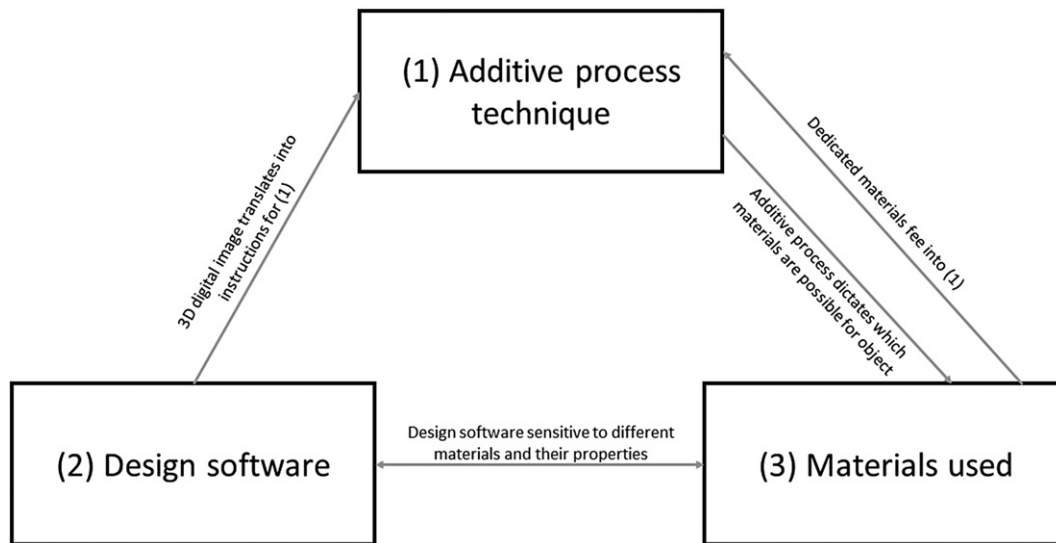


Figure 1 The three main elements of 3D printing. Source: Adopted from Robinson et al. (2019)

The first working 3D printer built in 1986 by Charles Hull was based on stereolithography. Since then, various additive process techniques [(1) in Fig. 1] have been introduced (Table 1). The movements of the printers, as well as the size and form of the successive layers that need to be printed to fabricate the 3D object, are dictated by 3D modeling software, often in the form of computer aided design (Griffiths, 2002).

Influence of 3D Printing on Transport and Logistics

The influence of 3DP extends from manufacturing to include changes in logistics and transport. The recent cost reduction and availability of easier-to-use printers have led to expanding the customer base to also include households, schools, libraries, and so on. Such an expansion might result in a decentralization of production (Holmström et al., 2010). The transition from centralized to decentralized manufacturing has repercussions for transport flows (Ford and Despeisse, 2016) and related supply chain management and business models of involved companies (Bogers et al., 2016).

To better understand the potential influence of 3DP on logistics, a literature review was conducted of 93 scientific articles and 12 reports of leading consultancies. Based on the review and a survey amongst experts on 3DP and logistics, a model was created that

Table 1 Summary of different 3D printing techniques

Technique	Short description	Application areas	Introduced since
Stereolithography	Photochemical process of UV laser causing monomers to link together to form polymers	Biomedical Prototyping	1986
Fused deposition modeling	Thermoplastic material filament fed through a moving printer extruder head, the shape of which is digitally controlled to define the printed object	Rapid prototyping Toys Advanced composite parts	1988
Powder bed fusion	Fusing thin layers of powder with use of lasers	Biomedical Electronics Aerospace Lightweight structures (lattices) Heat exchangers	1993
Inkjet printing and contour crafting	Powder pumped and deposited in the form of droplets via a nozzle to form a continuous pattern	Biomedical Large structures Buildings	1993/2011
Direct energy deposition	Energy source like lasers is focused on a substrate simultaneously material	Aerospace Retrofitting Repair Cladding Biomedical	1995
Laminated object manufacturing	Adhesive-coated material (paper, plastic) is glued together and cut into shape	Paper manufacturing Foundry industries Electronics Smart structures	1996

Source: Based on Ngo et al. (2018)

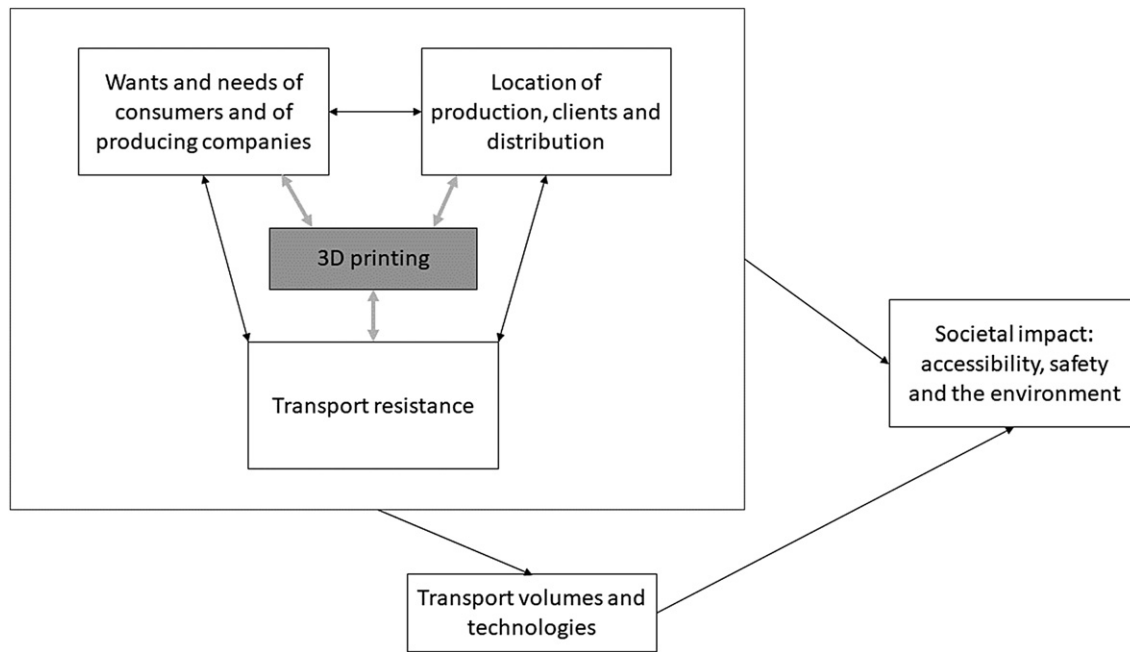


Figure 2 Conceptual model for the impact of 3D printing on logistics and eventually on accessibility, the environment and safety. Source: Adapted from Boon and van Wee (2018)

conceptualizes the influence of 3DP on logistics, and subsequently on societal impacts in terms of accessibility, traffic safety and sustainability (Boon and van Wee, 2018). The model is shown in Fig. 2.

Each element of the model can be elaborated upon, first zooming in on the influence of 3DP on wants and needs, location decisions and transport resistance. Subsequently, transportation volumes and technologies are addressed and, finally, the way in which accessibility, sustainability and safety are influenced.

Wants and Needs of Consumers and of Producing Companies

3DP was originally used to produce one-offs in professional settings like the abovementioned rapid prototyping in aerospace and car industries, or creating personalized models of skulls that can be used for preparing for complex surgery. Increasingly, 3DP is utilized to cover both prototypes and finalized products. The latter include examples like in-ear hearing aids and dental parts, as well as small consumer products, like jewelry and toys. Close to adoption are applications like spare parts in the aircraft and surgical industries. Consumer products are especially restricted in their impact and some authors claim that 3DP will remain having a limited effect (Fawcett and Waller, 2014; Gress and Kalafsky, 2015). More complex products remain comparatively rare, because of questionable efficiency and added-value to consumers (Mckinnon, 2016).

Literature and the experts who were surveyed were in agreement about the fact that mass-customization will become more important. There are different points at which the customer may become engaged in the supply chain, at each stage of the supply chain: design, purchasing, fabrication, assembling and distribution (Gosling et al., 2007). 3DP plays a significant part in customer-specific design and manufacturing, since it has the possibility to pursue product customization and a fast response to customer needs, as a result of flexibility and adaptability (Bhattacharjya et al., 2014; Chan and Chan, 2010; Christopher and Ryals, 2014; Janssen et al., 2014; Lopes da Silva and Vicente, 2013). 3D-printed prototypes can help increase user understanding early in an innovation process, decreasing time-to-market and risks of non-adoption (Kanto et al., 2014; Romero and Vieira, 2016). Companies might respond by making their supply chain and related production capacities more flexible and agile (Sasson and Johnson, 2016).

A related development is the rise of 3D platforms that offer, amongst other things, free-to-use design files (www.thingiverse.com) or offer a place to share 3DP capacity (www.3dhubs.com). Such platforms can increase the possibilities of personalization and the ability to deliver to the “long-tail” of consumer needs (Bogers et al., 2016). Moreover, these platforms are often two-sided, that is, consumers both use the presented products (e.g. designs) as well as offer their own creations. As such, 3DP ties into the development of consumers becoming producers, acting as “prosumers” (Chen et al., 2015; Rauch et al., 2016).

The experts interviewed (Boon and van Wee, 2018) were divided on the issue as to what extent consumers are willing to really manufacture a wide range of products at home, or whether new needs for different kinds of products will emerge. Also, opinions about the frequency of buying were split: some experts say that frequency will increase, because customized products follow trends

and are disposed of more quickly; additions are replaced more easily. Others claim frequency will decline, because customized products do not need to be replaced and if broken, small parts can be reprinted.

Location of Production, Clients and Distribution

3DP advances mass customization and personalization (Bhattacharjya et al., 2014; Mourtzis and Doukas, 2013). The question central to many articles and trend reports is to what extent 3DP leads to distributed production. Two extreme scenarios can be sketched: decentralized versus centralized manufacturing (Li et al., 2017). Most publications envisage that 3DP will result in more decentralized manufacturing, offering greater proximity to customers and responsiveness to market needs (Fornasiero et al., 2010; Manners-Bell and Lyon, 2012), possibilities to cut out middlemen (Jia et al., 2016), and even the reshoring of manufacturing to high-income countries (Campbell et al., 2011). Designs travel digitally (Janssen et al., 2014) and as such can replace the movement of products (Garrett, 2014).

In the decentralization scenario, the locations of 3D printers will vary, ranging from “print shops” serving markets at a continental scale, city- or neighborhood-level hubs to printers located at consumers’ homes (Rauch et al., 2016; United States Postal Service Office of Inspector General, 2014). Ryan et al. (2017) produced a different, yet related taxonomy of decentralized 3DP scenarios: mobile, in-transit 3DP; 3DP for standardized spare parts; local and regional factories; craft businesses, like “fablabs,” and personal manufacturing.

An in-between scenario is that 3D “mini-factories” are located at local service providers such as libraries, community centers and post offices, as such mimicking 2D printing of photos and books. It is also expected that 3D logistics service providers will enter the additive manufacturing market within the next ten years (Durach et al., 2017). Hub-level operations have the advantage over home-based printing, because 3D printer (machine) operation, post-processing, quality control and (consumer) packaging require investments and expertise that home-print consumers are not going to develop. Much is expected from the product customization, agility and even complexity that 3DP hubs can yield, which are associated with specialization, lower time-to-market and even lower costs (Rogers et al., 2016).

Some industry watchers still emphasize the importance of centralized production. High-volume, low-added-value production will remain more efficient in centralized factories (Holmström et al., 2010; Li et al., 2017). Also, low-volume and high-added-value manufacturing require support infrastructure and specialized human resources (Gress and Kalafsky, 2015). The degree of centralization involves a complex interplay of economies of scale, as well as user requirements, regarding flexibility, adaptability, demand uncertainty, supply uncertainty, and capacity utilization. For example, in spare part management in the field of aircraft maintenance there are parts that are not used very often, “slow-moving parts,” yet when required they are needed fast. In these cases, 3DP might provide a solution (Fawcett and Waller, 2014; Holmström et al., 2010; Huang et al., 2013).

In the end it may very well be that the two “extreme” scenarios are combined and are deemed complementary. A hybrid scenario could be that more advanced 3DP will (mainly) be done at centers, and that simpler printing will be done at home. For such hybrid scenarios to be pursued, it is important to think about communication, information exchange, intellectual property rights, and quality control between centralized and decentralized entities. Advancing digital innovation may facilitate managing this (Durão et al., 2017). Companies will vary in the business models they adopt, but experts expect the hub model to be favored over competing approaches and to emerge as the dominant business model (Boon and van Wee, 2018).

Transport Resistance

The location of 3DP dictates the nature and size of transport flows. 3DP taking place closer to home is related to a decrease in distribution between manufacturing and end users (Janssen et al., 2014; Tuck et al., 2007). Even shipping and air cargo volumes might come down (Barz et al., 2016; Manners-Bell and Lyon, 2012). As input to decentralized 3DP, shipments of raw materials will become more important. This would mean a shift from manufacturer–consumer distribution to the raw material supplier–manufacturer side (Gress and Kalafsky, 2015; United States Postal Service Office of Inspector General, 2014). Moreover, the transport of raw materials can be more efficient, since bulk delivery means less packaging and urgency, leading to a reduction of ton-kms and cubic meter-kms of freight movements (Mckinnon, 2016). Distribution networks will be organized more efficiently: with more nodes and more flexibility. Transport and logistics costs may come down as production happens closer to home and feedstock is transported more efficiently in the form of cartridges, leading to less air to be transported on the outbound-shipping part of the chain. Relatedly, 3DP has an impact on warehousing. Less inventory might result from less variety in input materials needed, less safety-stock, and consumers being dedicated in their needs (Mavri, 2015).

Still, some experts think that transportation costs may actually increase (Boon and van Wee, 2018; Durach et al., 2017). Raw materials will need to be shipped to 3D printers located at “mini-factories” or hubs. Then the distribution between hubs and consumers’ homes might spur short-distance “last mile” express delivery that is associated with higher flexibility in distribution and higher costs (Mohr and Khan, 2015). Even so, different types of costs can also balance each other out: for example, Westerweel et al. (2018) found that the higher design costs related to 3DP are offset by reduced after-sales logistical cost and lead-times.

Transport Volumes and Technologies

3DP might lead to the transportation of more raw materials, for example in the form of gels and powders contained in packaged cartridges. This, then, means a shift away from shipping ready-made consumer goods, the transport of bulk inputs for 3D printers will become more important and might mean less need for the use of containers. On the other hand, experts say that other emerging technologies, such as electrification and autonomous driving, are far more influential on how transport technologies of the future will look. An interesting, left-field idea is to have 3DP manufacturing onboard a driving truck, which opens up the possibility of creating products just-in-time for delivery (Ryan et al., 2017).

In terms of transported volume, there are reasons to assume that there will be a decrease because of extended product-life-in-use, less air to be transported, and a reduced number of kilometers in the last mile. There are also reasons for an increase in transport volumes as there is more material that needs to be cut out, and subsequently returned. The timing of deliveries to consumers might change. On the one hand, there may be more deliveries outside office hours, whereas on the other, brand new ways of dispatching can emerge, such as smart letterboxes.

Societal Impact: Accessibility, Safety, and the Environment

The literature has been silent about the effects of 3DP on traffic safety, accessibility, infrastructure needs, and the indirect effects on passenger transport. Concerning welfare, income, and employment, 3DP may be regarded as having a potential disruptive impact on the global transport sector (Bogers et al., 2016). The total system costs may not drop because labor and transport costs are communicating vessels (Chan and Chan, 2010; Manners-Bell and Lyon, 2012). For example, 3DP might lead to “reshoring” of manufacturing to high-income countries, especially as 3D printer hubs are to be located closer to customers (Frazier, 2014; Janssen et al., 2014). At the same time, this would require the creation of highly skilled jobs in these hubs (PWC, 2014). Traditional outsourcing countries may not be disadvantaged by reshoring, as experts claim that they are increasingly becoming capable to become frontrunners in 3DP as well (Gress and Kalafsky, 2015).

In terms of the environmental impacts of 3DP, there are model-based studies that hone in on material usage and waste (Campbell et al., 2011; Huang et al., 2016; Romero and Vieira, 2016; Barz et al., 2016; Garrett, 2014; Li et al., 2017; Lopes da Silva and Vicente, 2013). Less is written about logistics and transport-related environmental impact. Some publications make tentative predictions that emissions and oil use will be reduced (Bhattacharjya et al., 2014; Ford and Despeisse, 2016). Proposed hypotheses are that due to production closer to home, long-distance transport can be reduced and distribution of raw materials can be made more efficient. Last-mile movements might increase but if they are done by low-emission vehicles, then this means less

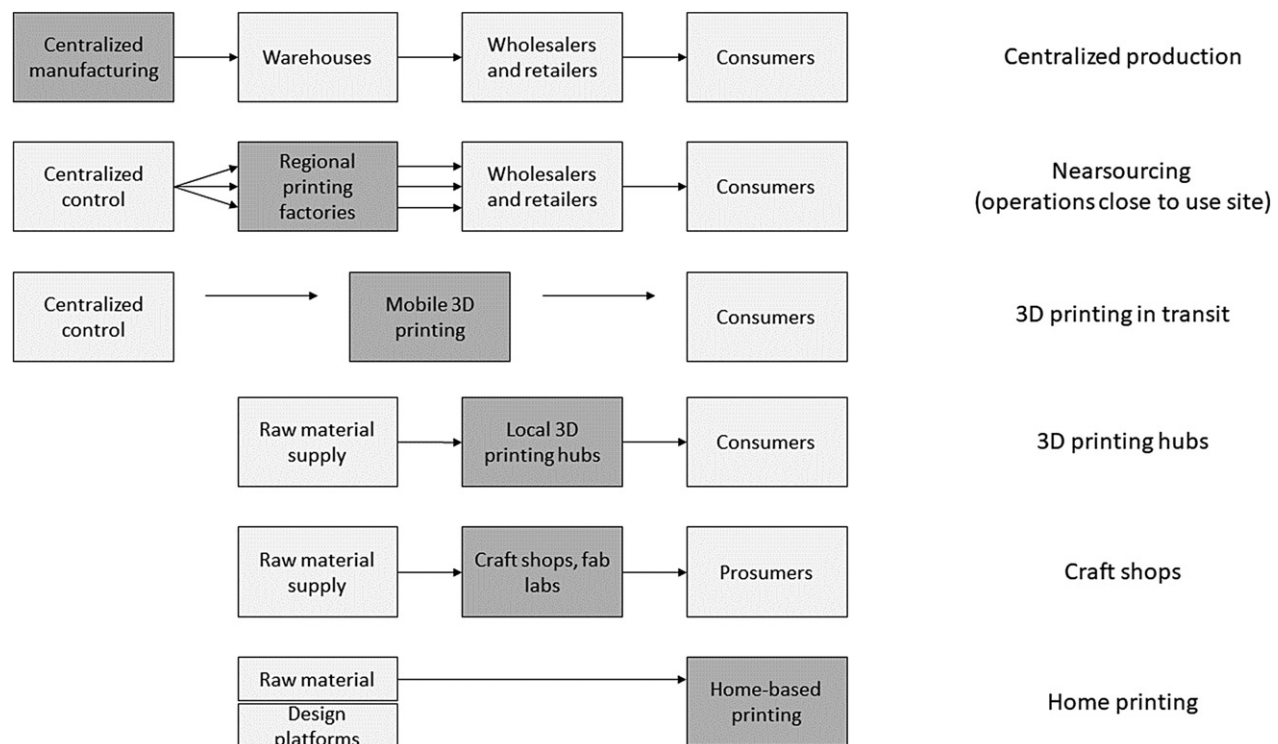


Figure 3 Spectrum of possible 3D printing locations.

emissions (Chen et al., 2015; Rauch et al., 2016). 3DP might, due to its production-on-demand model, lead to reduced inventory waste and greater potential for recycling (Despeisse et al., 2017; Ford and Despeisse, 2016). On the other hand, the logistical network might also become less efficient: “a change from standard to individualized, highly responsive logistics services [. . .] will lead to less sustainable logistics systems” (Tavasszy et al., 2003).

Discussion

How the future of 3DP will look is largely dependent on the interactions between end-user needs and wants, location decisions, and transport resistance. 3DP is associated with the personalization of products and might tie in with developments like mass-customization. Of course, not all types of products are equally suitable for personalization and in some cases small batches are not (cost)efficient. Still, in some sectors like aerospace and medicine, small batch 3DP is widely adopted, and other sectors might follow suit. As such, 3DP is expected to result in decentralization, although the exact nature of decentralization is still open for discussion. One may see the location of 3DP as part of a spectrum, ranging from centralized production all the way through to home-based printing (Fig. 3). Obviously, the exact outcome is dependent on factors like type of product to be manufactured, type of material, etc. Since there are a wide variety of consumer needs as well as types of materials and products, most likely a hybrid scenario will unfold in which different forms of 3DP in different locations are coexisting.

The emergence of 3DP and the associated decentralization of manufacturing will impact the transport and logistics sector in several ways:

- Transport volumes might not necessarily decrease, but delivery times and transport costs are expected to decrease.
- In terms of transport flows and movements, distribution networks are expected to be organized more efficiently. Distributed manufacturing still means that raw materials need to be shipped to decentralized locations. However, the transport of raw materials is expected to be more efficient, as they are moved in bulk in forms like cartridges, leading to less empty vehicles.
- The way inventories are kept will change: the size of inventories might decrease, amongst other things, because goods are now made to order and because safety-stock is not needed. Especially in the area of service parts, there are new opportunities for the instantaneous production of spare parts. Associated with this is the fact that suppliers run less risk of keeping products in stock that are never sold.
- The roles and positions of companies in the value chain will shift. In some scenarios some intermediaries, such as wholesalers or even retailers, might become superfluous. Some shops will vanish altogether or change into places to showcase products. At the same time, companies might strategically make upstream or downstream movements. Logistics companies might move into decentralized production by opening 3DP hubs. For example, postal services can turn their post offices into printshops and use their distribution networks to provide last-mile deliveries (United States Postal Service Office of Inspector General, 2014). Another example is wholesalers starting to emphasize stocking and delivering raw materials needed for home-based printing.
- Even if the roles of the companies in the value chain remain unchanged, the exchange of information will be affected. 3DP emphasizes agile development, on-demand production and customization, which means that customers are increasingly submitting personalized designs, either created on their own, or based on designs available on internet platforms. Design information needs to be sent to the 3D printer and probably checked before manufacturing can begin. Related to this, control over intellectual property rights and unwanted products is a challenge, as design platforms consist of propriety designs and there was commotion about the creation of a design template for a 3D-printed gun.
- Thus digitalization is becoming more important, because the design of products may move from end users to printing hubs, or from centralized companies to near-shored printing hubs (Fig. 2). There is a large role for quality control in this process: decentralized production still needs to adhere to quality standards imposed by the company and/or the government (e.g. in the form of safety and the environment).

Concluding Remarks

This contribution set out to explore the influence of 3DP on transport and logistics. More specifically, it dealt with how 3DP will influence consumer wants and needs, location choices and transport resistance. It has also addressed what are the repercussions for society and, specifically, the transport and logistics sector. Experts and industry watchers claim that 3DP has the potential to disrupt industries, including the transport and logistics sector. 3DP is an emerging technology, meaning that many aspects, like demands, networks and the technology itself, are under development, malleable and uncertain. Crucial factors of 3DP are the decentralization of manufacturing locations and the on-demand and agile nature of consumer demand. These two cornerstones will influence the way supply chains are organized and how logistics is organized. In turn this might then influence transport flows and related impacts on society in terms of safety, accessibility and the environment, as well as the nature and survival of firms in the value chain of many industries. Monitoring of the developments of 3DP in relation to transport and logistics is crucial, because it might have consequences for infrastructure, regulation, the capacity of logistics hubs like ports and harbors, and many more.

Biographies



Wouter Boon is Assistant Professor in Innovation. He works at the Innovation Studies group that is part of the Copernicus Institute of Sustainable Development, Utrecht University. His research in the field of innovation studies focuses on the dynamics and governance of emerging technologies in science-based sectors, such as transport and life sciences. The theoretical focus of his work is on the role of user innovations, user-producer interactions, regulation, and knowledge production in emerging technology fields.



Bert van Wee is Professor in Transport Policy at Delft University of Technology, the Netherlands, faculty Technology, Policy, and Management. In addition he is scientific director of TRAIL research school. His main interests are in long-term developments in transport, in particular in the areas of accessibility, land-use transport interaction, (evaluation of) large infrastructure projects, the environment, safety, policy analyses, and ethics.

References

- Barz, A., Buer, T., Haasis, H.D., 2016. Quantifying the effects of additive manufacturing on supply networks by means of a facility location-allocation model. *Logist. Res.* 9 (1).
- Bhattacharjya, J., Tripathi, S., Taylor, A., Taylor, M., Walters, D., 2014. *Additive Manufacturing: Current Status and Future Prospects*. Berlin, Heidelberg, Springer pp. 365–372. http://doi.org/10.1007/978-3-662-44745-1_36.
- Bogers, M., Hadar, R., Bilberg, A., 2016. Additive manufacturing for consumer-centric business models: implications for supply chains in consumer goods manufacturing. *Technol. Forecast. Soc. Chan.* 102, 225–239, 024 <http://doi.org/10.1016/j.techfore.2015.07>.
- Boon, W., van Wee, B., 2018. Influence of 3D printing on transport: a theory and experts judgment based conceptual model. *Transp. Rev.* 38 (5), 556–575.
- Bradshaw, S., Bowyer, A., Haufe, P., 2010. The intellectual property implications of low-cost 3D printing. *ScriptEd* 7 (1), 5–31.
- Campbell, T., Williams, C., Ivanova, O., Garrett, B., 2011. *Could 3D Printing Change the World?* Atlantic Council, Washington, DC.
- Chan, H.K., Chan, F.T.S., 2010. Comparative study of adaptability and flexibility in distributed manufacturing supply chains. *Decis. Supp. Sys.* 48 (2), 331–341, <http://doi.org/10.1016/j.dss.2009.09.001>.
- Chen, D., Heyer, S., Ibbotson, S., Salonitis, K., Steingrímsson, J.G., Thiede, S., 2015. Direct digital manufacturing: definition, evolution, and sustainability implications. *J. Clean. Product.* 107, 615–625, <http://doi.org/10.1016/j.jclepro.2015.05.009>.
- Christopher, M., Ryals, L.J., 2014. The supply chain becomes the demand chain. *J. Busin. Logist.* 35 (1), 29–35, <http://doi.org/10.1111/jbl.12037>.
- De Jong, J.P.J., De Bruijn, E., 2013. Innovation lessons from 3-D printing. *MIT Sloan Manage. Rev.* 54 (2), 43–52.
- Deloitte, 2019. *Deloitte Insights: 3D Printing Growth Accelerates Again – TMT Predictions 2019*. Available from: <https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/3d-printing-market.html>.
- Despeisse, M., Baumers, M., Brown, P., Charnley, F., Ford, S.J., Garmulewicz, A., Knowles, S., Minshall, T.H.W., Mortara, L., Reed-Tsochas, F.P., Rowley, J., 2017. Unlocking value for a circular economy through 3D printing: a research agenda. *Technol. Forecast. Soc. Chan.* 115, 75–84, <http://doi.org/10.1016/j.techfore.2016.09.021>.
- Durach, C.F., Kurpijuweit, S., Wagner, S.M., 2017. The impact of additive manufacturing on supply chains. *Int. J. Phys. Distribut. Logist. Manag.* 47 (10), 954–971.
- Durão, L.F.C., Christ, A., Zancul, E., Anderl, R., Schützer, K., 2017. Additive manufacturing scenarios for distributed production of spare parts. *Inter. J. Adv. Manufact. Technol.* 93 (1–4), 869–880.
- Economist, 2016. *A Printed Smile – 3D Printing is Coming of Age as a Manufacturing Technique*, 30 April.
- Fawcett, S.E., Waller, M.A., 2014. Can we stay ahead of the obsolescence curve? On inflection points, proactive preemption, and the future of supply chain management. *J. Busin. Logist.* 35 (1), 17–22, <http://doi.org/10.1111/jbl.12041>.
- Ford, S., Despeisse, M., 2016. Additive manufacturing and sustainability: an exploratory study of the advantages and challenges. *J. Clean. Product.* 137, 1573–1587, 150 <http://doi.org/10.1016/j.jclepro.2016.04>.

- Fornasiero, R., Chiodi, A., Carpanzano, E., Carneiro, L., 2010. Research Issues on Customer-Oriented and Eco-friendly Networks for Healthy Fashionable Goods. Springer, Berlin, Heidelberg pp. 36–44. http://doi.org/10.1007/978-3-642-14341-0_5.
- Frazier, W.E., 2014. Metal additive manufacturing: a review. *J. Mater. Eng. Perfor.* 23 (6), 1917–1928, <http://doi.org/10.1007/s11665-014-0958-z>.
- Garrett, B., 2014. 3D printing: new economic paradigms and strategic shifts. *Global Policy* 5 (1), 70–75, <http://doi.org/10.1111/1758-5899.12119>.
- Gausemeier, J.E., 2014. Thinking ahead the future of additive manufacturing e innovation roadmapping of required advancements. Paderborn, Germany.
- Gress, D.R., Kalafsky, R.V., 2015. Geographies of production in 3D: theoretical and research implications stemming from additive manufacturing. *Geoforum* 60, 43–52, <http://doi.org/10.1016/j.geoforum.2015.01.003>.
- Gosling, J., Naim, M.M., Fowler, N., Fearn, A., 2007. Manufacturers' preparedness for agile construction. In: IET International Conference on Agile Manufacturing, pp. 103–110.
- Griffiths, A., 2002. Rapid manufacturing – the next industrial revolution. *Materials World* 10 (12), 34–35.
- Holmström, J., Partanen, J., Tuomi, J., Walter, M., 2010. Rapid manufacturing in the spare parts supply chain. *J. Manufact. Technol. Manag.* 21 (6), 687–697, <http://doi.org/10.1108/17410381011063996>.
- Huang, R., Riddle, M., Graziano, D., Warren, J., Das, S., Nimbalkar, S., Cresko, J., Masanet, E., 2016. Energy and emissions saving potential of additive manufacturing: the case of lightweight aircraft components. *J. Clean. Product.* 135, 1559–1570, <http://doi.org/10.1016/j.jclepro.2015.04.109>.
- Huang, S.H., Liu, P., Mokasdar, A., Hou, L., 2013. Additive manufacturing and its societal impact: a literature review. *Inter. J. Adv. Manufact. Technol.* 67 (5–8), 1191–1203, <http://doi.org/10.1007/s00170-012-4558-5>.
- Janssen, R., Blankers, I., Moolenburgh, E., Posthumus, B., 2014. The impact of 3-D printing on supply chain management. TNO, Delft.
- Jia, F., Wang, X., Mustafee, N., Hao, L., 2016. Investigating the feasibility of supply chain-centric business models in 3D chocolate printing: a simulation study. *Technol. Forecast. Soc. Change* 102, 202–213, <http://doi.org/10.1016/j.techfore.2015.07.026>.
- Kanto, L., Alahuhta, P., Kukko, K., Pihlajamaa, J., Partanen, J., Vartiainen, M., Berg, P., 2014. How do customer and user understanding, the use of prototypes and distributed collaboration support rapid innovation activities? In: Management of Engineering & Technology (PICMET). Portland International Center for Management of Engineering and Technology, pp. 784–795.
- Li, Y., Jia, G., Cheng, Y., Hu, Y., 2017. Additive manufacturing technology in spare parts supply chain: a comparative study. *Internat. J. Product. Res.* 55 (5), 1498–1515, <http://doi.org/10.1080/00207543.2016.1231433>.
- Lopes da Silva, J.V., Vicente, J., 2013. 3D technologies and the new digital ecosystem. In: Proceedings of the Fifth International Conference on Management of Emergent Digital EcoSystems - MEDES '13. ACM Press, New York, New York, USA, pp. 278–284. <http://doi.org/10.1145/2536146.2536187>.
- Manners-Bell, J., Lyon, K., 2012. The implications of 3D printing for the global logistics industry. Transport Intelligence Ltd., Bath, UK.
- Mavri, M., 2015. Redesigning a production chain based on 3D printing technology. *Knowl. Process Manag.* 22 (3), 141–147.
- McKinnon, A., 2016. The possible impact of 3D printing and drones on last-mile logistics: an exploratory study. *Built Environ.* 42 (4), 617–629, <http://doi.org/10.2148/benv.42.4.617>.
- Mohr, S., Khan, O., 2015. 3D printing and supply chains of the future. In: Innovations and Strategies for Logistics and Supply Chains. epubli GmbH.
- Mourtzis, D., Doukas, M., 2013. Decentralized Manufacturing Systems Review: Challenges and Outlook. Springer, Berlin Heidelberg pp. 355–L369 http://doi.org/10.1007/978-3-642-30749-2_26.
- Ngo, T.D., Kashani, A., Imbalzano, G., Nguyen, K.T., Hui, D., 2018. Additive manufacturing (3D printing): a review of materials, methods, applications and challenges. *Composit. B Eng.* 143, 172–196.
- PWC, 2014. 3D Printing and the New Shape of Industrial Manufacturing. PricewaterhouseCoopers LLP, London.
- Rauch, E., Dallasega, P., Matt, D.T., 2016. Sustainable production in emerging markets through Distributed Manufacturing Systems (DMS). *J. Clean. Product.* 135, 127–138, <http://doi.org/10.1016/j.jclepro.2016.06.106>.
- Rayna, T., Striukova, L., 2016. From rapid prototyping to home fabrication: How 3D printing is changing business model innovation. *Technol. Forecast. Soc. Change* 102, 214–224, <http://doi.org/10.1016/j.techfore.2015.07.023>.
- Rifkin, J., 2011. The Third Industrial Revolution. Palgrave MacMillan, London.
- Robinson, D.K., Lagnau, A., Boon, W.P.C., 2019. Innovation pathways in additive manufacturing: methods for tracing emerging and branching paths from rapid prototyping to alternative applications. *Technol. Forecast. Soc. Change* 146, 733–750.
- Rogers, H., Baricz, N., Pawar, K.S., 2016. 3D printing services: classification, supply chain implications and research agenda. *Inter. J. Phys. Distribut. Logist. Manag.* 46 (10), 886–907, <http://doi.org/10.1108/IJPDLM-07-2016-0210>.
- Romero, A., Vieira, D.R., 2016. How Additive Manufacturing Improves Product Lifecycle Management and Supply Chain Management in the Aviation Sector? Springer, Cham pp. 74–85 http://doi.org/10.1007/978-3-319-33111-9_8.
- Ryan, M.J., Evers, D.R., Potter, A.T., Purvis, L., Gosling, J., 2017. 3D printing the future: scenarios for supply chains reviewed. *Inter. J. Phys. Dist. Log. Manag.* 47 (10), 992–1014.
- Sasson, A., Johnson, J.C., 2016. The 3D printing order: variability, supercenters and supply chain reconfigurations. *Inter. J. Phys. Distribut. Logist. Manag.* 46 (1), 82–94.
- Tavasszy, L.A., Ruijgrok, C.J., Thissen, M.J.P.M., 2003. Emerging global logistics networks: implications for transport systems and policies. *Growth Change* 34 (4), 456–472, <http://doi.org/10.1046/j.0017-4815.2003.00230.x>.
- Tuck, C., Hague, R., Burns, N., 2007. Rapid manufacturing: impact on supply chain methodologies and practice. *Inter. J. Serv. Operat. Manag.* 3 (1), 1, <http://doi.org/10.1504/IJOSM.2007.011459>.
- United States Postal Service Office of Inspector General, 2014. 3D Printing and the U.S. Postal Service. Mary Ann Liebert, Inc. 1 (3), 166–168.
- Westerweel, B., Basten, R.J., van Houtum, G.J., 2018. Traditional or additive manufacturing? Assessing component design options through lifecycle cost analysis. *Eur. J. Operat. Res.* 270 (2), 570–585.
- Wohlers Report, 2019. 3D Printing and Additive Manufacturing – State of the Industry. Fort Collins, Colorado, USA.

The Physical Internet and Logistics

Eric Ballot, Shenle Pan, MINES ParisTech, PSL University, Centre de gestion scientifique (CGS), i3 UMR CNRS, Paris, France

© 2021 Elsevier Ltd. All rights reserved.

Motivations for a New Theory of Logistics Networks	479
The State of the Art: Fragmented Networks and Operations	479
A Major Consequence: A Low Level of Efficiency	479
New Challenges: Service Improvement and Sustainability	480
The Physical Internet	481
Origins	481
Definition	482
Main Components	482
Physical	482
Data	483
Process	483
Legal	485
Implementation of the Physical Internet	485
Simulation	485
Start-ups: Routing, Storage, and Open Services	485
Container Development	486
Perspectives	486
Acknowledgment	486
References	486
Further Reading	487

Motivations for a New Theory of Logistics Networks

To understand why and how the Physical Internet was proposed, this part starts with the context of logistics networks, how and why they are organized. The analysis of their efficiency shows their limits especially in regards to new challenges such as e-commerce delivery and sustainability.

The State of the Art: Fragmented Networks and Operations

Logistics networks are a heritage of proprietary networks developed in the 19th and 20th centuries and dedicated to companies, a car manufacturer or a retailer as examples, or a specific service like mail with postal services over a limited territory. As a result, there are as many logistics networks as there are combinations of companies and services. This organization was not an issue as volumes became concentrated due to mass production or distribution or in monopolistic services as mail. With the evolution of both industry and distribution toward just-in-time and multiple channels all over the world, needs have drastically changed at the end of the 20th century with the creation of logistics networks by postal operators and carriers and the development of logistic service providers (LSP). With the notable exception of the maritime shipping industry, which is highly concentrated, the logistics sector remains vastly highly fragmented even though it is concentrating over time. Furthermore, even with major players in the markets, services also remain fragmented with dedicated resources. An LSP will dedicate a warehouse to a customer and operate it independently of other activities; a courier company will not share resources for different levels of service and so on.

Despite the ongoing concentration in the sector, operations remain highly fragmented as the need for shipping becomes more and more fragmented, with just-in-time deliveries and the proliferation of distribution channels. The effect is not well described but a survey made in France by IFSTTAR (the French transportation institute) showed that the median shipment weight dropped from 160 kg in 1988 to 30 kg in 2004. At that time, this trend did not take into account e-commerce activity as it was completely marginal. So this trend is even further reinforced now.

A Major Consequence: A Low Level of Efficiency

The fragmentation of flows makes any consolidation and therefore, the creation of positive scale or scope effects difficult. The fragmentation of flows is not directly measured and there is no globally accepted indicator. The phenomenon could, however, be illustrated by example. Fig. 1 represents the flows of goods between two main retailer distribution centers and their common top 100 suppliers.

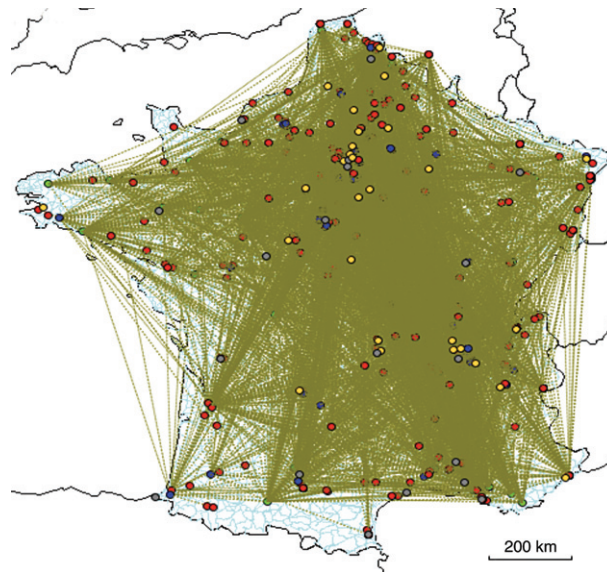


Figure 1 Flow map of two main retailers and their common top 100 suppliers located in France.

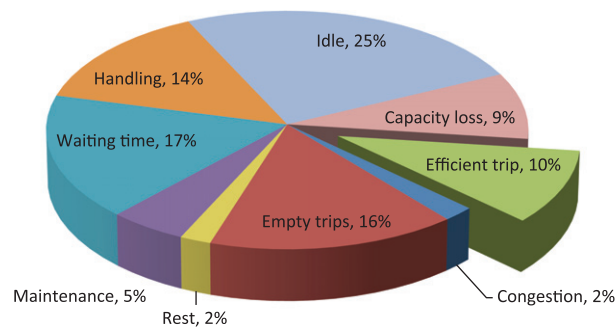


Figure 2 Analysis of truck's overall efficiency in the FMCG sector.

Another approach is to look at asset utilization rates. If there is no public information on the filling rate of warehouses, statistics can be found on the use of trucks. A 60% fill rate and 20% of empty km are common numbers in this regard. However, this falls far short of achieving overall efficiency. The overall efficiency for vehicles can be computed. As shown in [Fig. 2](#) based on [McKinnon et al. \(2003\)](#), the results show several main sources of losses and an astonishingly low level of 10% of efficient trips. [Fig. 2](#) indicates that compared to a perfect world, freight transportation could be performed with 10 times less trucks. Even if perfection is out of reach, the potential for progress remains considerable.

To cope with the inefficiencies now well known in logistics, one possible approach consists of pooling some resources—transportation means, warehouses, etc. This concept has been known for more than a decade, with real improvements in efficiency when applied. Generally, costs and CO₂ emissions are reduced by 25%, but the commitment required from companies to participate in such an organization is still a barrier against its generalization ([Cruijsen et al., 2007](#)).

New Challenges: Service Improvement and Sustainability

Despite low efficiency, logistics services are generally very good. They are affordable, reliable and delivery times have become shorter and shorter. So, they are more and more used. Economic development is coupled with the growth of flows. By 2050, economic development is anticipated to lead to an increase in freight traffic of more than 200% in Europe and North America and to nearly 600%. In Asia for maritime or surface freight ([ITF, 2017](#)). As a result, if nothing changes, this will lead to much more traffic than today and a much greater need for infrastructure. Even though more environmentally friendly vehicles are gradually being brought into service and products are being supplied closer to their places of consumption, it is clear that the current trend will not achieve the goals of sustainable development, starting with a 60% reduction in total CO₂ emissions. This would mean an almost complete decarbonization of the transport sector, including infrastructure. It seems out of reach on the basis of actual and foreseen technologies.

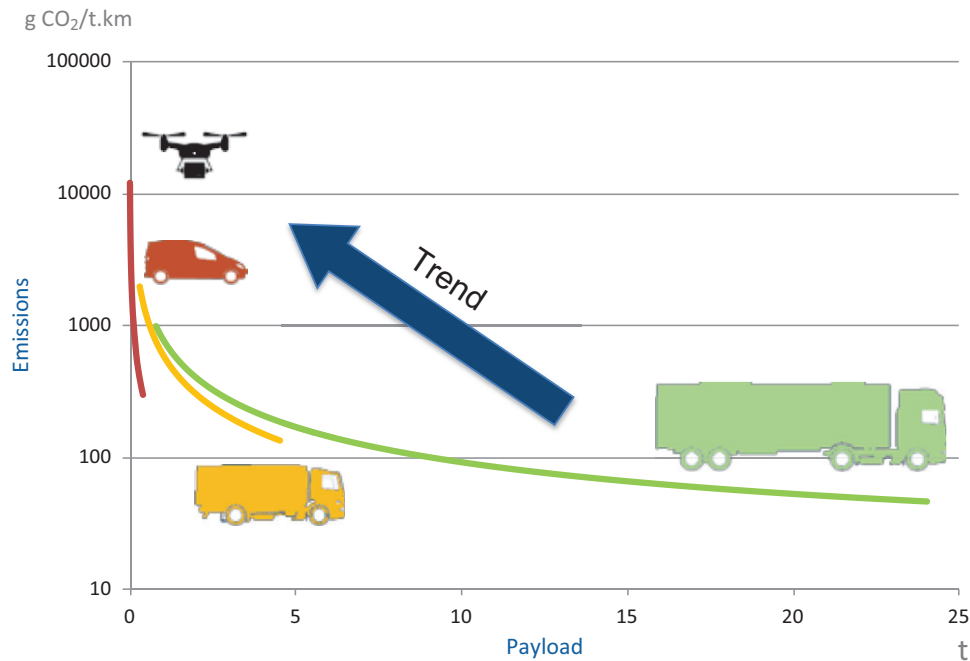


Figure 3 Emission factor and vehicle payload.

One might think that the solution to quickly delivering small quantities of goods would be to reduce the size of vehicles and accelerate their speed. This is, for example, the approach underpinning the potential use of drones. This approach, however, faces a major obstacle, that of scale effects. If it is better to have a small, well-filled vehicle than a bigger one, nothing can replace the efficiency of a well-filled big vehicle. This is not only true from the cost of transport operations point of view, it is also true for externalities and in particular for CO₂ emissions (Fig. 3).

This is where the antagonism between service improvement and impact on sustainability occurs. Indeed, if the acceleration of flows is reflected in the use of smaller scale means, the overall externalities of freight transport will increase. These include not only environmental impacts (air pollution, global warming, noise . . .), but also social costs (lost time, accidents...) which, ultimately, can represent a cost to society of the same order of magnitude as the cost of transportation itself in densely populated areas (European Commission, 2019). Under the current organization, therefore, there is a strong dilemma between a growing demand for fast deliveries, thus in smaller quantities, and the resulting environmental footprint. It is not a question of means of transportation, but a question of how demand is served by supply chains.

The Physical Internet

The previous analysis shows how logistics networks are organized and what results from them in terms of transport. This justifies finding new approaches to the organization of logistics activities that can help solve the challenges.

Origins

The origins of the Physical Internet are multiple. It has emerged naturally as the result of analyses carried out on the durability of logistic organizations and certain preexisting organizations described later. But the term Physical Internet comes from the title of a "classic" article on logistics published in *The Economist* in 2006 (Markillie, 2006) and once found by Benoit Montreuil was transformed into a research question. What would logistics be if it was designed as an internet? That is, as a set of interconnected networks and not as a set of independent heterogeneous networks. The generic question already has known answers in computing (the internet) but also in electrical energy, for example, electrical grid, and then smart grid. The question applied to logistics helps to rethink all logistics activities, but also helps to discover among existing practices those that could be considered as a beginning of the physical Internet.

The set of experiences and practices carried out in the field of horizontal collaboration to associate several companies to share logistics resources corresponds in part to the objectives of the physical internet to allow users to use shared resources with the benefits previously mentioned. However, each collaboration requires reinventing its content, which limits its development.

Another organization prefigures what could be the Physical Internet and the Dabbawallas. This refers to a distribution organization of home-cooked meals prepared in the morning and delivered for lunch at workplaces in Mumbai. The system

operates from identical boxes with a unique code that allows them to be routed to their destination in the midst of thousands of others and using all modes of urban transportation and multiple independent service providers (Thomke, 2012). This organization is very close to the internet physics, except that it provides only one service and is not connected to another place.

The last example concerns containerized marine shipments, which have evolved considerably since the standardization of containers and their fasteners, the twist-lock. This technical system has favored multimodal transport and the realization of logistical schemes using shared means. On the other hand, the logic of hub and shared transport usually stops at the port or is limited to railway lines in the countries where these networks are developed.

Definition

The core idea of the physical Internet is the interconnection of networks, which leads to a network of networks. The parallel with the digital Internet is obvious. The Internet is the network that encompasses all computer networks, but the means of this interconnection needs to be defined for logistics network to make it more tangible.

The definition was elaborated by Montreuil et al. (2012). It is the following: the Physical Internet is a global logistics system based on the interconnection of logistics networks by a standardized set of collaboration protocols, modular containers, and smart interfaces for increased efficiency and sustainability. The goal is efficiency, the means is the interconnection of logistics networks that involves standardization and protocols at four levels: physical, data, process, and legal to ensure compatible operations, regardless of the operator and the nature of the service provided.

This definition also limits the scope of the Physical Internet to logistics. The interconnection is established at (containerized) shipments in all aspects that affect them, but not beyond. Thus, a truck may be subject to a possible development to accommodate modular containers, but the road is not affected. However, a container-sorting center will have to be adapted to them for their handling and monitoring. In the same way, the commercial relationship between a supplier and his customer is not the object of the interconnection. On the other hand, an interconnection can strongly influence the design of truck trailers as well as how to manage a supply chain to take advantage of the potential offered.

Main Components

Physical

The equivalent of a digital packet for the physical Internet is a shipping unit: a container. Studies carried out in collaboration with industry, in the United States and in Europe, have shown that two layers of encapsulation between the products (packaged or not) and the means of transport or the manipulator are the maximum necessary. The first layer, the handling box is modular in external dimensions, stackable, and ensures the privacy and security of the goods, the second layer is how the transport container or interface fits with modular handling boxes, ensures the interface with the transportation means and protection from external conditions, if necessary. Both layers ensure handling productivity and are fully traceable. Fig. 4 describes the layers.

The standardization of containerization and associated handling tools is motivated by the current lack of widely accepted formats for cartons, the variety of pallet sizes and, consequently, the space lost in the stacks, the poor quality, the low productivity and the consumption of raw materials. In this respect, the study of the history of development of the maritime container is illuminating

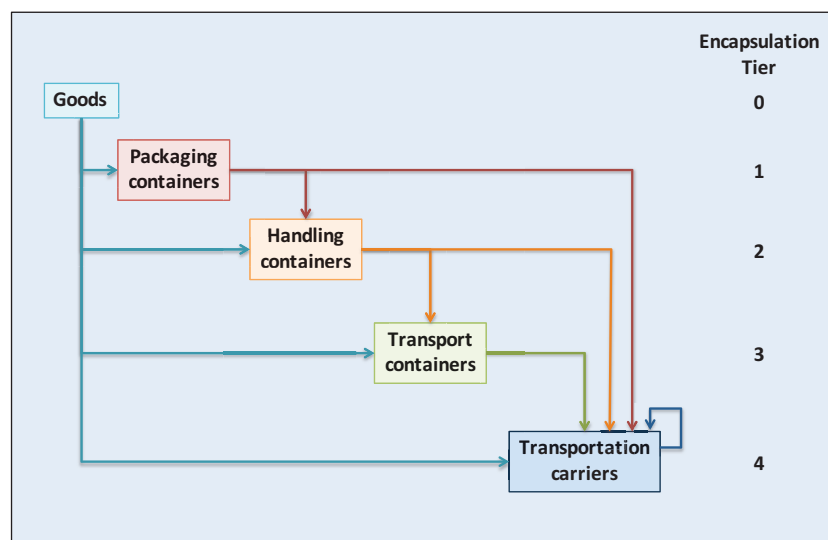


Figure 4 Encapsulation layers in the Physical Internet.

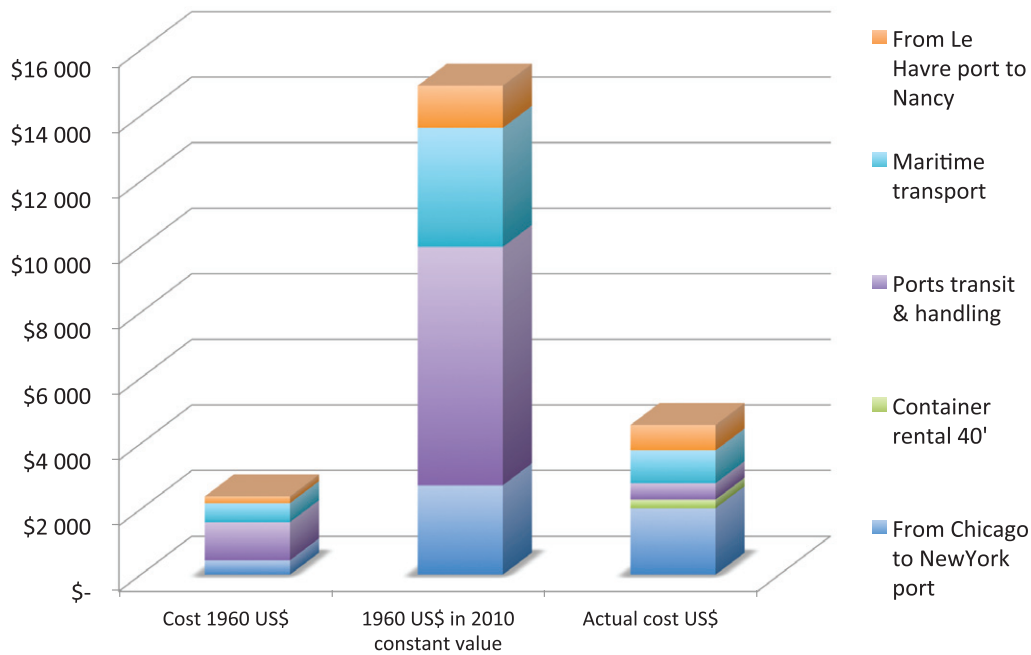


Figure 5 Comparison of cost components for a shipment equivalent to one full container from Chicago to Nancy. Source: Original data from Levinson (2006).

(Levinson, 2006). After a difficult start, the acceptance of both 20 and 40-foot formats and the twist-lock has enabled the shipping and port industries to develop impressive productivity gains in 50 years. The reconstitution of the productivity gains by the evolution of the prices of the components of a consignment between Chicago and Nancy before containerization and currently shows the progress made possible by the standardization of the container and, conversely, the absence of productivity gains for land transport. See Fig. 5 for the numbers.

Data

One of the advantages of using a dedicated solution is to facilitate the monitoring and control of operations by interfacing the information system of the service provider and its client. However, there are now standards, such as EPCglobal, which make it possible to designate products, handling objects, consignments, and places in a unique way on the one hand and, on the other hand, there are independent monitoring solutions from the information system for products and shipments.

The development of passive tags (RFID) or active (Bluetooth) for short-field communication solutions, coupled with long-range solutions (LORA) based on multisectoral telecommunication infrastructures, will in the near future provide almost permanent tracking of unit loads, if necessary. This can be done independently of logistics operators and their information systems that will rely on the same sources of information. Fig. 6 illustrates a possible scheme, already tested. Finally, the deployment of 5G technology will make a qualitative leap while keeping costs low.

A common language and technologies are now almost ready to bypass the interfaces between existing information systems for the needs of the Physical Internet. Clearly, data ownership, access permissions, and management remain. Blockchain technologies could also provide interesting support for logistics operations within the physical Internet, thus ensuring the authentication of logistics transactions between logistics providers and their customers.

Process

The interconnection of processes allows one of the main functionalities of the physical internet, namely the routing of the logistics units (containers) from their origin to their destination as requested during the dispatch. This may be different from the final destination, because the Containers can be brought to a hub, stored and later brought closer to the final destination. So routing with the physical internet must be able to transport, but also store, on demand. The determination of these needs, however, depends on the management of the supply chain.

Virtually, a customer indicates where to place each of his handling boxes in space and time. The request can be very precise (e.g., a given warehouse on a fixed date) or more flexible (e.g., before a deadline in a geographical area, with a short deadline or on the contrary a longer one), with environmental requirements, etc. These are the criteria of each container that defines the expected service that will be evaluated individually and aggregated for each service and provider. The points and means used at each intermediate stage of the routing are left to the discretion of the operators to enable them to optimize the routing. Routing can be broken down into two main steps.

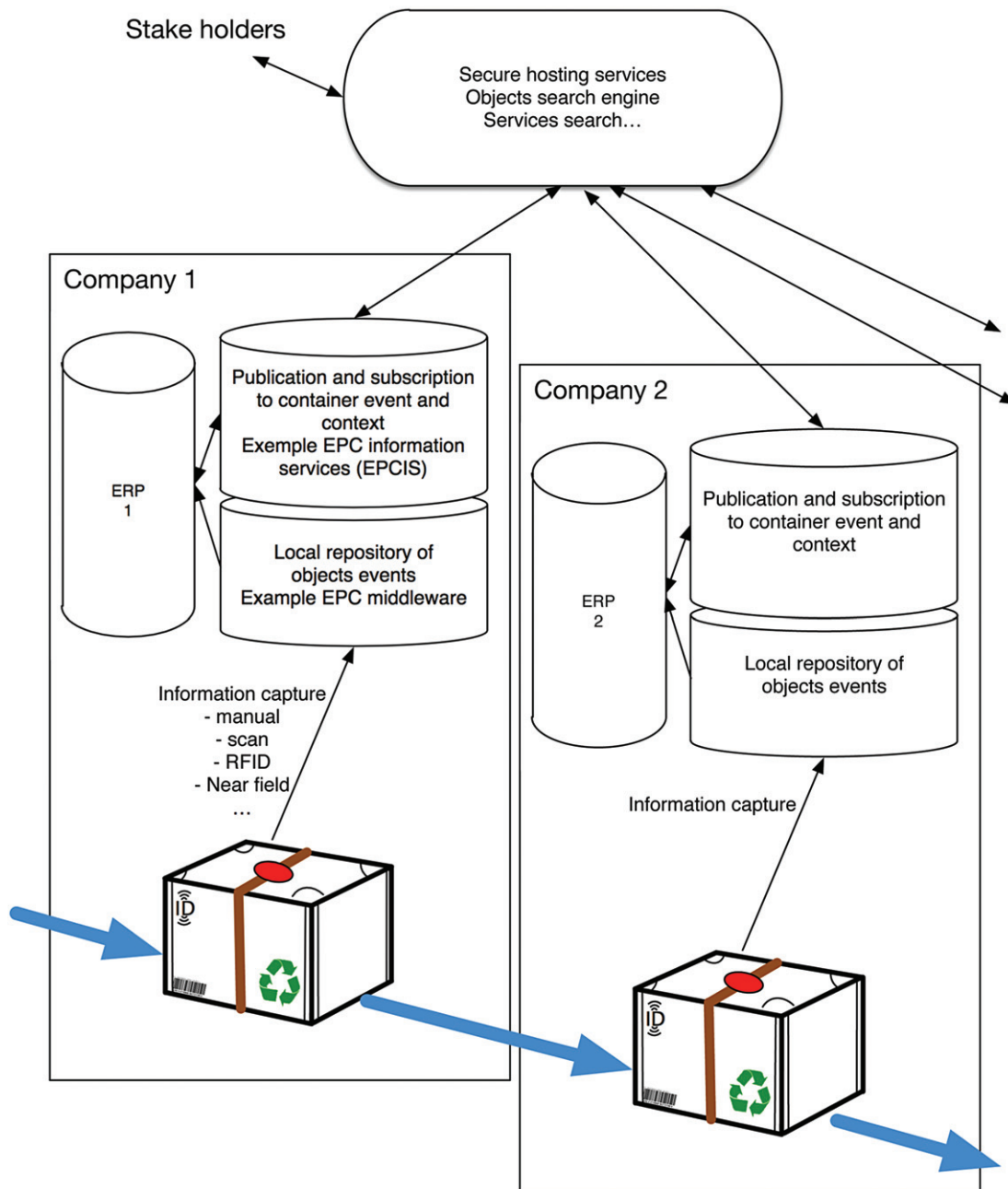


Figure 6 Digital tracking of containers in the Physical Internet.

The first step determines the path that includes transport services (network arcs) and storage and handling services (network nodes). For the verification of the feasibility and the fixing of a price offer, at least one end-to-end routing solution must be available. This solution may be frozen (booking) or otherwise remain flexible to allow adaptation to real conditions, or even an improvement of the initial routing solution. The second hypothesis implies a recalculation of the route according to the offers, but remains subject to the initial conditions. The initial solution and the following are calculated by conventional algorithms of shorter path for example. The services included in the search could be selected according to a minimum proven service level.

The second step concerns the actual allocation of the container to a specific service: transport, handling, storage, or a combination of these. At each stage of routing, a service provider must be determined and a transaction must occur to transfer each container with its specific needs. The selection of services for a set of containers arriving at a hub can be done with different purchasing solutions. Mechanism design, with combinatorial auction, has already been successfully tested in simulation. The main idea is that every time the best solution available is used, according to the requirements.

Legal

All logistic operations are covered by contracts in which the commitments and responsibilities of the stakeholders are established, including the case of outsourcing that is already widely used and internationally. In particular, this aspect has already been partly standardized by incoterms. The question is, therefore, whether these are sufficient or whether they will need to be amended and supplemented. It would seem that they are currently sufficient.

Another concern is the competition between LSP and the platforms that offer them. This point also mentioned in pooling is not yet resolved, but it is already noted in simulations that it could, on the contrary, make the market more dynamic and open.

Finally, the question of the governance of the physical Internet is often rightly raised. The Physical Internet being currently in the experimental phase and not yet deployed, the question can be postponed. However, the recent projects, pilots and exploitation of the concept suggest that centralized regulation already appears necessary, especially to label solutions as part of the Physical Internet. This will be the case for containers when a solution emerges, but also for identification codes or routing protocols.

The physical internet deployment represents a profound change in tools, systems, and processes as evidenced by the impact on the four levels described earlier.

Implementation of the Physical Internet

Simulation

The first implementations of the physical Internet were realized with simulation tools of supply chains and logistic networks, in order to understand how the theory could work on existing cases. That is to say, reproducing real cases of applications. This work was conducted in Europe, North America, and Hong Kong by independent research teams. The results obtained confirmed the hypothesis of a significantly more efficient organization. Thus, a large-scale simulation carried out in France and dealing with the flows of two distributors and their main suppliers, mentioned above, made it possible to show the impact on the flow structure, which is not only much more concentrated (Fig. 7), but also achieves improvements of more than 30% of filling rates, the possibility of making full trains or the reduction of 15% of t.km. In the end, there are almost 60% less emissions than with current technologies. The impact on inventories is identical, with gains of around 30% in quantity with distributed stocks that are supplying each other. As a result, the inventory network ensures product availability in place of a centralized safety stock. In all cases the service is preserved or improved.

Start-ups: Routing, Storage, and Open Services

Several start-ups and companies, most of them awarded during the annual International Physical Internet Conference or IPIC (www.pi.events), joined the approach and propose logistics services that implement part of the Physical Internet. In the case of routing, Mix-Move-Match and Collaborating Routing Center Services or CRC Services are used. They propose to operate routing services over existing logistics networks. Both companies have shown that multi-client, multi-recipient flow routing is working and that efficiency improvement ambitions are achieved even on limited scales.

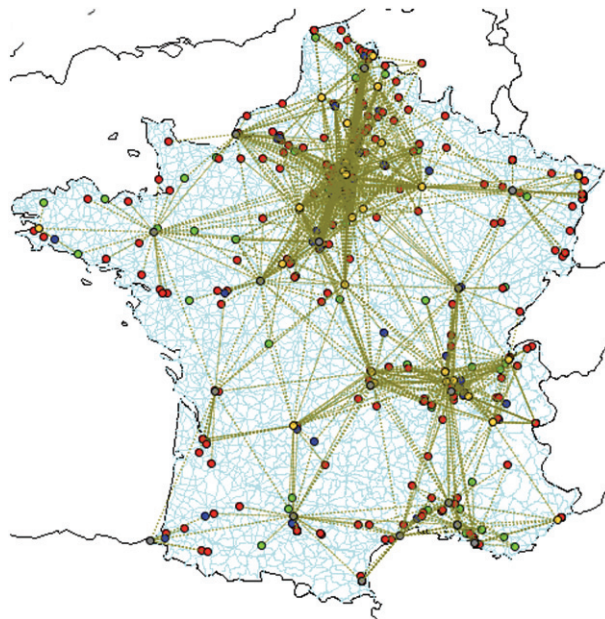


Figure 7 Flow map of freight in interconnected logistics networks enabled by the Physical Internet.

From a storage perspective, much like online booking sites, several companies, such as Flexe, Stock booking or OGOShip, offer on-demand permanent or temporary storage in a multitude of warehouses with unused capacity. Even if these companies are specialized at the beginning, they are moving toward a wider range of services: storage, preparation, and transport, in order to provide more comprehensive solutions that remain based on the same principles of shared and on-demand resources. When these companies open and sell their technology which allows other logistics service networks to operate in a consistent manner and have operations that move from one network to another, then interconnection becomes a reality, as is happening in 2019.

Container Development

The physical internet is often associated with containerization and the first European project Modulushca has directly worked on the benefits, design, implementation and testing of modular boxes. This effort continued with the help of the CGF then GS1 Germany, which carried out a successful test of boxes of 60×40 for care products. This box is now being rolled out by the companies that participated in the development, with the idea of a generalization at European level.

At the level of transport containers, several solutions exist, starting with sea or air containers. However, there is no suitable solution for multimodal terrestrial transport, particularly for urban deliveries. A European project, Clusters 2.0, focuses on the study and testing of a series of modular transport containers. The results are expected in 2020.

Standardized containerization is a subject that remains difficult because it involves significant investments to adapt logistics networks to these new tools. This approach is therefore gradual.

Perspectives

Current logistic organizations suffer from limitations related to network and flow fragmentation that leads to low efficiency at ever higher service levels. The organization promoted by the physical Internet, by the extent of the resources available and the flexibility of the routing, makes it possible to reach levels of efficiency that are very much superior from a theoretical point of view to the current organizations.

The partial experiments of the physical internet (containerization, monitoring, routing, etc.) carried out by companies, internally or at the level of groups of companies in a sector, have yielded convincing results. This means that it is not necessary to deploy all the elements of the physical internet to benefit from improvements in current operations. The fact that physical internet concepts can be used successfully by companies opens up interesting and potentially more accessible fields of application.

Experiments, ongoing operations and projects are yielding encouraging results. As such, the two next steps in the development of the Physical Internet can be envisaged within a few years. The first is to define a governance of the physical internet, in charge of collecting and synthesizing the results obtained, in order to turn them into technical and organizational recommendations. This step is fundamental to allow actors not involved in the developments to appropriate the results and to implement them. The European technological platform ALICE can play a major role at this level because it federates a large community of industrial and institutional players at the European level with two major objectives. The Physical Internet up and running in 2030 to reap gains from greater efficiency and zero emissions in 2050 to meet the objectives of sustainable development. The second step will be to extend to sectors other than that of fast-moving consumer goods that is currently at the cutting edge on this subject and to strengthen international collaboration around the Physical Internet.

Acknowledgment

The original development of the Physical Internet occurred during a set of workshops organized with Pr. Benoit Montreuil and Pr. Russell D. Meller. The authors also acknowledge the major role played by Sergio Barbarino from P&G and Fernando Liesa from the ALICE European Technology Platform in the development of the Physical Internet in Europe, with the strong and continuous financial support of the European Commission. Last but not least, the authors also want to sincerely thank all partners of the Chair Physical Internet (www.cip.mines-paristech.fr): P&G, GS1 France, FM Logistic, Carrefour Supply Chain, and Orange, for their trust, contributions, and financial support.

References

- Crujssen, F., Cools, M., Dullaert, W., 2007. Horizontal cooperation in logistics: opportunities and impediments. *Transp. Res. Part E* 43 (2), 129–142.
- European Commission, 2019. Handbook on the External Costs of Transport. Brussels. doi: 10.2832/27212.
- ITF, 2017. Capacity to Grow: Transport Infrastructure Needs for Future Trade Growth. International Transport Forum, Paris, France.
- Levinson, M., 2006. *The Box*. Princeton University Press, Princeton.
- Markillie, P., 2006. The physical internet: a Survey of logistics, *The Economist*, 17 June, pp. 1–14.
- McKinnon, A., Ge, Y., Leuchars, D., 2003. Analysis of Transport Efficiency in the UK Food Supply Chain. Logistics Research Centre, Heriot-Watt University, Edinburgh.
- Montreuil, B., Meller, R.D., Ballot, E., 2012. Physical internet foundations. *IFAC Proc. Vol.* 45 (6), 26–30.
- Thomke, S., 2012. Mumbai's Models of Service Excellence, *Harvard Business Review*, November.

Further Reading

- ALICE, 2019. Alliance for Logistics Innovation through Collaboration in Europe web site. Available from: <https://knowledgeplatform.etp-logistics.eu/>
- Ballot, E., Montreuil, B., and Meller, R.D., 2014. The physical internet: The network of logistics networks.
- Henrik, S., Norrman, A., 2017. The physical internet – review, analysis and future research agenda. *Int. J. Phy. Distrib. Logist. Manag.* 47 (8), 736–762.
- Huang, G.Q., Xu, S.X., 2013. Truthful multi-unit transportation procurement auctions for logistics e-marketplaces. *Transp. Res. Part B: Methodol.* 47, 127–148.
- Kong, X., Huang, G.Q., Luo, H., Yen, B.P., 2018. Physical-internet-enabled auction logistics in perishable supply chain trading: state-of-the-art and research opportunities. *Ind. Manag. Data Sys.* 18 (8), 1671–1694.
- McKinnon, A., 2018. *Decarbonizing Logistics: Distributing Goods in a Low-Carbon World*. Kogan Page Ltd.
- Mervis, J., 2014. The information highway gets physical: the physical internet would move goods the way its namesake moves data. *Science* 344 (6188), 1104–1107.
- Montreuil, B., 2011. Toward a physical internet: meeting the global logistics sustainability grand challenge. *Logist. Res.* 3 (2–3), 71–87.
- Pan, S., Ballot, E., Fontane, F., 2013. The reduction of greenhouse gas emissions from freight transport by pooling supply chains. *Int. J. Prod. Econ.* 143 (1), 86–94.
- Sarraj, R., Ballot, E., Pan, S., Hakimi, D., Montreuil, B., 2014. Interconnected logistic networks and protocols: simulation-based efficiency assessment. *Int. J. Prod. Res.* 52 (11), 3185–3208.
- Yang, Y., Pan, S., Ballot, E., 2016. Innovative vendor-managed inventory strategy exploiting interconnected logistics services in the physical internet. *Int. J. Prod. Res.* 55 (9), 2685–2702.

Bicycles for Urban Freight

Barbara Lenz, Johannes Gruber, German Aerospace Center, Berlin, Germany

© 2021 Elsevier Ltd. All rights reserved.

Challenges for Urban Freight	488
Urban Freight and Urban Deliveries—Characteristics and Volumes	488
The Technical Evolution of the Cargo Bike	489
The Development of Cycle Logistics and New Delivery Concepts	491
Infrastructure for Bicycle Freight	491
Experiences from Projects and Practical Examples	492
The Potential of Bicycles for Urban Freight	493
Outlook	493
References	493

Challenges for Urban Freight

Urban traffic is characterized worldwide by high densities and overloading, causing congestion as well as environmental harm such as air pollution and noise. Congestion is not only due to the use of the individual car, but also to commercial road transport that constitutes between 25% and 35% of the overall urban traffic in the cities of the industrialized world (Menge and Horn, 2014), thus contributing significantly to total urban transport emissions; A quarter of CO₂ emissions, a third of NO_x emissions and half of particulate matter emissions (Dabanc and Rodrigue, 2017). Urban freight has been increasing continuously during the last decade, with a particular growth in the number of parcel deliveries that constitute by far the largest share of urban deliveries. In Germany, for instance, the number of parcel deliveries in 2018 was about 3.5 billion (based on BIEK, 2019), in the UK it was 3.65 billion in the same year (Mintel, 2019), in France, 1.6 billion (Ecommerce News, 2019). Besides an increase in vehicles circulating on the streets, the growing amount of freight deliveries also generates obstacles as delivery vehicles are standing on the street and impair the traffic flow.

Against this background, the EU has set an ambitious goal for urban freight: in 2011 the Commission declared to “achieve essentially CO₂-free city logistics in major urban centers by 2030” (European Commission, 2011). Solutions to reduce CO₂ emissions and other environmental harm of urban freight deliveries are the optimization of delivery operations, new delivery options (for instance, pickup locations with parcel lockers) or the use of environment-friendly vehicles like electric vans or cargo bikes. There is no standard solution, but projects and “real-life” application of cargo bike delivery have shown that concepts fulfilling both the requirements of freight service providers and the expectations of cities are manageable by case-specific adaptations.

In addition, cargo bikes may not only be used for logistics services in the urban context, but also for internal, on-site transport of firms, for service trips or the transport of materials and tools that, for instance, a craftsman or a chimney sweep needs for his or her work, or for private purposes to transport errands (or children). This article, however, will focus on bicycles for urban freight in the logistics sector.

Urban Freight and Urban Deliveries—Characteristics and Volumes

There is no general definition for urban freight. As stated by Behrends et al. (2008), different perspectives may be applied to define “urban freight”—location, actors, or products. They use a definition that focuses on the combination of goods and their receivers, but also includes locational aspects, and suggest the following list:

- provision of industry with raw materials and semi-manufactured articles;
- provision of the wholesale trade with consumer goods;
- provision of shops with consumer goods;
- inbound and outbound consumer goods produced in the area;
- home deliveries made by professional delivery operators; and
- transit transport of goods.

There exists hardly any data about the shares in urban freight of different categories of goods. Rough estimations by Sonntag (2015), however, indicate for the year 2010 that a share of 38% of all truck trips in urban commercial transport in Germany have been carried out by courier, express and parcel (CEP) services. This is followed by an 18% share for wholesale and retail logistics, 15% freight forwarding/general cargo, 14% construction and building craft, 10% specialists’ trips, and 5% provision from industry and manufacturing.

Table 1 CEP volumes in German cities in 2016

	<i>Regional level</i>	<i>CEP deliveries (million) in 2016</i>	<i>Share of all German CEP deliveries</i>	<i>CEP deliveries per year per inhabitant</i>
Germany	National	3,160	100.0 %	38
Bavaria	State (Bundesland)	540	17.1 %	42
Munich	City	89	2.8 %	61
Berlin	City	130	4.1 %	37
Hamburg	City	90	2.9 %	51
Stuttgart	City	37	1.2 %	60
Düsseldorf	City	36	1.2 %	59

Source: BIEK (2018b)

The list of goods in urban freight reveals that the total of “urban freight” incorporates a broad range of goods, including goods for which cargo bike transport would not make sense or even be possible, due to obvious restrictions in driving range, carrying capacity, or payload volume. Other “typical” goods for bicycle messengers, such as legal documents or tickets, are increasingly becoming digitalized. Nonetheless, there is still a large variety of feasible goods, such as spare parts, medical samples, fresh produce transported in small load carriers.

The CEP services market is a main field of application for bicycle logistics, as it is characterized by lightweight shipments. In Germany, the average weight of a CEP shipment was 7.4 kg in 2016 and the total annual transport volume of the CEP industry was 23.5 million tons (BIEK, 2018a), which is less than 1% of the total road freight transport in Germany. Nonetheless, looking again at the 38% share of all urban commercial truck trips caused by CEP services, the need for action becomes evident.

With this general focus on parcels as the main type of freight, delivery concepts that use bicycle transport can be divided into four major groups:

1. Point-to-point/courier shipments
2. Express parcel deliveries (via microhub)
3. Home deliveries (e.g., grocery box, cooked meal)
4. Own-account/internal freight transport.

Cities are typically the top priority of CEP services. These markets have considerably changed during the last decade, above all by the huge increase in parcel deliveries to households, that is, B2C deliveries (business to consumer). The main driver has been online shopping for which cities are hot spots: The per capita delivery in German cities is far above the national or state average (Table 1). In Munich (capital of the German state of Bavaria), for instance, the per capita number is 61 while it is “only” 42 for Bavaria and 38 as the national German average (BIEK, 2018b). This underlines the particular challenge for alternative solutions in the urban delivery sector.

The Technical Evolution of the Cargo Bike

Due to EU regulation, cargo bikes commonly used for urban freight transport are legally equal to “conventional” bicycles (with all rights and obligations) as long as they do not exceed a pedal-assist engine of 250 W and an assisted speed of 25 kph (European Commission, 2020). Besides this legal classification, however, cargo bikes are certainly “more” than just bicycles. Other than a conventional bicycle, they offer transport possibilities by fixed cargo boxes, flexible containers or open cargo platforms of different sizes and different shapes. Dozens of construction types of cargo bikes exist: there are cargo bikes with 2, 3, 4, or even more wheels, diverse locations of the cargo and ways of steering. Information on the five main cargo bike construction types, as well as bicycle trailers can be found in Table 2.

The use of cargo bikes to transport goods has a long history. By the end of the 19th century, the high wheel with two wheels—a vehicle that was quite difficult to handle—was developed into a more robust three-wheel vehicle that offered new options for freight transport (Dodge, 2011; Rauck et al., 1979). In the beginning of the 20th century, the “Long John bike” was developed in Denmark. Cargo bikes were commonly used for home deliveries of milk or other goods and specific cargo bikes were developed for several purposes, for instance for postal services or for ambulance services. Due to the spread of automobiles, cargo bikes became nearly extinct for freight transport in the second half of the 20th century and were predominantly used for private mobility in some cities such as Copenhagen.

The recent development of cargo bikes has focused on the extension of transport weight and volumes, the comfort of the driver, but most importantly on new propulsion systems. The growing demand for conventional e-bikes (electrically-assisted bicycles) led to the development of powerful and reliable electric engines and flexible battery systems. This can be seen as one key factor for the renaissance of commercially applicable cargo bikes, together with a change of awareness toward the negative externalities of rapidly increasing urban freight trips. One of the most recent innovations to improve the propulsion of cargo bikes is H₂/fuel cell technology

Table 2 Characterization of typical construction types of cargo bikes

No. of wheels	Construction type	Side view	Payload volume	Carrying capacity Q1–Q3 range ^a	Vehicle characterization, driving behavior, features and typical transport tasks
2	Pizza delivery bike		90 L (back) + 40 L (front)	50–79 kg (n = 17)	- One track vehicle similar to conventional bicycles concerning frame and driving behavior - High cargo areas above front wheel and rear wheel - Feasible for small transports covering all trip distances
2	Long John bike		150–300 L	80–100 kg (n = 23)	- One track vehicle with extended wheelbase and low cargo area in front of the driver - Indirect steering of the smaller front wheel by thrust rod or cable pull - Longer but not necessarily wider than a conventional bicycle, unusual driving behavior - For fast light/medium-weight transports covering all trip distances
2	Longtail bike		Typically open platform (no closed cargo box)	70–105 kg (n = 8)	- One track vehicle with extended wheelbase and cargo area behind the driver - No sight obstruction even when transporting bulkier goods - Longer but not necessarily wider than a conventional bicycle, familiar driving behavior - For fast light/medium-weight transports covering all trip distances
3	Tricycle, front load		330 L	80–120 kg (n = 44)	- Two track vehicle with low cargo area in front of the driver - Wider than conventional bicycles - Traditional models with turntable steering: tip-over protection only whilst standing, no fast turns possible; for less time-critical medium-weight transports or multi-stop deliveries covering short and medium trip distances - Models with Ackerman steering (tilting technology: Considerably more agile and faster turning velocities; feasible for long distances, too)
3 (4)	Heavy-load tri-cycle (quadricycle)		1300–2300 L	125–200 kg (n = 27)	- Two track vehicle with cargo area behind the driver for high payloads, usually compatible with EPAL pallets - Tip-over protection whilst standing, slow operative velocities; for large and heavy-weight transports of 100 kg or more - Some models include weather protection covers
2-3	Bicycle trailer		1300–1700 L or open platform	(around 150 kg)	- Trailers are compatible with most conventional bicycles and cargo bikes with single rear wheel - Some models include electric assist of up to 25 kph while towing as well as up to 6 kph when operated as handcart using the trailer's drawbar

^aCarrying capacity: Based on a market research by Schenk et al. (2017) including carrying capacity information of 121 cargo bike models. Stated is the range between first and third quartile and the number of models.

Source: Own representation based on Behrensen and Gruber (2017) and Schenk et al. (2017).

that can provide a higher energy density than batteries. Furthermore, the hydrogen tank can be refueled within seconds and suffers no performance loss at low temperatures (DLR, 2020).

In the last decade, many cargo bikes have become available that allow for payloads far beyond the 25 kg of standard bikes, thus significantly opening up the potential range of tasks, and cargo bikes have started to satisfy the rising demand for city center point-to-point small-scale consignment delivery. Today's cargo bikes can operate 100–200 kg of total carrying capacity and up to two cubic meters of payload volume; they may have a container designed for the transport of a standard EPAL pallet. This significant increase in options and performance fostered the integration of cargo bikes into urban delivery concepts. Carrying capacity needs to be checked in accordance with the permissible gross vehicle weight, which also includes weight of the vehicle and its driver. Although a few cargo bikes may allow for loads up to 500 kg, their main purpose is to deliver parcels that have a weight of up to 31.5 kg and a maximum size of $120 \times 60 \times 60$ cm, but far less in most of the cases, as bicycles are operated by individual persons without assisting lifting devices. According to the German bicycle industry association ZIV, the amount of electric cargo bikes sold in Germany has risen from 10,700 in 2015 (there is no earlier data) to 39,200 in 2018 (ZIV, 2019).

The Development of Cycle Logistics and New Delivery Concepts

In Germany, the 1990s saw a peak in the foundation of bicycle courier companies providing immediate and fast point-to-point courier deliveries in the denser urban context. Most of these conventional business models were the provision of a platform to organize the interface between demand and individual couriers operating with a bicycle and a backpack. While the use of human-powered bicycles and cargo bikes will most likely remain limited to the courier market, the abovementioned electrification, as well as diversification of cargo bike models, allowed for an expansion of cycle logistics.

The use of cargo bikes for parcel shipments is often embedded in new delivery concepts that aim at reducing the environmental impact of urban freight and improving urban traffic flows, while simultaneously decreasing the cost for delivery service providers. An integral part of these concepts are urban distribution centers (micro-depots, microconsolidation centers, city cross docks). They are regarded as the central element of the new urban logistics concepts (Conway et al., 2011; Dablanc, 2011; Koning and Conway, 2014; Lenz and Riehle, 2013). Freight is carried to micro depots by larger motorized vehicles and then redistributed to smaller vehicles that provide a more environment-friendly distribution than conventional vehicles do. These small distribution centers are located close to, or even within, the inner city and may be installed as either permanent infrastructure (CityLog, 2012) or mobile infrastructure that allows for flexibility on a daily basis. Mobile depots are mostly containers or truck trailers that are located in designated areas (Verlinde et al., 2014).

Depending on the specific urban context in which the new concept is applied, it may also be combined with modes other than the truck to fill (or bring from) the depot; examples exist for the combination of ship and cargo bikes in central Paris (Hotel logistique Porte de la Chapelle). For the final 50 feet of the logistics chain, other low emission vehicles may be used as an alternative to cargo bikes, as for instance hand trucks that are used for the very short distances from curbside to the receiver or within the receiver's building (cf. University of Washington, 2020).

Besides point-to-point courier shipments and parcel shipments via microdepot, bicycles are often used for meal delivery or retail home delivery, both by large retail chains and niche providers (such as shippers of regionally grown fresh produce). Some "brick and mortar" stores started to collaborate and enter hybrid distribution channels by offering online order and emission-free last-mile delivery by a shared cargo bike.

The strengths and weaknesses of cargo bikes might also be seen as connected to the re-urbanisation of logistics areas (e.g., by Amazon's urban hubs) and the creation of more flexible delivery options, such as same day delivery or instant delivery (within the next 2 h). Cargo bikes are often implemented as parts of these concepts as they are less sensitive to congestion in the city centers. The most challenging element for these concepts is the cost decrease: Depending on the type of cargo bike and the type of conventional delivery vehicle, the payload volume of the conventional last-mile delivery vehicle is equal to 15–20 cargo bikes. This means that the switch to a concept which uses cargo bikes for a direct substitution of conventional delivery vehicles requires either more vehicles and drivers or a broader time window for deliveries (Raiber, 2015). From a business-economic perspective, the marginal transport costs for bikes today are higher than those of light commercial vehicles (€4.61 vs. €3.40 per stop), mostly due to higher labor costs. Using a welfare-economic viewpoint including the external costs (e.g., for emissions, noise and accidents), the balance is reversed and the marginal costs for bike deliveries would be lower than those of conventional vehicles (€4.52 vs. €6.05) (Maes, 2017). Another estimation looks at the cost parameters of cargo bikes for instant deliveries using a micro-depot and finds them economically viable compared to conventional approaches (Rudolph et al., 2018). It is obvious that business models implementing bicycles are part of dynamic markets both with expanding businesses, but also failures. In 2019, for instance, both Foodora and Deliveroo withdrew from the German meal delivery market (Fig. 1).

Infrastructure for Bicycle Freight

Cargo cycles have similar infrastructural demands as private bicycle transport, but there are some particularities (Conway et al., 2011; Gruber and Rudolph, 2016):

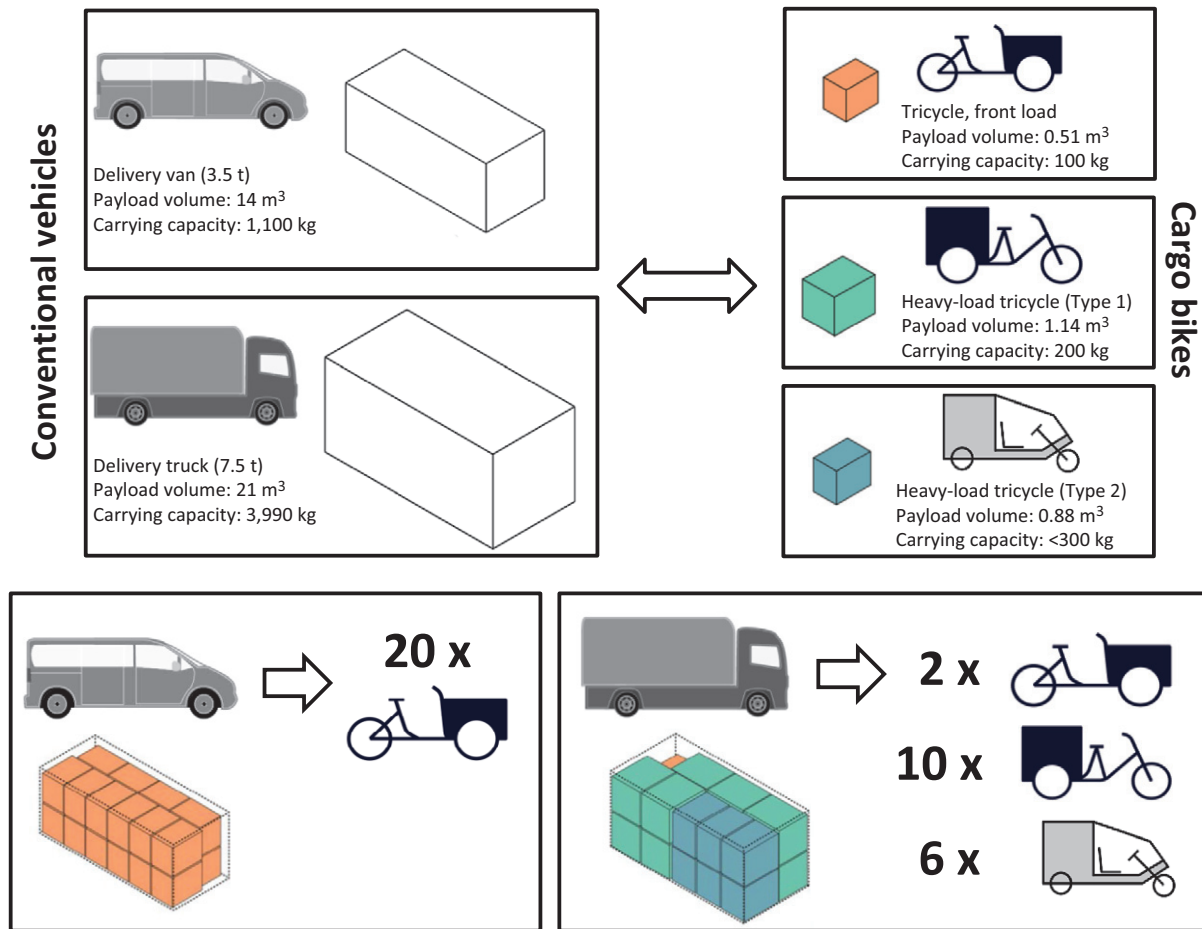


Figure 1 Comparison of payload volumes and carrying capacities of conventional vehicles and cargo bikes. *Source:* Own representation, based on Raiber (2015).

- As transporting goods is usually time-sensitive, cargo bikes often run at relatively high velocities.
- Broader spectrum of velocities.
- Larger cargo cycles need more space for parking.
- Heavy loaded cargo cycles have longer braking distances and larger turning radii.
- Higher vulnerability regarding slopes, steps and vibrations.

In addition, the implementation of cargo cycle concepts has specific requirements concerning the availability of inner-urban spaces and the design and usability of the urban (road) space (Conway et al., 2011; Gruber and Rudolph, 2016):

- There must be enough space to put up depots; this space needs to be permanently assigned for the new logistics concept, at least on weekdays.
- The depots—if permanent or mobile—must be accessible for trucks that deliver the freight.
- Ramps for cargo bikes must be provided to protect them from weather effects and vandalism.
- In case of electric cargo bikes loading infrastructure is needed (Raiber, 2015).
- Planning manuals for roads design should enclose the dimensions of cargo cycles.
- State-of-the-art infrastructure is needed.
- Redesigning the complete bicycle infrastructure of cities is neither possible nor necessary, also small constructive adjustments may have high impacts.
- Using cargo cycles should be legal on all urban roads.

Experiences from Projects and Practical Examples

In Germany, 800 companies and public institutions participated in a large-scale cargo bike testing scheme (project “Ich entlaste Städte”) with a total operative mileage of approximately 300,000 km. Without cargo bikes, two out of three of these kms would have

been conducted with a combustion engine vehicle. Around 30% of the testers acquired their own cargo bike after the trial. The project was particularly popular among self-employed persons, freelancers and small companies providing deliveries or services.

In the city of Hamburg, a CEP service provider (UPS) operates four microdepots to service the inner city area by cargo bike and handcarts. In the morning and in the evening, a truck trailer with parcels is brought to, or picked-up at, each microdepot location. An evaluation showed that 4–6 conventional 7.5 t delivery vehicles are substituted by 7 heavy-duty trikes, 4 front-load trikes and a number of handcarts. The before-after assessment of emissions showed a decrease from 25–30 t to 11 t of CO₂ and a decrease from 85–101 kg to 39 kg of NO_x (Ninnemann et al., 2017). UPS themselves state a commercially viable substitution of up to 10 conventional delivery vehicles by the ongoing concept. In Berlin, several parcel service providers participated in the project “KoMoDo”. While there was no company-overarching consolidation of shipments, it can still be viewed as one of the first examples where competing parcel companies collaborated under the roof of a municipal operator, in this case the Berlin freight port operator Behala.

A study on the use of heavy-duty bikes at two CEP services (DPD and GLS) in Nuremberg found an economically viable substitution of seven conventional delivery vehicles by eight cargo bikes, saving around a quarter of greenhouse gas and air pollutants. To this end, the authors carried out a micro-simulation with cargo bikes as part of a microdepot concept based on data of the current state, that is, conventional parcel delivery. On the basis of unspecified internal calculations, the authors indicate a potential shift from fuel-based vehicle to cargo bike of 30% of the total CEP deliveries in cities with over 100,000 inhabitants (Bogdanski et al., 2017).

The Potential of Bicycles for Urban Freight

As it was expressed earlier, cargo bikes are not the solution for urban freight in general, but rather for specific purposes within the urban context. Assessments concerning the potential of substitution for conventional vehicle use and, thus, the environmental effects of cargo bike use display considerable differences in their results. Most studies are based on the number of trips and kilometers driven per trip that currently are made by motorized vehicles (motorcycles, cars and light duty vehicles up to 3.5 t) compared to what can be substituted by cargo bikes.

Studies that take into account broader market segments of cargo bike use estimate that the overall substitution rate of cargo bikes for motorized vehicles will be between 8% and 23% of trips and between 1% and 4% of the mileage. Besides the influence of the share of electrification in the cargo bike fleet on the total substitution, a major factor is the readiness of drivers and companies to use cargo bikes and—directly related to this—to restructure logistics operations (Gruber and Rudolph, 2016). Although it may seem that the low share of mileage that can be substituted will not be relevant to reduce environmental harm in cities, these reductions occur in particular in inner city areas and in residential areas and thus are absolutely relevant on a local scale.

Studies that consider specific market segments, single courier companies (Gruber et al., 2014) or smaller areas inside the inner city (CityLog, 2012) result in much higher substitution potential. Other studies that show high potential of more than 50% of substitution consider all privately and commercially motivated transport of goods (Cyclelogistics, 2014).

Outlook

Cargo bicycles will play an increasingly important role in urban logistics. A major driver will be the cities' growing demands on environmentally friendly logistics, which logistics service providers will have to take into account in the future. In view of a European context in which the environmental impact of (urban) delivery traffic is more and more being considered by EU legislation, efforts to innovate are increasing, both in the technological development of cargo bikes and in the implementation of new concepts. The introduction of new framework conditions for inner-city traffic and the use of vehicles that are powered by fossil fuel could result in additional incentives for the use of cargo bicycles.

References

- Behrends, S., Lindholm, M., Woxenius, J., 2008. The impact of urban freight transport: a definition of sustainability from an actor's perspective. *Transport. Plan. Technol.* 31, 693–713.
- Behrens, A., Gruber, J., 2017. Lastenrad-Bauformen und Einsatzmöglichkeiten <https://www.lastenradtest.de/testraeder/> (accessed 6.02.20).
- BIEK (2019): KEP-Studie 2018 – Analyse des Marktes in Deutschland. Bundesverband Paket und Expresslogistik e. V. = German Parcel and Express Logistics Association, Berlin.
- BIEK (2018a): Kompendium Teil 1: Transportaufkommen und durchschnittliches Gewicht. Bundesverband Paket und Expresslogistik e. V. = German Parcel and Express Logistics Association, Berlin. <https://www.biek.de/download.html?getfile=1551>. Accessed 05/02/20.
- BIEK (2018b): Kompendium Teil 5: Regionale Verteilung des KEP-Sendungsvolumens. Bundesverband Paket und Expresslogistik e. V. = German Parcel and Express Logistics Association, Berlin. <https://www.biek.de/download.html?getfile=1564>. Accessed 05/02/20.
- Bogdanski, R., Bayer, M. & Seidenkranz, M. (2017). Pilotprojekt zur Nachhaltigen Stadtlogistik durch KEP-Dienste mit dem Mikro-Depot-Konzept auf dem Gebiet der Stadt Nürnberg. https://www.c-na.de/fileadmin/templates/global/media/Pedelistics/Download/Abschlussbericht_Mikro-Depot-Konzept_Nuernberg.pdf. Accessed 09/09/19.
- CityLog (2012). Background information to the press conference. Press release by Senate administration for urban development and environment on 25.01.2012. URL: <http://www.bentobox-berlin.de/presse/>. Accessed 07/02/20.
- Conway, A., Fatissou, P.-E., Eickemeyer, P., Cheng, J. & Peters, D. (2011). Urban Micro-consolidation and last mile goods delivery by freight-tricycle in Manhattan: Opportunities and challenges. In *Proceedings of the 91st Transportation Research Board Annual Meeting*. Washington D.C., 22–26.

- Cyclelogistics (ed.) (2014): D6.9 Final Public Report. Public Report about the CycleLogistics Project 2011-2014. http://one.cyclelogistics.eu/docs/119/D6_9_FPR_Cyclelogistics_print_single_pages_final.pdf. Accessed 13/12/19.
- Dabian, L., 2011. City distribution, a key element of the urban economy: guidelines for practitioners. In: Macharis, C., Melo, S. (Eds.), *City Distribution and Urban Freight Transport: Multiple Perspectives*. Edward Elgar, Cheltenham, pp. 13–36.
- Dabian, L., Rodrigue, J.-P., 2017. The geography of urban freight. In: Giuliano, G., Hansen, S. (Eds.), *The Geography of Urban Transportation*. The Guilford Press, New York, London, pp. 34–56.
- Dodge, P., 2011. *Faszination Fahrrad: Geschichte -Technik -Entwicklung*. Delius Klasing, Bielefeld.
- DLR Institut für Fahrzeugkonzepte (2020): Cargo-Pedelec mit modernster Brennstoffzellentechnologie. URL: https://www.dlr.de/fk/desktopdefault.aspx/tabid-3731/5826_read-41856/. Accessed 08/02/20.
- Ecommerce News (ed.) (2019). Europe: 9.3 billion parcels in 2018. URL: <https://ecommercenews.eu/europe-9-3-billion-parcels-in-2018/>. Accessed 06/02/20.
- European Commission (2011). Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system. White Paper. URL: <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52011DC0144&from=EN>. Accessed 20/07/18.
- European Commission (2020): Directives and regulations - two- and three-wheel vehicles and quadricycles. URL: https://ec.europa.eu/growth/sectors/automotive/legislation/motorbikes-trikes-quads_en. Accessed 06/02/20.
- Gruber, J. & Rudolph, C. (2016): Untersuchung des Einsatzes von Fahrrädern im Wirtschaftsverkehr (WIV-RAD). URL: <http://www.bmvi.de/SharedDocs/DE/Anlage/VerkehrUndMobilitaet/Fahrrad/wiv-rad-schlussbericht.pdf>. Accessed 05/02/20.
- Gruber, J., Kihm, A., Lenz, B., 2014. A new vehicle for urban freight? An ex-ante evaluation of electric cargo bikes in courier services. *Res. Transport. Bus. Manage.* 11, 53–62.
- Koning, M. & Conway, A. (2014). Biking for goods is good: An Assessment of CO2 savings in Paris. *Proceedings of the 94th Transportation Research Board Annual Meeting*.
- Lenz, B., Riehle, E., 2013. Bikes for Urban Freight? *Transport. Res. Rec.: J. Transport. Res. Board* 2379, 39–45.
- Maes, J., 2017. The Potential of Cargo Bicycle Transport as a Sustainable Solution for Urban Logistics. University of Antwerp, Antwerp.
- Menge, J., Horn, B., 2014. In: Bracher, T., Dziekan, K., Gies, J., Holzapfel, H., Huber, F., Kiepe, F., Reutter, U., Saary, K., Schwedes, O.O. (Eds.), *Fahrradnutzung im Wirtschaftsverkehr. HKV – Handbuch der kommunalen Verkehrsplanung*, Berlin.
- Mintel (Eds.), 2019. Delivering the Goods: British Courier and Express Delivery Market Hit £12. 6 billion in 2018. . <https://www.mintel.com/press-centre/retail-press-centre/delivering-the-goods-british-courier-and-express-delivery-market-hit-12-6-billion-in-2018> (accessed 05.02.20).
- Ninnemann, J., Hölter, A.-K., Beecken, W., Thyssen, R., Tesch, T., 2017. Last-Mile-Logistics Hamburg – Innerstädtische Zustelllogistik. Studie im Auftrag der Behörde für Wirtschaft, Verkehr und Innovation der Freien und Hansestadt Hamburg. https://www.hsba.de/fileadmin/user_upload/bereiche/forschung/Forschungsprojekte/Abschlussbericht_Last_Mile_Logistics.pdf (accessed 14.12.19).
- Raiber, S., 2015. Kurzstudie Innenstadtlogistik Stuttgart. In: *Räumliche Wechselwirkungen von Innenstadtlogistikkonzepten am Beispiel des Einsatzes von Lastenrädern in der Paketzustellung*, Industrie- und Handelskammer Region Stuttgart, Stuttgart.
- Rauk, M.J.B., Volke, G., Paturi, F.R., 1979. *Mit dem Rad durch zwei Jahrhunderte: Das Fahrrad und seine Geschichte*. AT-Verlag, Aarau.
- Rudolph, C., Gruber, J. & Liedtke, G. (2018). Simplified scenario based simulation of parcel deliveries in urban areas using electric cargo cycles and urban consolidation centers. 7th Transport Research Arena TRA 2018, April 16-19, 2018, Vienna, Austria.
- Schenk, M., Assmann, T., Behrendt, F., 2017. Intelligente Lastenradlogistik - Stand und Entwicklungsperspektiven für den effizienten logistischen Einsatz in urbanen Systemen. In: Werbeagentur, U. (Ed.), *Jahrbuch Logistik 2017*. unikat Werbeagentur, Wuppertal, pp. 84–89.
- Sonntag, H. (2015). Wirtschaftsverkehr in Metropolräumen – Herausforderungen und Lösungsansätze. Presentation held at IWIT Konferenz – Effizienter Wirtschaftsverkehr durch Verkehrsinformation und Telematik. Wildau, 23 June 2015.
- University of Washington (2020). Final 50 feet research program. URL: <https://depts.washington.edu/sct/cr/research-project-highlights/urban-goods-delivery-0>. Accessed 08/02/20.
- Verlindé, S., Macharis, C., Milan, L., Kin, B., 2014. Does a mobile depot make urban deliveries faster, more sustainable and more economically viable: results of a pilot test in Brussels. *Transport. Res. Procedia* 4, 361–373.
- ZIV (2019): AG Lasten- und Transportfahrräder. Sitzung am 10. Dezember 2019, Zweirad-Industrie-Verband e. V., Berlin.

China's Belt and Road Initiative

Paul Tae-Woo Lee, Maritime Logistics and Free Trade Islands Research Centre at Ocean College, Zhejiang University, Zhoushan, China

© 2021 Elsevier Ltd. All rights reserved.

Introduction	495
Economic and Transport Corridors in the BRI	495
Ports/Terminals Invested in by Chinese Stakeholders Under the “Going Global Strategy”	497
The Polar Silk Road (PSR)	499
Transport Infrastructure Projects in the BRI	499
Connecting China to Europe by Railway	500
Developing City Clustering and Restructuring the Transport System Through the BRI	501
The Asian Infrastructure Investment Bank (AIIB) and Funding for the BRI	502
Concluding Remarks	504
Acknowledgments	505
Biography	505
References	506
Further Reading	506

Introduction

Chinese President Xi Jinping announced the Silk Road Economic Belt (SREB) on September 7, 2013, in Astana, Kazakhstan, and the 21st century Maritime Silk Road (MSR) Initiative on October 3, 2013, in Jakarta, Indonesia. Initially, they were together called “One Belt and One Road (OBOR).” However, the Chinese government officially changed it to the “Belt and Road (B&R)” on September 24, 2015.^a Its more concrete contents and directions have been addressed in the *Vision and Actions on Jointly Building Silk Road Economic Belt and 21st-Century Maritime Silk Road* published on March 28, 2015 (NDRC, 2015), hereinafter referred to as the “BRI document.” As of April 30, 2019, China has signed 187 cooperation agreements on the BRI with 131 countries and 30 international organizations since its inception in 2013.

The BRI document consists of the MSR and the SREB. The former is designed to go from China’s coast to Europe through the South China Sea and the Indian Ocean on one route, and from China’s coast through the South China Sea to the South Pacific in the other. The latter focuses on railways and highways connecting China to Central Asia, Russia, and Europe (the Baltic). Therefore, the BRI aims to “promote the connectivity of Asian, European and African continents and their adjacent seas, establish and strengthen partnerships among the countries along the B&R, set up omni-dimensional, multi-tiered and composite connectivity networks, and realize diversified, independent, balanced and sustainable development in these countries” (NDRC, 2015, p. 2). The BRI document consists of a preface and eight chapters comprising (1) Background, (2) Principles, (3) Framework, (4) Cooperation Priorities, (5) Cooperation Mechanisms, (6) China’s Region in Pursuing Opening-Up, (7) China in Action, and (8) Embracing a Brighter Future Together. Some key words in the BRI document are, among others, China’s inland region, connectivity, economic corridor, transport corridor, city cluster, pilot free trade zone, marine economy development demonstration zone, marine economy pilot zone, core area, bay area, infrastructure, and financial mechanism (Lee et al., 2018a). This article will introduce China’s BRI, focusing on transportation connectivity and infrastructure.

Economic and Transport Corridors in the BRI

There are three ways to connect China to Europe, that is, (1) MSR, which is partly the same as the existing southwest bound maritime routes, (2) SREB, which is mostly the same as the existing rail routes via the Trans-China Railway (TCR) and the Trans-Siberian Railway (TSR), and (3) the Polar Silk Road (PSR). Recognizing China’s interest in developing Arctic shipping routes, in January 2018 the Chinese government announced a white paper on “China’s Arctic Policy” to establish a comprehensive B&R as the “three Silk Roads,” referring to the joint promotion of land (Belt), sea (Road), and the Arctic (Fig. 1).

Figs. 2 and 3 provide a summary of the economic and transport corridors outlined in the BRI document.^b These are as follows:

- China-Pakistan Economic Corridor (CPEC): Kashgar (Xinjiang Uygur autonomous region)—Gwadar port (Pakistan).

^a The Chinese government also announced that strategy, project, program, and agenda should not replace “initiative” (<http://news.sina.com.cn/c/sz/2015-09-24/doc-ifxhzevf1026412.shtml>).

^b Recently, Chinese Belt and Road Portal (2019) designates six international economic cooperation corridors, namely, (1) the New Eurasian Land Bridge, (2) the China-Mongolia-Russia Economic Corridor, (3) the China-Central Asia-West Asia Economic Corridor, (4) the China-Indochina Peninsula Economic Corridor, (5) the China-Pakistan Economic Corridor, and (6) the Bangladesh-China-India-Myanmar Economic Corridor (<https://eng.yidaiyilu.gov.cn/qywx/rdxw/88408.htm>).

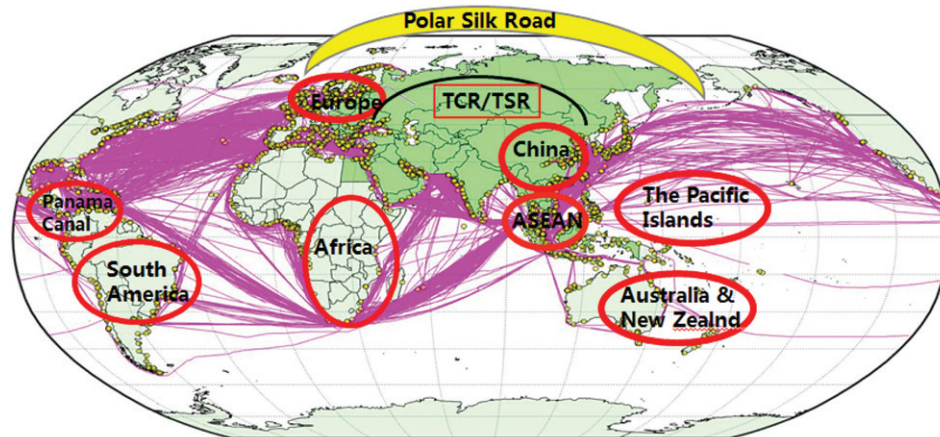


Figure 1 Connecting China to the rest of the world. Note: The width of the pink color indicates the amount of ship traffic flows. Source: Modified by the author based on Lee et al. (2016).

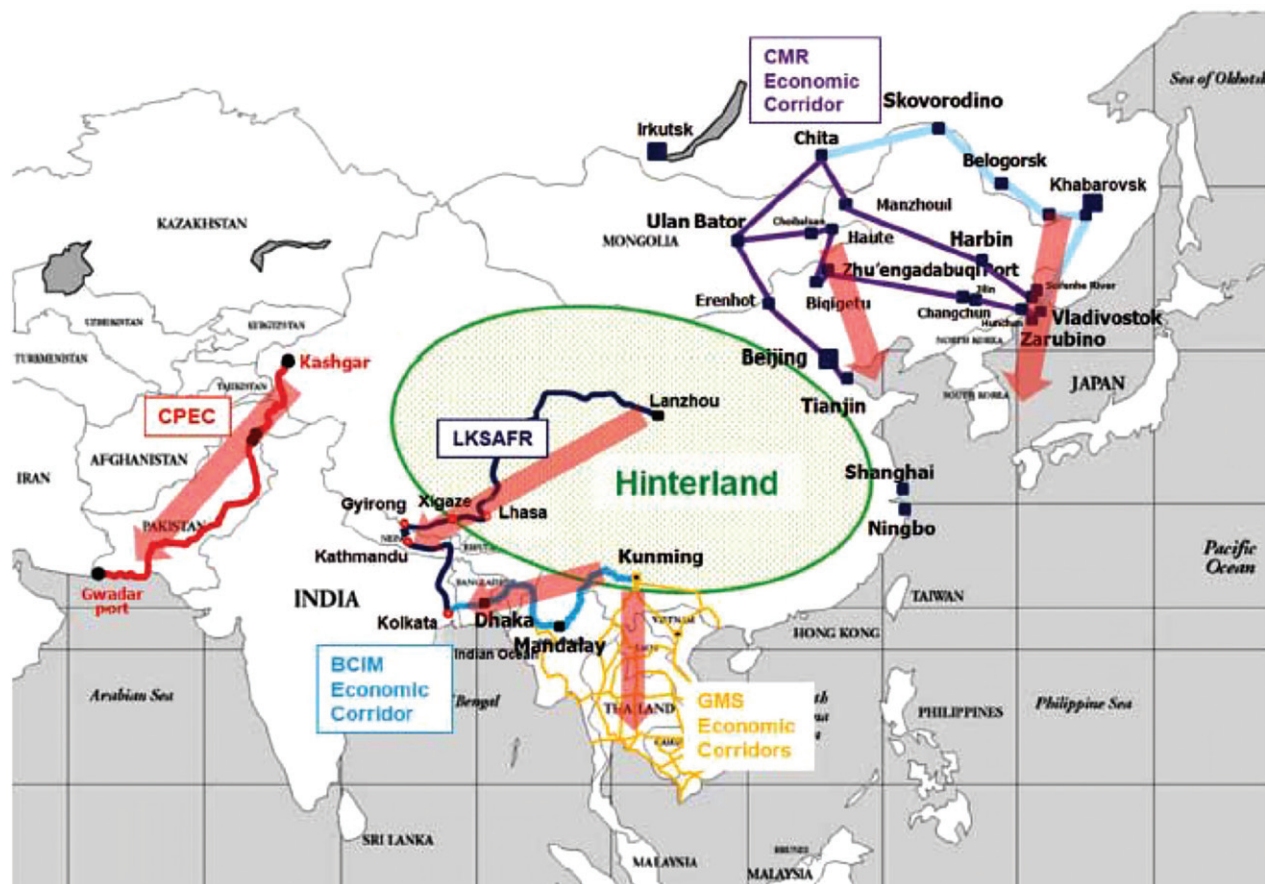


Figure 2 Summary of corridors in the Belt and Road (B&R). Source: Lee (2016).

- Bangladesh-China-India-Myanmar Economic Corridor: Kunming (Yunnan province, China)—Mandalay (Myanmar)—Dhaka (Bangladesh)—Kolkata (India).
- Four sub-corridors in Greater Mekong Subregion Economic Cooperation (GMSEC).
- China, Mongolia, and Russia Economic (CMR) Corridor: Heilongjiang Silk Road Belt.
- Beijing–Moscow Eurasian high-speed transport corridor: Beijing (China)—Khabarovsk (Russia)—Irkutsk (Russia)—Yekaterinburg (Russia)—Moscow (Russia)—Astara (Azerbaijan).
- Lanzhou–Kathmandu South Asia Freight Rail (LKSAFR).

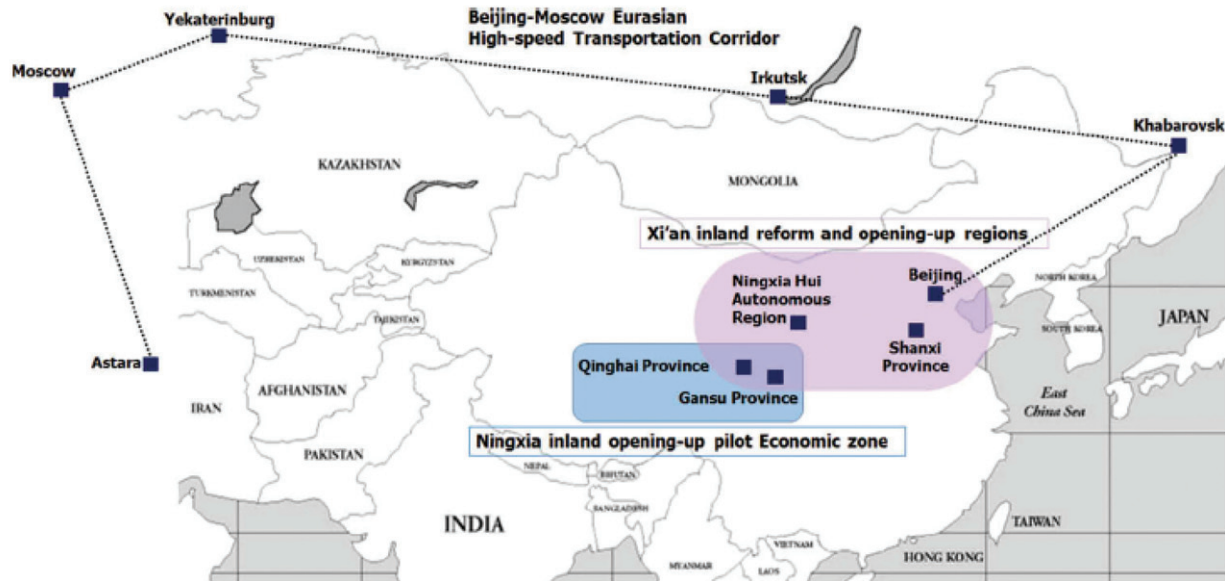


Figure 3 Beijing–Moscow Eurasian high-speed transport corridor. Source: Lee (2016).

The expected impacts of the corridors are multidimensional and include, but are not limited to, a change of energy supply chain, Chinese global port chain development, dry port development in inland China, inter-port competition, infrastructure development, and so on (Lam et al., 2018; Lee et al., 2018a; Wei et al., 2018; Chen et al., 2019). Out of the earlier six corridors, CPEC is expected to play a key role in generating port regionalization and network reconfiguration across the ports in the Indian Ocean. This is because Hambantota Port in Sri Lanka has been leased to China for 99 years, in tandem with the development of Colombo port in Sri Lanka and Gwadar Port in Pakistan within the context of the BRI (Lee et al. 2018a; Ruan et al., 2019). The planned 18 billion US dollars investment for building the CPEC includes the Karot hydropower project, which is the first project of the Silk Road Fund (SRF). In addition, the CMR economic corridor (the so-called Heilongjiang Silk Road Belt) is interrelated to Korea's "New Northward Policy," Russia's "New East Asian Policy," and the maritime logistics connectivity of ports and shipping networks in the East Sea Economic Rim within the context of the BRI, in which the development of trade transit transport corridors is critical in the northeast Asian region (Lim et al., 2017). All the aforementioned are interwoven within the BRI context in the northeast Asian region.

Ports/Terminals Invested in by Chinese Stakeholders Under the "Going Global Strategy"

China's seaports play a key role in connecting China to the rest of the world along the MSR. Therefore, President Xi Jinping has described a port as an "important pivot point" and an "important hub" on a number of occasions in international fora in order to highlight their importance. In addition, China claims the BRI as a growth enabler in the world economy, in association with transport infrastructure. Data from the Ministry of Commerce show that Chinese enterprises invested a total of 14.53 billion US dollars in countries along the B&R in 2016. As of the end of 2016, Chinese enterprises set up 56 inchoate cooperation zones in countries along the B&R, with an accumulated investment totaling 18.55 billion US dollars.

Fig. 4 shows 15 coastal city developments along the Chinese coastline. The coastal cities are Shanghai, Tianjin, Ningbo-Zhoushan, Guangzhou, Shenzhen, Zhanjiang, Shantou, Qingdao, Yantai, Dalian, Fuzhou, Xiamen, Quanzhou, Haikou, and Sanya. The 15 ports try to actively promote the "Going Global Strategy" (GGS), and to respond to China's BRI by becoming key gateways (Fig. 1). They are connected by two major hub airports, that is, Shanghai and Guangzhou. This strategy aims to connect China to the rest of the world by strengthening the functions of their international hub airports and by developing the two airports for efficient hub-and-spoke operations with the coastal cities.

The GGS is being implemented by three players in China, that is, COSCO Shipping and Ports, China Merchants Group, and Chinese local port groups. Tables 1–3 show the foreign ports/terminals in which these three players have invested in along the MSR since the inception of the BRI in 2013. Most of the ports/terminals have been acquired by them through acquisition and public-private partnership (PPP).



Figure 4 Coastal ports under the “Going Global Strategy” along the MSR. Source: Lee (2016).

Table 1 Foreign ports/terminals invested by COSCO shipping ports

Foreign port/terminal	Year	Country	Share %	Shareholding method
Zeebrugge Terminal	2017	Belgium	100%	Acquisition
	2014	Belgium	24%	Acquisition
Noatum Port	2017	Spain	51%	Acquisition
Reefer Terminal SPA	2016	Italy	40%	Acquisition
Abu Dhabi Khalifa Port	2016	United Arab Emirates	90%	Concession (PPP)
Euromax Terminal	2016	The Netherlands	47.5%	Acquisition
	2006	The Netherlands	12.5%	Joint venture
Piraeus Port	2016	Greece	67%	Acquisition (PPP)
	2008	Greece	—	Concession (PPP)
COSCO-PSA Terminal	2016	Singapore	—	—
	2003	Singapore	49%	Joint venture
Busan Port	2015	South Korea	20%	Acquisition
Kumport Terminal	2015	Turkey	26%	Acquisition
SSA Terminals (Seattle)	2008	United States	33.33%	Concession (PPP)
Suez Canal Container Terminal	2007	Egypt	20%	Acquisition
Antwerp Port	2004	Belgium	25%	Acquisition

Source: Huo et al. (2019).

Table 2 Foreign ports/terminals invested by Chin Merchant Group

Foreign port/terminal	Year	Country	Share (%)	Shareholding method
TCP	2017	Brazil	90%	Acquisition
Hambantota Port	2017	Sri Lanka	85%	Acquisition
				Acquisition (PPP)
				Concession (PPP)
Kumport Terminal	2015	Turkey	26%	Acquisition
Kyaukpyu Port	2015	Myanmar	—	BOT (PPP)
Zarubino Port	2014	Russia	—	—
Newcastle Port	2014	Australia	50%	Acquisition
				Concession (PPP)
Bagamoyo Port	2013	Tanzania	—	—
PDSA	2013	Djibouti	23.5%	Acquisition (PPP)
Terminal Link	2013	Around the world	49%	Acquisition
Lome Container Terminal	2012	Togo	50%	Acquisition
CICT	2011	Sri Lanka	85%	BOT (PPP)
TICT	2010	Nigeria	28.5%	Acquisition (PPP)
VICP	2008	Vietnam	49%	Joint venture

Source: Huo et al. (2019).

Table 3 Foreign ports invested by Chinese local port groups

Foreign port	Country	Year	Chinese local port group	Shareholding method
DMP & Djibouti FTZ	Djibouti	2016	Dalian Port Group	Joint venture (PPP)
Muara Port	Brunei	2017		Joint venture
Kuantan Port	Malaysia	2015	Guangxi Beibu Gulf Port Group	Acquisition/concession (PPP)
MCKIP	Malaysia	2013		Joint venture (PPP)
Melaka Gateway Port	Malaysia	2016	Shenzhen Yantian Port Group and Rizhao Port Group	Joint venture (PPP)
Jambi Industrial Park Port	Indonesia	2016	Hebei Port Group	—
Vado Port	Italy	2016	Qingdao Port Group	Acquisition
Guinea Boke Port	Guinea	2013	Yantai Port Group	Joint venture
Haifa Port	Israel	2015		Concession (PPP)
Zeebrugge Terminal	Belgium	2010	Shanghai International Port Group (SIPG)	Acquisition

Source: Huo et al. (2019).

The Polar Silk Road (PSR)

The Chinese government has announced *China's Arctic Policy* in January 2018 (State Council Information Office of the People's Republic of China, 2018), in which the PSR is similar to Arctic shipping routes comprising the Northeast Passage, the Northwest Passage, and the Central Passage. The document has the following contents:

- The Arctic situation and recent changes,
- China and the Arctic, and
- China's policy goals and basic principles on the Arctic.

As a result of global warming, the PSR is likely to become an important transport route for international trade. China's policies and positions on participating in the PSR are described as follows:

- Deepening the exploration and understanding of the Arctic,
- Protecting the eco-environment of the Arctic and addressing climate change,
- Utilizing Arctic resources in a lawful and rational manner,
- Participating actively in Arctic governance and international cooperation, and
- Promoting peace and stability in the Arctic.

Therefore, there are three ways available for China to be connected to Europe, that is, the existing southwest bound maritime routes, railway by TCR and TSR, and the PSR. China's interest in developing the PSR arises from her desire to establish a comprehensive B&R utilizing the "three Silk Roads," that refer to the joint promotion of land (One Belt), sea (One Road), and the Arctic (PSR).

Transport Infrastructure Projects in the BRI

Table 4 shows a summary of BRI projects between January 2013 and March 2019. Since its inception in 2013, the BRI has triggered 695 projects in 98 countries. 142 projects of these projects have been completed, 265 projects are under construction, 250 projects are under contract, 14 have been suspended and/or cancelled, and 24 projects are merely signed Memorandum of Understandings (MoUs).

Table 5 shows that out of the 695 projects in 98 countries, 597 projects can be classified as infrastructure projects. Among those, 217 projects are related to transport infrastructure, comprising 32 seaport/terminal projects, 16 airport projects, and 42 railway projects.

Table 4 Summary of BRI projects as of March 2019

Status	Number of project
Completed	142
Contract	250
MOU	24
Suspension/cancellation	14
Under construction	265
Total	695

Source: Compiled by the author based on Clarkson Research (2019).

Table 5 Category of BRI projects as of March 2019

<i>Sector</i>	<i>Number of project</i>	<i>Remarks</i>
Capital investment	73	
Building materials	30	
Energy and chemicals	29	
Metal processing	9	
Other	5	
Infrastructure	597	
Power plants	275	Energy sources for constructing transport infrastructure
Property	81	
Transportation	217	Seaport projects (32), airport projects (16), rail projects (42)
Other	34	
Raw materials	25	
Production	25	
Total	695	In total 98 countries

Source: Compiled by the author based on [Clarkson Research \(2019\)](#).

Table 6 Freight time and cost comparison between shipping routes and China Railway Express

	<i>Yu-Xin-Europe</i>	<i>Rong-Europe</i>	<i>Zheng-Europe</i>	<i>Han-Europe</i>	<i>Su-Man-Europe</i>
Original city	Chongqing	Chengdu	Zhengzhou	Wuhan	Suzhou
Domestic freight time (days)	12	5	4	5	0
Shipping time (days)			25		
Total freight time (days)	37	30	29	30	25
Frequency			At least seven per week		
Inland freight costs (USD/FEU)	645	1126	443	258	—
Shipping cost (USD/FEU)		1500 USD/FEU (A.P. Moller-Maersk A/S (MSK)'s freight)			
Total costs (USD/FEU)	2145	2626	1943	1758	1500

Source: [Jiang \(2018\)](#).

Connecting China to Europe by Railway

As discussed in the previous section, there are three possible routes to connect China to Europe. [Fig. 1](#) shows the existing shipping route, the rail routes via the TCR and TSR, and the Arctic shipping routes (the so-called PSR). The current shipping routes have been well established for logistics providers and carriers for China. The discussion on the railway mode to connect China and Europe requires a prerequisite assumption that the capacity of the railway and demand for the service are to be justified and that modal competition between the railway service and the shipping service in terms of freight and service quality is acceptable. The China-Europe freight rail service network, in the form of the SREB, has connected China with 50 cities in 15 European countries, thanks to the existing TCR. As of February 2019, the total freight rail service reached 14,000 trips. [Table 6](#) shows the five major routes connecting China to Europe, highlighting the comparative data on freight time and cost between the railway service and the shipping service. It should be mentioned that some rail routes are subsidized by Chinese central and regional governments.

[Table 6](#) shows that China has five main railway service routes for Europe by origin city that use an intermodal service. They have a weekly service, with each block train having more than 42 wagons. Their total freight costs range from 1500 to 2626 US\$ per forty-foot equivalent unit container (FEU). The highest freight rate 2626 US\$/FEU by intermodal service is still higher than the lowest railway freight, while the former's total transportation time is longer than the latter's. Apparently, from the viewpoint of the overall aspects of freight cost and service period between China and Europe, the Chinese railway service is more competitive than the shipping service, thanks to the subsidies received from central and local governments. As can be seen in [Fig. 1](#), if all corridors are working well in the future, the size of the hinterland in the green circle would be encroaching upon those of Shanghai and Ningbo Ports ([Lee et al., 2018a](#)). As a result, this would cause more severe inter-port competition between Shanghai and Busan Port to capture transshipment cargoes with price competition ([Anderson et al., 2008](#)).

China has made an effort to establish dry ports in Europe for enhancing the efficient connectivity of railway services between China and Europe. [Fig. 5](#) shows a development strategy to connect China to Europe by establishing a distribution center, the so-called "Great Stone" industrial park of 95 km², in Minsk in Belarus.^c This was invested in by the China Merchants Group and is

^cThis is based on the author's field trip to Lithuania with meetings and interviews on October 7–11, 2018: meetings with Vice Minister of Transportation, Lithuania; Director of Klaipeda port and CEO of Free Trade Zone (FTZ); CEO of Kaunas FTZ; interviews with China Merchants Groups from Minsk, Belarus; and Secretary-General of China International Freight Forwarders Association.

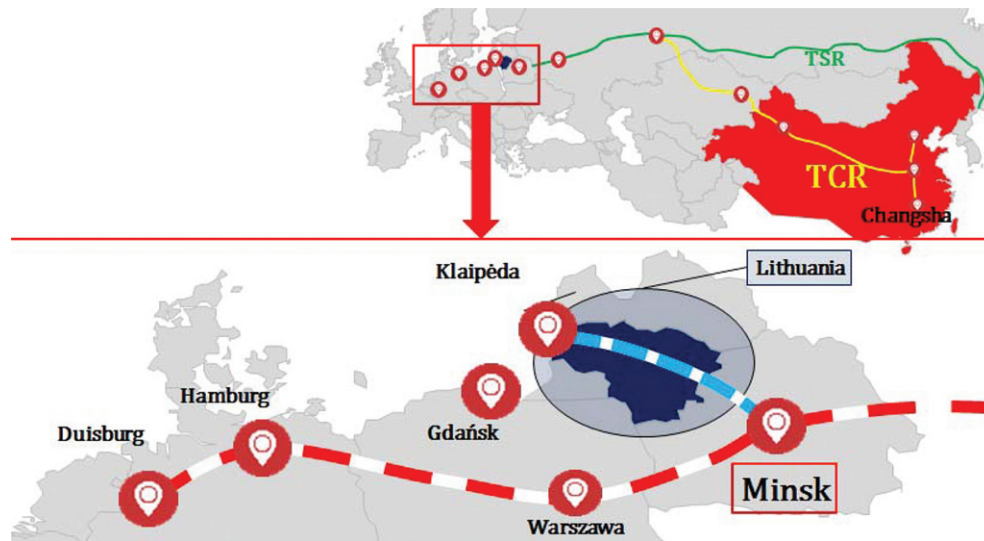


Figure 5 Connecting to China to Europe by railway. Source: Lee (2018).

adjacent to Lithuania, Germany, and Poland. This center aims to provide a distribution service for cargoes coming from China by railway, as well as to accommodate block trains. The characteristics and advantages of the location are as follows:

- A block train service is available between China and Belarus (e.g., Changsha-Great Stone Park).
- The gauge of the Lithuanian rail system is the same as Russia, China, and other CIS countries, so that the railway operator does not need to change the bogie system of the wagons.
- The border crossing time between Lithuania and the Great Stone industrial park in Minsk, Belarus takes 30 minutes for one block train; this is the so-called “Shuttle train Viking.”
- Lithuania offers a free economic zone in Klaipėda and Kaunas ports, with tax incentive policies to attract cargoes from China. Once the cargoes complete customs clearance in the country, they do not need any more customs clearance within the EU market because the country is a member of EU.
- Klaipėda Port expansion by CMG is under negotiation with an MoU.
- Port infrastructure is provided by central government and the superstructure by terminal operators.
- Major shipping lines are calling at Klaipėda, for example, MSC and Yang Ming. COSCO is planning to do the same, subject to China Merchant Groups’ investment decision.
- Minsk is expected to capture cargoes to and from the Scandinavia Peninsula through Klaipėda Port.

Despite the earlier advantages, China’s block train cannot be commercialized for Lithuania because Russia charges 3–4 times more than the normal usage rate of Russian railways for the direct block trains bound for Lithuania from China.

Having recognized the earlier merits and strategic location of Minsk, Lithuania has set up a global logistics hub strategy in the Baltic Region which aims to not only connect China by railway and seaport to Lithuania through the Great Stone industrial park in association with free economic zones in Klaipėda and Kaunas, but also capture cargoes to and from Western Europe and the Scandinavian countries. Therefore, the two countries are together trying to solve the higher usage rate issue of the Russian railway section, because both can get benefits from the direct block train services. This is an example of why governments are motivated to intervene to remove obstacles to the implementation of the BRI (Lee et al., 2018b).

The earlier Lithuania-China case also gives Chinese stakeholders a couple of insights. First, the rail operator in collaboration with logistics providers needs to consider expanding a strategic distribution center in Europe. For example, Klaipėda and dry ports in Lithuania could provide an alternative as the final rail terminus for rail freight movements from China Asia to Europe, instead of Duisburg. This is partly because the whole service from Northeast Asia to Lithuania has the same railway system and partly because Lithuania is an EU member. Last, but not least, the cargo imbalance between China and Europe should be considered. Klaipėda Port has signed an (MoU) on constructing container terminals with China Merchants Group. This may help mitigate the imbalance issue in association with Chinese shipping companies which try to capture cargoes to and from the Baltic and Scandinavian regions and to trigger inter-port competition with Gdansk Port.

Developing City Clustering and Restructuring the Transport System Through the BRI

One of the purposes of the Chinese BRI is to try to promote city clustering (Fig. 6) and introduce more advanced “pilot” free trade zone and big bay area (Fig. 7). The BRI designates the following in this respect: (1) Yangtze River City Cluster,



Figure 6 City clustering in the BRI. Source: Lee (2016).



Figure 7 Introducing pilot free trade zones and big bay areas into the BRI. Source: Lee (2016).

(2) Chengdu-Chongqing City Cluster, (3) Central Henan City Cluster, (4) Hohhot, Bator, Erdos, and Yulin City Cluster, and (5) Harbin-Changchun Metropolitan Area Cluster.

This clustering development aims to:

- set up coordination mechanisms in terms of railway transport and port customs clearance for the China-Europe corridor;
- cultivate the brand of “China-Europe freight trains”;
- construct a cross-border transport corridor connecting the eastern, central, and western regions;
- support inland cities such as Zhengzhou in building airports and international land ports (= dry ports);
- strengthen customs clearance cooperation between inland ports (= dry ports) and ports in the coastal and border regions; and
- launch pilot e-commerce services for cross-border trade.

The Asian Infrastructure Investment Bank (AIIB) and Funding for the BRI

The Asian Infrastructure Investment Bank (AIIB) was launched in January 2016 after 57 countries had signed its charter. The purpose of the AIIB is to foster sustainable economic development, create wealth, and improve infrastructure connectivity in Asia by investing in infrastructure and other productive sectors and to promote regional cooperation and partnership in addressing development

challenges, by working in close collaboration with other multilateral and bilateral development institutions (AIIB Article 1). Its paid capital is 100 billion US dollars. The BRI requires huge investment to construct infrastructure along the B&R (Table 5). Therefore, China has made efforts to establish financial integration as an important underpinning for implementing the BRI, so that China is able to establish a currency stability system, an investment and financing system, and a credit information system in Asia and to maintain the scope and scale of bilateral currency swaps and settlements with other countries along the B&R. The following additional funding sources are available for the BRI:

- Silk Road Fund
- BRICS New Development Bank
- Shanghai Cooperation Organization (SCO) financing institution
- Practical cooperation of China-ASEAN Interbank Association and SCO Interbank Association
- Multilateral financial cooperation in the form of syndicated loans and bank credit
- Bonds issues in both Renminbi and foreign currencies outside China by Qualified Chinese financial institutions and companies
- Bond market in Asia

The AIIB has two main group memberships, that is, “Members” and “Prospective Members.” Each group has regional and non-regional members, respectively. As of June 2019, the Bank has 70 members and 27 prospective members, totaling 97 members worldwide. Tables 7 and 8 show the top 10 regional and non-regional members of the AIIB and their shares and voting power. The top 10 regional members own 63.76 billion US dollars out of 96.4 billion US dollars of the total paid capital of the AIIB, equivalent to 58.3% of the voting power of all the AIIB members, while non-regional members have 22.5 billion US dollars, equivalent to 25.4% of the voting power of the total AIIB. China as a regional member owns the highest share ratio of 29.78%, while Germany as a non-regional member has the highest share of 4.65% of the total non-regional shares.

The BRI has been promoted by multilateral cooperation mechanisms under the Chinese strategy of making full use of existing mechanisms as follows:

- Shanghai Cooperation Organization (SCO),
- ASEAN Plus China (10 + 1),
- Asia-Pacific Economic Cooperation (APEC),
- Asia-Europe Meeting (ASEM),
- Asia Cooperation Dialogue (ACD),
- Conference on Interaction and Confidence-Building Measures in Asia (CICA),
- China-Arab States Cooperation Forum (CASCF),
- China-Gulf Cooperation Council Strategic Dialogue,
- Greater Mekong Subregion (GMS) Economic Cooperation, and
- Central Asia Regional Economic Cooperation (CAREC).

Table 7 Regional members and their shares and voting power of the AIIB

Rank	Regional members	Membership date	Total subscriptions: amount in million USD (% of total)	Voting power: number of votes (% of total)
1	China	December 25, 2015	29,780.4 (30.8913%)	300,338 (26.6167%)
2	India	January 11, 2016	8,367.3 (8.6794%)	86,207 (7.6399%)
3	Russia	December 28, 2015	6,536.2 (6.7800%)	67,896 (6.0171%)
4	Korea	December 25, 2015	3,738.7 (3.8782%)	39,921 (3.5379%)
5	Australia	December 25, 2015	3,691.2 (3.8289%)	39,446 (3.4958%)
6	Indonesia	January 14, 2016	3,360.7 (3.4861%)	36,141 (3.2029%)
7	Turkey	January 15, 2016	2,609.9 (2.7073%)	28,633 (2.5375%)
8	Saudi Arabia	February 19, 2016	2,544.6 (2.6395%)	27,980 (2.4797%)
9	Thailand	June 20, 2016	1,427.5 (1.4808%)	16,809 (1.4897%)
10	Iran	January 16, 2017	1,580.8 (1.6398%)	15,180 (1.3453%)
Subtotal (top 10 regional members)			63,637.3 (66.0113%)	658,551 (58.3625%)
Other regional members ^{a,b}			10,218.0 (10.5991%)	183,261 (16.241%)
Total (regional members)			Amount (million USD): 73,855.3	Number of votes: 841,812
			Percent of total: 76.6104%	Percent of total: 74.6035%
Grand total (total AIIB members)			Amount (million USD): 96,403.8	Number of votes: 1,128,381
			Percent of total: 100.0000%	Percent of total: 100.0000%

^aOther regional members include Afghanistan, Azerbaijan, Bahrain, Bangladesh, Brunei Darussalam, Cambodia, Cyprus, Fiji, Georgia, Hong Kong, China, Israel, Jordan, Kazakhstan, Kyrgyz Republic, Lao PDR, Malaysia, Maldives, Mongolia, Myanmar, Nepal, New Zealand, Oman, Pakistan, Philippines, Qatar, Samoa, Singapore, Sri Lanka, Tajikistan, Timor-Leste, United Arab Emirates, Uzbekistan, Vanuatu, and Vietnam.

^bKuwait was originally a regional member. The country's membership is currently under the category of prospective regional members as founding members of the prospective (Table 9).

Source: Compiled by the author based on AIIB (2019). Available from: <https://www.aiib.org/en/about-aiib/governance/members-of-bank/index.html>.

Table 8 Non-regional members and their shares and voting power of the AIIB

<i>Rank</i>	<i>Non-regional members</i>	<i>Membership date</i>	<i>Total subscriptions: amount in million USD (% of total)</i>	<i>Voting power: number of votes (% of total)</i>
1	Germany	December 25, 2015	4,484.2 (4.6515%)	47,376 (4.1986%)
2	France	June 16, 2016	3,375.6 (3.5015%)	36,290 (3.2161%)
3	United Kingdom	December 25, 2015	3,054.7 (3.1687%)	33,081 (2.9317%)
4	Italy	July 13, 2016	2,571.8 (2.6677%)	28,252 (2.5038%)
5	Spain	December 15, 2017	1,761.5 (1.8272%)	20,149 (1.7857%)
6	The Netherlands	December 25, 2015	1,031.3 (1.0698%)	12,847 (1.1385%)
7	Canada	March 19, 2018	995.4 (1.0325%)	11,888 (1.0535%)
8	Poland	June 15, 2016	831.8 (0.8628%)	10,852 (0.9617%)
9	Switzerland	April 25, 2016	706.4 (0.7328%)	9,598 (0.8506%)
10	Egypt	August 04, 2016	650.5 (0.6748%)	9,039 (0.8011%)
Subtotal (top 10 non-regional members)			19,463.20 (20.1893%)	219,372 (19.4413%)
Other non-regional members ^{a,b}			3,101.54 (3.2003%)	67,197 (5.9552%)
Total (non-regional members)			Amount (million USD): 22,548.5%	Number of votes: 286,569
			Percent of total: 23.3896%	Percent of total: 25.3965%
Grand total (total AIIB members)			Amount (million USD): 96,403.8	Number of votes: 1,128,381
			Percent of total: 100.0000%	Percent of total: 100.0000%

^aOther countries include Austria, Belarus, Canada, Denmark, Ethiopia, Finland, Hungary, Iceland, Ireland, Italy, Luxembourg, Madagascar, Malta, Norway, Portugal, Romania, Sudan, and Sweden.

^bBrazil and South Africa were originally non-regional members. Their membership is currently under the category of prospective non-regional members. They are founding members of the prospective (Table 9). Canada has joined AIIB on March 2018, having seventh voting power among the non-regional members.

Source: Compiled by the author based on AIIB (2019). Available from: <https://www.aiib.org/en/about-aiib/governance/members-of-bank/index.html>.

Table 9 Prospective members of AIIB

<i>Regional prospective members</i>	<i>Non-regional prospective members</i>
Armenia, Cook Islands, Kuwait ^a , Lebanon, Papua New Guinea, and Tonga	Algeria, Argentina, Belgium, Bolivia, Brazil ^a , Chile, Côte d'Ivoire, Ecuador, Ghana, Greece, Guinea, Kenya, Libya, Morocco, Peru, Serbia, South Africa ^a , Togo, Tunisia, Uruguay, and Venezuela

^aFounding member of the prospective membership.

Source: Compiled by the author based on AIIB (2019). Available from: <https://www.aiib.org/en/about-aiib/governance/members-of-bank/index.html>.

China has been playing a constructive role in international fora and exhibitions at regional and subregional levels, hosted by countries along the B&R.

- Boao Forum for Asia
- China-ASEAN Expo
- China-Eurasia Expo
- Euro-Asia Economic Forum
- China International Fair for Investment and Trade
- China-South Asia Expo
- China-Arab States Expo
- Western China International Fair
- China-Russia Expo
- Qianhai Cooperation Forum
- Silk Road (Dunhuang) International Culture Expo
- Silk Road International Film Festival
- Silk Road International Book Fair

Concluding Remarks

The Belt and Road Initiative (BRI) has still a short history of 5 years, which may not be enough to observe its formidable outcome. However, it can be said that its impact on maritime transportation, trade, global logistics patterns, and railway service development between China and Europe are potentially large. In addition, the BRI has been carved into the General Program of the Constitution of the Communist Party of China at the 19th National Congress of the Communist Party of China on October 24, 2017, as follows: "The Party shall constantly work to develop good neighborly relations between China and its surrounding countries and work to



Figure 8 Proposed new Maritime Silk Road. Note: The blue dotted lines depict the MSR. A combined route with red dot lines is called the new MSR. Source: Lee (2016).

strengthen unity and cooperation between China and other developing countries. It shall follow the principle of achieving shared growth through discussion and collaboration, and pursue the *Belt and Road Initiative*." This implies that China has paved a solid path to the implementation of the BRI over the coming years.

Since the inception of the BRI in 2013, 131 countries and 30 international organizations have signed an MoU with China on the BRI. In this circumstance, not a few countries are facing two strategic options for the BRI: either a "Connect Strategy" or a "Being Connected Strategy." The former is proactive, while the latter passive. The freight rail network, with block train services between China and Europe, has diverted a certain amount cargo flow from sea transport to land transport. The PSR has been incorporated into the BRI since January 2018. Therefore, China's BRI has three ways of connecting China to Europe and the rest of the world.

In addition, China's three players, that is, COSCO Shipping Ports, China Merchants Group, and Chinese local port groups, have acquired overseas ports/terminals along the MSR through acquisition and PPIs under the "GGS." This is an interesting development for China in establishing a port supply chain along the MSR. On top of that, dry port development in Europe and in inland China has promoted China's freight rail services. Finally, China's engagements in the sub-Saharan region of Africa and in South America has led to proposal of a new MSR as shown in Fig. 8 (Lee et al., 2018a).

Acknowledgments

The author would like to express his thanks to Korea Maritime Institute for granting copyright permission for the use of his paper that appeared in the *KMI International Journal of Maritime Affairs and Fisheries* for this article.

Biography



Paul Tae-Woo Lee is currently Professor of Maritime Transport and Logistics and Director of Maritime Logistics and Free Trade Islands Research Centre at Ocean College, Zhejiang University in Zhoushan, China. He holds a PhD from Cardiff University in United Kingdom. He was also Visiting Scholar at, among others, Faculty of Economics and Politics in University of Cambridge in United Kingdom, University of Plymouth, Dalian University, Hong Kong Polytechnic University, Chulalongkorn University, the University of the Thai Chamber of Commerce, Shanghai Maritime University, and MPA Visiting Professor at Nanyang Technological University (Singapore). His research interests include maritime transport and logistics, and the Belt and Road Initiative (BRI) issues. He is also a regular speaker at international conferences, including the Asia-Pacific Economic Cooperation (APEC), United Nations Development Programme, UN Economic and Social Commission of Asia and the Pacific (ESCAP), ASEAN-Australia-New Zealand Free Trade Agreement, China Academy of Social Sciences (CASS), and Vietnam Academy for Social Sciences (VASS). In particular, since 2015, he was invited to international conferences, fora, and seminars on the BRI around the world as a well-known speaker. He has published 7 books and more than 300 journal and conference papers, and edited 28 special issues of distinguished international journals, including 10 issues regarding the BRI. He is currently Editor-in-Chief of *Journal of International Logistics and Trade*, and Associate Editor of *Transportation Research Part E* and *Journal of Shipping and Trade*, Book Editor of Elsevier's *China Transportation Series* and Book Editor of Anthem Book Series of *Supply Chain Management* and *Maritime Transport and Logistics*. He is a founding member of Yangtze River Innovation and

Research Belt (Y-RIB, 2018), BRI—International Collaboration Network (BRI-ICN, 2016), Asian Logistics Round Table (ALRT, 2007), and International Association of Maritime Economists (IAME, 1992). He served Co-opt Vice President and Secretary of IAME, including council members for more than 10 years. He is currently Permanent Secretary of the ALRT and Secretary of the BRI-ICN.

References

- AiIB, 2019. Asian Infrastructure Investment Bank (AIIB). Available from: <https://www.aiib.org/en/index.html>.
- Anderson, C.M., Park, Y.-A., Chang, Y.-T., Yang, C.-H., Lee, T.-W., Luo, M., 2008. A game-theoretic analysis of competition among container port hubs: the case of Busan and Shanghai. *Maritime Policy Manag.* 35 (1), 5–26.
- Belt and Road Portal, 2019. Available from: <https://eng.yidaiyilu.gov.cn/>.
- Chen, J., Fei, Y., Lee, P.T.-W., Tao, X., 2019. Overseas port investment policy for China's central and local governments in the Belt and Road Initiative. *J. Contemp. Chin.* 28 (116), 196–215.
- Clarkson Research, 2019. China's Belt and Road Initiative—key project list. Available from: <http://www.clarksons.net/113867>.
- Huo, W., Zhang, W., Chen, P.S.-L., 2019. International ports investment of Chinese port-related companies. *Int. J. Ship. Transp. Logistics* 11 (5), 430–454.
- Jiang, Y., 2018. Evolution of hinterland patterns between China Railway Express and Seaborne Container Shipping under the B&R Initiative. 2019 IAME Conference, Mombasa, Kenya, September 11–14.
- Lam, J.S.L., Cullinane, K., Lee, P.T.-W., 2018. The 21st-century Maritime Silk Road: challenges and opportunities for transport management and practice. *Transp. Rev.* 38 (4), 413–415.
- Lee, P.T.-W., 2016. An anatomy of the “one belt and one road” from the viewpoint of structural changes in maritime transport. 2016 Annual Conference of International Association of Maritime Economists. Kuhne Logistics University, Hamburg Germany, August 23–26.
- Lee, P.T.-W., 2018. Connecting Korea to Europe in the context of the Belt and Road Initiative. *KMI Int. J. Maritime Affairs Fish.* 10 (2), 43–54.
- Lee, P.T.-W., Hu, Z.H., Lee, S.J., 2016. Research trend in the Belt and Road Initiatives. OBOR Conference. RMIT University, Melbourne, Australia, December 1–2.
- Lee, P.T.-W., Hu, Z.-H., Lee, S.-J., Choi, K.-S., Shin, S.-H., 2018a. Research trends and agenda on the Belt and Road (B&R) Initiative with a focus on maritime transport. *Maritime Policy Manag.* 45 (3), 282–300.
- Lee, P.T.-W., Feng, X., Lee, S.-W., 2018b. Challenges and chances of the Belt and Road Initiative at the maritime policy and management level. *Maritime Policy Manag.* 45 (3), 279–281.
- Lim, S.-W., Suthiwartnarueput, K., Abareshi, A., Lee, P.T.-W., Duval, Y., 2017. Key factors in developing transit trade corridors in Northeast Asia. *J. Korea Trade* 21 (3), 191–207.
- NDRC, 2015. Vision and Actions on Jointly Building Silk Road Economic Belt and 21st Century Maritime Silk Road. The National Development and Reform Commission (NDRC), Beijing.
- Ruan, X., Bandara, Y.M., Lee, J.-Y., Lee, P.T.-W., Chhetri, P., 2019. Impacts of the Belt and Road Initiative in the Indian subcontinent under future port development scenarios. *Maritime Policy Manag.*, doi:10.1080/03088839.2019.1594425.
- State Council Information Office of the People's Republic of China, 2018. China's Arctic Policy. State Council Information Office, Beijing.
- Wei, H., Sheng, Z., Lee, P.T.-W., 2018. The role of dry port in hub-and-spoke network under Belt and Road Initiative. *Maritime Policy Manag.* 45 (3), 370–387.

Further Reading

- Dong, G., Zheng, S., Lee, P.T.-W., 2018. The effects of regional port integration: the case of Ningbo-Zhoushan Port. *Transp. Res. Part E* 120, 1–15.
- Liu, M., Kronbak, J., 2010. The potential economic viability of using the Northern Sea Route (NSR) as an alternative route between Asia and Europe. *J. Transp. Geogr.* 18 (3), 434–444.
- Ng, A.K.Y., Jiang, C., Li, X., O'Connor, K., Lee, P.T.-W., 2018. A conceptual overview on government initiatives and the transformation of transport and regional systems. *J. Transp. Geogr.* 71, 199–203.

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

Page left intentionally blank

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

EDITOR-IN-CHIEF

Roger Vickerman

*School of Economics, University of Kent, Canterbury, United Kingdom
and*

Transport Strategy Centre, Imperial College, London, United Kingdom

VOLUME 4

**Traffic Management
Transport Modeling and Data Management**

SECTION EDITORS

Prof. Edward C.S. Chung

*Department of Electrical Engineering, Faculty of Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR*

Chandra R. Bhat

*Center for Transportation Research (CTR)
The University of Texas at Austin, TX, United States*



Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2021 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-08-102671-7

For information on all publications visit our
website at <http://store.elsevier.com>

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

Publisher: Oliver Walter
Acquisitions Editor: Oliver Walter
Content Project Manager: Natalie Lovell
Associate Content Project Manager: Manisha K and Ramalakshmi Boobalan
Designer: Matthew Limbert

EDITORIAL BOARD

Editor in Chief

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Section Editors

Maria Börjesson

Professor of Economics, VTI Swedish National Road and Transport Research Institute; Affiliated Professor at Linköping University, Sweden

Per Gårder

Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States

Prof. Kevin P.B. Cullinane

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

Prof. Edward C.S. Chung

Department of Electrical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

Chandra R. Bhat

Center for Transportation Research (CTR), The University of Texas at Austin, TX, United States

Edoardo Marcucci

Department of Political Sciences, University of Roma Tre, Rome, Italy; Department of Logistics, Molde University College, Molde, Norway

Prof. Maria Attard

Department of Geography, Faculty of Arts, University of Malta, Msida, Malta

Prof. Carlo G. Prato

School of Civil Engineering, The University of Queensland, Brisbane, Australia

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Page left intentionally blank

INTRODUCTION



Roger Vickerman

In an increasingly globalized world, despite reductions in costs and time, transportation has become even more important as a facilitator of economic and human interaction; this is reflected in technical advances in transportation systems, increasing interest in how transportation interacts with society and the need to provide novel approaches to understand its impacts. This has become particularly acute with the impact that Covid-19 has had on transportation across the world, at local, national, and international levels. This Encyclopedia brings a cross-cutting and integrated approach to all aspects of transportation from work in many disciplinary fields, engineering, operations research, economics, geography, and sociology to understand the changes taking place.

Transportation is both influenced by, and an influencer of, changes in the economy and society. Increasing speeds have reduced journey times and made the world a smaller place as globalization has affected both where people live and work and from where they source their goods and materials. Increasing volumes of traffic, often on old and life-expired infrastructure, lead to congestion and delays. Constraints on public budgets have led to increasing pressure on the private sector to fund improvements requiring new and innovative financial solutions. While there are clear differences in the nature of the pressures felt in the developed and less developed economies, there is an increasing recognition that in all societies, there is an accessibility problem such that certain groups become disadvantaged by the lack of access to reliable and cost-effective transport. While the problems are clearly multidimensional, research on transportation is often constrained by single disciplinary approaches and this carries over into the practice of transport planning and policy. The Encyclopedia cuts across these artificial boundaries by taking an approach that emphasizes the interaction between the different aspects of research and aims to offer new solutions to understand these problems. Each of the nine sections is based around a familiar dimension of work on transportation, but brings together the views of experts from different disciplinary perspectives. Each section is edited by an expert in the relevant field who has sought chapters from a range of authors representing different disciplines, different parts of the world, and different social perspectives. In this way, the work is not just a reflection of the state of the art that serves as a starting point for researchers and practitioners, but also a pointer toward new approaches, new ways of thinking, and novel solutions to problems.

The nine sections are structured around the following themes: Transport Modes; Freight Transport and Logistics; Transport Safety and Security; Transport Economics; Traffic Management; Transport Modeling and Data Management; Transport Policy and Planning; Transport Psychology; Sustainability and Health Issues in Transportation. Some of the chapters provide a technical introduction to a topic while others provide a bridge between topics or a more future-oriented view of new research areas or challenges. While there is guidance to cross-referencing between chapters, readers are encouraged to explore the tables of contents of all the sections to get a full understanding of the issues. Much of the Encyclopedia was completed before the Covid-19 pandemic and clearly this will have changed the situation in many areas covered by this work; the advantage of this type of reference work is that relevant updates will be possible in future editions.

The Encyclopedia has only been possible because of the cooperation of a large number of people. Robert Noland, Georgina Santos, Xiaowen Fu, and Dick Ettema served as an Editorial Advisory Board identifying possible editors of sections and advising on the overall structure. The Section Editors, Edoardo Marcucci, Kevin Cullinane, Per Garder, Maria Börjesson, Edward Chung, Chandra Bhat, Maria Attard, and Carlo Prato,

carried out the work of identifying potential authors of individual chapters, commissioning these, encouraging authors and reviewing drafts. More than 600 authors and co-authors of chapters are, however, ultimately responsible for the success of this venture. Thanks are also due to the key people at Elsevier, particularly the Publishers, Andre Wolff and Oliver Walter, and the Project Managers, Sophie Harrison and Natalie Bentahar; they have shown exemplary patience over more than 3 years in bringing this work to fruition.

LIST OF CONTRIBUTORS TO VOLUME 4

Abdul Rawoof Pinjari

Department of Civil Engineering, Centre for infrastructure, Sustainable Transportation and Urban Planning (CiSTUP), Indian Institute of Science, CV Raman Road, Bengaluru, Karnataka, India

Abolfazl Karimpour

Smart Transportation Lab Department of –Civil and Architectural Engineering and Mechanics, University of Arizona, Tucson, AZ –United States

Adam Wilkinson Davis

UC Davis Institute of Transportation Studies, Davis, CA, United States

Adekunle Adebisi

University of Cincinnati, Cincinnati, OH, United States

Amanda Fernandes Ferreira

Department of Civil Engineering, National Taiwan University, Taiwan

Amir Falamarzi

Civil and Infrastructure Engineering Discipline, –School of Engineering, RMIT University, Melbourne, Australia

Anil Kumar

Department of Civil and Environmental Engineering, IIT Patna, Patna, Bihar, India

Ankit Kumar Yadav

Transportation Systems Engineering, Department of Civil Engineering, Indian Institute of Technology (IIT) Bombay, Powai, Mumbai, India

Annesha Enam

Argonne National Laboratory, Lemont, IL, United States

Ashish Bhaskar

School of Civil and Environmental Engineering, Queensland University of Technology, Brisbane, QLD, Australia

Ashish Verma

Department of Civil Engineering, Indian Institute of Science, Bangalore, Karnataka, India

Azusa Toriumi

The University of Tokyo, Tokyo, Japan

Banavar Sridhar

NASA Ames Research Center, Moffett Field, CA, United States

Bart De Schutter

Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands

Becky P.Y. Loo

Department of Geography, The University of Hong Kong, Hong Kong, Hong Kong

Behrang Assemi

School of Built Environment, Queensland University of Technology (QUT), Brisbane, QLD, Australia

Bilal Farooq

Ryerson University, Toronto, ON, Canada

Catherine Morency

Department of Civil, Geological and Mining Engineering, Polytechnique Montreal, Montreal, QC, Canada

Chaitrali Shirke

Queensland University of Technology, Brisbane, QLD, Australia

Chaitrali Shirke

School of Civil and Environment Engineering, Queensland University of Technology, Brisbane, QLD, Australia

Chandra R. Bhat

Center for Transportation Research (CTR), The University of Texas at Austin, TX, United States

Charisma F. Choudhury

Institute for Transport Studies, Leeds, United Kingdom

Chintan Advani

School of Civil and Environmental Engineering, Queensland University of Technology, Brisbane, QLD, Australia

Chunliang Wu

Monash Institute of Transport Studies, Department of Civil Engineering, Monash University, Clayton, VIC, Australia

Cristiano P Garcia
University of Brasilia - UnB, Department of Computer Science - CIC, Brasilia, DF Brazil

Cynthia Chen
University of Washington, Seattle, WA, United States

David Boyce
University of Illinois at Chicago, Chicago, IL, United States

David López
Torre de Ingeniería piso 2-N, Instituto de Ingeniería Av., México

David T. Ory
WSP, New York, NY, United States

Deepa L
Department of Civil Engineering, Indian Institute of Science, CV Raman Road, Bengaluru, Karnataka, -India

Douglas Baker
School of Built Environment, Queensland University of Technology (QUT), Brisbane, QLD, Australia

Dr. Sören Groth
ILS-Research Institute for Regional and Urban Development gGmbH, Büderweg, Dortmund, Germany

Edward Chung
Department of Electrical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

Emma Frejinger
Department of Computer Science and Operations Research and CIRRELT, Université de Montréal, Montreal, QC, Canada

Farhana Naznin
Port Melbourne, Victoria, Australia

Feiyang Zhang
Department of Geography, The University of Hong Kong, Hong Kong, Hong Kong

Felipe de Souza
Argonne National Laboratory, Lemont, IL, United States

Felipe F. Dias
Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX, United States

Frances Sprei
Chalmers University of Technology, Göteborg, Sweden

Frank Witlox
Department of Geography, Ghent University, Ghent, Belgium

Gano B. Chatterji
NASA Ames Research Center, Moffett Field, CA, United States

Gayathri Shivaraman
WSP, New York, NY, United States

Gian-Claudia Sciara
School of Architecture, The University of Texas at Austin, Austin, TX, United States

Giovanni Circella
Institute of Transportation Studies, University of California, Davis, United States; School of Civil and Environmental Engineering, Georgia Institute of Technology, United States

Gowri Asaithambi
Department of Civil and Environmental Engineering, Indian Institute of Technology, Tirupati, India

Guang Yang
School of Transportation, Southeast University, Nanjing, China

H. Michael Zhang
One Shields Avenue, University of California, Davis, CA, USA

Hai Yang
Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

Hendrik Zurlinden
Department of Transport (DoT) - VicRoads, Kew, VIC, Australia

Hu Shao
School of Mathematics, China University of Mining and Technology, Xuzhou, China

Inhi Kim
Department of Civil and Environmental Engineering, Kongju National University, Cheonan, Republic -of Korea

Ipsita Banerjee
Department of Civil and Environmental Engineering, University of California, Berkeley, California, United States

Jacek Pawlak
Urban Systems Lab & Centre for Transport Studies, Imperial College London, United Kingdom

Jean-Simon Bourdeau
Department of Civil, Geological and Mining Engineering, Polytechnique Montreal, Montreal, QC, Canada

Jia Li
Texas Tech University, Lubbock, TX, USA

Jiani Xie
University of Southern California, Department of Civil and Environmental Engineering, Los Angeles, CA, United States

Jiaqi Ma
University of California, Los Angeles, CA,
United States

Jiarong Yao
Tongji University, Shanghai, China

Joel Freedman
RSG, Portland, OR, United States

John Gaffney
Department of Transport (DoT) - VicRoads, Kew, VIC,
Australia

José Ramón D. Frejo
Systems and Automation Engineering Department,
University of Seville, Seville, Spain

Joshua Auld
Argonne National Laboratory, Lemont, IL, United States

Jun Chen
School of Transportation, Southeast University, Nanjing,
China

Jun Xie
School of Transportation and Logistics, Southwest Jiaotong
University, Chengdu, China

Justin Geistefeldt
Institute for Traffic Engineering and Management, Ruhr
University Bochum, Universitaetsstr. 150, Bochum,
Germany

Kamal Khanali
The University of Queensland, Brisbane, QLD, –Australia

Kathleen Deutsch-Burgner
Data Perspectives Consulting, Santa Barbara, CA, United
States

Kentaro Wada
University of Tsukuba, Tsukuba, Ibaraki, –Japan

Keshuang Tang
Tongji University, Shanghai, China

Khandker Nurul Habib
Department of Civil & Mineral Engineering, University of
Toronto, Toronto, ON, Canada

Kian Keong Chin
Chief Engineer, Road & Traffic, Land Transport
Authority, Singapore

Koji Suzuki
Nagoya Institute of Technology, Gokiso-cho, Showa-ku,
Nagoya Japan

Konstadinos G. Goulias
Department of Geography, GeoTrans Laboratory,
University of California Santa Barbara, Santa Barbara,
CA, United States

Lelitha Vanajakshi
Department of Civil Engineering, IIT Madras, Chennai,
Tamil Nadu, India

Li Weigang
University of Brasilia - UnB, Department of Computer
Science - CIC, Brasilia, DF Brazil

Lijiao Wang
University of Southern California, Department of Civil
and Environmental Engineering, Los Angeles, CA, United
States

Ling Wang
The Key Laboratory of Road and Traffic Engineering of the
Ministry of Education, Tongji University, Shanghai, P. R.
China

Long Cheng
Department of Geography, Ghent University, Ghent,
Belgium

Maëlle Zimmermann
Department of Computer Science and Operations
Research and CIRRELT, Université de Montréal,
Montreal, QC, Canada

Madhuri Kashyap
Department of Civil and Environmental Engineering,
Indian Institute of Technology, Tirupati, India

Mahmoud Mesbah
Amirkabir University of Technology (Tehran Polytechnic),
Tehran, Iran

Mahmoud Salari
California State University Dominguez Hills, Department
of Accounting, Finance, and Economics, Carson, CA,
United States

Marco Guerrieri
DICAM, University of Trento, Trento, Italy

Maria G. Burns
BTI, a DHS Center of Excellence, University of Houston,
Spring, TX, United States

Michael A.P. Taylor
School of Natural and Built Environments, –University of
South Australia, Adelaide, SA, –Australia

Miho Iryo-Asano
Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi,
Japan

Milan Mathew Thomas
Department of Civil Engineering, Rajiv Gandhi Institute
of Technology Kottayam, Kerala, India

Milos Balac
Stefano-Francini-Platz 5, Institute for Transport
Planning and Systems, ETH Zurich, Zurich, –Switzerland

Min-Ho Ha
Graduate School of Logistics, Incheon National University, Incheon, Republic of Korea

Ming Zhang
The University of Texas at Austin, Austin, TX, United States

Monica Menendez
New York University Abu Dhabi (NYUAD), Abu Dhabi, UAE

Monique Stinson
Argonne National Laboratory, Lemont, IL, United States

Nagendra R Velaga
Transportation Systems Engineering, Department of Civil Engineering, Indian Institute of Technology (IIT) Bombay, Powai, Mumbai, India

Nancy McGuckin
Travel Behavior Analyst, Consultant, South Pasadena, CA, United States

Naveen Chandra Iraganaboina
Department of Civil, Environmental & Construction Engineering, University of Central Florida, Orlando, FL, –United States

Naveen Eluru
Department of Civil, Environmental & Construction Engineering, University of Central Florida, –United States

Omer Verbas
Argonne National Laboratory, Lemont, IL, United States

Prof. Dr. Dirk Wittowsky
University of Duisburg-Essen, Faculty of Engineering/ Institute for Mobility and Urban Planning, Universitätsstraße, Essen, Germany

Qiheng Lin
The Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai, P. R. China

Raffaele Mauro
DICAM, University of Trento, Trento, Italy

Rahim F. Benekohal
University of Illinois at Urbana-Champaign, Urbana, IL, United States

Rajesh Paleti
Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA, United States

Ram M. Pendyala
School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ, United States

Ramina Jahanbakhsh Javid
Shahid Beheshti University, Department of Urban Planning & Design, Tehran, Iran

Renato Guadamuz
Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA, United States

Roxana J. Javid
University of Southern California, Department of Civil and Environmental Engineering, Los Angeles, CA, United States

Rui-jun Guo
School of Transportation Engineering, Dalian Jiaotong University, Dalian, Liaoning, China

Ruimin Li
Department of Civil Engineering, Tsinghua University, Beijing, China

S.C. Wong
Department of Civil Engineering, The University of Hong Kong, Hong Kong, China

S.K. Jason Chang
Department of Civil Engineering, National Taiwan University, Taiwan

Sara Moridpour
Civil and Infrastructure Engineering Discipline, –School of Engineering, RMIT University, Melbourne, Australia

Sebastián Raveau
Department of Transport Engineering and Logistics, Pontificia Universidad, Católica de, Chile

Senthilmurugan Thyagarajan
ESC Inc, Turner-Fairbank Highway Research Center, McLean, VA, United States

Sharmili Banik
Department of Civil Engineering, IIT Madras, Chennai, Tamil Nadu, India

Shinji Tanaka
Yokohama National University, Tokiwadai, Hodogaya-ku Yokohama, Japan

Shiva Habibi
RISE, Research Institute of Sweden, Göteborg, –Sweden

Shlomo Bekhor
Faculty of Civil and Environmental Engineering, Technion-Israel Institute of Technology, Haifa, Israel

Shuhan Cao
Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

Shuyan Chen
School of Transportation, Southeast University, Jiangsu, China

Sigal Kaplan
Department of Geography, Hebrew University of Jerusalem, Jerusalem, Israel

Srijith Balakrishnan
Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX, United States

Stacey G. Bricka
Independent Consultant, Bastrop, TX, United States

Stephen D. Boyles
Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX, United States

Taha Hossein Rashidi
UNSW Sydney, Kensington, NSW, Australia

Takahiro Tsubota
3 Bunkyo-cho, Matsuyama, Ehime, Japan

Thomas F. Rossi
Cambridge Systematics, Inc., Medford, MA, United States

Timothy D. Hau
HKU Business School, The University of Hong Kong, Hong Kong

Toru Seo
The University of Tokyo, Bunkyo-ku, Tokyo, Japan

Venkatesan Kanagaraj
Department of Civil Engineering, Indian Institute of Technology Kanpur, Kanpur, India

Wael K.M. Alhajyaseen
Qatar Transportation and Traffic Safety Center, Qatar University, Doha, Qatar

Wanjing Ma
The Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai, P. R. China

Wenruifan Yang
University of Southern California, Department of Civil and Environmental Engineering, Los Angeles, CA, United States

William H.K. Lam
Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China

Yao-Jan Wu
Smart Transportation Lab Department of –Civil and Architectural Engineering and Mechanics, University of Arizona, Tucson, AZ –United States

Yasuhiro Shiomi
Ritsumeikan University, Kusatsu, Shiga, Japan

Yetis Sazi Murat
Pamukkale University, Faculty of Engineering, Civil Engineering Department, Denizli, Turkey

Yi Guo
University of Cincinnati, Cincinnati, OH, United States

Yingjiu Pan
School of Transportation, Southeast University, Jiangsu, China

Young-Joon Seo
School of Economics & Trade, Kyungpook National University, Daegu, Republic of Korea

Yu (Marco) Nie
Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL, United States

Yumin Cao
Tongji University, Shanghai, China

Yusak Susilo
Institute for Transport Studies, University of Natural Resources and Life Sciences, Vienna, Austria

Zaili Yang
Liverpool Logistics Offshore and Marine Research Institute (LOOM), Liverpool John Moores University, Liverpool, United Kingdom

Zhanmin Zhang
Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX, United States

Zhi-Chun Li
School of Management, Huazhong University of Science and Technology, Wuhan, China

Page left intentionally blank

CONTENTS OF ALL VOLUMES

<i>Editorial Board</i>	<i>v</i>
<i>Introduction</i>	<i>vii</i>
<i>Contributors to Volume 4</i>	<i>ix</i>

VOLUME 1

Introduction to Transportation Economics	1
Market Failures and Public Decision Making in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	2
Demand for Freight Transport <i>José Holguín-Veras and Diana G. Ramírez-Ríos</i>	7
Cost Functions for Road Transport <i>José Manuel Vassallo</i>	13
Future of Urban Freight <i>Russell G. Thompson</i>	20
Operation Costs for Public Transport <i>Marco Batarce</i>	26
Natural Monopoly in Transport <i>André de Palma and Julien Monardo</i>	30
Freight Costs: Air and Sea <i>Yulai Wan and Dong Yang</i>	36
Transport Production and Cost Structure <i>Ricardo Giesen and Darío Farren</i>	46
The Concept of External Cost: Marginal versus Total Cost and Internalization <i>Sofia F. Franco</i>	56
Value of Time <i>John J. Bates</i>	67
Valuation of Carbon Emissions <i>Svante Mandell</i>	72
Valuation of Travel Time Variability Using Scheduling Models <i>Katrine Hjorth</i>	76

Value of Crowding <i>Daniel Hörcher</i>	84
What Drives Transport and Mobility Trends? The Chicken-and-Egg Problem <i>Nathalie Picard</i>	89
Pricing Principles in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	95
Long-Run Versus Short-Run Valuations <i>Stefanie Peer</i>	102
Value of Noise <i>Jon P. Nelson</i>	106
The Value of Life and Health <i>Henrik Andersson</i>	114
The Value of Security, Access Time, Waiting Time, and Transfers in Public Transport <i>Raquel Espino, Juan de Dios Ortúzar, and Luis I. Rizzi</i>	122
Demand for Passenger Transportation <i>Kenneth A Small and Robin Lindsey</i>	127
Real-World Experiences of Congestion Pricing <i>Charles Raux</i>	134
Distributional Effects of Congestion Charges and Fuel Taxes <i>Jonas Eliasson</i>	139
The Bottleneck Model <i>Dereje Abegaz and Yili Tang</i>	146
Dynamic Congestion Pricing and User Heterogeneity <i>Kathrin Goldmann and Gernot Sieg</i>	150
Economics of Parking <i>Daniel Albalade and Albert Gragera</i>	159
Loss Aversion and Size and Sign Effects in Value of Time Studies <i>Andrew Daly</i>	165
Intertemporal Variation of Valuations <i>James Fox</i>	170
The Rebound Effect for Car Transport <i>Bruno De Borger, Ismir Mulalic and Jan Rouwendal</i>	174
Elasticities for Travel Demand: Recent Evidence <i>Fay Dunkerley, Charlene Rohr and Mark Wardman</i>	179
Parking Price Elasticities <i>Stephan Lehner</i>	185
The Pareto Criterion and the Kaldor Hicks Criterion <i>Harald Minken</i>	190
Social Discount Rates <i>Lars Hultkrantz</i>	195
Cross-Elasticities between Modes <i>Mark Wardman and Jeremy Toner</i>	201
First-Best Congestion Pricing <i>Moez Kilani</i>	209

Ethical Aspects-Can We Value Life, Health, and Environment in Money Terms? <i>Jose M. Grisolia and Ken Willis</i>	216
Car tolls, Transit Subsidies for Commuting, and Distortions on the Labor Market <i>Ioannis Tikoudis¹ and Kurt Van Dender¹</i>	221
Demand Management and Capacity Planning of Airports <i>Stephen Ison and Lucy Budd</i>	227
Dealing With Negative Externalities: Low Emission Zones Versus Congestion Tolls <i>Valeria Bernardo, Xavier Fageda, and Ricardo Flores-Fillol</i>	231
The Rule-of-a-Half and Interpreting the Consumer Surplus as Accessibility <i>Mogens Fosgerau and Ninette Pilegaard</i>	237
Producer Surplus <i>Chau Man Fung</i>	242
The Robustness of Cost-Benefit Analyses <i>Morten Welde and James Odeck</i>	249
The GDP Effects of Transport Investments: The Macroeconomic Approach <i>James Laird and Daniel Johnson</i>	256
The Mohring Effect <i>Hugo E. Silva</i>	263
Public Transport Fare and Subsidy Optimization <i>Qianwen Guo and Zhongfei Li</i>	267
Transportation Equity <i>Rafael H. M. Pereira, and Alex Karner</i>	271
Impact of Transport Cost-Benefit Analysis on Public Decision-Making <i>Niek Mouter</i>	278
Causal Inference for Ex Post Evaluation of Transport Interventions <i>Daniel J. Graham</i>	283
Transport Cost and Location of Firms <i>Adelheid Holl</i>	291
Commuting, the Labor Market, and Wages <i>Jan Rouwendal, and Ismir Mulalic</i>	297
How to Buy Transport Infrastructure <i>Johan Nyström</i>	302
Procurement of Public Transport: Contractual Regimes <i>Andrew Smith, and Chris Nash</i>	308
The Mono-Centric City Model and Commuting Cost <i>Zhi-Chun Li, and Ya-Juan Chen</i>	315
Value of Time in Freight Transport <i>Gerard de Jong</i>	321
The Economics and Planning of Urban Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	326
Incentives in Public Transport Contracts <i>Andreas Vigen</i>	332
Transportation Improvements and Property Prices <i>Alex Anas</i>	337

Transport Infrastructure Effects on Economic Output: The Microeconomic Approach <i>Patricia C. Melo</i>	347
Wider Economic Impacts of Transport Investments <i>Anthony J. Venables</i>	355
Employment Effects of Transport Infrastructure <i>Anna Matas, and Javier Asensio</i>	360
Regulatory Reforms and Competition in Public Transport <i>Didier van de Velde</i>	365
How to Finance Transport Infrastructure? <i>Tiziana D'Alfonso, and Giuseppe Catalano</i>	371
Congestion, Allocation and Competition on the Railway Tracks <i>John Armstrong, and John Preston</i>	378
Public Private Partnership <i>David Meunier, and Emile Quinet</i>	385
Airline Economics <i>Anming Zhang, Yahua Zhang, and Zhibin Huang</i>	392
Privatization and Deregulation of the Airline Industry <i>Achim I. Czerny, and Hao Lang</i>	397
Price Discrimination and Yield Management in the Airline Industry <i>Guoquan Zhang, Colin C.H. Law, Yahua Zhang, and Hangjun Yang</i>	404
Regulation and Competition in Railways <i>Chris Nash, and Andrew Smith</i>	409
Cycling Economics <i>Bert van Wee</i>	414
The Economic Rationale for High-Speed Rail <i>Ginés de Rus</i>	419
Rail Cost Functions <i>Kristofer Odolinski, and Phill Wheat</i>	425
Transport and International Trade <i>Siri Pettersen Strandenes</i>	431
Maritime Economics: Organizational Structures <i>Kevin P.B. Cullinane</i>	436
Port Planning and Investment <i>Jasmine Siu Lee Lam</i>	443
Estimating the Capital Stock of Transport Infrastructure <i>Heike Link</i>	449
The Economics of Reducing Carbon Emissions From Air and Road Transport <i>Olga Ivanova</i>	457
Regulation and Financing of Toll Roads <i>Marco Ponti</i>	464
Are Megaprojects too Transformational for Cost-Benefit Analysis? <i>Tom Worsley</i>	470
Economics of Transportation Safety <i>Ian Savage</i>	476

Cost Overruns of Transportation Infrastructure Projects <i>James Odeck, and Morten Welde</i>	483
The Downs-Thomson Paradox <i>Joel P. Franklin</i>	490
Policy Instruments for Plug-In Electric Vehicles: An Overview and Discussion <i>Jake Whitehead, Patrick Plötz, Patrick Jochem, Frances Sprei, and Elisabeth Dütschke</i>	496
Vertical and Horizontal Separation in the European Railway Sector and Its Effects on Productivity <i>Pedro Cantos-Sánchez</i>	503
How will Autonomous Vehicles Impact Car Ownership and Travel Behavior <i>Patrick M. Bösch, Felix Becker, Henrik Becker, and Kay W. Axhausen</i>	508
Policy Instruments to Reduce Carbon Emissions from Road Transport Computable General Equilibrium Analysis in Transportation Economics <i>Johannes Bröcker</i>	520
Estimation of Value of Time <i>Stefan Flügel, and Askill H. Halse</i>	527
The Taxation of Car Use in the Future <i>Griet De Ceuster, and Inge Mayeres</i>	534
Cost-Benefit Analysis and Other Assessment Techniques: Contrasts and Synergies <i>Paolo Beria</i>	540
Demand for Air Travel and Income Elasticity <i>Jing Lu, Yucan Meng, Changmin Jiang, and Cheng Lv</i>	547
Generalized Cost for Transport <i>Jepppe Rich</i>	555
The Impact of Electric Vehicles on Energy Systems <i>Patrick E.P. Jochem, Jake Whitehead, and Elisabeth Dütschke</i>	560
Uber versus Taxis <i>Georgina Santos</i>	566
Contract Efficiency in Public Transport Services <i>Philippe Gagnepain, and Marc Ivaldi</i>	572
Company Cars <i>Stefan Gössling</i>	580
Transportation Network Companies (TNCs) and the Future of Public Transportation <i>Susan Shaheen, and Adam Cohen</i>	584
Public Transport in Low Density Areas <i>Jani-Pekka Jokinen, Leif Sörensen, and Jan Schlüter</i>	589
Market Failures in Transport: Direct and Indirect Public Intervention <i>Federico Boffa, and Alberto Iozzi</i>	596
The Braess Paradox <i>Anna Nagurney, and Ladimer S. Nagurney</i>	601
VOLUME 2	
Introduction to Transportation Safety and Security <i>Per Gårder</i>	1

The Concept of “Acceptable Risk” Applied to Road Safety Risk Level <i>Claes Tingvall</i>	2
Crash Not Accident <i>Robert A. Scopatz</i>	6
Age and Gender as Factors in Road Safety <i>Marion Sinclair</i>	11
Aggressive Driving and Road Rage <i>James E.W. Roseborough, Christine M. Wickens, and David L. Wiesenthal</i>	17
Aircraft Maintenance and Inspection <i>Alan Hobbs</i>	25
Airport Security <i>Richard W. Bloom</i>	34
Transport Safety and Security: Alcohol <i>James C. Fell</i>	40
Animal Crashes <i>Michal Bil</i>	53
Attenuators <i>Simonetta Boria</i>	63
ATV, Snowmobile, and Terrain Vehicle Safety <i>David P. Gilkey, and William Brazile</i>	77
Automobile Safety Inspection <i>Subasish Das</i>	85
Aviation Safety: Commercial Airlines <i>Clarence C. Rodrigues</i>	90
Aviation Safety, Freight, and Dangerous Goods Transport by Air <i>Glenn S. Baxter and Graham Wild</i>	98
Bicycle Collision Avoidance Systems: Can Cyclist Safety be Improved with Intelligent Transport Systems? <i>Lars Leden</i>	108
Bicycle Infrastructure <i>Rock E. Miller</i>	115
Bicycles, E-Bikes and Micromobility, A Traffic Safety Overview <i>Höskuldur Kröyer</i>	125
Bicycles: The Safety of Shared Systems Versus Traditional Ownership <i>Mercedes Castro-Nuño and José I. Castillo-Manzano</i>	139
Bridge Safety <i>Per Erik Garder</i>	144
Carsharing Safety and Insurance <i>Elliot Martin and Susan Shaheen</i>	150
Carjacking <i>Terance D. Mieth, Christopher Forepaugh, Tanya Dudinskaya</i>	157
Collision Avoidance Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	164
Collision Avoidance Systems, Automobiles <i>Erick J. Rodríguez-Seda</i>	173

Connected Automated Vehicles: Technologies, Developments, and Trends <i>Azra Habibovic and Lei Chen</i>	180
Construction Zones <i>Jalil Kianfar</i>	189
Costs of Accidents <i>Ulf Persson</i>	196
Critical Issues for Large Truck Safety <i>Matthew C. Camden, Jeffrey S. Hickman, Richard J. Hanowski, and Martin Walker</i>	200
Demerit Points and Similar Sanction Programs <i>Matúš ťucha and Kristýna Josrová</i>	210
Driver State and Mental Workload <i>Dick de Waard and Nicole van Nes</i>	216
Drugs, Illicit, and Prescription <i>Rune Elvik</i>	221
Education, Training, and Licensing <i>Matúš ťucha and Kristýna Josrová</i>	228
Elderly Driver Safety Issues <i>Mark J King</i>	233
Emergency Response Systems <i>Frances L. Edwards</i>	240
Emergency Vehicles and Traffic Safety <i>Shamsunnahar Yasmin, Sabreena Anowar, and Richard Tay</i>	247
Encouragement: Awards and Incentives <i>Fred Wegman</i>	255
Enforcement and Fines <i>Matúš ťucha and Ralf Risser</i>	263
Epidemiology of Road Traffic Crashes <i>Sherrie-Anne Kaye, Judy Fleiter, and Md Mazharul Haque</i>	269
Evacuation Planning and Transportation Resilience <i>Karl Kim</i>	276
Exposure: A Critical Factor in Risk Analysis <i>Frank Gross, PhD, PE</i>	282
Passenger Ferry Vessels and Cruise Ships: Safety and Security <i>Wayne K. Talley</i>	290
Fuel Economy Standards: Impacts on Safety <i>Kenneth T. Gillingham and Stephanie M. Weber</i>	296
Hazardous Materials Transport <i>Dr. Arjan Vincent van der Vlies</i>	304
Head-on Crashes <i>John N. Ivan</i>	311
Helicopters in Emergency Medical Response <i>Stephen J.M. Sollid, M.D., PhD</i>	316
Horizontal and Vertical Geometry <i>Victoria Gitelman</i>	322

Human Factors in Transportation <i>Alison Smiley, Christina (Missy) Rudin-Brown</i>	331
In-Depth Crash Analysis and Accident Investigation <i>Yong Peng, Helai Huang, and Xinghua Wang</i>	346
Incident Detection Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	351
Inequality and Traffic Safety <i>Miles Tight</i>	358
Lighting <i>John D. Bullough</i>	361
Macroscopic Safety Analysis <i>Mohamed Abdel-Aty and Jaeyoung Lee</i>	367
Motor Vehicle Crash Reportability <i>John J. McDonough</i>	380
Nominal Safety <i>Per Erik Garder</i>	386
Parking Lots <i>Maxim A. Dulebenets</i>	392
Passenger Van Safety <i>Saksith Chalermpong and Apiwat Ratanawaraha</i>	401
Passive Prevention Systems in Automobile Safety <i>B. Serpil Acar</i>	406
Pedestrian Safety, Children <i>Mette Møller</i>	415
Pedestrian Safety, General <i>Muhammad Z. Shah, Mehdi Moeinaddini, and Mahdi Aghaabbasi</i>	420
Pedestrian Safety, Older People <i>Carlo Lui</i>	429
Visually Impaired Pedestrian Safety <i>Robert S. Wall Emerson</i>	435
Photo/Video Traffic Enforcement <i>Charles M. Farmer</i>	439
Powered Two- and Three-Wheeler Safety <i>Fangrong Chang, Helai Huang, and Md. Mazharul Haque</i>	443
Railroad Safety <i>Xiang Liu and Zhipeng Zhang</i>	451
Railroad Safety: Grade Crossings and Trespassing <i>Rahim F. Benekohal and Jacob Mathew</i>	466
Recreational Boating Safety: Usage, Risk Factors, and the Prevention of Injury and Death <i>Amy E. Peden, Stacey Willcox-Pidgeon, and Kyra Hamilton</i>	477
Refuge Islands <i>Christer Hydén</i>	487
Risk Perception and Risk Behavior in the Context of Transportation <i>Martina Raue and Eva Lerner</i>	494

Road Diets <i>Robert B. Noland</i>	500
Road Safety Audits <i>Xiao Qin</i>	508
Roadside Safety Barriers <i>Dean C. Alberson</i>	513
Road Safety Management in Selected Countries <i>Paul Boase and Brian Jonah</i>	519
Roadway Pavement Conditions and Various Summer and Winter Maintenance Strategies <i>Kamal Hossain, PhD P. Eng</i>	529
Safety of Roundabouts <i>Khaled Shaaban</i>	539
Rumble Strips, Continuous Shoulder, and Centerline <i>AnnaAnund and AnnaVadeby</i>	549
Safety Culture <i>Tor-Olav Nvestad</i>	554
Safety Data Quality Management <i>Robert A. Scopatz</i>	560
School Bus Safety <i>Yousif A. Abulhassan</i>	564
School Campus Traffic Circulation <i>Dimitrios Nalmpantis</i>	568
Sexual Violence in Public Transportation <i>Vania Ceccato</i>	576
Shared Space <i>Allison B. Duncan</i>	584
Side Area Safety and Side Slopes <i>José María Pardillo-Mayora</i>	593
Simulators <i>Nichole L. Morris, Curtis M. Craig, Jacob D. Achtemeier, and Peter A. Easterlund</i>	602
Sleep-Related Issues and Fatigue <i>Prof Marion Sinclair and Estelle Swart</i>	611
Speed Governors and Limiters <i>Christer Hydén</i>	617
Speed Limits on Rural Highways <i>Peter Tarmo Savolainen and Timothy Jordan Gates</i>	622
Speed Limits on Urban Streets <i>Anna Bray Sharpin, Claudia Adriazola-Steil, Ben Welle, and Natalia Lleras</i>	632
Speed-Reducing Measures <i>Vinod Vasudevan</i>	641
Striping, Signs, and Other Forms of Information <i>Kasem Choocharukul and Kerkritt Sriroongvikrai</i>	648
Suicides <i>Brendan Ryan</i>	656

Surrogate Measures of Safety <i>Nicolas Saunier and Aliaksei Laureshyn</i>	662
Targeting Transit: the Terrorist Threat and the Challenges to Security <i>Brian Michael Jenkins</i>	668
The Swedish Vision Zero: A Policy Innovation <i>Matts-Åke Belin</i>	675
Tire Safety <i>Saied Taheri</i>	681
Town Gates: Section on Transport Safety and Security <i>Charles Tijus</i>	685
Traffic Flow Volume and Safety <i>Athanasios Theofilatos and Apostolos Ziakopoulos</i>	692
Traffic Safety and Security of Taxis and Ride-Hailing Vehicles <i>Zhe Wang, Helai Huang, and Ye Li</i>	699
Traffic Signals and Safety <i>Andrew P. Tarko</i>	706
Tunnels, Safety and Security Issues-Risk Assessment for Road Tunnels: State-of-the-Art Practices and Challenges <i>Konstantinos Kirytopoulos, Panagiotis Ntzeremes, and Konstantinos Kazaras</i>	713
Understanding, Managing, and Learning from Disruption <i>Karl Kim</i>	719
Use/Analysis of Crash Data and Underreporting of Crashes <i>Mohammadali Shirazi and Dominique Lord</i>	726
Utility Poles <i>Lai Zheng and Tarek Sayed</i>	731
Value of Life and Injuries <i>David A. Hensher</i>	737
Effects of Weather <i>Maria Pregnolato, Amirhassan Kermanshah, and Wisinee Wisetjindawat</i>	742
Wrong-Way Driving on Motorways <i>Huaguo Zhou and Md Atiquzzaman</i>	751

VOLUME 3

Freight Transport and Logistics <i>Sharon Cullinane and Kevin Cullinane</i>	1
Expanding the Perspective of Logistics and Supply Chain Management <i>David J. Closs</i>	8
Logistics and Supply Chain Management Performance Measures <i>David B. Grant and Sarah Shaw</i>	16
Economic Regulation/Deregulation and Nationalization/Privatization in Freight Transportation <i>Wayne K. Talley</i>	24
Freight Transport Policy <i>Luca Zamparini and Aura Reggiani</i>	29

Planning and Financing Logistics Spaces <i>Nicolas Raimbault</i>	35
Supply Chain Risk Management: Creating the Resilient Supply Chain <i>Richard Wilding</i>	41
Transportation Safety and Security <i>Maria G. Burns</i>	47
Resilience in Freight Transport Networks <i>Zhuohua Qu, Chengpeng Wan, and Zaili Yang</i>	53
Environmental Sustainability in Freight Transportation <i>Lisa M. Ellram</i>	58
Sustainable Logistics, CSR in Logistics, and Sustainable Supply Chain Management <i>Maria Björklund and Maja Piecyk-Ouellet</i>	64
Information Sharing and Business Analytics in Global Supply Chains <i>Prof Usha Ramanathan and Prof Ramakrishnan Ramanathan</i>	71
Logistics Information Systems <i>Petri Helo and Javad Rouzafzoon</i>	76
Factors Affecting the Selection of Logistics Service Providers <i>Aicha AGUEZZOUL</i>	85
Logistics Service Performance <i>Kee-hung Lai, Jinan Shao, and Yongyi Shou</i>	89
The World Bank's Logistics Performance Index <i>Christina K. Wiederer cwiederer, Jean-François Arvis, Lauri M. Ojala, and Tuomas M. M. Kiiski</i>	94
Outsourcing Logistics Functions <i>Evi Hartmann, Hendrik Birkel, and Matthias Kopyto</i>	102
Freight Transport and Logistics in JIT Systems <i>James H. Bookbinder and M. Ali Aelkū</i>	107
Supply Chain Finance <i>Erik Hofmann</i>	113
Packaging Logistics <i>Jesús García-Arca, Alicia Trinidad González-Portela Garrido, and J. Carlos Prado-Prado</i>	119
The Bullwhip Effect <i>Jan C. Fransoo and Maximiliano Udenio</i>	130
Blockchain Applications in Logistics <i>Yingli Wang</i>	136
Logistics in Asia <i>Shong-lee Ivan Su</i>	143
Logistics in the Developing World <i>Charles Kunaka</i>	150
Freight Network Modeling <i>Lóránt Tavasszy and Yousef Maknoon</i>	157
National Freight Transport Models <i>Gerard de Jong</i>	162
Freight Flows in Cities <i>Genevieve Giuliano</i>	168

Urban Logistics and Freight Transport <i>Michael Browne, José Holguín-Veras, and Julian Allen</i>	178
Omni-Channel Logistics <i>Tom Van Woensel</i>	184
Humanitarian Logistics <i>Gyöngyi Kovács and Diego Vega</i>	190
Optimization of Humanitarian Logistics <i>M. Teresa Ortuño, Jose M. Ferrer, Inmaculada Flores, and Gregorio Tirado</i>	195
Event Logistics <i>Rev. Ruth Dowson and Dan Lomax</i>	201
Reverse Logistics <i>Dale S. Rogers and Ronald S. Lembke</i>	208
E-Tailing and Reverse Logistics <i>Sharon Cullinane</i>	219
Green Routing of Freight Vehicles <i>Tolga Bektaş</i>	224
Freight Mode Choice <i>Hyun Chan Kim and Alan Nicholson</i>	231
The Value of Time in Freight Transport <i>María Feo-Valero, Amaya Vega, and Bárbara Vázquez-Paja</i>	236
Behavioral Research in Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	242
Container (Liner) Shipping <i>Theo Notteboom</i>	247
Bulk Shipping Markets: An Overview of Market Structure and Dynamics <i>Manolis G. Kavussanos and Stella A. Moysiadou</i>	257
Ferries and Short Sea Shipping <i>Lourdes Trujillo and Alba Martínez-López</i>	280
Shipping and the Environment <i>Karin Andersson, Selma Brynolf, Lena Granhag, and J. Fredrik Lindgren</i>	286
Energy Efficiency of Ships <i>Harilaos N. Psaraftis</i>	294
Seaports <i>Mary R. Brooks and Geraldine Knatz</i>	299
Port Hinterlands <i>Francesco Parola, Giovanni Satta, and Francesco Vitellaro</i>	305
Seaports as Clusters of Economic Activities <i>Peter W. de Langen</i>	310
Port Efficiency and Effectiveness <i>Lourdes Trujillo, María Manuela González, Casiano Manrique-De-Lara-Peñate, and Ivone Pérez</i>	316
Container Port Automation <i>Michael G.H. Bell</i>	323
Optimizing Crane Operations in Ports Scheduling of Liner Container Shipping Services	335

<i>Yuquan Du and Qiang Meng</i>	
Dry Ports	344
<i>Gordon Wilmsmeier and Jason Monios</i>	
Arctic Shipping	349
<i>Yufeng Lin, David G. Babb, and Adolf K.Y. Ng</i>	
Airfreight and Economic Development	355
<i>Kenneth Button</i>	
Air Freight Logistics	361
<i>Keith Debbage and Neil Debbage</i>	
Air Freight Marketing	369
<i>Lucy Budd and Stephen Ison</i>	
Drones in Freight Transport	374
<i>Oliver Kunze</i>	
Duty of Care in the Selection of Motor Carriers	382
<i>Thomas M. Corsi</i>	
Carrier Selection for Less-Than-Truckload (LTL) Shipments	388
<i>Dinçer Konur, Gonca Yildirim, and Bahriye Cesaret</i>	
Decarbonizing Road Freight Transport	395
<i>Heikki Liimatainen</i>	
The Rebound Effect in Road Freight Transport	402
<i>Tooraj Jamasb and Manuel Llorca</i>	
Autonomous Goods Transport	407
<i>Heike Flømig</i>	
Rail Freight	413
<i>Dr Allan Woodburn</i>	
Rail Freight Vehicles	423
<i>Maksym Spiryagin, Qing Wu, Peter Wolfs, Colin Cole, Valentyn Spiryagin, and Tim McSweeney</i>	
Eurasia Rail Freight: Enablers and Inhibitors of Future Growth	436
<i>Hendrik Rodemann and Simon Templar</i>	
Intermodal and Synchromodal Freight Transport	456
<i>Tomas Ambra, Koen Mommens, and Cathy Macharis</i>	
Pipelines	463
<i>Matthew E. Oliver</i>	
3D Printers and Transport	471
<i>Wouter P.C. Boon and Bert van Wee</i>	
The Physical Internet and Logistics	479
<i>Eric Ballot and Shenle PAN</i>	
Bicycles for Urban Freight	488
<i>Barbara Lenz and Johannes Gruber</i>	
China's Belt and Road Initiative	495
<i>Paul Tae-Woo Lee</i>	

VOLUME 4

Introduction to Traffic Management <i>Edward C.S. Chung</i>	1
Urban Motorway Management <i>John Gaffney and Hendrik Zurlinden</i>	2
Ramp Metering Application <i>John Gaffney and Hendrik Zurlinden</i>	10
City Wide Coordinated Ramp Meters <i>John Gaffney and Hendrik Zurlinden</i>	21
Variable Speed Limits for Traffic Efficiency Improvement <i>José Ramón D. Frejo and Bart De Schutter</i>	33
Hard Shoulder Running <i>Justin Geistefeldt</i>	41
High-Occupancy Vehicle (HOV) and High-Occupancy Toll (HOT) Lanes <i>Roxana J. Javid, Jiani Xie, Lijiao Wang, Wenruifan Yang, Ramina Jahanbakhsh Javid, and Mahmoud Salari</i>	45
Reversible Lanes: Guidelines, Operation and Control, Research Directions <i>Gowri Asaithambi, Venkatesan Kanagaraj, and Madhuri Kashyap</i>	52
Electronic Toll Collection <i>Azusa Toriumi</i>	60
Road Pricing-Theory and Applications <i>Kian Keong Chin</i>	68
Road Pricing 1: The Theory of Congestion Pricing <i>Timothy D. Hau</i>	74
Road Pricing 2: Short- and Long-Run Equilibrium of Road Transportation <i>Timothy D. Hau</i>	83
Road Pricing 3: The Implications for Pricing Public Transportation <i>Timothy D. Hau</i>	90
Road Pricing 4: Case Study-The Implementation of Electronic Road Pricing in Hong Kong <i>Timothy D.</i>	103
Advanced Travelers Information Systems (ATIS) <i>Chintan Advani and Ashish Bhaskar</i>	106
Travel Time Reliability <i>Sharmili Banik, Anil Kumar, and Lelitha Vanajakshi</i>	109
Traffic Incident Management <i>Ruimin Li</i>	122
Traffic Incident Detection <i>Shuyan Chen and Yingjiu Pan</i>	128
Bottleneck <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	134
Flow Breakdown <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	143
Recurrent Congestion <i>Takahiro Tsubota</i>	154

Nonrecurrent Congestion <i>Takahiro Tsubota</i>	158
Freeway to Arterial Interfaces <i>Abolfazl Karimpour and Yao-Jan Wu</i>	162
Arterial Road Management <i>Jiaqi Ma, Yi Guo and Adekunle Adebisi</i>	169
Signalized Intersections <i>Chaitrali Shirke</i>	178
Protected Phase <i>Jiarong Yao, Yumin Cao, and Keshuang Tang</i>	185
Permitted Phase <i>Yumin Cao, Jiarong Yao, and Keshuang Tang</i>	196
Hook Turns: Implementation, Benefits, and Limitations <i>Sara Moridpour and Amir Falamarzi</i>	203
Traffic Signal Coordination <i>Rahim F. Benekohal</i>	213
Emergency Vehicle Priority (Preemption): Concept and Advancements <i>Chaitrali Shirke</i>	221
Signalized Roundabouts <i>Yetis Sazi Murat and Rui-jun Guo</i>	227
Turbo Roundabouts: Design, Capacity and Comparison With Alternative Types of Roundabouts <i>Marco Guerrieri and Raffaele Mauro</i>	238
Capacity of an Intersection <i>Ashish Verma and Milan Mathew Thomas</i>	247
Delay <i>Shinji Tanaka</i>	258
Queue Length <i>Shinji Tanaka</i>	263
Local Area Traffic Management <i>Michael A.P. Taylor</i>	268
On-Street Parking <i>Jun Chen and Guang Yang</i>	278
Off-Street Parking <i>Jun Chen and Guang Yang</i>	285
Parking Information Systems <i>Behrang Assemi and Douglas Baker</i>	289
Performance-Based Parking Management <i>Douglas Baker and Behrang Assemi</i>	299
Transit Priority <i>Wanjing Ma, Qiheng Lin, and Ling Wang</i>	307
Adaptive Bus Control <i>Monica Menendez</i>	315
Transit Fare Collection <i>Mahmoud Mesbah and Kamal Khanali</i>	325

Transit Information Systems <i>Ankit Kumar Yadav and Nagendra R Velaga</i>	331
Tram Lane Configurations and Driving Rules <i>Farhana Naznin</i>	338
Railway Crossing <i>Chunliang Wu and Inhi Kim</i>	341
Pedestrian Crossing (Crosswalk) <i>Miho Iryo-Asano, Wael K.M. Alhajyaseen, and Koji Suzuki</i>	346
Bike-Sharing System: Uncovering the “1Success Factors” <i>S.K. Jason Chang and Amanda Fernandes Ferreira</i>	355
Airspace Systems Technologies-Overview and Opportunities <i>Banavar Sridhar and Gano B. Chatterji</i>	363
Air Traffic Flow and Capacity Management <i>Li Weigang and Cristiano P Garcia</i>	380
Port Management <i>Maria G. Burns</i>	390
Port Performance Measurement from a Multistakeholder Perspective <i>Min-Ho Ha, Zaili Yang, and Young-Joon Seo</i>	396
Transport Modeling and Data Management <i>Chandra Bhat</i>	407
Computational Methods and Data Analytics <i>Bilal Farooq and David López</i>	408
Activity-Based Models <i>Renato Guadamuz and Rajesh Paleti</i>	414
Advanced Traveler Information Systems <i>Stephen D. Boyles</i>	418
Demand-Responsive Transit, Evaluation Studies <i>Sebastián Raveau</i>	423
Location Choice Models <i>Adam Wilkinson Davis</i>	428
Full Feedback and Equilibrium Modeling in Urban Travel Forecasting <i>Yu (Marco) Nie, Jun Xie, and David Boyce</i>	432
The Use and Value of Geographic Information Systems in Transportation Modeling <i>Ming Zhang</i>	440
Latent Demand and Induced Travel <i>Charisma F. Choudhury</i>	448
ICT, Virtual and In-Person Activity Participation, and Travel Choice Analysis <i>Jacek Pawlak and Giovanni Circella</i>	452
Public Transit Ridership Forecasting Models <i>Ipsita Banerjee, Deepa L, and Abdul Rawoof Pinjari</i>	459
Spatial Mismatch, Job Access, and Reverse Commuting <i>Gian-Claudia Sciarra</i>	468

Microsimulation and Agent-Based Models in Transportation <i>Milos Balac</i>	473
Choice Models in Transportation <i>Naveen Chandra Iraganaboina and Naveen Eluru</i>	477
Multi-Criteria Decision Analysis <i>Zhanmin Zhang and Srijith Balakrishnan</i>	485
The National Household Travel Survey Data Series (NPTS/NHTS) <i>Nancy McGuckin</i>	493
Route Choice and Network Modeling <i>Emma Frejinger and Maëlle Zimmermann</i>	496
Departure Time Choice Modeling <i>Khandker Nurul Habib</i>	504
Traveler Responses to Congestion <i>David T. Ory and Gayathri Shivaraman</i>	509
Origin-Destination Demand Estimation Models <i>William H.K. Lam, Hu Shao, Shuhan Cao, and Hai Yang</i>	515
Parking Demand Models <i>S.C. Wong, Zhi-Chun Li, and William H.K. Lam</i>	519
Pavement Management Systems <i>Senthilmurugan Thyagarajan</i>	524
Residential Location Choice Models <i>Shlomo Bekhor and Sigal Kaplan</i>	531
Transport Demand Management <i>Feiyang Zhang and Becky P.Y. Loo</i>	537
Traffic Flow Analysis <i>H. Michael Zhang and Jia Li</i>	544
Autonomous Vehicles and Transportation Modeling <i>Annesha Enam, Felipe de Souza, Omer Verbas, Monique Stinson, and Joshua Auld</i>	557
Ride-Hailing and Travel Demand Implications <i>Felipe F. Dias</i>	564
Transportation Modeling and Planning Software <i>Joel Freedman</i>	569
Transportation Statistics and Databases <i>Taha Hossein Rashidi</i>	574
Travel Surveys <i>Stacey G. Bricka</i>	587
Travel Demand Forecasting: Where Are We and What Are the Emerging Issues <i>Thomas F. Rossi</i>	590
Travel Model Calibration and Validation <i>Ram M. Pendyala</i>	596
Trip Chaining Analysis <i>Cynthia Chen and Yusak Susilo</i>	606
Vehicle Ownership Models <i>Dr. Sören Groth and Prof. Dr. Dirk Wittowsky</i>	612

Bicycle Sharing/Bikesharing <i>Catherine Morency and Jean-Simon Bourdeau</i>	617
Carsharing <i>Shiva Habibi and Frances Sprei</i>	623
Urban Recreational Travel <i>Long Cheng and Frank Witlox</i>	629
Place Perception and Travel Behavior <i>Kathleen Deutsch-Burgner and Konstadinos G. Goulias</i>	635

VOLUME 5

Transport Modes <i>Edoardo Marcucci</i>	1
Infrastructure Transport Investments, Economic Growth and Regional Convergence <i>Xavier Fageda and Cecilia Olivieri</i>	2
Transport Modes and an Aging Society <i>Charles B.A. Musselwhite and Theresa Scott</i>	6
Sustainable Mobility Paths <i>Erling Holden and Geoffrey Gilpin</i>	13
Vehicles that Drive Themselves: What to Expect with Autonomous Vehicles <i>Michele D. Simoni and Kara M. Kockelman</i>	19
Transport Modes and Tourism <i>Ila Maltese and Luca Zamparini</i>	26
Transport Modes and Accessibility <i>Bert van Wee</i>	32
Transport Modes and Globalization <i>Jean-Paul Rodrigue</i>	38
Transport Modes and Cities <i>Erick Guerra and Gilles Duranton</i>	45
Transport Modes and Remote Areas	51
Modeling Mode Choice in Freight Transport <i>Lóránt Tavasszy and Gerard de Jongb</i>	57
Travel Mode Choice as Reasoned Action <i>Sebastian Bamberg, Icek Ajzen, and Peter Schmidt</i>	63
Energy Consumption of Transport Modes <i>Zissis Samaras and Ilias Vouitsis</i>	71
Transport Modes and People With Limited Mobility <i>Roger L Mackett</i>	85
Transport Modes and Commuters <i>Colin G. Pooley</i>	92
Shopping and Transport Modes <i>Antonio Comi</i>	98
Transport Modes and Health <i>Jennifer S. Mindell and Sandra Mandic</i>	106

Multimodality in Transportation <i>Sören Groth and Tobias Kuhnimhof</i>	118
Transport Modes and Disasters <i>Brian Wolshon</i>	127
Big Data for Public Transport Planning <i>Jan-Dirk Schmöcker</i>	134
Active Transport: Heterogeneous Street Users Serving Movement and Place Functions <i>Regine Gerike, Stefan Hubrich, Caroline Koszowski, Bettina Schröter, and Rico Wittwer</i>	140
Electric Vehicles <i>Christine Eisenmann, Daniel Görges, and Thomas Franke</i>	147
Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service <i>Susan Shaheen, PhD and Adam Cohen</i>	155
Adoption of new travel information platforms <i>Sigal Kaplan</i>	160
ICT and Transport Modes <i>Galit Cohen-Blankshtain</i>	165
Mode Choice and Life Events <i>Joachim Scheiner</i>	171
Introduction to Air Transport <i>Milan Janić</i>	178
The History of Air Transportation <i>Richard P. Hallion</i>	192
The Geography of Air Transport <i>Lucy Budd and Stephen Ison</i>	198
The Future of Air Transport <i>Rico Merkert and James Bushell</i>	203
Air Transport and Its Territorial Implications <i>Lanfranco Senn</i>	208
Next Generation Travel: Young Adults' Travel Patterns <i>Tobias Kuhnimhof and Scott Le Vine</i>	215
Airport Network Planning and Its Integration with the HSR System <i>Francesca Pagliara, Juan Carlos Martín, and Concepción Román</i>	222
Airport Management <i>Peter Forsyth and Hans-Martin Niemeier</i>	229
Airport Regulation <i>Achim I. Czerny</i>	234
Air Route Planning and Development <i>Renan Peres de Oliveira and Gui Lohmann</i>	241
Airline Management <i>Sveinn Vidar Gudmundsson</i>	249
Air Cargo <i>Volodymyr Bilotkach</i>	258

Airline Regulation <i>Andrew R. Goetz</i>	263
Air Vehicles Classification <i>Vincenzo Torre</i>	269
Aircraft Manufacturing <i>Antonio Sollo</i>	290
A Geography of Road Transport in Cities <i>Cristian Domarchi and Juan de Dios Ortúzar</i>	300
The Future of Road Transport <i>Preston L. Schiller</i>	306
Road Transportation and Territorial Scale <i>Ana M. Condeço-Melhorado</i>	315
Road Modes: Walking <i>Kevin Manaugh, PhD Associate Professor</i>	320
Bus Public Transport Planning and Operations <i>Ehab Diab and Ahmed El-Geneidy</i>	326
Street Design for Active Travel <i>Bruce Appleyard</i>	333
Transit Planning and Management <i>Zakhary Mallett and Marlon G Boarnet</i>	349
Road Infrastructure: Planning, Impact and Management <i>Jos Arts, Wim Leendertse, and Taede Tillema</i>	360
Road Transport Planning at the Urban Scale <i>David A. King and Kevin J. Krizek</i>	373
Road Traffic Regulation: Road Pricing and Environmental Quality <i>Marco Percoco</i>	378
Car Ownership and Car Use: A Psychological Perspective <i>J.L. Veldstra, A.B. Ünal, E.M. Steg</i>	384
Road Transport: E-Scooters <i>Gysele Lima Ricci and Klaus Bogenberger</i>	391
Railway Station and Network Planning <i>Ingo Arne Hansen</i>	399
Introduction to Rail Transport <i>Chris Nash and Tony Fowkes</i>	406
The History of Rail Transport <i>Carlo Ciccarelli, Andrea Giuntini, and Peter Groote</i>	413
The Geography of Rail Transport <i>Frédéric Dobruszkes and Amparo Moyano</i>	427
Rail Transport and Territorial Scale <i>Prof. Andrés Monzón and Dr. Elena López</i>	437
Railway Management <i>Vassilios A. Profillidis</i>	444
Railway Terminal Regulation <i>Nacima Baron</i>	454

Service Network Design for Freight Railroads <i>Teodor Gabriel Crainic</i>	464
Subway Systems <i>Guillaume Monchambert, Daniel Hörcher, Alejandro Tirachini, and Nicolas Coulombel</i>	471
Railway Company Management <i>Vilius Nikitinas, Skaistė Miliauskaitė</i>	479
Regulation of Rail Infrastructure and Services <i>Javier Campos</i>	485
Rail Vehicle Classification <i>Christos Pyrgidis and Alexandros Dolianitis</i>	490
Introduction to Maritime Shipping <i>Christa Sys and Thierry Vanelslander</i>	508
The Geography of Maritime Transport <i>César Ducruet and Justin Berli</i>	517
The Future of Maritime Transport <i>Harilaos N. Psaraftis</i>	535
Maritime Transport and Territorial Scale <i>Brian Slack</i>	540
Containerization and the Port Industry <i>Hercules Haralambides</i>	545
Port Management <i>Lourdes Trujillo, Daniel Castillo Hidalgo, and Manuel Herrera</i>	557
Economic and Environmental Regulation in the Port Sector <i>Beatriz Tovar and Alan Wall</i>	563
Maritime Route Planning <i>Johan Woxenius</i>	570
Inland Waterway Transport and Inland Ports: An Overview of Synchromodal Concepts, Drivers, and Success Cases in the IWW Sector <i>Behzad Behdani, Bart Wiegmans, and Yun Fan</i>	577
Maritime Company Passenger Management/Liner Industry <i>Claudio Ferrari and Alessio Tei</i>	587
Cruise Industry <i>Athanasios A. Pallis and Aimilia A. Papachristou</i>	593
International Maritime Regulation: Closing the Gaps Between Successful Achievements and Persistent Insufficiencies <i>Laurent Fedi</i>	600
Ship Classification <i>Gareth C. Burton and Mimosa T. Miller</i>	607
The Shipbuilding Industry and its Interactions With Shipping <i>Paul William Stott</i>	617
Methods for Designing Public Transport Networks <i>Zain Ul Abedin and Avishai (Avi) Ceder</i>	625
Space Transportation <i>Mark Hempsell</i>	638

Pipelines	646
<i>Franco Cotana and Mattia Manni</i>	
Women and Transport Modes	656
<i>Priya Uteng, PhD and Yusak Susilo, PhD</i>	
Transport Modes and Big Data	665
<i>Hannah D Budnitz, Emmanouil Tranos, and Lee Chapman</i>	
Transport Modes and Inequalities	671
<i>Caroline Mullen</i>	
Railway Traffic Management	678
<i>Francesco Corman</i>	
Indoor Transportation	684
<i>Lutfi Al-Sharif</i>	
Bicycle as a Transportation Mode	697
<i>Raktim Mitra and Paul M. Hess</i>	
Urban Air Mobility: Opportunities and Obstacles	702
<i>Adam Cohen and Susan Shaheen, PhD</i>	
Transport Modes and Sustainability	710
<i>Long Cheng, Jonas De Vos, and Frank Witlox</i>	

VOLUME 6

Introduction to Transport Policy and Planning	1
<i>Maria Attard</i>	
Workplace Parking Levy	2
<i>Stephen Ison and Lucy Budd</i>	
Air Transport	7
<i>Lucy Budd and Stephen Ison</i>	
Mobility as a Service	12
<i>Milo N. Mladenović</i>	
Bicycle Sharing	19
<i>Cyrille Médard de Chardon</i>	
Light Rail	31
<i>Fiona Ferbrache</i>	
Planning Tourism Travel	39
<i>Luca Zamparini</i>	
Transport Planning and Management and its Implications in Chinese Cities	44
<i>Mengqiu Cao</i>	
Road Safety	51
<i>George Yannis and Eleonora Papadimitriou</i>	
Mobility Planning and Policies for Older People	59
<i>Charles B A Musselwhite</i>	
Electric Mobility	64
<i>Graham Parkhurst</i>	
Transport Policy and Governance	73
<i>Lisa Hansson</i>	

Transferability of Urban Policy Measures <i>Paul Martin Timms</i>	77
Accessibility Tools for Transport Policy and Planning <i>Benjamin Büttner</i>	83
Demand Responsive Transport <i>Marcus Enoch</i>	87
Transport and Climate Change <i>Robin Hickman and Christine Hannigan</i>	94
Taxicabs and Microtransit <i>David A. King</i>	101
Land-Use and Transport Planning <i>Luis A. Guzman</i>	107
Parking <i>Stephen Ison and Lucy Budd</i>	113
Transport Planning in the Global South <i>Daniel Oviedo and Mariajose Nieto-Combariza</i>	118
Planning for Children's Independent Mobility <i>E. Owen D. Waygood and Raktim Mitra</i>	125
The Politics of Mobility Policy <i>Geoff Vigar</i>	131
Technology Enabled Data for Sustainable Transport Policy <i>Susan M. Grant-Muller, Mahmoud Abdelrazek, Hannah Budnitz, Caitlin D. Cottrill, Fiona Crawford, Charisma F. Choudhury, Teddy Cunningham, Gillian Harrison, Frances C. Hodgson, Jinhyun Hong, Adam Martin, Oliver O'Brien, Claire Papaix, and Panagiotis Tsoleridis</i>	135
Car Sharing <i>Cyriac George and Tanu Priya Uteng</i>	142
Toward a More Holistic Understanding of Mega Transport Project (MTP) Success <i>John Ward</i>	147
Equity Considerations in Transport Planning <i>Karel Martens</i>	154
Planning for Rail Transport <i>Simon P. Blainey</i>	161
Connected and Autonomous Vehicles: Priorities for Policy and Planning <i>Dr. Alexandros Nikitas</i>	167
Gendered Mobility <i>Sheila Mitra-Sarkar</i>	173
Modeling and Simulation for Transport Planning <i>Michela Le Pira, Giuseppe Inturri, and Matteo Ignaccolo</i>	184
Externalities and External Costs in Transport Planning <i>Silvio Nocera</i>	191
Planning for Public Transport with Automated Vehicles <i>Gonçalo Homem de Almeida Rodriguez Correia</i>	198
Urban Congestion Charging in Transport Planning Practice <i>Ida Kristoffersson and Maria Börjesson</i>	206

Energy and Transport Planning <i>Debbie Hopkins and Christian Brand</i>	214
Customer Satisfaction as a Measure of Service Quality in Public Transport Planning <i>Laura Eboli and Gabriella Mazzulla</i>	220
Car Sharing and the Impact on New Car Registration <i>Mario Intini and Marco Percoco</i>	225
Evaluation Methods in Transport Policy and Planning <i>Niek Mouter</i>	230
Transport and Air Quality Planning and Policy <i>Dr Fabio Galatioto</i>	236
Cycling Policies <i>Esther Anaya-Boig</i>	241
Community Severance <i>Paulo Anciaes and Jennifer S. Mindell</i>	246
Planning for Bus Priority <i>Claus H. Sørensen, Fredrik Pettersson, and Joel Hansson</i>	254
High-Speed Rail and the City <i>Marie Delaplace</i>	261
Public Engagement in Transport Planning <i>Miriam Ricci</i>	266
Long-Distance Travel <i>Giulio Mattioli and Muhammad Adeel</i>	272
Urban Freight Policy <i>Laetitia Dablanç</i>	278
Regional Transport Planning <i>Chia-Lin Chen</i>	286
Transport Project Financing <i>Romeo Danielis and Lucia Rotaris</i>	292
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	298
Transitions and Disruptive Technologies in Transport Planning <i>Kate Pangbourne and Maria Attard</i>	309
Community Transport: Filling the Gaps for Those in Need of Mobility <i>Ian Shergold</i>	314
Fundamental Emerging Concepts and Trends for Environmental Friendly Urban Goods Distribution Systems <i>Sandra Melo</i>	320
Plateau Car <i>David Metz</i>	324
Emerging Trends in Transport Demand Modeling in the Transition Toward Shared Mobility and Autonomy <i>Patrizia Franco</i>	331
Policy and Planning for Walkability <i>Carlos Cañas Sanz and Maria Attard</i>	340

Public Transport Subsidy and Regulation <i>Jonathan Cowie</i>	349
Urban Regeneration and Transportation Planning <i>Thomas Vanoutrive</i>	356
Social and Distributional Impact Assessment in Transport Policy <i>Laura Walker and Angela Curl</i>	361
Mobility Planning for Healthy Cities <i>Ersilia Verlinghieri</i>	368
Transport Demand Management <i>Begoña Guirao</i>	374
Planning and the Global Movement of Goods and Commodities <i>Christopher Clott and Chris Petrocelli</i>	380
Public Transport Network Planning <i>Corinne Mulley and John D. Nelson</i>	388
A Timely Perspective on Planning for Ageing Infrastructure <i>Anthony Perl</i>	395
The role of media in transport planning and the transport policy process <i>Özgül Ardiç and J.A. Annema</i>	400
Travel Plans <i>Stephen Potter and Marcus Enoch</i>	408
Home Deliveries and their Impact on Planning and Policy <i>Rosário Macário</i>	413
Planning for Safe and Secure Transport Infrastructure <i>Per Erik Gårder</i>	418
Sensors and Data Driven Approaches in Transport <i>Mohammad Sadrani and Constantinos Antoniou</i>	426
Planning Active Travel and School Transport <i>Fahimeh Khalaj, Dorina Pojani, and Sara Alidoust</i>	432
VOLUME 7	
Introduction to Transport Psychology <i>Carlo Prato</i>	1
From Self-reports to Auto-Tech-Detect (ATD)-based Self-reports in Traffic Research <i>Türker Özkan and Timo Lajunen</i>	2
Observational Field Studies in Traffic Psychology <i>Tova Rosenbloom and Hodaya Levy</i>	8
Driving Simulators <i>Karel A. Brookhuis</i>	14
Naturalistic Driving Studies: An Overview and International Perspective <i>Johnathon P. Ehsani, Joanne L. Harbluk, Jonas Bärghman, Ann Williamson, Jeffrey P. Michael, Raphael Grzebieta, Jake Olivier, Jan Eusebio, Judith Charlton, Sjaanie Koppel, Kristie Young, Mike Lenné, Narelle Haworth, Andry Rakotonirainy, Mohammed Elhenawy, Gregoire Larue, Teresa Senserrick, Jeremy Woolley, Mario Mongiardini, Christopher Stokes, Paul Boase, John Pearson, and Feng Guo</i>	20

A Detailed Approach to Qualitative Research Methods <i>Sonja Forward and Lena Levin</i>	39
Behavioral Change <i>Sonja Haustein</i>	46
Habitual Behavior <i>Carlo G. Prato</i>	54
Driving Behavior and Skills <i>Timo Lajunen and Türker Özkan</i>	59
The Multidimensional Driving Style Inventory <i>Orit Taubman - Ben-Ari</i>	65
Risk Perception in Transport: A Review of the State of the Art <i>Trond Nordfjrn, An-Magritt Kummeneje, Mohsen F. Zavareh, Milad Mehdizadeh, and Torbjørn Rundmo</i>	74
Social-Symbolic and Affective Aspects of Car Ownership and Use <i>Birgitta Gatersleben</i>	81
Pitfalls of Statistical Methods in Traffic Psychology <i>J.C.F. de Winter and D. Dodou</i>	87
Data Analysis: Structural Equation Models <i>Marco Diana</i>	96
Data Analysis: Integrated Choice and Latent Variable Models <i>Carlo G Prato</i>	102
Explaining Data Analysis Using Qualitative Methods <i>Lena Levin and Sonja Forward</i>	107
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	113
Driver Aggression and Anger <i>Mark J.M. Sullman and Amanda N. Stephens</i>	121
Speeding: A "Tragedy of the Commons" Behavior <i>Bryan E. Porter, Thomas D. Berry, and Kristie L. Johnson</i>	130
Cycling as a Mode Choice: Motivational Psychology <i>Sigal Kaplan</i>	136
Motorcyclists <i>Narelle Haworth</i>	144
Drivers' Hazard Perception Skill <i>Mark S. Horswill and Andrew Hill</i>	151
Driver Education and Training for New Drivers: Moving beyond Current 'Wisdom' to New Directions <i>Teresa Senserrick, Oscar Oviedo-Trespalacios, David Rodwell, and Sherrie-Anne Kaye</i>	158
Road Safety Advertising: What We Currently Know and Where to From Here <i>Ioni Lewis, Barry Watson, Katherine M. White, and Sonali Nandavar</i>	165
Traffic Law Enforcement Theories and Models <i>Richard Tay</i>	171
Satisfaction with Travel and the Relationship to Well-Being <i>Tommy Gørling and Filip Fors Connolly</i>	177
	182

Electromobility: History, Definitions and an Overview of Psychological Research on a Sustainable Mobility System <i>Josef F. Krems and Isabel KreiÃŸig</i>	
Sharing: Attitudes and Perceptions <i>Yi Wen and Christopher R Cherry</i>	187
Women's Travel Patterns, Attitudes, and Constraints Around the World <i>Sandra Rosenbloom</i>	193
Driver Stress and Driving Performance <i>Lisa Dorn</i>	203
Introduction to Sustainability and Health in Transportation <i>Roger Vickerman</i>	225
Sustainable Development Goals and Health <i>Rosa Surinach</i>	226
Human Ecology <i>Roderick J. Lawrence</i>	234
Car- Free Cities <i>Haneen Khreis and Mark J. Nieuwenhuijsen</i>	240
Superblocks Base of a New Model of Mobility and Public Space. Barcelona as an Example <i>Salvador Rueda Palenzuela</i>	249
Resilience of Transport Systems <i>ErikJenelius and Lars-GöranMattsson</i>	258
Impact of Shipping to Atmospheric Pollutants: State-of-the-Art and Perspectives <i>Daniele Contini and Eva Merico</i>	268
Noise Pollution From Transport <i>Marianna Jacyna, Emilian Szczepański, Konrad Lewczuk, Mariusz Izdebski, Ilona Jacyna-Gotda, Michał, Kłodawski, Paweł, Gołda, Piotr Goł, and Ebiowski</i>	277
Visual Impacts From Transport <i>Paulo Rui Anciaes</i>	285
Light Pollution <i>John D. Bullough</i>	292
Wildlife Crossings and Barriers <i>Scott D. Jackson</i>	297
Environmental Justice, Transport Justice, and Mobility Justice <i>Devajyoti Deka</i>	305
Transport Noise and Health <i>Elisabete F. Freitas, Emanuel A. Sousa, and Carlos C. Silva</i>	311
Climate Change and Health, Related to Transport <i>Ersilia Verlinghieri</i>	320
Urban Greenspace, Transportation, and Health <i>Payam Dadvand and Mark J. Nieuwenhuijsen</i>	327
Transport Access and Health <i>Alireza Ermagun</i>	335
Social Exclusion and Health, Related to Transport <i>Roger L. Mackett</i>	341

Burden of Disease Assessment <i>David Rojas-Rueda</i>	347
Achieving a Near-Zero CO <i>Lewis M. Fulton</i>	353
Disabled Travelers <i>Bryan Matthews</i>	359
Health Impacts of Connected and Autonomous Vehicles <i>Soheil Sohrabi</i>	364
Electric Vehicles and Health <i>Kanok Boriboonsomsin</i>	372
Shared Mobility Opportunities and their Computational Challenges for Improving Health-Related Quality of Life <i>Cristiano Martins Monteiro, Cláudia Aparecida Soares Machado, Adelaide Cassia Nardocci, Fernando Tobal Berssaneti, José Alberto Quintanilha, and Clodoveu Augusto Davis</i>	376
Bike Sharing and Health <i>David Rojas-Rueda and Mark J. Nieuwenhuijsen</i>	384
E-Bikes and Health <i>Aslak Fyhri and Hanne Beate Sundfør</i>	393
<i>Index</i>	399

Introduction to Traffic Management

Edward C.S. Chung, Department of Electrical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

© 2021 Elsevier Ltd. All rights reserved.

Traffic management concerns the guidance and control of all types of vehicles, pedestrians and cyclists with the goal of providing a safe and efficient movements of people and goods. Road based traffic management can be broadly categorized under arterial roads management, freeway management, incident management, and parking management. Many diverse users share a limited road space in urban areas. The orderly and safe movement of traffic on arterial roads are attributed to good design of uncontrolled intersection (e.g., turbo roundabout), coordinated traffic signal control and designated lanes for public transport and cyclists. Freeways are high-capacity high-speed roads designed to carry high traffic volume between urban centers. Freeways are crucial in the functioning of cities. Often freeway capacity is degraded, for example through the development of bottlenecks, resulting in congestion and reduced mobility. Ramp metering and variable speed limit (VSL) are some of the controls used to regulate traffic entering freeways and to reduce vehicle speed approaching stationary queues. Air traffic management in broad term includes services provided to an aircraft during the flight, measures applied to reduce demand temporarily when demand exceeds capacity and airspace management. Similarly, port management concerns the strategic decision-making, monitoring and controlling of ports and terminals.

The articles in this section present a range of view across topics. Topics covered include common traffic phenomena such as capacity drop, flow breakdown, and recurrent and nonrecurrent congestion. Arterial traffic management such as signal coordination, permitted and permissive phases, hook turns, signalized roundabout, turbo roundabouts, and emergency vehicle priority. Public transport aspects such as adaptive bus control, transit priority, railway crossing, and tram lane configurations and driving rules. Articles on freeway management included city wide coordinated ramp meters, VSL, hard shoulder running, reversible lane, high occupancy vehicles lanes, and high occupancy toll lanes. Incidents contribute to a significant amount of road congestion and their impact can be mitigated by early detection and quick incident clearance. For this aspect, topics covered included incident detection technologies and incident management. Active transport such as bicycle sharing, and real time information systems such as parking information systems and advance travelers information systems are covered. Besides hardware control, road pricing as an economic mechanism to include externalities cost to reduce excess demand, and free flow electronic toll collection are covered. Non-road-based transport such as air traffic management and port management are also included.

Traffic Management

Urban Motorway Management

Keeping Urban Motorways Safe and Efficient With City Wide Coordinated Ramp Meters

John Gaffney, Hendrik Zurlinden, Department of Transport (DoT), VicRoads, Kew, VIC, Australia

© 2021 Elsevier Ltd. All rights reserved.

Optimizing Urban Motorways Networks With Ramp Meters	2
Introduction	2
Ramp Metering can Provide Mixed Findings	3
Global Trends Increasing the Importance of Traffic Control	3
Ramp Meters as the Only Practical Means to Sustain Maximum Flows	4
Why are Urban Motorways Important to Manage?	4
Introduction	4
Role of Managed Motorways in an Overall Transport Network	4
Role of Motorways in the Economy	5
Customer Expectations and the Role of the Road Operator	7
Effective Motorway Control is Complex and Elusive	7
Traditional Approaches to Design and Control of Urban Motorways usually do not Lead to Optimal Outcomes	7
A Short History of Ramp Metering	8
The US Experience	8
Pioneering Endeavours of Ramp Metering in Melbourne (1971–today)	8
Toward Contemporary Approaches to Design and Control of Urban Motorways Leading to Optimal Outcomes	9
Acknowledgment	9
See Also	9
References	9
Further Reading	9

Optimizing Urban Motorways Networks With Ramp Meters

Introduction

Urban motorways are one of the most critical road network assets and are usually the most expensive road assets to provide, maintain, and operate. It is important they are always managed to provide the best safety and economic outcomes for the community.

Unfortunately, these important road assets are often observed to perform poorly under congested conditions even when there is considerable deployment of intelligent transport systems (ITS) assets and systems. Substandard motorway operating conditions result in more casualty crashes, longer travel times, unreliable journeys, higher energy use, and higher vehicle emissions. When bottlenecks are allowed to form anywhere in the motorway, they quickly develop and can degrade performance over many kilometers, often affecting 10 or more kilometers upstream.

There are now thousands of ramp meters operating around the world although primarily in the USA. The following quote from the Federal Highway Administration brochure ([Federal Highway Administration, 2006](#)) demonstrates the value of ramp metering systems:

"Every evaluation of the system has shown reduced accidents, reduced delay and increased volumes when metering was installed. No other traffic management strategy has shown the consistently high level of benefits in such a wide range of deployments from all parts of the country".

Pete Briglia, Puget Sound Regional Council, Seattle, Washington and Chair of the TRB Freeway Operations Committee.

Outside of the USA other countries such as the United Kingdom have focused more on other ITS tools such as variable speed limits (VSL) to optimize traffic flows. The role of VSL application in managing traffic flows is discussed in [VicRoads \(2020a\)](#). The

authors of this chapter believe ramp metering is the most effective tool for managing traffic flows on urban motorways, it is only with appropriate mainline geometry and ramp design (both entry and exit ramp) combined with the appropriate application of traffic theory using advanced control systems that significant gains in both safety and efficiency can be achieved. Enhanced optimization will occur when City-Wide Coordinated Ramp Metering (CWCRM) is fully integrated with Dynamic VSL which is currently under development in VicRoads where the integrated DVSL and ramp metering control algorithms, which is a complex problem to research and understand, have been developed and the business case for full deployment in the field is currently underway.

Ramp Metering can Provide Mixed Findings

Unfortunately, there are many mixed findings in relation to ramp meters and there are many differences in the way ramp metering is implemented around the world, and in how performance measurement is undertaken. The authors recognize that achieving sustainable and reliable traffic flow with ramp metering is a fine art and it is possible without a thorough understanding of traffic theory relating to urban motorways (VicRoads, 2020a) to invest a lot of money and resources and achieve very little sustainable economic benefit for the community.

Sometimes ramp meters can shift the location of critical bottlenecks further downstream and in doing so create stronger bottlenecks. Some forms of ramp metering only delay the onset of flow breakdown. This in itself has an economic benefit (e.g. if it delays the onset of flow break down by 10–15 min each day this benefit can be calculated). However, in large modern urban cities with high motor car usage such as Melbourne (Population > 5 million) and in cities as defined in Section 3.2.1 of VicRoads (2019b), with a strong service-based economy with high numbers of commercial vehicles, where peak periods are of 3–5 h duration with little relief in traffic flows between peaks, this is not a viable solution.

Urban motorway networks typically have highly variable and random traffic patterns and demands which change minute to minute, hour to hour, day to day and season to season, and which also change markedly over time with the growth and development of the city. What is required is an effective ramp metering regime that can deliver sustainable traffic flows, reliably with minimal delays providing benefit over the entire day, and every day a week and which responds dynamically to the constantly variable nature of system demands.

Global Trends Increasing the Importance of Traffic Control

In the context of transport, motorways are the main arteries of any major conurbation worldwide. It is important that these arteries are as far as possible protected from influences impacting on their performance (like blood clots in this analogy). While other arteries are impacted by all sorts of influences from adjacent land uses that are difficult to control (e.g., driveways, local roads entering), motorways due to the limited access to them can relatively easily be protected.

The stress on motorways is steadily increasing. This is caused by the following global trends:

- Urbanization;
- Growing affluence and hence vehicle ownership;
- Increasing separation between home and the place of work and increasing number of trip purposes other than commuting; and
- Decreasing performance of other arteries directing more traffic toward the motorway network.

Local road authorities cannot (or only to a very limited extent) influence these global trends. Their area of influence is largely limited to the roads they plan, build, operate and maintain and not to planning and population growth, for example, immigration in Australia.

Generally, a *laissez-faire* approach to motorway management should not be taken because unconstrained access to a motorway will result in congestion characterized by different types of performance losses such as:

- “Capacity Drop”, “Throughput drop” or “Flow Loss” (a major economic loss due to less vehicles being serviced from the infrastructure across an entire motorway corridor or network)
- “Speed Drop” (a major economic loss due to lower average speed resulting in cumulative hours of delay)
- “Delay Escalation” (as delay increase is not linear as speed deteriorates, i.e., the economic loss increases exponentially as motorway speeds decay).
- “Safety Drop” (increased crash risk occurs around the density values associated with flow breakdown).
- “Increased Emissions” (from driving in congested and potentially stop–start conditions and from diversions which result in longer trips).
- “Increased Energy Use” (increased fuel consumption when driving in congested and potentially stop–start conditions and from diversions which result in longer trips).
- “Decreased Reliability” (when any bottleneck sets in and the travel times increase significantly, the flow and speed outcomes can be very variable).

Ramp metering to control access to the motorway is in the area of “damage control.” There may be some negative impacts such as waiting times at the on-ramps but this is generally far outweighed by positive impacts such as the avoidance of the above performance loss phenomena. It is an important point that, with the global trends as listed above, the net benefit will only increase everywhere,

whether it is in major conurbations in middle Europe or North America or in metropolises in China or India. In the above analogy with the human body, no doctor would ever allow a patient's main arterial getting clogged as a consequence of unhealthy diet, nor can any government afford letting their most important transport infrastructure (over time) getting dysfunctional.

Ramp Meters as the Only Practical Means to Sustain Maximum Flows

Effectively managing motorway networks requires limiting the flows to all sections of the motorway to sustainable flow values based on the temporal loading of the network or corridor (based on continuously measured sub-1 min-interval real-time data). This is achieved not through managing traffic flows, but rather traffic densities at every cross-section (typically 500 m apart) throughout the network. Ramp meters when designed and operated properly and accompanied by appropriate mainline geometric design using advanced automated real-time control systems is currently the most efficient and cost-effective way to manage traffic densities and thus the safety and efficiency of urban motorways.

When implemented on a city-wide basis as in Melbourne, it can deliver sustainable and reliable traffic flows in the order of 20% more veh/lane/h at a 20 km/h higher speed with a 30% reduction in casualty crashes. This is likely to be the case for many decades to come despite promises and predictions that autonomous vehicles will solve the problem and increase traffic flow markedly. These predictions are often model-based with minimal appreciation of road design elements (i.e., there is generally no appreciation of the geometry (physics) that keep the vehicles on the road at certain speeds, traffic engineering science, or the randomness of minute to minute traffic patterns in service-based economies). The variability of trip length, vehicle mix, and the need for mandatory lane changes required just to fill up all lanes to capacity are often overlooked in these models. For example, routinely measured on Melbourne's motorways are around 2000 lane changes per hour per km on a 4-lane motorway. This high number of lane changes is the minimum necessary to keep the motorway at optimal traffic levels (VicRoads, 2019b, 2020a). However, lane changing creates an internal friction within the carriageway (i.e., "lane shear") that needs to be measured continuously (e.g., density spikes (moving vehicle clusters)) and its effects managed dynamically, in real-time with effective ramp metering.

Why are Urban Motorways Important to Manage?

Introduction

It is the importance of urban motorways as highlighted by their contribution to daily economic activity in large urban areas that justifies investment in ramp metering. In Melbourne, urban motorways represent only 7% of the total urban arterial road pavement area; however, annually they typically facilitate 40% of travel as measured in vehicle kilometers travelled (VKT)). Reliance on the motorway network to provide arterial road travel is trending upward, with a 50% share likely in the relatively near future. Transport demand across the entire road network has increased over time placing more stress on these important assets, if uncontrolled resulting in regular flow breakdown and congestion over longer periods of the day. Commensurate with this increasing stress seems to be an increasing number of fatal and serious casualty crashes across many parts of the road network. Managed motorways have been effective in reducing and maintaining lower crash rates on urban motorways (VicRoads, 2017, 2020b).

Congestion with the consequences described above (e.g., "Throughput drop" by up to 20% and "Speed drop" by up to 50%) can impact several kilometers of the motorway upstream and can trap and delay hundreds and often thousands of motorists every hour. This can affect motorists even if they are not travelling as far as the location where the flow breakdown occurred (generally the most downstream point of the congestion). The performance of the motorway can also impact the broader surface road network.

Traffic congestion is not new. Since the introduction of uninterrupted flow facilities, in the form of freeways, flow breakdown has been present and recognized as problematic. With demand now regularly exceeding capacity for longer periods of the day, congested conditions spread temporally and geographically over time. Attempts to reduce or remove motorway congestion from urban networks have been implemented, usually targeted at increasing physical capacity (e.g., adding lanes to existing facilities and/or deploying some ITS assets and systems) to meet growing demands. However, whilst physical expansion does increase the ability to service more trips per unit time and tens of thousands more trips over the day, more lanes alone does not guarantee elimination of congestion or the assurance that available road space is utilized efficiently.

Role of Managed Motorways in an Overall Transport Network

One method used by VicRoads for understanding the effective output of a motorway system is to sum all the entry or exit flows over a day (refer to Fig. 1). Across the 75 km of the urban M1 corridor, it has been calculated that up to 900,000 vehicle journeys (trips) are serviced each day, moving more than 1 million people across Melbourne. This is the equivalent of more than 1000 fully loaded and packed train trips based on Melbourne trains (6–8 car train sets). Each trip on average travels around 15–25 km on the motorway, although the trip length varies considerably by motorway route in Melbourne and by time of day.

While this type of analysis has not traditionally been undertaken on road corridors, it is representative of person trips in a section of the transport system, similar to analysis undertaken for portions of the public transport system (e.g., boarding figures). The scale of these figures clearly demonstrates the significant task that urban motorways facilitate within the transport system. Travel surveys undertaken in 2012/13 (Department of Economic Development, Jobs, Transport and Resources, 2015) estimate that 12.3 million trips are made across Melbourne every weekday, including those using public transport and active modes (walking and cycling). The

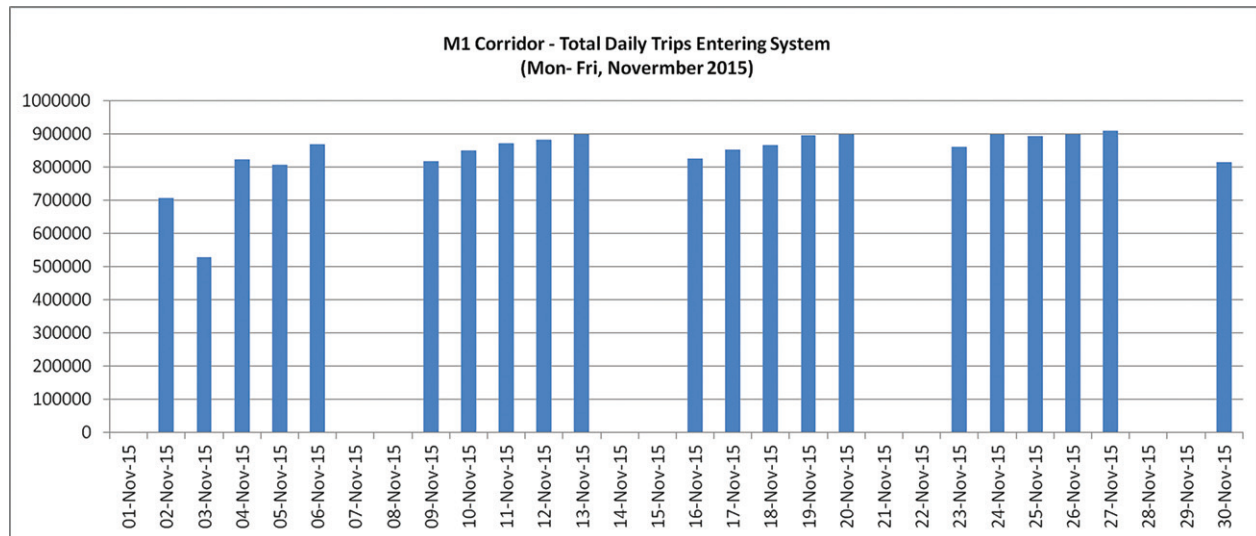


Figure 1 Weekday daily entry flows November on the M1 motorway corridor in Melbourne.

M1 motorway corridor in Melbourne alone, representing approximately 2% of the total urban arterial road pavement facilitates 6%–7% of the daily urban trips.

The cross-sectional annual average daily traffic volume (AADT) along the M1 Motorway corridor is approximately 160,000–170,000, in several locations peaking at around 220,000 vehicles/day. The above figures (i.e., the number of trips serviced when compared to AADT) imply that the average trip length is relatively short, that is trips turn over about 3–5 times along the 75 km corridor depending of the time of day. The high turnover rate combined with trip length variability over the day is just one factor revealing some of the complexity of the motorway operation due to different trip patterns being catered for as road users travel the M1 corridor.

Role of Motorways in the Economy

Due to their contribution to the broader transport task and the Victorian and Australian economies and employment and growth, the enormous scale of the work done by urban motorways and the need for sustained motorway productivity should not be underestimated. However, the lack of control on most urban motorways generally in Australia and globally, which allows flow breakdown to occur with regularity, indicates that the importance of these high cost facilities to the economy is not fully acknowledged by their asset owners.

Historically, heavy traffic flow and congestion in urban metropolitan areas used to be contained to typically 1–2-h peak periods in the morning and afternoon. As the economies in metropolitan areas have progressively shifted from manufacturing to service provision (service-based economy) the dominant travel purpose of commuting to and from a regular place of employment at the start and end of the day has also shifted. Dominant trips now include those that are associated with a range of working hours and activities that require specific road transport modes for employment (e.g., a tradesman’s mobile “workshop”, a courier’s van, a large freight vehicle or a Uber/taxi driver).

Roads are now “places of employment” with services to properties and product deliveries being a necessary component of most businesses operating efficiently—every premise (e.g., factory, retail, business, and private residence) is a potential service point every day. In a modern service-based economy, many vehicles on weekdays *usually* do not need to have more than one person in the vehicle to be considered efficient (e.g., freight, business, delivery and other services to properties *generally* only require a driver, e.g., a nanny, a painter or a cleaner). Hence measuring and having pre-defined targets for person occupancy per vehicle is not reflective of the true utilization, economic contribution, or efficiency of road transport systems (as compared to public transport with its limited origins and destinations—e.g., following fixed tracks and defined routes with limited trip purposes, e.g., commuting). This is particularly important to understand by those who measure and use person-occupancy data, as commuting trips typically comprise less than a quarter of the road transport task and thus reporting declining person-occupancy averages results in convenient stereotyping based on misunderstanding of how roads are used by the community to enable livability and the efficient supply and distribution of goods and services.

Across the day, approximately 20%–25% of trips on the motorway system and broader road network are solely commuting related (refer to Fig. 2). These figures are similar in many large cities around the world with large service based economies such as London. While there may be some periods of the day (e.g., 6 peak hours representing 25% of all hours of the day) when commuting represents a higher proportion of all trips than during other times of the day (during morning and afternoon peaks, commuting can represent 50%–60% of trips) many of these trips are chained with other activities (refer to Levinson et al., 2017).

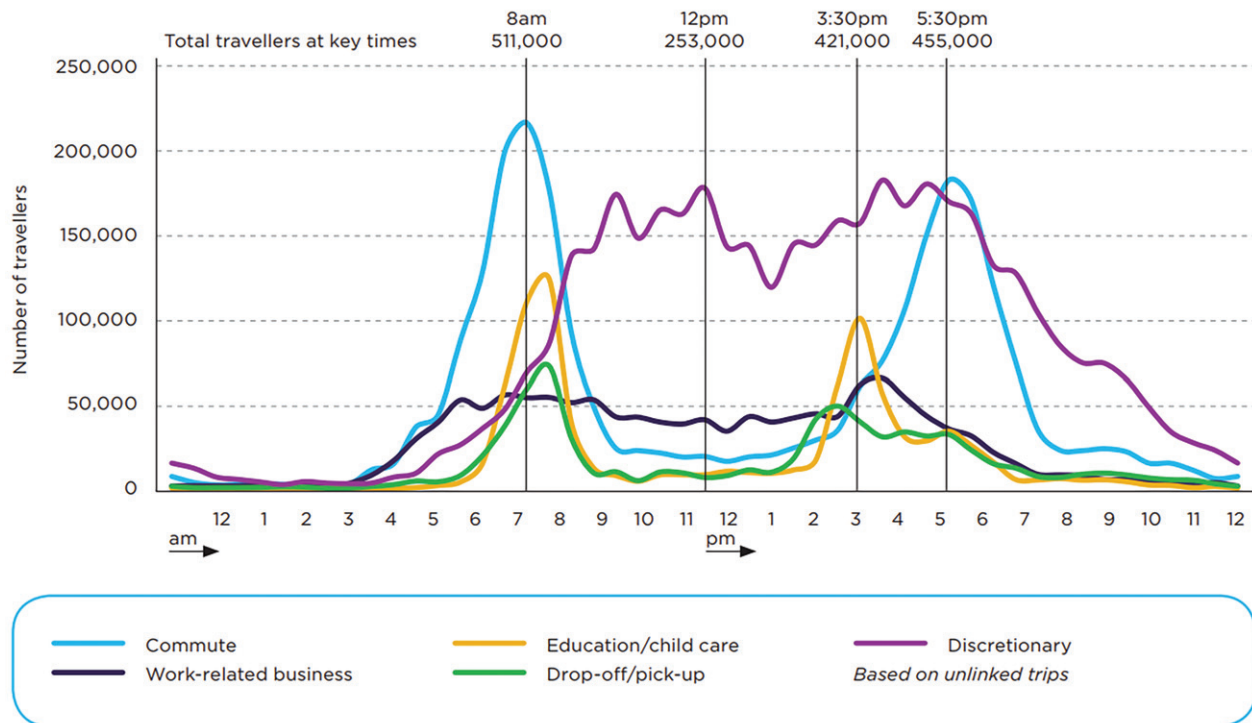


Figure 2 Distribution of trips by purpose in Sydney. Source: Figure 4.21 of NSW Long Term Transport Master Plan, December 2012.

The other significant development is the large contribution and steady growth in the freight transport task which can be illustrated by some numbers on freight transferred from the Port of Melbourne to the urban motorway network:

- Around 3000 ships annually docking at 34 berths;
- Over 2.64 million Twenty-foot equivalent units (TEU) handled annually (equivalent to over 7200 TEU per day); and
- 1000 new motor vehicles handled per day on average.

Although freight movements from the Port of Melbourne are significant they only represent about 5% of the freight movement in urban Melbourne. The urban freight task has also seen large growth in the “Just-for-Me” (personal consumer products as fast as possible to a noncentralized location of choice) parcel and service delivery via the growth in “small white van” phenomena where products and services are often purchased on-line and delivered at any time of the day, anywhere. This may be my home, my office or wherever I may be, for example flowers to the restaurant, a dog wash to my home and taxi/car hire services, etc.

The shift in road use and increase in overall demand has extended peak periods on all arterial roads, especially urban motorways. Periods between the traditional morning and afternoon peaks are now heavily trafficked (refer to Fig. 3) and various locations within the road network now experience peak period type demands, variable congestion and reduced performance throughout the day. This is especially true on some urban motorways in Melbourne such as the M1 corridor where more than 50,000 trips per hour are serviced all day, reaching 60,000 trips per hour in peak periods.

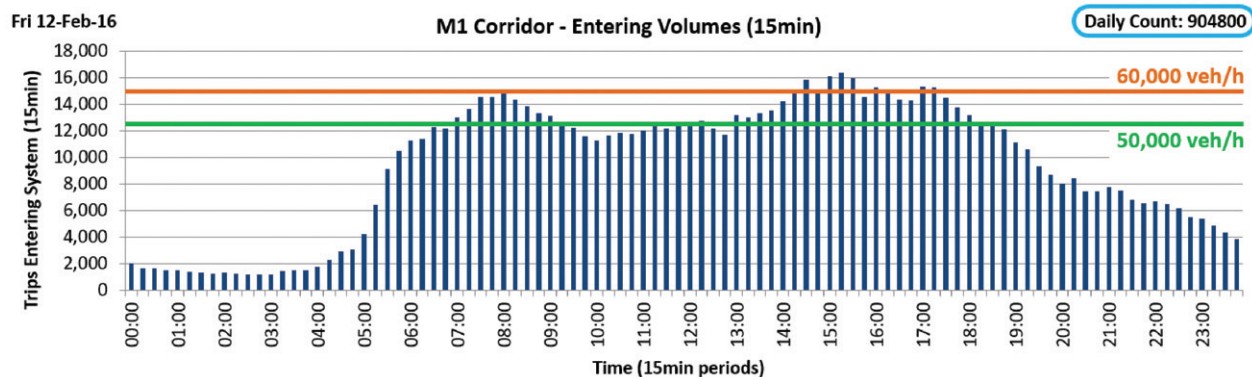


Figure 3 Total system flows entering the M1 corridor over 24 h-15 min periods.

As a result, for example, absolute truck numbers on the West Gate Bridge in Melbourne are in the order of 30,000 per day and small commercial “white vans” are in addition to this number. It is obvious that this activity is essential for the Victorian economy and road agencies have a focus on smoothly and seamlessly processing these trips on the urban motorway network as there is currently no alternative mode of transport available to facilitate and service these trips.

Customer Expectations and the Role of the Road Operator

Finally, most important are the expectations of the customer.

Motorway optimisation is about achieving “Good quality journeys to as many people as possible” (safe, fast, and reliable with high productivity/asset utilization). Planning and designing for these objectives needs understanding and requires determination and application of “Maximum Sustainable Flow” values (refer to [VicRoads, 2019b](#)). Utilizing such values achieves a good balance between higher speed and flows as well as a low probability of flow breakdown on the section/link/route/network levels (refer to [VicRoads, 2019b, 2020a](#)).

Many highway and road agencies have traditionally structured themselves and invested in intervention tools focused on responding to unplanned events, such as incidents and other disruptions that directly impact motorway capacity. In the past, it is acknowledged that incidents were a significant cause of congestion; however, recurrent congestion due to demand induced flow breakdown is now likely to be more prevalent on most urban motorways around the world.

Available sources suggest that more than half of the total congestion delays on urban motorways in large urban centers ([VicRoads, 2019b](#)) are due to recurrent congestion, and as it is acknowledged that growing recurrent congestion also increases the likelihood of triggering compounding nonrecurrent conditions ([Gamon, 2005](#)). This shift requires road operators to change their approach to investing in motorway assets—both civil and ITS—and develop capabilities in understanding the optimal performance of networks under all types of disruption.

This in no way reduces the requirement for road operators to have the necessary tools and resources to respond to incidents and events; however, there is an observed imbalance in the commitment of resources to maintaining optimal flows. Moving forward road agencies will require to have and adequately resource both the optimization function and the incident management function (i. e., proactive avoidance of congestion and crashes, followed by rapid recovery from incidents and breakdowns). This could also significantly reduce the number of incidents occurring across urban networks (refer to [VicRoads, 2020b](#)). Additionally, the tools required for optimization have significant synergies with those required for incident management.

Effective Motorway Control is Complex and Elusive

Traditional Approaches to Design and Control of Urban Motorways usually do not Lead to Optimal Outcomes

Traditional motorway design has been based on facilities that rarely operate at, or can sustain near capacity flows. In urban environments around the world, demands now regularly exceed the capacities of existing motorways which requires a changed approach to designing and operating various motorway elements in order to limit turbulence when flows are at or near capacity. Due to the variation of operating conditions on every segment of an urban motorway, applying the same solutions to all contexts is not appropriate.

In recent decades, in addition to providing expanded road space, road operators have also invested in technology solutions (ITS) as a means to reduce congestion and increase the efficient use of available road space. Some of these methods have achieved higher efficiency for a time; however, the ability to truly optimize the utilization of motorways and sustainably limit the onset of congestion is still broadly elusive. This is in part due to insufficient sophistication or poor implementation of the control system, for example simplistic algorithms for localized or coordinated ramp metering, or implementing dynamic VSL alone without supporting demand management. In some cases, due to the lack of understanding of detailed traffic engineering science within road agencies, the choice of tools and their application can be potentially detrimental to system performance, significantly reducing safety and efficiency from the outset. In fact, some ITS tools and operational practices can actively work against each other if used inappropriately.

Identifying the causes of congestion on motorways and quantifying its impacts requires detailed and accurate data from the motorway system. It has been observed that many road agencies have limited datasets and have lacked access to a comprehensive performance monitoring system. Often the performance monitoring system has limited visualization tools and uses data which lacks accuracy and has insufficient time and space resolution. It also often lacks ease of comprehensibility by practitioners and decision makers.

In many cases the available data has been the product of motorway detection systems originally being installed for identifying general use volumes (hourly and daily data) or identifying abnormal operations such as incidents. Whilst such systems may have served their purpose when first installed, the primary cause of congestion has shifted over time from being randomly induced by incidents to being substantially recurrent due to oversaturation at points or over extended lengths of motorway. Today in Australia casualty crashes are often likely to be caused by congestion itself (or near saturated conditions) rather than being the primary cause of congestion as was the case 20 or so years ago. Significantly improved data (additional relevant metrics, higher resolution and improved quality) is required for both recognizing the detailed causes of congestion and also to have the ability to maintain stable traffic flows.

It has been observed that whilst road operators have, with good intention, attempted to reduce or eliminate the occurrence of urban motorway congestion, the adopted practices have predominantly treated observed symptoms, not necessarily the actual cause (s) of recurrent congestion. The practices adopted by many road operators are often copied from one jurisdiction to another, whether they were successful in deployment or not (i.e., success not demonstrated through appropriate performance metrics showing sustainable long-term operational performance gains), and often without sufficient understanding of the actual impacts (or system operational principles) of adopted solutions to address the unique local conditions.

A Short History of Ramp Metering

The US Experience

The US Federal Highway Administration Ramp Management and Control Handbook ([Federal Highway Administration, 2006](#)) indicates that the first ramp metering was installed in 1963 on Chicago's Eisenhower Expressway. Ramp metering was developed as a technique to manage traffic demand following the launch of the Interstate Highway Program to address freeway flow problems associated with congestion and safety.

The initial application of entry ramp metering used a police officer stationed on the entrance ramp to stop traffic and release vehicles one at a time at a rate determined from a pilot detection program. This use of metering followed successful tests of the effectiveness of metering traffic entering New York tunnels and ramp closure studies in Detroit ([Piotrowicz and Robinson, 1995](#)).

The Minnesota Department of Transportation (Mn/DOT) has used ramp meters since 1969. In 2000 the Minnesota Legislature required Mn/DOT to study the effectiveness of ramp meters in the Twin Cities Region ([Cambridge Systematics Inc., 2001](#)) by conducting a shutdown study, that is comparison of the situations with and without ramp meters active. The evaluation report by [Cambridge Systematics Inc. \(2001\)](#) indicated significant benefits of ramp metering in the following areas:

- Traffic volumes and throughput
- Travel time
- Travel time reliability
- Safety
- Emissions
- Fuel consumption
- Benefit/cost analysis

There are now thousands of ramp meters operating in the USA. This is seen as a measure of the benefits of ramp metering installations. The following quote from the Federal Highway Administration brochure ([Federal Highway Administration, 2006](#)) demonstrates the value of ramp metering systems:

"Every evaluation of the system has shown reduced accidents, reduced delay and increased volumes when metering was installed. No other traffic management strategy has shown the consistently high level of benefits in such a wide range of deployments from all parts of the country".

Pete Briglia, Puget Sound Regional Council, Seattle, Washington and Chair of the TRB Freeway Operations Committee.

Pioneering Endeavours of Ramp Metering in Melbourne (1971–today)

The first ramp metering in Australia was provided in Melbourne on the South Eastern Freeway (now the M1 - Monash Freeway/Southern Link) at the Gibdon Street entry ramp. The ramp metering was initiated in 1971 by Mr. Kerras Burke of the Highways Division of the Melbourne and Metropolitan Board of Works (MMBW). The rate of vehicle entry to the freeway was based on data from detectors on the freeway. The traffic was regulated by varying the phase times at traffic signals at the entry ramp from the Gibdon Street/Barkly Avenue intersection. The metering had limited success due to the high freeway flows. The limited spare capacity resulted in low ramp phase time for allowing additional vehicles to enter from the ramp. The release of short platoons from the controlling signals at the top of the ramp (rather than signals close to the nose with one vehicle per green) was also less than desirable. The metering was eventually deactivated due to driver complaints about short phase times, lack of publicity to inform motorists, noncompliance and lack of enforcement.

Some 30 years later in 2002, VicRoads commenced the installation of ramp metering at motorway interchanges to reduce motorway traffic congestion and to improve merging. The entry ramp metering operated in an isolated manner with several fixed time cycles to manage the rate at which vehicles could join the motorway. Over the next few years, 11 more ramp meters were successfully installed, dependent on what the objectives of the assessment were. These ramp meters were initially seen as effective as they increased merge capacity and stability near the ramp merge and delayed the onset of flow breakdown sometimes by only a matter 10–15 min.

However, when a systemwide network performance analysis was undertaken, it was shown that improving the merge capacity alone and allowing more vehicles to enter the system could also during some peak periods increase the strength of other latent or critical bottlenecks further downstream, resulting in further degradation of the system that may not have occurred before the ramp meters were installed. Hence a new approach to ramp metering was needed which would take a helicopter view of the network and make sure that all traffic arriving at bottleneck areas could be effectively managed and also that ramp queues could be managed more effectively to limit interference with the arterial road network.

In 2008, a trial of six sequential dynamic coordinated ramp meters resulted in increased motorway performance and improved control to minimize flow breakdown. This was based on the effective management of inflows over a length of motorway to match the capacity of the mainline at each merge as well as at other critical bottlenecks. Due to the success of the trial by 2010 VicRoads installed over 60 coordinated ramp meters to manage the entire M1 Motorway within urban Melbourne. A key objective of this installation was to guarantee equity of access along the entire corridor by balancing queue lengths and waiting times. Subsequently Melbourne now has around 130 ramp meters installed on more the 50% of the network including numerous motorway-to-motorway ramp meters at system interchanges with many more presently under construction. Notably, the first system interchanges metered were associated with private toll roads interfacing with the state managed urban motorways. Today ramp metering and other managed motorway tools are supported by managed motorway policy such as it is mandatory on all new urban motorway and motorway upgrade projects in urban Melbourne.

Toward Contemporary Approaches to Design and Control of Urban Motorways Leading to Optimal Outcomes

Within the context described in above, sustained motorway optimization has remained elusive. In most urban networks, peak hour demand still exceeds capacity, that is not all trips can be facilitated at the desired time of day. The gap between *laissez-faire* performance (unmanaged operation resulting in flow breakdown and more crashes) and the desired state (stable and sustained high flow) can be closed by thorough intelligent investigation of the problem and deployment of solutions. The incorporation of state-of-the-art motorway design and traffic control maintains higher speeds at optimum flows for best productivity and also lower traffic density, the latter clearly showing a significant reduction in crash risk (refer to encyclopedia chapters on *Ramp Metering Application* and *City-Wide Coordinated Ramp Metering*).

Acknowledgment

VicRoads acknowledges the partnership and contributions of TRANSMAX as the owner and developer of the VicRoads Motorway Management system (STREAMS) and Prof. Markos Papageorgiou and Prof. Ioannis Papamichail from the Technical University of Crete in the development of the HERO Live coordinated ramp metering system used to manage Melbourne's motorways.

See Also

Ramp Metering Application; City-Wide Coordinated Ramp Metering

References

- Federal Highway Administration, 2006. "Ramp Management and Control Handbook", Federal Highway Administration (FHWA), Washington, D.C.
- Gamon, T., 2005. Tackling congestion. In: TRL News, Transport Research Laboratory (TRL), Crowthorne.
- Levinson, D., Marshall, W., Axhausen, K., 2017. Elements of Access - Transport Planning for Engineers - Transport Engineering for Planners. Network Design Lab.
- Cambridge Systematics Inc., 2001. Twin Cities Ramp Meter Evaluation Final Report. Minnesota Department of Transportation, Saint Paul.
- Piotrowicz, G., Robinson, J., June 1995. Ramp Metering Status in North America (FHWA), U.S. Department of Transportation (USDOT), Washington, D.C. Retrieved from Publication DOT-T-95-17.: http://www.itsdocs.fhwa.dot.gov/JPODOCS/REPTS_PR/3725.pdf.
- VicRoads, 2017. Managed Motorway Framework. VicRoads, Melbourne.
- VicRoads, 2019b. VicRoads MMDG Volume 1, Part 3. Motorway Capacity Guide. Department of Transport (VicRoads), Melbourne.
- VicRoads, 2020a. VicRoads MMDG Volume 1, Part 2: -Traffic Theory Relating to Urban Motorways. Department of Transport (VicRoads), Melbourne.
- VicRoads, 2020b. VicRoads MMDG Volume 1, Part 4: Road Safety on Urban Motorways. Department of Transport (VicRoads), Melbourne.

Further Reading

- Treiber, M., Kesting, A., 2013. Traffic Flow Dynamics – data, models and simulation. Springer, Berlin.
- Kerner, B., 2017. Breakdown in Traffic Networks - Fundamentals of Transportation Science. Springer, Berlin.
- Newman, L., Dunnet, A.M., Meis, G.J., 1969. An Evaluation of Ramp Control on the Harbour Freeway in Los Angeles. California Division of Highways, Los Angeles.
- VicRoads, 2016. VicRoads Operational Control of the Motorway Network – endorsed. VicRoads, Melbourne.
- VicRoads, 2017. Managed Motorway Framework. Department of Transport (VicRoads), Melbourne.
- VicRoads, 2019a. VicRoads MMDG Volume 2, Part 3. Motorway Planning and Design". Department of Transport (VicRoads), Melbourne.
- VicRoads, 2020c. VicRoads MMDG Volume 1, Part 5: - Linking Investment and Benefits Approach. Department of Transport (VicRoads), Melbourne.
- VicRoads, 2020d. VicRoads MMDG Volume 2, Part 2, Network Optimisation Tools". Department of Transport (VicRoads), Melbourne.

Ramp Metering Application - Keeping Urban Motorways Safe and Efficient With City Wide Coordinated Ramp Meters

John Gaffney, Hendrik Zurlinden, Department of Transport (DoT) - VicRoads, Kew, VIC, Australia

© 2021 Elsevier Ltd. All rights reserved.

The Journey Toward Improved Ramp Metering Application	10
Performance Measurement the Key to Understanding and Control	10
The Need for Illustrations of Fine-Grained Spatiotemporal Data	12
Understanding Very Small Time and Space Resolution of Data	12
Looking Deeper at the Vehicle Event Data	12
The Complex Problems Exhibited by Urban Motorways that Ramp Meters Need to Resolve	14
The Microdynamics at Play Differ Every Minute Everywhere	14
The Changing Shapes of the Fundamental Diagram	16
Moving on From Simplistic Numerate Constructs of Performance	17
How to Make Sense of the Ever-Increasing Amount and Granularity of Data (or How to Not Drown in it)	18
Acknowledgment	20
See Also	20
References	20
Further Reading	20

The Journey Toward Improved Ramp Metering Application

Performance Measurement the Key to Understanding and Control

Without the important pioneering experiences with ramp metering by Dolf May and many others and those associated with early ramp metering deployments in Melbourne, VicRoads would not have progressed to where ramp metering applications have advanced to today (refer to encyclopedia text on Urban Motorway Management).

In 2002, VicRoads realized it could “no longer afford to build as many new roads as in the past and still continue to meet peak localized demands or overall peak network demands.” At the same time, VicRoads had developed advanced motorway performance monitoring and assessment tools that showed VicRoads executive leadership, the potential gains to be achieved from efficient motorway operation. This only became obvious when the loss of productivity due to congestion occurring every day of the year on its urban motorways could be shown through synoptic charts showing the differences in the traffic volumes and speeds each day before, during and after flow breakdown and the widespread nature of the resulting congestion problem across Melbourne’s suburbs. Similarly, and as stated by Kurzhanskiy and Varaiya (2014):

“Empirical analysis of popular approaches such as High Occupancy Vehicle (HOV) and High Occupancy Toll (HOT) lane management show they are ineffective except on freeways that are inefficiently managed. New ITS technologies, such as “integrated corridor management” systems, appear to be founded on quicksand in the absence of a sufficient measurement system”.

VicRoads has found this statement to be insightful, and in many ways, it reflects the internal journey and experience where ideas and concepts often move to implementation without demonstrated evidence of sustainable beneficial outcomes.

In 2003, VicRoads developed a first-generation systematic and comprehensive motorway performance measurement and analysis tool for minute-by-minute analysis of route and system performance. These tools enable surface plots to be produced at 1-min and 500 m resolution for every day of the year, as well as one-minute time series of speed, flow and occupancy and fundamental relationships to be produced for every cross-section every day, and for each site to be integrated (e.g., minute-by-minute-traces) for the analysis of the chain of events in short time periods, that is 15 min or 1-hour periods (refer to Section 1.2 - Figs. 1 and 2 as examples).

In 2004–2005, building upon VicRoads previous ramp metering endeavors and the findings from the performance reporting, new understandings emerged of the enormous system losses (opportunity cost) of allowing motorway flow to collapse. The starkness of these findings brought VicRoads back to the drawing board to rethink through the numerous operational problem(s) of motorway networks from first principles with a view to improving system-wide efficiency and safety performance. Several years were subsequently taken out to analyse and rethink the complex problems motorways presented, and which included consultation with world leading academics in the traffic engineering field, traffic physicists and control engineering specialists. The key learnings

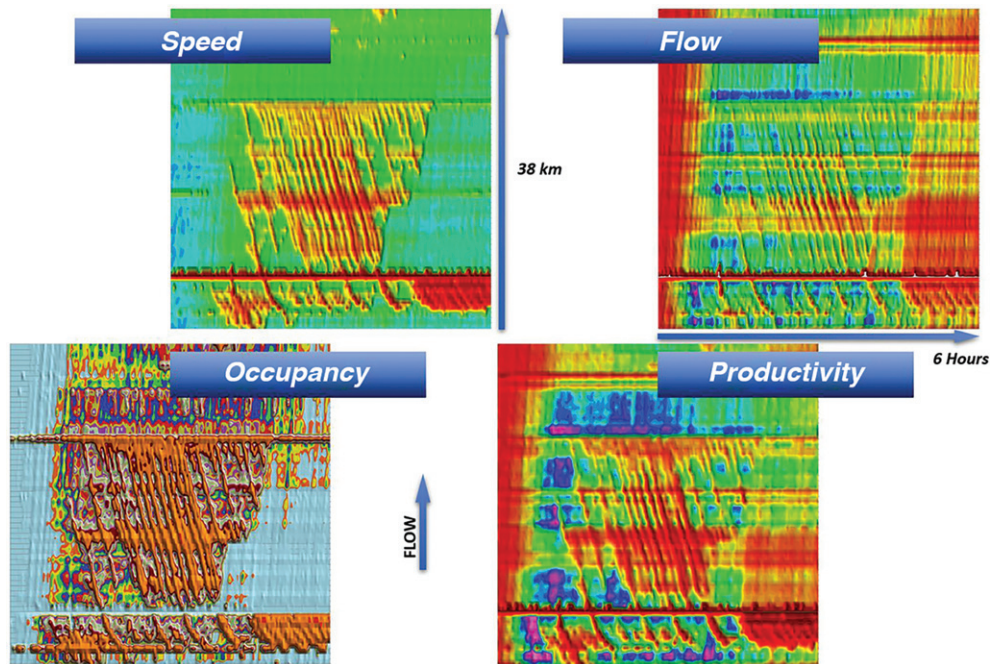


Figure 1 Example of electronically produced surface contours for Speed, Flow, Occupancy (SVO, occupancy as a surrogate measure for density), and productivity.

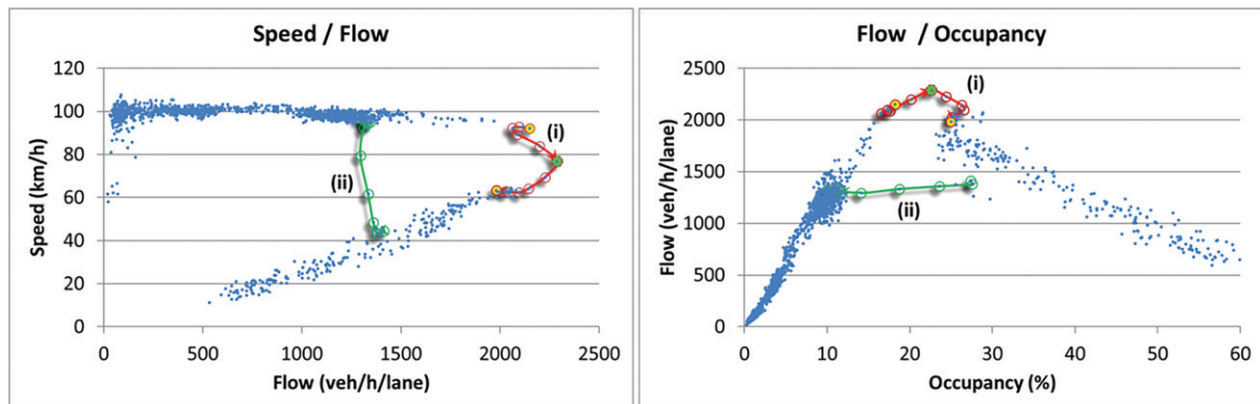


Figure 2 Speed/flow graphs for the same location showing the sequential path of (1) flow breakdown and (2) flow recovery (red line shows # consecutive 1-min intervals).

during this period which eventually lead to a new generation of motorway and ramp design and advanced control for VicRoads included:

- Recognizing the chaotic nature of traffic flows such as the randomness of traffic demands, (commercial) vehicle mix and traffic patterns (e.g., origin-destination patterns which influence lane changing rates—lane shear or carriageway friction) at the sub one-minute level;
- understanding the mechanisms involved in the formation of flow breakdown and flow recovery including understanding the various shockwave types and patterns, and their interactions, which included an understanding of both forward and backward moving waves;
- acknowledgement that traffic density along an entire route and not traffic flow was critical to understanding and improving motorway operation;
- understanding that ramp demand and ramp queue formation needed to be managed both locally and at a system level to avoid interference with the arterial road network;
- recognizing the need to develop advanced road network-wide control systems that can understand and resolve all-of-the-above complexity in real-time (e.g., 20 s intervals as a basis for measuring, assessing and responding to changing conditions);
- recognizing that accurate data was needed to drive the advanced control systems where the error in the real time data measurements had to be much lower than the control increments used by the algorithms; as a background, freeway data used by many road agencies was found to be imprecise around capacity flow values (e.g., vehicles in the act of lane changing were being double

counted, artificially increasing both carriageway flow counts and loop occupancy measurements) and the error in the flow counts was as high as 7%; and

- recognition that certain mainline and ramp geometry was inductive of motorway flow instability and thus the need to design motorways differently.

The Need for Illustrations of Fine-Grained Spatiotemporal Data

It is evident from early work in the 1960s, such as hand drawn contours of density, that new understanding occurs when we look at how motorways behave longitudinally, rather than just at a single point (detector) which represents less than 1% of the motorway length (a 2 m detected area at 500 m detector spacing) or a short motorway section like a merge area. When observed from above (i.e., taking a helicopter view), a motorway displays a lot of dynamic detail and complexity, such as vehicle interactions and varying behaviors in each motorway lane.

Since the 1960s we have witnessed large increases in the deployment of electronic measuring stations (commonly induction detector loops) and rapid developments in computer power and applications. It soon became possible to produce traffic density plots (synoptic charts) and similar plots based on other metrics (i.e., speed, flow, occupancy, and productivity) with greater ease, although not necessarily greater accuracy. These can be automatically produced each day and in some cases in real time. [Fig. 1](#) shows examples of surface contour plots produced for carriageway speed, flow, occupancy (density) and productivity. These figures illustrate their development over distance (space) and time on the Monash Freeway on one randomly selected day of the year.

Surface contour plots (e.g., a speed contour plot) of temporal data (measured over a peak period and along a corridor), where plots show metrics averaged over all lanes of a carriageway or lane based, have the potential to easily illustrate complex problems to practitioners such as traffic researchers, engineers and operators. These graphs can be analysed off-line or presented live in Traffic Management Centres (TMC) to illustrate the current and emerging traffic conditions along a route. VicRoads uses fine detailed SVO data (fine detail here is a reference to 1-min data on the carriageway level and high resolution 500 m detector spacing). This is because coarse data with low resolution hides (masks) the details of what happened, or if live what is happening now. When undertaking planning or design activities, different levels of resolution are required to determine suitable design values (refer to [VicRoads, 2019a, 2019b](#)).

It has been demonstrated that the higher the spatial and temporal resolution of the available traffic data, the more insights into the causes and occurrences of traffic flow problems the analyst can potentially get (although some level of appropriate aggregation, smoothing in presentation etc. is usually required).

There may be more than one approach to identifying cause and consequence that may indicate a potential operational problem at a motorway location, or within a corridor or network. Some potential diagnosis tools that may assist in identifying bottleneck location and in further understanding causation are listed below:

- “Surface Contour Plots” (SVO and productivity which is the mathematical product of speed and volume) over several days;
- Origin-Destination relations at intersections and along a corridor (averaged over a certain period such as an hour and temporal development); and
- Time series of speed and flow (various interval length(s)) for individual bottleneck sites (occurrence of breakdown and recovery paths) as shown in [Fig. 2](#) (hysteresis plots).

Understanding Very Small Time and Space Resolution of Data

Further examination of finer resolution data revealed that even more detail is evident in the traffic data, in particular if analysed on an individual lane level. For example, the 20 s data in [Fig. 3](#) for each of the four lanes in the top four graphs (red, blue, yellow, and green) shows they all have different lane occupancy values (where lane occupancy is used as a surrogate for density in traffic control) including many large spikes. When all lanes are displayed together (bottom left) it is possible to see that the spikes are not always simultaneously occurring in all the lanes. The black graph (bottom right) shows the result of averaging to a carriageway value—where much of the detail and complexity is lost and in some cases the magnitudes of the disturbances in the individual lanes are masked out.

The “spikes” in fine granularity data have often been interpreted as being “too noisy;” however, current investigations into such elements are indicating that the vehicle interactions that cause such spikes are likely to be a significant contributor to instability in motorway traffic flows and flow breakdown events, many of which are caused by lane changing events or aggressive drivers driving too close and suddenly braking or decelerating. Investigating the detailed information behind these spikes has helped to deepen the understanding of the inherent complexity of motorway traffic flow. Therefore, the new metering control algorithms needed to be based on understanding and working with this “noise” in order to understand and manage the real-time complexities within the motorway traffic flow.

Looking Deeper at the Vehicle Event Data

When data is observed at even finer levels, that is the vehicle event level (refer to [Figs. 4 and 5](#)) it is possible to see the details that contribute to many of the disturbances in the traffic flow including vehicles moving between lanes and the speeds of each individual vehicle. It is also possible to see how drivers place themselves within lanes with a wide scatter of wheel paths.

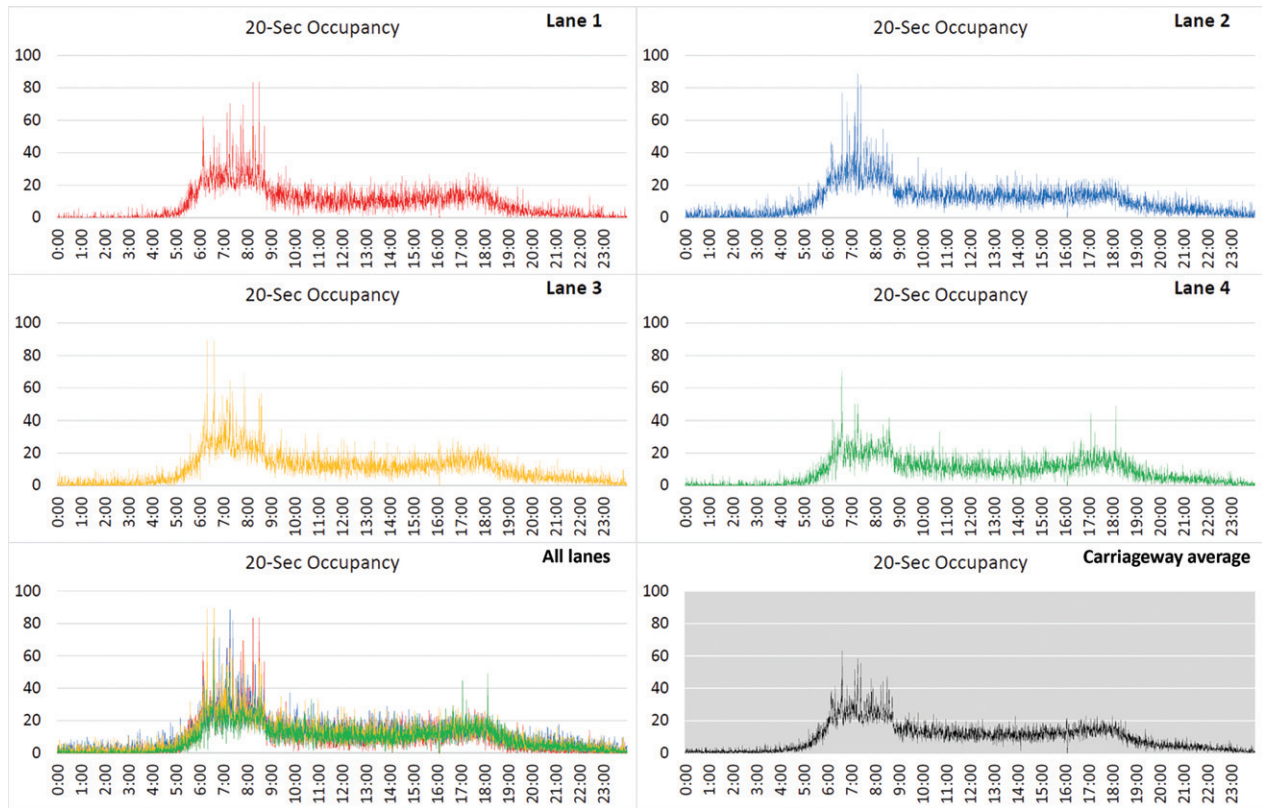


Figure 3 Lane and carriageway occupancy data at very fine resolution, i.e. 20 s data.

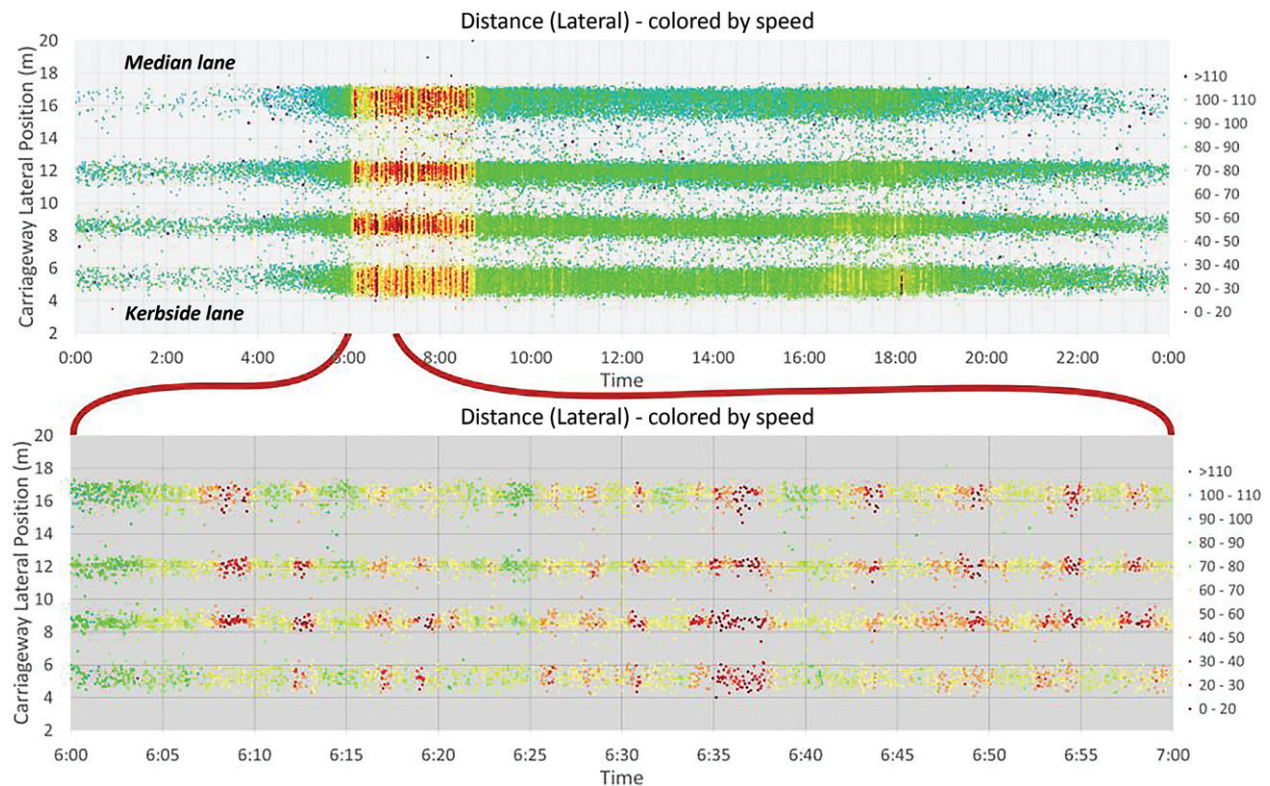


Figure 4 Lane and carriageway data at finer resolutions, i.e. vehicle event level over 24 h and 2 h.

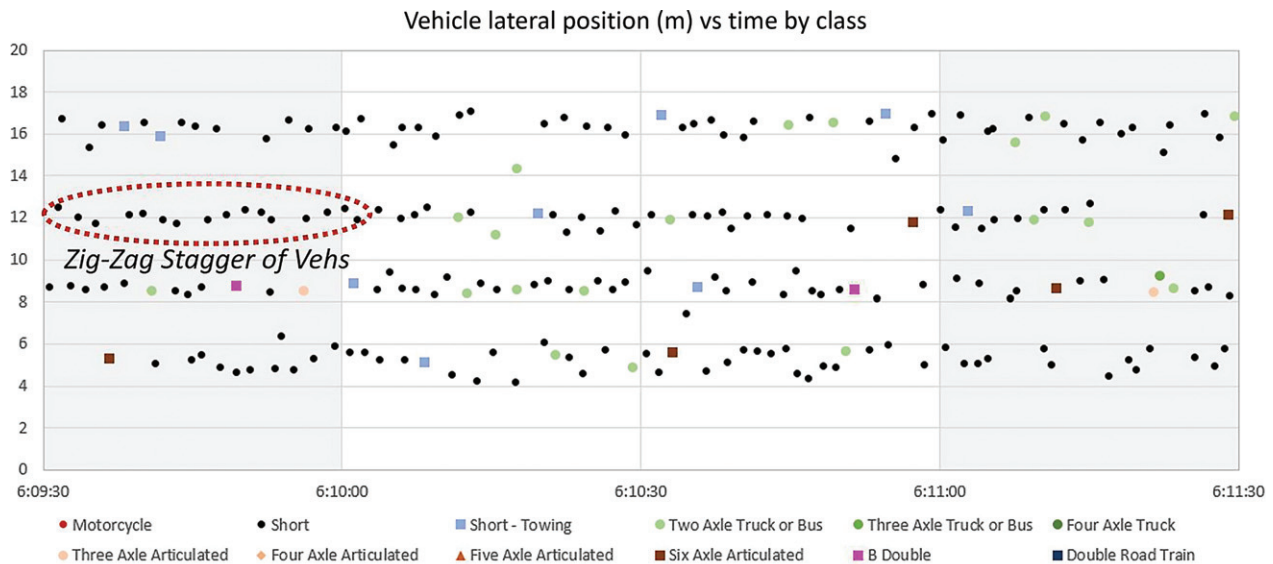


Figure 5 Lane and carriageway data at finer resolutions, i.e. vehicle event level showing vehicle lane position within lanes (including classification of vehicles and the Zig Zag effect).

In Fig. 4 the spread of vehicles in the Median Lane (top row of data points) tends to create a wide lateral gap to the adjacent lane. This may be a function of the emergency stopping lane and horizontal geometry at this location which allows drivers to perceive it is acceptable or safer to track closer to the unoccupied emergency stopping lane. Note that where a physical shoulder is quite narrow (i.e., less than 0.5 m) there will be a shy-line effect and vehicles will on average track differently.

For example, Fig. 5 shows the lateral positions of vehicles in more detail and staggering of vehicles within lanes is observed. This has been observed and described as the “zig-zag” effect which possibly occurs as a result of motorists wanting to see further ahead during dense traffic conditions or potentially as drivers prepare to lane change, thereby positioning their vehicle closer to a lane line.

When breaking down the 1 min data further (refer to Table 1), there are 122 vehicles of different lengths, traveling at different speeds, with different destinations, etc. It is new understanding at this level of data that provides further keys to understanding and managing the complexities which leads towards how ramp meters be more effectively used to optimize motorways. In this context, both the macro- and microproblems can be understood and described stochastically (e.g., with statistical distributions) and hence the real-time-variations can be measured and controlled to provide stable flows. This is discussed in the following section.

The Complex Problems Exhibited by Urban Motorways that Ramp Meters Need to Resolve

The Microdynamics at Play Differ Every Minute Everywhere

There are many factors that can influence how efficiently motorways operate. Some of the factors are static such as the number and width of lanes, grades (hills, sags, and crests), curves, ramp locations and separations, ramp geometry and the presence (or absence) of shoulders.

Other factors are dynamic such as traffic demand (varying over short time intervals, e.g. <1 min time resolution, and from day to day), changing trip patterns (origin-destination of vehicles resulting in fluctuations in trip length on the motorway), mix of vehicles (heavy vehicle proportions, presence of motorcycles), control actions (ramp metering, speed changes, lane closures), changing environment (time of day, wet and dry conditions, variable visibility, position of sun and shadows), driver impairment or distraction and incidents and events (roadworks, stationary vehicles or objects in the road environs).

Fig. 6 shows an example of some of the multiple factors that are constantly changing and interacting on the motorway, just as different rotation patterns and pitch can impact the trajectory of a baseball.

Most, if not all, of these factors, are likely to be constantly interacting on motorways and all can contribute in some part to the operational outcomes at any point in time. Due to the interplay of these factors, **it can be very difficult, or almost impossible, to separate out the impact of individual factors.** As a result, it needs to be acknowledged that the flow outcomes of a motorway are not deterministic but are stochastic (i.e., outcomes cannot be predicted precisely).

Fig. 7 indicates how this is happening every minute of every day on every section of motorway (and at every arterial road and intersection) and in different combinations on different scales, just as a baseball is rotating in flight. For example, the OD pattern (macro) is generalized by the motorist and defines the start and end of their journey. However, there are opportunities as they make a journey to modify route choice (OD Pattern micro) including where they enter or leave the motorway based on what they observe in

Table 1 Indicative breakdown of 122 vehicles per minute passing a single point on a 4-lane motorway

No.	Time	Dir/Lane	Speed (km/h)	Class Name	Axle Count	Length (m)	Headway (s)
1	6:10:00	W3	63.16	Short	2	5.243	1.757
2	6:10:00	W4	59.17	Short	2	4.981	0.981
3	6:10:00	W1	65.77	Short	2	4.943	1.579
4	6:10:01	W2	53.78	Short - Towing	4	9.881	5.072
5	6:10:01	W3	64.16	Short	2	4.742	1.199
6	6:10:01	W4	58.05	Short	2	4.236	1.497
7	6:10:02	W1	67.66	Short	2	4.279	1.515
8	6:10:03	W2	57.04	Short	2	5.004	2.381
9	6:10:03	W3	59.48	Short	2	5.262	2.165
10	6:10:03	W1	70.16	Short	2	4.459	1.424
11	6:10:04	W2	57.18	Short	2	4.282	1.218
12	6:10:05	W4	58.73	Short	2	4.698	3.314
13	6:10:05	W3	56.73	Short	2	4.356	2.293
14	6:10:05	W4	60.46	Short	2	5.012	0.818
15	6:10:05	W2	57.72	Short	2	5.223	1.244
16	6:10:06	W1	69.66	Short	2	5.346	2.487
17	6:10:07	W3	59.39	Short	2	4.479	1.362
18	6:10:07	W2	58.26	Short	2	5.357	1.361
19	6:10:07	W4	67.03	Short	2	5.352	1.723
20	6:10:08	W1	66.98	Short - Towing	3	9.448	2.005
21	6:10:08	W3	60.69	Short	2	4.99	1.187
22	6:10:09	W2	59.38	Short	2	5.236	1.925
23	6:10:09	W4	64.4	Short	2	5.238	1.697
24	6:10:10	W2	59.83	Short	2	4.163	1.204
25	6:10:11	W1	62.52	Short	2	4.678	2.8
26	6:10:11	W3	60	Two Axle Truck or Bus	2	7.396	3.113
27	6:10:11	W4	66.81	Short	2	4.936	2.401
28	6:10:12	W2	57.59	Two Axle Truck or Bus	2	11.309	1.904
29	6:10:12	W4	70.97	Short	2	5.028	1.086
30	6:10:12	W3	62.47	Short	2	5.233	1.328
31	6:10:13	W1	68.57	Short	2	3.777	2.47
32	6:10:13	W2	58.96	Short	2	4.577	1.503
33	6:10:15	W1	67.03	Short	2	4.409	1.62
34	6:10:15	W3	62.53	Two Axle Truck or Bus	2	8.296	2.493
35	6:10:15	W2	58.13	Short	2	5.225	1.616
36	6:10:17	W1	68.28	Short	2	4.466	2.207
37	6:10:17	W4	64.51	Two Axle Truck or Bus	2	7.897	4.574
38	6:10:17	W2	59.55	Two Axle Truck or Bus	2	5.808	1.985
39	6:10:19	W2	62.25	Short	2	4.824	1.767
40	6:10:19	W3	62.58	Short - Towing	4	12.122	4.188
41	6:10:20	W1	64.47	Short	2	5.066	2.948
42	6:10:20	W4	71.42	Short	2	4.905	2.861
43	6:10:20	W2	63.07	Short	2	5.223	1.362
44	6:10:21	W3	64.9	Short	2	4.422	1.533
45	6:10:21	W1	61.11	Two Axle Truck or Bus	2	10.48	1.07
46	6:10:22	W4	67.01	Short	2	5.434	1.982
47	6:10:22	W3	65.34	Short	2	4.554	1.43
48	6:10:22	W2	66.21	Short	2	5.234	2.091
49	6:10:22	W1	61.48	Short	2	4.426	1.393
50	6:10:24	W3	67.47	Short	2	5.157	1.483
51	6:10:24	W1	62.5	Short	2	4.441	1.36
52	6:10:24	W2	67.11	Two Axle Truck or Bus	2	5.596	1.498
53	6:10:24	W4	65.9	Short	2	5.235	2.189
54	6:10:25	W2	64.07	Short	2	4.471	1.409
55	6:10:25	W3	67.73	Short	2	5.108	1.848
56	6:10:26	W1	60.33	Short	2	5.082	2.323
57	6:10:26	W4	61.31	Short	2	4.495	2.12
58	6:10:27	W2	63.57	Short	2	5.218	1.495
59	6:10:27	W3	63.48	Short	2	4.576	1.516
60	6:10:28	W2	64.97	Short	2	4.878	1.246
61	6:10:28	W4	64.04	Short	2	5.207	1.846
62	6:10:29	W1	59.19	Two Axle Truck or Bus	2	7.324	2.667
63	6:10:29	W3	63.63	Short	2	5.126	2.155
64	6:10:30	W1	58.88	Short	2	4.807	1.421
65	6:10:30	W2	65.83	Short	2	5.099	2.288
66	6:10:30	W3	63.47	Short	2	4.352	1.427
67	6:10:31	W1	61.58	Short	2	5.113	1.215
68	6:10:31	W4	70.16	Short - Towing	4	11.888	3.556
69	6:10:32	W3	67.43	Two Axle Truck or Bus	2	5.714	1.887
70	6:10:33	W1	66.16	Six Axle Articulated	6	19.725	1.349
71	6:10:34	W4	70.45	Short	2	4.717	2.106
72	6:10:34	W2	75.83	Short	2	4.747	3.848
73	6:10:35	W3	68.21	Short	2	4.628	2.3
74	6:10:35	W4	68.95	Short	2	4.738	1.188
75	6:10:35	W2	62.69	Short - Towing	4	9.604	1.131
76	6:10:36	W1	66.03	Short	2	4.626	3.383
77	6:10:36	W3	70.86	Short	2	5.254	1.511
78	6:10:37	W4	71.81	Short	2	4.593	1.783
79	6:10:37	W2	62.82	Short	2	4.463	1.739
80	6:10:38	W3	71.06	Short	2	4.482	1.504
81	6:10:38	W4	69.9	Short	2	5.38	1.232
82	6:10:38	W2	59.69	Short	2	4.278	1.139
83	6:10:38	W1	63.23	Short	2	4.346	2.21
84	6:10:38	W3	69.8	Short	2	4.743	0.784
85	6:10:40	W4	75.43	Short	2	5.082	1.931
86	6:10:40	W1	65.24	Short	2	4.484	1.623
87	6:10:40	W3	73.24	Short	2	5.091	1.649
88	6:10:41	W2	63.72	Short	2	4.656	2.608
89	6:10:41	W4	73.17	Short	2	4.35	1.002
90	6:10:41	W1	64.63	Short	2	4.243	1.313
91	6:10:42	W3	67.12	Short	2	5.258	2.02
92	6:10:43	W1	62.39	Short	2	4.714	1.427
93	6:10:44	W2	61.37	Short	2	4.725	3.141
94	6:10:44	W1	53.98	Short	2	4.524	1.467
95	6:10:44	W3	66.55	Short	2	5.245	2.035
96	6:10:44	W4	75.32	Two Axle Truck or Bus	2	5.455	3.535
97	6:10:45	W2	57.06	Short	2	4.518	1.311
98	6:10:45	W1	56.08	Short	2	5.105	1.102
99	6:10:46	W3	69.71	Short	2	4.531	1.444
100	6:10:46	W1	57.74	Short	2	4.424	1.19
101	6:10:46	W4	72.77	Short	2	5.236	2.094
102	6:10:47	W2	57.74	Short	2	4.704	1.576
103	6:10:48	W2	57.28	Short	2	5.248	0.9
104	6:10:48	W1	59.19	Short	2	5.195	1.432
105	6:10:49	W4	73.91	Two Axle Truck or Bus	2	6.252	2.35
106	6:10:49	W1	61.68	Short	2	4.47	1.252
107	6:10:49	W2	62.24	Short	2	4.423	1.865
108	6:10:50	W1	62.95	Two Axle Truck or Bus	2	7.108	1.02
109	6:10:51	W3	92.27	Short	2	4.731	4.953
110	6:10:51	W2	63.92	8 Double	7	18.099	1.485
111	6:10:52	W4	75.93	Short	2	5.023	3.635
112	6:10:52	W1	65.33	Short	2	4.256	2.409
113	6:10:53	W2	69.77	Short	2	5.237	2.3
114	6:10:54	W4	71.28	Short - Towing	3	8.971	1.639
115	6:10:54	W1	63.92	Short	2	4.243	1.815
116	6:10:55	W4	82.11	Short	2	4.271	1.001
117	6:10:57	W3	59.6	Six Axle Articulated	6	18.841	6.136
118	6:10:57	W4	74.7	Short	2	4.425	1.91
119	6:10:58	W2	69.15	Short	2	5.254	4.54
120	6:10:58	W1	70.45	Short	2	4.205	3.627
121	6:10:59	W4	70.14	Short	2	5.252	1.794
122	6:10:59	W3	63.13	Short	2	4.415	2.787

Data source TIRTL data logger used by VicRoads for real time operations and control.

the traffic ahead, what information they see on the VMS signs, what they hear on their radio or a result of a decision they make on-route to modify the journey for another purpose, for example, fill up with petrol or go shopping, etc. Hence, in motorway control, short-term predictions are difficult and prone to error as there are no certainties.

If road operators are to have the ability to sustain stable traffic flows, motorway control systems need increased sophistication and ability to recognize and control (<1 min) the differing conditions both on individual sections (sub-links) and across whole systems. For example, VicRoads currently controls at the 20 s level with a system designed to have minimal latency between collecting field data to central control (city wide) and returning control decisions back to field (within seconds) (VicRoads, 2017).

This is demonstrated in Fig. 8 which shows 10 consecutive weekdays on a 45 km length of motorway over the morning peak period (6 h) and reveals quite different congestion patterns every day. Common patterns are visible, but to a trained eye many

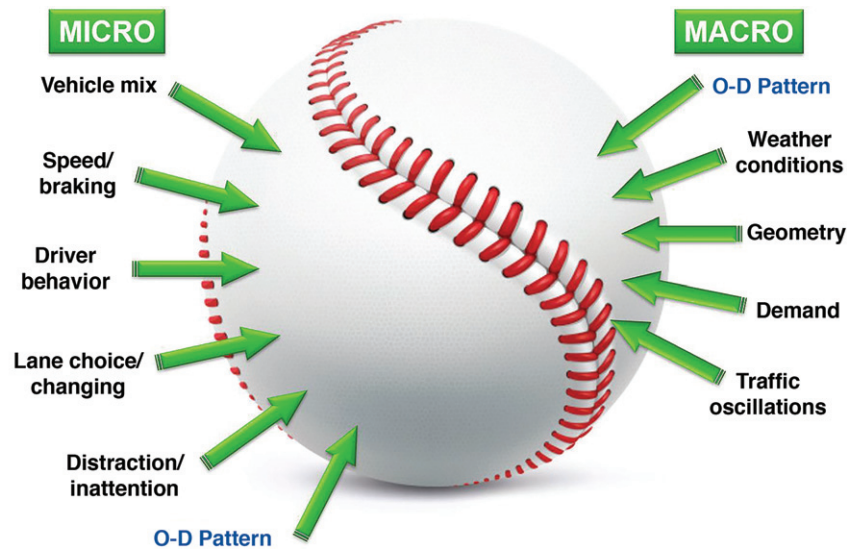


Figure 6 Multiple factors constantly interacting on the motorway.

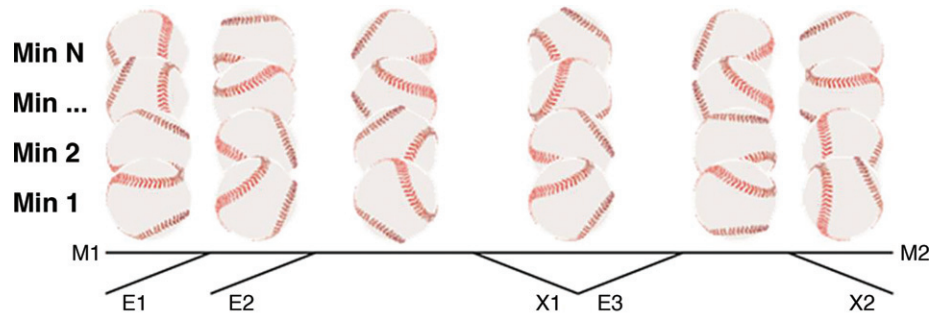


Figure 7 Multiple rotations interacting the motorway performance.

differences are apparent. Due to these differences the fundamental diagrams would vary significantly for each of the ten consecutive days. Also, it is these differences that need to be seen in the live real-time data and acted on by advanced control systems.

The Changing Shapes of the Fundamental Diagram

Measures of flow, speed and density (occupancy) from a motorway always shift from time period to time period (e.g., minute to minute) and no two 1-hour periods or days are the same. Observation indicates that the sequence of data points returned from the field are never on the same curve, or trajectory (from minute to minute), as was previously recorded (refer to Fig. 2). The resulting shapes of the fundamental relationships will also vary from section to section on a motorway and even adjacent sections 500 m apart can show markedly different generalized forms despite the fact that from a road design and visual perspective the sections look homogeneous to the road designer and road operator. *The differences from time-period to time-period have often been ignored or smoothed by averaging, yet they have major consequences for real time traffic control and optimization.*

Whilst the fundamental diagrams are influenced by road geometry, they are also influenced by the dynamic traffic loading of the motorway. Traffic engineers have undertaken traffic studies for many decades to determine traffic patterns such as Origin-Destinations (OD) studies; however, these studies are usually only useful for determining the general macro-patterns. The micro-patterns discussed above (e.g., individual manoeuvres of any of the 122 vehicles/min shown in Table 1) greatly influence the shape of the fundamental diagram.

For example, the number of lane changing manoeuvres resulting from a particular OD pattern at the 1-min time interval level may be three or more times, or even only one-third the number of lane changes that is inferred when the hourly number is divided by 60. This lane changing friction ("lane shear") which is also a microdynamic can influence the "traffic state" greatly and thus the shape of the fundamental relationships. In traffic operations care is therefore needed when using values from any fundamental relationship diagram.

The curves and their formulae often presented in text books and which are generally used in transport models are simple lines of best fit used to represent a much broader point cloud that comes from measured data, particularly when a large sample is used (i.e.,

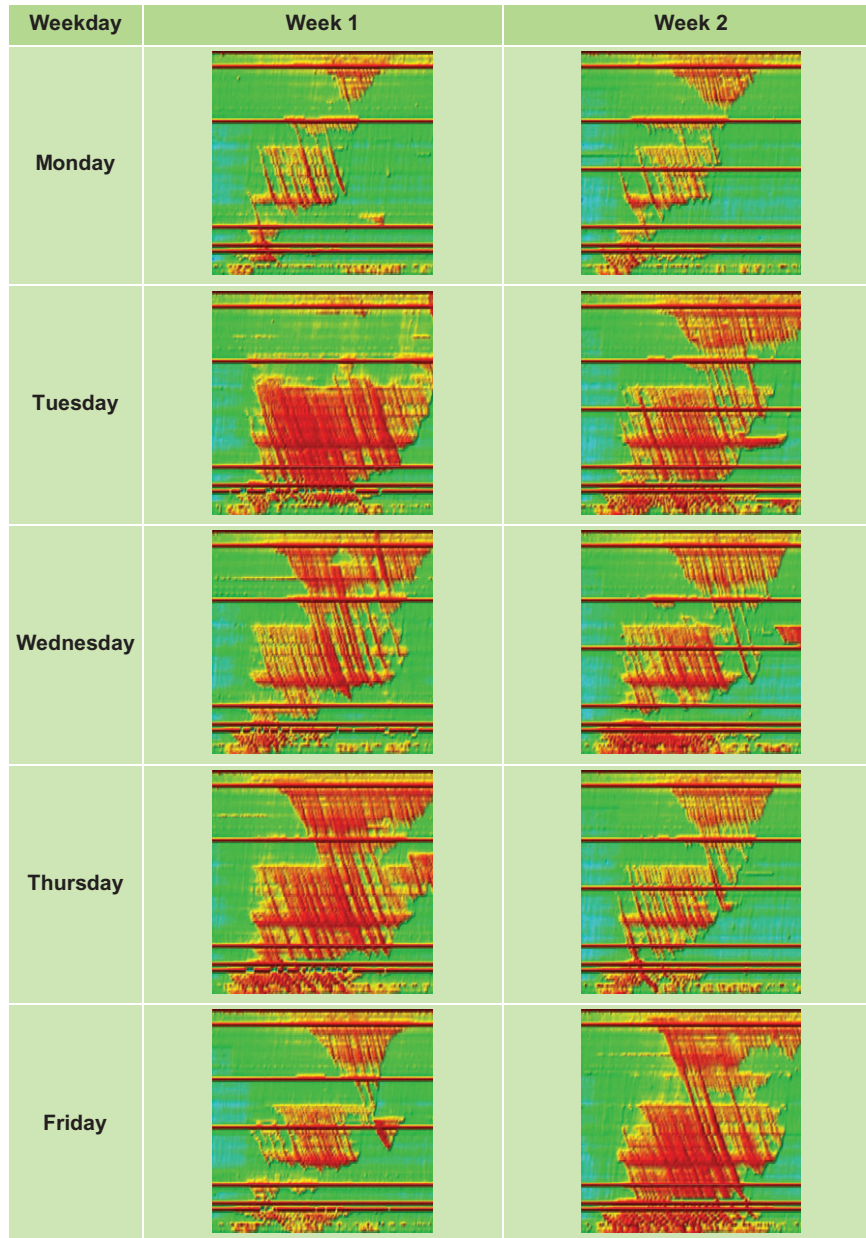


Figure 8 Speed surface contour plots for 10 consecutive weekdays showing different traffic patterns each day (direction of travel up the page).

30 days or greater). Therefore, the presentation of a curve in the absence of the point cloud masks the variability that is experienced in the real world by the motorist (refer to red fitted curve in [Fig. 9](#)).

[Fig. 9](#) utilizes 1-h aggregate speed and flow data measured over a 1-month period, revealing a wide range of values which represent the road user experience. The red curve shows a fitted line through the range of measured values. It should be noted that the 31 days of data used to produce this figure were analysed specifically to determine capacity curves in accordance with documented analysis processes associated with traditional capacity determination and has limited application beyond this use.

Moving on from Simplistic Numerate Constructs of Performance

The HCM ([Transportation Research Board, 2016](#)) uses numerate concepts such as Level of Service (LOS) to define the various operating characteristics for uninterrupted flow. Road user experience for each level of service and associated density ranges have been in use for many years to inform planning of uninterrupted flow facilities.

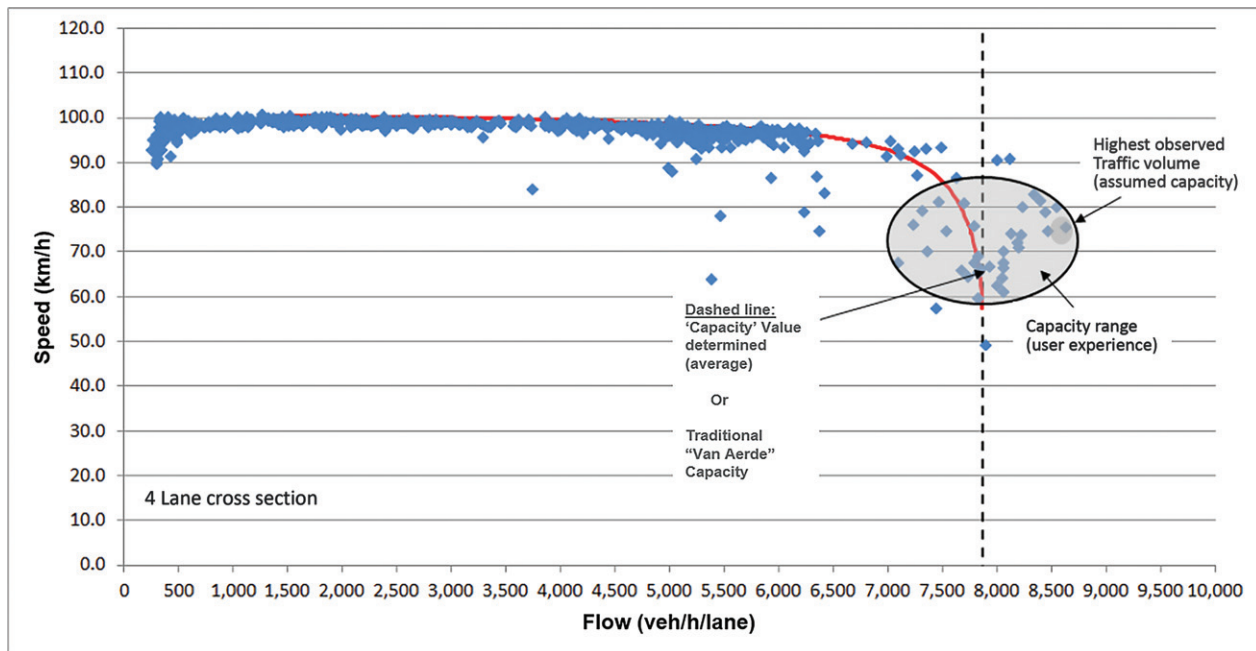


Figure 9 Typical graph of speed vs flow at a single location with a 4 Lane Cross Section (actual 1-h speed and flow data measured over 1 month). Dashed line indicates Traditional “Van Aerde” capacity.

In addition, researchers and practitioners have often used rather simple constructs such as volume-to-capacity ratios to assign single flow values to some sort of operational performance level or congestion target. It has been argued by [Lay \(2011\)](#) that such concepts are useless: “HCM LOS and volume-to-capacity ratios . . . falling into this numerate but meaningless category”. VicRoads no longer considers volume-to-capacity ratios or Level of Service (LOS) values to be useful for describing uninterrupted flow performance. Amongst practitioners there may be a misconception that LOS levels operate on a continuum (i.e., linear), however, transitions are not always smooth as the traffic condition can pass through various traffic states very quickly (sometimes in a matter of seconds, refer to [Fig. 10](#)). When it comes to density ranges or volume-to-capacity ratios, the corresponding user experience varies in the motorway domain as it depends on the number of motorway lanes and the micro- and macrofactors shown in [Fig. 6](#).

Congestion as a concept has been hard to describe and sometimes has focused solely on economically determined optimum flows uncoupled from the road user experience or any road operator’s measure of congestion. Arriving at a commonly agreed and useful definition of congestion has so far proved elusive and new metrics are needed to understand and optimize motorways. An attempt of targeting an operational optimum “Productivity” (speed $\times$ flow) is described in [VicRoads \(2019b\)](#).

How to Make Sense of the Ever-Increasing Amount and Granularity of Data (or How to Not Drown in it)

It has been stated above that traffic flow is chaotic or at least highly random because of the constant changes of influencing factors shown in [Fig. 6](#). While some of these factors follow (more or less) regular daily, weekly, or annual patterns (e.g., overall traffic demand), there are random short-term fluctuations that do not allow for the precise prediction of the overall outcome in terms of traffic flow quality.

Another way to put it is that these factors and hence the overall traffic flow outcome are (to the largest extent) not deterministic but stochastic. This means that each of them can be described by a statistical distribution or other statistical parameters and the occurrence of a particular value of such a parameter (e.g., traffic demand) has a certain likelihood. Therefore, statistical methodologies are needed, for example, to describe the entirety of the 8760 different hourly traffic demands that can be measured throughout a year.

Similarly, on the capacity side of things, the journey towards statistically describing the ever changing dynamic capacity as a result of the interactions between all the factors shown in [Fig. 6](#) that are in themselves random has not only begun but is progressively finding its way into Highway Capacity Manuals around the world (refer to [TRB, 2016](#), [Henkens and Heikoop, 2015](#); [VicRoads, 2019b](#)). Rather than on simplistic fundamental diagrams highlighting a single ‘Capacity’ value (refer to [Fig. 9](#)), this is now increasingly based on curves describing the flow breakdown probability as a function of traffic flow as shown on a per lane basis

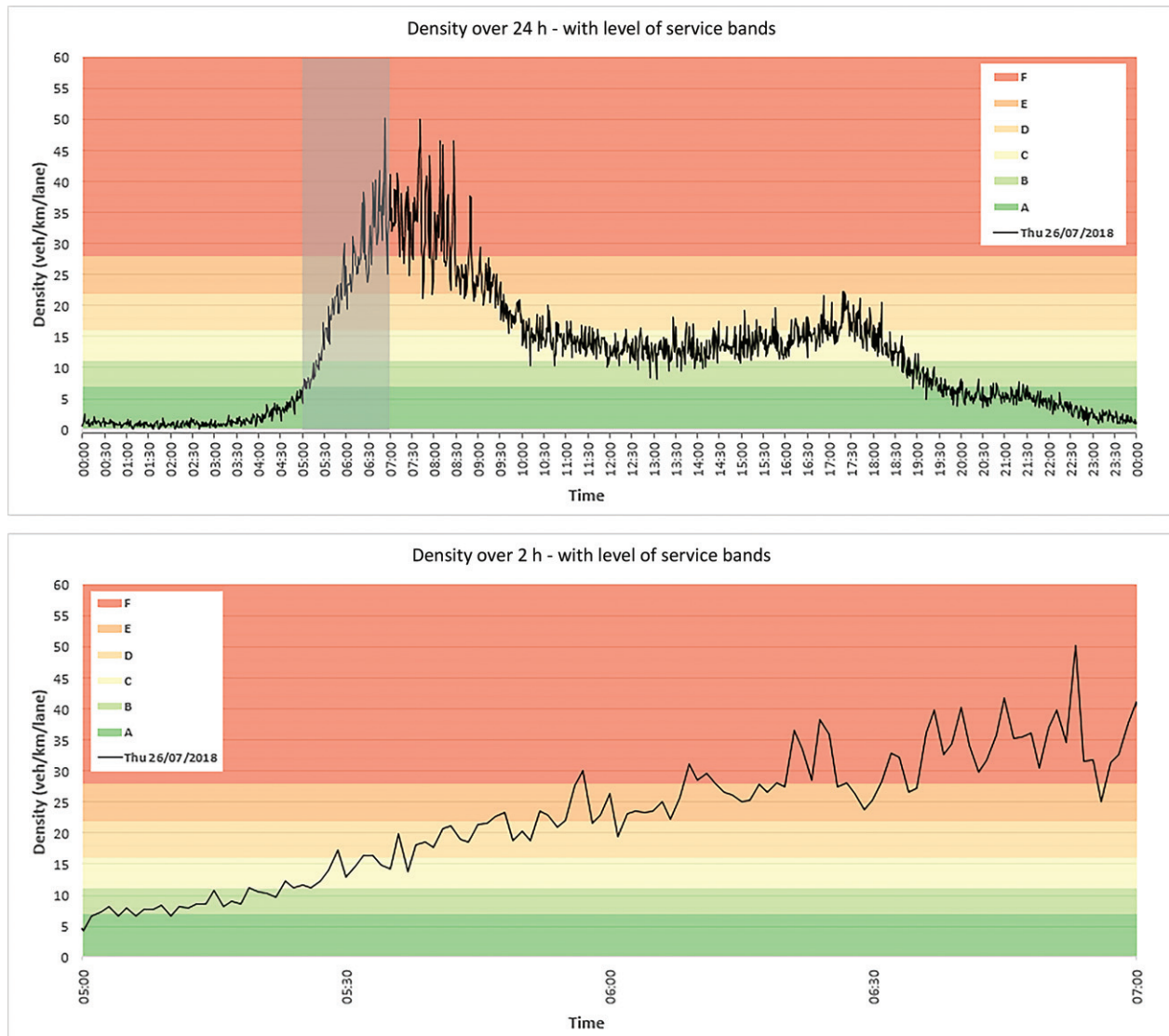


Figure 10 Variability of density with respect to traditional Levels of Service (LOS) (1-min resolution).

in **Fig. 11**. To reflect the changed thinking, VicRoads has chosen to widely replace the term “Capacity” by “Maximum Sustainable Flow Rate” (MSFR). The figure also shows the increase in productivity (speed $\times$ flow) with growing traffic flow up to a certain point before it decreases at around 90% of the traditional “Capacity.” As an additional information, **Fig. 11** illustrates the systematically decreasing Maximum Sustainable Flow Rates (MSFR) with increasing number of carriageway lanes which is particularly due to the exponential growth in lane changing activity (i.e., friction/lane shear) needed to fill all the lanes including the inner ones to “capacity.”

These examples that can be based on the traditional SVO data are important for motorway planning, design, and operations. Regarding operations, statistical methods have been applied that appropriately describe the entirety of different operational states over a long enough period and that recommend a certain state based on this comprehensive analysis.

With the finer individual vehicle data such methodologies are still to be developed. However, some practical consequences can already be seen such as the recommendation included in (VicRoads, 2019a, 2019b) that the number of carriageway lanes should potentially be limited to no more than 5 because of the very high number of measured lane changes that increases exponentially with increasing lane number.

In the context of real-time operations that this chapter is dealing with, a comprehensive ramp metering system must be able to quickly respond to critical changes (e.g., a too high density leading to high flow breakdown and crash risks) so that density at a bottleneck can be immediately reduced (refer to the chapter on *City Wide Coordinated Ramp Metering* in this Encyclopedia).

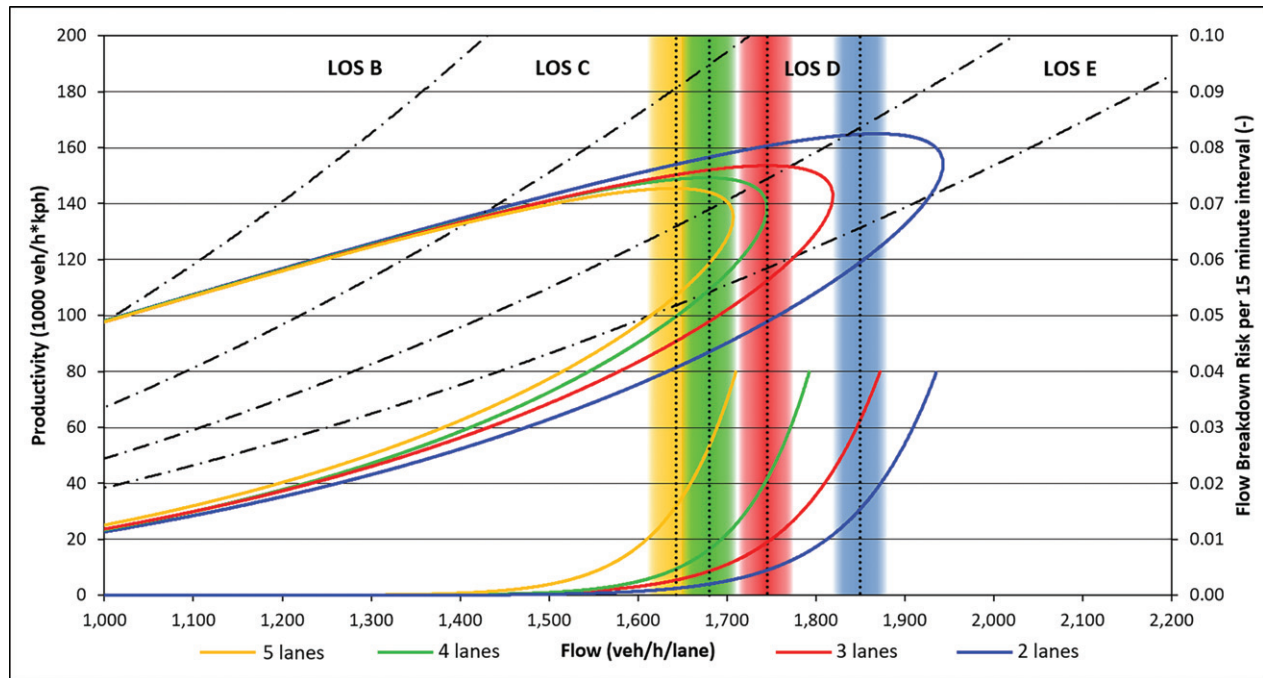


Figure 11 Breakdown probability and productivity plotted against per-lane flow rate for various cross-section types (i.e. number of carriageway lanes).

Acknowledgment

VicRoads acknowledges the partnership and contributions of TRANSMAX as the owner and developer of the VicRoads Motorway Management system (STREAMS) and Prof. Markos Papageorgiou and Prof. Ioannis Papamichail from the Technical University of Crete in the development of the HERO Live coordinated ramp metering system used to manage Melbourne's motorways.

See Also

Urban Motorway Management; Keeping Urban Motorways Safe and Efficient with City Wide Coordinated Ramp Meters
City Wide Coordinated Ramp Meters; Keeping Urban Motorways Safe and Efficient with City Wide Coordinated Ramp Meters

References

- Henkens, N., Heikoop, H., 2015. CIA (Capaciteitswaarden Infrastructuur Autosnelwegen), Handboek CIA versie 4, Rijkswaterstaat WVL, Rijswijk (ZH).
- Kurzanskiy, A., Varaiya, P., 2014. Traffic management: an outlook. *Econ. Transport.* 4 (3), 135–146.
- Lay, M.G., 2011. Measuring traffic congestion. *Road Transport Res* 20 (1), Australian Road Research Board (ARRB), Sydney.
- Transportation Research Board, 2016. Highway Capacity Manual, 6th ed. Transportation Research Board (TRB), Washington, D.C.
- VicRoads, 2017. Managed Motorway Framework. VicRoads, Melbourne.
- VicRoads, 2019a. VicRoads MMDG Volume 2, Part 3. Motorway Planning and Design". Department of Transport (VicRoads), Melbourne.
- VicRoads, 2019b. VicRoads MMDG Volume 1, Part 3. Motorway Capacity Guide. Department of Transport (VicRoads), Melbourne.

Further Reading

- Treiber, M., Kesting, A., 2013. *Traffic Flow Dynamics - Data, Models and Simulation*. Springer, Berlin.
- Kerner, B., 2017. *Breakdown in Traffic Networks - Fundamentals of Transportation Science*. Springer, Berlin.
- Papageorgiou, M., Hadj-Salem, H., Middelham, F., 1991. ALINEA: A Local Feedback Control Law for Onramp Metering, Transportation Research Record No 1320, Transportation Research Board (TRB), Washington, D.C., pp. 58–64.
- Papageorgiou, M., Hadj-Salem, M., Middelham, F., 1997. "ALINEA Local Ramp Metering Summary of Field Results", Transportation Research Record No 1603, Transportation Research Board (TRB), Washington, D.C., pp. 90–98.
- VicRoads, 2017. Managed Motorway Framework. Department of Transport (VicRoads), Melbourne.
- VicRoads, 2020a. VicRoads MMDG Volume 1. Part 2: -Traffic Theory Relating to Urban Motorways. Department of Transport (VicRoads), Melbourne.
- VicRoads, 2020b. VicRoads MMDG Volume 1, Part 5: - Linking Investment and Benefits Approach. Department of Transport (VicRoads), Melbourne.

City Wide Coordinated Ramp Meters - Keeping Urban Motorways Safe and Efficient with City Wide Coordinated Ramp Meters

John Gaffney, Hendrik Zurlinden, Department of Transport (DoT)—VicRoads, Kew, VIC, Australia

© 2021 Elsevier Ltd. All rights reserved.

Control Algorithms Used by VicRoads	21
HERO-LIVE Coordinated Ramp Metering Operation	22
HERO-LIVE Modules	22
Ramp Meter Operational Modes	24
Ramp Meters—Off (Default)	24
Times of Operation	24
Faults and Device Failures	25
Switching on/off Signs and Signals	25
Start-up Sequence	25
Close-down Sequence	25
Operating Sequence and Cycle Times (Not Used for Design)	26
Signal Timings	26
Minimum Cycle Time (Operations)	27
Maximum Cycle Time (Operations)	28
Mainline and Entry and Exit Ramp Design	28
Introduction	28
Performance-based Design	29
Design of Ramp Meters and Entry Ramps	30
Overview of the Design Process	30
Ramp Storage and Discharge Capacity for Design	31
Ramp Metering Physical and ITS Design Elements	31
Conclusion and Findings	32
Acknowledgments	32
See Also	32
References	32
Further Reading	32

Control Algorithms Used by VicRoads

The HERO-LIVE ramp metering control algorithms used by VicRoads on motorways in Victoria are based on the HERO/ALINEA suite of ramp metering control algorithms. The original ALINEA control philosophy which provides local control at an individual motorway entry ramp was developed by Markos Papageorgiou (refer to [Papageorgiou et al., 1991, 1997](#)).

HERO-LIVE today is an advanced suite of integrated control engineering tools which are robust in real-world (LIVE) operations where traffic conditions can be very complex. While some algorithms can perform well within offline traffic models, taking account of real-world problems is more complicated. Real-world operations need to consider and cope with the effects of rapid changes in traffic patterns (e.g., origin–destination patterns), such as those caused by incidents. There is also the need to handle real-time data errors, noisy data and data dropouts etc. which requires advanced levels of real-time statistical processing and evaluation for every 20-s time step.

HERO-LIVE incorporates a suite of ALINEA modules that were developed for coordination of ramp meters at a number of coordinated ramps along a length of motorway. In cooperation with Markos Papageorgiou and Associates, VicRoads has been involved in ongoing development of the algorithms with a significant number of enhancements to the algorithms and new modules, since the initial on-road trial in 2008. These enhancements specifically seek to solve unique problems encountered on complex urban motorways and have made significant advancements from a reactive coordinated ramp metering system towards a City Wide Coordinated Ramp Metering (CWRM) system that optimizes motorway networks within the urban road network (e.g., a network comprising motorways and arterial roads).

The HERO-LIVE suite, which is fully deployed in the field, has proven performance improvements based on utilization of the following features:

- Dynamic start up and shut down algorithms that ensure the system only operates when required;
- Consistency with contemporary traffic theory for optimizing motorway flow;
- The contemporary control logic is based on feedback from downstream conditions in real-time to dynamically adjust signal cycle times;
- Use of occupancy from the downstream motorway bottleneck locations as the optimizing measure;
- Transparent in operation with fully configurable parameters;
- Integrated operation of local ramp control within a coordinated system based on sound operating rationale;
- Incorporation of modules for adjustments to entry flow rates based on consideration of ramp queues, and arterial road queues in some cases, as well as ramp delays;
- Potential to manage flow at motorway-to-motorway interchanges by linking upstream ramps on separate motorways;
- Ability to manage bottlenecks many kilometers (3–4 km) downstream from the nearest ramp; and
- Potential for management of multiple bottlenecks to adaptively determine and target the critical bottleneck.

The adoption of an efficient control algorithm suite is of paramount importance for a successful ramp metering system. Designing and installing the necessary entry ramp layouts and field equipment is necessary and important, but these components are not sufficient in themselves for successful operations—effective control algorithms appropriately integrated are required. The management of any road system requires intentional intervention, particularly when the system is operating close to capacity. It should no longer be acceptable to road agencies to wait until the system has broken down (failed) before interventions occur. There are clearly definable metrics for which to design and operate motorway networks with a low chance of failure, and thus various components of the control system need to activate as certain target metrics are approached to limit flows moving towards bottleneck areas before the risk of failure rises beyond manageable limits.

HERO-LIVE Coordinated Ramp Metering Operation

The HERO-LIVE suite of algorithms is a comprehensive and complete concept (and corresponding implementation software) that applies coordinated ramp metering to motorway networks of arbitrary size, topology and characteristics. A central software installation at the coordinating road authority suffices for the coordinated ramp metering control of all equipped motorways, network-wide facilitated through entry/connector ramp signals that are directly controlled by the system and operated by ramp metering optimization specialists.

All specific data related to the network, the controllable entry-ramps and the real-time traffic conditions are communicated to the central software and made accessible via the STREAMS graphical user interface (GUI). Extensions of the motorway network under control or addition of new controlled ramps may be easily accommodated using the GUI. After its configuration, HERO-LIVE operates automatically without the need for operator interventions, except for specialist analysis, fine-tuning, checking for detector faults etc. Operator intervention is also possible for incident and emergency situations.

HERO-LIVE is fully traffic-responsive (with configurable update period that may be selected as short as 20 s) and adapts automatically to the prevailing traffic conditions aiming at maximizing stable motorway throughput; while, at the same time, monitoring the current situation at the on-ramps and limiting any incurred vehicle queues or waiting times. HERO-LIVE uses real-time measurements from multiple mainline locations on the motorway network and from the on-ramps and dynamically adjusts the ramp meter cycle times. Particular attention is paid to latent motorway bottlenecks, which are the typical triggers of flow breakdown and traffic congestion, and appropriate ramp metering actions are initiated, whenever necessary, to target the prevention of traffic breakdown and delivery of sustained near capacity flows based on the integrated robust feedback control methods.

The HERO-LIVE algorithms have been conceived based on the latest insights of contemporary traffic flow theory for optimizing motorway flow; moreover, they apply a variety of powerful and proven automatic control methods that guarantee stable, robust and efficient operation on the basis of feedback principles. This distinguishes HERO-LIVE from any approaches based on heuristic constructions which may lack the fundamental feedback control principles. In addition to the theoretically sound background, HERO-LIVE decisions are transparent in operation, which enables operators to be fully aware at any time about the reasons behind the employed control actions.

Motorway bottlenecks may be due to the merge of an entry ramp, but may also be a result of other geometric or traffic factors, such as merging at a lane drop, a steep upgrade, a tight-radius curve, a bridge, a tunnel, reduced lane widths or areas with high weaving or lane changing movements, for example, just upstream of a lane gain or high-flow exit ramp.

Traffic occupancy at mainline bottlenecks (detection locations in this control context) is the principal real-time measurement used to manage traffic flow on the motorway

HERO-LIVE Modules

The HERO-LIVE suite of algorithms manages the ramp traffic flow entering the motorway by monitoring and controlling the motorway flow at each ramp merge as well as other critical bottlenecks. Ramp discharge flow calculations are based on various modules providing outputs for isolated and coordinated operation as shown in [Fig. 1](#). It is noted that in order to protect intellectual

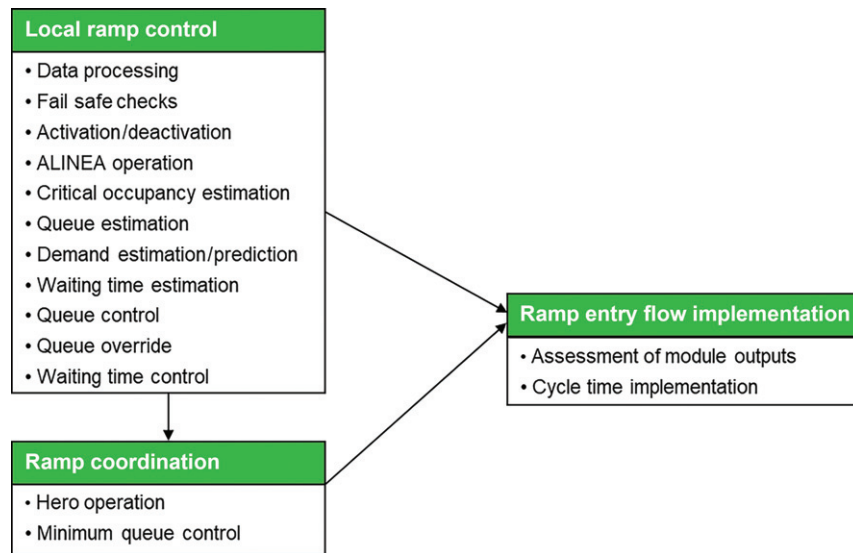


Figure 1 HERO-LIVE coordinated ramp control structure.

property (IP) the descriptions on the following pages are limited to what has already been published in the public domain for many years.

The algorithms use flow, speed, and occupancy data in various modules summarized below.

Activation/Deactivation

This module switches the motorway ramp meters on according to preset traffic flow, occupancy and speed thresholds at the mainline bottleneck locations downstream of the ramp. The ramp meters are switched off according to preset occupancy and speed thresholds. Activation and deactivation is independently controlled at each ramp within a coordinated system. The thresholds are set with the intention of turning on the signal control before the onset of flow breakdown and turning the metering signals off when traffic flow conditions indicate that flow breakdown is unlikely.

ALINEA Core Module

A form of this module calculates the necessary ramp flow at each entry ramp for local maximization of mainstream throughput according to the ALINEA feedback control algorithm, appropriately extended to enable the handling of multiple local bottlenecks simultaneously. The calculation uses the average mainstream occupancy measurement downstream of the merge (and at multiple downstream bottlenecks) relative to targeted critical occupancy values, at which the bottleneck throughput is maximized. The targeted critical occupancy at a mainline bottleneck is either set by the user or it is dynamically calculated by the critical occupancy estimation module.

Critical Occupancy Estimation Module

This module calculates the critical occupancy at motorway bottlenecks. Where and when this module is activated, the estimated value is used in the ALINEA core module as the targeted critical occupancy value. The adjustment of the critical occupancy value considers the flow/occupancy relationship that optimizes capacity and prevents flow breakdown, as well as the path of flow recovery if flow breakdown has occurred.

Queue Estimation Module

For each controlled entry ramp, this module uses the flow measurements of the ramp entrance and ramp exit detectors, as well as the average occupancy of the detectors in the middle of the ramp to calculate an estimate of the current queue length on the ramp.

Demand Estimation/Prediction Module

This module estimates the arriving ramp demand and makes a short-term prediction based on past values of flow measurements on the ramp.

Waiting Time Estimation Module

This module estimates the waiting time (ramp delay) caused due to ramp meters operation based on past values of the ramp demand and the recent signal control operation. Options for approximate or exact estimation are provided.

Queue Control Module

For each controlled ramp, this module uses the estimate of the queue length of the ramp (calculated by the queue estimation module) and the ramp demand in order to calculate a desired ramp exit flow to minimize the risk of queue overspill. This flow value needs to be determined so that traffic can still be absorbed into the mainline, so as to minimize the potential for causing flow breakdown on the mainline.

Queue Override Module

This module enables the ramp entrance/surface road interface to be managed. In the event that the ramp queue will exceed the available ramp storage, a prespecified ramp exit flow is activated to increase the metering rate. This ramp exit flow value needs to be determined so as to avoid an excessive inflow of traffic to the mainline that may trigger flow breakdown. While optimizing the motorway throughput has benefits to the road network as a whole, consideration may also need to be given to the implications of ramp queues extending onto the surface road.

Waiting Time Control Module

This module uses the waiting time estimate provided by the waiting time estimation module to calculate the desired ramp flow in order to achieve a waiting time on the ramp that is less than a pre-specified maximum waiting time value.

HERO Coordinated Ramp Operation

The HERO module coordinates local ramp metering actions in order to enable efficient mainline control despite limited ramp storage spaces, while balancing ramp queues and providing equity between ramps in a coordinated system.

HERO is activated when a ramp operating under local control experiences queues that meet pre-set thresholds, based on the available ramp storage. In this event, the ramp becomes a “master” and engages, step-by-step, upstream ramps as “slaves.” The algorithm then balances queues between the ramps. Such clusters of coordinated ramps are created wherever and whenever necessary and may be handled simultaneously at multiple motorway locations, as required.

While the system is operating in a coordinated manner under HERO, control using ALINEA and related algorithms at each individual ramp continues according to local needs.

Minimum Queue Control Module

When coordinated ramp metering is in operation, that is, HERO has been activated, this module uses the estimate of the ramp queue length and the ramp demand to calculate a desired metered flow rate so that the minimum queue calculated by HERO is implemented. This calculation aims to make best use of available storage on coordinated ramps and to balance queues between the ramps.

Final Ramp Flow Specification Module

This module calculates the final ramp exit flow to be applied to a ramp at the next control period. In the case of local ramp metering, the final ramp flow decision is based on exit flow values calculated by ALINEA as well as consideration of the various flow values to address ramp queue and waiting time on the ramp. In the case of coordinated ramp metering, the final ramp flow is based on ALINEA as well as individual ramp queues and the balancing of queues between ramps. The final flow rate may also be adjusted to be within prespecified minimum and maximum flow rates.

Implementation Module

The implementation module calculates the cycle time (sum of green, yellow, and red) that corresponds to the final ramp flow, considering the number of lanes at the stop line. The implemented cycle time is changed for each control period within prespecified increments for cycle time increases and decreases.

Ramp Meter Operational Modes

Ramp Meters—Off (Default)

When ramp meters are not operating (default situation), the ramp, including the auxiliary storage lanes, are managed with VSL signs set to default speed limits, i.e. the VSL signs will be always on and displaying the speed limit appropriate to the operating state at that time. On motorway to motorway ramp meters the display for warning VMS on the ramp (RC2-C) will be blank unless required for other relevant traffic situations, for example, during an incident and the display on the mainline warning VMS for exiting motorists (RC3-C) will be blank unless required for some other relevant traffic situation.

Times of Operation

Ramp metering operating at an isolated ramp or within a coordinated system may be activated in a number of ways.

Dynamic Activation and Deactivation

The dynamic switch-on and switch-off of ramp meters is based on the prevailing motorway traffic conditions. A dynamic system provides traffic responsive operation that activates the ramp meter signals at any time when warranted by the motorway traffic flow

conditions that could lead to the onset of flow breakdown. The activation and deactivation thresholds are set uniquely for each ramp/bottleneck during the manual fine tuning of the system.

The switch-on criteria are based on a combination of speed, occupancy and/or volume. Different criteria are used for starting up and switching off the ramp meter signals. The switching on criteria are usually set at a predetermined threshold (determined from analysis of historical conditions) to be sure that the signals start up to keep the motorway flow stable and well before the motorway flow collapses. The criteria need to be comprehensive to avoid the signals switching on at an inappropriate time, for example high occupancy and low speeds may occur at night due to a slow moving maintenance vehicle which may even sometimes stop on a detector. Usually, stronger criteria are used for switching off the signals to ensure the signals will not start up again soon after the deactivation.

Time of Day Activation

Although no longer broadly used by VicRoads for a primary control tool, time-of-day activation may be used according to critical periods, generally morning and evening weekday peak periods. Scheduled start-up and close-down times are chosen following an analysis of motorway and entry ramp flows during the peak periods and their respective shoulder periods. Typical times of operation would be 6:00–10:00 am for the AM peak period and between 3:00 pm and 7:00 pm for the PM peak period. Other times may also be scheduled to cover known occasions outside weekday peak periods where data shows that the motorway service is at risk, for example, Saturday shopping periods or special events.

Time of day parameters may also limit the times within which dynamic activation and deactivation may occur. Under this operation the metering signals may or may not switch on, depending on whether the criteria are met. This form of control for activation is advisable during the initial operation of a new coordinated system and when testing criteria for full dynamic activation.

During Incidents and Events

During periods of light traffic flow on the motorway (when the metering signals would normally be off), there may be advantages in using the signals to manage the headway of entering traffic or to manage the traffic flow. This may be necessary at times of a lane closure or traffic flow breakdown due to planned or unplanned incidents, for example, roadworks, crashes etc., to assist in traffic management and/or to facilitate flow recovery (refer to [VicRoads 2020a](#)).

Manual Operation

At any time when considered warranted, VicRoads is able to manually switch on the ramp metering signals, override the dynamic operation or switch the metering signals off.

Manual operation of other traffic management devices, for example, speed limit and lane control signs or VMS, including modification of displayed messages is also subject to VicRoads manual control, if warranted. Manual operation within the constraints of the system overrides default and dynamic operation.

Faults and Device Failures

During a communications failure (power still available), operation of the ramp metering would occur using pre-determined fixed time signal cycles which are suitable to the day of the week and time of day. In this situation VicRoads is to initiate close monitoring of the system as well as surveillance with CCTV cameras. These signal timings are reviewed periodically to ensure their appropriateness relative to changing traffic conditions over time.

With failure of advance devices at two or more locations upstream of the ramp meters, for example, warning signs (RC2-C), speed limit signs or VMS (RC3-C), the ramp metering signals will not switch on, or if already on, the ramp metering signals will switch off.

Switching on/off Signs and Signals

Start-up Sequence

The sequence for switching on typical ramp meters and associated ramp control and warning signs is shown in [Fig. 2](#). Prior to start up the signals and RC1 and RC2 signs have no display.

1. Switch Sign RC1 and Sign RC2 to “RAMP SIGNALS ON” and activate the signals to “flashing yellow” for 10 s. Activate the reduced speed limit (if applicable).
2. Switch on the alternating messages on Sign RC2, if provided, and switch traffic signals to “solid yellow” for 4 s.
3. Switch traffic signals to “solid red” for 6 s.
4. Commence the metering cycle with the initial green and continue the metering.

Separate sequences for switching on motorway to motorway ramps are included in [VicRoads \(2020b\)](#).

Close-down Sequence

The sequence for switching off the signals is shown in [Fig. 3](#) and described below.

1. Activate traffic signals to “flashing yellow” and switch Sign RC2, if provided, to “Ramp Signals ON” only (no alternating message) for 10 s.

	Device operation and time	RC1 sign	RC2 sign	Signals
1.	Prior to 'Start Up' Signals: Off Signs: Off			
2.	'Start up' Period (10 s) Signals: Flashing yellow Signs: 'Ramp Signals On'			 Flashing Yellow
3.	'Start up' Period (next 4 s) Signals: Solid Yellow Signs: 'Ramp Signals On' / 'Prepare to Stop' alternating		 Alternating Messages	 Solid Yellow
4.	'Start up' Period (next 6 s) Signals: Solid Red Signs: 'Ramp Signals On' / 'Prepare to Stop' alternating		 Alternating Messages	 Solid Red
5.	Signals commence metering Signals: Green-Yellow-Red Signs: 'Ramp Signals On' / 'Prepare to Stop' alternating		 Alternating Messages	 Green-Yellow-Red

Figure 2 Start-up control sequence—typical ramp meter.

- Switch off Sign RC1, Sign RC2, the “flashing yellow” of the signals and return the speed limit to default or another override value (if applicable).

Separate sequences for switching on motorway to motorway ramps are included in [VicRoads \(2020b\)](#).

Operating Sequence and Cycle Times (Not Used for Design)

Signal Timings

Ramp metering operation, which is considered differently to ramp metering design, has a variable cycle time generally in the range 4.0–20 s according to the determined metering rate. The sequence times based on one vehicle per green per lane are:

- Red variable—generally within the range 2.0–18 s
- Green 1.3 s
- Yellow 0.7 s.

Ramp metering design uses an average cycle time as the basis for design (e.g., an average cycle time of 7.5 s). In operations cycle times will be variable changing as often as every 20 s around an average to accommodate fluctuations in demand within the network

	Device operation and time	RC1	RC2	Signals
1.	Signals metering (Prior to 'Close Down') Signals: Green-Yellow-Red Signs: 'Ramp Signals On' / 'Prepare to Stop' alternating		 Alternating Messages	 Green-Yellow-Red
2.	'Close Down' Commences (10 s) Signals: Flashing yellow Signs: 'Ramp Signals On'			 Flashing Yellow
3.	Switch off Signals (Close down complete) Signals: Off Signs: Off			 Signals off

Figure 3 Close-down control sequence—typical ramp meter.

in real time. Designers shall not nominate a lower cycle time than specified in [VicRoads \(2019a\)](#), in an attempt to discharge higher flow rates into the motorway for the purposes of reducing storage provision. Such an approach reduces the necessary flexibility in operation and ramps are to be designed on the bases of ramp flows that can be achieved over a 1-h period (e.g., utilizing an average cycle times of 7.5 s—refer to [VicRoads 2019a](#), Section 6.2 for further details).

During ramp metering operations, when there are no vehicles waiting at the stop line, the signals are to be held on red. This prevents the signals cycling when there are no waiting vehicles and helps to avoid driver confusion in relation to timing their arrival and deciding whether to stop or not.

The operation of more than one vehicle per green per lane (e.g., two or three vehicles per green per lane) has not been permanently implemented in Australia, although it has been used internationally (refer to [VicRoads 2019a](#), Section 6.2). Generally, it is desirable to release a single vehicle per green per lane, even if shorter cycle times need to be adopted.

The entry ramp flows that result from a range of cycle times with varying lane arrangements at the stop line are shown in [Table 1](#). In practice, within a dynamic system the cycle time is based on the ability of the freeway to accommodate entering traffic. The signals apply to all lanes at the stop line.

Pre-set time of day signal cycle settings are used as a “fall back” mode when a dynamic system experiences a fault and fail safe mode is activated.

Minimum Cycle Time (Operations)

The general minimum cycle time that provides a high entry ramp flow is 5 s for a ramp leading to a motorway merge. A general minimum cycle time of 4 s could be considered when ramp traffic enters the motorway via an added lane(s)—shaded yellow in [Fig. 4](#). The minimum cycle time would generally be observed to operate under light mainline conditions, e.g. during the fringe of the peak periods or where the upstream mainline motorway flow is interrupted due to an unplanned incident. However, low cycle time values are not appropriate in design as an average value over the design hour (refer to [VicRoads 2019a](#), Section 6.2).

Cycle times lower than the minimum indicated are not generally recommended as this could approach a situation where the discharge of vehicles is almost continuous and the metering is relatively ineffective in managing headways. However, lower cycle times during operations under certain conditions may be appropriate subject to trial and assessment of driver behavior.

Table 1 Discharge capacity (stop line lanes) and ramp storage requirements

Indicative layout ^a	Ramp design flow ^a (pc/h)	Total storage required (lane-meters)	Ramp storage ^b and cycle time ^c relative to the number of lanes at the stop line							
			1 Lane	2 Lanes	3 Lanes	4 Lanes				
			Average storage per lane (m)	Average storage per lane (m)	Average storage per lane (m)	Average storage per lane (m)	Average cycle time (s)	Average cycle time (s)	Average cycle time (s)	Average cycle time (s)
Single lane merge ^d	200	113	113	18.0						
	300	170	170	12.0						
	400	227	227	9.0						
	500	283	283	7.2	142	14.4				
	600	340	340	6.0	170	12.0				
	700	397			198	10.3				
	800	453			227	9.0				
	900	510			255	8.0	170	12.0		
	1000	567			283	7.2	189	10.8		
	1100	623			312	6.5	208	9.8		
	1200	680			340	6.0	227	9.0		
Added lane entering the freeway or Two lane merge Ramp	1300	737					246	8.3	184	11.1
	1400	793					264	7.7	198	10.3
	1500	850					283	7.2	213	9.6
	1600	907					302	6.8	227	9.0
	1700	963					321	6.4	241	8.5
	1800	1020					340	6.0	255	8.0
Added lane entering the freeway plus a merging lane	1900	1077							269	7.6
	2000	1133							283	7.2
	2100	1190							298	6.9
	2200	1247							312	6.5
	2300	1303							326	6.3
	2400	1360							340	6.0
	2500	1417							354	5.8
	2600	1473							368	5.5
	2700	1530							383	5.3
	2800	1587							397	5.1
	2900	1643							411	5.0
	3000	1700							425	4.8

Max wait/vehicle (min.): 4.

Storage per vehicle (m): 8.5.

No. of vehicles/green/lane: 1.

^aRamp layout and design flow are subject to the mainline capacity and mainline/ramp configuration.^bAverage storage per lane assumes lanes of equal length. Not applicable with auxiliary lanes at the stop line.^cNumbers shown in bold would only be appropriate when the mainline analysis indicates ramp demands are accommodated with spare capacity for several downstream interchanges.^dA single lane merge layout may be satisfactory for higher flows, for example a ramp flow of 1400 vehicles/h entering mainline with adequate capacity.**Maximum Cycle Time (Operations)**

The general maximum cycle time that provides a low entry ramp flow is 16 s and would be implemented when freeway mainline flow is close to or above critical values of occupancy. Higher cycle times, shaded pink in [Fig. 4](#), may be implemented subject to adopted operational management strategies and driver acceptance, such as in situations of very heavy motorway congestion, during an incident or when the motorway is recovering from an incident (refer to [VicRoads, 2020a](#)).

Mainline and Entry and Exit Ramp Design**Introduction**

The era when road reservations for motorways were planned and set aside many years earlier for the purpose of building future motorways, or when motorways were progressively built and upgraded to meet traffic demand as the city expanded as populations grew, is receding in many large urban areas. In this era, it was possible to design new and upgraded motorways by targeting forecast demand for the design life, and then applying available road and traffic design standards. In the current era, projects are generally no

Cycle time (sec)	Cycle per hour (No.)	Equivalent ramp flow per hour (veh/h)			
		No. of lanes at Stop Line			
		1	2	3	4
4.0	900.0	900	1,800	2,700	3,600
4.5	800.0	800	1,600	2,400	3,200
5.0	720.0	720	1,440	2,160	2,880
5.5	654.5	655	1,309	1,964	2,618
6.0	600.0	600	1,200	1,800	2,400
6.5	553.8	554	1,108	1,662	2,215
7.0	514.3	514	1,029	1,543	2,057
7.5	480.0	480	960	1,440	1,920
8.0	450.0	450	900	1,350	1,800
8.5	423.5	424	847	1,271	1,694
9.0	400.0	400	800	1,200	1,600
9.5	378.9	379	758	1,137	1,516
10.0	360.0	360	720	1,080	1,440
10.5	342.9	343	686	1,029	1,371
11.0	327.3	327	655	982	1,309
11.5	313.0	313	626	939	1,252
12.0	300.0	300	600	900	1,200
12.5	288.0	288	576	864	1,152
13.0	276.9	277	554	831	1,108
13.5	266.7	267	533	800	1,067
14.0	257.1	257	514	771	1,029
14.5	248.3	248	497	745	993
15.0	240.0	240	480	720	960
15.5	232.3	232	465	697	929
16.0	225.0	225	450	675	900
16.5	218.2	218	436	655	873
17.0	211.8	212	424	635	847
17.5	205.7	206	411	617	823
18.0	200.0	200	400	600	800
18.5	194.6	195	389	584	778
19.0	189.5	189	379	568	758
19.5	184.6	185	369	554	738
20.0	180.0	180	360	540	720

Figure 4 Equivalent hourly ramp flows relative to cycle time and lanes at the stop line for operations.

longer just a civil design exercise. They require considerably more understanding of current problems and operation in the context of the broader road network as well as anticipated future operation and performance. The current reservations may have limited opportunities for widening of motorways or establishment of reservations to meet current demands, let alone meet forecast demands. Many corridors have reached their practical limits, especially during extended peak periods, creating a need for targeted civil upgrade improvements (based on traffic analysis) and retrofitting with traffic management devices to improve safety and efficiency and provide the ability to manage actual demand.

A further aspect of the current era is that it may not be possible to meet the traffic demands for a desirable design life for major projects where widening of existing motorways or building new links is technically feasible, but where practical limitations exist due to high cost or other constraints (for example, most tunnel options). This outcome then requires special consideration of optimizing design aspects and operational strategies to ensure they can be managed to Maximum Sustainable Flow Rates (MSFR) in real time.

Performance-based Design

In this context, performance-based design becomes a specific focus where new or upgraded infrastructure needs to bring together many elements of managed motorway operations into the design, with a focus on design and operation intent (such as safety and

productivity performance outcomes), as well as an understanding of operations and maintenance elements. New approaches require collection of traffic data, data analysis and refinement of traffic modelling which, while they may sound traditional, need to be combined with the adoption of new or appropriate design thinking. This thinking links physical features to likely operation performance such as traffic turbulence, bottleneck triggers and capacity impacts. It also requires an understanding of the managed motorway toolkit, including the relative benefits of the various traffic management systems that can guide designers and project managers in priorities. Motorway design is no longer just a geometric design and alignment exercise—it considers the impacts of multiple dynamic elements impacting driver responses and interactions in the context of the design. Designers need to be fully aware that there are now many integrated managed motorway real-time backend systems and automated decision support services requiring the provision of accurate data from the field for real-time feedback control, historical analysis and performance measurement.

The road designer and/or project delivery manager must be cognizant of the downstream operational activities that will occur in the future. These matters place particular attention on the design decision making processes and the technology provided, e.g. the precise location of vehicle detectors for richer data sets requiring high accuracy and high levels of availability. The performance-based design approach considers smooth entry into the motorway, smooth transition along the motorway (designing for variable traffic patterns and traffic demands) and smooth exit from the motorway network. Special advanced control systems are needed for both entry and exit ramp management and to ensure traffic is smoothly managed out of, and into the arterial road network and that the different traffic control systems (in Melbourne, STREAMS and SCATS) are effectively interfaced to manage demands. This section will focus on design for entry ramps including motorway to motorway connections.

Design of Ramp Meters and Entry Ramps

Overview of the Design Process

Every ramp meter has several unique roles to play to control traffic locally and as part of a larger network of ramps to ensure the supply of traffic downstream arriving at bottlenecks remains within sustainable flow limits across the entire day. The main components of the ramp meter design process at each entry ramp follows an investigation along the motorway to understand motorway mainline design, capacity and performance, including at each ramp entry to the mainline and other potential bottlenecks. The main components of the ramp metering design process are shown in Fig. 5.

The VicRoads managed motorway approach (different when compared to most other jurisdictions) is that it applies holistic principles to planning, design and operations, as well as state-of-the-art best practice (traffic theory, principles and analysis) to the traffic management control and algorithms that manage the system.

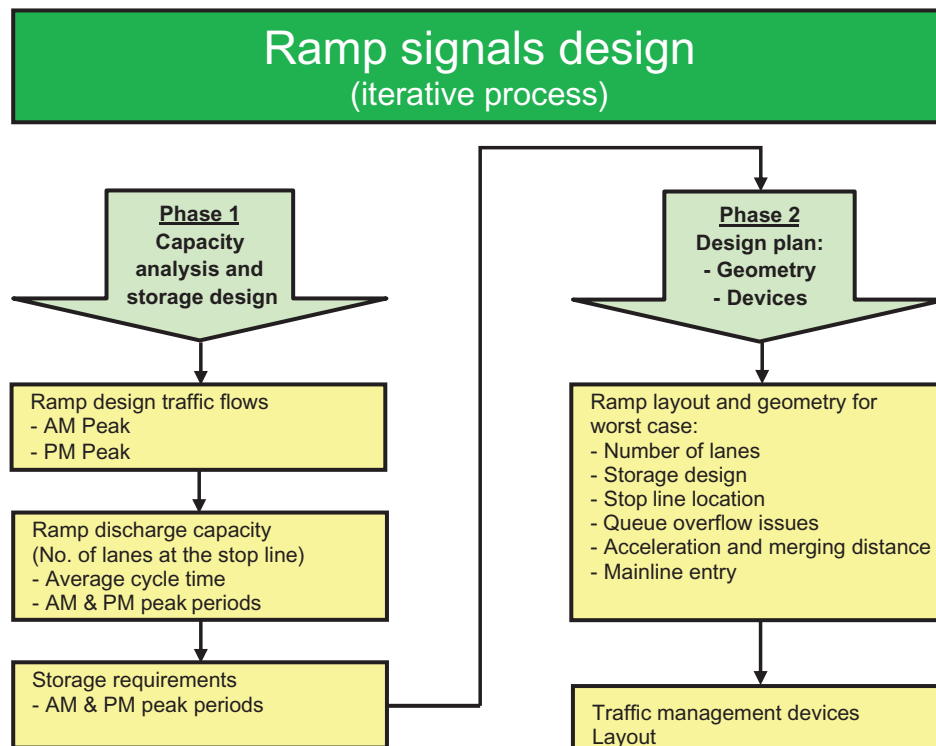


Figure 5 Ramp meter design process overview.

Historic examples exist where ramp metering signals were installed and subsequently did not meet operational expectations due to inadequate analysis, poor design, and/or misunderstanding of ramp metering principles. Performance problems may involve excessive queuing and delays or other adverse effects on both the ramp and motorway mainline. A good control algorithm may not be able to compensate for an inadequate design.

The ramp storage and discharge capacity is related to the number of lanes at the stop line as well as that of merging/added lanes, the ramp arrival flow, adopted for design, the number of vehicles per green per lane and an appropriate average design cycle time, to provide the metering of traffic into the mainline (refer to [Table 1](#)).

As discussed above, every ramp meter or motorway to motorway connection must be uniquely designed for its role in managing the motorway or broader road network. Therefore attention to detail is crucial such as ramp length and number of lanes at the stop line. Shown in [Fig. 6](#) is a typical 2 lane ramp meter, however VicRoads has many three and four lane meters with details as shown in ([VicRoads, 2019a](#)). Where the stop lines are located in relation to the ramp nose and mainline merge taper is important as vehicles need to be able to accelerate up to speed before entering the motorway. Also the placement of ITS devices and especially ramp control (VMS) signs and vehicle detectors are critical for many effective algorithm and control functions.



Figure 6 Typical motorway ramp metering signals layout—two metered lanes.

Conclusion and Findings

With the hindsight of today's knowledge and experience (i.e., 2020), the relatively simplistic beginnings of ramp metering (refer to the chapter on *Urban Motorway Management* in this encyclopedia—Section 3.2) were based on empirical evidence and have provided a solid foundation that has since lead to further advances in the design of motorways themselves to be more productive by limiting unnecessary turbulence, as well as leading to major advances in the way motorway control systems are designed and operate in real-time. This has been made possible by a combination of hands-on operational experience and the detailed data analysis using statistical techniques as described in the VicRoads comprehensive overview of contemporary traffic science in (VicRoads, 2020a) and the VicRoads Motorway Capacity Guide (VicRoads, 2019b). Significant and essential advances in data quality and availability have also led to statistical analysis techniques that can indicate the development of precursor bottleneck conditions that need to be managed before instability and flow breakdown occur (refer to VicRoads, 2019b).

Traffic congestion on urban motorways worldwide, with the corresponding impacts on safety, efficiency, CO₂ emissions, etc. occurs every day of the year, over multiple hours and many kilometers and, if uncontrolled, due to traffic growth this problem will only increase. There is often insufficient knowledge of the magnitude of the social and economic damage (usually equating to tens of millions of dollars per year on a single corridor and hundreds of millions per annum in a city the size of Melbourne) and this can be due to a lack of suitable measurement infrastructure installed and efforts to appropriately analyze the measured data.

With appropriate motorway design elements, correctly located and performing ITS devices coupled with a state-of-the-art control system and operations regime (policy and practice; refer to VicRoads 2016), it is possible to manage these assets to deliver greatly improved outcomes for the community and the economy. An evaluation of the VicRoads CWCRCM on a per lane basis has shown an increase in productivity (speed multiplied by flow) by 40%–50% and a reduction in crashes by 12% (19% reduction in fatal and 10% in serious/other crashes) compared to the before upgrade situation (VicRoads, 2017).

Investment in the advanced technologies required to improve road safety and network efficiency will at the same time provide enhanced functionality for traditional Traffic Management Centre services typically delivered by road agencies, i.e., incident management, access control, lane control and speed control enabling faster and safer incident response, clearance time and flow recovery time.

Acknowledgments

VicRoads acknowledges the partnership and contributions of TRANSMAX as the owner and developer of the VicRoads Motorway Management system (STREAMS) and Prof. Markos Papageorgiou and Prof. Ioannis Papamichail from the Technical University of Crete in the development of the HERO Live coordinated ramp metering system used to manage Melbourne's motorways.

See Also

Urban Motorway Management - Keeping urban motorways safe and efficient with City Wide Coordinated Ramp Meters
Ramp Metering Application - Keeping urban motorways safe and efficient with City Wide Coordinated Ramp Meters

References

- Papageorgiou, M., Hadj-Salem, H., Middelham, F., 1991. ALINEA: A Local Feedback Control Law for Onramp Metering. Transportation Research Record No. 1320. Transportation Research Board (TRB), Washington, D.C., pp. 58–64.
- Papageorgiou, M., Hadj-Salem, M., Middelham, F., 1997. ALINEA Local Ramp Metering Summary of Field Results. Transportation Research Record No. 1603. Transportation Research Board (TRB), Washington, D.C., pp. 90–98.
- VicRoads, 2016. VicRoads Operational Control of the Motorway Network—endorsed. VicRoads, Melbourne.
- VicRoads, 2017. Managed Motorway Framework. VicRoads, Melbourne.
- VicRoads, 2019a. VicRoads MMDG, vol. 2, Part 3. Motorway Planning and Design. Department of Transport (VicRoads), Melbourne.
- VicRoads, 2019b. VicRoads MMDG, vol. 1, Part 3. Motorway Capacity Guide. Department of Transport (VicRoads), Melbourne.
- VicRoads, 2020a. VicRoads MMDG, vol. 1, Part 2. Traffic Theory Relating to Urban Motorways. Department of Transport (VicRoads), Melbourne.
- VicRoads, 2020b. VicRoads MMDG, vol. 2, Part 2. Network Optimisation Tools. Department of Transport (VicRoads), Melbourne.

Further Reading

- Kerner, B., 2017. Breakdown in Traffic Networks—Fundamentals of Transportation Science. Springer, Berlin.
- Treiber, M., Kesting, A., 2013. Traffic Flow Dynamics—Data Models and Simulation. Springer, Berlin.
- VicRoads, 2017. Managed Motorway Framework. Department of Transport (VicRoads), Melbourne.

Variable Speed Limits for Traffic Efficiency Improvement

José Ramón D. Frejo*, Bart De Schutter†, *Systems and Automation Engineering Department, University of Seville, Seville, Spain;
†Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	33
Macroscopic Modeling of Variable Speed Limits on Freeways for Traffic Efficiency Improvement	34
VSL Model of Hegyi et al.	34
VSL Model of Carlson et al.	35
VSL Model of Frejo et al.	36
Model Comparison	37
VSL Control Algorithms for Traffic Efficiency Improvement by Mainline Metering	37
Optimization-Based Versus Easy-to-Implement VSL Controllers	37
Local Versus Global VSL Controllers	38
Discrete Versus Continuous VSL Controllers	38
Summary of VSL Control Algorithms	38
Integration With Other Freeway Traffic Control Measures	38
Conclusions	39
Acknowledgment	39
References	39

Introduction

Nowadays, many social and economic problems such as waste of time and fuel, greater accident risks, and an increase in pollution are caused by traffic jams on freeways. Due to economic or viability reasons, the construction of new freeways is not always the best solution. One promising alternative is the use of dynamic control signals such as ramp metering, reversible lanes, lane change control, and route guidance. Besides these traffic control measures, in the past years, Variable Speed Limits (VSLs) have emerged as a potential traffic management measure for increasing freeway efficiency.

The first implementation of VSLs dates back to 1970 in Germany (Zackor, 1972). During the subsequent decades, VSLs have been also applied in The Netherlands, United States, and other countries. Earlier studies assumed that, if a VSL is applied, the capacity of a freeway segment is increased due to speed harmonization within each lane and across different lanes. Thus, the main objective, beyond improving traffic safety, was to increase freeway efficiency by applying VSLs to congested bottlenecks. However, later works did not identify any capacity increase due to VSLs; even a capacity reduction is observed concluding that a substantial improvement of the traffic efficiency cannot be reached by applying VSLs to increase capacity at congested bottlenecks.

Indeed, the most recent empirical validations (Frejo et al., 2019) have shown that a reduction of the value of a VSL has the following effects on the fundamental diagram of traffic flow: an increase in the critical density, a decrease in the capacity, and a decrease in the free-flow speed. The most well-known macroscopic models that incorporate these effects of the VSLs on traffic flow are introduced in the section, Macroscopic modeling of variable speed limits on freeways for traffic efficiency improvement.

Nowadays, and based on the fact that it is not possible to directly increase the capacity of a segment by using VSLs, three main kinds of VSL control algorithms are considered in the recent literature:

- **Safety-oriented VSLs:** This kind of VSL control algorithms are applied in order to increase traffic safety by reducing the speed limit when a vehicle is approaching a congested segment or an incident or by homogenizing speeds to reduce fluctuations in traffic behavior. The safety benefits of these control algorithms have been reported in field studies. However, these applications have not yet achieved a significant improvement of the flow or the capacity.
- **VSLs for traffic efficiency improvement:** The goal of these VSL control algorithms is to avoid, resolve, or reduce traffic jams on freeways by limiting the arriving flow and, thereby, avoiding or mitigating the capacity drop. These techniques have been mainly tested by simulation and, therefore, their potential improvement in the traffic conditions relies on the model used. Moreover, the use of model-based optimal control algorithms (like Model Predictive Control) can considerably improve the Total Time Spent (TTS), emissions, fuel consumption, and other traffic performance indices, but their success highly depends on model accuracy.
- **Pollution-oriented algorithms:** Pollution-oriented VSL control algorithms, usually based on the structure of a controller designed for traffic efficiency improvement, provide a balanced trade-off between travel times, fuel consumption, and emissions. However, the reduction of the emissions that can be achieved without decreasing traffic efficiency is usually relatively low.

This work focuses on the second kind of algorithms since they have attracted most of the research attention during the last years and because they have a much higher potential (in terms of improvement of the traffic conditions). More concretely, the most

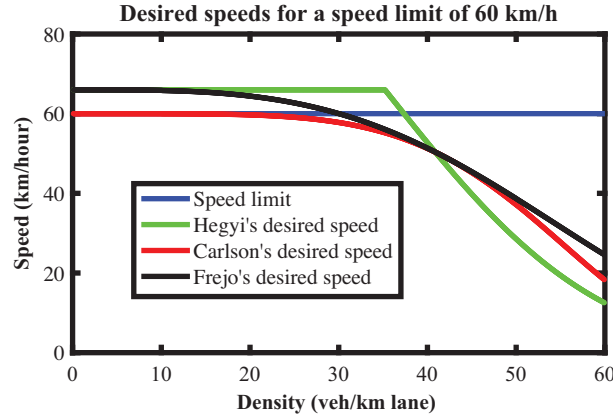


Figure 1 Desired speeds. The values used for the FD parameters are $v_{f,i} = 120$ km/h, $\rho_{c,i} = 30$ veh/(km lane), $a_i = 2.5$, $V_{c,i}(k) = 60$ km/h, $a_i = 1.4$, $A_i = 0.8$, $E_i = 3$ (for Carlson's model), $A_i = 0.889$ and $E_i = 1.78$ (for Frejo's model).

commonly used control approaches for the control of VSLs to improve traffic efficiency are presented in the section, VSL Control Algorithms for Traffic Efficiency Improvement by Mainline Metering. Furthermore, the main conclusions and major open challenges are drawn in section, Conclusions.

Macroscopic Modeling of Variable Speed Limits on Freeways for Traffic Efficiency Improvement

In model-based freeway traffic control using VSLs, an accurate, but significantly faster than real-time, prediction model is necessary for the computation of the control inputs. On the other hand, although easy-to-implement controllers do not usually need a traffic model for the computation of the control inputs, an accurate model is also necessary for calibration and validation purposes.

Two main approaches are mostly used for the traffic flow modeling:

- **Microscopic models**, which describe the longitudinal and lateral movement of individual vehicles.
- **Macroscopic models**, which mostly model traffic as a particular fluid with aggregate variables such as density, mean speed, and flow.

One of the main advantages of microscopic models is their ability to study individual vehicle motion visually, perceiving the real process using simulation software. However, microscopic models are difficult to use for model-based control due to the computational effort needed; thus, these models are usually only used for simulation and off-line tuning of controllers. On the other hand, macroscopic models are more suitable for model-based applications because they are relatively fast and have an analytical form. This work, therefore, focuses on the modeling of VSLs within the framework of macroscopic traffic models.

Within macroscopic traffic flow models, first-order models like the cell transmission model include a static speed-density relationship. On the other hand, second-order models like METANET address the speed as another state variable with an according state equation, which is capable of capturing additional dynamics and, more important, is able to model capacity drop. Therefore, this work focuses on the modeling of VSLs by modifying the second-order model METANET. It should be noted that in the literature a first-order model that includes the effects of VSL (Han et al., 2017) it can be also found. In the related literature, three main ways of including the effects of VSLs in METANET have been considered. These three approaches are presented in the further sections.

VSL Model of Hegyi et al.

The first macroscopic model for VSLs was proposed by Hegyi et al. (2004). In this model, VSLs are included by modifying the desired speed equation. When a VSL is applied, the desired speed $V_i(k)$ is computed by taking the minimum of two quantities: the speed based on the prevailing traffic conditions, and the speed caused by the VSL (affected by a compliance term):

$$V_i(k) = \min \left(v_{f,i} e^{-\frac{1}{a_i} \left(\frac{\rho_i(k)}{\rho_{c,i}} \right)}, (1 + \alpha_i) V_{c,i}(k) \right)$$

where $V_{c,i}(k)$ is the value of the speed limit on link i , α_i is a model parameter that reflects the driver compliance, $\rho_i(k)$ is the density on link i , $v_{f,i}$ is the free-flow speed, $\rho_{c,i}$ is the critical density, and a_i is a parameter of the METANET model. An example of the desired speed of a link obtained under the influence of a VSL can be seen in Fig. 1. The corresponding Fundamental Diagram (FD), the steady-state relation between flow and density, can be seen in Fig. 2.

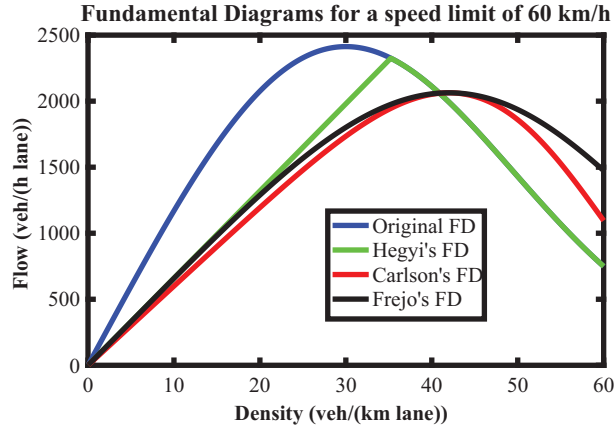


Figure 2 Fundamental diagrams with the same parameters as Fig. 1.

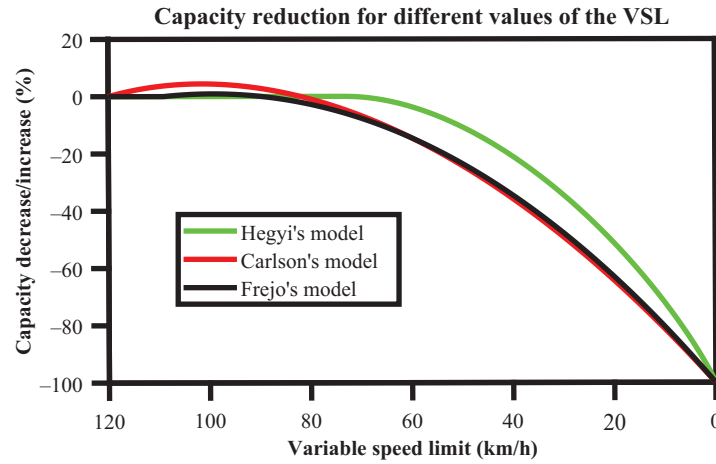


Figure 3 Capacity reductions (%) with respect to the original FD with the same parameters as Fig. 1.

Assuming that the parameters of the original METANET model are known, the effect of the application of VSLs on the three characteristic parameters of the FD (critical density, free flow speed, and capacity) according to Hegyi's model is the following:

- **VSL-induced free-flow speed:** Compliance can be adjusted by parameter a and, therefore, for a given value of $\nu_{f,i}$ and $\nu_{c,i}(k)$, any value (between $\nu_{c,i}(k)$ and $\nu_{f,i}$) for the VSL-induced free flow speed can be set according to the measurements.
- **VSL-induced critical density:** The new critical density is computed by intersecting the original FD with the speed line given by the VSL-induced free flow speed or by taking the original critical density if the intersection is on the uncongested part of the FD. For a given compliance and original FD, the VSL-induced critical density cannot be adjusted (i.e., calibrated according to measured data).
- **VSL-induced capacity:** The obtained capacity reduction (with respect to the capacity corresponding to a speed limit of 120 km/h) as a function of the value of the speed limit is shown in Fig. 3. As can be seen in the figure, the capacity reduction is usually relatively small or even zero for high and intermediate speed limits. For example, for the FD shown in Fig. 2, the capacity is reduced only by about 3.65% with the VSL being half (60 km/h) of the nominal speed limit (120 km/h).

VSL Model of Carlson et al.

The second model that is commonly used in the literature was proposed by Papamichail et al. (2008) in 2008 and it has been mainly used by Carlson et al. (2010). In this model, FD parameters are modified to incorporate the influence of the VSLs. The speed ratio $b_i(k)$ (defined between 0 and 1) is used instead of the implemented value of the speed limit $V_{c,i}(k)$:

$$V_i(k) = \nu_{f,i}^*(k) e^{-\frac{1}{a_i^*(k)} \left(\rho_i \frac{(k)}{\rho_{c,i}^*(k)} \right)^{a_i^*(k)}} \quad b_i(k) = V_{c,i} \frac{(k)}{V_{\max_{c,i}}(k)} \quad \nu_{f,i}^*(k) = \nu_{f,i} b_i(k)$$

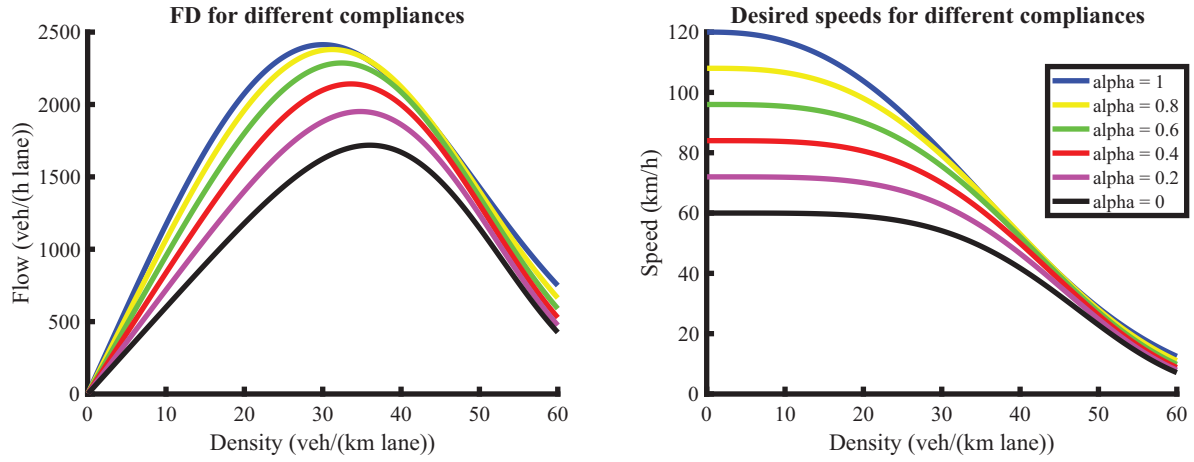


Figure 4 FD and desired speed for different levels of compliance using Frejo's model.

$$\rho_{c,i}^*(k) = \rho_{c,i}(1 + A_i(1 - b_i(k)))a_i^*(k) = a_i(E_i - (E_i - 1)b_i(k))$$

where A_i and E_i are parameters of Carlson's model.

An example of the desired speed and FD obtained with this model can be seen in **Figs. 1 and 2**.

According to Carlson's model, the three characteristic parameters of the FD change as follows:

- **VSL-induced free-flow speed:** Since compliance is not included, it is assumed that the VSL-induced free-flow speed is equal to the value of the corresponding speed limit (i.e., for steady states with low densities, the mean speed of a link will be equal the current VSL value in the corresponding link).

VSL-induced critical density: The critical density induced by a VSL can easily be adjusted in order to approach measured data by changing the parameter A_i .

VSL-induced capacity: The capacity induced by a VSL can be calibrated by adjusting the parameters E_i and A_i .

The obtained capacity reduction is shown in **Fig. 3**. As can be seen in the figure, for the given values of the model parameters, the capacity is slightly increased for high values of the VSL. On the other hand, for intermediate and low values of the VSL, there is a significant capacity reduction.

VSL Model of Frejo et al.

As previously explained, Hegyis model can be adapted for different VSL-induced free-flow speeds, but it lacks the capability of modeling different capacity reductions and critical densities due to the effects of VSLs. On the other hand, Carlson's model is able to model different VSL-induced capacities and critical densities, but it does not consider different levels of compliance.

The model proposed by Frejo et al. (2019) combines the advantages of both approaches: The effect of a VSL is modeled by modifying the FD parameters including a calibration parameter for each equation (a_i for $v_{f,i}$, A_i for $\rho_{c,i}$ and E_i for a_i):

$$V_i(k) = v_{f,i}^*(k)e^{-\frac{1}{a_i^*(k)}\left(\rho_{c,i}^*(k)\right)} \quad b_i^*(k) = \min\left(V_{c,i} \frac{v_{f,i}^*(k)}{V^{\max}}(1 + \alpha_i), 1\right)$$

$$v_{f,i}^*(k) = \min(V^{\max} b_i^*(k), v_{f,i})$$

$$\rho_{c,i}^*(k) = \rho_{c,i}(1 + A_i(1 - b_i^*(k))) \quad a_i^*(k) = a_i(E_i - (E_i - 1)b_i^*(k))$$

where $V^{\max}$ is the maximum allowed value for the VSLs.

An example of the desired speed and FD obtained with the proposed model can be seen in **Figs. 1 and 2**. According to the model, the three characteristics parameters of the FD change as follows:

- **VSL-induced free-flow speed:** Compliance with variable speed limits can be adjusted by the parameter a_i ; hence the VSL-induced free-flow speed can be freely set according to measured data. It is noted that compliance is also affecting the VSL-induced critical density and capacity. This allows to adapt an already calibrated model if the compliance changes (e.g., if the speed limit compliance is enforced by fines or other measures). In **Fig. 4**, the FDs and the desired speeds obtained for different values of the compliance are shown. The figure is equivalent to the one that would be obtained for different values of the VSL having the same compliance. It can be seen that, as expected from real traffic, the influence of the speed limits (and their compliance) is very low when the system is considerably congested.

- **VSL-induced critical density:** The critical density induced by a VSL can be adjusted in order to approximate measured data by changing the parameter A_i .
- **VSL-induced capacity:** The VSL-induced capacity can be also adjusted in order to fit measurements by using the parameters E_i and A_i .

Model Comparison

Hegyi's model can be adapted for different VSL-induced free-flow speeds, but it lacks the capability of modeling different VSL-induced capacity reductions and critical densities while Carlson's model is able to model different VSL-induced capacities and critical densities, but it does not consider different levels of compliance. On the other hand, Frejo's model can be adapted for different VSL-induced free-flow speeds, capacities, and critical densities. The main disadvantage of Frejo's model is that three parameters have to be adjusted for each link (while only one is necessary for Hegyi's model and two for Carlson's model). The additional parameter is not going to affect substantially simulation times or optimization computational load. However, the calibration effort will be accordingly increased.

As in any other modeling problem, model choice will depend on the circumstances of each case. For example, if the VSLs are only going to be applied for very low densities in a freeway with full compliance, the three models are going to perform similarly. In this case, Hegyi's model would be the best choice (since it is the simplest of the three considered models). However, if the compliance is not high and the capacity is considerably reduced when VSLs are applied or if the desired speeds are considerably changing for different values of uncongested densities the use of Frejo's model is justified (especially from a control point of view).

VSL Control Algorithms for Traffic Efficiency Improvement by Mainline Metering

Many different control algorithms have been proposed for VSLs during the last decades. This work classifies them according to their main characteristics.

Optimization-Based Versus Easy-to-Implement VSL Controllers

Two main approaches have been used in the literature for the control of VSLs in the context of traffic efficiency improvement: Optimization-based approaches and easy-to-implement controllers.

- **Optimization-based controllers:** This kind of control algorithms computes the values of the speed limits based on the minimization of a cost function. The most commonly used cost is the Total Time Spent (TTS), but also other traffic performance indices have been considered in the literature, such as emissions or fuel consumption. Optimization-based approaches using Model Predictive Control (MPC), which minimizes the cost function using a receding horizon approach, has shown to substantially improve the performance of the controlled traffic network in various simulation studies (Carlson et al., 2010). The main drawback of MPC is that the computation time quickly increases with the size of the network, making it difficult to apply centralized MPC for large traffic networks. The use of a distributed architecture (Frejo and Camacho, 2011) and/or of a suitable optimization algorithm such as the RPROP algorithm (Pasquale et al., 2016) may relieve these limitations but, unfortunately, the obtained controllers are still too complex to be implemented in real time for large networks. Moreover, MPC controllers for VSLs are not robust in case of communication or measurement errors. For this and other reasons, completely centralized control of large traffic networks is still viewed by most practitioners as impractical and unrealistic.
- **Easy-to-implement controllers:** In order to overcome the practical limitations of the optimization-based controllers, a few easy-to-implement control algorithms have been designed for the control of VSLs. The most well-known ones are SPECIALIST (Hegyi et al., 2008), MTFC (Carlson et al., 2011), and LB-VSL (Frejo and De Schutter, 2018).

Compared with optimization-based techniques, these VSLs controllers have the following advantages:

- Controller implementation is quite easy because the online computation is almost instantaneous in most cases.
 - Moreover, easy-to-implement controllers are robust against communication and measurement failures in other segments different from the bottleneck segment.
 - Finally, easy-to-implement controllers are usually more intuitive than MPC controllers for the traffic community, which is not always familiar with automatic control, as the easy-to-implement only provides the thresholds at which the VSLs have to be changed.
- **Intermediate approaches:** Other approaches, such as QL-VSL (Li et al., 2017) and SPERT (Frejo and De Schutter, 2019), propose an intermediate solution allowing an easy implementation while approaching the performance of an optimal controller (computed offline). However, unfortunately, the off-line effort needed for the training of the network (in the case of QL-VSL), or for the computation of the optimal solution (in the case of SPERT) makes that, for large-scale networks, these algorithms are still difficult to implement in practice.

Local Versus Global VSL Controllers

Some studies have shown that centralized controllers for VSLs allow, in some cases, to obtain a much greater improvement on the traffic conditions than fully decentralized (i.e., local) controllers (Frejo and Camacho, 2012). However, as previously explained, a centralized and optimization-based traffic control algorithm is quite difficult to implement in a real (i.e., medium or large) traffic networks due to the exponential growth of the computation times.

As a consequence, the use of an algorithm that properly uses communication and cooperation between different controllers to approach the centralized behavior without critically increasing computational effort can substantially improve the performance of the controlled traffic system. Accordingly, a few distributed control algorithms have been proposing for VSLs using optimization-based techniques (Frejo and Camacho, 2011).

Discrete Versus Continuous VSL Controllers

Although in the real implementations of VSL panels the displayed signals are just allowed to take a limited set of discrete values, in many previous references the VSL signals are assumed to have continuous values. However, some recent works have shown that, in some cases, converting the continuous unconstrained VSL signal into an implementable VSL signal (i.e., a discrete value respecting the safety constraints) can substantially reduce performance with respect to the ideal case.

In order to deal with this problem, in Frejo et al. (2014), the discretization of the VSL signals is obtained by solving a discrete optimization problem with a search space limited to a certain distance from the continuous MPC solution. However, this discretization entails a new optimization increasing the computational load of the controller.

On the other hand, other controllers (like SPERT or LB-VSL) directly compute a discrete implementable value for the speed limits, avoiding the need of a subsequent discretization.

Summary of VSL Control Algorithms

The following table summarizes some characteristics of some of the most well-known VSL control algorithms for increasing freeway efficiency. The second and third column refers to the computational load required during each control step and the fourth one shows which algorithms provide a directly implementable discrete solution.

The fifth and sixth columns analyze how optimal and intuitive (for the traffic community) the solution provided by each controller is. The seventh column shows the kind of congestion that can be solved by each controller. Finally, the last column comments the robustness of each controller against uncertainties and/or communication or measurement errors.

Controller	Computational Load		Solution provided	Optimality	Intuitiveness	Kinds of congestion	Robustness
	Off-line	On-Line					
Optimal (computed offline)	High	Very high	Continuous	Optimal solution (for the given demands)	Low	All	Lack of robustness
MPC	High	High	Continuous	Optimal solution (within a receding horizon)	Low	All	Lack of robustness against communication failures
Distributed MPC	High	Medium	Continuous	Optimal solution (within a receding horizon)	Low	All	Lack of robustness (only at agent level)
Hybrid MPC	High	Medium	Discrete	Optimal solution (within a receding horizon)	Low	All	Lack of robustness against communication failures
SPERT	High	Very low	Discrete	Suboptimal (based on the optimal solution)	Medium	Recurrent congestion	Robust (around typical demand)
SPECIALIST	Low	Very low	Continuous	Not based on optimal solution	Medium	Shock waves	Robust (for shock waves)
Local MTFC	Low	Very low	Continuous	Not based on optimal solution	Medium-low	Active bottlenecks	Robust (for jams caused by bottlenecks)
QL-based	Very	high	Low	Discrete	Based on reinforcement learning	Low	Recurrent bottlenecks
Robust (if properly trained)							
LB-VSL	Low	Very low	Discrete	Not based on optimal solution	High	Active bottlenecks	Robust

Integration With Other Freeway Traffic Control Measures

The combined use of VSLs and other traffic control measures has been studied in many simulation studies:

- **Optimization-based algorithms for the integrated control of VSLs and ramp metering installations:** The most widely implemented and studied freeway traffic control measure is ramp metering, which limits the traffic flow of traffic entering freeways according to current traffic conditions. There are many research works that propose integrated controllers for ramp metering installations and VSLs. In most cases, these controllers use MPC techniques, entailing a higher computational complexity than the isolated control of the VSLs but a significantly better control performance. A few relevant examples can be found in Hegyi et al. (2004), Carlson et al. (2010), and Frejo and Camacho (2012).
- **Easy-to-implement algorithms for the integrated control of VSLs and ramp metering installations:** On the other hand, only three easy-to-implement controllers have been proposed for the integrated control of VSLs and ramp metering installations: MTFC + PI-ALINEA (Carlson et al., 2014), SPECIALIST-RM (Schelling et al., 2011), and LB-TFC. In all cases, these controllers are extensions of previously proposed easy-to-implement controller for VSLs.
- **Integration with other traffic control measures:** VSLs have also been integrated with lane change control, which generates lane change recommendations or restrictions upstream of an incident or bottleneck, in Zhang and Ioannou (2017). VSLs may be also integrated with other traffic control measures, such as reversible lanes, route guidance or high-occupancy lanes. Moreover, in future works, VSLs algorithm should be adapted for a context with a high penetration of autonomous vehicles.

Conclusions

VSLs have shown to be a powerful measure for traffic efficiency improvement in many simulation studies and a few field tests.

The most common way to model VSLs in the context of freeway traffic control is by modifying a macroscopic traffic flow model, like METANET. This modification reproduces the effects of a reduction on the value of a VSL by increasing the critical density and by decreasing the capacity and the free flow speed of the corresponding segment.

Many control techniques have been proposed for VSLs for traffic efficiency improvement, which can be mainly divided into optimization-based techniques (which can achieve the best performance but are difficult to implement in real freeways) and easy-to-implement techniques (which provide a suboptimal performance but are easy to implement in real applications).

Nowadays, the most important open challenge in the field of VSLs is the real application and testing of the control algorithms that have been proposed in the literature. In fact, so far only two control algorithms for traffic efficiency improvement (SPECIALIST and MTFC) have been tested in real experiments. As a consequence, there is still a significant lack of data that hinder the analysis of the potential effects of the implementation in practice of VSLs.

Acknowledgment

This research was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 702579 and under the ERC Advanced Grant agreement No 789051.

References

- Carlson, R.C., Papamichail, I., Papageorgiou, M., Messmer, A., 2010. Optimal motorway traffic flow control involving variable speed limits and ramp metering. *Transp. Sci.* 44 (2), 238–253.
- Carlson, R.C., Papamichail, I., Papageorgiou, M., 2011. Local feedback-based mainstream traffic flow control on motorways using variable speed limits. *IEEE Trans. Intell. Transp. Syst.* 12 (4), 1261–1276.
- Carlson, R.C., Papamichail, I., Papageorgiou, M., 2014. Integrated ramp metering and mainstream traffic flow control on freeways using variable speed limits. *Transp. Res. C Emerg. Technol.* 46, 209–221.
- Frejo, J.R.D., Camacho, E.F., 2011. Feasible Cooperation based model predictive control for freeway traffic systems. In: *Proceedings of 50th IEEE Conference on Decision and Control and European Control Conference*, pp. 5965–5970.
- Frejo, J.R.D., Camacho, E.F., 2012. Global versus local MPC algorithms in freeway traffic control with ramp metering and variable speed limits. *IEEE Trans. Intell. Transp. Syst.* 13 (4), 1556–1565.
- Frejo, J.R.D., Núñez, A., De Schutter, B., Camacho, E.F., 2014. Hybrid model predictive control for freeway traffic using discrete speed limit signals. *Transp. Res. C Emerg. Technol.* 46, 309–325.
- Frejo, J.R.D., de Schutter, B., 2018. A logic-based speed limit control algorithm for Variable Speed Limits to reduce traffic congestion at bottlenecks, 2018 IEEE Conference on Decision and Control (CDC), Miami Beach, FL, 198–204.
- Frejo, J.R.D., De Schutter, B., 2019. SPERT: A SPEed limit strategy for Recurrent Traffic jams. *IEEE Trans. Intell. Transp. Syst.* 20 (2), 692–703.
- Frejo, J.R.D., Papamichail, I., Papageorgiou, M., De Schutter, B., 2019. Macroscopic modeling of variable speed limits on freeways. *Transp. Res. C Emerg. Technol.* 100, 15–33.
- Han, Y., Hegyi, A., Yuan, Y., Hoogendoorn, S., 2017. Validation of an extended discrete first-order model with variable speed limits. *Transp. Res. C Emerg. Technol.* 83, 1–17.
- Hegyi, A., De Schutter, B., Hellendoorn, H., 2004. Optimal coordination of variable speed limits to suppress shock waves. *Transp. Res. Record* 1852, 167–174.
- Hegyi, A., Hoogendoorn, S.P., Schreuder, M., Stoelhorst, H., Viti, F., 2008. SPECIALIST: a dynamic speed limit control algorithm based on shock wave theory. *Proc. 11th Int. IEEE Conf. Intell. Transp. Syst.* 827–832.
- Li, Z., Liu, P., Xu, C., Duan, H., Wang, W., 2017. Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks. *IEEE Trans. Intell. Transp. Syst.* 16 (1), 3204–3217, 2017.

- Papamichail, I., Kampitaki, K., Papageorgiou, M., Messmer, A., 2008. Integrated ramp metering and variable speed limit control of motorway traffic flow. Seventeenth IFAC World Congress 41 (2), 14084–14089.
- Pasquale, C., Anghinolfi, D., Saccone, S., Siri, S., Papageorgiou, M., 2016. A comparative analysis of solution algorithms for nonlinear freeway traffic control problems. Proc. 19th IEEE Conf. Intell. Transp. Syst. 1773–1778.
- Schelling, I., Hegyi, A., Hoogendoorn, S.P., 2011. SPECIALIST-RM - Integrated variable speed limit control and ramp metering based on shock wave theory. Proc. 14th IEEE Conf. Intell. Transp. Syst. 2154–2159.
- Zackor, H., 1972. Beurteilung verkehrsabhängiger Geschwindigkeitsbeschränkungen auf Autobahnen. Strassenbau Strassenverkehrstechnik 128, 1–61.
- Zhang, Y., Ioannou, P., 2017. Combined variable speed limit and lane change control for highway traffic. IEEE Trans. Intell. Transp. Syst. 18 (7), 1812–1823.

Hard Shoulder Running

Justin Geistefeldt, Institute for Traffic Engineering and Management, Ruhr University Bochum, Bochum, Germany

© 2021 Elsevier Ltd. All rights reserved.

Definition	41
Applications	41
Impacts on Traffic Flow and Road Safety	43
Conclusions	44
References	44

Definition

Hard shoulder running is an active traffic management (ATM) strategy that incorporates the use of the hard shoulder on freeways (motorways) as a travel lane during peak hours. It is primarily used on freeway segments or corridors affected by recurring congestion due to a lack of carriageway capacity during peak hours. Hard shoulder running is implemented either as an interim solution if widening of the carriageway cannot be realized within short periods of time or as a permanent solution if alternatives to improve traffic operation are infeasible. In contrast to permanently converting the hard shoulder into a travel lane, hard shoulder running provides the additional capacity only when needed to avoid congestion, while the hard shoulder can still be used as emergency lane or for road maintenance during off-peak traffic conditions.

Hard shoulder running is also referred to as “part-time shoulder use”, “temporary hard shoulder use”, “dynamic hard shoulder use”, and a number of similar expressions. Hard shoulder running schemes can be distinguished between

- dynamic (traffic-responsive) or static (fixed-time) control of opening and closing the hard shoulder for travel,
- use of the outer or inner shoulder,
- shoulder use between adjacent interchanges, where the shoulder acts as an auxiliary lane between an entrance ramp and the next exit ramp, or continuous shoulder use on a corridor over several interchanges, and
- opening the shoulder for all types of vehicles or for specific vehicle types like buses (bus-on-shoulder) or passenger cars only.

Today, the temporary use of the outer shoulder (i.e., the right shoulder when driving on the right) for all types of vehicles with traffic-responsive control is the most common type of hard shoulder running applications worldwide. Hard shoulder running typically involves the use of traffic monitoring and control systems with gantries that include speed and lane control signs as well as video surveillance of the hard shoulder. To ensure that traffic safety is not affected by the loss of the hard shoulder as emergency lane, the speed limit that is applied during the times when the hard shoulder is in operation is usually lower than during closed hard shoulder. Implementing hard shoulder running often requires a reallocation of the carriageway cross section in order to provide a sufficient width of the shoulder that is required to be safely used by running traffic, particularly trucks. Emergency refuge areas are also added to the carriageway.

Applications

While static part-time shoulder use has been applied in the United States since the 1970s, the first dynamic (traffic-responsive) hard shoulder running schemes were implemented in Germany and the Netherlands in the late 1990s, followed by a number of segments also in other European countries in the subsequent years. Based on the gained experience, deployment guidelines for hard shoulder running in Europe were published by [EasyWay \(2012\)](#).

In Germany, the first control systems for dynamic hard shoulder running were installed on freeway (Autobahn) A4 between Refrath and Cologne-Meerheim in 1996. Subsequently, major implementations of hard shoulder running on longer freeway corridors include the A3 between Offenbach and Hanau, the A5 between Frankfurt North-West interchange and Friedberg ([Figure 1](#)), and the A99 between Munich North interchange and Haar, all put into operation between 2001 and 2003. Today, control systems for dynamic hard shoulder running are being operated on more than 250 km of directional carriageways. While several plans for further applications of hard shoulder running on other freeway segments exist, a few former applications were already put out of operation due to the realization of freeway widening. During hard shoulder running, the maximum permitted speed on freeways is 100 km/h. On most sections, hard shoulder running is combined with line control systems including LED signs for displaying variable speed limits as well as road warnings in case of incidents.

In the Netherlands, hard shoulder running has been implemented on freeways since 1996, starting with the A28 near Utrecht. Today, almost 300 km of hard shoulders and so-called “plus lanes” can be dynamically opened for running traffic during peak hours ([CEDR, 2012](#)). “Plus lanes” are dynamic lanes on the left side of the carriageway, which are usually 2.75 m wide and hence narrower



Figure 1 Hard shoulder running on Autobahn A 5 north of Frankfurt, Germany



Figure 2 “Plus lane” on freeway A 12 west of Utrecht, The Netherlands

than normal travel lanes (Figure 2). Both hard shoulder running lanes on the right side of freeway carriageways and “plus lanes” on the left are dynamically opened during peak hours in combination with a speed limit of 100 km/h.

In England, the first hard shoulder running scheme was opened on motorway M42 in 2006. Based on the positive experience with this pilot application concerning its impacts on traffic flow and road safety, further hard shoulder running applications were put into operation on a number of motorways near Birmingham between 2009 and 2011 (Van Vuren et al., 2012). Hard shoulder running incorporates the installation of emergency refuge areas for broken-down vehicles as well as signals and variable message signs on gantries. The maximum speed is restricted to 60 mph when the hard shoulder is in operation. In recent years, hard shoulder running schemes are increasingly replaced by an “all lane running” strategy, which means that the hard shoulder is converted into a normal lane that can be closed by lane control signs in the event of an incident.

In France, hard shoulder running was implemented on the joint section of the freeways A4 and A86 in the east of Paris. A bus-on-shoulder scheme has been applied on freeway A48 north of Grenoble (Princeton and Cohen, 2011). In Flanders, Belgium, hard shoulder running lanes are being operated on freeways E313/E34 and E19 east of Antwerp as well as on freeway E40 from Brussels to Leuven. Further applications of dynamic hard shoulder running in Europe include the Hillerød Motorway M13 north of Copenhagen, Denmark, and the freeway A1 between Morges and Ecublens in Switzerland freeway A1 between Morges and Ecublens in and (CEDR, 2012).

The overall success of traffic-responsive hard shoulder running implementations particularly in Europe triggered their introduction in the United States since 2009, when dynamic hard shoulder use was implemented on Interstate I-35W in Minneapolis, Minnesota. Together with a large number of previous applications of bus-on-shoulder use and static shoulder use schemes, which are still most common, hard shoulder running is applied in more than 16 states. The applications of hard shoulder running in the United States cover quite a variety of different strategies and control systems, including hard shoulder use for buses, passenger cars, and all vehicle types, applications on freeways and arterials, as well as different speed limits and usage criteria. A comprehensive guideline on planning, design, implementation, and day-to-day operation of hard shoulder running was published by the FHWA (2016). This guideline also includes a compilation of existing hard shoulder running applications in the United States.

In South Korea, hard shoulder running lanes were implemented on seven freeway corridors between 2009 and 2013 (Choi et al., 2019). On these corridors, line control systems are used for the dynamic control of the hard shoulder and for ensuring a high level of traffic safety.

Impacts on Traffic Flow and Road Safety

The main objective of implementing hard shoulder running is to provide additional capacity during peak hours, resulting in less congestion and higher travel time reliability. The significant increase of carriageway capacity that is achieved by using the hard shoulder as an additional travel lane often leads to a high benefit-cost ratio of hard shoulder running projects. However, hard shoulder running can only be effective and economically efficient if congestion results from insufficient capacity of the carriageway itself and not from downstream bottlenecks like overloaded off-ramps. Hence, a detailed traffic analysis is useful when considering the implementation of hard shoulder running.

The major challenge of operating hard shoulder running is to maintain traffic safety despite suspending the hard shoulder as an emergency refuge area during peak hours. A high level of traffic safety is usually achieved by sophisticated traffic monitoring and control systems together with a low speed limit during hard shoulder running. The traffic monitoring systems typically include video surveillance of the hard shoulder to check whether the hard shoulder is free of obstacles and broken down vehicles and to be able to close the hard shoulder immediately in case of an incident on the hard shoulder. A number of studies have reported findings on the impact of hard shoulder running on traffic flow and road safety. A selection of relevant publications is summarized in the following.

[Geistefeldt \(2012\)](#) documented experience with dynamic hard shoulder running on two major freeways in the Federal State of Hesse, Germany. Based on loop detector data, it was shown that the capacity of three-lane carriageways can be increased by 20%–25 % if the hard shoulder is used as an additional travel lane during peak hours. The acceptance of hard shoulder running was found to be considerably high among truck drivers. Supporting previous findings, negative effects on road safety could not be determined. By limiting hard shoulder running to those times when the additional capacity of the shoulder lane is needed to prevent congestion together with continuous monitoring of the hard shoulder, the risk for broken down vehicles to be involved in a rear-end collision accident can effectively be reduced due to the fact that any car or obstacle that blocks the shoulder lane immediately causes a traffic breakdown involving a reduced speed level at the incident location. Hence, it was stated that the application of traffic-responsive hard shoulder running combined with dynamic traffic control systems is superior to the permanent transformation of the hard shoulder into an additional travel lane.

[Van Vuren et al. \(2012\)](#) examined the effects of managed motorway schemes around Birmingham considering earlier investigations of the first trialed managed motorway M42 in England. The empirical analysis was based on loop detector data, data from variable message signs, and incident reports from a monitoring period of five years. The comparison of the before and after cases showed that the managed motorway schemes involving hard shoulder running had significant positive effects on the average travel times and their variability as well as increasing traffic throughput. User consultations revealed a positive perception of road users in relation to driver comfort, journey time, and safety. A microscopic traffic simulation of the managed motorway scheme showed that hard shoulder running is particularly beneficial where large traffic volumes exit the motorway. In addition to conventional travel time benefits, an additional 15% of wider economic benefits were estimated, mainly resulting from commuting, business, and commercial trip purposes.

[Samoili et al. \(2013\)](#) analyzed the lane flow distribution with hard shoulder running based on traffic data from radar sensors of a four-lane freeway segment in Switzerland. The results of the lane choice behavior analysis were introduced into a short-term prediction model to describe traffic flow on the section with hard shoulder running. It was shown that opening the hard shoulder leads to an abrupt interruption of the lane flow equilibrium with an immediate transfer of 20% of the right-lane volume to the hard shoulder. An 11% capacity increase and maintenance of speed at 100 km/h were determined as positive effects of the system.

[Kononov et al. \(2012\)](#) examined the relationship between traffic flow and safety parameters to explain the effect of hard shoulder running on traffic safety. The empirical analysis of safety data from four-lane freeways in Denver, Colorado, over a 5-year period revealed that as flow increases, the crash rate initially remains constant until a certain critical threshold of flow parameters is reached. When exceeding the threshold, the crash rate rapidly increases, caused by an increase in density without a notable reduction in speed and resulting small headways that make it more difficult for drivers to compensate driving errors. Regarding the empirical results, it is supposed that crash rates decline during hard shoulder running because of the reduction of traffic density. It is concluded that a positive effect on traffic safety by hard shoulder running can only be reached when density can be reduced under a certain threshold.

[Aron et al. \(2013\)](#) investigated the traffic flow and safety effects of hard shoulder running on freeway A4/A86 in the east of Paris, France. Due to the insufficient capacity of this eight-lane segment carrying an AADT of 280,000 veh/d, hard shoulder running was implemented in 2006. When open to traffic, the hard shoulder operates as an auxiliary lane between the adjacent interchanges. A comparison of accident data from before and after periods revealed a slight, however statistically insignificant, reduction of the number of accidents by 3% during hard shoulder operation, which was attributed to the reduction of traffic density.

[Ma et al. \(2016\)](#) proposed simulation based incident management strategies for freeway segments with dynamic hard shoulder running. The strategies describe the shoulder use options outside peak hours to minimize the traffic flow effects of nonrecurring incidents. To fully use the potential of shoulder running during incidents, an open shoulder section length of 0.5 mile upstream and downstream of the incident location is sufficient. Furthermore, the shoulder can be closed as soon as possible after the incident is cleared, as no improvement in traffic conditions could be determined by a longer opening time. Applying these strategies, the average delay can be reduced by 30%–80% and the total traffic throughput can be increased by 15%–40%, depending on the traffic conditions and the number of blocked lanes.

[Guerrieri and Mauro \(2016\)](#) forecasted traffic flow and safety effects expected by implementing a hard shoulder running system on a 128 km section of freeway A22 in Italy. Safety effects were estimated by means of the Highway Safety Manual methodology ([AASHTO, 2010](#)) based on parameters derived from traffic data. To calibrate the model, real crash data from 3 years were evaluated

and traffic and operation conditions were assumed. According to the model results for motorway A22, no significant variation in the general safety conditions could be determined in the presence of a considerable capacity increase of 35%.

Chun and Fontaine (2017) analyzed the effects of the active traffic management (ATM) system characterized by advisory variable speed limits, queue warning systems, lane use control signs, and hard shoulder running on the Interstate I-66 in Northern Virginia. The authors compared the effects of static hard shoulder running in predefined peak hours before the implementation of the ATM system and dynamic hard shoulder running, which can be applied whenever an increase in roadway capacity was warranted. The investigated ATM system with dynamic hard shoulder running was found to significantly improve the mean travel time and travel time reliability during weekday off-peak periods and on weekends. However, an improvement of travel times in peak periods by the ATM system could not be determined.

Choi et al. (2019) analyzed the safety effects of hard shoulder running on freeways in South Korea. Comparing crash data from 4 years of 22 freeway sections with hard shoulder running, it was found that crash frequencies increased after the implementation of hard shoulder running. Regarding the crash severity, a higher proportion of serious crashes with personal damage was found for two-lane segments compared to three-lane segments. Applying a regression analysis, a positive relationship between the length of the hard shoulder running segment and the crash frequency was determined. However, a statistical relationship between the lane width or shoulder width and the crash frequency could not be verified. Based on the results, it was recommended that hard shoulder running should be avoided on long two-lane freeway corridors.

Conclusions

Hard shoulder running is an effective active traffic management measure to increase the capacity of frequently congested freeway segments by opening the hard shoulder for travel during peak hours. Hard shoulder running is mainly applied in the United States and central Europe. While the first applications in the United States were rather low-tech solutions with static control, including a number of bus-on-shoulder schemes, modern applications of hard shoulder running involve the use of sophisticated traffic monitoring and control systems. Safe operation is provided by continuously monitoring the shoulder and limiting shoulder use to times when the additional capacity is needed to avoid congestion, while the shoulder is still available as emergency lane at off-peak times. A number of scientific studies has proven the positive impact of dynamic hard shoulder running on traffic flow quality, whereas traffic safety is not significantly affected in most cases and can even be improved by the lower traffic density and the reduced extent of upstream congestion.

References

- AASHTO, 2010. Highway Safety Manual (HSM). American Association of State Highway Transportation Officials.
- Aron, M., Seidowsky, R., Cohen, S., 2013. Safety impact of using the hard shoulder during congested traffic. The case of a managed lane operation on a French urban motorway. *Transp. Res. Part C* 28, 168–180.
- CEDR, 2012. Traffic management to reduce congestion. Conference of European Directors of Roads.
- Choi, J., Tay, R., Kim, S., Jeong, S., Kim, J., Heo, T.-Y., 2019. Safety effects of freeway hard shoulder running. *Appl. Sci.* 9, 3614.
- Chun, P., Fontaine, M.D., 2017. Evaluation of operational effects of I-66 active traffic management system. *Transp. Res. Rec.* 2616, 91–103.
- EasyWay, 2012. Traffic Management Services: Hard Shoulder Running. Deployment Guideline. European Union, Trans-European Transport Network, Brussels.
- FHWA, 2016. Use of Freeway Shoulders for Travel – Guide for Planning, Evaluating, and Designing Part-Time Shoulder Use as a Traffic Management Strategy. U.S. Department of Transportation, Federal Highway Administration, Washington D.C.
- Geistefeldt, J., 2012. Operational experience with temporary hard shoulder running in Germany. *Transp. Res. Rec.* 2278, 66–73.
- Guerrieri, M., Mauro, R., 2016. Capacity and safety analysis of hard-shoulder running (HSR): A motorway case study. *Transp. Res. A* 92, 162–183.
- Kononov, J., Hersey, S., Reeves, D., Allery, B.K., 2012. Relationship between freeway flow parameters and safety and its implications for hard shoulder running. *Transp. Res. Rec.* 2280, 10–17.
- Ma, J., Hu, J., Hale, D.K., Bared, J., 2016. Dynamic hard shoulder running for traffic incident management. *Transp. Res. Rec.* 2554, 120–128.
- Princeton, J., Cohen, S., 2011. Impact of a dedicated lane on the capacity and the level of service of an urban motorway. *Proc. Soc. Behav. Sci.* 16, 196–206.
- Samoli, S., Elthymiou, D., Antoniou, C., Dumont, A.-G., 2013. Investigation of lane flow distribution on hard shoulder running freeways. *Transp. Res. Rec.* 2396, 133–142.
- Van Vuren, T., Baker, J., Ogawa, J., Cooke, D., Unwin, P., 2012. Managed motorways: Modeling and monitoring their effectiveness. *Transp. Res. Rec.* 2278, 85–94.

High-Occupancy Vehicle (HOV) and High-Occupancy Toll (HOT) Lanes

Roxana J. Javid*, Jiani Xie*, Lijiao Wang*, Wenruifan Yang*, Ramina Jahanbakhsh Javid†, Mahmoud Salari‡, *University of Southern California, Department of Civil and Environmental Engineering, Los Angeles, CA, United States; †Shahid Beheshti University, Department of Urban Planning & Design, Tehran, Iran; ‡California State University Dominguez Hills, Department of Accounting, Finance, and Economics, Carson, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	45
Objective of HOV Lane	45
Different Types of HOV Lanes	46
Past and Current Practices	46
Planning	48
Design	48
Implementation	48
Overhead and Roadside Signs	48
Pavement Marking	49
Performance Monitoring	49
Benefits and Costs	50
Enforcement	50
Acknowledgment	50
Relevant Websites	50
References	51
Further Reading	51

Introduction

High-occupancy vehicle (HOV) lanes, which have been introduced as an example of managed lane facilities, are reserved for vehicles carrying more than two people, including the driver and are designed and operated to improve travel conditions. The access is sometimes open to motorcycles, transit buses, hybrid, or electric cars regardless of the number of passengers. Commonly referred to as carpool lanes, HOV lanes are usually located on the inside (left) lanes on highways and are identified by signs along the road and painted diamond symbols on the pavement.

The restriction on HOV lanes is varied depending on the amount of traffic and commuter patterns, from 24 hours a day to limited peak congestion periods. When the capacity of the HOV lane is excessive, the lane may be changed to a high-occupancy toll (HOT) lane. The HOT lane can be utilized by vehicles carrying one person for a fee, which may vary depending on the time of day or the level of congestion to maximize the congestion mitigation effect of the HOT lane facility. A HOT lane can be used while the HOV lane is in operation but can also be restricted to use during peak congestion periods. **Figure 1** shows an example of HOV/HOT lane.

Objective of HOV Lane

The main concept of the HOV lane is to facilitate the transition from solo driver vehicle travel to ridesharing, and to eliminate congestion by providing two benefits to the driver: saving time and ensuring travel time (FHWA, 2016). Since HOV lanes carry vehicles with more occupants, more people can be moved during periods of congestion, even if the number of vehicles using the HOV lanes is less than adjacent general-purpose lanes. The concept of HOV lanes gradually developed in the 1970s in the United States and is now being introduced in many metropolitan areas as an effective means of reducing congestion.

HOV lanes also strive to increase the accessibility and reliability for drivers. HOV lanes change our understanding of mode choice behavior where public transit and self-driving are our original mode choices for short-haul land transportation. Public transit generally has low level of accessibility as well as reliability in arrival time. Parking is usually an important issue for self-driving. Parking availability in an acceptable time frame and distance from the destination is uncertain. Ridesharing and HOV lanes deal with these issues with relatively less travel time variability (Javid and Jahanbakhsh Javid, 2018) and need for parking.

Another objective of HOV lanes is to develop a sustainable transportation system, which supports the infrastructure, environment, and public health. HOV lanes are intended to increase the person carrying capacity of highways by encouraging ridesharing and transit ridership. This means a lower level of negative impact from traffic load to bridges and roads as well as less negative health impacts from air pollutants and greenhouse gases (GHGs). Emissions of various toxic gases and GHGs not only influence drivers involved in the congestion, but also impact nearby communities and the planet (Javid et al., 2019).



Figure 1 Example of HOV/HOT lane. Source: <https://ops.fhwa.dot.gov/publications> (FHWA, 2017b)

Table 1 Different classifications of HOV lanes

Classification basis	Types	Description
Traffic flow	Concurrent flow lane	The lane is in the same direction of the general operation lanes, which might be separated or not from the other freeway lanes and typically located in the inside lane or shoulder.
	Contraflow lane	The lane is designated for exclusive use by HOVs users traveling in the peak direction and separated from the off-peak (or opposite) flow by movable barriers.
Separation	Separated lane	The lane is separated from general operation lanes by physical barriers and can be either a one-way reversible or a two-way operation.
	Buffer-separated lane	The lane is not separated with general-purpose lanes by any physical barriers, but set a buffer zone with other lane by using marking on the pavement.
	Nonseparated lane	The lanes operate in the same direction and immediately adjacent to the general-purpose lanes (to the left or right). They are normally less expensive and provide easy access and more conflicts.
Operation condition	Bus lane	Busway lane is usually located on the right side lane so passengers can get on/off the bus safely. Bus Rapid Transit (BRT) lane is set on the left side lane with specific BRT station to reach a very high speed.
	HOT lane	HOT lanes are free to high-occupancy vehicles but single drivers can use them with a fee, which is variable by traffic demand.
	2+ or 3+ Lane	2+ HOV lane can be used by vehicle that have at least 2 occupants (including driver). While, 3+ HOV lane can be used by vehicle with 3 or more occupants.

The final objective of HOV lanes is to save the transportation economy. HOV lanes can encourage ridesharing, which in turn can benefit the transportation system by saving space required for parking and amount of fossil fuel used. Moreover, users who choose to share the ride, can split the fuel, vehicle maintenance, and parking costs.

Different Types of HOV Lanes

Different types of HOV lanes are developed, which can be categorized by traffic flow, lane separation, and operation condition. **Table 1** below summarizes the types of HOV lanes based on different classification basis.

Past and Current Practices

The first HOV lane was introduced in 1969 as a bus-only lane on the Shirley Highway (I-395) from Northern Virginia to Washington D.C. Closely following the bus-only lane in Northern Virginia and Washington D.C. was the bus-only HOV lane in New York-New Jersey's Lincoln Tunnel in 1970. In 1973 the bus-only lane expanded to include any vehicle with four plus (4 +) people, welcoming

Table 2 Examples of HOV/HOT lane facilities in the United States

Example facility	Specification	Explanation
Metro HOV system (Los Angeles, CA)	Concurrent flow lane, Full-time operation, mostly 2+ lanes	A total of 513 miles (823 km) of completed and 132.2 miles (213 km) of planned HOV lanes, make it the largest HOV system in the United States (Caltrans, 2009).
State Route 85 (San Francisco, CA)	Concurrent flow lane, Part-time operation, 2+ lanes	HOV lanes are available only during peak hours Monday through Friday 5 AM to 9 AM, and 3 PM to 7 PM (Metropolitan Transportation Commission, 2020).
I-278 Gowanus Expressway (New York City, NY)	Contraflow flow lane with movable barrier, Part-time operation, 3+	HOV lanes run from the Verrazano Bridge to the Brooklyn Battery Tunnel on the Gowanus Expressway for buses and vehicles weekdays from 6 AM to 10 AM (New York City, 2020).
I-394 Corridor (Minneapolis, MN)	Concurrent flow lane, HOV/HOT lane, combination of movable barrier and buffer-separated, part-time operation, 2+	HOV/HOT pricing varies depending upon the volume and occupancy values of the HOV/HOT lane. One third of the HOV/HOT lane is a reversible lane (offering 2 lanes of travel in the selected direction) operation eastbound 5:00 AM – 1:00 PM and westbound 2:00 PM – 4:00 AM. The rest of the facility is one lane buffer-separated operating eastbound 6:00 AM to 10:00 AM and westbound 2:00 PM to 7:00 PM (Liu et al., 2011).

vans and station wagons (Turnbull, 2005). Vehicle-occupancy requirements for carpools have changed over time and now most facilities use a two-person per vehicle (2 +) carpool designation.

The development of HOV lanes can be summarized as the two stages: The first stage from 1969 to 1990 referred to as the germination period of carpooling lanes, when the construction of HOV lanes dominated highway projects in the 1970s and 1980s lead to over 350 existing HOV facilities in the United States (Turnbull, 2005).

The second stage from 1991 to 2000 is the growth period, this stage involved the cooperation with the environmental protection department, transportation department, and legislation department to formulate a series of transportation measures that help reduce air pollution, such as HOV lanes. California implemented the first HOT lane in 1995 along the US 91 in Orange and Riverside County. HOT lanes were introduced after HOV lanes and are used in a way that employs market forces to improve the use of HOV lanes and collect tolls that can supplement the operations, enforcement, and maintenance costs of the facilities (Turnbull, 2005).

Today, HOV lanes remain the most prevalent form of managed lane in the United States, with lane-miles in service estimated at over 3300 miles (5311 km) (FHWA, 2020a). Today, there are over 350 existing HOV facilities in the 31 metropolitan areas across 20 states. California contains the most HOV facilities at 88, followed by Minnesota and Washington state with 83 and 43, respectively (as of 2009) (Chang et al., 2008). As of 2004 under the Clean Air Vehicle program, the California Department of Motor Vehicles, in partnership with the California Air Resources Board, allows single-occupied vehicles that meet specific emissions standards to use HOV lanes (California Department of Motor Vehicle, 2020). Thirteen other states have similar HOV and HOT lane facility exemptions for vehicles that meet specific emission standards (Alternative Fuel Data Center, 2016). Motorcycles are now also allowed in HOV lanes by federal law because it was safer for cyclists rather than traveling in regular lanes where there are typically stop-and-start traffic conditions, especially during peak travel times. As of 2017, there are 10 operating HOT facilities in either one- or two-lanes in each direction totaling over 100 miles (161 km) in California, Texas, Minnesota, Colorado, Utah, Washington, Florida, and Georgia (FHWA, 2017a). Many states have HOT facility projects in the planning stages as a response to increased congestion in communities. All of the operating projects are conversions of HOV lanes to HOT lanes, although some have extended the HOT lanes. Table 2 below shows some examples of HOV/HOT lane facilities in the United States.

HOV lanes are currently applied in several countries all over the world to solve the traffic problems and provide larger traffic capacity and higher efficiency. Canada has a total of 23 HOV sites in form of 150 km (93 mi) of highway HOV and over 130 km (81 mi) of arterial HOV lanes, in four different provinces. These sites are typically are 3+ HOV lanes, in operation during peak periods (Transport Canada, 2010). The first HOV lane was opened in Greater Vancouver and Toronto in early 1990s (Dixon and Alexander, 2005).

HOV lane was first introduced in Europe in the Netherlands in October 1993. The facility was a 7 km (4.3 mi) 3+ HOV lane with removable barrier and after a year due to low demand was converted to a reversible lane open to general traffic (Schijns and Eng, 2006). A median reversible HOV lane was opened in Spain in 1995 and is still in operation now. Austria opened some of the bus lanes to high occupancy vehicles with at least three occupants in 1998 (Berger, 2002). The United Kingdom opened the first HOV lane facility in Leeds in 1998 as an experimental project and it became permanent later. In 2001 the first ever HOV lane facility in Norway was introduced in the city of Trondheim. This short HOV lane [0.85 km (0.5 mi)] was available for buses, taxis, and vehicles with more than two occupants 24 hours a day (Sandelien, 2000). In New Zealand, HOV lanes are also known as priority lanes. There are a few types of priority lanes usually dedicated to vehicles that have more carriage such as high-occupancy vehicles and trucks.

In 1992, Indonesia implemented the first 3+ HOV lane in Jakarta, known as three-in-one lane, which is from the three-in-one policy (at least three occupants in one vehicle). In South Korea, HOV lanes were introduced in 1994. Only private vehicles carrying

more than six passengers can be access to HOV lanes. Buses also are allowed to use HOV lanes (Kwon et al., 2005). Australia opened their first HOV lane, also known as transit lane in February 1992 on the Eastern Freeway in Melbourne. There are two types of transit lanes called T2 and T3, corresponded to 2+ HOV lanes and 3+ HOV lanes respectively. In China's transportation HOV lane is a new concept. In 2014 Wuxi City, Jiangsu Province offered HOV lanes on Xinyuan Road. The vehicles that have 3 or more occupants can drive on this lane. Then Jinan city followed up and implemented HOV lanes in 2015. In Shenzhen, HOV 2+ has been implemented on Binhai Avenue in 2016. Small vehicles must carry at least two occupants to be qualified to use this lane. This HOV lane can be just operated on weekdays' peak commute time, which is from 7:30 AM to 9:30 AM, and from 5:30 PM to 7:30 PM in China. After 4 months of operation, the Shenzhen government announced a policy of punishment that the vehicles violating the lane will be fined. Based on the development of HOV lanes all over the world, HOV lanes are most commonly used in the United States and it has the most professional and systematic plan for HOV lanes.

Planning

HOV lanes should allow vehicles to travel at high speeds without the HOV facility being perceived by the public as congested or underutilized. To achieve this, planning criteria differ by state and city, with some states varying entry or operating rules by time of day and others operate their HOV facilities 24 hours a day. In terms of who and what vehicles are allowed to enter HOV facilities during operating hours, motorcycles, mass transit, and vehicles with either two (2 +) or three (3 +) occupants are usually allowed to access the HOV lanes. A higher requirement for number of occupants will result in more ridesharing including van sharing and corporate carpooling. If the regional policy makers aim to mitigate emissions and fuel consumption from road transportation certain alternative fuel vehicles can be exempted from the occupancy requirement (Caltrans, 2020). The type and number of lanes allocated to HOVs are also play an important role in their impacts on traffic. The two-lane HOV/HOT facility eliminates low-speed vehicles and reduces traffic on the road. On the ramp, HOV/HOT vehicles are allowed to pass first, thereby improving efficiency. Finally, in order to convert an HOV lane to HOT lane, the level of service, costs, and economic benefits of HOT lanes should be assessed and planned on a case-by-case basis.

Design

HOV lanes can be a part of new highway construction, whole or some parts of highway rebuilding. Generally speaking, HOV lanes, whether one or more, play a role as interior lanes located on the middle of freeway (Cothron et al., 2004).

The HOV lane design process consist of HOV lanes and its auxiliary facilities' design, including HOV direct connector, HOV ramps and local obstructions. Generally, the design considerations need to meet basic principles for highway road design manuals, involving horizontal stopping sight distance, decision stopping sight distance, and vertical clearance. For other requirements, it contains drainage, structural section, and lane width. It still has other options which must be evaluated on a case-by-case basis to satisfy safety, efficiency, and comfort.

Implementation

Public agencies, such as State Departments of Transportation and transit agencies, oversee construction and operation of HOV facilities, often with federal funding support (FHWA, 2020a). Standardization of signboard symbols, pavement marking, and design standards of HOV facilities is essential for proper operation. Due to factors such as multiculturalism and illiteracy, graphic signage is required as much as possible, not letter-based signage, and the internationally recognized symbol for HOV, a slender diamond, is used. If the HOV lane is not sufficiently distinguished from the mixed-flow lane, it may lead to misuse of the lane by an ineligible vehicle. Failure to do so will result in a loss of respect and understanding of the principles of HOV lanes and the potential benefits associated with them, as well as the ongoing issue of violations and enforcement. Therefore, it is very important to distinguish HOV lanes clearly by pavement signs and overhead and roadside signs (Rankin, 1994).

Overhead and Roadside Signs

HOV lane signs should generally follow the federal standards (FHWA, 2009). HOV lane signs should provide the drivers with the entrance location, occupancy requirement, business hours, and cost (for the HOT lane). In addition, the driver needs to be given sufficient time to decide whether to use a controlled lane facility and have safe access to the facility.

For HOT lanes with variable or dynamic pricing, the operator must place at least one variable message sign before every entry point in the managed lane to display fee information so that the user has enough time to decide whether to use the HOT lane (FHWA, 2017a). **Figure 2** provides some examples of HOV lane overhead signs.

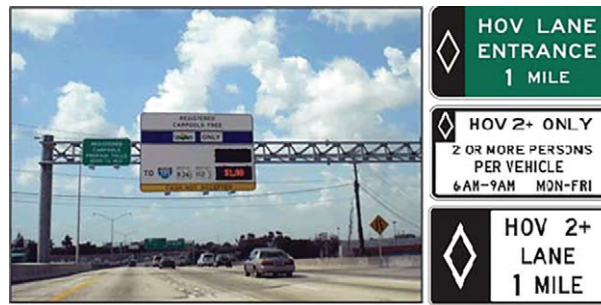


Figure 2 HOV lane overhead signs. Source: <https://mutcd.fhwa.dot.gov> (FHWA, 2020b)

HOV Lane Markings

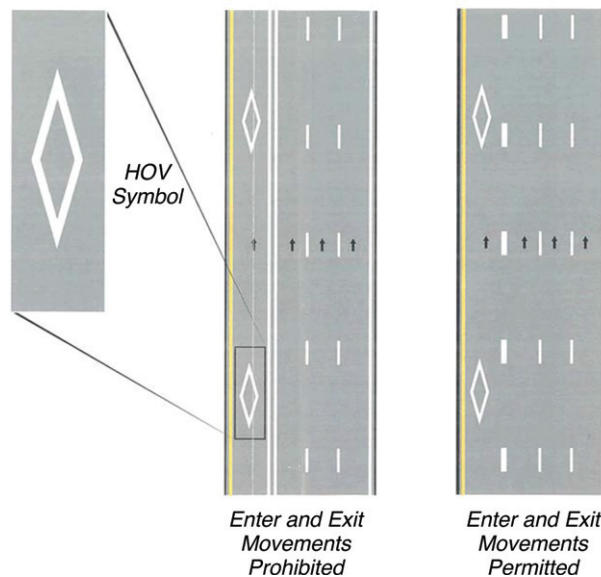


Figure 3 HOV diamond symbol and marking. Source from: <https://mutcd.fhwa.dot.gov> (FHWA, 2009)

Pavement Marking

It is common to draw HOV lanes on pavement marking highways with two parallel white dashed lines to highlight their presence and allow vehicles to join or leave the lanes with dashed lines marking. In addition to characteristic striping, the HOV diamond symbol or the word message HOV is usually marked on the pavement of the lane. **Figure 3** shows the symbol and marking for a regular HOV lane facility. Due to poor visibility from snow and limited durability, the operator cannot rely solely on pavement markings to provide information on HOV lanes. Sometimes, to emphasize the HOV message and alert the lane users, the whole lane is displayed in a different color.

Performance Monitoring

The State DOT in which the HOV facility operates is responsible for monitoring its performance and writing their own procedure and methodology for determining if the operational performance of an HOV facility is degraded (FHWA, 2016).

Programs are established to determine if HOV facilities are meeting its objectives and to monitor and evaluate the performance of HOV facilities. The results of the performance evaluation provide the basis for revising the HOV system operation in order to make improvements. Evaluating HOV facilities involves examining safety, vehicle volumes, level of service, impacts on the movement of people (i.e., how many people use the lane), modal shifts (i.e., how many people alter their travel behavior to use the HOV lane), reliability, travel-time savings, and the impact of the ecological environment (FHWA, 2016). There are minimum performance measures, such as average operating speed that must be met by each HOV facility.

Benefits and Costs

The way how HOV lanes restriction reduces congestion is to encourage carpooling and increase the people-moving capacity of the freeway system. HOV lanes can also increase the congestion on other non-HOV lanes, which may accumulate the overall delay on the road and make the time for congestion disappearance longer. The research on the effectiveness of HOV lanes towards reducing congestion should be positive but sometimes could also be limited and it really depends on the specific traffic condition.

The environmental benefits of HOV lane facilities depend on travelers' demand and their travel patterns, and vary significantly across counties and states according to population density (Javid et al., 2016). A recent research (Javid et al., 2017) found that the greatest projected reduction of air pollutant emissions from HOV lanes development are within areas of relatively high population density at the United States, this would mean that introducing HOV facilities in order to reduce emissions would be most effective in California and east coast states, and least effective in northwest Great Plains.

In addition to the advantages of HOV lanes, the advantage of HOT lanes over HOV lanes is that they allow single-passenger cars to pay for use, and revenue can be generated while increasing HOV lane utilization. With the conversion of HOV to HOT lanes, the number of vehicles on general purpose lanes is reduced and they will be safer with the improved driving environment (Cao et al., 2012). It seems that HOT lanes can be widely promoted and used. However, this is unreasonable because the HOT lanes allow single-passenger use, which is contrary to the original intention of the HOV lanes to promote ridesharing. Therefore, in order to fully and rationally use the HOV and HOT lanes, it is necessary to find a balance between carpooling and congestion pricing (Konishi and Mun, 2010).

The average cost of adding lanes in urban areas, depending on where they are planned, is about \$10 million per lane-mile (\$16 million per lane-km) (FHWA, 2020a). Construction, design engineering and admin, and capital cost for operations are major HOV/HOT related costs. Unlike HOV lanes, HOT lanes require the installation of multiple access points because of the exchange of tolls, and a large amount of money is required for communication facilities (FHWA, 2007). To justify these additional costs for the construction and operation of HOV lanes, especially HOT lanes, the delay saving and toll revenues must be sufficient.

Enforcement

Public awareness and enforcement are key in maintaining the effective and successful operation of HOV and HOT lane facilities. HOV enforcement programs aim to deter possible violators and promote the safe and efficient use of HOV lanes (FHWA, 2016). All HOV facilities are monitored by state or local police officers to enforce the HOV lane requirements and to do this, officers must visually confirm whether the number of passengers in each vehicle meets the minimum requirements. Some states have programs that add self-enforcement by encouraging commuters to report HOV lane violators to the police (FHWA, 2020a). If found violating HOV facility requirements by enforcement officers, violators can be cited or re-directed back into the general-purpose lanes. Fines accompanying citations vary from state to state and depend on the number of citations previously received by the violator. For example, the minimum HOV violation fine in California is \$490, and the maximum fine for first offenders in Georgia is \$75 (Caltrans, 2020; GDOT, 2020).

As technology advances and improves, State DOTs may move toward automated HOV lane enforcement systems for user compliance. Vehicle Passenger Detection systems, which use video analytics to determine the number of occupants in vehicles, may take the role of police officers for enforcing HOV facility regulations.

Acknowledgment

We are grateful to Ms. Desiree James and Mr. Takuya Marui, University of Southern California graduate students, for their constructive input, which helped to improve the quality of the manuscript.

Relevant Websites

Alternative Fuel Data Center: <https://afdc.energy.gov/laws/HOV>
California Department of Motor Vehicle: <https://www.dmv.ca.gov/portal/dmv>
Federal Highway Administration: <https://ops.fhwa.dot.gov>
Federal Highway Administration: <https://ops.fhwa.dot.gov/publications>
Georgia Department of Transportation: <https://dps.georgia.gov>
Metropolitan Transportation Commission: <https://511.org/carpool/lanes>
New York City: <https://portal.311.nyc.gov>
Transport Canada: <https://web.archive.org>

References

- Alternative Fuel Data Center, 2016. Alternative Fuel Vehicles and High Occupancy Vehicle Lanes. Available from: <https://afdc.energy.gov/laws/HOV>.
- Berger, W.J., 2002. The Austrian HOV-lane—experiences in implementation and operation. *J. Civ. Eng. Manag.* 8, 255–262.
- California Department of Motor Vehicle, 2020. Clean Air Vehicle Decals - High Occupancy Vehicle Lane Usage. Available from: <https://www.dmv.ca.gov/portal/dmv/detail/vr/decal>.
- Caltrans, 2020. High-Occupancy Vehicle (HOV) Systems. Available from: <https://dot.ca.gov/programs/traffic-operations/hov>.
- Caltrans, 2009. Data Integration and Reporting Branch (DIRB) State Highway Mileage and Other Data from: https://www.metro.net/projects/hov/#_edn1.
- Cao, X., Xu, Z., Huang, A.Y., 2012. Safety benefits of converting HOV lanes to HOT lanes: Case study of the I-394 MnPASS. *ITE J.* 82, 32–37.
- Chang, M., Wiegmann, J., Smith, A., Bilotto, C., 2008. A review of HOV lane performance and policy options in the United States (No. FHWA-HOP-09-029). United States. Federal Highway Administration. Available from: <https://ops.fhwa.dot.gov/publications/fhwahop09029/fhwahop09029.pdf>.
- Cothron, A.S., Ranft, S.E., Walters, C.H., Fenno, D.W., Lord, D., 2004. Guidance for Future Design of Freeways with High-occupancy Vehicle (HOV) Lanes Based on an Analysis of Crash Data from Dallas, Texas. Available from: <https://static.tti.tamu.edu/tti.tamu.edu/documents/0-4434-P1.pdf>.
- Dixon, C., Alexander, K., 2005. Literature review of HOV lane schemes- Highway Agency—Unpublished Project Report. Available from: https://www.standardsforhighways.co.uk/ha/standards/pilots_trials/files/trl2005a.pdf.
- FHWA, 2020a. Frequently Asked HOV Questions. Available from: <https://ops.fhwa.dot.gov/freewaymgmt/faq.htm#faq7>.
- FHWA, 2020b. Chapter 2G. Preferential and Managed Lane Signs. Available from: <https://mutcd.fhwa.dot.gov/html/2009/part2/part2g.htm#figure2G01>.
- FHWA, 2017a. HOT Lanes, Cool Facts. Available from: <https://ops.fhwa.dot.gov/publications/fhwahop12031/fhwahop12027/index.htm>.
- FHWA, 2017b. HOV-to-HOT Screening Checklist. Available from: <https://ops.fhwa.dot.gov/publications/fhwahop12031/fhwahop12024/index.htm>.
- FHWA, 2016. Federal-Aid Highway Program Guidance on High Occupancy Vehicle (HOV) Lanes. Available from: <https://ops.fhwa.dot.gov/freewaymgmt/hovguidance/>
- FHWA, 2009. Manual on Uniform Traffic Control Devices (MUTCD). Available from: <https://mutcd.fhwa.dot.gov/>
- FHWA, 2007. Considerations For High Occupancy Vehicle (Hov) Lane To High Occupancy Toll (HOT) Lane Conversions Guidebook. Available from: <https://ops.fhwa.dot.gov/publications/fhwahop08034/>.
- GDOT, 2020. High Occupancy Vehicle (HOV) Lanes. Available from: <https://dps.georgia.gov/high-occupancy-vehicle-lanes>.
- Javid, R.J., Jahanbakhsh Javid, R., 2018. A framework for travel time variability analysis using urban traffic incident data. *IATSS Res.* 42 (1), 30–38, doi:10.1016/j.iatssr.2017.06.003.
- Javid, R.J., Nejat, A., Hayhoe, K., 2017. Quantifying the environmental impacts of increasing high occupancy vehicle lanes in the United States. *Transp. Res. Part D Transp. Environ.* 56, 155–174, doi:10.1016/j.trd.2017.07.031.
- Javid, R.J., Nejat, A., Salari, M., 2016. The Environmental Impacts of Carpooling in the United States, In: *Proceedings of the Transportation, Land and Air Quality Conference*.
- Javid, R.J., Salari, M., Jahanbakhsh Javid, R., 2019. Environmental and economic impacts of expanding electric vehicle public charging infrastructure in California's counties. *Transp. Res. Part D Transp. Environ.* 77, 320–334.
- Konishi, H., Mun, S.I., 2010. Carpooling and congestion pricing: HOV and HOT lanes. *Reg. Sci. Urban Econ.* 40, 173–186.
- Kwon, Y.J., Oh, J., Kim, S., 2005. A Study on the weekday HOV-lane policy on Kyoungbu expressway in Korea. *Int. J. Urban Sci.* 9, 31–39.
- Liu, X., Zhang, G., Wu, Y.J., Wang, Y., 2011. Measuring the quality of service for high occupancy toll lanes operations. *Proc. Soc. Behav. Sci.* 16, 15–25.
- Metropolitan Transportation Commission, 2020. Carpool & Express Lanes. Available from: <https://511.org/carpool/lanes>.
- New York City, 2020. New York City HOV Lanes and Daily Restrictions. Available from: <https://portal.311.nyc.gov/article?kanumber=KA-02801>.
- Rankin, M., 1994. Operational Design Guidelines for High Occupancy Vehicle Lanes on Arterial Roadways. Prepared for the Municipal/Provincial HOV Committee of the Ontario Ministry of Transportation. Available from: <https://ops.fhwa.dot.gov/freewaymgmt/publications/hov/00101625.pdf>.
- Sandelien, B., 2001. The first Norwegian carpool lane opened in Trondheim, in: *Nordic Road and Transport Research* 13 (3).
- Schijns, S., Eng, P., 2006. High occupancy vehicle lanes—worldwide lessons for European practitioners, in: *WIT Transactions on the Built Environment*, 89.
- Transport Canada, 2010. High Occupancy Vehicle Lanes in Canada. Available from: <https://web.archive.org/web/20120419031447/http://www.tc.gc.ca/eng/programs/environment-utsp-hovlanescanada-886.htm>.
- Turnbull, K.F., 2005. HOV and HOT lanes in the United States. *PROCEEDINGS OF ETC 2005, STRASBOURG, France*, 18–20.

Further Reading

- California. Division of Traffic Operations, California. Department of Transportation and Technology Sharing Program (US), 1991. *High Occupancy Vehicle (HOV) Guidelines for Planning, Design, and Operations*. State of California, Business, Transportation and Housing Agency.
- Javid, R.J., Nejat, A., Hayhoe, K., 2017. Quantifying the environmental impacts of increasing high occupancy vehicle lanes in the United States. *Trans. Res. Part D* 56, 155–174.
- Chang, M., Wiegmann, J. and Bilotto, C., 2008. *A compendium of existing HOV lane facilities in the United States* (No. FHWA-HOP-09-030). United States. Federal Highway Administration.

Reversible Lanes: Guidelines, Operation and Control, Research Directions

Gowri Asaithambi*, Venkatesan Kanagaraj†, Madhuri Kashyap*, *Department of Civil and Environmental Engineering, Indian Institute of Technology, Tirupati, India; †Department of Civil Engineering, Indian Institute of Technology Kanpur, Kanpur, India

© 2021 Elsevier Ltd. All rights reserved.

Introduction	52
Warrants for Reversible Lanes	53
AASHTO Green Book	54
ITE (Institute of Transportation Engineers)	54
FHWA (Federal Highway Administration)	54
State-of-the Art	54
Ordered Traffic	54
Simulation and Optimization	54
Implementation, Operation, and Safety	55
Disordered Traffic	55
Simulation and Optimization	55
Implementation, Operation, and Safety	55
Advantages and Disadvantages of Reversible Lanes	55
Operational and Safety Issues of Reversible Lanes	56
Reversible Lanes Control	56
Signs and Signals	56
Pavement Markings	57
Physical Separations	57
Conclusions and Directions for Future Research	58
References	59

Introduction

Traffic congestion is one of the major issues in most of the urban areas in the world. FHWA states that peak-period trips take an average of about 7% longer in United States (TCR, 2005). Based on global traffic congestion index, Mumbai and Delhi in India, Moscow and Saint Petersburg in Russia have been assigned top most ranks among most congested cities of the world (McCarthy, 2019). The congestion mitigation should be a major concern of traffic engineers and transport planners as it contributes to the accidents, environmental pollution and fuel consumption. The recent statistics by WHO shows that 1.35 million road traffic deaths occur every year globally and there is an increasing trend in the fatality due to accidents (GSRORS, 2018). Road transport is estimated to be responsible for up to 30% of particulate emissions (PM) in European cities and up to 50% of PM emissions in OECD (Organization for Economic Cooperation and Development) countries and accounted for about 23% of global carbon dioxide emissions in 2010. Low- and middle-income countries suffer disproportionately from transport-generated pollution, particularly in Asia, Africa and the Middle East (HSD, 2000). In order to reduce traffic congestion, imparting physical changes on existing roadway facilities (such as road widening, construction of additional lanes) in most of the metropolitan cities is difficult due to the space constraints. Under these circumstances, traffic management techniques such as lane management help traffic engineers and policy makers in dealing with traffic congestion on roadways.

A managed lane facility is one that increases roadway efficiency by implementing various traffic control and management strategies. Reversible lanes (also known as contraflow lanes) are one of the strategies applied to the managed lanes. A reversible lane or roadway is defined as one in which the direction of traffic flow in one or more lanes or shoulders is reversed to the opposing direction for a temporary period of time (GPDOERLS, 2010) as shown in Fig. 1.

Reversible lanes change the directional capacity of a roadway to accommodate peak directional traffic demands. Implementation of reversible lanes does not require extra space or alterations in the geometric features of the roadway. The reversible lanes can be designed and operated with the help of control signs or movable barriers which help the driver to negotiate easily.

Reversible lane operation was found to be first applied in 1928 on 8th Street in Downtown Los Angeles, known as off-center lane movement (GPDOERLS, 2010). Three lanes were assigned for eastbound travel and one lane was for westbound traffic during morning. This was reversed for the evening with three lanes westbound and one lane eastbound. In 1937, this operation was implemented along six miles of Wilshire Boulevard in Los Angeles which was found to be successful. One of the latest reversible lanes implemented in Tyvola road in Charlotte, North Carolina is shown in Fig. 2.

The next section focuses on warrants for reversible lanes. Section State-of-the-Art reviews the state-of-the art on reversible lanes under ordered and disordered traffic conditions. Advantages and disadvantages of reversible lanes are discussed in Section

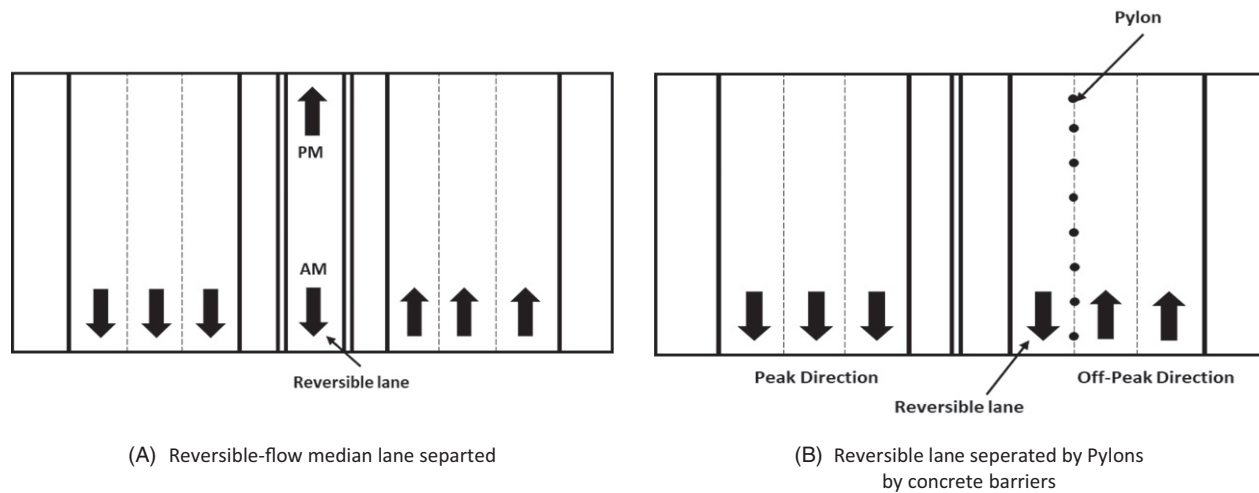


Figure 1 Reversible lane operation (A) Reversible-flow median lane separated (B) Reversible lane separated by Pylons by concrete barriers. *Source: Adapted from (GIML, 2016)*



Figure 2 Reversible lane on the lions gate bridge in Vancouver, Canada (Frejo, 2015).

Advantages and Disadvantages of Reversible lanes. Operational and safety issues of reversible lanes are provided in Section [Operational and Safety Issues of Reversible Lanes](#). The discussion on reversible lane control is presented in Section [Reversible Lanes Control](#) followed by conclusions and directions for future research in Section [Conclusions and Directions for Future Research](#).

Warrants for Reversible Lanes

The justification or warrants for providing reversible lanes given by various organizations are discussed in this section:

AASHTO Green Book

These manual states that reversible operations are justified when “65% or more of the traffic moves in one direction during peak hours” (APGDHS, 2001). It also suggests that a six-lane street with directional distribution of approximately 65%–35%, four lanes can be operated inbound (toward CBD) and two lanes outbound.

ITE (Institute of Transportation Engineers)

This organization although does not offer specific warrants, suggest a combination of criteria and studies that should be evaluated to justify the need of reversible lanes. The four ITE test criteria to determine the need for reversible lanes are as follows (NASEM, 2004):

- The average speed of the freeway should decrease by at least 25% over the normal speed, or there should be a noticeable backup at signalized intersections leading to vehicles missing one or more green signal intervals (i.e., demand should be greater than the capacity of the freeway)
- The traffic congestion problem under investigation should be both “periodic and predictable”
- The ratio of a major (predominant traffic) to minor traffic count should be at least 2:1 and preferably 3:1. Otherwise, the installation of a reversible lane could be the cause of a new traffic problem on the minor-flow side of the freeway
- The reversible lanes must be designed with adequate entrance and exit capabilities in addition to providing easy transition between the normal and reverse flow lanes. Otherwise, the reversible lane could be the cause of bottlenecks and other traffic problems in addition to the existing traffic congestion.

ITE provides few other guidelines to ensure the need of reversible lane implementation. Reversible lanes are desired in the locations where, the traffic studies show the following inferences:

- Lack of adequate adjacent side streets (not possible to implement one-way operation)
- A high proportion of commuter traffic is found that desires to traverse the area without turns or stops
- No other acceptable alternative improvement measures are possible, such as widening an existing facility or constructing a parallel roadway on a separate right-of-way due to right-of-way limitations and space constraints.

FHWA (Federal Highway Administration)

FHWA (Turnbull and Capelle, 1998) suggests a directional split of 60/40 as a threshold for the level of traffic imbalance needed to warrant a reversible facility.

State-of-the Art

This section focuses on the state-of-art studies related to evaluation of reversible lanes using simulation and optimization, field implementation, operational, and safety issues of reversible lanes under ordered and disordered traffic conditions.

Ordered Traffic

Simulation and Optimization

Zhou and coworkers (Zhou et al., 1993) proposed an intelligent-self learning dynamic optimal reversible lane control system for the George Massey Tunnel in southern Greater Vancouver. An optimization algorithm was developed to obtain real time optimal reversal schedule based on the predicted demand. A fuzzy modeling algorithm was developed to sort the real time data to identify the best matching pattern and a self-learning mechanism was used to modify the predicted demand incrementally. Meng and coworkers (Meng et al., 2008) developed a variation of genetic algorithm that embeds with microscopic traffic simulation model to find an optimal reversible lane configuration solution. Bi-level programming model was developed, where the upper level problem is a binary integer programming formulation that aims to minimize the total travel time of a study area and the lower level problem is a microscopic traffic simulation model (PARAMICS) simulating the dynamic reaction of the drivers which is caused by a reversible lane configuration scheme. A case study in Singapore is carried out to evaluate the proposed methodology. A bi-level programming model was presented by Wu and coworkers (Wu et al., 2009) in which the upper level model is to minimize the total system cost and flow entropy, while the lower level is a stochastic user equilibrium assignment with an advanced traveler information system. The objective function of the lower level model is to minimize the users’ travel cost under the stochastic user assignment with an ATIS. Result shows that, by adopting reversible lanes and an ATIS, the total system cost is greatly reduced. A framework for dynamic lane reversal was developed by Hausknecht and coworkers (Hausknecht et al., 2011), in which the lane directionality of reversible lanes is updated automatically with the information of instantaneous traffic conditions recorded by traffic sensors. Experimental results for assumed network indicate that reversible lanes provide increased network efficiency and dynamic lane reversals provide more benefit in travel time reduction inside the network. The study conducted by Liu et al. (2011) made an attempt to develop an optimization model using road impedance function (BPR function by FHWA) to determine the number of lanes to be converted for reversible use. The directional distribution factor of 2/3 was recommended for the criteria to activate the reversible lane system. A

macroscopic model has been presented by Frejo et al. (2015) for dynamic operation of reversible lanes. Two types of dynamic controllers were developed for lane control which involved, a logic based controller which works based on congestion lengths and a discrete model predictive control (MPC) which reduces the total time spent (TTS) of the network. The model and control algorithms were simulated and tested using loop detector data at SE-30 freeway Seville, Spain. The MPC system was found to be advantageous in reducing TTS on the network.

Implementation, Operation, and Safety

A unique solution that was developed in Utah to improve the efficiency of a major east-west urban arterial using existing road infrastructure was presented by Guebert et al. (2010). State Route 173 (urban arterial road) in Salt Lake has a week day traffic of 73/27 split in the morning peak hour and a 39/61 split in the afternoon peak. The recommended solution was to implement a reversible lane system that could shift the two-way left turn lane to accommodate lane reversal at different times of the day.

Operational and safety impacts of reversible lanes were analysed by Ma and Aden (Ma and Aden, 2011) on a reversible lane segment on Connecticut Avenue, NW. The reversible lane segment operated with four inbound (southbound) lanes and two outbound (northbound) lanes during the morning peak hours (7:00 am to 9:30 am) and it was reversed during the evening peak hours (4:00 pm to 6:30 pm). Three lanes are open to traffic in each direction at all other times. The peak direction accounted for 70 percent of the bidirectional traffic volume.

The effects of the reversible lane system on traffic flow and road safety was studied at a freeway work zone in Frankfurt, Germany by Waleczek et al. (2016). The road-safety analysis shows an increase of crash rates during road works. But, only 10% of the total number of crashes and no severe crashes could be linked to particular features of the reversible lane system. The evaluation study showed that, at the analyzed work zone, estimated travel time losses of 4,00,000 veh/h could be saved during two months of road works by the application of the reversible lane compared with a reduction of permanent lane in one direction.

Disordered Traffic

Simulation and Optimization

A dynamic control strategy and switch method for the operation of reversible lanes on freeways, bridges, or tunnels was designed by Ma et al. (2018) for both one and multiple reversible lane with continuous flows. Average delay of all the vehicles on the road were used to develop operation models of reversible lanes under different states. Enumeration algorithm were applied to solve the optimization models. The proposed methods and models were simulated and tested using data of a five lanes bridge in Guangzhou, China, via microscopic traffic simulation. The results show that the proposed method and models were able to reduce the delay and congestion. Kotagi and Asaithambi (2019) evaluated reversible lane operation using a microscopic simulation model developed specifically for urban undivided roads in disordered traffic. Results showed an increment of 568 veh/h during morning peak hour and 396 veh/h during evening peak hour after implementing reversible lanes.

Implementation, Operation, and Safety

Traffic conflict characteristics at a temporary reversible lane (1.4 km with 6 lanes going in two directions) at Huangpu Road (from Lingyuan Street to Gaoneng Street) in Dalian, China were investigated by Cao et al. (2014). Issues with temporary reversible lane such as inadequacy of control facilities, more number of conflicts during initial stage of direction changing and a lower utilization rate of temporary reversible lane were reported from the study. A case study on reversible lanes traffic controlling in Beijing, China was presented by Wang et al. (2014). The study stretch from Chedaogou Bridge to Sijiqing Bridge (Zizhuyuan Road) is 1.8 km long and 24 m wide (6 lanes) which is bidirectional was considered to design and implement the reversible lane operation. To implement reversible lane channeling, the central isolation barrier was moved and the main traffic lane was divided into 7 lanes, of which the most central lane is reversible lane. The estimated width of the reversible lane is 4 m with other traffic lanes being 3.25 m wide. The reversible lane operation was carried out every day between 6 am and 12 pm toward the city and for the rest of the day, the travel direction was away from the city.

Advantages and Disadvantages of Reversible Lanes

Major advantages of reversible lanes are decreased turbidity in adjacent flows, greater efficiency in toll collection for priced facilities and improved public perception of the system's efficiency (ADRL 2010). The other advantages of reversible lanes are listed below

- Reversible flow facilities are most suited for limited access through travel
- Efficient for longer distance trips
- Separation from normal lanes improves flow
- Maximizes V/C ratio utility by putting lanes in the direction having higher flow
- Potentially less expensive when compared to bi-directional facility
- Requires less right-of-way in general
- Provides air quality improvements by reducing congestion
- The construction period is lesser.

The disadvantages of reversible lanes when compared to normal bidirectional lanes are:

- Not well known to drivers and involves complex operations
- Requires studies to determine optimal hours of operation
- Less suited to short trips
- Requires additional signage and gates to prevent access to vehicles during off hours and needs more enforcement
- Transference onto a radial corridor may not be possible
- Variations in hours of operation can complicate accessibility
- Air quality benefits may not be significant in locations with lower directional splits

Operational and Safety Issues of Reversible Lanes

Reversible lane facilities though reduce congestion on peak hours, they may lead to some issues in traffic movement which trouble the commuters and sometimes the residents of the area. A study conducted in the United States ([Ma and Aden, 2011](#)) describes few issues associated with reversible lanes:

- The reversible section has a higher percentage of crashes during the hours of reversible operation than during regular operation. The major types of crashes that may occur on reversible roadways are sideswipe and head-on collisions.
- The reversible lanes are also not very pedestrian friendly due the absence of medians which increases pedestrians' exposure to vehicles when crossing the street. Also, reversible lanes make the crossing maneuver somewhat confusing for pedestrians, especially in areas that have large number of unfamiliar pedestrians.
- Reversible lanes can cause discomfort to the residents in the surrounding especially at those time periods when reversible lanes are used for one-way operations.

An analysis carried out on conflict analysis at temporary reversible in Dalian, China investigated the conflicts before and after the change in direction of movement on the reversible lane ([Cao et al., 2014](#)). The study stated that the traffic conflicts were more on the reversible lanes than the normal lanes. More number of conflicts are observed in the initial stages of direction change due to the adverse driving behavior due to the switching of movement directions.

A study conducted at a freeway work zone with reversible median lane in Frankfurt, Germany ([Waleczek et al., 2016](#)) showed that the marking of lane separation at the beginning of the reversible lane is necessary to ensure road safety. The study also stated that, in order to prevent crashes due to lane changes the barriers can be provided to separate lanes at both the sides of reversible lanes.

Transportation Association of Canada ([GPDOERLS, 2010](#)) suggests some unique requirements of reversible lane operations to address the issues associated with them:

- In case of reversible lane systems the through traffic is always given priority and hence turning movements are mostly prohibited along the roadway. Hence, there must be sufficient analysis made prior to the implementation of lane reversal so as to plan the provision of turning movements.
- The provision of pedestrian crossings along the reversible lanes should be limited to avoid the impact of lane reversal on pedestrians.

Reversible lanes do not contribute significantly to increased frequency or severity of accidents and additional care could be taken such as complete infrastructures, electronic information board and signal control ([Liu et al., 2011](#)). The properties of the barriers such as impact resistance, are also important in reducing the crash severity.

Reversible Lanes Control

The major methods of guiding and controlling traffic at the entrance, exit, and within reversible segments are roadway signs, signals pavement markings, and physical separation.

Signs and Signals

The earliest reversible segments were controlled only by signs and sometimes involved traffic enforcement personnel. The recent standards for the design and placement signs for reversible roadways are contained in the MUTCD 2009 ([MUTCDSH, 2009](#)). The use of both roadside ground-mounted signs and overhead signs for use on reversible arterial facilities is given in [Table 1](#). If overhead signs are used, they should be located at intervals not greater than 400 m (0.25 mi). The height of the overhead signs from the pavement grade should not be more than 5.8 m (19 ft). Where more than one sign is used at the termination of a reversible lane, the spacing between them should be at least 75 m (250 ft) apart. Longer distances between signs are suitable for roads with speeds over 60 km/h (35 mph), but the separation should be within 300 m (1,000 ft).

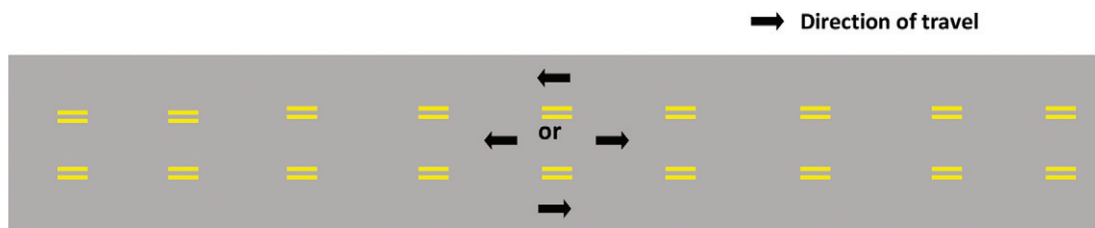
Lane-use control signals (refer [Table 2](#)) are used to indicate which lanes of a reversible roadway are available (or not available) for use in a particular direction and also, which lane might be in the process of changing operation ([NASEM, 2004](#)). MUTCD 2003

Table 1 Signs used in reversible lanes (MUTCDSH, 2009)

Symbol/word message	Meaning
Red X on white background	Lane closed
Upward pointing black arrow on white background	Lane open for through travel
Modified black left/through arrow on white background in case of permitted left turns	Lane open for through travel and any turns not otherwise prohibited
Black two-way left turn arrows on white background and legend ONLY	Lane may be used only for left turns in either direction (i.e., as a two-way left turn lane)
Black single left turn arrow on white background and legend ONLY	Lane may be used only for left turns in one direction (without opposing left turns in the same lane)

Table 2 Lane-use control signals used for reversible lanes (GIML, 2016)

Symbol	Name	Meaning
	Steady Downward Green Arrow	Permitted to drive in the lane
	Steady Yellow X	Prepare to vacate the lane
	Steady Red X	Not permitted to use the lane

**Figure 3** Reversible lane marking. Source: Adapted from (MUTCDSH, 2009).

(MUTCDSH, 2003) offers guidance on horizontally and vertically locating those devices along the roadway, stating that shall be clearly visible for 700 m (2,300 ft) at all times under normal atmospheric conditions, unless otherwise physically obstructed.

Pavement Markings

Pavement markings with words and symbols are considered supplementary to signs. MUTCD 2009 (MUTCDSH, 2009) suggested the following reversible lane pavement markings: “two normal broken yellow lines [on each side of the lane] to delineate the edges of lane in which the direction of travel is reversed” (refer Fig. 3). MUTCD 2003 (MUTCDSH, 2003) gives guidance on the use of raised pavement markers substituting or supplementing the pavement markings, including their lateral and longitudinal spacing.

Physical Separations

Various types of barriers have been used for reversible lane segments. A more innovative barrier system for reversible roadways is the movable barrier. Movable barriers have been used both on permanent bases for roadways and bridges and on temporary bases

Table 3 Summary of reversible lane control (GPDOERLS, 2010)

Control facility	Arterial (at-grade, minor or unrestricted access)	Freeway (limited access)
Signs	- Needs to be well signed to provide clarity to drivers; both main street and side street signage is recommended - Consider use of Dynamic message signs (DMS)	- Needs to be well signed to provide clarity to drivers - Consider use of DMS
Signals	- Green “down arrow”; flashing red “X”; solid red “X” - Where reversible lane signals are closer to intersection traffic signals, consideration should be given to turning off the green arrow when a red signal is displayed for that approach - Spacing is every 75 m or to fit situation - Signals should be placed such that they do not conflict with intersection traffic signals (not within 35 meters) - Drivers must be able to see at least 2 signal sets at a time in their direction of movement	- Green “down arrow”; flashing red “X”; solid red “X” - Drivers must be able to see at least 2 signal sets at a time - Only signals at transition zone may be needed in case of barrier separated reversible lanes
Pavement markings	Broken double yellow line dividing directional flows for different times of day	- Physical separation is preferred over lane markings to separate directional flows in freeway reversible lanes - Where physical separation is impractical, speed restrictions should be considered in conjunction with broken double yellow dividing lines
Separation barriers	Portable or none	- Barrier need to be considered to separate opposing traffic - Fixed or moveable concrete barriers have been used in different jurisdictions

within construction zones where unbalanced directional flows are experienced (GPDOERLS, 2010). The width of the buffer varies, and some are much wider and incorporate pylons or other channelizers to separate traffic flows (GIML, 2016).

Table 3 gives the summary of reversible lane controls to be adopted in different types of facility.

Conclusions and Directions for Future Research

Reversible lanes (contraflow lanes) is one of the traffic management measures suggested for reducing the traffic congestion on a roadway facility. In this chapter, warrants for reversible lanes suggested by few organizations like AASTO, ITE, and FHWA are discussed. Literature review shows that there are only few attempts made on the study of reversible lanes particularly on warrants, scheduling, operational, and safety issues. The advantages and disadvantages of reversible lanes and also, control of traffic on reversible lanes using different methods are explained in detail.

Further scope of the work are as follows: There are no clear warrants available for justification of reversible lanes for different countries. Standard warrants common for all the countries are difficult to formulate since the traffic conditions vary from country to country. However, few warrants (e.g., number of lanes, directional split) can be considered common for all the countries and country-specific warrants (e.g., traffic composition) can be added further based on local traffic conditions. Hence, there is a need for more number of studies to bring out the warrants for reversible lanes.

There are no rigorous studies available on finding optimal number of lanes and also, on optimal scheduling and operational hours of reversible lanes. More number of studies should be attempted to optimally utilize the reversible lanes both spatially and temporally using simulation and optimization techniques.

Real-time traffic management system can be employed to control traffic congestion more efficiently with the help of Internet of things (IOT) and cloud computing. IOT is connecting different application devices to one other through internet with the use of CCTVs, sensors and smart devices, which able to transmit a wide variety of data, location, movement, environment, etc. This system will be helpful to change the operational hours of reversible lanes intelligently according to traffic density, directional split on the roadway. It would assist road users to obtain prior contextual guidance with the help of dynamic message signs to avoid potential hazards on the reversible lanes.

In most of the developing economies, traffic is disordered in nature. In addition to that, traffic congestion is the most important problem due to increase in traffic volume with presence of side frictions like on-street bus stops, pedestrians, on street-parking, encroachments, etc. The extension of roads will not be sufficing to mitigate congestion. So, the available road width has to be effectively managed by implementing traffic management measures like reversible lanes. The existence of such lanes under disordered traffic conditions is hardly seen. Hence, there is a need for in depth study on reversible lanes in developing economies.

References

- Advantages and Disadvantages of Reversible Managed Lanes, 2010. Atlanta Regional Managed Lane System Plan Technical Memorandum 17C 2010 HNTB Corporation, Atlanta, Georgia.
- A Policy on Geometric Design of Highways and Streets (Green Book), 4th ed. 2001. American Association of State Highway and Transportation Officials, Washington D.C.
- Cao, Y., Zuo, Z., Xu, H., 2014. Analysis of traffic conflict characteristic at temporary reversible lane. *Period. Polytech. Transp. Eng.* 42 (1), 73–76.
- Frejo, J.R.D. 2015. Model predictive control for freeway traffic networks. Doctoral dissertation, PhD thesis, University of Seville.
- Frejo, J.R.D., Papamichail, I., Papageorgiou, M., Camacho, E.F., 2015. Macroscopic modeling and control of reversible lanes on freeways. *IEEE Transac. Intel. Trans. Sys.* 17 (4), 948–959.
- Guidelines for Implementing Managed Lanes. 2016. The National Academies Press National Academies of Sciences, Engineering, and Medicine, Washington D.C. <https://doi.org/10.17226/23660>.
- Guidelines for the Planning, Design, Operation and Evaluation of Reversible Lane Systems. 2010. Transport Association of Canada.
- Global status report on road safety. 2018. WHO, Geneva, Switzerland.
- Guebert, A.A., Carroll, D., Weston B., Kinnecom, D., 2010. Reversible Lanes in Utah - Adding Efficiency Safely. In Annual Conference of Transportation Association of Canada (TAC), Halifax, Nova Scotia.
- Hausknecht, M., Au, T.C., Stone, P., Fajardo, D. and Waller, T., 2011, October. Dynamic lane reversal in traffic management. In 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 1929–1934.
- Health and sustainable development: 2000 Air pollution and climate impacts. WHO, Geneva, Switzerland. <https://www.who.int/sustainable-development/transport/health-risks/air-pollution/en/>.
- Kotagi, P.B., Asaithambi, G., 2019. Microsimulation approach for evaluation of reversible lane operation on urban undivided roads in mixed traffic. *Transp. A: Trans. Sci.* 15 (2), 1–25.
- Liu, Y.S., Wang, X.H. and Liang, X.D., 2011. Conversion Mechanism of Reversible Lane System under Urban Tidal Flow Condition. In ICCTP 2011: Towards Sustainable Transportation Systems, pp. 1030–1041.
- Ma, J., Aden, Y., 2011. Reversible lane operation for arterial roadways: The Washington, DC, USA Experience. *ITE J.* 81 (5), 26.
- Ma, Y., Song, C., Wen, C., 2018. Dynamic Traffic Control and Direction Switch of Reversible Lanes for Continuous Flow Considering V2I. In CICTP 2017, Transportation Reform and Change—Equity, Inclusiveness, Sharing, and Innovation, American Society of Civil Engineers, Reston, VA, pp. 505–514.
- McCarthy, N. 2019. The World's Worst Cities for Traffic Congestion. *Forbes*. <https://www.forbes.com/sites/niallmccarthy/2019/06/05/the-worlds-worst-cities-for-traffic-congestion-infographic/#4c1e327612bc>.
- Meng, Q., Khoo, H.L., Cheu, R.L., 2008. Microscopic traffic simulation model-based optimization approach for the contraflow lane configuration problem. *J. Transp. Eng.* 134 (1), 41–49.
- Manual on Uniform Traffic Control Devices for Streets and Highways, 2003. U.S. Department of Transportation, Federal Highway Administration.
- Manual on Uniform Traffic Control Devices for Streets and Highways. 2003. U.S. Department of Transportation, Federal Highway Administration.
- National Academies of Sciences, Engineering, and Medicine, 2004. Convertible Roadways and Lanes. The National Academies Press. Washington, D.C. Available from: <https://doi.org/10.17226/23331>.
- Traffic Congestion and Reliability: Trends and Advanced Strategies for Congestion Mitigation. 2005. Cambridge Systematics, Inc., Massachusetts, United States.
- Turnbull, K.F. and Capelle, D.G., 1998. NCHRP Report 414: HOV Systems Manual. TRB, National Research Council, Washington, DC.
- Waleczek, H., Geistefeldt, J., Middendorf, D.C., Riegelhuth, G., 2016. Traffic flow at a freeway work zone with reversible median lane. *Transp. Res. Proc.* 15, 257–266.
- Wang, J.X., Liu, D., Zhao, Y., Wang, C., 2014. Study of tidal lanes on traffic controlling-taking Zizhuyuan road of Haidian district in Beijing as an example. *Appl. Mech. Mater.* 668, 1449–1452.
- Wu, J.J., Sun, H.J., Gao, Z.Y., Zhang, H.Z., 2009. Reversible lane-based traffic network optimization with an advanced traveller information system. *Eng. Opt.* 41 (1), 87–97.
- Zhou, W.W., Livolsi, P., Miska, E., Zhang, H., Wu, J. and Yang, D., 1993, October. An intelligent traffic responsive contraflow lane control system. In Proceedings of VNIS'93-Vehicle Navigation and Information Systems Conference. IEEE. pp. 174–181.

Electronic Toll Collection

Azusa Toriumi, The University of Tokyo, Tokyo, Japan

© 2021 Elsevier Ltd. All rights reserved.

Glossary	60
Nomenclature	60
Overview of Electronic Toll Collection	60
Technologies	61
Effect on Traffic on Toll Roads	62
Case Study in Japan	64
Traffic Management with Electronic Toll Collection	66
See Also	66
References	66
Further Reading	67

Glossary

Toll plaza an area along, at the entrance to, or at the exit from a tolled facility where tolls are collected, particularly areas consisting of a row of tollbooths across the roadway.

Nomenclature

ANPR Automatic Number Plate Recognition
AVI Automated Vehicle Identification
DSRC Dedicated Short Range Communication
ETC Electronic Toll Collection
EFC Electronic Fee Collection
GNSS Global Navigation Satellite System
ITS Intelligent Transport Systems
RFID Radio Frequency Identifier

Overview of Electronic Toll Collection

Electronic toll collection, often denoted as “ETC,” is a system capable of electronically charging a toll to a vehicle or its user by automatic identification and classification of vehicles on a tolled facility such as road, tunnel, or bridge.

Traditionally, toll was collected manually by cash, in some cases by token, pre-paid card, credit card, etc. Automatic payment machine was installed later in some cases instead of manual operation. Evolution of cashless and automatic transaction had contributed to reduce necessary time of transaction and labor costs. However, all of these payments require all vehicles to completely stop for a transaction at a tollbooth. Such conventional systems are not efficient for many reasons (Hensher, 1991):

- Each vehicle has to slow down, stop and wait for the transaction in toll plaza, which forces traffic flow to slow down, and results in congestion if amount of traffic comes in. Thus drivers incur delay and get frustrated. Furthermore, exhaust gas is emitted due to deceleration and acceleration.
- Necessary number of tollbooths (especially in open-toll system where tollbooths are on certain sections of roads) in toll plaza is typically larger than the number of traffic lanes, because traffic capacity of tollbooths is significantly lower than other general sections. That requires considerably high infrastructure costs including land requirements.
- Manual toll collection requires amount of labors, which is costly not only for their wage but also for more infrastructure, such as barrier or underground passage around tollbooths in order to protect the labors.
- Applicable tolling scheme becomes much limited. It is hard to apply complex tolls distinguished by time of day, type of vehicles, laden weight, distance traveled for example.

Electronic toll collection uses information communication technology (ICT) to automatically collect tolls without stopping vehicles at tollbooths, directly solving associated problems of congestion, delay, emissions, etc. It can reduce necessary infrastructure and land space in many cases, as well as reducing the labor cost. For example, in Norway, it was shown that a lane with electronic toll

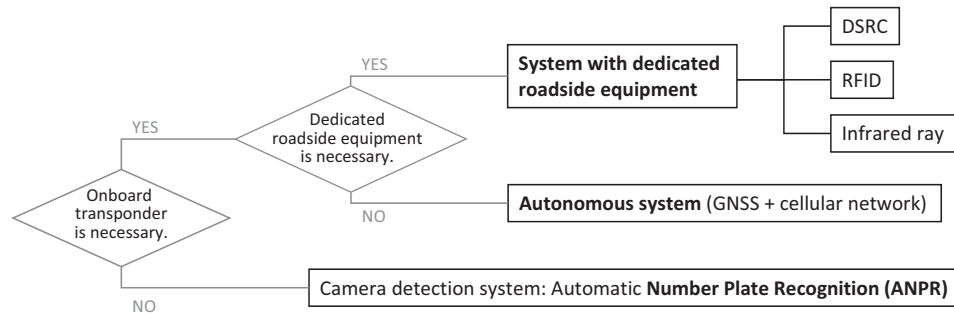


Figure 1 Types of AVI systems.

collection could handle three times of the traffic volumes handled on the manual lane, and the cost with full electronic collection could be between 33% and 50% of the one of a manual system (Hensher, 1991).

The cost to implementing and managing the electronic toll collection system depends on the case, but some empirical evidences have been demonstrated that it is more cost effective than manual operation. Thus, it has been widely spread and one of the most successful use case of the intelligent transport systems (ITS) around the world, since 1989 when it was fully deployed in motorways in Italy for the first time. According to the survey on the 43 different private and public entities involved in toll road operations across the Americas, Europe and Asia, 91% of the subject entities use some form of the electronic toll collection, and furthermore 27% uses only the electronic toll collection system and without any other alternatives as of around 2014–15 (KPMG International, 2015).

One of the most important advantage of the electronic toll collection is the flexible applicability to variety of tolling scheme. It makes it possible to differentiate tolls by time of day, type of vehicles, laden weight, distance traveled using an appropriate component in in-vehicle device, roadside equipment and back office. Because of this advantage, the system is currently considered as more widely applicable means of road pricing, thus sometimes called as the electronic fee collection (EFC), in which tolls are charged not simply on the individual routes or sections but rather inside the certain area such as city center (i.e., urban road pricing) or depending on traffic flow conditions (i.e., congestion pricing).

Technologies

There are four major components for the electronic toll collection: vehicle identification, vehicle classification, transaction processing, and violation enforcement. Currently, a variety of the technologies used for each component can be observed in different systems by country, state, or road operator.

Vehicle identification, abbreviated as "AVI" (Automated Vehicle Identification), is the most crucial process in electronic toll collection. It is a process of identifying the subject vehicles automatically. Basically, there are three different types of AVI: systems that uses dedicated roadside equipment, autonomous systems that uses satellite positioning, and camera detection system, as shown in Figure 1. The first two systems require onboard transponders.

The systems with dedicated roadside equipment are one of the most conventional and widely applied systems where subject vehicles are identified by a communication between onboard transponder located in the vehicles and roadside antennas when the vehicles pass through. Dedicated Short Range Communication (DSRC) has been used world widely, and Radio Frequency Identifier (RFID) is also getting popular in many countries. Some systems use infrared ray for the communication instead of these radio wave. Systems using DSRC and RFID are illustrated in Figure 2. In general, DSRC can achieve higher communication speed, larger transmission capacity and higher accuracy than the other systems, and enable bidirectional communications between onboard transponder and roadside antenna. In other words, both vehicle-to-infrastructure (V2I) and infrastructure-to-vehicle (I2V) communications are possible with DSRC. Thus, it is possible to expand the functions of the onboard unit using the same transponder (e.g., traffic information can be provided from roadside antenna to vehicle). On the other hand, RFID can be installed relatively cheaper because it requires only a small RFID tag for onboard transponder unlike the mechanical onboard unit for DSRC. Therefore, it is easier to introduce at lower cost, but the back office has to process much more task such as matching of the entry and exit data than the DSRC system where many tasks can be locally done by roadside equipment.

Apart from the above-mentioned systems, autonomous systems use satellite positioning for AVI, not requiring dedicated roadside equipment, as illustrated in Figure 3. This type of systems localizes the vehicle using Global Navigation Satellite System (GNSS) and find its position on the map contained in the onboard unit. Here, additional sensors such as gyroscopes, odometers, and accelerator meter are combined in many cases. Then, as the map in the onboard unit include geographic information of tolled area, vehicles can be identified when it enters and exit the tolled area, and these information is sent to the back office that makes payment transaction via cellular network. As mentioned, this system does not require dedicated roadside equipment for AVI, thus infrastructure cost can be lower. However, as means of violation enforcement, roadside equipment such as DSRC antenna or cameras are still often used for compliance checking. Great advantage of the autonomous systems is its wide applicability to various types of tolling scheme, such as distance-based tolls, route-based tolls, urban area pricing, zone pricing, etc., as its vehicle identification can be done continuously.

Automatic Number Plate Recognition (ANPR) is another way of AVI where roadside cameras capture images of vehicles passing through toll plazas, and the number plates of the vehicles are extracted through image recognition. This does not require any

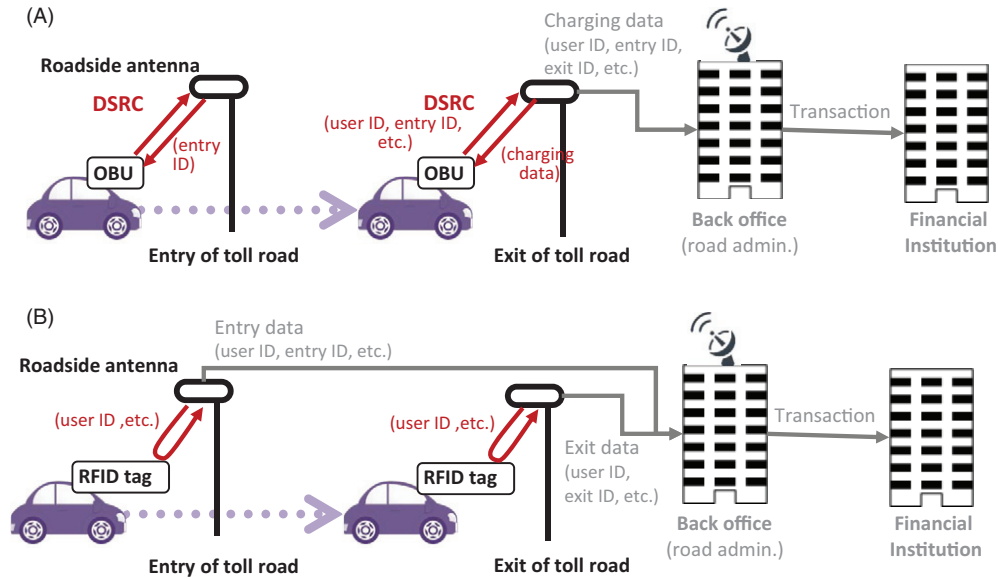


Figure 2 Systems using (A) DSRC and (B) RFID.

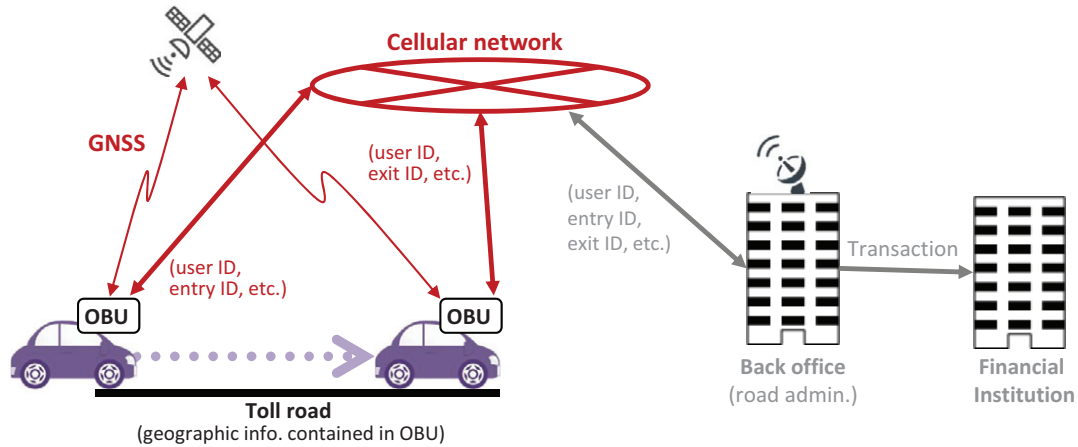


Figure 3 Autonomous systems using GNSS.

onboard transponder thus allows users to use toll roads without any preparation (i.e., purchase of onboard unit or RFID tag). However, at this moment, accuracy of fully automatic recognition is not as high as other systems. Manual review can be supplementary incorporated in order to improve the accuracy, but that requires additional labor costs. Privacy concern and jurisdictional problem are also issues which often raises for ANPR in many countries.

As mentioned, each system has advantages and disadvantages on different aspects such as accuracy, flexibility to tolling scheme, possible expansion of systems, interoperability, initial and running costs for road users as well as those for road operators. Thus, it is important for countries, states, or road operators to select the system based on their needs and requirements. New systems with new technologies are also emerging in some countries. Types of charging technologies (mainly about vehicle identification) and tolling schemes (in other words, charging policies) in different countries are listed in [Table 1](#).

Vehicle classification is necessary when different toll rates are charged for different types of vehicles. The technology is closely related to vehicle identification (AVI), as the simplest method is to store the vehicle type information in onboard transponder and send it together when vehicles are identified. Another way is to use the sensors, such as inductive sensors, laser sensors, etc. Weigh-in-motion (WIM) is a new technology which can capture the axle weight and gross vehicle weights when vehicles drive over the measurement site. Using this technology, it is possible to measure the weight without requiring vehicles to stop. WIM is often used to enforce a law on truck overloading, and further started to be utilized for toll-by-weight in China ([Hang et al., 2013](#)).

Effect on Traffic on Toll Roads

As mentioned, one of the most important benefit of electronic toll collection is reduction of congestion and associated negative impacts such as delay and emission. Several indices have been used to evaluate the effects of electronic toll collection on the

Table 1 Types of charging technologies and policy (JSAE, 2019)

Charging policy		Financing of road infrastructure			Traffic management	
		Toll road (ETC)	Inter-city road (Heavy goods vehicle charge)	Every road	Urban road (Rush hour charge)	Inter-city road
ANPR: Identification plate scan					London Stockholm	
DSRC		World wide (More than 50 countries)	Austria, Czech Republic Poland, (Slovenia)		Oslo, Bergen, etc. Singapore	
GNSS	Mobile phone network		Germany, Slovakia, Hungary, Belgium, Russia (Bulgaria)		(Singapore)	
	Odometer			USA Toll charge		
	DSRC					Japan Charge by guiding routes
RFID: Electronic tag		North America, South and Central America India, Taiwan, etc.				USA High-speed lane
WAVE: New DSRC		(South Korea)				
WIM: Dynamic load measuring apparatus		China				

Note: Countries in parentheses planning to introduce in near future

Source: https://www.jsae.or.jp/01info/org/its/its_2019_en.pdf

performance of toll plaza operations through empirical analysis and estimation models development (e.g., Al-Deek et al., 1997; Aycin, 2006; Woo and Hoel, 1991):

- (Traffic) capacity is defined as the maximum sustainable hourly flow rate at which vehicles can traverse a point or a section of a lane or roadway during a given time period. As for a lane with a tollbooth, it can be measured by counting vehicles departing from the tollbooth under saturated condition where queues are present upstream.
- Throughput per lane is the vehicle counted at the downstream of the tollbooth during a period of time. Unlike the capacity, throughput is measured regardless of the upstream queuing condition.
- Service time is the time that a vehicle spends to pay a toll at the tollbooth. Intervehicle time is the time difference between two consecutive departures of vehicles at the tollbooth. It is the sum of the service time and the vehicle headway. These two indices do not include the waiting time in the queue.
- Queuing delay is the time that each of individual vehicles spends in a queue upstream of the tollbooth. This does not include the service time. Average queuing delay is the average, and total queuing delay is the sum of queuing delay over all vehicles in the queue.
- Time spent in a toll collection system includes the sum of the effect of service time and queuing delay.
- Delay or lost time is the additional time experienced by the presence of tollbooths.

As mentioned, in Norway, a lane with electronic toll collection could handle three times of the traffic volumes handled on the manual lane (Hensher, 1991). In the United States, NCHRP (TRB, 1997) reported the capacities of a lane with manual collection and electronic toll collection are 350 veh/h and 1200 veh/h, respectively. Here, the lane with electronic toll collection means the one dedicated for electronic toll collection only, but it is located in the same toll plaza with different types of lanes (e.g., manual lane). The report also mentioned that if a lane with electronic toll collection is physically separated from other types of lanes and allows free-flow operation, which is called as an "express" lane, capacity can be further higher to 1800 veh/h. It is known that such a physical and operational conditions of electronic toll lanes (e.g., separation from manual lanes, presence of the gate, speed limit) affect capacity and operational performance. Another report in the United States by Al-Deek et al. (1997) mentioned that the capacity of a lane with electronic toll collection (not an express lane) was higher than the value shown in NCHRP (1200 veh/h) (TRB, 1997), which is considered due to the absence of the gate at the tollbooth. Furthermore, they reported that the service time, the average queuing delay and the maximum queuing delay could be decreased by five seconds per vehicle, one minute per vehicle and 2.5–3 minutes per vehicle, respectively by the electronic toll collection.

Mixed lanes which allow both manual and electronic payments are also installed in many cases. Performance of the mixed lane is not as high as the dedicated lane with electronic toll collection mentioned above but still higher than the manual lane [e.g., capacity is 700 veh/h according to NCHRP (TRB, 1997)]. It also depends on the proportion of users of the electronic toll collections; the more users pay electronically, the higher the lane performance becomes.

Besides, evaluation of the performance of as a whole of toll plaza often consisting of multiple lanes with different types of toll collections (manual, electronic, mixed, etc.) is also important. Different configuration of lanes may result in different levels of service, because it affects users' lane choice and conflicts between vehicles on the upstream of the tollbooth. It is recognized that adding another mixed lane does not necessarily increase capacity proportionally, because the capacity of the mixed lane depends on the proportion of electronic toll users as mentioned above but it is not known (Aycin, 2006). In general, users of electronic toll collection are likely to prefer the dedicated lane because it can process the queue faster (Al-Deek et al., 1997; Aycin, 2006), but it may also depend on the penetration rate of the electronic toll collection. Therefore, enough consideration has to be made for designing toll plaza.

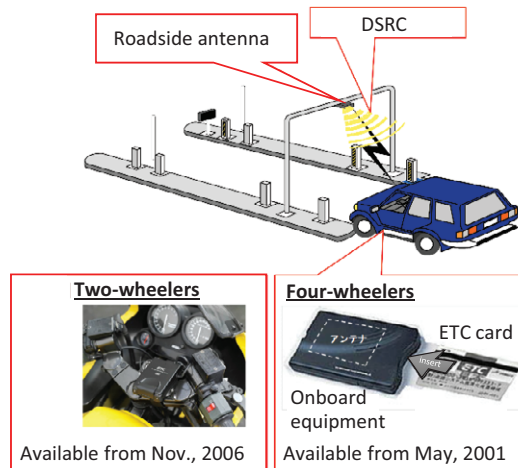


Figure 4 Systems of ETC in Japan (MLIT Japan, 2012). Source: <http://www.mlit.go.jp/road/ir/ir-council/pdf/7.pdf> (translated by the author)

Case Study in Japan

Japan is one of the few countries where a single universal system of electronic toll collection, which is called as “ETC” system, is applied to all of the toll expressways (controlled-access motorway) in the country. In Japan, research and development for the electronic toll collection had been conducted from 1994 to 1995 jointly by public sectors which are in charge of road construction and operation and private companies such as communication equipment, electric equipment, computer, and car manufacturers. Based on these results and experiments in the test field, the first test operation was done in the actual expressway sections in 1997, and specifications were established and published for the universal system in the country in 1998. In November 2001, it was finally deployed nationwide.

Figure 4 describes a schematic image of the ETC system in Japan. It uses DSRC for vehicle identification and classification, thus vehicle information is stored in the onboard unit and sent to the central office through the communication with roadside antenna. Onboard unit consists of an onboard equipment and a so-called “ETC card”; the former needs to be installed in a vehicle and store the vehicle information such as vehicle type and plate number, the latter is issued by credit card companies and used for payment. This is called as two-piece system, and makes it possible for the persons who are not the owner of the car (e.g., rent-a-car) to pay the fee by inserting his/her own ETC card into the onboard equipment. The onboard units for two-wheelers (i.e., motorcycles) are also sold from 2006.

Because users need to purchase and install onboard unit by themselves, government (i.e., MLIT Japan) and expressway companies put some incentives such as discount or mileage point on road users who buy the onboard unit in an early stage of the deployment. Thanks partly to these promotion, ETC penetration rate, percentage of those who paid their tolls by ETC system among all expressway users, have increased rapidly since its deployment in 2001; it reached above 15% in 2004, 56% in 2006, 83% in 2010, and 92% in 2019 (MLIT Japan, 2019).

As one of the effects of such increase of ETC users, congestion had been drastically reduced. It was reported that the total lost time due to congestion had been decreased with the increase of the ETC penetration rate since 2002; Actually, the total lost time could be reduced into less than one third with the ETC penetration rate of 25% within the first few years on the intercity expressways of east Japan (East Nippon Expressway Co., Ltd., 2020). Although there may be several additional factors for this reduction, we can identify that the introduction of ETC would have contributed so much. As one of the evidence, **Figure 5** shows the comparison of number of traffic breakdowns occurred at the major bottlenecks on expressways by segment type between before and after the deployment of ETC. As shown in the figure, toll plazas had been a major type of bottlenecks in expressway network, as breakdowns had occurred about four thousand times per year which accounts for 30% in total. However, it could be reduced to only sixty times per year which accounts for 0.8% after the introduction of ETC.

From the viewpoint of operational costs, ETC could be cheaper than manual collection, because it could reduce the large amount of labor costs. **Figure 6** shows that in FY2011 when 15% of road users paid manually and 85% of users used ETC, costs spent for manual collection was higher than those for ETC, due to the large amount of outsourcing costs for the manual collection done by labors.

Furthermore, ETC have made it possible to introduce new designs of interchanges with on- and off-ramps where toll plaza must be located between expressways and other highways. Such interchanges conventionally need larger space as illustrated in **Figure 7A**, because they usually require to put tollbooths for different directions into one toll plaza so that they can reduce labors and facility for manual toll collection. However, it became possible to operate unmanned tollbooths with the ETC, and which made it possible to separate toll plazas for different directions and need less space as illustrated in **Figure 7B** and **C**. Conventional interchanges without

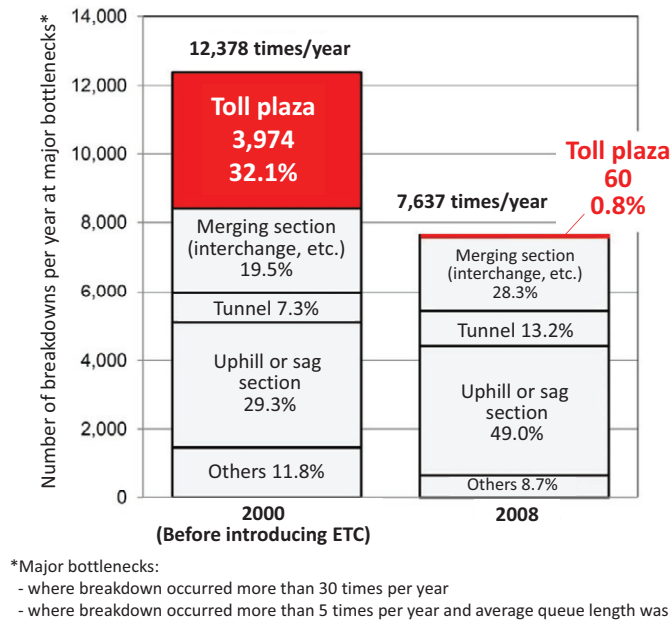


Figure 5 Comparison of occurrence and cause of breakdowns at major bottlenecks before (year 2000) and after (year 2008) introducing ETC in Japan (MLIT Japan, 2012). Source: <http://www.mlit.go.jp/road/ir/ir-council/pdf/7.pdf> (translated by the author)

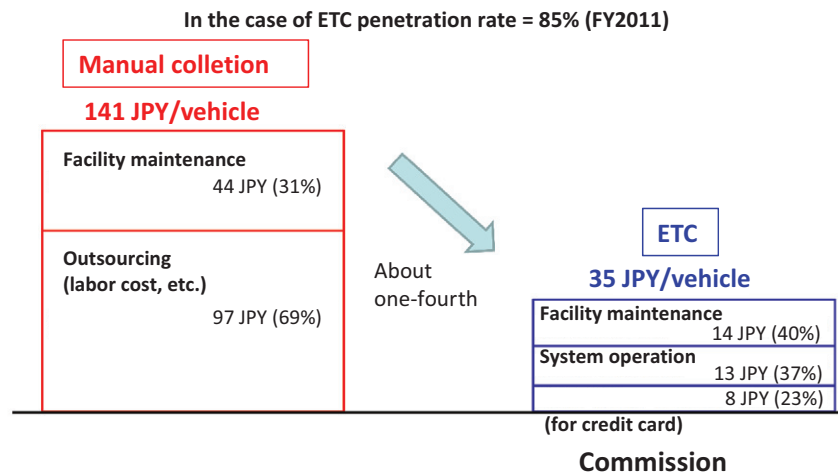


Figure 6 Comparison of the costs for manual and ETC payment per vehicle (MLIT Japan, 2012). Source: <http://www.mlit.go.jp/road/ir/ir-council/pdf/7.pdf> (translated by the author)

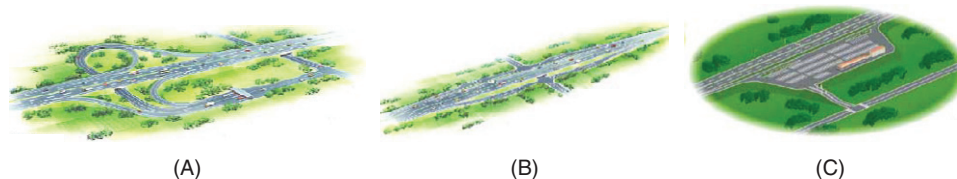


Figure 7 Schematic image of (A) conventional interchange, and new interchanges, (B) connected to the mainline directly and (C) connected to the rest area (MLIT Japan, 2012). Source: <http://www.mlit.go.jp/road/ir/ir-council/pdf/7.pdf> (adjusted by the author)

ETC costs 3.5 billion JPY for construction and 90 million JPY/year for maintenance, while new types of interchanges with ETC costs 2.0 billion JPY and 70 million JPY/year respectively as of November 2012 according to MLIT Japan (2012).

As mentioned above, one of the advantages of electronic toll collection is its flexibility of available tolling scheme. Here, one example in Tokyo is the discount for detours which contribute to relief congestion introduced in 2014; when users took longer

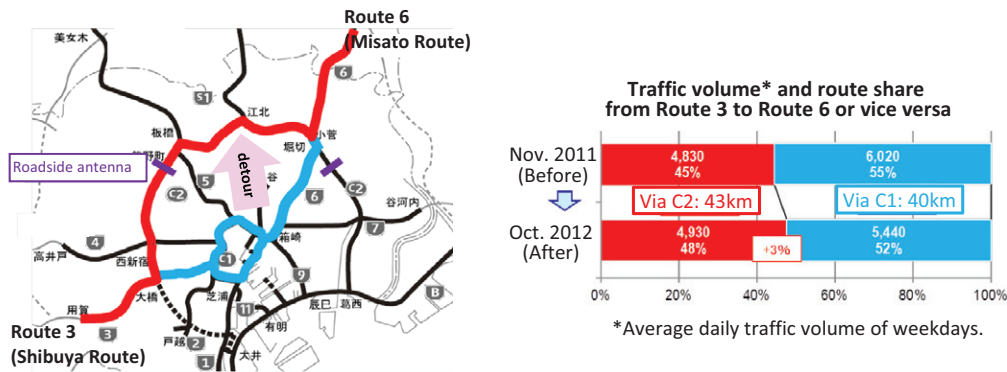


Figure 8 Discount for detour using ETC and its effect (MLIT Japan, 2012). Source: <http://www.mlit.go.jp/road/ir/ir-council/pdf/7.pdf> (translated by the author)

detour using ring road (C2, shown by the red line in Figure 8) and avoided from passing through the city center (C1, shown by the blue lines), toll were reduced by 100 JPY for passenger cars and 200 JPY for large vehicles. In order to give such a discount, it is necessary to identify route of vehicles from their origins to destinations, which can be hardly done by manual tolling but is possible by putting roadside antennas on the mainline of the detour route (as shown by purple lines in Figure 8). Right-hand side chart in Figure 8 shows an effect that the through traffic passing on the C1 road reduced and the route share was shifted in about 3%. Besides, flexible tolling scheme like discount in night time, discount for the companies which have frequent use of their vehicles, and giving mileage points for the certain time of day on the certain area have been introduced by utilizing ETC.

In 2014, the upgraded systems which is called “ETC 2.0” was deployed in Japan. The new types of onboard units can record vehicle trajectory using GPS and upload it to back office via roadside antennas along expressways and national highways using DSRC. It became also possible to provide area-wide road traffic information from roadside antennas to onboard units.

Traffic Management with Electronic Toll Collection

Road pricing has been considered as an efficient way to manage traffic demand. Due to the advantage of flexibility of tolling scheme, electronic toll collection has enabled to introduce road pricing with more choices of the charging objective, charging schemes, and charge levels.

One successful example is found in Singapore. Singapore introduced cordon-based road pricing named as Area Licensing Scheme (ALS) in 1975, which charged drivers who entered downtown in order to manage traffic demand. It was upgraded into so-called Electronic Road Pricing (ERP) using DSRC system of electronic toll collection. In this system, tolls are varied by time of day and by location of the gantry with DSRC antennas. The rate of tolls is reviewed every three months so that the average travel speeds on the roads can be maintained at the certain level (i.e., 20–30 km/h on city roads and 45–65 km/h on expressways). Singapore is currently developing the next generation road pricing system based on GNSS technologies, and expects to introduce around 2020. (ITF, 2019)

With the development of connected and automated vehicles and infrastructure, importance of traffic management using such technologies will be paid by much more attention in the future. Accordingly, electronic toll collection is expected to be evolved and integrated into other ITS systems which realize more efficient and sustainable transportation.

See Also

Regulation and Financing of Toll Roads; Road Pricing 4: Case Study - The Implementation of Electronic Road Pricing in Hong Kong; Road Pricing 1: The Theory of Congestion Pricing; Bottleneck; Local Area Traffic Management

References

- Al-Deek, H.M., Mohamed, A.A., Radwan, A.E., 1997. Operational benefits of electronic toll collection: Case study. *J. Transp. Eng.* 123 (6), 467–477.
- Aycin, M.F., 2006. Simple methodology for evaluating toll plaza operations. *Transp. Res. Rec.* 1988, 92–101.
- East Nippon Expressway Co., Ltd., 2020. “We analyze causes of traffic congestion and implement countermeasures to solve traffic congestion and mitigation”, Available from: https://www.e-nexco.co.jp/activity/safety/detail_07.html.
- Hang, W., Xie, Y., He, J., 2013. Practices of using weigh-in-motion technology for truck weight regulation in China. *Transp. Policy* 30, 143–152.
- Hensher, D.A., 1991. Electronic toll collection. *Transp. Res. Part A: General* 25 (1), 9–16.
- International Transport Forum (ITF), 2019. *Smart Use of Roads*, ITF Research Reports, OECD Publishing, Paris.
- KPMG International, 2015. “KPMG toll benchmarking study 2015 An evolution of tolling”, Available from: <https://assets.kpmg/content/dam/kpmg/pdf/2015/06/kpmg-toll-benchmarking-study-2015-v2.pdf>.
- Ministry of Land, Infrastructure, Transport and Tourism Japan (MLIT Japan), 2012. “Current status of the use of the ETC and its effects, etc.”, Available from: <http://www.mlit.go.jp/road/ir/ir-council/pdf/7.pdf>.
- Ministry of Land, Infrastructure, Transport and Tourism Japan (MLIT Japan), 2019. “Status of the use of the ETC.” Available from: <https://www.mlit.go.jp/road/yuryo/etc/riyou/index.html>.

Society of Automotive Engineers of Japan, Inc. (JSAE), 2019. ITS Standardization Activities of ISO/TC 204 2019, p.18, Society of Automotive Engineers of Japan, Inc., Tokyo.

Transportation Research Board (TRB), 1997. National Cooperative Highway Research Program NCHRP Synthesis 194, A Synthesis of Highway Practice, Electronic Toll and Traffic Management (ETTM) Systems, NATIONAL ACADEMY PRESS, Washington D.C., pp. 10–12.

Woo, T.H., Hoel, L.A., 1991. Toll plaza capacity and level of service. Transp. Res. Rec. 1330, 119–127.

Further Reading

CEN/TC 278 ITS standardization, 2020, "Electronic Fee Collection," Available from: <https://www.itsstandards.eu/efc>.

Road Pricing—Theory and Applications

Kian Keong Chin, Chief Engineer, Road & Traffic, Land Transport Authority, Singapore

© 2021 Elsevier Ltd. All rights reserved.

Introduction	68
The Economics Theory of Road Pricing	68
Early Road Pricing Studies	69
Purpose and Types of Road Pricing	70
Technologies for Road Pricing	70
Unique Identification of Vehicles	71
Payment Mechanism of the Road Pricing Scheme	71
Ensuring Compliance	71
Examples of Implemented Road Pricing Schemes	72
Singapore	72
London	72
Stockholm	73
Conclusion	73
References	73

Introduction

Transportation is a vital element in any community endeavor to strive and progress. Often, it is a means to an end, that is, it is an activity that connects two different functions or places such as from homes to workplaces, or homes to schools. Efficiency and safety are important attributes of transportation systems although other attributes such as comfort, cleanliness and environmental friendliness have also been considered important. Efficiency can mean reliable travel times and respectable travel speeds, but can also mean equity in terms of offering equal opportunities for all segments of a community to use and enjoy.

As travel patterns changes and demand increases with evolving life-styles or land-use intensification, the transportation infrastructure becomes under-capacity and not able to meet the needs of the community. It used to be a case that this increased demand is met with improved infrastructure capacity involving new or widened roads, but these often proved to be short-lived as demand quickly imposed itself on the increased supply. Hence, many communities and cities have realized the need to improve their public transport system, comprising both buses and subways (often refer to as metros or mass-transit systems).

While useful, the provision of high-capacity public transport systems do not, by itself, relieved the road network of traffic congestion as is evident from many large cities with a decent public transport system. There is a need for a mechanism, other than road congestion itself, to persuade car-users to move to public transport or to travel outside the peak periods. This mechanism is pricing and is often referred to as road pricing but can also be termed congestion-pricing, road-user charging, or similar.

The Economics Theory of Road Pricing

The economics theory to support the case for introducing road pricing is well documented in many literatures, for example, [Arnott \(2005\)](#) and [Wohl and Hendrickson \(1984\)](#). When traffic on a road is very light with no congestion, the average cost faced by each of the few individual motorists can be taken to be that incurred by that motorist alone. Since traffic is very light, there is not much interference or cost imposed from other motorists. When traffic increases, motorists will start to interact with each other and slow each other down. Each additional motorist will cause some slow-down to the traffic stream, imposing this delay to all other motorists already traveling on the road. This is the marginal cost introduced by the additional user, and will increase disproportionately as subsequent additional motorists join the traffic stream, as is indicated by the curve marked MC in [Fig. 1](#). Consequently, the average cost (AC in [Fig. 1](#)) of motorists will also increase as more motorists join the traffic stream. Also in [Fig. 1](#) is the demand curve or sometimes referred to as the marginal benefit curve, signifying the propensity for motorists to join the traffic stream as cost changes. As cost increases with more traffic joining the traffic stream leading to congestion eventually, there will be increasingly fewer motorists willing to join the traffic stream.

Motorists will continue to join the traffic stream as long as the benefit exceeds the cost perceived by that motorist. An equilibrium position is achieved when the benefit of this last motorist who joined in is equal that of the average cost faced by this motorist, and this is indicated as A, corresponding to the flow F_{eq} in [Fig. 1](#). However, at this traffic flow, the marginal cost is much higher because there are delay costs imposed on other motorists by that motorist joining the traffic stream. As a result, there is a social loss to the society and this is represented by the hatched zone. This social loss can be reduced if fewer motorists are using the road, and in fact

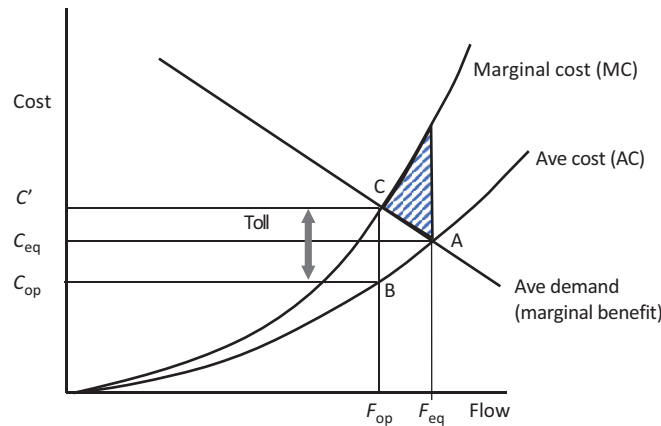


Figure 1 Economics of road pricing.

can be eliminated if traffic flow is reduced to F_{op} . A road pricing charge equal to BC can be imposed on motorists to reduce the demand to F_{op} .

To put it in layman perspective, this can be taken to mean that road-users consider only their own cost of travel in deciding whether to travel on that road. They would not consider the (societal) cost imposed on other fellow road-users or to the environment. The additional (societal) cost could be attributed to the little extra delay imposed on all other road-users as a result of that individual joining in the stream of traffic on a road. Likewise, the (societal) cost could be the environmental degradation introduced by that individual and imposed on all present along that travel route, including residents living along that route. Introducing a road pricing charge or toll can eliminate the (societal) cost, which basically means that traffic congestion is reduced or even eliminated.

When there is reduced demand arising from the imposition of the road pricing charge, some of these (marginal) road users will not use the road. They can either travel during another time or use another route with a lower road pricing charge or no charges at all, travel using other means of transport, for example, public transport where this road pricing charge does not apply directly, or even abandoning that trip altogether.

A popular query raised from the introduction of road pricing is on who the beneficiaries are. Obviously, all those who remained on that road and paid the road pricing charge can be viewed as financially worse off, and likewise those that are priced off would be worse off because they no longer enjoy using the road at the desired time, or enjoyed using the preferred transport (i.e., the car) of her choice. However, reduced congestion from the imposition of the road pricing charge meant that all those who remained are actually better off from reduced travel times. The other factor in this scenario is on how the collected road pricing charges would be used. If these are used to improve the road or the public transport option, then obviously there are societal benefits gained.

Early Road Pricing Studies

[Smeed \(1964\)](#) in a report published and named after him more than 50 years ago, provided practical advice for implementing road-pricing schemes. There were nine specified requirements for a workable road pricing scheme, and another eight desirable criteria.

The nine specified requirements listed below are still valid for the several road-pricing schemes that were implemented in the years that followed, after its publication in 1964.

1. Charges should be closely related to the amount of road use.
2. Charges should be variable.
3. Charges should be stable and ascertainable by road users before the start of their journey.
4. Payment in advance should be possible.
5. Incidence of system upon individual road user should be accepted as fair.
6. Method used should be simple for road users to understand.
7. Any equipment used should be highly reliable.
8. System should be reasonably free from fraud and evasion.
9. It should be capable of being applied, if necessary, in the whole country.

The Supplementary Licensing scheme proposed by the then [Greater London Council \(1974\)](#) for the city of London required drivers of certain vehicles to purchase special licenses to use their vehicles at specified times in designated areas. However, the Greater London Council eventually decided not to proceed with the supplementary licensing scheme after it assessed that there would be adverse effects on the lower income motorists and the problems of enforcement. It was also concerned on the possible effect on smaller businesses, industry and travel needs at times when the public transport option was not available, as was reported by [May \(1975\)](#).

Purpose and Types of Road Pricing

The early proposals for road pricing were designed to be essentially manually operated, mainly because the technologies available at that time do not permit more automated road pricing systems. It was only with the advent of electronic tags and wireless communications that electronic road pricing systems became possible.

Nevertheless, the purpose for implementing electronic road pricing needed to be decided upfront to design an appropriate and effective system. While the purpose of road pricing schemes is usually to mitigate traffic congestion, there can be other reasons for introducing it. In recent years, these schemes have a stronger association with environmental protection. Pricing is used to manage the use of higher emission vehicles, whose relatively higher pollutant emissions may be a problem for more sensitive communities, or just for general health reasons. It can also be extended to schemes to meet emission standards arising from climate change protocols. Another purpose for having road pricing may well do with raising revenue to meet various objectives, which may be transport-related, for example, for road maintenance or for improving the public transport system.

Designing a road-pricing scheme is not necessarily a straightforward task. There are many considerations, and it is necessary to have the intent of the road-pricing scheme in mind before the scheme get conceptualized and designed. Depending on the intent of the scheme, one of several types can be used and the basic options as proposed in [Walker and Pickford \(2018\)](#) are as follows:

1. Point-based charging where specified vehicle are charged a fixed amount each time a passing is made.
2. Cordon-based charging, where each entry into a zone defined by the cordon is charged. A variation to this is to also include the payment of a charge even when moving within the cordoned zone.
3. Distance-based charging where the charge is variable and dependent on the distance traveled. This is a refinement of the point-based and cordon-based types because with this, the amount of charge can become more connected and relevant to the actual degree of congestion contributed by each road user.

If the intention is to mitigate congestion, the charging will be applicable only when there is high traffic demand or where congestion would have resulted if there had been no road pricing. Hence, for all three types of road pricing options, there will be defined operating hours and these are usually during the peak morning or peak evening periods, although this can also be throughout the working hours of each weekday. An extension of this concept, in line with the requirements recommended in the Smeed Report, is to have the charges variable depending on the actual traffic demand for road use. Hence, the morning peak period will have a higher charge than during the mid-day period. In fact, [Chin \(2005\)](#) reported that for the Singapore's Electronic Road Pricing (ERP) system, the charges are in 30-min blocks and hence can change every half-hour. However, these variable charges should be made know upfront before the road user gets onto the road, and this is consistent with the requirements in the Smeed Report.

For distance-based charging, not only will the charge be variable depending on the time-of-day but also be dependent on the stretches of road traveled; those stretches with higher traffic demand or congestion will have a higher charge for the same time period.

Regardless of the type of road pricing scheme selected, there are some common issues that have to be considered. One of these is public acceptability, where the majority of stakeholders will have to assess the scheme as being fair, reasonable, and equitable. This means that there has to be a thorough study on who the stakeholders are and the extent to which each group will benefit or suffer. Appropriate complementary measures will then have to be introduced to get a wide enough support for that scheme. These complementary measures include:

1. Improving the public transport system as an alternative for those who are priced off the road.
2. Compensating rebates or reduction in transport-related fixed fees for targeted stakeholders.
3. Appropriate traffic schemes to improve efficiency of existing transport system, so that there is no argument that the road-pricing scheme was introduced because of inefficiencies caused by the failure of the transport agencies to act in the first instance. These include having traffic lights that are well-designed, coordinated, and reliable, having good enforcement action against on-street parking violations as these are often the cause of congestion and effective traffic management schemes.

Communicating the scheme to the many identified stakeholders is an important task, and these can range from town-hall type presentations to small customer-focused discussions to even one-to-one interactions with specific individuals. Hence, it is useful to keep in mind that the scheme should be as simple as possible, so that this is easily communicated, understood and accepted.

Technologies for Road Pricing

The Supplementary Licensing proposal for London in 1974 had (paper-based) licenses pre-purchased for all specified vehicles that intended to enter the priced zone, and it was reportedly not implemented because of (among others) difficulties with enforcement. Admittedly, the technologies at that time were limited and do present challenges for an automated system. However, it did not stopped the introduction of the first area-wide road pricing scheme in 1975, in the Central Business District (CBD) of Singapore.

Called the Area Licensing Scheme (ALS), [Chin \(2002\)](#) reported that the technologies used for this manually operated road-pricing scheme were not sophisticated. Many of its processes were manual. It relied on the pre-purchase of paper licenses from convenience stores, post-offices, and sale booths just outside the CBD. The CBD cordon was marked out with physical gantries, and carried the words "Restricted Zone" accompanied with a set of amber lights. These sets of traffic lights were switched on manually by the enforcement officers based on the time of their watches. Some flexibility was allowed for the switching on of the amber lights, as

it was not wise to do so when traffic was flowing through it because the traffic lights were showing green. Hence, these were switched on only after the traffic lights had turned red to stop the traffic entering the CBD. Enforcement was undertaken by enforcement officers stationed at each of the gantries, who would note down the registration numbers of vehicles that were seen without the pre-purchased license valid for that day.

Even though the ALS was a manually operated scheme, the processes can be grouped under three functional categories, that is, the unique identification of vehicles, the payment mechanism and ensuring compliance. These three functional categories are also relevant for electronic road pricing, and the possible technologies that could be adopted for each of these categories are described in the next sections.

Unique Identification of Vehicles

For road pricing to impose a charge on vehicles passing through a pricing point, it must be uniquely identified so that there is a proper record of there being a valid charge imposed for that particular vehicle. Traditionally, vehicles on the roads are identified by their vehicle registration number plate. For electronic road pricing, each vehicle has also to be identified uniquely in an automated manner. The identifier can be the vehicle's registration number on its vehicle plate, and these can be read by LPNR (license plate number recognition) techniques. This is the method used by the road pricing system in London and Stockholm to identify each vehicle uniquely. Alternatively, each vehicle can be fitted with an on-board electronic tag that carries a unique electronic code, and this can be read by roadside readers or antennae via radio frequency or other wireless communications. This is the system adopted by the Electronic Road Pricing (ERP) system in Singapore. A requirement is that the on-board electronic tag must be permanently linked to that vehicle (which is often the registration number of the vehicle's number plate). Checks can be randomly made to ensure that this remains so, from matching the electronic code from an installed on-board tag to that of the vehicle's license plate.

There have been some proposals to use a driver's mobile phone as the vehicle identifier for road pricing, but this is fraught with problems. Drivers can forget to bring their mobile phones when they get into their vehicle, or may have brought another mobile phone instead. There is also the possibility that the battery power can run out during a particular trip, and becomes inoperable. While the mobile phone is used as an identifier, there will be a need for an apps to be running to do all the functions required of road pricing, and the issue will be whether the driver will activate the apps before starting their journeys or in a worst case scenario, make illegal and unauthorized modifications to the apps to allow, for example, travel with no charges deducted.

Payment Mechanism of the Road Pricing Scheme

The second category of a road pricing function focuses on the acceptance and recognition of payments made by the road user. Payments can be made in advance, at the time of using or even after having used the priced roads. The important aspect of these payments is that it must be accurately associated with the vehicle for which this payment was made, that is, it should be linked to the identifier used for that vehicle.

For payments made in advance, the amount is linked to the identifier and kept in a backend database. Also recorded in that database is often the date or time for which the payment is valid for. Hence, when that vehicle identifier, for example, the vehicle registration number read off a LPNR system is picked up later, it is used to search against the payment records in the database. When the unique identifier is found, it is then taken that the payment has been made for that day or time period. This is the system used for the London's congestion pricing.

Payments can also be made at the point of using the priced road. This requires the system to uniquely identify the vehicle and made the deduction of the road pricing charge from a stored value account. The stored value could be on a smart card inserted into a reader or electronic tag, like that used in Singapore's ERP system or it could be in a backend account held with a financial entity or a bank. With the latter, this requires the transmission, usually wirelessly, of the unique identifier of the vehicle that used the priced road to this financial entity or bank. Together with other information like the time of day and location of the priced road, the correct amount of road pricing charge is deducted from that stored-value account.

It is also possible that the payment is made after the use of the priced road, and this is often useful when the road user decides to use the priced road at moment's notice for whatever reason. The payment system can be designed to accept payment for some time, usually by the end of that day or even the following day after the road user has used the priced road.

Ensuring Compliance

Road pricing schemes are often not popular with motorists and hence, there will always be the tendency of motorists to find shortcomings in the system to avoid paying the charge. Hence, this third functional category of ensuring compliance is just as important.

Ensuring compliance need not mean just an effective enforcement system, with detection devices like video cameras to capture non-payments or road users interfering with the normal operations. Publicity on the need for road pricing to benefit the community as a whole could also serve as a means for motorists to understand the rationale for paying the road pricing charges and willingly pay the charges when needed to. The publicity can also be on the high reliability of the detection devices to capture all non-payments, to the extent that road users will just pay since there would be no way to avoid not being detected.

Of course, the detection devices and the processes to capture non-payments must be highly reliable. Often the unique vehicle registration number on its license plate is the identifier used to determine the vehicle owner. The vehicle registration number captured by high-resolution video cameras would then be processed with optical character recognition (OCR) techniques at the

backend to automatically read and identify the vehicle registration number. There will always be a small number that could not be read by the OCR techniques and when this happens, it is often that the video image is then viewed manually by a human operator.

The legislation will often have the owner of the offending vehicle also the person liable to pay any penalty. However, it is conceivable that the driver of a vehicle is not the owner of that vehicle as in the case of fleet vehicles or rental vehicles. Hence, there must be a process to deal with this sort of situation. This could simply have the owner of an offending vehicle be provided with a channel to report on who the driver was at the time when the offence took place, and allow the penalty to be imposed on the named driver instead.

Examples of Implemented Road Pricing Schemes

Singapore

In 1998, Singapore had its manually operated ALS replaced by an Electronic Road Pricing (ERP) System. It used Dedicated Short-Range Communications technology for reading information from uniquely coded In-vehicle Units (IUs) that are fitted to vehicles that opted to do so. The fitting of the IU is not mandatory although almost all vehicles do so, because the same IU can be used for entry and exit into car parks. However, it is mandatory for an IU to be fitted when the vehicle passes through a pricing gantry during its operational hours.

The IUs, with its unique identification code, are used by the ERP system to recognize payments associated with it. This unique IU code does not carry the identity of the vehicle nor particulars of the vehicle's owner. This is in line with safeguarding the privacy needs of motorists and the link to vehicle's owner is only established for enforcement actions.

The IUs accept a national stored-value card issued by the local banks, and this is inserted into the IU to effect payments—for passage through road pricing gantries and also for payment on exit of car-parks. This stored-value card can be taken out from the IU to facilitate topping up of its value at banks' ATMs or at convenience stores. In recent years, the ERP system has also been modified to allow ERP payments to be deducted from back-end accounts of vehicle owners. After signing up for this mode of payment, there is no longer the need to insert the stored-value card into the IU. Instead it recognizes registered IU codes, and effects the back-end payments instead of taking it as a non-payment case.

To ensure compliance, all the road-pricing gantries are fitted with infrared enabled cameras to read license plate numbers of non-complying vehicles, that is, those that do not pay the road pricing charge. The OCR technique to automatically read the license plate numbers is not able to read all of them for a number of reasons (e.g., dirty license plates) and if so, the pictures taken by the cameras are used for manual processing. Vehicle owners are identified and issued with letters to make good the non-payment but with the addition of an administrative fee. However, if the vehicle owner does not make this payment asked for in the letters issued to them, they will be issued a traffic summons for which the non-payment can result in them being subsequently issued a court summons.

The ERP system started operations in 1998 with 33 road pricing gantries, with most of these forming a cordon around The Central Business District (CBD). Since then, more road pricing gantries were introduced at specific locations on the expressway network, and also about an outer cordon bordering the Outer Ring Road system of the CBD although this outer cordon is "leaky" in that there are several roads on the western part that are not priced. Hence, there are a total of 78 road-pricing gantries today (2019).

The road pricing charges vary throughout the day, with them changing as often as every half-hour. For the CBD, motorists are priced throughout the day from 8 a.m. to 8 p.m., with the exception of a 2-hour break from 10 a.m. to 12 noon. Elsewhere, the road pricing charges are mostly during the morning and/or evening peak periods.

The reason for varying road-pricing charges is because the intent is to keep operating speeds for each half-hour period within an optimal range based on speed-flow studies. For expressways, the optimal range is between 45 to 65 km/h. For other roads, this is between 20 to 30 km/h. The speeds for each half-hour are measured every 3 months to determine if the road pricing charges should be either increased or decreased. After years of operations, the road pricing charges do not vary often and have kept relatively stable.

While the ERP System has worked well for Singapore over the past 20 years, there are plans to replace this with a Global Navigation Satellite System (GNSS) based system where there is no need for physical road pricing gantries. Each on-board unit (OBU) in the vehicle uses satellite signals to determine its location, and based on this decides if a road pricing charge is to be imposed. Payment is expected to be either through a stored-value card or a backend account, similar to the practice used for the current ERP system. Enforcement is still necessary through cameras installed at strategic locations in the road network.

London

London's Congestion Pricing Scheme was implemented in 2003, and requires specified vehicles entering or remaining in Central London or the Charging Zone between 7 a.m. and 6 p.m. on weekdays, to pay the congestion charge. Details of this scheme are available in [Transport for London \(2007\)](#)'s report. The scheme was extended to the west of the city to an area that is known as the Western Extension Zone in 2007. This was, however, scrapped at the end of 2010.

The unique identification of vehicles is based on their license plate number, and this is read with cameras installed at entrances into Central London, as well as within Central London. The Congestion Charge is a fixed daily amount and this payment can be made in advance or on the day of use up to midnight that day. The user would provide the vehicle's license plate number together with the payment. There is also a provision for payments on the following working day, although this would attract a higher charge. All these payment information is stored in a backend database.

Following the detection and reading of vehicles' license plate number within the Charging Zone, this is compared to the list of vehicles that have paid in the backend database. If there is a match, the detected vehicle is taken to have paid the Congestion Charge. If there is no match, it will be taken that there has been a violation and a Penalty Charge Notice (PCN) is issued.

There are discounts and exemptions for specific vehicles in the Congestion Charging scheme. For example, residents are offered a discount of 90% while emergency vehicles and taxis are exempted from the payment. Private hire vehicles, which had enjoyed the exemption in the past, are no longer exempt and they pay the prevailing Congestion Charge like any other similar vehicle.

There is also a parallel Ultra Low Emission Zone (ULEZ) that operates for the same Congestion Charge zone, and vehicles that do not meet the minimum emission standards will have to pay a fee to enter and remain within the zone. Although this is strictly not a road-pricing scheme, it has the effect of managing the traffic within Central London.

Stockholm

Stockholm's charging scheme started in January 2006 as a pilot for 7 months and requires specified vehicles to pay a charge for entering or leaving the inner city from 6:30 a.m. to 6:30 p.m. on weekdays. Eliasson (2014) describes the scheme and its traffic impact, with the varying charge dependent on the time of day, with the maximum during the peak morning and evening periods. The pilot ended at the end of July 2006 with a referendum that had its residents voted in favor of the charging scheme, mostly because of the improvement in traffic congestion during the pilot. It started operations again in August 2007.

Like the London's Congestion Charging Scheme, the unique vehicle identifier for Stockholm's is the vehicle license number plate. It uses cameras to capture the vehicle's license number plate and imposed the relevant charge (based on the time-of-day) on that vehicle. Since the charge is time-dependent, it was not possible for payment in advance. Instead, an electronic invoice for all applicable charges for each month is sent to the vehicle owners. Payment can also be by direct debit from their bank accounts.

Since this is a post-paid scheme, there are technically no violations on the ground. The violation, if it occurs, comes from non-payment on receiving the electronic invoice.

Conclusion

Since the introduction of road pricing in Singapore, London, and Stockholm, there have been several other cities that have followed suit. These include Milan and Gothenburg, although the purpose for doing so in these cities have not been entirely to deal with traffic congestion.

While there are cities that introduced road-pricing schemes, there are also others that have attempted to do but did not get the support to do so. These include Manchester, United Kingdom where the referendum in 2008 had its residents not in favor of it and in New York City in 2007 where the NY State Assembly blocked it. Interestingly, New York City has now (2019) started to develop its road-pricing scheme again with the aim of not just relieving traffic congestion, but also to raise revenue to improve its public transport including its subway system.

Ultimately, the introduction of road pricing schemes is a complex one. It is often just one element in any communities' transport strategy, with the alternative of an effective and affordable public transport being an important complementary one.

References

- Arnott, R., 2005. City tolls – one element of an effective policy cocktail. CESifo DICE report. J. Institut. Comparison. 3 (3), 5–11.
- Chin, K.K., 2002. Road Pricing – Singapore's Experience. IMPRINT-EUROPE Thematic Network: Implementing Reform on Transport Pricing: Constraints and Solutions – Learning from Best Practice, Brussels.
- Chin, K.K., 2005. Road pricing – Singapore's 30 years of experience. CESifo DICE report. J. Institut. Comparison. 3 (3), 12–16.
- Eliasson, J., 2014. The Stockholm Congestion Charges: An Overview, CTS Working Paper 2014:7, KTH Royal Institute of Technology. Available from: <http://www.transportportal.se/swooped/CTS2014-7.pdf>.
- Greater London Council, 1974. Supplementary Licensing, London.
- May, A.D., 1975. Supplementary Licensing: An Evaluation, Traffic Engineering & Control.
- Smeed, R.J., 1964. Road Pricing: The Economic and Technical Possibilities. HMSO, London.
- Transport for London, 2007. Central London Congestion Charging. Impacts Monitoring – Fifth Annual Report. Available from: <http://content.tfl.gov.uk/fifth-annual-impacts-monitoring-report-2007-07-07.pdf>.
- Walker, J., Pickford, A., 2018. Types of road pricing, and measuring scheme cost and performance, Chapter 3 Road Pricing—Technologies, economics and acceptability, IET.
- Wohl, M., Hendrickson, C., 1984. Transportation Investment and Pricing Principles. John Wiley & Sons, Inc.

Road Pricing 1: The Theory of Congestion Pricing

Timothy D. Hau, HKU Business School, The University of Hong Kong, Hong Kong

© 2021 Elsevier Ltd. All rights reserved.

Introduction	74
An Analytical Framework for Charging Excessive Use of Private Transportation	74
Optimal Pricing	76
The Welfare Impact of Congestion Pricing	79
Distribution of the Values of Time	80
Congestion Toll as a User Fee	81
References	82

Introduction

This is the first of four articles that introduce the role of pricing in the establishment of an efficient transportation system for cities. In this article we develop the key theory of marginal cost pricing, subsequent articles use this: to develop the equilibrium conditions for highways in the short and long run; the complementary role of pricing public transportation; and finally, a case study of the application of electronic road pricing to Hong Kong.

Congestion pricing, popularly known as road pricing, aims to reduce excessive traffic during rush hours to the Central Business District (CBD).¹ Since the social cost of a trip typically diverges from its private cost, a congestion charge is imposed by an economic efficiency-enhancing authority to internalize the external effect brought about by a motorist. By doing so, total travel times by motor cars and buses to and from the CBD, together with their vehicle operating costs, are saved. In a wider context, road pricing is the application of market-oriented principles to curtail excessive automobile traffic and to encourage the use of public transportation. Our framework is to adopt the principle of marginal cost pricing on both private transportation via a congestion toll and to public transportation via an optimal transit subsidy. Without pricing as a quasi-market mechanism's signal, public transportation services by municipalities are likely to continue to suffer mounting deficits. Road pricing would constitute an indispensable part of a sustainable transport package of road and public transport provision.

An Analytical Framework for Charging Excessive Use of Private Transportation

Traffic congestion is well studied in the traffic engineering and the transportation economics literature (Dupuit (1844, 1849), Greenshields, 1935, Wardrop, 1952, Beckmann et al., (1956, Article 3), Vickrey (1955, 1969), Walters (1954, 1968)). We start with the technology of road congestion by assuming that homogeneous vehicles move uniformly on a stretch of urban road (Mohring and Harwitz (1962, Article 2)). By following a defensive driving rule and avoiding tailgating, a driver is advised to maintain a comfortable distance by counting 3 s behind the car in front of him. This so-called "three-second rule" roughly translates to the conventional safety education dictum of maintaining one car length for every 10 miles per hour (or 16 km/h) that one is cruising along. If all motorists adhere to such safe driving practice, a linear relationship between traffic density, D , in vehicles per kilometer per lane, and traffic speed, S , in kilometers per hour, follows:

$$F \equiv \text{Speed} \cdot \text{Density}$$

$$F \equiv S \cdot D$$

vehicles per hour per lane $\equiv$ kilometers per hour $\times$ vehicles per kilometer per lane

The identity says that traffic flow is the speed at which vehicles move times the number of vehicles in a lane-kilometer (Fig. 1A). Institutionally, a speed limit, $S^{\max}$, is often imposed by society on motor vehicles in order to lower traffic speed from the technologically achievable, S^f , as shown by the horizontal stretch leading up to the kink in Fig. 1A,B. At first, with a few entrants on the road, increased interactions amongst the vehicles would yield an increase in traffic flow but at a diminishing rate. At the largest traffic flow – as shown by a rectangular area in Fig. 1A – civil engineers' "optimal" density, D^m , and "optimal" speed, S^m , are reached. This point, at which normal congestion turns into hypercongestion, is the maximum capacity (or engineering capacity) at $F^{\max}$ (see Fig. 1B):

$$F^{\max} \equiv D^m \cdot S^m$$

However, with many more vehicles per hour entering the roadway, traffic speed continues to fall. Past a certain point—the mid-point of the linear speed-density relationship—traffic flow begins to fall rapidly until traffic flow and speed becomes zilch (see Fig. 1B). Instead of deriving the famous bullet-shaped speed-flow curve in this way, we could tell the story slightly differently: as more

¹ Here we deal mainly with congestion pricing as opposed to marginal social cost pricing, which is defined to include both the private costs and the external costs of congestion, air pollution, noise pollution, accidents, road damages and climate change (see Hau (2005a), Small and Kazimi (1995) for example).

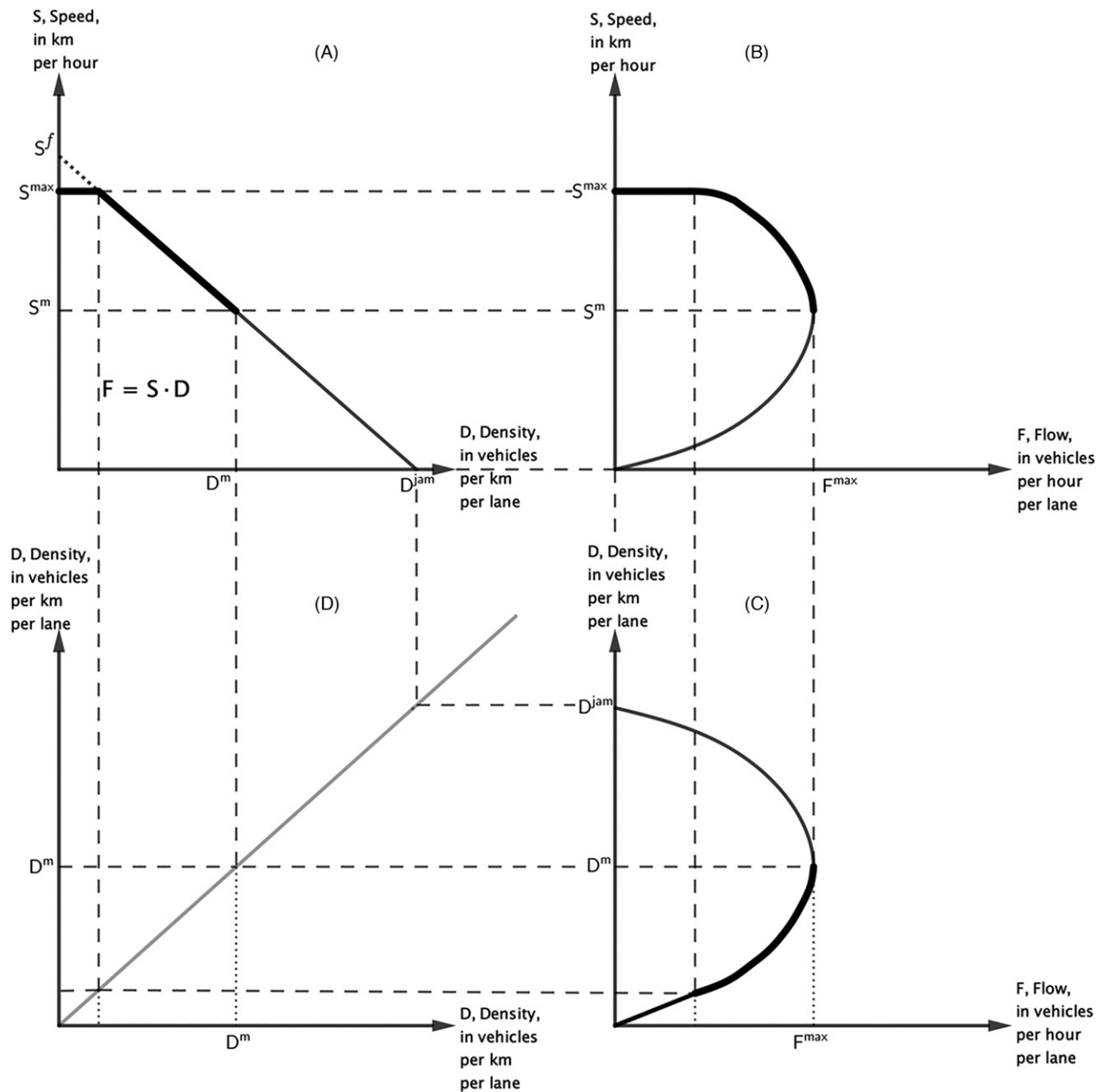


Figure 1 Fundamental Diagram of Road Traffic: Expressed in three forms based on the identity flow = $F = S \cdot D$

and more vehicles per distance of road space increases, traffic volume increases at first but falls until traffic comes to a standstill and jam density, D^{jam} , is reached, with the road becoming a parking lot (see Fig. 1C). The density-flow curve, or the flow-density curve with the axes reversed, in Fig. 1C, is “the fundamental diagram of road traffic” coined by Frank Haight (1963, article 3).² The initial downward-sloping part of the speed-density diagram is referred to as the congested or normally congested part (shown as the bolded portion of the line) whereas its latter part is called the hypercongested part (shown as the unbolded portion) (Fig. 1A).³ (The figure for

² As the fundamental relationship of such a steady-state traffic flow of identically-operated vehicles traversing a road uninterruptedly is an identity, the congestion technology could be expressed equivalently as a functional relationship between any two of the three variables: F , S and D (see Small and Verhoef, 2007, Chapter 3). As an illustration, consider the flow-speed relation: the third variable, density, D , is simply the slope connecting the origin to any point on the flow-speed relation. Similarly, for the flow-density relation, the third variable, speed, S , is the slope connecting the origin to a point on the flow-density relation.

³ Similarly, the normally congested part appears in a different form on the upper branch of the speed-flow curve (shown as the bolded portion of the curve) whereas the hypercongested part appears on its lower branch (shown as the unbolded portion) (Fig. 1b). Finally, the normally congested part once again appears on the lower branch of the density-flow curve (shown as the bolded portion of the curve) whereas the hypercongested part appears on the upper branch (shown as the unbolded portion) (Fig. 1c). The properties of such an identity have just been described. The last figure on the southwest quadrant is merely a 45 degree line that connects the values of density, thereby relating the speed-density diagram with the density-flow diagram: it is a mirror image of itself.

the engineering capacity for a typical expressway is about 1800 – 2000 vehicles per hour per lane at 50–55 km/h (or 30–35 miles/h) (see Gerlough and Huber (1975, article 4), Transportation Research Board (2011), Wong and Wong (2016) and any Highway Capacity Manual, for example, Transportation Research Board (2016)).

Drawing a parallel of our four-quadrant diagram with that of the principles of microeconomics is useful. With a conventional downward-sloping linear demand curve where price is placed on the vertical axis and quantity demanded is plotted on the horizontal axis in the northwest quadrant as in Fig. 1A, total expenditure comes about from the product of price and quantity, resulting in a price-total expenditure curve in the northeast quadrant as in Fig. 1B and a quantity-total expenditure curve in the southeast quadrant as in Fig. 1C. At the peak of the total expenditure–price curve and that of the total expenditure–quantity curve, the price elasticity of demand is unity. Pricing a good or service at a medium level, synonymous with the quantity demanded at a medium level for a linear demand curve, naturally yields the consumer's largest total consumption expenditure. Any price or quantity demanded that deviates from this middling level will result in a diminution of total expenditure and a price elasticity of demand that departs from unity.

Given a trip distance of 1 km of road stretch, say, the civil engineer's speed–flow curve (Fig. 1B), reproduced as Fig. 2A, is straightforwardly inverted to become an (average) travel time–flow curve, as travel time is the reciprocal of speed (Mohring, 1999). Travel time enters as the primary input into the production of a vehicle trip and, together with the money cost of travel, constitutes the costs of a trip. Travelers are frequently observed to trade off travel time with money so an estimate of the (marginal) value of time could be obtained from modal choice models' calibration of the shadow price on travel time.⁴ By using a representative value of time estimate, one could straightforwardly convert travel time (in time units) into time cost (in dollar units).

Money outlay is assumed to be independent of output. At low traffic density, motorists could drive fast, incurring high fuel costs. At high traffic density, vehicles slow down thereby lowering fuel costs, but with motorists having to weave in and out of heavy traffic and slow down, high fuel costs are incurred too. These two factors roughly cancel one another out, leading to the rough approximation that the costs of operating an automobile—which include fuel, engine oil, maintenance, and depreciation costs—are roughly independent of the level of traffic flow (Mohring (1976, article 3), AASHO's (1960) "Red Book"). The money cost, known as the vehicle operating cost, could therefore be summed up on a par with time cost, forming the traveler's generalized cost, GC (see Hau, 2005a). Writ small, the Average Travel Time curve, ATT, in Fig. 2B could then be thought of as a typical traveler's travel time of a trip depending on the traffic flow, ATT(F), and, after being rescaled by the value of time measure, becomes the traveler's average travel time cost, that is, the Average Variable Cost of a trip, AVC(q), in Fig. 3.⁵ Since the money cost component merely shifts the AVC curve up in a parallel manner to form the generalized cost of a trip, GC, henceforth we shall set the money cost aside to avoid cluttering our diagrams.⁶ So it is the time cost element alone that drives the AVC curve. The AVC curve climbs upwards due to significant adverse interactions amongst vehicles well before the traffic flow reaches the maximum capacity, $F^{\max}$, relabeled as maximal output $q^{\max}$, in Fig. 3.⁷ The shape of the AVC curve, being driven solely by the ATT curve, comes about from an inverted speed–flow curve. Hence in actuality the AVC curve is a technology or performance curve generated by all travelers rather than a single one. The concept of "transport supply" is nonexistent in such a setting: generalized cost depends on output, GC(q), and not the other way around. Causality only holds in one direction, as distinct from a conventional supply curve in economics.

When travelers formulate their decisions on the amount of travel to undertake, as consumers they are really basing their decisions on a combination of their travel time and travel costs, or the generalized price of a trip, commonly referred to as the full price of a trip, P (Mohring (1993), Mohring (1994, Introduction)). Obeying the law of demand, aggregate travel demand is responsive to the (full) price of a trip in a standard manner, $q^D(P)$. That is, when the full price of a trip is high, travel demand is low and vice versa, as reflected by a downward-sloping demand curve as depicted in Fig. 3. The demand curve is a typical traveler consumer's travel demand for the amount of trip taking depending on the full price of a trip..

Optimal Pricing

Whether or not a traveler makes a trip is contingent on the following calculus: the marginal benefit of undertaking a trip with the cost of undertaking that incremental trip to him, which is simply the amount of travel time that he incurs *on average*, that is, the average variable cost of the trip (and *not* its marginal cost). The marginal benefit of a trip is his marginal willingness to pay for an incremental trip as reflected by the inverse demand function of $P(q)$. With the traveler basing his decision to travel on the AVC curve, and with the functional relations of GC(q) and $P(q)$, the traffic equilibrium point—depicted as point U in Fig. 3—is characterized by $P(q) = GC(q)$, with q^0 and P^0 as its solution. Since GC is simply a translation of AVC, the interpretation of one is synonymous with the other. In equilibrium, we say that $P(q) = AVC(q) (=SRAVC(q))$ for simplicity.⁸ Point U is an equilibrium because there is a built-in tendency

⁴ There is a voluminous literature on estimating the values of travel time using disaggregate behavioral travel demand models in the transport literature (Beesley (1965), McFadden (1974, 1978), Train (1980), Hau (1987), Hensher (2001)).

⁵ Thus $t^{\min}$ is the minimum travel time of a trip (see Fig. 2b). When multiplied by the value of time, the fastest travel time of a trip, $t^{\min}$, gets converted into money terms and appears as t^M (see Fig. 3). Hence the ATT and AVC curves have identical shape.

⁶ Note that the magnitude of the money cost clearly does matter but it is not germane to our analysis.

⁷ When switching to the standard output notation that economists use, the traffic flow per hour unit could be expressed as a commensurate per unit of time measure for user benefit and cost of road capacity later on.

⁸ To avoid notational clutter, the italicized P symbolizing the inverse demand function shall be replaced by the regular P symbol, with the inverse demand function interpretation clear within the context. Note that the money cost appears on both the willingness to pay side and the time cost side, canceling one another out in our framework.

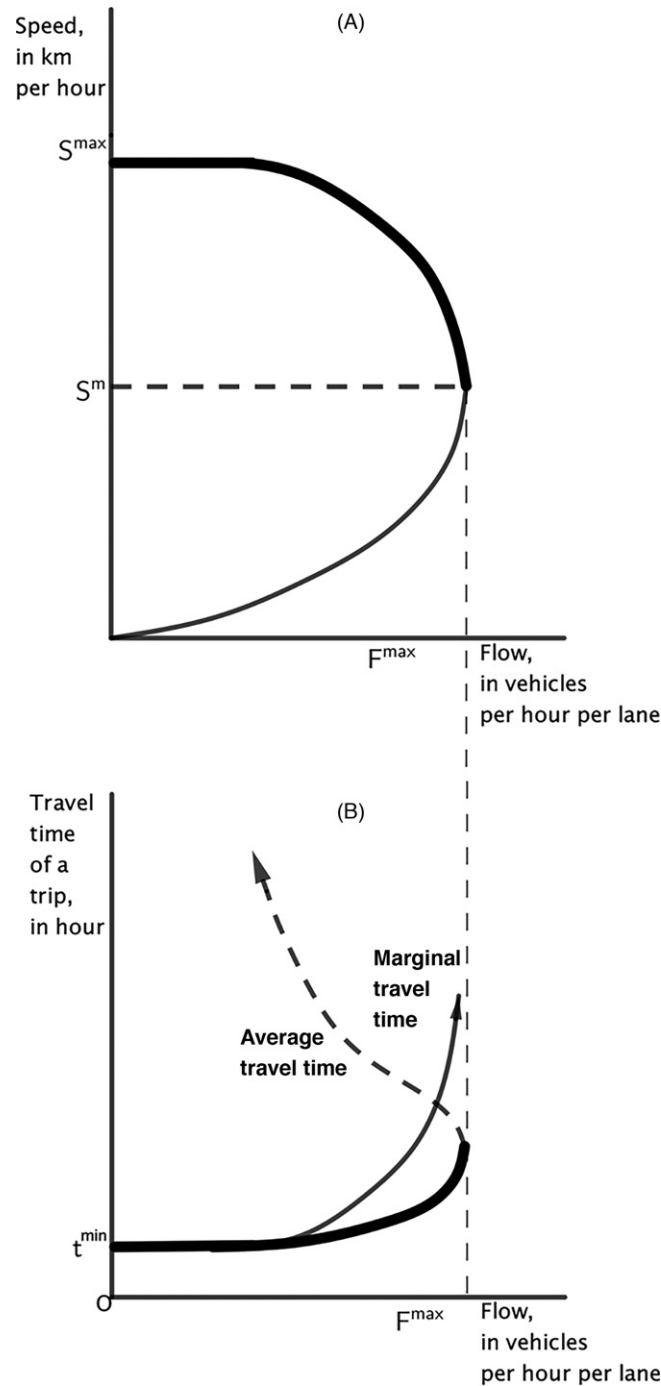


Figure 2 (A) A speed-flow curve; (B) Average travel time curve and marginal travel time curve.

for traffic to converge to that point.⁹ Thus far we have confined ourselves to considering a motorist's private costs. In the absence of congestion, private and social costs are the same. However, the social cost of one's private actions may diverge from one's private cost, due to an external effect that is left unaccounted for by the motorist.

The incremental cost to society of a motorist is that his entrance onto a road pushes up the average variable cost of a vehicle trip to everybody else. This is because the traveler is concerned only about the prevailing traffic condition as it affects himself. While he clearly accounts for the time that his own vehicle trip takes on average, the motorist is not mindful of the fact that he causes a definite delay, whilst a bit small, to each of the vehicle trips behind him when he joins a traffic stream, ending up as a nontrivial amount in

⁹ That points U and X are both stable equilibria and point W is an unstable equilibrium could be shown by perturbing the respective equilibrium using an output level slightly greater or lesser than the points U, X and W (see Hau (2005a)).

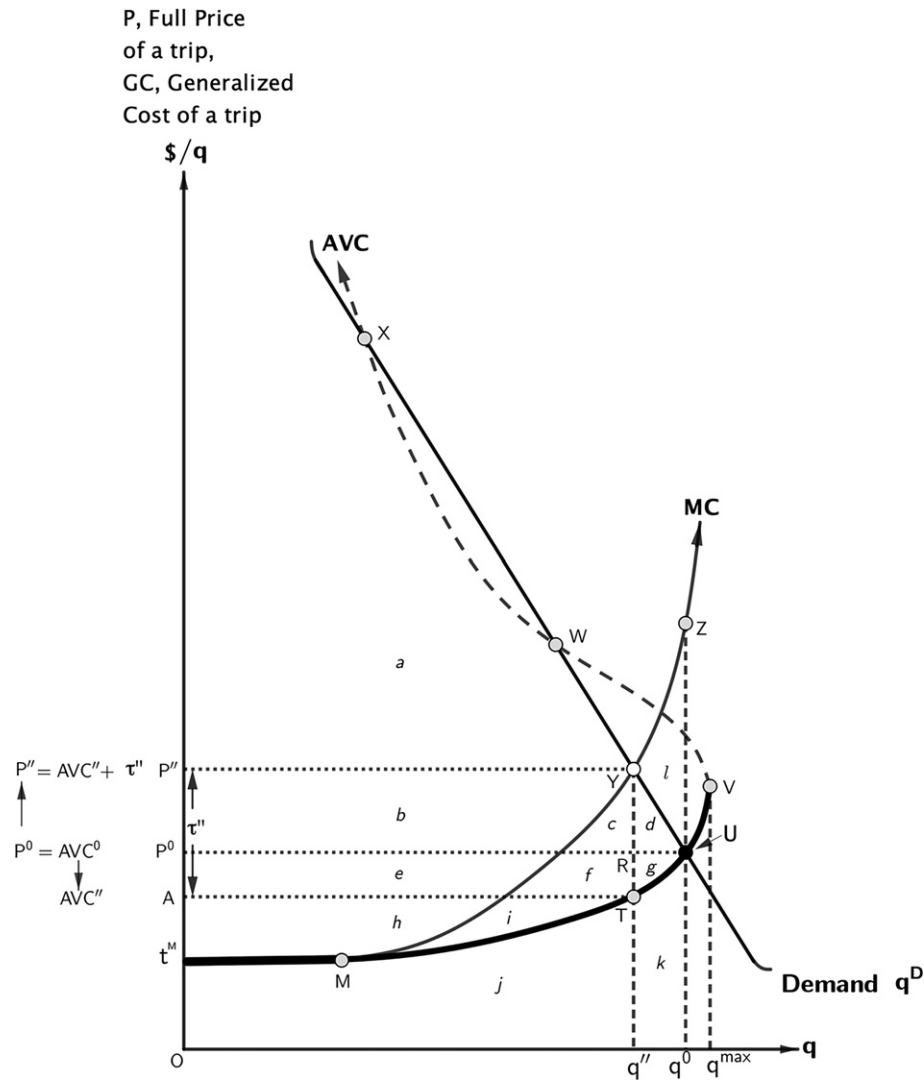


Figure 3 Welfare impact of road pricing.

the aggregate. In order to ensure that the motorist is responsible for his own behavior, the public-spirited road authority places an economic efficiency-enhancing toll on the *excess* of the marginal cost over the average variable cost at the traffic level where the travel demand meets the marginal cost curve.¹⁰ The congestion toll, shown as the distance YT in Fig. 3, is imposed to ensure equality of the

¹⁰The incremental cost or marginal cost of undertaking a vehicle trip, $MC(q)$, is: $MC(q) \equiv \frac{\Delta C(q)}{\Delta q} = \frac{\Delta TVC(q)}{\Delta q} = AVC(q) + q \cdot \frac{\Delta AVC(q)}{\Delta q} = AVC(q) \cdot (1 + \epsilon(q))$ where $\epsilon(q) \equiv \frac{\frac{\Delta AVC(q)}{\Delta q}}{\frac{AVC(q)}{q}} = \frac{\Delta AVC(q)}{\Delta q} \cdot \frac{q}{AVC(q)}$. $MC(q)$ is the Marginal Cost function, $C(q)$ is the cost function, $TVC(q)$ is the Total Variable Cost function, $AVC(q)$ is the Average Variable Cost function, is the output elasticity of the AVC function. Intuitively, marginal cost is the sum of two components: the first one is the traveler's Average Variable Cost, $AVC(q)$, and the second one is: $q \cdot \Delta AVC(q)/\Delta q$. The first term consists of the trip maker's own time cost, time in money terms converted from travel time via the valuation of travel time measure; the motorist himself incurs the time cost so we say that it is self-financed. The second term is the marginal external cost; it is the rate of change of the trip time cost to others due to his last vehicle entering the traffic stream, $\Delta AVC/\Delta q$, multiplied by the ensuing platoon of vehicles, q , that his incremental vehicle has slowed down. By multiplying and dividing the second term by AVC , the marginal external cost could be expressed as the product of the Average Variable Cost, AVC , and the output elasticity, ϵ . Because only the average variable cost, AVC , is accounted for by the trip maker himself, the external cost engendered by him – the second term – needs to be internalized by having a congestion toll required of him by the road authority. The public sector's call to pursue marginal cost pricing of a trip means that (full) price is set equal to marginal cost but with his time cost already self-financed, internalizing the externality calls for a congestion toll to be set at the *difference* of marginal cost and average variable cost. (It is on a per-unit-of-time basis just like traffic flow is.) Charging anything more than the external effect would be double-counting. Setting the two marginals equal would mean that the user benefits of highway-derived travel are maximized, given a road system. The congestion toll is also known as an optimal, economically efficient, efficiency-enhancing, net-benefit maximizing, Pareto Optimal, marginal cost or efficiency toll in the economics and transportation literature. To be valid, the toll wedge shifts the upward-sloping part of the AVC curve, i.e., the normally congested part of the AVC curve for the section from t^M leading up to point V only. Corresponding to the AVC curve, the MC curve rises asymptotically up to the capacity level of $q^{\max}$. Note that with a congestion toll, the capacity $q^{\max}$ would not be reached in equilibrium. It would be convenient to redefine the capacity $q^{\max}$ as K as different models of speed-flow relationships may not yield a backward-bending Average Travel Times curve with a capacity emerging just at the switching point from normal congestion to hypercongestion, so we can write: $AVC(q, K)$. The traffic flow on a per hour basis could then be expressed as being commensurate with a per-unit-of-time measure for the cost of road capacity and user benefit to be introduced later on. First-best pricing in the world of transport implicitly presumes that the rest of the economy is efficiently priced also.

marginal benefit of a trip and its marginal cost.¹¹ Such a toll wedge, τ' , shifts the upward-sloping AVC curve by the toll magnitude upwards in a parallel manner, resulting in a new equilibrium at point Y: $P(q) = MC(q) = AVC(q) + \tau'$. Price is driven up to P' and quantity demanded falls to q' . With the equimarginal principle satisfied, the net benefits of road use travel are maximized and indicated by the large triangular area $a + b + e + h$. Point Y is known as the Social Optimum (SO). This is called the first-best Optimal Pricing Rule, our First Optimality Rule. Once congestion pricing is introduced, not only would the final traffic level equilibrate below the initial equilibrium, q^0 , it would be at a level well below the capacity of the road, $q^{\max}$, the point at which hypercongestion sets in. The grossly inefficient backward-bending, hypercongested part of the average travel time curve in Fig. 2B and its AVC counterpart in Fig. 3 would be obviated completely under congestion pricing.

As an illustration, suppose a 10-mile trip from one place to another that takes half an hour during uncongested hours, denoted by $t^{\min}$ in Fig. 2B, and 45 min during rush hour at height Uq^0 . The motorist is fully cognizant that it would take him 45 min—a cost of which is *internal* to the driver—when traveling during the peak period but is oblivious to the fact that, consequential to his entry, he would be imposing an additional 20 min, say, of travel time, at height ZU, the *external* cost to the drivers behind him. Alternatively, if that last trip is eliminated altogether, the marginal motorist diverts his resources to other uses, including his 45 min of travel time, whilst *every* motorist staying on the road reaps a saving of travel time as a result of experiencing slightly lower traffic delay. It is the 20 min—of a trip that costs a whopping 65 min in total (at height Zq^0)—for society that is the congestion externality to be charged for in the form of a congestion charge were it to be evaluated at the initial traffic equilibrium, q^0 . Since travel demand responds to the (generalized) price at that high level, the number of vehicles on the road would be curtailed and congestion drastically reduced. The road authority should then respond accordingly and adjust the toll to a lower level in anticipation of the ensuing low traffic level and the corresponding smaller difference of marginal cost over average variable cost. After various trial-and-error iterations, the number of trips demanded could be viewed as moving in a cobweb manner, eventually converging to where demand intersects the marginal cost curve (at q'), and calling for an optimal toll-wedge: $YT = \tau'$, evaluated at the final equilibrium point q' , given the existing road capacity.¹² Note that the optimal toll, in the final analysis, would be at an intermediate level since the optimal level of congestion externality—akin to the notion of an optimal pollution level—should neither be zero nor an inordinate amount under normal travel demand conditions. In terms of either time unit or time cost, the marginal travel time delay at the final equilibrium traffic level (at q') would be less than 20 min since certain trips would be priced off the road, given a conventional travel demand that is downward sloping.¹³ In particular, the toll should be set equal to the marginal delay (of a fraction of a second, say) caused by an additional vehicle entering the roadway at the final equilibrium level multiplied by all the ensuing vehicles (at, say, 1000 vehicles per hour per lane). This results in a marginal delay of 10 min to others.¹⁴ Assuming that the value of time is HK\$60 an hour, or a dollar a minute, in our example, the numerical value of the congestion toll comes to HK\$6.¹⁵

The Welfare Impact of Congestion Pricing

Here we seek to explain the political failure of road pricing using the standard analytical tools of welfare economics. An economist might pose the question at hand as follows: why is it that the imposition of an externality-corrective tax in road transport, which we know yields a welfare gain overall to society, would end up hurting *all users*—not unlike the case of a distortionary excise tax would on the users and producers of a commodity.¹⁶ Putting it more mundanely, why would the imposition of a congestion toll fail to be accepted by the majority of people if the welfare gain is demonstrably positive?

For simplicity let us consider what happens when a congestion toll is imposed on road users in a moderately congestion situation without a preexisting toll or tax. The resulting increase in price means that certain motorists are tolled off the road. These motorists will value the reduction in those trips as the sum of the valuation, that is, the marginal willingness-to-pay, for trips between q^0 and q' , as indicated by the trapezoidal area YUq^0q' (area $d + g + k$). On the other hand, the released resources constitute the sum of the marginal costs of trips to society between q^0 and q' as shown by the larger trapezoidal area YZq^0q' (area $l + d + g + k$). At first glance, it appears that the incremental resource savings over and above the incremental benefits accrue to those users that are priced off the road when it turns out that the ones who benefit belong to those who had stayed on and paid the congestion toll. Another way of representing the change in variable costs of the q^0q' trips is shown by the inverted L-shaped area $P^0RUq^0q'TA$ (area $e + f + g + k$). The equivalence of these two simple approaches has an important economic interpretation: the savings in time costs to society from not undertaking ($q^0 - q'$) trips are broken down into two parts, with the first part, area RUq^0q' , being the (private) savings in resources accruing to those tolled off, and the second part, area RYZU, being the formerly savings in external travel time cost accruing instead to all those motorists who remain on the road: the tolled. The *bona fide* benefits from congestion pricing can be shown to be the familiar

¹¹ Note that the optimal toll is expressed on a per-unit-of-time basis so that it is commensurate with the time cost and traffic flow variables.

¹² Li (2002) uses the Drake model of speed-flow relationship to come up with the congestion toll and a value of the generalized cost even without having to specify a demand function.

¹³ p^0 is the money equivalent of 45 minutes, converted into money via the value of time, at the initial traffic level, and P' and A correspond to the travel time cost of 51 minutes and 41 minutes respectively. Note that the numerical example with round numbers are given for illustrative purpose only.

¹⁴ The numerical example implicitly assumes that the nonlinear average variable cost curve is assumed locally to be linear, with the linear demand drawn as being more unresponsive to price relative to the (technological) performance curve.

¹⁵ The official exchange rate is HK\$7.80 to US\$1.

¹⁶ Note that the results put forth here from congestion pricing do not carry over to the case of a corrective tax for pollution externality, say. This is because the curtailment of the excess amount of pollution results in savings in social cost by the group of incremental users over and above the valuation by the same group of marginal users. By contrast, with congestion pricing, the savings in external time costs released by the marginal users accrue technologically to the inframarginal users and *not* to the marginal users. It is in a sense such as this that some may find the concept of congestion pricing a bit elusive.

triangular welfare gain, area YZU (area l) or, equivalently, the rectangular area of the actual time savings accruing to the tolled, area P^0RTA (area $e + f$), less the welfare loss area YUR (area d), to those who are tolled off. As a result of curtailing the excess number of trips, q^0 less q^* , in this way, congestion is eased and travel time drops. The *incidental* time actually saved of one trip to those who remain on the road is shown by the distance RT in Fig. 3 and the aggregate time savings are valued at P^0RTA (area $e + f$), equal to \$400 in our example, while the aggregate toll payments are given by the area P^0YTA (area $b + c + e + f$), \$1000. Note that the time savings in a normally congested situation is *always* less than the toll payment (of distance YT), so that the tolled (i.e., those who chose to stay and pay) are worse off in the presence of congestion pricing than without, under the standard assumption of a uniform value of time (adopted universally by authors such as Walters, 1961, 1987, Vickrey, 1969). Were this not the case, that is, if people value their time savings greater than their toll payments, people would not be willing to switch mode or time of travel—and no decongestion would emerge! Those who are tolled off¹⁷ (i.e., those who avoided using the road to shun the toll) are clearly worse off also because they would be obliged to take a less preferred route or desired mode or time of travel, with the so-called deadweight loss shown by the triangular area YUR (area d) (see Hau (2005a, Appendix) for other derivations). Furthermore, those who are currently traveling during off-peak or on public transportation may find themselves being overcrowded—a group I call the “tolled on”—so that they too are made no better off than before. (Those on buses *could* experience a net gain in welfare if: (1) equilibrium travel times—in addition to line-haul travel times—along the transport corridor were sufficiently lowered as a result, or (2) if increased patronage were to come about due to increased frequency or service enhancement by the transit agency.) If all these groups are made no better off, who gains unambiguously most of the time? The answer is that the road authority—and *not* users—is the only party that gains unequivocally in the form of reaping the toll revenue collection (area YT q^*), which is \$1000 in our example). However, the government’s gain comes about actually because motorists who have opted to continue to travel during the peak period are required to surrender a toll payment for the right of access: total congestion toll revenue equals total expenditure. In economic jargon, the toll revenue collection is a pure “transfer payment”. The welfare gain of road pricing remains as before: the aggregate time savings of the tolled less the welfare loss to those who are tolled off. This is the genuine benefits of road pricing.

Note that implicit in this whole exercise is the common practice of adopting what at first appears to be a fairly innocuous assumption in road pricing: that the congestion toll revenues are redistributed in a “lump sum nondistortionary” fashion entirely back to the people, which in this case turns out to be the tolled.¹⁸ But adopting this seemingly innocuous assumption for public policy analysis runs counter to the general contention that the public regards road pricing as “just another tax” (as in several ill-fated Hong Kong electronic road pricing trials and other failed road pricing experiments elsewhere), as opposed to being properly regarded as a *price* for road use. Only by explicitly introducing road pricing as a *road use fee* and recycling the proceeds towards users’ direction would the tolled be made better off by the aggregate time savings area ($e + f$). If not, both the tolled and the tolled off are in fact worse off—as reflected in a diminution of their consumer’s surplus of area ($b + c + d$)—as a combined group. One should therefore not be surprised to find road users decrying so-called “congestion” pricing since their hunch is correct: that this governmental innovation is an instrument that adds insult—the tax—to their injury—being stuck in traffic. Actually, motorists generally realize that they are “not getting their money’s worth”, so to speak, especially if the congestion toll revenue collection is siphoned off to a more appealing or a higher purpose of partially paying off the interest on the national debt, say. Motorists are unlikely to perceive that the reduction in congestion is due to their behaving in an economically efficient manner when faced with the correct signaling from a road use price.

Distribution of the Values of Time

For policy analysis, we now ask whether or not the results obtained thus far based on the assumption of a constant value of time carry over when replaced by a distribution of values of time assumption. The strong conclusion reached above that in a moderately congested situation *all users* would be made worse off with the exception of the government entity needs modification:¹⁹ road users on average would be made worse off with congestion pricing, with travelers having high values of time made better off at the expense of those with low time values, who are now made worse off. In the analysis based on a distribution of values of time: (1) the magnitude of the time saved from a trip by everyone is still the same; and (2) the toll that all road users face is the same as before. Now instead of the congestion toll being computed on the basis of a constant (marginal) value of time, the latter is being replaced by a weighted average of the (marginal) values of time, with the weights being the number of trips taken by those who use the road.²⁰ Precongestion pricing, at the margin, the motorist is indifferent between taking a trip at the traffic equilibrium q^0 ; he compares his willingness-to-pay with the (average) time cost of a trip, the comparison being a straightforward one. Postcongestion pricing at the

¹⁷ These terms belong to Zettel and Carll (1964). By contrast, Hau (2005a) uses economic efficiency analysis to formally derive the welfare impact of road pricing on the tolled and the tolled off, etc., in a comprehensive short-run/long-run framework, drawing conclusions that are opposite to theirs.

¹⁸ In the foreword to the famous U.K. Ministry of Transport paper (1964) on road pricing, Reuben Smeed explicitly states that “the panel *assumed* that the introduction of road prices would be accompanied by a corresponding reduction in existing taxes so that the motoring population as a whole would pay no more than it would otherwise have done” (emphasis added). In fact, however, the road use price is converted into a ‘tax’ in that the toll revenues collected typically gets deposited into the government’s coffers and, at least as far as motorists are concerned, vanishes altogether (see the reasons for the failure of the proposed Hong Kong’s electronic road pricing pilot system of 1983-5 in Hau (1990)).

¹⁹ Weitzman (1974) was apparently the first to note that the variable factor employed would be better off under a free access equilibrium in the classic tragedy of the commons. For the case of roads, it implicitly assumes a constant value of time.

²⁰ Mohring (1976, Chapter 4 and Appendix) discusses the notion of the weighted values of time in relation to the optimization and pricing of transport services. Here I apply the concept of the weighted values of time to the welfare impact of a congestion toll: both optimal and nonoptimal.

new equilibrium q' , however, while the motorist still makes a similar comparison at the margin, but with a distribution of the values of time, the motorist who stays and pays is likely to have a higher value of time.²¹ The conclusion reached before that all parties except the government are worse off as a result of the toll holds whether the optimum toll is computed on the basis of a constant value or a distribution of values of time. Assuming that the toll revenues confer no benefit on road users, as is often perceived and perhaps justifiably so, it is clear that gainers cannot use their gains to more than compensate the losers' losses, that is, using a straightforward application of the compensation principle in welfare economics (see [Boadway and Bruce \(1984, Article 4\)](#)). To illustrate, suppose that a rich person is willing to pay much more than a dollar of toll to save a little time. His gain would be insufficient to overcompensate the rest of the tolled—supposing, as before, that the toll revenue is siphoned off or gets buried in the Treasury. This is because that optimal toll was computed on the basis of a weighted average of values of time. If the optimal traffic flow is reached and yet those with high values of time are willing to pay considerably more for time saved than the optimal toll as well as being able to overcompensate those with low values of time, then the "optimal" toll must not have been at a high enough level and could not have properly reflected the weighted average of the values of time. The impact of imposing a weighted congestion toll is in effect to raise the toll that everyone faces, so that the gains to those with high values of time are obtained at the expense of those with low time values.²² The law of demand must be obeyed.

Note that the *incidental* time saved (of 4 min) on average that comes about by charging the *marginal* externality (of 10 min or, rather, its money equivalent) is always less than the latter because of the downward sloping nature of the demand curve. We ask therefore, at the aggregate or average level as opposed to operating at the margin, whether it is possible for the rich to overcompensate the poor and still be better off. The answer is in the negative because while some of those with high time values are willing to pay more than the weighted congestion toll for time saved (which for them the toll payment amounts to only a fraction of the value of time savings), those with low time values who are not willing to pay the average toll payment for the average amount of time saved still have to make payment (because their willingness to pay for a trip exceeds the final equilibrium price inclusive of the congestion toll, P'). Thus, on average and in the aggregate, the gainers could not overcompensate the losers when faced with a weighted average congestion toll. Do bear in mind that here we are confining ourselves to the tolled and not considering those who have already been tolled off, who are clearly worse off by revealed preference.

A follow-up question is whether or not the aggregate value of time savings to the tolled would be sufficient to more than compensate the welfare loss of the tolled off, at least in a hypothetical compensation sense. After all, an in-kind transfer of this kind could not be feasibly carried out because one person's time saved could not possibly become someone else's gain in time units. The answer is clearly in the affirmative since the amount of actual time savings comes about by curtailing the excess number of trips: $q^0 - q'$. The welfare loss is only that part of the loss in valuation of those trips to the tolled off that is over and above the tolled off's own time costs (of area $g + k$). Putting it another way, area $e + f$ is always greater than area d .

Before leaving the point that road pricing hurts most parties except the government, it is important to point out that there are cases—one of which is the case of hypercongestion—analyzed in [Hau \(2005a, Fig. 6\)](#) which shows that road pricing benefits *everyone*: all users as well as the government entity. This occurs when traffic density is so high (at the point V and beyond) that the rate at which one more vehicle entering an area is greater than the rate at which one more vehicle leaving it. The resultant throughput is lower, yielding the backward-bending portion of the average variable cost curve. The initial traffic equilibrium would be observed at the intersection point of a high demand curve and a point on the backward-bending portion of the average variable cost curve. Implementing congestion pricing in traffic bottlenecks such as this one proves most promising since everyone is made better off: a clear case of a Pareto improvement. In other words, hypercongestion yields a truly win-win situation even without compensation ([Hau \(2005a, Fig. 6\)](#)). While it is empirically testable but debatable whether the duration of these bottlenecks is persistent enough in particular cases to justify promoting congestion pricing principally on these grounds, the rationale seems clear in many of the more seriously congested situations as seen in many megacities worldwide. Even so, hypercongestion is not generally the case on which road pricing proponents have built their case.²³

Congestion Toll as a User Fee

A question that stems from the above analysis is whether or not there is a theoretical basis for *actual* compensation, as opposed to the *hypothetical* compensation principle of treating "a dollar as a dollar to whomsoever it accrues to" implicit in the evaluation of the welfare gain from congestion pricing. The answer is "yes" if one is not confined to the notion of marginal cost pricing in the short run *per se*.

The following article applies these principles to the problem of identifying an equilibrium for traffic allocation to roads in the short- and long run.

²¹ At the margin, the motorist is indifferent between driving and not driving: driving involves incurring the time cost of distance Tq' (i.e., travel time) plus the toll payment of distance YT whereas not driving involves saving his (private) resources of Tq' and toll expense of YT by forgoing any benefits from taking the trip. If the motorist were to undertake the last trip at the margin, other motorists in the traffic stream would bear the amount of YT in terms of time delay (which equals 10 minutes in our example above) and the road authority would reap the gain in revenue from the (unit) toll payment. By forgoing this last trip, the motorist actually yields time savings for everyone who continues to use the road.

²² One might wonder whether the rich or urgent users, broadly speaking, would be able to overcompensate the poor or more leisurely users when a small nonoptimal toll is implemented at the observed output level, q^0 . It is possible to imagine that under such a scenario the gainers would be able to more than compensate the losers since no constraint would be placed on the numerical value of the nonoptimal toll. For a very small nonoptimal toll, indeed the loss to the tolled off will be extremely small; the average value of time savings of those remaining is likely to far exceed the toll in a very congested situation.

²³ William Vickrey, by contrast, strongly advocates using hypercongestion to build the case for road pricing.

References

- American Association of State Highway Officials (AASHO), 1960. "Road User Benefits for Highway Improvements," Committee on Planning and Design Policies, Washington, D.C., pp. 1–152.
- Beckmann, M., McGuire F. C.B., Winsten, C.B., 1956. *Studies in the Economics of Transportation*. Yale University Press, New Haven, Connecticut.
- Boadway, R., Bruce, N., 1984. *Welfare Economics*. Basil Blackwell, Oxford, England.
- Dupuit, J., 1844, "On the Measurement of the Utility of Public Works," *Annales des Ponts et Chaussées*, 2nd series, Vol. 8, 1844, translated from the French essay "De la mesure de l'utilité des travaux publics," by R. H. Barback for *International Economic Papers*, No. 2, 1952, pp. 83–110. Reprinted in Denys Munby, ed., *Transport, Penguin Modern Economics*, London, 1968, pp. 19–57.
- Dupuit, J., 1849, "On Tolls and Transport Charges," *Annales des Ponts et Chaussees*, 2nd series, 4th part, 1849, pp. 207–248, translated from the French essay "De l'influence des peages sur l'utilité des voies de communication," by Elizabeth Henderson for *International Economic Papers*, No. 11, 1962, pp. 7–31.
- Gerlough, D.L., Huber, M.J., 1975, *Traffic Flow Theory*, A Monograph, Special Report 165, Transportation Research Board, National Research Council, Washington, D.C.
- Greenshields, B.D., 1935, "A Study of Highway Capacity," *Proceedings Highway Research Record*, Washington D.C., Volume 14, pp. 448–477.
- Haight, F., 1963. *Mathematical Theories of Traffic Flow*. Academic Press, New York.
- Hau, T.D., 1990. Electronic road pricing: developments in Hong Kong 1983–89. *J. Transport Econ. Policy* 24 (2), 203–214.
- Hau, T.D., 2005a. Economic fundamentals of road pricing: a diagrammatic analysis, part i – fundamentals. *Transportmetrica* 1 (2), 81–117.
- Hensher, D.A., 2001. The valuation of commuter travel time savings for car drivers: Evaluating alternative model specifications. *Transportation* 28 (2), 101–118.
- Mohring, H.D., 1976. *Transportation Economics*. Ballinger Press, Cambridge, MA.
- Mohring, H., 1993. Maximizing, measuring, and not double counting transportation-improvement benefits: a primer on closed- and open-economy cost-benefit analysis. *Transport. Res. Part B Methodol.* 27 (6), 413–424.
- Small, K.A., Verhoef, E.T., 2007. *The Economics of Urban Transportation*, 2nd ed. Routledge, New York.
- Transportation Research Board, 2011. "75 Years of the Fundamental Diagram for Traffic Flow Theory," *Greenshields Symposium*, July 8–10, 2008, Woods Hole, Massachusetts, Traffic Flow Theory and Characteristics Committee, Transportation Research Circular E-C149, Washington, D.C., June.
- Transportation Research Board, 2016. "Highway Capacity Manual, Sixth Edition: A Guide for Multimodal Mobility Analysis," June, The National Academies of Sciences, Washington D.C.
- Vickrey, W.S., 1955. A proposal for revising New York's subway fare structure. *J. Oper. Res. Soc. Am.* 3 (1), 38–68, Reprinted in Richard Arnott, Kenneth Arrow, Anthony B. Atkinson and Jacques H. Dreze, eds., *Public Economics: Selected Papers by William Vickrey*, Cambridge University Press, Cambridge, England, 1994, pp. 277–306..
- Vickrey, W.S., 1969. Congestion theory and transport investment. *Am. Econ. Rev.* 59 (2), 251–260, Reprinted in Richard Arnott, Kenneth Arrow, Anthony B. Atkinson and Jacques H. Dreze, eds., *Public Economics: Selected Papers by William Vickrey*, Cambridge University Press, Cambridge, England, 1994, pp. 320–322..
- Walters, A.A., 1954. Track costs and motor taxation. *J. Ind. Econ.* 2 (2), 135–146.
- Walters, A.A., 1961. The theory and measurement of private and social cost of highway congestion. *Econometrica* 29 (4), 676–699, Reprinted in Herbert Mohring, ed., *The Economics of Transport*, Vol. I, The International Library of Critical Writings in Economics 34, An Elgar Reference Collection, Edward Elgar, Aldershot, Hants, pp. 24–47..
- Walters, A.A., 1968. *The Economics of Road User Charges*, World Bank Occasional Paper Number 5, International Bank for Reconstruction and Development, Johns Hopkins University Press, Baltimore, Maryland.
- Walters, A.A., 1987. Congestion. In: John Eatwell, Murray Milgate, Peter Newman, (Eds.), *The New Palgrave: A Dictionary of Economics*, Vol. 1. The Macmillan Press Limited, London, pp. 570–573.
- Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers PART II* 1, 325–378, Reprinted in Herbert Mohring, ed., *The Economics of Transport*, Vol. I, The International Library of Critical Writings in Economics 34, An Elgar Reference Collection, Edward Elgar, Aldershot, Hants, 1994, pp. 299–352..
- Wong, Wai, Wong, S.C., 2016. Network topological effects on the macroscopic bureau of public roads function. *Transportmetrica A* 12 (3), 272–296.
- Zettel, R.M., Carll, R.R., 1964. "The Basic Theory of Efficiency Tolls: The Tolloed, The Tolloed-Off and The Un-Tolloed". *Highway Research Record*, No. 47, paper presented at the 43rd Annual Meeting of the Highway Research Board, January 13–17, 1964, National Research Council, Washington, D.C., pp. 46–65.

Road Pricing 2: Short- and Long-Run Equilibrium of Road Transportation

Timothy D. Hau, HKU Business School, The University of Hong Kong, Hong Kong

© 2021 Elsevier Ltd. All rights reserved.

Introduction	83
Short-Run Equilibrium of Road Transportation	83
Long-Run Equilibrium of Road Transportation	84
Optimal Capacity	84
Relaxation of Assumptions	88
References	89
Further Reading	89

Introduction

In the previous article the theory of congestion pricing was developed to show how the application of road pricing could control congestion and achieve a welfare improvement. In this article, we develop the theory to show the equilibrium conditions for road traffic in the short- and long-run and subsequently show the effect of relaxing some of the assumptions.

Short-Run Equilibrium of Road Transportation

Thus, far we have characterized the nature of traffic congestion and discussed how its excess could be curtailed. Just as a truck whose axle weight damages the road—leaving behind potholes that raise vehicle operating costs and cause delays to other motorists—ought to be charged for the road damage inflicted, so also should a responsible road authority charge for the incremental external cost by equating the marginal cost of a trip to society to its marginal benefit. The rule is quite straightforward, and that is to always charge for the marginal external cost that a motorist rationally ignores, this regardless of the road infrastructure capacity. Given a particular level of travel demand, marginal cost pricing would call for a high congestion toll for a road with low capacity vis-à-vis a high one. One might ask, would not withholding investment in road capacity result in a congestion toll level considered by motorists to be too pricey? Posing the question this way suggests that an optimal investment in road capacity level exists: overinvestment in road capacity would possibly result in minimal congestion and a very low congestion toll. In contrast, underinvestment in road capacity would result in excessive congestion, calling forth an inordinate toll. In point of fact, up until now we had implicitly assumed a built-up road system that motorists utilize, as represented by adopting a given level of road capacity, say, at K' (see footnote 10 in the previous article): we are effectively in the short run. The short run is a time period during which some input, such as the capital input, could not be varied: the road, or associated transport infrastructure, is regarded as fixed exogenously at any point of time. A socially responsible road authority has to consider the fixed (opportunity) cost of constructing a road. The opportunity cost is the highest-valued use of the resource elsewhere when used here. In the absence of noncompetitive and external effects in road construction, the market price of the capital input used would be a good measure of the opportunity cost, which would include imputed interest on invested capital. As the fixed cost is a lump sum of money expenditure independent of output, the per-unit-of-time fixed cost of construction is depicted as a rectangular hyperbola in Fig. 1. The short-run average fixed cost (SRAFC), for a road with a given capacity, K' , is denoted by $SRAFC(K')$. It would be socially irresponsible for the road authority to merely minimize its dollar cost of road construction and provide a road with minimal capacity. To be socially accountable, the road authority as a producer of road capital needs to consider also the variable input supplied by its user: the time cost of the traveler. Thus the road authority sums up its per-unit-of-time construction cost, $SRAFC(K')$, with the traveler's (short-run) average variable cost, $SRAVC(K')$, to form the short-run average total cost, $SRATC(K')$, curve as shown in Fig. 1.

Here it comes in handy to view the traveler's dual role as a consumer–demander and producer–supplier of a finished trip. In order to decide whether or not to expand road capacity, the road authority compares the return on its facility, that is, the per-unit-of-time quasi-rent, with its per-unit-of-time capital cost. The quasi-rent—the difference of the full price of a trip less its average variable cost—turns out to be captured by the road authority via the instrument of a congestion toll.¹ As (economic) rent is a slippery concept, we can approach the problem of optimal road capacity provision by looking at the other side of the same coin: as a producer of road capital, the “revenue” received from the consumer as demander less the consumer as supplier of travel time input, in money terms, and the opportunity cost of road construction is the “economic profit”. The residual of quasi-rent less the opportunity cost is the economic profit, or, what amounts to the same thing, the economic profit is the per-unit-of-time optimal toll less the per-unit-of-

¹ ‘Rent’ is a payment to a factor of production over and above that which is necessary to draw forth its services and ‘quasi-’ refers to the fact that, if it earned no compensation, the flow of services from the fixed capital would be provided only until the road capital asset is depreciated (see Hau (2005a) and Mohring (1976, Chapter2)). Quasi-rent is the analog of producer's surplus in the input market.

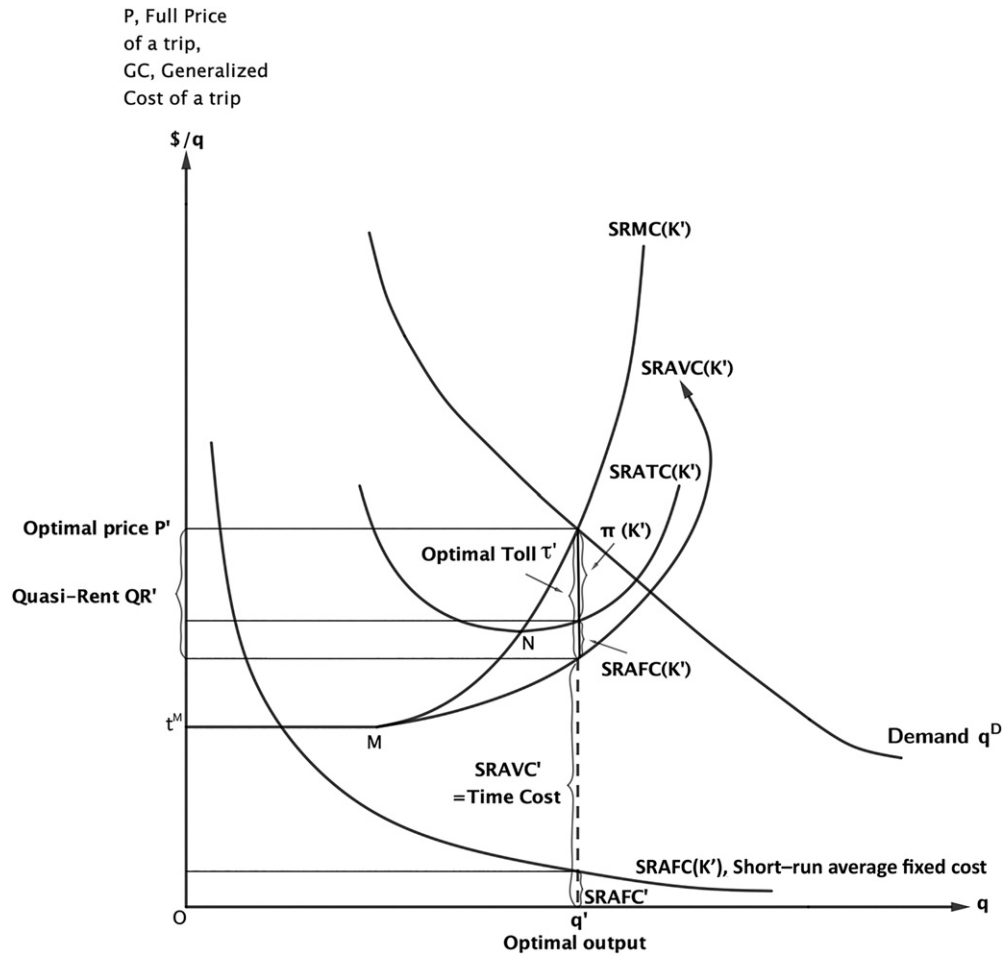


Figure 1 Short-run equilibrium of a quasi-market road authority.

time fixed cost of road capital, starkly shown in **Fig. 1**. With the level of road capital stock, K' , being an arbitrarily chosen one, the equilibrium of a road authority behaving in an efficiency-enhancing competitive manner in the short run is shown as the short-run equilibrium of a competitive, or quasi-market, rather, road authority in **Fig. 1**. Given any level of road capacity such as K' , the congestion toll is the difference of $SRMC(K')$ and $SRAVC(K')$: τ' .

Long-Run Equilibrium of Road Transportation

We now ask, would a road agency behaving competitively by mimicking such economic efficiency, efficiency-enhancing pricing principles discussed thus far be able to sustain itself over the long haul? That is, does the “competitive” road agency’s revenues cover its cost of road provision in the long run? The answer is “yes” under certain conditions. When the road agency’s economic profit is positive, pursuing economic efficiency calls for the expansion of the capital stock of a road. *Per contra*, when the congestion toll revenue is less than the opportunity cost of road capital, the capital stock of the road should be lessened. Varying road capacity from a road initially with nonoptimal capacity, K' , to one that has optimal capacity, K^* , is precisely the process by which a competitive road transport agency would arrive at in the long run when road capacity is optimized.

Optimal Capacity

How the road agency arrives in long-run equilibrium relies on the correct signal being emitted by optimally setting the full price of transportation to the short-run marginal cost of a trip for any given road capacity. This is how the quasi-market mechanism works: when economic profit is positive, starting off with an arbitrary, nonoptimal road capacity level that is low, say, K' (as shown in **Fig. 1**), the capacity of the road should be enlarged and the road partially decongested until economic profit is eroded. If so, the corresponding optimal price would be lowered from P' to P^* , output would rise from q' to q^* , the minimum level of traffic before congestion sets in would rise from $t^M M$ to $t^M M^*$. When economic profit is nil, the level of capacity is optimized, at an expanded level

of capacity at K^* . The minimum of $SRATC(K^*)$ would be achieved at N^* after a proportionate increase in capacity to K^* . At zero economic profit, the per-unit-of-time quasi-rent, QR^* , is equal to the per-unit-of-time opportunity cost of road construction, $SRAFC(K^*)$, that is, the optimal toll, τ^* , covers the per-unit-of-time cost of optimized road capital, $SRAFC^*$, as plainly shown in Fig. 1. The total congestion toll revenue collection for the optimized road capacity K^* would exactly cover the total opportunity cost of road construction. Equivalently, when economic profit is nil, long run equilibrium is reached. This path-breaking finding is known as the Mohring–Harwitz theorem, selffinancing or cost-recovery theorem. For the socially amenable road authority behaving in such a quasi-competitive market, note that this result would come about under stringent but plausible conditions of constant returns to scale.² When all inputs are proportionately increased, the same applies to output and costs also, resulting in a horizontal long-run average total cost (LRATC) that wraps itself as an envelope around all the minimal points of $SRATC(K')$ and $SRATC(K^*)$. Hence the long-run marginal cost (LRMC), is flat too and coincides with the LRATC curve (see Fig. 2).³ At whatever desired level of output a road with a particular capacity will have an associated SRMC curve associated with it that crosses the LRMC curve.

We ask what equimarginal principle is met when we try to determine the optimal road capital stock. Our objective now has turned to maximizing the net benefits of travel over the long run. Beforehand, with road capacity fixed at any given level, short-run net benefit maximization calls for the road authority's imposition of a congestion toll *wedge* when applying marginal cost pricing. In the long run, highway capacity can be relaxed and K varied as time passes. Since the benefits of travel do not depend on the capacity *per se*, long-run net benefit maximization implies cost minimization. The road authority's task is straightforward: jointly minimize the traveler's time cost and the road authority's capital cost of road capacity, KC . Given an arbitrary output level, the road authority aims to vary road capacity to the point where the marginal benefit of expanding capacity in the form of marginal savings in external time cost is equal to the marginal cost of road capacity.⁴ This is the Optimal Capacity Rule, our Second Optimality Rule. The net benefits of travel are not only maximized at any point in time, they are maximized in the long run by varying the road capacity level even if we had started off with a nonoptimal capital stock, K' . The per-unit-of-time quasi-rent earned from society's fixed investment in road capacity, QR^* , as captured by the per-unit-of-time optimal congestion toll, τ^* , is the same as the per-unit-of-time opportunity cost of road capital, $SRAFC^*$ (see Fig. 2). When overall constant returns is satisfied by a combination of the twin conditions of constant returns, the quasi-market mechanism of congestion tolling says that the size of road capacity should be adjusted to the point where optimal quasi-rent and average fixed cost is the same: $QR^* = SRAFC^*$, both on a compatible unit basis. This means that the total revenues from congestion tolling just cover the total economic cost of road construction (Fig. 2). This is the celebrated selffinancing or cost-recovery theorem advanced by Mohring and Harwitz, which has spawned half a century of research in transportation economics. The model is generalizable to all congestion-prone transport infrastructure networks such as airports and ports, for instance (Bennathan and Walters, 1979; Mohring, 1984; Morrison, 1983; Verhoef and Mohring, 2009; Zhang and Czerny, 2012).⁵

²To obtain the famous selffinancing result in roads, it turns out that two technical assumptions about constant returns are needed (Mohring and Harwitz (1962, Chapter 2)): Constant returns to scale in road construction (i.e., output is doubled when all inputs are doubled). We assume that the construction cost of road capacity, KC , defined to include capital, depreciation and "invariant" maintenance costs, is directly proportional to the capacity of the road alone, K , i.e., $KC(K) = aK$, where a is a constant, following Keeler and Small (1977). ("Invariant" maintenance costs are those nontraffic-dependent maintenance costs due to Walters (1968, p. 23).) The lane-width, treated continuously, could be viewed as a simple measure of capacity. Doubling road capacity, K , results in the construction cost being doubled. Technically, KC is homogeneous of degree one in road capacity. Diagrammatically, the rectangular area of road construction and invariant maintenance costs is doubled. When all inputs are doubled and output is doubled, we say there is constant returns to scale, and, if cost is doubled also, we say there is neutral or zero economies of scale. When the input market is assumed to be competitive, constant returns to scale in construction is synonymous with neutral economies of scale in construction. i) Constant returns to scale in road use (i.e., the travel time cost function remains unchanged when inputs and output are doubled). Clearly the time cost function, $AVC(q, K)$, rises with output but falls with capacity, *ceteris paribus*. Technically, AVC is homogeneous of degree zero in output and capacity. Diagrammatically, when capacity and output are doubled, the rectangular area of time cost double, whereas the per-unit-of-time time cost remains the same. Constant returns to scale in road use is synonymous with neutral economies of scale in road use. This technical condition could be achieved, for instance, by usually specifying that the travel time function depends on the volume-capacity ratio, q/K , as follows: $AVC(q/K)$. This condition is also known as constant returns to scale in congestion technology (Small (1992, Chapter 3)). e had argued that the (average) vehicle operating cost is independent of output. We further assume that the variable road maintenance cost too is independent of output. Together with the construction cost and invariant maintenance cost being proportional to capacity expansion, the average total cost could also be written as: $ATC(q, K) = ATC(q/K)$. These two mathematical conditions are vital to Mohring and Harwitz's theorem (1962, pp. 85–90). In actuality, it is the combination of both conditions that yields the exact selffinancing theorem. Assuming input prices remain the same, the condition of constant returns to scale is synonymous with neutral economies of scale. Small (1999) shows that the one-to-one correspondence between the production function and the cost function no longer holds when there is a rising supply curve for a capital input, say. We implicitly assume that prices in the input market, like the output market, are determined competitively, which allows us to use the production function and the cost function interchangeably.

³Pictorially, N and N^* are of the same height in Figs. 5 and 6. This comes about because constant returns imply a horizontal LRAC which envelops all SRAC curves regardless of the capital stock.

⁴Formally, cost minimization yields the following expression: $-q \frac{\Delta AVC(q, K)}{\Delta K} = \frac{\Delta KC(K)}{\Delta K}$ his says that the marginal decrease in travel time due to an incremental increase in capacity, $\Delta AVC(q, K)/\Delta K$, would have to be multiplied by all the vehicle trips behind it, q . With the negative sign, the inverse relationship between AVC and K turns the entire expression into a positive magnitude on the left-hand side. Intuitively, the left-hand side gives the marginal benefit of increasing capacity. The right-hand side gives the marginal increase in the capital cost, $\Delta KC(K)/\Delta K$, due to an incremental increase in capacity. This is the first-best Optimal Capacity Rule. Here capital cost includes both the cost of road capital and the invariant maintenance cost of road capacity. The capital cost, KC , of road capacity is in per-unit-of-time dollar terms, $KC(K)$.

⁵The Mohring–Harwitz theorem also reconciles the long-standing debate of short-run marginal cost pricing versus long-run marginal cost pricing at the World Bank. In the political economy arena, proponents who favor using pricing to tackle congestion relief favor short-run marginal cost pricing or congestion charging whereas supporters concerned about the cost recovery of a road agency in an increasingly tight fiscal environment favor long-run marginal cost pricing as all costs, including capital cost, are charged. Both views are in fact reconciled here. The correct dictum is to pursue short-run marginal cost pricing over the long run, first in the short run and then in the long run. Society's welfare would thereby be maximized. Pursuing long run marginal cost pricing under constant returns would be charging the same price regardless of the level of demand; it is in effect implemented by a fuel tax set to cover all road costs. Under this latter scenario congestion externalities would be excessive and road use inefficiently utilized.

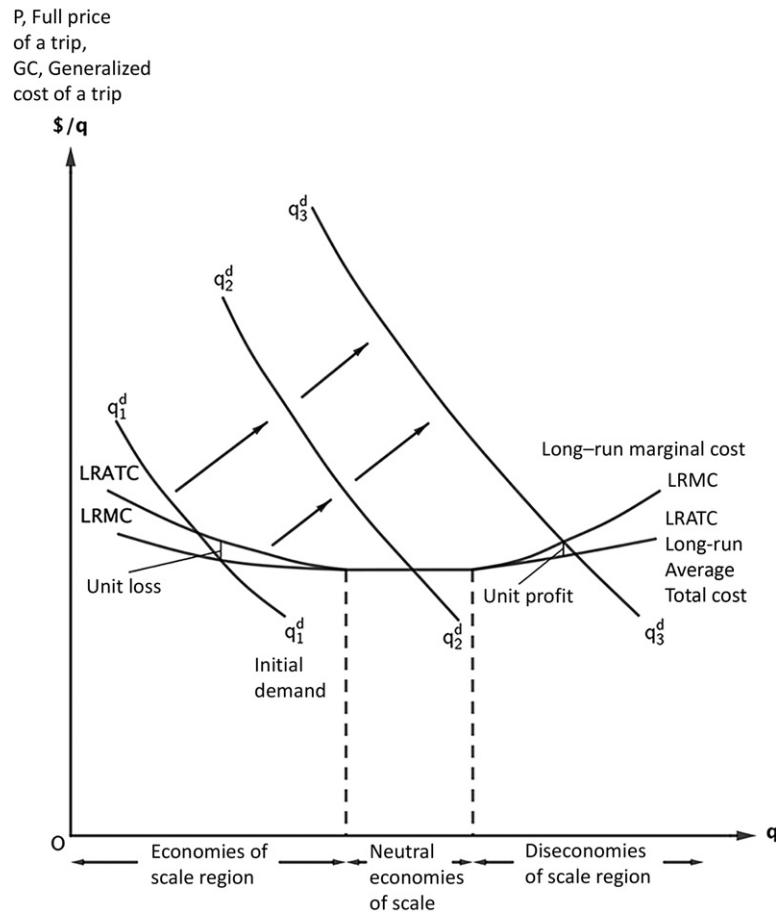


Figure 3 Economies and diseconomies of scale in the provision of road capacity with the growth of travel demand.

demand of q_3^d , the diseconomies of scale kick in. When traffic is low, a two-lane road is built. Increasing returns to scale due to (engineering) capacity clearly comes about when going from a two-lane road to a four-lane one. Beyond six or eight lanes, decreasing returns are likely to set in. In between is the case of constant returns to scale. Another way of looking at the three stages of traffic growth is to just treat each stage as a separate scenario independent of one another.

Consider first the case of decreasing returns of scale, for which the demand is high, at q_3^d , and where the long run marginal cost rises above its corresponding average cost curve. This scenario could come about because input is being bid up, resulting in a rising opportunity cost of capital and a divergence of marginal and average cost even in the long run. Pursuing short-run marginal cost pricing over the long run would definitely result in an economic profit. Urban roads with heavy traffic would characterize this region of diseconomies. Because downtown areas have high imputed land rents leading to high capital cost of roads, congestion pricing would mean that economic profits would be generated due to the increasing divergence of the pair of long run cost curves. Note that the optimal capacity chosen in the long run would be that capital stock that is less than the capital stock associated with the optimal output (see Hau (2005a, Fig. 4)). Writ large, decreasing returns to scale is likely to prevail not only with an urban road but with an urban road network. For instance, as the number of roads is doubled, the number of intersections and hence waiting times is quadrupled, a point attributable to William Vickrey (Mohring (1976, article 12), Hau (2005b, Fig. 3)). Either installing more traffic lights or building overpasses and interchanges would contribute toward bidding up the opportunity cost of land, a key ingredient in the cost of road capital, resulting in both rising average and marginal costs in the long run as well as their associated congestion tolls. These economic profits could naturally be used to finance the improvement of an alternative, being a cash-strapped public transport system, as those on public transport—both the tolled off and the tolled on—whose welfare would be well positioned to be improved.

Under the condition of increasing returns to scale or economies of scale, pursuing short-run marginal cost pricing over the long run would involve persistently incurring economic losses. Rural roads with low traffic would characterize this region of scale economies. Deficits would most likely prevail due to the technical condition of increasing returns to scale. The case of increasing returns is merely a case of insufficient demand, a point clearly seen in the economies of scale region where there is low traffic vis-à-vis the road capacity. The presence of scale economies would naturally beckon the road authority to seek economically efficient subsidization out of the public treasury. Before an optimal per unit subsidy is sought from the authority, a benefit-cost test needs

to be applied here, that is, the total benefit from travel must exceed the total cost from travel evaluated at the efficient output level, by adopting the standard compensation principle of treating a dollar of benefit or cost being a dollar regardless of to whomsoever it accrues to. Evaluated at the efficient output level,⁷ the optimal per unit subsidy is set at the difference of the long-run average cost over the LRMC to cover the per unit economic loss in Fig. 3 (see Hau (2005a, Fig. 5).)

Relaxation of Assumptions

The Mohring–Harwitz theorem is arrived at by assuming that the road authority mimics the workings of a competitive market by pursuing optimal pricing of road use and achieving optimal capacity of the road. By taking road capacity as given in the short run, road use is optimized by short-run marginal cost pricing. By optimizing road capacity in the long run, road use is doubly optimized. In so doing the sum of consumer's surplus and producer's surplus of the community being maximized would be achieved. The innovation of the economic approach adopted here, due to Professor Herbert Mohring (1994, Introduction), is to view the buyer of a transport service—the demander—as playing a dual consuming and producing role, bringing transport into the mainstream of microeconomic analysis and casting the use of road transport in a standard economics short-run and long-run framework for any good or service.

Besides considering all three cases of returns to scale, some simplifying assumptions of the Mohring–Harwitz model could be further relaxed. If the road authority mistakenly imposes a balanced-budget restriction when there are in fact increasing or decreasing returns to scale in road capacity, it still turns out that the selffinancing result is quite robust (Verhoef and Mohring, 2009). When land input price is considered when demand for road construction increases in an urban area, the exact selffinancing result still holds (Small, 1999). Verhoef and Mohring (2009) consider the robustness of the selffinancing theorem with various second-best scenarios using a simulation model.

1. *Static demand and time-varying demand:* It could be shown that the more realistic time-varying travel demand patterns that occur throughout the day—as opposed to the static one-period case—could be handled by varying the tolls optimally over time (Arnott and Kraus, 1998). Aggregating the time-of-day tolls over the day, month, and year would necessitate taking the present discounted value of the captured quasi-rents and comparing costs on the same per-unit-of-time basis.⁸ With growing traffic under constant returns to scale, the static selffinancing result is shown also to work, as shown by Arnott and Kraus (1998).
2. *Single road and road network:* When going from a single road to a full network, the selffinancing result holds for every individual road link in a road network, provided that all parts of the network are optimally priced (Yang and Meng, 2002).
3. *Variability of road thickness and road maintenance cost:* Heavy vehicles such as trucks and buses damage roads. When an additional attribute of a road—its thickness—is considered, its optimal durability needs to be considered. The optimal durability rule, a third rule, says that a road is optimally strengthened by investing up to the point where the marginal cost of durability is the same as the savings in vehicle operating costs to other motorists plus the associated savings in maintenance costs due to a thicker pavement (Hau, 2005b). When the road durability choice and road maintenance cost are considered, the selffinancing result still goes through. Under the conditions of constant returns to scale and road use, the efficient congestion toll will yield revenues that cover the capital cost of the highway but only the invariable (or nonallocable) portion of road maintenance cost attributable to weather (Newbery, 1989, Proposition 1). It then follows that the optimal road user charge, that is, the optimal congestion toll-cum-road damage charge, will then yield revenues that cover the capital cost of the road and the total road maintenance cost in a constant returns world (Newbery, 1989, Proposition 2). In other words, the optimal congestion toll covers both the capital cost of the road and the nonallocable fraction of road maintenance, whereas the allocable fraction of road maintenance is still chargeable to traffic loadings via the average variable road maintenance cost component per equivalent standard axle load (ESAL). In this way, the total expenditure on road maintenance is fully recoverable.
4. *Divisibility of road infrastructure:* Far from being perfectly divisible, in reality indivisibility is present in roads. A road requires a minimum lane-width of 12 feet, for instance, to accommodate an automobile, and a two-lane road for bi-directional travel. Coupled with the need for shoulders and drainage ditches, a significant proportion of road provision involves dead space. It turns out that indivisibility complicates the consideration of scale economies and economic losses.

With almost perfect divisibility, decreasing returns to scale would still yield economic profits at all output levels. With indivisibility, the upward-sloping LRMC hosts a series of finite, short-run marginal cost curves and is saw-toothed-shaped with the switchover (kinked) points occurring where the short-run average total cost curves cross one another (Hau (2005b, Fig. 4)). Any output level whose intersection of the short-run marginal cost curves (which cut the LRMC curve) greater than the corresponding short-run average cost curve would yield profitability and else not. However, the extent of profitability is lessened the greater the indivisibility due to the falling portions in the saw-toothed long-run average cost curve, forcing such roads to experience swings of losses and profits.

⁷ Applying short-run marginal cost pricing over the long run, we set the full price of a trip, P , to long-run marginal cost. The efficient level of output is found at the intersection of the demand curve q_1^d, q_1^d and the corresponding long run marginal cost curve. To be considered economically viable, the total benefit must exceed the total cost of travel at the efficient output level where the optimal capital stock is chosen. Note that there is excess capacity as the optimal capital stock selected should be higher.

⁸ Hau et al. (2011) estimate efficient time-varying tolls for Hong Kong's only three tunnels traversing the Victoria Harbour using hourly data from the Transport Department.

In contrast, with almost perfect divisibility, increasing returns to scale would still yield economic losses everywhere. With increasing returns, by symmetry; however, the loss regions are lessened the greater the extent of indivisibility. Just as large fixed costs of construction contribute to periodic losses, a steady increase in travel demand vis-à-vis the prevailing road capacity would contribute to periodic gains even in the presence of increasing returns. Hence the presence of indivisibility results in roads experiencing swings from losses to profits regardless of the extent of scale economies. The standard result of decreasing returns to scale being synonymous with profitability and increasing returns to scale corresponding with losses does not hold when indivisibility is taken into account (see [Hau \(2005b, Figs. 4 and 5\)](#)).

Empirically, [Keeler and Small \(1977\)](#) found that constant returns to scale exists in the San Francisco Bay Area. In *The Economics of Urban Transportation*, [Small and Verhoef \(2007, p. 112\)](#) conclude that:

“Altogether, the evidence supports the likelihood of mild scale economies for the overall highway network in major cities. Scale economies are probably substantial in small cities in which one or two major expressways are important, and may disappear altogether in very large cities where expanding expressways is extraordinarily expensive due to high urban density.”

Still, on top of the high opportunity cost of land in urban areas leading to a high cost of road capital, Vickrey's reasoning that there are considerable diseconomies of scale associated with an urban road network based on the geometry of road network would likely yield an urban road network that is economic profit-making.

References

- Arnott, R, Kraus, M., 1998. Self-financing of congestible facilities in a growing economy. In: David Pines, Efraim Sadka, Itzhak Zilcha, (Eds.), *Topics in Public Economics: Theoretical and Applied Analysis*, 1997 Cambridge University Press, Cambridge, England, pp. 161–184, Reprinted in Erik Verhoef, ed., *The Economics of Traffic Congestion*, Vol. II, The International Library of Critical Writings in Economics 244, An Elgar Reference Collection, Edward Elgar, Aldershot, Hants, 2010, article 16.
- Bennathan, E., Walters, A.A., 1979. Port Pricing and Investment Policy for Developing Countries. The World Bank, Oxford University Press, Washington, D.C.
- Hau, T.D., 2005a. Economic fundamentals of road pricing: a diagrammatic analysis, Part I – Fundamentals. *Transportmetrica* 1 (2), 81–117.
- Hau, T.D., 2005b. Economic fundamentals of road pricing: a diagrammatic analysis, Part II - relaxation of assumptions. *Transportmetrica* 1 (2), 119–149.
- Hau, T.D., Becky, P.Y., Loo, K.I., Wong Wong, S.C., 2011. An estimation of efficient time-varying tolls for cross harbor tunnels in Hong Kong. *Singapore Econ. Rev.* 56 (4), 467–488.
- Keeler, T.E., Small, K.A., 1977. Optimal peak-load pricing, investment, and service levels on urban expressways. *J. Polit. Econ.* 85 (1), 1–25, Reprinted in Tae Hoon Oum, John S. Dodgson, David A. Hensher, Steven A. Morrison, Christopher A. Nash, Kenneth A. Small, W.G. Waters II, eds., *Transport Economics: Selected Readings*, Transportation Series 103, Korea Research Foundation For the 21st Century, 1995, Seoul Press, Seoul, pp. 425–455.
- Mohring, H.D., 1976. *Transportation Economics*. Ballinger Press, Cambridge, MA.
- Mohring, H., 1994. Introduction. In: Herbert, Mohring (Ed.), *The Economics of Transport*, Vol. I, The International Library of Critical Writings in Economics 34, An Elgar Reference Collection, Edward Elgar, Aldershot, Hants, pp. ix–xlii.
- Mohring, H., Harwitz, M., 1962. In: *Highway Benefits: An Analytical Framework*. Northwestern University Press, Evanston, IL.
- Mohring, H., 1984. Profit maximization, cost minimization, and pricing for congestion-prone facilities. *Transport. Res. E Logist. Transport. Rev.* 21 (1), 27–36.
- Morrison, Steven A., 1983. Estimation of long-run prices and investment levels of airport runways. *Res. Transport. Econ.* 1, 103–130.
- Newbery, D.M.G., 1989. Cost recovery from optimally designed roads. *Economica* 56, 165–185, Reprinted in Herbert Mohring, ed., *The Economics of Transport*, Vol. I, The International Library of Critical Writings in Economics 34, An Elgar Reference Collection, Edward Elgar, Aldershot, Hants, 1994, pp. 255–277.
- Small, K.A., 1999. Economies of scale and self-financing rules with noncompetitive factor markets. *J. Public Econ.* 74 (3), 431–450.
- Small, K.A., Verhoef, E.T., 2007. *The Economics of Urban Transportation*, 2nd ed. Routledge, New York.
- Verhoef, E.T., Mohring, H., 2009. Self-financing roads. *Int. J. Sustain. Transp.* 3 (5–6), 293–311.
- Walters, A.A., 1968. The Economics of Road User Charges, World Bank Occasional Paper Number 5. In: International Bank for Reconstruction and Development, Johns Hopkins University Press, Baltimore, MD.
- Yang, H., Meng, Q., 2002. A Note on “highway pricing and capacity choice in a road network under a build-operate-transfer scheme”. *Transport. Res. Part A Policy Pract.* 34 (7), 659–663.
- Zhang, A., Czerny, A.I., 2012. Airports and airlines economics and policy: an interpretive review of recent research. *Econ. Transp.* 1 (1–2), 15–34.

Further Reading

- Mohring, H., 1999. Congestion. In: Gómez-Ibáñez, José, William, B., Tye, Clifford, Winston (Eds.), *Essays in Transportation Economics and Policy: A Handbook in Honor of John R. Meyer*, 1999. The Brookings Institution, Washington D.C, pp. 181–221.
- Small, K.A., 1992. Urban transportation economics. In: *Fundamentals of Pure and Applied Economics* 51. Harwood Academic Publishers, Chur, Switzerland, pp. 112–116.

Road Pricing 3: The Implications for Pricing Public Transportation

Timothy D. Hau, HKU Business School, The University of Hong Kong, Hong Kong

© 2021 Elsevier Ltd. All rights reserved.

Introduction	90
An Analytical Framework for Pricing Public Transportation Use	90
Base Case	95
Scenario 1: Marginal Cost Pricing of Auto/Private Transport Only	97
Scenario 2: Marginal Cost Pricing of Bus/Public Transport Only	98
Scenario 3: Marginal Cost Pricing of Both Auto/Private Transport and Bus/Public Transport	98
The Role of a Transport Fund	100
Conclusions	101
References	101

Introduction

In the previous two chapters the theory of marginal cost pricing has been applied to the use of roads under congestion and to determine the equilibrium use of road networks in the short- and long-run. In order to get a full picture of the application of pricing as part of an overall approach to urban transportation it is necessary to apply the same principles to public transportation. This chapter assumes an understanding of the basic principles behind the application of marginal cost and congestion pricing from the previous two chapters.

An Analytical Framework for Pricing Public Transportation Use

The pursuit of marginal cost pricing shall now be extended from private transport to public transport. It is a truism that the accounts of public transit agencies are typically in the red, the shortfall of which requires public sector intervention and support. Local public transit, be it bus or rail, is unsustainable for several reasons. First, transit is subject to substantial economies of scale. Basing transit fares on average cost would result in fares that are higher than are desirable efficiency-wise. In the presence of increasing returns to scale, the marginal cost pricing of public transport, while efficient, results in financial losses that cannot be sustained. The absence of the marginal cost pricing of transit's competing mode—the ubiquitous private car—means that private transportation or road use is overused. Consequently, excessive automobile traffic exacerbates the financial losses of public transit agencies by siphoning much needed revenue passengers away.

Indeed, the classic justification for subsidizing public transport is based on second-best pricing, that is, negatively pricing public transportation because private transportation is implicitly subsidized in the absence of congestion pricing. In this way, it is sometimes argued that a rough balance between mass transit and private transportation would therefore surface. Yet, when the cost of implementing congestion pricing falls due to the declining cost of electronic technologies over time, the first-best pricing of private transportation has now proven enticingly feasible more than ever.

We have shown that long run net benefit maximization of highway-derived travel implies cost minimization and that the road authority's task is thus to trade off the traveler's time cost and the road authority's capital cost of road capacity. We next consider a socially responsible public transport agency interested in minimizing costs between two alternatives by adopting a back-to-back diagrammatic analysis for doing so. The analytical and geometric framework elucidated here and paradoxes advanced below follow the works of Nash (1976), Mogridge (1990a, 1990b), Small (1992), Arnott and Small (1994), Vickrey (1980, 1994) and Ding and Song (2012). Only by marginal cost pricing both private and public transportation would the sustainability of public transportation—from financial all the way to environmental—be within reach.

We extend our one-market framework of a single road to the simplest of a road network: the classic case of two roads. By way of historical background, a debate took place in the 1920s between Cambridge Professor Arthur Cecil Pigou and Chicago Professor Frank Knight that yielded the archetypal two-road case. Pigou considered the case of two public roads connecting point A to point D: one road is narrow, direct, straight and smooth, whereas the other is wide, long, circuitous and bumpy (see Fig. 1).

"Suppose there are two roads, ABD and ACD both leading from A to D. If left to itself, traffic would be so distributed that the trouble involved in driving a 'representative' cart along each of the two roads would be equal. But, in some circumstances, it would be possible, by shifting a few carts from route B to route C, greatly to lessen the trouble of driving by those still left on B, while only slightly increasing the trouble of driving along C. In these circumstances a rightly chosen measure of differential taxation against road B would create an 'artificial' situation superior to the 'natural' one. But the measure of differentiation must be rightly chosen."

A.C. Pigou, *The Economics of Welfare*, 1st edition, 1920, p. 194.

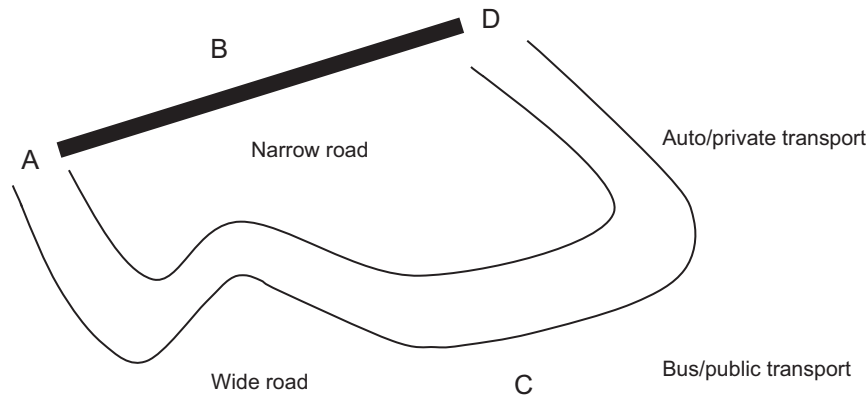


Figure 1 Pigou's classic two-road case.

A narrow road is short and direct and therefore fast but it has limited capacity; it attracts drivers eyeing for the fastest route. With many drivers taking the low-capacity road, congestion sets in quickly and the travel time, or time cost in money terms, rises geometrically (see Fig. 2A). A wide road is long and circuitous and therefore slow but it has ample capacity. The traffic on such a winding but high-capacity road is uncongested, with the travel time cost of a trip remaining the same throughout at AVC (see Fig. 2B).¹ Clearly drivers jockeying for the quickest route will divert to the road that is faster until it is just as fast, or as slow rather, as the narrow road. With each driver basing his decision solely on how long a time it takes for him to traverse that road, equilibrium will be reached where the travel time cost of a trip is the same on each road:

$$AVC_N = AVC_W$$

where AVC_N refers to the Average Variable Cost of trip taken on the Narrow Road and AVC_W refers to the Average Variable Cost of trip taken on the Wide Road. Given total traffic, Q_T , and each individual driver behaving rationally in a cost-minimizing manner on each road, the traffic equilibrium for this simple road network—the so-called User Equilibrium (U.E.)—will be reached as seen in Fig. 2C.² With two built roads, the intersection of AVC_N and AVC_W determines the short-run equilibrium: $SRAVC_N = SRAVC_W = AVC$, where AVC is the (initial) equilibrium average variable cost and the total traffic level of Q_T is distributed between equilibrium traffic on the narrow road, Q_N , and equilibrium traffic on the wide road, Q_W , at $Q = (Q_N, Q_W)$.

It is Knight's characterization of the conditions of such a traffic equilibrium for the two-link road network that was a precursor for what civil engineer Glenn Wardrop (1952, pp. 344-5) refers to as his first criterion of determining the distribution of routes based on travel times:

"The journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route."

Wardrop's first principle of route choice, widely referred to as User Equilibrium (U.E.), is also known as (selfish) Wardrop Equilibrium, the properties of which is that either of the two routes taken would end up being the same average travel time. If a route, say a third route, was not used, then that route must have a higher travel time; if not, there would be some traffic diverted to that third route. Wardrop's equilibrium and Knight's traffic equilibrium are the same, with the former expressed more formally in travel time units and the latter expressed in time cost units. We shall simply refer to both as User Equilibrium (UE) in our diagrams.

Further, Knight took Pigou to task by arguing that the introduction of a differential tax is misplaced because competition results in economic efficiency. Note that Knight's results are predicated upon the assumptions that: (1) road ownership is private and property rights are clearly delineated; (2) competition in road prevails and (3) scale economies are absent (Knight (1924), Hau (2005a, footnote 7)). However, even though Pigou was in fact considering the case of public roads and not private ones, regrettably Pigou withdrew the two-road case, now made famous, from subsequent editions of his book. Ever-present public roads the world over mean that negative externalities abound and the so-called "Pigouvian tax" or "Pigou tax"—attributed to none other than Professor Pigou himself—is indispensable in paving the way forward.³ After all, with the public sector being predominantly responsible for road provision and in spite of its monopoly position, it is the task of a benevolent road authority to mimic a competitive economy's efficiency-enhancing aspect by introducing a Pigouvian toll, more commonly known as a "congestion toll".

¹ Pigou and Knight assume that the wide road has sufficient capacity to handle total traffic (Knight (1924, Charts A, p. 588)).

² The term "user equilibrium" used here was subsequently attributed to Wardrop in 1952 based on his first criterion. Wardrop considered the general case of route choice with multiple routes whereas Pigou and Knight considered two routes in the 1920's.

³ We shall reserve the term 'Pigouvian tax' to include other externalities such as pollution on top of congestion. The term 'congestion tax' is a misnomer because it suggests that it is a governmental tax on congestion raised presumably for revenue purposes when it is clear that we should designate it as a 'congestion toll' as it is a price on congestion.

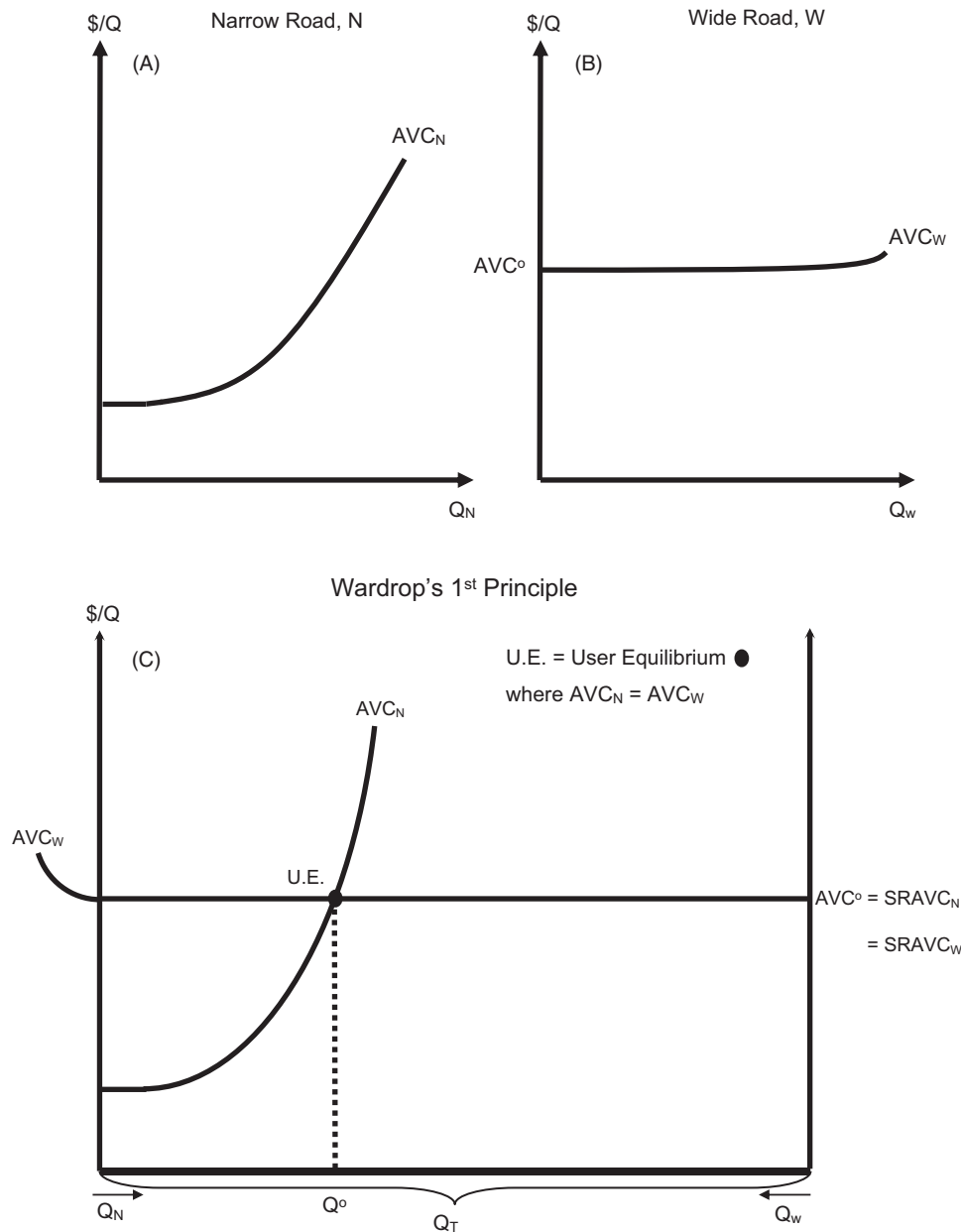


Figure 2 A back-to-back diagram of a narrow and wide road.

More occurrences of gridlock mean that there is a call for road capacity expansion of the fast road with either road construction or improvement. Would traffic woes be reduced? Following a lane expansion, say, the jump in traffic speed and a substantial improvement in travel time mean that traffic from the alternate wide route is attracted to the narrow road. The increase in road capacity is depicted by a horizontal fanning out of the average variable cost curve from AVC_N to AVC'_N (Fig. 3A).⁴ The resulting traffic equilibrium means that the expanded narrow road is now just as congested as before, with traffic equilibrating at the original average travel time cost of AVC^0 (Fig. 3B). The distribution of traffic has shifted from the wide road, Q , to the expanded narrow road, Q' . Road capacity expansion appears to be short-lived, a result known as the Pigou–Knight paradox. One way of overcoming the problem is to plow in a considerable amount of road investment resources so that the expanded narrow road is able to absorb the combined traffic of both roads to the point of being uneconomical at AVC' in Fig. 3B (see [Arnott and Small, 1994](#) and [Ding and](#)

⁴The fact that the shift in the average variable cost curve is anchored at the vertical intercept means that the average variable cost of a trip reflects the assumption of constant returns to road use in Figs. 4 and 5. For instance, the typical lane capacity of 1600 vehicles per hour per lane remains roughly the same when expanding from a 4-lane road to a 6-lane road.

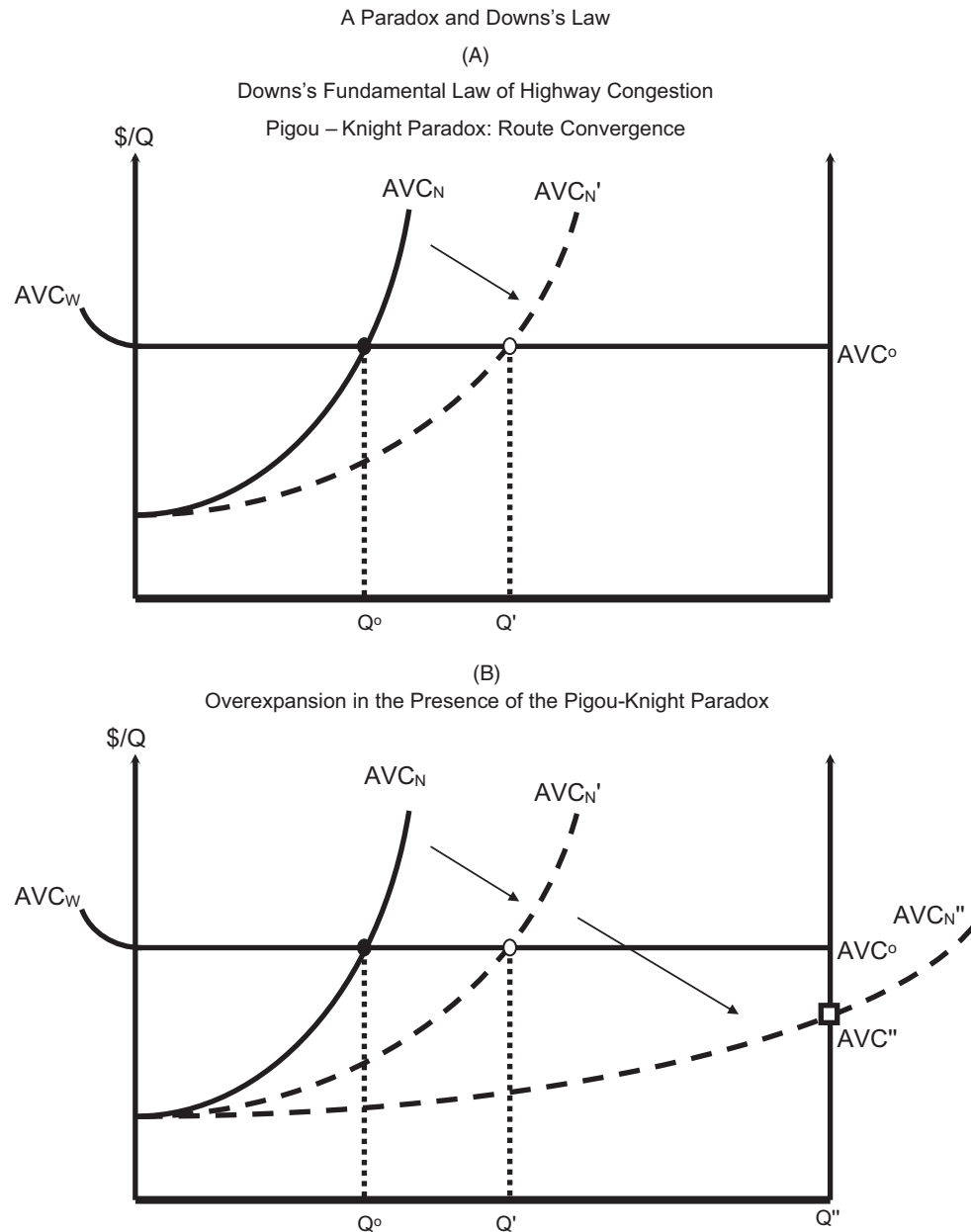


Figure 3 A Paradox and Downs's Law. (A) Downs's Fundamental Law of Highway Congestion Pigou–Knight Paradox: Route Convergence; (B) Overexpansion in the Presence of the Pigou–Knight Paradox.

Song, 2012 for numerical examples).⁵ It is uneconomical because a lowering of the average time cost from AVC to AVC' would suggest that the initial road had been underbuilt. Any optimally designed and used road should have *some* nonzero amount of congestion.

Anthony Downs (1962, 1992) advances the proposition that an increase in traffic capacity on commuter highways in urban areas results in an increase in urban travel demand during rush hour that erodes much of the capacity-enhanced traffic improvement. He calls it “the law of peak-hour expressway congestion” and “the iron law of congestion”, which states that, colloquially: “If you build them, they will come” as in the movie *The Field of Dreams*, 1989, starring Kevin Costner. The fundamental law of traffic congestion phenomenon arises because demand that has been suppressed up till now—otherwise known as latent demand—as a result of peak-hour congestion itself is released as soon as the traffic situation has improved.⁶ After all, we all opt to travel at our most convenient

⁵ The Pigou-Knight paradox is sometimes referred to as the Pigou-Knight-Downs paradox.

⁶ Downs's law is actually a transport example of a more general principle known in the literature as the tragedy of the commons: a resource that is unowned and unpriced would attract overexploitation and use.

time by our most favored mode and route were it not for having forced to queue up. So any betterment in travel time due to capacity enhancement will persuade hitherto early birds, for instance, to commute later and closer to their official work start time. Traffic will hereby increase until congestion worsens to a tolerable level—such as to that of an inferior local arterial—unless some sort of traffic restraint is imposed. In the Pigou–Knight paradox as depicted in the two-road case of Fig. 3A, the latent demand comes from the alternate route—the wide road—which by Pigou’s assumption has plentiful capacity.⁷ Downs coins this case “route convergence” or “spatial convergence”, the first of what he calls “triple convergence” occurring on a typical peak-hour weekday. Route convergence occurs because motorists would converge onto the most accessible and quickest route, thereby congesting it; motorists who formerly used alternate routes would search for the “best” route spatially. It is said that a variant of Say’s law in urban traffic therefore holds: (new) supply creates its own demand.

So how do we tackle this traffic conundrum? As in our one-market case, the economically efficient solution is to *price* for road use at marginal cost. For optimality, total travel time for the system via both routes has to be smallest, a result known as a System Optimal in time units or Social Optimum (S.O.) in money terms. Cost minimization for society could be met when the following condition holds for the case of two roads even when both are congested (Beckmann et al., 1956, Chapter 4)):

$$MC_N = MC_W$$

As before, the equimarginal principle would suggest that total travel cost would be lowest if the incremental costs end up being the same; else not. Civil engineers attribute this result to Wardrop’s (1952, p. 345) second principle:

“The average journey time is a minimum . . . the second criterion is the most efficient in the sense that it minimizes the vehicle-hours spent on the journey.”

Intuitively, if the marginal travel time of undertaking a trip via one road is lower than that of the other, total travel time for society could be lowered by moving over from one road to the other road until there is no longer a difference. Total travel time minimization could not be attained if the marginal cost of a trip undertaken via one road is higher than that of the other as a trip could be diverted from the higher cost road to a lower one with travel time saved.

When there is no divergence of the Average Variable Cost and the Marginal Cost of undertaking a trip via the wide road, no congestion toll needs be imposed. On the other hand, with a divergence of the Marginal Cost and the Average Variable Cost of undertaking a trip via the narrow road, a congestion toll τ^* needs to be imposed (Fig. 4). With marginal cost pricing, the collected

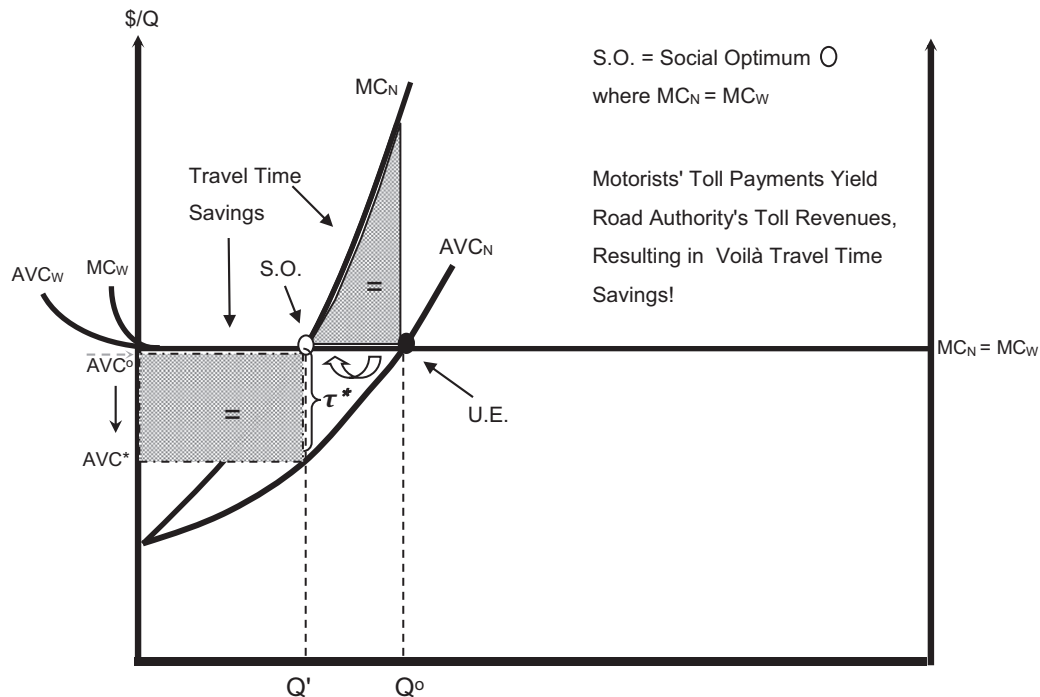


Figure 4 Wardrop's 2nd Principle.

⁷The analysis would still go through even if total traffic would result in congestion for both routes.

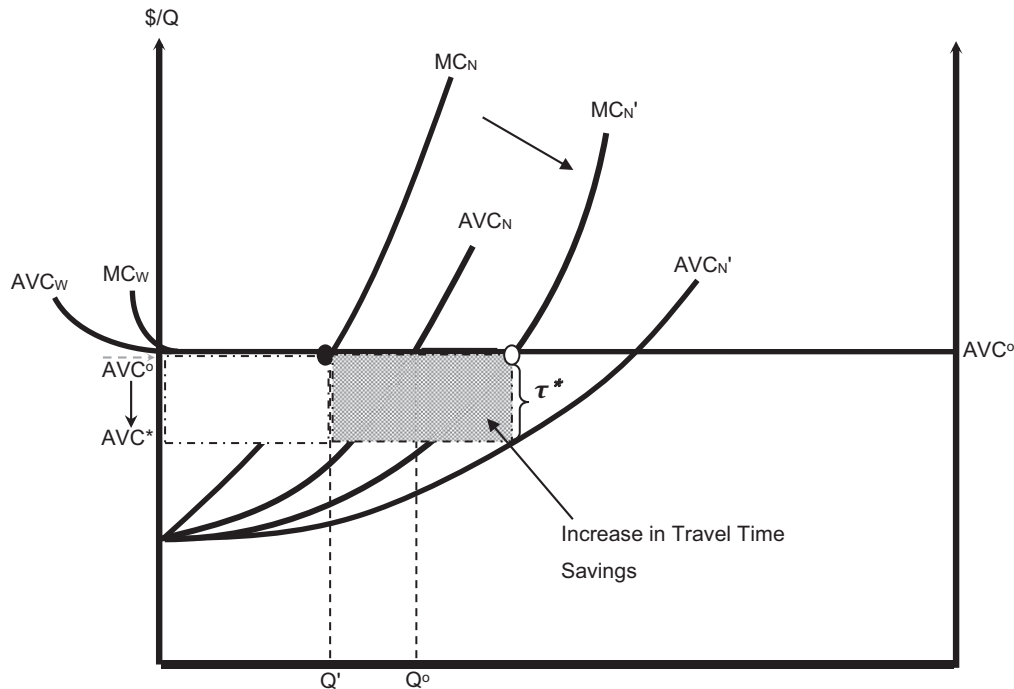


Figure 5 Reaping of the Bona Fide Benefits of Congestion Pricing

congestion toll revenues are simply the recouped quasi-rents. Note that the reaped travel time savings of a trip stem from applying marginal cost pricing, which lowers the Average Variable Cost of a trip from AVC to AVC^* . When multiplying the difference of the saving in Average Variable Cost by Q_N' , the traffic on the narrow road in the final equilibrium, the travel time savings result and is depicted as the shaded rectangle in Fig. 4.⁸ The congestion toll is able to—as if by magic—effectuate the travel time reduction of a trip via the narrow road by simply emitting the correct price signal to travelers to substitute money for travel time with the correct toll charged at the margin. Here we clearly see that it is the entire resultant travel time savings that is the *bona fide* benefits of congestion pricing; the road authority's congestion toll revenue collection as a transfer payment is further clarified. An alternative calculation of the resulting time savings: the welfare gain from the congestion toll, could be obtained by viewing the shaded triangular area below the Marginal Cost of undertaking a trip via the narrow road bounded by Q^0 and Q' .⁹ Looking at the argument the other way around, by *not* pursuing marginal cost pricing, the time cost incurred via both routes is the large rectangle in Fig. 4 and the difference in not going from U.E. to S.O. is the foregone shaded rectangle of time savings. Equivalently, the shaded triangular area is now the welfare loss of not going from U.E. to S.O.

So is road capacity expansion beneficial in such a case of marginal cost pricing? The answer is shown in Fig. 5, where the proportionate shifting of the pair of Marginal Cost and Average Variable Cost curves due to capacity expansion of the narrow road results in a proportionate increase in travel time savings.¹⁰

Duranton and Turner (2011) uses data from American cities to confirm Downs's "fundamental law of highway congestion". They find that vehicle kilometers traveled (VKT) increases one for one with interstate highways increase. They also uncover suggestive evidence that the law extends to major urban roads. They conclude that expansion of interstate highway and major urban road capacity is unlikely to bring about congestion relief. Surprisingly, the additional traffic does not come from other road types but from increased driving, among other factors, suggesting the presence and scale of induced demand.

The lesson here to be drawn here is that only by pursuing marginal cost pricing, or congestion pricing in the context of a road network, would road capacity expansion yield an increase in travel time savings or a gain in welfare without rent being dissipated. With proper pricing, the Pigou–Knight paradox and Downs's fundamental law of traffic congestion are now overcome.

Base Case

One key feature of public transport—both bus and rail—is that it is subject to increasing returns to scale due to large fixed cost, a feature we incorporate in our Base Case. Rail transportation is particularly capital-intensive and would require sufficient patronage

⁸The final equilibrium of the distribution of traffic is at $Q' = (Q_N', Q_W')$ in Fig. 4.

⁹The shaded triangular area and the shaded rectangular area is easily shown to be equivalent by first principles.

¹⁰While road expansion is no longer self-defeating, whether or not a road should be expanded should be governed by standard benefit-cost calculus (Hau, 2005b). As a transfer payment, the congestion toll revenues should be considered once but not double-counted (Mohring, 1993).

for it to survive. Although bus transport is less capital-intensive than rail, it too has to grapple with the issue of a declining average cost curve. In the long run, when a bus transit agency is allowed to vary the size of its capital stock, the average cost is lower with increased use. This means that the marginal cost of catering to more users lies below the average cost, frustrating a transit firm's efforts to survive in a market environment. On one hand, if average cost pricing were pursued, one financially breaks even but results in economic inefficiency. Imposing a hard budget constraint has a distinct advantage of incentivizing managerial efficiency in the transit agency. On the other hand, if marginal cost pricing were adopted, a transit agency will not be able to break even. After all, the case of economies of scale on the cost side, or increasing returns to scale on its production counterpart, is in fact a case of insufficient demand. Hence a Pigouvian subsidy or optimal per unit subsidy imposed on the user side is warranted following marginal cost pricing. But coming up with the funds to subsidize public transport involves considering the marginal cost of public funds that may make the economic case more difficult (Vickrey, 1980).

Another economic rationale in support of public transportation is due to Mohring (1963, 1970, 1972) and Turvey and Mohring (1975). Scheduled public transport differs from private transport in that the constant returns to scale condition does not hold. Suppose bus travel demand is doubled and if a transit agency were to respond by doubling the number of buses to be run, costs would appear to double and average costs would remain the same *prima facie*. However, with bus frequency doubling as a result, the waiting time is halved. Hence, the short-run average variable cost of a trip is lowered not only for the new arrivals, or "external" users rather, but also for the existing, or "internal" users, leading to a declining long run average cost curve and a long run marginal cost curve that lies below it at any output level. First observed by Professor Mohring, this effect is coined the Mohring effect in public transportation. The standard Short Run Average Variable Cost and Short Run Marginal Cost curves for bus transportation—which reflect elements internal to the bus firm and the tripmaker—rises, reflecting diminishing marginal productivity with respect to its own output of bus travel, but the Long Run Average Cost and Long Run Marginal Cost curves for bus travel—which reflect elements external to the transit firm but, because of the positive technological external economy engendered from more frequency, still internal to the industry as a whole—decline. This characterization is basically nothing more than the "system economy of scale effect" in microeconomics. Thus another first-best argument based on congestion pricing principles is brought forth to achieve economic efficiency on top of the traditional scale economy argument.

Given the fact that resource mobilization is increasingly difficult in a globalized world of financial shocks and pandemics, the question is how to come up with the funds to finance public transport. Putting it another way, the above considerations may lead one to wonder, how would transit agencies attract more users and become less cash-strapped? We shall appeal to various paradoxes to provide some guidance.

Consider a stylized urban transportation system where auto is the archetypal private transport mode and bus is the archetypal public transport mode. We can feature the salient aspects of urban transportation with a simplified model by depicting a rising Short Run Average Variable Cost curve for Auto as it faces diminishing marginal productivity and a falling Long Run Average Cost curve for Bus due to both aforementioned scale economy theories. By following the same reasoning as before with the two-road case, adopting Wardrop's 1st Principle on the behavior of both the auto and bus user would suggest that the condition for a new User Equilibrium is as follows (Fig. 6):

$$AVC_A = LRAC_B (= SRAC_B)$$

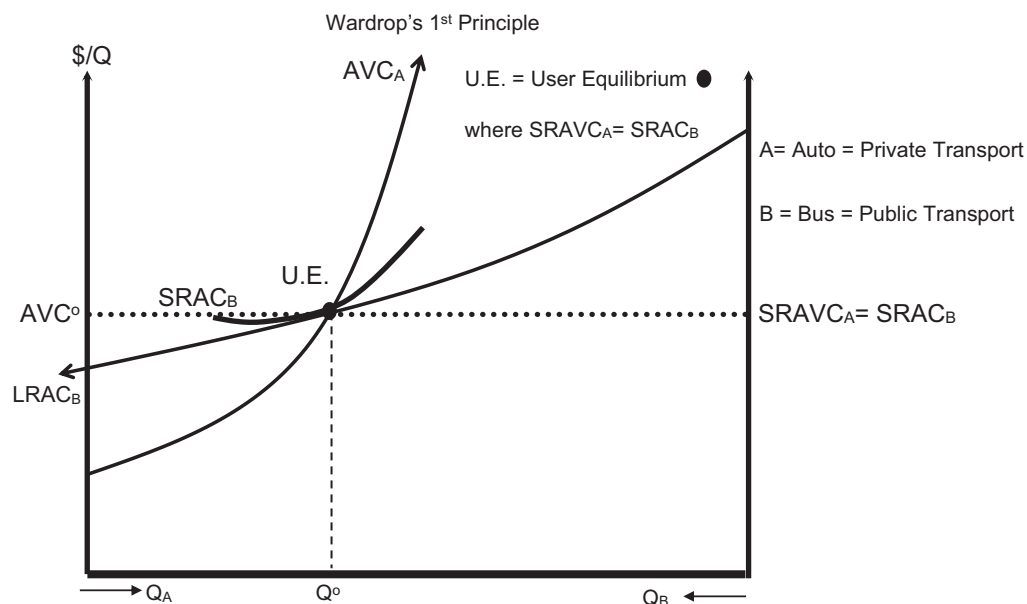


Figure 6 Scenario 0 Base Case of Average Variable Cost Pricing: $AVC_A = LRAC_B (= SRAC_B)$ Wardrop's 1st Principle

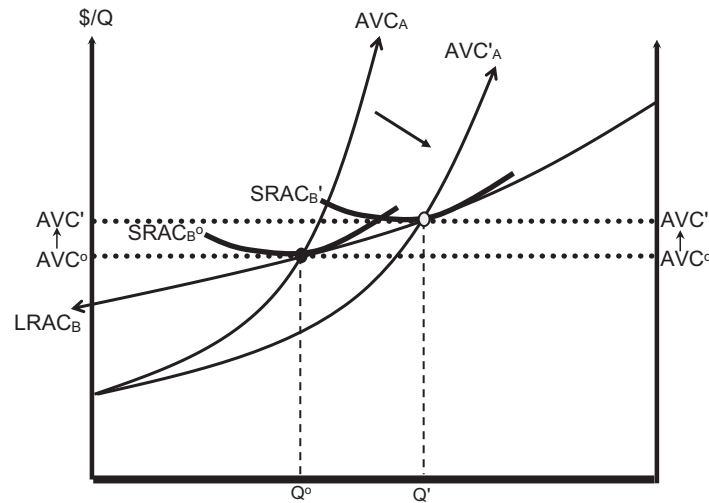


Figure 7 Downs–Thomson Paradox.

The rightly-chosen firm size that is determined by a bus agency is at the equilibrium traffic level of distribution of Q_B , where the *de facto* traffic distribution is: $Q = (Q_A, Q_B)$ at the status quo.¹¹ Note that at any point in time, one is always operating a particular firm size so the optimal firm size would be associated with the representative Short Run Average Cost curve for Bus, $SRAC_B$, tangent to the $LRAC_B$.¹² This is Scenario 0, also labeled as the Base Case, which I refer to as Average Variable Cost Pricing since at user equilibrium the average cost of a trip by either auto or bus is set equal to one another.¹³

Consider an increase in road capacity in the presence of both private transport and public transport. Private transport users, being the lion's share of road usage and the main beneficiaries of road expansion, benefit more; public transport users benefit too. Notice that an increase in road capacity now has a considerably greater impact—but negatively—on public transport than in the classic two-road case. This is because public transport relies on a large number of users to avoid being in the red. However, private transport, with its superior convenience and flexibility, outcompetes as well as siphons away much needed revenue passengers from public transport. Public transit agencies respond in two ways, both of which are detrimental to their own survival. With less passengers, transit agencies resort to raising fares to break even, likely leading to an even more of an exodus. And with less passengers, agencies cut transit frequencies to save on costs, which exacerbates the defection, given that the Mohring effect is now climbing up the long run curves, operating in the reverse direction. The traffic distribution changes from Q and Q' , reducing the public transport modal share vis-à-vis the private transport modal share. Not only that, the time cost of a trip perversely rises from AVC and AVC' for everyone going either by auto or bus! Indeed, the system travel time cost has increased as shown by an initial rectangle becoming yet an even larger rectangle (Fig. 7). The above phenomenon is known as the Downs–Thomson paradox (Mogridge et al., 1987; Thomson, 1977). Anthony Downs coins this result “modal convergence” as former public transport users would defect and switch to taking private transport, frustrating the congestion situation during rush hour.¹⁴

Scenario 1: Marginal Cost Pricing of Auto/Private Transport Only

We consider three scenarios that implement marginal cost pricing principles, with the first one on charging a positive price for a negative externality and the second one on charging a negative price for a positive externality. The marginal cost pricing of private transportation only—the conventional congestion pricing case—is our prime interest and forms Scenario 1 (Fig. 8). As the other alternative, public transportation, is not properly charged for, this scenario is considered a piecemeal approach to internalizing externalities. This is a realistic scenario in that public transit agencies are supposed to aim to covering their costs, despite their need for lump sum grants from governments to cover any shortfall. In recent decades there has been a call in many cities to introduce road pricing to not only tackle traffic woes but to use the collected revenue as a funding source for public transit. The condition to be met for Scenario 1 is as follows:

$$MC_A = LRAC_B (= SRAC_B)$$

By charging a congestion toll on private transportation/automobile, motorists trade off money with travel time, resulting in the time cost of a trip being lowered and the congestion toll revenues channeled into public coffers. As some motorists are priced off, they are

¹¹ After all, a transit agency would be concerned about its financial viability and firm survival rather than economic viability.

¹² Because of increasing returns nature of long-run average returns of a bus, excess capacity results.

¹³ A pre-existing gas tax for the auto mode, say, would merely shift the AVC_A curve upwards in a parallel manner.

¹⁴ The third convergence is “time convergence”. Motorists who used to travel just before or after the peak period due to congestion would move to their preferred peak time period following a road capacity increase in an urban area. Further, capacity expansion would result in the narrowing of the duration of the peak period, thereby benefiting commuters even though road capacity expansion appears to be futile diagrammatically. These urban phenomena occur because excessive congestion suppresses latent demand. It is this “triple convergence” that results in what Downs coins the fundamental law of peak-hour highway congestion.

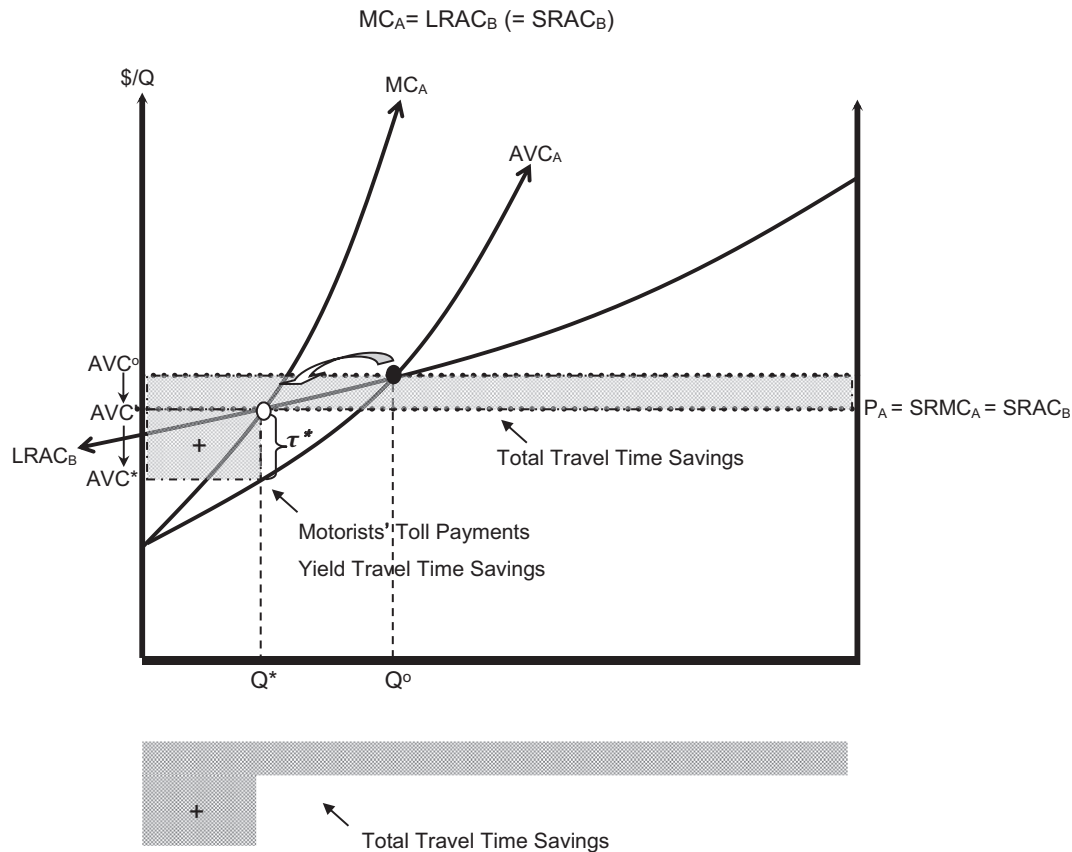


Figure 8 Scenario 1 Marginal Cost Pricing of Auto/Private Transport Only: $MC_A = LRAC_B (= SRAC_B)$

diverted to taking the alternate public transportation mode. With a declining Long Run Average Cost for Bus, $LRAC_B$, increasing returns are reaped, with a lower travel time cost of a trip via all modes at AVC' . Because the travel time cost of a trip is lowered from AVC to AVC' , travel time savings come about for both modes. This, coupled with the travel time savings brought about by the congestion toll wedge on motorists via auto yields the shaded rotated L-shaped area of the total travel time savings. The public transport modal share goes up from Q to Q^* , strengthening its sustainability. We find that it is the correct pricing signal that brings about the genuine benefits of congestion pricing and not the raising of revenues via the collected toll revenues. Unsurprisingly, public coffer does rise by the same amount of motorists' toll payments. The trouble is that the public's trust in the government's proper handling of the revenues collected is unlikely to improve.

Scenario 2: Marginal Cost Pricing of Bus/Public Transport Only

For Scenario 2, the marginal cost pricing of public transportation only would mean that economic efficiency is enhanced for public transport via the introduction of an optimal per unit subsidy of s^* (Fig. 9). Without congestion externality from private transport being internalized, let us consider the scenario of optimally subsidizing public transport, likely to be viewed as palatable politically. Maybe considered piecemeal, it is included here for comparative purpose. The condition to be met for Scenario 2 is as follows:

$$SRAVC_A = LRMC_B)(= SRMC_B)$$

Because of the downward-sloping nature of the Long Run Average Cost of Bus, $LRAC_B$, we see that the average time cost of a trip falls for both auto and bus from AVC to AVC' , whereas the modal share for public transport increases from Q to Q^{**} (Fig. 9). By the same token, there is a further reduction of travel time for both the motorist and the bus user, with the vertical difference being $(AVC' - AVC^*)$. Considering both modes, after netting out the resource cost of the subsidy, which has to be accounted for, the rectangular shaded total travel time savings area is reduced to the cross-hatched and shaded rotated L-shaped area of the net total travel time savings depicted in Fig. 9.

Scenario 3: Marginal Cost Pricing of Both Auto/Private Transport and Bus/Public Transport

We aim to achieve first-best pricing by applying marginal cost pricing to all modes; this forms Scenario 3 as shown in Fig. 10.

$$\text{SRMC}_A = \text{LRMC}_B (= \text{SRMC}_B)$$

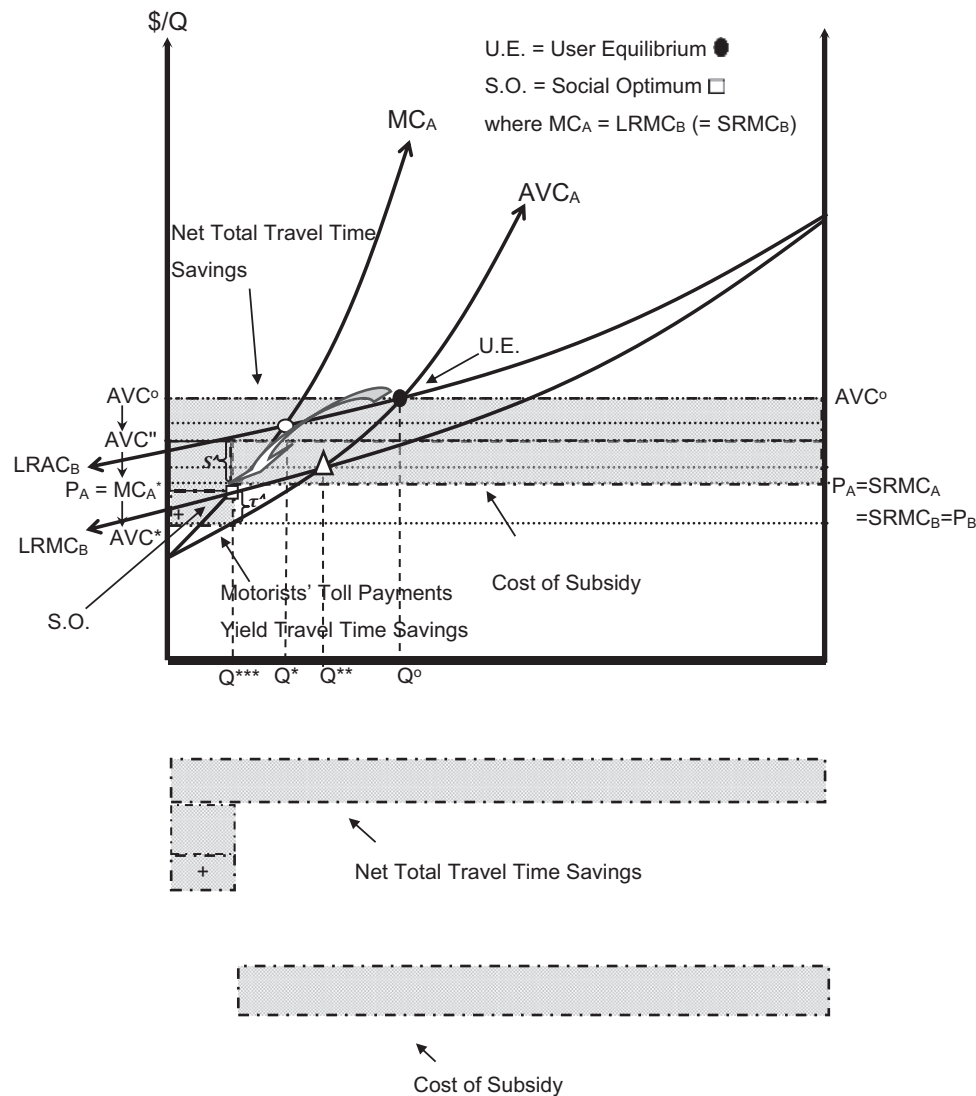


Figure 10 Scenario 3 Marginal Cost Pricing of Both Auto/Private Transport and Bus / Public Transport

Small (2005) argues that mass transit has failed because some view it as being the solution to our congestion problem. He models and quantifies the positive impact that road pricing has on public transport, contributing to its virtuous cycle (Small, 2004). Rather than viewing road pricing as an alternative policy to local transit provision, road pricing could be viewed as mass transit's savior and selling road pricing as a package approach to tackling gridlock.

The Role of a Transport Fund

With road pricing, our concern for compensating the affected parties—the tolled, the tolled off and tolled on—would need to be carried out in an indirect manner to ensure Pareto efficiency. One idea is to consider simply setting up a transport fund where the economic profits from heavily used roads and the diseconomies of a road network could be used to cover the economic losses from worthwhile lightly used roads as well as public transportation due to large fixed costs and the Mohring effect. Based on marginal cost pricing of all modes, we can appeal to Alfred Marshall's 'tax and bounty' system of financing:¹⁷

"One simple plan would be the levying of a tax by the community on their own incomes, or on the production of goods which obey the law of diminishing return, and devoting the tax to a bounty on the production of those goods with regard to which the law of increasing return acts sharply."

Alfred Marshall, *Principles of Economics*, 1920, Book V, Chap. XIII, Sec. 6, p. 392.

¹⁷ Marshall (1920) also observes that it is necessary to consider the cost of administering such a tax-subsidy scheme, whose electronic charging technology continues to fall and becoming more cost-effective (Hau, 2005b, 2006b).

The transport fund is a depository where the proceeds from congestion tolling are deposited and expenditures administered based strictly on the grounds of benefit-cost analysis. David Newbery (1994) shows that the replacement of current road taxes in the United Kingdom by a set of optimal road user charges would adequately cover the cost of the road infrastructure. This finding strengthens the case for establishing an independent public road authority.

Conclusions

Congestion pricing has not been successful in the past because it hurts road users on average. In fact, in the absence of compensation, most users of the transportation system area made worse off by congestion pricing. Among those who pay the toll, only a fraction of motorists who value their time very highly will feel that their time saved is worth more than the toll payment; the majority of motorists will be made worse off. Those who are tolled off clearly experience a reduction in welfare because they feel they are coerced to travel by a less desired mode, route or time of day. And those who are already on public transport may be made no better off because they are faced with more crowded buses and metros.

Without compensation to travelers, the main beneficiary of congestion pricing is the road authority that collects the tolls. With or without the establishment of a transport fund, it is only the government's *compensation* of the several groups that makes travelers themselves experience a net gain due to congestion pricing. Economic analysis suggests that, under certain conditions, compensation should take the form of:

1. road construction and improvement for those who stay on the road and pay (i.e., *the tolled*) based on an analytical framework that regards the congestion toll as a *user fee* in both the short run and the long run a la Mohring-Harwitz;
2. expanded and improved public transportation alternatives for those who are priced off and diverted from their preferred options because of the toll (i.e., for both *the tolled off* and *the tolled on*); and/or
3. lower gas taxes, purchase taxes and annual vehicle registration and license fees, etc. (for *the untolled*) if deemed excessive.

Without effective compensation schemes, congestion pricing is doomed to political failure.

Coupled with increasing concern over air and noise pollution, climate change as well as pandemics from Covid-19, the appetite for safe and socially-distanced transport—both private and public—will rise. The trend of working from home may offset the demand for face-to-face interaction in the workplace but not necessarily reduce the demand for socially-distanced automobile travel to nonwork destinations. It is incumbent that the first-best pricing strategy espoused here, or road pricing in common parlance rather, be pursued and externalities charged for. Based on a combined assessment of the conceptual underpinnings and international experiences (see Case Study below), congestion pricing offers a way out of our present gridlock, both trafficked and political.

References

- Arnott, Richard, Kenneth, A. Small, 1994. The Economics of Traffic Congestion. *American Scientist* 82, 446–455.
- Beckmann, Martin, C.B., McGuire, Christopher, B., Winsten, 1956. *Studies in the Economics of Transportation*. Yale University Press, New Haven, Connecticut.
- Ding, Chengri, Song, Shunfeng, 2012. Traffic Paradoxes and Economic Solutions. *Journal of Urban Management* 1 (1), 63–76.
- Anthony, Downs, 1962. The Law of Peak-Hour Expressway Congestion. *Traffic Quarterly* 16, 393–409.
- Anthony, Downs, 1992. *Stuck in Traffic: Coping with Peak-Hour Traffic Congestion*. The Brookings Institution, Washington D.C.
- Duranton, Gilles, Turner, Matthew A., 2011. The Fundamental Law of Road Congestion: Evidence from US Cities. *American Economic Review* 101 (6), 2616–2652.
- Hau Timothy, D., 2005a. Economic Fundamentals of Road Pricing: A Diagrammatic Analysis, Part I – Fundamentals. *Transportmetrica* 1 (2), 81–117.
- Hau Timothy, D., 2005b. Economic Fundamentals of Road Pricing: A Diagrammatic Analysis, Part II - Relaxation of Assumptions'. *Transportmetrica* 1 (2), 119–149.
- Hau, Timothy, D., 2006b. Congestion Charging Mechanisms for Roads, Part II - Case Studies. *Transportmetrica* 2 (2), 117–152.
- Knight Frank, H., 1924. Some Fallacies in the Interpretation of Social Cost. *Quarterly Journal of Economics* 38, 582–606, Reprinted in Kenneth J. Arrow and Tibor Scitovsky, eds., *Readings in Welfare Economics*, American Economic Association Series, George Allen and Unwin Ltd., London, 1969, pp. 213–227..
- Marshall, Alfred, 1920. *Principles of Economics*, 8th edition. Macmillan Press Limited, London.
- Mogridge Martin, J.H., 1990a. Travel in Towns: Jam Yesterday, Jam Today and Jam Tomorrow? Macmillan Press, London.
- Mogridge Martin, J.H., 1990b. Road Pricing and the Edgeworth Paradox. *Economic Affairs* 11–22.
- Mogridge, M.J.H., Holden, David, J., Bird, J., Terzis, G.C., 1987. The Downs–Thomson Paradox and the Transportation Planning Process. *International Journal of Transportation Economics* 14, 283–311.
- Herbert Mohring, 1963. The Place of Subsidies in an Optimum Transportation System. *Highway Research Record* (20), 103–113.
- Mohring, Herbert, 1970. The Peak Load Problem with Increasing Returns and Pricing Constraints. *American Economic Review* 60 (4), 693–705.
- Mohring, Herbert, 1972. Optimization and Scale Economies in Urban Bus Transportation. *American Economic Review* 62 (4), 591–604.
- Mohring, Herbert, 1993. Maximizing, Measuring, and *Not* Double Counting Transportation-Improvement Benefits: A Primer on Closed- and Open-Economy Cost-Benefit Analysis. *Transportation Research Part B: Methodological* 27 (6), 413–424.
- Nash Chris, A., 1976. *Public versus Private Transport* Macmillan Studies in Economics. The Macmillan Press Limited, London.
- David Newbery, M.G., 1994. The Case for a Public Road Authority. *Journal of Transport Economics and Policy* 28 (3), 235–253.
- Pigou, Arthur, C., 1920. *The Economics of Welfare*, 1st edition. Macmillan Company, London.
- Ramsey, Frank, P., 1927. A Contribution to the Theory of Taxation. *Economic Journal* 37, 47–61.
- Small, Kenneth A., 1992. In: *Urban Transportation Economics* Fundamentals of Pure and Applied Economics 51. Chur Switzerland, Harwood Academic Publishers, pp. 112–116.
- Small, Kenneth A., 2004. Road Pricing and Public Transport. In: Georgina Santos (Eds.), *Research in Transportation Economics Vol. 9 Road Pricing: Theory and Evidence*, Elsevier, pp. 133–158.
- Small, Kenneth A., 2005. Unnoticed Lessons from London: Road Pricing and Public Transit", *Access No. 26.Spring*, 10–14.
- Thomson, J.M., 1977. In: *Great Cities and Their Traffic*. Published in Peregrine Books, Gollancz, London, p. 1978.

- Train, Kenneth, E., 1977. Optimal Transit Prices under Increasing Returns to Scale and a Loss Constraint. *Journal of Transport Economics and Policy* 11 (2), 185–194.
- Turvey, Ralph, Mohring, Herbert, 1975. Optimal Bus Fares. *Journal of Transport Economics and Policy* 9 (3), 280–286, Reprinted In Tae Hoon Oum, John S. Dodgson, David A. Hensher, Steven A. Morrison, Christopher A. Nash, Kenneth A. Small, W.G. Waters II, eds., *Transport Economics: Selected Readings*, Transportation Series 103, Korea Research Foundation For the 21st Century, 1995, Seoul Press, Seoul, 303-312.
- Vickrey, William S., 1980. Optimal Transit Subsidy Policy. *Transportation* 9 (4), 389–409.
- Vickrey William, S., 1994. Reaching an Economic Balance between Mass Transit and Provision for Individual Automobile Traffic. *Logistics and Transportation Review* 30 (1), 3–19.
- Wardrop, Glen, John, 1952. Some Theoretical Aspects of Road Traffic Research. *Proceedings of the Institution of Civil Engineers* PART II 1, 325–378, Reprinted in Herbert Mohring, ed., *The Economics of Transport*, Vol. I, The International Library of Critical Writings in Economics 34, An Elgar Reference Collection, Edward Elgar, Aldershot, Hants, 1994, pp. 299-352.

Road Pricing 4: Case Study—The Implementation of Electronic Road Pricing in Hong Kong

Timothy D. Hau, HKU Business School, The University of Hong Kong, Hong Kong

© 2021 Elsevier Ltd. All rights reserved.

A Practicable Proposal for the Implementation of Electronic Road Pricing in Hong Kong References

103
105

A Practicable Proposal for the Implementation of Electronic Road Pricing in Hong Kong

The previous three articles have developed the theory of marginal cost or congestion pricing for roads and the implications for an efficient urban transportation system covering both private and public transportation. The article discusses the possibility of electronic road pricing, the history of its development in various cities across the world and the practical difficulties of applying it in one city, Hong Kong.

Technological breakthroughs in computers, radio frequency identification, global positioning systems, intelligent vehicle transportation system, and smart transportation have made automatic vehicle charging become a reality. The earliest proponents of Electronic Road Pricing (ERP) are Nobel Laureate [Vickrey \(1960\)](#) and panel members of the UK [Ministry of Transport \(1964\)](#), authors of the famous Smeed Report, both of which formed the precursors of Hong Kong's pioneering ERP feasibility study.

Hong Kong was the first city worldwide to undertake actual technical feasibility of ERP from 1983 to 1985 ([Transpotech, 1985](#)). Despite being tested and having attained over 99% effectiveness and reliability, the soldering of a video-cassette-sized electronic number plate on the underside of a private car—thus revealing a vehicle's whereabouts—helped spell its political demise. One lesson to be learnt from that thorough study was that the singling out of the private car motorist as the sole group to be charged ERP would doubly increase the burden and raise the ire of the private car motorist. The study also suggests that granting vehicular exemptions to certain vehicle types or groups would be a slippery slope. Viewed as “just another tax,” the promise of lowering annual vehicular license fees to offset ERP at the eleventh hour failed to assuage opponents of ERP. Nevertheless, the annualized benefit-ratio of implementing a pilot stage of ERP in Hong Kong is impressive, ranging from 14.7:1 to 17.8:1 ([Hau, 1990](#)).

With further technical advancements, the second feasibility study of ERP in Hong Kong from 1997 to 2001 reported significant time savings and vehicle operating cost savings and overcame privacy concerns by appealing to the use of Dedicated Short-range Radio Communication (DSRC) system ([Transport Department, 2001](#)). The report broaches the concept of revenue neutrality by proposing that the ERP gross revenue proceeds of HK\$0.4 to HK\$1.3 billion every year be ploughed back for transport infrastructure investment.

Meanwhile, in 1998, Singapore turned its paper-based road pricing system—the so-called Area Licensing Scheme, which had first begun in June 1975—into the world's pioneering ERP system. Using DSRC and smart card technology, Singapore adopted a flexible, cordon-based system in the charged area and imposed ERP rates according to vehicular size and type. The simplicity and beauty of setting a target speed range of 20–30 km/h on major roads in the cordoned charging area and 45–65 km/h on expressways for all to see and monitor has much to commend for, as opposed to appealing to the standard approach of calculating traffic volumes, volume–capacity ratios, service levels, road capacities, and so on.

Soon afterwards, on February 17, 2003, London defied informed critics by fast-tracking congestion charging with its Automatic Number Plate Recognition (ANPR) system, resulting in a costlier system which still yielded positive net benefits ([Santos et al., 2006](#)). London Mayor Ken Livingstone came up with important complementary measures such as using the congestion charging revenues to finance an expanded public transport fleet. London's resounding success with its opening of £5 congestion toll during daylight hours also put paid to skeptics' doubts on the feasibility of implementing road pricing in a major world city where the electorate holds sway.

Following that, Stockholm introduced a congestion charge for a 7-month trial period from January to July 2006. The concern of reverting to precharge congested traffic conditions, Stockholm city residents' preference for a better environment and air quality and the provision of ERP revenues for new road construction all contributed to the positive outcome of a referendum that made congestion charging in Stockholm permanent from August 2007 onwards to date ([Eliasson, 2009](#); [Simeonova et al., 2018](#)).

Other notable examples of the successful implementation of ERP include Milan, Italy in 2001, Durham, England in 2002, Valetta, Malta in 2007 (a one-of-a-kind, time-based system) and Gothenburg, Sweden in 2013. In recent years, a flurry of congestion charging systems has started, with the three historical Norwegian toll rings' flat tolling progressing to time-differentiated charging, together with several new toll rings being variable-tolled: Trondheim in 2012, Kristiansand in September 2013, Bergen in February 2016, Oslo in October 2017 (which has three cordons: city border, original cordon, and inner cordon), Nord-Jren in March 2019. To tackle both chronic traffic congestion and help finance its public transit system, a congestion pricing proposal to charge vehicles entering Manhattan south of 60th street in New York City in 2022 passed the New York State Budget in April 2019 but stalled at the federal level.

For most of the aforementioned ongoing congestion charging schemes worldwide, the net revenues are used for public transport infrastructure, services and improvements, road investments, busways, air pollution-reducing policies, and/or transport safety enhancements. A major finding of the British and Scandinavian road pricing desk studies and operational schemes is that motorists are against having to pay more than they already have but when the collected road pricing revenues are recycled for road construction and improvement and/or better public transport provision, there is a turnaround in support of road pricing as a package approach (Gu et al., 2018; Jones, 1991).

It is therefore right and prescient for the Financial Secretary Paul Chan Mo-po to promulgate – in a groundbreaking paragraph for transport in Hong Kong—in his annual Budget Speech 2019–20 on February 27, 2019: “Electronic road pricing does not aim to increase government revenue” and that Government will provide equivalent net revenues generated from a proposed ERP Pilot Scheme “for implementing measures to improve public transport services and encourage wider usage.” In April 2019, the Transport Department of the HKSAR Government initiated “Smart Mobility—Intelligent Traffic Management: Electronic Road Pricing Pilot Scheme in Central Core District: Concepts and Preliminary Ideas” (Transport Department, 2019).

Despite the high first registration taxes and annual license fees that rank amongst the highest in the world, travel demand still far outpaces society’s ability to provide sufficient transport infrastructure, resulting in traffic speeds in Central District that are a little faster than a slow jog. In order to fully exploit the benefits of the enhanced capacity provided by the recent opening of the HK\$36 billion Central Wan Chai Bypass in February 2019, ensuring the Bypass’s efficient use is of high priority. If road pricing is continually forestalled, suppressed travel demand long held back by traffic congestion itself would more sooner than not be released—due to the fundamental law of traffic congestion and traffic paradoxes—that would swamp the Bypass’s carrying capacity.

The promise to earmark the net revenues from the ERP Pilot Scheme to improve public transport services satisfies the *sine qua non* for ERP’s successful implementation in several overseas experiences mentioned above. It appears that Hong Kong needs to move one step further and redistribute the revenues in a way that directly benefits the public transport passengers by placing a part—such as a half—of the collected revenues in passengers’ pockets. As Central is the location where the charging zone is situated, a simple proposal—call this the flat rate subsidy proposal—is that a fixed amount be compensated for a public transport user whose destination is Central. Alternatively, a variant—call this the proportionate subsidy proposal—is that a certain percentage of a user’s public transport fare whose destination is Central be compensated. (The subsidy could be set conservatively to ensure ERP revenue neutrality.) What is the conceptual basis for such a compensation proposal?¹

1. The proportionate subsidy proposal is consistent with the *first-best pricing principle* for public transport (including both bus and rail) to be subsidized on the basis of both the increasing returns to scale due to its large fixed cost and the Mohring effect.
2. The subsidy proposal—be it flat rate or proportionate—is *progressive*.
3. Both subsidy proposals have a *lower relative price for public transport* as they lower the public transport fare for users whose destination is Central relative to private car use, so that travel demand using scarce road space is in the right direction with more public transport use and less private car use being incentivized, helping to further strengthen the alleviation of Central’s traffic bottlenecks.
4. Both subsidy proposals possess a *marginal economic incentive* that tilts travel mode choice from a less environmentally friendly mode to a more developmentally sustainable one.
5. Both subsidy proposals are along the lines of Nobel Laureate Richard Thaler’s *nudge theory* in behavioral economics in which the subsidy serves as the nudge—a negative surcharge—tilting the use of mode choice towards public transport and away from private car use.
6. Both subsidy proposals follow the *user pays principle*—just as the proposed charging of private car use does—by subsidizing a “good” (public transport use) and taxing (or, more accurately, pricing) a “bad” (private car use entering the CBD) and are fully consistent with the *polluter pays principle* behind Hong Kong’s soon-to-be-implemented mandatory waste charging scheme.
7. Both subsidy proposals—in conjunction with the proposed charging of private car use – would help forestall the Central Wan Chai Bypass from being overcome within a few years’ time due to latent demand *a la* Downs’s law. In October 2005, the Harbor-front Enhancement Committee’s “The Final Report of the Expert Panel on Sustainable Transport Planning and the Central—Wan Chai Bypass (CWB)” urges that ERP be pursued once CWB comes into being as the Bypass would serve as a viable escape route for the through traffic that avoids the charging zone completely.
8. To minimize administrative costs, it is proposed that the compensation scheme be piggybacked onto the innovative Public Transport Fare Subsidy Scheme (PTFS) already in place since January 2019 using the ubiquitous Octopus reusable contactless smart stored value card universally held by Hong Kongese. (In January 2020, the enhanced PTFS Scheme has an increased monthly subsidy cap of HK\$300, a subsidy rate of 33% in excess of HK\$400 in monthly public transport expenditures incurred by a user on franchised transport modes.²) By piggybacking onto the Transport Department’s hugely popular Subsidy Scheme, the compensation proposal benefits from the same properties as its PTFS scheme which has the side effect of *stimulating intermodal competition* by transit riders voting with their feet and revealing their preference for that public transport mode that yields the best combination of service performance and affordability. It is through increased intermodal competition that the traveling public’s well-being is served and even enhanced rather than via tightened regulatory rules (aimed at, for instance, the Mass Transit Railway Corporation), which may not be as responsive and effective whenever the market suddenly turns.

¹ The proposal is backed up by a group of 25 transportation researchers having Hong Kong’s long term well-being at heart. The group made up most of Hong Kong’s transportation experts.

² From July to December 2020, the HKSAR Government further lowered the monthly deductible by half to HK\$200 (from HK\$400) to participate in the PTFS Scheme because of the COVID-19 pandemic.

In 2006, the Government's Council for Sustainable Development released the results of a large-scale survey on "Clean Air and Blue Skies—the Choice is Ours" indicating that the majority of the survey respondents expressed support for a better environment via road pricing. The survey results are consistent with the well-known fact that over nine-tenths of trips made during peak-hour traffic are by public transport and only one-tenth is by private car.

The political difficulties of congestion pricing in the case of Hong Kong are not insurmountable if the proceeds of congestion pricing are made tangible and shared by all affected parties in the transport community: the tolled, the tolled off, and the tolled on. It is envisaged that road pricing implementation in major metropolises will hopefully arrive sooner rather than later.

References

- Eliasson, J., 2009. A cost-benefit analysis of the Stockholm congestion charging system. *Transport. Res. A: Policy Pract.* 43 (4), 468–480.
- Gu, Z., Liu, Z., Cheng, Q., Saberi, M., 2018. Congestion pricing practices and public acceptance: a review of evidence. *Case Stud. Transport Policy* 6 (1), 94–101.
- Hau, T.D., 1990. Electronic road pricing: developments in Hong Kong 1983–89. *J. Transport Econ. Policy* 24 (2), 203–214.
- Jones, P.M., 1991. Gaining public support for road pricing through a package approach. *Traffic Eng. Control* 32 (4), 194–196.
- Ministry of Transport, 1964. In: *Road Pricing: The Economic and Technical Possibilities*. Her Majesty's Stationery Office, London, pp. 1–61.
- Santos, G., Fraser, G., Newbery, D., 2006. Road pricing: lessons from London. *Econ. Policy* 21 (46), 264–310.
- Simeonova, E., Currie, J., Nilsson, P., Walker, R., 2018. Congestion Pricing, Air Pollution and Children's Health. NBER Working Paper No. 24410, March, pp. 1–34.
- Transport Department, 2001. *Feasibility Study on Electronic Road Pricing: Final Report*. April, Government of the HKSAR, pp. 1–27.
- Transport Department, 2019. Consultation with Central and Western District Council on 16 May 2019 on Smart Mobility – Intelligent Traffic Management: Electronic Road Pricing Pilot Scheme in Central Core District. HKSAR Government Website, May.
- Transpotech, 1985. *Electronic Road Pricing Pilot Scheme: Main Report*. May, Main Report prepared for the Hong Kong Government, pp. 1.1–4.40.
- Vickrey, W.S., 1960. Statement to the Joint Committee on Washington Metropolitan, DC, Problems, Transportation Plan for the National Capital Region, Hearings, Joint Committee on Washington Metropolitan Problems, 8–14 November 1959, pp. 454–490. Reprinted in *Journal of Urban Economics*, Vol. 36, Issue 1, July, 1994, pp. 42–65.

Advanced Travelers Information Systems (ATIS)

Chintan Advani, Ashish Bhaskar, School of Civil and Environmental Engineering, Queensland University of Technology, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	106
Need for Traveler Information	106
Information Generation	106
Information Dissemination	106
Recent Trends in ATIS	107
References	108

Introduction

Transport systems are the veins of the local, national, and global economy. Intelligent Transport Systems (ITS) aims to exploit information and communication technologies for efficient and safe mobility of people and goods. Advanced Traveler Information System (ATIS) is an important component of ITS, where the aim is to provide accurate and reliable information to the traveler for their travel over the network.

The information of interest to the traveler includes, not limited to, pretrip multimodal trip planning and route choice options, on route dynamic expected arrival time, traffic incidents, roadworks, hazards, and infrastructure restrictions. Generation of such information requires significant amount of monitoring data from the transport systems, which is primarily acquired and managed by public sector (transport agencies). Over the last decade, private sector has increasing business interest towards acquiring and storing road traffic conditions. The breadth of the information needed by different traveler and travel purpose requires multiple means to disseminate the information.

Need for Traveler Information

A significant amount of our precious time is wasted on non-optimized travel. Multiple reports have been published highlighting the “extra time” lost due to uncertainties in travel. Recently, TomTom (https://www.tomtom.com/en_gb/traffic-index/) has reported on average extra time spend in traffic 2019 was 87 min. The uncertainties in the travel leads to travel stress, aggressiveness, and road rage. Researchers have shown that an aggressive and stressed driver is more likely to have risky driving and lead to a crash (Bowen et al., 2020; Öz et al., 2010). Availability of real time and predictive traveler information should reduce traveler stress and related detrimental impacts.

Information Generation

Transport agencies collect significant amount of road network monitoring and related data through traditional loop detector, floating car data, manual survey, and traffic surveillance camera. With advancement in technology, new data sources such as Bluetooth and Wifi based MAC scanners (Advani et al., 2019; Bhaskar and Chung, 2013), GPS (Byon et al., 2006), and emerging drones (Barmounakis and Geroliminis, 2020; Kamnik et al., 2020) are being exploited for spatial temporal traffic monitoring. Private firms such as Google (<https://www.google.com/maps>), Waze (<https://www.waze.com/>), Here (<https://www.here.com/>), Intellematics (<https://www.intelematics.com/>), and INRIX (<https://inrix.com/>) are also collecting traffic congestion data on the network and these data is available (at cost) to the transport agencies. It is worth mentioning that the availability of the data from multiple sources can lead to over information. There is a need to fused data from multiple sources for a single point of truth. The fusion algorithm should consider the confidence and reliability of individual data source.

In addition to traffic congestion, information on road works, traffic crashes, hazards, flash flooding, fire, weather, etc. is also collected and integrated with the traffic conditions. Recently, there is increasing focus of the government agencies to open the data, and different initiatives are taken for real time sharing of the data through open data portals. The availability of such data provides opportunities to the independent developers to develop tools for traveler information.

Information Dissemination

After decades of development, ATIS is now capable of providing numerous kinds of information, covering almost all aspects of travel such as trip planning, route navigation, and real-time traffic conditions (Ben-Elia and Avineri, 2015; Chorus et al., 2006; Gössling, 2018). The focus, however, is usually quite narrow, targeting on either route choice (Bagloee et al., 2014; Ben-Elia and

Avineri, 2015; Liao and Chen, 2015), route/path switching behavior (Guo, 2011; Lou et al., 2017), mode and departure time choice (Gan and Ye, 2015) or activity planning (Liu, 2001).

The information supplied through various ATIS sources can be classified into three groups according to their function and needs: knowledge, risk reduction and efficiency. The provision of ATIS is used to gather knowledge of specific types of information (route guidance, descriptive information, etc.) and is delivered through various media sources (internet, text-messages, Variable Message Signs, etc.). The knowledge is applied in specific types of choices (departure time, route, or mode, etc.), made under different contexts (habitual trips, business trips, recreational trips, etc.) in various travel situations (pre- or in-trip, normal or accident conditions, etc.) to improve the overall efficiency of the trip.

For personal vehicle drivers, the provide information updates drivers on current roadway conditions—including delays, incidents, weather-related messages, travel times, emergency alerts, and alternate routes. It can be offered with value added options like sports scores, stock quotes, yellow pages, and current news. The ATIS is useful for various activities in pretrip, on-trip and posttrip phase as well as for all modes of travel including a private, public transport, or a multimodal traveler.

The pretrip travel information allows travelers to access real time inter modal transport information at home, work, and other major sites where the trip originates. Advanced pretrip traveler information system devices focus on providing real-time traffic information but often bundle it with transport and traveler services information. However, ATIS information is prominently used in during the trip to have a knowledge about traffic conditions, incidents, construction, transit schedules, and other mode choice options to drivers of personal, commercial, and public transit vehicles. For example, in case of using public transport, choosing an optimized itinerary requires knowledge about public transport network and services. On top of these, multi-modal trips, which involve multiple modes (e.g., plane, train, taxi) and highly probably various transport operators, are even more demanding.

The essentiality of the information in various modes and phases can be categorized as:

Multimodal trip planning: The ATIS information assists traveler during pre- and on-trip phase to find multimodal trip options/paths to reach their destination. Moreover, during nonrecurrent conditions, the on-trip planning computes and proposed alternate paths to reach the destination.

Route guidance service systems: The provided information helps traveler navigate throughout their trips for providing directions (drivers/pedestrians) or public transport navigation guidance (for passengers).

Advisory information: The advisory services includes information such as parking details, incident/congestion/ delay warning, weather information, public transport service schedule, traffic information, and other trip related information.

The disseminated information to drivers before and during trips allows them to make more effective travel decisions about changing routes, modes, departure times, or even destinations. The information also helps the travelers with planning, perception, analysis, and decision-making of their trip to improve convenience, safety, and efficiency of travel (Adler and Blue, 1998; Dong et al., 2010; Kumar et al., 2003).

Thus, the aim of the ATIS is to automatize the entire process of searching for relevant and useful resources (e.g., existing ATIS's, information sources, services) available on the Web, accessing them, and combining acquired information and services to build travel solutions tailored to each traveler's requests and needs.

Traveler information is delivered through various means which includes:

1. Television and radio channels: The information provided here is generally "near" real time travel conditions on the road with focus on major delays due to incidents.
2. Roadside variable message signs (VMS): The variable message signs are strategically placed along the major corridors to disseminate real time expected travel time along the corridor. For instance, on motorways the travel information can be to different exits. The VMS are also used to warn drivers of major incidents or hazards on the road.
3. Websites and apps related to route guidance and trip planning: These sources primarily focus on the route guidance and multimodal trip planning, and if connected to the network can provide dynamic updates during the travel. The information can be more personalized, where traveler has options to optimize the trip based on destination, departure time, and infrastructure usage (such as avoid toll roads). Google Maps, Waze, QldTraffic (<https://qldtraffic.qld.gov.au/>) are few examples of real time traveler information.
4. In vehicle time navigation systems, such as vehicle information and communication systems in Japan (<https://www.vics.or.jp/en/know/index.html>) and TomTom (tomtom.com).

Recent Trends in ATIS

The recent applications of ATIS are in vehicle routing and navigation systems that keep drivers informed about the traffic, vehicle, roadside condition all together in a single service. When integrated with an advanced traffic management system, information on recurrent traffic congestion can be obtained. This helps to select and display optimum routes based on real-time traffic data such as the distance/time/cost to the destination, and notification of incidents (route information navigation) can be provided. The integrated service is usually used for the in-vehicle guidance system for the recent applications.

The in-vehicle motorist services information provides motorists with commercial logos and signing for motel, eating facilities, service stations, and other signing displayed inside the vehicle to direct motorists to recreational areas, historical sites, etc.

In-vehicle signing information provide noncommercial routing, warning, regulatory, and advisory information inside the vehicle that are permanently available on the roadway.

In-vehicle safety advisory and warning systems provide warnings of unsafe conditions and situations on the vehicle/roadway ahead that could affect the driver.

The massive availability of data and services, resulted from expansion and use of internet opens a new perspective towards the design of ATIS that exploits large, dynamic, and open set of data sources as well as services coupled with advancement in artificial intelligence. Currently, numerous ATIS's are in operation. However, most, if not all, ATIS's are designed to target only certain types of travelers, travel modes, transport services provided by certain operators, and/or geographical coverage. Therefore, regardless of the significant number of existing ATIS's, travelers are quite often still required to consult multiple ATIS's and/or to seek for travel information from various sources to acquire enough information and assistance for their trips.

Though the systems possess several advantages, it faces budgeting and funding challenges. However, the same infrastructure that provides traveler information also enables more effective incident management and performance measurement—which can mean a greater return on the investment. Maintaining and upgrading these systems to reflect the most up-to-date technology requires implementation and maintenance funding. Another issue with the ATIS system usage is the safety concern by using system while driving and proper use of publicly collected traffic information when distributed by private interests. Moreover, issues are being faced in liability for distributed information that is either false, inaccurate, or unreliable and may be blamed for damage of some kind.

References

- Adler, J.L., Blue, V.J., 1998. Toward the design of intelligent traveler information systems. *Transport. Res. C: Emerg. Technol.* 6, 157–172.
- Advani, C., thakkar, S., shah, S., arkatkar, S., Bhaskar, A., 2019. A Wi-Fi sensor-based approach for examining travel time reliability parameters under mixed traffic conditions. *Transport. Dev. Econ.* 6, 1.
- Bagloee, S., Ceder, A., Bozic, C., 2014. Effectiveness of en route traffic information in developing countries using conventional discrete choice and neural-network models. *J. Adv. Transport.* 48.
- Barrpounakis, E., Geroliminis, N., 2020. On the new era of urban traffic monitoring with massive drone data: the pNEUMA large-scale field experiment. *Transport. Res. C: Emerg. Technol.* 111, 50–71.
- Ben-Elia, E., Avineri, E., 2015. Response to travel information: a behavioural review. *Transport Rev.* 35, 352–377.
- Bhaskar, A., Chung, E., 2013. Fundamental understanding on the use of Bluetooth scanner as a complementary transport data. *Transport. Res. C: Emerg. Technol.* 37, 42–72.
- Bowen, L., Budden, S.L., Smith, A.P., 2020. Factors underpinning unsafe driving: a systematic literature review of car drivers. *Transport. Res. F: Traffic Psychol. Behav.* 72, 184–210.
- Byon, Y., Shalaby, A., Abdulhai, B., 2006. Travel time collection and traffic monitoring via GPS technologies. In: 2006 IEEE Intelligent Transportation Systems Conference, 17–20 September 2006. pp. 677–682.
- Chorus, C., Molin, E., Wee, B., 2006. Travel information as an instrument to change car-drivers' travel choices: a literature review. *Social Res. Transport (SORT) Clearinghouse* 6.
- Dong, Z., Xiaoguang, Y., Haode, L., Jing, T., 2010. Internet-based advanced traveler information service: opportunities and challenges. In: 2010 International Conference on Optoelectronics and Image Processing, 11–12 November 2010, pp. 646–650.
- Gan, H., Ye, X., 2015. Whether to enter expressway or not? The impact of new variable message sign information. *J. Adv. Transport.* 49, 267–278.
- Gössling, S., 2018. ICT and transport behavior: a conceptual review. *Int. J. Sustain. Transport.* 12, 153–164.
- Guo, Z., 2011. Mind the map! The impact of transit maps on path choice in public transit. *Transport. Res. A: Policy Pract.* 45, 625–639.
- Kamnik, R., Nekrep Perc, M., Topolšek, D., 2020. Using the scanners and drone for comparison of point cloud accuracy at traffic accident analysis. *Accid. Anal. Prev.* 135, 105391.
- Kumar, P., Reddy, D., Singh, V., 2003. Intelligent transport system using GIS. In: 6th Annual International Conference, Map India, pp. 28–30.
- Liao, C.-H., Chen, C.-W., 2015. Use of advanced traveler information systems for route choice: interpretation based on a Bayesian model. *J. Intell. Transport. Syst.* 19, 316–325.
- Liu, Y.-C., 2001. Comparative study of the effects of auditory, visual and multimodality displays on drivers' performance in advanced traveler information systems. *Ergonomics* 44, 425–442.
- Lou, X.-M., Cheng, L., Chu, Z.-M., 2017. Modelling travellers' en-route path switching in a day-to-day dynamical system. *Transportmetrica B: Transport Dyn.* 5, 15–37.
- Öz, B., Özkan, T., Lajunen, T., 2010. Professional and non-professional drivers' stress reactions and risky driving. *Transport. Res. F: Traffic Psychol. Behav.* 13, 32–40.

Travel Time Reliability

Sharmili Banik*, Anil Kumar[†], Lelitha Vanajakshi*, ^{*}Department of Civil Engineering, IIT Madras, Chennai, Tamil Nadu, India; [†]Department of Civil and Environmental Engineering, IIT Patna, Patna, Bihar, India

© 2021 Elsevier Ltd. All rights reserved.

Introduction	109
Reliability Performance Measures	109
Classification of Measures	109
Statistical Range Measures	110
Buffer Measures	110
Tardy Trip Measures	111
Probabilistic Measures	111
Congestion Based Measures	112
Case Study	112
Agreement Among Reliability Measures	119
Acknowledgment	120
References	120
Further Reading	121

Introduction

Travel time, a fundamental measure in transportation, can be defined as the time taken to traverse a route between any two points of interest. Urban travel times are prone to higher degrees of variability due to the presence of traffic influencing events (traffic incidents, work zone activities, and weather), traffic demand (fluctuations in demand, special events, etc.), and presence of physical features (traffic control devices, inadequate base capacity). Due to this variability, travel times for a similar trip during different hours of a day or over the days become inconsistent. In order to capture these inconsistencies, researchers have defined several measures. Travel Time Reliability (TTR) and travel time variability, which are related, but different in their focus, are most commonly used to describe these inconsistencies. Based on literature, variability can be expressed as inconsistency in operating conditions (Lomax et al., 2003). On the other hand, TTR is defined as the consistency or dependency in travel times as measured from day to day and/or across different times in a day. In other words, it can be defined as a measure of the dispersion (or spread) of the travel time distribution (Federal Highway Administration, 2017).

While both of these measures are important, variability is related more to the facility and therefore more pertinent to transportation agencies (Lomax et al., 2003). Reliability is more pertinent to travelers since it is related to their travel experience and is significant to many transportation system users such as vehicle drivers, transit passengers, freight shippers, etc. For example, travelers and shippers want to know when they need to leave, or when the truck has to depart, in order to make an on-time arrival. Both user groups may also want to know which paths they should use to minimize the probability of unforeseen delays. Managing agencies/operators want to know where the problem spots lie; where the network segments are that make the travel times vary. If the transportation system was 100% reliable, with the same travel times all the time, none of this would be an issue.

Traffic engineers started recognizing the importance of travel time reliability as it better quantifies the benefits of traffic management and operation activities than simple averages. The initial research on reliability was mainly of qualitative in nature, mostly questionnaires ascertaining travelers' preferences with respect to the quality of their commute (Vaziri & Lam, 1983; Chang & Stopher, 1981; Prashker, 1979). However, advances in traffic data collection techniques generated huge amount of data in both real-time and historic contexts, which were used to quantify the reliability and various performance measures were reported. These are discussed in the next section.

Reliability Performance Measures

Travel time reliability measures the performance of an urban arterial. In a broader sense, reliability is an attribute of mobility and congestion. Traditionally, the dimensions of congestion are spatial (how much of the system is congested?), temporal (how long does the system got affected?), and severity-related (how much delay is there or how low are travel speeds?). Reliability adds a fourth dimension: How does congestion change from day to day? (SHRP Report, 2014). From a measurement perspective, reliability is quantified from the distribution of travel times, for any given section/network that occurs over a significant amount of time.

Classification of Measures

The travel time reliability measures can be broadly categorized as below (Lomax et al., 2003; Van Lint et al., 2008).

Table 1 Statistical range measures

Measure	Definition	Formulae
Standard Deviation	Deviation of travel times from the average travel time.	$\sqrt{\frac{(x_i - \mu)^2}{N}}$
Travel Time Window	Variation of travel time around the average expressed in a “plus-minus” form.	$Average\ Travel\ Time \pm Standard\ D$
Percent Variation	Variation of travel time around the average expressed in ratio	$\frac{Standard\ Deviation}{Average\ travel\ time} * 100$
Width (τ^{var})	Measure for width of the travel time distribution	$\tau^{var} = \frac{T90 - T10}{T50}$
Skew (τ^{skew})	Measure for skew of the travel time distribution	$\tau^{skew} = \frac{T90 - T50}{T50 - T10}$
Unreliability index (UI_r)	Measure for combination of width and skew of the travel time distribution	$UI_r = \tau^{var} \ln \left(\frac{\tau^{skew}}{L_r} \right), \text{ for } \tau^{skew} > 1 \frac{\tau^{var}}{L_r}, \text{ otherwise}$
Variability index	Ratio of the peak to off-peak variation in travel conditions	$\frac{(Upper\ 95\% - Lower\ 95\%)_{peak}}{(Upper\ 95\% - Lower\ 95\%)_{off-peak}}$

x_i is i^{th} travel time observation; μ is the average travel time, N is total number of observations; $T90$, $T50$, $T10$ are the 90th, 50th and 10th percentile travel time respectively, L_r is route length
 Note: For all the above discussed measures, higher the value of the measure, higher the unreliable condition

- Statistical range measures
- Buffer time measures
- Tardy trip measures
- Probabilistic measures
- Congestion/volume-based measures

The statistical range measures analyses travel time variability based on historical data or real-time information, while the buffer time type indicates the extra time to be allotted for a trip over the average conditions to combat the uncertainty in travel conditions. The tardy trip measures indicate unreliable conditions from the perspective of the worst trips or the unlucky travelers. Probabilistic reliability measures mainly reflect the probability that the travel time can match the specified conditions or thresholds in the form of probability distribution. Congestion/Volume-based measures represents unreliable condition in terms of congestion observed. These measures are discussed in detail below.

Statistical Range Measures

These measures generally use standard deviation or spread of the distribution to represent the travel conditions that are experienced by travelers (Lomax et al., 2003). These are often represented as a plus/minus times the standard deviation from the mean value. Measures that fall into this category include standard deviation, percent variation (coefficient of variation), travel time window, etc. Measures that fall into this category are briefly discussed in Table 1 (Lomax et al., 2003; Van Lint et al., 2008).

It has to be noted that standard deviation type of expression indicates the spread of travel time around some expected value, that is, average travel time, implicitly assuming travel times to be distributed symmetrically. For example, in case of normally distributed travel times, using a window of one standard deviation will encompass 68% of the days or peak periods or whatever time period is chosen for analysis. Analysis conducted by Van Lint and Van Zuylen (2005) showed that a symmetrical distribution can probably occur only under free flow conditions. Van Lint et al. (2008) proposed skew (τ^{skew}), width (τ^{var}) and a combination of skew and width (UI_r), that can indicate the likeliness of incurring a very bad travel time (large values for unreliable period and small otherwise) as robust indicator of reliability.

Buffer Measures

Buffer measures typically provide the extra time that travelers need to allot to their trip due to travel time variability, in order to make an on-time arrival to their destination. This extra time or “buffer” reflects the uncertainty in travel conditions. Measures in this category include buffer time, buffer time index (BTI), planning time and planning time index (PTI). Table 2 summarizes the existing buffer measures (Lomax et al., 2003; FHWA, 2017).

Previous studies showed similar trends from all of the above measures, with the planning time index as slightly exaggerated and buffer index as the most conservative (Lyman and Bertini, 2008). A study by Florida DOT (SIS Final Report, 2010) reported that buffer index could be an unstable indicator of changes in reliability as it can move in a direction opposite to the mean and percentile-based measures. This is mainly because buffer index uses both the 95th percentile and the mean travel time, and the percent change in these values can be different from time to time. If one changes more in relation to the other, counterintuitive results can appear. Fan and Gong (2017) used travel time reliability measures such as frequency of congestion (FOC) and PTI to identify and rank recurrent freeway bottlenecks and their results indicated that both FOC and PTI are capable of identifying and ranking bottlenecks.

Table 2 Buffer measures

Measure	Definition	Formulae
Buffer Time	Extra time that is allotted to ensure on-time arrival. This includes unexpected delays.	$T95 - \text{Average travel time}$
Buffer Time Index	Extra time that is allotted to ensure on-time arrival expressed in terms of percentage of the average travel time. This includes unexpected delays.	$\frac{T95 - \text{Average travel time}}{\text{Average travel time}}$
Planning Time	Total time that travelers should allocate towards their trip to ensure on time arrival. This includes buffer time as well.	$T95$
Planning Time Index	The planning time index compares near-worst case travel time to a travel time in light or free-flow traffic. -flow travel time. This Includes typical delay as well as unexpected delays	$\frac{T95}{\text{Free flow travel time}}$

T95 is the 95th percentile travel time

Note: For all the above discussed measures, higher the value of the measure, higher the unreliable condition

Table 3 Tardy trip measures

Measure	Definition	Formulae
On-time Arrival	Employs a threshold travel time to the trips to determine the percentage of trip that can be termed reliable. The higher value of the measure indicates good reliability.	$100 - (\% \text{ of trips greater than } 110\% \text{ of avg. TT})$ where, TT stands for Travel Time
Misery index	Measure that indicates how much time the worst trips exceeded the average travel time, expressed in percentage of the average travel time. Higher value indicates, greater unreliable conditions.	$\frac{\text{Avg. TT of longest 20\% of trips} - \text{Avg. TT}}{\text{Avg. TT}}$

Tardy Trip Measures

Tardy trip measures use the number of late trips to measure system unreliability. Key measures in this category are on-time arrival and misery index. **Table 3** explains the tardy trip measures in brief (Lomax et al. 2003).

Lomax et al. (2003) reported two issues related to the threshold value concept (1) threshold value for late arrival for all trips may not vary linearly and (2) it is not only related to trip duration but also to the previous and next activity. Also, they reported that “Being late may be more acceptable if the traveler is coming from an important activity and it may be less acceptable if they are going for an important activity.” Hence, more testing of the acceptable lateness concept needs to be done. Misery Index (MI) can evaluate the negative attribute of trip reliability and it can be judged by how much the average travel times of the unluckiest trips (20% of the worst trips in an hour or day) exceeded the mean travel time (Lomax et al., 2003). It provides an idea on how bad the worst trips are and to what extent of extreme travel times the travelers are exposed to.

Probabilistic Measures

In this category, travel time reliability is expressed as a probabilistic measure. Probabilistic travel measures often use a threshold travel time or a predefined window to differentiate between reliable and unreliable travel times (Van Lint et al., 2008). Asakura (1996) proposed a trip to be reliable if the probability of the trip between a given OD pair to be completed within a specified interval of time is high. The Dutch Ministry of Transport, Public Works and Water Management proposed a reliability measure which stated that all trips should be made within 20% bounds of median travel time (Verker and Waterstaat, 2004). **Table 4** shows probabilistic measures used for quantifying reliability (van Lint et al., 2008). Along with these two measures, concepts of cumulative curves can be used to assess the reliability by plotting travel times in x-axis and cumulative probability on y-axis.

Table 4 Probabilistic measures

Measure	Definition	Formulae
Pr(α)	Determines probability that travel times occur larger than multiple of some predefined travel time threshold. Higher value indicates greater unreliable condition	$Pr(\alpha) = P(TT_i > \alpha * TT_{50})$, e.g. $\alpha = 1.2$, where TT_{50} is the 50th percentile travel time A use of 1.2 as α refers to the probability that travel time is larger than the median travel time + 20 %
AVV 2004 Probabilistic Measures	Determines if at least 95% of all travel times does not deviate more than 10 min from the median travel time for short routes, while for longer routes, if 95% of the travel times are within a 20% margin around the median.	$Pr(TT_i \leq 10 \text{ mins} + TT_{50}) > 95\%$, for routes < 50 km $Pr(TT_i \leq 1.2 \text{ mins} + TT_{50}) > 65\%$ for routes > 50 km



Figure 1 Study corridor. Source: Open Street Maps.

Congestion Based Measures

Volume, speed, and congestion frequency graphs are reported to depict reliability (Lomax et al., 2003). These measures help to view the variability in traffic condition and determine the unreliable periods or days. Frequency of congestion (FOC) is a well-known reliability measure used in many studies to report reliability (FHWA, 2017; Fan & Gong, 2017). FOC is defined as the percent of days or time or trips during which the travel times exceed some threshold value. The threshold value can be chosen at the discretion of the analyst. Higher values of FOC indicate unreliable conditions. FOC is relatively easy to compute if continuous traffic data is available (FHWA, 2017). Each of these measures quantify the reliability in terms of certain selected variables and got advantages and limitations. Next section discusses a case study where these measures are calculated and analyzed for their performance.

Case Study

To better understand and depict the applicability of the reliability measures, a case study is carried out by collecting data from a 1.6 km urban arterial connecting the Bharathi Nagar and Vijaya Nagar intersection, located at southern part of the Chennai city as shown in Figure 1.

The data for the proposed analysis has been collected for a period of 1 month using Global Positioning System (GPS) fitted buses that are plying in the study corridor. The raw GPS data provided information about the position (latitude and longitude) of the bus at every 5 sec interval, which can be further processed to obtain travel times. For examining the variations of reliability along the study corridor, the 1.6 km corridor was segmented into eight subsections of 200 m length and the time taken to cover each subsection was calculated using linear interpolation technique. The travel time data thus generated is further used to generate reliability metrics.

Traffic in any roadway is expected to follow certain temporal and spatial patterns. Temporal patterns can include within day variation (peak and off-peak) and day-to-day variations (weekday vs weekend or day of the week variation), whereas spatial patterns may be specific to certain sections or adjacent sections. To understand these influences on travel time, the data collected from the study corridor was analyzed. To start with, the spatial analysis was done to identify the variation in travel time across various segments in the study corridor. Figure 2A is a sample plot showing travel time variation over various sections.

From Figure 2A, it can be observed that there exists a strong segment specific pattern in travel time. For example, section 7 shows higher central value and spread in travel times compared to other sections. Hence, it was decided to analyze each section separately. In the next level, analysis was conducted to identify the variations in travel time across days as shown in Figure 2B for a sample section. From Figure 2B, it can be observed that travel times observed across various days of the week does not show much difference. Both central values and spread are in similar range. This was observed to be true for all other sections as well. Hence, it was decided to consider the travel time data from all days together while calculating the reliability metrics. Similarly, within day variation, that is, hourly box plots are plotted for each section separately and a sample hourly box plot for a day is shown in Figure 2C. From Figure 2C, it can be observed that there are significant hourly variations over the day. This was observed to be true for all other sections for all days and hence, every hour is analyzed separately to capture the temporal variations.

Thus, based on the above observations, reliability metrics are calculated for each section and each hour separately after clubbing the travel time data of all days together to analyze the spatial and temporal reliability variations. Further, outliers are removed using median absolute deviation technique (Leys et al., 2013) before calculating the reliability measures. Reliability measures under each group (statistical, buffer, tardy trip, probabilistic, and congestion measures) calculated and related discussions are presented in the next section.

1. *Statistical range measures* All measures that are presented in Table 1 are calculated and the applicability of these measures is discussed. Figure 3 to Figure 9 show different statistical range reliability measures calculated using the data. In these figures, a continuous color coding is used to visually demonstrate reliability scenario along with the numbers. Under this color coding, green refer to very good reliability, yellow represent moderate reliability, and red a very poor reliability condition. From Figure 3

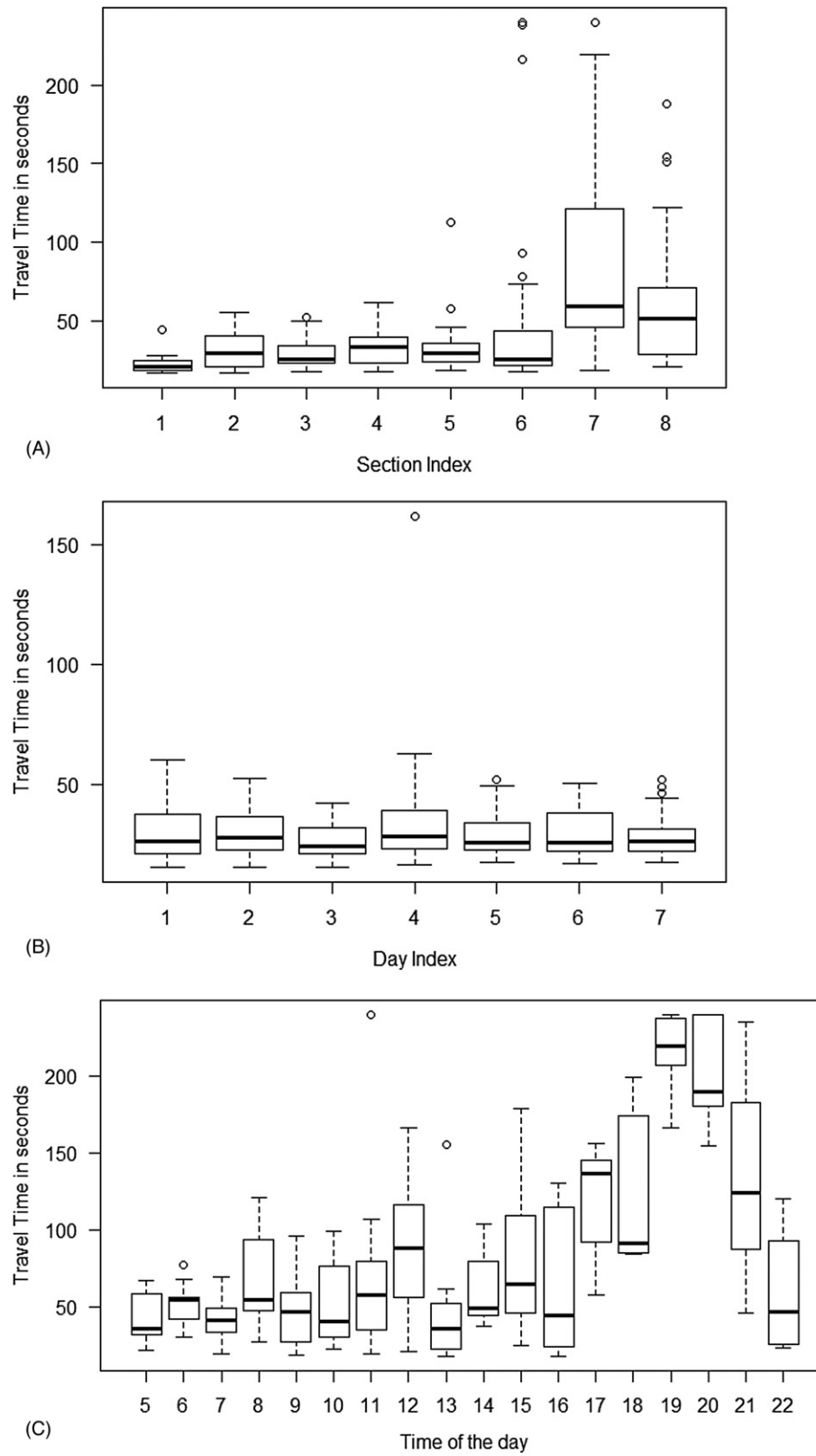


Figure 2 Preliminary analysis of travel times in the considered study route.

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	3.23	3.15	1.76	4.23	2.21	4.67	2.73	2.02	2.82	2.96	2.89	2.49	2.80	3.72	5.41	3.33	0.43
Section 2	6.17	10.42	7.26	9.40	9.70	6.24	6.08	9.80	7.62	12.08	7.52	11.18	8.42	11.43	11.54	9.21	6.30
Section 3	7.15	7.28	6.76	5.50	11.54	9.10	10.24	10.27	9.75	9.52	10.30	3.47	10.80	11.14	8.62	5.07	1.81
Section 4	3.66	6.68	3.81	13.86	6.25	9.12	3.65	8.00	7.51	9.58	7.53	8.33	10.31	9.77	7.57	10.20	9.11
Section 5	6.76	4.41	7.53	13.69	10.85	5.20	7.58	6.55	8.91	7.26	8.70	7.12	8.98	12.06	20.37	3.35	3.11
Section 6	6.53	4.59	1.39	7.10	5.68	6.87	3.70	12.54	8.11	4.97	9.79	6.80	5.07	7.64	79.04	83.07	16.58
Section 7	15.55	13.68	14.45	25.64	39.19	34.61	27.22	42.34	19.70	24.90	41.50	42.54	39.46	49.98	43.90	29.01	65.36
Section 8	16.71	9.57	15.45	32.97	19.06	16.81	36.97	20.63	31.64	20.38	27.37	32.37	28.34	37.69	39.65	15.48	21.83

Figure 3 Standard deviation (expressed in seconds).

	section1	section2	section3	section4	section5	section6	section7	section8
5:00 AM	19.79±3.23	25.53±6.16	24.58±7.14	21.29±3.66	25.92±6.75	24.81±6.53	39.75±15.55	39.68±16.71
6:00 AM	20.11±3.15	31.91±10.41	25.25±7.28	29.08±6.67	24.55±4.41	21.98±4.58	52.47±13.67	32.34±9.57
7:00 AM	18.14±1.76	25.81±7.25	26.82±6.76	23.14±3.81	31.51±7.52	20.14±1.39	40.91±14.44	48.34±15.44
8:00 AM	22.48±4.22	33.64±9.39	25.59±5.51	37.17±13.85	34.71±13.68	25.53±7.09	56.61±25.64	60.19±32.96
9:00 AM	20.99±2.21	29.62±9.69	37.36±11.53	27.91±6.25	38.71±10.84	24.82±5.67	55.09±39.19	55.15±19.05
10:00 AM	22.55±4.66	26.31±6.24	29.54±9.09	29.63±9.12	30.09±5.21	25.03±6.86	61.45±34.61	52.57±16.81
11:00 AM	19.84±2.72	24.11±6.07	30.25±10.24	22.66±3.65	31.81±7.58	22.78±3.71	60.21±27.22	71.09±36.97
12:00 PM	19.20±2.02	28.17±9.81	28.84±10.27	27.65±7.99	27.53±6.54	29±12.53	85.05±42.33	46.91±20.63
1:00 PM	20.91±2.81	22.74±7.62	27.99±9.74	29.11±7.51	27.62±8.91	28.68±8.11	41.87±19.71	60.01±31.64
2:00 PM	21.82±2.96	33.14±12.07	31.04±9.51	29.84±9.57	30.92±7.26	24.81±4.96	73.51±24.89	55.68±20.37
3:00 PM	19.45±2.88	29.91±7.52	28.63±10.29	26.98±7.53	30.88±8.71	28.91±9.78	79.45±41.49	50.85±27.36
4:00 PM	19.42±2.48	32.81±11.17	24.57±3.47	28.63±8.33	27.87±7.12	25.39±6.81	71.56±42.54	64.27±32.36
5:00 PM	20.95±2.79	34.98±8.41	29.74±10.79	33.33±10.31	36.59±8.98	25.43±5.07	100.71±39.46	62.23±28.34
6:00 PM	22.41±3.71	36.05±11.42	31.21±11.13	34.73±9.76	37.27±12.06	26.03±7.63	105.33±49.98	71.11±37.68
7:00 PM	25.26±5.41	35.61±11.53	32.86±8.61	28.01±7.57	44.67±20.37	111.71±79.03	168.31±43.91	83.83±39.64
8:00 PM	23.85±3.32	33.70±9.21	27.52±5.07	35.84±10.19	29.87±3.35	131.74±83.06	184.01±29.01	55.52±15.47
9:00 PM	18.48±0.42	34.01±6.31	22.04±1.81	34.21±9.11	25.68±3.11	34.55±16.57	123.14±65.35	47.11±21.82

Figure 4 Travel time window (expressed in seconds).

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	16.34	15.67	9.71	18.80	10.54	20.70	13.75	10.53	13.48	13.59	14.84	12.80	13.36	16.61	21.43	13.96	2.31
Section 2	24.15	32.65	28.12	27.93	32.74	23.73	25.20	34.80	33.52	36.43	25.14	34.06	24.07	31.70	32.39	27.33	18.53
Section 3	29.08	28.83	25.20	21.50	30.88	30.79	33.85	35.61	34.82	30.65	35.96	14.13	36.30	35.70	26.22	18.42	8.21
Section 4	17.20	22.95	16.48	37.27	22.40	30.79	16.13	28.92	25.78	32.08	27.91	29.09	30.92	28.13	27.02	28.45	26.62
Section 5	26.07	17.94	23.89	39.43	28.03	17.29	23.84	23.77	32.25	23.48	28.18	25.56	24.54	32.37	45.61	11.22	12.12
Section 6	26.33	20.88	6.92	27.80	22.87	27.43	16.24	42.39	28.27	20.02	33.84	26.77	19.94	29.33	70.75	63.05	47.96
Section 7	39.12	26.06	35.31	45.29	71.14	56.29	45.22	49.77	47.05	33.87	52.23	59.45	39.18	47.45	26.09	15.76	53.07
Section 8	42.12	29.60	31.95	54.76	34.55	31.98	52.01	43.98	52.73	36.59	53.81	50.36	45.54	52.99	47.29	27.87	46.34

Figure 5 Percent variation (expressed as percentage).

to Figure 6, it can be observed that (1) section 7 and 8 show poor reliability in almost all time periods and section 6 shows relatively poor reliability in the evening peak hours and (2) starting sections 2, 3 are more vulnerable in the afternoon peak hours. All these agree with the field condition observed. Although, these measures are able to represent field scenario, they suffer from (1) the assumption of symmetrical travel time distribution and (2) giving equal weight age to late and early arrivals, while travelers are concerned more about late arrivals. From Figure 7 to Figure 9, it can be observed that variability index, skew and unreliability index were not able to represent field condition. Data was analyzed further to understand this and it was observed that the left hand side of the distribution (T50-T10) is more steeper than the right hand side of the distribution (T90-T50), resulting in high value of skew. This means, skew only identifies a period as unreliable when the distribution is highly left-skewed, which occurs mostly during congestion onset or offset, that is, when the distribution is highly skewed. Thus, skew may be able to capture the congestion onset/offset better, but may not capture reliability during regular peak hours. A similar contradictory observation can be seen from Figure 9 as well indicating section 6 to be more unreliable than section 7 and 8. It may be

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	0.40	0.33	0.20	0.39	0.30	0.52	0.29	0.23	0.36	0.25	0.31	0.28	0.19	0.43	0.53	0.33	0.04
Section 2	0.64	0.75	0.70	0.71	0.91	0.57	0.67	0.98	0.89	1.00	0.63	0.71	0.49	0.74	0.73	0.63	0.41
Section 3	0.91	0.80	0.65	0.45	0.68	0.81	0.72	0.96	1.01	0.83	0.95	0.37	1.10	0.84	0.63	0.46	0.17
Section 4	0.50	0.61	0.38	0.87	0.59	0.81	0.38	0.73	0.61	0.73	0.72	0.65	0.50	0.69	0.75	0.71	0.56
Section 5	0.66	0.43	0.70	0.98	0.62	0.50	0.55	0.62	0.81	0.55	0.68	0.60	0.61	0.86	0.90	0.25	0.27
Section 6	0.61	0.60	0.15	0.61	0.52	0.61	0.44	0.93	0.70	0.34	0.94	0.69	0.39	0.46	1.83	1.52	1.23
Section 7	1.13	0.60	0.94	1.23	1.52	1.50	1.09	1.32	1.07	0.77	1.32	1.43	0.91	1.51	0.58	0.33	1.40
Section 8	1.12	0.69	0.65	1.30	0.79	0.78	1.18	1.26	1.23	0.95	1.42	1.02	0.99	1.21	0.68	0.46	1.24

Figure 6 Width (τ_{var}) (expressed as ratio).

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	3.88	1.95	1.37	0.51	0.93	1.33	2.53	2.03	1.51	1.10	2.50	1.08	1.34	1.87	4.12	2.09	1.47
Section 2	1.06	0.91	1.82	1.68	2.75	1.56	1.93	3.46	12.39	2.64	0.86	0.72	0.68	1.11	1.65	2.36	0.38
Section 3	3.16	2.72	2.70	1.00	0.56	1.64	0.80	1.95	3.38	1.90	2.88	2.01	4.82	2.99	2.43	3.57	0.47
Section 4	2.66	0.80	1.94	1.56	3.09	2.28	3.26	3.05	0.93	1.23	2.33	2.36	0.88	1.56	3.88	1.14	0.77
Section 5	4.03	1.56	1.37	1.95	1.58	2.07	1.46	1.31	2.34	1.10	1.01	1.99	0.72	3.25	3.28	0.53	1.47
Section 6	1.56	3.67	0.39	2.54	7.17	2.06	2.92	3.56	2.06	0.80	2.70	3.52	2.42	2.00	1.36	1.08	2.37
Section 7	2.74	0.79	0.95	1.66	2.19	2.13	1.00	1.60	1.22	0.77	1.94	0.98	1.72	3.18	1.36	0.81	1.31
Section 8	2.53	1.99	1.30	1.19	0.53	0.69	1.38	2.04	1.30	1.53	2.37	0.99	0.95	0.93	0.70	1.56	2.27

Figure 7 Skew (τ_{skew}) (expressed as ratio).

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	2.73	1.11	0.31	1.93	1.48	0.75	1.37	0.82	0.74	0.12	1.43	0.11	0.28	1.34	3.77	1.23	0.09
Section 2	0.19	3.76	2.10	1.83	4.58	1.26	2.22	6.10	11.16	4.84	3.13	3.55	2.44	0.37	1.81	2.71	2.07
Section 3	5.21	4.03	3.22	2.23	3.42	1.99	3.59	3.20	6.13	2.67	5.02	1.31	8.62	4.57	2.79	2.94	0.87
Section 4	2.43	3.06	1.27	1.93	3.34	3.34	2.25	4.09	3.05	0.75	3.07	2.78	2.50	1.54	5.11	0.48	2.80
Section 5	4.59	0.96	1.10	3.25	1.42	1.83	1.03	0.86	3.44	0.25	0.02	2.07	3.04	5.06	5.36	1.23	0.53
Section 6	1.35	3.91	0.76	2.87	5.12	2.20	2.38	5.89	2.55	1.68	4.70	4.37	1.71	1.60	2.80	0.56	5.32
Section 7	5.67	3.01	4.71	3.10	5.94	5.67	0.01	3.11	1.08	3.84	4.39	7.16	2.49	8.74	0.89	1.65	1.86
Section 8	5.20	2.37	0.85	1.16	3.97	3.88	1.90	4.52	1.64	2.01	6.14	5.12	4.97	6.06	3.39	1.03	5.10

Figure 8 Unreliability index (Ulr) (expressed as ratio).

Section 1	1.477187
Section 2	1.279468
Section 3	1.012271
Section 4	1.866031
Section 5	1.812528
Section 6	12.85249
Section 7	3.272336
Section 8	2.914822

Figure 9 Variability index (expressed as ratio).

because variability index captures only the difference between upper and lower 95% confidence value, but ignores the absolute value of the travel times. Overall, though unreliability index, variability index and skew were able to depict the reality to some extent, they are highly affected by the distribution of the collected data, that is, these measures were either penalized by the travel times in left hand side of the median (underestimating the unreliability) or right hand side of the median (overestimating the unreliability). On the other hand, though other metrics were able to completely depict the reality, they completely depend on spread of travel time around some expected travel time value and such information can only be an abstract indicator for the travelers.

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	6.91	5.41	3.04	5.78	3.04	6.12	4.20	3.55	4.44	4.12	4.92	4.03	4.22	5.55	8.78	6.00	0.59
Section 2	9.30	14.51	11.83	17.93	15.03	9.59	10.48	16.00	13.40	16.54	10.94	14.52	9.99	15.49	17.91	13.96	5.82
Section 3	14.42	12.83	12.69	8.66	14.17	16.40	13.29	18.90	13.93	16.55	16.30	6.63	20.14	17.39	13.07	8.68	1.97
Section 4	6.69	9.73	6.34	25.48	8.31	16.68	5.77	14.95	9.90	13.74	12.93	12.44	15.63	16.10	12.79	16.03	14.82
Section 5	13.00	6.58	11.06	21.25	17.80	8.31	11.67	9.56	13.14	10.17	12.36	12.55	11.74	22.97	33.56	4.56	5.15
Section 6	8.95	8.22	1.66	10.46	9.64	14.00	7.40	17.96	12.22	7.69	18.89	14.28	8.06	12.64	112.62	93.42	29.35
Section 7	23.84	19.79	26.19	42.61	66.66	47.02	39.67	75.77	30.68	34.61	64.69	54.84	50.15	87.43	57.97	40.98	101.92
Section 8	22.92	20.58	19.01	58.58	24.54	24.11	61.37	33.77	50.15	27.45	43.17	56.27	43.02	60.91	63.08	24.16	32.66

Figure 10 Buffer time (expressed in seconds).

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	34.91	26.90	16.74	25.69	14.50	27.14	21.14	18.48	21.21	18.87	25.30	20.75	20.15	24.79	34.75	25.16	3.20
Section 2	36.43	45.47	45.83	53.29	50.75	36.44	43.46	56.79	58.93	49.91	36.59	44.26	28.56	42.96	50.27	41.41	17.12
Section 3	58.66	50.81	47.31	33.85	37.92	55.51	43.92	65.54	49.77	53.32	56.93	27.00	67.71	55.73	39.76	31.55	8.93
Section 4	31.39	33.45	27.38	68.54	29.78	56.30	25.45	54.07	34.00	46.03	47.90	43.43	46.89	46.34	45.64	44.72	43.32
Section 5	50.15	26.78	35.12	61.20	45.98	27.62	36.67	34.72	47.56	32.88	40.02	45.00	32.08	61.62	75.12	15.28	20.04
Section 6	36.07	37.40	8.24	40.97	38.81	55.92	32.49	60.70	42.59	31.00	65.32	56.23	31.67	48.55	100.81	70.91	84.93
Section 7	59.98	37.70	64.01	75.26	120.99	76.48	65.90	89.08	73.27	47.08	81.42	76.63	49.79	83.00	34.44	22.27	82.77
Section 8	57.76	63.63	39.32	97.31	44.49	45.86	86.32	71.98	83.57	49.29	84.88	87.55	69.12	85.65	75.24	43.51	69.35

Figure 11 Buffer time index (expressed in percentage).

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	26.71	25.52	21.19	28.26	24.04	28.67	24.04	22.75	25.34	25.94	24.38	23.46	25.18	27.95	34.04	29.86	19.08
Section 2	34.84	46.42	37.63	51.57	44.65	35.89	34.58	44.18	36.15	49.69	40.86	47.34	44.98	51.54	53.52	47.67	39.82
Section 3	39.00	38.09	39.51	34.26	51.53	45.95	43.55	47.75	41.92	47.60	44.93	31.21	49.89	48.59	45.93	36.20	24.02
Section 4	27.98	38.82	29.48	62.66	36.23	46.32	28.43	42.61	39.02	43.58	39.92	41.07	48.97	50.83	40.81	51.87	49.04
Section 5	38.92	31.13	42.56	55.96	56.50	38.41	43.48	37.10	40.76	41.09	43.24	40.43	48.33	60.24	78.23	34.44	30.83
Section 6	33.75	30.20	21.81	36.00	34.46	39.03	30.19	47.54	40.91	32.51	47.81	39.68	33.50	38.68	224.33	225.16	63.91
Section 7	63.60	72.26	67.11	99.23	121.76	108.51	99.88	160.83	72.55	108.12	144.14	126.40	150.86	192.76	226.28	224.99	225.07
Section 8	62.61	52.93	67.36	118.77	79.69	76.69	132.46	80.69	110.16	83.13	94.03	120.54	105.25	132.03	146.92	79.69	79.76

Figure 12 Planning time (expressed in seconds).

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	1.67	1.59	1.32	1.77	1.50	1.79	1.50	1.42	1.58	1.62	1.52	1.47	1.57	1.75	2.13	1.87	1.19
Section 2	2.18	2.90	2.35	3.22	2.79	2.24	2.16	2.76	2.26	3.11	2.55	2.96	2.81	3.22	3.35	2.98	2.49
Section 3	2.44	2.38	2.47	2.14	3.22	2.87	2.72	2.98	2.62	2.97	2.81	1.95	3.12	3.04	2.87	2.26	1.50
Section 4	1.75	2.43	1.84	3.92	2.26	2.90	1.78	2.66	2.44	2.72	2.49	2.57	3.06	3.18	2.55	3.24	3.07
Section 5	2.43	1.95	2.66	3.50	3.53	2.40	2.72	2.32	2.55	2.57	2.70	2.53	3.02	3.76	4.89	2.15	1.93
Section 6	2.11	1.89	1.36	2.25	2.15	2.44	1.89	2.97	2.56	2.03	2.99	2.48	2.09	2.42	14.02	14.07	3.99
Section 7	2.89	3.28	3.05	4.51	5.53	4.93	4.54	7.31	3.30	4.91	6.55	5.75	6.86	8.76	10.29	10.23	10.23
Section 8	3.91	3.31	4.21	7.42	4.98	4.79	8.28	5.04	6.89	5.20	5.88	7.53	6.58	8.25	9.18	4.98	4.99

Figure 13 Planning time index (expressed in ratio).

2. *Buffer measures* Figure 10 to Figure 13 show different buffer reliability measures calculated for the data under consideration, where the reliability status is indicated by the previously used color coding. Here, buffer time considers only unexpected delays and planning time considers both unexpected and typical delays. From Figure 10 to Figure 13, it can be observed that all buffer time measures were able to depict the field condition to a great extent. However, it has to be noted that: (1) buffer time index has to be cautiously used because of the unstable indication of changes in reliability as it can move in a direction opposite to the mean and percentile-based measures (SIS Final Report, 2010) and (2) planning time index, that shows exaggerated trends when compared to buffer time index and planning time (Lyman and Bertini, 2008). For example, from Figure 13, it can be observed that at time slot 7–8 pm, planning time (Figure 12) for section 6 is 224.33 secs and section 7 is 226.28 secs and corresponding

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	76.47	81.25	85.71	76.47	81.82	66.67	76.47	80.00	80.00	76.92	81.25	85.71	83.33	76.92	70.00	81.25	100.00
Section 2	73.68	56.25	55.56	72.22	63.64	77.27	72.22	64.71	77.78	57.14	64.71	68.75	58.33	57.14	70.00	76.47	70.00
Section 3	73.68	75.00	75.00	81.25	54.55	70.83	70.59	66.67	63.64	71.43	70.59	78.57	72.73	66.67	72.73	73.33	100.00
Section 4	77.78	75.00	73.33	68.42	63.64	70.83	78.57	63.16	54.55	57.14	64.71	60.00	75.00	61.54	63.64	56.25	81.82
Section 5	52.63	62.50	66.67	68.42	63.64	60.87	60.00	63.16	60.00	53.85	61.11	62.50	72.73	69.23	55.56	90.91	81.82
Section 6	73.68	80.00	93.33	70.59	70.00	68.18	84.62	62.50	63.64	75.00	68.75	73.33	77.78	80.00	50.00	53.33	80.00
Section 7	65.00	68.75	66.67	70.59	69.23	58.33	61.54	65.00	70.00	61.54	55.56	56.25	54.55	76.92	66.67	83.33	64.29
Section 8	55.56	64.71	61.11	70.59	53.85	69.57	68.75	65.00	72.73	61.54	70.59	62.50	63.64	66.67	81.82	70.00	62.50

Figure 14 On-time arrival (expressed in percentage).

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	0.25	0.22	0.14	0.24	0.13	0.30	0.19	0.18	0.21	0.18	0.19	0.18	0.17	0.23	0.34	0.20	0.02
Section 2	0.35	0.42	0.38	0.41	0.44	0.32	0.39	0.52	0.56	0.49	0.31	0.42	0.27	0.40	0.48	0.41	0.17
Section 3	0.46	0.45	0.37	0.27	0.35	0.48	0.44	0.52	0.47	0.48	0.54	0.22	0.50	0.52	0.37	0.30	0.09
Section 4	0.29	0.30	0.36	0.58	0.29	0.44	0.25	0.43	0.31	0.43	0.40	0.42	0.40	0.39	0.36	0.37	0.32
Section 5	0.38	0.24	0.34	0.61	0.35	0.25	0.32	0.34	0.47	0.30	0.37	0.36	0.30	0.50	0.67	0.11	0.15
Section 6	0.39	0.37	0.08	0.40	0.38	0.35	0.24	0.56	0.38	0.26	0.49	0.45	0.29	0.45	0.86	0.70	0.82
Section 7	0.58	0.33	0.49	0.68	1.04	0.75	0.60	0.74	0.71	0.42	0.73	0.69	0.48	0.76	0.32	0.18	0.73
Section 8	0.56	0.42	0.42	0.74	0.40	0.40	0.69	0.66	0.71	0.47	0.81	0.63	0.54	0.70	0.53	0.41	0.65

Figure 15 Misery index (expressed in ratio).

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	0.18	0.13	0.07	0.06	0.00	0.25	0.12	0.07	0.20	0.08	0.13	0.07	0.08	0.23	0.30	0.13	0.00
Section 2	0.21	0.25	0.39	0.28	0.36	0.14	0.22	0.41	0.22	0.43	0.29	0.31	0.08	0.36	0.30	0.24	0.00
Section 3	0.26	0.25	0.25	0.13	0.18	0.29	0.29	0.33	0.45	0.21	0.41	0.14	0.27	0.33	0.27	0.27	0.00
Section 4	0.22	0.25	0.20	0.32	0.36	0.29	0.21	0.42	0.27	0.29	0.35	0.40	0.25	0.31	0.36	0.19	0.18
Section 5	0.37	0.25	0.33	0.32	0.36	0.26	0.40	0.32	0.40	0.46	0.39	0.25	0.36	0.23	0.22	0.18	0.18
Section 6	0.21	0.20	0.00	0.29	0.30	0.36	0.15	0.44	0.36	0.08	0.31	0.27	0.22	0.20	0.50	0.47	0.30
Section 7	0.40	0.19	0.33	0.29	0.31	0.46	0.38	0.35	0.30	0.23	0.44	0.44	0.45	0.38	0.33	0.08	0.36
Section 8	0.44	0.24	0.39	0.29	0.15	0.26	0.31	0.35	0.27	0.38	0.29	0.19	0.27	0.25	0.18	0.20	0.38

Figure 16 $Pr(\alpha)$.

planning time index (Figure 13) are 14.02 and 10.29 secs, respectively. Planning time at this period differs by only 0.87% between the sections whereas planning time index differs by almost 40%. This shows that even though the 95th percentile travel times are almost equal, section 6 is more unreliable at this period. This may be because of the presence of wide variation across a free flow travel and a worst case travel in section 6 compared to section 7. This aspect can be captured by the planning time index. Another benefit of this measure, is that it can be directly compared against travel time index on similar numeric scales, which is a measure of average congestion (Lomax et al. 2003). Overall, these measures connect to the way decisions are made by the travelers. The effect of uncertainty on decision-making can be more related by these measures as it evaluate the real traffic conditions against average condition. However, measures like buffer time index has to be cautiously used because of the unstable indication of changes in reliability over time.

3. *Tardy trip measures* Figure 14 and Figure 15 show different tardy trip measures calculated for the current data, where the reliability status is indicated by the same color coding. Figure 14 shows the percent of trips that lies within an acceptable lateness threshold. From Figure 14, it can be observed that a few time periods for some sections are highly unreliable due to relatively wide spread of travel time (between average and 110% of average travel time). However, the reliability metric is not consistent with the field congestion as well as other measures. This may be because on-time arrival considers only average and average+110% travel times whereas measures like planning time index considers the variability between free flow and near maximum travel time. Upon detailed investigation, it was also observed that the reliable periods indicated by on-time arrival measure shows worse travel times than the unreliable ones. For instance, at 8 pm, section 6 (marked as unreliable by on-time arrival) experience an average travel time of 131.74 secs whereas section 7 (marked as reliable by on-time arrival) experienced about 184.01 secs. This congestion condition could not be captured by the on-time arrival which may possibly due to the choice of the threshold value. On the other hand, the Misery Index that evaluates the negative attributes of trip reliability (how much the average travel times of the unluckiest trips exceeded the mean travel time) shows that section 7 and 8 has poor reliability in most of the time periods and

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%
Section 2	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%
Section 3	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%
Section 4	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%
Section 5	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%
Section 6	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%
Section 7	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%
Section 8	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%	>95%

Figure 17 $\Pr(TT_i \leq 10 \text{ mins} + TT50)$.

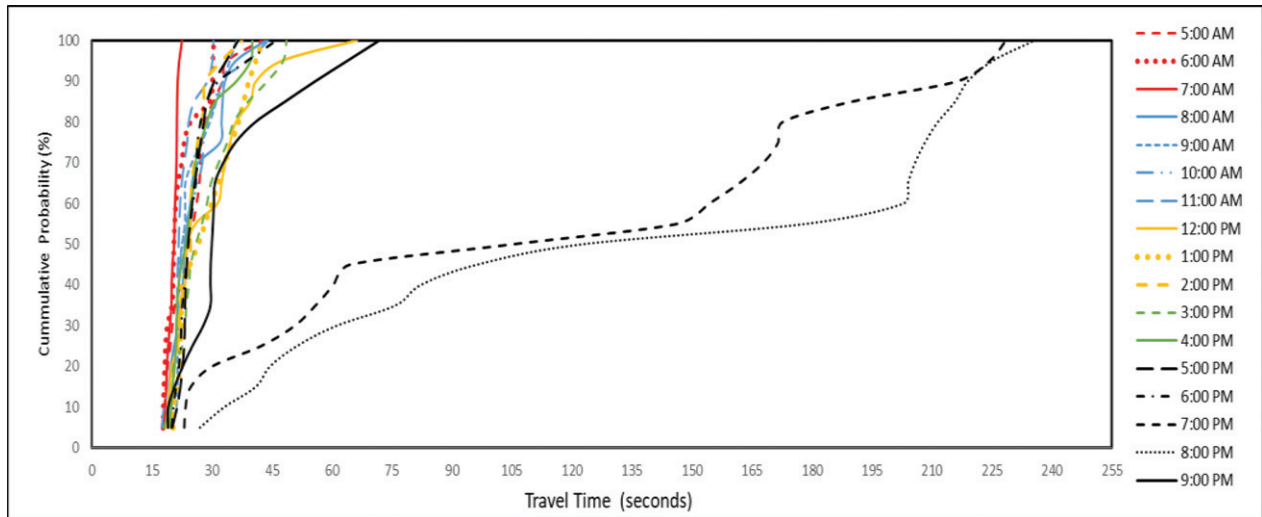


Figure 18 Hourly cumulative probability plot for section 6.

section 6 is critical during evening peak hours. The results are consistent with results from buffer measures and field scenario. Typically, this measure is useful for planning agencies and traffic analysts to identify critical areas and time periods. Overall, based on the metrics calculated using tardy trip measures, it can be observed that misery index was able to depict the field condition to a great extent and it has been recommended as a good measure of reliability (Lomax et al., 2003). However, the other tardy trip measure, “on-time arrival” could not depict the complete picture as it ignores the wide variation between the light traffic and heavy traffic condition while indicating reliability. On the other hand, the threshold used in on-time measures needs to be carefully defined as it does not vary linearly for every trip.

4. **Probabilistic methods** Figure 16 to Figure 19 show different probabilistic measures calculated for the case data, where the reliability status is indicated by the same color coding used earlier and cumulative curves. Figure 16 shows the $\Pr(\alpha)$ that determines the probability of trips that may take longer travel time than predefined travel time threshold where the higher value indicates greater unreliable condition. Van Lint et al. (2008) suggested 1.2 times median travel time as the threshold and also reported that it should be decided based on the interest of the analyst. In Figure 16, the probability was calculated by taking threshold as 1.1 median travel time, that is, $\Pr(\alpha = 1.1)$. From Figure 16, it can be observed that a few time periods in some sections are highly unreliable, with high relative spread of travel time (between median and median+10%). Also, it can be observed that the results are not consistent with statistical and buffer measures. This may be because it considers only median and 10% travel times above the median whereas buffer measures consider the variability in travel time between free flow and near maximum travel time. Figure 17 shows the AVV 2004 Probabilistic Measure [$\Pr(TT_i \leq 10 \text{ mins} + TT50)$], that determines if at least 95% of all travel times does not deviate more than 10 min from the median travel time for short routes (less than 50 km), while for longer routes, if 95% of the travel times are within a 20% margin around the median. Using the data under consideration, $\Pr(TT_i \leq 10 \text{ mins} + TT50)$ is calculated and observed that in all cases, the probability is greater than 95% indicating no unreliable periods. This measure could not capture the reliability condition as the threshold is too high for a finer level of spatial analysis as the case here (200 m sections). Hence, for very short routes or segment level analysis, the threshold has to be decided cautiously. Under probabilistic analysis of reliability, probability plots are also attempted. Figure 18 and Figure 19 show sample plots for a selected section to check hourly variations and for a selected hour to check spatial variations. From Figure 18, it can be observed that there is a wide spread in travel times, corresponding to evening peak hours and can be interpreted as unreliable periods as travelers may

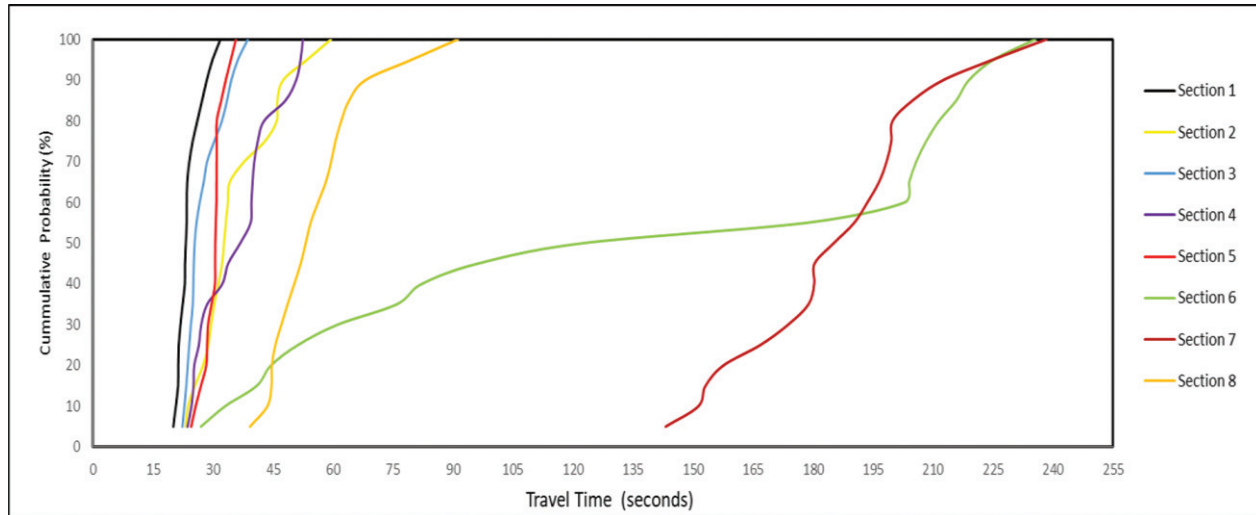


Figure 19 Section-wise cumulative probability plot.

	5:00 AM	6:00 AM	7:00 AM	8:00 AM	9:00 AM	10:00 AM	11:00 AM	12:00 PM	1:00 PM	2:00 PM	3:00 PM	4:00 PM	5:00 PM	6:00 PM	7:00 PM	8:00 PM	9:00 PM
Section 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Section 2	0.00	6.25	0.00	11.11	0.00	0.00	0.00	5.88	0.00	14.29	0.00	6.25	0.00	21.43	10.00	5.88	0.00
Section 3	0.00	0.00	0.00	0.00	27.27	4.17	5.88	0.00	0.00	7.14	5.88	0.00	18.18	8.33	9.09	0.00	0.00
Section 4	0.00	0.00	0.00	21.05	0.00	4.17	0.00	0.00	0.00	0.00	0.00	0.00	8.33	15.38	0.00	18.75	9.09
Section 5	10.53	0.00	27.78	31.58	63.64	17.39	40.00	10.53	20.00	23.08	27.78	12.50	54.55	38.46	66.67	0.00	0.00
Section 6	5.26	0.00	0.00	0.00	0.00	0.00	0.00	6.25	0.00	0.00	6.25	0.00	0.00	0.00	66.67	73.33	20.00
Section 7	35.00	62.50	33.33	58.82	30.77	50.00	61.54	70.00	30.00	76.92	72.22	62.50	100.00	92.31	100.00	100.00	78.57
Section 8	44.44	11.76	44.44	70.59	69.23	65.22	75.00	40.00	54.55	61.54	35.29	62.50	72.73	66.67	90.91	70.00	37.50

Figure 20 Frequency of congestion (expressed in percentage).

experience large variation in travel times during these periods. Figure 19 shows unreliable sections at the sample hour such as section 6 and 7.

Based on the metrics calculated using probabilistic measures, it can be observed that these measures ignore the wide variation between the light traffic and heavy traffic condition while indicating reliability. Hence, they report results different from that of other sets of measures. Also, these metrics have a binary construction (either within or exceeds threshold). As such, these can be insensitive to small changes in the reliability condition, and hence, multiple bounds should be used to detect the changes effectively. Probability plots also may help to visually observe reliability under multiple thresholds.

5. Congestion based measures

Figure 20 shows a congestion reliability measure, Frequency of Congestion (FOC), where the reliability status is indicated by the previously used color coding. The threshold used here is travel time corresponding to a speed of 15 kmph. Thus, the numbers in Figure 20 shows the percentage of times the travel time exceeded the threshold travel time in the corresponding hours.

From Figure 20, it can be observed that section 7 and section 8 shows higher congestion frequency for most of the time periods. Along with these sections, it can also be observed that section 5 and section 6 are also prone to congestion during peak hours which are upstream side of section 7. Key point to note here is so far though few measures were able to identify the unreliability of section 6 during peak hours, none of the measures were able to identify section 5. Though this measure implicitly indicates unreliability, it can be more related to performance assessment rather than variability or reliability.

Agreement Among Reliability Measures

The agreement or consistency among measures is evaluated by carrying out a correlation analysis. Correlation can identify the strength of relationship or association between a pair of reliability measure. The correlation coefficient r ranges from -1 to $+1$. Here, 1 indicates perfect correlation and 0 indicates absence of linear correlation. Positive values indicate that if one measure increases, other increases and vice versa. Negative values indicate that if one measure increases, other decreases and vice versa. High-positive values imply high strength of relation among a pair of reliability measure. This can be directly related to similarity among the two measures about the reliability condition. Figure 21 shows the correlation between different pairs of measures discussed in the

	Std Dev.	Percent Var.	Width	Skew	UI	Buffer Time	BTI	Planning Time	PTI	On-Time Arr.	Misery Index	Pr(α)	FOC
Std Dev.	1.00												
Percent Var.	0.81	1.00											
Width	0.77	0.96	1.00										
Skew	-0.15	0.01	0.10	1.00									
UI	0.22	0.49	0.56	0.58	1.00								
Buffer Time	0.98	0.84	0.80	-0.11	0.27	1.00							
BTI	0.71	0.95	0.92	0.13	0.54	0.78	1.00						
Planning Time	0.95	0.70	0.66	-0.17	0.18	0.94	0.62	1.00					
PTI	0.96	0.74	0.68	-0.18	0.17	0.94	0.65	0.97	1.00				
On-Time Arr.	-0.39	-0.55	-0.55	0.02	-0.23	-0.35	-0.44	-0.31	-0.34	1.00			
Misery Index	0.72	0.97	0.95	0.15	0.55	0.77	0.97	0.62	0.64	-0.47	1.00		
Pr(α)	0.44	0.63	0.69	0.19	0.37	0.43	0.58	0.36	0.37	-0.71	0.43	1.00	
FOC	0.80	0.62	0.66	-0.24	0.15	0.80	0.64	0.87	0.84	-0.35	0.64	0.32	1.00
Here UI is Unreliability Index, BTI is Buffer Time Index, PTI is Planning Time Index and FOC is Frequency of Congestion													
Yellow indicates high positive correlation and Red indicates low positive correlation or negative correlation													

Figure 21 Correlation between different pair of reliability measures.

previous sections. In the figure, yellow indicates high positive correlation and red indicates low correlation. Correlation coefficient above 0.6 is considered to indicate high correlation or strong agreement among measures.

From **Figure 21**, it can be observed that there is a strong agreement regarding the reliability condition among standard deviation, percent variation, width, buffer time, buffer time index, planning time, planning time index, misery index and frequency of congestion. These measures do agree well among each other. Skew, unreliability index, on-time arrival and $Pr(\alpha)$ do not agree well among each other and also with the other measures. However, this only shows linear relationship and if the measures are nonlinearly related, this will not identify that.

Overall, from the above measures and findings, it can be observed that many of the measures give different reliability indications. This may be due to the fact that some measures (1) use non-robust characteristics of the travel time distribution (2) do not explicitly take into account skewness of the distribution as an indicator of reliability. Future analysis can include comparison of private versus public transport reliability, link versus path reliability, and effect of intersection delay on reliability.

Acknowledgment

The authors acknowledge the support for this study as a part of the SPARC project funded by the Ministry of Human Resource Development, Government of India, through Project Number CIE1819289SPARLELI.

References

- Asakura, Y., 1996. Reliability measures of an origin and destination pair in a deteriorated road network with variable flows. In: Proceedings of the Fourth Meeting of the EURO Working Group in Transportation.
- Chang, X., Stopher, P., 1981. Defining the perceived attributes of travel modes for urban work trips. *Transp. Plan. Tech.* 7, 55–65.
- Fan W., Gong L., 2017. Developing a systematic approach to improving bottleneck analysis in North Carolina. North Carolina DOT. Available from: <https://connect.ncdot.gov/projects/research/RNAPProjDocs/201610%20Final%20Report.pdf>.
- Federal Highway Administration. 2017. Travel Time Reliability: Making It There On Time, All The Time. Available from: https://ops.fhwa.dot.gov/publications/tt_reliability/TTR_Report.htm. Accessed 10th March, 2020.
- Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L., 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* 49 (4), 764–766.
- Lomax, T., Schrank, D., Turner, S., and Margiotta, R., 2003. Selecting Travel Reliability Measures. Available from: <https://static.tti.tamu.edu/tti.tamu.edu/documents/TTI-2003-3.pdf>.
- Lyman, K., Bertini, R.L., 2008. Using travel time reliability measures to improve regional transportation planning and operations. *Transp. Res. Rec.* 2046 (1), 1–10.
- Prashker, J., 1979. Direct analysis of the perceived importance of attributes of reliability of travel modes in urban travel. *Transportation* 8, 329–346.
- Travel Time Reliability Implementation for the Freeway SIS Final Report, Florida Department of Transportation Systems Planning Office, Contract BDK77-931-04, December 31, 2010.
- The Keys to Estimating Mobility in Urban Areas: Applying Definitions and Measures That Everyone Understands. Sponsored by Departments of Transportation in California, Colorado, Kentucky, Minnesota, New York, Oregon, Pennsylvania, Texas and Washington, 1998. Available from: <http://mobility.tamu.edu>.
- Van Lint, J.W.C., Van Zuylen, H.J., 2005. Monitoring and predicting freeway travel time reliability: Using width and skew of day-to-day travel time distribution. *Transp. Res. Rec.* 1917 (1), 54–62.
- van Lint, J.W., Henk, C., Van Zuylen, J., Tu, H., 2008. Travel time unreliability on freeways: why measures based on variance tell only half the story. *Transp. Res. Part A: Policy Prac.* 42 (1), 258–277.
- Vandervalk, A., Louch, H., Guerre, J., Margiotta, R., 2014. Incorporating Reliability Performance Measures Into the Transportation Planning and Programming Processes: Technical Reference (No. SHRP 2 Report S2-L05-RR-3).
- Vaziri, M., Lam, T.N., 1983. Perceived factors affecting driver route decisions. *J. Transp. Eng.* 109, 297–311.
- Verker, Ministerie van, and Waterstaat. 2004. Mobility Note: Towards Reliable and Predictable Accessibility.

Further Reading

National Academies of Sciences, Engineering, and Medicine 2014. Incorporating Reliability Performance Measures into the Transportation Planning and Programming Processes: Technical Reference. Washington, DC: The National Academies Press. Available from: <https://doi.org/10.17226/22594>.

Texas Transportation Institute and Cambridge Systems Inc. 2006. Travel Time Reliability: Making It There on Time, All the Time. Available from: http://www.ops.fhwa.dot.gov/publications/tt_reliability/index.htm.

Traffic Incident Management

Ruimin Li, Department of Civil Engineering, Tsinghua University, Beijing, China

© 2021 Elsevier Ltd. All rights reserved.

Introduction	122
TIM Overview	122
TIM Development Key Points	123
Communication and Information Exchange	123
Traffic Incident Management System (TIMS)	124
Quick Clearance Legislation	124
Transportation Management Center (TMC)	125
Incident Command System (ICS)	125
Traveler Information	126
Others	126
Conclusion	127
References	127
Further Reading	127

Introduction

A traffic incident is “any non-recurrent event, such as a vehicle crash, vehicle breakdown, or other special event, that causes a reduction in highway capacity and/or an increase in demand” (I-95 Corridor Coalition, 2007). Traffic incidents are regarded as a significant cause of non-recurrent congestion delays that motorists encounter daily on roadways. Road incidents account for 10%–25% of congestion in Europe (General, 2011). Traffic incidents normally consist of two intervals. The first one is from the time of occurrence to the time when the incident is cleared, and the second one is from the end of the primary interval to the time when the facility has resumed normal operations. Adler et al. (2013) demonstrated that a one-minute duration reduction generates a €57 gain per incident and even considerably higher gains at locations with high levels of recurrent congestion (i.e., approximately €1,200/incident/minute at highly congested locations). Although nonrecurrent congestion is difficult to predict because of its stochastic nature, addressing it in a timely and effective manner is important in reducing its influence on traffic conditions.

Traffic incident management (TIM) is “a planned and coordinated program to detect and remove incidents and restore traffic capacity as safely and as quickly as possible” (Carson, 2010). The goals of TIM are (but not limited to) as follows: reduce the time for incident detection and verification, response time, and clearance time; promote traffic safety; and provide timely and accurate information to the public (Amer et al., 2015). In the past few decades, various strategies and tools have been developed and implemented worldwide to improve overall TIM efficiency. An effective TIM strategy can reduce not only safety-related costs, such as reducing secondary incidents, but also nonsafety-related ones, such as reducing the duration of congestion caused by traffic incidents.

TIM Overview

The major TIM stages include detection, verification, response, site management, traffic management, motorist information, clearance, and recovery (Neudorff et al., 2003).

The following text provides brief descriptions of each TIM stage (Carson 2010; Amer et al., 2015).

- **Detection** is “the determination that an incident of some type has occurred” (Carson, 2010).
- **Verification** is “the determination of the precise location and nature of the incident” (Carson, 2010).
- **Response** is “the activation of a ‘planned’ strategy for the safe and rapid deployment of the most appropriate personnel and resources to the incident scene” (Carson, 2010).
- **Site management** is “the coordination and management of resources and activities at or near the incident scene, including personnel, equipment, and communication links” (Carson, 2010).
- **Traffic management** is “application of traffic control measures onsite and in areas affected by an incident” (Amer et al., 2015).
- **Motorist information** is “the communication of incident-related information to motorists who are at the scene of the incident, approaching the scene of the incident, or not yet departed from work, home, or other location” (Carson, 2010).
- **Clearance** is “the safe and timely removal of any wreckage, debris, or spilled material from the roadway” (Carson, 2010).
- **Recovery** is “the restoration of the roadway to its full capacity” (Carson, 2010).

TIM's stages have several other division methods. For example, the different phases in incident management in Europe are divided into various categories, such as discovery, verification, initial response, scene management, recovery, restoration to normality, and normality (General, 2011).

Performance monitoring is an important issue to improve TIM efficiency. In 2004, the National Traffic Incident Management Coalition (NTIMC) was launched in the United States to coordinate TIM-related experiences, knowledge, practices, and ideas. Then, national unified goals (NUGs) for TIM were developed; these goals encompass responder safety; safe, quick clearance; and prompt, reliable, interoperable communication (NTIMC, 2009). Many performance indicators can be used to measure the efficiency of TIM and different TIM stages. The following are several of the common performance indicators used to measure the TIM program (Carson, 2010; Owens et al., 2010).

- Roadway clearance time: "The time between the first recordable awareness of an incident (detection, notification, or verification) by a responding agency and first confirmation that all lanes are available for traffic flow."
- Incident clearance time: "The time between the first recordable awareness of the incident and the time at which the last responder has left the scene."
- Secondary incidents: "The number of unplanned incidents beginning with the time of detection of the primary incident where a collision occurs either a) within the incident scene or b) within the queue, including the opposite direction, resulting from the original incident."

In Europe, national road administrations normally acquire data related to the following performance measures: vehicle flow/congestion, journey times and delay, response times, and safety. The Federal Highway Administration has launched a TIM performance measurement knowledge management system¹ to provide TIM professionals with the knowledge and tools needed in successfully measuring TIM performance.

TIM Development Key Points

A good TIM procedure is critical for successful TIM. The development of the TIM procedure involves several key points.

Communication and Information Exchange

Effective information exchange is an important basis for TIM success and mainly answers the question, "how do different agencies involved in TIM collect and use data related to TIM?" Identifying and agreeing on a particular standard or group of standards for data exchange are necessary to ensure timely and reliable information exchange among all stakeholders in TIM. Several existing related standards, such as the National Transportation Communications for ITS Protocol, can also be used in TIM.

The I-95 Corridor Coalition developed and deployed an information exchange network in the 1990s to promote communication and information exchange among member agencies and users of transportation facilities (I-95 Corridor Coalition, 2008). In 2008, the Southern Traffic Incident eXchange program was deployed by the I-95 Corridor Coalition to facilitate incident notification/information sharing among the southern states of North Carolina, South Carolina, Georgia, and Florida.

In Europe, DATEX II is a standard for digital information exchange between road traffic management centers; in addition, the Transport Protocol Experts Group is a data protocol group for traffic- and travel-related information, and it can be used for reference in TIM (General, 2011).

Several successful information exchange projects for TIM exist worldwide; however, achieving real-time, multi-agency communication and information exchange in TIM encounters the following challenges (Owens et al., 2010).

1. Proprietary systems in different agencies are difficult to integrate due to incompatible standards, systems, networks, and frequencies;
2. Lack of standards for different terminologies or data dictionaries and presence of multiple and inconsistent entries for the same data fields;
3. Different attitudes and concerns about sharing sensitive data among different agencies; and
4. Cost-sharing between agencies of shared systems.

The following tools and strategies, which have been proven effective in overcoming these challenges, have been deployed in different locations (Carson, 2010; Owens et al., 2010).

1. Establishing agreements among different agencies to protect sensitive data
2. Establishing technical committees to develop common data dictionaries or translators
3. Establishing agreements regarding a defined standard or group of standards for information exchange
4. Establishing agreements regarding methods of integrating multiple formats of data for information exchange, including text, video, and audio
5. Promoting consistent data collection practices among various agencies

¹ https://ops.fhwa.dot.gov/eto_tim_pse/preparedness/tim/knowledgebase/index.htm

Traffic Incident Management System (TIMS)

With the development of various technologies worldwide, many traffic control centers in cities and highways have deployed TIMS. TIMS is an effective tool to deal with traffic incidents and alleviate the influence of traffic incidents on traffic conditions (Schrank and Lomax 2009, Owens, Armstrong et al. 2010) by reducing the time to detect incidents, the arrival time of responding vehicles, and the time necessary for traffic to return to normal conditions. In general, TIMS has the following functions.

Automatic incident detection: automatically provides incident detection by using computers, traffic detectors, and surveillance equipment in all weather and light conditions at all times. Accordingly, the time needed to detect incidents is reduced.

Provision of traffic information to stakeholders: provides different stakeholders with necessary information related to incidents through various technologies automatically and in real time (e.g., a computer-aided dispatch system).

Dissemination of traffic and transport information to the public: provides road travelers with information on incidents and guidance information on paths to follow through the incident area via various media.

Coordination of existing and future traffic control and surveillance systems: a TIMS will coordinate several traffic management systems, such as data collection systems, Mayday and Automatic Collision Notification Systems, real-time traffic information service systems (e.g., changeable message signs, CMS), and response unit dispatch systems, to support efficient TIM.

A typical TIMS is the Coordinated Highways Action Response Team (CHART), which started in the mid-1980s as the “Reach the Beach” initiative and now is a full-scale intelligent transportation system managed through a statewide operations center (SOC) functioning 24/7 and satellite traffic operation centers (TOCs) around the state (CHART, 2013).

The subsystems of CHART mainly include the following:

Traffic monitoring subsystem: supports the SOC and TOCs’ monitoring and control activities and mainly includes data communication facilities, closed-circuit television (CCTV) system, and other traffic-detection systems.

Traveler information subsystem: provides travelers with various information related to incidents through various means, such as CMSs on roadsides, traveler advisory radio transmitters, and mobile phones.

Incident management: supported in part by CHART emergency patrols, who assist all allied law enforcement agencies and fire/rescue departments in TIM.

Traffic management: assisted by the freeway management system and arterial signal system to effectively balance the capacity and demand during a traffic incident.

Many benefit analyses have shown that CHART is effective in achieving its mission.

Quick Clearance Legislation

Safe, quick clearance is one of the three NUGs and a strong priority for TIM. Quick clearance is a program used to remove temporary obstructions from the roadway for increasing the safety of incident responders, reducing the probability of secondary incidents, and relieving overall congestion levels and delay resulting from traffic incidents. Such a program is a key feature of all TIM responder actions, and a number of countries worldwide have implemented various quick clearance laws to assist in TIM. Generally, three types of legislation provide TIM responder authority for removing vehicles and cargo from incident scenes as quickly as possible (Carson, 2008; Owens et al., 2010).

1) Driver Removal or “Move It”

A number of countries have enacted laws that require motorists involved in minor crashes to move their vehicles out of the travel lanes and off the road to avoid blocking traffic when the vehicle is drivable and when:

- Only property is damaged in a crash.
- The accident has not resulted in any serious injuries or fatalities.
- One or more lanes are blocked by the vehicles involved in the crash.
- The involved vehicles can be moved without risk of further damage to property or injury to any person.

Driver removal or “move it” laws are particularly applicable to limited-access expressways.

2) “Move Over”

To protect incident responders, move over laws provide specific requirements for motorists approaching an incident scene. Vehicles are required to slow down and move over to an adjacent lane, when possible, to provide an increased safety buffer. In 2007, “Move Over, America” was implemented in America to promote the national awareness and enactment of move over laws.

3) Authority Removal

This type of laws provides statutory authority to designated public agencies to remove vehicles or other roadway obstructions out of travel lanes for restoring traffic flow. The removal authority’s scope, as an important strategy for reducing incident-related congestion and delay, has expanded to generally include obstructions not only on the travel lanes but also on the shoulder or the roadway right-of-way in recognition of potential safety hazards (Carson, 2008). Similar to driver removal laws, authority removal laws are mainly limited to incidents occurring on limited-access, high-speed roadways.

Recently, the quick clearance program has been implemented broadly and provides many potential benefits, such as reduction of nonrecurrent congestion delay, reduction of secondary incidents, responder protection, and minimization of improper or delayed responder actions.

Transportation Management Center (TMC)

Successful TIM relies on effective data collection, a suitable traffic control strategy, and information service. TMC is a useful tool to support these functions. Generally, TMC is a key technical and institutional hub to bring together various jurisdictions, modal interests, and service providers for focusing on the common goal of optimizing the entire surface transportation system. TMC is an important TIM resource and can be used as a hub of transportation management and control that brings together human and technological components from various agencies to perform various TIM functions. In the TIM program, a TMC or TOC's functions include the following (Krechmer et al., 2012; Tantillo et al., 2014):

- serve as a hub for the collection of incident-related information and play a critical role in incident detection and verification, such as using a CCTV system;
- notify and/or coordinate with multi-disciplinary responders;
- assess damage and the capacity of the transportation system around an incident scene;
- implement various traffic operation strategies to mitigate or manage traffic flow in the incident area;
- implement a traffic diversion strategy (if appropriate);
- disseminate real-time, accurate, and predicted traveler information related to incident status and traffic conditions to the public and media, such as sending messages to the CMS;
- continuously monitor traffic conditions at and around an incident scene and adjust TMC response and support actions when necessary;
- coordinate with other TMC/TOCs and response agencies regarding incident status and conditions;
- support the need for additional responses, such as those required for spill cleanup or serious injuries;
- support on-scene responses with motorist assistance patrols, traffic response teams, or emergency traffic control by motorist patrols;
- log various information related to incidents, such as detection, response, and clearance times;
- Stand down or revise traffic management strategies, support personnel, and traveler information systems as the incident clears;
- participate in postincident reviews and joint training at a later date;
- plan for anticipated events, such as winter storms and hurricanes.

TMCs have operation professionals who perform TIM functions and are usually operational at all times or at least during hours that cover the most congested periods. Effective use of TMCs' resources is critical to minimize detection, response, and clearance time.

Incident Command System (ICS)

Various responders must execute their respective roles and responsibilities in incident clearance at the incident scene to achieve successful, safe, and quick clearance of the traffic incident scene. ICS, as defined in the National Incident Management System (NIMS), provides an effective framework.

ICS was originally developed in the 1970s to manage rapidly moving wildfires. The system has evolved into a well-established, flexible, standardized on-scene management protocol specific for managing incident response activities in TIM as a key component of NIMS. "ICS is a standardized approach to the command, control, and coordination of on-scene incident management that provides a common hierarchy within which personnel from multiple organizations can be effective" (FEMA, 2017). ICS, as an important support for scene management in TIM, allows multiple response agencies to work together at an incident scene and use common terminology and operating procedures for controlling facilities, personnel, equipment, procedures, and communication (Amer et al., 2015).

NIMS normally specifies an ICS organization consisting of the following six major functions (FEMA 2017).

- Command — provide on-scene management and control authority: A single incident command structure should be used when an incident occurs within a single jurisdiction and without jurisdictional or functional agency overlap. A unified command structure should be used to improve unity of effort in multijurisdictional or multiagency incident management.
- Operations — direct incident tactical operations, which can be organized and executed in many ways based on the nature and extent of the incident, the jurisdictions and agencies involved, and the incident's objectives, priorities, and strategies.
- Planning — prepare an incident action plan (IAP), facilitate the IAP process, maintain resource status, and display situation information.
- Logistics — "provide facilities, security (of the incident command facilities and personnel), transportation, supplies, equipment maintenance and fuel, food services, communications and IT support, and medical services for incident personnel" (FEMA 2017).
- Finance and administration — track and analyze incident costs, record personnel time, negotiate leases and maintain vendor contracts, and account for reimbursements.
- Intelligence — gather, manage, and share incident-related information to ensure that intelligence and investigative operations and activities are properly managed and coordinated.

Traveler Information

Traffic incidents, which are nonrecurrent events, usually exceed the traveler's expectation and cause travel disruption. During traffic incidents, provision of information on traffic incidents helps reduce the traffic demand, decrease the potential for secondary incidents involving motorists approaching the incident scene, and improve responder safety at the incident scene; it also allows motorists to make informed decisions on the basis of current traffic conditions for minimizing the impact of traffic incidents (Robinson et al., 2018).

Information dissemination is an important part of TIM and generally includes pretrip and enroute travelers. With the development of information and communication technology, an increasing number of common approaches have been used to provide travelers information on traffic incidents. The general methods of pre-trip information may include websites, TV, radio, text or email alerts, mobile mapping applications, and specific apps released by transportation agencies. For en-route information dissemination, aside from the previously listed media, the other methods usually used are CMSs, highway advisory radio, and in-vehicle devices.

In-depth understanding of the kind of information that travelers need, the time that they need it, and the manner that they wish to receive it is necessary to provide full play to the role of traveler information services. To ensure traveler cooperation and information effectiveness, traveler information tools or strategies should ensure the following (Carson, 2010):

- Provide travelers the accurate nature and extent of incidents to ensure that they may make intelligent choices (e.g., delay or cancel their trips or select alternative routes or travel modes).
- Provide necessary information on possible actions, such as alternative routes.
- Accurately and clearly describe the required actions of drivers, such as reducing speed and lane changing or diverting.

The effectiveness of traveler information has been proven in many applications (Clark et al., 2017). However, inaccurate traveler information remains a huge challenge in TIM. Acquiring real-time and accurate information on traffic incidents and traffic conditions around the incident scene is sometimes difficult due to the lack of effective traffic detection and/or reporting system, miscommunication, or lack of communication. Consequently, accurate traffic incident-related information is difficult to provide.

Traffic Incident Duration Estimation (Li et al., 2018)

Traffic operators must understand the main factors that influence traffic incident duration and accurately predict such a duration to improve TIM efficiency. This research field has been investigated in terms of two subfields with different techniques: analysis of the factors that influence traffic incident duration and prediction of traffic incident duration time with or without influence factor analysis.

Analysis and prediction of traffic incident duration in TIMS and intelligent transportation systems are currently important topics that have produced different results in previous studies. Incident duration time is related to various factors, such as temporal characteristics (e.g., time of day, day of the week, and/or season), incident characteristics (e.g., number of vehicles involved in an incident, truck/taxi/pedestrian involvement, and number of deaths and/or injured persons), road characteristics (e.g., incident location and road condition), traffic characteristics (e.g., traffic volume), and weather conditions (e.g., rain, fog, and/or snow).

Various statistical methods have been traditionally applied to analyze and predict traffic incident duration time. Among these methods are the following: linear/nonparametric regression, Bayesian classifier, hazard-based duration model, discrete choice model, structure equation model, and probabilistic distribution analyses. A new research field that is based on data-driven empirical algorithms and supported by unprecedented data availability has recently emerged for traffic incident duration prediction with an increasing amount of published literature. Different data mining-machine learning approaches have been used to estimate and predict traffic incident duration time. These approaches include the following: decision and classification tree models, artificial neural networks, genetic algorithm, and support/relevance vector machine. Several researchers have recently begun to utilize a hybrid method of predicting traffic incident duration and applying the advantages of the aforementioned methods.

Others

In Europe, the following 10 points form the backbone of TIM practices (General, 2011).

- Speedy detection and response
- Good information on location, severity, and attendant hazards
- Protection of the scene and ensuring the safety of responders, victims, and the public
- Coordinated response with a clear structure of authority, roles, and responsibility
- Reliable communication between responders and the public
- Provision of appropriate equipment, facilities, access paths, and control centers
- Sufficient backup services to ensure speedy clearance and minimize congestion
- Training and debriefing systems
- Written guidelines and formal agreements when necessary
- Monitoring, performance assessment, and feedback into practice

In the United States, NTIMC has identified the following nine “keys to success” for TIM programs (CORPORATION, 2010).

- TIM program strategic plans
- TIM administrative teams (TIM team/task force)
- Performance measures
- Responder and motorist safety
- Response and clearance policies and procedures
- Procedures for major incidents
- Integrated interagency communication
- Transportation management systems
- Traveler information

Conclusion

TIM practices vary between countries. However, the success of TIM at any state-, city-, or local-level programs primarily depends on the active involvement of and coordination among different TIM responders. TIM practice must have the support of various advanced technologies to achieve success. For example, the application of photogrammetry can significantly reduce the time required to document an incident scene (Clark et al., 2017).

The project Pro-Active Incident Management in Europe recommends several promising techniques in TIM. In the monitor and anticipate phase, preincident management techniques based on novel technologies can be used, and these include video incident detection systems, vehicle-based information report (e.g., eCall), probe vehicle data (e.g., vehicle-to-infrastructure communication and floating car data), or the simulation-based incident response strategy planning tool. In the prepare and respond phase, incident management techniques can be used, and these include advanced eCall, professional on-site incident reporting, model-based short-term response planning, optimal clearing strategy, incident screens to avoid rubbernecking, redirection strategy (contraflow), and optimal lane closing strategy or dynamic speed limits and driver information.

A successful TIM program will lead to safe, efficient, and automated strategies for handling traffic incidents in the future if TIM-related responders and evolving innovative technologies have excellent coordination. The final TIM goal is “Successfully Managing Traffic Incidents Is No Accident” (Vásconez, 2013).

References

- Adler, M.W., Ommeren, J.V., Rietveld, P., 2013. Road congestion and incident duration. *Econ. Transp.* 2 (4), 109–118.
- Amer, A., Roberts, E., Mangar, U., Kraft, W.H., Wanat, J.T., Cusolito, P.C., et al., 2015. *Traffic Incident Management Gap Analysis Primer*. Washington, DC, Federal Highway Administration Office of Operations: 90.
- Carson, J.L., 2008. *Traffic Incident Management Quick Clearance Laws A National Review of Best Practice*. College Station, Texas, Texas Transportation Institute.
- Carson, J.L., 2010. *Best Practices in Traffic Incident Management*. College Station, Texas 77843-3135, Texas Transportation Institute.
- Coordinated Highways Action Response Team (CHART), 2013. CHART Subsystems [online]. Available from: <https://chart.maryland.gov/about/subsystems.asp>.
- Clark, J., Neuner, M., Sethi, S., Bauer, J., Bedsole, L., and Cheema, A., 2017. *Transportation Systems Management and Operations in Action*. Reston, VA, Leidos.
- CORPORATION, D., 2010. *Traffic Incident Management Teams Best Practice Report*.
- FEMA, 2017. *National Incident Management System Third Edition*. Washington, DC, U.S. Department of Homeland Security.
- General, C. s. S. 2011. *Best practice in European traffic incident management*, CEDR's Secretariat General.
- Krechmer, D., A.S. III, Beer, P., Boyd, N., and Boyce, B., 2012. *Role of Transportation Management Centers in Emergency Operations Guidebook*. Bethesda, MD, Cambridge Systematics, Inc.
- Li, R., Pereira, F.C., Ben-Akiva, M.E., 2018. Overview of traffic incident duration analysis and prediction. *Eur. Trans. Res. Rev.* 2018 (10), 22.
- Neudorff, L.G., Randall, J.E., Reiss, R. and Gordon, R., 2003. *Freeway Management and Operations Handbook* New York, NY 10121, Siemens ITS.
- Owens, N., Armstrong, A., Sullivan, P., Mitchell, C., Newton, D., Brewster, R., and Trego, T., 2010. *Traffic Incident Management Handbook*, Federal Highway Administration, U.S. Department of Transportation.
- Robinson, E., Lubar, E., Singer, J., Kellman, D., Katz, B., and Kuznicki, S., 2018. *State of the Practice for Traveler Information During Nonrecurring Events*. Reston, VA, Leidos, Inc.
- Schrank, D. and Lomax, T., 2009. *2009 Urban Mobility Report*, Texas Transportation Institute.
- Tantillo, M.J., Roberts, E., and Mangar, U., 2014. *Roles of Transportation Management Centers in Incident Management on Managed Lanes*. Vienna, VA, Vanasse Hangen Brustlin, Inc (VHB).
- Vásconez, K.C., 2013. Successfully managing traffic incidents is no accident. *Public Roads* 77 (1).

Further Reading

- National Traffic Incident Management Coalition (NTIMC), 2009. Available from <http://timnetwork.org/national-traffic-incident-management-coalition/>.
- I-95 Corridor Coalition. 2007. *I-95 Corridor Coalition Coordinated Incident Management Toolkit for Quick Clearance*
- I-95 Corridor Coalition. 2008. *Corridor-Wide Center-to-Center Communications Study TECHNICAL MEMORANDUM Final Report*.

Traffic Incident Detection

Shuyan Chen, Yingjiu Pan, School of Transportation, Southeast University, Jiangsu, China

© 2021 Elsevier Ltd. All rights reserved.

Traffic Incident definition	128
Automatic Incident Detection	128
Methods of AID	129
Pattern Recognition Algorithms	129
Statistical Algorithms	129
Time Series and Filtering Algorithms	129
Traffic Model and Theoretical Algorithms	130
Fuzzy Logic-Based Algorithms	130
Data Mining-Based Algorithms	130
Imbalance Data Problem	131
Deep Learning	131
Performance Criteria for AID	131
References	132
Further Reading	133

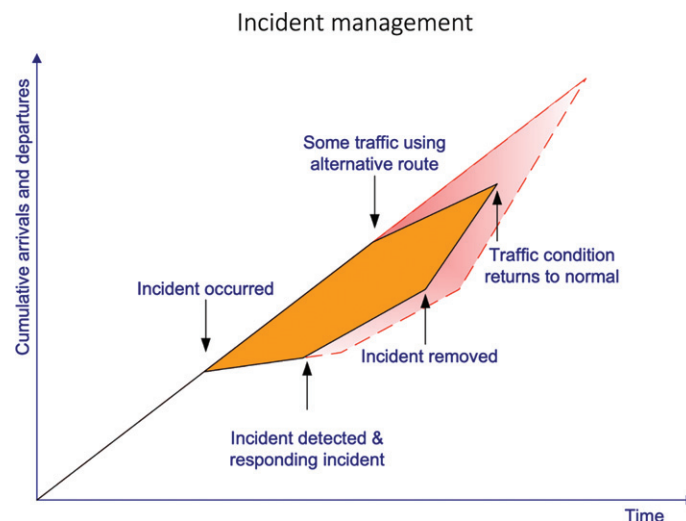
Traffic Incident definition

Traffic incidents are defined as nonrecurring events such as accidents, disabled vehicles, spilled loads, temporary maintenance and construction activities, signal and detector malfunctions, and other special or unusual events that disrupt the normal traffic flow and result in a capacity reduction of a facility (Srinivasan et al., 2003).

Incidents inevitably cause delays and reduce road capacity. According to one estimation, about 60% of the total vehicle-hours of delay on urban freeway was caused by traffic incidents (Lindley, 1987). In most urban areas, the situation is worsening with increasing traffic volume and limited expansion of the existing highway infrastructure. Moreover, incidents deteriorate road safety conditions. Congestions or traffic accidents caused by incidents have cost the world billions of dollars a year in the lost productivity, property damage, and personal injuries. According to the World Health Organization (WHO, 2018), road traffic injuries caused an estimated 1.35 million deaths worldwide in the year 2016, that is, one person is killed every 25 s. Thus, it is necessary to develop automated incident detection system to detect unexpected disturbance which are both reliable and quick to respond.

Automatic Incident Detection

Early detection of incidents is vital for formulating effective response strategy, and the benefits derived from effective and early incident detection that leads to prompt response can drastically reduce traffic delays, improve capacity and safety of road, and real-time traffic control. The previous studies have shown that 1 min early detected would reduce about 4–5 min delay (Srinivasan et al., 2004). The following diagram (Chung and Rosalion, 1999) illustrates the benefit of early detection



The function of Automatic Incident Detection (AID) is to automatically identify the occurrence of unpredictable incidents that affect the capacity of freeways so that appropriate response and clearance procedures can be executed to minimize the effects of the incident on traffic operation. AID algorithms work by detecting the change in traffic parameters measured at upstream and downstream detector stations. The working principle of the AID system consist of the following steps.

- Step 1. Receives real-time information from sensors;
- Step 2. Makes a decision whether an incident happening or not;
- Step 3. If so, give the alarms, and formulating effective response strategy;

A variety of sensor technologies are used to provide real-time traffic information for incident detection. These sensors commonly involve the use of inductive loop detectors (ILDs) and video sensors (Parkany and Xie, 2005). Video sensors become particularly important in traffic applications mainly due to their fast response, easy installation, operation and maintenance, and their ability to monitor wide areas (Kastrinaki et al., 2003). However, ILDs are the most commonly used sensors in traffic surveillance and management applications. In this chapter, most of AID algorithms introduced are based on traffic data collected from ILDs.

Methods of AID

AID Algorithms are a key component of a successful freeway traffic management system. Since the early 1970s, a variety of freeway incident algorithms have been developed based on traffic flow theory, pattern recognition, time series analysis, statistical techniques, and fuzzy logic, and Parkany and Xie (2005) reviewed the state of the art. After then, lots of new methods based on data mining and recently using deep learning were proposed to detect traffic incidents.

Pattern Recognition Algorithms

Pattern recognition-based algorithms rely on recognizing and differentiating unusual patterns of traffic from normal conditions. They compare the value of a traffic parameter, such as volume, occupancy, or speed, to preestablished thresholds. The algorithms trigger an alarm when parameters exceed the thresholds. Such kind of algorithms include California Algorithms, Berkeley (UCB) Algorithm, All-Purpose Incident Detection (APID) Algorithm, the Pattern Recognition Algorithm (PATREG), and Monica Algorithm developed in 1991.

California Algorithms (Payne and Tignor, 1978) were developed in the late 1960s for use in the Los Angeles freeway and surveillance control center, and ten modified versions of the California algorithm were identified and tested. The California #7 and California #8 provided the best overall results. Together with McMaster algorithms (Forbes, 1992), they are the most widely known and are often used as a standard for measuring the performance of other algorithms.

Statistical Algorithms

These algorithms use standard statistical techniques to determine whether observed detector data significantly differs from estimated or predicted values. It includes Standard Normal Deviate, and Bayesian Algorithm.

Standard Normal Deviate (SND) (Dudek et al., 1974) was developed by the Texas Transportation Institute (TTI) in the early 1970s. It is based on the premise that a sudden change in a measured traffic variable suggests that an incident has occurred on the freeway. Bayesian Algorithm uses Bayesian statistical techniques to compute the probability that an incident signal is caused by downstream lane blocking using historical data on the frequency of capacity-reducing events along a section of freeway.

Time Series and Filtering Algorithms

These algorithms smooth raw data over large time windows to eliminate short duration traffic disturbances such as random fluctuation, traffic pulses, and compression waves. Processed data are typically compared to predetermined thresholds. It includes Time Series ARIMA Algorithm, High Occupancy Algorithm (HIOCC), Exponential Smoothing Algorithm, Low-Pass Filtering, Dutch Algorithm and so on (Stephanedes and Chassiakos, 1993).

Time Series ARIMA Algorithm used an Auto-regressive Integrated Moving-Average (ARIMA) time series model to develop short-term forecasts and confidence intervals. Data that deviates from these predictions triggers an alarm. High Occupancy Algorithm inspects occupancy data for the presence of stationary or slow-moving vehicles, and examine several consecutive values of instantaneous occupancy to see if they exceed thresholds. Exponential Smoothing Algorithm utilized a single or double exponential smoothing function, which weights past observations to forecast future traffic conditions. It detects incidents using a tracking signal that is significantly different from that under nonincident conditions. Low-Pass Filtering algorithms removes sharp frequency data components, characteristic of noise in the data to produce a smoothed moving average. It uses direct comparison of differences in smoothed spatial occupancy over time to discriminate between recurrent congestion and an incident. Dutch Algorithm (Knibbe, 2004) detects the filtered average speed for the carriageway by means of an exponential smoothing filter.

Traffic Model and Theoretical Algorithms

These algorithms use complex theories of traffic flow to describe and predict traffic behavior during incident conditions. The model makes a decision whether an incident happened or not by comparing actual traffic parameters with the predicted output. It includes Dynamic Algorithm, McMaster Algorithm, and Low Volume Incident Detection Algorithms.

Dynamic Algorithm based on a dynamic model of traffic flow incorporates fundamental velocity-density and flow-density relationships into the detection analysis. The algorithm assumed that abrupt changes in a freeway system follow predictable patterns. McMaster Algorithm as well as modified version proposed a volume-occupancy template composed of four areas according to catastrophe theory (Persaud and Hall, 1989), each corresponding to a particular traffic profile. The algorithm performed two comparison tests between the template and actual loop data to decide whether traffic is congested and determined the source if congestion is detected. Low Volume Incident Detection Algorithms (Fambro and Ritch, 1979) developed by the Texas Transportation Institute used a speed input-output analysis of individual vehicles on freeway sections. The algorithm predicted the departure time of an entering vehicle and decided whether an incident happened based on the expected and actual output time.

Fuzzy Logic-Based Algorithms

Fuzzy Set Algorithms (Chang and Wang, 1994) incorporate inexact reasoning and uncertainty into the incident detection logic, which is designed to make decisions when data is imprecise or missing. This type of algorithm performed approximate reasoning by developing fuzzified boundaries. Whereas the "crisp" approach (e.g., California and McMaster) is basically a binary (0 or 1) decision process, the fuzzy approach shows the extent of the likelihood for an event. Some approaches use an occupancy trend factor and speed density comparison between neighboring detectors. However, membership functions are often set artificially, and this affects the performance of detection.

Data Mining-Based Algorithms

The rampant growth in traffic incidents has led to significant interest in the development of effective incident detection techniques in recent years, and various methods based on data mining have been proposed to address this problem, including partial least squares regression (PLSR), Neural Network, support vector machine (SVM), decision tree learning, rough set, ensemble learning, and so on.

PLSR (Wang et al., 2008) has been used as a classifier for AID. This approach builds a PLSR model to maps the relationship between the parameters depicting traffic flow and traffic state, that is the incident or nonincident state. The output of PLSR model tells whether an incident happened or not. In addition, a hybrid model which combines PLS and Neural Network has been developed to detect automatically traffic incident (Lu et al., 2012b).

Neural network approaches for incident detection have been thoroughly investigated and their potential superiority to other techniques has been demonstrated. Since 1990s, many different types of neural networks have been applied to solve incident detection problem (Ritchie and Cheu, 1993), including Bayesian-based Probabilistic Neural Network (PNN) (Abdulhai and Ritchie, 1995, 1999), probabilistic neural network (Wen et al., 2001), fuzzy-based neural network (Kim and Lee, 2005), fuzzy-wavelet radial basis function neural network (RBFNN) (Karim and Adeli, 2002), as well as combination of wavelet-based signal processing, statistical cluster analysis, and neural network pattern recognition (Ghosh-Dastidar and Adeli, 2003). The findings of previous studies indicate that neural network models have the potential to achieve better performance significantly in terms of detection rate (DR) and false alarm rate (FAR), as well as operational improvements in real-time incident detection over more conventional algorithms such as a series of California algorithms and McMaster algorithm.

However, all AID algorithms based on neural networks suffer from a number of shortcomings. First, it is a black-box model, thus hard to understand; second, it lacks transferability; another shortcoming is that its convergence depends on the selection of parameters, learning and momentum ratios, that have to be selected by trial and error. In addition, it is slow convergence and possibly being stuck in a local minima situation.

Incident detection is essentially a pattern classification problem, where the incident and incident-free traffic patterns are to be recognized. Any good classifier is a potential tool for the incident detection problem. Thus, decision tree learning, SVM, ensemble learning, rough set, as well as inductive logic programming, have been considered as alternative methods to develop some new AID algorithms.

Decision tree and grafted decision tree have been used to detect traffic incidents (Chen and Wang, 2009; Lu et al., 2014). As opposed to neural networks, decision trees have a top-down branched structure. Decision trees are not only efficient for classification, but also have an added advantage. Compared to other data mining algorithms, decision trees can easily interpret the rules used to filter datasets to their appropriate categories while simultaneously bringing to light the relative importance of different variables in the system being studied.

The rough set can cope with decision table and generate decision rules used for classification, therefore, the use of a rough set classifier has been investigated for freeway incident detection (Chen et al., 2007).

Inductive logic programming (ILP) is one of most famous inductive learning techniques, and the previous studies indicate that ILP is a competitive and promising method for traffic problems (Dzeroski et al., 1998; Roberts et al., 1998). Further study on ILP application to incident detection domain has been undertaken. Nfoil, one ILP system that combines naïve Bayes and foil, was applied to detect incidents, and their detection performance was compared with other existed AID algorithms (Lu et al., 2012a).

SVM classifiers have been investigated in AID problems with good results (Cheu et al., 2003). Like neural networks, SVM also has some drawbacks limiting its applications, since the performance of SVM depends on the choice of kernel function and several parameters, and setting the parameters of the SVM algorithm is a challenging task. Usually, an appropriate kernel function and parameters have to be chosen and tuned by trial and error. The ensemble learning is often more accurate than its individual components. Furthermore, it can avoid the burden of tuning the parameters for SVM or NN. Therefore, SVM ensembles, Neural network ensemble, random forest (Chen et al., 2007, 2009; Liu et al., 2013), and Nave Bayes Classifiers Ensemble (Liu et al., 2014) have been used to alleviate incident detection problem.

Imbalance Data Problem

In the real world, there are typically very few incident cases as compared to the large number of common traffic modes. Learning algorithms which do not consider class-imbalance, such as ANN, SVM, decision tree algorithms, tend to be overwhelmed by the major class and ignore the minor one. A heavily skewed data set often generates a classifier that predicts the minority class with a much higher rate of error than the majority class. This bias often makes the classifier highly accurate and completely useless at the same time (Chawla et al., 2004).

Traffic incident detection can be treated as a task of learning classifiers from imbalanced or skewed datasets, and the performance of traditional AID detector was restricted due to the imbalanced training set. Principal component analysis (PCA) as a one-class classifier for incident detection, which is constructed from the major and minor principal components of normal instances and a Support Vector Domain Description (SVDD) algorithm were brought into detect traffic incidents (Zheng et al., 2013). Another way is resampling by under sample, oversample or SMOTE to get a balanced training set, then learning some model from this training set, such as SVM, as a classifier to detect traffic incidents (Li and Chen, 2014, 2015, 2016).

Deep Learning

In recent years, deep learning technology has been applied to detect traffic incidents due to its excellent feature extraction and prediction capabilities. Zhang et al. (2018) compared and investigated deep belief network (DBN) and long short-term memory (LSTM) in detecting the traffic accident from social media data, revealing that DBN outperforms those of SVMs and supervised Latent Dirichlet allocation model. Singh and Mohan (2019) proposed a Stacked Autoencoder-based deep learning approach to detect road accidents using traffic surveillance video, which can automatically learn feature representation from the spatiotemporal volumes of raw pixel intensity. El Hatri and Boumhidi (2018) proposed a novel fuzzy deep learning-based traffic incident detection approach that considers the spatial and temporal correlations of traffic flow inherently, and the results revealed many advantages of the proposed model such as higher DR and lower FAR. Zhu et al. (2018) applied a convolutional neural network approach to detect incident that takes account of spatiotemporal network and traffic inherent structure, and validated that the detection accuracy of the CNN is generally superior to that of conventional alternatives.

Performance Criteria for AID

The performance of incident detection algorithms is often evaluated by the following statistical metrics, DR, FAR, mean time to detection (MTTD) (Cheu et al., 2003; Jin et al., 2001; Yuan and Cheu, 2003).

DR is defined as the ratio of the number of detected incidents to the actual number of incidents in the dataset, given as a percentage. If a single incident instance is identified within the period of actual occurrence, the incident case is regarded as detected.

$$DR = \frac{\text{number of incident cases detected}}{\text{total number of incident cases}} \times 100 \quad (1)$$

A pattern incorrectly classified as incident outside the incident occurrence period is considered as a false alarm. FAR is defined as the fraction of false alarm cases to the total number of nonincident instances.

$$FAR = \frac{\text{number of false alarm cases}}{\text{total number of nonincident instances}} \times 100 \quad (2)$$

Time to detection is the difference between the time the incident was detected and the actual time the incident occurred. MTTD is the average time to detect an incident over m incidents successfully detected:

$$MTTD = \frac{t_1 + \dots + t_m}{m} \quad (3)$$

Here, t_i is the length of time to detect the i th incident and m is the number of incident cases detected.

High DR, low FAR and MTTD are major requirements for AID models. In general, a more sensitive detection model has a higher DR, short MTTD but also higher FAR. There exist trade-offs between the DR, FAR, and MTTD. With these some conflicting performance measures, it is difficult to find the optimal incident detection model. For the purpose of comparing the detection

performance yielded by different AID models, a Performance Index (PI) that combines DR, FAR, MTTD, as well as CR has been suggested (Cheu et al., 2003):

$$PI = (1 - DR/100)^m \times FAR^n \times MTTD^p \times (1 - CR/100)^q \quad (4)$$

where CR is the classification rate, defined as the proportion of instances that were correctly classified based on total instances in dataset, written as

$$CR = \frac{\text{number of instances correctly classified}}{\text{total number of instances}} \times 100 \quad (5)$$

The coefficients $m > 0$, $n > 0$, $p > 0$ and $q \geq 0$, are used to emphasize the importance of each measure, respectively. Typical values for the coefficients are 1. Larger values denote greater importance of the particular parameter. A lower PI value indicates better performance.

Another PI modified is defined as follows (Chen et al., 2009):

$$PI = w_{DR} \cdot (1 - DR/100) + w_{FAR} \cdot FAR + w_{MTTD} \cdot MTTD/THD_{MTTD} + w_{CR} (1 - CR/100) \quad (6)$$

where w_{DR} , w_{FAR} , w_{MTTD} and w_{CR} are the normalized weights of corresponding measures, THD_{MTTD} can be interpreted as the threshold of MTTD, which can be derived from the statistic of MTTD in the past. This manipulation is to scale this measure into $[0, 1]$, the same scale to other measures.

Usually the weights of corresponding measures are taken equal to each other, except w_{CR} having a lower value, for DR, FAR and MTTD describe detection ability while CR describes classification ability, therefore, the formers are more meaningful and important than CR for evaluating the performance of AID algorithms, especially when the testing data set is highly screwed with rare incident instances.

The next criterion used is the area under the receiver operating characteristic curve, called AUC, which has been employed as a measurement of AID algorithms (Wang et al., 2008). The AUC is a numeric performance metric, which represents how separable two objects are. An advantage of this criterion is that it is not necessary to make any assumption on the prior probability of class distribution or specify the misclassification costs.

Relevant Websites

https://ops.fhwa.dot.gov/eto_tim_pse/docs/incident_mgmt_perf/section2.htm

Emergency Transportation Operations

<https://civitas.eu/measure/automatic-traffic-incident-detection>

Automatic traffic incident detection Homepage

References

- Abdulhai, B., Ritchie, S.G., 1995. Performance of artificial neural networks for incident detection in ITS. American Society for Civil Engineers, New York.
- Abdulhai, B., Ritchie, S.G., 1999. Enhancing the universality and transferability of freeway incident detection using a Bayesian-based neural network. *Transp. Res. Part C Emerg. Technol.* 7, 261–280, doi:10.1016/S0968-090X(99)00022-4.
- Chang, E.C.-P., Wang, S.-H., 1994. Improved freeway incident detection using fuzzy set theory. *Transp. Res. Rec. J. Transp. Res. Board* 1453, 75–82.
- Chawla, N.V., Japkowicz, N., Kotcz, A., 2004. Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newsl.* 6, 1, doi:10.1145/1007730.1007733.
- Chen, S., Wang, W., 2009. Decision tree learning for freeway automatic incident detection. *Expert Syst. Appl.* 36, 4101–4105, doi:10.1016/j.eswa.2008.03.012.
- Chen, S., Wang, W., Qu, G., Lu, J., 2007. Application of neural network ensembles to incident detection IEEE ICIT 2007–2007 IEEE. *Int. Conf. Integr. Technol.* 388–393, 10.1109/ICITECHNOLOGY.2007.4290502.
- Chen, S., Wang, W., van Zuylen, H., 2009. Construct support vector machine ensemble to detect traffic incident. *Expert Syst. Appl.* 36, 10976–10986, doi:10.1016/j.eswa.2009.02.039.
- Chen, S.Y., Wang, W., Qu, G.F., 2007. Traffic incident detection based on rough sets approach. *Proc. Sixth Int. Conf. Mach. Learn. Cybern. ICMLC 2007* (7), 3734–3739, doi:10.1109/ICMLC.2007.4370797.
- Cheu, R.L., Srinivasan, D., Teh, E.T., 2003. Support vector machine models for freeway incident detection. *IEEE Conf. Intell. Transp. Syst. Proceedings, ITSC 1*, 238–243, doi:10.1109/ITSC.2003.1251955.
- Chung, E., Rosalion, N., 1999. Review of freeway incident detection system. ARRB Research Report ARR 327. ARRB Transport Research, Vermont South, Victoria, Australia.
- Dudek, C.L., Messer, G.J., Nuckles, N.B., 1974. Incident detection on urban freeways. *Transport. Res. Rec.* 495, 12–24.
- Džeroski, S., Jacobs, N., Molina, M., Moure, C., Muggleton, S., Van Laer, W., 1998. Detecting traffic problems with ILP. *Proc. 8th Int. Conf. Inductive Log. Program. Lect. Notes Artif. Intell.* 1446, 281–290.
- El Hatri, C., Boumhidi, J., 2018. Fuzzy deep learning based urban traffic incident detection. *Cogn. Syst. Res.*, doi:10.1016/j.cogsys.2017.12.002.
- Fambro, D.B., Ritch, G.P., Automatic detection of freeway incident during low volume conditions. Report No. FHWA/TX-79/23+210-1. Texas Transportation Institute, Texas A&M University, College Station, TX, 1979.
- Forbes, G.J., 1992. Identifying incident congestion. *ITE J.* 62, 43–50.
- Ghosh-Dastidar, S., Adeli, H., 2003. Wavelet-clustering-neural network model for freeway incident detection. *Comput. Civ. Infrastruct. Eng.* 18, 325–338, doi:10.1111/1467-8667.t01-1-00311.
- Lindley, J.A., 1987. Urban freeway congestion: quantification of the problem and effectiveness of potential solutions. *ITE J.* 57, 27–32.
- Jin, X., Srinivasan, D., Cheu, R.L., 2001. Classification of freeway traffic patterns for incident detection using constructive probabilistic neural networks. *IEEE Trans. Neural Netw.* 12, 1173–1187, doi:10.1109/72.950145.

- Karim, A., Adeli, H., 2002. Comparison of fuzzy-wavelet radial basis function neural network freeway incident detection model with California algorithm. *J. Transp. Eng.* 128, 21–30, doi:10.1061/(ASCE)0733-947X(2002)128:1(21).
- Kim, D., Lee, S., 2005. Incident detection using a fuzzy-based neural network model. *J. East. Asia Soc. Transp. Stud.* 6, 2629–2638.
- Knibbe, W.J.J., 2004. Incident management in The Netherlands. AET European Transport Conference 2004. Strasbourg, France, October, 2004, 1–6.
- Li, M.H., Chen, S.Y., 2014. Use subsampling to solve imbalanced dataset problem for Automatic Incident Detection Algorithm. *Appl. Mech. Mater.* 587–589, 2114–2119, doi:10.4028/www.scientific.net/AMM.587-589.2114.
- Li, M.H., Chen, S.Y., 2015. Resampling methods for solving class imbalance problem in traffic incident detection. *Appl. Mech. Mater.* 744–746, 1985–1989.
- Li, M.H., Chen, S.Y., Lao, Y.C., 2016. Automatic incident detection algorithm based on under-sampling for imbalanced traffic data. *Green Build. Environ. Energy Civ. Eng.* 145–150, doi:10.1201/9781315375106-31.
- Liu, Q., Lu, J., Chen, S., 2013. Traffic incident detection using random forest, in: *Proceedings of the 92st Transportation Research Board Annual Meeting*.
- Liu, Q., Lu, J., Chen, S., Zhao, K., 2014. Multiple Naïve Bayes classifiers ensemble for traffic incident detection. *Math. Probl. Eng.*, doi:10.1155/2014/383671.
- Lu, J., Chen, S., Wang, W., Ran, B., 2012a. Automatic traffic incident detection based on nFOIL. *Expert Syst. Appl.* 39, 6547–6556, doi:10.1016/j.eswa.2011.12.050.
- Lu, J., Chen, S., Wang, W., Van Zuylen, H., 2012b. A hybrid model of partial least squares and neural network for traffic incident detection. *Expert Syst. Appl.* 39, 4775–4784, doi:10.1016/j.eswa.2011.09.158.
- Lu, J., Liu, Q., Yuan, L., Chen, S., 2014. Grafted decision tree for freeway incident detection. *Am. Soc. Civ. Eng.* 3743–3751.
- Parkany, E., Xie, C., 2005. A complete review of incident detection algorithms and their deployment: what works and what doesn't. Transportation Center, University of Massachusetts, Technical Report, NETCR37.
- Payne, H.J., Tignor, S.C., 1978. Freeway incident-detection algorithms based on decision trees with states. *Transp. Res. Rec.* 30–37.
- Persaud, B.N., Hall, F.L., 1989. Catastrophe theory and patterns in 30-second freeway traffic data- Implications for incident detection. *Transp. Res. Part A Gen.* 23, 103–113, doi:10.1016/0191-2607(89)90071-X.
- Ritchie, S.G., Cheu, R.L., 1993. Simulation of freeway incident detection using artificial neural networks. *Top. Catal.* 1, 203–217, doi:10.1016/S0968-090X(13)80001-0.
- Roberts, S., Laerand, W., Van, Jacobs, N., Muggleton, S.H., Broughton, J., 1998. A comparison of {ILP} and propositional systems on propositional data. *Proc. 8th Int. Work. Inductive Log. Program.* 291–299.
- Singh, D., Mohan, C.K., 2019. Deep spatio-temporal representation for detection of road accidents using stacked autoencoder. *IEEE Trans. Intell. Transp. Syst.*, doi:10.1109/TITS.2018.2835308.
- Srinivasan, D., Jin, X., Cheu, R.L., 2004. Evaluation of adaptive neural network models for freeway incident detection. *IEEE Trans. Intell. Transp. Syst.* 5, 1–11, doi:10.1109/TITS.2004.825084.
- Srinivasan, D., Loo, W.H., Cheu, R.L., 2003. Traffic incident detection using particle swarm optimization 2003. *IEEE Swarm Intell. Symp. SIS 2003 - Proc.* 144–151, doi:10.1109/SIS.2003.1202260.
- Stephanedes, Y.J., Chassiakos, A.P., 1993. Freeway incident detection through filtering. *Transp. Res. Part C* 1, 219–233, doi:10.1016/0968-090X(93)90024-A.
- Wang, W., Chen, S., Qu, G., 2008. Incident detection algorithm based on partial least squares regression. *Transp. Res. Part C Emerg. Technol.* 16, 54–70, doi:10.1016/j.trc.2007.06.005.
- Wen, H., Yang, Z., Jiang, G., Shao, C., Rode, L., 2001. A New Algorithm of Incident Detection on Freeways, in: *Proceedings of the IEEE International Vehicle Electronics Conference*. pp. 197–202.
- Yuan, F., Cheu, R.L., 2003. Incident detection using support vector machines. *Transp. Res. Part C Emerg. Technol.* 11, 309–328, doi:10.1016/S0968-090X(03)00020-2.
- Zhang, Z., He, Q., Gao, J., Ni, M., 2018. A deep learning approach for detecting traffic accidents from social media data. *Transp. Res. Part C* 86, 580–596, doi:10.1016/j.trc.2017.11.027.
- Zheng, C., Chen, S., Wang, W., Lu, J., 2013. Using principal component analysis to solve a class imbalance problem in traffic incident detection. *Math. Probl. Eng.*, doi:10.1155/2013/524861.
- Zheng, W., Chen, S., Wu, T., 2013. An Imbalanced Datasets Based SVDD-AID Algorithm, in: *ICTIS 2013-Proceedings of the 2nd International Conference on Transportation Information and Safety*. pp. 1178–1184.
- Zhu, L., Guo, F., Krishnan, R., Polak, J.W., 2018. The Use of Convolutional Neural Networks for Traffic Incident Detection at a Network Level, in: *Transportation Research Board 97th Annual Meeting*.

Further Reading

- Abdulhai, B., Ritchie, S.G., Chair, 1999. Towards adaptive incident detection algorithms, in: *The 6th World Congress on Intelligent Transport Systems*.
- Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24, 123–140, doi:10.1007/BF00058655.
- De Raedt, L., Fransconi, P., Kersting, K., Muggleton, S., 2008. Probabilistic Inductive Logic Programming - Theory and Applications. <https://doi.org/10.1007/978-3-540-78652-8>.
- Flach, P.A., 2004. Tutorial on the many faces of ROC analysis in machine learning, in: *The Twenty-First International Conference on Machine Learning*.
- Grzymala-busse, J.W., Wang, C.P.B., 1996. Classification and rule induction based on rough sets, in: *Proceedings of the Fifth IEEE International Conference on Fuzzy Systems*. pp. 744–747.
- Han, J., Pei, J., Kamber, M., 2001. *Data mining: concept and techniques*, Higher Education Press, Beijing, pp. 284–294.
- Kastrinaki, V., Zervakis, M., Kalaitzakis, K., 2003. A survey of video processing techniques for traffic applications. *Image Vision Computing* 21, 359–381.
- Kersting, K., De Raedt, L., 2001. Bayesian Logic Programs, in: *CoRRcs.AI/0111058*. pp. 1–52.
- Landwehr, N., 2007. Integrating Naive Bayes and FOIL, in: *Proceedings of the Twentieth National Conference on Artificial Intelligence*. pp. 481–507.
- Lee, D., Jeng, S., Chandrasekar, P., 2004. Applying data mining techniques for traffic incident analysis. *Civ. Eng.* 44, 90–102.
- Macskassy, S., Provost, F., 2004. Confidence bands for ROC curves: methods and an empirical study. *ROC Anal. Artif. Intell.* 1st Int. Work. 61–70.
- Muggleton, S.H., De Raedt, L., 1994. Inductive logic programming: theory and methods. *J. Log. Program.* 19, 629–679, doi:10.1016/0743-1066(94)90035-3.
- Rätsch, G., Mika, S., Schölkopf, B., Müller, K.R., 2002. Constructing boosting algorithms from SVMs: an application to one-class classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 1184–1199, doi:10.1109/TPAMI.2002.1033211.
- Weiss, G.M., 2004. Mining with rarity: a unifying framework. *ACM SIGKDD Explor. Newsl.* 6, 7–19, doi:10.1145/1007730.1007734.
- WHO, ed. (2018). *Global Status Report on Road Safety 2018 (official report)*. World Health Organization (WHO), Geneva. pp. xiv–xv, 1–13.
- Witten, I.H., Frank, E., 1999. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers. Elsevier B.V., San Francisco, pp. 89–97, doi:10.1016/j.future.2017.02.015, 125–127, 159–161.
- Yang, L., Jia, L., 2004. A fuzzy modeling method using rough sets and its applications to intelligent transportation systems, in: *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*. pp. 5338–5341.

Bottleneck

Kentaro Wada*, Toru Seo*, Yasuhiro Shiomi*, *University of Tsukuba, Tsukuba, Ibaraki, Japan; †The University of Tokyo, Bunkyo-ku, Tokyo, Japan; ‡Ritsumeikan University, Kusatsu, Shiga, Japan

© 2021 Elsevier Ltd. All rights reserved.

Introduction	134
Definition of Bottleneck	134
Macroscopic Mechanism of Congestion due to a Bottleneck	134
Sag and Tunnel Bottlenecks	136
What is Sag?	136
Features of Traffic Dynamics at Sags and Tunnels	136
Modeling Approach	136
Freeway Network Bottlenecks	138
Merge Bottleneck	138
Diverge Bottleneck	139
Weaving Bottleneck	139
Management of Bottlenecks	140
Current Managements	140
Future Management	142
Acknowledgment	142
References	142

Introduction

Definition of Bottleneck

A bottleneck is defined as a road segment whose traffic capacity is lower than that of the road segments in its vicinity. The definition of traffic capacity of a road segment is the maximum traffic flow rate that can reasonably pass through the segment. Some evident examples of bottlenecks are lane drops, lanes blocked by car accidents, and signalized intersections with small green time ratios. Detailed discussions on types and mechanisms of bottlenecks are described in the further sections.

The Highway Capacity Manual (HCM) (Transportation Research Board, 2016) provides a detailed discussion on traffic capacity and the concept of a bottleneck from practical perspectives. According to the HCM, traffic capacity is defined as “the maximum number of vehicles that can pass a given point during a specified period under prevailing roadway, traffic, and control conditions. This assumes that there is no influence from downstream traffic operation,” and a bottleneck is defined as “locations where the capacity provided is insufficient to meet the demand over a given period of time” where “demand” refers to the flow of vehicles arriving at the bottleneck. The typical capacity of a basic highway segment is between 2250 and 2400 veh/h/lane, and that of a basic arterial segment with a signalized intersection is between 790 and 1110 veh/h/lane. These values vary among locations and conditions. Consequently, a location with a low capacity becomes a bottleneck.

Bottlenecks play an essential role in the description and understanding of traffic behavior, especially in uninterrupted flows (e.g., freeways, expressways). To be more precise, almost all of traffic congestions are caused by bottlenecks. Consider a bottleneck in an uninterrupted traffic. If an arrival flow rate (i.e., flow from the upstream section) to the bottleneck is larger than its capacity, then only a flow rate equal to the capacity passes through the bottleneck, and the remaining traffic forms a waiting queue at the bottleneck. A bottleneck is said to be active when it causes a waiting queue without queue extension from the downstream section.

See Fig. 1 for an intuitive illustration of a bottleneck. It shows an upturned bottle. The width of each cross-section signifies the capacity of the corresponding section. If an inflow is sufficiently large, the outflow will be limited by the neck. Further, the medium (i.e., the vehicles) that cannot pass through the neck will be pooled (i.e., form an waiting queue) upstream from the neck.

Macroscopic Mechanism of Congestion due to a Bottleneck

Typical congestion phenomena in uninterrupted traffic with a bottleneck can be explained as follows. Consider a freeway road section with a bottleneck. The inflow to the section can take two possible values: a flow less than the capacity (denoted as F1) and a flow greater than the capacity (denoted as F2). Congestion phenomena with time-varying inflow can be illustrated as in Fig. 2, based on the kinematic wave theory (Lighthill and Whitham, 1955; Richards, 1956; Newell, 1993), which is the most standard and simplest traffic flow model. Figure 2A shows the relationship between flow and density, where the solid curve indicates the fundamental diagram (Greenshields, 1935) of the section except for the bottleneck, the dashed line indicates the bottleneck capacity, and the black dots indicate the observed traffic states. Figure 2B shows a time-space diagram of traffic, where the horizontal

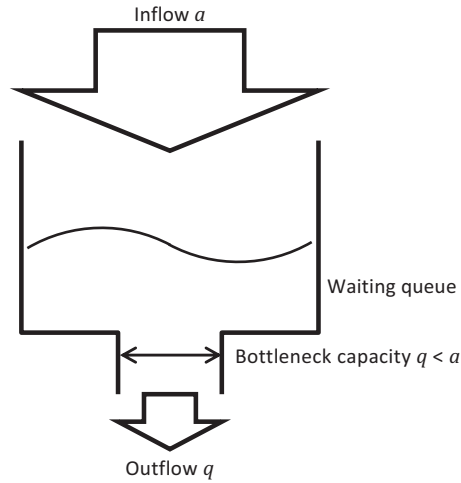


Fig. 1 Bottleneck and waiting queue.

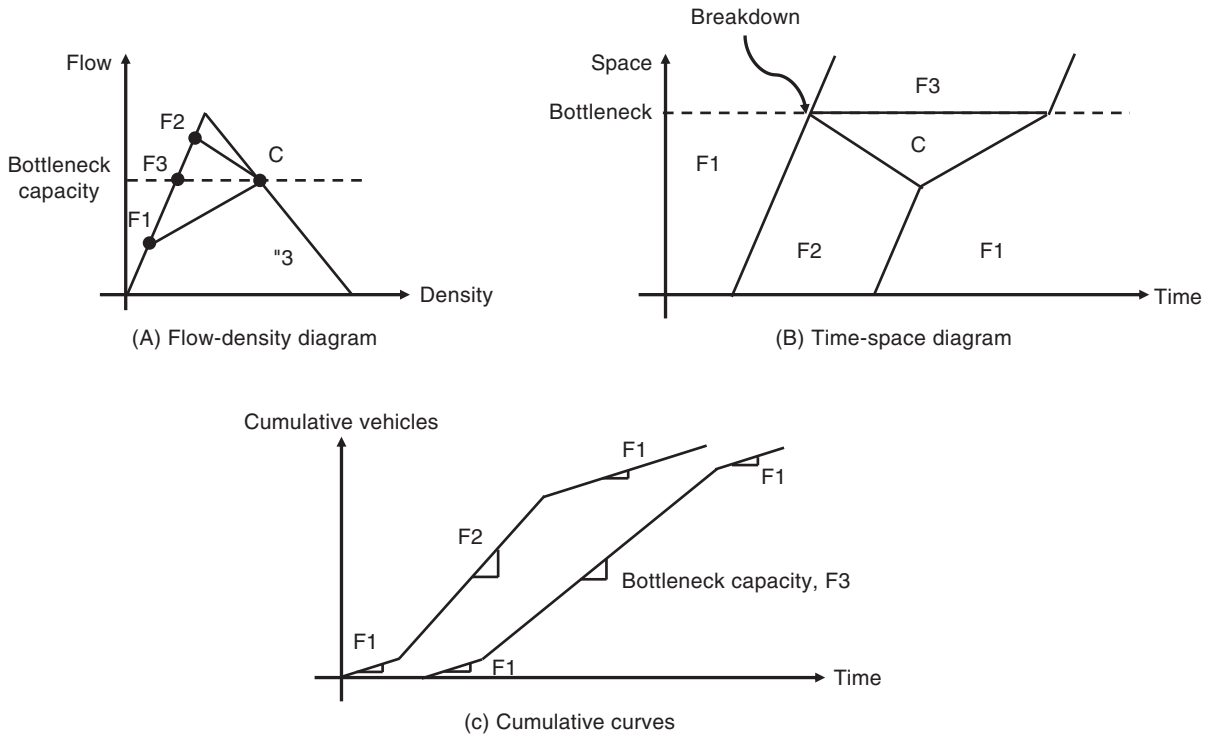


Fig. 2 Typical congestion phenomena in uninterrupted traffic with a bottleneck.

axis indicates time (and toward the right indicates the future), the vertical axis indicates the space (and up indicates the downstream direction of traffic), and solid lines indicate boundaries between different traffic states. Regions F1, F2, and F3 are free-flowing and region C is congested. **Fig. 2C** shows the cumulative curves of traffic, where the upper curve indicates the cumulative vehicle number counted at the upstream end of the section (referred to as the arrival curve) and the lower curve indicates the cumulative vehicle number counted at the bottleneck location (referred to as the departure curve).

As shown in **Fig. 2B**, traffic at the bottleneck breakdowns (see the chapter "Flow Breakdown" for the details) when the arrival flow exceeds the capacity. As a result, a waiting queue ("C" area in **Fig. 2B**) is formed, and it keeps growing as long as the arrival flow exceeds the capacity. As soon as the arrival flow becomes less than the capacity, the queue starts to diminish. Thus, the severity of the congestion is determined by the bottleneck capacity and the demand (or arrival flow) profile.

The same phenomenon can be illustrated in a different way by cumulative curves (**Fig. 2C**). Note that the slope of a cumulative curve is identical to flow, and the vertical distance between arrival and departure curves denotes the number of vehicles between the

two locations. Thus, if the slope of the arrival curve exceeds the bottleneck capacity, the vertical distance between two curves increases, signifying the development of a queue, and so on.

In the flow–density diagram (Fig. 2B), an arrival flow that is less than the capacity is represented as a dot labeled F1, and an arrival flow that is larger than the capacity is represented by dot F2, and the departure flow when the bottleneck is active is represented by dot F3, and the traffic state in the waiting queue is represented as C. When traffic data is collected by a traffic detector, the observed traffic state will be completely different depending on the location of the detector. If the detector is placed at upstream next to the bottleneck, states F1 and C will be observed. If the detector is placed far upstream of the bottleneck, states F1, F2, and occasionally C will be observed. If the detector is placed at downstream of the bottleneck, states F1 and F3 will be observed.

By leveraging this theory, we can detect a bottleneck from detector data. Suppose that the detector on a basic road segment observed (almost) free-flowing traffic only, and its upstream detector observed both free-flowing and congested traffic. Then, we can presume that a bottleneck exists between these two detectors and can roughly estimate its capacity as the upper bound of the downstream detector measurement.

Fig. 3 shows the actual traffic detector data collected along a road section near Takarazuka-West tunnel, Chugoku Expressway, an intercity highway in Japan. The area has no major merge or diverge sections. It is clear from the data that a waiting queue formed between 900 min and 1280 min due to a bottleneck. According to the flow–density diagrams, the location of the bottleneck appears to exist between 20.3 km and 19.3 km locations, and its capacity is approximately 1500 veh/km/lane.

Sag and Tunnel Bottlenecks

What is Sag?

As noted in the first edition of the HCM (Highway Research Board, 1965), it has been known since the dawn of the traffic flow theory that the terrain type and grade can affect the traffic capacity. The main concerns in the HCM were the speed reduction of large vehicles and other adjustment factors that are caused by the grades. Since the early 1980s, it has been found that some of the sags and tunnels¹ that are the vertical alignment of a downgrade section followed by an upgrade section can become active as bottlenecks on freeways (Koshi et al., 1992). The traffic congestion caused by sags and tunnels is common in hilly regions and subaqueous tunnels (e.g. Lincoln Tunnel and Holland Tunnel, see Edie and Foote, 1958). It is claimed that the capacity at a sag is up to 25% lower than the expected value for the uninterrupted flow on a freeway (Koshi et al., 1992). In Japan, almost 60% of traffic congestion on freeways occurs at sags or tunnels (Xing et al., 2014), and considerable academic and practical efforts have been made that attempt to reveal the congestion mechanism and find the effective countermeasures. The next section describes the features of traffic dynamics at sags and tunnels, and its modeling approach. Practical control measures that can be taken to mitigate traffic congestion can be found in Management of Bottlenecks.

Features of Traffic Dynamics at Sags and Tunnels

A possible explanation of the bottleneck mechanism at sags and tunnels is as follows (Koshi, 1986): (1) As traffic volume increases, the utilization ratio of the inner lane increases; (2) Some vehicles on the inner lane slightly decrease the speed on the upgrade section of sags or at the entrance of the tunnel; (3) Vehicles with a higher desired speeds catch up with slower vehicles and are forced to follow them; (4) The number of vehicles following behind the moving bottleneck increases, generating a platoon in which traffic density is locally high; (5) Once a disturbance within a platoon occurs at a bottleneck section for some reason, the deceleration wave propagates upstream, and a breakdown in traffic flow occurs. As is shown in the Movie 1, which is the driver's view when passing through a congestion due to a sag, drivers cannot find any apparent causation for the congestion queue such as merging, diverging, weaving, lane drops, and work zones. In other words, it is quite unclear for drivers where the head of the congestion queue (i.e., bottleneck) begins. For this reason and due to an increase in gradient, drivers may be late in starting their acceleration and result in insufficient and bounded acceleration operation (see Fig. 4), which may cause a considerable capacity drop after the traffic breakdown (for more details, see the article "Flow Breakdown").

Modeling Approach

Several models have been proposed to represent the congestion that occurs at sags from both microscopic and macroscopic perspectives. In the microscopic approach, a number of car-following models have been developed that consider the gradient of the road and the effect of gravity. For example, in order to discount the acceleration for General Motor (GM) based car-following model, Koshi et al. (1992) considered the gradient term ($-\gamma \sin \theta$), in which γ is a sensitivity parameter and θ is gradient difference at sag. Goñi Ros et al. (2016) considered the additive acceleration term in Intelligent Driver Model (IDM) that depicted a driver's compensation behavior and in particular regarding his operation of the accelerator pedal in response to a gradient change at the sag.

The macroscopic approach is rather limited when compared to the microscopic approach. Jin (2018) developed a behavioral kinematic wave (first-order) model taking the location-dependent time gaps and bounded acceleration into consideration to

¹ Although a sag and a tunnel can each be a bottleneck respectively and independently, sags and tunnels are located close together in many cases. It is therefore difficult to separate the effects of a sag and a tunnel, so we do not distinguish them explicitly in this manuscript.

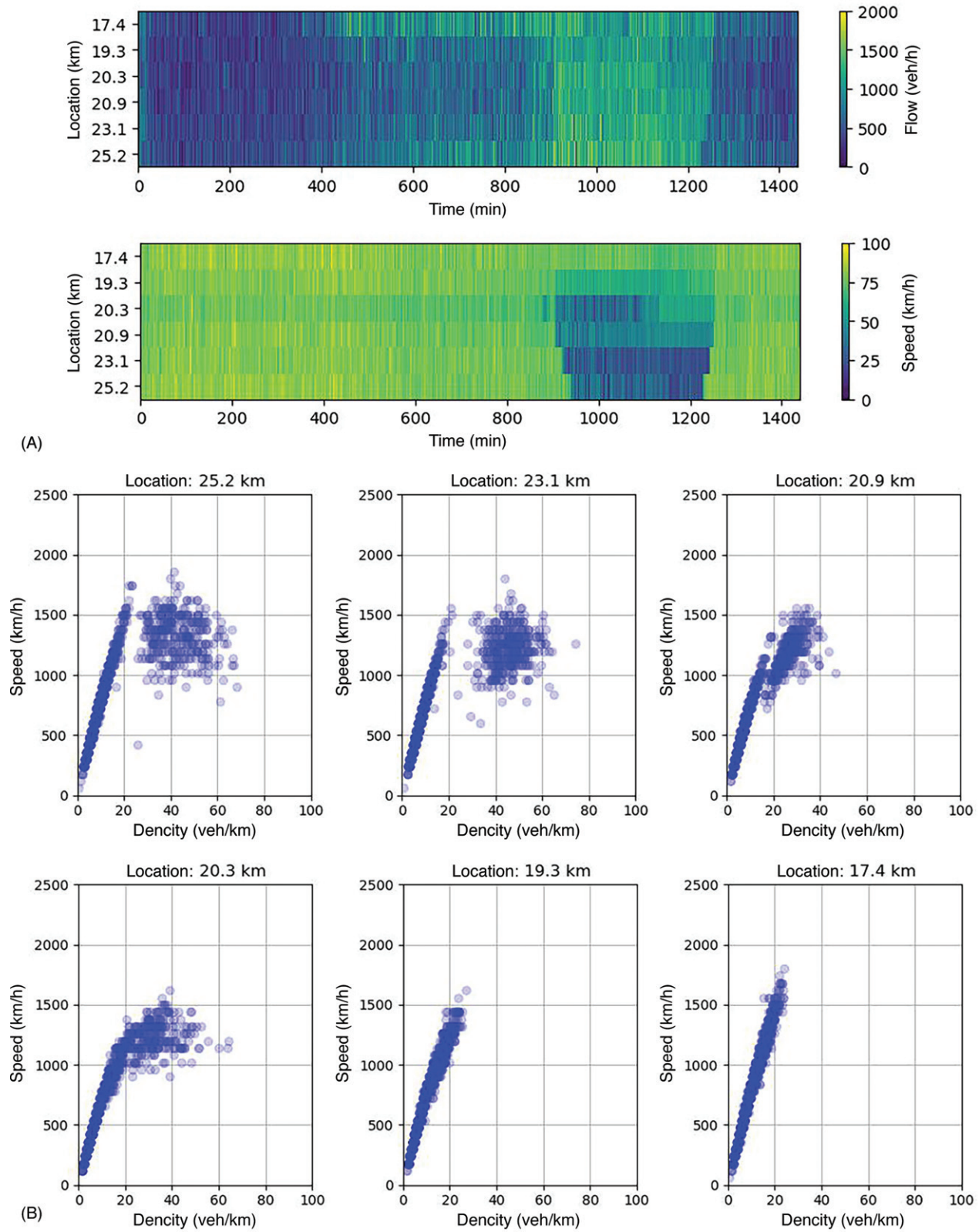


Fig. 3 Time-space diagrams and flow-density diagrams at Chugoku Expressway. Source: Data from *Shiomi et al. (2015)*.

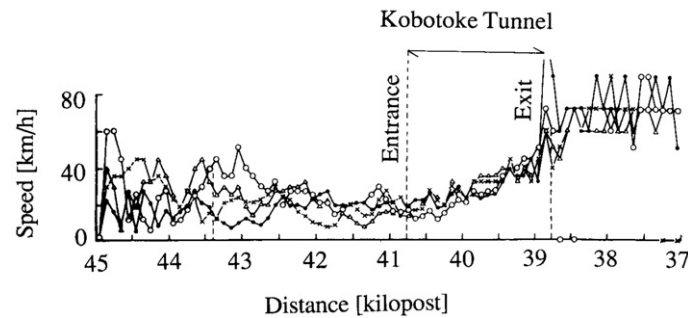


Fig. 4 Speed profile through the congestion at Kobotoke Tunnel. Note: The dashed vertical lines on the distance-axis denote entrance of the tunnel at 40.7 kp and the exit at 38.7 kp. It takes a distance of 2 km to increase the speed from 20 km/h to 60 km/h. Source: Adapted with permission from Koshi M., Kuwahara M., Akahane H., *Capacity of sags and tunnels on Japanese motorways*, Institute of Transportation Engineers Journal.

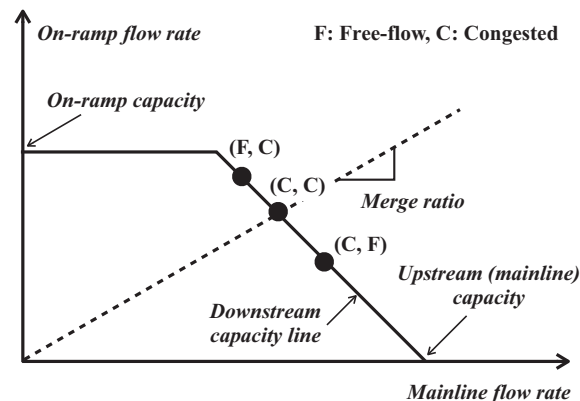


Fig. 5 Merge diagram. Note: Three black dots correspond to three possible states when the merge bottleneck is active. Each dot expresses a pair of mainline and on-ramp states, (mainline state, on-ramp state). Source: Reproduced from Daganzo (1995)

represent capacity reduction, capacity drop, and extremely low-acceleration rates. Wada et al. (2020) extended Jin's model to a second-order one to represent the temporal transition process and the formation mechanism of the capacity drop.

Freeway Network Bottlenecks

This section contains discussions on various types of freeway network bottlenecks such as "merge," "diverge," and "weave." The activation mechanisms using a macroscopic level modeling approach are discussed for each type. Please refer to the chapters such as "signalised intersections" and "signalised roundabouts" for urban network bottlenecks.

Merge Bottleneck

A merge bottleneck becomes active if the sum of traffic demands from upstream approaches (the total demand) exceeds the downstream traffic capacity. At an active merge bottleneck, a capacity drop likely occurs, and whether all or a subset of upstream approaches are congested depends on how the conflicting traffic streams share the downstream capacity. Such competing traffic flow behavior (under a constant downstream capacity assumption) can be explained by the following simple macroscopic theory (Daganzo, 1995), see Fig. 5.

Suppose a merging section has two upstream approaches (i.e., a mainline and an on-ramp), and the total demand exceeds the downstream capacity. In Fig. 5, a traffic state is expressed by a pair of mainline and on-ramp traffic flow rates. The solid lines indicate traffic states in which the traffic flow rate of both or either of the approaches equals to the capacity; if an initial state is located outside of these lines (i.e., demand exceeds capacity), the traffic state converges to a point on the lines as the initial state cannot be lasting. If both approaches have queues [(C, C) dot in Fig. 5], the traffic streams will merge according to some definite priority, that is, the traffic volume of each approach shares a constant percentage of the downstream traffic capacity. The ratio of the percentages of two approaches is called a merge ratio. Moreover, if the demand ratio is disproportionate to a fixed merge ratio, one of the approaches can be non-congested (i.e., the demand can be less than its capacity share). In this case, the other (congested) approach utilizes the remaining capacity [(F, C) and (C, F) dots in Fig. 5].

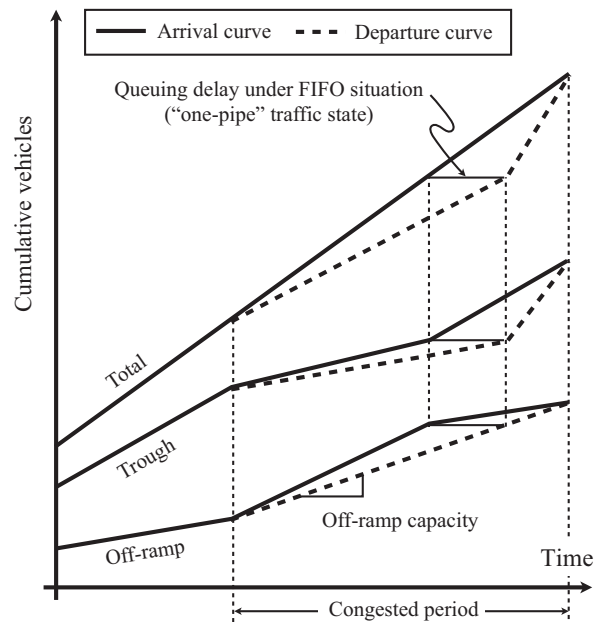


Fig. 6 Queue formation upstream of a diverge due to a fluctuation in traffic composition.

Is the merge ratio actually fixed, for example, irrespective of the severity of congestion? What is the main factor influencing the merge ratio? Cassidy and Ahn (2005) addressed the first question and they have observed that, at four merge sites in California, US and Toronto, Canada, upstream traffic streams enter a congested merge in some (nearly) fixed ratio. It suggests that the merge ratio mainly depends on merge geometry. Further, Bar-Gera and Ahn (2010) investigated 15 different merge sites in California and found that the merge ratio at each site can be reasonably estimated by its lane ratio (i.e., the ratio of the numbers of lanes of two approaches), which is similar to the capacity ratio proposed and (partially) validated by Ni and Leonard (2005).

Diverge Bottleneck

A diverge bottleneck becomes active if the traffic demand to either one of the downstream branches exceeds its traffic capacity. If all lanes of the upstream approach are queued (called a “one-pipe” traffic state) then the traffic that advances into not only the demand-exceeding branch, but additionally the others, experiences the delay equally due to the first-in-first-out (FIFO) discipline. Under FIFO, the specific arrangement and order (or composition) of vehicles with different destinations (i.e., different downstream branches) arriving at the back of the queue must be kept until the point of leaving the queue; thus, the discharge flows of non-congested branches can change significantly without a change in that of the congested one (Daganzo, 1995). The composition change can also cause a sudden traffic breakdown, even if traffic upstream of a diverge (i.e., the total demand) is steady, see Fig. 6.

While the above FIFO situation more likely occurs at an active diverge bottleneck with a narrow upstream approach, Muñoz and Daganzo (2002) showed that it could happen even at a wider one with a 5-lane upstream approach. The reference further showed that non-FIFO situations (called the “multi-pipe” traffic states) could persist for a long time. Congested multi-pipe traffic states, where queued lanes move at different speeds, possibly happen because drivers prefer different lanes depending on their destinations. Semi-congested traffic states, where some lanes are queued and others are not, also exist.

For modeling traffic behavior at a diverge, the simplified FIFO logic above (Daganzo, 1995) would be useful as a first approximation or for a large-scale network analysis. For a more detailed analysis, Daganzo (1997) and Daganzo et al. (1997) proposed a macroscopic multilane-multiclass theory that can approximate both FIFO and non-FIFO situations.

Weaving Bottleneck

Weaving is defined as the crossing of two or more traffic streams traveling in the same general direction. Weaving sections are formed when a merge area is closely followed by a diverge area, or when an on-ramp is closely followed by an off-ramp and the two are joined by an auxiliary lane (Transportation Research Board, 2016). At weaving sections, a lot of lane changes can occur, and their intensity and spatial distribution depend on the weaving demands and their geometrical features of weaving sections such as configuration, length, and width (the number of lanes) (see, Fig. 7 for examples). Several empirical studies showed that the lane changes (of the weaving traffic) are more likely to be concentrated close to the merge (e.g., Cassidy and May, 1991). Such a concentration makes the lane changes more disruptive and may trigger activations of weave bottlenecks.

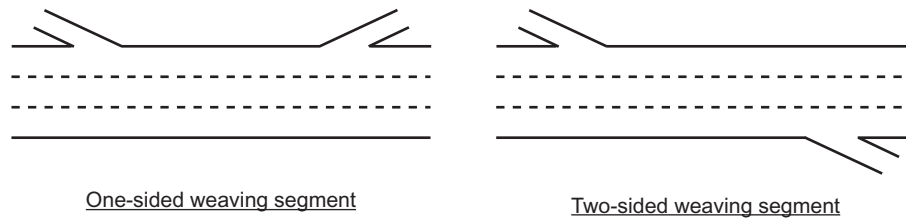


Fig. 7 Examples of weaving sections.



Fig. 8 Dynamic lane assignment: a variable indication of lane role by flood light at Tokyo Metropolitan Expressway.

The impact of systematic lane changes on the capacity (or on the fundamental diagram) can be captured at a macroscopic level, as follows (Jin, 2010). For the duration of a lane change, a lane-changing vehicle occupies the current and target lanes. Therefore, the contribution of lane-changing vehicles to the total density should be doubled. In other words, lane-changing traffic effectively causes additional density, or equivalently, it causes a reduction in the effective number of lanes; and thus, the capacity reduces. In the same study (Jin, 2010), the capability of the theory was demonstrated by using vehicle trajectories by the NGSIM project (U.S. Department of Transportation Federal Highway Administration, 2016). Another explanation for the impact that lane-changing can have on overall traffic flow is that cautious lane changers at a weaving sections can act as moving bottlenecks that can cause a decrease in the overall traffic speed. With the bounded acceleration capability of lane changers, the moving bottlenecks can further contribute to the capacity drop (Laval and Daganzo, 2006).

Management of Bottlenecks

Current Managements

Numerous measures have been proposed or implemented to eliminate or alleviate bottleneck congestion. In this section, some of the typical measures are reviewed. The measures can be roughly categorized into two groups: supply-side measures and demand-side measures. Generally speaking, the former increases the capacity of a bottleneck, whereas the latter decreases the arrival flow to a bottleneck to prevent breakdowns and capacity drops.

The most direct and simple measure to alleviate bottleneck congestion is to physically increase the capacity of a bottleneck. This can be achieved by number of ways involving infrastructure construction, such as road expansion, lane addition, road geometry improvement, implementation of electric toll collection systems, and optimization of traffic signals. However, infrastructure construction is not always practically feasible because of cost or space limitations.

By introducing advanced traffic operation schemes, traffic capacity can be improved or congestion can be prevented. Such schemes are sometimes called as active traffic management. They mainly aim to increase the supply capability of the road operation without largely altering the road infrastructure. In the following text, notable examples of active traffic management schemes are introduced. Dynamic lane assignment, sometimes referred to as interchange merge control, is a control scheme in which the role of lanes is dynamically controlled, depending on the presence of nearby merging and diverging traffic, so as to improve efficiency of lane-changing. In Fig. 8, an actual implementation of dynamic lane assignment is shown. Lane change optimization also aims to improve efficiency of lane-changing at weaving sections in a more actively manner. It designates or informs drivers of optimal lane-changing locations (Mai et al., 2016). Hard shoulder running allows drivers to travel through a shoulder lane near a bottleneck section. It directly improves the physical capacity of a bottleneck.



Fig. 9 (A) Level lines and a signboard (Route 3, Hanshin expressway). (B) Moving-light-guide-system at Hanshin Expressway. Source: Image adapted from Masumoto et al. (2018)



Light (flashing) Light (not flashing) Light (flashing)

Fig. 10 (lower photo) is "Moving-light-guide-system at Hanshin Expressway". Please add the source information. Source: Image adapted from Masumoto et al. (2018)

As sags are the common cause of traffic congestion on freeways, several countermeasures have been taken in practice. Among the several countermeasures, a common idea has been to promote driver awareness and encourage drivers to accelerate more when they are on upgrade sections. Fig. 9 is an example adopted in Hanshin Expressway Route 3 in Japan, in which horizontal blue lines are painted on the walls to notify drivers that they are on an upgrade section. In addition, a signboard has been installed that indicates in words as well as with an illustration that the upgrade section continues for a further 300 m from this point. As another countermeasure, there has recently been an increase in the number of installations of the moving-light-guide-system (MLGS) shown in Fig. 10 and Movie 2 that creates a flow of LED light traveling with constant speed alongside the car and which stimulates the driver to match the speed of the light. The significant positive effects of MLGS on the traffic capacity have been confirmed in many bottlenecks. Currently, the possibility of further elaboration in operation and application of MLGS is being intensively investigated.

A variable speed limit intentionally changes the speed limit of particular section, especially in upstream sections of a bottleneck, at particular times. The purpose is to decrease the arrival flow to a bottleneck in order to prevent breakdown at the bottleneck. Therefore, it maintains free flowing of traffic and prevents potential capacity drops. Ramp-metering controls inflow to a highway at on-ramps to decrease the arrival flow to a bottleneck, as in the variable speed limit. High-occupancy vehicles or toll (HOV/T) lane scheme is a demand-side measure to decrease traffic volume while maintaining passenger volume. It designates some lanes as HOV/T lanes, so that a vehicle with fewer passengers cannot enter these lanes freely. Although HOV/T lanes do not directly increase the capacity of a bottleneck, it does however encourage efficient usage of vehicles and thus, decreases traffic volume in the long term.

Traffic demand management aims to control traffic demand and limit the arrival flow to a bottleneck to a value that is below the capacity. This is a demand-side measure. Variable speed limit, ramp-metering, and HOV/T lane can also be considered as a traffic demand management. Furthermore, information provision and congestion pricing are also used to manage traffic demand on a larger scale. By providing travelers with information on current and predicted travel times for various routes, travelers' route choices can be improved. For instance, if travelers can avoid congested routes, the resulting network traffic assignment might be efficient. However, the efficiency of the information provision scheme is limited in the sense that it only aims to achieve the user equilibrium or user optimal assignment, not the system optimal. Congestion pricing aims to achieve the system optimal assignment. Under congestion pricing scheme, drivers who travel particular road sections need to pay a particular toll, which is designed to achieve political goals such as the alleviation of congestion.

Future Management

In the near future, more advanced management will become possible because of connected and automated vehicles (CAVs). CAVs will change the behaviors of traffic. Furthermore, we will be able to use CAVs to control traffic.

At an operational level, capacity will be increased by CAVs. For example, cooperative driving of CAVs will enable platooning and stable car-following behaviors. The platooning of CAVs will drastically increase the capacity by shortening the headway. Stable car-following behaviors will prevent breakdowns and the propagation of stop-and-go waves. Furthermore, merging will be optimized by controlling the CAVs' lane distribution and entrance timing (Roncoli et al., 2015).

At a strategic and tactical level, demand will be optimized by controlling CAVs. For example, the choice of route and departure time of a CAV could be optimized so as to achieve the system optimal dynamic traffic assignment. In fact, this is possible by introducing a demand management scheme called tradable bottleneck permits (Akamatsu and Wada, 2017). It has been shown that bottleneck congestion can be completely eliminated by implementing this scheme in which "the road manager issues a right that allows a permit holder to pass through the bottleneck at a prespecified time period ('bottleneck permits') and let the permits be traded in a free market. This can be considered as an ultimate management system for bottlenecks as it eliminates all congestions.

Acknowledgment

This research is partially supported by JSPS KAKENHI Grants (16K18164, 19K04637, 19H02268).

References

- Akamatsu, T., Wada, K., 2017. Tradable network permits: A new scheme for the most efficient use of network capacity. *Transp. Res. Part C* 79, 178–195.
- Bar-Gera, H., Ahn, S., 2010. Empirical macroscopic evaluation of freeway merge-ratios. *Transp. Res. Part C* 18, 457–470.
- Cassidy, M.J., Ahn, S., 2005. Driver turn-taking behavior in congested freeway merges. *Transp. Res. Rec.* 1934, 140–147.
- Cassidy, M., May, A., 1991. Proposed analytical technique for estimating capacity and level of service of major freeway weaving sections. *Transp. Res. Rec.* 1320, 99–109.
- Daganzo, C.F., 1995. The cell transmission model, part II: Network traffic. *Transp. Res. Part B* 29 (2), 79–93.
- Daganzo, C.F., 1997. A continuum theory of traffic dynamics for freeways with special lanes. *Transp. Res. Part B* 31, 83–102.
- Daganzo, C.F., Lin, W.H., Del Castillo, J.M., 1997. A simple physical principle for the simulation of freeways with special lanes and priority vehicles. *Transp. Res. Part B* 31, 103–125.
- Edie, L.C., Foote, R.S., 1958. Traffic flow in tunnels. *Highway Res. Board Proc.* 334–344.
- Greenshields, B.D., 1935. A study of traffic capacity. In: *Proceedings of the 14th Annual Meeting of the Highway Research Board*, Highway Research Board, Washington, D.C., 448–477.
- Gofñi Ros, B., Knoop, V.L., Shiomi, Y., Takahashi, T., van Arem, B., Hoogendoorn, S.P., 2016. Modeling traffic at sags. *Int. J. Intel. Transp. Sys. Res.* 14, 64–74.
- Jin, W.-L., 2010. A kinematic wave theory of lane-changing traffic flow. *Transp. Res. Part B* 44, 1001–1021.
- Jin, W.-L., 2018. Kinematic wave models of sag and tunnel bottlenecks. *Transp. Res. Part B* 107, 41–56.
- Koshi, M., 1986. Capacity of motorway bottlenecks. *J. Japan Soc. Civil Eng.* IV 5 (371), 1–7 (in Japanese).
- Koshi, M., Kuwahara, M., Akahane, H., 1992. Capacity of sags and tunnels on Japanese motorways. *ITE J.* 62, 17–22.
- Laval, J.A., Daganzo, C.F., 2006. Lane-changing in traffic streams. *Transp. Res. Part B* 40, 251–264.
- Lighthill, M.J., Whitham, G.B., 1955. On kinematic waves II. A theory of traffic flow on long crowded roads. *Proc. Royal Soc. A* 229 (1178), 281–316.
- Mai, T., Jiang, R., Chung, E., 2016. A cooperative intelligent transport systems (C-ITS)-based lane-changing advisory for weaving sections. *J. Adv. Transp.* 50 (5), 752–768.
- Masumoto, H., Higatani, A., Kodama, T., Kitazawa, T., Suzuki, K., Tomoeda, Y., Lee, Y.H., 2018. Assessment and effective application of the moving light guidance system in hanshin expressway. *JSTE J. Traffic Eng.* 4 (3), B_1–B_9 (in Japanese).
- Muñoz, J.C., Daganzo, C.F., 2002. The bottleneck mechanism of a freeway diverge. *Transp. Res. Part A* 36, 483–505.
- Newell, G.F., 1993. A simplified theory of kinematic waves in highway traffic, part I: General theory, part II: Queueing at freeway bottlenecks, part III: Multi-destination flows. *Transp. Res. Part B* 27 (4), 281–313.
- Ni, D., Leonard, J.D., 2005. A simplified kinematic wave model at a merge bottleneck. *Appl. Math. Model.* 29, 1054–1072.
- Richards, P.I., 1956. Shock waves on the highway. *Operat. Res.* 4 (1), 42–51.
- Roncoli, C., Papageorgiou, M., Papamichail, I., 2015. Traffic flow optimisation in presence of vehicle automation and communication systems – part II: Optimal control for multi-lane motorways. *Transp. Res. Part C* 57, 260–275.
- Shiomi, Y., Taniguchi, T., Uno, N., Shimamoto, H., Nakamura, T., 2015. Multilane first-order traffic flow model with endogenous representation of lane-flow equilibrium. *Transp. Res. Part C* 56, 198–215.
- Transportation Research Board, 1965/2016. *Highway Capacity Manual*, Transportation Research Board, Washington, D.C.
- U.S. Department of Transportation Federal Highway Administration 2016. Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data. [Dataset]. Provided by ITS DataHub through Data.transportation.gov. Available from: <http://doi.org/10.21949/1504477> (Accessed 2019-11-18).
- Wada, K., Martínez, I., Jin, W.-L., 2020. Continuum car-following model of capacity drop at sag and tunnel bottlenecks. *Transp. Res. Part C* 113, 260–276.
- Xing, J., Muramatsu, E., Harayama, T., 2014. Balance lane use with VMS to mitigate motorway traffic congestion. *Int. J. Intel. Transp. Sys. Res.* 12, 26–35.

Flow Breakdown

Kentaro Wada^{*}, Toru Seo[†], Yasuhiro Shiomi[‡], ^{*}University of Tsukuba, Tsukuba, Ibaraki, Japan; [†]The University of Tokyo, Bunkyo-ku, Tokyo, Japan; [‡]Ritsumeikan University, Kusatsu, Shiga, Japan

© 2021 Elsevier Ltd. All rights reserved.

Introduction	143
Characteristics of Flow Breakdown	143
Necessary Conditions for Flow Breakdown	143
High-Traffic Load	144
Bottlenecks and Local Disturbances	144
Stochastic Nature of Flow Breakdown	145
Long-term Variations of Breakdown Flow Rate	146
Associated Phenomena and Their Possible Explanations	146
Capacity Drop	146
Stop-and-Go Traffic	149
Future Direction	150
Acknowledgment	152
References	152

Introduction

Flow breakdown is defined as a spontaneous transition from a free-flow state to a congested state of a traffic state at a certain location. “Spontaneous” in this definition means that the transition is not caused by a queue extension from the downstream sections. This occurs when the arrival flow to the location exceeds the capacity of the location. Therefore, a breakdown happens at a bottleneck (see the chapter “Bottleneck”). In the Highway Capacity Manual (HCM) (Transportation Research Board, 2016), breakdown is explained as “the transition from uncongested to congested conditions. The formation of queues upstream of the bottleneck and the reduced prevailing speeds make the breakdown evident.” and “breakdown occurs when the ratio of existing demand to actual capacity . . . exceeds 1.00.”

The definitions of free-flowing and congested states of traffic are as follows. Traffic is in a free-flow state when most of the vehicles are traveling at their own desired travel speed. On the other hand, traffic is in congested state when most of the vehicles are forced to travel at speeds lower than their desired travel speed, because of too small spacing (i.e., the distance headway is too small). Macroscopic traffic states can be divided into two qualitative categories: free-flowing and congested states. In general, traffic with a small density is free-flowing and traffic with a large density is congested. The threshold density between the free-flowing and congested states is called critical density.

Breakdown is a spatial-temporal phenomenon. Fig. 1 illustrates breakdown at a bottleneck, based on kinematic wave theory (for a basic interpretation of this figure, see the chapter “Bottleneck”). In the flow–density (or fundamental) diagram (Fig. 1A), breakdown can be represented as the transition from the free-flowing states (F1 and F2) to the congested state (C). However, this representation is not complete, because it is not clear where and when the transition happens. In the time–space diagram (Fig. 1B), breakdown can be represented as the onset of congestion at the bottleneck location. If a traffic detector was placed at the direct upstream position of the bottleneck, it would observe the aforementioned transition of traffic state from free flowing to congested. Note that this breakdown is not caused by a queue extension from the downstream sections; the downstream section is always in the free-flowing state. On a microscopic scale (Fig. 1C), flow breakdown can be roughly understood as the sudden speed reduction of a particular vehicle for which its following vehicles must also reduce their speed, although its leading vehicle is free-flowing. This speed reduction grows into a macroscopic waiting queue.

Characteristics of Flow Breakdown

Traffic congestion occurs when traffic demand exceeds traffic capacity. This seems to be an unquestionable fact and, pragmatically, it is true. However, observations of real-world traffic reveal that traffic flow is not as simple, due to its stochastic nature. Traffic demand varies stochastically, and even traffic capacity does. That is why flow breakdown does not always happen at the same traffic demand level even at a same bottleneck point.

Necessary Conditions for Flow Breakdown

According to Treiber and Kesting (2013), traffic flow breakdowns are caused by “the simultaneous action of three factors: high-traffic load, a bottleneck, and disturbances of traffic flow caused by individual drivers.” Fig. 2 illustrates this interrelationship. When traffic flow with a high density encounters a bottleneck, some speed disturbance may occur for a variety of reasons. If the high-traffic

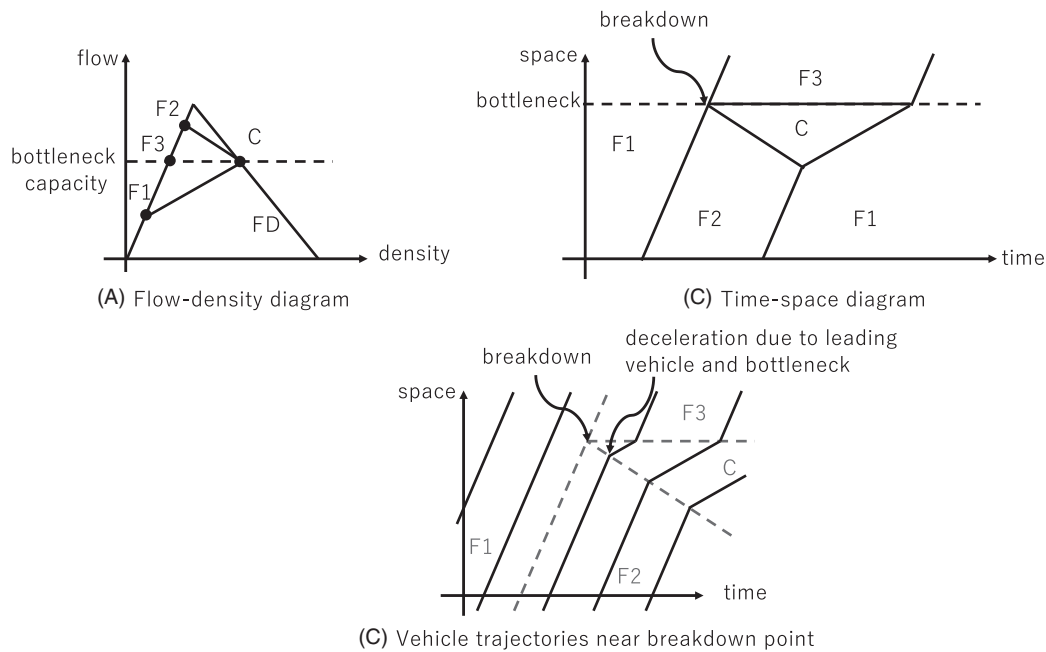


Fig. 1 Breakdown in time-space and flow-density diagrams.

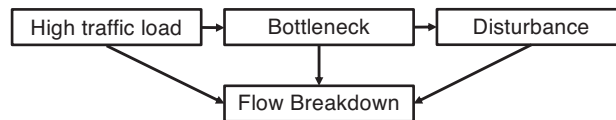


Fig. 2 Schematic representation of relationship between three factors and flow breakdown.

volume still continuously flows into the bottleneck, the disturbance remains there and is amplified. Then, a severe deceleration wave is generated and propagates upstream. Finally, traffic flow breaks down and the discharge flow rate from the bottleneck decreases (see section “Capacity drop” for details).

High-Traffic Load

Vehicles arrive at a bottleneck section randomly, not uniformly. This randomness is caused by the heterogeneity of time headways and desired speeds. A vehicle with a higher desired speed catches up with a slower vehicle and is forced to slow down and follow it. The number of following vehicles increases behind the slow vehicle, and then a platoon in which traffic load is locally and temporarily high is generated (Shiomi et al., 2011), as shown in Fig. 3. In a large platoon, speed disturbances at a bottleneck are likely to occur, propagate, and be amplified. This causes traffic flow breakdown.

On a multilane highway, traffic load is unevenly distributed lane by lane (Shiomi et al., 2015). When traffic density approaches critical density, more traffic is loaded onto the inner lane than the outer and center lanes, as shown in Fig. 4. Inequality in lane use results, at first, in a breakdown of traffic flow in the inner lane. Then, some of the overflow quickly moves to the less congested lane, thereby causing all the lanes to become congested almost simultaneously (Xing et al., 2014).

Bottlenecks and Local Disturbances

Local disturbances are likely to occur at a bottleneck or, conversely, we can say a bottleneck is the place where local disturbances frequently occur. On freeways, the cause of the disturbance at a bottleneck depends on the type of bottleneck. At a lane drop point, work zone, or accident site, where traffic capacity is apparently lower than the section behind the bottleneck, drivers are forced to change lanes and cut in at the bottleneck, which generate speed disturbances. At diverging and weaving sections, lane changes are necessary to get to the destination, which causes speed disturbances as well. At sags, tunnels, and even “rubbernecking” sites, drivers may unconsciously drop their driving speed, in turn stimulating any surrounding drivers to change lanes. Both speed drops and lane changes may cause considerable speed disturbance (Patire and Cassidy, 2011). For more details on a bottleneck refer to the article “Bottleneck.”

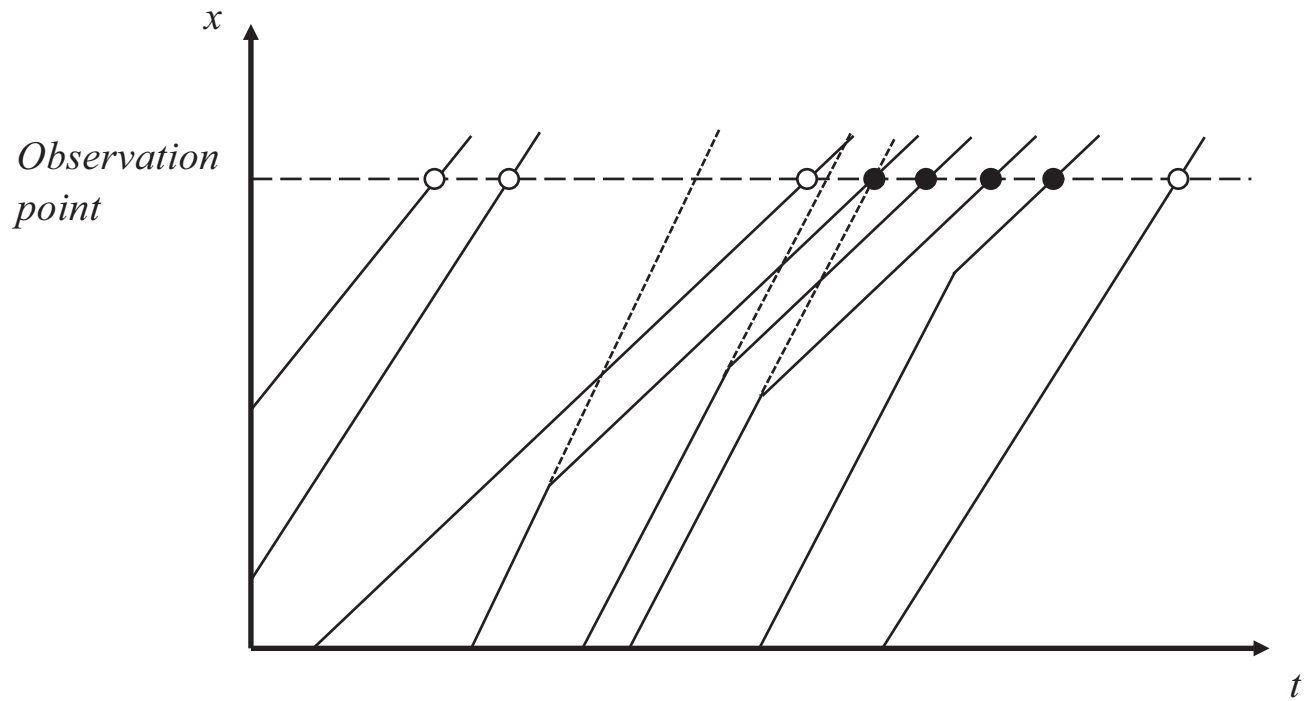


Fig. 3 Trajectories of traffic flow. *Solid lines* indicate real vehicle trajectories, *dotted lines* indicate the trajectories of vehicles driving at their desired speed, *open circles* indicate freely driving vehicles in the study section, and *solid circles* indicate following vehicles.

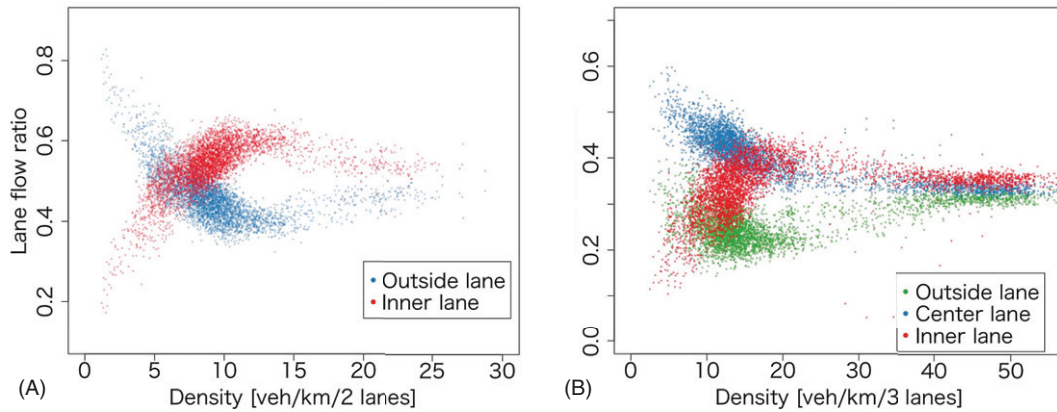


Fig. 4 Observation of lane-flow distribution on freeways. Source: Data from [Shiomi et al. \(2019\)](#)

Stochastic Nature of Flow Breakdown

The breakdown flow rate, which is the flow rate at which traffic flow falls into breakdown and a congestion queue appears, varies even under the same road and environmental conditions, because each necessary condition mentioned above is not deterministic, but stochastic, and vehicle behaviors including freely driving, car-following, and lane changing are heterogenous. Thus, flow breakdowns are probabilistic events, and the concept of stochastic capacity has been proposed ([Brilon et al., 2005](#)). In this concept of stochastic capacity, the capacity distribution function $F_c(q)$ is defined and represented as follows:

$$F_c(q) = p(c \leq q), \quad (1)$$

where $p(\cdot)$ is the probability, c is capacity, and q is flow rate. In accordance with the definition of “deterministic” traffic capacity, that every flow rate greater than the capacity causes a traffic breakdown, the capacity distribution function $F_c(q)$ represents the probability of a traffic breakdown dependent on the flow rate q .

According to the HCM (Transportation Research Boards, 2016), the breakdown probability is estimated by allocating the observed volumes into groups, determining the ratio of the number of breakdown intervals and the total number of intervals for each group, and fitting that to the Weibull distribution. Herein, the fifteenth-percentile value of the breakdown probability is recommended to be the traffic capacity.

However, observed traffic capacity is considered as a type of censored observation. That is, when a traffic flow rate is observed without breakdown, it implies that the traffic capacity at that time is greater than the observed flow rate. This “direct” estimation is sometimes criticized because it ignores bias. To remove bias, a method to estimate breakdown probability has been developed, based on models for lifetime data analysis (Brilon et al., 2005). A non-parametric method to estimate the distribution function of lifetime variables is the so-called “Product Limit Method” (PLM), in which the capacity distribution function $F_c(q)$ is written as Eq. (2).

$$F_c(q) = 1 - S(q) = 1 - \prod_{i: q_i \leq q} \frac{k_i - d_i}{k_i}, i \in \{B\} \quad (2)$$

where q is flow rate [veh/h], q_i is flow rate in interval i (veh/h), k_i is the number of the intervals with a flow rate of $q \geq q_i$, d_i is the number of breakdowns at a flow rate of q_i , and $\{B\}$ is a set of breakdown intervals. For a parametric estimation, the distribution parameters are estimated by applying the Maximum Likelihood Estimation (MLE). Eq. (3) defines the likelihood function L :

$$L = \prod_{i=1}^n f_c(q_i | \beta)^{\delta_i} \cdot [1 - F_c(q_i | \beta)]^{1-\delta_i},$$

where β is a parameter vector of the distribution function of the capacity; n is the number of intervals; and $\delta_i = 1$, if interval i contains an uncensored value; and, $\delta_i = 0$, if interval i contains a censored value. Comparisons between the direct estimation, PLM, and MLE are illustrated in Fig. 5 (Geistefeldt and Brilon, 2009), which shows that the PLM and MLE gives more robust results than direct estimation.

Long-term Variations of Breakdown Flow Rate

Recently, it was reported that the breakdown flow rate has been gradually decreasing, over time, in some countries. Fig. 6 shows the long-term variation of the fifth-percentile value of breakdown probability, from January 2008 to December 2016, at nine typical bottlenecks on freeways in Japan (Shiomi et al., 2019). Fig. 6 indicates that the fifth-percentile traffic volume of breakdown probability decreased over time at eight of the nine sites. On average, the fifth percentile of breakdown probability decreased by 10.8 veh/15 min each year (i.e., 86.4 veh/15 min over eight years), which is equivalent to 91.7% of the initial value. The environment at these bottlenecks has remained unchanged in the eight years, so the decreasing trend may be attributed to the characteristics of vehicles, drivers, or both. This suggests that local disturbances are more likely to occur today than in the past, because the characteristics of traffic flow have changed.

Associated Phenomena and Their Possible Explanations

Peculiar traffic flow phenomena, associated with a traffic breakdown, are usually observed near an active bottleneck, which has negative effects on traffic efficiency and safety. Here, we will discuss two peculiar phenomena: “capacity drop” and “stop-and-go traffic,” and their possible explanations.

Capacity Drop

Once traffic breaks down at a bottleneck, one would logically expect that the queue discharge flow rate from the bottleneck is equal to its traffic capacity. However, in real traffic, the observed queue discharge flow rate decreases by an order of 10%. This phenomenon is called “capacity drop.” This phenomenon has been recognized for many years (e.g., Lincoln and Holland Tunnels, Edie and Foote, 1958), and the shape of associated flow-density relation typically resembles a mirror image of the lambda character (λ), which exhibits a discontinuity between the free-flow and congested states (see, for example, Koshi et al., 1983). The decrease in the queue discharge flow rate prolongs the congestion period, and can incur a profound additional delay of drivers, which has motivated development of several traffic control strategies, such as ramp-metering and variable speed limits.

The hypotheses for the occurrence of the capacity drop phenomenon are mainly divided into two categories, depending on assumed driving behaviors, that is, a (longitudinal) sluggish car-following behavior, and a (lateral) disruptive lane-changing behavior. Let us look at the latter category first. In Laval and Daganzo (2006), it is postulated that lane-changing vehicles create voids in (congested) traffic streams, and that these voids reduce flow. They describe this mechanism with a hybrid model in which lane-changing particles, endowed with a bounded acceleration (BA) capability, are endogenously generated and incorporated into the kinematic wave model as moving bottlenecks. Additionally, they found that the model can reproduce the capacity drop

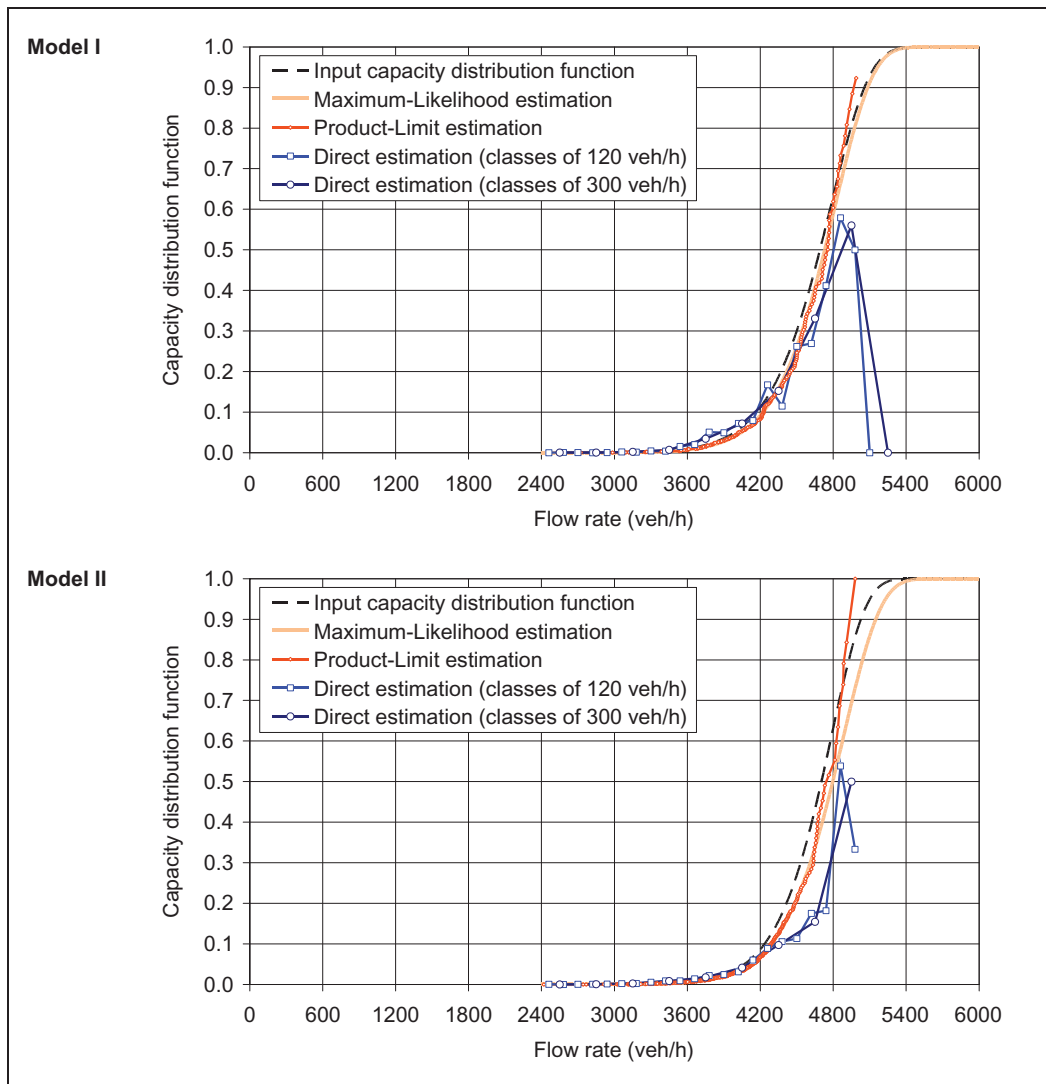


Fig. 5 Input and output capacity distribution functions for simulation model. Source: Adapted from Geistefeldt and Brilon (2009)

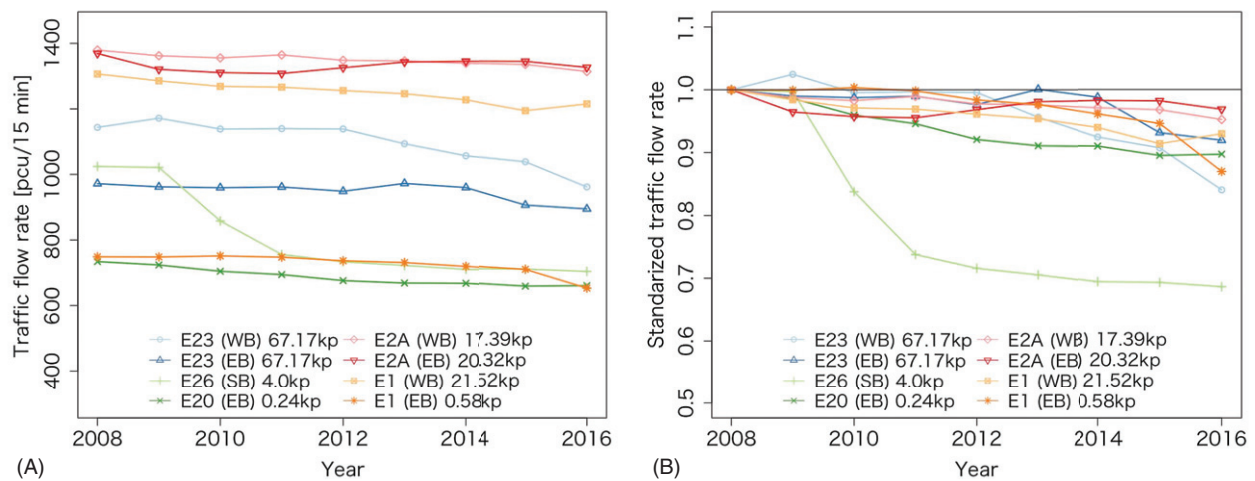


Fig. 6 Long-term variation of fifth-percentile value of breakdown probability Source: Data from Shiomi et al. (2019)

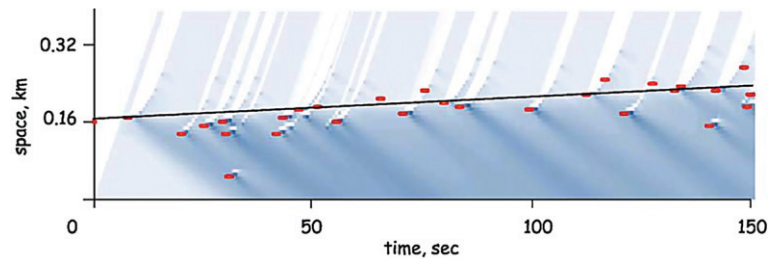


Fig. 7 Lane-changing particles (red dots) and voids (white regions) in traffic stream. Source: Adapted from [Laval and Daganzo \(2006\)](#)

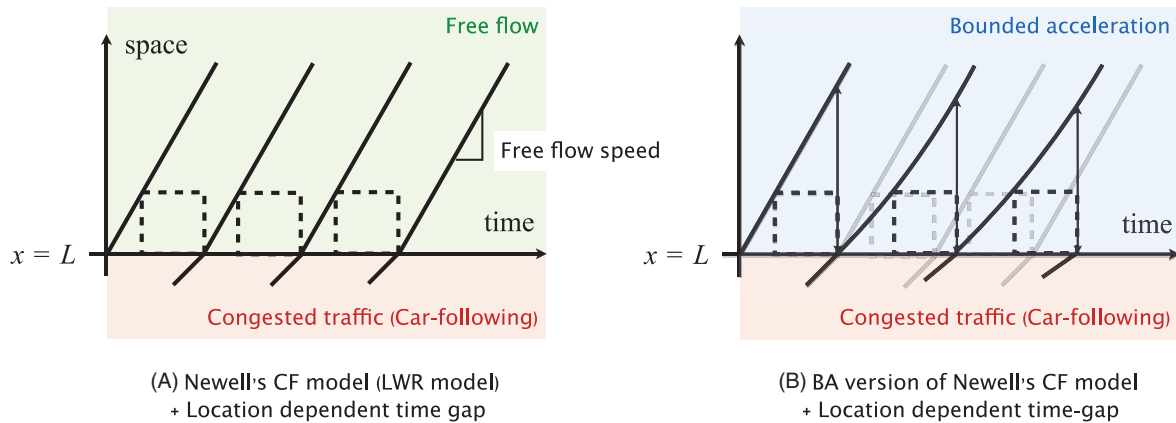


Fig. 8 Illustration of the formation of the capacity drop.

phenomenon (see [Fig. 7](#)), in agreement with empirical observations at a merge bottleneck ([Cassidy and Rudjanakanoknad, 2005](#)). However, in [Cassidy and Rudjanakanoknad \(2005\)](#), it was also argued that “lane changing alone might not explain the capacity drop.” In fact, a significant capacity drop has been observed even in a single-lane freeway bottleneck ([Okamura et al., 2000](#)).

Regarding car-following behavior, it is commonly conjectured that a reduction in the queue discharge flow rate is caused by a low acceleration and/or delayed response of vehicles that leave from queues or traffic disturbances (e.g., oscillations), upstream of bottlenecks ([Hall and Agyemang-Duah, 1991](#); [Koshi et al., 1992](#)). One modeling approach assumes state-dependent changes due to different car-following styles. For instance, [Zhang and Kim \(2005\)](#) describes less sensitive responses to the leading vehicle in car-following models, by a state-dependent time-gap parameter, that depend on the distance-gap (and speed). [Chen et al. \(2014\)](#) also consider some drivers, who become less aggressive and adopt longer response times and minimum spacing when passing traffic disturbances, in their multiclass traffic flow model. However, none of these models treat the capacity drop phenomenon endogenously. Furthermore, because their mechanisms are purely based on (intra- and inter-) drivers’ heterogeneities, insights to the relation between the capacity drop and bottlenecks—spatial heterogeneities—are lacking.

Another modeling approach, which can endogenously describe the capacity drop phenomenon, is founded on two main assumptions: spatial heterogeneities and the BA of vehicles ([Jin, 2017, 2018](#)). The spatial heterogeneity is expressed by a location-dependent fundamental diagram. For example, the jam density diminishes within lane-drop bottlenecks. Additionally, the time gap may increase within sag (and tunnel) bottlenecks, because drivers cannot fully compensate for gradient changes, but try to keep their spacing. Such a location-dependent fundamental diagram describes not only the local reduction in the (static) capacity (i.e., bottleneck effects), but also the delay in the speed adaption to the leading vehicle in the car-following situation ([Wada et al., 2020](#)). As illustrated in a lead vehicle problem by [Fig. 8B](#), the BA version of [Newell \(2002\)](#)’s car-following model ([Laval and Leclercq, 2008](#)), with a specific location-dependent fundamental diagram, leads to a capacity drop. Specifically, if the demand exceeds the capacity, (1) the speed of a vehicle arriving at the end of the bottleneck (at $x = L$) is less than that of the leading vehicle, due to delayed speed adaption; and, (2) unlike in [Fig. 8A](#), the vehicle cannot recover to the free-flow speed instantaneously, owing to the BA constraint, which results in an increase in the headway of the next vehicle at $x = L$ (i.e., the queue discharge flow rate decreases). This vicious cycle [steps (1) and (2)] continues to reach the capacity drop stationary state. Furthermore, this modeling approach can explain a very low acceleration rate, when passing through the bottleneck, during the capacity drop stationary state ([Persaud and Hurdle, 1988](#); [Koshi et al., 1992](#)). [Fig. 9](#) demonstrates that the speed profile in the capacity drop stationary state, by the (calibrated) model ([Jin, 2018](#); [Wada et al., 2020](#)), very well matches that observed in the Kobotoke tunnel in Japan ([Koshi et al., 1992](#)).

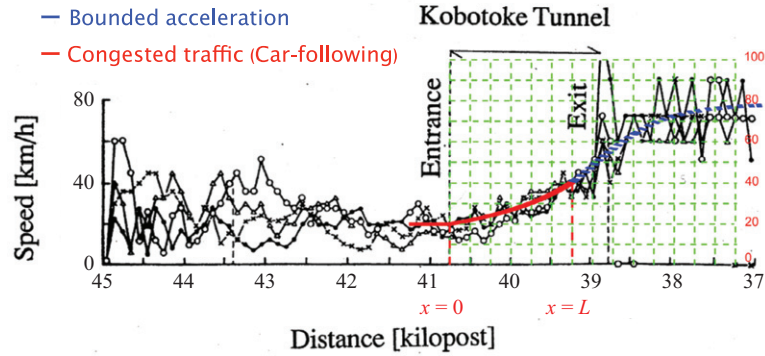


Fig. 9 Comparison between empirical and model's speed profiles in capacity drop stationary state near bottleneck. Source: Image adapted from Jin (2018)

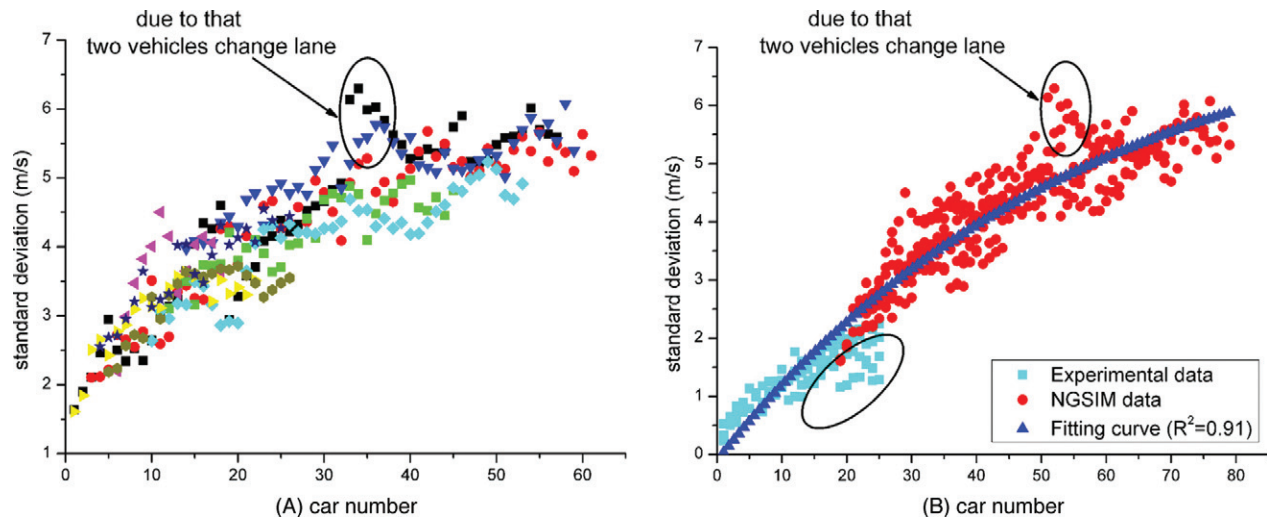


Fig. 10 The standard deviation of the speed of each vehicle. Source: Adapted from Tian et al. (2016)

Stop-and-Go Traffic

Once traffic flow breaks down, deceleration waves are formed around an active bottleneck and propagate regularly in the upstream direction, with an almost constant speed (on the order of 20 km/h). Vehicles are forced to decelerate and accelerate when passing through these traffic disturbances. This phenomenon is called “stop-and-go traffic” or “traffic oscillation,” and was first found in observations at the Lincoln tunnel (Edie and Foote, 1958). It is also known that the oscillations are rather small in amplitude immediately upstream of the bottleneck, but they are amplified as they propagate against the traffic stream; and, the oscillations across lanes are synchronized. Furthermore, Tian et al. (2016) recently found from empirical and experimental data that the standard deviation of the vehicles' speeds increases in a concave manner, during the oscillations (see Fig. 10). While the trigger of stop-and-go traffic can be caused by any traffic disturbance, such as lane changes and/or slow vehicles, there is still active discussion regarding the mechanism of the propagation and amplification of these disturbances.

Conventionally, stop-and-go traffic has been studied as a phase transition or pattern formation, due to the system instability (i.e., the string or flow instability) of car-following models, that are described as a system of (ordinary or delayed) differential equations (Kometani and Sasaki, 1958; Bando et al., 1995; Treiber and Kesting, 2013). This line of research indicates that, regardless of the details of the models, some common driving behaviors such as finite acceleration, deceleration, and reaction times lead to stop-and-go traffic (Treiber and Kesting, 2013). However, an empirical validation of these models is lacking (Laval and Leclercq, 2010). Furthermore, the growth pattern of oscillations, in the typical (homogeneous and deterministic) car-following models, is qualitatively different from the findings above (i.e., a convex growth pattern in the initial stage) (Tian et al., 2016).

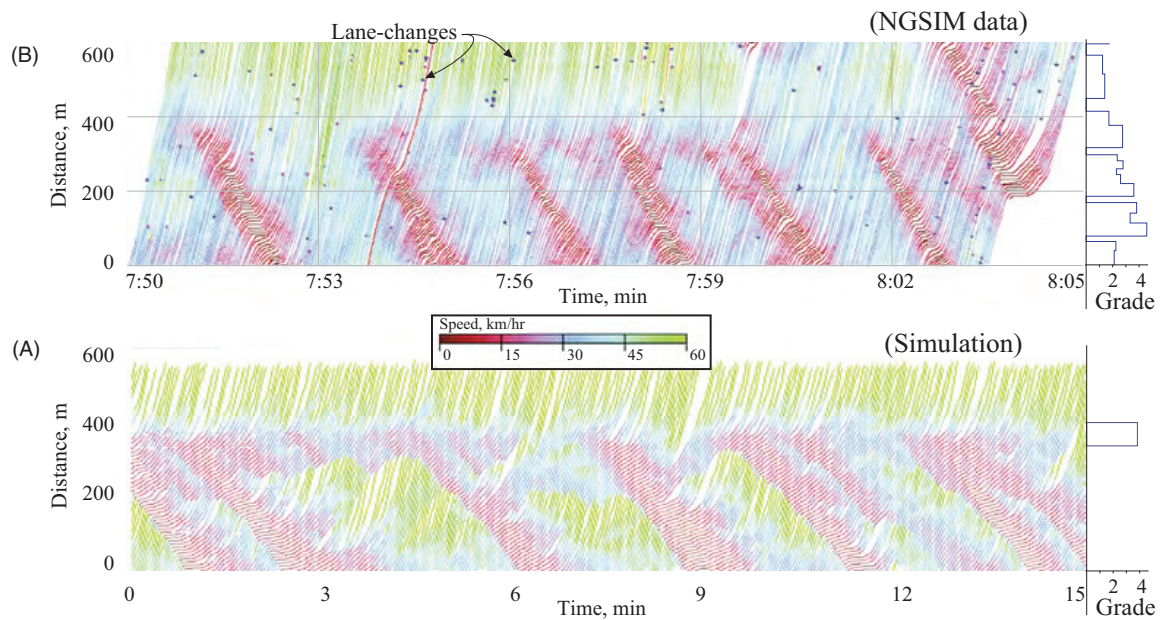


Fig. 11 Vehicle trajectories of NGSIM data and simulation results. Source: Adapted from [Laval et al. \(2014\)](#)

A more behavioral modeling approach is based on extensive empirical analyses of vehicle trajectory data that is available more recently (e.g., NGSIM data, [U.S. Department of Transportation Federal Highway Administration, 2016](#)). Specifically, it has been proposed that oscillations may be caused by changes in the car-following behaviors of drivers when facing traffic waves. [Yeo and Skabardonis \(2009\)](#) proposed an asymmetric traffic theory in which acceleration and deceleration behaviors have different characteristics due to the anticipation and overreaction of drivers. In contrast to the introduction of the intra-driver heterogeneity, [Laval and Leclercq \(2010\)](#) proposed a car-following model with “timid” and “aggressive” drivers, as observed in empirical trajectory data. This study showed that combining a model, without an instability mechanism (i.e., the BA version of Newell’s car-following model, [Laval and Leclercq, 2008](#)), with such an inter-driver heterogeneity model can produce realistic oscillations.

The last but simplest modeling approach is to introduce human errors as random noise. The earliest model by this approach is by [Nagel and Schreckenberg \(1992\)](#), which reproduces stop-and-go traffic by a cellular automata model, with a breaking probability. More recently, [Laval et al. \(2014\)](#) simply added an acceleration noise, which is a function of road geometry, to the BA version of Newell’s car-following model, and found that the oscillations produced accord well with observation (see [Fig. 11](#)). Furthermore, [Tian et al. \(2016\)](#) showed that this model can reproduce the concavity of the disturbance growth well.

Future Direction

In the future, two innovations may occur regarding flow breakdown. First, our understanding of breakdown may be improved by using new data. Second, we may be able to control breakdown by using connected and automated vehicles (CAVs).

Conventionally, traffic data were mainly collected by traffic detectors that are fixed at a location, and their pulse data or aggregated data were used for analysis. Although they enabled us to understand basic properties of breakdown as reviewed in the previous sections, they have clear limitations. That is, they do not capture the continuous spatial-temporal dynamics of traffic. Because breakdown is a continuous spatial-temporal phenomenon that covers approximately one hundred meters and a minute, it is impossible to reveal the detailed, full picture of a breakdown by conventional detectors.

Recent advancements in information and communication technology enable us to collect new data that capture continuous spatial-temporal traffic dynamics. Proper use of these data will improve the scientific understanding of breakdown. Notable examples are as follows. First, complete vehicle trajectory data, obtained by fixed cameras and image recognition technology, are useful experimental data. The data captured the complete, continuous spatial-temporal traffic dynamics of specific road sections where the cameras were installed. For example, NGSIM data covers an approximately 600 m highway section for 15 min, and Zen Traffic Data ([Hanshin Exp. Co. Ltd., 2018](#)) covers a 2 km highway section for 1 hour. An example from Zen Traffic Data is shown in [Fig. 12](#). Second, sampled vehicle trajectory data, obtained by probes or connected vehicles equipped with positioning devices, are becoming common owing to the widespread use of global navigation satellite systems and smartphones ([Herrera et al., 2010](#)). Although they do not include complete information on the traffic (i.e., only sampled trajectories and speed), they can cover wide-

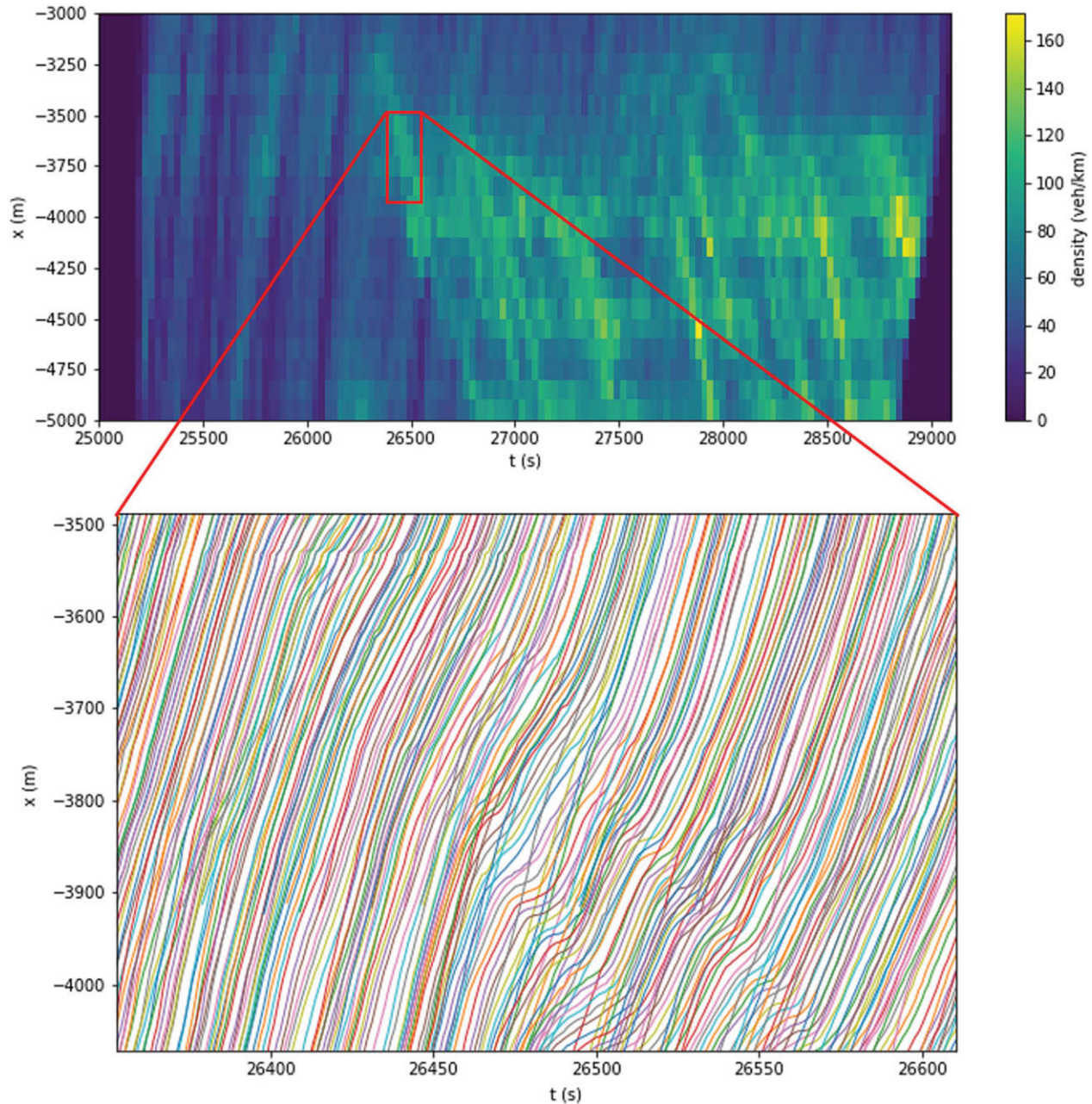


Fig. 12 Complete trajectory dataset obtained by series of video cameras on Hanshin Expressway, Japan. Top: aggregated traffic state. Bottom: trajectories near breakdown. Source: Data from *Hanshin Exp. Co. Ltd.*, (2018).

ranging areas. And third, sampled vehicle trajectory and spacing data, obtained by probe or connected vehicles equipped with ranging devices, will be useful (Seo et al., 2015). Those data capture local density and speed dynamics for wide-ranging areas, from which the complete, continuous traffic state can be estimated. An example of spacing data is shown in Fig. 13.

CAVs will enable breakdown to be controlled. As reviewed in the previous sections, breakdown is likely to be emerging from certain microscopic vehicle behaviors. Thus, by properly designing the driving behavior of CAVs, it may be possible to prevent or alleviate breakdowns. Especially, cooperative driving among CAVs may be useful, because it may enable efficient longitudinal and lateral movements which are impossible for ordinary human-driven vehicles. In-depth understanding of breakdown phenomena will be essential for breakdown prevention by CAVs

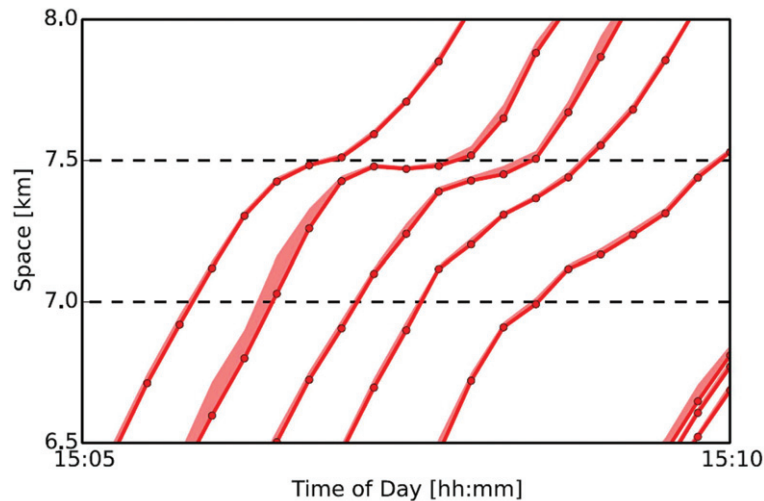


Fig. 13 Spacing data collected by probe vehicles near breakdown on Tokyo Metropolitan Expressway, Japan (Data from Seo et al., 2015). Red curves indicate trajectories of probe vehicles, and red areas indicate spacing between each probe vehicle and its leading vehicle.

Acknowledgment

This research is partially supported by JSPS KAKENHI Grants (16K18164, 19K04637, 19H02268).

References

- Bando, M., Hasebe, K., Nakayama, A., Shibata, A., Sugiyama, Y., 1995. Dynamical model of traffic congestion and numerical simulation. *Phys. Rev. E* 51, 1035–1042.
- Brilon, W., Geistfeldt, J., Regler, M., 2005. Reliability of freeway traffic flow: a stochastic concept of capacity. In: Mahmassani, H.S. (Ed.), *Proceeding of the 16th International Symposium on Transportation and Traffic Theory*, Elsevier, Maryland, pp. 125–143.
- Cassidy, M.J., Rudjanakanoknad, J., 2005. Increasing the capacity of an isolated merge by metering its on-ramp. *Transp. Res. Part B* 39, 896–913.
- Chen, D., Ahn, S., Laval, J., Zheng, Z., 2014. On the periodicity of traffic oscillations and capacity drop: The role of driver characteristics. *Transp. Res. Part B* 59, 117–136.
- Edie, L.C., Foote, R.S., Traffic flow in tunnels. In: *Highway Research Board Proceedings* 1958, Highway Research Board, Washington, D.C., 334–344.
- Geistfeldt, J., Brilon, W., 2009. A comparative assessment of stochastic capacity estimation methods. In: Lam, W.H.K., Wong, S.C., Lo, H.K. (Eds.), *Proceeding of the 18th International Symposium on Transportation and Traffic Theory*, Springer, Boston, MA, pp. 125–143.
- Hall, F.L., Agyemang-Duah, K., 1991. Freeway capacity drop and the definition of capacity. *Transp. Res. Rec.* 1320, 91–98.
- Hanshin Exp. Co. Ltd., 2018. Zen Traffic Data. Available from: <https://zen-traffic-data.net>.
- Herrera, J.C., Work, D.B., Herring, R., Ban, X.J., Jacobson, Q., Bayen, A.M., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transp. Res. Part C* 18 (4), 568–583.
- Jin, W.L., 2017. A first-order behavioral model of capacity drop. *Transp. Res. Part B* 105, 438–457.
- Jin, W.L., 2018. Kinematic wave models of sag and tunnel bottlenecks. *Transp. Res. Part B* 107, 41–56.
- Kometani, E., Sasaki, T., 1958. On the stability of traffic flow. *J. Operat. Res. Japan* 2, 11–22.
- Koshi, M., Iwasaki, M., Ohkura, I., 1983. Some findings and an overview on vehicular flow characteristics. In: Hurdle, V.F., Hauer, E., Stuart, G.N. (Eds.), *Proceedings of the Eighth International Symposium on Transportation and Traffic Theory*, University of Toronto Press, Toronto, pp. 403–426.
- Koshi, M., Kuwahara, M., Akahane, H., 1992. Capacity of sags and tunnels on Japanese motorways. *ITE J.* 62, 17–22.
- Laval, J.A., Daganzo, C.F., 2006. Lane-changing in traffic streams. *Transp. Res. Part B* 40, 251–264.
- Laval, J.A., Leclercq, L., 2008. Microscopic modeling of the relaxation phenomenon using a macroscopic lane-changing model. *Transp. Res. Part B* 42, 511–522.
- Laval, J.A., Leclercq, L., 2010. A mechanism to describe the formation and propagation of stop-and-go waves in congested freeway traffic. *Philos. Trans. Series A* 368, 4519–4541.
- Laval, J.A., Toth, C.S., Zhou, Y., 2014. A parsimonious model for the formation of oscillations in car-following models. *Transp. Res. Part B* 70, 228–238.
- Nagel, K., Schreckenberg, M., 1992. A cellular automaton model for freeway traffic. *J. Phys. I* 2, 2221–2229.
- Newell, G.F., 2002. A simplified car-following theory: A lower order model. *Transp. Res. Part B* 36, 195–205.
- Okamura, H., Watanabe, S., Watanabe, T., An empirical study on the capacity of bottlenecks on the basic suburban expressway sections in Japan. In: *Transportation Research Circular E-C018: Fourth International Symposium on Highway Capacity* 2000, Transportation Research Board, Washington, D.C., pp. 120–129.
- Patrie, A.D., Cassidy, M.J., 2011. Lane changing patterns of bane and benefit: Observations of an uphill expressway. *Transp. Res. Part B* 45, 656–666.
- Persaud, B., Hurdle, V., 1988. Some new data that challenge some old ideas about speed-flow relationships. *Transp. Res. Rec.* 1194, 191–198.
- Seo, T., Kusakabe, T., Asakura, Y., 2015. Estimation of flow and density using probe vehicles with spacing measurement equipment. *Transp. Res. Part C* 53, 134–150.
- Shiomi, Y., Taniguchi, T., Uno, N., Shimamoto, H., Nakamura, T., 2015. Multilane first-order traffic flow model with endogenous representation of lane-flow equilibrium. *Transp. Res. Part C* 59, 198–215.
- Shiomi, Y., Yoshii, T., Kitamura, R., 2011. Platoon-based traffic flow model for estimating breakdown probability at single-lane expressway bottlenecks. *Transp. Res. Part B* 45, 1314–1330.
- Shiomi, Y., Xing, J., Kai, H., Katayama, T., 2019. Analysis of the long-term variations in traffic capacity at freeway bottleneck. *Transp. Res. Rec.* 2673, 390–401.
- Tian, J., Jiang, R., Jia, B., Gao, Z., Ma, S., 2016. Empirical analysis and simulation of the concave growth pattern of traffic oscillations. *Transp. Res. Part B* 93, 338–354.
- Transportation Research Board, *Highway Capacity Manual* 1965/2016, Transportation Research Board, Washington, D.C.
- Treiber, M., Kesting, A., 2013. Traffic flow dynamics: Data. In: *Models and Simulation*, Springer-Verlag Berlin, Heidelberg.

- U.S. Department of Transportation Federal Highway Administration, 2016. Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data. [Dataset]. Provided by ITS DataHub through Data.transportation.gov. Available from: <http://doi.org/10.21949/1504477>.
- Wada, K., Martínez, I., Jin, W.-L., 2020. Continuum car-following model of capacity drop at sag and tunnel bottlenecks. *Transp. Res. Part C* 113, 260–276.
- Xing, J., Muramatsu, E., Harayama, T., 2014. Balance lane use with VMS to mitigate motorway traffic congestion. *Int. J. Intel. Transp. Sys. Res.* 12, 26–35.
- Yeo, H., Skabardonis, A., 2009. Understanding stop-and-go traffic in view of asymmetric traffic theory. In: Lam, W.H., Wong, S.C., Lo, H. (Eds.), *Proceedings of the 18th International Symposium on Transportation and Traffic Theory*, Boston, MA, pp. , 99–115.
- Zhang, H., Kim, T., 2005. A car-following theory for multiphase vehicular traffic flow. *Transp. Res. Part B* 39, 385–399.

Recurrent Congestion

Takahiro Tsubota, 3 Bunkyo-cho, Matsuyama, Ehime, Japan

© 2021 Elsevier Ltd. All rights reserved.

Congestion and Bottleneck	154
Peak in Traffic Demand and Congestion	154
Recurrent Bottlenecks	155
Bottlenecks on Freeways	155
Imbalanced Merging Section	155
Weaving Section	156
Sag and Tunnel	156
Bottlenecks on Surface Streets	156
Summary	156
See Also	157
References	157
Further Reading	157

Congestion and Bottleneck

Traffic congestion is one of the most concerning issues in the transportation system. Roadway sections are characterized by a maximum traffic volume safely passing within a specific time period (i.e., an hour); the maximum volume is the traffic capacity of the section. The capacity changes depending on various factors such as road geometry (e.g., number of lanes, grade) ([Transportation Research Board, 2010](#)), and along a roadway, different sections generally exhibit different capacity. The section with relatively lower capacity in comparison with upstream and downstream sections is identified as a bottleneck section. Typical values of roadway capacity are presented in roadway manuals. [Table 1](#) summarizes examples of freeway capacity and saturation flow rate of signalized intersections, respectively, cited from American and Japanese manual.

Traffic congestion occurs when the amount of traffic demand exceeds the bottleneck capacity, and the congestion forms a waiting queue towards upstream of the bottleneck. It usually is associated with the morning and the evening peak hours for commuting traffic ([Stopher, 2004](#)). As the capacity becomes insufficient for the traffic demand during the rush hours, traffic flow is no longer continuous but becomes interrupted at the bottleneck, then congestion occurs.

Efforts have been made to understand the cause of the bottleneck and congestion. A bottleneck can be roughly classified into two types: recurrent bottleneck due to static factors such as road geometries; and nonrecurrent bottleneck due to temporary conditions. The former includes junctions, intersections, sag sections, and tunnel. The latter includes traffic accidents, road works, adverse weather, and natural hazard. Consequently, traffic congestion is also categorized into two: recurrent congestion and nonrecurrent congestion ([Skabardonis et al., 2003](#)). Recurrent congestion occurs at the same place and the same time on a daily basis, especially on weekdays. It is the delay travelers regularly experience or expect at recurrent bottlenecks during known time of day such as morning and evening rush hour. Nonrecurrent congestion is less common and unexpected for road users; it occurs when nonrecurrent bottlenecks appear due to special events. Interest of this chapter is with the former: recurrent congestion ([Stopher, 2004](#)).

Recurrent congestion can be explained from two aspects. First, it causes due to the peak in traffic demand that exceeds the road capacity. Second, recurrent congestion arises because of a bottleneck. Remaining of this chapter describes the two causes of recurrent congestion.

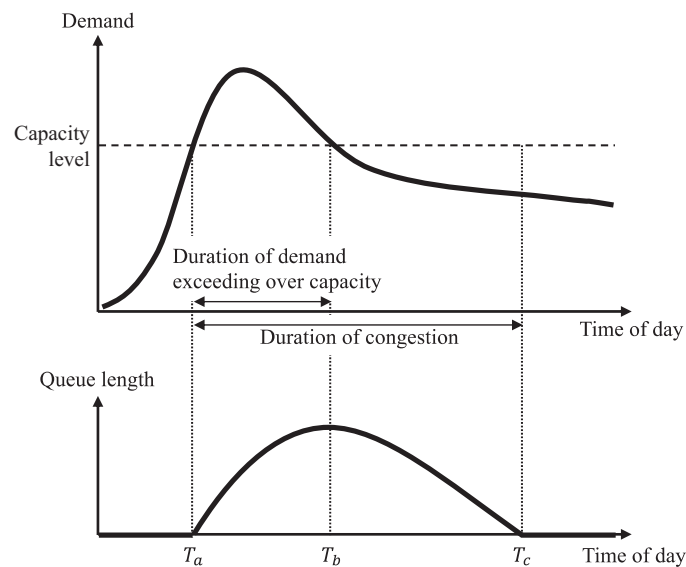
Peak in Traffic Demand and Congestion

Recurrent congestion arises when peak of traffic demand exceeds the road capacity. The excess of traffic demand over the bottleneck capacity forms a waiting queue which develops towards upstream from the bottleneck. [Fig. 1](#) illustrates the typical distribution of demand over time of day with a peak, and with very small demand in the early hours. The capacity is presented by the dashed line in [Fig. 1](#). The congestion starts when the demand exceeds the capacity at time T_a and the excess of the demand continues until time T_b . The excess of the demand accumulates on the road forming a waiting queue. The queue, or the congestion, keeps growing in length between time T_a and T_b . Therefore, the queue length is maximum at time T_b when the traffic demand is lower than the capacity. As the demand gradually decreases after time T_b , the queue length becomes shorter; eventually the congestion dissolves at time T_c . It is noteworthy that the duration of congestion is relatively longer than the duration of demand exceeding over the capacity; the congestion can continue for several hours even when the demand exceeds the capacity for only an hour ([Koshi et al., 1989](#)). It is also observed that the excess of traffic demand over the bottleneck capacity is not necessarily high, approximately 12%–15% in freeways and a few percent in surface streets even when the congestion lasts for a few hours ([Koshi et al., 1989](#)).

Table 1 Base capacity of basic freeway segments and base saturation flow rate of signalized intersections in American and Japanese manual

	USA	Japan
Base capacity of basic freeway segments (PCU per hour per lane)	2300 ^a	2200 ^b
Base saturation flow rate of signalized intersections (PCU per hour of green per lane)	1900	Through lane: 2000 Right-turn lane: 1800 Left-turn lane: 1800

^aValue for multilane freeway sections for free-flow speed of 100 km/h.

**Figure 1** Peak demand and congestion.

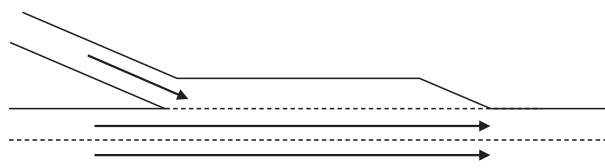
Recurrent Bottlenecks

Bottlenecks on Freeways

Major bottlenecks on European or American freeways are located at merging or diverging sections where intense lane changing is observed. On the other hand, traffic congestion in Japanese freeways is generally triggered at sag sections or at tunnel entrances. This section describes the major recurrent bottlenecks on freeways (Oguchi, 2015).

Imbalanced Merging Section

Bottlenecks can be created at merging areas, when several roadways converge without corresponding increase in drive lanes (Falcocchio and Levinson, 2015). Examples of such lane imbalance are presented in Fig. 2, in which three lanes merging into two. A common cause of freeway congestion comes from where the number of entering lanes on two merging freeways exceeds the number of departing lanes, often observed at on-ramp sections. It is obvious that this condition results in recurrent congestion when the traffic demand surges, because number of vehicles from the upstream lanes exceeds the capacity of downstream sections after the lane convergence. As a result, intense congestion could occur during peak hours if inflow control strategies or mainline traffic controls, such as variable speed limit, are not implemented.

**Figure 2** Imbalanced merging section.

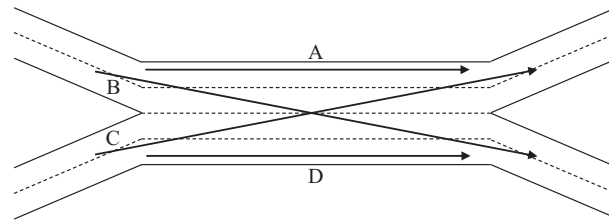


Figure 3 Weaving section and weaving traffic.

Weaving Section

Even when the number of lanes is balanced between upstream and downstream sections, recurrent bottlenecks near freeway junctions are major sources of traffic congestion. At the weaving sections, merging and diverging sections are closely placed and two or more traffic streams travel across each other (Transportation Research Board, 2010). Such lane formation requires intense lane changing behaviors for driver to access the lane appropriate for their desired directions. Fig. 3 depicts a typical formation of a weaving section. Drivers traveling the path B and path C must cross (weave) their ways. The vehicles traveling on the path B and C interact each other by the lane changings, and therefore cause speed reduction and the “capacity-drop” phenomenon. The significance of the capacity-drop depends on the weaving traffic volume; the higher the weaving volume is, the greater capacity-drop is observed at the section.

Sag and Tunnel

The bottleneck due to the capacity-drop phenomenon is also observed at basic freeway sections, where on-ramps or off-ramps are not present. It is reported that traffic congestion frequently occurs at sag sections and entrance of tunnels, accounting for 80% of all congestion in Japanese freeways (Xing et al., 2014). There is no merging or diverging at such sections, but slight speed perturbations due to slight changes in road geometries, such as gradients, cause speed reduction, and causes shock waves that propagates upstream, resulting in traffic congestion.

The sag sections have the vertical alignment with longitudinal slope changes from downgrade to upgrade. Since the sequence of slope changes causes disturbances in the speed of traffic flow, the capacity is lower at sag sections than in flat sections. The mechanism of the capacity-drop can be explained as follows (Koshi et al., 1992). A vehicle unintentionally slows down its speed at the start of the upgrade section due to insufficient acceleration. Then, the following vehicles are forced to decelerate accordingly, which generates a shockwave. The shockwave propagates towards upstream, and congestion occurs. As well, at a tunnel entrance, drivers tend to drive more carefully and slightly reduce their speed due to the sudden change of lighting conditions and the psychological pressure in the tunnel. Consequently, the capacity drop starts at the tunnel entrance, resulting in congestion as is the case at the sag sections.

Bottlenecks on Surface Streets

The most significant causes of congestion at surface streets are the bottlenecks at signalized intersections. The congestion at the signalized intersections is observed under oversaturated conditions, where traffic demand exceeds the capacity determined by signal control schemes. Oversaturated conditions can be defined as the state where part of vehicles in a waiting queue formed during a red light cannot be served during the next signal cycle. Once it occurs, the waiting queue keeps growing in length towards the upstream intersections; the queue blocks the exit of upstream intersections and reduce the throughput rate of the corridor.

Although several signalized intersections exist along a corridor, the bottleneck intersections that lead to traffic congestion are generally fixed. This is because the total green time per signal cycle must be shared among conflicting traffic streams. Traffic signals control the conflicting movements, and the share of green time is determined by the volume of different movements. If the higher traffic volume exists in the crossing direction, sufficient green time cannot necessarily be allocated to the other direction.

Furthermore, there is also a situation where on-street parking close to signalized intersections reduces the capacity of the intersection by blocking a lane available for serving traffic at the intersection. If an intersection has two lanes in a given direction, and when parked vehicles block one lane, there is approximately 50% loss in the capacity.

Summary

This chapter outlined the causes of traffic congestion and major recurrent bottlenecks. Roadway sections are characterized by traffic capacity, a maximum traffic volume that can safely passes within a specific time period (i.e., an hour); the section with relatively lower capacity is identified as a bottleneck section. Traffic congestion occurs at bottlenecks when traffic demand exceeds the bottleneck capacity. Recurrent congestion is the delay travelers regularly experience or expect during known time of day such as morning and evening rush hour, at recurrent bottlenecks where traffic capacity is reduced due to road geometries and traffic control devices. Major recurrent bottlenecks on freeway are located at merging and weaving sections, as well as sag and tunnel sections. At

merging sections, several roadways converge without corresponding increase in drive lanes as usually seen at on-ramp sections, and congestion occurs for the obvious reason; the capacity of the downstream lane cannot accommodate the number of vehicles from the upstream lanes. Weaving section is another source of recurrent congestion, where the capacity-drop phenomena occurs due to intense lane changing behaviors of driver. The capacity-drop is also observed at basic freeway sections, typically at sag and tunnel sections.

The most significant causes of congestion at surface streets are the bottlenecks at signalized intersections. Traffic signals control the conflicting movements, and the share of green time is determined by the volume of different movements. Signalized intersections can be recurrent bottlenecks when sufficient green time cannot be allocated because higher traffic volume exists in the crossing direction. Further, on-street parking vehicles close to signalized intersections also reduces the capacity of the intersection by blocking a lane available for serving traffic at the intersection.

See Also

Weather, Effects of; Non Recurrent Congestion; Bottleneck; Signalized Intersections; Queue Length

References

- Falcochchio, J.C., Levinson, H.S., 2015. *Road Traffic Congestion: A Concise Guide*. Springer, Switzerland.
- Koshi, M., Akahane, H., Kuwahara, M., 1989. Explanation of and countermeasures against traffic congestion. *IATSS Res.* 13 (2), 53–63.
- Koshi, M., Kuwarara, M., Akahane, H., 1992. Capacity of sags and tunnels on Japanese motorways. *ITE J.* 62 (5), 17–22.
- Oguchi, T., 2015. *Traffic engineering Traffic and Safety Sciences: Interdisciplinary Wisdom of IATSS*. (Chapter 4), pp. 40–48.
- Skabardonis, A., Varaiya, P., Petty, K.F., 2003. Measuring recurrent and nonrecurrent traffic congestion. *Transport. Res. Rec.* 1856, 118–124.
- Stopher, P.R., 2004. Reducing road congestion: a reality check. *Transport Policy* 11 (2), 117–131.
- Transportation Research Board, 2010. *Highway Capacity Manual 2010*. Washington, DC.
- Xing, J., Muramatsu, E., Harayama, T., 2014. Balance lane use with VMS to mitigate motorway traffic congestion. *Int. J. Intell. Transport. Syst. Res.* 12, 26–35.

Further Reading

- Margiotta, R.A., Spiller, N., 2009. *Recurring Traffic Bottlenecks: A Primer Focus on Low-Cost Operation Improvements*.

Nonrecurrent Congestion

Takahiro Tsubota, Matsuyama, Ehime, Japan

© 2021 Elsevier Ltd. All rights reserved.

Introduction	158
Causes of Nonrecurrent Congestion	158
Traffic Incidents	158
Inclement Weather	159
Work Zones	160
Special Events and Other Causes	160
Summary	160
See Also	161
References	161

Introduction

Nonrecurrent congestion results from a temporary decrease in capacity while the demand is unchanged, or it also occurs when the demand surges in nonperiodic manners. The primary difference between recurrent and nonrecurrent congestion is the predictability (Margiotta and Spiller, 2009). Recurrent congestion is predictable and typically occurs during the morning and evening peak hours in normal weekdays at bottlenecks due to static factors such as road geometries. On the other hand, nonrecurrent congestion is usually unexpected and hard to be planned for. It results from bottlenecks due to temporary capacity reduction caused by random or unplanned events, such as traffic incidents, inclement weather, and work zones. The nonrecurrent congestion also occurs by surge in demand during special events such as major sports events and recreational events. The occurrence of the nonrecurrent congestion varies both in time and location from day to day, and makes the transportation system unreliable. Obviously, when recurrent and nonrecurrent causes of congestion appear at the same time, their impact on delay are magnified. The amount of delay caused by recurrent causes and by and nonrecurrent causes is challenging to be estimated because the amount of delay produced by the two causes sometimes overlaps. Table 1 shows an estimate of the cause of recurrent and nonrecurrent congestion delay in the US freeways (Carvell et al., 1997; Falcocchio and Levinson, 2015). The estimate shows that nonrecurrent congestion generates a larger share of delay than recurrent congestion. Incidents are the major cause of nonrecurrent delay followed by bad weather, work zones, and special events

The focus of this chapter is on the following four main causes of nonrecurrent congestion:

- Traffic incidents
- Inclement weather
- Work zones
- Special events and other causes

Causes of Nonrecurrent Congestion

Traffic Incidents

Traffic incidents reduce roadway capacity and contribute to congestion occurrence. The amount of delay depends on the type and/or duration of the incident, the number of lanes blocked by the incident, the time to clear the incident location, and the time needed for the roadway to resume normal operation (Potts et al., 2014). Traffic incidents include not only vehicle crashes, but also refer to any events that remove one or more travel lanes from service, including stalled vehicles in travel lanes or on the shoulder. These events tend to block lanes, reducing the capacity of the roadway for vehicles driving by. Even incidents confined to the shoulder slightly reduce the capacity of the roadway, because drivers tend to slow down while passing by the vehicle on the shoulder. Estimates of the amount of roadway capacity available are provided in Table 2, as a function of number of lanes blocked by the incidents (Lindley, 1987). Even when an incident is located at the shoulder of the road, it reduces the roadway capacity.

Capacity reduction due to traffic incidents are also caused not only in the same direction of the incident location, but also in the opposite direction due to the drivers' behavior called "rubbernecking" that occurs when drivers slow down to observe a crash or anything else that may catch their eyes (Potts et al., 2014). The behavior decreases the operational speed of the roadway and also reduces capacity. An estimate conducted in American highways (Masinick et al., 2014) reported that the rubbernecking behavior at a crash site resulted in about 12.7% capacity reduction in the opposite direction of the accident where no lanes were blocked. Another observation of an accident site in Dutch highways (Knoop et al., 2008) revealed a sizable effect on traffic; the capacity decreased roughly by 50% in the opposite direction. At the position that provides the best view of the crash, the traffic begins congested, even though the road ahead is clear. The leader slows its speed to look at the accident site, creating a backward shockwave that propagates

Table 1 Sources of congestions in the US freeways and expressways

<i>Congestion causes</i>	<i>Recurrent (%)</i>	<i>Nonrecurrent (%)</i>
Bottlenecks	40	—
Poor signal timing	5	—
Traffic incidents	—	25
Bad weather	—	10
Work zones	—	15
Special events/others	—	5
Total	45	55

Table 2 Freeway capacity available under incident conditions

<i>Number of freeway lanes in each direction</i>	<i>Shoulder disablement</i>	<i>Shoulder accident</i>	<i>Lanes blocked</i>		
			<i>One</i>	<i>Two</i>	<i>Three</i>
2	0.95	0.81	0.35	0	—
3	0.99	0.83	0.49	0.17	0
4	0.99	0.85	0.58	0.25	0.13
5	0.99	0.87	0.65	0.40	0.20
6	0.99	0.89	0.71	0.50	0.25
7	0.99	0.91	0.75	0.57	0.36
8	0.99	0.93	0.78	0.63	0.41

Source: Lindley (1987)

toward the following vehicles. Once the backward wave reaches, the followers in the wave are also curious about the cause, so they too continue to slow down for a look. At the exact point of the crash, the traffic again begins to accelerate.

Inclement Weather

Driver behavior changes when weather conditions change. Rain and snow reduce visibility and cause drivers to reduce speed. The presence of snow and water on the road surface can also cause slick pavement that can increase slow traffic (Alfelor et al., 2009). Heavy rain and snow may cover the roadway, and hides road markings such as lane lines, resulting in longer headway as drivers move with caution; consequently, the roadway capacity significantly decreases. Among various weather conditions, the effect of rainfall on traffic has been intensively studied, and some reports on the capacity reduction effect. An empirical study conducted in Japanese metropolitan expressway clearly showed the decrease in free flow speed and in capacity with increasing amount of rainfall as presented in Fig. 1 (Chung et al., 2006). The capacity decreases by 4%–7% when light rain of 1 mm/h is present, and the larger

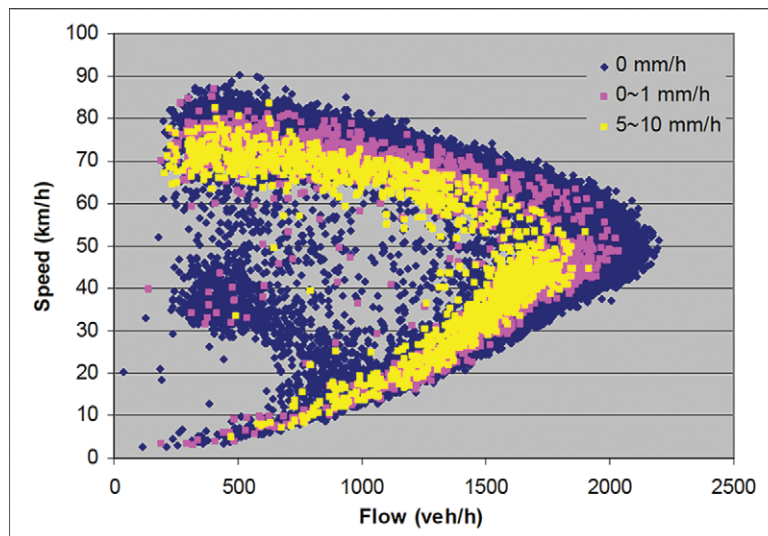
**Figure 1** Example of speed-flow curve for difference precipitation levels. Source: Chung et al. (2006).

Table 3 Capacities at work zone

Number of lanes		Capacity (vehicle per hour per lane)
Normal	Closed	
2	1	1515
3	1	1490
3	2	1440
4	1	1520
4	2	1480
4	3	1170

Source: Chin et al. (2004)

reduction up to 14% is observed for heavy rain condition. The Highway Capacity Manual (Transportation Research Board, 2010) provides capacity adjustments for different weather conditions. It shows that the capacity decreases by 2.01% in light rain with precipitation of up to 2.5 mm/h, and by 14.13% in heavy rain of over 6.4 mm/h, which largely agrees with the observation in Japan.

Work Zones

Work zones are sections on the roadway or roadside where construction, maintenance, or utility works are taking place. The degree to which work zones affect the capacity of the roadway is influenced by various factors; an investigation identified several important factors which can significantly affect work zone capacity (Weng and Meng, 2012), including the number of lane closures and work activity characteristics.

The capacity of a work zone is significantly affected by the number of lanes on the roadway as well as the number of lanes closed. The capacities of open lanes in work zones has been estimated as a function of the available lanes by using the data from literatures (Dixon et al., 1996; Transportation Research Board, 1985). Table 3 shows the example of the estimates of work zone capacity in urban area (Chin et al., 2004).

The work zone capacity is also affected by work activity characteristics, for instance work intensity. Work zone intensity has been classified into three levels: low, medium, or high (Karim and Adeli, 2003), and the work zone capacity can decrease as the work intensity increases from the lightest (e.g., guardrail installation) to the heaviest (e.g., bridge repair). The work time (daytime or nighttime) is another work activity characteristic which also affects work zone capacity. Night construction or maintenance could decrease work zone capacity because attention of drivers is reduced during the night time, according to an observation (Al-Kaisy and Hall, 2001). Work zone duration (i.e., short-term or long-term) is the third feature of work activity. In general, the capacity reduction in work zones is greater in short-term work zones than that in long-term work zones; short-term work zones can have a larger impact on the capacity because they are often unexpected, and drivers are not able to plan their trips to account for the delay they cause. However, the longer the work zone is present, the more drivers become aware of it and become familiar with congestion incurred by the work zone. Some drivers may start choosing alternative routes, and eventually the demand through the work zone is reduced. In addition, at the long-term work zone, frequent drivers become familiar with the configuration of the work activities, and thus the effect on the capacity can be minimized (Weng and Meng, 2013).

Special Events and Other Causes

Special events that cause nonrecurrent congestion often happen near arenas, stadiums, conference centers, and other large gathering spaces that host sports or music events, and large meetings. Special events tend to create large demand increase during the short period of time, usually just before the event starts and just after it ends. The higher traffic demand is also caused typically on weekends near vacation destinations. Another type of demand spike that could cause nonrecurrent congestions is that experienced during major evacuations.

Nonrecurrent congestion is also caused by malfunctions and/or suboptimal traffic control. Traffic control devices include traffic signals, ramp meters, speed limits, pavement markings, lane-use signs, and others. An example of nonrecurrent congestion due to a traffic control device that many drivers could experience occurs when a traffic signal malfunctions and goes to flashing operation, or when it becomes “out of plan” during a high demand period (Chin et al., 2004). Traffic may back up several road sections around the signal that is not functioning as it should. Traffic control devices can cause nonrecurrent congestion on freeways as well when traffic management devices such as ramp meters, or variable speed limit signs fail or malfunction.

Summary

This chapter described the causes of nonrecurrent congestion. Nonrecurrent congestion is incurred by temporary decrease in capacity due to unexpected events, or it also occurs due to sudden increase in traffic demand when major gathering of people takes place. Four main causes that have been identified in literatures are traffic incidents, inclement weather, work zones and special events/others.

Traffic incidents involves lane closures by crashed or stalled vehicles on travel lanes and the shoulder, reducing the traffic capacity at the incident location. As well, the incident causes capacity reduction on the opposite direction of roadway by the “rubbernecking” behavior of drivers observing the incident location. Inclement weather also reduces the capacity of roadway causing nonrecurrent congestion, because driving behavior changes under rain and snow due to reduced visibility and slick pavement. Third major cause of nonrecurrent congestion is work zones. The work zones usually block travel lane, and the capacity at the location decreases. Furthermore, drivers’ behavior varies depending on the characteristics of work activities, causing different effects on the roadway capacity. Finally, nonrecurrent congestion occurs during special events taking place; the events include sports or music events and large meetings where large traffic demand is created. Literatures have also identified that malfunctions or suboptimal traffic control could be a cause of unexpected congestion.

See Also

Construction Zones; Weather, Effects of; Bottleneck

References

- Al-Kaisy, A., Hall, F., 2001. Examination of effect of driver population at freeway reconstruction zones. *Transport. Res. Rec.: J. Transport. Res. Board* 1776 (1), 35–42.
- Alfeli, R., Mahmassani, H.S., Dong, J., 2009. Incorporating weather impacts in traffic estimation and prediction systems. *Inst. Transport. Eng. Annu. Meet. Exhibit* 2009, 443–457.
- Carvell, J.D., et al., 1997. *Freeway Management Handbook*, Washington, DC.
- Chin, S.M., Franzese, O., Greene, D.L., Hwang, H.L., Gibson, R.C. (2004), *Temporary Loss of Highway Capacity and Impact on Performance: Phase 2*, Oak Ridge National Laboratory, Report no. ORNL/TM-2004/209, November 2004.
- Chung, E., et al., 2006. Proceedings of the 5th International Symposium on Highway Capacity and Quality of Service. Country Rep. Spec. Session Pap. 1, 139–146.
- Dixon, K.K., Hummer, J.E., Lorscheider, A.R., 1996. Capacity for North Carolina freeway work zones. *Transport. Res. Rec.: J. Transport. Res. Board* 1529 (1), 27–34.
- Falcochchio, J.C., Levinson, H.S., 2015. *Road Traffic Congestion: A Concise Guide*. Springer, Switzerland.
- Karim, A., Adeli, H., 2003. Radial basis function neural network for work zone capacity and queue estimation. *J. Transport. Eng.* 129 (5), 494–503.
- Knoop, V.L., Hoogendoorn, S.P., Van Zuylen, H.J., 2008. Capacity reduction at incidents: empirical data collected from a helicopter. *Transport. Res. Rec.* (2071), 19–25.
- Lindley, J.A., 1987. Methodology for quantifying urban freeway congestion. *Transport. Res. Rec.* 1–7 (1132).
- Margiotta, R.A., Spiller, N., 2009. *Recurring Traffic Bottlenecks: A Primer Focus on Low-Cost Operation Improvements*.
- Masnick, J., Teng, H., Orochena, N., 2014. The impact of rubbernecking on urban freeway traffic. *J. Transport. Technol.* 4 (1), 116–125.
- Potts, I.B., et al., 2014. *Design Guide for Addressing Nonrecurrent Congestion*, Design Guide for Addressing Nonrecurrent Congestion. Washington, DC.
- Transportation Research Board, 1985. *Highway Capacity Manual*, Special Report 209. Washington, DC.
- Transportation Research Board, 2010. *Highway Capacity Manual* 2010. Washington, DC.
- Weng, J., Meng, Q., 2012. Ensemble tree approach to estimating work zone capacity. *Transport. Res. Rec.: J. Transport. Res. Board* 2286 (1), 56–67.
- Weng, J., Meng, Q., 2013. Estimating capacity and traffic delay in work zones: an overview. *Transport. Res. C: Emerg. Technol.* 35, 34–45.

Freeway to Arterial Interfaces

Abolfazl Karimpour, Yao-Jan Wu, Smart Transportation Lab Department of Civil and Architectural Engineering and Mechanics, University of Arizona, Tucson, AZ United States

© 2019 Elsevier Ltd. All rights reserved.

Introduction	162
Freeway Mobility	162
Arterial Mobility	165
Interface: Integrated Freeway and Arterial	167
References	167

Introduction

The rapid growth in population and car ownership per capita significantly increases the amount of congestion throughout the transportation network in rural and urban areas. Traffic congestion is a day-to-day phenomenon that directly causes an increase in delay and fuel consumption at both freeways and arterials. Freeways and arterials are two subnetworks of the transportation network that complement each other. Freeways accommodate the mobility of the transportation network, while arterials accommodate the accessibility of the transportation network. One of the primary strategies of transportation agencies is to use the available Intelligent Transportation System (ITS) technologies and innovative data-driven strategies to provide and maintain an efficient transportation system that keeps a balance between mobility on freeways and accessibility on arterials. Through the Moving Ahead for Progress in the 21st Century Act (MAP-21), the US Congress requires that all the state Departments of Transportation (DOTs) and Metropolitan Planning Organizations (MPOs) periodically monitor and evaluate the safety and mobility performance of their jurisdiction's roadways. Generally, the overall performance of a transportation network is defined based on mobility and safety measures.

The mobility of people and goods is one of the essential objectives of an efficient transportation network. Mobility measures provide a base for transportation agencies to evaluate the current operational condition of the transportation network. Mobility information, in conjunction with the ITS technologies, can provide valuable information for both traffic engineers and road users. Travel time and vehicle speed are two of the indicators of transportation network mobility that can help to develop strategies to alleviate delays and congestion. Improving roadway safety and decreasing traffic crash frequency is another essential objective of an efficient transportation network. The number of fatalities, severe injuries, and property damage only crashes are the metrics proposed by FHWA for measuring the safety performance of the transportation network. These metrics provide insight into the DOTs and MPOs to establish and report their safety targets (Neuman, 2003).

A simple and straightforward solution to enhance freeways and arterials safety and mobility would be constructing new roads or adding additional lanes to the existing roads. However, due to the limited infrastructure and high construction cost, this solution is not always viable. Recently, with the advent of ubiquitous technologies and the emergence of modern traffic sensors, new approaches, and ITS technologies for improving mobility, accessibility, and safety in the transportation network are emerging. Active traffic management (Sullivan and Fadel, 2010, Mirshahi et al., 2007), restriping the freeway (Kondyli et al., 2019), ramp metering control strategies (Asgharzadeh et al., 2019), automated traffic signal performance measures (Chamberlin and Fayyaz, 2019), real-time crash prediction (Ariannezhad et al., 2018, Parsa et al., 2020), microlevel (Razi-Ardakani et al., 2018, Ariannezhad et al., 2014, Ariannezhad and Wu, 2018) and macro-level (Ariannezhad et al., 2020) safety analysis, crash prediction and hotspot analysis (Mansourkhaki et al., 2016, Mansourkhaki et al., 2017), Advanced Driver Assistance Systems (ADAS), such as automatic emergency braking, blind-spot monitoring, and integrity risk analysis of autonomous vehicles (Hassani et al., 2018, Hassani et al., 2019) are some of the innovative emerging approaches recently adopted by transportation engineers to enhance roadways mobility and safety. This chapter will primarily cover the recent efforts done by transportation agencies for monitoring and improving mobility at freeway and arterial levels. However, the safety component of transportation and its relationships to traffic operations is out of the scope of this chapter.

Freeway Mobility

Freeways are the main arteries that are aimed to create the highest level of mobility between regions, cities, and states. Commuters expect to move from an origin to a destination at free-flow speed with minimum delay. However, due to the nonrecurrent and recurrent traffic congestion, this is not always possible (Papageorgiou and Kotsialos, 2002). To tackle the freeway traffic congestion problem and conduct robust performance assessments, transportation agencies rely heavily on accurate and complete traffic data. Traditionally, traffic data was collected using on-road sensors such as pressure hoses, piezo-electric sensors, and inductive loops. Although collecting traffic data with on-road sensors is effortless, but the drawbacks with these types of data collection are limited

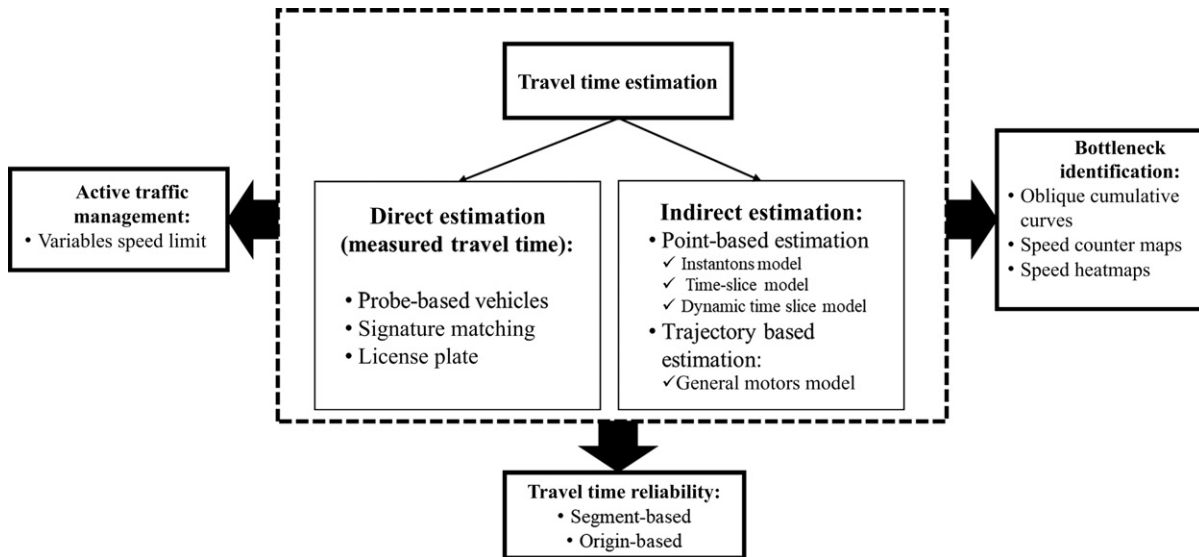


Figure 1 Travel time estimation methods and travel time applications.

coverage and high cost of implementation and maintenance (Leduc, 2008). In the past few decades, the emergence of new traffic sensors has significantly improved the traffic data collection process. Probe-based GPS data, video sensors, and floating car data are some of the most recent methods for collecting traffic data. However, the traffic data collected from these sources provide different levels of accuracy and completeness. Therefore, before utilizing the traffic data into analysis, data quality control tests need to be performed. Low quality and incomplete traffic data will negatively impact traffic projects' outcomes. The most common way to tackle the low quality and incomplete data is to use prediction (Duan et al., 2016), interpolation (Chen and Shao, 2000), statistical-learning (Ran et al., 2015) techniques, and hybrid models that include both prediction and statistical-learning techniques (Karimpour et al., 2019).

Travel time is one of the most crucial measures used for quantifying the mobility and efficiency of freeways and is directly impacted by freeway capacity. Oversaturated conditions severely disrupt traffic flow and increase the travel time by imposing massive delays on travelers (Asgharzadeh and Kondyli, 2018). However, travel time is not always available in broad spatial-temporal coverage to transportation agencies. Therefore, over the past decade, methods for estimating travel time have gained increasing attention. Regardless of the network size (freeway or arterial) or mode of transportation, travel time plays a significant role in transportation planning, design, and operation (Wu et al., 2004, Yang et al., 2016). Travel time data spans a variety of applications, including travel time reliability, quantifying freeway congestion, accessibility, and also for traffic safety studies (Yang et al., 2015). Fig. 1 illustrates a summary of different approaches for estimating travel time and its different applications.

Methods to estimate travel time could be categorized into direct travel time estimation (also known as measured travel time) and indirect travel time estimation. Traffic data collected from probe-based GPS data and test vehicles are commonly used for direct travel time estimation. In contrast, traffic data collected from existing traffic data collector infrastructures on the road, such as loop detectors and microwave radars, are generally used for indirect travel time estimation (Coifman, 2002). Traffic data collected from probe-based vehicles usually suffer from sparse and missing values and are usually labor-consuming. Therefore, indirect travel time estimation is preferred over direct travel time estimation (Yang et al., 2016).

Methods for estimating indirect travel time are classified into point-based and trajectory-based methods. Instantaneous model, time-slice model, and dynamic time slice model are three of the most known point-based methods used for estimating travel time (Cortes et al., 2002). Point-based methods assume that there exists a linear relationship between speed variability at upstream and downstream of a corridor. However, this assumption fails to consider the speed change within a corridor. Therefore, to improve the estimation accuracy, vehicle-trajectory-based methods were developed to improve the accuracy of travel time estimation. It is worthwhile to mention that both point-based and trajectory-based methods perform similarly under normal traffic conditions, but their accuracy will degrade significantly under congested traffic conditions (Yang et al., 2016).

Recently, researchers are focusing more on methods that are accurate even under congestion traffic conditions. This idea was initially proposed by Coifman (2002). Coifman estimated travel time by constructing vehicle trajectories using two-regime macroscopic models (Coifman, 2002). Later on, Yang et al. (2016), inspired by Coifman (2002), incorporated General Motors (GM) models to estimate travel time with less than seven percent MAPE under congested traffic conditions. Yang et al. (2016) evaluated their proposed model on the data collected from three corridors on I-270 in the Greater Saint Louis area. The results of their study showed that the GM-based travel time estimation model outperforms other conventional methods, such as the Instantaneous and Time Slice models. Fig. 2 is proposed by Yang et al. (2016) as the relationship between the GM-based model, macroscopic traffic models, vehicle trajectory, and travel time estimation models based on their work.

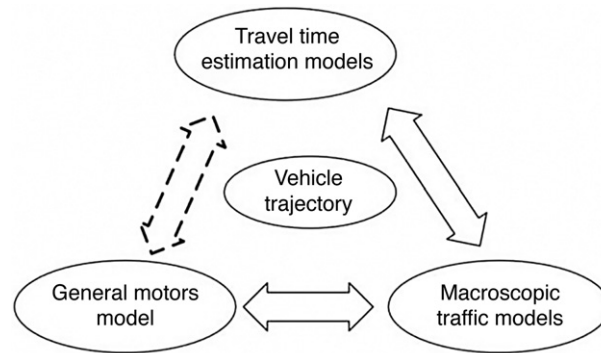


Figure 2 Relationship between general motors models, macroscopic traffic models, and vehicle trajectory and travel time estimation models. Source: Yang et al. (2016).

Recalling Fig. 1, estimated travel time could be used for estimating freeway travel time reliability (TTR), identifying freeway bottleneck, and as an input to Active Traffic Management (ATM) systems. TTR is a critical metric that is widely used for quantifying the mobility and efficiency of a freeway. TTR quantifies the level of satisfaction of the commuters and is the key to understanding on-time arrivals and delays. Generally, to estimate the amount of TTR segment-based or origin-destination-based (OD-based) methods are used. Segment-based methods are based on the estimated travel time data from fixed sensors (Yang and Wu, 2016, Yang et al., 2014). The OD-based methods estimate TTR based on the trip (origin and destination) travel time. For either approach, travel time, and travel time distribution need to be estimated (Yang et al., 2014). Gamma, Weibull, lognormal, Gaussian, and exponential (Polus, 1979, Emam and Al-Deek, 2006, Guo et al., 2010) are the most popular distributions used for fitting the travel time data. Guessousa et al. (2014) used six statistical distributions, Lognormal, Gamma, Burr (extended by Singh-Maddala), Weibull, a mixture of two Normal distributions, and a mixture of two Gamma distributions to estimate travel time. Then, they used the data collected from a two-lane urban motorway ring in France to conduct their study. The results of their study showed the superiority of using Singh-Maddala distribution in terms of quality of fit and practicality (Guessous et al., 2014). In another study, Yang and Wu (2016) used the data collected from an 8.1-mi segment of the I-270 southbound corridor and used Gaussian, Lognormal, and Gamma models to fit travel time data. The results of their statistical analysis showed that travel time could be more accurately fitted using mixture models. TTR, along with other reliability measures, has been widely used by different state departments of transportation for monitoring and evaluating the roadways. For instance, The Minnesota Department of Transportation used TTR to evaluate the effects of a ramp meter shutdown on Minneapolis St. Paul freeways (FHWA, 2017). Washington State Department of Transportation also used reliability measures for performance monitoring, reporting, and prioritizing their freeway's section improvement plans (Ishimaru and Hallenbeck, 1999).

Based on Fig. 1, estimated travel time data could also be utilized in identifying bottlenecks for quantifying the freeway congestion. Fluctuation in demand and the geometric design of the freeway are causes of recurrent bottlenecks. Meanwhile, nonrecurrent bottlenecks form due to traffic crashes, breakdowns, and other unpredicted events. Recently, extensive efforts have focused on identifying and characterizing both types of freeway bottlenecks. Based on the speed collected from a pair of upstream-downstream detector stations, the bottlenecks of a freeway can be identified simply by using oblique cumulative curves. Oblique cumulative curves are developed by subtracting a reference flow from the cumulative number of passing vehicles (Yuan et al., 2017). Speed counter maps are an alternative to the oblique cumulative curves that examine the overall picture of the traffic state in a segment of a freeway. Many transportation agencies, such as Caltrans Performance Measurement System (PeMS, 2017) and Portland Oregon Regional Transportation Archive Listing (PORTAL, 2017), are using speed counter maps. Another popular tool for identifying and tracking freeways bottlenecks is speed heatmaps. Fig. 3 is an example of a speed heatmap that visualizes the bottleneck hotspots over space and time. Fig. 3 visualizes the Spatiotemporal Congested Areas (STCA) along milepost 136 to 160 of I-10, Arizona (Ou et al., 2018).

Another application of the estimated travel time is in ATM systems (as shown in Fig. 1). ATMs are systems that control, monitor, and regulate traffic conditions with the primary goal of enhancing traffic safety and mobility (You et al., 2018). Variable Speed Limit (VSL) signs are one example of these systems that has gained much attention recently. Based on the traffic conditions, these signs dynamically adjust the posted speed limit. VSL signs are used to increase and homogenize the average headway, harmonize traffic flow, and reduce the speed variance (Ha et al., 2003, Abdel-Aty et al., 2006). The speed limit shown on a VSL is based on the traffic conditions upstream and downstream of the VSL (Grumert et al., 2018).

Other applications of travel time are real-time traffic monitoring, providing real-time route information to the commuters, and as a traffic condition indicator for traffic operation centers. It is worthwhile to mention that in order to have an efficient traffic network, transportation agencies need to develop strategies that can operate the freeways and arterials together as an interconnected system. The following section discusses the most recent efforts developed on enhancing arterial and local roadways.

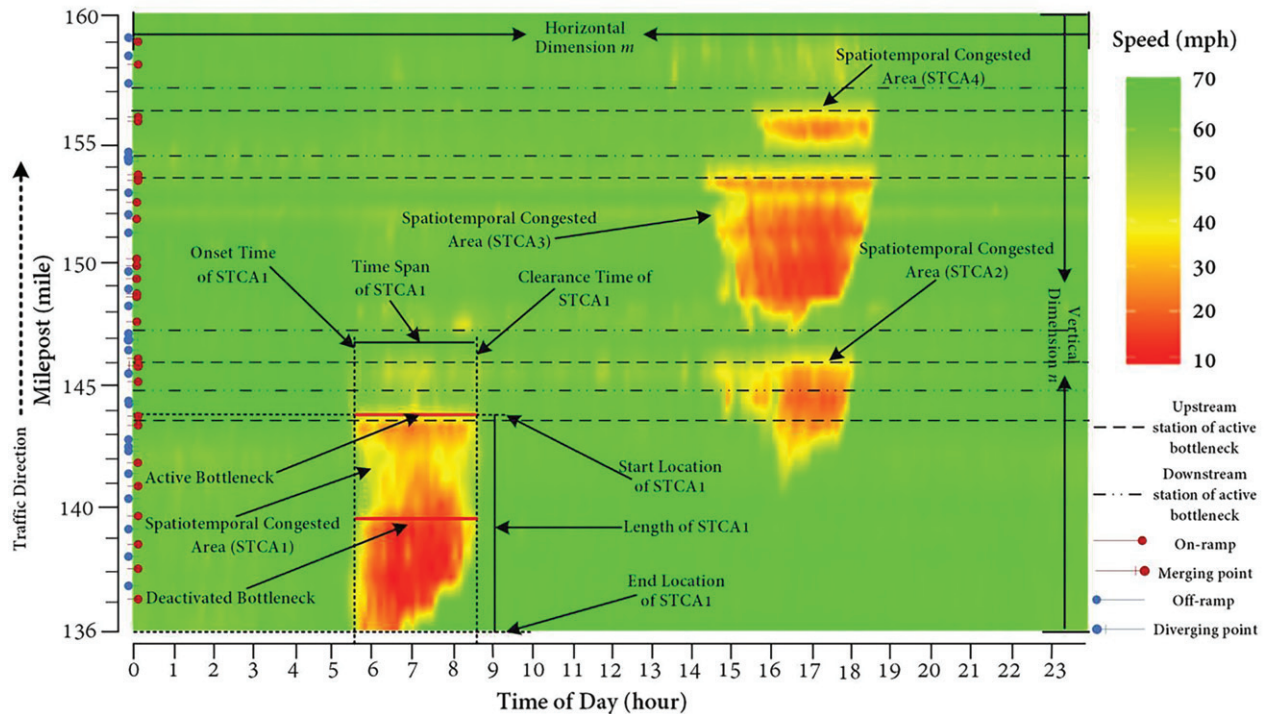


Figure 3 Traffic condition over space and time. Source: *Ou et al. (2018)*.

Arterial Mobility

Arterials are the connector of the national transportation system to regional mobility. Arterials play a significant role in providing accessibility to residential and commercial neighborhoods and are essential to the regional economy and residents' quality of life. In the United States, arterials connect more than one million lane miles of roadways to the national highway system. According to the Urban Mobility Report by Texas Transportation Institute, traffic congestion on arterials caused commuters an extra 8.8 billion hours of delays, which impacted the fuel consumption of the users by an extra of 3.3 billion gallons of gas (TTI, 2019). Therefore, improving the arterials traffic conditions is an essential part of every transportation improvement plan. Local transportation agencies are actively seeking strategies and technologies to monitor, maintain, and improve traffic conditions at intersection and corridor levels.

Currently, more than 330,000 intersections with active signal controllers are operating in the US transportation network (Fehon and O'Brien, 2015). Traffic signals play an essential role in distributing the right of way among pedestrians, cyclists, passenger vehicles, and transit vehicles. Implementing correct signal timing provides direct and indirect benefits. Reduction in the delay and travel time are two of the most tangible direct benefits of signal retiming for commuters. The indirect benefits of signal retiming are reduction in fuel consumption, air pollution, pavement wear and tear, and motorist frustration, and enhancement in traffic safety (Schneider, 2011, Srinivasa Sunkari, 2004). Due to the rapid change in the traffic patterns on the arterials, transportation agencies need to retime and synchronize traffic signals periodically (Li et al., 2019). As soon as the traffic patterns change significantly, the performance of the traffic signals will deteriorate. Therefore, to address the change in traffic patterns, traffic engineers need to ideally re-evaluate the efficiency of the signal controllers and potentially retime the signals annually (Srinivasa Sunkari, 2004). However, many factors, including cost, availability of resources, and data availability, restrict transportation agencies from retiming their signals. Traffic data is a fundamental base for doing any signal retiming studies. Traditionally, traffic data was collected manually by humans. However, recently, traffic data is collected using intrusive (e.g., Loop detectors), and non-intrusive traffic sensors (e.g., video-based detectors) (Tang et al., 2015). Collecting traffic data using intrusive sensors is much easier and more manageable. However, intrusive traffic sensors are saw-cut into the pavement and will significantly deteriorate the pavement condition (Klein et al., 2006). Therefore, transportation agencies are actively seeking newer sensors or data-driven approaches for achieving necessary traffic data for developing signal timing plans.

Recently, video-based sensors are receiving increasing attention and are used in a wide variety of applications, including collecting real-time traffic data, real-time traffic monitoring, and surveillance (Semertzidis et al., 2010, Zhang et al., 2007). In order for the video-based sensors to collect real-time traffic data, such as turning movement counts (TMC), virtual-loop detectors should be configured at each approach. However, configuring these virtual detectors is a trivial and time-consuming task for engineers. Therefore, to avoid configuring virtual detectors, Advance Event Detectors (AEDs) are used (Li et al., 2019). Fig. 4 shows a virtual



Figure 4 A sample AED virtual detector in Tucson, AZ. Source: Li et al. (2019).

AED detector in an intersection in Tucson, Arizona. These AED detectors will ease the process of collecting TMC for the transportation agencies and will allow them to retune signal plans more frequently.

Another approach for gathering traffic data for signal retiming is to predict data based on the available and historical data. Recently, researchers are leaning toward using predicted data as a substitute for the real-time data for developing traffic signal timing controls (Min and Wynter, 2011). Methods for predicting traffic volume are either univariate (only considers the temporal feature of traffic), or are multivariate prediction (considers both spatial and temporal feature of traffic) (Wu et al., 2016). Historical averaging, time-series-based methods such as autoregressive integrated moving average (ARIMA) (Karimpour et al., 2017) are some examples of univariate traffic volume prediction methods. Spatio-temporal random effects (Wu et al., 2016), Spatio-temporal (ST) ARIMA (Kamarianakis et al., 2005) are some examples of multivariate traffic volume prediction methods.

Monitoring and regulating traffic at the corridor level are other crucial tasks for local transportation agencies. At the corridor level, average vehicle speed and speeding behavior indicate the performance of the corridor mobility. National Highway Traffic Safety Administration (2017) reported that speeding-related crashes account for more than 25 percent of all the fatality crashes (NHTSA, 2017). With the infrastructure limitation and the growth in vehicle ownership, agencies are required to develop the most effective and practical solution for tackling the speeding problems for improving mobility and safety at the corridor level. Speed Management Strategies (SMS) are one of the emerging ITS solutions that recently received considerable attention in different states. SMS are multi-disciplinary approaches that control and regulate vehicle's speed and speeding behavior using enforcement, road design, and technology applications (Bagdade et al., 2012). The NHTSA defines speed management strategies as a balanced program that involves the relationship between speed, speeding, and safety (NHTSA, 2006). The main goal of every speed management strategy is improve traffic mobility, public health, and traffic safety (NHTSA, 2014). An effective SMS should concentrate on four primary countermeasures: Education, Engineering, Enforcement, and Emergency services. Over the last 40 years, by educating drivers, public awareness has significantly raised, and improvement has been achieved in reducing fatal and injury crashes (NHTSA, 2014).

Engineering countermeasures influence driver speed choice and are usually categorized into traffic control devices, road and street design, and traffic calming approaches (Bagdade et al., 2012). Traffic control devices, such as speed feedback signs, advisory signs, Optical speed bars, and pavement marketing, are cost-effective approaches to reduce drivers' speed (Karimpour et al. 2020, Santiago-Chaparro et al., 2012). Reducing lane width and raised median are some of the road and street design approaches that have been proven helpful in reducing average speed (Gross et al., 2009). Speed hump, roundabouts, traffic circles, signal timing, and corridor progression are some of the proven effective traffic calming approaches (Halkias et al., 2017). Enforcement countermeasures are usually implemented on the location where it is critical to achieve and increase the posted speed limit. Onsite law enforcement is a common enforcement strategy that has been proven effective in reducing the operating speed and increasing the speed limit compliance (Stevanovic et al., 2011). However, one major drawback to law enforcement is the limitation of resources in many rural and urban areas. Emergency services include the response time of the emergency medical units and the care received by the victim after a traffic incident or crash has happened.

For commuters to experience high mobility, accessibility to local neighborhoods and businesses, transportation agencies need to invest in innovative traffic solutions at both intersection and corridor levels. It is worthwhile to mention that pedestrians, cyclists, passenger vehicles, and transit vehicles are the primary users of arterials. Therefore, the solution adopted by the local transportation agencies must consider all the factors impacting the quality of trip for multimodal transportation commuters, rather than just focusing on a specific group of commuters. The next section provides insight into the interface between freeways, arterials, and commuters.

Interface: Integrated Freeway and Arterial

Transportation agencies around the world are rapidly adopting innovative ITS technologies and data-driven strategies for utilizing the existing capacity of transportation facilities in order to improve traffic efficiency. Such ITS technologies and strategies have been shown in many studies to be effective. For instance, ramp metering and incident management are effective in reducing travel time by up to 48% and decreasing crash rate by up to % on freeways (Urbanik et al., 2006). Additionally, arterial management strategies such as Advanced Traffic Management Systems are effective in reducing travel time by up to 15% and decreasing emission by up to 13% (FHWA, 1996). However, the operation of freeways and arterials are closely linked together.

Lack of coordination between the freeway and arterial usually causes extreme inefficiency in the transportation system (Mahmassani et al., 1998). Therefore, state and local transportation agencies must work closely together on developing integrated strategies for monitoring and regulating traffic on both freeways and arterials (Urbanik et al., 2006). An efficient strategy could quantitatively and qualitatively investigate the potential interaction between freeways and arterials to maximize the total network throughput and reduce congestion (Wu et al., 2010). This emerging concept is known as “Integrated Corridor Management (ICM)” (Hardesty and Hatcher, 2019). ICM is a set of systems that help the transportation sectors to operate their transportation networks at the optimal level. The main goal of every ICM system is to efficiently and safely move people and goods, by collaboration between different transportation sectors, and integrating the current transportation infrastructures (e.g., freeway, arterial, transit, etc.) (Christie et al., 2015). Compared to traditional management strategies, ICM systems emphasize on coordinated, multimodal, cross-network transportation along various corridors within a region. In addition, ICM systems provide solutions to solve a variety of challenges within a region, including incident management, special event management, emergency management, and recurrent delays caused by incidents.

In 2006 USDOT selected five states (eight sites) as the pioneers for implementing ICM systems. The USDOT required the pioneers to define and select the initial technologies and approaches (such as ramp metering, congestion pricing, signal optimization, transit priority) for implementing the ICM systems in their regionwide transportation network. In 2017, the evaluation results of ICM implementation in three pioneers’ sites, Dallas, TX, San Diego, CA, and Minneapolis, MN, revealed a significant improvement in corridor mobility, congestion, and incidents management and reporting (Hardesty and Hatcher, 2019). The evaluation results show the effectiveness of ICM systems for monitoring and regulating traffic on both freeways and arterials. Although national deployment of ICM is not possible at the moment, the motivation to deploy ICM systems in transportation improvement plans are being considered among many regional transportation sectors.

This chapter summarizes the efforts done by different levels of Departments of Transportation for collecting traffic data and utilizing ITS technologies for enhancing mobility and accessibility at freeway and arterial levels. One potential direction to expand the contents of this chapter would be including the safety aspect of transportation into the discussion and to further evaluate the potential relationships between mobility, safety, and accessibility of freeways and arterials.

References

- Abdel-Aty, M., Dilmore, J., Dhindsa, A., 2006. Evaluation of variable speed limits for real-time freeway safety improvement. *Accid. Anal. Prev.* 38, 335–345.
- Ariannezhad, A., Karimpour, A., Wu, Y.-J., 2020. Incorporating mode choices into safety analysis at the macroscopic level. *J. Transp. Eng. Part A: Sys.* 146 (4), 04020022.
- Ariannezhad, A., Razi-Ardakani, H. & Kermanshah, M., 2014. Exploring factors contributing to crash severity of motorcycles at Suburban roads, Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2014. (No. 18-05866).
- Ariannezhad, A., Wu, Y.-J., 2018. Effects of heavy rainfall in different light conditions on crash severity during Arizona’s monsoon season. *J. Transp. Safe. Secur.* 11 (6), 579–594.
- Ariannezhad, A., Wu, Y.-J. & Gofar, V. N. 2018. Real-time crash prediction using data mining techniques.
- Asgharzadeh, M., Gubbala, P.S., Kondyli, A., Schrock, S.D., 2019. Effect of on-ramp demand and flow distribution on capacity at merge bottleneck locations. *Transp. Lett.* 1–9.
- Asgharzadeh, M., Kondyli, A., 2018. Comparison of highway capacity estimation methods. *Transp. Res. Rec.* 2672, 75–84.
- Bagdade, J., Nabors, D., Mcgee, H., Miller, R. & Retting, R. 2012. Speed management: A manual for local rural road owners. No. FHWA-SA-12-027. 2012.
- Chamberlin, R. & Fayyaz, K. 2019. Using ATSPM Data for Traffic Data Analytics, Report No. UT-19.22.
- Chen, J., Shao, J., 2000. Nearest neighbor imputation for survey data. *J. Off. Stat.* 16, 113.
- Christie, B., Hardesty, D., Hatcher, G. & Mercer, M. 2015. Integrated Corridor Management: Implementation Guide and Lessons Learned (Final Report Version 2.0).
- Coifman, B., 2002. Estimating travel times and vehicle trajectories on freeways using dual loop detectors. *Transp. Res. Part A: Policy Prac.* 36, 351–364.
- Cortes, C.E., Lavanya, R., OH, J.-S., Jayakrishnan, R., 2002. General-purpose methodology for estimating link travel time with multiple-point detection of traffic. *Transp. Res. Rec.* 1802, 181–189.
- Duan, Y., LV, Y., Liu, Y.-L., Wang, F.-Y., 2016. An efficient realization of deep learning for traffic data imputation. *Transp. Res. Part C: Emerg. Tech.* 72, 168–181.
- Emam, E.B., Al-Deek, H., 2006. Using real-life dual-loop detector data to develop new methodology for estimating freeway travel time reliability. *Transp. Res.* 1959, 140–150.
- Fehon, K. & O’Brien, P., 2015. Traffic Signal Management Plans—An Objectives-and-Performance-based Approach for Improving the Design Operations and Maintenance of Traffic Signal Systems. United States: Federal Highway Administration.
- FHWA, 1996. Benefits, Intelligent Transportation Infrastructure: Expected and Experienced. *United States Department of Transportation, Washington, DC (Report No. FHWA-JPO-96-008)*.
- FHWA, 2017. Making it There on Time, all the Time. FHWA-HOP-06-070. Office of Operations, Federal Highway Administration
- Gross, F., Jagannathan, R., Hughes, W., 2009. Two low-cost safety concepts for two-way, stop-controlled intersections in rural areas. *Transp. Res. Rec.* 2092, 11–18.
- Grumert, E.F., Tapani, A., MA, X., 2018. Characteristics of variable speed limit systems. *Eur. Transp. Res. Rev.* 10, 21.
- Guessous, Y., Aron, M., Bhouiri, N., Cohen, S., 2014. Estimating travel time distribution under different traffic conditions. *Transp. Res. Proc.* 3, 339–348.
- Guo, F., Rakha, H., Park, S., 2010. Multistate model for travel time reliability. *Transp. Res. Rec.* 2188, 46–54.
- Halkias, M., Leng, T., Sorell, M., Parks, J. & Skabardonis, A. 2017. Arterial Speed Management with Control Measures: the Case of San Francisco, California.
- Hardesty, D. & Hatcher, G., 2019. Integrated Corridor Management (ICM) Program: Major Achievements, Key Findings, and Outlook.
- Hassani, A., Joerger, M., Arana, G. D. & Spenko, M., 2018. LiDAR Data Association Risk Reduction, Using Tight Integration with INS. ION GNSS + , The International Technical Meeting of the Satellite Division of The Institute of Navigation, 2018.

- Hassani, A., Morris, N., Spenko, M. & Joerger, M., 2019. Experimental integrity evaluation of tightly-integrated IMU/LIDAR including return-light intensity data. 32nd International Technical Meeting of the Satellite Division of the Institute of Navigation, ION GNSS+ 2019, 2019. Institute of Navigation, pp. 2637–2658.
- HA, T.-J., Kang, J.-G., Park, J.-J., 2003. The effects of automated speed enforcement systems on traffic-flow characteristics and accidents in Korea. *Institute of Transportation Engineers. ITE J.* 73, 28.
- Ishimaru, J. M. & Hallenbeck, M. E. 1999. Flow evaluation design technical report. Washington State Department of Transportation.
- Kamarianakis, Y., Kanas, A., Prastacos, P., 2005. Modeling traffic volatility dynamics in an urban network. *Transp. Res. Rec.* 1923, 18–27.
- Karimpour, A., Ariannezhad, A., Wu, Y.-J., 2019. Hybrid data-driven approach for truck travel time imputation. *IET Intel. Transp. Sys.* 13, 1518–1524.
- Karimpour, M., Karimpour, A., Kompany, K. & Karimpour, A., 2017. Online Traffic Prediction Using Time Series: A Case study. *Integral Methods in Science and Engineering*, Volume 2, pp. 147–156. Springer, Birkhäuser, Cham.
- Karimpour, A., Kluger, R., Wu, Y.J., 2020. Traffic sensor data-based assessment of speed feedback signs. *J. Transp. Saf. Secur.* 1–24, <https://doi.org/10.1080/19439962.2020.1731038>.
- Klein, L. A., Mills, M. K., & Gibson, D. R., 2006. Traffic detector handbook: Volume I. Turner-Fairbank Highway Research Center.
- Kondyli, A., Hale, D.K., Asgharzadeh, M., Schroeder, B., Jia, A., Bared, J., 2019. Evaluating the operational effect of narrow lanes and shoulders for the highway capacity manual. *Transp. Res. Rec.*, 0361198119849064.
- Leduc, G., 2008. Road traffic data: Collection methods and applications. *Working Papers on Energy. Transp. Clim. Change* 1 (55), 1–55.
- Li, X., Wu, Y.-J., Chiu, Y.-C., 2019. Volume estimation using traffic signal event-based data from video-based sensors. *Transp. Res. Rec.*, 0361198119842120.
- Mahmassani, H. S., Valdes, D. M., Machemehl, R. B., Tassoulas, J. & Williams, J.C., 1998. Integrated arterial and freeway operation control strategies for IVHS advanced traffic management systems. University of Texas at Austin. Center for Transportation Research.
- Mansourkhaki, A., Karimpour, A. & Sadoghi Yazdi, H., 2016. Non-stationary concept of accident prediction. *Proceedings of the Institution of Civil Engineers-Transport*, Thomas Telford Ltd, pp. 140–151.
- Mansourkhaki, A., Karimpour, A., Yazdi, H.S., 2017. Introducing prior knowledge for a hybrid accident prediction model. *KSCE J. Civil Eng.* 21, 1912–1918.
- Min, W., Wynter, L., 2011. Real-time road traffic prediction with spatio-temporal correlations. *Transp. Res. Part C: Emerg. Technol.* 19, 606–616.
- Mirshahi, M., Obenberger, J., Fuhs, C. A., Howard, C. E., Krammes, R. A., Kuhn, B. T., Mayhew, R. M., Moore, M. A., Sahebjam, K. & Stone, C.J., 2007. Active traffic management: the next step in congestion management. United States. Federal Highway Administration.
- Neuman, T.R., 2003. Guidance for implementation of the AASHTO strategic highway safety plan, *Transp. Res. Board.*
- NHTSA, 2014. Speed Management Strategic Initiative In , 028, D. H., (ed.) US Department of Transportation, Washington, DC
- NHTSA, 2017. Traffic safety facts: 2017 data: Pedestrians. *Ann. Emer. Med.* 53, 824.
- NHTSA. 2006. Uniform Guidelines for State Highways Safety Programs: Highway Safety Program Guideline No. 3.
- Ou, J., Yang, S., Wu, Y.-J., AN, C., Xia, J., 2018. Systematic clustering method to identify and characterise spatiotemporal congestion on freeway corridors. *IET Intel. Transp. Sys.* 12, 826–837.
- Papageorgiou, M., Kotsialos, A., 2002. Freeway ramp metering: An overview. *IEEE Trans. Intel. Transp. Sys.* 3, 271–281.
- Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.K., 2020. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accid. Anal. Prev.* 136, 105405.
- PeMS, C. P. M. S. 2017. Available from: <http://pems.dot.ca.gov> [Accessed accessed 6 June 2017].
- Polus, A., 1979. A study of travel time and reliability on arterial routes. *Transportation* 8, 141–151.
- PORTAL, P. O. R. T. A. L. 2017. Available from: <http://portal.its.pdx.edu/>, [Accessed accessed 6 June 2017].
- Ran, B., Tan, H., Feng, J., Liu, Y., Wang, W., 2015. Traffic speed data imputation method based on tensor completion. *Comp. Intel. Neurosci.* 2015, 22.
- Razi-Ardakani, H., Mahmoudzadeh, A., Kermanshah, M., 2018. A Nested Logit analysis of the influence of distraction on types of vehicle crashes. *Eur. Transp. Res. Rev.* 10, 44.
- Santiago-Chaparro, K.R., Chitturi, M., Bill, A., Noyce, D.A., 2012. Spatial effectiveness of speed feedback signs. *Transp. Res. Rec.* 2281, 8–15.
- Schneider, R., 2011. Comparison of turning movement count data collection methods for a signal optimization study. URS Corp. Tech. Rep.
- Semertzidis, T., Dimitropoulos, K., Koutsia, A., Grammalidis, N., 2010. Video sensor network for real-time traffic monitoring and surveillance. *IET Intel. Transp. Sys.* 4, 103–112.
- Sunkari, Srinivasa, 2004. The benefits of retiming traffic signals. *ITE J.*
- Stevanovic, A., Kergaye, C., Stevanovic, J., 2011. Evaluating robustness of signal timings for varying traffic flows. *Transp. Res. Rec.* 2259, 141–150.
- Sullivan, A. & Fadel, G., 2010. Implementing active traffic management strategies in the US. University Transportation Center for Alabama.
- Tang, J., Zhang, G., Wang, Y., Wang, H., Liu, F., 2015. A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation. *Transp. Res. Part C Emerg. Tech.* 51, 29–40.
- TTI 2019. 2019 Annual Urban Mobility Report. College Station, Texas.
- Urbanik, T., Humphreys, D., Smith, B., Levine, S., 2006. Coordinated freeway and arterial operations handbook. United States. Federal Highway Administration.
- Wu, C.-H., HO, J.-M., Lee, D.-T., 2004. Travel-time prediction with support vector regression. *IEEE Trans. Intel. Transp. Sys.* 5, 276–281.
- Wu, Y.-J., Chen, F., LU, C.-T., Yang, S., 2016. Urban traffic flow prediction using a spatio-temporal random effects model. *J. Intel. Transp. Sys.* 20, 282–293.
- Wu, Y.-J., Hallenbeck, M.E., Wang, Y., Watkins, K.E., 2010. Evaluation of arterial and freeway interaction for determining the feasibility of traffic diversion. *J. Transp. Eng.* 137, 509–519.
- Yang, S., Liu, X., Wu, Y.-J., Woelschlag, J., Coffin, S.L., 2015. Can freeway traffic volume information facilitate urban accessibility assessment?: Case study of the city of St. Louis. *J. Transp. Geogr.* 44, 65–75.
- Yang, S., Malik, A., Wu, Y.-J., 2014. Travel time reliability using the Hasofer–Lind–Rackwitz–Fiessler algorithm and Kernel density estimation. *Transp. Res. Rec.* 2442, 85–95.
- Yang, S., Wu, Y.-J., 2016. Mixture models for fitting freeway travel time distributions and measuring travel time reliability. *Transp. Res. Rec.* 2594, 95–106.
- Yang, S., Wu, Y.-J., Yin, Z., Feng, Y., 2016. Estimating freeway travel times using the general motors model. *Transp. Res. Rec.* 2594, 83–94.
- You, J., Fang, S., Zhang, L., Taplin, J., Guo, J., 2018. Enhancing freeway safety through intervening in traffic flow dynamics based on variable speed limit control. *J. Adv. Transp.*, 2018.
- Yuan, K., Knoop, V.L., Leclercq, L., Hoogendoorn, S.P., 2017. Capacity drop: a comparison between stop-and-go wave and standing queue at lane-drop bottleneck. *Transportmet. B Transp. Dyn.* 5, 145–158.
- Zhang, G., Avery, R.P., Wang, Y., 2007. Video-based vehicle detection and classification system for real-time traffic data collection using uncalibrated video cameras. *Transp. Res. Rec.* 1993, 138–147.

Arterial Road Management

Jiaqi Ma*, Yi Guo[†], Adekunle Adebisi[†], *University of California, Los Angeles, CA, United States; [†]University of Cincinnati, Cincinnati, OH, United States

© 2021 Elsevier Ltd. All rights reserved.

Concepts of Arterial Road Management	169
Elements of Signal Coordination	170
Time-Space Diagram	170
Master Clock, Local Clock, and Green Wave	170
Signal Offset	171
Bandwidth	171
Signal Coordination Complexities	171
Signal Coordination Frameworks and Tools	172
Coordinating Pretimed Signal Control Frameworks	173
MAXBAND	173
PASSER V	173
Coordinating Responsive/Adaptive Control Frameworks	174
SCATS	174
UTOPIA	174
SCOOT	174
RHODES	174
ALLONS-D	175
OPAC	175
Signal Timing and Coordination Software Tools	175
TRANSYT-7F	175
SYNCHRO	175
Integrating Emerging Connected Vehicle Technologies	175
Future Trends	176
References	176

Concepts of Arterial Road Management

Arterial road management techniques are implemented by transportation professionals for coordinating the traffic to improve safety and operational performance. Classical management techniques on arterial roads, including signalized intersections, have been in use and proven effective for decades. At signalized intersections, traffic signals are now more intelligent and responsive to traffic demands through the use of systems such as loop detectors, video imaging, and tube sensors for signal actuation (Hoel and Garber, 2007). These intelligent systems use detectors to sense the presence of vehicles on a typical roadway approach, send this detection information to the road-side controllers, and initiate an optimized plan that minimizes specific performance measure(s) at the intersection. The concept involves the use of elements such as signal phase max out, minimum green time, maximum green time, and gap-out in designing signal timing plans (Hoel and Garber, 2007). This intelligent type of signalization is usually applied to isolated intersections and is often effective, especially under low traffic demand. Another concept in this regard is the adaptive traffic signal control (ATSC), which works based on the dynamic demand at intersections. Traffic signals can work as an offline approach, where the optimized signal timing is designed based on pre-collected traffic demand information (time-of-day plans), or as an online system, where the system is designed to actively collect vehicle arrival information and optimize the signal phasing and timing in real time. While these management techniques are effective for isolated intersections, they do not effectively consider closely spaced intersections, which are usually the case on urban streets. Under this condition, the coordination of successive signals is usually implemented (Fred Mannering et al., 2007).

Generally, the concept of *signal coordination* refers to the time synchronization of closely-spaced successive signals along a corridor such that the platoon of vehicles discharged from an upstream intersection arrives at the green phase of the downstream intersections, thereby vehicles proceeding through the coordinated corridor without stopping. In essence, by strategically setting the time difference between the start of the green phase for each successive intersection (*termed signal offset*), it is possible to create a continuum of green phases (*termed green wave*). This reduces the signal control delay by allowing more vehicles to arrive at each intersection during the green phases and fewer vehicles during the red phases, thereby consequently improving the corridor performance. To achieve this, controllers at each intersection in a corridor are connected with a master controller acting as a reference for the other local controllers using wired (fiber optic cables, coaxial cables) or wireless systems (radio communications). A

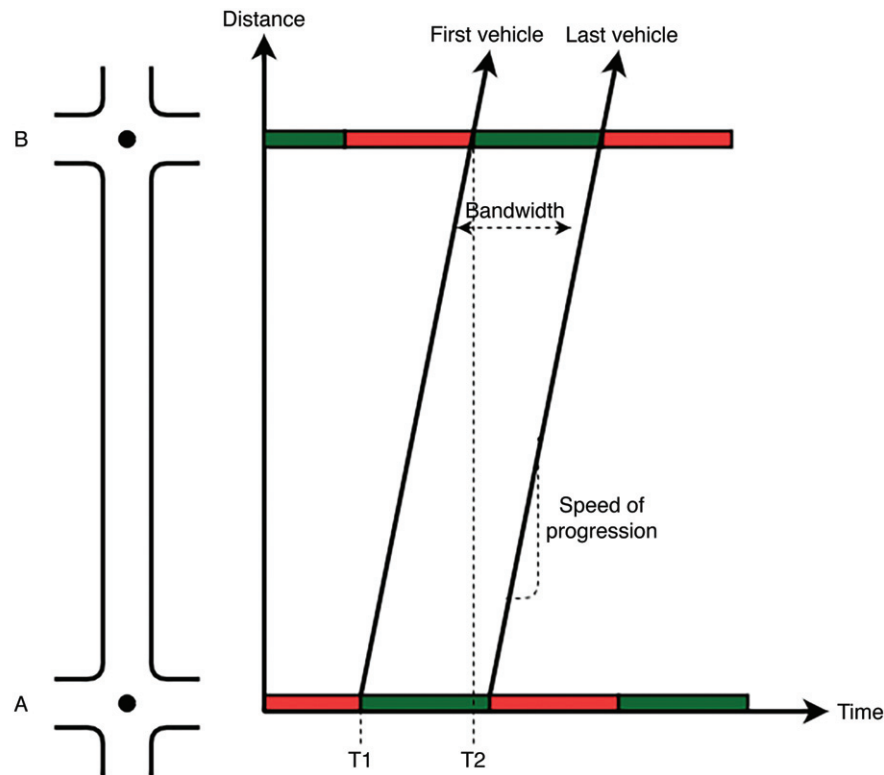


Figure 1 Signal coordination on a time-space diagram.

common method for time synchronization across intersections is using the master and local clocks, as will be described later (Urbanik et al., 2015). In the following sections, the elements of signal coordination, including the Time-space Diagrams, Signal Offsets, Bandwidth concepts, and Signal coordination complexities, are discussed. Thereafter, an outlook into various signal coordination frameworks and algorithms, including the currently implemented systems and emerging systems in terms of connected vehicle technologies, are provided. Finally, future trends and present limitations in the area of signal optimization and arterial traffic management are briefly discussed.

Elements of Signal Coordination

Time-Space Diagram

The Time-space diagram (TSD) shows the progression of signal indications and motion of vehicles along a corridor as a function of time (Hoel and Garber, 2007). It describes the technique for tracking the position of an individual vehicle or a platoon of vehicles over time, along with a one-dimensional specified distance. TSDs use the concept of velocity-time relationships to effectively evaluate the position of a subject vehicle within a segment, thereby allowing to coordinate the movement of vehicles with the phasing of successive signals. As a result, TSDs can be used to assess the relationship between vehicle movement, signal timing plans, and intersection spacing, and make modifications when necessary (Roess et al., 2011). They can also be used to assess other performance measures such as vehicle stops, green time arrivals, and queue lengths. Figure 1 shows a typical TSD (drawn to scale) connecting two successive intersections A and B. The time sequence is shown on the horizontal axis while the intersection distances are shown on the vertical axis. The line indicates the trajectory of a vehicle at each time t in the Northbound (NB) one-way street considered here. At A, the signal turns green at time $t = T1$. The subject vehicle moves a distance d_1 to intersection B, where the signal turns green at time $t = T2$. This progression continues for every possible intersection downstream of B. This phenomenon whereby vehicle platoons meet a green indication on arriving at each intersection is termed the *green wave*, and the time difference between each green indication is the *signal offset* (or simply *offset*).

Master Clock, Local Clock, and Green Wave

In a typically coordinated corridor, there are two types of clocks operating the system: the master clock (i.e., the controller, central system) and the local (individual intersection) clock. The master clock is the timing mechanism embedded within the controller logic relative to which the other downstream intersection local controllers are referenced in time, to generate the signal offset. The

clocks are expected to operate similar timeframes to maintain accuracy (Urbanik et al., 2015). Considering the TSD provided in Figure 1 and assuming the coordination starts at intersection A, the master clock will be located at A while the downstream local controllers will have individual local clocks. Each clock has a reference timestamp that vehicles are expected to get to the corresponding intersection. This synchronization of the master and the local clocks establishes the coordinated operations for the corridor. A closely related concept is the green wave. The green wave describes the continuum of green phases within a corridor (Hoel and Garber, 2007). Particularly for one-way streets, as shown in Figure 1, this indicates the start of the green phase at each downstream intersection on the arrival of vehicle platoons. Corridors with signal coordination have the same cycle lengths at each intersection to enable time synchronization for the start of green at each point, although the green time can vary at different intersections based on other factors. However, there are other slightly different signal progression concepts for two-way streets (Roess et al., 2011), which are described later.

Signal Offset

The signal offset is a major determining factor in the efficiency of signal coordination along a corridor. For every signal coordination design, the objective is to obtain an *ideal offset* for the corridor (Roess et al., 2011). The *ideal offset* is defined as the offset at which the first vehicle in the platoon reaches the downstream intersection exactly when the signal turns green. It defines the time relationship between the master clock and the coordinated downstream signals specified by the local clocks. Determining the offsets for a corridor usually considers factors such as the actual or desired travel speed of vehicles between successive signalized intersections, the distance, and the traffic demand on the corridor. The *ideal offset* ensures that vehicle platoons leaving an upstream intersection arrives at the downstream intersection towards the green phase or after minor street queues have been dissipated (Urbanik et al., 2015). To determine the *ideal offset* for each segment in a corridor, a combination of field review and TSD is recommended (Roess et al., 2011). For example, in the TSD provided in Figure 1, assuming no queue on the downstream intersection, the *ideal offset* between A and B can be calculated as the ratio of the distance (d_1) between the two intersections and the vehicle speed (S). However, this no-queue assumption is usually not the case as vehicles from nearby driveways or unserved vehicles from the last cycle may be queued at the downstream intersection. Under this condition, the *ideal offset* is adjusted for queues using Equation (1), where Q is the number of vehicles queued per lane, h is the discharge headway of queued vehicles, and l_1 is the start-up lost time measured for the downstream intersection (B) (Roess et al., 2011).

$$O_{adj} = \frac{L}{S} - (Qh + l_1) \quad (1)$$

Bandwidth

Closely related to Signal offset, Bandwidth (BW) refers to the time difference, measured in seconds, between when the first and the last vehicles in a platoon traverse the system without any of the vehicles stopping at any of the intersections (Hoel and Garber, 2007). The visualization of bandwidth on TSDs is effective for observing the behavior of signal coordination and assessing possible measures and their effectiveness in improving the system performance. While the TSD provided in Figure 1 shows the bandwidth for one-way street NB direction, it should be noted that if the same corridor were a two-way street, the bandwidth for the opposing direction could be different, and as additional intersections are added to the system, it becomes more complex to obtain a smooth progression like the one shown (Hoel and Garber, 2007). This, with other factors such as a change in traffic demand, or emergency vehicle pre-emption can affect the corridor Bandwidth. There are two important elements used to measure the performance of a Bandwidth: the bandwidth efficiency and the bandwidth capacity (Hoel and Garber, 2007). Bandwidth efficiency signifies how efficient the coordination is operating. It is measured as the ratio of bandwidth to the cycle length expressed in percentage, as shown in Equation (2). For conventional traffic, a BW efficiency of around 50% is considered reasonable. On the other hand, the BW capacity defines the number of vehicles per hour that can traverse a series of signals without stopping. Equation (3) gives the procedure for estimating BW capacity (Roess et al., 2011). An important parameter required for this is the BW itself, which can be estimated directly from the TSD as shown in Figure 1.

$$BW_{eff} = \left(\frac{BW}{C} \right) \times 100 \quad (2)$$

$$BW_{capacity} = \frac{3600 \times BW \times N}{C \times h} \quad (3)$$

where, BW = Bandwidth (s), N = number of through lanes in subject direction, C = cycle length (s), and h = saturation headway (s).

Signal Coordination Complexities

The TSD shown in Figure 1 depicts the signal progression on a typical one-way street. However, there are complexities and slight differences that occur with two-way streets. On one-way streets, signal progression can be described using four terms: simple progression, forward progression, flexible progression, and reverse progression (Roess et al., 2011). Simple progression occurs

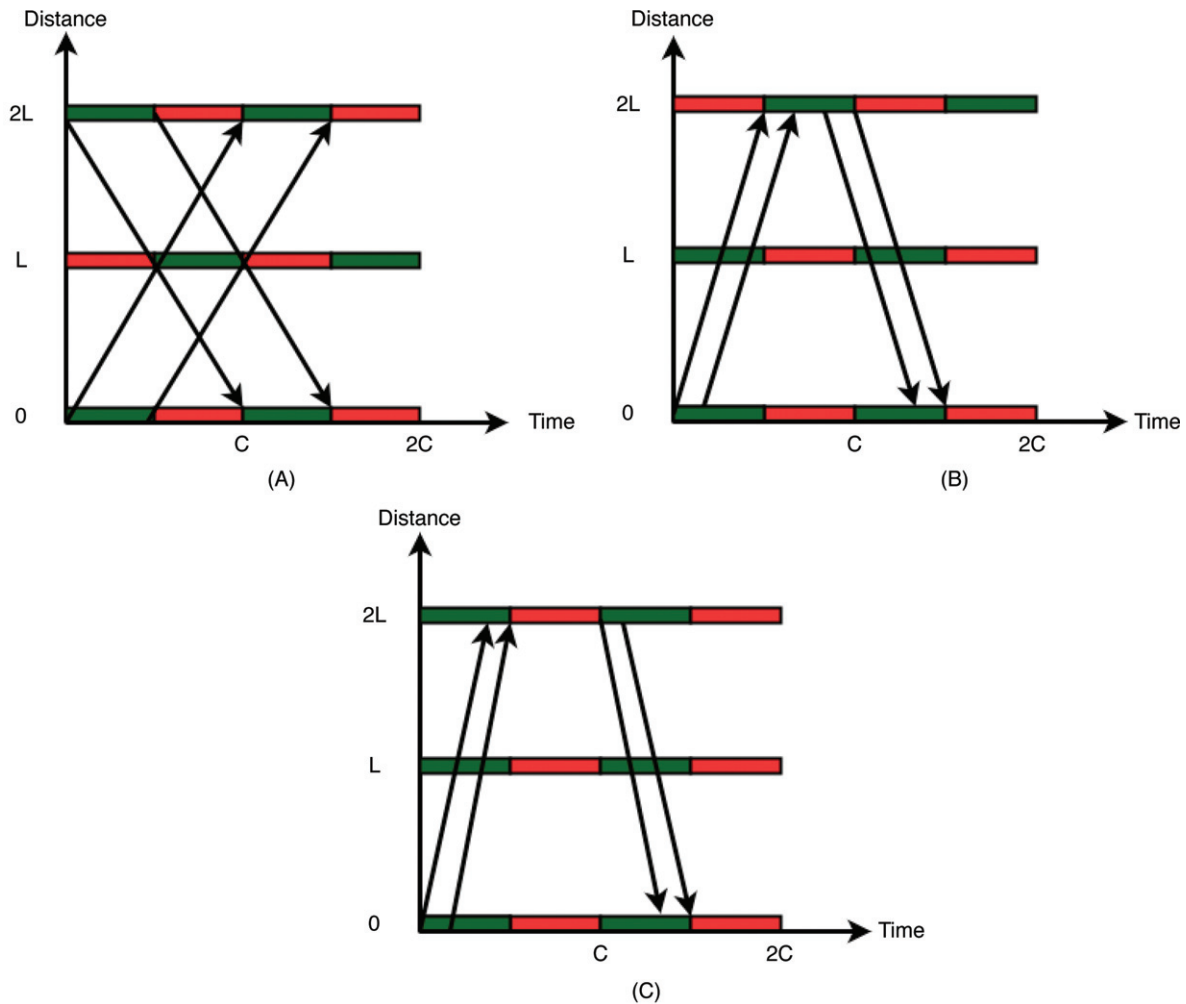


Figure 2 Two-way street TSD representations: (A) Alternating, (B) Double-alternating, and (C) Simultaneous progressions.

when vehicles arrive at the beginning of the green phase on the downstream intersection. As such, the green wave proceeds in the forward direction with the platoon of vehicles, and it is also often called the forward progression (Roess et al., 2011). In cases whereby the simple progression on a corridor is revised at some time in the day, for example, to optimize for varying traffic demand or direction of flow, the scheme is referred to as a flexible progression. Finally, the reverse progression occurs when the downstream signal must turn green before the upstream signal due to left-over queues at the downstream intersection being too large and needs to be dissipated first before the arrival of the upstream platoon (Roess et al., 2011). The cases of two-way streets are slightly different, and complex based on traffic and geometric characteristics of both directions of the streets. Terms such as Alternate progression, Double-alternating progression, and Simultaneous progression are relevant in this case (Roess et al., 2011). The Alternate progression (Figure 2A) describes a corridor condition where the *ideal offset* is equivalent to half the cycle length (C). In such a scenario, an observer at the first intersection looking downstream sees alternating signal phases red, green, red, green, respectively, at the following intersections. The Double-alternating progression (Figure 2B) describes when the ideal offset is equivalent to quarter the cycle length. In such a scenario, an observer at the first intersection looking downstream sees signal alternate in pairs red, red, green, green, red, red, respectively at the following intersections. In Simultaneous progression (Figure 2C), all the signals in the coordinated corridor turn green at the same time. This is mostly used in very closely spaced intersections or high-speed corridors (Roess et al., 2011).

Signal Coordination Frameworks and Tools

Some representative signal coordination frameworks and tools are introduced in this section. These frameworks and tools use optimization-related approaches to find the combination of signal design parameters (e.g., offsets, cycle lengths, and phase sequence) that minimize performance measures such as vehicle delays and stops along the arterial while satisfying the necessary

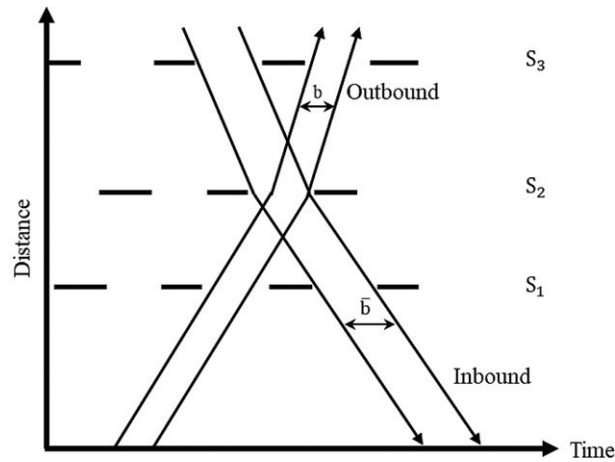


Figure 3 Inbound and outbound bandwidth.

constraints, and ultimately improving the overall arterial performance. In this section, these frameworks and tools are classified into four parts: coordinating pre-timed signal control, coordinating responsive/adaptive control, signal timing and coordination software tools, and emerging concepts that integrate connected vehicle technologies. Each of these is discussed based on the common frameworks and tools used in practice. It is worth noting that the term “framework” is a general concept in this section for the sake of brevity, and some frameworks are originally defined as “hierarchical algorithm” or “control scheme.” Specific definitions are introduced correspondingly in the following subsections.

Coordinating Pretimed Signal Control Frameworks

MAXBAND

MAXBAND (MAXimum BANDwidth) was first proposed in 1966 (Little, 1966), and then in 1981 with a FORTRAN computer program (Little et al., 1981). It considers both the two-way corridor and network signal coordination with n signals $S_1, S_2, \dots, S_n$. For signalized corridor, MAXBAND designs different offsets for each signal to maximize the number of nonstop vehicles while crossing the corridor. Splits of corridor signals and speed ranges of each road segment are regarded as known inputs of the model, and the splits can be replaced with traffic flows and capacities. MAXBAND attempts to maximize the inbound bandwidth b and outbound bandwidth b with the given information, as shown in Figure 3. In the MAXBAND model, the problem is formulated as a mixed-integer-linear-programming (MILP) problem, and a branch-and-bound algorithm is employed to reduce the computational burden and improve efficiency. For network-level coordination, Little (1966) appends cycle constraints that connect the streets to expand the basic MAXBAND.

MAXBAND-86 (Chang et al., 1988), an extension of MAXBAND, was initiated by FHWA (Federal Highway Administration) to adapt the multi-arterial closed networks. However, both MAXBAND and MAXBAND-86 have the fundamental limitation that they do not consider the actual volume of each segment but the average directional volumes along the arterial. In other words, the model assumes almost all vehicles cross the arterial as platoons with rare turn-in and turn-out, not considering the possibility that the traffic volume may vary remarkably between different segments. MULTIBAND (MULTIple BANDwidth) (Gartner et al., 1991) is a significant extension of MAXBAND, which expands the MAXBAND to adapt varying traffic volumes by generating a variable-bandwidth progression. It also incorporates left-turn phase sequence optimization, cycle length optimization, speed limit optimization, and initial queue clearance time to improve flexibility. The heuristic method is included to accelerate computation. Stamatiadis and Gartner (1996) expanded MULTIBAND and developed MULTIBAND-96 to control of closed-loop networks.

PASSER V

PASSER (Progression Analysis and Signal System Evaluation Routine) was first developed by the Texas Transportation Institute (TTI) in 1973 (Messer et al., 1973). It was designed as an online arterial bandwidth optimization computer program that considers the multi-phase sequence. The PASSER II (Messer et al., 1974) was developed by TTI in 1974, and it is the widely known and long-lasting version of PASSER, while PASSER III (Messer et al., 1977) mainly focuses on signal timing optimization of diamond interchanges. PASSER V (Chaudhary et al., 2002; Chaudhary and Chu, 2011) is the latest version in the PASSER series of computer programs for the signalized corridor, network, and diamond interchange optimization. PASSER V uses three main optimization algorithms, including genetic algorithm, interference minimization, and exhaustive search, and divides the optimization process into multiple stages. The first stage determines green splits of each signal. The second stage decides offsets and cycle times for maximizing the bandwidth. The third stage adjusts variables to further maximize progression or minimize total delay.

Coordinating Responsive/Adaptive Control Frameworks

SCATS

SCATS (Sydney Coordinated Area Traffic System) was developed in Australia by RTA¹ (Roads and Traffic Authority) in the early 1980s (Sims and Dobinson, 1980). It is a real-time hierarchical adaptive traffic signal control system with three levels: local controller data collection from detectors, data processing, and then signal operations decision. The controllers control up to 200 signals to collect and analyze detector information and implement a real-time operation. With SCATS, the control center usually does not implement, or influence signal operations directly but only store essential traffic data and monitor the operation of the entire system. SCATS adjusts cycle length, splits, and offsets to minimize delay and stops or maximize throughput. At each cycle, SCATS collects traffic data from detectors at first, and the degree of saturation (DS) and link flow (LF) are calculated. Note that the DS concept is the ratio of effectively used green time to the total green time in SCATS, which needs to be distinguished from the widely used concept of degree of saturation defined by Webster (1958). The updated cycle lengths, phase splits, and offsets are calculated based on DSs and LFs. Since SCATS can be installed to manage numerous signals, it divides the entire traffic system into several small sub-systems of signals. Adjacent subsystems can link together to form some larger systems or even one large system based on a voting mechanism. If the cycle time requirement of the subject subsystem is close to the requirement of the adjacent sub-system (the difference is less than 10 seconds), a positive vote is made (Stevanovic et al., 2009); otherwise a negative vote is made. Votes are cumulated over time, and the adjacent subsystem can be linked if the number of positive votes exceeds four (Shelby, 2001). The system and subsystems can also be decoupled based on the same voting mechanism, if the number of positive votes is lower than zero (Shelby, 2001). This system design and corresponding mechanism provide more flexibility to the signal control system.

UTOPIA

Developed by Italian researchers (Donati et al., 1984; Mauro and Di Taranto, 1990), UTOPIA (Urban Traffic Optimization by Integrated Automation) is a decentralized hierarchical traffic control algorithm involving rolling horizon. Since the process of optimization is implemented in two layers separately, the area level control mainly focuses on the signal coordination, and the intersection level control consists of single intersection modules. The single intersection module is also referred to as SPOT. For every single intersection, SPOT collects real-time traffic data and optimizes signal plans by using a branch and bound optimization. There are two key components of SPOT; observer and controller. The observer updates the traffic state estimation of the intersection and communicates with each other. Using this information, the upstream arrival estimation can be utilized to support queue estimation. The controller determines and implements the signal operation in a rolling horizon manner. Originally the roll period is designed as 6 seconds, and the planning horizon is 120 seconds. However, the parameter set varies from different implementations. The area-level control (consisting of more than one intersection) also consists of two parts of observer and controller. The observer obtains entire area data and then predicts demand levels with a 3-minute interval. The controller optimizes traffic flow with a 30-minute planning horizon while acting on fictitious decision variables, saturation flow, and average speed. However, the details of UTOPIA are vague in the published literature.

SCOOT

SCOOT (Split Cycle Offset Optimization Technique) was developed in the United Kingdom by the TRL (Transport Research Laboratory) in the early 1980s (Hunt et al., 1981). It is a centralized adaptive traffic signal control system and has been upgraded several times (Bing and Carter, 1995; Bretherton et al., 1998; Bretherton, 2007). The latest version of SCOOT is SCOOT MMX (Bretherton et al., 2011), which adapts the TRANSYT traffic model to online applications and is usually considered as the online version of TRANSYT. SCOOT can optimize cycle lengths, splits, and offsets to minimize delays or stops. At each cycle, it collects traffic data from detectors, calculates the degree of saturation, and predicts arrivals and queues based on in-built models. Then splits and offsets optimized based on the traffic information correspondingly. For cycle length optimization, SCOOT considers all links' degrees of saturation within the given region. It is worth noting that SCOOT tends to gradually update the signal plan to avoid overacting, aimed at reducing the side-effect on traffic while adjusting signal plans.

RHODES

RHODES (Real-time Hierarchical Optimizing Distributed Effective System) was developed by researchers from the University of Arizona (Head et al., 1992; Mirchandani and Head, 2001). It is a hierarchical traffic adaptive signal control system that employs dynamic programming, decision tree, and rolling horizon. RHODES consists of two levels of control: intersection level control and network-level control. The network-level control decomposes the control problem into several sub-problems, which intersection level controllers will finally solve those sub-problems correspondingly. RHODES collects real-time traffic data to predict arrivals and queues to support signal optimization. At the network level, the REALBAND (Dell'Olmo and Mirchandani, 1995) is utilized to decide the constraints for network signal coordination. Based on the traffic information, RHODES predicts platoon arrivals within the given planning horizon and forms a decision tree to resolve conflicts among different phases. Therefore, a set of constraints will be generated and sent to intersection level controllers. At the intersection level, the Controlled Optimization of Phases (Sen and Head, 1997), a dynamic programming-based algorithm, is used to minimize accumulative delay based on specified constraints.

¹The RTA was abolished in 2011 and now is known as NSW Roads and Maritime Services.

ALLONS-D

ALLONS-D (Adaptive Limited Look-Ahead Optimization of Traffic Signals-Decentralized) was developed by researchers from the University of Michigan (Porche et al., 1996; Porche 1998; Porche and Lafortune, 1999). It is a decentralized real-time hierarchical adaptive signal control scheme, which employs dynamic programming, decision tree, and rolling horizon. The original ALLONS-D was expanded to adapt the network optimization with a two-layer scheme (Porche and Lafortune, 1997). Instead of using queue estimation and arrival prediction, ALLONS-D uses the real-time state of the intersection (named “snapshot”) derived from complex detector technologies, such as AUTOSCOPE (Michalopoulos, 1991). ALLONS-D uses a depth-first branch and bound algorithm to search the optimal control in the decision tree at the signal level. To coordinate signals, two different approaches are used: the hierarchical approach, and the iterative approach. Under the hierarchical scheme, it generally divides the accumulative delay of each intersection into two parts based on directions. The higher-level controller (e.g., arterial level controller) keeps updating and sending weights of each phase to local controllers to coordinate the signals to minimize arterial or network level delay. Under the iterative scheme, it incorporates a simulation model and attempts to obtain a set of coordinated signal policies by iteratively interacting with the simulation model. The effects on adjacent intersections are recorded during the simulation period, and the simulation will be terminated if the pre-defined equilibrium criteria are met. The iterative scheme is also regarded as ALLONS-I (ALLONS-iterative).

OPAC

OPAC (Optimized Policies for Adaptive Control) was initially developed by Gartner. The initial version provides three different approaches to solve the signal control problem, including dynamic programming (OPAC I), sequential optimization (OPAC II), and rolling horizon (ROPAC, Rolling horizon OPAC, or OPAC III). In 1996, the first coordination version of OPAC was proposed (Gartner et al., 1996; Gartner et al., 2001), which is called VFC-OPAC (Virtual-Fixed-Cycle OPAC, or OPAC IV). Andrews et al. (1997) also implemented OPAC-RT (Real-time OPAC) version 3.0 on Route 16 in New Jersey to support signal coordination between adjacent intersections. Gartner et al. (2001) further expanded the OPAC to adapt network signal control by implementing OPAC in RT-TRACS (Real-time Traffic Adaptive Control System). OPAC IV is a real-time hierarchical demand-responsive adaptive signal control scheme. It consists of three layers, including the local control layer, coordination layer, and synchronization layer. The local control layer implements signal operations in a rolling horizon manner and calculates optimal switching sequences while considering the VFC constraint derived from the synchronization layer. The coordination layer optimizes the offsets of each intersection within the mini-network every cycle. The synchronization layer is responsible for calculating network-wide VFC every few minutes, which can be pre-defined by users.

Signal Timing and Coordination Software Tools**TRANSYT-7F**

Developed by Robertson in 1969 (Robertson, 1969), TRANSYT (Traffic Network Study Tool) is one of the most common series of signal optimization tools. It is a computer program of traffic signal system simulation and optimization that optimizes cycle lengths, phasing sequence, splits, and offsets. All signals in the network generally use the same cycle length, except for allowing specified intersections to double-cycle (Shelby, 2001). To coordinate signals in the network, TRANSYT tunes offsets and splits by optimizing given objectives. In early versions, TRANSYT provides two major optimization objectives of delay and number of stops. In the latest version, TRANSYT-7F, it supports customized optimization objective function with different “performance indexes” (PIs) in terms of delay, progression, number of stops, fuel consumption, queue length, and throughput. TRANSYT-7F employs a genetic algorithm as the primary optimization algorithm. It aims to find the globally optimal solution with the given inputs while also incorporating multi-period optimization and direct CORSIM (corridor simulation) optimization (i.e., microscopic optimization). It provides two different simulation options, including link-wise simulation and stepwise simulation, for various purposes. The link-wise simulation is convenient for undersaturated traffic conditions, while stepwise simulation aims to handle oversaturated situations.

SYNCHRO

Synchro was developed by Trafficware, which the latest version is Synchro Studio 11. It is a delay-based software for signal timing analysis and optimization for arterials and networks. Synchro provides cycle length, phasing sequence, split, and offset optimizations. The main objective of signal optimization in Synchro is to minimize delays and stops and increase the efficiency of the traffic network. For signalized corridor optimization, Synchro optimizes splits and cycle lengths of each intersection first and then coordinates cycle lengths at the arterial level. Finally, it decides optimal offsets and phasing sequence and generates optimal signal plans. Those signal plans can be used as inputs of other traffic simulation software to further analyze the performances.

Integrating Emerging Connected Vehicle Technologies

The emerging connected vehicle (CV) technologies enable an early and periodical detection of approaching vehicles within the vehicle to infrastructure (V2I) communication range. Priemer and Bernhard (2009) proposed a decentralized adaptive traffic signal control strategy for urban networks using V2I communication. Similar to traditional signal control approaches, this strategy is also phase-based in a rolling horizon manner. Instead of collecting data from detectors, this strategy collects traffic information from vehicles equipped with communication devices every 5 seconds. Then, it calculates or predicts the queue length for the next

20 seconds, which depends on the penetration rate of equipped vehicles: if the penetration rate is 100%, the queue length is directly calculated; otherwise, a queue length estimation algorithm (Priemer and Bernhard, 2008) is utilized to estimate the queue length. It employs dynamic programming and completes enumeration in the optimization process to determine the optimal phasing sequence, in which the main objective is to reduce the total queue length. To coordinate signals, a priority strategy is implemented for main streams to reduce the number of stops. Each vehicle platoon (under high penetration) or single vehicle (under low or medium penetration) in the coordinated direction will be assigned a weighted factor q , and the value of q will increase according to the elapsed waiting time. Therefore, the green phase can be extended or early switched for coordinated direction.

He et al. (2012) proposed a platoon-based multi-modal traffic signal control framework PAMSCOD (Platoon-based Arterial Multi-modal Signal Control with Online Data), for signals within the traffic network under V2I environment. PAMSCOD considers two traffic modes, including transit buses and passenger vehicles, and assumes that signal controllers can obtain traffic data via V2I communication. A platoon recognition algorithm is also developed to identify pseudo-platoons with collected traffic data. PAMSCOD coordinates signals by considering the progression through downstream intersections. PAMSCOD uses mixed-integer linear programming to optimize signal plans based on traffic conditions, signal status, platoon information, and priority requests of buses, and the objective function is to minimize total weighted delay and total green rest time.

Islam and Hajbabaie (2017) proposed a decentralized coordinated signal timing optimization method under the V2I environment. Signal controllers can communicate with connected vehicles and collect traffic data. The main objective is to maximize the throughput and minimize queue length while considering traffic flow and adjacent intersections' decisions. Mixed-integer linear programming is introduced to solve the optimization problem. However, this study assumes the market penetration is 100% and has not proposed an algorithm to estimate unequipped vehicles' states. Therefore, the low market penetration scenarios may not proceed well with the proposed method.

Future Trends

Emerging advanced vehicular technologies, such as vehicle connectivity and automation, have great potential in managing arterial road management and benefiting traffic performance. In the past decade, the connected and automated vehicle (CAV) technologies have been developed and tested, proving the benefit brought by allowing signal controllers to obtain a high-resolution vehicle data and therefore enhancing the capability of perception of the traffic environment (Guo et al., 2019a).

Also, vehicle trajectories can be precisely controlled in addition to infrastructure-side control. A concept of signal-free intersections was proposed by Dresner and Stone (2008) for scenarios with 100% CAV market penetration in the traffic stream. Each CAV can plan its space-time trajectory and coordinate with others to resolve conflicts at the approaches and within the intersection, and therefore physical traffic signals are no longer necessary. Yu et al. (2019) proposed a corridor-level cooperative trajectory optimization with all-CAV traffic. Trajectory coordination and lane change behavior are explicitly considered at the microscopic level. A mixed-integer linear programming model was formulated to optimize trajectories along the corridor and, therefore, overall system performance.

However, it may take decades to reach the 100% CAV market penetration. Prior to that, signal optimization still needs to be considered carefully for unequipped vehicles, especially in the near-term future. For example, Feng et al. (2018) integrated signal control and CAV trajectory optimization and formulated a two-stage optimization problem. Dynamic programming and optimal control theory are applied to optimize signal timings and CAV trajectories correspondingly. Guo et al. (2019b) proposed a joint optimization framework that simultaneously optimizes CAV trajectories and signal timings under different CAV market penetrations. This framework forms a dynamic programming optimization problem and incorporates trajectory planning with piecewise polynomials (TP3) (Zhou et al., 2017; Ma et al., 2017) as a subroutine to improve the efficiency. Integrating signal control and vehicle trajectory planning and optimization under the CAV environment is promising and has great potential to improve traffic system performance. However, many questions need to be solved, such as how to efficiently optimize signal timings and CAV trajectories simultaneously in real time and how to extend joint optimization algorithms to arterials and networks. Communication disruptions, malfunctions of control and communication equipment, and cybersecurity issues also need to be further considered in the implementation.

References

- Christina, M.A., Manzur Elahi, S., James, E., Clark, 1997. Evaluation of New Jersey route 18 OPAC/MIST traffic-control system. *Transp. Res. Rec.* 1603 (1), 150–155.
- Bing, Brian, Alan Carter. "SCOOT: The world's foremost adaptive traffic control system." *TRAFFIC TECHNOLOGY INTERNATIONAL '95* (1995).
- Bretherton, David. *SCOOT MC3 and current developments*. No. 07-1543. 2007.
- Bretherton, R.D., and G. T. Bowen., 1990. Recent enhancements to SCOOT-SCOOT version 2.4." In *Third International Conference on Road Traffic Control*, pp. 95-98. IET.
- Bretherton, R.D., K. Wood, and G. T. Bowen., 1998."SCOOT version 4." pp. 104-108.
- Bretherton, David, Mark Bodger, Jason Robinson, Ian Skeoch, Helen Gibson, and Gavin Jackman. "SCOOT MMX (SCOOT Multi Modal 2010)." In *18th World Congress on Intelligent Transport Systems*, pp. 16-20. 2011.
- Chang Edmond, C.P., Stephen L. Cohen, Charles, Liu, Nadeem A, Chaudhary, Carroll, Messer, 1988. MAXBAND-86: Program for optimizing left-turn phase sequence in multiarterial closed networks. *Transp. Res. Rec.* 1181 .
- Chaudhary, Nadeem A., and Chi-Leung Chu. Implementation Report on PASSER V-07 Training Workshops. No. FHWA/TX-11/5-5424-01-1. Texas Transportation Institute, Texas A & M University System, 2011.

- Chaudhary, N.A., Kovvali, V.G., and Mahabubul Alam, S.M., 2002. Guidelines for selecting signal timing software. No. FHWA/TX-03/0-4020-P2., Texas Transportation Institute, Texas A & M University System.
- Dell'Omo, Paolo, Mirchandani S Pitu, 1995. REALBAND: An approach for real-time coordination of traffic flows on networks. *Transp. Res. Rec.* 1494, 106–116.
- Donati, F., Vito Mauro, G., Roncolini, Vallauri, M., 1984. A hierarchical-decentralized traffic light control system. The first realization: "Progetto Torino. IFAC Proc. 17 (2), 2853–2858.
- Dressner, Kurt, Stone, Peter, 2008. A multiagent approach to autonomous intersection management. *J. Artif. Intel. Res.* 31, 591–656.
- Feng, Yiheng, Chunhui, Yu, Henry X, Liu, 2018. Spatiotemporal intersection control in a connected and automated vehicle environment. *Transp. Res. Part C: Emer. Tech.* 89, 364–383.
- Gartner, Nathan, H., 1983. A demand-responsive strategy for traffic signal control. *Transp. Res. Rec.* 906, 75–81.
- Gartner Nathan, H., Susan, F. Assman, Fernando, Lasaga, Dennis, L. Hou, 1991. A multi-band approach to arterial traffic signal optimization. *Transp. Res. Part B: Method.* 25 (1), 55–74.
- Gartner, N.H., Ch, Stamatidis, Zhou, C.S., 1996. Outline of the VFC-OPAC Model. *Report prepared for RT-TRACS Project.* UMass-Lowell.
- Gartner, Nathan H., Farhad J. Pooran, and Christina M. Andrews., 2001. Implementation of the OPAC adaptive control strategy in a traffic signal network." In *ITSC 2001. 2001 IEEE Intelligent Transportation Systems. Proceedings* (Cat. No. 01TH8585), pp. 195–200. IEEE.
- Guo, Qiangqiang, Li, Li, Xuegang, Jeff Ban, 2019a. Urban traffic signal control with connected and automated vehicles: A survey. *Transp. Res. Part C: Emer. Tech.* 101, 313–334.
- Guo, Yi, Jiaqi, Ma, Chenfeng, Xiong, Xiaopeng, Li, Fang, Zhou, Wei, Hao., 2019b. Joint optimization of vehicle trajectories and intersection controllers with connected automated vehicles: Combined dynamic programming and shooting heuristic approach. *Transp. Res. Part C: Emer. Tech.* 98, 54–72.
- He Qing, K., Larry, Head, Jun, Ding, 2012. PAMSCOD: Platoon-based arterial multi-modal signal control with online data. *Transp. Res. Part C: Emer. Technol.* 20 (1), 164–184.
- Head, K.L., Mirchandani, P.B., and Sheppard, D., 1992. *Hierarchical framework for real-time traffic control.* No. 1360.
- Hoel, Lester A., and Nicholas J. Garber., 2007. *Transportation infrastructure engineering: A multi-modal integration.* Cengage Learning.
- Hunt, P.B., D.I. Robertson, R.D. Bretherton, and R. I. Winton. *SCOOT-a traffic responsive method of coordinating signals.* No. LR 1014 Monograph.
- Islam, Al S.M.A. Bin, Hajbabaie, Ali, 2017. Distributed coordinated signal timing optimization in connected transportation networks. *Transp. Res. Part C: Emer. Tech.* 80, 272–285.
- Little John, D.C., 1966. The synchronization of traffic signals by mixed-integer linear programming. *Operat. Res.* 14 (4), 568–594.
- Little, John DC, Mark D. Kelson, and Nathan H. Gartner., 1981. MAXBAND: A versatile program for setting signals on arteries and triangular networks.
- Ma, Jiaqi, Xiaopeng, Li, Fang, Zhou, Jia, Hu, B. Brian, Park., 2017. Parsimonious shooting heuristic for trajectory design of connected automated traffic part II: Computational issues and optimization. *Transp. Res. Part B: Method.* 95, 421–441.
- Mannering, Fred, Walter, Kilareski, Scott, Washburn, 2007. *Principles of highway engineering and traffic analysis.* John Wiley & Sons, Hoboken, New Jersey, USA.
- Mauro, Vito, Di Taranto, C., 1990. Utopia. *IFAC Proc.* 23 (2), 245–252.
- Messer Carroll, J., Robert H., Whitson, Conrad L.F Dudek, Elio J., Romano, 1973. A variable-sequence multi-phase progression optimization program. *Highway Res. Rec.* 445, 24–33.
- Messer, C.J., Haenel, H.E., and Koeppel, E.A., 1974. *A Report on the User's Manual for Progression Analysis and Signal System Evaluation Routine-PASSER II.* No. TTI-218-72-165-14 Intrm Rpt.
- Messer Carroll, J., Daniel B., Fambro, Stephen H., Richards, 1977. Optimization of pretimed signalized diamond interchanges. *Transp. Res. Rec.* 644. .
- Michalopoulos, Panos G., 1991. Vehicle detection video through image processing: the autoscope system. *IEEE Trans. Veh. Tech.* 40 (1), 21–29.
- Mirchandani, Pitu, Head, Larry, 2001. A real-time traffic signal control system: architecture, algorithms, and analysis. *Transp. Res. Part C: Emerg. Technol.* 9 (6), 415–432.
- Porche, I., 1998. Dynamic Traffic Control: Decentralized and Coordinated Methods. ProQuest Dissertations Publishing, Dearborn, Michigan USA.
- Porche, I., and Lafortune, S., 1997. "Dynamic traffic control: Decentralized and coordinated methods." In *Proceedings of Conference on Intelligent Transportation Systems*, pp. 930-935. IEEE.
- Porche, Isaac, Lafortune, Stéphane, 1999. Adaptive look-ahead optimization of traffic signals. *J. Intel. Trans. Sys.* 4 (3-4), 209–254.
- Porche, I., Sampath, M., Sengupta, R., Chen, Y.-L., and Stephane, L., 1996. "A decentralized scheme for real-time optimization of traffic signals." In *Proceeding of the 1996 IEEE International Conference on Control Applications IEEE International Conference on Control Applications held together with IEEE International Symposium on Intelligent Control*, pp. 582-589. IEEE.
- Priemer, C., and Bernhard, F., 2008. A method for tailback approximation via C2I-data based on partial penetration." In *15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting ITS America/ERTICO/ITS Japan/TransCore.*
- Priemer, C., and Bernhard, F., 2009. A decentralized adaptive traffic signal control using V2I communication data." In *2009 12th International IEEE Conference on Intelligent Transportation Systems*, pp. 1-6. IEEE, 2009.
- Robertson, D.I., 1969. TRANSYT: a traffic network study tool. rrl report Ir 253. London, TRRL, UK.
- Roess, R.P., Prassas, E.S., and McShane, W.R., 2011. *Traffic engineering.* Pearson/Prentice Hall, Upper Saddle River, New Jersey, USA.
- Sen, Suvrajeet, Head, K. Larry, 1997. Controlled optimization of phases at an intersection. *Transp. Sci.* 31 (1), 5–17.
- Shelby, S. G., 2001. Design and Evaluation of Real-Time Adaptive Traffic Signal Control Algorithms. ProQuest Dissertations Publishing.
- Sims Arthur, G., Dobinson, Kenneth W.F, 1980. The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits. *IEEE Trans. Veh. Tech.* 29 (2), 130–137.
- Stamatidis, Chronis, Nathan H., Gartner, 1996. MULTIBAND-96: a program for variable-bandwidth progression optimization of multiarterial traffic networks. *Transp. Res. Rec.* 1554 (1), 9–17.
- Stevanovic, A., Cameron, K., and Martin, P.T., 2009. "Scoot and scats: A closer look into their operations." In *88th Annual Meeting of the Transportation Research Board. Washington DC.* 2009.
- Urbanik, T., Tanaka, A. Lozner, B., et al., 2015 *Signal timing manual.* Vol. 1. Washington, DC: Transportation Research Board, 2015.
- Webster, F.V., 1958. *Traffic Signal Settings.* Department of Scientific and Industrial Research, Her Majesty's Stationery Office, London Road Research Technical Paper No. 39, 1958.
- Yu, Chunhui, Yiheng, Feng, Henry X., Liu, Wanjing, Ma, Xiaoguang, Yang., 2019. Corridor level cooperative trajectory optimization with connected and automated vehicles. *Transp. Res. Part C Emerg. Technol.* 105, 405–421.
- Zhou, Fang, Xiaopeng, Li, Jiaqi, Ma., 2017. Parsimonious shooting heuristic for trajectory design of connected automated traffic part I: Theoretical analysis with generalized time geography. *Transp. Res. Part B: Method* 95, 394–420.

Signalized Intersections

Chaitrali Shirke, Queensland University of Technology, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Background	178
Traffic Signal Plan and Its Components	178
Goals of Designing Signal Plans	179
Methods of Designing a Signal Plan	180
Traffic Signal Control Modes	181
Performance Evaluation	182
Advances Due to Connected and Automated Vehicle Technology	182
References	183
Further Reading	184

Background

Traffic signals are used to control traffic and to avoid conflicts between road users, which are using the same road space (conflict area), for example, at grade intersections or pedestrian crossings. There are also several other implementation areas for traffic signals, such as traffic control at bottlenecks (e.g., at road construction sites), safeguarding of railway crossings, ramp metering at highway entrances or the exit of emergency service stations. Traffic signals can also be used to prioritize specific road-user groups and to meter the traffic entering an area or road section.

Traffic signal simultaneously guides all conflicting traffic streams of road users to pass a marked (or virtual) stop line or to stop and wait. A GREEN light authorizes vehicles to proceed. RED light is a negative signal indicating that road users are not allowed to proceed; they must stop and wait at the stop line. A YELLOW light (also called an amber light) gives road users a notice that the RED light will be signaled soon. It indicates that road users either have to stop at the stop line before the RED light appears or they have to pass and clear the conflict area as soon as possible.

Signal heads for motorized road traffic usually have three fields (RED, YELLOW, GREEN) which may be arranged vertically (RED on top) or horizontally (RED on the right or the left, depending on the driving side). In some special cases, signal heads with two fields or only one field can be applied. If the signal is addressing only specific traffic streams, then the signal light may show symbols for the respective road-user group (e.g., pedestrians or cyclists) or driving direction (e.g., left-turn arrow). Some of the examples of signal heads are shown in Fig. 1.

Traffic Signal Plan and Its Components

A signal plan consists of the signalization of conflicting traffic streams or modes at an intersection or crossing. A signal plan and its components for a typical four-leg intersection are shown in Fig. 2.

The GREEN time for each traffic stream depends on the cycle time, lost times (e.g., intergreen, start-up lost time) and saturation flow rate, which describes the number of vehicles that can theoretically pass the stop line within one hour of continuous GREEN time. The YELLOW time must correspond with driving dynamics (vehicle speed and braking performance), since, after the beginning of YELLOW, drivers need to decide either to stop safely in front of the stop line or to proceed and pass the stop line before the start of RED. Usually, the YELLOW time is set to a value between 3 s and 5 s, depending on the speed limit on the approach to the traffic signal. The RED time is not explicitly calculated, but it results from the other signal durations. Usually, the same signalization is repeated in a cycle with a duration defined as cycle time. During one cycle all traffic streams are given GREEN at least once.

Within one cycle, at least two or more stages (also referred as phases) can be defined as periods in which GREEN is given to one or more specific traffic streams, while all others have to wait. All streams given GREEN in one stage must be compatible streams (i.e., they do not have a common conflict area) or they must be at least partly compatible streams. For left-hand traffic shown in Fig. 2, the partly compatible streams are right-turn stream, that is, 3 and the opposing straight-through stream, that is, 8. In general, more stages allow a safer traffic operation. However, more stages also increase the number of stage transitions and the cycle time and therefore reduce the capacity of the intersection.

The stage transition (from one stage to another) must ensure that all road users of the ending stage have cleared the conflict area before any road user of the beginning stage reaches the same conflict area. In this context, the concept of intergreen times is applied.

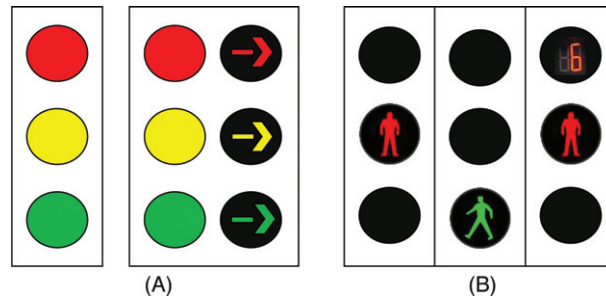


Figure 1 Example of signal heads for (A) vehicles and (B) pedestrians.

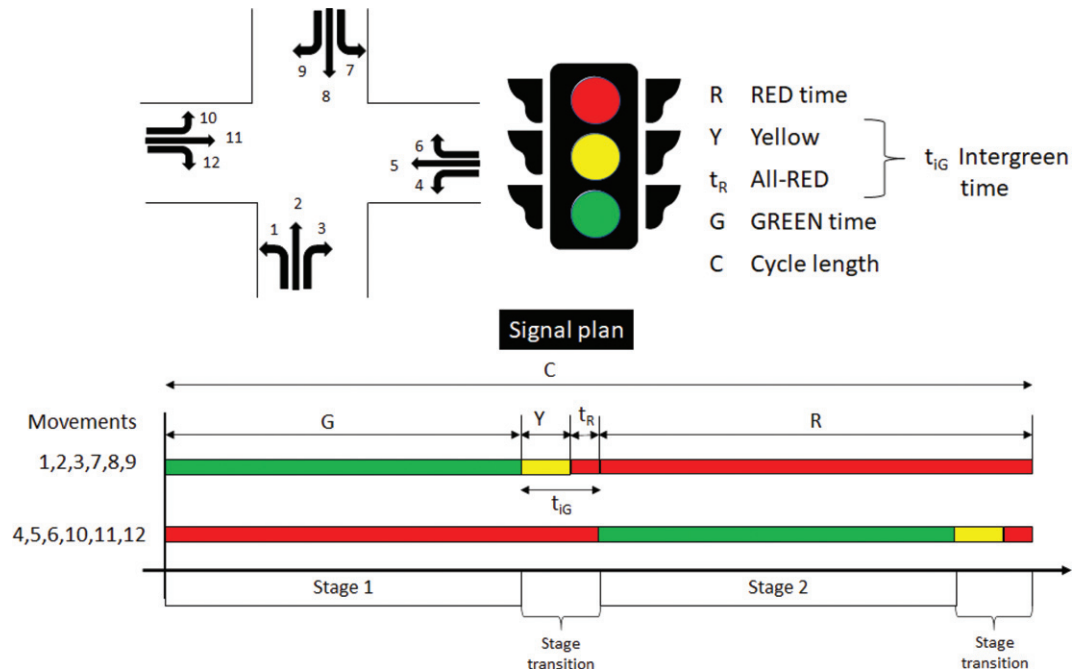


Figure 2 Example of a signal plan for typical four-leg intersection. Source: Adopted from Tang et al. (2019).

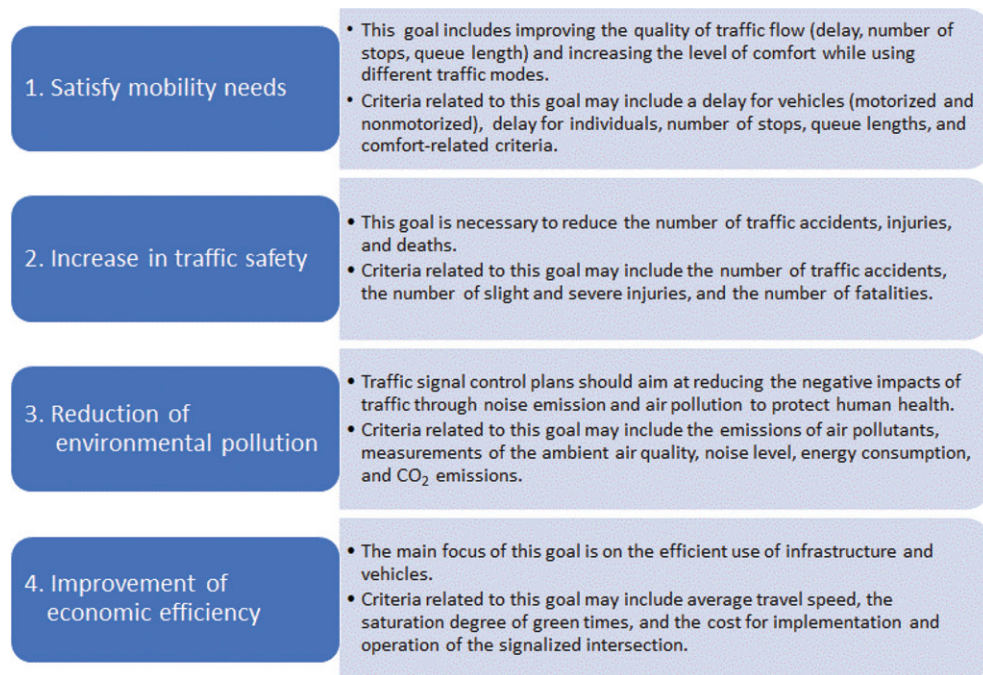
The intergreen time is defined as the time between the end of GREEN for one traffic stream and the beginning of GREEN for another stream, which is not compatible with the first stream.

The design of the signal plan has a serious impact on capacity, traffic safety, economic efficiency, and environmental compatibility. Statistics indicate that approximately 40% of total traffic accidents, about 80% of vehicle delays on urban roads, and nearly 20% of vehicle emissions occur at signalized intersections (World Health Organization, 2015; Federal Highway Administration (US) and Federal Transit Administration (US), 2017; Kan et al., 2018).

Goals of Designing Signal Plans

Usually, the impact of signals on all kinds of traffic modes such as motorized private transport, walking, cycling, and public transport should be considered for designing traffic signal plans. Furthermore, emergency vehicles and freight transport, especially heavy vehicle traffic, require special examination. The design of the traffic signal control should consider the various impacts on the different road-user groups as comprehensively as possible.

Following are four main goals within the framework of traffic management, which are generally relevant for all traffic modes (Boltze and Dinter, 1994):



Among aforementioned goals, due to the number of premature deaths caused by traffic-borne air pollution and traffic accidents, the prevention of emissions with adverse health effects and the improvement of traffic safety are recently gaining increased attention.

While designing a signal plan, the consideration of different goals or traffic modes can cause conflict. In some cases, the advantages regarding a specific goal or road-user group can only be achieved together with disadvantages for other goal or road-user groups. For a signal design, which would promote one traffic mode or a goal, besides the positive impact on this specific mode, the negative effects on other traffic modes or goals should be displayed and quantified, where possible.

For multimodal impact assessment, the same parameters regarding different traffic modes can be aggregated. For the evaluation of average delay, a person-related aggregation would be appropriate. Based on the average vehicle occupancy, the overall average delay for all road users at the intersection could be calculated (Hunter et al., 2011). Different weights could be assigned to the delay for different traffic modes to consider political preferences, comfort aspects of individual traffic mode, for example, operational aspects of public transport) (Boltze and Jiang, 2017). The signal design process gets further complicated when different modes have different priority levels of such as emergency vehicle priority, transit vehicle priority, etc.

Generally, it is also difficult to aggregate the impacts on different goals. However, it is necessary to find a fair balance of, for example, delay for different road users, number of traffic accidents, and health impacts of air pollution and noise. For fair trade-off among different objectives, multi-objective optimization traffic control models have been proposed in the literature. Researchers such as Sun et al. (2003); Hu et al. (2010); Costa et al. (2011); (Lertworawanich et al., 2011) proposed multi-objective models of fixed time that optimize different functions representing mobility needs such as average delay, the average number of stops, and maximum relative size of traffic jams. Similarly, studies (Wang et al., 2014; Zhou and Cai, 2014; Stevanovic et al., 2015) use multi-objective optimization to satisfy economic, environmental, health impacts and mobility needs simultaneous. Some research studies (Schmöcker et al., 2008; Stevanovic et al., 2011; Rakha and Abdelghaffar, 2019) apply multi-objective optimization to control different modes such as Bicycles, Buses and Cars. Various solution algorithms such as genetic algorithm (GA), ant colony optimization (ACO) algorithm, and particle swarm optimization (PSO) algorithm have been implemented to solve the multi-objective optimization problem.

Despite extensive research, a concept to combine and weight all the objectives and criteria—which is politically approved—lacks in practice. For that, more research is needed to test multi-objective approaches on a variety of networks and traffic conditions and to potentially find a good composite measure (of mobility, safety, and environment) that would enable good quality single-objective optimizations.

Methods of Designing a Signal Plan

Methods of signal plan designing, that is, obtaining values of different components of signal plan can be in general, divided into four types based on their core steps and the philosophy.

1. Critical-movement-based method with the NEMA ring-barrier constraint: It is mainly applied in the United States and Canada. Critical phase/ movement pairs are identified according to the sum of phase volumes, and those critical phases must be in one side of the barrier in the same ring. Refer to [Urbanik et al. \(2015\)](#) for more details.
2. Group-based/movement-based method: It is used in Germany, Austria, and Switzerland. It is characterized by calculating intergreen times for all the conflicting flows and determining the stage sequence to minimize the total intergreen times. The purpose of the combination of signal groups in Germany and Austria is to minimize the critical sum of traffic stream volumes. Conversely, in other types of methods, the stage structure and sequence are usually decided before calculating intergreen times.
3. Critical-movement-based method without the NEMA ring-barrier constraint: It is adopted in Australia and New Zealand. Two major tools, that is, the critical movement search table and the critical movement search diagram, are used to identify critical movements.
4. Stage-based method: It prevails in the United Kingdom, France, Turkey, Japan, China, India, South Korea, Qatar, and UAE. Stages are usually predefined according to intersection geometry and traffic volumes. Webster's method is mostly utilized with equal v/c ratios for all stages. Furthermore, signal timing plans are adjusted based on the requirements of the minimum green times for pedestrians and vehicles

Refer to [Tang et al. \(2019\)](#) for a comparative explanation of abovementioned methods. For more details on standard practices of calculating signal parameters, refer to [Akcelik \(1981\)](#); [FHWA \(2008\)](#). As suggested in [Tang et al. \(2019\)](#), the selection of the signal designing method depends on multiple factors, such as signal control objectives, functionalities of signal controllers, individual needs of public transport priority or signal coordination, traffic flow characteristics, and intersection geometry.

Traffic Signal Control Modes

There are various traffic signal control modes, which differ in the degree of variability of different control parameters (i.e., cycle time, stage number, stage sequence, green times, offset time). These control modes are listed below:

1. Fixed-time control: in this mode, none of these parameters in the signal plan is flexible. Signal plans are designed offline for various demands using historical traffic data. Suitable signal plans are selected and activated for a specific time of the day, which is referred to as the time-dependent (or time-of-day) signal control.
2. Actuated control: in this mode, at least one of the parameters in the signal program is actuated according to the current situation (e.g., green time adaptation based on traffic volume detection). Following two options are possible:
 - a. Semi actuated control: generally minor movements are actuated, which allows for an indefinite extension of green time for major movements until minor phase demand is actuated
 - b. Fully actuated traffic control: all movements (phases) are actuated, which allows for variation of phase sequence and duration depending on traffic demands

Refer to [Ni \(2020\)](#) for more details on vehicle actuation at a signalized intersection.
3. Traffic responsive plan selection: based on detector data returned from the field, traffic responsive algorithms select a preconfigured (i.e., predetermined) signal timing plan from a library of plans.
4. Real-time or adaptive control: In this mode, real-time detection data and algorithms are used to adjust signal-timing parameters for current conditions. Unlike traffic responsive plan selection systems, which use predetermined timing plans, in this mode, the system can adjust various timing parameters (within certain constraints) based on what the traffic requires.

Actuated or adaptive control is more common in most of the economically developed countries, including the United States, Canada, Germany, Austria, the United Kingdom, France, Switzerland, Australia, New Zealand, and Japan. Whereas, economically developing countries such as India, China, South Korea majorly rely on Fixed time control ([Tang et al., 2019](#)). There are various tools available to design offline fixed time control such as TRANSYT ([Robertson, 1969](#)), TRNASYT-7f ([Wallace et al., 1984](#)), SIDRA intersection ([Akcelik Associates Pty. Ltd., 1984](#)) etc. Due to rapid development of detection and computer technologies, most of the new generation signal controllers such as controllers by Siemens, Econolite, Swarco, Tyco, Nippon, etc. offer actuation.

Adaptive traffic control systems (ATCS) have been in continuous use since the 1980s. The first successes of ATCS include SCATS ([Hunt et al., 1981](#)) and SCOOT ([Lowrie, 1982](#)), which were soon followed by an FHWA initiative ([Tarnoff and Gartner, 1993](#)) to develop ATCSs in the United States. ATCSs currently in practice include SCATS, SCOOT, Centrac, ATMS, BALANCE, TACTICS, and VS-PLUS, etc. For more information on field experiences with ATCSs systems, refer to [Stevanovic \(2010\)](#).

Though a variety of centralized adaptive control systems have been developed and implemented in several countries, their practical effectiveness and cost-benefit have been argued over for a long time in industry and academia. These systems are generally considered to be capable of handling highly fluctuated traffic flow at the network level, mainly relying on good-quality traffic detectors and experienced engineers. However, due to the failure of traffic detectors and lack of experienced and knowledgeable people, as well as complex dynamics of traffic flow, the capabilities and functionalities of these systems cannot be fully utilized in many cases. Thus generally, their practical performance has not yet been compelling.

On the other hand, many academics and practitioners believe that well-designed fixed-time signal control, together with well-defined corridor coordination, is a good alternative in terms of cost and benefit, given limited resources for installing advanced signal controllers and systems as well as hiring experienced engineers for signal timing design and maintenance. Besides, the isolated actuated signal control with simple control logic is also considered as an efficient means for rural intersections. In some countries, such as the United States and Germany, actuated-coordination control has also proved to be quite useful. Researchers (Yin, 2008; Zhang and Yin, 2008; Zhang et al., 2010) have also proposed concept of robust signal plan to improve the performance of fixed time signals under varying demand. The signal plan that minimizes cost associated with different demand scenarios and whose performance is stable under different realizations of traffic volumes is called robust plan. The signal plan producing the optimal value of the objective functions such as total delay, total throughput etc. is known as an efficient signal plan for a given demand. On the other hand, robustness is a measure of deviation in the performance of a signal plan across various traffic demands. Given the conflicting nature of efficiency and robustness of signal control, a trade-off would be inevitable to make. Objective of robust signal plan is to achieve the reasonable trade-off between efficiency and robustness to achieve satisfactory performance of signal plan under varying demand.

Performance Evaluation

Performance evaluation is fundamental for continuous monitoring and improvement of signal timing plans at signalized intersections. Comprehensive and accurate assessment can help to diagnose signal control problems and providing insights for improvement. Thus, it is not only essential for the signal plan designing stage but also crucial for the operation and maintenance stage. From the viewpoint of cost-benefit, efficient monitoring and improvement of existing signal timing plans are sometimes more effective than using advanced control systems and devices.

Performance indicators for the evaluation of signalized intersections include control delay, capacity, degree of saturation, queue length, number of stops, travel time, number of traffic accidents or conflicts, fuel consumption, and so on. Though average control delay is widely used for determining LOS, the multicriteria and multimodal impact assessment has gained increasing concerns, which simultaneously accounts for the safety, efficiency, environment, and energy of multiple traffic modes. For instance, control delay per person might be a good integrated indicator in the context of multimodal operations, as it can account for the number of people accommodated by each type of vehicle. On the other hand, performance indicators should be appropriately selected based on signal control objectives, as too many indicators do not necessarily contribute to improvement. As discussed in Goals of designing signal plans, the importance of various performance indicators vary under different traffic circumstances.

In the past, the management of traffic signal systems was evaluated by labour-intensive processes, such as floating-car studies. In the mid-2000s, the concept of “high-resolution data” took shape (Smaglik et al., 2007a; Smaglik et al., 2007b; Liu and Ma, 2009). These studies required the real-time traffic states at individual detectors, which needed to be cross-referenced against the active signal state during each moment of operation. Based on such data, researchers soon developed Automated Traffic Signal Performance Measures (ATSPM) (Day and Bullock, 2011; Gettman et al., 2012; Wu and Liu, 2014). ATSPM are control-agnostic, discrete event-based performance measures for analyzing the actual operation of traffic signal systems, applicable within any signal system and for all types of operations (Day et al., 2014). ATSPM helps in automizing the process of performance evaluation and finding opportunities for improvement using real time data from traffic sensors and signal controller (Day et al., 2016). Soon many signal vendors started implementing high-resolution data collection in their controller products, which led to the development of software to deliver ATSPM to the user (UDOT, 2020).

The various sensors used to collect real-time traffic data for performance evaluation and operation of ATCSs are summarized below (Stevanovic, 2010):

- Inductive loop detectors: these detectors are classified based on their position at intersection approach such as stop-line, upstream mid-block, upstream far side.
- Radar detectors
- Video detection

These sensors are used to measure or estimate various traffic flow characteristics such as vehicle count, speed, occupancy (i.e., percentage of time detector is occupied), queue length etc. Since 2006, automatic vehicle identification (AVI) data in the form of Bluetooth MAC address matching emerged (Wasson et al., 2008). Thus, in the past few years, the raw vehicle position data has been applied to evaluate and adjust signal timing (Quayle et al., 2010; Remias et al., 2013; Sharifi et al., 2017). Future development of similar data sources and performance measures based on that data is a current research trend as discussed in the following section.

Advances Due to Connected and Automated Vehicle Technology

New wireless communication and computational technologies, such as connected and automated vehicles (CAVs) have inspired new methods of controlling traffic. The benefits of introducing CAVs include but not limited to crash reduction/elimination, travel time reduction, energy efficiency improvement, and others. Current research in the area of traffic control focus on achieving the following goals with the aid of CAVs:

- CAVs data can be used to improve the estimation accuracy of the performance indicators (Wan et al., 2016; Rompis et al., 2018; Xie et al., 2018)
- Better signal plans can be designed, as the arrivals of vehicles could be better predicted in advance (Goodall et al., 2013; Wu et al., 2015; Younes and Boukerche, 2016; Liang et al., 2018)
- Drivers and fully automated vehicles could better adapt their trajectories to cooperate with signal timing to reduce overall congestion and fuel consumption in urban areas (Katsaros et al., 2011; Li et al., 2012; Ubiergo and Jin, 2016; Tang et al., 2018)
- With CAVs, information between signals and individual vehicles can be exchanged in real-time, which can be used to simultaneously guide the vehicle's trajectory and plan better signal control. Such simultaneous planning is referred to as Signal vehicle coupled control (SVCC) (Xu et al., 2017; Yu et al., 2018)
- Recent studies (Hu et al., 2016; Wu et al., 2016; Wu et al., 2018) utilized CAVs to build efficient algorithms for transit priority control methods. CAVs provide more accurate and richer data for the control system, which help to generate an optimal solution to reduce the overall delay.

With this new technology, come new challenges such as variation in penetration rate and level of automation of CAVs, limited field data to test the developed methods, reliability of wireless communication and cyber security issues. Besides, technical and organizational obstacles also exist. The roles, responsibilities and business models of the different stakeholders are not clearly specified yet. From the technical point of view, since the currently installed infrastructure is expensive and partially aged, researchers need to think about simple ways for upgrading to connected environments. Thus, while the benefits of the CAVs are evident, there are still obstacles on the way to a large-scale functionally homogeneous CAV system deployment.

References

- Akelik Associates Pty. Ltd., 1984. Sidra Intersection. Available from: <http://www.sidrasolutions.com>.
- Akelik, R., 1981. Traffic signals: capacity and timing analysis ARR No. 123. ARRB Transport Research Ltd., Vermont South, Australia.
- Boltze, M., Dinter, M., 1994. The project FRUIT: a goal-oriented approach to traffic management in Frankfurt am main and the Rhein-Main region. *Traffic Eng. Control* 35 (7/8), 437–444.
- Boltze, M., Jiang, W., 2017. Berücksichtigung verschiedener Verkehrsteilnehmergruppen und Kriterien bei der Optimierung von Lichtsignalanlagen-ein Vorschlag zum verkehrspolitischen Rahmen fuer die Gestaltung der Lichtsignalsteuerung in Staedten. *Strassenverkehrstechnik* 61 (9).
- Costa, B.C., Almeida, P.E.M., Caldeira, E., 2011. Traffic lights timing inside microregion simulator using multiobjective optimization. Paper presented at the 2011 IEEE Congress of Evolutionary Computation (CEC).
- Day, C., Bullock, D., 2011. Arterial performance measures, volume 1: performance-based management of arterial traffic signal systems. NCHRP.
- Day, C.M., Bullock, D.M., Li, H., Lavrenz, S.M., Smith, W.B., Sturdevant, J.R., 2016. Integrating traffic signal performance measures into agency business processes.
- Day, C.M., Bullock, D.M., Li, H., Remias, S.M., Hainen, A.M., Freije, R.S., Brennan, T.M., 2014. Performance measures for traffic signal systems: An outcome-oriented approach (162260296X).
- Federal Highway Administration (US), Federal Transit Administration (US), 2017. 2015 Status of the Nation's Highways, Bridges, and Transit Conditions & Performance Report to Congress: Government Printing Office.
- FHWA, 2008. Traffic signal timing manual. Available from: <https://rosap.nhtl.bts.gov/view/dot/20661>.
- Gettman, D., Madrigal, G., Allen, S., Boyer, T., Walker, S., Tong, J., Hu, H., 2012. Operation of Traffic Signal Systems in Oversaturated Conditions, Volume 2, Final report.
- Goodall, N.J., Smith, B.L., Park, B., 2013. Traffic signal control with connected vehicles. *Transp. Res. Rec.* 2381 (1), 65–72, doi:10.3141/2381-08.
- Hu, H., Gao, Y., Yang, X., 2010. Multi-objective optimization method of fixed-time signal control of isolated intersections. Paper presented at the 2010 International Conference on Computational and Information Sciences.
- Hu, J., Park, B.B., Lee, Y.-J., 2016. Transit signal priority accommodating conflicting requests under connected vehicles technology. *Transp. Res.* 69, 173–192, doi:10.1016/j.trc.2016.06.001.
- Hunt, P., Robertson, D., Bretherton, R., Winton, R., 1981. SCOOT-a traffic responsive method of coordinating signals (0266-7045).
- Hunter, B., Wolfermann, A., Boltze, M., 2011. Multimodal assessment of signalized intersections considering the number of travellers. Paper presented at the TRB Annual Meeting Compendium of Papers.
- Kan, Z., Tang, L., Kwan, M.-P., Zhang, X., 2018. Estimating vehicle fuel consumption and emissions using GPS big data. *Int. J. Environ. Res.* 15 (4), 566.
- Katsaros, K., Kernchen, R., Dianati, M., Rieck, D., 2011. Performance study of a green light optimized speed advisory (GLOSA) application using an integrated cooperative ITS simulation platform. Paper presented at the 2011 7th International Wireless Communications and Mobile Computing Conference.
- Lertworawanich, P., Kuwahara, M., Miska, M., 2011. A new multiobjective signal optimization for oversaturated networks. *IEEE Trans. Intell. Transp. Syst.* 12 (4), 967–976, doi:10.1109/TITS.2011.2125957.
- Li, L., Wen, D., Zheng, N., Shen, L., 2012. Cognitive cars: a new frontier for ADAS research. *IEEE Trans. Intell. Transp. Syst.* 13 (1), 395–407, doi:10.1109/TITS.2011.2159493.
- Liang, X., Guler, S.I., Gayah, V.V., 2018. Signal timing optimization with connected vehicle technology: platooning to improve computational efficiency. *Transp. Res. Rec.* 2672 (18), 81–92, doi:10.1177/0361198118786842.
- Liu, H.X., Ma, W., 2009. A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials. *Transp. Res.* 17 (1), 11–26, doi:10.1016/j.trc.2008.05.002.
- Lowrie P.R., The Sydney Co-ordinated Adaptive Traffic System: Principles, Methodology, Algorithms 1982. Proceedings of the IEE international conference on road traffic signalling, pp. 67–70. London, UK
- Ni, D., 2020. Actuated control Signalized Intersections: Fundamentals to Advanced Systems, Springer International Publishing, Cham, pp. 211–224.
- Quayle, S.M., Koonce, P., Depencier, D., Bullock, D.M., 2010. Arterial performance measures with media access control readers: Portland, Oregon Pilot Study. *Transp. Res. Rec.* 2192 (1), 185–193, doi:10.3141/2192-18.
- Rakha, H.A., Abdelghaffar, H., 2019. Development of Multimodal Traffic Signal Control.
- Remias, S.M., Hainen, A.M., Day, C.M., Brennan, T.M., Li, H., Rivera-Hernandez, E., Bullock, D.M., 2013. Performance characterization of arterial traffic flow with probe vehicle data. *Transp. Res. Rec.* 2380 (1), 10–21, doi:10.3141/2380-02.
- Robertson, D.I., 1969. TRANSYT: a traffic network study tool.
- Rompis, S.Y.R., Cetin, M., Habtemichael, F., 2018. Probe vehicle lane identification for queue length estimation at intersections. *J. Intell. Transp. Syst.* 22 (1), 10–25, doi:10.1080/15472450.2017.1300887.

- Schmöcker, J.-D., Ahuja, S., Bell, M.G.H., 2008. Multi-objective signal control of urban junctions—Framework and a London case study. *Transp. Res.* 16 (4), 454–470, doi:10.1016/j.trc.2007.09.004.
- Sharifi, E., Young, S.E., Eshragh, S., Hamed, M., Juster, R.M., Kaushik, K., 2017. Outsourced probe data effectiveness on signalized arterials. *J. Intell. Transp. Syst.* 21 (6), 478–491, doi:10.1080/15472450.2017.1359093.
- Smaglik, E.J., Bullock, D.M., Sharma, A., 2007. Pilot study on real-time calculation of arrival type for assessment of arterial performance. *J. Transp. Eng.* 133 (7), 415–422, doi:10.1061/(ASCE)0733-947X.
- Smaglik, E.J., Sharma, A., Bullock, D.M., Sturdevant, J.R., Duncan, G., 2007b. Event-based data collection for generating actuated controller performance measures. *Transp. Res. Rec.* 2035 (1), 97–106, doi:10.3141/2035-11.
- Stevanovic, A., 2010. NCHRP synthesis 403: Adaptive traffic control systems: Domestic and foreign state of practice. Transportation Research Board of the National Academies, Washington, D.C.
- Stevanovic, A., Stevanovic, J., Kergay, C., Martin, P., 2011. Traffic control optimization for multi-modal operations in a large-scale urban network. Paper presented at the 2011 IEEE Forum on Integrated and Sustainable Transportation Systems.
- Stevanovic, A., Stevanovic, J., So, J., Ostojic, M., 2015. Multi-criteria optimization of traffic signals: Mobility, safety, and environment. *Transp. Res.* 55, 46–68, doi:10.1016/j.trc.2015.03.013.
- Sun, D., Benekohal, R.F., Waller, S.T., 2003. Multiobjective traffic signal timing optimization using non-dominated sorting genetic algorithm. Paper presented at the IEEE IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No.03TH8683).
- Tang, K., Boltze, M., Nakamura, H., Tian, Z., 2019. Global Practices on Road Traffic Signal Control: Fixed-time Control at Isolated Intersections. Elsevier.
- Tang, T.-Q., Yi, Z.-Y., Zhang, J., Wang, T., Leng, J.-Q., 2018. A speed guidance strategy for multiple signalized intersections based on car-following model. *Phys. A: Stati. Mech. Appl.* 496, 399–409, doi:10.1016/j.physa.2018.01.005.
- Tarnoff, P.J., Gartner, N., 1993. Real-time, traffic adaptive signal control. Paper presented at the Large Urban Systems. Proceedings of the Advanced Traffic Management Conference Federal Highway Administration.
- Ubiergo, G.A., Jin, W.-L., 2016. Mobility and environment improvement of signalized networks through vehicle-to-Infrastructure (V2I) communications. *Transp. Res.* 68, 70–82, doi:10.1016/j.trc.2016.03.010.
- UDOT, 2020. Automated traffic signal performance measures. Available from: <https://udottraffic.utah.gov/ATSPM/>.
- Urbanik, T., Tanaka, A., Lozner, B., Lindstrom, E., Lee, K., Quayle, S., Gettman, D., 2015. Signal timing manual: NCHRP report 812, Vol. 1, Transportation Research Board, Washington, DC.
- Wallace, C.E., Courage, K., Reaves, D., Schoene, G., Euler, G., 1984. TRANSYT-7F User's Manual.
- Wan, N., Vahidi, A., Luckow, A., 2016. Reconstructing maximum likelihood trajectory of probe vehicles between sparse updates. *Transp. Res.* 65, 16–30, doi:10.1016/j.trc.2016.01.010.
- Wang, J.Y.T., Ehrgott, M., Dirks, K.N., Gupta, A., 2014. A bilevel multi-objective road pricing model for economic. *Environ. Health Sustain.* 3, 393–402, doi:10.1016/j.trpro.2014.10.020.
- Wasson, J.S., Sturdevant, J.R., Bullock, D.M., 2008. Real-time travel time estimates using media access control address matching. *ITE J* 78 (6), 20–23.
- World Health Organization, 2015. Global status report on road safety 2015. World Health Organization.
- Wu, W., Head, L., Yan, S., Ma, W., 2018. Development and evaluation of bus lanes with intermittent and dynamic priority in connected vehicle environment. *J. Intell. Transp. Syst.* 22 (4), 301–310, doi:10.1080/15472450.2017.1313704.
- Wu, W., Ma, W., Long, K., Wang, Y., 2016. Integrated optimization of bus priority operations in connected vehicle environment 50 (8), 1853–1869, doi:10.1002/atr.1433.
- Wu, W., Zhang, J., Luo, A., Cao, J., 2015. Distributed mutual exclusion algorithms for intersection traffic control. *IEEE Trans. Parallel Distrib. Syst.* 26 (1), 65–74, doi:10.1109/TPDS.2013.2297097.
- Wu, X., Liu, H.X., 2014. Using high-resolution event-based data for traffic modeling and control: An overview. *Transp. Res.* 42, 28–43, doi:10.1016/j.trc.2014.02.001.
- Xie, X., van Lint, H., Verbraeck, A., 2018. A generic data assimilation framework for vehicle trajectory reconstruction on signalized urban arterials using particle filters. *Transp. Res.* 92, 364–391, doi:10.1016/j.trc.2018.05.009.
- Xu, B., Ban, X.J., Bian, Y., Wang, J., Li, K., 2017. V2I based cooperation between traffic signal and approaching automated vehicles. Paper presented at the 2017 IEEE Intelligent Vehicles Symposium (IV).
- Yin, Y., 2008. Robust optimal traffic signal timing. *Transp. Res.* 42 (10), 911–924.
- Younes, M.B., Boukerche, A., 2016. Intelligent traffic light controlling algorithms using vehicular networks. *IEEE Trans. Vehic. Technol.* 65 (8), 5887–5899, doi:10.1109/TVT.2015.2472367.
- Yu, C., Feng, Y., Liu, H.X., Ma, W., Yang, X., 2018. Integrated optimization of traffic signals and vehicle trajectories at isolated urban intersections. *Transp. Res.* 112, 89–112, doi:10.1016/j.trb.2018.04.007.
- Zhang, L., Yin, Y., 2008. Robust synchronization of actuated signals on arterials. *Transp. Res. Rec.* (2080), 111–119.
- Zhang, L., Yin, Y., Lou, Y., 2010. Robust signal timing for arterials under day-to-day demand variations. *Transp. Res. Rec.* (2192), 156–166.
- Zhou, Z., Cai, M., 2014. Intersection signal control multi-objective optimization based on genetic algorithm. *J. Traffic Transp. Eng.* 1 (2), 153–158, doi:10.1016/S2095-7564(15)30100-8.

Further Reading

- Guo, Q., Li, L., Ban, X., 2019. Urban traffic signal control with connected and automated vehicles: A survey. *Transp. Res.* 101, 313–334, doi:10.1016/j.trc.2019.01.026.
- Jing, P., Huang, H., Chen, L., 2017. An adaptive traffic signal control in a connected vehicle environment: a systematic review. *Information* 8 (3), 101.
- Kaths, J., Papanagiotou, E., Busch, F., 2015. Traffic Signals in Connected Vehicle Environments: Chances, Challenges and Examples for Future Traffic Signal Control. Paper presented at the 2015 IEEE 18th International Conference on Intelligent Transportation Systems.
- Namazi, E., Li, J., Lu, C., 2019. Intelligent intersection management systems considering autonomous vehicles: a systematic literature review. *IEEE Access* 7, 91946–91965, doi:10.1109/ACCESS.2019.2927412.
- Ni, D., 2020. Signalized Intersections: Fundamentals to Advanced Systems. Springer Nature.

Protected Phase

Jiarong Yao, Yumin Cao, Keshuang Tang, Tongji University, Shanghai, China

© 2021 Elsevier Ltd. All rights reserved.

Glossary	185
Definition of Protected Phase	185
Types of Protected Phase	185
When and Where to Implement a Protected Phase	188
Operational and Safety Performance of Protected Phase	192
Summary	193
Acknowledgment	194
Biography	194
References	194
Further Reading	195

Glossary

Movement Classified motorized and pedestrian traffic streams at an intersection according to direction, lane and right-of-way, etc.

Phase The part of the signal cycle allocated to any combinations of traffic movements receiving the right-of-way simultaneously during one or more intervals.

Signal group A group of one or more traffic flows which receive identical signal light indications.

Stage A part of the cycle during which a particular set of compatible streams receives green while all others have to wait.

Protected phase A separate phase for protected movement with no conflicting traffic flows.

Permitted phase A timing treatment allowing two or more conflicting movements to pass concurrently.

Lost time The time between signal phases during which an intersection is not used by any traffic.

Crosswalk A marked area for pedestrians to cross the street at an intersection or designated midblock location.

Definition of Protected Phase

- **Phase, signal group and stage** A traffic signal phase, or a **phase** for short, is a timing process within the signal controller, that facilitates serving one or more movements at the same time for one or more modes of users. A **signal group** refers to a group of one or more traffic flows which always receive identical signal light indications. A **stage** is a part of the cycle during which a particular set of compatible streams receives green while all others have to wait (Tang et al., 2019). Fig. 1 provides a description for these three terms in the scope of an intersection for better understanding. Note that the term “phase” used here should not be confused with the term “NEMA phase” widely used in North America, which is a green time interval accommodating only one movement, that is, a through movement or a left-turn movement (FHWA, 2009; Transportation Research Board, 2010).
- **Protected phase** Also known as “protected-only phase”, **protected phase** refers to providing a separate phase for protected movements with no conflicting traffic flows. In contrast, **permitted phase** (also known as “permitted-only phase”, “permissive-only phase”) refers to a timing treatment allowing two or more conflicting movements to pass concurrently. As a compromise, **protected/permitted phase** is a compound phase that displays the permitted phase after the protected phase. While, **permitted/protected phase** is a compound phase that displays the permitted phase before the protected phase. Note that these three terms are more commonly used for motor vehicles, while specific designations do exist for protected phase of public transport and non-motorized traffic, such as exclusive phase, pedestrian scramble and so on.

Types of Protected Phase

To avoid confusion, in the following text the protected phases for motor vehicles are introduced in the context of **right-hand traffic**.

- **Protected left-turn phase (motor vehicles) and its displays**
Traffic signal phasing that provides an exclusive phase for left turns and prohibits opposing through movements and pedestrian crossings is known as a **protected left turn phase**. There are two types of protected left-turn phases according to phase sequence, that is, leading left turn and lagging left turn. A leading left turn allows the green phase of left-turn movements ahead of the green phase of through movements, while a lagging left turn places the green phase of the left turn after the green phase of through movements.

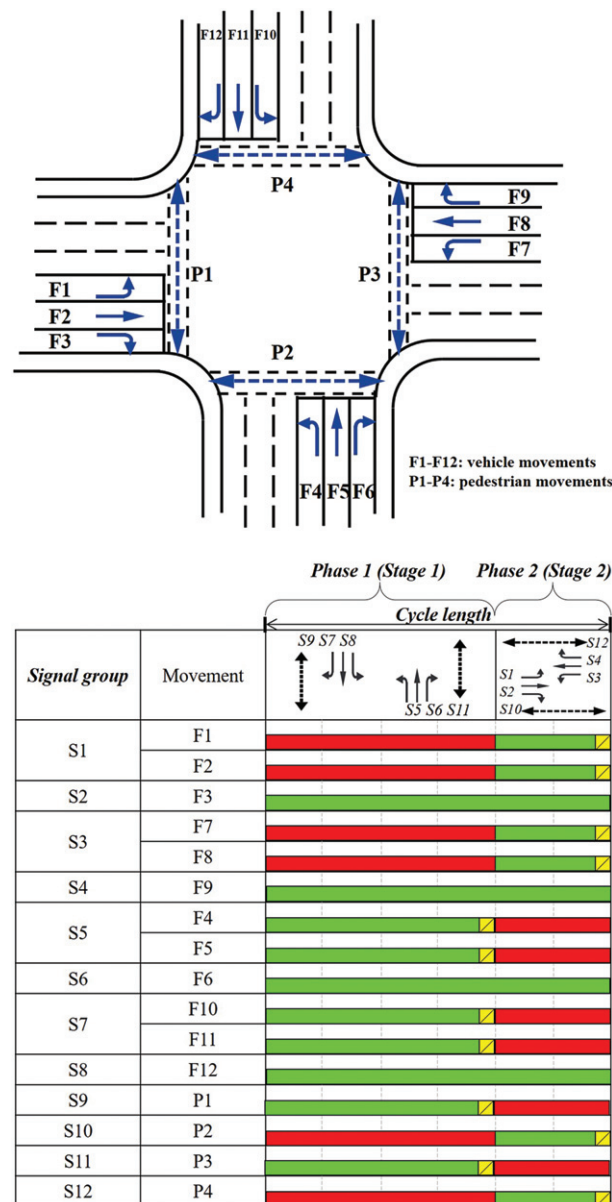
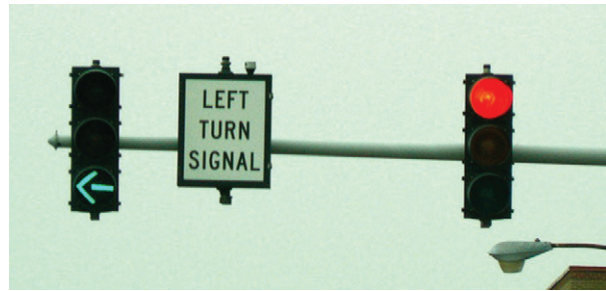


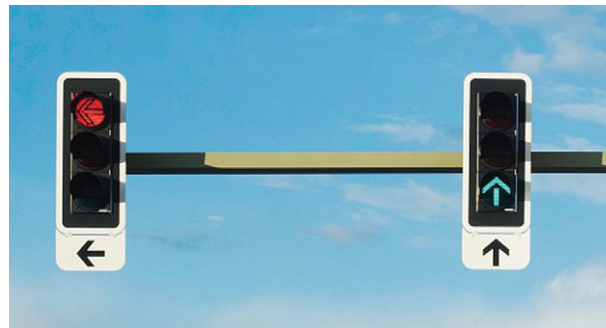
Figure 1 Illustration of phase, signal group and stage.

As shown in [Fig. 2](#), the left turn green arrow light is widely used as the exclusive display of left turn protected phase. Despite different signal head arrangements in various countries, green arrow light is used to direct vehicles to turn in the direction that the arrow indicates, when operated together with signal lights of other phases ([Tang et al., 2019](#)).

- **Protected right-turn phase (motor vehicles) and its displays** Traffic signal phasing that provides an exclusive phase for right turns and prohibits conflicting pedestrian crossings is known as a protected right turn. As shown in [Fig. 3](#), a right-turn green arrow signal indication is displayed under the protected right-turn phase ([Tang et al., 2019](#)).
- **Protected public transport phase and its displays** Traffic signal phasing that provides an exclusive phase for public transport (like BRT, LRT) and prohibits private vehicle movements and pedestrian crossings is known as a protected public transport phase. Examples of signal displays for the protected public transport phase in different countries are given in [Fig. 4](#). Note that the signal displays may be varied for different kinds of public transport in different countries, yet with the same function to indicate when to stop or go for public traffic ([FGSV, 2015](#); [JSTE, 2006](#); [The Ministry of Public Security of the People's Republic of China, 2015](#); [Urbanik et al., 2015](#)).
- **Protected pedestrian phase and its displays** Traffic signal phasing that provides an exclusive phase for pedestrians and prohibits conflicting motor vehicle movements is known as a protected pedestrian phase. Generally, protected pedestrian phase can be operated in pair at an intersection for pedestrians of both sidewalks along a road, together with parallel through phases for motor vehicles. Examples of signal displays for the protected pedestrian phase in some countries are shown in [Fig. 5](#) ([JSTE, 2006](#); [The Ministry of Public Security of the People's Republic of China, 2015](#); [Urbanik et al., 2015](#); [FGSV, 2015](#)).



(A)



(B)



(C)



(D)

Figure 2 Examples of signal displays for protected left-turn phase: (A) United States, (B) Germany, (C) China, and (D) Japan.

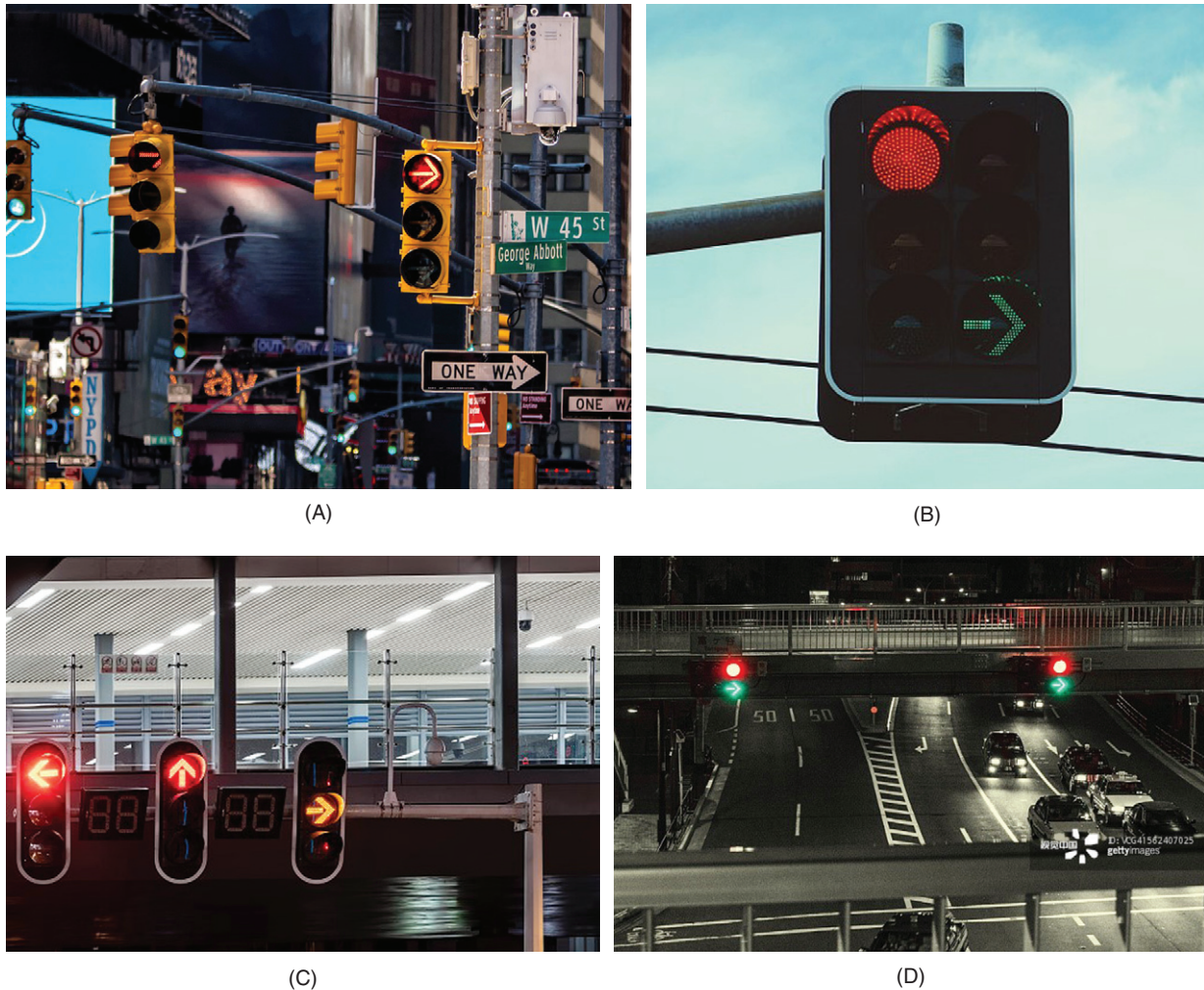


Figure 3 Examples of signal displays for protected right-turn phase: (A) United States, (B) Germany, (C) China, and (D) Japan.

- **Protected bicycle phase and its displays** Traffic signal phasing that provides an exclusive phase for bicycles and prohibits conflicting motor vehicle movements is known as a protected bicycle phase. Generally, bicycles share the same right-of-way as pedestrians, thus the same signal phase and even signal display. Examples of signal displays peculiar to bicycles implemented in some countries are shown in Fig. 6, which are sometimes also used to indicate exclusive right-of-way for both pedestrians and bicycles.

When and Where to Implement a Protected Phase

- **Protected left-turn phase**

When and where to implement a protected left-turn phase as well as the selection of its protection types (i.e., protected/permissive, permitted/protected, or protected only) must be determined before signal timing calculations. Generally, the protected left-turn phase for motor vehicles should be considered if any of the following conditions is satisfied (FSV, 1998; National Police Agency of Japan, 2015; Rodegerdts et al., 2004; The Ministry of Public Security of the People's Republic of China, 2016).

1. There is more than one exclusive left-turn lanes, or more than one opposite through lanes;
2. The left-turn volume is high, or the conflicting traffic volumes in the opposite through direction are high. For example, the left turn has a demand in excess of 240 veh/h, which is only be used for planning applications;
3. There are more than one streams conflicting with the left-turn movements;
4. The speed of the opposite flow is relatively high;
5. The left-turn lane is placed on right or a shared U-turn and left-turn lane exists

To implement a protected left-turn phase, many other factors should also be considered, including accident frequency, field observations, and conditions that may exist outside of the analysis period.

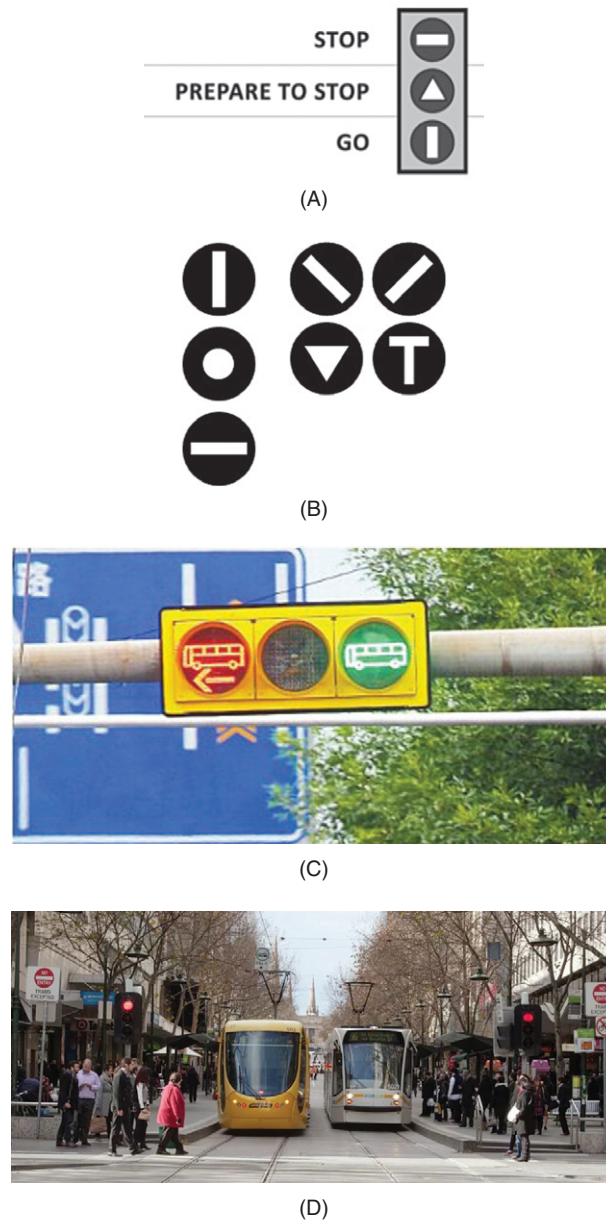


Figure 4 Examples of signal displays for protected public transport phase: (A) United States, (B) Germany, (C) China, and (D) Australia.

- **Protected right-turn phase**

In many cases, right-turning vehicles share a lane with through vehicles and should yield to crossing pedestrians and cyclists in many countries such as United States, Germany, and China. When the exclusive right-turn lane is available, implementation of protected right-turn phase may be considered. More in detail, the protected right-turn phase for motor vehicles should be considered if any of the following conditions is satisfied (FSV, 1998b; Rodegerdts et al., 2004).

1. Parallel pedestrian volumes are high and safety considerations predominate.
2. Right-turn traffic volumes are high or many conflicts with parallel pedestrians or cyclists occur and cannot be mitigated by other means;
3. Trams have separate through tracks on the right;
4. The right-turn traffic demand and the left-turn traffic demand at the crossing road are balanced.

Protected Public Transport Phase

Protected phase for public transport can be considered only when an exclusive lane bus lane is available. Factors like the number of bus routes passing the intersection and traffic demand of private vehicles should be taken into account to improve the efficiency of public transit without causing severe delay of social vehicles of other phases.



(A)



(B)



(C)



(D)

Figure 5 Examples of signal displays for protected pedestrian phase: (A) United States, (B) Germany, (C) China, and (D) Japan.



(A)



(B)



(C)



(D)

Figure 6 Examples of signal displays for protected bicycle phase: (A) United States, (B) Germany, (C) China, and (D) Japan.

- **Protected pedestrian phase**

Generally, pedestrians enjoy priority when crossing the intersection permitted with turning vehicles, so the pedestrian movement can be regarded as protected. In case that the protection for pedestrian requires to be more mandatory, protected phase exclusive for pedestrians should be considered, and the following conditions can be used as a reference ([FSV, 1998](#); [National Police Agency of Japan, 2015](#); [Rodegerdts et al., 2004](#); [The Ministry of Public Security of the People's Republic of China, 2016](#)).

1. The pedestrian demand is high and the impact of the conflict between pedestrians and turning vehicles are expected to be significant;
2. Delay for vehicular turning traffic is excessive due to the heavy pedestrian traffic;
3. There are a large number of vehicle-pedestrian conflicts involving all movements;
4. Multiphase signal indications (as with split-phase timing) would tend to confuse or cause conflicts with pedestrians using a crosswalk guided only by vehicular signal indications;
5. At an established school crossing at any signalized location;
6. A protected pedestrian phase for all directions can be used at intersections with diagonal pedestrian-crossing lane markings. Such exclusive pedestrian signal phase allows pedestrians to cross in all directions at the same time, including diagonally. Vehicle signals are red on all approaches of the intersection during the exclusive pedestrian signal phase. It is sometimes called a "barn dance" or "pedestrian scramble," whose objective is to reduce vehicle turning conflicts, decrease walking distance, and make intersections more pedestrian-friendly.
7. Under some special situations of unconventional intersection geometric design, no vehicular signal indications are visible to pedestrians, or the vehicular signal indications that are visible to pedestrians starting or continuing a crossing provide insufficient guidance for them to decide when it is reasonably safe to cross, such as on one-way streets, at T-intersections, or midblock crossing for links where neighboring intersections are far away for pedestrians to access.

- **Protected bicycle phase**

Generally, bicycle movement operate concurrently with pedestrian movement when passing the crosswalk at the intersections. Non-motor vehicles should yield to pedestrians when they have conflicts on the crosswalk while turning motor vehicles should yield to both the bicycles and pedestrians crossing the intersections. Thus, the guidelines for setting protected pedestrian phase can also be applied to protected bicycle phase when an exclusive lane for bicycles is available. Two additional conditions are listed as follows to provide specific suggestions for protected bicycle phase ([Urbanik et al., 2015](#)).

1. For two-phase intersections, exclusive signals or exclusive phases should be set up for left-turn non-motor vehicles if the average left-turn flow rate during the red interval is relatively heavy. Alternatively, other ways of street crossing organization for non-motor vehicles should be adopted;
2. For four-phase intersections, non-motor vehicle exclusive signals should be set up when the green interval needed by the non-motor vehicles is obviously larger than the motor vehicles in the parallel direction. And an "early break" mode should be adopted by the non-motor signals.

Operational and Safety Performance of Protected Phase

- **Operational performance**

According to the signal timing design procedure, the lost time will be a function of the cycle length and the number of phases. Thus, from the perspective of operational performance, the protected phase directly relates to the delay of users passing the intersection.

For protected turning phases of motorized traffic, protected-only phasing can improve capacity for the protected movements in many cases. However, by adding a protected left-turn phase, the cycle length would be increased and the proportion of green time for critical movements would be decreased. This change in the timing would lead to an increase in the number of vehicle stops for critical movements.

Regarding protected phase for non-motorized traffic, pedestrians and bicycles usually have more crossing opportunities and shorter waits, at the same time increasing vehicular capacity during other phases as conflicting non-motorized traffic is removed. However, an exclusive phase may increase the overall cycle length of the intersection, thus increasing delay for all users.

- **Safety performance**

Usually, treatments can only be expected to be successful when applied to a particular target group of collisions. From this perspective, the installation of protected phasing on one movement should substantially reduce conflicts or collisions involving that particular movement.

For right-hand traffic, left turns are widely recognized as the movement with the highest risk at signalized intersections. Collision types that would particularly benefit from a protected left-turn phase are rear-end and left-turn collisions. Provision of a left-turn lane in conjunction with protected left-turn phasing would appear to provide the most benefit of safety.

Substantial analyses have been done for the safety performance of protected phases of motor vehicles, leading to the following quantitative conclusions.

1. A study of Institute of Transportation Engineers (ITE) of the USA in 2004 indicated that employing split phasing could reduce crashes by 25% (Gan et al., 2005).
2. A study published in 2005 showed a 17% reduction in left-turn crashes with the use of a protected/permissive left-turn phase, and left-turn side-impact crashes was decreased by up to 25% (Lyon et al., 2005).
3. A report in 2008 evaluated crashes involving at least one left-turning vehicle on the treated roadway. Where fully protected phasing was added, left-turn crashes were virtually eliminated, but there was very little change in total crashes (Srinivasan et al., 2008).
4. One study of National Cooperative Highway Research Program (NCHRP) of the USA released in 2004 concluded that left-turn crashes could be reduced by 16% and right-angle crashes by 19% when a protected left-turn phase was provided (Antonucci, 2004).
5. An analysis conducted in 2018 indicated the installation of protected or protected/permissive left-turn phasing could be beneficial in reducing pedestrian crashes at sites with pedestrian-crossing volumes at or exceeding 5500 per day. The protected left-turn phasing evaluation used data from 27 treated sites in Chicago, 7 treated sites in NYC, and 114 treated sites in Toronto. The protected left turn phasing treatment also resulted in decreases of vehicle-vehicle injury crashes by 5% (FHWA, 2018).

At the intersection level, protected phase tends to increase the total number of conflicts, which can result from the following factors.

1. The increase in conflicts comes primarily from rear-end crashes and, under higher flows, from lane-changing maneuvers. As the proportion of green time for through phases is decreased, increasing number of vehicle stops which correlate with the number of rear-end crashes may lead to higher rear-end conflicts.
2. With a greater proportion of vehicles arriving to a standing or dispersing queue, there is an increased tendency of drivers to change to a lane with a shorter queue, despite all lanes having stopped traffic.
3. Vehicle-pedestrian crashes may increase at low pedestrian volumes, as pedestrians may not obey the “do not walk” signal during the protected left-turn phase. Besides, left-turning vehicles are turning left at too high a rate of speed immediately after their protected phase is over and pedestrians begin to cross, thus increasing the probability of vehicle-pedestrian crashes.

Thus, the effect of timing changes and driver behavior should be carefully evaluated before applying protected phase.

Regarding the protected phase of non-motorized traffic, it is noted that the safety benefits of protected-right phase are also covered due to the same starting points. Exclusive right-of-way makes it impossible for the interaction between motorized traffic and non-motorized traffic, thus fundamentally preventing the occurrence of vehicle-pedestrian crashes. Some findings are listed as follows to show the safety performance of protected phase for non-motorized traffic.

1. A paper in 2001 showed that using exclusive pedestrian intervals that stop traffic in all directions had been shown to reduce pedestrian crashes by 50 percent in some locations (i.e., downtown locations with heavy pedestrian volumes and low vehicle speeds and volumes) (Zegeer and Seiderman, 2001).
2. Using the data of all reportable crashes in the five boroughs of New York City in the USA from 1989 to 2008, a study published in 2012 found a 43% reduction in pedestrian crashes from changing the signal phasing from permissive only to protected/permissive or protected only (Bonneseon et al., 2012).
3. A report in 2014 showed that a 47% reduction in pedestrian crashes due to the protected pedestrian phase was found through the comparison of 30 treatment intersections and 493 untreated intersections in New York City in the USA through analysis of a 7-year period covering before and after period (Chen et al., 2014).
4. From data of 1297 signalized intersections involving a total of 2081 pedestrian crashes in 15 U.S. cities, a paper of 2014 showed that exclusive timing was associated with about half of the pedestrian crashes compared with concurrent timing or traffic signals with no pedestrian signals at locations with pedestrian volumes of more than 1200 people/day (Mead et al., 2014).

Another issue concerning the protected non-motorized phase is the utilization of the phase and the compliance of users. Unless a system more heavily penalizes motorists, non-motorized traffic will often have to wait a long time for their exclusive phase. As a result, many pedestrians or bicycles will simply choose to ignore the signal and cross when a gap in traffic occurs.

Summary

This article sheds light on the definition, category, and implementation of protected phase. As a signal treatment, protected phase is to provide exclusive spatial and temporal resources for a specific movement. Though traffic demand and conflictive degree are the main determinants of whether to adopt protected phase, other factors like intersection configuration, field observations of accident experience and sight distance condition also influence the choice of phasing option. For the protected movement, protected phase is the most effective way to improve its safety and progression efficiency, though it may seem kind of simple and crude to some extent. Considering the potential negative effect on the other movements, practitioners should try their best to balance the benefit and cost of applying protected phase through comprehensive actions from data gathering stage, phasing design, public education to field implementation.

Acknowledgment

The authors appreciate the Special Interest Group (SIG) C2 of World Conference on Transport Research Society (WCTRS), especially Zong Tian, Manfred Boltze, Hideki Nakamura, the editors of *Global Practices on Road Traffic Signal Control - Fixed-time Control at Isolated Intersections* for their support in this article.

Biography



Jiarong Yao received the B.S. degree in Traffic Engineering from Tongji University, Shanghai, China, in 2016. She is now a doctoral candidate at College of Transportation Engineering, Tongji University, Shanghai, China. Her main research interest is traffic signal control and Intelligent Transportation Systems.



Yumin Cao received his B.S. degree in Traffic Equipment and Control Engineering from Central South University in 2018. He is now a doctoral candidate at College of Transportation Engineering, Tongji University. His main research interests include traffic flow prediction and Intelligent Transportation Systems.



Keshuang Tang received his doctor's degree in traffic engineering from Nagoya University in 2008. Afterwards, he worked at The University of Tokyo as a postdoctoral research fellow, and then at Tohoku University as a Project Assistant Professor. He was also a visiting scholar of University of California, Berkeley in 2010. He is currently a full professor at College of Transportation Engineering, Tongji University, China. His main research interests include driver behavior, signal control and Intelligent Transportation Systems.

References

- Antonucci, N.D., NCHRP Report 500 Volume 12: A Guide for Reducing Collisions at Signalized Intersections. Transportation Research Board, Washington, DC, 2004.
- Bonneson, J., Pratt, M., Songchitruksa, P., Development of Guidelines for Pedestrian Safety Treatments at Signalized Intersections, Report No. FHWA/TX-11/0-6402-1, Texas Department of Transportation, Austin, TX, 2012.

- Chen, L., Chen, C., Ewing, R., 2014. The relative effectiveness of signal related pedestrian countermeasures at urban intersections-Lessons from a New York City Case Study. *Transport Policy* 32, 69–78.
- FHWA, The Manual on Uniform Traffic Control Devices (MUTCD): 2009 Edition. Federal Highway Administration, U.S. Department of Transportation, 2009.
- FHWA, 2018. Safety Evaluation of Protected Left-Turn Phasing and Leading Pedestrian Interval on Pedestrian Safety. Federal Highway Administration, U.S. Department of Transportation.
- FSV, RVS 05.04.32 Planen von Verkehrslichtsignalanlagen (Planning of traffic signals), Oktober 1998 (in German), 1988.
- FGSV, Richtlinien für Lichtsignalanlagen – RiLSA (German Standard for Traffic Signals). Edition 2015. FGSV-Verlag, Cologne (in German), 2015.
- Gan, A., Shen, J., Rodriguez, A., Update of Florida Crash Reduction Factors and Countermeasures to improve the Development of District Safety Improvement Projects. Florida Department of Transportation, 2005.
- Japan Society of Traffic Engineers (JSTE): Manual on Traffic Signal Control, Revised Edition. (in Japanese), 2006.
- Lyon, C., Haq, A., Persaud, B.N., Kodama, S.T., Development of Safety Performance Functions for Signalized Intersections in a Large Urban Area and Application to Evaluation of Left Turn Priority Treatment. 2005 TRB 84th Annual Meeting: Compendium of Papers CD-ROM, 05(2192), Washington, D.C., 2005.
- Mead, J., Zegeer, C., Bushell, M., Evaluation of Pedestrian-Related Roadway Measures: A Summary of Available Research, Pedestrian and Bicycle Information Center, Chapel Hill, N.C. [Online], 2014. Available: http://www.pedbikeinfo.org/cms/downloads/PedestrianLitReview_April2014.pdf.
- National Police Agency of Japan, Guideline of Traffic Signal Installation. (in Japanese), 2015.
- Rodegerdts, L.A., Nevers, B., Robinson, B., et al., Signalized Intersections: Informational Guide. Guides to Information, 2004.
- Srinivasan, R., Carter, D., Persaud, B., Eccles, K., Lyon, C., Evaluation of the Safety Effectiveness of Selected Treatments at Urban Signalized Intersections. Transportation Research Record: Journal of the Transportation Research Board, 2056, 70-76, Transportation Research Board, Washington, DC, 2008.
- Tang, K., Boltze, M., Nakamura, H., Tian, Z., 2019. Global Practices on Road Traffic Signal Control. Elsevier.
- The Ministry of Public Security of the People's Republic of China, Specifications for road traffic signal setting and installation, China. (in Chinese), 2016.
- The Ministry of Public Security of the People's Republic of China, Road Traffic Signal Control Mode, China. (in Chinese), 2015.
- Transportation Research Board, Highway Capacity Manual. Transportation Research Board, Washington, DC, USA, 2010.
- Urbanik, T., Tanaka, A., Lozner, B., Lindstrom, E., Lee, K., Quayle, S., Beaird, S., Tsoi, S., Ryus, P., Gettman, D., Sunkari, S., Balke, K., Bullock, D.M., 2015. Signal Timing Manual: Second Edition.. Transportation Research Board, Washington, D.C., USA.
- Zegeer, C.V., Seiderman, C.B., Designing for Pedestrians. In 2001 Compendium of Technical Papers, ITE, 71st Annual Meeting, Chicago, IL. ITE, Washington, DC, 2001.

Further Reading

- Federal Highway Administration, US Department of Transportation (FHWA, 2008): Traffic Signal Timing Manual.
- Federal Highway Administration, US Department of Transportation (FHWA 2016): 2015 Status of the Nation's Highways, Bridges, and Transit: Conditions & Performance.
- FGSV, Hinweise zum Qualitätsmanagement an Lichtsignalanlagen: H QML, (Advice on quality management at traffic signals), FGSV Verlag. (in German), 2014.
- Traffic Management Research Institute of the Ministry of Public Security, Report on Construction and Application of National Public Security Traffic Control System. China. (In Chinese), 2015.
- Japan Society of Traffic Engineers (JSTE) (forthcoming 2018): Basic Manual on Planning and Design of At-Grade Intersections and on Traffic Signal Control. (in Japanese).
- Australasian Road Transport and Traffic Agencies (AUSTROADS, 2003): Guide to Traffic Engineering Practice Series: Traffic Signals, Australia.
- UK DOT, Department of Transport, the United Kingdoms: Traffic Advisory Leaflet: General Principles of Traffic Control by Light Signals, the United Kingdoms, 2016.
- AFNOR, Régulation du trafic routier-Signaux lumineux de circulation routière-Caractéristiques techniques-R25. Norm. Fr. NF P99-200/A1 Janvier 2016. (in French), 2016.

Permitted Phase

Yumin Cao, Jiarong Yao, Keshuang Tang, Tongji University, Shanghai, China

© 2021 Elsevier Ltd. All rights reserved.

Definition of Permitted Phase	196
Phase, Signal Group, and Stage	196
Types of Permitted Phase	196
When and Where to Implement a Permitted Phase	197
Operational and Safety Performance of Permitted Phase	200
Summary	201
Acknowledgment	201
Biographies	201
References	202
Further Reading	202

Definition of Permitted Phase

Phase, Signal Group, and Stage

- A traffic signal phase, or a **phase** for short, is a timing process within the signal controller that facilitates serving one or more movements at the same time for one or more modes of users. A **signal group** refers to a group of one or more traffic flows which always receive identical signal light indications. A **stage** is a part of the cycle during which a particular set of compatible streams receives green while all others have to wait (Tang et al., 2019). Fig. 1 provides a description for these three terms in the scope of an intersection for better understanding. Note that the term “phase” used here should not be confused with the term “NEMA phase” widely used in North America, which is a green time interval accommodating only one movement, that is, a through movement or a left-turn movement (Transportation Research Board, 2010).
- Permitted phase **Permitted phase** (also known as permissive phase, permitted-only phase or permissive-only phase) is a kind of phase that gives proceeding permission to certain movements during a time in the cycle when the movement does not have the priority. And corresponding road users are going to proceed when there are available gaps in the conflicting flow after yielding to them. In contrast, **protected phase** (also known as protected-only phase) refers to providing a separate phase for protected movements with no conflicting traffic flows. And corresponding road users can proceed safely without any conflicts. As a compromise, **protected/permitted phase** is a compound phase that displays the permitted phase after the protected phase, while **permitted/protected phase** is a compound phase that displays the permitted phase before the protected phase. Note that these terms are more commonly used for motor vehicles, and could rarely be seen for nonmotorized traffic. Therefore, this chapter mainly discussed the permitted phase for motor vehicles.

Types of Permitted Phase

For signalized intersections, there are mainly two kinds of permitted phase according to the type of turning, namely permitted left-turn phase and permitted right-turn phase. To avoid confusion, in the following text the permitted phases for motor vehicles are introduced in the background of **right-hand traffic**.

- **Permitted left-turn phase and its displays** Permitted left-turn phase permits the left-turn vehicles of subjective approach to proceed after yielding to conflicting vehicular (opposing through vehicles) and pedestrians (parallel pedestrians) traffic before completing the turn. Left-turn vehicles could proceed when there are acceptable gaps between two consecutive conflicting users' movements. Typical signal indications of permitted left-turn movement across countries are shown in Fig. 2. Permitted Left-turns are often made on a circular green signal indication, a flashing left-turn yellow arrow signal indication, or a flashing left-turn red arrow signal indication after yielding to pedestrians or opposing traffic. During a permitted left-turn movement, the signal faces for through traffic on the opposing approach simultaneously display green or steady yellow signal indications (Tang et al., 2019).
- **Permitted right-turn phase and its displays** In general, the right-turn movement is served concurrently with adjacent through movement (i.e., by default permitted). Permitted right-turn phase permits the right-turn vehicles of subjective approach to proceed after yielding to conflicting pedestrians (parallel pedestrians) traffic before completing the turn. Right-turn vehicles could proceed when there are acceptable gaps in the pedestrian flows. Typical signal indications of permitted right-turn movement across countries are shown in Fig. 3. Permitted right-turns are often made on a circular green signal indication, a flashing right-turn yellow arrow signal indication, or a flashing right-turn red arrow signal indication after yielding to conflicting pedestrians (Tang et al., 2019).

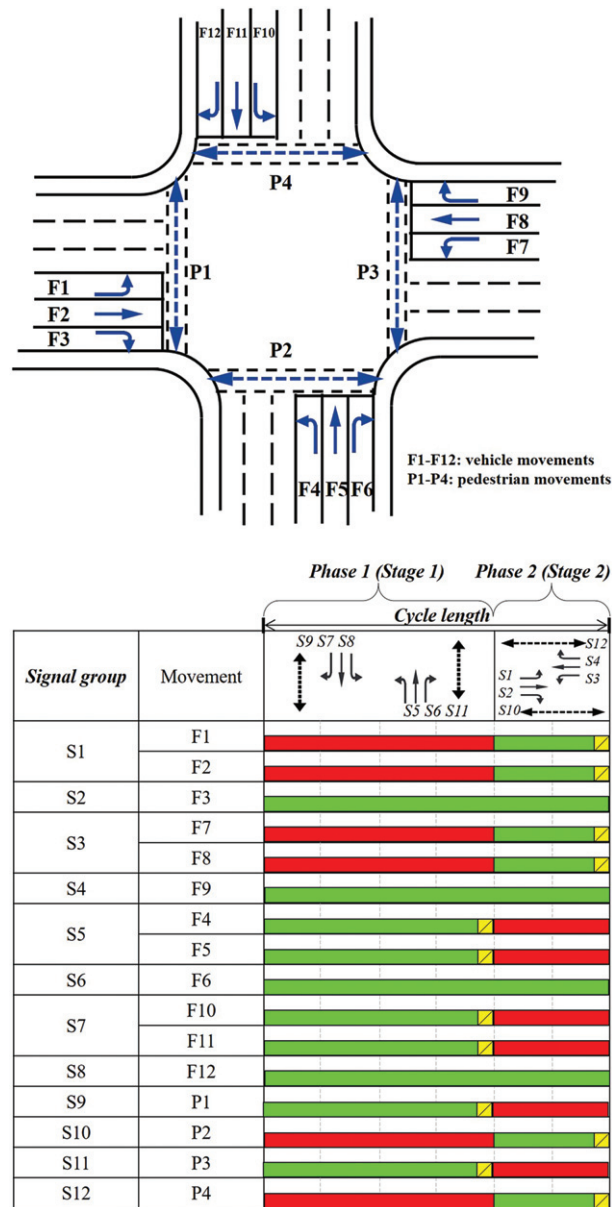
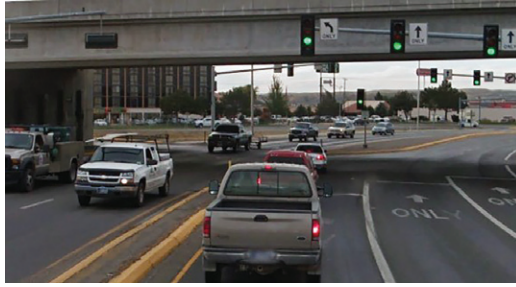


Figure 1 Sketch map for phase, signal group and stage.

When and Where to Implement a Permitted Phase

- Permitted left-turn phase** The type of signal phasing used for a left-turn movement directly affects safety and operational performance of the turn. In general, less-restrictive phasing schemes are preferable where appropriate because they result in lower delay to all users of the intersection (Koonce and Rodegerdts, 2008). A variety of guidelines exist that indicate conditions where the benefits of left-turn phases typically outweigh its adverse impact to overall intersection operation (Agent et al., 1995; Upchurch, 1986). Many of these guidelines indicate that a left-turn phase can be justified based on consideration of several factors that ultimately tie back to the operational or safety benefits derived from the phase. These factors mainly include: left-turn and opposing through volumes, number of opposing through lanes, cycle length, speed of opposing traffic, sight distance, and crash history. Generally, protected or semi-protected left-turn phasing (protected-permitted, permitted-protected, or protected only) should be considered if any one of the following criteria is satisfied, and permitted left-turn phasing may be considered at sites that do not satisfy any of these criteria (Rodegerdts et al., 2004):

1. A minimum of two left-turning vehicles per cycle and the product of opposing and left-turn hourly volumes exceeds certain value (45,000 per opposing lane).
2. The left-turning movement crosses three or more lanes of opposing through traffic.



(A)



(B)



(C)



(D)

Figure 2 Examples of permitted left-turn phase display: (A) United States, (B) Germany, (C) China, and (D) Australia.



(A)



(B)



(C)



(D)

Figure 3 Examples of permitted right-turn phase display: (A) United States, (B) Germany, (C) China, and (D) Australia.

3. The posted speed of opposing traffic exceeds 70 km/h.
4. Recent crash history for a 12-month period indicates 5 or more left-turn collisions that could be prevented by the installation of left-turn signals.
5. Sight distances to oncoming traffic are less than the standard minimum distances.

6. The intersection has unusual geometric configurations, such as five legs, when an analysis indicates that left-turn or other special traffic signal phases would be appropriate to provide positive direction to the motorist.
7. An opposing left-turn approach has a left-turn signal or meets one or more of the criteria here.

- **Permitted right-turn phase**In many countries, such as United States, Germany, and China, right-turning vehicles commonly share a lane with through vehicles and are realized together with the through vehicles, pedestrians and cyclists at the intersection. Therefore, right-turn phasing would by default be permitted phase unless the following conditions are satisfied:

1. The subject right-turn movement is served by one or more exclusive right-turn lanes.
2. The right-turn volume is high (e.g., 300 veh/h or more).
3. A left-turn phase is provided for the complementary left-turn movement.
4. U-turns from the complementary left-turn are prohibited.

If the aforementioned conditions are satisfied, then the appropriate operational mode can be determined. If the through movement phase for the subject intersection approach serves a pedestrian movement, then the right-turn phasing should operate in the protected-permitted mode, and the permitted operation would occur during adjacent through movement phase. If the through movement phase for the subject intersection approach does not serve a pedestrian movement, then the right-turn phasing should operate in the protected-only mode during both the adjacent through movement phase and the complementary left-turn phase (left-turn phase of the neighboring or intersecting approach).

Operational and Safety Performance of Permitted Phase

- **Operational performance**

Permitted phase provides the most efficient green time allocation to users within intersection, but the efficiency is dependent on the availability of gaps in the conflicting traffic. From the perspective of traffic operations, provision of permitted left-turns during the through green will always be beneficial regardless of the type of signal control or left-turn sequence pattern. Only in situations where safety concerns are an overwhelming influence should permitted left turns be prohibited.

- Trial for four intersections in 1979 demonstrated that the use of permitted-turn phasing resulted in a 50% reduction in left-turn delay and a 24% reduction in total delay compared to exclusive left-turn phasing (protected only phase). Meanwhile, questionnaire responses showed that over 90% of drivers were in favor of this type of signal ([Agent, 1981](#)).
- Study in 1981 found that permitted-only left-turn phase could reduce left-turn delays by approximately 30% and as much as 50%. Left-turn delays versus left-turn volumes were found to have a linear relationship up to capacity conditions ([Rocciola, 1981](#)).
- A study in 1995 included 264 intersections and data obtained from 518 approaches. These intersections were scattered across the United States with approaches included in 52 counties. At these intersections, protected-only left-turn phasing resulted in high left-turn delays while permitted phasing can reduce these delays. For approaches with one opposing lane, the use of permitted or protected/permitted left-turn phasing produced similar left-turn delays; for approaches with two opposing lanes, protected/permitted phasing produced lower delays than permitted. Moreover, the study pointed out that the desired left-turn phasing for both one and two-lane approaches is protected/permitted phase since it produces similar delays as permitted while it provides some level of safety for left-turning vehicles, and protected-only left-turn phasing should be considered only for locations with a high number of left-turn accidents or where there is a high left-turn accident potential ([Agent et al., 1995](#)).
- For permitted right-turns, right-turn-on-reds (RTORs) are developed as a fuel- and time-saving measure, vehicles could proceed to turn right without having to wait for green signal. In the United States, western states have allowed it for more than 50 years, and eastern states amended their traffic laws to allow it in the 1970s as a fuel-saving measure in response to motor fuel shortages in 1973. And all states have allowed right turns on red since 1980.

- **Safety performance**

Permitted phase can have an adverse effect on safety in some situations, such as applying permitted left-turn phase when the left-turning vehicles' view of conflicting traffic is restricted or when adequate gaps in traffic are not present.

- Trial for four intersections in 1979 demonstrated that the permitted left-turn phasing resulted in an increase in left-turn accidents. However, other accident types, such as rear-end accidents, did not increase. The number of left-turn accidents decreased as drivers became more familiar with the signals ([Agent, 1981](#)).
- Study in 1981 found that traffic conflicts were higher for permitted phase than protected phase while accident experience were not increased significantly, and the critical number of accidents at an permitted only phase has been determined to be 10 per year ([Rocciola, 1981](#)).
- For the study in 1995, considering all approaches, 69% had an average of 0.5 left-turn accidents per year or less. This percentage was 93% for protected-only phasing locations compared to 40% for protected/permitted phasing and 71% for permitted phasing approaches. Only 8% had an average of over two accidents per year with no protected-only approaches having this

number of accidents. This percentage was 18% for protected/permitted phasing and 7% for permitted phasing approaches. Besides, the average was only 0.20 left-turn accidents per approach per year for approaches with protected-only phasing and increased to 1.32 for protected/permitted phasing and 0.50 for permitted approaches (Agent et al., 1995).

- A surrogate safety assessment indicated that total number of conflicts for permitted left turn is significantly more than that of protected left turn; Protected left turn has less crossing conflicts but more rear-end and lane-change conflicts than permitted left turn. The comparison results for all other safety measures show no distinct safety preference between protected left turn and permitted left turn (Gettman et al., 2008). The following cited references (Gettman et al., 2008) are not included in the reference list. Either include the cited references with complete details in the reference list or delete their citations from the text..
- RTORs can also be detrimental for pedestrians and cyclists. A 1981 US Department of Transportation study determined that the frequency of motor vehicle collisions with bicyclists and pedestrians when the vehicle was turning right increased significantly after the adoption of RTOR. It also demonstrated that estimates of the magnitude of the increases ranged from 43% to 107% for pedestrian accidents and 72% to 123% for bicyclist accidents (Preusser et al., 1982). A 1984 study found that where RTOR was allowed, all right-turning crashes increase by about 23%, pedestrian crashes by about 60%, and bicyclist crashes by about 100% (Zador, 1984). A 1993 study also concluded that RTOR increased crashes for pedestrians and cyclists, by 44% and 59% respectively (Dussault, 1993).

In general, the operational benefits of permitted phase, for both left-turn and right-turn, are proved in early days, great delay reduction had been shown by replacing protected-only phases with permitted phase. The reductions in delay are encouraging toward additional but discriminating use of permitted left-turn phasing. Meanwhile, the accidental potential is indeed increased which limits its use to intersections that has suffered from accident problem. More recently, with increasing vehicles on the road, researchers have also increasingly investigated the gap acceptance theories, and the usage of left-turn permitted-only phase is decreasing for safety considerations, while no-right-turn-on-red has also been attempted in certain intersections to protect pedestrians and cyclists. Protected/permitted left-turns and prohibited/permitted right-turns are becoming potential alternatives.

Summary

This article focuses on the definition, types, implementations as well as operational and safety performances of permitted phase. As a widely-used signal treatment, permitted phase provides certain road users of permission to proceed without right-of-way, mainly including left-turn and right-turn motor vehicles. It is developed as an effective measure to decrease the vehicle delays and increase capacity of intersections only when safety issues do not occur. However, increased conflicts are also bounded for permitted phase and so as accidental potentials. Though several investigations demonstrated a decreased rear-end crashes and insignificant change in accidents, the implementation of permitted phases shall be cautious and take varied factors into consideration.

Acknowledgment

The authors appreciate the Special Interest Group (SIG) C2 of World Conference on Transport Research(WCTR), especially Zong Tian, Manfred Boltze, Hideki Nakamura, the editors of *Global Practices on Road Traffic Signal Control* for their support in this article.

Biographies



Yumin Cao received his B.S. degree in Traffic Equipment and Control Engineering from Central South University in 2018. He is now a doctoral candidate at College of Transportation Engineering, Tongji University. His main research interests include traffic demand estimation, network modeling, and Intelligent Transportation Systems (ITS).



Jiarong Yao received his B.S. degree in Traffic Engineering from Tongji University, Shanghai, China, in 2016. She is now a doctoral candidate at College of Transportation Engineering, Tongji University, Shanghai, China. Her main research interest is traffic signal control and Intelligent Transportation Systems.



Keshuang Tang received his doctor's degree in traffic engineering from Nagoya University in 2008. Afterwards, he worked at The University of Tokyo as a postdoctoral research fellow, and then at Tohoku University as a Project Assistant Professor. He was also a visiting scholar of University of California, Berkeley in 2010. He is currently a full professor at College of Transportation Engineering, Tongji University, China. His main research interests include driver behavior, signal control and Intelligent Transportation Systems.

References

- Agent, K., 1981. An evaluation of permissive left-turn phasing. *Ite J. Inst. Transport. Eng.* 51 (12), 16–20.
- Agent, K.R., Stamatiadis, N., Dyer, B., Guidelines for the Installation of Left-turn Phasing, 1995.
- Dussault, C., Safety Effects of Right Turn on Red: A Meta-Analysis. In *Proceedings of the Canadian Multidisciplinary Road Safety Conference VIII*. Saskatoon, Saskatchewan, June 1993.
- FHWA, The Manual on Uniform Traffic Control Devices (MUTCD): 2009 Edition. Federal Highway Administration, U.S. Department of Transportation, 2009.
- Gettman, D., Pu, L., Sayed, T., Shelby, S., Siemens, I.T., 2008. Surrogate safety assessment model and validation (No. FHWA-HRT-08-051). United States. Federal Highway Administration. Office of Safety Research and Development.
- Koonce, P., Rodegerdts, L., . Traffic signal timing manual (STM), 2008.
- Preusser, D.F., Leaf, W.A., DeBartolo, K.B., Blomberg, R.D., Levy, M.M., 1982. The effect of right-turn-on-red on pedestrian and bicyclist accidents. *J. Safety Res.* 13 (2), 45–55.
- Rocciola, J., 1981. An Evaluation of Exclusive and Exclusive-Permissive Left Turn Signal Phasing. Maryland Department of Transportation, Baltimore.
- Rodegerdts, L.A., Nevers, B.L., Robinson, B., Ringert, J., Koonce, P., Bansen, J., Suggett, J., Signalized intersections: informational guide, 2004.
- Tang, K., Boltze, M., Nakamura, H., Tian, Z., 2019. Global Practices on Road Traffic Signal Control: Fixed-time Control at Isolated Intersections. Elsevier.
- Transportation Research Board, Highway capacity manual (HCM). Washington, DC, USA, 2010.
- Upchurch, J.E., 1986. Guidelines for selecting type of left-turn phasing. *Transport. Res. Rec.* 1069, 30–38.
- Urbanik, T., Tanaka, A., Lozner, B., Lindstrom, E., Lee, K., Quayle, S., Gettman, D., Signal Timing Manual. Transportation Research Board, Washington, DC, 2015.
- Zador, P.L., 1984. Right-turn-on-red laws and motor vehicle crashes: a review of the literature. *Accid. Anal. Prevent.* 16 (4), 241–245.

Further Reading

- Federal Highway Administration, US Department of Transportation (FHWA), 2015 Status of the Nation's Highways, Bridges, and Transit: Conditions & Performance, 2016.
- FGSV, Hinweise zum Qualitätsmanagement an Lichtsignalanlagen: H QML, (Advice on quality management at traffic signals), FGSV Verlag, 2014.
- Traffic Management Research Institute of the Ministry of Public Security, Report on Construction and Application of National Public Security Traffic Control System. China. (In Chinese), 2015.
- Japan Society of Traffic Engineers (JSTE), Basic Manual on Planning and Design of At-Grade Intersections and on Traffic Signal Control. (in Japanese), 2018.
- Australasian Road Transport and Traffic Agencies (AUSTROADS), Guide to Traffic Engineering Practice Series: Traffic Signals, Australia, 2003.
- Department of Transport, the United Kingdoms (UK DOT), Traffic Advisory Leaflet: General Principles of Traffic Control by Light Signals, the United Kingdoms, 2006.
- Akçelik, R., A review of gap-acceptance capacity models. In 29th Conference of Australian Institutes of Transport Research (CAITR), University of South Australia, Adelaide, Australia, 2007.
- Qi, Y., Guoguo, A., 2017. Pedestrian safety under permissive left-turn signal control. *Int. J. Transport. Sci. Technol.* 6 (1), 53–62.
- Srinivasan, R., Baek, J., Smith, S., Sundstrom, C., Carter, D., Lyon, C., Hamidi, A., Evaluation of Safety Strategies at Signalized Intersections. Nchrp Report, 2011.
- Stevanovic, A., Radivojevic, D., Manual on Performance of Traffic Signal Systems: Assessment of Operations and Maintenance, 2017.

Hook Turns: Implementation, Benefits, and Limitations

Sara Moridpour, Amir Falamarzi, Civil and Infrastructure Engineering Discipline, School of Engineering, RMIT University, Melbourne, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	203
Background	203
Different Types of Hook Turns in Melbourne	205
Hook Turn for Cyclists and Motorcyclists	207
Impact on Traffic Safety	208
Traffic Management Improvement	210
Limitations and Disadvantages	211
Summary	211
References	212

Introduction

In recent decades, light rail transport systems and trams have drawn the attention of traffic engineers and urban planners due to their advantages as a popular mode of public transport. The light rail transport is more accessible than heavy rail transport as they share the road with other road users. Tram tracks are usually located in the middle of the streets. Thus, the interaction between trams and different modes of transport is always a challenge for traffic engineers and safety experts, especially at busy intersections. This problem is more common when the allocated space for the right-turn traffic is limited. While waiting for a safe gap in opposing traffic to make a turn, the right-turn traffic can block the through the movement of trams. Hook turns have been introduced as a novel approach to minimize the impact of right-turn traffic on not only the tram movements but also the other modes of transportation. Hook turns have been implemented to buses, bicycles and motorcycles to improve their safety.

In this article, the history of hook turns, and the background of hook turn implementation will be described first. Then, by providing examples, different types of hook turns in metropolitan Melbourne will be presented followed up by the examples of implementing hook turns for bicycles and motorcycles. It is followed by investigating the impact of hook turns on traffic safety. Afterwards, different illustrations of the traffic management improvements resulted from the implementation of the hook turns will be presented. Then, the limitations and disadvantages of hook turns will be discussed. Finally, the summary of the article will be presented.

Background

Hook turn is a traffic management approach that is originally implemented in the Melbourne Central Business District (CBD) more than 60 years ago. The primary purpose of introducing this approach was to clear the CBD road network for the movement of trams. Tram system in Melbourne is one of the most crucial components of the public transport system in inner Melbourne. Trams are the second most popular mode of public transport in metropolitan Melbourne and comprise the highest passenger demand after the heavy rail system.

The Melbourne tram network is one of the largest tram networks in the world and consists of 250 kilometers of double tracks (25 routes and more than 1,700 stops). In 2019, more than 205 million trips were served by the Melbourne tram network. **Figure 1** shows the map of the Melbourne tram network.

At present, the Melbourne road network comprises 49 hook turns (Currie and Reynolds, 2011; Falamarzi et al., 2018). **Figure 2** shows an aerial view of an intersection with a hook turn in Melbourne CBD. It must be noted that hook turns are also used by cyclists and motorcyclists at intersections to make a right turn. Hook turn as a safe approach may help these road users avoid crossing multiple lanes for reaching the middle of the road and the right turn lane.

Hook turns are not limited to Melbourne's tram network. Some bus routes have adopted the hook turn in Australia and worldwide. Hook turns have been introduced for buses in Adelaide, Australia (at two intersections), and in Shanghai, China. Hook turns have been used for different road users in Illinois, United States of America, China, and Taiwan (Hounsell and Yap, 2013). The idea behind introducing the hook turn was originated from the conflict between trams and the traffic intending to make right turns at intersections. The right turn traffic can potentially embed the movement of trams while looking for a safe and sufficient gap in opposing traffic. Instead on implementing hook turns, there are two potential solutions to deal with this problem and avoid blocking trams, including (1) restricting the right turn traffic movements, or (2) modifying the traffic signals to accommodate a protected right turn phase and assigning an exclusive traffic signal phase to right turn traffic which means an additional signal phase at the intersections. The second solution brings a significant disadvantage to the trams, as it is imposing



Figure 1 Map of the Melbourne tram network (Yarra Trams, 2020).



Figure 2 Intersection of La Trobe Street and Elizabeth Street, Melbourne, Australia (Google, 2020).

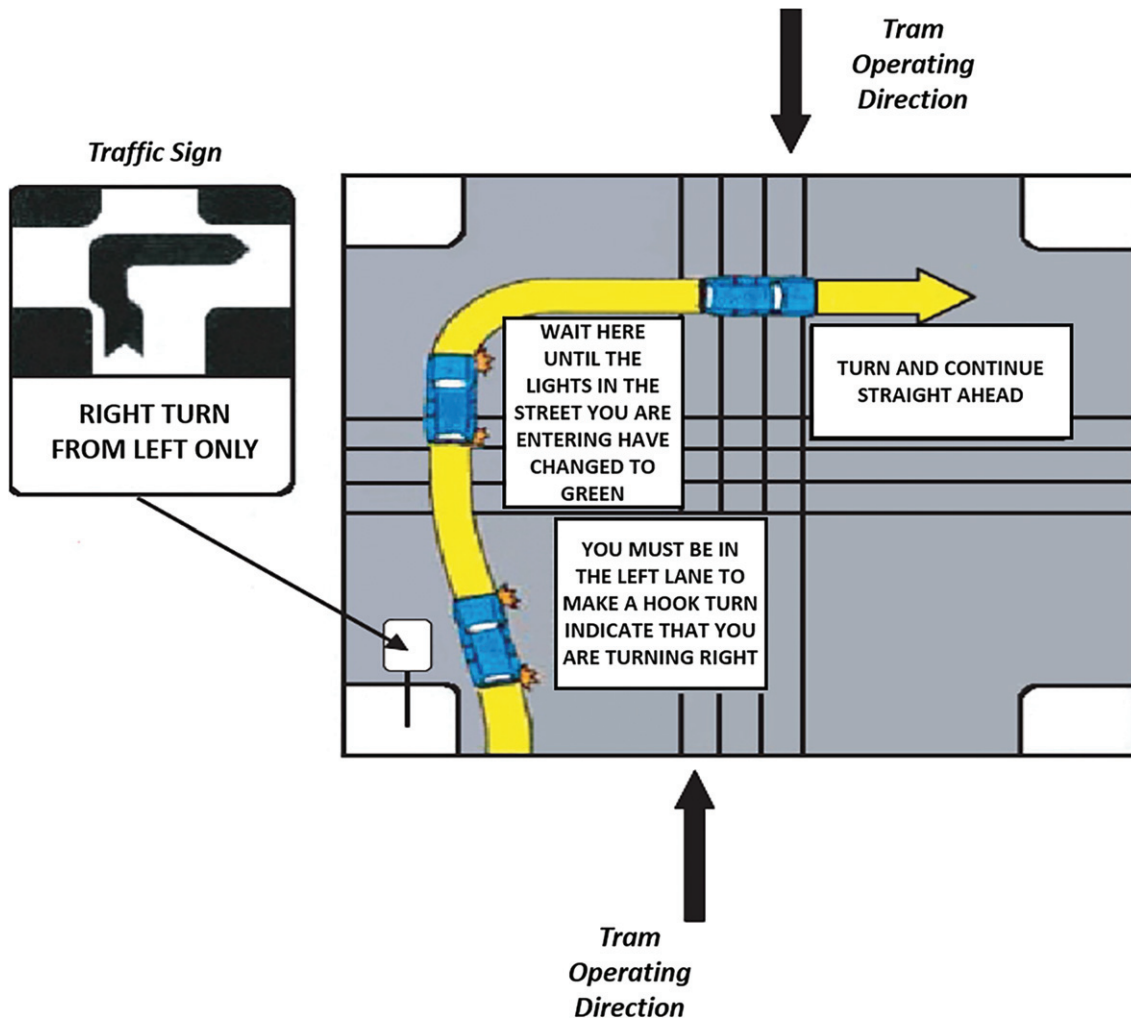


Figure 3 Hook turn function in a left-hand traffic context (Currie and Reynolds, 2011).

additional delays to tram systems, and on a broader aspect, reducing the reliability and attractiveness of this mode of public transport. Both solutions result in reducing the capacities of intersections. By using the hook turns, there is no need for extra lanes dedicated for turning vehicles (Currie and Shalaby, 2007).

According to VicRoads Guidelines (VicRoads, 2020), a few steps should be undertaken to safely perform a hook turn (Figure 3). First, vehicles need to approach the hook turn intersection from the far left lane and turn on the vehicle indicator for turning right. Then, the vehicles need to move toward the most left side of the intersection and keep clear of any pedestrian crossings. The right-turning vehicles need to stay stopped and wait until the traffic lights for the road they are planning to turn to, become green.

Different Types of Hook Turns in Melbourne

There are four main types of hook turns implemented in the Melbourne metropolitan area, which have been integrated with tram movements. In Figure 4, typical hook turn designs along with non-hook turn designs are illustrated. In Figure 5, different examples of hook turn implementations in metropolitan Melbourne are illustrated.

Type 1 is related to the typical hook turn implementation within the Melbourne CBD, which includes two traffic lanes, the tram right-of-way and a tram stop or a safety zone stop. The safety zone is located between the through lane and the tram track. Safety zone stops can be either at the same level with the streets or be raised. Both stops are designed for improving the safety of pedestrians boarding or alighting a tram. They can protect pedestrians from the general traffic by a railing. An example of this type of hook turn has been implemented at the intersection of La Trobe Street and Swanston Street in Melbourne CBD.

Type 2 is similar to Type 1 but without a tram stop or traffic island. An example of this type of hook turn has been implemented at the intersection of La Trobe Street and Elizabeth Street in Melbourne CBD.

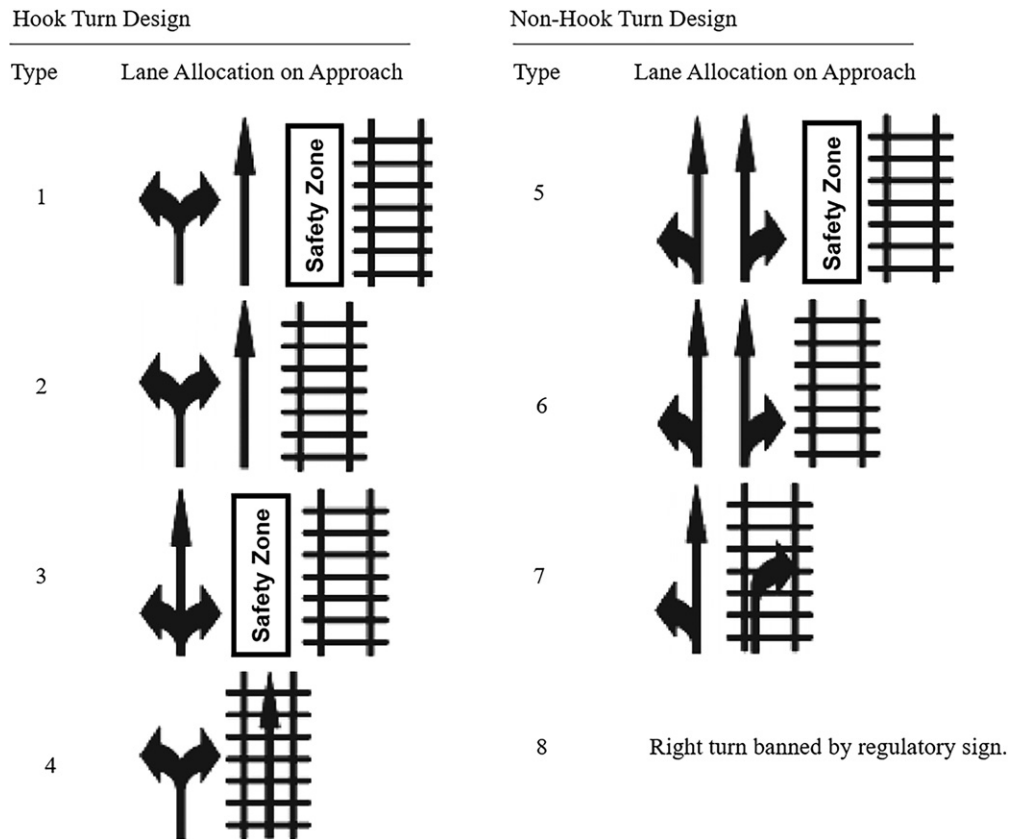


Figure 4 Different types of hook turns in metropolitan Melbourne (Currie and Reynolds, 2011).



Figure 5 Different types of hook turn implementations (Types 1, 2, 3, and 4) in metropolitan Melbourne, (Google, 2020).

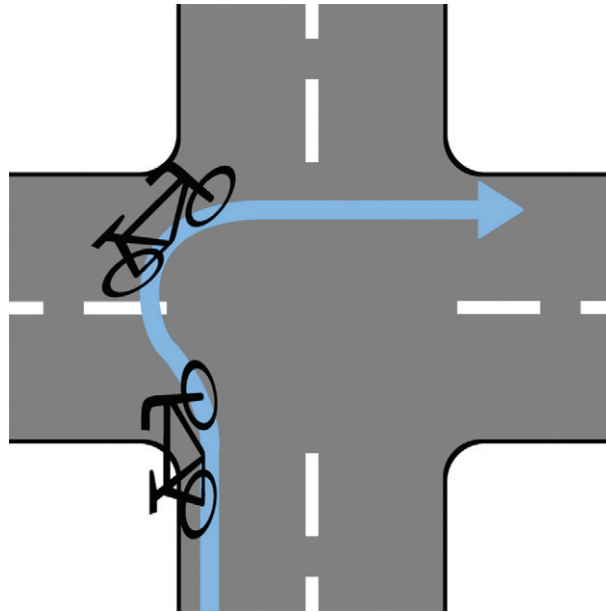


Figure 6 Performing a hook turn at an unsignalised intersection by a cyclist (TMR, 2020).

Type 3 is similar to Type 1 but is used at locations without sufficient space for two lanes on the approach to the intersection. Vehicles still undertake a hook-turn movement, with through vehicles overtaking the right-turning cars on the right within the intersection. An example of this type of hook turn has been implemented at the intersection of Collins Street and William Street in Melbourne CBD.

Type 4 has been implemented on Clarendon Street, a strip shopping center outside the CBD area. This treatment is accompanied by curb-side tram stops.

Hook Turn for Cyclists and Motorcyclists

In addition to cars, bicycles, and motorcycles use hook turns in some countries including Australia, the Netherlands, New Zealand, and Taiwan. In these countries, hook turn as a practical traffic management approach has provided a safe turning movement by reducing the exposure of cyclists and motorcyclists to other vehicles and multi-lane traffic flow. In this section, different examples of the hook turn approach designed for cyclists and motorcyclists are presented.

For instance, in Queensland, Australia, the transport authorities have specified the hook turn rules for cyclists at intersections with and without traffic lights. If the intersection has a traffic light, the right-turning cyclists need to keep to the far left side when approaching the intersection and then stay at the left-hand side of the intersection. They need to wait and give way to through traffic and move forward across the intersection when the road is clear (at unsignalized intersections) or when the signal light turns green (at signalized intersections). The performance of hook turns at an unsignalized intersection has been depicted in [Figure 6](#). In some intersections, a hook turn storage box ([Figure 7](#)) has been implemented. This area provides a safe, dedicated area for right-turning cyclists to stop and wait (TMR, 2020).



Figure 7 Hook turn storage box at the Whiteleigh-Lincoln intersection, Christchurch, New Zealand (Cycling in Christchurch, 2020).

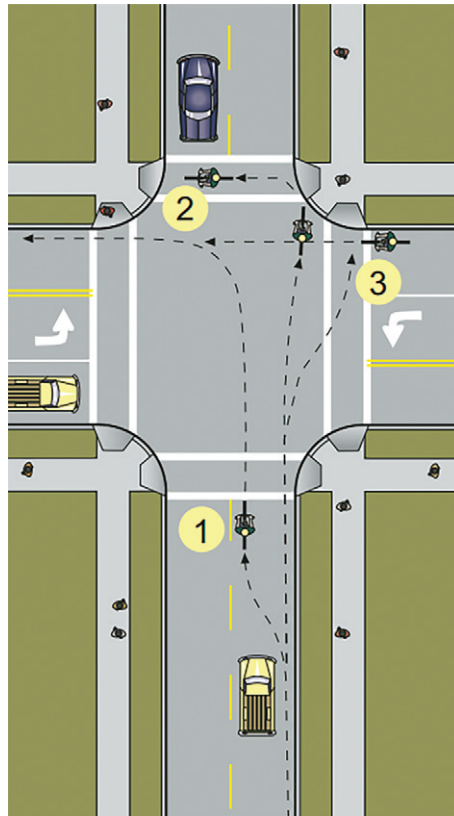


Figure 8 A schematic view of making a left turn at an intersection in British Columbia (ODOT, 2017).

In British Columbia, Canada, where vehicles move on the right side of the road, cyclists have three main methods of performing a left turn maneuver at signalized intersections (Figure 8). In the first method (movement 1), similar to cars, cyclists make a left turn at intersections from the middle of the road and the far left side of the intersection. In this method, cyclists and other vehicles merge while turning left. In the second method (movement 2), the cyclists can use the pedestrian crossings, dismount and wait for the green walk signal at the crossing point together with the pedestrians. The third method (movement 3) is similar to approaches implemented in Australia and New Zealand (hook turn is called Perimeter style in British Columbia, Canada). In the third method, the cyclists move to the far right of the intersection and wait at the right side of the intersection, yield to pedestrians in a crosswalk and move forward when the traffic light turns green (ODOT, 2017).

Taiwan is a country, which ranks high in terms of the proportion of motorcycles to the other motorized vehicles. According to the crash statistics data, the majority of motorcycle accidents are related to the right of way accidents where motorcyclists violate the right of way of approaching automobiles. To reduce the negative impact of motorcycles, hook turns have been introduced in Taiwan as an engineering solution. As seen in Figure 9, the Hook Turn Area (HTA) is located in the middle of two parallel pedestrian crosswalks, and motorcyclists are required to perform a hook turn. The motorcyclists violating the hook turn rule will be fined. This measure is basically applied on roadways with two lanes or more (roadways wider than 8 m). This measure can reduce the risk of sideswipe and rear-end collisions which are common in left-turn crashes involving motorcycles and automobiles (Pai et al., 2013).

Similar to hook turns, two-stage left turn boxes have been implemented at several signalized intersections in Taiwan (Figure 10). The left turn boxes are installed after the pedestrian crossings. The rationale behind this approach is to provide a safe approach to make a left turn for motorcyclists at multi-lane signalized intersections (Le and Nurhidayati, 2016).

Impact on Traffic Safety

According to VicRoads investigations, a hook turn intersection has a lower rate of accidents and can operate better in terms of traffic safety compared to intersections with no hook turns (Marti et al., 2016). Hook turns can reduce traffic conflict points. Conflict points are the intersecting spots between the paths of different traffic movements. The conflict points can be classified into merging, diverging and crossing types. In this regard, hook turns can eliminate the conflict points between turning traffic, through traffic and tram movement, which results in the improvement of the intersection safety. The collision between trams and vehicles turning across the tramways is one of the common accident types at non-hook turn intersections, accommodating tram traffic (Figure 11). As tram

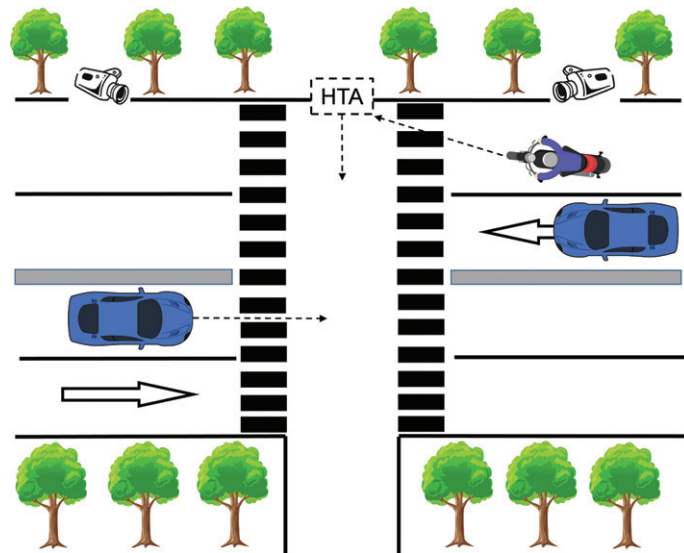


Figure 9 A schematic depiction of HTA for making a left turn (Pai et al., 2013).



Figure 10 An example of two-stage left turn box in Taiwan (YouTube, 2014).



Figure 11 A tram and car accident in Amsterdam (Flickr, 2020).

cars are much more massive than standard passenger cars, the severity of the collisions between trams and passenger cars is severe (Marti et al., 2016).

There are two main advantages of using hook turns and consequently removing the conflict points between through traffic and right-turn traffic. These advantages include the removal of right-turn head-on crashes and right turn-pedestrian crossing crashes (Currie and Smith, 2006; Currie and Reynolds, 2011; Naznin et al., 2018). The right turn head-on crashes are crashes that occur between right-turning traffic and the oncoming opposing traffic. This type of crash is dangerous due to the nature of the accident and

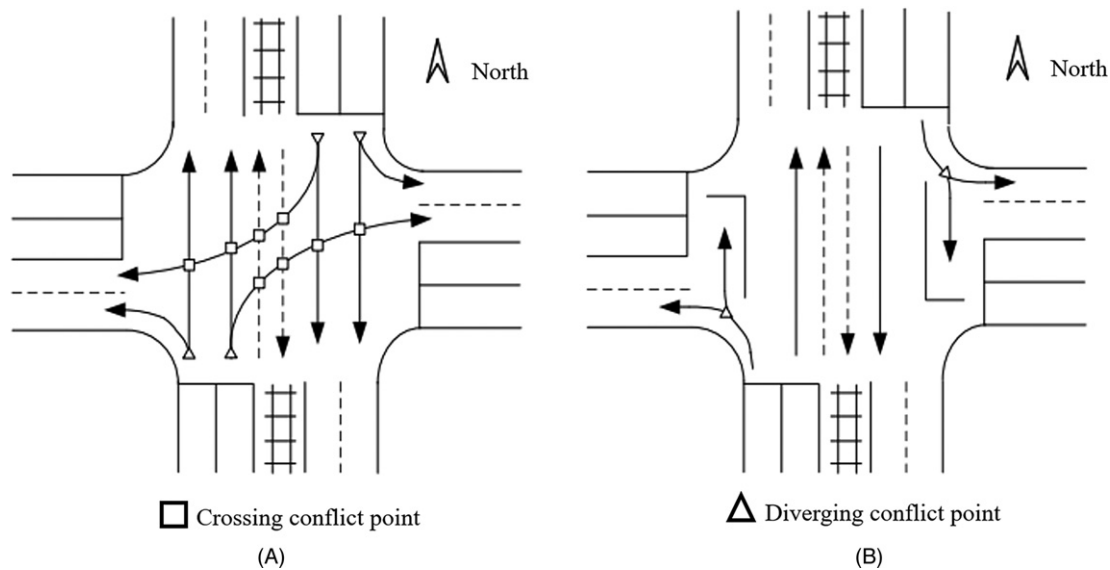


Figure 12 Comparison of conflict points at intersections with and without a hook turn (Bie and Liu, 2015).

the speed of involving vehicles. In addition, the investigations highlighted that conventional right turn traffic could increase the risk of collisions between pedestrians and right-turn traffic. In such situations, drivers of right-turning vehicles concentrate on finding the gap of sufficient size in the opposing traffic and leaving the intersection as soon as possible instead of focusing on pedestrians who may be crossing the streets in which right-turning vehicles are traveling. This study also suggests that the crashes occurred at hook turn intersections tend to be less severe than the ones that occurred at non-hook turn intersections.

Figure 12 shows a hypothetical four-leg intersection during the green time of the north-south signal phase. There are eight crossing conflict points between the right-turn traffic and the through movement. Also, four diverging conflict points resulted from the separation of left and right movements from the through traffic. Meanwhile, at intersections with hook turn implementation, there are only two diverging conflict points that resulted from the separation of right-turn traffic from left-turn traffic. Generally, the existence of a hook turn can remove the crossing conflict points at an intersection, which are considered as the main cause of traffic accidents (Bie and Liu, 2015).

A study was carried out analyzing the crash statistics data in metropolitan Melbourne between 2006 and 2008, and the conflict points associated with the normal right turns versus hook turns at tram intersections in metropolitan Melbourne. According to this investigation, the number of crashes at intersections with hook turn is lower than the corresponding number at the ones without hook turns. Vehicles at intersections with hook-turn treatment are less exposed to road safety hazards than those at intersections without the hook turn treatment (VicRoads, 2010). Also, Pai et al. (2013) suggested that based on official crash statistics in Taiwan, after implementing HTA, motorcyclists' death rate has been decreased noticeably. In this regard, in the first year of hook turn implementation, the number of deaths decreased from 719 to 593 and in the second year dropped to 529 deaths.

Traffic Management Improvement

Analysis of the turn volumes at a number of hook turn intersections within the Melbourne CBD between 1986 and 1999 reveals that the right-turn traffic volume had decreased in most cases. The capacity of the hook turn to discourage turning right, which can ease the left turn has been studied in a series of surveys conducted for urban and rural/tourist drivers. According to this survey, 39% of urban drivers and 38% of rural/tourist drivers stated that they had avoided right turn movement through hook turns.

To evaluate the potential impact of hook turns on the travel time of trams, a series of field surveys carried out during the morning peak period at a number of intersections in metropolitan Melbourne in 2010. In this survey, the traffic volume of right turning movement of hook turn intersections and non-hook turn intersections have been investigated. According to the outcome of this experiment, hook turn intersections can reduce the delay experienced by trams between 11 and 16 seconds per tram-intersection movement. This reduction in traffic delay is equal to saving three to five trams from being deployed to the public transport system compared to non-hook turn intersections. If the cost of each tram is considered as \$5 million, then the total saving cost would be between \$15 and \$25 million (O'Brien, 2000).

Limited studies have been conducted to analyze the traffic performance in terms of capacity and delay before and after implementing the hook turns. Meanwhile, there are some studies that investigated the impact of hook turns on intersections by using microscopic traffic simulation packages and providing before and after scenarios for hook turn implementation. O'Brien (2000) conducted traffic simulation studies to evaluate the performance of hook turns in metropolitan Melbourne. To analyze the performance of hook turns in Melbourne,

seven major intersections with hook turn were selected in those studies. For this purpose, SIDRA software was used. The analysis compares the performance of those intersections with hook turns against the performance of the same intersections without hook turns when the right-turn traffic use and occupy the tram line to make right turns. For this comparison, the degree of saturation has been used as a measurement for the intersection capacity. According to the results, the degree of saturation was much lower in four of seven hook turn intersections compared to non-hook turn intersections. For the other three intersections, removing the hook turns caused a slight improvement in the traffic operation. However, this study suggests that the overall traffic performance within the Melbourne CBD has been improved by implementing the hook turns. This study suggests that hook turns may improve traffic capacity and reduce traffic congestion by separating the right-turn traffic from through traffic and consequently enhancing the movement of through traffic. However, it must be mentioned that hook turns may account for the delay in left-turn traffic. Therefore, the performance of hook turns depends on the volume of turning vehicles. In addition, due to the unfamiliarity of drivers with the hook turn or other unknown reasons, some drivers avoid making a hook turn. This can increase the capacity of the hook turn intersections.

Hounsell and Yap (2013) studied the impact of hook turns on four-leg intersections using a microscopic traffic simulating package. In that study, two hypothetical four-leg intersections with and without a hook turn in London were modeled. For the model development, various traffic scenarios based on different traffic volumes, turning ratios, green splits, and signal timings were examined. According to the findings of this study, the hook turn can provide positive outcomes in terms of reducing the total intersection delay when the right-turn and left-turn traffic volumes are low since the hook turns result in less delay to left-turning vehicles, and fewer left-turning vehicles are required to queue up. Hounsell and Yap (2013) also suggested that the hook turns can be beneficial when the volume of through traffic is high as the hook turn approach can avoid blocking the through traffic at intersections. Also, the opposing traffic flow can take advantage of the hook turn as the obstruction of opposing through traffic caused by right turn traffic is eliminated. In general, when the overall traffic is high, the simulation results demonstrated better traffic performance when the hook turn is implemented to an intersection. This study suggests that outside of the above conditions, hook turn implementation may cause additional traffic delays compared to the conventional intersections without hook turns, mainly because of blocking the left-turn movement.

Bie and Liu (2015) studied the traffic impact of hook turn implementation at two similar intersections with a tram line using microscopic traffic simulation software packages. In this study, both intersections are equipped with Actuated Control Method (ACM) for signal controlling and adjustment. This feature can adjust the signal timing dynamically. This system actively monitors the traffic volumes at all approaches using sensors and loop detectors. ACM is beneficial to cope with the spillover of turning vehicles. ACM can assist intersections in taking the full advantage of the hook turn structure/performance. Using VISSIM microscopic traffic simulation software in this study, the traffic performance of hook turn and non-hook turn intersections were analyzed and compared. In this study, the average delay and vehicle capacity were used as the main performance indicators. The simulation results express that when average and low traffic volumes are observed at intersections, applying hook turns result in a larger average delay compared with the non-hook turn intersections. However, the simulation results for intersections with high total traffic volumes or the high right turn volume ratios show that the hook turn intersections experience lower average delays compared with non-hook turn intersections. This study suggests that the advantage of introducing hook turns is more visible when the ratio of right-turning vehicles is high. When the ratio of right-turning vehicles is low, the capacity of the hook turn and non-hook turn intersections are similar. In general, the hook turn approach improved the performance of right-turning and through vehicles in terms of the total vehicle capacity of the intersection.

Limitations and Disadvantages

Right-turn traffic volume is an important factor in the successful operation of hook turn intersections. The volume of right and left turns needs to be carefully counted/estimated as the right-turn traffic can block the left-turn approach, and this may lead to a long queue for the left-turning vehicles. Moreover, as Currie and Reynolds (2011) suggested in their study, lack of understanding and driver knowledge is a major constraint on the extensive implementation of hook turns in metropolitan Melbourne. Although hook turns are implemented in Melbourne for more than 60 years, around 6% of urban drivers struggled to understand and perform hook turns. It could be worsened when rural or tourist drivers who are not familiar with Melbourne CBD roads are confronted with hook turns. According to this study, 19% of non-commuters struggled to understand how the hook turns work.

According to Le and Nurhidayati (2016), the hook turn implementation can be a challenge because of the potential conflict between through traffic and left-turning vehicles. During the peak hours, the number of motorcyclists waiting to make left turn can overflow the left turn box and subsequently spill into other traffic lanes and the pedestrian crossings. This overflow can increase traffic congestion and reduce traffic safety at the pedestrian crossings of the intersections. As a result, it is clear that prior to the implementation of hook turn approach in new locations and environments, improving the knowledge and understanding of drivers with regard to hook turn rules and regulations as well as drivers' acceptance is essential.

Summary

In this study, the performance and impact of hook turns on intersections and tram movements have been investigated. Hook turns as an alternative way for right turn approach has first been introduced in metropolitan Melbourne. The primary purpose of introducing

hook turns was to keep tramways clear of the right-turning vehicles. Hook turns can have positive safety impacts on intersections as they can effectively reduce the number of conflict points at intersections. By applying the hook turns, the conflict between the right turn traffic and the trams as well as opposing through traffic, can be eliminated. The right-turn head-on crashes are server crashes that may occur at intersections where the right-turn movement is permitted. Overall, vehicles passing through intersections with a hook turn treatment are less prone to the traffic safety hazards compared to the vehicles that pass through the same intersection without a hook turn. In terms of the intersection performance, different studies and traffic simulation experiments have carried out to assess the performance of hook turns. In summary, hook turns may improve the capacity of the intersections. The role of hook turns is more highlighted when the right of way of tram and the reduction in their experienced delay time are targeted. With this regard, the through traffic can take advantage of hook turns as the right-turning vehicles can obstruct the opposing through traffic. However, the hook turns have some limitations. The ratio of right-turn traffic needs to be considered in hook turn implementation as they can obstruct the left-turn traffic and impose excess delays to this approach. Moreover, lack of understanding and driver knowledge can be a major obstacle to the proper implementation of this approach.

Finally, due to the lack of comparative research and before-after studies related to the hook turn implementation, it is recommended to conduct more studies to provide more accurate assessments of the hook turn approach.

References

- Bie, Y., Liu, Z., 2015. Evaluation of a signalized intersection with hook turns under traffic actuated control circumstance. *J. Transp. Eng.* 141 (5), 1–10.
- Currie, G., Reynolds, J., 2011. Managing trams and traffic at intersections with hook turns: Safety and operational impacts. *Transp. Res. Rec.* 2219 (1), 10–19.
- Currie, G., Shalaby, A.S., 2007. Success and challenges in modernizing streetcar systems: experiences in Melbourne, Australia, and Toronto, Canada. *Transp. Res. Rec.* 2006 (1), 31–39.
- Currie, G., Smith, P., 2006. Innovative design for safe and accessible light rail or tram stops suitable for streetcar-style conditions. *Transp. Res. Rec.* 1955 (1), 37–46.
- Cycling in Christchurch 2020. Retrieved from <http://cyclingchristchurch.co.nz/2013/03/13/clever-cycling-stuff-hook-turn-boxes>
- Falamarzi, A., Moridpour, S., Nazem, M., Hesami, R., 2018. Rail degradation prediction models for tram system: Melbourne case study. *J. Adv. Transp.* 2018, 1–8, doi:10.1155/2018/6340504.
- Flickr, 2020. Retrieved from <https://www.flickr.com/photos/rorytai/94797345>.
- Google, 2020. Retrieved. <https://www.google.com/maps/place/Elizabeth+St+%26+La+Trobe+St,+Melbourne+VIC+3000/@-37.8102971,144.9592407,17z/data=!3m2!4b1!5s0x6ad642-cae96e5483:0xe2c955d3c4c52ef5!4m5!3m4!1s0x6ad642cacda01d6f:0xce8edf98a686d0bc!8m2!3d-37.8103014!4d144.9614294>
- Hounsell, N.B., Yap, Y.H., 2013. Hook turns as a solution to the right-turning traffic problem. *Transp. Sci.* 49 (1), 1–12.
- Le, T.Q., Nurhidayati, Z.A., 2016. A study of motorcycle lane design in some Asian countries. *Proc. Eng.* 142, 292–298.
- Marti, C.M., Kupferschmid, J., Schwertner, M., Nash, A., Weidmann, U., 2016. Tram safety in mixed traffic: Best practices from Switzerland. *Transp. Res. Rec.* 2540 (1), 125–137.
- Naznin, F., Currie, G., Logan, D.B., 2018. Exploring road design factors influencing tram road safety—Melbourne tram driver focus groups. *Accid. Anal. Prev.* 110, 52–61.
- O'Brien, A., 2000. Review of Hook Turns (Right Turn from Left). Research and Development Project 729. VicRoads, Melbourne, Australia.
- ODOT 2017. Oregon Department of Transportation, Oregon Bicyclist Manual. Salem, OR, US.
- Pai, C.W., Hsu, J.J., Chang, J.L., Kuo, M.S., 2013. Motorcyclists violating hook-turn area at intersections in Taiwan: An observational study. *Accid. Anal. Prev.* 59 (3), 1–8.
- TMR 2020. Bicycles Road Rules and Safety. Department of Transport and Main Roads, Queensland, Australia.
- VicRoads. 2010. VicRoads Crash Statistics Data. Melbourne, Australia.
- VicRoads. 2020. Turning Road Rules. <https://www.vicroads.vic.gov.au/safety-and-road-rules/road-rules/a-to-z-of-road-rules/turning>. Accessed May 25, 2020.
- YouTube, 2014. Where Scooters Rule—Taiwan Scooter System at an Intersection. <https://www.youtube.com/watch?v=g4i47vTsVR8>.
- Yarra Trams. 2020. Map of Melbourne Tram Network. <https://yarratrams.com.au/network-map>. Accessed May 22, 2020.

Traffic Signal Coordination

Rahim F. Benekohal, University of Illinois at Urbana-Champaign, Urbana, IL, United States

© 2021 Elsevier Ltd. All rights reserved.

Signal Coordination Definition	213
Signal Coordination Goals	213
When to Coordinate Traffic Signal	214
Signal Coordination Elements	214
Cycle Length	214
Split and Green Time Allocation	214
Offsets	215
Computing Offsets	215
Effects of Vehicle in Queue on Offset	215
Fine-Tuning Offsets	216
Movement Numbering Scheme	216
Actuating Coordinated Phase	216
Transition Logic	217
Bandwidth	217
Single Band Versus Multiband	218
One-Way and Two-Way Signal Coordination on Arterial	218
Decisions Affecting Coordination	218
Signal Progression Schemes	219
Special Signal Progression Schemes	219
Simultaneous System	219
Alternate System	219
X-ple Alternating	219
Force Off	220
Offset Reference Point	220
Signal Coordination in Oversaturated Conditions	220
References	220

Signal Coordination Definition

Adjacent traffic signals can be timed such that a group of vehicle (a platoon) would go thru the intersections without stopping. Traffic signal coordination (or traffic signal synchronization) is a process by which the beginning of green time at the subject intersection is determined such that the platoon of vehicles arriving from the upstream intersection would traverse without stopping. When traffic signal are well-coordinated, travelers would see green signal indications as they approach successive intersections. A proper signal coordination plan can result in safe and efficient intersection operation.

Signal Coordination Goals

There may be one or more of the following goals for signal coordination:

1. Optimize traffic signal operation;
2. Minimize delay, number of stops, travel time, etc.;
3. Maximize traffic throughput, speed, green utilization, etc.;
4. Reduce frequency or severity of crashes and improve traffic safety;
5. Allow pedestrian crossing at midblock crosswalk.

Efficiency of intersection operation can be high when vehicles are platooned and they are allowed to pass thru intersections with minimal delay. Vehicles in platoon maintain shorter headways and that increases the number of vehicles that go thru the intersection in a given time.

When a platoon travels a longer distance, it is more likely that the vehicles in platoon would disperse more and that would reduce the efficiency of traffic signal operation. Vehicles in platoon can stay pretty much intact miles or longer (Figs. 1 and 2).



Figure 1 Example of a dense platoon of vehicles.



Figure 2 Example of a partially dispersed platoon of vehicles.

When to Coordinate Traffic Signal

Signals that are located close to each other should be coordinated. What is considered “close”? It depends on travel time and/or the distance between the two adjacent intersections. The higher the travel speed the longer the distance between the two signals can be. Generally, if it is under 1 mile on streets/highways with higher speed limits (55 mph), it may be considered as “close”. The Manual on Uniform Traffic Control Devices ([MUTCD, 2003, 2009](#)) considers 0.5 miles as “close” and the Signal Timing Manual ([Urbanik et al., 2015](#)) and Traffic Signal Timing Manual ([Koonce et al., 2008](#)) considers miles as “close”. If traffic signals are too close to each other (e.g. about 1000 ft or less), they may be treated as “one” signal. In deciding to coordinate or not one should consider traffic volume, operating speed and travel time between intersections, elements causing platoon dispersion, and feasibility of keeping a dense platoon.

When traffic volume is low, signal coordination may not be very advantageous since the number of vehicles in queue would be low. When speed is high and travel time is short, platoon does not dissipates much. However, when speed is low and travel time is long, the vehicles in the platoon may disperse.

Signal coordination may be done along an arterial or on a network of intersections. The coordination may be on one direction of travel or on both directions. While signal coordination on one-way arterial is relatively straightforward, it gets complex on two-way coordination; and even more complex when a network of intersections are coordinated.

Signal Coordination Elements

Three important elements need to be determined:

1. Cycle length
2. Splits (green times)
3. Offsets

Cycle Length

In general, a common cycle length is used for all coordinated signal. The common cycle optimizes the operation on arterial (network) level and may not be the optimal cycle length at every intersection. Each intersection may have an optimal cycle length that is different than the common cycle length. In unusual cases and special conditions, the cycle length may vary at different intersections. The cycle length for unusual cases may be even multiple of the common cycle (1/3, 1/2, 2, 3 times). Using different cycle length on an arterial (network) makes the signal coordination more complicated, but that has been used ([Abu-Lebdeh and Benekohal, 1997a,b, 2000a,b](#)). The half common cycle (1/2 of common cycle) may be used when an intersection is far from the rest of the intersections and coordination is more important in one direction. The double cycle may be used at the key intersection when it is located in the middle of two groups of intersections.

To determine the common cycle length, a lower bound (minimum cycle length) and an upper bound (maximum cycle length) is decided. Often, the minimum cycle length is 50 s and the maximum is around 180 s. Then, the common cycle length is determined by finding the cycle length that optimizes the performance measure for the entire corridor (or network). The common cycle length is found by trying different cycle lengths within the specified range at small increments (like 10 s). The performance measure often used is delay. In some occasions, a combination of delay with another performance measure such as number of stops is used. Using the combination of two or more performance measure gets complicated because the performance measures use different variables and have different units.

Split and Green Time Allocation

Green time available during a cycle is divided among the phases. Split shows the portion of cycle that is allocated to a given phase. Split is composed of green, yellow change interval, and red clearance times. Most part of the split is the green time. The green time

should be long enough to process the vehicles that arrive in one cycle when the intersection is not oversaturated (under saturated). When intersection is oversaturated, the demand becomes larger than the capacity of the intersections. The green time should not be wasted. Green time is wasted when it is not utilized efficiently. Efficient green utilization is when vehicles go thru intersection at a steady departure rate. A reasonable range for the steady departure rate is 1.8–2.2 s per vehicle. The default saturation flow rate for signalized intersection in Highway Capacity Manual is 1900 passenger cars per hour per lane. This translates to a departure headway of 1.89 s.

To estimate the amount of green time needed to process “ N ” vehicle per cycle, one may use “2 + 2 rule”. The 2 + 2 rule implies that give 2 s of green to start with and for every vehicle give 2 s of green. It is mathematically may be expressed as: Estimated green time = $2 + 2(N)$.

This expression gives a good estimate for the lower bound of the green time. When start-up lost time is longer or there are vehicles that use longer headways (such as trucks) the estimated green time should be increased. When vehicles are departing at a headway less than 2 s, the estimated green time may be decreased. This estimate may be used to allocate the green time to the minor street. The major street may get the remaining green time.

A split time indicates the proportion of cycle length allocated to that phase. Split time = green time + yellow change interval + red clearance time. Summation of split times gives the cycle length.

Offsets

An offset is the time difference between the beginnings of green times (usually) for the phases to be coordinated. Offset is often expressed relative to adjacent signal. It can be referenced to any point in cycle or a reference point in time. Offset is often less than a cycle, but it can be greater than cycle if the reference point is not adjacent signal. Offset can be positive, zero, or negative.

A positive offset indicates that the downstream phase will start after the corresponding upstream phase. A zero offset indicate the corresponding phases start at the same time. A negative offset indicates that the coordinated phase at the downstream intersection will start earlier than the corresponding phase green at the upstream intersection. The negative offset is used to clear the vehicles are at the downstream intersection before the vehicles from the upstream intersection arrive.

Computing Offsets

The two important variables in offset determination are: speed and distance between the two intersections. An “ideal” offset is computed by dividing the distance by speed.

$$t_0 = L/S$$

where t_0 is the offset when there is no queued vehicles (ideal offset) (s), L is the distance from one stop bar to the next (ft), S is the average speed of vehicles on that section of the road (ft/s).

Often the posted speed limit is used for computing the offset. Other speed such as average operating speed or a given percentile speed may also be used. Using a speed higher than the speed limit yields a shorter offset and that may encourage speeding. On the other hand, using a lower speed yields a longer offset and that result in an inefficient operation because the arriving vehicles in platoon encounter a red light and are forced to stop.

Effects of Vehicle in Queue on Offset

Offset computation is straightforward when there is no vehicle in queue at the beginning of green at the downstream intersection. However, when there are vehicles in queue at the downstream intersection, the time to process the queued vehicles (Z_q) is subtracted from the offset

$$t_q = L/S - Z_q$$

where t_q is the offset when there are vehicles in queue at downstream intersection (s), Z_q is the time required to process the vehicles in queue at downstream intersection (s).

Z_q depends on no of vehicles in queue at the end of red phase (Q_r). These vehicles may come from minor streets or driveways between the intersections during red, or they may be left over vehicles from the previous cycle.

$$Z_q = Q_r(h) + l_r$$

where Q_r is the number of vehicles in queue at the end of red phase, h is saturation headway (s). l_r is the start-up lost time for the queued vehicles (s).

Fine-Tuning Offsets

Often the computed offsets are fine-tuned in the field to account for the “factors” that were not considered in computation of the offsets. For example, pedestrian crossing in midblock, bus stops, or vehicles entering/exiting from the minor streets or driveways can affect the travel speed and thus the offset. In addition to the factors not considered in computing offset, the offsets are affected by the errors and inaccuracies in the variables used in offset calculation.

$$t_{qf} = L/S - Z_q + t_1 - t_2 - t_3 + t_0$$

where t_{qf} is the fine-tuned offset when there are queued vehicles, t_1 Time (s) that takes for the first vehicle in platoon to reach the speed used in finding offset. If $t_1 = 0$, it indicates the lead vehicle instantaneously reached the speed and traveled zero distance, t_2 Time it takes for the leader of platoon to travel the distance over which it reached the speed used in finding the offset. If $t_1 = 0$, then $t_2 = 0$, t_3 Time (s) the green at downstream intersection starts a little earlier to prevent the platoon leader from slowing down due to seeing a red light ahead, t_0 Time added or subtracted to take care of factors that are not presented by other adjustments.

If we assume $t_1 - t_2 - t_3 = 0$, and $t_0 = 0$, then no adjustment is needed due to acceleration, traverse time, and deceleration.

Movement Numbering Scheme

Each vehicular and pedestrian movement are numbered, as shown in the numbering scheme. Left turns are numbered 1, 3, 5, and 7. Through movements are numbered as 2, 4, 6, and 8. The right turners are numbered as 12, 14, 16, and 18. Pedestrian movement, regardless of its direction of travel (e.g. from east to west or from west to east) are numbered 2P, 4P, 6P, and 8P. Often phase 2 and phase 6 are the major street through movements (Fig. 3).

Actuating Coordinated Phase

Normally, the coordinated phases do not have vehicle detectors since the other phases have detectors. However, the coordinated phase can be actuated during the last portion of its green time. The actuation allows to terminate the coordinated phase earlier when there is no demand on it (Fig. 4).

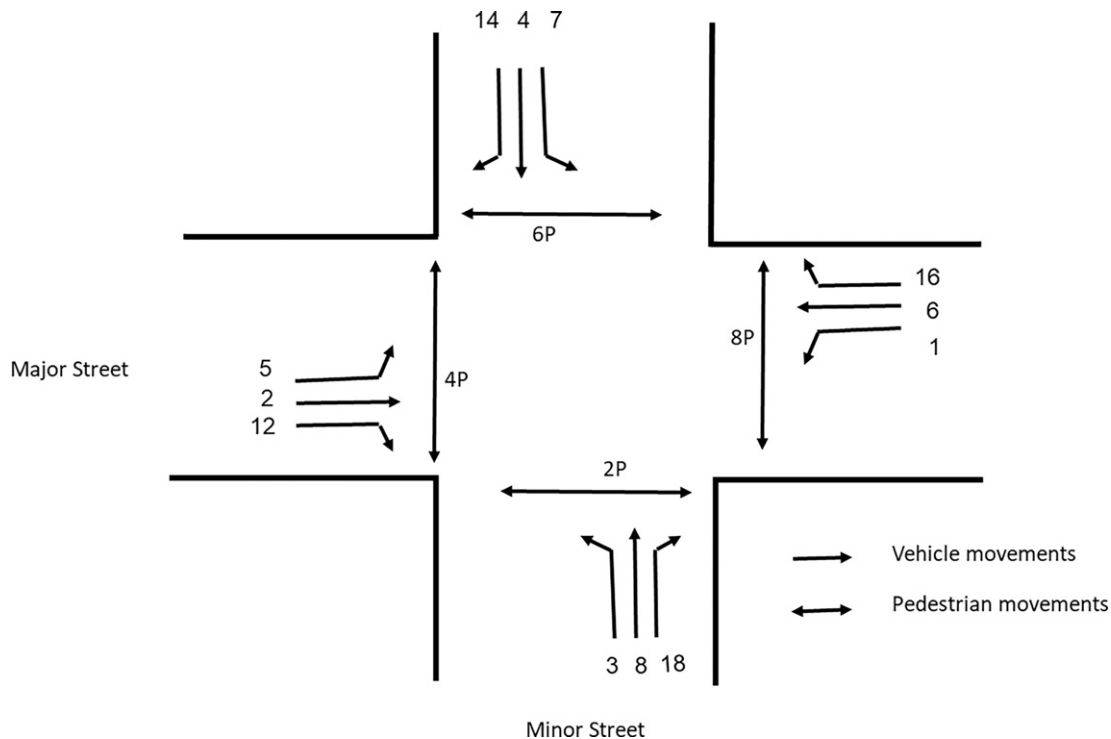


Figure 3 Movement and phase numbering scheme.

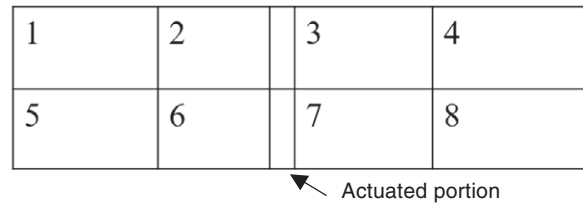


Figure 4 Actuated portion of coordinated phases.

In the sketch phases 2 and 6, which are coordinated phases will be allowed to gap out to terminate the phase earlier. This allows the minor street get green quickly. If “actuated time” is too large, it may cause the coordinated movements to gap out just prior to the arrival of the platoon. If “actuated time” is too small, it may have inconsequential result.

Transition Logic

Traffic signal operation pattern (cycle, splits, offsets, phase sequence) may change due to time of day schedule, emergency vehicles or train arrival interruption of normal operation, pedestrian signal activation that last longer than the vehicular green time, or other reasons. When the pattern is changed, operation goes thru a transition time and a logic to switch to the new pattern. The transition time often is over a few to several cycles. The transition logic takes operation from one pattern to another pattern. Excessive changes in pattern of operation is not good because of the transition time involved. Generally recommended to keep the pattern for at least 30 min (STM2). Changing patterns during congested conditions should be avoided.

Bandwidth

Bandwidth is the green time available for the vehicles in a platoon to go through a group of intersections without stopping. Bandwidth in a “good” one-way coordination plan can be equal to the shortest green time of the coordinated phases. Thus, green split at the intersection can affect the bandwidth (**Fig. 5**).

Band width (BW) efficiency (EB_{BW})

It is defined as ratio of bandwidth (BW) to cycle length (C). It is expressed as decimal or percent.

$$E_{BW} = (BW/C) \times 100$$

The BW efficiency often is not in 80's% or 90's% since green time is allocated to two or more phases. An efficient BW may be achieved at 40%–60% levels and it depends on the on shortest green time of the coordinated phase.

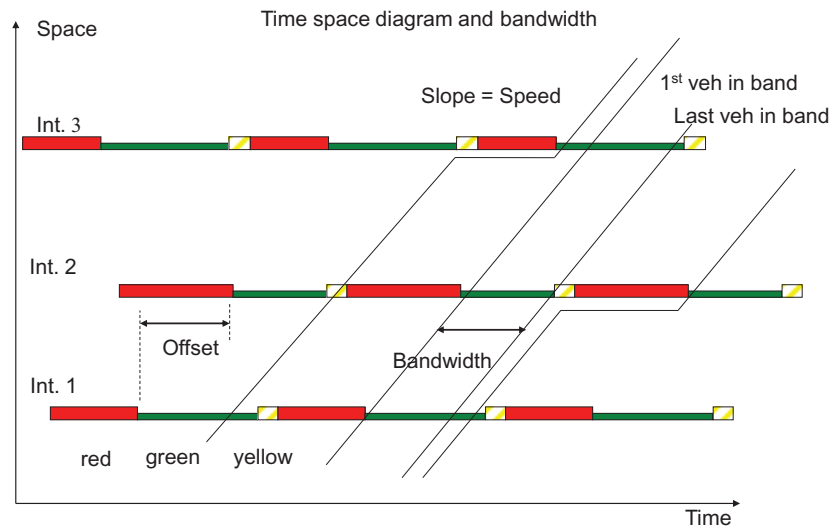


Figure 5 Offset and bandwidth at three consecutive intersections.

Capacity of bandwidth (C_{BW}) is defined as the number of vehicles that can travel within the bandwidth and go through all the coordinated intersections without stopping. Bandwidth capacity in units of vehicles/hour/lane can be expressed as

$$C_{BW} = (BW/h)(3600/C)$$

where h is saturation headway (s), C is cycle length (s), and BW is bandwidth (s). Bandwidth capacity cannot be greater than the capacity of the coordinated phase with the least green time. Queued vehicles at the downstream intersections can adversely affect the bandwidth capacity, if they are not processed before arrival of the platoon of vehicles.

Single Band Versus Multiband

A single bandwidth can always be maintained among the signal located on a one-way street. When several intersections are coordinated, more than one band (multiband) may be possible among a sub-group of the intersections (see Fig. 9).

One-Way and Two-Way Signal Coordination on Arterial

An efficient signal coordination is possible when the signals are along a one-way street. However, two-way coordination the bandwidth in one direction may be reduced to accommodate the coordination in the other direction. In some cases, this accommodation may not be possible. Multibandwidths may be used along a long arterial, but it gets more complicated to implement.

Signal coordination in two directions is more complicated than in one direction mainly because offsets separately computed for one direction often messes up the progression in the opposite direction.

The offset in one direction between intersection 1 and 2 (t_{12}) plus the offset in the other direction (t_{21}) is equal to cycle length (one or multiple cycle lengths).

$$t_{12} + t_{21} = nC$$

This shows that the offset are interdependent (Fig. 6).

Decisions Affecting Coordination

- Selecting a different bandwidth in main direction
- Changing speed
- Changing cycle length
- Chang green splits
- Changing phase plan to change green time
- Restricting minor street access to right in right out
- Removing the traffic signal (eliminate the access)

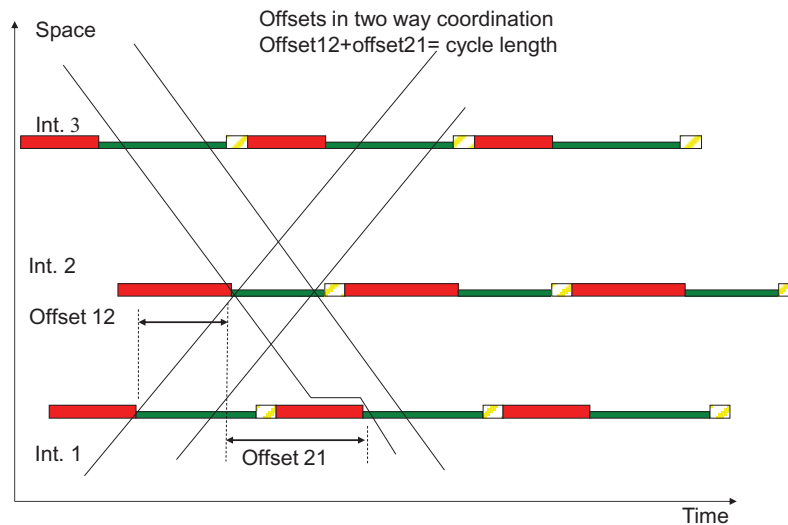


Figure 6 Offset in one direction and its relation to offset in other direction.

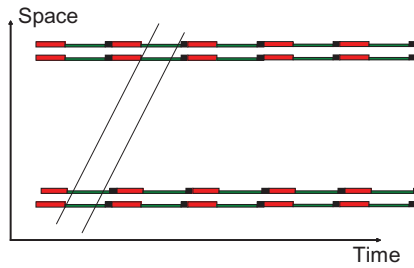


Figure 7 Simultaneous signal coordination scheme.

Signal Progression Schemes

The most common type of signal progression is what was discussed previously. Some call this “simple” progression when there are no queues at intersections. Simple progression is also called “forward” progression. If progression is revised during the day to provide a different priority some call it “flexible” progression. When upstream intersections start earlier than subject intersection (negative offset) it is referred as “reverse” progression (Roess et al., 2011).

Special Signal Progression Schemes

Simultaneous System

That is when the coordinated phases at different intersections start at the same time. This can be used for intersections that are very close to each other, or the offset is equal to the cycle length (Fig. 7).

Alternate System

The progression is such that every other intersection starts simultaneously (Fig. 8).

X-ple Alternating

Double (triple, quadruple) alternating—if there are two (three, four) intersections are considered as a group and you have alternating system between the groups (Fig. 9).

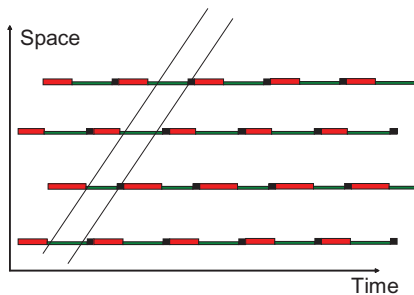


Figure 8 Alternating signal coordination scheme.

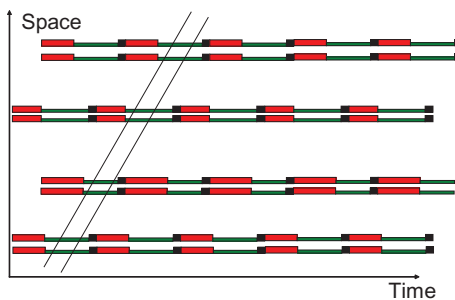


Figure 9 Double alternating signal coordination scheme.

Force Off

Refers to a time during cycle that the uncoordinated phases must end (even if there is demand for longer green) to ensure that coordinated phase would start as planned. There are two types of force offs: Floating type or fixed type. The type indicates how unused actuated green time of uncoordinated movement is allocated to the subsequent phases. In floating force off, the unused green is given to the coordinated phase (force off point of the uncoordinated phase to coordinated phase is floating (not fixed)). In fixed force off, the unused green is given to the following phase (force off point of the uncoordinated phase to coordinated phase is fixed).

Offset Reference Point

Defines the point in time, relative to master clock, the cycle begins. It can be variety of points in cycle. Modern controller allows options such as: Strat of green for the first coordinated phase (used in TS2); Strat of green for both coordinated phases (used in TS 1); Strat of FDW for the first coordinated phase; Strat of yellow for the first coordinated phase (used in 170).

Signal Coordination in Oversaturated Conditions

Signal Coordination in oversaturated conditions becomes more complex due to queue spill back and lane blockage preventing enough demand to reach downstream (starvation). In oversaturated conditions, queue management is more appropriate measure of effectiveness than delay (Abu-Lebdeh and Benekohal, 1997a,b, 2000a,b, 2007). Abu-Lebdeh and Benekohal developed a dynamic traffic signal control algorithm for oversaturated arterials to dynamically manage queue formation and dissipation. For a one-way arterial, their method provided dynamic time-dependent traffic control. Offsets and green times were dynamically changed as a function of demand and queue lengths. They found similar results for a two-way arterial (Abu-Lebdeh and Benekohal, 1997a, b, 2000a,b).

For real-time dynamic signal coordination, Intelligent search and optimization technique need to be used to find the optimal solutions due to the large search space of the solutions. Girianna and Benekohal (2004) and Hajbabaie and Benekohal (2013) used genetic algorithm (GA) and Medina et al. (2013) used reinforcement learning to find optimal signal coordination plan for oversaturated networks. For oversaturated conditions, the signal timing optimization may be done over multiple cycles (Abu-Lebdeh et al., 2007).

References

- Abu-Lebdeh, G., Benekohal, R.F., 1997. Development of a traffic control and queue management procedure for oversaturated arterials. In: Transportation Research Record 1603. Transportation Research Board, No. 119–127. National Research Council, Washington, DC.
- Abu-Lebdeh, G., Benekohal, R.F., 1997b. Development of a traffic control and queue management procedures for oversaturated arterials. Transport. Res. Rec. 1603, 119–128. TRB, NRC.
- Abu-Lebdeh, G., Benekohal, R.F., 2000a. Genetic algorithms for traffic signal control and queue management of oversaturated two-way arterials. J. Transport. Res. Board 1727, 61–67, NRC.
- Abu-Lebdeh, G., Benekohal, R.F., 2000b. Signal coordination and arterial capacity in oversaturated conditions. J. Transport. Res. Board 1727, 68–76.
- Abu-Lebdeh, G., Chen, H., Benekohal, R.F., 2007. Modeling traffic output for design of dynamic multi-cycle control in congested conditions. J. Intell. Transport. Syst. 11 (1), 25–40, doi:10.1080/15472450601122371.
- Girianna, M., Benekohal, R.F., 2004. Using genetic algorithms to design signal coordination for oversaturated networks. J. ITS: Technol. Plann. Oper. 8 (2), 117–129.
- Hajbabaie, A., Benekohal, R.F., 2013. Traffic signal timing optimization: choosing the objective function. J. Transp. Res. Board 2355, 10–20.
- Koonce, P., Rodegerdts, L., Lee, K., Quayle, S., Scott, B., Braud, C., Jim, B., Tarnoff, P., Urbanik, T., 2008. Traffic Signal Timing Manual. FHWA-HOP-08-024, FHWA, Washington, DC, 2008.
- Manual on Uniform Traffic Control Devices, December 2003. U.S. Department of Transportation, Washington, D.C. https://mutcd.fhwa.dot.gov/htm/2009/html_index.htm.
- Medina, J.C., Hajbabaie, A., Benekohal, R.F., 2013. Effects of metered entry volume on an oversaturated network with dynamic signal timing. J. Transp. Res. Board 2356, 53–80.
- Roess, R., Prassas, E., McShane, W., 2011. Traffic Engineering, fourth edition. Prentice Hall.
- Urbanik T., et al., 2015. Signal Timing Manual, second edition. NCHRP 812, TRB. <http://www.trb.org/OperationsTrafficManagement/Blurbs/173121.aspx>.

Emergency Vehicle Priority (Preemption): Concept and Advancements

Chaitrali Shirke, School of Civil and Environment Engineering, Queensland University of Technology, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	221
EVP State of the Practice	221
Communication Technologies for EVP	222
EVP Systems of the Real World	223
Transition Out of EVP	223
EVP State of the Art	225
Multiple Preemption Requests	225
Route-Based Strategy	225
Conclusion	225
References	226
Further Reading	226

Introduction

Emergency vehicle priority or emergency vehicle preemption (EVP) is a form of preferential treatment used to give emergency vehicles (EV) a priority green light at an intersection. EVP interrupts the normal operation of traffic signals for EV to cross an intersection and then returns to normal operation after the EV passes the intersection. Vehicles that are designated and authorized to respond to an emergency in a life-threatening situation such as police vehicles, ambulance, and firefighting vehicles are referred to as EVs. EVP aims to reduce the EV's travel time and enables it to respond to an emergency call for an incident with a minimum delay.

Although signalized vehicle preemption is a relatively recent development resulting from advancements in Intelligent Transportation Systems (ITS), the concept of prioritizing emergency vehicle movement is not. Emergency vehicles were and are still prioritized on streets using sirens and strobe lights. However, intersections remained to hinder it from moving uninterrupted. In the 1960s, hardware technology to detect vehicles using vehicle-based emitters and signal-based detectors emerged which provoked the concept of EVP at signals.

Though EVP interrupts and might increase the delay for general traffic, it has several benefits such as:

- Reduce the probability for an EV to receive a red signal, and as such, reduce travel time.
- Reduce the likelihood of crashes at intersections by clearing the traffic queues from the path of the EV. Thereby, it assists in safely and effectively responding and attending to emergency incidents
- Increase the potential for saving lives by providing faster access to incidents
- Decrease the potential disruption to the road network by clearing the incident faster
- It helps to support the emergency response personnel to meet response targets
- Assists EVs to manage increasing demands
- EVP can also be adapted for other modes of transport, such as buses and heavy vehicles
- It gives a policy-based approach for giving priority to road users
- It provides a safe work environment for front-line officers.

This article presents a brief review of EVP systems deployed in practice as well as the state-of-the-art research on EVP operation.

EVP State of the Practice

Automatic vehicle location (AVL) systems such as global positioning systems (GPS) are used to determine the location of an EV and calculate estimated times of arrival at intersections. Further, a message is sent to the traffic controller that an emergency vehicle is approaching using vehicle to roadside communication (VRC). After receiving the preemption call, which may include the expected time of arrival at the intersection and signal phase requested by EV, traffic controller generally follows the steps shown in Fig. 1. First, after receiving the preemption call, the current status of the phase requested by EV is checked. If the status of the requested phase is red without or with pedestrian movements active in conflicting phase, then the requested phase is tuned green only after serving yellow, and all red or minimum pedestrian clearance time, yellow and all red to conflicting phase. If the status of phase requested by EV is green, then the green time for the phase is extended to serve the EV request. Once the request is served, the transition algorithm is applied to exit out of preemption.

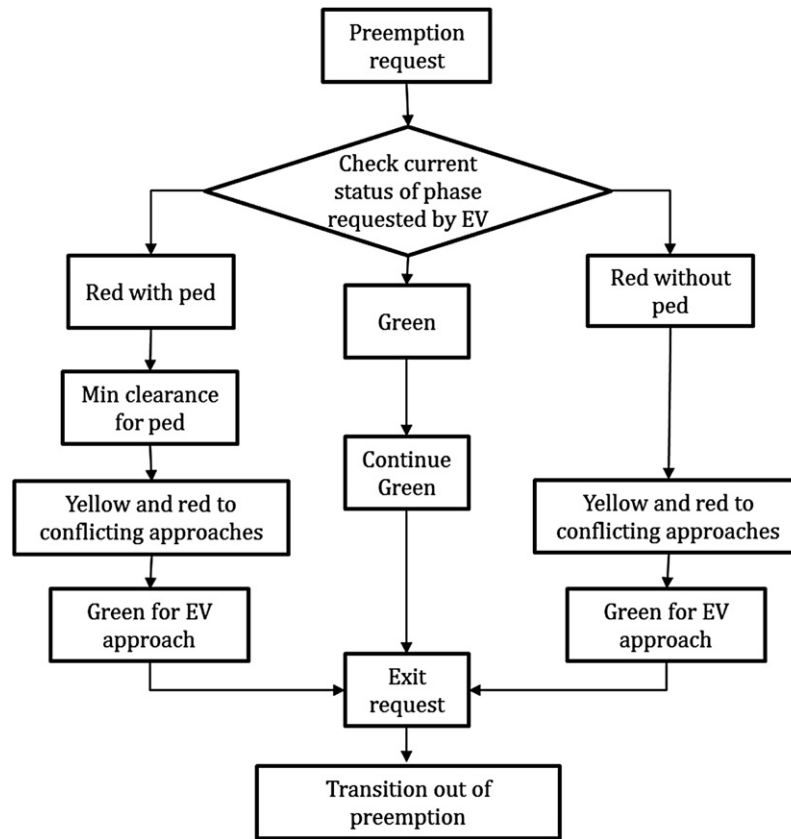


Figure 1 General framework for preemption logic. EV, emergency vehicle

It is essential to understand that EVP is different from transit signal priority (TSP), which is a treatment used to give priority to transit vehicles such as buses, trams, etc., at a signalized intersection. EVs have an absolute (or highest) priority whereas TSP offers priority to transit vehicles only when it is possible to do so without disrupting normal signal operations. For example, TSP operation is not allowed to skip or override signal phases to serve the phase requested by a transit vehicle. In contrast, EVP operation can do so to offer a priority to EV.

Communication Technologies for EVP

The major technologies currently used for communicating with traffic controller include light or infrared-based systems, sound-based systems, and radio-based systems (Paniati and Amoni, 2006). Each of these systems has its advantages and disadvantages as shown in Table 1.

Table 1 Communication technologies used for EVP operation

Technology	Dedicated vehicle emitter required	Susceptible to electronic noise interference	Clear line of sight required	Affected by weather	Preemption possible on other approaches
Light or infrared system	Yes	No	Yes	Yes	No
Sound-based system	No	No	No	No	Yes
Radio-based system	Yes	Yes	No	No	Yes

EVP, emergency vehicle priority or emergency vehicle preemption.

EVP Systems of the Real World

Examples of EVP systems deployed around the globe include EVP practice in Maricopa Association of Governments (MAG) region (MAG, 2018), the SAFER system deployed in China (Tokuda & Ohmori, 2011), EVP system in Victoria, Australia (Intelligent Transport Systems Group, 2005), STREAMS EVP by Transmax (2018), Tongji Emergency Vehicle Signal Preemption System (TJ-EVSP) (Wang et al., 2013), Plano Texas EVP System and Fairfax County Virginia EVP System (Paniati and Amoni, 2006). Other real-world EVP systems are OPTICOM emergency response by Global Traffic Technologies LLC, (2020), STROBECOM II system by TOMAR Electronics Inc. (2020). All these systems use infrared signals for communication between EV and signal control infrastructure.

Abovementioned preemption systems operate on an intersection-by-intersection basis. Sensors detect an EV at each controller and each controller switches to a predefined preemptive phase. This result in preemption of each intersection only after the EV reaches it which may cause the EV to wait at an intersection for vehicles to clear. This can also confuse drivers of other vehicles about whether to pull over or proceed in the presence of an EV at a preempted green. Local detection of an EV is further complicated by peak hour traffic or after-event traffic when the corridor is congested. In such conditions, pre-emption can create increased delays at local intersections due to lack of clearance at downstream intersections.

Very few route-based dynamic preemption systems have been found in the practice. The only route-based signal clearance research found from the literature was the Fast Emergency Vehicle Pre-emption System (FAST) developed by a group of Japanese researchers (Miyawaki et al., 1999; Shibuya et al., 2000). When an incident occurs, FAST is activated in support of the EV's operation. The traffic control center calculates the recommended route by which the EV can reach the scene in the shortest time; it uses the infrared beacon's position through which the vehicle is expected to pass and transmits the route information to the beacon. Then it conducts priority control, giving maximum signal green time and adjusting the offset at each intersection on the recommended route to make it easy for the vehicle to reach the scene. When it passes through the infrared beacon, the EV is informed of the recommended route and is given details about the situation at the scene in response to transmission of the vehicle identification (ID) information. The EV heads toward the scene along the recommended route. The infrared beacon notifies the intersection downstream of the direction taken by the EV and of the EV's approach, and it initiates another priority control extending green time or shortening red time to allow the EV to pass with a green light. The information board installed near the intersection warns civilian vehicles of the approaching emergency vehicle. The traffic control center uses advanced mobile information systems (AMIS) to provide traffic information about the surroundings and the occurrence of incidents to in-vehicle units mounted on civilian vehicles through the infrared beacons. As of the end of 2013, 15 prefectures had adopted FAST (Intelligent Transport Systems, 2014).

Transition Out of EVP

As presented in Fig. 1, after preempting a signal for serving a priority request from EV, some recovery measures/ transition strategies are necessary to transition from a preemption control plan back to the coordinated operation of the normal timing plan. If a traffic signal is operating in isolated or uncoordinated mode (e.g., fully actuated or free), decisions on how to exit from a preemption control plan can be made without considering the traffic progression between adjacent signals. To identify how the signal controller will exit or recover from a preemption control plan requires specifying a return interval or sequence of intervals that must be served. The sequence that is specified should be based on the conditions and constraints specific to each intersection.

However, to maintain coordination with adjacent traffic signals, exit transition strategies provide the capability to return the controller to a point in the signal timing plan where it would have been if it had not been preempted. Such a strategy should automatically perform the necessary adjustments to reach the desired reference or coordination point in the normal signal-timing plan. The adjustments are generally accomplished by holding in the desired return phase, or reducing or expanding the length of the signal timing plan until the desired coordination or offset point is reached.

Table 2 summarizes some of the widely used transition strategies in practice. Column one presents the name of the strategy, column two provides a brief introduction of the strategy, column three lists the signal controllers that support the particular transition strategy, and finally, the last column identifies the studies that evaluate the specific strategy. Few of the researchers (Obenberger & Collura, 2007; Yun et al., 2008) also performed a comparative evaluation of the transition strategies described in Table 2. These studies suggest that several traffic characteristics, conditions, and factors significantly influence the potential impacts of different exit transition strategies on traffic flow. Some of these factors are summarized below:

- Frequency and direction of travel for the vehicles requesting preemption
- Roadway capacity to accommodate turning movements to avoid any further disruption of traffic on all approaches to the intersection
- Travel demand and available roadway capacity (e.g., volume to capacity ratio)
- Presence and frequency of pedestrian phases or actuated calls
- Green time available in cycle length to be reallocated to accomplish the desired exit transition strategy
- Cycle length, number of phases, and intervals to serve in the signal plan; and
- Intersection spacing and quality of progression of traffic between successive traffic signals

Table 2 State of practice transition strategies to exit from preemption

<i>Transition strategy</i>	<i>Description</i>	<i>Controllers</i>	<i>References</i>
Dwell/ Hold	Corrects a misaligned coordinated phase (also known as the sync phase) by holding the coordinated phase green (i.e., dwelling) until it is aligned with the schedule in the new plan or is in sync	170E, 2070 ATC, Eagle EPAC300, Econolite ASC2S, Siemens NextPhase, Naztec Model 980	Shelby et al. (2006); Obenberger and Collura (2007); Yun et al. (2008)
Max Dwell	It is similar to Dwell logic, but the extra time spent dwelling in the sync phase is constrained each cycle—typically 20% of the cycle length. Thus, it may take multiple cycles to complete the offset correction.	Eagle EPAC300, Econolite ASC2S, Siemens NextPhase, Naztec Model 980	Shelby et al. (2006)
Long way/Add	Corrects the offset by lengthening the cycle, typically increasing all phase durations proportionally to their splits. The maximum amount of additional time added to each transition cycle is generally constrained to a specified percentage of the cycle length.	2070 ATC, Eagle EPAC300, Econolite ASC2S, Siemens NextPhase, Naztec Model 980	Shelby et al. (2006); Obenberger and Collura (2007); Yun et al. (2008)
Subtract	Corrects the offset by shortening phases (typically all phases), subject to minimum green time, pedestrian intervals, and an allowed decrease in the cycle length	Eagle EPAC300, Econolite ASC2S, Siemens NextPhase, Naztec Model 980	Shelby et al. (2006)
Minimax/ smoothed/ Shortest/best way	Implements either Add or Subtract transition, choosing whichever method is determined to get into sync earlier	170E, 2070 ATC, Eagle EPAC300, Econolite ASC2S, Siemens NextPhase, Naztec Model 980	Shelby et al. (2006); Obenberger and Collura (2007); Yun et al. (2008)
Short way	Decreases proportionally the green interval for each phase by a predetermined time prior and cycles through the signal timing plan until the coordination or offset point is reached; this process is repeated with the decreased cycle length until this offset point is achieved		Obenberger and Collura (2007)
Fixed exit phase control	Traffic controllers can select only one set of exit phases after emergency pre-emption	ASC/3-2100	Yun et al. (2011)
Dynamic exit phase control	The exit phases can be changed dynamically after the EVP operation according to the location of the EVP call within the local cycle timer using the programming function available in advanced traffic controllers	ASC/3-2100	Yun et al. (2011)

EVP State of the Art

A fair amount of new theoretical approaches are proposed in the literature to improve the basic concept of EVP presented in Fig. 1 such as route-based dynamic preemption. Summary of such studies is presented below.

Multiple Preemption Requests

Head et al. (2006) presented that in case of multiple priority requests at an intersection, first-come, first-served policy results in a higher total delay to the vehicles receiving priority than would result, if no attempt to provide priority were made at all. Head et al. (2006) and He et al. (2011) proposed an analytical formulation to best serve multiple priority requests by deciding phase times for each cycle and by allocating different priority request to different signal cycles, respectively. Both these studies aimed at minimizing the total priority delay and do not guarantee the highest priority vehicle to be served first. Thus, Asaduzzaman and Vidyasankar (2017) proposed to first sort multiple conflicting EVP requests based on their priority weight and higher priority requests get the schedule first.

Route-Based Strategy

Most of the state of practice EVP systems operate on a single-intersection basis and depend on of the EV at each intersection—also referred to as the local detection—to activate preemption request at each intersection. However, on a saturated or oversaturated network, after locally detecting an EV approaching the intersection, there may not be enough time available to clear the intersection for an EV to pass without stopping or reducing its speed and this results in a delay at intersections.

For avoiding such situations and improving the efficiency of EVP system, Kwon and Kim (2003) and Gedawy et al. (2008) proposed a route-based signal preemption method that identifies an optimal route for an EV in real-time and clears intersections on a designated EV route. The data from the loop detectors are used to quantify the travel cost of each link through time and to find minimum-cost-path. These studies assumed the existence of a reliable congestion detection system that can measure the congestion level on the roads of the network and report it to the central server. However, practicality and reliability of congestion detection system and eventually route-based EVP system has not been tested on the field.

Once the optimum route is known, Mu et al. (2018) proposed to use a multi-objective programming model to obtain phase durations at each intersection along the route. The objectives can be the minimal residence time of EV at all intersections and the maximal passing numbers of general vehicles. Researchers like He et al. (2014); Kang et al. (2014) suggest that providing maintaining coordination among signalized intersections along EV's route can improve the travel time of EV.

Contrary to predetermining EV's route, Kamalanathsharma and Hancock (2010) proposed sequential preemption, that is, when the EV enters the corridor, it is detected, and preemption is requested to the first intersection controller which then transmits the request to downstream controllers. Using real-time or historic queue measurements, preemption offset for each intersection on response route is calculated and sequential preemption is set to provide an uninterrupted way to EV. Similar logic has been implemented in STREAMS EVP system (Transmax, 2018).

Research also suggests that the emergence of connected vehicle (CV) technologies such as vehicle to vehicle (V2V) and vehicle to infrastructure (V2I) communication can also benefit the EVP operations. For example, Noori et al. (2016) proposed an EVP strategy considering V2I communication is available in EV as well as in all passenger cars on the road. For a given route of EV the proposed strategy considers all traffic signals along the EV route, and finds the number of vehicles stopped at each traffic signal, and based on these number, changes the signal status to provide early green before the EV reaches the signal. Noori et al. (2016) proposed preemption strategy showed a significant reduction in EV's travel time when evaluated in a large-scale realistic simulation environment.

Other than the aforementioned research on various aspect of EVP, some of the researchers have also focused on evaluating the performance of EVP operation. Hanbali and Fornal (2004) assessed the performance of the EVP system operating in the city of Milwaukee, US, that is, OPTICOM emergency response by Global Traffic Technologies LLC. (2020). Though the study found that the optimal priority control efficiently serves the need of EV minimizing the disruption to other traffic on the road, several constraints exist while implementing the optimal priority control strategy that is unique to each intersection depending on the geometry. For instance, the existence of lift bridge, unusual pedestrian phasing, etc. Teng et al. (2012) used GPS data from fixed-route transit vehicles in the Las Vegas area to evaluate the impact of EVs on arterial traffic speeds. The results indicate that speed variance is significantly higher during EVP and the mean speeds of traffic flowing in the same direction as the emergency vehicle and on crossing streets are lower during preemption than during normal conditions. Moreover, signal preemption during peak periods and duration of pre-emption had a significant negative impact on traffic speeds. Also, the transition time has a negative impact on traffic speeds.

Conclusion

Emergency vehicle pre-emption algorithm is used to give EV a priority green light at an intersection. The three critical goals of EVP operation are to reduce EV's response time, ensure the safety of EV as well as other vehicles on the road and minimize interruption to normal traffic on the road. EVP operation includes three major steps, namely, detection of EV, the transition of traffic signal into

preemption and out of preemption. EVP technology has been deployed in various developed countries that have appropriate infrastructure such as infrared signals for communication between EV and signal control, AVL systems in EVs. Evaluation of EVP systems at a various location has shown a significant reduction in travel time of EV.

Research has been conducted to further improve performance of EVP systems such as optimally serving multiple EVP requests at an intersection, avoid EVs getting caught behind the queued vehicles at an intersection due to failing to clear queues at the intersection before EV's arrival. Research suggests optimizing the allocation of priority requests in case of multiple requests instead of serving on first come first basis. Researchers have also suggested applying EVP strategies to route instead of detecting EV and using EVP strategy at each intersection individually. However, these advancements in EVP operation and corresponding improvements lacks real-world validation. Also, interruption to general traffic due to EVP and its impact on the performance of the road network is rarely evaluated. Consequently, further research is required to improve the strategies for transition out of pre-emption which can lead to minimum interruption to general traffic. Lastly, the impact of EVP on the safety of EV as well as other vehicles on the road network needs more attention.

References

- Asaduzzaman, M., & Vidyasankar, K. (2017). A Priority Algorithm to Control the Traffic Signal for Emergency Vehicles. Paper presented at the Vehicular Technology Conference (VTC-Fall), 2017 IEEE 86th.
- Gedawy, H.K., Dias, M., Harras, K., 2008. Dynamic path planning and traffic light coordination for emergency vehicle routing. *Comp. Sci. Dept., Carnegie Mellon Univ., Pittsburgh, Pennsylvania, USA*.
- Global Traffic Technologies LLC. (2020). OPTICOM emergency response. Available from <https://www.gtt.com/emergency-response/>.
- Hanbali, R.M., Fornal, C., 2004. Emergency vehicle preemption signal timing: Constraints and solutions. *Inst. Transp. Eng J.* 74 (7), 32.
- He, Q., Head, K.L., Ding, J., 2011. Heuristic algorithm for priority traffic signal control. *Transp. Res. Rec.* 2259 (1), 1–7.
- He, Q., Head, K.L., Ding, J., 2014. Multi-modal traffic signal control with priority, signal actuation and coordination. *Transp. Res. Part C: Emerg. Technol.* 46, 65–82.
- Head, L., Gettman, D., Wei, Z., 2006. Decision model for priority control of traffic signals. *Transp. Res. Rec.* 1978, 169–177.
- Intelligent Transport Systems.* (2014). Available from https://www.sae.org/journals/publications/yearbook_e/2014/docu/26_intelligent_transport_systems.pdf.
- Intelligent Transport Systems Group., 2005. *Vic roads's specification for Emergency Vehicle Pre-emption System*. Available from <https://www.vicroads.vic.gov.au/business-and-industry/technical-publications>
- Kamalanathsharma, R.K., Hancock, K.L., 2010. *Congestion-based emergency vehicle preemption (No. VT-2009-04). Mid-Atlantic Universities Transportation Center*.
- Kang, W., Xiong, G., Lv, Y., Dong, X., Zhu, F., Kong, Q., 2014. Traffic signal coordination for emergency vehicles. Paper presented at the Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference.
- Kwon, T.M., Kim, S., 2003. Development of dynamic route clearance strategies for emergency vehicle operations, Phase I.
- MAG.(2018). *Emergency Vehicle Preemption Study: Regional Coordination for Unified Operations*. Available from http://azmag.gov/Portals/0/Documents/MagContent/EVP-Study-Report_Final_Nov2018.pdf?ver=2018-11-21-142348-897.
- Miyawaki, M., Yamashiro, Z., Yoshida, T., 1999. Fast emergency pre-emption systems (fast). Paper presented at the Proceedings 199 IEEE/IEEJ/JSAI International Conference on Intelligent Transportation Systems (Cat. No. 99TH8383).
- Mu, H., Song, Y., Liu, L., 2018. Route-based signal preemption control of emergency vehicle. *J. Cont Sci Eng* 2018.
- Noori, H., Fu, L., Shiravi, S., 2016. A connected vehicle based traffic signal control strategy for emergency vehicle preemption. Paper presented at the Transportation Research Board 95th Annual Meeting.
- Obenberger, J., Collura, J., 2007. Methodology to assess traffic signal transition strategies for exit preemption control. *Transp. Res. Rec.* 2035 (1), 158–168.
- Shelby, S.G., Bullock, D.M., Gettman, D., 2006. Transition methods in traffic signal control. *Transp. Res. Rec.* 1978 (1), 130–140.
- Shibuya, S., Yoshida, T., Yamashiro, Z., Miyawaki, M., 2000. Fast emergency vehicle preemption systems. *Transp. Res. Rec.* 1739 (1), 44– L50., doi:10.3141/1739-06.
- Teng, H.H., Kwigizile, V., Xie, G., Kaseko, M., Gibby, A.R., 2012. The impacts of emergency vehicle signal preemption on urban traffic speed. *J. Transp. Res. Forum* 49, (1), 69–79.
- Tokuda, K., Ohmori, S., 2011. Demonstration experiments of SAFER (Speedy Ambulance First-aid, Emergency, Rescue operations supporting) system. Paper presented at the 2011 The 14th International Symposium on Wireless Personal Multimedia Communications (WPMC).
- TOMAR Electronics Inc. (2020). STROBECOM II Optical Preemption and Priority Control System. Available from <https://www.tomar.com/products/traffic-control/strobecom-ii/>.
- Transmax. (2018). Case Study: STREAMS Emergency Vehicle Priority (EVP). Available from https://www.transmax.com.au/wp-content/uploads/2018/03/FINAL_EVP-Case-Study_web.pdf
- Paniati, J.F., Amoni, M., 2006. Traffic signal preemption for emergency vehicle: a cross-cutting study. US Federal Highway Administration.
- Wang, Y., Wu, Z., Yang, X., Huang, L., 2013. Design and implementation of an emergency vehicle signal preemption system based on cooperative vehicle-infrastructure technology. *Adv. Mech. Eng.* 5, 834976.
- Yun, I., Best, M., Brian Park, B., 2008. Evaluation of transition methods of the 170E and 2070 ATC traffic controllers after emergency vehicle preemption. *J. Trans. Eng.* 134 (10), 423–431.
- Yun, I., Park, B.B., Lee, C.K., Oh, Y.T., 2011. Investigation on the exit phase controls for emergency vehicle preemption. *KSCE J. Civil Eng.* 15 (8), 1419–1426, doi:10.1007/s12205-011-1326-2.

Further Reading

- Qin, X., Khan, A.M., 2012. Control strategies of traffic signal timing transition for emergency vehicle pre-emption. *Transp. Res. Part C: Emerg. Technol.* 25, 1–17.

Signalized Roundabouts

Yeti Sazi Murat*, Rui-jun Guo[†], *Pamukkale University, Faculty of Engineering, Civil Engineering Department, Denizli, Turkey; [†]School of Transportation Engineering, Dalian Jiaotong University, Dalian, Liaoning, China

© 2021 Elsevier Ltd. All rights reserved.

Introduction	227
Types of Signalized Roundabouts	228
Entrance Metering at Roundabouts	228
Nearby Vehicular and Pedestrian Signals	229
Entrance Signalization	229
Full Signalization of the Circulatory Lane	229
The Design Parameters	232
Geometric Elements	232
Storage Area of Central Island for Left Turn Vehicles	232
Traffic Volume of Left-Turning Vehicles	232
Signal Timing and Phase Plan	233
Design Methods	233
Biographies	235
Relevant Websites	236
References	236
Further Reading	236

Introduction

Traffic congestion and environmental pollution are among the most common problems that are encountered in many cities in the world. To minimize the negative effects of these problems, different types of traffic management approaches have been taken into account for many years. Roundabouts are one of the intersection types used for these purposes. The priority control approach is the main idea of traditional roundabouts. However, traditional roundabouts are suitable for low-traffic volumes. When traffic volume increases and flow is unbalanced, a roundabout cannot automatically adjust to meet traffic demand and queue length and delays are greatly increased and even the junction can be locked (Guo and Lin, 2011). Unbalanced traffic flow problem is one of the most important problems for roundabouts and needs some additional control application such as metering signals, etc. Unbalanced traffic flow conditions are observed when the traffic volume values in the junction entrance are very different from each other (Akcelik et al., 1996, 1997, 1999). In several countries, including China, United Kingdom, France, Sweden, Australia, Netherlands, and Turkey, some roundabouts have signal control, especially where the vehicle and pedestrian traffic are heavy and unbalanced. This application can be defined as a combination of signalized intersections and roundabouts (Murat et al., 2019).

Signalized roundabouts are used in cases of an increase in demand for traffic flows or violation of traffic rules by drivers in general. In the first case, because of an increase in demand on some arms (unbalanced flow), the capacity of the intersection may decrease dramatically and excessive delays may occur. Vehicle delays can be decreased and traffic safety can be improved by the use of signalized roundabouts (Pilkko et al., 2017). Another main reason to use signalized roundabouts is related to driver behavior at roundabouts without signals. In some developing countries, drivers are not accustomed to roundabouts and often violate the traffic rules. Because of this violation, bottlenecks occur. The bottlenecks of a roundabout with two or more circulatory lanes are the weaving sections, where the vehicles enter or leave the roundabout. Thus, signal control is integrated to prevent possible traffic accidents and increase intersection capacity.

Signalized roundabouts have been used in the world since the 1950s. The use of signal control and design procedure for roundabouts is somewhat different for the countries.

In Europe, the use of signalized roundabouts is common. However, design criteria and reasons for use are different. In the United Kingdom, the design criteria of signal-controlled roundabouts have been summarized in a document as a series of local transportation notes (TSO, 2009). In this note, the first implementation is stated in 1959 for the UK. The reasons for the use of signal control for roundabouts are investigated and the safety issue for both the pedestrians/cyclists and the cars are stated as one of the main reason in this document. Partial control of a roundabout is often employed where delays do not occur on all arms. It is also stated that, in some cases, signal control is used a part-time basis in the UK (TSO, 2009).

The capacity and traffic safety issues are the main reasons for traffic signal implementation on roundabouts in the Netherlands (DHV, 2009). Some reasons include:

- The insufficient overall capacity of the roundabout;
- Heavy left-turning traffic flow;
- Unacceptable delays and/or queues on one or more of the connecting legs.

Sometimes the problems to be solved are (also) traffic safety-related:

- Due to high-circulating speeds drivers experience difficulties when merging with the traffic on the multilane roundabout;
- For pedestrians and cyclists, it is difficult to cross the multilane legs of the roundabout.

In the Netherlands, the implementation of a traffic signal on roundabouts is classified into four categories:

1. Traffic signals meter traffic on one or more entries;
2. Traffic signals control entries and exits for the benefit of crossing pedestrians and cyclists;
3. Traffic signals control both the entries and the circulating lanes;
4. Traffic signals on a turbo roundabout

Two-phase control is generally preferred to provide alternating green waves for east-west and north-south through traffic. In the case of heavy left-turn movements, four-phase control is used.

The use of roundabouts is common in France. There are now more than 25,000 roundabouts in France (Guichet, 2005). The French policy is to favor roundabouts over signalized intersections. The French guidelines for the design of urban intersections allow for signalization of roundabouts for pedestrian safety.

In the United States, signalization of roundabouts is discouraged. The roundabout informational guide of Federal Highway Administration states “roundabouts should never be planned for metering or signalization.” (FHWA, 2000). However, the guide does concede that “unexpected demand” may require signalization after a roundabout is constructed. The FHWA guide goes on to describe three signalization alternatives to be considered: (1) metering, (2) nearby pedestrian signals, and (3) full signalization of the circulatory roadway. In addition to the three warrants for signalization suggested in the roundabout information guide, the guide also points out that signals may be used where “disabled pedestrians and/or school children are present at high volume” (FHWA, 2000).

In China and Turkey, signalized roundabouts are quite common and are familiar to drivers. For both countries, the reasons for using signalized roundabouts can be expressed as traffic safety and congestion problems in general. The problem of traffic congestion and junction gridlock may be observed when drivers violate the priority rule at the roundabout during peak hours. However, traffic accidents sometimes occur due to a violation of the rule. Signalized roundabouts are used as solutions to both problems. Generally, full signaling is preferred.

Types of Signalized Roundabouts

Roundabouts by priority control are self-organized intersections. The signs at the intersection indicate that vehicles entering the roundabout need to decelerate and circulatory vehicles already on the circulatory lanes have priority. Although yield control of entries is the default at roundabouts, traffic flows at roundabouts have been signalized by metering one or more entries, or signalizing the circulatory lane at each entry when necessary. Especially with the increase in traffic volume, priority-controlled roundabouts tend to be congested and difficult to meet the high traffic demand. The use of a signal system for roundabouts may be classified in many ways. In some applications, it is only used for pedestrians (Azhar and Svante, 2011). On the other hand, vehicle traffic is targeted in many applications. Types of signalized roundabouts are examined in the following Fig. 1

Entrance Metering at Roundabouts

Metering signals is one of the applications of signalized roundabouts. Roundabout metering signals help to create gaps in the circulating stream to solve the problem of excessive queuing and delays caused by unbalanced flow patterns and high-demand flow levels. (Akcelik, 2005)

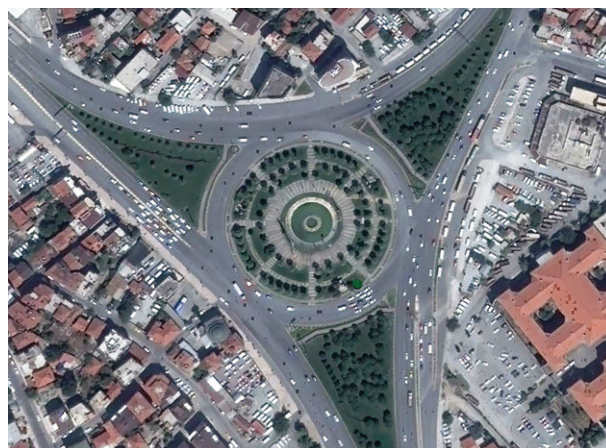


Figure 1 Sample signalized roundabout from Denizli, Turkey. Source: Google Earth, 2015.

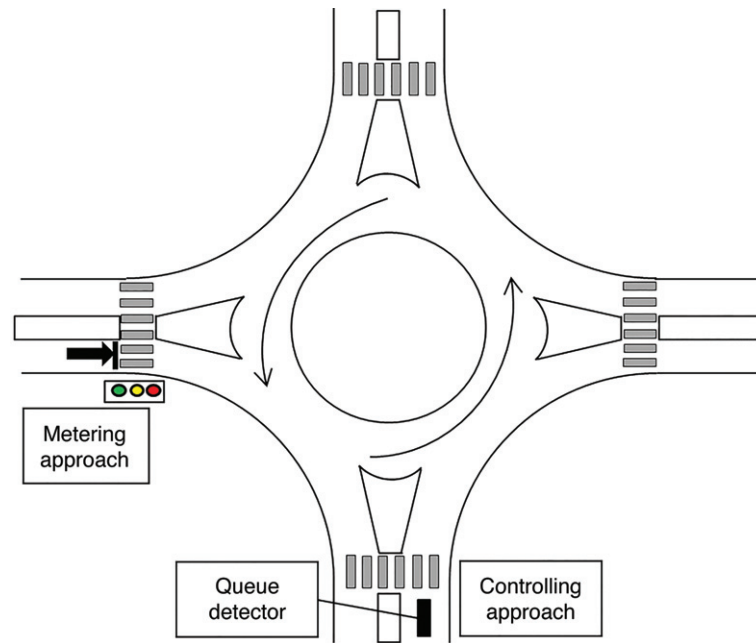


Figure 2 Entrance metering at roundabouts.

When low capacity conditions occur during peak periods, for example, due to unbalanced flow patterns, the use of metering signals is a cost-effective measure to avoid the need for a fully signalized intersection treatment. Roundabout metering signals are often installed on selected roundabout approaches and used on a part-time basis since they are required only when heavy demand conditions occur during peak periods. The concept of metering at roundabouts is similar to ramp metering on freeways. In general, signal light and queue detectors are set on two adjacent entrances, which named as metered approach and controlling approach (Akcelik, 2006). The entrance metering at roundabouts is depicted in Fig. 2.

When the queue on the controlling approach extends back to the queue detector, the signals on the metered approach display red to create a gap in the circulating flow. This helps the traffic flows on the controlling approach to enter the roundabout. When the red display is terminated on the metered approach, the roundabout reverts to normal operation (Akcelik, 2011).

Metering signals have been used in Australia, United Kingdom, and United States to alleviate the problem of excessive delay and queuing by creating gaps in the circulating stream. The necessity and effectiveness of the application of roundabout metering are considered in the literature (Natalizio, 2006, Brabender and Vereeck, 2007). In these studies, it is concluded that roundabouts are an important option to reduce traffic accidents and the reduction of traffic accidents is closely related to speed limits at main roads and bypass roads. It is also stated that signalized roundabouts are the safest type of intersection.

Nearby Vehicular and Pedestrian Signals

Another method of metering (refer to Robinson et al., 2000) is the use of a nearby upstream signalized intersection or a signalized pedestrian crossing on the subject approach road. Unlike entry metering, such controls may stop vehicles from entering and leaving the roundabout, so expected queue lengths on the roundabout exits between the metering signal and the circulatory lane should be compared with the proposed queue space (FHWA, 2000). Sample application is shown in Fig. 3.

Entrance Signalization

In this application, traffic flows on each entrance are controlled by a single phase. All vehicles on each entrance are released simultaneously. There is no conflicting traffic flows in this control method. The green light duration can be set according to the traffic volumes of each entrance. It is possible to better balance the traffic flow.

A four-phase release mode is used in entrance signalization generally to minimize the loss at the four-entrance roundabouts. The phase diagram is shown in Fig. 4.

Full Signalization of the Circulatory Lane

In the case of full application, traffic signals implemented on all arms of the roundabout and around the main circle.

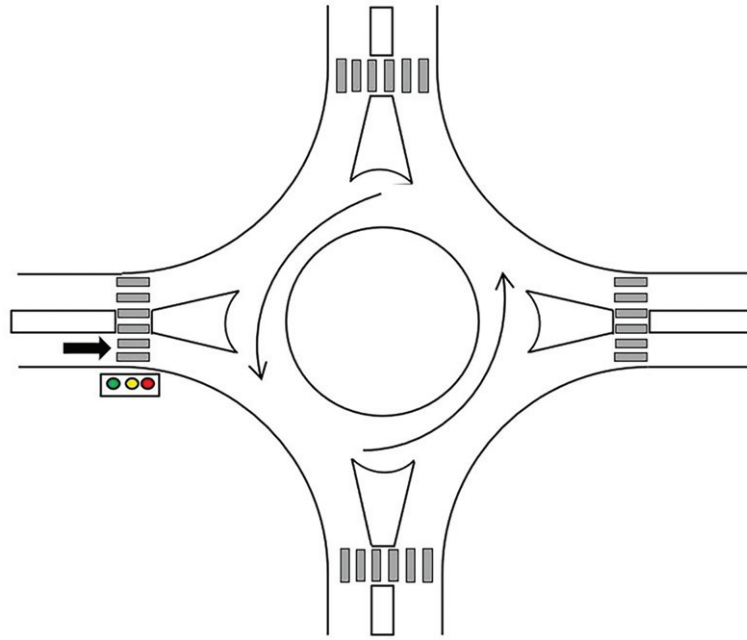


Figure 3 Nearby pedestrian signal at roundabout.

Full signalization including control of circulating traffic is possible at large-diameter multilane roundabouts that have adequate storage space on the circulatory lanes. The full signalized control eliminates all conflict points and weaving sections and adapts to roundabouts with two or more circulatory lanes.

Generally, it is used in the cases of an increase in demand for traffic flows, violation of traffic rules by drivers and traffic safety for pedestrians. In the second case, because of the violation of traffic rules by drivers, grid-lock is occurred, the capacity of intersection may decrease dramatically and excessive delays may occur. Besides violation of traffic rules, left-turning volumes, and traffic composition (heavy vehicle rate) have an adverse effect on the capacity of signalized roundabouts. To overcome these deficiencies, a proper signal timing design should be needed. On the other hand, storage area of the central island should be determined appropriately for traffic volumes of left turn vehicles. The storage area is defined as the space that is determined regarding geometric dimensions of the left turn lanes around the central island. In [Figure 5](#), a sample application and storage area are shown.

In the field, different control types may be used for full signalization of roundabout. The full signalization may be used in the rush time of the morning and evening periods. The signal light may be turned off in the off-peak period and priority control may be applied to improve the LOS. However, the partial signalization at roundabouts will probably increase potential safety problems.

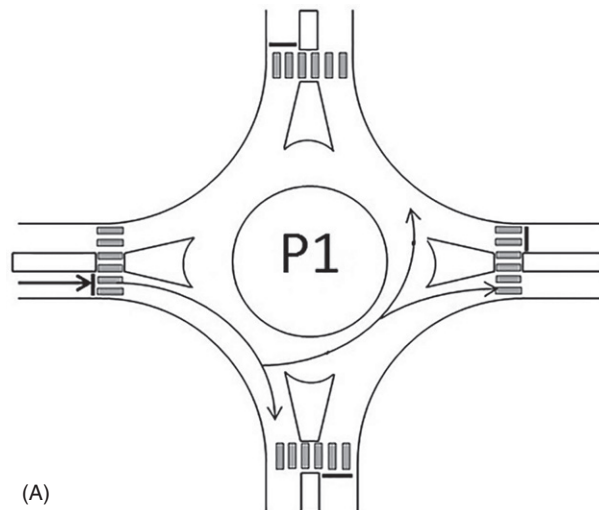
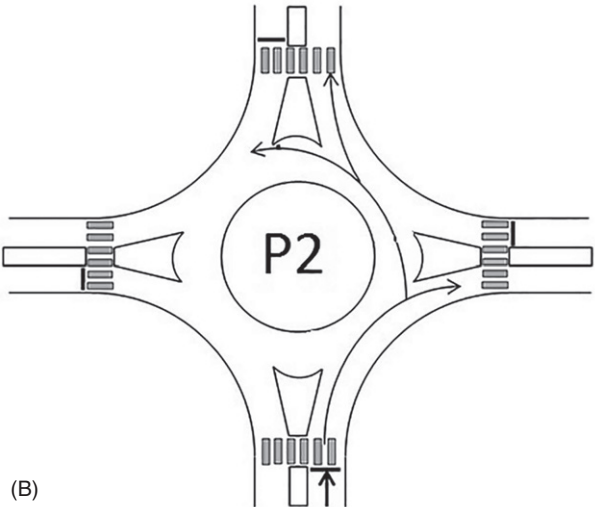
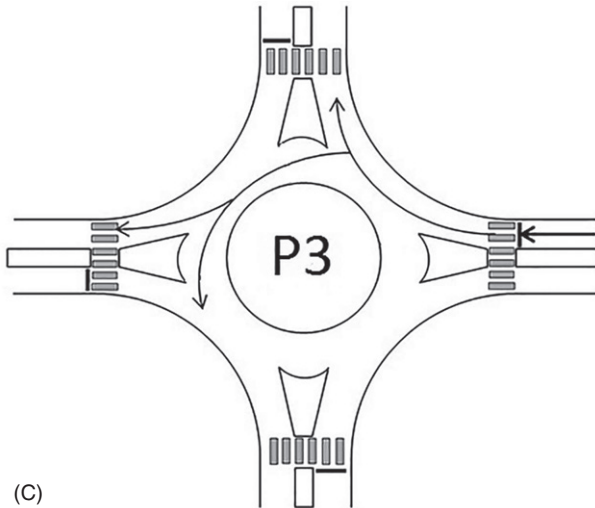


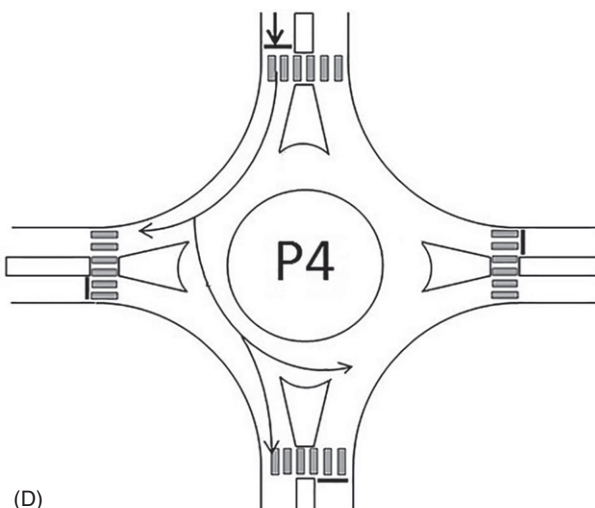
Figure 4 Four phase control in entrance signalization: (A) Phase 1, (B) Phase 2, (C) Phase 3, (D) Phase 4



(Continued)



(Continued)



(Continued)

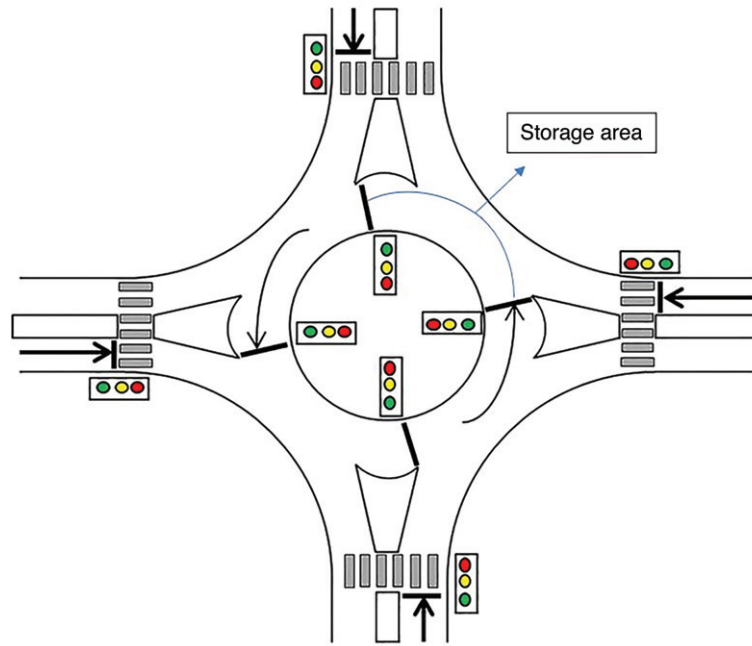


Figure 5 Full application of signalized roundabout.

The Design Parameters

In the design and operation stages of signalized roundabouts, certain parameters should be taken into consideration. These parameters may be summarized as follows:

- Geometric elements
- Traffic volume of left-turning vehicles
- Storage area of the central island for left-turn vehicles
- Signal timing and phase plan

In the design stage, the parameters stated above should be regarded in detail by field observations. Otherwise, the designer has to make some assumptions about the values of these parameters. In this case, the values assumed might not be compatible with those obtained from the field, which can result in an improper design. In the following, the parameters listed above are examined and discussed.

Geometric Elements

Geometric elements of a roundabout have a significant effect on design. The geometric elements or components such as turning radius, lane width, the radius of the central island, etc., should meet the required standards. The design of an intersection may have some faults when using non-standardized elements. On the other hand, the intersection may operate below capacity because of these design faults. Many studies in literature support these findings. An investigation about the effects of the diameter of signalized roundabouts and the cycle time on vehicle delay revealed that the increased radius of the central island leads to the rise of average delay.

Storage Area of Central Island for Left Turn Vehicles

The storage area of the central island for left-turning vehicles is shown in Fig. 5 and defined in Section Full Signalization of the Circulatory Lane. This area should have some specifications to serve the traffic flows with minimum effects. The number of circulation lanes and the dimensions of the circulation lanes is the main elements to consider. The designer should consider these elements and collect the required data from the field, such as the dimensions of the space for left-turning traffic flows.

Traffic Volume of Left-Turning Vehicles

The left-turning traffic volumes have some considerable effects on the design of signalized roundabouts. The number of vehicles and traffic composition should be observed for a proper design. The signal timing for circulating flows should be determined based on

the traffic volumes and especially the percentage of heavy vehicle. On the other hand, the variations of traffic volumes should also be observed both in peak and off-peak periods and the trend of this variation should be obtained. Otherwise, signal timing assigned may not be sufficient for the traffic volumes and some vehicles may have additional delays. Moreover, traffic jams may also occur because of this improper signal timing design.

The left-turn heavy traffic volumes should be taken into account carefully. The intersection may be congested if the percentage of heavy vehicles is higher than expected. On the other hand, because of the heavy vehicles, the departure headways of vehicles may be affected and grid-lock may be occurred (Nikitin et al., 2017; Murat et al., 2019). The observations and estimations of heavy vehicles should be done well in the design stage.

Signal Timing and Phase plan

The signal timing assignment and phase plan selection has a great impact on the performance of the intersection control system. Although much software for the simulation and design of intersections is developed, signal timings of signalized roundabouts are limited. On the other hand, some microsimulation software may be used only for performance analysis, not for signal timing design. Therefore, signal timing design is a challenge for designers and it may be varied based on the designer's ability and foresight.

Signalized roundabouts are generally controlled by two, three, or four-phased plans. Two-phased management is preferred when the left-turn flows are not significant. Three-phased management is generally used at the intersection of secondary roads and main roads. The phase plans have a considerable effect on the performance of signal control systems. It may be determined by field observations and some specific analysis. Left turning flows, pedestrians, and traffic composition (heavy vehicle rates) are some important parameters that should be considered in phase plan selection. Any negligence of these parameters may lead to unexpected performance results. Thus, design engineers need certain information or a guide to develop more sustainable solutions for roundabouts with signals.

Design Methods

The design of signalized roundabouts generally needs experience in traffic management. In the design, some software such as SIDRA Intersection, TRANSYT, ARCADY, VISSIM, LinSig, etc. may be helpful. In the past, approaches developed for conventional signalized junctions have been used to control signalized roundabouts. In this context, two-phase management and four-phase management have been adopted more. Two-phase management is preferred in the case of low left-turn traffic flows; if left-turn flows increase, four-phase management is preferred. In some cases where conventional design approaches are not appropriate, new approaches are needed. The change in left-turn traffic flows during the day or the traffic safety problems in the weaving areas can be stated as the main reasons for the need for new management approaches. However, the design procedure including signal timing is not properly defined in the literature. In some limited researches, this issue is taken into consideration and design methods are proposed. Yang et al., 2005 proposed a two-stage left-turn control method of roundabout based on the minimum average delay. The method is suitable to calculate the optimal cycle time under the unsaturated traffic flow. Cakici and Murat (2016) proposed an iterated signal timing method in which the optimal cycle length can be calculated by iterative progress. Both of them are full signalization methods of a roundabout.

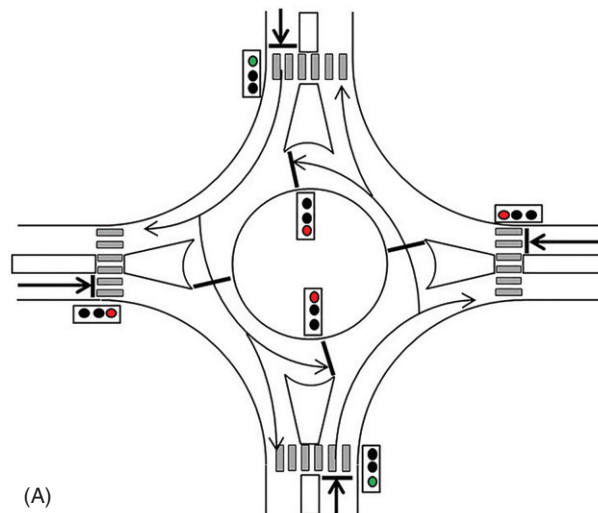
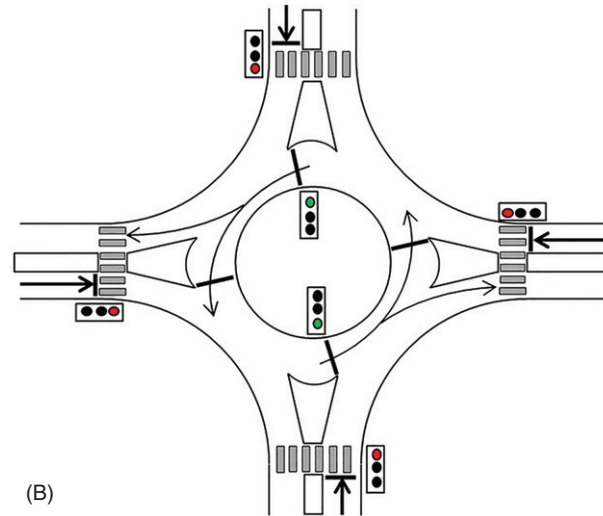
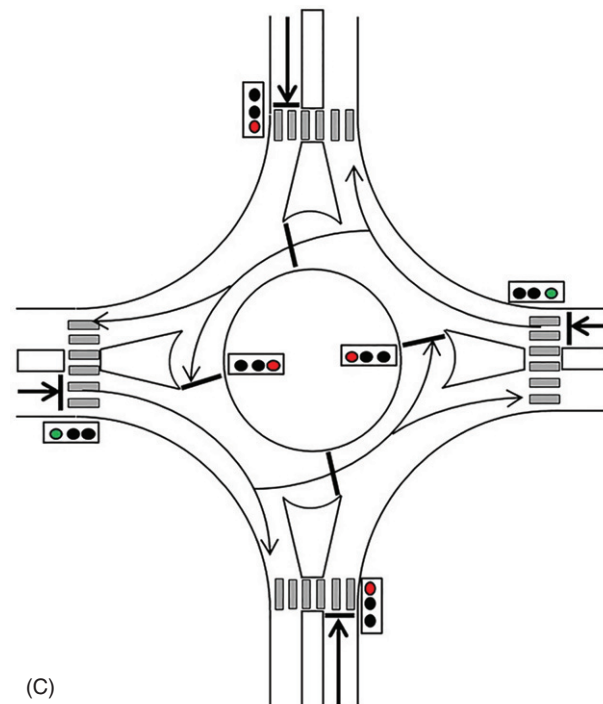


Figure 6 Four phases in the two-stage left-turn control: (A) Phase 1, (B) Phase 2, (C) Phase 3, (D) Phase 4.



(Continued)



(Continued)

Two-stage left-turn control is typical full signalization of circulatory lane at roundabouts. In addition to the stop line on the approach, the second stop line on the circulatory lane is set up to control the left-turn traffic flow. Traffic signals instruments are installed before each stop line to eliminate the weaving flow conflicts on the circulatory lane. Left-turn vehicles on the circulatory lane will stop before red signals to avoid weaving.

The two-stage left-turn control method has four phases. In phase I, vehicles on the south and north entrances enter the roundabout. Turn-left vehicles stop in front of the second stop line. In phase II, the waiting vehicles before the second stop-line, from the north and south entrances, run out of the circulatory lanes. In phase III, vehicles on the east and west entrances enter the roundabout. Left-turn vehicles stop in the front of the corresponding second stop line. In phase IV, the waiting vehicles before the second stop-line, from the east and west entrances run out of the circulatory lanes. The phases diagram is shown in [Fig. 6](#).

In the method proposed by [Cakici and Murat \(2016\)](#), a formula (Equation 1) is developed for the design of signal timings regarding the storage area of the central island and circulating traffic flows. Signal timing of central island is calculated by this

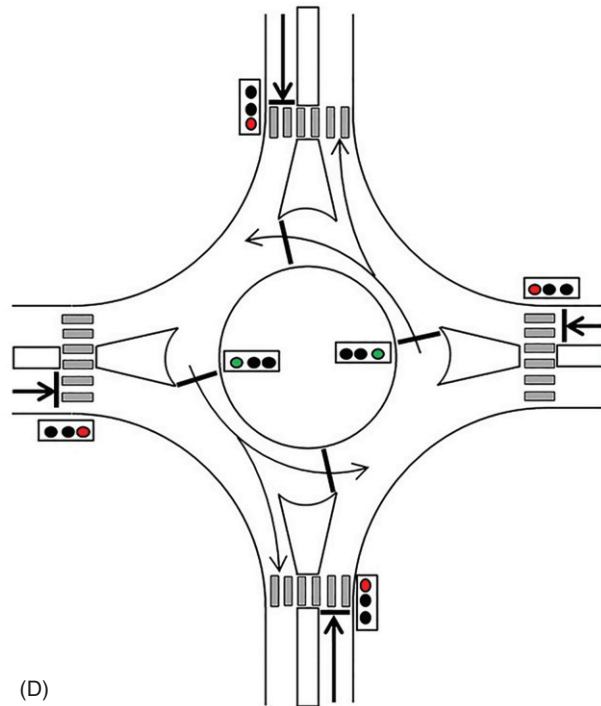


Figure 6 (Continued)

formula. Different phase options are investigated in the research regarding some left-turn traffic flow scenarios and encouraging results are obtained by the proposed method. Detailed information about the method can be found in [Murat et al. \(2019\)](#).

$$\phi = \alpha + [(n - 1) \times \lambda] + \varepsilon \quad (1)$$

Where;

(ϕ): green time of central island (sec),

α : the lost times of departing vehicles in the storage area (sec),

n : the number of lines in storage area,

λ : departing headways of vehicles (sec),

ε : calibration coefficient for signal timing

For the past two decades, although the numbers of signalized roundabouts have been increased worldwide, traffic researchers still have many questions and uncertainties related to the operation and design of signalized roundabouts. Further research is needed to address improving capacity and LOS by considering environmental, economic and social effects. On the other hand, the variation of pedestrians and vehicle traffic flows may be considered in future works. Investigation of different signal timings and phase plans may also be interesting for future studies.

Biographies



Dr. Yetis Sazi MURAT graduated from Civil Engineering Dept. of Dokuz Eylül Univ., İzmir, Turkey in 1992. He has received MSc Degree from Pamukkale University in 1996 and PhD Degree from Istanbul Technical University in 2001. He worked as a visiting scholar at Virginia Tech University, Falls Church, United States in 2006 (6 months) and University of Nevada, Reno, United States in 2017 (6 months). He has published 85 research papers, 7 book chapters, and conducted many projects on different subjects of Transportation Engineering (especially about Traffic Engineering). He is currently working at Pamukkale University, Denizli, Turkey as a full-time Professor of Civil Engineering Department.



Guo Ruijun got a Bachelor's and Master's degree from the Wuhan University of Technology in 1999 and 2003; the PhD in transportation planning and management from Beijing Jiaotong University in 2013. He worked as a Post-Doctoral Scholar at Research Institute of Highway Science, China Ministry of Transportation from 2013–16. He worked as a visiting scholar at the University of Nevada, Reno, United States from 2017–18. He has published 40 research papers, two books, and two invention patents. Now he is an Associate Professor, Department of Transportation Engineering, Dalian Jiaotong University.

Relevant Websites

<https://earth.google.com>
Google Earth

References

- Akcelik, R., Chung, E., Besley, M., 1996. Performance of roundabouts under heavy demand conditions. *Road Transp. Res.* 5 (2), 36–50.
- Akcelik, R., Chung, E. and Besley, M., 1997. Analysis of roundabout performance by modelling approach flow interactions. *Proceedings of the Third International Symposium on Intersections Without Traffic Signals*, July 1997, Portland, Oregon, USA, pp 15–25.
- Akcelik, R., Chung, E. and Besley, M., 1999. Roundabouts: Capacity and Performance Analysis. Research Report ARR No. 321. ARRB Transport Research Ltd, Vermont South, Australia (2nd Ed. 1999).
- Akcelik, R., 2005 Capacity of Performance Analysis of Roundabout Metering Signals, TRB National Roundabout Conference, Vail-Colorado, pp. 1–19.
- Akcelik, R. 2006. Operating Cost, Fuel Consumption and Pollutant Emission Savings at a Roundabout with Metering Signals, 7th International Congress on Advances in Civil Engineering (ACE 2006), Istanbul-Turkey, pp. 1–14.
- Akcelik, R., 2011. Roundabout metering signals: Capacity, performance, and timing. *Proc. Soc. Behav. Sci.* 16, 686–696.
- Azhar, Al-M., Svante, B., 2011. Signal control of roundabouts. *Proc. Soc. Behav. Sci.* 16, 729–738.
- Brabender, B.D., Vereeck, L., 2007. Safety effects of roundabouts in Flanders: Signal type, speed limits and vulnerable road users. *Accid. Anal. Prev.* 39 (3), 591–599.
- Cakici, Z., Murat, Y.S., 2016. A new calculation procedure for signalized roundabouts and performance analysis. *Tech. J. Turk. Cham. Civil Eng.* 27 (4), 7569–7592 (in Turkish).
- DHV, 2009. Roundabouts: Application and Design, A Practical Manual, Ministry of Transport, Public Works and Water management Partners for Roads, The Netherlands. p. 104.
- FHWA, 2000. Roundabouts: An Informational Guide, US Dept of Transportation, Federal Highway Administration, McLean Virginia, USA, p. 277.
- Guichet, B. 2005. Roundabouts in France: Safety and New Uses. Presentation at National Roundabout Conference, Vail, CO.
- Guo, R.J., Lin, B.L., 2011. Gap acceptance at priority-controlled intersections. *J. Transp. Eng.* 137 (4), 269–276.
- Murat, Y.S., Cakici, Z., Tian, Z., 2019. A signal timing assignment proposal for urban multi-lane signalized roundabouts. *J. Croat. Asso. Civil Eng. Grad.* 71 (2), 113–124.
- Natalizio, E., 2006. Roundabouts with Metering Signals, Institute of Transportation Engineers 2005 Annual Meeting, Melbourne-Australia, pp. 1–10.
- Nikitin, N., Patskan, V., Savina, I., 2017. Efficiency analysis of roundabout with traffic signals. *Transp. Res. Proc.* 20, 443–449.
- Webb, P.J., 1994. SIG-NABOUT- the development and trial of a novel junction design, *IEE Conf. Pub.*, 391; 106–110.
- Pilko, H., Mandzuka, S., Baric, D., 2017. Urban single-lane roundabouts: A new analytical approach using multi-criteria and simultaneous multi-objective optimization of geometry design, efficiency, and safety. *Transp. Res. Part C Emer. Tech.* 80, 257–271.
- Robinson B W, Rodegerdts L, Scarborough W, et al., 2000. Roundabouts: An Informational Guide. NHCPR Report, 70.
- The Stationery Office (TSO), 2009. Signal Controlled Roundabouts, Local Transport Note 1/09, Department for Transport, UK.
- Yang, X., Li, X., Xue, K., 2005. A new traffic signal control for modern roundabouts: Method and application. *IEEE Trans. Intel. Transp. Sys.* 5 (4), 282–287.

Further Reading

- Anderson and K. W. Martin., 1994. Traffic flow improvements at Newbridge Roundabout. *IEE Conf. Pub.*, 391:101–105.
- Bai, Y., Chen, W., Xue, K., 2010. Association of Signal-Controlled Method at Roundabout and Delay, 2010 International Conference on Intelligent Computation Technology and Automation (IEEE), Changsha, pp. 816–820.
- Bie, Y., Mao, C., Yang, M., 2016. Development of vehicle delay and queue length models for adaptive traffic control at signalized roundabout. *Proc. Eng.* 137, 141–150.
- Cheng, W., Zhu, X., Song, X., 2016. Research on capacity model for large signalized roundabouts. *Proc. Eng.* 137, 352–361.
- Coelho, M.C., Farias, T.L., Roupail, N.M., 2006. Effect of roundabout operations on pollutant emissions. *Transp. Res. Part D Transp. Environ.* 11 (5), 333–343.
- Gardziejczyk, W., Motylewicz, M., 2016. Noise level in the vicinity of signalized roundabouts. *Transp. Res. Part D Transp. Environ.* 46, 128–144.
- Halami, H., Aghayan, I., 2017. Traffic efficiency evaluation of elliptical roundabout compared with modern and turbo roundabouts considering traffic signal control. *Prom. Traffic Transp.* 29 (1), 1–11.
- Johnnie, B.E., Ahmed, A., Iman, A., 2012. Extent of delay and level of service at signalized roundabout. *Int. J. Eng. Technol.* 2 (3), 419–424.
- Ma, W., Liu, Y., Head, L., Yang, X., 2013. Integrated optimization of lane markings and timings for signalized roundabouts. *Transp. Res. Part C Emer. Technol.* 36, 307–323.
- Ma, W., Sun, X., Huang, W., 2015. Comparative study on the capacity of a signalised roundabout. *IET Intel. Transp. Sys* 10 (3), 175–185.
- Maher, M., 2008. The optimization of signal settings on a signalized roundabout using the cross-entropy method. *Comp.-Aided Civil Infrastruct. Eng.* 23 (2), 76–85.
- Qian, H., Li, K., Sun, J., 2008. The Development and Enlightenment of Signalized Roundabout, 2008 International Conference on Intelligent Computation Technology and Automation (IEEE), Hunan, pp. 538–542.

- Shawaly, E.A.A., Li, C.W.W., Ashworth, R., 1991. Effects of entry signals on the capacity of roundabout entries. A case-study of Moore Street roundabout in Sheffield. *Traffic Eng. Contr.* 32 (6), 297–301.
- Sisiopiku, V., Oh, H., 2001. Evaluation of roundabout performance using SIDRA. *J. Transp. Eng.* 127 (2), 143–150.
- Tracz, M., Chodur, J., 2012. Performance and Safety of Roundabouts with Traffic Signals, SIIV-5th International Congress – Sustainability of Road Infrastructures (*Procedia – Social and Behavioral Sciences*), 53, pp. 788-799.
- Vaughan W.I., Davis G.W., 2007. Synthesis of Literature Relevant to Roundabout Signalization to Provide Pedestrian Access. FHWA Report.

Turbo Roundabouts: Design, Capacity, and Comparison With Alternative Types of Roundabouts

Marco Guerrieri, Raffaele Mauro, DICAM, University of Trento, Trento, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	238
Performance Analysis of Turbo Roundabouts	238
Values of the Critical Gap and Follow-Up Time in Turbo Roundabouts	240
Capacity Calculation	240
Delays and Level of Service Estimation	243
Comparison With Other Types of One and Two-Level Roundabout Intersections	244
References	246

Introduction

In turbo roundabouts traffic entry, exit, and circulatory lanes are physically delimited by kerbs (see Fig. 1A). This implies a lower number of potentially conflicting points (Fig. 1B) between vehicle trajectories than conventional roundabouts (Fortuijn, 2009). By way of an example, in a turbo roundabout with four arms and two circulatory lanes, the number of conflicting points is equal to ten, while in a conventional roundabout with similar geometric characteristics there are 22 conflicting points (see Table 1). In many European countries, turbo roundabouts are utilized more and more. The geometric guidelines have been published in the Netherlands, Slovenia, Germany, Serbia, Slovakia, and Croatia.

Several turbo-roundabout layouts may be used: egg, basic turbo, knee, spiral and rotor, stretched-knee, and star. Undoubtedly, the most used is the *basic* turbo roundabout (Fig. 1A). The typical layout of the central island and circulatory lanes is generally designed through arcs of circumferences with different centers and radii, as shown in Fig. 2 for mini, standard, medium, and large *basic* turbo roundabouts (CROW, 2008). A lot of design guidelines lay down that the axis connecting the centers C_1 and C_2 in Fig. 2 must be inclined by around 30 degree to the axis of the minor road. In addition, ellipse arcs (Fig. 1C) or the Archimedean spiral can be used with the purpose for limiting the variation of the centrifugal acceleration around the circulatory lanes.

The radii values (Fig. 2) and the lane widths must be selected in such a way that the traffic speed along the turbo roundabout does not exceed 40 km/h.

Experimental measures on the Slovenian turbo roundabout of Fig. 1C ($R_{\max} \approx 25$ m) show that instantaneous vehicle speeds at entry are rather moderate (lower than 25 km/h, 15 m well ahead of the yield line) and accelerations are always lower than 2 m/s^2 at both entry and circulatory lanes (where they are generally below 1.5 m/s^2). Only at exit lanes accelerations prove to be slightly higher than 2 m/s^2 . As a rule, the following accidents may occur at double-lane roundabouts: failure to yield collisions, crashes for vehicle control loss, rear-end at entry, carriageway run-off, circulating-exiting collisions, and side-by-side collisions. The main difference between turbo roundabouts and conventional layouts lies in the lack of circulating-exiting and side-by-side conflicts, which are avoided with lane dividers, both at entries and on the circulatory lane. Compared to double-lane roundabouts, turbo roundabouts can provide the following potential benefits:

- reductions in the number of total potential accidents to between 40% and 50%;
- reductions in the number of potential accidents with injuries to between 20% and 30%.

Therefore, turbo roundabouts are more appropriate than conventional double-lane roundabouts in the cases requiring a higher safety level, for instance, in urban contexts where the traffic level of pedestrians and two-wheeler riders is very significant (Mauro et al., 2015).

Performance Analysis of Turbo Roundabouts

Turbo roundabouts are appropriate for daily traffic volumes of the order of 20,000–32,500 veh/day entering an intersection. The turbo-roundabout capacity is well higher than mini and single-lane roundabouts; it can be compared to the two-lane roundabout capacity and is lower than the signalized roundabout capacity, as can be seen from Fig. 3 (Brilon et al., 2014).

In order to study intersection functionality, it generally needs to determine either the total roundabout capacity or the simple single entry capacity, average vehicle delays, entry queue lengths, levels of service (LoS), and where required, pollutant emissions from traffic.

Functionality analyses can be carried out by traffic microsimulation softwares (Aimsun, Vissim, etc.) or closed-form models.

The calculation procedure implying the adoption of closed-form capacity models is schematized in the flowchart of Fig. 4. The physical lane division requires that the functionality of every driving lane be studied, in other words, it involves a lane-by-lane

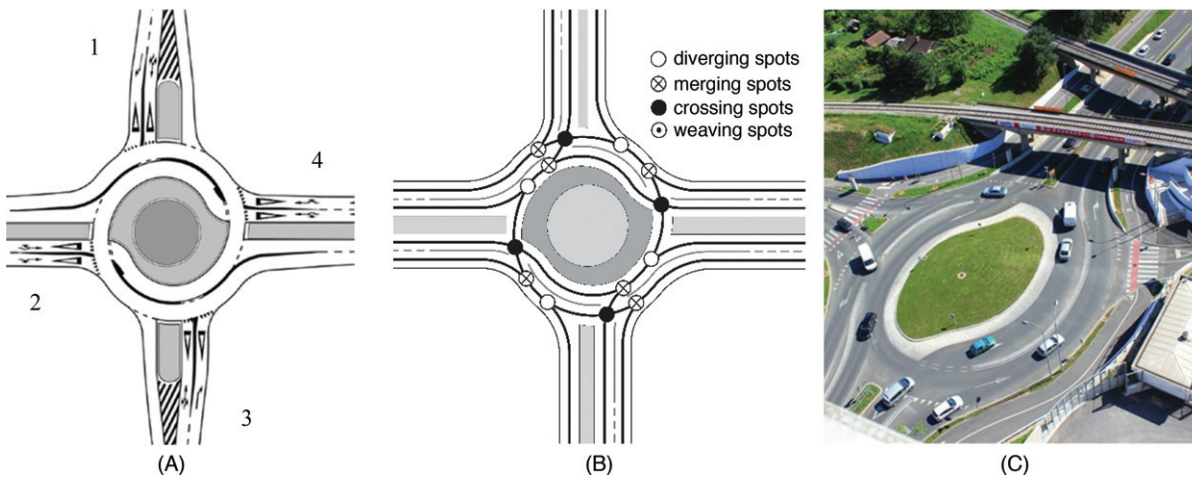
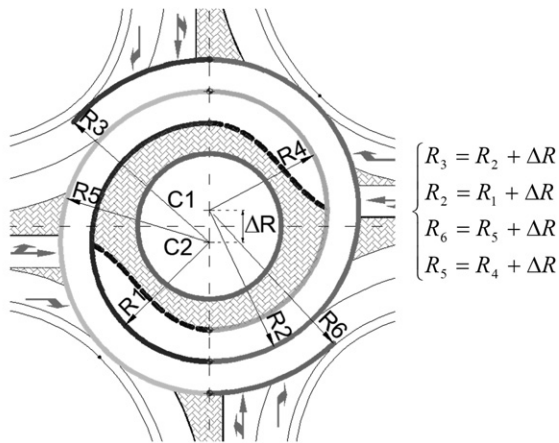


Figure 1 (A) Basic turbo-roundabout layout; (B) potential conflict points; (C) first turbo roundabout in Slovenia.

Table 1 Conflict points at unsignalized intersection, two-lane roundabout, and turbo roundabout

Number of arms	Number of conflicting points		
	Unsignalized intersection	Two-lane roundabout	Turbo roundabout
3	9	16	7
4	32	22	10



$\Delta R = 4.20$ m (Lane width = 3.50 m)				
Radius	Mini	Standard	Medium	Large
R_1 [m]	10.50	12.00	15.00	20.00
R_2 [m]	14.70	16.20	19.20	24.20
R_3 [m]	18.90	20.40	23.40	28.40
R_4 [m]	10.50	12.00	15.00	20.00
R_5 [m]	14.70	16.20	19.20	24.20
R_6 [m]	18.90	20.40	23.40	28.40
$\Delta R = 4.45$ m (Lane width = 3.75 m)				
R_1 [m]	10.50	12.00	15.00	20.00
R_2 [m]	14.95	16.45	19.45	24.45
R_3 [m]	19.40	20.90	23.90	28.90
R_4 [m]	10.50	12.00	15.00	20.00
R_5 [m]	14.95	16.45	19.45	24.45
R_6 [m]	19.40	20.90	23.90	28.90
$\Delta R = 4.70$ m (Lane width = 4.00 m)				
R_1 [m]	10.50	12.00	15.00	20.00
R_2 [m]	15.20	16.70	19.70	24.70
R_3 [m]	19.90	21.40	24.40	29.40
R_4 [m]	10.50	12.00	15.00	20.00
R_5 [m]	15.20	16.70	19.70	24.70
R_6 [m]	19.90	21.40	24.40	29.40

Figure 2 Example of turbo-roundabout design criterion and typical radii values.

analysis. More specifically, through the capacity models based on the gap-acceptance theory, the preliminary step is to identify formulations for calculating the single-lane capacity at each entry (in function of the corresponding circulating flow). The second step is:

- To assess the impact of heavy vehicles on traffic (generally dependent on heavy vehicle types, heavy vehicle percentages, and total flow);
- To properly choose drivers' behavioral parameters, with special regard to the critical gap and the follow-up time;
- To estimate how a pedestrian flow may impact on single entry capacities.

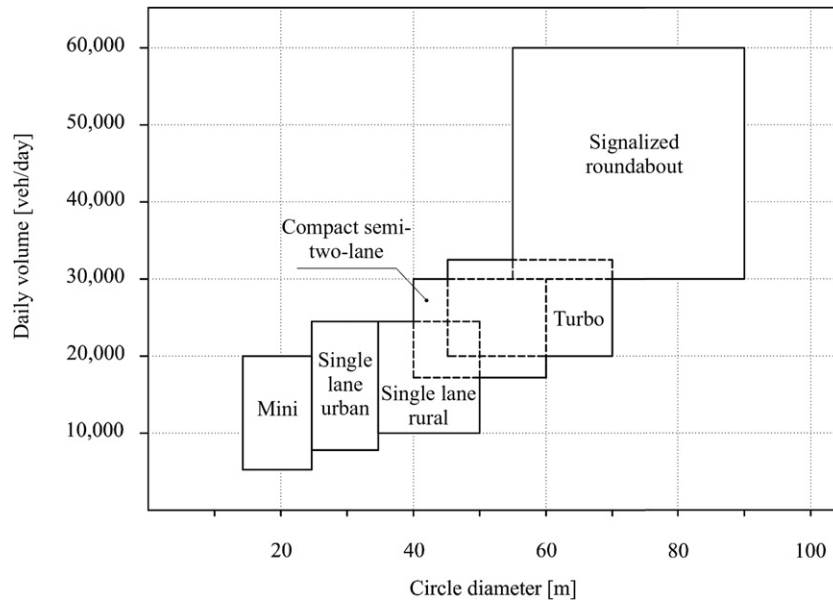


Figure 3 Choice of roundabout types according to of daily volume.

These aspects are set out in detail in the further sections.

Values of the Critical Gap and Follow-Up Time in Turbo Roundabouts

The critical gap t_c and the follow-up time (or follow up headway) t_f play a crucial role in analyzing the functionality of unsignalized at-grade intersections in both conventional and turbo roundabouts (see further sections). These parameters are involved in the capacity expressions based on the gap-acceptance theory (the Harders formula and its derivatives). The t_c and t_f values estimated in studies performed in Europe since 2009 are shown in Fig. 5 (Guerrieri et al., 2018; Guerrieri et al., 2019; Macioszek, 2017). A recent research on how users behave in a roundabout (Fig. 1C) was conducted in Slovenia by analyzing vehicle trajectories on major and minor entry lanes with the digital image processing technique and estimated the following characteristic headways:

- critical gap: $t_c = 4.03\text{--}5.48$ s (obtained from 832 sample data);
- follow up time: $t_f = 2.52\text{--}2.71$ s (obtained from 1452 sample data)

Such values agree well with those provided by previous studies and shown in Fig. 5. In addition, the average values of psychotechnical times associated to heavy vehicles are always higher than the corresponding values in light vehicles. The histograms in Fig. 6 show t_c and t_f entry values for the minor and major arms (see Fig. 1C), distinguished for light and heavy vehicles. Depending on the entry lane, it can be observed that t_c values of heavy vehicles are higher by 15% ÷ 47% than those of light vehicles, instead t_f values of heavy vehicles are higher by 18% ÷ 25% than those of light vehicles.

Capacity Calculation

The layout of turbo roundabouts and the separation of circulatory lanes and entry arms require that the capacity of each entry lane be estimated. As turbo roundabouts are more appropriate in urban contexts, the capacity estimation model here suggested takes into account the different traffic components allowed to circulate in urban contexts (i.e., light vehicles, heavy vehicles, and pedestrians). Pedestrian crossings, located at arms near roundabout entries, can significantly influence the entry capacity. The models based on the gap-acceptance theory assume that circulatory lanes are always undersaturated. Thus, there is an essential need for the flow Q_i crossing the i th lane to be always lower than its capacity C_i . In other words, the saturation degree must be: $x_i = Q_i/C_i < 1$. The capacity of the circulatory lane is 1400–1600 veh/hour in single-lane sections (i.e., opposite entries 2 and 4 in Fig. 1A) and below 2500 veh/hour in two-lane sections (i.e., opposite entries 1 and 3 in Fig. 1A). Turbo roundabout entry lane capacity can be estimated through the following formulas:

- *Right-turn lane*: Tanner equation (adjusted by Brilon-Wu)

$$C_{E,R} = 3600 \cdot \left(1 - \frac{t_{\min} \cdot q_k}{3600}\right) \cdot \frac{1}{t_f} \cdot e^{-\frac{q_k}{3600} \left(t_c - \frac{t_f}{2} - t_{\min}\right)} \quad (1)$$

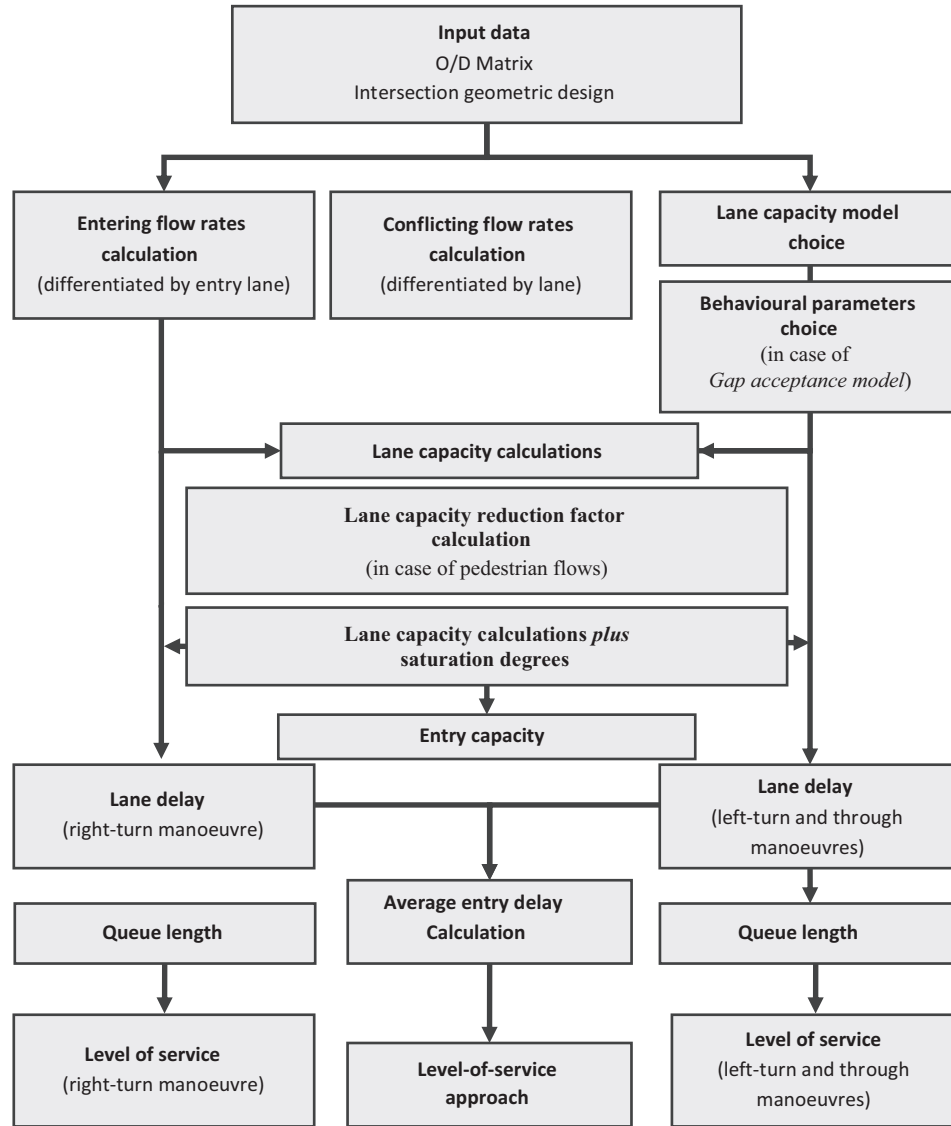


Figure 4 Functionality analysis flowchart for basic turbo roundabouts.

where: $C_{E,R}$ = the right-turn lane capacity at entry E [veh/h]; q_K = circulating vehicles opposite entry E, just next to the right-turn lane R [veh/h]; t_c = critical gap [s]; t_f = follow-up time [s] (t_c and t_f values have been specified in the above section), and $t_{\min} = 2.1$ seconds denotes the least headway between vehicles moving along the circulatory lanes [s].

- Through and left-turn lanes: Harders Model

$$C_{E,TLT} = q_k \cdot \frac{e^{-q_k \frac{t_{c,x}}{3600}}}{1 - e^{-q_k \frac{t_{f,x}}{3600}}} \quad (2)$$

with $C_{E,TLT}$, capacity of through and left-turn lanes; q_k , conflicting flow rate; $t_{c,x}$ and $t_{f,x}$ critical gap and follow-up time, respectively; $q_k = Q_{c,e} + Q_{c,i}$; $Q_{c,e}$ = circulating flow on the outer circulatory lane; $Q_{c,i}$ = circulating flow on the inner circulatory lane. The critical gap ($t_{c,x}$) and the follow-up time ($t_{f,x}$) can be assumed as: $t_{c,x} = 6,4$ s; $t_{f,x} = 3,5$ s.

The passenger car equivalents (PCEs) can reasonably assume the following values: 1 heavy vehicle = $1.5 \div 2$ cars (depending on heavy vehicle types and traffic flow), 1 heavy truck = 2.5 cars; 1 articulated truck = 3 cars, 1 bicycle/motorbike = 0.5 cars.

- Effect of pedestrian flow on entry capacity

Study name	Entry type description	Figure	Site number /country	Sample size (for t_c)	Critical headway t_c [s]	Follow-up time t_f [s]
Fortuijn, 2009	- two entry lanes, and two circulating lanes		1 / Netherlands	580	3.40 ± 0.31 (left lane)	2.26 (left lane)
			2 / Netherlands	363	3.54 ± 0.25 (left lane)	2.13 (right lane)
			3 / Netherlands	131	3.60 ± 0.28 (left lane)	
	- two entry lanes, and two circulating lanes		1 / Netherlands	115	3.28 ± 0.23 (right lane)	2.13
			2 / Netherlands	424	3.35 ± 0.32 (right lane)	2.13
			3 / Netherlands	280	3.55 ± 0.48 (right lane)	
Brilon, 2014	- two entry lanes, and two circulating lanes		n/a Germany	n/a	4.0 (left lane)	2.6 (left lane)
					4.5 (right lane)	2.7 (right lane)
	- one entry lane and one circulating lane		n/a Germany	n/a	4.50	2.5
	- one lane at the entry and two circulating lanes		n/a Germany	n/a	4.30	2.8
	- two lanes at the entry, one circulating lane		n/a Germany	n/a	4.50 (left lane)	2.5
					4.50 (right lane)	2.5
Macioszek, 2017	- two entry lanes, and two circulating lanes		n/a Poland	n/a	3.21–4.27 (left lane)	3.23 (left lane)
			n/a Poland	n/a	3.59–4.35 (right lane)	3.83 (right lane)
	- one entry lane and one circulating lane		n/a Poland	n/a	3.07–3.85	4.35–4.61
	- one lane at the entry and two circulating lanes		n/a Poland	n/a	3.07–4.50	4.35–4.62
	- two lanes at the entry, one circulating lane		n/a Poland	n/a	3.28–4.66 (left lane)	5.08–5.61 (left lane)
					3.59–4.35 (right lane)	4.37–4.76 (right lane)

Figure 5 Critical headways and follow-up times at turbo roundabouts.

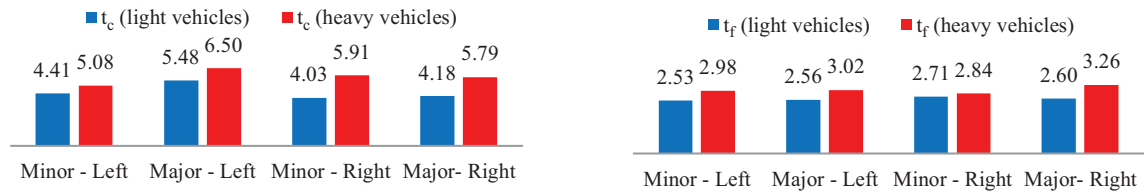


Figure 6 Average values of the critical gap (t_c) and follow up time (t_f) for light and heavy vehicles.

The German method can be used (Mauro, 2010). The model implies the determination of a capacity reduction factor M . For a single entry lane the reduction factor M is as follows:

$$M = \frac{1119.5 - 0.715 \cdot q_k - 0.644 \cdot Q_{ped} + 0.00073 \cdot q_k \cdot Q_{ped}}{1069 - 0.65 \cdot q_k} \quad (3)$$

where: q_k = conflicting flow rate (pcu/h); Q_{ped} = pedestrian flow (ped/h).

By particularizing to turbo roundabouts, it results:

$$M_{E,R} = \frac{1119.5 - 0.715 \cdot Q_{c,e} - 0.644 \cdot Q_{ped} + 0.00073 \cdot Q_{c,e} \cdot Q_{ped}}{1069 - 0.65 \cdot Q_{c,e}} \quad (4)$$

$$C_{E,R}^{ped} = C_{E,R} \cdot M_{E,R} \quad (5)$$

$$M_{E,TLT} = 1119.5 - 0.715 \cdot (Q_{c,e} + Q_{c,i}) - 0.644 \cdot Q_{ped} + 0.00073 \cdot \frac{(Q_{c,e} + Q_{c,i}) \cdot Q_{ped}}{1069 - 0.65 \cdot (Q_{c,e} + Q_{c,i})} \quad (6)$$

$$C_{E,TLT}^{ped} = C_{E,TLT} \cdot M_{E,TLT} \quad (7)$$

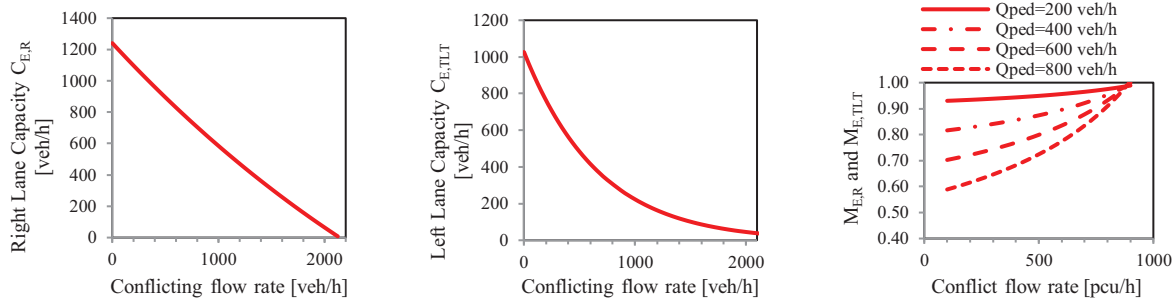


Figure 7 Comparison between models for entry lane capacities and capacity reduction factor values.

where: $M_{E,R}^{ped}$ = right-turn lane pedestrian capacity reduction factor; $M_{E,TLT}^{ped}$ = through and left-turn lane pedestrian capacity reduction factor; $C_{E,R}^{ped}$ = right-turn lane vehicle capacity considering impact on pedestrians [veh/h]; $C_{E,TLT}^{ped}$ = through and left-turn lane vehicle capacity considering impact on pedestrians [veh/h]; $C_{E,R}$ = right-turn lane vehicle capacity (no pedestrian crossings, only vehicles) [veh/h]; $C_{E,TLT}$ = through and left-turn lane vehicle capacity (no pedestrian crossings, only vehicles) [veh/h].

Fig. 7 shows the entry lane capacities and reduction factors in function of the circulating flow rate.

Each entry lane at a turbo roundabout is characterized not only by different capacity values (C_i), but also by different flow rates (Q_i); it results that the saturation degree ($x_i = Q_i/C_i$) can differ between lanes of the same entry and the total entry capacity is not a simple sum of the single lane capacities. For these reasons the effective entry capacity C_E^{ped} can be obtained from:

$$C_E^{ped} = \frac{(Q_{E,R} + Q_{E,TLT})}{\max \left[\frac{Q_{E,R}}{C_{E,R}^{ped}}, \frac{Q_{E,TLT}}{C_{E,TLT}^{ped}} \right]} \quad (8)$$

The previous equation shows that capacity of each entry depends on the capacity of the single lane and thus on: (1) conflicting flow; (2) circulating flows at the circulatory carriageway; (3) user behavior (through parameters t_c , t_f , t_{min}); (4) traffic demand balance at entry arms; and (5) pedestrian flow.

Delays and Level of Service Estimation

After calculating the capacity and saturation degree of each lane (in case of lane undersaturation and pedestrian flow), the average delay per vehicle can be estimated at the lane under examination, by means of the following equations (in [Tollazzi et al., 2016](#)):

$$D_{E,TLT}^{ped} = \frac{3600}{C_{E,TLT}^{ped}} + 900 \cdot T \cdot \left[\frac{Q_{E,TLT}}{C_{E,TLT}^{ped}} - 1 + \sqrt{\left(\frac{Q_{E,TLT}}{C_{E,TLT}^{ped}} - 1 \right)^2 + \left(\frac{3600}{C_{E,TLT}^{ped}} \right) \cdot \frac{\left(\frac{Q_{E,TLT}}{C_{E,TLT}^{ped}} \right)}{450 \cdot T}} \right] + 5 \cdot \min \left[\frac{Q_{E,TLT}}{C_{E,TLT}^{ped}}, 1 \right] \quad (9)$$

$$D_{E,R}^{ped} = \frac{3600}{C_{E,R}^{ped}} + 900 \cdot T \cdot \left[\frac{Q_{E,R}}{C_{E,R}^{ped}} - 1 + \sqrt{\left(\frac{Q_{E,R}}{C_{E,R}^{ped}} - 1 \right)^2 + \left(\frac{3600}{C_{E,R}^{ped}} \right) \cdot \frac{\left(\frac{Q_{E,R}}{C_{E,R}^{ped}} \right)}{450 \cdot T}} \right] + 5 \cdot \min \left[\frac{Q_{E,R}}{C_{E,R}^{ped}}, 1 \right] \quad (10)$$

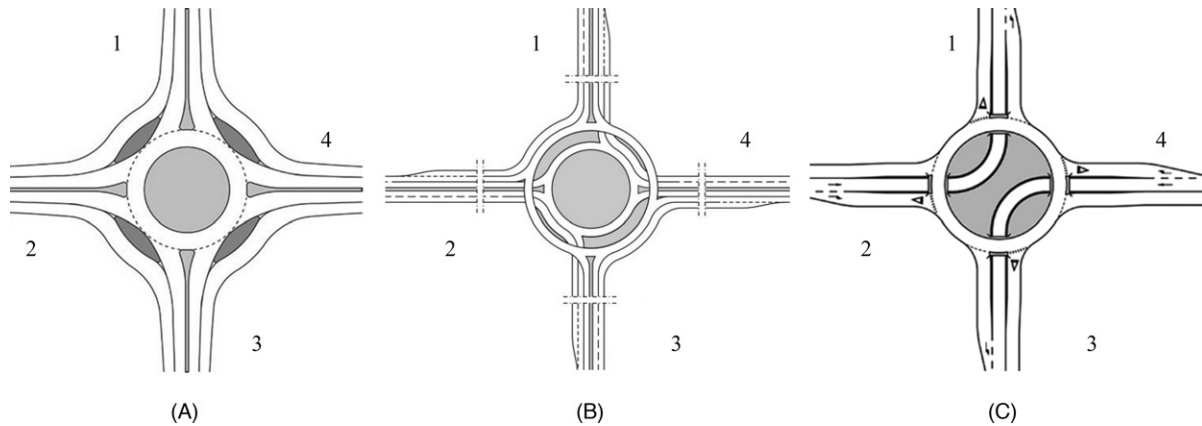
where: $D_{E,R}^{ped}$ = average control delay for the right-turn lane (s/veh); $D_{E,TLT}^{ped}$ = average control delay for through and left-turn lanes (s/veh); T = reference time (h), ($T = 1$ for a 1-h analysis, $T = 0.25$ for a 15-min analysis). Generally speaking, delays will differ at the two entry lanes; so the level of service of the right-turn lane is to be differentiated from the corresponding level of service at the through and left-turn lanes. The total average delay at entries is expressed by the following relation:

$$D_E^{ped} = \frac{D_{E,R}^{ped} \cdot Q_{E,R} + D_{E,TLT}^{ped} \cdot Q_{E,TLT}}{Q_{E,R} + Q_{E,TLT}} \quad (11)$$

where $D_{E,R}^{ped}$, $Q_{E,R}$, $D_{E,TLT}^{ped}$, $Q_{E,TLT}$ are respectively delays and flow rates at the two lanes of entry E. In order to define the level of service at each entry lane, valuable indications can be given by [HCM \(2010\)](#), as shown in [Table 2](#).

Table 2 Level of service

Delay (s/veh)	Level of service by volume-to-capacity ratio	
	$v/c \leq 1$	$v/c > 1$
0–10	A	F
10–15	B	F
15–25	C	F
25–35	D	F
35–50	E	F
> 50	F	F

**Figure 8** (A) Flower roundabout; (B) target roundabout; (C) four flyover roundabout

Comparison With Other Types of One and Two-Level Roundabout Intersections

A turbo roundabout may be preferable in specific traffic conditions, which can be also identified by comparing the following roundabout types:

- Basic turbo roundabout (Fig. 1A)
- Flower roundabout with right-turn bypass lane with yield sign (Flower-yield) or with free-flow right-turn (Flower-free) (Fig. 8A)
- Target roundabout (Fig. 8B)
- Four flyover roundabout (Fig. 8C)
- Standard roundabout with an entry lane and a circulatory lane (1 + 1)
- Standard roundabout with an entry lane and two circulatory lanes (1 + 2)
- Standard roundabout with two entry lanes and two circulatory lanes (2 + 2).

The capacity models presented in the previous sections were used for traffic simulations of turbo roundabouts. On the other hand, the models described in the HCM 2010 manual were used for standard roundabouts; the capacity models used for Flower, Target, and Four flyover roundabouts are explained in Tollazzi et al. (2016).

Three test matrices ρ_1 , ρ_2 , and ρ_3 , for traffic demand, were taken into account. The total entry ranging between 225 veh/h and 4,775 veh/h (equally distributed among the four arms of each intersection) flowed into the road intersections. The origin destination (OD) matrices were:

- OD Matrix ρ_1 = 72% of vehicles turn right, 13% cross, and 15% turn left;
- OD Matrix ρ_2 = 13% of vehicles turn right, 72% cross, and 15% turn left;
- OD Matrix ρ_3 = 15% of vehicles turn right, 13% cross, and 72 % turn left.

$\rho_1 =$	0	0.72	0.12	0.15	$\rho_2 =$	0	0.13	0.72	0.15	$\rho_3 =$	0	0.15	0.13	0.72
	0.15	0	0.72	0.13		0.15	0	0.13	0.72		0.72	0	0.15	0.13
	0.13	0.15	0	0.72		0.72	0.15	0	0.13		0.13	0.72	0	0.15
	0.72	0.13	0.15	0		0.13	0.72	0.15	0		0.15	0.13	0.72	0

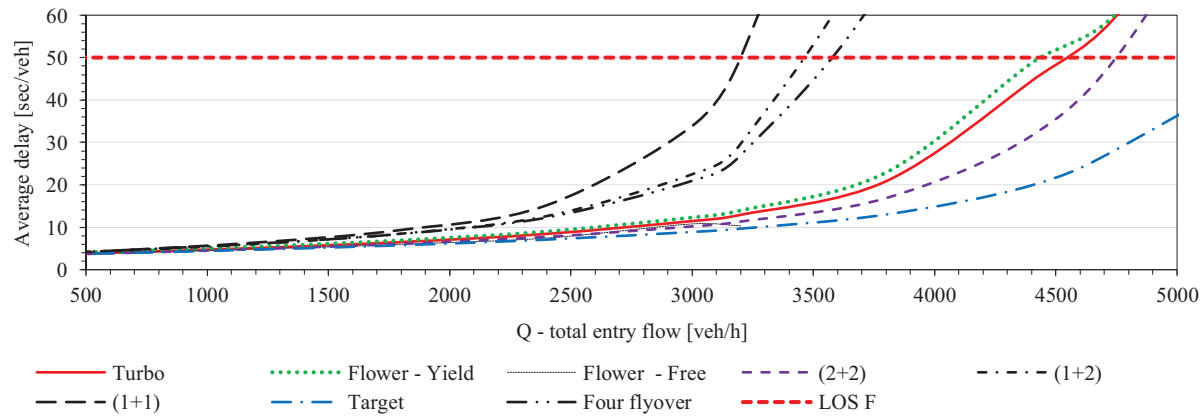


Figure 9 Average delays at roundabouts for OD Matrix p1.

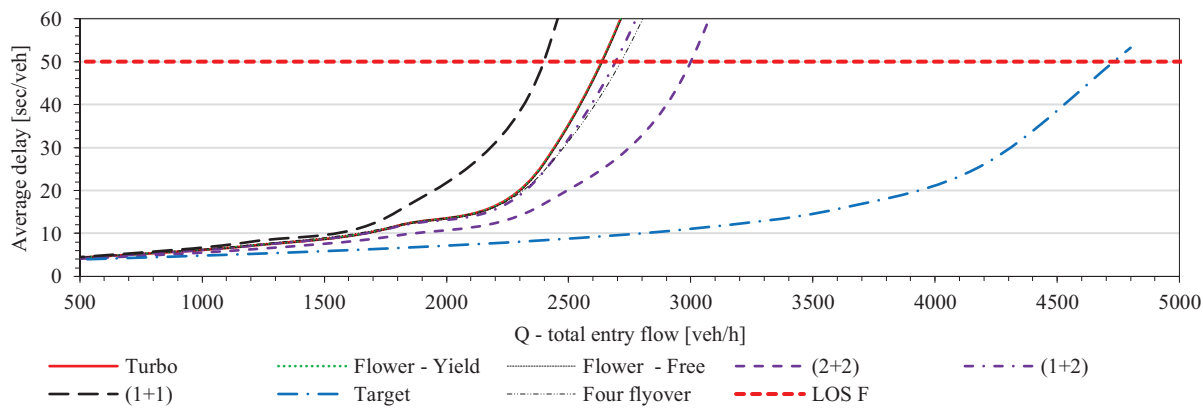


Figure 10 Average delays at roundabouts for OD Matrix p2.

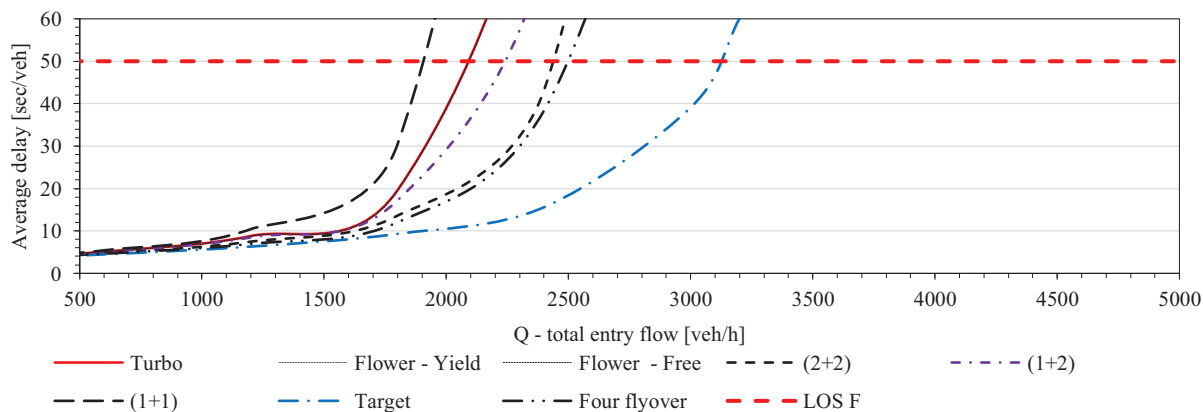


Figure 11 Average delays at roundabouts for OD Matrix p3.

Figs 9, 10, and 11 show the average delays in function of the total entering traffic. As clearly seen, the target roundabout (two-level roundabout, Fig. 8B) produces lower delays in all traffic conditions examined. Among the at-grade intersections, standard roundabouts (2 + 2) provide lower delays than other standard types [(1 + 1) and (1 + 2)]. On the other hand, turbo roundabouts guarantee lower average delays, and thus they would be preferable to at-grade intersections if the entering flow turns mainly right (i. e., OD Matrix p1, see Fig. 9).

References

- CROW: Turborotondes, Publication No. 257. CROW, Netherland (2008).
- Brilon, W., Bondzio, L., Weiser, F. 2014. Experiences with turbo- roundabouts in Germany. In: 5th Proc., Rural Roads Design Meeting, Copenhagen. Bochum, Germany: Brilon Bondzio Weiser.
- Fortuijn, L.G.H., 2009. Turbo roundabouts: Estimation of capacity. *Transp. Res. Rec.* 2130 (1), 83–92.
- Guerrieri, M., Mauro, R., Parla, G., Tollazzi, T., 2018. Analysis of kinematic parameters and driver behavior at turbo roundabouts. *J. Transp. Eng. Part A Sys.* 144 (6), 1–12.
- Guerrieri, M., Mauro, R., Tollazzi, T., 2019. Turbo-roundabout: case study of driver behavior and kinematic parameters of light and heavy vehicles. *J. Transp. Eng. Part A Sys.* 145 (6), 1–11.
- Highway Capacity Manual. 2010. *TRB, National Research Council*, Washington, DC.
- Macioszek, E., 2017. Analysis of significance of differences between psychotechnical parameters for drivers at the entries to one-lane and turbo roundabouts in Poland. *Adv. Intell. Syst. Comput.* 505, 149–161.
- Mauro, R., Cattani, M., Guerrieri, M., 2015. Evaluation of the safety performance of turbo-roundabouts by means of a potential accident rate model. *Baltic J. Road Bridge Eng.* 10, 28–38.
- Mauro, R., 2010. *Calculation of Roundabouts*. Springer-Verlag, Berlin, Heidelberg.
- Tollazzi, T., Mauro, R., Guerrieri, M., Renčelj, M., 2016. Comparative analysis of four new alternative types of roundabouts: “Turbo”, “flower” “target” and “four-flyover” roundabout. *Periodica Polytech. Civil Eng.* 60 (1), 51–56.
- .

Capacity of an Intersection

Ashish Verma*, Milan Mathew Thomas[†], ^{*}Department of Civil Engineering, Indian Institute of Science, Bangalore, Karnataka, India;
[†]Department of Civil Engineering, Rajiv Gandhi Institute of Technology Kottayam, Kerala, India

© 2021 Elsevier Ltd. All rights reserved.

Introduction	247
Concepts of Intersection Capacity	247
The Capacity of Signalized Intersection	248
United States of America	249
India	249
Indonesia	250
Australia	250
United Kingdom	251
Summary	251
The Capacity of Unsignalized Intersection	251
United States of America	252
India	253
Indonesia	254
Summary	254
Capacity of Roundabouts	255
United States of America	255
India	255
Indonesia	256
Summary	256
Conclusion	257
References	257
Further Reading	257

Introduction

The vehicular flow along a given road section is akin to the water flowing through a pipe. The volume of water that can flow through a section of pipe for a given period depends on the geometry of the pipe and the speed of the water. By the same token, along a given road section, the volume of vehicles that can travel is confined by certain factors like road width and speed of the traffic. In general, the capacity of a road at a particular time can be described as the number of vehicles that can travel along the section of the road. This is a key element in governing the planning, analysis, and operation of a road facility. A clear understanding of the capacity helps in a safe, smooth, and economic operation of traffic through the facility. **Figure 1** illustrates the dichotomy of basic traffic stream parameters. The quantitative measure of the traffic stream explains the capacity while the Level-of-Service (LOS)—the level of quality one can derive from the traffic facility—is reflected by the qualitative measures of the traffic stream.

During the initial era of traffic flow research, the concept of capacity was associated with the issue of congestion, and the researchers defined capacity in terms of the maximum traffic volume and density. Even though there exists a complex relationship between demand, volume, and capacity, they are a different measure of traffic. **Figure 2** differentiates the concept of volume, demand, and capacity.

Like other basic concepts, the sense of capacity has evolved. In 1950, Highway Capacity Manual (HCM) formally defined the concept of highway capacity for the first time. It did so by defining three unique capacity values: basic capacity, possible capacity, and practical capacity. They are characterized by the roadway and traffic conditions under which the capacity is estimated. The basic capacity considers the most nearly ideal conditions that can be attained. The prevailing conditions are considered in the possible capacity and the practical capacity also accounts for the prevailing conditions without the traffic density being so high as to cause unreasonable delay, hazard, or restrictions to maneuverability. The 1965 HCM replaced these three capacity values with a single one and revised in the subsequent editions of HCM. According to HCM, the capacity of a system element is the maximum sustainable hourly flow rate at which persons or vehicles reasonably can be expected to traverse a point or uniform section of a lane or roadway during a given time period under prevailing roadway and traffic conditions.

Concepts of Intersection Capacity

An intersection is an element of a road network, where two or more vehicular movements share a common space. This space enables the vehicles to turn to their desired direction and continue to the destination. But the uncontrolled movement of traffic in this area

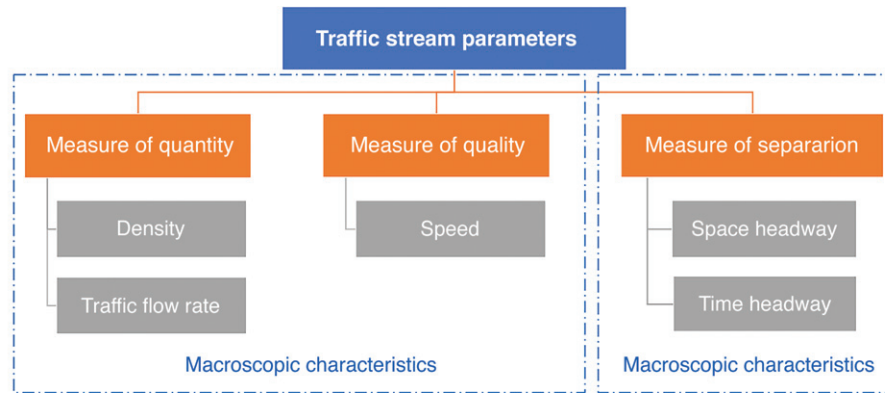


Figure 1 Traffic stream parameters.

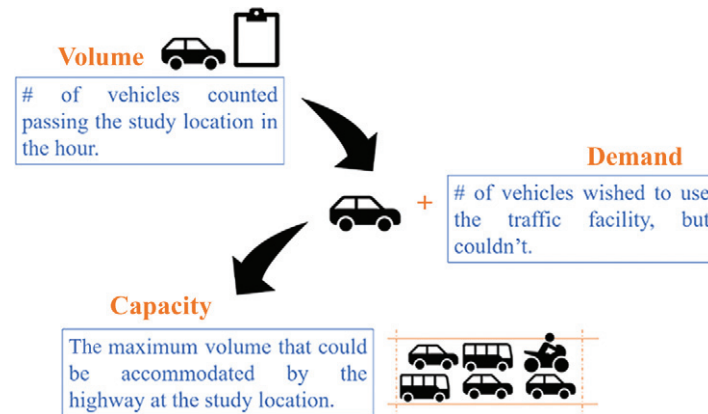


Figure 2 Volume, demand, and capacity.

results in the conflict of vehicles. It is vital to segregate the vehicles so that the potential conflict points are removed. This can be achieved either spatially or temporally, where vehicles find their safe time slot or vehicles are allocated a time slot to make their maneuver. Based on the traffic controlling measures, intersections are categorized into at-grade intersections and grade-separated intersections (Figure 3). Every traffic facility has a specific measure of effectiveness to quantify the quality one can derive from the traffic facility. For grade-separated intersections, the measure of effectiveness used is density and for at-grade intersection, it is control delay.

Just as the road section, intersections also have the capacity. As mentioned earlier, intersection capacity is also influenced by different factors like geometric conditions, traffic flow, vehicular composition, and signal timing. It is more appropriate to discuss the capacity of the intersection as the capacity of individual approaches. The analysis of grade-separated intersection is similar to that of road links. In this chapter, we will focus on the capacity analysis of the at-grade intersections, under the headings of a signalized intersection, unsignalized intersection, and roundabouts.

The Capacity of Signalized Intersection

Signal operations in a road section interrupt the vehicular flow and which causes delays. Saturation flow will reflect this effect of the signalized intersection. The concept of saturation is used for estimating the capacity of the signalized intersection. Saturation flow can be defined as the rate at which the vehicles that have been waiting in a queue during the red interval cross the stop line during the green interval. The steps for estimating signalized intersection capacity can be simplified as follows:

1. Collect the information regarding the geometric and traffic characteristic of intersection along with the control characteristics. Geometric characteristics include the lane width, number of lanes, gradient, parking condition, etc. Traffic characteristics include demand by movement, base saturation flow, peak hour factor, approach speed, pedestrian activity, percentage of heavy vehicles, etc. Cycle length, green time phase plan, etc., comes under the control characteristics.
2. Calculate the saturation flow s_i for each lane group in the intersection.
3. Compute the capacity based on the saturation flow and effective green ratio for the lane group. If the effective green time on an approach i is g_i and the cycle time is C then the capacity on the given approach is given by equation (1).

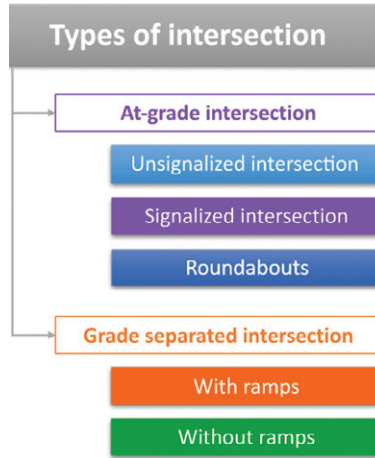


Figure 3 Types of intersection.

$$\text{Capacity of lane group } i = s_i \frac{g_i}{C} \quad (1)$$

Where,

S_i = Saturation flow of the lane group i .

g_i = Effective green time on an approach i .

C = Cycle time

Saturation flow can be either measured directly from the field or estimated theoretically. The later method, estimates the saturation flow by adjusting the base saturation value (under ideal condition) according to the roadway and traffic condition. Estimation of saturation flow is highly location specific and also affected by the driving behavior of people. This spurred several nations to develop their guidelines for measuring and estimating the saturation flow.

United States of America

According to the HCM (2016), saturation flow rate for each lane-group is given by equation (2). By considering a saturation headway of 1.9 seconds, a base saturation flow rate of 1900 passenger cars per hour per lane is usually adopted.

$$s = s_0 \times f_w f_{HV} f_g f_p f_{bb} f_a f_{LU} f_{LT} f_{RT} f_{Lpb} f_{Rpb} \quad (2)$$

Where,

s = adjusted saturation flow rate for the subject lane group in PCU per hour per lane.

s_0 = base saturation flow rate per lane in PCU per hour per lane.

f_w = lane width adjustment factor.

f_{HV} = adjustment factor for heavy vehicles in the traffic stream.

f_g = approach grade adjustment factor.

f_p = adjustment factor for the existence of a parking lane and parking activity adjacent to lane group.

f_{bb} = adjustment factor for the blocking effect of local buses that stop within the intersection area.

f_a = adjustment factor for area type.

f_{LU} = lane utilization adjustment factor.

f_{LT} = adjustment factor for left turns in lane group.

f_{RT} = adjustment factor for right turns in lane group.

f_{Lpb} = pedestrian adjustment factor for left-turn movements.

f_{Rpb} = pedestrian-bicycle adjustment factor for right-turn movements.

India

The base saturation flow in Indian intersection is estimated based on the width of the approach, Indo-HCM (2017) discusses the relationship between the base saturation flow and the width of the approach. Equation (3) gives the base saturation flow rate per unit width of the road. The prevailing saturation flow of the intersection approach for the lane group under consideration is then obtained from equation (4).

Table 1 Base saturation flow by environment class and lane type

<i>Environment class</i>	<i>1</i>	<i>Lane type</i> <i>2</i>	<i>3</i>
A	1850	1810	1700
B	1700	1670	1570
C	1580	1550	1270
<i>Environment classes</i>			
<i>Class A:</i>	<i>Ideal or nearly ideal conditions for free movement of vehicles</i>		
<i>Class B:</i>	<i>Average conditions</i>		
<i>Class C:</i>	<i>Poor conditions</i>		
<i>Lane type</i>			
<i>Type 1:</i>	<i>Through lane</i>		
<i>Type 2:</i>	<i>Turning lane</i>		
<i>Type 3:</i>	<i>Restricted turning lane</i>		

Source: Austroads, 2020

$$USF_0 = \begin{cases} 630; & \text{for } w < 7.0 \text{ m} \\ 1140 - 60w; & \text{for } 7.0 \leq w \leq 10.5 \text{ m} \\ 500; & \text{for } w > 10.5 \text{ m} \end{cases} \quad (3)$$

Where,

 USF_0 = unit base saturation flow rate in PCU per hour per meter. w = effective width of approach in meters.

$$S = w \times USF_0 \times f_{bb} f_{br} f_{is} \quad (4)$$

Where,

 S = adjusted saturation flow rate in PCU per hour. w = effective width of the approach in meter used by the lane group. USF_0 = unit base saturation flow rate. f_{bb} = adjustment factor for bus blockage due to the curbside bus stop. f_{br} = adjustment factor for the blockage of through vehicles by standing right-turning vehicles waiting for their turn. f_{is} = adjustment factor for the initial surge of vehicles due to approach flare and anticipation effect.**Indonesia**

The base saturation flow for different approach types is obtained from the graphs reported in the [Indonesian HCM \(1993\)](#). Then the prevailing saturation flow value is calculated using the equation (5).

$$S = S_o \times f_{CS} \times f_{SF} \times f_G \times f_p \times f_{RT} \times f_{LT} \quad (5)$$

Where,

 S = prevailing saturation flow in vehicles per hour. S_o = base saturation flow. f_{CS} = adjustment factor for city size. f_{SF} = adjustment factor for side friction. f_G = adjustment factor for the gradient. f_p = adjustment factor for parking. f_{RT} = adjustment factor for right turns in the lane group. f_{LT} = adjustment factor for left turns in the lane group.**Australia**

Australian Road Research Board defined base saturation flow by environment class and lane type as shown in [Table 1 \(Austroads, 2020\)](#). This base saturation flow is adjusted to match the prevailing condition to estimate the saturation flow. This method is summarized in equation (6).

Table 2 Summary of the base saturation flow estimation and adjustment factors for prevailing saturation flow

Sl. No	Country	Base saturation flow	No. of adjustment factors
1	United States	Metropolitan area with population > 250,000: 1900 pc/h/ln Otherwise: 1750 pc/h/ln	11
2	Indonesia	Based on approach types (opposed/protected discharge)	6
3	Canada	Based on the location (nine cities) and direction of movement (through and	13
4	India	Based on the width of the lane	3
5	Australia	Based on the environmental class and lane type	3
6	United Kingdom	Prevailing saturation flow is predicted using the model developed by Kimber and coworker.	7

Source: Austroads (2020); Teply et al. (2008); HCM (2016); Indo-HCM (2017); Indonesian Highway Capacity Manual (1993); Kimber et al. (1986)

$$S = \frac{f_w f_g}{f_c} S_0 \quad (6)$$

Where,

S = adjusted saturation flow in vehicles per hour.

S_0 = base saturation flow from **Table 1**.

f_w = lane width adjustment factor.

f_g = gradient adjustment factor.

f_c = Traffic compensation factor.

United Kingdom

The prevailing saturation flow at signalized intersections in the United Kingdom is obtained using the model developed based on the lane geometry and composition of traffic flow. This predictive model is evolved out of the study done at multiple signalized intersection at various sites throughout the United Kingdom. According to (Kimber et al., 1986) saturation flow at signalized intersections in the United Kingdom have a significant influence on the condition of the road surface (wet/dry), the proportion of turning traffic, radius of turn, gradient, lane position (nearside, non-nearside), lane width and the number of lanes at the stop line.

Summary

The intersection capacity of a signalized intersection depends upon the prevailing saturation flow rate and the effective green time ratio. The factors affecting the prevailing saturation flow differs from place to place, it is difficult to have a common methodology for estimating capacity. **Table 2** summarizes the base saturation flow estimation procedure practiced by some nations .

The Capacity of Unsignalized Intersection

The unsignalized intersection can be broadly categorized into priority intersection and uncontrolled intersection. In priority intersections, the intersecting roads are given a definite priority over the other, while in the uncontrolled intersection, all the intersecting roads have equal importance. The traffic flow through the unsignalized intersection is governed by signs namely, stop or yield sign. Based on these traffic control measures; the intersection can be a two-way stop-controlled intersection (TWSC) or all-way stop-controlled intersection (AWSC). In the TWSC intersection, the stop sign is provided on the minor street.

The movement of vehicles through the intersection follows the gap acceptance theory. Any traffic stream at an unsignalized intersection moves by accepting gaps in the conflicting streams. The gap availability is inversely related to the number and volume of conflicting streams and the size of the critical gap. In an unsignalized intersection, there exists a hierarchy of movement, **Figure 4** illustrates this priority of movements at a typical four-leg and a typical T-intersection. Vehicles traversing at an unsignalized intersection accepts the gaps in this prioritized manner.

Often, at unsignalized intersections, an approach may be shared by more than one movement. In such cases, the vehicles of a particular movement might be caught behind the vehicles of other movement groups, and its gap of interest never gets fully utilized. That is, even if a gap is available it may not be used because the vehicle, which could have used it had other vehicles ahead of it in the queue. These factors influence the capacity and are sometimes incorporated in capacity calculations.

The traffic behavior at an unsignalized intersection is different across the world in terms of give-way rules, lane discipline, queuing rules, and the enforcement of traffic regulation. It is very difficult to describe such practices with behavioral models like gap-acceptance models. This makes the adoption of a single methodology for capacity estimation infeasible. The analysis of unsignalized intersection can be broadly categorized into two regression (empirical) models and analytical models. The regression models establish a statistical relationship between geometric characteristics and the capacity based on the field data. The analytical models are developed based on the traffic flow theory and actual driver behavior.

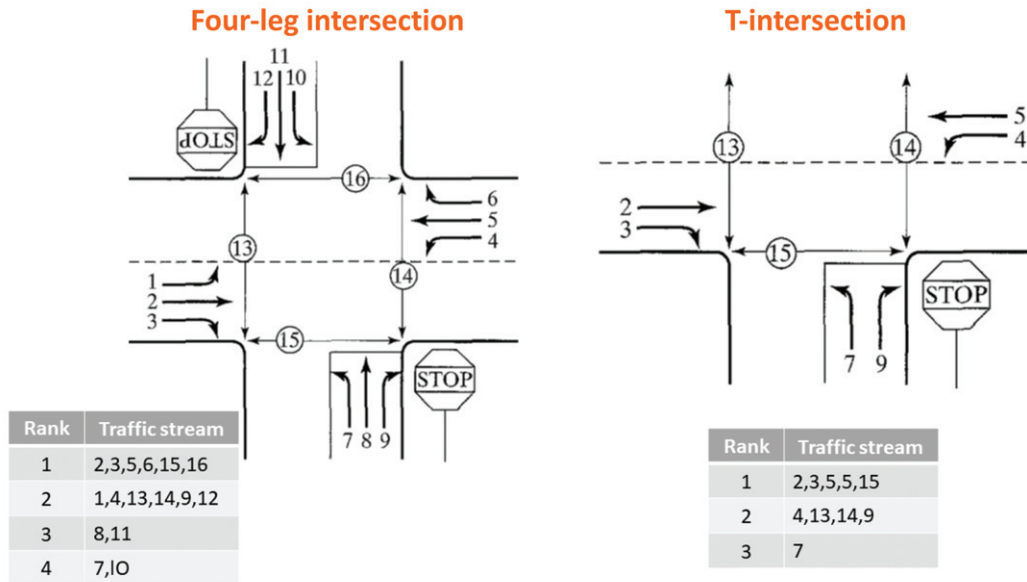


Figure 4 Priority of movements. Source: HCM, (2016)

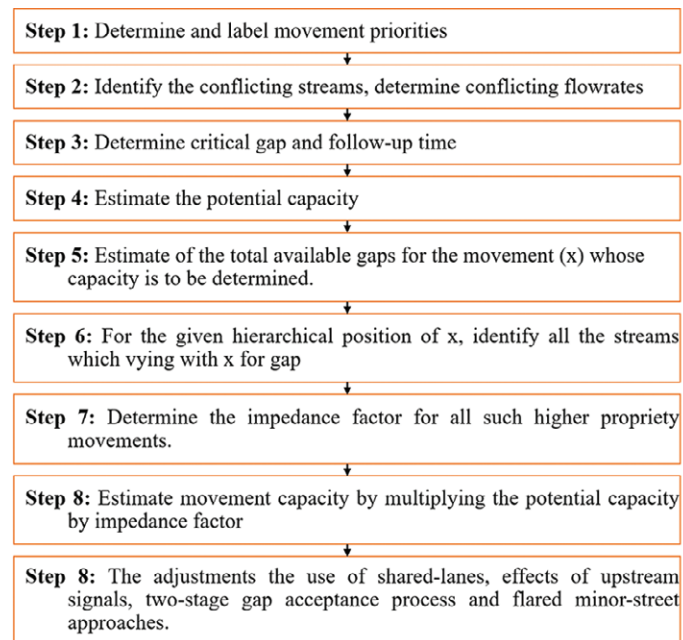


Figure 5 Steps for estimating the unsignalized intersection capacity Source: HCM (2016).

United States of America

Figure 5 illustrates the methodology used for estimating unsignalized intersection capacity in the United States the capacity is estimated based on the conflicting volume, critical gap, and follow-up headway calculated. The effect of the gradient, heavy vehicle, two-stage gap acceptance process, and type of geometry is considered in the calculation of critical gap and follow-up time. The computation of critical gap and follow-up headway is as shown in equation (7) and (8), respectively (HCM, 2016).

$$t_{c,x} = t_{c,base} + t_{c,HV}P_{HV} + t_{c,G}G - t_{3,LT} \quad (7)$$

where

$t_{c,x}$ = critical gap/headway for movement x.

$t_{c,base}$ = base critical headway.

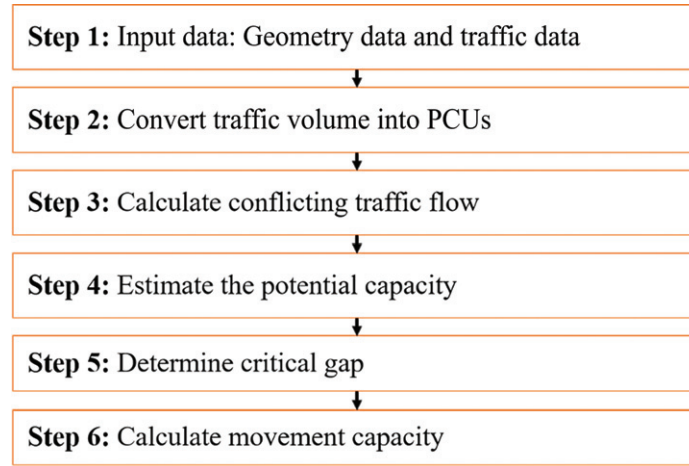


Figure 6 Methodology for estimating the capacity of an unsignalized intersection in India. Source: *Indo-HCM (2017)*

$t_{c,HV}$ = heavy vehicles adjustment factor.

P_{HV} = proportion of heavy vehicles for movement.

$t_{c,G}$ = adjustment factor for the grade.

G = percent grade.

$t_{3,LT}$ = adjustment factor for intersection geometry

$$t_{f,x} = t_{f,bass} + t_{f,HV}P_{HV} \quad (8)$$

where

$t_{c,x}$ = follow-up headway for movement x.

$t_{c,bass}$ = base follow-up headway.

$t_{c,HV}$ = adjustment factor for heavy vehicles.

P_{HV} = proportion of heavy vehicles for movement.

Using the conflicting volume, critical gap, and follow-up time for a given movement, its potential capacity is then estimated using gap acceptance models. According to *HCM (2016)*, the potential capacity of a movement is computed by equation (9). This potential capacity is then adjusted to reflect the impedance from higher-priority movements, sharing of lanes, and other factors.

$$C_{p,x} = v_{c,x} \frac{e^{-v_{c,x}t_{c,x}/3600}}{1 - e^{-v_{c,x}t_{f,x}/3600}} \quad (9)$$

Where,

$C_{p,x}$ = potential capacity of movement x in vehicles per hour.

$v_{c,x}$ = conflicting flow rate for movement x in vehicles per hour.

$t_{c,x}$ = critical headway for minor movement x.

$t_{f,x}$ = follow-up headway for minor movement x.

The movement capacity is then calculated by adjusting the potential capacity. The adjustment factor is obtained by multiplying the probability that each impeding movement will be blocking the subject vehicle. Movement capacity is found by equation (10) (*HCM, 2016*).

$$C_{m,x} = C_{px} \prod_{i,j} P_{vi} P_{pj} \quad (10)$$

Where,

$C_{m,x}$ = Movement capacity, movement x. in vehicles per hour.

C_{px} = potential capacity, movement x in vehicles per hour.

P_{vi} = probability that impeding vehicular movement i is not blocking the subject flow.

P_{pj} = probability that impeding pedestrian movement j is not blocking the subject flow.

India

In India, due to weak enforcement of traffic regulations and lack of understanding of priority rules among road users, no distinction is made in the Indo-HCM between a TWSC and AWSC intersection (*Indo-HCM, 2017*). The manual clearly defines the conditions for a base intersection. If the candidate intersection does not conform to these conditions (non-base intersection), adjustment

Table 3 The capacity of unsignalized intersection summary

Sl. No	Country	Flow rate	Capacity estimation methodology	Critical gap estimation	Follow-up time
1	United States	veh/h	Gap acceptance method	Adjusting the base value	Adjusting the base value
2	Indonesia	pcu/h	Empirical method	N.A.	N.A.
4	India	pcu/h	Gap acceptance method	Occupancy time method	60% of the critical gap

Source: HCM (2016); Indo-HCM (2017); Indonesian Highway Capacity Manual (1993)

factors need to be applied for the deviations from the base conditions. The follow-up time is assumed as 60% of the critical gap. Figure 6 illustrates the steps involved in the estimation of unsignalized intersection capacity in India.

Occupancy Time Method (OTM) is used for the calculation of critical gaps. OTM considers the actual driver behavior on unsignalized intersections. It captures the actual clearing pattern of the conflict area and traffic interactions. The critical gap for different movements is then adjusted for the proportion of heavy vehicles in the conflicting traffic streams. According to the Indo-HCM (2017), the prevailing critical gap for any movement can be obtained using the equation (11).

$$t_{c,x} = t_{c,bass} + f_{LV} \times \ln(P_{LV}) \quad (11)$$

Where,

$t_{c,x}$ = critical gap for vehicle type x.

$t_{c,bass}$ = base critical gap value for corresponding vehicle type executing the same movement.

f_{LV} = adjustment factor for large vehicles.

P_{LV} = proportion of large vehicles in the conflicting traffic stream.

Once the value of the critical gap, follow-up time, and conflicting flow rates is estimated, then the capacity of the movement is calculated from equation (12) (Indo-HCM, 2017).

$$C_x = a \times v_{c,x} \frac{e^{-v_{c,x}(t_{c,x}-b)/3600}}{1 - e^{-v_{c,x}t_{f,x}/3600}} \quad (12)$$

Where,

C_x = capacity of movement 'x' in PCU per hour

$v_{c,x}$ = conflicting flow rate corresponding to movement x in PCU per hour

$t_{c,x}$ = critical gap of standard passenger cars for movement 'x'

$t_{f,x}$ = follow-up time for movement 'x'

'a' and 'b' = adjustment factors based on intersection geometry

Indonesia

Indonesia follows the regression model approach, the prevailing capacity of unsignalized intersections in Indonesia is estimated by applying the corrections to the base capacity. The base capacity for different types of intersections is tabulated in the Indonesian Highway Capacity Manual (1993). The total capacity for all arms of the intersection is calculated from equation (13).

$$C = C_o \times F_w \times F_M \times F_{CS} \times F_{RF} \times F_{LT} \times F_{RT} \times F_N \quad (13)$$

Where,

C_o = Capacity for an intersection type for a predetermined ("base") set of influencing conditions.

F_w = Correction factor for base capacity due to intersection entry width.

F_M = Correction factor for base capacity due to major road median type.

F_{CS} = Correction factor for base capacity due to city size.

F_{RF} = Correction factor for base capacity due to road environment type and side friction level.

F_{LT} = Correction factor for base capacity due to left turning.

F_{RT} = Correction factor for base capacity due to right turning.

F_N = Correction factor for base capacity due to road flow split.

Summary

The capacity of a movement depends on (1) the number of conflicting movements and conflicting volumes, (2) critical gap size, and (3) hierarchical position of the movement among all movements vying for a gap at the intersection. Different techniques are practiced across the globe to estimate the critical gap and capacity of different movements at an unsignalized intersection. The most popular technique is the gap acceptance approach, that assumes a minor street vehicle to come to a halt on its arrival at the intersection and wait for a suitable gap in the conflicting traffic stream to complete its desired maneuver. Table 3 summarizes the capacity estimation method practiced in the United States, India, and Indonesia.

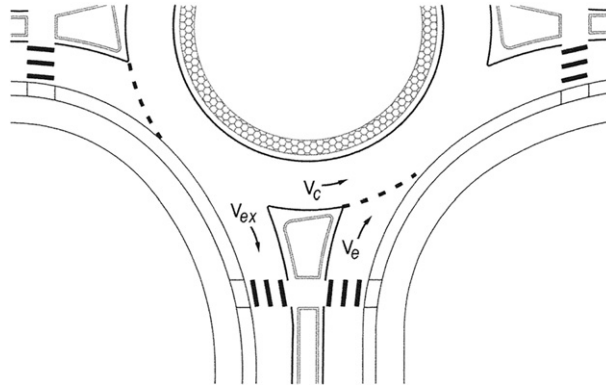


Figure 7 Roundabout approach. Source: HCM (2016)

Table 4 Capacity equation for roundabouts in the United States

Entering lanes	Conflicting circulating lanes	Capacity equation
1	1	$C_{e,pce} = 1130 \text{Exp}[-1.0 \times 10^{-3} v_{c,pce}]$
2	1	$C_{e,pce} = 1130 \text{Exp}[-1.0 \times 10^{-3} v_{c,pce}]$
1	2	$C_{e,pce} = 1130 \text{Exp}[-0.7 \times 10^{-3} v_{c,pce}]$
2	2	$C_{e,R,pce} = 1130 \text{Exp}[-0.7 \times 10^{-3} v_{c,pce}]$
		$C_{e,L,pce} = 1130 \text{Exp}[-0.75 \times 10^{-3} v_{c,pce}]$

Source: HCM (2016)

Capacity of Roundabouts

A roundabout is a class of at-grade intersection where the vehicles approaching the intersection are forced to move around a central island in one direction and move out of the roundabout into their desired direction. The capacity of the roundabout depends upon the circulating flow (V_c) that conflicts with the entry flow (V_e) and the geometric elements of the roundabout (see Figure 7).

The rate at which vehicles can enter the roundabout is influenced by the circulating flow. The size of the gap and the circulating flow is inversely proportional, thus a low-circulating flow results in a higher gap, which helps the vehicles to enter the roundabout without much delay. As the circulating flow increases, the rate at which vehicles entering the roundabout decreases. The entry flow is also affected by the geometric features of the roundabout, such as the entry width, width of the circulatory roadways, and the number of lanes at the entry on the roundabout.

United States of America

According to HCM (2016), the roundabout analysis in the United States is a combination of simple, lane-based regression, and gap-acceptance. The capacity equation for single-lane and multi-lane roundabouts is mentioned in Table 4 the capacity values are adjusted for the presence of heavy vehicles.

Where,

$C_{e,pce}$ = Lane capacity, adjusted for heavy vehicles.

$C_{e,R,pce}$ = Capacity of the right entry lane, adjusted for heavy vehicles.

$C_{e,L,pce}$ = Capacity of the left entry lane, adjusted for heavy vehicles.

$v_{c,pce}$ = Conflicting flow rate.

India

According to Indo-HCM (2017), more than 70% roundabouts in India possess 20–70 m diameter and the average diameter of roundabouts in Indian cities/town is 35 m. Thus, the method mentioned in Indo-HCM confines to roundabouts having this diameter range. Indo-HCM adopted the exponential model from HCM, as mentioned in equation (14–16). The values of the critical gap and follow-up time parameters for roundabouts of different diameters are presented in Indo-HCM. The follow-up time is considered as 0.75 times the critical gap. Using these values, the capacity equations for a varying range of diameters is tabulated in Table 5.

$$\text{Capacity, } C = A \times \text{Exp}(-B \times Q_c) \quad (14)$$

Table 5 Roundabout capacity equations in India

Diameter, D (m)	Critical gap, T_c (s)	Follow-up time, T_f (s)	$A = 3600/T_f$	$B = (T_c - 0.5 \cdot T_f)/3600$	Capacity, $C = A \cdot \text{Exp}(-B \cdot Q_c)$ (pcu/h)
$20 < D \leq 30$	2.01	1.51	2384	0.00035	$C = 2384 \cdot \text{Exp}(-0.00035 \cdot Q_c)$
$30 < D \leq 40$	1.87	1.40	2571	0.00032	$C = 2571 \cdot \text{Exp}(-0.00032 \cdot Q_c)$
$40 < D \leq 50$	1.65	1.24	2903	0.00029	$C = 2903 \cdot \text{Exp}(-0.00029 \cdot Q_c)$
$50 < D \leq 70$	1.61	1.21	2975	0.00028	$C = 2975 \cdot \text{Exp}(-0.00028 \cdot Q_c)$

Source: *Indo-HCM (2017)*

$$A = 3600/T_f \quad (15)$$

$$B = (T_c - 0.5 \times T_f)/3600 \quad (16)$$

Where,

 T_f = Follow-up time in seconds T_c = Critical gap in seconds Q_c = Circulating flow in PCU per hour**Indonesia**

According to the [Indonesian Highway Capacity Manual \(1993\)](#), the capacity under the prevailing condition is obtained by adjusting the base capacity. Base capacity is calculated using the equation (17). Where the input variables are weaving width W , weaving width/entry width ratio WE/W , weaving percentage $W\%$, and weaving width/length-ratio W/L . The actual capacity is then computed by equation (18).

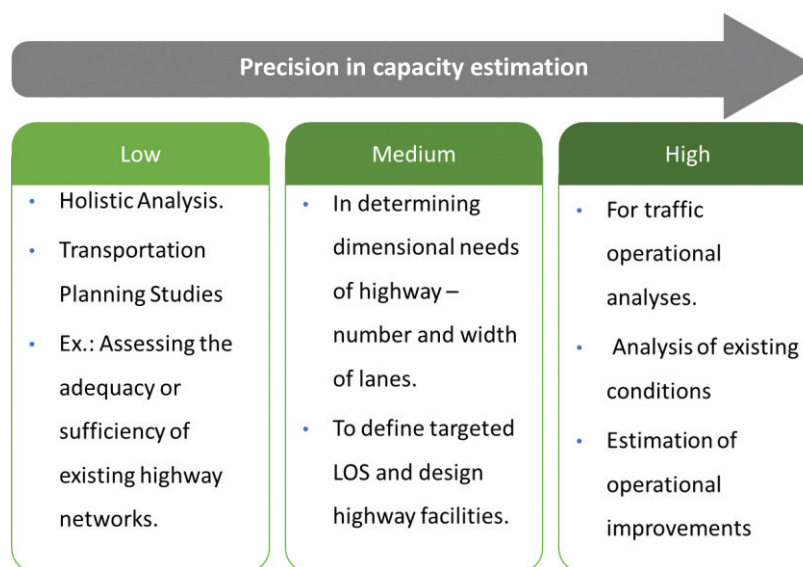
$$C_0 = 135 \times W^{1.3} \times (1 + WE/W)^{1.5} \times (1 - W\%/300)^{0.5} / (1 + W/L)^{1.8} \quad (17)$$

$$C = C_0 \times F_{CS} \times F_{RF} \quad (18)$$

Where,

 C_0 = Base capacity F_{CS} = City size correction factor F_{RF} = Road environment type and side friction correction factor**Summary**

The capacity of a roundabout is influenced by the flow pattern and the geometry of the roundabout. As in the unsignalized intersection, the capacity analysis of roundabouts also based on the gap acceptance theory. In [HCM \(2016\)](#), the capacity equation

**Figure 8** Application of capacity estimation.

is selected based on the number of the lane and in [Indo-HCM \(2017\)](#), the capacity equation is defined by the diameter of the roundabout.

Conclusion

Modeling the capacity of a traffic facility will give an upper hand in the planning, analysis, and operation of the facility. The estimation capacity is influenced by the prevailing roadway, traffic, and control conditions. Thus, the actual capacity observed may vary day to day depending upon several factors like the weather condition, ambient light, and familiarity of drivers, but still, it is important to have a good estimate of capacity values. Most of the capacity analysis models include the determination of capacity under the ideal roadway, traffic, and control conditions, later, it is adjustments for prevailing conditions. According to the intended purpose of future analysis, the precision of capacity estimation varies as shown in [Figure 8](#).

References

- Austroroads, 2020. Guide to Traffic Management Part 3: Transport Study and Analysis Methods. Sydney.
- Tepley, S., Allingham, D.I., Richardson, D.B., et al., 2008. Canadian Capacity Guide for Signalized Intersection third ed.. In: Gough, J.W. (Ed.), The Institute Of Transportation Engineers, Canada.
- HCM, 2016: Highway Capacity Manual. Transportation Research Board, National Research Council, Washington, DC., USA.
- Indo-HCM, 2017. Indian Highway Capacity Manual (Indo-HCM), CSIR-Central Road Research Institute, New Delhi.
- Indonesian Highway Capacity Manual, 1993, Directorate General of Highways. Ministry of Public Works, Indonesia.
- Kimber, R.M., McDonald, M., Hounsell, N.B., 1986. The Prediction Of Saturation Flows For Road Junctions Controlled By Traffic Signals. Crowthorne, Berkshire.

Further Reading

- Roess, R.P., Prassas, E.S., McShane, W.R., 2008. Traffic Engineering. Pearson Education, Inc. third ed. New Jersey
- Akçelik, R., 1981. Traffic Signals: Capacity and Timing Analysis. Australian Road Research Board. Research Report ARR No. 123 (7th reprint: 1998).

Delay

Shinji Tanaka, Yokohama National University, Tokiwadai, Hodogaya-ku Yokohama, Japan

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	258
Definition	258
Theoretical Explanation by Cumulative Curve	258
Assumption to Use Cumulative Curve	259
Methodologies to Measure Delay	260
(1) License Plate Matching	260
(2) Measuring Vehicle Run	260
(3) Probe Vehicle	260
Delay as an Evaluation Index	261
An Example to Measure Delay Using Cumulative Curve	261
Biography	262
See Also	262
References	262

Nomenclature

FFTT free-flow travel time of the target section

ATT actual travel time

$\mu_{\max}$ capacity of the bottleneck

$\lambda(t)$ arrival rate of traffic demand

$A(t)$ cumulative count of arrival from upstream

$A'(t)$ cumulative count of arrival to the bottleneck

$D(t)$ cumulative count of departure from the bottleneck

When vehicles travel on roadway, they experience delay by several reasons. The most typical delay is caused by traffic jam, which occurs when many vehicles try to use a certain roadway section in a short-time period. Delay also occurs at signalized intersections, where vehicles have to stop and wait during red signal. Moreover, delay can be seen not only on roadway but also in any kinds of queueing system, such as, cashier, ticket booth, passport control, and so on. Delay is normally considered as negative output from a queueing system, therefore it is often used to evaluate the performance of the system.

Definition

Delay is defined as increment of actual travel time from free-flow travel time. Here, free-flow travel time is time duration that a vehicle can travel along a certain section of roadway at its desired speed without any influence of other vehicles, control devices such as traffic signals. On the other hand, actual travel time is time duration that a vehicle actually spends for the same roadway section. (May, 1990; Transportation Research Board, 2016)

When we consider delay at a signalized intersection, vehicle's travel time without any acceleration, deceleration or stop in passing the intersection is used as free-flow travel time. Therefore, even though there are no other vehicles and no queue at the stop line of the intersection, a vehicle may experience signal control delay if it stops by red signal.

In some cases in practice, vehicle-stopping time by red signal is used as signal control delay. It becomes smaller than the original definition as it does not include deceleration status of vehicles by signal or congestion. However, they are usually very closely correlated and this method is practical because it is easy to identify the status of vehicles, that is, moving or stopping.

Theoretical Explanation by Cumulative Curve

Suppose a simple roadway section that has a bottleneck whose capacity is $\mu_{\max}$ in the downstream as shown in Figure 1. And, observation point A is set in the upstream enough where queue tail does not reach and another observation point B is set immediately downstream of the bottleneck. Now traffic demand $\lambda(t)$ arrives from the upstream and flows into the bottleneck. While $\lambda(t)$ is less than $\mu_{\max}$, vehicles can travel this section by free flow travel time (FFTT), which is assumed almost constant value. When $\lambda(t)$ exceeds the capacity $\mu_{\max}$, queue is formed at the bottleneck and vehicles have to spend more travel time (actual travel time: ATT) than FFTT. Here, delay is formulated as ATT-FFTT according to the definition.



Figure 1 Simple roadway section.

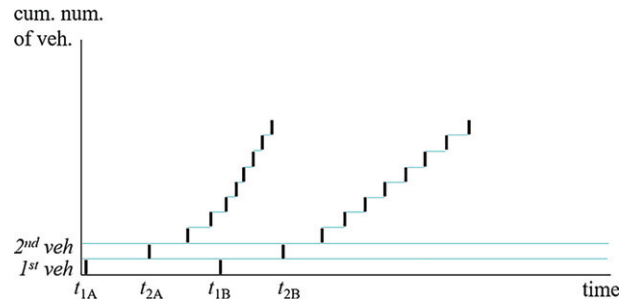


Figure 2 Plot of vehicle passing time.

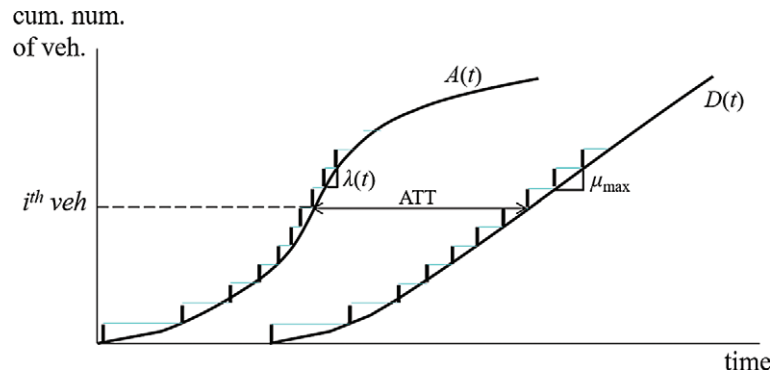


Figure 3 Arrival curve and departure curve.

Now, let's consider vehicle's passing time at the upstream point A and the downstream point B. The recorded data is plotted as shown in **Figure 2**. Here, the horizontal axis is time and the vertical axis is the cumulative number of passing vehicles. When the first vehicle passes, two points are plotted at time t_{1A} and t_{1B} on the bottom line. Next, the second vehicle passes at time t_{2A} and t_{2B} on the line above the previous one.

By repeating this process and connecting the plots, the cumulative number of passing vehicles at the point A and B and their passing time can be obtained as cumulative curves as shown in **Figure 3**. The upstream curve is called as "Arrival Curve," denoted as $A(t)$, whereas the downstream curve is called as "Departure Curve," denoted as $D(t)$. The arrival rate $\lambda(t)$ is shown as the slope of the arrival curve $A(t)$, whereas the capacity of the bottleneck $\mu_{\max}$ is shown as the maximum gradient of the departure curve $D(t)$.

As the i th vehicle is shown as a certain height in this figure, its actual travel time (ATT) can be obtained by the horizontal difference of the two curve $A(t)$ and $D(t)$. Here, ATT is composed of FFIT and delay. As FFIT is common for all the vehicles, let's shift $A(t)$ to the right by this amount. Then, new curves $A'(t)$ and $D(t)$ are obtained as shown in **Figure 4**. Here, the horizontal difference of $A'(t)$ and $D(t)$ is the delay of the i th vehicle. And when there is no delay, these curves become the same line, that is $A'(t) = D(t)$. Then, we can understand that the delay occurs only the period that $A'(t)$ and $D(t)$ are apart. As vehicles that experience delay pass during this time period, the area surrounded by $A'(t)$ and $D(t)$ can be regarded as the total delay experienced by the passing vehicles by this queue. Therefore, we can evaluate the amount of the total delay by calculating the area from this graph. The unit of the total delay is (vehicle-hours), because it is multiplication of time and number of vehicles. (Daganzo, 1997)

Assumption to Use Cumulative Curve

As explained, cumulative curve is useful representation to describe arriving and departing traffic flow. However, it should be noted that cumulative curve assumes the following two conditions.

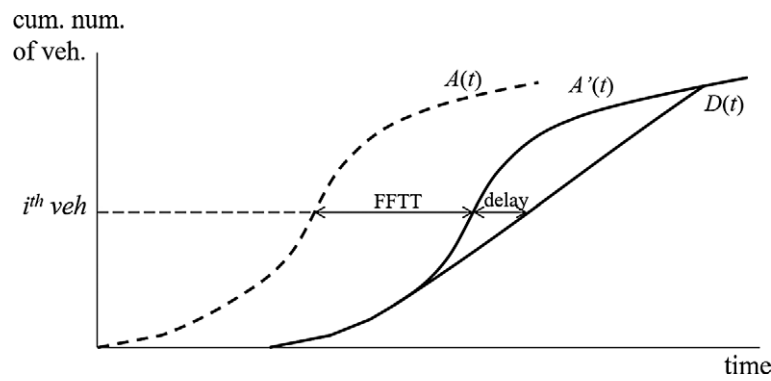


Figure 4 Arrival curve and departure curve.

The first assumption is “First-In First-Out,” which is often abbreviated as “FIFO.” It means that an earlier arriving vehicle to the target section should depart the section earlier. In other words, all the vehicles have to keep their order of travel and there is no overtaking inside the section. By this assumption, it is possible to measure the individual vehicle’s delay by the horizontal difference of $A'(t)$ and $D(t)$ at the same height of the graph.

Another assumption is “point queue” concept, which assumes no physical length of queueing vehicles and their space clearance. In other words, arriving vehicles directly reach the bottleneck position and wait for their turn at this point for departure. It is also denoted as “vertical queue,” because the queue is represented not by horizontal formation of vehicles but by a vertical pile at the bottleneck position. Based on this concept, the shifted arriving curve $A'(t)$ means the vehicles’ arrival time at the bottleneck, and we can simply consider the delay as the separated region of $A'(t)$ and $D(t)$.

Methodologies to Measure Delay

To measure delay, travel time of a certain roadway section have to be measured. Therefore, it is basically the same as to measure travel time, that can be done various kinds of methods as follows.

(1) License Plate Matching

Setting observation stations at the both ends of the roadway section, surveyors record license plate number of passing vehicles and their passing time. It is also possible by taking video image at the observation stations. After the observation, the recorded data of the both stations are matched and then the travel time of the identified vehicles can be obtained by taking the difference of their passing time at the both stations. Although this method is easy to conduct and efficient to collect a lot of data samples, manual license plate matching is generally time consuming work. However, automated technology using image processing for license plate matching is available and makes it becomes a feasible real time solution.

(2) Measuring Vehicle Run

This method prepares a vehicle to measure travel time by running the target roadway section actually. To measure representative travel time, the measuring vehicles should travel in average speed of the traffic flow, that is, not too fast nor too slow compared with other vehicles. This method is simple and clear, however, it is difficult to collect many data samples unless many measuring vehicles and drivers are prepared.

(3) Probe Vehicle

“Probe vehicle” is an idea to utilize general running vehicles on the road network as moving sensors. Thanks to the mobile network communication technology development, it becomes possible to collect information of a lot of individual vehicles to the central server. By accumulating those information, road network condition such as speed, travel time etc. can be collected. It is also called as “floating car data (FCD).” This method requires relatively complicated and large-scale system including information transfer by communication technology. However, once the system is established, large amount and continuous information can be obtained automatically, which is very advantageous to monitor long-term transition of the performance. Another issue is privacy aspect that means drivers’ consents have to be obtained before collecting probe data.

In any methods, sometimes it may be difficult to obtain free flow travel time if delay occurs most of the time period in the target section or the section contains multiple signalized intersections. In such cases, free-flow travel time can be estimated by dividing the section length by the assumed free flow speed (e.g., speed limit).

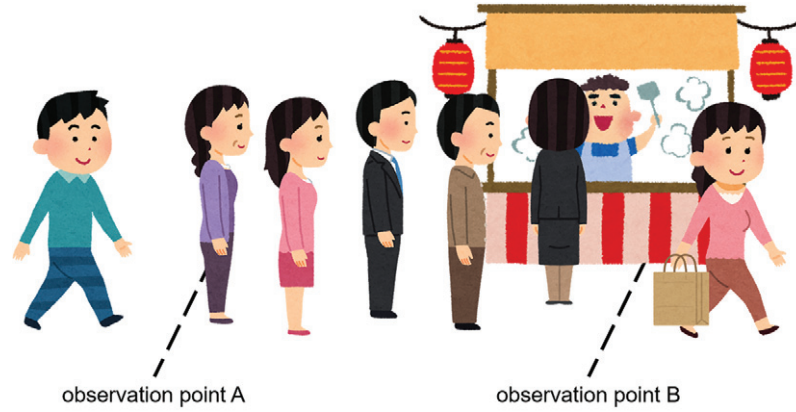


Figure 5 A queue to buy lunchbox at street vendor.

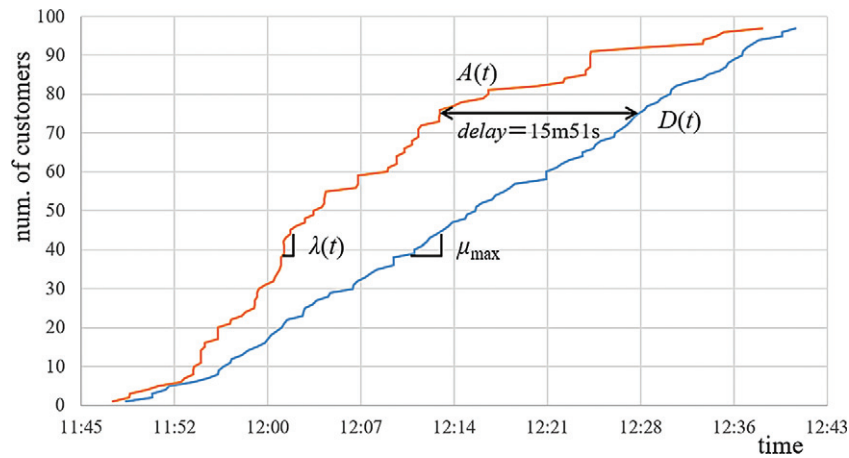


Figure 6 Arrival and departure curve during lunchtime.

Delay as an Evaluation Index

As delay is a simple and intuitive idea that can be easily measured by recording time, it is widely used to describe the influence caused by some disturbance, such as roadwork, traffic signal and so on. In that sense, delay is a good index to evaluate queueing systems. Furthermore, total delay considers volume of users or travelers, therefore it is more appropriate to quantify the negative influence of queueing systems. Generally, smaller total delay means better performance of the system.

For example, if you would like to evaluate a certain roadwork project to alleviate traffic congestion, total delay before and after the project should be measured and the reduction of them (in other words, total travel time saving) is regarded as the effect of the project. This idea is applicable to evaluate the performance of other queueing systems that may include delay, such as highway sections, customer services, and so on.

An Example to Measure Delay Using Cumulative Curve

Here, let's see an example to measure delay by an actual queue using cumulative curves. This is not a queue of vehicles but a queue of people. A popular street vendor cooks and sells take-out lunchboxes during the lunchtime and people make a long queue to buy them everyday as shown in **Figure 5**. Now two observation points are set in the upstream and the downstream of the queue. The upstream point A is at the tail of the queue, recording the time when a new customer arrives at the tail. The downstream point B is at the vending counter, recording the time when a customer receives his/her lunchbox and leaves there.

The result is shown in **Figure 6**. Here, the cooking time which is necessary for every customer even without a queue (regarded as FFT) is already subtracted. The orange line shows the arrival curve $A'(t)$ and the blue line shows the departure curve $D(t)$. From this figure, we can understand that there were approximately 100 customers on that day. There was almost no delay until 11:53, because the orange curve and the blue curve coincided each other. After that, the two cumulative curves were separate and came close again at 12:40. That means the queue was formed during this period and the delay of a person can be measured as horizontal difference of the

two curves. The slope of the arriving curve is the arrival rate $\lambda(t)$, which varied time to time, whereas the slope of the departure curve almost stayed constant. Measuring this slope, we can evaluate the maximum service rate $\mu_{\max}$ of this vendor as approximately 2 customers/min.

The vertical difference of the two curves means the number of waiting people in the queue. The largest difference occurred at 12:13, when 32 people were waiting their turns. On the other hand, the largest delay (= the longest waiting time) also occurred to a customer arriving at this timing, that is 15 min 51 sec. Finally, the area that is surrounded by the arriving curve $A'(t)$ and the departure curve $D(t)$ is the total delay by the queue of that day and its amount was 13 hours 11 min 42 sec (man-hours). Thus, the influence of a queue can be quantified and evaluated from a simple observation, which is comparable to other cases.

Biography



Shinji TANAKA is an Associate Professor of Yokohama National University, JAPAN. He received a degree of Doctor of Engineering from The University of Tokyo in 2007. His research interests are traffic engineering, traffic operation, and intelligent transport systems.

See Also

Queue Length

References

- Daganzo, C.F., 1997. Fundamentals of transportation and traffic operations. Pergamon-Elsevier, Oxford, U.K.
- May, A.D., 1990. Traffic flow fundamentals. Prentice Hall, Englewood Cliffs, NJ.
- Transportation Research Board, 2016. Highway Capacity Manual. Transportation Research Board, Washington, D.C..

Queue Length

Shinji Tanaka, Yokohama National University, Tokiwadai, Hodogaya-ku, Yokohama, Japan

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	263
Definition	263
Theoretical Explanation of Queue Length in Number	263
Theoretical Explanation of Queue Length in Physical Distance	264
Methodologies to Measure Queue Length	265
Queue Length as an Evaluation Index	266
Queue Length at Traffic Signal	266
An Example to Measure Queue Length Using Cumulative Curve	267
Biography	267
See Also	267
References	267

Nomenclature

$A'(t)$ cumulative count of arrival to the bottleneck

$D(t)$ cumulative count of departure from the bottleneck

μ capacity of the bottleneck

λ arrival rate of traffic demand

k_1, k_2, k_3 traffic density

u shockwave speed

R duration of red phase of traffic signal

G duration of green phase of traffic signal

C cycle time of traffic signal

Q_{max} maximum queue length

A queue is formed at a bottleneck when demand exceeds capacity or arrival rate is higher than service rate. Queue length is the number of vehicles/people in a queue, which shows how large a queue is. It is a very easy and intuitive way to show the size of a queue, therefore it is widely used to describe the influence of a queue or congestion. In some cases, especially when a moving queue is discussed, same number of queueing vehicles/people may occupy different spatial distance. In such a situation, queue length in physical distance would also be considered.

Definition

Queue length is defined as accumulated number of vehicles/people in the upstream of a bottleneck that joined but not yet left a queue. Therefore, to obtain the queue length, a queue itself have to be identified. Queue is a part of traffic flow whose speed is lower than the upstream and the downstream flow. The tail of the queue is the point where arriving vehicle/person join the queue, whereas the head of the queue is the point where queueing vehicle/person leave the queue. (May, 1990; Transportation Research Board, 2016)

Normally, the point queue concept is employed as shown in Chapter 10320, that is, there is no physical length of vehicles as well as their space clearance. Then, these vehicles are accumulated vertically at the point of the bottleneck and the queue length can be measured as the number of the vehicles there.

In some context, queue length in physical distance has to be discussed literally. In this case, physical queue concept is employed and queue length is defined that a queue occupies in the space from its head to its tail in the unit of distance, such as meter, kilometer, etc. The difference of point queue and physical queue is shown in [Figure 1](#). As the traffic states in queues are not always the same, the same queue length might sometimes have different influence on the performance such as travel time, etc.

Theoretical Explanation of Queue Length in Number

Similar to queueing delay in Chapter 10320, let's use cumulative curves by employing the point queue concept to illustrate queue length. We can draw arrival curve $A'(t)$ and departure curve $D(t)$ as shown in [Figure 2](#). Please refer to Chapter 10320 for the detail to draw them. If we set the current time as τ , $A'(\tau)$ means the cumulative number of vehicles that already arrived at the bottleneck and D

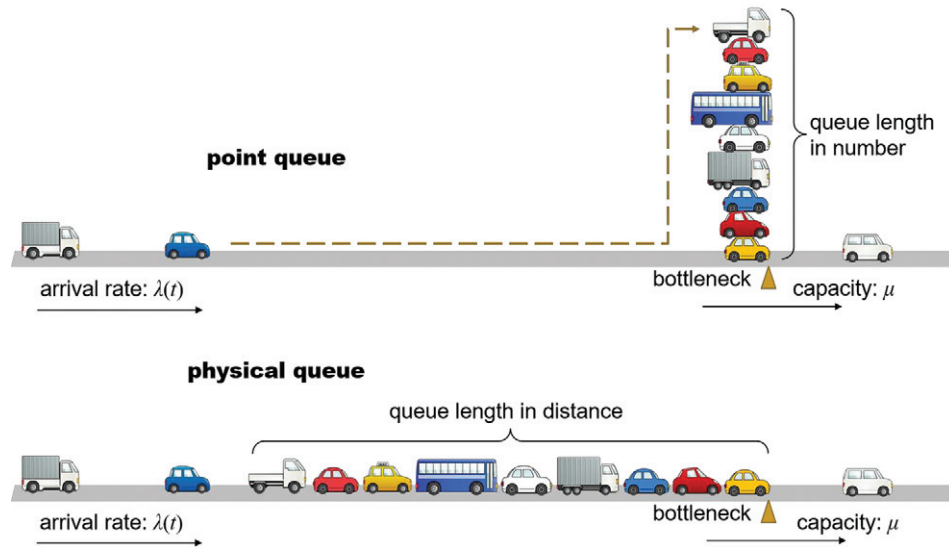


Figure 1 Point queue and physical queue.

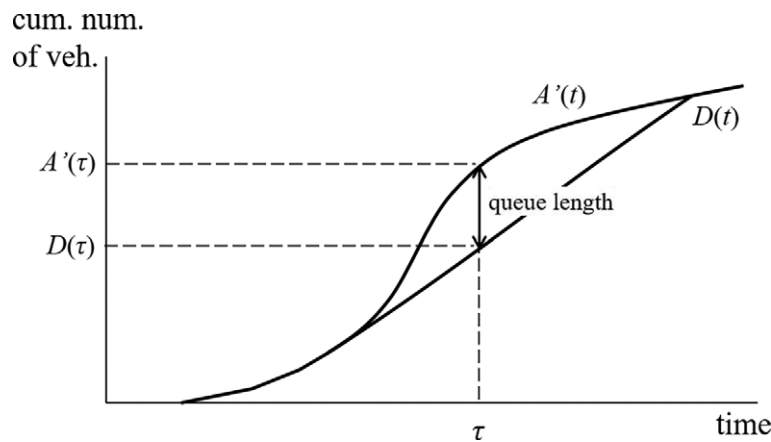


Figure 2 Queue length by cumulative curves.

(τ) means the cumulative number of vehicles that already departed from the bottleneck by that time. Then, the vertical difference at this time $A'(\tau) - D(\tau)$ shows the number of vehicles remaining in the queue, that is, queue length. (Daganzo, 1997)

Theoretical Explanation of Queue Length in Physical Distance

If we consider queue length in physical distance, we need to assume not point queue but physical queue. And traffic flow characteristics in the queue such as flow rate, density, and speed should be considered. Let's see how to obtain the physical queue length using shockwave analysis. Figure 3 shows simplified vehicle trajectories that form a physical queue at a bottleneck. Each line shows a trajectory of a vehicle and its slope shows the speed of the vehicle.

Vehicles arrive from the upstream in the flow rate of λ , which is the traffic demand. However, as the bottleneck capacity is μ which is lower than λ , the flow rate in the downstream of the bottleneck is regulated to μ . And, the difference of the demand and the capacity is accumulated as a queue in the upstream of the bottleneck, whose flow rate is also μ . The vehicle speed in the queue is relatively slow compared with the upstream and the downstream section. In this figure, flow rate λ or μ can be obtained by the inverse value of time headway, which is the horizontal interval of the vehicle trajectories. On the other hand, the vertical interval of the trajectories shows space headway, which is the inverse value of traffic density.

Here, let's classify this traffic flow into three region based on their states, that is, region I, II and III as shown in different colors in Figure 4. Region I is arriving traffic from the upstream in free flow condition, and its flow rate and density are λ and k_1 . Region II is queue at the bottleneck in congested flow condition, and its flow rate and density are μ and k_2 . Region III is departing traffic from the bottleneck in free flow condition, and its flow rate and density are μ and k_3 . Then, the queue length is obtained by measuring the

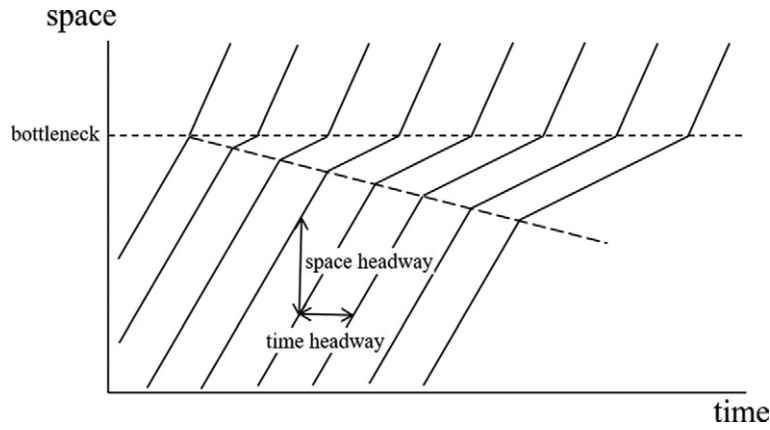


Figure 3 Vehicle trajectories at a bottleneck.

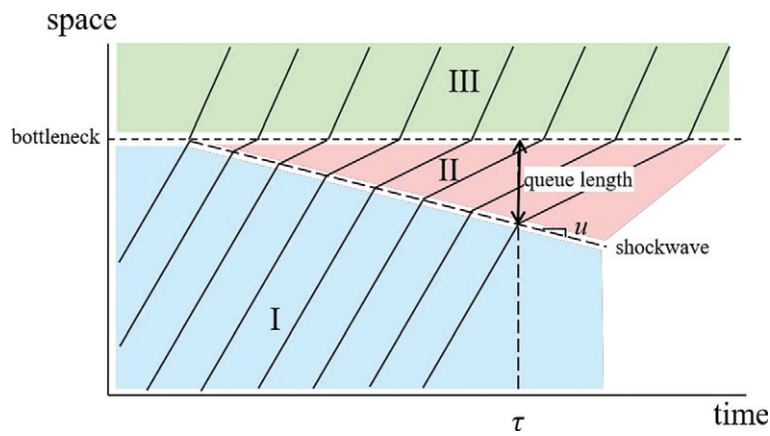


Figure 4 Different states of traffic flow.

vertical distance of the region II in the upstream of the bottleneck. As the head of the queue stays at the same position, the queue length is determined if the boundary movement between region I and II is obtained.

Figure 5 is enlarged trajectories between region I and II. The slope of this boundary is denoted as u ($u < 0$). Suppose we consider a rectangular region as shown in the figure that has a unit time width t , then the height of this rectangle is $-ut$. Based on the flow conservation law, the volume of vehicles entering this unit region should exit without any loss. Then, the following equation should be established.

$$\lambda \times t + k_1 \times (-ut) = \mu \times t + k_2 \times (-ut)$$

$$\therefore u = \frac{\lambda - \mu}{k_1 - k_2}$$

Here, λt and μt are entering and exiting traffic volume from the bottom and the top edge, because λ and μ are flow rate (vehicles per unit time). Similarly, $k_1(-ut)$ and $k_2(-ut)$ are entering and exiting traffic volume from the left and the right side edge, because k_1 and k_2 are density (vehicles per unit distance). And, u shows the moving speed of the boundary, which is called shockwave. Using this shockwave speed, the position of the queue tail can be calculated and the queue length can be obtained.

Methodologies to Measure Queue Length

If the entire queue is visible from the observer, it is easy to measure the queue length in number. However, if it is not visible from one position, queue length has to be estimated by the difference of the cumulative arrival and the cumulative departure of the queue as explained in the previous section. To conduct this method, at least one traveler has to be identified both in arrival and departure to set the starting point of the cumulative curve. Another possible option is to use the traffic density inside a queue. If it is known and the queue length in physical distance is measured by the methodology as shown in the next paragraph, queue length in number is also calculated.

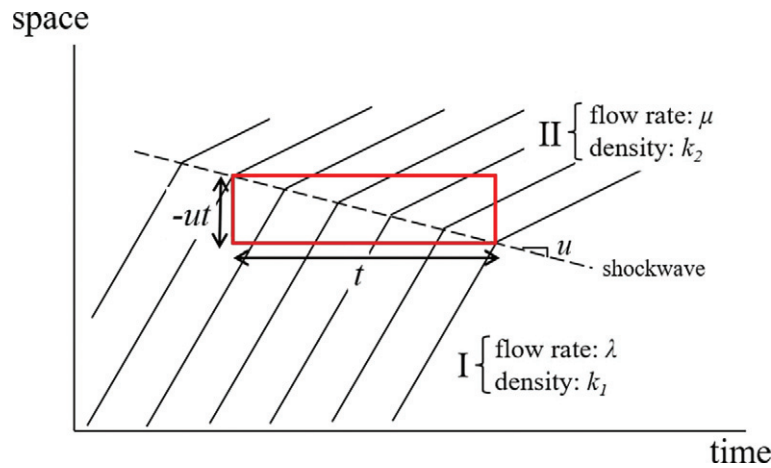


Figure 5 Enlarged vehicle trajectories and shockwave.

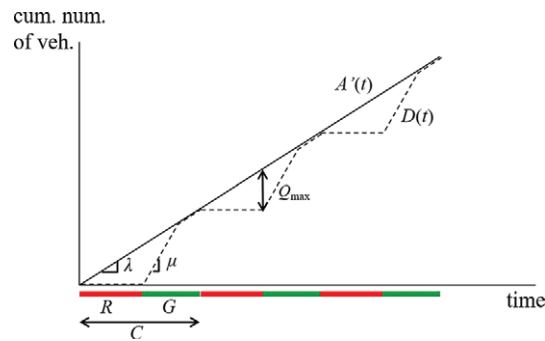


Figure 6 Arrival and departure curve at traffic signal.

To measure queue length in physical distance, the positions of the head and the tail of the queue have to be identified and recorded. Normally, the head of the queue is the bottleneck and its position is stable in many cases. On the other hand, the tail of the queue often moves a lot, therefore the observer has to follow it to record its position. It may sometimes become difficult if the tail moves quickly.

Queue Length as an Evaluation Index

Queue length is a very simple and easy way to describe how large a queue is, therefore it is commonly used in traffic information, newspapers, and so on. When some kinds of investments to improve the system were implemented, reduction of queue length would be a good index to show the effect.

In some cases, queue length in physical distance is also important. For example, when we need to design some components of infrastructure such as left/right-turn bays in intersections, storage length has to be properly designed to keep the queue inside it. Otherwise, the queue may block other traffic movements and deteriorate the system performance when its length becomes longer than that.

Queue Length at Traffic Signal

A typical case to consider queue length is commonly observed at traffic signals. Let's consider a simple traffic signal that repeats red and green phase alternately. Here, yellow phase is ignored for simplicity. The duration of red phase, green phase, and cycle time are denoted by R , G and $C(=R + G)$, respectively. Regarding the traffic flow, constant vehicle arrival rate λ and saturation flow rate μ are assumed. Suppose this intersection is undersaturated, that is, all the queued vehicles at the end of the red can leave the intersection during the next green and there is no queue when the signal turns to red. **Figure 6** shows arrival curve $A'(t)$ and departure curve $D(t)$ in this situation. $A'(t)$ has a constant slope λ , whereas the slope of $D(t)$ takes three values; zero (horizontal line) during red phase, μ during green phase until the queue disappears and λ during green phase after the queue disappears.

As explained in the earlier section, the queue length is shown as the vertical difference of $A'(t)$ and $D(t)$ in this figure. Therefore, the queue length starts to increase when red signal begins. The queue length become the maximum value (Q_{max}) at the end of red.

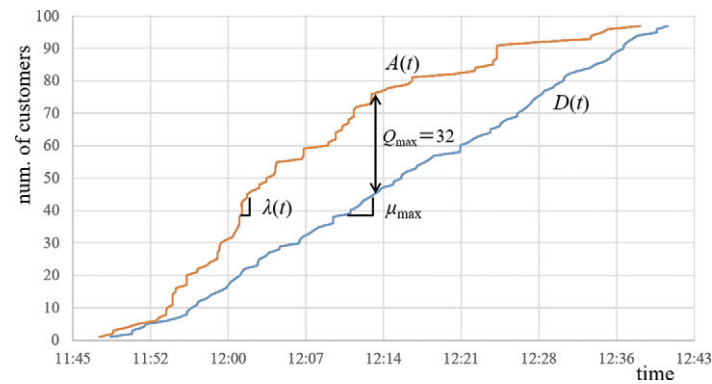


Figure 7 Arrival and departure curve during lunchtime.

When the signal turns to green, the queue length starts to decrease and finally it becomes zero when $D(t)$ catches up $A'(t)$. It remains zero until the end of green.

An Example to Measure Queue Length Using Cumulative Curve

Let's use the same example shown in Chapter 10320 to demonstrate how to measure queue length by an actual queue using cumulative curves. **Figure 7** is the cumulative curves obtained from the observation of the queue in front of the street vendor. The orange line shows the arrival curve $A'(t)$ and the blue line shows the departure curve $D(t)$. Then, the vertical difference of the two curves means the number of waiting people in the queue. The largest difference, that is the longest queue length, occurred at 12:13, when 32 people were waiting their turns.

Biography



Shinji TANAKA is an Associate Professor of Yokohama National University, JAPAN. He received a degree of Doctor of Engineering from The University of Tokyo in 2007. His research interests are traffic engineering, traffic operation, and intelligent transport systems.

See Also

Delay

References

- Daganzo, C.F., 1997. Fundamentals of transportation and traffic operations. Pergamon-Elsevier, Oxford, U.K.:
- May, A.D., 1990. Traffic flow fundamentals. Prentice Hall, Englewood Cliffs, NJ.
- Transportation Research Board, 2016. Highway Capacity Manual. Transportation Research Board, Washington, D.C..

Local Area Traffic Management

Michael A.P. Taylor, School of Natural and Built Environments, University of South Australia, Adelaide, SA, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	268
Typical Traffic Problems in Local Areas	268
Good Practice	269
LATM Design Solutions	269
Through Traffic	273
Critical Values	275
Conclusions	276
See Also	276
References	277
Further Reading	277

Introduction

Local area traffic management (LATM) describes a range of traffic engineering measures applied to the planning and management of traffic movement in local areas (precincts or neighbourhoods). It is generally applied as remedial treatments for local street networks, usually but not always in residential precincts, which are experiencing adverse traffic impacts such as excessive volumes of traffic and high travel speeds. LATM often involves the use of physical devices, streetscape design treatments, and other measures (e.g., regulatory measures including lower speeds limits and turning restrictions), aimed at reducing through traffic movements and controlling vehicle speeds. More broadly it may be seen as a process used to plan and control the traffic usage of a local street network to achieve improvements in the residential environment of the local area, meeting goals for that end determined by relevant stakeholders.

For the purpose of distinguishing between LATM and other aspects of traffic management, a “local (traffic) area” may be taken as an area containing only local streets and collector roads, and usually bounded by arterial roads or other roads serving a significant road transportation function, or other physical barriers such as waterways, railways, reserves, or impassable terrain.

LATM implementations are usually undertaken by local government, generally and most successfully with the collaboration and involvement of community groups representing residents and local business. Neighbourhood groups concerned with pedestrian safety and traffic impacts are often the initiators of LATM plans.

Typical Traffic Problems in Local Areas

The main issues encountered in LATM studies are typically speeding, through traffic, parking, corner cutting and traffic congestion, and generally decreasing in that order of importance.

Speed management is perhaps thus an overarching concern, generally approached by the installation of midblock speed control devices (e.g., humps), but there are mixed views on the effectiveness and acceptability of different devices. For instance, speed humps are seen as effective but controversial, largely because of adverse public opinion. The effectiveness of other devices such as raised platforms, road narrowing, and kerb extensions is also widely debated. An important issue for speed control devices is their impact on buses and emergency vehicles. Some bus operators may refuse to allow buses to use road links with such devices. Alternatives such as speed cushions that permit ready bus operation may then be considered. “Blister island” treatments are also seen as an effective speed reduction device in some jurisdictions.

Table 1 lists a number of common LATM treatments and devices. The table shows two different orderings of the listed treatments, based on recent research on local government LATM practice in Australasia (Damen, 2007; Damen and Ralston, 2015). The first column lists the treatments in order of effectiveness, as perceived by municipal traffic engineers. The second column lists the treatments in terms of actual usage.

Table 1 indicates that roundabouts, installed at internal junctions in the local area street network, were ranked as the most effective device and were widely used. The second most effective device was a full road closure, but it was seldom employed.

The LATM devices expressly designed for speed limitation along a street section showed some interesting results. These devices were: flat-topped hump, road (Watts profile) hump, slow point, and speed cushion. They were ranked 5th, 5th, 13th, and 14th, respectively, in terms of effectiveness, but there was clear evidence of a growing reluctance by municipalities to use them. Less than half of the surveyed authorities rarely or never used flat-topped humps, Watts profile humps or slow points. The speed cushion is perhaps a more specialized installations only suited to specific cases (e.g., cater for bus movements along a street). The Watts profile hump was found to be the least popular in the community, with three times the level of complaints received about the next most unpopular device, the flat-topped hump. Most other devices appeared to attract little or no complaint.

Table 1 LATM devices employed by local government in Australasia, listed in terms of effectiveness and usage

<i>Effectiveness of treatment (most effective to least effective)</i>	<i>Usage of treatment (most common to least common)</i>
Roundabout	Roundabout
Full road closure	One way, stop and give way signs
Median treatment	Lower speed limit
One way, stop and give way signs	Kerbside lane narrowing/kerb extension
Flat top road hump	Median treatment
Blister island	Blister island
Modified T junction	Bicycle lane
Road hump (Watts profile)	Flat top road hump
Kerbside lane narrowing/kerb extension	Pedestrian (zebra) crossing
Bicycle lane	Road hump (Watts profile)
Partial road closure	Turn ban
Pedestrian (zebra) crossing	Slow point (angled or straight)
Slow point (angled or straight)	Tactile surface treatment
Speed cushion	Modified T junction
Lower speed limit	Speed cushion
Raised intersection pavement	Raised intersection pavement
Shared zone	Perimeter threshold treatment
Turn ban	Full road closure
Tactile surface treatment	Driveway link
Driveway link	Partial road closure
Bus only link	Shared zone
Perimeter threshold treatment	Bus only link

Source: Data from Damen and Ralston (2015)

Good Practice

Good practice in LATM applications may be taken as the formulation of LATM policy specifically aimed at:

1. maintaining amenity in local, residential access streets by deterring inappropriate levels of through traffic from using those roads and controlling the speed of traffic that does use them;
2. recognizing the varied uses of the road space and optimizing the safety and viability of using walking, cycling or public transport as alternatives to the private car; and
3. determining the circumstances in which traffic calming works are warranted and prioritizing them equitably to address the aims 1 and 2 above.

The need for LATM investigations should then be triggered by the observation that one or more of the following factors exceeds acceptable limits:

- speeding (85th percentile speed greater than the relevant speed limits of, for example, 40, 50, or 60 km/h);
- volume (annual average daily traffic, considered in relation to surrounding streets);
- through traffic (which is to be considered in the context of the total traffic volumes); and
- safety (examined through consideration of crash statistics, street geometry and roadside environment, and identified risks).

Data collection and assembly for the previous factors provide basic information on the likely extent of traffic problems facing a given local area.

Barriers to implementation may exist. In particular strong community support is vital. Sometimes local residents may be opposing to a LATM scheme, often relating to the suggested use of specific types of LATM devices rather than an overall concept plan. Strong local community support is essential for the success of a LATM scheme, so that community participation in the planning and design of LATM and the compatibility of a proposed scheme with local aspirations and objectives is essential. Financial costs may be an issue for some municipalities. Historically, some traffic engineers objected to LATM plans when concerned about traffic mobility rather than overall street design.

LATM Design Solutions

LATM is applied to an area through the use of physical devices, streetscaping treatments, and other measure (including regulations to influence vehicle operation on local streets), in order to create safer and more liveable local streets. The basic concept for good LATM design can be summarized by the following rules:

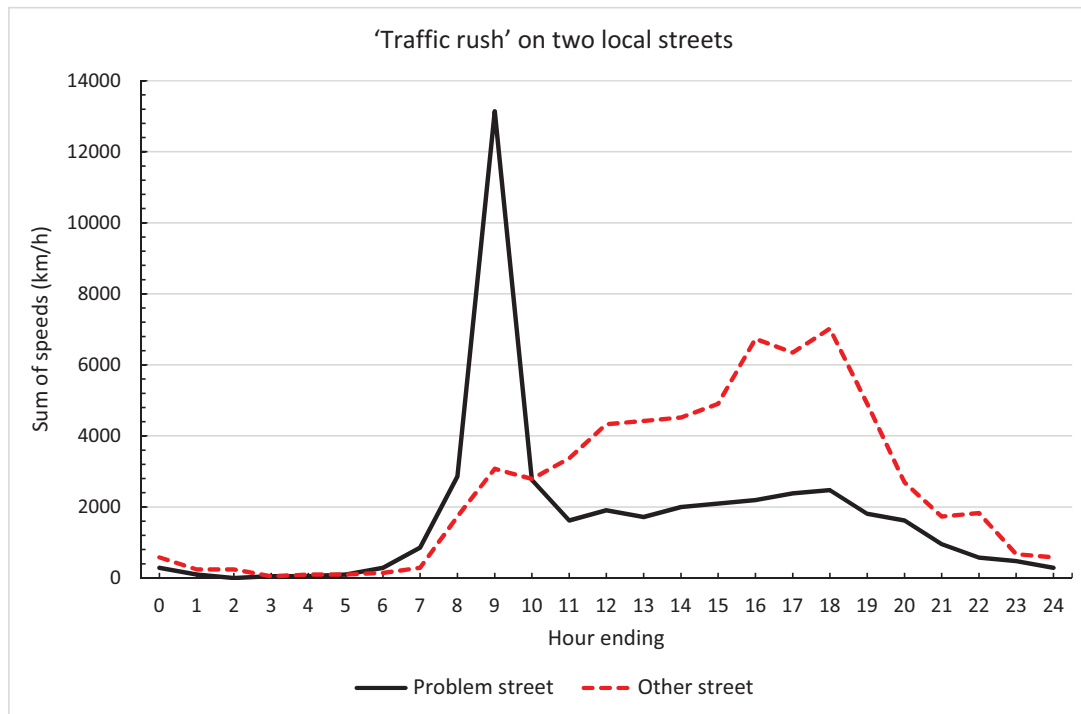


Figure 1 Traffic rush on two streets, one with resident complaints about traffic problems, the other with no complaints.

- good local street design should aim for *poor long distance visibility* (for visual enclosure) but *good short distance visibility*, especially between knee and head height;
- devices should be forgiving on those drivers who misjudge them, for example, solid immovable objects should be set back well behind deflective kerbs;
- construction materials, road geometry, and texture should indicate the local nature of the street and its multiple use;
- good night time visibility is critical, especially for strangers to the area; and
- *the more local the street, the greater the importance of landscape architects and urban designers*. They, rather than the traffic engineer, should have the major say in the design of devices in streets, which carry less than 2000 veh/day.

One useful parameter indicating the potential susceptibility of a street to local traffic problems is its “traffic rush.” A traffic rush metric (rush per hour) may be defined as the sum of the speeds of all vehicles passing in an hour, in either one direction of interest, or in both directions as a measure of total disturbance from passing traffic (Woolley et al., 2002). Fig. 1 shows contrasting traffic rush patterns for two different streets in one suburb (following the introduction of the 40 km/h local area speed limit in that area). One of the streets had complaints voiced by local residents—it attracted rush hour traffic from a nearby crowded collector road. The second street had no specific complaints raised about traffic. The contrast shown in Fig. 1 for the problem street between morning rush hour and the rest of the day is significant. While the total rush on the other street (the total rush on a street is shown by the area under the traffic rush curve) is greater than that on the problem street, the rush hour contrast to 9:00 a.m. is far less, by a factor of at least six, that is, the contrast is far less pronounced on the transverse street.

The concept of speed-based design, in which the geometric features of the street and its road pavement are adjusted to provide a driving environment that encourages lower traveling speeds by vehicles, is now seen as the primary design tool for LATM. This approach is less concerned with the use of specific speed control devices, but rather seeks to ensure an overall street speed environment, which is compatible with community expectations. It involves flattening the speed profile of the street, that is the variation of free speeds along the street length. The speed management objective is then given as a target speed profile. The speed profile is the graph of instantaneous vehicle speed along the street. Fig. 2 Fig. 2 shows an example of the “before and after” speed profiles along a street from which humps were removed and a roundabout then installed. The two speed profiles represent the actual traversal of a length of local street by a vehicle, which runs in to the street from a main road and travels some 800 m along the street, to exit it at a stop sign. In the “before” case the street contains seven Watts profile humps, approximately 100 m apart. In the “after” case, the humps have been removed and a single roundabout installed at an intersection about 300 m from the entry point. The data presented in Fig. 2 are based on actual observations of traffic movement on the street in question (Zito and Taylor, 1996).

The “before” speed profile in Fig. 2 shows a succession of decelerations and accelerations as the vehicle negotiates each hump in turn, with the driver attempting to regain a cruising speed of about 45 km/h between the humps. This leads to a confusing speed

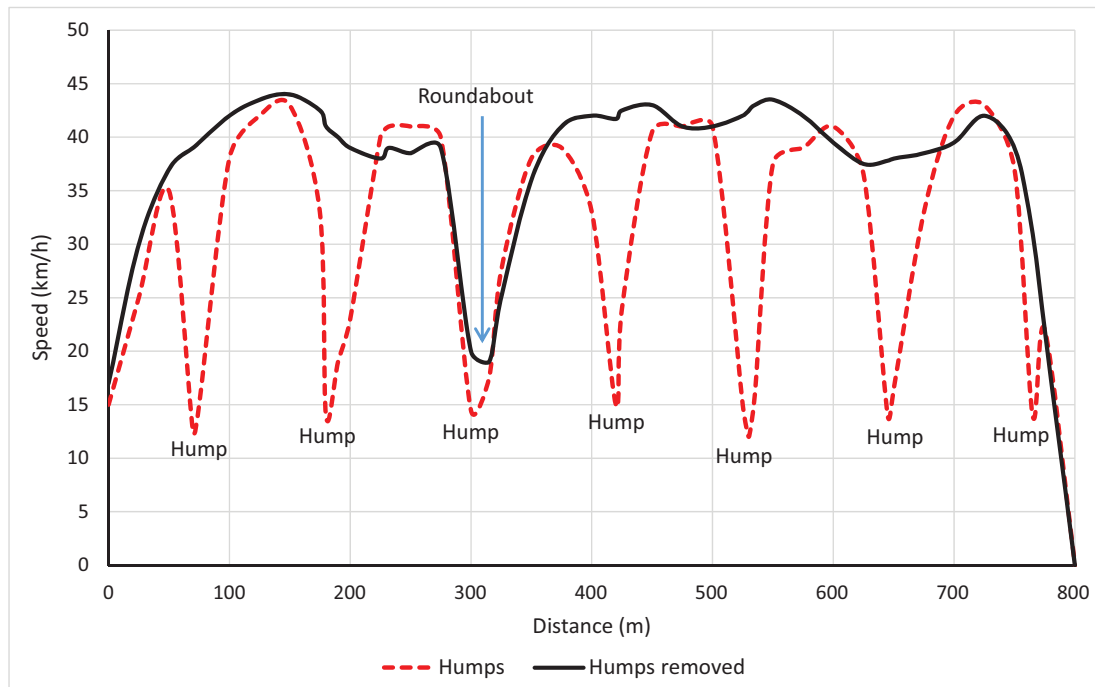


Figure 2 Comparison of speed profiles before and after the removal of road humps.

environment as the instantaneous speed of the vehicle is changing continually. In the “after” case, the speed profile is more stable, with a single deceleration and then acceleration as the vehicle negotiates the roundabout.

Speed management should therefore aim to lower the speed profile based on a realistic estimation of driver behavior. It requires consideration of the sequence of speed control devices along the street, not each device in isolation. Each device can be assumed to have a “zone of influence,” the distances either side of the device where the traversing vehicle’s speed is affected by the device (Taylor and Rutherford, 1986). These zones of influence can be seen in Fig. 3. Figure 3 indicates the ‘zone of influence’ and the speed differential for a speed control device. Note that the design speed variable for the speed profile is the 85th speed, that is, the speed exceeded by 15% of the vehicles traversing the street section where the device is located. Outside the zone of influence, drivers will try to maintain their free speeds, whatever the speed at which they have to traverse the speed control device itself. The aim of speed-based LATM design is thus to ensure that the new street speed is below a target speed and thus minimize fluctuations in speed along the street. This may be accomplished by considering the “speed differential,” which is the difference between free speed and the traversal speed of the device at that point—with all other treatments in place. A maximum value for the speed differential of 20 km/h is recommended.

The speed differential can then be used to select, design, and locate the sequence of devices along a street. High-speed differentials imply large speed changes within the zones of influence and thus along the length of the street, with consequent risks of increased noise and hazard.

The speed-based design process is summarized by the following steps:

1. measure or estimate current free speeds on the street section;
2. specify target street speed, and thus the required speed differential; and
3. design the scheme to meet the target speed with speed differential below the maximum allowed value at each device in turn.

There is a wide range of traffic control devices available for use in LATM. Table 2 indicates the application of different devices and identifies them according to the device categories defined by Ogden (1996):

1. regulatory devices,
2. network modifications,
3. intersection devices,
4. vertical displacement devices,
5. horizontal displacement devices, and
6. gateways.

Fig. 4 illustrates some of these devices.

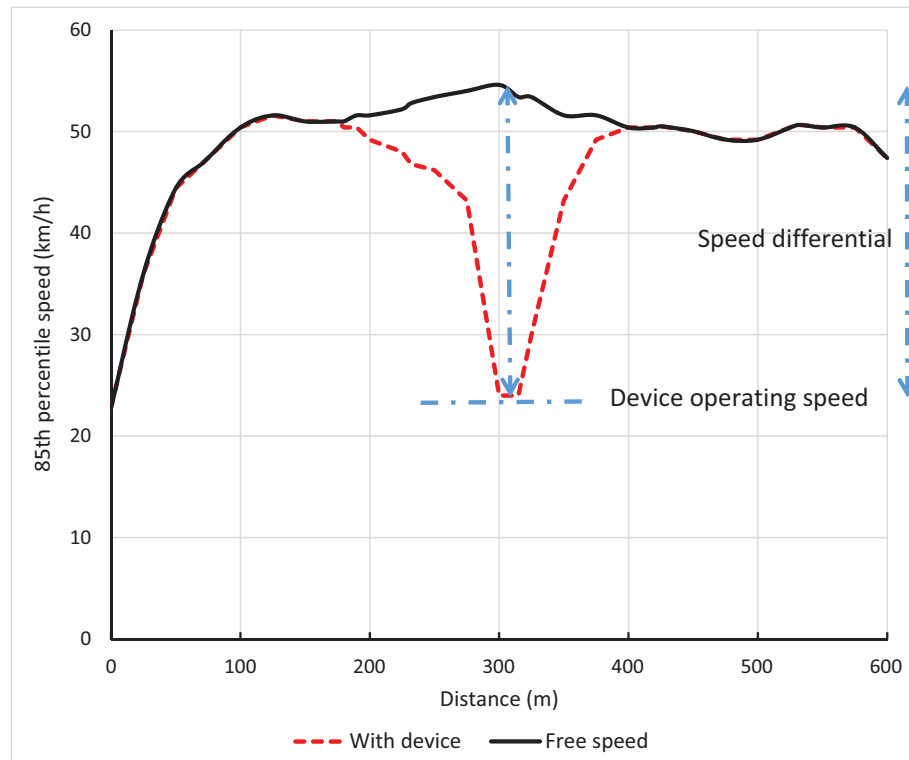


Figure 3 The “zone of influence” and the speed differential value for a speed control device such as a road hump.

Table 2 Description and use of LATM control devices

Device type category	Device	Application		
		Reduce speeds	Reduce traffic volume	Reduce crash risk
Regulatory devices	Lower speed limits	✓✓✓	?	✓✓✓
	Prohibitions on heavy vehicles (trucks)		✓✓✓	✓✓✓
Network modifications	Street closures—full	✓✓✓	✓✓✓	✓✓✓
	Street closures—partial		✓✓✓	✓✓✓
Intersection devices	One way streets		✓✓✓	✓✓✓
	Roundabouts	✓✓✓	✓✓✓	✓✓✓
Vertical displacement devices	Turn bans		✓✓✓	✓✓✓
	Road (Watts profile) humps	✓✓✓	✓✓✓	✓✓✓
	Flat top humps	✓✓✓	✓✓✓	✓✓✓
	Raised pavements	✓✓✓	✓✓✓	✓✓✓
	Road cushions	✓✓✓	✓✓✓	✓✓✓
Horizontal displacement devices	Lane narrowings and kerb extensions	✓✓✓		
	Slow points	✓✓✓	✓✓✓	
	Blister islands	✓✓✓	✓✓✓	
	Driveway links	✓✓✓	✓✓✓	
Gateways	Midblock medians	✓✓✓		✓✓✓
	Threshold treatments	✓✓✓	✓✓✓	✓✓✓
	Left turn in/left turn out ^a		✓✓✓	✓✓✓

^aFor traffic driving on the lefthand side of the road, as in the United Kingdom, Australia, Japan, India, and New Zealand (amongst others)

Observations of traffic behavior on local streets suggest that the more devices that are installed along a street, the milder the effects of each device (in terms of speed differential) need to be. At the limit, the least speed differentials will be obtained on streets with continuous speed control devices. One continuous speed control device is the use (and enforcement) of lower speed limits.



(A) Lower speed limit, area-wide



(E) Flat top hump



(B) Street closure



(F) Slow point



(C) Roundabout



(G) Blister island



(D) Road (Watts profile) hump



(H) Left in/left out gateway

Figure 4 Some typical LATM devices: (A) lower speed limit, area-wide, (B) street closure, (C) roundabout, (D) road (Watts profile) hump, (E) flat top hump, (F) slow point, (G) blister island, and (H) left in/left out gateway.

Through Traffic

Besides excessive speed, through traffic intrusion is also a recognized traffic issue for traffic precincts. Local traffic, being traffic with either or both an origin and a destination in the area, would be considered to be part of the area and its activity and more likely to behave in ways compatible with the local street environment in terms of speed, and would therefore be seen as an acceptable part of the local life.

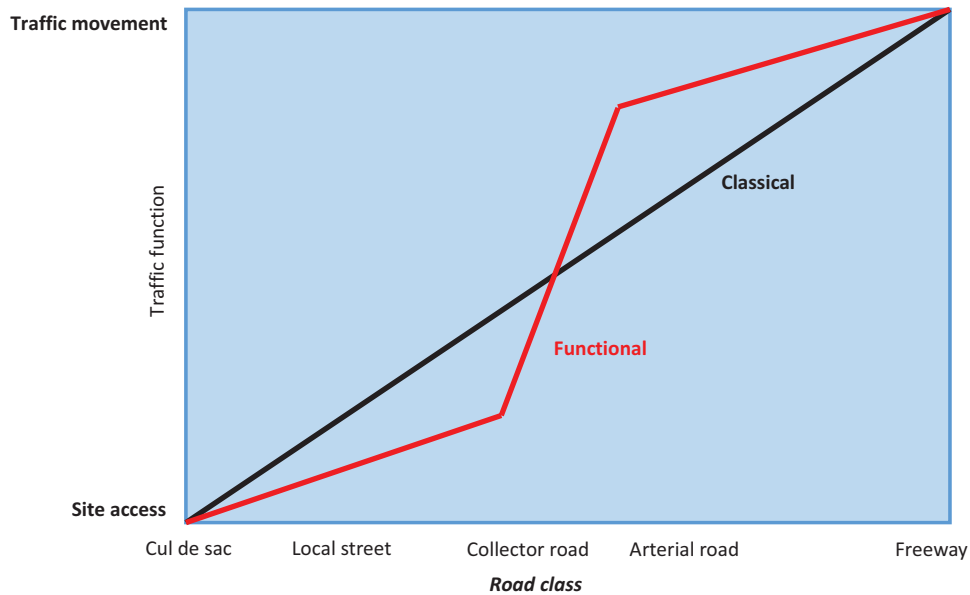


Figure 5 Road hierarchy classification, including the classical and functional conceptual models of the split of traffic movement and site access functions for the road classes.

The issue is to consider what actually constitutes through traffic, either in a local area or on a particular street section. For example, the objectives of an LATM study can be to:

- reduce vehicle speeds in the area;
- reduce through traffic volume on local streets and encourage it to use more appropriate roads, and
- effectively manage on-street parking in areas in which this is a problem.

A common definition of through traffic is “traffic with neither an origin nor a destination within the local area. Depending on the definition of the local area, this may not be traffic which diverts between two arterial roads.” This definition includes the possibility that traffic originating in one part of a local area may in fact be seen as through traffic in another part of it. However, the definition is not particularly useful in assessing the extent of this issue. What it does provide is the indication of a longstanding principle in LATM that local streets should only be available for the terminal ends of journeys and for local circulation and access, and not be regarded as a part of the general urban road transport network. This principle is embodied in the concept of the hierarchy of roads in which roads are classed in terms of their traffic-carrying or access functions, with local streets serving a predominant access function. A typical urban road hierarchy would include the following road classes: urban freeways, urban arterial roads, subarterial roads, primary collector roads, secondary collector roads, and local streets. The freeway is solely for traffic movement. The ultimate local street, which can provide no more than access, is a cul-de-sac. All other road classes involve some proportioning of traffic movement and site access function. **Fig. 5** indicates the conceptual differences in traffic function (movement or access) between the road classes. Note that this figure presents the both the historical “classical” view of a continuous change in function proportion between the road classes, and the more recent “functional” classification which suggests a break between access and movement, for the case of collector roads. The functional classification model is based on the observation that network design including roads which attempt to provide an even balance between movement and access generally fail to provide satisfactory service for either function on those roads.

The hierarchy of roads may be used to suggest a definition of through traffic that can be applied across a study area including the level of the individual street (**Brindle, 1989**). The idea is that of an “ideal” journey structure for a vehicle moving from its origin, as directly as possible, via increasingly “important” roads (in traffic-carrying terms), until it reaches the arterial road network on which the principal part of the journey takes place. If the destination is in a local street, the vehicle then moves back down the road hierarchy, only entering the local access road network when necessary. Thus, in the ideal journey, traffic moves upwards then downwards in the road hierarchy, so that it minimizes the length of travel in the local access street network. It follows that through traffic is traffic which, on some part of its journey, has dropped down then goes back upwards in the hierarchy. Traffic which, for whatever reason, has dropped down sooner into the “lower-order” network so that length of travel on those roads is greater than the minimum is also through traffic on that part of its journey, even if the route choice is logical to the driver. **Fig. 6** summarizes this definition of through traffic.

This figure identifies an “ideal” journey through a network, which minimizes the use of the “lower-order” roads in the hierarchy, and a hypothetical “rat running” trip in which a through vehicle penetrates a residential area. Through traffic is thus any traffic movement involving the use of a lower class road section in mid journey. American studies of traffic precinct design have suggested

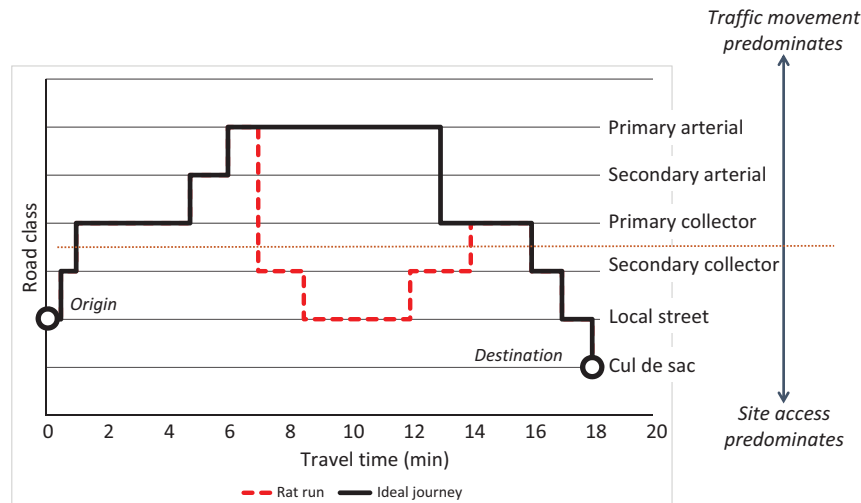


Figure 6 Definition of “through traffic” in terms of the usage of different classes of roads in the road hierarchy, including an “ideal” journey and a “rat running” journey. Source: Adapted with permission from Brindle (1989)

that the maximum driving time required to reach a collector road from any address on a local residential street should be one minute (Ewing, 1999), as a means of encouraging drivers to use the “higher class” roads for the bulk of their journeys.

Critical Values

While there is no established “best” practice for setting warrants or priorities for LATM implementations, there are a variety of mixed systems of warrants, priority rankings based on aggregate points scores on a set of traffic factors, and accepted “threshold” values. Points scores for priority rankings may be based on both objective (observed data) values and subjective values, using expert opinion or the results of community surveys. These are likely to be particular to specific communities at given points in time, so there can be no firmly established values for general use. However, the reported values may be used in an indicative fashion when considering the likely severity of perceived or actual traffic problems in local areas.

Typical factors for inclusion in a rating system include:

- character and function of the street,
- general speed limit.
- traffic volumes and speeds,
- level of “nonlocal” traffic,
- level of commercial traffic,
- presence or absence of major traffic generators or non-residential users,
- availability of crash data or safety review inspection assessments,
- availability of lighting, and
- degree of local support.

Note that some factors may interact, so that an adjustment in one factor may change the level of acceptability of another. Traffic speed and volume are a case in point, where the combined effect may be the key driver to the development of a traffic problem and to its resolution. For example, a local street carrying more than 2000 veh/day with an 85th percentile speed of 55 km/h is likely to have a traffic problem, whereas if the 85th percentile speed were 45 km/h the likelihood of there being a recognized problem would be significantly reduced.

A particular concern for some LATM studies is the possible level of through traffic intrusion in the area, largely concerning “rat running” behavior through a local street grid, or access to specific traffic generators located on the periphery or nearby the study area and for which access (including parking in the study area) may be sought through the local area.

One “rule of thumb” is that if the peak hour traffic volume exceeds 15% of daily total traffic then there is a strong likelihood of “rat running” behavior in the street, implying the presence of through traffic. This percentage could be taken as a trigger value. It can also be related to the “traffic rush” metric (Fig. 1).

On the basis of the functional model for road classification under the hierarchy of roads and using recent research results, the set of threshold values in Table 3 can be suggested.

The threshold values in Table 3 may then be taken as triggers to alert the traffic engineer to the possibility of speed or volume-related traffic problems on a particular street or road section.

Table 3 Recommended threshold values as a guide to identifying traffic problems in a local area

Threshold value	Sub arterial	Primary collector	Secondary collector	Local street
1. Minimum traffic volume: daily traffic	>9000 veh/day	>6000 veh/day	>3000 veh/day	>1000 veh/day
peak hour traffic	>900 veh/h	>600 veh/h	>300 veh/h	>100 veh/h
2. 85th percentile speed	5 km/h above speed limit	5 km/h above speed limit	5 km/h above speed limit	Above speed limit
3. Anticipated through traffic (from origin-destination data, if available)	65%	60%	40%	25%
4. Peak hour factor	>15%	>15%	>15%	>15% or >15 veh/h ^a

^aFor very low volume streets, the percentage value is less useful and a component traffic volume is a better guide, hence a combined threshold value for local streets is more appropriate.

Conclusions

LATM involves the planning and management of road space usage in a local traffic area, often to modify streets and street networks which were originally designed in ways that are now no longer considered appropriate to the needs of residents and users of the local area. LATM can be seen as a tool of traffic calming at the local area (precinct) level. It involves the use of physical devices, streetscaping designs and treatments, and other measures (including regulations and other non-physical measures) to influence driver behavior, with the aim of creating safer and more pleasant streets in local areas.

LATM is system-based and area-wide. It considers neighbourhood traffic-related problems and their proposed solutions in the context of the local area or a group of streets within it, rather than only at isolated locations. In addition, it requires that physical traffic management measures be seen as a sequence of interrelated devices rather than individual treatments. Traffic engineers need to be alert to the potential problems of isolated speed management devices.

The primary target of LATM is to change driver behavior, both directly by physical influence on vehicle operation and indirectly by influencing the driver's perceptions of what is appropriate behavior in that street. The objective is to reduce traffic volumes and speeds in residential areas to improve amenity, increase liveability, and improve safety and access for pedestrians and cyclists.

The need for LATM usually arises from:

- an intent to reduce traffic-related problems,
- orderly traffic planning and management,
- a need to modify "transport" behavior,
- a desire to improve the community space,
- a desire to improve environmental, economic and social outcomes,
- traffic interventions associated with new development or the implementation of pedestrian and bicycle plans and other local policies.

Traffic-related problems concern mainly:

- improved traffic safety and security, leading to programs for speed moderation and other changes in driver behavior; and
- protection or improvement of local amenity focusing on appropriate allocation, design, and use of street space, as well as driver behavior.

The following planning and design principles, which help to build and ensure attractive streets and healthy communities, should therefore be applied in LATM implementations:

- build the street for everyone, for streets have multiple uses that must be recognized and balanced;
- create many linkages so that pedestrians and cyclists have multiple routes to destinations;
- provide comfortable footpaths, and streets that are easily crossed;
- keep motor vehicle traffic dispersed and moving at low speeds;
- tree planting and landscape design help to foster the attractiveness of a street;
- ensure the presence of public space, for the street is a main component of the public realm in which people and interact and build community spirit; and
- build to a proper scale and size, focusing on people not motor vehicles.

See Also

Pedestrian Safety, General; Road Diets; Shared Space; Speed Limits on Urban Streets; Speed-Reducing Measures; Arterial Road Management; Street Design for Active Travel; Road Traffic Regulation: Road Pricing and Environmental Quality; Methods for Designing Public Transport Networks; Road Safety; Parking

References

- Brindle, R.E., 1989. SOD the distributor! Multidisciplinary transactions. IEAust 13 (2), 99–112.
- Damen, P., 2007. The whole box'n'dice of traffic calming experiences in local government. Proceedings of IPWEA Conference, Cairns, Institute of Public Works Engineering Australasia, Sydney. Available from: https://www.level5design.com.au/uploads/8/9/3/3/8933720/damen_pj_ipwea_cairns_2007_v2.pdf.
- Damen, P., Ralston, K., 2015. The latest on local area traffic management practice in Australia and New Zealand – an update and comparison for local government. Proceedings of IPWEA Conference, Rotorua. Institute of Public Works Engineering Australasia, Sydney. Available from: <https://www.ipwea.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=50062593-552e-4281-8b3b-38b223e75f28&forceDialog=0>.
- Ewing, R., 1999. Residential street design: Do the British and Australians know something Americans do not? Trans. Res. Rec. 1455, 42–49.
- Ogden, K.W., 1996. Safer Roads: A Guide to Road Safety Engineering. Avebury, Aldershot.
- Taylor, M.A.P., Rutherford, L.M., 1986. Speed profiles at slow points on residential streets. Proc. Aust. Road Res. Board 13 (9), 65–77.
- Woolley, J.E., Dyson, C.B., Taylor, M.A.P., Zito, R., Stazic, B., 2002. Impacts of lower speed limits in South Australia. IATSS Res. 26 (2), 6–17.
- Zito, R., Taylor, M.A.P., 1996. Speed profiles and vehicle fuel consumption at LATM devices. Proceedings of the Combined 18th ARRB Transport Research Conference and Transit New Zealand Land Transport Symposium, Christchurch, New Zealand 18 (7), 391–406.

Further Reading

- Austroroads, 2016. Guide to Traffic Management Part 8: Local Area Traffic Management. Austroroads, Sydney.
- Brindle, R.E., 1997. Traffic calming in Australia—more than neighborhood traffic management. ITE J. 67 (7), 26–33.
- Damen, P., 2007. The whole box'n'dice of traffic calming experiences in local government. Proceedings of IPWEA Conference, Cairns. Institute of Public Works Engineering Australasia, Sydney. Available from: https://www.level5design.com.au/uploads/8/9/3/3/8933720/damen_pj_ipwea_cairns_2007_v2.pdf.
- Damen, P., Ralston, K., 2015. The latest on local area traffic management practice in Australia and New Zealand – an update and comparison for local government. Proceedings of IPWEA Conference, Rotorua. Institute of Public Works Engineering Australasia, Sydney. Available from: <https://www.ipwea.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=50062593-552e-4281-8b3b-38b223e75f28&forceDialog=0>.
- Elvik, R., 2001. Area-wide traffic calming schemes: a meta-analysis of safety effects. Accid. Anal. nd Prevent. 33, 327–336.
- Herrstedt, L., 1992. Traffic calming design – a speed management method: Danish experiences on environmentally adapted through roads. Accid. Anal. Prevent. 24 (1), 3–16.
- ITE/FHWA, 1999. Warrants, project selection procedures and public involvement. In: Traffic Calming: State of the Practice. Institute of Transportation Engineers, Washington DC.
- Selberg, K., 1996. Road and traffic environment. Lands. Urban Plan. 35, 153–172.
- Transportation Association of Canada, 2018. Primer on traffic calming. Transportation Association of Canada, Ottawa Available from: www.tac-atc.ca
- UK Department for Transport, 2007. Manual for Streets. Thomas Telford, London.
- VTPI, 2017. Traffic calming – roadway design to reduce traffic speeds and volumes. Victoria Transport Policy Institute, Victoria Transport Policy Institute, Victoria BC. Available from: <https://www.vtpi.org/tm/tm4.htm>

On-Street Parking

Jun Chen, Guang Yang, School of Transportation, Southeast University, Nanjing, China

© 2021 Elsevier Ltd. All rights reserved.

Influencing Mechanism of On-Street Parking on Dynamic Traffic Flow	278
Supply Allocation	278
Parking Spaces Scale and Location Layout	278
Spatial Design Method of Parking Spaces	278
Parking Supply and Demand Regulation Strategies	280
Intelligent and Informational Parking Management	280
See Also	283
References	284
Further Reading	284

Influencing Mechanism of On-Street Parking on Dynamic Traffic Flow

The concern about the interference of on-street parking on dynamic traffic flow has been longstanding and pervasive. Relevant study involves parking search and traffic flow theory. Before the 1980s, scholars studied the traffic operation and safety problems caused by on-street parking. Before entering the 21st century, the influencing mechanism of on-street parking on dynamic traffic is still in its infancy. Scholars were still focusing on basic traffic flow theory. The development of traffic flow theory and models has gone through a process from free flow theory to non-free flow theory, then to mixed urban traffic flow theory. In this stage, a lot of discussions have been made on which forms (parallel, vertical and oblique) of the parking lot should be setting in the road. For many years, traffic engineers believed that parking along major roads should be prohibited at least during rush hours, while angle parking can only be set on low-speed roads. Since 2000, the research on this issue has become multi-objective. More research emerged with the focuses on improving road safety and operation efficiency, reducing delays, considering stakeholders, setting on-street parking belts, achieve the comprehensive optimization of overall social benefits (Zhenyu and Jun, 2012). Automatic parking system is another new research focus. Algorithms has been designed for vehicles to park in existing parking spaces automatically, while the design methods of on-street parking spaces have not been updated for automatic driving.

Supply Allocation

Considering that unreasonable supply allocation of on-street parking lots (parking spaces scale, location layout, and the spatial design of parking spaces) will occupy urban road space (motor lane, non-motor lane, and pedestrian passage), and then affect urban traffic flow operation, it is necessary to study the supply allocation methods of on-street parking lots.

Parking Spaces Scale and Location Layout

Supply allocation (parking spaces scale and location layout) of on-street parking lots is closely related to road conditions, so the influencing factors of supply allocation mainly include road conditions and road network gradation, road network capacity, traffic conditions and road network service level, off-street parking facilities, and traffic management level.

The parking supply analysis of on-street parking mainly includes the determination of capacity and location. Among them, the determination methods of parking capacity can be summarized into three types: (1) the determination method based on user travel cost; (2) the threshold determination method based on multiple constraints; (3) the determination method based on stochastic gap theory and queuing theory. The location determination of on-street parking is mainly based on the establishment of multi-level and multi-objective optimization model, which comprehensively considers the impact of on-street parking on dynamic traffic, detour distance, walking distance, and spaces supply.

However, considering that on-street parking spaces around different types of buildings (schools, residential areas, offices, shopping malls, etc.) often show different parking demand in different time periods, there are few research results about the supply allocation of on-street parking spaces in different time periods. Therefore, in the future, different parking supply allocation strategies (parking spaces scale control, parking time limit, parking guidance, etc.) should be researched and implemented to avoid illegal parking in the peak hour and waste of resources in the off-peak hour according to different on-street parking demand.

Spatial Design Method of Parking Spaces

Each country has proposed spatial design methods of on-street parking spaces according to its own actual situation. China's "Setting Specification for Parking Spaces in Urban Road (The Industry Standard of Public Security, 2009)" gives general requirements, setting

conditions, road sections and areas where parking spaces should not be set, and design methods of parking spaces on urban roads, the design layouts of Chinese on-street parking spaces are shown in Fig. 1 (The Industry Standard of Public Security, 2009). Rochford District Council's "Parking Standards Design and Good Practice Supplementary Planning Document" also gives a variety of parking styles including Square Parking (or 90° Square Parking), Angled Parking, Parallel or "End to End" Parking, as shown in Fig. 2.

The following conditions should be considered in the design of on-street parking spaces: (1) road conditions, involve road width and cross-sectional form; (2) traffic volume conditions, involve motor vehicle flow, non-motor vehicle flow and pedestrian flow in the section; (3) rationality of the planning location, involves coordination between parking belts on road and road environmental conditions such as signalized intersections, road openings, and pedestrian crossings.

In the future, with increasingly serious environmental problems, the design of on-street parking spaces needs to be considered as an important aspect. In addition, with the rise and promotion of automatic driving technology, the research of on-street parking spaces design method will certainly face changes. Since computer-controlled vehicles could perform more stably and accurately than human drivers in parking process. And it can be assumed that the interference of on-street parking on traffic flow can be reduced to a certain extent through the communication and cooperation between vehicles.

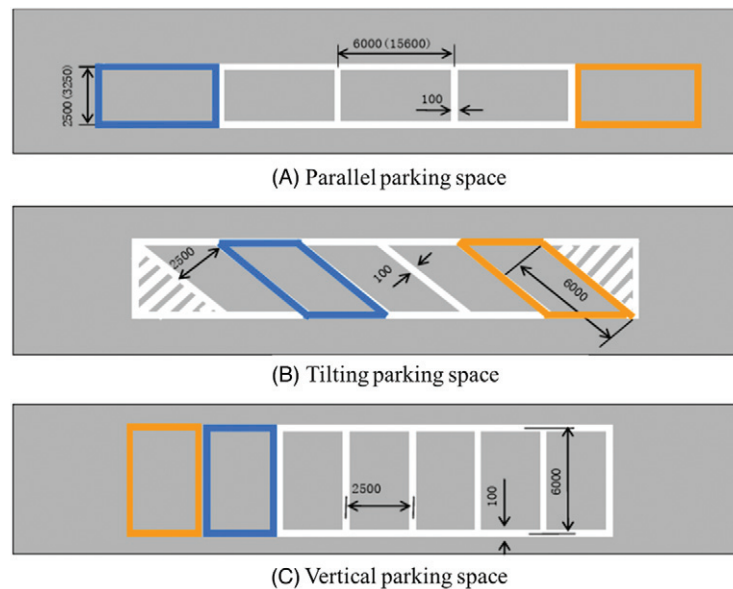


Figure 1 Design layout of Chinese on-street parking spaces (GA/T850-2009).

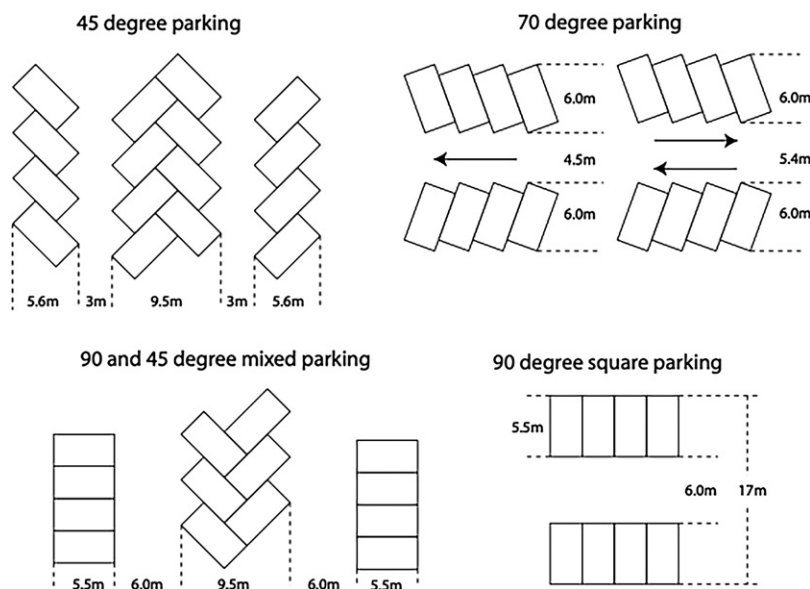


Figure 2 Examples of parking arrangements.

75 State Street Parking Rates	
Each new business day begins at 5am. Parkers who stay beyond 5am will be charged additional fees for another day.	
20 Minutes or Less	\$ 8.00
40 Minutes or Less	18.00
1 Hour or Less	26.00
Maximum Daily Rate	36.00
Weekend Rate	\$ 17.00
Friday 5pm - 5am	per day
Saturday & Sunday 5am - 5am	
Credit Cards Accepted:	
Parking charges are for a license to park only; no judgment is intended or created at this facility. Customers without tickets will be charged the maximum daily rate. Proof of identity will be required.	
617-443-2817 BostonParking.spplus.com	
Sponsored by 	

Figure 3 Parking rates increase strategy.

Parking Supply and Demand Regulation Strategies

If the urban on-street parking resource is not under management, more drivers will park in on-street parking lots for a long time or illegally park on urban roads and pedestrian passages, which deviates from the positioning function (temporary short parking) of on-street parking lots, seriously interferes with urban motor vehicle flow, non-motor vehicle flow and pedestrian traffic, and affects the urban landscape. In order to maximize the utilization of parking resources in limited land, some scholars and relevant urban management departments discuss the regulation of on-street parking demand from two aspects: parking demand regulation strategies and parking supply regulation strategies.

In terms of parking demand regulation strategies, they mainly rely on law enforcement management, parking pricing policy, and management of supply-demand relationship to regulate on-street parking demand. There are different measures to increase the turnover of on-street parking. For example, parking rates increase exponential with the duration of parking. **Fig. 3** is one of such examples to discourage long-term parking. As for parking rates, the parking demand, the parking price of off-street parking lots, the parking spaces occupancy and other factors should be comprehensively considered to determine the pricing standard of on-street parking lots. San Francisco has introduced the "SF Park," which are sensors installed on parking spaces to detect when the parking spaces are occupied, and formulates parking prices based on the data of parking spaces in various regions, so as to control on-street parking demand (**Chatman and Manville, 2014**). Analogously, the utilization rate of parking spaces in different blocks were investigated in Seattle to find the optimal parking price. (**Pu et al., 2017**).

In the aspect of parking supply regulation strategies, the supply strategies are mainly carried out on the premise of ensuring the level of road service (traffic flow and running speed) and reducing network congestion, including determining the maximum supply scale of on-street parking spaces, the parking time limit, etc. There are also some schemes to prohibit on-street parking during peak period and parking is only allowed during off-peak and/or inter-peak periods, as shown in **Fig. 4**.

Considering the air and noise pollution caused by on-street parking, the impact on ecological environment should also be taken into account when setting parking supply and demand regulation strategies. In addition, future parking supply and demand regulation strategies (parking price adjustment, parking spaces scale adjustment, and parking time adjustment) should be dynamically implemented according to real-time parking demand prediction results and traffic flow condition of road network. At the same time, parking spaces utilization information collected by intelligent multi-source big data (GPS, video, vehicle network, and other data) can be used to analyze the impact of dynamic regulation strategies on the on-street parking demand, and then the effectiveness of dynamic regulation strategies can be evaluated.

Intelligent and Informational Parking Management

With the rapid development of smart phones, parking information monitoring and communications media, more parking information services (e.g., vacancy information inquiry, parking space reservation, parking guidance, self-service payment, charging of



Figure 4 Prohibit on-street parking during peak period.

new energy vehicles, etc.) and parking resource management demand have been derived, and a lot of parking Apps and intelligent management system or platform has accordingly appeared. The parking App mainly serves drivers, the intelligent management system or platform mainly serves relevant management departments.

The premise for the parking Apps companies to enter the market is to establish a cooperation mechanism with the managements of parking resource (e.g., government, the operator of the parking lot), so as to obtain the real-time parking information (the number of remaining parking spaces, charging information, etc.) of parking lots. The real-time information of on-street parking lots is generally obtained by the geomagnetic detector, high-level video detection equipment, etc. The real-time information of off-street parking lots is generally obtained by the access control system, etc. After that, the background system of the parking Apps transforms the real-time parking information into visual information which is released to the driver through the interface of parking Apps. The relevant information includes the real-time number of remaining parking spaces (Fig. 5), parking charge information (Fig. 6),



Figure 5 The real-time number of remaining parking space (APP: PP Parking).

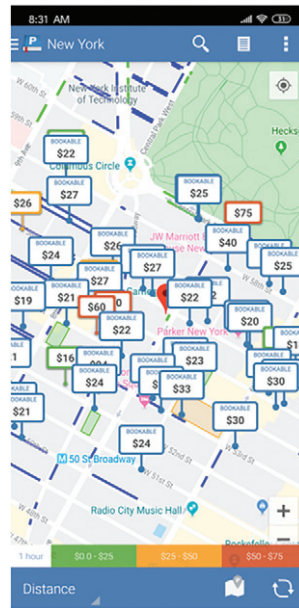


Figure 6 Parking charge information (APP: park opedia).



Figure 7 Electric vehicle charging station (APP: PP Parking).

electric vehicle charging station (Fig. 7) and other information of the parking lots near the destination. When the driver selects a parking lot, some Apps can provide the services of parking space reservation (Fig. 8) and parking path navigation (Fig. 5). When the driver arrives at the parking lot, some Apps can also provide the service of indoor navigation (Fig. 9). When the driver leaves the parking lot, the parking fee can be paid to the parking APPs by swiping the card, scanning the code (Fig. 9) and unconscious pay. The key to the development of parking App is to get enough parking resources as much as possible.

Parking intelligent and information technology can not only serve the driver, but also facilitate the management of parking resources by relevant parking departments. On the one hand, it is necessary to improve the standardized management of parking behavior on the on-street parking lots, such as collecting evidences of fee evasion and vehicles scraping, etc. On the other hand, excessive utilization of on-street parking resources will interfere with dynamic traffic flow, relevant management departments also hope to use intelligent on-street parking management system or platform to timely release parking information and road traffic information to drivers, facilitating drivers to make decisions (choose the off-street parking lots, change the travel mode, etc.) as well as maintaining a good on-street parking order.

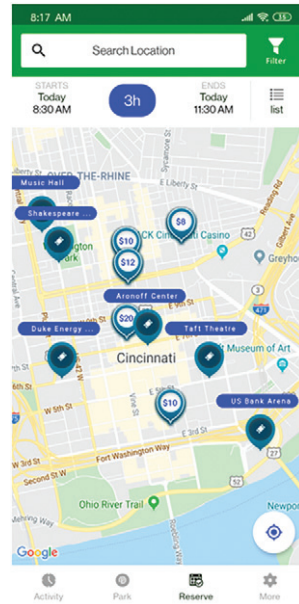


Figure 8 parking space reservation service (APP: park mobile).

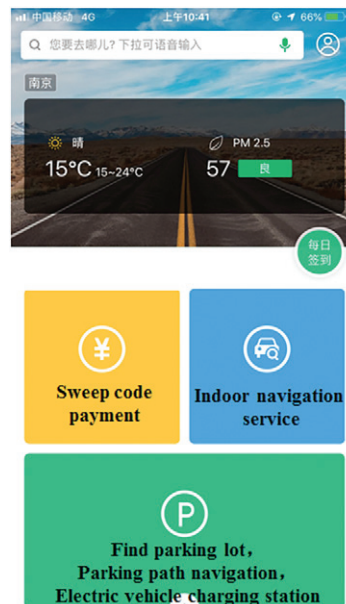


Figure 9 Indoor navigation service (APP: PP Parking).

However, due to the reality that parking information of all kinds of parking facilities (on-street, off-street, and building accessorial parking lots) is not shared, uncritical on-street parking control, drivers' parking preference and other reasons, too many drivers still choose on-street parking or illegal parking on roads, result in low utilization of off-street parking resources. Therefore, in order to make full use of the overall parking resources of the city, an urban intelligent parking platform integrating on-street, off-street and building accessorial parking lots should be constructed. The parking availability information of various parking lots should be timely released to drivers so as to facilitate drivers to make parking decisions and improve parking convenience. Thus, the social and economic benefits of urban parking facilities resources will be improved.

See Also

Parking Price Elasticities; Parking Lots; Parking Information System

References

- Chatman, D.G., Manville, M., 2014. Theory versus implementation in congestion-priced parking: An evaluation of SFpark, 2011–2012. *Res. Transp. Econ.* 44 (1), 52–60.
- Pu, Z., Li, Z., Ash, J., Zhu, W., Wang, Y., 2017. Evaluation of spatial heterogeneity in the sensitivity of on-street parking occupancy to price change. *Transp. Res. Part C* 77, 67–79.
- The Industry Standard of Public Security of the People's Republic of China, 2009. Setting Specification for Parking Spaces in Urban Road (GA/T850-2009).
- Zhenyu, M., Chen, J., 2012. Modified motor vehicles travel speed models on the basis of curb parking setting under mixed traffic flow. *Math. Prob. Eng.* 2012, 1–14.

Further Reading

- Arnott, R., Inci, E., 2010. The stability of downtown parking and traffic congestion. *J. Urban Econ.* 68 (3), 260–276.
- Chatman, D.G., Manville, M., 2014. Theory versus implementation in congestion-priced parking: An evaluation of SFpark, 2011–2012. *Res. Transp. Econ.* 44, 52–60.
- Ma, C.Q., Wang, Y.P., Wei, G., 2011. Curb parking supply scale with conditions of dynamic traffic. *Appl. Mech. Mater.* 71–78, 4876.
- Kong, D., Li, F., Zhang, B., 2018. Design and implementation of intelligent management system for urban road parking. *J. Phys.* 1087, 062061, doi:10.1088/1742-6596/1087/6/062061.
- Horng, G.J., 2019. The cooperative on-street parking space searching mechanism in city environments. *Comp. Elect. Eng.* 74, 349–361.
- Millard-Ball, A., Weinberger, R.R., Hampshire, R.C., 2013. Is the curb 80% full or 20% empty? Assessing the impacts of San Francisco's parking pricing experiment. *Transp. Res. Part A Policy Prac.* 63, 76–92.
- Zheng, N., Geroliminis, N., 2016. Modeling and optimization of multimodal urban networks with limited parking and dynamic pricing. *Transp. Res. Part B: Method.* 83, 36–58.
- Pu, Z., Li, Z., Ash, J., Zhu, W., Wang, Y., 2017. Evaluation of spatial heterogeneity in the sensitivity of on-street parking occupancy to price change. *Transp. Res. Part C* 77, 67–79.
- Rochford District Council, 2010. Parking Standards Design and Good Practice Supplementary Planning Document.
- Ye, X., Chen, J., 2013. Impact of curbside parking on travel time and space mean speed of nonmotorized vehicles. *Transp. Res. Rec.* 2394, 1–9.
- Mei, Z., Chen, J., 2012. Modified motor vehicles travel speed models on the basis of curb parking setting under mixed traffic flow. *Math. Prob. Eng.* 139–1139.

Off-Street Parking

Jun Chen, Guang Yang, School of Transportation, Southeast University, Nanjing, China

© 2021 Elsevier Ltd. All rights reserved.

Parking Supply Allocation	285
Management Strategy	285
Parking Charging Research	286
Parking Sharing Research	286
Intelligent and Informational Parking Management	286
Parking Guidance Information Technology	287
Parking Information Service Platform	287
Parking Support Technology for New Automobile Industry	287
See Also	288
References	288
Further Reading	288

Parking Supply Allocation

Off-street parking lots occupy a large amount of urban space resources, and unreasonable supply allocation of off-street parking lots will result in a waste of parking resources. Therefore, it is necessary to rationally determine the supply allocation of off-street parking lots (parking spaces scale, location layout) based on parking demand prediction theory and parking supply theoretical method.

Parking demand prediction theory is one of the basis for determining the parking spaces scale of off-street parking lot, it is mainly based on the parking spaces utilization characteristics of different types of land in different periods of time (working days, weekends, and peak or off-peak parking periods of a day), parking demand influencing factors, parking behavior characteristics (vehicles arrival and departure, parking time, parking times, etc.) and parking management strategies (parking sharing, etc.), and to comprehensively analyze the parking demand of all kinds of land-use within the influence scope of the off-street parking lot.

On the basis of research of parking demand prediction theory, some scholars and relevant management departments further consider extra factors such as urban parking management strategy (zonal differentiated parking supply, etc.), the interaction mechanism between dynamic and static traffic. They put forward theoretical methods of parking supply for off-street parking lots, including determining the parking spaces scale and location layout.

In terms of the determination method of parking spaces scale, there are differences between building accessorial parking lots and public parking lots. For building accessorial parking lots, the main research contents are establishing principles and methods of parking requirement standard. Before 1970s, large cities in Europe, America, and other countries built parking facilities in strict accordance with the minimum requirement standard of off-street parking facilities. After 1990s, the parking requirement gradually changed to the highest standard (Weinberger et al., 2010). At present, according to the concept of "zonal differentiated parking supply", some cities in China have formulated differentiated parking requirement standards for different types of buildings in different zones. When calculating the reasonable parking spaces scale of public parking lots, many factors should be considered, such as parking demand, parking management strategy (parking spaces control index of public parking lots, zonal differentiated parking supply), city's socioeconomic attributes (population, motor vehicle ownership, etc.), urban land development and utilization condition, urban traffic development strategy, the interaction mechanism between dynamic and static traffic, etc.

For the determination method of the location layout of off-street parking lots, the main optimization objects involve public parking lots and P + R (Park-and-Ride) parking facilities. For public parking lots, the restrictions such as parking land, the impact of location layout on dynamic traffic and drivers' travel cost (driving distance, walking distance) are mainly considered. For P + R parking facilities, location layout optimization focuses on combining with urban traffic transfer hubs, the optimization goals are to minimize the park-and-ride cost of users and improve the utilization of urban traffic transfer hubs (Farhan and Murray, 2008; Aros-Vera et al., 2013). At present, the research on the coordinated layout optimization of multi-type parking facilities is relatively weak, such as the coordinated location optimization of on-street and off-street parking facilities.

Management Strategy

Off-street parking lots occupy huge urban space resources and require high construction and maintenance costs. In order to give full play to the social benefits (meeting parking needs, ensuring a certain proportion of parking spaces occupancy rate or turnover rate) and economic benefits (parking operation income) of off-street parking lots, some management strategies need to be adopted to improve the parking spaces utilization of off-street parking lots. On the one hand, the parking profits can be realized by setting

reasonable parking charges in order to ensure stable customer flow and orderly parking of vehicles. On the other hand, through parking sharing measures, the vacant parking resources can be activated, and the parking resources in the areas can be fully utilized over different period of time.

Parking Charging Research

When the parking costs of off-street parking lots exceed on-street parking lots, drivers will cruise to urban roads for an on-street parking space. Since cruising can affect dynamic traffic, many researchers have estimated that on-street cruise vehicles account for about 40% of the total traffic volume in some neighborhoods (Garrett, 2014). When the parking costs of off-street parking lots are lower than on-street parking lots, drivers will not be motivated to look for an on-street parking space. However, low income can hardly improve the economic benefits of off-street parking lots, and also lower the enthusiasm of social capital to invest the construction of off-street parking lots. Therefore, it is necessary to develop a reasonable pricing system for off-street parking lots to improve the efficiency of parking resources and the economic benefits of parking lots.

The influencing factors for off-street parking prices mainly include the comprehensive costs, economic incentives, economic affordability of drivers, relationship between parking supply and demand, on-street parking prices, parking time, and parking policy. In addition, the parking price is inseparable from the type of off-street parking lots. With regards to the pricing target of public parking lots, maximization of the multi-party benefits (drivers, investment operators, and parking management departments) is considered to achieve parking convenience, market development, and traffic management; whereas for the industrial investment parking lots, the most important consideration is the maximization of economic benefits. Due to the high-investment costs of off-street parking lots and the long cycle of capital recovery, scientific operation, development mode, and subsidy method will become the research focus; the low charge of P + R parking facilities mainly attracts park-and-ride drivers to use public transport to achieve the purpose of traffic management; the pricing standard of building accessorial parking lots for parking sharing mainly considers the balance of interest between parking spaces owners, users, and operating companies. In the future, based on the analysis of impact of dynamic parking charges on parking demand and benefits through intelligent multi-source big data, solving the dynamic optimal parking pricing will be a hot issue.

Parking Sharing Research

In recent years, with the rapid increase in car ownership in cities, limited-parking facilities can no longer meet the growing demand for parking. In view of this, a number of scholars and relevant urban management departments have proposed alleviating the contradiction between parking supply and demand by sharing parking resources. The main feature of parking sharing is the dynamic matching of parking resource supply and parking demand for various purposes by opening the parking space use authority (permit) while not changing the total supply of parking spaces in the area and ensuring the primary demand of parking in the various buildings. In order to implement parking sharing, it is necessary to research parking sharing resource allocation and parking sharing management policies.

The allocation of shared parking resources is determining the type of shared parking facilities, the number of shared parking spaces, and the opening time windows by using space-time technology and optimization model. In the above process, the matching degree of parking supply and demand, the sharing willingness, the travel characteristics of parking users and the sharing economic benefits should be considered. In order to promote the development of parking sharing policy, some relevant departments and scholars have carried out research on parking sharing management measures, including the development of universal parking spaces sharing step, the establishment of parking sharing system (management and supervision subjects, operating mode), gradual reduction or elimination of the parking spaces in the surrounding roads, and determining the cost of parking sharing and the number of reserved parking spaces. The above research may provide a theoretical basis for relevant departments, operating companies to purchase shared parking spaces, analyze customer groups, and formulate operational management strategy.

Intelligent and Informational Parking Management

The off-street parking lots' parking spaces status have time-varying characteristics, and the large parking lots have a large number of parking spaces and a complex spatial structure. In consideration of the parking guidance needs of drivers in the above scenarios, it is necessary to investigate the relevant parking guidance information technology. In addition to the rapid development of vehicle equipment, smart phones, parking supply and demand information identification, matching and transmission technology, more parking information services (information inquiry, parking guidance, parking reservation, self-service payment, etc.) have been derived. In order to make it convenient for drivers to enjoy diversified parking information services, it is necessary to build parking information service platform. With the development of new automotive industry (driverless technology, time-sharing rental vehicles, new energy vehicles, etc.), the parking demand of new automotive industry needs to be addressed through advanced technical means and model algorithm.

Parking Guidance Information Technology

Drivers often need parking guidance information during travel and arrival of large parking lots: (1) Before or during the travel, drivers often choose an appropriate parking lot around the destination based on their parking preferences (acceptable maximum driving distance, walking distance, charge, etc.), road congestion, and other factors. However, the availability of parking spaces in the parking lot has time-varying characteristics, and drivers need to know about availability information of parking spaces during the travel, in order to make timely decision (continue to drive, change parking lot, park-and-ride, etc.) to prevent parking failure; (2) When drivers arrive at a large parking lot, due to the large number of parking spaces and the complex spatial structure, it is necessary to use parking guidance information to induce vehicles from the entrance to the appropriate parking space and to provide the function of searching vehicles reversely. Considering the drivers' parking guidance information demand in the two travel stages, it is necessary to research the parking guidance information technology in corresponding stage.

1. Parking guidance method in the travel

Urban Parking Guidance Information System (PGIS) can provide accurate and real-time information on parking spaces for drivers in the travel phase. At the same time, PGIS can balance the spatial-temporal distribution of parking vehicles in various urban parking facilities, and make full use of urban parking facilities. The VMS (Variable Message Sign) design is the main research content of PGIS, consisting the research on information release and location selection. In the aspect of information release, as the remaining parking spaces have time-differentiated characteristics, it is necessary to predict the number of vacant parking spaces in a short period of time in future, and release relevant information to drivers through VMS to facilitate drivers to make reasonable decisions (continuation of driving, change parking lot, park-and-ride, etc.) in accordance with their travel time. In terms of location selection research, the VMS location scheme should be comprehensively identified from the perspectives of vehicles guidance demand, VMS service level and operation effect, road network traffic flow, system investment, and environmental impact.

2. Vehicles guidance and other service functions in large parking lots

In order to better realize vehicles guidance in large parking lots, it is necessary to accurately judge the occupancy status of each parking space, and then transmit the parking space status information to the management system. The management system assigns parking spaces and recommends paths to drivers. Furthermore, large parking lots should provide reverse parking search service to guide drivers to walk from the pedestrian entrance to their parking spaces and exit to leave the parking lot. At present, the status recognition of parking spaces, guidance of vehicles and pedestrian paths and other service functions (real-time monitoring, parking charge, parking information inquiry, intelligent anti-theft protection) can be achieved through geomagnetic sensors, RFID (Radio Frequency Identification Devices) technology, image recognition, RSU (Road Side Unit) technology, map matching algorithm, and other technologies.

Parking Information Service Platform

VMS is a relatively traditional parking guidance technology. However, with the rapid development of vehicle equipment, smart phones, parking supply and demand information identification, matching, and transmission technology, more complex parking information service platforms have been formed. The parking information service platform can provide multiple parking service information (vacancy, reservation, guidance, and charging) through the Internet, Internet of Things, Vehicle Network, and other media, and realize more convenient and intelligent parking service. For example, [Safi et al. \(2018\)](#) proposed a parking information system based on cloud platform, which provides parking information and reservation services to drivers, taking into account driving and walking time, traffic congestion, and other factors.

However, based on the parking survey of some cities in China (Wuhan, Xiamen, Nanjing, and Shenyang), there are differences in the management departments of different types of parking lots (off-street, on-street, and building accessorial parking lots). Different types of parking lots belong to the government or private property, so the government can not integrate all kinds of parking lots' parking information (parking spaces utilization status, parking charge, etc.) into the parking information service platform. In order to further enhance the balanced use of different types of urban parking resources and the convenience of drivers, relevant management departments should adopt administrative means, incentive mechanisms, tax policies and other measures to integrate all kinds of parking lots into the parking information service platform, giving full play to the social and economic benefits of urban parking resources. In addition, the current parking information service platform mainly serves traveling individuals and parking lots operators, and the future service functions should also serve the government and related industries, such as the realization of location tracking, alarm for special vehicles, and providing basic data for the coordinated management of dynamic and static traffic.

Parking Support Technology for New Automobile Industry

With the development of new automobile industry (driverless technology, time-sharing rental vehicles, new energy vehicles, etc.), it is necessary to solve the parking demand of new automobile industry through advanced technical means and model algorithm. For driverless technology, the future system framework of intelligent parking lots should consider the parking process, parking rules,

parking demand of driverless vehicles and coexistence mode between traditional vehicles and driverless vehicles, and design an driverless module which integrates parking demand response, vehicle access control identification and parking space allocation. For time-sharing rental vehicles, it is necessary to consider factors such as vehicles rental demand, the convenience for users to pick up and return vehicles, vehicle scheduling strategy, etc., so as to determine the location layout and scale of time-sharing rental parking lots. For new energy vehicles, we need to consider the parking and charging demand of vehicles, the increasing trend of the number of new energy vehicles and other factors, and discuss the location layout of charging piles in the city, as well as the scale standards of charging piles in different zones and different types of parking lots.

See Also

Economics of Parking; Parking Price Elasticities; Parking Lots; Parking Information System; Parking Demand Models

References

- Aros-Vera, F., Marianov, V., Mitchell, J.E., 2013. P-hub approach for the optimal park-and-ride facility location problem. *Eur. J. Operat. Res.* 226 (2), 277–285.
 Farhan, B., Murray, A.T., 2008. Siting park-and-ride facilities using a multi-objective spatial optimization model. *Comp. Operat. Res.* 35 (2), 445–456.
 Garrett, M., 2014. *Encyclopedia of Transportation*. SAGE Publications, Inc. Thousand Oaks, California, 2014, pp. 1038.
 Safi, Q.G.K., Luo, S., Pan, L., et al., 2018. SVPS: Cloud-based smart vehicle parking system over ubiquitous VANETs. *Comp. Network* 138, 18–30.
 Weinberger, R., Kaehny, J., Rufo, M., 2010. *US parking policies: An overview of management strategies*. The Institute for Transportation and Development Policy, New York.

Further Reading

- Alejandro, C., Guillem, B., Antoni, M., et al., 2017. Autonomous car parking system through a cooperative vehicular positioning network. *Sensors* 17 (4), 848.
 Aros-Vera, F., Marianov, V., Mitchell, J.E., 2013. P-hub approach for the optimal park-and-ride facility location problem. *Eur. J. Operat. Res.* 226 (2), 277–285.
 Auchincloss, A.H., Weinberger, R., Aytur, S., Namba, A., Ricchezza, A., 2015. Public parking fees and fines: A survey of U.S. cities. *Public Works Manag. Policy* 20 (1), 49.
 Michael, C.H., Hadisaputra, K. R., et al., 2017. Smart Parking Management System: An integration of RFID, ALPR, and WSN. 2017 IEEE 3rd International Conference on Engineering Technologies and Social Sciences (ICETSS). IEEE.
 Shao, Haoyi, Yang, Hai, Zhang, Yi, Ke, Jintao., 2016. A simple reservation and allocation model of shared parking lots. *A simple reservation and allocation model of shared parking lots. Transp. Res. Part C* 71, 303–312.
 Institute O. T. E., 2018. *Parking generation*, Fifth ed. Institute of Transportation Engineers, Washington, DC.
 Ji, Y., Tang, D., Blythe, P., et al., 2015. Short-term forecasting of available parking space using wavelet neural network model. *IET Intel. Transp. Sys.* 9 (2), 202–209.
 Hao, J., Chen, J., Chen, Q., 2018. Floating charge method based on shared parking. *Sustainability* 11 (72), 1–14.
 Safi, Q.G.K., Luo, S., Pan, L., et al., 2018. SVPS: Cloud-based smart vehicle parking system over ubiquitous VANETs. *Comp. Network.* 138, 18–30.
 Weinberger, R., Kaehny, J., Rufo, M., 2010. *US parking policies: An overview of management strategies*. The Institute for Transportation and Development Policy, New York.
 Waldemar Parkitny, 2017. Using the theory of games to modelling the equipment and prices of car parking. *Conference Series: Materials Science and Engineering*. 245.

Parking Information Systems

Behrang Assemi, Douglas Baker, School of Built Environment, Queensland University of Technology (QUT), Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	289
Introduction	289
PIS at a Glance	290
PIS Deployment	290
System Architecture	290
Detection/Monitoring	291
Data Collection and Processing	292
Information Dissemination	292
Communication Technologies	292
Requirements of PIS	292
Categories of PIS	292
Common Features and Functions of PIS	293
Intelligent Algorithms, Booking and Customer Management	294
Access Control, Automation, and Emerging Features	296
Benefits and Advantages of PIS	296
Improved Traffic Efficiency	296
Enhanced Customer Satisfaction	296
More Effective Parking Management	296
Acknowledgment	296
Biography	297
References	297
Further Reading	298

Nomenclature

API Application Programming Interface
ATIS Advanced Traveler Information Systems
CAPS Centralized Assisted Parking Search
COINS Car Park Occupancy Information System
IoT Internet of Things
IPIS Intelligent Parking Information System
IR Infrared
ITS Intelligent Transport Systems
LPR License Plate Recognition
PGIS Parking Guidance (and) Information System
PIS Parking Information Systems
RFID Radio Frequency Identification
RSPS Reservation-based Smart Parking System
VMS Variable Message Sign

Introduction

Parking is usually considered a significant contributor to the overall performance of the traffic network, especially in dense urban areas that are heavily car-oriented. Cruising for parking can adversely impact on traffic by increasing congestion and likelihood of incidents (Wang and He, 2011). It lengthens travel times, increases fuel consumption, and worsens the environmental effects of cars.

Previous research has shown that pre-trip information about parking can help customers reduce their parking search time, the corresponding fuel consumption and the likelihood of road accidents, while enhancing customers' overall parking experience (Bock et al., 2019; Kim et al., 2014; Rajabioun et al., 2013; Yang and Lam, 2019). Parking information systems (PIS) provide drivers with real-time information about parking availability and navigate them to a vacant space to assist them with their trip planning (Milosavljevic and Simicevic, 2019).

PIS are considered a crucial part of smart city management frameworks, specifically intelligent transport systems (ITS) (Yang and Lam, 2019; Yang et al., 2003). Given their role in provision of information to assist customers with their travel planning, PIS can be integrated into advanced traveler information systems (ATIS) (Shiue et al., 2010). ATIS, as the most comprehensive trip planning solutions, aim at providing road users with accurate and reliable real-time information about traffic, parking and available alternatives enabling them to make more informed decisions. ATIS in general, and PIS in particular, help improving the overall traffic flow, reducing congestion, incidents and environmental effects, and enhancing user experience (Li, 2019; Milosavljevic and Simicevic, 2019).

PIS at a Glance

One major objective of PIS is to provide drivers with real-time availability details, specifically to reduce cruising for parking (Teng et al., 2008; Yang and Lam, 2019). PIS may also include routing, fee collection and payment management, and availability forecasting features, in addition to the system's common information provision function (Hui-ling et al., 2003; Rajabioun et al., 2013; Yang et al., 2003). PIS may incorporate advanced algorithms for routing drivers to the most suitable parking space, considering each user's preferences and past parking behavior (Rajabioun et al., 2013). Furthermore, these systems can provide users with estimated cruising time, given real-time/forecasted parking availability (Chen et al., 2016), and add it to their total travel time. PIS may also provide customers with space reservation capability, where the infrastructure supports it (Wang and He, 2011).

The information provided by PIS is either static or dynamic, as summarized by Hui-ling et al. (2003), one of the early studies on PIS. Static information may include parking lot location, total capacity, fees, physical restrictions (such as height and length limits), and usage regulations (such as maximum length of stay). Dynamic information includes real-time space availability and future occupancy forecast.

PIS use historical data, real-time data or both (Chen et al., 2016). Parking was usually neglected by conventional traffic management systems; however, more attention is being paid to parking and its impact on traffic in recently developed traffic management systems. Traditionally, traffic management systems relied heavily on historic data and well-established models. These systems are now increasingly integrating real-time data for more accurate and effective traffic management by gaining access to data produced by advanced ITSs. The integration of real-time data on traffic and parking enhances evidence-based and timely decision-making, which is necessary for efficient and safer use of traffic networks (Milosavljevic and Simicevic, 2019).

PIS Deployment

Most PIS rely on a network of space occupancy sensors and/or cameras to determine parking availability and to function effectively (Chen and Chang, 2011; Wang and He, 2011). Data from sensors and cameras transferred through a dedicated network to a central system are consumed by PIS based on the system objectives (Chen et al., 2016; Hui-ling et al., 2003).

PIS are often utilized along with variable message signs (VMSs) and/or web interfaces to inform customers of parking availability (Kim et al., 2014; Sun et al., 2016; Teng et al., 2008). VMSs usually guide customers through "en-route parking search," while a web interface can assist customers with their "pretrip parking planning" (Teng et al., 2001, 2008). Emerging interfaces such as mobile apps and in-vehicle parking guidance systems can help customers with their search and planning at both stages, pretrip and en-route (Teng et al., 2008).

The main factors recommended by the literature to be considered before deployment of a parking information system include (Teng et al., 2008):

- difficulty of finding parking space as perceived by customers,
- customers' technology preference, and
- system costs.

Many PIS, relying on smart parking technologies, have been developed over the past decade. These systems, especially deployed for on-street parking, can play a crucial role in effective implementation of parking policies (Wang and He, 2011), such as reducing total vehicle mileages, increasing parking turnovers and encouraging use of off-street parking and more sustainable modes of transport. Table 1 lists examples of smart on-street parking trials and compares them in terms of their underlying technologies and overall performance.

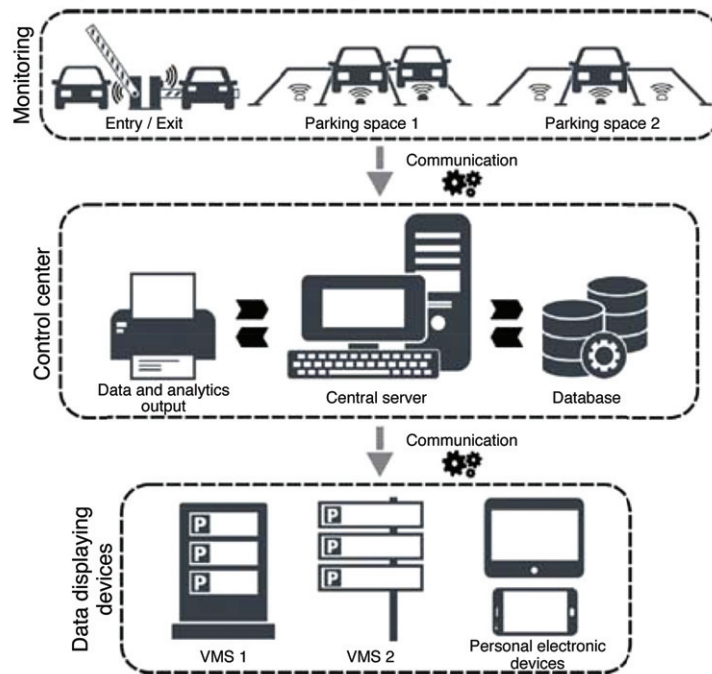
System Architecture

A typical parking information system consists of four main elements, including detection (monitoring) mechanism, data collection and processing unit (control center), information dissemination (display) mechanism, and communication technologies (Kotb et al., 2017; Lee et al., 2016; Ni and Cao, 2017). The detection mechanism often includes a range of inter-related sensors and smart technologies used for monitoring parking utilization and controlling access-payments. The data collection and processing unit is responsible for aggregating all historic and real-time parking-related data obtained through the detection mechanism as well as other ITS technologies. This unit also processes the data, forecasts future occupancies and handles user requests such as reservations. The information dissemination mechanism is performed either through VMSs and/or web/smartphone interfaces. Finally, the communication technologies handle the transmission of data/information between the aforementioned system components. Figure 1 illustrates the abstract architecture of a common parking information system.

Table 1 Examples of smart parking trials

System	Years of operation	Technology	Reported accuracy	Notes
ParkNet	2010	Sonar	95%	Used Crowdsources data collectuin too-used less than one sensor per bay
ParkSense	2013	Mobile phones Wi-Fi	83%	Used beacons between mobile phones and Wi-Fi to infer parking status
Street parking System	2013	Magnetic	98%	
SFPark	2014	Magnetometer	86%	On and off-street solution-more than one sensor per bay
Smart Santander	2014	Ferromagnetic	Not Specified	Part of Smart City project
FASTPAK	Continuous	Magnetic	95%	includes a smartphone application
Geomii	Continuous	Magnetometer	86%	Provides real-time and forecasted parjubg availability
Integrated smart parking	Continuous	Radar	Not Specified	Traffic flow is sensed too -less than one sensor per bay required
Parking spotter	Continuous	Sonar/Radar	Not Specified	Uses crowdsourced data collection too -uses less than one sensor per bay
Smart parking	Continuous	Infrared and magnetic	Not Specified	

Source: Adapted from Roman et al. (2018)

**Figure 1** PIS architecture. Source: Adapted from Milosavljevic and Simicevic (2019)

Detection/Monitoring

Smart sensors and technologies are the key element in the architecture of PIS; they provide the system with the data required for effective planning and management of parking (Paidí et al., 2018). The detection mechanism, relying on sensors, may collect real-time data at a parking facility's entry/exit and/or at each single bay (Ni and Cao, 2017). The widespread use of smart sensing technologies, as the basis of the Internet of Things (IoT), may also support the detection/monitoring mechanism (Lee et al., 2016); sensors deployed for other ITS services may also be able to collect data on parking utilization.

Historical data or sample observations might also be used in lieu of, or in conjunction with, smart technologies collecting real-time data (Paidí et al., 2018). However, a lack of access to real-time and accurate data may adversely impact on the reliability and effectiveness of PIS.

Data Collection and Processing

Another important element of PIS is a data collection and processing unit, which includes a central data warehouse or data lake, where all parking data are aggregated. Other relevant data, such as traffic and weather conditions, might also be integrated with parking data to provide a more comprehensive basis for processes performed by PIS (Li, 2019). The aggregated, and potentially processed data, can efficiently be shared with other transport management systems and third-party applications (Dalangauskas et al., 2016), often through application programming interfaces (APIs).

The data collection and processing unit is also responsible for data analysis, forecasting, and handling user requests. When a parking information system navigates users to a specific parking place, this unit also handles optimal route generation and user navigation. Big data methods and cloud-based technologies are increasingly used by PIS to efficiently store and process large volumes of data (Fang et al., 2016).

Information Dissemination

Dissemination of information is usually handled through VMSs, web/smartphone apps or both. VMSs may be used in two different positions helping users with their en-route decision-making and navigation to the optimal parking bay (Ni and Cao, 2017): (1) on the road or at the entrance of parking facilities showing availabilities and other relevant information such as prices, and (2) along the path in a parking facility guiding users to best available alternatives. Web interfaces and smartphone applications, on the other hand, provide users with more detailed information about parking and help them with pretrip planning as well as en-route decision-making and navigation (Ni and Cao, 2017).

Communication Technologies

Finally, PIS rely on a range of wired and wireless communication technologies to relay real-time data from sensors and smart technologies to data warehouses and processing units. Such technologies are also needed to transmit system outputs and commands to parking management technologies and VMSs. Similar to smart sensors, these communication technologies might be part of a bigger ITS.

Requirements of PIS

As mentioned earlier, smart sensors and technologies which collect data on parking occupancy and utilization are essential for effective functioning of PIS. These sensors range from magnetometers to LPR, and vary widely in terms of their cost, accuracy and collected details. This is especially important, as the reliability and effectiveness of PIS depends on the accuracy and timeliness of the data collected by these sensors (Milosavljevic and Simicevic, 2019). Table 2 shows the range of sensors which are commonly used in conjunction with PIS.

Overall, these sensors and smart technologies are categorized into “on-roadway” and “off-roadway” (Kotb et al., 2017). On-roadway sensors, such as loop detectors, magnetic sensors, and radio frequency identification (RFID) sensors are installed on a road’s surface or embedded into it. Off-roadway sensors, such as ultrasonic sensors, cameras, and ceiling-mounted infrared (IR) sensors are installed above the roadway. Sensors might also be classified as active if they depend on an external power source for their operation, and passive, otherwise (Al-Turjman and Malekloo, 2019).

Most smart sensors and technologies are still expensive and might need regular maintenance. Therefore, cost of real-time data collection is an important consideration for deployment of PIS (Paidí et al., 2018). Widespread adoption of ITS and the growing application of the IoT can significantly reduce such cost, as smart sensors and technologies are simultaneously used by more than one system (Lee et al., 2016).

Another important requirement of PIS is wireless communication technologies which transmit real-time data from sensors to data warehouses used by the PIS (El-Seoud et al., 2016). Smartphone applications may be developed to facilitate customer interactions, such as reservation, payment and navigation (Al-Turjman and Malekloo, 2019). Finally, security and privacy should be taken into account in data collection, storage and sharing; secure, encrypted communication is a significant requirement of PIS which determines their adoption by users (Al-Turjman and Malekloo, 2019).

Categories of PIS

PIS are usually categorized depending on their main objectives, although these categories may not be mutually exclusive. A parking information system—in its simplest form—is often referred to as a car park occupancy information system (COINS), and provides parking operators and users with real-time information on parking availability (Al-Turjman and Malekloo, 2019). This category of PIS, acting as a decision support system, mainly serves customers while searching for the best available parking option by providing them with basic information (Paidí et al., 2018). Table 3 presents examples of commercial PIS, which provide customers with availability and location of parking through a web and/or smartphone app interface.

Parking guidance (and) information systems (PGIS) is another major category of PIS, which mainly focuses on navigating users to a predetermined (and potentially reserved) parking bay. PGISs are often used for off-street parking. Centralized assisted parking search (CAPS) is a special type of PGISs often used in garages, which dedicates a parking lot to each incoming vehicle and navigates them to the right place (Al-Turjman and Malekloo, 2019).

Table 2 Parking occupancy monitoring technologies

<i>Sensor type</i>	<i>Intrusive^a</i>	<i>Count^b</i>	<i>Speed detectio</i>	<i>Multi-lane detection</i>	<i>Weather sensitive</i>	<i>Accuracy</i>	<i>Cost</i>
Passive Infrared	–	Y	–	–	Y	**	**
Active Infrared	Y	Y	Y	Y	–	**	**
Ultrasonic	–	Y	–	–	Y	***	*
Inductive loop	Y	Y	Y	–	Y	****	**
Magnetometer	Y	Y	Y	–	–	****	*
Piezoelectric	Y	Y	Y	–	Y	****	***
Pneumatic road tube	Y	Y	Y	–	–	****	****
Weigh in Motion (WIM)	Y	–	–	Y	–	**	****
Microwave	–	Y	Y	Y	–	***	***
Camera (CCTV)	–	Y	Y	Y	Y	****	****
RFID	–	–	–	–	–	**	*
Light Dependent Resistor (LDR)	–	Y	–	–	Y	**	*
Acoustic	–	Y	Y	Y	Y	**	**

^aIntrusive or payment-intensive sensors require changes in the pavement surface for proper installation

^bWhether a sensor can count the number of vehicles or not

Source: Adapted from *Al-Turjman and Malekloo (2019)*

Table 3 Examples of commercial PIS

<i>PIS</i>	<i>Country</i>	<i>Technology</i>
Park.ME	Austria and Germany	Underground sensors
SmartParking	New Zealand	Underground sensors and RFID
ParkMe	Japan, USA, UK, Germany and Brazil	Underground sensors
ParkAssist	USA	M4 Smart sensors and LPR
SpotHero	USA	Underground sensors
EasyPark	Canada	Underground sensors
PaybyPhone	France	Underground sensors
ParkMobile	USA	Underground sensors
AppyParking	UK	Magnetometer
EasyPark Group	Sweden and Denmark	Transactional data and crowdsourcing
Parker	USA	Underground sensors and machine vision
ParkFi	USA	Magnetometer
Best parking	USA	Underground sensors
Parkopedia	USA, Germany and Sweden	Predictive analytics and underground sensors
SFPark	USA	Underground sensors
Open Spot	USA	Crowdsourcing

Source: Adapted from *Paidi et al. (2018)*

Reservation-based smart parking systems (RSPS) is another major category of PIS that enable users to book parking space in advance (*El-Seoud et al., 2016*). Such a parking information system often relies on historic trends and also accommodates forecasting of parking availability for a user's requested time, to guarantee the vacancy of the reserved space at the given time (*Ni and Cao, 2017*).

Automated parking, with vehicle lifts and robotized vehicle displacing technologies, also uses a customized category of PIS which manages the automatic dispatching of vehicles to/from a dedicated bay, in addition to common PIS features (*Al-Turjman and Malekloo, 2019*). These PIS incorporate advanced algorithms to optimize the use of space and maximize operation safety.

Common Features and Functions of PIS

Depending on the purposes of a parking information system and the main category into which it fits, different types of features and functions are offered by the system. **Figure 2** represents the common features of typical PIS, as well as the actors using such a system. These common features can be classified into administrative and operation-related. Administrative features, mostly used by parking operators, provide access to intelligent parking management technologies, payment and transactions data as well as system configuration and troubleshooting. Operation features, used by operators, customers and third parties (such as smartphone

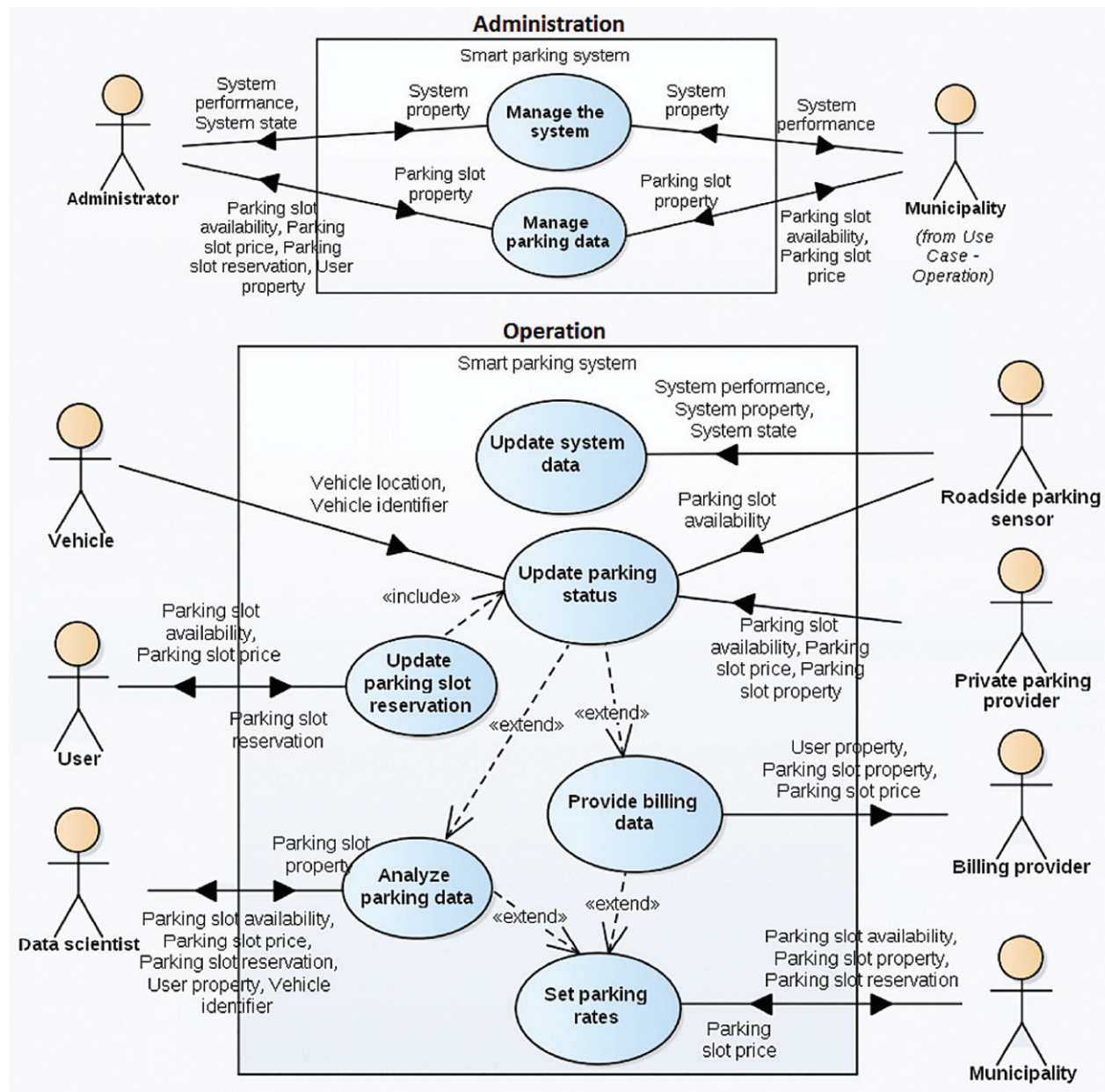


Figure 2 PIS actors and use cases Source: Adapted from Lin et al. (2017)

applications), provide access to analyzed data, facilitate parking utilization and management, and enable automated, smart operations such as contactless entries/exits.

Figure 3 illustrates common features and flow of information in typical PIS. Along with providing real-time information on parking space availability, PIS may also guide users to the optimal parking bay (Yang et al., 2003). This can be done from users' origin, on the road, and/or inside parking facilities. Physical signs and guiding systems, often in the form of VMSs, are commonly used to communicate availability and locations of vacant parking both on-street and in off-street garages (Milosavljevic and Simicevic, 2019). Figure 4 illustrates common information that can be presented through a PIS-connected VMS.

Intelligent Algorithms, Booking and Customer Management

Intelligent algorithms are also being used in advanced PIS. These algorithms can optimize the choice of parking for users, considering their objectives and constraints, such as price, availability likelihood and distance to final destination (Dalangauskas et al., 2016). PIS may even collect personal preferences and use machine learning algorithms to develop user-specific profiles including their final destinations, regular parking patterns and payment choices. Table 4 summarizes potential factors which may be

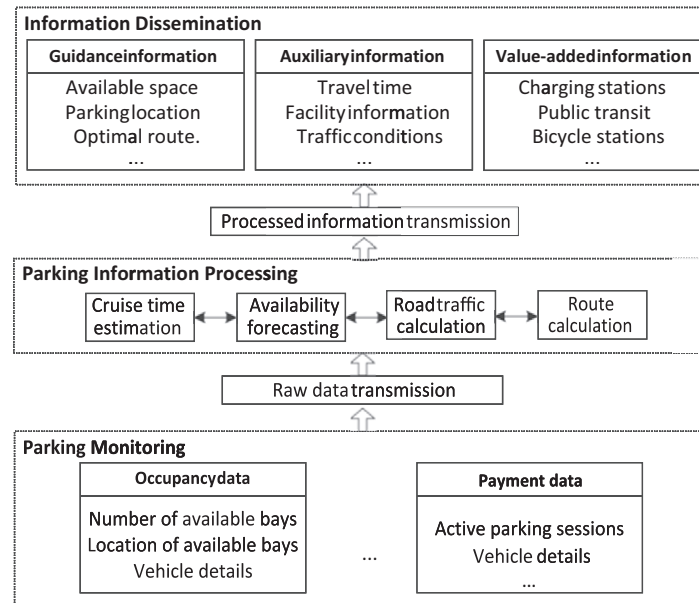


Figure 3 Main features and flow of information in a typical parking information system. Source: Adapted from Li (2019)

Current date and time: 21 May 2020 15:45				
Parking	Total Spaces	Available Spaces	Occupancy	Trend
1	290	100	66%	↑↑
2	480	230	52%	⇒
3	100	20	80%	↓↓

Figure 4 Parking occupancy information. Source: Adapted from Milosavljevic and Simicevic (2019)

Table 4 Summary of potential factors to consider for personalized parking choices

Category	Factor	Description
Parking constraints	Availability	Current availability of parking spots at each location
	Parking rules	Parking duration limits, bin collection, height and length limits
	Rates	Hourly (\$/hour), flat rates, free period
Driver preferences	Parking type	On-street, off-street, valet
	Time or expense	Driver's preference of paying less or reaching their final destination sooner
	Parking duration	Driver's preferred length of stay
Traffic condition	Arrival time	Driver's preferred time of arrival at final destination
	Driving time	Travel time from the current location to each parking alternative
	Walking distance	Walking distance between each parking alternative and final destination
	Total travel time	Sum of driving time and walking time

Source: Adapted from Rajabioun et al. (2013)

considered by PIS while generating and updating user profiles. These profiles are later used to determine the best customized alternative for each user (Rajabioun et al., 2013).

Booking is another feature offered by PIS, which is especially relevant for off-street parking (Yang and Lam, 2019). Accordingly, PIS may also incorporate billing and account management to automate customer management, handle payments and generate invoices. Furthermore, PIS also provide accounting and auditing features to effectively manage parking cashflows (International Parking Institute et al., 2018).

Another main feature of PIS is the ability to forecast the likelihood of parking availability (Bock et al., 2019; Liu et al., 2006). In addition to historic and real-time parking occupancy, a PIS may also integrate weather, traffic flow, road conditions, and special events to improve the accuracy of forecasts (Yang et al., 2003).

PIS often include data analytics features providing parking operators with summaries, models and insights on parking utilization (Liu et al., 2006). These features can be used to generate “actionable” information; the information about trends and system performance which facilitates incremental operation improvements and change implementations (International Parking Institute et al., 2018).

PIS often include features to differentiate potential usages and customers. In addition to time-dependent fee structures, parking regulations and fees might change due to special days and events. Furthermore, parking facilities usually serve different groups of customers with specific needs and preferences (e.g., valet users or mobility-impaired customers). PIS should allow easy specification and implementation of such differentiating structures (International Parking Institute et al., 2018).

PIS may also be capable of determining optimal parking price based on demand and operation objectives (International Parking Institute et al., 2018; Wang and He, 2011). Some PIS incorporate features to detect illegal parking and facilitate efficient enforcement (International Parking Institute et al., 2018; Sadhukhan, 2017).

Access Control, Automation, and Emerging Features

PIS can also be linked to parking gates, license plate recognition (LPR) systems, and other physical instruments, such as weight and height sensors used in automated car parks (Dalangauskas et al., 2016). Contactless payments of parking fees is possible only with the support of PIS (El-Seoud et al., 2016; Yang and Lam, 2019). Therefore, PIS often plays a central role in the automation and enhanced sustainability of modern parking facilities.

Several features and function are also being integrated into advanced PIS to address emerging needs and trends in ITS and sustainable parking management (Ni and Cao, 2017). An important feature is the capability of communicating with self-driving vehicles, specifically to guide them to a predetermined parking bays (Al-Turjman and Malekloo, 2019). Another emerging feature is providing and managing extra services at parking bays needed by advanced electric/hybrid vehicles, such as charging their batteries. Finally, scheduling the use of shared autonomous vehicles and dispatching them to customers is a potential feature needed in future PIS.

Benefits and Advantages of PIS

Hassoune et al. (2016) provides a summary of advantages and disadvantages of multiple smart parking systems, given their underlying technologies and main PIS features. Overall, PIS enable more efficient use of parking space, reduces cruising for parking, and enhances user experience (Milosavljevic and Simicevic, 2019; Roman et al., 2018). This is especially relevant in situations where parking capacity is used unevenly across time and space. The unbalanced parking utilization is often due to incomplete information about alternatives and real-time/future availability. Balanced parking utilization optimizes the usage of existing capacity and reduces the overall cost of searching for parking for all customers (Chai et al., 2019).

Improved Traffic Efficiency

Previous research has shown that pretrip information about parking significantly impacts on users’ decision-making and results in more efficient use of parking and traffic network (Bock et al., 2019; El-Seoud et al., 2016; Tasserou, 2017). Providing such information, combined with navigation features, PIS can save customers’ time, while reducing their vehicle mileage and depreciation (Yang and Lam, 2019). Furthermore, pretrip knowledge about high level of occupancy and difficulty of finding parking close to a user’s destination may even motivate them to use a more sustainable mode of transport such as public transport or cycling instead of car (Milosavljevic and Simicevic, 2019).

Enhanced Customer Satisfaction

Availability of the reservation feature in a parking information system can also enhance user satisfaction, while enabling parking operators to increase their revenue and to optimize their capacity utilization (El-Seoud et al., 2016; Yang and Lam, 2019). Automated and contactless payment along with connected entry/exit controls also reduce congestion around parking facilities and improve user experience (Yang et al., 2003). PIS can also reduce cruising in off-street parking by navigating users to vacant parking, which in turn decrease congestion and vehicle emissions and likelihood of incidents (Dalangauskas et al., 2016).

More Effective Parking Management

Finally, effective PIS can result in lower illegal parking rates (Roman et al., 2018). PIS can also enhance car safety in off-street facilities through integrated safety and enhanced monitoring features (Hassoune et al., 2016).

Acknowledgment

This research is funded by iMOVE CRC (project 3-002) and supported by the Cooperative Research Centres program, an Australian Government initiative.

Biography



Dr. Behrang Assemi is an iMove postdoctoral research fellow at Queensland University of Technology, Australia, researching parking utilization, choice, and related congestion. He is a lead data scientist with 18 years of experience in both industry and academia. He has published papers in leading journals, for example, *Transportation Research Part A*, *Transportation Research Part C*, and *IEEE Transactions on Intelligent Transportation Systems*.



Dr. Douglas Baker is Professor in the School of Built Environment at the Queensland University of Technology, Brisbane, Australia. Dr. Baker's areas of expertise include land use planning, performance-based measurement, airport management, transit-oriented development, and regional governance. He has worked in the public and private sectors and conducts research in applied infrastructure and logistics solutions.

References

- Al-Turjman, F., Malekloo, A., 2019. Smart parking in IoT-enabled cities: A survey. *Sustain. Cities Soc.* 49, 101608, doi:10.1016/j.scs.2019.101608.
- Bock, F., Di Martino, S., Sester, M., 2019. What Is the Impact of On-street Parking Information for Drivers? SpringerLink, in: *Web and Wireless Geographical Information Systems, Lecture Notes in Computer Science*. Springer Nature, Switzerland, pp. 75–84.
- Chai, H., Ma, R., Zhang, H.M., 2019. Search for parking: A dynamic parking and route guidance system for efficient parking and traffic management. *J. Intel. Transp. Sys.* 23, 541–556, doi:10.1080/15472450.2018.1488218.
- Chen, M., Chang, T., 2011. A parking guidance and information system based on wireless sensor network, in: 2011 IEEE International Conference on Information and Automation. Presented at the 2011 IEEE International Conference on Information and Automation, pp. 601–605. <https://doi.org/10.1109/ICINFA.2011.5949065>.
- Chen, X., Qian, Z. (Sean), Rajagopal, R., Stiers, T., Flores, C., Kavalier, R., Floyd Williams, I.I.I., 2016. Parking Sensing and Information System: Sensors, Deployment, and Evaluation: *Transportation Research Record*. <https://doi.org/10.3141/2559-10>.
- Dalangauskas, G., Džiaugys, V., Adomaitis, D., Šajev, I., 2016. Car parking information system, in: 2016 Open Conference of Electrical, Electronic and Information Sciences (EStream). Presented at the 2016 Open Conference of Electrical, Electronic and Information Sciences (eStream), pp. 1–4. Available from: <https://doi.org/10.1109/eStream39242.2016.7485927>.
- El-Seoud, S.A., El-Sofany, H., Taj-Eddine, I., 2016. Towards the development of smart parking system using mobile and web technologies, in: 2016 International Conference on Interactive Mobile Communication, Technologies and Learning (IMCL). Presented at the 2016 International Conference on Interactive Mobile Communication, Technologies and Learning (IMCL), pp. 10–16. Available from: <https://doi.org/10.1109/IMCTL.2016.7753762>.
- Fang, J., Chang, Z.-Q., Cheng, Y.-J., Cai, M., Li, J., Chen, Z.-J., 2016. Design of Intelligent Parking System Based on Cloud Platform. Presented at the 3rd International Conference on Wireless Communication and Sensor Networks (WCSN 2016), Atlantis Press, pp. 576–579. Available from: <https://doi.org/10.2991/icwscn-16.2017.116>.
- Hassoune, K., Dachry, W., Moutaouakkil, F., Medromi, H., 2016. Smart parking systems: A survey, in: 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA). Presented at the 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA), pp. 1–6. Available from: <https://doi.org/10.1109/SITA.2016.7772297>.
- Hui-ling, Z., Jian-min, X., Yu, T., Yu-cong, H., Ji-feng, S., 2003. The research of parking guidance and information system based on dedicated short range communication, in: *Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*. Presented at the Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems, pp. 1183–1186. Available from: <https://doi.org/10.1109/ITSC.2003.1252671>.
- International Parking Institute, Fernandez, K., Yoka, R. (Eds.), 2018. *A Guide to Parking*, First. ed. Routledge, Taylor & Francis Group, New York, NY, USA.
- Kim, N., Jing, C., Zhou, B., Kim, Y., 2014. Smart parking information system exploiting visible light communication. *Int. J. Smart Home* 8, 251–260.
- Kotb, A.O., Shen, Y., Huang, Y., 2017. Smart parking guidance monitoring and reservations: A review. *IEEE Int. Transp. Sys. Mag.* 9, 6–16, doi:10.1109/MITS.2017.2666586.
- Lee, C., Han, Y., Jeon, S., Seo, D., Jung, I., 2016. Smart parking system for Internet of Things, in: 2016 IEEE International Conference on Consumer Electronics (ICCE). Presented at the 2016 IEEE International Conference on Consumer Electronics (ICCE), pp. 263–264. Available from: <https://doi.org/10.1109/ICCE.2016.7430607>.

- Li, Y., 2019. Research on Parking Guidance Information System Based on "Internet Plus". In: Abawajy, J., Choo, K.-K.R., Islam, R., Xu, Z., Atiquzzaman, M. (Eds.), *International Conference on Applications and Techniques in Cyber Security and Intelligence ATCI 2018 Advances in Intelligent Systems and Computing*, Springer International Publishing, Cham, pp. 452–461, doi:10.1007/978-3-319-98776-7_49.
- Lin, T., Rivano, H., Mouël, F.L., 2017. A survey of smart parking solutions. *IEEE Trans. Intel. Transp. Sys.* 18, 3229–3253, doi:10.1109/TITS.2017.2685143.
- Liu, Q., Lu, H., Zou, B., Li, Q., 2006. Design and Development of Parking Guidance Information System Based on Web and GIS Technology, in: 2006 6th International Conference on ITS Telecommunications. Presented at the 2006 6th International Conference on ITS Telecommunications, pp. 1263–1266. Available from: <https://doi.org/10.1109/ITST.2006.288857>.
- Milosavljevic, N., Simicevic, J., 2019. *Sustainable Parking Management: Practices, Policies, and Metrics*. Elsevier, Amsterdam, Netherlands.
- Ni, X.-Y., Cao, L.-Y., 2017. Current Situations and Development Trends of Urban Parking Guidance and Information System. Presented at the 3rd Annual 2017 International Conference on Management Science and Engineering (MSE 2017), Atlantis Press, pp. 142–144. <https://doi.org/10.2991/mse-17.2017.36>.
- Paidi, V., Fleyeh, H., Håkansson, J., Nyberg, R.G., 2018. Smart parking sensors, technologies and applications for open parking lots: A review. *IET Intel. Transp. Sys.* 12, 735–741, doi:10.1049/iet-its.2017.0406.
- Rajabioun, T., Foster, B., Ioannou, P., 2013. Intelligent parking assist, in: 21st Mediterranean Conference on Control and Automation. Presented at the 21st Mediterranean Conference on Control and Automation, pp. 1156–1161. Available from: <https://doi.org/10.1109/MED.2013.6608866>.
- Roman, C., Liao, R., Ball, P., Ou, S., de Heaver, M., 2018. Detecting on-street parking spaces in smart cities: Performance evaluation of fixed and mobile sensing systems. *IEEE Trans. Intel. Transp. Sys.* 19, 2234–2245, doi:10.1109/TITS.2018.2804169.
- Sadhukhan, P., 2017. An IoT-based E-parking system for smart cities, in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). Presented at the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1062–1066. <https://doi.org/10.1109/ICACCI.2017.8125982>.
- Shiue, Y.-C., Lin, J., Chen, S.-C., 2010. A study of geographic information system combining with gps and 3g for parking guidance and information system. *City* 65, 9–15.
- Sun, D.J., Ni, X.-Y., Zhang, L.-H., 2016. A discriminated release strategy for parking variable message sign display problem using agent-based simulation. *IEEE Trans. Intel. Transp. Sys.* 17, 38–47, doi:10.1109/TITS.2015.2445929.
- Tasserou, G., 2017. *Urban parking information provision: an in-depth effect analysis* (PhD Thesis). Radboud University Nijmegen, Delft, The Netherlands.
- Teng, H., Falcocchio, J.C., Lapp, F., Price, G.A., Prassas, S., Kolsal, A., 2001. Parking information and technology for a parking information system. Presented at the Transportation Research Board (TRB) 80th Annual Meeting, Washington D.C., USA.
- Teng, H.(Harry), Qi, Y.(Grace), Martinelli, D.R., 2008. Parking difficulty and parking information system technologies and costs. *J. Adv. Transp.* 42, 151–178, doi:10.1002/atr.5670420204.
- Wang, H., He, W., 2011. A Reservation-based Smart Parking System - IEEE Conference Publication, in: 2011 IEEE Conference on Computer Communications Workshops. Presented at the IEEE Conference on Computer Communications Workshops, IEEE, Shanghai, China, pp. 690–695.
- Yang, W., Lam, P.T.I., 2019. Evaluation of drivers' benefits accruing from an intelligent parking information system. *J. Clean. Prod.* 231, 783–793, doi:10.1016/j.jclepro.2019.05.247.
- Yang, Z., Liu, H., Wang, X., 2003. The research on the key technologies for improving efficiency of parking guidance system, in: The 2003 IEEE International Conference on Intelligent Transportation. Presented at the IEEE International Conference on Intelligent Transportation, IEEE, Shanghai, China, pp. 1177–1182.

Further Reading

- Books and book chapters:

Parking guidance and information system (Chapter 11): N. Milosavljevic and J. Simicevic, *Sustainable Parking Management: Practices, Policies, and Metrics*. Amsterdam, Netherlands: Elsevier, 2019.

Technology (Chapter 5): M. Drow and B. Laufer. International Parking Institute, K. Fernandez, and R. Yoka, Eds., *A Guide to Parking*, First. New York, NY, USA: Routledge, Taylor & Francis Group, 2018.

- Review articles:

Smart parking in IoT-enabled cities: A survey: F. Al-Turjman and A. Malekloo, *Sustain. Cities Soc.*, vol. 49, p. 101608, Aug. 2019, doi: 10.1016/j.scs.2019.101608

A survey of smart parking solutions: Lin, T., Rivano, H., Mouël, F.L., 2017. *IEEE Transactions on Intelligent Transportation Systems* 18, 3229–3253. doi: 10.1109/TITS.2017.2685143

- Standards:

Alliance for Parking Data Standards: Standard for the Data Domains (Parking Place, Rate, Occupancy, Right, Session, and Observation), available on request at <https://www.allianceforparkingdatastandards.org/versions>

DATEx II – European Commission's ITS (including parking) Standards: <https://docs.datex2.eu/>

- Websites (news and information about commercial PIS):

International Parking and Mobility Institute: <https://www.parking-mobility.org/>

Parking Network: <https://www.parking-net.com/>

Performance-Based Parking Management

Douglas Baker, Behrang Assemi, School of Built Environment, Queensland University of Technology (QUT), Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	299
Literature Review	300
Performance-Based Planning	300
Parking Performance	301
A Framework for Performance-Based Parking Management	302
Land Use Performance Indicators	302
Building Performance Indicators	303
Transportation Performance Indicators	304
Conclusion	304
Acknowledgment	304
Biographies	305
References	305
Further Reading	306

Introduction

The paradox of parking and future sustainability lies in how we have tied parking to the automobile, which has traditionally been used for commuting — causing pollution, greenhouses gases, and congestion on roads. As of 2018, there were more than 1.2 billion vehicles in the world (Wessel and Yoka, 2018), mostly comprised of private cars; these cars are parked on an average 95% of the time (Morris, 2016). Cruising for parking, especially in busy urban areas with limited parking availability, adversely impacts on traffic congestion and exacerbates the resulting air pollution (Lee et al., 2017). Parking, as a use of space, is associated with the use of single occupancy vehicles and the impacts created from their use. Large cities around the world are struggling to make a balance between the demands for more parking and the capacity to supply parking, while minimizing the abovementioned adverse consequences (Cavaldón, 2014).

Parking is a critical transport resource in urban centers that services a wide range of clients and uses. Wessel and Yoka (2018) suggest that the relationship between parking and sustainability lies in parking as a land use, building type, and form of transportation. We argue that parking needs to be conceptualized as a link between mobility and transport, and that space for parking is changing in its role as a destination in urban centers.

Therefore, the management challenge becomes changing parking from a prescriptive land use, where land uses are prescribed by zoning, to a link in mobility that is controlled by performance-based standards. Traditionally, in most cities and councils, parking utilization is controlled by the zone in which it resides; thus, the priority use that determines parking is the type of zone. For example, in a residential zone, on-street parking is regulated for local residents and guests. This may vary from all-day parking to prescribed time limits (time of day, length of stay, permit, or meter) depending on the demand and surrounding density. In a commercial zone, parking is often allocated for short-term customer access. The point is that in prescriptive zoning there are multiple management strategies for on- and off-street parking designed for a specific land use type. However, as Elliott (2012, p. 2) notes, “. . . many current zoning systems adjust poorly to changed circumstances; and reflect and encourage poor systems of city governance.” A performance-based approach can address such shortcomings (Baker et al., 2006; Frew et al., 2016).

How should parking then be managed as a link in the transport chain, or be conceptualized as part of an individual's mobility? The argument forwarded here is that parking is a destination point and part of a mode sequence, where performance standards can be adapted to suit the required modes of transport. Furthermore, such standards can be defined in a way to support sustainable urban development (Cavaldón, 2014). However, a different approach is required to manage parking space that does not tie parking to stringent, nonflexible zoning standards. Currently, there is no coherent framework for measuring parking performance considering its role in the transport mode sequence. Moreover, little is known about how to implement parking performance standards as a management tool due to both a lack of empirical data and an overarching theoretical framework.

Accordingly, the focus of this chapter is to introduce performance standards as a means to manage parking in a more efficient method and as a part of the transport journey. Research on performance-based planning has not integrated parking as a land use, yet parking is one of the dominant land uses in the urban core. The main challenge is how we define and enforce the performance indicators used for parking. As the first stepping stone, we build on the current literature to propose a conceptual framework for performance-based parking management in this study. In this framework, we incorporate well-known, mostly economic measures of parking performance. This framework quantifies the performance of parking spaces, and thus it enables planners and parking operators to consider relevant measures and achieve management objectives. Such a framework can play a crucial role in developing

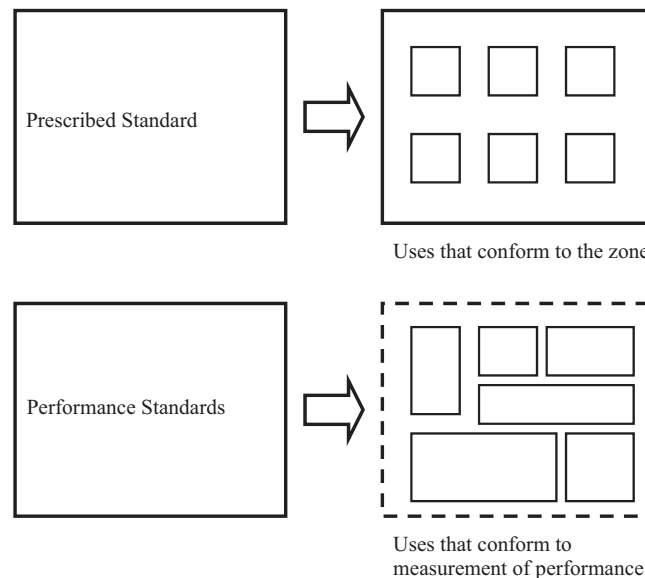


Fig. 1 Performance-based versus prescriptive planning.

actionable strategies for achieving long-term sustainability goals, especially as the majority of human beings will be urban dwellers by 2100 (West, 2017).

The remainder of this chapter is organized as follows. Section Literature Review reviews the literature on performance-based planning and parking performance measurement. Section A Framework for Performance-Based Parking Management presents the proposed the performance-based framework for parking management. Finally, Section Conclusion concludes this study.

Literature Review

Performance-Based Planning

Performance-based planning is increasingly applied to the public sector as a means to improve the efficiency and effectiveness of decision-making (Frew et al., 2016). Performance-based planning is built upon the assumption that the impact of land use is a function of intensity rather than specific use. Thus, performance-based approaches set standards that describe the desired end results and acceptable limits of impacts (Aigwi et al., 2019). This type of planning is supposed to be more flexible, require fewer regulations, speed up the approval process, and encourage a greater dialogue among stakeholders, as opposed to traditional prescriptive planning (Blumenauer, 2011). Fig. 1 provides an abstract representation of these two types of planning to highlight their main differences.

However, there are considerable challenges in applying performance-based planning to regulatory contexts (see Baker et al., 2006; Frew et al., 2016). According to Blumenauer (2011), there are five major considerations when developing and implementing a performance-based approach:

1. solutions should be comprehensive;
2. accountable partnerships should be formed;
3. explicit, precise outcomes should be defined and the corresponding performance indicators should be developed;
4. accountability should be encouraged to ensure the integrity of the process and the likelihood of achieving the outcomes; and
5. participation should be voluntary.

A successful implementation of performance-based planning is dependent on the application of indicators to measure performance standards (Vullo et al., 2018). More specifically, the desired outcomes within the plan are contingent on applying indicators at various levels to evaluate impacts and regulate activities. Indicators are used in a performance-based planning framework as a means to monitor baseline conditions to ensure that regulatory standards are being met. As a result, these indicators can be used to evaluate the relative success of the planning process against the predetermined planning objectives. In addition, the strategic and normative directions of the plan can constantly be measured using an indicator approach. Indicators can be vertically integrated to measure the planning process at different performance levels.

The application of indicators to land use planning is a complex and conceptually difficult problem. The primary difficulty arises from the evaluation of the planning process. What do we evaluate and how do we do it? Almost two decades ago, Briassoulis (2001) wrote on the complexity of measuring sustainable development in planning by using indicators: indicators are “a long way from making a substantial contribution to planning” due to their lack of integration, fragmented assessment of sustainable development,

and poor theoretical assumptions in the application of indicators; the dimensions of the problem incorporate basic questions such as “what do we measure?”, “how do we evaluate what we have measured?”, and “is it relevant to where we need to go?”. These are tough conceptual questions that relate to the evaluation of the performance of planning and are not easily answered. This continues to be the case today, although changing technology is allowing a broader application of indicators in planning and regulation.

Part of the answer to these questions about indicators lies in the definition of a model of planning and the application of an indicator framework to evaluate the planning process. The correlation of indicators to what they measure is essential. Indicators are a means of performance evaluation. They define: (1) what to measure; (2) how to measure; and (3) how to communicate measurement results.

Most often, indicators provide a method to measure output, efficiency, outcome, and/or productivity. An indicator “is something that points to an issue or condition” and “helps you understand where you are, which way you are going, and how far you are from where you want to be” (Hart, 1999, p. 26). Piasecki et al. (1999) suggest the purpose of any performance measurement in business is to change behavior. Thus, the main functions of indicators are to provide (1) an assessment of conditions or trends, (2) a comparison across places or situations, (3) an assessment in relation to goals or targets, (4) early warning information, and (5) an anticipation of future conditions or trends (Gallopín, 1997).

The literature on the use of indicators is extensive and spans a wide range of topics. Despite the differences in the application of indicators, there are criteria that are internal to the use of indicators common to all fields (see Alberti, 1996; Bossel, 1999; Gallopín, 1997; Hart, 1999; Milosavljevic and Simicevic, 2019). Generically, indicators should:

1. be appropriate to the context and reflect the system properties;
2. be easy to interpret and understand;
3. be analytically sound, incorporating reliability, and validity;
4. have a timeliness in order to respond;
5. be compatible within the system; and
6. be comparable outside the system.

Each of these characteristics defines a necessary criterion for the successful application of indicators as a means to measure performance and inform decision-making.

Finally, it is worth highlighting that there are two types of indicators. The first type consists of indicators used for measurement; these indicators set compliance levels or conformance to industry standards. This type of indicators is empirical and includes measurement in decibels, parts per million, pH levels, and other standards required by legislation or internal company standards. The second type of indicators measure the effectiveness of performance. Indicators of this nature evaluate internal programs, desired outcomes, productivity, and performance (Jasch, 2000). These indicators evaluate internal outputs such as “how are we doing inside the system?” and “are we achieving what we set out to achieve?”, as well as external outputs such as “how do we compare with similar entities?”. Both types of indicators have to be applied congruently in order to measure performance.

Parking Performance

Traditionally, the performance of parking has been measured and evaluated using localized survey-based indicators. Furthermore, given that mostly ad-hoc and uncoordinated procedures are used for parking management, parking use has usually been inefficient in many large cities (Willson, 2018). However, by introducing the Internet of Things (IoT) and innovative data collection and analytics technologies, a variety of tools and methods (sensors, pricing mechanisms, and real-time information systems) are becoming widely accessible, which enable more effective and accurate measurement and management of parking performance (Al-Turjman and Malekloo, 2019; Willson, 2018).

Within the context of parking planning and management, performance indicators can be defined and implemented to enable and facilitate achieving sustainable outcomes (Milosavljevic and Simicevic, 2019). This is especially relevant, as the most recent parking management practices consider parking as a service rather than an object, and thus the ultimate goal of such practices is to maximize parking service efficiency given both its financial and environmental aspects (Shoup, 2018).

As highlighted in the earlier section, careful development and effective implementation of performance indicators play a crucial role in the success of a performance-based planning framework for parking. Such indicators should enable a simple evaluation of compliance/conformance of parking with a predetermined financial standard, while providing the opportunity of assessing the effectiveness of parking performance toward higher sustainability (Milosavljevic and Simicevic, 2019). Parking performance, however, has often been evaluated through simple financial measures (Willson, 2018), such as the parking space occupancy over 24 hours, addressing the financial objective alone. It is important to define and implement parking sustainability indicators in order to develop an integrated framework for sustainable performance-based parking planning (Yoka, 2014). For example, the total number of parking spaces in the CBD of Portland, US, has been capped since the 1970s as a means to decrease private vehicle usage specifically to comply with an air quality maintenance plan focusing on carbon monoxide and ozone emissions (Blumenauer, 2011).

Every element of parking practice, from planning and designing the parking structures to searching for and booking a parking place, plays a crucial role in realizing potential management goals (Leinart, 2015). Air pollution is the most obvious impact of parking on the environment, caused by the emissions of vehicles traveling to parking facilities and cruising to find parking

Table 1 Parking performance measures

<i>Parking aspect</i>	<i>Measure</i>	<i>Description</i>
Land use	Land use cost-effectiveness	Level of efficiency in using the land dedicated to parking considering other potential uses of the land
	Turnover / Index	Turnover is the total number of vehicles parked in a parking bay during a given period of time Index is the percentage of time where a given parking bay is occupied
	Average duration	Average duration of parking utilization for all vehicles parked in a parking facility
	Land use mix	How mixed and/or multi-modal is the land
	Shared use	The extent to which the parking is shared for multiple uses (e.g., residential, office and entertainment)
Building	Land use effectiveness	How well the land is preserved, it provides community spaces and it promotes sustainable urban living
	Capacity	Efficiency in parking supply given the total available space
	Building cost-effectiveness	Cost-effectiveness of the parking considering its construction and operational costs as well as its revenue
	Overflow plan efficiency	Efficiency of overflow management plans for occasional extra parking demands
	Spill-over level	Average spill-over over 24 hours
Transportation	Availability	Space availability likelihood
	Walkability index	Range of destinations which can be serviced by the parking facility through walking
	Mobility index	Extent to which the parking facility enables mode change as a mobility hub
	Usage diversity index	Diversity of user types and/or destinations served by the parking
	Value usage ratio	Share of space dedicated to higher-value uses (e.g., service vehicles, deliveries, and people with special needs)
	Enforcement efficiency	The extent to which parking regulation enforcement is “efficient,” “considerate,” and “fair”
	Shared economy index	The extent to which the parking facility enables shared mobility services
	Alternative modes index	The extent to which alternative, more sustainable modes of transportation are supported

(Čuljković, 2018; Wessel and Yoka, 2018). Other environmental effects of parking include energy use, such as for lighting, although this is often not significant (Wessel and Yoka, 2018). Cruising for parking has also significant negative economic consequences, accounting for 8%–74% of total traffic in some areas and averaging between 3.5 and 14 minutes in each occasion (Lee et al., 2017).

Parking as a mobility service relies on a balanced combination of solutions and practices which address economic, public health, and environmental sustainability. The ultimate goal of such solutions and practices is to minimize fossil fuel use, vehicle emissions, and unsustainable land use (Geneletti et al., 2017; International Parking Institute, 2017).

Wessel and Yoka (2018) argue that the relationship between parking and sustainability can be investigated considering parking as a “land use,” “building type,” and “transportation” hub. From the land use perspective, valuable, sometimes very scarce, land resources are used to build parking facilities, whereas these resources could be used for a wide range of other purposes (Wessel and Yoka, 2018). Furthermore, the use of advanced materials and technologies in parking structures can significantly reduce the energy consumption in parking, while simultaneously enabling more efficient methods of finding parking with less driving (Yoka, 2014). Finally, Leinart (2015) argues that parking must be considered as a fully integrated element of the transportation system, which encourages active and sustainable modes of transport instead of single-occupancy vehicles (SOVs). This view or approach can lead to higher sustainability goals.

A Framework for Performance-Based Parking Management

Table 1 provides an overview of a proposed performance-based framework for sustainable parking management. As outlined in Table 1, we propose performance measures across the three major aspects of parking as: (1) a land use, (2) a building, and (3) an element of transportation (Wessel and Yoka, 2018). We have included common financial parking performance measures in the framework as defined and used in the literature (e.g., Litman, 2006; Milosavljevic and Simicevic, 2019; Shoup, 2011). The framework also includes sustainability performance measures across the three aspects of parking proposed in this study, and thus it is a multi-criteria performance-based framework (Vullo et al., 2018). The next subsections elaborate on the proposed framework and its performance measures.

Land Use Performance Indicators

We recommend land use cost-effectiveness as the main land use performance indicator of parking from a financial perspective. This indicator concerns how efficiently the land dedicated to parking is being used, considering all other potential uses of the land. The land dedicated either to on-street or off-street parking in dense urban areas is scarce and a valuable public resource in most cities. Thus, it is important to measure how cost-effective is to use it for parking rather than any other alternative (Willson, 2018).

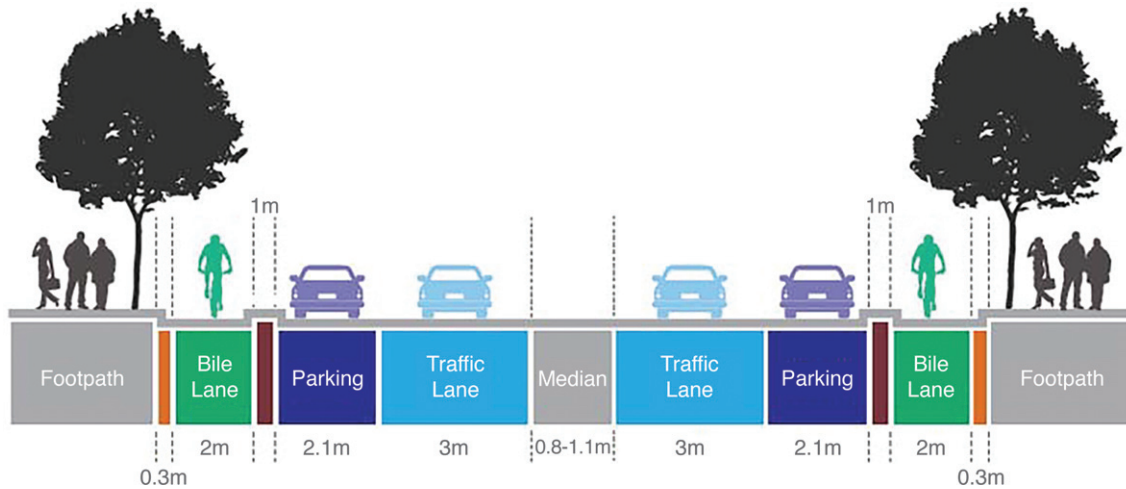


Fig. 2 Moray street bike path design as a part of the Victoria's government "Metro Tunnel" project (image courtesy of the State Government of Victoria, 2018).

The literature has also used parking turnover and/or parking index as other indicator(s) of the efficiency of parking space utilization (K. Al-Obaidi et al., 2018; Milosavljevic and Simicevic, 2019). Parking turnover considers the use of both space and time; it is the total number of vehicles parked in a parking bay during a given period of time. Thus, the higher turnover indicates a better land use performance, because a larger number of customers have used the parking during a specific period.

The parking index, on the other hand, is the percentage of time where a given parking bay is occupied (K. Al-Obaidi et al., 2018; Mathew, 2014). While it also indicates the extent to which the parking land is used by customers, it does not consider the shared utilization of the land by "different" customers. Therefore, it is a weaker indicator of efficient land use, compared to parking turnover.

Land use mix is a common indicator of a parking lot's performance from a management perspective. This indicator reflects the extent to which a parking facility is used for purposes other than its primary objective (parking). A higher land use mix often indicates multi-modal transit-oriented development of the parking facility, which integrates a balanced mixture of services and facilities to minimize the need for driving. A high land use mix can promote the idea of "park once and walk" to multiple destinations, enhance place making, and better integrate the community (Timothy Haahs and Associates, 2013). Shared use ratio is a similar indicator, which has been used in the literature; it indicates how much the parking space is shared for different uses, including residential, office, and entertainment (Timothy Haahs and Associates, 2013).

Finally, average parking duration, as the average duration of parking utilization for all vehicles parked in a parking facility, is another common indicator of land use performance (K. Al-Obaidi et al., 2018; Mathew, 2014). City planners often prefer shorter average parking duration for on-street and longer average parking duration for off-street parking facilities, supporting a higher shared utilization of often scarce on-street parking.

Land use effectiveness can be used as an indicator for the land dedicated to parking. However, this indicator can be expanded to other uses within the context of urban living (Timothy Haahs and Associates, 2013). The land used for parking need not be exclusive and can be integrated into other uses. Fig. 2 provides a cross-sectional example of parking land use, which effectively enhances the bike lanes' safety and thus promotes more sustainable modes of transport.

Building Performance Indicators

Building performance indicators are concerned with how efficient and effective a parking facility (either on-street or off-street) can be used. Parking capacity provides an indicator which corresponds to the total capacity of a parking facility. It becomes an effective performance measure if used as a relative indicator: the ratio of the actual capacity of a facility to the maximum potential capacity of the same facility if it was designed in a different way. Such relative indicators demonstrate the extent to which the parking space is wasted due to design and technological issues; while it could be used more effectively through a smarter design, more compact stalls and use of advanced car stackers.

Another indicator of building performance is its cost-effectiveness. This indicator reflects the trade-off between a parking facility's revenue and its construction and operation costs, regardless of being on-street or off-street. It shows how well the parking costs are managed through the use of advanced building material and parking operation technologies (for example, solar-powered lighting, and access control). It is also indicative of how well the corresponding parking price strategies work in practice.

Overflow plan efficiency is a building performance indicator, relevant for off-street parking facilities only. This indicator reflects how the overflow management plan and its supporting practical elements (real-time guidance information, signs, and paths)

perform in efficiently handling occasional extra-parking demands. An efficient overflow plan results in smoother traffic in and around the parking facility, enhances the customer experience, and impedes potential negative effects of mismanaged overflow traffic on other users of the surrounding area. Parking spill-over, measured as the average spill-over during a specific period of time, is a similar indicator used in the literature.

Transportation Performance Indicators

As a key element in the transport system, parking performance has often been examined through a wide range of transportation performance indicators. A parking availability index is an example of an indicator which corresponds to the probability of finding a parking space in a specific area at a specific time (Milosavljevic and Simicevic, 2019). Parking search/cruise time is a similar measure used in the literature (e.g., Milosavljevic and Simicevic, 2019), indicating the average time spent by drivers to find an available parking space in a specific area at a specific time. Accordingly, this indicator also reflects parking availability and demand; however, it is also affected by overall status of the transportation system other than parking (traffic flow) in a particular situation of interest.

The walkability index is another transportation performance indicator of a parking facility, reflecting the extent of destinations that are easily accessible by walking from the parking facility. It plays a central role in the emerging trend of park once and walk facilities, which help reducing motorized trips (Timothy Haahs and Associates, 2013). The average/shortest walking distance/time from a parking facility to a user's final destination is a similar performance indicator used in the literature (Milosavljevic and Simicevic, 2019).

A mobility index indicates how well a parking facility acts as a mobility hub by enabling mode change for its users. A parking facility with a high mobility index often provides easy access to bike-sharing facilities and/or bicycle storage areas. Furthermore, such parking also facilitates the use of public transit to multiple destinations, and thus it encourages more efficient and sustainable travel patterns. Therefore, parking with a higher walkability and/or mobility index promotes environmentally friendly transportation, reduces private car usage and congestion, and improves air quality.

The usage diversity index is similar to the mobility index; however, it is mainly concerned about the diversity of a parking facility's users as well as their destinations (Milosavljevic and Simicevic, 2019). The value usage ratio reflects the share of space dedicated to higher value usage (e.g., service vehicles, deliveries, and people with special needs). Therefore, it is an important performance indicator of parking as a common, shared resource, and identifies how well the whole community is served by the facility (rather than just private car owners).

Enforcement efficiency is an indicator of parking performance demonstrating how successful the enforcement policies and practices are in realizing the predetermined objectives of a particular parking planning scheme from the transportation perspective (Milosavljevic and Simicevic, 2019). This indicator is also concerned about parking users' perceptions of fairness in policy implementation.

A shared economy index is another indicator of a parking's environmental transportation performance that reflects the extent to which the parking facility enables shared mobility services. This includes the availability of car- and bike-sharing vehicles in the facility. Similarly, an alternative modes index determines the extent to which more sustainable modes of transportation are supported by a parking facility. This includes, but is not limited to, the availability of electric vehicle charging stations and electric scooter/bicycle docks (Timothy Haahs and Associates, 2013).

Conclusion

Performance-based management offers a different view of urban space dedicated to parking and provides a more flexible approach to parking. The benefit of this method for management is to consider parking as a service rather than an object, and thus one of the goals is to maximize parking service efficiency. In addition, by incorporating a wide range of indicators, a performance-based framework can integrate wider stakeholder engagement outside of parking specialists and private car users.

The primary advantage of a performance-based approach is that it is adaptable — with a wide range of financial and sustainability indicators. The indicators have the opportunity to be objective and service-oriented (for the client), and the focus is outcome-based. We argue that this offers a new direction in which to view land use for parking and to change the focus of space for cars to space for mobility opportunities. This framework also enhances cost efficiency for all stakeholders, including developers, parking lot owners and users. An indicator-based approach helps to adapt existing parking facilities to the needs of smarter, more sustainable cities (Michalec et al., 2019).

Acknowledgment

This research is funded by iMOVE CRC (project 3-002) and supported by the Cooperative Research Centres program, an Australian Government initiative.

Biographies



Dr. Douglas Baker is Professor in the School of Built Environment at the Queensland University of Technology, Brisbane, Australia. Dr. Baker's areas of expertise include land use planning, performance-based measurement, airport management, transit-oriented development, and regional governance. He has worked in the public and private sectors and conducts research in applied infrastructure and logistics solutions.



Dr. Behrang Assemi is an iMove postdoctoral research fellow at Queensland University of Technology, Australia, researching parking utilization, choice, and related congestion. He is a lead data scientist with 18 years of experience in both industry and academia. He has published papers in leading journals, namely, Transportation Research Part A, Transportation Research Part C, and IEEE Transactions on Intelligent Transportation Systems.

References

- Aigwi, I.E., Egbelakin, T., Ingham, J., Phipps, R., Rotimi, J., Filippova, O., 2019. A performance-based framework to prioritise underutilised historical buildings for adaptive reuse interventions in New Zealand. *Sustain. Cities Soc.* 48, 101547.
- Alberti, M., 1996. Measuring urban sustainability. *Environ. Impact Assess. Rev.* 16 (4), 381–424.
- K. Al-Obaidi, M., M. Ahmed, A., N. Aboud, S., M. Khalaf, A.-H., W. Ibrahim, I., and A. Abdullah, B., 2018. Analysis of Parking Performance of Public Off-Street Parks in Baghdad City. *Int. J. Adv. Sci. Res. Eng.* 4(7): 111–125. Available from: <https://doi.org/10.31695/IJASRE.2018.32806>
- Al-Turjman, F., Malekloo, A., 2019. Smart parking in IOT-enabled cities: A survey. *Sustain. Cities Soc.* 49, 101608.
- Baker, D.C., Sipe, N.G., Gleeson, B.J., 2006. Performance-based Planning: Perspectives from the United States, Australia, and New Zealand. *J. Plan. Edu. Res.* 254, 396–409.
- Blumenauer, E., 2011. Beyond the backlash: Using performance-based regulations to produce results through innovation. *J. Environ. Law Litigat.* (26), 351–366.
- Bossel, H., 1999. Indicators for Sustainable Development: Theory, Method, Applications, Winnipeg. International Institute for Sustainable Development Winnipeg, Canada.
- Briassoulis, H., 2001. Sustainable development and its indicators: Through a (planner's) glass darkly. *J. Environ. Plan. Manag.* 443, 409–427.
- Đuljković, V., 2018. Influence of parking price on reducing energy consumption and CO₂ emissions. *Sustain. Cities Soc.* 41, 706–710.
- Frew, T., Baker, D., Donehue, P., 2016. Performance based planning in Queensland: A case of unintended plan-making outcomes. *Land Use Policy* 50, 239–251.
- Gallop, G.C., 1997. Indicators and Their Use: Information for Decision-Making. In: Moldran, B., Billharz, S., Matravars, R. (Eds.), *Sustainability Indicators: A Report on the Project on Indicators of Sustainable Development*. John Wiley and Son Ltd, Chichester, U.K., pp. 69–83.
- Gavaldón, A.N. S. 2014. Less parking, more city: Case study in Mexico city, México City, México: Institute for Transportation and Development Policy México.
- Geneletti, D., La Rosa, D., Spyra, M., Cortinovis, C., 2017. A review of approaches and challenges for sustainable planning in urban peripheries. *Landscape Urban Plan.* (165), 231–243.
- Hart, M. 1999. Guide to Sustainable Community Indicators, Second ed., North Andover, U.S.: Hart Environmental Data.
- International Parking Institute. 2017. International Parking Institute's Framework on Sustainability for Parking Design, Management, and Operations, International Parking Institute. Available from: <https://www.parking.org/wp-content/uploads/2017/04/2017-IPI-Sustainability-Framework.pdf>.
- Jasch, C., 2000. Environmental performance evaluation and indicators. *J. Clean. Prod.* 81, 79–88.
- Lee, J. (Brian), Agdas, D., Baker, D., 2017. Cruising for parking: New empirical evidence and influential factors on cruising time. *J. Transp. Land Use* (101).
- Leinart, M., 2015. Sustainability and parking: What's (really) in it for you? *Park. Prof. Mag.* 36–39.
- Litman, T., 2006. *Parking Management Best Practices*. Routledge, New York, NY, USA.
- Mathew, T.V., 2014. *Parking Studies*. In: *Transportation Systems Engineering*, Bombay, India: IIT
- Michalec, A. (Ola), Hayes, E., and Longhurst, J., 2019. Building Smart Cities, the Just Way. A Critical Review of 'Smart' and 'Just' Initiatives in Bristol, UK. *Sustain. Cities Soc.* 47: 101510. Available from: <https://doi.org/10.1016/j.scs.2019.101510>
- Milosavljevic, N., Simicevic, J., 2019. *Sustainable Parking Management: Practices, Policies, and Metrics*. Elsevier, Amsterdam, Netherlands.
- Morris, D.Z., 2016. Today's Cars Are Parked 95% of the Time, *Fortune*, Available from: <http://fortune.com/2016/03/13/cars-parked-95-percent-of-time/>.
- Piasecki, B., Fletcher, K.A., and Mendelson, F.J., 1999. *Environmental Management and Business Strategy: Leadership Skills for the 21st Century*, Wiley, New York, USA
- Shoup, D.C., 2011. *The High Cost of Free Parking*, First ed. New York, USA: Routledge.
- Shoup, D.C., 2018. *Parking and the City*, First ed. New York; Routledge, Taylor & Francis Group.

- Timothy Haahs and Associates, Inc., 2013. From Obstacle to Opportunity: The Evolution of Sustainability and Parking, Timothy Haahs and Associates, Inc. Newsletter (13), pp. 1-5.
- Vullo, P., Passera, A., Lollini, R., Prada, A., Gasparella, A., 2018. Implementation of a multi-criteria and performance-based procurement procedure for energy retrofitting of facades during early design. *Sustain. Cities Soc.* 36, 363–377.
- Wessel, P., and Yoka, R. 2018. Sustainability, first ed. In: Fernandez, K. and Yoka, R. (Eds.), *A Guide to Parking*, International Parking Institute, New York, NY, USA: Routledge, Taylor & Francis Group, pp. 50-59. Available from: <https://www-taylorfrancis-com.ezp01.library.qut.edu.au/books/e/9780429947865>.
- West, G.B., 2017. *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies*. Weidenfeld & Nicolson, London, UK.
- Willson, R., 2018. "Parking management for smart growth, first ed. In: Shoup, D.C. (Ed.), *Parking and the City*, New York, NY, USA: Routledge, Taylor & Francis Group, pp. 222-227.
- Yoka, R. (Ed.), 2014. *Sustainable Parking Design and Management: A Practitioner's Handbook*, Alexandria, International Parking Institute, USA.

Further Reading

- Baker, D.C., Neil, G.S., Brendan, J.G., 2006. Performance-based planning. *J. Plan. Edu. Res.* 25 (4), 396–409.
- Elliott, D.L., 2012. *A Better Way to Zone: Ten Principles to Create More Livable Cities*. Island Press, Washington D.C., USA.
- Frew, T., Douglas, B., and Paul, D., 2016. "Performance Based Planning in Queensland: A Case of Unintended Plan-Making Outcomes." *Land Use Policy* 50:239-51.
- Milosavljevic, N., and Simicevic, J., 2019. *Sustainable Parking Management: Practices, Policies, and Metrics*, Amsterdam, Netherlands: Elsevier. Available from: <https://www.sciencedirect.com/book/9780128158005/sustainable-parking-management>.

Transit Priority

Wanjing Ma, Qiheng Lin, Ling Wang, The Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai, P. R. China

© 2021 Elsevier Ltd. All rights reserved.

Overview	307
Passive Transit Priority	308
Active Transit Priority	308
Real-Time Transit Priority	309
Multiple Priority Requests	310
Signal Coordination for Transit	310
Transit Priority Combined With Other Facilities	311
Summary and Other Issues	312
Acknowledgment	313
References	313
Further Reading	314

Overview

Transit priority, also known as the Transit Signal Priority (TSP), is a signal control strategy that grants priority to transit vehicles at signalized intersections. The purpose of transit priority is to assist transit vehicles in traveling through intersections quickly. Transit vehicles are given priorities by assigning green time to them so that they may travel through the intersection with a lower delay compared with vehicles of other movements (Lin et al., 2015).

Transit priority is a long-standing research topic that has been evolving for over half a century. The first transit priority was conducted in Los Angeles, California, in the United States in 1967 (Elias, 1974). The reason for introducing the transit priority was to maintain the competitiveness of transit service, which is mostly needed in populous urban areas (Baker et al., 2004). If the travel time of transit is short and stable, more residents will travel by transit instead of driving. This shift in travel mode helps to mitigate urban road congestion.

Transit priority improves the efficiency and reliability of transit service. By providing priorities to transit vehicles at intersections, transit priority helps transit vehicles reduce their waiting time. Therefore, they have shorter travel time along the route. If the transit priority is not provided, congestion at intersections may hold transit vehicles at intersections, preventing them from following the schedule by increasing travel time between stops and thereforereducing the punctuality of transit service. However, if the transit priority is granted, transit vehicles may not experience unpredicted intersection congestions, thereby improving punctuality and make transit service more reliable to passengers (Baker et al., 2004).

However, granting priority to a particular movement (i.e., through, left-turn, right-turn, U-turn, etc.) at intersections may adversely impact the efficiency of the nonprioritized traffic (Balke et al., 2000; Lin et al., 2015; Ma et al., 2010). Therefore, transit priority should only be granted when it is necessary and beneficial. As sensor and communication technologies evolve, transit priority strategies are becoming more responsive to real-time transit demand and more flexible in providing priority in order to obtain a better balance between transit and non-prioritized traffic. These strategies can be classified into the following three categories (Baker et al., 2004).

- *Passive transit priority*: Passive transit priority only optimizes signal timing of intersections based on given transit information and generates a static signal control plan. Most research on passive transit priority were conducted from 1967 to the early 1990s.
- *Active transit priority*: Active transit priority entails detecting transit vehicles' travel demand and granting priority to these vehicles using one or a combination of basic strategies. This category of transit priority strategies was extensively researched from 1967 to the early 1990s.
- *Real-time transit priority*: Real-time transit priority optimizes signal timing of intersections for a specific performance index based on real-time information of both transit and nonprioritized vehicles. Research on real-time transit priority began in the early 1990s, partly replacing active transit priority.

The aim of this article is to present the basic concept of transit priority and analyse its strategies. The details of each category of transit priority strategies are outlined in separate sections. Then, the article addresses the handling of multiple priority requests, the signal coordination for transit, and the combination of transit priority with other facilities in the following sections. A summary and future directions for transit priority are provided in the last section.

Passive Transit Priority

Passive transit priority strategy optimizes signal timing of intersections based on given transit information in an off-line mode and generates a static signal control plan. This priority strategy does not require the detection of approaching transit vehicles (Skabardonis, 2000). All transit-related information, including the distribution of bus arrival time and dwelling time at bus stops, is input for optimization.

There are several approaches to modify the original signal control plan in order to achieve transit priority (Baker et al., 2004; Skabardonis, 2000). The first approach is to adjust cycle lengths and signal timings. Reducing cycle length may reduce the waiting time of transit vehicles at red signals and therefore reduce transit delays. As cycle length changes, the length of each phase should change accordingly. However, reducing the cycle length may negatively impact the overall capacity and vehicle delay at the intersection. On the other hand, reducing the number of phases may reduce delay for all traffic at the intersection, including transit vehicles, since it saves total lost time.

The second approach is to split the phases. By splitting and repeating the phase corresponding to the prioritized movement, transit vehicles with different arrival times may encounter shorter red times and experience lower delays. However, the frequent phase transitions may increase the total lost time and reduce the overall capacity of the intersection. In addition, the corresponding phase should not be over-split since pedestrians need a certain period of green time to cross the road safely.

Passive transit priority is suitable for intersections with high and stable transit demand. Take the Bus Rapid Transit (BRT) system in Guangzhou, China as an example (Deng and Nelson, 2011; Wirasinghe et al., 2013). This is a busy system with over 30 transit routes operating on the same road. The travel speed of transit vehicles in Guangzhou BRT has been improved from 15 km/h to 21 km/h during peak hours due to dedicated lanes to transit vehicles, a reduced number of phases, and shortened cycle lengths at signalized intersections (Lin et al., 2015). In London, UK, passive transit priority provided better performance than other transit priority strategies when bus volume is higher than 60 vehicles per hour (Hounsell and Wu, 1996). However, a field test on Queen Street, Toronto, Canada in the 1990s showed no evident improvement in transit delay since transit volume was not stable (Yagar, 1993).

The advantages of passive transit priority are distinctive. Passive transit priority is simple and inexpensive to implement since it does not require sensors to detect transit vehicles. The signal control plan produced by off-line optimization is compatible with existing signal controllers (Hounsell et al., 1996; Ma et al., 2019; Skabardonis, 2000). However, passive transit priority does not consider the randomness of transit arrival time, making the system unable to respond to the time change of the transit arrive. Moreover, if no transit vehicle arrives during the prioritized phases, the passive transit priority would still unnecessarily give the green time to transit vehicles, which causes delays for nonprioritized traffic.

Active Transit Priority

Active transit priority detects transit vehicles' travel demands then grants priority to these vehicles using one or a combination of basic strategies. Travel demands are detected by sensors and sent to signal controllers, then signal controllers respond (Baker et al., 2004; Lin et al., 2015).

Compared with passive transit priority, active transit priority can respond to real-time transit demand. It considers the arrival time of transit vehicles at intersections and grants priority only when transit demand is present, which partly remedies one shortcoming of passive transit priority, namely, the granting of priority when priority is not needed.

Active transit priority needs some extra devices to obtain real-time transit information. The sensor for detecting transit vehicles could be a loop detector on a dedicated bus lane, a video camera combined with visual recognition techniques, Global Positioning System (GPS), or a Road Side Unit (RSU) that receives RFID or other radio signals from transits (Fan and Gurmu, 2015; Zhou et al., 2006). The signal controllers should be able to communicate with these sensors and respond to real-time information (Lin et al., 2015).

Active transit priority uses one or a combination of basic strategies to grant priority, including green extension, red truncation, and phase insertion. If the current phase corresponds to the transit vehicle's movement, the green extension strategy extends the green time of current phase when detects an approaching transit vehicle. In this way, the transit vehicle may travel through the intersection without encountering red light of the next phase. If the current phase does not correspond to the transit vehicle's movement but the following phase is for transit vehicles, red truncation terminates the red light of current phase in advance and provides green time to approaching transit vehicle so that the vehicle may experience less or no waiting time at the intersection. The phase insertion strategy inserts a dedicated phase when an approaching transit vehicle is detected (Lin et al., 2015).

To reduce the adverse impact of active transit priority on nonprioritized traffic, some strategies are proposed (Lin et al., 2015). Green compensation strategy provides extra green time to nonprioritized traffic after the transit vehicle left the intersection. However, there is still an extra delay for nonprioritized traffic during the compensation cycle (Lin et al., 2015). Another strategy restricts transit priority under some conditions. The British Department for Transport recommended a rule: only allowing red truncation, if there was no red truncation in the last cycle. The purpose of this rule is to prevent transit priority strategies from interrupting nonprioritized traffic too frequently (Department of Transport, 1977). The framework of active transit priority is summarized in Fig. 1.

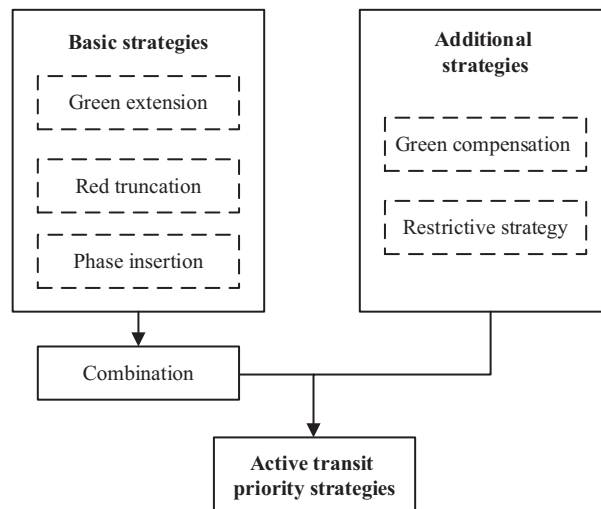


Figure 1 Framework of active transit priority.

Active transit priority strategies can be further classified into two categories, namely unconditional strategies and conditional strategies (Lin et al., 2015). Unconditional strategies grant priority to every approaching transit vehicle, but conditional strategies consider schedule adherence of each transit vehicle. For conditional strategies, if the approaching transit vehicle is behind schedule, priority shall be granted to help the vehicle to catch up. However, transit vehicles on schedule or ahead of schedule do not receive priority (Hu et al., 2015; Skabardonis, 2000). Meanwhile, conditional strategies may introduce additional rules that prevent basic strategies from disrupting arterial or network signal coordination.

Active transit priority is suitable for a lower volume of transit traffic compared with passive transit priority. Unconditional active transit priority is more compatible with isolated signalized intersections since extending, truncating, or inserting phases may disrupt coordinated signal control. However, conditional strategies may mitigate this problem by adding coordination constraints to control rules. Moreover, active transit priority can handle high fluctuation in dwelling time and arrival time at intersections (Lin et al., 2015).

Active transit priority is a more sophisticated transit priority strategy compared with passive transit priority. It shortens the delay of nonprioritized traffic because it only grants priority on real-time transit demand. Active transit priority may also involve compensation and restrictive strategies that prevent delay of non-prioritized traffic from radically increasing. However, the basic strategies employed by active transit priority may severely impact the nonprioritized traffic and possibly disrupt the arterial signal coordination.

Real-Time Transit Priority

Real-time transit priority is a model-based strategy that optimizes signal control for a specific performance index based on real-time information of transit and nonprioritized vehicles (Skabardonis, 2000). The frequently used performance indexes are average transit vehicle delay, nonprioritized traffic delay, and average person delay (Ma et al., 2014, 2013). This approach requires the integration of transit priority and signal optimization so that the signal control plan would respond to the traffic demands (Lin et al., 2015). Real-time transit priority may consider factors from both transit and nonprioritized traffic, seeking a better balance between them.

Real-time transit priority requires more advanced sensors and communication technologies than the previously mentioned strategies. It could utilize Automatic Vehicle Location (AVL) systems (Hounsell et al., 2000) equipped on transit vehicles or Connected Vehicle (CV) technologies (Hu et al., 2015, 2014; Miucic, 2019) to obtain precise real-time information on transit vehicle locations and speeds, facilitating the prediction of arrival time at intersections. Furthermore, real-time transit priority may obtain the number of onboard passengers on each transit vehicle to support the optimization that uses person delay as the objective (Ma et al., 2014). Information from sensors on General Purpose (GP) lanes could also be considered, depending on the objective of real-time transit priority.

If data on the state of non-prioritized traffic is not available, minimizing total bus delay or average bus delay is a reasonable objective to improve transit efficiency (Head et al., 2006). However, if there are enough sensors to provide reliable and real-time information on nonprioritized traffic, considering the increase of nonprioritized traffic delay when transit priority is granted would be more comprehensive. Furthermore, if passenger count information on transit and non-transit vehicles are available, the impact of transit priority on passenger delay of nonprioritized traffic could be used as the optimization objective (Hu et al., 2014; Ma et al., 2014). This impact is two-fold, depending on the movement of non-transit traffic at the intersection. The person delay of non-transit traffic travelling in prioritized movements decreases, but the person delay of non-transit traffic of non-prioritized movement

increases. On the other hand, a multi-objective optimization model is also possible, which transit passenger delay, nonprioritized traffic delay, and transit schedule adherence delay in a comprehensive manner (Chang et al., 1996; Ma et al., 2013). Overall, if enough information is available, real-time transit priority should aim at reducing transit passenger delay without disproportionately increasing passenger delay of non-prioritized traffic. The optimization objective should reflect the overall benefit to the performance of the intersection, covering both transit and non-transit traffic.

The advantages of real-time transit priority are evident. It covers as many factors as possible, including the current states of transit and nonprioritized traffic. It may further optimize priority at the passenger level, producing the most precise results of balancing transit and nonprioritized traffic performance. Additionally, the variety of optimization models ensures flexibility in determining how to grant transit priority. However, in order to provide the information needed for optimization, the investments in infrastructure will increase (Lin et al., 2015). The cost-effectiveness of the real-time transit priority strategies should be thoroughly investigated before implementation.

Multiple Priority Requests

When transit traffic volumes increase and transit routes overlap each other, an intersection may need to simultaneously handle multiple priority requests from different approaches and different movements at the same time. There may be conflicts between these requests. For example, transit priority strategy cannot grant priority to two transit vehicles at the same time if their movements at the intersection conflict with each other. Therefore, it is vital to accommodate multiple priority requests while developing transit priority strategy, especially for active transit priority and real-time priority, which grant priority on real-time transit demand (He et al., 2014; Head et al., 2006; Lin et al., 2015; Ma and Bai, 2007).

Priority requests sent to an intersection could be arranged into a request queue in the order of received time, and it is the task of transit priority strategy to determine determining the order in which to respond to the requests in the queue and which requests to ignore. Two methods to handle the queue of multiple priority requests are the first-come-first-served (FCFS) policy and trade-off the overall performance of transit and nonprioritized traffic delay (Lin et al., 2015). The FCFS policy is an intuitive way, which is convenient to include in rule-based active transit priority. However, multiple studies have proven that the FCFS policy may not yield the best performance. Furthermore, FCFS ignores the schedule adherence of each transit vehicle (Head et al., 2006; Lin et al., 2015).

The second method, that is, trade-off strategy, may solve these problems. The order in which priority requests are served is determined by the model-based real-time transit priority, which optimizes the sequence by trading-off the performance of both transit and nonprioritized traffic. Performance indexes could be person delay and schedule adherence of transit vehicles. Multiple models, including decision trees (Ma and Bai, 2007), operations research models (Head et al., 2006), and mixed-integer linear program (MILP) (He et al., 2014), have been investigated for handling multiple priority requests, and the results have confirmed the advantage of real-time transit priority in accommodating multiple priority requests.

Signal Coordination for Transit

Implementation of transit priority at isolated intersections is not enough to reduce the overall transit vehicle travel time along the transit route. The travel time saved at one prioritized intersection might be neutralized by the downstream intersection if the two intersections are not coordinated for transit vehicles (Hu et al., 2015; Skabardonis, 2000). Therefore, transit priority strategies at intersections along the transit route should be coordinated to achieve systematic efficiency improvement for transit service.

All of the three categories of transit priority strategies can coordinate signal control plans for transit vehicles. For passive transit priority, signal coordination for transit may be achieved by either manually modifying the original signal control plan based on the transit headway setting (i.e., the interval of time between two neighboring transit vehicles moving in the same direction on the same route) of the transit schedule or including transit priority into the coordination optimization process (Lin et al., 2015). For the latter approach, take the off-line signal timing optimization software TRAFFIC NETWORK STUDY TOOL-7F (TRANSYT-7F) as an example. This software optimizes signal control plans for arterials and networks. Transit movements are modelled as separate links and assigned greater weighing factors for delay and stops. Then, these factors are included in the objective of the optimization (Skabardonis, 2000). Similar to single intersection strategies, passive transit priority performs better when transit traffic volume is high and stable when signals are coordinated (Lin et al., 2015).

For active transit priority, the control rules should be carefully designed to avoid oversaturated non-prioritized movements or disruptions to signal coordination (Lin et al., 2015; Skabardonis, 2000). Schedule adherence could also be used as the constraints for granting priority similar to in the single intersection scenario (Hu et al., 2015, 2014). The logic of active transit priority has been integrated with online signal control systems such as the Sydney Coordinated Adaptive Traffic System (SCATS) and the Split Cycle Offset Optimization Routine (SCOOT) (Ma et al., 2013). A field test of transit priority with SCOOT in London, UK, in 1999 demonstrated that transit vehicle delay could be reduced by 5–10 seconds per intersection without evident adverse impacts on nonprioritized traffic (Hounsell et al., 2000).

Real-time transit achieves better performance for both transit and non-prioritized traffic with the aid of advanced technology such as AVL and CV technology. First, the group of intersections to be coordinated should be determined. One can group all intersections on an arterial together within a pre-set distance or travel time threshold (Hu et al., 2015, 2014), or all intersections

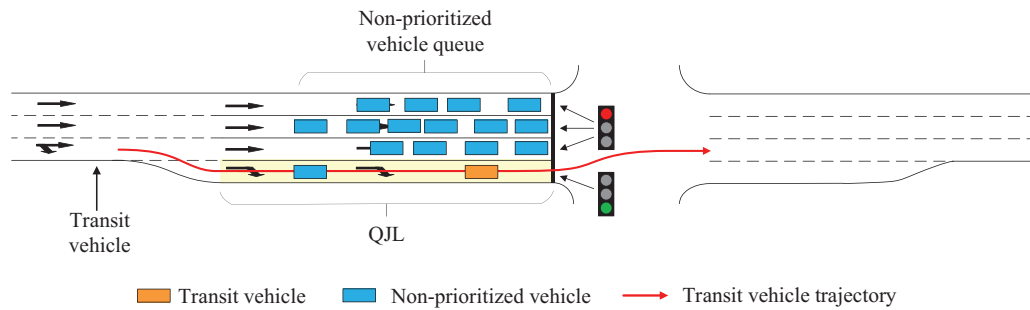


Figure 2 An illustration of QJL.

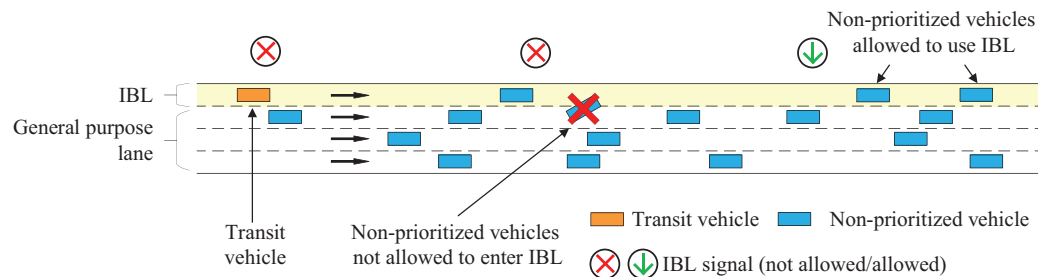


Figure 3 An illustration of IBL on urban road segment.

between two transit stops could be grouped together (Ma et al., 2013). Then, the planners use the optimization model to determine the signal control plan of all intersections within the group based on transit priority requests. The objective could be the minimization both transit delay and nonprioritized traffic delay increased due to transit priority. The model should be able to predict the arrival time of each transit vehicle at all intersections (Lin et al., 2015). Both analytical evaluation and simulation results showed that signal coordination under real-time transit strategy could improve transit efficiency and schedule adherence at a low cost of nonprioritized traffic efficiency.

Transit Priority Combined With Other Facilities

Transit priority could be combined with other facilities that give transit priority on urban roads to enhance the effect of transit priority (Truong et al., 2017). Four types of facilities investigated in this section are Dedicated Bus Lane (DBL), Queue Jump Lane (QJL), Intermittent Bus Lane (IBL), and Bus Lane with Intermittent Priority (BLIP).

A DBL is the most common facility that gives transit vehicle priority on the road segments. Since there are no disturbances from non-prioritized traffic, the speed of transit vehicles is stable, and therefore, transit arrival time estimation at each intersection becomes more precise, which further improves the accuracy of optimization (Deng and Nelson, 2011). DBLs have been adopted by many BRT systems. A major shortcoming of DBL is that it reduces the capacity for non-prioritized traffic on the road segments (Viegas and Lu, 2004, 2001). In addition, the spatial-temporal resource may be wasted when the transit demand is low. Therefore, DBLs only suit urban roads with high transit volumes (Lin et al., 2015).

In order to assign spatial-temporal resources to both transit and non-prioritized traffic more accurately, unconventional facilities such as QJL, IBL, and BLIP have been proposed. The basic idea of QJL is to utilize a short stretch of lane, such as the right-turn pocket lane, so that transit vehicles may bypass the queue on the approach of the intersection (Farid et al., 2015; Zhou et al., 2006). An illustration of QJL is presented in Fig. 2.

In a QJL, the transit vehicle travels through the intersection. A right turn is allowed on the red signal; therefore, there will not be a queue in the right-turn lane. A special phase called the "jumper phase" is inserted when an approaching transit vehicle is detected. During the jumper phase, the non-prioritized through traffic that travels in the same movement as the transit vehicle is stopped, but a right turn is still allowed. The transit vehicle then uses the right-turn lane as a QJL to bypass the queue of non-prioritized vehicles. It merges back in front of the non-prioritized vehicle queue after it passes the stop line. Analytical and simulation results have demonstrated a significant decrease in transit delays for systems combining transit priority with QJLs (Farid et al., 2015; Zhou et al., 2006).

An IBL allows non-prioritized vehicles to use IBL when no transit vehicles present (Viegas and Lu, 2004, 2001; Viegas et al., 2007). An illustration of IBL is shown in Fig. 3. When a transit vehicle appears, non-prioritized vehicles downstream of the transit vehicle are not allowed to enter IBL by changing lane, but those already inside IBL may continue to use it. IBL signals are set up along the

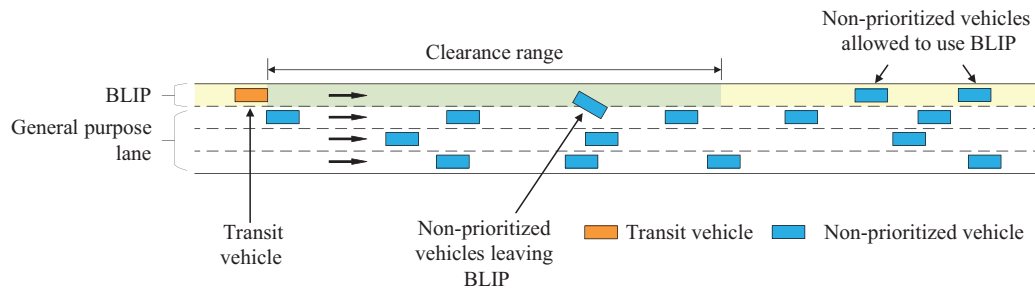


Figure 4 An illustration of BLIP on urban road segment.

arterial to indicate whether non-prioritized vehicles are allowed to enter IBL, which are integrated with conventional signal controllers (Viegas and Lu, 2004). When a transit vehicle is detected, the signal controller at the local intersection controls IBL signals and intersection signals at the same time (Viegas and Lu, 2001). In this way, priority is granted both on the approach of the intersection and at the intersection itself.

An illustration of BLIP is shown in Fig. 4. Similar to IBL, BLIP also allows non-prioritized vehicles to use BLIP lane when no transit vehicles present. However, when a transit vehicle approaches, non-prioritized vehicles downstream of the transit vehicle and within a certain distance (i.e., the Clearance Range in Fig. 4) of that transit vehicle are instructed to vacate the BLIP lane, either by signals or by connected vehicle technologies (Carey et al., 2009; Eichler, 2005; Eichler and Daganzo, 2006). The BLIP is also open to non-prioritized vehicles upstream of the transit vehicle. This facility is suitable for transit routes with headway around 10 to 15 minutes or more. The popularization of connected vehicle technologies would enhance the effect of BLIP in reducing transit delays.

Summary and Other Issues

This article has reviewed three categories of transit priority strategies: passive transit priority, active transit priority, and real-time transit priority. The evolution of sensors and communication technologies has been supporting the development of more sophisticated transit priority strategies. Two crucial issues regarding transit priority are multiple priority requests at an intersection and signal coordination for transit. Systems can also combine transit priority with other facilities in order to achieve better performance. The advantages, shortcomings, and applicable conditions for the three categories of transit priority strategies are summarized in Table 1.

The core idea of transit priority is to reduce transit delay at the expense of the efficiency of non-prioritized traffic at the intersection. In order to grant the priority for transit and also to reduce the adverse impacts on other traffic, transit priority strategies try to balance the performance of transit and non-prioritized traffic. For transit, the frequently used performance indexes are transit delay or travel time stability along the transit route. For the non-prioritized traffic, frequently used indexes are delay reduction of non-prioritized traffic in prioritized movements, delay gain of non-prioritized traffic in non-prioritized movements, and capacity for

Table 1 Summary of three categories of transit priority strategies

Category	Characteristics	Applicable conditions
Passive transit priority	• Easy implementation • Low investment in infrastructure • Irresponsive to real-time transit demand	• High and stable transit volumes
Active transit priority	• Requires detection of transit vehicles • Grants priority on transit demand • Potential disruption to signal coordination • Transit vehicles in need of priority may be ignored due to restrictive strategies	• Lower transit volume with fluctuations • Fluctuations in dwelling time and arrival time at intersections
Real-time transit priority	• Requires advanced sensors and communication technologies • High investment in infrastructure • Supports optimizations on the passenger level • Considers the overall performance of the intersection • High flexibility	• Lower transit volume with fluctuations • Fluctuations in dwelling time and arrival time at intersections

non-prioritized traffic at intersections or road segments. Other performance indexes measure the overall impact of transit priority, such as the reduction in average person delay and the person capacity.

Future transit priority strategies may exploit emerging connected and automated vehicle (CAV) technology to enable more precise control of transit and non-prioritized vehicles. A cooperative bus priority system in the CV environment had been implemented and tested in Taicang, China. In addition to transit priority, the system instructed transit vehicles to drive at a suggested speed. The field test showed promising results in the reduction of transit delay, stops, and average fuel consumption (Wang et al., 2014). In CAV environment, transit vehicles are not only connected but also automated, which makes them respond to speed suggestions in a more precise way (Hu et al., 2015, 2014; Wang et al., 2014). Future transit priority strategies may consider interestingly controlling both signal and vehicle trajectories with the help of CAV technology in order to achieve better efficiency and environmental gain.

Furthermore, transit vehicles are simultaneously impacted by transit schedules and traffic signals. Current transit priority strategies more focus on traffic signal control. However, optimizing transit schedules together with transit signal priority may further improve the efficiency of transit without adversely impact non-prioritized traffic (Zhou et al., 2019). Future transit priority strategies may also consider the integrated optimization of transit schedules and signal control plans at intersections.

Acknowledgment

The article is supported by the National Natural Science Foundation of China (No. 51722809) and the National Key Research and Development Program of China 2018YFB1601000).

References

- Baker, R.J., Collura, J., Dale, J.J., et al., 2004. An overview of transit signal priority. ITS America, Washing, D.C.
- Balke, K.N., Dudek, C.L., Urbanik, T., 2000. Development and evaluation of intelligent bus priority concept. *Transp. Res. Rec.: J. Transp. Res. Board* 1727, 12–19, doi:10.3141/1727-02.
- Carey, G., Bauer, T., Giese, K., 2009. Bus Lane with Intermittent Priority (BLIMP) Concept Simulation Analysis, Final Report (No. FTA-FL-26-7109.2009.8). US Department of Transportation.
- Chang, G., Vasudevan, M., Su, C., 1996. Modelling and evaluation of adaptive bus-preemption control with and without automatic vehicle location systems. *Transp. Res. A* 30, 251–268, doi:10.1016/0965-8564(95)00026-7.
- Deng, T., Nelson, J.D., 2011. Recent developments in bus rapid transit: a review of the literature. *Transp. Rev.* 31, 69–96, doi:10.1080/01441647.2010.492455.
- Department of Transport, 1977. Bus priority at traffic signals using selective detection, Technical Memorandum H2/77. HMSO, London, UK.
- Eichler, M., 2005. Bus lanes with intermittent priority: assessment and design. University of California, Berkeley, Berkeley, California, United States.
- Eichler, M., Daganzo, C.F., 2006. Bus lanes with intermittent priority: strategy formulae and an evaluation. *Transport. Res. B: Meth.* 40, 731–744, doi:10.1016/j.trb.2005.10.001.
- Elias, W.J., 1974. The greenback experiment: signal pre-emption for express buses: a demonstration project (No. DMT-014). California Department of Transportation, Sacramento.
- Fan, W., Gurmu, Z., 2015. Dynamic travel time prediction models for buses using only GPS data. *Int. J. Transp. Sci. Technol.* 4, 353–366, doi:10.1016/S2046-0430(16)30168-X.
- Farid, Y.Z., Christofa, E., Collura, J., 2015. Dedicated bus and queue jumper lanes at signalized intersections with nearside bus stops: person-based evaluation. *Transp. Res. Rec.: J. Transp. Res. Board* 2484, 182–192, doi:10.3141/2484-20.
- He, Q., Head, K.L., Ding, J., 2014. Multi-modal traffic signal control with priority, signal actuation and coordination. *Transp. Res. Part C: Emerg. Technol.* 46, 65–82, doi:10.1016/j.trc.2014.05.001.
- Head, L., Gettman, D., Wei, Z., 2006. Decision model for priority control of traffic signals. *Transp. Res. Rec.: J. Transp. Res. Board* 1978, 169–177, doi:10.1177/0361198106197800121.
- Hounsell, N.B., Wu, J.P., 1996. Public transport priority in real-time traffic control systems, in: *Applications of Advanced Technologies in Transportation Engineering*. Presented at the 4th International Conference, Capri, Italy, pp. 71–75.
- Hounsell, N.B., McLeod, F.N., Bretherton, R.D., Bowen, G.T., 1996. PROMPT: field trial and simulation results of bus priority in SCOOT. Presented at the Eighth International Conference on Road Traffic Monitoring and Control, London, UK, pp. 90–94. <https://doi.org/10.1049/cp:19960297>.
- Hounsell, N.B., McLeod, F.N., Gardner, K., Head, J.R., Cook, D., 2000. Headway-based bus priority in London using AVL: first results. Presented at the Tenth International Conference on Road Transport Information and Control, pp. 218–222. <https://doi.org/10.1049/cp:20000136>.
- Hu, J., Park, B., (Brian), Parkany, A.E., 2014. Transit signal priority with Connected Vehicle Technology. *Transp. Res. Rec.: J. Transp. Res. Board* 2418, 20–29, doi:10.3141/2418-03.
- Hu, J., Park, B.B., Lee, Y.-J., 2015. Coordinated transit signal priority supporting transit progression under Connected Vehicle Technology. *Transp. Res. Part C: Emerg. Technol.* 55, 393–408, doi:10.1016/j.trc.2014.12.005.
- Lin, Y., Yang, X., Zou, N., Franz, M., 2015. Transit signal priority control at signalized intersections: a comprehensive review. *Transp. Lett.* 7, 168–180, doi:10.1179/1942787514Y.0000000044.
- Ma, W., Bai, Y., 2007. Serve sequence optimization approach for multiple bus priority requests based on decision tree, in: *Plan, Build, and Manage Transportation Infrastructure in China*, Proceedings. Presented at the Seventh International Conference of Chinese Transportation Professionals Congress (ICCTP), pp. 605–615. [https://doi.org/10.1061/40952\(317\)59](https://doi.org/10.1061/40952(317)59).
- Ma, W., Yang, X., Liu, Y., 2010. Development and evaluation of a coordinated and conditional bus priority approach. *Transp. Res. Rec.: J. Transp. Res. Board* 2145, 49–58, doi:10.3141/2145-06.
- Ma, W., Ni, W., Head, L., Zhao, J., 2013. Effective coordinated optimization model for transit priority control under arterial progression. *Transp. Res. Rec.: J. Transp. Res. Board* 2356, 71–83, doi:10.1177/0361198113235600109.
- Ma, W., Head, K.L., Feng, Y., 2014. Integrated optimization of transit priority operation at isolated intersections: a person-capacity-based approach. *Transp. Res. Part C: Emerg. Technol.* 40, 49–62, doi:10.1016/j.trc.2013.12.011.
- Ma, W., Zou, L., An, K., Gartner, N.H., Wang, M., 2019. A partition-enabled multi-mode band approach to arterial traffic signal optimization. *IEEE Trans. Intell. Transp. Syst.* 20, 313–322, doi:10.1109/TITS.2018.2815520.
- Miucic, R., ed., 2019. *Connected Vehicles: Intelligent Transportation Systems, Wireless Networks*. Springer International Publishing.
- Skabardonis, A., 2000. Control strategies for transit priority. *Transp. Res. Rec.: J. Transp. Res. Board* 1727, 20–26, doi:10.3141/1727-03.
- Truong, L.T., Currie, G., Sarvi, M., 2017. Analytical and simulation approaches to understand combined effects of transit signal priority and road-space priority measures. *Transp. Res. Part C: Emerg. Technol.* 74, 275–294, doi:10.1016/j.trc.2016.11.020.
- Viegas, J., Lu, B., 2001. Widening the scope for bus priority with intermittent bus lanes. *Transport. Plan. Technol.* 24, 87–110, doi:10.1080/03081060108717662.

- Viegas, J., Lu, B., 2004. The Intermittent Bus Lane signals setting within an area. *Transp. Res. Part C: Emerg. Technol.* 12, 453–469, doi:10.1016/j.trc.2004.07.005.
- Viegas, J.M., Roque, R., Lu, B., Vieira, J., 2007. Intermittent Bus Lane System: Demonstration in Lisbon, Portugal. Presented at the Transportation Research Board 86th Annual Meeting, Washington DC, United States, pp. 1–10.
- Wang, Y., Ma, W., Yin, W., Yang, X., 2014. Implementation and testing of cooperative bus priority system in Connected Vehicle environment: case study in Taicang City, China. *Transp. Res. Rec.:J. Transp. Res. Board* 2424, 48–57, doi:10.3141/2424-06.
- Wirasinghe, S.C., Kattan, L., Rahman, M.M., et al., 2013. Bus rapid transit—a review. *Int. J. Urban Sci.* 17, 1–31, doi:10.1080/12265934.2013.777514.
- Yagar, S., 1993. Efficient transit priority at intersections. *Transp. Res. Rec.* 1390, 10–15.
- Zhao, B., Zang, X.-D., 2016. Research on the Policy of Private Cars Travelling at Peak Hours on BRT Roads in Guangzhou City. Presented at the The 16th COTA International Conference of Transportation Professionals (CICTP2016), Shanghai, P.R. China, pp. 2154–2163. <https://doi.org/10.1061/9780784479896.194>.
- Zhou, G., Gan, A., Zhu, X., 2006. Determination of optimal detector location for transit signal priority with queue jumper lanes. *Transp. Res. Rec.:J. Transp. Res. Board* 1978, 123–129, doi:10.1177/0361198106197800116.
- Zhou, W., Bai, Y., Li, J., Zhou, Y., Li, T., 2019. Integrated optimization of tram schedule and signal priority at intersections to minimize person delay. *J. Adv. Transport.* 2019, 1–18, doi:10.1155/2019/4802967.

Further Reading

- Hunt, P.B., Robertson, D.I., Bretherton, R.D., Winton, R.I., 1981. SCOOT—a traffic responsive method of coordinating signals (No. 1014). Transport Road Research Laboratory, p. 28.
- Smith, H.R., Hemily, P.B., Ivanovic, M., 2005. Transit Signal Priority (TSP): A Planning and Implementation Handbook. ITS America, Washington, DC.

Adaptive Bus Control

Monica Menendez, New York University Abu Dhabi (NYUAD), Abu Dhabi, UAE

© 2021 Elsevier Ltd. All rights reserved.

Introduction	315
Adaptive Control Strategies at the Infrastructure Level	317
Presignals	317
Presignals at Intersections With Single-Lane Approaches	317
Presignals at Intersections With Multiple-Lane Approaches	317
Perimeter Control	318
Perimeter Control as a Substitute for Dedicated Bus Lanes	318
Bimodal Perimeter Control	319
Adaptive Control Strategies During the Dispatching Process	319
Bus Holding	319
Bus Stop Skipping	320
Bus Substitution and Insertion	320
Adaptive Control Strategies at the Vehicle Level	321
Enroute Guidance Based on Signal Timing Information	321
Enroute Guidance Based on Headways or Timetables	321
Future Bus Adaptive Control	322
Acknowledgments	323
Biography	323
References	323
Further Reading	324

Introduction

Public transportation vehicles (e.g., buses), with the ability to carry a large number of passengers, can use limited road space more efficiently than private cars. Thus, increasing the share of passengers on buses can ultimately contribute to reducing traffic congestion and its associated externalities in urban areas. To this end, it is crucial for public transportation to offer high-quality services. This includes high reliability, competitive travel times, comfortable rides, dense network for high accessibility, competitive prices, unified payment system in the case of multiple operators, ease of transfers, etc. Some of these factors are very related to the design, while others are more specific to the operations of the system. In particular, service reliability, which has been shown to be a major predictor of public transportation ridership (Bates et al., 2001; Currie and Wallis, 2008), can be very much adjusted with operational strategies. Service reliability might refer to the guarantee that planned services are actually run, and that they adhere to the schedule. While the first aspect is an issue when there are budget cuts, broken vehicles, or other major disruptions to the system, the second aspect can be an issue on day-to-day operations, simply because of the normal fluctuations in traffic and demand. In this Chapter we focus on adaptive bus control as a means to increase the reliability of transportation systems for day-to-day operations. Such bus control can be implemented at the infrastructure level, during the dispatching process, or at the vehicle level (see Fig. 1).

Strategies implemented at the infrastructure level aim to provide priority to public transportation over other modes of transport. These priority strategies normally fall into one of two categories: those related to facility design and those related to traffic control (Skabardonis, 2000). In terms of facility design, it is clear that the allocation of road space across the different modes (e.g., cars vs buses) is key to the operations of the individual modes (Roca-Riu et al., 2020). Recently developed concepts, such as the three-dimensional macroscopic fundamental diagram (3D-MFD) (Geroliminis et al., 2014; Loder et al., 2017) allow to better understand and quantify the trade-offs between buses and cars, and how they affect the overall performance of a network in terms of vehicle or passenger throughput (Loder et al., 2019). This can be particularly valuable when considering different public transportation layouts (Loder et al., 2019; Dakic et al., 2020a, 2020b), including dedicated bus lanes (DBL) (Jacques and Levinson, 1997; Basso et al., 2011), intermittent bus lanes (IBL) (Viegas and Lu, 2001; Eichler and Daganzo, 2006), queue jumper lanes (Farid et al., 2015), and other types of flexible space allocation between buses and cars (Guler and Cassidy, 2012; He et al., 2018). In general, a higher proportion of managed bus lanes (i.e., those mentioned earlier) lead to more reliable travel times for buses (Arnet et al., 2015). More information about some of these lanes is provided in another chapter. In terms of signal control, it is clear that different algorithms can be used to assign different portions of the intersection capacity to the different modes. Thus, signal control is an important lever to increase the reliability of public transportation. Delays at traffic signals make up 10%–20% percent of the average bus travel times and about 50% of the overall bus delays (Evans and Skiles, 1970). Transit signal priority (TSP), for example, is an inexpensive way of reducing this delay for buses (Smith et al., 2005). TSP strategies are common in cities across the globe, and

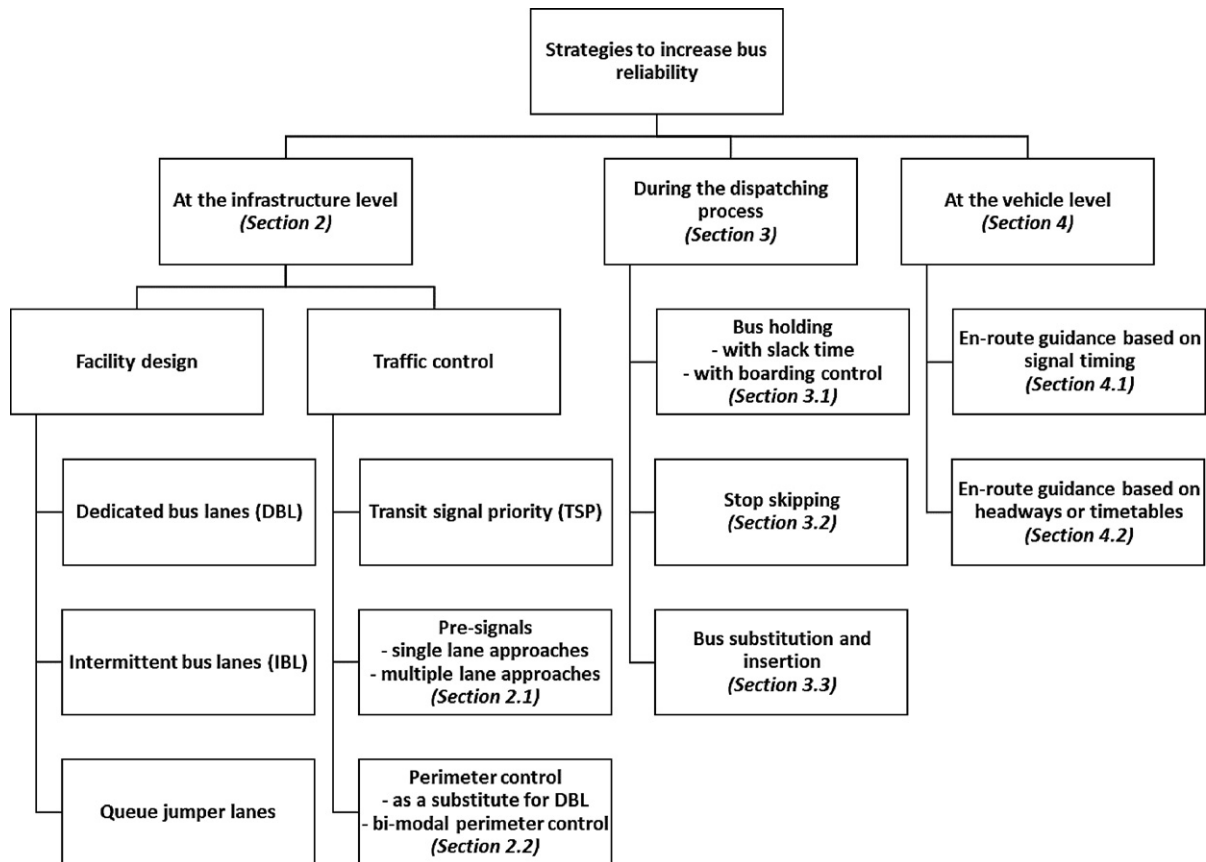


Figure 1 Strategies to increase reliability of public transportation vehicles.

they have been well-studied and well-documented for a long time (Levinson et al., 1975). They can be implemented to provide full or adaptive priority (Wolput et al., 2015), take into account the presence of nearby bus stops (Wu et al., 2017), or take into account only schedule-related delays (Yang et al., 2019a). An in-depth review of TSP will be provided in another chapter. More advanced signal control strategies, which can fully adapt to real-time conditions include the use of presignals at intersections with single-lane approaches (Guler et al., 2016), presignals at intersections with multiple-lane approaches (Guler and Menendez, 2014a,b), and bimodal perimeter control (He et al., 2019). More information about these strategies will be provided later in this Chapter in the [Section Adaptive Control Strategies at the Infrastructure Level](#).

Strategies during the dispatching process (i.e., affecting the whole route) can be used to address either changes in traffic conditions along a link and/or demand variations at bus stops. They mostly aim at mitigating bus bunching, a phenomenon related to the instability of the public transportation system (Newell and Potts, 1964). Bus bunching arises when a bus gets somehow delayed, and then it starts encountering more passengers than normal in subsequent stops. These additional passengers increase the dwell time which causes further delays. Simultaneously, the subsequent bus encounters fewer passengers and reduced travel times. Thus the two buses tend to come together (i.e., bunch) as they move along their route. Some dispatching strategies can be introduced during the planning stage and some during the operational stage. They include bus holding (Osuna and Newell, 1972), stop skipping (Vuchic, 1973), and bus substitution and insertion (Petit et al., 2018; Morales et al., 2019). More information about these strategies will be given later in this Chapter in the [Section Adaptive Control Strategies During the Dispatching Process](#).

Last, en-route control strategies (i.e., control strategies at the vehicle level) normally rely on Driver Advisory Systems (DAS), and in the future, could potentially exploit some of the advantages provided by Connected and Automated Vehicles (CAVs). These technologies allow us to switch from infrastructure-based strategies or dispatching strategies to vehicle-based strategies. They can work independently of TSP or other signal control algorithms, yet still provide benefits to buses at signalized intersections. They can also work independently of bus holding or insertion, and bus stop skipping strategies, while still reducing bus bunching. Typical strategies within this category include the optimal speed advisory based on intersection green times (Stahlmann et al., 2018), and vehicle control based on headways or timetables (Daganzo and Pilachowski, 2011). More information about these strategies will be provided later in this Chapter in the [Section Adaptive Control Strategies at the Vehicle Level](#).

As new technologies and information sources emerge, these strategies will undoubtedly evolve, and new strategies will be developed. A brief look into the future is provided in the [Section Future Bus Adaptive Control](#) at the end of this Chapter.

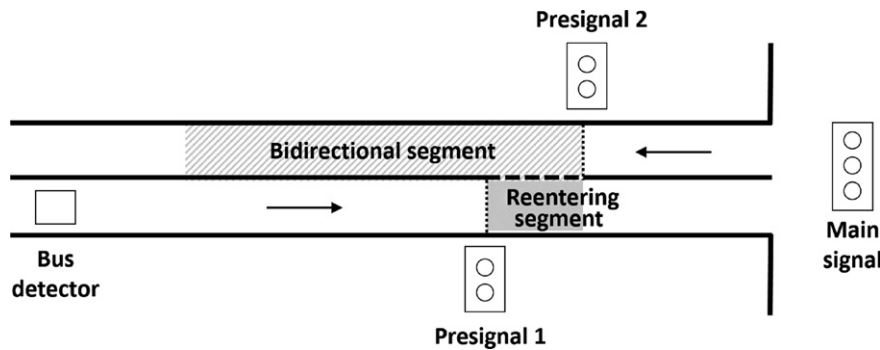


Figure 2 Schematic representation of a presignal at an intersection with a single-lane approach.

Adaptive Control Strategies at the Infrastructure Level

In this section, we focus on advanced signal control strategies that can adapt to real time conditions. In particular, we will focus on the use of presignals (Section Presignals) and perimeter control (Section Perimeter Control).

Presignals

Presignals are additional signals located upstream of an intersection. They are used to provide priority to buses, and can function together with (or independently of) TSP or any other signal control at the intersection itself. They are relatively uncommon in practice, but they are very effective at increasing the reliability of public transportation services. They can be used at intersections with single-lane approaches (Section Presignals at Intersections With Single-Lane Approaches), or with multiple-lane approaches (Section Presignals at Intersections With Multiple-Lane Approaches).

Presignals at Intersections With Single-Lane Approaches

At intersections with single-lane approaches, it is very difficult to provide bus priority. A dedicated bus lane means no cars can use that link, and TSP is not effective if buses get trapped in congestion behind cars in a mixed traffic lane. Thus, presignals are a relatively inexpensive way of allowing the bus to bypass the queue and adhere to its schedule without significantly limiting the movement of cars.

To provide this type of priority, two presignals are placed upstream of the main signal in the travel direction of the bus (see Fig. 2). One of the signals controls traffic in the bus lane (presignal 1), while the other one controls traffic in the opposite lane (presignal 2). Further upstream an additional sensor is used to detect incoming buses. When there are no buses, the two presignals remain green, so traffic moves as usual. On the other hand, when a bus is detected at the sensor location, the two presignals turn red, stopping cars traveling in the two directions. This clears both the bidirectional segment (in the opposite lane) and the lane reentering segment (in the bus travel direction). The goal is for the bus to be able to use the bidirectional segment (or a portion of it) to bypass any queue on its lane, and reenter its lane closer to the main intersection so that it can discharge within a single cycle with minimum delay (Guler et al., 2016). Notice that the signal settings at the intersection do not need to be modified.

The locations and timings of the two presignals need to be carefully designed to make sure that (1) the bus has room to maneuver between presignals 1 and 2 when reentering its lane, (2) all traffic in the bidirectional segment can clear that segment before the bus comes in, (3) the queue created when presignal 2 turns red does not spill back into the intersection, and (4) the bus can skip as much of a queue as possible (Guler et al., 2016). Some of these conditions compete with each other, so a proper setting for presignals at intersections with single-lane approaches must consider those trade-offs. In other words, any implementation should consider both the savings provided to buses as well as the extra delays imposed onto cars, ultimately trying to minimize the total passenger delay across the two modes.

Presignals at Intersections With Multiple-Lane Approaches

In the case of intersections with multiple-lane approaches presignals are used in combination with a dedicated bus lane, especially in cases where the low bus flows do not justify dedicated infrastructure at the intersection itself. The layout and operational strategy in this case are simpler than for single-lane approaches. Only one presignal for the car lanes and a sensor to detect incoming buses in the bus lane are needed (see Fig. 3). This presignal, located upstream of the main intersection, is used to discontinue the dedicated lane so that we can increase the car capacity of the intersection if no buses are present (Guler and Menendez, 2015). As before, the signal settings at the intersection do not need to be modified.

When a bus is detected, the presignal turns red, stopping all car traffic and allowing the bus to maneuver in front of those cars. In this way, the bus can discharge from the intersection within the same cycle with minimum delay. Once the bus passes the presignal location, the presignal can turn green again, if it had been green without the bus. In the case where there are no buses at all, the presignal turns green in advance of the main signal, so that cars can move down to all lanes at the intersection and when the main

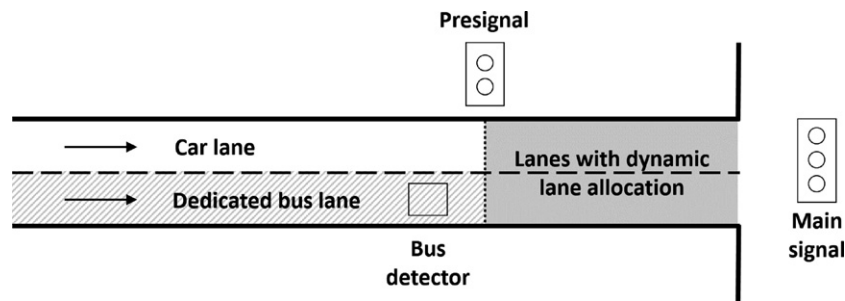


Figure 3 Schematic representation of a presignal at an intersection with a multiple-lane approach.

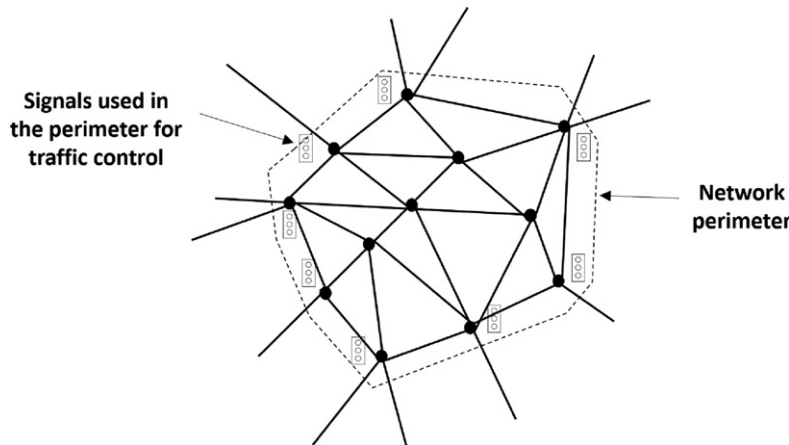


Figure 4 Schematic representation of an area with perimeter control.

signal turns green, cars can discharge using the intersection's full capacity. The presignal should then turn red in advance of the main signal to make sure the area in between the two remains clear of cars for most of the duration of the main red. This way, if a bus arrives, it can move to the front of the queue without any disruption (Guler and Menendez, 2014a).

The location and timing of the presignal must be carefully designed to make sure that (1) the capacity at the main intersection is fully used, that is, the intersection does not starve especially as the main signal turns green, (2) buses can move to the front of the queue as often as possible so they can adhere to their schedule even if the dedicated lane is discontinued, and (3) the queue of cars upstream of the presignal does not grow too long as to spill back into the upstream intersection. As before, some of these conditions compete with each other. Thus close attention should be paid to these trade-offs when designing the system. In particular, presignals for multiple-lane approaches are not effective when the system is close to saturation (Guler and Menendez, 2014b, 2015), as they can lead to very long queues. For varying demands, it is then possible to use adaptive presignals (He et al., 2016) that turn on and off in response to real time traffic conditions, and always aim to minimize the overall passenger delay through the intersection.

Perimeter Control

Perimeter control is used to regulate the inflow of vehicles into a specific area (e.g., downtown). Traffic signals in the perimeter of such area control the rate at which vehicles enter, and this in turn can be used to ensure that the area remains uncongested (see Fig. 4). The idea has gained popularity in the last decade with the development of new macroscopic monitoring and modeling tools for urban networks. Perimeter control can be used instead of dedicated bus lanes under certain scenarios (Section Perimeter Control as a Substitute for Dedicated Bus Lanes) or simply to control the inflow of the different types of modes, providing priority to public transportation (Section Bimodal Perimeter Control).

Perimeter Control as a Substitute for Dedicated Bus Lanes

Although dedicated bus lanes can be very effective for providing priority to public transportation vehicles, they are also rather controversial, as they consume space reducing the available capacity for cars and potentially increasing congestion. Recently, it has been shown that under certain conditions it is possible to maintain the reliability of public transportation vehicles by substituting the controversial dedicated bus lanes with a multiscale perimeter control scheme (Chiabaut et al., 2018). The perimeter control scheme tries to maintain optimal traffic densities within the protected area, while simultaneously reducing delay at the gating

intersections (Yang et al., 2018). This guarantees that buses can move through the network at the desired speeds, yet cars still can use all lanes. While this type of statement still needs more research, it seems a promising direction.

Bimodal Perimeter Control

The idea of perimeter control has become quite popular in the years since the concept of the macroscopic fundamental diagram (MFD) emerged (Geroliminis and Daganzo, 2008). The MFD relates the total accumulation of vehicles in the network to its total throughput. An extension to this concept is the 3D-MFD (Geroliminis et al., 2014; Loder et al., 2017), which relates the individual accumulation of buses and cars in the network to the overall vehicle or people throughput across the two modes. The 3D-MFD allows to better understand and quantify the trade-offs between the two modes as well as their interactions. This opens the door to more advanced perimeter control strategies, that regulate the inflows of the two modes individually (He et al., 2019). It can also be extended to include other priority schemes in the perimeter (Yang et al., 2019b). The overall goal is to maximize the total people throughput in the network.

Adaptive Control Strategies During the Dispatching Process

In this section, we focus on strategies that are related to the dispatching of vehicles. They tend to affect the performance along the whole route or at least a portion of it. These include bus holding strategies (Section Bus Holding), bus stop skipping strategies (Section Bus Stop Skipping), and bus substitution and insertion strategies (Section Bus Substitution and Insertion).

All these strategies can be used independently or in combination with each other, for example, bus holding can be combined with stop skipping (Fu et al., 2003; Cortés et al., 2010). Given that they all have both advantages and disadvantages, they should be evaluated carefully, taking into account the specific settings of the system, before any real implementation. Important metrics to consider include (1) waiting times, (2) commercial bus speeds with (3) the resulting riding times, (4) bus loads, (5) bus trajectories, (6) cycle time distributions, and (7) loss of service for specific passengers. The goal is to increase (2), reduce the average and the variance of (1) and (3), make sure (4) are as uniform as possible across buses (i.e., avoid a mix of full and empty buses), minimize deviations in (5) so to avoid bus bunching (i.e., try to keep headways as constant as possible throughout the route), reduce both the average and the variance of (6) in order to reduce operating costs while maintaining reliability, and minimize (7).

Bus Holding

It is well-known that when buses along a route are not properly controlled, they cannot adhere to their schedule and they tend to bunch (Newell and Potts, 1964). Initial strategies to counteract this phenomenon included holding buses at the dispatching station. This was initially done when no real time information was available (Osuna and Newell, 1972), and then with real time information (Eberlein et al., 2001; Hickman, 2001). With time, these strategies came to include multiple control points along the route (i.e., bus stops) where buses could be held (Daganzo, 2009; Bartholdi and Eisenstein, 2012) (see Fig. 5).

Some of these holding mechanisms were combined with an additional (slack) time built into the schedule so vehicles could catch up, if needed. Slack times reduce the commercial speed of the buses, increasing in turn passenger riding times and potentially the

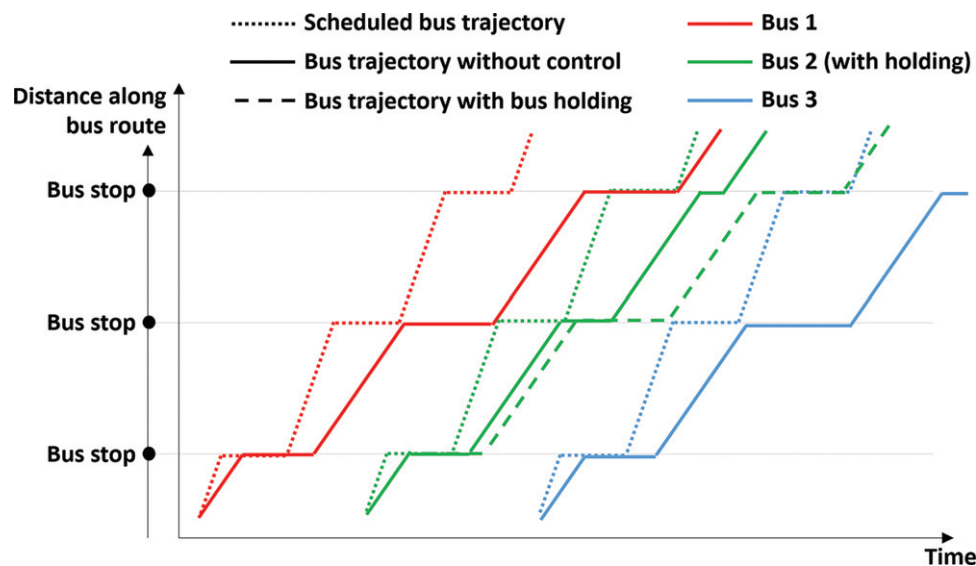


Figure 5 Schematic representation of bus trajectories with and without bus holding.

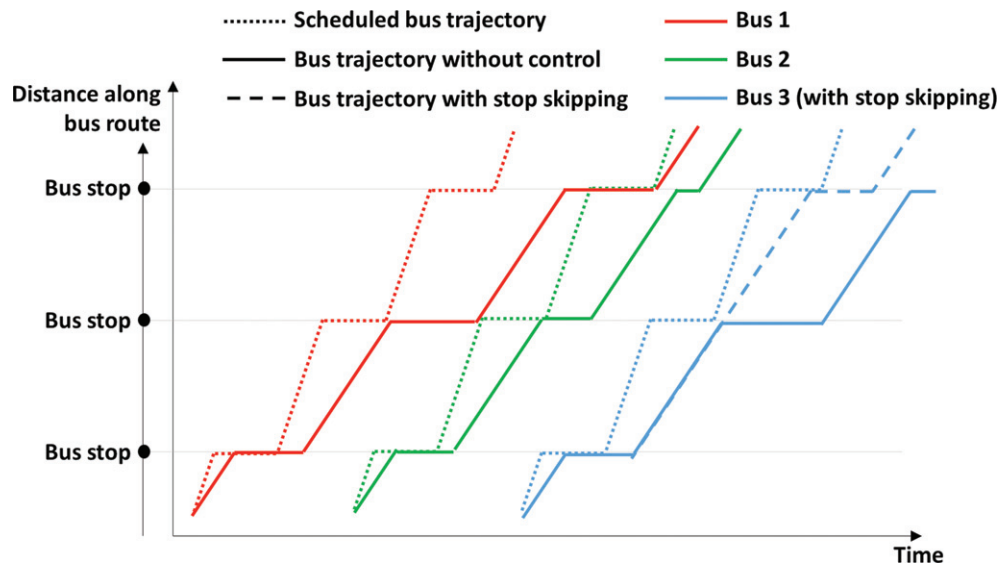


Figure 6 Schematic representation of bus trajectories with and without stop skipping.

operating costs. Therefore, they must be relatively small (Daganzo, 1997). These small slack times, when implemented alone, cannot compensate for many system disruptions or a large one (Newell, 1977).

Bus holding can also be combined with boarding limits (Delgado et al., 2012). Boarding limits can emerge spontaneously when passengers chose to wait for the next bus expecting it to have more space. They can also be enforced when the driver asks all passengers or a portion thereof to wait for the next bus.

Bus holding is, in general, one of the most common strategies to mitigate bus bunching, as it is a relatively easy strategy to implement, at a relatively low cost for both the operator and the users.

Bus Stop Skipping

This strategy calls for buses to skip one or more stops along a route. This can be decided in the planning phase to design express services or during real time operations to deal with some disruptions to the system (see Fig. 6).

The idea of introducing stop skipping in the planning phase, offering two types of service (normal vs express), is not new (Vuchic, 1973). It is typically done when there is a high demand associated with specific origin-destination (OD) pairs. It aims at reducing riding times for some portion of the demand, and potentially reduce fleet size requirements. However, for certain frequency settings, it is possible that express services generate more demand than what they can actually handle, leading to additional queues and increased travel times, even when the aggregate capacity of the route is enough (Larrain and Muñoz, 2020).

Bus stop skipping introduced during the normal operations aims to increase the adherence to predefined timetables, or maintenance of the headways. The unpredictability associated with this mechanism can make it an inconvenience for passengers whose origin or destination is skipped. However, even after accounting for that, there are control algorithms that can actually improve the overall system (Sun and Hickman, 2005). As new technologies to share information with the passengers emerge, the use of bus stop skipping strategies will potentially increase.

Bus Substitution and Insertion

Bus substitution strategies are also used to address bus bunching. When a bus is detected to be very late, a new bus (previously in standby) is inserted to take over the schedule of the late bus, and the late bus is marked out of service so it only drops the passengers on-board (does not pick up any new passengers). Once the late bus finishes dropping its passengers, it goes to a standby location or is re-inserted somewhere else along the route (Petit et al., 2018). This strategy is very effective in reducing bus bunching but requires a larger vehicle fleet as some buses must always be in standby.

Alternatively, it is possible to just insert buses without taking any out of service. Similar to the previous strategy, some buses are kept in standby at one or multiple locations along a route. Once a headway is detected to be longer than a given threshold, a new bus is inserted to shorten that headway. For some conditions, this method has been proven to be more beneficial to the system than having the whole fleet of buses constantly circulating (Morales et al., 2019).

In general, bus substitution and insertion strategies can be a very effective way to increase the reliability of the system (see Fig. 7). However, a careful analysis is needed before implementation to make sure they are also cost efficient.

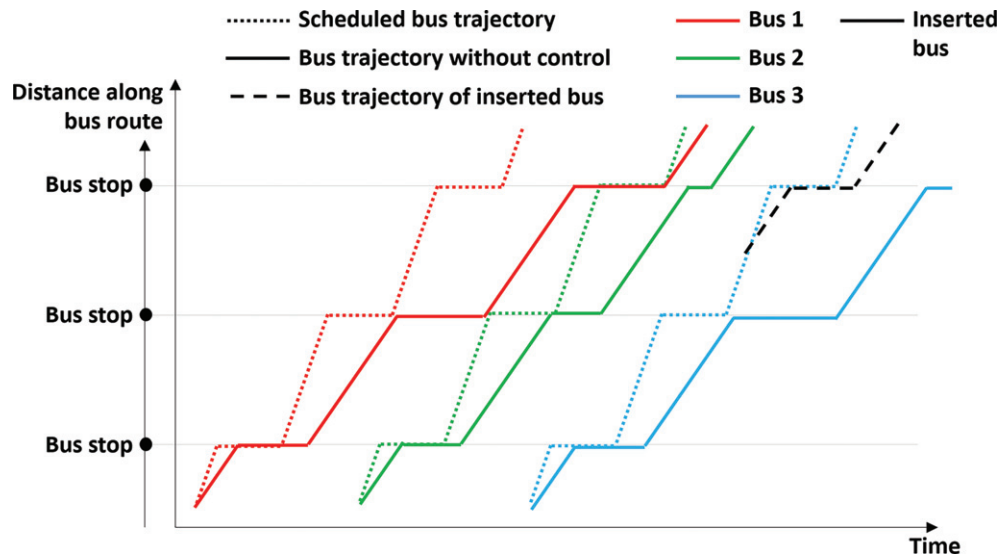


Figure 7 Schematic representation of bus trajectories with and without inserting a bus.

Adaptive Control Strategies at the Vehicle Level

In this section, we focus on strategies that employ DAS (or potentially vehicle connectivity) to regulate the speed profile of a vehicle. These include those related to the timing of the intersection signals (e.g., speed advisory based on green time), and those related to the headways between vehicles or their actual timetables (e.g., to mitigate bus bunching). The former are discussed in the Section Enroute Guidance Based on Signal Timing Information; and the latter are discussed in the Section Enroute Guidance Based on Headways or Timetables.

Enroute Guidance Based on Signal Timing Information

One of the main goals of this type of strategies is to mitigate stop and go maneuvers for buses. Information related to the upcoming signal is given to the public transportation vehicles, so they can better time their arrival at the intersection and reduce the overall number of stops and/or acceleration/deceleration maneuvers.

In the case of speed advisory (Stahlmann et al., 2018), the suggested speed profile for the bus is given based on the phase and timing of the intersection, as well as the actual location of the bus. The goal is for buses to modify their speed, so they can go through the intersection without actually stopping, potentially entering the intersection with a higher speed. In the case of dwell time advisory (Seredynski and Khadraoui, 2014) the suggested dwell times are also given based on the phase and timing of the intersection, and the actual location of the bus stop. The goal is for the buses to wait at near-side bus stops so they can later go through the intersection without stopping. Some of the outcomes might be similar to those of bus holding strategies, but not necessarily.

These two strategies can be implemented independently of TSP, although recent studies have suggested to combine them with some type of priority at the signals (Seredynski et al., 2019). Notice that without TSP or some other type of signal priority, these strategies do not necessarily reduce the delay (as they just move it upstream of the intersection) (see Fig. 8). Combined with TSP, however, they could be used to increase the reliability of the system.

Enroute Guidance Based on Headways or Timetables

Regulating the speed of buses is also relevant for ensuring either constant headway across buses and/or adherence to the timetables. In particular, it can be a very valuable method to minimize bus bunching. Systems running on predefined timetables are typically used to serve low demands, with low bus frequencies (i.e., long headways). Systems serving high demands, with high bus frequencies, normally run on a headway-based schedule.

For public transportation systems that have predefined timetables, automatic vehicle location (AVL) technologies can be used to compare the exact position of a bus at a given time with its target position. This, in turn, can be used to make constant adjustments to the speed along the whole route. Alternatively, for public transportation systems that have a headway-based schedule, each bus can be regulated based on the spacing with the bus in front and the spacing with the bus behind (Daganzo and Pilachowski, 2011) (see Fig. 9). Extensions to this strategy have proven to be useful also for buses following predefined timetables (Xuan et al., 2011), especially when combined with bus holding (Argote-Cabanero et al., 2015). Other applications (e.g., the Transantiago's bus system

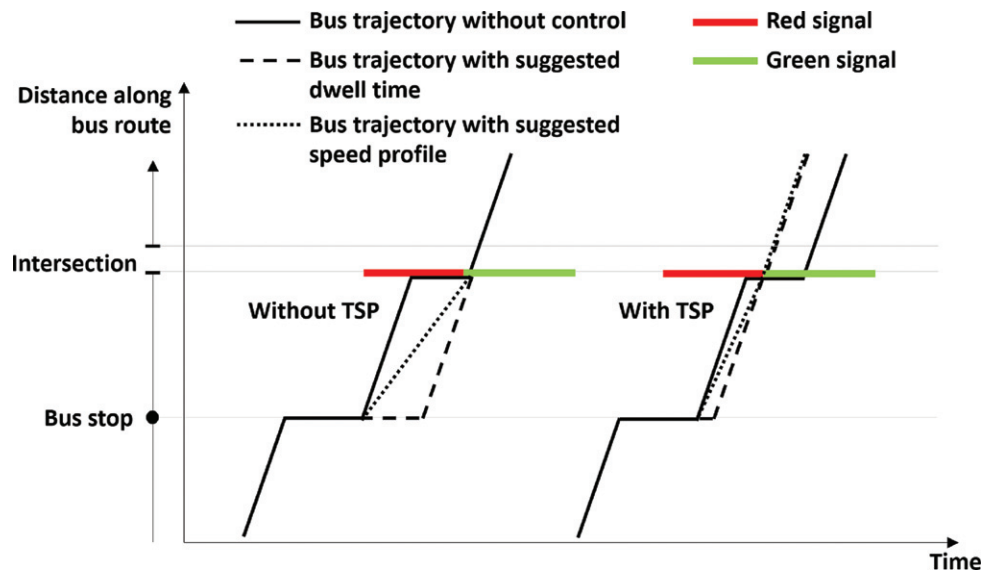


Figure 8 Schematic representation of bus trajectories with and without control using signal timing information. TSP refers to transit signal priority

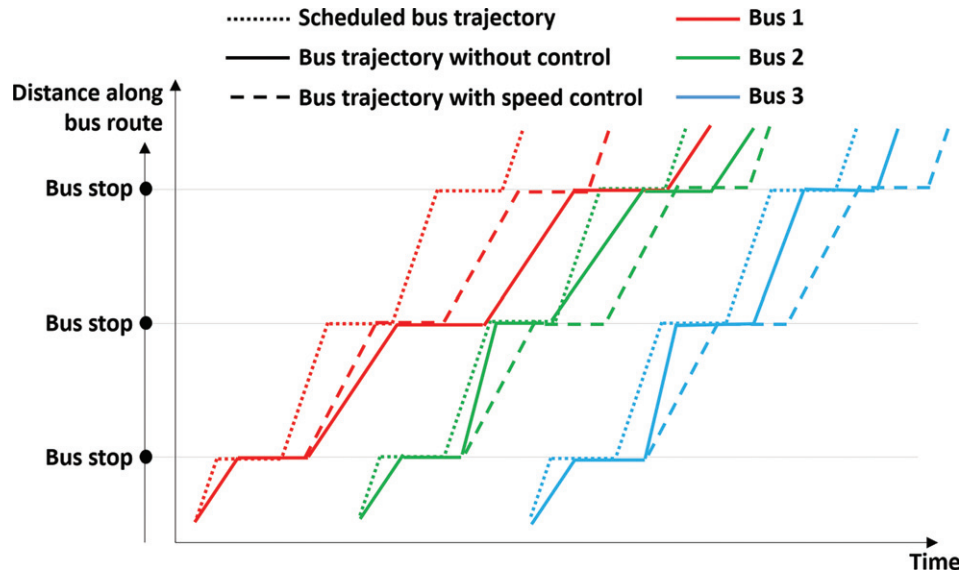


Figure 9 Schematic representation of bus trajectories with and without speed control based on headways or timetables.

in Santiago de Chile) use a combination of a maximum advised speed (so buses can catch up) with bus holding schemes for high-frequency systems (Lizana et al., 2014).

To simplify the enroute speed adjustment process for the drivers, buses can have either a clock that shows how early or late the bus is, a display that shows whether the driver needs to accelerate or decelerate, or a display that shows the maximum speed at which the bus should drive at any given instant. Ideally, buses should maintain relatively constant speeds, and acceleration/deceleration maneuvers should be kept to a minimum.

Future Bus Adaptive Control

Looking forward, new information and communication technologies (ICT), smart infrastructure, and connected and automated vehicles (CAV) hold the promise of more reliable public transportation systems. First, ICT and smart infrastructure could be used to allocate space dynamically to buses while minimizing the disruptions to cars (Wang et al., 2016). Second, these technologies can also be used to develop smarter dispatching strategies, responsive to real time information regarding public transportation demands, as well as traffic information (Dakic et al., 2020c). Third, both the connectivity and computer driving capabilities of the vehicles

could be exploited to easily adapt the speed profile of buses in coordination with optimized signal settings that provide bus priority and take into account the presence of near-side or far-side bus stops, as well as the bus schedules (Yang et al., 2019a). Fourth, ICT and the connectivity across vehicles could also be exploited to predict and minimize the occurrence of bus bunching (Andres and Nair, 2017). In general, we expect that new technologies and information sources not only change the type of vehicles (e.g., to include for example automated modular bus units), but also the flexibility of the system (e.g., to increase coverage and address demand requirements on real time), and its efficiency and reliability (e.g., to react to real time traffic conditions while minimizing both the user and the operator's costs). With all of these resources, both current and upcoming, public transportation planners and engineers have never been better positioned to craft optimal bus services for the benefit of city dwellers everywhere.

Acknowledgments

This work was partially supported by the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001 and by the Swiss Re Institute under the Quantum CitiesTM initiative.

The author also wants to thank Philip Rodenbough and the NYU Abu Dhabi Scientific Writing Program for the feedback on manuscript development.

Biography



Monica Menendez is, since January 2018, an Associate Professor of Civil and Urban Engineering at New York University in Abu Dhabi; and a Global Network Associate Professor at the Tandon School of Engineering in New York University. She is also the Director of the Research Center for Interacting Urban Networks (CITIES). Between 2010 and 2017, Monica was the Director of the research group Traffic Engineering at ETH Zurich. Prior to that, she was a Management Consultant at Bain & Company. She holds a PhD and a MSc in Civil and Environmental Engineering from UC Berkeley. During her studies there she received an NSF Fellowship and the Gordon F. Newell Award. In total, she is the recipient of more than 20 scholarships and awards from well-known and prestigious organizations, professional societies, and universities. Monica also holds a dual degree in Civil Engineering and Architectural Engineering from the University of Miami, from where she graduated Summa Cum Laude.

Her research interests include multimodal transportation systems paying special attention to new technologies and information sources. She is an active reviewer for over 20 journals, and a member of multiple editorial boards for top journals in Transportation. Monica is the author of over 70 peer reviewed journal publications and over 160 conference proceedings and technical reports.

References

- Andres, M., Nair, R., 2017. A predictive-control framework to address bus bunching. *Transp. Res. Part B* 123–148.
- Argote-Cabanero, J., Daganzo, C.F., Lynn, J.W., 2015. Dynamic control of complex transit systems. *Transp. Res. Part B: Method.* 81, 146–160.
- Arnet, K., Guler, S.I., Menendez, M., 2015. Effects of multimodal operations on urban roadways. *Transp. Res. Rec.* 2533 (1), 1–7.
- Bartholdi III, J.J., Eisenstein, D.D., 2012. A self-coordinating bus route to resist bus bunching. *Transp. Res. Part B: Method.* 46 (4), 481–491.
- Basso, L.J., Guevara, C.A., Gschwender, A., Fuster, M., 2011. Congestion pricing, transit subsidies and dedicated bus lanes: Efficient and practical solutions to congestion. *Transp. Policy* 18 (5), 676–684.
- Bates, J., Polak, J., Jones, P., Cook, A., 2001. The valuation of reliability for personal travel. *Transp. Res. Part E: Logist. Transp. Rev.* 37 (2–3), 191–229.
- Chiabaut, N., Küng, M., Menendez, M., Leclercq, L., 2018. Perimeter control as an alternative to dedicated bus lanes: A case study. *Transp. Res. Rec.* 2672 (20), 110–120.
- Cortés, C.E., Sáez, D., Milla, F., Núñez, A., Riquelme, M., 2010. Hybrid predictive control for real-time optimization of public transport systems' operations based on evolutionary multi-objective optimization. *Transp. Res. Part C: Emerg. Technol.* 18 (5), 757–769.
- Currie, G., Wallis, I., 2008. Effective ways to grow urban bus markets—a synthesis of evidence. *J. Transp. Geogr.* 16 (6), 419–429.
- Daganzo, C.F., Pilachowski, J., 2011. Reducing bunching with bus-to-bus cooperation. *Transp. Res. Part B: Method.* 45 (1), 267–277.
- Daganzo, C.F., 1997. Schedule instability and control Daganzo, CF *Fundamentals of Transportation and Transportation Operations*, Elsevier, New York, NY, pp. 304–309.
- Daganzo, C.F., 2009. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transp. Res. Part B: Method.* 43 (10), 913–921.
- Dakic, I., Ambühl, L., Schümperlin, O., Menendez, M., 2020a. On the modeling of passenger mobility for stochastic bi-modal urban corridors. *Transp. Res. Part C: Emerg. Technol.* 113, 146–163.
- Dakic, I., Leclercq, L., Menendez, M., 2020b. On the optimization of the bus network design: an analytical approach based on a bi-modal macroscopic fundamental diagram. Preprints, 2020060312.
- Dakic, I., Yang, K., Menendez, M., Chow, J.Y., 2020c. On the design of an optimal flexible bus dispatching system with modular bus units: using the three-dimensional macroscopic fundamental diagram. Preprints, 2020060349.
- Delgado, F., Munoz, J.C., Giesen, R., 2012. How much can holding and/or limiting boarding improve transit performance? *Transp. Res. Part B: Method.* 46 (9), 1202–1217.
- Eberlein, X.J., Wilson, N.H., Bernstein, D., 2001. The holding problem with real-time information available. *Transp. Sci.* 35 (1), 1–18.
- Eichler, M., Daganzo, C.F., 2006. Bus lanes with intermittent priority: Strategy formulae and an evaluation. *Transp. Res. Part B: Method.* 40 (9), 731–744.
- Evans, H.K., Skiles, G.W., 1970. Improving public transit through bus preemption of traffic signals. *Traffic Quart.* 24, (4).

- Farid, Y.Z., Christofa, E., Collura, J., 2015. Dedicated bus and queue jumper lanes at signalized intersections with nearside bus stops: Person-based evaluation. *Transp. Res. Rec.* 2484 (1), 182–192.
- Fu, L., Liu, Q., Calamai, P., 2003. Real-time optimization model for dynamic scheduling of transit operations. *Transp. Res. Rec.* 1857 (1), 48–55.
- Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transp. Res. Part B: Method.* 42 (9), 759–770.
- Geroliminis, N., Zheng, N., Ampountolas, K., 2014. A three-dimensional macroscopic fundamental diagram for mixed bi-modal urban networks. *Transp. Res. Part C: Emerg. Technol.* 42, 168–181.
- Guler, S.I., Cassidy, M.J., 2012. Strategies for sharing bottleneck capacity among buses and cars. *Transp. Res. B* 46 (10), 1334–1345.
- Guler, S.I., Menendez, M., 2014a. Analytical formulation and empirical evaluation of pre-signals for bus priority. *Transp. Res. B* 64, 41–53.
- Guler, S.I., Menendez, M., 2014b. Evaluation of presignals at oversaturated signalized intersections. *Transp. Res. Rec.* 2418 (1), 11–19.
- Guler, S.I., Menendez, M., 2015. Pre-signals for bus priority: Basic guidelines for implementation. *Public Transp.* 7 (3), 339–354.
- Guler, S.I., Gayah, V.V., Menendez, M., 2016. Bus priority at signalized intersections with single-lane approaches: A novel pre-signal strategy. *Transp. Res. Part C: Emerg. Technol.* 63, 51–70.
- He, H., Guler, S.I., Menendez, M., 2016. Adaptive control algorithm to provide bus priority with a pre-signal. *Transp. Res. Part C: Emerg. Technol.* 64, 28–44.
- He, H., Yang, K., Liang, H., Menendez, M., Guler, S.I., 2019. Providing public transport priority in the perimeter of urban networks: A bimodal strategy. *Transp. Res. Part C: Emerg. Technol.* 107, 171–192.
- Haitao, He, Menendez, M., Guler, S.I., 2018. Analytical evaluation of flexible-sharing strategies on multimodal arterials. *Transp. Res. Part A: Policy Prac.* 114, 364–379.
- Hickman, M.D., 2001. An analytic stochastic model for the transit vehicle holding problem. *Transp. Sci.* 35 (3), 215–237.
- Jacques, K.S. and Levinson, H.S., 1997. TCRP Report 26: operational analysis of bus lanes on arterials. *TRB, National Research Council, Washington, DC.*
- Larraz, H., Muñoz, J.C., 2020. The danger zone of express services: When increasing frequencies can deteriorate the level of service. *Transp. Res. Part C: Emerg. Technol.* 113, 213–227.
- Levinson, H.S., Adams, C.L. and Hoey, W.F., 1975. Bus use of highways: Planning and design guidelines. *NCHRP Report*, (155).
- Lizana, P., Muñoz, J.C., Giesen, R., Delgado, F., 2014. Bus control strategy application: case study of Santiago transit system. *Proc. Comp. Sci.* 32, 397–404.
- Loder, A., Ambühl, L., Menendez, M., Axhausen, K.W., 2017. Empirics of multi-modal traffic networks—Using the 3D macroscopic fundamental diagram. *Transp. Res. Part C: Emerg. Technol.* 82, 88–101.
- Loder, A., Ambühl, L., Menendez, M., Axhausen, K.W., 2019. Understanding traffic capacity of urban networks. *Sci. Rep.* 9 (2019), 16283.
- Loder, A., Dakic, I., Bressan, L., Ambühl, L., Bliemer, M.C., Menendez, M., Axhausen, K.W., 2019. Capturing network properties with a functional form for the multi-modal macroscopic fundamental diagram. *Transp. Res. Part B: Method.* 129, 1–19.
- Morales, D., Muñoz, J.C., Gazmuri, P., 2020. A stochastic model for bus injection in an unscheduled public transport service. *Transp. Res. Part C: Emerg. Technol.* 113, 277–292.
- Newell, G.F. and Potts, R.B., 1964. Maintaining a bus schedule. In *Australian Road Research Board (ARRB) Conference*, 2nd, 1964, Melbourne (Vol. 2, No. 1).
- Newell, G.F., 1977. Unstable Brownian motion of a bus trip. In *Statistical Mechanics and Statistical Methods in Theory and Application*, Springer, Boston, MA, pp. 645–667.
- Osuna, E.E., Newell, G.F., 1972. Control strategies for an idealized public transportation system. *Transp. Sci.* 6 (1), 52–72.
- Petit, A., Ouyang, Y., Lei, C., 2018. Dynamic bus substitution strategy for bunching intervention. *Transp. Res. Part B: Method.* 115, 1–16.
- Roca-Riu, M., Menendez, M., Dakic, I., Buehler, S., Ortigosa, J., 2020. Urban space consumption of cars and buses: an analytical approach. *Transportmetrica B* 8 (1), 237–263.
- Seredynski, M. and Khadraoui, D., 2014, October. Complementing transit signal priority with speed and dwell time extension advisories. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)* (pp. 1009–1014). IEEE.
- Seredynski, M., Laskaris, G. and Viti, F., 2019. Analysis of Cooperative Bus Priority at Traffic Signals. *IEEE Transactions on Intelligent Transportation Systems.*
- Skabardonis, A., 2000. Control strategies for transit priority. *Transp. Res. Rec.* 1727 (1), 20–26.
- Smith, H.R., Hemily, B. and Ivanovic, M., 2005. Transit signal priority (TSP): A planning and implementation handbook. ITS America, Department of Transportation.
- Stahlmann, R., Möller, M., Brauer, A., German, R., Eckhoff, D., 2018. Exploring GLOSA systems in the field: Technical evaluation and results. *Computer Comm.* 120, 112–124.
- Sun, A., Hickman, M., 2005. The real-time stop-skipping problem. *J. Intel. Transp. Sys.* 9 (2), 91–109.
- Viegas, J., Lu, B., 2001. Widening the scope for bus priority with intermittent bus lanes. *Transp. Plan. Technol.* 24 (2), 87–110.
- Vuchic, V.R., 1973. Skip-stop operation as a method for transit speed increase. *Traffic Quart.* 307.
- Wang, C., David, B., Chalou, R., Yin, C., 2016. Dynamic road lane management study: A smart city application. *Transp. Res. Part E: Logist. Transp. Rev.* 89, 272–287.
- Wolput, B., Tegenbos, R., Skabardonis, A. and Tampère, C.M., 2015. Control strategies for transit priority: Comparing adaptive priority with full transit signal priority (No. 15-2662).
- Wu, K., Guler, S.I., Gayah, V.V., 2017. Estimating the impacts of bus stops and transit signal priority on intersection operations: queuing and variational theory approach. *Transp. Res. Rec.* 2622 (1), 70–83.
- Xuan, Y., Argote, J., Daganzo, C.F., 2011. Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis. *Transp. Res. Part B: Method.* 45 (10), 1831–1845.
- Yang, K., Menendez, M., Guler, S.I., 2019a. Implementing transit signal priority in a connected vehicle environment with and without bus stops. *Transp. B: Transp. Dyn.* 7 (1), 423–445.
- Yang, K., Menendez, M., Zheng, N., 2019b. Heterogeneity aware urban traffic control in a connected vehicle environment: A joint framework for congestion pricing and perimeter control. *Transp. Res. Part C: Emerg. Technol.* 105, 439–455.
- Yang, K., Zheng, N., Menendez, M., 2018. Multi-scale perimeter control approach in a connected-vehicle environment. *Transp. Res. C* 94, 32–49.

Further Reading

- Daganzo, C.F. and Yanfeng, O., 2019. *Public Transportation Systems: Principles of System Design, Operations Planning and Real-time Control*. World Scientific.
- Desaulniers, G., Hickman, M.D., 2007. *Public transit Handbooks in operations research and management science*, 14. pp. 69–127.
- Guler, S.I., Gayah, V.V., Menendez, M., 2016. Bus priority at signalized intersections with single-lane approaches: A novel pre-signal strategy. *Transp. Res. Part C: Emerg. Technol.* 63, 51–70.
- Ibarra-Rojas, O.J., Delgado, F., Giesen, R., Muñoz, J.C., 2015. Planning, operation, and control of bus transport systems: A literature review. *Transp. Res. Part B: Method.* 77, 38–75.
- Loder, A., Ambühl, L., Menendez, M., Axhausen, K.W., 2017. Empirics of multi-modal traffic networks—Using the 3D macroscopic fundamental diagram. *Transp. Res. Part C: Emerg. Technol.* 82, 88–101.
- Morales, D., Muñoz, J.C., Gazmuri, P., 2019. A stochastic model for bus injection in an unscheduled public transport service. *Transp. Res. Part C: Emerg. Technol.*
- Petit, A., Ouyang, Y., Lei, C., 2018. Dynamic bus substitution strategy for bunching intervention. *Transp. Res. Part B: Method.* 115, 1–16.
- Skabardonis, A., 2000. Control strategies for transit priority. *Transp. Res. Rec.* 1727 (1), 20–26.
- Xuan, Y., Argote, J., Daganzo, C.F., 2011. Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis. *Transp. Res. Part B: Method.* 45 (10), 1831–1845.

Transit Fare Collection

Mahmoud Mesbah^{*,†}, Kamal Khanali^{*}, ^{*}Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran; [†]The University of Queensland, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	325
Fare Structures	325
Fare Policies	325
Credit Plans	326
General Credit Types	326
Special Tickets and Credits	327
Payment Methods	327
Ticketing Technologies	329
Equipment	329
Integration of Fare Payment	330
Conclusions	330
References	330

Introduction

Individuals are required to pay a fare to board public transport services which is used to (partially) recover the service costs. This chapter presents a taxonomy, which can be used to characterize a fare collection system of public transport. The presented characterization can be useful for classifying, creating, or modifying components of a Fare Collection System (FCS).

This chapter identifies six components of an FCS including fare structures, fare policies, credit plans, payment methods, ticketing technologies, and equipment. For each component, feasible options are listed and a comparison is made between their attributes. Sections are arranged in such a way that one can understand the key characteristics of a fare structure system and can match a desirable FCS by sequentially selecting the appropriate options from each component. Depending on what has chosen in a previous section, the options in the next section can be selected.

Fare Structures

The fare structure can be divided into two categories: flat (or uniform) and differential (or metered). In a flat structure, passengers pay a fixed rate regardless of their trip characteristics (e.g., Mexico City, Mexico, buses in London, UK, Adelaide and Canberra, Australia, Sao Paulo, and Rio de Janeiro, Brazil). In a differential structure, the fare may vary depending on trip characteristics such as time of day (peak/off peak), start/end location, or distance (TCRP, 2007, 2015, Mezghani, 2008).

The flat structure is easier for the passengers to understand and is more straightforward for the operator to implement than the differential structure. However, the differential structure can better apply the principles of social justice and customer care, as well as adapt a Travel Demand Management (TDM) policy, for example in monitoring and controlling long-distance trips (example distance-base systems in Sydney, Australia, and Amsterdam, the Netherlands; zone-based fares for regional travel in Brisbane, Australia, Dresden, Germany and London underground, UK). **Table 1** summarizes the attributes of flat and differential fare structures.

Another component that affects shaping the fare structure is the management of passenger transfers. The common transfer options include a free transfer (e.g., Hobart, Australia, between rail lines in Washington DC, USA and London, UK), a low-cost transfer, and not recognizing a transfer at all which means separate tickets are required for each and every trip. Transfer management is interconnected with the fare policy which is elaborated in the next section. One should note that all three types of transfer options can be adapted in a flat or a differential structure (TCRP, 2007).

Fare Policies

Maximizing revenue is one of the most important goals of any operator, and is directly related to the number of passengers and the amount of fare charged for each trip. A considerable part of the literature is devoted to studying the variation of demand with a change in the cost of transit trips which is usually carried out by means of mode choice modeling or an elasticity analysis. The operator's revenue is calculated as the net effect of the change in the number of passengers and the fare per passenger. The policies,

Table 1 Attributes of flat and differential fare structures

Fare Structure	Social Justice	Customer focused	Simplicity for managers	Attributes		Income adjustability	Controlling long trips
				Simplicity for passengers			
Flat	Low	Low	High	High		Low	Low
Differential	High	High	Low	Low		High	High

measures, and initiatives that operators use to either influence the number of passengers or adjust the fare usually in order to increase ridership or revenue are called fare policies. These policies may reduce operators' revenue especially when fares are cut and the number of passengers remain unchanged. However, if carefully designed, discount strategies may reduce the fare and increase the number of passengers and thus the revenue.

Examples of these strategies are "discounted fares," which include services designed to encourage a particular group of people such as students or senior citizens to use the public transportation system. Some operators also offer discounts for people using newer payment methods. For example, fare rates can vary depending on paying by a smart card or cash. These discounts may also apply to the specific time of day that public transport has spare capacity; for instance, the fare in off-peak hours may be lower than the peak hours. The off-peak hour discounts are discussed in details in the "credit plan" section.

Some changes in strategies are designed to encourage a higher public transport patronage. The three most commonly used strategies are "free or reduced fare zone," "elimination of fare zones," or "elimination of express/rail surcharge." In the first strategy, some operators introduce free (or discounted) areas such as the Central Business District (CBD) area to increase ridership (e.g., Seattle's downtown in 1973, Portland Oregon in 1975, and Albany New York in 1978). Boarding at the stations of these areas is free or very cheap. At many of these stations, the transit vehicles are crowded. Therefore, removing the fare collection requirement increases facility capacity and improves speed and reliability in these areas which are critically needed. Further to fare collection, inspection of fare is also costly to the operator and can be detrimental to the service performance.

In the case of "elimination of fare zones" or "elimination of express/rail surcharge," some costs such as a surcharge at peak times or a surcharge due to the use of express/rail services are removed or reduced. The abovementioned strategies can be effectively implemented in a differential fare structure because of its embedded flexibility. Within the differential structure, it is possible to define different fares at different times, stations, routes, and days, while in the flat fare structure, fares are already fixed and flat (TCRP, 2007).

Some strategies are easier to implement in a flat structure such as reducing the base fare. The base fare is the minimum price that is charged to every passenger regardless of the time of day and usage duration. This strategy is not common, and if not supported by sufficient studies, may incur a significant loss to the relevant service operators.

If the transfer charge is already low, a less-risky approach is the simultaneous implementation of a low-cost fare scheme and the elimination of a free (or a low-cost) transfer (see the fare structures section). Although eliminating the free transfer is easy to implement, it can also lead to a decline in the ridership (depending on the reduction amount in the base fare and the number of transfers) and reduce operator's revenue.

In the case that the current transfer charge is already high, a "transfer price reduction" strategy can be adapted to attract passengers. Removing or reducing transfer charges are easier in systems that issue one ticket for an entire trip (usually on boarding) or use electronic methods (see "payment methods") since the transfer charge can be deducted from the total fare when calculated by the system (TCRP, 2007). A partial removal of transfer charges is also possible when the "same mode transfer" is made free. Examples are train-train or BRT-BRT (Bus Rapid Transit) transfers since passengers can move in a close system without passing a payment gate to transfer within the same mode. Another common option to reduce (or remove) transfer charges is using a *duration-based* ticket. Hourly or daily tickets are examples of such an option by which passengers can travel for as many times as they want during the designated period. The complete range of duration-based tickets are reviewed in the "credit plans" section.

Credit Plans

General Credit Types

There are various types of credits issued for transit tickets. Each transit operator offers a range of tickets depending on the history of the ticketing system and how it is evolved as well as the existing "fare structure" and "fare policies." These credit plans can be classified to three general categories of duration-based, trip-based, and value based plans.

Duration-based tickets are travel permits for the validation period of the ticket. Unlimited trips are possible during this validation period which can be as short as one or two hours to a day or even to a few months. Common types are hourly tickets (usually two hours), daily, weekly, and monthly tickets. A longer period such a six-months or a yearly pass also fall in this category. These tickets can be activated from the moment of purchase or the first transaction. The use of this type of credit became more prevalent with the introduction of electronic payment options which is reviewed in the "ticketing technologies" section (TCRP, 2007, 2015).

Contrary to a duration-based ticket for which the time to use the ticket is limited, the other two types of tickets put a limit on the number of trips or the fare value itself. A trip-based ticket allows one to make a certain number of trips, but without limiting the time of those trips. Common types are return tickets (two-way tickets), ten-trips or the like. This type of ticket can be an attractive option for people who use transit without a fixed schedule or make long-distance trips in a differential fare structure. Value-based tickets provide passengers with a credit to pay in a complex fare structure. By using the ticket, the trip cost will be deducted from the available credit. A common type of this ticket is the smart card technology that attributes a credit value to the card and adjusts it with each transaction. Duration-based and trip-based credits are more compatible with a flat fare structure while the value-based credit can better be adapted in a differential fare structure.

The above credit types can be combined in one ticket too. For example, combined duration-based and value-based options are one of the types of tickets that can operate in a multimodal system. In the metro network, the cost can be paid by credit based on zone and distance, and on the local buses based on time. In other words, from the passenger's perspective, one ticket is used in all modes but the operators may charge it as a duration-based, trip-based, or value-based (TCRP, 2007, 2015).

Special Tickets and Credits

In addition to the above categorization, credit plans and respective ticket types can be viewed from other aspects including off-peak tickets, group tickets, special event tickets, special strata tickets, tickets promoting a specific ticketing technology, and advanced pattern-specific promotions.

Off-peak tickets are offered at a cheaper rate during the off-peak periods when the transit vehicles are less crowded. These types of tickets may encourage people to travel at specific times of the day, weekends, or a combination. Where more than one type of off-peak ticket is offered, cheaper tickets have more restrictions. These cheaper tickets are sometimes called super off-peak (National rail of UK, 2019).

Another type of special credit ticket is a group ticket. A group ticket is a discounted ticket that some transport operators offer to a group of passengers traveling together. These discounts are mainly offered during the off-peak times, and the carrier determines the minimum and the maximum number of the group members that can receive this type of ticket.

A special event ticket is offered in conjunction with other purchased services (such as sports matches, community festivals, fairs, and off-site parking areas) to allow a convenient access to the destination by transit, rather than using a personal vehicle. Special event tickets are not only a Transport Demand Management (TDM) strategy to mitigate crowding of a big event, but also an opportunity to attract new customers to the transit system who do not use transit in their weekly routine.

Special strata tickets are offered to a special sociodemographic class of citizens. The common strata are children, students, and senior citizens/pensioners. Special strata tickets can easily be incorporated into credit plan categories of duration-based, trip-based, and value-based either by issuing a special strata ticket or when a credit is loaded to the ticket.

The next aspect on the credit plan is its capability to promote a ticketing technology. A variety of discounts and concessions are proposed to encourage passengers to buy a *fare card* (see "ticketing technology"). One of the most common of these concessions is that the fare card is preloaded with an initial credit. Second and more importantly, the fare can be cheaper if paid by a fare card compared to cash. Third, a guaranteed last ride (or a negative balance) can be included in a ticket. In this case, if the passenger's fare card does not have enough balance to pay for the trip, a negative balance is registered to enable the passenger to make the trip. The passenger should load the card for future trips. One reason to promote a fare card over cash is the higher efficiency of fare payment by a fare card since the average boarding time by a fare card is much shorter than by cash. This reduces vehicle dwell time at stations and can potentially increase transit facility capacity and speed (TCRP, 2007).

There are other advanced options available that are specific to some trip patterns. One of these pattern-specific options is refunding the passengers for a partial cost of their trip. The refund is loaded as credit to encourage them to make more trips within a certain period of time. This bonus credit can expire after a defined period if not used. A similar option is offering free trips if the number of trips exceeds a given threshold in a week (also called a frequency-based discount). A third example is called a "guaranteed lowest fare" plan which is applicable to a fare system that offers multiple promotions such as off-peak, frequency based, or special event tickets. The guaranteed lowest fare ensures that passengers are charged the minimum amount while considering the combination of all promotions. A value-based fare structure enables operators to design complex marketing schemes such as the above pattern-specific promotions to increase ridership and revenue.

Any of the duration-based, trip-based, and value-based credit plans to some extent can adopt the options of off-peak tickets, group tickets, special event tickets, special strata tickets, tickets promoting a specific ticketing technology, and pattern-specific promotions.

Payment Methods

This section reviews the fare collection options and defines its relation to the fare structure, credit plan, and ticketing technology. Payment methods can be classified with respect to place, time, and mode of payment as depicted in Figure 1.

In terms of place of payment, passengers may pay before boarding (off-board) or may pay in-vehicle after boarding (on-board). The method of payment and the technology used have an important effect on the capacity of a transit facility as well as other performance indicators (TCRP, 2017). Thus, depending on the transit authority's policy and the available budget, the best payment

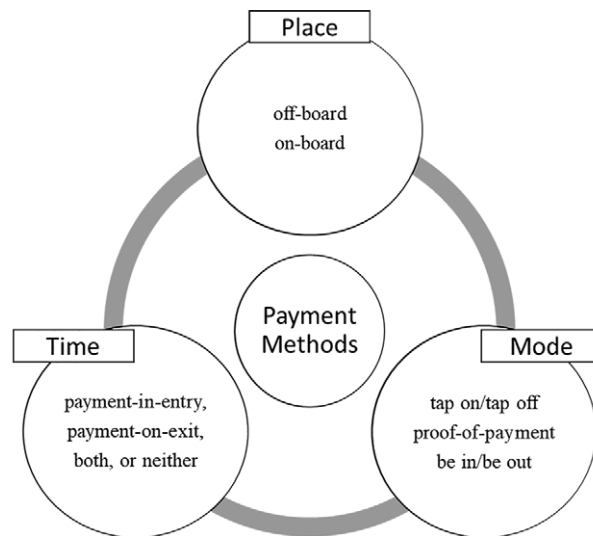


Figure 1 Classification of payment methods.

Table 2 Payment modes and their properties

Mode	Passengers	System/Operator
Tap on/tap off (TOTO)	Presenting tickets (active)	Checking tickets (active)
Proof-of-payment	Not presenting tickets (passive)	Not checking tickets (passive)
Be in/be out (BIBO)	Not presenting tickets (passive)	Checking tickets (active)
Not an available mode	Presenting tickets (active)	Not checking tickets (passive)

method can be selected. Each of the payment methods have their own advantages and limitations. In general, off-board payment methods enable the transit facility to offer a faster boarding and thus a higher capacity than the on-board methods. In return for the high capacity, the off-board methods occupy a larger station space in which the fare collection equipment should be installed. Having the equipment installed at stations, reduces the flexibility of the transit line to relocate stops. As for the on-board method, monitoring the payment is challenging if multiple doors are available and boarding is allowed from all doors (such as the case of a tram/streetcar). Both payment methods can be implemented in a flat or a differential fare structure.

In terms of payment time, the boarding, in-vehicle, and alighting phases should be considered. Some systems may need to check passengers on entry, on exit, on both, or neither. If a flat rate is charged (i.e., a flat fare structure), it is sufficient to check the fare on entry (called pay-on-entry) or on exit (called pay-on-exit) while the former is more common around the world (Currie and Reynolds, 2016). If a differential rate based on distance or zones is charged (i.e., a differential fare structure), passengers should be checked on both entry and exit to calculate the fare. There is a special payment system that does not require passengers to present their ticket at all, which means neither at the entry nor at the exit. Proof-of-payment allows passengers to acquire a form of ticket and to carry it at all times onboard. The operators only use random checks by a few inspectors to enforce ticket purchasing. Pay-on-entry, pay-on-exit, or differential payment which requires checking at both entry and exit, can happen off-board or on-board.

The mode of payment divides the payment methods to “tap on (or check in)/tap off (or check out)” (TOTO or CICO), proof-of-payment, and “be in/be out” (BIBO) based on what roles are played by the passengers and the fare collection system. TOTO is when passengers need to present their ticket actively and the system checks all tickets. Proof-of-payment does not require passengers to present their ticket nor the system to check all passengers. Only random checks may be performed by inspectors. The third mode of payment is BIBO in which passengers do not present a ticket actively, but the system checks all passengers using the BIBO technology (e.g., in Dresden, Germany). Special gates at the entry and/or exit points monitor whether or not passengers carry a valid ticket without asking them to take the ticket out of their pocket. Since it does not make sense for the passengers to present the ticket actively but for the system not to check the tickets, a fourth type of payment mode is not defined (TCRP, 2015). Table 2 summarizes these payment modes and their properties.

A “tap on” or “be in” is utilized when a pay-on-entry fare is charged. Similarly, a “tap off” or “be out” corresponds to pay-on-exit. It is also possible to have a hybrid system with “tap on” at the entry and “be out” at the exit. A TOTO or a BIBO matches a differential fare structure when both entry and exits are to be recoded.

Ticketing Technologies

In decades, various technologies have been developed to collect transit fares. Implementation of an efficient fare system relies on choosing the right ticketing technology. Some of the capabilities listed in the previous sections can only be expected from a certain technology. In this section, a wide range of technologies are reviewed to enable the transit agencies to assess them in conjunction with the agency's goals and objectives, technology availability, and budget. This section covers from traditional methods of cash, paper ticket, and ticket token to more modern tickets of magnetic cards, smart cards, Radio Frequency Identification (RFID), Near Field Communication (NFC), Short Message Service (SMS) tickets, QR code tickets, and smartphone ticketing.

- Paper ticket and cash: The fare could be paid by purchasing a paper ticket or a token before boarding and presenting the ticket to ride a transit vehicle. Alternatively, the fare could be paid by cash to board. Paying cash increases the boarding time compared to a paper ticket since the station or vehicle operator may need to return a change. These traditional methods work well with a flat fare structure although examples of systems accepting cash in a differential fare structure exist around the world (e.g., Brisbane, Australia). A cash accepting system is especially convenient for tourists given that they can readily board a transit vehicle without prepurchasing any form of ticket.
- Magnetic card: Magnetic cards are a simple paper (or plastic) card with a magnetic strip which make a transaction when passed through a reader slot. Magnetic cards generally come in two types: read only and read & write. The credit of read-only cards are set when issued and the credit can only be deduced from them. They can be used with the "credit plans" of duration-based and trip-based. Read & write cards can be used with all "credit plans" including a value-based plan given that credit can be deduced or added on them. Although magnetic cards are a relatively cheap ticket to produce, they are not durable and can easily malfunction if folded or exposed to a magnetic field. These cards were introduced in periods on the London underground and the Long Island Railroad (Mezghani, 2008, TCRP, 2015).
- Smart card: These fare cards have an electronic chip in them that can store and encode a considerable amount of information such as the card holder's identity, credit value, and trip history. The three main types of smart cards are contact, contactless, and proximity (BIBO) cards. The contact cards need to be inserted into a reader while the contactless cards can be read when placed closed to a reader (a range of 5 cm). Proximity cards can be read when the card holder is in the vicinity of its special reader (a BIBO reader). Proximity cards can also work as a contactless card if a special BIBO reader is not available. (The World Bank, 2011, TCRP, 2015, Faroqi et al., 2018, Mezghani, 2008). Both magnetic and smart card technologies can cover a variety of credit options. However, magnetic cards are cheaper than smart cards, but their reader equipment is more expensive to maintain. The storage and processing capacity of smart cards are significantly higher than magnetic cards. Generally, magnetic cards are easier to sell, while smart cards are easier to use (Mezghani, 2008).
- RFID (Radio Frequency Identification) technology: RFID tags fares are functionally similar to proximity smart cards since both can be read when placed in the vicinity of a reader/gate. However, RFIDs are mainly Read-only, and their radio-based barcodes are mostly used for identification. They are cheaper, have much less memory, and are smaller in size than smart cards. However, smart cards are more secure than RFID tags. (Octopus cards in Hong Kong) (Mezghani, 2008).
- NFC (Near Field Communication) technology: The NFC technology allows an electronic device to communicate with a reader in a short range of 2–10 cm. Similar to a contactless smart card, the exchange of data is contactless and can happen in a short range. NFC can perform a two-way transmission between the reader (receiver) and the device (transmitter). The technology is powered and is already installed on some smart devices, especially smart phones. This enables NFC to be programmed by smartphone apps, while smart cards and RFID tags are unpowered and rely on the master device for any interaction (Secure Tech Alliance, Chandler, 2014).
- SMS (Short Message Service) tickets: Short Message Service (SMS) is a standard communication service on the fixed and mobile phones. A SMS ticket can be issued upon request from the ticketing server by an SMS text or an Unstructured Supplementary Service Data (USSD) service. The text can be used as an electronic travel document to be presented to the transit ticket inspector (Florida Department of Transportation BDV, 2016).
- QR Code tickets: QR code is a two dimensional barcode that can encode some information. The ticket information (e.g., its value and validation time) can be converted to a QR code and given to a passenger on a smart device or printed on paper. The QR code reader, by reading this code, can access the stored information and allow an entry. QR Code tickets can be issued by web services, smartphone applications, and ticketing machines.
- Smart phone ticketing: In addition to NFC, SMS, and QR code that are available on smartphones and some other media, some ticketing options are exclusive to smartphones since they have integrated abilities of processing power, a large memory, and network connection means. This allow travelers to book trips and get tickets through smartphone applications (e.g., the contactless card or smart phone cashless buses in London) (Florida Department of Transportation BDV, 2016). A specially interesting case is the emerging concept of Mobility as a Service (MaaS). Passengers may purchase their tickets from a MaaS smartphone application and present it to board a service if the transit infrastructure accepts one of the available technologies on smartphones. For more information on MaaS please see the MaaS Entry (Florida Department of Transportation BDV, 2016).

Equipment

A fare collection system requires equipment to issue a ticket, collect a ticket, and recharge users' accounts. The type of this equipment is determined by the ticketing technology, the payment method, and the fare structure. The equipment is a part of the transit infrastructure system and plays a key role in the quality of services provided to passengers. Equipment types are as follows:

- Ticket vending machine: It is a device that sells tickets and if applicable, recharges passengers' tickets. Depending on the type of ticketing technology, vending machines offers compatible services to passengers. These devices are usually placed in main transit stations.
- Online ticket service: with the dominance of web-based and mobile ticketing systems, online sale of tickets is an emerging venue. A notable example in this area is Software as a Service (SaaS) which is a business model to contract with a transit operator to arrange the sale of tickets through a standard mobile application. The vendor usually receives a portion of sales in return for the service.
- Fare box: A device at the station (off-board payment) or in-vehicle (on-board payment) that is used by passengers to pay the fare.
- Fare gate: A fare gate requires passengers to validate their ticket before they are allowed to pass. A fare gate is usually used off-board and is compatible with magnetic card, smart card, QR code, and smart phone ticketing.
- Fare gantries: A fare gantry allows passengers to seamlessly get through the gantry and automatically pay the travel fare. It is a replacement for a farebox or a fare gate and is compatible with a BIBO payment method. It can be applied in off-board and on-board systems.

Integration of Fare Payment

A transit network in a conurbation may have several modes, operators, or even jurisdictions. On a daily commute, one passenger may use multiple services across those divisions but it is inconvenient for passengers to acquire multiple tickets. From the passengers' perspective, a door to door mobility service is expected and it is not important how it is operated and managed in the background. Therefore, many transit authorities offer an integrated fare payment system that mediates between the passengers and the operators (or even jurisdictions). The integration is made possible with the advent of new ticketing technologies such as smart cards and smartphone ticketing. However, integration of a fare system does not mean integrating all of its components. With a flexible fare system, operators can partly retain their own fare structure, credit plans, or even payment methods (see the respective sections above). However, a consistent ticketing technology and equipment should be used. For example, using a smart card (the same ticketing technology) and a set of fare boxes and fare gates (the same equipment), passengers may pay a flat fare on buses but a differential fare on metro (different fare structures) (TCRP, 2007).

Integration of fare payments provides passengers with a higher convenience in purchasing and presenting tickets, easier transfers, and increased capacity, speed, and reliability for the services due to shortening the dwell time and its variability. The operators also benefit from the integration since a higher patronage leads to an increased revenue, a higher capacity, speed, and reliability can reduce operating costs and increase passenger satisfaction. The general public would also benefit from the potential congestion relief, environmental improvements, a better quality of life, and economic efficiency resulted from an improve public transport system with a higher patronage.

Conclusions

In this chapter, a transit fare collection system is characterized by identifying six components namely fare structure, fare policy, credit plan, payment method, ticketing technology, and equipment. The goal was to present an inclusive taxonomy which provides a unified basis to compare existing and future fare collection systems. The six components are presented in order, that means a reader should first determine the fare structure (flat or differential), then the credit plan (duration-based, trip-based, or value-based) and the payment equipment (off/on-board, pay-on-entry/exit, TOTO/proof-of-payment/BIBO). After that, the ticketing technology (from paper tickets to smartphone apps) and the equipment (from a humble farebox to fare gantries) are to be defined. The final remarks highlighted the importance of moving toward an integrated fare collection system in which passengers can travel seamlessly.

References

- Chandler, N., 2014. *What's the difference between RFID and NFC?* [Online]. Available from: <https://electronics.howstuffworks.com/difference-between-rfid-and-nfc1.htm>.
- Currie, G., Reynolds, J., 2016. Evaluating pay-on-entry versus proof-of-payment ticketing in light rail transit. *Transp. Res. Rec.* 2540, 39–45.
- Faroqi, H., Mesbah, M., Kim, J., 2018. Applications of transit smart cards beyond a fare collection tool: A literature review. *Adv. Transp. Stud.* 45, 107–122.
- Florida Department of Transportation BDV, 2016. Assess of Mobile Fare Payment Technology for Future Deployment in Florida.
- Mezghani, M., 2008. Study on electronic ticketing in public transport. European metropolitan transport authorities.
- National rail of UK, 2019. *ticket types* [Online]. Available from: https://www.nationalrail.co.uk/times_fares/ticket_types/46548.aspx.
- Secure Tech Alliance. *NFC Resources* [Online]. Available from: <https://www.securetechalliance.org/smart-cards-applications-nfc/>.
- TCRP, 2007. TCRP REPORT 111 (Elements Needed to Create High Ridership Transit Systems). The Federal Transit Administration.
- TCRP, 2015. TCRP REPORT 177 (Preliminary Strategic Analysis of Next Generation Fare Payment Systems for Public Transportation). The Federal Transit Administration.
- TCRP, 2017. Transit Capacity and Quality of Service Manual, 3rd Ed.-Ch03_Operations-Concepts. Transportation Research Board.
- The World Bank, 2011. *Toolkit on Fare Collection Systems for Urban Passenger Transport; Smart card* [Online]. Available from: <https://www.ssatp.org/sites/ssatp/files/publications/Toolkits/Fares%20Toolkit%20content/fare-collection-technologies/smart-cards.html>.

Transit Information Systems

Ankit Kumar Yadav, Nagendra R Velaga, Transportation Systems Engineering, Department of Civil Engineering, Indian Institute of Technology (IIT) Bombay, Powai, Mumbai, India

© 2021 Elsevier Ltd. All rights reserved.

Glossary	331
Introduction	331
Factors Influencing the Transit-Oriented Decisions of Passengers	331
Information Demand–Supply Trade-Off of RTTIS	332
Key Impacts of Transit Information Systems on Travel Behavior	333
Waiting Time/Travel Time	333
Transit Use	334
Perception of Safety and Personal Security	334
Personal Satisfaction With Transit Service	334
Psychological Effects of RTTIS on Ridership	334
Social Effects of RTTIS	335
Technological Advancements in Transit Information Systems	335
Conclusions and Directions for Future Research	336
References	336
Further Reading	337

Glossary

DMS dynamic message signs

ICT information and communication technologies

Path choice route taken by the passengers to minimize their travel time during the journey

RTTIS real-time transit information systems. They include all types of transit information providing systems such as RTI display boards, web-enabled services and mobile devices

Ridership the number of passengers using a particular mode of transit

RTI S Display Boards used to provide real-time information to passengers through display boards at transit stops and stations

Transit providers agencies or organizations which facilitate the transit services

Introduction

In the present digital era, technological advancements are promoting the transit industry across the world and assisting the passengers by providing appropriate transit information (such as route details, traffic conditions, schedule of availability, time and fare details). Transit information enables the users to make travel choices according to their suitability. The effectiveness of “transit information system” relies on the availability of real-time information about the position as well as prediction of arrival time of the transit mode at a particular place of access or station. The real-time transit information (RTTI) can be distributed in various forms such as mobile phones, roadside kiosks at stations and dynamic message signs (DMS). Due to the information and communication technologies (ICT)-based dissemination of real-time transit information to the passengers, they can access the transit details at anytime and anywhere during their travel (Papangelis et al., 2013). It has made travel flexible as travel decisions can be taken/modified at any stage of the trip (i.e. pre-trip, en-route, and post-trip) since the transit information acquisition in real time is possible.

Factors Influencing the Transit-Oriented Decisions of Passengers

Fig. 1 shows the complex links between the various factors influencing the behavioral responses of the travelers (Zhang, 2010). Along with the real-time transit information obtained from the transit service providers, demographic factors (such as age, income, car ownership, etc.) and situational factors (such as weather, environment, etc.) influence the transit-oriented behavior/decisions of the passenger as well as their psychological responses. The positive attitudinal changes and positive perceptions toward transit services are advantageous to the smooth functioning of transportation systems. For these positive changes to occur, travelers might be able to obtain the real-time information easily, and subsequently the obtained information would lead to favorable psychological and behavioral results.

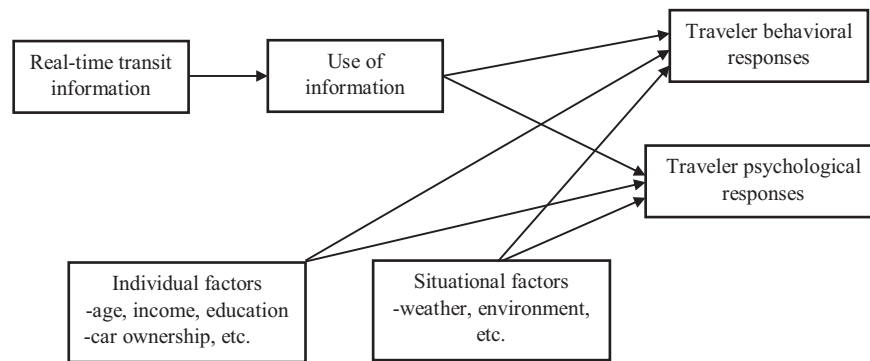


Figure 1 Factors influencing the traveler behavioural responses (Zhang, 2010).

Information Demand–Supply Trade-Off of RTTIS

The efficiency of transit information systems would be maximum when the information demand of passengers meets the information supply from the transit providers (Velaga et al., 2012a). With respect to information dissemination to the passengers, dynamic message signs (DMS) were found to be the most popular among the different methods in Ireland (Caulfield and O'Mahony, 2007). They also observed that the use of RTTIS increased with the increase in trip complexity. A study in Germany investigated the gap between information demand and supply by interviewing the transit service providers and conducting a survey among the passengers (Beul-Leusmann et al., 2013). They found that majority of the passengers rely on websites to obtain real-time information with respect to transit services, and also criticized the methods of providing transit-related information by the transit providers. Research on the information supply side showed that factors such as agency size, demographics, modal service and increasing competition in the market influenced the information dissemination to its customers. Other reasons that were observed over time were lack of funding and technical expertise (American Public Transportation Association, 2015).

However, information demand and supply preferences are likely to change with time and growth in technology. A recent study in the United States examined the user preferences in the information demand and identified the most preferred type of information required by the passengers, along with preferred medium of information dissemination (Harmony and Gayah, 2017). Furthermore, the extent to which the transit providers were able to satisfy the passenger needs was investigated along with the reasons for the differences observed in demand and supply of transit information. The observed comparisons of information demand and supply are summarized in Table 1 (Harmony and Gayah, 2017). The first column in the table lists out the information demand of the

Table 1 Comparison of transit-related information demand and supply (Harmony and Gayah, 2017)

Information type	Comments	Adequately meets needs?
Arrival/departure times for next vehicle	The most desired type of information by passengers is the best provided information by agencies	Yes
Real-time location of transit vehicles	- Information well provided for in many different ways - Highly valued by passengers and provided by many agencies - The two most desired mediums for providing information are the two best supported mediums provided by agencies	Yes
Seating availability	- Relatively low importance for passengers but more important for bus passengers; hence too few agencies offer information - More information needs to be offered on smartphones and DMS	No
Number of cars on next train or type of next bus	- One of the least valued types of information and least provided by agencies - Information is provided through multiple platforms when it is available	Yes
Real-time trip planning	- Very important to passengers and many agencies provide information - Well provided for in most desired mediums for providing this information	Yes
Information for planned detours and special events	- Very important to passengers and many agencies provide information - More information needs to be provided on DMS	No
Identification of unplanned service disruptions	- Very important to passengers and many agencies provide information but more need to do so - More information needs to be provided on smartphone applications	No
Emergency information	- Somewhat important information that is well provided by agencies - Needs to be better provided on DMS	No
Information on availability and operational status of elevators/escalators	- Lower importance by passengers but it is relatively well provided by agencies - Information needs to be better provided for on smartphone applications	No
Parking availability	- One of the least important types of information and least provided by agencies - Information needs to be better provided for on smartphone applications	No

passengers. The second column describes its order of preference/priority and the third column reports whether these information demands are satisfied by the transit providers. It was found that information related to the arrival/departure times of transit vehicle was the most desired information by the transit users; whereas information on seating availability in the vehicle was the lowest desired information (Harmony and Gayah, 2017). With respect to the medium of providing information, smartphones were the most preferred medium followed by websites and dynamic message signs. The comparison presented in Table 1 showed that the transit-related information supply by service providers is mostly in consensus with the priority of information demanded by the passengers (Harmony and Gayah, 2017). Nevertheless, some information gaps exist which are required to be bridged with time to improve the transit efficiency. These gaps may be identified as the information with respect to seat availability, emergency information, details about any detours, availability of elevators, etc. The implementation of RTTIS can be improved by maximizing the public involvement by considering their feedback about their experience and issues faced, along with working on the alternative sources of providing information.

Key Impacts of Transit Information Systems on Travel Behavior

RTTIS is known to influence the travel decisions of commuters, thereby impacting the decision to undertake a trip, to select a preferred mode of transit over the other modes, and to select a designated path to reach the destination. Moreover, it also affects the decisions to board the transit as well as alight from the transit during the trip. All these decision-making procedures with respect to RTTIS can be compiled in a trip choice framework from the user perspective as presented in Fig. 2. The framework shows the various influencing factors and behavioral outcomes which are likely to be influenced by RTTIS services (Brakewood and Watkins, 2019). All of them are discussed in detail in the subsequent sub-sections.

Waiting Time/Travel Time

The large body of research shows that RTI display boards reduce the perceived waiting time; whereas, RTI facilities provided through personal devices reduce the perceived as well as actual waiting time since the users can stay at their place of residence or work, and leave to their pickup stop when the transit facility is about to arrive (Velaga et al., 2014). The relationship between the RTTIS and the waiting time of riders have been examined in various studies (Brakewood et al., 2014; Dziekan and Vermeulen, 2006; Watkins et al., 2011). A study by Watkins et al. (2011) investigated the perceived waiting times of RTI users and non-RTI users at bus stops in Washington (USA), and reported that self-reported average waiting time for RTI users was 30% less than the riders who did not use RTI facilities. Another study by Dziekan and Vermeulen (2006) in Hague (Netherlands) analyzed the change in waiting times of passengers at a tramline before and after the installation of RTI systems. They found about 20% decrement in the waiting time of passengers after the RTI system was installed.

The passengers tend to select the shortest path from their origin to destination; this information about the optimum path choice (or route choice) can be obtained using RTTIS facilities. These facilities are more useful in cases where transit services do not comply with a particular fixed schedule. In such cases, passengers can use RTTIS and make a decision of selecting an alternative mode/route of transit if the scheduled mode/route is not available. The extent to which RTTIS influence the path choice and associated travel time of passengers has been studied in past (Cats et al., 2011; Estrada et al., 2015; Hickman and Wilson, 1995). A simulation study in Uruguay revealed that RTTIS has the potential to reduce the total travel times up to the extent of 45% compared to unavailable time

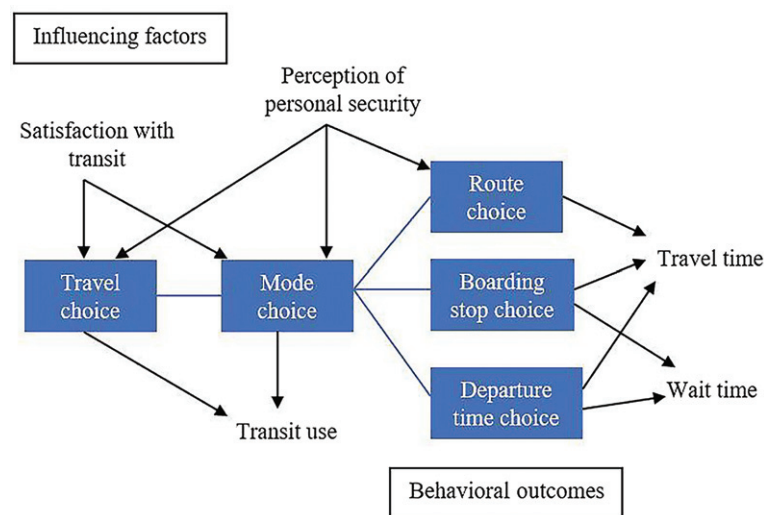


Figure 2 Influence of RTTIS on transit service (Brakewood and Watkins, 2019).

schedules of buses (Estrada et al., 2015). A study conducted in Sweden on subway transit service reported that implementation of comprehensive RTTIS facilities resulted in savings in total travel times (Cats et al., 2011).

The research indicates that savings in total travel time of passengers using RTTIS services range from 3% to 45%. However, there is huge scope of conducting future research where travel time reductions can be analyzed with different frequencies of transit service, and different levels of network coverage and route options (such as sparse networks versus dense networks).

Transit Use

Due to the implementation of RTTIS applications, reduced waiting times and efficient route choices minimizing the total travel time would eventually lead to an increased use of the transit mode. Moreover, new riders may also be attracted toward the use of the transit systems. Various researchers had observed significant increase in the ridership of RTTIS users in transit vehicles such as buses and trains (Ferris et al., 2010; Ge et al., 2017; Gooze et al., 2013; Politis et al., 2010). After the installation of RTI display boards on an office building in Seattle, transit users were found to increase by about 4% (Ge et al., 2017). With respect to RTI services provided on personal devices, 30% users reported higher frequency of noncommute trips and 10% reported higher commute trips (Ferris et al., 2010; Gooze et al., 2013). Another study in Greece observed 19.7% of the people mentioning that they have made new trips after the implementation of RTI display boards (Politis et al., 2010).

Some survey-based studies reported that significant proportion of respondents would like to increase the frequency of transit use if the RTTIS services are provided at transit stops/stations (Tang and Thakuriah, 2011, 2007). A stated preference survey was conducted in Chicago (USA) before the implementation of RTTIS, where about 76% of the participants were willing to increase their use of transit when the RTTIS services were provided (Tang and Thakuriah, 2011). In a survey-based study, it was reported that 67% of the participants were willing to increase the frequency of transit use if the transit stations were equipped with RTI display boards (Tang and Thakuriah, 2007). However, several other studies did not find any significant impact of RTTIS services on the increase in transit use (Brakewood et al., 2014; Zhang et al., 2008). A major limitation observed in the past studies is that they are mostly focused on larger cities such as Chicago, New York and Seattle. Future investigations are required to emphasize on medium and small-scale cities to examine the influence of RTTIS on transit use among the existing users as well as the new users.

Perception of Safety and Personal Security

As RTTIS enables the commuters to devote less time for waiting at transit stops/stations, it is likely to increase the perception toward safety and personal security while using transit facility, especially during night time. A behavioral study on passengers using a bus as a transit mode in Tampa, USA reported that safety perceptions were significantly higher for RTTIS users with respect to the control group during day time. However, no such significant increment in perceived safety was observed during night time (Brakewood et al., 2014). Alternatively, a study in the University of Maryland, USA reported statistically significant positive influence of RTTIS on personal security feelings during night time, but no such significance was observed during daytime (Zhang et al., 2008).

A few web-based surveys enquired about the safety perceptions of RTI users. A study by Ferris found that 21% respondents felt safer by using RTI facilities on their personal devices (Ferris et al., 2010) while the other study reported 32% of the users with increased safety perceptions (Gooze et al., 2013). Another study in Hague, Netherland analyzed the passengers' security ratings (on a scale of 1 to 10) at tram stations, and observed no significant difference in the security ratings before and after the implementation of RTI display boards at the stations (Dziekan and Vermeulen, 2006).

Personal Satisfaction with Transit Service

Reduced travel times and waiting times at transit stations due to RTTIS may also lead to increased individual satisfaction with the transit service. The effects of RTTIS on personal satisfaction with transit services had been widely explored by various researchers (Brakewood et al., 2014; Ferris et al., 2010; Ge et al., 2017; Gooze et al., 2013; Zhang et al., 2008). However, the literature has observed mixed findings with respect to personal satisfaction, where majority of the studies reported significant increase in personal satisfaction with transit service after RTTIS implementation (Brakewood et al., 2014; Ferris et al., 2010; Gooze et al., 2013; Zhang et al., 2008) with an exception of a study which observed decrease in personal satisfaction after RTI services were provided (Ge et al., 2017).

Psychological Effects of RTTIS on Ridership

The influence of Real-time transit information systems (RTTIS) on the ridership are generally studied by comparing the transit ridership statistics prior and post implementation of transit information systems. The ridership effects have been observed positive on the routes where RTTIS was implemented. However, analyzing the before and after effects of RTTIS may not be sufficient to successfully predict the variations in ridership as it depends on various factors such as transit fares, fuel price, population changes, etc. which fluctuates with time and may influence the ridership.

Positive psychological effects of RTTIS have been observed on the commuters, such as reduced anxieties, higher safety perceptions, and higher satisfaction levels. To investigate the effects of cognitive psychology (i.e., attitudinal and behavioral factors) along

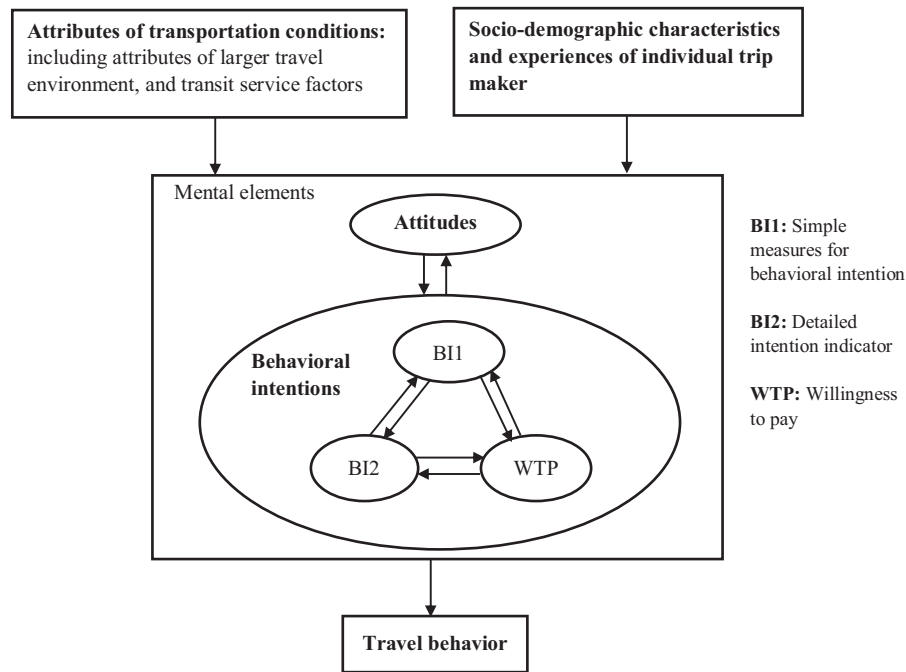


Figure 3 Relationship between attitudes, behavioral intentions and travel behavior (Tang and Thakuriah, 2011).

with transit service-related factors and socioeconomic factors on ridership, researchers have developed conceptual frameworks, such as one shown in Fig. 3 (Tang and Thakuriah, 2011). Their findings suggested that RTTIS is successful in increasing the transit ridership and has potential in serving as an intervention for breaking the travel habits of transit nonusers', thereby increasing the transit use mode share. Passengers' attitude toward travel time is a significant factor influencing the psychological decisions related to the use of transit service. Moreover, past experience with RTTIS also influences their intentions of using the transit service. Therefore, by incorporating methods which enhance the exposure of RTTIS facilities to the commuters to familiarize them with the usage, it can prove to be even more beneficial. Also, the information provided to them should be easily understandable, effortlessly digestible, more importantly reliable, and should be able to answer the demands of commuters.

Social Effects of RTTIS

RTTIS improves the accessibility to transit services for individuals with disabilities and empowers them to be independent and enhances their quality of life. Since disabled people have various physical and environmental constraints, use of emerging technologies helps them in adapting to the smoother way of travel. Especially, the increasing use of smartphones has resulted in increased use of transit related applications which provide real-time information related to transit services to the users. These applications use the smartphone location and provide an estimate on the expected time of arrival or departure of the transit. Consequently, they are quite useful for the disabled population as well; however, use of these applications demand the users to be well-equipped with the use of technology. In a recent report on accessibility of disabled population in the United States, it was found that the count of commuters with certain disabilities using public transit increased rapidly compared to the other non-disabled users during 2005–2015 (National Council on Disability, 2015). However, more efforts are required to provide supplementary disabled-friendly services with special emphasis on the disabled population which is likely to reduce the accessibility barriers and assist them during their travel.

Technological Advancements in Transit Information Systems

Ridership data related to origin and destination of trips are important for transit agencies to understand the travel patterns of transit users. One of the widely used methods to obtain this information is through linking open data and crowdsourcing with the help of smartphone-based applications (Corsar et al., 2017; Traut and Steinfeld, 2019). The ridership data is collected when the transit users select their place of origin and destination, along with transit stops, and available mode while using the applications. Moreover, the origin-destination and transfer data of transit users can also be extracted from the smartcards which are used for the purpose of automated fare collection (Pelletier et al., 2011).

Conclusions and Directions for Future Research

This chapter provides a distinguished review on the impacts of real-time transit information systems on passengers, their travel behavior and ridership. On one hand, the positive sides of RTTIS makes transit easier and comfortable for the commuters; on the other hand, the information demand-supply difference exists which needs to be bridged in order to improve the efficiency and effectiveness of the overall transit system.

Various stakeholders are indulged at each stage of providing real-time transit information (RTTI) to the users, i.e. during the development, dissemination and operation stages. These stakeholders decide the nature of information demand (i.e. what information is required by the users) and information supply (i.e. what information needs to be provided to the users). However, researchers have examined both the information demand and information supply separately; therefore, the investigation on trade-off between the demand and supply of information may serve as a scope for future research. Furthermore, the comparison of user benefits from RTTIS with the transit service provider costs of providing RTTIS (i.e. cost-benefit analysis) is an interesting area for future research. Another direction for future research could be to investigate the situation-wise impacts of RTTIS on the transit ridership, for instance, with increasing transit frequency and service availability. With increasing usage of smartphone applications for connecting to transit service providers, the users provide their location data from which their travel patterns are estimated. The influence of travel behavior of commuters on the area-wise real-time traffic crowding/traffic calming has scope for future investigations.

Uber transit is a recent feature provided by the Uber in various cities around the world, with the aim to improve the accessibility of public transit by providing real-time information of the availability of the transit service at nearby transit stops. For instance, in the city of Denver, Uber riders can plan their subsequent transit ride by following the directions available in the Uber app. This integration of public transit information in the applications of aggregator taxi services is a positive step toward the increased use of transit service; however, the influence of such technology on the transit ridership shall be an important area for future investigations.

References

- American Public Transportation Association, 2015. Update on Public Transportation Agencies Providing Real – Time Data. Washington, DC.
- Beul-Leusmann, S., Jakobs, E.M., Zieffe, M., 2013. User-centered design of passenger information systems. In: IEEE International Professional Communication Conference, Vancouver, BC, pp. 1–8.
- Brakewood, C., Barbeau, S., Watkins, K., 2014. An experiment evaluating the impacts of real-time transit information on bus riders in Tampa, Florida. *Transport. Res. A: Policy Pract.* 69, 409–422.
- Brakewood, C., Watkins, K., 2019. A literature review of the passenger benefits of real-time transit information. *Transport Rev.* 39 (3), 327–356.
- Cats, O., Koutsopoulos, H., Burghout, W., Toledo, T., 2011. Effect of real-time transit information on dynamic path choice of passengers. *Transport. Res. Rec.: J. Transport. Res. Board* 2217, 46–54.
- Caulfield, B., O'Mahony, M., 2007. An examination of the public transport information requirements of users. *IEEE Trans. Intell. Transp. Syst.* 8 (1), 21–30.
- Corsar, D., Edwards, P., Nelson, J.D., Baillie, C., Kapangelis, P., Velaga, N.R., 2017. Linking open data and the crowd for real-time passenger information. *J. Web Semantics* 43 (1), 18–24.
- Dziekan, K., Vermeulen, A., 2006. Psychological effects of and design preferences for real-time information displays. *J. Public Transport.* 9 (1), 1–19.
- Estrada, M., Giesen, R., Mauttone, A., Nacelle, E., Segura, L., 2015. Experimental evaluation of realtime information services in transit systems from the perspective of users. In: *Proceedings of the Conference on Advanced Systems in Public Transport (CAPST)*, pp. 1–20.
- Ferris, B., Watkins, K., Borning, A., 2010. Onebusaway: results from providing real-time arrival information for public transit. *Proc. CHI* 1807–1816.
- Ge, Y., Jabbari, P., MacKenzie, D., Tao, J., 2017. Effects of a public real-time multi-modal transportation information display on travel behavior and attitudes. *J. Public Transport.* 20 (2), 40–65.
- Gooze, A., Watkins, K., Borning, A., 2013. Benefits of real-time transit information and impacts of data accuracy on rider experience. *Transport. Res. Rec.: J. Transport. Res. Board* 2351, 95–103.
- Harmony, X., Gayah, V., 2017. Evaluation of real-time transit information systems: An information demand and supply approach. *Int. J. Transport. Sci. Technol.* 6 (1), 86–98.
- Hickman, M., Wilson, N.H.M., 1995. Passenger travel time and path choice implications of realtime transit information. *Transport. Res. C: Emerg. Technol.* 3 (4), 211–226.
- National Council on Disability, 2015. Transportation Update: Where We've Gone and What We've Learned. National Council on Disability, Washington DC, USA. <http://www.ncd.gov/publications/2015/05042015/>.
- Papangelis, K., Sripada, S., Corsar, D., Velaga, N.R., Edwards, P., Nelson, J.D., 2013. Developing a real time passenger information system for rural areas. *LNCS: Human Interface Manage. Inform.* 8017, 153–162.
- Pelletier, M., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: a literature review. *Transport. Res. C* 19, 557–568.
- Politis, I., Papaioannou, P., Basbas, S., Dimitriadis, N., 2010. Evaluation of a bus passenger information system from the users' point of view in the city of Thessaloniki, Greece. *Res. Transport. Econ.* 29 (1), 249–255.
- Tang, L., Thakuriah, P., 2007. Relationship of attitudes towards road and transit capital investments and propensity to ride transit given traveler information. In: *Proceedings of the Annual Meeting of the Transportation Research Board*, Washington, DC.
- Tang, L., Thakuriah, P., 2011. Will psychological effects of real-time transit information systems lead to ridership gain? *Transport. Res. Rec.: J. Transport. Res. Board* 2216, 67–74.
- Traut, E.J., Steinfeld, A., 2019. Identifying commonly used and potentially unsafe transit transfers with crowdsourcing. *Transport. Res. A* 122, 99–111.
- Velaga, N.R., Beecroft, M., Nelson, J.D., Corsar, D., Edwards, P., 2012a. Transport poverty meets the digital divide: accessibility and connectivity in rural communities. *J. Transport Geogr.* 21, 102–112.
- Velaga, N.R., Rotstein, N.D., Oren, N., Nelson, J.D., Norman, T.J., Wright, S., 2012b. Development of an integrated flexible transport systems platform for rural areas using argumentation theory. *Res. Transport. Bus. Manage.* 3, 62–70.
- Velaga, N.R., Nelson, J.D., Papangelis, K., Corsar, D., Edwards, P., 2014. Designing a passenger-centric information eco-system for integrated and flexible transport systems in rural areas. *J. Sustain. Mobil.* 1 (2), 73–82.
- Watkins, K., Ferris, B., Borning, A., Rutherford, G.S., Layton, D., 2011. Where is my bus? Impact of mobile real-time information on the perceived and actual wait time of transit riders. *Transport. Res. A: Policy Pract.* 45 (8), 839–848.

- Zhang, F., 2010. Traveller responses to real-time transit passenger information systems. PhD thesis. Faculty of the Graduate School of the University of Maryland.
- Zhang, F., Shen, Q., Clifton, K., 2008. Examination of traveler responses to real-time information about bus arrival using panel data. *Transport. Res. Rec.: J. Transport. Res. Board* 2082, 107–115.

Further Reading

- Caicedo, B., Ocampo, M., Sanchez-Silva, M., 2012. An integrated method to optimise the infrastructure costs of bus rapid transit systems. *Struct. Infrastruct. Eng.* 8 (11), 1017–1033.
- Delbosc, A., Vella-Brodrick, D., 2015. The role of transport in supporting the autonomy of young adults. *Transport. Res. F: Traffic Psychol. Behav.* 33, 97–105.
- Kaplan, S., Abreu e Silva, J., Di Ciommo, F., 2014. The relationship between young people's transit use and their perceptions of equity concepts in transit service provision. *Transport Policy* 36, 79–87.
- Légal, J.B., Meyer, T., Csillik, A., Nicolas, P.A., 2016. Goal priming, public transportation habit and travel mode selection: the moderating role of trait mindfulness. *Transport. Res. F: Traffic Psychol. Behav.* 38, 47–54.
- Spears, S., Houston, D., Boarnet, M.G., 2013. Illuminating the unseen in transit use: a framework for examining the effect of attitudes and perceptions on travel behaviour. *Transport. Res. A: Policy Pract.* 58, 40–53.

Tram Lane Configurations and Driving Rules

Farhana Naznin, Port Melbourne, VIC, Victoria, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	338
Tram Lane Configurations	338
Tram Lanes and Road Rules	338
Tramways	338
Tram Lanes	339
Mixed Traffic Tram Lane	339
References	339
Further Reading	340

Introduction

Trams are light rail transit vehicles, which run on fixed rail tracks on roads and mostly share the road space with other traffic and pedestrians. The term “streetcar” used in North America share the similarity of the tram system in the rest of the world. Tram systems have a number of attractive features including their high passenger capacity, good comfort, and very low emission of pollutants compared to other transport systems. Therefore, many cities and countries are extending their tram networks as well as introducing new networks to reduce traffic congestion and to improve the urban environment.

Trams sharing the road with general road traffic is often known as “mixed traffic tram operation.” Mixed traffic tram operation is the least desirable operational type due to the impact of road traffic on tram travel time and reliability (Vuchic, 2007). Melbourne has the largest tram network in the world and it includes the largest mixed traffic tram network with a track length of 167 km (100 miles) (Currie and Shalaby, 2007). The majority of streetcar routes in Toronto also operate in mixed traffic, which is the second largest mixed traffic tram operation. Many cities and countries are implementing tram lane priority measures on road corridors with an aim to transform mixed traffic tram or streetcar system to a light rail comparable service “as far as practical” by providing more protected and prioritised right of way from other road traffic (Yarra Trams, 2010).

Tram Lane Configurations

The types of tram lane priority measures vary for different cities around the world, but their differences lie essentially in the amount of road space for trams and the types of separation provided for trams. At the highest level, trams are given grade-separated or physically segregated right-of-way (ROW); classified as “Category B type ROW” for trams by Vuchic (2007) and generally known as “tramways.” Tramways are mostly available in European, American, and Australian cities, and give trams the speed, safety, and time advantage necessary to compete effectively with the automobile. At a lower level of tram lane priority, tram lanes are designed for exclusive use by trams, either for the entire day (full-time tram lane) or for a specific time of the day (peak-hour tram only lanes/part-time tram lanes). These lanes have demarcation lines beside the tracks and overhead signs. There is no physical separation for these types of tram lanes. This is classified as “Category C type ROW” for trams (Vuchic, 2007) and often known as “full-time” and “part-time” tram lanes. These tram lanes are mostly available on Australia and North American tram/streetcar networks. A recent study by Naznin et al. (2018) investigated tram drivers’ perception of road safety for different tram lane configurations in Melbourne. The study found strong support from tram drivers for the raised tramways followed by the raised yellow strips beside tram tracks.

Tram Lanes and Road Rules

Tramways

A tramway is a part of a road with tram tracks that is between an overhead tramway sign and an end tramway sign as shown in Fig. 1A. Tramways are usually designed by raised tram tracks, raised dividing strips or double yellow lines beside the tram tracks. Fig. 1B shows an example of raised tram tracks and Fig. 1C presents a section of tramways with raised plastic yellow dividing strips beside tram tracks in Melbourne.

A driver (except the driver of a tram, tram recovery vehicle, or public bus) is not allowed to drive on a tramway and even not permitted to drive over raised dividing strips or double yellow lines except it is necessary to avoid an obstruction (VicRoads, 2019).



Figure 1 Tramway configurations. (A) Overhead Tram Only Lane sign for tram ways. (B) Raised tram track. (C) Tram way separated by raised yellow strip.



Figure 2 Tram lane configurations.



Figure 3 Mixed traffic tram operation. Source: <https://images.app.goo.gl/y7abBirm68CJJ2s1A>.

Tram Lanes

Tram lanes are divided from traffic by a single yellow line beside the tram tracks. They can be full-time (24-h) or part-time. Overhead signs indicate the hours of operation (Fig. 2A). Fig. 2B illustrates an example of tram lane section in Melbourne with overhead tram lane operation sign and a single yellow line beside tram track. A driver (except the driver of a tram, tram recovery vehicle or public bus) may only drive in a tram lane for up to 50 m to turn right, or to avoid an obstacle, providing that an approaching tram has not been delayed. The driver must have to move out of the path of the tram as soon as the driver can do so safely.

Mixed Traffic Tram Lane

On mixed traffic tram lanes, trams share the road with general road traffic and trams do not have any priority on tram lanes over general traffic and pedestrians (Fig. 3).

References

- Currie, G., Shalaby, A., 2007. Success and challenges in modernizing streetcar systems: experiences in Melbourne, Australia, and Toronto, Canada. *Transport. Res. Rec.: J. Transport. Res. Board* 2006, 31–39.
- Naznin, F., Graham, C., David, L., 2018. Exploring road design factors influencing tram road safety—Melbourne tram driver focus groups. *Accid. Anal. Prev.* 110, 52–61.
- Vuchic, V.R., 2007. *Urban Transit Systems and Technology*. John Wiley & Sons, Hoboken, New Jersey.
- VicRoads, 2019. Driving with trams. Viewed on 11 October 2019. <https://www.vicroads.vic.gov.au/safety-and-road-rules/road-rules/a-to-z-of-road-rules/trams>.
- Yarra Trams, 2010. Tram priority – research, modelling & analysis of tram priority on the Melbourne network.

Further Reading

- Adams, S., Commissioner, C., 2008. Portland Streetcar Development Oriented Transit. Office of Transportation, Portland, Oregon, and Portland Streetcar, Inc.
- American Public Transportation Association, 2015. Atlanta Streetcar Enters Service. Passenger Transport, January 9, 2015
- Association of German Transport, 2000. Light Rail in Germany. Germany.
- Cornwell, P.R., Luk, J., Negus, B., 1986. Tram priority in SCATS. Traffic Eng. Control 27
- Currie, G., Burke, M., 2013. Light Rail in Australia—Performance and Prospects. Australasian Transport Research Forum, Brisbane, Australia
- Currie, G., Goh, K., Sarvi, M., 2012. Exploring the Impacts of Transit Priority Measures Using Automatic Vehicle Monitoring (AVM) Data. Australasian Transport Research Forum (ATRF), 35th, 2012, Perth, Western Australia, Australia
- Gormick, G., 2004. The Streetcar Renaissance. Greg Gormick and On Track consulting
- Shalaby, A., Abdulhai, B., Lee, J., 2003. Assessment of streetcar transit priority options using microsimulation modelling. Can. J. Civil Eng. 30, 1000–1009.
- Skabardonis, A., 2000. Control strategies for transit priority. Transport. Res. Rec.: J. Transport. Res. Board 1727, 20–26.
- Topp, H.H., 1999. Innovations in tram and light rail systems. Proc. Inst. Mech. Eng. F: J. Rail Rapid Transit 213, 133–141.

Railway Crossing

Chunliang Wu*, **Inhi Kim†**, *Monash Institute of Transport Studies, Department of Civil Engineering, Monash University, Clayton, VIC, Australia;
†Department of Civil and Environmental Engineering, Kongju National University, Cheonan, Republic of Korea

© 2021 Elsevier Ltd. All rights reserved.

Railway Crossing	341
Types of Traffic Control	341
Safety	342
Design	343
Policy	343
Education and Enforcement	344
Technology	344
Knowledge Management	344
Risk Management	344
References	344
Further Reading	345

Railway Crossing

Railway crossings are an integral part of transport networks in many countries. In Australia, there are more than 23,500 railway level crossings (Larue et al., 2018a). The United States has approximately 212,000 railway crossings (Soleimani et al., 2019). However, their existence causes a severe risk to the safety of road users. Statistics show that there are around 100 incidents at railway level crossings every year in Australia (Australian Transport Council, 2010). In Sweden between 2008 and 2012, 59 accidents occurred at road-rail crossings (Jonsson et al., 2019). Although the number of crashes is relatively low, the impacts of these crashes on lives, both fatalities and serious injuries, and financial costs have the potential to be catastrophic. In Australia, these accidents result in an average of 37 deaths annually and more than \$100 million in losses (Australian Transport Council, 2010). Thus, over the past few decades, countries have made great efforts to improve railway crossing safety.

Types of Traffic Control

There are two types of traffic control systems that are used at operational railway crossings in worldwide. They are passive and active railway crossings (see Fig. 1).

Passive railway crossings have no active warning devices. Only signs and line markings are placed to attract road users' attention and warn them that there is a railway crossing ahead. They need to decide when it is safe to cross by estimating the arrival time and speed of a train at the railway crossing. The safety of road users entirely relies on their awareness of warning signs and their judgment. Passive railway crossings are normally used in rural areas, where the speed of trains is relatively faster than that in urban areas.

Active railway crossings provide a warning through active devices. These devices include flashing lights, alarms, and boom gates. They can react to the presence of an approaching train, which is different from the passive control. This kind of control system is mainly used on high-traffic volume roads.

The proportion of these two traffic control systems is different in various countries. In Australia, 79% of railway level crossings are passive (Australian Rail Track Corporation, 2016). The majority of them use either a "Stop" or "Give Way" sign, so road users are expected to check the presence of an approaching train. The remaining 21% of railway level crossings are active. They usually used flashing lights or boom gates to alert road users when a train is coming. In Finland, there were 22% active railway level crossings (Laapotti, 2016). Most of them were equipped with both flashing lights and bells. Also, all active devices were controlled automatically without the intervention of railway staffs in Finland, which is different from Britain (Evans and Hughes, 2019). They have manual control railway crossings. However, in some European countries, the proportion of passive railway crossings is less than the active one. For example, 15% of railway crossings in Belgium were passive in 2011, 30% in France, and 36% in Germany (Laapotti, 2016).

Currently, with the development of intelligent transport systems (ITS), some innovative technologies are emerging to complement existing traffic control systems at railway crossings. They include vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication technologies. By adopting these technologies, the information from trains can transmit to road vehicles and warning infrastructure at railway crossings. The vehicles equipped with appropriate devices such as Global Positioning System (GPS) navigation systems and dedicated short-range radio communication systems will receive advanced warnings in time to avoid collisions. Compared with traditional control systems, they could significantly improve the safety of road users at railway crossings.



Figure 1 (A) A passive railway crossing; and (B) An active railway crossing. Source: NSW Government (2017)

Safety

Based on the Australian Transport Safety Bureau (ATSB) report, there were 601 collisions between road vehicles and trains at railway crossings between 2002 and 2012 (Australian Transport Safety Bureau, 2012). Safety at railway crossings has become a significant concern to the Australia Transport Safety Bureau. They investigated some railway level crossing incidents, and found these incidents have some similar characteristics (1) fatal level crossing incidents most often occurred in daylight on a dry road; (2) nearly half of the fatal incidents are attributed to unintended road user error; (3) fatal incidents frequently occurred at active railway crossings; (4) most of the fatalities at railway level crossing are to pedestrians. In addition, they also pointed out that although collisions normally occur between light vehicles and trains, the collisions between heavy vehicles and trains have a growing trend recently. These collisions will cause severe consequences.

Over the past decades, a number of studies analyzed the factors influencing the safety of railway crossings. Some studies found that passive railway crossings pose a significant safety risk. The reason is that at a passive railway crossing, road users are required to decide when it is safe to cross. However, there is usually no information provided for road users to make a decision. The errors of human judgment constitute a significant factor associated with crashes at passive railway crossings. For example, although some drivers recognize that there is a railway crossing, it is hard for them to notice warning signs. It leads to that they are not sure whether they need to stop at a railway crossing or not, which might increase the possibility for rear-end collisions. Also, drivers have difficulty in estimating the speed of an approaching train. Larue et al. (2018a) conducted a field-based experiment to test drivers' ability to estimate train speeds. The results showed that most participants underestimated the train speed by 30%. It indicated that they might be less likely to detect their potential risks of traversing a railway crossing. They might cross it in a limited safe space. It will increase the potential for near misses.

Recently, some passive crossings have upgraded to active crossings in many countries to improve safety. Although active railway crossings reduce the errors in judgment, there is also a risk of collisions because other forms of behavior errors happen at this kind of crossings. For example, drivers and pedestrians might not comply with active railway crossing rules. Some active control systems are designed based on the principle of ensuring the safety of road users to the greatest extent possible. It means that when a train is coming, the railway crossing will be closed for a long time. Some road users are reluctant to wait a long time. Thus, they attempted to traverse a railway crossing when the boom gates were down because their experience told them that the closing time of railway crossings would be extended to ensure safety. Statistics show that these impatient behaviors often occur at a congested level crossing (Larue et al., 2018b). Also, drivers with sensation-seeking tendencies are more likely to take such action. Larue et al. (2020) further studied the relationship between these risk behaviors and waiting times. They found that waiting time less than three minutes can effectively reduce the likelihood of risky behavior. In addition, road users, especially pedestrians, deliberately choose to ignore the flashlights at railway crossings to save time, though these controllers are visible. Most incidents happened to those people who were very familiar with railway crossings and violated the active warning systems. Moreover, Larue et al. (2018b) pointed out that if a pedestrian disregards the flashing lights, other pedestrians are more likely to follow. It will significantly increase risks. Note that unlike other road crashes, crashes at railway crossings are more likely to be attributed to unintentional errors caused by road user behaviors, and less likely to be affected by fatigue, alcohol, or drugs (Australian Transport Safety Bureau, 2008).

In addition to road user-related factors, there are other environment-related factors associated with crashes at railway crossings. Some studies found that sighting distance is highly related to crashes between trains and road vehicles at crossings. The reason is that short sighting distances do not provide sufficient time and space for road users to make a decision. Also, the speed limit of the road at passive crossings most often was higher than at active crossings. It results in that the fatal incidents often took place at the passive crossing where the speed limit is high. A lower speed limit provides a sufficient time for road users to pay attention to whether a danger ahead exists. Besides, the road gradient also affects safety at crossings. Crashes are more likely to occur on uphill or downhill

roads. It is difficult for drivers to make a complete stop when vehicles go uphill or downhill to railway crossings compared with flat roads.

In order to reduce the risks at railway crossings, much effort has been devoted to developing accident prediction models. Earlier studies preferred to apply generalized linear models to predict the likelihood of crashes at railway crossings. These models are suitable for discrete crash data, and they are a powerful tool to examine the relationship between road accidents and explanatory variables (Zheng et al., 2016). However, a potential limitation of these models is that the correlation between variables needs to be predetermined. If there is no relationship between variables, the estimation model might lead to erroneous results. In recent years, data mining, as a new technique, has demonstrated great potential for exploring large data sets. Because this technique does not lie on assumptions, it has been adopted by experts to estimate crashes. However, limited studies used them to predict accident rates at railway crossings. Within these studies, decision tree models are the most commonly used data mining techniques because of its ease of use and comprehensiveness (Zheng et al., 2019). More recently, with the rapid development of deep learning, it might provide a new tool for examining the factors influencing safety at railway crossings and improving the accuracy of prediction models.

Moreover, analyzing real field data is the best way to understand road user behaviors at railway crossings. However, the number of collisions between trains and vehicles is low. Thus, there is not enough field data to study. Currently, the simulation-based method is a good alternative for generating a large amount of reliable data to assess the safety of different control systems at railway crossings. Kim et al. (2015) used this method to assess the influence of ITS devices on driving behavior. The results show that driver behavior is more likely to be influenced by passive crossings with ITS intervention than by active crossings.

Design

Railway crossings pose a significant safety risk to road users. Based on the above review of factors contributing to crashes at crossings, the following elements need to be vigorously considered before undertaking the design of a railway crossing.

The level crossing surface should be in good condition with no potential to cause a hazard to road users. Also, the selection of crossing surface type shall fully consider the road vehicle type, volume, and track configuration to make road vehicles smoothly traverse tracks (Australian Rail Track Corporation, 2016). The crossings should be situated on the flat ground as much as possible (Laapotti, 2016). It allows drivers to stop more easily before crossing and have sufficient time to make a decision.

Concerning the design of passive railway crossings, sight distance is the most critical element that needs to be considered. Its design can be based on the road traffic mix. In general, the sight distance requirements for heavy vehicles are more stringent than those for light vehicles. Also, trees that might restrict visibility should be removed or trimmed. Buildings are forbidden to be established in the line of sight area at a railway crossing because they would influence existing visibility. Moreover, operators need to enhance the conspicuity of warning signs (Yeh and Multer, 2007). The signs can be illuminated or reflectorized to remind road users that they are approaching a railway crossing. Lowering the height of warning signs also can enhance its conspicuity. In addition, the speed at a passive crossing should be limited to a specific range. A higher speed limit might result in that road users are not aware of the danger and deliberately disregard warning signs (Laapotti, 2016). It would escalate risks at railway crossings. Furthermore, operators need to develop in-vehicle warning systems further. Combining them with passive countermeasures can provide more information for road users.

Waiting time is a critical factor that needs to be considered when designing an active crossing. Reducing waiting time would improve compliance (Larue et al., 2020). Also, the reliability of active warning systems shall be improved. It means that false alarms and failures of warning signals should be avoided. Besides, lowering and extending gate arms might reduce road user risk behaviors. Moreover, flashing lights can be upgraded to traffic signals. In comparison to flashing light, road users have familiarity with this type of signal; thus, they will be more likely to comply with traffic signals (Lenne et al., 2011).

Regarding the design of pedestrian crossings, in addition to the above aspects, particular users need to be considered. The crossing should make sure that the wheelchair users and children can access the surrounding infrastructures, such as public transport stations (De Gruyter and Currie, 2016). Also, local weather patterns such as fog shall be considered. It might influence pedestrians' line of sight. Moreover, if pedestrian crossings have high pedestrian usage rates, the lowest speed limit of rail traffic should be used, or grade separation shall be seriously considered. Besides, cyclists are sensitive to the traversing angle at skewed railway crossings (Ling et al., 2017). A higher angle can dramatically reduce crashes.

In July 2019, Federal Highway Administration published the 3rd edition of the Highway-Rail Crossing Handbook, which provides best practices as well as adopted standards relative to highway-rail grade crossings. The handbook contains the designs of geometry, illumination, crossing surfaces, sight distance, markings, treatments, signals, warnings, gates, access, and pedestrian treatments (Brent and Chelsey, 2019).

Policy

To reduce the likelihood of crashes at railway crossings, some strategies were proposed and implemented in various countries. These strategies can be summarized as four key areas, including education and enforcement, technology, knowledge management, and risk management.

Education and Enforcement

Educational measures are used to change road user behavior. Safety education campaigns conducted in some countries, which can increase the safety awareness of road users and improve community understanding of the risks. The Australian state government, New South Wales (NSW), conducted a campaign called “Don’t rush to the other side” from 2016 to 2017. The campaign was broadcasted through television, outdoor billboards and petrol pumps, radio, digital, and social media. The campaign was very successful for particularly the drivers of the light vehicles claiming that their behaviors and attitudes toward the railway crossings were changed (NSW Government, 2017).

Also, local law enforcement can remind road users of the penalties for violating the road rules and deter risk behaviors. Police are responsible for the detection of violations at level crossings in most countries. This should be shifted to technology since the police cannot provide sufficient enforcement coverage where technology such as cameras, sensors, and emerging detectors provide a wide-scale permanent and consistent detection coverage. The technology might detect road users’ plate numbers, face, behavior, and whether they violate the rules (Economic Commission for Europe, 2017).

Technology

New engineering and technological solutions are adopted to improve safety and efficiency at railway crossings. Also, research efforts should continue to develop new types of crossing controls and warning infrastructure.

Very recently, a vehicular communication system has been emerging in railway crossings too. The specific types of communication include V2I (vehicle-to-infrastructure), V2V (vehicle-to-vehicle), and V2X (vehicle-to-everything). Connected vehicle-based warning systems warn road users to stop for crossing to mitigate the risk of grade-crossing collisions. The warning is delivered by visual and/or audio alerts and the display of the estimated waiting time (Wang et al., 2019).

Knowledge Management

Data is vital to support decision-making and accident data of crossings is relatively sparse. It is desirable to create a national database, including road traffic data at railway crossings, to reduce data gaps. Each jurisdiction can access these data and analyze them to understand road user behavior better and examine characteristics surrounding incidents.

The decision making process can be driven by dynamic and static datasets from different sources, including service failure data, signal data, ballast history, grinding history, remedial action history, traffic data, inspection data, as well as curve and grade data (Ghofrani et al., 2018). The data sources include acoustic bearing detectors, hotbox detector (HBD) temperature trending, hot/cold wheel detectors, truck performance detectors, hunting detectors, wheel impact load detectors (WILD), cracked axle detectors, cracked wheel detector, and machine vision (Wang et al., 2018, Stratman et al., 2007, Li et al., 2014).

NSW collected traffic controls, level crossing characteristics, and other related risks on all public level crossings. Two streams of data collection projects store ALCAM (Australian Level Crossing Assessment Model) field assessments at 212 road and pedestrian level crossings on the ARTC (Australian Rail Track Corporation) network and road traffic data collection (including type, volume, and speed of vehicles traversing a level crossing) at 315 level crossings in NSW (NSW Government, 2017).

Risk Management

Utilizing incident data and innovative modeling techniques is to develop an effective risk assessment model. Based on these models, estimating marginal accident costs is essential for operators to make an investment decision. Also, operators can use simulation methods such as driving simulator to compare different crossing controls and assess the risks before and after taking counter-measures. In addition, recent years witness countries implemented level crossing removal and grade separation programs to eliminate collision risks. It is necessary to develop a model to determine the suitability of these programs.

The Dutch Ministry released the 2016 framework policy after several revisions in 1999, 2004, and 2010. Since the safety performance in the Netherlands was satisfied in 2016, the Dutch level crossing policy changed a direction from safety performance indicators to risk factors (Dutch Safety Board, 2018).

The risk management framework in Victoria, Australia, was improved in 2010 by, in partnership with the road and rail managers who are able to provide knowledge and perspectives in level crossing safety and develop a comprehensive plan to manage the strategy (Victorian Auditor-General, 2010). Incorporating rail managers were critical since they are liable to assess level crossing safety and accredited as capable of managing the risks.

References

- Australian Rail Track Corporation, 2016. Level crossing safety. Australia: Australian Rail Track Corporation.
- Australian Transport Council, 2010. National railway level crossing safety strategy 2010/2020. Australia.
- Australian Transport Safety Bureau 2008. Railway level crossing safety bulletin. Australia: Australian Transport Safety Bureau.
- Australian Transport Safety Bureau, 2012. Australian Rail Safety Occurrence Data 1 July 2002 to 30 June 2012. Canberra: Australian Transport Safety Bureau.
- Brent, D.O., Chelsey, C. 2019. Railroad-highway grade crossing handbook, third ed. United States: Federal Highway Administration.
- De Gruyter, C., Currie, G., 2016. Rail-road crossing impacts: an international synthesis. *Transp. Rev.* 36, 793–815.
- Dutch Safety Board, 2018. Level crossing safety. Dutch Safety Board, Netherlands.
- Evans, A.W., Hughes, P., 2019. Traverses, delays and fatalities at railway level crossings in Great Britain. *Accid. Anal. Prev.* 129, 66–75.
- Ghofrani, F., He, Q., Goverde, R.M., Liu, X., 2018. Recent applications of big data analytics in railway transportation systems: A survey. *Transp. Res. Part C: Emerg. Technol.* 90, 226–246.

- Economic Commission for Europe, 2017. Assessment of safety at level crossings in UNECE member countries and other selected countries and strategic framework for improving safety at level crossings. In: Working Party on Rail Transport. Economic and Social Council, United Nations.
- Jonsson, L., Bj Rklund, G., Isacson, G., 2019. Marginal costs for railway level crossing accidents in Sweden. *Transp. Policy* 83, 68–79.
- Kim, I., Larue, G.S., Ferreira, L., Rakotonirainy, A., Shaaban, K., 2015. Traffic safety at road–rail level crossings using a driving simulator and traffic simulation. *Transp. Res. Rec.* 2476, 109–118.
- Laapotti, S., 2016. Comparison of fatal motor vehicle accidents at passive and active railway level crossings in Finland. *IATSS Research* 40, 1–6.
- Larue, G.S., Blackman, R.A., Freeman, J., 2020. Frustration at congested railway level crossings: How long before extended closures result in risky behaviours? *Appl. Ergon.* 82, 102943.
- Larue, G.S., Filtness, A.J., Wood, J.M., Demmel, S., Watling, C.N., Naweed, A., Rakotonirainy, A., 2018a. Is it safe to cross? Identification of trains and their approach speed at level crossings. *Safe. Sci.* 103, 33–42.
- Larue, G.S., Naweed, A., Rodwell, D., 2018b. The road user, the pedestrian, and me: Investigating the interactions, errors and escalating risks of users of fully protected level crossings. *Safe. Sci.* 110, 80–88.
- Lenne, M.G., Rudin-Brown, C.M., Navarro, J., Edquist, J., Trotter, M., Tomasevic, N., 2011. Driver behaviour at rail level crossings: responses to flashing lights, traffic signals and stop signs in simulated rural driving. *Appl. Ergon.* 42, 548–554.
- Li, H., Parikh, D., He, Q., Qian, B., Li, Z., Fang, D., Hampapur, A., 2014. Improving rail network velocity: A machine learning approach to predictive maintenance. *Transp. Res. Part C: Emerg. Technol.* 45, 17–26.
- Ling, Z., Cherry, C.R., Dhakal, N., 2017. Factors influencing single-bicycle crashes at skewed railroad grade crossings. *J. Transp. Health* 7, 54–63.
- Nsw, 2017. Level crossing strategy council yearly report 2016–17. In: N.S.W., T.F., (Ed.). Nsw Government 2017. Level crossing strategy council yearly report 2016–17. New South Wales: Transport for NSW, Centre for Road Safety.
- Victorian Auditor-General 2010. Management of Safety Risks at Level Crossings. In: Victorian Government, (ed.). Victorian Auditor-General's Office, Victoria.
- Soleimani, S., Mousa, S.R., Codjoe, J., Leitner, M., 2019. A comprehensive railroad-highway grade crossing consolidation model: A machine learning approach. *Accid. Anal. Prev.* 128, 65–77.
- Stratman, B., Liu, Y., Mahadevan, S., 2007. Structural health monitoring of railroad wheels using wheel impact load detectors. *J. Fail. Anal. Prev.* 7, 218–225.
- Wang, W., He, Q., Cui, Y., Li, Z., 2018. Joint prediction of remaining useful life and failure type of train wheelsets: Multitask learning approach. *J. Transp. Eng. Part A: Sys.* 144, 04018016.
- Wang, X., Li, J., Zhang, C., Qiu, T.Z., 2019. Active warning system for highway–rail grade crossings using connected vehicle technologies. *J. Adv. Transp.* 2019.
- Yeh, M., Multer, J., 2007. Traffic control devices and barrier systems at grade crossings: Literature review. *Transp. Res. Rec.* 2030, 69–75.
- Zheng, Z., Lu, P., Pan, D., 2019. Predicting highway–rail grade crossing collision risk by neural network systems. *J. Transp. Eng. Part A: Sys.* 145, 04019033.
- Zheng, Z., Lu, P., Tolliver, D., 2016. Decision tree approach to accident prediction for highway–rail grade crossings: Empirical analysis. *Transp. Res. Rec.* 2545, 115–122.

Further Reading

- Australian Transport Council, 2010. National railway level crossing safety strategy 2010/2020. Australia.
- Evans, A.W., Hughes, P., 2019. Traverses, delays and fatalities at railway level crossings in Great Britain. *Accid. Anal. Prev.* 129, 66–75.
- Kim, I., Larue, G.S., Ferreira, L., Rakotonirainy, A., Shaaban, K., 2015. Traffic safety at road–rail level crossings using a driving simulator and traffic simulation. *Transp. Res. Rec.* 2476, 109–118.
- Laapotti, S., 2016. Comparison of fatal motor vehicle accidents at passive and active railway level crossings in Finland. *IATSS Res.* 40, 1–6.
- Larue, G.S., Filtness, A.J., Wood, J.M., Demmel, S., Watling, C.N., Naweed, A., Rakotonirainy, A., 2018b. Is it safe to cross? Identification of trains and their approach speed at level crossings. *Safe. Sci.* 103, 33–42.
- Larue, G.S., Naweed, A., Rodwell, D., 2018b. The road user, the pedestrian, and me: Investigating the interactions, errors and escalating risks of users of fully protected level crossings. *Safe. Sci.* 110, 80–88.

Pedestrian Crossing (Crosswalk)

Associate Professor Miho Iryo-Asano*, **Associate Professor Wael K.M. Alhajyaseen[†]**, **Associate Professor Koji Suzuki[‡]**, ^{*}Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, Japan; [†]Qatar Transportation and Traffic Safety Center, Qatar University, Doha, Qatar; [‡]Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya Japan

© 2021 Elsevier Ltd. All rights reserved.

General Introduction	346
Classification of Pedestrian Crossings	346
Unsignalized Crossing	347
Zebra Crossing	347
Other Unsignalized Crossing Facilities	348
Signalized Crossing	348
Crossing and Clearance Times for Pedestrians	348
Pelican Crossing (PEdestrian LIght COntrolled crossings)	349
Puffin Crossing (PEdestrian User Friendly INtelligent crossings)	349
Hawk Crossing (High intensity Activated crossWalk)	351
Crossings at Signalized Intersections	351
Pedestrian–Vehicle Separation Controls	351
Scramble Crossings	352
Additional Considerations	352
Consideration for Disabled People	352
Countdown Signals	353
References	353
Further Reading	354

General Introduction

A pedestrian crossing (crosswalk) is an area designed specifically for pedestrians to cross a road that is being used by vehicular traffic. The core function of these facilities is to provide safe crossing opportunities for pedestrians with fewer delays. Pedestrian crossings are set in locations where they are clearly visible to drivers, and where pedestrians can cross the road most safely across vehicle traffic flow. Pedestrian crossings are a critical part of walkway networks, whose performance significantly affects the accessibility and walkability of pedestrian networks.

Marked pedestrian crossings can be seen at intersections and on busy roads, where crossing without them would be unsafe because of the large volume of vehicles, road width, and high speeds. Furthermore, pedestrian crossings are usually installed in areas that have considerable demand of crossing pedestrians, such as shopping areas, business districts, and near schools or hospitals where vulnerable road users regularly cross.

If vehicle demand is high, pedestrian crossings (especially with traffic signal control) may increase vehicle delay. Some advanced signal control methods enable delay minimization using pedestrian detection systems. Some structural devices have functions to minimize pedestrian exposure time to possible conflict with vehicles. This article introduces the components of pedestrian crossings and different control methods for improving efficiency and safety.

Classification of Pedestrian Crossings

Pedestrian crossings are divided into two types: unsignalized and signalized. The former is installed at locations with low demand of vehicles and pedestrians, whereas the latter are installed at locations with a relatively higher demand. Unsignalized crossings can be further classified according to the existence of pavement markings and types of facilities. Unsignalized crosswalks with zebra-shaped pavement markings are called Zebra Crossings, in which pedestrians have full priority to cross the road. Crossings without markings do not prioritize pedestrians, but instead include refuge islands, humps, and other physical facilities to ensure the visibility of crossing locations for safety reasons. In order to provide pedestrians with safe crossing opportunities, in cases of both priority settings, it is necessary to consider the followings to provide safe crossing opportunities for pedestrians; (1) both pedestrians and drivers can easily recognize each other and avoid conflicts at crossings, and (2) the crossing time of pedestrians (pedestrian exposure time to vehicle flow) is minimized.

Signalized crossings are classified by their control methods and pedestrian detection functions. In Pelican crossings mainly used in the United Kingdom, fixed green and flashing amber/green periods are given to pedestrians according to the demand requested by



Figure 1 Zebra crossing with hump and refuge island.

the push-button. Hawk crossings used in the United States or Canada have similar functions as Pelican crossings. In addition to push-buttons, Puffin crossings detect pedestrians at crossings and adjust the time such that pedestrians can safely complete crossing before the subsequent green for vehicles starts. Toucan crossings are designated for both pedestrians and cyclists, and Pegasus crossings are for horse-riders.

Recently, several guidelines or tools are published to assist practitioners to select proper types of pedestrian facilities taking into account the traffic and geometry conditions at crossings. [Federal Highway Administration \(2018\)](#) provides a guidance to choose geometric layout of unsignalized crossings from safety viewpoint. [Austroads \(2015\)](#) developed a tool to support the selection of crossing types by the combination of the feasibility check and the cost-benefit analysis, considering delay, safety, and walkability index perceived by crossing pedestrians as benefits obtained by construction of crossing.

Unsignalized Crossing

Zebra Crossing

Zebra crossings are crossings with zebra-shaped pavement markings ([Fig. 1](#)), where pedestrians always have the right of way to cross over vehicles. In the UK and some other countries, poles with a flashing globe lamp must be deployed to ensure visibility of the crossings as in [Fig. 1](#). Zebra crossings are widely used not only at midblock crossings but also at unsignalized intersections. [Fig. 2](#) is an example of an unsignalized intersection in Switzerland, where yellow-colored markings are applied.

As described in the note by Department for Transport, The National Assembly of Wales, The Scottish Executive and The Department for Regional Development ([Department for Transport, 1995a,b](#)), pedestrian delays are minimal because they have priority to cross at Zebra crossings. However, if vehicle demand is high, pedestrians may feel threatened to proceed as there would only be small gaps between vehicle arrivals, and vehicles may not yield to pedestrians. This may increase pedestrian delay. In



Figure 2 Zebra crossings with refuge islands at an unsignalized intersection.



Figure 3 Refuge island without zebra-shaped pavement markings.

contrast, if pedestrian demand is too high, vehicles have limited opportunity to pass the crossing and their delays increase significantly. Therefore, it is not reasonable to install Zebra crossings at locations with high pedestrian demand conditions.

If vehicle approach speeds are high, Zebra crossings are not suitable because pedestrians can be exposed to a higher risk of more serious injury. In such cases, installation of signalized crossings or physical facilities to reduce the speed of approaching vehicles can be considered.

Other Unsignalized Crossing Facilities

If the road width is quite wide and pedestrians have to cross the road at once without stopping, the pedestrian crossing time is increased, and the possibility of crossing pedestrians meeting oncoming traffic is also increased. In order to avoid possible risks, several physical facilities can be installed at pedestrian crossings. The facilities also serve to improve the visibility of the crossing to drivers such that they will be aware of crossing pedestrians on the road.

Raised medians or refuge islands are crossing facilities installed between vehicle lanes. These facilities enable pedestrians to cross the road with several stops. For two-way streets, pedestrians usually need to check for vehicles coming from both directions. If raised medians are installed, pedestrians need to check for vehicles from only one direction at once. In this manner, pedestrians can concentrate on vehicle arrivals from specific lanes, thus easing their decision for an appropriate time to cross. **Fig. 3** is an example of a refuge island, where there are no zebra markings. Pedestrians using the island are not guaranteed to have priority but they are required to search for a gap between vehicle arrivals. This type of crossing is installed at locations with less vehicle and pedestrian demands. For further discussions, refer to the chapter of "Stage Crossing."

Narrowing vehicle lanes also works effectively to reduce the crossing distance. For example, in a road section with high parking demand, outer most lanes can be utilized as parking spaces and these lanes can be removed at the location of pedestrian crossing. **Fig. 4** shows an example with the vehicle lane narrowed at a Zebra crossing. This strategy also improves the visibility of the crossing such that drivers can stop the vehicle at the crossing.

Humps are another type of physical facility, which reduce vehicle speed by their vertical alignment. Sometimes, humps are installed at zebra-marked areas of Zebra crossings. As sidewalks are usually slightly elevated than vehicle lanes, crossings with humps can be designed such that there is no difference in the level of the crossings and sidewalks. This barrier-free treatment is advantageous for disabled pedestrians.

Signalized Crossing

Crossing and Clearance Times for Pedestrians

Generally, pedestrian crossing time consists of a discharge time and a clearance time as summarized by [Iryo-Asano and Alhajyaseen \(2014\)](#). The discharge time is the time necessary for pedestrians waited for the green to leave the curb for crossing. Pedestrians need to have some time to start crossing by responding to the green traffic light indication. If there is high pedestrian demand, additional times need to be provided so that all pedestrians waiting in the queue are able to leave the curb.

The clearance time is the time that pedestrians can have sufficient time or crossing before the subsequent vehicle green phase. Usually the clearance time is provided so that pedestrian can cross the entire crosswalk length. In order to calculate the clearance time, we have to assume pedestrian walking speed. However, pedestrian speeds vary significantly. To ensure the safety, slow walking pedestrians such as elderly and disabled should be considered in setting the walking speed for the estimation of the needed clearance interval. This also means that in many cases the given clearance times are too long for most of the pedestrians and there will be the time that is unused by anybody. This will decrease the efficiency of the traffic signal control. The control systems introduced below (Pelican, Puffin, and Hawk crossings) have advanced functions to remove such unused times.



Figure 4 Zebra crossing (narrowing the road).



Figure 5 Pelican crossing. (A) Push-button of Pelican crossing, (B) Location of traffic lights and push-buttons.

Pelican Crossing (PEdestrian Light CONTROLLED crossings)

Pelican (previously Pelicon) crossings are equipped with push-buttons at each edge of the crossing as well as pedestrian signal indicators at the far-side of the crossing (Fig. 5). Pedestrians wishing to cross the road push the button, after which the green lights for vehicles turn into amber followed by red, and then the steady green (green man) is indicated for pedestrians to cross. Pedestrians are allowed to cross only when this steady green is indicated. After a few seconds, pedestrian traffic lights start flashing. At the same time (often a few seconds after the onset of pedestrian flashing green), flashing amber is indicated to vehicles. During this period, pedestrians on the crossing can keep crossing with priority while pedestrians outside of the crossing should not start to cross. If there are no pedestrians on the crossing during flashing green/amber, vehicles are permitted to pass through the crossing. Pelican crossings do not include any sensors for actively detecting crossing pedestrians, and the durations of the steady and flashing green are fixed. Pelican crossings are installed in the United Kingdom and several other countries.

Pelican crossings can reduce unnecessary vehicle delay because vehicles are allowed to pass through the crossing during the flashing amber period after all pedestrians complete crossing. Meanwhile, it may not be able to provide sufficient crossing time to slowly moving pedestrians because there are no sensors to detecting the existence of pedestrians on the crossing.

Under light vehicle traffic demand, a scenario may also arise where pedestrians who pressed the button may cross the road before traffic lights turn to green. In this case, Pelican crossings will provide unnecessary pedestrian green times that are not used by any pedestrians.

Puffin Crossing (PEdestrian User Friendly INtelligent crossings)

Puffin crossings have pedestrian traffic lights located on the same side of the road for pedestrians, diagonal to the road edge, while most of other types of pedestrian traffic lights are placed at the far (opposite) side of the crossing. This means that pedestrians at



Figure 6 Facilities of Puffin crossing.



Figure 7 Push-buttons and traffic lights of Puffin and Toucan crossings. (A) Puffin crossing, (B) Toucan crossing.

Puffin crossings cannot see the traffic lights when they are crossing (as in Fig. 6). This will give the chance for pedestrians to monitor the traffic while they are waiting for the pedestrian green signal indication to cross the street.

Most Puffin crossings comprise sensors installed on the top of traffic lights, and the sensors are sometimes embedded in the ground, especially in the waiting area. These sensors can detect the presence of pedestrians waiting to cross. This will allow canceling of the crossing demand and skipping to provide the pedestrian green when there are no such pedestrians. Some sensors use infrared, microwave, and other types of detection techniques to extend the crossing time if there are pedestrians still crossing the road. Vehicle drivers waiting at puffin crossings are not permitted to pass until all pedestrians have finished crossing the road.

Puffin crossings do not provide flashing green/amber indications. Pedestrian steady green immediately turns into steady red. Pedestrians can start crossing during the steady green period. After the signal turns into red, only pedestrians on the crossing can continue to cross. Red lights for vehicles turn to green after sufficient clearance time that allows pedestrians on the road to complete crossing. This clearance time is adjusted according to the detection of pedestrians by sensors. This allows the system to accommodate slow pedestrians such as elderly and people with disabilities.

Puffin crossings can overcome some issues of Pelican crossings by utilizing pedestrian detectors. The crossing time is automatically extended according to the existence of pedestrians. Additionally, Puffin crossings cancel the call if there are no more pedestrians on the sidewalks. However, due to the equipment of sensors, maintenance, and construction costs are higher than those of other control systems.

Puffin crossings are used in the United Kingdom and other countries. As a similar type of crossing, Toucan crossings are also used in the United Kingdom. Toucan crossings have similar traffic signal heads as Puffin crossings, but bicycle signs are indicated next to pedestrian signs (Fig. 7).

Hawk Crossing (High intensity Activated crossWalk)

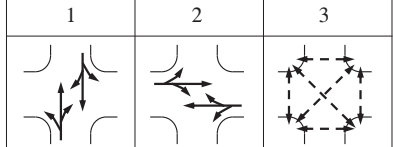
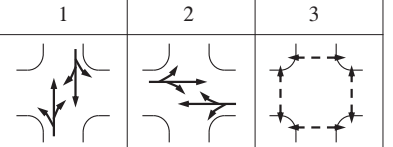
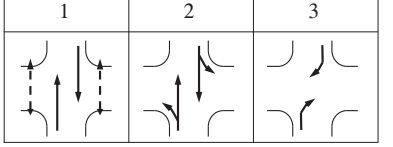
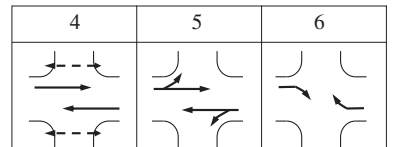
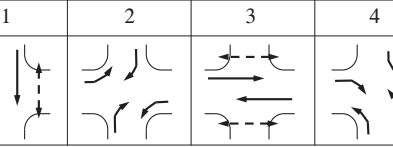
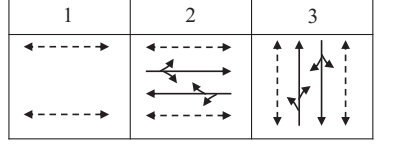
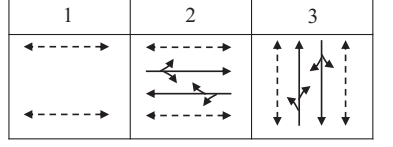
Hawk crossings, also known as Pedestrian Hybrid Beacon (PHB), are another type of crossing used in the United States and Canada. The system is similar to the Pelican crossing. The vehicle traffic signal is dark until the crossing is activated by pedestrians (usually by push-buttons). Once activated, the vehicle signal starts flashing yellow lights to warn drivers of pedestrian presence. Then, the flashing yellow changes to steady yellow and then steady red, instructing drivers to stop. In the meantime, "WALK" signs are given to pedestrians for crossing. After "WALK" signs, the pedestrian signal turns to flashing "DON'T WALK" signs and the vehicle signal turns to flashing red. The flashing red signal has the same meaning as the flashing green/amber in Pelican crossing, during which vehicles can proceed if there are no more pedestrians crossing. Finally, the vehicle signal turns back to dark and the pedestrian signal turns to steady red "DON'T WALK" until any next activation by pedestrians.

Crossings at Signalized Intersections

Pedestrian–Vehicle Separation Controls

At signalized intersections, the pedestrian green signal is often provided simultaneously with parallel vehicle green, permitting vehicles to turn across the crossings. Although vehicles have to yield to pedestrians, there is a potential for hazardous conflict between turning vehicles and pedestrians when drivers do not recognize pedestrians.

Table 1 Pedestrian–vehicle separation controls.

Control types	Examples of phasing plan (ϕ_i : phase number i)
1. Dedicated pedestrian phase	Dedicated pedestrian phase is provided. During this phase, vehicles cannot enter the intersection and pedestrian can cross any crosswalks. Scramble crossing: diagonal crossings are allowed   
2. Turning vehicle separation	Pedestrian phase and conflicting turning vehicle phase are separately provided. (Note: dedicated lanes for every movement are required.)   
3. Leading pedestrian interval	Pedestrian phase starts a few seconds earlier than conflicting turning vehicle phase. 

As a pedestrian-oriented control type, pedestrian–vehicle separation controls have been gaining attention recently. Examples of pedestrian–vehicle separation controls are shown in [Table 1](#). They are installed when capacity or safety problems are expected because of conflicts between pedestrians and turning vehicles. Dedicated pedestrian phases are additional phases where only pedestrians are allowed enter the intersections. Scramble crossings, which even allow diagonal crossing at intersections, are also classified into this type. Turning vehicle separation is a phasing plan that separately provides the right of way to pedestrians and conflicting turning vehicles. Various patterns of phasing can be considered in this type, but two typical examples are given in [Table 1](#). It should be noted that dedicated vehicle lanes are required for all directions to apply these phasing plans (no shared lanes) because all directions have a right of way at different timings.

Leading Pedestrian Interval (LPI) is a phasing plan in which the pedestrian green starts a few seconds prior to the vehicle green. In ordinary control with simultaneous phasing, there is a high probability of conflict between pedestrians and turning vehicles at the beginning of the green indications because both of them start to discharge at the same time. In LPI, platoons of queued pedestrians can start crossing earlier and the probability of conflict decreases. Even if pedestrians are still crossing at the onset of vehicle green, their visibility from drivers are high enough and they can cross safely.

Stage crossings with refuge islands are also utilized to improve the efficiency and/or safety of intersections. See the cross references for more details.

Scramble Crossings

A scramble crossing (also called a Barnes Dance) is a type of pedestrian–vehicle separation phase that provides an exclusive pedestrian phase for all pedestrians. In this phase, no vehicle can enter the intersection. All pedestrians are allowed to cross towards any direction, even diagonal paths in the intersection. This phasing plan can reduce conflicts between pedestrians and vehicles.

Scramble crossings are often installed in central business districts with high pedestrian demand. In these districts, scramble crossings are also regarded as landmarks or symbols of areas for walking places. The scramble crossing located in front of Shibuya station in Tokyo, Japan ([Fig. 8](#)), enables more than 1000 pedestrians to cross at once. Another use of scramble crossing is to achieve safer crossing for children. [Fig. 9](#) shows an example of scramble crossings located in front of an elementary school.

Although scramble crossings can fully separate the right of ways between pedestrians and vehicles, it should be noted that the pedestrian green time in scramble crossings is considered as lost time for vehicles. This period completely prohibits vehicle flow, which deteriorates the efficiency of vehicle traffic. In general, the cycle length (duration to complete a sequence of signal phasing) with scramble crossing tends to be longer than that of an ordinary phasing plan. Furthermore, with higher vehicle demand, the cycle length is also increased to ensure capacity. Longer cycle lengths correspond to longer vehicle delay, and there is a possibility that drivers in hurry may rush into the intersection just before the traffic signal turns red. Pedestrians may also get irritated waiting for the long cycle and take illegal crossings. Therefore, to install scramble crossings, vehicle demand should be carefully examined such that cycle length is kept within an acceptable range.

Additional Considerations

Consideration for Disabled People

Crossings with push-buttons, such as Pelican or Puffin crossings, support blind/visually impaired pedestrians by providing non-visual signs that notify pedestrians that it is safe to cross. There are various types of equipment, such as beepers, vibrating buttons, and rotating knobs installed under the wait box.



Figure 8 Scramble crossing in a business district (Shibuya, Japan).



Figure 9 Scramble crossing near a school (Nagoya, Japan).



Figure 10 Example of countdown pedestrian signal head. Source: From *International Association of Traffic and Safety Sciences (IATSS) (2007)*.

At signalized intersections in Japan, two different types of sounds or melodies are used at one intersection to announce pedestrian green. The two types of sounds correspond to different phases of the intersection such that visually impaired pedestrians can distinguish the green signal.

Countdown Signals

To improve pedestrian safety and comfort, countdown pedestrian signals are widely utilized. There are several types of countdown signals worldwide, and the two most common types display the remaining time either in seconds or dots. The former type is typically used, for example, in the United States, with the remaining seconds being shown along with flashing “DON’T WALK” signs. An example of the latter type is that used in Japan, as shown in **Fig. 10**. At the beginning of the pedestrian green, all dotted lights are displayed as in the left hand side of **Fig. 9**. The dots disappear one by one as time proceeds, and all dots disappear when the pedestrian flashing green starts. Similarly, countdown red dots are displayed during the red time.

References

- Austroroads, 2015. Development of the Australasian Pedestrian Facility Selection Tool, Report No. AP-R472-15.
 Department for Transport, 1995a. The National Assembly of Wales, The Scottish Executive and The Department for Regional Development. The Assessment of Pedestrian Crossings. Local Transport Note 1/95.

- Department for Transport, 1995b. The National Assembly of Wales, The Scottish Executive and The Department for Regional Development. The Design of Pedestrian Crossings. Local Transport Note 2/95.
- Federal Highway Administration, 2018. Guide for Improving Pedestrian Safety at Uncontrolled Crossing Locations, Report No. FHWA-SA-17-072.
- International Association of Traffic and Safety Sciences (IATSS), 2007. H856 Project Report "Research into practical optimized signal control with an emphasis on actual conditions on pedestrian crossings". (In Japanese).
- Iryo-Asano, M., Alhajyaseen, W.K.M., 2014. Analysis of pedestrian clearance time at signalized crosswalks in Japan. *Proc. Comp. Sci.* 32, 301–308.

Further Reading

- American Association of State Highway and Transportation Officials, 2004. Guide for the Planning, Design, and Operation of Pedestrian Facilities.
- Austrroads, 2013. Guide to Traffic Management, Part 6: Intersections, Interchanges and Crossings.
- CERTU, 2010. Guide de conception des carrefours à feux, technical report. Centre d'Etudes sur les Réseaux, les Transports, l'Urbanisme et les constructions publiques. (In French).
- Federal Highway Administration, 2009. Manual on Uniform Traffic Control Devices for Streets and Highways (MUTCD), Part 4 - Highway Traffic Signals.
- Tang, K., Boltze, M., Nakamura, H., Tian, Z (Eds.), 2019. Global Practices on Road Traffic Signal Control. Elsevier, Amsterdam, Netherlands.

Bike-Sharing System: Uncovering the “Success Factors”

S.K. Jason Chang, Amanda Fernandes Ferreira, Department of Civil Engineering, National Taiwan University, Taiwan

© 2021 Elsevier Ltd. All rights reserved.

Introduction	355
The Public Bike-sharing Systems: Then and Now	355
The “Success Factors” for Public Bike Sharing Systems	358
Provision of Infrastructure	359
Integration With Other Modes of Transport	359
Service Regulation	360
Use of New Technologies	360
Provision of Information and Education	360
Conclusions	360
References	361
Further Reading	362

Introduction

The active mobility, particularly walking and cycling, has been heavily promoted as a sustainable alternative to motorized vehicles. The rise in the participation of active transportation modes in the modal split has been pointed out as an economically viable way to improve the overall mobility in the cities. However, its benefits can go far beyond the reduction of traffic congestion levels, optimization of urban infrastructure use, and, in the case of cycling, the reduction of total travel time and cost. It also may comprises lowering the noise, pollutants, and greenhouse gases (GHG) emissions that are generated by transportation activity. More than this, it may help to improve public health as well as to reduce the associated costs of it, as it increases the individuals’ physical activity levels, which ultimately contributes to enhancing the population quality of life and lower mortality rates. Moreover, the increase in the active mode share can contribute to decrease the number and severity of traffic accidents, reduce crime rates, facilitate social interaction and networking, positively affect housing prices and the local economy in general. All these factors make cycling an excellent choice for cities that aim to be livable and sustainable (Bieliński and Ważna, 2019; DeMaio, 2009; Ellison and Greaves, 2011; Granville et al., 2001; Jacobsen, 2015; Kabak et al., 2018; Liu et al., 2012; Ogilvie et al., 2004; Rojas and Wong, 2017; Schabus, 2020; Toronto, 2012; Zhang et al., 2015).

As a mode of transport, bicycles are a promising alternative to partially replace the use of private automobiles due to their flexibility and low cost, being a competitive option especially for travels under 5 km in urban areas (Buehler, 2012; Kumar et al., 2016; Li et al., 2013; Ma et al., 2018; Millward et al., 2013; Van Wee et al., 2006). Due to its numerous positive characteristics for the users and the city environment and considering all the negative impacts caused by the unrestricted use of private automotive vehicles in urban centers, cycling has emerged as an attractive option to urban transport being used both alone for short distances in daily commuting trips as well as integrated with public transportation creating a door-to-door network (Kumar et al., 2016). Thus, cycling has been globally promoted through the provision and improvement of cycling infrastructure and by the raising of population awareness on its individual and collective benefits. In this context, public bicycle schemes, also known by other names such as bike-sharing schemes, bike-sharing systems, public bicycle share, or public bike-sharing system, appeared in line with the principles of economic, social, and environmental sustainability.

The public bike-sharing systems (PBSS) comprise a scheme of self-service public bicycles and pedelecs that are shared, usually in urban areas, by individuals for a certain period with or without the payment of a rental fee. At present, there are three types of PBSS: the station-based bike system (SBBS), the free-floating bike system (FFBS), which is also referred to as the dockless bike system, and, recently, the hybrid bike-sharing system (HBSS), which combine the first two models.

The SBBS comprises bicycles that can be rented and returned after usage in any of the system’ docking stations. Whereas, the bicycles used in the FFBS can be rented and returned anywhere within the operating area, being necessary simply park and lock them using the mobile app. Finally, the HBSS gathers features of both schemes. It offers the possibility to track the nearest bicycle and unlock it using the app as well as to return it anywhere in the same way that users can do while using the FFBS. However, in this scheme, there are also stations from where the bicycles can be rented and where they can be returned to after usage just like in the SBBS (Albiński et al., 2018; Bieliński and Ważna, 2019; Caggiani et al., 2018; Pal and Zhang, 2017; Shen et al., 2018).

The Public Bike-sharing Systems: Then and Now

The first public bike-sharing system was launched in Amsterdam by a group of activists a few decades ago, in 1965, under the name of *Witte Fietsen* (White Bikes). The system counted with 50 unlocked white bicycles to be used anonymously and free of charge by

anyone who intended to, as part of a social/shared economy. Even though this experience was unsuccessful as the bicycles were stolen or damaged within days, it was later known as the 1st generation of the public bike system (Chardon, 2019; DeMaio, 2009; Kabak et al., 2018; Shen et al., 2018). After the failure of the White Bikes, other cities such as La Rochelle (France), in 1974; Cambridge (United Kingdom) in 1993, and Portland (USA), in 1994, also implemented public bike-sharing schemes with the characteristics of the 1st generation but only the La Rochelle systems had not faced a fast fade (Shaheen et al., 2010; Melo and Maia, 2013).

It was in the cities of Farsø and Grenå (Denmark) that, in 1991, a system emerged with characteristics that would later be recognized as belonging to the 2nd generation. Although another system was opened in the city of Næskov in the same country 2 years later, it was only in 1995 that a large-scale system was launched. This system, which was located in Copenhagen, Denmark, begun to operate in 1995 with 110 stations and 1100 bicycles. Unlike in the 1st generation, the systems of 2nd generation implemented stations from where the bicycles should be taken and returned upon a refundable deposit. As a coin was required as a deposit, the systems of this generation started to be referred to as "coin-based-systems". Due to the need to return the bicycle to a station to get the deposit back, the incidence of theft and vandalism decrease, however, it was still a problem as the user remained anonymous (DeMaio, 2009; Shaheen et al., 2010). Over the years many other cities implemented this type of system. The *Bycykler* in Sandnes (Norway), in 1996; the *City Bikes* in Helsinki (Finland) in 2000; the *Bycykel* in Aarhus (Denmark) in 2005; the *Yellow Bike Project* in Minneapolis and Saint Paul in 1996 and Austin in 1997; the *Red Bike* in Madison in 1995 and the *Freewheels* in Princeton in 1998, all in the USA, where some of the systems that operated using the 2nd generation model. Even overcoming some of the problems of the previous generation through the reduction of the stealing and vandalism the systems that used this operational scheme did not last as there were still problems due to the lack of control over the return and usage time (Melo and Maia, 2013).

The 3rd generation brought some innovations that helped to considerably lessen the occurrence of robbery and vandalism. The use of information communication technology (ICT) was the main improvement incorporated into the systems of this generation, which are usually referred to as "SmartBike" or "IT-based system". Instead of a coin, the bicycles of this generation were unlocked at the stations through the use of a personal card, which made the registration in advance mandatory, thus, ending anonymity. The first system with these characteristics received the name *Bikeabout* and appeared inside the campus of the Portsmouth University (England) in 1996. However, it was only in 1998, in Rennes (France), that systems with these features became available for general public under the name *Vélo à la carte* (later known as *Le vélo Star*). In contrast with the previous generations, this system was free of charge only for the first 30 min of use, being the exceeded time charged on the credit card indicated for the user during the registration (DeMaio, 2009; Melo and Maia, 2013; Shaheen et al., 2010). In Asia, the first system was opened in 1999. The Singaporean system received the name of *TownBike* (later renamed as *SmartBike*). No longer after, Japan launched the *Taito Bicycle Sharing Experiment* in 2002.

The 3rd generation was a success and made the PBSS internationally known. The bike-sharing scheme attracted even more attention when in 2005, 2007, and 2008 three new systems were launched in the cities of Lyon (*Vélo'v*), Paris (*Vélib'*), and La Rochelle (*Yélo*), respectively, incorporating new technological innovations. These three French systems were equipped with Global Positioning System (GPS) tracking. The use of GPS technology made it possible for the service providers to assess real time data about the bicycles' location as well as the status of the stations and information about the users (DeMaio, 2009; Melo and Maia, 2013). Due to these improvements, some authors even categorize these systems as part of a so-called 3rd generation plus (Assim Mohammed, 2017; Bieliński and Ważna, 2018). These innovations represented a significative advance for the systems' operation and management, which also contributed to raise users' satisfaction. Thereafter, the system spread rapidly across the world.

South Korea started their system *Nubija* in Changwon city in 2008. In the same year, China implemented the *HZ Bike* in Hangzhou city. Taiwan, then, introduced in the cities of Kaohsiung and Taipei the *C-Bike* and *YouBike* systems in 2009. The first American system was debuted in Washington (USA) in 2008 and received the name *SmartBikeDC* (later renamed as *Capital Bikeshare*) (Shaheen et al., 2010; Melo and Maia, 2013).

No longer after this, systems started to blowout in other continents. Auckland (New Zealand), in 2008, inaugurated the *NextBike*, Australia, in 2010, launched the *CityCycle* in Brisbane and the *Melbourne Bike Share*, in Melbourne. In Latin America, the *B'easy* was introduced in Santiago, Chile, in 2010; the *EcoBici* was presented in Mexico City, Mexico, in 2010; the *Mejor enBicici* (which became *EcoBici* later) was announced in Buenos Aires, Argentina, in 2010; the *EnCicla*, was created in Medellín, Colombia in 2011; the *UseBike*, was introduced in Sao Paulo, Brazil in 2008 and then replaced by the *Bike Sampa* in 2012; what also happened in Rio de Janeiro, Brazil, with the *PedalaRio*, which has started in 2008 and was substituted by the *BikeRio* in 2011 (Melo and Maia, 2013).

It was in Montreal (Canada) that, in 2009, the system, called *BIXI*, that was going to become the first system of the 4th generation appeared. The 4th generation of PBSS introduced some new features into the systems being the main one the use of smart cards. The use of smart cards enabled the implementation of a unified payment structure and provision of real time information, not only about bicycles and stations, but also on the integration with other transit modes. More than this, the improvements included technologies that facilitated the demand-responsive redistribution and the value-pricing to stimulate self-rebalancing. Ultimately, these advances gave to the users the chance to better plan their trips as they could easily access all the information through their mobile-phone and, at the same time, it increased the control of the operators over their systems.

Aside from these IT-based innovations, the 4th generation was responsible to reintroduce the FFBS operation, eliminating the need of stations. The use of smart lock mechanisms accessed by mobile-phone applications brought back the flexibility and convenience of the systems of the 1st generation and, at once, maintained some level of control over the bicycles and users, which previously would only be possible in systems with stations (Melo and Maia, 2013; Shaheen et al., 2010; Shaheen et al., 2013).

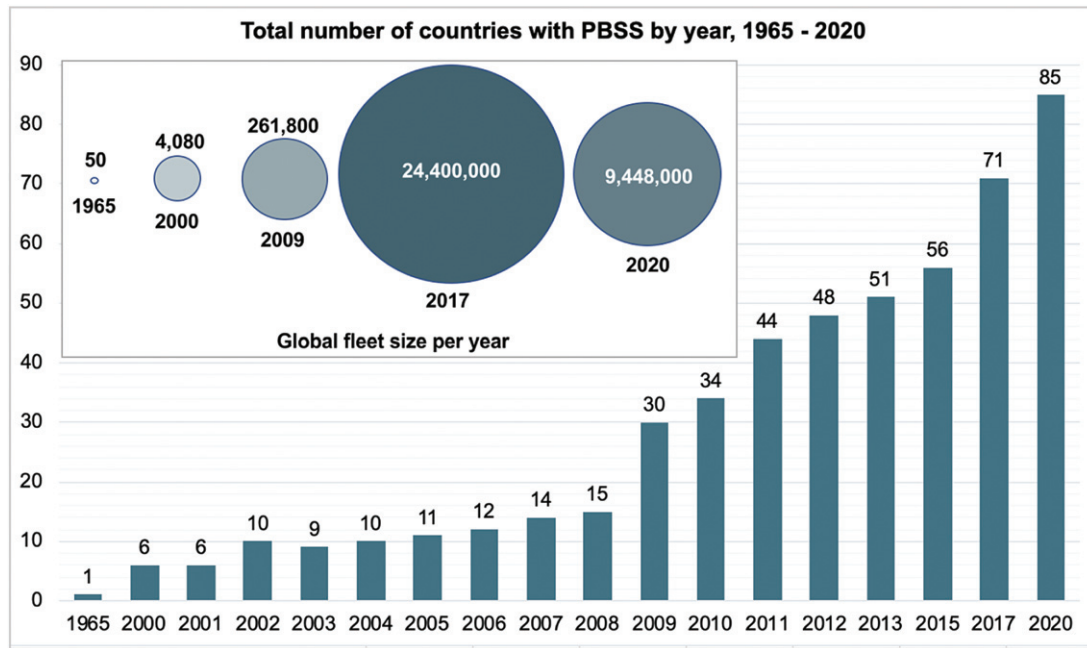


Figure 1 Number of countries with PBSS (1965–2020) and global fleet size variations. Source: Meddin et al. (2020), Earth Policy Institute (2013) and Ferreira et al. (2020).

Besides the before mentioned flexibility and convenience that the absence of docking stations represented, the new version of FFBS brought a number of other advantages in comparison with the SBSS. From the operators' point of view, the absence of stations and payment kiosks made the implementation cost lower. The GPS technology enabled real time management and reallocation. It also helped to reduce the occurrence of theft. From the users' viewpoint, the systems became much more attractive. If before the users had to find a station and would face the uncertainty regarding the availability of bicycles or of free docks to return their bicycles after use, now with the FFBS, parking became no longer an issue. The possibility of parking anywhere made the FFBS a perfect alternative for door-to-door travel (Pal and Zhang, 2017).

Due to all these benefits, the systems of FFBS type became very popular and experienced an explosive growth. In 2017, there were more than 24 million PBSS in the world being a huge parcel of them in China. Nevertheless, the rapid growth brought a lot of problems. The lack of government control led to a chaotic scenario. Littering, inappropriate parking and oversupply caused an increase in risk of accidents and a reduction of the capacity of the walkways and roads, which ultimately contributed to lower city's walk ability, lesser users' satisfaction and give a bad image to cycling (Bieliński and Ważna, 2018). In order to overcome these issues, many cities implemented regulations, which led to a better equilibrium between the demand and supply, thus causing a drop in the total number of shared bicycles, from more than 24 million in 2017 to less than 10 million in 2020 (as shown in Fig. 1).

Fig. 1 presents an overview of the historical growth in the numbers of PBBS. As can be seen, the number of countries with PBBS has been growing consistently over the years. The growth curve became sharper after the 3rd generation of bicycles were launched in 2005 doubling in 2009 when the 4th generation emerged. The rise in the number of countries with this kind of system has a direct connection with the development and popularization of technologies as smartphones and the Internet.

Recently, the 4th generation also incorporated electric-assisted bicycles and hybrid operation (HBSS), which consists of the combination of the characteristics of the two types of PBSS. *Norisbike* in Nürnberg or the *MVGRad* in Munich are examples of HBSS. In the systems of this type the users can still start and finish a ride anywhere through the mobile phone application, just as in the FFBS, however, the system also provides some stations. The main functions of the stations are (1) recharge the batteries of the electric-assisted bicycles and (2) guarantee more availability of bicycles in strategic locations such as close by transit stations making it easier for the users to find a bicycle (or a place to park it) in these locations (Albiński et al., 2018; Bieliński and Ważna, 2019).

Even though there is no unanimity on what a 5th generation may be like, the fact is that the PBSS are still in ascension. Globally, it is estimated that, by August 2020, there were 2018 PBSS in operation. These systems were located in 1583 cities in 85 different countries around the globe. They counted together with around 9.4 million selfservice public bicycles and pedelecs (electrically assisted bicycles). There were also 318 systems planned or under construction and 691 systems that were no longer operating or that were temporarily closed (Meddin et al., 2020). The spatial distribution of these systems around the globe can be seen in Fig. 2.

As can be seen in Fig. 2, most of the systems are located in Europe and Asia. Europe counts with 41.9% of the systems in the world. It represents 845 systems distributed in 749 cities of 40 different countries. Considering that there are 44 countries in Europe according to the United Nations (1999), one can see that 91% of the European countries have at least one PBBS. Asia, on the other hand, has the biggest portion of the systems (43.6% of total), however, unlike in Europe, in Asia, the systems are not evenly distributed over the territory. This continent has 48 countries but only 24 of them (50%) have PBSS, being the vast majority of them

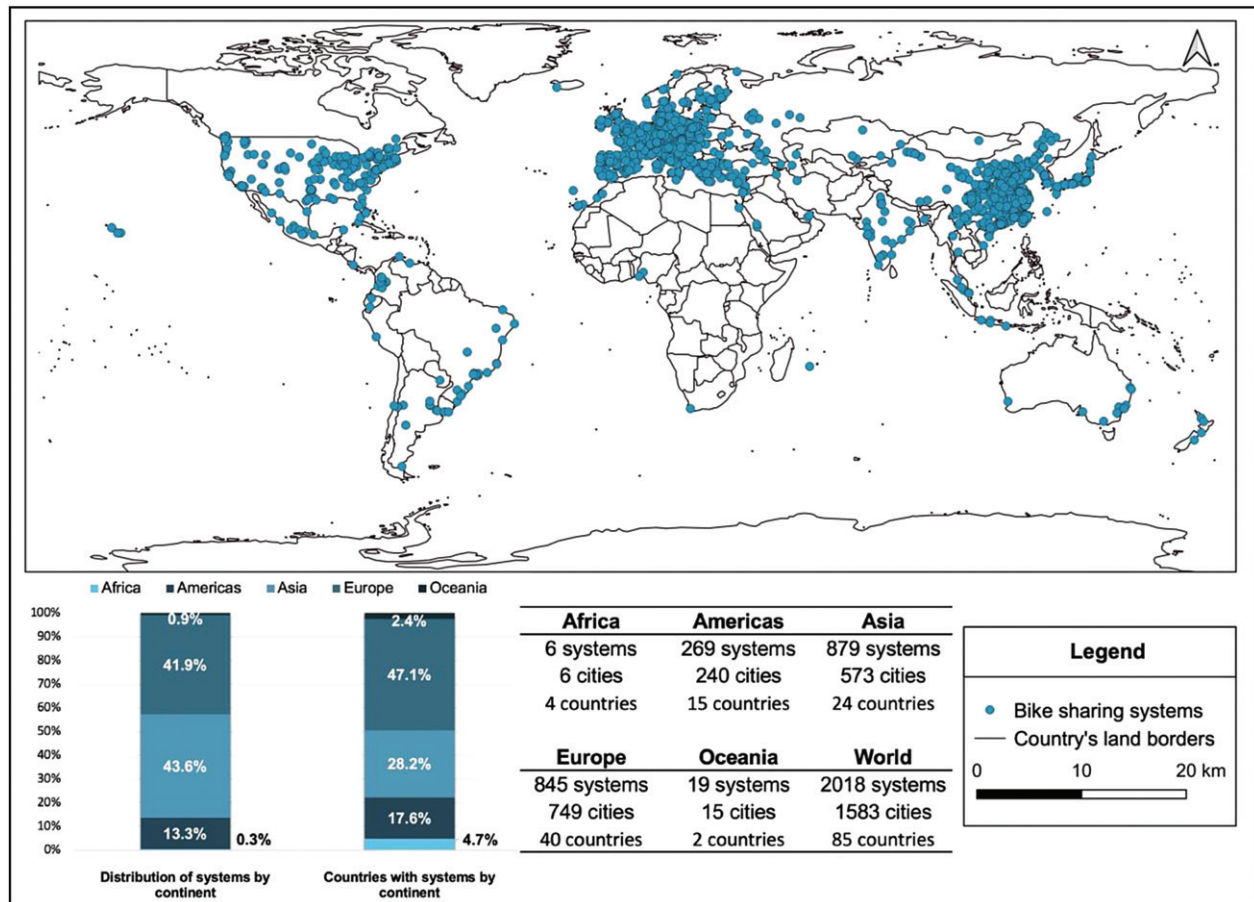


Figure 2 Spatial distribution of the bike-sharing systems in the world. Source: *Meddin et al. (2020)*.

located in Eastern Asia. China alone has 709 systems, which represent almost 81% of all the systems of this continent. The Americas have 269 systems located in 15 different countries, which means that 43% of the countries of these continents have at least one system. The majority of these systems are located in Northern America (73%), specifically in the USA (176 systems). Africa and Oceania are the continents with the lowest number of systems. Despite having 54 countries, only four of them have PBSS. Oceania, on the other hand, has systems only in two of its 14 countries, however, it is interesting to notice that there are 19 systems in the only two densely populated countries and all the other countries that have no systems are small islands.

In brief, we can say that almost half of the countries in the world have at least one PBSS. It gives us the idea of how much this kind of scheme spread globally and gained relevance in a short period of time. As previously stated, the first PBSS was launched a little over five decades ago and since then, many systems have been created but many have also ended up failing. Hereupon, it is important to understand that the rise and fall of systems are an intrinsic part of the process of development of this type of system. Thus, even if there is not a formula for success, it is possible to learn from previous experiences in order to foresee possible issues that can be avoided and be aware of factors that may contribute to the success.

The “Success Factors” for Public Bike Sharing Systems

When we talk about “the success factors” for bike-sharing it is important to understand that these factors are not universal. There is not a formula of success that can be applied to any context and that can guarantee the success of the system. There are, though, some essential elements, such as the provision of infrastructure, that are, indeed, key factors for success (in fact, they are almost a precondition to make cycle possible) that must be taken into consideration during the planning process. The elements that will contribute to increasing the system’s chances of success will deeply depend of the purpose of the system and of the particular characteristics of the city where it is going to be implemented (*Chardon et al., 2017*). All these parameters as well as the system feasibility, its detailed design and its business and financial model will be defined during the planning process (*Gauthier, 2013*).

The purpose of the system refers to what was the main intention behind the system implementation. For instance, if the bike-sharing system was projected aiming to reduce pollutant emissions and traffic congestion in the city center, it can be considered successful by the local government even if its benefits are restricted to a specific part of the city or if the system is deficient in other aspects.

Besides the purpose of the system, the factors that can lead to its success are strongly related to the characteristics of the city where it is going to be implemented. If on the one hand, bike-sharing systems are considered a more affordable mode of transport, on the other hand, studies have shown that the systems do not necessarily promote social equity. More than this, depending on how it is designed it may highlight and deepen some preexisting class, gender, and race inequalities (Chardon et al., 2017; Ricci, 2015; Ursaki and Aultman-Hall, 2015). The same logic applies to health benefits. If it is undeniable that the physical activity promoted by cycling can contribute to enhancing the population quality of life and health, researches have concluded that in some contexts the benefits of cycling do not overcome the harm to health that the increased exposure to pollutants and to the risk of traffic accidents or even the reduction of physical activity levels that is perceived when cycling replaces walking represents (Chardon et al., 2017; Fishman et al., 2015; Qian and Wu, 2019). Even some commonly mentioned benefits of bike-sharing such as reduction of pollutant emissions, improvement in the traffic congestion, an increase of public transit ridership may not be true in all contexts (Fishman et al., 2014; Ricci, 2015).

Therefore, we can say that the success of a system will heavily rely on its context and that the solutions will have to be adjusted accordingly. However, it is possible to list, in a general way, some factors that are basic and can be considered as the ground where the systems should be built over. Below we list and briefly discuss some of the factors that we consider fundamental for the success of a PBSS.

Provision of Infrastructure

It is unquestionable that the success of cycling heavily relies on the provision of adequate infrastructure. On the one hand, the PBSS itself must guarantee that both the stations and the bicycles are easy to use and have good quality. It is also fundamental that the stations, in the case of SBBS, are well located and properly integrated into the urban environment (placed in safe locations and close to points of interest, not interfering in the free movement of pedestrians, protected from the general traffic, etc). The Institute for Transportation and Development Policy offers some guidelines on that matter stating that a well-located station should be no more than 300 m from each other (10–16 stations for every square kilometer) and a minimum of 10km² must be covered. The number of bicycles per 1000 residents should vary between 10 and 30 for optimal usage (ITDP, 2018).

On the other hand, it is also crucial that the municipality offers adequate public infrastructure such as bicycle lanes and parking areas that allow and facilitate cycling, thus, creating a friendly environment for PBSS use. The provision of an urban environment properly designed to accommodating cycling can substantially affect the ridership (Marqués et al., 2015; Ricci, 2015). Thus, the infrastructure has to connect the users not only with other stations but also with main landmarks and points of interest such as commercial centers, schools, universities, and residential clusters. Besides the convenience of the location, it also has to offer a safe environment for cycling protecting the cyclist from general traffic (Fishman et al., 2012).

In addition, to convenience and security, the infrastructure must consider the users' comfort. In cities where the natural environment is not encouraging to cycling (due to weather or hilly terrain), for instance, the issue can be minimized with the provision of special infrastructures such as sheltered, shaded, cooled, or heated cycle lanes or other solutions such as wide cycle lanes or even cyclocables, lifts or electric-assisted bicycles (Campbell et al., 2016). Hereupon, it is important to notice that all the infrastructures and operations are related to financial sustainability. Public Private Partnership approach is normally applied in many cities such as in New York, Paris, and London. YouBike system in Taipei, for example, has a 9 years concession to Giant Group with PPP approach in which the Taipei City Government provides funding of USD20 million for the infrastructure (vehicles and rental stations) while Giant Group takes the responsibility for operation (Chung and Huang, 2017).

Integration With Other Modes of Transport

The possibility to use bicycles is well known as a factor that may positively impact the attractiveness of public transportation, improving the users' convenience and efficiency perception (Martens, 2007; Krizek and Stonebraker, 2010; Sebok and Chang, 2018), especially when the distance between the origin or destination points and the public transportation station/stops is not considered short enough to be done by walk (Kittelson et al., 2003; Shaheen et al., 2010).

In this sense, to make the PBSS attractive, it is necessary to integrate it with the existing transit system creating an intermodal network that can provide door-to-door transportation. Thus, the increase of ridership and user satisfaction with PBSS can be achieved by the provision of docking stations or parking spaces connected or conveniently placed close by transit stations, especially when it is combined with a well-designed and consisted cycling infrastructure (Martens, 2004; Sebok and Chang, 2018). Moreover, the integration of PBSS with public transit can contribute to increasing the coverage of both systems, reducing in some cases the need for public transit systems expansion, hence, lowering the use of public resources. The provision of this intermodal network can enhance accessibility as people can reach places that are distant to be reached by walking (ITDP, 2018). In this sense, include PBSS into the Mobility as a Service plan is also an essential strategy for the success of both PBSS and MaaS (Chang et al., 2019).

Beyond the physical integration, it is relevant to consider the unification of payment. The use of universal rechargeable smart cards that can be used interchangeably to pay bus, rail, and PBSS trips is a good solution that is already being used in some cities to improve the service and user convenience (ITDP, 2018; Ji et al., 2017). In fact, the implementation of an easy-to-use registration and payment structure could be considered as a “success factor” itself as it is decisive for the success of a system. Registration and payment process should not only enable the integration of PBSS with the transportation network, but it should also bring solutions that favor small systems, tourists, and low-income users in a way that optimizes the usage while fostering equity.

Service Regulation

Another key element to increase the chances of success of a PBSS is the involvement of the government. As previously mentioned, the public power plays an important role as the provision of infrastructure for cycling is its responsibility. However, the infrastructure such as bicycle lanes and parking areas is not the only thing that public authorities should be in charge of. Governments should consider cycling as a fundamental part of urban planning and should take cycling as an indisputable part of the cities' Mobility Master Plan (Grafl et al., 2018; Liu et al., 2012).

In this sense, it is the government's responsibility to create, implement, and enforce regulations and policies that can guarantee the proper use of the urban environment. Besides the effective management of the public space, the regulations should establish the minimum requirements and standards to ensure the quality of the service that is provided by the private companies that operate PBSS schemes while fostering equity and accessibility as well as protect the users (ITDP, 2018).

As most systems operate through a Public-Private Partnership framework, the government can actively participate in all the stages of its implementation. Nevertheless, as before mentioned, there is no universal formula that can be applied to all the contexts and it is also valid when we talk about regulations. Thus, policies should address local issues.

In this sense, application of penalties and fines for users that did not comply with the regulations; delimitation of regions to be served and of maximum distance between stations; definition of desired density and the minimum number of bicycles per resident; description of minimum quality requirements for bicycles and stations, control of fare price or even permission to service providers implement a scheme of incentives to encourage the users to return the bicycles uphill when the city is not built over flat terrain or to a certain location as a way to foster social equity are valid (ITDP, 2018; Chardon, 2019).

Use of New Technologies

One of the reasons behind the quick spread of the PBSS over the world is the advance and popularization of technology. The development of new technologies and concepts such as the Internet of Things (IoT), Artificial Intelligence (AI), Information Communication Technology (ICT), Mobility as a Service (MaaS), Transit-Oriented Development (TOD) will strongly affect the future of PBSS.

The development of new technologies will facilitate the collection and processing of real-time big data. The possibility to use this type of data and the improvement and democratization of the use of existing technologies such as GPS, Internet, radiofrequency, and Bluetooth will enable a smart and customized management of the whole system including the real-time reallocation of bicycles, better integration of the system with the other modes of transportation or even among different systems of PBSS that operate within the same area, use of smart cards, solutions for easy registration and payment, implementation of electric or virtual fences to delimit parking areas, installation of smart lockers on the bicycles and use of e-bikes, and other smart solutions (Zhao et al., 2018; Zhang et al., 2019; Zhou et al., 2020). The use of these technologies may help to overcome some challenges that PBSS often face, thus enhancing user experience and service provider/government control. It is also worth noting that private sector involvement in PBSS development has allowed the flexibility of adopting new technologies for PBSS infrastructures and operations.

Provision of Information and Education

Cycling can bring numerous benefits, both to the individuals and the city as a whole. PBSS can play an important role, positively affecting the image of cycling, due to its potential to normalize bicycles as a mode of transport and cycling in general (López and Wong, 2017; Ricci, 2015).

Provision of information and education can be considered as a factor that may potentially contribute to the success of PBSS because it, directly and indirectly, affects the public perception about the system as well as the sense of safety, convenience, and satisfaction of the users. It is up to both the government and the service provider to promote campaigns and to employ marketing strategies to educate, inform, and encourage the use of PBSS, creating a good image of cycling.

Besides the provision of information and education to promote cycling and encourage the use of PBSS, it is also important to educate the general population, especially the drivers of motorized vehicles, as their disregard about cyclists can represent an increase in the accident levels involving cyclists and raises the PBSS risk perception (López and Wong, 2017).

Thus, governments and service providers should actively promote cycling, inform about its benefits, and educate all road users to respect the traffic laws and the regulations regarding the use of public space. It also includes the education of the cyclist about the safety of pedestrians, especially the ones that belong to vulnerable groups such as elders, people with disabilities, children, and pregnant women, for example. These initiatives can contribute to creating a safer and friendlier environment for cycling, which can increase ridership and the users' sense of safety (Grafl et al., 2018).

Beyond raising awareness about road safety and the benefits of cycling for the whole society, it is also fundamental that the public power acts assertively enforcing and punishing the road users that disrespect the laws or engage in negative behaviors towards the cyclists or the cycling infrastructure (Grafl et al., 2018; Unwin, 1995).

Conclusions

PBSS is a scheme that appeared for the first time less than six decades ago as a small initiative of a group of activists in the Netherlands. Since then, this scheme has passed for many improvements and it is now present in 85 out of the 195 countries

around the world. The system has evolved over these years passing through four generations. Along these years three different configurations of operation were proposed: with docks, without docks, and a hybrid version that combines the other two. All of these three configurations are still in operation, and in some cases, they coexist within the cities.

The quick dissemination of this type of system around the globe shows how popular this scheme has become. However, the systems also faced many challenges and in several situations, they ended up failing. Even though it is not possible to point out exactly which factors can lead to success in all the cases, it is possible to evaluate which general characteristics may have contributed to this success.

Given this, five factors have been identified as the key elements that may affect positively the chances of success of a system. The factors are the provision of infrastructure, the integration with other modes of transportation, the service regulation, the use of new technologies, and the provision of information and education. It is also worth mentioning that Private–Public Partnership approach is normally applied in development of PBSS.

Finally, we emphasize that the development of the PBSS is an ongoing process and that there is much more to learn and to understand about it while the system keeps evolving and the success cases consolidate over the years. Nevertheless, we believe that the overview of the system's importance, its historical overview, and these five factors that we listed here may contribute to paving the way in which new knowledge will be built over.

References

- Albiński, S., Fontaine, P., Minner, S., 2018. Performance analysis of a hybrid bike sharing system: A service-level-based approach under censored demand observations. *Transport. Res. E: Log. Transport. Rev.* 116, 59–69.
- Assim Mohammed, M., 2017. Next Generation Bike Sharing Design Concept using axiomatic design theory. Master thesis. University of Vaasa, Vaasa, Finland.
- Belinski, T., Wazna, A., 2018. New generation of bike-sharing systems in China: lessons for European cities. *J. Manage. Financia* 11 (33).
- Bielinski, T., Wazna, A., 2019. Hybridizing bike-sharing systems: the way to improve mobility in smart cities. *Transp. Econ. Log.* 79, 53–63.
- Buehler, R., 2012. Determinants of bicycle commuting in the Washington, DC region: the role of bicycle parking, cyclist showers, and free car parking at work. *Transp. Res. D* 17, 525e531.
- Caggiani, L., Camporeale, R., Ottomaneli, M., Szeto, W.Y., 2018. A modeling framework for the dynamic management of free-floating bike-sharing systems. *Transp. Res. Part C: Emerg. Technol.* 87, 159–182.
- Campbell, A.A., Cherry, C.R., Ryerson, M.S., Yang, X., 2016. Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. *Transp. Res. Part C: Emerg. Technol.* 67, 399–414.
- Chang, S.K.J., Chen, H.Y., Chen, H.C., 2019. Mobility as a service policy planning, deployments and trials in Taiwan. *IATSS Res.* 43 (4), 210–218.
- Chardon, C.M., 2019. The contradictions of bike-share benefits, purposes and outcomes. *Transp. Res. Part A: Policy. Pract.* 121, 401–419.
- Chardon, C.M., Caruso, G., Thomas, I., 2017. Bicycle sharing system 'success' determinants. *Transp. res. Part A: Policy. Pract.* 100, 202–214.
- Chung, C.L., Huang, C.L., 2017. The effects of fare adjustment on bikesharing—a case study of big data analysis on Taipei YouBike, Proceedings of the 15th ITS Asia Pacific Forum, Hong Kong.
- DeMaio, Paul., 2009. Bike-sharing: history, impacts, models of provision, and future. *J. Public. Trans.* 12 (4), 41–56.
- Ellison, R.B., Greaves, S., 2011. Travel time competitiveness of cycling in Sydney, Australia. *Transport. Res. Rec.* 2247 (1), 99–108.
- Ferreira, A.F., Chang, S.K.J., Liu, Z., Greiner, M., Chen, Y., 2020. O sistema dockless de compartilhamento de bicicletas na Ásia: desafios e lições. In: *Bicicleta nas cidades: experiências de compartilhamento, diversidade e tecnologia* / Jacques Rancière; organized by Victor Andrade, Letícia Quintanilha, Juliana Decastro. Belo Horizonte, Relicário.
- Fishman, E., Washington, S., Haworth, N., 2012. Barriers and facilitators to public bicycle scheme use: A qualitative approach. *Transp. Res. Part F: Traffic Psychol. Behav.* 15 (6), 686–698.
- Fishman, E., Washington, S., Haworth, N., 2014. Bike share's impact on car use: Evidence from the United States, Great Britain, and Australia. *Transp. Res. Part D: Transp. Environ.* 31, 13–20.
- Fishman, E., Washington, S., Haworth, N., 2015. Bikeshare's impact on active travel: evidence from the United States, Great Britain, and Australia. *J. Transp. Health* 2 (2), 135–142.
- Graff, K., Bunte, H., Dziekan, K., Haulbold, H., Neun, M., 2018. Framing the third cycling century. Bridging the gap between research and practice. *European Cyclists' Federation -ECF-*, Brüssel.
- Granville, S., Rait, F., Barber, M., Laird, A., 2001. Sharing road space: Drivers and cyclists as equal road users. Scottish Executive Central Research Unit, Edinburgh Scotland.
- Gauthier, A., 2013. The Bike-Share Planning Guide. Institute for Transportation & Development Policy, New York.
- Institute for Transportation and Development Policy, 2018. Optimizing Dockless Bikeshare for Cities. New York: Institute for Transportation & Development Policy. <https://como.org.uk/wp-content/uploads/2018/06/ITDP-Optimizing-Dockless-Bikeshare-for-Cities-1.pdf> (accessed on 12 June of 2019).
- Jacobsen, P.L., 2015. Safety in numbers: more walkers and bicyclists, safer walking and bicycling. *Inj. Prev.* 21 (4), 271–275.
- Ji, Y., Fan, Y., Ermagun, A., Cao, X., Wang, W., Das, K., 2017. Public bicycle as a feeder mode to rail transit in China: The role of gender, age, income, trip purpose, and bicycle theft experience. *Int. J. Sustain. Transp.* 11 (4), 308–317.
- Kabak, M., Erbaş, M., Çetinkaya, C., Özceylan, E., 2018. A GIS-based MCDM approach for the evaluation of bike-share stations. *J. Clean. Prod.* 201, 49–60.
- Krizek, K.J., Stonebraker, E.W., 2010. Bicycling and transit: A marriage unrealized. *Transp. Res. Rec.* 2144 (1), 161–167.
- Kittelson & Associates, KFH Group, Parsons Brinkerhoff Quade and Douglas Inc., & JHunter-Zaworski, K. (2003). *Transit Capacity and Quality of Service Manual*, Second ed. Transit Cooperative Research Program TCRP, Washington, DC.
- Kumar, A., Nguyen, V.A., Teo, K.M., 2016. Commuter cycling policy in Singapore: a farecard data analytics-based approach. *Ann. Oper. Res.* 236 (1), 57–73, doi:10.1007/s10479-014-1585-7.
- Li, Z., Wang, W., Yang, C., Jiang, G., 2013. Exploring the causal relationship between bicycle choice and journey chain pattern. *Transp. Policy* 29, 170e177.
- Liu, Z., Jia, X., Cheng, W., 2012. Solving the last mile problem: ensure the success of public bicycle system in Beijing. *Procedia-Soc. Behav. Sci.* 43, 73–78.
- López, M.C.R., Wong, Y.D., 2017. Attitudes towards active mobility in Singapore: a qualitative study. *Case Stud. Transp. Policy* 5 (4), 662–670.
- Ma, Y., Lan, J., Thornton, T., Mangalagiu, D., Zhu, D., 2018. Challenges of Collaborative Governance in the Sharing Economy: The case of free-floating bike sharing in Shanghai. *J. Clean. Prod.* 197, 356–365.
- Marqués, R., Hernández-Herrador, V., Calvo-Salazar, M., García-Cebrián, J.A., 2015. How infrastructure can promote cycling in cities: Lessons from Seville. *Res. Transp. Econ.* 53, 31–44.
- Martens, K., 2004. The bicycle as a feeder mode: experiences from three European countries. *Transp. Res. Part D: Transp. Environ.* 9 (4), 281–294.
- Martens, K., 2007. Promoting bike-and-ride: The Dutch experience. *Transp. Res. Part A: Policy. Pract.* 41 (4), 326–338.

- Melo, M.F.S., Maia, M.L.A., 2013. Sistema de bicicletas públicas: um balanço de sua evolução e sua integração na rede de transporte público. In: XXVII Congresso Nacional de Pesquisa e Ensino em Transportes, Belém. XXVII ANPET.
- Millward, H., Spinney, J., Scott, D., 2013. Active-transport walking behavior: destinations, durations, distances. *J. Transp. Geogr.* 28 (0), 101e110.
- Ogilvie, D., Egan, M., Hamilton, V., Petticrew, M., 2004. Promoting walking and cycling as an alternative to using cars: systematic review. *BMJ* 329 (7469), 763.
- Pal, A., Zhang, Y., 2017. Free-floating bike sharing: solving real-life large-scale static rebalancing problems. *Transp. Res. Part C: Emerg. Technol.s* 80, 92–116.
- Qian, X., Wu, Y., 2019. Assessment for health equity of PM2.5 exposure in bikeshare systems: The case of Divvy in Chicago. *J. Transp. Health* 14, 100596.
- Ricci, M., 2015. Bike sharing: A review of evidence on impacts and processes of implementation and operation. *Res. Transp. Bus. Manag.* 15, 28–38.
- Rojas, M.C., Wong, Y.D., 2017. Attitudes towards active mobility in Singapore: a qualitative study. *Case Stud. Transp. Policy* 5, 662–670.
- Schabus, E., 2020. Promoting active travel for all in European urban regions. A review of evaluated initiatives. In *International Cycling Conference 2017. Bridging the gap between research and practice 19-21 September 2017, Mannheim, Germany. Conference Proceedings*.
- Sebok, M., Chang, S.K.J., 2018. Assessment of Infrastructure and Institutional Framework for Integration of Cycling and Mass Rapid Transit System: The case study of Taipei. *Velo-city 2018. Rio de Janeiro, Brazil*.
- Shaheen, S.A., Cohen, A.P., Martin, E.W., 2013. Public bikesharing in North America: early operator understanding and emerging trends. *Transp. Res. Rec.* 2387 (1), 83–92.
- Shaheen, S.A., Guzman, S., Zhang, H., 2010. Bikesharing in Europe, the Americas, and Asia: past, present, and future. *Transp. Res. Rec.* 2143 (1), 159–167.
- Shen, Y., Zhang, X., Zhao, J., 2018. Understanding the usage of dockless bike sharing in Singapore. *Int. J. Sustain. Transp.* 12 (9), 686–700.
- The Meddin Bike-sharing World Map. Russell Meddin, Paul DeMaio, Oliver O'Brien, Renata Rabello, Chumin Yu, Rahil Gupta, Jess Seamon <http://bikesharingworldmap.com/> (accessed on August 10, 2020).
- Toronto (Ont.), & Toronto Public Health, 2012. Road to Health: Improving Walking and Cycling in Toronto. Toronto, Ontario, Canada, Available at <http://www.toronto.ca/health> (accessed on August 10, 2020).
- United Nations, 1999. Countries or areas/geographical regions. <https://unstats.un.org/unsd/methodology/m49/> (accessed on August 2, 2020).
- Unwin, N.C., 1995. Promoting the public health benefits of cycling. *Public Health*, 109(1), 41–46.
- Ursaki, J., Aultman-Hall, L., 2015. Quantifying the equity of bikeshare access in US cities (No. TRC Report 15-011). University of Vermont. Transportation Research Center.
- Van Wee, B., Rietveld, P., Meurs, H., 2006. Is average daily travel time expenditure constant? in search of explanations for an increase in average travel time. *J. Transp. Geogr.* 14 (2), 109–122.
- Zhang, L., Zhang, J., Duan, Z.Y., Bryde, D., 2015. Sustainable bike-sharing systems: characteristics and commonalities across cases in urban China. *J. Clean. Prod.* 97, 124–133.
- Zhang, Y., Lin, D., Mi, Z., 2019. Electric fence planning for dockless bike-sharing services. *J. Clean. Prod.* 206, 383–393, doi:10.1016/j.jclepro.2018.09.215.
- Zhao, N., Zhang, X., Banks, M.S., Xiong, M., 2018. Bicycle Sharing in China: Past, Present, And Future. *SAIS 2018 Proceedings, USA 23–24*, <https://aisel.aisnet.org/sais2018/11> (accessed on June 10, 2019).
- Zhou, T., Law, K.M., Yung, K.L., 2020. An empirical analysis of intention of use for bike-sharing system in China through machine learning techniques. *Enterp. Inf. Syst.* 1–22.

Further Reading

- Berger, R., 2016. Bike Sharing 5.0. Market Insights and Outlook. Available online: https://www.rolandberger.com/publications/publication_pdf/roland_berger_study_bike_sharing_5_0.pdf (accessed on 10 August 2020).
- Chen, F., Turon, K., Ktos, M., Czech, P., Pamula, W., Sierpiński, G., 2018. Fifth-generation bikesharing systems: examples from Poland and China. *Sci. J. Silesian Univ. Technol.* 99, 05–13, doi:10.20858/sjsutst.2018.99.1, ISSN:0209-3324.
- O'Brien, O., Cheshire, J., Batty, M., 2014. Mining bicycle sharing data for generating insights into sustainable transport systems. *J. Transp. Geogr.* 34, 262–273.

Airspace Systems Technologies—Overview and Opportunities

Banavar Sridhar, Gano B. Chatterji, NASA Ames Research Center, Moffett Field, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	363
Airspace Operations	364
Air Traffic Control	365
Traffic Flow Management	366
National Traffic Flow Management	366
Regional Traffic Flow Management	366
Terminal Area Operations	368
Surface Traffic Operations	368
Advanced Airspace Operations Concepts	369
Integrated Demand Management	370
Integrated Arrival/Departure/Surface Metroplex Traffic Management	370
Applied Traffic Flow Management (ATD-3)	371
Environmentally Friendly Operations	371
NextGen Implementation	372
Automatic Dependent Surveillance-Broadcast (ADS-B)	372
Performance-Based Navigation (PBN)	373
Weather Integration	373
Data Communications	374
Operations	374
Optimal Descent Trajectories (ODT)	374
Wind-Optimal and User-Preferred Routes	374
Surface	375
Integrated Technology	375
Global Harmonization	375
Newly Emerging Technologies	376
Unmanned Aerial System Traffic Management	376
Urban Air Mobility (UAM)	376
Unmanned Aircraft Systems	377
Conclusions	377
Acknowledgments	377
References	377

Introduction

Civil Aviation is a vital sector of the U.S. economy. In 2013, air transportation and related industries in the U.S. generated an output of 1.55 trillion dollars employing 10.5 million people (TEICAUE, 2011). The National Airspace System (NAS) provides a network of airspace, air navigation facilities, equipment, services (including information services), airports and landing areas, aeronautical charts, rules and regulations, procedures, technical information and personnel needed for the operation of civil aviation in the U.S. Included in the NAS are system components shared jointly with the military. In 2018, the NAS managed the progress of nearly 16.1 million flights with approximately 5000 flights in the air at any given moment. The capacity of the airspace in the current system is primarily limited by the ability of the air traffic controllers to maintain situational awareness and provide separation services to the aircraft. Additionally, wind, weather, and operational uncertainties add to the inefficiencies that result in traffic demand exceeding airspace and airport capacity. Traffic flow management initiatives such as holding aircraft on the ground, metering aircraft in the air and rerouting flights for mitigating demand-capacity imbalances result in significant delays, lost productivity, greater amount of noise pollution and increase in carbon dioxide and other greenhouse gas emissions. With the expected increase in traffic density and greater degree of automation in the future, airspace capacity will more likely be limited by the ability of the automation to detect conflicts between aircraft and to resolve them in a safe manner. Airport capacity will continue to be limited by the existing number and layout of runways and taxiways and by the inability to extend or construct new ones because of availability of land, public support and construction cost. Federal Aviation Administration (FAA) estimated the cost of delays to airlines in 2018 at 6.4 billion dollars (TEICAUE, 2011). The air traffic demand generally follows trends in the national economic growth. Air traffic demand went down during the years 2000–02, when the number of enplanements fell from 640 million to 574 million. It took another 2 years for the traffic to bounce back to its 2000 level. Currently, there is an unprecedented decline of between 56 and 60% in air traffic demand

due to the outbreak of corona virus (COVID-19) pandemic in China and its rapid spread to the rest of the world (ENC, 2020). The recovery of traffic from the effects of the pandemic depends on the ability of airlines to assure passengers that air travel is safe. Before the pandemic, the 2020 FAA forecast predicted the air traffic demand to grow over the next 20 years with an annual growth rate of about 2% (FAA, 2020). This anticipated increase in traffic will put a further strain on the airports and the airspace; it will result in large delays, airline schedule breakdown and adverse environmental impact.

The capacity improvements needed for meeting the future air transportation demand might not be achievable by incremental changes to the current operations. Major changes, including increased level of automation in a safety-critical environment with new roles for automation, controllers and pilots, are needed. To lead this transformation with the charter of planning and coordinating the development of the Next Generation Air Transportation System (NextGen), the United States Congress established the public--private multiagency Joint Planning and Development Office (JPDO) in 2003. This office included Department of Transportation, Department of Defense, Department of Commerce, Department of Homeland Security, FAA, National Aeronautics and Space Administration (NASA) and White House Office of Science and Technology. JPDO ended its mission in 2014. JPDO developed a comprehensive concept-of-operations (JPDO, 2009) that advocates operations based on four-dimensional aircraft trajectories. NASA continues to collaborate with the FAA and other industry partners to develop advanced automation tools to provide air traffic controllers, pilots and other airspace users accurate real-time information about traffic flow, weather and routing. The greater precision of this information is a key enabler of NextGen.

This article provides background about the current air traffic system and discusses new airspace technologies and efforts to introduce them to modernize the system to meet new challenges facing the air traffic system. The rest of the article is organized as follows. Section II provides an overview of operations typical of the air transportation system at the end of 2019, introduces air traffic control and traffic flow management and describes issues facing the 2019-era system. New concepts in air traffic management and the key challenges in designing the future system are described in Section III. Section IV describes the implementation of the advanced concepts in NextGen. Concluding remarks are provided in Section V.

Airspace Operations

Airspace operations are a distributed, hierarchical process with multiple decisionmakers. Airline flights operating under Instrument Flight Rules (IFR) are accomplished by the interaction of three participants: pilot, controller, and dispatcher. Fig. 1 shows the interaction between these three participants. The pilot is primarily responsible for safe operation of the aircraft. The controller is responsible for keeping the aircraft separated from all other aircraft in all phases of flight, including taxi on the airport surface, climb and descent in the terminal area, and climb, cruise and descent in en route airspace. Separation is assured by controllers talking to pilots via two-way radio, and pilots following their separation advisories. Like the pilot, the dispatcher is a certified airman. The dispatcher is responsible for flight, crew and maintenance schedules, creating and filing flight-plans, weight and balance, fuel, flight monitoring, separation of flight from severe weather, getting passengers to their destination, interacting with traffic flow managers and assisting the pilot with information and responses to their requests. Dispatcher and pilot communicate by voice via two-way radio and electronic messages via Aircraft Communications Addressing and Reporting System (ACARS), a digital datalink using airband radio or satellite.

The airspace over the U.S. is divided into 24 regions: 20 Air Route Traffic Control Centers (ARTCC a.k.a. center) in the Continental United States (CONUS), Anchorage Center (ZAN) and three Combined Center & Radar Approach Controls (CERAP)—Honolulu CERAP (ZHN), Guam CERAP (ZUA), and San Juan CERAP (ZSU). The centers in the CONUS are shown in Fig. 2. The center airspace is subdivided into sectors (ETMS, 1999; Nolan, 2000). For example, the high-altitude sectors in the

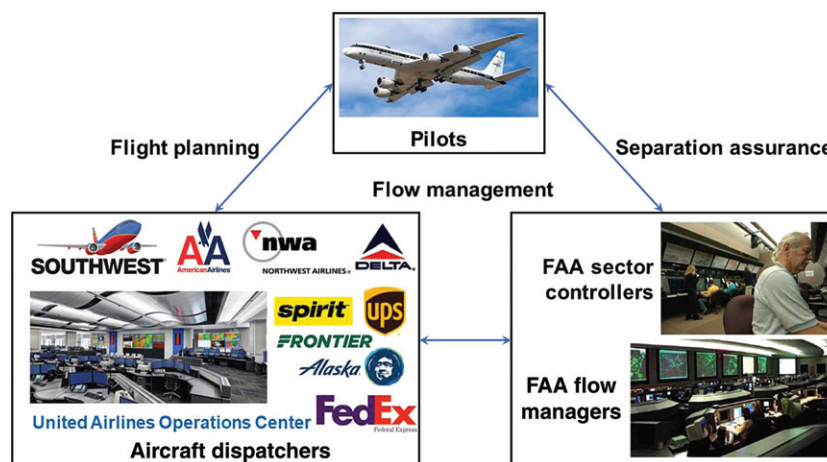


Figure 1 Air traffic management operations.

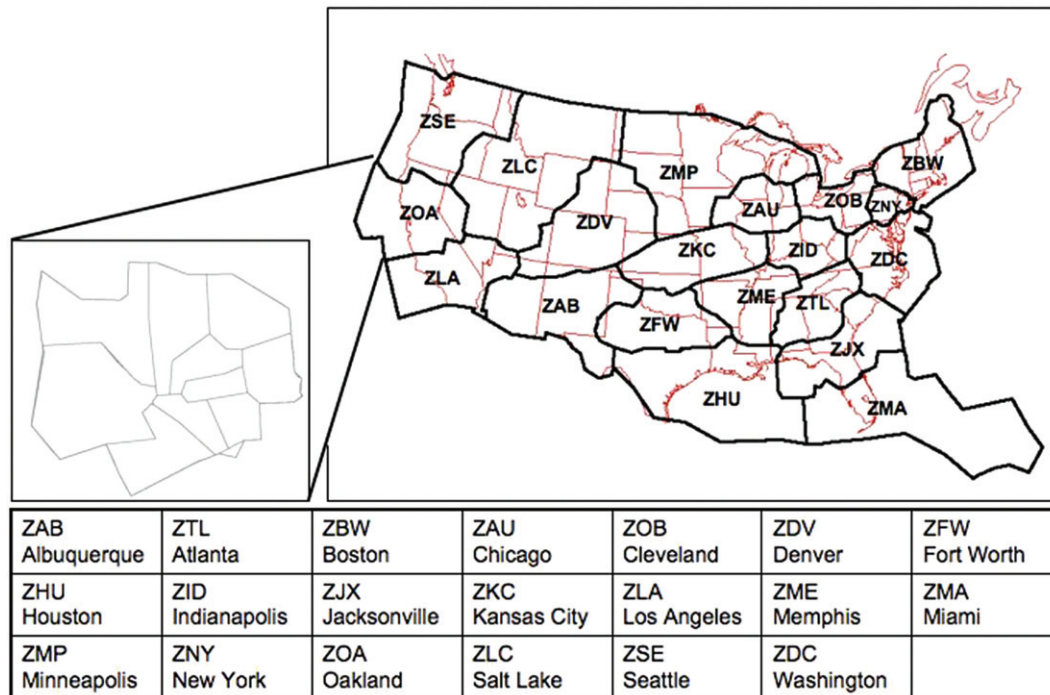


Figure 2 Centers in the Continental US with Sectors shown for Oakland Center (ZOA).

Oakland Center (ZOA) are depicted in the insert appearing on the left side of **Fig. 2**. Air traffic controllers always keep the aircraft in their sectors separated from each other; they handoff aircraft from one sector to another within the center and then to a sector in the neighboring center as flights traverse the center airspace. Traffic flow managers in the center impose traffic flow controls to protect the air traffic controllers in the sectors from being overwhelmed by traffic demand. They coordinate with the neighboring centers and the Air Traffic Control System Command Center (ATCSCC a.k.a. Command Center), located in Warrenton, Virginia, to impose metering restrictions on inbound traffic from neighboring centers. They also interact with sector controllers within their center to ensure that the controller's handoff outbound traffic according to the metering schedules coordinated with the receiving centers. Traffic flow managers at the centers are also responsible for coordinating metering of arrival and departure traffic to and from airports transitioning through the terminal area.

Two major functions associated with Air Traffic Management (ATM) are Air Traffic Control (ATC) for keeping aircraft separated from each other and Traffic Flow Management (TFM) for maintaining efficient flow of traffic. These two functions must be performed while (1) preventing airport and airspace capacity constraints from being exceeded by demand and (2) reducing environmental impact of aircraft emissions. The task of detecting conflicts (potential loss of separation) and resolving them is referred to as "Separation Assurance," which is performed by air traffic controllers with the help of decision support tools. TFM is the planning of initiatives to prevent traffic demand from exceeding airport and airspace capacities and redistribute demand for effective use of the available capacity. Strategic plans for TFM are developed throughout the day by the ATCSCC following a Collaborative Decision Making (CDM) process with centers, terminal facilities, airports and aircraft operators guided by predictions of national traffic flows resulting from forecast demand and weather conditions. Airline schedules, route preferences and operational objectives of the different stakeholders are factored into the CDM process (Ball et al., 2001). Resulting plans might call for delaying some flights at airports and rerouting others. Dispatchers interacting with air traffic coordinators at airlines respond to Traffic Management Initiatives (TMI) by amending flight plans, rescheduling, substituting and canceling flights.

Air Traffic Control

Air traffic controllers have always used cognitive analysis supported by tools for display of track history and trend vectors (short duration forecasts) for determining conflicts between pairs of aircraft and developing strategies for resolving them. Using voice communication, they issue an advisory to the pilot, such as climb, reduce speed, or change heading to resolve conflicts. Modern Decision Support Tools (DSTs) deployed in recent years provide trajectory-based advisory information to assist controllers with conflict detection and resolution (Erzberger, 2004, 2006) and arrival metering. Trajectories are predicted by employing models of aircraft performance and the atmosphere, wind data, flight plan and track data. The flight plan defines the horizontal path from origin to destination, cruise altitude and cruise speed. The aircraft performance model specifies takeoff weight, thrust, drag, fuel burn, climb and descent calibrated airspeed and speed and altitude thresholds for transitioning from one flight mode to the next such as from climb to cruise and approach to landing.

The atmosphere model provides parameters such as density of air, temperature, speed of sound and static pressure as a function of altitude that are used in the aircraft performance model. The airspeed of the aircraft is obtained from the aircraft performance model in the climb and descent phases. It is estimated using groundspeed and course heading, derived from track data, and wind data in the cruise phase of flight. The estimated airspeed and the wind forecast are then used to forecast groundspeed along the planned route of flight. The course heading for the horizontal component of the forecast trajectory is estimated assuming (1) the aircraft will continue flying along the flight plan route starting from its current position or (2) it will continue flying along its current path as indicated by its recent track history. The horizontal component of the trajectory is predicted by propagating the latitude and longitude forward in time using the course heading and groundspeed in the equations of motion. The vertical component of the trajectory is similarly predicted by propagating the altitude forward in time using the climb and descent rates obtained from the aircraft performance model for climb to cruise altitude and for descent past the top of descent point. Climb rate is set to zero past the top of climb point and prior to the top of descent point for predicting the vertical component of the trajectory in the cruise phase of flight.

Conflicts between aircraft are determined by computing the distance between the forecast trajectories at discrete intervals along the trajectories and checking if the minimum separation between them is less than or equal to the minimum separation standard. Once a conflict is determined, a process called trial planning is initiated to resolve the conflict. Trial planning, as the name suggests, uses the trajectory prediction capability repeatedly with different values of control variables—speed, heading, altitude and combinations of them—for creating alternative trajectories. These trajectories are examined for conflicts using the conflict detection procedure. Control variables associated with trajectories that do not result in conflicts are chosen as acceptable conflict resolution maneuvers. These options are then displayed on the DST as choices to the controller for providing advisories to the pilots.

Although these DSTs improve safety and reduce delays, human-centered automation limits the capacity of a sector to about 15 aircraft, a number far below that required by legal separation, because of a human controller's cognitive ability to maintain situational awareness and safety of traffic in the sector (Sridhar et al., 2011). A fundamental transformation in the way separation assurance is provided is necessary to achieve the capacity gains necessary for meeting future growth in demand.

Traffic Flow Management

The complexity of the TFM problem and the duration of the planning interval leads to a natural decomposition of the problem into national TFM decisions performed by the Command Center providing guidance to centers or groups of centers (regions) and regional TFM decisions performed in the regions. Regional TFM plans are detailed plans for the region for a short duration of time. They are implemented using near-term traffic and weather information.

National Traffic Flow Management

The goal of national traffic flow management is to prevent the traffic demand from exceeding the available capacity of NAS resources because exceeding capacity directly affects safety. TFM imposes flow restrictions following the CDM process for accommodating user preferences. TFM is often required during severe weather conditions because weather reduces both airspace and airport capacity. Predictions of weather and traffic demand in the airspace and at airports are employed by the ATCSCC to form strategic plans from an hour to 24-hour time horizon (FAA, 2008). Some of the mechanisms employed for national TFM are Airspace Flow Program (AFP), Ground Stop (GS), Ground Delay Program (GDP), and National Playbook. AFP identifies flights scheduled to travel through a capacity constrained airspace such as a region impacted by severe weather. The impacted flights on the ground can either take delay waiting at the airports or takeoff and take the delay while airborne. Depending on the extent of the constraint, flights might be allowed to travel through the constrained area or be required to circumvent it. While AFP is primarily used for responding to en route constraints (it can be used for constraining flows into and out of airports), GDP and GS are used for responding to airport constraints. GS holds all flights destined to an affected airport at their airport of origin for the duration of the GS initiative. Unlike GS, GDP issues a departure clearance time to aircraft scheduled to arrive at the capacity constrained airport using a Ration-By-Schedule (RBS) procedure. RBS spaces out aircraft arrivals in time to reduce the arrival rate according to the limited airport capacity. The difference between the RBS scheduled arrival time and the airline provided arrival time—delay—is added to the airline provided departure time to determine the departure clearance time. Aircraft are expected to takeoff within 5-minutes early or late (a 10-minute window) with respect to the assigned departure clearance time.

The FAA employs the National Playbook (FAA, 2008), which is a compendium of standardized alternative routes for circumventing defined regions of airspace based on historically validated data, as a part of its Severe Weather Avoidance Plan (SWAP) and for mitigating equipment outage impact. Fig. 3 shows the *CAN RUBKI EAST 1* West to East transcontinental route from the Playbook for routing eastbound traffic through the Minneapolis Center when a large portion of airspace in the Midwest is affected by weather. The large rectangular region in the southern portion of Minneapolis Center (ZMP) represents a predicted severe weather region. The routes represented with a solid line in this figure represent alternative routes for aircraft originating on the West Coast and traveling to East Coast destinations such as Boston (BOS), La Guardia (LGA), and Dulles (IAD).

Regional Traffic Flow Management

Regional traffic flow management, which operates on a 20-min to 2-hour time horizon, adjusts flow control actions based on near-term forecast of air traffic demand, airspace capacity and weather while preserving the national flow management objectives set by the Command Center. These adjustments are made by Re-Routing (RR) flights, distancing flights spatially with Miles-In-Trail (MIT)



Figure 3 Graphical representation of the CAN RUBKI EAST 1 route with a conceptual region of severe weather depicted by the shaded polygon.

and distancing them temporally with Minutes-In-Trail (MINIT). Distancing has the effect of reducing the number of aircraft entering the region; the number entering is inversely proportional to the MIT and MINIT. MITs are set in increments of five miles with a typical value between 10 and 30 miles. MIT and MINIT restrictions are used both in the en route and terminal airspaces. They are routinely used along center boundaries to control (meter) the inbound and outbound flow rates. However, they can be applied at any location where metering is desired including along lines and arcs demarcating airspace. The centers requesting the restriction set the inbound rates while the centers enforcing the restriction constrain the outbound traffic flow rate to comply with the requested restriction. The local MIT and MINIT restrictions are coordinated with the Command Center for a common system-wide awareness of TFM initiatives.

Fig. 4 shows the outcome of the national and local TFM actions on flights on a day affected by severe weather. Each dot in the display indicates a flight in the NAS. Gray, blue, and red dots indicate on-time flights, flights delayed between 15-minutes and 2-hours and flights delayed by more than 2 h, respectively. While delays due to TFM initiatives are unavoidable especially on days affected by severe weather, the efficiency of TFM decisions can be improved by timely sharing of information between decision-makers (Wambsganss, 1997). Airlines being fully aware of the traffic conditions and the status of the NAS in making their rerouting, rescheduling and cancellation decisions and the TFM decision support tools having the data associated with these decisions in a timely manner result in greater accommodation of user preferences and reduction of delays. This is the main motivation for the FAA and the aviation industry for developing the CDM process. The FAA has also invested in providing real-time and near real-time system-wide air traffic data to the stakeholders via the System-Wide Information Management (SWIM) data feed starting with the completion of Segment 1 in 2015—to foster common understanding and to aid decision-making. While the CDM process is effective for strategic decision-making, user participation is somewhat limited for short-term planning because of the time available for decision-making and for implementing tactical decisions, leading to inefficiency.

Other reasons for inefficiency are business and human decision-making related. For example, takeoff weight of the aircraft, which is an important parameter especially for climb trajectory prediction (Krozel et al, 2003), is not provided by the operator for business reasons because knowledge of the takeoff weight along with the type of aircraft and route of flight could be used by a competitor to estimate the load factor—number of passengers carried by the flight. TFM decisions made by human traffic managers, while aided by computational tools, are dictated primarily by experience and safety considerations and not by optimality. It might be possible to



Figure 4 Traffic display with on-time flights in gray, aircraft delayed between 15 min and two hours in blue, and aircraft delayed more than 2 h in red.

better utilize the capacity of the NAS by considering the national and regional initiatives as parts of a single integrated optimization problem. TFM optimization has been formulated as a mixed-integer linear problem and efficient computational techniques have been developed for solving it (Bertsimas and Patterson, 1998; Rios and Lohn, 2009; Sengupta et al., 2005). The main difficulty in using these solutions operationally, however, is that these solutions provide controls such as delay and speed reduction for each aircraft; it is difficult to translate the solution to flow controls. In fact, the optimization problem for transforming the disaggregated controls for individual aircraft to aggregated flow control is of the same size and complexity as the original problem for generating the disaggregated controls (Sun et al., 2010). The scalable computational hardware and software platforms based on cloud technologies promise to make these optimization solutions more practical in the future.

Terminal Area Operations

The average spacing or distance between aircraft for ensuring safety with human-centered control limits capacity at the airports and within en route airspace. Spacing is influenced by the separation standards and the ability of the Air Navigation Service Provider to precisely control to those standards. Both are influenced by the technologies and procedures employed by controllers for tactical (20-minute time horizon) air traffic management and by pilots for complying with the resulting advisories. Improving the efficiency in the terminal area, which is the volume of airspace within a radius of about 50 miles surrounding airports, is especially difficult due to the differences in the operating characteristics compared to that in the en route airspace. Terminal area controllers manage both climbing and descending aircraft operating in the proximity of terrain and close to each other in a high traffic density environment. Due to the small size of the terminal area compared to the high-altitude airspace, both the time and space available for aircraft maneuvers for conflict resolution and for absorbing delays, especially for metering arrivals to airports, is short.

In today's terminal area arrival operations, as an aircraft descends for landing, controllers track and guide the aircraft all the way to the runway using visual and data aids provided by terminal automation systems such as the Standard Terminal Automation Replacement System (STARS) as well as their skills and judgment. They issue turn-by-turn instructions (a process known as vectoring) to the pilot via voice communications. As aircraft fly towards the airports, controllers instruct the pilots to merge aircraft into arrival flows and sequence them for arrival. Busy terminal area conditions often force the aircraft to fly inefficient arrival routes requiring frequent changes in direction, altitude and speed. Controllers are forced to ask aircraft to fly a longer path—path stretching advisory—or remain in a holding patterns to absorb larger amounts of delay. These maneuvers cause additional fuel burn and noise. They also increase controller workload by exacerbating traffic congestion and requiring them to monitor the aircraft during these off-nominal maneuvers and return them back to their intended routes. The uncertainties in the system force controllers to add spatiotemporal buffers for separation, which decreases airspace capacity, leading to further delays.

While more efficient arrival paths can be flown, currently deployed technology and procedures limit their use to light traffic conditions such as during night. As density of traffic increases, maintaining safe separation with a human-centered system will continue to take precedence over efficiency. Making efficient arrival procedures a common practice during heavy traffic, when they are needed most, while ensuring safety remains a technical challenge.

The arrival and departure capacities at the busiest airports will continue to play a key role in determining the efficiency of the NAS and will ultimately limit the growth of air traffic. To accommodate the traffic anticipated by NextGen, the arrival/departure capacity must increase at the large airports and regional airports, and number of operations must increase at the underutilized and smaller airports. The JPDO envisioned new technologies for enabling significant growth at large airports.

Surface Traffic Operations

The surface operation at many major airports in the U.S. is conducted by airlines controlling the ramp area (also known as the "nonmovement area") and by the FAA controllers in the Air Traffic Control Tower (ATCT) controlling traffic on taxiways and runways. Airline ramp controllers coordinate pushback of aircraft from their gate based the scheduled departure time and pilot readiness for departure. The difference of the actual gate pushback time with respect to the scheduled gate pushback time is tracked as a metric of system performance. Ramp control is responsible for the aircraft leaving the gate to stay separated prior to entering the movement areas. Because gate departure time determines the entry time at the spot (the location demarcating the boundary between the ramp area and taxiway), into the taxiways and eventually arrival at the runway, uncoordinated gate release times especially during busy conditions result in taxiway congestion and large runway entry queues (Balakrishnan and Jung, 2007; Malik et al., 2010; Rathinam et al., 2008).

Once the ground controller gives permission to the pilot to enter the movement area, the aircraft leaves the spot and enters the taxiway. The ground controller is responsible for separation of aircraft and vehicles on airport movement areas except on the active runways. The ground controller issues instructions to the pilots to safely taxi to the runway. Pilots contact the local controller prior to crossing an active runway and for permission to enter the designated runway for takeoff. The local controller sequences this flight into the local flow while ensuring that the aircraft is separated from the other inbound and outbound aircraft. The local controller clears the flight for takeoff and instructs the pilots to contact the departure controller at the Terminal Radar Approach Control (TRACON).

Like the gate pushback delay, taxi-out time delay is also a metric of airport performance. Taxi-out time is defined as the difference between when the aircraft leaves the gate/spot and its takeoff time. The excess taxi-out time with respect to the average of unimpeded taxi-out time for the airport is defined as the taxi-out delay. A study in 2007 estimated the taxi-out delay to be 6.8 min per departure

considering all U.S. airports and higher than 15 min per flight for the three major airports in New York (Guilding et al., 2009). During January 2016, the FAA Aviation System Performance Metrics database reported taxi times that, in Los Angeles airport (LAX), 1% of the flights experienced more than 40 min of taxi times and the average taxi times was 15.73 min per flight. An analysis of fuel consumption at the Dallas/Fort Worth (DFW) airport showed approximately 7000 gallons/day is spent because of stop-and-go during taxi.

The efficiency of NAS can be improved by better coordination of arrival, departure and surface traffic. The coordination is complicated by the fact that all airports are unique, and their distribution of responsibility amongst airlines, airport authorities and the FAA, makes the coordination more challenging, and difficult to apply a monolithic solution. Pre-departure coordination between several operational positions within the ATCT and Traffic Management Units (TMUs) at both the associated TRACON and the ARTCC serving the departure airport is currently done by phone calls, manual data entry and the use of restrictions such as a Miles-In-Trail (MIT) or Minutes-In-Trail (MINIT) for flights crossing a departure fix (Doble et al., 2009). The extent of coordination needed depends on the complexity and activity level of the departure airport, number of other airports within the terminal area, merging of departures with overhead traffic flows, the presence of convective weather in the terminal area or en route and constraints at the destination airports.

The FAA has developed a suite of capabilities to improve departure congestion management. One of them is the TFM Surface Data Integration (TSDI), which processes airport data from various automation and surveillance systems (e.g., Airport Surface Detection Equipment (ASDE), Electronic Flight Strip Transfer System, Automatic Dependent Surveillance-Broadcast (ADS-B) and airline-provided event data) to produce information on actual surface events, predicted events and metrics. This standardized information is intended for integration into TFM decision support tools.

Advanced Airspace Operations Concepts

The ATM system supports operations during all phases of flight starting from gate pushback at the departure airport, taxi, takeoff, climb, cruise, descent, approach, landing and taxi to the gate at the arrival airport. This is accomplished by using surveillance, separation standards and decision support systems tailored for airport surface movement, flight in the terminal areas and in en route airspace. The ATM system has embraced digital transformation for quite some time upgrading and modernizing hardware and software-based systems for managing air traffic. The transition from a largely manual system to a computer-based system has been an ongoing endeavor. With increasing automation and especially that required for accommodating highly-automated pilotless aircraft operations in the NAS, major challenges in the future will be: (1) the appropriate human/automation mix, (2) graceful degradation with transfer to human control when automation fails, and (3) reliance on aircraft operators and their service providers with roles and responsibilities for managing the safety of air traffic operations. The expected evolution from a centralized—single provider of ATM service—to a decentralized system with multiple providers is shown in Fig. 5. Successful transition will depend on meeting the human-automation challenges.

The near-term focus of the FAA will stay on the NextGen modernization program, which is responsible for implementing major new technologies and capabilities with the objectives of improving safety, efficiency and predictability. The multiyear NextGen investment and implementation plan is almost halfway done; it calls for continued deployment of cutting-edge technologies, procedures, and policies through 2025 and beyond. Several NextGen technologies are described below in Section IV.

The future ATM system will need to accommodate operations with new types of vehicles such as very light jets, unmanned aircraft systems (UAS) and commercial space launch vehicles. The developments in Urban Air Mobility (UAM)—on demand flights with two to six passenger aircraft—and Unmanned Aircraft System (UAS)—package delivery drones—will introduce a large amount of low altitude air traffic into the NAS. NASA and the FAA have been developing operational concepts for the safe and efficient integration of the operations of these new classes of vehicles. The architecture developed under NASA's UAS Traffic Management (UTM) project with air traffic services provided by multiple organizations is considered a model for developing UAM air traffic

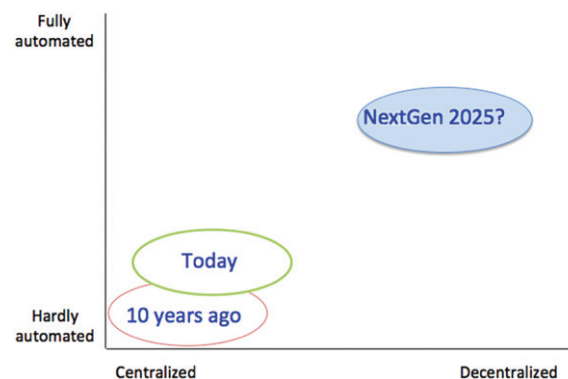


Figure 5 Transformation of ATM systems.

services. The FAA has recently released the concept-of-operations for UAM flight operations (ConOps 1.0) (FAANGO, 2020). UAM is a subset of the Advanced Air Mobility (AAM) initiative of NASA, FAA and industry for developing the future air transportation system for serving local, regional, intraregional and urban areas with new electric vertical takeoff and landing aircraft. The envisioned end state of UAM operations depends heavily on autonomy. Initial operational experience will be gained with new vehicles certified under the current regulatory frameworks, which are expected to evolve as UAM operations increase. Eventually, routine operations are expected to introduce new rules and regulations, infrastructure and automated vehicles, operational control and traffic management systems. Several emerging technologies are summarized in Section V.

In addition to enabling operations with new types of vehicles, enhancements will continue for improving safety and efficiency with growing demand for commercial high-altitude traffic in the future. Better sharing of data and decisions across different systems employed for airport surface, terminal and en route airspace traffic flow management and air traffic control offer the potential for eliminating unnecessary restrictions and uncoordinated decisions, thereby, reducing delays and fuel consumption and increasing utilization of capacity constrained airport and airspace resources. Several of these technologies in various stages of development and evaluation are outlined in this section.

Integrated Demand Management

Integrated Demand Management (IDM) (Smith et al., 2016) is a near to midterm Trajectory Based Operation (TBO) concept to collaboratively organize aircraft trajectories into well-managed flows that match traffic demand to available capacity. The concept seeks to leverage FAA and NASA pre-departure, en route and arrival technologies. IDM uses Traffic Flow Management System (TFMS) tools to pre-condition traffic into the Time-Based Flow Management (TBFM) system, enabling TBFM to better manage delivery to the capacity-constrained destination. TFMS is used operationally by the FAA to (1) forecast air traffic demand in the airspace sectors and at airports based on airborne flights and flights scheduled to depart in the future and (2) to develop traffic flow management initiatives such as altering departure time, increasing the temporal or spatial separation and increasing the length of the flight route to balance the demand with respect to the available capacity. As opposed to TFMS, which is employed for longer-term planning, TBFM is used by the FAA for short-term arrival and departure traffic flow management. TBFM assigns arrival slots at metering locations such as meter fixes, meter arcs and runway threshold based on estimated time of arrival of flights, separation rules and the arrival rate set by the airport. TBFM also enables air traffic controllers to reschedule internal departures—from airports located within some distance from major airport—to fit into the overhead stream of arrivals to the major airport based on their calculated Scheduled Times of Arrivals at the metered locations.

The IDM approach is expected to benefit airlines by reducing fuel consumption because of holding aircraft on the ground as opposed to in the air. One component of IDM that is currently being investigated is the use of Flow Constraint Areas (FCA) at the TBFM boundary to which Required-Times-of-Arrival (RTA) are generated (Yoo et al., 2016). These FCA rates are set based on downstream airspace and airport capacity constraints. IDM accommodates user preferences using the Collaborative Trajectory Options Program (CTOP) capability of TFMS. CTOP is a procedure for managing traffic demand using FCAs while accounting for user route and delay preferences specified via Trajectory Options Sets (TOS). As opposed to a single flight plan request, TOS enables the flight operator to request multiple flight plans with different routes, altitudes and speeds. Options within the TOS might contain “start” and “end” times for operational acceptability of each flight plan option. This enables CTOP to rank the TOS options based on the extent of delay required for mitigating demand-capacity imbalance. CTOP then assigns either a route to go around the FCA or a route and departure time—Expect Departure Clearance Time (EDCT)—to arrive at the scheduled slot at the FCA. The operator is expected to file a flight plan with the assigned option. Aircraft assigned an EDCT in CTOP are exempt from additional delays by policy except as required by miles-in-trail and departure and en route spacing approved by the ATCSCC.

Integrated Arrival/Departure/Surface Metroplex Traffic Management

Some of the inefficiency in today’s system stems from impediments in information sharing for effectively managing traffic in busy terminal areas. While technologies for improving the arrival, departure and airport surface traffic operations have been under development for quite some time, NASA’s recent Airspace Technology Demonstration-2 (ATD-2) initiative seeks an integrated approach with information sharing across airlines, air traffic service providers, airport authorities and technology vendors to improve predictability of aircraft movement times on the airport surface and in the airspace to reduce fuel burn and emissions. Several of the inefficiencies targeted by ATD-2 include untimely pushback from the gate and uncertainties in prediction of taxi, takeoff, and climb times. Uncoordinated pushback from the gate can cause congestion on the taxiways and long runway queues resulting in delays. Inability to predict taxi, takeoff and climb times reasonably accurately leads to inaccurate prediction of time of arrival at metered locations, and therefore, to inaccurate traffic demand at constrained resources. This often results in overly conservative traffic flow management initiatives.

ATD-2 integrates arrival, departure and surface (IADS) concepts for traffic management in a metroplex environment (Jung et al., 2019). The technologies will increase predictability of the air traffic system and enhance operational efficiency, while maintaining or improving throughput. IADS combines multiple concepts and technologies, including the FAA’s three operational

decision support tools (TFMS, TBFM, and Terminal Flight Data Management (TFDM)), as well as the NASA's Spot and Runway Departure Advisor (SARDA) and Precision Departure Release Capability (PDRC) technologies. SARDA was developed to improve the efficiency of airport surface operations through time-based metering of aircraft and improved situational awareness amongst stakeholders, while PDRC coupled a trajectory-based decision support functionality with the tactical departure scheduling capabilities of TBFM, resulting in more precise scheduling of surface departures into constrained overhead flows. Based on ready times of aircraft at the gate, the Surface Trajectory-Based Operations (STBO) tool, which incorporates both SARDA and PDRC, schedules pushback times for departures for communication to a tool for the airline operators in the form of a Ramp Traffic Console (RTC). Outbound scheduling of departures under TBFM scheduling to another airport (typically internal departures) is coordinated with the TBFM system scheduling those departures.

Applied Traffic Flow Management (ATD-3)

The technical challenge for NASA's ATD-3 subproject (Gong, 2020) was to "reduce weather-induced delays through integration of weather information to better manage aircraft, traffic flow, airspace and schedule constraints by delivering air-ground procedures and user-tool technologies." The key technologies for achieving the ATD-3 technical challenges were: Multi-Flight Common Route (MFCR), Traffic Aware Strategic Aircrew Requests (TASAR) and Dynamic Routes for Arrival in Weather (DRAW). MFCR is concerned with automatically searching for efficient weather avoidance routes (in terms of fuel and time) and unnecessary weather avoidance routes imposed by earlier traffic management initiatives that are no longer needed due to changes in weather conditions. Avoiding unnecessary reroutes reduces delay. TASARs focus is flight-deck-based automated continuous search for efficient reroutes in the airborne phase for reduced fuel consumption and flight time while avoiding traffic, weather and restricted airspace. DRAW seeks to find efficient en route weather avoidance routes for arrival flights while complying with meter-fix capacity constraints. The DRAW algorithm computes weather avoidance routes for flights that are predicted to be in conflict with weather and provides these routes to a scheduler. The timeline generated by the scheduler shows the estimated arrival times of the rerouted flights in relation to the other flights and the available slots. The trial planning process consists of iteratively creating routes and determining the schedule impact until the desired route and schedule are found to be acceptable to traffic flow managers. An application of DRAW is balancing the demand between the arrival meter-fixes. The ATD-3 Integrated Concept utilizes MFCR, TASAR, and DRAW to create a departure to arrival route solution.

Environmentally Friendly Operations

Aviation affects the environment in terms of noise, air quality, water quality and climate (IPCC, 1999). It is estimated that aviation is responsible for 13% of transportation-related fossil fuel consumption and two percent of all anthropogenic Carbon Dioxide (CO₂) emissions (Brasseur and Gupta, 2010; Lee et al., 2009). Although this contribution is small, a large portion of the emissions takes place at higher altitudes where they remain longer compared to that emitted at the surface. The desire for growth in air traffic while limiting the impact of aviation on the environment has led to research in green aviation. The goals of green aviation are better scientific understanding, utilization of alternative fuels, introduction of new aircraft technologies—engine and airframe—and operational changes. Aviation operations affect the climate by injecting the by-products of the combustion of fossil fuel—Carbon Monoxide (CO), CO₂, water vapor, oxides of nitrogen (NOX) and particulate matter—into the atmosphere. Among the constituents of engine exhaust, CO₂ and water vapor are Green House Gases (GHG) that perturb the balance between incoming solar radiation and outgoing infrared radiation at the top of the troposphere. This perturbation is called Radiative Forcing (RF); GHGs result in positive RF. Higher concentration of GHGs causes higher amount of infrared radiation to be trapped in the atmosphere. Global mean surface temperature change varies approximately linearly with RF. Because of its abundance and long lifetime, CO₂ has a long-term effect on climate change compared to the non-CO₂ emissions. Water vapor and the particulate matter in the exhaust cause the formation of condensation trails (contrails) (Duda et al, 2003), which can then lead to the formation of cirrus clouds. Contrails and cirrus clouds reflect the infrared radiation received from the Earth's surface back to the surface. Some estimates suggest that contrails might be causing more climate warming today than all the residual CO₂ emitted by aircraft (Boucher, 2011). Methods have been proposed for eliminating contrail formation such as flying at a different cruise altitude and rerouting around regions of airspace with temperature and humidity conditions favorable for persistent contrail formation (Williams et al., 2002). However, these solutions however come at the cost of longer flight time and cruise at nonoptimal altitude, resulting in increased fuel consumption and therefore CO₂ emissions.

To estimate the impact of aviation contrails in the U.S., simulation-based studies have been conducted. For example, Ref. 34 employed the process summarized in Fig. 6. This process consists of (1) simulating air traffic, (2) predicting contrails, (3) predicting emissions, and (4) generating alternative trajectories to reduce the environmental impact. Air traffic is generated by simulating flight trajectories along the path prescribed in the flight plans using aircraft performance models, derived from Eurocontrol's Base of Aircraft Data (BADA), and wind data. These trajectories and predicted temperature and humidity data are used to predict the contrails. The fuel burn and emissions model are then used to generate estimates for the simulated trajectories. Finally, alternative trajectories are created to circumvent regions that are conducive to contrail formation. The procedure in Fig. 6 can be applied to present and future traffic scenarios. It can also be used to evaluate the energy efficiency of contrail reduction strategies and tradeoff between persistent contrail mitigation and fuel consumption for policy makers to set aviation operation guidelines.

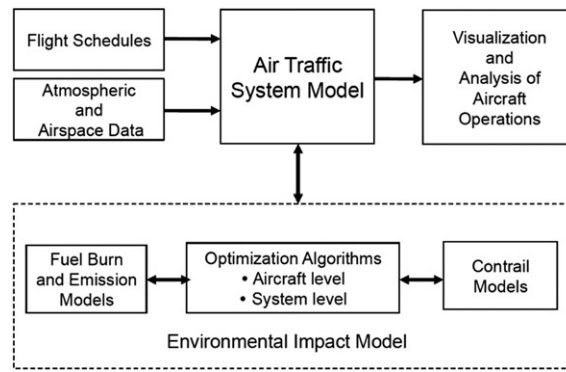


Figure 6 Simulation capability to evaluate impact of aviation emissions and contrails.

Performance Based Navigation (PBN) enabled by satellite and other communication technologies being introduced in the NAS frees aircraft from ground-based navigation, which makes it possible to fly fuel-efficient and environmentally friendly routes adjusting dynamically with changing atmospheric conditions. Part of the motivation for introducing new technologies and procedures is a carbon neutral growth in the fzaand reduction of climate impact of aviation emissions in the long-term (2050).

In addition to emissions and impact of contrails, noise reduction is an important environmental goal. Noise from aircraft has direct health, social and economic impacts to local residents and communities near the airport vicinity. It is one of the leading aviation-induced environmental issues that have been studied for decades. In U.S., about 500,000 people today are affected by aircraft noise based on the FAA's Office of Environment and Energy's standard, i.e. people that live in the area where the Average Day-and-Night Sound Exposure Level over 24 h (DNL) is 65dB or more. 65 dB DNL is recommended as threshold of compatibility with residential land use. However, experts have suggested a lower level of 55 dB DNL as "level of environmental noise requisite to protect the public health and welfare with an adequate margin of safety" (1972 Noise Control Act) (SAE, 1986). The anticipated growth in air traffic will result in an increased number of people affected by aircraft noise. There are mainly two ways to reduce aircraft noise impact on the environment. One is through developing noise-reduction aircraft engine technologies (Hileman et al., 2007). The other is by redesigning currently operated terminal routes to mitigate the aircraft noise impact while ensuring safety and throughput. To reduce the impact of noise near airports, arrival and departure routes and procedures are prescribed to ensure that the flights taking-off and landing leave a small noise footprint on the population below (Alam et al., 2011; Xue and Athins, 2006).

NextGen Implementation

The NextGen portfolio of projects seeks to plan and implement improvements to infrastructure, new innovative technologies and procedures to modernize the air transportation system. This section describes some of the mature concepts and technologies that are being considered for implementation (FAA, 2019).

Automatic Dependent Surveillance-Broadcast (ADS-B)

ADS-B (FAA, 2013) is expected to transform air traffic control by providing surveillance information directly from the aircraft as opposed to indirectly from the current radar-based systems, providing surveillance information in oceanic airspace and in remote locations lacking radar coverage such as the Gulf of Mexico and parts of Alaska, providing traffic data for display on the onboard Cockpit Display of Traffic Information (CDTI) system for improving pilot situational awareness and by promoting better controller and pilot decisions with common situational awareness. ADS-B capable aircraft derive their position from the onboard systems such as the Global Positioning System (GPS) and broadcast their position along with state variables such as speed, heading and altitude, and short-term intent information—consisting of locations of the next several waypoints describing the planned route—using satellite communication. The broadcast information can be received by other ADS-B capable aircraft and ATC in real time. ADS-B consists of two different services: ADS-B Out and ADS-B In. ADS-B Out is the capability for periodically broadcasting information and ADS-B In is the capability for receiving traffic and weather information data from nearby-aircraft and ground-based systems. Vehicle position data provided by ADS-B, and displayed on cockpit and controller displays, has been found to reduce the risk of runway incursions and collisions especially during nighttime and low visibility conditions on the airport surface. ADS-B is also expected to help reduce separation between aircraft, which will lead to increasing airspace capacity.

The FAA has integrated ADS-B into automation platforms at all 24 en route air traffic control facilities and into most of the TRACON facilities, including the 30 busiest terminal areas in the U.S. They plan to integrate into the remaining terminal area automation platforms as the platforms are updated. The FAA has engaged with several airlines to obtain ADS-B data to validate the business case for early adoption. The FAA mandated aircraft need to be equipped with ADS-B Out capability to operate in Class A, B, C and E airspace after January 1, 2020. Currently, there is no mandate for aircraft to be equipped with ADS-B In capability.

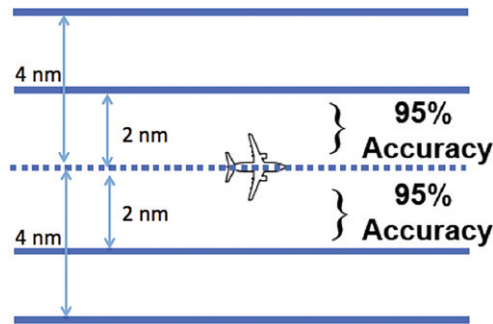


Figure 7 Aircraft with RNAV capability of RNP 2.

Table 1 PBN requirements during different phases of flight

Oceanic	RNP 10, RNP 4
U.S. En route	RNP 2
Terminal Area	RNP 2, RNP 1
Approach in Instrument Meteorological Condition (IMC)	RNP 0.3

Performance-Based Navigation (PBN)

Area Navigation (RNAV) is a method of navigation that enables an aircraft to fly along a desired flight path within the coverage of the navigational aids, within the limits of the aircraft or a combination of both. RNAV capable aircraft can fly directly between points in the airspace rather than from navigation aid to navigation aid. Safety along an RNAV route is ensured by a combination of aircraft navigation accuracy, route separation, and ATC radar monitoring and communications. Required Navigation Performance (RNP) is a navigation accuracy requirement that RNAV capable aircraft—equipped with onboard avionics—can monitor and comply with during flight. The Performance Based Navigation (PBN) concept assumes the navigation accuracy specification is met using a combination of ground-based, satellite-based and aircraft-based hardware and software. RNP specifies the cross-track accuracy between the desired and actual trajectory of the aircraft. As shown in [Fig. 7](#), an aircraft with RNAV capability of RNP 2 should be able to follow the desired trajectory with a cross-track accuracy of two nautical-miles 95% of the time and within a lateral containment region of four nautical-miles 99.999% of the time. RNP values ranging from 10 to 0.1 nautical miles are typically used. Lower values such as RNP 0.1 are specified for Authorization Required (AR) approaches for instrument landing. [Table 1](#) shows some of the commonly used RNP during different phases of flight.

RNAV routes not based on VOR routes in low altitudes are preceded with the letter “T”; high altitude routes are designated with the letter “Q”. In addition to the PBN requirements listed in [Table 1](#), aircraft on Q-routes on flight levels (FL—altitude in 100s of feet) between 180 (18,000 feet altitude) to 450 and T-routes below FL 180 have a requirement of RNP 2. In the en route airspace, PBN provides more efficient routes around convective weather, increased airspace capacity using multiple Q-routes, direct routes between city/terminal area pairs and reduction in delays. T-routes provide benefits by enabling flight through or around Class B and C airspace.

The benefits of increased capacity and reduction in noise and emissions in the terminal area are accrued by using integrated Standard Terminal Arrival Route (STAR) and RNP approaches and optimized profiles and by decoupling flows between primary and satellite airports.

The realization of the benefits of PBN depends partly on aircraft equipage ([Devlin et al., 2009](#)). PBN requirements can be realized using navigational systems such as RNAV capable Flight Management System (FMS), Non-RNAV system capable of supporting RNP and non-FMS RNAV navigation capability and sensors such as Distance Measuring Equipment (DME), Inertial Reference Unit (IRU) and GPS. The RNP Special Aircraft and Air Crew Authorization Required (SAAAR) requires dual FMS, dual GPS and dual IRU systems. PBN is increasingly possible with more aircraft equipped with FMS. The number of aircraft with FMS increased from 79% to 90% during the years 2004–09. The FAA estimated 96% of the 6653 aircraft conducting Part 121 operations in July 2011 were capable of RNAV operations with RNP 1 ([Devlin et al., 2009](#)).

Weather Integration

Severe weather has been identified as the cause of 67% of the traffic delays in the United States ([TEICAUE, 2011](#)). Weather factors such as wind, icing and visibility reduce the arrival and departure capacity at the airports. Convective weather and turbulence reduce the number of aircraft that can travel through a given airspace. Weather information is also important for imposing airspace and airport constraints ([Krozel et al., 2011](#)) in advanced planning algorithms for TFM. The NextGen Network Enabled Weather (NNEW)

will provide a single authoritative source of weather information to all the users of the NAS for enabling dynamic and collaborative planning especially during severe weather conditions by interfacing with the 4-Dimensional Weather Data Cube. The National Weather Service is primarily responsible for populating the Weather Data Cube. The Weather Data Cube will link FAA, National Oceanic and Atmospheric Administration (NOAA), Department of Defense (DOD) and commercial weather data provider databases. The system will provide the ability to convert the data between standard formats, units and coordinate systems. Finally, it will provide interfaces for retrieval of large volumes of weather data such as in a region and along the planned flight trajectory.

Data Communications

Currently, voice is the primary communications method for exchanging critical ATM information between the pilots and ground-based controllers. Communication via voice while efficient is error prone. To guard against errors, ATC communication requires read-back from the receiver, which can be time consuming especially if the read-back is incorrect. Voice communications for supporting the expected future growth of air traffic with new entrants such as Urban Air Mobility (UAM) is also not scalable because of the radio-frequency spectrum limitations. Digital data communications are less prone to errors and systems can be built with error checking and correction mechanisms. Digital data communication has the potential of improving system safety by relieving both pilots and controllers from routine tasks and enabling them to focus on strategic tasks. Increased use of digital communication in dense traffic environment will relieve radio-frequency congestion, thus freeing up voice communication for critical communications.

Data communications were first introduced in operations as a part of the Future Air Navigation System (FANS) program. FANS-1 and FANS-A developed by Boeing and Airbus, respectively, provide the ability to autonomously send some data from the aircraft to the air traffic control system using Automatic Dependent Surveillance-Contract (ADS-C) capability. The introduction of these capabilities in the oceanic airspace enabled separation distance between aircraft to be reduced from 100 to 50 nautical miles. A modified version of FANS—FANS-1/A+—that uses VHF Digital Link (VDL) mode-2 is designed for use in the higher traffic density domestic airspace. A new data communication standard harmonizing the global needs of civil aviation, Aeronautical Telecommunication Network (ATN) Baseline-2, is currently under development by ICAO (RTCA, 2013).

Operations

Some concepts and technologies for improving the efficiency in the descent, en route and surface phases of flight are discussed below.

Optimal Descent Trajectories (ODT)

ODT tools provide decision support data to air traffic controllers (Green and Vivona, 1996) for enabling continuous descents at near-idle thrust, ensuring conformance to arrival schedule constraints for maximizing throughput, avoiding traffic and airspace constraints along the arrival path and providing clearance delivery—by voice or data link—guided by knowledge of flight deck capabilities for precision guidance and control available onboard the aircraft. Different implementations of the ODT tools, Efficient Descent Advisor (EDA) (Coppenbarger et al., 2009), Continuous Descent Approach (Clarke, 2006; Ren and Clarke, 2007), and Optimal Profile Descent (Shrestha et al., 2000), have been tested in field-evaluations. These tests have demonstrated that ODT technology can save fuel and reduce emissions while avoiding conflicts with other aircraft in busy terminal areas (Shrestha et al., 2000).

The feasibility of use of three-dimensional trajectory clearances issued over data link for automated guidance and control using onboard flight management system was evaluated in Tailored Arrivals (TA) to San Francisco operational trials. Tests of this NASA developed technology were conducted by a partnership between FAA, Boeing and United Airlines during 2006 and 2007. Boeing estimated use of the TA process could save between 400 to 800 pounds of fuel per arrival. In 2012, NASA transferred the results of the research for defining and validating the EDA concept to the FAA for further evaluation, development and operational use as a part of NextGen technologies.

Wind-Optimal and User-Preferred Routes

Airlines operations are focused on reducing the cost during cruise—considering cost of fuel and crew time—because most of the time is spent and fuel is consumed in the cruise phase. In the current system, aircraft cruise at a fixed altitude and airspeed along the planned horizontal route from origin to destination. The optimal route, cruise altitude and cruise speed are amended to comply with terminal area and en route restrictions due to airport and airspace capacity constraints and for circumventing regions of severe weather. Flying this non-optimal trajectory results in higher fuel consumption and greater emissions. A study using air traffic data covering flights to/from the top 34 airports in the continental United States during 2007 estimated the routes used by aircraft were 2.9% longer than the direct routes between the city-pairs. The corresponding distance for traffic between the top 34 city-pairs in Europe was 4% longer (Ren and Clarke, 2007). The extra distance traveled over great-circle routes is significantly greater over US-Europe oceanic airspace because of reliance on procedural separation due to the lack of radar surveillance and VHF radio communication coverage. This could change with more aircraft equipped with the ADS-B system. Flights from Europe to Asia need to fly longer routes due to large amount of restricted airspace, strict entry points and steep terrain. Several studies have determined the limitations of the current routing structure and estimated the benefits of using wind-optimal routes (Hewell et al., 2003;

Reynolds, 2008; Kettunen 2005). Even though wind optimal routes are longer compared to direct routes, the direct operating cost is reduced by taking advantage of the winds.

Surface

Airport Surface Detection Equipment–Model X (ASDE-X) deployed at 35 major U.S. airports provides position information for tracking the movement of aircraft and vehicles on the airport surface. ASDE-X acquires data using radar, multilateration and ADS-B. The acquired data are provided to systems used by ramp operators, air traffic controllers, traffic managers, flight operators and aviation system managers for managing surface traffic movement—scheduling gate departure, predicting conflicts and runway entry times, preventing runway incursions, resolving conflicts, determining taxi route and schedule compliance—and evaluating airport performance—delay and throughput metrics and causal factors. Data sharing promotes common situational awareness and collaborative decision-making for enhancing safety and traffic flow on runways, taxiways and in ramp areas. These data are also important for airport incident and accident investigations. In addition to ASDE-X at 35 airports, the FAA plans call for deployment of Airport Surface Surveillance Capability (ASSC), a multilateration and ADS-B based system, at additional nine busy airports—Portland International (PDX), Ted Stevens Anchorage International (ANC), Kansas City International (MCI), Louis Armstrong New Orleans International (MSY), Pittsburgh International (PIT), San Francisco International (SFO), Cincinnati/Northern Kentucky International (CVG), Cleveland Hopkins International (CLE) and Andrews Field (ADW). ASSC tracks transponder-equipped aircraft and ADS-B-equipped ground vehicles on the surface and aircraft flying within five nautical miles (nm) of airports. Surface data acquired by these systems are also provided in the SWIM feed to the different stakeholders. SWIM is the NextGen's digital data-sharing backbone. SWIM Terminal Data Distribution System (STDDS) processes the raw surface and terminal surveillance data and sends surface information from airport towers to the corresponding TRACON facility through NAS Enterprise Messaging Service (NEMS) for internal and external NAS users (RTCA, 2013).

Integrated Technology

NASA and FAA are collaborating to develop and demonstrate an integrated set of NextGen technologies (Jung et al., 2013) that provide an efficient arrival solution for managing aircraft beginning from just prior to the top-of-descent and continuing down to the runway. These technologies are: ADS-B, RNAV arrival routes, Optimized Profile Descent (OPD) Procedures, Terminal Metering, Flight Deck Interval Management (FIM) and Controller Managed Spacing (CMS) tools. NASA and the FAA have demonstrated the feasibility of sustained high throughput of fuel-efficient arrival operations using precision time-based scheduling which provides runway arrival times and fix crossing times for arriving aircraft. Aircraft are delivered to meter fixes according to schedule, but with small spacing errors that need to be reduced to maximize throughput and avoid spacing violations. Most flight crews use FMS to fly OPDs along RNAV/RNP routes—largely without controller intervention. Terminal controllers correct residual spacing errors and cope with disturbances and off-nominal events using tools and display enhancements based on 4D trajectories. The integration of these technologies and the associated tools will enable aircraft to fly closer together on more fuel-efficient routes, thereby providing even greater benefits in terms of increased capacity and reduced delay, fuel burn, noise and greenhouse gas emissions.

Global Harmonization

The FAA works with other international Air Navigation Service Providers (ANSP) to ensure the interoperability of technologies developed in the U.S. and in other countries under the auspices of the ICAO to enable equipped aircraft to operate globally. Collaboration with the Single European Sky Air Traffic Management Research (SESAR) continues to play an important role in coordinating research and technology development in the U.S. and Europe.

The Asia and Pacific Initiative to Reduce Emissions (ASPIRE) (Aspire, 2000) is a collaboration between the FAA and ANSPs in Australia, New Zealand, Singapore, Japan and Thailand. ASPIRE conducted a series of flights in 2008–11 to successfully demonstrate the potential for fuel and emissions savings in the region. These flights employed modified gate-to-gate operations, including reduced separation, more efficient flight profiles and tailored arrivals. The best practices from these tests can be applied to flights—with properly equipped aircraft—between some cities in the United States and in the Asia Pacific region.

During 2011, the FAA, the European Commission, several European ANSPs and 40 European airlines participated in the demonstration of NextGen and SESAR capabilities on transatlantic flights. As a participant of the Atlantic Interoperability Initiative to Reduce Emissions (AIRE) (Aspire, 2000; AIRE, 2007), Air France flew several flights between John F. Kennedy International (JFK) and Paris Charles De Gaulle (CDG) as well as between Paris Orly (ORY) and Guadeloupe's Pointe-a-Pitre Airport (PTP) during 2010–11. The flights procedures were designed to reduce environmental impact without special equipment. Various improvements to taxiing, real-time lateral and vertical optimization and fuel saving RNAV approaches produced fuel savings ranging from 200 to 300 gallons per flight. Global harmonization must allow for the evolution of aviation systems around the world in a way that allows ANSPs and Airspace Users to invest in advanced capabilities.

The harmonization work should identify key components that will guide prototype testing, functionality, and prioritizing implementation efforts to solve the roadblocks to global interoperability. ICAO has defined standards for global interoperability. Aviation System Block Upgrade capabilities (Lutte, 2015) facilitate strategic planning and investment decisions with a goal of global aviation system interoperability. Initiatives around the world, like NextGen and SESAR, are addressing their own local or regional issues with initiatives to work together.

Newly Emerging Technologies

Unmanned Aerial System Traffic Management

Many beneficial civilian applications of small UAS weighing less than 20 kg have been proposed. Applications include package delivery, infrastructure surveillance, monitoring for search and rescue and agricultural inspection. An established infrastructure for enabling and safely managing widespread UAS operations in low-altitude airspace doesn't exist at present. A UTM system that is being envisioned now is expected to make extensive use of automation for reducing the cost of such operations. The vehicles are also expected to be largely autonomous. It will be difficult for human operators to monitor every vehicle continuously, therefore, automation will be needed for ensuring that these vehicles stay separated from each other and from commercial and general aviation aircraft. The automated system will also need to provide information to humans for making strategic decisions for fleet management and flow management driven by demand, capacity and wind and weather conditions.

NASA in collaboration with the FAA, industry and academic partners has conducted a series of increasingly complex flight tests to develop requirements for the UTM system. The first field test conducted in August 2015 was focused on UAS operations for agriculture, firefighting and infrastructure monitoring. It enabled UAS operators to file flight plans reserving airspace for the duration of each flight. This spatiotemporal reservation was employed for separation assurance by preventing other flights from entering the same airspace segment at the same time. The second field test in October 2016 demonstrated beyond visual line-of-sight flight in sparsely populated areas. It examined technologies for dynamic adjustments based on availability of airspace and contingency management. The third field test in May 2018 examined capabilities for tracking cooperative and uncooperative UAS flights over moderately populated areas. This ability to track is essential for keeping the manned aircraft separated from the unmanned aircraft and vice-versa. The final—fourth—field test during May–August 2019 focused on UAS operations in higher-density urban areas for tasks such as news gathering and package delivery. These activities also tested technologies that could be used to manage large-scale contingencies. NASA's research on UTM has established the innovative concept that various navigational and planning functions will be provided as services by distributed entities. NASA provided results from these tests in the form of airspace integration requirements to the FAA in 2019 for maturing UTM.

Urban Air Mobility (UAM)

Road transportation was the mode of transportation for short distances in the 20th century and air transportation for distances greater than 500 miles (Schafer, 2000). The development of Short/Vertical Takeoff and Landing (S/VTOL) aircraft in the 1970s was unsuccessful in gaining broad acceptance due to noise and ride comfort issues (Fry and Zabinsky, 1967). Helicopters remained special purpose vehicles for use in emergency situations, firefighting and applications where fixed wing aircraft could not be operated due to terrain and space limitations. This model is changing with advances in light-weight composite materials, electric propulsion (including distributed electric and hybrid—gas-electric) and battery technology. These technologies have enabled construction of commercial Electric VTOL (eVTOL) aircraft for carrying one to six persons that promise unprecedented growth for moving people in urban metropolitan areas with the potential for decreasing road traffic congestion. Advances in cloud computing, data science, open source software and satellite technology can be expected to make UAM operation feasible, efficient and cost effective.

The actual growth, however, will depend on the public perception of the noise and visual scene of hundreds of aircraft flying at low altitude in densely populated areas in addition to the economics of operations. A recent market study for NASA (Goyal, 2019) considered the demand and barriers for Airport Shuttle, Air Taxi and Air Ambulance. This study concluded Air Ambulance is currently not viable using eVTOL. While it valued the Air Taxi and Air Ambulance market at \$250 billion, the available market size was reduced to \$2.5 billion based on overcoming legal/regulatory, certification, public perception and weather constraint challenges.

The UTM and UAM systems must provide all the services—safe separation between aircraft, resolution of conflicts and efficient path planning—enjoyed by traditional air traffic at a fraction of the cost. The ability to model and predict winds at low altitudes presents a challenge for low altitude flights especially for safety of power-limited smaller vehicles. These systems also need to be effectively integrated with the traditional ATC and TFM decision support systems for a seamless air transportation system.

The current vision of UAM, like UTM, is a community-based, collaborative traffic management system with operators coordinating, managing and executing operations within the established legal framework. The UAM system will provide federated services for enabling collaborative management of operations between UAM-vehicle fleet operators, supported by their service suppliers, exchanging information digitally using data-communication networks. In addition to the wired data-communication networks, data will be exchanged wirelessly—using airband radio and satellite communication—between the UAM-vehicles and between UAM-vehicles and ground-based systems. This cyber-physical world of interconnected and interdependent systems increases the vulnerability of aircraft and air traffic management systems to cyberattacks. The impact of security breaches includes potential of both economic harm and loss of life.

Some of the cyber-physical system challenges highlighted by the aviation community are: (1) security and authentication for guarding against malicious activities and for preventing unlawful access to operator and government systems, (2) information/data integrity for ensuring information and data are conforming and trusted, and (3) data access for providing data to law enforcement, state and federal authorities enabled using remote ID information and correlation. Future information management system designs need to leverage International Aviation Trust Framework (IATF) framework and policies.

Unmanned Aircraft Systems

Unmanned Aircraft Systems (UAS) such as General Atomics Predator and Reaper drones and the Northrop Grumman Global Hawk drone are longer and have a greater wingspan compared to a Cessna 172. Global Hawk is about half the length of a Boeing 737 and has a greater wingspan. These aircraft can fly at high altitudes where commercial jet aircraft fly. While these aircraft are well equipped with sensors and largely autonomous, they are remotely piloted. Due to this reason, these aircraft are also called Remotely Piloted Vehicles (RPV). Because flying these vehicles on-demand by the military is disruptive to the scheduled commercial flights, coordination with the civilian air traffic system is needed. In the past several years, the FAA, DOD and NASA have been developing technologies for integrating UAS operations in the NAS. One of the critical enabling technologies for successful integration is Detect and Avoid (DAA) system.

DAA systems provide surveillance information, alerts and maneuver commands to pilots for keeping their aircraft separated—well clear—from other aircraft (Cook et al., 2015). The data needed for defining DAA Well Clear (Cool et al., 2015; Goyal, 2019) (DWC) and performance requirements for alerting and maneuvers have been derived from flight tests and simulations. References 64 and 65 describe prototype DAA methods. RTCA Special Committee 228 (SC-228) published the Phase 1 Minimum Operational Performance Standards (MOPS) for DAA systems (Johnson et al., 1993) and air-to-air radar (Cook et al., 2015) in 2017 based on the outcome of research. The FAA published the corresponding Technical Standard Orders (TSO): TSO-C211 and TSO-C212 in October 2017. Phase 1 MOPS are for operations outside the terminal areas. These MOPS suggest combinations of radar, ADS-B and Tra-c Alert and Collision Avoidance System (TCAS) II technologies. The major objective for Phase 2 MOPS is defining requirements for low cost, size, weight, and power (low C-SWaP) sensors for use by slower UAS aircraft—speed less than 200 knots. The challenge of low C-SWaP sensors is they must provide adequate surveillance volume and accuracy for the DAA system. An additional objective of Phase 2 is DWC while conflicting with aircraft without a functioning transponder. Because Phase 1 permits use of TCAS II, DWC was defined to enclose the TCAS alerting volume to prevent TCAS advisories from triggering. TCAS II advisories cannot be triggered in encounters with aircraft without transponders, therefore, the DWC volume can be reduced for such encounters. Reference 68 presents analysis of the surveillance volume effect on DAA alerting performance for SC-228 Phase 2 MOPS development. Surveillance volume is important for providing enough time for UAS pilots to maneuver according to DAA advisory. While additional surveillance volume benefits the DAA system performance, the sensor cost, size, weight and power increase beyond an extent makes it impractical.

Conclusions

This article begins with the description of the air traffic management system in the United States in which the pilots, controllers and dispatchers interact with each other to ensure safe and efficient air traffic operations. The U.S. airspace was described as organized in 24 regions with 20 Air Route Traffic Control Centers in the continental United States, one in Anchorage and three Combined Center and Radar Approach Controls in Honolulu, Guam and San Juan, respectively. The two main functions of traditional air traffic management, separation assurance and traffic flow management were discussed. The essential steps of trajectory-based separation assurance, trajectory prediction, conflict detection and conflict resolution were described. Traffic Flow Management was described in terms of national traffic flow management and regional traffic flow management. Next, terminal area operations and surface operations were briefly described. Several advanced airspace operations concepts being pursued by the Federal Aviation Administration, National Aeronautics and Space Administration and industry for better data sharing and decisions across decision support tools were highlighted. These included Integrated Demand Management, Integrated Arrival, Departure and Surface Metroplex Traffic Management, Applied Traffic Flow Management and Environmentally Friendly Operations. The NextGen portfolio of projects in various stages of implementation noted in the article included Automatic Dependent Surveillance-Broadcast, Performance-Based Navigation, weather integration, data communications, integrated technologies and global harmonization. Technical challenges and emerging technologies related to Unmanned Aerial System Traffic Management, Urban Air Mobility and large Unmanned Aircraft System for the future air traffic management system were outlined. The major challenge for the U.S. air traffic management system in the future will be safely accommodating the operations of new types of aircraft including electric vertical takeoff and landing aircraft for applications such as mobility of people and cargo in the urban areas. Other challenges will be increased use of automation in the vehicles and systems for managing air traffic, role of humans in highly automated environment and greater role of aircraft operators in the safety and efficiency of operations.

Acknowledgments

The authors are grateful for several NASA colleagues for reviewing the paper and providing feedback for improving the manuscript.

References

- Abramson, M., Refai, M., Santiago, C., 2017. "A Generic Resolution Advisor and Conflict Evaluator (GRACE) in Applications to Detect-And-Avoid (DAA) Systems of Unmanned Aircraft," 17th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference.
- AIRE, 2007. "Atlantic Interoperability Initiative to Reduce Emissions Kick-Off Meeting," October 26, 2007. URL: www.faa.gov/about/office-org/headquarters-offices/ato/publications/071024.

- Alam, S., Nguyen, M.H., Abbass, H.A., Lokan, C., Ellejmi, M., Kirby, S., 2011. Multi-aircraft dynamic continuous descent approach methodology for low noise and emission guidance. *J. Aircraft* 48 (4).
- ASPIRE, "Asia and South Pacific Initiative to Reduce Emissions," URL: <http://www.aspire-green.com/default.asp>.
- Balakrishnan, H., Jung, Y., 2007. "A Framework for Coordinated Surface Operations Planning at Dallas-Fort Worth International Airport," AIAA Guidance, Navigation, and Control Conference, Hilton Head, SC, August 2007.
- Ball, M.O., Chen, C.Y., Hoffman, R., Vossen, T., 2001. Collaborative decision making in air traffic management: current and future research directions. In: Bianco, L., Dell'Omo, P., Odoni, A.R. (Eds.), *New Concepts and Methods in Air Traffic Management*. Springer-Verlag, Germany.
- Bertsimas, D., Patterson, S.S., 1998. The air traffic flow management with en route capacities. *Oper. Res.* 46, 406–422.
- Boucher, O., 2011. Atmospheric science: seeing through contrails. *Nat. Clim. Change* 1, 24–25.
- Brasseur, G.P., Gupta, M., 2010. Impact of aviation on climate: research priorities. *Bull. Amer. Meteor. Soc.* 91, 461–463.
- Clarke, J.-P., 2006. "Development, Design, and Flight Test Evaluation of a Continuous Descent Approach Procedure for Nighttime Operation at Louisville International Airport", Cambridge, MA: PARTNER-COE-2005-002.
- Cook, S.P., Brooks, D., Cole, R., Hackenberg, D., Raska, V., 2015. "Defining Well Clear for Unmanned Aircraft Systems," AIAA Infotech@Aerospace.
- Coppenbarger, R., Mead, R., Sweet, D., 2009. Field evaluation of the tailored arrivals concept for datalink-enabled continuous descent approach. *J. Aircraft* 46 (4), 1200–1209.
- Daytime Operations," Eighth USA/Europe Air Traffic Management Research and Development Seminar, 2009.
- Devlin, C.J., Herndon, A.A., McCourt, S.F., Williams, S.S., 2009. "Performance-Based Navigation Fleet Equipage Evolution," Digital Avionics Systems Conference, 2009.
- Doble, N.A., et al., 2009. "Linking Traffic Management to the Airport Surface Departure Flow Management and Beyond," 8th USA/Europe Air Traffic Management RD Seminar, Napa, CA.
- Duda, D.P., Minnis, P., Costulis, P.K., Palikonda, R., 2003. "CONUS Contrail Frequency Estimated from RUC and Flight Track Data," European Conference on Aviation, Atmosphere, and Climate, Friedrichshafen at Lake Constance, Germany, June-July 2003.
- Effects of Novel Coronavirus (COVID-19) on Civil Aviation: Economic Impact Analysis, Air Transport Bureau, International Civil Aviation Organization (ICAO) Montréal, Canada, September 2020.
- Erzberger, H., 2004. "Transforming the NAS: The Next Generation Air Traffic Control System," Proceedings of the International Congress of the Aeronautical Sciences (ICAS), Yokohama, Japan, August 30.
- Erzberger, H., 2006. "Automated Conflict Resolution for Air Traffic Control," 25th International Congress of the Aeronautical Sciences (ICAS), Hamburg, Germany, September 3-8.
- "Enhanced Traffic Management System (ETMS)," Report No. VNTSC-DTS6-TMS-002, Volpe National Transportation Center, U.S. Dept. of Transportation, Cambridge, MA, March 31, 1999.
- Federal Aviation Administration, 2008. "Air Traffic System Command Center, National Severe Weather Playbook," URL: <http://www.faa.gov/PLAYBOOK/pbindex.html>, February 14.
- Federal Aviation Administration, 2013. "Automatic Dependent Surveillance-Broadcast (ADS-B)," www.faa.gov, Washington, DC, August 2013.
- Federal Aviation Administration, "FAA NextGen Implementation Plan 2018-2019," Washington, DC.
- Federal Aviation Administration, 2020. "FAA Aerospace Forecast, Fiscal Years 2020-2040."
- Federal Aviation Administration NextGen Office, "Urban Air Mobility (UAM) Concept of Operations," V1.0, 800 Independence Avenue, S.W., Washington, Dc 20591, June 26, 2020. URL: https://assets.evtol.com/wp-content/uploads/2020/07/UAM_ConOps_v1.0.pdf. (Accessed 08/26/2020).
- Fry, B.L., Zabinsky, J.M., 1967. "Feasibility of V/STOL Concepts for Short-haul Transport Aircraft (No. 743)," National Aeronautics and Space Administration.
- Gong, C., 2020. "Airspace Technology Demonstration 3 (ATD-3): Applied Traffic Flow Management Project Overview," NASA Document 20160014325, August 31, 2016. URL: <https://ntrs.nasa.gov/citations/20160014325>. (Accessed 8/26/2020).
- Goyal, R., 2019. "Urban Air Mobility (UAM) Market Study," Technical Report HQ-E-DAA-TN65181, National Aeronautics and Space Administration, March.
- Green, S., Vivona, R., 1996. "Field Evaluation of Descent Advisor Trajectory Prediction Accuracy," AIAA Guidance, Navigation, and Control Conference, San Diego, CA, July 1996.
- Guiding, J., et al., 2009. "US/Europe Comparison of ATM-related Operational Performance," 8th USA/Europe Air Traffic Management RD Seminar, Napa, CA.
- Howell, D., Bennett, M., Bonn, J., Knorr, D., 2003. "Estimating the En Route Efficiency Benefits Pool", 5th USA/Europe Air Traffic Management Seminar, Budapest.
- Hileman, J., Spakovszky, Z., Dreia, M., Sargeant, M., 2007. Airframe design for "silent aircraft". In 45th AIAA Aerospace Sciences Meeting and Exhibit (p. 453), January 2007.
- Intergovernmental Panel on Climate Change's (IPCC) Working Groups, "Aviation and the Global Atmosphere," Tech. Report, Cambridge University Press, Cambridge, UK, 1999.
- Johnson, M., Mueller, E.R., Santiago, C., 2015. "Characteristics of a Well Clear Definition and Alerting Criteria for Encounters between UAS and Manned Aircraft in Class E Airspace," Eleventh UAS/Europe Air Trac Management Research and Development Seminar, pp. 23-26.
- Joint Planning and Development Office (JPDO), "Concept of Operations for the Next Generation Air Transportation System," Version 3.0, Washington, DC, October 2009.
- Jung, J., et al., 2013. "Evaluation of the Controller-Managed Spacing Tools, Flight-deck Interval Management and Terminal Area Metering Capabilities for the ATM Technology Demonstration," Tenth USA/Europe Air Traffic Management Research and Development Seminar, Chicago, IL.
- Jung, Y.C., Coupe, W.J., Capps, A., Engelland, S., Sharma, S., 2019. "Field Evaluation of the Baseline Integrated Arrival, Departure, and Surface Capabilities at Charlotte Douglas International Airport," Thirteenth USA/Europe Air Traffic Management Research and Development Seminar, Vienna, Austria, June 17-21.
- Krozel, J., Kicing, R., Andrews, R., 2011. "Classification of Weather Translational Models for Nextgen," American Meteorological Society, Seattle, WA, January 2011.
- Kettunen, T., Hustache, J.-C., Fuller, I., Howell, D., Bonn J., Knorr, D., 2005. "Flight Efficiency Studies in Europe and the United States," 6th USA/Europe Air Traffic Management Seminar, Baltimore.
- Krozel, J., Rosman, D., Grabbe, S., 2003. "Analysis of En Route Sector Demand Error Sources," AIAA Guidance, Navigation and Control Conference, Austin, TX, August 11-14.
- Lee, D.S., et al., 2009. Transport impacts on atmosphere and climate: aviation. *J. Atmos. Environ.*
- Lee, S.M., Park, C., Thipphavong, D.P., Isaacson, D.R., Santiago, C., 2016. NASA-TM-2016-219067, "Evaluating Alerting and Guidance Performance of a UAS Detect-And-Avoid System," NASA Ames Research Center.
- Luthe, R.K., 2015. ICAO aviation system block upgrades: A method for identifying training needs. *Int. J. Aviation Aeronaut. Aerosp.* 2 (4).
- Malik, W.A., Gupta, G., Jung, Y., 2010. "Managing Departure Aircraft Release for Efficient Airport Surface Operations," AIAA Guidance, Navigation and Control Conference, Toronto, Canada.
- Muñoz, C., Narkawicz, A., Hagen, G., Upchurch, J., Dutle, A., Consiglio, M., Chamberlain, J., 1993. "DAIDALUS: Detect and Avoid Alerting Logic for Unmanned Systems," IEEE/AIAA 34th Digital Avionics Systems Conference (DASC), 2015, pp. 5A1-1.
- Murphy, J.R., Hayes, P.S., Kim, S.K., Bridges, W., Marston, M., 2016. "Flight Test Overview for UAS Integration in the NAS Project," Atmospheric Flight Mechanics Conference AIAA SciTech.
- Nolan, M., *Fundamentals of Air Traffic Control*, 4th ed., Thomson Brooks/Cole Belmont, CA.2000.
- Rathinam, S., Montoya, J., Jung, Y., 2008. "An Optimization Model for Reducing Taxi Times at the Dallas Fort Worth International Airport," 26th International Congress of the Aeronautical Sciences (ICAS), Anchorage, Alaska.
- Ren, L., Clarke, J.-P., 2007. "Flight Demonstration of the Separation Analysis Methodology for Continuous Descent Arrival", Seventh USA/Europe ATM RD Seminar, Barcelona, Spain.
- Reynolds, T.G., 2008. "Analysis of Lateral Flight Inefficiency in Global Air Traffic Management," 8th AIAA Aviation Technology, Integration and Operations Conference, Anchorage, Alaska, September 14-19.
- Rios, J., Lohn, J., 2009. "A Comparison of Optimization Approaches for Nationwide Traffic Flow Management," AIAA 2009-6010, AIAA Guidance, Navigation, and Control Conference, Chicago, Illinois, August 10–13.
- RTCA Inc., "Minimum Operational Performance Standards (MOPS) for Air-to-Air Radar for Tra-c Surveillance," DO-366, 2017.
- RTCA Inc., "Minimum Operational Performance Standards (MOPS) for Detect and Avoid (DAA) Systems," DO-365, 2017.
- Radio Technical Commission for Aeronautics, "RTCA SC-214/ EUROCAE WG-78 Standards for Air Traffic Data Communication Services," www.faa.gov, August 8, 2013.

- SAE Committee A-21, "SAE-AIR-1845: Procedure for the Calculation of Airplane Noise in the Vicinity of Airport", March 1986.
- Schafer, A., 2000. Regularities in travel demand: an international perspective. *J. Transport Stat.* 3 (3).
- Sengupta, P., Kwan, J., Menon, P.K., 2005. Optimal traffic flow scheduling using high-performance computing. *J. Aerospace Inform. Syst.* 12 (10), 661–671.
- Shresta, S., Neskovic, D., Williams, S.S., 2000. "Analysis of Continuous Descent Benefits and Impacts During." AIAA 2000-2221, 16th AIAA Aviation Technology, Integration, and Operations Conference, Washington, DC, pp. 4221.
- Sprong, K.R., Klein, K.A., Shiotsuki, C., Arrighi, J., Liu, S., 2008. "Analysis of Atlantic Interoperability Initiative to Reduce Emissions Continuous Descent Arrival Operations at Atlanta and Miami," 27th Digital Avionics Systems Conference, Minneapolis.
- Sridhar, B., Kulkarni, D., Sheth, K.S., 2011. Impact of uncertainty on the prediction of airspace complexity of congested sectors. *ATC Quart.* 19 (1).
- Sridhar, B., Chen, N., Ng, H., 2013. "Energy Efficient Contrail Mitigation Strategies for Reducing the Environmental Impact of Aviation," 10th USA/Europe Air Traffic Management RD Seminar, Chicago, IL, May 2013.
- Sun, D., Sridhar, B., Grabbe, S., 2010. Disaggregation method for an aggregate traffic flow management model. *J. Guid. Control Dyn.* 33 (3), 666–676.
- The Economic Impact of Civil Aviation on the U.S. Economy, U.S. department of Transportation, Federal Aviation Administration, August 2011.
- Wambsganss, M.C., 1997. Collaborative decision making through dynamic information transfer. *Air Traffic Control Quart.* 4, 107–123.
- Williams, V., Noland, R.B., Toumi, R., 2002. Reducing the climate change impacts of aviation by restricting cruise altitudes. *Transport. Res. Part D: Transport Environ.* 7 (5), 451–464.
- Wu, M.G., Cone, A.C., and Lee, S., "Detect and Avoid Alerting Performance with Limited Surveillance Volume for Non-Cooperative Aircraft," *AIAA SciTech Forum*, San Diego, California, January 7–11, 2019.
- Xue, M., Atkins, E.M., 2006. "Noise-Minimum Runway-Independent Aircraft Approach Design for Baltimore-Washington International Airport", *J. Aircraft*, Vol. 43, No. 1, January 2006.
- Yoo, H.-S., et al., 2016. "Required Time of Arrival as a Control Mechanism to Mitigate Uncertainty in Arrival Traffic Demand Management," 35th IEEE Digital Avionics Systems Conference, Sacramento, CA, 2016.

Air Traffic Flow and Capacity Management

Li Weigang, Cristiano P Garcia, University of Brasilia - UnB, Department of Computer Science - CIC, Brasilia, DF Brazil

© 2021 Elsevier Ltd. All rights reserved.

Glossary	380
Capacity Management	381
Airspace Capacity Assessment	381
Working Positions Management	381
Flexible Use of Airspace	382
Airport Capacity Assessment	383
Air Traffic Flow Management	384
ATFM Measures	384
Slots	385
Miles-In-Trail	385
Minutes-In-Trail	385
Rerouting	385
Ground Delay	385
Linear Holding	385
Holding Pattern	386
Diversions	386
List of Flights Exempted From Flow Control Measures	386
Phases	386
Strategic	387
Pre-Tactical	387
Tactical	387
Post-Operational	387
CDM	387
The Future of ATFCM	387
Acknowledgment	388
Relevant Websites	388
References	388
Further Reading	389

Glossary

Airspace Users (AU) Civil aircraft operators (including airlines, charter flights, cargo flights, general aviation), military operators, hot air balloon operators, and unmanned aerial vehicle operators.

Air Traffic Control (ATC) Service provided by an air traffic controller in which the aircraft receives a series of instructions to ensure a *safe, orderly, and expeditious* flow of traffic. ATC Includes the service provided by air traffic controllers in the Tower during take-off and landing, approach control during climb and descent, and area control center during the en-route phase of flight.

Air Traffic Controller (ATCO) Professional responsible for providing air traffic services to all aircraft in the airspace under their control.

Air Traffic Flow Management (ATFM) Set of ATFM measures applied tactically when a demand/capacity imbalance takes place. The main goal of these measures is to reduce demand temporarily to meet the reduced capacity.

Air Traffic Management (ATM) A broad term that includes Air Traffic Services (ATS), Air Traffic Flow and Capacity Management (ATFCM), and Airspace Management.

Air Traffic Services (ATS) All services provided to an aircraft during the flight, including Air Traffic Control (ATC), Flight Information Service (FIS), and Alert Service.

Area Navigation (RNAV) Navigation method composed of a series of ground or space-based stations to allow navigation in any desired course within the coverage area of the stations.

ATC sectors Portions of airspace where an ATCO is responsible for providing air traffic services.

Distance Measuring Equipment (DME) Ground-based navigation device used for distance measuring between the aircraft and the transmitter.

Flow Management Positions (FMP) Cell located in ATC facilities responsible for connecting the local ATC needs with network manager instructions.

International Civil Aviation Organization (ICAO) United Nations agency created in 1944 to regulate the international civil aviation.

Network Manager Central organization responsible for ATFCM and coordinating the application of flow control measures in a country or set of countries.

NOTAM Notice to airmen used to alert pilots of any aeronautical information that might affect the safety, regularity, or efficiency of the flight.

Very High Frequency Omnidirectional Range (VOR) Ground-based navigation device that emits radio beacons which are interpreted by the receiving equipment onboard, allowing an aircraft to determine its relative position to the transmitter.

Air traffic has been increasing in the past years, and the predictions indicate a more significant growth in the future. On the other side, the infrastructure required to provide a safe, orderly, and expeditious flow of traffic does not grow as fast because it is costly. The high price tag encourages the authorities to make sure the existing infrastructure is optimized as much as possible before making any investment to increase capacity. Air Traffic Flow and Capacity Management (ATFCM) is part of Air Traffic Management (ATM) that seeks to balance the increasing demand with a limited capacity (Cinar and Demirel, 2017).

Demand and capacity are on opposite ends of the scale, but proper management of both is a necessary step to provide a quality service for all airspace users.

Capacity Management

To correctly manage the airspace capacity, it is paramount to have an in-depth knowledge of the existing capacity of airports and air traffic control facilities. It is also essential to know the limitations involved in the expansion of this capacity to cope with the rise of demand.

The ideal scenario would be real-time monitoring of the existing capacity and current demand where the capacity could be increased quickly whenever necessary. In the real world, this expansion of airspace and airport capacity is not so simple.

Airspace Capacity Assessment

Many factors can affect the capacity of an Air Traffic Control sector (ICAO, 2005a).

When assessing capacity values, factors to be taken into account should include:

1. the type of Air Traffic Services (ATS) provided,
2. the complexity of the control sector or the aerodrome concerned,
3. controller workload as measured by the capacity assessment team, including control and coordination tasks to be performed,
4. the types of communications, navigation and surveillance systems in use, their degree of technical reliability and availability as well as the existence of backup systems, and
5. existence of ATC systems providing controller support and alert functions.

The most crucial factor which determines the capacity of an ATC sector is the air traffic controller workload (Majumdar and Ochieng, 2002). Therefore the capacity of a given sector is a measure of all factors which affect the mental and physical activities of the air traffic controller. The goal is to keep the workload below an acceptable level where safety is not jeopardized.

The methods for measuring controller workload are usually categorized as subjective or objective. Subjective self-assessment of controllers is carried out either by instantaneous techniques or non-instantaneous techniques. Objective methods involve direct observation of controllers, usually by other controllers or ATC system experts.

Although a lot of effort has been made to define and measure the workload of an air traffic controller, there is not yet one definitive method to assess the airspace capacity based on the air traffic controller workload. The primary concern is transferability, as the techniques developed for the specific scenario of a country can hardly be applied directly on other countries (Majumdar et al., 2005).

Airspace capacity can be improved in the long term by redesigning the airspace to create optimized routes that will accommodate more aircraft without increasing the ATCO workload. In the short term, better management of the working positions is enough to alleviate the burden on overloaded sectors.

Working Positions Management

The amount of aircraft being serviced by an air traffic controller on a given sector plays a significant role in the workload of the Air Traffic Controller (ATCO). The management of working positions might be enough to reduce the workload and consequently increase the capacity of a portion of the airspace.

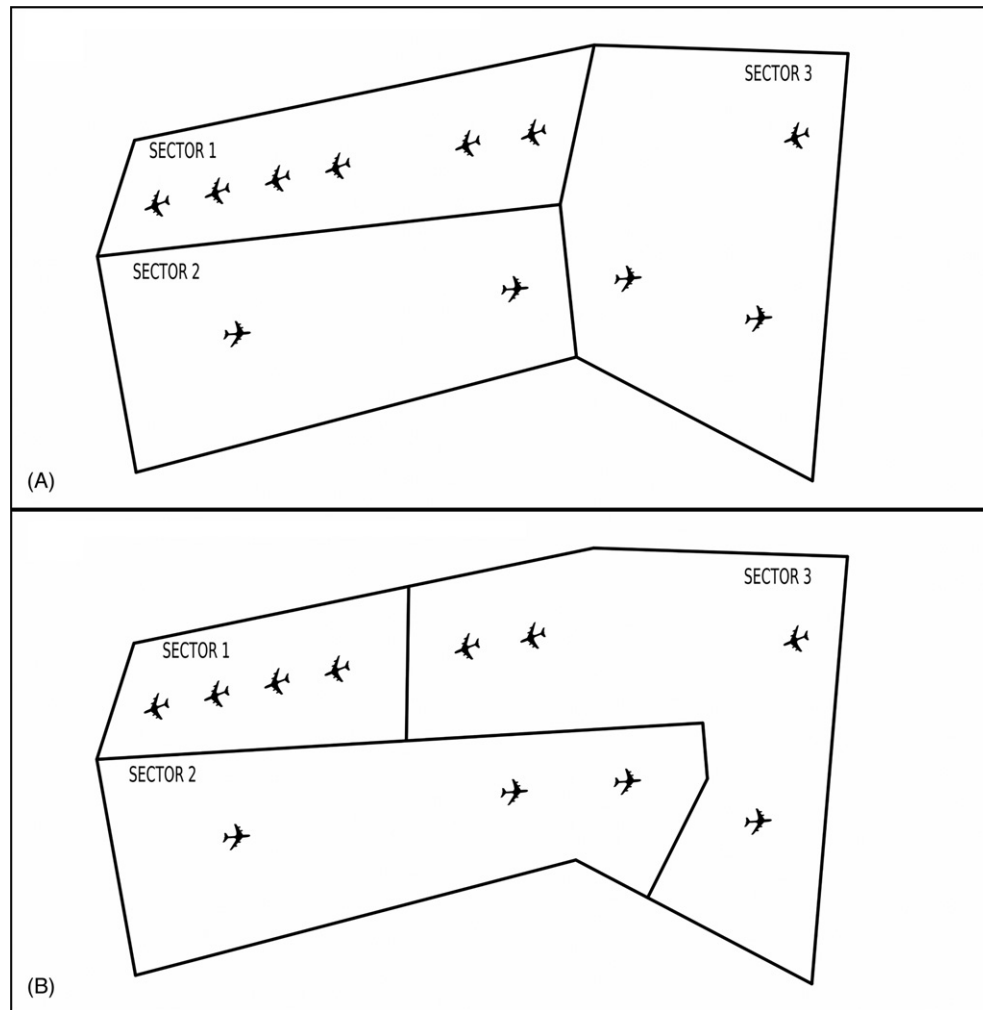


Figure 1 Dynamic sector boundaries can help reduce the workload of some sectors. (A) Before dynamic airspace configuration. (B) After applying dynamic airspace configuration.

It is a common practice in ATC facilities to group two or more sectors when the demand is low. Once the demand starts to rise, the ATCO can monitor the workload of the ATCOs and ungroup the sectors as a tactical solution to increase the airspace capacity. In some cases, a specific sector is only ungrouped if the demand is expected to reach a peak due to some unusual circumstances.

ATC sectors traditionally have well defined boundaries. But some recent research suggests (Martinez et al., 2007; Zelinski and Lai, 2011) that dynamic sector configuration according to the evolution of traffic gives better results on smoothing out of peaks of demand and results in a higher capacity. Fig. 1 shows an example of how the dynamic changing of boundaries can reduce the workload on some sectors. On Fig. 1A, sector 1 is overloaded while sector 2 has available capacity. Fig. 1B shows the result of dynamic sectorization, where all sectors have similar workloads.

Flexible Use of Airspace

Airspace is a limited resource and the consensus states that certain portions of the airspace should not be restricted to a single type of use.

The key to avoiding unnecessary reduction of the airspace capacity is to limit the duration of the restrictions on a portion of the airspace only for the period in which they are required.

Restricted airspaces are portions of the airspace wherein activities must be confined because of their nature, or wherein limitations are imposed upon aircraft operations that are not a part of those activities, or both. Permanently restricted airspaces are depicted on aeronautical charts, while temporarily restricted airspace is published via NOTAM. Fig. 2 shows an example of a restricted area published on an aeronautical chart.

There are three main types of restricted airspace:

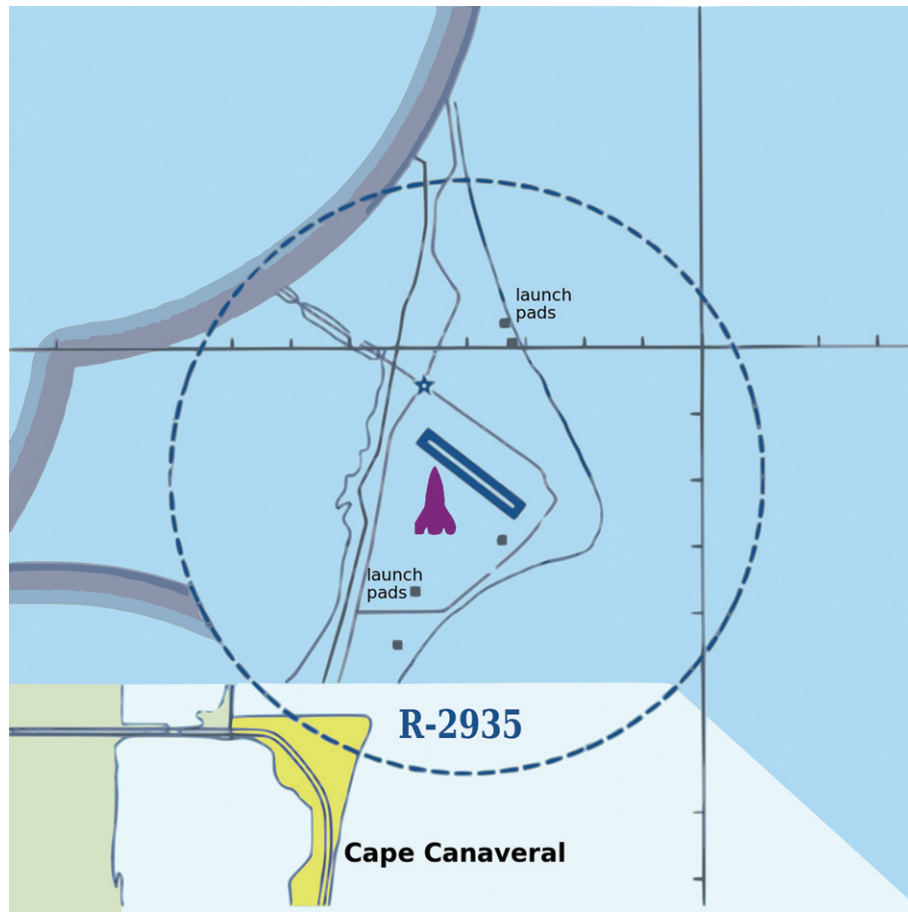


Figure 2 Restricted area 2935 used for rocket launches at Cape Canaveral.

- Danger area: An airspace within which activities dangerous to the flight of aircraft may exist at specified times.
- Prohibited area: An airspace above the land areas or territorial waters of a State within which the flight of aircraft is prohibited.
- Restricted area: An airspace above the land areas or territorial waters of a State, within which the flight of aircraft is restricted following certain specified conditions (ICAO, 2005a).

Flexible use of airspace (FUA) is a concept that allows restricted airspace to be activated when necessary and makes that same airspace available for the use by all aircraft for the rest of the time.

Airport Capacity Assessment

The capacity of an airport can be divided into two main domains, namely the airside and the landside. The airside capacity includes the Terminal Manoeuvring Area (TMA) approach and departure interface, the runways, the taxiway configuration and aprons up until the gates. The landside capacity involves everything that happens after a passenger leaves the plane, including terminal structure, baggage handling, and transportation links to and from the airport. Fig. 3 depicts the airside and landside separation present in most airports. Delays can also be caused by systemic processes that affect landside capacity but they are out of the scope of this work (O'Flynn, 2016).

The airside capacity assessment can be viewed as the maximum sustainable throughput of a runway system which, in practice answers the question, "How many aircraft operations can an airfield reasonably accommodate in a given period of time when there is a continuous demand for service during that period?" (Fisher, 2012).

Airport capacity enhancement usually starts with an efficiency assessment. If demand continues to surpass the capacity regularly and the systems are already operating at maximum efficiency, the solution to increase capacity will be the construction of new and more efficient infrastructure.

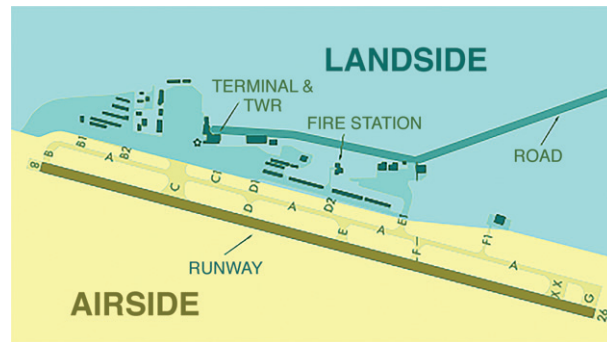


Figure 3 Areas that affect airside capacity and landside capacity.

Air Traffic Flow Management

Airspace capacity can't always be increased as fast as it would be desirable. Since capacity can't always be increased, Air Traffic Flow Management (ATFM) has been long recognized as a principal instrument to deal with capacity shortfalls and delay phenomena in the transportation system (Madas and Zografos, 2008).

The problem is to determine how to optimally control aircraft by rerouting, delaying, or adjusting the speed of the aircraft in the air traffic control system to avoid airspace regions that have reduced capacities. The overall objective is the minimization of delay costs that include all applicable fuel costs, safety costs, and taxes (Bertsimas and Patterson, 2000).

ATFM Measures

A variety of tools are available to help modulate the demand. Each ATFM measure can be used alone or in combination with other measures. The most common ATFM measures are described further. Fig. 4 shows the decision flow usually followed when applying ATFM measures.

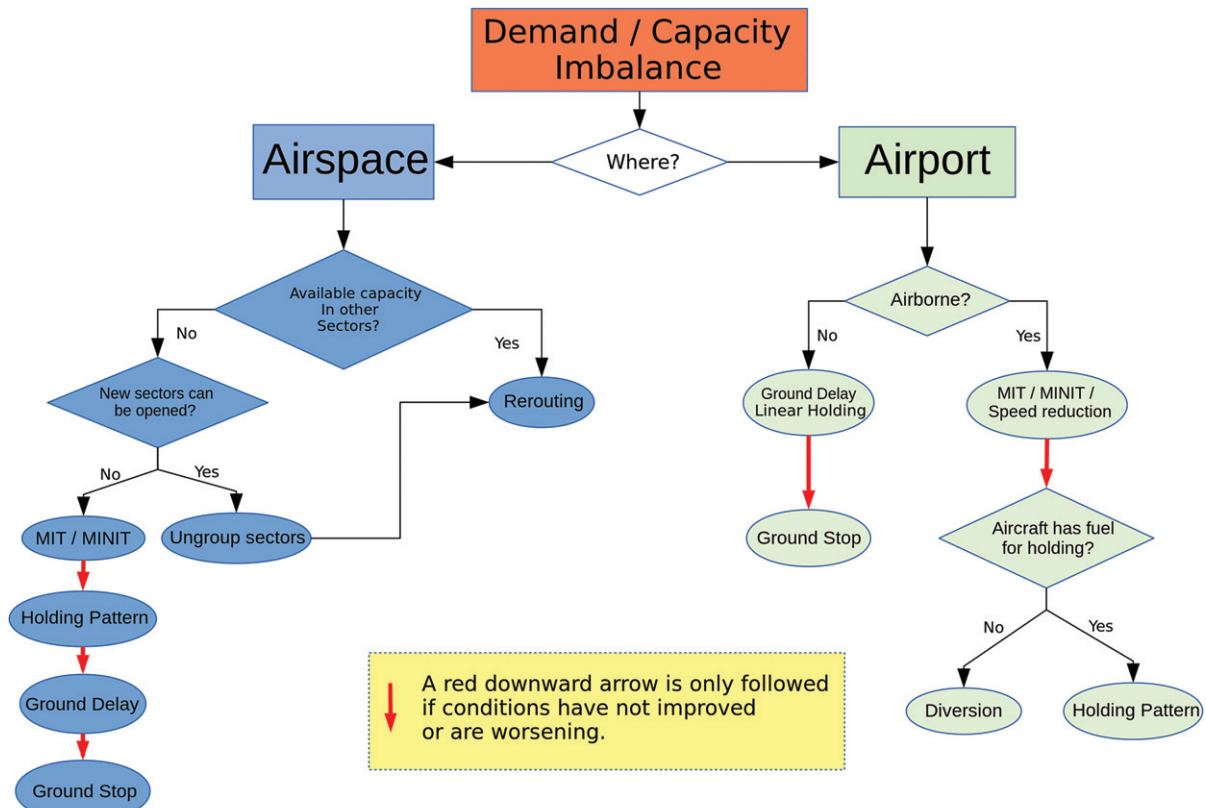


Figure 4 Decision flowchart for ATFM measures that will be applied according to the situation.

Slots

Slots are used by the airport administration to allocate a specific time window for an aircraft to take-off or land. When slots are applied, the airport is considered as a coordinated airport. Aviation is the most global of industries. Where capacity constraints exist, the guidelines should be uniform around the world. These guidelines are published by the International Air Transport Association (IATA).

Coordination is the allocation of constrained or limited airport capacity to aircraft operators to ensure a viable airport and air transport operation. Coordination seeks to maximize the efficient use of airport infrastructure. It is not a solution to the fundamental problem of a lack of airport capacity. In all instances, coordination should be seen as an interim solution to manage congested infrastructure until the long term solution of expanding airport capacity is implemented (IATA, 2017).

Miles-In-Trail

One way to impose restriction criteria on a stream of aircraft approaching the same airport, to a specific sector or a portion of airspace is to determine a minimum distance that will be acceptable between two aircraft. This type of ATFM measure is known as miles-in-trail (MIT). It requires the transferring ATCO to use other measures to obtain the desired separation. For example, speed restriction, holding pattern, or vectoring.

Research around this topic include the development of an MIT Impact Assessment Capability that could allow a flow manager to tailor each MIT, accurately setting start time, stop time, spacing value, and thus the affected flight set (Ostwald et al., 2006).

Minutes-In-Trail

Instead of requesting a specific distance between two consecutive aircraft, the receiving sector might require a specific time spacing over the sector boundaries.

This ATFM measure is called Minutes-in-trail (MINIT). It can be used when the receiving sector is not capable to provide RADAR surveillance. MINIT is very similar to the MIT as both are considered flow-rate limits but the latter is usually applied when there is a need for a lower flow.

Rerouting

When the weather conditions are poor, the capacities of some airports and ATC sectors drop significantly. Depending on the conditions the capacity can be reduced to zero. Aircraft must then fly alternative routes if they were scheduled to pass through airspace regions of reduced capacity (Bertsimas and Patterson, 2000). This enables the use of airspace capacity that is not fully used at the moment and relieves the burden on an overloaded sector. The challenge is to define a new route that is far enough from the original one to alleviate the sector that is facing a temporary peak of demand without increasing too much the time and cost of the new route.

Aircraft usually depart with a predefined route that is programmed to be flown if nothing unexpected happens. This route is declared in the flight plan and can be cleared as proposed or changed by the ATC before departure. The complexity of the problem of rerouting often requires the use of decision support tools (DST) to provide a better solution than the one given by the human operators. For example, a linear integer-programming model to support strategic ATFM decisions related to ground delays and dynamic rerouting of flights proposed by Mukherjee and Hansen (2009) has resulted in 10%–15% delay cost reduction compared to static rerouting. This model also achieved about 50% less delay cost compared to a pure ground holding strategy.

Ground Delay

The use of ground delays sometimes is also referred to as Ground Delay Program (GDP). The Network Manager (NM) issues GDPs whenever it expects sustained periods of congestion at an airport. Usually, this is caused by a sudden decline in the destination airport arrival capacity due to adverse weather conditions. In a GDP, flights that are scheduled to arrive during the expected periods of congestion are assigned delays before their departure, hence, the term ground delay. The reason is that it is both safer and less expensive to delay a flight on the ground (i.e., before its take-off) than to delay an airborne flight (Vossen and Ball, 2006). If the capacity drops even further or if it stays lower for long periods, the GDP might evolve to a Ground Stop Program (GSP), where all aircraft destined to one airport are not allowed to take off until the capacity is restored.

GDP is a tactical measure, therefore the decision to use it is made mostly on the day of the operation. Slots are similar to GDP but they are used as a strategic measure since one airport becomes coordinated only due to circumstances that are foreseen months before the operation.

The simplest form of implementing GDP is by establishing a fix distance or time between two aircraft heading to the same destination, also known as Minimum Departure Interval (MDI). The use of MDI doesn't require any decision support tools since it relies solely on the expertise of the ATCO. But there are very sophisticated algorithms being developed that aim to improve the results of GDP. One example is the Dynamic Stochastic Ground Holding (DSGH) algorithm that dynamically adapts to weather forecasts by revising, if necessary, departure delays (Mukherjee et al., 2012).

Linear Holding

Linear holding (LH) can be seen as a postponing of the delay that would be applied on the ground as a part of a GDP. The aircraft is allowed to take off a few minutes earlier but has to fly at a slower speed. The advantage over ground holding is that once the aircraft is airborne, the speed can be adjusted dynamically within a specific range if the conditions change during the flight. For this reason,

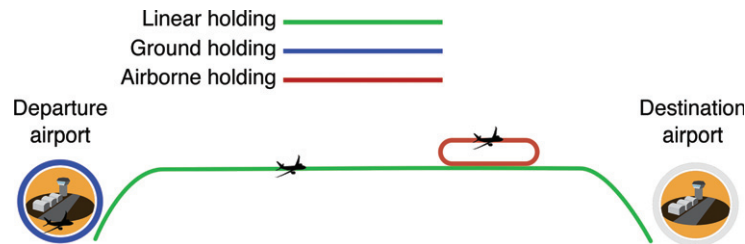


Figure 5 Ground delay can be applied before take-off. Linear holding can be applied for the whole flight. Holding pattern can be applied as long as the aircraft has enough fuel.

linear holding can be considered more flexible when compared to a ground hold. LH also has some advantages when compared to conventional airborne holding.

Typical airborne holding would consume more fuel due to the extended flight track (assuming no specific speed adjustment), while ground holding has no impact on fuel consumption in most cases. Due to the increased extra fuel, the airborne holding time is fairly limited, taking account that safety-related issues may arise from a reduction of the on-board reserve fuel (Xu and Prats, 2017).

Holding Pattern

Conventional Holding Pattern (HP) is used to maintain the aircraft in a specified space while the adjacent sector or the destination airport has a limitation in capacity. It is probably the oldest of the ATFM measures.

Every holding pattern is designed with a fix that will define the beginning of the holding. It can be a navigational aid like a VOR, a DME distance or an RNAV fix.

The speed used during the holding pattern is usually lower than the speed used during the cruise phase of the flight to decrease the turn radius and the size of the airspace occupied (ICAO, 2016).

Corrections for wind in the heading and timing can be applied to maintain the holding pattern as specified. These corrections usually are performed by the aircraft FMS automatically.

A holding pattern is not the most efficient ATFM measure but sometimes it is the last resort available to ATCOs.

Ground delay, linear holding, and holding pattern have similar effects, but the moment of application and the flexibility are slightly different. Fig. 5 shows where these ATFM measures can be applied.

Diversions

Diversions are not considered ATFM measures but are included here just for completeness, as they can be used in extreme cases when the capacity of the destination airport drops severely and unexpectedly.

There can be many causes for this sudden drop in capacity, like a damaged aircraft on the runway, unserviceable navigational aids or adverse weather conditions. If the estimated time for restoring the capacity of the airport is greater than the time available for holding according to the fuel of the aircraft, the flight is diverted to an alternate airport. This is very costly for the airline company and brings a lot of inconvenience to the passengers. Therefore a diversion is used only when there is no other option available.

List of Flights Exempted From Flow Control Measures

The International Civil Aviation Organization (ICAO) determines that the following types of flights should be granted exemption from flow control measures (ICAO, 1984):

- flights in a state of emergency, including flights subjected to unlawful interference,
- flights operating for humanitarian reasons,
- medical flights specifically declared by medical authorities,
- flights on search and rescue missions,
- flights with "Head of State" status, and
- other flights as specifically required by state authorities.

Phases

On a time scale, Air Traffic Flow Management (ATFM) can be divided into four phases. Firstly, the strategic phase deals with questions that require a change in infrastructure to be solved. The pre-tactical phase prepares the final adjustments for the day of the operations when the tactical phase is carried out. After the operation, the data collected is analyzed on the post-operational analysis. Fig. 6 resumes the moment of each phase and the actions that are taken (ICAO, 2005b).

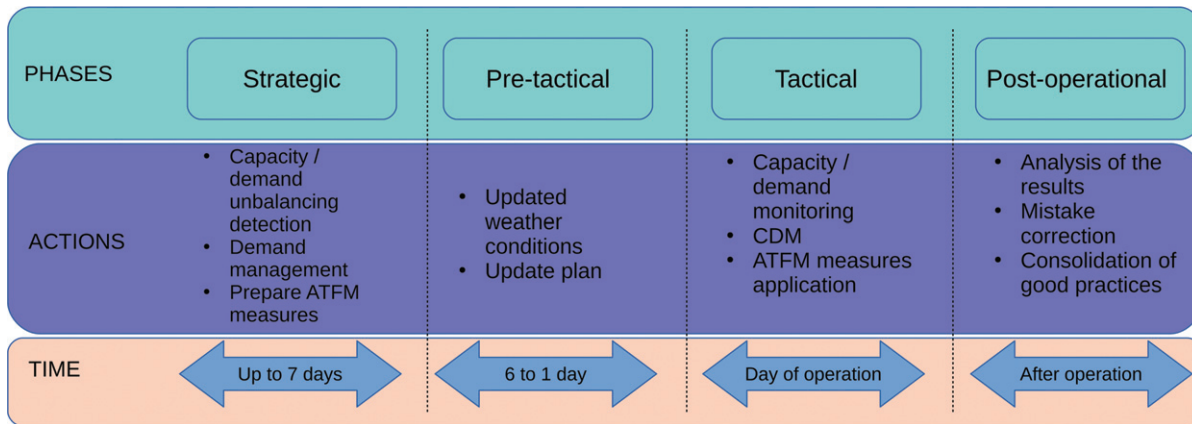


Figure 6 Phases of ATFM and the actions taken on each phase.

Strategic

The strategic phase of the flow management comprises the period seven days or more prior to the day of a specific operation. During this phase data are collected in order to detect any major imbalance between ATC capacity and the foreseen demand. Seasonal or significant events like national holidays are examples of a possible cause of a spike in demand. Military operations, significant weather conditions, or construction work in key airports could be responsible for a drop in demand. If such imbalance is detected, the measures taken to reestablish the balance can include: optimization of the existing capacity, use of other available capacities and demand regulation.

Pre-Tactical

The pre-tactical phase of the flow management takes place between 6 days and 1 day before the operation. It consists of a fine-tuning upon the planning that was prepared during the strategic phase. Based on the outcome of the ATFM measures taken on the strategic phase and the updated information about capacity and demand, corrections are made and some of the possible tactical ATFM measures are prepared.

Tactical

On the day of the operation, the NM and the Flow Management Positions (FMPs) carefully monitor the evolution of the traffic in a timely manner. Special attention is given to those sectors or moments in which the capacity is going to be reduced or the demand is expected to rise. If necessary, new restrictions are applied.

Post-Operational

During the previous phases, data are collected in order to allow a post-operational analysis of the actions.

This is a cyclic process and the outcome of one iteration is used to improve the actions taken on the next one. This is important to correct mistakes and to consolidate good practices.

CDM

Many complex processes require the cooperation of multiple parts to function properly. The outcome of these processes can be improved if each stakeholder is allowed take part on important decisions.

The ATFM is one of these complex systems which require different parts working together synchronously to produce the desired effect. The decisions taken for a specific flight or a group of flights have huge impacts for every party involved. In this context, the Collaborative Decision-Making (CDM) is applied whenever possible in ATFCM.

CDM embodies a new philosophy for managing air traffic. It is based on the belief that air traffic flow management can be improved if there is a closer collaboration between the stakeholders, including the Network Manager (NM), the Flow Management Positions (FMPs), and the airspace users (AUs) with large benefits for all parties. This collaboration takes the form of mutual exchange of data and more flexible and efficient collaborative procedures (Ball and Lulli, 2004).

The Future of ATFCM

One of the most promising technologies for the future of ATM is four-dimensional trajectory based operation (4D TBO). This concept allows a precise description of an aircraft's path in space and time during the whole flight. Conflict detection and resolution

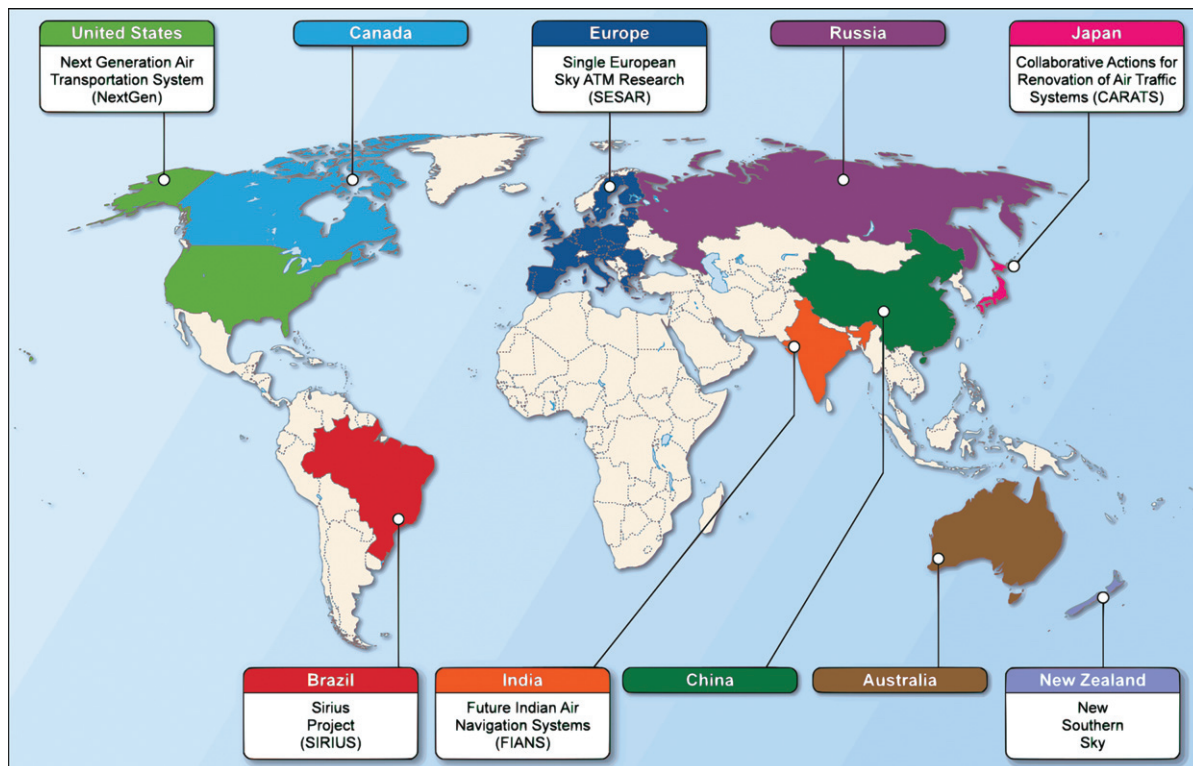


Figure 7 4D trajectory based operations programs of each country/region.

will be possible even before the aircraft starts the engines, reducing congestion in airspace and the workload of Air Traffic Controllers (Ayhan et al. 2018; Ribeiro et al., 2018). Trajectory prediction will be more precise because the onboard Flight Management System (FMS) will receive updated information about wind and temperature available for the intended route (Legrand et al. 2019).

Many countries/regions are developing or starting the implementation of their own 4D TBO programs like Next Generation Air Transportation System (NextGen) in the United States, Single European Sky ATM Research (SESAR) in Europe, Collaborative Action for Renovation of Air Transport Systems (CARATS) in Japan, Sirius in Brazil and Future Indian Air Navigation System (FIANS) in India. Fig. 7 shows a map of the countries and their respective programs. Once 4D TBO is fully implemented, it will have a huge potential to entirely extinguish delays or providing optimal solutions in the cases where ATFM measures have to be applied.

The CDM is going to be much more efficient in the future as the System Wide Information Management (SWIM) comes into practice. This effective information management and sharing enables each participant to be aware of information of relevance to other participants' decisions (ICAO, 2018).

Acknowledgment

This work has been partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under the grant number 311441/2017-3.

Relevant Websites

Federal Aviation Administration: <https://www.faa.gov/>

European Organization for the Safety of Air Navigation: <https://www.eurocontrol.int/>

International Civil Aviation Organization: <https://www.icao.int/>

Australian Civil Aviation Safety Authority: <https://www.casa.gov.au/>

References

- Ayhan, S., Costas, P., Samet, H., 2018. Prescriptive analytics system for long-range aircraft conflict detection and resolution. In: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 239–248.
- Ball, M.O., Lulli, G., 2004. Ground delay programs: optimizing over the included flight set based on distance.. Air Traffic Cont. Q. 12 (1), 1–25.

- Bertsimas, D., Patterson, S.S., 2000. The traffic flow management rerouting problem in air traffic control: A dynamic network flow approach. *Transp. Sci. INFORMS* 34 (3), 239–255.
- Çinar, E., Demirel, S., 2017. Eurocontrol slot allocation methodology and new approach for calculated take off times (CTOT): alternative CTOT versus CTOT. *Anadolu Üniversitesi Bilim Ve Teknoloji Dergisi A-Uygulamalı Bilimler ve Mühendislik* 18 (1), 69–77.
- Fisher, L., 2012. Inc. ACRP Report 79: Evaluating Airfield Capacity, Transportation Research Board of the National Academies, Washington, DC.
- IATA, 2017. Worldwide Slot Guidelines, 2017. International Air Transport Association.
- ICAO, 1984. Doc 9426 - Air Traffic Services Planning Manual. Montreal.
- ICAO, 2005a. Annex 2 - Rules of the Air. Montreal.
- ICAO, 2005b. Doc 9854 - Global Air Traffic Management Operational Concept. Montreal.
- ICAO, 2016. Doc 4444 - Procedures for Air Navigation Services - Air Traffic Management. Montreal.
- ICAO, 2018. Doc 9971 - Manual on Collaborative Air Traffic Flow Management (ATFM), Montreal, Montreal.
- Legrand, K., Rabut, C., Delahaye, D., 2019. Correction and Optimization of 4D Aircraft Trajectories by Sharing Wind and Temperature Information. *Université de Toulouse*.
- Madas, M.A., Zografos, K.G., 2008. Airport capacity vs. demand: mismatch or mismanagement. *Transp. Res. Policy Pract.* 42 (1), 203–226.
- Majumdar, A., et al., 2005. En-route sector capacity estimation methodologies: an international survey. *J. Air Transp. Manage.* 11 (6), 375–387.
- Majumdar, A., Ochieng, W.Y., 2002. Factors affecting air traffic controller workload: multivariate analysis based on simulation modeling of controller workload. *Transp. Res. Rec.* 1788 (1), 58–69.
- Martinez, S. et al., 2007. A weighted-graph approach for dynamic airspace configuration. In: *Aiaa Guidance, Navigation and Control Conference and Exhibit*, p. 6448.
- Mukherjee, A., Hansen, M., 2009. A dynamic rerouting model for air traffic flow management. *Transp. Res. Methodol.* 43 (1), 159–171.
- Mukherjee, A., Hansen, M., Grabbe, S., 2012. Ground delay program planning under uncertainty in airport capacity. *Transp. Plan. Technol.* 35 (6), 611–628.
- O'Flynn, S., 2016. Airport capacity assessment methodology, ACAM Manual.
- Ostwald, P. et al., 2006. The miles-in-trail impact assessment capability. In: *6th AIAA Aviation Technology, Integration and Operations Conference (ATIO)*, p. 7780.
- Ribeiro, V.F. et al., 2018. Conflict Detection and Resolution with Local Search Algorithms for 4D-Navigation in ATM. In: *International Conference on Intelligent Systems Design and Applications*, pp. 129–139.
- Vossen, T., Ball, M., 2006. Optimization and mediated bartering models for ground delay programs. *Naval Res. Logist.* 53 (1), 75–90.
- Xu, Y., Prats, X., 2017. Including linear holding in air traffic flow management for flexible delay handling. *J. Air Transp. Am. Instit. Aeronaut. Astronaut.* 25 (4), 123–137.
- Zelinski, S., Lai, C.F., 2011. Comparing methods for dynamic airspace configuration. In: *2011 IEEE/AIAA 30th Digital Avionics Systems Conference*, pp. 3A1–1.

Further Reading

- Hopkin, V.D., 1995. Situational awareness in air traffic control. In: *Situational awareness in complex systems: Proceedings of a CAHFA conference*, pp. 171–178.

Port Management

Maria G. Burns, BTI, a DHS Center of Excellence, University of Houston, Spring, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Port Management: An Overview	390
The Significance of Maritime Trade	390
Port Management Encompasses Strategy and Operations	390
Port Ownership Styles	391
Port Efficiency as a Central Theme of Port Management	393
The Key Drivers for Successful Port Development Decisions	393
Conclusions	394
Acknowledgment	395
Biography	395
References	395

Port Management: An Overview

Port management is defined as the strategic decision making, and the subsequent monitoring and controlling of ports and terminals. It encompasses the management of port facilities, including infrastructures, superstructures, and substructures, which may be privately or publicly owned. Ports serve a significant role in the global economy, trade, transport, but also global and national security: Their role is to facilitate the global flows of trade and passengers, thus promoting the economy. They help global and national security by safeguarding our borders, and thus prevent illegitimate activities (Burns, 2014a).

Ports are the convergence points of supply chains that facilitate ships' loading and unloading operations, but also offer them a safe shelter during adverse circumstances, for example, extreme weather conditions, etc. They are the nodal points that facilitate the global movement of people and cargoes.

Throughout the history of humankind, ports have been reflecting the level of economic prosperity, technological innovation, and progress. Developed economies demonstrate their dominance by expanding seaports. In fact, the rise and fall of economic empires can be predicted by observing the growth of national seaports, and trade activities.

The Significance of Maritime Trade

Although global transportation consists of multimodal nodal points (i.e., ports for sea, land, and air), seaports hold a prominent position for many reasons. First, they represent about 80% of global trade and transport. At a global level, there are about 9,000 large and medium sized global ports, and about 100,000 ships, among which about 50% are ocean-going (i.e., 3,000 deadweight or larger), and 50% are smaller ships involved in short-sea shipping. Second, two-thirds of planet earth are covered by water. Third, the ship size is much larger compared to the land and air transportation modes: the length of modern ships is comparable to the height of the Empire State building, that is, exceeding 1,000 feet. The volume of cargoes a large ship can carry, could fit in over 230 Olympic swimming pools. Fourth, maritime trade is the most cost effective transportation mode; the longer the voyage, the lower the cost per unit is, since ships benefit greatly from economies of scale (Burns, 2014b).

The global maritime industry is highly competitive, and therefore port management requires a multitude of skills for the management of different departments and tasks, encompassing all executive and operational activities. Port managers need to employ all the factors of production including labor, land, capital, and everything that capital can buy: immobile assets such as new terminals, mobile assets, such as technologies, etc. High-level port management entails strategic decision making regarding the market needs, competition, investment, and others.

Port Management Encompasses Strategy and Operations

In our modern, fast paced global supply chains, port managers have become a hybrid of strategists and operators, due to the increasingly industry, with a plethora of critical decisions that need to be made. As seen in Fig. 1, port managers are in charge of business development and investment decisions such as port expansion, etc. They negotiate contracts oversee concession agreements between industry stakeholders and port authorities. They are in charge of asset management, and thus select banking institutions and financing terms and conditions. They decide to which extent the port should invest in new terminals and new storage areas. They are responsible for asset management, port operations, cargo loading and discharging, cargo handling, etc. They conduct cost and

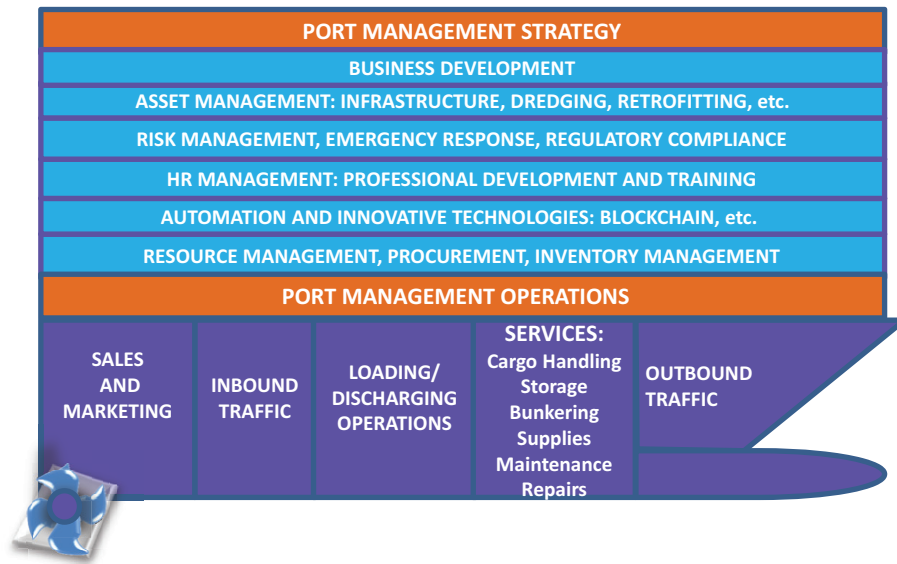


Figure 1 Port management strategy and operations. *Source: By Prof. Maria Burns.*

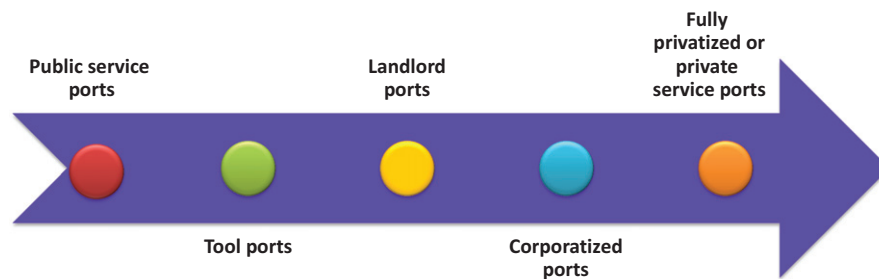


Figure 2 Port ownership models from public service to full privatization. *Source: By Prof. Maria Burns.*

benefit analyses to decide how much dredging is enough (i.e., the deepening and widening of the sea bottom, by removing debris and sediments). They must decide on the port's level of automation, while training their personnel, to upgrade their skills. They need to recruit, train, and supervise thousands of port workers, such as stevedores (longshoremen), terminal operators, engineers, crane operators, maintenance and repair teams, cyber and physical security professionals, transportation planners, and various other specialists (Burns, 2015).

Port Ownership Styles

Port ownership involves decisions on the land, infrastructure, superstructures, and substructures, and any management and operational activity involving the terminals, cargo handling equipment, and various other services.

Until the 1960s, government-centric strategies and management decisions about ports. After the 1960s, especially due to the globalization, port decisions were made at a state level, as only the local authorities had a better understanding of the strength, weaknesses, challenges, and opportunities. Gradually, the control at a public level was shifted with the public sector becoming more prominent.

There are five key models of port management, classified in accordance to the ownership by the public or private sectors. As depicted in Fig. 2, the two extreme strategies involve full privatization or total control by the public sector.

1. **Public service ports:** Typically the federal or state government owns these ports, and appoint port officials in key management positions. The port authority owns, controls, and operates equipment, and offers a wide array of services. Port hires the labor in charge of cargo operations. Cargo handling is a service typically offered by service ports, yet performed by independent companies, which are managed by the port authorities. This arrangement has been prevailing in developing and rapidly developing economies.

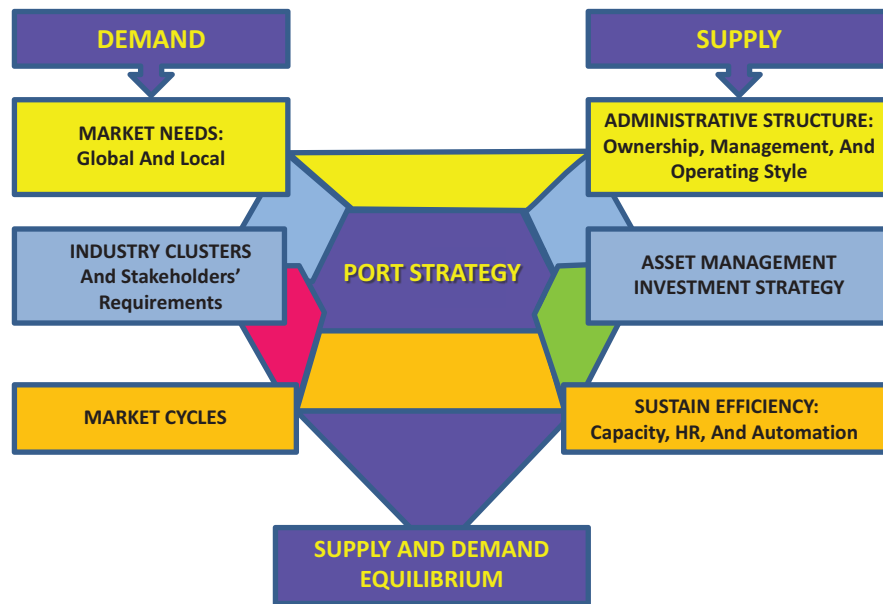


Figure 3 The port management diamond model. Attaining supply & demand equilibrium. *Source: By Prof. Maria Burns.*

Many developing economies transform their ports from a service port model into a landlord model. The public sector undertakes cargo handling operations, pilotage, storage, and other services.

The case of China

The case of China, one of the global economic leaders, has two interesting practices worth sharing: While in most developed countries ports are no longer controlled by the central government, China's central government takes particular interest in its ports, recognizing their role in the nation's prosperity. China's Five-Year Plans, that is, 5-year development strategies, often name the ports and regions to materialize specific national goals. The Chinese government has divided its port authorities into five coastal regions, which correspond to nine supply chain networks, in accordance to nine main cargo types. The Chinese government works closely with these ports' municipalities, to set goals, allocate resources, decide on Free Trade Zones and industrial zones, shipbuilding activities, trade routes, and inland interconnectivity.

Therefore, the fact that Chinese ports of Shanghai, Shenzhen, Qingdao, Tianjin, and others are ranked among the world's top ports, is neither a lucky break, nor a coincidence.

This impressive national strategy makes Chinese ports unique.

2. **Tool ports:** Typically, the port authority owns, develops, and manages the port's land, infrastructure, and most assets. The private sector -appointed or approved by the port authority- undertakes the cargo handling operations, whereas the port authorities still own the equipment. This model is an attractive choice for port authorities seeking to benefit from public-private partnerships (PPP), as they can use initial capital investment to expand their market share, while retaining up the public sector ownership. This simple and fast model is ideal for ambitious port expansion plans, and is typically adopted by smaller and medium ports, or developing regions.

Many ports are shifting from this model toward a service model, for example, by allowing shipping companies and contractors to be more involved in the cargo handling operations and other services.

As seen in Fig. 3, The Tool Ports model is a hybrid between service and landlord ports, whereas the semi-privatized ports are hybrids between tool ports and Fully Privatized ports. There is a need to make a distinction between service and tool ports, as over the past years these types have been overlapping, especially in terms of financing and public control.

3. **Landlord ports:** These are characterized by a public-private coalition, where the port authority provides the legal and administrative foundation, while private companies invest in cargo handling equipment, and other facilities (Burns, 2018a). Private companies have the option to lease land, terminals, storage areas, on a short or longer-term basis. Thus, concession agreements may have a duration of several decades, with many energy, chemical and other company types leasing for 50 or 100 years (Burns, 2018b). This model is very common for mega-ports and ambitious middle-sized ports. It is worth observing that the long-term concession agreements provide the port with financial and commercial stability, whereas the partnering companies contribute to the regional economic growth, through clustering activities. These companies act as money magnets, as their dynamic supply chains consisting of thousands of suppliers and buyers, divert a part of their business near the hub port. The ports of New York/New Jersey, Houston, and Antwerp, among others, fall into this category.
4. **Corporatized ports:** This is a popular choice for nations not wishing to fully privatize their ports. A hybrid between a landlord and a privatized port model, the corporatized option allows a government-owned, autonomous arrangement, while the private sector manages terminal operations. The public sector remains a major stakeholder, while the shareholders make administrative

and strategic decisions. Although airports have successfully implemented this model, the maritime industry has not advanced as much. Most nations implement public administration protocols to manage their ports. On one hand, the private sector provides insight to strategic decisions, as they are industry-savvy and adaptable enough to meet the short-term and longer-term market needs. In addition, the affluent private sector can always provide financial support to seaports, by enabling access to capital markets, have the commercial liberty to regulate prices. Investments will be driven by market demand. In addition, operating costs can be reduced through the elimination of bureaucracy, a better understanding of the typical market performance (input/output). Port authorities that have adopted this model include ports in the Netherlands, Australia, etc.

5. **Fully privatized or private service ports:** This model reflects the ultimate port privatization, with the port's ownership and policies being fully transferred from the public to the private sector. Agreements stipulate which, if any, functions are still retained by the public sector. Ports in the United Kingdom, Australia, and New Zealand, have selected this ownership style. Despite the commercial benefits of this option, it may not adequately protect a nation's borders in case of war, and may impose increasing security risks. Contractual agreements between the port's government and the private owners, should stipulate both parties' duties in case of a future security threat, while defining each parties' role in issues of national security.

Port managers and maritime professionals must keep in mind certain strategic considerations:

- A port's ownership style is closely related to its nation's political regime, its geopolitical role (e.g., warfare, tensions with neighboring countries), and its economic development level. For example, the Chinese central government controls all national ports and terminals in a Public Service arrangement. On the other hand, privatization has been adopted by countries like England, Australia, and New Zealand, among others. These nations' geographical location, combined with lack of territorial disputes at the time of privatization, encouraged their governments to privatize.
- The golden means here entails the Landlord ports style: this is the most common style, and is preferred by many advanced and emerging economies as it combines the best of both worlds.
- The global market conditions, and stakeholders' pressures or preferences, may instigate the change from one ownership style to another.
- While some decisions are irrevocable (e.g., once a port is privatized, it is not likely to be returned to the public sector), it is common for port authorities to change ownership type, that is, shift from a tool port to a landlord port.
- Modern ports combine private and public functions. They adopt economic multiplier strategies to maximize efficiency. Successful performance derives from increase of trade volume, productivity in cargo handling, and other services. Often the role of public and private sectors is not well defined in the contractual stipulations, thus generating strategic and operational complications.

Port Efficiency as a Central Theme of Port Management

The 21st century is an era of unprecedented changes in the global landscape, introducing geopolitical, technological, regulatory, banking, and operational changes. These radical changes have dictated the need of revisiting the role of port management. Globalization has increased port competition, forcing managers to intensify their business development strategies. Their goal is to allure their stakeholders, be it financial institutions, shippers, supply chain partners, or the community.

As depicted in Fig. 3, port strategy seeks to attain supply and demand equilibrium, by evaluating the demand, and adopting a customer-based supply strategy.

The Key Drivers for Successful Port Development Decisions

To make the best possible strategic decisions, an optimum port strategy should focus on the following drivers:

1. **Market needs:** To satisfy the global market and location-based market needs, ports need to fine-tune the ownership, management, and operating styles. Changes in the port's structure may involve public/private ownership, management, and operating arrangements;
 - a. **Key players**, namely the new versus existing key players, including both potential partners and competitors;
 - b. **Geographical trends**, that is, the think global and act global notion is required, that is, the examination of regional versus global supply chain trends and development;
2. **Growth and industrial clusters:** Clusters of economic growth and industrial activities are created in port-cities as a result of:
 - a. Policies aiming at regional development.
 - b. When several companies, part of the same industry or supply chain, decide to locate in a particular port-city, to reap specific benefits (e.g., specialization, technology, infrastructure, vicinity to exporters, producers, buyers, etc).

The presence of a cluster in a port is extremely beneficial, for all entities involved. It ensures sustainable development, and mutual growth through parallel activities. It enhances the port's growth, and prolongs global prominence. What many mega-ports have in common is the presence of industrial, financial, or other production clusters.

As specific industry clusters are created in the region, or within the port's global supply chain, the ports' stakeholders have specific requirements, for example, new equipment, terminal expansion, increased security, etc. In addition to procedural and operational improvements, stakeholders' requirements may involve asset management and new investments.

3. **Market cycles:** Business cycles reflect the supply-demand equilibrium at any given stage of the economy. There is a need to adopt strategic decisions to market cycles, accurate forecasting, and being proactive in view of market fluctuations. Market cycles are supply-demand driven fluctuations, driven by a myriad of factors, including geopolitical tensions, trade agreements, new regulations or technologies, and others. While ports cannot always leverage market demand, or change the trends of market cycles, they can always adopt the supply side of the equation by sustaining efficiency through optimizing capacity. This includes decisions on the optimum size, configuration, and decisions on land acquisition, number of terminals, and types of ships to be hosted. An efficient and well-organized port enjoys an optimum balance between automation and labor.

4. **Administrative structure: Ownership, management, and operating style:** Administrative decisions include land acquisition, port and terminal configuration, investment, maintenance and operations of the port's infrastructure, superstructures, and substructures, as well as activities involving terminals, cargo handling equipment, storage, supplying of bunkers, maintenance and repairs, and various other services.

As discussed in the port ownership section, high-level strategic decisions will seek to strengthen the port's global role, at a commercial, economic, and geopolitical level.

Interconnectivity and holistic infrastructure: decisions is another administrative responsibility. Balancing port development with the regional infrastructures, ensuring the port's interconnectivity. Port development and investment can only be commercially successful in the presence of efficient supply chain networks.

5. **Port development: Asset management and investment strategies:** Port development is an ongoing commitment evolving around asset management and port investment strategies. An efficient and effective port serves as an enhancer of the regional or national economy. Once a single entity, that is, a port commits to a significant investment, it attracts other investments that overlap or align with the cargo handling operations, supply chain, transport, etc. Port development plans are always an attractive marketing tool, which give a strong message to our global markets as to the port's successful past performance, leading to an ongoing growth. Thorough investment decisions are required, as port development involves high-risk investments, often leading to a questionable return-on-investment.

Risks of investment: The three risks port managers should consider are:

- a. **Overinvestment**, that is, being overambitious about the regional market's potential to utilize the port's new development facilities;
 - b. **Underinvestment**, that is, inadequately expanding port development plans, and missing potential opportunities for growth and commercial success, and
 - c. **Mis-investment**, that is, investing in the wrong industry segment, wrong technology, etc.
6. **Port capacity and throughput:** Ports facilitate the movement of people and cargoes from the sea to inland transportation up to the cargoes' final destination. Port throughput is typically measured by combining the number of vessel calls and cargo volumes handled within a specific time, that is, on a monthly or annual basis. Namely, port capacity is measured in terms of:
 - a. Port size, measured in acres, number of terminals, number of berths per terminal, draft, etc.;
 - b. Number of vessel calls and fleet tonnage handled per year; container ships are measured by TEUs (20-foot-equivalent unit);
 - c. Annual cargo volume handled, measured in metric tons for most commodity types, and TEUs for containerized cargoes;
 7. **Automation and technology, versus labor decisions:** Automation is the new buzz word, asking port managers to consider the soft spot between becoming fully automated (at the expense of thousands of displaced workers), or retaining their core manpower. This is one of the toughest decisions modern port managers need to make, as this is a transition stage in automation, where some technologies may not always deliver as promised, whereas port employees may be inadequately prepared to embed high-tech solutions into their existing working protocols. Personnel training is also a key impact factor.
 8. **Technologies: new cyber platforms: ACE, Blockchain technology, etc.:** Without a doubt, blockchain platforms, single window Automated Commercial Environment (ACE), and paperless trade is the new norm for modern ports. Cyber space capabilities now bring a new dimension in conducting global business (Burns, 2019a).
 9. **Regulatory compliance, and embracing the new HSQE regulatory framework**, that is, regarding Health, Safety, Security, Quality, and the Environment. The maritime industry is highly regulated, and the decades from the 2020s through the 2050s are becoming increasingly regulated, especially in terms of Environmental integrity. Modern ports must embrace these changes.
 10. **IMO2020, IMO2050, greener fuels, and the new bunkering era:** Modern ports must invest in low Sulfur fuel, in compliance with the IMO2020 regulations. LNG fuels, and the new generation of fuels are the new "business as usual," and modern ports need to proactively invest in collaboration with energy majors and shipowners (Reuters, 2019; Burns, 2019b).

Conclusions

This is an exciting era for maritime transportation professionals. Modern port managers need to rapidly adapt into this new global environment. Innovative technological advancements, and the creation of new supply chains, are shaping a new global setting for decades to come.

To retain their market share and global/regional prominence, modern ports have to keep abreast of market developments, technologies, and implement efficiency strategies. Port efficiency and expansion do not only involve high levels of investment, but also entail high risk, in a global setting where “business as usual” is no longer the norm. We must adopt to the new era of groundbreaking changes, involving both tremendous opportunities and unique challenges.

Acknowledgment

The author wishes to acknowledge the support and collaboration of the editors and publishers of the Encyclopedia, for making this work possible.

Biography

Professor Maria Burns is a National Academies scholar, the Director of the Logistics and Transportation Policy Program, University of Houston, and an Honorary Member of the US Coast Guard Auxiliary. Her graduate studies entail International Trade and Transport (London Metropolitan University) and Political Science (University of Houston). A most productive security researcher and data scientist, she has developed numerous research and training manuals for the Government and the energy and maritime industries. Since 2014 she is serving as a lead researcher (Principal Investigator) in Department of Homeland Security Research and Training Grants. She has authored the books “Port Management & Operations” (2014); “Logistics & Transportation Security” (2015), and “Managing Energy Security” (2019). She has obtained several prestigious awards including the “Excellence in Emergency Management Award,” and Membership in the Private Sector Advisory Council (PSAC). She is a Certified Maritime Auditor for Security (ISPS), Safety (ISM), Quality (ISO9001), and the Environment (ISO14001).

References

- Burns, M.G., 2019a. *Managing Energy Security. An All Hazards Approach to Critical Infrastructure*, first ed. Routledge. p. 446.
- Burns, M.G., 2019b. IMO2020, Addressing Emissions Reduction by Shipping—The 2020 0.5% Global Sulfur Cap. Annual Mare Forum Conference, Houston, November 2019.
- Burns, M.G., 2018a. Participatory operational and security Assessment on homeland security risks: an empirical research method for improving security beyond the borders through public/private partnerships. *J. Transp. Secur.* 11 (3–4), 85–100, Online Journal published in July 2018 & Special Issue in December 2018.
- Burns, M.G., 2018b. Containerized Cargo Security at the U.S.—Mexico Border: How supply chain vulnerabilities impact processing times at land Ports of Entry. Online Journal published in December 2018. Special Issue in December 2018.
- Burns, M.G., 2015. *Logistics and Transportation Security: A Strategic, Tactical, and Operational Guide to Resilience*, first ed. CRC Press. p. 380.
- Burns, M.G., 2014a. Estimating the impact of maritime security: financial tradeoffs between security and efficiency. *J. Transp. Secur.* 6 (4), 329–338.
- Burns, M.G., 2014b. *Port Management and Operations*, first ed. CRC Press.
- Reuters, 2019. Vessels will avoid areas lacking cleaner maritime fuel: Maria Burns quoted by Reuters. Mare Forum Panel. November 20, 2019. Available from: <https://www.reuters.com/article/us-shipping-imo-2020-fueloil/vessels-will-avoid-areas-lacking-cleaner-maritime-fuel-panel-idUSK>.

Port Performance Measurement from a Multistakeholder Perspective

Min-Ho Ha^{*}, Zaili Yang[†], Young-Joon Seo[‡], ^{*}Graduate School of Logistics, Incheon National University, Incheon, Republic of Korea; [†]Liverpool Logistics Offshore and Marine Research Institute (LOOM), Liverpool John Moores University, Liverpool, UK; [‡]School of Economics & Trade, Kyungpook National University, Daegu, Republic of Korea

© 2021 Elsevier Ltd. All rights reserved.

Introduction	396
A Conceptual Discussion on the Port Performance Measurement (PPM) Method	397
Port Performance Measurement Process and Methodology	397
The Conceptual Development of the Port Performance Indicators (PPIs)	398
The Use of the Hybrid Methodology for Port Performance Measurement in South Korea	400
The Use of DEMATEL and ANP to Analyze PPIs' Interdependent Weights	400
Port Performance Measurement in South Korea Using FER	400
Discussion and Conclusion	404
Acknowledgments	404
References	404

Introduction

Seaports are key nodes in global logistics networks and consequently contribute to the efficiencies of the associated global supply chains. Changes in supply chains have forced ports and terminals to seek effective integration in these supply chains when delivering value to shippers and third-party logistics service providers (Robinson, 2002). Ports are thus part of complex systems operating in an uncertain logistics environment. They are also places where a number of port stakeholders provide products and services and create value together. The interests of different port stakeholders, that is, port authorities, port users, service providers and related communities, in economic, social, and environmental issues, are sometimes in conflict (Notteboom and Winkelmans, 2003). Port managers increasingly rely on stakeholder relation management practices to secure long-term relations with key stakeholders (Dooms and Verbeke, 2007). To this end, port performance measurement (PPM) has become an important tool in stakeholder relation management and to achieve a sustainable competitive position.

Over the past three decades, PPM has become a well-established segment in port-related academic literature (see Pallis et al., 2011; Woo et al., 2012) but there are still research gaps to be filled within the context of the emerging development of port logistics. First of all, the existing literature tends to focus on limited dimensions of PPM or specific areas of ports. Such a fragmented approach may fail to take into account new issues and challenges faced by ports. The extant port literature mainly introduces lists of PPIs to measure the productive and allocative efficiency of port/terminal operations (i.e., operational efficiency), focusing more on terminal quayside operations via the application of data envelopment analysis (DEA) and stochastic frontier models (Cullinane et al., 2002; Tongzon, 1995). Compared to port efficiency studies, existing studies focusing on port effectiveness (i.e., Brooks and Schellinck, 2013) are mostly restricted to the dimension of customer satisfaction using qualitative PPIs (i.e., service effectiveness). In this regard, PPM should consider the different natures of PPIs. Using only quantitative PPIs is not sufficient to measure and diagnose performance (Beamon, 1999).

Secondly, there is little literature available on the development of a systematic approach to address the multi-stakeholder dimension in PPM. PPM demands a stakeholder-driven approach to cover the wide-ranging objectives and desired results of stakeholders. This can be achieved by integrating a multi-stakeholder dimension in a PPM framework which takes into account the corresponding port performance indicators (PPIs). These stakeholder-specific PPIs need to be aligned with organizational goals and strategies (Kaplan and Norton, 2004; Neely et al. 1995) and present a clear picture of the organizational performance (Gunasekaran et al., 2001). Moreover, the range of port activities that port stakeholders are concerned with, requires a focus on a multi-dimensional set of quantitative and qualitative PPIs. Using only one dimension (e.g., financial measures) in a performance measurement setting is no longer sufficient to cover all related issues for the new business environment (Miller and Vollmann, 1985). The importance of non-financial (i.e., intangible assets) measures is enhanced and the integral application of multi-dimensional measures (i.e., both financial and non-financial) for performance measurement have been continuously acclaimed (Neely et al., 1995).

In order to fill these research gaps, we present a hybrid approach of a fuzzy logic based evidential reasoning (FER) method (Yang and Xu, 2002), a decision-making trial and evaluation laboratory (DEMATEL) tool (Gabus and Fontela, 1973) and an analytic network process (ANP) technique (Saaty, 1996) to measure four major container ports in South Korea. It answers the related questions of "what to measure" and "how to measure" port performance. Such a PPM method is needed not just to meet the needs of port stakeholders, but also to enrich the diagnostic tools available to support decision-making in complex port/terminal systems operating in an uncertain environment. In the next section, the proposed port performance measurement (PPM) method and

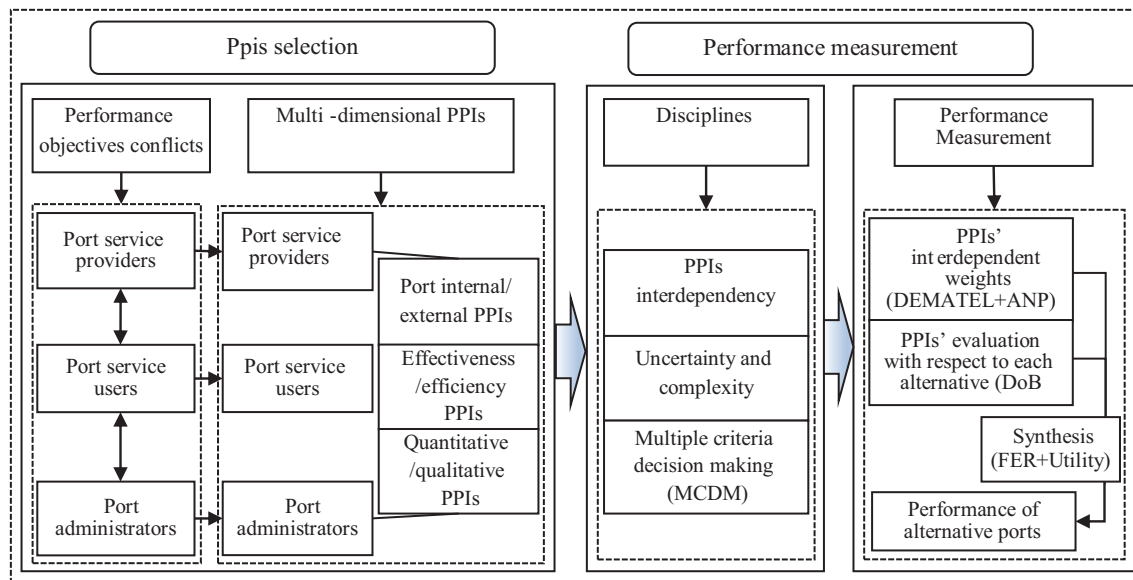


Figure 1 Port performance measurement (PPM) framework. Source: Ha et al. (2017)

theoretical justification to select PPIs are described in detail. In Section 3, an overview of the hybrid approach is applied to a case study of Korean container ports for practical insights. Finally, the chapter concludes with a discussion of the results the industrial and academic implications and recommendations for further research.

A Conceptual Discussion on the Port Performance Measurement (PPM) Method

Port Performance Measurement Process and Methodology

The objective of the proposed PPM framework is to identify most crucially needed PPIs for each group of port stakeholders and to develop a powerful performance measurement tool. In the framework, various disciplines such as uncertainty and interdependency among the PPIs are considered to deliver more practical applications in PPM. As seen in Fig. 1, the needs of different stakeholders were investigated in the first phase and their associated PPIs were derived in the second phase. To this end, we identified crucial interests in major (container) ports investigating stakeholders' goals and objectives, and discussed them with 10 port stakeholders.¹ Conflicts of interests between stakeholders require them to interpret others' assertiveness rightly. Consequently, the analysis of their interests and needs on various dimensions of port activities becomes essential. The six dimensions defined in this chapter cover the range of port activities to cope with new evolutionary changes, to measure and communicate their impacts on society, economy and environment and to be consistent with their goals. Then, through a literature review and an analysis of industrial practices the associated PPIs were identified and then confirmed by the panel of ten experts to assess the suitability of the potential indicators and to test the feasibility of the selected indicators (Bagozzi et al., 1991). The majority of the experts commented on all dimensions for content validation. Previous studies on port performance generally consider the PPIs as independent attributes. However, considering PPIs as independent and irrelevant to each other is not realistic any more to solve multiple criteria decision making (MCDM) problems in complex port activities and operations (Lee et al., 2013). In order to address this issue, we use DEMATEL to quantitatively identify whether there are interdependent relationships among the PPIs, while ANP is applied to determine the intensity of the relationships among the PPIs. Furthermore, FER is applied for dealing with uncertainties presented in the evaluations of the selected PPIs. In summary, the proposed method for a hybrid PPM system consists of four steps²:

1. Investigate the performance needs and interests of different stakeholders and translate them into corresponding PPIs for each group of stakeholders.
2. Find proper methodologies to deal with various features of the PPIs (i.e., interdependency, uncertainty and MCDM).
3. Assign PPIs' interdependent weight (using DEMATEL and ANP) and evaluate PPIs' performance (degrees of belief (DoB)) with respect to each alternative.
4. Synthesize the evaluations of PPIs with their weights using fuzzy rule-based ER algorithm and utility technique (IDS software).

¹The group included six industrial experts who have been working in shipping and port industry for more than 15 years with PhD (one expert from a shipping line), MSc (three experts from terminal operators, a shipping line and a forwarder), and BA (one from a terminal operator and a forwarder, respectively) degrees participated in the judgements. Two professors who have more than 15 years teaching and research experience as well as they are (ex)member of port committee participated in the survey. Lastly, two experts from government/port authority (one department manager and one managing director) who have been working for port logistics departments participated in the survey.

²The detailed description of the employed methods is discussed in Ha and Yang (2018).

The Conceptual Development of the Port Performance Indicators (PPIs)

Seaports are integrated process platforms where a number of port stakeholders interact in port activities related to cargos, vessels and other transport modes. Ports need an alignment of seaside, intermodal/multimodal, and landside logistics to achieve an efficient movement of the physical (i.e., cargos) and non-physical (i.e., information) flows (UNCTAD, 2004). To this end, PPIs in a PPM framework need to reflect these performance aspects. Moreover, PPIs evaluations need to be conducted with inputs from associated stakeholders. This may assist decision makers not only in diagnosing both the efficiency and effectiveness aspects of performance but also in identifying the strengths and weaknesses of ports. Based on this, six dimensions with their associated 16 principal-PPIs and 60 PPIs are defined as particularly relevant factors for port stakeholders. These dimensions include core activities (CA), supporting activities (SA), financial strength (FS), users' satisfaction (US), terminal supply chain integration (TSCI), and sustainable growth (SG) (Table 1). It is noteworthy that the discussions on each dimension are aimed at identifying PPIs for container terminals, focusing on their internal and external activities, but the PPIs are assessed by associated port stakeholders. Thus, the term container port performance refers to the performance of a collection of container terminals in port area.

The core activities (CA) relate to the core function of seaports, for example, vessel operation, cargo operation and other activities regarding container transfer or transit from ports to vessels and other transport modes (or vice versa) in container terminal area. The

Table 1 The hierarchy of port performance indicators (PPIs)

<i>Dimensions</i>	<i>Principal-PPIs</i>	<i>PPIs</i>	<i>Literature</i>	<i>Note^a</i>
Core activities (CA)	Output (OPC)	Throughput growth, vessel call size growth	UNCTAD, 1976; De Monie, 1987; Tongzon 1995; Cullinane et al., 2002; Brooks, 2006; Woo et al., 2011	QT; data input from TO, PA and GOV database
	Productivity (PDC)	Ship load rate, berth utilization, berth occupancy, crane productivity, yard utilization, labor productivity		
	Lead time (LTC)	Vessel turnaround, truck turnaround, container dwell time		
Supporting activities (SA)	Human capital (HCS)	Knowledge and skills, capabilities, training and education, commitment and loyalty	Marlow and Paixão Casaca, 2003; Kaplan and Norton 2004; Albadvi et al., 2007; Brown et al., 2011; Woo et al., 2013	QL; data input by TO
	Organization capital (OCS)	Culture, leadership, alignment, teamwork		
	Information capital (ICS)	IT systems, database, networks		
Financial strength (FS)	Profitability (PFF)	Revenue growth, operating profit margin, net profit margin	Su et al., 2003; Bitchou and Gray, 2004; Brooks, 2006	QT; data input from TO and GOV database
	Liquidity & solvency (LSF)	Current ratio, debt to total asset, debt to equity		
Users' satisfaction (US)	Service fulfilment (SFU)	Overall service reliability, responsiveness to special requests, accuracy of documents & information, incidence of cargo damage, incidence of service delay	Marlow and Paixão Casaca, 2003; Woo et al., 2011; Brooks and Schellinck, 2013	QL, data input by SL, FF and 3PL QL, data input by TO, SL, FF and 3PL
	Service costs (SCU)	Overall service cost, cargo handling charges, cost of terminal ancillary services		
Terminal supply chain integration (TSCI)	Intermodal transport systems (ITST)	Sea-side connectivity, land-side connectivity, reliability for multimodal operations, efficiency of multimodal operations	Bichou and Gray, 2004; Notteboom and Rodrigue, 2005; Panayides and Song, 2009; ESPO, 2010; Woo et al., 2013	
	Value-added services (VAST)	Facilities to add value to cargoes, service adaptation to customers, capacity to handle different types of cargo, tailored services to customers		
	Information/communication integration (ICIT)	Integrated EDI for communication, integrated IT to share data, collaborate with channel members for channel optimization, latest port IT systems		
Sustainable growth (SG)	Safety and security (SSS)	Identifying restricted areas and access control, formal safety and security training practices, adequate monitoring and threat awareness, safety and security officers and facilities	De Lagen, 2002; IMO, 2002; Peris-Mora et al., 2005; Darbra et al., 2009; ESPO 2010; Woo et al., 2011	QL, data input by TO, PA and GOV
	Environment (EVS)	Carbon footprint, water consumption, energy consumption, waste recycling, environment management programs		
	Social engagement (SES)	Employment, regional GDP, disclose of information		By PA and GOV

^aNote: FF, freight forwarder; GOV, government; PA, port authority; QL, qualitative PPI; QT, quantitative PPI; SL, shipping liner; TO, terminal operator.

Source: Ha et al. (2017)

CA have traditionally, and most frequently, been assessed by scholars and industry practitioners (Cullinane et al., 2002; Tongzon, 1995; UNCTAD, 1976) using different types of taxonomies, such as output, productivity, capacity utilization, and efficiency. Productivity is one of the most important criteria guiding port choice by shipping lines (Murphy et al., 1992). The term productivity refers to how efficiently resources (i.e., labor, equipment and land) are being used. The outputs generally considered include production, throughput and profit (Bichou, 2006). Output thus refers to the total quantity of work performed in a port over a period of time without considering the resources utilized (De Monie, 1987). The lead-time refers to the speed at which activities are performed. This term gained in importance by the introduction of just-in-time (JIT) production (De Treville et al., 2004), where it is defined as the time that elapses between the start of a process and its completion. Schmenner (2004) stressed that companies achieving a higher competitiveness through a combination of speed and variability reduction and productivity improvement would have a higher performance than when focusing on only one aspect.

Supporting activities (SA) refer to the maintenance of internal resources to improve an organization's effectiveness and/or efficiency. Kaplan and Norton (2004) stressed that desired strategic outcomes could be achieved by appropriate deployment and effective utilization of intangible assets in the information era. A value-oriented organization based on collaboration, trust, sharing, learning, and openness tends to achieve desirable outcomes such as efficiency, effectiveness, and innovation (Alavi et al., 2006). There is a need of reliable human resources (HRs) that cannot be easily imitated by competitors (Marlow and Paixão Casaca, 2003). Employees who have the right skills, talent and knowledge contribute to enhancing the organization's internal processes and performance (Kaplan and Norton, 2004). A higher worker commitment and loyalty leads to a higher workplace performance (Brown et al., 2011). Albadvi et al. (2007) found a statistically significant correlation between ICT and firm performance.

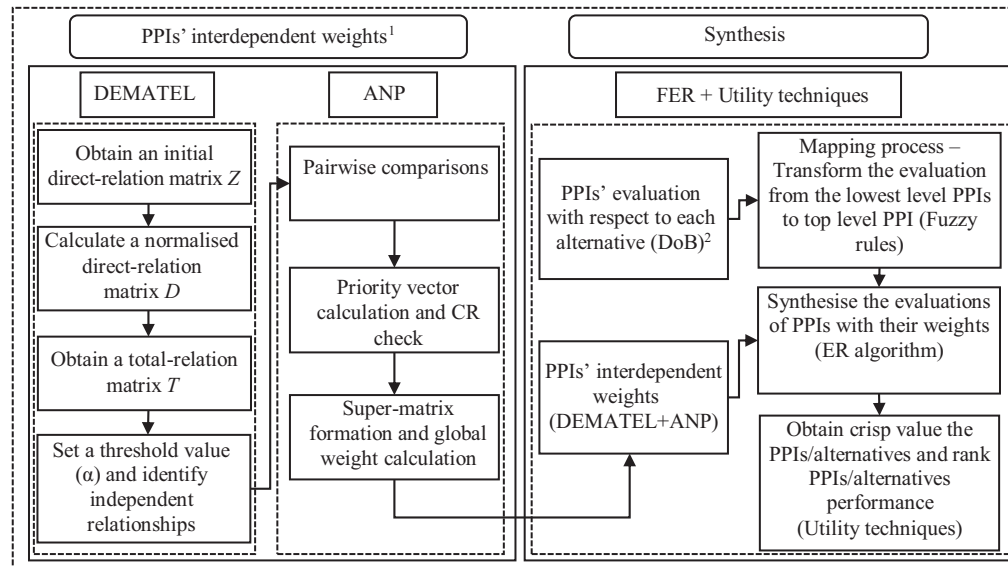
The financial strength (FS) dimension concerns financial profitability and stability. Profitability measures a firm's ability to generate profit relative to land, labor and capital inputs. Liquidity is to measure the firm's ability to pay its short liability whilst solvency covers its long-term liabilities. Irrespective of the type of industries, financial performance is very important for managers and investors. This is reflected by the many PPM approaches in literature. UNCTAD (1976) introduced revenue and cost items and classified major port cost items into labor costs, equipment costs and capital costs. Marlow and Paixão Casaca (2003) suggested the measures of cost items in the lean port process. Su et al. (2003) used profitability, solvency and return on investment as financial indicators for the balanced scorecard (BSC) based comprehensive performance measurement systems. Brooks (2006) identified specific revenue and cost items that are widely used by port operators in 42 ports located in ten different countries.

The aforementioned PPIs can be measured based on the information gained internally from terminal operators. Therefore, we take into account externally generated information to represent port users' stance for performance measurement. Previous literature mainly investigated whether a service quality delivered by ports meets port users' needs in terms of timing, quantity and quality (Brooks and Schellinck, 2013). In addition, we suggest to include an indicator to measure port agility, or the speed with which the port service provider responds to customers' special requests (Marlow and Paixão Casaca, 2003). These are underpinned by the growing number of studies using the SERVQUAL on service quality in the port industry (Pantouvakis et al., 2008). Woo et al. (2011) used various port service prices as a service quality measure. Service cost is considered as one of the most important criteria which affects port selection by port users and determines port competitiveness when service quality is ascertained (Yeo et al., 2014). Consequently, low port service charges are a key driver to attract customers (Woo et al., 2011).

Ports have a key role to play in supply chains. In this context, higher integration and coordination between the players in supply chains lead to a higher competitiveness (Panayides and Song, 2009). Seaport terminals should provide a reliable and adequate multimodal process such as sea/land side connectivity, multimodal transport integration in order to attract trade flows and increase a port's competitiveness (Woo et al., 2013). In addition, they should provide value-added services to better meet the objectives of the supply chain system (Panayides and Song, 2009). Furthermore, information & communication systems (ICS) integration cannot be excluded for TSCI; it measures the establishment and use of seamless communication systems and the degree of collaboration with other partners (Bichou and Gray, 2004; Notteboom and Rodrigue, 2005). Marlow and Paixão Casaca (2003) demonstrated that integrated IT systems would contribute to total cost reduction in supply chains.

Sustainability refers to the intersection of social, environmental and economic contributions that deliver long-term effectiveness for the natural environment, society and firms (Carter and Rogers, 2008). Despite the wide adoption of the term sustainability, the application of the concept to the maritime industry is of recent date (Lam, 2015). Due to legislations and the requirement to fulfil corporate social responsibility (CSR), ports are urged to reduce environmental impacts and to enhance safety, security and social and economic responsibility (ESPO, 2010). Hence, ports need to pay more attention to the promotion of long-term sustainable growth with ecological health and social and economic contributions. In the long term, port efficiency and competitiveness can be gained from the appropriate safety and security scheme (Woo et al., 2011) and monitoring and implementing environmental management systems (EMS) in maintaining port operations (Darbra et al., 2009; Peris-Mora et al., 2005). Furthermore, the contribution of ports to society and the economy is important in the context of CSR (De Langen, 2002; Grewal and Darlow, 2007).

Based on the insights presented earlier, there is a need for a high degree of excellence of modern ports in a multi-stakeholder perspective which can deliver internal and external satisfaction. The six dimensions for PPM are intertwined in practice. With regard to "employment" in SG, for example, an alternative port can be judged with a good performance on the "employment" PPI when the port has a huge contribution to create an employment opportunity or maximizes employment to fulfil SCR. However, the situation could simultaneously deteriorate the FS of the firm, leading to increased costs and can have an adverse effect on "labor productivity (throughput/number of employee)" in CA. In this regard, the mixed use of benefit and cost PPIs needs to be taken into account to represent interests of different port stakeholders in the context of port performance measurement.



Note: ¹The input data for the weight evaluations was obtained by the panel of expert group while ²the evaluations of PPIs' performance with respect to each container terminal were conducted using inputs gained from different stakeholders (i.e. questionnaire survey for qualitative PPIs; secondary data for quantitative PPIs)

Figure 2 A hybrid methodology for port performance measurement. Note: ¹The input data for the weight evaluations was obtained by the panel of expert group while ²the evaluations of PPIs' performance with respect to each container terminal were conducted using inputs gained from different stakeholders (i.e., questionnaire survey for qualitative PPIs; secondary data for quantitative PPIs). Source: Ha et al. (2017)

The Use of the Hybrid Methodology for Port Performance Measurement in South Korea

Based on the methodology (Fig. 1) in section "Port Performance Measurement Process and Methodology," this section uses an analysis of four major container ports in South Korea to demonstrate the detailed quantitative modelling for PPM using a hybrid approach of DEMATEL and ANP incorporating FER and utility technique, which is shown in Fig. 2. The case analysis is conducted to identify and prioritize their PPIs from multiple stakeholder perspectives. In order to remain competitive, these ports need to meet the requirements of (global) shippers, shipping lines and third-party logistics service providers while withstanding the competitive forces exerted by rival ports in the region. The methodology could equally be applied to other performance-driven port regions around the world. As seen in Table 1, the PPIs include various types of numeric and subjective data to reflect the complexity of port/terminal business environments. The data collection for the evaluation of all PPIs' performance, see Ha et al. (2017).

The Use of DEMATEL and ANP to Analyze PPIs' Interdependent Weights

An integrated method of DEMATEL and ANP has been proven to be a successful tool for measuring dependence and feedback among elements in complex decision problems in various applications (Buyukozkan and Cifci, 2012). In this section, the local weights of 60 PPIs obtained by AHP are shown in Table 2. Further computation has been conducted to obtain global weights of the bottom level PPIs by multiplying their local weights with the ones of their associated upper level criteria. For instance, the global weight of "throughput growth (OPC1)" can be obtained as 0.083 (=0.12 (the global of output (OPC)) × 0.696 (the local weight of throughput (OPC1))). The global weights of 60 PPIs at the bottom level are demonstrated in Fig. 3. The results derived from ANP suggest that throughput growth (OPC 1) is the most important PPI, which has a relative importance value of 0.083, followed by vessel turnaround (LTC 1, 0.071), crane productivity (PDC 4, 0.048), overall service reliability (SFU 1, 0.037), vessel call size growth (OPC 2, 0.036), and IT systems (ICS 1, 0.036). In contrast, waste recycling (EVS 2, 0.002), water consumption (EVS 4, 0.002), and carbon footprint (EVS 1, 0.002) under environment (EVS) are the least important PPIs. The top 10 rank PPIs in the ANP results include four PPIs under core activities (CA), three PPIs under supporting activities (SA), two PPIs under financial strength and one PPI under users' satisfaction (US). The global weights obtained are used as inputs for FER method.

Port Performance Measurement in South Korea Using FER

The synthesis of four ports against all PPIs together with their weights in this chapter was conducted by IDS software incorporating the ER algorithm and utility technique. The results derived from IDS are shown in Tables 3–5 and Fig. 4 with respect to different ranking criteria. The performance of individual PPIs with respect to the alternative ports is shown in Table 3. They provide direct information on analyzing performance of each port activity driven by each stakeholder, which makes port managers possible to interpret the performance results straightforwardly. For example, both the container throughput growth and vessel

Table 2 Local weights for 60 PPIs

PPIs	OPC1	OPC2	PDC1	PDC2	PDC3	PDC4	PDC5	PDC6	LTC1	LTC2	LTC3	HCS1	HCS2	HCS3	HCS4
LW	0.696	0.304	0.158	0.132	0.107	0.345	0.103	0.155	0.602	0.185	0.213	0.246	0.243	0.354	0.157
PPIs	OCS1	OCS2	OCS3	OCS4	ICS1	ICS2	ICS3	PFF1	PFF2	PFF3	LSF1	LSF2	LSF3	SFU1	SFU2
LW	0.175	0.296	0.198	0.330	0.364	0.301	0.335	0.318	0.328	0.354	0.342	0.349	0.309	0.361	0.147
PPIs	SFU3	SFU4	SFU5	SCU1	SCU2	SCU3	ITST1	ITST2	ITST3	ITST4	VAST1	VAST2	VAST3	VAST4	ICST1
LW	0.134	0.188	0.170	0.549	0.315	0.137	0.466	0.159	0.197	0.178	0.369	0.172	0.197	0.262	0.291
PPIs	ICST2	ICST3	ICST4	SSS1	SSS2	SSS3	SSS4	EVS1	EVS2	EVS3	EVS4	EVS5	SES1	SES2	SES3
LW	0.261	0.232	0.216	0.298	0.206	0.231	0.265	0.158	0.145	0.248	0.149	0.300	0.578	0.272	0.150

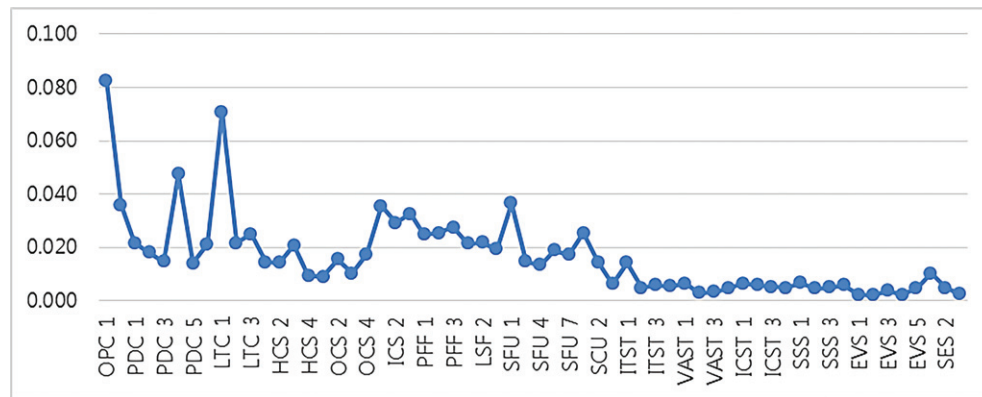
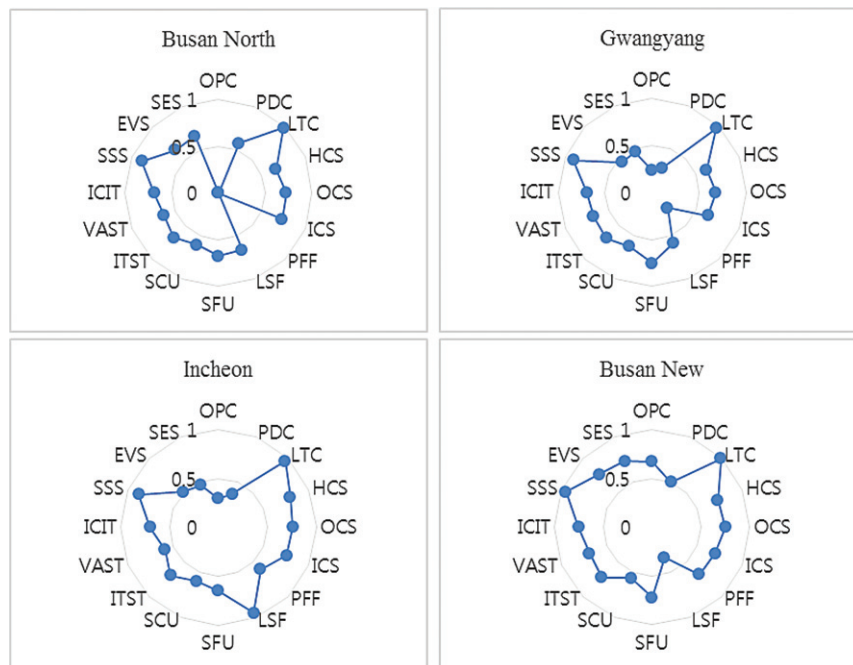
**Figure 3** Global weights of 60 PPIs.**Figure 4** Performance score on 16 principal-PPIs. Source: [Ha et al. \(2017\)](#)

Table 3 Performance score of each port against the 60 PPIs

<i>PPIs</i>	<i>Busan North</i>	<i>Gwangyang</i>	<i>Incheon</i>	<i>Busan New</i>
Throughput growth (OPC1)	0.0000	0.1630	0.3203	0.7317
Vessel call size growth (OPC2)	0.0000	0.6160	0.2436	0.3824
Ship load rate (PDC1)	0.8471	0.0000	0.0496	0.1817
Berth utilization (PDC2)	0.7040	0.1177	0.6042	0.9272
Berth occupancy (PDC3)	1.0000	0.0000	0.1055	0.0000
Crane productivity (PDC4)	0.4353	0.5359	0.5359	0.6797
Yard utilization (PDC5)	0.1055	0.0105	0.0000	0.6381
Labor productivity (PDC6)	0.5967	0.5000	0.5000	0.5823
Vessel turnaround (LTC1)	1.0000	1.0000	1.0000	1.0000
Truck turnaround (LTC2)	0.9114	0.8516	0.7656	1.0000
Container dwell time (LTC3)	0.9367	0.8301	0.9051	0.9525
Knowledge and skills (HCS1)	0.8270	0.7402	0.8609	0.8701
Capabilities (HCS2)	0.6531	0.6236	0.7736	0.6973
Training and education (HCS3)	0.5345	0.5397	0.7302	0.6136
Commitment and loyalty (HCS4)	0.6686	0.6205	0.7796	0.7365
Culture (OCS1)	0.6613	0.6186	0.7494	0.7517
Leadership (OCS2)	0.7294	0.6944	0.7615	0.7746
Alignment (OCS3)	0.7081	0.6978	0.7368	0.6959
Teamwork (OCS4)	0.7081	0.6415	0.7420	0.7289
IT systems (ICS1)	0.7828	0.6162	0.7615	0.7280
Database (ICS2)	0.6807	0.5957	0.7470	0.6620
Networks (ICS3)	0.6857	0.6849	0.7118	0.6531
Revenue growth (PFF1)	0.0000	0.7194	1.0000	1.0000
EBIT(operating profit) margin (PFF2)	0.0000	0.0000	0.4856	0.7201
Net profit margin (PFF3)	0.0000	0.0705	0.4209	0.3929
Current ratio (LSF1)	0.8047	1.0000	0.8047	0.8047
Debt to total asset (LSF2)	0.1953	0.1953	1.0000	0.1953
Debt to equity (LSF3)	1.0000	0.5000	1.0000	0.0000
Overall service reliability (SFU1)	0.6891	0.7457	0.6634	0.7084
Responsiveness to special requests (SFU2)	0.6247	0.7510	0.6355	0.6307
Accuracy of documents & information (SFU3)	0.6831	0.7134	0.6439	0.7518
Incidence of cargo damage (SFU4)	0.7152	0.7294	0.6136	0.7560
Incidence of service delay (SFU5)	0.5634	0.7299	0.6113	0.6720
Overall service cost (SCU1)	0.6060	0.6002	0.5876	0.5760
Cargo handling charges (SCU2)	0.5842	0.6476	0.6179	0.5395
Cost of terminal ancillary services (SCU3)	0.5702	0.6279	0.5687	0.5079
Sea-side connectivity (ITST1)	0.6526	0.6784	0.6960	0.7128
Land-side connectivity (ITST2)	0.6915	0.6426	0.6405	0.7049
Reliability for multimodal operations (ITST3)	0.7002	0.7078	0.6896	0.7299
Efficiency of multimodal operations (ITST4)	0.6741	0.6605	0.6473	0.7178
Facilities to add value to cargoes (VAST1)	0.6300	0.6015	0.5513	0.6470
Service adaptation to customers (VAST2)	0.6181	0.6820	0.5676	0.7310
Capacity to handle different types of cargo (VAST3)	0.6321	0.6981	0.6210	0.6863
Tailored services to customers (VAST4)	0.6031	0.7078	0.6365	0.6849
Integrated EDI for communication (ICIT1)	0.6963	0.7228	0.6870	0.7628
Integrated IT to share data (ICIT2)	0.6963	0.6744	0.6620	0.7389
Collaborate with Channel members for channel optimization (ICIT3)	0.6721	0.6513	0.7196	0.7218
Latest port IT systems (ICIT4)	0.6189	0.6576	0.6641	0.7181
Identifying restricted areas and access control (SSS1)	0.8841	0.8860	0.9344	0.9502
Formal safety and security training practices (SSS2)	0.8478	0.8791	0.7552	0.9178
Adequate monitoring and threat awareness (SSS3)	0.8486	0.8602	0.8494	0.9326
Safety and security officers and facilities (SSS4)	0.8915	0.9139	0.8941	0.9679
Carbon footprint (EVS1)	0.4269	0.3943	0.3785	0.6492
Water consumption (EVS2)	0.7086	0.4266	0.5124	0.8446
Energy consumption (EVS3)	0.8023	0.4661	0.5847	0.9207
Waste recycling (EVS4)	0.7228	0.5332	0.6471	0.7512
Environment management programs (EVS5)	0.5479	0.4779	0.4590	0.6308
Employment (SES1)	0.6552	0.4250	0.4329	0.7184
Regional GDP (SES2)	0.6915	0.5355	0.4803	0.8504
Disclose of information (SES3)	0.5473	0.6021	0.6655	0.4961

Source: *Ha et al. (2017)*

Table 4 Performance score on six dimensions

6 dimensions	Busan North	Gwangyang	Incheon	Busan New	Ranking
Core activities	0.5313	0.4716	0.5274	0.7146	BN > B > I > G
Supporting activities	0.7305	0.6601	0.7848	0.7306	I > BN > B > G
Financial strength	0.2432	0.3488	0.7707	0.5527	I > BN > G > B
User satisfaction	0.6667	0.7347	0.6458	0.6973	G > BN > B > I
Terminal supply chain integration	0.6822	0.6987	0.6858	0.7442	BN > G > I > B
Sustainable growth	0.7744	0.6633	0.6705	0.8580	BN > B > I > G

Source: Ha et al. (2017)

Table 5 Performance score of each port

Ports	Performance	Ranking index	Ranking
Busan North	VP 0.23; P 0.1; M 0.03; G 0.22; VG 0.42	0.61	4
Gwangyang	VP 0.21; P 0.14; M 0.03; G 0.21; VG 0.40; UK 0.01	0.61	3
Incheon	VP 0.11; P 0.14; M 0.04; G 0.22; VG 0.48; UK 0.01	0.70	2
Busan New	VP 0.10; P 0.11; M 0.04; G 0.25; VG 0.51	0.74	1

Note: (1) G, good; M, medium; P, poor; UK, unknown; VG, very good; VP, very poor.

(2) UK has arisen due to unavailable quantitative data.

Source: Ha et al. (2017)

call size growth PPIs in Busan North Port are shown as a negative growth. However, the ship load rate, a ratio of the combined two PPIs of container throughput volume (TEU) and an average vessel call size (GT), is performing well with a performance score of 0.8471. Even though the number of vessel calls to Busan North Port saw a relatively small decrease from 7702 in 2012 to 7386 in 2013 (−4.1%), the total gross tonnage (GT) of the vessels decreased radically from 136,448k GT to 113,405k GT (−16.9%). This indicates that smaller sized vessels came into Busan North Port in 2013 compared to the vessel size in 2012. Accordingly, container throughput decreased dramatically (−12.5%) but the decline was not as dramatic as the drop in vessel capacity calling the port (−16.9%). This leads to the remarkable performance result of the ship load rate in 2013 (81.69 TEU/GT). Moreover, the higher number of small sized vessel calls leads to a higher berth occupancy rate (the ratio of time that a vessel is occupying a berth, 1.0000) because the vessel berthing practices in general are conducted in terms of berth (identity) number regardless of berth capacity, but relatively lower the berth utilization performance (TEU/berth length, 0.704). Busan North Port performs moderately in other PPIs but shows a very poor performance on all profit PPIs. This is an expected result due to the poorest performance on the container throughput PPI that generates revenues for terminal operators. It might be noted from the result that terminal operators in Busan North Port are required to make an effort to create the coopetition strategy recommended by Song (2002) to face intensified port competition.

Beyond the individual PPI's performance, port managers can analyze performance at a higher level of abstraction (i.e., 16 principal PPIs and six dimensions). Fig. 4 demonstrates the performance scores of the 16 principal-PPIs, which is derived from the transformed values through the mapping process from the lowest level PPIs to their associated principal-PPIs, and the lowest level PPIs' weights. The results can lead to performance scores for the six dimensions (Table 4) and the overall performance results of each alternative port (Table 5). From the case study results the following conclusions can be drawn. The results suggest that Busan New Port shows the best results, followed by Incheon Port (Table 5). The difference is significant especially between the adjacent ports of Busan New Port and Busan North Port. Busan North Port is assessed to be the least competitive port with the lowest performance especially in terms of output and profitability (see Fig. 4). Busan New Port outperforms the other ports in terms of output, lead-time, profitability, intermodal transport systems, value-added services, information and communication integration, safety and security, environment and social engagement but is less competitive at the level of two principal-PPIs such as liquidity & solvency and service costs. This is because the five terminal operators in Busan New Port started up operations from 2005 to 2011 respectively. So there has been a rather recent heavy initial capital spending for port superstructure, state-of-the-art systems and equipment. The required capital is generally raised from financial institutions and investors through project finances. With regard to the service costs, the adjacent port, Busan North Port, lowered its service price to secure its market share from the moment Busan New Port started operations (based on interviews with terminal operators in Busan Port). The "lower price" strategy is the more preferential strategy when port operators adjust themselves in a changing business environment characterized by intense port competition. On the other hand, Incheon Port has its strengths in terms of human capital, organization capital, information capital and liquidity & solvency, accordingly in supporting activities and financial strength. Another striking feature of ports demonstrates that a very similar trend but a clear difference in performance score and ranking. For example, they show relatively poor performance results on output, productivity, environment, social engagement and profitability while they outperform in terms of lead-time and safety and security

(see Fig. 4). These results can provide a validation of the proposed methodology as the case ports are in pursuing similar objectives³ under a similar logistics environment (i.e., similar organizational structure, port governance, policy and economic condition, etc.). At the same time, they are also a weakness of this chapter as the results are only drawn from Korean port cases which indicates further empirical study in different regions/areas needs to be conducted. From the case study results it is possible to provide the strengths and weaknesses of the four ports. Accordingly, decision makers in the ports can identify the particular areas for improvement to enhance their competitiveness. These results provide an important contribution for decision makers to enhance their port performance based on any necessary comparisons. Furthermore, it can be used for a longitudinal study to investigate the improvement of ports within different timeframes.

Discussion and Conclusion

This chapter presented a PPM study with two new features, one is to take into account PPIs' interdependency while the other is to measure the interdependent PPIs in a quantitative manner by taking the perspectives from different port stakeholders. The result represents an effective performance measurement tool in complex port/terminal systems, which is validated through the case study of four major container ports in South Korea. This offers diagnostic instruments to decision makers in identifying the particular areas for improvement. In order to strengthen the research validation and its contribution to industry and academic research, we adopt the feedback from four senior managers in terminal operating companies located in each port. The development of the measurement instruments and the hybrid quantitative model to measure port performance is perceived as a major contribution. On the basis of the multiple objectives of key stakeholders, the PPIs crucially needed for measuring port performance were identified. They include both quantitative and qualitative items to offer diagnostic instruments to port managers, aiming to meet the different needs of port stakeholders. The identified PPIs within multidimensional conceptualizations characterize that (1) they are very specific measures to represent direct information on port activities; (2) they are very balanced measures to represent both internal/external aspects and efficiency/effectiveness aspects. According to the feedback from the senior managers, a set of quantitative PPIs have generally been utilized because the data can be readily available, or because the qualitative PPIs are too ambiguous to interpret them in a meaningful way. Accordingly, they cannot be used frequently together despite the need for performance measurement. The implication is that the hybrid approach can successfully deal with both quantitative and qualitative PPIs within a single framework. The results yielded by the hybrid approach provide for the ranking of the ports in terms of their overall performance with respect to multiple PPIs as well as a PPI's ranking with a single performance value. The interpretation of the results of the hybrid approach is very straightforward. This feature enables port managers to identify the strengths and weaknesses of the ports in terms of each individual PPI, principal-PPI and dimension, and offers insights into them to find the particular areas in need of special attention. Therefore, on the basis of the conclusions derived from the case study, this chapter provides port managers with valuable insights as this framework allows them (1) to recognize current strengths and weaknesses of each port; (2) to better understand the conditions and status of their competitive ports; (3) to prioritize investment to improve competitiveness and customers' satisfaction by focusing on the poor performing areas or by adjusting their strategies based on the relative importance of PPIs.

Acknowledgments

This work is based on the previous work by Ha, M.H., Yang, Z., Notteboom, T., Ng, A.K. and Heo, M.W., 2017. Revisiting port performance measurement: A hybrid multi-stakeholder framework for the modelling of port performance indicators. *Transportation Research Part E: Logistics and Transportation Review*, 103, pp. 1–16. The authors would like to thank publisher (Elsevier) for permission to reuse the materials of this work.

References

- Alavi, M., Kayworth, T.R., Leidner, D.E., 2006. An empirical examination of the influence of organizational culture on knowledge management practices. *J. Manage. Inform. Syst.* 22 (3), 191–224.
- Albadvi, A., Keramati, A., Razmi, J., 2007. Assessing the impact of information technology on firm performance considering the role of intervening variables: organizational infrastructures and business processes reengineering. *Int. J. Prod. Res.* 45 (12), 2697–2734.
- Bagozzi, R.P., Yi, Y., Phillips, L.W., 1991. Assessing construct validity in organizational research. *Admin. Sci. Q.* 36 (3), 421–458.
- Beamon, B.M., 1999. Measuring supply chain performance. *Int. J. Oper. Prod. Manage.* 19 (3), 275–292.
- Bichou, K., 2006. Review of port performance approaches and a supply chain framework to port performance benchmarking. *Res. Transport. Econ.* 17, 567–598.
- Bichou, K., Gray, R., 2004. A logistics and supply chain management approach to port performance measurement. *Maritime Policy Manage.* 31 (1), 47–67.
- Brooks, M.R., 2006. Issues in measuring port devolution program performance: a managerial perspective. *Res. Transport. Econ.* 17, 599–629.
- Brooks, M.R., Schellinck, T., 2013. Measuring port effectiveness in user service delivery: what really determines users' evaluations of port service delivery? *Res. Transport. Bus. Manage.* 8, 87–96.
- Brown, S., McHardy, J., McNabb, R., Taylor, K., 2011. Workplace performance, worker commitment, and loyalty. *J. Econ. Manage. Strategy* 20 (3), 925–955.

³The port governance of the case ports in Korea is located somewhere between the PRIVATE and the PRIVATE/public model, which is in pursuing the maximisation of the port profits or market shares (Cullinane et al., 2002).

- Buyukozkan, C., Cifci, G., 2012. A novel hybrid MCDM approach based on fuzzy DEMATEL, fuzzy ANP and fuzzy TOPSIS to evaluate green suppliers. *Expert Syst. Applic.* 39 (3), 3000–3011.
- Carter, C.R., Rogers, D.S., 2008. A framework of sustainable supply chain management: moving toward new theory. *Int. J. Phys. Distrib. Logist. Manage.* 38 (5), 360–387.
- Cullinane, K., Song, D.-W., Gray, R., 2002. A stochastic frontier model of the efficiency of major container terminals in Asia: assessing the influence of administrative and ownership structures. *Transport. Res. A* 36 (8), 743–762.
- Darbra, R.M., Pittam, N., Royston, K.A., Darbra, J.P., Journee, H., 2009. Survey on environmental monitoring requirements of European ports. *J. Environ. Manage.* 90 (3), 1396–1403.
- De Langen, P., 2002. Clustering and performance: the case of maritime clustering in The Netherlands. *Maritime Policy Manage.* 29 (3), 209–221.
- De Monie, G., 1987. Measuring and Evaluating Port Performance and Productivity. United Nations.
- De Treville, S., Shapiro, R.D., Hameri, A.-P., 2004. From supply chain to demand chain: the role of lead time reduction in improving demand chain performance. *J. Oper. Manage.* 21 (6), 613–627.
- Dooms, M., Verbeke, A., 2007. Stakeholder management in ports: a conceptual framework integrating insights from research in strategy, corporate social responsibility and port management. *International Association of Maritime Economists (IAME)*, July 4–6, Athens, Greece.
- ESPO, 2010. Port performance indicators selection and measurement work package 1: pre-selection of an initial set of indicators.
- Gabus, A., Fontela, E., 1973. Perceptions of the world problematique: communication procedure, communicating with those bearing collective responsibility. DEMATEL Report No. 1, Battelle Geneva Research Centre, Geneva, Switzerland.
- Grewal, D., Darlow, N.J., 2007. The business paradigm for corporate social reporting in the context of Australian seaports. *Maritime Econ. Logist.* 9 (2), 172–192.
- Gunasekaran, A., Patel, C., Tirtiroglu, E., 2001. Performance measures and metrics in a supply chain environment. *Int. J. Oper. Prod. Manage.* 21 (1/2), 71–87.
- Ha, M.H., Yang, Z., 2018. Modelling interdependency among attributes in MCDM: its application in port performance measurement. *Multi-Criteria Decision Making in Maritime Studies and Logistics*, Springer, Cham, pp. 323–354.
- Ha, M.H., Yang, Z., Notteboom, T., Ng, A.K., Heo, M.W., 2017. Revisiting port performance measurement: a hybrid multi-stakeholder framework for the modelling of port performance indicators. *Transport. Res. E: Logist. Transport. Rev.* 103, 1–16.
- IMO, 2002. ISPS code conference of contracting governments to the International Convention for the SOLAS 1974: SOLAS/ CONF.5/34. IMO, London, 17 December 2002.
- Kaplan, R.S., Norton, D.P., 2004. Measuring the strategic readiness of intangible assets. *Harvard Business Rev.* 82 (2), 52–63.
- Lam, J.S.L., 2015. Designing a sustainable maritime supply chain: a hybrid QFD–ANP approach. *Transport. Res. E* 78, 70–81.
- Lee, P.T.W., Wu, J.Z., Hu, K.C., Flynn, M., 2013. Applying analytic network process (ANP) to rank critical success factors of waterfront redevelopment. *Int. J. Shipping Transport Logist.* 5 (4), 390–411.
- Marlow, P.B., Paixão Casaca, A.C., 2003. Measuring lean ports performance. *Int. J. Transport Manage.* 1 (4), 189–202.
- Miller, J.G., Vollmann, T.E., 1985. The hidden factory. *Harvard Business Rev.* 63 (5), 142–150.
- Murphy, P., Daley, J., Dalenberg, D., 1992. Port selection criteria: an application of a transportation research framework. *Logist. Transport. Rev.* 28 (3), 237–255.
- Neely, A., Gregory, M., Platts, K., 1995. Performance measurement system design: a literature review and research agenda. *Int. J. Oper. Prod. Manage.* 15 (4), 80–116.
- Notteboom, T.E., Rodrigue, J.-P., 2005. Port regionalization: towards a new phase in port development. *Maritime Policy Manage.* 32 (3), 297–313.
- Notteboom, T., Winkelmann, W., 2003. Dealing with stakeholders in the port planning process. In: Dullaert, W., Jourquin, B. (Eds.), *Across the Border: Building Upon a Quarter of Century of Transport Research in the Benelux*. De Boeck, Antwerp, pp. 249–265.
- Pallis, A., Vitsounis, T., De Langen, P., Notteboom, T., 2011. Port economics, policy and management: content classification and survey. *Transport. Rev.* 31 (4), 445–471.
- Panayides, P.M., Song, D.-W., 2009. Port integration in global supply chains: measures and implications for maritime logistics. *Int. J. Logist.: Res. Applic.* 12 (2), 133–145.
- Pantouvakis, A., Chlomidis, C., Dimas, A., 2008. Testing the SERVQUAL scale in the passenger port industry: a confirmatory study. *Maritime Policy Manage.* 35 (5), 449–467.
- Peris-Mora, E., Orejas, J.D., Subirats, A., Ibáñez, S., Alvarez, P., 2005. Development of a system of indicators for sustainable port management. *Mar. Pollut. Bull.* 50 (12), 1649–1660.
- Robinson, R., 2002. Ports as elements in value-driven chain systems: the new paradigm. *Mar. Policy Manage.* 29 (3), 241–255.
- Saaty, T.L., 1996. *Decision Making with Dependence and Feedback, the Analytic Network Process*, RWS Publication, Pittsburgh.
- Schmenner, R.W., 2004. Service businesses and productivity. *Decis. Sci.* 35 (3), 333–347.
- Shieh, J.I., Wu, H.H., Huang, K.K., 2010. A DEMATEL method in identifying key success factors of hospital service quality. *Knowl.-Based Syst.* 23 (3), 277–282.
- Song, D.-W., 2002. Regional container port competition and co-operation: the case of Hong Kong and South China. *J. Transport Geogr.* 10 (2), 99–110.
- Su, Y., Liang, G.-S., Chin-Feng, L., Tsung-Yu, C., 2003. A study on integrated port performance comparison based on the concept of balanced scorecard. *J. Eastern Asia Soc. Transport. Stud.* 5, 609–624.
- Tongzon, J.L., 1995. Determinants of port performance and efficiency. *Transport. Res. A* 29 (3), 245–252.
- UNCTAD, 1976. *Port Performance Indicators*, TD/B/C.4/131/Supp.1/ Rev.1. UNCTAD, New York, USA.
- UNCTAD, 2004. *Assessment of a Seaport Land Interface: An Analytical Framework*. UNCTAD Secretariat.
- Woo, S.-H., Pettit, S., Beresford, A.K., 2011. Port evolution and performance in changing logistics environments. *Mar. Econ. Logist.* 13 (3), 250–277.
- Woo, S.-H., Pettit, S., Beresford, A., Kwak, D.-W., 2012. Seaport research: a decadal analysis of trends and themes since the 1980s. *Transport. Rev.* 32 (3), 351–377.
- Woo, S.-H., Pettit, S.J., Beresford, A.K., 2013. An assessment of the integration of seaports into supply chains using a structural equation model. *Supply Chain Manage.* 18 (3), 235–252.
- Yang, J.-B., 2001. Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties. *Eur. J. Oper. Res.* 131 (1), 31–61.
- Yang, J.-B., Xu, D.-L., 2002. On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty. *IEEE Trans. Syst. Man Cybernet. A: Syst. Hum.* 32 (3), 289–304.
- Yeo, G.-T., Ng, A.K., Lee, P.T.-W., Yang, Z., 2014. Modelling port choice in an uncertain environment. *Maritime Policy Manage.* 41 (3), 251–267.

Page left intentionally blank

Transport Modeling and Data Management

Chandra R. Bhat, Center for Transportation Research (CTR), The University of Texas at Austin, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

In recent years, we have witnessed tremendous technology progress in almost every aspect of our lives. Smartphones, GPS navigators, Bluetooth devices, tablets, and other wireless devices are becoming ubiquitous. Innovations in sensors, cameras, computers, wireless communications, and visualization techniques, such as virtual reality and augmented reality, now make it possible to gather, analyze, and project data forward (in near real-time and otherwise) in ways that have never before been possible earlier. At the same time, the quick evolution of technology poses interesting challenges by way of modeling (predicting) travel needs into the future as well as to understand travel behavior considerations in a fast-changing landscape. For example, autonomous vehicles (or driverless vehicles) are likely to emerge in the marketplace within the next few years, even though their adoption and use in mainstream transportation may take longer. Such vehicles can act as probes for a variety of transportation operations, and collect voluminous amounts of data that can be harnessed for effective transportation management. The key issue is how to deal with such voluminous and diverse amounts of incoming data per unit of time, and translate them into usable information for near-real time operations purposes or for long-term planning purposes. This is a challenge, given the data reliability required to translate data into actionable intelligence, especially for such safety applications as collision avoidance. In addition, predictive analytics to translate data into information requires the ability to deal with data that may be from multiple sources, highly noisy, heterogeneous, and high-dimensional with complex interdependencies. There is also a need for a data architecture to store such data and retrieve quickly. Thus, it is important that scholars are knowledgeable about the diversity of data sources, data-collection and integration approaches, data storage and retrieval, simulation and optimization techniques, mathematical, statistical, and econometric approaches, visualization techniques, and prediction methods/algorithms used in transportation analysis.

The articles in this section deal with a range of topics associated with transport data collection, data management, and transport modeling, including (1) methodology-oriented contributions (agent-based microsimulation methods, computational methods, spatial choice and general choice models, demand-supply feedback loops, use of Geographic Information Systems in data architectures and analysis, forecasting methods, multicriteria decision-making, and emerging new analytic tools), (2) topic-oriented contributions (trip departure and route choice, trip-chaining, vehicle ownership, shared mobility, recreational travel, commuting patterns, demand management and pavement demand strategies, parking demand, public transit systems, autonomous vehicle adoption/use, and ride-hailing patterns), and (3) data collection/software-related contributions (including discussions of a national travel survey database and transportation statistics/databases more generally, and a review of transport modeling/planning software). Many of the articles have cross-cutting themes with articles in other sections of the Encyclopedia, especially those dealing with Transport Planning and Policy. The collection of cross-cutting and inter-disciplinary contributions should be valuable to scholars focused on modeling and analysis, transport applications, data architecture design, and algorithm development for translating data into information in near-real time and planning, just to identify a few areas.

Transport Modeling and Data Management

Computational Methods and Data Analytics

Bilal Farooq*, David López†, *Ryerson University, Toronto, ON, Canada; †Torre de Ingeniería piso 2-N, Instituto de Ingeniería Av., México

© 2021 Elsevier Ltd. All rights reserved.

Computational Methods	408
Deterministic Computational Methods	408
Stochastic Computational Methods	409
Data Analytics	411
Descriptive Analytics	411
Diagnostic Analytics	411
Predictive Analytics	412
Prescriptive Analytics	413
Acknowledgment	413
References	413

Computational Methods

Computational methods in transportation involve developing and using mathematical models of a certain complex phenomenon or system, for example, flow of traffic at an exit ramp or individuals using public transit system. Transportation analysts can use these models in a computer program to numerically predict the behavior of the modeled system in “what-if” scenarios as well as optimize the system for various objectives under certain constraints. Computational methods used in transportation can broadly be classified into two categories: deterministic and stochastic (Möller, 2014). Fig. 1 shows a more detailed hierarchy of various types of computation methods employed in transportation.

Deterministic Computational Methods

Given a particular set of inputs, deterministic computation methods will always produce the same set of outputs. Typically, the deterministic methods are used to simplify the complexity of the problems, for easily finding the solution, or when the scenario we are trying to model is only being observed in a very controlled environment. In reality, phenomena in transportation systems are not purely deterministic, but some are close enough and can be approximated using deterministic computational methods. For example, train schedules on dedicated tracks are often considered deterministic since operators control the flow of trains and usually the demand is more or less stable, that is, trains are encapsulated in a controlled environment. If a train leaves stop A at 6:45 a.m., under normal operating conditions, it will arrive at B at 7:00 a.m., and this will happen most of the days during a year.

Deterministic computational methods are static if the phenomena we are solving do not change over time and are dynamic if the phenomena change over time. For instance, in the static equilibrium assignment, the flows between origin and destination are considered static (Ortuzar and Willumsen, 2011). This is not very realistic as it does not reflect the fluctuations of the vehicles on the roads over the day. However, for long-term planning of transportation infrastructure, this assumption may not introduce a very large error. The dynamic equilibrium assignment, on the other hand, is more suitable for representing how flow changes during the day and should be considered in the operational aspects of transportation system (Ortuzar and Willumsen, 2011). Static models are useful when the system we are trying to model does not change considerably over the observed period of time. On the contrary, if the system changes are significant over the considered period of time, it is better to use a dynamic model. In terms of computations, generally a dynamic deterministic computational method has higher complexity when compared with the equivalent static method. The variables for solving static deterministic models can be considered constant for the entire process or they may change in each step. Steady-state variables in deterministic static models are used when the modeled process remains unchanged. This type of models can be viewed as a photograph of a phenomenon at certain instant of time.

Time series deterministic computational method involves a sequence of points equally spaced over time and is somewhere in between dynamic and static, depending on how we layout the model. Time-series methods, in static form, are a set of possible initial states of our model such that each point in the series is evaluated in the model each time. In static user equilibrium with time-series variables, we assign a flow at each point in the series. Using this set of input flows, we then find an equilibrium for each point in time. On the other hand, time series in dynamic models is the evolution of the initial state over time where every point is dependent upon

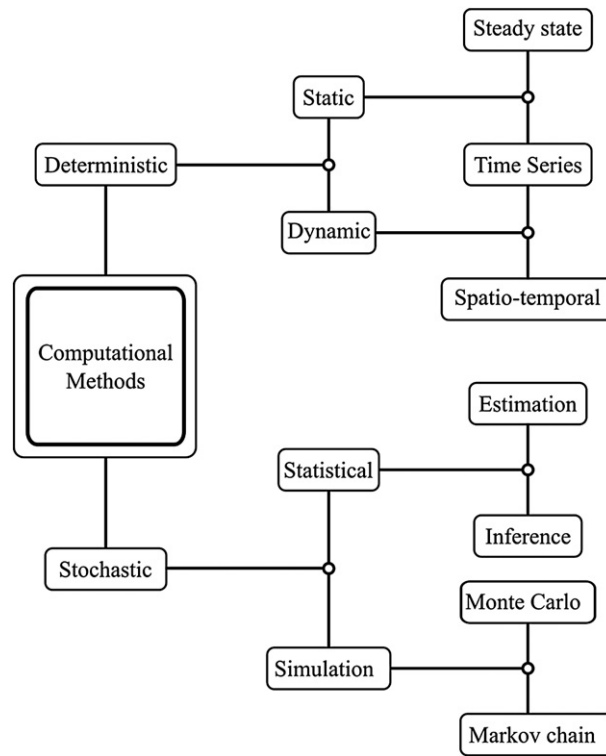


Figure 1 Computational methods used in transportation.

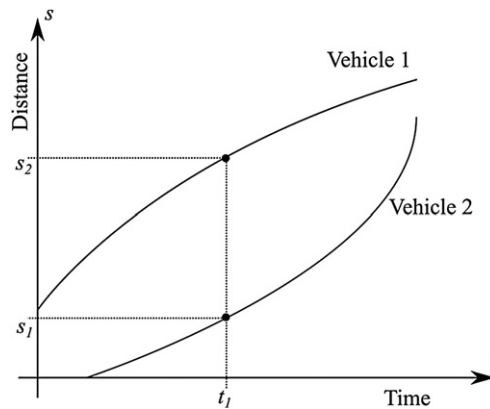


Figure 2 Time space diagram of two vehicles along a road segment.

the value of the previous point. In dynamic user equilibrium, we assign the vehicles arriving during certain time step to the start of the time step or randomly distribute them over the time step using certain distribution, for example, uniform or Poisson. The equilibrium is then computed over time, while accounting for the arrival sequence of the vehicles.

Spatio-temporal variables in dynamic deterministic computational methods reflect the changes of the variables (states) in space and time. Traffic flow models of a road segment are often modeled with a time-space diagram, which plots the movements of one or more vehicles in time along a given road segment. Fig. 2 shows a time-space diagram of two vehicles along a road segment, where a spatio-temporal variable is defined with a time-distance tuple (s, t) . Observe that at time t_1 , vehicle 1 has traveled s_1 meters of the road, while vehicle 2 has traveled s_2 meters, that is, vehicle 2 is closer to the end of the road segment.

Table 1 provides a list of classic deterministic computational methods commonly used in transportation.

Stochastic Computational Methods

While in reality, almost every aspect of transportation is stochastic, there are some cases that may involve higher randomness than others, for instance, the nonrecurrent congestion levels after a baseball game. Transportation analyst can model such phenomena

Table 1 Example of classical computation methods

<i>Deterministic</i>	<i>Static</i>	<i>Steady state</i>	<i>Dijkstra algorithm</i>
	Dynamic	Time series Time series Spatio-temporal	Deterministic Wardrop equilibrium Time-dependent shortest paths Dynamic traffic assignment Microscopic models Mesoscopic models Macroscopic models
Stochastic	Statistical	Estimation	Maximum likelihood Curve fitting Bayesian estimation Least squares Stochastic gradient descent
	Simulation	Monte Carlo Markov chain Monte Carlo	Antithetic draw Importance sampling Metropolis-Hasting sampler Gibbs sampler

using stochastic computational methods (Molugaram et al., 2017). Given certain state of the system, the best a transportation analyst can do is to predict the outcome by assigning a probability to it. With stochastic methods, the outcome of a state can be different even for the same set of inputs. To account for this randomness, the analyst usually evaluates the stochastic model several times for the same input and develops a mean outcome.

Let us assume a simple scenario where a person leaves home every day at 6:00 a.m. and drives to work. Here, leaving for work in a vehicle at 6:00 a.m. can be considered as the initial state. Given that the traffic on certain route is random and may vary because of weather conditions, accidents, maintenance work, number of vehicles on the route, etc., the best a transportation analyst could do is to estimate the arrival time for the person based on the initial state and approximation of the route conditions. No matter if a person leaves home every day at exactly 6:00 a.m., some days they will arrive at their office at 6:44 a.m. due to low traffic on the route, but then they could also arrive at 7:30 a.m. due to some accident on the route. Thus, the exact arrival time at the office cannot be predicted with certainty.

Stochastic computational methods can be solved by using statistical models or simulation techniques (Kutz, 2013). Statistical methods are parametric mathematical expressions, models, and techniques that extract relevant information from the raw observations in a concise form. The relevant information is usually extracted in the form of a set of parameters, for example, mode, scale, and average rate. It is assumed that the observed data are representative of the stochastic process. In the example above, let us assume that the person logs their arrival times for a whole year. By looking at the data, they observed that a probability distribution can be fitted such that the arrival time between 6:50 a.m. and 7:10 a.m. is 70%, while the probability of arriving later or earlier is 15% each. They also conclude that the mode for travel time is 1 h.

In general, the statistical model can be parametric, semiparametric, and nonparametric. In the case of a parametric model, a parametrized distribution is fitted to the observed data. For instance, the phenomenon of vehicles arriving at a signalized intersection is usually modeled by a Poisson process and by finding the headway distribution between successive vehicles. Nonparametric models are used where fitting a parametrized distribution to the data is not possible because of the high level of stochasticity. Here, methods like histogram, running average, and kernel fitting are used to estimate model. Semiparametric approaches are a combination of the previous two; part of the phenomena is modeled using parametrized distribution, while the rest is modeled using nonparametrized model.

A common parameter estimation method for statistical models is the maximum likelihood approach, where the algorithms search for a set of parameter values that maximizes the joint probability of reproducing the data using the proposed statistical model. If the search space is very complex and/or the available data size is too low, then Bayesian estimation techniques can be used to maximize the likelihood. In the case of very large amount of data, stochastic gradient descent method is used to find the optimal parameters.

Simulation methods are computer programs that abstract a real-life phenomenon using mathematical formulations and probability distributions (Ross, 1997). The idea of a simulation is to have a computer model, where transportation analysts can test different scenarios and assess their outcomes before the actual implementation, for example, studying the effects of constructing an additional lane on an existing highway. Simulation plays a useful role where the outcomes of a model are difficult to obtain with analytical expressions. Computer implementation of a simulation involves drawing from probability distributions. This is achieved by first generating a pseudo-random¹ number and then transforming it to a draw from the empirical probability distribution, which is part of the simulation. This process is known as Monte Carlo simulation. Suppose that we want to analyze the wait time at a four-

¹ Note that computers are discrete and have limiting memory, due to which generation of a true random number is not possible.

way intersection for different conditions. This phenomenon can be modeled as a queue with certain arrival time distribution for vehicles and service time distribution at the intersection for each direction. Suppose that the arrival rate is modeled as a Poisson distribution and service time as an exponential distribution. The parameter for the two distributions can be estimated from the historic data. Monte Carlo simulation can then be used to simulate the arrival time of every vehicle arriving at an intersection in each direction. Once a vehicle arrives and the intersection is empty, it will be serviced immediately and will pass through. The service time is represented by the value drawn from the exponential distribution. Service time may vary due to the initial speed of the vehicle, acceleration/deceleration, which direction the vehicle is going to, etc. In the case where there are other vehicles in the intersection, the arriving vehicle will join at the end of the queue and wait for its turn to cross the intersection. If we keep track of the time for each vehicle in the simulation, we can develop statistics like average and standard deviation of wait time. This simulation can also be used to test various what-if scenarios, for example, what will be the effect on wait time of North-bound traffic, if the East-bound and West-bound traffic increases by 30%?

There are various phenomena, where the underlying distribution is very complex and operationally it is not feasible to use Monte Carlo simulation for sampling from such distributions. Markov chain Monte Carlo (MCMC) methods are the alternative that can be used in such cases. MCMC methods use the properties of the Markov chain to indirectly sample from the underlying distribution. For instance, Gibbs sampling, a type of MCMC simulation, has been used to synthesize the population and their characteristics for use in the activity-based travel demand modeling system. It should be noted that MCMC methods involve a random walk in the search space, which could be computationally wasteful, but the properties of the Markov chain ensure that once the MCMC simulation reaches the steady state, it will be able to sample from the underlying distribution. Techniques like Hamiltonian dynamics are used to make the MCMC simulation more efficient.

Operationally, simulation methods can be classified into discrete event or discrete time simulations. In the case of discrete event simulation, the computer simulation updates its state only if an event happens. This means that the time between two consecutive state updates is not constant. For instance, in the discrete event version of intersection simulation, the state of the intersection will be updated only when a vehicle arrives. Whereas, in the case of discrete time simulation, the state will be updated every constant time interval. In the case of the intersection simulation, we will process all the vehicles arriving in an interval, for example, during 10 s, at the start of that interval. Discrete event simulation is used where the events are more randomly distributed over time and their occurrence rate is low to moderate. Discrete time event is suited for a more uniformly distributed event over time with high rate of occurrence.

Stochastic computational methods can also operate at different spatio-temporal scales, depending on the complexity of the phenomena, available computational resources, and the context of their usage. [Table 1](#) provides a list of classic stochastic computational methods commonly used in transportation.

Data Analytics

Data analytics involves examining the data observed on a complex phenomenon or system to understand, classify, characterize, and predict better. Such an analysis involves developing descriptive, diagnostic, predictive, and prescriptive models ([Chowdhury et al., 2017](#)). [Fig. 3](#) shows four main types of data analytics methods and tools available for analyzing transportation data. Both deterministic and stochastic computational methods are employed in developing these data analytics.

Descriptive Analytics

Such analytics studies the current state of a phenomenon and helps in identifying issues as well as the underlying distributions in the data. It helps as a transportation analyst in revealing the desirable and undesirable outcomes. With descriptive analytics, we can identify bottlenecks on the highways, share of the commuters using public transport, or the spatial distribution of cyclists in a city.

Transportation analyst uses classic statistical methods to analyze the historical data and provide qualitative and quantitative insights related to the transportation system. The most used techniques in transportation system analysis are tendency, distribution, and dispersion analysis. These types of analysis involve tables and graphical representations of the data that can help in the organization and understanding of the state of a system. Visualization techniques and tools like geographic information systems (GIS) are readily used in developing descriptive analytics.

Diagnostic Analytics

Diagnostic analytics aim to explain the circumstances that lead to a certain phenomenon. While descriptive analytics show the current state of the phenomenon, diagnostic analytics tries to explain *why* a certain phenomenon happens. The focus here is on finding the causation and relationship between different variables. Typical techniques used to develop such analysis include correlation analysis, pattern mining techniques like association rule mining, and multidimensional contingency table, also known as hypercubes, developed in a data warehouse.

Correlation analysis is typically applied where a small representative sample is periodically collected, for example, travel survey conducted by a municipality every 5 years. The focus here is on testing the hypothesis related to causations. Pattern mining techniques are applied on larger datasets with higher frequency, but having less information available on the actual traveler, for

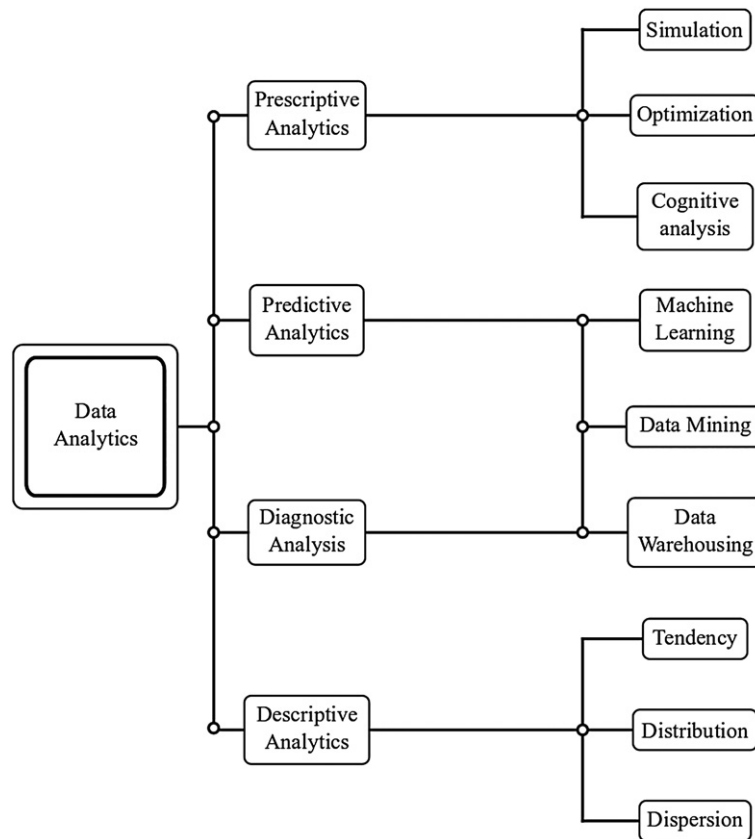


Figure 3 Data analytics used in transportation.

example, bike usage data in a bike sharing system. In such cases, other datasets are overlaid, for example, land use data are overlaid on bike usage data to mine and understand the spatial usage patterns of bike sharing system. Transportation agencies may also employ large data warehouses to host information coming from heterogeneous sources, for example, smartcards, loop detectors, travel survey, traffic cameras, etc. Hypercubes are developed over these datasets to conduct diagnostic analytics.

Predictive Analytics

Predictive analytics help a transportation analyst to foresee the future state of a phenomenon given certain future conditions, for example, how will the wait time at an intersection increase with an increase in the population in the neighborhood by 15%? Given the present state and the historical data with predictive analytics, it is possible to develop a generalization, for example, parametric mathematical expression that can be used to predict future conditions. Predictive analytics mainly use hypothesis-driven approaches, data-driven approaches, or a mix of both.

In the case of hypothesis-driven approach, theoretical foundations are employed to develop a mathematical relationship for the phenomena. The form of the mathematical relationship is a priori defined. Available data are then used to prove or disprove the causal relationship defined in the mathematical relationship. For instance, the behavior of a commuter in choosing mode can be expressed as a multinomial logit (MNL) model, which has a strong foundation in consumer behavior theory and econometrics. At core, the attractiveness of a choice is expressed in form of a linearly additive parametric function and an extreme value distribution that accounts for the errors. Particular hypotheses would result in an exact form of this parametric function—it may be hypothesized that with an increase in the travel time, a mode may not be as attractive to the commuter as other modes with lower travel times. Statistical estimation methods, for example, maximum likelihood method can be used to estimate the parameters and indicators identifying their significance (e.g., *t*-statistics) in the data. The estimated significance is used to judge if certain hypothesis holds or not. Once the model estimation is finalized, it can be applied using Monte Carlo simulation to give us predictions based on the future conditions. For instance, an MNL model estimated for mode choice using the travel survey data from a municipality can be used in Monte Carlo simulation to predict the change in the share of public transit if the fare is increased by 5%. Hypothesis-driven approaches are particularly useful in the situation where predictive models need to be developed using small datasets. They also provide a well-established foundation for economic analysis, for example, investigating the willingness to pay of a population.

Data-driven machine learning approaches, on the other hand, do not assume a particular a priori form on the underlying distribution in the data and learn the distribution with an abstract structure. For instance, a deep neural network uses a structure of

hidden layers with hidden units and connections between the hidden units to learn the relationship between variables in the data. Unlike hypothesis-driven approaches, here the interpretability of the estimated model is not trivial. However, these approaches are more efficient when large amount of data is available. They are also better at capturing the higher order correlations and non-linearities in the data.

Prescriptive Analytics

Prescriptive analytics provide recommendations about how to achieve the desired outcome of a phenomenon or system. These types of analytics use the outcomes of predictive analytics to build guidelines or trace a plan for getting to the preferred outcome. Most of the time, the process may involve the use of optimization, where the objective is to find the optimal decision variables that will achieve the desired outcome. For instance, simulation-based optimization can be used to set the signal timing for intersections in a corridor. In a more complicated phenomenon, the optimization may not be operationally tractable, so a range of carefully developed scenarios are developed by the domain experts and are simulated. The outcomes from such scenarios are then compared based on different predefined criteria and the best scenario is selected. For instance, if a municipality is interested in developing a transit-oriented development, it may ask the planners and architects to recommend various scenarios based on the best practices in the design of urban built space and mix of land use. These scenarios are then simulated for the future years and their performance based on the public transit shares is evaluated.

Acknowledgment

The authors are grateful to Arash Kalatian for participating in the initial discussions on the contents of this chapter.

References

- Chowdhury, M., Apon, A., Dey, K., 2017. *Data Analytics for Intelligent Transportation Systems*, 1st ed. Elsevier, Amsterdam, The Netherlands.
- Ortuzar, J., Willumsen, L., 2011. *Modelling Transport*, 4th ed. Wiley, Chichester, West Sussex.
- Möller, D.P.F., 2014. *Introduction to Transportation Analysis, Modeling and Simulation, Simulation Foundations, Methods and Applications*. Springer, London.
- Molugaram, K., Rao, G.S., Shah, A., Davergave, N., 2017. *Statistical Techniques for Transportation Engineering*. Butterworth-Heinemann, Oxford, UK.
- Kutz, M., 2013. *Handbook of Transportation Engineering*. McGraw-Hill, New York.
- Ross, S.M., 1997. Simulation. In: *Statistical Modeling and Decision Science*, Elsevier, Amsterdam, The Netherlands.

Activity-Based Models

Renato Guadamuz, Rajesh Paleti, Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	414
Relative Advantages Over Trip-Based Models	414
Alternate Activity-Based Modeling Systems	415
Survey Data Needs	416
Supplementary Modules	416
Integrated Models	416
Conclusion	417
References	417

Introduction

Historically, the main goal of transportation planning was to predict aggregate travel demand under different long-term socio-economic factors, transportation infrastructure conditions, and regional land-use configurations. These aggregate travel demand models (TDMs) were used to prioritize and build transportation infrastructure projects to meet the expected travel demand. However, constrained budgetary situation coupled with raising awareness about the importance of sustainability shifted the focus to demand management strategies that aim to change individual travel habits as opposed to building new infrastructure. As a result, disaggregate travel demand models that provide greater insights into individual travel-related decisions became increasingly popular in recent times. Activity-based models (ABMs) are a class of disaggregate TDMs that are based on the premise that aggregate travel demand is a derived outcome of the people's needs and desires to participate in activities distributed in space and time (Castiglione et al., 2015). So, the primary focus of ABMs is on understanding the factors that influence individual 'activity-patterns' including (1) the number of and type (e.g., work vs. non-work) of activities undertaken, (2) the activity locations, (3) the time and duration of each activity, (4) the travel mode used to reach the activity location, and (5) with whom the activity was undertaken. In a way, ABMs attempt to predict the complete activity diary of every traveler in the study region. Quite understandably, the development of ABMs is a data-intensive and computationally complex endeavor. Therefore, transportation planners have attempted to develop ABMs that can reasonably mimic the true behavior of travelers while maintaining practical feasibility. So, different simplifying behavioral assumptions (e.g., the sequence in which different activity-pattern decisions are made) can lead to different ABMs with different policy implications. Unlike aggregate TDMs such as the traditional trip-based four-step model that has become the standard practice for long-term travel demand forecasting, ABMs are constantly evolving to reflect new behavioral research findings and innovative transportation technologies, and to incorporate efficient computational processes.

Relative Advantages Over Trip-Based Models

Trip-based models commonly include four primary components—trip generation, trip distribution, mode choice, and traffic assignment. To some extent, these components mimic the activity frequency, destination, mode, and route choice decisions made by travelers. However, given the aggregate nature of trip-based models, they do not account for realistic constraints of space and time that dictate the choices made by individual travelers (Pinjari and Bhat, 2011). Every trip is assumed to be made independently without recognizing that all attributes (e.g., location, mode, and time) of any given trip as well as different trips made by an individual are inter-related. For instance, linked trips that begin and end at the traveler's home have fewer feasible travel mode combinations compared to three independent trips. Along similar lines, trips made by individuals belonging to the same household are also assumed to be independent in trip-based models. For instance, a full-time worker who drops-off her/his child at school during commute would continue to do so even after an additional toll is imposed because of the child-care responsibility. A trip-based model is prone to over-estimate the reduction in car mode usage under congestion pricing scenario because it does not account for such household and person-specific space-time constraints. Also, the "time" dimension is either completely ignored or treated in an ad-hoc static manner in trip-based models by assuming that a fixed percentage of trips are made during different aggregated time-periods (e.g., peak vs. off-peak). Due to these simplistic assumptions, trip-based models are prone to provide biased travel demand forecasts under demand management strategies. The disaggregate activity-based models, on the other hand, explicitly account for space-time constraints among trips made by an individual as well as trips made by other people belonging to the same household and are capable of continuous representation of time. A related advantage of ABMs is that policy impacts can be easily summarized for different population segments to support equity analysis (Castiglione et al., 2015). Also, any performance metric generated by

trip-based models (e.g., mode shares and trip-length distribution) can be produced by the activity-based models by aggregating individual travel choices.

Alternate Activity-Based Modeling Systems

Given that the focus is on the individual traveler who is the decision-making agent, activity-based modeling systems are ensembles of multiple sequential and inter-connected sub-modules that predict different activity and travel choices of individual travelers. These modeling systems are applied sequentially to all travelers in the study region to populate the activity and travel diaries of the entire population. A key step that defines the heart and soul of an ABM is the design of the behavioral framework which defines the sequence/order in which different activity-travel decisions are made by individual travelers within each household. Most ABMs have two primary steps—(1) the activity generation step in which the activity participation decisions of all individuals in each household are predicted, and (2) the activity scheduling step in which the generated activities are packaged into tours that begin and end at the home or work place of the traveler along with a host of additional predicted attributes including travel mode, destination location, time-of-day, and travel party (Pinjari and Bhat, 2011). Typically, mandatory activities such as work or school are predicted first that later serve as spatial and temporal pivots around which other non-mandatory activities (e.g., shopping) with relatively more flexibility are scheduled. With the advent of effective telecommunication technologies, the spatial and temporal lines between work and non-work activity windows are getting blurred necessitating more nuanced design frameworks in recent times. Irrespective of the design framework adopted, most of the activity-travel choices that must be predicted within the behavioral framework are discrete in nature, that is, the traveler is faced with choice situations where one option must be picked from a set of mutually exclusive and collectively exhaustive alternatives (Train, 2003). For instance, in the commute mode choice context, the traveler chooses one travel mode from the set of car, transit, and walk/bike alternatives. So, these sub-modules are either statistical models that predict the probability of different choice outcomes or set of rules in the form of condition-action pairs that represent context-specific decision heuristics used by travelers (Pinjari and Bhat, 2011).

In the first group of ABMs that employ statistical models, the most commonly used sub-module structure is a discrete choice model that is based on the utility maximization theory of consumer choice that assumes that the traveler makes activity-travel choices that maximize his/her utility or satisfaction derived from that choice (McFadden, 2000; Bhat, 2003). For instance, in the mode choice context, every competing modal alternative is assumed to have some quantifiable utility associated with it (which is a function of the modal attributes such as travel time and travel cost as well as traveler characteristics such as income and gender). The traveler is then assumed to choose the modal alternative that maximizes his/her utility. Given that it is extremely unlikely for the analyst to know all the factors affecting the traveler's utility function, stochasticity is introduced in the utility functions leading to probabilistic predictions. Different assumptions about the nature of stochasticity underlying utility functions lead to different statistical models. For instance, the commonly used multinomial logit (MNL) and nested logit models are obtained by assuming the utility functions have a multivariate Gumbel distribution (Koppelman and Bhat, 2006). One of the critiques of utility-maximization consistent choice models is that travelers are not always perfectly rational and may make sub-optimal choices due to their limited cognitive capabilities. However, in recent times, choice models that can approximate a wide variety of decision heuristics that are not necessarily consistent with utility maximization are developed in the research community (Hess et al., 2012). In choice contexts where the underlying outcome is continuous (e.g., activity participation duration), regression and hazard durations models are also used. Examples of statistical models-based ABMs that have seen applicability in different metropolitan areas include CEMDAP (Comprehensive Econometric Microsimulator for Activity-Travel Patterns) (Bhat et al., 2004), CT-RAMP (Coordinated Travel Regional Activity Modeling Platform) (Davidson et al., 2010), TASHA (Travel/Activity Scheduler for Household Agents) (Miller and Roorda, 2003), and openAMOS (open source activity-mobility simulator) (Pendyala et al., 2012).

Rule-based ABMs are composite systems of if-else conditions which resolve activity-travel tasks that individual travelers confront daily. The underlying idea of rule-based ABMS is that the activity-travel decisions in the real world are too complex to handle within a strict utility-maximizing regimen and are based on several context-dependent heuristic rules working in tandem. However, one of the main difficulties with rule-based ABMs is that the number of if-else conditions that must be enumerated can be too large for complex decisions involved during activity generation and scheduling. In fact, most rule-based ABMs assume the generation of activity episodes as an exogenous input and primarily focus on the activity scheduling step. Also, the statistical significance of factors affecting different activity-travel decisions is not readily apparent in rule-based ABMs (Pinjari and Bhat, 2011). ALBATROSS (A Learning-BAsed TRansportation Oriented Simulation System) is an example of rule-based ABMs in which the activity-travel decisions are derived from decision trees that are trained using observed data (Arentze and Timmermans, 2004).

Hybrid ABMs that use a combination of statistical models and decision rules as well as other modeling approaches are becoming increasingly common. For instance, TASHA uses statistical models for predicting the number of activity episodes for each person along with the desired start time and duration of each episode. In the next step, the activity travel schedule is built from scratch by adjusting start times and durations to ensure feasibility. So, activity generation is not completely prior to the scheduling step because real-life schedules change, and some activity episodes may be removed as well as new activity episodes may be added during the scheduling phase (Miller and Roorda, 2003). Along similar lines, openAMOS uses statistical models for predicting activity participation and scheduling decisions in combination with the concept of space-time prisms that define the open boundaries within which the activities can take place (Pendyala et al., 2012). More recently, agent-based modeling concepts in which travelers are treated as autonomous agents who interact with their travel environment and can learn and adapt over time are incorporated in the

rule-based ABM—ALBATROSS (Arentze and Timmermans, 2004). Another example of a hybrid ABM is ADAPTS (Agent-based Dynamic Activity Planning and Travel Scheduling) which incorporates both statistical models for predicting activity-travel behavior and agent-based modeling concepts for handling dynamic interaction between the travelers and the travel environment (Auld and Mohammadian, 2009).

Survey Data Needs

The statistical models or decision rules that govern the behavior of synthetic population agents within ABMs are trained based on the observed data. To do this, extensive household travel survey (HTS) that collect detailed activity-travel diary information for a 24-hour period of a representative sample of households and individuals within those households in the study region is undertaken. The information recorded in these surveys include the activity purpose, location, departure and arrival times, duration, and means of transportation. For example, the National Household Travel Survey (NHTS) is a nationally representative periodic national survey to understand travel patterns in United States. For futuristic conditions that do not exist in the market currently (e.g., autonomous vehicles or new tolling scenarios), stated preference surveys that elicit responses from travelers under controlled choice experiments are used. In some cases, researchers develop custom-designed surveys that elicit more information on the decision process rather than choice outcomes to support the activity scheduling and conflict resolution decision rules used within ABMs (Doherty et al., 2004). In some cases, multi-day surveys that track habitual activity-travel behavior of households over time instead of a single 24-hour period are also used (Cherchi and Mohammadian, 2014).

Supplementary Modules

ABMs simulate the activity-travel patterns of every traveler residing within the study region. Therefore, a key input to ABMs are synthetic agents that are simplified microscopic representation of the actual resident population. For instance, Southern California has about 6 million households with 19 million residents. So, any ABM for Southern California should create 19 million synthetic population agents who belong to 6 million residents such that key socio-demographic characteristics of this synthetic population closely replicate those in the real-world within small geographies (such as Traffic Analysis Zones or Census Tracts). The creation of these synthetic agents (i.e., travelers) is typically handled within a supplementary module known as the “Synthetic Population Generator” (SPG). Synthetic agents are created by expanding a much smaller survey sample (e.g., Public Use Microdata Sample (PUMS) data) that includes socio-demographic information of households and their constituent people. This expansion is done iteratively such that the statistical distributions of key demographic factors in the generated synthetic population closely replicate those in the actual population while maintaining the household and population control totals (Ye et al., 2009). However, SPGs typically generate synthetic agents with only select socio-demographic factors (e.g., age, gender, and residential location). Other medium-to-long term attributes that influence activity-travel patterns (e.g., employment status, occupation industry, work location, housing tenure and type, household auto ownership, and household income) are another supplementary module because the data generated from SPGs lack variation due to expansion from a smaller sample (Pendyala et al., 2012). Another key input to ABMs is the regional land-use that is surrounding the resident population. To do this, land use models are used to predict the development patterns including the home and business locations (Waddell, 2002). Lastly, the transportation network infrastructure along with multi-modal level-of-service skims (e.g., travel times and costs) generated using traffic assignment models are also needed as input to ABMs.

Integrated Models

There are potentially strong inter-dependencies between the inputs used in the ABMs and activity-travel patterns generated by the ABMs. An obvious inter-dependency is between the network level of service (LOS) and activity-travel patterns. While activity-travel patterns at any given time depend on the network travel times and costs, it is also true that changes in the travel patterns will change the congestion patterns and associated travel times. So, there is a dynamic interaction between travel demand and transportation supply at smaller time scales. To account for this inter-dependency, ABMs are integrated with Dynamic Traffic Assignment (DTA) models in such a way that any given time the travel times between different origin-destination pairs in the activity-travel pattern are consistent with realistic network travel times (Auld et al., 2012, 2016). Along similar lines, over longer time scales, the activity-travel patterns can produce changes in the regional business and residential location patterns. For instance, transportation pricing policies can change where travel leading to residential and business relocation decisions in the long-term. Therefore, integrated land-use and transportation models that capture demand and supply interactions over medium-long term scales are being developed (Salvini and Miller, 2005; Waddell, 2011). However, the design as well as implementation of such integrated modeling systems is a challenging endeavor requiring expertise spanning multiple disciplines including transportation, urban planning, economics, human dynamics, and computer science.

Conclusion

The core idea underlying activity-based models (ABMs) is that people travel to partake in different types of activities dispersed in space and time. So, future travel demand and associated traffic conditions under different scenarios can be predicted more accurately (relative to trip-based models), if the individual decision processes underlying these activity-travel decisions can be captured better. In the past decade, ABMs have come a long way both in terms of behavioral accuracy and implementation efficacy. Recent ABMs encompass several behavioral enhancements including tour formation (i.e., where tours emerge from activity needs as opposed to fitting activities into tour templates) (Paleti et al., 2017), joint activity participation among household members (Bhat et al., 2013), and allocation and tracking of household vehicles (Hatzopoulou et al., 2007). Several metropolitan areas around the world are successfully transitioning to ABMs for developing their regional transportation plans (RTPs). Ongoing efforts are focused on (1) enhancing the policy sensitivity and forecasting accuracy of ABMs, (2) development of computationally efficient integrated modeling systems for capturing demand and supply interactions over different time scales, and (3) improving the behavioral realism of the sub-modules that govern the actions of synthetic agents within ABMs.

References

- Arentze, T.A., Timmermans, H.J.P., 2004. A learning-based transportation oriented simulation system. *Transp. Res.* 38 (7), 613–633, doi:10.1016/j.trb.2002.10.001.
- Auld, J., Mohammadian, A., 2009. Framework for the development of the agent-based dynamic activity planning and travel scheduling (ADAPTS) model. *Transp. Lett.* 1 (3), 245–255, doi:10.3328/TL.2009.01.03.
- Auld, J., Javanmardi, M., Mohammadian, A.(Kouros), 2012. Integration of activity scheduling and traffic assignment in the ADAPTS activity-based model, Transportation Research Board 91st Annual Meeting, (No. 12-4225).
- Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., Zhang K., 2016. POLARIS: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. *Transp. Res.* 64, 101–116, doi:10.1016/j.trc.2015.07.017.
- Bhat, C.R., 2003. Random utility-based discrete choice models for travel demand analysis. *Transp. Syst. Plan.* 10, 1–30.
- Bhat, C.R., Guo, J.Y., Srinivasan, S., Sivakumar A., 2004. Comprehensive econometric microsimulator for daily activity-travel patterns. *Transp. Res. Rec.* 1894 (1), 57–66, doi:10.3141/1894-07.
- Bhat, C.R., Goulias, K.G., Pendyala, R.M., Paleti, R., Sidharthan, R., Schmitt, L., Hu, H.-H., 2013. A household-level activity pattern generation model with an application for Southern California. *Transportation* 40 (5), 1063–1086, doi:10.1007/s11116-013-9452-y.
- Castiglione, J., Bradley, M., Gliebe, J., 2015. Moving to activity-based travel demand models. *Activity-Based Travel Demand Models: A Primer*. Available from: [https://doi.org/Strategic Highway Research Program \(SHRP 2\) Report S2-C46-RR-1](https://doi.org/StrategicHighwayResearchProgram(SHRP2)ReportS2-C46-RR-1).
- Cherchi, E., Cirillo, C., 2014. Understanding variability, habit and the effect of long period activity plan in modal choices: a day to day, week to week analysis on panel data. *Transportation* 41 (6), 1245–1262.
- Davidson W. Vovsha P.Freedman J. Donnelly R. CT-RAMP family of activity-based models. In: *Proceedings of the 33rd Australasian Transport Research Forum (ATRF)*, Vol. 29, p. 29, 2010.
- Doherty, S.T., Nemeth, E., Roorda, M., Miller, E.J., 2004. Computerized household activity-scheduling survey for Toronto, Canada, area: Design and assessment. *Transp. Res. Rec.* 1894 (1), 140–149, doi:10.3141/1894-15.
- Hatzopoulou, M., Miller, E.J., Santos, B., 2007. Integrating vehicle emission modeling with activity-based travel demand modeling: case study of the Greater Toronto, Canada, Area. *Transp. Res. Rec.*, doi:10.3141/2011-04.
- Hess, S., Stathopoulos, A., Daly, A., 2012. Allowing for heterogeneous decision rules in discrete choice models: An approach and four case studies. *Transportation* 39 (3), 565–591, doi:10.1007/s11116-011-9365-6.
- Koppelman F.S. Bhat C. A self instructing course in mode choice modeling: multinomial and nested logit model Accession Number: 01036604, Federal Transit Administration. U.S. Department of Transportation 2006 <http://doi.org/10.1002/stem.294>.
- McFadden, D., 2000. Disaggregate behavioral travel demand's RUM side A 30-year retrospective. *Travel Behav. Res.* 17–63, <https://doi.org/10.1.1.26.6944..>
- Miller, E.J., Roorda, M.J., 2003. Prototype model of household activity-travel scheduling. *Transp. Res. Rec.* 1831 (2), 114–121, doi:10.3141/1831-13.
- Paleti, R., Vovsha, P., Vyas, G., Anderson, R., Giaimo, G., 2017. Activity sequencing, location, and formation of individual non-mandatory tours: application to the activity-based models for Columbus, Cincinnati, and Cleveland, OH. *Transportation* 44 (3), doi:10.1007/s11116-015-9671-5.
- Pendyala, R., Konduri, K., Chiu, Y.C., Hickman, M., Noh, H., Waddell, P., Wang, L., You, D., Gardner, B., 2012. Integrated land use-transport model system with dynamic time-dependent activity-travel microsimulation. *Transp. Res. Rec.* 2303 (1), 19–27, doi:10.3141/2303-03.
- Pendyala, R., Bhat, C., Goulias, K., Paleti, R., Konduri, K., Sidharthan, R., Hu, H.-H., Huang, G., Christian K., 2012. Application of socioeconomic model system for activity-based modeling: experience from Southern California. *Transp. Res. Rec.* 2303 (1), 71–80, doi:10.3141/2303-08.
- Pinjari, A.R., Bhat, C.R., 2011. Activity-based travel demand analysis. In: *A Handbook of Transport Economics*, Edward Elgar Publishing, Texas, doi:10.4337/9780857930873.00017.
- Salvini, P., Miller, E.J., 2005. ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks Spatial Econ.* 5 (2), 217–234, doi:10.1007/s11067-005-2630-5.
- Train, K., 2003. Discrete choice methods with simulation. *Discrete Choice Methods with Simulation*, 9780521816. Available from: <https://doi.org/10.1017/CBO9780511753930>.
- Waddell, P., 2002. Urbansim: modeling urban development for land use, transportation, and environmental planning. *J. Am. Plan. Assoc.* 68 (3), 297–314, doi:10.1080/01944360208976274.
- Waddell, P., 2011. Integrated land use and transportation planning and modelling: addressing challenges in research and practice. *Transp. Rev.* 31 (2), 209–229, doi:10.1080/01441647.2010.525671.
- Ye, X., Konduri, K.C., Pendyala, R.M., Sana, B., Waddell P., 2009. Methodology to match distributions of both household and person attributes in generation of synthetic populations. *Transp. Res. Board Ann. Meeting* 2009.

Advanced Traveler Information Systems

Stephen D. Boyles, Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	418
Technology of ATIS	418
Technologies for Traffic Monitoring	419
Technologies for Reporting Information	419
Predicting Traffic States	420
Planning for ATIS	420
Modeling Individual Choices	420
Modeling Collective Choices	421
Future Directions and Research Needs	421
References	422

Introduction

Transportation systems are inherently uncertain. Travel conditions vary from day to day, both due to “supply-side” factors (incidents, poor weather, or construction work reducing capacity) or to “demand-side” factors (special events, or natural variations in travel patterns). In fact, research indicates that reliability of systems can impact travel choices just as much as their average condition (Pinjari and Bhat, 2006). This is particularly true in mode choice (many travelers are reluctant to take public transit trips, particularly those involving transfers, because of the risk of missing a connection). This uncertainty is particularly acute during disasters, when the transportation system is operating in an unfamiliar state (e.g., during evacuation, or when multiple bridges are closed due to flooding).

As a result, transportation systems need to be designed and managed considering *reliability* in addition to “typical” operating conditions. While it is impossible to eliminate all of these factors causing uncertainty, one strategy is to *provide travelers with up-to-date information on network conditions so they can make better choices*. Advanced traveler information systems (ATIS) is an umbrella term to describe such efforts to inform the public on current conditions.

ATIS has been implemented in many forms, and is growing as technology advances and becomes more accessible. Early forms of ATIS (still used today) include highway advisory radio, variable message signs, and phone lines the public can call to receive information. Later, Internet sites were created to further disseminate information. ATIS is now moving into the use of smartphones and connected vehicles to provide information. All of these forms of ATIS require information on the network condition to be gathered in some way, and ATIS thus is one part of a larger intelligent transportation system that integrates sensing and distribution of information.

While travelers are often grateful for the information provided by ATIS, it is not a panacea for addressing travel reliability. Indeed, there are cases where providing more information (or more accurate information) to travelers can actually degrade the performance of a network. Consider a message sign reporting an accident on a freeway and suggesting an alternate route. If the alternate route has a small capacity, and the accident is relatively minor, encouraging a large number of travelers to switch routes may actually cause more congestion than would have occurred in the first place. Transportation systems are complex, and simply making one part of a complex system better need not improve the performance of the larger system due to these interactions. ATIS must therefore be deployed carefully. Considerable research effort has been invested in developing strategies to forecast these impacts and design effective strategies.

The remainder of this article is organized as follows. The “Technology of ATIS” section provides an overview of the most commonly used technologies in ATIS, as well as some of the supporting technologies used to monitor the network state, and challenges in trying to *predict* a future state from present data. Section, Planning for ATIS, discusses at length several modeling tools (in particular, online shortest paths to represent how individual travelers will use ATIS, and user equilibrium with recourse to describe the collective behavior of many such travelers) that are designed to forecast the impacts that an ATIS system will have when implemented, to ensure maximum benefits to the public and avoid the type of scenario discussed in the previous paragraph. Section, Future Directions and Research Needs, concludes by discussing ongoing efforts in ATIS, and pointing specifically to research needs in this area.

Technology of ATIS

This section describes technologies, which play a role in ATIS—either in monitoring traffic conditions to generate the data for ATIS, or for reporting the data to the public. As technologies advance, the options available for these continue to grow. Therefore, the aim is

to provide an overview of the most common technologies and some emerging options, rather than to provide a comprehensive directory. A separate discussion at the end of the section highlights the differences between simply trying to observe the current state of a traffic system, and trying to *predict* its evolution over the near term. The latter problem is clearly more challenging than the former, but can play an important role in designing effective and trusted information strategies.

Technologies for Traffic Monitoring

Before travel information can be reported to the public, it must be collected in some way by the agency or organization implementing ATIS. Examples of the kind of data that can be reported in ATIS systems include:

- **Travel times** along selected corridors or roadway segments
- **Density** or **speed** of vehicular traffic (e.g., for congestion heatmaps)
- **Weather** information, such as the presence of fog or blowing snow
- Location and severity of **incidents**, **construction work**, or **roadway closures**.
- Quantity and/or location of **available parking spaces**

Although some of these data types typically require manual input (e.g., work zone locations and closures), to the extent possible ATIS systems aim to collect data automatically, possibly synthesized by human intervention, as when traffic management center employees decide what information to include in a recorded traffic information message. A number of technologies exist for continuously monitoring roadway systems to gather such information. Some examples of these technologies are:

- **Inductive loops** which detect the presence of vehicles by recording the change in impedance in a circuit when a large metallic object passes by. Inductive loops are very common on urban freeways, and can directly report vehicle flows and occupancy. By assuming a vehicle length and using traffic flow principles, these values can be converted to speed, density and other traffic variables. Installing and maintaining these devices can be difficult, because they are located within the roadway and cannot be easily accessed without closing a lane and disturbing the pavement surface.
- **Radar, infrared, acoustic, and laser sensors** can also be placed above or alongside the road to measure flows.
- **Video cameras** can be combined with image detection software to provide real-time monitoring of traffic conditions. In principle, video cameras can obtain a wider range of data than inductive loops (vehicle classification, or providing speeds and densities directly, rather than through calculation involving additional assumptions). However, they suffer from occlusion/blocking effects, and image detection algorithms are less reliable in poor weather, in darkness, or when glare or shadows obscure the image.
- **Road weather information systems (RWIS)** are automatic weather stations, recording information such as temperature, wind speed and direction, visibility, and precipitation, and communicating it to a central location.
- **Probe vehicles** can provide travel time measurements directly, rather than estimating them based on speed observations at multiple locations. Historically, official vehicles driving within the traffic stream were used as probes, with specially trained drivers instructed to travel at the average traffic speed. With advancing technology, it is possible to make use of civilian traffic to obtain equivalent data. Examples of this include using **cellular phone GPS** locations to observe speeds or travel times (many smartphones default to an “opt in” for sharing this data to a provider), or using **Bluetooth readers** to record the IDs of devices that pass the station. By locating these readers in multiple locations and matching IDs, the travel times of specific vehicles can be inferred.
- **Automated vehicle location** systems for reporting the location of public transit vehicles. GPS is currently the most common technology for this application. Alternatives (which are suitable in places with poor GPS reception) include the use of “checkpoint” systems, which receive a radio signal from a passing vehicle, or “dead reckoning” systems based on inertial measurements (acceleration, deceleration) from the vehicle to infer position over time.

Technologies for Reporting Information

After traffic data has been collected, it can be distributed to the public in a variety of modes. ATIS systems often use a combination of these modes of communication, to reach a broader segment of the population and accommodate various needs (e.g., more information can be conveyed at once through a visual message, but an audio message may be less distracting to drivers on the road). Common technologies for disseminating information to the public include:

- **Highway advisory radio**, sometimes called **traveler information stations**, broadcast a pre-recorded message on a radio frequency. A sign indicates when a message is being broadcast, commonly using lighted beacons. These devices are commonly used to report information about severe incidents or road closures.
- **Variable message signs** (also called **dynamic** or **changeable message signs**) broadcast information on a highway sign. These signs are commonly used to report corridor travel times, to report roadway incidents, or other special events.
- **Phone hotlines** can provide on-demand information to travelers, for instance playing messages pre-recorded by traffic management center employees. In the United States and Canada, the 511 number is often reserved for this purpose.
- **Websites** or **smartphone apps** can report traffic information on road closures or special events. A common feature is a “congestion heatmap,” reporting density or speed information visually through a color-coded roadway map. Some of these apps are integrated with routing software, to propose alternate routes to travelers in case of unexpected congestion.

- **SMS (texting)** systems can be used to disseminate information to travelers who have provided a mobile phone number.
- **Connected vehicles** can receive information either from roadway infrastructure (V2I) or from other connected vehicles (V2V), and can share information in turn. Onboard vehicle systems can interpret this data in various ways.
- **Transit information** can be provided at transit stops or online. This information commonly includes the estimated arrival times of upcoming buses or trains.

Predicting Traffic States

It is important to distinguish reporting *current* travel observations from reporting *predictions* of future traffic conditions. The former is easier to do; the latter may be more useful in providing route guidance: for instance, during the beginning or end of peak periods, travel conditions throughout a network may change significantly over a 30 minute trip, and anticipating these changes can lead to a better choice of routes even at the start.

There are several approaches to predicting traffic states, based on observations at a current time. Broadly speaking, these approaches can be divided into “model-based” approaches (which use some mathematical model for traffic flow over a network), and “data-driven” approaches (which rely primarily on historical data to search for patterns, using neural networks or other statistical techniques). For some recent examples of model-based approaches, see [Ji et al. \(2009\)](#) or [Kumar et al. \(2017\)](#). For examples of data-driven approaches, see [Duan et al. \(2016\)](#), [Jenelius and Koutsopoulos \(2017\)](#), or [Fan et al. \(2018\)](#).

In all cases, accuracy of the information transmitted to travelers is very important. The effectiveness of ATIS requires that travelers trust the information they receive.

Planning for ATIS

The previous section discussed a number of technological options for implementing ATIS. Planning an *effective* deployment strategy requires additional work. For instance, given a limited budget, where should a city locate variable message signs to have the greatest impact? Where should a state locate highway advisory radio transmitters to report information about weather closures? Should these devices be placed near the locations with highest traffic, the locations where travelers have the most number of viable alternatives, or somewhere else?

There are even more dramatic situations, in which providing information actually can make problems worse. That is, providing more information to more travelers does not always result in improved system performance. This is largely because transportation systems involve complex interactions among many travelers and the physical infrastructure.

In economics, an *externality* occurs when a choice made by an individual affects others who played no part in that choice. Externalities can either be positive or negative; an example of a positive externality is when a homeowner landscapes their yard nicely, in that the neighbors may also benefit from improved home values. In transportation externalities are often negative; a common example is the *congestion externality*, which arises when travelers enter an already-crowded freeway. In doing so, they cause delay to other travelers on the freeway. Another example is in turn prohibitions on busy streets—a driver waiting to turn can cause large delays to vehicles behind them who wish to continue driving through.

When externalities exist, individuals optimizing their behavior based on their own objectives may not optimize the overall system performance; that is, the so-called “invisible hand” fails to guide the system to an optimal state. This is why it may make sense to ban left turns on high-traffic arterials—even though drivers who want to turn left may suffer under such a policy, the extra distance they must travel is outweighed by the overall benefit of smoother traffic across all travelers. A more dramatic example is the Braess paradox ([Braess, 1969](#)), in which constructing a new roadway link degrades network performance for all travelers, because of congestion externalities associated with how drivers choose routes.

Analogous situations appear in ATIS deployment. Advising too many travelers to avoid an accident may overload an alternative route. A direct analogue of the Braess paradox also exists ([Ukkusuri and Patil, 2007](#)), where providing information increases the travel time for all travelers in the network.

All of these considerations underscore the need for careful planning and implementation of ATIS strategies, and subsequent monitoring and performance evaluation to ensure they are having the intended benefits. The remainder of this section discusses two categories of mathematical models, which are designed to predict the impacts of candidate ATIS strategies before implementation. The first category, discussed in Section, Modeling Individual Choices, addresses changes in behavior for *individual* travelers who receive information. The second category, discussed in Section, Modeling Collective Choices, considers the *collective* behavior of a large number of travelers who receive information. These latter models are the most useful for planning ATIS systems, using the former kind of models as building blocks.

Modeling Individual Choices

For concreteness, this section will discuss models for how drivers choose routes in the presence of real-time information. Similar models exist in the transit literature ([Rambha et al., 2016](#)), in parking ([Tang et al., 2015](#)), and other settings. In this setting, the classical assumption is that drivers choose a route with minimum generalized cost (including travel time, a monetary toll, and

possibly other factors). This route can be identified using a *shortest path* algorithm. There are very efficient algorithms that can identify such paths within seconds even on country- or continent-sized networks.

When drivers receive information *en route*, they may change their route choice in response to what they have learned. This leads to a class of *online shortest path* algorithms. Rather than generating a single path, these algorithms generate a set of *policies*, containing all of the paths the driver hypothetically might take, along with the information scenarios under which those paths would be chosen. For instance, a policy might consist of three routes (A, B, and C); where route A is followed under typical conditions, route B is followed when there is an incident at one location, and route C when there is an incident at a different location.

Online shortest path algorithms vary according to the assumptions they make regarding the nature of the information (is it local or distant? do network conditions change between when the information is learned and when the driver reaches them? is it perfect or noisy?), the underlying uncertainties in the network state (are link travel times independent or correlated?), and the behavior of drivers receiving the information (are they risk-neutral or risk-averse?). The simplest algorithms—and the most efficient—assume local information, independent link travel times, and risk-neutral travelers. These can be solved in a comparable amount of time to classical shortest paths. The most complex instances scale poorly with network size and are not yet suitable for large-scale implementation. For more details and examples of such formulations, see Polychronopoulos and Tsitsiklis (1996), Miller-Hooks (2001), Waller and Ziliaskopoulos (2002), Provan (2003), or Boyles and Rambha (2016).

Using such an algorithm, we can determine the impact of an ATIS strategy on any given traveler, by comparing the policy under ATIS to a path chosen a priori in the absence of information, or by comparing policies under different ATIS implementations. By examining the policies, one can compute the mean and standard deviation of travel time, or other reliability measures that can serve as performance metrics.

Modeling Collective Choices

Building on the models described in the previous section, we can estimate the benefits of an ATIS strategy in aggregate, over an entire population. If congestion levels are low, then the impact of one driver's route choice will not impact that of another, and one of the individual models from Section, Modeling Individual Choices, can be applied separately to each traveler, and the results summed (although aggregation by origin or destination can reduce the amount of computation; see Boyles and Waller, 2011). In such cases, there is no risk of ATIS inadvertently worsening travel conditions, because there is effectively no congestion externality.

The situation is rather different in congested networks, because of the presence of externalities. Furthermore, in congested networks drivers may not take ATIS information at face value, but rather try to anticipate how other drivers will respond to the same information. (For instance, upon receiving notification of an accident downstream, a driver seeing many other drivers exit for an alternate route may choose to remain on the freeway, anticipating that enough drivers are diverting to make the delay on the freeway manageable.) Such behavior is especially likely in situations involving experienced drivers who have traveled on the network in many different states in the past, as with regular commuters during the peak period.

In these situations, traffic assignment algorithms can be used to anticipate the effect of an ATIS system. The “base case” scenario without ATIS can be modeled using the classical formulation of the problem; see Boyles et al. (2020) for an overview of such methods. Such methods seek a solution where every traveler is assigned to a path in such a way that they cannot switch to a path with lower travel time. The extension for ATIS systems is to assign travelers to *policies* with the least expected travel time; such methods are termed *user equilibrium with recourse*. In situations where it is easy to find policies (e.g., independent link travel times and risk-neutral drivers), user equilibrium with recourse can be solved using many of the same methods used for classical assignment (Unnikrishnan and Waller, 2009; Rambha et al., 2018). Candidate ATIS deployments can then be compared based on the expected total travel time over all travelers, or using other risk measures.

It is worth noting that user equilibrium with recourse models are relatively new, and date to the last decade. The historical view is that the assumptions of an equilibrium model (first developed to reflect long-term behavior) were incompatible with the real-time perspective of ATIS (which reflects short-term fluctuations in network conditions and behavior). However, the presence of ATIS induces long-term changes in behavior—for example, one of the main arguments for providing real-time transit information is to attract passengers away from driving alone, a change in habitual mode choice. User equilibrium with recourse is designed to capture the changes in habitual behavior (as exhibited by the set of policies), as well as the short-term realization on any given day (which can be obtained by sampling the policy for a given network state and corresponding information).

Agent-based simulation is an alternative to user equilibrium with recourse. In contrast to the above models, agent-based techniques can simulate a wider variety of information strategies and driver behaviors. For instance, in situations involving unfamiliar drivers, the equilibrium assumption may be invalid, and an agent-based model can describe other behaviors. It is also easier to model correlations in the network state. However, it is harder to obtain the data necessary to calibrate agent-based models, since they often involve a larger number of parameters that need tuning, and a greater number of assumptions requiring validation.

Future Directions and Research Needs

Reliability and other effects, caused by nonrecurring sources of congestion, are major issues in transportation systems. Providing information to travelers is a relatively inexpensive strategy for mitigating these types of uncertainty. Recent technological advances make it easier to collect data on traffic conditions throughout a network, and easier to transmit this data to the public. The prospect of

providing information to a large number of travelers introduces new challenges in designing effective information strategies; it is important to anticipate how travelers will use this information, and in particular whether there may be unintended consequences that need to be addressed.

As connected and autonomous vehicle technology continues to progress, ATIS will likely see continued growth in deployment and range of applications. The range of information collected by vehicles, and that vehicles can make use of, will enable new mechanisms for addressing nonrecurring congestion; for instance, rather than simply being passive receivers of information, connected and automated vehicles may coordinate their response to network disruptions, rather than relying on an individual driver to make effective use of a limited set of information.

The prospect of ubiquitous connectivity and traveler information underscores the need for research in better understanding the behaviors, preferences, and values that travelers have regarding reliability in travel; new algorithms to make use of this emerging technological context; and new planning methods that can maximize the benefits of these new opportunities while minimizing any unintended consequences.

References

- Boyles, S.D., Lowmes, N.E., Unnikrishnan, A., 2020. Transportation Network Analysis, Volume I, Version 0.85. Available from: <http://sboyles.github.io/blubook.html>.
- Boyles, S.D., Rambha, T., 2016. A note on detecting unbounded instances of the online shortest path problem. *Networks* 31, 86–99.
- Boyles, S.D., Waller, S.T., 2011. Optimal information location for adaptive routing. *Netw. Spat. Econ.* 11, 233–254.
- Braess, D., 1969. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 12, 258–268.
- Duan, Y., Lv, Y., Wang, F.-Y., 2016. Travel time prediction with LSTM neural network. Presented at the 19th International Conference on Intelligent Transportation Systems (ITSC).
- Fan, S.-K.S., Su, C.J., Nien, H.T., Tsai, P.F., Cheung, C.Y., 2018. Using machine learning and big data approaches to predict travel time based on historical and real-time data from Taiwan electronic toll collection. *Soft Comput.* 22, 5707–5718.
- Jenelius, E., Koutsopoulos, H.N., 2017. Urban network travel time prediction based on a probabilistic principal component analysis model of probe data. *IEEE Transact. Intell. Transp. Syst.* 19, 436–445.
- Ji, Y., Zhang, X., Daamen, W., Sun, L., 2009. Traffic incident recovery time prediction model based on cell transmission model. Presented at the 12th International Conference on Intelligent Transportation Systems (ITSC).
- Kumar, B.A., Vanajakshi, L., Subramanian, S.C., 2017. Bus travel time prediction using a time-space discretization approach. *Transp. Res. Part C* 79, 308–332.
- Miller-Hooks, E.D., 2001. Adaptive least-expected time paths in stochastic, time-varying transportation and data networks. *Networks* 37 (1), 35–52.
- Pinjari, A.R., Bhat, C., 2006. Nonlinearity of response to level-of-service variables in travel mode choice models. *Transp. Res. Rec.* 1977, 67–74.
- Polychronopoulos, G.H., Tsitsiklis, J.N., 1996. Stochastic shortest path problems with recourse. *Networks* 27 (2), 133–143.
- Provan, J.S., 2003. A polynomial-time algorithm to find shortest paths with recourse. *Networks* 41 (2), 115–125.
- Rambha, T., Boyles, S.D., Unnikrishnan, A., Stone, P., 2018. Marginal cost pricing for system optimal traffic assignment with recourse under supply-side uncertainty. *Transp. Res. Part B* 110, 104–121.
- Rambha, T., Boyles, S.D., Waller, S.T., 2016. Adaptive transit routing in stochastic time-dependent networks. *Transp. Sci.* 50, 1043–1059.
- Tang, S., Rambha, T., Hatridge, R., Boyles, S.D., Unnikrishnan, A., 2015. Modeling parking search on a network using stochastic shortest paths with history dependence. *Transp. Res. Rec.* 2467, 73–79.
- Ukkusuri, S.V., Patil, G., 2007. Exploring user behavior in online network equilibrium problems. *Transp. Res. Rec.* 2029, 31–38.
- Unnikrishnan, A., Waller, S.T., 2009. User equilibrium with recourse. *Netw. Spat. Econ.* 9, 575–593.
- Waller, S.T., Ziliaskopoulos, A.K., 2002. On the online shortest path problem with limited arc cost dependencies. *Networks* 40 (4), 216–227.

Demand-Responsive Transit, Evaluation Studies

Sebastián Raveau, Department of Transport Engineering and Logistics, Pontificia Universidad, Católica de, Chile

© 2021 Elsevier Ltd. All rights reserved.

Understanding Demand-Responsive Transit	423
Operational Characteristics of Demand-Responsive Transit	424
DRT Applications	424
Australia: CoastConnect	424
Austria: Go-Mobil	424
Belgium: FlexiTec	425
Bulgaria: FMS	425
Germany: Bürgerbus, Muldental in Fahrt and Rebus	425
Italy: ProntoBus	425
Ireland: Ring a Link & Local Link	425
The Netherlands: RegioTaxi	425
Portugal: Médio Tejo	426
Slovenia: Sopotniki	426
Spain: Shotl & Castilla y León	426
Switzerland: Alpine Bus	426
United Kingdom: Suffolk Links	426
United States: ITNAmerica & MetroAccess	426
Evaluation and Discussion	426
Biography	427
Further Reading	427

Understanding Demand-Responsive Transit

Demand-responsive transit (DRT) is a flexible form of public transport where vehicles adjust their service (in terms of pick-up/drop-off locations and hours of operation) based on travelers' needs rather than having fixed routes or timetables. Vehicles' capacities tend to be low, and the fleet can include taxis, vans, minibuses, or buses. DRT seeks to overcome some shortcomings of conventional transit systems in two main aspects: providing transit service in low density or poorly connected areas and providing accessibility for all travelers regardless of their necessities.

DRT services are usually presented as a complement for traditional fixed transit services. DRT services can work as feeder services that connect travelers with high-demand services (such as buses, trams, or trains). Integration with other modes (such as bike-sharing systems) has also risen as an opportunity in recent years. Alternatively, in low-density or rural areas DRT can fully replace traditional fixed transit services, aiming to provide better levels-of-service and, in some cases, lower operational costs. Empirical evidence shows that this can happen in areas where the hourly demand is around 40 passengers per mile² or less.

As urbanization levels increase worldwide, rural areas experience a decrease in population. This can lead to a decrease in level-of-services (such as reductions in the number of offered routes and their frequency) due to the high costs of maintaining low demand services. In such cases, DRT can be an attractive alternative to adjust the transit operation to its demand, offering a better level-of-service, without necessarily increase the operational costs.

DRT has also risen as an alternative for travelers with special needs (such as travelers with reduced mobility), even in high-density areas. When traditional fixed transit services do not provide an accessible service for all travelers, DRT can act as an alternative and complementary system designed for key segments of the population.

DRT are designed to serve low-demand areas and their operation is based around the flexibility on the routes and stops. Routes may be based around corridors or service areas. Stops may be predefined (i.e., fixed stops along the route), semi-flexible (i.e., combining fixed stops and on-demand stops), or fully flexible (i.e., no stops are defined in advanced). As DRT flexibility increases, access and egress times decrease, offering door-to-door service.

DRT service can be supplied by a wide variety of operators: local governments, community organizations, transit providers, and private companies, among others. This way, the level of integration and formality of DRT varies. If DRT is designed as a complement to traditional fixed transit, serving as a feeder, cooperation between both systems are key aspects in order to effectively provide a good level-of-service (e.g., by coordinating transfers and integrating fares).

Although DRT services are not a new concept (they have been proposed and discussed for many decades, mainly in developed countries), only in recent years there has been a significant increase in implementations worldwide due to technological advancements for on-line communications and management. Nowadays, travelers can easily book DRT services in real time, providing a temporal component to the DRT flexibility.

Operational Characteristics of Demand-Responsive Transit

In this section, the main characteristics of DRT, in contrast to traditional fixed transit services, are discussed. Although the implementation and characteristics of DRT systems vary, it is possible to identify different aspects of the level-of-service that they tend to provide:

1. *Connectivity*: The attractiveness of any transport mode as a travel alternative for a trip is based on the places it connects. This way, DRT should be flexible enough to serve the needs of the travelers. The route flexibility can range from fixed configurations of routes serving predetermined stops along corridors to fully flexible door-to-door services.
2. *Capacity*: DRT fleets often consider low capacity vehicles to serve low demand areas. The results of different applications around the world recommend using vehicles for 8–10 passengers. The fleet size is a key design aspect of DRT, as it determines the number of routes that can be served by the system.
3. *Hours of operation*: While DRTs flexibility is usually described in a spatial manner (adjusting routes and stops), it can also provide flexibility in a temporal manner (adjusting hours of operation). DRT services can run whenever they are needed, reducing some of the operational costs associated with running empty vehicles. Nevertheless, as in any transit system, relocation costs are always present. Moreover, DRT can be an attractive alternative for serving low demand in early/late periods that are not served by traditional fixed transit.
4. *Information and booking*: Current technologies allow for real-time bookings, as a complement for advanced booking. The former requires the travelers to wait for the service and imposes a planning challenge for the service provider, as the systems must react to real-time information. This usually leads to a higher fleet size.

Travel times are highly dependent on the flexibility of the service. As any shared mode of transport, DRT travelers are affected by each other's travel structure, particularly if the stops are flexible and the service covers an area without predetermined corridors. This way, travelers may not know in advanced how long the trip would take. Although this is true for any transit mode, where there are boarding and alighting times along the trip, the effects on DRT are increased due to deviations made along the way to serve travelers' necessities.

DRT services tend to have low frequencies and to operate without timetables. If the service runs on a corridor with predetermined stops, then travelers can access those stops and wait for the service. But in general, DRT services tend to operate with advanced and/or on-line reservations. The former can eliminate waiting times and provide a dependable level-of-service. The latter leads to a waiting times which, depending on the density of services and fleet size, can largely exceed those of traditional fixed transit services.

If the DRT system operates close to its capacity, there is a possibility that on-line reservations are declined (or at least largely delayed) and that demand is not served. The criteria for declining reservations depend on vehicle availability and deviation to pick-up or drop-off the requesting traveler. Declining a trip implies one of the most undesired outcomes of DRT operation. While denied boardings can occur on traditional fixed transit services, their higher frequency can alleviate the effects.

As DRT vehicles are smaller than traditional fixed transit, passengers ride seated. This way, comfort can be a key differentiating element as high crowding levels are nonexistent. This is particularly important for travelers with reduced mobility that highly depend on such level-of-service.

Depending on the structure of the DRT system and its integration with other modes of transport, travelers might need to transfer to reach their destinations. This happens when the DRT is designed to act as a feeder service to mass transit. Transferring does not only increase the total travel time but also can add a source of uncertainty in terms of waiting times.

DRT Applications

In this section, some implementations of DRT worldwide are described. As the characteristics and potential success of DRT systems highly depend on the context, in many cases it is not possible to compare different implementations. Nevertheless, it is possible to draw some good practices and establish some recommendations.

Australia: CoastConnect

DRT services in Australia have been designed mainly as a feeder system to connect travelers with other mass transit modes. It is aimed at commuters living in rural or low density areas and to reduce their car usage. This way, the main objective of the system is to achieve a more sustainable mobility. The system offers flexible services with door-to-door connectivity, where the trips must be booked in advanced through a smartphone application.

Austria: Go-Mobil

DRT systems in Austria offer door-to-door mobility alternatives for people in suburban and rural areas where traditional fixed transit services offer a low level-of-service or are even absent. The services are designed to be used by different segments of the population, focusing on off-peak travelers and noncommuters.

The services operate with a mixed fleet of cars, taxis, and minibuses, which accommodate to the demand, under a public private partnership scheme. Hours of operation extend until late in the night, where the availability of transport modes decreases. Trips must be booked in advance. The services are designed to offer door-to-door connectivity, while connection with other regional public transport alternatives is also possible.

Belgium: FlexiTec

DRT has been envisioned in Belgium as a last resort service for users without access to other transit alternatives, enhancing the accessibility of individuals living in poorly connected or remote areas. A key target demographic are elders and other travelers with reduced mobility that lack ease of access to regular transit and do not own a car. The majority of users are female noncommuters from households where the car is used by someone else.

Trips must be booked in advanced and the system establishes a limit of trips a traveler can make within a year, in order to manage the capacity and provide access for all. Services provide door-to-door connectivity at a significantly lower fare than other potential bus alternatives.

Bulgaria: FMS

DRT applications in Bulgaria have been targeted for tourists, in order to provide better accessibility to touristic sites, boost the economy and reduce the car usage of tourists. This focus on tourism has implied several particular characteristics of these systems: the use of electric and even horse-powered vehicles, the seasonal variation of routes and frequencies, the need to operate services with short notices, and the possibility to offer tours as part of the service.

Germany: Bürgerbus, Muldental in Fahrt and Rebus

DRT systems in Germany are diverse. While many operate in rural areas that lack connectivity through other means of public transport, other systems operate in urban areas for specific segments of the population. In general, they tend to be a complement to traditional urban and regional mass transit.

Services tend to be offered by local governments and municipalities, aimed at providing accessibility for the entire community, particularly in rural areas. Volunteer-based services are also present. While the services are available for everyone, users tend to be students, elders and people with reduced mobility, as is the case in many other countries. The fleets operate generally operate with fixed routes and stops, handling both advanced and real-time reservations.

Italy: ProntoBus

DRT in Italy has been implemented for connecting urban and suburban areas with rural areas that lack other transit services. While the system is designed to serve low density areas without alternative connectivity options, it has a special interest on travelers with reduced mobility, providing accessibility for all individuals. Services are fully flexible and allow traveling between any pairs of predetermined stops.

One of the most interesting characteristics of the DRT services is the real-time information platform provided to the users. Through a smartphone app the travelers can manage their bookings and access to information regarding the hours of operation, route availability and travel options. This interface also allows the operators to collect valuable information regarding travel structure for planning purposes.

Ireland: Ring a Link & Local Link

DRT services have been implemented in rural counties with small towns. While the counties are well interconnected, DRT provides internal services. The services are community-based and aim to reduce social exclusion. The system requires advanced booking and is mainly used by commuters and students, as well as people living in isolated places or having special needs, which now at least have a weekly opportunity for conducting activities. This way, reducing social exclusion has been one of the key drivers of DRT in Ireland.

The DRT system is comprised of different services: fixed route services that operate on a regular schedule, on-demand services that provide flexible travel options for travelers, school transport for students living in isolated places, late services to reduce drink-driving, services that connect residents with other transport hubs and social trips for members of the community.

The Netherlands: RegioTaxi

DRT services have been implemented in several regions in The Netherlands, serving low population and rural municipalities that lack other means of transport. The system aims to provide connectivity mainly in the evening, when other transit options do not operate, particularly for travelers with reduce mobility or special needs.

The services are fully flexible without predetermined routs or stops, offering door-to-door connectivity for all users. The system operates with a mixed fleet of minibuses and taxis, depending on the levels of demand. Travelers must book their trips in advanced.

Portugal: Médio Tejo

DRT services have been implemented in medium- and low-size towns in mountainous areas of Portugal, in order to promote sustainable development and enhance the connectivity for people living in low density or remote areas. No specific target demographic was defined by the operators, serving all trip purposes.

The services are operated by taxis and are fully integrated with other public transport modes, serving as feeder services. The system operates with fixed routes and stops, where trips are booked in advanced, during peak hours.

Slovenia: Sopotniki

DRT services have been implemented in Slovenian rural municipalities for elders without other travel alternatives. The service is free of charge and provides door-to-door connectivity in both directions, while accommodating for any special mobility needs. Volunteer drivers can even accompany the users on their activities, providing additional support if needed. The system is designed to provide mobility solutions for elders and avoid their social exclusion.

Spain: Shottl & Castilla y León

DRT services cover low demand origin–destination pairs in rural or poorly connected areas in different municipalities of Spain. The systems are designed to replace traditional fixed transit services due to their financial and operational restrictions. Although door-to-door connectivity is possible in some cases, the services are designed to provide connectivity to other public transport modes, acting as feeder services.

While trips can be booked in advanced, on-line booking has been implemented in recent years in some of the municipalities. Services operate with both flexible and predetermined routes and stops. A hub-and-spoke structure has been adopted in locations where a significant number of trips (mostly going to and from the urban areas) can be coordinated to reduce operational costs and increase efficiency.

Switzerland: Alpine Bus

As is the case in other countries, DRT services in Switzerland are offered where no other transit alternatives are available. The system covers mountain areas across different Swiss cantons where the population density is extremely low. But the main difference is that these services have an emphasis on tourism, the main economic activity in the areas where they operate. This way, service routes and frequencies change seasonally according to the number of tourists visiting the area.

United Kingdom: Suffolk Links

DRT services have been implemented in low density counties of the United Kingdom where connectivity is low, as no regular bus services are available. The services require advanced booking and operate 6 days a week during working hours. The DRT system provides opportunities for different travelers; while students and workers without access to car are the main demographic, the system is also used by people shopping, running errands, and even sightseeing. The system also provides access to other modes of public transport.

One of the main objectives of the DRT system is to provide sustainable travel options for the communities. Not only the vehicles are environmentally friendly in terms of emissions, but the system also reduces the necessity of households of owning a second car. The DRT services have the same fare structure than regular buses and can provide door-to-door connectivity if needed.

United States: ITNAmerica & MetroAccess

The DRT system operates as community service for elders and people with reduce mobility that lack other means of transport, in many communities across the United States where regular transit connectivity is poor or nonexistent. This way, the system provides accessibility to a segment of the population that can be otherwise excluded of society. The system is operated through privately owned vehicles and provides door-to-door connectivity for the travelers. In some cases, eligibility cards are required to use the system.

Evaluation and Discussion

Although DRT services can operate in any low demand context, most of the implementations worldwide have been in rural areas, where traditional fixed transit is absent or offers low level-of-service. The reason for such a situation is mainly an economic one: it is not viable nor convenient to operate traditional fixed transit services. While DRT can better adjust to the demand levels, providing efficient services in terms of vehicle occupancy, revenues are generally low and providers can experience losses.

The limited economic attractiveness of DRT has led to municipalities and communities to provide such services. The goal is not a private one, but a social one. DRT has proven to be a mobility solution for all segments of the population, with an emphasis on

people with reduced mobility and people without access to other means of transport. This focus on the disadvantaged is key element of DRT applications.

The social component of some DRT systems has led to free-of-charge services offered by volunteers. This is certainly based on a strong sense of collaboration within local communities, building intergenerational solidarity. Such circumstances lead to a strong support and acceptance of the DRT services within their respective areas of operation. At the same time, the involvement of the communities helps to further develop the systems.

DRT have also played a significant role in promoting sustainable transport. Many systems worldwide have set as a goal the reduction of car ownership rates. Providing a travel alternative for households instead of buying a second vehicle has helped in that sense. This can explain some mobility trends observed in many different contexts: DRT travelers tend to be noncommuters, elders, and students. Additionally, the introduction of electric vehicles to DRT fleets can further help in providing more sustainable means of travel.

Other positive externalities have been observed in some implementations of DRT. Providing late services, when traditional fixed transit does not operate, can help reduce drink-driving and provide safe mobility for vulnerable segments of the population. DRT has also shown to be an attractive alternative for promoting and easing tourism in rural areas.

Biography



Sebastián Raveau received his B.S. degree in Civil and Industrial Engineering in 2009 and his Ph.D. degree in Engineering Sciences (Transportation program) in 2014 from Pontificia Universidad Católica de Chile (PUC). He is an assistant professor at the Department of Transport Engineering and Logistics at PUC, and a researcher at the Center for Sustainable Urban Development (CEDEUS) and the BRT+ Centre of Excellence. His research interests include travel behavior, travel demand analysis and public transport systems.

Further Reading

- Alonso-González, M.J., Liu, T., Cats, O., Van Oort, N., Hoogendoorn, S., 2018. The potential of demand-responsive transport as a complement to public transport: an assessment framework and an empirical evaluation. *Transport. Res. Rec.* 2672, 879–889.
- Atasoy, B., Ikeda, T., Song, X., Ben-Akiva, M.E., 2015. The concept and impact analysis of a flexible Mobility on Demand System. *Transport. Res. Part C* 56, 373–392.
- Cayford, R., Yim, Y.B., 2004. Personalized Demand-Responsive Transit Service. California PATH Research Report UCB-ITS-PRR-2004-12.
- ITF, 2015. International Experiences on Public Transport Provision in Rural Areas. International Transport Forum, Discussion Paper 2015-7.
- Khattak, A.J., Yim, Y., 2004. Traveler response to innovative personalized Demand-Responsive Transit in the San Francisco Bay Area. *J. Urban Plan. Develop.* 130, 42–55.
- Li, X., Quadrioglio, L., 2010. Feeder transit services: choosing between fixed and demand responsive policy. *Transport. Res. Part C* 18, 770–780.
- Mageean, J., Nelson, J.D., 2003. The evaluation of demand responsive transport services in Europe. *J. Transport Geogr.* 11, 255–270.
- Palmer, K., Dessouky, M., Zhou, Z., 2008. Factors influencing productivity and operating cost of demand responsive transit. *Transport. Res. Part A* 42, 503–523.
- Quadrioglio, L., Dessouky, M.M., Ordóñez, F., 2008. A simulation study of demand responsive transit system design. *Transport. Res. Part A* 42, 718–737.
- Uehara, K., Akamine, Y., Toma, N., Nerome, M., Endo, S., 2014. A proposal of a transport system connecting demand responsive bus with mass transit. In 2014 IEEE International Conference on Consumer Electronics-Taiwan, pp. 143–144.

Location Choice Models

Adam Wilkinson Davis, UC Davis Institute of Transportation Studies, Davis, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Location Choice Models	428
What are Location Choices?	428
How Do We Model Location Choices?	428
Discrete Choice Models in a Spatial Context	428
Addressing Spatial Autocorrelation in Location Choice Modeling	429
Measurement Units and the Modifiable Areal Unit Problem	429
Choice Sets for Location Choice Modeling	430
Summary	431
See Also	431
References	431

Location Choice Models

What are Location Choices?

People and households choose places to live, places to work, and places and times to shop, dine, and be entertained. Firms choose where to put stores, factories, and other facilities. Public agencies choose locations for large infrastructure pieces and routes for highways.

In making location choices, decision makers weigh attributes of each potential choice in order to come to the decision that works best for them. These attributes can be inherent to a particular location (like rent, size, or attractiveness) or dependent on the relationship between that location and other locations that matter to the decision maker (like proximity to work or distance to a distribution center). Location choices are also fundamentally about time: when making a long-term choice like where to buy a house or where to put a power plant, decision makers must choose a location that meets their needs both now and in the future; when making a shorter-term choice like where to buy groceries, decision makers must choose a location that is reachable within the available time using available means of transportation.

How Do We Model Location Choices?

Two of the methods transportation researchers commonly use to understand complex spatial choices are discrete choice models and spatial optimization. While both models assume that an optimal location can be identified, they come at this question from opposite directions based on whether the researcher wants to determine what underlying criteria decision makers use to make decisions or whether they want to identify what choice should be made given a known set of criteria. In somewhat more technical terms, discrete choice models use observed or stated choices to estimate what factors underlie the latent utility function that each decision maker is seeking to optimize; in contrast spatial optimization approaches start with information about what goes into the utility calculation and choose the location that provides the greatest benefit. This chapter focuses on discrete choice modeling applied to location questions, but some of the issues presented here (particularly those relating to spatial aggregation and the modifiable areal unit problem) apply equally to spatial optimization problems.

Discrete Choice Models in a Spatial Context

Discrete choice models are used to understand and predict people's behavior when choosing a single option from a menu of two or more opportunities. These models are based on the theory that when trying to decide what to buy, where to go, or how to get there, people consider all the available options and choose the one that brings them the most benefit, measured through a latent variable called "utility" (Ben-Akiva and Lerman, 1985). Discrete choice models separate utility into a systematic component and a random component, with the systematic component of utility modeled as a linear combination of attributes of each option and individual and a set of coefficients that correspond to people's preferences. Decision makers are assumed to choose whatever option has the highest value of utility, but because the true value of the random component of utility is unknown, the models instead provide the probability of each person choosing each option based on the assumed distribution of the error term and the relative values of the systematic component of utility provided by the various choices. While there are ways in which these models may not accurately or fully represent human decision-making, they can provide useful and meaningful analysis and predictions of the choices that people make.

Location choice involves choosing a single site from a menu of options, making discrete choice models a natural fit for them, but spatial applications of discrete choice models face two problems. First, the independence of irrelevant alternatives assumption is

violated under spatial autocorrelation, necessitating either a nesting structure or an autoregressive component to account for relationships between alternatives. Second, since space can be measured at a range of scales, some degree of aggregation is in order to produce a discrete choice set. Spatial aggregation tasks must take into account the modifiable areal unit problem and ecological fallacy. Additionally, the choice sets used for location choice modeling can be quite large and differ significantly between decision makers due to variable space-time constraints and variable choices can entail very large choice sets, which presents a further challenge to modelers.

Addressing Spatial Autocorrelation in Location Choice Modeling

Conventional choice models work under the assumption of Independence of Irrelevant Alternatives (IIA), which means that an individual's relative likelihood of choosing one of two options is exclusively a function of the utility they would derive from those two options and that there is no correlation in unobserved variables between alternatives. While this assumption has been addressed in various ways throughout the field of choice modeling, it is particularly problematic in spatial choice problems because of spatial autocorrelation, the tendency of nearby observations to be similar in both measured and unmeasured ways (Anselin, 2003). Spatial autocorrelation can be made less of an issue in models if it can be accounted for by explanatory variables, but residual spatial autocorrelation violates observation independence assumptions in many families of models.

One solution to the issue of spatial autocorrelation in choice models is the nested logit, which is designed to handle some degree of correlation between alternatives by relaxing the IIA assumption (Train, 2009). In this model, nearby locations could be grouped, in essence recasting the choice into a two- (or more) step process in which the decision maker first chooses among a set of larger areas and then chooses a specific location within each one. The nested logit solution is attractive because nested models are commonly used to understand other sorts of choices, but its rigid structure produces two significant shortcomings in this context. First, this model assumes that there is no relationship between choices in different branches of the hierarchy. This is acceptable when the choices can be easily divided into spatially distinct groups, but does not work when the choices are spread continuously over an area without major divisions, as in the case of modeling housing choices as falling into different neighborhoods in a city. Second, nested models assume that a single nesting structure works for the choice sets faced by all decision makers (Pellegrini and Fotheringham, 2002; Sener et al., 2011). Cross-classified choice models that assign each option to a position within multiple separate hierarchies may be sufficient to address some of these concerns (Bhat, 2000), but in general, a model that accounts for correlations between individual pairs of options is preferable.

Spatially correlated choice models may be a preferred solution because they account for the similarity of nearby observations in a more realistic fashion. These models include a term to estimate the strength of the relationship between the utility offered by neighboring choice options. Various formulations of spatially correlated choice models have been proposed and implemented, including the Spatially Correlated Logit, which extends the nested logit framework (Bhat and Guo, 2004); the Structured Spatial Effects Model, which uses spatial weights terms in a similar way to spatial econometrics models developed from time series models (Miyamoto et al., 2004); and the Generalized Spatially Correlated Logit, which allows for neighboring locations to be modeled on a continuous scale rather than all-or-nothing (Sener et al., 2011).

A key question for spatially correlated choice models is how to think about proximity. It is convenient to model spatial relationships as binary, such that for each location there is a fixed set of neighbors determined by polygon adjacency or a K nearest neighbors search, with all other locations classified as non-neighbors. In contrast, geographers generally think of spatial relationships as a continuous variable, often according to some decay function that decreases with distance, with nearer locations more closely related and more distant points less closely related (Anselin, 2003; Goodchild, 2010; Sener et al., 2011). The decay function used can have a profound impact on the model results, making it important to choose a function that accurately represents the spatial relationships present in the data. Guo and Bhat (2007) test housing location choice models based on three different neighbor weighting schemes, one based on adjacency of census spatial units, one based on Euclidean distance, and one based on network distance, and find that using network distance is the most realistic.

Measurement Units and the Modifiable Areal Unit Problem

Aggregation of individual data points into larger spatial units is a necessary step for many sorts of geographic analysis and is particularly important for discrete choice models for home, work, and activity locations, which require limited choice sets containing meaningful and distinct spatial units. In these sorts of location choices, the actual alternatives can be thought of as specific points in space that nest within and are modeled as polygonal spatial units in order to incorporate information about the surrounding area and limit the size of the choice set for reasons of practicality.

The process of aggregating locations to areas encounters two related geostatistical problems that present a challenge to accurately modeling spatial processes: the ecological fallacy and the modifiable areal unit problem. The ecological fallacy describes the case where the value of variables aggregated by groups is inaccurately taken to reflect the value of all observations within this group. This fallacy applies to any sort of aggregation, but it can be particularly problematic for spatial/location choices since the analysis scale chosen can have a tremendous bearing on what information is available. In general, there is a tradeoff between information availability (generally greater at coarser scales) and how representative that information is (generally more at finer scales). The modifiable areal unit problem (MAUP) describes how the way in which boundaries are drawn between units can change the results of analysis, and has been identified as a potential source of bias in location choice models (Guo and Bhat, 2004). MAUP is particularly a concern in location choice modeling if the true location choices are spatially clustered or tend to be located near the edges of spatial

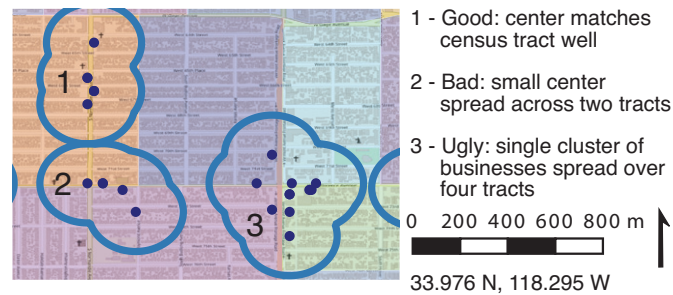


Figure 1 Three potential outcomes when businesses concentrated along roadways get mapped to census units. Census tracts shown in shaded colors.

units, which is discussed in more detail below. Since spatial aggregation is generally necessary for location choice modeling, care must be taken to choose spatial units that mitigate both of these effects.

Census spatial units (such as block groups and tracts) are commonly used in social and health sciences to define neighborhoods both because they are predefined and standardized and because they match the spatial resolution of other available datasets, but their size, shape, and boundaries may not be well-suited to measuring phenomena of interest to researchers (Bates, 2006; Weiss et al., 2007). Although they are designed to measure where people live, census units are often even a poor choice for delineating neighborhoods when studying housing in part because the census splits units at major roads, which assigns houses on opposite sides of the same street to separate units (Clapp and Wang, 2006).

In travel demand modeling, traffic analysis zones (TAZs) built up from census blocks are the typical spatial unit of analysis, but a variety of analytical issues arise from their use. Páez and Scott (2004) highlight two major issues with TAZs that relate to the MAUP: the scale effect, by which the same analysis can lead to different conclusions depending on the resolution of the spatial units; and the zoning effect, by which different spatial partitioning leads to a wide range of possible analytical outcomes. This casts doubts upon models that incorporate zone-level spatial relationships. Solutions to the problem of developing the “right” zoning system have been proposed that attempt to minimize some of the negative impacts of spatial aggregation, including some that focus on capturing travel within a single zone using a variety of scales (Moeckel and Donnelly, 2015; Viegas et al., 2009). Fig. 1 illustrates MAUP using business data and census tracts in Central Los Angeles.

Boundaries of these units are set based on “visible and identifiable features,” like arterial roads, which are often the site of many businesses. Commercial districts built along major roads and intersections are often split among multiple units. Fig. 1 shows three potential cases of business aggregation to tracts from an area in central Los Angeles. The first example is the best case for small commercial centers, because all the businesses are located near the centroid of a single tract. In example 2, the businesses in an area cluster along a street dividing two tracts. In example 3, a set of businesses clusters along the intersection of two major roads that divide four tracts, which means that a tract-based spatial aggregation would effectively separate those businesses by as much as half a kilometer in an area with high population density, and much more in an area with larger tracts. The bubbles on the map correspond to the search radius for destinations in each of the clusters that result from our clustering result.

Choice Sets for Location Choice Modeling

Spatial autocorrelation and the choice of spatial aggregation present problems for all modeling of spatial data, but developing comprehensive choice sets for location choice problems can present an additional difficulty that is not present in other modeling settings. Discrete choice models were originally developed to improve our understanding of mode choice, for which the range of alternatives is fairly tractable, but for location choice questions, the number of alternatives can quickly make model estimation infeasible. If measured at the scale of individual neighborhoods or census block groups, there can be hundreds of potential home locations in a major metropolitan area. Likewise potential destinations for shopping or dining can easily number in the hundreds.

The issue of choice set definition is innately tied to the choice of spatial scale and the issues of MAUP and the ecological fallacy discussed above. Modeling location choices presents a tradeoff between reality and practicality because the choice generally cannot be modeled at the finest and most realistic scale. The choice of home or destination is likely better understood as being made at the level of an individual house or a specific business establishment with information about the surrounding area as context, rather than at the level of the surrounding area with consideration for all of the opportunities available in that area, but it can be difficult to gather much usable data about specific houses or specific destinations. Given that spatial aggregation is generally necessary for data availability reasons, the question of scale becomes paramount. The ecological fallacy becomes more pronounced for larger aggregations, but it may be impractical to model choice sets with the large menu of alternatives provided by using smaller aggregations.

One proposed solution for the issue of overly large choice sets is to sample from the full choice set. Care must be taken to ensure that the sample choice set is sufficiently large and representative, and it may be necessary to apply a correction to the model to account for the subsampling (Guevara et al., 2014; Lee and Waddell, 2010; Lemp and Kockelman, 2012; Nerella and Bhat, 2004). Choice sets for destinations should also be limited by considering the temporal and mobility constraints facing the decision maker (Chen and Kwan, 2012; Wang and Miller, 2014).

Summary

Discrete choice models have been used to understand a wide range of location choices. While models for location choice encounter many of the same issues faced by choice models in other contexts, a number of problems are unique to choice modeling. Some of these issues, like the difficulty of handling the large choice sets present in many location analyses, make themselves unavoidable, whereas spatial autocorrelation, aggregation-related biases, and the links between space and time are often ignored.

Spatial autocorrelation, the tendency of nearby locations to be similar in unmeasured ways, violates assumptions about the independence of observations. In a location modeling context, spatial autocorrelation can be addressed using a nesting structure or incorporating a spatial weighting structure into the model. At minimum, it is vital to investigate spatial patterns in prediction accuracy or bias in order to determine whether further corrections for spatial autocorrelation are necessary.

The relationship between space and time (and between these and mobility) is particularly important for understanding location choices made on shorter time scales, like activity locations chosen over the course of a day. In these contexts, it's important to consider first what destinations are reachable within the available time and given the available means of transportation.

While location choices are often made at the level of an individual house or business establishment, it is generally necessary to aggregate locations into larger spatial units in order to decrease the size of the potential choice set and access data for a wider range of attributes. Using spatially aggregated data can introduce biases into models, particularly when the boundaries between units are chosen poorly or the units chosen are too large, but data availability can force the choice of a specific aggregation scale.

See Also

Activity-Based Models; The Use and Value of Geographic Information Systems in Transportation Modeling; Route Choice and Network Modeling; Departure Time Choice Modeling; Residential Location Choice Models

References

- Anselin, L., 2003. Spatial econometrics. In: Baltagi, B.H. (Ed.), *A Companion to Theoretical Econometrics*. Blackwell Publishing Ltd, Malden, MA, USA, pp. 310–330. <https://doi.org/10.1002/9780470996249.ch15>.
- Bates, L.K., 2006. Does neighborhood really matter? Comparing historically defined neighborhood boundaries with housing submarkets. *J. Plan. Educ. Res.* 26, 5–17.
- Ben-Akiva, M., Lerman, S.R., 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press Series in Transportation Studies. The MIT Press, Cambridge, MA.
- Bhat, C.R., 2000. A multi-level cross-classified model for discrete response variables.. *Transport. Res. B* 34 (7), 567–582. [https://doi.org/10.1016/S0191-2615\(99\)00038-7](https://doi.org/10.1016/S0191-2615(99)00038-7).
- Bhat, C.R., Guo, J., 2004. A mixed spatially correlated logit model: formulation and application to residential choice modeling. *Transp. Res. Part B Methodol.* 38, 147–168. [https://doi.org/10.1016/S0191-2615\(03\)00005-5](https://doi.org/10.1016/S0191-2615(03)00005-5).
- Chen, X., Kwan, M.-P., 2012. Choice set formation with multiple flexible activities under space–time constraints. *Int. J. Geogr. Inf. Sci.* 26, 941–961. <https://doi.org/10.1080/13658816.2011.624520>.
- Clapp, J.M., Wang, Y., 2006. Defining neighborhood boundaries: Are census tracts obsolete? *J. Urban Econ.* 59, 259–284. <https://doi.org/10.1016/j.jue.2005.10.003>.
- Goodchild, M.F., 2010. Twenty years of progress: GIScience in 2010. *J. Spat. Inf. Sci.* 2010, 3–20. <https://doi.org/10.5311/JOSIS.2010.1.32>.
- Guevara, C.A., Chorus, C.G., Ben-Akiva, M.E., 2014. Sampling of alternatives in random regret minimization models. *Transp. Sci.* 50, 306–321. <https://doi.org/10.1287/trsc.2014.0573>.
- Guo, J., C., Bhat, 2004. Modifiable areal units: problem or perception in modeling of residential location choice? *Transp. Res. Rec.* 1898, 138–147. <https://doi.org/10.3141/1898-17>.
- Guo, J.Y., Bhat, C.R., 2007. Operationalizing the concept of neighborhood: application to residential location choice analysis. *J. Transp. Geogr.* 15, 31–45. <https://doi.org/10.1016/j.jtrangeo.2005.11.001>.
- Lee, B.H.Y., Waddell, P., 2010. Residential mobility and location choice: a nested logit model with sampling of alternatives. *Transportation* 37, 587–601. <https://doi.org/10.1007/s11116-010-9270-4>.
- Lemp, J.D., Kockelman, K.M., 2012. Strategic sampling for large choice sets in estimation and application. *Transp. Res. Part Policy Pract.* 46, 602–613. <https://doi.org/10.1016/j.tran.2011.11.004>.
- Miyamoto, K., Vichiensan, V., Shimomura, N., Páez, A., 2004. Discrete choice model with structuralized spatial effects for location analysis. *Transp. Res. Rec.* 1898, 183–190. <https://doi.org/10.3141/1898-22>.
- Moeckel, R., Donnelly, R., 2015. Gradual rasterization: redefining spatial resolution in transport modelling. *Environ. Plan. B Plan. Des.* 42, 888–903. <https://doi.org/10.1068/b130199p>.
- Nerella, S., Bhat, C.R., 2004. Numerical analysis of effect of sampling of alternatives in discrete choice models. *Transp. Res. Rec.* 1894, 11–19. <https://doi.org/10.3141/1894-02>.
- Páez, A., Scott, D.M., 2004. Spatial statistics for urban analysis: a review of techniques with examples. *GeoJournal* 61, 53–67. <https://doi.org/10.1007/s10708-005-0877-5>.
- Pellegrini, P.A., Fotheringham, A.S., 2002. Modelling spatial choice: a review and synthesis in a migration context. *Prog. Hum. Geogr.* 26, 487–510. <https://doi.org/10.1191/0309132502ph382ra>.
- Sener, I.N., Pendyala, R.M., Bhat, C.R., 2011. Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior. *J. Transp. Geogr.* 19, 294–303. <https://doi.org/10.1016/j.jtrangeo.2010.03.013>.
- Train, K.E., 2009. *Discrete Choice Methods with Simulation*, second ed. Cambridge University Press, New York.
- Viegas, J.M., Martínez, L.M., Silva, E.A., 2009. Effects of the modifiable areal unit problem on the delineation of traffic analysis zones, effects of the modifiable areal unit problem on the delineation of traffic analysis zones. *Environ. Plan. B Plan. Des.* 36, 625–643. <https://doi.org/10.1068/b34033>.
- Wang, J., Miller, E.J., 2014. A prism-based and gap-based approach to shopping location choice. *Environ. Plan. B Plan. Des.* 41, 977–1005. <https://doi.org/10.1068/b130063p>.
- Weiss, L., Ompad, D., Galea, S., Vlahov, D., 2007. Defining neighborhood boundaries for urban health research. *Am. J. Prev. Med.* 32, S154–S159. <https://doi.org/10.1016/j.amepre.2007.02.034>.

Full Feedback and Equilibrium Modeling in Urban Travel Forecasting

Yu (Marco) Nie*, Jun Xie*, David Boyce*, *Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL, United States; *School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China; ‡University of Illinois at Chicago, Chicago, IL, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	432
Problem Setting	432
The Four-Step Procedure With Feedback	433
User Equilibrium Traffic Assignment	434
Combined Equilibrium Models	435
Combined Mode and Route Choices	435
Combined Model of Destination, Mode, and Route Choice	437
Future Research	439
References	439

Introduction

Designing and managing urban transportation systems require a capability to predict the future travel that will use these systems. How far in the future depends on what decision-making processes that such forecasts aim to support. For large-scale facilities such as urban freeways or rail transit lines, 20 or 25 years in the future have been regarded as appropriate, even though such a facility may well have a longer physical and economic life. At present developing countries, especially those in Asia, are actively developing these transportation infrastructure systems at large scale in order to accommodate their rapidly expanding urban population. This is in stark contrast with the mission of the developed world, which is largely focused on travel demand management measures, as well as repairs and incremental additions to their existing and aging transportation systems. These smaller investments and operational/management plans may require forecasts for shorter time horizons, typically ranging between 1 and 5 years. Effective and efficient decisions for planning and managing these systems require an advanced evaluation framework to provide information on the distribution of impacts on travelers, employers, and public agencies. Travel forecasts are central to such a framework.

Urban transportation studies began in Detroit in 1953, then in Chicago in 1955 and later in Pittsburgh and the tristate (New York City) region, all under the leadership of Carroll and Creighton (1957). Similar studies were initiated during this period, stimulated in part by the work of Glenn Brokke, William Mertz, Walter Hansen, and others at the Bureau of Public Roads (BPR), which belonged to US Department of Commerce (Boyce and Williams, 2015). Out of these early studies emerged a travel forecasting methodology known today as the sequential procedure or the four-step procedure in the United States. These steps include:

1. *Trip generation* creates the total amount of travel per time period (hour, day) that begins and ends at each location;
2. *Trip distribution* determines the amount of travel from every origin to every destination, forming an origin–destination (O-D) trip matrix;
3. *Mode split* defines the proportion of trips by private cars, trains, buses, cycles, walking, and other modes of travel; and
4. *Traffic assignment* allocates modal trip matrices to a set of routes to determine flows on road and transit links.

Of the earlier four steps, traffic assignment lies at the core of travel forecasting and is also the focus of this paper. It determines link flows on the network, which is the goal of the entire procedure. Traffic assignment also models congestion, which has become a dominating factor in planning urban transportations systems. Importantly, it is from the representation of congestion in traffic assignment that the concept of equilibrium and feedback loop first originates.

The remainder of this paper is organized as follows. First, we summarize the basic definitions, notation, and assumptions that are common to travel forecasting and network equilibrium modeling. Then in Section “The Four-Step Procedure With Feedback” we review the four-step procedure for travel forecasting and introduce the concept of “feedback.” Section “User Equilibrium Traffic Assignment” presents the basic equilibrium model for the traffic assignment problem. Section “Combined Equilibrium Models” introduces the so-called combined models to integrate the last three steps in the four-step procedure into a single model. Essentially, combined models replace the external feedback with internalized interactions between destination, mode, and route choices. The last section concludes this paper with a discussion of future research.

Problem Setting

Consider an urban region divided into small, relatively homogeneous zones. Zone size varies with the density of development, so that activity levels per zone are relatively similar. These zones will be referred to as traffic analysis zones (TAZ) hereafter. Urban

activities in TAZs are described in terms of: (1) residential population and households; (2) employment, education (primary, secondary, and higher), or day care; and (3) shopping, personal and business services, recreation, etc. Facilities that support urban activities consist of land and buildings for (1) residences, (2) workplaces, (3) schools, (4) shopping and service centers, and (5) parks and recreational facilities.

Travel occurs on two types of transportation systems or modes, namely roadway and public transport, which consist of three elements: equipment, infrastructure, and control policy. The roadway system is characterized by private vehicles (equipment), ways for driving-biking-walking (infrastructure), and traffic control system (control policy). For public transport, the three elements are, respectively, bus or train, roadway or railway, and operations plan. The roadway systems are represented as networks of nodes and links, and for public transport, routes of scheduled services. Service characteristics of links depend on fixed parameters related to physical characteristics such as (1) length, number, and width of lanes by type including bikeways and walkways; (2) grade; (3) public transport station spacing and vehicle performance; (4) control and operations plans, for example, speed limits, signal settings, road tolls; and (5) service frequencies, operating speeds, and public transport fares. Other variables are related to travel flows (demand), which also determine service characteristics. These include flows of cars, trucks, bikes, and pedestrians and public transport boarding and alighting at stops per unit time. Taken together, these variables determine the performance characteristics of individual links, conceptualized as:

$$\text{Link travel time} = f(\text{flows} | \text{fixed vehicle/way characteristics, and operations plans}).$$

The earlier relationship is commonly known as a cost-performance function or *link performance function*. It is worth noting here that such a performance function represents traffic congestion at an aggregated level for the average condition in a relatively long analysis period. It is not intended to capture the dynamic evolution of traffic flow.

Travel between daily activities (e.g., residing, working, eating, shopping, schooling, recreating) may be described in terms of pairs of activities linked by trips. According to the activity purpose, a *trip* may be classified as a home-work trip, a work-shop trip, a shop-home trip, and so on. In the trip-based approach, individual trips are aggregated by purpose and forecast as separate groups. Whether travel occurs by private car (either alone or with others), by public transport, bike, or walking, depends on the availability of modes, their relative service times and monetary costs, as well as intangible factors like comfort and convenience.

The timing of travel during the day depends on constraints imposed by activity schedules and the prevailing travel conditions in the transportation systems. Trips occurring in a 24-hour weekday are typically aggregated into several periods, such as the morning peak commuting period or the off-peak period. A transformation is usually needed to convert observed trips corresponding to a specific departure or arrival time into flows in each period (measured by person/unit time).

For transportation systems planning, facility design, operations planning, or conformance with air quality regulations, forecasts of the following variables are required for each scenario of transportation system/activity pattern:

1. Flows of private cars, trucks, bikes, pedestrians and public transport vehicles on the road network by morning and evening peak commuting periods, and for longer periods when travel conditions are relatively stable;
2. Flows of persons on the public transport network by sub-mode and by time period; and
3. Flows of persons from origin zone to destination zone by private vehicles and public transport and by time period.

A capability to examine changes in these flows in response to changes in network layout, capacity and service attributes, monetary costs (e.g., fuel, tolls, fares, parking fees), and changes in zonal activity levels is an essential requirement for planning urban transportation systems.

The Four-Step Procedure With Feedback

An early prototype of the four-step procedure was developed in Detroit and Chicago transportation studies in 1950s. The procedure consists of three parts: estimating trips generated from and attracted to each TAZ according to land use, predicting flows between each O-D zone pair, and transforming O-D flows into predicted traffic flows on links in each major type of transportation network. At the time, such studies primarily concerned roads, so only travel by car was forecasted. The procedure was soon extended to include travel by transit, and the “splitting” of those trips by mode, but nonmotorized travel was ignored.

Estimating the number of trips from land use became *trip generation* (Fig. 1) and is typically based on linear regression methods. Predicting future O-D flows was performed by a *trip distribution* procedure. Growth factor methods were initially employed for trip distribution, but they were later replaced by the concept of spatial interaction between points separated by travel distance or travel time, which were analogous to the gravity law of spatial interaction, and indeed became known as the *gravity model*. Early innovators in trip distribution modeling were Voorhees (1955), who introduced the empirical gravity concept into urban travel forecasting, and Schneider (1959), who derived and implemented an *intervening opportunities* model. *Mode choice* was initially based on the so-called diversion curves, until replaced by more sophisticated discrete choice models. The last step, that is, loading O-D flows onto the transportation networks, came to be known as *traffic assignment*.

Even in the 1950s, the creators of the early sequential procedure realized that an iterative solution process—which they termed *feedback*—is required for their method. They noted that (1) O-D flows are typically modeled as a function of travel time in trip distribution, and, when assigned to the road network, result in link flows and link travel times; but (2) these link times determine the

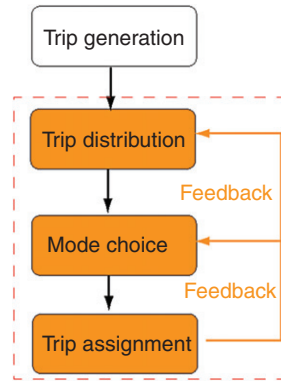


Figure 1 The four-step procedure with feedback.

travel times between any O-D pair, which in turn affect O-D flows. Thus, a feedback from traffic assignment to trip distribution is required. Since mode choice also employs the O-D travel times from traffic assignment as inputs, it too should receive a feedback. The feedback loops must be repeated to reach equilibrium and consistency between the input data and the final results. That is, the O-D travel times fed into the trip distribution or mode choice model should produce the same travel times after the procedure runs its course. However, a key question regarding the feedback scheme is whether or not it converges to a stable equilibrium when iterated. Guaranteeing the convergence poses a significant theoretical challenge that calls for different modeling approaches. We will discuss such an approach in Section “Combined Equilibrium Models.”

User Equilibrium Traffic Assignment

John Wardrop (1952), a British traffic scientist, proposed the following criterion to describe traffic flows on a network:

“The journey times on all routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route. . . . (this) criterion is quite a likely one in practice, since it might be assumed that traffic will tend to settle down into an equilibrium situation in which no driver can reduce his journey time by choosing a new route.” (Patriksson, 1994, p. 31)

The first sentence is now known as Wardrop’s first principle of *user equilibrium* (UE). By “routes actually used” Wardrop meant the routes used to travel from a given origin to a given destination, or an O-D pair. If routes consist of sequences of links, then route costs may be defined as the sum of the link costs along the route. Since link costs depend on link flows through the link performance function, and each link may serve many routes connecting different O-D pairs, identifying the route costs that satisfy Wardrop’s principle involves solving the route choice problem simultaneously for the entire network of many O-D pairs.

Link flows have units of vehicles/hour (vph), so route flows and O-D flows also have units of vehicles/hour or persons/hour. The resulting route choice model is a steady-state (or static) flow model that does not consider the physical movement of an individual trip from its origin to its destination. Rather, O-D flows occur with corresponding flows on the links of each used route. This formulation leads to a relatively simple concept of congestion: steadily flowing vehicles traveling at speeds determined by link performance functions. Clearly, it cannot accurately capture dynamic traffic phenomena, such as queuing behind bottlenecks, or delays propagating from one link to another.

The UE link flows and costs, and a set of route flows, corresponding to fixed O-D flows may be determined by solving the following constrained optimization problem (Beckmann et al., 1956; Sheffi, 1985, p. 60).

$$\min z(f) = \sum_{a \in A} \int_0^{x_a} c_a(w) dw \quad (1)$$

subject to:

$$\sum_{k \in K_{rs}} f_k^{rs} = q_{rs}, \quad \forall r \in R, s \in S, \quad (2)$$

$$f_k^{rs} \geq 0, \quad \forall k \in K_{rs}, r \in R, s \in S, \quad (3)$$

$$\sum_{r \in R} \sum_{s \in S} \sum_{k \in K_{rs}} f_k^{rs} \delta_{ak}^{rs} = x_a, \quad \forall a \in A, \quad (4)$$

where:

- x_a = flow of all vehicles on link a (vph),
- $c_a(x_a)$ = generalized travel cost function for link a , a nondecreasing function of link flow x_a ,
- f_k^{rs} = flow of vehicles on route k in the set of all routes K_{rs} connecting O-D pair r and s (vph),
- q_{rs} = exogenous flow of vehicles between O-D pair r and s (vph),
- $\delta_{ak}^{rs} = 1$, if link a belongs to route k connecting O-D pair r and s , and 0 otherwise,
- R, S = sets of origin and destination zones, respectively, and
- A = set of links in the network.

The unknown variables are vehicle route flows $f = (f_k)$. The vehicle link flows (x_a) are defined in terms of the route flows, through the exogenous link-route correspondence matrix (δ_{ak}^{rs}) .

The optimality conditions for the earlier problem may be stated as:

$$\sum_{a \in A} c_a(x_a) \delta_{ak}^{rs} - u_{rs} \geq 0, \quad \forall k \in K_{rs}, r \in R, s \in S, \quad (5)$$

$$f_k^{rs} \left(\sum_{a \in A} c_a(x_a) \delta_{ak}^{rs} - u_{rs} \right) = 0, \quad \forall k \in K_{rs}, r \in R, s \in S, \quad (6)$$

where u_{rs} is a dual variable associated with the conservation of flow constraint defined on the exogenous O-D flow q_{rs} . Conditions (6) may be interpreted as follows for O-D pair rs :

$$f_k^{rs} > 0 \Rightarrow c_k^{rs} = u_{rs}, \quad \forall k \in K_{rs}, \quad (7)$$

$$f_k^{rs} = 0 \Rightarrow c_k^{rs} \geq u_{rs}, \quad \forall k \in K_{rs}, \quad (8)$$

where c_k^{rs} is the travel cost on route k , the sum of the costs of the links comprising route k . Hence, every used route between an O-D pair rs has a (minimum) generalized travel cost equal to u_{rs} , and no unused route has a lower travel cost. Thus, this formulation corresponds to Wardrop's first principle.

If the generalized cost functions are strictly increasing with link flows, then the objective function in Eq. (1) is strictly convex, guaranteeing that the solution is unique in the link flows. However, the solution is not unique in the route flows, since the objective function is not strictly convex with respect to the route flows in general. The earlier formulation also requires each link performance function be *separable*, that is, it depends only on the link's own flow. If there are cross-link interactions, the applicability of the formulation depends on whether or not such interactions are *symmetric*, namely each link performance function depends on a specified vector of link flows such that the effect of a change in link a 's flow on link b equals the effect of a change in link b 's flow on link a for all links specified in the cost function. Models not exhibiting such symmetries are called *asymmetric*, and they can no longer be formulated as a minimization problem. Instead, more general formulations, such as those based on variational inequality or nonlinear complementarity problems, are often used.

Modeling route choice on a public transport network is generally more complex. An important difference between roadway and transit systems is that, whereas travel by car generally can begin on demand, travel by transit begins when the transit vehicle (bus, train) arrives at the point of service or service node (bus stop, station). Hence, transit riders incur waiting and transfer times that depend on the service timetable and the arrival times of passengers at the service node. These waiting times contribute to a different type of stochastic attribute than the one found in route choice by car drivers.

Generally, the objective of each transit rider is assumed to minimize his or her generalized travel cost, a weighted linear function of in-vehicle travel time, waiting and transfer times, and fare. Since the waiting and transfer times are not deterministic but depend on the service frequency and passenger arrival pattern, each passenger's objective is to minimize the expected generalized travel cost. If transit services are frequent, it is reasonable to assume that travelers arrive at their service nodes randomly with respect to the service schedule and take the first vehicle belonging to the set of "desired" transit lines. This set of lines and the rules that allow the traveler to reach his/her destination is usually called a *strategy* (Spiess and Florian, 1989). Developing strategy-based (or hyperpath-based) transit assignment models remains an active research area, and a particular challenge here is to account for congestion and crowding effects, which exhibit very different characteristics than those found in roadway systems. These more complex transit route choice models are not considered in combined equilibrium models discussed in the next section.

Combined Equilibrium Models

Combined Mode and Route Choices

If travel choices are viewed as a hierarchy, then route choice would seem to fall naturally at the lowest level. Route choices are specific to each transportation mode and may vary in every trip in response to current travel conditions. The choice of mode may be hypothesized to be the next higher choice in this hierarchy. Mode choices may be made on a daily basis, assuming a car is available. Choice of cycle and walking may also be considered, if available as options. Note that the representation of choices as a hierarchy does not imply that choices are made sequentially.

Instead of linking route and mode choices through a feedback as described in Fig. 1, a combined modeling approach attempts to integrate these choices in a same optimization problem (Boyce and Bar-Gera, 2003). To begin, the deterministic car route choice model is extended to include choice of mode. Let us assume that only two modes are considered: car and transit; the transit between any O-D pair has a fixed, flow-independent cost γ_{rs} , and transit vehicles do not impact road networks. Then, the aforementioned route choice problem can be generalized to the following combined mode and route choice problem:

$$\min z(\mathbf{f}, \mathbf{q}^t) = \sum_{a \in A} \int_0^{x_a} c_a(w) dw + \sum_{r \in R} \sum_{s \in S} q_{rs}^t \gamma_{rs} \quad (9)$$

subject to:

$$\sum_{k \in K_{rs}} f_k^{rs} + q_{rs}^t = q_{rs}, \quad \forall r \in R, s \in S, \quad (10)$$

$$\sum_{r \in R} \sum_{s \in S} \sum_{k \in K} f_k^{rs} \delta_{ak}^{rs} = x_a, \quad \forall a \in A, \quad (11)$$

$$f_k^{rs} \geq 0, q_{rs}^t \geq 0, \quad \forall k \in K_{rs}, r \in R, s \in S, \quad (12)$$

where q_{rs}^t denotes the flow using the transit mode between O-D pair rs . This formulation seeks to find the allocation of exogenous O-D flows to the two modes, and within the car mode to routes, so as to minimize an objective function constructed according to desired equilibrium conditions. Because car route flows should satisfy the UE conditions, the total cost of car travel is not minimized; rather, the sum of the integrals of the links cost functions is retained. The optimality conditions pertinent to route and mode choice include Eqs. (10–12) and:

$$q_{rs}^t (\gamma_{rs} - u_{rs}) = 0, \quad \forall r \in R, s \in S. \quad (13)$$

To interpret these conditions, note that the UE costs of the used car routes not only determine the O-D cost, but also determine whether any of the fixed cost transit mode has sufficiently low costs to be used: if $u_{rs} < \gamma_{rs}$, then $q_{rs}^t = 0$ (no one uses transit). If $u_{rs} = \gamma_{rs}$, then the O-D cost of transit and car are equal, and use of transit may occur. If $u_{rs} > \gamma_{rs}$, then all O-D flow occurs by transit, and no one uses car. That is, either there is transit flow or the transit cost sets a maximum level for the car costs for each O-D pair. Hence, the solution is “all-or-nothing” with respect to mode choice.

In survey data, one often observes travel by two or more modes (car, transit, cycle, or walk) between many O-D pairs. Therefore, the formulation of the mode and car route choice model as a deterministic optimization problem, while instructive, is unrealistic. The relaxation of this deterministic formulation can be achieved through the addition of a modal dispersion constraint. A function representing modal dispersion may be imposed to make mode choices more dispersed than the UE level, that is, some choices are allocated to higher cost modes. For example, the mode split between each O-D pair may be predicted using the logit model as follows:

$$q_{rs}^t = \frac{q_{rs} e^{-\theta \gamma_{rs}}}{e^{-\theta \gamma_{rs}} + e^{-\theta u_{rs}}}, \quad (14)$$

where θ is a dispersion parameter. The earlier equation can be rearranged as:

$$u_{rs} = \gamma_{rs} + \frac{1}{\theta} \ln \frac{q_{rs}^t}{q_{rs} - q_{rs}^t}. \quad (15)$$

If we view the transit line as a special link between the O-D pair with the travel cost defined by $\gamma_{rs} + \frac{1}{\theta} \ln \frac{q_{rs}^t}{q_{rs} - q_{rs}^t}$, then the combined mode choice/traffic assignment problem can be reformulated as follows:

$$\min z(\mathbf{f}, \mathbf{q}^t) = \sum_{a \in A} \int_0^{x_a} c_a(w) dw + \sum_{r \in R} \sum_{s \in S} \int_0^{q_{rs}^t} \left(\gamma_{rs} + \frac{1}{\theta} \ln \frac{w}{q_{rs} - w} \right) dw \quad (16)$$

subject to

$$\sum_{k \in K_{rs}} f_k^{rs} + q_{rs}^t = q_{rs}, \quad \forall r \in R, s \in S, \quad (17)$$

$$\sum_{r \in R} \sum_{s \in S} \sum_{k \in K_{rs}} f_k^{rs} \delta_{ak}^{rs} = x_a, \quad \forall a \in A, \quad (18)$$

$$f_k^{rs} \geq 0, \quad \forall k \in K_{rs}, r \in R, s \in S. \quad (19)$$

The second term in Eq. (16) is interpreted as a dispersion term for mode choice.

Combined Model of Destination, Mode, and Route Choice

The combined mode and route choice formulation can be further extended to include an O-D dispersion function in the same manner as described earlier (Boyce and Daskin, 1997). In the version presented here, constraints are added to the mode and route choice formulation to derive a classical gravity trip distribution model. These constraints consist of origin and destination constraints and another function representing dispersion of trips to higher cost destinations, separately from the modal dispersion constraint. We can similarly convert these constraints into a dispersion term (as in Eq. (15)) and add it into the objective function, which lead to the following formulation:

$$\begin{aligned} \text{Min } z(\mathbf{f}, \mathbf{q}) = & \sum_{a \in A} \int_0^{x_a} t_a(w) dw + \frac{1}{\theta_1} \sum_{r \in R} \sum_{s \in S} q_{rs} (\ln q_{rs} - 1) \\ & + \frac{1}{\theta_2} \sum_{m \in \{a, t\}} \sum_{r \in R} \sum_{s \in S} q_{rs}^m (\ln q_{rs}^m - 1) + \sum_{r \in R} \sum_{s \in S} \gamma_{rs} q_{rs}^t \end{aligned} \quad (20)$$

subject to:

$$\sum_{k \in K_{rs}} f_k^{rs} = q_{rs}^a, \quad \forall r \in R, s \in S, \quad (21)$$

$$\sum_{m \in \{a, t\}} q_{rs}^m = q_{rs}, \quad \forall r \in R, s \in S, \quad (22)$$

$$\sum_{s \in S} q_{rs} = O_r, \quad \forall r \in R, \quad (23)$$

$$\sum_{r \in R} q_{rs} = D_s, \quad \forall s \in S, \quad (24)$$

$$\sum_{r \in R} \sum_{s \in S} \sum_{k \in K_{rs}} f_k^{rs} \delta_{ak}^{rs} = x_a, \quad \forall a \in A, \quad (25)$$

$$f_k^{rs} \geq 0, \quad \forall k \in K, r \in R, s \in S, \quad (26)$$

$$q_{rs}^m \geq 0, \quad \forall m \in \{a, t\}, r \in R, s \in S, \quad (27)$$

where q_{rs}^a and q_{rs}^t represent the demand for car and transit modes, respectively, and θ_1 and θ_2 are the dispersion parameters corresponding to trip distribution and mode choice, respectively. Let u_{rs} , k_{rs} , μ_r , λ_s denote the multipliers of the constraints (21), (22), (23), (24), respectively. Then, the optimality conditions pertinent to mode choice and trip distribution are stated as follows:

$$\frac{1}{\theta_1} \ln q_{rs} + k_{rs} - \mu_r - \lambda_s = 0, \quad \forall r \in R, s \in S, \quad (28)$$

$$\frac{1}{\theta_2} \ln q_{rs}^a + u_{rs} - k_{rs} = 0, \quad \forall r \in R, s \in S, \quad (29)$$

$$\frac{1}{\theta_2} \ln q_{rs}^t + \gamma_{rs} - k_{rs} = 0, \quad \forall r \in R, s \in S. \quad (30)$$

Eqs. (29) and (30) lead to:

$$q_{rs}^a = e^{\theta_2(k_{rs} - u_{rs})}, \quad (31)$$

$$q_{rs}^t = e^{\theta_2(k_{rs} - \gamma_{rs})}. \quad (32)$$

The modal flows q_{rs}^m are constrained to sum to the O-D flow by the mode conservation of flow constraint. Thus, inserting q_{rs}^m into Eq. (22) yields:

$$e^{\theta_2 k_{rs}} = \frac{q_{rs}}{e^{-\theta_2 u_{rs}} + e^{-\theta_2 \gamma_{rs}}}, \quad (33)$$

from which the multiplier k_{rs} can be derived as:

$$k_{rs} = \frac{1}{\theta_2} \ln q_{rs} - \frac{1}{\theta_2} \ln(e^{-\theta_2 u_{rs}} + e^{-\theta_2 \gamma_{rs}}). \quad (34)$$

Inserting (34) back to (31) and (32) leads to the logit mode choice:

$$q_{rs}^a = \frac{q_{rs} e^{-\theta_2 u_{rs}}}{e^{-\theta_2 u_{rs}} + e^{-\theta_2 \gamma_{rs}}}, \quad (35)$$

$$q_{rs}^t = \frac{q_{rs} e^{-\theta_2 \gamma_{rs}}}{e^{-\theta_2 u_{rs}} + e^{-\theta_2 \gamma_{rs}}}. \quad (36)$$

Furthermore, by replacing k_{rs} in Eq. (28) with (34) we have:

$$\frac{1}{\bar{\theta}} \ln q_{rs} - \Gamma_{rs} - \mu_r - \lambda_s = 0, \quad (37)$$

where $\bar{\theta} = \frac{\theta_1 \theta_2}{\theta_1 + \theta_2}$ and $\Gamma_{rs} = \frac{1}{\bar{\theta}} \ln(e^{-\theta_2 u_{rs}} + e^{-\theta_2 \gamma_{rs}})$ may be interpreted as the total disutility associated with both modes for the O-D pair rs . An expression for the O-D flow q_{rs} can then be derived from (37) as:

$$q_{rs} = e^{\bar{\theta}} (\Gamma_{rs} + \mu_r + \gamma_s). \quad (38)$$

By substituting q_{rs} into Eqs. (23) and (24) we have:

$$\begin{aligned} e^{\bar{\theta}} \mu_r &= \frac{O_r}{\sum_{l \in S} e^{\bar{\theta}} (\lambda_l + \Gamma_{rl})} \equiv A_r O_r, \\ e^{\bar{\theta}} \lambda_s &= \frac{D_s}{\sum_{l \in R} e^{\bar{\theta}} (\mu_r + \Gamma_{ls})} \equiv B_s D_s, \end{aligned} \quad (39)$$

which allows us to rewrite q_{rs} in the form of the conventional gravity model as:

$$q_{rs} = A_r B_s O_r D_s e^{\bar{\theta}} \Gamma_{rs}. \quad (40)$$

The earlier formula predicts that the flow between an O-D pair increases with the product of the total trip generation and attraction, and decreases with the disutility associated with travel between the O-D pair. Alternatively, we can also express q_{rs} only using the first relationship in Eq. (39), that is,

$$q_{rs} = \frac{O_r e^{\bar{\theta}} (\lambda_s + \Gamma_{rs})}{\sum_{l \in S} e^{\bar{\theta}} (\lambda_l + \Gamma_{rl})}. \quad (41)$$

Then, by substituting the O-D function for q_{rs} into the O-D-mode function for q_{rs}^m , the combined O-D-mode function may be stated as:

$$\begin{aligned} q_{rs}^a &= \frac{O_r e^{\bar{\theta}} (\lambda_s + \Gamma_{rs})}{\sum_{l \in S} e^{\bar{\theta}} (\lambda_l + \Gamma_{rl})} \cdot \frac{e^{-\theta_2 u_{rs}}}{e^{-\theta_2 u_{rs}} + e^{-\theta_2 \gamma_{rs}}}, \\ q_{rs}^t &= \frac{O_r e^{\bar{\theta}} (\lambda_s + \Gamma_{rs})}{\sum_{l \in S} e^{\bar{\theta}} (\lambda_l + \Gamma_{rl})} \cdot \frac{e^{-\theta_2 \gamma_{rs}}}{e^{-\theta_2 u_{rs}} + e^{-\theta_2 \gamma_{rs}}}. \end{aligned} \quad (42)$$

The earlier result can be interpreted as a two-level nested logit choice model on the part of the traveler (Fig. 2), where the first layer determines the destination and the second determines the mode.

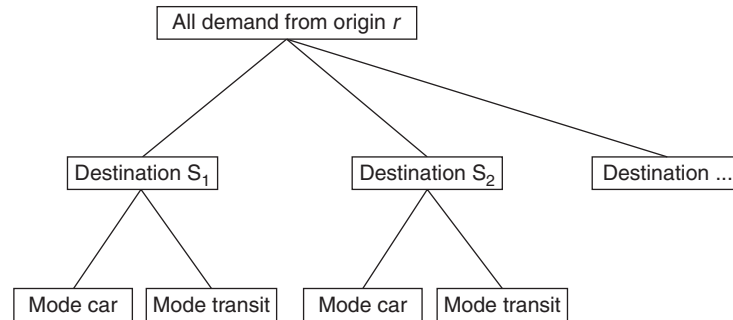


Figure 2 Interpretation of the combined model as a two-level nested logit choice model.

Future Research

Despite more than 60 years of research on urban transportation network equilibrium, many problems remain unsolved, and practice lags increasingly behind research knowledge. Research problems may be broadly classified according to travel choices, network representation, network design, and solution procedures, among others.

1. Until now, the modeling of travel choices has mainly followed the trip-based paradigm in practice. Increasingly, travel demand modelers view travel in terms of daily tours or daily activities. Generally, prediction of tours or activities has been approached from a micro-simulation point of view. Yet, the simulation-based approach encounters many analytical and practical challenges—computation resources, data requirements, calibration issues, to just name a few—that have so far hindered its broad adoption.
2. The representation of travel cost functions in network equilibrium models has advanced little beyond the original separable formulation of Beckmann. Although asymmetric models can be formulated as variational inequality problems, the lack of uniqueness of solutions has discouraged serious efforts to investigate this approach further. The paradigm shift to a dynamic representation of the underlying physical world has been pursued with enthusiasm by both researchers and practitioners in the past 3 decades. Yet, the real impact of this movement on the urban travel forecasting practice remains unclear.
3. Another problem that has received little attention, at least in practice, is transportation network design. In a combinatorial sense, the network design problem is intractable because of its large size. One way to address this challenge is to choose from a set of alternative designs generated using a systematic and semiautomated approach. Such a method would require an ability to distinguish among the merits of closely related scenarios. Given the recent advances in high-performance solution algorithms to network models, this capability may now be achievable. Other approaches are possible, however, such as *continuous approximation* (an example is the spacing of freeways in a grid, as was considered early in the history of this field).
4. Although practitioners are required by government regulations in the United States to solve their sequential travel forecasting procedures in a way that achieves an internal consistency of travel costs, there is no general agreement on how this should be done. No practitioner or software developer has described, tested, and demonstrated that one procedure is best among alternatives for the complex models typically applied in practice. Moreover, the errors introduced in forecasts that do not achieve consistency remain unknown. This relatively straightforward research problem should be tackled.
5. Academic interest in the solution of combined model formulations of travel choice has influenced travel forecasting practice, but only to a limited extent. Except for Santiago, Chile, combined models have rarely been applied in practice. The formulation of a combined model clearly enhances the understanding of the challenges facing the practitioner in solving the sequential procedure. Few practitioners, however, seem equipped by their training or mathematical ability to gain from insights from these formulations. Moreover, software developers have not incorporated tools in their software systems to facilitate the application of this approach. Based upon past experience, they will not do so until interest among practitioners strongly induces them to proceed.

Pursuit of the earlier research agenda requires a knowledge of optimization methods, computer skills, data, and perseverance. Many similar problems could be identified, especially from the viewpoint of other perspectives. For those so inclined, the journey will be challenging, but always interesting, and hopefully rewarding.

References

- Beckmann, M., McGuire, C.B., Winsten, C.B., 1956. *Studies in the Economics of Transportation (Part I)*. Yale University Press, New Haven, CT.
- Boyce, D., Bar-Gera, H., 2003. Validation of urban travel forecasting models combining origin-destination, mode and route choices. *J. Reg. Sci.* 43, 517–540.
- Boyce, D.E., Daskin, M.S., 1997. Urban transportation. In: ReVelle, C., McGarity, A. (Eds.), *Design and Operation of Civil and Environmental Engineering Systems*, Wiley, New York, pp. 277–341.
- Boyce, D., Williams, H., 2015. *Forecasting Urban Travel*. Edward Elgar, Cheltenham.
- Carroll Jr., J.D., Creighton, R.L., 1957. Planning and urban area transportation studies. Highway Research Board Proceedings of the Thirty-Sixth Annual Meeting, Highway Research Board, Washington, DC, pp. 1–7.
- Patriksson, M., 1994. The Traffic Assignment Problem—Models and Methods. VSP, Utrecht.
- Schneider, M., 1959. Gravity models and trip distribution theory, *Papers Reg. Sci. Assoc.* 5, 51–56.
- Sheffi, Y., 1985. *Urban Transportation Networks*. Prentice-Hall, Englewood Cliffs, NJ.
- Spieß, H., Florian, M., 1989. Optimal strategies: a new assignment model for transit networks. *Transp. Res.* 23B, 83–102.
- Voorhees, A.M., 1955. A general theory of traffic movement. 1955 Proceedings of the Institute of Traffic Engineers, ITE, New Haven, CT, pp. 46–56.
- Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. *Proc. Inst. Civil Eng. Part II* 1, 325–378.

The Use and Value of Geographic Information Systems in Transportation Modeling

Ming Zhang, The University of Texas at Austin, Austin, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	440
GIS Data Integration and Handling	440
GIS Enhancing Four-Step Modeling and Other Modeling Applications	442
GIS in/for Four-Step Transportation Modeling Procedures	442
GIS for Accessibility Modeling	443
LUTI Models and GIS	444
GIS Visualization for Transportation Modeling	444
New GIS for New Mobility Modeling	446
Conclusions	446
References	446

Introduction

Geographic information systems (GIS) refer to computer systems for capturing, storing, managing, analyzing, and presenting geographically referenced data. Transportation modeling has an inherent geographic component in it as it aims to characterize and predict (using mathematical and statistical tools) the movement of people and goods over space and time. While GIS and transportation modeling have been developing in their respective fields separately, the geographic focus of GIS and the geographic nature of transportation make the use of GIS in transportation modeling both a logical choice and a motivation for enhancement (Lewis, 1990; Waters, 1999; Loidl et al., 2016). Public agencies and private consulting firms in transportation have adapted GIS widely for the interest of sharing and managing data sources, supporting transportation modeling and plan-making, and informing public engagement and investment decisions (e.g., USDOT, 2016).

GIS may apply to transportation modeling in different stages of modeling work, including data assembling from multiple input sources, calibration of model parameters and search for best-fit model specifications, and communication of modeling results with the public and policy makers (Slavin, 2004). There are generic commercial GIS packages, for example, ArcGIS developed by ESRI (2019). Institutional or private sector users may customize the system to develop extensions to the built-in functions and modules for specific applications. Specialized GIS packages, for example, TransCAD (Caliper, 2019), PTV (PTV Group, 2019), and CUBE (Citilabs, 2019), offer fully integrated procedures of transportation modeling in addition to the basic GIS data processing and visualization functions. GIS scholars and transportation professionals have also developed GIS applications for modeling specific topics of transportation interest, for instance, accessibility modeling and land use-transportation integrated (LUTI) models. These GIS applications may operate as stand-alone programs or may be loosely coupled with existing GIS packages.

This chapter introduces the use and value of GIS in transportation modeling. The rest of the chapter includes four parts. Part 1 explains how the topological data structure underlying vector GIS design fits well the needs of data handling for transportation modeling. Part 2 describes GIS applications to the standard procedures of four-step transportation models and to specific transportation modeling interests, including accessibility modeling and LUTI modeling. Part 3 highlights the visualization feature of GIS and illustrates how geo-visualization has contributed to advancing transportation modeling. Lastly, Part 4 discusses the potential of and challenges to GIS in and for transportation modeling amid rapid development of new information and communications technologies (ICTs) and burgeoning new markets for shared economy and shared mobility as well as autonomous/driverless vehicles (AV/DV). The chapter ends with concluding remarks.

GIS Data Integration and Handling

The fundamental concern of transportation is the movement of people and goods. Transportation modeling takes a quantitative approach to assess the condition of people and goods movement in the existing transport system, predict the system performance under assumed future conditions, and identify needs for policy interventions. Hence, adequate data input is crucial to transportation modeling. GIS presents a capable tool to integrate spatial and nonspatial data in a unified platform for enhanced transportation modeling.

Typically transportation modeling utilizes three types of data (Cambridge Systematics, Inc., et al., 2012): (1) socioeconomic and demographic data that typically exist in aggregate areal units, for instance, traffic analysis zones (TAZs), zip codes, and counties; (2) infrastructure and system network data that display linear features, for instance, highway and street segments, transit routes, and pipelines, or in point format, for instance, street intersections as network nodes, transit station or stop placements, and locations of traffic incidents; (3) performance data such as vehicle flows, transit ridership, congestions, accidents, and emissions that can be reported in various

geographic locations and spatial scales. Data for model validation are also critical to transportation modeling. Quality transportation modeling relies on quality data input from various sources. It is the variety of input data in formats, metrics, and standards that creates a start-up challenge before essential modeling work begins. GIS integrates spatial and nonspatial data from various sources through the application of a topologically referenced relational data structure (Galdi, 2005; Longley et al., 2011).

Essentially GIS represent spatial features in three basic types of data objects (Chang, 2018): point (node), line (arc), and polygon (area). Depending on the spatial scale, a geographic feature can be represented in different type of features. For example, a city may appear as a point in the large scale of national map layer. In a detailed map scale, the city may become a polygon object to show the area of city limits. The three basic types of objects are geographically referenced and topologically encoded: points (nodes) are defined by their geographic coordinates (longitude and latitude) through geographic projections; lines are vectors defined by the starting and ending points (nodes); multiple vector lines define polygons (areas). Each of the spatial objects has its unique data table to record the spatial attributes, for instance, the X-/Y-coordinates of points (nodes), the IDs of the From-/To-nodes and the linear length of lines, and the area of polygons (Fig. 1). The three basic types of objects are interrelated as defined by their topological

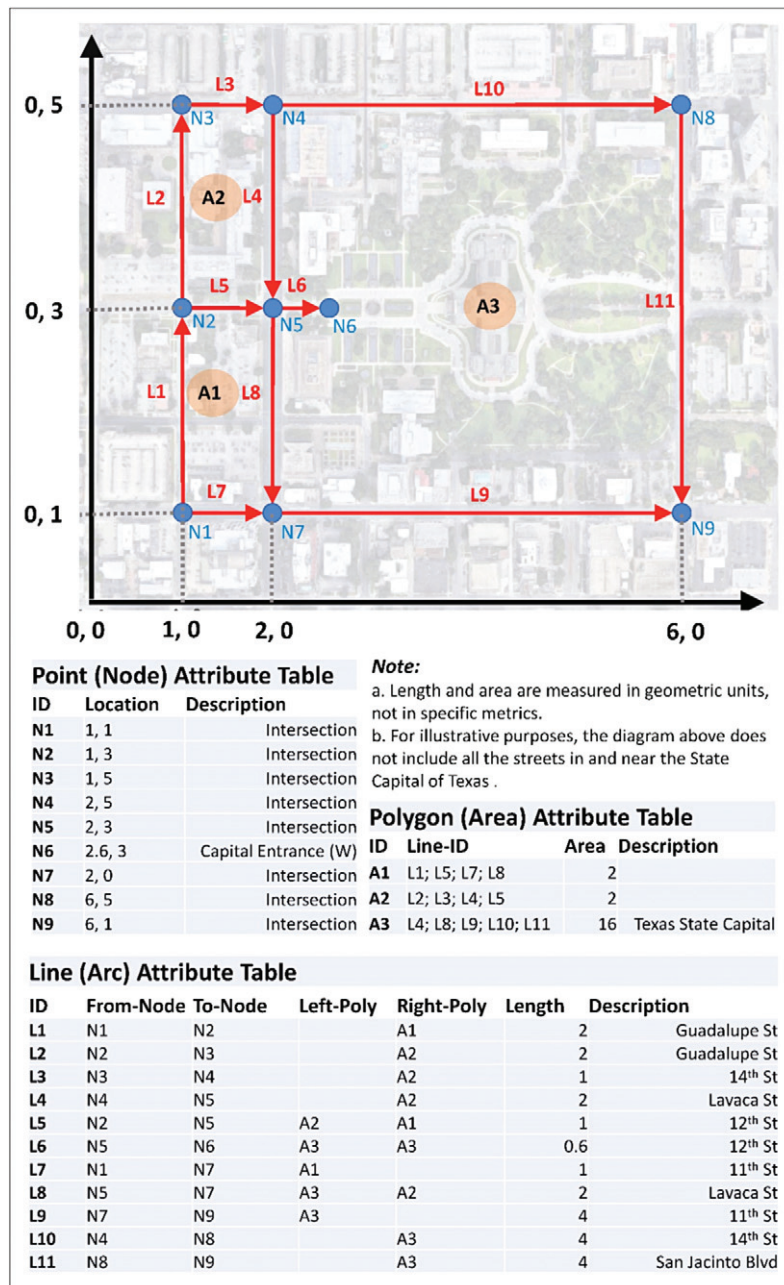


Figure 1 Illustration of topological representation of selected city blocks near the state capital of Texas. Source: Author.

relationship (Hoel, 2009). This data topology allows for representation and examination of spatial relationship within and between these objects in terms of dimensionality, adjacency, connectivity, contiguity, and containment.

The relational database structure also enables the attribute tables for the spatial objects to be related to auxiliary data tables, for example, information on the name, capacity, and service hours of a train station attached to the point where the station is located, bus frequency and service capacity information attached to the attribute tables of the lines with transit routes; total population and employment for census tracts or TAZs related to polygon features (e.g., zipcode areas and TAZs). Data integration achieved by GIS enables cross-data source referencing, integrating data between spatial and nonspatial and different spatial data types of the same spatial layer, or datasets from different spatial layers. For instance, for transit catchment area analysis, GIS's built-in tools allow different ways to perform spatial buffering analyses, by overlaying line-to-polygon, point-to-polygon, or polygon-to-polygon analysis. The analyses can be conducted using Euclidian distances or network-based distances. The buffering function allows spatial buffering analysis to be carried out using any variable of the analyst's interest. For modeling applications, the cross-referencing capability of GIS offers flexible choices between disaggregate travel modeling, for example, individual based trip generation and mode choice analyses, and aggregate, area-based operation such as gravity modeling of origin-destination interactions.

There are specialized software tools developed for specific data management, graphics, and statistical analysis functions. For example, dBase, MS Access, and Oracle Database specialize in data processing and management. AutoCAD and Rhino are popular computer-aided design and drafting tools. SAS, MATLAB, and Stata are long-standing powerful tools for statistical analysis. The value of GIS lies in its capability to integrate the three main functions of database management, graphics, and statistical analysis into one platform, although each of the three functions does not offer the full capacities that their respective specialized software tools do (Sharma, 2011).

GIS Enhancing Four-Step Modeling and Other Modeling Applications

GIS in/for Four-Step Transportation Modeling Procedures

Specialized GIS such as TransCAD, PTV, and Cube, target transportation planning and modeling applications. These packages make the prevailing transport modeling procedures, for example, the four-step models fully operable in a GIS environment. Data processing, especially spatial data, was not readily available in the past for transportation modeling. In the early stage of transportation modeling, when spatial data representation and processing was quite constrained, spatial data input was simplified. For example, land use characteristics were represented in a limited way by the density of population and employment at the TAZ level. Another value of GIS for model improvement is that GIS makes it less onerous to perform spatial statistics to address explicitly the issues such as spatial dependence and the modifiable area unit problems (MAUP) (Loidl et al., 2016).

The first step, trip generation, models the number of trips for the predefined geographic unit of analysis, typically TAZ, for the study region. This step reports two types of trip-making outcome: trips produced from and attracted to each TAZ.

For estimating trip generation rates, Metropolitan Planning Organizations (MPOs) often conduct activity-travel surveys that record and report the exact location and time information for each activity taken by the survey respondent during the survey day. The location information may come in as X-/Y-coordinates or street addresses. GIS enables identification and mapping of the activities through the built-in address-matching and other geo-coding procedures. By linking traveler activities with fine-grained urban form measures, GIS offers an effective tool to improve trip generation modeling greatly compared with the TAZ-based aggregate estimates (Clifton et al., 2015).

The second step, trip distribution, takes trip production–attraction tables as input and converts them into matrices of trip origin–destination (O-D). Most commonly, the step applies gravity-type models to distribute the trips estimated for each TAZ to other TAZs in the rest of the region according to the relative attractiveness of the TAZs and the travel time impedance among them. While the relational database structure employed by most GIS package enables integration of datasets from multiple, different data sources, the flat table-based data storage and retrieval makes it inefficient to handle matrix data commonly seen in transportation modeling applications. For example, a 1000 by 1000 O-D matrix would take one million rows in a flat table. Generic GIS packages in the market remain rather limited in providing matrix operation functions. Specialized GIS for transportation applications have developed quite sophisticated matrix handling capabilities.

The third step, mode split/mode choice, models shares or probabilities of various travel modes, that is, driving alone, carpool, transit, and walking or biking, for the trips estimated from step two. The output of this step contains O-D matrices of trips by different modes for different trip purposes. Modeling for mode split/mode choice may be carried out at the individual or household level or at the aggregate level such as TAZ. Hagen-Zanker and Jin (2013) introduce an adaptive zoning method to enhance the accuracy of mode choice modeling at the city or city-region scale. Rather than using a fixed boundary definition for TAZs, the method utilizes GIS to define interactively TAZs and generate TAZ data for modeling travel mode choice. The authors tested and showed that their GIS-enabled adaptive zoning approach retained information during rezoning and improved modeling accuracy by a factor of six to eight.

The last step, traffic assignment, of the four-step modeling process involves simulating the performance of transportation networks that are assigned with the trips of different modes and reports network conditions, mostly measured by level of service (LOS) as a function of traffic volume over capacity ratio, when all trips are loaded to the network. A variety of algorithms are available for the modeler to choose from for the task. Transportation networks consist of interconnected nodes and links that represent system components such as intersections and stops, highway lanes and transit lines. The network representation builds upon classic graph

theory. Vector GIS' topological data structure corresponds well with the graph theory-based network representation of transportation systems for model building.

For traffic assignment modeling, finding the shortest path (based on time, distance, cost, and other impedances) must take into consideration of turn impedances at network nodes and transfer impedances between routes. The model performance relies on efficient handling of spatial and statistical data across multiple spatial data layers. GIS presents unique capacities to strengthen and enhance the modeling performance compared with the purely statistical procedures (Shaw, 1993).

The four-step modeling procedures were originally developed without GIS. Many transportation professionals and consulting firms continue to use non-GIS computer packages for the four-step modeling. A converging trend is that these traditional transportation modeling packages evolve with enhanced visualization and spatial data processing capabilities while the generic GIS packages grow with more and more analytical modules for transportation applications (Slavin, 2004).

GIS for Accessibility Modeling

While the fundamental concern of transportation is the mobility of people and goods, the ultimate goal of transportation lies in providing access to services and opportunities essential to the economy and people's quality of life. Accessibility has thus been a central concept and indicator for assessing transportation system performance and for making investment and policy decisions. The evolving GIS capabilities contribute to accessibility modeling for a wide range of policy interest and applications (Liu and Zhu, 2004; Benenson et al., 2010; Lei and Church, 2010; Mavoa et al., 2012; Long, 2017). Conceptually, accessibility denotes to the ease to reach destination opportunities such as jobs, stores, schools, and health services. While social, economic, and political factors could become access barriers or facilitators, modeling accessibility typically considers two factors: destination attractiveness and transportation costs to reach the destination. Depending on modeling purposes, destination attractiveness can be measured by the number of jobs, the size of schools, stores, or medical facilities, etc. Measures of transportation costs for accessibility modeling may take a variety of forms (Miller, 2018).

A simplified accessibility model considers transportation system performance only, quantifying accessibility by access time or monetary cost without taking into consideration of variations in origin and destination attributes (Allen et al., 1993). The most widely used approach to model accessibility takes the form of gravity model (Hansen, 1959), incorporating the quantity and quality of destination opportunities. Discounting the attractiveness of the opportunities by the cost of reaching them, while the discounting function, also called friction function, varies in math expressions. Individual-based accessibility modeling quantifies a traveler's accessibility as the total utility associated with all travel means available to her/him. The value of accessibility can be derived from disaggregate choice models, that is, the logsum (Ben-Akiva and Lerman, 1985). Miller (1991, 1999) made a major breakthrough in accessibility research. He reconceives accessibility around Hägerstrand's (1970) space-time prism concept and operationalizes it in a GIS environment. Jarv et al. (2018) proposed and tested a conceptual framework of dynamic location-based accessibility modeling (Fig. 2).

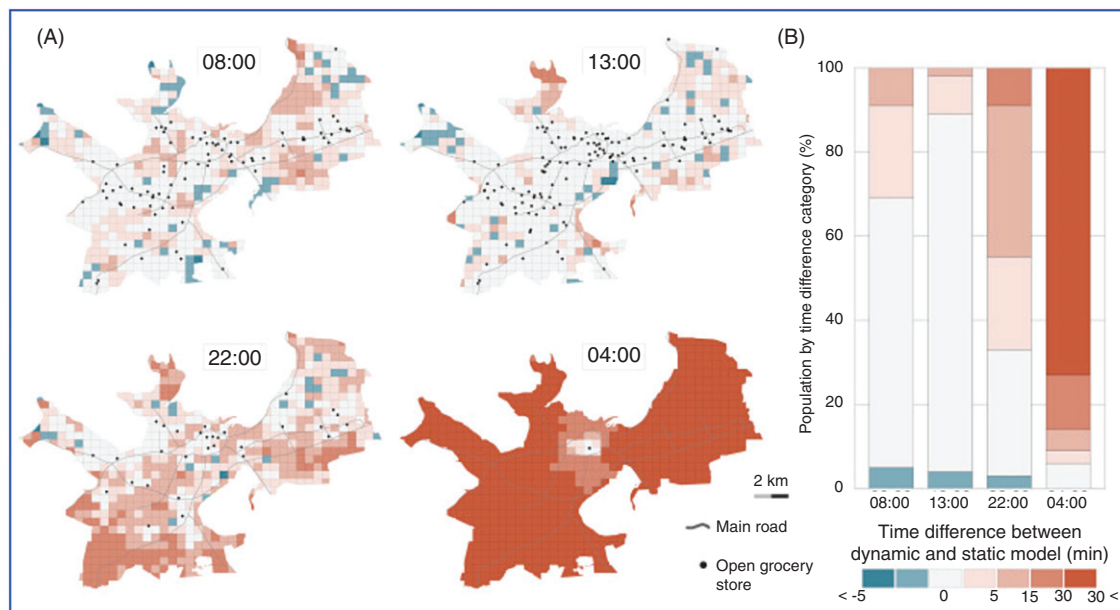


Figure 2 Spatial-temporal modeling of accessibility. The maps (A) show differences in travel time to the closest grocery store at different times of the day. The graphs (B) show the population in the study area aggregated to given travel time difference categories. Source: Jarv et al. (2018).

While these various forms of accessibility models have been implementable computationally, GIS-based accessibility modeling tools have been developed to facilitate the application of them. Ford et al. (2015) also utilized OSM data, along with other data input from conventional sources, to develop an accessibility analysis tool. This tool takes advantages of many built-in features in ArcGIS and enables practically convenient and flexible operations to assess accessibility conditions by different transportation modes at current conditions and future infrastructure scenarios. O'Sullivan et al. (2000) developed a platform on desktop GIS for accessibility modeling. This tool uses input from shortest path in a transportation network to analyze patterns of spatial access (Fig. 2).

LUTI Models and GIS

Increasingly, state and local authorities are required to assess broad issues beyond vehicle movement to address the social and environmental dimensions of mobility challenges such as equity, emissions, air quality, and energy security. LUTI models are expected to offer analytical and modeling capacities to assist decision-makers for accountable assessments of mobility conditions and better-informed investment choices. LUTI models that sprang up initially in the 1960s have gained renewed interest recently, mainly due to the availability of spatial data and advances in computing technologies to process the data (Moeckel et al., 2018). These models offer sophisticated procedures not available in the existing commercial GIS packages such as ArcGIS, PTV, or TransCAD. Yet geographic information is essential to the new generations of LUTI models.

Traditional transportation modeling does not consider adequately the influence of land uses and other aspects of the built environment on travel outcome. Population and employment densities in most cases serve as the only representation of land use. Growth forecasts for future population and employment are done externally. In most cases, the modeling procedures consider no feedback of land-use changes (households' and firms' location and relocation moves) resulting from transportation improvements such as travel demand management (TDM), increased capacities of highway or transit, and applications of new technologies. Furthermore, traditional transport models operate largely with aggregate data at the TAZ level; they are rather insensitive to changes in local built-environmental conditions that have significant impacts on cycling and walking travel. LUTI modeling offers great potential to overcome the many limitations of traditional transportation modeling. Known examples of LUTI models include MEPLAN (Echenique et al., 1969; Echenique, 2004), MUSSA (Martinez, 1996), IRPUD (Wegener, 1996), PECAS (Hunt and Arabaham, 2003), and UrbanSIM (Waddell, 2002).

GIS Visualization for Transportation Modeling

The embedded cartographic functions in GIS enable effective and efficient visual presentation and communication of input data, intermediate analysis results, and the final output of various transportation modeling tasks. Traditional transportation modeling systems do offer graphing tools. Most of them, however, are limited to dot and line graphs. GIS offers way more options to visualize spatial and nonspatial data input and output. For example, thematic mapping allows the analyst to generate maps to visually present the attributes associated with transportation system components or performance conditions. For instance, the analyst may create dot-density maps to illustrate and compare the spatial distribution of population and jobs for the base year and the forecasting year. Desire-line maps can show O-D flows of passenger and goods movement, more intuitively informing than the original data stored in matrix files (Fig. 3). Traditionally, the analyst has used a letter scheme A~F to indicate highway LOS and congestion conditions. With GIS, LOS can be vividly presented using line-width maps based on the modeled volume-to-capacity (V/C) ratios.

Aside from 2-D mapping, GIS offers 3-D visualization functions. Kwan (2000) introduced several GIS-based methods to geovisualize in a 3D environment the spatial and temporal patterns of individual activities. She illustrates the application of these methods with travel diary data collected in the Portland, OR metropolitan region. Datasets obtained from activity-travel surveys record quite precise locations (e.g., X-/Y-coordinates) of activities performed by individuals in different time of day. GIS enables visualization of the disaggregate space-time data in a 2-D or 3-D environment (Kwan, 2000).

GIS-based transportation simulation is another area in which visualization plays an essential role. Transportation simulation refers to the imitation of people and vehicle movement in transportation systems under a virtual environment. There are many stand-alone transportation simulation software packages developed for academic research or professional applications. Examples of GIS-integrated tools include Cube Dynasim, PTV Vissim, and TransModeler, which are commercially available. These simulation tools model the behavior of transportation system and recreate travel reality in a two-dimensional or three-dimensional GIS environment. The range of transportation simulation applications vary widely. For example, one may simulate in a great detail and with high fidelity the performance of a street intersection involving through and turning vehicle traffic, bicycle and pedestrian flows, and their interactions with each other on the road space as well as their responses to traffic control devices such as traffic lights, stop and yield signs, and surface pavements.

These simulation tools may also apply to region-wide systems to simulate operational performance of proposed projects and plans. Examples of simulation applications include dynamic route choice behavior, public transportation uses as well as car and truck traffic under different traffic management policy schemes such as electronic toll collection, route, and traffic guidance under real-time traffic detection and surveillance. Traffic simulation results can also feedback to travel demand forecasting to improve transportation planning.

The power of GIS visualization goes beyond making beautiful maps. Visualization enables visual thinking to detect spatial patterns, generate hypotheses, and construct knowledge (Nollenburg, 2007). Visual thinking and communication involve cognitive

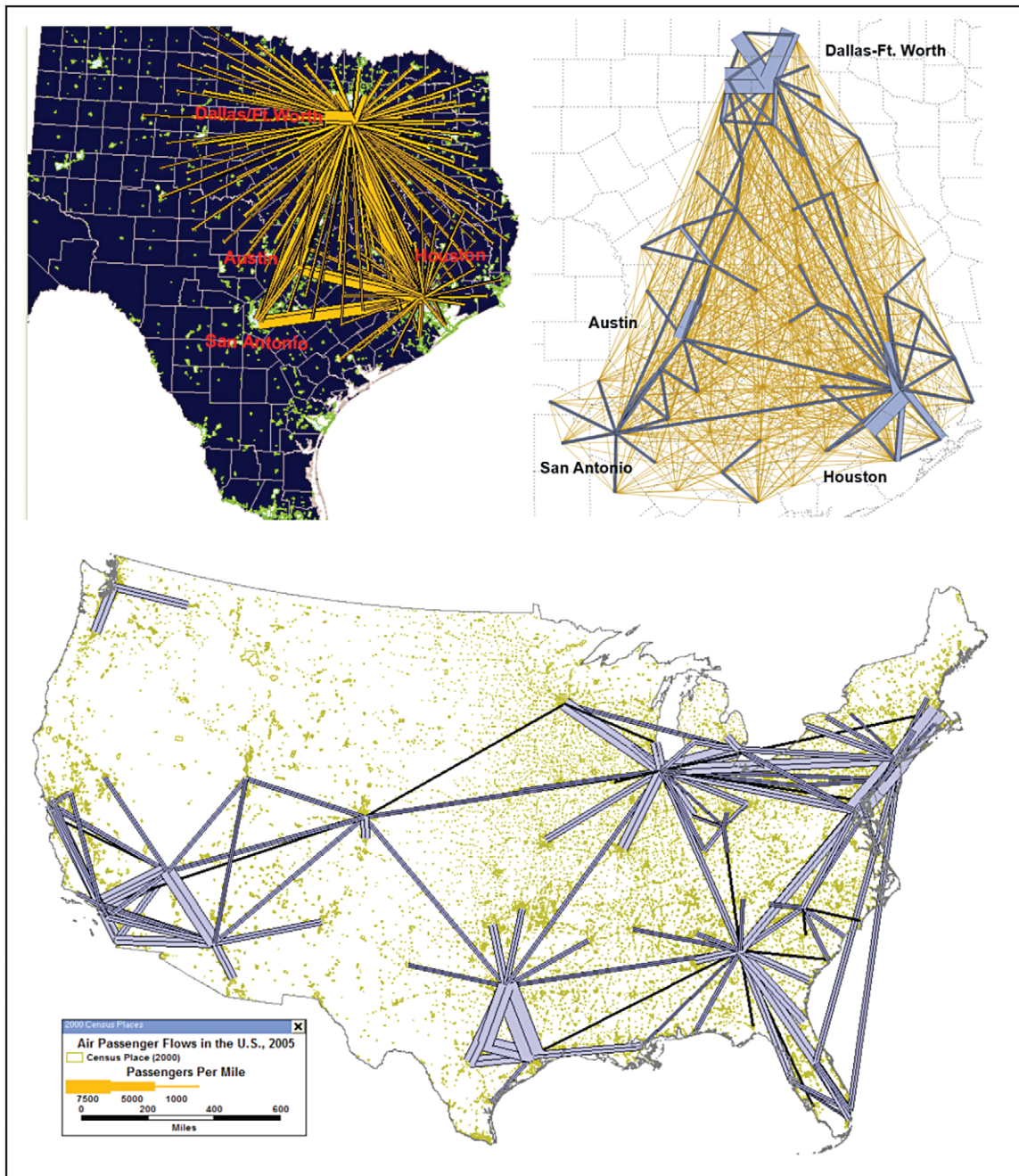


Figure 3 Visualization of transportation data in GIS: truck flow in Texas (*upper left*), super commute in Texas metros (*upper right*), and air passenger density in the United States. Source: Author.

processing of visual objects and information. Using GIS to display efficiently spatial data sets, the analyst can explore, understand, and convey spatial phenomena in ways that using words and statistics may find quite limited (**Fig. 3**).

Growing interest in geovisualization in the last 30 years have been driven by three forces (**Nollenburg, 2007**): (1) rapid advance in graphics and display technology owing to the availability of low-cost 3D graphics hardware and the development of highly immersive 3D virtual environments; (2) the dramatically increasing amount of geospatial data in a variety of forms, numbers, texts, images; and (3) the rise of the Internet-based medium for public engagement, which is important for both governmental agencies and business companies.

New GIS for New Mobility Modeling

Fast developing ICTs have facilitated the rise of shared economy and new mobility, which in turn propel the advance of GIS and transportation in their respective fields and present both challenges and opportunities for application of GIS to transportation modeling.

The first new challenge and opportunity is the growth of big data. Big data has emerged due to the widespread use of devices such as smartphones, GPS, transit smart cards, credit cards as well as the sensors, video cameras, and user-volunteered geographic information (VGI) available on the web via social media, Facebook, Twitter, and Instagram. These data have space–time information stored directly or inferable, offering the potential to broaden the scope and improve the quality of transportation modeling. Ziemke et al. (2017) demonstrate advantages of use VGI to overcome the data barriers to accessibility analysis and application by conventional accessibility models. This measure does not require population and land use (measures of opportunities at destinations) and network traffic performance data input as the conventional accessibility models do. Instead, the measure developed by Ziemke et al. (2017) relies exclusively on information extracted from OpenStreetMap (OSM). From a case study of Nelson Mandela Bay in South Africa, Ziemke et al. (2017) compare his approach with that using conventional approach and show comparable quality of insights into accessibility patterns but with operational and analytical advantages. Aside from lower input data requirements, the OSM-based accessibility measure offers the properties of higher interpretability, portability, and applicability to policy discussion and planning applications.

By definition, public transit (e.g., buses and subways) and taxi are traditional forms of shared mobility. New forms of shared mobility refer to a mobility service strategy that enables users to gain temporal access to transportation modes on an as-needed basis with no ownership of the transportation means (Shaheen and Chan, 2015). A variety of new forms of shared mobility services have emerged and grown, for instance, carsharing, stationed and dockless bicycle sharing and scooter services, ridesharing, ride-hailing, and transportation network companies (TNC) services, as well as future autonomous vehicles. Mobility-As-A-Service (MaaS) offers yet another mode of new mobility. MaaS is data-driven, powered by the growth of smartphone-based App development. MaaS integrates all modes of transportation such as carshare, bus, subway, taxi, and bikeshare and offers real-time optimized travel combinations as a single service package to the end users. Real-time geographic information, often crowdsourced, is key to the MaaS infrastructure (Goodall et al., 2017).

These new forms of mobility services rely on real-time, fast, and reliable information infrastructure. Real-time GIS offers information infrastructure support to the new forms of transportation and business services. Different from traditional GIS platforms that are static and stationed, real-time GIS offers information services capable of gathering, analyzing, and displaying streaming data from the Internet of Things (IoT) (Li et al. 2019). These data streams generated continuously by sensors, devices, and social media feeds are of high location and temporal granularity.

Conclusions

Transportation modeling has benefited from GIS greatly, while they evolve and advance in their own respective fields. GIS offers valuable contributions to transportation modeling mainly in three aspects. The first lies in the GIS platform that integrates the spatial and nonspatial input data from multiscale and multifunctional sources; quality and rich data input are essential to enhancing model performance. Second, GIS facilitates model calibration through refined spatial and temporal representation of transportation objects (e.g., people, vehicles, and infrastructure) and movements. In addition, GIS enables the modeler to consider explicitly the influence of spatial phenomena (e.g., spatial autocorrelation) on modeling results, which has been largely ignored in the traditional modeling work. Third, GIS offers powerful visualization capabilities for the modeler to communicate effectively the intermediate and final modeling output. GIS visualization is valuable not only in better informing the public and decision maker, but also empower the analyst to blend spatially oriented visual thinking with quantitative reasoning, gaining insights into complex travel behavior and transportation phenomenon.

The rapid growth of big-data industry and shared economy/shared mobility places new and increasing challenges and opportunities to transportation studies. GIS is also evolving from data processing as information tools to knowledge building as information science (GIScience) in an increasingly information-rich time (Goodchild, 2009; Onsrud and Kuhn, 2016). GIScience will continue to contribute to transportation modeling in a wide range of aspects, from basic data integration to dynamic, real-time optimization, and service delivering.

References

- Allen, W.B., Liu, D., Singer, S., 1993. Accessibility measures of U.S. metropolitan areas. *Transp. Res. B* 27B (6), 439–449.
- Ben-Akiva, M., Lerman, S.R., 1985. *Discrete Choice Analysis: Theory and Application to Predict Travel Demand*. MIT Press, Cambridge.
- Benenson, I., Martens, K., Rofé, Y., Kwartler, A., 2010. Public transport versus private car GIS-based estimation of accessibility applied to the Tel Aviv metropolitan area. *Ann. Reg. Sci.* 47, 499–515. doi:10.1007/s00168-010-0392-6.
- Caliper, Corp., 2019. TransCAD: Transportation Planning Software. <https://www.caliper.com/tcovu.htm> (Accessed 30 October 2019).
- Cambridge Systematics, Inc., Vanasse Hangen Brustlin, Inc., Gallop Corp., Bhat, C., Shapiro Transportation Consulting, LLC, Martind/Alexiou/Bryson, PLLC. 2012. NCHRP Report 716: Travel Demand Forecasting: Parameters and Techniques. Transportation Research Board: Washington, D.C. Available from: <https://www.nap.edu/download/14665>.

- Chang, K., 2018. Introduction to Geographic Information Systems, 9th ed. McGraw-Hill Education, New York.
- Citilabs, 2019. CUBE: Transportation & Land-Use Modeling. <https://www.citilabs.com/software/cube/> (Accessed 30 October 2019).
- Clifton, K., Currans, K., Muhs, C., 2015. Adjusting ITS's trip generation handbook for urban context. *J. Transp. Land Use* 8 (1), 5–29.
- Echenique, M.H., 2004. Econometric models of land use and transportation. In: Hensher, D.A., Button, K.J. (Eds.), *Transport Geography and Spatial Systems. Handbook 5 of Handbooks in Transport*. Pergamon/Elsevier Science, Kidlington, UK, pp. 185–202.
- Echenique, M.H., Crowther, D., Lindsay, W., 1969. A spatial model for urban stock and activity. *Reg. Stud.* 3, 281–312.
- ESRI, 2019. What is ArcGIS? . <https://developers.arcgis.com/labs/what-is-arcgis/> (Accessed 30 October 2019).
- Ford, A., Barr, S., Dawson, R., James, P., 2015. Transport accessibility analysis using GIS: assessing sustainable transport in London. *ISPRS Int. J. Geo-Inf.* 4, 124–149, <https://doi.org/10.3390/ijgi4010124>.
- Galdi, D., 2005. Spatial Data Storage and Topology in the Redesignated MAF/TIGER System. U.S. Census Bureau, Geography Division.
- Goodall, W., Dorey, F., Bornstein, J., Bonthron, B., 2017. The rise of mobility as a service: reshaping how urbanites get around. *Deloitte Rev.* (20), 113–129. <https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/consumer-business/deloitte-nl-cb-ths-rise-of-mobility-as-a-service.pdf>.
- Goodchild, M.F., 2009. Geographic information systems and science: today and tomorrow. *Ann. GIS* 15 (1), 3–9, doi:10.1080/19475680903250715.
- Hagen-Zanker, A., Jin, Y., 2013. Adaptive zoning for transport mode choice modeling. *Trans. GIS* 17 (5), 706–723.
- Hägerstrand, T., 1970. What about people in regional science? *Pap. Region. Sci.* 24, 7–24.
- Hansen, W.G., 1959. How accessibility shapes land use. *J. Am. Inst. Planners* 25, 73–76.
- Hoel, E., 2009. Topological data models. In: Liu, L., Özsu, M.T. (Eds.), *Encyclopedia of Database Systems*, Springer, Boston, MA.
- Jarv, O., Tenkanen, H., Salonen, M., Ahas, R., Toivonen, T., 2018. Dynamic cities: location-based accessibility modeling as a function of time. *Appl. Geogr.* 95, 101–110.
- Kwan, M.-P., 2000. Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. *Transp. Res. Part C: Emerg. Technol.* 8, 185–203.
- Lei, T.L., Church, R.L., 2010. Mapping transit-based access: integrating GIS, routes and schedules. *Int. J. Geogr. Inf. Sci.* 24, 283–304.
- Lewis, S., 1990. Use of geographic information systems in transportation modeling. *ITE J.* 3, 34–38.
- Li, W., Batty, M., Goodchild, M., 2019. Real-time GIS for smart cities. *Int. J. Geogr. Inf. Sci.*, doi:10.1080/13658816.2019.1673397.
- Liu, S., Zhu, X., 2004. An integrated GIS approach to accessibility analysis. *Trans. GIS* 8, 45–62.
- Loidl, M., Wallentin, G., Cyganski, R., Graser, A., Scholz, J., Haslauer, E., 2016. GIS and transport modeling—strengthening the spatial perspective. *ISPRS Int. J. Geo-Inf.* 5, 84, <https://doi.org/10.3390/ijgi5060084>.
- Long, J., 2017. Modelling accessibility. In: Wilson, John P. (Ed.), *The Geographic Information Science & Technology Body of Knowledge 3rd Quarter 2017 ed.*, doi:10.22224/gistbok/2017.3.7.
- Longley, P., Goodchild, M., Maguire, D., Rhind, D., 2011. *Geographic Information Systems and Science*, 3rd ed. John Wiley & Sons, Inc, New Jersey.
- Mavoa, S., Witten, K., McCreanor, T., O'Sullivan, D., 2012. GIS based destination accessibility via public transit and walking in Auckland. *J. Transp. Geogr.* 20, 15–22.
- Miller, E., 2018. Accessibility: measurement and application in transportation planning. *Transp. Rev.* 38 (5), 551–555.
- Miller, H.J., 1991. Modelling accessibility using space-time prism concepts within geographical information systems. *Int. J. Geogr. Inf. Syst.* 5, 287–301.
- Miller, H.J., 1999. Measuring space-time accessibility benefits within transportation networks: basic theory and computational procedures. *Geogr. Anal.* 31 (2), 187–212.
- Moeckel, R., Garcia, C., Chou, A., Okrah, M., 2018. Trends in integrated land-use/transport modeling: an evaluation of the state of the art. *J. Transp. Land Use* 11 (1), 463–476.
- Nollenburg, M., 2007. Geographic visualization. In: Kerren, A., Ebert, A., Meyer, J. (Eds.), *Human-Centered Visualization Environments*. Springer-Verlag, Berlin, Heidelberg, pp. 257–294.
- Onsrud, H., Kuhn, W. (Eds.), 2016. *Advancing Geographic Information Science: The Past and Next Twenty Years*. GSDI Association Press, Needham, MA.
- O'Sullivan, D., Morrison, A., Shearer, J., 2000. Using desktop GIS for the investigation of accessibility by public transport: an isochrone approach. *Int. J. Geogr. Inf. Sci.* 14, 85–104.
- PTV Group, 2019. PTV Visum. <https://www.ptvgroup.com/en-us/solutions/> (Accessed 30 October 2019).
- Shaheen, S., Chan, N., 2015. Mobility and the Sharing Economy: Impacts Synopsis. Shared-Use Mobility Definitions and Impacts, Special Edition. Transportation Sustainability Research Center. http://innovativemobility.org/wp-content/uploads/2015/11/SharedMobility_WhitePaper_FINAL.pdf (Accessed 31 October 2019).
- Sharma, R., 2011. GIS versus CAD versus DBMS: what are the differences? *Int. J. Comput. Sci. Commun. Technol.* 3 (2), 686–691.
- Shaw, S.L., 1993. GIS for urban travel demand analysis: requirements and alternatives. *Comput. Environ. Urban Syst.* 17, 15–29.
- Slavin, H., 2004. The role of GIS in land use and transport planning. In: Hensher, D., Button, K., Haynes, K., Stopher, P. (Eds.), *Handbook of Transport Geography and Spatial Systems*. Elsevier, Amsterdam, pp. 329–356.
- USDOT, 2016. GIS Strategic Plan for the U.S. Department of Transportation. <https://www.transportation.gov/mission/open/gis-strategic-plan> (Accessed 30 October 2019).
- Waddell, P., 2002. UrbanSim: modeling urban development for land use, transportation and environmental planning. *J. Am. Plann. Assoc.* 68 (3), 297–314.
- Waters, N.W., 1999. Transportation GIS: GIS-T. In: Longley, P.A., Goodchild, M.F., Maguire, D.J., Rhind, D.W. (Eds.), *Geographical Information Systems: Principles, Techniques, Management and Applications*. Wiley, New York, pp. 827–844.
- Ziemke, D., Joubert, J., Nagel, K., 2017. Accessibility in a post-apartheid city: comparison of two approaches for accessibility computations. *Netw. Spatial Econ.* 1–31, doi:10.1007/s11067-017-9360-3.

Latent Demand and Induced Travel

Charisma F. Choudhury, Institute for Transport Studies, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Definition	448
Implications	449
Methods for Predicting Induced Travel	449
Econometric Method	449
Transport Model-Based Analyses	450
Case Study-Based Analyses	450
Empirical Findings	450
Outstanding Issues	451
References	451
Further Reading	451

Definition

In economics literature, the term latent (or suppressed) demand is used to refer to the situation in which a consumer is unable to satisfy his/her needs due to nonexistence of the product or lack of awareness of the existence of the product and/or lack of affordability for purchasing the product. In the context of transport, latent demand specifically refers to the activities and travel that are desired but unrealized because of external constraints (Clifton and Moura, 2017). For example, non-car users may have a desire to visit a destination only accessible by car that constitutes the latent demand for that destination; the latent demand can be realized after instating public transport connectivity to the destination. Similarly, there can be a latent demand to purchase a car which is realized as the price of the car drops and/or the affordability of the consumer increases; there can be a latent demand to make more frequent social trips which is realized if there is a decrease in the congestion level; there may be a latent demand for increased mobility among the elderly and/or disabled population which can be realized when the self-driving cars become widely available (Wadud et al., 2016).

A closely related term in economics is induced demand which is the increase in demand observed as a result of the increase in supply and the associated reduction in generalized cost^a. In the context of travel, investments in transport typically lead to improvements in the levels-of-service in the short term which encourages people to travel more. This can occur through a variety of behavioral mechanisms including mode shifts, route shifts, redistribution of trips, generation of new trips, and long-run land-use changes that create new trips and/or longer trips (Noland, 2001). The “new” travel spurring from new or improved transport infrastructure and operational measures is known as induced travel. It may be noted that, in the network level, the “new” travel should only include the travel above and beyond those made originally to avoid overestimation.

Although the transportation literature uses the terms latent and induced demand somewhat interchangeably, in strict definitions, latent demand represents currently desired demand that is not realized because of a wide variety of constraints (including nonexistence of a mode or technology, lack of purchasing power, etc.) while induced travel would not occur without transportation system improvements and is stimulated specifically by the capacity enhancements.

The graphical representation of induced travel after a transport investment leading to enhancement of capacity (i.e., supply) is shown in Fig. 1. In Fig. 1A, it is assumed that the population and demographic effects (such as employment rates), which can also lead to an increase in demand levels, are held constant. The line S_1 represents the original supply, and C_1 and Q_1 are the corresponding costs and quantity of travel (e.g., vehicle miles traveled (VMT)), respectively. Capacity expansion (or another form of transport investment) lowers the cost of travel to C_2 and the supply is shifted downward (S_2) (since the same demand is met at a lower cost) leading to an increase in travel (Q_2). The increase in the quantity of travel from Q_1 to Q_2 represents the induced travel effect.

In Fig. 1B, the changes in population and/or economy are also accounted for. The demand curve is assumed to shift outward from D_1 to D_2 (e.g., due to population and/or economic growth in the area). Simultaneous to this, the supply curve shifts down. As a result, the quantity of travel (Q_3) increases even more than in Fig. 1A. In this case, the increase in quantity due to exogenous growth is the difference between Q_3 and Q_2 and the induced travel is the difference between Q_2 and Q_1 .

^a In transport economics, the generalized cost is the sum of the monetary and nonmonetary costs (e.g., travel time) associated with travel translated to monetary unit.

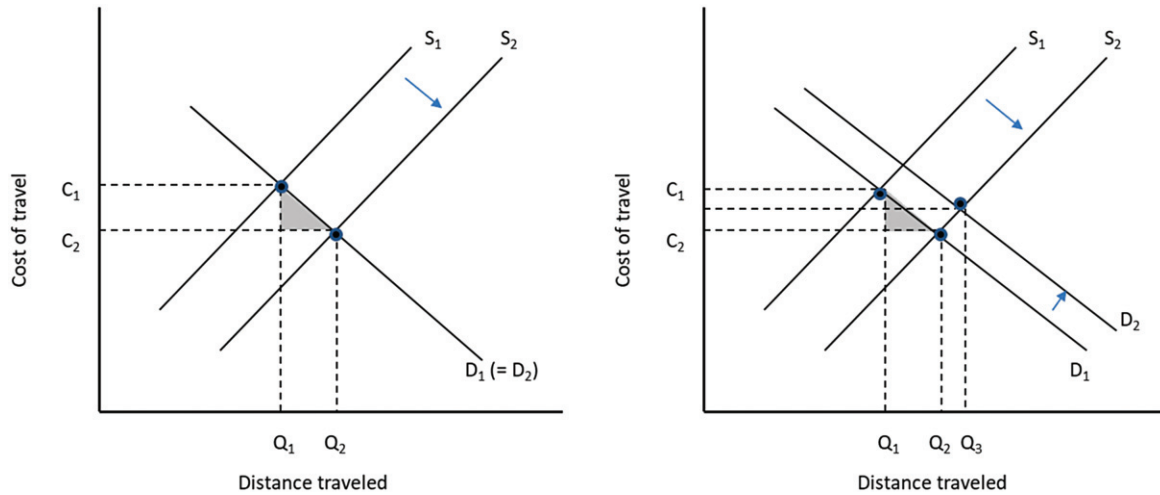


Figure 1 Graphical representation of induced travel. (A) With no population and/or economic growth, (B) accounting for population and/or economic growth.

Implications

In the short run, when residential and employment locations are fixed, improvements in transport infrastructure and/or operational measures can lead to changes in travel departure times and/or destinations, route switches, mode switches, and/or additional trips. Among these, changes in mode and increase in trip numbers have a direct contribution to induced travel. Change in departure time (usually a shift from off-peak to peak) can free up capacity at other times of the day leading to new trips being made at those times and have an indirect contribution to induced travel. On the other hand, choosing a different destination and/or route can result in either shorter or longer distances being traveled. If the net effect of these changes is increased travel, it also contributes to induced travel (Noland, 2001).

Further, in the long run, faster speeds encourage additional social and economic activities in areas made more accessible by the new highway capacity. This can lead to changes in land use, residential and employment locations, and in many cases, an increase in trip lengths in the long run.

The theoretical benefits due to the transport investment are hence offset and often completely diminished due to the latent demand and induced travel. The finding has even been dubbed the “Fundamental Law of Road Congestion,” which asserts that the elasticity of VMT with respect to lane mileage is equal to one, implying that driving increases in exact proportion to highway capacity additions (Downs, 1962). It may be noted though that if the traffic is diverted from other routes, it is expected to lead to a decrease in congestion elsewhere and in network-level analyses; this should be factored in to avoid overestimation.

In terms of wider benefits, given transport is a derived demand, meeting latent demand can contribute to improving the quality of life of the travelers by offering increased mobility and better access to employment and other opportunities. These net benefits of the travelers from transport investments can be measured by measuring the changes in consumer surplus^b (the shaded triangle in Fig. 1).

Methods for Predicting Induced Travel

The latent demand and induced travel can be predicted by econometric and transport model-based approaches. Further, historical evidence based on before–after analyses is also used to approximate latent demand in similar projects.

Econometric Method

In the econometrics studies, the relationship between induced travel (measured by VMT) and road capacity (generally measured by an increase in lane-miles) is established, typically using regression approach (Department of Transport, 2018):

$$\text{Log}(\text{VMT}_{it}) = a_i + \beta_t + \gamma \text{Log}(\text{lane-mile}_{it}) + \sum_k \delta_k X_{kit} + \varepsilon_{it} \quad (1)$$

^b Consumer surplus is the difference between the price that consumers pay and the price that they are willing to pay. On a supply and demand curve, it is the area between the equilibrium price and the demand curve.

where subscripts i and t refer to area and time period, X represents a set of k control variables. Using logarithmic forms of the dependent VMT and lane-mile variables means that the coefficient γ can be interpreted as the elasticity of VMT with respect to lane-mile, the proportionate response in demand to an increase in road capacity^c.

However, transportation planners, for example, do not randomly select which highways to widen. Instead, they prioritize improving highway segments with unacceptable levels of traffic congestion or in areas expecting economic growth. Therefore, though travel is highly correlated with capacity, it is not plausible to assume that causality flows in a single direction. Rather, highway capacity itself is endogenously^d determined by the volume of vehicle travel and other factors (Hymel, 2019). Failing to control for endogeneity in a travel demand model will likely generate biased estimates of the induced demand elasticity. Instrumental variables (IVs) are used to avoid endogeneity bias in predicting VMT. Commonly used IVs include lagged values of highway capacity growth (Fulton et al., 2000), the amount of urban land area (Noland and Cowart, 2000), and a combination of political and environmental measures (Cervero and Hansen, 2002).

Transport Model-Based Analyses

In transport model-based analysis, models to simulate land use and/or travel demand behavior, calibrated with existing data, are used to predict the travel behavior in post-investment scenarios. The induced travel is reported as percentage changes in traffic relative to the baseline in most cases rather than the change in vehicle-miles (though the latter can be calculated from most models as well). This makes the model-based studies difficult to compare with the elasticities reported in econometric studies. It may be noted though the background traffic growth and reassigned traffic need to be controlled for to get meaningful estimates from the model outputs. Further, the models typically involve some assumptions and there are often concerns associated with the spatial and temporal transferability of model parameters. These make it challenging to generalize the values derived from modeling studies.

Case Study-Based Analyses

Similar to the modeling studies, in the case studies approach, the induced travel associated with a specific road network improvement is reported as a percentage change in traffic relative to the baseline (i.e., change in vehicle numbers) since quantifying changes in VMT is extremely challenging in the real world. Due to potential confounding with other factors affecting demand, the case studies-based approach also needs to be carefully controlled for other associated changes and traffic reassignment. The challenges associated with the transferability of the values also persist.

Empirical Findings

The extent of latent demand and induced travel tends to vary considerably with the location and type of transport improvement and depends on the time horizon, unit of measurement, and method deployed. Some indicative values are presented later, though, due to the issues associated with transferability, the values should be considered with caution:

- The “Fundamental Law of Road Congestion” puts forward the proposition that the highway capacity expansion generates an exactly proportional increase in vehicle travel ($\gamma = 1$). Most of the empirical studies report this as the upper limit, particularly after controlling for the background growth and factoring in the reassignment.
- The short- and long-run elasticities tend to vary significantly. For instance, in the context of US highways, the short- and long-run elasticities are reported to be in the range 0.46–0.59 and 0.9, respectively (Fulton et al., 2000).
- Modeling studies reporting at the city region scale find smaller percentage changes in traffic flows due to induced traffic (0.7%–3.84%), compared with 5% on the main road link (Department of Transport, 2018).
- The highest increases in traffic volumes were found to occur for schemes where the baseline traffic volume was high or the route is close to capacity. Assuming that the percentage change in volume relative to the percentage change in capacity provides a rough elasticity estimate, the data from this study indicates elasticities of 0.5–1 for the congested and high volume routes but much smaller elasticities for other schemes (Davies, 2015; with data based on seven schemes in Australia).
- The relative contribution of absolute increase compared to increase due to changes in mode, destination, and route are typically difficult to identify. However, in the context of a major motorway improvement scheme (Manchester Motorway Box), a 15%–17% increase in car trips was observed of which the majority were attributed to changes in the destination (9%) and less to mode shift (4%). The demographic and employment effects were explicitly accounted for using detailed land-use data for the area and accounted for between 2.5% and 3% of the changes (Rohr et al., 2012).

^c Differentiating (1) $\gamma = \frac{\text{lane-mile}}{\text{VMT}} \cdot \frac{\partial \text{VMT}}{\partial (\text{lane-mile})}$.

^d Endogeneity refers to situations in which an explanatory variable is correlated with the error term.

Outstanding Issues

The importance of accounting for latent demand and induced travel is widely acknowledged. However, the challenge in incorporating it in transport cost–benefit analyses or other planning decisions lies in accurately predicting the amount of induced travel. Given the values being highly location dependent, it is not straightforward to simply “borrow” values from other similar studies. The lack of suitable data often makes it impossible to calculate it in a rigorous manner. Further, there are challenges associated with controlling for the background growth and the potential traffic reassignments. Moreover, induced travel very often occurs due to land use and/or economic growth stimulated by the transport project; this land-use change benefits and/or costs also need to be captured in the calculations which are a nontrivial task. The lack of evidence of latent demand and guidance for calculating it is more acute in case of public transport where it is often more challenging to measure the VMT equivalent.

Another well-acknowledged, yet under-researched area is the calculation of the cross-modal interactions that lead to induced demand. For example, a new high-speed rail connection between two cities may lead to increased economic activities between them leading to an increase in car trips. This also remains an area of ongoing development.

References

- Cervero, R., Hansen, M., 2002. Induced travel demand and induced road investment: a simultaneous equation analysis. *J. Transp. Econ. Policy* 36 (3), 469–490.
- Clifton, K.J., Moura, F., 2017. Conceptual framework for understanding latent demand: accounting for unrealized activities and travel. *Transp. Res. Rec.* 2668 (1), 78–83.
- Davies, W., 2015. What Effect do Queensland's Major Road Infrastructure Projects have on Traffic Volumes and Growth Rates? Australasian Transport Research Forum.
- Department of Transport, 2018. Latest evidence on induced travel demand: an evidence review, Report No. 70038415, Manchester.
- Downs, A., 1962. The law of peak-hour expressway congestion. *Traffic Q.* 16 (3), 393–409.
- Fulton, L.M., Noland, R.B., Meszler, D.J., Thomas, J.V., 2000. A statistical analysis of induced travel effects in the US mid-Atlantic region. *J. Transp.* 3, 1–14.
- Hymel, K., 2019. If you build it, they will drive: measuring induced demand for vehicle travel in urban areas. *Transp. Policy* 76, 57–66.
- Noland, R.B., 2001. Relationships between highway capacity and induced vehicle travel. *Transp. Res. Part A Policy Pract.* 35 (1), 47–72.
- Noland, R.B., Cowart, W.A., 2000. Analysis of metropolitan highway capacity and the growth in vehicle miles of travel. *Transportation* 27 (4), 363–390.
- Rohr, C., Daly, A., Fox, J., Patruni, B., van Vuren, T., Hyman, G., 2012. Manchester motorway box: post-survey research of induced traffic effects. *Plan. Rev.* 48 (3), 24–39.
- Wadud, Z., MacKenzie, D., Leiby, P., 2016. Help or hindrance? The travel, energy and carbon impacts of highly automated vehicles. *Transp. Res. Part A Policy Pract.* 86, 1–18.

Further Reading

- Cervero, R., 2002. Induced travel demand: research design, empirical evidence, and normative policies. *J. Plan. Lit.* 17 (1), 3–20.
- Downs, A., 2005. *Still Stuck in Traffic: Coping with Peak-Hour Traffic Congestion*. Brookings Institution Press, Washington, DC.
- Goodwin, P.B., 1996. Empirical evidence on induced traffic. *Transportation* 23 (1), 35–54.

ICT, Virtual and In-Person Activity Participation, and Travel Choice Analysis

Jacek Pawlak*, **Giovanni Circella^{†,*}**, ^{*}Urban Systems Lab & Centre for Transport Studies, Imperial College London, United Kingdom; [†]Institute of Transportation Studies, University of California, Davis, United States; ^{*}School of Civil and Environmental Engineering, Georgia Institute of Technology, United States

© 2021 Elsevier Ltd. All rights reserved.

Information and Communication Technology	452
Related Concepts in the Analysis of Activity Decisions and Travel Choices	453
Tele-, Online, Virtual, Digital, and e-activities	453
Smart and Mobile Capabilities	454
Digital Divide and Digital Skills	455
Intelligent Transport Systems (ITS)	455
Big Data	455
Frameworks for Analyzing ICT Interactions With Activity Decisions and Travel Choices	456
Outlook for the Future	457
References	457

Information and Communication Technology

Information and communication technology (ICT)—also used in the plural form, that is, information and communication technologies, or ICTs—has a central role in a growing number of contexts related to activity decisions and travel choices. The concept has entered the broader scientific discourse in the 1980s reflecting the convergence and ever-tighter coupling between computing (information processing) and telecommunications that started in the late 1970s and early 1980s. The subsequent decades saw the mass adoption of personal computing- and telecommunications-capable devices (hardware), with a big expansion in the access to the internet and, more recently, the adoption of mobile devices, together with all their capabilities and services (software). All of those have placed ICT among the key factors shaping societies, hence also warranting the analysis of their important role in the contexts of activity decisions and travel choices.

Despite its popularity, the term ICT has lacked a single, authoritative definition. For instance, the Dictionary of Media and Communication ([Chandler and Munday, 2012](#), p. 120) defines the meaning of ICT in a rather broad sense:

“An umbrella term for all of the various media employed in communicating information: for example, in an educational context ICT may include computers, the internet, television broadcasts, and even printed or handwritten notes.”

While flexible, such a general interpretation is very broad and not very conducive to the analysis of the role of ICT in relation to activity decisions and travel choices. The International Telecommunication Union (ITU), a specialized agency of the United Nations responsible for issues that concern ICT, groups the indicators describing the state of ICT into four main technological families ([ITU, 2011](#)):

- fixed-telephone networks;
- mobile-cellular networks;
- the internet;
- broadcasting.

Within each group, the indicators are defined so that they are measurable (quantifiable) and thus allow the comparison of the state of ICT sectors in various regions and countries. The ITU approach is strongly hardware- and communications-centric, though. This may be limiting when it comes to acknowledging changes occurring in the “softer” layer of ICT, that is, capabilities, services and applications. From that standpoint, the most appealing definition has been offered by the [OECD \(2004, p. 3\)](#), which has also been employed by the [ITU \(2011, p. 133\)](#), to describe ICT goods and services as those:

“ . . . intended to fulfil the function of information processing and communication by electronic means, including transmission and display, or use electronic processing to detect, measure and/or record physical phenomena, or to control a physical process.”

Such definition is focused on the functionalities and capabilities offered by ICT while explicitly acknowledging the role of electronic means, thus excluding various non-electronic (“pen and napkin”) information processing and communication possibilities. The definition is also particularly attractive in its emphasis on the role of ICT as a means of interacting with the reality, thus enabling new forms of activity participation, themselves of crucial importance in the context of activity decisions and travel choices.

In that sense, ICT solutions more concretely include:

- fixed and mobile (cellular) telephones, including smartphones;
- computers, including desktop, portable (laptops, tablet), high-performance computers, and servers of various kinds;
- any other stationary or portable devices capable of information processing or communication, including portable digital assistants (PDA), TV and radio sets, music players, game consoles, wearable technologies, home appliances;
- the internet, including technologies enabling its operation, that is, data routing, and accessing, for example, WiFi, fixed broadband, mobile broadband (including 3G, 4G, and 5G standards);
- other networks, for example, local area networks (LAN), intranet, adhoc networks;
- various sensing technologies with information processing and communication capabilities, including closed-circuit television, satellite-based sensing, automatic traffic counters;
- software utilized by the aforementioned hardware, both based directly on the local devices and on remote servers (the cloud), to enable the services, applications, and communication through information processing and transmission.

The list above is not exhaustive, but acts as a convenient “entry point” to understanding what ICT means in the context of activity decisions and travel choice analysis. Any specific analysis of ICT requires establishing a working definition tailored to the context, including consideration of the scope of technologies interpreted as ICT, the level of detail in their specification, time horizon, or the range of applications and capabilities considered.

In the past three decades, the transportation research community has shown a strong interest in understanding the impacts of ICT adoption on activity participation and travel choices. This interest has been motivated by early hopes of mitigating travel demand and reducing congestion through substitution with an ICT-based means of activity participation. Despite the ICT-travel substitution having been often below expectations, the topic regularly regains prominence in the transportation planning research agenda during periods of large-scale and prolonged disruptions to transportation systems. Such disruptions can be man-made, for example, large-scale events in densely populated areas, for example, London 2012 Olympics, or nature-made, such as air travel disruptions following the 2010 Eyjafjallajökull volcano in Iceland. However, not until the COVID-19 pandemic in the first half of 2020 has the potential of ICT been shown vividly at a scale. With lockdown measures that limited physical travel within and between countries, large portions of populations were forced to operate from their homes. Where possible, individuals started to rely on ICT tools for work activities and social interactions, causing the adoption rate and frequency of many ICT-enabled activities to soar. It remains to be seen to what extent these unprecedented circumstances might translate into long-lasting changes in the relationships among ICT use, activity participation, and travel choices (and eventually requiring revisiting the existing evidence and knowledge).

Related Concepts in the Analysis of Activity Decisions and Travel Choices

A number of concepts in the field of the analysis of activity decisions and travel choices are often seen alongside ICT in a variety of modeling, policy, and planning contexts, including:

- tele-, online, digital, and e-activities;
- smart and mobile capabilities;
- digital divide and digital skills;
- intelligent transport systems (ITS);
- big data.

Unfortunately, a detailed understanding of their respective interactions with ICT is often hampered by the lack of clarity on the terms’ meaning and often adhoc and lax interchange of the terms in use. To facilitate clarity, below we provide working definitions to distinguish the concepts and more directly relate them to the context of ICT, activity decisions, and travel choices.

Tele-, Online, Virtual, Digital, and e-activities

In the broadest sense, *teleactivities* (or tele-activities) refer to activities conducted remotely (from the Greek word *tēle* meaning “far off”). The concept has remained in use since at least 1970s, when the phrases of *teleworking* and *telecommuting* were coined and popularized in the wake of the energy crisis of 1970s (Bailey and Kurland, 2002; Mokhtarian, 2009). A comprehensive review from a decade ago (Andreev et al., 2010) highlighted the existence of (already then) substantial research concerning teleactivities and their interactions with travel behavior, including teleworking and telecommuting, teleconferencing, teleshopping, teleservices, teleleisure.

Notably, teleactivities are focused on participation in an activity that is spatially separated from the location in which the activity would traditionally take place. This original meaning implicitly embodies ICT-facilitated participation. The concept does not specify, however, either the need for existence of communication between the locations. For instance, word processing or writing a piece of a computer program can be standalone teleactivities. If/when such communication occurs via a modern telecommunication means, typically the internet, the activity becomes an *online activity*.

On the other hand, participation in a *virtual activity* invokes the notion of virtual reality. Steuer (1992, p.6) offers a convenient conceptual link with teleactivities through the notion of *telepresence* defined as “the experience of presence in an environment by

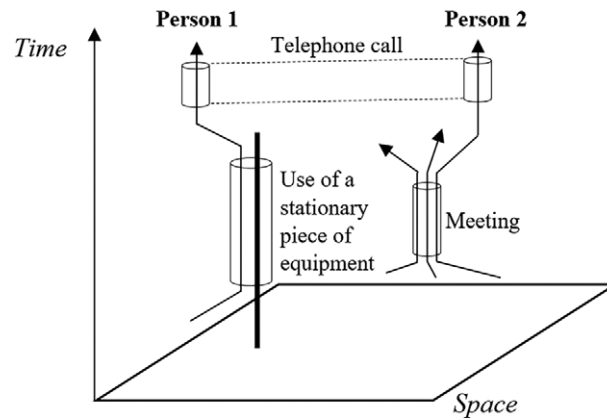


Figure 1 Hågerstrand's representation of a telephone call in the context of activity decisions and spatiotemporal constraints. *Source:* Created by the authors, based on Figure 2 in Hågerstrand (1970)

Table 1 ICT-related activity concepts

Activity concept	Focus implied by the definition	The relationship to ICT
Teleactivity	Spatial separation between the activity participant and location of the activity in its traditional version.	ICT as an enabler of the remote activity participation.
Online activity	Use of a telecommunication medium.	ICT as modern telecommunication technologies, enabling exchange of information for activity participation.
Virtual activity	Relying on the ability to create a sensory-rich experience of presence in an environment (virtual reality).	ICT as a means of creating sensory rich experience, including communication.
Digital activity	Relying on technologies for digital processing and communication of the information.	ICT as technologies using digital processing of information to enable activity participation.
E-activity	Use of electronic technologies.	ICT as technologies that are electronic and enable activity participation.

Source: Created by the authors

means of a communication medium." Following from there, a virtual activity can be defined as one conducted in virtual reality, that is, "a real or simulated environment in which a perceiver experiences telepresence" (Steuer, 1992, p.9).

The reference to the use of a particular technology, hardware or software, that processes information digitally (i.e., as binary digits) is a defining feature of a *digital* activity. Similarly, the use of *electronic* technology, whether for computing or communications, to conduct an activity defines the term *e-activity*. As nowadays ICT are almost exclusively digital and electronic, this helps explaining an almost synonymous use of the terms in modern contexts. A summary of the above concepts is presented in Table 1.

In the analysis of the activity decisions and travel choices, early research hypotheses (and hopes from policy makers) assumed the emergence of teleactivities to be a panacea for transportation (road, in particular) congestion, enabling substantial reduction in the demand for travel. This reduction was supposed to be fueled by the expanding range and sophistication of the activities facilitated by ICT, including those enabling online and virtual reality experience. Nowadays, this interpretation is known to be incomplete, with interactions being much more elaborate. Section 3 of this chapter discusses the conceptual and modeling frameworks and taxonomies that have been proposed to analyze such interactions.

Smart and Mobile Capabilities

The notion of *smartness* is increasingly prevalent in the mobility discourse, that is, concerning activity decisions and travel choices. Smart ICT devices are those that are capable of processing information and of communicating with other devices with a substantial degree of autonomy, while also providing a considerable means of interaction with the user. In the mobility discourse, the notion of a smart ICT device is nowadays almost always accompanied by the notion of *mobile devices*. A mobile device is designed to allow the realization of its information processing and communication capabilities without the continuous reliance on fixed infrastructure, for example wired connection or power supply, and with the device dimensions facilitating portability.

The mass adoption of smart mobile ICT devices such as smartphones, portable computers, and wearable technologies, has been a major driver behind the emergence of new business models and technological solutions in mobility. Those range from advanced and real-time trip planning and navigation, to service offering comparison, frictionless payment and ticketing, shared mobility, and/or on-demand mobility services.

Digital Divide and Digital Skills

The adoption and use of ICT require access to resources (devices, infrastructure) as well as the acquisition of the appropriate level of skills. The uneven access to or ability to use ICT among groups defined on social or geographical grounds is referred to as *digital divide* (OECD, 2006). The term can be observed in absolute terms, for example, access to ICT or lack of it, as well as in relative terms, for example, varying quality of connectivity. The term in the mobility context has particular relevance as it highlights variations in people's ability to benefit from advances in ICT in the context of mobility.

The accompanying concept is that of *digital skills* which refers to individual capabilities and knowledge concerning the use of ICT. ITU (2018) defines a number of such skills along which they measure the ability to use ICT:

- copying or moving a file or folder
- sending e-mails with attached files
- using copy and paste tools
- transferring files between a computer and other devices
- finding, downloading, installing, and configuring software
- connecting and installing new devices
- using basic arithmetic formulas in a spreadsheet
- creating electronic presentations
- writing a computer program

While regrettably often overlooked in travel-behavior analysis, the concept is key to comprehensively understand any postulated impacts from ICT on activity decisions and travel choices. Using a parallel to car access and driving license possession, the potential neglect of the digital divide and digital skills in the analysis of ICT in the context of activity decisions and travel choices would be the counterpart of assuming a universal access to a car and driving license possession. While the assumption holds in some contexts, in others, such as for certain segments of the populations (e.g., elderly, rural dwellers, etc.), it clearly does not, possibly resulting in ineffective or unequitable policies.

Intelligent Transport Systems (ITS)

Even if they are not the main object of discussion of this chapter (but are discussed in other chapters of this book), intelligent transport systems (ITS) warrant a brief mention in the present context due to their close relation to the concept of ICT, and hence a potential conceptual confusion. In probably the simplest terms, ITS applications make use of ICT to deliver advanced transportation systems operation and management capabilities. Thus, ITS are focused primarily on providing efficient *transportation supply*, toward which end they make use of ICT to process the relevant data and allow communication and coordination between transportation systems' components (hence ITS are increasingly referred to as Cooperative ITS, or C-ITS).

The use of ICT for activity decisions and travel choices does not need to involve ITS. An example is a decision to make use of ICT to enable an online activity participation and not travel, which affects *transportation demand*. Still, the deployment of various types of ICT-based ITS solutions have important implications also in terms of their effects on activity participation and travel choices.

Big Data

Big data conventionally refers to the notion of analysis of very large datasets using modern, large-scale computation capabilities for the purposes of obtaining insights that could not have been obtained using a more traditional analytical means. Big data are typically characterized by so-called 3Vs: large *volume*, substantial *variety*, and high *velocity*. The role of ICT in the space of big data is two-fold: it enables big data analysis itself as well as the creation of big data in the forms of *digital footprints* and *data fumes*. The former function of ICT has a very direct role in the field of activity decisions and travel choices, enabling new methods of inquiry.

The role of ICT as enabler of digital footprints and data fumes requires further elaboration. *Digital footprints* are traceable data points created by individuals as the result of their participation in digital activities. Such footprints can take the form of content generated by a person online, including textual or visual, or their choices, including website selection or goods and services purchase. Datasets generated in this manner can be collected and analyzed to better understand people's activity decisions and travel choices, in light of the growing role of digital activities in people's lifestyles.

A related concept is that of *data fumes* which also refers to user-generated data, but more often passively collected and often without user awareness. The records of online activities as well as those generated by the devices on the network, such as traces, transactions or incidents, can also be harvested for the purposes of mobility analysis.

Following experimental approaches in the 2000s, big data applications have now made it to the mainstream analysis of activity decisions and travel choices. The broad opportunities allowed by the analysis of big data in transportation planning are not the main focus of this chapter, but are discussed in details in other chapters.

Frameworks for Analyzing ICT Interactions With Activity Decisions and Travel Choices

Arguably what can be seen as the earliest reference to how ICT may be interacting with activity decisions and travel choices was made in Torsten Hägerstrand's presidential address (1970) in which he stated:

"Telecommunication allows people to form bundles [joining other individuals in a common activity] without (or nearly without) loss of time in transportation." (Hägerstrand, 1970, p.15)

More broadly, Hägerstrand's theory developed the idea of *space time paths* along which individuals move (see Figure 1). By embodying the idea of a telephone, which escapes what are otherwise the spatiotemporal constraints on individuals having to travel to conduct an activity, he proposed what was effectively the first link among the predecessor of modern ICT, activity decisions and travel choices. Remarkably, the conceptualization is still largely applicable to modern teleactivities.

In the 50 years since then, many research studies have discussed how ICT can affect many aspects of human life, ranging from long-term choices, for example, relocation and migrations, to medium-term choices, for example, vehicle ownership, to short term-choices, such as the organization of activities, including those performed for work/business and/or leisure/social purposes, and trip-related choices, including destination, travel mode, departure time, and route choices. All these impacts of ICT can have plenty of direct or indirect effects, sometimes in conflicting directions, on various components of travel demand (Circella and Mokhtarian, 2017).

According to the taxonomy introduced by Salomon and Mokhtarian in the late 1980s and 1990s (Mokhtarian, 1990; Salomon, 1986), one which has remained remarkably stable over the years, interactions between ICT and travel behavior can occur at different levels. *Substitution* takes place when an individual chooses to conduct activity by means of ICT, that is, as a teleactivity, rather than in physical reality, thus reducing or even eliminating the need to travel. In its pure conceptual form, substitution should result from an individual still undertaking an activity, but in its ICT-based form instead of traveling to a location where the physical activity would have been undertaken. In practice, however, research studies frequently identify net substitution as measured by a reduction in the amount of travel.

An opposite effect to that of substitution is *complementarity* (sometimes referred to as *generation*), in which case the use of ICT leads to an increased amount of travel. Such situation is supposed to result from ICT-induced easier access to information and cheaper communication leading to increased awareness of various opportunities, also at more distant locations. Moreover, cheaper communication could provide a means for maintaining larger social networks, potentially translating into additional demand for physical meetings, and thus travel. Such effects could emerge at different scales, ranging from local leisure-related meetings to global professional networks and potential expansion into new markets, with the consequent need for international business travel.

The substitution and complementarity constitute extreme cases of interactions between ICT and travel behavior, focusing on the sheer amount of travel. In situations when travel still takes place, at least partly, ICT can lead to *modifying* activity-travel patterns in terms of, for instance:

- trip purpose(s) and/or destination(s);
- trip timing;
- route;
- mode(s) of travel;
- use of travel time (travel-based multitasking).

The final effect of *neutrality* is a limiting case of the other three types either not existing or canceling each other, and thus resulting in *net* neutrality.

In addition to the earlier, there are also higher order interactions that involve ICT influencing other aspects of individual lifestyles and long-term choices (residential choice, mobility and time use preferences, and vehicle ownership) or even society and economy (social norms, urban form, smart infrastructure provision, and market organization). Such impacts can subsequently propagate to ultimately affect activity-travel patterns. At the same time, such higher order interactions take place over longer time due to various rigidities and inertia in socioeconomic systems, including durations of rental contracts, investment time spans, regulations' design, introduction, and implementation, as well as habits, customs, and lack of trust in novel technologies among others. It should be noted, however, that the two different levels of interaction are artificial constructs to systematize understanding. In reality, interactions at different levels would be closely interconnected, for example, long-term substitution of activity could only occur if its tele-counterpart is trusted.

It should be also noted that ICTs as a travel replacement strategy have been seen for many years as a solution to many societal problems, including urban congestion, dependence on non-renewable energy sources, air pollution, and greenhouse gas emissions. The role of ICT in overcoming physical remoteness has been suggested as a way to overcome physical boundaries and limitations such as rural underdevelopment, reduced economic opportunity for the mobility-limited, and the struggle to balance job and family responsibilities of many segments of society (Mokhtarian et al., 2005; Salomon and Mokhtarian, 1998). The evidence in the literature has shown that use of ICT certainly can replace "a lot" of travel, but at the same time can generate additional and modify other travel. It contributes to relaxing some physical boundaries and constraints, while also imposing new ones, such as the dependence on the availability of appropriate hardware and software (and charging facilities and batteries), and access to reliable communication networks.

In addition to the above classical view on of ICT in the context of activity decisions and travel choices, ICT has also been discussed as offering increased opportunities for the adoption of *multitasking* behaviors, increasing the ability to overlay multiple activities performed “at the same time” but also causing *activity fragmentation*. These time allocation phenomena have been hypothesized as being enhanced by the proliferation of ICT. In the case of multitasking, Kenyon and Lyons provided a concise definition of the concept (Kenyon and Lyons, 2007, p. 162):

“[multitasking is] simultaneous conduct of two or more activities during a given time period.”

The importance of the multitasking phenomenon lies in its potential of relaxing the conventional microeconomic time allocation theory assumption that the total duration of activities an individual can conduct during a specified period is fixed. This assumption underlies most of the modern transport investment appraisal methodologies and its relaxation may have implications for methods for travel time valuation. The incorporation of multitasking in the analysis of activity choices and travel decisions, and particularly in the context of so-called travel-based multitasking, has become of particular importance recently, fueled by the proliferation of mobile ICTs as well as by the expectations regarding new transportation technologies, and in particular connected and automated vehicles. Circella et al. (2012) provides a comprehensive theoretical discussion on multitasking and related concepts, while the framework of Pawlak et al. (2015) demonstrates how ICT, teleactivities, and multitasking are interrelated in the context of activity decisions and travel choices. Lastly, Keseru and Macharis (2018) and Pawlak (2020) have provided systematic reviews of empirical studies, also in reference to ICT use in the context of travel-based multitasking.

The phenomenon of *fragmentation* describes the relaxation of the traditional boundaries defining spaces and times characteristic to certain activities (Couclelis, 2009, p. 1560):

“Instead of occupying compact chunks inside the daily prism, ICT-aided activities tend to disintegrate into sets of discrete tasks which get spread out across places and over time.”

In other words, fragmentation refers to conditions in which conventional hours and locations typical for activities, for example, work between 9 am and 6 pm in an office, become increasingly invalid as a result of ICT-enabled flexibility in the form of teleactivities.

Outlook for the Future

During recent decades, the society has experienced a sharp growth in the adoption of ICT. Initially constrained to only a handful of industrialized countries, ICT-based solutions have become widespread also in developing countries and emerging economies, thanks to the adoption of mobile communication and smartphone technologies that helped overcome the eventual lack of physical infrastructure. ICT has generated a variety of new products and services—including new shared mobility and on-demand mobility services - and novel solutions for business-to-business and business-to-consumer transactions, the dematerialization of the goods, fast communication and social connections at distance, with complex implications on the resulting activity patterns and travel demand, as discussed by Pawlak et al. (2020). Moving forward, transportation researchers question to what extent past trends and established knowledge based on the former will hold for the future.

Certain ICT applications, such as e-commerce, have already altered the modern life as well as the physical shape of cities and regions significantly. For example, the growing internet penetration and the consequent stimulation of e-commerce altered individuals' shopping behaviors and spending patterns. It has also significantly grown into the retail sector and generated structural changes to supply chains and the organization of retail space itself, through the disappearance of many stores themselves, and the emergence of combinations of “bricks-and-clicks” forms of retail that include both physical presence with traditional stores complemented by online e-shopping.

In the year of publication of this chapter, the COVID-19 pandemic has caused large portions of the world population to abruptly reduce their amount of travel and their participation in in-person activities. During the pandemic, ICTs have been used as substitutes for travel for many activities at much higher rates, and across larger segments of the population, than in the past, causing significant reductions in vehicle miles traveled (VMT) for all trip purposes but also a large drop in the use of public transportation or shared mobility, i.e. the travel modes that might be more conducive to the transmission of the disease. This could create a substantial challenge for transportation demand management, making private car a particularly attractive alternative, as activities gradually reopen but travelers remain hesitant about using “shared” modes of travel. It remains to be seen, though, whether the extended time over which individuals have been forced to modify their habits and embrace ICT-based alternatives might have caused changes in the way that ICT interact with activity decisions and travel choices, with consequently long-lasting impacts on individuals' behaviors and lifestyles.

References

- Andreev, P., Salomon, I., Pliskin, N., 2010. State of teleactivities. *Transp. Res. Part C Emerg. Technol.* 18 (1), 3–20.
 Bailey, D.E., Kurland, N.B., 2002. A review of telework research: Findings, new directions, and lessons for the study of modern work. *J. Organ. Behav.* 23 (4), 383–400.
 Chandler, D., Munday, R., 2012. *Dictionary of Media and Communication*. Oxford University Press, Oxford.

- Circella, G., Mokhtarian, P.L., 2017. Impact of information communication technology on transportation. In: Susan, Hanson, Genevieve Giuliano, C. (Eds.), 4th Ed. *The Geography of Urban Transportation*, The Guilford Press, New York, pp. 86–109.
- Circella, G., Mokhtarian, P.L., Poff, L.K., 2012. A conceptual typology of multitasking behavior and polychronicity preferences. *Elec. Int. J. Time Use Res.* 9 (1), 59–107.
- Couclelis, H., 2009. Rethinking time geography in the information age. *Environ. Plan. A* 41 (7), 1556–1575.
- Hägerstrand, T., 1970. What about people in regional science? *Papers Region. Sci. Asso* 24, 7–21.
- ITU, 2011. Handbook for the collection of administrative data on Telecommunications/ICT. International Telecommunication Union (ITU), Geneva. Available from: https://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-ITC_IND_HBK-2011-PDF-E.pdf. (Visited 11 Nov 2019).
- ITU, 2018. Measuring the Information Society Report 2018, International Telecommunication Union (ITU), Geneva. Available from: <https://www.itu.int/en/ITU-D/Statistics/Pages/publications/misr2018.aspx>. (Visited 11 Nov 2019).
- Kenyon, S., Lyons, G., 2007. Introducing multitasking to the study of travel and ICT: Examining its extent and assessing its potential importance. *Transp. Res. Part A: Policy and Practice* 41 (2), 161–175.
- Keseru, I., Macharis, C., 2018. Travel-based multitasking: Review of the empirical evidence. *Transp. Rev.* 38 (2), 162–183.
- Mokhtarian, P.L., 1990. A typology of the relationships between telecommunications and transportation. *Transp. Res. Part A* 24 (3), 231–242.
- Mokhtarian, P.L., 2009. If telecommunication is such a good substitute for travel why does congestion continue to get worse? *Transp. Lett.* 1 (1), 1–17.
- Mokhtarian, P.L., Salomon, I., Choo, S., 2010. Measuring the measurable: Why can't we agree on the number of telecommuters in the US? *Qual. Quant.* 39 (4), 423–452.
- OECD, 2004. A Proposed Classification of ICT Goods, OECD Working Party on Indicators for the Information Society. The Organisation for Economic Co-operation and Development, Paris. Available from: <https://stats.oecd.org/glossary/detail.asp?ID=6274>. (Visited 11 Nov 2019).
- OECD, 2006. Glossary of Statistical Terms: Digital Divide. The Organisation for Economic Co-operation and Development, Paris. Available from: <https://stats.oecd.org/glossary/detail.asp?ID=4719>. (Visited 11 Nov 2019).
- Pawlak, J., 2020. Travel-based multitasking: review of the role of digital activities and connectivity. *Transp. Rev.* 40 (4), 429–456.
- Pawlak, J., Circella, G., Mahmassani, H.S., and Mokhtarian, P.L., 2020. Information and Communication Technologies (ICT), Activity Decisions, and Travel Choices: 20 Years into the Second Millennium and Where Do We Go Next? Centennial Paper, Transportation Research Board, Washington DC.
- Pawlak, J., Polak, J.W., Sivakumar, A., 2015. Towards a microeconomic framework for modelling the joint choice of activity–travel behaviour and ICT use. *Transp. Res. Part A: Policy and Practice* 76, 92–112.
- Salomon, I., 1986. Telecommunications and travel relationships: A review. *Transp. Res. Part A: General* 20 (3), 223–238.
- Salomon, I., Mokhtarian, P.L., 1998. What happens when mobility-inclined market segments face accessibility-enhancing policies? *Transport. Res. D* 3 (3), 129–140.
- Steuer, J., 1992. Defining virtual reality: Dimensions determining telepresence. *J. Comm.* 42 (4), 73–93.

Public Transit Ridership Forecasting Models

Ipsita Banerjee*, **Deepa L[†]**, **Abdul Rawoof Pinjari[‡]**, *Department of Civil and Environmental Engineering, University of California, Berkeley, California, United States; [†]Department of Civil Engineering, Indian Institute of Science, CV Raman Road, Bengaluru, Karnataka, India; [‡]Department of Civil Engineering, Centre for infrastructure, Sustainable Transportation and Urban Planning (CiSTUP), Indian Institute of Science, CV Raman Road, Bengaluru, Karnataka, India

© 2021 Elsevier Ltd. All rights reserved.

Why Model Transit Ridership?	459
Approaches to Modeling Transit Ridership	460
Direct Demand Models for Forecasting Transit Ridership	460
Design Considerations for Transit Ridership Forecasting Models	461
Temporal Resolution	461
Spatial Resolution	461
Complementary and Competing Routes	462
Transfer Boardings and Alightings	462
Interdependency Between Demand and Supply	462
Factors Influencing Transit Ridership	463
External or Exogenous Factors Influencing Transit Ridership	463
Internal or Endogenous Factors Influencing Transit Ridership	463
Accessibility of the Transit System	464
Accessibility Provided by the Transit System	464
Factors Not Considered in Direct Demand Models	464
Model Structure	464
Emerging Trends and Future Directions	466
Acknowledgments	466
References	466
Further Reading	467

Why Model Transit Ridership?

Ridership is a key measure of a transit system's performance. It is a direct indicator of the extent of use of a transit system and an important determinant of the revenue generated. Understanding and modeling transit ridership is useful primarily for two purposes:

1. To forecast transit ridership for evaluating long-term urban planning strategies, such as transportation project investments (e.g., introduction of a new transit service for the city) and land-use strategies (e.g., transit-oriented developments). Such long-term forecasts help in assessing project feasibility, evaluating alternative plans and scenarios, and estimating environmental and social impacts of transportation projects.
2. To provide objective means of assessing ridership and revenue changes following improvements by the transit agencies. Public transit agencies continuously endeavor to improve their services through a variety of strategic and operational improvements. Such improvements include addition of new routes, increases in service frequency, changes in schedules and service network, and pricing. These also include connectivity to and from other modes of travel, such as feeder services to metro stations, improvements in amenities at stations and stops, designation of transit priority corridors, and provision of information to travelers. These improvements are aimed at increasing service efficiency, ridership, and revenue, and enhancing the service coverage. Ridership forecasting models can be used by transit agencies to assess the possible impacts of the improvements on ridership and revenue.

This chapter begins with an overview of the different approaches to modeling transit ridership. It then delves into direct demand models that are used to model aggregate ridership as a function of the factors influencing it. The challenges in developing and applying transit ridership forecasting models are discussed in the section on model design considerations. Following that is a discussion on the model structures and explanatory variables used in ridership prediction models. The chapter concludes with a discussion of emerging trends that impact transit ridership analysis and future directions.

The focus of this chapter is on fixed-route, fixed schedule bus transit systems. However, many of the considerations also apply to rail transit and to bus transit systems with flexible routes and schedules.

Approaches to Modeling Transit Ridership

Modeling transit ridership for a study region involves estimating the number and demographic distribution of riders who use different transit services. It also involves estimating the spatial, temporal, and route distribution of ridership as a function of the socio-demographic, land-use, and transportation service characteristics of the study region, including the service characteristics of the transit system. Most approaches to modeling transit demand may be classified into two broad categories: (1) City-scale or regional travel demand model systems, and (2) Direct demand models. Each of these is briefly discussed further, with direct demand models being the focus of the rest of the chapter.

1. City-scale or regional travel demand model systems are aimed at predicting a region's aggregate travel patterns as a result of the trip makers' choices of whether to travel and how much, as well as why, when, where, and how to travel. These models can be structured in two broad forms:
 - a. aggregate, trip-based travel demand models
 - b. disaggregate activity-based travel microsimulation models

The aggregate, trip-based travel demand models often involve the four-step approach used to estimate (i) trip generation or total trips generated for each spatial zone in the study area; (ii) spatial distribution of the trips between each pair of zones; (iii) modal split of trips, including transit ridership between each pair of zones; and (iv) assignment of traffic to highway and transit networks. The disaggregate, activity-based models focus on microsimulating the travel choices of each individual in the study area toward estimating aggregate travel patterns. Using such disaggregate models, one can estimate the aggregate spatial, temporal, route, and demographic distribution of transit ridership (Pinjari and Bhat, 2011). Mode choice models form the basis for estimating transit ridership in these models.

2. Direct demand models, as the name suggests, are aimed at directly modeling aggregate transit ridership, typically in the form of statistical regression models that relate ridership in an area over a specific time period to the various factors influencing it. It is worth noting here that, even if there were access to a regional travel demand model system, transit agencies that operate large transit systems would benefit from developing their own ridership forecasting models in the form of direct demand models. This is because direct demand models can be used to answer questions of interest to transit agencies that many regional travel demand models in use might not be helpful for. For example, regional travel demand models might not include enough detail to forecast ridership impacts of changes in transit stop-level amenities and accessibility. Even if they do, it might be easier to run a simpler direct demand model than to run a large-scale regional travel demand model. Further, direct demand models can be structured to utilize information from regional travel demand models. Once developed, these models are quick response in nature and less expensive in terms of time, cost, and data requirements when compared to large-scale travel demand models. A large body of literature exists in this area (Peng et al., 1997; Chu, 2004; Chu, 2007; Polzin et al., 2011; Gutiérrez et al., 2011; and Chakour and Eluru, 2016). Most of the literature, however, is based on developed economies and not much exists in the context of developing countries.

Direct Demand Models for Forecasting Transit Ridership

Ortuzar and Willumsen (2011) describe a direct demand model as one that subsumes trip generation, distribution, and mode choice. They cite the SARC (Kraft 1968) model in which demand between a pair of zones (i and j) is estimated as a multiplicative function of the activity and socioeconomic variables at each of those zones, and the level-of-service attributes of the modes serving them, as shown in the following equation:

$$T_{ijk} = \phi_k (P_i P_j)^{\theta_{kp}} (I_i I_j)^{\theta_{kl}} \prod_m \left[\left(t_{ij}^m \right)^{\alpha_{km}} \left(C_{ij}^m \right)^{\beta_{km}} \right]$$

In the above equation, P_i and P_j are population or other surrogates of trip generative characteristics in zones i and j , respectively; I_i and I_j are employment or other trip attractive characteristics; t_{ij} and C_{ij} are travel time and travel cost between i and j by mode k ; $k \in [1, m]$, respectively; and ϕ_k , θ_{kp} , θ_{kl} , α_{km} , β_{km} are parameters of the model. Specifically, θ_{kp} and θ_{kl} are demand elasticities with respect to population and income, respectively; and α_{km} and β_{km} are demand elasticities with respect to time and cost of travel, respectively (they are direct elasticities when $k = m$ and cross-elasticities otherwise).

For transit ridership prediction, the above framework may be used to estimate stop-to-stop (origin stop to destination stop) ridership flows as a function of the catchment area characteristics of the origin and destination stops and the transit service characteristics between the stops. Such a framework helps not only in estimating the stop-to-stop ridership but also in estimating the revenues earned from ridership. However, many direct demand models of transit ridership focus only on trip generation for transit without considering the spatial distribution of those trips. These are typically in the form of route-level or stop-level ridership models. Route-level models relate ridership to service characteristics at the route-level. Some of the route-level models are designed to relate ridership and service characteristics for specific route segments defined by fare-levels (Peng et al., 1997).

Although service changes typically occur at the route level, ridership on a route varies across the different city zones that the route covers. To consider variation in ridership along a route, stop-level models may be considered. These models relate boardings and

(or) alightings at each transit stop to the catchment area characteristics of that stop (Chu, 2004; 2007; Polzin et al., 2011; Cervero et al., 2010; Gutiérrez et al., 2011; Chakour and Eluru, 2016). Such models make it possible to consider the influence of stop-level infrastructure, amenities, and the population and land-use around the transit stops. Separate models may be developed for boardings and alightings as the influence of service and catchment area characteristics can be different for boardings and alightings. Stop-level models may also be further disaggregated by the routes serving the stops, as route-stop models (Mucci and Erhardt, 2018) in which both route-level service and stop-level catchment area characteristics can be used to explain the stop-level boardings or alightings. Note that stop-level models also allow one to quantify and analyze transfer trips (if such data is available) separately from direct trips.

Transit services in many cities are differentiated into general, subsidized, or luxury categories, with each of these services differing in convenience and comfort, frequency, fare, and in the subset of stops served along a route. In such cases, modeling by service types might be a consideration. Of course, modeling disaggregate, stop-route level ridership separately for different types of services warrants the need to consider competition and complementarity between different services as well as competition or complementarity relationships among different routes in the network.

It is worth noting here that stop-level transit trip generation models are more commonly used than either route-level trip generation or stop-to-stop ridership flow models. Since revenue forecasting is an important reason for transit ridership forecasting, the stop-level ridership forecasts may be used for subsequent estimation of stop-to-stop ridership flows by route and service type or fare class.

Design Considerations for Transit Ridership Forecasting Models

Many considerations go into designing and building an effective transit ridership forecasting model. These include:

- determining the temporal and spatial resolution at which ridership is modeled,
- addressing the complementary and competitive nature of service along different routes,
- treating transfers separately from direct boardings or alightings,
- addressing the interdependency between transit ridership (demand) and service (supply),
- incorporating the different factors that influence ridership, and
- determining the type of mathematical/statistical models to be used.

Each of these model design considerations should be guided by a combination of theoretical as well as practical factors, such as data availability and simplicity for easy usage and explainability. Some of these design considerations are discussed in the remaining section.

Temporal Resolution

The purpose of prediction and data availability guides the temporal resolution of the model. For real-time operational planning purposes, such as provision of on-demand last-mile services, temporal intervals of 15 minutes to 1 hour are useful. While it is desirable to have fine temporal resolution, such as minutes, for purposes such as crowding estimation for real-time operational planning, it is important to have access to real-time sensor data on crowding for such purposes. Models for different time-of-day periods, for example AM peak, AM off-peak, mid-day, PM peak, evening, and night, may be used for near-term planning purposes such as schedule and route structure changes, and fleet maintenance scheduling. Predictions by days of the week, or broadly by weekdays, weekends, and public holidays, are useful for updating weekly schedules and service plans as well as for all of the above-mentioned purposes. For medium to long-term applications, such as resource allocation or substantial increases in bus fleet size, monthly or annual demand may be predicted.

Spatial Resolution

Predicting transit ridership at the typical traffic analysis zone (TAZ) resolution does not provide helpful information to transit agencies for service planning and operational purposes. As discussed earlier, modeling ridership at the route-level, route segment-level, stop-level, or between origin and destination stops is more useful. However, modeling ridership at finer, stop-level spatial resolution is associated with its own challenges and uncertainties. First, closely spaced stops sometimes have overlapping catchment areas. Besides, ridership data collection in many transit systems happens at a fare-stage or fare-segment level, where the number of boardings (or alightings) at a cluster of consecutive stops on a route are aggregated into those of a designated stage-stop. Since fare setting and issue of tickets to riders typically happens at a stage-to-stage aggregation, ridership data is also collected at a stage-level. Therefore, it might be pragmatic to model stage-level boardings or alightings, or stage-to-stage ridership, than to do so at a disaggregate stop-level. However, the analyst will have to disaggregate stage-level catchment area characteristics to stop-level for stop-level ridership prediction, as stage-level ridership is too aggregate to include meaningful stop-level characteristic variables. Additionally, ridership at spatially proximate stages or stops and on a same route might be correlated due to the influence of unobserved factors that influence ridership at these stages (or stops). Considering these issues can potentially improve model

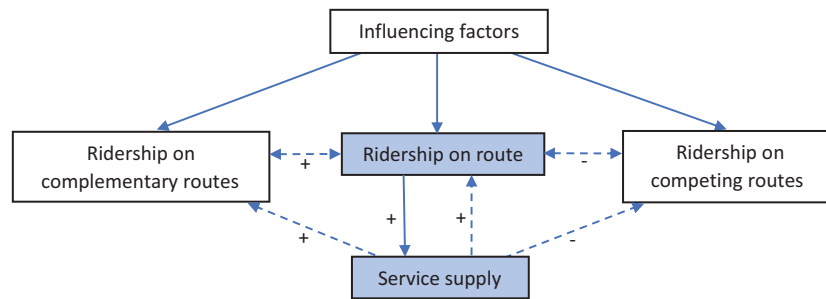


Figure 1 Illustration of ridership-service interactions between complementary and competing routes.

predictions. For example, when linear models are used to model stop-level ridership, [Cervero et al. \(2010\)](#) suggest using hierarchical linear models that account for correlations among stops belonging to the same route.

Complementary and Competing Routes

Service along different routes in a transit network can be competing or complementary. Routes, or portions of routes, that share the same catchment area and serve the same set of origins and destinations compete. For example, two routes that run parallel to each other for any section compete for passengers along that section. In a set of competing routes, service improvements on one route takes away some ridership from the other routes. Therefore, ridership models need to accommodate that predicted increases in ridership on a route is not solely due to new riders attracted to the system but also due to drawing some riders from other routes. Complementary routes are those among which people transfer to get to their destination. The increase in ridership due to service improvements on one route will bring about an increase in ridership on its complementary routes. [Fig. 1](#) below illustrates some of these relationships (also, see [Peng et al., 1997](#)). When modeling transit ridership, it is important to take such substitutive, competitive and complementary effects into account.

Competing and complementary relationships can exist within a same form of transit (for example, between different routes or service categories of a bus service) or across different forms of transit (e.g., between metro and bus transit routes). Of course, such relationships exist between transit and other modes of travel as well. For example, app-based on-demand mobility services can potentially compete with or complement transit systems depending on the spatial and temporal coverage of their services vis-à-vis the transit network and schedules. Therefore, it is important to consider such relationships not only within a transit network but also across different networks and travel modes.

Transfer Boardings and Alightings

Transfers are inevitable in large-scale transit systems. Therefore, not all boardings (alightings) at a stop are due to trip generative (attractive) characteristics of the stop's catchment area. Since the catchment area characteristics do not influence the number of transfers at a stop, modeling transfer boardings (alightings) together with those of direct boardings (alightings) will lead to biased estimation of the influence of service and catchment area characteristics on transit ridership. Therefore, it is important to estimate transfers separately from direct boardings and alightings. This can be done by developing separate models for direct boardings and alightings and transfers ([Chu, 2007](#)). Alternatively, transfers can be estimated at the transit assignment stage after estimating direct boardings and alightings, and the spatial distribution of trips due to direct boardings and alightings.

In several transit systems, a challenge with separating transfer boardings (alightings) from direct boardings (alightings) is that the data collection does not differentiate between the two. Then it becomes necessary to collect additional information on the spatial and temporal distribution of the proportion of boardings (alightings) that are transfers. Such information needs to be used to separate transfers from direct boardings (alightings).

Interdependency Between Demand and Supply

The number of boardings and alightings at bus stops, or the number of riders on a bus route are indicators of transit ridership or demand. Bus frequency or headway on a route, seat miles operated, number of seats per bus, and daily service span are different indicators of the service or supply variables.

It is worth noting here that ridership and service are interdependent. Ridership tends to increase in response to an improvement in service. Sizeable increase in ridership is, in turn, catered to by a consequent increase in service. Besides, some routes have a greater level of service than others because of anticipated greater level of ridership along those routes. For increasing the accuracy of transit ridership prediction, it is useful to consider such mutual interdependency or endogeneity of transit ridership and service attributes. Models that do not account for this relationship overestimate the influence of service improvements on ridership ([Peng et al., 1997](#); [Rahman et al., 2019](#)).

Simultaneous equation models may be used to account for endogeneity between transit demand and supply. [Peng et al. \(1997\)](#) build such a model for Portland's Tri-Met transit agency operations in which major service and route changes occur once a year while minor service and route changes occur whenever needed. Simultaneous transit demand and supply equations for such situations are represented as:

$$R_{iz}^d = F(S_{iz}^s, X_{iz}^d)$$

$$S_{iz}^s = F(R_{iz}^d, R_{-1i}, X_{iz}^s)$$

in which R_{iz}^d and S_{iz}^s are transit ridership (demand) and level of service (supply), for route i in fare zone z . R_{-1i} is the ridership in the previous planning period. X_{iz}^d is the vector of other variables that explain demand. X_{iz}^s is the vector of other variables that explain supply based on the transit operator's decisions. In addition to the above two equations, they also consider another equation to represent demand for other competing routes. The three stage least squares (3SLS) method is used to estimate the parameters of the above simultaneous equations model system as all the endogenous variables under consideration are treated as linear. [Chu \(2007\)](#) states that stop-level models address simultaneity, as simultaneity breaks down at the stop level with service decisions being made at the level of the route. However, even in stop-level ridership models, there is value in considering simultaneity between route-level service characteristics and stop-level ridership for all stops on a given route.

Factors Influencing Transit Ridership

The ridership attracted to transit systems depends on the following two broad factors: (1) external (or exogenous) factors that the transit agency cannot control (at least in the short-term), such as population demographics and spatial land-use of the study area and (2) internal (or endogenous) factors that the transit agency can try to control such as the spatial, temporal, and operational characteristics of the transit service ([Taylor et al., 2009](#)). While a transit agency might not be able to control the external factors, the agency will benefit from monitoring them, anticipate their possible impacts on transit demand, and devise strategies to mitigate or leverage them to their advantage. The external and internal factors together determine transit ridership in the form of two major types of accessibility of the transit system: (1) accessibility of the transit system from a patron's origin or destination; that is, access to population and activity locations within the transit stop's catchment area, and (2) accessibility or connectivity provided by the transit system to major population and employment locations. A brief summary of various factors influencing transit ridership is presented next. In addition, a brief discussion is provided on the factors that influence ridership but not typically considered in direct demand models.

External or Exogenous Factors Influencing Transit Ridership

The exogenous factors influencing transit ridership are listed as follows:

- *Socio-demographic characteristics* of the study area population such as age, gender, race, ethnicity, citizenship or immigrant status, education, employment, income, vehicle ownership, and population attitudes and perceptions;
- *Land-use characteristics* of the study area such as population density, spatial structure and distribution of employment locations, and other trip generators such as entertainment or tourist locations, hospitals, and educational institutions;
- *Attributes of competing modes* such as ownership costs of private modes and usage costs such as gasoline prices, toll prices, maintenance costs, parking availability and pricing; presence of app-based ride hailing modes and the cost and ease of travel by those modes; and commuter travel benefits from employers such as shuttle services and subsidies in transit fares or parking fees;
- *Travel conditions* that include climate and weather patterns, traffic-congestion levels, and disruptions; and
- *Transportation and land-use policies, and funding initiatives* such as zoning regulations, transit-oriented development, air quality mandates, auto emission standards, and state transit funding for capital and operating expenditures.

Internal or Endogenous Factors Influencing Transit Ridership

The endogenous factors may be categorized as follows:

- *Spatial characteristics* such as network coverage, route structure (that is, the extent of transfers, and of competition and complementarity among different routes), stop location and spacing, integration with other modes through feeder services, access and egress infrastructure for intermodal and last-mile connectivity.
- *Temporal characteristics* such as transit frequency, reliability, duration of service, origin-to-destination in-vehicle travel time, waiting and transfer time, access and egress time.
- *Operational characteristics* such as fare structure, ticketing integration, scheduling strategies, and operation of transit priority lanes and signalization.
- *Perceptual characteristics* such as ride quality, comfort, cleanliness, reliability, affordability, safety and security, and the overall level of customer service ([Outwater et al., 2013](#)). Some of these characteristics are determined by factors such as maintenance, extent of crowding, likelihood of denied boarding, provision of real-time information and passenger trip planning applications, on-time arrival, and travel time variance. They are also determined by the effectiveness of infrastructure and procedures for safety and

security. Station or stop-level attributes such as lighting, ventilation, cleanliness, shelter from weather, comfortable waiting, and seating also influence the perception of transit service. Additionally, ease of access to services, safety and surveillance, in-station patron assistance, and ease of buying tickets and passes determine travelers' perceptions of transit service.

Accessibility of the Transit System

The more population and activity centers a transit system is accessible to, the more would be its ridership. The following variables may be used to characterize accessibility of the transit system:

- population density and demographics (defined by income, car ownership, age, gender, etc.) around transit stops or routes within an acceptable walking distance;
- land-use and economic activity around stops or routes in the form of employment in different sectors, measures of activity such as floor area at major trip generators including educational institutions, hospitals, and malls; and
- Measures of connectivity of the transit system to locations that are not accessible by walk, such as connectivity due to first/last-mile feeder services and provision of park-and-ride facilities.

Accessibility Provided by the Transit System

The more activity locations or destinations that a transit system provides access to, the more would be its ridership. Such accessibility can be measured in the form of connectivity of a transit stop (or a route) to various destinations in the network. Stops or routes that offer access or connectivity to a larger number of destinations within a shorter time duration are likely to attract greater levels of ridership. This depends on: (1) the spatial structure of land-use in the study region, (2) the transit route network structure (network coverage, directness of routes, stop location and spacing, ease of transfer among different transit routes), (3) temporal characteristics of transit service such as frequency, reliability (on time arrival performance, travel time reliability), and duration of service, and (4) integration with other modes through feeder services, access and egress infrastructure, etc. for intermodal and last-mile connectivity to enhance transit catchment area.

Note that accessibility of a transit system, discussed in the earlier section, is different from the accessibility provided by the transit system, as discussed in this section. In a stop-level ridership model, for example, the former is a measure of how much population and employment is a given stop accessible to whereas the latter is a measure of how many other destinations can one access from the given stop.

Factors Not Considered in Direct Demand Models

Many direct demand models do not consider the influence of several factors on ridership. One such set of factors is traveler perceptions and attitudes. Research is underway on ways to assess traveler perceptions of transit services (e.g., reliability, safety, comfort, cleanliness, affordability, amenities, customer friendliness) as well as to consider the influence of traveler attitudes (e.g., environment-friendliness and status-seeking behavior) in promoting transit ridership (Outwater et al., 2013). Ridership forecasting models must be enhanced to consider these factors. Equally important is to consider the role of contemporary issues such as that of the current pandemic on travelers' perceptions and its influence on public transit ridership.

Pricing is not included in many ridership models. Transit fare is an important determinant of transit ridership. However, it is challenging to capture the effect of fare changes (in transit or in its competing modes) in direct demand models. This is because of the lack of variation in prices in the time range of typically available transit ridership data. Most studies are conducted using data for a couple of months to at most a year, and transit fares may not necessarily be revised in that duration. Yet, transit agencies need to understand the influence of changes in price structure (not only of the transit systems but also of other modes) on transit ridership.

Finally, it is not easy to incorporate the influence of changes in level-of-service characteristics of other competing modes in direct demand models. In fact, most implementations of direct demand models for transit ridership forecasting do not consider competition with other modes. One way to overcome these limitations is to integrate information from disaggregate mode choice models, which capture the influence of pricing and characteristics of other modes. For example, measures such as log-sum variables from a logit-based disaggregate mode choice model can be used as explanatory variables in a direct demand model. Doing so might help embed the influence of pricing as well as the characteristics of other modes on aggregate level transit ridership.

Model Structure

A variety of different statistical/econometric model structures may be used to model transit ridership. The choice of the model structure depends on the spatial and temporal resolution at which ridership is modeled, the nature of available data, the various model design considerations (as discussed in earlier sections), and a trade-off between model complexity and ease of application.

Data for transit ridership prediction can be of the following types:

- time series or historical data of ridership of a city or an independent transit region;
- real-time data for 1-step (e.g., 15 minute) ahead predictions for arrivals at stations;
- cross-sectional data spanning multiple locations (e.g., bus stops) in the same time duration and
- panel data of a same set of locations from multiple time periods.

Table 1 Summary of transit ridership models used in the literature

<i>Author(s)/ Year</i>	<i>Model approach</i>	<i>Data</i>	<i>Spatial resolution</i>	<i>Time duration for the model</i>	<i>Explanatory variables used in the model</i>
Peng et al. (1997)	Simultaneous equation model	Tri-Met (Portland) service area	Route-segment (by fare zones)	Morning peak, midday, afternoon, evening, and night	Transfers, service duration, headway, service supply, population, income, employment density, parking spaces, previous year ridership, percent population in overlap area of route buffers, frequency, route typology
Chu (2004)	Poisson regression model	Jacksonville, Florida	Stop level	Average weekday boarding	Catchment area variables: median household income, jobs, zero vehicle households, share of persons of different age, gender, race; Composite TLOS variable (transit service, stop catchment area characteristics, stop level pedestrian environment) pedestrian factor, persons and jobs up and downstream without transfer, trolley stop, TLOS stops (four stops at an intersection)
Chu (2007)	Poisson and negative binomial regression models	Jacksonville, Florida	Stop level (direct and transfer boarding models)	Weekday morning and afternoon peak, weekends	Direct boarding models: natural log of frequency, daily service span, workers, population in origin buffer, vehicle ownership, service and commercial employment, park-n-ride lot spaces, stop on route typology, accessibility to population and employment
Cervero et al. (2010)	Hierarchical linear regression	BRT services in Los Angeles county	Boardings (and /or exits) at a stop/station	Daily recordings for October 2008	Service attributes, location and neighborhood attributes, bus stop attributes
Sarlas and Axhausen (2014)	Spatial autoregressive models	Road network of North-East Switzerland	Link level	Morning peak hour (8–9 p.m.) of a typical weekday	Spatially resolved road density, public transport network densities, densities of ramp links. Population and employment, kernel densities
Chakour and Eluru (2016)	Composite marginal likelihood based ordered response probit model	Public Transit Network of Montreal	Stop level	Morning peak, off-peak, evening peak, night	Stop-level transit operational variables (e.g., average headway for time-period), public transit accessibility indices (e.g., number of bus/metro/train stops around each bus-stop), transportation infrastructure attributes (e.g., road length by functional classification), stop-level built environment (e.g., employment density and walk-score).
Mucci and Erhardt (2018)	Linear regression model	GTFS data (2009) from MUNI transit system operating in San Francisco	Route-stop level model	Daily average number of bus/rail boarding and alighting	Demand variables: employment near stop, income, housing and population density, share of households with zero vehicles and on-street parking cost. Supply variables: average number of buses, number of stops within walking distance of route-stop, terminal stop of bus line, route configuration, reliability.
Rahman et al. (2019)	Simultaneous equation model	Greater Orlando region	Stop level	Average weekday boarding, alighting, and headway	Stop level attributes, transportation and transit infrastructure variables, built environment and land use attributes, demographic and socio-economic variables in vicinity of the stop: income, vehicle ownership, age and gender distribution.

In naïve time-series models, a curve is fitted to the historical ridership data and projected to obtain estimates for the future. It may also be regressed against explanatory variables such as jobs, income, fare, and fleet size (McLeod et al., 1991). Cross-sectional data across bus stops or routes of a single city is widely used to estimate transit ridership (Chu, 2004; 2007; Peng et al., 1997; Cervero et al., 2010; Chakour and Eluru, 2016). Panel data models are used to recognize that the data may have come from multiple time periods from the same set of locations (Rahman et al., 2019).

Although intended for continuous data, linear regression models are often used in estimating direct demand models of transit ridership counts. Poisson and negative binomial models are used to recognize the count data nature of transit ridership. In the case of over-dispersed (variance larger than mean) data, negative binomial models are preferred. It is also possible to use ordered response models, where ridership is grouped into different levels such as low, moderate, high, and very high.

Regardless of the type of the dependent variable used for modeling ridership, considering dependencies in ridership among proximate stops would need one to incorporate spatial correlations (Chakour and Eluru, 2016; Sarlas and Axhausen, 2014). Further, accommodation of endogeneity between demand (ridership) and supply (service) variables needs econometric approaches such as three stage least squares or simultaneous equations (if the demand and supply variables are treated as continuous) or control function approaches (if the variables are analyzed using non-linear models). The past decade has witnessed the use of machine learning, particularly for predictions using near real-time data (Koutsopoulos et al., 2019) such as crowding prediction (Noursalehi, 2017).

The reader is referred to **Table 1** for some examples of different transit ridership models used in the literature, their spatial and temporal resolutions, and the variables used to explain ridership in those studies.

Emerging Trends and Future Directions

The emerging mobility landscape, particularly the market penetration of on-demand ride hailing and ride sharing services, the rapid rise of automation in transportation, and the influences of information and communication technologies on travel influence travel behavior and transit ridership. Some notable trends with anticipated influence on transit ridership include the following:

- the availability of real-time passenger information and trip-planning apps,
- automated and integrated fare collection and payment systems,
- use of app-based on-demand Mobility as a Service (MaaS) modes for enhancing first- and last-mile connectivity of fixed route transit systems, and
- technology-assisted safety features such as in-vehicle cameras and remote monitoring.

Unlike the socio-demographic and land use changes (and their influence on transit ridership), which occur relatively slowly, some of the above-mentioned factors have altered travel behavior in a very short time and will have a substantial influence on transit ridership. Transit ridership forecasting models must be enhanced to accommodate the influence of such trends.

Many transit systems are now equipped with automated vehicle location (AVL) and automated fare collection (AFC) systems that generate large amounts of historical as well as real-time data on transit ridership and system performance. There is scope for combining such “big data” sources with traditional data sources for understanding and predicting ridership. For example, ridership and AVL data can be combined to distinguish between locations of designated or informal stops for passenger boarding or alighting and stops due to traffic congestion. There is also scope for developing transit ridership models combining land-use, weather, events, and other big data sources. Existing sources of big data often do not provide information such as socio-demographic characteristics of passengers. Integrating data from small surveys would be a potential direction to undertake for efforts needing such information (Zannat and Choudhury, 2019).

Acknowledgments

This work was funded by the Ministry of Electronics and Information Technology (MeitY) of the Government of India under a grant titled “*Collaborative Intelligent Transportation Systems Endeavour for Indian Cities – InTranSE Phase II*” and by the Department of Science and Technology (DST) of the Government of India under the “*IMPacting Research, INnovation and Technology (IMPRINT)-II*” scheme.

References

- Cervero, R., Murakami, J., Miller, M., 2010. Direct ridership model of Bus Rapid Transit in Los Angeles County, California. *Transp. Res. Rec.* 2145, 1–7.
- Chakour, V., Eluru, N., 2016. Examining the influence of stop level infrastructure and built environment on bus ridership in Montreal. *J. Transp. Geogr.* 51, 205–217.
- Chu, X., 2004. Ridership models at the stop level. NCTR-473-04, BC137-31. National Center for Transit Research, University of South Florida. Research report prepared for the Florida Department of Transportation (FDOT). Available from: <https://www.nctr.usf.edu/2004/12/transit-ridership-models-at-the-stop-level/>.
- Chu, X., 2007. Validating T-BEST models with 100% APC counts. NCTR-576-07, BD549-19. National Center for Transit Research, University of South Florida.
- Gutiérrez, J., Cardozo, O.D., García-Palomaresa, J.C., 2011. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *J. Transp. Geogr.* 19, 1081–1092.

- Koutsopoulos, H.N., Ma, Z., Noursalehi, P., Zhu, Y., 2019. Transit data analytics for planning, monitoring, control, and information. In: Antoniou, C., Dimitriou, L., Pereira, F. (Eds.), *Mobility Patterns, Big Data and Transport Analytics: Tools and Applications for Modeling*. Elsevier, Amsterdam Netherlands, pp. 229–261.
- McLeod Jr., M.S., Flannelly, K.J., Flannelly, L., Behnke, R.W., 1991. Multivariate time-series model of transit ridership based on historical aggregate data: the past, present and future of Honolulu. *Transp. Res. Rec.* 1297, 76–83.
- Mucci, R.A., Erhardt, G.D., 2018. Evaluating the ability of transit direct ridership models to forecast medium term ridership changes: evidence from San Francisco. *Transp. Res. Rec.* 2672 (46), 21–30.
- Noursalehi, P., 2017. Decision support platform for urban rail systems: real-time crowding prediction and information generation. Ph.D. dissertation, Department of Civil and Environmental Engineering, Northeastern University.
- Ortuzar, J.D., Willumsen, L.G., 2011. *Modelling Transport*, fourth ed. Wiley Publications, Chichester.
- Outwater, M.L., Sana, B., Ferdous, N., Bhat, C.R., Sidharthan, R., Pendyala, R.M., Hess, S., Woodford, W., 2013. TCRP Project H-37 - Characteristics of premium transit services that affect mode choice: key findings and results. Technical paper, Resource Systems Group, White River Junction, VT.
- Peng, Z., Dueker, K.J., Strathman, J., Hopper, J., 1997. A simultaneous route-level transit patronage model: demand, supply, and inter-route relationship. *Transportation* 24, 159–181.
- Pinjari, A.R., Bhat, C.R., 2011. Activity-based travel demand analysis. In: de Palma, A., Lindsey, E., Quinet, E., Vickerman, R. (Eds.), *A Handbook of Transport Economics* 10. Edward Elgar, Cheltenham, pp. 213–248.
- Polzin, S., Chu, X., Bunner, R., Pinjari, A.R., 2011. TBEST model enhancements—parcel level demographic data capabilities and exploration of enhanced trip attraction capabilities. Research report prepared for the Florida Department of Transportation (FDOT).
- Rahman, M., Yasmin, S., Eluru, N., 2019. Controlling for endogeneity between bus headway and bus ridership: A case study of the Orlando region. *Transp. Policy* 81 (C), 208–219.
- Sarlas, G., Axhausen, K.W., 2014. Towards a direct demand modeling approach. In 14th Swiss Transport Research Conference. Conference paper STRC, Monte Verita/Ascona.
- Taylor, B.D., Miller, D., Iseki, H., Fink, C., 2009. Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas. *Transp. Res. Part A Policy Pract.* 43 (1), 60–77.
- Zannat, K.E., Choudhury, C.F., 2019. Emerging big data sources for public transport planning: a systematic review on current state of art and future research directions. *J. Indian Institute Sci.* 99 (4), 601–619.

Further Reading

Florida Department of Transportation (FDOT)'s TBEST Transit Planning Software. Available from: <https://tbest.org/>.

Spatial Mismatch, Job Access, and Reverse Commuting

Gian-Claudia Sciara, School of Architecture, The University of Texas at Austin, Austin, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Spatial Mismatch, Job Access, and Reverse Commuting	468
Emergence of the Spatial Mismatch Hypothesis	468
Serving the Reverse Commute	469
Expanding, Evolving, and Challenging Spatial Mismatch	469
Measuring Spatial Mismatch	470
Job Access	471
Spatial Mismatch in Other Contexts	471
References	471

Spatial Mismatch, Job Access, and Reverse Commuting

Since the latter 20th century, transportation policymakers, planning practitioners, and researchers have maintained steady interest in the impact that access to jobs may have on the employment outcomes of individuals—particularly economically or socially disadvantaged individuals. To what extent, if any, does the spatial arrangement of employment and housing opportunities support or hinder one's ability to find and keep a job? Do restrictions on residential mobility resulting from racial discrimination diminish job access and employment? How important are challenges in transportation and commuting to employment outcomes? Such questions are persistent concerns in an abundant literature examining such related topics as spatial mismatch, job access, reverse commuting, jobs-housing balance, and transportation mismatch.

Interest in these areas is rooted largely, but not exclusively, in the context of U.S. transportation planning in the latter 20th century, and in the distinctive spatial patterns of postwar American metropolitan growth and employment of persistent residential segregation. This essay traces the emergence of these concepts in the wider domain of urban policy and planning and highlights how practitioners and scholars have continued to observe and measure them and to debate their salience.

Emergence of the Spatial Mismatch Hypothesis

Policymakers and scholars first began to consider the importance of potential linkages between transportation and employment in the 1960s. Economist John Kain articulated the first threads of what is known as the spatial mismatch hypothesis in a seminal conference paper (1965, 1992). He aimed to explain the lower levels of employment among American Blacks concentrated in the urban core and Blacks' underrepresentation in employment zones far from Black urban enclaves.

Kain highlighted how racial discrimination in the housing market limited Blacks' ability to exercise residential choice. This not only led Blacks to be spatially concentrated in "urban ghettos" but also consequently impeded their ability to share in the employment growth spreading from cities to suburbs at the time. The spatial mismatch between the "the Negro ghetto" and the locations of expanding employment cost Blacks jobs, Kain argued.

Kain's ideas about spatial mismatch gained purchase in wider discussions of urban policy in the mid-1960s to early 1970s. During this period, an unprecedented wave of Black-initiated urban rioting and violence erupted in hundreds of cities across the U. S., symbolized most vividly perhaps by the Watts Riots in Los Angeles (1965), the Newark riots (1967), and the Detroit riots (1967). Though specific causes of individual riots were debated, cumulatively the riots were important means through which urban Blacks protested the fundamental conditions of ghetto life. These included high rates of unemployment and rigid residential segregation (Fogelson, 1967, 353).

A Federal commission appointed to study the causes of Black civil unrest pointed in part to the challenges of the urban ghetto and its physical isolation from wider economic opportunity (U.S. Kerner Commission, 1968). Federal and state policymakers directed attention to how transportation might better connect urban Blacks with suburbanizing jobs. Interest in "reverse commuting" first arose in this context (Rosenbloom (1992).

As post-WWII employment and population growth concentrated increasingly in suburbs, many central city workers had to seek jobs in the suburbs or further exurbs. Construction of the Interstate Highway System and rising levels of household automobile ownership enabled employers to establish themselves in the suburbs, which would quickly transition from mere bedroom communities dependent on central city jobs to centers of employment and commerce in their own right (Muller, 2017).

At the same time, holdover racial covenants and exclusionary zoning present in many suburban communities precluded Blacks from establishing residence there (Whittemore, 2017). Further, discriminatory mortgage lending practices like redlining made it difficult for blacks to secure legitimate home loans from government lending programs. Consequently, many turned to exploitative "contract sales" to purchase their own urban homes, albeit on highly unfavorable terms. Together, these and other practices all but

ensured that most American suburbs remained white, that few Blacks would amass the equity needed for home purchases, and that most Blacks would remain in inner city neighborhoods, far from where new jobs were being added.

Serving the Reverse Commute

Access to suburban jobs was essential to un- and underemployed inner city residents, but it was also potentially challenging. For workers unable or unwilling to move beyond the city to follow the jobs, reverse commuting from central city to suburb was one answer. Reverse commutes accounted for roughly 7% of all trips in the 1960s. Yet, reverse commutes to suburban job locations would typically be longer in distance than CBD-bound commutes and would require more time and expense. Dispersed suburban job sites were also less likely than urban jobs to be served by public transit. Many workers turned to the private car, if they had the means to do so (Rosenbloom, 1992).

Acknowledging these challenges, the Federal government sponsored a number of demonstration projects in the mid- to late 1960s to establish better transit links between the inner city and suburban jobs. Well documented by Rosenbloom (1992), examples included the Shirley Highway Express Bus in Washington D.C. and the Chicago express bus route carrying airport employees from the end of a rapid transit line to O'Hare.

Still, these demonstrations generally yielded modest benefits, revealing how difficult it was to improve employment outcomes for inner city residents through transportation. Project assessments suggested that a range of factors beyond transportation might be keeping inner city minorities from better suburban jobs, including job search and dissemination practices, high employee turnover, mismatch between vacancies and applicant skills, not to mention naked discrimination. It was also hard for local transit operators to make reverse commute service operationally viable. Traditional transit operators already faced declining ridership and rising costs, and they struggled to serve reverse commutes given "the sprawling automobile orientation of suburban . . . work locations, [and workers'] shift times, and overtime requirements" (Rosenbloom, 1992, 11).

By the 1980s, direct Federal efforts to address unemployment through transportation had diminished. Instead, though transportation was outside of their mission, not-for-profit social and human service agencies increasingly provided free or low cost transportation to connect inner city residents with suburban jobs. Private employers also funded and coordinated employee shuttles to suburban job sites, acting through nonprofit Transportation Management Associations (TMAs). Tenant management associations organized transportation services helping public housing residents to access jobs. To a limited extent, some public transit operators provided feeder buses to and from suburban rail stations and other niche services to support reverse and off-hours commutes. Most of these nonprofit and transit operator services, however, had limited success in reversing high joblessness among inner city residents (Rosenbloom, 1992).

In the 1990s, as federal welfare reform transitioned recipients of public aid into local labor markets, the U.S. government began to encourage local collaborations to enhance job access for disadvantaged urban populations via transportation. The signature Job Access Reverse Commute (JARC) program, which remained in place through 2012, offered Federal grants to support a range of planning, capital and operating costs for public transit projects connecting to suburban workplaces; funds for reverse commutes services and travel vouchers were available too. Projects demonstrating collaboration and cost sharing among state and regional transportation agencies, transportation providers, county welfare and employment agencies, and other community stakeholders were highly favored.

Experience with JARC and earlier reverse commute efforts suggested that community transportation services and were not enough to improve employment outcomes for unemployed and minority inner city residents. JARC successfully stimulated nontraditional partnerships to address transportation needs of low-income job seekers, but the diverse organizational perspectives among collaborators made it difficult to identify and plan for the transportation needs of targeted groups (Blumenberg, 2002). More significantly, JARC funds could be used only for public transportation, not for automobile-related expenses (Blumenberg and Schweitzer, 2006). Program observers highlighted this limitation, heralding rising skepticism that public transit could adequately bridge the spatial mismatch between inner city residents and spatially dispersed job opportunities.

Emerging demographic trends further suggest limited potential for public transit in general or reverse commute options in particular to improve employment outcomes for low-income populations. Fast growing poverty in American suburbs since the 2000s has captured the attention of policymakers and planners. Poverty in the U.S. has traditionally been associated with the urban core and rural communities, and it continues to expand in those places; however, the number of suburban residents living in poverty has increased at a starkly faster rate in the 21st century. Expanding suburban poverty and continuing decentralization of jobs together pose new transportation challenges. Most transit systems provide central city-oriented hub-and-spoke service and do not provide the suburb-suburb connections important for reaching dispersing jobs (Kneebone, 2017). Such trends may add to calls for enhanced automobile access for low-income job seekers.

Expanding, Evolving, and Challenging Spatial Mismatch

In the decades since Kain first articulated the theory, researchers have questioned the extent to which poor labor market outcomes experienced by inner city residents, particularly those of color and/or low income, are attributable to "spatial mismatch." Most researchers have agreed that the spatial disconnect between job rich areas and urban neighborhood residents who need those jobs was real. Still, many have disagreed about the relative importance of spatial mismatch and used different approaches to measure it.

Some have viewed the emphasis on spatial causes of poverty as downplaying the role of racial discrimination. To understand Black urban unemployment, Ellwood emphasized, “It’s not space, it’s race” (1983). Other work has questioned whether an emphasis on race obscured important insights into the operation of labor markets, particularly the operation of those markets in the face of suburbanizing employment and housing discrimination. How do workers with certain characteristics and residing at a certain location ultimately match with jobs with their own specific characteristics and locations?

Others concluded that mismatch had been oversold as a locational problem and that other reasons for poor employment outcomes among blacks and low-income groups were important. Declines in employment rates of blacks and black males in the 1970s should be viewed in the context of “industrial shifts, technological change, declining relative quality of black education, [and] weakening antidiscrimination efforts by government” (Holzer, 1991, 108).

A prominent stream of research has suggested that low-income inner city residents are challenged less by the spatial mismatch distancing them from jobs, and more by the “transportation mismatch” or “modal mismatch” preventing them from bridging the distance to work sites, far or near. Job seekers who lived in job-poor neighborhoods and who relied on public transit experienced significantly less access to employment *because of* long and unreliable transit commutes. Studying welfare participants in Los Angeles, Blumenberg and Ong (2001) found that—despite suburban employment growth—the bulk of jobs were actually located *in the central city*; yet, residents of job-poor neighborhoods had trouble accessing those jobs if they depended only on public transit.

Further insights called attention to the diversity of settings in which welfare recipients and other low-income job seekers live. Traditional conceptualizations of “spatial mismatch” focused narrowly on the central city. What about low-income households and job seekers living in older inner-ring suburbs and non-urbanized areas and challenged by low concentrations of jobs? Such settings made access to private vehicles important for employment. Other transportation and policy approaches might be appropriate as well, including dynamically routed transit, express buses, as well as local economic development (Blumenberg and Hess, 2003).

Such insights catalyzed calls for programs to increase automobile access for welfare recipients and low-income job seekers. Later research substantiated such concerns. One longitudinal evaluation of U.S. welfare recipients transitioning to employment found that access to an automobile was a better predictor of subsequent employment than was access to public transportation or housing assistance (Blumenberg and Pierce, 2017). Automobile access also proved important in a longitudinal study of the U.S. government’s landmark Moving to Opportunity program, which provided housing assistance for participating low-income households to move out of low-income neighborhoods in order to improve their employment outcomes. While housing mobility proved inconsequential, households that gained or retained access to an automobile during the study period were more likely to be employed, suggesting that “having a car provides multiple benefits that facilitate getting and keeping a job” (Blumenberg and Pierce 2014, 52). Further, transit access did not necessarily produce job gains, though it seemed to help already employed individuals retain their jobs.

Another longitudinal study by Hu (2015) examined the Los Angeles region between 1990 and the late 2000s and concluded that when job accessibility is measured by the spatial distribution of jobs and job-seekers traveling by car, that the inner-city poor in fact do not face spatial mismatch and that they enjoy higher job accessibility than most suburban poor (2015). Accordingly, she questions whether traditional answers to spatial mismatch, including moving people to the suburbs, bringing jobs to the inner city, or providing mobility options will be effective (Hu, 2015, 44). Conversely, a separate 2018 study of person-specific job accessibility and the duration of joblessness provides compelling support for the mismatch hypothesis, suggesting that scholars will continue to debate how best to state and test for the spatial mismatch problem (Andersson et al., 2018)

Measuring Spatial Mismatch

In the abundant literature around spatial mismatch, researchers have used a variety of approaches to identify ‘mismatch’ and measure job access. Houston provides a thoughtful summary (2005). Early research into spatial mismatch used simple measures of residential segregation to suggest that, where a highly segregated community also experienced high unemployment, spatial mismatch was at work. Subsequent studies aimed to account for employment access. Researchers made cross sectional comparisons of average commuting times and distances for Blacks and Whites, and observed mismatch where average commute times (or distances) for Blacks exceeded those for Whites. Critics, however, suggested that the average commute times and distances approach failed to account for different travel modes used by Blacks and Whites, for income differences between the two groups, or for their different abilities to trade off housing costs and proximity to employment. Alternatively, some studies have compared the average wages of African-Americans in the city with those in the suburbs, though data limitations have made this difficult to do well.

Studies in the 1990s began to estimate measures of job proximity, a clear step forward though one still limited by data and technical constraints (Houston, 2005). For instance, it is difficult to capture proximity accurately, as measures must take account of competition for jobs, travel modes and mode-specific travel times, as well as the labor market segment of interest.

Houston’s review provides perhaps most clearly distills the basic direction for spatial mismatch research. “In methodological terms, the first, and inescapable, stage in testing the spatial mismatch hypothesis is to measure the extent of spatial mismatch by examining the location of jobs in relation to the location of people, differentiating between labor market segments in terms of occupation, skill level, gender, race, and those with and without access to a car. The second stage is to determine what effect spatial mismatch has on these groups in terms of unemployment and wage levels” (2005, 430–431).

Job Access

Measuring job accessibility has been a key focus of the spatial mismatch literature and has also attracted attention from the fields of urban sociology, regional science, and geography in studies of the geography of opportunity or “opportunity structures” across multiple dimensions, including labor. Such sustained interest in job access has yielded numerous innovations and improvements in its measurement.

The earliest approaches to gauging job access used gravity-based and cumulative opportunity accessibility measures, derived for a given spatial unit, such as a census tract or traffic analysis zone (TAZ). Cumulative opportunity measures are perhaps the simplest; these tally the number of employment opportunities that reachable within a defined commute-shed, based on time or distance. All destinations that fall within the commute-shed threshold contribute equally to accessibility. Gravity-based measures typically estimate job accessibility by calculating the sum of employment opportunities reachable within a travel distance or time threshold, discounted by an impedance factor representing the generalized cost of travel, including time or distance. Thus, employment opportunities count unequally in gravity-based job accessibility (El-Geneidy & Levinson, 2006; Handy and Niemeier, 1997).

Contributions to the spatial mismatch literature have developed increasingly specific measures of job access. By finely-tuning job accessibility measures, researchers have aimed variously to more accurately represent workers or job-seekers, the locations of travel origin and destination, travel modes used, and job opportunities of interest (e.g., by wage-level or economic sectors).

Transit-based job accessibility has been a particular point of interest. Blumenberg and Ong (2001), for instance, measured levels of access for (largely female) welfare recipients in Los Angeles specifically to low-wage, feminized jobs, reachable by automobile and by public transit. More recently, others have sought to document patterns of transit usage to low-wage employment zones by workers in different occupations, finding that different occupations align with different patterns of transit usage (Legrain et al., 2016). The emergence of General Transit Feed Specification (GTFS) data, published by growing numbers of transit properties and capturing schedule and geographic information in a common format, has enabled additional measurement advances, including more universal assessments of job access by transit. Such data have been used to derive route-level accessibility measures for different demographic groups (Karner, 2018), to examine how transit network changes impact job access of demographic subgroups (Swayne and Miller, 2018), and to account for temporal variation in transit accessibility over a 24-h period (El-Geneidy et al., 2016). Other advances in accessibility measurement include person-specific job accessibility, in contrast to conventional cross-sectional approaches that aggregate individuals by neighborhood (Andersson et al., 2018).

Spatial Mismatch in Other Contexts

While concern with spatial mismatch first arose in the context of suburbanizing and segregated postwar America, the topic—and related issues of job access and reverse commuting—has emerged in other national contexts too. Examples mentioned below, while not comprehensive, illustrate the extended interest in mismatch. Recent scholarship has applied the spatial mismatch framework to rapidly urbanizing China. Different urbanization processes in China have not produced the inner same city–suburban contrasts or pronounced income-based residential segregation as are visible in the U.S. However, the spatial concentration of essential services and opportunities in central areas, compounded in importance by residential permitting systems or *hukou*, may be problematic for low-income populations and rural migrants living in peripheral areas. In Britain, a stream of literature related to spatial mismatch emerged in the 1970s to examine the “entrapment hypothesis” that low-income residents of public housing were disadvantaged in the labor market (for an overview, see Houston 2005). Finally, absent the U.S. context of segregation and inequality, reverse commuting by regional rail from central Stockholm to satellite new towns facilitated economic integration as the region transformed from a monocentric city to a polycentric metropolis in the latter 20th century (Cervero, 1998).

References

- Andersson, F., Haltiwanger, J.C., Kutzbach, M.J., Pollakowski, H.O., Weinberg, D.H., 2018. Job displacement and the duration of joblessness: the role of spatial mismatch. *Rev. Econ. Stat.* 100 (2), 203–218.
- Blumenberg, E., 2002. Planning for the transportation needs of welfare participants: institutional challenges to collaborative planning. *J. Plan. Educ. Res.* 22 (2), 152–163.
- Blumenberg, E., Hess, D.H., 2003. Measuring the role of transportation in facilitating welfare-to-work transition: evidence from three California counties. *Transport. Res. Rec.* 1859 (1), 93–101.
- Blumenberg, E., Ong, P., 2001. Cars, buses, and jobs: welfare participants and employment access in Los Angeles. *Transport. Res. Rec.* 1756 (1), 22–31.
- Blumenberg, E., Pierce, G., 2014. A driving factor in mobility? Transportation's role in connecting subsidized housing and employment outcomes in the moving to opportunity (MTO) program. *J. Am. Plan. Assoc.* 80 (1), 52–66.
- Blumenberg, E., Pierce, G., 2017. The drive to work: The relationship between transportation access, housing assistance, and employment among participants in the welfare to work voucher program. *J. Plan. Educ. Res.* 37 (1), 66–82.
- Blumenberg, E., Schweitzer, L., 2006. Devolution and transport policy for the working poor: the case of the U.S. job access and reverse commute program. *Plan. Theory Pract.* 7 (1), 7–25.
- El-Geneidy, A., Buliung, R., Diab, E., van Lierop, D., Langlois, M., Legrain, A., 2016. Non-stop equity: Assessing daily intersections between transit accessibility and social disparity across the Greater Toronto and Hamilton Area (GTHA). *Environ. Plan. B: Plan. Design* 43 (3), 540–560.
- El-Geneidy, A.M., Levinson, D.M., 2006. Access to destinations: Development of accessibility measures. <https://conservancy.umn.edu/handle/11299/638>.
- Ellwood, D.T., 1983. The spatial mismatch hypothesis: Are there teenage jobs missing in the ghetto? Working Paper No. 1188. National Bureau of Economic Research, Cambridge, Mass.
- Fogelson, R.M., 1967. White on Black: a critique of the McCone Commission Report on the Los Angeles riots. *Polit. Sci. Quart.* 82 (3), 337–367.
- Handy, S.L., Niemeier, D.A., 1997. Measuring accessibility: an exploration of issues and alternatives. *Environ. Plan. A* 29 (7), 1175–1194.
- Holzer, H.J., 1991. The spatial mismatch hypothesis: what has the evidence shown? *Urban Stud.* 28 (1), 105–122.

- Houston, D.S., 2005. Methods to test the spatial mismatch hypothesis. *Econ. Geogr.* 81 (4), 407–434.
- Hu, L., 2015. Job accessibility of the poor in Los Angeles: has suburbanization affected spatial mismatch? *J. Am. Plan. Assoc.* 81 (1), 30–45.
- Kain, J.F., 1965. The Effect of the Ghetto on the Distribution and Level of Nonwhite Employment in Urban Areas (No. P-3059-1). Rand Corp. Santa Monica, California.
- Kain, J.F., 1992. The spatial mismatch hypothesis: three decades later. *Housing Policy Debate* 3 (2), 371–460.
- Karner, A., 2018. Assessing public transit service equity using route-level accessibility measures and public data. *J. Transport Geogr.* 67, 24–32.
- Kneebone, E., 2017. The changing geography of US poverty. Brookings Institutions February.
- Legrain, A., Buliung, R., El-Geneidy, A.M., 2016. Travelling fair: Targeting equitable transit by understanding job location, sectorial concentration, and transit use among low-wage workers. *J. Transport Geogr.* 53, 1–11.
- Muller, P., 2017. Transportation and Urban Form: stages in the spatial evolution of the American Metropolis. In: Hanson, S., Giuliano, G. (Eds.), *The Geography of Urban Transportation* (57–85). Guilford Press, New York.
- Rosenbloom, S., 1992. Reverse commute transportation: Emerging provider roles (No. DOT-T-93-01). U.S. Department of Transportation, Washington, DC.
- Swayne, M., Miller, M., 2018. Innovation on Job Accessibility with General Transit Feed Specification (GTFS) Data. Final Report METRANS Project 16-08. Metrans Transportation Center. https://www.metrans.org/assets/research/Tier-1-UTC_16-08_Final-Report.pdf.
- U.S. Kerner Commission, 1968. Report of the National Advisory Commission on Civil Disorders. U. S. Government Printing Office.
- Whittemore, A.H., 2017. The experience of racial and ethnic minorities with zoning in the United States. *J. Plan. Lit.* 32 (1), 16–27.

Microsimulation and Agent-Based Models in Transportation

Milos Balac, Stefano-Francini-Platz 5, Institute for Transport Planning and Systems, ETH Zurich, Zurich, Switzerland

© 2021 Elsevier Ltd. All rights reserved.

The Microlevel in Transportation Simulations	473
Why Microlevel?	474
Examples of Microlevel Transport Simulation Models	475
Nothing is Perfect	475
Final Remarks	476
References	476
Further Reading	476

The transportation system in our society today is made up of complex large-scale interactions that are driven by behaviors of travelers as they interact with each other and their dynamic environment in a bid to get from one place to the other. Complex systems can generally be explained on a macrolevel, together with their functionality; however, many of the explanatory details are lost. This becomes especially important if improvements or changes to the system are to take place. To explain the concept we can take a laptop computer as an example. You know how to start the laptop, operate it, use it to write text, watch movies, design webpages, develop applications, or even design transport policies. Most of you reading this are probably not concerned how a computer is able to execute the operations a user requests. However, for those curious enough or those that need to perform fundamental changes or improvements to the system, a lower level of understanding is required. On a macrolevel one might want to add additional memory to improve the performance of the laptop. However, to increase the performance of the memory itself, one needs to go to microlevel, to circuit design. Therefore, one can start from a macro perspective, through meso-, to finally arrive at a microscopic perspective, depending on the focus of usage or improvements one wants to be able to perform.

The same analogy can be applied to transportation. As the general users of laptop computers the same applies to everybody stepping out of their homes and using the transport network to travel. Adding road lanes to increase the capacity of the network is similar to adding memory to a computer. Finally, understanding the effects of vehicle or individual behavior on the system shifts the view to the microlevel similar to the circuit design for computers.

This leads us to the first important observation in this chapter: Understanding a system on a macrolevel is in general important and can be useful to both users and engineers; however, complete understanding and possibility to do fundamental changes come from a micro perspective.

In traffic modeling, traditional models have in general been aggregated and have observed travel on a macrolevel. This results in the loss of the link to the individual travelers or vehicles. This is where four-step models, which model traffic flows on an aggregate level between the predefined zones start to exhibit limitations. Microsimulations of traffic and agent-based models of individual travelers tend to overcome this limitation by focusing on individual vehicles and travelers, respectively.

The rest of this chapter will guide you through the microlevel of modeling in transportation, the reasons behind its relevance, and some of the current micro-modeling approaches both commercial and academic. The final section will highlight some of the shortcomings and best practices for the use of simulation in transportation.

The Microlevel in Transportation Simulations

Before we start we need to define two simulation approaches in transportation that will be the focus of this chapter: microsimulation of traffic flows and agent-based models.

Microsimulation of traffic flow considers vehicles as particles on the road links which are moving through the network according to the rules of traffic. These rules include, but are not limited to the following:

- traffic lights, which define whether a vehicle is allowed to cross an intersection
- turning rules defining to which road a vehicle can go on after approaching an intersection
- vehicle separation rules, defining how far apart vehicles need to be at different speeds or congestion levels
- lane changing rules, defining when and how it is acceptable to change the driving lane
- parking rules defining an approach to for example on-street parking spot and maneuvers for parallel, diagonal or perpendicular parking, which can affect the traffic in different ways, etc.

Some of the rules or regulations in traffic are imposed by infrastructure, like traffic lights or turning restrictions, while some are informal like vehicle separation, lane changing or merging behavior. The latter are usually not uniform across the drivers and simulations try to capture this by including stochasticity in their models. The rules and driving behavior explained above are a part of

the rules that define how a driver would behave in the physical world. Traffic microsimulation therefore aims to mimic these rules in the simulated world. Through this, microsimulation allows to study the complex nature of vehicle interaction on road networks and their impacts on a system as a whole.

Each vehicle is also characterized by certain performance metrics, making each vehicle unique in the simulation. Some examples of these metrics include: driving characteristics like maximum speed, acceleration or deceleration, length and width which define the space it occupies on the road, etc.

Vehicle origins, destination, and routes are generally generated externally and assigned to the network, where traffic conditions are then analyzed.

Agent-based models in transportation consider both individual travelers and vehicles on a microlevel. However, in contrast to the microsimulation of traffic, agent-based models are more focused on persons and their behavior than on the vehicles and traffic rules that they are required to follow. Each individual person in an agent-based model is represented by a so called agent. Each agent usually contains a sequence of activities it needs to perform within a certain period of time. This period of time can be hours, days, weeks, or even months. To reach all of the activities an agent has to travel and in this way interact with other travelers and compete for the resources on the transportation network. The resources that travelers compete for can be in the form of space on the road or on the bus, or also for limited number of car-sharing vehicles available in the system.

To allow the agents to make choices in the system based on the state of transportation system and choices of other agents, mode-choice and location-choice models are usually fully integrated in agent-based models.

Each agent, or in a physical world a traveler, contains certain information that makes it unique, as is the case for vehicles in the case of traffic microsimulations. Some examples of the various information an agent can contain is: unique id, age, gender, income, certain preferences or attitudes toward transportation modes, etc. Since agent-based models in transportation are more focused on travelers behavior, they usually model traffic on a higher level, where many of the traffic rules are approximated.

It is also important to note that they can also be combined into a single simulation platform. In this way both vehicles and travelers would be considered as particles that follow certain rules, be it traffic or behavioral. Due to the large complexity of simulating both worlds on a microlevel, these simulations are computational very slow and are currently rarely seen in practice due to this limitation.

Why Microlevel?

Now that we have gone through the differences between the two types of transport simulation, it is essential to understand why they are important.

Before the advent of the car, the speed of individual travelers in densely populated areas was bound to the speed of foot and hoof.¹ As the speed of travel was slow, modeling traffic or pedestrian flows was unnecessary. With the invention of a car, the speed on the roads substantially increased and with their increased affordability the number of vehicles on the roads became substantial. This has consequently created the need for more sophisticated traffic rules. As the experimental approach to test innovative traffic rules, traffic light signaling, or intersection types in different road networks is both time and cost inefficient, a simulation was seen as a better alternative. As the power of computers increased, it allowed for transport planners to develop efficient simulation traffic models. These simulations allow for the analysis of a multitude of design solutions before implementing them in practice, reducing both time and cost.

The progress of technology and an advent of the smartphone has also brought many changes to how people move between points in the network. In recent decades we have witnessed the start of car-sharing and bike-sharing services, followed by ride-hailing, and more recently e-scooter sharing services. Much research has also been focused on automated vehicles and their operations. Aggregated models were just not ready to handle these emerging mobility solutions, as they are neither available at all times like a private car nor are they scheduled like public transport. These services usually cover certain operating areas and their availability is space and time dependent.

Bonabeau (2002) very nicely explained that the main benefit of agent-based models is the capability to model emerging behavior. That is why agent-based models have seen great success in many fields, and not just in transportation. Being able to incorporate new modes of transportation easily and being flexible allows to answer many of the newly arising operational topics in transportation, like relocation of vehicles, dispatching of taxi services (be it automated or conventional), parking constraints, pooling of passengers, route constraints, intermodality, or any combination of the above.

We can also take an example of a transport policy that is already used in practice for some time: the cordon pricing strategy, usually introduced to reduce congestion in city cores. This strategy implements a charging scheme, which makes all or certain drivers pay a toll to enter a city with a car. With aggregated models the impact of this strategy on individuals with different activity patterns or income is lost and thus limits the level of insights we can gain on for instance equity analysis.

This brings us to conclude that agent-based models are necessary when designing transport policies for modern transportation systems as they are more equipped to deal with microconstraints imposed by the emerging transportation modes and research questions we want to answer.

¹ Disregarding a short period of time between the start of the urban trains era and invention of the automobile.

Examples of Microlevel Transport Simulation Models

Microsimulation represents the movements of all vehicles on the road network. These movements, however, can be modeled using different methods. Two more widely used methods are car-following and cellular automata models. Car-following model is a continuous time model of traffic that is based on differential equations. It is potentially very realistic, can model spillback effects,² but is computationally very intensive. Cellular automata similarly to car-following models uses a set of rules to determine the behavior of each vehicle on the road network. It, however, divides roads into cells with two states, empty or occupied, to define whether a vehicle is currently occupying a cell or not. It can still model spillback, and is potentially computationally less intensive than a car-following model.

Agent-based models, on the other hand, tend to use simplified or higher level representations of traffic flow, like volume-delay functions or queue models. Volume-delay functions are used to estimate travel time on the road link based on the demand. They do not model spillback, and are not really realistic, but are computationally very fast. Queue models use queue theory to represent traffic dynamics. They are one of the simplest models that can still represent spillback, but do not model any vehicle behavior inside of the road link.

This section would not be complete without mentioning some examples of the microsimulation and agent-based frameworks. The list below is definitely not exhaustive, but should build a good foundation for further reading.

Some examples of traffic microsimulation software are: VISSIM (Fellendorf and Vortisch, 2010), Aimsun (Casas et al., 2010), Simulation of Urban Mobility (SUMO, Behrisch et al., 2011), and MITSIMLab (Ben-Akiva et al., 2010). All these focus on a more detailed representation of the interaction of vehicles on the road network and use some version of a car-following model. While VISSIM and Aimsun are both commercial software, Sumo and MITSIMLab are fully open-source software developed at DLR³ and MIT⁴, respectively.

Agent-based models are mostly arising from academic work. While some of the approaches to agent-based modeling in transportation have been discontinued for different reasons, many are still around today and some examples of those are: POLARIS (Auld et al., 2016), SimMobility (Adnan et al., 2016), MATSim (Horni et al., 2016), mobiTopp (Mallig et al., 2013), and many others. All these agent-based models also include a traffic simulation part. Polaris uses a variant of a volume-delay function approach, SimMobility utilizes MITSIMLab for its traffic simulation, MATSim uses a queue model while mobiTopp relies on VISUM⁵ for an aggregated traffic assignment. As hinted previously, most agent-based models tend to approximate vehicle interactions on the network for the sake of reducing computational complexity and mostly focus on other aspects of transportation modeling, but some still depend on more detailed traffic simulation like SimMobility.

Agent-based models are often paired or fully integrated with activity-based or land-use models to provide additional features to the user. Polaris uses an activity-based model called ADAPTS (Auld and Mohammadian, 2009) to generate activities for people in the study area, while MATSim has been paired previously with an activity-based model called TASHA (Roorda et al., 2008) or a land-use model simple integrated land-use orchestrator (SILO).⁶ Integrating or pairing activity-based, agent-based, and land-use modes into a single framework can create a powerful planning tool capable of answering many transportation policy questions.

Nothing is Perfect

While microsimulation and agent-based model frameworks can be very powerful they are not perfect as they are in the end both models or approximations of a certain manifestation of the physical world. Therefore, we also have to look at some of the shortcomings of these simulation approaches.

Since the first microsimulation and agent-based modeling techniques in transportation appeared during the previous century, computing power has substantially increased. However, both approaches are still limited by high computational complexity and sometimes slow computing speeds. With new operational questions arising, new transportation modes appearing, and as vehicle-to-vehicle communication is starting to pave its way into traffic systems, these models have to deal with more computational heavy tasks than ever before. Researchers and planners usually overcome this problem by approximating certain parts of the simulation to gain in speed.

Agent-based modeling frameworks are usually very data hungry. Users of these frameworks need to invest considerable time to set up their simulation models. In the process they have to deal with many datasets, their cleaning, processing, and finally incorporating them in the model. When not available, data need to be obtained through surveys or data purchase from commercial providers. Data sources can also come from different time periods, where the time difference between collection points of datasets can be more than a few years. All this implies that substantial work is generally needed to prepare necessary inputs to an agent-based model. Additionally, as data-sources contain errors in itself, input to the models can be of various accuracy. Therefore, one must be careful

² Spillback occurs when downstream road link has reached its capacity so the congestion builds up on the upstream road link.

³ Das Deutsche Zentrum für Luft und Raumfahrt, www.dlr.de.

⁴ Massachusetts Institute of Technology, www.mit.edu.

⁵ www.ptvgroup.com.

⁶ www.silo.zone.

when interpreting the output generated by the model. Possible implications of the input accuracy on the model output and its usefulness be it qualitative or quantitative should be pointed out.

A comparison of performance of different traffic simulation or agent-based modeling platforms is also lacking. Almost no attempts were made to compare the outputs of different models and their forecasting abilities. Agent-based models are already being developed and used for some time, so it is becoming hard to neglect the fact that there are no studies aiming to verify their forecasting power. The powerful features that agent-based models have should be strengthened by a careful analysis of their ability to predict and forecast policy outcomes.

Reproducibility of studies is as well one of the stumbling blocks of simulation. It is very rarely the case that studies conducted with transport simulation software are reproducible, as the chain of steps from input generation to final output is rarely documented in a reproducible manner. This currently decreases the value of simulation outcomes. This decreases the value of the studies performed with simulation tools. To avoid this, one should make sure that input generation process is well documented, data is available for re-use and a clear documentation of steps taken to perform the study is available. This is not a simple task, but one that is necessary to increase the fidelity and reproducibility of simulation studies.

Most simulation tools are not perfect, but even with some shortcomings, these transportation simulation models provide a powerful tool to analyze possible transportation policies.

Final Remarks

Microsimulation of traffic flows and agent-based models are powerful techniques that simulate transportation phenomena observed in the physical world. Both techniques are today used by practitioners and researchers to model changes in the transportation infrastructure and forecast their impacts on the whole transportation system. This chapter provides some of the backbone information on these techniques, why they are important modeling tools, why they are necessary, and some of the shortcomings that you should be aware of. On this you should be able to build upon and extend your knowledge with either the literature provided in the next section or by becoming one of the many users of these techniques.

References

- Bonabeau, E., 2002. Agent-based modeling: methods and techniques for simulating human systems. *Proc. Natl. Acad. Sci.* 99 (Suppl 3), 7280–7287.
- Fellendorf, M., Vortisch, P., 2010. Microscopic traffic flow simulator VISSIM. In: *Fundamentals of Traffic Simulation*. Springer, New York, NY, pp. 63–93.
- Casas, J., Ferrer, J.L., Garcia, D., Perarnau, J., Torday, A., 2010. Traffic simulation with aimsun. In: *Fundamentals of Traffic Simulation*. Springer, New York, NY, pp. 173–232.
- Behrisch, M., Bieker, L., Erdmann, J., Krajewicz, D., 2011. SUMO—simulation of urban mobility: an overview. In: *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*. ThinkMind.
- Ben-Akiva, M., Koutsopoulos, H.N., Toledo, T., Yang, Q., Choudhury, C.F., Antoniou, C., Balakrishna, R., 2010. Traffic simulation with MITSIMLab. In: *Fundamentals of Traffic Simulation*. Springer, New York, NY, pp. 233–268.
- Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., Zhang, K., 2016. POLARIS: agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. *Transp. Res. Part C: Emerg. Technol.* 64, 101–116.
- Adnan, M., Pereira, F.C., Azevedo, C.M.L., Basak, K., Lovric, M., Raveau, S., Zhu, Y., Ferreira, J., Zegras, C., Ben-Akiva, M., 2016. SimMobility: a multi-scale integrated agent-based simulation platform. In: *95th Annual Meeting of the Transportation Research Board Forthcoming in Transportation Research Record*, January.
- Horni, A., Nagel, K., Axhausen, K.W. (Eds.), 2016. *The Multi-Agent Transport Simulation MATSim*, Ubiquity Press, London.
- Mallig, N., Kagerbauer, M., Vortisch, P., 2013. Mobitopp – a modular agent-based travel demand modelling framework. *Proc. Comput. Sci.* 19, 854–859.
- Auld, J., Mohammadian, A., 2009. Framework for the development of the agent-based dynamic activity planning and travel scheduling (ADAPTS) model. *Transp. Lett.* 1 (3), 245–255.
- Roorda, M.J., Miller, E.J., Habib, K.M., 2008. Validation of TASHA: a 24-h activity scheduling microsimulation model. *Transp. Res. Part A: Policy Practice* 42 (2), 360–375.

Further Reading

The following books and articles in addition to references cited within the chapter could be a useful source of further information on the topic covered in this chapter:

- Barceló, J. (Ed.), 2010. *Fundamentals of Traffic Simulation*. Springer, New York.
- de Dios Ortuzar, J., Willumsen, L.G., 2011. *Modelling Transport*. John Wiley & Sons, Chichester.
- Gilbert, N., 2008. *Agent-Based Models*. SAGE Publications, Inc., Thousand Oaks.

Choice Models in Transportation

Naveen Chandra Iraganaboina, Naveen Eluru, Department of Civil, Environmental & Construction Engineering, University of Central Florida, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	477
Choice Model Architecture and Mathematical Formulations	477
Unordered Dependent Variable Models	478
Multinomial Logit Model	478
Mixed Logit Model	479
Latent Class Multinomial Logit Model	479
Regret-based Multinomial Logit Model	480
Ordered Dependent Variable Models	480
Ordered Logit Model	480
Generalized Ordered Logit Model	481
Case Studies	481
Mode Share Models	481
Conclusion	483
References	483

Introduction

Decision makers (DM)—individuals, households, and firms among others—make several decisions as part of their life cycle. These decisions could be short time decisions such as *what mode of travel to consider to arrive at the next activity location* (for individuals) or long-term decisions such as *what house to purchase* (for a household) or *where to establish a place of residence or commercial enterprise* (for a household or a firm respectively). In understanding these decision processes that are usually discrete outcome variables, researchers across various fields including economics, psychology and transportation have developed different frameworks, referred to as Choice models. These frameworks relying on assumptions about the choice environment and the decision maker provide a mathematical representation to study choice behavior. The formulation of the choice model will vary based on the assumptions incorporated within the representation structure. These mathematical formulations are employed to analyze data from wide-ranging fields including economics, statistics, marketing, transportation, psephology, political science, and biostatistics.

In the transportation field, a large number of diverse examples of choice behavior is often encountered. At the individual level, the various choices situations encountered include: (1) travel mode choice—an individual chooses a transportation mode from a set of available transportation modes (such as car, bus, and walk), (2) route choice—a road user chooses a route from multiple alternatives for a trip (such as different Google Maps generated routes), (3) activity choice—an individual chooses an activity from a series of activities for the day (such as leisure and shopping), and (4) freight shipment size—an individual shipper determines the size of their shipment (such as under 50 pounds, 50–200 pounds and greater than 200 pounds). At the household level a sample of choices considered include (1) residential location choice—a household chooses a residential unit to buy or rent (spatial aggregated alternatives from the urban region such as traffic analysis zones or parcels), and (2) vehicle type choice—a household chooses to buy a vehicle defined by a combination of type (such as sedan minivan or coupe), make (such as Chevy, or Honda), and fuel type (electric or gasoline). At the firm level sample of choices evaluated include: (1) firm size and structure—a firm selects a size defined by the number of employees and hierarchy structure, and (2) firm location choice—a firm determines a location based on various incentives or tax implications. The examples illustrate the range and complexity of these choices.

In this article, we illustrate how choice or outcome models provide data supported analysis mechanisms for such choice contexts. The rest of the chapter is organized as follows: Section 2 summarizes mathematical frameworks commonly used for choice analysis in transportation. In Section 3, we provide a brief overview of case studies examining different choice contexts in transportation, Section 4 provides details of choice model application for travel mode share applications. Finally, Section 5 concludes the article.

Choice Model Architecture and Mathematical Formulations

Across the various choice contexts described earlier, DMs have to process the available information regarding the choice environment in arriving at their preferred choice. A DMs choice is dependent on how the DM defines the decision at hand, processes the alternatives (and their attributes) and the decision rule applied to arrive at the choice. The choice problem definition determines the potential alternatives of the choice problem. For example, if the DMs choice context is to determine the travel mode for an out of

home activity, the alternatives will possibly include all transportation modes (universal choice set of transportation modes) available in the study region. Of these alternatives, based on the DMs characteristics only a subset are feasible (feasible choice set). However, based on the attributes of the various alternatives it is possible that the decision maker will only consider a subset of alternatives from the feasible choice set (evoked choice set). The DM evaluates the alternatives in the evoked choice set based on the alternative attributes. For example, in travel mode choice context, alternative attributes include travel time and travel cost.

In choice contexts with smaller number of alternatives in the evoked choice set, DM can compare across alternatives using various decision rules to arrive at the final choice. In other cases, decision makers may consider a small subset of the alternatives to determine their choice. Earlier choice frameworks attempted a process of elimination (using satisfaction criteria [Murtaugh and Gladwin, 1980](#)) or ranking approach using lexicographic criteria ([Foerster, 1979](#)) to determine the DMs choice. However, these approaches might provide non-unique choices and/or choice outcomes that are impractical. Toward addressing these concerns, a scalar alternative utility-based approach was proposed. In this approach, each alternative is accorded a score based on a utility function (usually a linear additive function) of various alternative attributes. The alternative with the highest utility is considered to be the chosen alternative. Through the utility function, the approach allows for trade-offs (compensatory effects) between various attributes. For example, in a utility approach for travel mode choice, increase in travel cost can be compensated by improvements in travel time. While DMs are completely aware of the various attributes considered, analysts only observe the choice process partially. Hence, to allow for the influence of missing or unobserved information, utility is partitioned into two parts—observed component and unobserved component. The observed component accommodates for the impact of variables compiled in the data collection. The unobserved component takes the form of a stochastic error term to represent missing information. The alternative with the largest utility—computed as the sum of the two components—is considered the chosen alternative. Given the inherent stochasticity, the approach is referred to as the Random Utility Maximization (RUM) approach.

RUM choice models represent the most commonly employed frameworks for choice model development. The approach allows for the consideration of trade-offs across various attributes affecting the choice process. This implicit compensatory nature of the formulation allows for a poor performance on an attribute to be compensated by a positive performance on another attribute ([Chorus et al., 2008](#)). Several researchers, motivated by research in behavioral economics, have considered alternative decision rules for developing discrete choice models such as relative advantage maximization ([Leong and Hensher, 2015](#)), contextual concavity ([Kivetz et al., 2004](#)), fully-compensatory decision making ([Arentze and Timmermans, 2007](#); [Swait, 2001](#)), prospect theory (PT) ([Kahneman and Tversky, 2013](#); [Tversky and Kahneman, 1992](#)), and random regret minimization (RRM) ([Chorus, 2010](#); [Chorus et al., 2008](#)). Of these approaches RRM offers a framework that parallels RUM based models and is emerging as alternative modeling framework.

The reader would note that while all the choice examples presented in the introduction are discrete/categorical variables, some of them are potentially ordered discrete variables (for example number of employees) (Other examples of ordered discrete variables in the field of transportation include: (1) driver and passenger injury severity in traffic collisions, (2) household vehicle (automobile and bicycle) ownership, and (3) activity participation indicators (such as number of tours, number of stops, activity episode participation frequency, and activity duration)). The RUM and RRM approaches are applicable for analyzing all discrete variables. However, for modeling ordered dependent variables, another class of models referred to as ordered response frameworks also emerge as a potential model structure. In the following discussion, we present the mathematical details of the widely adopted frameworks organized by dependent variable characterization: (1) unordered dependent variable models and (2) ordered dependent variable models.

Unordered Dependent Variable Models

The unordered model frameworks are applicable for modeling all discrete choice or outcome dependent variables. In this section, we describe the RUM-based multinomial logit model and its extensions (mixed logit and latent segmentation based multinomial logit model) followed by a discussion of RRM-based multinomial logit model.

Multinomial Logit Model

Consider that a DM n has to choose an alternative i from a set of alternatives ($j = 1, 2, \dots, j$). With these notation, the utility in the RUM framework is defined as:

$$v_{ni} = \beta_i x_{ni} + \varepsilon_{ni} \quad (1)$$

where v_{ni} is the utility obtained by individual n by selecting alternative i from the choice set j . x_{ni} is the vector of alternative and DMs characteristics, β is a corresponding vector of parameters (including a constant) and ε_{ni} is the stochastic term that captures the unobserved part of the utility. The DM is expected to select an alternative that provides the highest utility. Given the inherent stochasticity involved, the choice of an alternative can only be arrived as the probability that a particular alternative (i in our case) provides the highest utility. The mathematical formula for an alternative probability of choice is affected by the assumptions about the stochastic term (ε_{ni}). Assuming ε_{ni} to be independent and identically Type I Standard Extreme Value distributed across the dataset simplifies the probability evaluation to the following multinomial logit form

$$P_{ni} = \frac{e^{v_{ni}}}{\sum_j e^{v_{nj}}} \quad (2)$$

Under different stochastic assumptions, such as ε_{ni} following a multivariate normal distribution would give rise to multinomial probit model.

The estimation of the parameters (β) is achieved by maximizing the likelihood function across the dataset, that is maximizing the probability of the chosen alternative across all records. The likelihood function for individual n is defined as

$$L_n = \sum_{\forall j} P_{nj}^{\gamma_{nj}} \quad (3)$$

where P_{nj} is the probability of the individual n choosing j and γ_{nj} is an indicator variable that takes the value 1 for chosen alternative, 0 otherwise.

The likelihood function for the entire data is computed as the product of likelihood over all the responses in the data as follows

$$L = \prod_{n=1}^N \sum_{\forall j} P_{nj}^{\gamma_{nj}} \quad (4)$$

For mathematical convenience, the natural logarithm of the likelihood function—log-likelihood—is maximized. The log-likelihood function takes the following form

$$LL = \sum_{n=1}^N \sum_{\forall j} \gamma_{nj} P_{nj} \quad (5)$$

The maximization of the function from Eq. (5) is achieved using different nonlinear optimization methods. With the increased access to computational resources, several proprietary and open-source software are available for estimating the models such as SPSS, SAS, ALOGIT, and R. The reader would note that the likelihood approach and software are similar for all models presented in the remaining discussion.

Mixed Logit Model

In the multinomial logit model, the vector of parameters estimated (β) is restricted to be the same across all DMs. The restriction is often referred to as the population homogeneity assumption. The mixed multinomial logit model accommodates for heterogeneity effects across DMs by allowing for the vector of parameters to follow a distribution across the dataset (as opposed to a fixed value). Based on the distributional assumption, the parameters that define the distribution are estimated.

In mixed logit models, the random utility formulation takes the following form (building on the formulation from Eq. (1)):

$$v_{ni} = (\beta_i + \eta_{ni})x_{ni} + \varepsilon_{ni} \quad (6)$$

where η_n is another column vector representing unobserved factors, usually considered to be independent realizations from a normal population distribution ($\eta_n \sim N(0, \sigma^2)$). Then, the probability that any DM will select alternative i for a given value of η can be expressed as:

$$P_{ni}|\eta = \frac{e^{[(\beta + \eta_{ni})x_{ni}]}}{\sum_{\forall j} e^{[(\beta + \eta_{ni})x_{nj}]}} \quad (7)$$

The unconditional probability then can be written as:

$$P_{ni} = \int_{\eta} (P_{ni}|\eta_{ni})dF(\eta_{ni}|\sigma) \quad (8)$$

where F is the multivariate cumulative normal distribution and is a vector of parameters. The integral in the probability expression cannot be evaluated using an analytical approach. Hence, the mixed logit model estimation is approximated using simulation methods for integral evaluations. The most commonly used method of simulation are Pseudo-Monte Carlo (PMC) simulation and Quasi-Monte Carlo (QMC) simulation (Bhat, 2003). The wide adoption of QMC approaches have resulted in ubiquitous adoption of mixed logit models for analysis in transportation and several other fields.

Latent Class Multinomial Logit Model

In the mixed logit model system, while the mean of the random coefficients can be allowed to vary across DMs based on observed variables, the approach usually restricts the variance and the distributional form of a random coefficient to be the same across all drivers. An alternative approach to relax the population heterogeneity assumption is the latent (or sometimes also referred to as endogenous) segmentation approach. In this approach, the DMs are allocated probabilistically to different segments, and segment-specific multinomial models are estimated. Such a segmentation approach is appealing in many respects: (1) each segment is allowed to be identified with a multivariate set of exogenous variables, (2) the probabilistic assignment of DMs to segments explicitly acknowledges the role played by unobserved factors in moderating the impact of observed exogenous variables, and (3) there is no need to specify a distributional assumption for the coefficients (Greene and Hensher, 2003; Yasmin et al., 2014).

Let us consider S homogenous segments of DMs (the optimal number of S is to be determined). We need to determine how to assign the DMs probabilistically to the segments for the segmentation model. The utility for assigning a DM n ($1, 2, \dots, N$) to segment s follows the multinomial logit structure as:

$$U_{ns} = \gamma_s z_n + \xi_{ns} \quad (9)$$

where z_n is a matrix of attributes that influences the propensity of belonging to segment s , γ_s is a vector of coefficients and ξ_{ns} is an idiosyncratic random error term assumed to be identically and independently Type 1 Extreme Value distributed across DMs n and segment s . Then the probability that DM n belongs to segment s is given as:

$$P_{ns} = \frac{\exp(\gamma_s z_n)}{\sum_s \exp(\gamma_s z_n)} \quad (10)$$

Within the latent segmentation approach, the probability of DM n choosing alternative i is given as:

$$P_{ni} = \sum_{s=1}^S (P_n(i)|s)(P_{ns}) \quad (11)$$

where $P_n(i)$ represents the multinomial logit probability for selecting alternative i within segment s following notation from Eq. (2).

Regret-based Multinomial Logit Model

The prevalent framework for developing discrete choice models is the RUM approach. RUM based approaches assume that decision makers prefer routes that provide the highest utility or satisfaction (Ben-Akiva et al., 1985; McFadden, 1974; Train, 2009). Among these approaches, the regret minimization approach offers a viable alternative approach due to its mathematical simplicity within a semi-compensatory decision framework.

The random regret associated with the choice of alternative i among j alternatives, each characterized by m attributes by the DM n is given as

$$RR_{ni} = \sum_{j \neq i} \sum_m \ln \{1 + \exp[\tau_m(x_{njm} - x_{nim})]\} \quad (12)$$

where τ_m denotes the estimable parameter associated with attribute x_m . x_{njm} and x_{nim} denote the values associated with attribute x_m for chosen alternative i and considered alternative j for DM n . In random regret models, the error term is assumed to be identically and independently Type 1 Extreme Value distributed across the dataset, which yields a closed form of probability expression similar to utility based multinomial logit model.

$$P_{ni} = \frac{e^{-RR_{ni}}}{\sum_{j \neq i} e^{-RR_{nj}}} \quad (13)$$

The model estimation follows a similar procedure of maximizing likelihood described earlier. Finally, the reader would note that similar extensions described in Sections 2.1.2 and 2.1.3 can be accommodated for RRM frameworks to relax population homogeneity assumptions.

Ordered Dependent Variable Models

The ordered response models represent the decision process under consideration using a single latent propensity. The choice probabilities are determined by partitioning the uni-dimensional propensity into as many categories as the dependent variable alternatives through a set of thresholds. The reader would note that the distributional assumptions of the unobserved component of the latent propensity determines the exact formulation of the model. The prevalent mechanism to analyze ordered discrete variables including ordered logit and generalized ordered logit models are presented in this section.

Ordered Logit Model

Let n ($n = 1, 2, \dots, N$) and j ($j = 1, 2, \dots, J$) be the indices to represent decision makers and alternatives, respectively. In the traditional OL model, alternative levels (y_n) are assumed to be associated with an underlying continuous latent variable (y_n^*). This latent variable is typically specified as the following linear function:

$$y_n^* = \alpha z_n + \varepsilon_n, \quad y_n = i, \quad \text{if } \tau_{i-1} < y_n^* < \tau_i \quad (14)$$

where y_n^* is the latent propensity for DM choosing an alternative level i , z_n is a vector of exogenous variables, α is a vector of coefficients to be estimated and ε_n is a random disturbance term assumed to be standard logistic. The latent propensity y_n^* is mapped to the observed ownership levels y_n by τ thresholds ($\tau_0 = -\infty, \tau_J = +\infty$) with the following ordering conditions:

$(-\infty < \tau_1 < \tau_2 < \dots < \tau_{j-1} < +\infty)$. Given these relationships across the different parameters, the resulting probability expression takes the following form:

$$P_{ni}(y_n = j) = \Lambda(\tau_n - \alpha x_n) - \Lambda(\tau_{j-1} - \alpha x_n) \quad (15)$$

where $\Lambda(\cdot)$ is the standard logistic cumulative distribution function (see [Greene and Hensher, 2010](#); [Train, 2009](#) for more details). A standard normal distributional assumption for ε_n would result in an ordered probit model system.

Generalized Ordered Logit Model

The generalized ordered response model relaxes the constant threshold across population restriction to provide a flexible form of the traditional OL model. The basic idea of the GOL is to represent the threshold parameters as a linear function of exogenous variables ([Eluru et al., 2008](#); [Maddala, 1983](#); [Srinivasan, 2002](#); [Terza, 1985](#)). Thus, the thresholds are expressed as:

$$\tau_{nj} = \text{function of } (Z_{nj}) \quad (16)$$

where Z_{nj} is a set of exogenous variable (including a constant) associated with j th threshold. Furthermore, to ensure the accepted ordering of observed discrete severity $(-\infty < \tau_{i,1} < \tau_{i,2} < \dots < \tau_{i,j-1} < +\infty)$. We employ the parametric form employed by ([Eluru et al., 2008](#)):

$$\tau_{nj} = \tau_{n,j-1} + \exp(\delta_{nj} Z_{nj}) \quad (17)$$

where δ_{nj} is a vector of parameters to be estimated. The remaining structure and probability expressions are similar to the OL model. For identification reasons, we need to either suppress the latent propensity of one of the δ_{nj} vectors. The model estimation follows a similar procedure of maximizing likelihood described earlier. Also, the reader would note that similar extensions described in Sections 2.1.2 and 2.1.3 can be accommodated for ordered response frameworks to relax population homogeneity assumptions.

Case Studies

To illustrate the diverse application of choice models in transportation, we present a brief summary of choice model application for different transportation empirical contexts (The reader would note that an exhaustive review of studies employing choice models is beyond the scope of the chapter). [Table 1](#) provides the summary of studies with information on study region, transportation topic of interest, data elicitation approach employed in the study, dependent variable used along with the details of the alternatives considered, choice model estimated, and the independent variables found to affect the choice process. Several observations can be made from the summary presented in [Table 1](#). First, choice models are applied for examining different transportation dimensions such as passenger and freight travel model choice, route choice, activity type choice, residential location choice, electric and autonomous vehicle purchase decisions, activity destination choice and comparison of travel times by mode. Second, the dependent variables vary considerably in terms of the number of alternatives ranging from 2 through 30. Third, the methodologies considered in these studies span the spectrum of choice models including simple binary logit models to panel mixed multinomial logit models. The studies also encompass RUM and RRM based models. Finally, a detailed summary of the independent variables considered for each study is included in the table. This summary on factors highlight how the adoption of these choice model frameworks can offer insights on a host of independent variables such as (1) DM's socio-economic and socio-demographics, (2) alternative specific characteristics such as mode characteristics, and dwelling unit characteristics, and (3) choice environment attributes such as transportation network infrastructure attributes, built environment factors, and regional and environmental factors.

Mode Share Models

In the transportation field, travel mode share models represent the most common application of choice models. Among choice models, the most commonly employed frameworks include unordered model systems such as RUM based multinomial logit and RUM based Mixed Logit. Recently RRM-based multinomial logit and mixed logit have also been applied. These model systems provide an understanding of the impact of various attributes affecting mode choice.

Traditionally, travel mode choice models are estimated using household travel survey data. In these surveys, individual travel diaries compile information on travel mode choice for every out of home travel episode. These travel episodes can be modeled individually as a trip or can be considered as a chain of trips in the form of trip chain or tour. Depending on the modal aggregation by trip or tour, the model developed will either represent a trip level mode choice or tour level mode choice. In all these models, the data compiled from the survey provides only a subset of information necessary for model estimation. To elaborate, the level of service information (such as travel time and cost by mode) is usually provided by respondents only for the chosen alternative. In the model estimation process, the analyst will need to build the other alternatives under consideration for the specific record (trip or tour) and then generate level of service measures for all these alternatives. The generation of level of service measures is likely to be challenging. For instance, the travel cost of a trip by public transit is not likely to be easy to estimate if the individual uses a monthly or annual pass. The analyst will need to estimate the number of potential instances the individual will use transit for in determining the record

Table 1 Literature review of choice models in transportation

<i>Study</i>	<i>City/Country</i>	<i>Field of study</i>	<i>Data source</i>	<i>Dependent variable</i>	<i>Modeling approach</i>	<i>Factors affecting the choice</i>
Puan et al. (2019)	Johor Bahru (Malaysia)	Mode choice	Survey	Travel Mode (car or bus)	Binary Logit	Demographic and socio-economic characteristics: age, gender, education, employment, income, household size, and vehicle ownership, Mode characteristics: travel time, travel cost, toll cost, parking cost, parking availability, number of transfers, comfort, reliability, overall quality of bus service, and bus coverage
Zhao et al. (2019)	China	Route choice	Survey	Driver's route choice response to incident information through VMS	Multinomial logit model	Demographic and socio-economic characteristics: gender, driving experience, education, occupation, income Trip characteristics: trip purpose, Road information: alternate route information (delay, congestion message by text) VMS characteristics: color, text, graph
Shabanpour et al. (2017)	Chicago, USA	Activity type	URACS Survey	Activity start times with 6 time periods of the day	Hybrid RUM-RRM model	Travel time, age, gender, income, employment, travel mode, activity location, and activity duration.
Keya et al. (2018)	USA	Freight mode choice	CFS	Shipping mode (hire truck, private truck, air, courier and others)	Hybrid Regret-Utility based MNL, Latent Class model	Mode Characteristics: Shipping cost and shipping time Freight Characteristics: Type of shipment and value of shipment Transportation Network and Demographic variables: Type of region, climate, road network density, population density, employment density, poverty level and seaports at origin and destination location.
Marois et al. (2019)	Montreal (Canada)	Residential location	National Household Survey data	Residential location choice (from 30 alternatives)	Mixed Logit	Demographic and socio-economic characteristics: Budget, previous residence, housing cost, income, and family structure. Dwelling characteristics: Type of the unit, number of rooms, state of repair and age of the house Neighborhood characteristics: Urban morphology and amenities.
Nickkar et al. (2019)	Maryland, USA	Electrical vehicle choice	Survey	Ownership of electrical vehicles by female drivers	Multinomial Logit	Demographic and socio-economic characteristics: Age, education, income, marital status, race, political affiliation, household size, number of vehicles in household.
Jiang et al. (2019)	Japan	Autonomous vehicle	Survey	Ownership of autonomous vehicles	Mixed Logit	Alternative characteristics: Additional purchase cost, Insurance reduction rate, parking cost reduction, Choice environment attributes: release time of AVs to market, penetration rate of AVs, driving experience (short and long terms), age and income of the respondent.
Hasnat et al. (2019)	Central Florida Region, USA	Destination choice	Location based social media data	Destination choice among a choice set of 30 census tracts	Panel Latent Segmentation Multinomial Logit	Origin and destination characteristics: residential, industrial, recreational, office, agricultural, land use mix, per capita income, number of schools, hospitals and civic centers.
Faghih-Imani et al. (2017)	New York, USA	Bike share	New York City bike data	Faster mode of travel (bicycle or cab)	Panel Mixed Multinomial Logit	Distance between origin and destination Trip characteristics: Time of day, trip distance, whether the trip includes crossing a bridge or not Origin and destination attributes: Station capacity, length of cycling facilities, length of streets, number of restaurants, population density, employment density, presence of transit station.

level cost. Similar challenges exist in evaluating automobile record level travel costs as maintenance and operational costs vary substantially based on the vehicle used for travel as well as the frequency of use. Thus, in transportation, analysts typically use urban regional traffic analysis zone (TAZ) level travel time and travel cost matrices for analysis. These matrices are generated using traffic assignment results from regional travel demand models. To be sure, the approach does result in some conceptual challenges. For example, changes to travel mode preferences will modify the level of service matrices and vice versa. Hence, it is prudent to consider multiple iterations between mode choice and the traffic assignment process to ensure the level of service estimates used in mode choice are realistic.

The model estimation will consider several variables including level of service attributes, DM socio-demographic and socio-economic attributes, transportation infrastructure and built environment variables. The specification of mode choice models for level of service attributes is different from other discrete choice contexts. Usually, the coefficients for independent variables are estimated specific to each alternative. However, for level of service attributes, it is useful to consider a generic parameter across all alternatives. For example, the generic parameter ensures that level of service parameters for variables such as travel time and travel cost are same across all modes. Further, the presence of generic parameters allows us to evaluate the trade-off between various mode specific level of service attributes. The trade-off between travel time and cost, defined as the value of time (VOT) or willingness to pay (WTP) is generated using mode choice estimates. For RUM-based multinomial model system, the measure is defined as

$$VOT_{RUM} = \frac{\beta_{tt}}{\beta_{tc}} \quad (18)$$

where β_{tt} and β_{tc} are estimates of travel time and travel cost respectively from the RUM based multinomial logit model.

For RRM framework, the trade-offs are dependent on levels of attributes and generated as:

$$VOT_{RRM} = \frac{\sum_{j \neq i} -\beta_{tt} \left(1 + \frac{1}{\exp[\beta_{tt}(t_j - t_i)]} \right)}{\sum_{j \neq i} -\beta_{tc} \left(1 + \frac{1}{\exp[\beta_{tc}(c_j - c_i)]} \right)} \quad (19)$$

where β_{tt} and β_{tc} are estimates of travel time and travel cost respectively from the RRM based multinomial logit model, t_i and t_j represent the travel time attributes for the chosen route i and considered route j , respectively. c_i and c_j are represent the travel time attribute for the chosen route i and considered route j , respectively. To be sure, these value of time expressions will need to be appropriately modified for mixed logit and latent segmentation models.

Conclusion

The current chapter provides an overview of model frameworks employed to model Decision Maker choice behavior in the context of transportation. Given the wide range of choice outcomes examined in transportation, multiple choice models useful for analyzing unordered and ordered discrete variables are presented. The unordered frameworks described include random utility-based multinomial logit model and random regret based multinomial logit model, and their mixed and latent segmentation variants. For ordered discrete variables, we present the traditional ordered model and its generalized variant. A brief overview of case studies covering a diverse set of choice contexts is also presented. The description of choice models was restricted due to space constraints. Several advanced modeling approaches (such as panel models, generalized extreme value extensions of traditional models) that build on the frameworks described in the chapter are developed in recent years in economics and transportation fields.

References

- Arentze, T., Timmermans, H., 2007. Parametric action decision trees: Incorporating continuous attribute variables into rule-based models of discrete choice. *Transport. Res. Part B: Methodol.* 41 (7), 772–783.
- Ben-Akiva, M.E., Lerman, S.R., Lerman, S.R., 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press.
- Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transport. Res. Part B: Methodol.* 37 (9), 837–855.
- Chorus, C.G., 2010. A new model of random regret minimization. *Eur. J. Transport Infrastructure Res.* 10 (2).
- Chorus, C.G., Arentze, T.A., Timmermans, H.J., 2008. A random regret-minimization model of travel choice. *Transport. Res. Part B: Methodol.* 42 (1), 1–18.
- Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accid. Anal. Prevention* 40 (3), 1033–1054.
- Faghih-Imani, A., Anwar, S., Miller, E.J., Eluru, N., 2017. Hail a cab or ride a bike? A travel time comparison of taxi and bicycle-sharing systems in New York City. *Transport. Res. Part A: Policy Practice* 101, 11–21.
- Foerster, J.F., 1979. Mode choice decision process models: a comparison of compensatory and non-compensatory structures. *Transport. Res. Part A: Gen.* 13 (1), 17–28.
- Greene, W.H., Hensher, D.A., 2003. A latent class model for discrete choice analysis: contrasts with mixed logit. *Transport. Res. Part B: Methodol.* 37 (8), 681–698.
- Greene, W.H., Hensher, D.A., 2010. *Modeling Ordered Choices: A Primer*. Cambridge University Press.
- Hasnat, M.M., Faghih-Imani, A., Eluru, N., Hasan, S., 2019. Destination choice modeling using location-based social media data. *J. Choice Model.* 31, 22–34.
- Jiang, Y., Zhang, J., Wang, Y., Wang, W., 2019. Capturing ownership behavior of autonomous vehicles in Japan based on a stated preference survey and a mixed logit model with repeated choices. *Int. J. Sustain. Transport.* 13 (10), 788–801.
- Kahneman, D., Tversky, A., 2013. In: *Handbook of the Fundamentals of Financial Decision Making: Part I*. World Scientific, pp. 99–127.

- Keya, N., Anowar, S., Eluru, N., 2018. Freight mode choice: a regret minimization and utility maximization based hybrid model. *Transport. Res. Rec.* 2672 (9), 107–119.
- Kivetz, R., Netzer, O., Srinivasan, V., 2004. Alternative models for capturing the compromise effect. *J. Market. Res.* 41 (3), 237–257.
- Leong, W., Hensher, D.A., 2015. Contrasts of relative advantage maximisation with random utility maximisation and regret minimisation. *J. Transport Econ. Policy (JTEP)* 49 (1), 167–186.
- Maddala, G., 1983. Methods of estimation for models of markets with bounded price variation. *Int. Econ. Rev.* 361–378.
- Marois, G., Lord, S., Morency, C., 2019. A mixed logit model analysis of residential choices of the young-elderly in the Montreal metropolitan area. *J. Housing Econ.* 44, 141–149.
- McFadden, D., 1974. The measurement of urban travel demand. *J. Public Econ.* 3 (4), 303–328.
- Murtaugh, M., Gladwin, H., 1980. A hierarchical decision-process model for forecasting automobile type-choice. *Transport. Res. Part A: Gen.* 14 (5–6), 337–347.
- Nickkar, A., Shin, H.-S., Farkas, A., 2019. Analysis of ownership and travel behavior of women who drive electric vehicles: the case of Maryland. *arXiv preprint arXiv*, 1911.03943..
- Puan, O., Hassan, Y., Mashros, N., Idham, M., Hassan, N., Warid, M., Hainin, M., 2019. In: *Transportation Mode Choice Binary Logit Model: A Case Study for Johor Bahru City*. IOP Publishing, p. 012066.
- Shabanpour, R., Golshani, N., Auld, J., Mohammadian, A., 2017. Dynamics of activity time-of-day choice. *Transport. Res. Rec.* 2665 (1), 51–59.
- Srinivasan, K.K., 2002. Injury severity analysis with variable and correlated thresholds: ordered mixed logit formulation. *Transport. Res. Rec.* 1784 (1), 132–141.
- Swait, J., 2001. A non-compensatory choice model incorporating attribute cutoffs. *Transport. Res. Part B: Methodol.* 35 (10), 903–928.
- Terza, J.V., 1985. A Tobit-type estimator for the censored Poisson regression model. *Econ. Lett.* 18 (4), 361–365.
- Train, K.E., 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press.
- Tversky, A., Kahneman, D., 1992. Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertainty* 5 (4), 297–323.
- Yasmin, S., Eluru, N., Bhat, C.R., Tay, R., 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. *Analyt. Methods Accid. Res.* 1, 23–38.
- Zhao, W., Quddus, M., Huang, H., Lee, J., Ma, Z., 2019. Analyzing drivers' preferences and choices for the content and format of variable message signs (VMS). *Transport. Res. part C: Emerging Technol.* 100, 1–14.

Multi-Criteria Decision Analysis

Zhanmin Zhang, Srijith Balakrishnan, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	485
Stages in Multi-Criteria Decision Analysis	486
Delineate Scope of the Decision-Making Problem (Project)	486
Identify the Project Alternatives	486
Define Criteria for Evaluating Alternatives	486
Assess the Performance of Alternatives	486
Assign Weights for Different Criteria	487
Calculate the Overall Score of Alternatives	487
Analyze the Results and Make Recommendations	487
Sensitivity Analysis and Final Selection of Alternative	488
Classifications of MCDA Techniques	488
Commonly Used MCDA Techniques for Transportation-Related Decision-Making	488
Weighted Global Criterion Method	488
Goal Programming	489
Weighted Sum Model	489
Analytical Hierarchical Process (AHP)	489
Analytical Network Process (ANP)	490
Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE) Methods	490
Elimination of Choice Expressing Reality (ELECTRE)	491
Technique for Order Preference by Similarity (TOPSIS)	491
Summary	491
See Also	492
References	492
Further Reading	492

Introduction

Transportation infrastructure projects (hereon, projects) are primarily meant for mitigating existing transportation problems such as traffic congestion, or to improve accessibility, mobility, or connectivity in terms of routes or modes of transport. However, there are several concerns, such as safety, equity, environmental sustainability, and availability of funds that determine the selection of a particular project for implementation. In addition, there are other considerations, such as short- and long-term economic growth and job creation resulting from the completion of the project. Not only that some of these objectives conflict or compete with each other, but the value of project alternatives perceived by different stakeholders based on the stated objectives could also be vastly different. Exclusion of important objectives could lead to wrong choice of alternatives, resulting in unanticipated project overruns due to political and legal hurdles. Foreseeing these issues, legislation often mandates that several project alternatives must be explored before a project's implementation and a holistic analysis must be carried out with due consideration to the concerns of stakeholders and the public in general. The challenge for the agencies is to identify and apply appropriate methods to select the alternatives that deviate the least from the stated project objectives and concerns of the stakeholders. Multiple criteria decision analysis (abbreviated as MCDA; also known as multiple criteria decision-making) methods are increasingly becoming popular tools for transportation-related project selection due to their ability in handling competing and conflicting interests of various stakeholders.

MCDA techniques are a set of decision-support tools available to transportation agencies to evaluate, assess, and rank project alternatives based on a multitude of relevant stakeholder concerns (Kiker et al., 2005). The concerns, in a formal decision-making problem setting, are otherwise called "criteria." In general, MCDA methods break down a complex decision-making problem (project selection) into several subproblems, analyze each of the available alternatives based on the subproblem, and then integrate the findings to choose the best alternative among the available options for implementation. MCDA also provides a systematic and transparent procedure for evaluating alternatives with due consideration to all concerns and perspectives, however, insignificant in scope they may be.

The key advantage of MCDA methods is that they are capable of incorporating subjective information. In this way, MCDA methods can consider both tangible and intangible benefits and costs of project alternatives leading to more meaningful decisions

supported by sound rationale. The two aspects that make different MCDA methods distinct are (1) the methods by which weights for preferences of stakeholders are assigned; and (2) the way in which uncertainties in subjective preferences are handled. The main focus of this article is to outline the general framework of MCDA and introduce some of the widely adopted MCDA techniques used for addressing transportation-related decision-making problems.

Stages in Multi-Criteria Decision Analysis

Any MCDA involving multiple stakeholders can be divided into eight major stages ([Department of Communities and Local Government, 2009](#)). In the rest of the section, each of these stages is discussed in detail.

Delineate Scope of the Decision-Making Problem (Project)

This stage involves establishing the basic setting for the MCDA problem to be solved. Specifically, four aspects of the problem are defined: (1) the aims of the project; (2) participants (decision-makers and key players concerned with the project); (3) the socio-technical apparatus of the MCDA; and (4) the context of the MCDA. The aims of the MCDA are broad intentions that the agency (which is responsible for the project) desires to achieve by the implementation of the projects. The next step involves identifying the decision-makers and key players (MCDA participants) who can contribute to the assessment of various perspectives and concerns related to the project during the MCDA process. The MCDA participants must be able to bring in the concerns of all stakeholders who could be directly or indirectly affected as well as the opinions of the subject matter experts who have adequate knowledge in the domains pertinent to the project. In the third step, the methods and time frame for collecting relevant information from the participants are decided. Facilitated workshops and surveys are widely adopted methods for collecting stakeholder inputs. In addition, the various MCDA techniques appropriate for the current decision-making problem must be identified and the corresponding advantages and limitations must be ascertained. In the last step, the resources for the MCDA are allocated appropriately. External factors such as political, economic, social, and technological systems that could influence the project may also be considered so that further fine-tuning of the setup could be done before the commencement of the MCDA.

Identify the Project Alternatives

For a project to be implemented, several alternatives may exist that can achieve the stated aims and objectives; however, resource constraints make implementation of only one or a few of the alternatives viable. While some projects may have predefined alternatives, in some others, the MCDA must develop alternatives in consultation with the key players and stakeholders. Along with new alternatives, “do-nothing” must also be considered; in many cases, continuing the status quo may have more value over all other alternatives. The development of alternatives may be done in several steps, as appraising the short-term and long-term implications of alternatives could motivate the stakeholders to come up with better alternatives. Flexibility in modifying the initial pool of alternatives is therefore key in a successful MCDA endeavor.

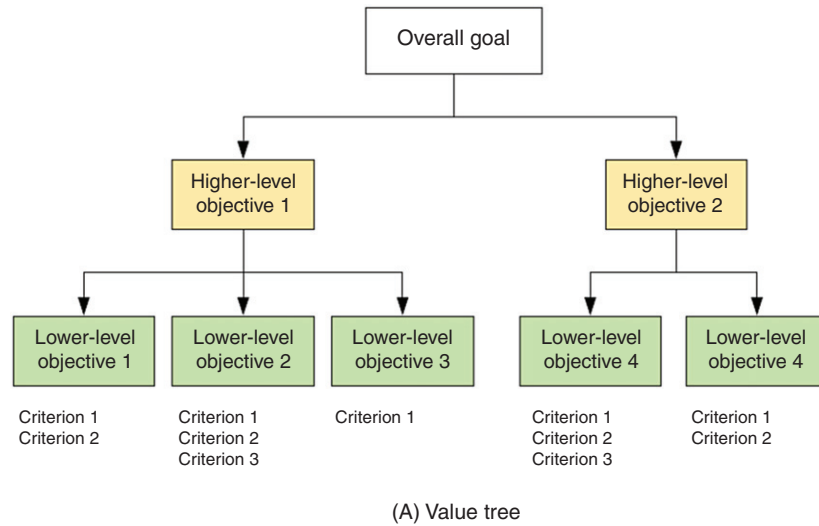
Define Criteria for Evaluating Alternatives

A criterion, in MCDA context, is defined as “an individual measurable indicator of a key value dimension” ([Department of Communities and Local Government, 2009](#)) or “a particular perspective according to which alternative technologies may be compared” ([Belton and Stewart, 2002](#)). The broader objectives of the project are achieved by ensuring desired level of the selected criteria. The criteria evaluate the value and the consequences (both positive and negative) of the project alternatives, leading to the selection of the best alternative. Criteria can be either quantitative or qualitative in nature based on their measurability. The criteria must reflect the undertaking agency’s stated values and goals, the specific objectives of the project, and the concerns of the participants and the general public. Once the criteria are identified, a systematic representation of the objectives and criteria is developed depending on the complexity of the problem ([Fig. 1](#)).

In complex MCDA problems involving multiple levels of objectives, a value tree is used. A value tree refers to organizing criteria by clustering them under respective higher- and lower-level objectives of the project. This representation facilitates the evaluation of alternatives in terms of various objectives and sub-objectives. In a value tree, the top-level objectives are the costs and benefits of the project, under which other lower-level objectives are clustered. In the case of simple MCDA problems, a simple performance matrix is appropriate for representing alternative-criterion relations.

Assess the Performance of Alternatives

The next step in MCDA is to assess the various alternatives based on the final set of criteria. MCDA techniques use preference scores to normalize various qualitative and quantitative performance measures for comparison across criteria. The preference scores are articulated during workshops and surveys. The alternatives are assigned with a score based on a relative scale in which a low score refers to a low level of preference and a high score indicates high level of preference of a particular alternative. Preference scores indicate the relative strength of preference of one alternative to others for each of the criteria. The preference scores obtained must be



Alternatives	Criterion 1 (Objective 1)	Criterion 2 (Objective 2)	Criterion 3 (Objective 3)
Alternative 1			
Alternative 2			
Alternative 3			
Alternative 4			

(B) Performance matrix

Figure 1 Formats of value tree and performance matrix.

checked for consistency as part of input validation. Ensuring consistency of the scores within and across criteria is crucial and requires several iterations of preference score assessments until the scores are agreed upon by all the key players and stakeholders.

Assign Weights for Different Criteria

While preference scores are effective in comparing alternatives based on a particular criterion, it requires further processing for cross-criterion comparisons. After all, the goal of MCDA is to aggregate the preferences based on various criteria to assign overall scores for alternatives. Weights are assigned to each criterion to denote their (criteria) relative importance. Consequently, the “weighted” preference scores can be combined so that the final score would reflect not only the relative preference of alternatives based on the criteria, but also their importance in achieving the project objectives.

Calculate the Overall Score of Alternatives

Once the weights are derived, the overall weighted preference scores at each level in the value tree are calculated as follows:

$$S_i = \sum_{j=1}^n w_j s_{ij}$$

where S_i is the overall score corresponding to alternative i , w_i is the derived weight for criteria j , and s_{ij} is the preference score corresponding to criteria j and alternative i . One key issue often encountered in this stage is the presence of dependence among some of the criteria. The preference of an alternative may be dependent on the preference of another criteria, leading to unreasonable scores for alternatives. However, there are several alternative mathematical techniques available to deal with such situations.

Analyze the Results and Make Recommendations

The top-level weighted scores of alternatives can be used to rank them based on the order of suitability for the project objectives. In addition, cross-alternative comparisons can also be carried out based on the scores at each level of the hierarchical clustering of objectives. In the case of unexpected results from the ranking exercise that are contrary to the intuition of the participants, further

Table 1 Major classifications of multi-criteria decision-making techniques

<i>Classification basis</i>	<i>Categories</i>	<i>Unique characteristics</i>	<i>Example techniques</i>
Selection of alternatives (Triantaphyllou, 2013)	Multi-objective decision-making (MODM)	Alternatives are not predefined, and criteria are expressed in continuous form	Weighted global criterion method, goal programming
	Multi-attribute decision-making (MADM)	Alternatives are predefined and discrete in nature	Analytical hierarchical process, technique for order preference by similarity
Range of application (Belton and Stewart, 2002)	Value measurement methods	Numerical measures for representing preference of one alternative over another	Analytical hierarchical process, weighted sum method
	Goal, aspiration, or reference-level methods	Acceptable level of a criterion is predefined	Goal programming, technique for order preference by similarity
	Outranking methods	Pairwise comparison of alternatives and aggregation of results across criteria	Elimination of choice expressing reality, preference ranking, organization method for enrichment evaluation methods
Level of aggregation of alternatives (Mardani et al., 2015)	Complete aggregation	Comparison of all alternatives in a single step	Technique for order preference by similarity
	Partial aggregation	Pairwise comparison of alternatives	Analytical network process, analytical hierarchical process

formal discussions must be organized to review the results and their implications following implementation. Stakeholder discussions on the MCDA results may open up new possibilities in terms of alternatives or may be helpful in identifying the pitfalls in the decision-making process.

Sensitivity Analysis and Final Selection of Alternative

Sensitivity analysis provides insights into the implications resulting from any uncertainty in the information articulated from the participants of the MCDA. The final scores may be highly sensitive to some of the subjective inputs and MCDA method used. Sensitivity analysis is capable of pinpointing the root causes of uncertainties in implications of project alternatives, may be due to poorly defined criteria weights or uncertain model inputs (Saltelli et al., 1999).

Classifications of MCDA Techniques

Several classifications exist for categorizing MCDA methods. The most important among these are with respect to selection of alternatives range of application and level of aggregation of alternatives, as presented in Table 1.

Commonly Used MCDA Techniques for Transportation-Related Decision-Making

The following are some of the major MCDA techniques that are used for transportation-related decision-making involving multiple stakeholders and project objectives.

Weighted Global Criterion Method

The weighted global criterion method is a multi-objective decision-making framework in which the utility functions corresponding to the various objectives of the project are combined into a single global function. The alternatives are generally not predefined and the criteria are represented using a continuous hypercube. The utility functions are constructed such that the global function minimizes the distance between the solution point and the “utopia point.” The utopia point refers to the most-desirable (ideal) solution that may or may not lie within the criterion space. While several weighted utility functions exist, the following are the most-widely adopted ones.

$$U = \left\{ \sum_{i=1}^k w_i^p [F_i(\mathbf{x}) - F_i^0] p \right\}^{\frac{1}{p}}$$

$$U = \left\{ \sum_{i=1}^k w_i^p [F_i(x) - F_i^0] p \right\}^{\frac{1}{p}}$$

where $w_i \in \mathbf{w}$ is the weights corresponding to each of the project objective i such that $\sum_{i=1}^k w_i = 1$ and $w_i > 0 \forall i$, $\mathbf{x}$ is the set of criteria, $F_i(\cdot)$ is the function that defines the objective i , F_i^0 is the ideal value of the objective i , and $p = 1, 2, 3, \dots$

The weights are directly proportional to the importance of each of the project objectives. The method is commonly employed in resource allocation problems.

Goal Programming

Goal programming, first introduced by Charnes et al. (1955), is a commonly used multi-objective decision-making technique to find a project alternative which is "acceptable" for all the key players and stakeholders. This is a type of global criterion method in which the absolute deviation from ideal solution points corresponding to each of the objectives are minimized as follows:

$$\begin{aligned} \min \quad & \sum_{i=1}^k (d_i^+ + d_i^-) \\ \text{s.t.} \quad & F_i(\mathbf{x}) - d_i^+ + d_i^- = b_i, \quad i = 1, 2, 3, \dots, k \\ & d_i^+, d_i^- \geq 0, \quad i = 1, 2, 3, \dots, k \\ & d_i^+ d_i^- = 0, \quad i = 1, 2, 3, \dots, k \end{aligned}$$

where b_i is the desired goal corresponding to project objective i , $\mathbf{x}$ is the set of alternatives, $F_i(\mathbf{x})$ is the utility function corresponding to objective i , and d_i^+ and d_i^- represent the overachievement and underachievement of j . There are three types of goal programming formulations based on how the bounds of goals are defined, namely, lower one-sided goal, upper one-sided goal, and two-sided goal.

The extension of goal programming with differential weights assigned to goal functions based on their relative importance is called Archimedean goal programming. While goal programming can be effective in obtaining a solution acceptable to the key players and stakeholders, there is no assurance that the solution is Pareto optimal. Another major limitation is that computation becomes complex if there are additional variables and nonlinear equality constraints in the optimization problem.

Weighted Sum Model

Weighted sum model is the simplest multi-criteria decision-making method introduced by Fishburn (1967) in which the utility functions of various project objectives are weighted according to their importance and are summed up as follows:

$$A_i^{\text{WSM_score}} = \sum_j^n w_j a_{ij}, i = 1, 2, \dots, m, j = 1, 2, \dots, n$$

where $A_i^{\text{WSM_score}}$ is the weighted score for alternative i , w_j is the weight corresponding to objective j , and a_{ij} is the performance value of i corresponding to j . The performance values must be depicted in the same unit across criteria. The alternative that has the maximum weighted score is the most preferred one. In the case of multi-objective decision-making problems, the solution point can be obtained by maximizing the weighted sum of utility functions.

$$U = \sum_j^n w_j F_j(\mathbf{x}), j = 1, 2, \dots, n$$

The weights in the weighted sum model formulation are nonarbitrary and must be derived using methods such as ranking methods, rating methods, paired comparison methods, etc. In addition, the criteria selected must also be independent of each other to ensure that the scores are not amplified by the presence of correlated criteria.

Analytical Hierarchical Process (AHP)

Analytical hierarchical process (AHP), introduced by Saaty (1990), has been an effective tool for multi-criteria decision-making problems involving ranking, selection, evaluation, optimization, and prediction. AHP involves four major steps. In the first step, the decision-making problem is structured in the form of a value tree, where the top level denotes the goal of the project, followed by the objectives and sub-objectives in intermediate levels and the corresponding criteria take up the bottom layer (Fig. 1A). In the second step, pairwise comparison of criteria is carried out by taking one pair at a time and obtaining the ratio of importance of one criterion over another from the participants. The ratios are later arranged in a comparison matrix. Usually, a scale ranging from 1 to 9 is used, where the score 1 denotes a situation where the criteria are equally important and a score of 9 denotes that the first in the pair is extremely important over the second. The intermediate values refer to intermediate preference states. Reciprocal of the scores is used if the second element is more preferred over the first. The consistency of the articulated comparison matrix is checked by finding the consistency ratio (CR), which is a function of the consistency index (CI) of the matrix (the principal eigenvalue of the matrix) and the random consistency index (RI). If the consistency ratio is not near to the ideal value of zero (less than 10%), the articulation may be repeated until desired level of consistency is achieved. Then, the weights of the criteria are obtained by calculating the normalized principal eigenvector corresponding to the highest eigenvalue of the comparison matrix.

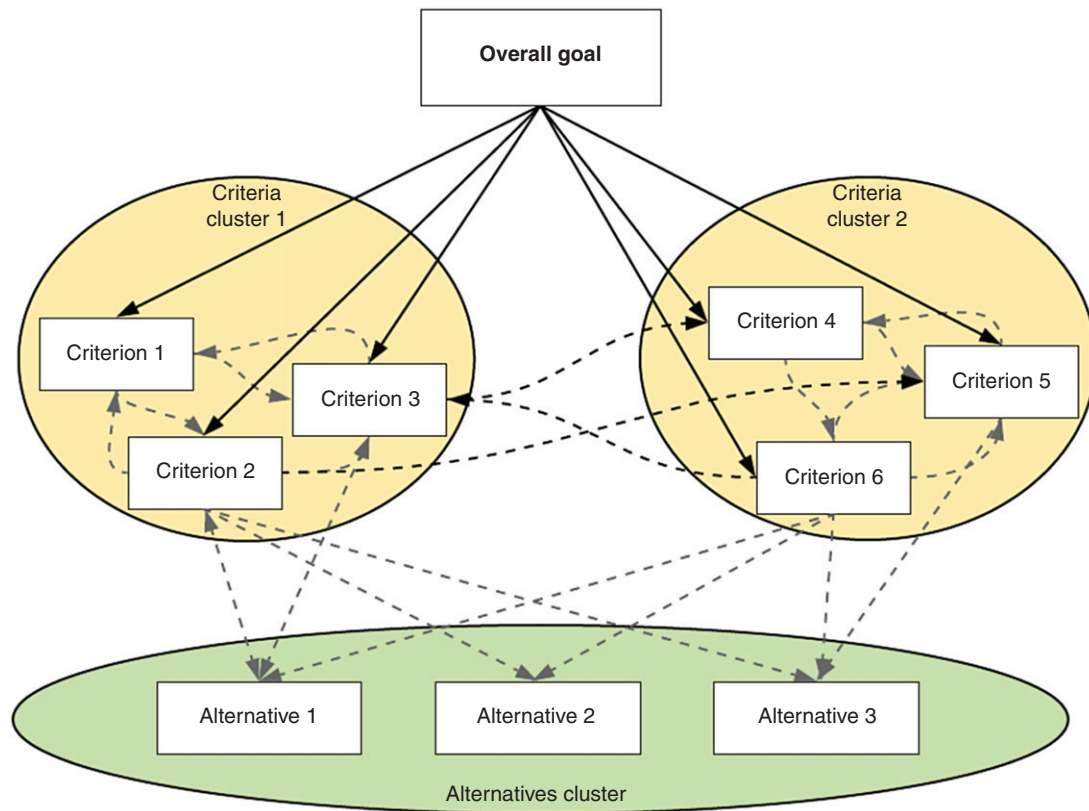


Figure 2 Structure of an analytical network process problem.

In the third step, a similar procedure as in the previous step is implemented to obtain the priorities based on each of the criteria. The comparison matrix denotes the preference of each alternative over other corresponding to a specific criterion. If several levels of objectives and sub-objectives are to be dealt with, a bottom-up approach must be followed for pairwise comparison of alternatives. The normalized principal eigenvector corresponding to each of the comparison matrices in this step reflect the preference of various alternatives with regard to a specific criteria. In the final step, the priority vectors of both criteria and alternatives are appropriately combined to obtain the composite scores of each alternative indicating their strength of preference.

Analytical Network Process (ANP)

Analytical network process (ANP) is a generalized version of AHP where the decision-making problem is structured as a network (Fig. 2) and not as a hierarchy of goal, objectives, and criteria (Saaty and Vargas, 2013). This method is appropriate when dependencies exist among decision criteria. The dependencies and feedback loops represented by the network can compute the criteria weights more precisely, eliminating any overestimation of influence of a criterion. ANP involves four major steps.

In the first step, the network consisting of the goal cluster (node G), criteria nodes (nodes $C_1, C_2, \dots, C_n$) and alternatives cluster (nodes $A_1, A_2, \dots, A_m$) along with the dependencies and feedback links. The criteria are clustered according to the objectives. In the next step, the weights of nodes are obtained by implementing second and third steps described in AHP. Four types of pairwise comparisons are done in this step to obtain the unweighted super matrix of the network comparison of criteria with goal, comparison among criteria, comparison of alternatives with respect to each goal, and comparison of criteria in each cluster with respect to each alternative. In the third step, the unweighted super matrix is converted into weighted super matrix. For this purpose, three types of cluster-level pairwise comparisons are carried out—comparison of criteria clusters with respect to goal, comparison of all criteria and alternatives clusters with respect to each of the criteria clusters, and comparison of criteria clusters with respect to each of the alternative clusters. In the final step, the weighted super matrix is multiplied to itself until all the columns become equal. The column values in the limiting matrix represent the weights of the criteria and the alternatives.

Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE) Methods

PROMETHEE methods also use pairwise comparisons between alternatives to sort them in the order of preference based on a finite number of criteria (Brans and Vincke, 1985). The methods are implemented in three steps. In the first step, such as AHP and ANP, alternatives are compared in pairs for each of the criterion. However, a preference scale (ranging from 0 representing no preference or

indifference to 1 representing strict preference) is used for representing the preference of one alternative over the other in each pair. A generalized criterion function is used to convert the preference of the decision-maker into a preference score. In the next step, for each pair of alternative a and b , a multi-criteria preference index $\Pi(a, b)$ is computed as the weighted average of preference scores corresponding to the criteria to represent the relative overall preference of alternatives in each pair. In the final step, the leaving flow $\phi^+(a)$ and the entering flow $\phi^-(a)$ for each of the m alternatives are calculated as follows.

$$\phi^+(a) = \frac{1}{m-1} \sum_{i=1, i \neq a}^m \Pi(a, i)$$

$$\phi^-(a) = \frac{1}{m-1} \sum_{i=1, i \neq a}^m \Pi(i, a)$$

The leaving flow value indicates the extent of preference of a over other alternatives, whereas the entering flow represents the extent of preference of all other alternatives over a . The various methods within the family of PROMETHEE methods differ in how alternatives are sorted based on the leaving flow and entering flow values. For example, in PROMETHEE I, a is preferred over β if $\phi^+(a) \geq \phi^+(b)$ and $\phi^-(a) < \phi^-(b)$. In the case of PROMETHEE II, the net flow ($\phi^+(a) - \phi^-(a)$) is used to develop a complete ranking of alternatives in one single step.

Elimination of Choice Expressing Reality (ELECTRE)

ELECTRE methods use the concepts of concordance and discordance (non-concordance) to develop outranking relations among the project alternatives (Figueira et al., 2005). Three levels of outranking relations are used—strictly preferred, indifferent, and incomparable. The ELECTRE methods essentially have two steps, namely aggregation and exploitation. In aggregation phase, overall preference of one alternative over another is computed using pairwise outranking assertions. For an alternative a to be strictly preferred over b , both concordance and discordance conditions must be satisfied. The concordance condition for the assertion that a is strictly preferred over b (aSb) is satisfied if a majority of such outranking relations between a and b follows the same assertion. The discordance condition is satisfied if none of the criteria in the minority subset shall strictly oppose the assertion aSb . The outranking relations can be crispy, fuzzy, or embedded in nature. In the exploitation phase, a small set of alternatives is selected based on the outranking relations among alternatives. Several methods, such as graph kernel method, are used for this purpose.

ELECTRE methods are broadly classified into choice problematic, ranking problematic, and sorting problematic, based on the nature of the decision-making problem.

Technique for Order Preference by Similarity (TOPSIS)

TOPSIS is a multi-attribute decision-making method that checks for similarity of each alternative to the ideal solution to rank the alternatives in the order of preference (Tzeng et al., 2011). The best alternative is the one which is the nearest to the positive ideal solution and the farthest from the negative ideal solution. The positive ideal solution is the one that maximizes the benefit criteria and minimizes the cost criteria. On the other hand, the negative ideal solution is the one that minimizes the benefit criteria and maximizes the cost criteria. TOPSIS has essentially three steps for implementation. In the first step, the normalized weighted preference matrix is obtained by multiplying the elements in the normalized preference matrix with the corresponding criteria weights. In the second step, the positive and negative ideal solutions are computed. As the names suggest, the positive ideal solution is calculated using the best possible values for the criteria and the negative ideal solution is obtained using the worst possible values for the criteria. In the last step, the distance from each of the project alternatives to the positive and negative ideal solutions is computed using Euclidean distance. The relative closeness index for each alternative, computed as the ratio of the distance to the negative ideal solution to the sum of positive and negative ideal solutions, is used to identify the best alternative.

Summary

In this article, a brief introduction to the concept of MCDA is provided and some of the popular techniques used for MCDA are discussed. MCDA methods are appropriate for transportation-related project prioritization, when consideration of opinions, concerns, and demands of a diverse group of stakeholders become integral to the decision-making process. While only a very few MCDA techniques are introduced, there are several other techniques which may become equally handy while dealing with MCDA problems. In many past studies, application of hybrid MCDA methods (combination of two or more methods) has also been investigated. Nevertheless, the aim of any MCDA technique remains the same—choose a desired number of project alternatives from a large pool which are least deviating from the expectations of the stakeholders and key players.

While MCDA has numerous advantages over conventional decision-making systems, there are several major challenges as well. MCDA methods could produce misleading results in some cases, as they are vulnerable to personal biases in preferences. In addition, the selection of MCDA method is very crucial and depends on the problem under consideration, as different methods have different

approaches for deriving criteria weights and alternative selection. Some of the aspects that must be considered while choosing MCDA techniques are enlisted as follows (Bonissone, 2008).

1. Uncertainty and imprecision in the subjective preferences of the MCDA participants and the ability of the MCDA methods to handle them.
2. High dimensionality of criteria space (more number of criteria makes preference articulation complex).
3. Ability to identify inconsistencies in stakeholder preferences and the flexibility of MCDA methods update them.
4. Ease of development, deployment, and management of adaptive and distributed decision support systems using the MCDA methods.

In addition, transportation projects have huge impact on the society and regional economy and therefore adequate attention must also be given to the uncertainties in implications of the alternatives selected by the MCDA procedures. Because of these reasons, the decision-makers must be aware of the merits and demerits of the MCDA approaches available before undertaking the analysis.

See Also

Cost Over-Runs of Transport Infrastructure Projects; Cost-Benefit Analysis and Other Assessment Techniques: Contrasts and Synergies; Information Sharing and Business Analytics in Global Supply Chains

References

- Belton, V., Stewart, T., 2002. *Multiple Criteria Decision Analysis: An Integrated Approach*. Springer Nature Book Archives Millennium, Springer US, Boston, MA.
- Bonissone, P.P., 2008. Research issues in multi criteria decision making (MCDM): the impact of uncertainty in solution evaluation. In: Magdalena, L., Ojeda-Aciego, M., Verdegay, J.L. (Eds.), *12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Málaga, Spain, pp. 1409–1416. Available from: <https://www.semanticscholar.org/paper/Research-Issues-in-Multi-Criteria-Decision-Making-Bonissone/66ec73253dbcb92c11fb1cc7778f967e585d218f>.
- Brans, J.P., Vincke, P., 1985. A preference ranking organisation method. *Manag. Sci.* 31 (6), 647–656, doi:10.1287/mnsc.31.6.647.
- Charnes, A., Cooper, W.W., Ferguson, R.O., 1955. Optimal estimation of executive compensation by linear programming. *Manag. Sci.* 1 (2), 138–151, doi:10.1287/mnsc.1.2.138.
- Department of Communities and Local Government, 2009. *Multi-Criteria Analysis: A Manual*. Communities and Local Government Publications, London, UK. Available from: http://eprints.lse.ac.uk/12761/1/Multi-criteria%7B%5C_%7DAnalysis.pdf.
- Figueira, J., Mousseau, V., Roy, B., 2005. Electre methods. In: *Multiple Criteria Decision Analysis: State of the Art Surveys*. Springer-Verlag, New York, pp. 133–153, doi:10.1007/0-387-23081-5_4.
- Fishburn, P.C., 1967. Additive utilities with incomplete product sets: application to priorities and assignments. *Oper. Res.* 15 (3), 537–542, doi:10.1287/opre.15.3.537.
- Kiker, G.A., Bridges, T.S., Varghese, A., Seager, T.P., Linkov, I., 2005. Application of multicriteria decision analysis in environmental decision making. *Integr. Environ. Assess. Manag.* 1 (2), 95–108, doi:10.1897/IEAM_2004a-015.1.
- Mardani, A., Jusoh, A., Zavadskas, E.K., 2015. Fuzzy multiple criteria decision-making techniques and applications: two decades review from 1994 to 2014. *Expert Syst. Appl.* 42 (8), 4126–4148, doi:10.1016/j.eswa.2015.01.003.
- Saaty, T.L., 1990. How to make a decision: the analytic hierarchy process. *Eur. J. Oper. Res.* 48 (1), 9–26, doi:10.1016/0377-2217(90)90057-I.
- Saaty, T.L., Vargas, L.G., 2013. *Decision making with the analytic network process*. In: *International Series in Operations Research & Management Science*. Springer US, Boston, MA, doi:10.1007/978-1-4614-7279-7.
- Saltelli, A., Tarantola, S., Chan, K., 1999. A role for sensitivity analysis in presenting the results from MCDA studies to decision makers. *J. Multi Criteria Decision Anal.* 8 (3), 139–145, doi:10.1002/(SICI)1099-1360(199905)8:3<139::AID-MCDA239>3.0.CO;2C.
- Triantaphyllou, E., 2013. *Multi-Criteria Decision Making Methods: A Comparative Study*. Springer, Boston, MA.
- Tzeng, G.-H., Huang, J.-J., Huang, J.-J., 2011. *Multiple Attribute Decision Making*. Chapman and Hall/CRC, Boca Raton, FL, doi:10.1201/b11032.

Further Reading

- Beinat, E., Nijkamp, P. (Eds.), 1998. *Multicriteria analysis for land-use management*. Environment & Management. Springer Netherlands, Dordrecht, doi:10.1007/978-94-015-9058-7.
- Durbach, I.N., Stewart, T.J., 2012. Modeling uncertainty in multi-criteria decision analysis. *Eur. J. Oper. Res.* 223 (1), 1–14, doi:10.1016/J.EJOR.2012.04.038.
- Ishizaka, A., Nemery, P., 2013. *Multi-Criteria Decision Analysis: Methods and Software*. Wiley, Chichester, West Sussex, UK.
- Jankowski, P., 1995. Integrating geographical information systems and multiple criteria decision-making methods. *Int. J. Geogr. Inf. Syst.* 9 (3), 251–273, doi:10.1080/02693799508902036.
- Mardani, A., Zavadskas, E.K., Khalifah, Z., Jusoh, A., Nor, K.M., 2016. Multiple criteria decision-making techniques in transportation systems: a systematic review of the state of the art literature. *Transport* 31 (3), 359–385, doi:10.3846/16484142.2015.1121517.
- Marler, R., Arora, J., 2004. Survey of multi-objective optimization methods for engineering. *Struct. Multidiscipl. Optim.* 26 (6), 369–395, doi:10.1007/s00158-003-0368-6.
- Massam, B.H., 1988. Multi-criteria decision making (MCDM) techniques in planning. *Prog. Plan.* 30, 1–84, doi:10.1016/0305-9006(88)90012-8.
- Velasquez, M., Hester, P.T., 2013. An analysis of multi-criteria decision making methods. *Int. J. Oper. Res.* 10 (2), 56–66.
- Zeleny, M. (Ed.), 1976. *Multiple criteria decision making Kyoto 1975*. In: *Lecture Notes in Economics and Mathematical Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-45486-8.

The National Household Travel Survey Data Series (NPTS/NHTS)

Nancy McGuckin, Travel Behavior Analyst, Consultant, South Pasadena, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Overview	493
Uses of the Data	493
Biography	495
References	495

Overview

Title 23, United States Code, Section 502 authorizes the US Department of Transportation (USDOT) to carry out transportation research to measure the performance of the surface transportation systems in the United States, including the efficiency, energy use, air quality, congestion, and safety of the highway and intermodal transportation systems. The USDOT has overall responsibility to obtain current information on national patterns of travel, establish a database to better understand travel behavior, evaluate the use of transportation facilities, and gauge the impact of the USDOT's policies and programs.

The Federal Highway Administration (FHWA) Office of Policy Information sponsors the National Household Travel Survey (NHTS, previously called the Nationwide Personal Travel Survey) to collect information on the travel of US residents. The survey has been conducted periodically since 1969 and has evolved in methods and uses across the decades. The 2017 NHTS, the most recent data released in April 2018, contains information for 264,000 completed responses from people aged 5 and older in 130,000 households (Fig. 1). Every household member recorded their travel for a single, assigned travel day (adults proxied for children 13 and under). This allows the analysis of the interactions between household members and the linking of travel to the individual and household characteristics, such as income, number of vehicles, and the presence of children. The findings of the survey are documented in the Summary of Travel Trends, which is available at <https://nhts.ornl.gov/>.

Uses of the Data

Data from the NHTS are widely used to support research needs within the USDOT, and state and local agencies, in addition to responding to queries from Congress, the research community, and the media on important issues for transportation policy and planning. A compendium of uses of the most recent NHTS data (compiled by Oakridge National Laboratories for the 2018 calendar year) is shown in Fig. 2.

Within the USDOT, the FHWA holds responsibility for technical and funding coordination and is a primary user of the data. Other primary data users include the National Highway Traffic Safety Administration (NHTSA), Federal Transit Administration (FTA), and the Bureau of Transportation Statistics (BTS); these agencies have historically participated in project planning and financial support. For example, NHTS data are integral to the calculation of the model year Corporate Average Fuel Economy (CAFE) standards, which are regulations issued by the National Highway Traffic Safety Administration (Chajka-Cadin et al., 2017). The data are commonly used to estimate vehicle miles of travel by specific groups, such as age groups or men and women drivers (Claude Ouimet et al., 2010), and the Bureau of Transportation includes analysis of personal travel from the NHTS data in the Transportation Statistics Annual Report (USDOT BTS) (US Department of Transportation, 2019).

The NHTS informs other policy research, especially energy consumption. For example, the Energy Information Administration (EIA) uses the NHTS data as a source of information on household fuel consumption for transportation (USDOE EIA) (US DOE Energy Information Agency, 2015) and the Environmental Protection Agency (EPA) to provide default values for the Moves2010 model (Beardsley, 2011) and to estimate the value of various policy initiatives related to fuel consumption and greenhouse gas emissions (US DOI EPA) (Department of Interior Environmental Protection Agency Green Vehicle Guide, 2015).

Travel behavior studies are a common use of the data, for instance analysis related to the impact of transportation policies on various population groups. Some of these studies are related to social equity (Sanchez et al., 2003), gendered travel (Hsu and Saphores, 2013), or have a focus on tracking age cohorts—like baby boomers (McGuckin and Lynot, 2012) or millennials (Dutzik et al., 2014). And the NHTS is used for pedestrian studies, including by the Safe Routes to School National Partnership (Randolph, 2015). A relatively new area of use is to assess the way that active transportation can improve public health: the Centers for Disease Control (CDC, 2010a) uses the data on the percent of children who walk to school (among other indicators) from the NHTS as part of their Healthy People 2020 (CDC, 2010b).

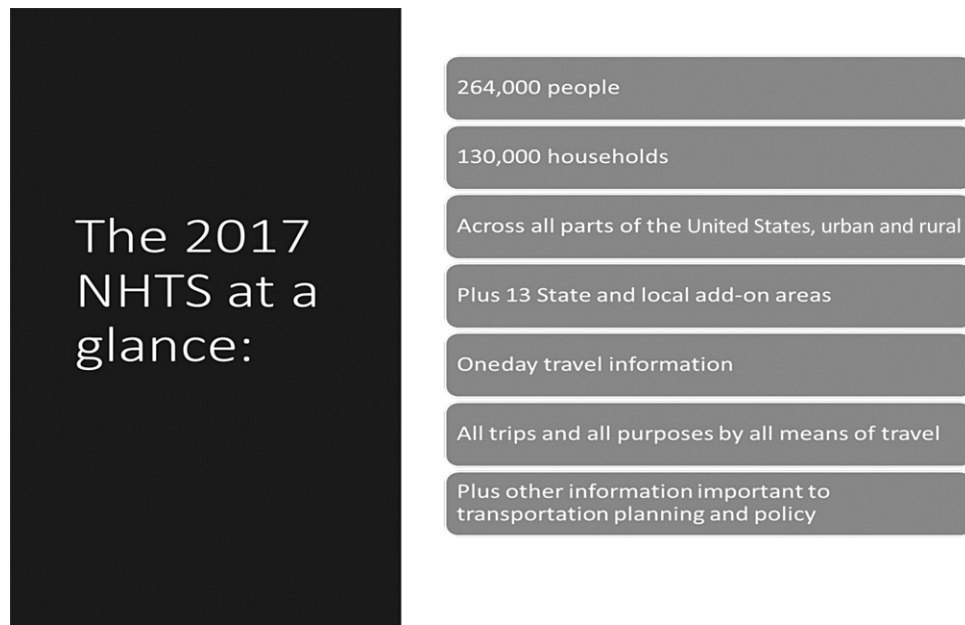


Figure 1 Summary of the 2017 National Household Travel Survey. *Source: USDOT FHWA National Household Travel Survey.*

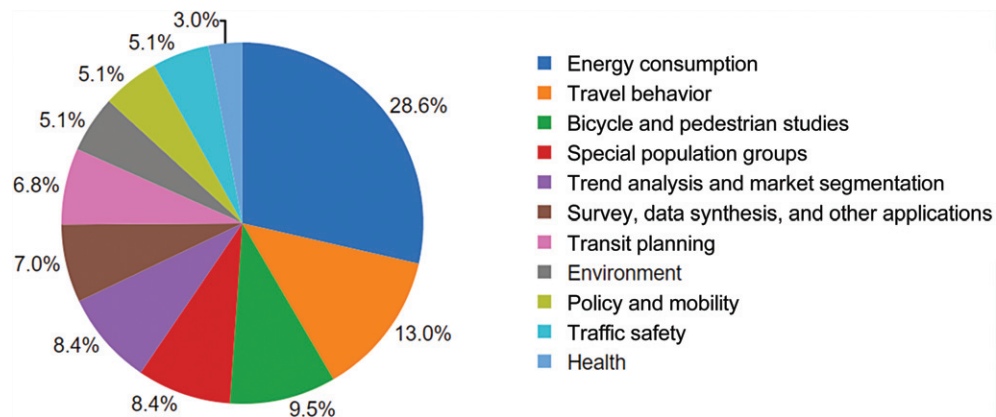


Figure 2 Compendium of uses of the National Household Travel Survey. *Source: USDOT FHWA National Household Travel Survey compendium of uses. Available from: <https://nhts.ornl.gov/compendium>.*

Advocacy groups and nonprofit organizations are another set of data users. For example, the American Association of Retired People (AARP) and the American Association of Automobiles Foundation for Traffic Safety (AAAFTS) often use NHTS data in their reports and presentations to increase awareness about priority topics and to lobby Congress for action. And, finally, transportation modelers use the NHTS data to provide context for regional and metro-area level travel behavior. Overall, policymakers, transportation planners and modelers, and decision-makers use the data and statistics generated from the NHTS extensively to understand the travel of the American public.

Biography



Nancy McGuckin is a senior advisor and consultant to federal, state, and local public agencies and private clients. She conducts primary research, data mining, and insightful statistical analysis to interpret and forecast travel behavior, examine special populations and equity issues, and help develop robust performance measures on these and other policy initiatives. She is well known for her work identifying technological, demographic, and cultural impacts on travel behavior, her rigorous data analysis, and her ability to communicate complex issues in a clear and concise manner.

Social, technological, and economic dynamics have all significantly affected travel rates—sometimes in surprising ways. Ms. McGuckin offers new insights gleaned from data mining the 40 years of NHTS, Census iPUIMS, and ACS/Journey-to-Work data to develop a context for understanding the recent unprecedented shifts in travel. For example, she has recently completed the NHTS 2017 Summary of Travel Trends (starting with the 1969 data, available from: https://nhts.ornl.gov/assets/2017_nhts_summary_travel_trends.pdf) and was a major contributor to the Transportation Statistics Annual Report (a summary of data findings important to policy makers, available from: <https://www.bts.gov/sites/bts.dot.gov/files/docs/browse-statistical-products-and-data/transportation-statistics-annual-reports/TSAR-2018-Web-Final.pdf>).

As part of her efforts to reinforce evidence-based policies, she frequently develops and posts briefs and presentations on critical issues related to travel behavior, demographic trends, and other topics at: www.travelbehavior.us.

References

- Beardsley, M., 2011. Fleet & activity data in MOVES2010. Available from: <https://www.epa.gov/sites/production/files/2016-06/documents/fleet-activity-moves-2011.pdf>.
- Centers for Disease Control and Prevention (CDC), 2010. CDC recommendations for improving health through transportation policy. Available from: <https://www.cdc.gov/transportation/docs/FINAL-CDC-Transportation-Recommendations-4-28-2010.pdf>.
- Centers for Disease Control and Prevention, President's Council on Fitness, Sports, and Nutrition (CDC), 2010. Healthy people 2020 midcourse review. Table 33-2. Available from: <https://www.cdc.gov/nchs/data/hpdata2020/HP2020MCR-C33-PA.pdf>.
- Chajka-Cadin, L., Petrella, M., Timmel, C., Fletcher, E., Mittleman, J., 2017. Federal highway administration research and technology evaluation: national household travel survey program final report. FHWA-HRT-16-082. Available from: <https://www.fhwa.dot.gov/publications/research/and/evaluations/16082/16082.pdf>.
- Claude Ouimet, M., Simons-Morton, B.G., Zador, P.L., Lerner, N.D., Freedman, M., Duncan, G.D., Wang, J., 2010. Using the U.S. National Household Travel Survey to estimate the impact of passenger characteristics on young drivers' relative risk of fatal crash involvement. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S0001457509002966>.
- Department of Interior Environmental Protection Agency Green Vehicle Guide, 2015. Citation 2. Available from: <https://www.epa.gov/greenvehicles/what-if-we-kept-our-cars-parked-trips-less-one-mile>.
- Dutzik, T., Inglis, J., Baxandall, P., 2014. U.S. PIRG millennials in motion—changing travel habits of young Americans and the implications for public policy. Available from: <https://uspirg.org/sites/pirg/files/reports/Millennials%20in%20Motion%20USPIRG.pdf>.
- Hsu, H.-P., Saphores, J.-D., 2013. Impacts of parental gender and attitudes on children's school travel mode and parental chauffeuring behavior: results for California based on the 2009 National Household Travel Survey. Available from: <https://link.springer.com/article/10.1007/s11116-013-9500-7>.
- McGuckin, N., Lynot, J., 2012. Impact of baby boomers on U.S. travel, 1969 to 2009. Available from: https://www.aarp.org/content/dam/aarp/research/public_policy_institute/liv_com/2012/impact-baby-boomers-travel-1969-2009-AARP-ppi-liv-com.pdf.
- Randolph, M., 2015. National Center for Safe Routes to school, how children get to school: school travel patterns from 1969 to 2009. Available from: <https://www.saferoutespartnership.org/resources/report/school-travel-patterns-1969-2009>.
- Sanchez, T., Stolz, R., Ma, J., 2003. Moving to equity: addressing inequitable effects of transportation policies on minorities. Available from: <https://escholarship.org/content/qt5qc7w8qp/qt5qc7w8qp.pdf>.
- US Department of Transportation, 2019. Bureau of transportation statistics. Transportation Statistics Annual Report 2019. Available from: <https://doi.org/10.21949/1502596>.
- US DOE Energy Information Agency, 2015. Today in energy, 2015: households with more vehicles travel more. Available from: <https://www.eia.gov/todayinenergy/detail.php?id=20832>.

Route Choice and Network Modeling

Emma Frejinger, Maëlle Zimmermann, Department of Computer Science and Operations Research and CIRRELT, Université de Montréal, Montreal, QC, Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction	496
Models for Analyzing and Predicting Path Choice Behavior	496
Preliminaries	497
Deterministic Versus Stochastic Models	497
Discrete Choice Models	497
Path-Based Models	498
Arc-Based Models	499
Traffic Flow Prediction and Related Problems	500
Traffic Assignment and Equilibrium	500
Bilevel Optimization and Network Design	500
Advantages and Drawbacks: Path-based Versus Arc-Based Models	501
Conclusion and Perspectives	502
References	502

Introduction

People make different choices because their preferences are different. This is true in many contexts, including transportation. Understanding and predicting traveler behavior is therefore crucial in order to design and operate transport systems that meet heterogeneous demand while minimizing effects negative to the environment as well as inefficiencies, such as congestion.

This article is devoted to route choice analysis and network modeling of heterogeneous traveler behavior. More precisely, we focus on analyzing and predicting path choice behavior with disaggregate stochastic models — discrete choice models — whose parameters are estimated using observed path choices (revealed preference data). The models are essentially used for two purposes. First, the interpretation of travel behavior by analyzing the parameter estimates. Such interpretations can for example be used as inputs in cost-benefit analysis of infrastructure investments. Second, predicting network traffic flows. Given an aggregate demand — number of people/vehicles traveling between each origin–destination (OD) pair — a disaggregate stochastic model distributes this demand over paths in the network. In this article we focus on disaggregate demand models and take aggregate demand as given and known. Of particular interest for flow prediction is the analysis of network changes. That is, the predictions should not only be accurate for the same network structure as during the data collection, but also generalize to new structures. This is why structural prediction models are used with attributes that aim to characterize the network, for example, type of infrastructure, travel times, and traffic signaling.

This article provides an overview of two, fundamentally different, modeling approaches to route choice analysis. We refer to them as path-based and arc-based and devote the third section to an overview of them. The purpose of this article is not to provide a detailed review of all types of path and arc-based models, rather we focus on advantages and drawbacks of the different modeling approaches from the perspective of the ultimate use of the models: behavioral interpretation, prediction, scenario analysis, or network design problems (i.e., optimization). To do so we provide a high-level description of related problems in the fourth section followed by the fifth section entirely devoted to the discussion of advantages and drawbacks of arc-based versus path-based models. Finally, Section 6 concludes.

Models for Analyzing and Predicting Path Choice Behavior

The problem of predicting how an aggregate OD demand distributes over paths in a network is a statistical learning problem. Having collected observations of individuals' trajectories in the network, the modeler's goal is to formulate a model, which identifies patterns in the data in order to accurately predict actual behavior. Nevertheless, modeling path choice should be driven not only by data but also by assumptions regarding how we expect individuals to behave. In the case of path choice in a transportation network, it is plausible to assume that decisions are not random but driven by a rational intent from individuals to minimize to some extent the disutility of traveling. Therefore path choice modeling can be cast as a problem of learning a utility function explaining a decision maker's behavior. In this section, we present different models to learn such a utility function and hence predict path choice behavior.

Preliminaries

We consider an oriented graph $G = (\mathcal{V}, \mathcal{A})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{A}$ is the set of arcs. Arcs $a \in \mathcal{A}$ are characterized by a head node i_a and a tail node j_a , and are associated to a utility v_a . A path in the graph G is a sequence of arcs such that the head node of each arc is the tail node of the next. Deterministic attributes, such as travel time, road, and traffic signaling information, are associated with arcs and nodes in the network. In this article we do not cover the more complex case of stochastic attributes that has been studied, for example, in [Gao et al. \(2010\)](#).

Such graphs may serve as representations for all kinds of urban transportation networks, where nodes include origins and destinations, and a path represents at some level of detail an alternative to travel between an OD pair. Typically, a sequence of arcs could define a precise itinerary in a physical network, where nodes $v \in \mathcal{V}$ represent intersections between road segments. More generally, paths may also represent traveling alternatives in space and time. For instance, a multimodal trip composed of several legs, each consisting of a distinct transport mode, intermediate OD, and departure/arrival time, could be represented as a path in a time-expanded network. In such a graph, a node v represents a point in space and time, while arcs a define more generally a means to connect two nodes with a specific transportation service.

Deterministic Versus Stochastic Models

Distinct assumptions regarding user behavior lead to either deterministic or stochastic models to distribute the aggregate demand over paths. The simplest path choice model in the literature is in fact a deterministic shortest path algorithm, which assigns individuals to the shortest (i.e., maximum utility) path between each OD pair. However, using this model to predict route choice in practice would rely on very strong behavioral assumptions, which are rarely valid. Implied is that individuals all perceive the same path utilities, have perfect discriminatory power, and are able to select the single best path 100% of the time. In reality, we observe on the contrary that individuals behave heterogeneously for several reasons, such as distinct preferences or varying perception of path attributes.

A way to relax this strong behavioral assumption is to resort to *stochastic* path choice models. The hypothesis underlying such models is that travelers are random utility maximizers: each individual perceives his/her own utility, which may be based on personal preferences or criteria unavailable to the modeler. Hence the modeler does not know each individual utility, but only the distribution among the population. This leads to representing path utilities as random variables (consisting of a deterministic and a random term) and distributing the aggregate demand on the network's paths according to the probability of each path being optimal, that is, the one with maximum utility.

Whether a deterministic or stochastic model is used, in practice we *do not know* the utilities of the network's paths nor their distributions. However, observed trajectories relay information regarding what paths are used in the network. This is why path choice modeling is in fact an inverse problem, which consists in identifying from the data the true utility function within the path choice model under consideration.

This inverse problem can be approached in two ways. On the one hand, under the assumption of deterministic path utilities, the learning problem is simply an inverse shortest path, that is, identifying the values associated to each arc, which make observed paths optimal (see e.g., [Burton and Toint, 1992](#)). This problem is considered to be difficult because it is under determined. In other words, there does not necessarily exist unique arc utilities ensuring that observed paths are optimal. On the other hand, when assuming that utilities are random variables, the inverse problem becomes one of learning the parameters of a probability distribution over a set of paths. Observations can be interpreted as noisy, that is, deviating randomly from deterministic shortest paths according to the probability function. Inverse problems with noisy data pose a double challenge. The difficulty stems not only from nonuniqueness, but also from the fact that there may not be any nontrivial value of arc utilities, which accounts for observed behavior. For more information on such inverse problems, the reader is referred to [Zimmermann and Frejinger \(2019\)](#).

In this paper we focus on learning the utility function of stochastic path choice models, which are based on more realistic assumptions of human behavior than the deterministic counterparts. In the next section, we present the most widely used methodology in this context, namely discrete choice modeling.

Discrete Choice Models

Discrete choice models are stochastic demand models grounded in random utility maximization theory. They provide a methodology to statistically estimate parametric models of behavior where an individual makes a choice between a discrete number of alternatives, under the assumption that alternatives are associated with a random utility, and an individual's choice is the outcome of utility maximization. In the context of path choice, alternatives correspond to paths between an individual's OD pair.

Discrete choice models assume that the utility function is a parametric function of several attributes, usually linear-in-parameters. This parameterization allows to resolve the under-determinism which otherwise characterizes the inverse problem. Fitting the noisy path observations corresponds to finding the parameters which minimize the error of the data, which is measured by the log-likelihood loss. In other words, the model's parameters are selected by maximum likelihood estimation, such that the estimated parameters $\hat{\beta}$ maximize the probability of the observations $n = 1, \dots, N$,

$$\hat{\beta} = \arg \max_{\beta} \ln \prod_{n=1}^N P_n(i; \beta), \quad (1)$$

where $P_n(i; \beta)$ is the probability of path alternative i chosen by individual n . Note that we use the same notation for individuals and observations implicitly assuming that we do not have panel data.

Inherent to the problem of path choice is a combinatorial challenge, due to the exponential number of paths connecting each OD pair in real networks. Indeed, in practice it is impossible to enumerate all the paths an individual could choose from between an OD pair. Hence it is difficult to specify the form of the probability distribution: over which paths should the probability be defined? If all paths, how to define the support for probability distribution without resorting to enumeration?

The literature is well acquainted with this issue. As a result, there exist two kinds of discrete choice models — path-based and arc-based — which propose each a different solution, introduced in the subsequent sections. In the former, the action of choosing a path is modeled in two steps, while in the latter, it is modeled as a sequential process of arc choices. In the following sections, we detail how a discrete choice model can be estimated and used for prediction with each approach.

Path-Based Models

In path-based models, the probability distribution over paths for an individual n traveling between a given OD is defined in two steps as follows.

1. The modeler restricts the number of paths to a certain individual specific choice set, denoted C_n . Most of the choice set generation algorithms (see, e.g., [Prato, 2009](#)) generate path alternatives by iteratively computing shortest paths according to different generalized cost functions or on different network structures.
2. Paths not belonging to C_n are assigned a zero choice probability, while paths in C_n obtain a probability normalized by a constant computed over all paths in C_n .

The model's formulation relies on path-based variables. The utility of a path i is a random variable $u_i = v_i + \mu \varepsilon_i$, where ε_i is a random error term, μ is its scale, and v_i is the deterministic component of the utility parametrized by attributes, which can be additive over arcs or not. While not required, the deterministic utility is often linear in parameters, that is, $v_i = \beta^T x_i$, where β is a vector of parameters to be estimated and x_i is a vector of attributes of the path i . Individuals are assumed to maximize random utility, therefore the probability of choosing an alternative $i \in C_n$ is given by:

$$P(i|C_n; \beta) = P(u_i > u_j, \forall i \neq j \in C_n). \quad (2)$$

The form of the path choice probability $P(i|C_n; \beta)$ depends on the distributional assumption on the error terms ε_i . The simplest model assumes i.i.d. Extreme Value Type I distributed error terms and is known under the name multinomial logit (MNL) or softmax in machine learning,

$$P(i|C_n; \beta) = \frac{e^{\frac{1}{\mu} v_i(\beta)}}{\sum_{j \in C} e^{\frac{1}{\mu} v_j(\beta)}}, \quad (3)$$

while more complex models account for correlation between path utilities in the stochastic part of the utility function. In the latter category, we note that there exists among others the paired combinatorial logit ([Pravinongvuth and Chen, 2005](#)) the cross-nested logit ([Vovsha and Bekhor, 1998](#)) and the mixed logit models (e.g., [Bekhor et al., 2002](#); [Frejinger and Bierlaire, 2007](#)).

In terms of complexity, the estimation of models, which introduce correlated error terms, is more costly, due to the additional model parameters and the nonconvexity introduced in the log-likelihood function. Furthermore, mixed logit models require to resort to simulated maximum likelihood estimation ([Train, 2009](#)). However, accounting for correlation between error terms allows to more correctly represent the perceived similarity between overlapping paths and thus improve the predicting accuracy of the models.

There exists an alternative approach to estimate path-based models based on the assumption that individuals may choose any path in the network, known as sampling of alternatives. It consists in correcting the path utilities for the sampling protocol, which generated the restricted choice sets ([Frejinger et al., 2009](#)). The correction requires an algorithm that samples paths with known probability, which is a difficult problem in itself ([Flötteröd and Bierlaire, 2013](#)). [Guevara and Ben-Akiva \(2013a, 2013b\)](#) show that it is possible to estimate mixed logit and multivariate extreme value logit models using sampling of alternatives. To the best of our knowledge, in a route choice context the use has been limited to the logit model with path size attribute ([Frejinger et al., 2009](#)) and cross-nested logit ([Lai and Bierlaire, 2015](#)). While the sampling approach allows to correct the utilities to obtain consistent estimates, it is unclear how to compute the path sampling correction for prediction. In the following we therefore discuss prediction with path-based models without sampling correction.

The estimated models, $P(i|C_n; \hat{\beta})$ where we denote the maximum likelihood estimates $\hat{\beta}$, can be used to predict path choice (simulation from the estimated probability distribution) or, based on a given OD matrix, to compute expected arc flows. Path-based models require to decompose the prediction problem by OD pair, and possibly further by category of individual. The latter is necessary if path utilities are a function of socio-economic characteristics of individuals. We denote OD specific choice sets C^{od} (which may, in addition, be specific to individual category). The corresponding demand g^{od} is then assigned proportionally to the

predicted choice probabilities for each path in C^{od} . Flow on arc a for a given OD pair can be obtained by summing the flows on all paths traversing arc a ,

$$f_a^{od} = \sum_{i \ni a} P(i|C^{od}; \hat{\beta}) g^{od}. \quad (4)$$

Total expected flow f_a on arc a can be obtained by summing over OD pairs,

$$f_a = \sum_{od} f_a^{od}. \quad (5)$$

Arc-Based Models

In arc-based models, the process of choosing a path i is modeled as a sequential process of arc choices. To be more precise, it is formulated as an undiscounted finite horizon Markov decision process with an absorbing state, which is the destination. Instead of reducing the number of options to a given choice set, each path is associated to a nonzero probability, corresponding to the product of the probability of choosing each arc composing the path. The choice of arc of an individual is independent of the trajectory followed up to this point, but may depend on the incoming arc. One reason to allow it is to account for turning angles. For this reason, we denote an ingoing arc k , and outgoing arcs from k are denoted $a \in \mathcal{A}(k)$.

The model is formulated with arc variables. The instantaneous deterministic utility of an arc a when coming from k is denoted $v(a|k)$. Implied is that there exists another component to an arcs's utility. The latter is called the value function, which represents the maximum expected utility to be gained from arc a until destination is reached. For an individual traveling to destination d , the random utility $u(a|k)$ of an outgoing arc a to k is then the sum of its instantaneous deterministic utility $v(a|k) = \beta^T x(a|k)$, an error term $\mu \varepsilon_a$, and the destination-specific value function V_a^d , that is,

$$u(a|k) = v(a|k) + \mu \varepsilon_a + V_a^d. \quad (6)$$

The value function is defined recursively according to the Bellman equation:

$$V_k^d = \begin{cases} 0, & k = d, \\ E_\varepsilon [\max_{a \in \mathcal{A}(k)} \{v(a|k) + \mu \varepsilon_a + V_a^d\}], & \forall k \in \mathcal{A}. \end{cases} \quad (7)$$

Similarly to path-based models, arc-based recursive models can allow random terms to be correlated or not. The standard assumption of i.i.d. Extreme Value Type I error terms yields the recursive logit (RL) model (Fosgerau et al., 2013). However there exists several arc-based models accounting for correlation between path utilities, notably the nested recursive logit (NRL), the mixed recursive logit (MRL), and the recursive crossnested logit (RCNL) models (Mai, 2016; Mai et al., 2015; Mai et al., 2018). Moreover, Oyama and Hato (2017) use a discounted RL model to analyze travel behavior in a network going into gridlock.

Both the RL and the NRL are based on an MNL model over the outgoing arcs, the difference being that the NRL possesses arc-specific scale parameters μ_k . In more complex models such as the RCNL, each arc choice situation is modeled with a multivariate extreme value model. In general, the choice probability $P(i; \beta)$ of a path i consisting of a sequence of arcs $k_0, k_1, \dots, k_T$ is given by the product of arc choice probabilities:

$$P(i; \beta) = \prod_{j=0}^{T-1} P^d(k_{j+1}|k_j; \beta) \quad (8)$$

When arc choice probabilities $P^d(k_{j+1}|k_j; \beta)$ correspond to an MNL with fixed scale μ , the RL path choice probability takes a simple form and is equivalent to that of a path-based model with unrestricted choice set. In this case, arc choice probabilities are given by:

$$P^d(a|k; \beta) = \frac{e^{\frac{1}{\mu} v(a|k; \beta) + V_a^d(a| \beta)}}{\sum_{a' \in \mathcal{A}(k)} e^{\frac{1}{\mu} v(a'|k; \beta) + V_{a'}^d(a' | \beta)}} \quad (9)$$

The value function in this case can also be conveniently computed by solving a system of linear equation. In contrast, all models which relax the assumption of i.i.d error terms are computationally more costly to estimate, in part because the value functions are more complex to solve and because the log-likelihood becomes non convex.

In order to compute the log-likelihood in Eq. (1), path choice probabilities in Eq. (8) can be used. However, those require to solve the value function in Eq. (7), since arc choice probabilities depend on it. Hence the model is usually estimated using the nested fixed point algorithm, which consists of an outer loop which searches over the parameter space, and an inner loop which solves the value function for each trial value of β (Rust, 1987).

In order to predict arc flows from an OD matrix, recursive route choice models offer two approaches, namely (i) simulation, and (ii) solving a system of linear equations, as in, for example, Baillon and Cominetti (2008). The latter allows to decompose the problem by destination only. Indeed, arc flows f^d to destination d can be obtained as the result of the destination specific system:

$$f^d = (I - P^d(\hat{\beta})^T)^{-1} g^d, \quad (10)$$

where g^d is the demand vector from all origins to destination d , and $P^d(\hat{\beta})$ is the estimated matrix of arc choice probabilities to destination d , where $P^d_{k,a}$ represents the choice probability of outgoing arc a from arc k . It then suffices to sum the destination specific flows in order to obtain the vector of total arc flows f ,

$$f = \sum_d f^d. \quad (11)$$

Traffic Flow Prediction and Related Problems

Path choice predictions — Eqs. (2) and (8) — or predicted arc flows — Eqs. (5) and (10) — are rarely useful on their own, rather, they are central components in models of other problems that depend on path choice predictions. In this context, network design (Magnanti and Wong, 1984) constitutes an important class of problems and encompass, for example, facility location, network pricing and flow capture. We devote section, “Bilevel Optimization and Network Design”, to bilevel optimization linking this class of problem to path choice modeling. First, however, it is important to discuss the role of congestion when computing traffic flows in networks with capacity restrictions. Indeed, if the network of interest is congested, then demand and supply interact and flows at a traffic equilibrium provide more useful information than the flows assuming an uncapacitated network as in Eqs. (5) and (10). A detailed discussion of network design and traffic equilibrium is clearly out of the scope of this article. Instead, we give a brief high-level description with the purpose to provide guidance on the trade-offs involved in the choice of a path choice model and its specification, that is, choice sets, distribution of error terms, and functional form of the utilities. We first focus on traffic assignment and equilibrium solutions followed by bilevel optimization.

Traffic Assignment and Equilibrium

Stochastic traffic assignment models are used to predict flow in a network under the assumption that travelers choose paths according to an additive random utility maximization model. Through attributes that depend on flow — for example, a link performance function that gives travel time as a function of flow — the utilities depend on capacity and interaction with other travelers in the network. A stochastic user equilibrium is reached when no user can unilaterally improve his/her utility by changing path (Daganzo and Sheffi, 1977). The equilibrium can in some cases be characterized as a solution to a minimization problem (e.g., Sheffi, 1985) and the associated solution algorithm is typically based on iterative stochastic traffic assignments.

As can clearly be seen in Eqs. (5) and (10), the assignment models rely on (i) a choice model and (ii) information on aggregate demand in the form of OD matrices. They state how many vehicles need to travel between each OD pair in a given time interval. We note that if the utilities of the path choice models depend on socioeconomic attributes of the travelers, then the OD matrices must be defined accordingly. In practice, the assignment is often limited to a few user categories each having an OD matrix and the utilities otherwise depend on network attributes only.

Through the path choice model, the traffic assignment crucially depends on the choice set assumption. The flows can be limited to a restricted network defined for each OD pair, either explicitly through the path choice sets (Bekhor et al., 2008), or implicitly as proposed by Dial (1971). A restricted network should be defined so that it comprises all alternatives having a utility high enough to attract nonzero flow. This means that the definition of a restricted network needs to be updated if network attributes, such as travel times change, or if the structure of the network itself change, for example by the addition or removal of arcs. As opposed to this restricted network assumption, Baillon and Cominetti (2008) propose a Markovian traffic equilibrium (MTE) for unrestricted networks (MTE extends other works, for example, Akamatsu, 1996; Bell, 1995, that also avoid the assumption of a restricted network). The model relies on an arc-based path choice model defined by destination specific arc choice probability matrix. It is sufficient that the latter is a stochastic matrix; hence, it is not limited to a specific choice model.

An important practical use of assignment and equilibrium solutions is the analysis and comparison of network scenarios, for example: How do the flows in the network change if we increase speed on certain links? How do the flows in the network change if we add new infrastructure? The computation of an equilibrium solution (or assignment if the network is uncongested) allows to compare the different scenarios. Network design problems go beyond such simple scenario analysis and aim at finding the optimal network structure given demand.

Bilevel Optimization and Network Design

Bilevel optimization is suitable to model a hierarchical relationship between decision makers with possibly conflicting objectives (see, e.g., Colson et al., 2007, for an overview). The leader (upper-level problem) first makes decisions, then the followers (lower-level problem) make decisions according to their objective function that depends on the leader’s decisions. In our case, the leader takes network design decisions and the followers make path choices in the network. Important to emphasize is the link between the two optimization problems. The leader can anticipate the reactions of the followers which leads to a nested optimization problem where constraints in the upper-level problem involves decision variables from both levels. Bilevel optimization problems are non-convex and non-differentiable and even the simplest case when upper and lower problems are linear programs is known to be NP-hard (Jerusalem, 1985).

Bilevel optimization formulations of network design problems where users are assigned to deterministic shortest paths or where users achieve a deterministic user equilibrium are well studied (a few examples are Abdulaal and LeBlanc, 1979; Brotcorne et al., 2008; Labbé et al., 1998; Marcotte, 1986; Nair and Miller-Hooks, 2014). On the contrary, the literature on the stochastic counterpart where users are assigned to paths according to a discrete choice model is scarce. Exceptions are the series of works by Gilbert et al. (2014a, 2014b, 2015) devoted to the network pricing problem where the leader selects revenue-maximizing tolls on a subset of arcs in the network and the followers select random utility maximizing paths where the utilities depend on tolls, among other attributes. The stochastic variant leads to challenging nonlinear optimization problems with strong combinatorial features. Solution algorithms rely on reformulations to a single level and approximations which, in turn, depend on the analytical properties of the choice model. The aforementioned studies therefore use the nice properties of the logit and mixed logit models with utilities that are linear, at least in upper-level decision variables.

Given the importance of network design problems for applications where travelers are known to exhibit heterogeneous behavior, we believe that there are several promising avenues for future research related to bilevel optimization. First, modeling followers reactions according to choice models with better prediction accuracy than logit. Second, model followers' reaction through a stochastic user equilibrium solution, in which case the problem belongs to the class of mathematical programs with equilibrium constraints (MPEC).

Advantages and Drawbacks: Path-based Versus Arc-Based Models

In this section we discuss the advantages and drawbacks of the path and arc-based approaches to model path choice behavior in a network using discrete choice. The essential points are summarized in Table 1.

The majority of studies in the literature are based on the first approach, namely path-based models with generated choice sets, which are treated as the true ones. The major drawback of this approach is its reliance on path choice sets, which are costly to generate and may lead to biased parameter estimates (Frejinger et al., 2009). In turn, this can result in counterintuitive results and an inaccurate behavioral analysis. Its main advantage is that once choice sets of paths are specified, the model can be estimated with a standard software for maximum likelihood estimation. There exists several extensions of path-based models to incorporate correlation between path utilities, and any imaginable path attribute — additive over arcs or not — can be incorporated in the utility function, which makes this approach flexible in terms of model specification. Another drawback is yet related to prediction. Since predicted choice probabilities vary depending on the generated choice sets which are defined arbitrarily (i.e., without support from data) this approach can lead to predictions that are arbitrarily bad. Moreover, particular care has to be taken when path-based models are used for scenario analysis or as part of optimization. In this case choice sets need to be regenerated each time changes to the network occur. This is costly and also impact the comparison of scenarios as it can be unclear if a change in traffic flow is simply due to a change in the choice sets or a change to the network. In this context, counterintuitive results have been reported in the literature (Nassir et al., 2014) where an improvement to the network leads to worse accessibility (the reason for the results is explained in Zimmermann et al., 2017).

Path-based models with sampled choice sets belong to the approaches based on an unrestricted network. By estimating the models with a correction term that accounts for the path sampling protocol, the resulting parameter estimates are unbiased.

Table 1 Summary of advantages (+) and drawbacks (–) for each approach

Restricted network		Unrestricted network	
Path-based models with generated choice sets		Path-based models with sampled choice sets	Arc-based models
Estimation	+ Standard software – Parameter estimates may be biased – Choice sets costly to generate + Flexibility regarding models (correlated utilities) + Flexible utility specification	+ Standard software – Computation of a correction term – Choice sets costly to generated and limited choice of algorithms – Models limited because of correction term + Flexible utility specification	– Non standard software + Consistent parameter estimates – Computation of value functions + Some flexibility regarding models – Arc-additive utility specification
Prediction	– Potential re-generation of choice sets (per OD pair) – Predictions can be arbitrarily bad (depending on choice sets). Limitations when comparing network scenarios. – Predictions computed per OD pair	– Unclear how to use for prediction	– Computation of value functions (per D or O) + Accurate predictions that can be directly compared across scenarios + Predictions computed per D or O – Predicts paths with cycles with non-zero probability

Nevertheless, the costly necessity to generate choice sets remains, and the algorithms at disposal are restricted to those with known path sampling probabilities. A major drawback is that the model cannot predict according to the true estimated choice probabilities, since sampling choice sets is required for prediction as well. Therefore this approach is not recommended for applications, which go beyond the analysis of estimated parameters. We note that it is, however, possible to estimate models using sampled choice sets and then using an arc-based model for prediction, as in Zimmermann et al. (2018) who study activity-travel scheduling behavior.

Arc-based models overcome the main drawbacks of path-based models. The models can be consistently estimated without requiring any path choice sets. This necessity is replaced by the need to solve value functions at each step of the maximum likelihood estimation algorithm, which are quick to solve for the RL model, but more costly for the recursive counterparts which account for correlation. Since the estimation algorithm requires to solve the value functions as part of an inner algorithm when searching over the parameter space, non-standard estimation software is required. Some of the recursive models are implemented in MATLAB, in a code available online, supplemented by a tutorial detailing how to use it (<http://intermodal.iro.umontreal.ca/software.html>).

A potential downside is that the arc-based models require attributes that are additive over arcs, so nonadditive attributes, such as slope, need to be transformed into additive ones (Zimmermann et al., 2017). This can potentially be challenging and increase the size of the state space. Nevertheless, we note that although path-based models seem more flexible in this respect, their choice set generation step involve shortest path algorithms which, in turn, require arc-additive attributes.

Regarding prediction, the arc-based approach is the most straightforward and the only one which allows to predict path choices according to the true estimated probabilities even under network changes. Although value functions need to be solved for prediction, it need only be done once, as opposed to numerous times during estimation. Prediction is faster thanks to the fact that value functions are destination (or origin) specific as opposed to OD specific path choice sets. While arc-based models on an unrestricted network have several major advantages over path-based models, they have the drawback of assigning nonzero probability to paths with cycles (provided, of course, that the network contains cycles). It is clearly behaviorally unrealistic that travelers choose to make numerous cycles. Fosgerau et al. (2013) numerically investigate the issue and conclude on real data that the probability of even one cycle is very small. We note that if simulated path predictions are of interest and paths with cycles should not be considered, it is easy to detect cycles and truncate the distribution to noncyclic paths.

Conclusion and Perspectives

The problem of route choice modeling consists of learning the utility function of a rational cost-minimizing decision maker from data. Deterministic or stochastic models can be used to model route choice depending on the application and assumptions made on the data. Among stochastic models, discrete choice models are the most widely used in the transportation research community. This article gave an overview of that literature with the purpose to highlight advantages and drawbacks of existing discrete choice models, depending on their ultimate use. We detailed two main modeling approaches, namely path-based and arc-based models.

Path-based models essentially restrict the available choices a traveler can make in the network. This allows to keep users' modeled behavior within sensible limits, and typically prevents the improbable occurrence of cycles. Nevertheless, path-based models have important theoretical drawbacks both regarding estimation and prediction, which cannot be fully overcome by using sampling corrections as in Frejinger et al. (2009).

Arc-based models are based on an unrestricted network assumption by modeling the choice of paths arc by arc. It leads to a mathematically advantageous formulation, with unbiased parameter estimates as well as fast and accurate predictions.

Each approach possesses advantages and drawbacks. However, we believe that the efficiency and accuracy of the arc-based formulation largely overcome its flaws. Implementing the estimation and prediction of such models in standard software would therefore benefit transportation researchers and practitioners.

References

- Abdulaal, M., LeBlanc, L.J., 1979. Continuous equilibrium network design models. *Transport. Res. B Methodol.* 13 (1), 19–32.
- Akamatsu, T., 1996. Cyclic flows, markov process and stochastic traffic assignment. *Transport. Res. B Methodol.* 30 (5), 369–386.
- Baillon, J.-B., Cominetti, R., 2008. Markovian traffic equilibrium. *Mathemat. Program.* 111 (1–2), 33–56.
- Bekhor, S., Ben-Akiva, M.E., Ramming, M.S., 2002. Adaptation of logitkernel to route choice situation. *Transport. Res. Rec.* 1805 (1), 78–85.
- Bekhor, S., Toledo, T., Prashker, J.N., 2008. Effects of choice set size and route choice models on path-based traffic assignment. *Transportmetrica* 4 (2), 117–133.
- Bell, M.G., 1995. Alternatives to dial's logit assignment algorithm. *Transport. Res. B Methodol.* 29 (4), 287–295.
- Brotcorne, L., Labbé, M., Marcotte, P., Savard, G., 2008. Joint design and pricing on a network. *Operat. Res.* 56 (5), 1104–1115.
- Burton, D., Toint, P.L., 1992. On an instance of the inverse shortest paths problem. *Mathemat. Program.* 53 (1–3), 45–61.
- Colson, B., Marcotte, P., Savard, G., 2007. An overview of bilevel optimization. *Annal. Operat. Res.* 153 (1), 235–256.
- Daganzo, C.F., Sheffi, Y., 1977. On stochastic models of traffic assignment. *Transport. Sci.* 11 (3), 253–274.
- Dial, R.B., 1971. A probabilistic multipath traffic assignment model which obviates path enumeration. *Transport. Res.* 5 (2), 83–111.
- Flötteröd, G., Bierlaire, M., 2013. Metropolis-hastings sampling of paths. *Transport. Res. B Methodol.* 48, 53–66.
- Fosgerau, M., Frejinger, E., Karlström, A., 2013. A link based network route choice model with unrestricted choice set. *Transport. Res. B* 56, 70–80.
- Frejinger, E., Bierlaire, M., 2007. Capturing correlation with subnetworks in route choice models. *Transport. Res. B Methodol.* 41 (3), 363–378.
- Frejinger, E., Bierlaire, M., Ben-Akiva, M., 2009. Sampling of alternatives for route choice modeling. *Transport. Res. B Methodol.* 43 (10), 984–994.
- Gao, S., Frejinger, E., Ben-Akiva, M., 2010. Adaptive route choices in risky traffic networks: A prospect theory approach. *Transportat. Res. C Emerg. Technol.* 18 (5), 727–740.

- Gilbert, F., Marcotte, P., Savard, G., 2014a. Logit network pricing. *Comput. Operat. Res.* 41, 291–298.
- Gilbert, F., Marcotte, P., Savard, G., 2014b. Mixed-logit network pricing. *Computat. Optimizat. Applicat.* 57 (1), 105–127.
- Gilbert, F., Marcotte, P., Savard, G., 2015. A numerical study of the logit network pricing problem. *Transport. Sci.* 49 (3), 706–719.
- Guevara, C.A., Ben-Akiva, M.E., 2013a. Sampling of alternatives in logit mixture models. *Transport. Res. B Methodol.* 58, 185–198.
- Guevara, C.A., Ben-Akiva, M.E., 2013b. Sampling of alternatives in multivariate extreme value (MEV) models. *Transport. Res. B Methodol.* 48, 31–52.
- Jeroslow, R.G., 1985. The polynomial hierarchy and a simple model for competitive analysis. *Mathema. Program.* 32 (2), 146–164.
- Labbé, M., Marcotte, P., Savard, G., 1998. A bilevel model of taxation and its application to optimal highway pricing. *Manag. Sci.* 44 (12-part-1), 1608–1622.
- Lai, X., Bierlaire, M., 2015. Specification of the cross-nested logit model with sampling of alternatives for route choice models. *Transport. Res. B Methodol.* 80, 220–234.
- Magnanti, T.L., Wong, R.T., 1984. Network design and transportation planning: Models and algorithms. *Transport. Sci.* 18 (1), 1–55.
- Mai, T., 2016. A method of integrating correlation structures for a generalized recursive route choice model. *Transport. Res. B Methodol.* 93, 146–161.
- Mai, T., Bastin, F., Frejinger, E., 2018. A decomposition method for estimating recursive logit based route choice models. *EURO J. Transport. Logist.* 7 (3), 253–275.
- Mai, T., Fosgerau, M., Frejinger, E., 2015. A nested recursive logit model for route choice analysis. *Transport. Res. B Methodol.* 75, 100–112.
- Marcotte, P., 1986. Network design problem with congestion effects: a case of bilevel programming. *Mathema. Program.* 34 (2), 142–162.
- Nair, R., Miller-Hooks, E., 2014. Equilibrium network design of shared vehicle systems. *Eur. J. Operat. Res.* 235 (1), 47–61.
- Nassir, N., Ziebarth, J., Sall, E., Zorn, L., 2014. Choice set generation algorithm suitable for measuring route choice accessibility. *Transport. Res. Rec.* 2430 (1), 170–181.
- Oyama, Y., Hato, E., 2017. A discounted recursive logit model for dynamic gridlock network analysis. *Transport. Res. C Emer. Technol.* 85, 509–527.
- Prato, C.G., 2009. Route choice modeling: past, present and future research directions. *J. Choice Model.* 2 (1), 65–100.
- Pravinovongvuth, S., Chen, A., 2005. Adaptation of the paired combinatorial logit model to the route choice problem. *Transportmetrica* 1 (3), 223–240.
- Rust, J., 1987. Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher. *Econometrica: J. Econometr. Soc.* 55 (5), 999–1033.
- Sheffi, Y., 1985. *Urban transportation networks* vol. 6. Prentice-Hall, Englewood Cliffs, NJ.
- Train, K., 2009. *Discrete Choice Methods with Simulation*, second ed. Cambridge University Press, Cambridge, United Kingdom.
- Vovsha, P., Bekhor, S., 1998. Link-nested logit model of route choice: overcoming route overlapping problem. *Transport. Res. Rec.* 1645 (1), 133–142.
- Zimmermann, M., Frejinger, E., 2019. A tutorial on recursive models for analyzing and predicting path choice behavior, Accepted for publication in *EURO J. Transport. Logist.*
- Zimmermann, M., Mai, T., Frejinger, E., 2017. Bike route choice modeling using GPS data without choice sets of paths. *Transport. Res. C Emer. Technol.* 75, 183–196.
- Zimmermann, M., Västberg, O.B., Frejinger, E., Karlström, A., 2018. Capturing correlation with a mixed recursive logit model for activity-travel scheduling. *Transport. Res. C Emer. Technol.* 93, 273–291.

Departure Time Choice Modeling

Khandker Nurul Habib, Department of Civil & Mineral Engineering, University of Toronto, Toronto, ON, Canada

© 2021 Elsevier Ltd. All rights reserved.

Background	504
Departure Time Choices	504
Modeling Techniques	505
Data for Empirical Modeling	506
Policy Evaluations	507
Conclusions	507
Acknowledgment	507
Biography	508
References	508
Further Reading	508

Background

Departure time choice models are either parts of travel demand modeling systems or stand-alone policy analysis tools necessary to investigate the temporal distribution of transportation demands. Modeling departure time choices intends to capture the preferences of travelers to specific time slots of the day to start out-of-home activities. Out-of-home activities require movement from one location to another location (defined as a trip). The choice of the departure time for an out-of-home activity is, in fact, the choice of the trip start time to the activity location (destination). Conventionally, departure time choice is modeled as opposed to arrival time (at destination location to start the activity) to recognize the uncertainty in transportation network performance and resulting randomness of travel time to reach the destination. Naturally, travelers tend to choose the time of departure for a trip with the expectation of travel time it would take to reach the destination. Depending on the purposes of destination activity and travel time uncertainty (resulting from dynamics of transportation network performance), different time-of-day can be feasible. For example, for work and school activities, activity start times are often fixed by the type and nature of work and school. In such a case, corresponding departure time choices may not have an extended period to choose from. Regular work and school start in the morning and ends in the afternoon. This can be different for people with flexibilities in work start times and work durations. In nonwork activities, such as shopping, recreational, and social activities, people may have more flexibility to choose departure times. In any case, modeling departure time choice is modeling the traveler's choice of a time-of-day to start a trip.

Departure time choice of the different trips of urban residences has profound implication to the performances of the urban transportation system (Ashiru et al., 2004). The peak and off-peak period of travel demand and corresponding congestion effects in transportation network are results of clustering of departure time choices at particular parts of the day. Objectives of departure time choice models include investigating daily travel behavior for future demand forecasting to specific transportation policy evaluations, for example, peak spreading, congestion pricing, high occupancy vehicle (HOV) lane usage, etc. Departure time choice model is pivotal to any demand modeling system in accurately capturing time-varying vehicle flow patterns on urban transportation networks, which is also called transportation system performance. Transportation system performances define the difference between the departure time of the trip and the arrival time to start the activity (the travel time of the trip). Schedule delay in terms of early arrival or late arrival and corresponding penalty (dissatisfaction) influence the dynamics of departure time choices. It has become customary that departure time choice is explicitly modeled and arrival time and early/late arrivals are estimated based on travel time to reach the destination.

Modeling the choice of departure time has been an important research question since the early 1970s. Earlier efforts to capture the effects of influencing factors of departure time choices are predominantly for car users. Later, the departure time choices of transit users are also being considered to investigate transit demand. Modeling departure time choices of commuting trips by car has been receiving continuous research focus. Researchers investigated whether transportation network performances (variation of travel time and cost for different starting points) alone or other factors influence the choice of departure time alternatives. Other factors include job-related attributes (arrival departure time flexibility, flexible duration of work), job category, job type, age, gender, income, and other socioeconomic factors.

Departure Time Choices

Departure time choice models can be different for different trip types as activity-scheduling constraints vary for different activity types. Also, the departure time choices of any trip can be correlated with the travel mode of the trip. Departure time choices of car

drivers can be different from the departure time choices of transit users. Similarly, departure time choice for carpool users can be different from the departure time choices of taxi and ride-sharing or ride sourcing services. The type and approach of modeling departure time choices can vary by the context of the application of the models. Generally, dealing with departure time choices can be part of:

1. A developing specific and stand-alone policy analysis tool
2. A traffic assignment model development
3. A demand modeling system

To develop a stand-alone policy analysis tool, explicit modeling of departure time choices of different modes for different activity type is conducted. Such models allow capturing effects of transportation system performances (travel time and cost), parking restrictions, business office hours, activity schedule flexibility, and various socioeconomic factors on the choice of departure time alternatives.

The significant difference between a static and a dynamic traffic assignment model to predict transportation network flow is that static assignment overlooks distribution of departure time choices of the travelers. Modeling dynamic traffic assignment requires explicit specification departure time choices either through aggregate departure time distribution or disaggregate departure time choice modeling. Aggregate departure time distribution is defined by patterns of departure time choices of urban residences as observed through travel surveys or the changing patterns of traffic volume on the networks. Disaggregate departure time models for dynamic traffic assignment can be discrete choice models or continuous time allocation (to different activity type) choice model.

Departure time choices are explicitly modeled for activity-based travel demand modeling systems (Habib, 2013, 2018). Typical four-stage urban transportation models usually consider static traffic assignment and so departure time choices for different trips are often overlooked. However, for activity-based travel demand model, it is customary that departure time choices of different activities that require making trips to the destination location are modeled through modeling activity scheduling. Alternative activity-scheduling approach includes:

- Use of univariate or joint econometric models for departure time choices of trips to form daily travel schedule.
- Use of behavioral rules to capture complicated processes of scheduling various activities with different flexibility, the priority of scheduling at specific parts of the day.
- Use of econometric models for the choices of departure times and time allocation to different activity types and then use of behavioral rules to form daily travel schedules.

In activity-scheduling model departure time, choices are often modeled jointly with travel mode choice to capture substitution patterns among a choice of different modes and different time-of-day options for departure choices. Similarly, a joint model of departure time choice and destination location choice can be with or without consideration of alternative mode choices. Such joint models are used to investigate complicated and multidimensional trade-off in daily activity-travel scheduling. Joint model refers to the use of probabilistic mathematical formulations of choice dimension with consideration of correlations among the choices that are modeled.

However, in any departure choice model (either stand-alone model or a part of demand modeling system), travel time to destination is considered as an expenditure to the scheduling process. So, travel time (as well as travel cost) is an exogenous variable to departure time choice model if travel route choice is not jointly modeled. In standard models, route choice is modeled separately. This raises the issue that the choices of departure times define traffic flow and thereby transportation network performances (travel time and cost) in turn influence the choices of departure time alternatives. In standard modeling practice, such cyclical relationship is tackled through iterative feedback between departure time choice and route choice models. However, it is possible to model departure time choices jointly with route choice with or without considering choices of alternative modes but is less common because of computational complexity it poses.

Modeling Techniques

Specific modeling techniques that are used for modeling departure time choices in various contexts and applications depend on whether the time is discretized into alternative time slots to choose from or the continuous nature of time variable needs to be preserved. Time discretization of a 24-hour day can be of:

- Big time chunks: Peak period, off-peak periods, shoulder (between peak and off-peak), etc. Such classification is based on efforts to capture patterns of traffic congestion throughout the day.
- Smaller time chunks: It can be of 15 minutes to a 1-hour time slot. Such classification intends to capture the fine-grained distribution of traffic load onto the network.

Discrete choice models that are used for modeling are multinomial logit model, nested logit model, generalized extreme value model, and mixed logit model (Bhat, 1998a, 1998b). Discrete choice models are based on the assumption of random utility

maximizing choice-making behavior that refers to the fact that a traveler maximizes satisfaction/utility in making a choice and a portion of such utility is random (cannot be explained by systematic variables). The difference among these discrete choice models is on the accommodation of correlation between two or multiple discrete departure time alternatives. Multinomial logit model considered all alternatives are independent, which may not always correct when adjacent time slots can be a positive correlation. Nested logit and generalized extreme value models allow group/nesting of alternatives with the possibility of having correlations in the choice model structure to overcome the limitations of the multinomial logit model. Mixed logit model provides a very flexible model structure to capture all sorts of correlation and substitution patterns in alternative departure time choices.

Considering time as a continuous entity and departure time choices are indirectly modeled through modeling time allocation choice:

- One approach of such a model can be based on the microeconomic theory of time allocation/consumption. The primary purpose of such an approach is to derive the perceived value of time (VOT). Using the assumption of marginal cost factors, VOT can be derived considering the choices of departure time alternatives. It requires a time expenditure function, which needs to be optimized under several time–space–money constraints. The continuous-time model can also be based on the hazard rate of activity duration. Dynamics of time expenditure to an ongoing activity can eventually define the departure time to the trip leading to the next activity. Such models are called hazard model, and the primary purpose of such models is to capture time expenditure dynamics.
- Another approach of continuous time model is based on activity-scheduling optimization under daily time budget (24 hours) constraints. Using Kuhn–Tucker optimality condition and microeconomic theory of the consumption process, such a model can capture various trade-off involved in daily activity scheduling. Such models are known as Multiple Discrete-Continuous Extreme Value (MDCEV) model.

In any of these econometric approaches of modeling departure time, choices can also give a useful economic evaluation measurement, which is the value of travel time savings (VOTTS). VOTTS is the marginal rate of substitution between travel time and money, which can be estimated by taking the ratio of the coefficients travel time and travel cost variables in the discrete or continuous choice model. However, since travel time is attributed to a travel mode, a generalized VOT derived from the departure time choice models needs to be weighted over all modes. Mode-specific departure time choices (modeled as separate departure time choice models for separate modes or a joint mode and departure time choice model) would give mode-specific VOTTS. The VOTTS, in context of departure time choices, is also attributed to the travel time uncertainty leading to early to late arrival compared to the desired/expected arrival time. Many of empirical econometric model primarily address the early arrival, and late arrival issue through the representation of reference-dependent choice behavior (e.g., prospect theory) as travelers response to these two may not be symmetric.

Data for Empirical Modeling

Departure time choice model that is explained in this paper is primarily empirical models and requires observed choice information (data) (Arellana et al., 2012). Departure time choice model estimation requires data that are usually collected through:

- Revealed preference (RP) surveys: Typically, travel surveys conducted among a sample of individuals from the study area. RP travel surveys can be of a specific type of trips, for example, commuting trips, shopping trips, social trips, etc., or general travel diary survey—RP travel diary surveys. Such data provide information on chosen departure times by the respondents in the sample.
- Stated Preference (SP) survey data are used to investigate sensitivities of various policy (pricing, time-of-day restriction, etc.) on the choice of departure time. SP surveys can be designed to capture various aspects of departure time choices including the effects of scheduled delays, congestion pricing, peak-period parking restriction, in-transit crowding (during peak hours) effects, etc. SP surveys use hypothetical scenarios of alternative departure time options with all attributes that differentiate or equate the alternatives. A sample of individuals is required to collect stated responses to those hypothetical scenarios, which provide data on behavioral trade-offs in departure time choices. Also, in recent year, joint RP-off-SP surveys are increasingly being used where the SP survey is built on an RP travel diary. First, RP travel diary is collected, and then the SP scenarios are built based on observed departure time choices in the RP data. Such a survey can much minimize hypothetical biases in SP data and also enrich the RP data through specific SP contexts.
- Mode-specific aggregate origin–destination (OD) level trip table data are often used for jointly modeling departure time choice and traffic assignment model. Such OD data can be collected through household travel surveys, transit smart card data, or other sources, for example, roadside counts, intercept survey, etc. However, these data can allow capturing the distribution of alternative departure time choices than disaggregate modeling of departure time choices.

RP data might not include specific information about the scheduled delay and corresponding effects on the chosen departure time choices unless that information was explicitly collected through the survey. In any case, estimating the parameters of a departure

time choice model through the use of RP data requires imputation of alternative level-of-service information (specifically travel time and travel cost) for non-chosen departure time alternatives. Typically, a calibrated and validated traffic assignment model for the study area is used to generate such information. Of course, the traffic assignment model needs to have the temporal specification that matches the specification of departure time choice alternatives. For example, if the departure time choice model considers peak and off-peak periods as alternatives, the traffic assignment model should be able to generate travel separate time and travel cost for those alternative periods too. However, if the departure time choice model considers hours or even shorter time slots as choice alternatives, the traffic assignment model needs to generate level-of-service variables at that fine grade too. Otherwise, departure time choice alternatives may become quality indifferent (not characterized by differences in level-of-service variables) alternatives. In such case, modeling choices are still possible but would capture more of the market share of departure choices than disaggregate choice trade-offs.

Policy Evaluations

Departure time choice models are used:

- To understand travel behavior of residents that needs to be captured in regional travel demand modeling system for accurately predicting transportation demands in the future or of changes in transportation or land-use system.
- To investigate demand-side policies, to curve traffic congestion, especially peak-period overuse of the overall road network, specific network elements (e.g., bridges, tunnel, etc.) transit service, etc.

Various travel demand management policies of land-use policies require a specific model of departure time choice of urban residences. Frequently, urban parking restrictions (on-street parking) and parking fees are used to facilitate commuting travel. For example, on-street parking restrictions on commuting corridors are often used to facilitate travel during the peak periods of the day. Evaluation of traffic management tools, for example, HOV lane usage and high occupancy tolled (HOT) lane usage, requires models that can capture the switching probability of departure times (with or without travel mode or routes) by the travelers in response to any such policies. Activity-based travel demand models are based on the concept of time-space constraints that refer to the facts that travelers can adjust activity schedules (activity start time and/or trip departure time) in response to transportation network performances. So, departure time choices are explicitly captured in activity-based models.

Conclusions

Departure time choice models are essential policy analysis tools used by transportation planners, traffic engineers, and transportation and land-use policy analyst. The objective of a departure time choice model is to capture how (or whether) travelers change their choice of trip scheduling (trip departure time). Departure time choices of all travel mode users (car user and transit users) have relevance to urban transportation performances as these define the peak-period traffic congestion levels on road networks and peak-period transit crowding on the transit network. Different type of modeling approaches is used for departure time choice. However, the predominantly econometric approach is prevalent. Econometric approaches used behavioral theory to define the choice model specification and observed data to develop an empirical model. Both aggregate and disaggregate choice models are used to investigate departure time behavior. Data from RP travel diary, as well as SP surveys, are used to estimate departure time choice model parameters. Departure time choice model can work as a stand-alone policy analysis tool to investigate specific transportation policies, for example, parking policy, congestion pricing, lane restrictions, transit crowding, etc. However, activity-based travel demands modeling systems explicitly capture departure time choices through activity-scheduling models. Various types of activity-scheduling models exist, for example, as the rule-based approach, econometric modeling approach, and hybrid rule and econometric approach. For the econometric model of departure time choices, modeling techniques vary based on the specification of time to define departure time choices. The discretization of time allows the use of discrete choice models, and continuous time specification requires optimization of time allocation approach. With a wide variety of possible departure time choice models, users of the model require to choose an approach and technique that suits the purpose of using the model as well as data availability for empirical model development.

Acknowledgment

The author acknowledges the support from Percy Edward Hart Professorship awarded by the Faculty of Applied Science and Engineering of the University of Toronto for this work.

Biography

Khandker Nurul Habib is the Percy Edward Hart Professor in Civil & Mineral Engineering at the University of Toronto. His main research focus is “Econometric Models of Transportation Demands and Land Use-Transportation Interactions.” He focuses on transportation and urban planning methodology and qualitative evaluation of planning/policy options. He works on bringing knowledge from various disciplines (psychology, geography, economics, computer science, and system control) to develop appropriate planning methodology and policy analysis tools that can give realistic prediction and forecasting capacity for short-, medium-, and long-term planning exercises.

References

- Arellana, J., Daly, A., Hess, S., Ortúzar, J.D., Rizzi, L.L., 2012. Development of surveys for study of departure time choice: two-stage approach to efficient design. *Transp. Res. Rec. J. Transp. Res. Board No. 2303*. Transportation Research Board of the National Academies, Washington, DC, pp. 9–18.
- Ashiru, O., Polak, J.W., Noland, R.B., 2004. Utility of schedules: theoretical model of departure-time choice and activity time allocation with application to individual activity schedules. *Transp. Res. Rec. J. Transp. Res. Board No. 1894*, TRB, National Research Council, Washington, DC, pp. 84–98.
- Bhat, C.R., 1998a. Analysis of travel mode and departure time choice for urban shopping trips. *Transp. Res. B* 32 (6), 361–371.
- Bhat, C.R., 1998b. Accommodating flexible substitution patterns in multidimensional choice modeling: formulation and application of travel mode and departure time choice. *Transp. Res. B* 32 (7), 455–466.
- Habib, K.M.N., 2013. A joint discrete-continuous model considering budget constraint for the continuous part: application in joint mode and departure time choice modeling. *Transportmetrica* 9 (2), 149–177.
- Habib, K.M.N., 2018. A comprehensive utility-based system of travel options modeling (CUSTOM) considering dynamic time-budget constrained Potential Path Area (PPA) in activity scheduling process: application in modeling worker's daily activity-travel schedules. *Transp. Part A Transp. Sci.* 14 (4), 292–315.

Further Reading

- Arnott, R., de Palma, A., Lindsey, R., 1990. Departure time and route choice for the morning commute. *Transp. Res. B* 24 (31), 209–228.
- Bhat, C.R., Steed, J., 2002. A continuous time departure time choice for urban shopping trips. *Transp. Res. Part B* 36, 207–224.
- de Palma, A., Lefèvre, C., 2019. Bottleneck models and departure time problems. In: Briassoulis, H., et al. (Eds.), *The Practice of Spatial Analysis*. Springer International Publishing AG, part of Springer Nature 2019, Cham, pp. 161–165.
- Guo, R.-Y., Yang, H., Huang, H.-J., 2018. Are we really solving the dynamic traffic equilibrium problem with a departure time choice? *Transp. Sci.* 52 (3), 603–620.
- Ozbay, K., Yanmaz-Tuzel, O., 2008. Valuation of travel time and departure time choice in the presence of time-of-day pricing. *Transp. Res. Part A* 42, 577–590.
- Sasic, A., Habib, K.M.N., 2013. A heteroskedastic GEV model with cross-nested/overlapping choice set: application in modeling commuting departure time choices in GTHA. *Transp. Res. Part A* 50, 15–32.
- Steed, J., Bhat, C.R., 2000. On modeling departure time choice for home-based social/recreational and shopping trips. *Transp. Res. Rec.* 1706. Paper No. 00-0706.
- Sumi, T., Matsumoto, Y., Miyaki, Y., 1990. Departure time and route choice model of commuters on mass transit systems. *Transp. Res. B* 24 (4), 247–262.
- Timmermans, H., Ettema, D., 2003. Modeling departure time choice in the context of activity scheduling behavior. *Transp. Res. Rec.* 1831, Paper No. 03-2425.

Traveler Responses to Congestion

David T. Ory, Gayathri Shivaraman, WSP, New York, NY, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	509
Inventory of Behavioral Responses and Traveler Implications	509
Changes in Route	509
Changes in Departure Time/Scheduling	510
Changes in Itineraries/Stop-Making	510
Change in Temporary Workplace Location	511
Changes in Travel Mode	511
Vehicle/Equipment Upgrade	511
Within-Region Residential or Employment Location Change	512
Outside-Region Residential and Employment Location Change	512
Understanding and Ability to Predict	512
Conclusions	513
References	513
Further Reading	513

Introduction

A transportation system is a supply–demand framework seeking equilibrium, wherein transportation services and infrastructure serve as the supply side and individual travel needs and desires serve as the demand side. Degradation on either side of the equilibrium results in imbalance leading to undesirable effects on the system as a whole. Congestion caused by, for example, morning commute travel demand exceeding roadway capacity is the most perceptible example of such undesirable effects of equilibrium imbalance in transportation systems (Vickrey, 1969). Any event interpreted as undesirable or cumbersome on personal resources motivates change. Change as such is characteristic of a response-mitigation measure against the undesirable effect(s) caused by the imbalance. This text explores individual level responses to congestion categorized into tactical and strategic measures and their culminating effects on the urban form.

Let us define “transportation supply” as the infrastructure and services that seek to satisfy travel demand. Let us define “congestion” as the degradation, due to crowding, of any transportation supply element. Congestion can therefore refer, as it most often does, to the reduction in travel speeds of a roadway due to vehicle demand exceeding available roadway supply. But it can also refer to a public transportation vehicle that no longer accommodates every passenger that would like to ride or to sit. Or a dedicated bicycle path in which travelers must moderate their speed due to the presence of other cyclists.

Congestion is a key driver of changes in traveler behavior, both tactical and strategic. For the purposes of this entry, let us define tactical changes in travel behavior as decisions that may vary day to day and have a relatively low switching cost whereas strategic changes have high switching costs. In aggregate, individual responses to congestion have profound impacts on urban form and land consumption. In fact, responses to congestion are arguably the most important influence on urban settlement patterns over the prior 80 years.

An inventory of potential tactical and strategic changes are introduced and defined in the next section, paired with a discussion of the implications of the response, both to individuals and society. Thereafter, a section discusses the degree to which the relationship between congestion and each behavior is well understood and readily predicted. A concluding section completes the entry.

Inventory of Behavioral Responses and Traveler Implications

Table 1 introduces potential behavioral responses to an increase in congestion and maps these responses on a qualitative scale of tactical to strategic (National Academies of Sciences, Engineering, and Medicine, 1994; Deakin, 1994).

An explication of each of the earlier responses is provided later, along with commentary regarding the impact of the change to an individual and the collective impact on society.

Changes in Route

A roadway traveler faced with steadily increasing congestion on a freeway may experiment with alternate routes. Similarly, a transit commuter faced with a rail or bus service that is uncomfortably crowded and/or difficult to board due to crowding may seek an

Table 1 Introduction and mapping of traveler responses

Item	Response	Tactical → strategic
1.	Changes in route	
2.	Changes in departure time/scheduling	
3.	Changes in itineraries/stop-making	
4.	Change in temporary workplace location, that is, satellite office or working from home	
5.	Changes in travel mode	
6.	Vehicle upgrade	
7.	Within-region residential or employment location change	
8.	Outside-region residential and employment location change	

alternate route. The switching cost here is typically low, hence the mapping in **Table 1** of this behavior to the far end of the tactical side of the chart.

The implications to individuals of switching routes are minor, for example, traveling on local roads rather than freeways and having to endure more frequent encounters with intersection controls. With modern navigation applications on smartphones assisting travelers in finding lower cost routes, any gain from changing routes is likely to be short-lived, as other travelers are likely switching as well, shifting congestion accordingly.

The collective implications to a society of individuals switching routes are broader and generally negative. Roadways that are designed to accommodate small volumes of vehicles traveling at low speeds are not well suited for daily commuters seeking to move faster than, say, an adjacent congested freeway. Formerly quiet neighborhood streets may become less quiet and less safe for children traveling to school, or myriad other human activities that are incompatible with fast moving vehicular movements. Communities are therefore motivated to accommodate peak travel conditions on limited access freeways or other high-capacity facilities. This desire to accommodate peak travel, in turn, results in the construction of roadway facilities designed to try and meet peak demand, which, in turn, consumes land and shapes urban form—often expanding the footprint of urban settlements, separating neighborhoods with difficult-to-cross infrastructure, forming new communities, etc.

Changes in Departure Time/Scheduling

Congestion tends to occur during the morning and evening commute periods due to the convenience of a majority of work taking place during a common period of time—preferably when the sun is out. Travel demand, therefore, frequently exceeds available supply during these time periods; travelers with the ability to do so can travel outside of commute hours to avoid or mitigate the impacts of congested conditions.

Travelers with this flexibility often gain time after shifting their commute to earlier (or later) in the morning inbound and earlier (or later) in the evening outbound. Public transportation commuters shifting time can benefit from more available and/or convenient parking at suburban park and ride stations, as well as a place to sit on crowded transit lines.

The collective spreading of travel through the day is generally seen as a worthy public policy goal, with numerous governments implementing surcharges (often called “congestion pricing”) to cross bridges or board transit vehicles or enter a central business district during commute hours. A more even distribution of demand through the day makes it easier for transport authorities to manage their systems. Congestion pricing does have its critics, as for those without the flexibility at home or at work to change the time during which they travel, a congestion fee is an unavoidable tax on travel. This tax pushes against the utility toward common travel during the commute periods, as evidenced by its popularity—which motivated the congestion price in the first place.

As the congested travel period continues to expand, shifts farther and farther from common work hours begin to have disruptive and negative impacts on travelers’ lives. These can include:

- Reduction in time spent with family;
- Reduction in exposure to daylight; and
- Reduction in exposure to colleagues.

Changes in Itineraries/Stop-Making

After dropping the kids at school, a traveler may have just enough time to read the paper and have a coffee at a nearby coffee shop prior to traveling to work. An increase in congestion may motivate the elimination of the stop at the coffee shop. Similarly, a commuter may choose, if given the option, a so-called 9/80 schedule, in which, over a 2-week period, 8 days are spent working 9 hours per day, a 9th working 8 hours, and a 10th, typically every other Friday, taken off. As congestion increases, itineraries may change, with more time devoted to travel and less time devoted to activities.

As with the changes in departure time/scheduling, government actions that change itineraries are generally seen as a good public policy. Ample research exists on the benefits of motivating trip chaining behavior, where travelers are nudged to, for example, stop at

the grocery store on the way home from work rather than return home and then make a separate trip to the grocery store. In a vehicle fleet dominated by internal combustion engines, such trip chaining can result in reduced vehicle emissions, by reducing travel as well as the so-called “cold” engine starts. These collective benefits do come at a negative cost to individuals. Returning home prior to grocery shopping may allow a commuter to take the dog for a quick walk, use the restroom, or check on the garden. There is a reason a nudge is often needed for a trip chain to form: doing so is, for one reason or another, not ideal.

Change in Temporary Workplace Location

As congestion makes commuting increasingly onerous, travelers may opt to work 1 or 2 days a week in another location. This location may include home in a telecommuting arrangement, and it may include a satellite office. Many firms are now keeping small, headquarter addresses in high-status locations such as Manhattan or San Francisco, and moving the majority of operations to lower cost, often suburban, locations. These suburban locations provide less expensive real estate for the firm and—often—shorter commutes for workers. For workers that must spend most of their time in a city center location, satellite offices may offer a temporary respite from congestion.

The collective implications of employers moving to the suburbs—either to form satellite offices or to abandon the central city altogether—are profound. As employers move to the suburbs, households follow, allowing for the previously idealized situation of both living and working in a suburban setting. Working and living in the suburbs was the original promise of Silicon Valley. Once located in suburban locations, employment centers motivate the development of residential suburbs a bit farther away from the city center. When faced with commuting to two separate suburban office locations, two worker families may choose to live in yet another suburban community between the respective employment sites, motivating the creation of yet another residential suburban community. Many cities that came of age in the 1960s—Atlanta, Phoenix, Silicon Valley, Houston—reveal this development pattern. While such satellite workplace arrangements have the ability to relieve congestion from city centers and create jobs in suburban jurisdictions, they tend to redirect significant traffic toward the outlying suburban thoroughfares incapable of handling such congestion (Cervero, 1989). Between-suburb commuting is difficult to serve efficiently with public transportation, hence the need for very large investments in roadway infrastructure in these types of places over the last 50 years.

Changes in Travel Mode

As congestion makes traveling by automobile less convenient, public transportation often becomes relatively more convenient. In large, dense urban areas, such as New York City, traveling by public transport is often faster than driving and, in fact, is often the reason cities like New York have become as dense and prosperous as they are.

Increasing the encumbrance of automobile travel in order to motivate travelers to select alternate travel modes, including public transport, is generally viewed as a worthy public policy goal, as these modes generally have a smaller environmental impact than traveling by automobile. Individually, switching to a less crowded mode is generally undesirable, as there is a reason the other mode was chosen by the traveler as well as the numerous other travelers that make the switched-from mode crowded in the first place. Travelers who struggle to walk, climb stairs, or find confinement in shared spaces unpleasant face significant challenges when nudged away from automobiles, as do those traveling with young children. Overall, the benefits of high-quality infrastructure aimed at mitigating the impact of congested roadways can be seen in our densest cities, which provide high levels of accessibility to large numbers of individuals. Over time, these dense areas must continue to invest in nonautomobile infrastructure to avoid crowding on public transit routes—particularly subways—as crowding on desirable transit services shifts travelers to even less desirable modes (for most), including bicycles and buses.

Vehicle/Equipment Upgrade

When faced with an automobile commute that continues to increase as congestion increases, many travelers may mitigate the impacts of spending a lot of time in a vehicle by purchasing a new vehicle. For the past few decades, common amenity upgrades included automatic transmissions, climate control, and entertainment/communication systems. In decades to come, these amenities may include electric motors and automation. A present question of keen interest to travel behavior researchers is how travelers will value time spent in an automated vehicle. On the one hand, many travelers have the option of riding—rather than driving—in a vehicle today and express little enthusiasm in doing so, for example, it is not common to hear what a great time a friend/colleague had riding in a car on the way to work. On the other hand, automation will allow travelers to be alone in their vehicles and this isolation, paired with a freedom to move around, may make all the difference, as the ability to sleep, work, eat, exercise, etc., without the social pressures of a fellow traveler sitting 2 feet away or the restrictions of sitting in a crash-ready position, may be a highly tolerable-to-enjoyable experience.

Collectively, vehicle upgrades have mixed impacts on society. On the positive side of the ledger, new vehicles are generally safer and pollute less than older vehicles. On the negative side, automobile travel is still generally harmful to the environment and urban form.

In the same category as a vehicle upgrade is any other manner of equipment that may make commuting on any mode of travel more tolerable, including an electric bicycle, audiobook subscription, or high-quality headphones.

Within-Region Residential or Employment Location Change

When travel becomes unpleasant, a more dramatic and strategic change is to move—changing your home or work location. As noted in the temporary workplace section earlier, changing jobs to a firm with an office location closer to your home, particularly if your home is in a suburban location, has profound impacts on urban form and may, in fact, be the most important driver of changes in urban form that we have experienced over the last 60 years.

Changes in residential location also come with high transaction costs. The urbanization of formerly industrial areas of dense cities has mitigated the impacts of the suburbanization of employment. Consider, for example, the large number of condominiums now present in the South of Market neighborhood in San Francisco, the Pearl District neighborhood in Portland Oregon, and the Hudson Yards neighborhood in New York City.

The individual impact of changing jobs or homes can be significant, as it has the potential to disrupt a positive family dynamic often composed of the school and work locations of other household members. Further, moving to a work location away from a dense urban job center has the potential to reduce career earnings, as commute convenience is favored over proximity to other potential employers. The change can also be positive, as inertia may be responsible for families to remain in suboptimal housing situations after, for example, children leave home.

The collective impacts of changing residential or employment locations are generally negative, as these responses are a signal that the community has failed to provide sufficient transport services. Abundant inefficiencies result, including the formation of new cities, with the commensurate administrative overhead, the construction of costly and underutilized suburban freeways, increased self-segregation by socioeconomic class, and/or consumption of open-space/natural habitats.

Outside-Region Residential and Employment Location Change

A logical escalation of within-region residential or employment change is an outside-region residential and employment change. As noted earlier, an important motivation of the large increase in population of previously low population areas such as Phoenix, Atlanta, and Dallas is the ability to travel in an automobile with low levels of congestion (affordable housing and air conditioning's ability to make a warm climate pleasant are also key drivers). Had coastal cities such as Boston, Los Angeles, Washington, and San Francisco (1) invested in the types of public transport system that allows cities such as New York to grow larger and more dense, and (2) allowed the tall buildings necessary to accommodate a growing city, the populations of cities with less desirable amenities would, all else equal, be lower—potentially much lower. One could argue that the existence of all large, metropolitan areas in the Southwest, South, and Southeast regions of the United States is due in part by the desire of individuals to travel in an automobile in uncongested conditions.

The individual impact of relocating to a less congested area is generally negative, as the region with less congestion is logically of lower amenity than the region of higher congestion. Further, moving to a lower population area may result in few employment opportunities.

The collective impact of coastal cities failing to accommodate larger populations is difficult to understate—estimates have placed the cost at around \$1 trillion annually (Hsieh and Moretti, 2019). Transportation plays a key role in the decision not to get larger: opponents often cite increases in traffic or difficulty in parking as a reason to oppose development.

Understanding and Ability to Predict

Sound public policy can mitigate the impacts of congestion on individuals and society if the relevant phenomena are well understood and readily predicted. Table 2 summarizes the same behavioral responses identified in Table 1 and discussed earlier. The symbols now suggest the difficulty in making predictions in response to congestion.

The field of travel demand forecasting has embraced the challenge of predicting responses to congestion for the past several decades, motivated by a mandate from the United States Federal Government to comment on the likelihood that transport

Table 2 Introduction and mapping of traveler responses

Item	Response	Hard to predict → easy to predict
1.	Changes in route	□□□□□□□□□□
2.	Changes in departure time/scheduling	□□□□□□□□□□
3.	Changes in itineraries/stop-making	□□□□□□□□□□
4.	Change in temporary workplace location, that is, satellite office or working from home	□□□□□□□□□□
5.	Changes in travel mode	□□□□□□□□□□
6.	Vehicle upgrade	□□□□□□□□□□
7.	Within-region residential or employment location change	□□□□□□□□□□
8.	Outside-region residential and employment location change	□□□□□□□□□□

infrastructure and policies will reduce vehicle emissions. Travel modeling systems built for regional governments have long focused on making accurate predictions in regards to changes in route (item 1 in Table 2) and mode choice (item 5). The other tactical responses to congestion are more nuanced and difficult to predict. While changes in departure time/scheduling (item 2) and changes in itineraries (item 3) are easy to predict in the aggregate, simulating individual responses is a challenge, as individual preferences, employment considerations, and family considerations are difficult to observe. Understanding of such constraints has led to a shift in travel analysis and policy research from system management to demand management, with increased interest in ridesharing incentives, congestion pricing, and employer-based demand management schemes (Pinjari and Bhat, 2011). One such advance that has seen significant increase in transportation research and development interests is the prospects of autonomous vehicles. Research findings indicate that autonomous vehicles are expected to not only mitigate congestion but also improve vehicle safety and fuel efficiency, with an expected cost reduction ranging from USD\$2000 to USD\$4000 USD per vehicle (Bagloee et al., 2016; Fagnant and Kockelman, 2015; Lamotte et al., 2017). While autonomous vehicles may emerge as an ideal solution to urban congestion, their implementation, interaction with existing systems, mass-market penetration, and legal liabilities remain highly uncertain.

The field of land-use modeling deals with the more strategic aspects of congestion responses, including changing residential or employment location (item 7) and migration (item 8). Research suggests a strong correlation between household-level automobile ownership and residential density (Pushkarev and Zupan, 1977; Mogridge, 1997; Kitamura and Mokhtarian, 1997). Growing advocacy among land-use planners to promote residential density and mixed use developments stem from findings that travel patterns are derived from individual needs and demands for activities and accessing services (Cervero, 1995; Salomon and Mokhtarian, 1998). In other words, the change in spatial distribution of such attraction hubs within close proximity to residential locations could result in reduced travel time, congestion, and ultimately reducing or eliminating the need to consider a change in residential and/or employment location or migration.

Conclusions

This entry identifies and discusses the implications of travel congestion on traveler behavior and, in turn, urban form and other urban phenomena. Congestion is a significant driver of urban form at the local, regional, state, and national level. A brief section then comments on our ability to accurately predict these responses.

References

- Bagloee, S.A., Tavana, M., Asadi, M., Oliver, T., 2016. Autonomous vehicles: challenges, opportunities, and future implications for transportation policies. *J. Modern Transp.* 24 (4), 284–303, doi:10.1007/s40534-016-0117-3.
- Cervero, R., 1989. Suburban employment centers: probing the influence of site features on the journey-to-work. *J. Plan. Edu. Res.* 8 (2), 75–85.
- Cervero, R., 1995. Planned communities, self-containment and commuting: a cross-national perspective. *Urban Stud.* 32 (7), 1135–1161. Available from: <https://journals.sagepub.com/doi/10.1080/00420989550012618>.
- Deakin, E., 1994. Urban transportation congestion pricing: effects on urban form. *Curbing Gridlock: Peak-Period Fees to Relieve Traffic Congestion—Special Report 242*, Vol. 2. The National Academies Press, Washington, DC, pp. 334–355.
- Fagnant, D.J., Kockelman, K., 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transp. Res. Part A Policy Pract.* 77, 167–181, doi:10.1016/j.tra.2015.04.003.
- Hsieh, C.-T., Moretti, E., 2019. Housing constraints and spatial misallocation. *Am. Econ. J. Macroecon.* 11 (2), 1–39.
- Kitamura, R., Mokhtarian, P.L., 1997. A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. *Transportation* 24, 125–158, doi:10.1023/A:1017959825565.
- Lamotte, R., de Palma, A., Geroliminis, N., 2017. On the use of reservation-based autonomous vehicles for demand management. *Transp. Res. Part B Methodol.* 99, 205–227, doi:10.1016/j.trb.2017.01.003.
- Mogridge, M.J.H., 1997. The self-defeating nature of urban road capacity policy: a review of theories, disputes and available evidence. *Transp. Policy* 4 (1), 5–23. Available from: <https://www.sciencedirect.com/science/article/abs/pii/S0967070X96000303?via%3Dihub>.
- National Academies of Sciences, Engineering, and Medicine, 1994. *Curbing Gridlock: Peak-Period Fees to Relieve Traffic Congestion—Special Report 242*. The National Academies Press, Washington, DC.
- Pinjari, A.R., Bhat, C.R., 2011. Activity-based travel demand analysis. *Handbook Transp. Econ.* 10, 213–248.
- Pushkarev, B., Zupan, J.M., 1977. *Public Transportation and Land Use*. Regional Plan Association, New York, NY.
- Salomon, I., Mokhtarian, P.L., 1998. What happens when mobility-inclined market segments face accessibility-enhancing policies? *Transp. Res. Part D Transp. Environ.* 3 (3), 129–140, doi:10.1016/S1361-9209(97)00038-2.
- Vickrey, W., 1969. Congestion theory and transport investment. *Am. Econ. Rev.* 59 (2), 251–260. Available from: www.jstor.org/stable/1823678.

Further Reading

- Arnott, R., Small, K., 2014. The economics of traffic congestion. *Am. Sci.* 82 (5), 446–455. Available from: <https://about.jstor.org/terms>.
- Cervero, R., 2018. America's suburban centers: the land use-transportation link. In: *America's Suburban Centers: The Land Use-Transportation Link*, doi:10.4324/9781351048040.
- Cervero, R., 2003. Coping with complexity in America's urban transport sector. 2nd International Conference on the Future of Urban Transport, September 22–24, pp. 1–18. Available from: <https://escholarship.org/uc/item/4wf4n16r>.
- Hennessy, D.A., Wiesensthal, D.L., 1997. The relationship between traffic congestion, driver stress and direct versus indirect coping behaviours. *Ergonomics* 40 (3), 348–361, doi:10.1080/001401397188198.

- Hennessy, D.A., Wiesenthal, D.L., 1999. Traffic congestion, driver stress, and driver aggression. *Aggress. Behav.* 25 (6), 409–423, 10.1002/(SICI)1098-2337(1999)25:6<409::AID-AB2>3.0.CO;2-0.
- Khattak, A.J., De Palma, A., 1997. The impact of adverse weather conditions on the propensity to change travel decisions: a survey of Brussels commuters. *Transp. Res. Part A Policy Pract.* 31 (3), 181–203, doi:10.1016/S0965-8564(96)00025-0.
- Lesteven, G., 2016. Behavioral responses to traffic congestion—findings from Paris, São Paulo and Mumbai. In: *Traffic Management*, pp. 121–138, doi:10.1002/9781119307822.ch9.
- Salomon, I., Mokhtarian, P.L., 1997. Coping with congestion: understanding the gap between policy assumptions and behavior. *Transp. Res. Part D Transp. Environ.* 2 (2), 107–123, doi:10.1016/S1361-9209(97)00003-5.
- Stokols, D., Novaco, R.W., Stokols, J., Campbell, J., 1978. Traffic congestion, type A behavior, and stress. *J. Appl. Psychol.* 63 (4), 467–480, doi:10.1037/0021-9010.63.4.467.

Origin–Destination Demand Estimation Models

William H.K. Lam*, **Hu Shao[†]**, **Shuhan Cao[†]**, **Hai Yang^{*}**, ^{*}Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China; [†]School of Mathematics, China University of Mining and Technology, Xuzhou, China; ^{*}Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

© 2021 Elsevier Ltd. All rights reserved.

Survey-Based Models	515
Traffic-Count-Data-Based Models	516
Static OD Estimation Model	516
Dynamic OD Estimation Model	517
Multiple Sources of Data-Based Model	517
Multiple Sources of Traffic Data	517
OD Estimation Models	517
Remarks	518
References	518
Further Reading	518

Origin–destination (OD) demand, which is also referred to as OD travel or traffic demand, corresponds to the traffic flow (in terms of total number of travelers or vehicles) between two traffic zones (namely origin and destination zones) during a study period. In a spatial manner, the OD matrix consists of defining a two-dimensional table whose rows and columns represent the traffic zones, in which the entry of the OD matrix is the OD demand between the corresponding two traffic zones. OD demand is regarded as the essential data required for transportation planning and traffic management. Therefore, researchers and practitioners have continued to explore new models for OD demand estimation.

In the middle of 20th century with launch of the computers, some developed countries in Europe and America began to investigate the traffic demand forecasting models in order to predict the needs of large-scale transportation infrastructure projects for urban development. By the 1970s, a representative “four-stages” urban traffic demand forecasting model had been proposed. The four-stages model divides the process of traffic demand forecasting into four stages, namely (1) trip generation, (2) trip distribution, (3) modal split, and (4) trip assignment. OD estimation, corresponding to the trip distribution stage, has attracted an increasing attention since then.

In reality, the complete information of actual OD demand cannot be accessed easily. In the past decades, researchers have proposed various models to estimate the OD demand on the basis of different approaches. These existing models can be classified by several criteria later. In the temporal manner, the OD estimation models can be classified as static and dynamic models. Considering the network uncertainty issue, the OD estimation model can be categorized as deterministic and stochastic OD estimation models. Regardless of different criteria for grouping the OD estimation models, one common feature of OD demand estimation model is to use the observed data, which contain the partial information of OD demand, in order to estimate the complete OD demand matrix under different scenarios. In view of this, this chapter addresses the up-to-date existing OD estimation models with respect to different sources of observed data used for various models.

Fig. 1 shows classification of these available OD estimation models with different sources of data and modeling approaches. With the fast-growing development of communication technologies, more and more sources of relevant data are available for OD demand estimation. At the early stage for development of the traditional top-down approach for OD estimation, survey-based models were used in practice in spite of its limited sample size and the vast amount of time and monetary consuming for collection of sampled OD data. Later, benefiting from the traffic flow data collected by traffic counting sensors (e.g., loop detector), a large body of bottom-up models for OD estimation from partial traffic counts were proposed and widely used by the researchers and practitioners. Nowadays, owing to the progress of Big Data science, multiple sources of data are now available for estimation of the OD demand matrix. A brief account of these existing OD estimation models is given later for review.

Survey-Based Models

Intuitively, the OD demand information can be obtained by the top-down approach with use of the survey-based methods and limited sample size, such as home-based survey, roadside interview, and license plate recognition technique. Home-based surveys together with census tracts have been empirical foundation for traffic demand forecasting models to capture many important demographic and travel characteristics of the households. For example, household interview surveys that have higher response rates and larger sample sizes can provide us reliable information on trip generation and distribution by traffic zones. With these sampled information, it follows with estimation of the complete OD trip matrix between traffic zones. These methods can yield accurate results when sufficient sampled households have been selected systematically for collection of the relevant travel

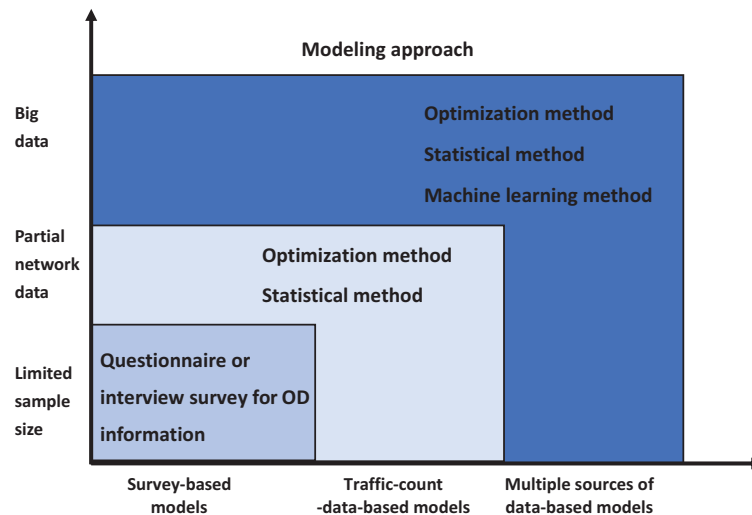


Figure 1 Classification of OD estimation models.

characteristics information. However, it is often costly and requires a large amount of resources for implementation of these household interview surveys.

On the contrary, roadside interviews have also been used to collect travel data on people passing through a specific location (OD station). By locating a series of these stations around the perimeter of the study area, it is possible to measure the travel into and/or through the selected area. With this survey method, drivers are stopped on the road and questioned as to their origin, destination, trip purpose, and associated travel characteristics data. However, this network-wide survey-based method is difficult for larger study areas and/or impossible to undertake frequently, notwithstanding the advancement in traffic survey technologies such as license plate recognition techniques.

Traffic-Count-Data-Based Models

Static OD Estimation Model

In view of the lower cost of the bottom-up approach, traffic flows collected by traffic counting sensors (e.g., loop detectors) have been widely used for OD demand estimation. The traffic-count data motivate transportation researchers to develop advanced models for OD estimation, which are well received by transportation practitioners. The key idea of OD demand estimation model based on traffic-count data is to estimate the OD demand so as to reproduce the resultant link flows which are consistent with the observed link flows at selected sensor locations. Most of these traffic-count-data-based models early developed are concentrated on the static OD demand estimation using optimization and/or statistical methods with partial network data (traffic counts on partial of the network links).

Optimization method: The OD estimation models using optimization method aim to minimize the difference between observed and estimated traffic flows. Such difference is usually formulated as the objective function. For example, suppose that the OD demand is the deterministic variable to be estimated, an optimization model can be proposed that yields the most likely OD matrix using an entropy-type objective function, in which the estimated link flows are consistent with measurements of traffic flows on selected links installed with traffic counting sensors. With consideration of the network congestion effects, [Yang et al. \(1992\)](#) extended Bell's model ([Bell, 1991](#)) and proposed a bi-level programming model for estimation of the OD demand from traffic counts. The upper-level problem is to minimize an entropy/or generalized least squares objective function, and the lower level is a traffic assignment problem to model the travelers' path choice behavior under congestion. In order to assess the performance of these bottom-up models, the quality of the OD estimation can be evaluated using different measures, such as the maximum possible relative error for OD demand estimates which is the relative deviation of the estimated OD demands from the true ones.

Statistical method: In view of the statistical properties of the traffic count data, some researchers adopted statistic methods to address the OD estimation problem. For example, the likelihood function is used in the maximum likelihood estimation method based on some probability distribution, such as Poisson distribution, to estimate the OD demand matrix. Some issues are considered in these statistical models, such as the presence of measurement and sampling errors in the observed link flows. The Bayesian estimator for OD estimation is another kind of statistical model. [Hazelton \(2000\)](#) considered the problem of estimating the OD traffic intensities with the given link flow data from an uncongested network over a number of time periods. The variation of route choice proportions could be captured with use of Poisson distributed OD flows. In view of the network uncertainties due to congestion, the mean and covariance of OD demand in a spatial manner can be simultaneously estimated using the least squares method or Lasso method.

Dynamic OD Estimation Model

Based on various static OD estimation models from traffic counts, some researchers investigated the dynamic OD demand estimation problem. Different from the static OD estimation model, the temporal information of the OD traffic flows is also considered in these dynamic models. For example, dynamic path flow estimators using time-varying traffic counts were developed to estimate the time-varying OD flows by path within the study networks. If more information, besides the traffic counts, can be obtained, the dynamic OD demand matrices can be estimated using different ways of the static models. For example, a path flow-based optimization model, which incorporates heterogeneous sources of traffic measurements and does not require explicit dynamic link-path incidences, was proposed to estimate the dynamic OD demand. Alternatively, Zhou and Mahmassani (2006) proposed a dynamic OD estimation model by extracting valuable point-to-point split-fraction information from automatic vehicle identification (AVI) counts without estimating market-penetration rates and identification rates of AVI tags.

Multiple Sources of Data-Based Model

Although various OD estimation models have been proposed in the literature, it is well known that a common difficulty of the existing OD demand estimation models is the solution identifiability problem, that is, it is impossible to uniquely identify the optimal OD demand matrix as the number of observed links (from which the traffic counts are available) is usually less than the number of elements to be estimated in the OD demand matrix. It is generally believed that the more information provided, the more accurate is the OD estimates. Thus, more and more sources of traffic data can be used for improving the OD estimation accuracy.

Multiple Sources of Traffic Data

In the age of Big Data, the AVI technologies, including license-plate-based AVI systems and cellular-phone-based AVI systems, have been studied extensively. The license-plate-based AVI system collects the vehicle license information and the time stamps that the vehicles pass by the detection point. Although the corresponding AVI data are more convenient to derive the OD demand matrix, the network-wide data collection is difficult due to low coverage of the license-plate-based AVI sensors. Currently, cellular phones have been widely used by travelers. Thus, the cellular-phone-based AVI data may be more convenient to collect for estimation of the OD demand matrix.

Global positioning system (GPS) can collect data for multiple-day trips, including departure and arrival times at origins and destinations, respectively, which directly records the OD demands by time of day. The aggregated GPS trajectory data are regarded as low-cost survey data, which can be used to reproduce the OD trip flows in both the spatial and temporal dimensions. However, attention should be given on two important issues when using GPS data, that is, the high measurement errors and low market penetration rates.

Radio frequency identification (RFID) is an emerging technology to collect AVI data. The RFID is a technology deployed in some cities to monitor traffic through detecting RFID tag-equipped vehicles passing by RFID detectors installed along the major road links. Information from RFID is more accurate, real time, and less influenced by the external environment as compared with the license-plate-based or the cellular-phone-based AVI sensors. The RFID data were utilized to derive the sampled OD demand matrix with use of the dynamic traffic assignment by matching the RFID data detected at origins and destinations along the selected road links.

OD Estimation Models

Methodologically, there are three types of the bottom-up models for OD demand estimation with use of multiple sources of traffic data. Apart from statistical and optimization methods, machine learning method can also be utilized to solve the OD estimation problem in view of the advanced development of artificial intelligence techniques.

Statistical method: This approach is based on the knowledge and methods of statistics. Researchers adopted statistical methods to dig out more information for estimation of the stochastic OD demand. For example, the particle filter has been used to estimate the probability of a vehicle's trajectory from all possible candidate trajectories using the license-plate-based AVI data on a large-scale urban road network. The destination choice model with pairwise district-level constants was proposed using the Quasi-Newton maximum likelihood estimates or the maximum likelihood estimates. On the contrary, the hierarchical Bayesian model was capable of estimating passenger-based OD flow matrices and the period-level OD flow matrix using the sampled OD flow data collected by cellular-phone-based AVI sensors and boarding data provided by passenger fare boxes. The iterative generalized least squares method has been proposed to estimate the mean and covariance matrix of OD demand with taking account the day-to-day variation induced by travelers' independent route choices using day-to-day travel time/speed data.

Optimization method: Similar to the earlier traffic-count-data-based models, the optimization method is still widely adopted for OD estimation with multiple sources of data. Researchers proposed various data-driven objective functions based on multiple sources of data in order to obtain the static and dynamic OD demands. Different from the traffic-count-data-based models, these models focus on the observed path (or) path segment flows and travel times, which provide more information for the OD demand estimation. The corresponding methods include least square method, itinerary-based nonlinear optimization model, sequential

updating method based on the maximum entropy principle, four-location models, heuristic expectation maximization method, bi-level programming model, etc.

Machine learning method: It was found in some recently developed OD estimation models that machine learning models were better at capturing the nonlinear relationships of available data as compared with the statistical models. The machine learning method is also less restrictive in terms of the assumptions regarding the error structures, making them arguably more appropriate for this application. The machine learning methods used in OD estimation includes K -nearest neighbors regression, support vector regression, neural networks: multilayer perceptron, ordinary least square regression, lasso regression, ridge regression, random forest regression, points of interest and K -means clustering method, etc. It was recently reported in the literature (Krishnakumari et al., 2019) that the multilayer perceptron, a neural networks approach, would lead to an outperforming model for the OD demand estimation purpose.

Remarks

This chapter discusses the comprehensive scope of OD estimation issues in order to provide a better understanding of the model development with different sources of traffic data and different modeling approaches. Although we have classified the existing models into different categories, some models may simultaneously belong to more than one category. For example, the maximum likelihood model utilizes both the statistical and optimization methods. In this chapter, we hope that the up-to-date review of the OD estimation models would be helpful for inspiring and stimulating new research initiatives to propose robust OD demand estimation models by adapting or developing innovative modeling approaches in the arena of Big Data.

Nowadays, due to the fast-growing development of traffic data collection technologies, OD estimation is increasingly convenient and more accurate as more traffic data can be available and used for OD estimation. In the age of Big Data, OD estimation studies will come into a new avenue with advanced and efficient technologies using multiple sources of different data (e.g., weather data) that are not restricted to the traffic-related data only.

The work described in this chapter was financially supported by grants from National Natural Science Foundation of China (Project Nos. 71671184 and 71271205), together with grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 152628/16E) and from the Research Committee of the Hong Kong Polytechnic University (Project No. 4-ZZFY).

References

- Bell, M., 1991. The estimation of origin-destination matrices by constrained generalized least-squares. *Transp. Res. Part B* 25 (1), 13–22.
 Hazelton, M.L., 2000. Estimation of origin-destination matrices from link flows on uncongested networks. *Transp. Res. Part B* 34 (7), 549–566.
 Krishnakumari, P., Lint, H.V., Djukic, T., Cats, O., 2019. A data driven method for OD matrix estimation. *Transp. Res. Part C*, doi:10.1016/j.trc.2019.05.014.
 Yang, H., Sasaki, T., Iida, Y., Asakura, Y., 1992. Estimation of origin-destination matrices from link traffic counts on congested networks. *Transp. Res. Part B* 26 (6), 417–434.
 Zhou, X.S., Mahmassani, H.S., 2006. Dynamic origin-destination demand estimation using automatic vehicle identification data. *IEEE Trans. Intell. Transp. Syst.* 7 (1), 105–114.

Further Reading

- Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transp. Res. Part C* 58, 240–250.
 Asakura, Y., Hato, E., Kashiwadani, M., 2000. Origin-destination matrices estimation model using automatic vehicle identification data and its application to the Han-Shin expressway network. *Transportation* 27 (4), 419–438.
 Cascetta, E., Inaudi, D., Marquis, G., 1993. Dynamic estimators of origin-destination matrices using traffic counts. *Transp. Sci.* 27 (4), 363–373.
 Castillo, E., Maria Menendez, J., Sanchez-Cambronero, S., 2008. Traffic estimation and optimal counting location without path enumeration using Bayesian networks. *Comp. Aided Civil Infrastr. Eng.* 23 (3), 189–207.
 Ma, W., Qian, Z., 2018. Estimating multi-year 24/7 origin-destination demand using high-granular multi-source traffic data. *Transp. Res. Part C* 96, 96–121.
 Rao, W., Wu, Y.J., Xia, J., Ou, J., Kluger, R., 2018. Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. *Transp. Res. Part C* 95, 29–46.
 Shao, H., Lam, W.H.K., Sumalee, A., Chen, A., Hazelton, M.L., 2014. Estimation of mean and covariance of peak hour origin–destination demands from day-to-day traffic counts. *Transp. Res. Part B* 68, 52–75.
 Watling, D.P., Maher, M.J., 1992. A statistical procedure for estimating a mean origin-destination matrix from a partial registration plate survey. *Transp. Res. Part B* 26 (3), 171–193.
 Zhou, X.S., List, G.F., 2010. An information-theoretic sensor location model for traffic origin-destination demand estimation applications. *Transp. Sci.* 44 (2), 254–273.
 Zhu, J.Y., Ye, X., 2018. Development of destination choice model with pairwise district-level constants using taxi GPS data. *Transp. Res. Part C* 93, 410–424.

Parking Demand Models

S.C. Wong*, **Zhi-Chun Li†**, **William H.K. Lam‡**, *Department of Civil Engineering, The University of Hong Kong, Hong Kong, China; †School of Management, Huazhong University of Science and Technology, Wuhan, China; ‡Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China

© 2021 Elsevier Ltd. All rights reserved.

Parking Demand Forecasting Models	519
Parking Behavior Analysis Models	520
Factors That Influence Parking Location Choice	520
Parking Location Choice Models That Do Not Consider Travel Choices	520
Integrated Models for Parking Location and Travel Choices	521
Parking Demand Management Strategies	521
Parking Pricing	521
Parking Supply-Side Strategies	521
Park-and-Ride Schemes	522
Remark	522
Acknowledgment	522
References	522
Further Reading	522

Parking has become an increasingly serious problem in many large cities around the world because of the dramatic increase in the number of motorized vehicles. Traffic conditions have tended to deteriorate in many urban areas, particularly in densely populated development areas during peak periods. And hence, the time needed to search for parking has increased substantially in central business districts. An efficient parking demand management tool is required to improve the efficiency of urban transportation systems. This paper concerns the main aspects of parking demand study, including parking demand forecasting, parking behavior analysis, and parking demand management strategies.

The following parking statistics are important in parking demand research:

- *Parking accumulation* is defined as the number of vehicles parked at a given moment, which can be expressed by an accumulation curve (obtained by plotting the number of parking spaces occupied over time).
- *Average parking duration* is the ratio of the total parking time of the vehicles parked to the number of vehicles parked.
- *Parking turnover* is the ratio of the number of vehicles parked during a particular period to the number of parking bays available. For example, if a parking lot has 100 spaces and 300 different vehicles occupy the lot over the day, then the turnover is 3.0.
- *Parking occupancy* is the number of vehicles parked in a specific parking location, represented as the percentage of spaces occupied. The parking occupancy is an important index for the measurement of parking performance because it can help us understand the dynamics of parking demand during a day.
- *Parking occupancy rate* is the percentage of all parking spaces occupied by vehicles at a given time.

Parking Demand Forecasting Models

Advanced parking demand forecasting models are required to forecast the temporal and spatial distribution of future parking demand in a densely populated urban area. Reliable parking demand forecasting is the basis of efficient planning and management of a city's parking facilities. Overestimation of the parking demand inevitably results in a waste of funds and land resources devoted to parking facilities. Underestimation may lead to a severe shortage of parking spaces, which may restrict urban economic activities because of excessively long times needed to search for an available parking space in an urban area. Parking demand forecasting issues have recently attracted considerable attention because of their increasing importance particularly under the big data arena for smart parking demand management.

Parking demand forecasting usually adopts a combination of a disaggregate modeling approach and a regression approach, such as multiple linear regression. Such a combined method was successfully applied in a study of Hong Kong's parking demand by the Transport Department of Hong Kong in 1995 (Lam et al., 1998; Wong et al., 2000; Tong et al., 2004b).

In the disaggregate approach, the districts of the city concerned are divided into many zones, and each zone is further subdivided into many subzones according to land-use type (e.g., residential, institutional, business, industrial, agricultural, parkland). Parking activities are classified by various attributes, such as trip purpose (home-based work, home-based school, other home-based, nonhome-based, employer's business), parking space ownership (ownership-related, usage-related), vehicle type (private car, goods vehicle), and parking characteristics (on-street metered parking, on-street nonmetered parking, off-street parking, and illegal parking).

The multiple linear regression approach is used to estimate the demand for each parking activity in each zone. Such parking demand is related to various factors, such as demographic and socioeconomic characteristics, land-use patterns, vehicle ownership, parking supply, parking duration, parking turnover rate, parking searching time, and parking charges. The regression approach is used to reveal each factor's relationship with parking demand.

A city's population growth and socioeconomic development are closely related to its vehicle ownership and the trip frequency of its residents. As the level of socioeconomic development increases, the number of motorized vehicles and the frequency of their vehicle use increase, and parking demand thus increases with the level of demographic and socioeconomic development.

Land-use patterns show the use of the city's available land as dictated by urban and regional planners, including the land-use type and the development intensity of the unit land area. Urban land resources are usually allocated for the development of various functional facilities, such as commercial districts, schools, hospitals, and residential neighborhoods. The functional type of a zonal area determines its parking demand characteristics, including the number of trips generated or attracted, parking generation rate, parking duration, and parking accumulation. The intensity of land development also affects the parking demand in each zone. In general, the greater the intensity of the land development (or equivalently, the higher the floor area ratio), the greater the demand for parking in that area.

Vehicle ownership directly affects the implementation of trip activities and thus has an important influence on parking demand. Vehicle ownership is closely relevant to the supply of parking facilities. An increase in vehicle ownership usually requires the provision of more parking facilities in an urban area. It has been reported that the addition of a car to a city requires 1.2–1.5 additional parking spaces.

The parking turnover of a parking space is its rate of use. Turnover indicates how often a parking stall is used by different vehicles throughout the day. A higher average turnover rate is due to the fact of a shorter average parking duration. As a result, more vehicles can be parked and thus a higher parking service level, and the inverse is also true.

Parking search time is the time spent on looking for an available parking space. The parking search time, the walking time from the parking space to the final destination, and the parking charges affect a driver's choice of parking location and thus the distribution of parking demand.

Parking Behavior Analysis Models

Parking behavior analysis has great importance in exploring the spatial and temporal distribution of parking demand. The previous related models may be classified into two categories in terms of their objectives. One focuses only on the choice of parking location or lot (ignoring travel choice behavior on the road network) and the other considers the combined parking and travel choices. The former determines the parking location choice based on the parking facility's characteristics alone, whereas the latter aims to reveal the interaction between parking congestion and travel choices on the road network. In the following section, the factors that influence the choice of parking location are described, and the parking location choice models are introduced.

Factors That Influence Parking Location Choice

The characteristics or attributes of the parking locations greatly influence drivers' parking location choices. The parking location characteristics include the access time to the parking area, the parking charges, the hours of operation, the parking occupancy rate, and the walking time to the final destination. In a large central business district, a wide range of parking location types are available, including surface parking lots, multistory or underground garages, metered and unmetered on-street parking lots, and off-street parking locations.

Drivers' preferences for a parking location are also a significant factor. Drivers generally have preferences for their desired type of parking facilities and locations of parking lots. These preferences vary by the purpose of the trip, the drivers' perceptions of parking location characteristics, and inherent differences in their tastes and attitudes.

Parking cruising is another important factor (Shoup, 1999). When a driver arrives at a parking lot and finds that it is fully occupied, he or she must cruise to find the next parking lot. This phenomenon is quite common in daily life, especially during peak hours or at busy locations. Parking cruising has a significant effect on the search time for an available parking space and is a major contributor to road traffic congestion in the downtown areas of major cities.

Parking Location Choice Models That Do Not Consider Travel Choices

Discrete choice modeling is a popular method of modeling drivers' parking location choice behavior. In this family of models, various parking locations are considered as discrete alternatives. Drivers choose their parking locations according to the attributes of the parking alternatives and/or their preferences for parking locations under various behavioral decision rules. A multinomial logit model is usually used (Hess and Polak, 2004), which can also be extended as a nested or mixed type to consider correlations between the alternatives (Hunt and Teplý, 1993). The parameters in the multinomial logit model or its variations can be calibrated with various estimation methods (such as the maximum likelihood method or least-squares method) based on the stated preference and/or revealed preference survey data.

Integrated Models for Parking Location and Travel Choices

Parking is closely tied to road usage. The parking service level and parking policies affect not only drivers' parking location choices, but also their travel choices on the road network (e.g., route, departure time, and transport mode). One class of models integrates parking location and travel choices to consider the interaction between road traffic and parking congestion. Such integrated models usually adopt a network equilibrium modeling approach, which accounts for the demand–supply equilibrium of the parking services and travel choices (e.g., route, departure time, and mode choices) of the drivers on the network (Gur and Beimborn, 1984; Bifulco, 1993; Lam et al., 1999; Tong et al., 2004a; Lam et al., 2006). The generalized cost of a route–parking alternative from an origin to a destination consists of the following cost components: the in-vehicle travel time from the origin to the parking location, the search time for an available parking space, the parking charge, and the walking time from the parking location to the destination.

When making travel and parking choices, drivers consider the generalized costs of all route–parking alternatives between an origin–destination pair and choose one route–parking alternative with minimal generalized cost for the trip. When a user equilibrium state of the simultaneous route and parking location choice is achieved, for each origin–destination pair, the used route–parking alternative has the minimal generalized cost, and the generalized cost of any unused route–parking alternative is greater than or equal to the minimum. The static integrated model of parking location and travel choice can be extended to a dynamic version with the addition of a time dimension (e.g., departure time and parking duration).

Parking Demand Management Strategies

Parking demand management strategies include a number of policies and programs designed to reduce parking demand, preserve parking for certain trip types and users, and promote a shift from single-occupant vehicle trips to transit, pedestrian, and bicycling trips. These parking policies and programs offer a potentially strong instrument to influence road traffic flow and parking demand, including: (1) giving up on a trip plan because of heavy parking congestion, (2) changing the parking location or parking duration because of a parking time limit or high parking fee, (3) changing the trip destination because of parking control, (4) changing the travel mode (e.g., bus or rail), (5) changing the departure time (e.g., from peak period to off-peak period) or travel route, and (6) changing the trip frequency. Parking demand management strategies include parking pricing, supply-side strategies, and park-and-ride (P&R) schemes with an integrated multimodal service. Parking pricing involves the fee charged for parking, parking supply-side strategies involve restricting the supply of available parking spaces to achieve a desired objective, and P&R schemes involve an integrated service between multiple travel modes.

Parking Pricing

Parking pricing, or the fee charged for parking, is commonly used to control parking demand in busy parking lots during rush hours, including the application of on-street parking meters and pricing for garage parking (Anderson and de Palma, 2004; Van Ommeren and Russo, 2014; Qian and Rajagopal, 2015). The basic idea behind this policy is to control the spatial and temporal distribution of parking demand by properly setting parking rates to maintain an ideal level of parking occupancy such that any newly arriving drivers can always find a suitable open space without circling around.

Typical examples of parking pricing schemes include (1) differential pricing implemented over space and time, such as raising parking charges in areas with high parking demand (such as central business districts, employment areas) and during peak hours of parking demand (dynamic pricing responsive to the real-time parking occupancy level); (2) parking tariffs designed to discourage long-term parking and promote parking turnover (e.g., to prioritize parking for shopping customers rather than employees); (3) a downtown employee parking payroll tax; and (4) parking subsidies, including policies such as employer-paid parking, free shopping mall parking, parking coupons, and park-and-shop discount programs.

Rapid advances in information and communication technologies under the big data arena have allowed the use of intelligent parking systems to determine parking prices dynamically at various parking locations, leading to more efficient use of limited parking resources for smart parking demand management.

Parking Supply-Side Strategies

Parking supply-side strategies involve restricting the supply of available parking locations to achieve a desired level. Some typical strategies include (1) setting the maximum number of parking spaces and parking time limits (e.g., 2-hour maximum parking) for each zone (Arnott and Rowse, 2013); (2) designing parking permit programs in which only vehicles with permits are allowed to park in a parking permit area, such as residential parking and employer-provided parking (the distribution and trading of parking permits are important issues in this theme) (Zhang et al., 2011); (3) shared parking strategies, in which a parking space can serve two or more individuals at different times without conflict (an efficient shared parking scheme can significantly improve the usage rate of parking spaces) (Xu et al., 2016); (4) preferential parking for carpools or vanpools or alternative-energy vehicles; and (5) parking guidance systems, which guide drivers to the nearest available parking space by displaying real-time occupancy information.

Park-and-Ride Schemes

P&R schemes have been recognized as an effective means of addressing traffic congestion problems in densely developed urban areas (Lam et al., 2001; Wang et al., 2004; Li et al., 2007). These schemes encourage commuters to transfer from private cars to public transport modes to enter the city's central areas. Accordingly, the journey is divided into two parts: driving a private car to a car park near a transit station and making the remaining journey through public transit. Determining the optimal location and size of each P&R site is very important for successful implementation of these schemes. In practice, P&R sites are located at the periphery of built-up areas, near major radial road corridors and metro stations, and at or near the urban fringe.

P&R services can serve to supplement other transport policies, such as the restriction of parking supplies in urban central areas, increasing parking charges in central-area car parks, and the promotion of public transport, cycling, and walking. P&R schemes are becoming popular in some large Asian cities, such as Hong Kong, Beijing, and Shanghai, with the rapid development of metro systems.

Remark

This paper illustrates the wide scope of parking issues in order to achieve a better understanding of parking demand and its impacts on urban traffic conditions. Apart from providing timely reviews of recent methodological advances on various parking issues, we hope that this paper will inspire and stimulate new research initiatives and efforts in the application of various parking demand models, parking behavior analysis methods, intelligent parking systems, and parking demand management strategies for the design of more efficient and reliable transportation networks in future.

Acknowledgment

The work described in this paper was jointly supported by grants from the National Key Research and Development Program of China (2018YFB1600900), the National Natural Science Foundation of China (71525003, 71890970/71890974), the NSFC-EU Joint Research Project (71961137001), and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. R5029-18).

References

- Anderson, S.P., de Palma, A., 2004. The economics of pricing parking. *J. Urban Econ.* 55 (1), 1–20.
- Arnott, R., Rowse, J., 2013. Curbside parking time limits. *Transp. Res. Part A Policy Pract.* 55, 89–110.
- Bifulco, G.N., 1993. A stochastic user equilibrium assignment model for the evaluation of parking policies. *Eur. J. Oper. Res.* 71, 269–287.
- Gur, Y.J., Beimbom, E.A., 1984. Analysis of parking in urban centers: equilibrium assignment approach. *Transp. Res. Rec.* 957, 55–62.
- Hess, S., Polak, J., 2004. Mixed logit estimation of parking type choice. Presented at the 83rd Annual Meeting of the Transportation Research Board, Washington, DC.
- Hunt, J.D., Teply, S., 1993. A nested logit model of parking location choice. *Transp. Res. Part B Methodol.* 27 (4), 253–265.
- Lam, W.C.H., Fung, R.Y.C., Wong, S.C., Tong, C.O., 1998. The Hong Kong parking demand study. *Proc. Inst. Civil Eng. Transp.* 129, 218–227.
- Lam, W.H.K., Li, Z.C., Huang, H.J., Wong, S.C., 2006. Modeling time-dependent travel choice problems in road networks with multiple user classes and multiple parking facilities. *Transp. Res. Part B Methodol.* 40 (5), 368–395.
- Lam, W.H.K., Nicholas, M.H., Lo, H.P., 2001. How park-and-ride schemes can be successful in Eastern Asia. *ASCE J. Urban Plan. Develop.* 127, 63–78.
- Lam, W.H.K., Tam, M.L., Yang, H., Wong, S.C., 1999. Balance of demand and supply of parking spaces. In: Ceder, A. (Ed.), *Transportation and Traffic Theory*. Elsevier, Oxford, pp. 707–731.
- Li, Z.C., Lam, W.H.K., Wong, S.C., Zhu, D.L., Huang, H.J., 2007. Modeling park-and-ride services in a multimodal transport network with elastic demand. *Transp. Res. Rec.* 1994, 101–109.
- Qian, Z., Rajagopal, R., 2015. Optimal dynamic pricing for morning commute parking. *Transp. A Transp. Sci.* 11 (4), 291–316.
- Shoup, D.C., 1999. The trouble with minimum parking requirements. *Transp. Res. Part A Policy Pract.* 33 (7–8), 549–574.
- Tong, C.O., Wong, S.C., Lau, W.W.T., 2004a. A demand-supply equilibrium model for parking services in Hong Kong. *Hong Kong Inst. Eng. Trans.* 11 (1), 48–53.
- Tong, C.O., Wong, S.C., Leung, B.S.Y., 2004b. Estimation of parking accumulation profiles from survey data. *Transportation* 31 (2), 183–202.
- Van Ommeren, J.N., Russo, G., 2014. Time-varying parking prices. *Econ. Transp.* 3 (2), 166–174.
- Wang, J.Y., Yang, H., Lindsey, R., 2004. Locating and pricing park-and-ride facilities in a linear monocentric city with deterministic mode choice. *Transp. Res. Part B Methodol.* 38 (8), 709–731.
- Wong, S.C., Tong, C.O., Lam, W.C.H., Fung, R.Y.C., 2000. Development of parking demand models in Hong Kong. *J. Urban Plan. Develop.* 126 (2), 55–74.
- Xu, S.X., Cheng, M., Kong, X.T., Yang, H., Huang, G.Q., 2016. Private parking slot sharing. *Transp. Res. Part B Methodol.* 93, 596–617.
- Zhang, X., Yang, H., Huang, H.J., 2011. Improving travel efficiency by parking permits distribution and trading. *Transp. Res. Part B Methodol.* 45 (7), 1018–1034.

Further Reading

- Arnott, R., de Palma, A., Lindsey, R., 1991. A temporal and spatial equilibrium analysis of commuter parking. *J. Public Econ.* 45 (3), 301–335.
- Arnott, R., Inci, E., 2006. An integrated model of downtown parking and traffic congestion. *J. Urban Econ.* 60 (3), 418–442.
- De Borger, B., 2011. Optimal congestion taxes in a time allocation model. *Transp. Res. Part B Methodol.* 45 (1), 79–95.
- Fosgerau, M., de Palma, A., 2013. The dynamics of urban traffic congestion and the price of parking. *J. Public Econ.* 105, 106–115.

- Hensher, D.A., King, J., 2001. Parking demand and responsiveness to supply, pricing and location in the Sydney central business district. *Transp. Res. Part A Policy Pract.* 35 (3), 177–196.
- Inci, E., 2015. A review of the economics of parking. *Econ. Transp.* 4 (1–2), 50–63.
- Liu, W., 2018. An equilibrium analysis of commuter parking in the era of autonomous vehicles. *Transp. Res. Part C Emerg. Technol.* 92, 191–207.
- Shoup, D.C., 2006. Cruising for parking. *Transp. Policy* 13 (6), 479–486.
- Thompson, R.G., Richardson, A.J., 1998. A parking search model. *Transp. Res. Part A Policy Pract.* 32, 159–170.
- Verhoef, E., Nijkamp, P., Rietveld, P., 1995. The economics of regulatory parking policies: the (im)possibilities of parking policies in traffic regulation. *Transp. Res. Part A Policy Pract.* 29 (2), 141–156.
- Young, W., Thompson, R.G., Taylor, M.A., 1991. A review of urban car parking models. *Transp. Rev.* 11 (1), 63–84.

Pavement Management Systems

Senthilmurugan Thyagarajan, ESC Inc, Turner-Fairbank Highway Research Center, McLean, VA, United States

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	524
Importance	524
Pavement Management Approaches	524
Components of Pavement Management System	525
Data Collection	525
Inventory Information	525
Pavement Condition Evaluation	525
Traffic Counts	526
Data Management	526
Performance Modeling and Prediction	527
Engineering and Economic Analysis	527
Reporting and Feedback	528
Future Directions	528
Closing the Gap Between Pavement Design and Pavement Management for Better Pavements	528
Technology	528
Evolution of Transportation Asset Management (TAM)	528
Performance Management and Accountability	529
Life-Cycle Planning (LCP)	529
Incorporation of Sustainability and Risk	529
Biography	530
References	530
Further Reading	530

Nomenclature

FHWA Federal Highway Administration
AASHTO American Association of State Highway and Transportation Officials
ASTM American Society for Testing and Materials
NCHRP National Cooperative Highway Research Program

Importance

The 2016 American Infrastructure Report Card from the American Society of Civil Engineers gave D+ grade for America's infrastructure and estimates a 1.1 trillion USD funding gap on surface transportation investment (ASCE, 2016). This highlights the urgent responsibility that highway agencies face to develop strategies to achieve and maintain their assets in a state of good repair at minimum practical cost over the life cycle. At the same time, agencies are also facing severe funding shortfalls due to gas tax and other revenues that is not keeping pace with the growth vehicle miles traveled and inflation that require them to establish performance-based metrics to defend funding requests and provide accountability. Implementation of a systematic and strategic approach of managing pavements using a PMS can lead to improved network condition (Fig. 1), a higher level of service to the public, a better understanding of trade-off analysis between different assets, and a more streamlined decision process. While a PMS cannot make final decisions, it can provide the basis for an informed decision and possible consequences of alternative strategies (AASHTO, 2012). Many transportation agencies are also developing integrated approaches for making investment decisions that consider impacts on multiple assets. In the United States for example, recent law requires States to develop Transportation Asset Management Plans (TAMP) that act as a focal point for information about the assets, their management strategies, long-term expenditure forecasts, and business management processes. TAMPs are essential management tools that bring together all related business processes and stakeholders to achieve a common understanding and commitment to improve performance.

Pavement Management Approaches

Pavement management supports agencies decision at three levels: strategic, network (tactical), and project (operational). Strategic decisions are made at the upper management level within the agency by staffs responsible for making policy and investment

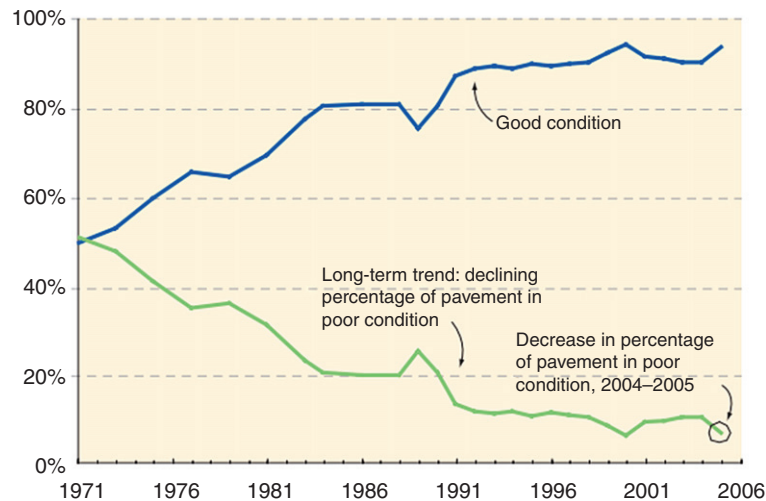


Figure 1 Trends in poor and good pavement condition in Washington State highways following adaptation of a pavement condition survey in 1969 and a pavement management system in 1982 (FHWA, 2010a). Source: Washington State Department of Transportation Materials Lab.

decisions, such as elected officials, transportation boards and commissions, and city councils. The task at this level is to make long-term strategic decisions that reflect agency priorities and setting performance targets. The decisions at this level are generally less structured than decisions made at other levels.

Network-level decisions are concerned with achieving agency goals for an entire road network through multiyear improvement programs. These decisions include: establishing pavement preservation policies, identifying priorities, developing recommended list of maintenance, rehabilitation, and reconstruction (MR&R) treatment type, location and timing over the multiyear analysis horizon and estimating corresponding funding needs, and allocating budgets based on prioritization that meets the agency policies and priorities. The results of the network-level analysis are used in strategic level to set realistic performance targets and to evaluate investment options. Project-level decisions address selection of site-specific MR&R actions for individual projects and groups of projects based on site-specific conditions. For instance, cores and material testing may be conducted to determine in situ structural conditions at project level, but an indirect nondestructive testing is usually used at network level.

Components of Pavement Management System

The basic component of pavement management system (PMS) includes inputs, database, analysis components and modules, reporting module, and feedback loop. The features of each component vary with individual agencies requirement and the subsequent section provides a general overview of each component.

Data Collection

The data upon which all decisions are made form the basis for any PMS. The quality of decisions is largely function of the data available to support those decisions. Inputs include information on inventory, pavement condition, age of last and/or previous treatments, and traffic counts. The information needed are often collected or maintained by different divisions of the agencies. Traffic or planning division staff are responsible for traffic data while construction and maintenance staff are responsible for maintenance and rehabilitation history of the pavements.

Inventory Information

The level of detail per section depends on each agency's goals and methodology to support their decision. A typical inventory feature of the pavement segment includes start and end reference point and/or global positioning system (GPS) coordinates, route designation along with route type, functional classification, segment length, width and type of pavement and shoulder, number of lanes in each traffic direction, construction history (year, treatment type, layer(s) type, material properties, and layer(s) thicknesses), subgrade type, and climate and traffic information. In addition, for jointed plain concrete pavements, joint spacing and transverse joint load transfer are also included when possible.

Pavement Condition Evaluation

Pavement condition evaluation is a key element of pavement management. Systematic and regular pavement condition evaluation provides the data an agency needs to make informed decisions on treatment type, location and timing for the efficient management

Table 1 Pavement distresses (FHWA, 2004)

<i>Pavement</i>	<i>Hot mix asphalt</i>	<i>Jointed plain and jointed reinforced</i>	<i>Continuously reinforced</i>	<i>Composite</i>
States reporting	50	47	21	41
Roughness	98%	100%	100%	100%
Rutting	100%	N/A	N/A	N/A
Fatigue or alligator cracking	96%	N/A	N/A	N/A
Longitudinal cracking	86%	81%	90%	100%
Transverse cracking	94%	83%	N/A	100%
Transverse joint faulting	N/A	74%	N/A	N/A
Transverse joint spalling	N/A	74%	N/A	N/A
Punch-outs	N/A	N/A	86%	N/A
Surface friction	68%	66%	81%	68%
Other	46%	47%	57%	61%

of the network, report current pavement condition, assess past, and forecast future performance. Network-level pavement condition evaluations are conducted on a 1–3-year cycle based on agency policies, legislative requirements, and highway classification. The type of pavement condition data normally collected by transportation agencies can be generally fit into three categories:

1. **Surface distresses:** Characterizing visible condition along the pavement surface can provide direct indication of the performance problem and thereby helpful in selecting suitable treatment and planning long-term strategies. The cause can be either traffic load or non-traffic load related. The distresses are categorized by pavement type and rated by severity (e.g., [WSDOT, 1992](#)). The common distresses collected in the United States are presented in [Table 1](#). The other widely used guide to define pavement distresses is ASTM D6433, Standard Practice for Roads and Parking Lots Pavement Condition Index Surveys. Distress surveys can be manual or automated. Manual surveys are visual assessments of field conditions conducted through the windshield of a vehicle or by walk along the section. Automated surveys are conducted using instrumented vehicles and then computers are used in data processing to extract distress information with little or no human intervention (NCHRP Synthesis 334) ([National Academies of Sciences, Engineering, and Medicine, 2004](#)).
2. **Surface characteristics:** This involves measured surface smoothness (longitudinal profile, surface texture) and noise. Longitudinal profile is typically reported using International Roughness Index (IRI), an index originally developed by World Bank for use in Highway Development Model. Present serviceability index (PSI) is another index commonly used to characterize roughness and distresses. Surface texture is subdivided into three distinct groupings by their wavelength and amplitude: microtexture, macrotexture, and megatexture. The microtexture and macrotexture influence pavement's wet weather friction characteristics and hydroplaning potential. Noise is still not mostly collected on routine basis.
3. **Structural capacity:** Functional condition is not necessarily linked to structural condition. Structural evaluation estimates the adequacy of an existing pavement to meet future traffic loadings. Destructive testing such as coring and nondestructive testing such as pavement deflection devices are the common methods to estimate structural adequacy. Falling weight deflectometer (FWD) is the most commonly used device to assess pavement responses. The structural evaluation is typically conducted at project-level investigation due to the limitation of technology to conduct network-level investigation. However, the latest advancement in traffic speed deflection device technology has enabled deflection measurement at network level without the necessity of traffic control for testing. Traffic speed deflectometer and RAPTOR are the two commercially available devices for network-level pavement structural evaluation.

Traffic Counts

Past and projected traffic volumes and loads are critical in making both network- and project-level pavement management decisions. Traffic monitoring guide ([FHWA, 2001](#)) provides a basic traffic data collection framework for agencies. Automated and manual are two general methods of traffic data collection. A weigh-in-motion (WIM) detector is the common automated device used to measure dynamic tire forces of a moving vehicle and estimates corresponding static tire loads. Manual method refers to visually observing number of vehicles and classification, vehicle occupancy, turning movement counts, or direction of traffic using tally sheets or electronic counting boards.

Data Management

Design, construction, rehabilitation, and maintenance information used with in pavement management are often drawn from different sections within an agency. Location referencing system (LRS) is the foundation of a road data management system and hence the key to data integration. A LRS contains multiple location reference methods, each referring assets and/or attributes for a single asset within their reference framework. For pavement management, data can be referenced using linear features, point features, or a combination of the two. For example, smoothness data (IRI) are based on fixed linear segments, while friction data or FWD are based on point data. Geographic information system (GIS) containing spatial and attribute data is used for storing, managing, retrieving, querying, analyzing, and presenting data.

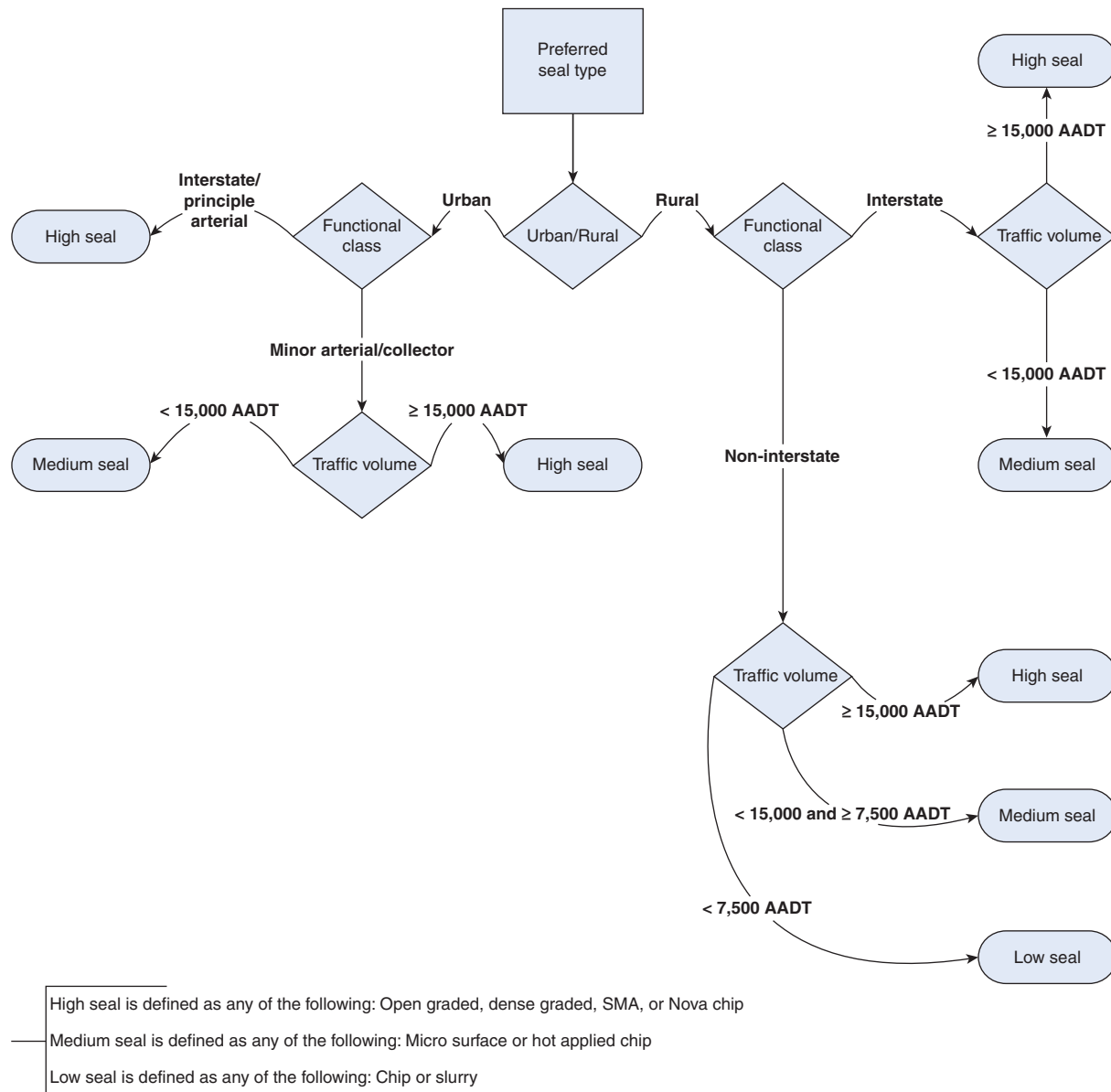


Figure 2 Sample decision tree (FHWA, 2008).

Performance Modeling and Prediction

The treatment and project selection methodology involve pavement deterioration models, treatment rules, and cost model. In an ideal case, the effect of any treatment should be mechanistically evaluated considering current pavement condition and future traffic loading and climate in determining the treatment type and timing for any pavement section to minimize long-term or life-cycle cost. However, current pavement deterioration models are largely empirical and the agency develops decision tree/matrix (Fig. 2) based on experience with treatment types and their performance when applied to pavements at different conditions. The decision tree includes performance models, treatment trigger rules, and impact rules or reset values that define the jump in pavement condition after each treatment application. The treatment costs are also required in estimating budget needs.

Engineering and Economic Analysis

Depending on the technical and business requirements and policy of the agency, the analysis may involve identification of projects and treatments, and consequence of different investment levels. Most agencies employ proactive pavement preservation activities to thriftilly extent pavement service life. Under fiscal constrained analysis, approaches such as ranking, prioritization, or optimization are used in project selection at single- or multiyear analyses. Life-cycle cost analysis (LCCA) is common method of comparing the

cost-effectiveness of a number of different strategies over a multiyear analysis period. This module also estimates future network condition for different investment levels that would be instructive in planning and reporting to elected officials.

Reporting and Feedback

Stakeholders use pavement management information for various purposes. Conventional reports are the common method of distributing information after completion of each round of condition surveys. Agencies use GIS-based tools or web-based programs to access and view pavement management information. Graphics are usually used to communicate the primary message quickly and effectively.

Feedback loop is a critical step in keeping PMS abreast. Updates shall include changes in pavement performance or conditions under which treatments are used, or changes in material properties or construction. Ideal scenario would be to create a feedback mechanism between pavement designs and observed performance that will lead to more cost-effective pavement designs.

Future Directions

Pavement management continues to evolve to meet progressive goals of the agencies. In 2010, FHWA sponsored the development of pavement management road map to identify steps needed to address current gaps in pavement management ([FHWA, 2010b](#)). Short- and long-term needs are grouped under the use of existing technology and tools, institutional and organizational issues, broad role of pavement management, and new tools, methodologies, and technologies. In addition, the following topics are identified as some of the ongoing researches to better serve the progressive goals.

Closing the Gap Between Pavement Design and Pavement Management for Better Pavements

PMS concept was conceived as framework for pavement design (project level) but has since evolved, and in many cases disassociated from pavement design, as a system to optimally manage and maintain pavements (network level). Contributing to this divergence were the different performance indicators used for pavement design in previous design procedures and pavement management. Development of the current mechanistic-empirical-based pavement design procedure based on similar performance indicators as that used for pavement management has brought to forefront the need and the benefits of relinking pavement design and pavement management. Pavement design and management share many features and can be the feedback loop for each other. For instance, integration of pavement design inputs into pavement management can lead to better performance prediction, while past performance in the management can be used to arrive at better design. Calibration/validation can be done ad hoc. Spatial variability on each pavement section and the variability among pavement sections with similar traffic-climate-materials can be used in transforming deterministic performance prediction models into probabilistic models. Closing the gap would reconcile the maintenance and rehabilitation (M&R) treatments regime that is derived from LCCA during project-level design with the M&R regime that is the result of optimal budget allocation during network-level PMS.

Technology

In recent years, significant improvement has been made in data collection technology. Automated pavement condition surveys are more efficient and safer than manual surveys. The laser crack measurement system (LCMS) is one of the recent, high-performing, and accurate technology available commercially for pavement crack detection. Continuous friction measurement equipment (CFME) such as the Sideway-Force Coefficient Routine Investigation Machine (SCRIM) and proactive pavement friction management programs are being developed to significantly reduce the number and severity of crashes by decreasing pavement friction and texture-related crashes.

Integrated units are currently available that can collect at highway speeds the pavement surface distress and structural data as well as other characteristics such as transverse profile/rutting, grade, cross-slope, pavement texture, GPS coordinates, right-of-way video, and pavement video.

With advancement in vehicle automation technology comes the challenge of understanding its impacts on roadway infrastructure. The connected vehicle and truck platooning that might reduce wheel wander and spacing between trucks thus increasing concentrated wheel loads and reducing pavement relaxation time are currently being investigated to understand their effect on pavement performance.

Evolution of Transportation Asset Management (TAM)

TAM is a process that enables agencies to use engineering and business practices to make more informed and transparent investment decisions. Several processes are required for successful TAM practices, including effective decision analysis methodologies that are informed by quality data and performance measures. The interaction between models, data, and performance measures is a critical one for the success of TAM implementation. There is a focus on incorporating cross-asset and trade-off analysis in management

systems that can be beneficial for coordinating total highway programs, determining performance measure targets, and allocating funding among different asset classes.

There is international effort on use of forward-looking metrics that measure the sustainability of infrastructure conditions. These metrics encourage a long-term, asset-management-based approach to managing infrastructure, not just to meet condition targets today, but also to sustain those targets into the future. Asset Sustainability Index is one such composite metric computed by dividing the amount budgeted on infrastructure maintenance and preservation over time by the amount needed to achieve a specific infrastructure condition target.

Performance Management and Accountability

Performance-based management program allows stakeholders to invest resources in projects that collectively will make progress toward the achievement of its strategic goals. For example, a new rule enacted in the United States, Moving Ahead for Progress in the 21st century (MAP-21) Act, is about establishment of a performance- and outcome-based program (MAP-21, 2012). The rule set national performance goals for federal-aid highway program in seven areas: safety, infrastructure condition, congestion reduction, system reliability, freight movement and economic vitality, environmental sustainability, and reduced project delivery delays. New performance measures are being established to quantify the new and changing agency goals and objectives and to set performance targets.

FHWA strategic plan for FY 2019-22 set safety, infrastructure, innovation, and accountability as the four strategic goals. There is increased effort to improve program and project decision-making by using a data-driven approach, asset management principles, and a performance-based program that will lead to better conditions and more efficient operations.

Life-Cycle Planning (LCP)

LCP is an approach to manage transportation assets over their whole life, covering the time each asset goes into service after construction to the time it is retired or disposed of. The Asset Management Rule (See 81 Federal Register 73196) by US DOT emphasizes the importance of LCP in its definition for asset management, as shown later: *A strategic and systematic process of operating, maintaining, and improving physical assets, with a focus on both engineering and economic analysis based on quality information, to identify a structured sequence of maintenance, preservation, repair, rehabilitation, and replacement actions that will achieve and sustain a desired state of good repair over the life cycle of the assets at a minimum practical cost.*

The rule defines LCP as “a process to estimate the cost of managing an asset class, or asset sub-group, over its whole life with consideration for minimizing cost while preserving or improving the condition.” LCP should not be confused with LCCA at the project level. An LCCA considers all agency expenditures and user costs throughout the life of strategy. LCP is a network-level analysis that encompasses many aspects of an engineering economic analysis, including opportunity cost, noneconomic factors, and funding constraints (FHWA, 2017). Remaining service interval (RSI) is one such performance measure that can unify the outcome of different life-cycle approaches to determine needs by focusing on when and what treatments are needed while maintaining the service level (Elkins et al., 2013).

Incorporation of Sustainability and Risk

Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs. Agencies strive to improve the triple bottom line (social, economic, and environmental factors) over the life cycle of an asset through life-cycle assessment (LCA). The economic component has been the dominant decision factor for many years, but recent years have seen the emergence of both the environmental and social components. Though all triple-bottom-line components are considered important, but the relative importance of these factors and how each are considered are driven by the goals, demands, characteristics, and constraints of a given project (Van Dam et al., 2015). Current research focuses on developing comprehensive LCA tool based on systematic framework, compiling unbiased database, product category rule for common materials, and quantifying contribution from different phases of the pavements including construction, maintenance, and use phase.

Risk assessment involves identifying and evaluating sources of risk and developing mitigation strategies. MAP-21 Act adopted a requirement for States to develop and implement risk-based asset management plans for the National Highway System to improve or preserve the condition of the assets and the performance of the system.

The infrastructure should be resilient to extreme weather, sea-level change, and changes in environmental conditions to preserve the considerable investment. Resilience as defined by FHWA is *the ability to anticipate, prepare for, and adapt to changing conditions and withstand, respond to, and recover rapidly from disruptions*. The Fixing America's Surface Transportation (FAST) Act in the United States requires agencies to take resiliency into consideration during transportation planning processes (FAST Act, 2015). US Department of Transportation Strategic Plan for FY 2018-2022 includes development of new tools to improve transportation infrastructure durability and resilience as a priority innovation area (Dix et al., 2018). The stakeholders have started working on assessing vulnerabilities, incorporate resilience in asset management plans, addressing resilience in project development and design, and optimizing operations and maintenance practices.

Biography



Senthilmurugan Thyagarajan is a Research Civil Engineer, ESC Inc., working at Turner-Fairbank Highway Research Center, McLean, Virginia, for over 10 years. His areas of interest include pavement design, pavement performance modeling, pavement structural evaluation, life-cycle cost analysis, life-cycle assessment, and pavement management-related topics. He has obtained his Bachelor's degree in Civil Engineering from Madurai Kamaraj University, India, and a Master's degree in Soil and Foundation Engineering from Anna University, India, and PhD in Civil Engineering from Washington State University, United States. He is licensed professional engineer in the state of Maryland and Virginia, USA.

References

- American Association of State Highway Transportation Officials (AASHTO), 2012. Pavement Management Guide, 2nd ed. AASHTO, Washington, DC.
- American Society of Civil Engineers (ASCE), 2016. Failure to Act: Closing the Infrastructure Investment Gap for America's Economic Future. American Society of Civil Engineers, Washington DC. Available from: <https://www.asce.org/failuretoact/>.
- Dix, B., Zgoda, B., Vargo, A., Heitsch, S., Gestwick, T., 2018. Integrating resilience into the transportation planning process. FHWA-HEP-18-050. Federal Highway Administration, Washington, DC.
- Elkins, G.E., Rada, G.R., Groeger, J.L., Visintine, B., 2013. Pavement remaining service interval implementation guidelines. Report No. FHWAHRT-13-050. Federal Highway Administration, Washington, DC.
- FAST Act, 2015. Fixing America's Surface Transportation Act. Available from: <https://www.fhwa.dot.gov/fastact/>.
- Federal Highway Administration (FHWA), 2001. Traffic Monitoring Guide. Federal Highway Administration, Washington, DC.
- Federal Highway Administration (FHWA), 2004. Summary of State Pavement Management Systems. Federal Highway Administration, Washington, DC.
- Federal Highway Administration (FHWA), 2008. Pavement Management System Peer Exchange Program Report: Sharing the Practices of the California, Minnesota, New York, and Utah Departments of Transportation. Federal Highway Administration, Washington, DC.
- Federal Highway Administration (FHWA), 2010. Pavement Management Systems: The Washington State Experience. Office of Asset Management, Federal Highway Administration, Washington, DC.
- Federal Highway Administration (FHWA), 2010. Pavement management roadmap. Report No. FHWA-HIF-11-011. FHWA, Washington, DC.
- Federal Highway Administration (FHWA), 2017. Using a Life Cycle Planning Process to Support Asset Management. Federal Highway Administration, Washington, DC. Available from: https://www.fhwa.dot.gov/asset/pubs/life_cycle_planning.pdf.
- MAP-21, 2012. Moving ahead for progress in the 21st century. Available from: <https://www.fhwa.dot.gov/map21/>.
- National Academies of Sciences, Engineering, and Medicine, 2004. Automated pavement distress collection techniques. NCHRP Synthesis 334. The National Academies Press, Washington, DC. Available from: <https://doi.org/10.17226/23348>.
- Van Dam, T., Harvey, J., Muench, S., Smith, K., Snyder, M., Al-Qadi, I., Ozer, H., Meijer, J., Ram, P., Roesler, J., Kendall, A., 2015. Towards sustainable pavement systems: a reference document. Federal Highway Administration, Washington, DC.
- WSDOT, 1992. Pavement surface condition rating manual. Northwest Technology Transfer Center, Washington State Department of Transportation, Local Programs Division. Available from: <https://www.wsdot.wa.gov/NR/rdonlyres/1AB0E29D-72D7-466A-9547-C9F631B4CE6C/0/PavementSurfaceConditionRatingManual.pdf>.

Further Reading

- Federal Highway Administration (FHWA), 2018. FHWA strategic plan FY 2019-2022. FHWA-PL-18-025 2018. Washington, DC. Available from: <https://www.fhwa.dot.gov/policy/fhwaplan.cfm>.
- Haas, R., Hudson, W.R., Cowe Falls, L., 2015. Pavement Asset Management. Wiley Publication, New York.
- Pavement Interactive, 2010. Online repository for complete articles, news, and information on all topics related to pavements. Developed by the pavement tools consortium, a partnership between several state DOTs, the FHWA, and the University of Washington. Available from: <https://www.pavementinteractive.org/>.

Residential Location Choice Models

Shlomo Bekhor*, Sigal Kaplan†, *Faculty of Civil and Environmental Engineering, Technion—Israel Institute of Technology, Haifa, Israel;
†Department of Geography, Hebrew University of Jerusalem, Jerusalem, Israel

© 2021 Elsevier Ltd. All rights reserved.

Introduction	531
Theoretical Foundation	531
Empirical Choice Modeling	532
Defining the Residential Choice Outcome and Choice Set Formation	532
Joint Choices	533
Population Strata	533
Decision-Makers	533
Decision Rules and Model Form	534
Data Sources for Estimation	534
Implementation in Transport Models	535
Further Research	535
References	535
Further Reading	536

Introduction

Being a cornerstone of spatiotemporal urban processes in the modern era, residential choice has been a focus of interest for urban planners, economists, and transport planners alike for almost 50 years. Residential choice can be viewed from the real-estate market perspective, the urban growth and development perspective, and the urban flow perspective. *Looking through the lens of the real-estate market*, residential choice serves to understand spatiotemporal dynamics, consumer preferences, and constraints. Unlike other consumer goods, buildings are durable assets that have an exceptionally long-lasting life span from decades to centuries, meaning that the current generation housing supply and the nature of the built environment is a mediator between past and future generations. Moreover, residential choice is the most important determinant of land consumption, defining the depletion of regional and national land resources. Further, a mismatch between housing supply and demand leads to surges of spontaneous, informal housing markets consisting of poorly built unsafe housing structures causing environmental, health, and life concerns. *From the regional development perspective*, residential housing markets represent bottom-up and middle-out market forces that create pressure for land-use development and top-down policy implementation. From the traditional policy and practice perspective, residential choice can determine the long-term success or demise of regional ventures, and the four stages of regional development: urbanization, expansion, growth, and decline and revitalization. Residential location choice helps to explain processes driving regional development such as inter- and intra-urban migration, suburbanization, segregation, ghost apartments and towns, labor market success and failure, recovery from natural hazards, and urban resilience. Therefore, understanding residential choice is a basic unit of analysis in regional master plans as well as in disaster management and increasing urban resilience. *From the urban flow perspective*, residential location choice affects the individual space–time prism in and is closely tied to lifestyle choices of land consumption, workplace, car ownership, modal choice, activity participation and travel patterns, as well as long-term life opportunities. Thus, residential choice drives the daily rhythms and flows of urban transport and activity patterns that allow planners to plan efficient, economic and sustainable transport, urban functions, and other vital infrastructures. Accordingly, residential choice is a basic building block in traditional four-stage transport models as well as the most advanced state-of-the-art agent-based urban development models, activity-based, urban transport models, as well as combined dynamic transport and land-use models.

Theoretical Foundation

The representation of residential choice in regional science was theorized in William Alonso's bid-rent theory (Alonso, 1964). The theory postulates households competing on land plots in a monocentric city, trading off distance from the city center with land-plot size. According to the theory assumptions, location matters with accessibility to the central business district (CBD) and main city function is the most attractive feature of urban location. However, households are willing to trade their accessibility to the CBD with a larger amount of land consumed. The theory produces a theoretical bid-rent curve with higher values near the CBD and lower values near the city outskirts. The theory established the two main factors in residential choice—the relative location from resources and the amount of land consumed. The theory explains residential, firm, and retail location. The theory was developed in the 1960s and its basic dynamics stem from agriculture and medieval city structure, but its rigorous economic principles hold for today's

modern cities. Nevertheless, the theory has undergone updates and amendments due to the growing complexity in urban form, and gained knowledge of human choice dynamics and the growing need to increase behavioral realism and modeling goodness-of-fit. Modern-day urban environments are multicentric and more complex and dispersed with declining and resurging trending urban locations, variety of housing types, and policy measures such as zoning and urban growth boundaries, which alter land-use dynamics. Yet, recent research of real-estate databases from the United States shows that the bid-rent curve holds at the local level when considering polycentrism and transportation infrastructure resulting in travel time as a distance measure. Starting from the 1970s, the bid-rent theory has led to the evolution of two separate research branches, which act as two-sided coin. The first branch takes the supply-side perspective by focusing on hedonic evaluation of dwelling units and land-use parcels for development. The second branch takes the demand perspective by looking at residential choice from a utilitarian perspective following the principles of microeconomics. The move to utility-based models in the 1970s was mainly driven by the need to develop and implement urban simulation models that represent the interaction between residential markets, labor markets, and transportation interact. Half a century later, in current agent-based models of urban development, which simulate urban development processes based on a utility-based development probability of land-use parcels, the two branches coincide to represent urban dynamics overtime.

Empirical Choice Modeling

Residential choice is one of the most complex choices in terms of its mathematical representation, due to three reasons. Hence, from the 1970s residential location choice has been extensively modeled both in transportation and regional science. The complexity is fourfold. First, location choice is characterized by an immense number of viable alternatives, considering both location and dwelling unit type. A metropolitan scale considering only location, the number of alternative zones as units of analysis could range from several hundred to over a thousand alternatives. When considering precise location and dwelling unit characteristics, the numbers become inflated and the choice problem could become unmanageable. Moreover, the choice is conducted under partial information, thus violating the assumption of full information embedded in the utilitarian approach. Second, residential choice is naturally highly influenced by population heterogeneity as a result of its importance, long-term impact, and high costs. Thus, in addition to mere socioeconomic background and constraints, the choice is influenced by individual life stage, lifestyle, attitudes toward housing, personal values, and norms. Third, location choice alternatives are spatially correlated because of geographical proximity, construction style similarities, location amenities, and sociodemographic characteristics. Last, while housing rental choices are conducted a few times in a lifetime, housing purchase and sales are rare events in the life of the individual and thus are difficult to trace. Because of these reasons residential choice is an elusive process, which is difficult to estimate and predict with high accuracy. Bearing this in mind, starting from Daniel McFadden's nested logit model (McFadden, 1978), the research efforts identifying and addressing the complexities embedded in residential choice helped gaining new insights and pushed forward advancements in econometric models. Since the 1970s, the use of probabilistic discrete choice models gradually became a main research tool exploring a wide range of aspects in residential choice. Due to the wide scope of residential choice and its interdisciplinary nature, encompassing and charting the full extent of current practice is essential for envisioning and undertaking new research directions. The following subsections describe the state of the practice in discrete choice modeling of residential decisions from several perspectives: definition of the residential choice bundle as the dependent variable, investigated themes, decision strategies, choice set formation, and model form and survey methods.

Defining the Residential Choice Outcome and Choice Set Formation

The choice of a dwelling unit as a bundle of attributes is inherently complex due to the multi-attribute nature of housing. Each dwelling unit is characterized by numerous attributes including geographical location, public services, social and environmental amenities, education (school) quality, accessibility, external and internal structural features and costs. However, true to the old saying that "location, location and location are the three most important factors in the real-estate market," the residential choice literature focused mainly on location and related amenities. A wide range of location variables were represented: size and density of areas, population characteristics and employment levels, land-use destinations, accessibility measures, commute and travel information, local transportation network, environmental amenities, social amenities, local public goods and services, and average housing costs. About 25% of the studies consider residential choice as a bundle of location-related features and dwelling unit characteristics, while less than 20% focused only on dwelling unit features. The dwelling attributes that were considered are costs, tenure, structural features (e.g., size, number of rooms, direction, and floor), appliances, and utilities. While most studies focused on residential choice as a single decision, some studies divide it into a series of correlated choices differentiating the decision to stay or move, the choice of tenure type, location choice, and dwelling type.

Since the compensatory approach assumes a choice from a given and well-specified choice set, the main approach for choice set formation requires defining the choice set prior to the model estimation either by application of deterministic availability rules, aggregation of alternatives, or sampling from the universal realm. Deterministic availability rules in the residential choice literature consisted usually of spatial constraints that were applied to the whole sample. Namely, the analysis was confined to one or more well-defined spatial units, such as county, city, community, or metropolitan area. Deterministic availability constraints were also applied per observation on the basis of travel time and housing affordability. Aggregation applied to residential choice concerned mainly housing market sectors by building type, tenure, size, or quality. Another aggregation method concerned spatial units by

proximity or by similarity. Sampling methods applied to residential choice comprised mainly stratified and non-stratified random sampling (Lee and Waddell, 2010). In terms of choice set-size, the literature shows great variation with the number of alternatives spanning from several alternatives in the early modeling stages to several hundred alternatives with more recent modeling capabilities.

Joint Choices

Understanding urban equilibrium of residential markets, labor markets, and transportation has largely driven the need for joint modeling of residential location choice and other lifestyle decisions. The earliest examples of joint decisions, already in the beginning of the 1980s, are an integrated choice of residential location choice and travel mode and a simultaneous decision of residential and workplace location choice. Today, there is also a growing tendency to jointly model residential location choice with several other lifestyle decisions with the most commonly modeled joint decisions are coupling residential location with workplace location, car/bicycle ownership, activity patterns, and transport mode. An example of a model integrating six lifestyle dimensions is a model incorporating residential location choice, work location choice, automobile ownership, commuting distance, commute mode, and number of stops on commute tours. These joint choices are typically represented as hierarchical choice processes, which carry both the problem of the defining the choice order, which can be solved by means of latent market segmentation which respect to the choice structure, and a rapidly increasing realm of possible choices, in particular in the case of residential and workplace choices.

Population Strata

The majority of the studies that focused on population groups covered differences among groups. The differentiation among groups was conducted according to income and number of employees in the household, ethnicity, and gender. A few studies concentrated on residential choice of specific groups, particularly elderly, youth migrating from rural to urban areas, dual-earner households, students, and high-skill knowledge-workers.

The main population strata of interest from the 1980s until the beginning of the new millennium were mainly the move from the traditional bread-winner housekeeper household type to dual-earner households, and the role of job perks and the ability to telecommute on residential mobility and location decisions of the household. More recently, research in this direction has taken a turn toward understanding self-selection processes guiding joint residential and lifestyle choices. In the 1970s, researchers were mainly concerned with spatial segregation on the basis of ethnicity and socioeconomic status and the impact of such segregation on equity. Nowadays, while equity is still an important issue, the main discussion is voluntary self-selection in correlation with other lifestyle choices. For planners, understanding residential self-selection is important for improving the representation of the household characteristics in activity-based models and integrated transport and land-use models. For policy-makers, self-selection may lead to voluntary segregation and generate spatial and intergenerational inequities. From the beginning of the millennium, the growing role of education in the economic basis of regions and cities and the positive role of students as revitalization catalyzers versus the negative externalities of large masses of students in medium-sized cities (i.e., studentification) have motivated research toward understanding the residential choice of students in local housing markets. Richard Florida's theorem of learning regions sparked interest in migration of knowledge-workers as a growth engine for knowledge-cities (Florida, 2002). Following Florida's theorem that knowledge-workers are characterized by high residential mobility and high sensitivity to lifestyle, new research has surged to understand international and regional migration of knowledge-workers as well as their intra-metropolitan location choices. This research stream is closely tied to innovation and regional development, with the main aim of understanding the attraction and retention of young, highly skilled knowledge-workers. This research line is important also because it is deviating from the traditional view of a tight knit between residential location and job opportunities toward a more sustainable approach considering also social and environmental perspective. It has also contributed to extending the definition of urban amenities from "hard" physical amenities to "soft" cultural amenities and has highlighted the role of otherwise neglected city features such as nightlife, street art, street ambience, architecture, walk ability, green scenery, and sport amenities.

Decision-Makers

Households are the most common decision-maker unit in transportation. Namely, the main assumption is that the household can be regarded as a non-separable unit with a single utility function. This representation is based on the assumption that once negotiations end, the household can reach an agreement regarding the desired location choice and that the selected choice largely reflects a consensus. It is also based on the assumption that the household has a single utility function upon which all members agree. While this could fit to some extent traditional breadwinner-housekeeper households, this assumption loses its validity in two-earner households and power couples as well as with the growing role of childcare and education. Starting from the end of the 1990s, the commuting needs of two-earner households were incorporated both into the theoretical specification of the household location geometry and into the utility function. The new perspective of household no longer views the household as an integral unit but rather as a group of individuals with different objectives that could vary in their degree of conformity. This perspective is particularly important in the joint modeling of residential location and lifestyle, because household members can significantly vary in their

lifestyle choices. Therefore, the household unit is now viewed as a group of decision-makers that needs to negotiate their interests. Thus, instead of a single utility function, a conjoint analysis of preferences and group dynamics are explored.

Decision Rules and Model Form

The models applied to residential choice can be classified according to their underlying choice strategy, choice set formation process, and model form. As a general rule, residential choice models follow the utilitarian approach and assume a compensatory choice process guided by the principle of additive utility. This approach carries important advantages with respect to the mathematical representation of residential choice: ease of estimation, reproduction of market shares, ability to account for spatial correlation patterns and population heterogeneity, ability to model jointly correlated choices, and the possibility of model estimation from standard data collected by municipal and government authorities.

Among the estimated models, the most prominent one is the multinomial logit (MNL). Roughly one-third of the studies attempted to resolve the constraints of the MNL: independence of irrelevant alternatives, population homogeneity and homoscedasticity. To resolve the first constraint, the estimation of the Nested Logit is the most popular approach, but other Generalized Extreme Value models have also been estimated, with the most advanced being the Spatially Correlated Logit (Bhat and Guo, 2004) and the Generalized Spatially Correlated Logit (Sener et al., 2011). The second constraint of population preference homogeneity was resolved with various application of mixed logit with random coefficients, and both constraints have been resolved with mixed versions of the Generalized Spatially Correlated Logit. Joint choice models are estimated as means to address residential choice and other lifestyle decisions that guide residential self-selection. This line of models started by combining the MNL with ordered response models (MNL-OR) for modeling the combination of residential location and car ownership and continues with a series of models combining the MNL with Discrete-Continuous Extreme Value (MDCEV) models for representing the joint decision of residential choice, car ownership, activity, and time use patterns.

Only a few studies addressed non-compensatory decision rules, although research efforts steadily continued from the 1980s and until this decade. This approach stems from the bounded rationality approach criticizing the assumptions of full information and the ability to even proximally calculate the utility of alternatives deriving from a large choice sets. Due to the complexity embedded in the efficient estimation of non-compensatory and two-stage models, as well as the intensive data requirements, this line of research is slowly advancing. The first non-compensatory model to be applied to residential choice is the Elimination-By-Aspects model, which was applied to an example of three alternatives, because of the model limitation in handling large choice sets. Other studies estimated two-stage semi-compensatory choice models, which combine probabilistic choice set formation based on random thresholds followed by an MNL model based on Manski's approach of jointly estimating the probability to choose a choice set, and the probability of an alternative to be chosen from the choice set. The models' main focus is the representation of threshold selection as a function of individual characteristics and jointly accounting for the selection of multiple thresholds, both of continuous and discrete nature. The focus on choice set selection has started with socioeconomic constraints to account for threshold selection of location characteristics such as dwelling unit price, attributes, and proximity measures. Starting from 2015, the focus is shifting toward understanding the role of lifestyle, attitudes, and norms in the specification of threshold selection, with the corresponding model structure being the generalized heterogeneous data model (GHDM) combining latent variables underlying the joint choice of multiple thresholds from various types. The modeling advantages and the superiority of the mathematical representation of choice set formation and threshold selection in residential choice compared to the traditional models that assume a single unified choice set have been consistently proven using various data sources.

Data Sources for Estimation

The vast majority of residential choice models have been estimated by using revealed-preference (RP) data with a limited number of studies using the stated-preference (SP) or the combined RP-SP approach. The advantage of RP surveys is that they reflect actual choices in real-world situations. However, RP surveys are subject to two disadvantages when applied to residential choice. The first disadvantage lies in the difficulty to observe revealed residential choice preferences at the point of decision, since residential choice decisions are infrequently made during the average household life cycle and the residential decision process may span over long periods (from several weeks to several years). The second disadvantage consists in the limited ability to represent trade-offs among attributes, since RP surveys are limited to existing residential choice bundles and thus may not adequately represent the full range of attribute values. The advantage of SP surveys is the relative ease of data collection during the choice task and the possibility to control the range of attribute values. However, their disadvantage in the context of residential choice is that they are subject to hypothetical response bias, since individuals do not bear the consequences of their chosen alternative. Early research efforts involved mainly recruiting current home buyers/renters or households who have recently moved, for example, through newly built neighborhoods and real-estate agents. Existing data sources that are commonly used for residential choice model estimation are census data, national travel habit surveys, economic data sources, and regional or national housing surveys. An example is the American Housing Survey (AHS), sponsored by the Department of Housing and Urban Development. Because of the difficulty of observing revealed residential choice preferences at the point of decision, most of the studies consider the actual place residence at the time of the analysis. These data sources are mostly employed for cross-sectional analysis or pooled panel data, thus ignoring the time dimension. Moreover, households choose residences at different time points, so the same location may change over time. The disadvantage of this approach is that changes in environmental and societal conditions between the time of the decision and the

time of the analysis are not accounted for. Approximating the choice at the point of the decision can be done by restricting the analysis to choices of recent movers or new residents or alternatively referring to moving status.

Implementation in Transport Models

Residential choice models are not an integral part of traditional four-stage transportation models. In four-stage models, the location of the households is predetermined according to population forecasts on the basis of socioeconomic trends. In four-stage transport models, the data units of analysis are traffic analysis zones. In contrast, residential choice models are a key component in agent-based activity-based models and in integrated dynamic land-use and transportation models. In agent-based models, the unit of analysis comprises individual agents representing individual decision-makers. Synthetic populations representing current and future generations are compiled, consisting of attributing socioeconomic characteristics, residential location, car ownership, and activity patterns to each agent. The travel choices of each agent are derived from its inherent characteristics. While the generated synthetic population follows predetermined population distributions, unlike trend forecasting, the synthetic population is simulated at the individual level. The joint estimation of residential choice, car ownership, activity, and time use is a primary requisite for modeling derived travel demand. When short-term simulation is performed, residential choice is employed only to define the preconditions of the synthetic population. When long-term simulation is performed, integrated land-use and transport dynamics take place, through the iterative change of residential choice, job location and travel choices, their evaluation, and reassessment. The long-term integrated land-use and transportation approach allows understanding codependencies between transport and land-use development, understanding housing price fluctuations, and predicting urban processes of expansion, deterioration, and renewal.

Further Research

Online sources for modeling residential choices: RP residential choice studies did not manage to observe residential choice at the point of purchase/rent decision. The majority of the studies referred to the actual residence at the time of the analysis as the choice at the decision point. Such an analysis is retrospective and susceptible to unknown changes in the location specification of each household. Moreover, the choice set used at the time of the decision is unknown to the researchers. With the advancement of online real-estate databases, public access to real-estate transaction records, and computer-based databases of real-estate agents, analysis of real-estate transactions at the time of the decision is becoming a realistic option, which will vastly improve the representation of residential choice processes.

Lifestyle, hedonism, attitudes, and norms: Residential choice has been explained mostly from the functional gain (utilitarian) perspective. However, lifestyle, attitudes, and norms, as well as normative versus hedonic goal framing, are latent constructs that motivate housing choice and their affect can interact or override pragmatist housing choice. Lifestyle components represented by the individual's activity pattern are known to be correlated with residential choice. Both green attitudes and luxurious lifestyle as normative and hedonic goal framing have also been suggested as location choice motivators but were measured only indirectly through physical indicators. Further exploration of this topic is merited to better understand the residential choice of young people, people with high residential mobility and economic resources, and home renters. This line of research will also help understanding self-selection processes and counterintuitive residential choices such as in the case of long commuters.

Shared economy and rent-own models: Shared economy brings about new ways of community living and ownership patterns. One example is the Scandinavian rent-own model of shared apartments and houses, in which a few individuals or families from different households decide to share a dwelling unit, which is common space. The tenants are chosen based on group agreement and the apartments are designed from multiple households. The American example is the concept of companies offering fully furnished apartments in a community environment for short-term rental, which replaces the traditional rent agreements. Both models include private and commonly shared spaces, which redefine housing needs. They also redefine sharing as an amenity rather than a burden. From the residential location choice modeling, these new models are important because they allow affordable accessibility to the city center, where they are commonly located. Thus, understanding the trade-off between central shared dwelling units to suburban private dwelling units.

Soft location amenities: While much is known about the impact of physical location amenities and public services, very little is known regarding the impact of "soft" amenities including cultural and creative amenities, street art and vibrant street scene, digital amenities, sense of community and belonging, and coolness and networking factors. The literature on regional innovation and growth has long established valid indicators for "soft" amenities and linked "soft" amenities to economic growth, urban development, and regional attraction. The next step is to embed these amenities into residential choice models.

References

- Alonso, W., 1964. *Location and Land Use: Toward a General Theory of Land Rent*. Harvard University Press, Cambridge MA.
- Bhat, C.R., Guo, J., 2004. A mixed spatially correlated logit model: formulation and application to residential choice modeling. *Transp. Res. Part B* 38 (2), 147–168.
- Florida, R., 2002. *The Rise of the Creative Class*. Basic Books, New York.

- Lee, B.H., Waddell, P., 2010. Residential mobility and location choice: a nested logit model with sampling of alternatives. *Transportation* 37 (4), 587–601.
- McFadden, D., 1978. Modeling the choice of residential location. *Transp. Res. Record* 673, 72–77.
- Sener, I.N., Pendyala, R.M., Bhat, C.R., 2011. Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior. *J. Transp. Geogr.* 19 (2), 294–303.

Further Reading

- Adnan, M., Pereira, F.C., Azevedo, C.L., Basak, K., Lovric, M., Raveau, S., Zhu, Y., Ferreira, J., Zegras, C., Ben-Akiva, M., 2016. SimMobility: a multi-scale integrated agent-based simulation platform. *Transportation Research Board 95th Annual Meeting*, Washington, DC.
- Barrios-Garcia, J.A., Rodriguez-Hernandez, J.E., 2007. Housing and urban location decisions in Spain: an econometric analysis with unobserved heterogeneity. *Urban Stud.* 44 (9), 1657–1676.
- Bhat, C.R., 2015. A comprehensive dwelling unit choice model accommodating psychological constructs within a search strategy for consideration set formation. *Transp. Res. Part B* 79, 161–188.
- Bhat, C.R., Astroza, S., Bhat, A.C., Nagel, K., 2016. Incorporating a multiple discrete-continuous outcome in the generalized heterogeneous data model: application to residential self-selection effects analysis in an activity time-use behavior model. *Transp. Res. Part B* 91, 52–76.
- Borgers, A., Timmermans, H., Veldhuisen, J., 1986. A hybrid compensatory-noncompensatory model for residential preference structures. *J. Housing Built Environ.* 1 (3), 227–234.
- Chen, C., Chen, J., Timmermans, H., 2009. Historical deposition influence in residential location decisions: a distance-based GEV model for spatial correlation. *Proceedings of the 88th Annual Meeting of the Transportation Research Board*, Washington, DC.
- Earnhart, D., 2002. Combining revealed and stated data to examine housing decisions using discrete choice analysis. *J. Urban Econ.* 51 (1), 143–169.
- Freedman, O., Kern, C.R., 1997. A model of workplace and residential choice in two-worker households. *Reg. Sci. Urban Econ.* 27 (3), 241–260.
- Frenkel, A., Bendit, E., Kaplan, S., 2013. Residential location choice of knowledge workers: the role of amenities, workplace and lifestyle. *Cities* 35, 33–41.
- Galilea, P., de Dios Ortúzar, J., 2005. Valuing noise level reductions in a residential location context. *Transp. Res. Part D Transp. Environ.* 10 (4), 305–322.
- Habib, K.N., Kockelman, K.M., 2008. Modeling the choice of residential location and home type: recent movers in Austin, Texas. *Proceedings of the 87th Annual Meeting of the Transportation Research Board*, Washington, DC.
- Kaplan, S., Bekhor, S., Shifan, Y., 2012. Development and estimation of a semi-compensatory model with a flexible error structure. *Transp. Res. Part B* 46, 291–304.
- McFadden, D.L., 1978. Modeling the choice of residential location. In: Karlqvist, A., et al. (Eds.), *Spatial Interaction Theory and Residential Location*. North Holland, Amsterdam, The Netherlands, pp. 75–96.
- Molin, E., Oppewal, H., Timmermans, H., 1999. Group-based versus individual-based conjoint preference models of residential preferences: a comparative test. *Environ. Plan. A* 31, 1935–1947.
- Quigley, J.M., 1985. Consumer choice of dwelling neighborhood and public services. *Reg. Sci. Urban Econ.* 15 (1), 41–63.
- Tiwari, P., Hasegawa, H., 2004. Demand for housing in Tokyo: a discrete choice analysis. *Reg. Stud.* 38 (1), 27–42.

Transport Demand Management

Feiyang Zhang, Becky P.Y. Loo, Department of Geography, The University of Hong Kong, Hong Kong

© 2021 Elsevier Ltd. All rights reserved.

Introduction	537
The State of the Art of Transport Demand Management	537
Addressing “How Many Trips are Made?”	538
Promoting Mixed-Use Development	538
Telecommunications and E-Working	539
Addressing “When are Trips Made?”	539
Traffic Information	539
Flexible Work Schedule	539
Addressing “Where are Trips Made?”	539
Congestion Charging	539
Better Job-Housing Balance	540
Distance-Based Charging	540
Addressing “How are Trips Made?”	540
Taxation	540
Vehicle Quota System	540
Parking Controls and Pricing	540
Encouraging Ride-Sharing (Carpooling)	540
Public Transport Subsidization	541
Transit-Oriented Development	541
Urban Design and Site Design	541
Park and Ride Facilities	541
Mobility as a Service	541
School and Tourist Transport Management Programs	541
Conclusion and Discussion	542
References	542

Introduction

The adoption of automobiles in modern society has caused traffic congestion in cities across the world, leading to problems such as long commutes, high-energy consumption, sedentary lifestyle, air pollution, and climate change. As the urban population continues to increase, accommodating the mobility of people and goods within the limited space of cities having highly competitive land uses has become a huge challenge for local governments.

Traditionally, countries such as the United States had tried to tackle the problem by increasing the “supply” of road capacity, by building more highways, bridges and tunnels to alleviate traffic congestion. While some developing countries, such as China, had been following this model, the approach had been largely criticized during recent years because the increased capacity actually induced more demand (Lee et al., 1999), resulting in same or more severe congestion. Under this background, the concept of “Transport Demand Management” (TDM, in many cases also referred as “Travel Demand Management”) was introduced since the 1970s in the United States (Meyer, 2016). As is suggested by its name, TDM promotes a new mindset to deal with the traffic congestion by focusing more on the better management of the “demand” side of transportation. The following sections provide an overview about the categories, measures, effects, and some selected case studies of TDM.

The State of the Art of Transport Demand Management

Ferguson (2000) identified three major categories of TDM solutions as “voluntarism”, “markets”, and “regulation” (Ferguson, 2000, pp. 57–270). His analysis has largely focused on the situation of US cities. Ison and Rye (2008) focused on the actual implementation and impact analysis of a range of TDM measures. They further refined the categories of TDM solutions into five, that is, “economic”, “land use”, “information for travelers”, “substitution of communications for travel,” and “administrative” (Ison and Rye, 2008, p. 1). The categorization is helpful but some categories like the “information for travelers” can be rather narrow. More recently, Meyer (2016) listed TDM goals, objectives, and performance measures from a more practical perspective of transport planning. Numerous other researchers have also looked into specific measures deemed successful in cities such as Singapore and London (Banister, 2008; May, 2004; Seik, 2000).

Table 1 General categories and effects of TDM measures

Categories	Measures	How many trips are made?	When are the trips made?	Where are the trips made?	How are the trips made?
		Reduce the number of trips	Shift trips from to nonpeak hours	Reduce trip distance and avoid congestion hotspot	Encouraging sustainable modes of transportation
Economic	Taxation	✓			✓
	Congestion charging		✓	✓	✓
	Vehicle quota system				✓
	Parking controls and pricing		✓		✓
	Encouraging ride-sharing (carpooling)				✓
	Public transport subsidization				✓
	Distance-based charging			✓	✓
Urban form and land use	Development guidance for transit-oriented and mixed-use development	✓		✓	✓
	Job-housing balance			✓	
	Urban design and site design that promote transit use, walking, and cycling				✓
	Park and ride facilities				✓
Technological	Traffic information		✓	✓	✓
	Telecommunications (substitution for travel)	✓			
	E-working	✓			
	Mobility as a service (MaaS)				✓
Organizational	Flexible work schedule		✓		
	School transport management programs		✓		✓
	Tourist transport management programs		✓		✓

Most of the common measures of TDM are implemented in the form of public policy deeply rooted in the local cultural and political context of different countries. Therefore, the satisfactory implementation of one TDM measure in one country does not necessarily guarantee its success in another country. With this regard, it is more important to understand the nature of people's travel demand, as most human beings have the need to move themselves and their goods across space for the purpose of daily life. Specifically, four fundamental questions are asked to structure the analysis of TDM in this article: How many trips made? When are trips made? Where are trips made? How are trips made? These guiding questions are listed in [Table 1](#). Correspondingly, the four main goals of TDM measures are ways of reducing the number of trips, shifting trips from peak hours to other time, reducing average trip distance and avoiding congestion hotspot, as well as encouraging more sustainable modes/means of transportation. Specific policies from both developed and developing countries will be included in each section to provide an international perspective. In order to further enrich the analysis, we suggest a more general categorization of TDM measures, list out some common measures under each category, and highlight their relationships in answering the four key questions.

Addressing “How Many Trips are Made?”

When there is a high degree of concentration of people and goods, as is the case in most large cities today, it would be helpful to alleviate congestion by reducing the absolute number of trips made in the society. This can be achieved by more integrated land use and transport planning, and by promoting new work styles, such as e-working.

Promoting Mixed-Use Development

Compared to other short-term TDM measures, the management of land use is considered to be more of a long-term solution of easing congestions ([Meyer, 2016](#)). Generally, high-density, mixed-used urban environment would allow the citizens to fulfill their daily needs with the minimum number of trips per day with trip chaining at the same or nearby locations. For instance, it is typical for Hong Kong citizens to find restaurant and grocery stores at the ground floor of their apartment, meaning he/she will be able to fulfill the need of dining or buying groceries at his/her place of residence without making additional trips, while it is usual for those who live in detached single-family housing neighborhood in the United States to make a single-purposed trip to achieve the same purposes.

Telecommunications and E-Working

Telecommunications have been providing alternatives for traveling since the invention of fax machines (Ferguson, 2000). With the fast development of Internet and smartphones, many necessary trips in the past could now be simply replaced by a couple of clicks on smartphone applications. For example, one could easily use mobile banking service offered by most of the banks today to manage his/her accounts, and does not necessarily need to go the local bank branch for the same purpose.

It is also typical for companies today to hold teleconferences, which reduce the need for employees to travel to the same place. Meyer (2016) described a telework program operated by US Patent and Trademark Offices to allow the agency's employees to e-work almost every week. Other similar programs in the United States have reported both reduction of drive-alone trips and increase of employee's productivity (Meyer, 2016).

Although face-to-face interactions are still important (Hall, 2003), the fast advancement of technology provides some degree of substitution for travel. With the new generations of 5G and virtual reality technologies, it is reasonable to believe the experience of e-working would be further improved, which will attract more people to work at home, especially for those who need to make super-long commute in metropolitan areas. Therefore, there is a huge potential for telecommunications and e-working to help reduce the number of trips and ease traffic congestion.

Addressing "When are Trips Made?"

The capacity of transport infrastructure usually faces the greatest challenge during the peak hours. Therefore shifting some trips from peak hours to other regular hours would help alleviate the peak-hour congestion. Traffic information, as well as flexible work schedule, can help to achieve this goal.

Traffic Information

Information and communication technologies (ICT), or more specifically, services such as Google Map is now able to provide historic and real-time traffic information for travelers, allowing them to make smart decisions about the departure time of their trip to avoid time periods when congestion happens, and consequently spend less time traveling (Cohen-Blankshtain and Rotem-Mindali, 2016). As some informed users would shift their trips to noncongested periods, it is reasonable to argue that the peak-hour pressure on transportation infrastructure capacity will be alleviated by services of this sort.

Flexible Work Schedule

Since most of the commuting happens during peak hours, flexible work schedule can help reduce peak-hour traffic volume. In the United States, flexible work hours have been one of the major TDM programs (Meyer, 2016). A typical flexible work hours program can contain "core periods" and "flex periods" (Ferguson, 2000, p. 88), meaning that employees must be present at the office during designated core work hours, but could be absent during other time during the day. Staggered work hours is similar to flexible work hours in that employees may arrive or leave the work place at different time of the day, but this approach is rather unique in that it is the employer who decides and arranges the exact work shifts time, instead of the employee (Ferguson, 2000). Another popular approach is compressed weeks, meaning that employees could work more hours per day and work less number of days per week (Ferguson, 2000; Meyer, 2016).

As ICT continues to improve human connections without being together physically, flexible work schedule can become more popular as people may very well work with smart devices even when they are traveling (Wang and Loo, 2017). For instance, one may just write or reply to email on a bus to home, although he/she leaves office earlier than other coworkers.

Addressing "Where are Trips Made?"

Where people make their trips matters because recurrent congestion is the most common at some areas of the city. In addition, whether people have accessibility to employment and other services in nearby neighborhood will affect their trip distance. Congestion pricing can be used to reduce the volume of traffic in congestion-prone areas. Similarly, a better job-housing balance at the city level can also help reduce travel distance. In addition, distance-based charging can be applied to further discourage long trips.

Congestion Charging

Cordon-based congestion charging (also named congestion pricing) means that vehicles entering a designated area are charged. This policy is commonly applied to reduce the volume of traffic in the urban center, which is usually a congestion hotspot. An example of congestion pricing is the electronic road pricing (ERP) system in Singapore. If a vehicle enters the "Restricted Zone" (Seik, 2000, p. 36) during certain time periods, fees will be automatically charged with the help of the communication between the in-vehicle unit and the ERP system at the entry points of the restricted area (Seik, 2000). After the implementation of this system, the traffic flow fell by 20% to 24% during weekdays in the designated area, which covers the CBD of the city (Seik, 2000).

It is important to note that the ERP system were made possible by ICTs such as "radio-frequency, optical-detection, imaging, and smart-card technologies" (Seik, 2000, p. 36), which allows much more flexibility such as changing rates of charging throughout the day, when compared to the previous area license scheme in Singapore (May, 2004).

Similar policies have also been implemented in European countries. According to [Banister \(2008\)](#), London is the first city in Europe to practice congestion charging. First implemented in 2003, the measure reduced traffic level by 20% in the charging area ([Banister, 2008](#)).

Better Job-Housing Balance

Many modern cities have their employment centers located in the centers of the urban area, with most of the residence located in the peripheral area. As the metropolitan areas continue to grow, more and more people are involved in “super commute”. Therefore it is important to plan for job-housing balance to provide better accessibility to employment opportunities, especially for the peripheral area, and to reduce commuting time and distance.

Job-housing balance, however, is not listed in most existing literature as a measure of TDM, partly because it is more related to the field of land use planning. However, it is important to point out that, similar to transit-oriented development (TOD), coordinated land use planning can have great impact on transport demand and travel behavior, especially for developing countries where rapid urbanization is happening.

Distance-Based Charging

According to the [Ministry of Transport \(2019\)](#) in Singapore, the city is planning to integrate the distance-based charging mechanism into its current ERP system in 2020. This means drivers will be charged based on the distance they drive on congested road sections, whereas the original system charge the driver the same amount of fees regardless of their travel distance in the congested area ([Ministry of Transport, 2019](#)). It is reasonable to expect that this approach will further discourage long-distance travel in congested area, especially during peak hours. Again, this vision will be supported with the advanced ICT such as the global navigation satellite system ([Ministry of Transport, 2019](#)).

Addressing “How are Trips Made?”

In recent years, much effort has been spent on transport planning in different countries to promote the use of alternative modes of transport, instead of the use of private cars. As land resources become scarcer in big cities, the government needs to promote transport modes that have higher “space efficiency” ([Litman, 2014](#)). For the same number of people to move around, it is obvious that private cars take up much more road space, as well parking space, when compared to other modes of transportation, such as transit, cycling, and walking. Common measures to encourage the sustainable modes of transports are listed as follows.

Taxation

[Potter \(2008, p. 14\)](#) specified three types of taxation related to the use of automobiles, which are “tax on the initial purchase of a vehicle”, “ownership of vehicles”, and “the use of vehicles”. Taxation increases the cost of using private cars, and consequently promotes a modal shift from car to other modes of transportation. According to [Potter \(2008\)](#), this measure is widely applied in most of European Union countries, but are used more for the promotion of cleaner car technologies and less for the management of travel demand.

Vehicle Quota System

Compared to western countries, a rather strict policy called vehicle quota system has been implemented in Singapore, which directly limits the increase of the vehicle ownership by setting a predetermined “ceiling” each year ([Seik, 2000](#)). Similar policies have also been implemented in major cities in China, such as Shanghai, and have been proven to be effective ([Xiao et al., 2017](#)).

Parking Controls and Pricing

The use of private cars always leads to the problem of finding a parking space in crowded cities. [Shoup \(2005\)](#) pointed out the negative impacts of providing free roadside parking, as was the typical case in major cities in the United States. Furthermore, [Shoup \(2008\)](#) promoted an approach to charging the fee for parking space, which varies by time and location to resolve the problem of parking space shortage. Using this method, some of the original drivers will be discouraged to drive and will use other modes of transport instead, especially when provided with other incentives such as free transit pass ([Shoup, 2008](#)).

Encouraging Ride-Sharing (Carpooling)

Furthermore, in places where automobile-oriented infrastructure is already in place, TDM measures may be taken to promote carpooling to reduce the number of trips. A typical approach is to give priority to high occupancy vehicles, such as dedicating one lane on the highway for cars that has at least two passengers. In addition, there are also public agencies and private firms that provide “ride-matching services” ([Meyer, 2016, p. 648](#)).

Currently, a popular option for ride-sharing is through a common platform like UberPool, provided by the transport network company (TNC) known as Uber, which offers a lower price if the passenger is willing to carpool with other people going in a similar direction ([Uber, 2019](#)). As both the public and private sectors are now envisioning a future with autonomous vehicles, ride-sharing would be very important in reducing single-occupancy vehicle trips in future cities. It is possible that TNCs may play a key role in the promotion of ride-sharing if autonomous vehicles are to be operated by those types of companies in the future.

Public Transport Subsidization

Although some scholars have challenged the justifications of public transport subsidies, public transport is now heavily subsidized both in developed countries such as the United States (Van Reeve, 2008) and in developing countries such as China. Lowering the fares or public transport by subsidies helps encourage people to use transit instead of automobiles for their daily travel. This approach also helps to promote social equality by providing affordable transport options for socially disadvantaged groups (Preston, 2008). It is reasonable to argue that a strong local economy is required to provide sustainable funding to keep the subsidized public transport system running once it is built.

Transit-Oriented Development

TOD means putting intensive commercial and mixed-used development near or on top of transit stations or corridors. People living in such kind of environment may easily fulfill their daily needs around transit stations, therefore are more willing to use transit as their means of transport.

This type of transport-land use coordination can be promoted through regional planning organizations, such as the Metropolitan planning organization (MPO) in the United States. For example, the Southern California Association of Governments is the largest MPO in the United States, which is responsible for making regional transport and land use policies for six counties in California (Southern California Association of Governments, 2019). Through making the regional transportation plan/sustainable communities strategy, the agency aims to channel the new land use development to the transit station or transit corridors (Southern California Association of Governments, 2019).

Urban Design and Site Design

Loo and du Verle (2017, pp. 55–66) pointed out that future TOD strategies should not only focus on the patronage of transit system itself, but also consider the design features such as “greenery” and “vibrancy” near transit stations. Urban design and site design features can influence people’s travel behavior in a very subtle but significant way. Based on the above understanding, government has started to make design guidelines that encourage the use of the alternative modes of transportation.

In Europe, countries such as the United Kingdom and Sweden have developed planning and design guidelines that require the promotion of sustainable modes of transport during the project planning and approval process (Black and Shreffler, 2010). In London, specific methods are described in the Streetscape Guidance—a set of street design guidelines to create high-quality street environment and to make walking, cycling, and transit use more attractive (Transport for London, 2019).

In New York, a policy document named Active Design Guidelines is used to encourage active transport (Meyer, 2016). While the design guidelines focus more on improving public health, the promotion of walking and cycling will also help to manage traffic demand and reduce the dependency on automobiles (Center for Active Design, 2010).

Park and Ride Facilities

Park and ride facilities give the drivers opportunities to park their vehicles and switch to public transport, usually at the edge of urban areas (Meek, 2008). This measure is mainly used for promoting the use of public transport and has been widely applied in countries such as the United States and the United Kingdom (Lam et al., 2001). As large transport hub continues to emerge in big cities, this concept may further evolve at the age of autonomous driving, such as giving people opportunities to switch from autonomous automobiles to transit when traveling into urban center.

Mobility as a Service

In recent years, there are increasing discussions about the mobility as a service (MaaS), which aims to provide mobility options as services that could be subscribed through integrated online platforms. The concept was first introduced by Sampo Hietanen, who defined MaaS as “a mobility distribution model in which a customer’s major transportation needs are met over one interface and are offered by a service provider” (Hietanen, 2014, p. 2). Under this concept, people’s daily travel demand would be fulfilled by an integrated package of transportation services, which guarantee the level of the reliability and convenience comparable to having a car, which as a result will reduce citizen’s car ownership. It is important to acknowledge that this new concept is largely enabled by the advancement of ICT technologies. Jittrapirom et al. (2017) summarized nine core characteristics of MaaS, almost all of which are enabled by current data integration technology, location-based services, as well as web and mobile interfaces technologies (Hietanen, 2014). This paradigm has gained popularity since it promises to bring benefits to different parties, including transport users, private service providers, and public sectors, by taking full advantage of the ICT technology to better match supply and demand (Hietanen, 2014). Although limited research had been done to investigate the real impact of the pilot MaaS programs around the world (KiM Netherlands Institute for Transport Policy Analysis, 2018), it is reasonable to expect the potential of well-designed MaaS programs to tackle automobile-dependency and alleviate congestion.

School and Tourist Transport Management Programs

As many trips are generated by students traveling between home and school, specific programs are also developed to better manage this type of travel activities. In the United States, various measures have been implemented at different school levels to reduce automobile dependency, such as providing free access to transit, mobility education programs, etc. (Washington State Transportation Commission, 2009).

Tourist transport management programs can also be used to alleviate traffic congestion while accommodating the mobility needs of tourist. For example, based on the prediction from travel forecasting model, the Salt Lake City Olympic Committee provided a transportation guidebook for people visiting the city for the Winter Olympic Game, which specified tips on how they could navigate through the city efficiently (Transportation Research Board, 2004).

Conclusion and Discussion

By answering the four fundamental questions, common TDM measures being implemented in cities across of the world were discussed. As the problem of congestion becomes more severe in metropolitan area, learning from the successful experience of managing transport demand will be helpful. These measures, though described separately, should be better integrated into holistic policy packages (Meyer, 2016) and fundamental city transformations at different spatial levels (Loo and Tsoi, 2018).

Although technological TDM measures may be listed as a separate category, they actually support and underlie different TDM measures. For instance, congestion pricing, distance-based charging, and ride-sharing would not be successful without the support of the state-of-art ICT. As Gillen (2008) has pointed out, ICT has helped to provide information, direct traffic, automate tasks, and monitor the performance of the road-pricing efforts. Looking ahead, the next generation of ICT (such as the 5G technology), combined with robotics, automation, and artificial intelligence technology, are likely to bring new tools, as well as new challenges, to the field of TDM.

With all the technological advancement mentioned above, the world is expected to enter an age of ubiquitous connectivity, potentially resulting in a dramatic change of people's lifestyle. Some traditional trips can be further replaced by e-working or other online activities. Some trips can be further directed to nonpeak hours as work schedules could become increasingly flexible. On the other hand, activities that need frequent face-to-face interaction are still likely to concentrate in urban centers (Hall, 2003), and cities would still be challenged with the problem of space scarcity in the foreseeable future. Therefore encouraging people to switch to transport modes with higher space efficiency (i.e., transit, cycling, walking) is necessary. All these may be achieved by thoughtful urban planning and economic measures, as well as innovative MaaS programs.

Last but not least, as the transport industry is increasingly focusing on the implementation of autonomous driving, some scholars and practitioners (Calthorpe and Walters, 2017; Litman, 2019) have warned about the dystopia scenario of implementing autonomous driving technology, where convenient autonomous vehicle services can induce people to travel more and travel longer distance, and consequently leading to urban sprawl and congestion problem at a larger scale than what we witness today. Therefore regulations by the public sector would be essential to harness the benefits of these types of technology and to avoid the potential negative impacts. With this regard, careful research and policy design for TDM will continue to be an important field in the years to come.

References

- Banister, D., 2008. The land use and local economic impacts of congestion pricing. In: Ison, S., Rye, T. (Eds.), *The Implementation and Effectiveness of Transport Demand Management Measures, An International Perspective*. Ashgate Publishing Ltd, Aldershot.
- Black, C., Shreffler, E., 2010. Understanding transport demand management and its role in delivery of sustainable urban transport. *Transport. Res. Rec.* 2163 (9), 81–87.
- Calthorpe, P., Walters, J., 2017. Autonomous Vehicles: Hype and Potential. Available from: <https://urbanland.uli.org>.
- Center for Active Design, 2010. Active Design Guidelines. Available from: <https://centerforactivedesign.org>.
- Cohen-Blankshtain, G., Rotem-Mindali, O., 2016. Key research themes on ICT and sustainable urban mobility. *Int. J. Sust. Trans.* 10 (1), 9–17.
- Ferguson, E., 2000. *Travel Demand Management and Public Policy*. Ashgate Publishing Ltd, Aldershot.
- Gillen, D., 2008. The role of intelligent transportation system (ITS) in implementing road pricing for congestion management. In: Ison, S., Rye, T. (Eds.), *The Implementation and Effectiveness of Transport Demand Management Measures, An International Perspective*. Ashgate Publishing Ltd, Aldershot.
- Hall, P., 2003. The end of the city? "The report of my death was an exaggeration". *City* 7 (2), 141–152.
- Hietanen, S., 2014. 'Mobility as a Service' – the new transport model? *Eurotransport* 12 (2), 2–4.
- Ison, S., Rye, T., 2008. Introduction: TDM measures and their implementation. In: Ison, S., Rye, T. (Eds.), *The Implementation and Effectiveness of Transport Demand Management Measures, An International Perspective*. Ashgate Publishing Ltd, Aldershot.
- Jittrapirom, P., Caiati, V., Feneri, A.M., Ebrahimigharehbaghi, S., Alonso González, M.J., Narayan, J., 2017. Mobility as a service: a critical review of definitions, assessments of schemes, and key challenges. *Urban Plan.* 2 (2), 13–25.
- Kim Netherlands Institute for Transport Policy Analysis, 2018. Mobility-as-a-Service and Changes in Travel Preferences and Travel Behaviour: A Literature Review. Available from: <https://english.kimnet.nl/>.
- Lam, W.H., Holyoak, N.M., Lo, H.P., 2001. How park-and-ride schemes can be successful in Eastern Asia. *J. Urban Plan. Dev.* 127 (2), 63–78.
- Lee Jr., D.B., Klein, L.A., Camus, G., 1999. Induced traffic and induced demand. *Trans. Res. Rec.* 1659, 68–75.
- Litman, T., 2014. Economically Successful Cities Favor Space-efficient Modes. Planetizen. Available from: <https://www.planetizen.com/>.
- Litman, T., 2019. Autonomous Vehicle Implementation Predictions: Implications for Transport Planning. Available from: <https://www.vtpi.org>.
- Loo, B.P.Y., Tsoi, K.H., 2018. The sustainable transport pathway: a holistic strategy of five transformations. *J. Trans. Land Use* 11 (1), 961–980.
- Loo, B.P.Y., du Verle, F., 2017. Transit-oriented development in future cities: towards a two-level sustainable mobility strategy. *Int. J. Urban Sci.* 21 (S1), 54–67.
- May, A.D., 2004. Singapore: the development of a world class transport system. *Trans. Rev.* 24 (1), 79–101.
- Meek, S., 2008. Park and ride. In: Ison, S., Rye, T. (Eds.), *The Implementation and Effectiveness of Transport Demand Management Measures, An International Perspective*. Ashgate Publishing Ltd, Aldershot.
- Meyer, M.D., 2016. Travel demand management. In: Meyer, M.D. (Ed.), *Transportation Planning Handbook: Institute of Transportation Engineers*. John Wiley & Sons, Inc, New York.
- Ministry of Transport, 2019. ERP. Available from: <https://www.mot.gov.sg/about-mot/land-transport/motoring/erp>.

- Potter, S., 2008. Purchase, circulation and fuel taxation. In: Ison, S., Rye, T. (Eds.), *The Implementation and Effectiveness of Transport Demand Management Measures, An International Perspective*. Ashgate Publishing Ltd, Aldershot.
- Preston, S., 2008. Public transport subsidisation. In: Ison, S., Rye, T. (Eds.), *The Implementation and Effectiveness of Transport Demand Management Measures, An International Perspective*. Ashgate Publishing Ltd, Aldershot.
- Seik, F.T., 2000. An advanced demand management instrument in urban transport: electronic road pricing in Singapore. *Cities* 17 (1), 33–45.
- Shoup, D.C., 2005. *The High Cost of Free Parking*. Planners Press: American Planning Association, Chicago.
- Shoup, D.C., 2008. The politics and economics of parking on campus. In: Ison, S., Rye, T. (Eds.), *The Implementation and Effectiveness of Transport Demand Management Measures, An International Perspective*. Ashgate Publishing Ltd, Aldershot.
- Southern California Association of Governments, 2019. About SCAG. Available from: <https://www.scag.ca.gov/>.
- Transport for London, 2019. Streetscape Guidance. Available from: <http://content.tfl.gov.uk/>.
- Transportation Research Board, 2004. Integrating tourism and recreation travel with transportation planning and project delivery. Available from: <https://www.mytrb.org/>.
- Uber, 2019. UberPool. Available from: <https://www.uber.com/>.
- Van Reeve, P., 2008. Subsidisation of urban public transport and the Mohring effect. *J. Trans. Econ. Policy* 42 (2), 349–359.
- Wang, B., Loo, B.P.Y., 2017. Travel time use and its impact on high-speed-railway passengers' travel satisfaction in the e-society. *Int. J. Sust. Trans.* 13 (3), 197–129.
- Washington State Transportation Commission, 2009. Transportation Demand Strategies for Schools Phase II Report: Reducing Auto Congestion Around Schools. Available from: <https://www.wsdot.wa.gov>.
- Xiao, J., Zhou, X., Hu, W., 2017. Welfare analysis of the vehicle quota system in China. *Int. Econ. Rev.* 58 (2), 617–650.

Traffic Flow Analysis

H. Michael Zhang*, Jia Li†, *One Shields Avenue, University of California, Davis, CA, United States; †Texas Tech University, Lubbock, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Basic Tools and Definitions	544
Vehicle Trajectory	544
Cumulative Curves	544
Traffic Flow Measurements and Variable Definitions	546
Stationary Traffic Flow Models	547
Dynamic Traffic Flow Models	548
Particle Tracing Traffic Models	549
Car-Following Models	549
Numerical Solution and Traffic Simulation	550
Piped Traffic Flow Models	551
The LWR Model	551
Nonequilibrium Models	553
Heterogeneous Traffic Flow	553
Reservoir Traffic Flow Models,	
Comments on the Piped Traffic Flow Models	554
Reservoir Traffic Flow Models	554
Remarks and Further Reading	555
Biography	555
References	555
Further Reading	556

Basic Tools and Definitions

In this section, we introduce two basic tools to describe traffic flow at the individual vehicle level, namely the time-distance trajectory of a vehicle and the cumulative "curve" of vehicle counts at a specific location over time or at a specific time over space. We will use these geometric constructs to show how variables of interest, such as time headway, spacing, flow rate, speed, and density are defined and measured.

Vehicle Trajectory

Imagine that one can install a Global Positioning System (GPS) device on each vehicle, and from which one can deduce at each time instant, the distance traveled by that vehicle. Let $x_n(t)$ denotes the distance traveled by vehicle n , then the plot of $x_n(t)$ against time t is called the vehicle trajectory of n . With this trajectory, one knows the speed and acceleration of each vehicle:

$$\text{Speed : } v_n(t) = \frac{dx_n(t)}{dt}, \quad \text{Acceleration : } a_n(t) = \frac{dv_n(t)}{dt} = \frac{d^2x_n(t)}{dt^2}$$

One can also compute the vehicle-miles-traveled (VMT) of this vehicle from $t = t_1$ to $t = t_2$: $\text{VMT} = x_n(t_2) - x_n(t_1)$, or vehicle-hours-traveled from x_1 to x_2 : $\text{VHT} = x_n^{-1}(x_2) - x_n^{-1}(x_1)$, where $x_n^{-1}(x)$ is the inverse function of $x_n(t)$ (see Fig. 1).

If we collect the trajectories for a group of vehicles, as shown in Fig. 1, we have a complete description of the traffic dynamics of this group of vehicles. Note that $x_n(t)$ is the distance traveled by a vehicle, these trajectories cannot bend backward. In one-lane traffic without passing, these trajectories will not cross each other, but in multilane traffic, vehicle trajectories do cross each other and can pop up or disappear due to passing, lane changes, entries and exits.

Cumulative Curves

A cumulative curve of traffic arrivals to a fixed location of the highway (x) is constructed by recording the arrival time of each vehicle at that location and increment a counter by one for each arrival. This "curve" is actually a staircase function with the vertical axis being the total vehicle counts N and the horizontal axis being observation time t , as shown in Fig. 2 for locations x_1, x_2 . But as arrivals become more frequent (i.e., the flow is high), the steps become narrower and one can approximate this staircase by a smooth curve $N(x, t)$, hence the name cumulative curve. Similarly, a cumulative curve of traffic passing any location of the road at a time instant t can be constructed by recording the location of each vehicle at that time, and counting the number of vehicles that has passed that location, as shown in Fig. 3 for two time instants t_1, t_2 .

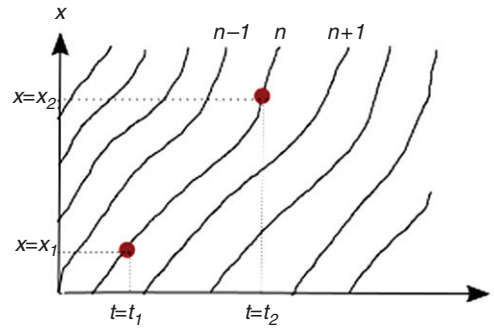


Figure 1 Vehicle trajectory diagram.

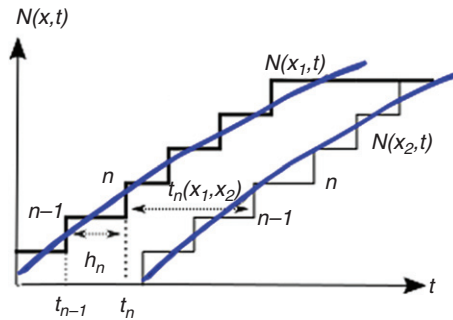


Figure 2 Cumulative vehicle counts passing fixed locations, observed over time.

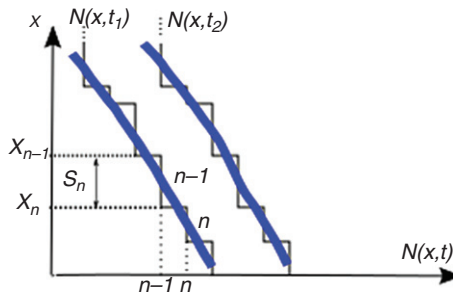


Figure 3 Cumulative vehicle counts passing any location, observed over space.

From the cumulative curve 28.2, one can know the average intensity of arriving flow in time period $t = [t_1, t_2]$: $q(x; t_1, t_2) = \frac{N(x, t_2) - N(x, t_1)}{t_2 - t_1} = \frac{\Delta N}{\Delta t}$. When $t_2 \rightarrow t_1 = t$ and $N(x, t)$ is smooth, we have the point intensity, or flow rate, as $q(x, t) = \frac{\partial N(x, t)}{\partial t}$. Another important variable that can be obtained from the arrival cumulative curve is the inter arrival times between successive vehicles, known as time headway or simply headway: $h_n = t_n - t_{n-1}$, where t_n is the arrival time of vehicle n at the observed location. For a sufficiently large Δt , we have $\sum_{n=1}^N h_n = t$, or $q = \frac{N}{\sum_{n=1}^N h_n} = \frac{1}{h}$.

Form the cumulative curve 28.3, one can get the concentration or density of vehicles on the road in a given road segment $x = [x_1, x_2]$: $k(x_1, x_2; t) = \frac{N(x_2, t) - N(x_1, t)}{x_2 - x_1} = \frac{\Delta N(x, t)}{\Delta x}$. Again, when $N(x, t)$ is treated as a smooth curve, one has $k(x, t) = -\frac{\partial N}{\partial x}$, with the negative sign reflecting the fact the total number of vehicles passing a downstream point (here x_2) is no more than the total number of vehicles passing an upstream point (here x_1). From Fig. 3, one can also obtain the space headway between successive vehicles on the road, or spacing (usually measured from head to head or tail to tail): $s_n = x_{n-1} - x_n$. For a sufficiently large Δx , we have $\sum_{n=1}^N s_n = x$, or $k = \frac{N}{\sum_{n=1}^N s_n} = \frac{1}{s}$.

Vehicle counts and headway are of fundamental interest to many transportation problems, such as traffic safety and traffic signal timing. While in heavy traffic situations it is common to treat the arrivals to a location as deterministic and use cumulative curve $N(t)$ (here x is suppressed) to model the arrival process. In light traffic situations the stochastic nature of the arrival process dominates and $N(t)$ is often treated as a random variable. A common probability distribution used for traffic counts for a given time interval $[0, t]$ is

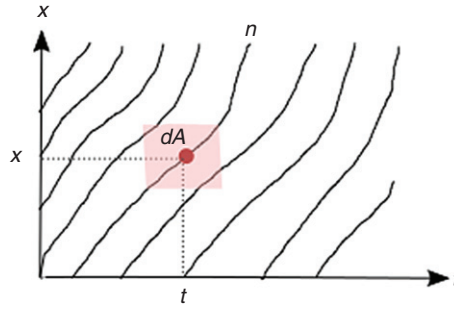


Figure 4 The definition of flow variables on a trajectory map.

the Poisson distribution with mean arrival rate q :

$$P(N(t) = n) = \frac{(t)^n e^{-qt}}{n!} \quad (1)$$

whose headway h has a negative exponential distribution:

$$P(h < t) = 1 - e^{-qt} \quad (2)$$

The Poisson distribution can also be applied to vehicle counts in a road section $[0, x]$ in light traffic situations with q replaced by k and t by x . The Poisson model is extensively used in traffic delay analysis for light traffic situations.

Traffic Flow Measurements and Variable Definitions

As we have illustrated in the vehicle trajectory concept, the ultimate way of measuring traffic is to track each vehicle continuously with GPS or other devices. For economic, privacy and other considerations, this is often not feasible. In contrast to such mobile sensing, the most common ways of measuring traffic flow are through installation of nonmobile sensors at fixed locations, such as inductive loops, video or microwave sensors. It is not the interest of this chapter to cover the specific sensing technologies. It is suffice to say that these nonmobile sensors map out a "measurement area" dA in the time-distance trajectory diagram, as shown in **Fig. 4**, and for each vehicle n inside this area, its travel time dt_n and travel distance dx_n . We then can define the average traffic flow rate, density, and speed as follows:

$$\text{Flow rate : } q = \frac{\sum_n dx_n}{dA} \quad (3)$$

$$\text{Density : } k = \frac{\sum_n dt_n}{dA} \quad (4)$$

$$\text{Speed : } u = \frac{\sum_n dx_n}{\sum_n dt_n} \quad (5)$$

By definition, we have $q = ku$. The average speed thus defined is also known as the space-mean speed. For a sensor that measures traffic within a fixed length dx , all vehicles measured within a given time interval T travels the same distance $dx_n = dx$ (we assume that dx is small enough that vehicles do not leave in the middle of it). Then Eq. (5) gives the following formula:

$$\text{Speed : } u = \frac{\sum_n dx}{\sum_n dt_n} = \frac{N}{\sum_{n=1}^N \frac{1}{v_n}} \quad (6)$$

which is known as the harmonic mean of individual speed $v_n(dx_n/dt)$. The arithmetic mean

$$v = \frac{\sum_n dx/dt_n}{N} = \frac{1}{N} \sum_{n=1}^N v_n \quad (7)$$

is known as the time-mean speed when individual speeds are measured at a fixed location over time. Time mean speed is always larger or equal to space speed and usually $q \neq kv$. Although time-mean speed is often used in spot speed studies, it is much less relevant in traffic flow analysis. Unless stated otherwise, the mean speed we use throughout this chapter is the space mean speed according to definition (5).

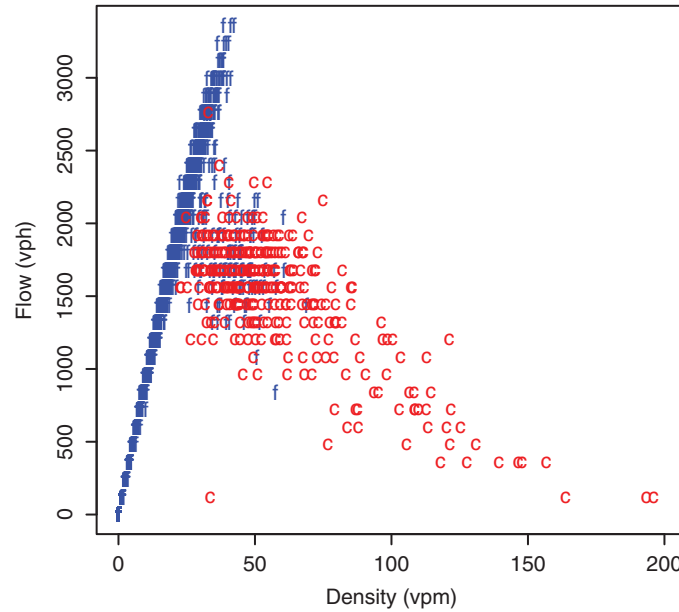


Figure 5 The flow-density plot from measured data on the fast lane of I-180 near Davis, California.

Another important measurement of interest is *travel time* or *travel delay*. For an individual vehicle n , its trajectory provides a direct way of getting its travel time between any two locations x_1 and x_2 ($x_2 > x_1$): $t_n(x_1, x_2) = x_n^{-1}(x_2) - x_n^{-1}(x_1)$ where $x_n^{-1}(x)$ is the inverse function of the trajectory $x_n(t)$. From $t_n(x_1, x_2)$, one can compute the population or sample mean of travel times, depending whether all the vehicles in the traffic stream or a sample of vehicles in the traffic stream have been measured with a vehicle positioning sensor.

The concept of delay depends on the definition of a reference travel time t_r , for example, the time it takes to traverse the $[x_1, x_2]$ section with the maximum travel speed (say the speed limit). Once this reference time is decided, the delay of vehicle n can be computed as

$$\text{Delay : } d_n(x_1, x_2) = t_n(x_1, x_2) - t_r \quad (8)$$

When traffic is not measured by mobile sensors, one can use the cumulative arrival curve to obtain the vehicle travel times, as shown in Fig. 2, which requires the construction of cumulative arrival curve $N(x_1, t)$ at location x_1 and $N(x_2, t)$ at location x_2 . The horizontal separation between these two arrival curve at $N(x_1, t_1) = N(x_2, t_2) = n$, is vehicle n 's travel time. The delay of vehicle n is its travel time subtract t_r , which is equivalent to shift the $N(x_1, t)$ curve to the right by t_r . It must be pointed out that this way of obtaining travel times or delays over a road segment depends on the assumption that vehicles do not pass each other (or vehicles obey first-in-first-out [FIFO] rule in queuing jargon). It can still be used to estimate travel times or delays when FIFO is obeyed by most vehicles.

Last but not least, the *vehicle-miles-traveled* (VMT) and *vehicle-hours-traveled* (VHT) in measurement area dA are simply $\text{VMT}(dA) = \sum_n dx_n$ and $\text{VHT}(dA) = \sum_n dt_n$, and that for the total road system would be $\sum_{dA} \text{VMT}(dA)$ and $\sum_{dA} \text{VHT}(dA)$ when all road segments of the system are measured in a given time period.

Stationary Traffic Flow Models

From empirical data, the flow rate, density and speed measured at highway locations have shown to exhibit certain relations when plotted against each other, as shown in Fig. 5. Such plots often contain stationary and transient traffic conditions. Here we define stationarity as the condition that flow rate, density and speed remain nearly constant for a period of time (usually longer than a minute). Studies have shown that the flow rate, density, and speed at a location under stationary conditions observe well defined relations (Castillo et al., 1995; Cassidy, 1998). These pair-wise relations are known as traffic stream models, with the flow-density relation in particular being referred to as the fundamental diagram of traffic flow, although from a driver's perspective, the speed-spacing or speed-headway relation is more natural and fundamental to understand her driving behavior.

There have been various models proposed for the fundamental diagram of traffic flow, with the earliest being the model proposed by Greenshields et al. (1935)

$$q = v_f(k - k^2/k_j) \quad (9)$$

where v_f is the free-flow speed (or speed limit since it is obtained at $k = 0$), and k_j is the jam density (when traffic is not moving or $u = 0$). The corresponding speed-density relation of Greenshield's model is given as $u = v_f(1 - k/k_j)$, a linear function of density.

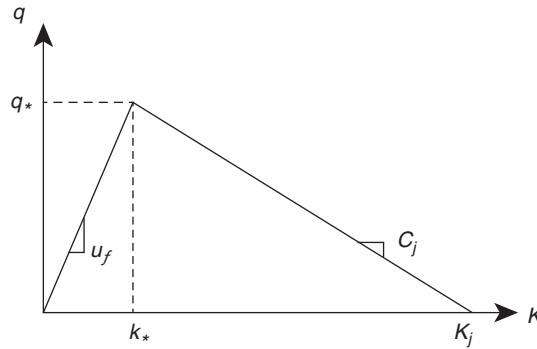


Figure 6 The triangular fundamental diagram.

Typical of the fundamental diagram (FD), the Greenshields FD has a maximum flow q_* at a particular density k_* . q_* is known as the capacity of the road at that location and k_* is known as the critical density. In the case of Greenshields FD, $k_* = k_j/2$ and $q_* = (u_f k_j)/4$. For a road with a free flow speed of 95 km/h and jam density of 143 car per km (5 m car length plus 2 m buffer between stopped cars), the maximum flow $q_* = 3396$ cars per hour greatly exceeds what one observes from real-life traffic. In the literature, unrealistically low jam densities are used to make the Greenshields FD produce realistic flow capacities.

There is another important quantity associated with the FD, which is the derivative of the flow with respect to density at $k = k_j$: $w_j = \frac{dq}{dk}|_{k=k_j}$. This quantity is known as the wave speed at jam density, which captures how quickly the front of a discharging queue with jam density moves upstream, and is known to be around -20 km/hr in most situations (see Chapter 5 of [Gartner et al., 2001](#)). Again for the Greenshields FD, this speed $w_j = -u_f$, which is in the range of -88 km/h to -120 km/h for most freeways, and clearly inconsistent with field observations.

While the Greenshields' FD played a pioneering role in the development of traffic flow theory, its drawbacks have long been recognized. Subsequent research on traffic stream characteristics produced a large number of traffic stream models, ranging from smooth Greenshields like models, models with a jump at or near the critical density, to even multi-valued functions with hysteresis (see [Gartner et al., 2001](#), Chapter 2 for a comprehensive review). The more exotic models, however, were developed to fit traffic data that contain both stationary and transient states, and are often under the influence of active bottlenecks. They do not reflect the traffic stream characteristics under stationary conditions.

Regardless which FD one chooses, one should check if the FD has reasonable free-flow speed and jam density, and produces reasonable flow capacity, critical density, and jam wave speed, all measurable quantities from field that can be used to validate the FD model. It turns out that there is a simple FD that meets all these requirements. It is the so-called triangular FD, given by the following equations:

$$q = \begin{cases} ku_f & k \leq k_* \\ |w_j|(k_j - k) & k \geq k_* \end{cases} \quad (10)$$

This FD has four parameters, u_f, k_j, w_j, k_* , all can be directly measured (the first three) or estimated from measurements (k_*). Only three of them are needed to determine the FD. The flow capacity is controlled by u_f, k_j , and w_j . As in the Greenshields' example, a road location with $u_f = 95$ km/h and $k_j = 143$ cars per km would have the capacity $q^* = 2363$ cars per hour for $w_j = -20$ km/h, which is in accordance with observed capacities for such roads.

Fig. 6 shows the triangular FD. As we shall show in the next section, this FD is particularly attractive when one studies traffic congestion using the piped traffic flow models, particularly the kinematic wave model known as the Lighthill-Whitham-Richards (LWR) model ([Lighthill and Whitham, 1955](#); [Richards, 1956](#)). Nevertheless, this FD is not without its own drawbacks. For example, the free-flow branch has only one speed, which is not fully consistent with observed traffic stream characteristics. The same is true for the congested branch. These can be overcome by using piece-wise linear FDs, as shown in **Fig. 7**, if field observations call for the use of such FDs.

The FDs capture the bulk of our driving behavior and is an essential element of traffic flow analysis. They or their variants are extensively used in dynamic traffic flow modeling (see the next section), transportation planning, and highway capacity and level of service analysis ([TRB, 2016](#)).

Dynamic Traffic Flow Models

This section presents three families of dynamic traffic models that describe traffic in different levels of detail. Using a fluid flow analogy, we call them (1) *particle tracing models* where each individual vehicle's movements are modeled, this family of models contain the so-called car-following models and cellular automaton models, and are known as microscopic models; (2) *piped traffic flow models* where highway traffic is treated like a fluid flowing in a pipe, this family of models are known by various names in the

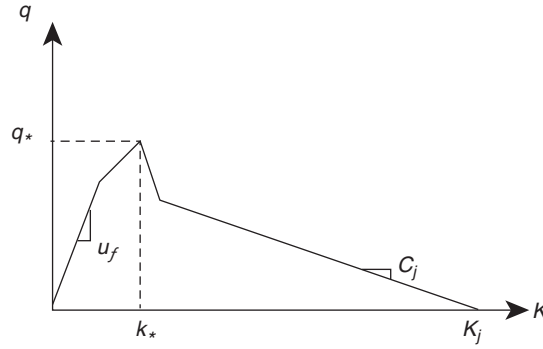


Figure 7 An example of a piece-wise linear fundamental diagram.

literature, including hydro dynamic models, kinematic wave models, and higher-order models, to name a few; and (3) *reservoir traffic flow models* where an entire city or blocks of a city are treated as a reservoir with entering and exiting flows through both the boundary and internally, some of these models known by names such as two-fluid town models or town models with a macrofundamental diagram (MFD).

Particle Tracing Traffic Models

Car-Following Models

A car-following model describes the movement of a vehicle constrained by its interactions with nearby vehicles. Although a driver pays attention to all surrounding vehicles, her predominant focus is often on the vehicle ahead of her. Most car-following models, therefore, were developed for a pair of vehicles, the following vehicle n and the front vehicle $n - 1$. Let $x_n(t), x_{n-1}(t)$ denote the positions (by its traveling distance from a reference point) of vehicle n and $n - 1$ respectively, and $v_n(t), v_{n-1}(t)$ their respective velocity, and $a_n(t), a_{n-1}(t)$ their respective accelerations. A car-following model predicts the movement of vehicle n described by the triplet $\{x_n(t), v_n(t), a_n(t)\}$, depending on the movement of the preceding vehicle $(n - 1) - \{x_{n-1}(t), v_{n-1}(t), a_{n-1}(t)\}$ from $t = 0$ to some terminal time $t = t_e$.

The simplest car-following model is perhaps the one derived from a common recommendation to drivers, which are known as the one-car length or 2-second rule:

One-car length rule: For every 10-mph speed of travel, leave one-car length space between your car and your front car.

2-second rule: Keep 2 s of time gap between you and your front car.¹

We can construct the trajectory of car n if we know the trajectory of car $(n - 1)$ for drivers who strictly follow these rules. In the case of the one-car-length rule, this is equivalent to

$$x_{n-1}(t) - x_n(t) - l_n = \left(\frac{v_n(t)}{10}\right) l_n \quad (11)$$

where l_n is the length of car n . And in the case of the 2-s rule, it is equivalent to

$$(x_{n-1}(t) - x_n(t) - l_n)/v_n(t) = 2 \quad (12)$$

In fact, if we define a new parameter $\tau = l_n/10$ in Eq. (11) and $\tau = 2$ in Eq. (12), both equations can be translated into the same equation:

$$v_n(t) = \frac{x_{n-1}(t) - x_n(t) - l_n}{\tau} = \frac{S_n(t) - l_n}{\tau} \quad (13)$$

$$a_n(t) = \frac{v_{n-1}(t) - v_n(t)}{\tau} \quad (14)$$

Eq. (14) tells us that according to these driving rules, the following driver should accelerate with a rate proportional to their velocity difference, that is, if the car in front travels faster than you, you accelerate; if it travels slower than you, you decelerate; and if it travels at the same speed as you, you cruise.

Since drivers have to respect the speed limit of the road, we impose a speed limit u_f to Eq. (13) to get

$$v_n(t) = \min \left\{ u_f, \frac{S_n(t) - l_n}{\tau} \right\} \quad (15)$$

¹ In practice, a driver is recommend to identify a target spot on the road and reach that spot two seconds after her front car. This is the time headway as we defined earlier.

Then the position of vehicle n can be obtained by integrating speed over time:

$$x_n(t) = \int_0^t v_n(t) dy \quad (16)$$

Using car-following models like this, one can obtain the vehicle trajectories of a stream of vehicles $n = 1, 2, \dots, n, \dots, N$ by solving a system of equations consisting Eqs. (15) and (16), as long as the initial position and velocity of each vehicle is known, and the trajectory of the leading vehicle is also known (in the case of a closed system such as vehicles traveling on a ring road, only the initial conditions need to be known). From the vehicle trajectories, one can compute the flow variables such as flow rate, density, speed, travel time/delay, and VMTs and VHTs as described in Section Basic Tools and Definitions.

Besides some simple car-following models like Eq. (15), one has to resort to numerical solutions to obtain approximate vehicle trajectories by discretizing the car-following equations in time. Under special situations, one does not need to solve these equations explicitly to gain insights of the behavior of such models. One such a special condition is when traffic is stationary. Without loss of generality, let's assume the cars are of the same length l . Further denote the stationary speed as u and spacing as s . Under stationary conditions, Eq. (15) becomes

$$u = \min \left\{ u_f, \frac{s-l}{\tau} \right\} \quad (17)$$

and taking into account that $s = 1/k$, and $l = 1/k_j$, the stationary flow rate is

$$q = ku = \min \left\{ ku_f, k \frac{1/k - 1/k_j}{\tau} \right\} = \min \left\{ ku_f, \frac{l}{\tau} (k_j - k) \right\} \quad (18)$$

Eq. (18) gives exactly the triangular FD, and is identical to Eq. (10) with $w_j = -l/\tau$. This also provides a way to estimate τ if w_j is known or w_j when τ is known. This connection between the triangular FD and the car-following model also provides the behavioral grounding for the triangular FD.

The study of car-following has a rich history in traffic flow analysis and what was discussed above shows a simple example of how such models can be developed based on the understanding of driving rules. In real life drivers could not perceive the relative speeds or distances precisely, and have different levels of cognitive and driving capabilities as well as different driving habits. These factors lead to more complex traffic dynamics and nonlinear car-following models to describe them. A fairly general nonlinear car-following model is given by the following equation (Gazis et al., 1961):

$$a_n(t+T) = b(v_n(t))^\alpha \frac{v_{n-1}(t) - v_n(t)}{(x_{n-1}(t) - x_n(t))^\beta} \quad (19)$$

where T is a time lag that reflects a driver's reaction time to respond to changes in relative spacing and speeds, and b, α, β are parameters to be determined through fitting car-following data. Interestingly, this model reduces to a linear model almost identical to (14) with $\alpha = 0, \beta = 0, b = 1/\tau$. There is a fundamental difference, however, between this reduced model and (14). Due to the introduction of a time delay (T), the reduced linear model can be unstable if the values of bT is large enough, while the model given by (14) is always stable. By instability we refer to the property that small changes introduced in the speed of the leading vehicle get magnified over time through a string of vehicles behind it. Time delay is not the only factor that can cause instability in car-following. Certain types of nonlinearity can also produce instability, such as in the [Bando's model](#). Another interesting special case of this model is when $\alpha = 0, \beta = 2$. This reduced model leads to the Greenshields FD under stationary traffic conditions.

Numerical Solution and Traffic Simulation

The car-following models described above are all in the form of ordinary differential equations (ODE), some are linear and others are nonlinear. Besides the linear and some special nonlinear car-following models that can be solved analytically, most of them has to be solved numerically. A common practice is to divide the time into intervals with equal increment, say Δ : $\{t_0, t_1, t_2, \dots, t_j, t_{j+1}, \dots\}$, $t_{j+1} - t_j = \Delta, \forall j$, and use difference equations to replace the original differential equations with a suitable approximation of the derivatives. One simple such scheme is the so-called Euler difference scheme, which uses $\frac{v(t_j) - v(t_{j-1})}{\Delta}$ to approximate $\frac{dv}{dt}$. Using this scheme, the difference equation for the general nonlinear model (19) is

$$\frac{v_n(t_{j+1} + T) - v_n(t_j + T)}{\Delta} = b(v_n(t_j))^\alpha \frac{v_{n-1}(t_j) - v_n(t_j)}{(x_{n-1}(t_j) - x_n(t_j))^\beta} \quad (20)$$

One can then update the position of each vehicle from time t_j to time t_{j+1} as follows:

$$x_n(t_{j+1}) = x_n(t_j) + v_n(t_{j+1}) \quad (21)$$

Suppose we know the initial conditions at time t_0 : $(x_n(t_0), v_n(t_0)), \forall n$ and the speed of the leading vehicle $v_1(t)$, then one can advance the position of each vehicle n from one time interval to the next, until the end of the specified time. Such difference solution schemes give rise to microscopic traffic simulation. In these simulations, only the time t is discretized into equal increments. There is

a family of models, known as particle hopping or cellular automaton models, that discretize both time and distance. The unit of discrete distance x is called a cell and its length is usually set to 7 m, a car length (with safety buffer when stopped). A car therefore occupies one cell and hops into another cell based on its current speed, which is measured in cells per time tick. Rules can then be specified on how cars hop among cells (Nagel and Schreckenberg, 1992). The advantages of this family of models are their simplicity, flexibility in specifying driving rules, and amenity to parallel computation.

Of course there is much more to a traffic simulation of a highway network than what is shown here. For example, one needs to specify lane change rules, paths taken by vehicles, and right-of-way rules at junctions, to name a few. Interested readers are referred to Gartner et al. (2001) for further reading.

A traffic simulation produces a piece-wise linear trajectory of each simulated vehicle. One can therefore use the methods in Section Basic Tools and Definitions to obtain performance measures such as travel time, travel delay, and VMTs for each vehicle or the entire simulated highway network from a microscopic traffic simulation.

Piped Traffic Flow Models

The LWR Model

Proposed by Lighthill and Whitham (1955) and Richards (1956), the LWR model borrows the idea from hydrodynamics: traffic propagation can be described as waves, which are called kinematic waves (KW).

Model formulation The LWR model is essentially a conservation equation (i.e., conservation of vehicles) coupled with the flow-density fundamental diagram obtained from stationary flow conditions:

$$k_t + Q(k, x)_x = 0 \quad (22)$$

where the subscripts t and x denote partial derivatives, and $Q(k, x)$ is a suitably chosen fundamental diagram that can vary by location (x). This is a partial differential equation derived from the integral equation of flow conservation

$$\frac{d}{dt} \int_x^{x+dx} k(x, t) dx = Q(k(x, t), x) - Q(k(x + dx, t), x + dx) \quad (23)$$

For the clarity of presentation, we show the properties of the LWR model with the simpler case where the FD is not dependent on location, that is, $Q(k, x) = Q(k)$ (the same approach can be applied to the inhomogeneous case with $Q(k, x)$ but it is more involved and is beyond the scope of this book). The solution uniquely exists for the LWR model when proper initial and boundary conditions are specified, and an additional condition called *entropy condition* is applied. The LWR model is a hyperbolic partial differential equation, whose properties are well studied. Characteristics $x_*(t)$ of the LWR model are defined as:

$$x_*(t) : \frac{dx}{dt} = Q'(k) \quad (24)$$

where $'$ denotes the derivative of a scalar function (e.g., $Q'(k) = dQ/dk$).

It can be verified that the characteristics defined above are straight lines, because along the $x_*(t)$, there is

$$\frac{dk}{dt} = k_x + k_t x'_*(t) = k_x + k_t Q'(k) = k_t + Q(k)_x = 0 \quad (25)$$

that is, k is constant on the characteristic $x_*(t)$. This property may be interpreted as characteristic lines carry the information about density. This requires $Q'(k)$ to be bounded for the LWR model to make physical sense, meaning information in traffic flow must propagate at finite speed.

Shock waves and rarefaction waves: If one examine the characteristic lines of the LWR model further, one may find in many cases, characteristic lines will cross each other. Therefore, to determine the solution, one must choose a proper value for density at this point. This basic observation motivates two most important concepts in the hyperbolic conservation equations (the mathematical theory behind the LWR model), namely weak solution and entropy condition. Rather than presenting the entire theory (see e.g., LeVeque, 1992), we illustrate properties of the LWR model through the so-called Riemann problem, which considers traffic evolution with piece-wise constant initial data (k_L, k_R) and a discontinuity at $x = 0$, as illustrated in Fig. 8.

The Riemann problem of LWR model leads to two types of waves, namely shock wave and rarefaction (expansion) wave. They arise under the following conditions:

- Shock wave: if and only if $k_L < k_R$
- Rarefaction wave: if and only if $k_L > k_R$

The first condition means upstream traffic is lighter than downstream traffic, therefore traveling faster. It models the situation when fast traffic catches slow traffic. In this case, a queue forms, and a shock wave describes how queue-ends move. A vehicle enters a shock will decrease its speed abruptly. The second condition describes the opposite case. In this case, the discontinuity in initial

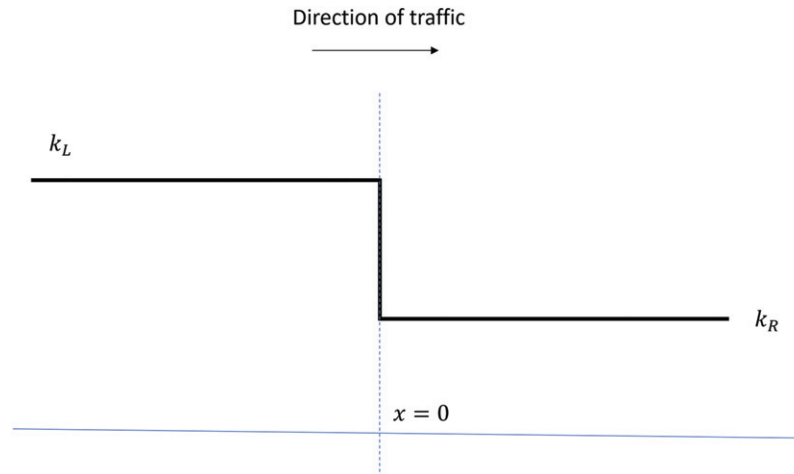


Figure 8 Illustration of a Riemann problem.

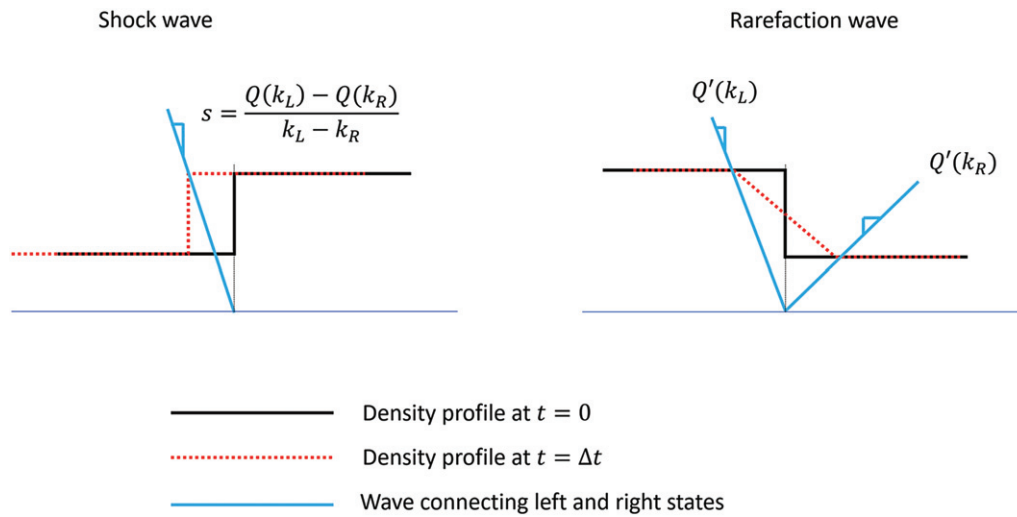


Figure 9 Two basic waves associated with the LWR model.

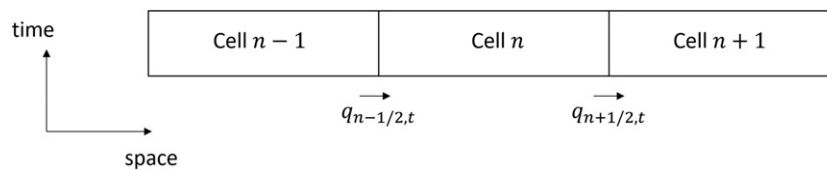


Figure 10 Cell discretization of traffic flow problems.

density profile will disappear, and the two states will be connected by a rarefaction fan. A vehicle enters the fan will accelerate smoothly until it travels out of the fan. Fig. 9 illustrates the difference between these two waves and depicts how the wave speeds are determined.

The Riemann problem with these two waves form the basis of the LWR theory to solve more complicated conditions, when general initial data are given. However, to solve such general problems analytically are tedious, if not difficult. Numerical solution schemes are therefore developed, with special care to treat shock conditions.

Numerical scheme: The common approach to solve the LWR model is through discretization of space into cells and update traffic state in each cell over discrete time steps (see Fig. 10). Then the LWR model can be numerically solved by

$$k_{n,t_j+1} = k_{n,t_j} + \frac{t}{x} (q_{n-1/2,t_j} - q_{n+1/2,t_j}) \quad (26)$$

This translates the problem into determining the corresponding numerical fluxes $q_{n-1/2,t_j}$ and $q_{n+1/2,t_j}$ based on the solutions to the corresponding Riemann problems (the initial density, when given in a continuous form, is approximated by a piece-wise constant staircase function).

It turns out that the numerical flux to the LWR model can be analytically derived from the solution of the Riemann problem at the cell boundaries. In transportation literature, this is known as the supply-demand scheme (Daganzo, 1994; Lebacque, 1996), written as follows (subscripts of time are omitted here)

$$q_{n-1/2} = \min\{D_{n-1}(k_{n-1}), S_n(k_n)\} \quad (27)$$

Here, $D_{n-1}(\cdot)$ and $S_n(\cdot)$ are respectively demand and supply functions of cell density for cells $(n-1)$ and n . For any cell, these functions are defined as,

$$D(k) = Q(\min\{k, k_*\}), S(k) = Q(\max\{k, k_*\}) \quad (28)$$

where k_* is critical density of the fundamental diagram. These functions can be expressed more explicitly as

$$D(k) = \begin{cases} Q(k) & k \leq k_* \\ Q(k_*) & \text{otherwise} \end{cases} \quad (29)$$

$$\text{and } S(k) = \begin{cases} Q(k) & k \leq k_* \\ Q(k_*) & \text{otherwise} \end{cases} \quad (30)$$

It is noted that when $Q(k)$ takes the form of the triangular FD, the LWR model is greatly simplified and can be solved easily without resorting to numerical solutions. The solution space is comprised of polygons, with each polygon defining an area of constant density/speed/flow. The boundaries of the polygons are the shock or rarefaction waves determined by the density on both sides of the edge of a polygon. The construction of these polygons are made particularly easy if one takes a variational approach by using the cumulative count $N(x,t)$ instead of density $k(x,t)$ in the LWR model (Daganzo, 2005; Newell, 1993).

Nonequilibrium Models

At the core of the LWR model is the FD, $Q(k)$, which captures the behavior of traffic under stationary conditions. Real traffic flow contains transient conditions and the FD is inadequate to represent such conditions. Nonequilibrium traffic flow models such as the one developed by Payne (1971) and Whitham (1974), adds a so-called equation of motion to the conservation equation $k_t + (ku)_x = 0$:

$$\frac{du}{dt} = u_t + uu_x = \frac{u_*(k) - u}{\tau} - \frac{c^2(k)}{k} k_x \quad (31)$$

The first term on the right hand represents relaxation of speed to the equilibrium state $u_*(k)$, and the second term represents a driver's response to gradient of density, which is known as the anticipation term.

Unlike the LWR model, the PW model comprises a system of PDEs with variables k, u (or k, q). The properties of nonequilibrium models can be subtle, which triggered extensive discussions on this topic. Among others, one puzzle was that there is a family of waves in the PW model that can travel faster than traffic (Castillo et al., 1993; Daganzo, 1995b), which leads to the situation that vehicles can "push" or "pull" vehicles from behind to speed up or slow down (Zhang, 1998, 2000). This puzzle was later resolved in Aw and Rascle (2000) and Zhang (2002). In their models, the acceleration of traffic depends on the speed gradient u_x instead of the density gradient k_x .

Nonequilibrium models allow traffic states to deviate from the FD and produce hysteresis loops in flow-density or speed-density plots (e.g., Jin and Zhang, 2003), which are often observed in such plots of real life traffic. Unlike the LWR model, which always produces stable traffic flow patterns, the nonequilibrium models can produce unstable traffic flow patterns due to the interaction between relaxation and anticipation (e.g., Zhang, 1999). While the nonequilibrium models provide a more realistic description of traffic flow (e.g., Fan and Seibold, 2013; Fan et al., 2017), solving it is more involved than solving the LWR model (e.g., LeVeque, 1992; Zhang, 2001), and are not as widely used as the LWR model in practice.

Heterogeneous Traffic Flow

When traffic comprises quite different vehicles/drivers, such as cars and trucks, extension of the LWR model to address heterogeneous traffic is necessary. In traffic flow, heterogeneity can be a result of different vehicle maneuverability (e.g., passenger car vs. truck), level of automation (e.g., human-driven vehicles vs. autonomous vehicles) or driving style (e.g., aggressive vs. timid drivers).

Most existing models consider two classes of vehicles, typically passenger cars and trucks. In recent works, attentions are paid to modeling the mixed flow of human-driven and autonomous vehicles. A flux function is defined for each class, and the interactions of different traffic streams are captured through coupling of the two flux functions, $Q_1(k_1, k_2)$ and $Q_2(k_1, k_2)$. This leads to a general formulation of two-class traffic flow models,

$$U_t + Q(U)_x = 0 \quad (32)$$

where $U = (k_1, k_2)$, and $Q = (Q_1, Q_2)$ is the flux pair. In the vector notation, its form resembles the single-class LWR model. However, its properties and solution can be much more complex.

The flux pair may be specified through either a descriptive or a behavioral approach. The descriptive approach prescribes flux functions based on assumptions of aggregate flow attributes, for example, regimes when the two traffic streams synchronize speed and traffic volumes. Many earlier multi-class traffic flow models adopt this approach (see e.g., [Wong and Wong, 2002](#)). The major limitation of this approach lies in the negligence of detailed driver/vehicle interactions as well as the role of lane use.

In contrast, the behavioral approach explicitly considers interactions of different vehicle classes and lane settings (such as special lanes) and leverages such information to construct the pair of fluxes. The construction usually starts from nominal fundamental diagrams each class is endowed with. This procedure can lead to more interpretable heterogeneous flow model, but the flux pairs usually do not have closed-form expressions, unless in simple cases. As an example, [Daganzo \(1997\)](#) used this approach to model two streams of traffic on a road with a special lane. Another type of models that share similarities with both descriptive models and behavioral models. These models incorporate behavioral rules for drivers/vehicles, such as Wardrop's User Equilibrium principle ([Wardrop, 1952](#)), but do not model lane use policies or detailed lateral distribution of drivers/vehicles across lanes. A representative model in this category is [Logghe and Immers \(2008\)](#).

Comments on the Piped Traffic Flow Models

The piped traffic flow models are among the most versatile and powerful traffic flow analysis models developed to date. They provide enough detail to describe and model the important features of traffic flow, such as congestion formation, propagation, and dissipation, and the workings of bottlenecks, without the need to model thousands or millions of vehicles in a city individually. The numerical schemes developed to solve these models, coupled with junction models (e.g., [Daganzo, 1995a](#); [Lebacque, 2005](#)), form the core of flow rather than vehicle based network traffic simulation models. As in a microscopic traffic simulation, flow-based simulations can be used to obtain travel times, travel delays, VMTs and VHTs using the methods described in Section Basic Tools and Definitions. Unlike in a microscopic traffic simulation, only a limited number of parameters need to be calibrated (for example, the location-dependent $FD-Q(k, x)$ is the only function to calibrate in the LWR family of models) in a flow-based simulation. Flow based simulation can provide good estimates of the performance measures such as travel times/delays, VMTs and VHTs. On the other hand, vehicle based simulation provides more accurate estimate of fuel use, emissions, and individual delay due to their finer granularity. Which simulation to use in a specific application depends on the size of the network, the level of detail one needs to model the traffic environment, the types of performance measures one is interested in, and the amount of labor and financial resources at hand for the application.

Reservoir Traffic Flow Models

The basic spatial unit of analysis in the models presented in the previous sections is a road segment or junction. Using these models to assess how well city traffic moves would require the modeling of each road and junction in the city. Yet in some occasions there is a desire to know how well a city's traffic moves on average without zooming in to each road, just like a physician takes the pulse and/or temperature of a person to get the first impression of the health of a person without X-rays or MRIs. The first known effort towards this is due to [Smeed \(1966\)](#), who investigated the capacity of traffic entering the city center (Q_c) as a function of the area of the city center (A), the proportion of the area devoted to roads (f), and the capacity of the roads (Q_r):

$$Q_c = \alpha f Q_r \sqrt{A} \quad (33)$$

where α is a parameter specific to a city. A subsequent effort by [Thomson \(1967\)](#) to examine the average speed and flow data collected over the center of London yields a linear speed-flow relation:

$$U = U_0 - bQ \quad (34)$$

and found that the speed U_0 for inner zones of London is lower than that for the outer zones of London.² Replacing U with Q/K in the above equation we get the town center flow-density FD:

$$Q = U_0 \frac{K}{1 + bK} \quad (35)$$

This is the first known macrofundamental diagram (MFD) of city traffic. Yet since dQ/dK is positive, that is, flow always increases with density, this MFD is not suitable to describe town traffic under highly congested situations. It is almost four decades later that more extensive data were collected and a more complete picture of MFD is revealed. [Geroliminis and Daganzo \(2008\)](#) showed the MFD of town traffic resembles that of a road, with a concave shape where flow reaches zero when traffic is jammed throughout the

² Here we use upper case letters to denote that these variables are for a town center, to distinguish them from those of a specific road location.

city. From a quite different approach, [Herman and Prigogine \(1979\)](#), based on the gas-kinetic theory of traffic flow ([Priogine and Herman, 1971](#)), stipulated that the average speed of a town's traffic is a function of the fraction of moving or stopped vehicles at a given moment, and carried out simulation studies to verify this so-called two-fluid town traffic model with good accordance (e.g., [Williams et al., 1987](#)).

Both empirical evidence and theoretical analysis strongly indicate that for the town center with densely laid roads with similar characteristics, a macro fundamental diagram, $Q^*(K)$ exists. Suppose the amount of the flow reaching their destinations inside the town center is proportional to $Q^*(K)$: $rQ^*(K)$, with r being the percentage of the flow reaching their destinations, and that the traffic generated from within the town center and from outside the town center is $Q(t)$, one can develop a reservoir like model

$$L \frac{dK}{dt} = Q(t) - rQ^*(K) \quad (36)$$

where L is the total length of the roads inside the town center. This model tells how the average density and speed in the town center evolves over time without the need to study how traffic evolves in each road segment and at each junction. An important feature of this town traffic model is that, due to the concave shape of the MFD and the existence of an absorbing state at $K = K_j$, the town traffic has a tendency to gridlock after the density (or the total number of vehicles inside the town) passed a critical value (e.g., [Daganzo et al., 2011](#)).

Remarks and Further Reading

This chapter presented some of the main traffic flow models developed to date, from static to dynamic, and from individual vehicles to traffic in an entire town. Due to space limitations, several aspects of traffic flow analysis was not discussed in this chapter. The first is the role of lane use and lane changing in particle tracing traffic models. Interested readers can refer to [Zheng \(2014\)](#) for a recent review on this topic. The second is the role of randomness in traffic flow dynamics, which has been shown to play a significant role in causing traffic instability (e.g., [Tian et al. 2019](#)). Last but not least, the emergence of connected and autonomous vehicles (CAV) is likely to affect traffic flow analysis profoundly. For example, the communication between vehicles enable information sharing with neighboring vehicles and beyond, so vehicles can respond to traffic conditions both behind and ahead of them in a greater range beyond their immediate neighbors, and can follow each other more closely than human driven vehicles. This means that for CAV traffic both the capacity and wave speeds can be higher. A more significant challenge lies in modeling the mixed flow containing both human driven vehicles and CAVs. Since CAVs are still in their infancy, limited data are available to the transportation research community to understand the interactions between human driven vehicles and CAVs. Much research is anticipated in the development the CAV technology, its role in transportation, and the flow analysis tools to model traffic mixed with CAVs and human driven vehicles in decades to come. Since this is a rapidly developing area of research, there is no definitive reference for this topic and readers are recommended to read the latest articles in journals and conference proceedings such as the Transportation Research B & C, Transportation Science, the IEEE Transactions series, and Transportation Research Records.

Biography

H. Michael Zhang is a professor in the Department of Civil and Environmental Engineering at the University of California Davis. He is also a faculty member in the Graduate Programs of Applied Mathematics and Transportation Technology and Policy at UC Davis. His research focuses on applications of systems theory to transportation systems analysis and operations. Specific topics include traffic flow theory, traffic control, traffic safety, transportation network modeling, and intelligent transportation systems such as connected and autonomous vehicles. He is an area editor of the Journal of Networks and Spatial Economics and Associate Editor of Transportation Research, Part B: Methodological and Transportation Science. His awards include an NSF CAREER (2000) and a best paper award from Vehicular Communications (2018), an Elsevier journal. He received his BS degree in Civil Engineering from Tongji University (Shanghai, China) and his MS and PhD degrees in Civil Engineering from University of California Irvine. Prior to his UC Davis appointment, he has taught at the University of Iowa from 1995 to 1998.

Jia Li is a research assistant professor at Texas Tech University. His research focuses on the intersection of traffic flow theory, transportation systems modeling, and connected and autonomous vehicles. He received his PhD in Transportation Engineering from University of California Davis in 2013, and conducted his postdoctoral research with Professor Michael Walton at the University of Texas at Austin from 2013 to 2016.

References

- Aw, A.A.T.M., Rascle, M., 2000. Resurrection of "second order" models of traffic flow. *SIAM J. Appl. Math.* 60 (3), 916–938.
 Bando, M., Hasebe, Nakayama A., Shibata Akihiro, Sugiyama, Y., 1995. Dynamical model of traffic congestion and numerical simulation. *Phys. Rev. E* 51 (2), 1035.
 Cassidy, M.J., 1998. Bivariate relations in nearly stationary highway traffic. *Transport. Res. Part B: Methodol.* 32 (1), 49–59.

- Castillo, J.M.D., Benitez, F.G., 1995. On the functional form of the Speed-density relationship: empirical investigation. *Transport. Res. Part B: Methodol.* 29 (5), 391–406.
- Castillo, J., Pintado, P., Benitez, F., 1993. A formulation for the reaction time of traffic flow models. *Transport. Traffic Theory* 12, 387–405.
- Daganzo, C.F., 1994. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transport. Res. Part B: Methodol.* 28 (4), 269–287.
- Daganzo, C.F., 1995a. The cell transmission model, part II: network traffic. *Transport. Res. Part B: Methodol.* 29 (2), 79–93.
- Daganzo, C.F., 1995b. Requiem for second-order fluid approximations of traffic flow. *Transport. Res. Part B: Methodol.* 29 (4), 277–286.
- Daganzo, C.F., 1997. A continuum theory of traffic dynamics for freeways with special lanes. *Transport. Res. Part B: Methodol.* 31 (2), 83–102.
- Daganzo, C.F., 2005. A variational formulation of kinematic waves: basic theory and complex boundary conditions. *Transport. Res. Part B: Methodol.* 39 (2), 187–196.
- Daganzo, C.F., Gayah, Vikash, V., Gonzales, E.J., 2011. Macroscopic relations of urban traffic variables: bifurcations, multivaluedness and instability. *Transport. Res. Part B: Methodol.* 45 (1), 278–288.
- Fan, S., Seibold, B., 2013. Data-fitted first-order traffic models and their second-order generalizations: Comparison by trajectory and sensor data. *Transport. Res. Rec.* 2391 (–1), 32–43.
- Fan, S., Sun, Y., Piccoli, B., Seibold, B., Work, D.B., 2017. A collapsed generalized Aw-Rascle-Zhang model and its model accuracy arXiv preprint arXiv:1702.03624.
- Gartner, N.H., Messer, C.J., Rath, A., 2001. Traffic flow theory-A state-of-the-art report: revised monograph on traffic flow theory. *Transport. Res. Board.*
- Gazis, D.C., Herman, Robert, Rothery, R.W., 1961. Nonlinear follow-the-leader models of traffic flow. *Oper. Res.* 9 (4), 545–567.
- Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. *Transport. Res. Part B: Methodol.* 42 (9), 759–770.
- Greenshields, B.D., Channing, W., Miller, H.O., 1935. A study of traffic capacity. *Highway Research Board Proceedings* 1935 .
- Herman, R., Prigogine, I., 1979. A two-fluid approach to town traffic. *Science* 204 (4389), 148–151.
- Jin, W.-L., Zhang, H.M., 2003. The inhomogeneous kinematic wave traffic flow model as a resonant nonlinear system. *Transport. Sci.* 37 (3), 294–311.
- Lebacque, J.-P., 1996. The godunov scheme and what it means for first order traffic flow models, *Transportation and Traffic Theory. Proceedings of the 13th International Symposium on Transportation and Traffic Theory*, Lyon, France, 24–26 July 1996.
- Lebacque, J.-P., 2005. First-order macroscopic traffic flow models: Intersection modeling, network modeling, *Transportation and Traffic Theory. Flow, Dynamics and Human Interaction. 16th International Symposium on Transportation and Traffic Theory* University of Maryland, College Park.
- LeVeque, R.J., 1992. *Numerical Methods for Conservation Laws*, vol. 132. Springer.
- Lighthill, M.J., Whitham, G.B., 1955. On kinematic waves i. flood movement in long rivers. *Proc. R. Soc. Lond. Series A. Math. Phys. Sci.* 229 (1178), 281–316.
- Logghe, S., Immers, L.H., 2008. Multi-class kinematic wave theory of traffic flow. *Transport. Res. Part B: Methodol.* 42 (6), 523–541.
- Nagel, K., Schreckenberg, M., 1992. A cellular automaton model for freeway traffic. *J. Phys. I* 2 (12), 2221–2229.
- Newell, G.F., 1993. A simplified theory of kinematic waves in highway traffic, Part I: General theory. *Transport. Res. Part B: Methodol.* 27 (4), 281–287.
- Payne, H., 1971. Models of freeway traffic and control. *Mathematical Models of Public Systems.*
- Prigogine, I., Herman, R., 1971. *Kinetic Theory of Vehicular Traffic*. American Elsevier.
- Richards, P.I., 1956. Shock waves on the highway. *Oper. Res.* 4 (1), 42–51.
- Smeed, R.J., 1966. The road capacity of city centers. *Highway Res. Rec.* (169).
- Thomson, J.M., 1967. Speeds and flows of traffic in central london: 1. sunday traffic survey. *Traffic Eng. Control* 8 (11), 672–676.
- Tian, J., Zhang, H., Treiber, M., Jiang, R., Jia, Z.G.B., 2019. On the role of speed adaptation and spacing indifference in traffic instability: Evidence from car-following experiments and its stochastic model. *Transport. Res. Part B: Methodol.* 129, 334–350.
- TRB, 2016. *Highway Capacity Manual (6th Edition) A Guide for Multimodal Mobility Analysis*. Transportation Research Board, Washington, DC.
- Wardrop, J.G., 1952. Road paper. some theoretical aspects of road traffic research. *Proc. Inst. Civil Eng.* 1 (3), 325–362.
- Whitham, G.B., 1974. *Linear and Nonlinear Waves*, vol. 42. John Wiley & Sons.
- Williams, J.C., Mahmassani, H.S., Herman, R., 1987. Urban traffic network flow models. *Transport. Res. Rec.* 1112.
- Wong, G.C.K., Wong, S.C., 2002. A multi-class traffic flow model—an extension of lwr model with heterogeneous drivers. *Transport. Res. Part A: Policy Practice* 36 (9), 827–841.
- Zhang, H.M., 1998. A theory of nonequilibrium traffic flow. *Transport. Res. Part B: Methodol.* 32 (7), 485–498.
- Zhang, H.M., 1999. Analyses of the stability and wave properties of a new continuum traffic theory. *Transport. Res. Part B: Methodol.* 33 (6), 399–415.
- Zhang, H.M., 2000. Structural properties of solutions arising from a nonequilibrium traffic flow theory. *Transport. Res. Part B: Methodol.* 34 (7), 583–603.
- Zhang, H.M., 2001. A finite difference approximation of a non-equilibrium traffic flow model. *Transport. Res. Part B: Methodol.* 35 (4), 337–365.
- Zhang, H.M., 2002. A non-equilibrium traffic model devoid of gas-like behavior. *Transport. Res. Part B: Methodol.* 36 (3), 275–290.
- Zheng, Z., 2014. Recent developments and research needs in modeling lane changing. *Transport. Res. Part B: Methodol.* 60, 16–32.

Further Reading

- Zhang, H.M., Jin, W.L., 2002. Kinematic wave traffic flow model for mixed traffic. *Transport. Res. Rec.* 1802 (–1), 197–204.

Autonomous Vehicles and Transportation Modeling

Annesha Enam, Felipe de Souza, Omer Verbas, Monique Stinson, Joshua Auld, Argonne National Laboratory, Lemont, IL, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction and Motivation	557
Demand Impacts of CAVs	557
Land Use and Other Long-Term Impacts of CAV	558
Modeling Traffic Flow Impacts of CAV	559
System Control Under CAV	560
Conclusion	561
Biography	561
Relevant Websites	562
References	562
Further Reading	563

Introduction and Motivation

A new era of connected and autonomous vehicle (CAV) technology is emerging due to rapid advancements in communication technology, machine learning and artificial intelligence, vehicle-based sensing, and computation. However, as autonomous vehicles near commercial availability, there exists much uncertainty about the implications of this transformative technology including the ownership costs, mode of deployment (shared versus private), adoption rate and technical capabilities, which will all influence potential changes in travel behavior, mobility, and energy outcomes. National, regional and local governments are facilitating research to develop a better understanding about potential outcomes of CAV in terms of travel demand, regional mobility, energy use, emissions, and so on. This has led to numerous studies looking at different aspects of CAV short- and long-term outcomes, by both adapting traditional transportation models and developing new transportation simulation techniques.

This article reviews research that models CAV from the following perspectives: (1) modeling demand impacts, (2) modeling land use and other long-term changes, (3) modeling traffic flow impacts, and (4) modeling system control impacts. For each area, a general discussion about the plausible impact is presented first. A review of pertinent modeling studies, findings and research gaps for each perspective follows.

Demand Impacts of CAVs

The automation aspect of CAVs is expected to drive changes in travel behavior through the elimination of the human driver, as identified in multiple high-level analyses, including (Brown et al., 2014). This has the possibility to increase travel demand in multiple ways, including expanding mobility, inducing additional travel demand, changing long-term travel patterns and altering land use. Mobility could be expanded in two ways. First, CAV offers a new and convenient mobility option that operates like privately owned automobiles to individuals who currently are underrepresented in the driving population: persons with disabilities, minors, elderly citizens, and so on. As such, CAV is likely to lead to increased travel by these groups. In the same vein, elimination of driving duties can alter the experience of driving altogether for individuals that currently drive. By reducing or removing the need for driver attention (in partial and full automation, respectively), the experienced burden of travel in an automobile can be substantially reduced for drivers. This can lead to a reduced disutility of travel, which is further enhanced by the ability to engage in productive and relaxing activities in lieu of driving. For drivers, therefore, mobility is expected to expand due to a reduction in value of travel time (VOTT). This in turn could induce greater travel demand, with people making more trips and/or accessing more distant but more desirable activity locations. This would result in longer travel durations and increased vehicle miles travelled (VMT). However, travelers initially may be unfamiliar with or apprehensive about CAV technology, so they may attach high VOTT to CAV travel at first. There is also much debate about the deployment mode of the CAV technology: a shared fleet system versus privately ownership. While shared AV is more favorable from the VMT and energy use perspective, private AV can generate a notable increase in VMT due to low marginal cost and increased empty miles traveled. However, private ownership of CAV can also be influenced by its income generating potential through participation in the sharing economy and thereby minimizing negative externalities. Addressing these factors is important for assessing travel demand impacts of CAV.

Attempts to investigate the impacts of CAV or shared automated vehicle (SAV) systems on travel behavior range from simple rule-based models to complex agent-based and integrated activity-based and dynamic traffic simulation frameworks. On the simpler end,

several studies have used methods including cross-classification based on demographic characteristics or assignment of fixed demand with a simplified grid network to study the impacts of CAV. More detailed transportation modeling efforts have paired regional demand inputs with dynamic traffic assignment. Multiple studies using a variety of transportation modeling platforms for various cities, including Austin, Chicago, Berlin, Zurich, and others have explored the impact of SAV fleets for various fleet sizes and potential replacement rates to serve existing demand. In general, these studies predict that each SAV could replace between 5 and 12 privately owned vehicles (Bischoff and Maciejewski, 2016; Fagnant and Kockelman, 2014). These detailed analyses also generally predict that SAV systems will generate a substantial increase in overall travel largely due to vehicle repositioning and pickup trips. To date, most studies of SAV on a regional level have used fixed demand inputs or have allowed some characteristics of demand, such as mode choice, to change. The induced demand effect in such studies, then, is limited to that derived from mode changes and route assignment.

Other studies have explored modifying existing trip-based or activity-based travel demand models to simulate one or more aspects of CAV deployment. In activity-based models, the VOTT change effect can often be introduced by modifying travel time parameters in various choice modeling components. Enhanced mobility can be included in the mode choice process by allowing nonlicensed drivers to utilize privately owned AVs or to have AVs perform household tasks such as escorting children. Multiple examples of existing activity-based models at MPOs being modified in this manner to explore some CAV impacts have been observed in the literature. Agent-based models are more frequently being used to explore CAV impacts as well. These share many features of activity-based travel demand models, with more emphasis on modeling agent interactions in the system, including persons, households, vehicles, system managers, and fleets. These types of models allow the representation of the interactions between agents in the system, which is especially relevant in exploring CAV impacts on travel demand and traffic flow and the resultant feedbacks (Auld et al., 2018).

From studies observed in the literature, it is found that most of the research explores the impact of private or SAVs with very few studies investigating the impact of both together (i.e., mixed modes of AV operation). There is also little focus in current transportation models on the connectivity aspect of CAVs. Moreover, the studies that do exist have had to make critical assumptions about the travel characteristics to enumerate different future scenarios under the market penetration of autonomous and connected vehicles. Major assumptions are made regarding road capacity (+10% to +100%), VOTT (−20% to −100%), parking cost (−50% to −100%), market penetration of private AVs (+7.5% to +100%), operating cost of shared AVs (\$0.13/mile to \$1.65/mile), travel time increase due to sharing (+20% to +50%), and replacement of vehicle trips by AVs (+1.3% to +100%). Consequently, the outcome of different studies seems to vary considerably depending on the range of assumptions.

To offer insights into system performance impacts of CAV, existing studies typically report the change in vehicle miles (kilometers) traveled (VMT/VKT), vehicle hours traveled (VHT), and shift in modal share in response to CAV. Studies have generally predicted that VMT/VKT will increase, with values ranging up to 79% for private AV (Auld et al., 2018). Reported empty miles traveled due to zero occupancy vehicle (ZOV) trip ranges up to 16% (Loeb et al., 2018). Some studies report reduction in vehicle hours traveled (VHT) up to 41% (Childress et al., 2015), largely due to assumed increases in road capacity. In contrast, some research shows a decrease in VMT when all private vehicles are replaced by a SAV system with high operating costs, while generally most studies of SAV systems show overall VMT increases due to repositioning moves.

In terms of mode shifts, the studies report mixed outcomes depending on the assumptions regarding private vehicles. When it is assumed that private vehicle ownership and utilization will still exist, transit mode share generally decreases depending on location, quality of existing transit service, and assumptions about VOT in AVs. Studies that instead assume complete disposal of private vehicles with replacement by SAV systems find an increase in transit and walk share.

Research gaps in modeling travel demand impacts of CAV pertain mainly to underlying assumptions that are used in analysis and how the assumptions are developed. A review of transportation demand modeling studies reveals widespread use of simplified assumptions regarding future scenarios. Almost all the studies assume a fixed rate for household vehicle disposal irrespective of household location and SAV levels of service such as operating cost and wait time. Some studies do not consider any changes in activity participation for different future CAV scenarios resulting in fixed demand for all future conditions. Others do not consider the trade-off in level of service (LOS) between different transport modes, instead assuming a fixed modal share or rule-based assignment to different transport modes. Simplification in traffic assignment is also often seen, with studies as using an idealized network. Most have also used unimodal network assignment instead of multimodal assignment. Finally, many studies only focus on one specific effect of CAV, such as the effect of dynamic ride sharing. However, as transportation is a complex adaptive system with many potential ownership and deployment possibilities as well as multiple interactions between demand, vehicle operation, system operation, traffic flow, and so on, focusing on only one aspect can often lead to uninformative or misleading results.

Land Use and Other Long-Term Impacts of CAV

CAV deployment is expected to trigger changes in longer term choices due to changes in accessibility and reduction in VOTT. Changes in accessibility might influence people to choose different home and workplace locations, creating shifts in population density. Also, assuming that inter- and intrahousehold sharing of CAVs reduces parking demand, opportunities for new

infrastructure/land use development may be created. Changes in accessibility also have the potential to lead to firm relocation, which would also result in new land use patterns.

Very few studies have attempted to study the long-term impacts of CAV. Some studies have used land use and transportation interaction models to study the impact of CAV on transportation infrastructure and urban form. Typically, these tools use discrete choice models to (re)distribute population across the study area as a function of accessibility to employment and four step models for simulating transportation system impacts. Some researchers have used macroeconomic models to understand the impacts of private and shared automation on land use. Macroeconomic formulations model consumers' choice of household location, job location, and commute mode using discrete choice specification. One distinguishing feature of the macroeconomic formulations is that they ensure price equilibrium for the residential market. Integrated SAV simulation and disaggregate choice models for residential and firm locations have also been used to study the impact of CAV on parking and urban sprawl, where SAV operation is usually simulated as discrete event engine and residential/firm location choices are generated using discrete choice models. Some researchers have used macroscopic transportation simulation tools to investigate the impact of AV on accessibility. Initial findings from the research on long-term impact of CAV can be grouped into three main categories: change in parking requirements, change in (sub)urban population, and change in land use.

Assuming that SAV systems serve all or part of current travel demand, studies predict a reduction of up to 90% in parking space requirements (Bischoff and Maciejewski, 2016). The estimated reduction in parking space requirement is much smaller assuming private AV ownership, with several studies showing less than 10% reduction, although this could be changed if dedicated AV parking facilities for private AV were introduced (Zhang et al., 2018).

Private AV ownership is predicted to increase accessibility and consequently population in the suburban and rural areas. Studies predict around a 5% increase in suburban population and a similar reduction in the urban population due to increased accessibility of the suburban areas resulting from private AV ownership. Mixed effects are reported for SAV scenarios. According to some studies, SAV has the potential to curb urban sprawl. However, depending on the assumption, SAV also has the potential to increase urban sprawl. For example, an assumption that all private vehicle trips are replaced by SAVs generates a 4% increase in urban population and 3% reduction in suburban population. Efficient public transportation systems complemented by SAV can increase urban population by 3% and decrease suburban population by the same margin.

Land-use changes are also expected to be driven by CAV deployment, and have often been studied with dedicated land use models, such as UrbanSim or others. Studies have indicated an increased trend of deindustrialization in urban areas, with changes in the job density based on the type of industry. Such predictions of impacts in future scenarios are sensitive to the assumptions that are made. Uncertainty about these assumptions (such as shared versus private ownership of autonomous vehicles, changes in VOT, and level of service characteristics of SAVs) currently forms a major hindrance to predicting future impacts of CAV. Also, many of the studies assume no change in the short-term travel behavior—in particular, several studies assume that activity patterns do not change. This is a major limitation since activity patterns in general probably will experience noticeable shifts, as the new mobility options alter trip making tendencies and hence activity patterns of those who currently do not drive as well as those who currently drive. Also, studies that focus on changes in short-term decision-making use simplified inputs such as grid networks instead of detailed, local transportation network, which limits the realism of the analysis.

Modeling Traffic Flow Impacts of CAV

Along with inducing changes in travel demand, the increase in vehicle automation and communication capability enables a plethora of applications that can affect different aspects of traffic flow, that are also critical to transportation system modeling. Depending on the vehicle capability and the infrastructure available, such systems can improve safety, increase traffic capacity (throughput), and reduce energy consumption and emissions. A summary of the implications of CAV technology on traffic flow is provided below. Readers can refer to Diakaki et al. (2015) for a thorough description of each.

- **Adaptive cruise control (ACC):** Adaptive cruise control is a low-level automation that allows vehicles to maintain a desired speed while maintaining a safe distance between successive vehicles. ACC can increase roadway capacity by reducing headway as the penetration rate of ACC increases. Increased capacity of up to 20% at high penetration rates is predicted;
- **Cooperative Adaptive Cruise Control (CACC):** Adding to the functionalities of ACC, CACC systems exchange information with neighboring vehicles to avoid traffic oscillations following accelerations and decelerations of the lead vehicle. Several studies have indicated that CACC can increase capacity up to 60% (Shladover et al., 2012). The communication also improves string stability (i.e., small oscillations are not amplified), which can further lead to energy savings.
- **Cooperative Merging:** vehicles assist drivers or perform lane-changing maneuvers leveraged by vehicle-to-vehicle (V2V) communication between the maneuvering vehicle and vehicles in the target lane. It can reduce travel times and energy consumption, but the extent of reduction is unclear.
- **Eco (Green)-driving:** Eco-driving encompasses a variety of strategies that attempt to reduce fuel consumption and greenhouse gas emissions without affecting overall travel times. Different levels of automation and connectivity can improve eco-driving benefits. In general, the vehicle gathers information of an expected future driving horizon through sensors, V2V data (e.g., current speeds of preceding vehicles), vehicle-to-infrastructure (V2I) data (e.g., green and red times of traffic lights) then plans vehicle controls to

reduce energy consumed along the trajectories. Several studies have found energy savings of up to 20% through the utilization of eco-driving.

- **Platooning:** This application attempts to keep multiple vehicles closely spaced for convenience, safety and fuel efficiency, primarily through aerodynamic drag reduction (of up to 50% according to some studies).

There is considerable uncertainty regarding the estimated benefits of CAV to traffic flow, as field data from such applications are limited. Most estimates are obtained through microsimulation by adding the specific driving behavior (car-following and lane-changing) into existing microsimulation packages. Additionally, the communication network and message exchanges between the vehicles are sometimes considered. Impacts are assessed by varying the share of vehicles with the specific capability (e.g., CACC, eco-driving, etc.).

Although microsimulation models can be used to analyze CAV impacts, they are usually limited to a small region as computational cost prevents application in a broader area. For this reason, several mesoscopic and macroscopic models have been developed to evaluate the impacts of CAVs at an aggregated level. The most common approach involves deriving the Lighthill–Witham–Richards (LWR) model and its popular discretizations such as the Cell or Link Transmission Models. The fundamental diagram is modified based on the share of enabled vehicles within the cell or link in addition to the density.

The market share-dependent fundamental diagram can be obtained in different ways. One can derive baseline fundamental diagrams from data as in common practice. For increasing technology penetration, a steeper derivative is assumed for the congested segment of the fundamental diagram, reflecting a smaller gap between enabled vehicles. Alternatively, fundamental diagrams for varying penetration rates can be obtained through microsimulation runs with different penetration rates of automation technology (Kun Li and Ioannou, 2004; Shladover et al., 2012). Several studies have considered a two-class cell transmission model for modeling mixed traffic. These models maintain a good compromise between computational time and accuracy and are suitable for large-scale metropolitan area analysis, especially for capturing the interaction between demand and supply (capacity).

In summary, increased level of automation and increased penetration can lead to improvement in traffic flow for similar demand levels, but the extent of the benefits is uncertain due to different aspects. First, while the traffic dynamics are somewhat well understood for the current conditions (i.e., human-driven vehicles) and for 100% penetration of CAV, the benefits for intermediate level of automation are uncertain. This is primarily due to the lack of data for mixed traffic conditions and for different vehicles with variable levels of automation and enabled features. Second, the coordination among vehicles and infrastructure depends on whether all different equipment (actuated signals, vehicles from different OEMs, sensors, etc.) can effectively communicate among themselves. Finally, some aspects of AV traffic flow such as automated merging have been largely understudied to this point.

System Control Under CAV

As discussed above, the presence of CAV features can improve traffic flow through modification in the driving behavior and communication exchange with surrounding vehicles. However, these are local impacts with uncoordinated control. As is well known in the transportation literature, a local change can trigger downstream impacts in the network because of changes in the route, mode and departure time choices. This motivates the development of management and control strategies for CAVs to generate benefits at the system level.

The key aspect of management and control at the system level is coordinated decision-making by various elements of the infrastructure or by a centralized authority, in contrast to decisions being performed only by individual vehicles. Note, however, that this level of coordination does not forfeit the benefits provided by increasing automation level; it further improves their benefits by coordinating vehicle level control with infrastructure decisions. Some potential applications discussed in the literature include:

- **Eco-approach and Departure (EAD):** The eco-driving approach can be extended by enabling the infrastructure to exchange information with individual vehicles. A recent field test (Hao et al., 2014) in which the actuated signal communicates Signal, Phase and Timing (SPaT), a system within the vehicle computes optimal speed trajectory and provides a suggested speed to the driver. They found an energy reduction of about 6%. Microsimulation studies have found energy savings of around 20%.
- **Fully Automated Intersection:** A new intersection timing is based on the concept of allocating time to specific movements for a scheme in which vehicles are coordinated to cross the intersection at specific intervals, thereby avoiding stops while maintaining safety. This approach is often referred to as a signal-free scheme, and can give travel time reduction up to 50% (Malikopoulos et al., 2018).
- **Charging Scheduling:** As the share of electric vehicles increases, an important element is to determine when and where to charge vehicles (including CAV) and at what price. Through V2I, the infrastructure can mediate this interaction using information on the state of charge, location, and intended trip schedule of electric vehicles. Congestion pricing or tradable networks help to constrain increases in congestion by charging a higher fee in periods of high demand. Alternatively, the demand can be limited by a fixed quota of road usage given to citizens that could trade these permits based on their necessity.

In summary, different infrastructure-based management and control strategies can be leveraged by connected vehicles. It is worth mentioning that improvements can also be achieved on the current traditional Intelligent Transportation Systems (ITS) applications such as traffic signal control and ramp metering. The impacts of the various strategies will depend on technological aspects, such as

automation levels, as well as market penetration and deployed infrastructure. The potential of these strategies is also dependent on the CAV adoption and household AV ownership.

Conclusion

Understanding the implications of CAV deployment on the transportation system is still in its infancy. There exists much uncertainty about how CAVs will be deployed, the strategies that will be used for their operation, the public's willingness to adopt the technology, and consequent impacts for both individuals and the system as a whole. Moreover, a lack of data regarding these and other elements has caused heavy reliance on assumptions to produce plausible future scenarios. The outcomes of the studies vary considerably depending on the assumptions used and their ranges, which vary widely across studies. Implications for travel behavior appear to be much less certain than implications for network flow. Another issue is that although CAV is likely to have numerous effects that interact with one another, many studies rely on relatively simple models that focus only on a limited number of system subcomponents. However, a few studies have used agent-based models with dynamic traffic assignment to generate tenable results at the system level, demonstrating that sufficiently comprehensive analytical methods are within reach. As the deployment of the technology becomes imminent, a better understanding about the assumptions will be needed to narrow down the plausible implications and to help formulate effective policies regarding CAV operation.

Biography



Annesha Enam is currently serving as a postdoctoral appointee at Argonne National Laboratory where she is working on travel behavior modeling with focus to emerging and disruptive technologies. Annesha Enam completed her undergraduate and master's degree from Bangladesh University of Engineering and Technology (BUET). She completed her doctoral studies from University of Connecticut (UConn), USA in activity and travel behavior modeling. Her dissertation research was awarded the 2017 CUTC Charley V. Wootan Memorial Award by the Council of University Transportation Centers in the USA. Her research interests include econometric modeling of choice behavior, activity time use and well-being, transportation planning, use of data mining and machine learning methodologies for understanding choice behavior.



Felipe de Souza is a Postdoctoral Appointee in the Vehicle and Mobility Systems Group at Argonne National Laboratory. He received a bachelor degree in Control and Automation Engineering and master of science in Systems Engineering, both at Universidade Federal de Santa Catarina, Brazil. He completed his doctoral studies at University of California, Irvine in Transportation Systems Engineering. His main research interests are traffic modeling and control, transportation systems simulation, and fleet dispatch optimization.



Monique Stinson is a Computational Transportation Scientist at Argonne National Laboratory where she is developing agent-based models of freight transportation and behavioral models of passenger transportation. Her research covers both commercial markets as well as interactions between commercial and passenger markets, including aspects related to AV and SAV adoption in a Sharing Economy context. Monique is a member of the Freight Planning and Logistics Committee and a friend of other freight and passenger behavior committees. She earned her Bachelor of Science and Master of Science degrees in Civil Engineering from the University of Texas at Austin, and is currently earning her PhD in Civil Engineering at the University of Illinois at Chicago.



Omer Verbas is a Computational Transportation Engineer in the Vehicle and Mobility Simulations Group in the Center for Transportation Research at Argonne National Laboratory. His primary research areas are in transportation network modeling; multimodal routing, assignment, and simulation; and transit network design and scheduling. He completed his doctoral studies at Northwestern University in 2014 under the supervision of Prof. Hani S. Mahmassani with whom he worked as a Post-Doctoral Researcher until 2016. Prior to joining the PhD program at Northwestern, he received his Master of Science degree in Transportation Engineering from Istanbul Technical University.



Joshua Auld is a Principal Computational Transportation Engineer in Argonne's Vehicle and Mobility Simulations Group, and technical manager of transportation systems simulation. He completed his Doctorate in August 2011, in the Civil and Materials Engineering Department of the University of Illinois at Chicago with a concentration in transportation. He has experience in a variety of areas in transportation, with a primary focus on dynamic activity-based travel demand models and the interactions between travel demand and intelligent transportation systems operations. He is a member of the TRB Travel Forecasting Resource and Transportation Demand Forecasting Committees.

Relevant Websites

<https://www.fhwa.dot.gov/planning/tmip/>
http://tfresource.org/Category:Autonomous_vehicles
<https://www.its.dot.gov/pilots/>
<https://www.rand.org/topics/autonomous-vehicles.html>
<https://ertico.com/focus-areas/connected-automated-driving/>

References

Auld, J., Verbas, O., Javanmardi, M., Rousseau, A., 2018. Impact of Privately-Owned Level 4 CAV Technologies on Travel Demand and Energy. *Procedia Comput. Sci.*, The 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018) / The 8th International Conference on Sustainable Energy Information Technology (SEIT-2018) / Affiliated Workshops 130, pp. 914–919. <https://doi.org/10.1016/j.procs.2018.04.089>.

- Bischoff, J., Maciejewski, M., 2016. Simulation of City-wide Replacement of Private Cars with Autonomous Taxis in Berlin. *Procedia Comput. Sci.*, The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops 83, pp. 237–244. <https://doi.org/10.1016/j.procs.2016.04.121>.
- Brown, A., Gonder, J., Repac, B., 2014. An analysis of possible energy impacts of automated vehicle. In: Meyer, G., Beiker, S. (Eds.), *Road Vehicle Automation*, Lecture Notes in Mobility. Springer International Publishing, pp. 137–153. doi:10.1007/978-3-319-05990-7_13.
- Childress, S., Nichols, B., Charlton, B., Coe, S., 2015. Using an Activity-Based Model to Explore Possible Impacts of Automated Vehicles. Presented at the Transportation Research Board 94th Annual Meeting.
- Diakaki, C., Papageorgiou, M., Papamichail, I., Nikolos, I., 2015. Overview and analysis of vehicle automation and communication systems from a motorway traffic management perspective. *Transp. Res. Part Policy Pract.* 75, 147–165. doi:10.1016/j.tra.2015.03.015.
- Fagnant, D.J., Kockelman, K.M., 2014. The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. *Transp. Res. Part C Emerg. Technol.* 40, 1–13. doi:10.1016/j.trc.2013.12.001.
- Hao, P., Boriboonsomsin, K., Wu, G., Barth, M., 2014. Probabilistic model for estimating vehicle trajectories using sparse mobile sensor data, in: *Intelligent Transportation Systems (ITSC)*, 2014 IEEE 17th International Conference On, pp. 1363–1368. <https://doi.org/10.1109/ITSC.2014.6957877>.
- Kun Li, Ioannou, P., 2004. Modeling of traffic flow of automated vehicles. *IEEE Trans. Intell. Transp. Syst.* 5, 99–113. doi:10.1109/TITS.2004.828170.
- Loeb, B., Kockelman, K.M., Liu, J., 2018. Shared autonomous electric vehicle (SAEV) operations across the Austin, Texas network with charging infrastructure decisions. *Transp. Res. Part C Emerg. Technol.* 89, 222–233. doi:10.1016/j.trc.2018.01.019.
- Malikopoulos, A.A., Cassandras, C.G., Zhang, Y.J., 2018. A decentralized energy-optimal control framework for connected automated vehicles at signal-free intersections. *Automatica* 93, 244–256. doi:10.1016/j.automatica.2018.03.056.
- Shladover, S.E., Su, D., Lu, X.-Y., 2012. Impacts of cooperative adaptive cruise control on freeway traffic flow. *Transp. Res. Rec.* 2324 (1), 63–70.
- Zhang, W., Guhathakurta, S., Khalil, E.B., 2018. The impact of private autonomous vehicles on vehicle ownership and unoccupied VMT generation. *Transp. Res. Part C Emerg. Technol.* 90, 156–165. doi:10.1016/j.trc.2018.03.005.

Further Reading

- Auld, J., Javanmardi, M., Freyermuth, V., Islam, E., Rousseau, A., 2019. "Simulation Analysis of the Energy and Mobility Impacts of Privately Owned Fully Autonomous Vehicles." In *Proceedings of the 98th annual meeting of the Transportation Research Board*, Washington DC.
- Boesch, Patrick M., Ciari, Francesco, Axhausen, Kay W., 2016. Autonomous vehicle fleet sizes required to serve different levels of demand. *Transport. Res. Rec.* 2542 (1), 111–119.
- Burghout, W., Rigole, P.J., Andreasson, I., 2015. Impacts of shared autonomous taxis in a metropolitan area." In *Proceedings of the 94th annual meeting of the Transportation Research Board*, Washington DC.
- Burns, L.D., Jordan, W.C., Scarborough, B.A., 2013. Transforming personal mobility. *The Earth Institute* 431, 432.
- Chen, T.D., Kockelman, K.M., Hanna, J.P., 2016. Operations of a shared, autonomous, electric vehicle fleet: implications of vehicle & charging infrastructure decisions. *Transport. Res. Part A: Policy Practice* 94, 243–254.
- Fagnant, D.J., Kockelman, K.M., 2015. Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. *Transportation* 45 (1), 143–158.
- Heilig, M., Hilgert, T., Mallig, N., Kagerbauer, M., Vortisch, P., 2017. Potentials of autonomous vehicles in a changing private transportation system—a case study in the Stuttgart region. *Transport Res. Procedia* 26, 13–21.
- Heinrichs, D., 2016. "Autonomous driving and urban land use," in: Maurer, M., Gerdes, J.C., Lenz, B., Winner, H. (Eds.), *Autonomous Driving: Technical, Legal and Social Aspects*, Springer, Berlin Heidelberg, pp. 213–231.
- Kröger, L., Kuhnimhof, T., Trommer, S., 2019. Does context matter? A comparative study modelling autonomous vehicle impact on travel behaviour for Germany and the USA. *Transport. Res. Part A: Policy Pract.* 122, 146–161.
- Martinez, L.M., Viegas, J.M., 2017. Assessing the impacts of deploying a shared self-driving urban mobility system: An agent-based model applied to the city of Lisbon, Portugal. *Int. J. Transport. Sci. Technol.* 6 (1), 13–27.
- Meyer, J., Becker, H., Bösch, Patrick M., Axhausen, Kay W., 2017. Autonomous vehicles: the next jump in accessibilities? *Res. Transport. Econ.* 62, 80–91.
- Simoni, M.D., Kockelman, K.M., Gurumurthy, K.M., Bischoff, J., 2019. Congestion pricing in a world of self-driving vehicles: An analysis of different strategies in alternative future scenarios. *Transport. Res. Part C: Emerging Technol.* 98, 167–185.
- Spieser, K., Treleaven, K., Zhang, R., Frazzoli, E., Morton, D., Pavone, M., 2014. Toward a systematic approach to the design and evaluation of automated mobility-on-demand systems: A case study in Singapore. In *Road vehicle automation*. Springer, Cham, pp. 229–245.
- Thakur, P., Kinghorn, R., Grace, R., 2016. Urban form and function in the autonomous era. In: *Australasian Transport Research Forum (ATRF)*, 38th, 2016, Melbourne, Victoria, Australia.
- Xu, X., Mahmassani, H.S., Chen, Y., 2019. Impact of Autonomous Vehicles on Household Activity and Travel Scheduling: An Integrated Dynamic Network Modeling Approach. In *Proceedings of the 98th annual meeting of the Transportation Research Board*, Washington DC.
- Zhang, W., Guhathakurta, S., Khalil, E.B., 2018. The impact of private autonomous vehicles on vehicle ownership and unoccupied VMT generation. *Transport. Res. Part C: Emerging Technol.* 90, 156–165.
- Zhang, W., Guhathakurta, S., Fang, J., Zhang, G., 2015. Exploring the impact of shared autonomous vehicles on urban parking demand: An agent-based simulation approach. *Sustain. Cities Soc.* 19, 34–45.

Ride-Hailing and Travel Demand Implications

Felipe F. Dias, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, Austin, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	564
Disruptiveness	565
Regulation	565
Travel Demand	565
Automation	566
Conclusions	567
See Also	567
References	567

Introduction

It is undeniable that the world's transportation landscape has undergone substantial changes in the 2010s, one of the most notable ones being the birth and growth of ride-hailing services provided by companies such as Uber, Lyft, Didi, and Ola (often referred to as transportation network companies or TNCs). Ride-hailing has been referenced by several different terms, including ride-sourcing, ride-splitting, ride-sharing, real-time ride-sharing, ride-booking, parataxis, on-demand rides, and e-hailing. These services can be classified as being a form of mobility-on-demand and as being part of the larger mobility-as-a-service trend, which allows individuals to purchase mobility by the trip instead of owning (and using) their own vehicles.

Simply put, ride-hailing companies provide a platform that connects willing drivers and riders. These services take advantage of the popularity of smartphones and their embedded technology to provide users with a reliable on-demand door-to-door transportation service that is usually cheaper than traditional taxis. Another key element that appeals to users is the seamlessness of the experience: rides can be requested (hailed), tracked, and paid using one single smartphone app (Dias et al., 2017). Ride-hailing companies usually offer multitiered services to their users, ranging from cheaper services with regular vehicles to more premium services with luxury vehicles and professional drivers. This provides more options to potential riders based on their need for comfort and sensitivity to price. Recently, most ride-hailing companies have also made a notable new addition to their service rosters: pooled rides, sometimes referred to as shared rides or ride-splitting. This service offers ride-hailing users a discount if they accept sharing their rides with other passengers who have similar pick-up and drop-off locations, further enhancing the ride-hailing option set. Further, typical taxis have fares calculated using a fixed formula based on mileage and time which can vary by time of day. Conversely, in the context of ride-hailing, fares are usually multiplied by a "surge" during periods of high demand of ride-hailing. Ride-hailing companies claim this is done to moderate demand and to stimulate an influx of drivers to serve high-demand locations. Also of note is that Uber has recently started engaging in "route-based pricing," in which they charge customers based on what an Artificial Intelligence algorithm predicts the potential riders are willing to pay, instead of its older pricing scheme which only involved mileage, time and surge multipliers (Newcomer, 2017).

Ride-hailing services differ from classic taxis in a few key aspects in addition to their cost/price structures. The first difference has been the center of several clashes between the ride-hailing and taxi industries: the issue of licensing and regulation. In most cities, the taxi industry is heavily regulated: drivers need to have city-issued permits, some form of insurance, and vehicles need to be inspected on a routine basis. Ride-hailing companies, on the other hand, while claiming to uphold high driver and vehicle quality standards, generally do not need to adhere to the same standards of regulations. There have been a few exceptions, as in the city of Dusseldorf in Germany, where after ceasing operations in 2015 due to regulatory disputes, Uber relaunched their services in late 2018 after agreeing to employ only professional drivers (see Bershidsky 2018). Another difference between the two services is that ride-hailing vehicles are typically hailed exclusively through smartphone apps, while traditional taxis can be hailed on the street and, in some cities, through smartphone apps as well.

The dramatic exponential growth of ride-hailing services is often depicted using a now famous statistic: since the day it was created, it took Uber 6 years to serve its one billionth ride, but only 6 more months to serve two billionth. Just two years after that, in 2018, Uber reported serving its 10 billionth ride (Kokalitcheva, 2016; Uber, 2015, 2018). An interesting summary of the company's timeline can be found in Hartmans and Leskin (2019).

A comprehensive literature review of all aspects surrounding ride-hailing services can be found in Wang and Yang (2019). In this paper, we will present the reader with a short summary of some of the main topics surrounding ride-hailing and how its use has been studied in recent years.

Disruptiveness

Since their creation, ride-hailing companies have disrupted the transportation industry similarly to how Amazon and AirBnB have, respectively, disrupted the retail and hotel industries. This disruption has brought multiple positive changes, such as increased accessibility in transit deficient locations, job opportunities to lower-skilled and unemployed individuals, opportunities for people to earn money “on the side” by driving for ride-hailing companies outside of their regular working hours, and giving passengers a transportation option that has the same comforts as private automobiles and is slightly cheaper (and often more convenient) than traditional taxis.

This disruption, however, can have (and has had) unintended negative consequences. One such example is New York City’s taxi industry and its medallion system. “Medallions” are city-issued licenses to operate taxis and hail rides. At the time of the medallion system’s creation in 1937, each of the city’s 16,900 medallions cost \$10 (Van Gelder, 1996). As time passed, the number of medallions remained almost constant while their cost climbed, reaching \$100,000 in 1985 and \$200,000 in 1997, until they reached their peak price of \$1.3 million in 2014 (Hu, 2017; Rosenthal, 2019). In 2015 just 4 years after the introduction of ride-hailing services in New York City, the number of Uber cars was already larger than the number of licensed taxis (Licea et al., 2015). Shortly thereafter, medallion prices fell dramatically, reaching prices as low as \$160,000 in 2018, sending thousands of people into debt and even leading some to commit suicide (Fitzsimmons, 2018; Hu, 2017; Rosenthal, 2019; Walker, 2018). It has been defended that the actual effect of ride-hailing companies was minimal, reducing taxi pickups by only 10%, and that this was simply the small nudge needed to burst a market bubble that was already long overdue (Rosenthal, 2019). Regardless of the actual magnitude of ride-hailing companies’ impacts, it is clear that these services have become an important segment of our transportation system and deserve to be closely studied.

Regulation

Another important aspect about ride-hailing companies and the services they provide is centered around the issue of regulations. Companies such as Uber and Lyft have, for the most part, not been subject to the same regulatory requirements that typically apply to transportation services such as taxis and limousine companies. Taxi driver demonstrations demanding law-makers to enforce regulations on ride-hailing companies and to ensure fair market conditions have been a common occurrence around the globe (Che, 2015). In most cases, ride-hailing companies have been successful in their push back efforts and have been able to avoid new regulations. A good example of such efforts is the case of Austin, Texas. Uber and Lyft (the two biggest ride-hailing companies in the city) spent approximately \$8 million to support Proposition 1, a bill that prohibited the city from requiring ride-hailing companies to perform fingerprint background checks on their drivers. In May 2016 the bill was put to public vote and was rejected by 56% of voters, making it clear that Austin’s residents wanted ride-hailing companies to perform the background checks on their drivers. Three days after the vote’s results were published, Uber and Lyft expressed disappointment with the political defeat and officially ceased operations in the city. The companies then decided to focus their lobbying efforts at the state level: approximately one year later, in a 21–9 vote, the Texas Senate approved a statewide bill that eliminated local ordinances for ride-hailing companies (such as Austin’s fingerprint background check requirements), effectively overruling the decision Austin’s residents’ has voted on just one year prior. Uber and Lyft then relaunched their operations in Austin mere hours after the Texas Governor, Greg Abbott, signed the bill into law two weeks later (Herskovitz, 2017; Lee, 2017; Mekelburg, 2016; Samuels, 2017; Office of the Texas Governor, 2017; Wear, 2017).

Travel Demand

Due to the ever-growing global presence of ride-hailing companies, several studies have attempted to study this newcomer in the transportation landscape. In this section, we briefly cover some of the research efforts that have tried to understand the characteristics of ride-hailing users and their trips.

Despite the dramatic growth in the demand for and usage of ride-hailing services, their mode shares are still quite small in almost all the cities in which they operate. Most recent travel surveys in the United States, including the 2017 National Household Travel Survey and a number of metropolitan area travel surveys (e.g., San Diego, Sacramento, Phoenix), indicate mode shares of under one percent for ride-hailing services (Federal Highway Administration, 2017; RSG Inc., 2018). Even though ride-hailing services may have generated some new latent demand and reduced some personal automobile travel, particularly in a few dense urban markets (Lavieri and Bhat, 2019a), some authors defend that they have served as a substitute to transit and traditional taxi services (Schaller, 2018), both of which already had very low mode shares in most American cities. Meanwhile, other researchers have found that ride-hailing services actually complement transit services (Zhang and Zhang, 2018). Another such example is Babar and Burtch (2017), who found that rail-based and long-haul transit services, such as subway and commuter rail, are also associated with increased ride-hailing use. It seems that, at least for the time being, ride-hailing will continue to represent only a small fraction of all trips unless ride-hailing companies start making significant impacts on individuals’ current personal vehicle travel trends. Regardless of their small mode shares, metropolitan planning organizations (MPOs) and transportation agencies around the world still need to

consider ride-hailing services in their modeling efforts because they might have significant implications for future transit investments, parking capacity needs, curbside management, safety, and vehicle miles of travel (Dias et al., 2019).

Unfortunately, there are only a handful of publicly available data sources on ride-hailing usage. Due to understandable privacy and competitiveness issues, ride-hailing companies are quite reluctant to share any kind of ridership data. This has lead researchers and transportation professionals to rely on secondary data sources (e.g., MPO's household travel surveys that have some ride-hailing specific questions) and to develop their own surveys to characterize ride-hailing use, such as Lavieri and Bhat (2019a) and Rayle et al. (2016). Regardless of their data-related difficulties, these research efforts have provided valuable insights into understanding how people interact with ride-hailing services.

The pioneering works in the behavioral analysis of ride-hailing revealed that the service's users tend to be younger, more educated, live in urban areas, earn higher incomes, and own fewer cars than the general population (Clewlow and Mishra, 2017; Dias et al. 2017; Lavieri et al. 2017; Rayle et al. 2016; Vinayak et al. 2018). These studies also revealed that driving while intoxicated and parking are the two main issues users are trying to avoid when they decide to use ride-hailing services. Furthermore, Clewlow and Mishra (2017) found that about 9% of the individuals they surveyed claimed that the existence of ride-hailing services had lead them to dispose of one or more of their household vehicles. Curiously, this specific finding was later corroborated by Hampshire et al. (2018): they found that approximately 9% of ride-hailing users purchased an additional private vehicle after Uber and Lyft suspended their activities in Austin, TX in May 2016 (see more about this service interruption in the third section). More recent studies that focus specifically on millennials have shown that older, more educated and employed millennials tend to use ride-hailing services more frequently than other millennials (Alemi et al. 2018, 2019). In another recent study, researchers collected data on a sample of commuters in the Dallas-Fort Worth area and found that certain attitudinal factors (specifically pro-environmental, technology-embracing, and variety-seeking attitudes) contribute to increased ride-hailing use, and therefore should be considered alongside individual sociodemographics as key determinants in ride-hailing use models. They also found that the delays caused by picking up additional passengers are greater barriers to sharing rides (i.e., using the "pooled" service) than the actual presence of strangers in the vehicle (Lavieri and Bhat, 2019a, 2019b).

The studies mentioned so far used fairly traditional data collection processes. Kooti et al. (2017) employed a significantly less orthodox (yet no less effective) data collection process: the authors partnered with Yahoo to investigate the receipts sent by Uber to riders' email addresses after a trip was completed. This gave the researchers access to trip-level information from approximately 59 million rides. The authors then coupled the data mined from the receipts with Yahoo's own data, which contained users' demographics. Using this composite data set containing trip-level data as well as user demographics, the authors showed that the average active Uber rider is a female individual in her mid-20s with an above-average income. They also found that older riders tend to make longer trips and use more expensive ride-hailing services than younger riders. Unsurprisingly, they also found that users with higher income levels are more likely to use the more expensive services. In their study, Kooti et al. (2017) found that users with higher incomes are only slightly more likely to take rides during surge pricing periods, but it seems that age plays a much more significant role in this respect, making younger adults much more likely to use ride-hailing services during surge periods.

A more recent trend in the behavioral analysis of ride-hailing has been to make use of disaggregate trip-level ride-hailing data sets, such as City of Chicago (n.d.), City of New York (n.d.), DiDi (n.d.), and RideAustin (2017). Some of the studies that use these data include Dias et al. (2018), Gerte et al. (2018), Komanduri et al. (2018), Lavieri et al. (2018), and Wenzel et al. (2019). Using this disaggregate data, authors have found that ride-hailing trips are heavily concentrated on weekend evenings (Friday, Saturday, and Sunday), and that trips to and from the airport represent about 5% of all ride-hailing trips. Researchers have also found that ride-hailing trip generation has a "geographic spill-over effect" (i.e., areas with high levels of ride-hailing trip generation are likely to increase their neighbors' trip generation rates). The data have also served to show that deadheading (i.e., driving without a passenger) constitutes approximately 37% and 50% of ride-hailing drivers' miles driven and travel time, respectively. It has also been estimated that ride-hailing has had a 41%–90% net increase in energy consumption when comparing current personal travel patterns to those before ride-hailing companies were widely available. The disaggregate data have also been used to show that the unbounded growth of ride-hailing services observed on the macroscopic level is not observed on the local scale: there does seem to be a point when ride-hailing use stops growing and simply stabilizes.

Automation

It is no secret that ride-hailing companies have been interested in expanding their business models to incorporate self-driving autonomous vehicles that do not require drivers. All four of the major ride-hailing companies (Uber, Lyft, Didi, and Ola) have already announced significant investments and partnerships to reach this goal. One of their main objectives is to cut costs significantly by reducing labor wages: once drivers are no longer needed, the financial resources that once went toward paying them for their labor can be redirected to reducing customer fares, maintaining the autonomous fleet's upkeep, and potentially increasing profit margins. Another purported benefit is that autonomous vehicles have the potential of being safer than human-driven vehicles due to their decreased reaction times and their more accurate sensing capabilities.

The path to fully autonomous fleets, however, is not entirely clear since it represents a significant departure from ride-hailing companies' current business model. Presently, the individuals that drive for ride-hailing companies are (in most cases) not legally seen as the ride-hailing companies' employees—drivers are instead seen as contractors, and the vehicles they use during rides are their own private vehicles, and are not owned by the ride-hailing companies. In an autonomous future where drivers are not needed,

this model would be absolutely unsustainable. In that case, it is most likely that ride-hailing companies will have to own their autonomous fleet. Another possibility is that there will be enough people who privately own autonomous vehicles and are willing to lease these vehicles out to ride-hailing companies during the time they do not use them (which, for nonautonomous vehicles has been estimated to be around 95% of the time, according to [Shoup, 2017](#)). This would allow the vehicle owners to earn some extra cash while not having to actively perform any kind of work and, in turn, allow ride-hailing companies to reduce the size of company-owned autonomous fleets.

This shift toward automation might also represent a potential threat to the hegemony of the private vehicle's mode share seen in most US cities. The exponential growth ride-hailing services have experienced since their creation exemplifies the population's wide interest in private vehicle door-to-door transportation services. If ride-hailing companies do indeed find a way to offer an autonomous alternative that is both cheaper and offers even more privacy than current taxis and ride-hailing services (due to the absence of drivers), this might actually shift the current paradigm of privately-owned vehicles to another where most vehicles are owned by an oligopoly of ride-hailing companies and their vehicles are shared across large swaths of the population. Some authors defend that the confluence of ride-hailing services with autonomous vehicles might bring forth several other synergistic benefits, such as improvements in vehicle-miles traveled per person, vehicle occupancy levels, vehicle sizes, and commute lengths ([Greenblatt and Shaheen, 2015](#)). Further reviews of the implications of the use of autonomous vehicles can be found in [Bagloee et al. \(2016\)](#), [Chen et al. \(2016\)](#), and [Fagnant and Kockelman \(2014\)](#).

Whatever direction ride-hailing companies decide to take, it seems inevitable that autonomous vehicles are going to play a central role in their future. That being the case, transportation agencies need to consider scenarios where ride-hailing services use both human-driven and autonomous vehicles.

Conclusions

Ride-hailing services have shown to be widely successful across the globe in attracting new customers. This is likely due to a confluence of several factors, some of the most important ones being the convenience and comfort they bring and their arguably low costs to users. The introduction of this new service into the transportation landscape has brought forth both positive and negative consequences, such as providing increased accessibility to transit deficient locations, while also disrupting and upsetting the taxi market. They seem to have curtailed private vehicle travel for certain use cases (e.g., avoiding driving while intoxicated), but their impact on overall private vehicle usage remains comparatively negligible. Ride-hailing can have both complementary and substitutive effects on transit, but the dominance of one effect over the other seems to vary on a case-by-case basis. The future of ride-hailing systems is still somewhat unclear: companies seem to be strongly invested in developing self-driving autonomous vehicles, but it is currently impossible to tell if this will effectively lead to a significant shift away from private vehicle ownership and toward a true sharing economy. Academics are starting to piece together what currently drives people to use ride-hailing services and how they engage in differently priced tiers of service (such as private versus pooled), but understanding the long term effects of these services on travel demand is likely going to require much more research.

See Also

Uber Versus Taxis; TNCs and the Future of Taxi Public Transport; Taxi Services and Microtransit; Traffic Safety and Security of Taxis and Ride-Hailing Vehicles; Autonomous Goods Transport; Autonomous Vehicles and Transportation Modeling; Vehicles that Drive Themselves: What to Expect with Autonomous Vehicles; How will Autonomous Vehicles Impact Car Ownership and Travel Behaviour

References

- Alemi, F., Circella, G., Handy, S., Mokhtarian, P., 2018. What influences travelers to use Uber? Exploring the factors affecting the adoption of on-demand ride services in California. *Travel Behav. Soc.* 13, 88–104.
- Alemi, F., Circella, G., Mokhtarian, P., Handy, S., 2019. What drives the use of ridehailing in California? Ordered probit models of the usage frequency of Uber and Lyft. *Transport. Res. C Emer. Technol.* 102, 233–248.
- Babar, Y., Burtch, G., 2017. Examining the Impact of Ridehailing Services on Public Transit Use. Available from: papers.ssrn.com.
- Bagloee, S.A., Tavana, M., Asadi, M., Oliver, T., 2016. Autonomous vehicles: challenges, opportunities, and future implications for transportation policies. *J. Modern Transport.* 24 (4), 284–303.
- Bershidsky, L., 2018. Europe Teaches Uber to Do Business Better. Bloomberg. Available from: www.bloomberg.com.
- Che, J., 2015. 9 Countries that aren't Giving Uber an Inch. HuffPost. Available from: www.huffpost.com.
- Chen, T.D., Kockelman, K.M., Hanna, J.P., 2016. Operations of a shared, autonomous, electric vehicle fleet: Implications of vehicle & charging infrastructure decisions. *Transport. Res. A Policy Pract.* 94, 243–254.
- City of Chicago, n.d. Transportation Network Providers - Trips. Available from: data.cityofchicago.org.
- City of New York, n.d. TLC Trip Record Data. Available from: www.nyc.gov.
- Clewlow, R.R., Mishra, G.S., 2017. Disruptive transportation: the adoption, utilization, and impacts of ridehailing in the United States, Technical Report UCD-ITS-RR-17-07, University of California, Davis, Institute of Transportation Studies.

- Dias, F.F., Kim, T., Bhat, C.R., Pendyala, R.M., Lam, W.H. K., Pinjari, A.R., Srinivasan, K.K., Ramadurai, G., 2019. Modeling the evolution of ride-hailing adoption and usage: a case study of the Puget Sound region. Submitted to the Transportation Research Board 99th Annual Meeting.
- Dias, F.F., Lavieri, P.S., Garikapati, V.M., Astroza, S., Pendyala, R.M., Bhat, C.R., 2017. A behavioral choice model of the use of car-sharing and ride-sourcing services. *Transportation* 44 (6), 1307–1323.
- Dias, F.F., Lavieri, P.S., Kim, T., Bhat, C.R., Pendyala, R.M., 2018. Fusing multiple sources of data to understand ride-hailing use. *Transport. Res. Rec.* 2673 (6), 214–224.
- DiDi, n.d. GAIA Open Dataset. Available from: outreach.didichuxing.com.
- Fagnant, D.J., Kockelman, K.M., 2014. The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. *Transport. Res. C Emer. Technol.* 40, 1–13.
- Federal Highway Administration, 2017. 2017 National Household Travel Survey. Available from: nhts.ornl.gov.
- Fitzsimmons, E.G., 2018. Suicides Get Taxi Drivers Talking: 'I'm Going to be One of Them'. *The New York Times*. Available from: www.nytimes.com.
- Gerte, R., Konduri, K.C., Eluru, N., 2018. Is there a limit to adoption of dynamic ridesharing systems? Evidence from analysis of Uber demand data from New York City. *Transport. Res. Rec.* 2672 (42), 127–136.
- Greenblatt, J.B., Shaheen, S., 2015. Automated vehicles, on-demand mobility, and environmental impacts. *Curr. Sustainable/Renewable Rnergy Rep.* 2 (3), 74–81.
- Hampshire, R.C., Simek, C., Fabusuyi, T., Di, X., Chen, X., 2018. Measuring the impact of an unanticipated suspension of ride-sourcing in Austin, Texas. Submitted to the Transportation Research Board 97th Annual Meeting.
- Hartmans, A., Leskin, P., 2019. The History of How Uber Went from the Most Feared Startup in the World to Its Massive IPO. *Business Insider*. Available from: www.businessinsider.com.
- Herskovitz, J., 2017. Texas Lawmakers Clear Way for Uber, Lyft Return to Major Cities. *Reuters*. Available from: www.reuters.com.
- Hu, W., 2017. Taxi Medallions, Once a Safe Investment, Now Drag Owners into Debt. *The New York Times*. Available from: www.nytimes.com.
- Kokalitcheva, K., 2016. Uber Completes 2 Billion Rides. *Fortune*. Available from: www.fortune.com.
- Komanduri, A., Wafa, Z., Proussaloglou, K., Jacobs, S., 2018. Assessing the impact of app-based ride share systems in an urban context: Findings from Austin. *Transport. Res. Rec.* 2672 (7), 34–46.
- Kooti, F., Grbovic, M., Aiello, L.M., Djuric, N., Radosavljevic, V., Lerman, K., 2017. Analyzing Uber's ride-sharing economy. In: *Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee*, pp. 574–582.
- Lavieri, P.S., Bhat, C.R., 2019a. Investigating objective and subjective factors influencing the adoption, frequency, and characteristics of ride-hailing trips. *Transport. Res. C Emer. Technol.* 105, 100–125.
- Lavieri, P.S., Bhat, C.R., 2019b. Modeling individuals' willingness to share trips with strangers in an autonomous vehicle future. *Transport. Res. A Policy Pract.* 124, 242–261.
- Lavieri, P.S., Dias, F.F., Juri, N.R., Kuhr, J., Bhat, C.R., 2018. A model of ridesourcing demand generation and distribution. *Transport. Res. Rec.* 2672 (46), 31–40.
- Lavieri, P.S., Garikapati, V.M., Bhat, C.R., Pendyala, R.M., Astroza, S., Dias, F.F., 2017. Modeling individual preferences for ownership and sharing of autonomous vehicle technologies. *Transport. Res. Rec.* 2665 (1), 1–10.
- Lee, D., 2017. What Happened in the City that Banned Uber. *BBC*. Available from: www.bbc.com.
- Licea, M., Ruby, E., Harshbarger, R., 2015. More Uber Cars than Yellow Taxis on the Road in NYC. *New York Post*. Available from: www.nypost.com.
- Mekelburg, M., 2016. Austin's Proposition 1 defeated. *The Texas Tribune*. Available from: www.texastribune.org.
- Newcomer, E., 2017. Uber Starts Charging What It Thinks You're Willing to Pay. *Bloomberg*. Available from: www.bloomberg.com.
- Office of the Texas Governor, 2017. Governor Abbott signs bill ending local regulations of transportation network companies. Available from: www.gov.texas.gov.
- Rayle, L., Dai, D., Chan, N., Cervero, R., Shaheen, S., 2016. Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. *Trans. Policy* 45, 168–178.
- RideAustin, 2017. Rideaustin Data. Available from: www.data.world.
- Rosenthal, B.M., 2019. They were Conned: How Reckless Loans Devastated a Generation of Taxi Drivers. *The New York Times*. Available from: www.nytimes.com.
- RSG Inc., 2018. San Diego regional transportation study. vol. 1. Technical Report. Available from: www.sandag.org.
- Samuels, A., 2017. Senate Sends Bill Creating Statewide Ride-hailing Regulations to Governor. *The Texas Tribune*. Available from: www.texastribune.org.
- Schaller, B., 2018. The new automobility: Lyft, Uber and the future of American cities. Available from: <https://trid.trb.org/view/1527868>.
- Shoup, D., 2017. *The High Cost of Free Parking: Updated Edition*. Routledge, New York.
- Uber, 2015. One in a Billion. Available from: www.uber.com.
- Uber, 2018. 10 Billion. Available from: www.uber.com.
- Van Gelder, L., 1996. Medallion Limits Stem from the 30's. *The New York Times*. Available from: www.nytimes.com.
- Vinayak, P., Dias, F.F., Astroza, S., Bhat, C.R., Pendyala, R.M., Garikapati, V.M., 2018. Accounting for multi-dimensional dependencies among decision-makers within a generalized model framework: an application to understanding shared mobility service usage levels. *Trans. Policy* 72, 129–137.
- Walker, A., 2018. In NYC, 139 Prized Yellow Taxi Medallions will Hit the Auction Block. *Curbed New York*. Available from: ny.curbed.com.
- Wang, H., Yang, H., 2019. Ridesourcing systems: a framework and review. *Trans. Res. B Methodol.* 129, 122–155.
- Wear, B., 2017. Lyft and Uber return to Austin as Abbott signs ride-hailing bill. *Austin American-Statesman*. Available from: <https://www.statesman.com>.
- Wenzel, T., Rames, C., Kontou, E., Henao, A., 2019. Travel and energy implications of ridesourcing service in Austin, Texas. *Transport. Res. D Trans. Environ.* 70, 18–34.
- Zhang, Y., Zhang, Y., 2018. Exploring the relationship between ridesharing and public transit use in the United States. *Int. J. Environ. Res. Public Health* 15 (8).

Transportation Modeling and Planning Software

Joel Freedman, RSG, Portland, OR, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	569
Historical Overview	569
Transportation Modeling Software Data and Features	570
Core Data Elements	570
Core Software Features	571
Additional Features	572
Activity-Based Models and the Open-Source Revolution	572
Future Directions for Transportation Planning and Modeling Software	573
References	573
Further Reading	573

Introduction

Transportation modeling and planning software is as vast as the field of transportation planning itself. Software packages have been developed to aid practically every aspect of transportation modeling and planning:

- the application of four-step and activity-based travel demand models;
- the application of traffic microsimulation models and to aid in the timing of traffic signals;
- the estimation of demand for and benefits of transportation demand management strategies;
- the estimation of air quality impacts of planned transportation infrastructure; and
- the application of strategic planning models.

This chapter focuses on software developed specifically for application of four-step and activity-based travel models. Such software goes back to the very beginnings of transport modeling, which is described in the first section on the history of travel modeling and planning software. The chapter then describes the main features of modern transport modeling software. The chapter concludes with recent developments and future directions of modeling software.

Historical Overview

The Detroit Metropolitan Area Transportation Study (DMATS) in the early to mid-1950s is cited as the first application of a travel demand model to a regional transportation plan in the United States (Boyce and Williams, 2015). The models were implemented on the IBM 407 accounting machine, capable of simple calculations and utilizing punch cards as input and paper output. Model calculations were only partially implemented using the computer; the rest of the calculations were performed manually. The Chicago Area Transportation Study of the mid to late 1950s advanced travel forecasting practice significantly beyond DMATS by utilizing Moore's (1957) shortest route algorithm () to the Chicago area network and zone system, coupled with an incremental assignment procedure (Boyce and Williams, 2015). The models were implemented on an IBM 704 mainframe computer. A capacity constrained 24-hour one-iteration assignment with 584 zones took 7 h to complete.

In the 1960s, the Bureau of Public Roads (BPR) developed a set of procedures and "batteries" of mainframe computer programs for the implementation of travel demand models. These focused mostly on road planning but included rudimentary mode split models. These programs drew heavily from the DMATS and CATS work as well as other travel models developed for Washington DC and other regions. The US Department of Housing and Urban Development developed a suite of BPR-compatible programs specifically for transit modeling, including transit network path building, transit skimming and assignment, as well as mode split capabilities. These programs became the basis of the Urban Transportation Planning System (UTPS) software in the 1970s. UTPS consisted of 13 software modules that ran in batch mode on IBM 360/370 series of computers. It included highway and transit network analysis models, demand forecasting models, matrix manipulation, and limited graphics capabilities. These programs were refined and added to over time, forming the basis for several other mainframe-based modeling software programs.

It should be noted that during the initial development stage of these modeling tools, mainframe computers were too expensive to be owned by planning agencies. This required the agency to purchase time on mainframes owned by service bureaus, corporate computing facilities, large universities, or federal agencies. Such requirements surely limited both the complexity of early travel demand models as well as their widespread adaptation.

The availability of microcomputers starting in the late 1970s was a turning point for the availability of travel demand models and choices of travel modeling software packages. Modeling software was now being developed and sold to government agencies,

research institutions, and consulting firms by private entities. Among the packages initially developed in the 1980s for microcomputers were EMME/2, MINUTP, QRS II, SATURN, System II, TMODEL2, TRANPLAN, TransCAD, TRIPS, and VISUM/VISEM. Most of these software packages were available for IBM-PC, but some (EMME/2 and VISUM/VISEM) were programmed for the Unix operating system due to the increased computing power of Unix compared to MS-DOS at that time. Interestingly, nearly all of these packages are available today in some form. For example, INRO consultants now sell EMME; MINUTP, TRANPLAN, and TRIPS became the basis for Citilabs Cube software. SATURN is sold by Atkins, PTV sells VISUM/VISEM, and Caliper sells TransCAD. Another transportation modeling package, [Aimsun](#), grew out of work to develop a traffic microsimulation tool by faculty at the Polytechnic University of Catalonia in the 1990s, and now offers both microscopic and mesoscopic modeling capabilities similar to other packages mentioned earlier.

Transportation Modeling Software Data and Features

Many if not all transportation modeling software packages have several core features and functionalities required for modern transportation planning and modeling. Core data elements of transportation software packages are listed as follows.

Core Data Elements

- **Zones:** Travel models typically aggregate space into Transportation Analysis Zones, or TAZs. These are typically around the size of a Census Tract or a Census Block Group, though potentially smaller. Zonal data describes the quantity or quality of land-use and travel opportunities in each TAZ. TAZ data typically include households by type, employment by type, population, school enrollment, square footage of space by type, parking costs, terminal times (time required for picking up and dropping off passengers), and fields used to identify special destinations such as airports, hospitals, and major universities. Zones can also be used to group smaller sets of zones into larger areas referred to as districts.
- **Network:** A representation of the transport system. Common network elements include:
 - **Nodes:** Representation of a link end point. Typically located at road intersections, changes in road attributes (speed, number of lanes, etc.), transit stops, and to connect regular links to centroid connectors. Nodes used to be used to give “stick” networks shape, but this is typically not necessary any longer due to the use of shape points in most modern transport software packages. Common node attributes include the Node ID, XY location, an indicator identifying whether the node is a centroid, and one or more other attributes useful for modeling purposes (whether the intersection is signalized, permitted turning movements, etc.).
 - **Centroids:** A special type of node used to represent demand loading points on the network. Typically located at the center of a zone. Some software packages (e.g., Cube) require that centroid nodes be sequentially numbered from 1 to N where N is the total number of zones.
 - **Transit stops:** A special type of node used to represent a boarding and/or alighting location for transit.
 - **Links:** A representation of the road network, bike network, and/or pedestrian network. A link is defined by its endpoints, typically referred to as the “A” node and the “B” node. Link attributes include the number of the endpoint nodes, number of lanes, facility type, free-flow speed, road name, link capacity, traffic count (used for validating model outputs), and so on. Some software require a separate link for each direction of travel, whereas other software allows for special codes to identify one-way versus two-way links. Special links are also often created for transit; for example, transit-only links are used for routing grade-separate transit routes, and transfer links are used by transit path-builders to find a transfer connection between transit stop nodes.
 - **Transit routes:** Transit routes typically are specified via a route table and a route itinerary table. The route table describes the route metadata, and typically includes the route ID, route name, direction, fare, and headway for each time period. The route itinerary table describes the route alignment by listing the links and/or nodes that the route traverses from its starting point to its ending point. The route itinerary can optionally include the specific boarding and/or alighting movement allowed at each stop node, the travel time between nodes (particularly for grade-separated transit), and other relevant data.
 - **Matrices:** Two-dimensional data tables where each row and column correspond to a TAZ. Used for storing trips between TAZs output from trip distribution and mode choice models, as well as travel time, distance, and cost between zones output from path-building procedures. The latter are commonly referred to as network “skims.” Trip tables are often segmented by some combination of socioeconomic market segment, trip purpose, mode, and/or time period. It should be noted that trip tables from four-step travel models could be represented in “production-attraction format”¹ or “origin-destination format,” depending upon the

¹ The production end of a trip is the home end of home-based trips and the origin of a non-home-based trip. The attraction end of a trip is the non-home end of a home-based trip and the destination of a non-home-based trip. A production-attraction format trip table is one in which rows are production zones and columns are attraction zones.

type of model or process that created it. Skims are often segmented by time of day. Transit skims typically separate in-vehicle time from out-vehicle time, and include number of boardings or transfers as additional attributes.

- Volume-delay functions: Volume-delay functions are used to reflect the effects of congestion on travel time. One of the first and perhaps still the most common volume-delay function (VDF) was published by the BPR ([Bureau of Public Roads, 1964](#)). Most modern software allows the user to specify any type of VDF, of which there are many.
- Turning movement prohibitors: Specifies disallowed turning movements, or travel time penalties for specific movements. Typically specified by from node, through node, and to node.

Core Software Features

The following section lists main features and capabilities of modeling software.

- Data input/output: Most software has capabilities to import and export networks from shapefile format, comma-separated value (CSV) format, database format (DBF), Excel (XLSX) Open Street Map (OSM), General Transit Feed Specification, and/or other transportation software programs. Tabular and matrix data can often be imported and exported from/to CSV, DBF, XLSX, and other formats. Background layers (OSM, Google Maps, satellite images) can be loaded from the web or disk.
- Network editing: A graphical tool for displaying and editing network features. This tool is used for adding and removing links, nodes, and transit routes, adding to or changing the attributes of links, nodes or routes, etc.
- Network calculations: A tool for calculating network attributes, such as link capacity, from other attributes, such as number of lanes and other attributes. Network calculations are often used to automate modeling procedures, such that the user can rely on programmed calculations rather than manually change multiple data items.
- Matrix calculations: A tool for creating and calculating matrix data. For example, performing mathematical operations on matrices (adding, subtracting, multiplying matrices), and transposing matrices². Matrix calculations can be used to apply travel demand models, such as gravity models and logit mode choice models. Iterative proportional fitting (IPF), known as Fratar adjustment for two-dimensional matrices, is a special type of matrix calculation required for doubly constraining gravity models but has other applications in travel demand modeling as well.
- Traffic path building and assignment: Tools for finding paths between nodes and/or TAZs through road, bike, or pedestrian networks, and assigning demand to those networks. Inputs include networks, demand matrices, and volume-delay functions. Outputs may include congested travel times on links, network skims, and link volumes.
- Transit assignment: Tools for finding paths between nodes and/or TAZs through transit networks, and assigning demand to those networks. Inputs include road and transit networks, transit demand matrices, and transit delay functions. Outputs may include transit skims, route level and stop level boardings and alightings, and transfer matrices.
- Select link/route tracing: The analyst may be interested in identifying the demand associated with one or more selected links or transit routes. Or they may be interested in tracing the routes used by one or more zone pairs through the assignment process. Such analysis is typically required in the case of traffic impact studies or project-specific analysis. In either case, most software packages provide this capability.
- Visualizing and reporting results: Travel demand models are a tool that, if successful, provides useful insights into the alternatives and policies being modeled. To that end, the visualization and reporting of results is a valuable feature of modern transportation modeling software packages. Following are a few common types of reports and graphics available from most modeling software packages:
 - Link bandwidth plots: Plots in which the thickness of the link is displayed according to the link flow, and the color of the link is displayed according to the volume to capacity ratio on the link. This plot is useful for understanding roads with high demand and/or congestion.
 - Route passengers and boardings: Plots of the transit network where each route segment is scaled to passenger flow. Transit stops can also be scaled or colored by boardings/or alightings.
 - Thematic maps: Typically used to display zonal or other network data not mentioned earlier (i.e., household density, trips by zone, link capacity, etc.)
 - Trip length frequency distributions: A matrix histogram or bar chart, in which the x-axis is trip distance and the y-axis is the absolute number or percent of trips in each distance bin. These plots are useful for understanding how far people are traveling, often by trip purpose.
 - Desire lines: A visualization of a matrix showing flows between zones or districts as lines whose width is scaled to demand.
 - Other statistics and reports: Vehicle miles of travel, vehicle hours of travel, transit boardings by route, trips by mode, trips by distance, and many other statistics are useful outputs from travel demand models and are commonly available from transport modeling software.
- Automation: Travel models can take hours if not days to run from start to finish. Therefore automation is a required core feature to ensure replicability, accuracy, and efficiency and reduce labor costs and fatigue. Various software packages have implemented

² A transpose matrix is one in which the rows and columns of the original matrix are flipped. A common application of matrix transposition is when changing a matrix format from production-attraction to origin-destination format.

automation in different ways. For example, both Cube and TransCAD have developed their own scripting languages, though TransCAD also provides an application programming interface (API) that can be called from other modern programming languages. EMME and VISUM/VISEM rely upon Python as the main scripting language, and provide an API for software data program functionality. The model developer and/or user must implement each model component either in the modeling software scripting language or in the API language when developing or enhancing the model. The model developer must ensure that the model steps can run for different scenarios and provide outputs that provide useful insights for decision-making.

Additional Features

Over time, as software has evolved and user needs have become more complex, software developers have added a number of additional features to enhance the usability of their software. A few examples are provided as follows.

Geographic Information System (GIS) integration: Modern transportation modeling software allows the user to perform common GIS functionality: intersect, union, buffer, tag to nearest, etc. across various point, line, and polygon layers. This functionality is either provided as a core feature of the software or through an add-on such as an ESRI ArcMap license.

Scenario management: Analysts often code dozens or even hundreds of scenarios when working on a metropolitan transportation plan or project analysis. Scenario management tools have been developed to organize different versions of networks and inputs for each scenario. The user often has the ability to select various bundles of projects to include in a specific scenario, limiting the amount of recoding required to apply the same project across multiple scenarios.

Subarea extraction: Often planners working in larger metropolitan regions need to carve out a subarea from the regional model, adding network and zonal detail to that smaller area for a specific planning study such as a road project, traffic impact study, or community plan. To enable this procedure, software vendors often provide a subarea procedure where the analyst identifies the boundaries of the area of interest and the zones to be included in the subarea, and the software extracts a subarea network from the regional network along with demand associated with that subarea. The software may additionally allow disaggregation from more aggregate to more disaggregate TAZ systems within the subarea as well as aggregation of demand at external stations outside the subarea, based on the regional assignment.

Origin-destination matrix estimation (ODME): ODME is a process whereby a base demand matrix is adjusted to match traffic counts through an iterative process consisting of highway assignment, comparison to counts, and matrix algebra to improve goodness-of-fit. Although there is considerable debate about the usefulness and validity of such automated adjustments, ODME functionality has been added to most software packages to provide this option to those who want it.

Expansion of choice models and algorithms: Software vendors are constantly moving research into practice in order to maintain their competitive edge and provide their users increased capabilities and faster runtimes. Software packages now support a multitude of highway assignment techniques: Frank-Wolfe static equilibrium, stochastic, origin-based assignment, etc. Transit assignment methods have also increased to include shortest path, stochastic, and frequency-based assignments. Some software packages have added automated tools for applying multinomial and nested logit choice models.

Support for disaggregate models: Due to the increasing adoption of activity-based models, software vendors have added support for various aspects of disaggregate travel models, such as record-based processing, expanding their data models to consider household and person data, simulating from probabilities, and so on. Despite these features, most activity-based models continue to be implemented in stand-alone software.

Cloud computing: In order to reduce runtime and agency IT costs for large-scale travel models, many software packages are offering either cloud computing services through the company or network licenses, which allow the software to be deployed in the cloud. Typically additional configuration is required to support distributed computing and not all algorithms are distributable at this time.

Activity-Based Models and the Open-Source Revolution

As discussed earlier, many of the software packages in use today have their roots in the four-step models of the 1970s and 1980s. Starting in the 1990s, activity-based models began to emerge as a practical alternative to the four-step approach. Their adoption was motivated by a desire to model behaviors and policies that are not addressed by the majority of trip-based models; trip-linking, activity (re)scheduling in response to congestion, travel demand management, certain pricing policies, and so on. As these models required model forms and application procedures that did not exist or at best were very cumbersome to implement in most transport modeling packages (list-based processing, simulation from probabilities, choice models with thousands of alternatives, etc.), activity-based modelers developed their own software to implement their models. In such cases the software became intrinsically tied to the specific model implementation—unlike commercial transport software in which any given package can be used in dozens or hundreds of different regions with significantly different model features. Some of the more commonly applied activity-based models for planning agencies in the United States are DaySim, CT-RAMP, and TourCast. Each software package relies on one of the commercial transport modeling packages for network and matrix calculations, such as network skimming and assignment, as well as application of aggregate models for commercial vehicles and/or other special markets (Davidson et al., 2007).

We are now seeing an increasing interest in the development of open-source tools in the transport modeling industry. Two popular software packages for data science and statistics—Python and R—are increasingly being applied to analyze transport data

and implement travel demand models. Both software tools are firmly rooted in the open source paradigm, with large user communities who develop and share libraries for specific applications that can be downloaded and applied by other developers. The ActivitySim project is an example of how the open-source philosophy is gaining traction in the modeling community. The mission of the ActivitySim project is to create and maintain advanced, open-source, activity-based travel behavior modeling software based on best software development practices for distribution at no charge to the public. The ActivitySim project is led by a consortium of metropolitan planning organizations and other transportation planning agencies, as opposed to a commercial entity (Association of Metropolitan Planning Organizations Research Foundation, 2019). MatSim, an open-source framework for implementing large-scale agent-based transport simulations is another example (Horni et al., 2016).

Future Directions for Transportation Planning and Modeling Software

Travel modeling software vendors face significant challenges due to the evolving demands of transportation planners and decision-makers. Increasingly the focus of transportation planning is turning from analysis of additional capacity to management, maintenance, and reliability of the existing system. The emergence of Mobility-as-a-Service, the increasing use of transportation network companies as an alternative to the use of the privately owned automobile, the substitution of internet shopping and home delivery for shopping travel, and the potential availability of autonomous vehicles poses additional questions about equity, congestion, and quality of life. Travel modelers increasingly recognize that four-step travel models coupled with static equilibrium assignments are incapable of accurately measuring congestion due to increasing demand, system reliability, fleet management, and how emerging technologies could potentially revolutionize how, when, and how much people travel. To address these needs, software vendors are increasingly investing in new methods to model travel. Such features include integration of dynamic traffic assignment software and/or traffic microsimulation software and vehicle routing models within their modeling platforms. It remains to be seen, however, whether the traditional transportation modeling software with roots in the planning processes of the 50s and 60s can evolve to keep up with the planning questions of tomorrow.

References

- Aimsun. Available from: <https://www.aimsun.com/aimsun-next/>.
- Association of Metropolitan Planning Organizations Research Foundation, 2019. Available from: <https://activitysim.github.io/>.
- Boyce, D.E., Williams, H.C.W.L., 2015. *Forecasting Urban Travel: Past, Present and Future*. Edward Elgar Publishing, Cheltenham, UK.
- Davidson, W., Donnelly, R., Vovsha, P., Freedman, J., Ruegg, S., Hicks, J., Castiglione, J., Picado, R., 2007. Synthesis of first practices and operational research approaches in activity-based travel demand modeling. *Trans. Res. A* 41, 464–488.
- Horni, A., Nagel, K., Axhausen F.K.W., 2016. *The Multi-Agent Transport Simulation MATSim*. Ubiquity Press, London UK.
- Moore, E.F., 1957. The shortest path through a maze, *International Symposium on the Theory of Switching*, Harvard University; Proceedings, Part II, *Annals of the Computation Laboratory of Harvard University*, vols. 29–30, pp. 285–292.
- Bureau of Public Roads, 1964. *Traffic Assignment Manual*, US Department of Commerce, Urban Planning Division, Washington, DC.

Further Reading

- Caliper Corporation, 2019. *TransCAD Program Help*, Caliper Corporation, Newton, MA..
- Citilabs, 2018. *CubeVoyager Reference Guide*, Version 6.4.4, Citilabs, Tallahassee, FL..
- INRO, 2019. *Model User Guide*, INRO, Montreal, Canada..
- PTV, 2019. *PTV VISUM Manual*, PTV, Karlsruhe, Germany..

Transportation Statistics and Databases

Taha Hossein Rashidi, UNSW Sydney, Kensington, NSW, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	575
Statistical Methods	575
Akaike's Information Criterion	575
Analysis of Variance (ANOVA)	575
Anderson–Darling Tests	576
Autocorrelation in Regression	576
Bayesian Nonparametric Statistics	576
Bayesian <i>P</i> Values	577
Best Unbiased Estimation	577
Binomial Distribution	577
Bivariate Distributions	577
Bootstrap Methods	578
Chi-square Distribution	578
Confidence Interval	578
Copulas	578
Degrees of Freedom	578
Dummy Variables	579
Durbin–Watson Test	579
Entropy	579
Expected Value	579
Extreme Value Distributions	579
F Distribution	579
Factor Analysis and Principal Component Analysis	579
Hypothesis Testing	580
Gamma Distribution	580
Geometric and Negative Binomial Distributions	580
Heteroscedasticity	580
Identifiability	581
Kolmogorov–Smirnov Test	581
Likelihood Ratio Test	581
Lognormal Distribution	581
Markov Chain Monte Carlo	581
Mixture Models	582
Moving Averages	582
Multicollinearity	582
Normal Distribution, Univariate, or Multivariate	582
Poisson Distribution and its Application in Transportation	583
<i>P</i> Values	583
Sample Size Determination	583
Sampling Algorithms	583
Student's <i>t</i> -Tests	583
Uniform Distribution	584
Weibull Distribution	584
Databases	584
Crosssectional Data	584
Panel Data	584
Revealed Preference	584
Stated Preference Data	585
Survival Data	585
References	586

Introduction

Transportation modeling as a science, or a profession is entangled with statistical methods being used to discover patterns in the data provided by users of the transportation system or the system itself. Transportation data and models are used to develop understanding about the performance of the system when characteristics of the system and users are changed, mainly for forecasting purposes. Therefore, enhancing the quality of data and modeling techniques which are used for development of transportation models, that can then assess the potential effectiveness of policies for forecasting, is essential for planning purposes. This significance gets further amplified when we know that the transportation network is the most expensive infrastructure that is still being managed by the public sector. Therefore, it is crucial to identify, and measure the value of the most beneficial strategies for improving the level of service and experience on the transportation network as well as their impacts on the well being of users and nonusers. Evolving and constantly advancing statistical methods and data sources are the backbone of the process of development of transportation models for forecasting transportation related indicators such as travel time, reliability of network, traveling cost, and number of users on the network. In this article the most fundamental approaches used in different discipline of the large field of transportation modeling and engineering are discussed. The first section of the article provides a list of major tests, concepts, models, and techniques, which are commonly used by transportation modelers as well as some discussion on the surveying methods to be used for collecting these data sources. The second section discusses the types of databases that are used.

Statistical Methods

Akaike's Information Criterion

As an information criterion and a goodness-of-fit estimator, AIC measures the quality of a statistical model based on the best (maximum) log-likelihood value that can be obtained given a specific number of parameters. Therefore, AIC falls into the set of estimators that are adjusted based on the number of parameters to avoid overfitting. The main use of AIC is for model selection for a given data, where lower values of AIC are considered better, however, AIC is silent about how better a model with lower AIC is. For a given model, with p parameters and log-likelihood value of $LL(\beta)$ for β as the vector of parameters of size p , $AIC = -2[LL(\beta) - p]$.

Analysis of Variance (ANOVA)

The analysis of variance is closely related to linear regression analysis. So, first we consider a simple linear regression model as below, and then discuss ANOVA from the linear regression point of view.

$Y_i = \alpha + \beta X_i + \varepsilon_i$, where Y_i is the dependent variable, α is the constant or intercept of the linear function, β is a $[1 \times p]$ vector of parameters for X_i , which is a $[p \times 1]$ vector of explanatory variables and ε_i is the unobserved residual for individual i . α and β can be estimated using methods such as ordinary least square.

One property of the linear regression model is that the line passes through the sample means of $\bar{Y}$ and $\bar{X}$. In other words, $\bar{Y} = \alpha + \beta \bar{X}$. Therefore, we can subtract both sides of the main regression equation from the sample mean equation to have $Y_i - \bar{Y} = \beta(X_i - \bar{X}) + \varepsilon_i$. By defining new variables that are based on the deviation from the mean, we have $y_i = \beta x_i + \varepsilon_i$, where $y_i = Y_i - \bar{Y}$ and $x_i = X_i - \bar{X}$. Then by squaring both sides of the regression function for deviation from the mean model and summing over the sample data we have:

$$\sum_i y_i^2 = \beta^2 \sum_i x_i^2 + \sum_i \varepsilon_i^2 + 2 \sum_i \beta x_i \times \varepsilon_i.$$

Then we can say $\sum_i y_i^2 = \beta^2 \sum_i x_i^2 + \sum_i \varepsilon_i^2$ because $\sum_i \beta x_i \times \varepsilon_i = 0$. Now we define $\sum_i y_i^2 = \sum_i (Y_i - \bar{Y})^2$ as the total sum of squares (TSS), $\beta^2 \sum_i x_i^2$ as the explained sum of squares (ESS), and $\sum_i \varepsilon_i^2$ as the residual sum of squares (RSS). This means TSS to be equal to RSS + ESS. The study of these three terms is called ANOVA. Assuming a sample of size n , TSS has a degree of freedom of $n - 1$ because it reflects the deviation of the sample data from its mean. RSS has $n - p - 1$ degrees of freedom where p is the size of the vector of parameters that are estimated. ESS has p degrees of freedom, which is the number of parameters estimated. By dividing these three terms by their degrees of freedom, we obtain the mean sum of squares, which can then help us define the F variable, which follows an F-distribution.

$$F = \frac{MSS \text{ of ESS}}{MSS \text{ of RSS}} = \frac{\beta^2 \sum_i x_i^2 / p}{\sum_i \varepsilon_i^2 / (n - p - 1)} = \frac{\beta^2 \sum_i x_i^2 / p}{\sigma^2},$$

where σ^2 is the variance of the error term. The F ratio can be used to test the hypothesis that β is meaningfully contributing to explain the dependent variable. The well-known ANOVA table is commonly generated by commercial statistical packages (Gujarati, 2009)

ANOVA Table

Source of variation	Sum of square	df	Mean sum of squares
ESS	$\beta^2 \sum_i x_i^2$	p	$\beta^2 \sum_i x_i^2 / p$
RSS	$\sum_i e_i^2$	$n - p - 1$	$\sum_i e_i^2 / n - p - 1$
TSS	$\sum_i y_i^2$	$n - 1$	

$$F = \frac{\beta^2 \sum_i x_i^2 / p}{\sum_i e_i^2 / n - p - 1}$$

Anderson–Darling Tests

Anderson–Darling (AD) test belongs to the category of distance-based tests for examining the similarity of two distributions or a sample belonging to a distribution.

If X is a random variable or $(x_1, \dots, x_n)$ with cumulative density function of $F_s(X)$ which it is tested against another distribution $F_b(X)$, then the distance between the two distributions can be estimated using the following statistic:

$$D_w^2 = n \int_{-\infty}^{\infty} [F_s(x) - F_b(x)] w[F_b(x)] dF_b(x),$$

where $w[F_b(x)]$ is a positive weight function.

The AD statistic considers $w[F_b(x)]$ to be equal to $1/[F_b(x)[1 - F_b(x)]]$. Further, $F_b(x)$ and $F_s(x)$ should be transformed to uniform distributions to form a simple formulation for the AD statistic when a sample is being tested for a specific distribution. By defining a new random variable $U = F_b(x)$ with $\Pr\{U \leq u\} = u, 0 \leq u \leq 1$ and defining an empirical distribution function for the sample $(x_1, \dots, x_n)$ as $F_s(x) = k/n$ for k from $(x_1, \dots, x_n)$ and the sample being ordered then it can be shown (Anderson and Darling, 1954) that the AD statistic can be estimated by:

$$D_{AD}^2 = -n - \sum_{k=1}^n \frac{2k-1}{n} \{\log[F_b(x_k)] + \log[1 - F_b(x_{n+1-k})]\}.$$

They also showed that the critical value for 5% significance is 2.492 and for 1% is 3.880.

Autocorrelation in Regression

When there is correlation between ordered observations in time or space resulting in the covariance between the error terms being nonzero, autocorrelation occurs. In the context of OLS, where the dependent variable (even if it is observed for multiple times) is regressed only against the explanatory variables (and not the dependent variable observed for the previous time intervals), the estimated parameters remain unbiased (their expected value is equal to the true values), and asymptotically normally distributed. However, the coefficients are not any more efficient (not having the minimum variance) which means the t and F statistics are not any more reliable.

$Y_t = \alpha + \beta X_t + \varepsilon_t$, which can be simply written as $Y_{t-1} = \alpha + \beta X_{t-1} + \varepsilon_{t-1}$ for one time interval earlier than t and can be written for s time intervals earlier as $Y_{t-s} = \alpha + \beta X_{t-s} + \varepsilon_{t-s}$. To examine the first order autocorrelation, $\varepsilon_t = \rho \varepsilon_{t-1} + \varepsilon_t$, ρ can be estimated which is the main factor for examining the strength of autocorrelation as well as correcting the estimated coefficients.

There are some applications of spatial autocorrelation in the area of transport on traffic flow and origin-destination models (Bolduc et al., 1989). In the context of discrete choice modeling, there has been studies trying to capture the spatial autocorrelation into the structure of the utility function (Bhat and Guo, 2004; Bolduc, 1992), either as a systematic component or into the random error component. However, these works do not much discuss the properties of the estimated coefficients such as unbiasedness and efficiency.

When the dependent variable is present on the right-hand side of the equation, a typical time series model is of interest. The commonly known equation for times series can be written as $Y_t = \alpha Y_{t-1} + \varepsilon_t$, which can get complicated by adding a drift (a fixed value) or trend (time as t) to the right hand side. Time series models are implicitly assumed to be developed on the underlying stationary (loosely saying when the mean and variance of the data are fixed over time) data. Then by using autoregressive regression and a moving average (MA) mechanism, unbiased parameters can be estimated. If the data are not stationary, there are ways to transform it to stationary data such as by differencing i times which is denoted in the literature as $I(i)$ which is called integrated of order i .

Bayesian Nonparametric Statistics

Conditional probability of an event given occurrence of another event can be explained by the Bayes' Theorem as: $P\{A|B\} = P\{AB\}/P\{B\}$. This rule is applicable on occurrence of B given A , which then results in $P\{A|B\} = P\{B|A\}P\{A\}/P\{B\}$. As the probability of B can be written as the probability of the product of B and disjoint events whose union cover the whole space, then the previous conditional probability can be reexpressed as $P\{A_k|B\} = P\{B|A_k\}P\{A_k\}/\sum_{j=1}^N P\{B|A_j\}P\{A_j\}$ where $B = \cup_{j=1}^N A_j B$,

$\cup_{j=1}^N A_j = \text{Space}$, and $A_j \forall j = 1, \dots, N$ are disjoint. The last form of Bayes' theorem is the core of Bayesian methods with growing applications in the field of transportation.

Statistical models with infinite number of parameters in their model, which are developed based on Bayesian methods, are called Bayesian nonparametric statistics. Following the Bayesian analysis of statistical models, the parameter space Θ , which reflects the analyst's knowledge, is captured through $\pi(\theta)$. Given available data, shown as Y for the dependent variable and X for the explanatory variables, θ is aimed to be estimated. Y being a random variable is expected to be generated using a random mechanism explained through the posterior distribution. The posterior distribution is conditional on Y and provides updated information for θ as $\pi(\theta|Y)$. In the Bayesian nonparametric context, prior, and posterior distributions are defined in an infinite dimensional parameter space.

A Dirichlet process is parameterized using the concentration parameter α , which is a positive, real-valued to control the variance of the process and the base probability distribution G_0 on Θ . As α approaches zero, the process concentrates around the mean. An important property of the Dirichlet process is that if G denotes a sample generated from a Dirichlet process, $G \sim DP(\alpha, G_0)$, for any finite, mutually exclusive partition $A_1, \dots, A_k$ of the parameter space Θ , exhibits a Dirichlet distribution as $G(A_1), \dots, G(A_k) \sim \text{Dirichlet}(\alpha G_0(A_1), \dots, \alpha G_0(A_k))$, where $G(A_1)$ is the probability mass in partition A_1 . Two main properties of the Dirichlet process for any realizations of it are being discrete and clustered with non-zero probabilities. The formal definition of the Dirichlet process is not applicable to practical applications, which results in three alternative representations of the Dirichlet process have been proposed in the literature as the Blackwell–MacQueen urn scheme, the Chinese Restaurant process and the stick-breaking process construction. For more information on these processes and application of the stick breaking process in the context of shared automated vehicle service choice Krueger et al. (2019) can be consulted.

Bayesian P Values

Borrowing the P values concept from the classical frequentist domain can help Bayesians to assess the goodness of fit of a single model using the previous notation for the prior as $\pi(\theta)$, Θ as the parameter space, $\pi(\theta|Y)$ as the posterior and $g(Y)$ as the goodness-of-fit measure for the data and model $m(Y|\theta)$.

If we define the prior predictive density as $\tilde{p}(Y|y) = \int_{\Theta} m(Y|\theta)\pi(\theta|Y)d\theta$ then the posterior predictive P value can be defined as $p^* = \int_{g(Y) > g(y)} \tilde{p}(Y|y)dy$. As the observed data are used twice in the estimation of p^* , a corrected posterior predictive P value can be estimated using the method discussed by Bayarri and Berger (1999).

Best Unbiased Estimation

An *unbiased* estimator, for example parameters of a linear regression model, is an estimator whose expected value is equal to the true value of that estimator, like mean of a sample. Among all unbiased estimator the one(s) with the minimum variance is called an *efficient* estimator. Also an efficient unbiased estimator is called a *best* estimator.

Binomial Distribution

As one of the popular distributions, binomial distribution explains the chance of s successes in n independent trials where the chance of success is p and the chance of failure is therefore $q = 1 - p$. Then as random variable X , which is a nonnegative integer, is said to follow a binomial distribution $X \sim B(n, p)$ with probability mass function of

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x}, \text{ where } \binom{n}{x} = n! / (x!(n-x)!) \text{ for } 0 \leq x \leq n.$$

Then the cumulative density function of binomial distribution would be $F(x) = P(X \leq x) = \sum_{i=0}^x \binom{n}{i} p^i q^{n-i}$.

The Poisson distribution is derived from the negative binomial distribution for events with low chance of success with so many instances, which is discussed in detail in the Poisson distribution section. Poisson distribution and its generalization of negative binomial are commonly used in modeling count data such as accident and safety analysis.

Bivariate Distributions

Having two random variables correlating with one another requires a joint/bivariate distribution to explain the interaction/dependency between the two variables. The concepts are quite similar but we first start by discussing two continuous random variables where the duplet (X, Y) shows two random variables with a joint probability density function (PDF) of $f(X, Y) > 0$. The joint cumulative density function can be estimated by $F(X, Y) = P(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f(z, w) dz dw$. The marginal pdf with regard to X can be written as $f_x(x) = \int_{-\infty}^{\infty} f(x, w) dw$, and the conditional probability of X given $Y = y$ can be written as $f(x|y) = f(x, y)/f(y)$. In a similar fashion, the probability mass function of X and Y as discrete random variables can be written as $f(x, y) = f(X = x, Y = y)$ and the cumulative mass function as $(X, Y) = P(X \leq x, Y \leq y) = \sum_{-\infty}^x \sum_{-\infty}^y f(z, w)$.

The most commonly used bivariate distribution is normal, which has the following function for the standardized normal function: $f(x, y) = \frac{1}{(2\pi\sqrt{1-\rho^2})} \exp\{-(x^2 - 2\rho xy + y^2)/[1(1-\rho^2)]\}$. In case the joint function does not have a closed form, it can be approximated using copulas, which will be explained later.

Bootstrap Methods

The bootstrap method is a statistical technique to indirectly infer properties of a population by forming small random samples. The random samples are constructed by drawing observations from a large data sample, or some prior information about the distribution of the population, one at a time and returning them to the data sample. This process is called sampling with replacement, which then allows one observation to appear in the sample multiple times. Therefore, the resampled sets are used to infer true properties of the population.

Chi-square Distribution

Chi-square distribution is one of the most popular distributions used in statistics and consequently transportation modeling. It is closely related to the normal distribution and formed as sum of squares of j independent standardized normal variables, which is called as the χ^2 distribution with k degrees of freedom (df). The PDF of χ^2 distribution is:

$$f(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{1}{\Gamma(\frac{n}{2})} \frac{1}{2^{\frac{n}{2}}} \frac{1}{x^{\frac{n}{2}}} e^{-\frac{x}{2}} & \text{if } x > 0 \end{cases},$$

where Γ refers to the Gamma function.

For example, when estimating the variance of a random sample of n independent and identically distributed random variables shown as $\{X_1, \dots, X_n\}$, the variance of the sample can be estimated as $S^2 = \left(\sum_{i=1}^n (X_i - \bar{X})^2\right)/(n-1)$, which has a χ^2 distribution with $n-1$ degrees of freedom.

Confidence Interval

A $100 \times (1-\alpha)$ percent confidence interval is an interval around a sample parameter θ that with $\hat{\theta}$ as the true parameter of the population. It is expected that if repeated random samples of size N are drawn, $100 \times (1-\alpha)$ percent of the time, the interval around θ includes $\hat{\theta}$. This interval is a random variable of type *range*, which may or may not include the true parameter. The interval is determined by the sampling distribution, which is used to specify the probability of observing each value in the sample.

Therefore, for a given sample size N and α , the confidence interval determines the accuracy of the estimated statistic/parameter θ . The statistics determines what method to be used for constructing the confidence regions.

Copulas

Hoefding (1940) introduced the best possible boundaries for such functions as one of the earliest attempts of introducing copulas. However, the term of copula was first used by Sklar (1959). Sklar's work was followed by several others, which led to pervasive usages of copula functions in different fields such as economics (Smith, 2003), finance (Embrechts et al. 2002), medical sciences (Clayton, 1978), statistics (Frees and Valdez, 1998), and more recently transport modeling (Bhat and Eluru, 2009).

The popularity of copula functions owes to four attributes. First, although the normal distribution is the backbone of constructing multivariate distributions, the copula formulations allow for a diversity of other types of functions to capture multivariate distributions among which Gaussian distribution is one. Second, copula distributions allow for linear and nonlinear dependencies, providing both symmetric and asymptotic options. Third, the simple closed form copula functions significantly facilitate their usage in complex contexts. Fourth, various correlation structures can be handled with copula functions (Bhat and Eluru, 2009; Trivedi and Zimmer, 2007). Several copulas have been introduced among which Gaussian, Farlie-Gumbel-Morgenstern, and Archimedean class of copulas such as Frank, Joe, Gumbel, and Clayton are the most popular ones (Bhat and Eluru, 2009). For detailed information and introduction to copulas refer to the work by Bhat and Eluru (2009). More advanced copula functions such as Vine copulas (Joe et al., 2010) and nested copulas (Hossein and Mohammadian, 2016) are becoming interesting in different fields.

Degrees of Freedom

Generally speaking, the concept of degrees of freedom is referred to as the number of times that a statistic can freely vary. Therefore, every required parameter or necessary relations estimated using data for that statistic, reduces the degree of freedom of the random variable by one degree.

Dummy Variables

A dummy variable is a binary variable that can take values of either 0 or 1. Dummy variables are commonly used in regression models to transform categorical variables to numerical values. Dummy variables, especially when interacted with other variables, allow capturing seasonality in the data, taste variation, and sudden shifts in the constant parameter.

Durbin–Watson Test

The Durbin–Watson test is one of the most well-known tests to examine autocorrelation in linear regression. This test can be used to test first-order autoregression $\varepsilon_t = \rho\varepsilon_{t-1} + \varepsilon_t$, where ε_t is the error term of the regression model $Y_t = \alpha + \beta X_t + \varepsilon_t$. To test the hypothesis that $\rho = 0$ or not, given $d = \left(\sum_{t=2}^n (\varepsilon_t - \varepsilon_{t-1})^2 \right) / \left(\sum_{t=1}^n \varepsilon_t^2 \right)$ being the DW statistic, two critical values for a specific confidence interval are required. Assuming that for significance level of $(1-\alpha\%)$, critical values of dw_l and dw_u are provided, statistically significance positive autocorrelation exists if $0 < d < dw_l$ and statistically significance negative autocorrelation exists if $4 > d > dw_u$. We cannot reject the hypothesis of no autocorrelation if $4 - dw_u > d > dw_l$. For other values of d the test is indecisive which is the main drawback of the test.

Entropy

Considering a set of mutually exclusive and exhaustive of random events $\{E_1, \dots, E_N\}$ where each of the j event can be associated with a random variable x_j , the classical definition of entropy can be written as $-\sum_{i=1}^N p_i \log p_i$ where p_i is the probability of event j which can be associated with random variable X_j such that $\sum_{i=1}^N p_i = 1$. If x is a continuous random variable, entropy is defined as $\int_x P(x) \log P(x) dx$. In general entropy reaches its maximum when all probabilities are equally likely. Different usages for entropy have been discussed in the field of transportation such as in land use (Kockelman, 1997) and choice modeling Swait and Adamowicz (2001).

Expected Value

Expected value of a discrete/continuous random variable is defined as the sum/integral of that variable over all possible values of the variable weighted by their probability of happening. Therefore, it can be estimated as follows for discrete and continuous random variables.

$$\text{Discrete for } N \text{ observations : } \sum_{i=1}^N p_i x_i$$

$$\text{Continuous over } x : x \int_{-\infty}^{\infty} x f(x) dx$$

Extreme Value Distributions

Similar to the central limit theorem that talks about one property of a sample with a finite variance, the extreme value distributions are relevant to a sample with a converging normalized maximum. For a sample with a converging normalized maximum (as the sample size goes to infinity), the normalized maximum (which is a random variable) has a PDF, which belongs to either the Gumbel (Type I), the Fréchet (Type II) or the Weibull (Type III) family. These three distributions are grouped together into the generalized extreme value distribution family. Gumbel and Weibull family distributions are commonly used in the field of transportation in discrete choice modeling (Train, 2009) and hazard-based analysis (Ghasri et al., 2018).

F Distribution

F Distribution is closely related to normal and chi-square distributions. In simple words, let X_1 and X_2 be two independent random variables with chi-square distributions. Define a new random variable based on the ratio $F = (X_1/df_1)/(X_2/df_2)$, where df_1 and df_2 are degrees of freedom of X_1 and X_2 , respectively. Then F follows an F distribution with df_1 and df_2 degrees of freedom. F distribution is commonly used in ANOVA analysis and linear regression modeling. The PDF of the F distribution is not directly used while the corresponding cumulative density function tables are readily available in statistics textbooks.

Factor Analysis and Principal Component Analysis

Factor analysis (Explanatory Factor Analysis, EFA) is a type of latent variable modeling, which is developed by using observed/measured variables (x_i). A linear relationship between the latent variable/factors are considered in the EFA where an unobserved error term is also present. Therefore, latent variables and their weights (their contributions in explaining the observed values in the measurement equations) are estimated in the factor analysis, in a way that $x_i = \alpha_0 + \beta f + \varepsilon_i$ in which β is the vector of coefficients/

weights (of size p) to be estimated and f is the vector of factors (of size s). The number and value of the factors is unknown. The error term ε_i follows a normal distribution with mean zero and variance σ_i . There is a possibility to consider relationships between the factors in latent variable modeling, which is not common in EFA.

In case the purpose of the analysis is to reduce the dimension of the data (x_i) and there is an assumption that there is no unobserved error term, the principal component analysis (PCA) is used. In PCA $x_i = \beta s$, score (s) are and weights β are estimated based on minimizing the sum of squared perpendicular distance to the component axis such that the correlation between the scores are minimized.

Examples of using EFA and PCA in transport modeling can be found on topics such as travel attributes (Rashidi and Mohammadian, 2011a, 2011b), emission, (Rashidi et al., 2015), and land use (Kockelman, 1997).

Hypothesis Testing

Model estimation and hypothesis testings are always discussed together, because parameters estimated using a sample are not necessarily close enough to the true parameters estimated based on the population. Therefore, a hypothesis is developed to examine whether the estimated parameters are sufficiently close to true parameters based on statistical measurements, knowing that all is available in the sample. Typically, the *null hypothesis* is formed around the hypothesis that the estimated parameter is equal (close) to the true parameter and the *alternative hypothesis* negates this hypothesis.

To test the null hypothesis, the sample is used to calculate a *test statistic*, which is associated with a probability which is then used to use the *confidence interval* or *test of significance*.

Assuming X as a random variable based on a sample, and parameter μ measuring some information about X , for example mean of X , μ becomes a random number with a specific PDF. Then for any confidence level (α), using the PDF of μ , a range for μ can be obtained such that $\mu_l \leq \mu \leq \mu_u$ (this is called acceptance region, with μ_l, μ_u being the critical values). This means for a null hypothesis, for example, $\mu = \tilde{\mu}$, if $[\mu_l, \mu_u]$ does not include $\tilde{\mu}$ then the null hypothesis can be rejected, otherwise, the null hypothesis cannot be rejected (note that it is not accepted, it is just not rejected).

Gamma Distribution

Gamma distribution is commonly used in the field of transportation for estimation of heterogeneity (Lord, 2006; Rashidi and Mohammadian, 2011b), trip distribution (McNally, 2000), trip generation (Wilmot and Stopher, 2001), and travel time variability (Susilawati et al., 2013).

Gamma distribution with three parameters for a random variable X can be written as:

$$\Pr(X = x, l, s, \gamma) = \frac{1}{s\Gamma(\gamma)} \left(\frac{x-l}{s} \right)^{\gamma-1} \exp\left(-\frac{x-l}{s}\right), l < x < \infty, s > 0, \gamma > 0,$$

where s is the scale parameter, l is the location parameter and γ is the shape parameter. In this notation, mean is equal to $s + \gamma + l$ and variance is $s^2 \times \gamma$, exponential distribution is a special case of gamma distribution where the scale parameter is one and location is zero.

Geometric and Negative Binomial Distributions

The negative binomial distribution is commonly used in accident analysis (Washington et al., 2010). This distribution is used to measure in a discrete probability of the number of successes in a sequence of iid Bernoulli trials (only two outcomes are viable, success and failure) before a specified and nonrandom number of successes occurs. If the number of success is 1, the distribution is called geometric design.

If the probability of success is p and the probability of failure is $q = 1 - p$, the probability of s successes and f failures for a random variable X :

$$\Pr(X = s) = \binom{f+s-1}{s} p^s q^f, \text{ where } \binom{f+s-1}{s} = \frac{f+s-1}{s!(f-1)!}.$$

The expected number of trials before observing s successes (or f failures) is

Heteroscedasticity

Heteroscedasticity (or heteroskedasticity) is used when the variance of a random variable (dependent variable) changes across another variable that is used to predict (explanatory variable or independent variable) it. Its opposite is called homoscedasticity, which can be written in mathematical format as: $\text{var}(Y|X) = \sigma^2$ which is fixed and not varying across different values of X s.

In linear regression, where ordinary least square is used, at presence of heteroscedasticity, the estimations are not anymore BLUE. BLUE estimations can be obtained when instead of OLS, the weighted least squares method is used, provided the fact that σ_i^2 are known. In this method, the dependent variable, the independent variables, and the residuals are weighted by the known σ_i .

To detect presence of heteroscedasticity, White's general heteroscedasticity test is the most well-known approach where the square of residuals are regressed against the explanatory variables, interactions of explanatory variables and/or higher powers of

them. When σ_i s are not known, White's heteroscedasticity-corrected standard errors are estimated where squared residuals are used in place of $\text{var}(Y|X) = \sigma^2$ in estimation of the variance of the coefficients.

Identifiability

A parameter of a model for a random variable, say Y is identifiable if for two different values of the parameter, different probability values are obtained.

In practice, for a given sample drawn for the random variable, a unique value can be estimated for the parameter. If this is the case, then the parameter is called identifiable. Identifiability can be empirical or theoretical. For example, [Bhat and Eluru \(2010\)](#) state that there are empirical identification issues for jointly estimating satiation parameters in the MDCEV formulation. In the random utility maximization formulations, it is quite well known that only the difference between utilities matters. This results in a theoretical identification issue for estimation of alternative specific parameters for all alternatives ([Train, 2009](#)). More details on identification issues of the logit kernel models is available in the paper by [Ben-Akiva et al. \(2001\)](#).

Kolmogorov–Smirnov Test

KS test is a nonparameter test to examine the similarity of two probability distributions, an empirical sample and a probability distribution, or two samples. For brevity, for a two-sided case, the null hypothesis is defined as the maximum distance between the two cumulative distributions. If this distance is greater than the critical value of the Smirnov table, $S_{m,n,1-\alpha}$, for parameters m, n , and $1-\alpha$ (where α is the significance level), then we can reject the null hypothesis of no difference between the two distributions.

Likelihood Ratio Test

The likelihood ratio test is commonly used in the field of transportation to compare the goodness-of-fit of two models developed on the same sample for different sets of parameters (one set being a subset of the other one, Θ_1, Θ_2). In this case, $LR = -2(LL(\Theta_1) - LL(\Theta_2))$, and LR follows a χ^2 distribution with degree of freedom $m = |\Theta_2| - |\Theta_1|$.

Lognormal Distribution

A random variable follows a lognormal distribution if its values are equal to the exponential of another random variable that is normally distributed.

Lognormal distribution is commonly used in the mixed logit formulation of discrete choice modeling where a specific sign is expected for a variable (such as price and travel time). The PDF of X as a random variable with lognormal distribution is:

$$f(X = x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(-\frac{(\log(x) - \mu)^2}{2\sigma^2}\right),$$

Where μ and σ are mean and standard deviation of $\log(x)$.

Markov Chain Monte Carlo

Markov chains are commonly used to predict the next stage of a variable/system/agent conditional on its current state. Therefore, transitions are random and conditional on the distribution of the current state of agent/system/variable.

Borrowing concepts from the Markov chain approach, the Markov chain Monte Carlo (MCMC) approach is typically used to estimate a hard-to-estimate expectations like $\sum_{m=0}^M x(\tau^{m_0+m}) f\left(\frac{\tau^{m_0+m}}{M}\right) \cong \int x(t) f(t) dt$, where f is the PDF of x , but challenging to draw from. In such cases, the MCMC approach is used to generate a sequence of random draws (a Markov chain) to approximate f and eventually $\bar{x}$. If the sequence of these random draws for a large M is defined as $\{\tau^m, m \geq 0\}$, and the initial m_0 draws are discarded for burn-in effect, the set $\{\tau^{m_0}, \dots, \tau^{m_0+M}\}$ can be used to approximate such that:

$$\sum_{m=0}^M x(\tau^{m_0+m}) f\left(\frac{\tau^{m_0+m}}{M}\right) \cong \int x(t) f(t) dt.$$

When $f(t)$ is available from which random draws can be easily obtained, it can be used to approximate $\bar{x}$ using the Metropolis–Hasting sampling, or Gibbs sampling which are popular MCMC approaches. Gibbs sampling is a special case of MH. Gibbs sampling is used when it is possible to directly sample from the conditional distribution. MH is used when the conditional distribution does not belong to recognizable families of distributions or when direct sampling from the conditional is difficult. Metropolis–Hastings sampling is useful for Bayesian procedures, but is affected by challenges such as inefficiency, tuning and serial correlation. In the context of MH algorithms, f is called the target density and g is called the proposal. The MH algorithm is formed in two steps. In the first step, a proposal is generated. In the second step, the proposal is accepted or rejected. If the proposal is Gaussian, the Random–Walk Metropolis algorithm is obtained. In this case, the proposal is $\hat{\tau}$ is sampled from $N(\bar{\tau}, \Sigma)$, where τ denotes the current value and Σ is a scale matrix. Then, the acceptance probability is calculated as $A = \min(1, f(\hat{\tau})/f(\bar{\tau}))$.

Mixture Models

To estimate heterogeneity in parameter estimation, it is common to use a combination of components of a distribution. For example, in discrete choice modeling parametric mixed logit models, and non-parametric/semi-parametric latent class logit models are widely used in disciplines studying individual choice behavior.

If decision makers' decisions are consistent with the random utility maximization assumption, a mixed random utility model such as the mixed multinomial logit/probit model can capture random heterogeneity distribution by marginalizing the discrete choice kernel over some mixing distribution which captures the unobserved tastes variation of the decision makers (Train, 2009). However, such an assumption might be requiring strict assumptions about the taste variation distribution.

Despite the aforementioned limitations, all mixture models follow the following formulation:

$$f(X, \beta) = \int f(X, \psi) d\beta(\psi).$$

Distribution β is called the mixing distribution, where mixture models of this structure can be parameterized over the parameter space of $\mathbf{B}$. When the mixing distribution has discrete components as p_i , $\sum_{i=1}^M p_i = 1$, the probability can be written as:

$$f(X, \beta) = \sum_{i=1}^M p_i f(X, \beta_i),$$

which is the same structure as latent class models with M classes.

Moving Averages

Time series can be transformed by taking averages of several sequential values. In a general two-sided averaging scheme, the original time series variable e_t can be transformed to μ_t with order k as:

$$\mu_t = \frac{1}{2k+1} \sum_{i=-q}^q e_{t+i}.$$

When μ_t is defined for one side, that is, dependent on the past, and it is weighted by parameters, say θ_q , a MA model is constructed as:

$$\mu_t = \frac{1}{2k+1} \sum_{i=-q}^q \theta_i e_{t+i},$$

where $\theta_0=0$ and $\{e_t\}$ is white noise series. In this case, μ_t is observed and θ_i is estimated. The main difference between MA and the autoregressive model is that several instances of white noise appear on the right hand side while past instances of the dependent variable do not appear on the right hand side of the equation.

Multicollinearity

For crosssectional data, and more specifically for linear regression, multicollinearity violates some basic assumptions of OLS. When perfect collinearity exists, the model coefficients cannot be estimated, unidentifiability. When partial collinearity exists, typically, acceptable goodness-of-fit is obtained but there are not many statistically significant coefficients in the model. This might be due to pairwise correlation between the explanatory variables, which then results in wider confidence intervals with smaller t-values. Since multicollinearity is a sample data issue, if the purpose of model development is prediction only, then the small goodness-of-fit should not be an issue, and the model can be used with no changes. However, if possible, omitting highly colinear variables or combing and transforming colinear variables might be a useful technique to reduce the impact of multicollinearity.

Normal Distribution, Univariate, or Multivariate

Also called Gaussian distribution, normal distribution is the most commonly used distribution in statistics. With PDF as $f(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(X-\bar{X})^2}{2\sigma^2}\right\}$, it is noted that $X \sim N(\bar{X}, \sigma)$. This distribution is symmetric and unimodal with mean and median being equal.

When multiple random variables that are individually normally distributed are correlated, a multivariate normal distribution can be used. If X and Y are normally distributed with variances σ_X and σ_Y with covariance σ_{XY} then the joint PDF can be written as:

$$f(X, Y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[\left(\frac{X-\bar{X}}{\sigma_X}\right)^2 \left(\frac{Y-\bar{Y}}{\sigma_Y}\right)^2 - 2\frac{X-\bar{X}}{\sigma_X} \frac{Y-\bar{Y}}{\sigma_Y} \right] \right\},$$

$$\text{where } \rho = \frac{\sigma_{XY}}{(\sigma_X\sigma_Y)}$$

If more than two random variables are considered as $X=[X_1, \dots, X_n]$ $\rho = \frac{\sigma_{XY}}{(\sigma_X\sigma_Y)}$ as then the PDF of X can be written as:

$$f(X) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Omega}|}} \exp\left\{-\frac{1}{2}(X-\bar{X})^T \mathbf{\Omega} (X-\bar{X})\right\},$$

where $\mathbf{\Omega}$ is the covariance matrix being equal to the expected value of $(X-\bar{X})(X-\bar{X})^T$. Then it can be said that $X \sim N(\bar{X}, \mathbf{\Omega})$.

Poisson Distribution and its Application in Transportation

Count data consist of nonnegative integer values. In the area of transportation there are several examples of such data such as: number of weekly/daily trips per person or household, number of driver route changes per week, number of trip departure changes per week, drivers' frequency-of-use of technologies for traveling per week, number of accidents observed on road segments per year. The most common count data models are Poisson and negative binomial regression models. Consider the number of accidents occurring per year at various intersections in a city. The probability of location i having Y_i accidents per year can be written as $P(Y_i) = \frac{e^{-\lambda_i} \lambda_i^{Y_i}}{Y_i!}$, where λ_i is the parameter of location i which can be parameterized as $\lambda_i = e^{\beta X_i}$ where X_i is a vector of attributes of location i . λ_i provides the expected number of accidents at i .

If parameterization of λ_i includes an unobserved random term, other than the attributes of location i , then λ_i can be re-written as $\lambda_i = e^{\beta X_i + \varepsilon_i}$. Then by assuming a gamma distribution for ε_i with mean 1 and variance σ , then the probability of Y_i can be rewritten as:

$$P(Y_i) = \frac{\Gamma\left[\frac{1}{\alpha} + Y_i\right]}{\Gamma\left(\frac{1}{\alpha}\right) Y_i!} \left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + \lambda_i}\right)^{\frac{1}{\alpha}} \left(\frac{\lambda_i}{\frac{1}{\alpha} + \lambda_i}\right)^{Y_i},$$

where Γ is the gamma function and α is the overdispersion parameter, and $\text{Var}[Y_i] = E[Y_i] + \alpha E[Y_i]^2$. This PDF with gamma distributed unobserved term is called negative-binomial function, which is commonly used in modeling accidents (Washington et al.,2010).

P Values

The p-value (the most commonly used value is 5%) is lowest level for rejecting the null hypothesis, while it is true. In other words, for a 5% P value, the chance of committing a Type I error (the probability of rejecting the true hypothesis) is 5 in 100. There are some confusions about P value such as:

- P value (or $1 - p$) measures the probability of the null hypothesis (alternative hypothesis) being true (or false);
- P value is to assess the importance, significance, generalizability or replicability of a result;
- P value is the rate of the Type I error.

Sample Size Determination

Sample size is typically determined in the field of transportation based on the required variance of a random variable (a dependent variable). Then the sample size can depend on three factors: (1) variability of the parameter in the population, (2) degree of accuracy required for the variable, and (3) population size. By considering a normal distribution for the mean of the variable (based on the central limit theorem) for a 5% acceptance level, we can write $-1.96 \times s < \bar{x} - \mu \leq 1.96 \times s \Rightarrow \frac{s}{\sqrt{n}} - 1.96 \times \frac{s}{\sqrt{n}} < \bar{x} - \mu \leq 1.96 \times \frac{s}{\sqrt{n}}$. Then it can be formatted as $d = |\bar{x} - \mu| \leq 1.96 \times \frac{s}{\sqrt{n}}$. Based on the highest acceptable value of d , let's call it w , the sample size can be estimated as $n = 15.36 \times \sigma^2 / w^2$. However, it is common that no information is available for w . Instead the relative ratio of w to the true mean is known $w' = w / \mu \times 100 \Rightarrow w = \mu \times w'$. Then the sample size can be estimated based on w' as: $n = 15.36 \times \sigma^2 / (\mu \times w')^2 = \frac{15.36}{w'^2} \times \left(\frac{\sigma}{\mu}\right)^2 = \frac{15.36}{w'^2} \times CV^2$ where CV is called the coefficient of variation (Ortuzar and Willumsen). If the population size is not large enough, then it should be adjusted as $n / (1 + \frac{n}{N})$.

Sampling Algorithms

Samples are used to represent the population. Therefore, it is essential to assure that the attributes of the population are reflected when using the sample. If we can afford having a large sample, a pure random sample can easily represent the population. However, there are several issues that may arise if a pure random sample is considered. First, there might be minorities from whom finding a representative sample would increase the overall sample size. Second, sampling is a costly process. For example, the cost of having household travel surveys is over \$200 for each completed sample. Therefore, it is crucial to maintain the sample size at an efficient level.

Stratified sampling is a commonly practiced approach to reduce the sample size while having representatives from minorities within the population. In stratified sampling random samples are obtained with specific sizes from each stratum. Then the overall sample is formed by pooling the strata.

Student's t-Tests

This test is a parametric test based on the t-student distribution. It is generally symmetric and flatter than the normal distribution.

For two random variables of $X_1 \sim N(0, 1)$ and $X_2 \sim \chi_k^2$ that are independent we can construct another random variable as $t_k = X_1 / \sqrt{\frac{X_2}{k}}$, which is called to follow Student's t distribution with k degrees of freedom. As k increases, the t distribution approximates the normal distribution.

This distribution is commonly used in regression models to test the null hypothesis of the estimated variables being any different than zero. In such cases, because the expected value of the estimated parameter is of interest, it is assumed that $\bar{\beta}$ is normally distributed and its standard deviation is unknown to be estimated using the sample. Then a statistic is formed as $\beta = \bar{\beta}/(S/\sqrt{n})$. Here S follows a χ^2 distribution and $\bar{\beta}$ is normally distributed. For large samples ($n \geq 30$) this statistic collapses to a normally distributed random variable.

Another application of t-student test is to compare the mean of two independent samples (with sample sizes of n_1 and n_2) by constructing the following statistic:

$$t_k = (\bar{X}_1 - \bar{X}_2) / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

which approximately follows a t-distribution with k being equal to the following:

$$\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left(\left(\frac{s_1^2}{n_1} \right)^2 / (n_1 - 1) + \left(\frac{s_2^2}{n_2} \right)^2 / (n_2 - 1) \right).$$

Uniform Distribution

One of the most basic distributions has a fixed probability in an interval and zero otherwise. The general PDF has the form of: $f(x) = \{ 1/(U - L) \text{ if } L \leq x \leq U \}$ otherwise for values of the random variable in the $[L, U]$ interval the cumulative density function is $(x - L)/(U - L)$, and the expected value of this distribution is $(U + L)/2$.

Weibull Distribution

Weibull distribution is the most commonly used function in hazard-based modeling. The PDF of the Weibull distribution for positive values, with shape parameter γ can be written as $f(x, \gamma) = \gamma x^{\gamma-1} e^{-x^\gamma}$. Then the cumulative density, survival and hazard functions are written as $1 - e^{-x^\gamma}$, e^{-x^γ} , and $\gamma x^{\gamma-1}$.

Databases

Crosssectional Data

In one categorization, the most commonly used data for transportation modeling is called crosssectional data which is collected for a specific *crosssection* of time. Therefore, the temporal variation in such data is negligible. Household travel surveys are the most commonly used data sources of this kind which may be collected during a year or on a specific day of the year which is assumed to be a typical day (Meyer and Miller, 1984). Data on network (highway and transit) performance can also be collected for a section of time (Meyer and Miller, 1984) while it can be recorded for multiple time slots, which can be then used for time series modeling (Ma et al., 2015).

Panel Data

Panel data, also called longitudinal data, are datasets, where for multiple times, the same respondent/individual/subject is interviewed/surveyed. In transportation applications, respondents might be households, firms, individual persons, locations, network links or network nodes. The main advantage of panel data over crosssection data is that they can be used to study dynamics and variation of tastes of individuals over time. In a typical regression model $Y_{it} = \alpha + \beta X_{it} + \epsilon_{it}$ where t refers to the time periods and i refers to the individuals/respondents/etc. The random term in this case may not be simply iid, meaning that it can be split into two parts as $\epsilon_{it} = \epsilon_i + u_{it}$, where u_{it} is the random distance and is the time invariant unobserved effect of individual i . If it is assumed that $\epsilon_i \sim IID(0, \sigma_\epsilon^2)$ and $u_{it} \sim IID(0, \sigma_u^2)$, then the correlation between ϵ_{it} and $\epsilon_{it'}$ for the same individual is equal to 1 if t and t' are the same and $\sigma_\epsilon^2 / (\sigma_\epsilon^2 + \sigma_u^2)$, otherwise. Similarly, if another error term is considered in ϵ_{it} as $\zeta_t \sim IID(0, \sigma_\zeta^2)$ then for the same individual, the correlation between ϵ_{it} and $\epsilon_{it'}$ is 1 if t and t' are the same and $\sigma_\epsilon^2 / (\sigma_\epsilon^2 + \sigma_u^2 + \sigma_\zeta^2)$, otherwise. However, if t and t' are the same but we are interested in attributes of two individuals i and j then the correlation is between the error terms ϵ_{it} and $\epsilon_{jt'}$ and would be $\sigma_\zeta^2 / (\sigma_\epsilon^2 + \sigma_u^2 + \sigma_\zeta^2)$. It is straightforward to show that such random effect structure implies homoscedasticity with an additive variance format for the three/two random terms. Therefore, the correlation between individual and period effects can be captured using these correlation (variance, covariance) values. If not properly controlled for, the model estimations in presence of panel data can be biased.

Revealed Preference

In one categorization, the most commonly analyzed data in the field of transportation is called revealed preference data, which are collected from respondents about their choices in observed situations. In the area of travel demand modeling, preference of people with regard to the mode of transport, traveling route, destination, and time of day (or departure time) are commonly modeled using information extracted from revealed preference of travelers.

An important challenge before using revealed preference data is to complement the observed preferences with attributes of nonchosen alternates. For example, in the case of mode choice, people reveal that they select a specific mode of transportation, say car. However, to be able to develop a choice model, other competing alternatives and their attributes should be also available.

Self-reported alternatives, especially nonchosen alternatives' attributes are associated with perceptions of the respondent, therefore, it is common to obtain the attributes of all alternatives including the chosen one, from external sources such as the transportation model or Google API.

There are two main issues limiting the use of revealed preference data. First, revealed preference data reflect the real world as such the variation of variables are limited to context. For example, transit fare is typically set by authorities and does not annually change or possibly minimally changes. Therefore, when crosssectional data are collected limited variation on cost of transportation is observed depending what exists in the transportation system. In other words, for a specific trip cost of transportation does not vary for specific mode of transportation, or possibly changing with limited variance. Second, hypothetical scenarios cannot be studied using revealed preference data. For example, what would be the preference of travelers when a new highway is built in the city? Or how people react to emerging mobility options, such as driverless vehicles, when they are ubiquitous? These two issues can be controlled for by collecting stated preference data.

Stated Preference Data

To study hypothetical contexts or examine larger variance of alternative attributes which might be too costly to be examined in the real world, stated preference surveys are designed. Some of the alternatives may be hypothetical while one (or even more) of them be an existing one. Therefore, the main purpose for designing stated preference surveys is to observe how people trade off when attributes of alternative change to large extent compared to what exists out in the transportation network, or to explore their behavior when presented with policies which are entirely new.

There are three main types of stated preference surveys: contingent valuation, conjoint analysis, and stated choice. Contingent valuation deals with elicitation of willingness-to-pay in which respondents are typically asked to directly report their willingness-to-pay for new options or policies (Fagnant and Kockelman, 2015). In the conjoint analysis respondents are presented with several options/policies and are asked to rank them. The best-worst analysis can be considered as a special case of conjoint analysis. In the most commonly used type of stated preference surveys, respondents are presented with several alternatives among which they select the preferred option (it is also possible to select more than one in special cases depending on the context of the survey). The choice for selecting the preferred options is assumed to be made based on the utility that is expected to be gained through the attributes of the preferred option. The stated choice data are commonly used to develop models based on random utility maximization (Train, 2009). Full factorial, fractional factorial, efficient, and optimal designs are some of the popular methods to design stated preference surveys (Hensher et al., 2005).

Survival Data

Survival data measure time from an origin, which can be a reference point in time beyond which data are censored (left censorship) or from the birth of the respondent, until a particular event occurs or the end of the study (right censorship). The event can be related to a decision changing the lifestyle such as changing job or residential location (Ghasri et al., 2018), vehicle transaction decisions (Rashidi et al., 2011), generating an activity (Auld et al., 2011), or a change in the lifestyle such birth of a child (Jenkins, 2005).

Survival analysis or hazard-based models are formed around four primary functions: (1) *Failure function* is defined as the probability of failure before a specific time, (2) *Survival function* is defined as the probability of not failing before a specific time, (3) *Failure PDF* is defined as the probability of failure in a minuscule time interval, which is the slope of failure function (4) *Hazard rate*, which is defined as the conditional probability of failure at time t , given the fact that the failure has not occurred before this time.

$$f_0(t) = \lim_{\Delta t \rightarrow 0} \Pr \frac{(t \leq T \leq t + \Delta t)}{\Delta} t = \frac{\partial F_0(t)}{\partial t} = -\frac{\partial S_0(t)}{\partial t},$$

$$h_0(t) = \lim_{\Delta t \rightarrow 0} \Pr \frac{(T + \Delta t > t \geq T | T \leq t)}{\Delta} t = \frac{f_0(t)}{S_0(t)} = \frac{f_0(t)}{1 - F_0(t)},$$

where F_0 denotes the failure function, S_0 shows the survival function, f_0 represents the failure PDF, and h_0 denotes the hazard rate.

A nonparametric function of survival and hazard rate can be estimated for any dataset using Kaplan-Meier estimators (Jenkins, 2005). Nonparametric formulations are quite widely used in practice given their flexibility.

There are two defined methods to incorporate the impact of exogenous variables called proportional hazard (PH), and accelerated failure time (AFT) (Jenkins, 2005). In AFT, covariates are incorporated in the definition of failure PDF. In the PH models, the baseline can be separated from the covariate part and the hazard function follows the same distribution as the baseline hazard function. The PDF and hazard function of AFT specifications are provided follows:

$$f(t, x_i) = f_0(t) \exp \left(- \sum_{j=1}^J \beta_j x_{ij} \right),$$

$$h(t, x_i) = h_0 \left(t \exp \left(- \sum_{j=1}^J \beta_j x_{ij} \right) \right) \exp \left(- \sum_{j=1}^J \beta_j x_{ij} \right).$$

References

- Anderson, T.W., Darling, D.A., 1954. A test of goodness of fit. *JASA* 49 (268), 765–769.
- Auld, J., Rashidi, T.H., Javanmardi, M., Mohammadian, A., 2011. Dynamic activity generation model using competing hazard formulation. *Transp Res Rec* 2254 (1), 28–35.
- Bayarri, M.J., Berger, J.O., 1999. Quantifying surprise in the data and model verification. *Bayesian Stat* 6, 53–82.
- Ben-Akiva, M., Bolduc, D., Walker, J., 2001. Specification, identification and estimation of the logit kernel (or continuous mixed logit) model.
- Bhat, C.R., Eluru, N., 2009. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Trans. Res. B Meth.* 43 (7), 749–765.
- Bhat, C.R., Eluru, N., 2010. The multiple discrete-continuous extreme value (MDCEV) model: formulation and applications. In: *Choice Modelling: The State-of-the-Art and The State-of-Practice: Proceedings from the Inaugural International Choice Modelling Conference*, Emerald Group Publishing Limited, pp. 71–99.
- Bhat, C.R., Guo, J., 2004. A mixed spatially correlated logit model: formulation and application to residential choice modeling. *Trans. Res. B Meth.* 38 (2), 147–168.
- Bolduc, D., 1992. Generalized autoregressive errors in the multinomial probit model. *Transp. Res. Part B Methodol* 26 (2), 155–170.
- Bolduc, D., Dagenais, M.G., Gaudry, M.J., 1989. Spatially autocorrelated errors in origin-destination models: a new specification applied to aggregate mode choice. *Trans. Res. B Meth.* 23 (5), 361–372.
- Clayton, D.G., 1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 65 (1), 141–151.
- Embrechts, P., McNeil, A., Straumann, D., 2002. Correlation and Dependence in Risk Management: Properties and Pitfalls. *Risk Management: Value at Risk and Beyond*, pp. 176–223.
- Fagnant, D.J., Kockelman, K., 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Trans. Res. A Policy Pract.* 77, 167–181.
- Frees, E.W., Valdez, E.A., 1998. Understanding relationships using copulas. *N. Am. Actuar. J.* 2 (1), 1–25.
- Ghasri, M., Rashidi, T.H., Saberi, M., 2018. Comparing survival analysis and discrete choice specifications simulating dynamics of vehicle ownership. *Trans. Res. Rec.* 2672 (49), 34–45.
- Gujarati, D.N., 2009. *Basic Econometrics*. Tata McGraw-Hill Education.
- Hensher, D.A., Rose, J.M., Greene, W.H., 2005. *Applied Choice Analysis: A Primer*. Cambridge University Press.
- Hoeffding, W., 1940. Scale-invariant correlation theory. In: Fisher, N.I., Sen, P.K. (Eds.), *The Collected Works of Wassily-Hoeffding*, Springer-Verlag, New York, pp. 57–107.
- Hossein Rashidi, T., Mohammadian, K., 2016. Application of a nested trivariate copula structure in a competing duration hazard-based vehicle transaction decision model. *Transportmetrica A Trans. Sci.* 12 (6), 550–567.
- Jenkins, S.P., 2005. *Survival Analysis*. Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK.
- Joe, H., Li, H., Nikoloulopoulos, A.K., 2010. Tail dependence functions and vine copulas. *J. Multivar. Anal.* 101 (1), 252–270.
- Kockelman, K.M., 1997. Travel behavior as function of accessibility, land use mixing, and land use balance: evidence from San Francisco Bay Area. *Trans. Res. Rec.* 1607 (1), 116–125.
- Krueger, R., Rashidi, T.H., Vij, A., 2019. Semi-Parametric Hierarchical Bayes Estimates of New Yorkers' Willingness to Pay for Features of Shared Automated Vehicle Services. *arXiv preprint arXiv:1907.09639*.
- Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accid. Anal. Prev.* 38 (4), 751–766.
- Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y., 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Trans. Res. C Emer. Technol.* 54, 187–197.
- McNally, M.G. The four step model. 2000.
- Meyer, M.D., Miller, E.J., 1984. *Urban Transportation Planning: A decision-oriented Approach*.
- Rashidi, T.H., Kanaroglou, P., Toop, E., Maoh, H., Liu, X., 2015. Emissions and built form – an analysis of six Canadian cities. *Trans. Lett.* 7 (2), 80–91.
- Rashidi, T.H., Mohammadian, A., Koppelman, F.S., 2011. Modeling interdependencies between vehicle transaction, residential relocation and job change. *Transportation* 38 (6), 909.
- Rashidi, T., Mohammadian, A., 2011a. A competing hazard model of household vehicle transaction behavior with discrete time intervals and unobserved heterogeneity. *Trans. Lett.* 3 (3), 219–229.
- Rashidi, T.H., Mohammadian, A., 2011b. Household travel attributes transferability analysis: application of a hierarchical rule based approach. *Transportation* 38, 697, <https://doi.org/10.1007/s11116-011-9339-8>.
- Sklar, A. (1959), Fonctions de répartition à dimensions et leurs marges, *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.
- Smith, M.D., 2003. Modelling sample selection using Archimedean copulas. *Econ. J.* 6 (1), 99–123.
- Susilawati, S., Taylor, M.A., Somenahalli, S.V., 2013. Distributions of travel time variability on urban roads. *J. Adv. Trans.* 47 (8), 720–736.
- Swait, J., Adamowicz, W., 2001. The influence of task complexity on consumer choice: a latent class model of decision strategy switching. *J. Cons. Res.* 28 (1), 135–148.
- Train, K.E., 2009. *Discrete Choice Methods with Simulation*. Cambridge University Press.
- Trivedi, P.K., Zimmer, D.M., 2007. *Copula Modeling: An Introduction for Practitioners*. Now Publishers Inc.
- Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC.
- Wilmot, C.G., Stopher, P.R., 2001. Transferability of transportation planning data. *Transportation Research Record* 1768 (1), 36–43.

Travel Surveys

Stacey G. Bricka, Independent Consultant, Bastrop, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	587
Household Surveys	587
Establishment Surveys	588
Transit On-Board Surveys	588
The Future of Travel Surveys	588
Biography	589
References	589

Introduction

Travel surveys used to support transportation planning and policy efforts. In a very basic sense, these surveys are used to understand the current and future demand for transportation. Data from the surveys feed directly into long-range transportation plans, travel demand models, and policy analyses, among other uses. While the methods and technologies used to conduct these surveys have evolved over time, the core data obtained remain largely unchanged.

The main types of travel surveys include household, establishment (or workplace), commercial vehicle, on-board transit, visitor, and special event surveys. Each survey documents the origin and destination of travel for one slice of the traveling public along with demographic, geographic, and attitudinal characteristics of the travelers. Household and on-board surveys are the most common. The following provides more details about each survey type.

Household Surveys

Household survey data form the core input into most travel demand models. These surveys document the origin, destination, purpose, mode, and time of daily travel for residents of a region. Over time, these surveys have evolved from capturing only the work trip, only auto trip, only travel by those ages 16+ to comprehensive inventories of travel by all members of participating households, using all modes and for all purposes (Weiner, 2008).

Household surveys typically document details about the household and its members (demographics), inventory the household vehicle fleet, sometimes capture attitudinal data on topics important to the sponsoring agency, then capture the specific travel of all household members for a specific travel period, which can range from one day to a week or longer. Most household surveys focus on documenting weekday travel during the spring and fall, when schools are in session. Some surveys, such as the 2017 National Household Travel Survey (FHWA, 2017) obtain travel details for 365 days (including weekends, summer, and holidays).

As the methods have evolved, so have the sampling frames. Telephone-based samples (such as random-digit-dial or RDD frames, then listed and unlisted telephone frames) migrated to address-based sampling (ABS) frames based on the US Postal System address delivery sequencing files. The ABS sampling frame is often supplemented now by frames of cellular phone numbers, consumer data (including name, email and address as well as other demographic details). Often the sampling approach calls for over-sampling specific geographies or behaviors to ensure a robust data set to support modeling. The surveys focus on populations not in group quarters, so supplemental surveys are necessary to understand travel behavior of military personnel living on base, university students living in dormitories, and the elderly in assisted living communities. Research is underway to evaluate how consumer data can be used to refine sampling plans, including using vendor-generate demographic characteristics about the households to target specific types of households (such as low-income household) for oversampling.

The current survey process follows a two-stage design: the recruitment survey obtains basic household, personal, and attitudinal data, and the retrieval survey focuses on capturing the travel behavior details. Over time, the design has varied from single stage to two-stage surveys. Earlier studies were conducted in person in one interview (thus a single-stage design). As the use of the telephone to conduct surveys became more commonplace, the single-stage survey was thought to under-collect travel data so the industry moved to the current two-stage survey. As response rates have continued to decline, professionals are testing the return to a one-stage survey with controls to mitigate the concerns of trip underreporting.

In the past, households were provided detailed travel diaries or logs to capture the details of their travel. While travel logs are still used in some surveys today, more emphasis is placed on the use of smartphones to passively capture travel details. The surveys offer phone, mail, and web options to complete, with more advanced surveys offering respondents the option to track their travel using a smartphone application (app). Smartphone apps use the built-in global positioning system (GPS) technology to capture the movement of the phone and algorithms are used to decode those traces into trips, which are then validated by the traveler. Most

travel log-based surveys collect data for a 24-hour period. The smartphone-based surveys collect data for 3–7-day periods, with supplemental surveys of one month or longer to capture the rarer long-distance travel that occurs infrequently.

As with most surveys in the United States, household surveys have experienced declining response rates, nonresponse bias (missing critical population segments), and increasing costs. Agencies traditionally have conducted the household surveys once a decade (funding permitting), spending sometimes up to \$1 million for a sample of less than 1% of the population. The cost of these large-scale surveys includes both the financial cost as well as the staff time that must be dedicated to getting up to speed on current methods and then managing the effort for about 1 year.

A more recent trend adopted by some agencies is to conduct smaller surveys more frequently (annually or biannually). For example, the Ohio Department of Transportation's current survey of 10,000 households is comprised of 10 annual surveys of 1,000 households each (Anderson et al., 2015). The Federal Highway Administration is undertaking a similar conversion of its long-standing National Household Travel Survey. Formerly conducted every 5–8 years since 1969 with national samples of approximately 25,000 households, plans are to conduct the NextGen NHTS every other year starting in 2002 with a national sample of approximately 7,500 households.

Establishment Surveys

The counterpart to household surveys is establishment or workplace surveys. These surveys capture details about the firm (number of employees, hours of operation, industry classification, location) and about the travel to and from that establishment by both the employees and customers. As compared to the household surveys, establishment surveys are rarely conducted in the United States. The two most recent establishment surveys were conducted in New York City and Phoenix, both in the 2014–16 time frame.

The establishment survey process includes recruiting the firm and assigning it to a specific data collection period (typically weekday operating hours during fall and spring when school is in session). On the actual survey day, physical counts are taken to document the “universe” of site visits, which are used to expand the survey data.

The surveys are conducted in person using tablet-based survey programs. The surveys are short, focusing on only a few key demographic characteristics and the trip to and from that location. Respondents are asked to locate their origins and destinations using real-time mapping features in the survey software. Employees may be offered the option to complete their survey online.

Sampling plans for establishment surveys are drawn from business-frames that are commercially available. The databases list each business in a region, its address, sometimes a contact person, along with the industry classification and firm size. For the recent Phoenix survey, the stratification included industry classification, size, and geography (FHWA TMIP webinar, 2018).

There are two additional surveys often conducted as part of establishment surveys: special generator surveys and commercial vehicle surveys.

- Special generator surveys are establishment surveys conducted at locations that tend to be high volume and specialized in terms of their impact on regional transportation. Often, these surveys focus on hospitals, universities, shopping malls, and sporting venues. The same type of data are collected in terms of person counts and in-person intercept surveys. Some agencies have successfully used passive data products to capture travel associated with specific sporting events. The passive data provide comprehensive origin-destination details and an imputed trip origin (home) location but tends to lack any demographic details about the traveler. This issue is discussed more in the next session.
- Commercial vehicle survey—these surveys collect the detailed daily travel made in commercially owned vehicles, including a range of vehicle fleet from automobiles to vans and light-, medium-, and heavy-duty trucks. As with household surveys, the detailed paper logs are being replaced with smartphone apps. Unlike household surveys, some commercial vehicle surveys consist of obtaining administrative data from a company representative and fusing that to GPS-data already collected in the company-owned vehicles.

Transit On-Board Surveys

The most common mode-specific survey is the transit on-board survey. This survey focuses on the rider-reported usage of the transit system, including boarding and alighting locations as well as ultimate origins and destinations of travel for a specific transit trip. Supplemental demographic data are collected and sometimes a few attitudinal questions are also asked, space permitting. Because the survey is conducted on-board the transit vehicle or as intercepts while riders are waiting for the vehicle to arrive, the surveys are short and focused.

The methods for this type of survey are transitioning from in-person paper-based administration to in-person tablet administered surveys. In addition, agencies and researchers are mining fare card data and turnstile transaction data (where available) to obtain the overall O-D flow patterns (Zalewski et al., 2019).

The Future of Travel Surveys

With the availability of passive data products that summarize origin-destination flows in and about a specific region for a much lower cost than a traditional survey, the role of travel surveys is changing. In this section, the future of travel surveys is explored by

first reviewing the challenges, then the opportunities. Three examples of how passive data products have changed the travel survey landscape include the end of cordon line surveys, the quest for long-distance travel data, and the application of passive data to capture visitor and special event markets.

In the past, cordon line surveys (also called “external station surveys”) were conducted to capture details regarding trips entering a region, either to a destination within the region or to pass through the region en-route to another destination. This type of trip is important, as it influences traffic volume counts as well as congestion levels but is not captured in the household, establishment or transit surveys. Cordon line surveys were conducted roadside, by intercepting drivers at or near the regional boundaries. Drivers were asked a short series of questions to capture origin and destination of travel, trip purpose, vehicle occupancy, and, if commercial, cargo.

As traffic volumes increased, safety concerns led to the end of this practice and transportation planners turned to technology. First, roadside interviews were replaced with “license plate capture”—cameras that recorded vehicle license plates, which were then transcribed and matched against state vehicle registration records. Recorded plates were compared across capture points to determine if the vehicle exited the area and if so, time stamps were used to determine duration of time spent in the region. The cost of transcribing the license plate data led researchers to install portable Bluetooth readers along major routes to capture unique device identification numbers. Those details were matched across collection locations (similar to the license plate capture analysis) to determine if trips were to a local destination or through travel.

As analysts mine the passive data products to identify through versus local travel, the research has expanded into using the data to better understand travel patterns among low incidence travel groups, most notably long-distance travelers and special event travel. Using imputed home location of each device, researchers can identify when a device is used outside the home region and for how long. This data can be used to understand patterns of long-distance travel. In addition, by focusing on specific locations on specified dates and times, researchers can evaluate travel surrounding special events. These special events can be sporting events, community events, and even disasters.

The more these passive data products are used for known applications, the increased likelihood that more applications will be identified. The passive data products “narrow but deep,” meaning that there are only a few variables in these data products but a large volume of cases. Travel surveys, on the other hand, are “wide but shallow,” meaning that the data collected is significant (often hundreds of variables in the data sets) but the sample sizes are small (ranging from 1,500–25,000 households). Research is underway to evaluate the best methods for fusing these products and to identify alternative uses for both data sources. While some believe the passive data products, most working in this specialty see the passive data products as explaining the “why”—the traffic volumes, the origin-destination pairs, etc. However, for the “why” and the underlying demographics that explain travel behavior, the travel surveys remain the only option and thus are expected to continue to be administered.

Biography

Dr. Stacey Bricka has 30 years of experience in travel behavior research and analysis, travel surveys, and associated survey methods and emerging technologies and data sources. She is known for her extensive and innovative contributions to the state of the art in travel behavior methods, technologies, and analyses. She has managed or directed more than 100 surveys pertaining to travel behavior and transportation in general and moderated numerous conference sessions aimed at advancing the integration of passive data into travel behavior data programs for transportation agencies. She routinely advises agencies on various methodological and technological aspects of travel surveys and leveraging complementary passive data products. Dr. Bricka currently supports FHWA’s NHTS program, including the NextGen NHTS design activities. Her research priorities currently focus on integrating traditional travel survey data with emerging passive data sources. She holds a PhD in Community and Regional Planning from the University of Texas at Austin.

References

- Anderson, R., Giaimo, G., Greene, E., Geilich, M., Hathaway, K., Flake, L. (2015). First Large-Scale Smartphone-Based Household Travel Survey in USA—Pilot Results. *Travel Forecasting Resource* wiki entry. Available from: http://www.tfresource.org/images/bb/ITM16_First_Large-Scale_Smartphone-Based_Household_Travel_Survey_in_USA_%E2%80%9393_Pilot_Results.pdf.
- Department of Transportation and Federal Highway Administration, U.S. Department of Transportation, Federal Highway Administration, 2017. National Household Travel Survey. Available from: <https://nhts.ornl.gov>.
- U.S. Department of Transportation, Federal Highway Administration, Travel Model Improvement Program, 2018. MAG Large-Scale Regional Travel Surveys and Bottleneck Survey. Available from: <https://tmip.org/content/mag-large-scale-regional-travel-surveys-and-bottleneck-survey-october-18-2017>.
- Weiner, E., 2008. *Urban Transportation Planning in the United States*. Springer.
- Zalewski, A., Sonenklar, D., Cohen, A., Kressner, J., Macfarlane, G., 2019. Public Transit Rider Origin-Destination Survey Methods and Technologies: A Synthesis of Transit Practice. Transit Cooperative Research Program Synthesis 138.

Travel Demand Forecasting: Where are we and What are the Emerging Issues

Thomas F. Rossi, Cambridge Systematics, Inc., Medford, MA, United States

© 2021 Elsevier Ltd. All rights reserved.

Summary of the State of the Practice	590
Travel Demand Model Components	591
Demand Model Components and Their Parameters	592
Trip Assignment Methods	592
Model Validation	593
Model Application	593
Four-Step Modeling	593
Activity-Based Modeling	594
How and Where Different Modeling Approaches are Used Today	594
Addressing Emerging Issues	595
References	595

Travel demand models are analysis tools used in a variety of transportation planning studies. These models began to appear in the 1950s, and the analytical approaches have become more sophisticated as they took advantage of research in such fields as behavioral analysis, statistics, and operations research, as well as increases in computational power (Boyce and Williams, 2015). They are used in short and long-range transportation planning, analyses of proposed transportation infrastructure investments, public policy analyses, and studies of the effects of changing demographics and technology on transportation system use.

The first models were developed for use in large cities and metropolitan areas at a time when many jurisdictions were investing heavily in transportation infrastructure and needed tools to evaluate various policy and construction options (Committee for Determination of the State of the Practice in Metropolitan Area Travel Forecasting, 2007). Over time, models began to be used in a wider group of geographic contexts, ranging from relatively small areas to long distance interurban areas or corridors, and even entire countries, states, or provinces. The types of analytical methods used in the models became more varied in response to the various analysis contexts.

Moreover, the types of analyses for which travel demand models are used have evolved, as many mature jurisdictions began to focus less on large infrastructure projects such as new highways and more on smaller scale improvements and efficient use of existing transportation systems. In many places, there are statutory or regulatory requirements for transportation planning that affect analyses for which travel demand models can be valuable tools. For example, the United States requires that metropolitan areas that are designated as not meeting air quality goals must demonstrate that transportation plans conform to plans for meeting statutory standards. Recently, some jurisdictions have set ambitious goals for reducing carbon footprints that could require substantial changes in the ways that people travel.

Other transportation planning issues that have emerged since the first travel demand models appeared include the following:

- investment in new urban rail systems (models first emerged during a time of disinvestment in urban rail);
- an increased interest in managed lanes and priced roadways, especially dynamic pricing that varies by time of day and/or demand level;
- interest in promoting active transportation modes such as walking or bicycling, as a means of reducing vehicular congestion and emissions of pollutants, as well as improving public health;
- the emergence of transportation network companies (TNCs) that use mobile technology to match riders with drivers;
- the emergence of “micromobility,” or shared vehicles such as bicycles, scooters, and automobiles available in public settings that make use of remote payment technology; and
- the recognition that autonomous vehicle technology will soon make automobile travel available to persons who do not drive or who prefer to use their in-vehicle time doing something besides driving.

Summary of the State of the Practice

The state of the practice in travel demand modeling appropriately varies depending on the application context (Committee for Determination of the State of the Practice in Metropolitan Area Travel Forecasting, 2007). Larger problems, naturally, require more sophisticated tools, and so in many parts of the world the models in large metropolitan areas, as well as some larger scale models (national and statewide), use more complex approaches (though older, simpler practices persist in some countries). The state of the practice can be briefly summarized as follows:

- On the demand side, in larger planning areas with complex issues to analyze, disaggregate activity-based models (Castiglione et al., 2015) can be considered the state of the practice. In smaller planning areas where the issues do not require as detailed

analysis, aggregate trip-based models (Ortúzar and Willumsen, 2011) are used. In some larger areas, the older aggregate approaches are still in use.

- On the network assignment side, the state of the practice is still to use aggregate, static assignment procedures (Cambridge Systematics, Inc. et al., 2012) in most contexts. While disaggregate dynamic traffic assignment (DTA) (Chiu et al., 2011) approaches have existed for many years, they are not widely used in large regional contexts; they are more frequently applied for smaller areas or corridors.

In disaggregate approaches, typically the activities and travel choices of individual agents (persons, households, vehicles) are individually simulated; when aggregate results are required, the results from the individual simulations are summed. In aggregate approaches, the traveler population is segmented into a relatively small number of groups, and the travel made by each segment is modeled in an aggregate manner.

Travel Demand Model Components

Regardless of whether the underlying model uses an aggregate or disaggregate application approach, it is useful to consider the major components of a travel demand model using the concepts of the demand side and the supply side. Highway, transit, pedestrian, and bicycle networks, which are spatial representations of the transportation system in the model region, comprise the supply side of travel demand models. Highway networks include the locations and characteristics (such as capacities and speeds) of roadways. Transit networks represent the bus and rail systems (routes, stations, transfer points, and supporting parking facilities). In a general sense, the outputs of travel demand models that are used for planning purposes are facility level volumes (persons or vehicles). For vehicular travel, these volumes represent the number of vehicles using roadway segments in the model's highway network. For person travel, the volumes represent individual persons using public or private transit systems, walking, or bicycling.

Most personal travel is not truly a good that consumers desire but rather a means of reaching different locations where activities, such as working, education, shopping, recreation, or personal business, are performed (Castiglione et al., 2015). Early models (as do many current models) treated travel as a good to be consumed; many recent models treat the activities people wish to undertake as the desired goods and simulate the need for travel as a derived demand. Either way, the amount, locations, times, and modes of travel represent the demand side of models.

Representing the demand side of the model means simulating the movement of persons and goods throughout the region along the networks. Whether the travel is modeled directly or as a derived demand from activity undertaking, the demand side estimates the amount of travel (how many persons are traveling), the locations (origins and destinations of trips or journeys), the modes used to travel between origins and destinations, and the times at which the persons are traveling. This process is necessarily an oversimplification of the complex ways in which people make decisions about whether, where, how, and when to travel. It is impossible to model the myriad factors (some of which cannot be quantified easily) that affect human travel behavior decisions and the interactions among travelers, especially within households, that influence these decisions.

The vast majority of travel demand models simulate travel for a typical weekday (Cambridge Systematics, Inc. et al., 2012), in many cases segmented into different time periods within the day. Depending on the type of model, the demand components of the model may consist of the following components:

- characteristics of the traveler population, such as auto ownership;
- the amount of travel made by individuals in the travel population or in an activity-based framework, the number of activities requiring travel to their locations;
- the locations of the activities or the travel locations (trip ends);
- the modes used to travel between origins and destinations; and
- the time of day at which the travel takes place, or when the activities requiring travel begin and end.

The final step in the modeling process is assignment, which performs route choice using the supply side of the model (networks) for the trips simulated on the demand side of the model. All models perform highway assignment for automobile and truck trips; in regions where transit is a significant component of travel, transit assignment is also performed. The units for highway assignment are vehicles, or vehicle trips, while the units for transit assignment are persons, or person trips using transit. While there are a few recent applications where pedestrian or bicycle trips are assigned to networks, in most models these trips are not assigned.

A travel demand model produces a set of outputs that can be used for further analyses, based on a set of inputs that can be more readily obtained or estimated. There are two basic sets of inputs to the model:

- The transportation networks, from which estimates of travel distances, times, and costs can be obtained, and
- Socioeconomic and land use data, which are used to quantify the size and characteristics of the traveler population and to identify the relative attractiveness of various locations to which people might travel. Usually, these data are aggregated to "zones"—typically a model region has hundreds or thousands of zones—but in some cases, more detailed geographic resolution is used.

There are also a few data inputs that are technically not part of either of these two categories. These include fuel and other auto operating costs and geographically specific parking costs, which are not strictly tied to the transportation networks.

While it is recognized that travel behavior is a complex set of interconnected choices, travel demand models separate those choices into individual components so that the estimation of travel demand is a tractable problem. This means that the choice model components are applied sequentially, and while the effects of downstream components can affect upstream choices in simple, aggregate ways, for the most part the effects filter mainly in one direction.

Demand Model Components and Their Parameters

Travel demand model components use mathematical formulations to relate the model outputs to the model inputs (which often include outputs of previous components in the model sequence). The type of mathematical formulation varies depending on the particular component and the overall structure of the model. Commonly used formulations include simple factoring, linear regression, discrete choice (multinomial, nested, or ordered response logit, or probit) (Ben-Akiva and Lerman, 1985), and hazard duration.

The parameters for each model component may be estimated using observed travel behavior data, such as travel surveys, or asserted based on previous experience in the model region or elsewhere. Both estimation from observed data and assertion are common practices. When parameters are estimated, statistical methods are used to produce the estimates using maximum likelihood or other optimization processes.

When parameters are asserted, it is typical to use values from other contexts, such as similar areas to the model region, or averages of parameters from multiple contexts. One of the main motivations for asserting parameters is that the data to estimate parameters may not be available and may be difficult or expensive to collect. The sample sizes required to obtain statistically significant parameters given the population segmentation used in the model can be quite large. Despite the widespread use of transferred parameters, there is not strong evidence that model parameters are truly transferable in all cases (Rossi and Bhat, 2014).

Trip Assignment Methods

Highway assignment methods are based on estimating least cost paths between trip origins and destinations. Most of the cost is attributable to travel time; however, other costs, notably toll charges, may also be considered.

Because aggregated highway trips can cause congestion and therefore affect roadway travel times, it is not possible (except under uncongested conditions) to simply calculate shortest paths and allocate all trips to these paths. It is therefore necessary to consider the effects of congestion on travel times by relating traffic volumes, capacities, and travel times. Since driving behavior is complex and varies substantially among drivers, and because network representations in models do not consider all roadway characteristics that can affect travel time and capacity, these relationships are necessarily simplified. It is also important to note that travel times vary within assignment periods, and many drivers choose routes based on considerations other than minimizing costs (e.g., route reliability); having to ignore these effects further oversimplifies the route choice process.

There are two main approaches to highway assignment, static assignment (Sheffi, 1985), and dynamic traffic assignment (Chiu et al., 2011). In static assignment, origin-destination trip matrices are loaded onto the highway network, usually for various subperiods of the modeled day (e.g., the morning and afternoon peak periods). The most common static assignment approach is user equilibrium, which is based on Wardrop's principle that drivers choose to minimize their own costs, and equilibrium is reached when no driver can improve her travel time by switching routes (Wardrop, 1952). Model users can determine the settings for the assignment process (e.g., the function relating travel time to volume and capacity) using the modeling software. The main limitations of static assignment include the following:

- A quite limited set of network characteristics (facility type, area type, and number of lanes) is available for use in calculating capacity, and considerations such as signal timing and phasing, turning lanes, parking, and pedestrian activity are usually unavailable.
- The travel time on each link is computed using the volume and capacity for only that link, and interactions among individual highway links (downstream queuing, turns across oncoming traffic) are not considered.
- Volumes and travel times are averaged over the course of the assignment time period; variations in demand across the period are not considered.
- There is no requirement that demand not exceed capacity; all trips in the period must be loaded onto the network.

DTA methods directly simulate the movement of vehicles through the network over time. DTA can consider additional network characteristics, including turning lanes and signal timing (though maintaining all of this information for large networks can be problematic). In DTA applications, vehicles are processed through the network respecting capacity constraints and interactions with other vehicles. This allows them to simulate queuing when all demand cannot be processed through specific points in the network.

Despite the advantages of DTA over static assignment, its use is mainly confined to area or corridor studies rather than the large metropolitan regions (or bigger areas) for which many travel demand models are applied. There are several research-oriented examples of regionwide DTA applications, but they have been associated with long run times and insufficient data for proper application and validation (though in some cases input data are synthesized). Static assignment remains the standard practice in most urban area travel demand models.

Although there has been research into dynamic transit assignment methods, the transit assignment methods used in existing travel demand models are static. In models of areas where transit crowding can affect path choices, capacity constraints may be introduced.

Model Validation

Validation of a travel demand model is an essential step in ensuring that the model can produce reasonable outputs for the planning analyses that it will be used for (Cambridge Systematics, Inc., 2010; Pendyala and Bhat, 2008). Model validation generally consists of:

- comparisons of model results for a scenario that represents conditions that can be observed to empirical data for these conditions;
- reasonableness checks of model results in cases where the information represented by outputs cannot be observed; and
- tests of the sensitivity of the model to changes in inputs that represent the types of scenarios for which the model will be applied.

Model Application

Travel demand models are applied using specialized software packages, usually on personal computers or servers, or, more recently, using cloud-based services. In the early days of models, when mainframe computers were the only hardware options, the United States federal government, specifically the Urban Mass Transportation Administration (now the Federal Transit Administration) and the Federal Highway Administration, developed the Urban Transportation Planning System, a software package designed to apply the four-step modeling process. As personal computers came into common use in the 1980s, proprietary model application packages developed by private companies began to appear. Such packages are still in common use although they are often supplemented with custom software built for particular applications or, more recently, open source software that can be adapted for specific models. The latter are often used in the application of activity-based and other advanced travel demand models.

The sequential nature of model application results in some dilemmas in cases where choices may affect one other. For example, travel demand in a corridor depends on the travel times in the corridor, but these times are affected by the level of demand if the corridor becomes congested. To compensate for this, models are often applied in an iterative manner, where the inputs to demand side components are updated based on the travel times results from the highway assignment. In theory, the iteration continues until “convergence,” that is, when the output travel times are equal to the inputs; in practice, due to model run time considerations, the process is often truncated prior to true convergence.

Four-Step Modeling

Until the 1990s, the “four-step” modeling approach was ubiquitous (Boyce and Williams, 2015), and it remains the most commonly used approach. The four steps refer to trip generation, trip distribution, mode choice, and trip assignment—the first three steps representing the demand side. Four-step models are aggregate in nature, meaning that large segments of the traveler population are modeled together. As a result, it is not possible to identify travel behavior for individuals or for groups smaller than the segments used in the model. Spatial resolution is nearly always at the zone level.

Four-step models are trip-based, meaning that they model trips between two points without considering stops along the way (a stop would result in the end of a trip, after which a new trip would begin). Trips are modeled without consideration of other trips made by the same person.

The four-step process can be summarized as follows:

- Trip generation—Trips are segmented by purpose; usually there are three to eight purposes, such as home-based work, home-based shopping, nonhome-based, etc. Trip ends are classified as productions, which are the home ends of home-based trips, or attractions, the nonhome ends. Trip production models typically use crossclassification structures (Cambridge Systematics, Inc. et al., 2012), where the households in each zone are segmented by two or possible three dimensions, and the number of households in each zone is multiplied by an estimated or asserted trip rate. Trip attraction models are typically linear regression models relating the number of trips to the employment by type and the population in the zone.
- Trip distribution—For each trip purpose, this process matches up the trip productions and trip attractions to create trip tables (matrices) based on a formulation that relates the attractiveness of a zone to the number of attractions from the trip generation model and the impedance (time, cost, and/or distance) from the production zone. The model parameters relate the attractiveness of a zone to the impedance. Early models used the gravity model, an aggregate procedure where the productions for each zone are allocated to attraction zones based on the attractiveness function. While gravity models are still popular, most larger four-step models use multinomial logit trip distribution models, usually called “destination choice” or “location choice” models, where the attractiveness of each destination is represented in a logit utility function. The outputs of the trip distribution process are production–attraction trip matrices representing the total person trips for each purpose.
- Mode choice—The person trip table outputs from the trip distribution step are split into trips by mode, including various types of auto, transit, walk, and bicycle travel. There could be anywhere from two to a few dozen modes; these represent various submodes, including auto occupancy levels as well as transit types and access/egress modes. Typically, mode choice uses a nested

logit formulation although multinomial logit is often used when there are few modal alternatives. Besides the trip tables, other inputs can include the level of service (time, cost, number of transit transfers, etc.), traveler characteristics, and land use related variables. In the mode choice step, auto vehicle trip tables and transit person trip tables are produced at the origin-destination level for use in assignment.

- Trip assignment—Four-step models use static, aggregate assignment procedures, consistent with the aggregate outputs from the demand side.

While some models assign trips for the entire weekday, in most four-step models for larger regions, trip assignment (at least highway assignment) is usually done for several periods shorter than the entire modeled weekday. At some point in the process (the point varies by model), the daily trips or trip tables are divided into trips for each of the time periods used in assignment.

Activity-Based Modeling

Activity-based models (Bowman, 1998; Castiglione et al., 2015) have several fundamental differences from four-step models:

- The demand side is disaggregated; the activities for each person in the population are explicitly simulated.
- Travel is simulated as a way to accommodate the spatial and temporal differences in activity locations and timing.
- Since each person's activity pattern is simulated, the interdependencies among a person's trips are considered.
- Activity-based models are tour-based, rather than trip-based. Entire tours starting and ending at home or work are simulated as individual units, followed by simulation of the trips (stops) that comprise the tours. In other words, trip chaining is explicitly considered.
- In most activity-based models, the major interactions among members of households are considered (e.g., joint tour making among multiple household members, or escorting children to and from school).

There is more variation among the structures of different activity-based models than is the case for four-step models. Some basic features common to most activity-based models include the following:

- Synthetic population—This is generated to represent the true population of the model region. It is used as the basis for the simulation of individual activities and travel. The synthetic population is generated using fitting methods based on control totals that represent the assumed or observed demographics in the modeled scenario, with seed distributions often coming from large sample surveys (in the United States, the American Community Survey is often used).
- Long-term choices—These components represent choices or characteristics of the population that do not vary from day to day. Examples include vehicle ownership, workplace and school location, and, in some models, characteristics of the members of the population such as income levels and worker/student status. Common model forms include various discrete choice formulations though simpler methods are sometimes used.
- Activity generation—The set of activities performed by each individual in the population is simulated; this process considers the traveler's household structure and, in many models, the activities undertaken by other household members. These components are among the more complex in activity-based models due to the various spatial and temporal constraints on activity patterns. Common model forms include multinomial logit and multiple discrete-continuous extreme value. Activity generation is roughly equivalent to trip generation in four-step models though is far more complex.
- Activity timing, location, and mode of travel, and generation of stops on tours—In this series of components, the timing of activities (start and end times/durations), their locations, and the modes used to travel from home on the associated tours are simulated. Again, the spatial and temporal constraints on the various activities and the travel between their locations are considered. Common model forms include multinomial, nested, and ordered logit as well as (for time of day choice models) hazard duration.
- Stop locations and timing, and travel modes between stops—These models simulate the activities that take place at intermediate stops on the tours previously simulated. They are somewhat similar to the tour level components, but the alternatives are heavily constrained by the previously simulated tour level choices.

The outputs of the demand side simulation are tour and trip rosters, including the origins, destinations, times of day, and modes of the tours and trips; the characteristics of the travelers who make them are known and are included in the rosters. These rosters represent all of the personal travel in the model region made by residents of the region. Despite the disaggregate approach on the demand side, most activity-based models use static aggregate assignment procedures. This means that the trip rosters must be aggregated to create trip tables for assignment. There have been several research efforts into the integration of disaggregate demand DTA models, but these models have not yet made it into mainstream application.

How and Where Different Modeling Approaches are Used Today

While no model is a completely accurate depiction of behavioral reality, activity-based models are more realistic since they consider more of the factors that affect behavior and travel choices than four-step models do. Four-step models ignore the interactions among

multiple trips made by the same person (trip chaining) and among members of a household. Activity-based models are also more flexible because they are disaggregated; not only can a richer set of variables be considered (since there is no need to estimate statistically significant models for each of a large set of market segments defined by multiple variables), but the disaggregate outputs can be summarized using any variables used in the model. For example, outputs can be summarized by income level, geographic subregion, vehicle availability, ethnicity, age, or gender.

Despite these advantages, four-step models are still much more common than activity-based models. The latter can be more expensive to develop and maintain, and model run times are longer due to the increased computation required. In addition, their complexity relative to four-step models can make it more difficult for many practitioners to develop and maintain them, especially in contexts where planning resources are tight. Even though activity-based models have been used for more than two decades, they are still relatively new in many places. Furthermore, for many applications the simpler four-step model provides adequate information, especially for smaller, less diverse modeling environments.

Presently, activity-based models are used in most of the large metropolitan areas in the United States, as well as in many other countries around the world such as the Netherlands and Israel. In many countries, the four-step approach is still the standard, as well as in most smaller planning regions throughout the world.

Addressing Emerging Issues

While the types of planning analyses for which model results are used has evolved over time, recent years have brought an even greater pace to the changing needs of modeling. This is a direct result of changes in the way transportation supply is provided (e.g., open road tolling, TNCs, micromobility) as well as anticipated “game changers” such as autonomous vehicles. Additionally, traveler behavior may be changing as many cities are experiencing increases in residential development in or near downtowns, and greater percentages of the population are older. Furthermore, as computing power has increased, models can more feasibly deal with finer levels of detail in demographics and temporal and spatial resolution.

The aggregate nature of four-step models makes it difficult for them to adequately deal with these recent and expected changes in transportation service, which affect travelers at an individual level and can have substantial impacts on things like auto ownership. This makes activity-based models better suited to deal with these changes. However, the effects of these changes may be of greater planning significance in larger, more diverse planning environments, where activity-based models are already more likely to have been implemented, and smaller regions may choose not to invest in more sophisticated models to address them.

A modeling-related topic that has been gaining attention recently is dealing with forecasting uncertainty. Whether they are applied in an aggregate or disaggregate manner, travel demand models are essentially deterministic. It has always been recognized that the point estimates that models output have a degree of uncertainty that is unquantifiable since the error associated with much of the model input data (such as demographic or cost forecasts) as well as with the model parameters and the data used to estimate them may not be easily quantified.

In recent years, there has been increasing recognition of the effects of uncertainty around forecasts on planning decisions, especially as the effects of changes in transportation supply and travel behavior are becoming more difficult to predict (e.g., due to the introduction of autonomous vehicles). There have been efforts to incorporate risk analysis into the ways in which models are applied, through means such as exploratory modeling analysis. New tools are being developed that can work with existing models of different types to estimate probabilities of outcomes and of performance measures being achieved.

References

- Ben-Akiva, M., Lerman, S., 1985. *Discrete Choice Analysis*. MIT Press, Cambridge, MA.
- Bowman, J.L., 1998. *The Day Activity Schedule Approach to Travel Demand Analysis*. PhD dissertation. Massachusetts Institute of Technology.
- Boyce, D., Williams, H., 2015. *Forecasting Urban Travel: Past, Present and Future*. Edward Elgar, Cheltenham, UK.
- Cambridge Systematics, Inc., 2010. *Travel Model Validation and Reasonableness Checking Manual*, second ed., Federal Highway Administration, Washington, DC.
- Cambridge Systematics, Inc., Vanasse Hangen Brustlin, Inc., Gallop Corporation, Bhat, C.R., Shapiro Transportation Consulting, LLC, Martin/Alexiou/Bryson, PLLC, 2012. NCHRP Report 716 - travel demand forecasting: parameters and techniques, Transportation Research Board.
- Castiglione, J., Bradley, M., Gliebe, J., 2015. Activity-based travel demand models: a primer. SHRP 2 Report S2-C46-RR-1, Transportation Research Board.
- Chiu, Y.-C., Bottom, J., Mahut, M., Paz, A., Balakrishna, R., Waller, T., Hicks, J., 2011. *Transportation Research Circular E-C153: Dynamic Traffic Assignment: A Primer*. Transportation Research Board of the National Academies, Washington, DC.
- Committee for Determination of the State of the Practice in Metropolitan Area Travel Forecasting, 2007. *Metropolitan Travel Forecasting: Current Practice and Future Direction*. Transportation Research Board, Washington, DC.
- Ortúzar, J. de D., Willumsen, L., 2011. *Modelling Transport*, fourth ed. John Wiley & Sons, Ltd., Chichester, West Sussex, UK.
- Pendyala, R., Bhat, C., 2008. Validation and assessment of activity-based travel demand modeling systems. *Innovations in Travel Demand Modeling: Summary of a Conference*, vol. 2, papers.
- Rossi, T.F., Bhat, C.R., 2014. *Guide for Travel Model Transfer*. Federal Highway Administration, Washington, DC.
- Sheffi, Y., 1985. *Urban Transportation Networks*. Prentice Hall, Englewood Cliffs, NJ.
- Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. *Proceedings, Institute of Civil Engineers, Part II*, vol. 1, pp. 325–378.

Travel Model Calibration and Validation

Ram M. Pendyala, School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, AZ, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	596
Key Terminology Related to Model Calibration and Validation	596
Sources of Error and Uncertainty in Travel Model Calibration and Validation	598
Calibration and Validation of Model Components and Entire Model System	599
Sensitivity Analysis of Model Components and Entire Model System	602
Summary	604
References	605
Further Reading	605

Introduction

Travel demand forecasting models have been developed to aid in short- and long-range transportation planning efforts in metropolitan areas around the world for more than 50 years. Over the past few decades, travel models have become increasingly sophisticated – both in terms of the behavioral processes that they seek to represent and in terms of the methodological approaches to model and predict activity-travel patterns. In the recent past in particular, travel demand forecasting models have increasingly strived to predict household and person activity-travel patterns by microsimulating individual mobility choices and movements through time and space, while explicitly recognizing the many constraints and opportunities that govern these behavioral phenomena. These models, often called activity-based or tour-based models, are increasingly being implemented in metropolitan areas around the world. At the same time, more traditional four-step travel demand models continue to be in wide use. More recently, the availability of big data – often constituting trajectories of millions of individuals collected through passive monitoring of cellular phone locations and movements – has made it possible to develop more data-driven modeling approaches and opened up new possibilities for validating the predictions of travel demand forecasting models.

Even though a variety of modeling tools are increasingly finding their way into agency practice, the principles of model calibration and validation are quite robust and applicable across modeling methods and platforms. When a travel demand forecasting model is developed, it is desirable to ensure that the model is able to replicate current base-year real-world conditions within an acceptable tolerance, and is able to provide forecasts of travel demand and activity patterns that are behaviorally intuitive, consistent with documented evidence (if available), and useful for planning purposes. It is in this context that this chapter purports to offer an overview of travel model calibration and validation methods, fully recognizing that these terms are often interpreted and use somewhat interchangeably in the profession.

Planning transportation investments and formulating transportation policies relies on the ability to model and replicate activity-travel demand under a wide variety of policy scenarios and system conditions. Travel models often have many components, covering the wide range of activity-travel and mobility choices that constitute travel demand. These may include choices such as activity generation, type of activity, timing (scheduling of activity), trip departure time choice, route choice, mode choice, destination choice, activity sequencing (trip chaining), activity duration (time use), and joint activity-travel engagement (accompaniment). The model components capturing each of these (and other) choices are then linked together, often with appropriate feedback loops, to comprise an entire travel model system. Each individual model component needs to be calibrated and validated in isolation, and the model as a whole needs to be calibrated and validated as well. Errors in calibration and validation of a specific model component could lead to downstream issues in the model chain, as the propagation of errors may will inevitably impact the accuracy and usefulness of the travel forecasts obtained from the model. Calibration and validation are therefore critical steps in the model development and deployment process.

This chapter is intended to offer an overview of travel model calibration and validation processes and considerations. The next section offers explanations of the various terms and a distinction between disaggregate and aggregate model calibration and validation. The third section identifies sources of error and uncertainty in the travel model calibration and validation arena. The fourth section offers an overview of calibration and validation of travel model components and the model as a whole. A few examples are furnished in this section as well. Because sensitivity analysis is a critical part of establishing the validity and usefulness of a model, a separate section is dedicated to sensitivity analysis. Finally, the chapter closes with a concluding section.

Key Terminology Related to Model Calibration and Validation

The model development and deployment process may be viewed as consisting of four major steps (Cambridge Systematics, Inc., 2010). This section offers an overview of these steps with a view to help delineate the steps and draw a distinction between calibration and validation. Although there may be slightly differing definitions and interpretations of these terms in the literature

and model documentation, the essential distinctions are well-recognized and the descriptions presented in this chapter are quite generalizable to most model development efforts.

The first step in the model development process is model estimation. Model estimation (which is not within the scope of this chapter) involves the calculation of model coefficients, parameters, and goodness-of-fit statistics. This is typically accomplished using travel survey data sets, although secondary data sources may be used to augment and enhance the travel survey data set. A variety of statistical methods including regression-based techniques, discrete choice models, and other econometric methods are typically used to estimate coefficients and parameters. Model coefficients are calculated such that the model fits the data as closely as possible, with the fitting function defined by the specific econometric methodology being used. Model coefficients are often examined for their behavioral intuitiveness, relative magnitudes and signs, and potential sensitivity to changes in explanatory variables. Many different model specifications are often tested before a final model specification is adopted. In instances where a coefficient value consistent with expectations is not being obtained, or local data are sparse, the model developer may choose to assert a coefficient value; this is often referred to as coefficient assertion.

Model calibration (which is within the scope of this chapter) often involves the adjustment of parameters so that model predictions replicate observed patterns in the data set used to develop the models in the first place. Estimated or asserted coefficients may be used to predict distributions and averages of various mobility choices and travel characteristics (for the estimation data set sample) and compared against the observed distributions in the data set. This is similar to comparing a predicted value to an observed value in a regression modeling effort. The difference between these values constitutes a difference or an error that the model estimation step attempts to minimize in calculating the coefficients. Comparisons of predicted versus observed travel characteristics provide the ability to assess the extent to which the model fits the data. If a certain comparison is viewed as subpar, it is possible to change the specification of the model and re-estimate model coefficients with a potentially richer model specification. Even after repeated attempts at model specification fine-tuning, if the match between model predictions and observed sample characteristics is not deemed acceptable, then parameter adjustment and coefficient assertion may be done to “calibrate” the model.

Following model calibration, model validation is undertaken. Model validation may take a number of forms, with two key forms described here. First, model validation involves comparing model predictions to real-world conditions, where the real-world conditions are derived from external data sources not used in model estimation and calibration. A typical validation exercise may involve comparing predictions of traffic volumes on major corridors to observed traffic counts on those corridors for a comparable time period. In this case, the traffic counts are an external source of data and not used in the model estimation and calibration process to begin with. Similarly, transit boardings, alightings, and route-level ridership numbers may serve as an external data source to validate the predictions coming out of a travel demand model. Care should be exercised to ensure that the external data used to check the validity of a travel model are accurate and correspond to the outputs of the travel model in terms of spatial and temporal resolution. Sometimes, model estimation and calibration may be performed using a subset of a travel survey sample, keeping the hold-out sample for validation purposes. In this case, model estimation and calibration is performed on the estimation (sub)sample, and then the model is applied to the hold-out validation (sub)sample. The predictions are compared to observed characteristics for the validation sample to determine the ability of the model to replicate observed patterns in an independent sample; because this subsample was not included in the model estimation data set, some regard this effort as a model validation exercise. This would, however, require that the survey sample data set is large (thus offering a sizeable estimation subsample and a hold-out validation sample). In addition, even if such a validation exercise is conducted, it is advisable to perform additional validation checks against external data sources such as traffic counts, census journey to work flow data, and other secondary sources that are completely independent of the travel survey data set.

The second major element of model validation is sensitivity testing. The sensitivity of models to changes in input conditions can be assessed through the computation of elasticity measures (derived from coefficient values) and other metrics of interest (e.g., computation of value of time, or ratio of out-of-vehicle to in-vehicle time coefficient values, in a mode choice model) that convey the degree to which a model reflects tradeoffs in behavioral choice phenomena. In addition, sensitivity testing involves the application of models to many different scenarios defined by a variety of input conditions. Socioeconomic variables, network conditions, modal level of service attributes, and land use variables may all be changed to different degrees in various combinations; running the model on alternative hypothetical scenarios will provide useful insights on the ability of the model to provide behaviorally intuitive, reasonable, and useful forecasts for planning and policy making. In some cases (but not all), the degree of sensitivity exhibited by the model (to changes in inputs) can be compared against real world empirical evidence to check the validity of the forecast (e.g., Evans, 2004; Milam et al., 2017).

Another key aspect of model validation is the need to distinguish between disaggregate and aggregate validation. This distinction may be particularly drawn in the context of recent microsimulation-based activity-based models of travel demand. Such models often provide individual activity-travel patterns (records) as outputs. When interfaced or integrated with a dynamic traffic assignment (DTA) model, they are able to provide individual vehicle trajectories as they traverse the network from origin to destination. It is potentially feasible to check the validity of these individual activity-travel patterns and vehicle trajectories for their reasonableness. Activity-travel patterns should not violate laws of physics, spatio-temporal constraints cannot be breached (in most circumstances), children (under a certain age) cannot be left abandoned without adult supervision, and the same person cannot be at two locations at the same time. Similarly, the route, travel time, and speed associated with a vehicle trajectory on the network should be reasonable and consistent with real-world path choice behavior. Many of these checks can be performed to examine disaggregate level outputs and ensure that the model is producing patterns of travel that are plausible. This reasonableness and logic check is a critical part of disaggregate model validation.

Aggregate model validation applies to both traditional four-step travel demand models and newer activity-based microsimulation models of travel demand. In this case, aggregate measures of travel demand are checked against known external data sources that represent ground truth. Such external data sources may include census socioeconomic and demographic data, census journey-to-work flow data, traffic volume data, travel time and speed data, and regional measures of vehicle miles of travel (VMT), trip rates, and transit ridership. Comparisons of aggregate measures of travel demand output by the model with external data sources provides a means of determining whether the model is able to replicate overall patterns of travel within the specified time period, geography, and market segment of interest. Aggregate validation can be done for specific demographic and socioeconomic groups, specific corridors in the network, or specific geographical subareas or markets to ensure that the model is capturing the heterogeneity that may be prevalent in a region. Simple region wide comparisons may mask the variation that exists across markets, space, and time.

Sources of Error and Uncertainty in Travel Model Calibration and Validation

It should be recognized that there are a number of sources of error and uncertainty associated with travel model calibration and validation. These sources of error and uncertainty often contribute to the need for model adjustments so that model predictions more effectively replicated observed ground truth conditions. However, there may be instances where model validity checks can reveal underlying issues with the land use data, network coding, or travel survey data; the model developer can then fix underlying data issues without the need to make ad-hoc or arbitrary model adjustments. Before making corrections to a model specification through a calibration process, it would be advisable to first check all input data for any anomalies or issues and ensure that the data sets do not have errors that could adversely impact a model validation exercise. The sources of error and uncertainty are described in this section.

One of the key sources of error is the model specification error. Despite decades of research into traveler behavior and values across a myriad of disciplines, much remains to be learned about travel decision-making processes under different conditions. In addition, travel characteristics present both dynamics and stationarity over time, rendering it difficult to determine elements that should remain static versus those that should evolve. Model specification is an abstract representation (and an approximation) of traveler choice processes, and hence any calibration and validation exercise should involve a close look at the model structure, sequencing of model components, the causal decision hierarchy implied by the model component sequence, and the array and form of explanatory variables included in various model components. In addition, the forms of the statistical models should be also be revisited to ensure that the methodological approach to modeling different choice phenomena are appropriate.

There may also be errors in the validation data, that is, the secondary external data against which model predictions are compared. As noted earlier in this chapter, external validation data may include traffic counts, link/corridor volumes, speed, and travel time data offered by an external source, transit ridership, travel survey data, mobile phone trajectory data, and census socioeconomic and journey-to-work flow data. All of these data sets are used extensively in model validation exercises and hence it is critical to ensure that the methods used to collect and compile the data are appropriate, correct, and devoid of elements that could lead to biases in the data. Moreover, it is critical to have appropriate metadata so that details about the validation data are well-documented. In this way, it is possible to ensure that comparisons between model forecasts and validation data are appropriate and refer to the same time period, geography, underlying conditions, and output variable. In some cases, it may be possible to identify and correct errors in the validation data (or check against multiple sources) without having to make corrections to the model structure and specification itself.

Sensitivity analysis is a key component of any model calibration and validation exercise. When conducting a sensitivity analysis, it would be useful to construct a range of scenarios for which the capabilities of the model can be tested. Some scenarios should be constructed in such a way that the general magnitude and direction of change in the forecast is known by virtue of available empirical evidence or prior experience. For example, it is generally known how transit ridership responds to changes in fare or frequency of service. Similarly, it is likely that an area has some evidence (from the past) on how traffic volumes and travel times change when a lane closure occurs (say, in a workzone) or a capacity expansion project (e.g., addition of a highway lane) is accomplished. There is also evidence documented in the literature regarding changes in travel demand that result from changes in system conditions, and model forecasts (of change) can be compared against documented evidence to verify model validity.

On the other hand, there may be a number of scenarios for which real-world evidence is absent. This is particularly true in the context of emerging modes of transportation and mobility services, new traveler information systems and mobile apps, and futuristic transportation technologies such as autonomous vehicles. In such instances, there is no basis to determine whether a model forecast is reasonable or valid. In other words, there is considerable uncertainty as to how the market may adopt and adapt to the new mobility options, modes of transportation, transportation policy, or traveler information system. This uncertainty renders model validation difficult, especially when the model is expected to be used to forecast long-term transportation trends when new rapidly evolving technologies may find their way into the marketplace. The analyst should recognize the uncertainty associated with such scenarios and explain all of the assumptions underlying model applications involving future transportation technologies for which no past or current evidence or data is available. Appropriate reasonableness and logic checks may be applied to assess whether the model is generally offering forecasts consistent with behavioral intuition, but there is little basis to assess whether the magnitude of change is of the appropriate order.

Finally, as noted at the beginning of this section, there may be errors in the input data that drive the model estimation and development process. Models rely on a suite of input data sets including land use data, socioeconomic and demographic data

(census data, for example), travel survey data, and highway and transit networks (described by a series of attributes such as speeds, capacities, and link lengths). These data sets and the variables they contain are critical to model specification, estimation, and calibration. If any of these data sets have errors, the resulting model is likely to be erroneous and offer forecasts that are invalid (do not match ground truth) and do not pass reasonableness and logic checks. In that case, it would be of value to try and isolate the potential error in the input data sets and make necessary checks and corrections that may be warranted.

Calibration and Validation of Model Components and Entire Model System

This section offers a brief overview of calibration and validation of individual model components that are commonly found in most travel demand forecasting model systems. As noted earlier, most assessments of model predictions can be performed for specific market segments, times of day, purposes of travel, geographic (sub) areas, and network corridors or links. As long as observed real-world data about system conditions are available for these specific segments and markets, model validation can be conducted at the level of fidelity for which data is available.

Before embarking on any model development effort, it is important to check the validity of the input data that will drive the model development process. This may include travel survey data, land use and socioeconomic data, population census data, and highway and transit network data. Likewise, care should be taken to assemble and check validation data including traffic counts by time of day and vehicle type for key corridors and links, transit ridership data by route and time of day, travel time and speed data from third party vendors or detectors, and origin-destination flow data that may be available from external sources. In the interest of brevity, validation checks for these data elements are not described here, as this chapter focuses on calibration and validation of model components; details on checking validity of input and validation data are available elsewhere ([Cambridge Systematics Inc., 2010](#)). The remainder of this section will focus on key components of typical travel model systems.

1. Socioeconomic or Synthetic Population Models

Travel demand model systems rely on socio-economic data. In the case of activity-based microsimulation models, a synthetic population of the region is needed so that individual-level activity-travel patterns can be simulated. For more traditional four-step travel demand models, the socio-economic models usually aim to derive joint distributions of various socio-economic characteristics (household size, number of workers, household income, and vehicle ownership) based on available marginal distributions of such variables. The joint distributions may be derived through iterative proportional fitting (IPF) techniques or through the use of curves that represent population proportions of various subgroups. Synthetic population models aim to generate a population of agents for the region such that the characteristics of the synthetic population mirror those of the true population as captured in the census data. In either case, the resulting distributions of various socio-economic characteristics should be compared against census data or other socio-economic data that may be available from an external source. Checks should be made at the regional level as well as a more fine-grained spatial level, such as census tract, traffic analysis zone (TAZ), or census public use microdata area (PUMA). In the case of a synthetic population generation process, the resulting synthetic population should be checked (for validity) with respect to both variables that were controlled in the synthesis as well as variables that were not controlled in the synthesis. If the match to (important) uncontrolled variables is not adequate, then the set of control variables should be modified. Checks should also be performed by comparing multidimensional distributions against corresponding distributions derived from census sources.

2. Vehicle Ownership and Fleet Composition Models

Most travel demand model systems incorporate a vehicle ownership model that takes the form of a discrete choice model such as a multinomial logit or ordered logit/probit model. More recently, a few activity-based travel models are also incorporating multiple discrete-continuous extreme value (MDCEV) models of vehicle fleet composition and primary driver allocation (to each vehicle), so that vehicle type choice can be explicitly modeled. The calibration step would involve comparing model predictions (of distributions of vehicle ownership by socioeconomic or geographic market segment) against travel survey data and making adjustments to alternative specific constants to better replicate the distributions. Validation may be performed by comparing vehicle ownership distributions against external sources such as census data. A vehicle fleet composition model may likewise be calibrated against travel survey data, and validated against an external source such as data from a motor vehicle registration agency that may be able to share aggregate distributions of vehicle makes, models, and vintage in the region or subregion (e.g., zipcode) level. Measures of elasticity can be computed and sensitivity tests may be performed to ensure that the model provides appropriate levels of responsiveness when input conditions change.

3. Activity-Travel-Tour Generation and Daily Schedule/Pattern Models

Most travel demand modeling processes start with an activity, trip, or tour generation process. In traditional four-step travel demand models, cross-classification matrices of trip rates (by market segment) are computed or regression-based models of trip generation are developed using survey data. If regression-based models are adopted, then model calibration may be conducted by comparing predicted trip rates (for different markets) against observed trip rates in the survey data. Intercept terms may be adjusted and model specifications may be enhanced to better replicate observed trip generation rates for different socio-demographic groups. In the case of activity-based models, various schemes may be adopted to represent generation of travel. These may include activity generation, tour generation, or daily activity-pattern models (for different market segments). These models, often taking the form of discrete choice models, may be calibrated against survey data. For example, in a recent model validation exercise for the Baltimore Metropolitan Council (BMC) activity-based model (called InSITE), the model predicted that 41.2% of full-time workers make one work tour with no stops and 33.8% make one work tour (in the day) with intermediate stops. The corresponding values in the

expanded survey data are 42.6% and 31.5% (Cambridge Systematics Inc., 2017). Deviations between survey data and model predictions would necessitate model recalibration by adjusting alternative specific constants and asserting coefficients (as sparingly as possible). If minor adjustments are not able to yield an improved replication of observed behavior, then it is advisable to re-estimate the model with a richer specification and to check the survey data for possible errors, outliers, or illogical responses.

4. Location/Destination Choices

Both traditional four-step travel demand models and more recent activity-based travel models include an array of location and destination choice models. Location models often purport to identify work and school locations (which tend to be longer term location decisions) while destination choice models are intended to determine places visited for various activity and travel purposes (more day-to-day decisions). These models often take the form of multinomial logit models with appropriate size terms to ensure robustness regardless of how zonal systems are configured. In general, these models do not include alternative specific constants, but they may include calibration terms to help the models replicate observed patterns. Work location choice models can be calibrated to replicate work trip length distributions in the travel survey data set and validated against census journey to work flow data. Appropriate shadow price terms can be calibrated to help replicate work flow data and ensure that work trips are reasonably consistent with the number of jobs in each zone. Similarly, school location choice models can be calibrated using school trip length distributions in the survey data and validated against school enrollment data typically available from a Department of Education or similar external source. Destination choice models are generally calibrated such that observed (in survey data) and predicted trip length distributions are replicated for various trip purposes and market segments. Calibration terms may be introduced to help bring about closer matches in observed and predicted trip length distributions. Measures of closeness include the chi-square test to examine the significance of difference between two distributions and the coincidence ratio that measures the degree of overlap of two distributions. The coincidence ratio is formulated as follows:

$$CR = \frac{\sum_{n=1}^N [\min(P_n^m, P_n^o)]}{\sum_{n=1}^N [\max(P_n^m, P_n^o)]}$$

where,

CR = coincidence ratio

N = number of intervals in the distribution (e.g., time or distance bins)

P_n^m = proportion of modeled distribution in interval n

P_n^o = proportion of observed distribution in interval n

A coincidence ratio of 1.0 indicates a perfect match while that closer to zero indicates a poor fit or match of trip length distributions. In general, coincidence ratios above 0.6 may be considered acceptable. When the match is not good, calibration terms can be introduced into the destination choice model to bring about a closer match. The calibration terms may help capture specific markets (e.g., central business district or CBD term) and entertainment districts, better replicate the proportion of intra-zonal trips (by including an intra-zonal variable or modifying intra-zonal travel times up or down for zones where issues exist), and improve the ability to match proportions within specific distance bands.

5. Mode Choice Models

Determining modal splits is often accomplished through the use of multinomial logit or nested logit models of mode choice. These discrete choice models are estimated on travel survey data sets, which are further augmented with transit on-board survey data because regular travel survey data may not include enough transit observations. Mode choice models should be calibrated by comparing against mode shares in a household travel survey data set. There is rarely an external data source that can provide population wide metrics of mode shares. Even if transit ridership counts are available, it is difficult to convert that to a true population mode share unless the total trips are well-established. If the activity-travel generation and destination choice steps are well calibrated, then it is potentially possible to derive transit mode shares and compare model predictions against observed mode shares. Recent availability of passive GPS/smartphone trace data and other location-based service (LBS) data also opens up opportunities for comparing model predicted vehicle trips against population level vehicle trips.

In general, however, most mode choice calibration and validation is done against household travel survey data for various market segments and trip purposes. The alternative specific constants are typically adjusted to better replicate observed mode shares; coefficients associated with explanatory variables should be adjusted only when absolutely necessary and in a behaviorally sound way. One robust way to check the validity of mode choice models is to use the coefficients on cost and travel time variables to compute values of time. These values of time can be checked against published literature to see if they are reasonable and consistent with expectations. Similarly, a ratio of coefficients associated with out-of-vehicle travel time (OVTT) and in-vehicle travel time (IVTT) can shed light on the relative penalty assigned to each of these travel time components. If the ratio is consistent with documented evidence and expectations, then the mode choice model may be appropriate for application. This ratio, for example, is expected to lie between 2.0 and 3.0. Essentially, computing elasticities, conducting mode choice specific sensitivity analysis, and evaluating trade-offs in attributes can serve as powerful mechanisms to validate mode choice models.

6. Time of Day Choice and Activity Scheduling/Timing Models

Travel demand models need to capture the temporal distribution of activity-travel engagement effectively to ensure that peak period travel estimates are accurate and policy decision-making to help relieve congestion is based on sound forecasts of time of day

distributions of travel demand. In traditional four-step travel demand models, simple time-of-day factors may be applied to split trip tables by time of day. Alternatively, time of day choice models may be estimated to compute the splitting proportions for each period of the day. Similarly, in activity-based travel demand models, time of day choice models are often implemented to determine timing and scheduling of activities and trips. In more continuous-time activity-based models (that do not split the day into discrete periods of the day, but rather simulate activities and trips along the continuous time axis while respecting time-space prism constraints), time of day is often an outcome of time use behavior and activity-travel duration decisions on the part of the agent.

Regardless of the exact methodology, the output of a travel model can be analyzed to construct distributions of activity and trip start times, end times, or midpoints. These distributions should be compared against time of day distributions in household travel survey data to determine whether the observed and predicted distributions match closely. The closeness of the match can be measured using chi-square test of the comparison of distributions or the coincidence ratio presented earlier. When the match is deemed less than acceptable, the time of day choice models or the activity and travel duration models can be re-calibrated to obtain a closer fit to the observed distribution. Typically, it should be possible to alter constants for different time periods of the day. In continuous duration models, intercept terms may be calibrated, or model specifications can be enriched to enhance fit. As a last resort, some fine tuning of model coefficients and parameters can be done through assertion. An important consideration is the ability of the model to respond appropriately, and reflect changes in activity-trip scheduling in response to changes in system conditions; for example, the model should be able to capture peak-spreading that may occur when peak period travel conditions (travel times) worsen.

7. Joint Activity-Travel Engagement and Vehicle Occupancy

There is considerable interest in representing intra-household interactions and inter-dependencies effectively in travel demand models. With emerging mobility services emphasizing sharing of vehicles, there is renewed interest in understanding ridesharing and car/van-pooling behaviors and the factors that could enhance shared mode usage. Vehicle occupancy has traditionally been captured in the mode choice modeling step, with explicit choice alternatives dedicated to auto travel with two passengers and 3+ passengers. In traditional travel demand models, the mode choice model captures shared/joint travel and the mode choice model calibration and validation procedures can be used to ensure the model replicates pooling effectively.

In more recent activity-based microsimulation models, joint tours are simulated with a view to represent household members traveling together, either to pursue an activity together or for chauffeuring passengers (e.g., drop-off/pick-up children at school). Household members may also travel separately, but meet at a location to pursue an activity jointly. Capturing these interactions is critical to representing coupling constraints in travel models. Once again, it is difficult to identify external sources of data that can be used to calibrate and validate joint tour engagement models (that generally take the form of discrete choice models or tour frequency/count models). The alternative specific constants and intercept terms would need to be adjusted to ensure a good fit with observed frequency distributions for different tour purposes. The household travel survey data is the best source of data for model calibration and validation; appropriate sensitivity tests can be performed to further verify model validity and reasonableness.

8. Highway Assignment Models

All travel models assign trip tables or trip chains to a model network comprised of links and nodes. Links are characterized by a number of attributes including length, speed, number of lanes, and facility and area type. Nodes may be characterized by turn penalties, intersection delays, and signal control information. Highway assignment results can be analyzed to obtain measures of vehicle miles of travel (VMT) by highway functional classification or facility type. In addition, most highway assignments will output traffic volumes on links by time of day and vehicle class. The traffic volumes can be used to compute link speeds and travel times, as well as origin-destination travel times. Comparisons are generally made against observed traffic counts, speed and travel time data available through detectors or a third party data source, and highway performance monitoring system (HPMS) data that is collected and reported by states on a regular basis.

A common method of validating highway assignment results involves comparing observed link traffic counts against model estimated values. This can be done for various link groups where groups can be defined by functional class, geographic market, area type, or traffic volume interval. For each group, the difference between observed and predicted link volumes can be measured using the Root Mean Square Error (RMSE) or %RMSE. Other measures of difference such as Mean Absolute Error (MAE) may also be used. The RMSE and %RMSE are formulated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N [(Obs_i - Pred_i)^2]}{N}}$$

and

$$\%RMSE = \frac{RMSE}{\left(\sum_{i=1}^N Obs_i / N\right)} \times 100$$

where, Obs_i = observed traffic count on link i

$Pred_i$ = model predicted traffic volume on link i

N = number of links in group under consideration

Another useful measure of the fit between observed and modeled traffic counts is the coefficient of determination (R^2) associated with a scatterplot of observed versus modeled volumes. Scatterplots can be generated for different link groups; such scatterplots not only offer an ability to assess goodness-of-fit of the highway assignment results, but also provide the ability to identify outliers. Many jurisdictions have set criteria or benchmarks that should be met to consider highway assignment results acceptable (Cambridge Systematics Inc., 2010).

In general, the quality of assignment results depends on the ability of the model to accurately represent origin-destination flows in the region (by mode, vehicle class, and time of day) and the quality of the highway network. When poor validation results are obtained, the highway network should be checked for errors. In addition, comparisons of speeds and travel times should be conducted, preferably using third party data sources and detector (traffic management center or TMC) data. By comparing speeds and travel times, it will be possible to identify links and corridors where problems exist and make the necessary corrections. Finally, screenlines (that often represent natural boundaries across a region) or cutlines that cut across multiple facilities that comprise a corridor may be drawn, and the amount of traffic traversing these lines may be compared to ensure that the model is replicating traffic patterns appropriately. Checking volume-to-capacity (V/C) ratios is another useful exercise that can help identify network links where assignment results are not likely to be consistent with reality. Conducting extensive sensitivity analysis should be an element of any highway assignment validation exercise. The speeds or capacities of specific links and nodes should be altered and the change in assignment results for specific links (select link analysis) and the network as a whole should be assessed for reasonableness.

9. Transit Assignment Models

Transit assignment involves assigning transit trip tables by time of day and transit submode to the transit network. The path choice in the transit assignment process is often able to represent transfers involved in executing transit trips between origins and destinations. The path choice may also be used to capture choice between alternative transit submodes that may be serving the same origin-destination pairs, thus obviating the need to include all transit submodes as part of the mode choice model.

The validation of transit assignment results is usually done by comparing observed and model predicted stop-level boardings and alightings and route level ridership (by time of day period). As it may be difficult to replicate boardings and alightings at specific stops within various time periods, the comparisons of boardings and alightings may need to be done at a greater level of aggregation (groups of stops compared for broader time periods). Transit agencies generally have good route-level ridership data, and in many cases, stop-level automated passenger count (APC) data may also be available. In addition to these sources of data, many areas have comprehensive transit on-board survey data. On-board surveys conducted on a periodic basis can help provide much needed data for travel model estimation, calibration, and validation. The expanded and weighted on-board survey data can be used to construct transit trip tables against which model predicted transit trip tables can be compared for various time periods. Before any assignment exercise is conducted, it would be invaluable to ensure that the trip tables represent transit flows accurately.

The quality of transit assignment is highly dependent on getting the transit trip tables right, which requires the destination and mode choice modeling steps to provide accurate results. If the transit assignment results are not consistent with observed ridership counts and on-board survey data, then it would be advisable to revisit the mode choice step for sure and ensure that the transit mode shares and estimates are consistent with expectations and ground truth in the region. In addition, it would be advisable to check the accuracy of the transit network. One particular area that tends to be an approximation is the walk and auto access information associated with each zone. This aspect of travel models is recently gaining higher degrees of fidelity through the use of micro-analysis zones (MAZs) that are much smaller than TAZs. By using a finer zone resolution for transit access representation, it is possible to capture proportions of individuals with walk and auto access more accurately. Finally, any validation exercise should include a sensitivity analysis to see how ridership changes in response to fare, service frequency, and travel time changes. There is empirical evidence documenting how transit ridership changes in response to service changes, and this information can be used as the basis for assessing transit model validity (Evans, 2004).

Sensitivity Analysis of Model Components and Entire Model System

Sensitivity analysis is an important element of any model calibration and validation exercise. Depending on the reasonableness of the results of a sensitivity analysis, it is possible to loop back to any upstream steps in the model development process. It is possible to iterate back to the model specification and estimation step; the model form can be altered, specifications can be enhanced, and coefficients can be re-estimated or re-asserted. It is possible to iterate back to the model calibration step where the model form and specification are retained, but coefficients and parameters are adjusted to offer a level of sensitivity that is more in line with expectations and any available empirical evidence.

Quite often, the terms sensitivity analysis and scenario analysis are used rather interchangeably. For most practical purposes, that is indeed appropriate; by applying the model to a set of alternative scenarios, it will be possible to assess the sensitivity of the model to changes in input conditions. Sometimes, however, a sensitivity analysis refers to an elasticity or marginal effect computation, i.e., the change in dependent variable or choice probability in response to a unit change in an independent variable. Marginal effects and elasticity values serve as measures of sensitivity. When the scenario involves a change in input variable(s) that has a cascading influence through a series of model components, then the sensitivity analysis may be more appropriately badged as a scenario analysis to clearly indicate that the model outputs reflect changes in travel demand captured by the model system as a whole.

Table 1 Scenario descriptions for model sensitivity analysis

Scenario	Description	Changes to base year
Base scenario	2003 is the analysis year	—
25% increase in population and employment densities	Population and employment in the study area are increased by 25%	Population and employment density measures were increased by 25%
100% increase in cost—drive alone mode	A 100% increase in cost for drive-alone for all time periods	Cost tables were altered by multiplying the drive alone auto cost by two in the AM, PM, and off-peak files
25% increase in IVTT— auto mode for all periods	A 25% increase in IVTT for the drive alone and shared ride for all time periods	IVTT tables were altered by multiplying the auto IVTT by 1.25 in all time periods
25% increase in IVTT— auto mode and peak periods	A 25% increase in IVTT for the drive alone and shared ride for the AM and PM peak time periods	IVTT tables were altered by multiplying the auto IVTT by 1.25 in the AM and PM peak files

Scenarios should be constructed such that the outputs can be assessed for their reasonableness and potential accuracy (through comparisons with available empirical evidence). In some cases, however, it may be useful to check extreme (potentially unrealistic) scenarios so that the limits of the model system are tested and the analyst can determine the boundary conditions within which the model is able to offer plausible and defensible forecasts. In the context of future transportation technologies, this becomes more involved as there is considerable uncertainty on how these technologies will evolve and find their way into the transportation marketplace.

An illustration of a scenario analysis is furnished in this section, in the context of a first-generation activity-based microsimulation model of travel demand developed and applied for the Southern California region several years ago (Goulias et al., 2011). The initial model development effort involved synthesizing and simulating the activity-travel patterns of about 18 million people in just over 4000 traffic analysis zones (TAZs). Following the model development process, there was a desire to determine how systemwide changes in cost and travel time would impact those below and above the poverty line. Activity-based microsimulation models facilitate comparisons across socio-economic and demographic groups, thus enabling equity analysis of alternative transportation policies.

Table 1 presents the system-wide scenarios that were tested using the model. The base year was 2003, and the scenarios essentially constituted a 25% increase in population and employment densities, a doubling in the cost of drive alone mode, a 25% increase in in-vehicle travel time (IVTT) for the auto mode in all time periods of the day, and a 25% increase in IVTT for the auto mode during peak periods. These scenarios were applied in isolation (not as a combination of changes). After applying the model system, changes are predicted in all aspects of activity-travel patterns. Activity-trip frequencies, destination locations, and mode choices change; scheduling and chaining of activities and trips is altered; and joint activity-travel engagement with other members of the household is altered as well.

In the interest of brevity, illustrative results are presented for an aggregate and a disaggregate sensitivity analysis. **Table 2** presents the changes in average trip lengths for different trip types. The three trip types are home-based work (HBW), home-based other (HBO), and non home-based (NHB). In all cases, it is expected that average trip lengths will decrease. An increase in cost or time (impedance measures or travel skims) will have the effect of reducing trip lengths. Similarly, an increase in employment and population density will also engender a shift towards shorter trips. It can be seen that the results in **Table 2** are consistent with these expectations and reflect shifts in trip lengths that are behaviorally intuitive. In general, work trips are largely inflexible and as a consequence, the trip lengths for HBW show the smallest percent change. On the other hand, HBO trips exhibit the greatest level of change in trip lengths, consistent with the notion that these are mostly personal business, social-recreational, and shopping trips.

Table 2 Changes in average trip lengths by trip type and poverty group

Trip type	Base case Avg (mi)	DA cost increase 100% Avg (mi)	% Reduction	DA IVTT increase 25% Avg (mi)	% Reduction	Peak period DA IVTT increase 25% Avg (mi)	% Reduction	25% Increase in empl & pop densities Avg (mi)	% Reduction
Below poverty									
HBW	11.17	10.21	8.6%	10.68	4.3%	10.67	4.5%	10.70	4.2%
HBO	10.52	6.91	34.3%	5.48	47.9%	6.37	39.5%	6.95	33.9%
NHB	7.49	7.06	5.8%	5.74	23.3%	6.58	12.2%	7.12	5.0%
Total	9.69	7.28	24.9%	6.11	37.0%	6.87	29.1%	7.40	23.7%
Above poverty									
HBW	11.19	10.53	5.9%	10.95	2.1%	10.92	2.5%	10.99	1.8%
HBO	10.15	8.37	17.5%	6.72	33.8%	7.77	23.5%	8.42	17.1%
NHB	7.99	7.71	3.5%	6.36	20.4%	7.27	8.9%	7.84	1.9%
Total	9.68	8.52	11.9%	7.39	23.6%	8.16	15.7%	8.72	9.9%

Note: DA, drive alone; HBW, home-based work; HBO, home-based other; IVTT, in-vehicle travel time; NHB, non home-based

Table 3 Illustration of impact of drive alone cost increase on activity-travel pattern of an individual

<i>Base case</i>		
Overall pattern		Commute and additional tour
Total miles		47.17
	Home → Work commute	
Number of non-work stops		0
Mode		Drive alone
Activity at non-work stops		—
	Work → Home commute	
Number of non-work stops		0
Mode		Drive alone
Activity at non-work stops		—
	Tour 1	
Number of stops		1
Mode		Drive alone
Activity at stops		Eat out
<i>Policy case (Doubling of drive alone cost)</i>		
Overall pattern		Commute tour
Total miles		35.21
	Home → Work commute	
Number of non-work stops		0
Mode		Drive alone
Activity at non-work stops		—
	Work → Home commute	
Number of non-work stops		1
Mode		Drive alone
Activity at non-work stops		Eat out

Finally, NHB trips show rather modest changes in trip lengths as well; these are likely to be work-based trips (say, during the lunch hour) or intermediate trips within a longer trip chain. Indeed, the percent change is smaller than that for HBO trips and more than that for HBW trips (except for the case where drive alone cost doubles). Since these trips are relatively shorter to begin with, it is reasonable to find that NHB trip lengths show lower sensitivity to cost than to travel time. In all instances, the percent change is larger for the “below poverty” market segment when compared with the “above poverty” market segment. This is behaviorally intuitive as one would expect those below the poverty line to be more adversely impacted and sensitive to changes in impedance (cost and travel time).

Finally, **Table 3** presents an illustration of the change in activity-travel behavior for an individual agent in the population, thus representing a very disaggregate validation exercise. By examining individuals in the synthetic population, it is possible to see how their activity-travel patterns changed from one scenario to another. This Table shows what happens to an individual when the drive alone cost doubles.

It is found that a doubling of the cost results in the individual adopting a more efficient activity-travel pattern. In the base case, the individual undertakes a commute tour (Home→Work→Home) with no intermediate stops on either journey to or from work. The individual undertakes a separate home-based tour to eat out, resulting in a total distance traveled of 47 miles. In the policy case, when drive alone costs are doubled, the individual makes a stop to eat out on the way back home from work, presumably gaining efficiencies in distance (and hence cost). This modification of the activity-travel pattern results in the individual traveling only 35 miles in the policy case. This change in mileage and the change in activity-travel pattern are behaviorally intuitive, suggesting that the model is able to predict changes in travel demand in response to changes in system conditions without necessarily violating basic rules and constraints. The decrease in mileage (of about 12 miles) may be checked against historical evidence of how person or vehicle miles of travel have changed in response to changes in driving cost. In general, it is known that travel mileage is quite insensitive to driving costs (Brons et al., 2008), and so a change of this magnitude can only be realized with a dramatic increase (doubling) in travel cost.

Summary

Model calibration and validation is a very critical step in the model development process. Unless a model has been appropriately calibrated and verified to be “valid,” capable of providing reasonable and useful forecasts for transportation planning and policy making, it should not be adopted. To the extent possible, model validation should be done at a fine-grained level, checking model

outputs separately for various market segments, geographic subareas, network corridors and links, and times of day. In this way, it is possible to ensure that the model is capable of reflecting the heterogeneity that often characterizes activity-travel demand in a region.

Given the importance of these steps, it is critical that enough resources and time be dedicated to these tasks in the model development process. Quite often, resources and time are spent developing the model specification and estimating models, leaving very little to go through extensive calibration and validation exercises. Recognizing that calibration and validation is at least as important, if not more than, model estimation is critical to ensuring that appropriate resources and time are allocated to these efforts. Calibration and validation should be done for each model component and for the model system as a whole. Validation data should be assembled with care and checked extensively for accuracy to ensure their appropriateness as a basis to assess model performance.

Finally, model calibration and validation exercises should be fully documented. Documentation of these efforts can help model users understand how the models came to be specified and structured the way they are, how the models are expected to respond to changes in input variables, and appropriate limits beyond which model capabilities may be stretched in ways that are not fully valid or adequately tested. The model calibration and validation documentation should present scenarios tested, predictions obtained, and an assessment of the reasonableness and potential value of the forecasts. Metrics used to assess validity should be documented for each model component and the model as a whole. Care should be exercised when applying models to new and emerging transportation and mobility scenarios featuring technologies and services that do not currently exist anywhere in the real-world. Given the large degree of uncertainty as to how emerging transportation technologies and shared mobility services may evolve in the years ahead, forecasts are heavily dependent on underlying assumptions of technology capabilities and features and market behaviors. Thus, any and all forecasts should be accompanied with clear documentation of underlying assumptions so that decision-makers can assess the validity and usefulness of the model outputs.

An exercise that the profession should increasingly adopt as part of standard model calibration and validation efforts is that of backcasting or forecasting from the base year to the current year. The base year is often a year that is a few years in the past (when compared to the year in which the model development is completed). By applying the model to forecast to the present year, it will be possible to assess the ability of the model to replicate ground truth conditions. Similarly, the model can be applied to backcast to a past year by using the socio-economic and network inputs of the past year. Because ground truth conditions are known for the prior year, it will be possible to assess the ability of the model to replicate real-world traffic patterns when absolutely correct inputs are fed into the model (any error must be largely due to model specification errors). Although time intensive, exercises that involve backcasting and forecasting can prove invaluable in determining validity of travel models and identifying model components that need to be re-calibrated to better replicate real-world conditions.

References

- Brons, M., Nijkamp, P., Pels, E., Rietveld, P., 2008. A meta-analysis of the price elasticity of gasoline demand: A SUR approach. *Ener. Econ.* 30, 2105–2122.
- Cambridge Systematics, Inc., 2017. Baltimore Metropolitan Council In: SITE Activity Based Travel Model: Model Validation Report. Baltimore Metropolitan Council, Baltimore, MD.
- Cambridge Systematics, Inc., 2010. Travel Model Validation and Reasonableness Checking Manual. Second Edition. Federal Highway Administration, US Department of Transportation, Washington, DC.
- Evans, J.E., 2004. Traveler Response to System Changes: Chapter 9 – Transit Scheduling and Frequency. TCRP Report 95, Transit Cooperative Research Program, Transportation Research Board, Washington, DC.
- Goulias, K.G., Bhat, C.R., Pendyala, R.M., Chen, Y., Paleti, R., Konduri, K.C., et al., 2011. Simulator of Activities, Greenhouse Emissions, Networks, and Travel (SimAGENT) in Southern California: Design, Implementation, Preliminary Findings, and Integration Plans. *IEEE Forum on Integrated and Sustainable Transportation Systems*, Vienna, pp. 164–169, doi: 10.1109/FISTS.2011.597362.
- Milam, R.T., Birnbaum, M., Ganson, C., Handy, S., Walters, J., 2017. Closing the induced vehicle travel gap between research and practice. *Transp. Res. Rec.* 2653 (1), 10–16, doi:10.3141/2653-02.

Further Reading

- Cambridge Systematics Inc., 2008. FSUTMS-Cube Framework Phase II: Model Calibration and Validation Standards. Florida Department of Transportation, Systems Planning Office, Tallahassee, Florida.
- Childress, S., 2015. SoundCast Calibration and Sensitivity Test Results. Puget Sound Regional Council, Seattle, WA.
- COMPASS, 2017 Regional Travel Demand Forecast Model Calibration and Validation Report for Ada and Canyon County, Idaho. Report Number 06-2017. Community Planning Association of Southwest Idaho, Idaho.
- Erhardt, G., et al., 2020. Traffic Forecasting Accuracy Assessment Research. NCHRP Report 934, National Cooperative Highway Research Program, Transportation Research Board, Washington, DC.
- Huntsinger, L.F., 2017. The Lure of Big Data: Evaluating the Efficacy of Mobile Phone Data for Travel Model Validation. 96th Annual Meeting of the Transportation Research Board, Washington, DC.
- Proussaloglou, K., Popuri, Y., Tempesta, D., Kasturirangan, K., and Cipra, D., 2007. Wisconsin passenger and freight statewide model: Case study in statewide model validation. *Transp. Res. Rec.* 2003: 120–129.
- Roorda, M., Miller, E.J., Habib, K.N., 2008. Validation of TASHA: A 24-h activity scheduling microsimulation model. *Transp. Res. Part A* 42 (2), 360–375.
- San Diego Association of Governments, 2016. Activity-Based Travel Model Calibration and Validation for Base Year 2012. San Diego Association of Governments, San Diego, CA.
- Wegmann, F. and Everett, J., 2008. Minimum Travel Demand Model Calibration and Validation Guidelines for State of Tennessee. Center for Transportation Research, University of Tennessee, Knoxville, TN.
- WSP, 2017. Activity-Based Model Calibration Report: Coordinated Travel – Regional Activity Based Modeling Platform (CT-RAMP). Atlanta Regional Commission, Atlanta, GA.

Trip Chaining Analysis

Cynthia Chen*, Yusak Susilo†, *University of Washington, Seattle, WA, United States; †Institute for Transport Studies, University of Natural Resources and Life Sciences, Vienna, Vienna, Austria

© 2021 Elsevier Ltd. All rights reserved.

Definition	606
From Trips to Trip Chains	606
Trip Chaining in Relation to Congestion, Emissions, and Public Health	606
The Decision Theory and Models Explaining Trip Chaining	607
Factors Affecting Trip Chaining	608
Trip Chaining, Transit Use, and Emerging Mobility	609
Trip Chaining Analysis in an Era of Big Data	609
Concluding Thoughts	609
Acknowledgments	610
Biographies	610
References	610

Definition

A trip is simply travel from one place to another by any means of travel. A trip chain is a series of trips chained together. Although there is no formal definition on the definition of a trip chain, there are common terms and expectations to describe whether a series of trips should be considered part of a chain (Primerano et al., 2008). In many cases, a trip chain comprises a series of trip that start and end at one's anchor locations, which may or may not be same. An anchor location is one that is not only regularly visited but also is a key (anchor) point for other trips conducted in a day. Home and work places are typically considered as one's anchor locations because other trips are often made around them. A related concept is tour, which is also defined as a series of trips chained together. The key difference between a trip chain and a tour is that the former is a general reference to a series of trips chained together and the latter is a specific reference for trip chains that start and end at the same anchor location. Common tours are home-based or work-based tours, referring to a series of trips that start and end at home or work, respectively.

Fig. 1 illustrates the differences between trip, trip chaining, and tour. In this example, the individual drives from home in the morning to a Park and Ride location, parks her car, gets on a bus and arrives at work. At noon, she walks to a restaurant for lunch and then walks back to office. In the afternoon, she carpool with colleagues to another restaurant for dinner. Afterwards, her colleague drops her off at the Park and Ride location where she picks up her own car and drives home. Here, there are five separate trips: home to work, work to lunch, back to work from lunch, work to restaurant, and restaurant to home. There are two tours: one home-based tour capturing the entire daily mobility pattern from home in the morning and then back home in the evening and the other work-based tour for two trips at noon: from work to lunch and then back to work. Both of these tours can also be viewed as trip chains with the same origin and destination. Additionally, there are two other trip chains: the first is from home to Park and Ride location and then to work in the morning and the second one is from work to restaurant for dinner and then back home in the evening. In both cases, the start and end points are different and both also involve a stop in a middle.

From Trips to Trip Chains

It is a landmark recognition in the travel behavior research (an area that studies how individuals make travel-related decisions such as mode choices, route choices, and departure time choices, etc.) that trips shall not be studied independently but together (Adler and Ben-Akiva, 1979; Kitamura, 1984). This recognition is a stark departure from decades' running four-step models that has treated trips independently. The key reason for this shift is the recognition of history and future dependence between trips. The fact that the car is parked at a Park and Ride location prevents one from driving to the restaurant for dinner in the evening (history dependence) and the expectation that driving to and parking at the restaurant in downtown in the evening makes one decide to drive to a Park and Ride location and take the bus to work in the morning (future dependence).

Trip Chaining in Relation to Congestion, Emissions, and Public Health

Transportation researchers have longed recognized the role that trip chaining plays in societal phenomena that affect everyone's welfare, such as congestion, emissions, and public health. The overwhelming reliance on automobiles in the United States is viewed

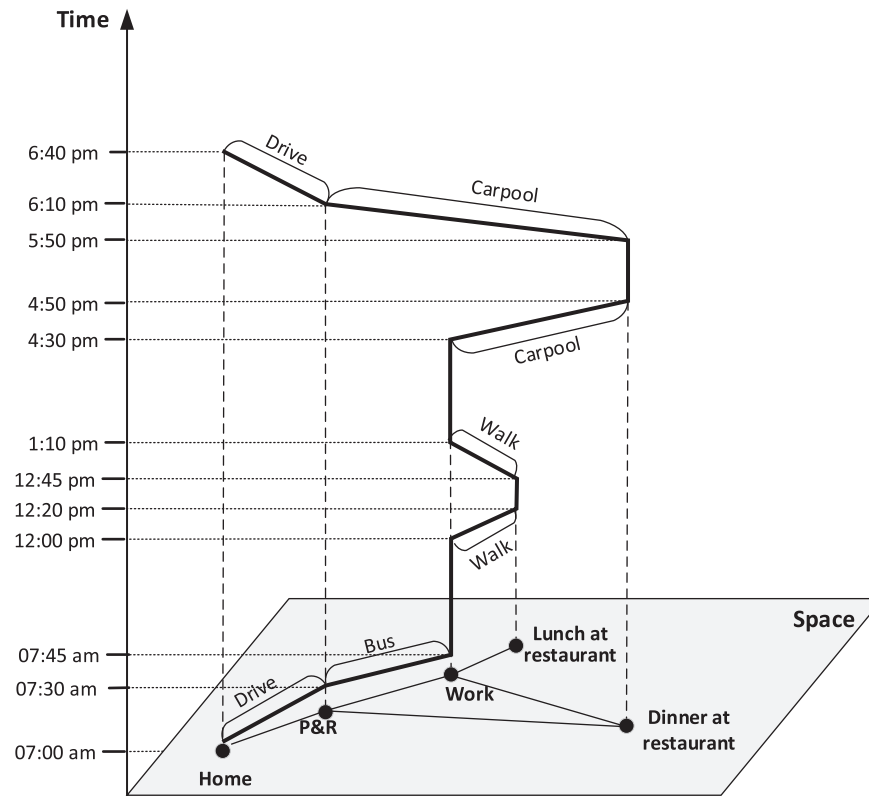


Figure 1 A 3D illustration of one individual's mobility pattern in a day.

as a significant cause behind widespread and increasing congestion in many metropolitan regions, declining transit ridership, CO₂ emissions, and prevalence of obesity. Automobiles still surpass other modes in its ability of chaining trips to accomplish desired activities that are often distributed in space. In other words, while reducing America's reliance on automobiles represents potentially the most effective strategy in reducing congestion and CO₂ emissions and improving public health, it is also the most difficult one. There are significant barriers in the use of alternative modes (other than autos) in completing the trip chains. Take the example of transit use: the most critical barrier relates to the first and the last mile of a trip'the difficulty in accessing a transit station from an origin and reaching a destination from a transit station prevents many from using transit.

The Decision Theory and Models Explaining Trip Chaining

The dominant decision theory underlying trip chaining is random utility maximization, where an individual is assumed to be a rational decision maker. When faced with a set of alternatives, she is to evaluate and trade-off between the various decision factors and select one that maximizes her satisfaction (utility). Since trips are primarily made to conduct activities at the destinations, the utilities are expressed in negative terms and the goal is then to minimize one's disutility.

The classic model based on random utility maximization for explaining trip chaining is the discrete choice model where each alternative is discrete and finite and the decision maker is said to choose the one that maximizes her utility (Ben-Akiva and Lerman, 1985). Each alternative has a utility function that explains how different factors may come together to approximate the amount of (dis)satisfaction that may incur if that alternative is chosen. Inherent in the utility function is the trade-off nature of the decision-making process, meaning that one unattractive feature, for example, long travel time, can be compensated by an attractive feature, for example, being inexpensive. Because of this, the utility function is often expressed as a weighted sum of various factors, with the weights being the decision maker's preferences, or values toward different factors.

Because more than one decision is involved in a trip chain, complex discrete choice models have been developed. There have been two modeling approaches: simultaneous (Bowman and Ben-Akiva, 2001) vs. sequential (Kitamura et al., 2000). The simultaneous approach assumes that prior to the start of a trip chain, one develops a few alternatives regarding its general pattern composition, for example, a trip chain with one stop or two stops. The individual will first decide on which pattern is to be adopted, followed by decisions that need to be made within a particular pattern. For example, one may first select a trip chain with one stop, and then proceed to decide on the mode choices to and from the stop within the chain. In modeling, the simultaneous approach is often formulated as a hierarchical model with multiple decisions involved and the estimation of these multiple decisions is

simultaneous. The sequential approach assumes that individuals are myopic and do not first develop a general idea of how a trip chain may be conducted. Instead, decisions are made in a sequential way, one trip chained after another. In modeling, the sequential approach is often formulated by a sequence of discrete choice models, conditioned one after another.

Factors Affecting Trip Chaining

Three categories of factors have been consistently identified to affect trip making: structural (alternative related) factors, the built environment, and socio-demographic attributes. Structural factors directly affect the (un)attractiveness of an alternative. The most common structural factors are travel time and travel cost. There are also alternative-specific factors that only matter to certain alternatives, for example, parking availability and cost associated with driving; waiting time, number of transfers and distance to transit associated with transit; availability of sidewalk for walking; and elevation for biking. The structural factors have long been considered as most important factors explaining trip making behaviors, given that individuals are rational decision makers. Though many behavioral nuances have been found to suggest many irrational aspects in individuals' decision-making processes, the assumption that in trip making, individuals are mostly rational is still considered valid.

The built environment determines the availability of options and their levels of accessibility for trip making. For example, if there is not a transit stop within a close range (e.g., 0.5 mile), transit cannot be considered a viable option. In the context of trip chaining, the availability and accessibility of a particular mode choice (e.g., transit) for any part of a trip chain could be a deciding factor for selecting a mode for the entire trip chain. Temporally, the availability of an opportunity at a destination can both dictate the departure times associated with the trips in a chain and the mode choices used. The opportunity and accessibility effects afforded by the built environment have been measured with metrics in five D categories (Ewing and Cervero, 2010): density, diversity, design, destination accessibility, and distance to transit. These effects are not uniform across all: for example, research has found that the built environment affects the travel behavior of men and women differently.

Socio-demographic factors include both one's personal attributes (e.g., age, gender, and physical disability) and demographic characteristics (e.g., being a part of a household, a neighborhood, a company, or a community). These factors often shape one's responsibilities and constraints at work, in families, and in society. They also reflect long-standing traditions, culture, and practices in a society. As an example, the obligation to pick up children prevents one from taking transit and dictates her to depart from work by a certain time. These factors (personal and social) are often intertwined together. Women are found to have more complex trip chaining patterns in both time and space due to their continuing primary care responsibilities in the household, despite a significant rise in women's participation in the workforce (Hanson, 2010; Kwan, 1999). Interestingly, women's higher shares of family obligations are often coupled with their lower mobility levels, as reflected in fewer access to vehicles and drivers' licenses. Previous studies (Susilo et al., 2019) show that low mobility levels and time-space constraints, such as childcare obligation, are two key factors explaining why women tend to work closer to home and use more public transport, compared with their male counterparts.

These three categories of factors can also be understood in Hägerstrand's space-time prism framework (Hägerstrand, 1970), which illustrates various elements in a trip chain in a 3D time-space comprising a vertical time dimension and a two-dimensional physical space (see Fig. 1). The constraints one faces in completing the trip chaining are of three kinds: (1) capacity constraints due to our biological limits (e.g., one must allocate time for eating and sleeping) and physical constraints (e.g., one cannot run faster than a vehicle); (2) coupling constraints that arise from limits in one's resources (e.g., lack of a car) or from the built environment (e.g., lack of public transportation nearby); and (3) authority constraints due to rules in a society (e.g., one shall not drive faster than the speed limit or stores are not open at night).

Aside from the above three categories of factors, there are other factors that affect trip chaining. Weather, for example, has been found an important factor in use of nonmotorized modes such as walking and biking: cold and rain are severe barriers for biking and walking. Attitudinal factors are also found to be important: their roles in affecting trip making behaviors are complex (Wang and Chen, 2012)—interacting with other factors, as latent factors forming a hierarchical decision structure, and acting as nonlinear thresholds that could distort the trade-off nature assumed in the classic random utility theory.

Additionally, there is increasing recognition that factors that often suggest behavioral irrationality can also play a role in explaining trip making behavior. This is where a number of other prominent behavioral and social and psychological theories coming into play. Examples include prospect theory (Kahneman and Tversky, 1979) and peer influence. Prospect theory identifies three aspects in explaining behavior: (1) framing effect: a scenario stated positively vs. negatively results in different behavioral choices; (2) risk seeking vs. avoidance: risk seeking is associated with loss and risk avoidance is with gain; and (3) when faced with uncertain scenarios, people attribute excessive weights to low probability events and insufficient weights to high-probability events. Peer influence refers to the decisions made to conform with those made by peers, even though the choice itself may derive lower utility for the decision maker. In a modern society characterized by ubiquitous access to the internet (and thus information) and prevalence of mobile devices, it is expected that these factors will play an increasingly important role in trip making behaviors.

It is important to note that all factors including the trip chaining behaviors evolve over time and this is due to changes at all levels, individual, family, community, and society. For example, recent studies (Susilo et al., 2019) have shown that older generations of partnered/married young women have fewer trips per work trip chain than their male counterparts, but this gender difference is becoming smaller among younger generations. The declining gender difference is mainly due to changes in gender effects in the number of trips and activity duration through indirect effects. The women from younger generations, including those who have children, are becoming more active in out-of-home non-work activities and having more complex trip chaining, compared to their male counterparts.

Trip Chaining, Transit Use, and Emerging Mobility

As noted earlier, trip chaining plays a critical role in transit use: significantly increasing transit use requires solving the first and last mile problem in trip chaining. In the past, access to transit is limited to those who live within a short distance (e.g., 0.5 mile) to the station and those who do not live within walking distance to a transit station but can drive and park near a station. In the United States, the former are largely in a few densely populated urban areas and the latter are primarily in suburban area whose primary use of transit is for commuting purposes. Together, these two populations account for a small percentage in the nation, resulting in a low share of transit use in most of the metropolitan areas except a few large cities (e.g., New York City and San Francisco).

The rise of emerging mobility services such as Uber and Lyft has aroused new worries and hope for transit use and both are related to trip chaining. The nearly ubiquitous presence of those new services in urban areas has prompted multiple studies that seek to answer related questions such as “whether the new mobility services are generating new trips, and/or attracting trips from solo-driving (which is encouraging) or from transit (which is discouraging)?” and “to what extent do these new mobility services contribute to congestion?” Today there seems to be substantial evidence from multiple studies in different locales pointing to the contribution of the new mobility services to congestion as well as their role in attracting transit users and thus declining transit ridership.

There is also new hope that the emerging mobility services may potentially boost transit ridership by providing an efficient and cost-effective means for the first- and the last-mile. Recent years have seen a number of cities (e.g., NYC, Seattle, San Diego, DC, and LA, etc.) that have or are experimenting with using new mobility services to solve the first- and last-mile problem (Marshall, 2019). The various types of experiments include, for example, providing free or discounted rides for economically disadvantaged riders or those within certain geographic areas, providing discounts (absolute \$ amount or in percentage terms) for rides to/from certain stations or for particular corridors, and integration of Uber/Lyft services in transit mobile apps. Some agencies (e.g., Livermore-Amador Valley transit and Arlington County) even experimented with replacing the traditional dial-a-ride and certain low-demand fixed-route services with new mobility services. Since this is a very recent phenomenon, there have not been conclusive results pointing in either direction. Some very early experiments (e.g., Kansas City, Missouri and Centennial, Colorado) recently ended after disappointing results, but they may be attributed to reasons such as lack of marketing effort for raising the awareness.

Trip Chaining Analysis in an Era of Big Data

Analysis of trip chaining gained its prominence after the landmark recognition of its important role in travel behavior analysis. Today, activity-based models in place of the traditional four-step models for travel demand forecasting directly recognize the importance of trip chaining and build individuals' daily mobility patterns as a series of trips chained together. These models are typically built based on household travel survey data, which contains information on all trips conducted during 24 h for a small but representative sample of a region's population. In the household travel survey data, how a series of trips are chained together within 24 h is directly observed through either self-reporting or a combination of active GPS logging at regular intervals (e.g., every 10 s) and confirmation by survey participants.

The last 10 years have also witnessed a surge of studies using passively generated, big data (e.g., phone data, app-based data, social media data) for deriving individuals' mobility patterns (Chen et al., 2016). Passively generated, because such data is generated as a by-product of attaining primary purposes such as billing, operation, and maintenance. Trips and how a series of trips are chained together cannot be directly observed from such data; they can only be inferred. It is thus apparent to see that since the data is not generated to capture one's trip making behavior throughout a day like the household travel survey does, it is possible or quite likely that certain trips within a chain will be missed. In fact, under-estimation of trips has been observed in a number of studies using such data. It is however also worthy to note that the emphasis of trip chaining seems to have diminished in studies using big, passively generated data to derive mobility patterns. This is reflected in three aspects: (1) trips are the dominant unit of interest in vast majority of the studies as opposed to trip chains; (2) little understanding of what types of trips tend to be missed in trip chains and how they affect our ability in deriving mobility patterns; and (3) there is also little work that attempts to recover missed trips by leveraging the history and future dependences between trips in a chain.

Concluding Thoughts

It is the trip chains (a series of interdependent trips chained together), not independent trips, that underlie our daily mobility patterns. Therefore, studies on individuals' trip making behavior or mobility patterns shall recognize the linkages between trips. This notion is important in all inquiries, whether it is to understanding individuals' travel behaviors, predicting them, or seeking to change them. It shall also remain important no matter what type of data is used, actively generated data such as household travel surveys or passively generated big data (e.g., phone data, app-based data, or social media data).

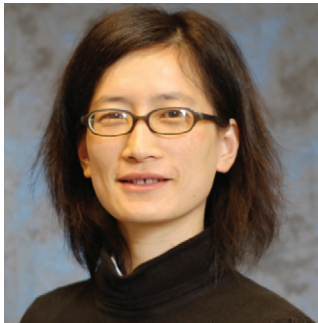
As noted earlier, individuals' trip chaining behaviors directly contribute positively or negatively to several phenomena of interest at the society level, such as congestion, emissions, and obesity. The ability to trip chain transit trips with other modes together will potentially boost transit ridership, reduce emissions, and obesity (by increasing physically active travels); on the other hand, the inability to chain transit trips with other modes loses ground to automobiles, thus directly adding to congestion, emissions and public health crisis including obesity. From this perspective, how to help individuals achieve effective and sustainable trip chaining

may be an important opening for potential behavioral changes. Recent studies (Duncan, 2016) have shown that trip chains (in particular, work-based tours) have the potential of reducing VMT by 20%–50%, when compared to multiple trips with separate destinations. This suggests that offering recommendations to people on how to schedule and reschedule their trips together may be a viable option for sustainability purposes.

Acknowledgments

Funding for this research is provided by a NSF (National Science Foundation) grant (CMMI 1536340) and a NIH (National Institute of Health) grant (1R01GM108731-01A1) to Cynthia Chen. The authors thank the help of Feilong Wang, a PhD student in the Department of Civil and Environmental Engineering, University of Washington (Seattle), for his help in creating the figure.

Biographies



Cynthia Chen is a professor in the Department of Civil and Environmental Engineering at the University of Washington, Seattle. She received her doctoral degree from University of California, Davis. Her main research interests are understanding human mobility patterns with big and small data, cascading processes on networks, and intervention designs for behavioral modification.



Yusak Susilo is a BMVIT Endowed Professor in Digitalisation and Automation in Transport and Mobility Systems at the University of Natural Resources and Life Sciences (BOKU), Vienna, Austria. He received his doctoral degree from Kyoto University, Japan. His main research interest lies in the intersection between transport and urban planning, transport policy, decision-making processes, and behavioral interactions modeling.

References

- Adler, T., Ben-Akiva, M., 1979. A theoretical and empirical model of trip chaining behavior. *Transp. Res. B: Methodol.* 13, 243–257, doi:10.1016/0191-2615(79)90016-X.
- Ben-Akiva, M., Lerman, S., 1985. *Discrete Choice Analysis: Theory and Applications to Travel Demand*. MIT Press.
- Bowman, J.L., Ben-Akiva, M.E., 2001. Activity-based disaggregate travel demand model system with activity schedules. *Transp. Res. A: Policy Pract.* 35, 1–28, doi:10.1016/S0965-8564(99)00043-9.
- Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transp. Res. C: Emerg. Technol.* 68, 285–299, doi:10.1016/j.trc.2016.04.005.
- Duncan, M., 2016. How much can trip chaining reduce VMT? A simplified method. *Transportation* 43, 643–659, doi:10.1007/s11116-015-9610-5.
- Ewing, R., Cervero, R., 2010. Travel and the built environment. *J. Am. Plan. Assoc.* 76, 265–294, doi:10.1080/01944361003766766.
- Hägerstrand, T., 1970. What about people in Regional Science? *Pap. Reg. Sci. Assoc.* 24, 6–21, doi:10.1007/BF01936872.
- Hanson, S., 2010. Gender and mobility: new approaches for informing sustainability. *Gend. Place Cult.* 17, 5–23, doi:10.1080/09663690903498225.
- Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47, 263–291, doi:10.2307/1914185.
- Kitamura, R., 1984. Incorporating trip chaining into analysis of destination choice. *Transp. Res. B: Methodol.* 18, 67–81, doi:10.1016/0191-2615(84)90007-9.
- Kitamura, R., Chen, C., Pendyala, R.M., Narayanan, R., 2000. Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation* 27, 25–51, doi:10.1023/A:1005259324588.
- Kwan, M.-P., 1999. Gender, the home-work link, and space-time patterns of nonemployment activities. *Econ. Geogr.* 75, 370–394, doi:10.1111/j.1944-8287.1999.tb00126.x.
- Marshall, A., 2019. Transit agencies turn to Uber for the last mile. *Wired*.

- Primerano, F., Taylor, M.A.P., Pitaksringkarn, L., Tisato, P., 2008. Defining and understanding trip chaining behaviour. *Transportation* 35, 55–72, doi:10.1007/s11116-007-9134-8.
- Susilo, Y.O., Liu, C., Börjesson, M., 2019. The changes of activity-travel participation across gender, life-cycle, and generations in Sweden over 30 years. *Transportation* 46, 793–818, doi:10.1007/s11116-018-9868-5.
- Wang, T., Chen, C., 2012. Attitudes, mode switching behavior, and the built environment: a longitudinal study in the Puget Sound Region. *Transp. Res. A: Policy Pract.* 46, 1594–1607, doi:10.1016/j.tra.2012.08.001.

Vehicle Ownership Models

Dr. Sören Groth*, **Prof. Dr. Dirk Wittowsky[†]**, ^{*}ILS—Research Institute for Regional and Urban Development gGmbH, Büberweg, Dortmund, Germany; [†]University of Duisburg-Essen, Faculty of Engineering/Institute for Mobility and Urban Planning, Universitätsstraße, Essen, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	612
Aggregated Vehicle Ownership Models	612
Disaggregated Vehicle Ownership Models	613
Objective Factors: The Example of Life Stages	614
Subjective Factors: The Example of Mode-Related Psychologies	614
Blurred Lines Between Objective and Subjective Factors: The Role of Built Environment in Ownership Decisions	615
New Challenges: New Technologies, New Transport Modes, No Ownership Anymore	615
References	616
Further Reading	616

Introduction

Vehicle ownership represents a material precondition for transport-related participation in society. In this sense, vehicle ownership both enables and restricts the realization of spatially differentiated activities. With regard to the individual subject, vehicle ownership has a high behavioral relevance, which is often described as a commitment between vehicle ownership and use (Simma and Axhausen, 2003). The decision for specific vehicle ownership has long-term and short-term consequences with regard to individual travel behaviors. The long-term nature of the consequences is reflected, for example, in certain Modality Styles, that is, behavioral predispositions characterized by a certain mode or set of modes that an individual habitually uses and which are expressed in long-term anticipated spatial activity patterns of the traveler. For example, (monomodal) car users organize their locations and all associated distances around their home, workplace, leisure activities, etc. from behind the steering wheel of their car. The short-term consequences of owning a vehicle are equally effective in the situational moments of everyday activities, that is, spontaneous or planned activities including frequency, duration, geographical location, etc.

Over the past decades, a complex and heterogeneous variety of vehicle ownership models has developed that can be used. Many model approaches are unknown, for example, due to the fact that the private sector generally operates within the framework of trade secrets. Vehicle ownership models are used for different purposes in all sectoral areas (politics, business, science, nongovernmental organizations, etc.). What all these approaches have in common is that they seek to explain and/or understand vehicle ownership in some way, in order to draw appropriate conclusions. Governments and administrations at the political-administrative level, for example, use vehicle ownership models to simulate traffic demands, calculate tax revenues, or also estimate climate-impacting emissions and draw up action plans. The manufacturers of specific transport modes from the private sector in turn use vehicle ownership models to predict the demand for specific vehicle types — for example, for specific car brands or bicycle types — and thereby, profit expectations. For transport and mobility research, the focus is on concrete epistemological approaches to understanding and explaining ownership patterns and the associated consequences.

Many methodologically sophisticated vehicle ownership models can be located in the context of the global north. Since these countries are largely characterized by automobile societies, modeling often occurs according to car ownership. With the growing questioning of the automobile, however, a number of other focal points have emerged in recent years, such as the trends toward multimodality or new transport systems. Therefore, the terminology used differs across studies with regard to different questions and the interdisciplinary composition of the actors. For example, in addition to vehicle ownership, other terms too, such as “mobility resource holdings” (Le Vine et al., 2013) or “mobility tool ownership” (Scott and Axhausen, 2006) are used.

Aggregated Vehicle Ownership Models

The different vehicle ownership models are distinguished according to aggregated and disaggregated approaches. Aggregated models are based in different ways on general social structural data for modeling vehicle ownership. The application of aggregated vehicle ownership models has a very long tradition, often based on general demographic or economic parameters. An example that has attracted much attention beyond the boundaries of transport and mobility research are the “patterns of automobile dependence in cities” in various regions of the world, identified by Kenworthy and Laube (1999), in which they use the GDP (and also density factors) to explain car ownership. Due to their limited data quality, such aggregated models are particularly suitable for simple analyses. A typology of aggregated model types is provided by de Jong et al. (2004), *inter alia*, by comparing studies of different car ownership models to which belong aggregate cohort models and aggregate time series models.

Cohort models make use of aggregated demographic data on vehicle ownership. Here, the population is divided into certain age groups [e.g., according to (1) minors up to 17 years, (2) young adults from 18 to 27, (3) adults between 28 and 49, (4) older adults between 50 and 65, etc.]. A common goal of grouping the population by age is to extrapolate future trends in ownership structures using scenario techniques. In the 1980s and 1990s, for example, cohort models were used in countries of the global north to forecast rising motorization rates for the future (Jansson 1989). This assumption was derived from the analyses as follows: the older generations born before the Second World War grew up at a time when car-owning lifestyles were not yet established. At the time of the survey, this age cohort had significantly lower motorization rates than the younger generations of the subsequent car era. On this basis, it was assumed that the degree of motorization would increase ubiquitously with future generations. However, current studies on Generation Y indicate that in most countries of the global North, motorization of societies in this Generation Y will be curbed by a significant decrease in car ownership and driving license rates (Delbosc and Currie, 2013). It is not yet clear how stable this shift away from tools for car use is. In political planning practice, this development is seen as crucial for promoting alternatives to the private car even more strongly than before and thus promoting sustainable development trends for the transport sector.

In time series models, aggregated economic data on the ownership of transport modes are commonly used. In many cases, saturation processes are illustrated using the sigmoid function and various special cases such as the Gompertz curve. As explanatory variables for specific vehicle ownership in societies, the time series models use, for example, GDP, buying power data, oil prices, population densities, etc. These models are based on economic assumptions of diffusion theory and analyze diffusion processes of ownership. The tendency toward low data requirements makes it possible, for example, to model specific vehicle ownership structures in one or more countries, in order to examine and problematize the unequal degree of motorization of regions in the global south and global north with regard to economic parameters. Particularly noteworthy are the works of Dargay and Gately (1999) and Dargay et al. (2007), who use an econometrically estimated model to forecast the growth of the automobile and vehicle population in countries of the global north and the global south, taking into account short- and long-term income elasticities. The studies have a high impact, because they assume that the global vehicle inventory will grow from around 800 million in 2002 to more than 2 billion in 2030. These forecasts raise political questions about the extent to which the expansive demand for oil can be met.

Disaggregated Vehicle Ownership Models

Disaggregated models aim to understand transport ownership more precisely and in more detail, starting from the individual or household level, than would be possible with aggregated models. In recent decades, an extensive and complex variety of disaggregated vehicle models has been developed that differs in terms of (1) theoretical-conceptually deposited explanatory variables, (2) methodological approaches, and (3) political implications. Due to the increasing number of scientific publications, meta-analyses of published data and systematic literature reviews on vehicle ownership models and their effects on behavior and transport systems already exist. The literature review by Anowar et al. (2014), which initially roughly differentiates between exogenous and endogenous models, deserves special mention: exogenous models treat vehicle ownership as being independent of other decisions. Endogenous models, on the other hand, include other decision-making processes for vehicle ownership, in addition to subjective decision processes for specific vehicle ownership. This is based on the assumption that vehicle ownership can by no means be treated only as an exogenous factor for vehicle use, for example, but depends on various decision-making processes (e.g., positive emotions toward a vehicle or a specifically pronounced awareness of environmental problems).

The application of exogenous vehicle ownership models or endogenous vehicle ownership models is suitable for answering very different questions. Exogenous models may be used, for example, to model greenhouse gas emissions in terms of available fleet composition by car type. Methodologically, classical forms of discrete decision experiments are often applied. In many works of the last few decades, especially logistic regressions based on disordered mechanisms have been established in order to prove the behavioral relevance of vehicle ownership. Endogenous vehicle ownership models, in turn, aim to better understand the complex interrelationships for vehicle ownership decisions of the individual or the household. Beyond all used consideration, structural equation models have proven their worth in uncovering more complex relationships. They are usually based on two components: (1) a preceding factor analysis with which correlating background variables are combined into factors and (2) the structural equation with which the directional relationships of the factors and observed variables are checked. The direct, indirect, and total relationships of the explanatory variables are made visible in order to explain the corresponding effects on the specific mode of vehicle ownership and use.

In understanding changes in vehicle ownership based on the individual and/or the household, time plays an important role in the models. Models can be static or dynamic. Static vehicle ownership models examine vehicle ownership at a certain point in time, usually based on crosssectional data, taking explanatory variables into account. They represent a kind of snapshot in time. Dynamic vehicle ownership models, on the other hand, examine vehicle ownership taking into account dynamics over time. They assume that vehicle ownership is by no means static. For this purpose, researchers usually draw on panel data sets based on longitudinal surveys. Pseudo panels can also be used, in which crosssectional surveys were carried out at different points in time. Methodologically, in the crosssection of many studies, different methods are used to model vehicle ownership. The approaches range from mixed multidimensional choice modeling techniques, discrete continuous models, copula-based models and Bayesian models, to simultaneous equation models and structural equation models.

Finally, in the disaggregated modeling of vehicle ownership, objective and subjective factors are used to explain ownership patterns (van Acker et al., 2011). Behind this lie specific theoretical-conceptual assumptions, according to which ownership structures are to be explained on the basis of objective determinants (e.g., life situations) or subjective determinants such as mode-related psychologies. The fact that the objective and subjective factors cannot always be easily separated can be illustrated by the example of the built environment, which structurally foresees specific types of transport use, but which is also reproduced by specific individual preferences for neighborhoods and related transport modes.

Objective Factors: The Example of Life Stages

The concept of life stages plays an implicit role in disaggregated models of vehicle ownership. In the social sciences, life stages represent the interaction of social roles, available resources, and corresponding restrictions. This is based on the theoretical assumption that in the course of their lives people find objective/external living conditions that should be understood against a cultural background and which are determinants of expected decisions (e.g., vehicle ownership decisions): For example, while working people drive to work and need a car, students go to university and own a “semesterticket”/a student public transport pass, the poor have no money and therefore only have inexpensive transport modes available and so on.

In vehicle ownership modeling, life stages are often simplified and translated into sociodemographic or socioeconomic variables, which are intended to explain vehicle ownership, for example, with regard to fleet sizes, vehicle types or different modes including mode usage. Explanatory variables for vehicle ownership are, for example, occupational position, income, age group, family status, or sex. In addition to such sociodemographic and socioeconomic variables, variables on land use or infrastructural accessibility indicators (e.g., distance to the nearest train station) are also used (not always without contradiction).

In the course of time, it is shown that vehicle ownership changes with specific life events such as the birth of a baby, the transition into working life or when moving. One example is the multimodal tool ownership of young adults in Generation Y, which is increasingly associated with life situations beyond car captivity: compared to earlier generations of young adults, Generation Y is more often located in cities with good accessibility structures, studies more frequently at universities, is not at work, and lives more frequently in single households than in family households. With regard to the dynamic ownership models, however, there is concern that a later entry into employment and/or a delayed family formation could lead to the postulation of the necessity of owning cars, which in turn will trigger car use patterns (Article 10429: Young Adults’ Travel: Patterns and Evolution).

Such observations on life stages are important for political implications, because they also provide information about the key events in which a “window of opportunity” opens for the purchase of a new transport mode and possible behavioral changes (Article 10423: Mode Choice and Life Events). For example, if policymakers want to counter the imminent car purchase by young future parents, they could explicitly address parents through political initiatives and offer them attractive alternatives to private cars (e.g., family cabins or discounted family tickets for public transportation).

Subjective Factors: The Example of Mode-Related Psychologies

More recent disaggregated models are increasingly integrating psychological perspectives to explain vehicle ownership and, as a consequence, mode choice. In contrast to the models with objective explanatory variables, the acquisition of certain transport modes or vehicle types can by no means be described as a utilitarian phenomenon (e.g., as a reduced result of the size of the household or the geographical location in space). This is based on the assumption that degrees of freedom have increased as a result of individualization and localization in modern societies and that vehicle ownership differs primarily according to pluralized mode-related psychologies. In these vehicle ownership models, subjective factors are used for explanation.

Psychological factors used to explain vehicle ownership are theoretically and methodologically complex. Here, action-theoretically founded constructs from transport psychology and microeconomics exist to explain the respective claims of ownership structures. The economic models focus on the idea that people act on the basis of rational economic decisions (time and available budget). From a psychological perspective, this approach is considered to be “subcomplex”, because it does not take into account the manifold inner psychic processes. In transport psychology, at least three constructs based on action theory can be differentiated, which have an influence on the ownership decisions and mode use: these are (1) control beliefs as subjective evaluation of persons by means of vehicle ownership to be converted into certain behavior, (2) attitudes that include cognitive, affective, and behavioral elements related to vehicle ownership, and (3) norms that relate to expectations of social groups or individuals that are considered relevant to one’s own person (Hunecke, 2015). All three dimensions have been shown to play a decisive role in ownership decisions and the associated long-term and short-term vehicle uses. In order to explain individual decision-making processes for ownership and use of transport modes as coherently as possible, the constructs are often integratively translated into action models from psychology, such as the theory of planned behavior or the norm activation model, which in turn can be tested using structural equation models (Article 10408: Modeling Passenger Transport Mode Choice).

The use of psychological variables not only plays an important role in understanding vehicle ownership and use, but is also of political and planning relevance. Bamberg et al. (2003) offer an insightful example of the extent to which the introduction of new mobility tools can result in a psychological upgrading of transport modes, which is relevant to travel behaviors. With the help of the theory of planned behavior in a structural equation model, they use a before-and-after survey to show that the introduction of the

semester ticket—an affordable season ticket based on a solidarity payment system for students—at a German university led to a change in attitudes, subjective norms, and perceived behavioral control, expressed in a stronger intention to use public transportation.

Another example of how psychological variables can be used to affect ownership structures in terms of policy implications is the segmentation of individuals by mode-related psychological factors. Of interest is, for example, which attitude-based groups of people can be addressed for new mobility offers (carsharing, MaaS, etc.). Multivariate cluster analysis methods are suitable for this, with which persons can be combined into groups with homogeneous psychological constitutions. They may already have a specific mobility toolkit at their disposal (e.g., a season ticket for public transport) and may also be attracted to other specific mobility options (e.g., by being open to membership of bike sharing providers). Political decision makers and transport companies could then place their offers specifically where these groups move (e.g., at the physical interfaces of major train stations).

Blurred Lines Between Objective and Subjective Factors: The Role of Built Environment in Ownership Decisions

One problem is that the aspects of vehicle ownership that are treated as objective or subjective cannot always be viewed independently of each other and are even blurred. In recent years, there has been an increased effort to address this problem, as for example the category of built environment. Many studies can establish a connection between the built environment and a specific vehicle ownership structure. However, it often remains unclear to what extent vehicle ownership is a result of the built environment or arises from positive attitudes or individual preferences for residential locations and the transport modes associated with the residential locations [models were prominently offered by [Bhat and Guo \(2007\)](#) and most recently by [Macfarlane et al. \(2015\)](#)].

A common approach is to consider the built environment *a priori* as an explanatory variable for vehicle ownership. Methodologically, for example, using multiple linear regression models, specific spatial indicators — for example, “for example density or diversity” — are analyzed simultaneously with other sociodemographic, socioeconomic, and psychographic factors as explanatory variables. However, in this analysis structure, the complex interactions with the other influencing factors that potentially exist remain invisible. The result is an interpretation that can be criticized as spatial determinism, that is, when vehicle ownership—as it were a natural phenomenon—is to be derived from the structural conditions of the residential location, without additionally taking into account the individual socioeconomic or psychographic conditions behind the decision on the residential location.

Methodologically more sophisticated vehicle ownership models (e.g., the earlier-mentioned structural equation approaches), on the other hand, can show that individuals with mode-related preferences (or other psychological categories) harmonize their ownership structures with the built environment of their residential location; that is, people with an affinity for cars are attracted to a locality that is geared to cars, people critical of cars and with an affinity for active or public transportation are attracted to a multimodal built environment, etc. Here, the vehicle ownership models can sew the content to the concept of residential self-selection, which is based on a spatial self-selection of people according to desired forms of life and mobility (see also the remarks in Article 10379: Residential Location Choice Models). Here, the desired ownership of transport modes is satisfied in complementary spatial structures, that is, in residential locations where the affective connection between human being and transport modes is most in harmony. In this sense, the built environment is reproduced by individual preferences for residential locations and associated transport modes.

Likewise, certain vehicle ownership can also be a result of the built environment. Especially in societies with a strong socioeconomic divide and a tense housing market, social exclusion processes can result in specific groups of people being unable to realize their preferred vehicle ownership on-site. The processes of gentrification in many prospering cities of the global North, in which socially marginalized groups of people in particular are pushed into peripheral locations, should be mentioned as an example. A residential mismatch describes the discrepancy between the preferred and the real mode choice depending on the place of residence. The concept of “forced car ownership” ([Banister, 1994](#)), for example, describes the way poor individuals are forced to buy a car in order to participate in society at all, even though they cannot afford it.

The evidence of concepts such as residential self-selection or a residential mismatch can be methodically considered within the framework of vehicle ownership models if they are implemented, for example, within the framework of the structural equation models mentioned earlier.

New Challenges: New Technologies, New Transport Modes, No Ownership Anymore

New challenges in modeling and predicting vehicle ownership effects exist with respect to the transformation of hegemonic, autocratic transport systems. In the automobile-influenced societies of the global north, research has long focused on the modeling of structures and effects of private car ownership. With the increasing interest in alternative concepts to the automobile such as multimodality—that is, the use of more than one transport mode—further mode options have already been included in the models in recent years. An interest of research in the context of transformation processes of the automobile societies exists in the understanding of new structures and effects with regard to (1) new driving technologies, (2) new transport modes, and (3) new transport systems beyond ownership.

The modeling of new driving technologies is based on purely technical solutions to meet (national and international) climate protection targets and at this level aims to reduce emissions of globally effective greenhouse gases. The models usually focus on hydrogen-based fuel cell technologies or the electrification of vehicles. Since, for example, electric vehicles differ fundamentally from

fossil fuels in certain characteristics (shorter distances, long battery charging times, etc.), modelers are often interested in specific questions as to the extent to which electric vehicles can satisfy the daily transport needs of certain population groups or which specific groups of people are interested in ownership and use with regard to their attitudes and preferences. This interest is usually based on simple economic assumptions, according to which early adopters for emerging market segments are to be identified.

New vehicle ownership structures arise with the emergence of new transport modes. This applies, for example, to the modeling of potential ownership structures of autonomous vehicles with regard to possible effects on individual routes or general mode choices and routing. In this respect, it is assumed that an autonomous vehicle era can result in a reorganization of individual vehicle ownership, which could fundamentally change spatial movement patterns. Researchers are interested, for example, in the extent to which households and individuals could replace their current private transport modes, adapt residential locations in relation to places of work and leisure, and what kind of (social, ecological and economic) consequences this has for existing transport infrastructures (e.g., road capacities, parking needs). Also, individual preferences with regard to lifestyle factors seem to be of interest for the ownership of autonomous vehicles, in order to identify possible early adopters.

Finally, a third challenge in the modeling of transport ownership concerns the emergence of new mobility services, in which vehicle ownership is successively replaced by vehicle usership. Individuals no longer necessarily own their own transport modes, but decide flexibly—that is, entirely in the sense of smart mobility on the basis of interconnected mobility services (carsharing, bikesharing, scootersharing, trains, trams, buses, etc.) using modern information and communication technologies such as the smartphone—on the basis of their memberships which mode is the most suitable for the respective situation, on the basis of their memberships. The autonomous vehicles mentioned earlier can also be a component of such a system. In many economically prosperous cities, technically sophisticated mobility services are already being offered and used. A central question for researchers is to what extent individual behavior patterns produced within the framework of interconnected mobility services can lead to the dissolution of ownership structures in general, thereby achieving green effects in transport, or which social groups participate in the new services in terms of sociodemographic or psychographic characteristics.

References

- Anowar, S., Eluru, N., Miranda-Moreno, L.F., 2014. Alternative modeling approaches used for examining automobile ownership. A comprehensive review. *Trans. Rev.* 34 (4), 441–473, doi:10.1080/01441647.2014.915440.
- Bamberg, S., Ajzen, I., Schmidt, P., 2003. Choice of travel mode in the theory of planned behavior. The roles of past behavior, habit, and reasoned action. *Basic Appl. Social. Psych.* 25 (3), 175–187, doi:10.1207/S15324834BASP2503_01.
- Banister, D., 1994. Equity and acceptability question in internalising the social costs of transport. Paris.
- Bhat, C.R., Guo, J.Y., 2007. A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels. *Transport. Res. B Meth.* 41 (5), 506–526, doi:10.1016/j.trb.2005.12.005.
- Dargay, J., Gately, D., 1999. Income's effect on car and vehicle ownership, worldwide. 1960–2015. *Transport. Res. A Policy Pract.* 33 (2), 101–138, doi:10.1016/S0965-8564(98)00026-3.
- Dargay, J., Gately, D., Sommer, M., 2007. Vehicle ownership and income growth, worldwide 1960–2030. *Energy J.* 28 (4), 143–170, doi:10.5547/ISSN0195-6574-EJ-Vol28-No4-7.
- de Jong, G., Fox, J., Daly, A., Pietres, M., Smit, R., 2004. Comparison of car ownership models. *Trans. Rev.* 24 (4), 379–408.
- Delbosc, A., Currie D.G., 2013. Causes of youth licensing decline. A synthesis of evidence. *Trans. Rev.* 33 (3), 271–290.
- Hunecke, M., 2015. Mobilitätsverhalten verstehen und verändern. Psychologische Beiträge zur interdisziplinären Mobilitätsforschung. Springer VS, Wiesbaden.
- Jansson, J.O., 1989. Car demand modelling and forecasting: a new approach. *J. Trans. Econ. Policy* 23 (2), 125–140.
- Kenworthy, J.R., Laube, F.B., 1999. Patterns of automobile dependence in cities. An international overview of key physical and economic dimensions with some implications for urban policy. *Trans. Res. A Policy Pract.* 33 (7–8), 691–723, doi:10.1016/S0965-8564(99)00006-3.
- Le Vine, S., Lee-Gosselin, M., Sivakumar, A., Polak, J., 2013. A new concept of accessibility to personal activities. Development of theory and application to an empirical study of mobility resource holdings. *J. Trans. Geogr.* 31, 1–10, doi:10.1016/j.jtrangeo.2013.04.013.
- Macfarlane, G.S., Garrow, L.A., Mokhtarian, P.L., 2015. The influences of past and present residential locations on vehicle ownership decisions. *Transport. Res. A Policy Pract.* 74, 186–200, doi:10.1016/j.tra.2015.01.005.
- Scott, D.M., Axhausen, K.W., 2006. Household mobility tool ownership. Modeling interactions between cars and season tickets. *Transportation* 33 (4), 311–328.
- Simma, A., Axhausen, K.W., 2003. Commitments and modal usage. Analysis of German and Dutch panels. *Transport. Res. Rec. J. Transport. Res. Board* 1854, 22–31.
- van Acker, V., Mokhtarian, P.L., Witlox, F., 2011. Going soft: on how subjective variables explain modal choices for leisure travel. *Eur. J. Trans. Infrastruct. Res.* 11 (2), 115–146.

Further Reading

- Clark, B., Chatterjee, K., Melia, S., 2016. Changes in level of household car ownership. The role of life events and spatial context. *Transportation* 43 (4), 565–599, doi:10.1007/s11116-015-9589-y.
- Dargay, J.M., 2001. The effect of income on car ownership. Evidence of asymmetry. *Transport. Res. A Policy Pract.* 35 (9), 807–821, doi:10.1016/S0965-8564(00)00018-5.
- Lavieri, P.S., Garikapati, V.M., Bhat, C.R., Pendyala, R.M., Astroza, S., Dias, F.F., 2017. Modeling individual preferences for ownership and sharing of autonomous vehicle technologies. *Transport. Res. Rec. J. Transport. Res. Board* 2665 (1), 1–10, doi:10.3141/2665-01.
- Levinson, D.M., Marshall, W., Axhausen, K., 2017. Elements of access. Transport planning for engineers. In: *Transport Engineering for Planners*, Network Design Lab, Sydney.
- Pinjari, A.R., Eluru, N., Bhat, C.R., Pendyala, R.M., Spissu, E., 2008. Joint model of choice of residential neighborhood and bicycle ownership. *Transport. Res. Rec. J. Transport. Res. Board* 2082 (1), 17–26, doi:10.3141/2082-03.
- van Acker, V., Witlox, F., 2010. Car ownership as a mediating variable in car travel behaviour research using a structural equation modelling approach to identify its dual relationship. *J. Trans. Geogr.* 18 (1), 65–74, doi:10.1016/j.jtrangeo.2009.05.006.
- Vij, A., Carrel, A., Walker, J.L., 2011. Capturing modality styles using behavioral mixture models and longitudinal data. 2nd International Choice Modelling Conference. Leeds.

Bicycle Sharing/Bikesharing

Catherine Morency, Jean-Simon Bourdeau, Department of Civil, Geological and Mining Engineering, Polytechnique Montreal, Montreal, QC, Canada

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	617
Introduction	617
Overview of Bikesharing	617
Designing a Bikesharing Service	619
Operational Management: The Challenge of Rebalancing	619
Understanding Bikesharing Usage	620
How Bikesharing Fits in the Multimodal Urban Bouquet of Transportation Modes	620
What are the Impacts of Bikesharing in Cities?	621
Next Steps	621
Acknowledgment	621
See Also	621
References	621
Further Reading	622

Nomenclature

CBD It refers to the Central Business District of an agglomeration where there is important concentration of employment, commercial, and leisure locations.

Dock A dock is a physical system that allows parking and locking a shared bike.

Dockless It refers to a bikesharing system that operates without stations and docks to lock the shared bikes.

First and last mile It refers to the access segments of a transit trip, that is, the distance between the origin and the transit station or between the transit station and the destination.

Geo-fence or electric fence A geo-fence or electric fence refers to the concept of defining a zone with virtual frontiers relying on global positioning system (GPS), radio frequency identification device (RFID), or Bluetooth signals to determine whether bikes are within or near a certain area.

Loop A trip chain starting and ending at the same station.

One-way trip A trip that has different origin and destination stations.

Rebalancing It refers to the operation by which shared bikes are redistributed among the stations typically using trucks.

Trip generation It refers to estimation of the number of trips that are produced or attracted by a station depending on its features (number of docks) and the features of the surroundings (population density, transit stations, etc.).

Introduction

This paper first proposes an overview of bikesharing, namely its origin, its most important evolutions over time, the main components of a bikesharing system, as well as a description of how such system typically works. It then discusses the challenge of designing a system in a city (where and how many stations and how many bikes for instance) as well as the most important challenge with respect to operational planning of such a system, which is the rebalancing of bikes throughout the stations. This leads to the issue of modeling the use of the systems (how many trips, between which stations and at what time) as well as of the behaviors of users, both members and occasional ones. Factors which are the most determinant on the use of a bikesharing systems are afterward discussed, followed by some views on how such transportation mode fits in the multimodal transportation systems of cities as well as on the impacts it has at the city level, namely with respect to sustainable development collective objectives. Key emerging issues are finally discussed.

Overview of Bikesharing

Bicycle sharing or bikesharing refers to systems in which people can rent bicycles for short periods of time to travel either between stations, for systems with docks, or within zones for dockless ones.

The first official bikesharing service appeared in 1965 in Amsterdam under the name Witte Fietsen thanks to the initiative of a citizen which decided to provide 50 bicycles all painted in white for free usage by all citizens. Unfortunately, without any sort

of management, the entire fleet of bikes disappeared within few days. This is said to be the first generation of bikesharing systems.

The second generation of bikesharing systems appeared in the beginning of the 1990s in Denmark, first with small-scale systems and then with City bikes, the Copenhagen system, bringing many improvements compared with the previous generation, namely more robustly designed bikes, usage management through coin deposit and stations. Still, such system was facing important theft issue due to the low cost and anonymity of users.

The third generation of systems addressed this issue with improved users (and bikes) monitoring and introduction of pricing of usage. People must register or at least provide some sort of identification and bank account to be able to rent a bike, consequently discouraging bad usages or thefts. It is the 2005 Lyon's Velo'v system, with 1500 bikes, that really consolidated this third generation of bikesharing system, due to its large scale and noticeable impacts. Many other systems followed in Europe, Americas, and Asia, and all contributed to increase the interest toward this innovative urban transportation mode.

Improvements for the fourth generation include mobile docking stations that can be moved around to fit user demand or that can be stored when weather conditions are not suitable to operate such system (as it is the case for the Montreal's bixi system which only operates from April 15 to November 15 due to cold weather and snow during the other months) as well as the use of local energy sources for the stations such as solar panels.

Readers are invited to read [De Maio \(2009\)](#) who provides further insights into the history of bikesharing.

The main components of station-based bikesharing systems are the bikes, the stations, and the docking points. Bikes typically are constructed to prevent access to internal components and limit the risk of breakage or piece theft. They are robustly designed and made accessible to a diversity of cyclist. A robustly design, fourth-generation shared bike from the Montreal bixi bikesharing system, is presented in [Fig. 1](#).

Stations (example presented in [Fig. 2](#)) are composed of a kiosk to allow users to make transactions such as buying a day pass or choosing a bike and docking points. New systems also allow to directly take a bike without interacting with the kiosk while some stations do not provide a kiosk and users must make their transactions with their phones.

In the recent years, we have seen the comeback of bikesharing systems without stations. But such systems have nothing in common with the first generation of systems. They are called dockless bikesharing systems and use bikes equipped with GPS



Figure 1 Example of a share bike from the Montreal bixi bikesharing service. Source: Image provided by bixi Montreal.



Figure 2 Example of a station from the Montreal bixi bikesharing service. Source: Image provided by bixi Montreal.

(or other technologies) allowing to locate the bikes at any time as well as onboard systems to allow starting/ending a rental without requiring a physical docking point. Hence, bikes are monitored through technology and users access the bikes using their phone. It must be noted that dockless systems are more vulnerable to vandalism and theft. While they are appealing for users since they do not need to bother about finding an empty dock to leave their bike or walking from the station to their destination, such system may create chaos in cities if rules regarding where bikes can be left are not clear and enforced. One way of managing the areas where bikes can be returned after usage is through geo-fences. Instead of allowing bikes to be left anywhere within a certain area, it may be useful to prevent misbehaviors and to control specific small zones where bikes can be left using geo-fences (or electric fence).

Designing a Bikesharing Service

Depending on the systems, the number of stations and docks per station will vary, from a few in lower density areas to few hundreds or more in central parts of the city or near important trip generators and transit stations. The design will depend on the market potential in the various areas, in terms of membership (hence related mostly to population density and sociodemographic composition) and trip generation (depending on types and density of interest points). Various attributes will determine the production and attraction potential of a station. Determining where to implement the stations, how many to install, as well as how many docking points to include in each station are important design questions that will have an impact on the number of members that will adopt the service as well as on how many trips each station will support. This problem is often referred to as the location-allocation problem and proposed models to assist in the design will aim typically to minimize impedance (minimize distance between stations and demand) and maximize coverage (maximize the population covered by the stations). The set of stations also needs to provide good coverage and connectivity in the sense that stations should not be too far away from each other, especially when expanding the network in more residential areas. According to US and Canadian bikesharing operators, optimal distance between stations varies between 275 and 400 meters (Shaheen et al., 2012). The distribution of stations throughout a system can be assessed using indicators such as density of stations, area covered by stations, compactness, and distance between stations and number of stations (Médard de Chardon et al., 2017).

Since bikesharing most plausibly will not be able to provide an answer to all transportation needs, it often works well when implemented where other transportation alternatives are available, namely transit. Moreover, since it allows users to do either one-way trips or loops, it can be used in combination with other modes, to access transit for instance or to come back from work in the afternoon while traveling by transit in the morning. There is no consensus among operator with respect to optimal distance between a bikesharing station and a transit station but it is as low as 25 meters for large-scale systems with good public transit systems (Shaheen et al., 2012).

Regarding the number of bikes, it is also an important feature of the bikesharing system. Actually, the ratio between the number of bikes and the number of docks is a key indicator to account for when designing a system. Too many bikes per dock will facilitate access to shared bikes but will make it harder to find an empty dock at the end of the ride while too few bikes per docks will make it hard for users to find a bike but those who do will typically not have issues finding an empty dock to return it. Most system will aim for a ratio of two docks per bike, but it can vary a lot, namely if we account for system failures (damaged bikes or docks) (Shaheen et al., 2012).

Another important design component is the pricing scheme. In most systems, users of the bikesharing systems are either members or occasional users. In the first case, people register as member for various time periods (month, year) and can use the shared bikes as often as they want without further cost as long as their trips are under a time threshold (typically 30–45 minutes). In the second case, they either buy a short-time access, 24–72 hours for instance, and or have free rides as long as they are under the time threshold or pay for a single use.

In the context of network development, the design question will narrow down to identifying potential zones for extension. In such case, potential market must be assessed and impacts on the generation of new trips as well as on the spatial-temporal structure of the trips require modeling.

Operational Management: The Challenge of Rebalancing

On a day-to-day basis, a bikesharing operator must continuously assess whether the bikes are in the right stations or whether some docks need to be emptied. This challenge, faced by most if not all bikesharing operators, is referred to as rebalancing or more precisely bikesharing rebalancing problem (BRP). Many developments have been made with respect to algorithm and approach to assist operators in anticipating where bikes must be relocated to meet travel demand. These algorithms perform short-time predictions of the optimal state of the system and optimize the boarding/alighting movement over a truck route while accounting for various constraints such as the number of bikes in the system, the capacity of each station, the number of vehicles available, as well as their capacity (in number of bikes they can transport).

The reason why systems must be rebalanced is due to spatial heterogeneity of demand over time. Typically, in the morning period, trips will dominantly start in stations located in residential areas, usually further from the CBD, and end in stations located near points of interest, namely centrally located. This is typically referred to as commuting patterns. Hence, stations used for departure will be emptied by users while those located nearer to important activity locations (work, study) will be filled by users

returning their bikes. Rebalancing will thus be required to ensure users in residential locations can access bikes and users heading to central areas find empty docks. Topography of the city can further enhance the issue of origin–destination asymmetry since stations located on top of hills will be less likely to be chosen as destination stations and will continuously require to be filled, the reverse being true for down the hill stations that will be more rapidly filled. Of course, this issue will lessen if the bikes have electric assistance.

Various indicators can be estimated to assess the quality of service from the user point of view such as the proportion of time a station is empty and the proportion of time a station is full. In both cases, the user may have difficulty either finding a bike or finding an empty dock to return the one he has. Other performance indicators can be estimated such as the ratio of bicycles docked versus capacity per station, the proportion of damaged bikes daily, the utilization rate, that is, the total bicycle-minutes of usage versus the total bicycle-minutes of availability per temporal unit. With respect to the operator, operating cost must be minimized since daily redistribution of bikes using trucks is an important component of these costs and also affects the perception of bikesharing since some argue that the environmental gains of making bikes more available for daily travel is cancelled by the motorized vehicle-kilometers traveled by these redistribution trucks.

Understanding Bikesharing Usage

The way bikesharing systems are used is examined either from a system-wide perspective or from a user perspective. Administrative data from bikesharing systems, when available at the trip and member level, allow conducting both types of analysis.

Analysis at the system-wide level focuses on the estimation of daily trips generated per station (either trips starting, ending, or the sum of both), depending on various features of the stations. Station features which were found to be more determinant of the number of trips observed per stations are number of docking points, altitude of the stations, population density, employment density, proximity to a university/college, distance to nearest stations, density of points of interest, proximity to transit, and availability of cycling infrastructures. Global trips per day per bike and its variability throughout the months can, for instance, be used to compare the performance of various systems (Médard de Chardon et al., 2017). Various types of models can be used to explain and forecast total number of trips per station such as regression models, ordered logit models, or spatial expansion models. Models were also developed to understand and forecast origin–destination matrix between stations. Travel time or distance, on the bike network, is used as input in distribution models that aim to reproduce the spatial structure of trips.

Analysis at the user level focuses on the understanding of usage patterns by users. But first, study using both administrative data and surveys helps assess the dominant features of users. While characteristics vary across systems, the most typically reported attributes for becoming a member are: typically men, younger than the average population (typically below 30–35 years old increases probability of being a member), home location near to a station, both students and working people.

Hence, it is frequent that users will be classified based on their usage patterns using clustering methods. Typologies often segment users based on their frequency of use (from hardly use to frequently use), their seasonal pattern of use (mainly during the 2 summer months or during the entire season), their weekly behaviors (weekdays during peaks or weekends), the diversity of stations they use, the structure of their trips (one way or loop), the distance/duration they travel, or the fact that they combine bikeshare with other modes such as transit. All these indicators are useful to apprehend usage, identify potential markets, and forecast the impacts of design scenarios.

Hence, various factors have been confirmed as having an important impact on bikesharing usage. Some factors are structural and will promote the use of cycling, both private bicycles and shared bikes. These factors are built environment (namely route pattern and diversity of usages), topography (hilly areas make cycling more challenging), cycling infrastructures, safety of the available route (related not only to available of infrastructures but also to traffic speed and flow), and features of the trip itself such as its length and the time period when it is done. Other factors will cyclically or punctually affect cycling and bikesharing usage such as weather. Typically, the use of cycling will increase as the weather gets warmer but it will start to decline again when temperature exceeds 30°C or so.

Due to unfavorable weather conditions, many systems in the northern hemisphere will either stop operating during winter or reduce their capacity. Some fourth-generation systems will actually be fully removed from the streets during the winter period due to cold and important snowfalls. While some users will make bikesharing throughout the operating season, clement weather will favor increased usage of bikesharing. Additionally, rain will, on the contrary, significantly reduce the number of trips occurring on the network (both rain forecast and current rain). Other types of events, either planned (match or show for instance) or unexpected (subway breakdown), will also change the typical behaviors and will increase demand punctually in specific locations. This will complexify the rebalancing issue; still, with good short-term forecasting models including planned events, network can be planned according to expected travel demand.

How Bikesharing Fits in the Multimodal Urban Bouquet of Transportation Modes

With the diversification of alternative modes of transportation, there has been increase interest into how to coherently integrate all these services to provide the required tools to meet the entire set of travel needs, as a global aim to reduce the place of private cars in cities. Concepts such as Mobility as a Service (MaaS) or integrated mobility have emerged to frame the interaction between modes

and envision services that are planned and provided in batch. Hence, the question of the role bikesharing plays in an urban bouquet of transportation modes has become more and more relevant. Both competitiveness and complementary interactions have been examined to understand which modes in which contexts should be favored.

Results of these analyses point to the fact that bikesharing may be competing with taxi for short trips during the summer months while transit and bikesharing have a more complex relationship, bikesharing sometimes facilitating access to important transit nodes (first and last mile transit issue), replacing transit trips, or compensating for transit unavailability during night period, closures, and disruptions for instance. Research has found that bikesharing both decreases and increases transit ridership, while the decrease, when observed, was very low. Hence, long-term perspectives must be added to trip-based analysis since, in the long run, availability of more than one option may prevent someone from buying and using a private car. Further discussion on the interaction between transit and bikesharing can be found in TCRP 132 ([National Academies of Sciences Engineering and Medicine, 2018](#)).

What are the Impacts of Bikesharing in Cities?

Bikesharing has reinvented the way people use bicycles to move around cities. It has also confirmed that cycling can be part of the transportation systems that cities can rely on to reduce dependence toward the car. Various studies have assessed the impacts of bikesharing in cities such as health impacts, environmental benefits, GHG reduction, decreased automobile use, increased mobility, promotion of cycling as a travel tool not only a leisure activity, and congestion reduction. An important component of such assessment is to be able to estimate what would happen with the trips currently done using bikesharing if there was no bikesharing system available: would they be traveled using another mode of transportation and if so which mode would be selected. Hence, results vary according to system size (stations and bikes), city scale, demographic composition (namely income and car ownership levels), and quality of the available transit system. Regarding the health impacts, estimations show that the health benefits of the physical activity related to cycling are more important than the increased risk of traffic fatalities and air pollution for those who cycle ([Otero et al., 2018](#)). Hence, at the individual level, in addition to its contribution in staying fit, bikesharing is a cheapest travel option than private car or even transit and can contribute to reducing travel expenditures for household.

Next Steps

The sharing economy has changed the way people access to services. Many cities have seen the proliferation of station-based carsharing services, free-floating car services, and dock and dockless bikesharing systems. More recently, e-scooters have been invading cities, requesting them to seriously address the issue of road sharing and shared vehicle managements. People are rapidly adopting these agile transportation modes that offer them high flexibility and accessibility at low cost, but integrated planning still is a step behind. Travel choices have become highly dynamic as they adapt to available options at the time of travel. Cities will need to plan the transportation systems in an integrated way and account for the travel needs and system requirements while aiming for the optimal combination from the global sustainability point of view.

Acknowledgment

The authors wish to acknowledge the collaboration of *bixi* Montreal for providing access to data for research purposes. They wish to acknowledge the *Mobilité* Chair partners for their support and financing: City of Montreal, Montreal Transit Authority, Ministère des transports du Québec, Autorité régionale de transport métropolitain, Exo—Réseau de transport de Montréal. The authors finally acknowledge the financial support of NSERC (Natural Sciences and Engineering Research Council of Canada) and the Canada Research Chair program.

See Also

Cycling economics
Bicycle infrastructure
Bicycles: The Safety of Shared Systems vs. Traditional Ownership

References

- De Maio, P., 2009. Bike-sharing: history, impacts, models of provision, and future. *J. Public Transp.* 12 (4), 41–56.
- Médard de Chardon, C., Caruso, G., Thomas, I., 2017. Bicycle sharing system 'success' determinants. *Transp. Res. Part A* 100, 202–214.
- Otero, I., Nieuwenhuijsen, M.J., Rojas-Rueda, D., 2018. Health impacts of bike sharing systems in Europe. *Environ. Int.* 115, 387–394.

Shaheen, S., Martin, E.W., Cohen, A.P., Finson, R.S., 2012. Public bikesharing in North America: early operator and user understanding. Mineta Transportation Report 11-26. Available from: <http://transweb.sjsu.edu/sites/default/files/1029-public-bikesharing-understanding-early-operators-users.pdf>.
National Academies of Sciences, Engineering, and Medicine, 2018. Public Transit and Bikesharing. The National Academies Press, Washington, DC, doi: 10.17226/25088.

Further Reading

Fishman, E., 2016. Bikeshare: a review of recent literature. *Transp. Rev.* 36 (1), 92–113, doi:10.1080/01441647.2015.1033036.
IDTP, 2018. The bikesharing planning guide. Institute for Transportation and Development Policy, 110 pp. Available from: <https://bikeshare.itdp.org/>.
NACTO, 2019. Shared micromobility in the U.S.: 2018. National Association of City Transportation Officials, 16 pp. Available from: https://nacto.org/wp-content/uploads/2019/04/NACTO_Shared-Micromobility-in-2018_Web.pdf.
Vogel, P., 2016. Service Network Design of Bike Sharing Systems: Analysis and Optimization. Springer, New York 172 pp.
Shaheen, S., Cohen, A., 2019. Shared micromobility policy toolkit: docked and dockless bike and scooter sharing. UC Berkeley Transportation Sustainability Research Center. Available from: <https://escholarship.org/uc/item/00k897b5>.

Carsharing

Shiva Habibi*, **Frances Sprei†**, *RISE, Research Institute of Sweden, Göteborg, Sweden; †Chalmers University of Technology, Göteborg, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Background	623
Data	623
Revealed Data	624
Surveys	624
Travel Surveys	624
Stated Preference	624
Methods	624
Demand	624
Membership Models	624
Travel Demand	625
Supply	626
Gaps, Further Research and Final Words	626
Biographies	626
References	627
Further Reading	628

Background

Carsharing is a type of shared mobility that allows short-term access to a car. Carsharing services can be traced back to 1948 in Germany. However, they increased in popularity with advances in information and communications technology (ICT) (Shaheen et al., 2016). Carsharing can influence transportation systems, energy systems and city development through changes in travel behavior, land use as well as the creation of new businesses.

There are several different schemes for carsharing in the market. Round-trip station-based carsharing was the first form of carsharing introduced and is still the most widespread. In this carsharing scheme, users pick up a car from a station and return it to the same station after use. This carsharing scheme can be resource-efficient for operators, but less flexible for users. One-way carsharing services (station-based and free-floating) are more flexible schemes offered to the market by operators. In one-way station-based carsharing systems, users do not need to return cars to the station of origin. In free-floating services, users can pick up and return cars to any location within the business area of these services, which is usually the city center. One disadvantage of this flexibility is the lower reliability in terms of availability of shared cars. Due to different pricing and service schemes, the usage patterns (specifically in terms of rental duration and distance traveled) and user groups of free-floating services might be different from those of station-based services (see e.g., Becker et al., 2017a; Kopp et al., 2015). The research on free-floating services, including their impact on sustainable transport indicators, has been less-developed. The relative newness of the services, the lack of relevant data, and the additional complexity presented by the fact that the cars could be anywhere, pose major obstacles to research. Currently, four different carsharing markets exist: business-to-consumer, business-to-business, peer-to-peer, and not-for-profit. In the business-to-business model, employees are provided access to a carsharing organization's fleet through their employer. Peer-to-peer carsharing services allow car owners to make their personal cars available to be shared on a short-term basis. In this chapter, the focus will only be on business-to-consumer carsharing since this is the type that has been most extensively studied.

Despite its small mode share, carsharing is growing throughout the world and becoming recognized as a new transport mode. It is therefore crucial to develop transport models in which carsharing and other new mobility services are explicitly modeled in order to assess their impacts. These models can be used by policymakers to include these services in future planning processes and predict different future scenarios. Moreover, these models help operators to design efficient and frequently used services. This chapter overviews the modeling efforts in the area, addresses the limitations and research gaps and discusses future research. It starts with data used for modeling and continues with the modeling of different sectors of carsharing, that is, demand and supply and their interaction. The focus of the chapter is mainly on the demand side.

Data

Different data sources have been used in the literature to study carsharing services. These sources are introduced below.

Revealed Data

Revealed data in this field usually comprise membership and usage data obtained from the operators. This source of data is limited in that it does not provide any information about nonmembers. Therefore, it does not provide a basis for comparison and is consequently not appropriate to studying changed travel behavior or to estimating choice models. However, this type of data could be suitable for validation of simulation models.

There is a relatively new source of data, specifically in the field of free-floating carsharing, that is obtained by web scrapping (see e.g., [Kortum et al., 2016](#)). Web-scraping methods enable the collection of real-time data from publicly available information. The technique simulates a user requesting available information to book a car from operators' website, but is performed in an automated and repeated way, for example, every 60 s. This source of data is good for studying trip characteristics and spatiotemporal distribution; however, it does not provide any information on users or purpose of trips.

Surveys

One of the most common approaches to studying carsharing travel behavior is to use surveys. The questions of interest in surveys are usually users' characteristics, lifestyles and attitudes as well as changes in travel behavior after joining carsharing such as changes in private car ownership, vehicle miles driven, or transport mode taken. However, in the majority of behavioral change studies, the observed behavioral changes are attributed to carsharing membership and the correlation is assumed causation. Yet there could be a common underlying factor, such as a life event, or structural issues, such as access to public transport, that affects both carsharing membership and travel behavior changes (endogeneity issue). The possibility of endogeneity issue could lead to systematically overestimating the impacts of carsharing ([Le Vine and Polak, 2015](#)). A few studies have controlled for these issues by including a sample of noncar sharers (see e.g., [Cervero, 2003](#); [Kopp et al., 2015](#)).

Another methodological issue with surveys in the field is recall bias ([Becker et al., 2017a](#); [Kopeck and Esdaile, 1990](#)). In these surveys, respondents are usually asked about their travel behavior (e.g., vehicle miles traveled) before joining carsharing, and the assessment is therefore based on self-reported data. One way to overcome this bias is by designing a survey in multiple waves of before-and-after joining carsharing (see e.g., [Cervero and Tsai, 2004](#)).

Travel Surveys

Travel surveys or travel diaries are often the largest and most reliable data sources available to transportation researchers. Travel surveys include data on individual travel behavior, capturing all activities and trips during a predefined period. However, carsharing has not been included in travel surveys in many countries. An essential future effort in modeling carsharing services is to explicitly include them in travel surveys.

Stated Preference

Stated preference (SP) choice experiment is another type of survey used to study carsharing services. These experiments are designed to measure preferences for different attributes of carsharing services. Small scale and being based on a nonrepresentative sample of the population are the major limitations of the majority of the stated preference choice experiments in the field.

Methods

There is a large descriptive body of literature that deals with carsharing membership, user characteristics and travel behavior, but modeling approaches are still limited. The most important challenge is data. Models are expected to be policy-sensitive so that they can be used by practitioners and policymakers to test different future scenarios. Thus, extensive data sources on individuals' socio-demographics, travel behavior, travel needs, and attitudes are required. All these data are not necessarily accessible and available, especially data on newer services. In many cases, stated preference surveys are required, which are costly and might not be representative. Below, modeling efforts in the field for both the demand and supply sides are discussed. The focus of the chapter is mainly on the demand side.

Demand

Although the popularity of carsharing is increasing worldwide, the estimation of travel demand for this mode has only rarely been addressed by researchers. In this section, the literature on demand modeling and its limitations and gaps is reviewed.

To be able to choose carsharing as a transport mode, one first has to join a carsharing organization. This section thus starts with different membership models and continues on to trip demand models that deal with the day-to-day demand for carsharing services.

Membership Models

Membership models explore the market potential for carsharing services by identifying potential members. These models investigate the factors affecting the intention to join a carsharing organization. Similar to car ownership models, these models are estimated at the person level and are used to predict mid-term strategic travel choices that influence trip-level decisions.

In the majority of studies, membership models are estimated using discrete choice models. In the absence of member characteristics, Kortum and Machemehl (2012) use regression models to predict the proportion of population of a census block that are members of free-floating carsharing given the probability of that block to contain any members. To the best of our knowledge, this is the only study modeling the membership of free-floating services.

Early attempts to identify carsharing membership potential mostly focused on socio-demographics and annual private car mileage to find a breakeven point between ownership costs and sharing costs. Later on, aspects such as neighborhood factors, individuals' travel patterns, carsharing attributes, individuals' attitudes, and environmental concerns have been added.

However, most models in the field have limiting assumptions. The main limitations are described in the following paragraphs.

Supply-side attributes are known to affect potential membership for carsharing. However, only a small number of studies have included the attributes of carsharing services or supply-side aspects in membership models (see e.g., Juschten et al., 2019).

Another common limitation of the majority of membership models is that they do not include other alternative mobility options. The choice of becoming a carsharing member is a choice of a mobility resource, that is, product, service or knowledge that enables travel (Le Vine et al., 2014). To incorporate other mobility resources as explanatory variables in membership models might lead to endogeneity issue, as discussed earlier. The challenge of modeling the choice between different mobility resources is that mobility resources are not mutually exclusive. One way to overcome this problem is to estimate every possible combination of mobility resources and define a mobility portfolio, as in Le Vine et al. (2014). Another approach is to account for correlation in error terms between choices by employing, for example, a multivariate Probit model (see Becker et al., 2017b).

Another limitation is that most researches ignore the role of individuals' attitudes and lifestyles in carsharing decisions. Attitude and satisfaction have been addressed by a few studies employing hybrid choice models (see e.g., Kim et al., 2017). Hybrid choice models (Ben-Akiva et al., 2002) incorporate latent attributes such as values, social norms and lifestyle based on a set of observable indicators.

Finally, becoming a carsharing member does not necessarily mean becoming an active member, that is, using carsharing as a transport mode. Morency et al. (2012) and Habib et al. (2012) identify active members with dynamics of habit formation, that is, linking the decision to use the carsharing service to the user's previous usage decisions in a defined period, that is, 4 months.

Travel Demand

Travel demand refers to day-to-day trips made using carsharing services. These decisions are made at the trip-level and are conditional on the decision of becoming a carsharing member. There are a few studies that model the modal split of these services. The majority of these studies are based on stated preference experiments due to the lack of suitable data. However, trip-level demand models might not be appropriate tools to model travel demand for carsharing services. Studies show that car-sharers are purpose-oriented, flexible, multimodal and intermodal. Additionally, characteristics like reservation prior to use and distance-time price structure of carsharing services further encourage users to plan their activities and their daily schedules. Therefore, considering only one trip, as in trip-level models, might not be representative for this type of travel behavior and the entire daily tour might need to be considered, as in activity-based models. Moreover, carsharing is a service with limited supply in terms of both fleet size and stations. Therefore, the accessibility and reliability of these services have a determining effect on the choice of these services (De Lorimier and El-Geneidy, 2013). Accessibility is determined by the distance between the position of the user and the positions of the stations/cars. Reliability is the availability of cars when they are needed. Furthermore, shared cars compete with private cars for infrastructure in terms of both roads and parking spaces. All these attributes make the demand for these services supply-sensitive, and accounting for these attributes requires high spatial and temporal resolution.

A new research direction was introduced by Ciari et al. (2008), who address the above-mentioned aspects by employing an activity-based agent-based microsimulation using MATSim (Balmer et al., 2008) as the platform. This type of models allows for a very rich description of both individuals and infrastructure and therefore explicitly accounts for the complex links between travel needs and supply attributes in order to predict carsharing demand. It simulates the daily activities and travel of each agent. All of these come at the cost of being computationally intensive. Moreover, the agent-based modeling approach is mainly based on the behavioral rules defined at the microlevel and therefore does not rely much on past information to make predictions (Ciari et al., 2013). This modeling approach is thus an appropriate tool for new mobility services, for which past data is often not readily available. This also presents a big challenge regarding calibration and validation of the models. One limitation of the majority of carsharing simulation studies is that they only address one day of simulation. However, studies show that the demand for carsharing change over different days, especially weekdays versus weekends. The study of Heilig et al. (2018) is the first to simulate carsharing for more than one day using the agent-based approach and therefore analyzes the variability of carsharing usage.

Although simulation gives a very detailed representation of the carsharing system in general, it only provides a snapshot of the situation. The long-term effects are not directly within the scope of the simulation. Regression models have been used to explain long-term demand (see e.g., Becker et al., 2017c; Schmöller et al., 2015). These models relate carsharing demand to service attributes, neighborhood/city characteristics and demographic factors. These models are at zonal spatial resolution.

Finally, data mining techniques are another method employed to identify spatiotemporal patterns of usage characteristics such as usage frequency, traveled distance, usage variability as well as characteristics of users (see e.g. Schmöller et al., 2015). Recently, advances in data methods and access to big data have made data mining techniques more popular.

Supply

The supply of carsharing services has a huge impact on the demand for these services. Despite the existence of very detailed carsharing simulation models, models to accurately represent supply are rare. In this section, the overall challenges in modeling the supply side are covered.

Optimization of infrastructure, that is, location and capacity of carsharing stations, and fleet management, that is, fleet size planning, relocation strategies, pricing and refueling/charging are the main objectives in a carsharing operation system to maximize the profitability of the service (Ferrero et al., 2018). Much of the literature on one-way systems (station-based or free-floating) is focused on vehicle relocation. Vehicle relocation is the transfer of cars from stations or areas (in the case of free-floating carsharing) with higher vehicle accumulation (unused cars) to stations or areas experiencing a vehicle shortage in order to improve the performance of services. Relocation models investigate the trade-off between cost and availability. Dynamic pricing policies can increase profits for operators as well as improve the balance of the system (Jorge et al., 2015). Determining the operating area is another important issue on the supply side.

The supply models are, in general, developed for one-way carsharing systems. Modeling the supply side of free-floating services can be particularly challenging since cars can be everywhere. The two major methodologies used in supply models are simulation (e.g., agent-based or discrete-event) and optimization (see e.g., Correia and Antunes, 2012), as well as a few mixed approaches (see e.g., Jorge et al., 2012). Most of the simulation models are at the disaggregate level of users and vehicles but do not necessarily provide optimal solutions, while optimization models provide optimal solutions but at an aggregate level (Nourinejad and Roorda, 2015).

A small number of studies investigate the supply-demand relation (see e.g., Ciari et al., 2016; Martínez et al. 2017). A future direction of research in the supply side would be to determine the optimal solutions based on the demand model. In the majority of supply studies, demand is given, and origins and destinations are considered random.

Gaps, Further Research and Final Words

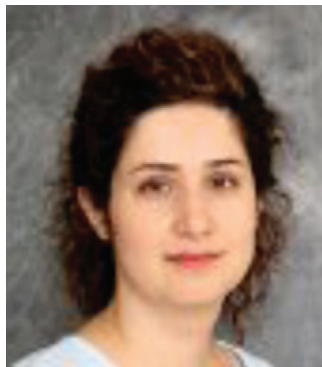
The main objective of this chapter is to highlight the importance of developing modeling tools for carsharing in the near future. Extensive literature exists on carsharing travel behavior but understanding its effects on the transportation system requires additional modeling efforts. Forecasting the impacts of these services on travel demand, transport network, and land use patterns requires the development of policy-sensitive models. The increasing popularity of these services, their expected future growth and the possible integration of shared autonomous vehicles into the picture make the need for such tools even more essential. The main obstacles to such models are lack of data, which is usually not shared by companies, as well as behavioral complexity. Using sophisticated models to answer complex planning questions requires large datasets incorporating extensive independent variables.

A further step in carsharing demand modeling is to integrate these models into larger travel forecasting model systems used for long-term transportation planning. To do so, travel surveys should explicitly include these services as distinct modes. Further development of disaggregate choice models that incorporate these services explicitly as choice options are necessary.

Further research is needed to address user short-term and mid-term response to uncertainty in availability of carsharing system as well as user preference for spontaneous booking versus prior booking.

To conclude, carsharing as other new mobility services will benefit from any methodological contribution to both supply and demand sides to verify its application as a transportation mode and its contribution to sustainable transport indicators.

Biographies



Shiva Habibi is a senior researcher at RISE - Research Institute of Sweden. She holds a Ph.D. in Transport Science from KTH - Royal Institute of Technology, Stockholm, Sweden. Her research aims to understand the decision-making process of different agents in transportation systems and develop models to predict their responses to different policy and marketing intervention. Her research interest include transport planning, transport demand modeling, transport policy as well as innovative mobility services.



Frances Sprei is an Associate Professor in Sustainable Mobility. She has a PhD in Energy and Environment from Chalmers University of Technology in Sweden. Her research assess different mobility services such as electric vehicles, carsharing, and autonomous vehicles and studies their adoption and diffusion. Her methods are interdisciplinary and she takes into consideration technical, economical and behavioral aspects.

References

- Balmer, M., Meister, K., Nagel, K., Axhausen, K.W., 2008. Agent-based simulation of travel demand: Structure and computational performance of MATSim-T. ETH. Eidgenössische Technische Hochschule Zürich, IVT Institut für Verkehrsplanung und Transportsysteme.
- Becker, H., Ciari, F., Axhausen, K.W., 2017a. Comparing car-sharing schemes in Switzerland: user groups and usage patterns. *Transport. Res. Part A: Policy Pract.* 97, 17–29.
- Becker, H., Loder, A., Schmid, B., Axhausen, K.W., 2017b. Modeling car-sharing membership as a mobility tool: a multivariate Probit approach with latent variables. *Travel Behav. Soc.* 8, 26–36.
- Becker, H., Ciari, F., Axhausen, K.W., 2017c. Modeling free-floating car-sharing use in Switzerland: a spatial regression and conditional logit approach. *Transport. Res. Part C: Emerging Technol.* 81, 286–299.
- Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D.S., Daly, A., 2002. Hybrid choice models: progress and challenges. *Market. Lett.* 13 (3), 163–175.
- Cervero, R., 2003. City CarShare: First-year travel demand impacts. *Transport. Res. Rec.* 1839 (1), 159–166.
- Cervero, R., Tsai, Y., 2004. City CarShare in San Francisco, California: second-year travel demand and car ownership impacts. *Transport. Res. Rec.* 1887 (1), 117–127.
- Ciari, F., Weis, C., Balac, M., 2016. Evaluating the influence of carsharing stations' location on potential membership: a Swiss case study. *EURO J. Transport. Logistics* 5 (3), 345–369.
- Ciari, F., Schuessler, N., Axhausen, K.W., 2013. Estimation of carsharing demand using an activity-based microsimulation approach: model discussion and some results. *Int. J. Sustain. Transport.* 7 (1), 70–84.
- Ciari, F., Balmer, M., Axhausen, K.W., 2008. Concepts for a large scale car-sharing system: Modelling and evaluation with an agent-based approach. Working paper/IVT, 517.
- Correia, G.H.d.A., Antunes, A.P., 2012. Optimization approach to depot location and trip selection in one-way carsharing systems. *Transport. Res. Part E: Logistics Transport. Rev.* 48 (1), 233–247.
- De Lorimier, A., El-Geneidy, A.M., 2013. Understanding the factors affecting vehicle usage and availability in carsharing networks: a case study of Communauto carsharing system from Montréal, Canada. *Int. J. Sustain. Transport.* 7 (1), 35–51.
- Ferrero, F., Perboli, G., Rosano, M., Vesco, A., 2018. Car-sharing services: An annotated review. *Sustain. Cities Soc.* 37, 501–518.
- Habib, K.M.N., Morency, C., Islam, M.T., Grasset, V., 2012. Modelling users' behaviour of a carsharing program: Application of a joint hazard and zero inflated dynamic ordered probability model. *Transport. Res. Part A: Policy Pract.* 46 (2), 241–254.
- Heilig, M., Mallig, N., Schröder, O., Kagerbauer, M., Vortisch, P., 2018. Implementation of free-floating and station-based carsharing in an agent-based travel demand model. *Travel Behav. Soc.* 12, 151–158.
- Jorge, D., Molnar, G., de Almeida Correia, G.H., 2015. Trip pricing of one-way station-based carsharing networks with zone and time of day price variations. *Transport. Res. Part B: Methodol.* 81, 461–482.
- Jorge, D., Correia, G., Barnhart, C., 2012. Testing the validity of the MIP approach for locating carsharing stations in one-way systems. *Procedia-Soc. Behav. Sci.* 54, 138–148.
- Juschten, M., Ohnmacht, T., Thao, V.T., Gerike, R., Hössinger, R., 2019. Carsharing in Switzerland: identifying new markets by predicting membership based on data on supply and demand. *Transportation* 46 (4), 1171–1194.
- Kim, J., Rasouli, S., Timmermans, H., 2017. Satisfaction and uncertainty in car-sharing decisions: An integration of hybrid choice and random regret-based models. *Transport. Res. Part A: Policy Pract.* 95, 13–33.
- Kopec, J.A., Esdaile, J.M., 1990. Bias in case-control studies. A review. *J. Epidemiol. Community Health* 44 (3), 179.
- Kopp, J., Gerike, R., Axhausen, K.W., 2015. Do sharing people behave differently? An empirical evaluation of the distinctive mobility patterns of free-floating car-sharing members. *Transportation* 42 (3), 449–469.
- Kortum, K., Machemehl, R.B., 2012. Free-floating carsharing systems: innovations in membership prediction, mode share, and vehicle allocation optimization methodologies (No. SWUTC/12/476660-00079-1). Southwest Region University Transportation Center, Center for Transportation Research, University of Texas at Austin.
- Kortum, K., Schönduwe, R., Stolle, B., Bock, B., 2016. Free-floating carsharing: city-specific growth rates and success factors. *Transport. Res. Procedia* 19, 328–340.
- Le Vine, S., Polak, J., 2015. Introduction to special issue: new directions in shared-mobility research. *Transportation* 42 (3), 407–411.
- Le Vine, S., Lee-Gosselin, M., Sivakumar, A., Polak, J., 2014. A new approach to predict the market and impacts of round-trip and point-to-point carsharing systems: case study of London. *Transport. Res. Part D: Transport Environ.* 32, 218–229.
- Martínez, L.M., Correia, G.H.d.A., Moura, F., Mendes Lopes, M., 2017. Insights into carsharing demand dynamics: outputs of an agent-based model application to Lisbon, Portugal. *Int. J. Sustain. Transport.* 11 (2), 148–159.
- Morency, C., Habib, K.M.N., Grasset, V., Islam, M.T., 2012. Understanding members' carsharing (activity) persistency by using econometric model. *J. Adv. Transport.* 46 (1), 26–38.
- Nourinejad, M., Roorda, M.J., 2015. Carsharing operations policies: a comparison between one-way and two-way systems. *Transportation* 42 (3), 497–518.
- Schmöller, S., Weikl, S., Müller, J., Bogenberger, K., 2015. Empirical analysis of free-floating carsharing usage: The Munich and Berlin case. *Transport. Res. Part C: Emerging Technol.* 56, 34–51.
- Shaheen, S., Cohen, A., Zohdy, I., 2016. Shared Mobility: Current Practices and Guiding Principles (No. FHWA-HOP-16-022). Federal Highway Administration, United States.

Further Reading

- Boyacı, B., Zografos, K.G., Geroliminis, N., 2015. An optimization framework for the development of efficient one-way car-sharing systems. *Eur. J. Oper. Res.* 240 (3), 718–733.
- Ciari, F., Balac, M., Axhausen, K.W., 2016. Modeling carsharing with the agent-based simulation MATSim: State of the art, applications, and future developments. *Transport. Res. Rec.* 2564 (1), 14–20.
- Cervero, R., Golub, A., Nee, B., 2007. City CarShare: longer-term travel demand and car ownership impacts. *Transport. Res. Rec.* 1992 (1), 70–80.
- Dias, F.F., Lavieri, P.S., Garikapati, V.M., Astroza, S., Pendyala, R.M., Bhat, C.R., 2017. A behavioral choice model of the use of car-sharing and ride-sourcing services. *Transportation* 44 (6), 1307–1323.
- Efthymiou, D., Antoniou, C., Waddell, P., 2013. Factors affecting the adoption of vehicle sharing systems by young drivers. *Transport Policy* 29, 64–73.
- Jorge, D., Correia, G., 2013. Carsharing systems demand estimation and defined operations: a literature review. *Eur. J. Transport Infrastructure Res.* 13 (3).
- Jorge, D., Correia, G.H., Barnhart, C., 2014. Comparing optimal relocation operations with simulated relocation policies in one-way carsharing systems. *IEEE Trans. Intell. Transport. Syst.* 15 (4), 1667–1675.
- Kim, J., Rasouli, S., Timmermans, H.J., 2017. The effects of activity-travel context and individual attitudes on car-sharing decisions under travel time uncertainty: A hybrid choice modeling approach. *Transport. Res. Part D: Transport Environ.* 56, 189–202.
- Le Vine, S., 2011. Strategies for personal mobility: A study of consumer acceptance of subscription drive-it-yourself car services.
- Le Vine, S., Polak, J., 2017. The impact of free-floating carsharing on car ownership: Early-stage findings from London. *Transport Policy*.
- Li, Q., Liao, F., Timmermans, H.J., Huang, H., Zhou, J., 2018. Incorporating free-floating car-sharing into an activity-based dynamic user equilibrium model: A demand-side model. *Transport. Res. Part B: Methodol.* 107, 102–123.
- Millard-Ball, A., 2005. Car-sharing: Where and How it Succeeds, vol. 108. Transportation Research Board.
- Shaheen, S.A., Chan, N.D., Micheaux, H., 2015. One-way carsharing's evolution and operator perspectives from the Americas. *Transportation* 42 (3), 519–536.
- Vosooghi, R., Puchinger, J., Jankovic, M., Sirin, G., 2017. October. A critical analysis of travel demand estimation for new one-way carsharing systems. In: In 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC) (199–205). IEEE.
- Weikl, S., Bogenberger, K., 2013. Relocation strategies and algorithms for free-floating car sharing systems. *IEEE Intell. Transport. Syst. Mag.* 5 (4), 100–111.

Urban Recreational Travel

Long Cheng, Frank Witlox, Department of Geography, Ghent University, Ghent, Belgium

© 2021 Elsevier Ltd. All rights reserved.

Definition	629
Travel Patterns	630
Temporal and Spatial Distribution	630
Travel Mode Choice Characteristics	630
Influencing Factors	630
Recreation Demand	630
Human Factors	630
Social Influences	630
Government Policies	631
Economic Environment	631
Climate and Atmospheric Impacts	631
Recreation Supply	631
Transport	631
Benefits and Risks	632
Social and Economic Benefits	632
Health and Well-Being	632
Environmental Impacts	632
ICT and Recreational Travels	632
Substitution	632
Generation	633
Modification	633
References	633
Further Reading	634

Definition

The *Macquarie Dictionary* (2017) refers to “recreate” as “to refresh by means of relaxation and enjoyment, as after work; to restore or refresh physically or mentally.” Yukic (1970, p. 5) states that “recreation is an act or experience, selected by the individual during his/her leisure time, to meet a personal want or desire, primarily for his/her own satisfaction.” Mercer (2003, p. 412) shows that “recreation refers to activities, either active or passive, enjoyed either outdoors or indoors, which take place during leisure—as opposed to non-work—time.” In addition, the Fifth Edition of *The Dictionary of Human Geography* (Gregory et al., 2009, p. 624) defines recreation as “pursuits or activities (even inactivity) undertaken voluntarily outside paid employment for the primary purpose of pleasure, enjoyment and satisfaction.”

Confusion may arise over the use of the terms “recreation” and “leisure,” which are closely related and often used interchangeably. The simplest difference identifies recreation with activity and leisure with time. Recreation is an activity, for example, play, that amuses, stimulates, or diverts while leisure emphasizes the freedom in time obtained through the cessation of activities. In other words, you could be at leisure, but not necessarily engage in recreational activities. For example, you could just sit somewhere and do nothing. In addition, we need to distinguish between “recreation” and “tourism.” Tourism is commonly discussed in the context of recreation and leisure: the majority of tourism is recreational, in that tourism takes place during leisure time and often for the purpose of personal satisfaction and pleasure. Britton (1979) makes a distinction that recreational travel inclines to be private, self-service engaging, whereas tourism concerns services provided by diverse agencies. Gunn and Var (2002) assign an emphasis on profit-making and economic aspects to tourism, while link recreation basically with noncommercial purposes.

Recreational travel involves travel for recreation. Urban recreational travel shares with commuting travel regarding short-term and reversible characteristics. They differ from each other in that commuting travels take place, for most people, at fixed times on each working day, and between fixed origins and fixed destinations, whereas recreational travels have only fixed origins—the home in most cases, and in every other aspect they are discretionary. The 2017 US National Household Travel Survey (NHTS) (US Federal Highway Administration, 2019) classifies social/recreational trip purposes into three categories: performing recreational activities (e.g., visit bars, parks, museums, or movies), exercising (e.g., walk, walk the dog, go for a jog, or go to the gym), and buying meals (e.g., go out for meals, snacks, or carry-out).

Urban recreational travel is by no means an insignificant segment of total trips made by citizens. In many cities, recreational travel accounts for a relatively high proportion of total individuals’ trips, from one-third to one-half (Anable, 2002; Ohnmacht

et al., 2009). The 2017 US NHTS Report (US Federal Highway Administration, 2019) shows that trips for social/recreational activities increased in proportion every year, rising from 22% to 32% of the total discretionary trips between 1990 and 2017. The European Council of Ministers of Transport (European Conference of Ministers of Transport, 2000) observes that growth in recreational travel (and activities) could be related to three main reasons: increasing living standards, earlier retirement, and the trend toward shorter working hours. Therefore, with the continued rise in individuals' leisure time and economic prosperity worldwide, the demand for recreational travel is expected to increase (Misra and Bhat, 2000).

Travel Patterns

Temporal and Spatial Distribution

Essential to the recreation experience is the freedom of selecting the activity, location, and time. Recreational travel starts any time during the day, with the majority occurring during the off-work time. Certain definite temporal patterns on an hourly, daily, and seasonal basis can be identified, depending on the specific activities and their spatial locations relative to individuals' place of residences. The spatial distribution of trips is determined by the location of recreation attractions and the location of individuals' homes. The place of production is essentially an individual's home; the place of attraction usually refers to the recreation location. An exception is possible, for instance, pleasure driving, with no single point of attraction. Travel distance is characterized by the time available and cannot exceed a certain maximum.

Both numbers of recreational trips and distance traveled vary by day of the week, with much more trips and longer distances of travel on the weekend. An analysis of New Zealand Household Travel Surveys 2003–6 shows that 40% of social and recreational trips and 47% of the distance traveled take place on the weekend (Witten and Mavoa, 2011). According to the 2017 US NHTS Report (US Federal Highway Administration, 2019), there was not a clearly discernible trend in terms of trip distance or duration over the past decades: social/recreational trips remain almost the same (averaged at 8.6 miles (13.8 km)) in trip length but vary in duration (20.8 min in 2001, 19.7 min in 2009, and 22.2 min in 2017). But temporal distribution by the time of day keeps consistent: social/recreational travel peaks at the 12 a.m. and 6 p.m. hours.

Travel Mode Choice Characteristics

It is observed that the car is commonly used for recreational travel (McGuckin and Fucci, 2017); it offers the most freedom in selecting any destination and in timing a trip, and provides a relatively high degree of comfort. The use of public transit depends on different criteria: timing of a trip is an important one because public transit is operated on a fixed schedule. This is usually acceptable for attending sport/concert events, and other schedule-bound activities, or activities confined in a limited area such as low emission zones. Another criterion is the convenience of carrying equipment. Public transit will not be a good option unless the equipment is packed in a way that is easily carried. The recreational activity itself is also a criterion in the choice of travel mode. Sightseeing, for instance, leads to riding public transit. Sightseeing tours in cities are well known and might attract people from those who are fed up with driving.

Travel mode choice characteristics are different, contextualized in the physical and social environments of a country/region. An analysis of New Zealand Household Travel Surveys 2003–6 (New Zealand Ministry of Transport, 2007) shows that 45% of trips were made as car drivers and 32% as car passengers, 19% by walking and 2% by bicycle. According to the 2017 US NHTS (McGuckin and Fucci, 2017), American people conducted 77.1% of social/recreational trips by private vehicles, 1.6% by public transit, and 18.1% by walking. As expected, people living in households with few vehicles reported few social/recreational trips. However, residents in Nanjing (China) were more apt to use walking (77%) and public transit (10%) for recreational trips, on the basis of data analysis from 2013 Nanjing Travel Survey (Nanjing Institute of City and Transport Planning, 2013).

Influencing Factors

Recreation Demand

Human Factors

Personality and cultural background are of significant relevance to recreation demand (Johnson et al., 2001). Kaplan (1960) observed that residents in the same town with different cultural backgrounds (e.g., Africans and Jews) conducted different recreational activities. Africans tended to chat with friends and hangout near the neighborhood while Jews participated more in reading, listening to music, and other intellectual activities. Individual's sociodemographics (including age, gender, income, household structure, etc.) also exert influences on activity participation. Liu et al. (2008) found that the elderly people seem to be more involved with passive activities (e.g., playing chess and cards, chatting), while the young people are more likely to conduct active ones (e.g., sporting, singing).

Social Influences

Social pressure (or peer pressure) is an important factor selecting recreational activities (Shaw et al., 1996). If playing tennis is the current popularity, many people will tend to join tennis clubs. If attending cultural events is "the thing to do" in the neighborhood,

many people will follow doing the same since an immediate neighborhood usually has a common social background. In addition, the social environment might also influence recreation participation. Social environment refers to the neighborhood sociodemographic composition (e.g., unemployment rate, the percentage of low-income people) and social ties (e.g., trust, cohesion) of people living in the neighborhoods (Wang and Lin, 2013). Research findings in the field of sociology suggest that the social environment has strong impacts on individuals' behaviors, for example, the intensity of physical travel (Zastrow and Kirst-Ashman, 2006). Studies in public health also indicate that people residing in an environment with intensive social interactions are more likely to gather regularly with neighbors and friends (Leyden, 2003).

Government Policies

Government has an interest in how its residents make use of their free time because recreation is important to their quality of life. Government policies including educational efforts, financial subsidy, and distribution of free time are important to recreational activities (Leonhardt, 1971). School programs are able to influence recreation participation, for example, a tennis program can contribute a great deal to the growing number of tennis players. Financial subsidies for developing recreational areas and enhancing their accessibility can improve an area's attractiveness. Government policies also impact the distribution of free time. Creating several 3-day weekends when local holidays are celebrated on a Friday or Monday will increase the number of recreational activities. The proposal of "flexible work schedule" may also shift the way people spend their free time.

Economic Environment

Recreation cost and personal income are an individual's primary economic considerations. Clawson and Knetsch (2013) identified positive correlations between disposable personal income and recreation participation. Directly related to income is the cost of participation. For example, low-income people cannot afford high transport costs. User fees of recreational amenities are subject to the type of activity and the ownership of recreational facilities. Public facilities, for example, parks, are usually offered at low cost or free of charge, while privately owned areas operate at market price, for example, golf courses.

Climate and Atmospheric Impacts

The type and amount of seasonal participation depend largely on climatic factors (Leonhardt, 1971). Too hot or cold/wet seasons are not attractive for outdoor recreational activities; the preferable outdoor recreation season is spring or fall in most countries. The impact of weather is also powerful on recreation participation (Smith, 1993), in particular at weekends. The longer the planning period, the less influential the weather on the determination of recreation demand. Air pollution, for example, smog, also affects individuals' outdoor recreation.

Recreation Supply

The role of urban form, land use, and the physical aspect of the living environment—the built environment—in shaping travel demand has consistently been a topic of great interests in travel behavior analysis. Ewing and Cervero (2010) suggest that the following 5Ds of the built environment affect travel: density, diversity, design, destination accessibility, and distance to public transport. A review of 17 studies shows that walking as a travel mode of recreational trips is strongly associated with living in high-density residential neighborhoods and close proximity to destinations (Saelens and Handy, 2008). The type of recreational supply available with the urban area depends on not only the geographic environment (e.g., flat/hilly surfaces, sunny/shaded sidewalks, green spaces and parks), but also man-made facilities (e.g., chess/card rooms, basketball courts). Quite several studies document the connection between the number of recreational facilities and the level of participation in recreation. For example, Li et al. (2005) conducted a survey of 500 elderly people from 56 neighborhoods in Portland, Oregon, observing that the area of green or open spaces and the number of recreation facilities are highly associated with levels of walking.

Transport

Recreation demand and supply are usually generated at different spatial locations. To link them, some means of transport is needed. Surface transport—car, public transit (bus and rail), walking—are the predominant modes. The ability to provide accessibility is an important characteristic of transport. The better the transport service, the more intense the interchange of demand and supply. Enhanced transport service facilitates and encourages participation in recreational activity.

Car plays a leading role in recreational travel (McGuckin and Fucci, 2017). Cars allow for the flexibility in the selection of location and the timing of recreational trips. They provide the convenience of carrying large amounts of baggage and equipment from "door to door" with the minimum trouble for recreationists. Yet, recreation sites need to build enough parking lots, which sometimes is difficult due to space limitations and/or environmental considerations. Public transit may deal with part of this problem; it will not have the annoying parking problems. However, a disadvantage is that public transit limits the freedom of choice of when and where to go, which is of much significance to recreational travels. Still, public transit can play a more important role in recreational travel if it is combined with cars. "Park & ride" service is an option, which is particularly useful in the densely traveled corridors between urban regions and major recreation attractions.

Benefits and Risks

Social and Economic Benefits

Donnelly and Coakley (2002) conclude the positive effects of recreation on social inclusion. Recreational travel and activity offer face-to-face contact/interaction with others and ensure that individuals have opportunities to identify common interests as well as chances to have fun together. In addition, recreational travel and activity can play a significant role in local urban economies (US National Recreation and Park Association, 2018). They contribute to economic prosperity in that related agencies employ hundreds of thousands of people and their operations and capital spending create considerable economic activities.

Health and Well-Being

There are clear evidences confirming the connections between recreation and positive health outcomes (Godbey, 2009). For example, walking, a common travel mode of recreational trips, is associated with a lower risk of heart disease, managing weight, boosting good cholesterol, preventing depression, elevating overall mood, and sense of well-being (Colditz et al., 1997). Stress reduction seems to be another major benefit: negative moods decrease after recreation, and recreationists report lower levels of sadness and anxiety during recreational travel and activity (Godbey et al., 1992). Grahn and Stigsdotter (2003) noted significant relationships between stress reduction and the visits of urban green spaces; the more often an individual visits urban green spaces, the less often he or she reports self-perceived illness. This result is applicable to the entire population, regardless of respondents' gender, age, or socioeconomic status. Moreover, De Vos (2019) concluded that the experience of leisure/recreational travel—the mood during travel and the evaluation of the trip—could contribute to one's life satisfaction.

Environmental Impacts

Automobiles are of growing importance in urban recreational travel, contributing to the increased car ownership/use and greenhouse gas emissions. Over the last decade 2005–15, the number of passenger cars (per 1000 inhabitants) in the world has grown by 20% (Eurostat, 2018). By far the fastest growth in the ratio of passenger cars to population was recorded in China, as this ratio was more than 6 times as high as in 2015 (around 100 passenger cars for every 1000 inhabitants) as in 2005. However, North America and the European Union continued to outpace most other regions in the world regarding the density of car ownership. Ownership of passenger cars in Canada was more than 600 passenger cars per 1000 inhabitants in 2015 while the EU had a density of 500.

Traffic congestion resulting from participation in recreational activities such as concerts, festivals, and sporting events is a frequent occurrence in many cities. The 2017 US NHTS shows that recreational (including social) trips account for 30% of all urban discretionary trips on weekdays. This percentage increases to around 32% on weekends when more recreational trips are performed. The NHTS also shows that (social and) recreational trips account for a substantial 27% of annual person miles of travel in the United States; this figure is 29% for New Zealand (Witten and Mavoa, 2011). Given that recreational trips are mainly made by means of car in many countries, urban recreational travel contributes a great deal to the number of passenger vehicle miles of travel on urban streets today. For instance, the NHTS data show that the annual vehicle miles of travel for recreation are 4825.5 miles (7765.9 km), 65% higher than the annual vehicle miles of travel for shopping. This contribution is only likely to get larger in the future given that disposable income and leisure time continue to grow, and as the sociodemographic characteristics of the population change over time (Misra and Bhat, 2000).

ICT and Recreational Travels

In the era of mobile Internet, the growing use of information and communication technology (ICT) has profound impacts on recreational travel. Three types of effects could be identified: substitution, generation, and modification (for a review, see Mokhtarian et al., 2006).

Substitution

ICT may provide an alternative means to perform recreational activities. If ICT-based forms of activities are chosen over location-based forms, travel is expected to be reduced. The degree of location- and time-independence of activities may impact the likelihood of being replaced by ICT-based forms. Location- and/or time-dependent activities are less likely to be substituted by ICT. Some activities are spatial-fixed (e.g., camping/backpacking or mountain climbing) or temporal-fixed (e.g., family visits on certain holidays), and hence not readily be affected by ICT. Other activities, such as playing cards and watching movies, are location- and time-independent, which makes them more likely to be influenced by ICT.

ICT may also generate new activities, for example, watching movies on the Internet. If people spend more time on ICT-based activities, it is likely that they spend less time on non-ICT-based activities (Nie et al., 2002). Even though the displacement effect could be conscious and prompt, it might also happen subconsciously and over longer time. Due to the fact that the abandoned recreational activities involve travel, this effect might also reduce travel.

Generation

Rose et al. (2009) argue that ICT is not a reasonable replacement for recreational travel since face-to-face interactions with family and friends are essential. In contrast, the use of ICT facilitates activity and travel generation (Adams, 2005; Miller, 2006). For instance, the Internet allows a last-minute reservation that was not easy previously. By providing readily available information about a variety of activity and travel options, ICT prompts making the arrangements for recreational trips. On the other hand, ICT might reduce the time and/or cost required to perform a certain activity and associated travel, with the saved time or cost used for conducting other activities. For example, the travel time saved by videoconference might be spent on recreational activities. The money saved by group buying may be spent on other recreational activities. If the new activities involve longer travels, the reallocated time to other activities derives more demand for travels.

Modification

The impact of ICT on travel might also be modification in some cases. For example, a mobile phone call is likely to change the planned travel route, and its consequences for travel distance, timing, and duration. ICT also has direct influences on travel behavior of other individuals, including generation and reduction effects. For example, arranging a social activity on the fly would not have occurred without the mobile phone (generation); mobile navigation system helps to avoid congested road (reduction).

References

- Adams, J., 2005. Hypermobility: a challenge to governance. In: Lyall, K., Tait, J. (Eds.), *New Modes of Governance: Developing an Integrated Policy Approach to Science, Technology Risk and the Environment*, Routledge, London, pp. 123–139.
- Anable, J., 2002. Picnics, pets and pleasant places: the distinguishing characteristics of leisure travel demand. In: Black, W., Nijkamp, P. (Eds.), *Social Change and Sustainable Transport*, University Press, Bloomington, IN, pp. 181–190.
- Britton, R., 1979. Some notes on the geography of tourism. *Can. Geogr.* 23 (3), 267–282.
- Clawson, M., Knetsch, J.L., 2013. *Economics of Outdoor Recreation*. RFF Press, New York.
- Colditz, G.A., Manson, J.E., Hankinson, S.E., 1997. The nurses' health study: 20-year contribution to the understanding of health among women. *J. Womens Health* 6 (1), 49–62.
- De Vos, J., 2019. Analysing the effect of trip satisfaction on satisfaction with the leisure activity at the destination of the trip, in relationship with life satisfaction. *Transportation* 46 (3), 623–645.
- Donnelly, P., Coakley, J.J., 2002. *The Role of Recreation in Promoting Social Inclusion*. Laidlaw Foundation, Toronto.
- European Conference of Ministers of Transport (ECMT), 2000. *Transport and Leisure: Round Table 111*. ECMT, Paris.
- Eurostat, 2018. *The EU in the World—2018 Edition*. European Commission Publication, Luxembourg.
- Ewing, R., Cervero, R., 2010. Travel and the built environment: a meta-analysis. *J. Am. Plan. Assoc.* 76 (3), 265–294.
- Godbey, G., 2009. Outdoor recreation, health, and wellness: understanding and enhancing the relationship. RFF Discussion Paper No. 09-21. Available from: <http://dx.doi.org/10.2139/ssrn.1408694>.
- Godbey, G., Graefe, A.R., James, S.W., 1992. *The Benefits of Local Recreation and Park Services: A Nationwide Study of the Perceptions of the American Public*. National Recreation and Park Association, Arlington, VI.
- Grahn, P., Stigsdottir, U.A., 2003. Landscape planning and stress. *Urban Forest. Urban Greening* 2 (1), 1–18.
- Gregory, D., Johnston, R., Pratt, G., Watts, M., Whatmore, S. (Eds.), 2009. *The Dictionary of Human Geography*, fifth ed. John Wiley & Sons, Inc., New York.
- Gunn, C.A., Var, T., 2002. *Tourism Planning: Basics, Concepts, Cases*. Psychology Press, Hove, UK.
- Johnson, C.Y., Bowker, J.M., Cordell, H.K., 2001. Outdoor recreation constraints: an examination of race, gender, and rural dwelling. *South. Rural Sociol.* 17, 111–133.
- Kaplan, M., 1960. *Leisure in America: A Social Inquiry*. John Wiley & Sons, Inc., New York.
- Leonhardt, K., 1971. *Weekend Recreation Travel: Development of a Concept*. Puget Sound Governmental Conference, Seattle, WA.
- Leyden, K.M., 2003. Social capital and the built environment: the importance of walkable neighborhoods. *Am. J. Public Health* 93 (9), 1546–1551.
- Li, F., Fisher, K.J., Brownson, R.C., Bosworth, M., 2005. Multilevel modelling of built environment characteristics related to neighbourhood walking activity in older adults. *J. Epidemiol. Commun. Health* 59 (7), 558–564.
- Liu, H., Yeh, C.K., Chick, G.E., Zinn, H.C., 2008. An exploration of meanings of leisure: a Chinese perspective. *Leis. Sci.* 30 (5), 482–488.
- McGuckin, N., Fucci, A., 2017. *Summary of Travel Trends: 2017 National Household Travel Survey*. Federal Highway Administration, Department of Transportation, Washington, DC.
- Mercer, D.C., 2003. Recreation. In: Jenkins, J.M., Pigram, J.J. (Eds.), *Encyclopedia of Leisure and Outdoor Recreation*, Routledge, London.
- Miller, H.J., 2006. Social exclusion in space and time. In: Axhausen, K.W. (Ed.), *Moving Through Nets: The Physical and Social Dimensions of Travel—Selected Papers from the 10th International Conference on Travel Behaviour Research*. Elsevier, Oxford, pp. 353–380.
- Misra, R., Bhat, C., 2000. Activity-travel patterns of nonworkers in the San Francisco Bay area: exploratory analysis. *Transp. Res. Rec.* 1718 (1), 43–51.
- Mokhtarian, P.L., Salomon, I., Handy, S.L., 2006. The impacts of ICT on leisure activities and travel: a conceptual exploration. *Transportation* 33 (3), 263–289.
- Nanjing Institute of City and Transport Planning, 2013. *Nanjing Travel Survey 2013 Dataset*. NICTP, Nanjing, China.
- New Zealand Ministry of Transport, 2007. *Household Travel Surveys 2003–2006 Datasets*. Ministry of Transport, Wellington.
- Nie, N.H., Hillygus, D.S., Erbring, L., 2002. Internet use, interpersonal relations, and sociability: a time diary study. In: Wellman, B., Haythornthwaite, C. (Eds.), *The Internet in Everyday Life*, Blackwell Publishing, Malden, MA, pp. 215–243.
- Ohnmacht, T., Götz, K., Schad, H., 2009. Leisure mobility styles in Swiss conurbations: construction and empirical analysis. *Transportation* 36 (2), 243–265.
- Rose, E., Witten, K., McCreanor, T., 2009. Transport related social exclusion in New Zealand: evidence and challenges. *Kōwhiri: New Zealand J. Social Sci. Online* 4 (3), 191–203.
- Saelens, B.E., Handy, S.L., 2008. Built environment correlates of walking: a review. *Med. Sci. Sports Exercise* 40 (7), S550–S566.
- Shaw, S.M., Caldwell, L.L., Kleiber, D.A., 1996. Boredom, stress and social control in the daily activities of adolescents. *J. Leis. Res.* 28 (4), 274–292.
- Smith, K., 1993. The influence of weather and climate on recreation and tourism. *Weather* 48 (12), 398–404.
- The Macquarie Dictionary, 2017. *The Macquarie Dictionary*, seventh ed. The Macquarie Library, Sydney.
- US Federal Highway Administration, 2019. *National Household Travel Survey Report: Trends in Discretionary Travel*. Department of Transportation, Washington, DC.
- US National Recreation and Park Association, 2018. *Promoting Parks and Recreation's Role in Economic Development*. Centre for Regional Analysis, The George Mason University, Fairfax, VI.

- Wang, D., Lin, T., 2013. Built environments, social environments, and activity-travel behaviour: a case study of Hong Kong. *J. Transp. Geogr.* 31, 286–295.
- Witten, K., Mavoa, S., 2011. Social and recreational travel: the destinations, travel modes and CO₂ emissions of New Zealand households. *Social Policy J. New Zealand* 37, 1–13.
- Yukic, T.S., 1970. *Fundamentals of Recreation*, second ed. Harper & Row, New York.
- Zastrow, C., Kirst-Ashman, K., 2006. *Understanding Human Behaviour and the Social Environment*. Cengage Learning, Boston, MA.

Further Reading

- De Vos, J., Schwanen, T., Witlox, F., 2017. The road to happiness: from obtained mood during leisure trips and activities to satisfaction with life. Presented at the 2017 World Symposium on Transport and Land Use Research (WSTLUR), Brisbane, Australia.
- Gosens, T., Rouwendal, J., 2018. Nature-based outdoor recreation trips: duration, travel mode and location. *Transp. Res. Part A Policy Pract.* 116, 513–530.
- Jenkins, J., Pigram, J., 2007. *Outdoor Recreation Management*, second ed. Routledge, London.
- Keith, J., Fawson, C., Chang, T., 1996. Recreation as an economic development strategy: some evidence from Utah. *J. Leis. Res.* 28 (2), 96–107.
- Leitner, M.J., Leitner, S.F., 2004. *Leisure Enhancement*, third ed. Haworth Press, New York.
- Page, S.J., Hall, C.M., 2014. *The Geography of Tourism and Recreation: Environment, Place and Space*, fourth ed. Routledge, London.
- Pozsgay, M.A., Bhat, C.R., 2001. Destination choice modelling for home-based recreational trips: analysis and implications for land use, transportation, and air quality planning. *Transp. Res. Rec.* 1777 (1), 47–54.
- Prosser, R., 2000. *Leisure, Recreation and Tourism*. Collins, London.
- Thompson, B., 1967. Recreational travel: a review and pilot study. *Traffic Q.* 21 (4), 527–542.

Place Perception and Travel Behavior

Kathleen Deutsch-Burgner*, **Konstadinos G. Goulias***, *Data Perspectives Consulting, Santa Barbara, CA, United States; †Department of Geography, GeoTrans Laboratory, University of California Santa Barbara, Santa Barbara, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	635
Space, Place, Scale, and Sense of Place	635
Quantifying Sense of Place for Centers of Activity	636
Quantifying Citywide Sense of Place	638
Closing Remarks	639
References	640
Further Reading	641

Introduction

During the course of our lives we all develop various likes and dislikes. We develop favorite movies, foods, colors, and beaches, ski resorts, and hikes. We can have favorite material possessions, favorite relationships or people, and favorite places! On the other hand, we also develop adverse reactions to places, artifacts, and people we do not like. In the process of developing favorites, not so favorites, and the in-betweens, each person uses his or her experiences with the place, artifact, or person in question. When faced with a choice, humans consider the options, their attributes, and a variety of other factors to arrive at a decision. We name all these factors motivations along similar strands of thinking as in [Mokhtarian et al. \(2015\)](#), in other travel behavior literature. Among these motivations are the attitudes and preferences that have been formed about the choice at hand as well as other attitudes that may not be strictly related to the object of the choice ([Wang, 2015](#)). Here we focus on how we perceive places and how these perceptions correlate with how we travel.

More recent advances in travel behavior modeling and simulation have greatly increased our ability to successfully understand, analyze, and predict travel behavior as the derived demand for activity participation. This approach is dubbed the activity-based approach to travel demand forecasting ([Axhausen and Gärling, 1992](#); [Bhat and Koppelman, 1999](#); [Ettema and Timmermans, 1997](#); [Kitamura, 1988](#)). Activities are pursued in locations, and they are strongly correlated with the types of land uses at these locations, creating the necessity to integrate land use patterns with travel behavior ([Goulias, 1996](#); [Timmermans, 2003](#)). Analytical advancements have also increased the accuracy and reliability of prediction and simulation efforts and offer new opportunities to incorporate previously unmeasured behavioral determinants. These include integration of latent constructs in discrete choice methods ([Ben-Akiva et al., 2002](#)), enriching traditional models with personality indicators ([Goulias and Henson, 2006](#); [Prevedouros, 1992](#); [Steg et al., 2001](#)), account for social networks ([Arentze and Timmermans, 2008](#); [Carrasco and Miller, 2006](#)), incorporate activity scheduling and planning intentions ([Auld et al., 2008](#); [Ruiz and Roorda, 2008](#)), expanding the envelope of modeling behavior in hierarchies of lifestyles ([Van Acker et al., 2014](#)), and accounting for subjective well-being ([Mokhtarian, 2019](#); [Wang and Wang, 2016](#)). Attitudinal variables can enhance the explanatory power of behavioral models and are correlated with urban design ([Kitamura et al., 1997](#)). Moreover, spatial perception and perceived attributes are recognized as an important factor in understanding travel behavior ([Adamowicz et al., 1997](#); [Ahuja et al., 2018](#); [Goulias et al., 1998](#); [Golledge and Garling, 2003](#); [Hoehner et al., 2005](#)). Evidence is also increasing about the need to combine built environment metrics with psycho-attitudinal metrics about places and perceptions of the environment ([Clifton and Perez, 2015](#); [Hannes et al., 2012](#)).

Space, Place, Scale, and Sense of Place

In this chapter we distinguish between space and place following a typical geographic definition that defines place as a qualified space with distinct spatial attributes. With this in mind, travel behavior analysis has modeled the influence of places on behavior using measured spatial attributes such as accessibility and access to facilities ([Geurs and Van Wee, 2004](#); [Handy and Niemeier, 1997](#)). These spatial attributes are summarized in indicators that include the attractive features of a place and the difficulty to reach them ([de Abreu et al., 2012](#); [Lin et al., 2014](#)). These metrics are then used to define attractiveness of a place as explanatory variables of daily behavior. In activity-based models we often use availability of opportunities for different types of places including travel time by different modes to reach opportunities and time of day availability of these opportunities ([Chen et al., 2011](#)). Individuals' perceptions of places reflect these measurable attributes of places, as a starting point for making destination choice models that better reflect human spatial decision making ([Pike and Ryan, 2004](#), and a chapter in this encyclopedia). This is not enough to understand behavior because people perceive attributes of these locations in different ways and because they attribute meanings to places that change in their life course. The nature of the meanings assigned to places is still debated ([Agarwal, 2018](#)). However, a

group of meanings in the form of attitudes (beliefs coupled with affective stances) have been tested as a summary of these meanings and named *Sense of Place*. The term *Sense of Place* (also defined as the powerful connection between people and places or the affective ties of people with places) has a long history in human geography, and the theory generally operates in realms that are deeply personal and social but invisible. Numerous efforts to pin down specific aspects of places are found in the geography and environmental psychology literature (Proshansky, 1978; Stedman, 2003; Stokols and Shumaker, 1981). Similarly, in the tourism research literature one aspect of *Sense of Place* called place image is used for marketing destinations (Campelo et al., 2014; Chon, 1990; Elliot et al., 2011) and recognize the importance of correlating tourists and residents *Sense of Place* for design of destinations (Kianicka et al., 2006; Styliadis et al., 2016).

Quantifying *Sense of Place* with the specific goal to include it as a determinant of daily travel behavior can be done through the use of attitudinal questions in human surveys and examination of their correlation with travel behavior (Chen and Sekar, 2018; Deutsch and Goulias, 2010). This quantified sense of place can also be used in land use policy implementation to identify resistance to change due to *Sense of Place* but also to design and improve urban design quality (Hu and Chen, 2018). A place can be a house, a store, a neighborhood, a section of a city, or even an entire region or country making quantification of sense of place and the formation of metrics scale dependent (Montello, 1993). *Sense of Place* is generally only studied for particularly meaningful places like the home, but studying place at broader scales can greatly expand our understanding of people's day-to-day activities outside the home. In translating quantified *Sense of Place* to policies we also need to develop metrics at the appropriate scale. In this chapter we review quantification of place perception and *Sense of Place* at two different scales.

Quantifying Sense of Place for Centers of Activity

Quantification of sense of place was done in a Santa Barbara, CA, experiment using 719 persons from an intercept survey conducted at two outdoor shopping centers (malls). Both malls are situated along one of Santa Barbara's main arterial streets, State Street, one being downtown (Paseo Nuevo) and one uptown (La Cumbre). At the time of the study, both malls offered two anchor big box stores, and a variety of smaller retail stores. In addition, both places offer restaurants and additional nearby shopping. Details of the study areas and data collection process and instrument are discussed in Deutsch and Goulias (2010). The survey consisted of questions developed to capture the respondents' sense of place views toward each mall. Following work conducted by Jorgensen and Stedman (2001) and Stedman (2003) in environmental psychology a series of questions were adapted to this study which were focused on the respondents' place attachment, place identity, and place dependence. In addition, several additional questions were included to examine other aspects of sense of place. All sense of place questions used a seven-point Likert scale, ranging from strongly disagree to strongly agree. Additionally, several socioeconomic and demographic questions were included as well as questions about travel behavior before, during, and after the visit to the shopping mall to capture the role of visiting a specific place in the context of daily behavior. Because of the intercept format of the survey, travel behavior variables are limited to a small snapshot of the day's activities. In this work, we found that three additional constructs create a more complete sense of place representation, which are correlated with individual characteristics and travel behavior indicators (Fig. 1). The additional constructs are based on questions relating to one's satisfaction with the place, a construct of sense of place previously developed but not as widely known (Stedman, 2003), as well as questions about the physical/aesthetic surroundings of the place and the social environment. Researchers have discussed the importance of physical, social, and cultural attributes of place in developing a sense of place (Stokols and Shumaker, 1981). The inclusion of these indicators in a factor analysis provides insight into how these concepts manifest themselves in everyday activity locations, and which place attributes are more apparent in everyday settings. In this example six factors (attachment, dependence, identity, satisfaction, atmosphere, and community) represent sense of place. Each factor is measured by three to six indicators (answers to the attitudinal questions), and the factors are explained by sociodemographic variables and characteristics of the day. In this comparison between two shopping malls in Santa Barbara we also find that: (1) places with similar affordances have different sense of place meaning to patrons; (2) different travel behavior facets are influenced in different ways by sense of place; and (3) sociodemographics are correlated the *Sense of Place* factors, but they are not able to exhaustively map differences among people and therefore cannot substitute sense of place metrics. Different socio-demographic factors influence sense of place at the two sites. For example, marital status and income are important factors at Paseo Nuevo, as are gender and marital status at La Cumbre. For both sites, important explanatory variables include income, age, gender, ethnicity, and ownership of a bus pass and possession of a driver's license. Several *Sense of Place* indicators influence travel behavior variables. Of note, the use of a car is associated with lower place dependence and satisfaction, but a higher place identity. Additionally, walking is associated with high ranking of a place's atmosphere. However, causality should not be inferred in this case—Do those who have a positive view of the atmosphere choose to walk, or do those who walk take notice and view the surrounding atmosphere more positively? Additionally, walkers have a lower attachment score. Atmosphere was again significant for those who have no companion in their visit to the mall. The model diagram in Fig. 1 shows that persons of different gender and ethnicity and at different stages of their life cycle (e.g., having children) also have different sense of place attitudes. It also shows that these same attitudes are different depending on the day of the week and time of day presumably due to the different roles places play in peoples' schedule of activities in a week and within a day. Similar influence of day of the week and time of day is also found for sequencing of activities (see the dashed line to the left hand side of the figure). Sense of place is a dynamic concept that changes within a day, within a week, and within a life course. This makes it hard to measure and at the same time is a rich informant of behavior. In a subsequent study Chen and Sekar (2018) used the same six *Sense of Place* factors to study the frequency of visiting places and nonmotorized mode choice for

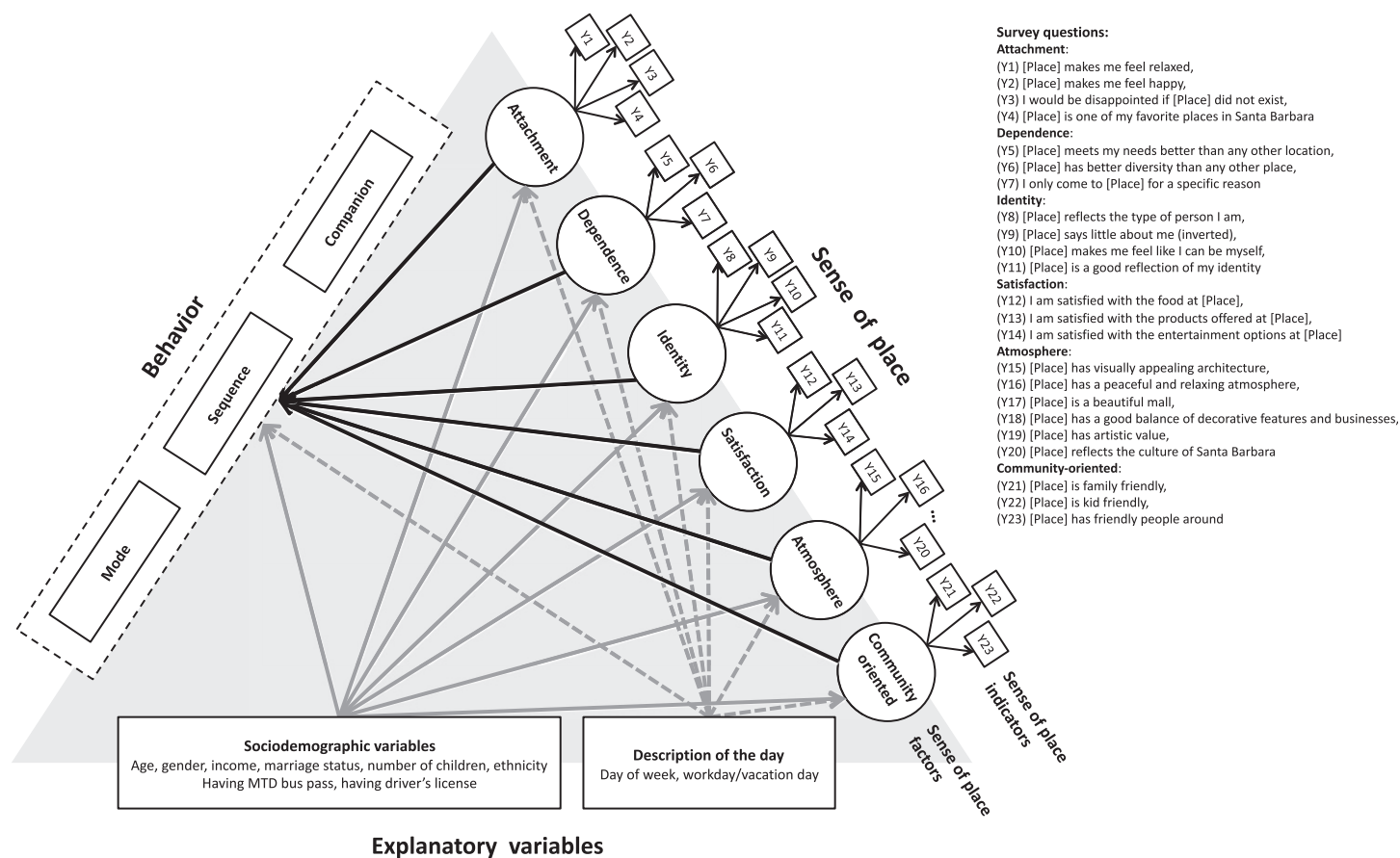


Figure 1 Conceptual and correlational sense of place diagram.

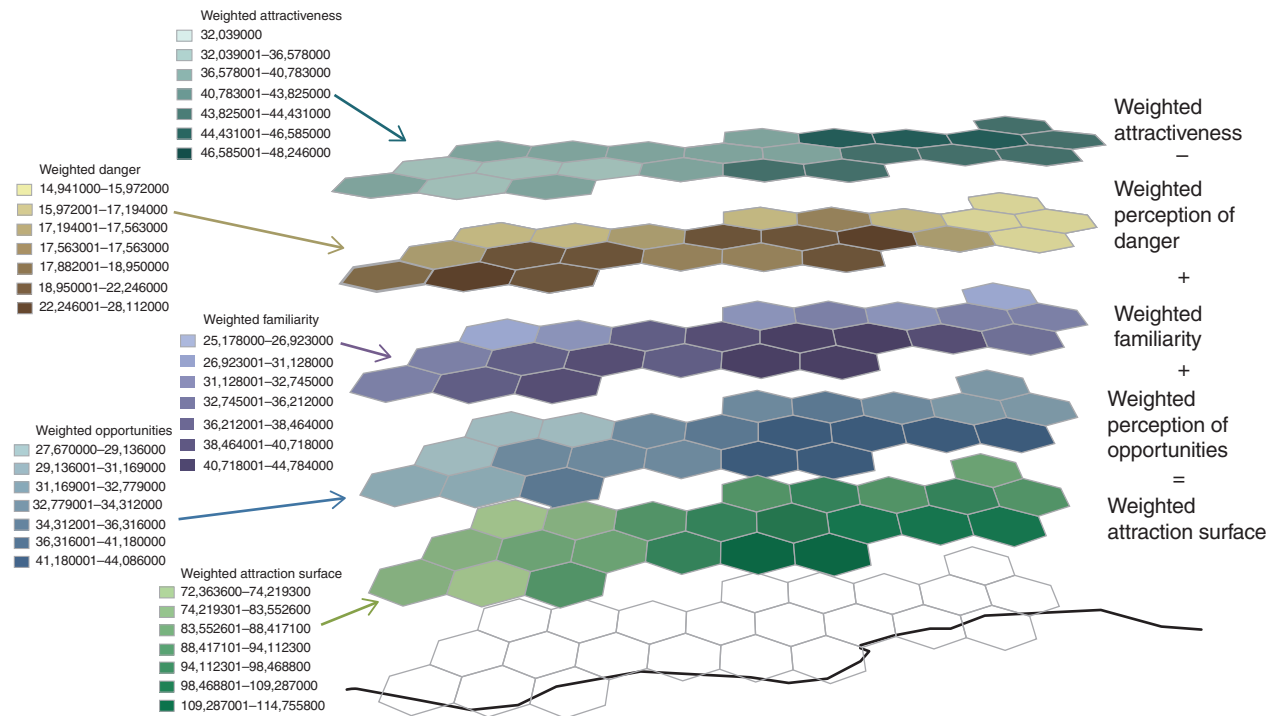


Figure 2 Attractiveness, danger, opportunities, familiarity, and a composite attractiveness index for Santa Barbara hexagons.

three sites in Rochester, NY, proving repeatability of this technique of behavioral measurement. They also studied the use of information and communication technology (ICT) about the three sites. In terms of the relationship between *Sense of Place* and travel behavior they found *Sense of Place* influencing the propensity of people not only to visit places more often, but also to use nonmotorized modes. In addition, getting information about places via ICT changes peoples *Sense of Place*.

Quantifying Citywide Sense of Place

The primary tool to link citywide land use with transportation services and travel behavior is accessibility through a metric representing the ease of reaching opportunities, amenities, and services (Geurs and Van Wee, 2004; Handy and Niemeier, 1997). Distinguishing between experienced accessibility versus objectively measured accessibility is the subject of more recent studies that use travel behavior data to inform place perception (Thériault and Des Rosiers, 2004). Similarly, comparisons between perceived and objectively measured mode-specific attributes are also used to link propensity of using modes and place-specific risks (Manton et al., 2016). These are methods that measure perception of place affordances and represents one dimension of place quality. In contrast, one can quantify *Sense of Place* citywide using attitudinal questions about each location in a city and develop maps of cities that capture not only perceptions of affordances but also feelings, emotions, and place-specific perceptions of quality of life.

In a pilot study in South Santa Barbara County, CA, a direct measurement of place perceptions was implemented as a mapping exercise dividing the study area in 23 hexagons (tessellating the study region) asking respondents to rank their agreement on a Likert-like scale (of -3 being strongly disagree to +3 being strongly agree), with respect to four descriptive questions. The number of hexagons tessellating the study area as well as the number of questions used to capture place perception requires deliberate thought and attention to the balance between level of detail of the data and respondent burden. The four key questions (as well as their keywords in parentheses) for each hexagon are: (1) this is an attractive area of Santa Barbara (attractiveness); (2) this is a dangerous area of Santa Barbara (perception of danger); (3) this area provides me with a lot of opportunities to do things I like to do (opportunities); (4) I am very familiar with this area of Santa Barbara (familiarity). These attributes combine a variety of discussions on topics such as the development of cognitive maps and the influence of the physical landscape on psychological perspectives and human spatial interaction. Attractiveness of the landscape can be seen as an element that contributes to cognitive aspects such as the legibility of places (Lynch, 1960), development of sense of place (Bjørn and Bjerke, 2002), differences that exist between perception and reality of danger (Rengert and Pelfrey, 1997); and the role of exposure to the region as a factor of attaching meaning of places and organizing mental representations of spatial information (Golledge and Spector, 1978; Portugali, 2018).

Following this mapping exercise, respondents were also asked to score the importance of each of these four aspects when choosing a destination for an activity on a 1 through 10 scale (with 1 being not important, and 10 being very important). A full description of the methodology and reasoning behind the use of this hexagonal method can be found in (Deutsch, 2013). Based on

the respondent data an index of subjective attractiveness for each hexagon was created that in essence is the sum of importance by perception of each of the four questions and the weight or importance that individual gives to each aspect of place (Fig. 2). Attractiveness and opportunities were thought to be attractors and therefore were given positive influence in the place equation, danger was seen as a negative, and therefore had a negative influence, and familiarity was determined to magnify these perceptions and therefore was treated as a multiplier of the three aforementioned indicators (other types of indices are documented in Deutsch (2013). Deutsch-Burgner and Goulias (2015)). Our team also used latent class cluster analysis on these data and found noticeable differences among attractiveness of the hexagons. Residents of the Eastern side of town (Montecito and Downtown/midtown region) had higher geographic preference toward the downtown and Montecito areas. Goleta residents (the western portion of the study region) however were less geographically biased and more consistent with indicators of attraction based on density of opportunities (the downtown area). All this supports the use of subjective measures of accessibility in models to improve realism.

Subjective well-being is a more recent interest in travel behavior research (Mokhtarian, 2019) and in the wider regional science research (Ballas and Tranmer, 2012). Deutsch et al. (2013, 2014) and Deutsch and Goulias (2013) examined the attractiveness of the hexagons and their relationship with subjective well-being and happiness jointly with other characteristics. In this portion of the survey, participants rated how different aspects matter to them in the selection of a destination for grocery shopping, other shopping, activities with family, outdoor recreation, social activities, eating out, and entertainment. The questions iterated through each of the seven activity types, and responses were on a seven point Likert scale from strongly disagree to strongly agree. The attributes include: (1) cost of goods or services at the place—"cost"; (2) whether the place is a good reflection of the type of person I am—"identity"; (3) the quality of the products or services offered—"quality"; (4) Whether the place has a positive social atmosphere—"social"; (5) how much time it will take me to travel to the place—"travel time"; (6) how well the place reflects the Santa Barbara lifestyle—"culture"; (7) how close the place is to my home—"distance"; (8) the safety of the surrounding area—"safe"; (9) if there are other places close by where I can do other activities—"proximity"; (10) whether the place meets all my (fill in the activity type) needs—"dependence"; (11) whether the place makes me feel happy—"happy".. Analysis of these aspects shows that individuals have different expectations and criteria to evaluate a destination, and therefore the activities at each place. For instance, happiness matters less for shopping activities than it does for family, social or recreation activities. Additionally, proximity is consistently ranked in the middle of all aspects of place, which challenges current practices of modeling accessibility as one of the most important factors in choosing one destination over another. It also shows that sense of place indicators as stronger than the usual cost and time variables to explain preferences. Moreover, anticipated subjective well-being is a motivator for activity location choice.

To test further these findings, Davis (2015) and Davis et al. (2015) also studied the correlation between attitudes measured with this hexagonal tessellation with objectively measured spatial opportunities (enumeration of business establishments by type) and online reports on Twitter (see the sentiment review by Hollander et al., 2016). They found that the opportunities people believe a place provides are higher when business establishments are more numerous, areas that provide more natural amenities are considered more attractive, and area residents are generally more familiar with areas that are more attractive and opportunity-rich. This portion of comparison with Twitter data shows that online social media provide a different facet of place attitudes with a rich vocabulary of positive and negative aspects of different places that are not readily translatable to travel behavior variables.

Closing Remarks

In summary, place perceptions have a strong influence on decision making and choices in travel behavior such as the choice of a destination. The development of place perceptions, or *Sense of Place*, is a function of individual preferences as well as a variety of elements of the physical, social, and cultural environment of a place. Place perceptions are important to recognize in our attempts to understand, measure, model, and predict travel behavior using quantitative measures of *Sense of Place*. In this way the more traditional attributes of distance, cost or time, and summaries of affordances such as accessibility can be complemented by a variety of subjective measures of place attractiveness and their decision making weights.

The examples presented here are a small portion of the rich body of literature describing and theorizing the interaction of people and place and the affective ties that one develops. The examples here show how measurement can be done at different scales applying *Sense of Place* concepts to the measurement of everyday locations and the application in travel behavior shows modeling is possible and beneficial as a multiscale process. The *Sense of Place* development and analysis described here suggests that *attachment, dependence, identity, satisfaction, atmosphere and community* are the six most important factors for small-scale application. For larger scales perceptions of opportunities and attractiveness combined with perception of danger and familiarity can be used as succinct summaries of place perception. These cannot be indirectly measured by other variables or represented using socio-demographics due to heterogeneity of perception among people.

As we have discussed, place perceptions influence the production and generation of travel and as such they are therefore integral to include in urban design efforts. As we strive to design for all modes of transportation, it is important to include design elements that influence the use of those modes. Lighting, inviting open spaces or storefronts, and vegetation, benches or other street furniture can all contribute to a positive development of atmosphere perceptions and therefore *Sense of Place*. Elements of place perceptions should therefore be included in metrics we use to design complete streets (streets that enable the use of all modes comfortably, equitably, and safely, Hui et al., 2018). Beyond the complete streets idea, Hu and Chen (2018) present a framework for integrating landscape design with place perceptions in order to create urban environments that foster positive

places. This should extend to the prescriptions of human-centered landscape architecture that appeals to the physiological aspects of every-day life (Gehl, 2011, 2013). Combining place perceptions with traditional design standards and metrics can enhance the environment, *Sense of Place*, and therefore quality of life of people who dwell or frequent these urban environments.

References

- Adamowicz, W., Swait, J., Boxall, P., Louviere, J., Williams, M., 1997. Perceptions versus objective measures of environmental quality in combined revealed and stated preference models of environmental valuation. *J. Environ. Econ. Manag.* 32 (1), 65–84.
- Agarwal, P., 2018. Chapter 16. In: Montello, D.R. (Ed.), *Handbook of Behavioral and Cognitive Geography*, Edward Elgar Publishing, Cheltenham, UK, pp. 291–306.
- Ahuja, C., Ayers, C., Hartz, J., Adu-Brimpong, J., Thomas, S., Mitchell, V., Johnson, A., 2018. Examining relationships between perceptions and objective assessments of neighborhood environment and sedentary time: data from the Washington, DC Cardiovascular Health and Needs Assessment. *Prevent. Med. Rep.* 9, 42–48.
- Arentze, T., Timmermans, H., 2008. Social networks, social interactions, and activity-travel behavior: a framework for microsimulation. *Environ. Plann. B: Plann. Design* 35 (6), 1012–1027.
- Auld, J., Mohammadian, A., Doherty, S., 2008. Analysis of activity conflict resolution strategies. *Transport. Res. Re. J. Transport. Res. Board* 2054 (1), 10–19.
- Axhausen, K.W., Gärling, T., 1992. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Trans. Rev.* 12 (4), 323–341.
- Ballas, D., Tranmer, M., 2012. Happy people or happy places? A multilevel modeling approach to the analysis of happiness and well-being. *Inter. Region. Sci. Rev.* 35 (1), 70–102.
- Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D.S., Daly, A., De Palma, A., Gopinath, D., Karlstrom, A., Munizaga, M.A., 2002. Hybrid choice models: progress and challenges. *Market. Lett.* 13 (3), 163–175.
- Bhat, C.R., Koppelman, F.S., 1999. Activity-based modeling of travel demand. In: *Handbook of Transportation Science*, Springer, Boston, MA, pp. 35–61.
- Bjørn, P.K., Bjerke, T., 2002. Associations between landscape preferences and place attachment: a study in Røros, Southern Norway. *Landsc. Res.* 27 (4), 381–396.
- Campelo, A., Aitken, R., Thyne, M., Gnoth, J., 2014. Sense of place: The importance for destination branding. *J. Travel Res.* 53 (2), 154–166.
- Carrasco, J.A., Miller, E.J., 2006. Exploring the propensity to perform social activities: a social network approach. *Transportation* 33 (5), 463–480.
- Chen, R.B., Sekar, A., 2018. Investigating the impact of Sense of Place on site visit frequency with non-motorized travel modes. *J. Trans. Geograph.* 66, 268–282.
- Chen, Y., Ravulaparthi, S., Deutsch, K., Dalal, P., Yoon, S.Y., Lei, T., Hu, H.H., 2011. Development of indicators of opportunity-based accessibility. *Transport. Res. Rec.* 2255 (1), 58–68.
- Chon, K.S., 1990. The role of destination image in tourism: a review and discussion. *Tourist Rev.* 45 (2), 2–9.
- Clifton, K., Perez, P., 2015. Workshop synthesis: Built environment and contextual variables. *Transport. Res. Procedia* 11, 452–459.
- Davis, A.W., 2015. Investigating Place Attitudes in Santa Barbara, CA. MA Thesis. Department of Geography, UCSB, Santa Barbara, CA. Available from: <https://cloudfront.escholarship.org/dist/prd/content/qt8b6922n/qt8b6922n.pdf>.
- Davis, A.W., Lee, J.H., Deutsch-Burgner, K., McBride, E., Goulias, K.G., 2015. Perception and reality: linking metrics of place perception to measurable attributes of place using cross-classified SEM analysis. In: *International Choice Modeling Conference*, Austin, TX, vol. 20, 1. Available from: http://geog.ucsb.edu/geotrans/publications/Davis_Hexagons_GeoTransReport.
- de Abreu, E.S.J., Morency, C., Goulias, K.G., 2012. Using structural equations modeling to unravel the influence of land use patterns on travel behavior of workers in Montreal. *Transport. Res. A Policy Pract.* 46 (8), 1252–1264.
- Deutsch, K., 2013. An Investigation in Decision Making and Destination Choice Incorporating Place Meaning and Social Network Influences. Doctoral Dissertation. University of California, Santa Barbara, Santa Barbara, CA.
- Deutsch-Burgner, K., Goulias, K.G., 2015. Measuring heterogeneity in spatial perception for activity and travel behavior modeling. In: *Transportation Research Board 94th Annual Meeting*. Available from: <http://www.geog.ucsb.edu/geotrans/publications/>.
- Deutsch, K., Goulias, K., 2010. Exploring sense-of-place attitudes as indicators of travel behavior. *Transport. Res. Rec. J. Transport. Res. Board* 2157, 95–102, doi:10.3141/2157-12.
- Deutsch, K., Goulias, K.G., 2013. Decision makers and socializers, social networks and the role of individuals as participants. *Transportation* 40 (4), 755–771, doi:10.1007/s11116-013-9465-6.
- Deutsch, K., Ravulaparthi, S., Goulias, K., 2014. Place Happiness: its constituents and the influence of emotions and subjective importance on activity type and destination choice. *Transportation* 41 (6), 1323–1340.
- Deutsch, K., Yoon, S.Y., Goulias, K., 2013. Modeling travel behavior and sense of place using a structural equation model. *J. Trans. Geograph.* 28, 155–163, doi:10.1016/j.jtrangeo.2012.12.001.
- Elliot, S., Papadopoulos, N., Kim, S.S., 2011. An integrative model of place image: exploring relationships between destination, product, and country images. *J. Travel Res.* 50 (5), 520–534.
- Ettema, D., Timmermans, H.J.P., 1997. Activity-based approaches to travel analysis. Pergamon, Oxford, UK.
- Gehl, J., 2011. *Life Between Buildings: Using Public Space*. Island press, Washington, DC, USA.
- Gehl, J., 2013. *Cities for People*. Island press, Washington, DC, USA.
- Geurs, K.T., Van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. *J. Trans. Geograph.* 12 (2), 127–140.
- Golledge, R., Spector, A., 1978. Comprehending the urban environment: theory and practice. *Geograph. Analysis* 10 (4), 403–426.
- Golledge, R.G., Gärling, T., 2003. Spatial behavior in transportation modeling and planning. In: Goulias, K.G. (Ed.), *Transportation Systems Planning: Methods and Applications*, CRC Press, Boca Raton, FL, USA.
- Goulias, K., Brog, W., Erl, E., 1998. Perceptions in mode choice using situational approach: trip-by-trip multivariate analysis for public transportation. *Transport. Res. Record* 1645 (1), 82–93.
- Goulias, K., Henson, K., 2006. On altruists, egoists in activity participation, travel, who are they, do they live together? *Transportation* 33 (5), 447–462.
- Goulias, K.G., 1996. Activity-based travel forecasting: what are some issues. In: *Activity-Based Travel Forecasting Conference*, vol. 25, p. 3749.
- Handy, S.L., Niemeier, D.A., 1997. Measuring accessibility: an exploration of issues and alternatives. *Environ. Plann. A* 29 (7), 1175–1194.
- Hannes, E., Kusumastuti, D., Espinosa, M.L., Janssens, D., Vanhoof, K., Wets, G., 2012. Mental maps and travel behaviour: meanings and models. *J. Geograph. Syst.* 14 (2), 143–165.
- Hoehner, C.M., Ramirez, L.K.B., Elliott, M.B., Handy, S.L., Brownson, R.C., 2005. Perceived and objective environmental measures and physical activity among urban adults. *Am. J. Prevent. Med.* 28 (2), 105–116.
- Hollander, J.B., Graves, E., Renski, H., Foster-Karim, C., Wiley, A., Das, D., 2016. A (short) history of social media sentiment analysis. In: *Urban Social Listening*, PalgraveMacmillan, London, pp. 15–25.
- Hu, M., Chen, R., 2018. A framework for understanding sense of place in an urban design context. *Urban Sci.* 2 (2), 34.
- Hui, N., Saxe, S., Roorda, M., Hess, P., Miller, E.J., 2018. Measuring the completeness of complete streets. *Trans. Rev.* 38 (1), 73–95.
- Jorgensen, B.S., Stedman, R.C., 2001. Sense of place as an attitude: Lakeshore owners attitudes toward their properties. *J. Environ. Psychol.* 21 (3), 233–248.
- Kianicka, S., Buchecker, M., Hunziker, M., Müller-Böcker, U., 2006. Locals' and tourists' sense of place. *Mount. Res. Dev.* 26 (1), 55–64.
- Kitamura, R., 1988. An evaluation of activity-based travel analysis. *Transportation* 15 (1–2), 9–34.
- Kitamura, R., Mokhtarian, P.L., Laidet, L., 1997. A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. *Transportation* 24 (2), 125–158.
- Lin, T.G., Xia, J.C., Robinson, T.P., Goulias, K.G., Church, R.L., Olaru, D., Han, R., 2014. Spatial analysis of access to and accessibility surrounding train stations: A case study of accessibility for the elderly in Perth, Western Australia. *J. Trans. Geograph.* 39, 111–120.

- Lynch, K., 1960. *The Image of the City*. MIT Press, Cambridge, Mass.
- Manton, R., Rau, H., Fahy, F., Sheahan, J., Clifford, E., 2016. Using mental mapping to unpack perceived cycling risk. *Accid. Analysis Prevent.* 88, 138–149.
- Mokhtarian, P.L., 2019. Subjective well-being and travel: retrospect and prospect. *Transportation* 46 (2), 493–513.
- Mokhtarian, P.L., Salomon, I., Singer, M.E., 2015. What moves us? An interdisciplinary exploration of reasons for traveling. *Trans. Rev.* 35 (3), 250–274.
- Montello, D., 1993. Scale and multiple psychologies of spaces. *Spatial information theory: a theoretical basis for GIS*. In: Frank, A.U., Campari, I. (Eds.), *Lecture Notes in Computer Science*, 716. Springer-Verlag, Berlin, pp. 312–321.
- Pike, S., Ryan, C., 2004. Destination positioning analysis through a comparison of cognitive, affective, and conative perceptions. *J. Travel Res.* 42 (4), 333–342.
- Portugali, J., 2018. History and theoretical perspectives of behavioral and cognitive geography. In: Montello, D.R. (Ed.), *Handbook of Behavioral and Cognitive Geography*, Elgar Publishing, Cheltenham, UK.
- Prevedouros, P.D., 1992. Associations of personality characteristics with transport behavior and residence location decisions. *Transport. Res. A Policy Pract.* 26A, 5.
- Proshansky, H.M., 1978. The city and self-identity. *Environ. Behav.* 10 (2), 147–169.
- Rengert, G.F., Pelfrey, W.Jr, 1997. Cognitive mapping of the city center: comparative perceptions of dangerous places. In: Weisburd, D., McEwen, T. (Eds.), *Crime Mapping and Crime Prevention*, Criminal Justice Press, Monsey, New York, pp. 193–218.
- Ruiz, T., Roorda, M., 2008. Analysis of planning decisions during the activity- scheduling process. *Transport. Res. Rec. J. Transport. Res. Board* 2054, 46–55.
- Stedman, R.C., 2003. Is it really just a social construction? The contribution of the physical environment to sense of place. *Societ. Nat. Resour.* 16 (8), 671–685.
- Steg, L., Vlek, C., Slotegraaf, G., 2001. Instrumental-reasoned and symbolic-affective motives for using a motor car. *Transport. Res. F Traffic Psychol. Behav.* 4 (3), 151–169.
- Stokols, D., Shumaker, A., 1981. People in places: a transactional view of settings. In: Harvey, J. (Ed.), *Cognition, Social Behaviour and the Environment*, Erlbaum, New Jersey.
- Stylidis, D., Sit, J., Biran, A., 2016. An exploratory study of residents' perception of place image: the case of Kavala. *J. Travel Res.* 55 (5), 659–674.
- Thériault, M., Des Rosiers, F., 2004. Modelling perceived accessibility to urban amenities using fuzzy logic, transportation GIS and origin-destination surveys. In: *Proceedings of AGILE 2004 7th Conference on Geographic Information Science*, Heraklion, Greece. Crete University Press, pp. 475–485.
- Timmermans, H., 2003. The saga of integrated land use-transport modeling: How many dreams before we wake up? Paper presented at 10th International Conference on Travel Behavior Research, Lucerne, Switzerland.
- Van Acker, V., Mokhtarian, P.L., Witlox, F., 2014. Car availability explained by the structural relationships between lifestyles, residential location, and underlying residential and travel attitudes. *Trans. Pol.* 35, 88–99.
- Wang, D., 2015. Place, context and activity–travel behavior: to the special section on geographies of activity–travel behavior. *J. Trans. Geograp.* (47), 84–89.
- Wang, F., Wang, D., 2016. Place, geographical context and subjective well-being: state of art and future directions. In: *Mobility, Sociability and Well-Being of Urban Living*, Springer, Berlin, Heidelberg, pp. 189–230.

Further Reading

- Altman, I., Low, S.M., 1992. *Place attachment*. Plenum Press, New York.
- Golledge, R.G., Stimson, R.J., 1997. *Spatial Behavior: A Geographic Perspective*. Guilford Press, New York.
- Mondschein, A. S., 2012. *The Personal City: The Experiential, Cognitive Nature of Travel and Activity and Implications for Accessibility*. Doctoral Dissertation. UCLA.
- Tuan, Y.-F., 1974. *Topophilia: A Study of Environmental Perception, Attitudes, and Values*. Prentice-Hall, Englewood Cliffs, NJ.

Page left intentionally blank

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

Page left intentionally blank

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

EDITOR-IN-CHIEF

Roger Vickerman

*School of Economics, University of Kent, Canterbury, United Kingdom
and*

Transport Strategy Centre, Imperial College, London, United Kingdom

VOLUME 5

Transport Modes

SECTION EDITORS

Edoardo Marcucci

*Department of Political Sciences, University of Roma Tre, Rome, Italy
and*

Department of Logistics, Molde University College, Molde, Norway



Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2021 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-08-102671-7

For information on all publications visit our
website at <http://store.elsevier.com>

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

Publisher: Oliver Walter
Acquisitions Editor: Oliver Walter
Content Project Manager: Natalie Lovell
Associate Content Project Manager: Manisha K and Ramalakshmi Boobalan
Designer: Matthew Limbert

EDITORIAL BOARD

Editor in Chief

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Section Editors

Maria Börjesson

Professor of Economics, VTI Swedish National Road and Transport Research Institute; Affiliated Professor at Linköping University, Sweden

Per Gårder

Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States

Prof. Kevin P.B. Cullinane

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

Prof. Edward C.S. Chung

Department of Electrical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

Chandra R. Bhat

Center for Transportation Research (CTR), The University of Texas at Austin, TX, United States

Edoardo Marcucci

Department of Political Sciences, University of Roma Tre, Rome, Italy; Department of Logistics, Molde University College, Molde, Norway

Prof. Maria Attard

Department of Geography, Faculty of Arts, University of Malta, Msida, Malta

Prof. Carlo G. Prato

School of Civil Engineering, The University of Queensland, Brisbane, Australia

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Page left intentionally blank

INTRODUCTION



Roger Vickerman

In an increasingly globalized world, despite reductions in costs and time, transportation has become even more important as a facilitator of economic and human interaction; this is reflected in technical advances in transportation systems, increasing interest in how transportation interacts with society and the need to provide novel approaches to understand its impacts. This has become particularly acute with the impact that Covid-19 has had on transportation across the world, at local, national, and international levels. This Encyclopedia brings a cross-cutting and integrated approach to all aspects of transportation from work in many disciplinary fields, engineering, operations research, economics, geography, and sociology to understand the changes taking place.

Transportation is both influenced by, and an influencer of, changes in the economy and society. Increasing speeds have reduced journey times and made the world a smaller place as globalization has affected both where people live and work and from where they source their goods and materials. Increasing volumes of traffic, often on old and life-expired infrastructure, lead to congestion and delays. Constraints on public budgets have led to increasing pressure on the private sector to fund improvements requiring new and innovative financial solutions. While there are clear differences in the nature of the pressures felt in the developed and less developed economies, there is an increasing recognition that in all societies, there is an accessibility problem such that certain groups become disadvantaged by the lack of access to reliable and cost-effective transport. While the problems are clearly multidimensional, research on transportation is often constrained by single disciplinary approaches and this carries over into the practice of transport planning and policy. The Encyclopedia cuts across these artificial boundaries by taking an approach that emphasizes the interaction between the different aspects of research and aims to offer new solutions to understand these problems. Each of the nine sections is based around a familiar dimension of work on transportation, but brings together the views of experts from different disciplinary perspectives. Each section is edited by an expert in the relevant field who has sought chapters from a range of authors representing different disciplines, different parts of the world, and different social perspectives. In this way, the work is not just a reflection of the state of the art that serves as a starting point for researchers and practitioners, but also a pointer toward new approaches, new ways of thinking, and novel solutions to problems.

The nine sections are structured around the following themes: Transport Modes; Freight Transport and Logistics; Transport Safety and Security; Transport Economics; Traffic Management; Transport Modeling and Data Management; Transport Policy and Planning; Transport Psychology; Sustainability and Health Issues in Transportation. Some of the chapters provide a technical introduction to a topic while others provide a bridge between topics or a more future-oriented view of new research areas or challenges. While there is guidance to cross-referencing between chapters, readers are encouraged to explore the tables of contents of all the sections to get a full understanding of the issues. Much of the Encyclopedia was completed before the Covid-19 pandemic and clearly this will have changed the situation in many areas covered by this work; the advantage of this type of reference work is that relevant updates will be possible in future editions.

The Encyclopedia has only been possible because of the cooperation of a large number of people. Robert Noland, Georgina Santos, Xiaowen Fu, and Dick Ettema served as an Editorial Advisory Board identifying possible editors of sections and advising on the overall structure. The Section Editors, Edoardo Marcucci, Kevin Cullinane, Per Garder, Maria Börjesson, Edward Chung, Chandra Bhat, Maria Attard, and Carlo Prato,

carried out the work of identifying potential authors of individual chapters, commissioning these, encouraging authors and reviewing drafts. More than 600 authors and co-authors of chapters are, however, ultimately responsible for the success of this venture. Thanks are also due to the key people at Elsevier, particularly the Publishers, Andre Wolff and Oliver Walter, and the Project Managers, Sophie Harrison and Natalie Bentahar; they have shown exemplary patience over more than 3 years in bringing this work to fruition.

LIST OF CONTRIBUTORS TO VOLUME 5

Edoardo Marcucci

Department of Political Sciences, University of Roma Tre, Rome, Italy; Department of Logistics, Molde University College, Molde, Norway

Xavier Fageda

Department of Applied Economics and GIM-IREA, Universitat de Barcelona, Barcelona, Spain

Cecilia Olivieri

Department of Economics, Faculty of Social Sciences, Universidad de la República, Montevideo, Uruguay

Charles B.A. Musselwhite

Centre for Innovative Ageing, Swansea University, Swansea, United Kingdom

Theresa Scott

University of Queensland, St. Lucia, QLD, Australia

Erling Holden

Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Aas, Norway

Geoffrey Gilpin

Department of Environmental Sciences, Western Norway University of Applied Sciences, Sogndal, Norway

Michele D. Simoni

Department of Civil, Architectural, and Environmental Engineering, The University of Texas at Austin, Austin, TX, United States

Kara M. Kockelman

Center for Transportation and Logistics, Massachusetts Institute of Technology, Cambridge, MA, United States

Ila Maltese

Dipartimento di Scienze Politiche, Università degli Studi Roma3, Roma, Italy

Luca Zamparini

Dipartimento di Scienze Giuridiche, University del Salento, Lecce, Italy

Bert van Wee

Delft University of Technology, Delft, The Netherlands

Jean-Paul Rodrigue

Department of Global Studies and Geography, Hofstra University, Hempstead, NY, United States

Erick Guerra

University of Pennsylvania, Stuart Weitzman School of Design, Department of City and Regional Planning, Philadelphia, PA United States

Gilles Duranton

University of Pennsylvania, The Wharton School, Real Estate Department

Lóránt Tavasszya

Delft University of Technology, Delft, The Netherlands

Gerard de Jongb

Institute for Transport Studies, University of Leeds, United Kingdom and Significance, The Hague, The Netherlands

Sebastian Bamberg

Department of Social Sciences, University of Applied Sciences, Bielefeld, North Rhine-Westphalia, Germany

Icek Ajzen

Department of Psychological and Brain Sciences, University of Massachusetts Amherst, Amherst, MA, United States

Peter Schmidt

Department of Political Science and Centre for Environment and Development (ZEU), University of Giessen, Ludwigstrasse, Germany

Zissis Samaras

Department of Mechanical Engineering, Aristotle University of Thessaloniki, Greece

Ilias Vouitsis

Department of Mechanical Engineering, Aristotle University of Thessaloniki, Greece

Roger L Mackett

Centre for Transport Studies, University College London, London, Great Britain

Colin G. Pooley
Lancaster Environment Centre, Lancaster University,
Lancaster, United Kingdom

Antonio Comi
Antonio Comi, Department of Enterprise Engineering,
University of Rome Tor Vergata, via del Politecnico 1,
Rome, Italy

Jennifer S. Mindell
Health and Social Surveys Research Group, Research
Department of Epidemiology & Public Health, UCL,
London, England, United Kingdom

Sandra Mandic
Active Living Laboratory, School of Physical Education,
Sports, and Exercise Sciences, University of Otago,
Dunedin, Otago, New Zealand

Sören Groth
ILS-Research Institute for Regional and Urban
Development, Dortmund, Germany

Tobias Kuhnimhof
RWTH Aachen University, Institute for Urban and
Transport Planning, Aachen, Germany

Brian Wolshon
Department of Civil and Environmental, Louisiana State
University, Baton Rouge, LA, United States

Jan-Dirk Schmöcker
Department of Urban Management, Kyoto University, Japan

Regine Gerike
Technische Universität Dresden, Faculty of Transport and
Traffic Sciences "Friedrich List", Chair of Integrated
Transport Planning and Traffic Engineering, Dresden,
Germany

Stefan Hubrich
Technische Universität Dresden, Faculty of Transport and
Traffic Sciences "Friedrich List", Chair of Integrated
Transport Planning and Traffic Engineering, Dresden,
Germany

Caroline Koszowski
Technische Universität Dresden, Faculty of Transport and
Traffic Sciences "Friedrich List", Chair of Integrated
Transport Planning and Traffic Engineering, Dresden,
Germany

Bettina Schröter
Technische Universität Dresden, Faculty of Transport and
Traffic Sciences "Friedrich List", Chair of Integrated
Transport Planning and Traffic Engineering, Dresden,
Germany

Rico Wittwer
Technische Universität Dresden, Faculty of Transport and
Traffic Sciences "Friedrich List", Chair of Integrated

Transport Planning and Traffic Engineering, Dresden,
Germany

Christine Eisenmann
German Aerospace Center, Institute of Transport
Research, Berlin, Germany

Daniel Görges
University of Kaiserslautern, Electromobility Research
Group, Kaiserslautern, Germany

Thomas Franke
Universität zu Lübeck, Institute for Multimedia and
Interactive System, Engineering Psychology and Cognitive
Ergonomics, Lübeck, Germany

Susan Shaheen
McLaughlin Hall, University of California, Berkeley, CA,
United States

Sigal Kaplan
Department of Geography, Hebrew University of
Jerusalem, Mount Scopus, Jerusalem, Israel

Galit Cohen-Blankshtain
School of Public Policy, Department of Geography, The
Hebrew University, Jerusalem, Israel

Joachim Scheiner
Technische Universität Dortmund, Faculty of Spatial
Planning, Department of Transport Planning, Dortmund,
North Rhine-Westphalia, Germany

Milan Janić
Department of Transport & Planning, Faculty of Civil
Engineering and Geosciences, Delft University of
Technology, Delft, The Netherlands

Richard P. Hallion
Independent Scholar, Shalimar, FL,
United States

Lucy Budd
Leicester Castle Business School, De Montfort University,
Leicester, United Kingdom

Stephen Ison
Leicester Castle Business School, De Montfort University,
Leicester, United Kingdom

Rico Merkert
Institute of Transport and Logistics Studies,
The University of Sydney Business School, Sydney,
NSW, Australia

James Bushell
Institute of Transport and Logistics Studies,
The University of Sydney Business School, Sydney,
NSW, Australia

Lanfranco Senn
Bocconi University, Milano, Italy

Scott Le Vine
Department of Geography, SUNY New Paltz, New Paltz, NY, United States

Francesca Pagliara
Department of Civil, Architectural and Environmental Engineering, University of Naples Federico II, Naples, Italy

Juan Carlos Martín
Faculty of Economics, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Concepción Román
Faculty of Economics, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Peter Forsyth
Department of Economics, Monash University, Clayton, Australia

Hans-Martin Niemeier
School of Business and Tourism, Southern Cross University, Australia; University of Applied Sciences, Bremen, Germany

Achim I. Czerny
Department of Logistics and Maritime Studies, Hong Kong Polytechnic University, Kowloon, Hong Kong

Renan Peres de Oliveira
Latin American Center for Transportation Economics, Aeronautics Institute of Technology, São José dos Campos, SP, Brazil

Gui Lohmann
School of Engineering and Built Environment, Griffith University, Nathan, QLD, Australia

Sveinn Vidar Gudmundsson
2 Rue des Pyrenees, Escalquens, France

Volodymyr Bilotkach
Singapore Institute of Technology, Singapore, Singapore

Andrew R. Goetz
Department of Geography & the Environment, University of Denver, Denver, CO, United States

Vincenzo Torre
Center for Near Space, Italian Institute for the Future, Naples, Italy

Antonio Sollo
Piaggio Aerospace, Villanova d'Albenga (SV), Italy

Cristian Domarchi
Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, Santiago, Chile

Juan de Dios Ortúzar
Department of Transport Engineering and Logistics, Instituto Sistemas Complejos de Ingeniería (ISCI), BRT+ Centre of Excellence, Pontificia Universidad Católica de Chile, Santiago, Chile

Preston L. Schiller
Department of Civil and Environmental Engineering, University of Washington (Seattle), Bellingham, WA, United States.

Ana M. Condeço-Melhorado
Complutense University of Madrid, Geography Department, Madrid, Spain

Kevin Manaugh
Department of Geography & McGill School of Environment McGill University, Montreal, QC, Canada

Ehab Diab
University of Saskatchewan, Department of Geography and Planning, Saskatoon, Saskatchewan, Canada

Ahmed El-Geneidy
McGill University, School of Urban Planning, Montréal, QC, Canada

Bruce Appleyard bappleyard@sdsu.edu
San Diego State University, San Diego, CA, United States

Zakhary Mallett
University of Southern California, Los Angeles, CA, United States

Marlon G Boarnet
University of Southern California, Los Angeles, CA, United States

Jos Arts
Department of Planning, Faculty of Spatial Sciences, University of Groningen, The Netherlands

Wim Leendertse
Department of Economic Geography, Faculty of Spatial Sciences University of Groningen, The Netherlands

Taede Tillema
Department of Economic Geography, Faculty of Spatial Sciences University of Groningen, The Netherlands

David A. King
Arizona State University, Tempe, AZ, United States

Kevin J. Krizek
University of Colorado Boulder Boulder, CO, United States

Marco Percoco
Department of Social and Political Sciences and GREEN Università Bocconi Milano, Italy

J.L. Veldstra

Department of Psychology, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, Netherlands

A.B. Æœnal

Campus Fryslan, University of Groningen, Leeuwarden, The Netherlands

E.M. Steg

Department of Psychology, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, Netherlands

Gysele Lima Ricci

Department of Civil, GEO and Environmental Engineering at Technical University of Munich, Munich, Germany

Klaus Bogenberger

Department of Civil, GEO and Environmental Engineering at Technical University of Munich, Munich, Germany

Ingo Arne Hansen

Delft University of Technology, Delft, The Netherlands

Chris Nash

Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

Tony Fowkes

Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

Carlo Ciccarelli

Department of Economics and Finance, University of Rome Tor Vergata, Rome, Italy

Andrea Giuntini

Department of Economics, University of Modena and Reggio Emilia, Modena, Italy

Peter Groote

Department of Cultural Geography, University of Groningen, Groningen, The Netherlands

Frédéric Dobruszkes

Université libre de Bruxelles (ULB), Brussels, Belgium

Amparo Moyano

Universidad de Castilla La Mancha (UCLM), Ciudad Real, Spain

Prof. Andrés Monzón

Transport Research Centre, TRANSyT-Universidad Politécnica de Madrid, Spain

Dr. Elena López

Transport Research Centre, TRANSyT-Universidad Politécnica de Madrid, Spain

Vassilios A. Profillidis

Democritus University of Thrace, School of Civil Engineering, Section of Transportation, Xanthi, Greece

Nacima Baron

Université Paris Est, Laboratoire Ville Mobilité Transport - ENPC, France

Teodor Gabriel Crainic

Analytics, Operations and Information Technology Department, School of Management, Université du Québec À Montréal, Montréal, QC, Canada

Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Montréal, QC, Canada

Guillaume Monchambert

Univ Lyon, Université Lyon 2, LAET, Lyon, France

Daniel Hörcher

Transport Strategy Centre, Imperial College London, London United Kingdom

Alejandro Tirachini

Civil Engineering Department, Universidad de Chile and Instituto Sistemas Complejos de Ingeniería (Complex Engineering Systems Institute), Chile

Nicolas Coulombel

LVMT, Ecole des Ponts, Université Gustave Eiffel, Champs-sur-Marne, France

Vilius Nikitinas

Law firm NIKITINAS LEGAL, Vilnius, Lithuania

Skaistė Miliauskaitė

Law firm NIKITINAS LEGAL, Vilnius, Lithuania

Javier Campos

University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Christos Pyrgidis

Aristotle University of Thessaloniki, Thessaloniki, Greece

Alexandros Dolianitis

Aristotle University of Thessaloniki, Thessaloniki, Greece

Christa Sys

University of Antwerpen, Antwerpen, Belgium

Thierry Vanelander

University of Antwerpen, Antwerpen, Belgium

César Ducruet

Centre National de la Recherche Scientifique (CNRS), Paris, France

Justin Berli
Centre National de la Recherche Scientifique (CNRS), Paris, France

Harilaos N. Psaraftis
Department of Technology, Management and Economics, Technical University of Denmark Lyngby, Denmark

Brian Slack
Department of Geography, Planning and Environment, Concordia University, Montreal, QC, Canada

Hercules Haralambides
Dalian Maritime University, Dalian, China; Texas A&M University, College Station, TX, United States; Erasmus University Rotterdam, Rotterdam, The Netherlands

Lourdes Trujillo
Department of Applied Economics, University of Las Palmas de Gran Canaria (Campus Tafira), Faculty of Economics, Las Palmas de Gran Canaria, Spain

Daniel Castillo Hidalgo
Department of History, University of Las Palmas de Gran Canaria (Campus Tafira), Faculty of Economics, Las Palmas de Gran Canaria, Spain

Manuel Herrera
Louvain School of Management (LSM), Université Catholique de Louvain, Louvain-la-Neuve, Belgium

Beatriz Tovar
Tourism and Transport Research Unit, Institute of Tourism and Sustainable Economic Development, University Las Palmas, Las Palmas University, Las Palmas de Gran Canaria, Spain

Alan Wall
Oviedo Efficiency Group, Department of Economics, University of Oviedo, Oviedo, Spain

Johan Woxenius
Department of Business Administration, University of Gothenburg, Gothenburg, Sweden

Behzad Behdani
Operations Research and Logistics Group, Wageningen University, Wageningen, The Netherlands

Bart Wiegmans
Department of Transport and Planning, Delft University of Technology, Delft, The Netherlands

Yun Fan
Operations Research and Logistics Group, Wageningen University, Wageningen, The Netherlands

Claudio Ferrari
Department of Economics and Business, University of Genoa, Genoa, Liguria, Italy

Alessio Tei
School of Engineering, Newcastle University, Newcastle upon Tyne, Tyne and Wear, United Kingdom

Athanasios A. Pallis
Department of Shipping, Trade and Transport, University of the Aegean, Chios, Greece

Aimilia A. Papachristou
Department of Shipping, Trade and Transport, University of the Aegean, Chios, Greece

Laurent Fedi
KEDGE BS, CESIT, Maritime Governance, Trade and Logistics Lab, Marseille, France

Gareth C. Burton
American Bureau of Shipping, Spring, TX, United States

Mimosa T. Miller
American Bureau of Shipping, Spring, TX, USA

Paul William Stott
Newcastle University, Newcastle, United Kingdom

Zain Ul Abedin
TUMCREATE Ltd. 1 CREATE Way, Singapore

Avishai (Avi) Ceder
Transportation Research Institute, Technion-Israel Institute of Technology, Haifa, Israel

Mark Hempsell
Hempsell Astronautics Limited, Herts, United Kingdom

Franco Cotana
Department of Engineering, University of Perugia, Via Goffredo Duranti, Perugia, Italy

Mattia Manni
Interuniversity Research Center on Pollution and Environment "Mauro Felli", Via Goffredo Duranti, Perugia, Italy

Priya Uteng
Department of Sustainable Urban Development and Mobility, Institute of Transport Economics (TOI) Gaustadalléen Oslo, Norway; Department of Landscape, Spatial and Infrastructure Sciences, Institute for Transport Studies (IVe), University of Natural Resources and Life Sciences, Gregor-Mendel-Straße Vienna, Austria

Yusak Susilo
Department of Landscape, Spatial and Infrastructure Sciences, Institute for Transport Studies (IVe), University of Natural Resources and Life Sciences, Gregor-Mendel-Straße Vienna, Austria

Hannah D Budnitz
Transport Studies Unit, School of Geography and the Environment, University of Oxford, Oxford, Oxfordshire, United Kingdom

Emmanouil Tranos

*School of Geographical Sciences, University of Bristol,
Bristol, United Kingdom*

Lee Chapman

*School of Geography, Earth & Environmental Sciences,
University of Birmingham, Birmingham,
United Kingdom*

Caroline Mullen

*Institute for Transport Studies, University of Leeds, Leeds,
United Kingdom*

Francesco Corman

Stefano Frascini Platz, Zürich, Switzerland

Lutfi Al-Sharif

*Mechatronics Engineering Department, The University of
Jordan, Amman, Jordan; Al-Hussein Technical University,
Amman, Jordan; Peters Research Ltd., Great Missenden,
United Kingdom*

Raktim Mitra

*School of Urban and Regional planning, Ryerson
University, Canada*

Paul M. Hess

*Department of Geography and Planning, University of
Toronto, Canada*

Adam Cohen

University of California, Berkeley, CA, United States

Long Cheng

Department of Geography, Ghent University, Ghent, Belgium

Jonas De Vos

*Bartlett School of Planning, University College London,
London, United Kingdom*

Frank Witlox

*Department of Geography, Ghent University, Ghent,
Belgium*

CONTENTS OF ALL VOLUMES

<i>Editorial Board</i>	<i>v</i>
<i>Introduction</i>	<i>vii</i>
<i>Contributors to Volume 5</i>	<i>ix</i>

VOLUME 1

Introduction to Transportation Economics	1
Market Failures and Public Decision Making in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	2
Demand for Freight Transport <i>José Holguín-Veras and Diana G. Ramírez-Ríos</i>	7
Cost Functions for Road Transport <i>José Manuel Vassallo</i>	13
Future of Urban Freight <i>Russell G. Thompson</i>	20
Operation Costs for Public Transport <i>Marco Batarce</i>	26
Natural Monopoly in Transport <i>André de Palma and Julien Monardo</i>	30
Freight Costs: Air and Sea <i>Yulai Wan and Dong Yang</i>	36
Transport Production and Cost Structure <i>Ricardo Giesen and Darío Farren</i>	46
The Concept of External Cost: Marginal versus Total Cost and Internalization <i>Sofia F. Franco</i>	56
Value of Time <i>John J. Bates</i>	67
Valuation of Carbon Emissions <i>Svante Mandell</i>	72
Valuation of Travel Time Variability Using Scheduling Models <i>Katrine Hjorth</i>	76

Value of Crowding <i>Daniel Hörcher</i>	84
What Drives Transport and Mobility Trends? The Chicken-and-Egg Problem <i>Nathalie Picard</i>	89
Pricing Principles in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	95
Long-Run Versus Short-Run Valuations <i>Stefanie Peer</i>	102
Value of Noise <i>Jon P. Nelson</i>	106
The Value of Life and Health <i>Henrik Andersson</i>	114
The Value of Security, Access Time, Waiting Time, and Transfers in Public Transport <i>Raquel Espino, Juan de Dios Ortúzar, and Luis I. Rizzi</i>	122
Demand for Passenger Transportation <i>Kenneth A Small and Robin Lindsey</i>	127
Real-World Experiences of Congestion Pricing <i>Charles Raux</i>	134
Distributional Effects of Congestion Charges and Fuel Taxes <i>Jonas Eliasson</i>	139
The Bottleneck Model <i>Dereje Abegaz and Yili Tang</i>	146
Dynamic Congestion Pricing and User Heterogeneity <i>Kathrin Goldmann and Gernot Sieg</i>	150
Economics of Parking <i>Daniel Albalade and Albert Gragera</i>	159
Loss Aversion and Size and Sign Effects in Value of Time Studies <i>Andrew Daly</i>	165
Intertemporal Variation of Valuations <i>James Fox</i>	170
The Rebound Effect for Car Transport <i>Bruno De Borger, Ismir Mulalic and Jan Rouwendal</i>	174
Elasticities for Travel Demand: Recent Evidence <i>Fay Dunkerley, Charlene Rohr and Mark Wardman</i>	179
Parking Price Elasticities <i>Stephan Lehner</i>	185
The Pareto Criterion and the Kaldor Hicks Criterion <i>Harald Minken</i>	190
Social Discount Rates <i>Lars Hultkrantz</i>	195
Cross-Elasticities between Modes <i>Mark Wardman and Jeremy Toner</i>	201
First-Best Congestion Pricing <i>Moez Kilani</i>	209

Ethical Aspects-Can We Value Life, Health, and Environment in Money Terms? <i>Jose M. Grisolia and Ken Willis</i>	216
Car tolls, Transit Subsidies for Commuting, and Distortions on the Labor Market <i>Ioannis Tikoudis¹ and Kurt Van Dender¹</i>	221
Demand Management and Capacity Planning of Airports <i>Stephen Ison and Lucy Budd</i>	227
Dealing With Negative Externalities: Low Emission Zones Versus Congestion Tolls <i>Valeria Bernardo, Xavier Fageda, and Ricardo Flores-Fillol</i>	231
The Rule-of-a-Half and Interpreting the Consumer Surplus as Accessibility <i>Mogens Fosgerau and Ninette Pilegaard</i>	237
Producer Surplus <i>Chau Man Fung</i>	242
The Robustness of Cost-Benefit Analyses <i>Morten Welde and James Odeck</i>	249
The GDP Effects of Transport Investments: The Macroeconomic Approach <i>James Laird and Daniel Johnson</i>	256
The Mohring Effect <i>Hugo E. Silva</i>	263
Public Transport Fare and Subsidy Optimization <i>Qianwen Guo and Zhongfei Li</i>	267
Transportation Equity <i>Rafael H. M. Pereira, and Alex Karner</i>	271
Impact of Transport Cost-Benefit Analysis on Public Decision-Making <i>Niek Mouter</i>	278
Causal Inference for Ex Post Evaluation of Transport Interventions <i>Daniel J. Graham</i>	283
Transport Cost and Location of Firms <i>Adelheid Holl</i>	291
Commuting, the Labor Market, and Wages <i>Jan Rouwendal, and Ismir Mulalic</i>	297
How to Buy Transport Infrastructure <i>Johan Nyström</i>	302
Procurement of Public Transport: Contractual Regimes <i>Andrew Smith, and Chris Nash</i>	308
The Mono-Centric City Model and Commuting Cost <i>Zhi-Chun Li, and Ya-Juan Chen</i>	315
Value of Time in Freight Transport <i>Gerard de Jong</i>	321
The Economics and Planning of Urban Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	326
Incentives in Public Transport Contracts <i>Andreas Vigren</i>	332
Transportation Improvements and Property Prices <i>Alex Anas</i>	337

Transport Infrastructure Effects on Economic Output: The Microeconomic Approach <i>Patricia C. Melo</i>	347
Wider Economic Impacts of Transport Investments <i>Anthony J. Venables</i>	355
Employment Effects of Transport Infrastructure <i>Anna Matas, and Javier Asensio</i>	360
Regulatory Reforms and Competition in Public Transport <i>Didier van de Velde</i>	365
How to Finance Transport Infrastructure? <i>Tiziana D'Alfonso, and Giuseppe Catalano</i>	371
Congestion, Allocation and Competition on the Railway Tracks <i>John Armstrong, and John Preston</i>	378
Public Private Partnership <i>David Meunier, and Emile Quinet</i>	385
Airline Economics <i>Anming Zhang, Yahua Zhang, and Zhibin Huang</i>	392
Privatization and Deregulation of the Airline Industry <i>Achim I. Czerny, and Hao Lang</i>	397
Price Discrimination and Yield Management in the Airline Industry <i>Guoquan Zhang, Colin C.H. Law, Yahua Zhang, and Hangjun Yang</i>	404
Regulation and Competition in Railways <i>Chris Nash, and Andrew Smith</i>	409
Cycling Economics <i>Bert van Wee</i>	414
The Economic Rationale for High-Speed Rail <i>Ginés de Rus</i>	419
Rail Cost Functions <i>Kristofer Odolinski, and Phill Wheat</i>	425
Transport and International Trade <i>Siri Pettersen Strandenes</i>	431
Maritime Economics: Organizational Structures <i>Kevin P.B. Cullinane</i>	436
Port Planning and Investment <i>Jasmine Siu Lee Lam</i>	443
Estimating the Capital Stock of Transport Infrastructure <i>Heike Link</i>	449
The Economics of Reducing Carbon Emissions From Air and Road Transport <i>Olga Ivanova</i>	457
Regulation and Financing of Toll Roads <i>Marco Ponti</i>	464
Are Megaprojects too Transformational for Cost-Benefit Analysis? <i>Tom Worsley</i>	470
Economics of Transportation Safety <i>Ian Savage</i>	476

Cost Overruns of Transportation Infrastructure Projects <i>James Odeck, and Morten Welde</i>	483
The Downs-Thomson Paradox <i>Joel P. Franklin</i>	490
Policy Instruments for Plug-In Electric Vehicles: An Overview and Discussion <i>Jake Whitehead, Patrick Plötz, Patrick Jochem, Frances Sprei, and Elisabeth Dütschke</i>	496
Vertical and Horizontal Separation in the European Railway Sector and Its Effects on Productivity <i>Pedro Cantos-Sánchez</i>	503
How will Autonomous Vehicles Impact Car Ownership and Travel Behavior <i>Patrick M. Bösch, Felix Becker, Henrik Becker, and Kay W. Axhausen</i>	508
Policy Instruments to Reduce Carbon Emissions from Road Transport Computable General Equilibrium Analysis in Transportation Economics <i>Johannes Bröcker</i>	520
Estimation of Value of Time <i>Stefan Flügel, and Askill H. Halse</i>	527
The Taxation of Car Use in the Future <i>Griet De Ceuster, and Inge Mayeres</i>	534
Cost-Benefit Analysis and Other Assessment Techniques: Contrasts and Synergies <i>Paolo Beria</i>	540
Demand for Air Travel and Income Elasticity <i>Jing Lu, Yucan Meng, Changmin Jiang, and Cheng Lv</i>	547
Generalized Cost for Transport <i>Jepppe Rich</i>	555
The Impact of Electric Vehicles on Energy Systems <i>Patrick E.P. Jochem, Jake Whitehead, and Elisabeth Dütschke</i>	560
Uber versus Taxis <i>Georgina Santos</i>	566
Contract Efficiency in Public Transport Services <i>Philippe Gagnepain, and Marc Ivaldi</i>	572
Company Cars <i>Stefan Gössling</i>	580
Transportation Network Companies (TNCs) and the Future of Public Transportation <i>Susan Shaheen, and Adam Cohen</i>	584
Public Transport in Low Density Areas <i>Jani-Pekka Jokinen, Leif Sörensen, and Jan Schlüter</i>	589
Market Failures in Transport: Direct and Indirect Public Intervention <i>Federico Boffa, and Alberto Iozzi</i>	596
The Braess Paradox <i>Anna Nagurney, and Ladimer S. Nagurney</i>	601
VOLUME 2	
Introduction to Transportation Safety and Security <i>Per Gårder</i>	1

The Concept of “Acceptable Risk” Applied to Road Safety Risk Level <i>Claes Tingvall</i>	2
Crash Not Accident <i>Robert A. Scopatz</i>	6
Age and Gender as Factors in Road Safety <i>Marion Sinclair</i>	11
Aggressive Driving and Road Rage <i>James E.W. Roseborough, Christine M. Wickens, and David L. Wiesenthal</i>	17
Aircraft Maintenance and Inspection <i>Alan Hobbs</i>	25
Airport Security <i>Richard W. Bloom</i>	34
Transport Safety and Security: Alcohol <i>James C. Fell</i>	40
Animal Crashes <i>Michal Bil</i>	53
Attenuators <i>Simonetta Boria</i>	63
ATV, Snowmobile, and Terrain Vehicle Safety <i>David P. Gilkey, and William Brazile</i>	77
Automobile Safety Inspection <i>Subasish Das</i>	85
Aviation Safety: Commercial Airlines <i>Clarence C. Rodrigues</i>	90
Aviation Safety, Freight, and Dangerous Goods Transport by Air <i>Glenn S. Baxter and Graham Wild</i>	98
Bicycle Collision Avoidance Systems: Can Cyclist Safety be Improved with Intelligent Transport Systems? <i>Lars Leden</i>	108
Bicycle Infrastructure <i>Rock E. Miller</i>	115
Bicycles, E-Bikes and Micromobility, A Traffic Safety Overview <i>Höskuldur Kröyer</i>	125
Bicycles: The Safety of Shared Systems Versus Traditional Ownership <i>Mercedes Castro-Nuño and José I. Castillo-Manzano</i>	139
Bridge Safety <i>Per Erik Garder</i>	144
Carsharing Safety and Insurance <i>Elliot Martin and Susan Shaheen</i>	150
Carjacking <i>Terance D. Mieth, Christopher Forepaugh, Tanya Dudinskaya</i>	157
Collision Avoidance Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	164
Collision Avoidance Systems, Automobiles <i>Erick J. Rodríguez-Seda</i>	173

Connected Automated Vehicles: Technologies, Developments, and Trends <i>Azra Habibovic and Lei Chen</i>	180
Construction Zones <i>Jalil Kianfar</i>	189
Costs of Accidents <i>Ulf Persson</i>	196
Critical Issues for Large Truck Safety <i>Matthew C. Camden, Jeffrey S. Hickman, Richard J. Hanowski, and Martin Walker</i>	200
Demerit Points and Similar Sanction Programs <i>Matúš ťucha and Kristýna Josrová</i>	210
Driver State and Mental Workload <i>Dick de Waard and Nicole van Nes</i>	216
Drugs, Illicit, and Prescription <i>Rune Elvik</i>	221
Education, Training, and Licensing <i>Matúš ťucha and Kristýna Josrová</i>	228
Elderly Driver Safety Issues <i>Mark J King</i>	233
Emergency Response Systems <i>Frances L. Edwards</i>	240
Emergency Vehicles and Traffic Safety <i>Shamsunnahar Yasmin, Sabreena Anowar, and Richard Tay</i>	247
Encouragement: Awards and Incentives <i>Fred Wegman</i>	255
Enforcement and Fines <i>Matúš ťucha and Ralf Risser</i>	263
Epidemiology of Road Traffic Crashes <i>Sherrie-Anne Kaye, Judy Fleiter, and Md Mazharul Haque</i>	269
Evacuation Planning and Transportation Resilience <i>Karl Kim</i>	276
Exposure: A Critical Factor in Risk Analysis <i>Frank Gross, PhD, PE</i>	282
Passenger Ferry Vessels and Cruise Ships: Safety and Security <i>Wayne K. Talley</i>	290
Fuel Economy Standards: Impacts on Safety <i>Kenneth T. Gillingham and Stephanie M. Weber</i>	296
Hazardous Materials Transport <i>Dr. Arjan Vincent van der Vlies</i>	304
Head-on Crashes <i>John N. Ivan</i>	311
Helicopters in Emergency Medical Response <i>Stephen J.M. Sollid, M.D., PhD</i>	316
Horizontal and Vertical Geometry <i>Victoria Gitelman</i>	322

Human Factors in Transportation <i>Alison Smiley, Christina (Missy) Rudin-Brown</i>	331
In-Depth Crash Analysis and Accident Investigation <i>Yong Peng, Helai Huang, and Xinghua Wang</i>	346
Incident Detection Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	351
Inequality and Traffic Safety <i>Miles Tight</i>	358
Lighting <i>John D. Bullough</i>	361
Macroscopic Safety Analysis <i>Mohamed Abdel-Aty and Jaeyoung Lee</i>	367
Motor Vehicle Crash Reportability <i>John J. McDonough</i>	380
Nominal Safety <i>Per Erik Garder</i>	386
Parking Lots <i>Maxim A. Dulebenets</i>	392
Passenger Van Safety <i>Saksith Chalermpong and Apiwat Ratanawaraha</i>	401
Passive Prevention Systems in Automobile Safety <i>B. Serpil Acar</i>	406
Pedestrian Safety, Children <i>Mette Møller</i>	415
Pedestrian Safety, General <i>Muhammad Z. Shah, Mehdi Moeinaddini, and Mahdi Aghaabbasi</i>	420
Pedestrian Safety, Older People <i>Carlo Lui</i>	429
Visually Impaired Pedestrian Safety <i>Robert S. Wall Emerson</i>	435
Photo/Video Traffic Enforcement <i>Charles M. Farmer</i>	439
Powered Two- and Three-Wheeler Safety <i>Fangrong Chang, Helai Huang, and Md. Mazharul Haque</i>	443
Railroad Safety <i>Xiang Liu and Zhipeng Zhang</i>	451
Railroad Safety: Grade Crossings and Trespassing <i>Rahim F. Benekohal and Jacob Mathew</i>	466
Recreational Boating Safety: Usage, Risk Factors, and the Prevention of Injury and Death <i>Amy E. Peden, Stacey Willcox-Pidgeon, and Kyra Hamilton</i>	477
Refuge Islands <i>Christer Hydén</i>	487
Risk Perception and Risk Behavior in the Context of Transportation <i>Martina Raue and Eva Lerner</i>	494

Road Diets <i>Robert B. Noland</i>	500
Road Safety Audits <i>Xiao Qin</i>	508
Roadside Safety Barriers <i>Dean C. Alberson</i>	513
Road Safety Management in Selected Countries <i>Paul Boase and Brian Jonah</i>	519
Roadway Pavement Conditions and Various Summer and Winter Maintenance Strategies <i>Kamal Hossain, PhD P. Eng</i>	529
Safety of Roundabouts <i>Khaled Shaaban</i>	539
Rumble Strips, Continuous Shoulder, and Centerline <i>AnnaAnund and AnnaVadeby</i>	549
Safety Culture <i>Tor-Olav Nvestad</i>	554
Safety Data Quality Management <i>Robert A. Scopatz</i>	560
School Bus Safety <i>Yousif A. Abulhassan</i>	564
School Campus Traffic Circulation <i>Dimitrios Nalmpantis</i>	568
Sexual Violence in Public Transportation <i>Vania Ceccato</i>	576
Shared Space <i>Allison B. Duncan</i>	584
Side Area Safety and Side Slopes <i>José María Pardillo-Mayora</i>	593
Simulators <i>Nichole L. Morris, Curtis M. Craig, Jacob D. Achtemeier, and Peter A. Easterlund</i>	602
Sleep-Related Issues and Fatigue <i>Prof Marion Sinclair and Estelle Swart</i>	611
Speed Governors and Limiters <i>Christer Hydén</i>	617
Speed Limits on Rural Highways <i>Peter Tarmo Savolainen and Timothy Jordan Gates</i>	622
Speed Limits on Urban Streets <i>Anna Bray Sharpin, Claudia Adriazola-Steil, Ben Welle, and Natalia Lleras</i>	632
Speed-Reducing Measures <i>Vinod Vasudevan</i>	641
Striping, Signs, and Other Forms of Information <i>Kasem Choocharukul and Kerkritt Sriroongvikrai</i>	648
Suicides <i>Brendan Ryan</i>	656

Surrogate Measures of Safety <i>Nicolas Saunier and Aliaksei Laureshyn</i>	662
Targeting Transit: the Terrorist Threat and the Challenges to Security <i>Brian Michael Jenkins</i>	668
The Swedish Vision Zero: A Policy Innovation <i>Matts-Åke Belin</i>	675
Tire Safety <i>Saied Taheri</i>	681
Town Gates: Section on Transport Safety and Security <i>Charles Tijus</i>	685
Traffic Flow Volume and Safety <i>Athanasios Theofilatos and Apostolos Ziakopoulos</i>	692
Traffic Safety and Security of Taxis and Ride-Hailing Vehicles <i>Zhe Wang, Helai Huang, and Ye Li</i>	699
Traffic Signals and Safety <i>Andrew P. Tarko</i>	706
Tunnels, Safety and Security Issues-Risk Assessment for Road Tunnels: State-of-the-Art Practices and Challenges <i>Konstantinos Kirytopoulos, Panagiotis Ntzeremes, and Konstantinos Kazaras</i>	713
Understanding, Managing, and Learning from Disruption <i>Karl Kim</i>	719
Use/Analysis of Crash Data and Underreporting of Crashes <i>Mohammadali Shirazi and Dominique Lord</i>	726
Utility Poles <i>Lai Zheng and Tarek Sayed</i>	731
Value of Life and Injuries <i>David A. Hensher</i>	737
Effects of Weather <i>Maria Pregnolato, Amirhassan Kermanshah, and Wisinee Wisetjindawat</i>	742
Wrong-Way Driving on Motorways <i>Huaguo Zhou and Md Atiquzzaman</i>	751

VOLUME 3

Freight Transport and Logistics <i>Sharon Cullinane and Kevin Cullinane</i>	1
Expanding the Perspective of Logistics and Supply Chain Management <i>David J. Closs</i>	8
Logistics and Supply Chain Management Performance Measures <i>David B. Grant and Sarah Shaw</i>	16
Economic Regulation/Deregulation and Nationalization/Privatization in Freight Transportation <i>Wayne K. Talley</i>	24
Freight Transport Policy <i>Luca Zamparini and Aura Reggiani</i>	29

Planning and Financing Logistics Spaces <i>Nicolas Raimbault</i>	35
Supply Chain Risk Management: Creating the Resilient Supply Chain <i>Richard Wilding</i>	41
Transportation Safety and Security <i>Maria G. Burns</i>	47
Resilience in Freight Transport Networks <i>Zhuohua Qu, Chengpeng Wan, and Zaili Yang</i>	53
Environmental Sustainability in Freight Transportation <i>Lisa M. Ellram</i>	58
Sustainable Logistics, CSR in Logistics, and Sustainable Supply Chain Management <i>Maria Björklund and Maja Piecyk-Ouellet</i>	64
Information Sharing and Business Analytics in Global Supply Chains <i>Prof Usha Ramanathan and Prof Ramakrishnan Ramanathan</i>	71
Logistics Information Systems <i>Petri Helo and Javad Rouzafzoon</i>	76
Factors Affecting the Selection of Logistics Service Providers <i>Aicha AGUEZZOUL</i>	85
Logistics Service Performance <i>Kee-hung Lai, Jinan Shao, and Yongyi Shou</i>	89
The World Bank's Logistics Performance Index <i>Christina K. Wiederer cwiederer, Jean-François Arvis, Lauri M. Ojala, and Tuomas M. M. Kiiski</i>	94
Outsourcing Logistics Functions <i>Evi Hartmann, Hendrik Birkel, and Matthias Kopyto</i>	102
Freight Transport and Logistics in JIT Systems <i>James H. Bookbinder and M. Ali Aælkü</i>	107
Supply Chain Finance <i>Erik Hofmann</i>	113
Packaging Logistics <i>Jesús García-Arca, Alicia Trinidad González-Portela Garrido, and J. Carlos Prado-Prado</i>	119
The Bullwhip Effect <i>Jan C. Fransoo and Maximiliano Udenio</i>	130
Blockchain Applications in Logistics <i>Yingli Wang</i>	136
Logistics in Asia <i>Shong-lee Ivan Su</i>	143
Logistics in the Developing World <i>Charles Kunaka</i>	150
Freight Network Modeling <i>Lóránt Tavasszy and Yousef Maknoon</i>	157
National Freight Transport Models <i>Gerard de Jong</i>	162
Freight Flows in Cities <i>Genevieve Giuliano</i>	168

Urban Logistics and Freight Transport <i>Michael Browne, José Holguín-Veras, and Julian Allen</i>	178
Omni-Channel Logistics <i>Tom Van Woensel</i>	184
Humanitarian Logistics <i>Gyöngyi Kovács and Diego Vega</i>	190
Optimization of Humanitarian Logistics <i>M. Teresa Ortuño, Jose M. Ferrer, Inmaculada Flores, and Gregorio Tirado</i>	195
Event Logistics <i>Rev. Ruth Dowson and Dan Lomax</i>	201
Reverse Logistics <i>Dale S. Rogers and Ronald S. Lembke</i>	208
E-Tailing and Reverse Logistics <i>Sharon Cullinane</i>	219
Green Routing of Freight Vehicles <i>Tolga Bektaş</i>	224
Freight Mode Choice <i>Hyun Chan Kim and Alan Nicholson</i>	231
The Value of Time in Freight Transport <i>María Feo-Valero, Amaya Vega, and Bárbara Vázquez-Paja</i>	236
Behavioral Research in Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	242
Container (Liner) Shipping <i>Theo Notteboom</i>	247
Bulk Shipping Markets: An Overview of Market Structure and Dynamics <i>Manolis G. Kavussanos and Stella A. Moysiadou</i>	257
Ferries and Short Sea Shipping <i>Lourdes Trujillo and Alba Martínez-López</i>	280
Shipping and the Environment <i>Karin Andersson, Selma Brynolf, Lena Granhag, and J. Fredrik Lindgren</i>	286
Energy Efficiency of Ships <i>Harilaos N. Psaraftis</i>	294
Seaports <i>Mary R. Brooks and Geraldine Knatz</i>	299
Port Hinterlands <i>Francesco Parola, Giovanni Satta, and Francesco Vitellaro</i>	305
Seaports as Clusters of Economic Activities <i>Peter W. de Langen</i>	310
Port Efficiency and Effectiveness <i>Lourdes Trujillo, María Manuela González, Casiano Manrique-De-Lara-Peñate, and Ivone Pérez</i>	316
Container Port Automation <i>Michael G.H. Bell</i>	323
Optimizing Crane Operations in Ports Scheduling of Liner Container Shipping Services	335

<i>Yuquan Du and Qiang Meng</i>	
Dry Ports	344
<i>Gordon Wilmsmeier and Jason Monios</i>	
Arctic Shipping	349
<i>Yufeng Lin, David G. Babb, and Adolf K.Y. Ng</i>	
Airfreight and Economic Development	355
<i>Kenneth Button</i>	
Air Freight Logistics	361
<i>Keith Debbage and Neil Debbage</i>	
Air Freight Marketing	369
<i>Lucy Budd and Stephen Ison</i>	
Drones in Freight Transport	374
<i>Oliver Kunze</i>	
Duty of Care in the Selection of Motor Carriers	382
<i>Thomas M. Corsi</i>	
Carrier Selection for Less-Than-Truckload (LTL) Shipments	388
<i>Dinçer Konur, Gonca Yildirim, and Bahriye Cesaret</i>	
Decarbonizing Road Freight Transport	395
<i>Heikki Liimatainen</i>	
The Rebound Effect in Road Freight Transport	402
<i>Tooraj Jamasb and Manuel Llorca</i>	
Autonomous Goods Transport	407
<i>Heike Flømig</i>	
Rail Freight	413
<i>Dr Allan Woodburn</i>	
Rail Freight Vehicles	423
<i>Maksym Spiryagin, Qing Wu, Peter Wolfs, Colin Cole, Valentyn Spiryagin, and Tim McSweeney</i>	
Eurasia Rail Freight: Enablers and Inhibitors of Future Growth	436
<i>Hendrik Rodemann and Simon Templar</i>	
Intermodal and Synchromodal Freight Transport	456
<i>Tomas Ambra, Koen Mommens, and Cathy Macharis</i>	
Pipelines	463
<i>Matthew E. Oliver</i>	
3D Printers and Transport	471
<i>Wouter P.C. Boon and Bert van Wee</i>	
The Physical Internet and Logistics	479
<i>Eric Ballot and Shenle PAN</i>	
Bicycles for Urban Freight	488
<i>Barbara Lenz and Johannes Gruber</i>	
China's Belt and Road Initiative	495
<i>Paul Tae-Woo Lee</i>	

VOLUME 4

Introduction to Traffic Management <i>Edward C.S. Chung</i>	1
Urban Motorway Management <i>John Gaffney and Hendrik Zurlinden</i>	2
Ramp Metering Application <i>John Gaffney and Hendrik Zurlinden</i>	10
City Wide Coordinated Ramp Meters <i>John Gaffney and Hendrik Zurlinden</i>	21
Variable Speed Limits for Traffic Efficiency Improvement <i>José Ramón D. Frejo and Bart De Schutter</i>	33
Hard Shoulder Running <i>Justin Geistefeldt</i>	41
High-Occupancy Vehicle (HOV) and High-Occupancy Toll (HOT) Lanes <i>Roxana J. Javid, Jiani Xie, Lijiao Wang, Wenruifan Yang, Ramina Jahanbakhsh Javid, and Mahmoud Salari</i>	45
Reversible Lanes: Guidelines, Operation and Control, Research Directions <i>Gowri Asaithambi, Venkatesan Kanagaraj, and Madhuri Kashyap</i>	52
Electronic Toll Collection <i>Azusa Toriumi</i>	60
Road Pricing-Theory and Applications <i>Kian Keong Chin</i>	68
Road Pricing 1: The Theory of Congestion Pricing <i>Timothy D. Hau</i>	74
Road Pricing 2: Short- and Long-Run Equilibrium of Road Transportation <i>Timothy D. Hau</i>	83
Road Pricing 3: The Implications for Pricing Public Transportation <i>Timothy D. Hau</i>	90
Road Pricing 4: Case Study-The Implementation of Electronic Road Pricing in Hong Kong <i>Timothy D.</i>	103
Advanced Travelers Information Systems (ATIS) <i>Chintan Advani and Ashish Bhaskar</i>	106
Travel Time Reliability <i>Sharmili Banik, Anil Kumar, and Lelitha Vanajakshi</i>	109
Traffic Incident Management <i>Ruimin Li</i>	122
Traffic Incident Detection <i>Shuyan Chen and Yingjiu Pan</i>	128
Bottleneck <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	134
Flow Breakdown <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	143
Recurrent Congestion <i>Takahiro Tsubota</i>	154

Nonrecurrent Congestion <i>Takahiro Tsubota</i>	158
Freeway to Arterial Interfaces <i>Abolfazl Karimpour and Yao-Jan Wu</i>	162
Arterial Road Management <i>Jiaqi Ma, Yi Guo and Adekunle Adebisi</i>	169
Signalized Intersections <i>Chaitrali Shirke</i>	178
Protected Phase <i>Jiarong Yao, Yumin Cao, and Keshuang Tang</i>	185
Permitted Phase <i>Yumin Cao, Jiarong Yao, and Keshuang Tang</i>	196
Hook Turns: Implementation, Benefits, and Limitations <i>Sara Moridpour and Amir Falamarzi</i>	203
Traffic Signal Coordination <i>Rahim F. Benekohal</i>	213
Emergency Vehicle Priority (Preemption): Concept and Advancements <i>Chaitrali Shirke</i>	221
Signalized Roundabouts <i>Yetis Sazi Murat and Rui-jun Guo</i>	227
Turbo Roundabouts: Design, Capacity and Comparison With Alternative Types of Roundabouts <i>Marco Guerrieri and Raffaele Mauro</i>	238
Capacity of an Intersection <i>Ashish Verma and Milan Mathew Thomas</i>	247
Delay <i>Shinji Tanaka</i>	258
Queue Length <i>Shinji Tanaka</i>	263
Local Area Traffic Management <i>Michael A.P. Taylor</i>	268
On-Street Parking <i>Jun Chen and Guang Yang</i>	278
Off-Street Parking <i>Jun Chen and Guang Yang</i>	285
Parking Information Systems <i>Behrang Assemi and Douglas Baker</i>	289
Performance-Based Parking Management <i>Douglas Baker and Behrang Assemi</i>	299
Transit Priority <i>Wanjing Ma, Qiheng Lin, and Ling Wang</i>	307
Adaptive Bus Control <i>Monica Menendez</i>	315
Transit Fare Collection <i>Mahmoud Mesbah and Kamal Khanali</i>	325

Transit Information Systems <i>Ankit Kumar Yadav and Nagendra R Velaga</i>	331
Tram Lane Configurations and Driving Rules <i>Farhana Naznin</i>	338
Railway Crossing <i>Chunliang Wu and Inhi Kim</i>	341
Pedestrian Crossing (Crosswalk) <i>Miho Iryo-Asano, Wael K.M. Alhajyaseen, and Koji Suzuki</i>	346
Bike-Sharing System: Uncovering the “1Success Factors” <i>S.K. Jason Chang and Amanda Fernandes Ferreira</i>	355
Airspace Systems Technologies-Overview and Opportunities <i>Banavar Sridhar and Gano B. Chatterji</i>	363
Air Traffic Flow and Capacity Management <i>Li Weigang and Cristiano P Garcia</i>	380
Port Management <i>Maria G. Burns</i>	390
Port Performance Measurement from a Multistakeholder Perspective <i>Min-Ho Ha, Zaili Yang, and Young-Joon Seo</i>	396
Transport Modeling and Data Management <i>Chandra Bhat</i>	407
Computational Methods and Data Analytics <i>Bilal Farooq and David López</i>	408
Activity-Based Models <i>Renato Guadamuz and Rajesh Paleti</i>	414
Advanced Traveler Information Systems <i>Stephen D. Boyles</i>	418
Demand-Responsive Transit, Evaluation Studies <i>Sebastián Raveau</i>	423
Location Choice Models <i>Adam Wilkinson Davis</i>	428
Full Feedback and Equilibrium Modeling in Urban Travel Forecasting <i>Yu (Marco) Nie, Jun Xie, and David Boyce</i>	432
The Use and Value of Geographic Information Systems in Transportation Modeling <i>Ming Zhang</i>	440
Latent Demand and Induced Travel <i>Charisma F. Choudhury</i>	448
ICT, Virtual and In-Person Activity Participation, and Travel Choice Analysis <i>Jacek Pawlak and Giovanni Circella</i>	452
Public Transit Ridership Forecasting Models <i>Ipsita Banerjee, Deepa L, and Abdul Rawoof Pinjari</i>	459
Spatial Mismatch, Job Access, and Reverse Commuting <i>Gian-Claudia Sciara</i>	468

Microsimulation and Agent-Based Models in Transportation <i>Milos Balac</i>	473
Choice Models in Transportation <i>Naveen Chandra Iraganaboina and Naveen Eluru</i>	477
Multi-Criteria Decision Analysis <i>Zhanmin Zhang and Srijith Balakrishnan</i>	485
The National Household Travel Survey Data Series (NPTS/NHTS) <i>Nancy McGuckin</i>	493
Route Choice and Network Modeling <i>Emma Frejinger and Maëlle Zimmermann</i>	496
Departure Time Choice Modeling <i>Khandker Nurul Habib</i>	504
Traveler Responses to Congestion <i>David T. Ory and Gayathri Shivaraman</i>	509
Origin-Destination Demand Estimation Models <i>William H.K. Lam, Hu Shao, Shuhan Cao, and Hai Yang</i>	515
Parking Demand Models <i>S.C. Wong, Zhi-Chun Li, and William H.K. Lam</i>	519
Pavement Management Systems <i>Senthilmurugan Thyagarajan</i>	524
Residential Location Choice Models <i>Shlomo Bekhor and Sigal Kaplan</i>	531
Transport Demand Management <i>Feiyang Zhang and Becky P.Y. Loo</i>	537
Traffic Flow Analysis <i>H. Michael Zhang and Jia Li</i>	544
Autonomous Vehicles and Transportation Modeling <i>Annesha Enam, Felipe de Souza, Omer Verbas, Monique Stinson, and Joshua Auld</i>	557
Ride-Hailing and Travel Demand Implications <i>Felipe F. Dias</i>	564
Transportation Modeling and Planning Software <i>Joel Freedman</i>	569
Transportation Statistics and Databases <i>Taha Hossein Rashidi</i>	574
Travel Surveys <i>Stacey G. Bricka</i>	587
Travel Demand Forecasting: Where Are We and What Are the Emerging Issues <i>Thomas F. Rossi</i>	590
Travel Model Calibration and Validation <i>Ram M. Pendyala</i>	596
Trip Chaining Analysis <i>Cynthia Chen and Yusak Susilo</i>	606
Vehicle Ownership Models <i>Dr. Sören Groth and Prof. Dr. Dirk Wittowsky</i>	612

Bicycle Sharing/Bikesharing <i>Catherine Morency and Jean-Simon Bourdeau</i>	617
Carsharing <i>Shiva Habibi and Frances Sprei</i>	623
Urban Recreational Travel <i>Long Cheng and Frank Witlox</i>	629
Place Perception and Travel Behavior <i>Kathleen Deutsch-Burgner and Konstadinos G. Goulias</i>	635

VOLUME 5

Transport Modes <i>Edoardo Marcucci</i>	1
Infrastructure Transport Investments, Economic Growth and Regional Convergence <i>Xavier Fageda and Cecilia Olivieri</i>	2
Transport Modes and an Aging Society <i>Charles B.A. Musselwhite and Theresa Scott</i>	6
Sustainable Mobility Paths <i>Erling Holden and Geoffrey Gilpin</i>	13
Vehicles that Drive Themselves: What to Expect with Autonomous Vehicles <i>Michele D. Simoni and Kara M. Kockelman</i>	19
Transport Modes and Tourism <i>Ila Maltese and Luca Zamparini</i>	26
Transport Modes and Accessibility <i>Bert van Wee</i>	32
Transport Modes and Globalization <i>Jean-Paul Rodrigue</i>	38
Transport Modes and Cities <i>Erick Guerra and Gilles Duranton</i>	45
Transport Modes and Remote Areas	51
Modeling Mode Choice in Freight Transport <i>Lóránt Tavasszya and Gerard de Jongb</i>	57
Travel Mode Choice as Reasoned Action <i>Sebastian Bamberg, Icek Ajzen, and Peter Schmidt</i>	63
Energy Consumption of Transport Modes <i>Zissis Samaras and Ilias Vouitsis</i>	71
Transport Modes and People With Limited Mobility <i>Roger L Mackett</i>	85
Transport Modes and Commuters <i>Colin G. Pooley</i>	92
Shopping and Transport Modes <i>Antonio Comi</i>	98
Transport Modes and Health <i>Jennifer S. Mindell and Sandra Mandic</i>	106

Multimodality in Transportation <i>Sören Groth and Tobias Kuhnimhof</i>	118
Transport Modes and Disasters <i>Brian Wolshon</i>	127
Big Data for Public Transport Planning <i>Jan-Dirk Schmöcker</i>	134
Active Transport: Heterogeneous Street Users Serving Movement and Place Functions <i>Regine Gerike, Stefan Hubrich, Caroline Koszowski, Bettina Schröter, and Rico Wittwer</i>	140
Electric Vehicles <i>Christine Eisenmann, Daniel Görges, and Thomas Franke</i>	147
Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service <i>Susan Shaheen, PhD and Adam Cohen</i>	155
Adoption of new travel information platforms <i>Sigal Kaplan</i>	160
ICT and Transport Modes <i>Galit Cohen-Blankshtain</i>	165
Mode Choice and Life Events <i>Joachim Scheiner</i>	171
Introduction to Air Transport <i>Milan Janić</i>	178
The History of Air Transportation <i>Richard P. Hallion</i>	192
The Geography of Air Transport <i>Lucy Budd and Stephen Ison</i>	198
The Future of Air Transport <i>Rico Merkert and James Bushell</i>	203
Air Transport and Its Territorial Implications <i>Lanfranco Senn</i>	208
Next Generation Travel: Young Adults' Travel Patterns <i>Tobias Kuhnimhof and Scott Le Vine</i>	215
Airport Network Planning and Its Integration with the HSR System <i>Francesca Pagliara, Juan Carlos Martín, and Concepción Román</i>	222
Airport Management <i>Peter Forsyth and Hans-Martin Niemeier</i>	229
Airport Regulation <i>Achim I. Czerny</i>	234
Air Route Planning and Development <i>Renan Peres de Oliveira and Gui Lohmann</i>	241
Airline Management <i>Sveinn Vidar Gudmundsson</i>	249
Air Cargo <i>Volodymyr Bilotkach</i>	258

Airline Regulation <i>Andrew R. Goetz</i>	263
Air Vehicles Classification <i>Vincenzo Torre</i>	269
Aircraft Manufacturing <i>Antonio Sollo</i>	290
A Geography of Road Transport in Cities <i>Cristian Domarchi and Juan de Dios Ortúzar</i>	300
The Future of Road Transport <i>Preston L. Schiller</i>	306
Road Transportation and Territorial Scale <i>Ana M. Condeço-Melhorado</i>	315
Road Modes: Walking <i>Kevin Manaugh, PhD Associate Professor</i>	320
Bus Public Transport Planning and Operations <i>Ehab Diab and Ahmed El-Geneidy</i>	326
Street Design for Active Travel <i>Bruce Appleyard</i>	333
Transit Planning and Management <i>Zakhary Mallett and Marlon G Boarnet</i>	349
Road Infrastructure: Planning, Impact and Management <i>Jos Arts, Wim Leendertse, and Taede Tillema</i>	360
Road Transport Planning at the Urban Scale <i>David A. King and Kevin J. Krizek</i>	373
Road Traffic Regulation: Road Pricing and Environmental Quality <i>Marco Percoco</i>	378
Car Ownership and Car Use: A Psychological Perspective <i>J.L. Veldstra, A.B. Ünal, E.M. Steg</i>	384
Road Transport: E-Scooters <i>Gysele Lima Ricci and Klaus Bogenberger</i>	391
Railway Station and Network Planning <i>Ingo Arne Hansen</i>	399
Introduction to Rail Transport <i>Chris Nash and Tony Fowkes</i>	406
The History of Rail Transport <i>Carlo Ciccarelli, Andrea Giuntini, and Peter Groote</i>	413
The Geography of Rail Transport <i>Frédéric Dobruszkes and Amparo Moyano</i>	427
Rail Transport and Territorial Scale <i>Prof. Andrés Monzón and Dr. Elena López</i>	437
Railway Management <i>Vassilios A. Profillidis</i>	444
Railway Terminal Regulation <i>Nacima Baron</i>	454

Service Network Design for Freight Railroads <i>Teodor Gabriel Crainic</i>	464
Subway Systems <i>Guillaume Monchambert, Daniel Hörcher, Alejandro Tirachini, and Nicolas Coulombel</i>	471
Railway Company Management <i>Vilius Nikitinas, Skaistė Miliauskaitė</i>	479
Regulation of Rail Infrastructure and Services <i>Javier Campos</i>	485
Rail Vehicle Classification <i>Christos Pyrgidis and Alexandros Dolianitis</i>	490
Introduction to Maritime Shipping <i>Christa Sys and Thierry Vanelslander</i>	508
The Geography of Maritime Transport <i>César Ducruet and Justin Berli</i>	517
The Future of Maritime Transport <i>Harilaos N. Psaraftis</i>	535
Maritime Transport and Territorial Scale <i>Brian Slack</i>	540
Containerization and the Port Industry <i>Hercules Haralambides</i>	545
Port Management <i>Lourdes Trujillo, Daniel Castillo Hidalgo, and Manuel Herrera</i>	557
Economic and Environmental Regulation in the Port Sector <i>Beatriz Tovar and Alan Wall</i>	563
Maritime Route Planning <i>Johan Woxenius</i>	570
Inland Waterway Transport and Inland Ports: An Overview of Synchromodal Concepts, Drivers, and Success Cases in the IWW Sector <i>Behzad Behdani, Bart Wiegman, and Yun Fan</i>	577
Maritime Company Passenger Management/Liner Industry <i>Claudio Ferrari and Alessio Tei</i>	587
Cruise Industry <i>Athanasios A. Pallis and Aimilia A. Papachristou</i>	593
International Maritime Regulation: Closing the Gaps Between Successful Achievements and Persistent Insufficiencies <i>Laurent Fedi</i>	600
Ship Classification <i>Gareth C. Burton and Mimosa T. Miller</i>	607
The Shipbuilding Industry and its Interactions With Shipping <i>Paul William Stott</i>	617
Methods for Designing Public Transport Networks <i>Zain Ul Abedin and Avishai (Avi) Ceder</i>	625
Space Transportation <i>Mark Hempsell</i>	638

Pipelines	646
<i>Franco Cotana and Mattia Manni</i>	
Women and Transport Modes	656
<i>Priya Uteng, PhD and Yusak Susilo, PhD</i>	
Transport Modes and Big Data	665
<i>Hannah D Budnitz, Emmanouil Tranos, and Lee Chapman</i>	
Transport Modes and Inequalities	671
<i>Caroline Mullen</i>	
Railway Traffic Management	678
<i>Francesco Corman</i>	
Indoor Transportation	684
<i>Lutfi Al-Sharif</i>	
Bicycle as a Transportation Mode	697
<i>Raktim Mitra and Paul M. Hess</i>	
Urban Air Mobility: Opportunities and Obstacles	702
<i>Adam Cohen and Susan Shaheen, PhD</i>	
Transport Modes and Sustainability	710
<i>Long Cheng, Jonas De Vos, and Frank Witlox</i>	

VOLUME 6

Introduction to Transport Policy and Planning	1
<i>Maria Attard</i>	
Workplace Parking Levy	2
<i>Stephen Ison and Lucy Budd</i>	
Air Transport	7
<i>Lucy Budd and Stephen Ison</i>	
Mobility as a Service	12
<i>Milo N. Mladenović</i>	
Bicycle Sharing	19
<i>Cyrille Médard de Chardon</i>	
Light Rail	31
<i>Fiona Ferbrache</i>	
Planning Tourism Travel	39
<i>Luca Zamparini</i>	
Transport Planning and Management and its Implications in Chinese Cities	44
<i>Mengqiu Cao</i>	
Road Safety	51
<i>George Yannis and Eleonora Papadimitriou</i>	
Mobility Planning and Policies for Older People	59
<i>Charles B A Musselwhite</i>	
Electric Mobility	64
<i>Graham Parkhurst</i>	
Transport Policy and Governance	73
<i>Lisa Hansson</i>	

Transferability of Urban Policy Measures <i>Paul Martin Timms</i>	77
Accessibility Tools for Transport Policy and Planning <i>Benjamin Büttner</i>	83
Demand Responsive Transport <i>Marcus Enoch</i>	87
Transport and Climate Change <i>Robin Hickman and Christine Hannigan</i>	94
Taxicabs and Microtransit <i>David A. King</i>	101
Land-Use and Transport Planning <i>Luis A. Guzman</i>	107
Parking <i>Stephen Ison and Lucy Budd</i>	113
Transport Planning in the Global South <i>Daniel Oviedo and Mariajose Nieto-Combariza</i>	118
Planning for Children's Independent Mobility <i>E. Owen D. Waygood and Raktim Mitra</i>	125
The Politics of Mobility Policy <i>Geoff Vigar</i>	131
Technology Enabled Data for Sustainable Transport Policy <i>Susan M. Grant-Muller, Mahmoud Abdelrazek, Hannah Budnitz, Caitlin D. Cottrill, Fiona Crawford, Charisma F. Choudhury, Teddy Cunningham, Gillian Harrison, Frances C. Hodgson, Jinhyun Hong, Adam Martin, Oliver O'Brien, Claire Papaix, and Panagiotis Tsoleridis</i>	135
Car Sharing <i>Cyriac George and Tanu Priya Uteng</i>	142
Toward a More Holistic Understanding of Mega Transport Project (MTP) Success <i>John Ward</i>	147
Equity Considerations in Transport Planning <i>Karel Martens</i>	154
Planning for Rail Transport <i>Simon P. Blainey</i>	161
Connected and Autonomous Vehicles: Priorities for Policy and Planning <i>Dr. Alexandros Nikitas</i>	167
Gendered Mobility <i>Sheila Mitra-Sarkar</i>	173
Modeling and Simulation for Transport Planning <i>Michela Le Pira, Giuseppe Inturri, and Matteo Ignaccolo</i>	184
Externalities and External Costs in Transport Planning <i>Silvio Nocera</i>	191
Planning for Public Transport with Automated Vehicles <i>Gonçalo Homem de Almeida Rodriguez Correia</i>	198
Urban Congestion Charging in Transport Planning Practice <i>Ida Kristoffersson and Maria Börjesson</i>	206

Energy and Transport Planning <i>Debbie Hopkins and Christian Brand</i>	214
Customer Satisfaction as a Measure of Service Quality in Public Transport Planning <i>Laura Eboli and Gabriella Mazzulla</i>	220
Car Sharing and the Impact on New Car Registration <i>Mario Intini and Marco Percoco</i>	225
Evaluation Methods in Transport Policy and Planning <i>Niek Mouter</i>	230
Transport and Air Quality Planning and Policy <i>Dr Fabio Galatioto</i>	236
Cycling Policies <i>Esther Anaya-Boig</i>	241
Community Severance <i>Paulo Anciaes and Jennifer S. Mindell</i>	246
Planning for Bus Priority <i>Claus H. Sørensen, Fredrik Pettersson, and Joel Hansson</i>	254
High-Speed Rail and the City <i>Marie Delaplace</i>	261
Public Engagement in Transport Planning <i>Miriam Ricci</i>	266
Long-Distance Travel <i>Giulio Mattioli and Muhammad Adeel</i>	272
Urban Freight Policy <i>Laetitia Dablanç</i>	278
Regional Transport Planning <i>Chia-Lin Chen</i>	286
Transport Project Financing <i>Romeo Danielis and Lucia Rotaris</i>	292
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	298
Transitions and Disruptive Technologies in Transport Planning <i>Kate Pangbourne and Maria Attard</i>	309
Community Transport: Filling the Gaps for Those in Need of Mobility <i>Ian Shergold</i>	314
Fundamental Emerging Concepts and Trends for Environmental Friendly Urban Goods Distribution Systems <i>Sandra Melo</i>	320
Plateau Car <i>David Metz</i>	324
Emerging Trends in Transport Demand Modeling in the Transition Toward Shared Mobility and Autonomy <i>Patrizia Franco</i>	331
Policy and Planning for Walkability <i>Carlos Cañas Sanz and Maria Attard</i>	340

Public Transport Subsidy and Regulation <i>Jonathan Cowie</i>	349
Urban Regeneration and Transportation Planning <i>Thomas Vanoutrive</i>	356
Social and Distributional Impact Assessment in Transport Policy <i>Laura Walker and Angela Curl</i>	361
Mobility Planning for Healthy Cities <i>Ersilia Verlinghieri</i>	368
Transport Demand Management <i>Begoña Guirao</i>	374
Planning and the Global Movement of Goods and Commodities <i>Christopher Clott and Chris Petrocelli</i>	380
Public Transport Network Planning <i>Corinne Mulley and John D. Nelson</i>	388
A Timely Perspective on Planning for Ageing Infrastructure <i>Anthony Perl</i>	395
The role of media in transport planning and the transport policy process <i>Özgül Ardiç and J.A. Annema</i>	400
Travel Plans <i>Stephen Potter and Marcus Enoch</i>	408
Home Deliveries and their Impact on Planning and Policy <i>Rosário Macário</i>	413
Planning for Safe and Secure Transport Infrastructure <i>Per Erik Gårder</i>	418
Sensors and Data Driven Approaches in Transport <i>Mohammad Sadrani and Constantinos Antoniou</i>	426
Planning Active Travel and School Transport <i>Fahimeh Khalaj, Dorina Pojani, and Sara Alidoust</i>	432
VOLUME 7	
Introduction to Transport Psychology <i>Carlo Prato</i>	1
From Self-reports to Auto-Tech-Detect (ATD)-based Self-reports in Traffic Research <i>Türker Özkan and Timo Lajunen</i>	2
Observational Field Studies in Traffic Psychology <i>Tova Rosenbloom and Hodaya Levy</i>	8
Driving Simulators <i>Karel A. Brookhuis</i>	14
Naturalistic Driving Studies: An Overview and International Perspective <i>Johnathon P. Ehsani, Joanne L. Harbluk, Jonas Bärghman, Ann Williamson, Jeffrey P. Michael, Raphael Grzebieta, Jake Olivier, Jan Eusebio, Judith Charlton, Sjaanie Koppel, Kristie Young, Mike Lenné, Narelle Haworth, Andry Rakotonirainy, Mohammed Elhenawy, Gregoire Larue, Teresa Senserrick, Jeremy Woolley, Mario Mongiardini, Christopher Stokes, Paul Boase, John Pearson, and Feng Guo</i>	20

A Detailed Approach to Qualitative Research Methods <i>Sonja Forward and Lena Levin</i>	39
Behavioral Change <i>Sonja Haustein</i>	46
Habitual Behavior <i>Carlo G. Prato</i>	54
Driving Behavior and Skills <i>Timo Lajunen and Türker Özkan</i>	59
The Multidimensional Driving Style Inventory <i>Orit Taubman - Ben-Ari</i>	65
Risk Perception in Transport: A Review of the State of the Art <i>Trond Nordfjrn, An-Magritt Kummeneje, Mohsen F. Zavareh, Milad Mehdizadeh, and Torbjørn Rundmo</i>	74
Social-Symbolic and Affective Aspects of Car Ownership and Use <i>Birgitta Gatersleben</i>	81
Pitfalls of Statistical Methods in Traffic Psychology <i>J.C.F. de Winter and D. Dodou</i>	87
Data Analysis: Structural Equation Models <i>Marco Diana</i>	96
Data Analysis: Integrated Choice and Latent Variable Models <i>Carlo G Prato</i>	102
Explaining Data Analysis Using Qualitative Methods <i>Lena Levin and Sonja Forward</i>	107
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	113
Driver Aggression and Anger <i>Mark J.M. Sullman and Amanda N. Stephens</i>	121
Speeding: A "Tragedy of the Commons" Behavior <i>Bryan E. Porter, Thomas D. Berry, and Kristie L. Johnson</i>	130
Cycling as a Mode Choice: Motivational Psychology <i>Sigal Kaplan</i>	136
Motorcyclists <i>Narelle Haworth</i>	144
Drivers' Hazard Perception Skill <i>Mark S. Horswill and Andrew Hill</i>	151
Driver Education and Training for New Drivers: Moving beyond Current 'Wisdom' to New Directions <i>Teresa Senserrick, Oscar Oviedo-Trespalcacios, David Rodwell, and Sherrie-Anne Kaye</i>	158
Road Safety Advertising: What We Currently Know and Where to From Here <i>Ioni Lewis, Barry Watson, Katherine M. White, and Sonali Nandavar</i>	165
Traffic Law Enforcement Theories and Models <i>Richard Tay</i>	171
Satisfaction with Travel and the Relationship to Well-Being <i>Tommy Gørling and Filip Fors Connolly</i>	177
	182

Electromobility: History, Definitions and an Overview of Psychological Research on a Sustainable Mobility System <i>Josef F. Krems and Isabel KreiÃŸig</i>	
Sharing: Attitudes and Perceptions <i>Yi Wen and Christopher R Cherry</i>	187
Women's Travel Patterns, Attitudes, and Constraints Around the World <i>Sandra Rosenbloom</i>	193
Driver Stress and Driving Performance <i>Lisa Dorn</i>	203
Introduction to Sustainability and Health in Transportation <i>Roger Vickerman</i>	225
Sustainable Development Goals and Health <i>Rosa Surinach</i>	226
Human Ecology <i>Roderick J. Lawrence</i>	234
Car- Free Cities <i>Haneen Khreis and Mark J. Nieuwenhuijsen</i>	240
Superblocks Base of a New Model of Mobility and Public Space. Barcelona as an Example <i>Salvador Rueda Palenzuela</i>	249
Resilience of Transport Systems <i>ErikJenelius and Lars-GöranMattsson</i>	258
Impact of Shipping to Atmospheric Pollutants: State-of-the-Art and Perspectives <i>Daniele Contini and Eva Merico</i>	268
Noise Pollution From Transport <i>Marianna Jacyna, Emilian Szczepański, Konrad Lewczuk, Mariusz Izdebski, Ilona Jacyna-Gotda, Michał, Kłodawski, Paweł, Gołda, Piotr Goł, and Ebiowski</i>	277
Visual Impacts From Transport <i>Paulo Rui Anciaes</i>	285
Light Pollution <i>John D. Bullough</i>	292
Wildlife Crossings and Barriers <i>Scott D. Jackson</i>	297
Environmental Justice, Transport Justice, and Mobility Justice <i>Devajyoti Deka</i>	305
Transport Noise and Health <i>Elisabete F. Freitas, Emanuel A. Sousa, and Carlos C. Silva</i>	311
Climate Change and Health, Related to Transport <i>Ersilia Verlinghieri</i>	320
Urban Greenspace, Transportation, and Health <i>Payam Dadvand and Mark J. Nieuwenhuijsen</i>	327
Transport Access and Health <i>Alireza Ermagun</i>	335
Social Exclusion and Health, Related to Transport <i>Roger L. Mackett</i>	341

Burden of Disease Assessment <i>David Rojas-Rueda</i>	347
Achieving a Near-Zero CO <i>Lewis M. Fulton</i>	353
Disabled Travelers <i>Bryan Matthews</i>	359
Health Impacts of Connected and Autonomous Vehicles <i>Soheil Sohrabi</i>	364
Electric Vehicles and Health <i>Kanok Boriboonsomsin</i>	372
Shared Mobility Opportunities and their Computational Challenges for Improving Health-Related Quality of Life <i>Cristiano Martins Monteiro, Cláudia Aparecida Soares Machado, Adelaide Cassia Nardocci, Fernando Tobal Berssaneti, José Alberto Quintanilha, and Clodoveu Augusto Davis</i>	376
Bike Sharing and Health <i>David Rojas-Rueda and Mark J. Nieuwenhuijsen</i>	384
E-Bikes and Health <i>Aslak Fyhri and Hanne Beate Sundfør</i>	393
<i>Index</i>	399

Transport Modes

Edoardo Marcucci, Department of Political Sciences, University of Roma Tre, Rome, Italy; Department of Logistics, Molde University College, Molde, Norway

© 2021 Elsevier Ltd. All rights reserved.

“Transport Modes” constitutes one of the nine sections composing the Encyclopedia on Transport.

Be it passenger or freight, transport takes place using different modes to move people/goods from A to B. This distinction is blurred when both passengers and freight (e.g., free luggage capacity in passenger planes is used to ship goods) jointly use the same mode. The transport modes typically considered are road, rail, air and maritime.

The distinctions among the modes relates to issues that are technical (e.g., speed, capacity), technological (e.g., ICE/Electric), operational (e.g., speed/weight limits, safety/security conditions, operating hours) and commercial (e.g., transport demand, ownership).

Mode choice and mode shift are fundamental topics to investigate since they have relevant implications for different transport--related aspects such as cost, time, reliability, frequency, punctuality, resilience, emissions and accidents. Agents' mode choice strictly correlates with other topics the other Sections of the Encyclopedia investigate in detail. To facilitate the detection of these connections, each article proposes, at the end, a suggested “reading map” by indicating the conceptually closer articles in the Encyclopedia (i.e., cross references).

To maximize readers' benefit when using the Encyclopedia, what follows briefly illustrates the planning phase of this Section by discussing the logic adopted in defining the article types and content the Section strived to put together.

When I accepted the invitation to coordinate this section of the Encyclopedia I decided to adopt a potential-reader-perspective and asked myself about the main type of issues/categories somebody reverting to an Encyclopedia might be interested in discovering. After a lengthy period of critical thinking, inspiration taken from other Encyclopedias, conversations with colleagues and friends (e.g., Prof. Valerio Gatta, and Dr. Ila Maltese) and, of course, many changes brought to the original structure, I adopted the following framework that was, at least from my standpoint, useful in defining the type and number of articles to include.

The structure of the section rests on the definition of the most important transport modes, which are the “classical” ones: road, rail, air and maritime.

With the intent of providing a systematic and consistent treatment of each mode, the Section unfolds by adopting a hierarchical structure so to provide a clear perspective and avoid jeopardizing consistency. In more detail, the hierarchy adopted is the following.

Type A papers can be defined generic since they deal with issues from a broad transport modes perspective (e.g., ageing society and transport modes). These papers have a “bridging” nature with respect to other sections in the Encyclopedia.

Type B papers deal with a single transport mode. They set the overall picture of a given mode by introducing the current situation, history, future, geography and its territorial scale.

Type C papers deal specifically with “infrastructure” and “vehicles” with attention paid to both terminals and lines. Whenever appropriate, both for infrastructure and vehicles, the papers investigate: (1) planning, (2) management, (3) regulation, (4) design, (5) manufacturing, and (6) ITS issues.

The full combination of all these aspects would have produced approximately 145 papers that were way above the maximum number of papers allotted to this Section and above the potential number of articles any reader would bear to read just on transport modes. Due to various reasons linked both to the level of detail and availability of potential Authors, some “leaves” of this tree were pruned leaving us with a total of 93 papers in its final version. This is still a large number of papers to read on transport modes. The articles in this Section provide a comprehensive, updated, coherent and well-articulated set of sources on transport modes that will represent a reference point for all those readers that need an introduction to a specific issue and a guidance for additional readings.

I do hope you will enjoy reading the articles (and linking them to others within this Section and across the Encyclopedia) as much as I did enjoy conceiving, commissioning and revising them.

Infrastructure Transport Investments, Economic Growth and Regional Convergence

Xavier Fageda*, **Cecilia Olivieri†**, *Department of Applied Economics and GIM-IREA, Universitat de Barcelona, Barcelona, Spain; †Department of Economics, Faculty of Social Sciences, Universidad de la República, Montevideo, Uruguay

© 2021 Elsevier Ltd. All rights reserved.

Introduction	2
Transport Infrastructure Investments and Economic Growth	2
Spatial Spillovers	3
The Impact of Transport Investments on Regional Convergence	4
Conclusions	4
References	5

Introduction

Investments in transport infrastructures may have strong economic effects since better transport infrastructures lead to a greater outlay of public capital and, hence, the higher productivity of private factors, fewer transport costs for firms, and greater accessibility to territories. In this regard, there is an extensive theoretical literature on identifying the multiple causal mechanisms that link transport and economic growth (see [Lakshmanan, 2011](#) for an in-depth review). The main arguments rely on market expansion, cheaper inputs for production, competition, gains from trade, technological changes, agglomeration forces, and processes of innovation and commercialization of new knowledge in spatial clusters. However, the analysis of transport infrastructure investments on regions requires that the effects related to growth be distinguished from those related to the reorganization of economic activity ([Redding and Turner, 2014](#)).

Indeed, we may expect that they will have a positive aggregate economic effect being the magnitude of the impact an empirical question. However, it is less clear whether transport investments contribute to reduce regional disparities. Recent theoretical literature, under the assumption of increasing returns to scale and imperfect competition, have helped to explain, in a common framework, both divergence and convergence processes in a context of a decrease of transportation costs. The main contribution of the so-called theory of the “new economic geography” was that it provided an explanation of the way in which, a priori similar regions, may diverge in their production structures and income levels due to their different access to markets. These location approaches, basically analyzed the ambiguity of the effects on different regions of the reduction of transportation costs ([Puga, 2002](#)). In this regard, the contribution of infrastructure investments—with the consequent effect on the reduction of transport costs—to regional convergence (or divergence) is related to the degree of reduction of such costs and to the conditions of the labor market.

A related issue is the sign and magnitude of the spatial spillovers since it can be expected that investments in transport infrastructures not only have an effect on the territory where they take place but also in nearby areas. Furthermore, we may expect heterogeneous impacts of transport investments according to the concerned mode (roads, rails, ports, and airports) and the characteristics of regions and sectors affected.

In this chapter, we review the empirical literature on the economic effects of transport investments. The rest of this chapter is organized as follows. Section 2 examines the studies that examine the link between transport investments and economic growth. Section 3 analyzes the results in the existing literature on the spatial spillovers of transport investments. Section 4 reviews studies that examine the contribution of transportation to reduce regional disparities. In all cases, we put particular attention on the results obtained for specific transportation modes. Finally, Section 5 provides some final considerations.

Transport Infrastructure Investments and Economic Growth

Several empirical studies have used production functions to examine the impact of infrastructures on economic growth. The geographical unit of analysis has been quite diverse and so studies conducted at the country, regional, and local levels are found. In general, the focus of these studies is to estimate the aggregate impact of the stock of public capital on gross domestic product per capita. Early examples of studies using production functions include [Aschauer \(1989\)](#), [Munnell \(1990\)](#), [García-Milà and McGuire \(1992\)](#), [Holtz-Eakin and Schwartz, \(1995\)](#), and [Fernald \(1999\)](#). Some recent contributions have focused on a case study or a single region study, such as those undertaken by [Cohen \(2010\)](#) or [Berechman et al. \(2006\)](#) for the United States; [Crafts \(2009\)](#) for the United Kingdom; [Fan and Chan-Kang \(2008\)](#), [Lean et al. \(2014\)](#), [Yu et al. \(2012\)](#) for China; [Park and Seo \(2016\)](#) for Korea; [Lall](#)

(2007) for India; Yamaguchi (2007) for Japan; Crescenzi and Rodríguez-Pose (2012) for the European Union, and Farhadi (2015) for the OECD. However, the analysis of comparative data for a large number of countries can be valuable in assessing how transport impacts on the development process in heterogeneous economic contexts (Park et al., 2019).

Given that roads represent a high proportion of the total stock of public capital, studies using production functions typically focus the attention on roads. While it is generally accepted that public capital (above all that related with roads) has positive effects on gross domestic product per capita, the magnitude of these effects is unclear. In general, the impact found is smaller in those studies that use more sophisticated econometric techniques like cointegration models that deal with spurious correlation, instrumental variables models that deal with endogenous biases or models that include time and spatial lags.

In contrast to the extensive literature examining the link between public capital and output, few studies focus on the impact of transport infrastructures on other relevant aspects like costs at the sector level, innovation, location choices of firms, or regional employment. The focus on production functions may be explained by the more widely availability of data. In this regard, a few studies estimate cost functions (Morrison and Schwartz, 1996; Nadiri and Mamuneas, 1994) and the analysis on regional employment concentrates on one specific mode of transportation at the region or city level. While some studies about the impact of road investments on regional employment do not find a strong effect (Clark and Murphy, 1996; Jiwattanakulpaisarn et al., 2010), those works employing sophisticated identification strategies find a relevant effect (Akyelken, 2015; Duranton and Turner, 2012; Möller and Zierer, 2018; Percoco, 2016; Rephann and Isserman, 1994; Xu and Nakajima, 2017). A positive and significant impact of roads in studies that examine the effects on, for example, the location of manufacturing establishments (Holl, 2004), wages (Matas et al., 2013), regional innovation (Agrawal et al., 2017), or population (Duranton, 2016) have also been found.

Regarding the specific impact of high-speed rails, if it can be said to exist, seems to be limited to the urban core and the area around the station (see Givoni, 2006 and Albalade and Bel, 2012 for detailed reviews). In contrast, Donaldson (2018) and Banerjee et al. (2020) find that the extension of the railroad network in India and China, respectively, has had a positive causal effect on income.

The analyses of the impacts of ports and airports have generally focused on the causal link between different indicators of traffic and some measure of regional or urban performance, so that the impact of investments is indirectly measured through their expected positive effects on those indicators of traffic. In this regard, Bottasso et al. (2013) and Ferrari et al. (2010) report that port throughput has a positive impact on employment while that Brueckner (2003), Bel and Fageda (2008), Percoco (2010), Blonigen and Cristea (2015), Bilotkach (2015), or Albalade and Fageda (2016) find evidence of a strong effect of air traffic on knowledge intensive services.

In general, the evidence about the impact of transport infrastructure has focused on the specific effect of one single mode. Thus, only a few studies have compared the role of different types of transport infrastructure together. Along these lines, Melo et al. (2013) conduct a meta-analysis based on a sample of 563 estimates from 33 previous studies. Their results indicate that the impact of transport infrastructure may vary among countries and type of transport mode, tending to be higher for the United States and for roads. Holmgren and Merkel (2017) conducted a meta-analysis to examine the association between transport infrastructure and economic growth based on 776 estimates from previous studies. They found that the economic sectors are affected differently by the investment in modes of transport infrastructure. In particular, roads have a positive impact on manufacturing and construction, while seaports affect the agricultural sector. Park et al. (2019) examine the role of various types of transport infrastructure on the economic growth of OECD and non-OECD countries. The findings show stronger significance of seaports on economic growth than airports and surface transportation modes, particularly in developing countries.

Spatial Spillovers

Given that the spatial spillovers between regions may be especially relevant in the analysis of economic impacts of transport infrastructures, some studies employ spatial econometric techniques. In this regard, the omission of the spatial interactions may lead to substantial biases in the estimated economic impact of transport infrastructure investments (Berechman et al., 2006; Cohen, 2010). The consideration of spatial interdependencies usually reduces the magnitude of the impacts associated to transport investments and may explain that those impacts are smaller when the analysis focuses on regions instead of countries.

The most common approach is to estimate production functions including as covariates spatial lags of the dependent and explanatory variables. Spatial econometric techniques allow examining the direct effects on the areas in which the infrastructure is located and the spillover effects on neighboring regions. To this point, the sign of the impact of a better transport infrastructure endowment in one region on its neighbors is not, a priori, clear. Indeed, improved infrastructure may give rise to a competition effect associated with the agglomeration of activities in the region with better infrastructure and a complementary effect associated with improved access to other regions or international markets.

In contrast to the extensive literature on direct effects of transport investments, few studies analyze their indirect effects. For example, Delgado and Álvarez (2007) and Moreno and López-Bazo (2007) find evidence of negative indirect effects of road investments in Spain. A negative indirect effect of roads is also found in the study of Lo Cascio et al. (2019) for Italy. This latter study and Percoco (2010) find evidence of positive indirect effects of airports for Italy. Results of Yu et al. (2013) for China are not conclusive since the sign and magnitude of the spatial spillovers due to transport investments differ by region.

In most cases, the analysis is made without any prior disaggregation by type of infrastructure. However, some exceptions are studies that compare the role of different transport modes simultaneously. For the case of Spain, Arbués et al. (2015) find that roads positively affect the output of the region in which the infrastructure is located and its neighboring provinces, while the direct effect of seaports is negative. Fageda and González-Aregall (2017) examine the direct, indirect, and total impacts of all transport modes on

industrial employment and find that only ports are able to generate positive total effects, and that motorways generate indirect negative effects. [Chen and Haynes \(2015\)](#) examine the regional impact of public transportation infrastructure (highways, public railways, public transit, and public airports) in the U.S. Northeast Megaregion. The result suggests that transport infrastructure (except public transit) have significant positive impacts on regional economic growth, most of which is from spillover effects.

The Impact of Transport Investments on Regional Convergence

Economic convergence at national or regional level refers to an inverse relationship between the growth rate of income per capita and the starting level of income per capita. Specifically, it is a situation where the gap in output per capita between regions (or countries) tends to decrease over time. Additionally, the literature distinguishes between absolute and conditional convergence. In the absolute convergence specification, the growth rate is explained only by the initial level of per capita income. The conditional convergence specification includes some other variables in order to control for attributes that may produce different steady-state growth rates among regions, including variables about the capital stock in transport infrastructures. In this literature, transport investments contribute to regional convergence if the negative relationship between the growth of income per capita and the starting level of income per capita is stronger when adding the transport infrastructure variables.

Studies that use samples of several countries generally find a positive effect of surface transportation on regional convergence. [Calderón and Chong \(2004\)](#) use country-level data to show that the endowment of roads and railways (in terms of both quantity and quality) are negatively linked with income inequality. [Del Bo and Florio \(2012\)](#) find a positive effect of motorways on regional convergence using data of regions within the European Union. [Lessmann and Seidel \(2017\)](#) use luminosity data to examine the determinants of regional inequality for a sample of countries from all over the world. They find that higher transport costs lead to more regional inequality in large countries.

In contrast, studies that use regional data within a country generally do not find evidence about a positive influence of transport infrastructure on regional convergence. [Rodríguez-Pose et al. \(2012\)](#) show that public investments in transport infrastructure have not contributed to the regional convergence in Greece. [Cosci and Mirra \(2018\)](#) show that motorway investments in Italy have contributed to the reduction of regional disparities in some period but the opening of the Autostrada del Sole has just contributed to the economic growth of the richer regions. Finally, [Fageda and Olivieri \(2019\)](#) measure the direct, indirect, and total impact of roads, railways, ports, and airports, and they find that only roads appear to have a positive impact on regional convergence.

Some works do not explicitly test for regional convergence, but their analyses have implications on the role of transportation in reducing regional disparities. [Costa-Font and Rodríguez-Oreggia \(2005\)](#) find that public investments have only been able to reduce regional inequalities among the richest regions in Mexico. [Pereira and Andraz \(2006\)](#) find that public investments in transportation in Portugal have contributed to a concentration of the activity in Lisbon. [Checherita \(2009\)](#) shows that the regional convergence process in the United States is not explained by the public capital stock in each state. Finally, several studies show the role of roads in promoting processes of suburbanization or decentralization of population and economic activity within an urban area ([Baum-Snow, 2007](#); [Baum-Snow et al., 2017, 2018](#); [García-López et al., 2015](#)) so that they provide relevant results linking transport investments and inequalities within an urban area.

Conclusions

Many studies have examined the role of transport investments on economic growth. However, some gaps in the current knowledge can be mentioned. In this regard, most of studies have focused on one specific country, particularly when the unit of observation is the region or an urban area. Furthermore, most of studies either consider aggregate transport investments or focus on one single mode. Hence, cross-country studies using data at the region level and studies that consider separately all transport modes may provide new insights and may be useful for policy-makers. Indeed, heterogeneous impacts of transport investments according to region attributes and the specific mode considered can be expected and no clear patterns can be inferred from the current literature.

Furthermore, not many studies consider spatial spillovers that are essential to fully understand the broader effects of transport investments. The omission of the spatial spillovers may lead to an overestimation of the positive impacts of transport investments and may hide some perverse effects of the investment, particularly when the investment is not intended to correct clear bottlenecks. A clear shortcoming of the spatial econometric techniques is the difficulties in addressing the simultaneous determination bias. Thus, studies that took into account both spatial interactions and endogeneity biases would represent a technological advance in methodological terms. In fact, evidence is not conclusive even in the sign of the spatial spillovers so that much more work is needed along these lines.

The literature on the contribution of transportation to regional convergence is scarce, and very few studies provide clues about the contribution of each specific mode to such regional convergence process. Hence, more studies about the link between transport investment and regional disparities that disaggregate by modes are needed. Moreover, it is not at all clear that investments in transport infrastructures really reduce regional disparities and this could be explained by the drivers of such investments (efficiency, equity, redistribution, tactical politics, etc.). Thus, studies that provided a bridge between the literature on the contribution of transportation to regional disparities and the literature about the drivers of such investments may be particularly helpful for policy-makers. A final remark is the need of more studies that examined the role of transport investments to reduce (or increase) inequalities within urban areas.

References

- Agrawal, A., Galasso, A., Oettl, A., 2017. Roads and innovation. *Rev. Econ. Stat.* 99, 417–434.
- Akyelken, N., 2015. Infrastructure development and employment, the case of Turkey. *Regional Stud.* 49, 1360–1373.
- Albalade, D., Bel, G., 2012. The Economics and Politics of High Speed Rail, Lessons from Experiences Abroad. Rowman and Littlefield Publishers (Lexington Books), Lanham.
- Albalade, D., Fageda, X., 2016. High-tech employment and transportation: evidence from the European regions. *Regional Stud.* 50, 1564–1578.
- Arbués, P., Baños, J., Mayor, M., 2015. The spatial productivity of transportation infrastructure. *Transport. Res. A* 75, 166–177.
- Aschauer, D., 1989. Is public expenditure productive? *J. Monet. Econ.* 23, 177–200.
- Banerjee, A., Duflo, E., Qian, N., 2020. On the road: access to transportation infrastructure and economic growth in China. *J. Dev. Econ.* 102442.
- Baum-Snow, N., 2007. Did highways cause suburbanization? *Q. J. Econ.* 122, 775–805.
- Baum-Snow, N., Brandt, L., Henderson, J., Turner, M., Zhang, Q., 2017. Roads, railroads, and decentralization of Chinese cities. *Rev. Econ. Stat.* 99, 435–448.
- Baum-Snow, N., Henderson, J.V., Turner, M.A., Zhang, Q., Brandt, L., 2018. Does investment in national highways help or hurt hinterland city growth? *J. Urban Econ.* 103124.
- Bel, G., Fageda, X., 2008. Getting there fast: globalization, intercontinental flights and location of headquarters. *J. Econ. Geogr.* 8, 471–495.
- Berechman, J., Ozmen, D., Ozbay, K., 2006. Empirical analysis of transportation investment and economic development at state, country and municipality levels. *Transportation* 33, 537–551.
- Bilotkach, V., 2015. Are airports engines of economic development? A dynamic panel data approach. *Urban Stud.* 52, 1577–1593.
- Blonigen, B., Cristea, A., 2015. Airports and urban growth, evidence from a quasi-natural policy experiment. *J. Urban Econ.* 86, 128–146.
- Bottasso, A., Conti, M., Ferrari, C., Merk, O., Tei, A., 2013. The impact of port throughput on local employment: evidence from a panel of European regions. *Transport Policy* 27, 32–38.
- Brueckner, J.K., 2003. Airline traffic and urban economic development. *Urban Stud.* 40, 1455–1469.
- Calderón, C., Chong, A., 2004. Volume and quality of infrastructure and the distribution of income: an empirical investigation. *Rev. Income Wealth* 50, 87–106.
- Checherita, C., 2009. Variations on economic convergence: the case of the United States. *Pap. Regional Sci.* 88, 259–279.
- Chen, Z., Haynes, K., 2015. Regional impact of public transportation infrastructure: a spatial panel assessment of the US Northeast megaregion. *Econ. Dev. Q.* 29, 275–291.
- Clark, D.E., Murphy, C.A., 1996. Countywide employment and population growth: an analysis of the 1980s. *J. Regional Sci.* 36 (2), 235–256.
- Cohen, J.P., 2010. The broader effects of transportation infrastructure: spatial econometrics and productivity approaches. *Transport. Res. E* 46, 317–326.
- Cosci, S., Mirra, L., 2018. A spatial analysis of growth and convergence in Italian provinces: the role of highways. *Regional Stud.* 52, 516–527.
- Costa-Font, J., Rodríguez-Oreggia, E., 2005. Is the impact of public investment neutral across the regional income distribution? Evidence from Mexico. *Econ. Geogr.* 81, 305–322.
- Crafts, N., 2009. Transport infrastructure investment: implications for growth and productivity. *Oxf. Rev. Econ. Policy* 25, 327–343.
- Crescenzi, R., Rodríguez-Pose, A., 2012. Infrastructure and regional growth in the European Union. *Pap. Regional Sci.* 91, 487–513.
- Del Bo, C., Florio, M., 2012. Infrastructure and growth in a spatial framework: evidence from the EU regions. *Eur. Plann. Stud.* 20, 1393–1414.
- Delgado, M., Álvarez, I., 2007. Network infrastructure spillover in private productive sectors: evidence from Spanish high capacity roads. *Appl. Econ.* 9, 1583–1597.
- Donaldson, D., 2018. Railroads of the Raj: estimating the impact of transportation infrastructure. *Am. Econ. Rev.* 108, 899–934.
- Durantón, G., 2016. Determinants of city growth in Colombia. *Pap. Regional Sci.* 95, 101–132.
- Durantón, G., Turner, M., 2012. Urban growth and transportation. *Rev. Econ. Stud.* 79, 1407–1440.
- Fageda, X., González-Aregall, M., 2017. Do all transport modes impact on industrial employment? Empirical evidence from the Spanish regions. *Transport Policy* 55, 70–78.
- Fageda, X., Olivieri, C., 2019. Transport infrastructure and regional convergence: a spatial panel data approach. *Pap. Regional Sci.* 98, 1609–1631.
- Farhadi, M., 2015. Transport infrastructure and long-run economic growth in OECD countries. *Transport. Res. A* 74, 73–90.
- Fan, S., Chan-Kang, C., 2008. Regional road development, rural and urban poverty: evidence from China. *Transport Policy* 15, 305–314.
- Fernald, J.G., 1999. Roads to property? Accessing the link between public capital and productivity. *Am. Econ. Rev.* 89, 619–638.
- Ferrari, C., Percoco, M., Tedeschi, A., 2010. Ports and local development: evidence from Italy. *Int. J. Transport Econ.* 37 (1), 9–30.
- García-Milà, T., McGuire, T., 1992. The contribution of publicly provided inputs to states' economies. *Regional Sci. Urban Econ.* 22, 229–241.
- García-López, M.A., Holl, A., Viladecans-Marsal, E., 2015. Suburbanization and highways in Spain when the romans and the bourbons still shape its cities. *J. Urban Econ.* 85, 52–67.
- Givoni, M., 2006. Development and impact of the modern high-speed train, a review. *Transport Rev.* 26, 593–611.
- Holl, A., 2004. Manufacturing location and impacts of road transport infrastructure: empirical evidence from Spain. *Regional Sci. Urban Econ.* 34 (3), 341–363.
- Holmgren, J., Merkel, A., 2017. Much ado about nothing?—a meta-analysis of the relationship between infrastructure and economic growth. *Res. Transport. Econ.* 63, 13–26.
- Holtz-Eakin, D., Schwartz, A., 1995. Infrastructure in a structural model of economic growth. *Regional Sci. Urban Econ.* 25 (2), 131–151.
- Jiwattanakulpaissarn, P., Noland, R.B., Graham, D.J., 2010. Causal linkages between highways and sector-level employment. *Transport. Res. A* 44, 265–280.
- Lakshmanan, T.R., 2011. The broader economic consequences of transport infrastructure investments. *J. Transport Geogr.* 19, 1–12.
- Lall, S.V., 2007. Infrastructure and regional growth, growth dynamics and policy relevance for India. *Annu. Regional Sci.* 41, 581–599.
- Lean, H.H., Huang, W., Hong, J., 2014. Logistics and economic development: evidence from China. *Transport Policy* 32, 96–104.
- Lessmann, C., Seidel, A., 2017. Regional inequality, convergence, and its determinants – a view from outer space. *Eur. Econ. Rev.* 92, 110–132.
- Lo Cascio, I., Mazzola, F., Epifanio, R., 2019. Territorial determinants and NUTS 3 regional performance: a spatial analysis for Italy across the crisis. *Pap. Regional Sci.* 98, 641–677.
- Matas, A., Raymond, J., Roig, J.L., 2013. Wages and accessibility, the impact of transport infrastructure. *Regional Stud.* 49, 1–19.
- Melo, P.C., Graham, D.J., Brage-Ardao, R., 2013. The productivity of transport infrastructure investment: a meta-analysis of empirical evidence. *Regional Sci. Urban Econ.* 43, 695–706.
- Möller, J., Zierer, M., 2018. Autobahns and jobs: a regional study using historical instrumental variables. *J. Urban Econ.* 103, 18–33.
- Moreno, R., López-Bazo, E., 2007. Returns to local and transport infrastructure under regional spillovers. *Int. Regional Sci. Rev.* 30, 47–71.
- Morrison, C.J., Schwartz, A.E., 1996. State infrastructure and productive performance. *Am. Econ. Rev.* 86, 1095–1111.
- Munnell, A., 1990. How does public infrastructure affect regional economic performance? *N. Engl. Econ. Rev.*, Federal Reserve Bank of Boston, September/October: 11–32.
- Nadiri, M.I., Mamuneas, T.P., 1994. The effects of public infrastructure and R&D capital on the cost structure and performance of US manufacturing industries. *Rev. Econ. Stat.* 76, 22–37.
- Park, J.S., Seo, Y.J., Ha, M.H., 2019. The role of maritime, land, and air transportation in economic growth: panel evidence from OECD and non-OECD countries. *Res. Transport. Econ.* 78, 100765.
- Park, J.S., Seo, Y.-J., 2016. The impact of seaports on the regional economies in South Korea: panel evidence from the augmented Solow model. *Transport. Res. E* 85, 107–119.
- Percoco, M., 2016. Highways, local economic structure and urban development. *J. Econ. Geogr.* 16, 1035–1054.
- Percoco, M., 2010. Airport activity and local development evidence from Italy. *Urban Stud.* 47, 2427–2443.
- Puga, D., 2002. European regional policies in light of recent location theories. *J. Econ. Geogr.* 2, 373–406.
- Pereira, A., Andraz, J., 2006. Public investment in transportation infrastructures and regional asymmetries in Portugal. *Ann. Regional Sci.* 40, 803–817.
- Redding, S.J., Turner, M.A., 2014. Transportation Costs and the Spatial Organization of Economic Activity, NBER Working paper, No. 20235, pp. 1–59.
- Rephann, T., Isserman, A., 1994. New highways as economic development tools: an evaluation using quasi-experimental matching methods. *Regional Sci. Urban Econ.* 24, 723–751.
- Rodríguez-Pose, A., Psycharis, Y., Tselios V., 2012. Public investment and regional growth and convergence: evidence from Greece. *Pap. Regional Sci.* 91, 543–568.
- Xu, H., Nakajima, K., 2017. Highways and industrial development in the peripheral regions of China. *Pap. Regional Sci.* 96, 325–356.
- Yamaguchi, K., 2007. Inter-regional air transport accessibility and macro-economic performance in Japan. *Transport. Res. E* 43, 247–258.
- Yu, N., De Jong, M., Storm, S., Mi, J., 2012. Transport infrastructure, spatial clusters and regional economic growth in China. *Transport Rev.* 32, 3–28.
- Yu, N., De Jong, M., Storm, S., Mi, J., 2013. Spatial spillovers effects of transport infrastructure: evidence from Chinese regions. *J. Transport Geogr.* 28, 56–66.

Transport Modes and an Aging Society

Charles B.A. Musselwhite*, Theresa Scott†, *Centre for Innovative Ageing, Swansea University, Swansea, United Kingdom; †University of Queensland, St. Lucia, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	6
Are Older People Safe as Drivers?	7
Age-Friendly Transport System	7
Active Travel	8
Walking	8
Cycling	8
Public and Community Transport	9
Public Buses	9
Train Travel	9
Community Transport	9
Taxis and Paratransit	10
Governance, Policy, and Societal Norms	10
Conclusion	10
See Also	10
References	10
Further Reading	12

Introduction

We live in an aging society. Across the globe, people are living longer and coupled with a lower birth rate the percentage of older people is increasing in number. Between 2015 and 2050, the proportion of the world's population over 60 years will nearly double from 12% to 22%. By 2020, the number of people aged 60 years and older will outnumber children under 5 years of age. At the moment, this is more pronounced in high-income countries but is also noticeable to a lesser but growing extent in many low- to middle-income countries. It is forecast that by 2050, 80% of older people will be living in low- and middle-income countries (LMIC).

Mobility is important to older people (Schlag et al., 1996) and a lack of mobility is linked to poorer health and well-being, and increases in depression and loneliness (Fonda et al., 2001; Ling and Mannion, 1995). Some mobility does far more good for health than others. Active travel (walking and cycling) is good for older people's physical health. Regular walking or cycling reduces the risk of coronary heart disease, stroke, cancer, obesity, and type 2 diabetes (NICE, 2013), overall reducing the risk of cardiovascular disease by around 30% and all-cause mortality by 20% (Hamer and Chida, 2008). Driving a car has less direct benefits due to it being a rather sedentary activity. Using public transport can have health benefits if there is walking involved in getting to and from bus stops, which is a protective factor in obesity for older people (Webb et al., 2011). Bus use is especially high among older people where there are concessionary (reduced) or free fares, as in the United Kingdom.

Much of the research in the area of older people and mobility has focused on western society. In such countries the motor vehicle dominates mobility. It is the most used form of transport among older people, of which the pattern of car use has been steadily growing over the last 30–40 years for those aged 65 and over. This seems set to continue with the 50–59 year olds being heavy users of cars, who are likely to take this pattern into old age. Why has there been such a growth in car use among older people? There are push factors—western society has become geared around the car, so that planning and urban design forces people to use vehicles for accessing basic healthcare, shops, and services by, for example, placing them on the edge of city or town centers on major trunk roads, highways, or motorways. In addition, the western world is championed by a hypermobile society where people are highly mobile throughout their life, living in different places, moving for jobs or recreation, moving between different communities, and away from family and friends. In later life, older people may want to stay connected to different communities and to family and friends and hence high levels of mobility are needed to achieve this. Older people are more likely to still be working now than in previous generations and hence need mobility to enable them to access workplaces. There are also factors that pull people to car use, the status symbol and desirability of the car, sold to them by manufacturers for years as a symbol of freedom and independence. The convenience and door-to-door element of using a car and that it is there ready to be used at the whim or need of the owner is hard to match in other modes. Consequently, giving up driving in later life is also associated with multiple negative outcomes. Research has shown that giving up driving is related to restricted community participation, reduced well-being, and an increased risk of depression, anxiety, and other health-related problems, including feelings of stress, loneliness and isolation, and also increased mortality (Edwards et al., 2009; Fonda et al., 2001; Ling and Mannion, 1995; Marottoli et al., 2000; Marottoli et al., 1997; Mezuk and

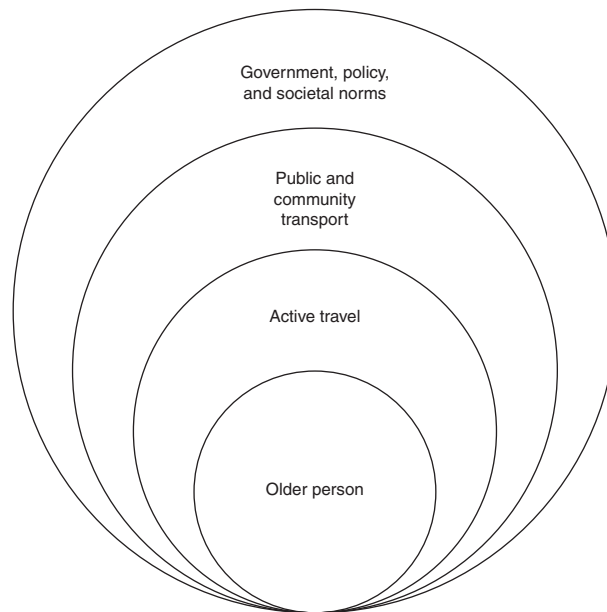


Figure 1 A model of age-friendly transport system (after [Musselwhite, 2016, 2018a](#)).

[Rebok, 2008](#); [Musselwhite and Haddad, 2010, 2018a, 2018b](#); [Musselwhite and Shergold, 2013](#); [Peel et al., 2002](#); [Ragland et al., 2005](#); [Windsor et al. 2007](#); [Ziegler and Schwanen, 2011](#)).

In LMIC, patterns of car ownership are not as straightforward. Although cars make up a growing amount of transport use in LMICs, in almost all countries they are a luxury that most cannot afford. The difference between the rich and the poor and car use is significant, and those from higher incomes are much more likely to own and drive a car than those from a less well-off background ([Monteiro et al., 2004](#)). This difference is not as pronounced in high-income countries.

Are Older People Safe as Drivers?

In the normal aging process, older people face many physiological challenges that may affect their ability to drive. These include cognitive changes resulting in increased difficulty in sustained and selective attention, cognitive overload, and slower processing ([Zanto and Gazzaley, 2014](#)). Medical conditions such as dementia can impact driving safety. However, a diagnosis of dementia is never a sufficient reason to immediately withdraw driving privileges because there are varied presentations and rates of progression. Changes in eyesight, in particular changes in recovery from glare and reduced vision in low light, can hamper driving ability ([Musselwhite, 2017a](#)). Between 15 and 65 years of age, not only does susceptibility to glare increase, the recovery time from glare increases from 2 to 9 s ([Holland, 2001](#)). Research suggests that by the age of 75 years old drivers may require 32 times the brightness than a 25-year old does in order to be able to see effectively. Finally, changes in muscle tone and ability to detect feedback affect mobility ([Musselwhite, 2017a](#)).

On the whole, older people are safe drivers. Statistics suggest there is an increase in collisions involving older people over the age of 75 in most countries. However, studies tend to attribute that to frailty, older people are more likely to be injured or die from collisions that younger people would walk away from and statistics do not include noninjury collisions. Older people do have issues with eyesight, fatigue, and detecting feedback that may result in different injuries. However, testing does not seem to make any difference to the collision rates, both for older drivers or for the population as a whole ([Mitchell, 2018](#); [Siren and Haustein, 2015](#); [Siren and Meng, 2012](#)).

Age-Friendly Transport System

To help the decision-making process and encourage older people out of the car, alternatives need to be improved. [Musselwhite \(2016, 2018a\)](#) offers a model of age-friendly transport, based on an ecological system, with the individual at the center and then covering walking and cycling, followed by public and community transport and surrounded by laws, rules, and strategy ([Fig. 1](#)). The model highlights that in an age-friendly transport system it is important to get the walking right before other forms of transport can be considered, since accessing the walking environment is needed to access public and community transport. Then overarching all of the modes is formalized policy and informal norms and rules that affect how services and infrastructure are provided.

Active Travel

Walking

Walking is the main mode among the aging population in the majority of LMIC, with long journeys often being the norm especially those from lower socioeconomic backgrounds (Prohaska et al., 2012). It is difficult to obtain comparable figures, but data suggest South Africa, Mexico, Thailand, and Russia have especially high levels of walking among their older people. In higher-income countries, population-wide walking is on the decrease including walking undertaken by older people. For example, in the United Kingdom, Germany, and France, walking has decreased by around 9% in the last 25 years (Musselwhite, 2018b). Walking makes up around 9% of all trips for over 65 in the United States. This is largely as a result of car-centric policies, reducing services and shops within walking distances and the growth of the motor vehicle as the default mode of transport. However, in some countries walking is more prevalent among older people. In the Netherlands, walking accounts for 15% of all trips made by people over 65, and in Denmark around 23% of all trips are by foot for those over 65 years of age (Pucher and Buehler, 2008, 2012). These figures have been stable for a number of years. Quality infrastructure, planning for walking and cycling, and cultural norms are generally seen as reasons for this (Pucher and Buehler, 2008, 2012).

It is relatively dangerous for older people to be pedestrians. In the United Kingdom, they make up 20% of the population and around 21% of overall distance traveled on foot, yet account for 41% of road traffic collisions. In Australia, 4 times the number of pedestrians aged 60 years and over were killed, compared to children under 16 years of age representing around 37% of all-aged fatalities. There are higher numbers in LMIC. Data from Britain (DfT, 2019), Australia (Job, 1998), France (Domnes et al., 2014), and the Netherlands (Dijkstra and Bos, 1997) suggest that older people are overrepresented in injuries resulting from incidents where failure to judge vehicle speed is a significant factor, where they look less at traffic and accept significantly smaller gaps in traffic when crossing the street. Looking at the barriers to walking for older people may help reveal where safety improvements can occur. Frailty is an issue here, but moreover the walking environment is not one that is pleasant for older people.

There are a number of known barriers to walking. Significantly in all countries, poorly maintained pavements, for example, that are frequently cracked or damaged can severely inhibit walking (Musselwhite, 2018b; Newton and Ormerod, 2012) and there are difficulties walking and balancing on tactile pavements (Newton and Ormerod, 2012). Poor surfaces, caused by fallen leaves, rain, ice, or snow, for example, can also be a barrier when not cleared (Wennberg, 2009). Older people also need continuous dedicated space for walking, free from clutter and with minimal sharing with other modes, especially cycling (Musselwhite, 2018b). There should be safe and accessible public conveniences and seating (Newton et al., 2010). To enable walking, older people need a dedicated pavement (or sidewalk) that is often absent in rural areas of high-income countries and very likely to be absent in LMIC (Srichuae et al., 2016).

Difficulty in crossing the road is often cited as an issue for older people. Dedicated crossing facilities must be sited to give pedestrians fixed or temporary right of way across roads, especially in areas where older people need them, along their “desire lines.” Again these are less likely in LMIC, leaving crossing dangerous fast roads at the hands of older people. As people age, their walking speed slows up and so ability to cross roads becomes harder. Dedicated time to cross the road should be allowed for older people. Previous research suggests older people cross the road at between 0.7 and 0.9 meters per second (Asher et al., 2012; Newton and Ormerod, 2008), yet crossing times are set for a walking speed of 1.22 meters per second (m/s) in most countries, and this is too fast for older people. Musselwhite (2015) found that 88% of older people could not complete the crossing in this time. In addition, changes in gait as people age means stepping down from the kerb is harder to negotiate. Hence, crossing places need to have minimal drop or better still be at grade with the road. Older people tend to spend more time looking at their step and less time looking at the traffic, compared to younger people, adding to the potential for dangerous crossing maneuvers or a slower crossing pattern that pedestrian crossings must take into account (Avineri et al., 2012).

Dark or poorly lit streets have a negative effect on older people using the street, both in terms of concerns about falling and because of safety fears (Shumway-Cook et al., 2003). Smog is a deterrent for older people to go out walking, especially in LMIC, but also cities in higher-income countries. Cities in India are particularly a no-go area for older people at certain times of high pollution. Wealthier nations can also suffer similar problems; older people can be at risk in Japanese cities and at times European cities (Deguen and Zmirou-Navier, 2010).

Cycling

Typically, in high-income countries, there is a big decrease in cycling for older people. For example, distance cycled per person per year in England was 15 miles for those over the age of 70 compared to 52 miles for 60–69 year olds and 53 miles across all age groups (Musselwhite, 2018a). Although the number of miles cycled per person per year has increased over the last 13 years for 60–69 year olds, this has not occurred for those aged 70 or over years of age (DfT, 2015). However, the amount of cycling in the United Kingdom among the older (and indeed the younger) population is low compared to other northern European countries. For example, cycling accounts for 23% of trips made by older people in the Netherlands, 15% in Denmark, and 9% in Germany, against less than 1% in the United Kingdom (Pucher & Buehler, 2012; DfT, 2019). In LMIC, cycling is much higher. Around 56% of older Chinese men cycle daily (Lee et al., 2012), with similar numbers in Uganda (Hill and Heesch, 2018).

Barriers to cycling include built environment and personal and cultural factors (Musselwhite, 2018b). One of the major factors putting older people off from cycling is road danger, sharing the road with fast moving, and heavy traffic (Pooley et al., 2013). In Denmark and the Netherlands, where there are higher levels of cycling among older people, dedicated cycling infrastructure is typical, allowing people to cycle off road. Often these spaces are for cyclists only; older people also feel vulnerable

sharing spaces where pedestrians might be present and find shared use paths for walking and cycling difficult to navigate (Jones et al., 2016).

Older people are often put off cycling feeling that they are not fit enough (Davies et al., 1997). Again this is something less likely to be felt in countries where cycling across the life course is the norm and people have continued to cycle throughout their life (Jones et al., 2016).

Public and Community Transport

Public Buses

A major barrier to using public transport is the cost. Where it is provided at reduced rate for older people or indeed for free to older people, an increase in use and in satisfaction is found. In the United Kingdom, around 60% of older people reported an increase in their bus use due to free travel being introduced for older people (Baker and White 2010; Mackett, 2013a, 2013b). In addition, around 74% of respondents stated having a pass enabled them to visit new places and to engage in new leisure pursuits. Andrews (2011) found that 74% of his respondents thought their free bus pass improved their quality of life. Dargay et al. (2010) modeled bus use against what would have happened if no free bus pass had been introduced and suggests the number of bus stages (groups of bus stops) traveled through by older people increased by 45.4% in rural areas and 26.5% in urban areas. While becoming eligible for a free bus pass is associated with increased use of public transport (Webb et al., 2011), this has not reduced the amount of driving among the older population, though it is worth noting that driving may have increased even more without free bus travel.

The driver of the bus can make or break journeys for older people. An awareness of older people's issues is needed. A major concern noted in studies across the world was that the driver drove away from the bus stop before the older person had sat down, or it was missing the older person's stop and the older person then having to walk much further than anticipated (Broome et al., 2010; Musselwhite, 2018c; Transport for London, 2009, 2013). Age UK London (2011) surveyed bus driving behavior and, across 550 journeys, found that in 42% of cases passengers were not given enough time to sit down before the bus was driven away from the stop and in 25% of the cases the bus did not pull up tight to the kerb at the bus stop. Gilhooly et al. (2002) note a major barrier being personal safety and security, especially in the evening and at night. Again this is an issue found globally. Buses in high-income countries are usually bound by regulation to make them accessible, with lowering floors to ease entry and exit and grab rails to help maneuver inside the bus, along with standards of information provision and lighting, all which help older people. However, these standards are not the case in many other countries, making the use of public transport dangerous for older people, or excluding them altogether (Broome et al., 2010).

Bus services often tend to be poorer in rural communities, with few buses or in some cases no buses at all. Even in urban areas buses often tend to follow radial routes into the city center, which favors workers' patterns of travel. Older people are just as likely to travel between intraurban across cities to access the services and shops they want to (Musselwhite and Curl, 2018).

Train Travel

There is not much research on train travel among older people. Use of railways tends to decline in later life, mostly because the necessity to use railways for commuting or work declines. However, rail travel could provide an excellent way to travel long distances for older people.

Previous research has also highlighted issues with overcrowding on stations and platforms as a particular issue for older people (Gilhooly et al., 2002). In addition, ticketing can be confusing, leading to anxiety, or older people failing to get the cheapest or even the correct ticket (Musselwhite, 2011). Stations themselves can also appear unsafe, especially if they do not have staff present at them (Cozens et al., 2004).

In the study of one train operating company in the United Kingdom, Musselwhite (2019) found older people were overrepresented in passenger accidents across the railway, and were especially overrepresented compared to other age groups in accidents boarding, slips, trips and falls on trains and slips, trips and falls at the station frontage, car park, and concourse. Audits reveal the need to see the railway station to train service as a continuous entity with better signage, lighting, and places to sit and rest throughout the station from frontage, to concourse, to platform, and to boarding the train. The areas from arriving at the station from other transport or in the car to the platform itself were often neglected spaces. To reduce accidents boarding more level access between platform and train is needed at standard.

Community Transport

As an alternative to conventional public bus services, in some countries there are specialist transport services, sometimes also known as community transport or transit, which operate more or less door to door for people who cannot access public or private transport. Services are often provided by a charity or a third sector organization, sometimes under license from local authorities. Such services can be the difference between staying in and going out (ECT, 2016). There are direct improvements on people's health through affording greater access to GP and hospital services and reductions in missed appointments, improvements in diagnosis, and therefore lower healthcare costs (ECT, 2016). It can also mean people are discharged earlier as they have access between hospital appointments and home (ECT, 2016). Importantly, drivers can act as informal carers and can help identify early warning signs of illness or of loneliness and isolation, as well as offering social support to the passengers (ECT, 2016).

However, there are some barriers to community transport use (Musselwhite, 2018c). Because services are very dependent on third sector and charity provision they can be somewhat fragmented across the country. They might be short-lived often existing around a particular one-off or small scale-funding offering and around key individuals. People who may well benefit from such a service can sometimes feel the service is not for people like them; there is sometimes the perception that it is for people with disabilities, rather than for everyone with accessibility issues (Musselwhite, 2018c). Frequently, there is a lack of information and, as a result, much misunderstanding of the service (Parkhurst et al., 2014; Ward et al., 2013). Journeys typically are based around providing transport to shops, services, and doctors and hospitals, but there needs to be more “discretionary” journeys provided to places of leisure and fun (Musselwhite, 2017b).

Taxis and Paratransit

To help people in many LMIC, motorcycle taxi services are found to transport people around, especially into congested and busy city centers. In some countries, this service is still too expensive for the majority of people to use. For example, in Tanzania many people cannot afford to use the motorcycle service as discussed by Porter et al. (2018). People in one district (Kilolo) would leave their village shortly around midnight to walk 5 hours to the bus stop to get the one bus per day into town. In another district (in Meru district) older people could not physically get onto a motorcycle or on to the back of a pickup to go to town (Porter et al., 2018).

In high-income countries, the uses of taxis are often seen as too expensive, so they are avoided or rationed and so only partly fulfill transport needs. There are concerns over the safety of modern paratransit services such as Uber and Lyft, with concerns over data being shared, safety of the ride itself, and regulation of drivers.

Governance, Policy, and Societal Norms

There is a tendency for older people to be ignored in transport policy and governance. This is because transport provision is driven largely by economic interests in most countries which results in provision around the working adult and inter- and intraurban mobility with a focus on provision in core work hours 9–5 (Musselwhite, 2018a). Older people are less likely to fit this pattern of work and where they work, they often work part time and avoid rush hour travel if they can. Even though there is scope for greater impact from policy if directed around life stages (e.g., retirement from work, becoming a grandparent, becoming a carer), age is very rarely mentioned in policy contexts (Avineri and Goodwin, 2010).

Conclusion

There is no quick fix solution to producing age-friendly transport solutions and making alternatives to the car more attractive. At the very top level, there is a need to recognize needs and issues older people have that are unique and have strategies, and plans in place to support the difficulty that older people have with transport. Multiple stakeholders often control public space and have different interests. There is a need to recognize that older people's mobility may fit different patterns than those that transport policies are usually formed around. They are less likely to be working, if they are more likely to be part time or have more flexible hours and they are less likely to be moving in interurban spaces. A comprehensive strategy bringing together the different stakeholder needs with a focus on aging is required. Overall, walking and cycling for older people must be given greater priority among decision-makers. Provision of services for older people needs to take these elements into account. Crucial to public and community transport use is a mix of people and infrastructure-based support. At the active travel level, infrastructure is crucial to support walking and cycling.

See Also

Mobility Planning/Policies For Older People

References

- Age UK London, 2011. On the Buses: Older and Disabled People's Experiences on London Buses. Age UK, London.
- Andrews, G., 2011. Just the Ticket? Exploring the Contribution of Free Bus Fares Policy to Quality of Later Life. A thesis submitted in partial fulfilment of the requirements of the University of the West of England, Bristol, for the degree of Doctor of Philosophy. Bristol.
- Asher, L., Aresu, M., Falaschetti, E.A., Mindell, J., 2012. Most older pedestrians are unable to cross the road in time: a cross-sectional study. *Age Age*. 41, 690–694.
- Avineri, E., Goodwin, P., 2010. Individual behaviour change: evidence in transport and public health. Project Report. Department for Transport, London.
- Avineri, E., Shinar, D., Susilo, Y., 2012. Pedestrians' behaviour in cross walks: the effects of fear of falling and age. *Acc. Anal. Prev.* 44 (1), 30–34.
- Baker, S., White, P., 2010. Impacts of concessionary travel: case study of an English rural region. *Transp. Policy* 17 (1), 20–26.
- Broome, K., Nalder, E., Worrall, L., Boldy, D., 2010. Age-friendly buses? A comparison of reported barriers and facilitators to bus use for younger and older adults. *Australas. J. Ageing* 29 (1), 33–38.
- Cozens, P.M., Neale, R., Whitaker, J., Hillier, D., 2004. Tackling crime and fear of crime while waiting at Britain's railway stations. *J. Public Transp.* 7 (3), 23–41.
- Dargay, J., Ronghui, L., Toner, J., 2010. Concessionary bus fares in England: the impact on bus travel by the over-60s. Presented at European Transport Conference, Glasgow, Scotland.
- Davies, D.G., Halliday, M.E., Mayes, M., Pocock, R.L., 1997. Attitudes to Cycling: A Qualitative Study and Conceptual Framework. Transport Research Laboratory, Crowthorne.
- Deguen, S., Zmirou-Navier, D., 2010. Social inequalities resulting from health risks related to ambient air quality—a European review. *Eur. J. Public Health* 20, 27–35.

- DfT (Department for Transport), 2015. National Travel Survey. Department for Transport, London, UK.
- DfT (Department for Transport), 2019. National Travel Survey. Available from: <https://www.gov.uk/government/collections/national-travel-survey-statistics>.
- Dijkstra, A., Bos, J.M.J., 1997. ACEA—Dutch Contribution: Road Safety Effects of Small Infrastructural Measures With Emphasis on Pedestrians. SWOV, Leidschendam, Netherlands.
- Domnes, A., Cavallo, V., Dubuisson, J.B., Tournier, I., Vienne, F., 2014. Crossing a two-way street: comparison of young and old pedestrians. *J. Safety Res.* 50, 27–34.
- ECT (Ealing Community Transport), 2016. Why community transport matters? ECT Transport Report. ECT, Greenford.
- Edwards, J.D., Perkins, M., Ross, L.A., Reynolds, S.L., 2009. Driving status and three year mortality among community-dwelling older adults. *J. Gerontol. A Biol. Sci. Med. Sci.* 64, 300–305.
- Fonda, S.J., Wallace, R.B., Herzog, A.R., 2001. Changes in driving patterns and worsening depressive symptoms among older adults. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 56 (6), 343–351.
- Gilhooly, M.L.M., Hamilton, K., O'Neill, M., Gow, J., Webster, N., Pike, F., Bainbridge, C., 2002. Transport and ageing: extending quality of life via public and private transport. ESCR Report L48025025. Brunel University Research Archive. London, UK.
- Hamer, M., Chida, Y., 2008. Walking and primary prevention: a meta-analysis of prospective cohort studies. *Br. J. Sports Med.* 42 (4), 238–243.
- Hill, R.L., Heesch, K., 2018. The problem of physical inactivity worldwide among older people. In: Nyman, S.R., Barker, A., Haines, T., Horton, K., Musselwhite, C., Peeters, G. (Eds.), *The Palgrave Handbook of Ageing and Physical Activity Promotion*. Palgrave Macmillan, Cham, Switzerland, pp. 25–41.
- Holland, C.D., 2001. Older Drivers: A Review. Department for Transport, London.
- Job, R.F., 1998. Pedestrians at traffic light controlled intersections: crossing behaviour of the elderly and non-elderly. *Proceedings of the Conference on Pedestrian Safety*, Melbourne, Victoria, Australia, 29–30 June, pp. 40–44.
- Jones, T., Chatterjee, K., Spinney, J., Street, E., Van Reekum, C., Spencer, B., Jones, H., Leyland, L.A., Mann, C., Williams, S., Beale, N., 2016. cycle BOOM: Design for Lifelong Health and Wellbeing. Summary of Key Findings and Recommendations. Oxford Brookes University, UK.
- Lee, I.-M., Shiroma, E.J., Lobelo, F., et al., 2012. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *Lancet* 380, 219–229.
- Ling, D.J., Mannion, R., 1995. Enhanced mobility and quality of life of older people: assessment of economic and social benefits of dial-a-ride services. *Proceedings of the Seventh International Conference on Transport and Mobility for Older and Disabled People* (Vol. 1). DETR, London.
- Mackett, R., 2013. The impact of concessionary bus travel on the wellbeing of older and disabled people. Paper presented at the Transportation Research Board 92nd Annual Meeting, Washington, DC, 13–17 January (Transp. Res. Rec., 2352, 114–119).
- Mackett, R., 2013. The benefits of a policy of free bus travel for older people. *Proceedings of 13th World Conference on Transport Research*, Rio de Janeiro, 15–18 July.
- Marottoli, R.A., Mendes de Leon, C.F., Glass, T.A., Williams, C.S., Cooney, L.M., Berkman, L.F., 2000. Consequences of driving cessation: decreased out-of-home activity levels. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 55B (6), 334–340.
- Marottoli, R., Mendes de Leon, C., Glass, T., Williams, C., Cooney, L., Berkman, L., Tinetti, M., 1997. Driving cessation and increased depressive symptoms: prospective evidence from the New Haven EPESE. *J. Am. Geriatr. Soc.* 45 (2), 202–206, doi:10.1111/j.1532-5415.1997.tb04508.
- Mezuk, B., Rebok, G.W., 2008. Social integration and social support among older adults following driving cessation. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 63B, 298–303.
- Mitchell, C.G.B., 2018. Are older people safe drivers on the roads: testing and training? In: Musselwhite, C. (Ed.), *Transport, Travel and Later Life*. Emerald Publishing Limited, Bingley, pp. 37–63.
- Monteiro, C.A., Moura, E.C., Conde, W.L., Popkin, B.M., 2004. Socioeconomic status and obesity in adult populations of developing countries: a review. *Bull. World Health Organ.* 82, 940–946.
- Musselwhite, C., 2011. The importance of driving for older people and how the pain of driving cessation can be reduced. *J. Dementia Mental Health* 15 (3), 22–26.
- Musselwhite, C.B.A., 2015. Environment-person interactions enabling walking in later life. *Transp. Plan. Technol.* 38 (1), 44–61.
- Musselwhite, C.B.A., 2016. Vision for an age friendly transport system in Wales. *EnvisAGE, Age Cymru* 11, 14–23.
- Musselwhite, C.B.A., 2017a. Assessment of computer-based training packages to improve the safety of older people's driver behaviour. *Transp. Plan. Technol.* 40 (1), 64–79.
- Musselwhite, C.B.A., 2017b. Exploring the importance of discretionary mobility in later life. *Work. Older People* 21 (1), 49–58.
- Musselwhite, C.B.A., 2018a. Community connections and independence in later life. In: Peel, E., Holland, C., Murray, M. (Eds.), *Psychologies of Ageing: Theory, Research and Practice*. Palgrave Macmillan, Cham, Switzerland, pp. 221–252.
- Musselwhite, C.B.A., 2018b. Transportation and promoting physical activity among older people. In: Nyman, S.R., Barker, A., Haines, T., Horton, K., Musselwhite, C., Peters, G., Victor, C., Wolfe, J.K. (Eds.), *The Handbook of Ageing and Physical Activity Promotion*. Palgrave Macmillan, London, UK, pp. 507–526.
- Musselwhite, C.B.A., 2018c. Public and community transport. In: Musselwhite, C. (Ed.), *Transport, Travel and Later Life* (Transport and Sustainability, Vol. 10). Emerald Publishing Limited, Bradford, UK, pp.117–128.
- Musselwhite, C.B.A., 2019. Older people's mobility, new transport technologies and user-centred innovation. In: Müller, B., Meyer, G. (Eds.), *Towards User-Centric Transport in Europe—Challenges, Solutions and Collaborations*. Lecture Notes Series. Springer, Cham, Switzerland, pp. 87–103.
- Musselwhite, C.B.A., Curl, A., 2018. Geographical perspectives on transport and ageing. In: Curl, A., Musselwhite, C. (Eds.), *Geographies of Transport and Ageing*. Palgrave Macmillan, Cham, Switzerland, pp. 3–24.
- Musselwhite, C., Haddad, H., 2010. Mobility, accessibility and quality of later life. *Qual. Ageing Older Adults* 11 (1), 25–37.
- Musselwhite, C.B.A., Haddad, H., 2018a. Older people's travel and mobility needs: a reflection of a hierarchical model 10 years on. *Qual. Ageing Older Adults* 19 (2), 87–105.
- Musselwhite, C.B.A., Haddad, H., 2018b. The travel needs of older people and what happens when people give-up driving. In: Musselwhite, C. (Ed.), *Transport, Travel and Later Life* (Transport and Sustainability, Vol. 10). Emerald Publishing Limited, Bingley, UK, pp. 93–115.
- Musselwhite, C., Shergold, I., 2013. Examining the process of driving cessation in later life. *Eur. J. Ageing* 10 (2), 89–100.
- NICE (National Institute of Clinical Excellence), 2013. Walking and cycling: local measures to promote walking and cycling as forms of travel or recreation. NICE Public Health Guidance 41. NICE, London, UK.
- Newton, R., Ormerod, M., 2008. The design of streets with older people in mind: design guide—materials of footways and footpaths. IDGO Inclusive Design for Getting Outdoors. Sheffield, UK.
- Newton, R.A., Ormerod, M.G., 2012. The design of streets with older people in mind: tactile paving. IDGO. Inclusive Design for Getting Outdoors. Available from: www.idgo.ac.uk/design_guidance/streets.htm.
- Newton, R.A., Ormerod, M.G., Burton, E., Mitchell, L., Ward-Thompson, C., 2010. Increasing independence for older people through good street design. *J. Integr. Care* 18 (3), 24–29.
- Parkhurst, G., Galvin, K., Musselwhite, C., Phillips, J., Shergold, I., Todres, L., 2014. Beyond transport: understanding the role of mobilities in connecting rural elders in civic society. In: Hennessey, C., Means, R., Burholt, V. (Eds.), *Countryside Connections: Older People, Community and Place in Rural Britain*. Policy Press, Bristol, UK, pp. 125–175.
- Peel, N., Westmoreland, J., Steinberg, M., 2002. Transport safety for older people: a study of their experiences, perceptions and management needs. *Inj. Control Saf. Promot.* 9, 19–24.
- Pooley, C., Jones, T., Tight, M., Horton, D., Scheldeman, G., Mullen, C., Jopson, A., Strano, E., 2013. Promoting Walking and Cycling: New Perspectives on Sustainable Travel. Policy Press, Bristol, UK.
- Porter, G., Tewodros, A., Gorman, M., 2018. Mobility, transport and older people's well being in sub-Saharan Africa: review and prospect. In: Curl, A., Musselwhite, C. (Eds.), *Geographies of Transport and Ageing*. Palgrave Macmillan, pp. 75–100.
- Prohaska, T., Anderson, L., Binstock, R. (Eds.), 2012. *Public Health for an Aging Society*. The Johns Hopkins University Press, Baltimore, MD.
- Pucher, J., Buehler, R., 2008. Making cycling irresistible: lessons from The Netherlands, Denmark and Germany. *Transp. Rev.* 28 (4), 495–528.
- Pucher, J., Buehler, R., 2012. *City Cycling*. MIT Press, Cambridge, MA.
- Ragland, D.R., Satariano, W.A., MacLeod, K.E., 2005. Driving cessation and increased depressive symptoms. *J. Gerontol. A Biol. Sci. Med. Sci.* 60, 399–403.

- Schlag, B., Schwenkhausen, U., Trankle, U., 1996. Transportation for the elderly: towards a user friendly combination of private and public transport. *IATSS Res.* 20 (1), 75–82.
- Shumway-Cook, A., Patla, A., Stewart, A., Ferrucci, L., Ciol, M.A., Guralnik, J.M., 2003. Environmental components of mobility disability in community-living older persons. *J. Am. Geriatr. Soc.* 51, 393–398.
- Siren, A., Haustein, S., 2015. Driving licences and medical screening in old age: review of literature and European licensing policies. *J. Transp. Health* 2, 68–78.
- Siren, U., Meng, A., 2012. Cognitive screening of older drivers does not produce safety benefits. *Acc. Anal. Prev.* 45, 634–638.
- Srichuae, S., Vilas, N., Ranjith P., 2016. Aging society in Bangkok and the factors affecting mobility of elderly in urban public spaces and transportation facilities. *IATSS Res.* 40, 26–34.
- TfL (Transport for London), 2009. Older people's experience of travelling in London. Transport for London, June 2009. Available from: <http://www.tfl.gov.uk/cdn/static/cms/documents/older-peoples-transport-experiences-report.pdf>.
- TfL (Transport for London), 2013. TfL and accessibility charities launch new awareness training for bus drivers. Available from: <http://www.tfl.gov.uk/info-for/media/press-releases/2013/october/tfl-and-accessibility-charities-launch-new-awareness-training-for-bus-drivers>.
- Ward, M., Somerville, P., Bosworth, G., 2013. 'Now without my car I don't know what I'd do': the transportation needs of older people in rural Lincolnshire. *Local Econ.* 28 (6), 553–566.
- Webb, E., Netuveli, G., Millett, C., 2011. Free bus passes, use of public transport and obesity among older people in England. *J. Epidemiol. Community Health* 66 (2), 176–180.
- Wennberg, H., 2009. Walking in old age: a year-round perspective on accessibility in the outdoor environment and effects of measures taken. Doctoral thesis. Institutionen för Teknik och samhälle, Trafik och väg.
- Windsor, T.D., Anstey, K.J., Butterworth, P., Luszcz, M.A., Andrews, G.R., 2007. The role of perceived control in explaining depressive symptoms associated with driving cessation in a longitudinal study. *Gerontologist* 47, 215–223.
- Zanto, T.P., Gazzaley, A., 2014. Attention and ageing. In: Nobre, A.C., Kastner, S. (Eds.), *The Oxford Handbook of Attention*. Oxford University, Oxford, UK, pp. 927–971.
- Ziegler, F., Schwanen, T., 2011. I like to go out to be energised by different people: an exploratory analysis of mobility and wellbeing in later life. *Ageing Soc.* 31 (5), 758–781.

Further Reading

- Broome, K., Worrall, L., Fleming, J., Boldy, D., 2013. Evaluation of age-friendly guidelines for public buses. *Transp. Res. Part A Policy Pract.* 53, 68–80.
- Burton, E., Mitchell, L., 2006. *Inclusive Urban Design: Streets for Life*. Routledge, London.
- Curl, A., Musselwhite, C. (Eds.), 2018. *Geographies of Transport and Ageing*. Palgrave, London, UK.
- Eby, M., Louis, R.S., 2018. *Perspectives and Strategies for Promoting Safe Transportation Among Older Adults*. Elsevier, Amsterdam, Netherlands.
- Green, J., Jones, A., Roberts, H., 2014. More than A to B: the role of free bus travel for the mobility and wellbeing of older citizens in London. *Ageing Soc.* 34 (3), 472–494.
- Nyman, S., Barker, A., Horton, K., Musselwhite, C., Peeters, G., Victor, C., Wolff, J. (Eds.), 2018. *The Palgrave Handbook of Ageing and Physical Activity Promotion*. Palgrave Macmillan, London, UK.
- Musselwhite, C. (Ed.), 2018. *Transport, Travel and Later Life*. Emerald, Bingley, UK.

Sustainable Mobility Paths

Erling Holden^{*,†}, Geoffrey Gilpin[†], ^{*}Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Aas, Norway; [†]Department of Environmental Sciences, Western Norway University of Applied Sciences, Sogndal, Norway

© 2021 Elsevier Ltd. All rights reserved.

Introduction	13
An Unsustainable Transport System	14
A Typology for Achieving Sustainable Mobility	15
The Main Strategies	15
The Main Agents	16
The Sustainable Mobility Paths	16
Concluding Remarks	17
See Also	18
Biography	18
Further Reading	18

Introduction

In 1992, the European Union (EU) published a *Green Paper on the Impact of Transport on the Environment*.

'This Green Paper provid[ed] an assessment of the overall impact of transport on the environment and presents a Common strategy for 'sustainable mobility' which should enable transport to fulfil its economic and social role while containing its harmful effects on the environment.'

This paper responded to the challenges raised a few years earlier by the United Nations (UN) report *Our Common Future*. The *Green Paper* took aim at the rapidly increasing environmental impacts from the transport sector, and was a clear statement of transport policy in the EU. The *Green Paper* introduced the concept of 'sustainable mobility' to the international agenda—a bold but much needed appearance.

Bold in its acknowledgement of the intertwined and exceedingly complex relationship between transport's positive effects and its negative social and environmental impacts. Indeed, over the succeeding years, it has proven difficult to find a comfortable match between transport and sustainable development. It was *necessary* as it was (and still is) widely acknowledged that the increasingly negative impacts of transport were beginning to exceed the overall positive ones, that is, business-as-usual was unacceptable.

In 2015, the UN presented the 2030 Agenda for Sustainable Development comprising 17 Sustainable Development Goals (SDGs) and 169 targets—almost three decades after *Our Common Future* and the *Green Paper*. Again, this was another *bold* step. However, it is both striking and alarming that transport and its impacts are almost absent in the *2030 Agenda*. Admittedly, the *2030 Agenda* states that "we will adopt policies which increase [...] sustainable transport systems . . .," but is this helpful? SDG target 11.2 is somewhat more specific stating: "By 2030, provide access to safe, affordable, accessible and sustainable transport systems for all, improving road safety, notably by expanding public transport, with special attention to the needs of those in vulnerable situations, women, children, persons with disabilities and older persons." In summary, the SDG targets are diffuse and deal with transport at a very general level—and one strains to draw any clear conclusions as to how progress will be made towards the achievement of target 11.2.

This chapter is organized as follows; in the second section, the currently unsustainable status of the transport system will be presented. The third section outlines some of the dominant paths towards sustainable mobility. In the fourth section, we present some thoughts on further work.

Before we proceed, three clarifications are necessary. First, the terms "sustainable transport" and "sustainable mobility" can be found in the literature, and are often used synonymously. In North America "Sustainable transport" seems to be preferred, and in Europe "sustainable mobility." These terms embody both revealed- (actual- physical displacement) and potential (attribute of being mobile) mobility—in addition to entailing the same ideas and policy implications. As such, we acknowledge that- and use these terms interchangeably.

The second clarification emphasizes that this chapter focuses on *passenger* mobility in *developed* countries, since it is here that most of the resources are consumed, and pollution emitted. However, we acknowledge the importance of achieving sustainable- goods transport and mobility in developing countries. Additionally, the conclusions reached here are transferable to the freight sector, and with suitable adaptations for sustainable mobility in developing countries.

The third and final clarification states that this chapter focuses on the *environmental* imperatives associated with sustainable mobility, as these are the developed countries' main challenges. However, the equally important imperatives concerning needs and equal access to transport should not be forgotten, and the interpretations presented here should be relevant for the concerns of developing countries.

The number of books, articles and reports about sustainable mobility has increased steadily since 1992. This reflects the increasing unsustainability of the transport sector. A large and rich selection of literature now exists covering this field—and since projections forecast that the unsustainability of transport will continue to grow, the number of books, articles and reports is likely to increase too.

From a review of the academic literature, one can acknowledge that the mainstream understanding and interpretation of sustainable mobility has evolved since 1992, and can be grouped into four generations of studies. The first generation of studies spans the era 1992–93, and were both techno-centric and focused on how to limit transport’s negative environmental impacts by improving then-existing technology. The second, third, and fourth generations of studies span the eras 1993–2000, 2000–10, and 2010 present, respectively. These last three generations of studies, increasingly acknowledge the limitations of preceding efforts toward achieving sustainable mobility, and open for an increasingly greater diversity of alternatives. These generations of studies have become gradually more interdisciplinary in nature—reflecting the inter-relatedness of mobility with all other aspects of society.

An Unsustainable Transport System

Since the human migrations out of Africa millions of years ago, travel has been part of collective experience. Our motivation for traveling has varied, though for the most part, we have traveled to improve our lives. Whatever ones motivation, early travel was uncomfortable, dangerous and enormously time-consuming—today, some might argue that little has changed. However, there are two indisputable differences between past and present travel. First, in modern times we have experienced an extraordinary growth in mobility (measured in trips made and distance traveled), second, is the staggering increase in mobility’s associated environmental and social consequences.

Population growth in the 20th century was remarkable—with the world’s population increasing by a factor of 4. Growth in mobility was remarkable too—with motorized passenger kilometers and ton-kilometers by all modes growing by a factor of approximately 100 (e.g., [Fig. 1](#)). However, whereas population growth shows signs of leveling-off, growth in mobility does not—particularly, growth in mobility has been extensive the past half century. For example, in 1952, the average Norwegian traveled 5.5 km/day by motorized vehicle. In 2000, Norwegians’ traveled 48.7 km daily—and is every increasing into this century, with similar patterns visible in the rest of the developed world. Furthermore, comparisons of global scale—baseline (business-as-usual) scenarios foresee that this trend, particularly of increased travel by road-freight and air are likely to increase in the future.

While transport and mobility are acknowledged as important elements towards economic growth and accessibility, the negative social and environmental impacts of increased motorized mobility—particularly road and air travel—have been broadly acknowledged:

- Transport is a major consumer of energy and material resources, with around 31.6% of the world’s final energy consumption used for transport, mostly from nonrenewable energy resources.
- The production of motor vehicles requires large amounts of materials, for example, ferrous and nonferrous metals. Currently, in Organization for Economic Co-operation and Development (OECD) and non-OECD countries, motor vehicle production consumes 7% and 3% of ferrous metals respectively, this is similar for non ferrous. With demand for metals in both regions expected to grow by a factor of 2.2 and 3.5, respectively, between 2017 and 2060.

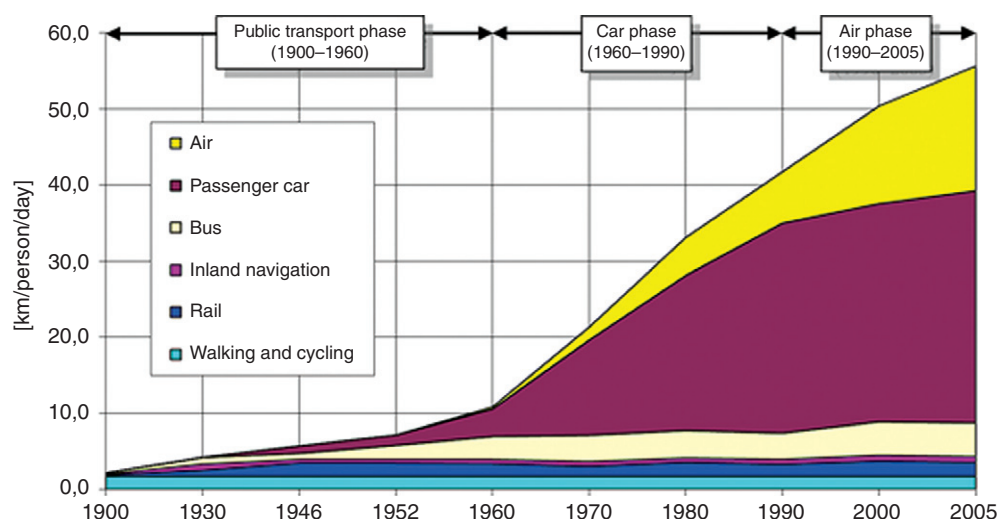


Figure 1 Passenger transport in Norway, 1900–2005. Nonmotorized and motorized transport. Source: [Holden \(2007\)](#).

- Transport is a major contributor to air-, soil- and water pollution at the local, regional, and global scale. For example, transport is currently the source of 24% of global CO₂ emissions.
- Despite the fact that transport networks and infrastructures cover only 3% of artificial land in Europe, its associated impacts can have dire consequences for biodiversity due to land fragmentation.
- Furthermore, annual casualties from road traffic are 1.35 million people worldwide, and cost most countries 3% of their GDP.
- Finally, access to mobility services is unevenly distributed, resulting inequality issues with respect to public and private services access and potentially social exclusion.

The previous points lead to diagnosing the mobility system as unsustainable. Furthermore, without major changes in mobility policies and practices, this unsustainability will continue well toward the end of this century.

Unfortunately, the persistence and complex nature of the transport system makes influencing mobility patterns difficult. The persistent nature lies in the fact contemporary investments in transport—both infrastructure for example, roads, bridges and airports, and vehicles, for example, cars, trains and planes—will influence mobility patterns long into the future. Thus, changes to the transport system and subsequently mobility patterns requires a long period.

The transport system is complex, and consist of three sub-systems that are interconnected. The first subsystem covers motorized means of transport. Here, sustainable mobility requires an evaluation of both the technology relating to the means of transport and the total distance traveled by each transport mode. In other words, claims of sustainability are not limited to artifacts such as vehicles but pertain to the level of mobility in society. The second subsystem covers transport infrastructure. Here, sustainable mobility requires an additional evaluation of all impacts incurred during construction, use, maintenance, and decommissioning for each mode of transport, including the energy and resources embedded in the infrastructure itself (e.g., concrete, asphalt, and control systems). The third subsystem covers the energy system; particularly those energy carriers used to propel the aforementioned modes of transport. Here, sustainable mobility requires a comparative evaluation of both conventional energy systems with possible improvements, and alternative energy systems—covering both modes of transport and infrastructure. In summary, assessing the sustainability of a transport system requires the long-term integrated assessment—including both positive and negative impacts—of all three subsystems.

A Typology for Achieving Sustainable Mobility

Sustainable mobility means being able to address three elements simultaneously. The first element focuses on *what* strategies are relevant, and the second focuses on *who* the agents are that could take the lead. We combine these first two elements, thus creating a matrix populated by paths (element 3) of *how sustainable mobility* can be achieved and which actions could be taken to achieve them. **Table 1** presents the aforementioned matrix; with *what* (rows), *who* (columns), and *how* (cells) elements. Combined, these rows, columns and cells provide a typology for sustainable mobility paths.

The simplicity of this matrix conceals the inevitable gaps and overlaps in practical policy between strategies, agents and paths. Nevertheless, each strategy, agent, and path reflects an important area of contemporary interest. Furthermore, this differentiation promotes understanding and facilitates analysis.

The Main Strategies

It is logical to think that we can achieve sustainable mobility if we: travel more efficiently, travel differently and/or travel less. From this, one can glean three main strategies for sustainable mobility, which are: efficiency-, alteration-, and reduction strategies (elaborated on below).

The efficiency strategy entails improving the sustainability of mobility through more efficient-novel technologies. Here, the concept of “technology” is used in a broad sense; it includes the use of both *hard*- and *soft* technology (e.g., vehicle technology and fuel shifts vs. information, trip linking, apps, and logistics). More efficient technologies can be implemented in three parts of the transport system, that is, modes, infrastructure, and energy system.

Table 1 A typology of sustainable mobility paths

		<i>Agents (Who?)</i>		
		<i>The experts</i>	<i>The people</i>	<i>The firms</i>
Strategy (What?)	Efficiency (Improve)	“The green government”	“The green purchaser”	“The clean vehicles”
	Alteration (Shift)	“The public transport provider”	“The responsible traveler”	“The shared mobility schemes”
	Reduction (Avoid)	“The compact city”	“The essential life”	“The traveling electrons”

Source: Holden et al. (2020)

The *alteration strategy* entails the implementation of modal shifts to change existing transport patterns. Currently, cars and planes dominated the prevailing transport patterns; and the necessary shift is towards more collective, affordable and well-functioning forms of transport, namely public transport, for example, buses, trains, and trams, along with increased shared mobility, walking and cycling. An added benefit of which is that many of these alternative forms are more energy efficient than cars and planes. Furthermore, an affordable and well-functioning (public)-transport system increases accessibility for low-mobility groups.

The *reduction strategy* does not diminish the roles of the efficiency and alteration strategies—which both result in cuts to energy consumption and emissions. However, these cuts are not sufficient on their own to achieve sustainable mobility. Moreover, expected growth in transport demand can negate these reductions achieved by the first two strategies alone. Thus, a collective strategy must be formulated which includes an absolute *reduction* in motorized travel—which at the same time allows for those whose basic transport needs are not yet met to do so. Such *reductions* can be achieved by traveling less and/or performing shorter trips, for example, from compact land-use planning, increased internet shopping and telecommuting, and changing established travel preferences.

The Main Agents

The aforementioned strategy for achieving sustainable mobility is imperative, but someone (an agent) must take the lead in enacting the strategy (or strategies). It should be noted that no single agent may have either the responsibility or the possibility of implementing a particular strategy alone. Instead, finding strategic alliances and alignment of interests between agents is more likely. Thus, several agents at all levels should contribute, including national and local authorities, small and large companies, and the public. Though—with this said—someone must *take the lead*, notably key agents possessing the necessary powers (and resources) to implement the first steps driving the changes forward. These key agents can be classified into three main groups:

The experts: Politicians (doing politics) and their bureaucrats (implementing policies) populate the first group—these experts put administrative rationalism at the forefront. In this group, a close collaboration between science, professional administration, and a hierarchical bureaucratic structure forms the foundation for implementation. Here, the experts' solutions and the bureaucrat's implementation of these solutions trump the potentially preferred solutions of the *people* and *firms*—the two remaining groups. These experts draw on the collective experiences of established institutions, instruments and practices, such as regulatory agencies and policy instruments, environmental assessment, expert advisory commissions, policy analysis techniques, and public R&D programs.

The people: The second group is the people exercising democratic pragmatism at the forefront. While the previous group (*experts*) offers a top-down approach, the *people* promote a bottom-up approach, recognizing conditions such as learning, trial-and-error, and incremental progress toward achieving their strateg(y/ies). The *people's* democratic pragmatism is grounded on dialogue between networks of local agents, for example, local officials, nongovernmental organizations, lobbyists, activists, journalists, corporations, and international organizations; as opposed to administrative rationalism which is based on central leadership and hierarchy. The *people* draw on devices such as public consultation, alternative dispute resolution, policy dialogue, lay citizen deliberation, public inquiries, right-to-know legislation, and the power of social movements.

The firms: The third group (*firms*) embrace market mechanisms based on voluntary transactions between sometimes competing players towards achieving their sustainable mobility strategies. Good solutions offered by such markets can often be overlooked by administrative rationalism and democratic pragmatism that are based on political or social priorities respectively. It should be noted that the *firm's* relationship to administrative rationalism is twofold; hardline economic rationalists would prefer to minimize the role of the authorities, introduce private property rights (incl. natural resources and pollution), and place transactions between market actors at the forefront. More moderate economic rationalists acknowledge that authorities should *design*—or create even—markets through the use of financial instruments, thus authoring the rules to which market transactions must comply. Negotiable quotas, green taxes, fines, and incentives are some of the instruments the authorities can implement to provide for a favorable market. Here, economic rationalism appeals to the consumer, whereas democratic pragmatism appeals to the citizen. Consumers can use their influence in choosing environmentally friendly products and solutions, thus inadvertently influencing those products and solutions that on-offer by firms. In such a market, it is the most environmental-friendly firms who will prosper.

The Sustainable Mobility Paths

Derived from these three main strategies (*efficiency, alteration, and reduction*), and the three main agents (*experts, people and firms*), we derive nine sustainable mobility paths—presented in **Table 1** and the forthcoming text. Unavoidably, some will claim that other/additional paths have lost our attention, and it is likely that others would propose an alternate design for identifying these paths. Nevertheless, we argue that the paths presented here are firmly anchored in basic concepts from which any discussion of sustainable mobility can venture. Furthermore, the paths presented here cover the range of ideas found in the literature – building on a single strategy and one agent, these are:

The green government (1): Here, a top-down approach by governing bodies at national, regional or local level who impose regulations and standards or taxes and subsidies on the existing modes of mobility is suggested. This embodies the most frequent contemporary approach focused on emission control (e.g., European Emission Standards and the Motor Vehicle Air Pollution Control Act (Clean Air Act)), though would also include subsidizing of technology that is more efficient.

The green purchaser (2): Here, new- greener and efficient technologies are introduced to fulfill the demands of an informed, but slightly hesitant population. Hesitant, since the green purchaser does not necessarily wish to sacrifice their travel habits for a more

sustainable mobility system. Accordingly, the green purchaser seeks cleaner- more efficient options, as opposed to alternatives or reducing their travel. Some examples of this path include the proliferation of EVs in affluent areas, and voluntary carbon offsets for air travel.

The clean vehicles (3): Companies take the lead role towards achieving sustainable mobility in this path. These companies can range from the opportunistic (i.e., bidding to appease regulatory bodies and/or appeal to individual's green ambitions) to the altruistic. As opposed to throwing existing business models out the window—in this path—safe-incremental improvements are implemented towards cleaning existing vehicle technology (e.g., 83 and 276).

The public transport provider (4): Here, governments at the national, regional and local levels provide the basic infrastructure- and services for collective transport – opting to own and operate such services themselves, or purchasing these services from private companies. Additive to this is the government's promotion of public transport and discouragement of private vehicle use. For such governments, several instruments are available including: subsidizing travel fares by public transport, dedicating lanes for public transport, and tolling private cars in specific areas.

The responsible traveler (5): As opposed to the *green purchaser* path, here, the population is both informed and willing to choose more sustainable alternatives for mobility, that is, exchanging old mobility habits for new routines (e.g., bicycling instead of driving). These slight modifications in lifestyle do not come at the expense of satisfy general *needs* and *wants* involving mobility, and are often accompanied with added benefits (e.g., health). These decisions in some instances, are part of a grassroots innovation where bottom-up solutions are developed by activists and organizations responding to the interests and values of the community.

The shared mobility schemes (6): Here, the companies at the forefront are those who introduce disruptive technologies, that is, technologies that significantly alter the prevailing mobility system. This is as opposed to path three (clean vehicles) and it's incremental improvements to existing vehicle technology. These can be either intra- or intermodal disruptions, or as completely novel modes. Uber's introduction of car sharing, and autonomous road vehicles are examples of the former, and drone delivery vehicles the latter.

The compact city (7): As the name suggests, this path proposes a high-density built environment and intensification of activities, efficient land-use planning, and diverse and mixed land uses, with clear (non sprawling) boundaries. Here, transportation systems are efficient and rely primarily on public- and active forms of transportation (e.g., cycling and walking) made possible by shorter trip lengths—and which are often more attractive and desirable. As such, the compact city seeks to increase quality of life in its city quarters, (re-) devoting road space to urban parks, cafés and stores.

The essential life (8): The changes in lifestyle incurred by the population in this path are much deeper than the previous two (paths 2 and 5), and requires a transformation in how we define and fulfil *needs*. Here, the population must consider the full scope of Sustainable Development Goals and their perspectives on sociality, equity, work, and environment. The essential life requires extending the current 'local' trend to new dimensions—both in breadth (e.g., entertainment and tourism) and depth (e.g., the low carbon economy, or the 'tight' circular economy). Digital solutions will allow us to remain global citizens.

The traveling electrons (9): Here, allowing electrons to travel instead of people is made possible by information and communications technologies. In other words, telecommuting, the use of video conferences, and exchanging all types of information on the internet reduces people's need for physical travel. Despite studies showing the large embedded energy use and emissions associated with the necessary infrastructure, and the potential need (and want) to increase travel resulting from greater access to information and communication technologies, this is nevertheless a path which bears the *potential* to offer great reductions in energy use and emissions.

Concluding Remarks

The nine paths presented in this chapter outline the dominant paths towards achieving sustainable mobility. They introduce key elements into the sustainable mobility debate, and emphasize the plethora of strategies and agents involved. Individually, they are unlikely to make a real difference, and it is only in combination that a truly sustainable mobility system can be realized. Future work on these paths must include three steps.

First, this is a conceptual review of sustainable mobility paths. Thus, we do not provide detailed descriptions of how to achieve a sustainable energy system. Such descriptions must be accompanied with a set of concrete action in terms of governmental means and policies, individual behavior changes, and new business models and strategies. Still, choosing paths is important because they form and legitimize policy and behavior, and can create the necessary momentum for political movement and behavior change.

Second, each path must be empirically assessed against its credibility. There are three criteria by which paths must be assessed: their feasibility (are they possible?), acceptability (do we approve them?), and centrality (do they deliver sustainability?). All criteria are crucial. On one hand, a path that delivers sustainability but is neither feasible nor acceptable is of little help. On the other hand, a path that is feasible and acceptable, but fails to deliver sustainability is of little use.

Third, we need to study the relationship between the paths. They can be complementary (each focusing on different aspects of sustainable development), competing (each claiming to be the best solution to a particular aspect), and substitutional (each reducing the prospect of the other). They all belong, however, to the wider set of narratives that make up the sustainable energy literature and therefore deserve attention here.

See Also

Equity Concerns in Transport; Transport Modes and Accessibility; Energy Consumption of Transport Modes; Knowledge-base on Sustainable Urban Land use and Transport; Transport Air Pollution and Health

Biography

Erling Holden is a professor in Renewable Energy at the Norwegian University of Life Sciences, and professor II at the Western Norway University of Applied Sciences. He has worked with issues related to energy, transport and sustainable development since 1988. In this work, Holden combines technological-oriented environmental studies, sociological and socio-psychological behavioural studies, and physical planning studies. Since 2008 he has also worked with projects related to renewable energy projects' impacts on local economies, local societies, and local environments in pursuing a sustainable energy policy. He is author of *The Imperatives of Sustainable Development* (2017. Abingdon: Routledge) and *Achieving Sustainable Mobility* (2007. Farnham: Ashgate).

Geoffrey Gilpin is an associate professor in the Renewable Energy Program at the Western Norway University of Applied Sciences. He holds a PhD (2016) Environmental Physics and Renewable Energy from the Department of Mathematical Sciences and Technology, at the Norwegian University of Life Sciences (NMBU). He worked as a researcher at NMBU (2006–2008), at the Western Norway Research Institute (2008–15), and in his current position works primarily in the application of industrial ecology methods to the fields of renewable energy and sustainability.

Further Reading

- Banister, D., 2008. The sustainable mobility paradigm. *Transport Policy* 15 (1), 73–80.
- Banister, D., 2018. *Inequality in Transport*. Alexandrine Press, Marcham.
- Berger, G., Feindt, P., Rubik, F., Holden, E., 2014. Sustainable Mobility—Challenges for a Complex Transition. *J. Environ. Policy Plan.* 16 (3), 303–320.
- Black, W.R., 2003. *Transportation: A Geographical Analysis*. Guilford Press, London, UK.
- Cervero, R., Guerra, E., Al, S., 2017. *Beyond Mobility: Planning Cities for People and Places*. Island Press, London.
- Gough, I., 2017. *Heat, Greed and Human Need: climate Change, Capitalism and Sustainable Wellbeing*. Edward Elgar, Cheltenham.
- Gössling, S., 2017. *The Psychology of the Car. Automobile Admiration, Attachment, and Addiction*. Elsevier.
- Holden, E., 2007. *Achieving Sustainable Mobility: Everyday and Leisure-time Travel in the EU*. Ashgate, Aldershot.
- Holden, E., Gilpin, G., Banister, D., 2019. Sustainable Mobility at thirty. *Sustainability* 11 (7), 1965, doi:10.3390/su11071965.
- Holden, E., Gilpin, G., Banister, D., 2020. Grand narratives for sustainable mobility: A conceptual review. *Energy Res. Soc. Sci.* 65, 101454.
- Santos, G., 2018. Sustainability and shared mobility models. *Sustainability* 10, 3194.
- Schiller, P.L., Kenworthy, J.R., 2018. *An Introduction to Sustainable Transportation: Policy, Planning and Implementation*, 2nd ed. Routledge.
- UN Habitat (2013) *Planning and Design for Sustainable Urban Mobility, Global Report on Human Settlements 2013*, London: Routledge, <https://unhabitat.org/planning-and-design-for-sustainable-urban-mobility-global-report-on-human-settlements-2013/More>
- UN, 2015. *Transforming Our World: The 2030 Agenda for Sustainable Development*. Resolution adopted by the General Assembly on 25 September 2015, A/RES/70/1. United Nations General Assembly
- World Commission on Environment and Development (WCED), 1987. *Our Common Future*. Oxford University Press, Oxford.

Vehicles That Drive Themselves: What to Expect With Autonomous Vehicles

Michele D. Simoni^{*,†}, **Kara M. Kockelman^{*}**, ^{*}Department of Civil, Architectural, and Environmental Engineering, The University of Texas at Austin, Austin, TX, United States; [†]Center for Transportation and Logistics, Massachusetts Institute of Technology, Cambridge, MA, United States

© 2021 Elsevier Ltd. All rights reserved.

Background	19
AV Adoption	19
Travel Behavior Impacts	20
Traffic Impacts	20
Shared Mobility: SAVs + DRS	21
Freight Transportation	22
Conclusion	23
References	23
Further Reading	25

Background

Humans are headed into a mobility revolution, thanks to the emergence of self-driving vehicles. Tesla's "Autopilot" feature has been used by many consumers in freeway conditions, and testing of fully automated or "autonomous" vehicles (AVs) has become rather common in many settings, like the City of San Francisco, Phoenix metro area, and Detroit region. Technological improvements and investments by vehicle manufacturers and technology companies suggest that AVs will be publicly accessible in many locations within 10 years, although design, production, and licensing challenges remain.

AV technologies may radically change several aspects of personal trip making and freight mobility, including crash counts and congestion, energy use and emissions, and land use decisions. For certain impact areas, one may be confident that benefits will outweigh costs, such as in safety and vehicle efficiency per mile driven. In the case of other impact areas, it is not clear whether the automation of vehicles ultimately will be viewed as beneficial, since making "driving" easier adds demand to already congested roadways and can support further regional sprawl. Predicting AVs' direct and indirect effects is a daunting task, given the uncertainties inherent in AV technology development, public attitudes, policymaking, and safety standards. Nevertheless, it is valuable at this early stage to try understanding all potential implications in order to move toward more sustainable/less unsustainable travel choices and urban systems.

This paper provides an overview of such topics by surveying recent research results and publicly available reports. After a discussion of AV adoption forecasts, we examine travel choice, traffic and congestion impacts, shared vehicle systems, and shared rides. We conclude with a discussion on potential impacts for freight movement. Readers are referred to many references throughout these sections, including a section on "Further Reading" for more details.

AV Adoption

There has been much news hype regarding release of AVs to the public, driven in part by competition within the manufacturing space, for investment and returns. Machine learning algorithms have advanced quickly, while sensor prices have fallen, but "corner cases" in design remain daunting (unusual situations that could bring to wrong detection) and LiDAR remains expensive. Some of the most detailed work in household-fleet forecasting, recognizing the long lifetime (17 years on average) of existing household vehicles, includes that by [Bansal and Kockelman \(2017\)](#), who surveyed thousands of Americans regarding their vehicle holding choices and simulated, year by year, US household fleet evolution. Those predictions do not suggest 80% adoption until year 2050 or later, without public policy interventions (like AV-only requirements in sales and/or use at certain points in time). Using a bass diffusion model, with parameters based on the adoption of other technologies in the United States, [Lavasani et al. \(2016\)](#) suspect that technology cost will play little role. [Clements and Kockelman \(2017\)](#) estimate the value-of-time and crash-avoidance benefits to be so valuable to most travelers in developed countries (with high wages) that the choice may be easy (saving drivers effectively 100 + hours of time or almost \$2,000 in time-cost every year, with similar property and injury benefits), once vehicles become available in showrooms. But experience can be critical, and manufacturers do not want to release vehicles that are not yet ready for prime-time success. As to the potential initial customers of AVs, they will likely be young, more educated, and more frequent travelers ([Haboucha et al., 2017](#)). The major determinants behind their adoption are consumer perceptions of their improved usefulness and trustworthiness ([Choi and Ji, 2015](#).)

People's opinions can and do change, based on what they and their colleagues, friends, family members, and neighbors are experiencing (like being served by a shared AV (SAV) while on business travel). So preferences will evolve alongside AV cost

reductions and policymaking, presumably shifting those survey-data-based adoption curves to higher ownership and use rates. In addition, most or nearly all early AV releases may be to professional fleet managers (like existing car-rental companies and ride-hailing companies), who sign lengthy contracts agreeing to restricted use cases (like specific corridors and weather contexts, under specific lighting situations) and special maintenance provisions (like daily equipment reboots, algorithm checks, and sensor cleaning). So household AV adoption may not be nearly as important as SAV release and adoption, raising roadways' AV vehicle-miles traveled (VMT) quickly. Attitudinal and lifestyle factors, such as a person's tech-savviness and past use of ride-hailing services, together with safety concerns and expectations of driving automation, will play important roles in AV and SAV adoption (Bansal et al., 2016; Lavieri et al., 2017; Nazari et al., 2018). The extensive US survey work and household simulation over time suggest that one-third of US personal ground travel in year 2050 may be by SAVs (with about one-third of that in true ride-sharing—with a stranger) and that 60% or more of VMT should be in self-driving mode.

Travel Behavior Impacts

Vehicle automation is widely agreed to affect travelers' travel cost and experiences—both monetary and perceived. The fixed costs of owning an AV will be high (at least in the early, adoption phase) to accommodate expensive equipment, including computers and cameras, radar and presumably LiDAR, automated controls and wireless communications for safe, accurate, and rapid navigation. Initial prices are expected to be \$20,000 or more above base comparable vehicles (Fagnant and Kockelman 2018, Litman, 2019). Maintenance costs may rise as well, thanks to the added vehicle complexity, specialized sensor maintenance, and more expensive equipment replacement (Litman, 2019). On the other hand, AV operating costs should fall, thanks to sizable reductions in collision insurance costs (Bösch et al., 2018; Fagnant and Kockelman, 2015) and per-mile fuel savings of 5%–20%, depending on driving environment and user preferences, thanks to speed synchronization, inter-vehicle coordination, and eco driving (Lee and Kockelman 2019; Stephens et al., 2016). Parking search and use costs may also fall, as travelers rely on SAVs for many trips to high-parking-cost locations (like central business districts) and/or if AVs are permitted to park away from where they drop off their owners/users (Litman, 2019) and/or at least better anticipate available spaces (Shoup, 2006).

There is not yet strong agreement on the value of travel time (VOTT) effects of using AVs. Most experts expect significant VOTT reductions since AVs allow former drivers (though not former passengers, presumably) to suddenly engage in leisure and work activities (Auld et al., 2017; Bansal and Kockelman, 2017; van den Berg and Verhoef, 2016; Zhao and Kockelman, 2018) and to reduce driving stress (Singleton, 2019). Of course, being stuck in a vehicle is still limiting, so VOTTs are not expected to fall more, say, 50% for drivers (Cyganski et al., 2015; Lenz et al., 2016; Yap et al., 2016). Given the importance of VOTT assumptions in travel demand modeling and network performance predictions, further research involving revealed preference data (rather than stated preference surveys) is needed.

A key outcome of lower VOTTs is higher VMT. This addition to VMT comes alongside empty driving by SAVs, better “driving” access for those with disabilities (Harper et al., 2016) and lowered monetary costs of “driving” that translates in higher frequencies and longer distance traveled to the detriment of public transit and active modes competitiveness (Loeb and Kockelman, 2019; Simoni et al., 2019). All together, these behavioral changes portend a potentially dramatic rise in VMT demand and thus very serious congestion issues. Different simulation-based studies anticipate overall VMT increases between 10% and 50% from the adoption of (privately owned) AVs, depending on their cost and competition with other modes (Auld et al., 2017; Kim et al., 2015; Zhang et al., 2018). Significant increases are also evident in SAV-focused studies (Childress et al., 2015; Fagnant and Kockelman 2018; Liu et al., 2017; Loeb and Kockelman, 2019), due to about 5%–25% extra VMT from empty driving between passengers. Finally, AVs may dramatically affect long-distance travel choices (modes and destinations), by reducing air and rail travel by passengers and freight, while increasing highway travel between regions (Huang and Kockelman, 2018; Perrine et al., 2017).

In order to limit the externalities related to the increase of motorized trips, different travel demand management strategies could be adopted. Lamotte et al. (2017) propose a reservation-based strategy to manage automated vehicles' departures in Vickrey's bottleneck model. Simoni et al. (2019) assess alternative congestion pricing schemes to curb AV and SAV demand in alternative future scenarios. Tscharaktschiew and Evangelinos (2019) evaluate tolls to affect travelers' decision to switch between autonomous and manual driving and limit congestion. Conversely, Lee and Kockelman (2019) suggest that traffic flow benefits from AV adoption might obviate the need of congestion pricing schemes.

Traffic Impacts

Connected and automated driving systems (CADS) may increase traffic flow along lanes and through intersections and other points of conflicts by improving reaction maneuvers and intervehicle coordination decisions, reducing headways and crashes, and facilitating merging and mini-platoons. Benefits over time ultimately depend on manufacturers' design decisions, government regulations, and consumer adoption rates.

Thanks to having information on more vehicles around them, along with shorter reaction times, highly automated vehicles can reduce headways and increase road capacity (Hoogendoorn et al., 2014; Tientrakool et al., 2011; van Arem et al., 2006). For example, Shladover et al.'s (2012) simulations suggest basic-freeway link capacity increases up to 80%, simply thanks to full adoption and application of cooperative adaptive cruise control (CACC). Most existing studies anticipate considerable flow improvements only at

high adoption rates (e.g., above 50%), but [Jerath and Brennan \(2012\)](#) and [Kesting et al. \(2008\)](#) anticipate benefits at lower rates (e.g., above 30%). [Talebpoor and Mahmassani \(2016\)](#) by means of microsimulation identify automation as the main factor behind the capacity improvement (thanks to a significant increase of traffic stability) especially in comparison to communication technologies implemented alone. However, the highest benefits will be mostly likely achieved by the possibility of quickly sharing traffic information among vehicles, processing it by means of advanced algorithms, and identifying efficient solutions without involving humans.

Some recent theoretical traffic-flow work emphasizes general formulations for feasible lane policies ([Chen et al., 2016](#)) and adaptations of existing models ([Levin and Boyles, 2016](#)) to try and reproduce mixed-flow behaviors involving both, conventionally driven and fully automated vehicles. By investigating the influence of different factors (e.g., minimum inter-vehicle spacing, headway options, and market penetration, researchers like [Bujanovic and Lochrane \(2018\)](#) developed analytical models for predicting platoons' traffic effects based on platoon lengths and CADS penetration rates.

Other studies have focused on achieving better road network use via innovative control strategies using advanced automation and communication technologies. For example, [Dresner and Stone \(2008\)](#) proposed, and then microsimulated autonomous intersection management, where vehicles approaching an intersection can coordinate arrivals and movements via reservation, greatly reducing travel delays. [Sharon et al. \(2017\)](#) developed link-based "micro-tolls" across a downtown network that dynamically vary to reflect and thus lower congestion costs from AVs' routing choices. Several (CACC)-based strategies have been developed to stabilize traffic flows and thereby help avoid triggering certain types of congestion ([Milanés and Shladover, 2014](#); [Naus et al., 2010](#); [Wang et al., 2014](#)). The estimated capacity gain from CACC-equipped vehicles can almost correspond to 90% of the original throughput for low headways (e.g., below 1 s) and high penetration rates (above 90%). [Zhou et al. \(2017\)](#) proposed an AV-based strategy to maximize highway on-ramps' efficiency by smoothening traffic oscillations with halved speed dispersion rates for 25% AV penetration rates. [Stern et al. \(2018\)](#) show how a single AV can be effectively used to dampen stop-and-go waves (with 10%–15% throughput increases for a single AV controlling the flow of 20 "traditional" vehicles). [Wu et al. \(2018\)](#) obtained similar results using artificial intelligence-based control techniques for AV flows. It is worth mentioning that in these studies, only vehicular flows are investigated and conflicts with other modes such as pedestrians and cyclists are not considered.

Finally, CADS are expected to significantly reduce crashes rates ([Li and Kockelman, 2018](#)), since over 90% of public-roadway crashes involve human error ([NHTSA, 2008](#)). Since such crashes currently account for about 25% of roadway delay costs in the United States and presumably elsewhere ([Systematics, 2005](#)), CADS may reduce congestion costs by that much in the long run (assuming very high market penetration rates, which may come with regulations, much like seat belts and electronic stability control requirements on all new vehicles being sold in most developed countries). It should also be noted that reduction in nonrecurring congestion from crashes and from reduced congestion overall should improve reliability of the entire transportation system (by reducing uncertainty or variability in door-to-door travel times). This does require that AV technologies plus variable road pricing strategies, for example, are able to avoid the added congestion and potential gridlock that many expect from having too many AVs and other vehicles on the road). Moreover, improved safety and congestion delay reductions often involve a trade-off: in order to reduce crash likelihood, larger spacings between vehicles and lower speeds are often desired. Finally, long-term testing and programming will still be needed before this technology can guarantee zero (or close) risk of failure.

Shared Mobility: SAVs + DRS

High initial costs for vehicle automation, special maintenance needs, and use restrictions on early AVs will favor deployment of heavily used shared fleets of AVs or "SAVs". Since automation can significantly reduce operating costs for taxi and ride-hailing fleet owners or managers, transportation network companies (like Lyft, Uber, and Didi) are running AV tests ([Hawkings, 2017](#); [Kang, 2016](#)), investing considerable resources ([Buhr, 2017](#)) and developing strategic collaborations with automakers and governments ([Russell, 2017](#)) for future, large-scale SAV deployments. Automakers like Ford, GM, Fiat Chrysler, BMW, Daimler and Volvo are also considering the possibility of serving as SAV providers ([Stocker and Shaheen, 2017](#)). Even cities and local transportation authorities are considering SAVs as potential development for the future mobility services. For example, in 2016, Switzerland's largest bus company ([PostBus](#)) launched an AV shuttle service trial in the city of Sion to serve less accessible areas ([Lavars, 2016](#)). That same year, a local transit operator tested a similar service in Lyon, France ([Pultarova, 2016](#)). Since then, several cities worldwide have been launching similar pilot tests (e.g., Paris, Helsinki, and London in Europe, Ann Arbor and Las Vegas in the United States) in order to provide innovative "first-/last-mile" solutions (linking local vehicles with longer-distance, existing, fixed-rail investments, for example). Waymo, a Google-initiated program of SAVs, has been serving households around the Phoenix, Arizona region as well, using plug-in hybrid-electric minivans ([DeBord, 2018](#)).

From a sustainability perspective, self-driving, on-demand mobility can offer an attractive alternative to privately owned and, often, less fuel-efficient vehicles while ideally enabling efficient integration with existing transit system investments. Nevertheless, SAVs also often act in direct competition to existing public transport services, and can generate new VMT. In order to steer the development of SAV services in less unsustainable directions, it is important to better understand SAVs users' perceptions, along with likely operations, performance metrics, passenger-mile costs, and other SAV-fleet implications for mobility and the natural environment.

A few studies have explicitly explored SAV use tendencies, using stated preference surveys. [Bansal et al.'s \(2016\)](#) survey found that Austin, Texas respondents more inclined to state a willingness to use SAVs are males, full-time workers, tech-savvy individuals, and

those living in densely populated urban areas. Similarly, [Krueger et al. \(2016\)](#) identified relatively young males as the most interested demographic for use of SAVs via a survey across several major Australian cities. Meaningfully, a distinction between car drivers and car passengers emerged from this survey, with the latter more inclined to make use of SAVs with dynamic ride-sharing (DRS). Using paired comparisons of current and future travel modes in Germany, [Pakusch et al. \(2018\)](#) anticipated a growth of car sharing thanks to automation, and mainly at the expense of public transit. Although such studies can provide useful indications of trends and key factors at play in SAV adoption, reliance on stated preferences is always tricky. Other studies use existing parameters and estimates of VOTT changes to anticipate adoption rates largely via existing parameters and mode-cost estimates in transport planning models ([Davidson and Spinoulas, 2015](#); [Huang and Kockelman, 2019](#); [Lavieri et al., 2017](#); [Zhao and Kockelman, 2018](#)).

Fleet performance metrics are also crucial in anticipating SAV competitiveness and sustainability. Using simulation of vehicles and travelers, [Fagnant and Kockelman \(2015\)](#) demonstrated how SAVs may be able to replace up to 10 conventional vehicles in a small region or town while reducing carbon monoxide and various other emissions, at the expense of a 10%–20% increase in VMT. [Chen et al. \(2016\)](#) then explored the influence of different SAV-use pricing strategies with a fleet of range-constrained all-electric SAVs and highlighted trade-offs between fleet performance (including response times and empty VMT), fares, and overall demand for SAV services. In the case of all-electric SAVs, charging infrastructure investments and battery size/vehicle range can be crucial to the fleet's competitive performance ([Chen et al., 2016](#); [Loeb et al., 2018](#)). Furthermore, sophisticated customer assignment strategies (involving reassignment of travel requests among different vehicles) can be developed to improve fleet performance by lowering response times during the most congested hours ([Hyland and Mahmassani, 2018](#)).

Of course, dynamic/real-time matching of overlapping rides for distinct travelers via DRS should also reduce SAVs' congestion impacts and environmental footprint. A few studies, mainly based on simulation, have explicitly explored the issue of DRS for SAVs and examined the propensity to share rides based on fares and levels-of-service ([Fagnant and Kockelman 2018](#); [Farhan and Chen, 2018](#); [Loeb and Kockelman, 2019](#)). Recently, [Gurumurthy et al. \(2019\)](#) found that pricing strategies reflecting overall levels of congestion and fleet usage can be particularly useful in incentivizing users to adopt DRS and thereby reduce SAVs' VMT.

Freight Transportation

Automation will affect freight transport via the use of platooned trucks relying on CACC technologies, self-driving trucks, drone deliveries, and ground-based robots. AV research in freight transport has focused on long-distance trucking applications and drone-based delivery systems.

Assisted highway trucking may be one of the first AV applications in public use, thanks to high trucking costs (for labor and vehicles) and long-distance use, making the economic case for AV investment clear and rather immediate. Using a random-utility-based multi-regional input-output mode of production and trade across the United States, [Huang and Kockelman \(2019\)](#) estimate a long-term shift away from conventional trucks and railroad to automated heavy trucks for routes over 500 miles. Smart, safe, truck-driving, and vehicle-following algorithms can improve safety, give operators needed rest for longer drives, and decrease fuel consumption by about 10%, when platooned ([Alam et al., 2015](#); [Bergenheim et al., 2012](#); [Tsugawa et al., 2011](#)). Operators may take control for the shortest and most complex sections of long journeys while dealing with other tasks and supervising the cruise during the long legs of the trip. Although truck platooning has been successfully tested worldwide in the last few years ([Tsugawa et al., 2016](#)) several safety issues (e.g., V2V communication failures, hijacking, infrastructure constraints) and regulations (e.g., platoon size limits, road prohibitions, liability rules) will need to be addressed before implementation ([Janssen et al., 2015](#)).

Smaller, autonomous, or semiautonomous ground vans and single-unit trucks may also be employed for delivery, with a person on board to perform loading/unloading tasks and/or with different compartments that can be unlocked by customers at destinations ([DHL, 2014](#)). Drones or Unmanned Aerial Vehicles (UAVs) may soon improve supply chains and freight operations by delivering very small (weighting less than 3 kilograms) packages over the "last mile" of urban trips and/or by performing deliveries in less accessible areas. UAV use offers several advantages over conventional vehicles. For example, drones can typically perform quicker deliveries than trucks since they are not as constrained by street networks (although they might be mandated to fly over public right of ways in cities) and hindered by congestion. They also can more easily deliver lightweight packages in rural areas with low demand and in difficult terrains. Various supply chain businesses (like Amazon, DHL, and Google) have been testing drone-based solutions in an effort to reduce last-leg distribution costs ([Hern, 2014](#)). In 2019 UPS started the first commercial drone-delivery service in order to deliver medical supplies in North Carolina ([Stradling, 2019](#)). Several recent studies have explored the profitability of drones and their integration in current delivery systems using optimization and business approaches ([Otto et al. \(2018\)](#) for a comprehensive review). Depending on the speed, flight range, carrying capacity, and deployment strategy drones can yield up to 30% savings (over the traditional TSP) in total delivery time ([Agatz et al., 2018](#); [Murray and Chu, 2015](#)). However, drones will probably be employed for only a small share of total delivery volumes because of fly-height and battery-range restrictions. Public safety concerns can also slow their adoption.

Another automotive innovation with potential applications in the area of last-mile freight distribution is delivery robots. Several companies are exploring the possibility of delivering groceries and parcels via robots operating on sidewalks (Starship, Dispatch, Marble) and roads (Nuro). Similarly to drones, robots could be employed in combination with larger delivery trucks that cannot so easily penetrate neighborhoods or park in front of delivery addresses. Before large-scale implementation, however, issues related to their operations in crowded areas and regulations will need to be addressed ([Hoffmann and Prause, 2018](#)).

Conclusion

AVs are expected to transform the passenger and freight mobility landscapes in several ways. This chapter discusses their expected impacts on travel choices, traffic flows, shared mobility, and freight distribution, by reviewing various research findings. Several decades may pass before AVs' large-scale impacts become visible in most locations, due the technological challenges, costs, and other complex issues that surround AV use—including public perceptions, demonstrated safety statistics, liability (for manufacturers as well as policy makers), market trends, and specific deployment decisions and settings. Nevertheless, it is important to gain a better understanding of AVs' many potential effects at this early stage, in order to foster more sustainable travel choices and smarter transport-system investments.

While autonomous and connected driving has great potential to ultimately improve traffic efficiency, its extent will strongly depend on demand decisions, public policy, and user trust, in addition to technical challenges from algorithm and sensor development. As discussed here, travel costs for persons and freight are expected to fall thanks to once-drivers being able to perform other activities en route and travel longer distances without driver fatigue and sleep requirements. One downside of lower VOTTs, however, is a rise in trip making, motorized vehicle use, travel distances, and suddenly empty-vehicle travel. If empty VMT is not limited by law or pricing policies and local-intersection operational overhauls, congestion can become much worse in urban areas. Stated preference surveys and extensions of existing travel models offer meaningful indications of potential vehicle-choice, travel-choice, and network-condition changes, but additional research based on real observations will be needed in the future to better anticipate long-term shifts while developing more appropriate policies.

While added VMT may swamp any traffic benefits from AVs and SAVs, in the near term, highway and major arterial's streets' traffic performance may eventually benefit from the possibility of shorter headways and improved coordination at intersections. Within this context, advanced cellular and radio-based communication technologies will presumably play a key role in enabling effective coordination among vehicles and hopefully with pedestrians and cyclists. Moreover, shared mobility options may greatly facilitate inter-modal trips, higher-occupancy (shared-ride) vehicle use, and a regular reliance on battery-only vehicles without range anxiety for users. Of course, automation via drones and robots may also considerably change freight-delivery landscapes, in terms of last-mile distribution.

References

- Agatz, N., Bouman, P., Schmidt, M., 2018. Optimization approaches for the traveling salesman problem with drone. *Transp. Sci.* 52 (4), 965–981.
- Alam, A., Besselink, B., Turri, V., Martensson, J., Johansson, K.H., 2015. Heavy-duty vehicle platooning for sustainable freight transportation: a cooperative method to enhance safety and efficiency. *IEEE Contr. Syst. Mag.* 35 (6), 34–56.
- Auld, J., Sokolov, V., Stephens, T.S., 2017. Analysis of the effects of connected–automated vehicle technologies on travel demand. *Transp. Res. Rec.* 2625, 1–8.
- Bansal, P., Kockelman, K.M., 2017. Forecasting Americans' long-term adoption of connected and autonomous vehicle technologies. *Transp. Res. A Policy Pract.* 95, 49–63.
- Bansal, P., Kockelman, K.M., Singh, A., 2016. Assessing public opinions of and interest in new vehicle technologies: an Austin perspective. *Transp. Res. C: Emer. Technol.* 67, 1–14.
- Bergenheim, C., Shladover, S., Coelingh, E., Englund, C., Tsugawa, S., 2012. Overview of platooning systems. In: *Proceedings of the 19th ITS World Congress*, 22–26 October, Vienna, Austria.
- Bösch, P.M., Becker, F., Becker, H., Axhausen, K.W., 2018. Cost-based analysis of autonomous mobility services. *Trans. Policy* 64, 76–91.
- Buhr, S., 2017. Lyft Launches a New Self-driving Division and Will Develop Its Own Autonomous Ride-hailing Technology. *Tech Crunch.com*. Available from: <https://techcrunch.com/2017/07/21/lyft-launches-a-new-self-driving-division-called-level-5-will-develop-its-own-self-driving-system/>.
- Bujanovic, P., Lochrane, T., 2018. Capacity predictions and capacity passenger car equivalents of platooning vehicles on basic segments. *J. Transp. Eng. A Sys.* 144 (10).
- Chen, T.D., Kockelman, K.M., Hanna, J.P., 2016. Operations of a shared, autonomous, electric vehicle fleet: implications of vehicle and charging infrastructure decisions. *Transp. Res.* 94, 243–254.
- Childress, S., Nichols, B., Charlton, B., Coe, S., 2015. Using an activity-based model to explore the potential impacts of automated vehicles. *Transp. Res. Rec.* 2493 (1), 99–106.
- Choi, J.K., Ji, Y.G., 2015. Investigating the importance of trust on adopting an autonomous vehicle. *Int. J. Hum. Comp. Inter.* 31 (10), 692–702.
- Clements, L.M., Kockelman, K.M., 2017. Economic effects of automated vehicles. *Transp. Res. Record* 2606.1, 106–114.
- Cyganski, R., Fraedrich, E., Lenz, B., 2015. Travel time valuation for automated driving: a use-case-driven study. In: *94th Annual Meeting of the Transportation Research Board*, Transportation Research Board, Washington, DC.
- Davidson, P., Spinoulas, A., 2015. Autonomous vehicles: what could this mean for the future of transport. Australian Institute of Traffic Planning and Management (AITPM) National Conference, Brisbane, Queensland.
- DeBord, M., 2018. Waymo has Launched Its Commercial Self-driving Service in Phoenix — and It's Called 'Waymo One'. *Business Insider*. Available from: <https://www.businessinsider.com/waymo-one-driverless-car-service-launches-in-phoenix-arizona-2018-12>.
- DHL Customer Solutions & Innovation, 2014. Self-driving vehicles in logistics: A DHL perspective on implications and use cases for the logistics industry, 2014. Publisher represented by Matthias Heutger, Senior Vice President, DHL Customer Solutions & Innovation.
- Dresner, K., Stone, P., 2008. A multiagent approach to autonomous intersection management. *J. Artif. Intell. Res.* 31, 591–656.
- Fagnant, D.J., Kockelman, K., 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transp. Res. A Policy Pract.* 77, 167–181.
- Fagnant, D.J., Kockelman, K.M., 2018. Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. *Transportation* 45 (1), 143–158.
- Farhan, J., Chen, T.D., 2018. Impact of ridesharing on operational efficiency of shared autonomous electric vehicle fleet. *Transp. Res. C Emerg. Technol.* 93, 310–321.
- Gurumurthy, K.M., Kockelman, K.M., Simoni, M.D., 2019. Benefits & costs of ride-sharing in shared automated vehicles across Austin, Texas: opportunities for congestion pricing. *Transp. Res. Rec.* 2673.
- Haboucha, C.J., Ishaq, R., Shifan, Y., 2017. User preferences regarding autonomous vehicles. *Transp. Res. C Emerg. Technol.* 78, 37–49.
- Harper, C.D., Hendrickson, C.T., Mangones, S., Samaras, C., 2016. Estimating potential increases in travel with autonomous vehicles for the non-driving, elderly and people with travel-restrictive medical conditions. *Transp. Res. C Emerg. Technol.* 72, 1–9.
- Hawkings, A.J., 2017. Uber's Self-driving Cars are Now Picking Up Passengers in Arizona. *The Verge*. Available from: <https://www.theverge.com/2017/2/21/14687346/uber-self-driving-car-arizona-pilot-ducey-california>.

- Hern, A., 2014. Amazon Claims First Successful Prime Air Drone Delivery. The Guardian. Available from: <https://www.theguardian.com/technology/2016/dec/14/amazon-claims-first-successful-prime-air-drone-delivery>.
- Hoffmann, T., Prause, G., 2018. On the regulatory framework for last-mile delivery robots. *Machines* 6 (3), 33.
- Hoogendoorn, R., van Arrem, B., Hoogendoorn, S., 2014. Automated driving, traffic flow efficiency, and human factors: literature review. *Transp. Res. Record* 2422 (1), 113–120.
- Huang, Y., Kockelman, K., 2019. How will self-driving vehicles affect U.S. megaregion traffic? The case of the Texas triangle. In: 98th Annual Meeting of the Transportation Research Board.
- Huang, Y., Kockelman, K., 2018. What will autonomous trucking do to U.S. trade flows? Application of the random-utility-based multi-regional input-output model. *Transportation*, 1–28.
- Hyland, M., Mahmassani, H.S., 2018. Dynamic autonomous vehicle fleet operations: optimization-based strategies to assign AVs to immediate traveler demand requests. *Transp. Res.* 92, 278–297.
- Janssen, R., Zwijsenbergh, H., Blankers, I., de Kruijff, J., 2015. Truck Platooning Driving the Future of Transportation. TNO. Available from: <https://www.logistiekprofs.nl/uploads/content/file/janssen-2015-truck.pdf>.
- Jerath, K., Brennan, S.N., 2012. Analytical prediction of self-organized traffic jams as a function of increasing ACC penetration. *IEEE Trans. Intell. Transport. Sys.* 13 (4), 1782–1791.
- Kang, C., 2016. Self-Driving Cars Gain Powerful Ally: The Government. The New York Times. Available from: <https://www.nytimes.com/2016/09/20/technology/self-driving-cars-guidelines.html>.
- Kesting, A., Treiber, M., Schönhof, M., Helbing, D., 2008. Adaptive cruise control design for active congestion avoidance. *Transp. Res. C Emerg. Technol.* 16 (6), 668–683.
- Kim, K., Rousseau, G., Freedman, J., Nicholson, J., 2015. The travel impact of autonomous vehicles in metro Atlanta through activity-based modeling. In: 15th TRB national transportation planning applications conference.
- Krueger, R., Rashidi, T.H., Rose, J.M., 2016. Preferences for shared autonomous vehicles. *Transp. Res. C Emerg. Technol.* 69, 343–355.
- Lamotte, R., De Palma, A., Geroliminis, N., 2017. On the use of reservation-based autonomous vehicles for demand management. *Transp. Res. B Methodol.* 99, 205–227.
- Lavars, N., 2016. Autonomous Buses Hit the Road in Switzerland. New Atlas. Available from: <https://newatlas.com/switzerland-autonomous-bus/44041/>.
- Lavasani, M., Jin, X., Du, Y., 2016. Market penetration model for autonomous vehicles based on previous technology adoption experiences. In: Proceedings of the Transportation Research Board Annual Meeting (16-2284), Washington, DC.
- Lavieri, P.S., Garikapati, V.M., Bhat, C.R., Pendyala, R.M., Astroza, S., Dias, F.F., 2017. Modeling individual preferences for ownership and sharing of autonomous vehicle technologies. *Transp. Res. Rec.* 2665 (1), 1–10.
- Lee, J., Kockelman, K.M., 2019. Development of traffic-based congestion pricing and its application to automated vehicles. *Transp. Res. Rec.* 2673 (6), 536–547.
- Lenz, B., Fraedrich, E.M., Cyganski, R., 2016. Riding in an autonomous car: What about the user perspective? Available from: http://elib.dlr.de/108174/1/Lenz_Fraedrich_Cyganski_HKSTS_2016_RIDING%20IN%20AN%20AUTONOMOUS%20CAR%20WHAT%20ABOUT%20THE%20USER%20PERSPECTIVE.pdf.
- Levin, M.W., Boyles, S.D., 2016. A multiclass cell transmission model for shared human and autonomous vehicle roads. *Transp. Res. C Emerg. Technol.* 62, 103–116.
- Li, T., Kockelman, K., 2018. Valuing the safety benefits of connected and automated vehicle technologies. In: Kara Kockelman, Stephen Boyles, (Eds.), *Smart Transport for Cities & Nations: The Rise of Self-Driving & Connected Vehicles*, Create Space, The University of Texas, Austin.
- Litman, T., 2019. Autonomous Vehicle Implementation Predictions: Implications for Transport Planning. Victoria Transport Policy Institute, Victoria, Canada. Available from: <https://www.vtpi.org/avip.pdf>.
- Liu, J., Kockelman, K.M., Boesch, P.M., Ciari, F., 2017. Tracking a system of shared autonomous vehicles across the Austin, Texas network using agent-based simulation. *Transportation* 44 (6), 1261–1278.
- Loeb, B., Kockelman, K.M., 2019. Fleet performance and cost evaluation of a shared autonomous electric vehicle (SAEV) fleet: a case study for Austin, Texas. *Transp. Res.* 121, 374–385.
- Loeb, B., Kockelman, K.M., Liu, J., 2018. Shared autonomous electric vehicle (SAEV) operations across the Austin, Texas network with charging infrastructure decisions. *Transp. Res.* 89, 222–233.
- Murray, C.C., Chu, A.G., 2015. The flying sidekick traveling salesman problem: optimization of drone-assisted parcel delivery. *Transp. Res.* 54, 86–109.
- Milanés, V., Shladover, S.E., 2014. Modeling cooperative and autonomous adaptive cruise control dynamic responses using experimental data. *Transp. Res. C Emerg. Technol.* 48, 285–300.
- Naus, G.J., Vugts, R.P., Ploeg, J., van de Molengraft, M.J., Steinbuch, M., 2010. String-stable CACC design and experimental validation: a frequency-domain approach. *IEEE Trans. Vehicul. Technol.* 59 (9), 4268–4279.
- Nazari, F., Norzoliaee, M., Mohammadian, A.K., 2018. Shared versus private mobility: modeling public interest in autonomous vehicles accounting for latent attitudes. *Transp. Res. C Emerg. Technol.* 97, 456–477.
- NHTSA, 2008. Traffic safety facts 2006. Report no. DOT HS 810 809. National Highway Traffic Safety Administration, Washington, DC. Available from: <http://www-nrd.nhtsa.dot.gov/Pubs/810809.pdf>.
- Otto, A., Agatz, N., Campbell, J., Golden, B., Pesch, E., 2018. Optimization approaches for civil applications of unmanned aerial vehicles (UAVs) or aerial drones: a survey. *Networks* 72 (4), 411–458.
- Pakusch, C., Stevens, G., Boden, A., Bossauer, P., 2018. Unintended effects of autonomous driving: a study on mobility preferences in the future. *Sustainability* 10 (7), 2404.
- Perrine, K., Kockelman, K., Huang, Y., 2017. Anticipating long-distance travel shifts due to self-driving vehicles. In: 97th Annual Meeting of the TRB.
- PostBus, 2019. Start of public testing of autonomous shuttles. <https://www.postauto.ch/en/news/start-public-testing-autonomous-shuttles>. Available from: <https://www.postauto.ch/en/news/start-public-testing-autonomous-shuttles>.
- Pultarova, T., 2016. Lyon Launches Fully Electric Autonomous Public Bus Service. E&T Engineering and Technology. Available from: <https://eandt.theiet.org/content/articles/2016/09/lyon-launches-fully-electric-autonomous-public-bus-service/>.
- Russell, J., 2017. China's Didi Chuxing Opens U.S. Lab to Develop AI and Self-driving Car Tech. Tech Crunch. Available from: <https://techcrunch.com/2017/03/08/didi-us-research-lab/>.
- Sharon, G., Levin, M.W., Hanna, J.P., Rambha, T., Boyles, S.D., Stone, P., 2017. Network-wide adaptive tolling for connected and automated vehicles. *Transp. Res.* 84, 142–157.
- Shladover, S.E., Su, D., Lu, X.-Y., 2012. Impacts of cooperative adaptive cruise control on freeway traffic flow. *Transp. Res. Rec.* 2324, 63–70.
- Shoup, D.C., 2006. Cruising for parking. *Trans. Policy* 13 (6), 479–486.
- Simoni, M.D., Kockelman, K.M., Gurumurthy, K.M., Bischoff, J., 2019. Congestion pricing in a world of self-driving vehicles: An analysis of different strategies in alternative future scenarios. *Transp. Res. C Emerg. Technol.* 98, 167–185.
- Singleton, P.A., 2019. Discussing the “positive utilities” of autonomous vehicles: Will travellers really use their time productively? *Trans. Rev.* 39 (1), 50–65.
- Stephens, T.S., Gonder, J., Chen, Y., Lin, Z., Liu, C., Gohlke, D., 2016. Estimated Bounds and Important Factors for Fuel Use and Consumer Costs of Connected and Automated Vehicles (No. NREL/TP-5400-67216). National Renewable Energy Lab (NREL), Golden, CO, United States.
- Stern, R.E., Cui, S., Delle Monache, M.L., Bhadani, R., Bunting, M., Churchill, M., Seibold, B., 2018. Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments. *Transp. Res. C Emerg. Technol.* 89, 205–221.
- Stocker, A., Shaheen, S., 2017. Shared automated vehicles: review of business models. International Transport Forum Discussion Paper.
- Stradling, R., 2019. WakeMed in Raleigh is the Spot as UPS Makes Its First Deliveries with Drones in the U.S. The News&Observer. Available from: <https://www.newsobserver.com/news/local/article228373214.html>.
- Systematics, Cambridge, 2005. Traffic congestion and reliability: Trends and advanced strategies for congestion mitigation. No. FHWA-HOP-05-064. Federal Highway Administration, United States.
- Talebpour, A., Mahmassani, H.S., 2016. Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transp. Res. C Emerg. Technol.* 71, 143–163.
- Tientrakool, P., Ho, Y.C., Maxemchuk, N.F., 2011. Highway capacity benefits from using vehicle-to-vehicle communication and sensors for collision avoidance. In: IEEE Vehicular Technology Conference (VTC Fall), IEEE, pp. 1–5.
- Tscharktschiew, S., Evangelinos, C., 2019. Pigouvian road congestion pricing under autonomous driving mode choice. *Transp. Res. C Emerg. Technol.* 101, 79–95.

- Tsugawa, S., Jeschke, S., Shladover, S.E., 2016. A review of truck platooning projects for energy savings. *IEEE Trans. Intell. Veh.* 1 (1), 68–77.
- Tsugawa, S., Kato, S., Aoki, K., 2011. An automated truck platoon for energy saving. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems* (4109–4114), IEEE.
- van Arem, B., Van Driel, C.J.G., Visser, R., 2006. The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Trans. Intell. Transp. Sys.* 7 (4), 429–436.
- van den Berg, V.A., Verhoef, E.T., 2016. Autonomous cars and dynamic bottleneck congestion: the effects on capacity, value of time and preference heterogeneity. *Transp. Res. B Methodol.* 94, 43–60.
- Wang, M., Daamen, W., Hoogendoorn, S.P., van Arem, B., 2014. Rolling horizon control framework for driver assistance systems. Part II: Cooperative sensing and cooperative control. *Transp. Res. C Emerg. Technol.* 40, 290–311.
- Wu, C., Bayen, A.M., Mehta, A., 2018. Stabilizing traffic with autonomous vehicles. In: *IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 1–7.
- Yap, M.D., Correia, G., Van Arem, B., 2016. Preferences of travellers for using automated vehicles as last mile public transport of multimodal train trips. *Transp. Res. A Policy Pract.* 94, 1–16.
- Zhang, W., Guhathakurta, S., Khalil, E.B., 2018. The impact of private autonomous vehicles on vehicle ownership and unoccupied VMT generation. *Transp. Res. C Emerg. Technol.* 90, 156–165.
- Zhao, Y., Kockelman, K., 2018. Anticipating the regional impacts of connected and automated vehicle travel in Austin, Texas. *J. Urban Plan. Dev.* 144 (4), 04018032.
- Zhou, M., Qu, X., Jin, S., 2017. On the impact of cooperative autonomous vehicles in improving freeway merging: a modified intelligent driver model-based approach. *IEEE Trans. Intell. Transp. Sys.* 18 (6), 1422–1428.

Further Reading

- Anderson, J.M., Nidhi, K., Stanley, K.D., Sorensen, P., Samaras, C., Oluwatola, O.A., 2014. *Autonomous Vehicle Technology: A guide for Policymakers*. RAND Corporation.
- Kockelman, K., Boyles, S., 2018. *Smart Transport for Cities & Nations: The Rise of Self-Driving & Connected Vehicles*. Amazon Create Space. Available from: http://www.caee.utexas.edu/prof/kockelman/public_html/CAV_Book2018.pdf.
- Milakis, D., Van Arem, B., Van Wee, B., 2017. Policy and society related implications of automated driving: a review of literature and directions for future research. *J. Intell. Transp. Sys.* 21 (4), 324–348.

Transport Modes and Tourism

Ila Maltese*, Luca Zamparini†, * Dipartimento di Scienze Politiche, Università degli Studi Roma3, Roma, Italy; † Dipartimento di Scienze Giuridiche, Università del Salento, Lecce, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	26
Historical Evolution of Tourism and Transport Modes	26
Transport and Tourism	27
Transport as Tourism	27
Transport for Tourism	27
Tourism for Transport	28
Conclusions	29
Biography	29
See Also	30
References	30
Further Reading	31

Introduction

The tourism industry, unlike other economic sectors, have proved to be particularly resilient to the various economic crises that have marked the last decades. Up to the rising of the COVID-19 pandemic in 2020, the sector, on a global scale, did not seem to have suffered drastic drops in flows, and showed trends of constant growth, also at an international level (UNWTO, 2018a, 2018b). It is questionable how long the effects of COVID-19 will characterize tourism. However, the upward long-term trend suggests the need and the opportunity to investigate the different elements of the phenomenon, focusing here, primarily, on the transport sector and its linkages with tourism.

This chapter is structured in two sections. The first one discusses the evolution of tourism and the consequent developments in the transport sector while the second one explores the complex relationships between transport and tourism. Some conclusions follow, mainly dealing with sustainability and methodological issues.

Historical Evolution of Tourism and Transport Modes

Transport toward a specific place has often been a crucial element of the accessibility of a tourist destination. Within this context, the historical evolution in the transport sector, especially in terms of innovation and technological development, by increasing the modal choice set, on one side, and by reducing costs (in terms of both time and tickets) and risks, on the other, has also scaled up the geographical tourism expansion, making it more and more a globalized phenomenon, affordable to ever largest shares of the population (Sorupia, 2005).

In this section, some brief notes (For a more in-depth and detailed historical description of the transport innovations that enhanced tourism improvement, see, for example, Gierczak, 2011) on tourism history will be the base to assess the mutual causality between transport and tourism.

The first tourists ever were travelers, prompted by curiosity toward unknown places, or actually discovering new trade routes, or driven by religious matters in pilgrimage; the transport modes at their disposal were mainly animals, like horses, provided that roads were built, for medium-long haul land traveling, or sailing ships, for maritime journeys, or feet, for short or spiritual trips.

Apart from commercial or religious matters, in the Grand Tour era (XVII-XIX century) people traveled throughout European cities in order to increase their culture and knowledge; due to the long time spent on traveling (more than a year), carriages for luggage and riches were necessary (Towner, 1985, 2002).

After the industrial revolution (second half of XVIII century) journeys became shorter, and transport sector worked hard to improve the quality of services and increase the speed of transport modes in order to cover space quickly: new roads, railroads and channels were built, while trains, stagecoaches and sailing ships kept on increasing the travel comfort.

Specifically, in the XIX century, the widespread use of the steam engine, invented by Watt in 1769, made rail and steamships very popular and cheap (Gierczak, 2011), while the invention of the gasoline engine and the consequent car diffusion in XX century has pushed motorized and flexible (door2door) tourism (Gierczak, 2011); air transport, operated by both scheduled and chartered airlines, also let tourists cover larger distances, making remote destinations more accessible.

Moreover, in more recent times (XXI century), with the introduction of low cost carriers, first reaching sunny coastal destinations, a simpler and more affordable air transport has definitely allowed many places to become a tourist destination. Furthermore, high-

speed rail (HSR) has come against air transport for medium-haul trips, due to the increased waiting times and inspections suffered by air travel, but also to the ever improved travel comfort in commercial rail journeys, while maritime trips have become more and more interesting as tourism itself (cruises) as much as a proper transport mode for getting from a place to another (hovercrafts for islands more than transatlantic lines); indeed, in this case, it is tourism that has pushed innovation into transport. Another example is the “roTEL”, a bus equipped for making travelers sleep onboard.

At the same time, the widespread concern for the environment and health has made active travel in tourism very popular, with many tourists choosing nonmotorized transport modes, such as walking (Hall et al., 2017) and cycling (den Hoed, 2019) not only at but also toward destination.

Finally, in the next future, space tourism could arise, given the right development of both transport vehicles and proper infrastructures.

Last but not least, the sudden growth of excursionists’ demand, together with the increased interest in transport within destination, call attention to local public transport modes, such as urban light rail, metro, bus and their infrastructure and services.

Interestingly, it must be noticed that nowadays tourists, traveling for both leisure or business, can choose among a large set of modes of transport (water, land, including rail and rail, or air), depending on their needs but also following their attitudes. Indeed, transport toward a destination or within destinations may also become an opportunity for “experiencing” the travel itself (e.g., sailing cruises), as well as for working (meeting area on the trains or planes) or enjoying a movie (boat cinema or devices at seat on HSR).

Transport and Tourism

It is widely acknowledged that transport and tourism are strictly interconnected (see, for example, Maltese, 2018; Zamparini and Maltese, 2021). Transport can normally be considered as one of the various complementary goods and services that characterize the tourism experience. There is clear evidence of a strong mutual causality. However, it is difficult to discriminate between tourists and resident users of the transport system. Therefore, the link between transport and tourism is very complex to analyze (Page and Ge, 2009). Consequently, most of the studies dealing with this connection end up privileging one of the two perspectives (tourism or transport) (Palhares, 2003). On the contrary, Page (2005) suggests that scholars should consider three main aspects of this link:

- the trip itself can become a tourism experience (transport as tourism);
- transport clearly encourages and facilitates tourism (transport for tourism);
- tourism promotes the development of the transport sector (tourism for transport);

In the following sections each of these issues will be analyzed and discussed.

Transport as Tourism

The first dimension that can be considered in the relationship between tourism and transport concerns the possibility that the trip is not functional to reach the tourist destination; it rather becomes a tourist experience itself.

In this case, the transport demand ceases to be a derived demand and provides tourists with the possibility of covering various ranges of distances. Long-distance trips are normally related to cruise ships (which have a diversified supply for all market segments) or to train journeys (e.g. the Indian Pacific line between Perth and Sidney in Australia or the famous Transiberian railway between Moscow and Vladivostok). Medium-haul trips are generally performed in camper vans, on trains, or on sailing boats. Short trips may also involve hikers or excursionists—for example on tourist buses, river boats, mountain trains or panoramic cable car or ropeway systems.

This category also includes some experiences on particular aircrafts (e.g., balloons, gliders or two-seater) and cycle tourism, as bikers or passengers, while in the last few years, space aircraft expensive “tours” have also been organized.

Moreover, the last decades have witnessed the diffusion of active travel and in particular of “walking” tourism. The latter can represent the individual tourist experience *per se*, in the medium-long distance, in the case of pilgrimage “walking”, or at the urban scale, to best “enjoy” the tourist destination. In this case, it is important that the considered town or neighborhood has a good “walkability” score, that can be measured by taking into account several variables such as: road safety, landscape aesthetics, the possibility of finding refreshment points and shelter (from the sun as rain), information poles (Le Pira et al., 2021).

Transport for Tourism

Transport is definitely a key component of the tourism product basket (Speakman, 2005), together with catering, accommodation, activities (such as visits to parks and museums, festivals, seminars and sport events) and shopping. Actually, the very definition of “tourism” (“Travelling to and staying in places outside one’s own usual environment for not more than one consecutive year for leisure and not less than 24 hours, business and other purposes” (WTO, 1995)), requires leisure or work

activities to be carried out in a place other than that of residence. Consequently, the need to go “elsewhere” (i.e., to cover the distance between the place of origin and the tourism destination) can only be satisfied by the use of any possible transport means, be it public or private.

Moreover, it could be worthwhile to distinguish between three different types of transport services that can be used by tourists, as they deal not only with different mobility patterns and different transport modes, but also with different aspects of the tourism experience.

Specifically, a tourist can make use of:

- Transport from the origin (e.g., place of residence) to the tourist destination and back;
- Transport within the tourist destination;
- Transport between multiple tourist destinations (during the same trip).

In the first case, which has attracted most of the literature about transport for tourism both at the domestic and at the international or intercontinental scale, accessibility to the tourist destination is the main theme of analysis. It heavily depends on a series of factors that include infrastructures, services (both unimodal and multimodal), a good degree of information and affordable tariffs (Van Wee et al., 2013). It has emerged that infrastructures and services play a momentous role to reach destinations that are particularly isolated or in any case distant from urban communities. In this context, low cost airlines have been able to expand their networks by including point to point connections to the most sought after tourist destinations.

More broadly speaking, the development of mass tourism has been determined by the conjoint effects of the evolution in the transport sector (that has reduced distances, times and costs, and has increased efficiency, travel comfort and service capacity) and of some socio-economic and regulatory phenomena (such as an increased leisure time availability and higher income levels, on one side, and the liberalization of passenger movement, on the other). This evolution has also attained the international scale and it has affected and has been affecting all the different modes of transport (Culpan, 1987; Duval, 2007).

Moving to the second point, when tourists choose a unique destination for their journey, it is important to consider transport within tourist destinations, given that it becomes a part of the tourism product offered by such destinations (Masiero & Zoltan, 2013; Prideaux, 2000). In these instances, that are mainly related to the urban scale, the local public transport (LPT from now on) plays a crucial role (Le-Klähn and Hall, 2015). It deals with several interesting aspects for tourism, such as: (1) the diffusion of the service that may allow to reach all the amenities that are present in the tourist destination; (2) the management of feeder services to and from the major transport hubs (airports, railway and maritime stations, etc.) in order to efficiently connect local transport with the national and international ones; (3) the proposal of specific offers or packages for tourists and the provision of tariff integration, possibly with some of the tourist activities that it is possible to perform at the tourist destination; (4) the management of the communication and information services, which may allow to digitally display maps, and pedestrian and cycle paths connected and adjacent to the LPT network. In the case in which the destination has a relevant share of foreign tourist demand, it is necessary to provide information in several different languages. A recent trend has been constituted by the substitution of LPT by active travel options (walking or cycling). This has been due to two main factors: (1) normally, the distances to be covered are relatively short and (2) the scarce attractiveness of the LPT, due to possible overcrowding or lower service regularity, as during and in the aftermath of the COVID-19 pandemic.

Finally, as concerns the third point, it may be taken into account that tourists normally want to make excursions to places that are in proximity to their main destination (Plog, 1974). In this case, agreements among several local administrations are necessary in order to improve the efficiency and effectiveness of interurban transport services (bus and railways, for the most, but also maritime routes) running from/connecting one location to another, to finance the upgrade of transport infrastructures, and to strengthen the degree of connectivity among the different locations especially in terms of joint planning, marketing, information and communication about tariffs, timetables and mobility options.

Tourism for Transport

If the contribution of transport to tourism is clear, the causal link between tourism and transport must certainly be better investigated. Indeed, the need to improve and specialize the transport service in order to cater for tourists' needs has led to numerous innovations and improvements in the transport sector, with many positive externalities for residents and city users (as commuters and elderly people). For example, new charter services have been set up based on the specific tourists demand, new routes have been induced by the development of a specific tourist destination, some lines and routes have been strengthened, through shuttle services and more generally by feeders that serve fundamental infrastructures (also) for tourism induced demand.

In some cases, infrastructures built from scratch or expanded compared to the existing ones and services that guarantee multi-modality are generated from the need to meet the tourists' demand. It then emerges that they have a momentous impact on the territorial and economic development (in terms of employment and of gross domestic product) of the region that encompasses the tourist destinations where these investments have been deployed. Moreover, some intangible aspects can also be induced by the tourism industry to the transport one, such as new marketing strategies, especially in terms of communication and digitization that

simplify the acquisition of information and the possibility of remote booking. Finally, as regards management, new services can be developed, such as offers based on demand segmentation and new tariff integration systems.

Conclusions

This chapter investigated the historical evolution of tourism and transport modes and the relationships between tourism and transport. The analytical framework for scrutinizing the mutual causality between transport and tourism has identified three different aspects: transport (toward or within destinations) for tourism, tourism for transport and tourism as transport, that can also occur simultaneously, depending on the considered transport mode.

Within this context, it can be concluded that the distinctions among transport modes and among the causality direction between transport and tourism are crucial for two main reasons.

The first one is methodological and concerns the (new) possibility to discriminate tourists from residents and city users. For example, transport toward destination, especially using those modes who need a personal identification of the travelers (airlines or HSR), is easier than walking within destination. Moreover, big data and technological devices (such a smartphone) could help in data collection for researchers (De Cantis et al., 2016; Dickinson et al., 2018; Lau and McKercher, 2006; Orellana et al., 2012; Xu et al., 2019; Zheng et al., 2017).

The second one regards the unavoidable sustainability issue concerning tourism transport (Hopkins, 2020). Taking into account that the transport mode towards the destination—and the motivations for choosing it—mostly affects the transport mode chosen at destination (Cohen and Kantenbacher, 2019; Gössling et al., 2019; Hergesell and Dickinger, 2013), it is firstly worthwhile noticing that an increased accessibility to some remote and isolated places can be harmful in terms of the derived negative externalities that may lead to long term problems for the destination region. Furthermore, also an over-crowded public vehicle (bus or local train) can be a negative externality for residents to minimize. Lastly, and more broadly speaking, it has been proved that environmental degradation reduces the destination attractiveness (Chen et al., 2017) and that transport is responsible for most of the negative externalities coming from tourism (Gössling, 2002). As a consequence, the concept of “sustainable tourism” should be enhanced (Becken, 2005; Boluk et al., 2017; Panzer-Krause, 2019; Peeters et al., 2019), making tourists aware of their impact even in terms of transport modal choice both toward and within destination(s) (Buijendijk et al., 2018; Maltese and Zamparini, 2021; McLennan et al., 2014), together with a suitable offer of public services and active travel infrastructures, adequately communicated, thus stimulating a more sustainable mobility (Banister, 2008; Barr and Prillwitz, 2012; Barr et al., 2010; Hall, 1999; Høyer, 2000; Prillwitz and Barr, 2011).

Biography



Ila Maltese holds a PhD in Urban Policies and Projects and is currently senior researcher at TRELab (Transport Research Lab) at the University of RomaTre in Rome. She also teaches Economics at the Faculties of “Political, Economic and Social Sciences” and “Humanities” at the University of Milan and at the Faculty of “Management and Production Engineering” at the Polytechnic of Milan. Her primary research interests include sustainable mobility at urban scale, both for passengers—tourists and residents—and freight, and smart city dimensions and drivers.



Luca Zamparini is an associate professor of Economics and currently chairing the courses of Economics and of Tourism Economics at the University of Salento (Italy). He is a coordinator of the cluster on Leisure, Recreation and Tourism for NECTAR (Network on European Communications and Transport Activity Research). His main fields of expertise are transport and tourism economics with an emphasis on the qualitative attributes of transport, on transport security, and on the relationships between transport and tourism development. He has edited various volumes, three special issues and has published many articles and chapters for international journals and books.

See Also

Mode Share/Choice Models; Urban Recreational Travel; Place Perception and Travel Behavior; Sustainable Mobility Paths; Transport Modes and Accessibility; Transport Modes and Globalization; Transport Modes and Cities; Transport Modes and Remote Areas; Active Transport: Heterogeneous Street Users Serving Movement and Place Functions; Electric vehicles; Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service; Adoption of New Travel Information Platforms; ICT and Transport Modes; The History of air Transport; Next Generation Travel: Young Adults' Travel Patterns; Airport and Airport Network Planning; Road Modes: Walking; Street Design For Active Travel; Transit Planning and Management; Car Ownership; Road Transport: E-Scooters; The History of Rail Transport; The History of Maritime Transport; Cruise Industry; Bus Route Planning; Space Transport; Bicycle as a Transportation Mode; Active Travel; Planning Tourism Travel; Walkability

References

- Banister, D., 2008. The sustainable mobility paradigm. *Transport Policy* 15 (2), 73–80. doi:10.1016/j.tranpol.2007.10.005.
- Barr, S., Prillwitz, J., 2012. Green travellers? Exploring the spatial context of sustainable mobility styles. *Appl. Geogr.* 32 (2), 798–809.
- Barr, S., Shaw, G., Coles, T., Prillwitz, J., 2010. A holiday is a holiday: Practicing sustainability, home and away. *J. Transport Geogr.* 18 (3), 474–481. doi:10.1016/j.jtrangeo.2009.08.007.
- Becken, S., 2005. Towards sustainable tourism transport: An analysis of coach tourism in New Zealand. *Tourism Geographies* 7 (1), 23–42. doi:10.1080/1461668042000324049.
- Boluk, K.A., Cavaliere, C.T., Higgins-Desbiolles, F., 2017. Critical thinking to realize sustainability in tourism systems: Reflecting on the 2030 sustainable development goals. *J. Sustain. Tourism* 25 (9), 1201–1204. doi:10.1080/09669582.2017.1333263.
- Buijtenijk, H., Blom, J., Vermeer, J., van der Duim, R., 2018. Eco-innovation for sustainable tourism transitions as a process of collaborative co-production: The case of a carbon management calculator for the Dutch travel industry. *J. Sustain. Tourism* 26 (7), 1222–1240. doi:10.1080/09669582.2018.1433184.
- Chen, C.-M., Lin, Y.-L., Hsu, C.-L., 2017. Does air pollution drive away tourists? A case study of the Sun Moon Lake National Scenic Area, Taiwan. *Transport. Res. Part D: Transport Environ.* 53, 398–402. doi:10.1016/j.trd.2017.04.028.
- Cohen, S.A., Kantenbacher, J., 2019. Flying less: personal health and environmental co-benefits. *J. Sustain. Tourism*, doi:10.1080/09669582.2019.1585442.
- Culpan, R., 1987. International tourism model for developing economies. *Ann. Tourism Res.* 14 (4), 541–552.
- De Cantis, S., Ferrante, M., Kahani, A., Shoval, N., 2016. Cruise passengers' behavior at the destination: Investigation using GPS technology. *Tourism Manage.* 52, 133–150.
- den Hoed, W., 2019. Where everyday mobility meets tourism: An age-friendly perspective on cycling in the Netherlands and the UK. *J. Sustain. Tourism* 28 (2), 185–203. doi:10.1080/09669582.2019.1656727.
- Dickinson, J.E., Filimonau, V., Cherrett, T., Davies, N., Hibbert, J.F., Norgate, S., Speed, C., 2018. Life-sharing using mobile apps in tourism: The role of trust, sense of community and existing lift-sharing practices. *Transport. Res. Part D: Transport Environ.* 61 (B), 397–405. doi:10.1016/j.trd.2017.11.004.
- Duval, D.T., 2007. *Tourism and Transport: Modes, Networks and Flows*. Channel View Publications, Clevedon, UK.
- Gierczak, B., 2011. The history of tourist transport after the modern industrial revolution. *Polish J. Sport Tourism* 18 (4), 275–281.
- Gossling, S., Hanna, P., Cohen, S., Higham, J.E.S., Hopkins, D., 2019. Can we fly less? An evaluation of the 'necessity' of air travel. *J. Air Transport Manage.* 81, 101722. doi:10.1016/j.jairtraman.2019.101722.
- Gössling, S., 2002. Global environmental consequences of tourism. *Global Environ. Change* 12, 283–302.
- Hall, D.R., 1999. Conceptualising tourism transport: Inequality and externality issues. *J. Transport Geogr.* 7 (3), 181–188. doi:10.1016/S0966-6923(99)00001-0.
- Hall, C.M., Ram, Y., Shoval, N. (Eds.), 2017. *The Routledge International handbook of walking*. 1st ed. Routledge, Abingdon, UK.
- Hergesell, A., Dickinger, A., 2013. Environmentally friendly holiday transport mode choices among students: the role of price, time and convenience. *J. Sustain. Tourism* 21 (4), 596–613.
- Hopkins, D., 2020. Sustainable mobility at the interface of transport and tourism. *J. Sustain. Tourism* 28 (2), 129–143.
- Høyer, K.G., 2000. Sustainable tourism or sustainable mobility? The Norwegian case. *J. Sustain. Tourism* 8 (2), 147–160. doi:10.1080/09669580008667354.
- Lau, G., McKercher, B., 2006. Understanding tourist movement patterns in a destination: A GIS approach. *Tourism Hospitality Res.* 7 (1), 39–49.
- Le-Klähn, D.-T., Hall, C.M., 2015. Tourist use of public transport at destinations – a review. *Curr. Issues Tourism* 18, 785803. doi:10.1080/13683500.2014.948812.
- Le Pira, M., Gatta, V., Marcucci, E., 2021. A stated preference survey to analyse tourist walking satisfaction. *J. Transport Geogr.* (forthcoming).
- Maltese, I., Zamparini, L., 2021. Mobility choices of young tourists within destinations: a survey of Northern Italian students. *J. Transport Geogr.* (forthcoming).
- Maltese, I., 2018. *Trasporti e Turismo*. In: Beria (Eds.), *Atlante tematico dei trasporti italiani*. Milano, Libreria Geografica, pp. 218–221.
- Masiero, L., Zoltan, J., 2013. Tourists intra-destination visits and transport mode: A bivariate probit model. *Ann. Tourism Res.* 43, 529–546.
- McLennan, C.J., Becken, S., Battye, R., So, K.F., 2014. Voluntary carbon offsetting: Who does it? *Tourism Manage.* 45, 194–198. doi:10.1016/j.tourman.2014.04.009.
- Orellana, D., Bregt, A.K., Ligtenberg, A., Wachowicz, M., 2012. Exploring visitor movement patterns in natural recreational areas. *Tourism Manage.* 33 (3), 672–682.
- Page, S., 2005. *Transport and Tourism: Global Perspectives*. Pearson education –Prentice Hall, Upper Saddle River, New Jersey.
- Page S., Ge Y.(G.), 2009. Transportation and Tourism: A Symbiotic Relationship?, https://www.researchgate.net/publication/285827144_Transportation_and_tourism_A_symbiotic_relationship [accessed Sep 21, 2018].
- Palhares, G.L., 2003. The role of transport in tourism development: Nodal functions and management practices. *Int. J. Tourism Res.* 5 (5), 403–407.
- Panzer-Krause, S., 2019. Networking towards sustainable tourism: Innovations between green growth and degrowth strategies. *Regional Stud.* 53 (7), 927–938. doi:10.1080/00343404.2018.1508873.
- Peeters, P., Higham, J.E.S., Cohen, S., Eijgelaar, E., Gössling, S., 2019. Desirable tourism transport futures. *J. Sustain. Tourism* 27 (2), 173–188. doi:10.1080/09669582.2018.1477785.
- Plog, S.C., 1974. Why destination areas rise and fall in popularity. *The Cornell Hotel and Restaurant Administration Quarterly* 14 (4), 55–58.
- Prideaux, B., 2000. The role of the transport system in destination development. *Tourism Manage.* 21 (1), 53–63.
- Prillwitz, J., Barr, S., 2011. Moving towards sustainability? Mobility styles, attitudes and individual travel behaviour. *J. Transport Geogr.* 19 (6), 1590–1600.
- Sorupia, E., 2005. Rethinking the role of transportation in tourism. *Proceedings of the Eastern Asia Society for Transportation Studies* 5, 1767–1777.
- Speakman, C., 2005. Tourism and transport: Future prospects. *Tourism Hospitality Plan. Develop.* 2 (2), 129–135. doi:10.1080/14790530500173647.
- Towner, J., 2002. Literature, Tourism and the Grand Tour. In: Robinson, M., Andersen, H.C. (Eds.), *Literature and Tourism*. Thomson Learning, London, UK.
- Towner, J., 1985. The grand tour: A key phase in the history of tourism. *Ann. Tourism Res.* 12 (3), 297–333. doi:10.1016/0160-7383(85)90002-7.
- UNWTO, 2018a. UNWTO World Tourism Barometer, Available from: <http://marketintelligence.unwto.org/content/unwto-world-tourism-barometer>, [accessed Sep 21, 2018].
- UNWTO, 2018b. *unwto-tourism-highlights-2018* Available from: <http://marketintelligence.unwto.org/publication/unwto-tourism-highlights-2018>, [accessed Sep 21, 2018].
- Van Wee, B., Geurs, K., Chorus, C., 2013. Information, communication, travel behavior and accessibility. *J. Transport Land Use* 6 (3), 1–16.
- Xu, F., Nash, N., Whitmarsh, L., 2019. Big data or small data? A methodological review of sustainable tourism. *J. Sustain. Tourism*, doi:10.1080/09669582.2019.1631318.

- WTO, 1995. Concepts, Definitions and Classifications for Tourism Statistics. World Tourism Organization.
- Zamparini, L., Maltese, I., 2021. Introduction. In: Zamparini, L. (Ed.), *Sustainable Transport and Tourism Destinations*. Emerald publishing, Bingley, UK, pp. 1–9.
- Zheng, W., Huang, X., Li, Y., 2017. Understanding the tourist mobility using GPS: where is the next place? *Tourism Manage.* 59, 267–280.

Further Reading

- Baranowski, S., 2007. Common ground: Linking transport and tourism history. *J. Transport Hist.* 28 (1), 120–124.
- Khadaroo, J., Seetanah, B., 2007. Transport infrastructure and tourism development. *Ann. Tourism Res.* 34 (4), 1021–1032.
- Lohmann, G., Pearce, D.G., 2012. Tourism and transport relationships: The suppliers' perspective in gateway destinations in New Zealand. *Asia Pacific J. Tourism Res.* 17 (1), 14–29.
- Lumsdon, L.M., Page, S.J. (Eds.), 2007. *Tourism and Transport*. Routledge, Abingdon, UK.
- Page, S., Connell, J., 2014. Transport and tourism. *The Wiley Blackwell Companion to Tourism* 155–167.
- Prideaux, B., 2001. Links between transport and tourism—past, present and future. *Tourism in the twenty-first century: reflections on experience* 91–109.
- Prideaux, B., 2007. Transport and destination development. In: Lumsdon, L.M., Page, S.J. (Eds.), *Tourism and Transport*. Routledge, Abingdon, UK, pp. 94–107.

Transport Modes and Accessibility

Bert van Wee, Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

The Concept of Accessibility	32
Definition and Components	32
ICT and Accessibility	34
Challenges for Future Research	34
Transport Modes, Travel, and Activities-Related Challenges	34
ICT	35
Short Distances and Slow Modes	35
Multiple Modes	35
Robustness	35
Perception of Accessibility	35
Air Transport	35
Goods Transport	35
Transport Innovations	36
Developing Countries	36
Evaluation-Related Challenges	36
The Integration of Accessibility in Wider Evaluations	36
The Logsum as an Accessibility Indicator	36
Distribution Effects, Equity, and Social Exclusion	36
Option Value	36
Freedom of Choice	36
Comparisons of Indicators for Real-World Cases	37
The Wishes of the Client	37
Concluding Remarks	37
Biography	37
References	37
Further Reading	37

The Concept of Accessibility

Accessibility is a, if not: the, core concept in transport geography and transport planning. The transport system mainly aims to provide access to destinations of locations where people want to carry out activities such as living, working, education, shopping, recreating, and visiting other people. In case of goods, transport accessibility aims to allow companies to transport goods in different stages of production (from raw materials), via several intermediate stages to final (consumer) products, and (lastly waste materials) between locations (Geurs and van Wee, 2004). Briefly summarizing, it expresses the potential for interaction (Hansen, 1959; El-Geneidy and Levinson, 2011). Consequently, the main benefits of improvements in the transport system follow from providing access to destinations. Therefore, accessibility plays a dominant role in about all transport policy documents. But what is accessibility, how can it be defined, and which components influence final levels of accessibility? What is the state of knowledge, and which open challenges remain? This paper aims to answer these questions.

A main message is that multiple categories of indicators measuring accessibility exist, and that the choice between all those options depends on the purpose and context of the study. In addition, many research challenges could be a source of inspiration for researchers.

Definition and Components

The most downloaded paper in the *Journal of Transport Geography* is the paper on accessibility written by Geurs and van Wee (2004, p. 128) who define accessibility as “the extent to which land-use and transport systems enable (groups of) individuals to reach activities or destinations by means of a (combination of) transport mode(s).” Fig. 1 shows a conceptual model explaining which factors influence levels of accessibility.

In line with Fig. 1, accessibility levels depend on four components:

- The land-use component (visualized as activity locations): the land-use system describing what is located where (e.g., jobs, other people, shops, health services, schools, and recreation facilities).
- The transport component (travel resistance): characteristics of the transport system determining how much time, money, and effort it takes to travel between locations.
- The individual component (wants, needs, abilities, opportunities, etc.) expressing what people need or want, and are able to do.
- The temporal component, expressing that it matters at which time which (locations of) activities are available. For example, shops and kindergartens have specific opening times.

The temporal component is not explicitly included in Fig. 1, but applies to all the other three components: people might have time restrictions, and activities transport services might be available during a selection of hours.

Fig. 1 shows that accessibility levels depend on the locations of activities, travel resistance, expressed in terms of time, costs, and effort, and the wants and needs of people. And those factors influence each other. As recognized by the so-called land-use transport interaction (models), land use influences the transport system and vice versa. For example, if residential areas are built adjacent to a railway line, a new station may be built along that line. And vice versa: locations near stations or motorway entrances and exits are relatively attractive for urbanization. The interactions between land use and transport also play a role at the level of individuals and households. For example, if new infrastructure is build, it reduces transport resistance, and that reduced resistance can make people change their residential location or destination choice. Characteristics of the transport system might influence people's wants and needs, for example, preferences to own vehicles, and the wants and needs of people might influence their residential location, and vice versa. Opening hours of locations that attract many people might influence timetables of a public transport system, and vice versa.

Almost all accessibility measures make use of one or multiple of the four components. Depending on the selection of components, these measures can be clustered in families of measures (Geurs and van Wee, 2004):

- Infrastructure-based measures, focusing on the level of service of the transport system. Examples are travel times to destinations, congestion levels on roads, and the number of stations accessible within a time or distance threshold (transport component).
- Location-based measures (also called place-based measures), focusing on access to spatially distributed activities. Examples are contour measures such as the number of jobs accessible from residential locations within a certain time threshold value. More advanced examples correct for distance or time using decay functions expressing that locations further away are valued less (transport and land-use component).
- Person-based measures, focusing on the individual level, such as the activities available for a person within time or other constraints. These measures originate from time, or time-space geography (transport, land use, individual, and temporal component) (Kwan, 1998; Neutens, 2010).
- Utility-based measures, expressing the utility that people address to the options of having access to spatially distributed activities, the so-called logsum being an important example. The logsum is the log of the denominator of a so-called logit model. This denominator includes the choice options available (transport, land use, and individual component; temporal component sometimes via time of day, but only partly).

The use of one or multiple accessibility indicators should primarily be based on the aim of the research and the options available. More specifically for a specific purpose the choice can be based on next criteria:

- The quality of the theoretical underpinnings

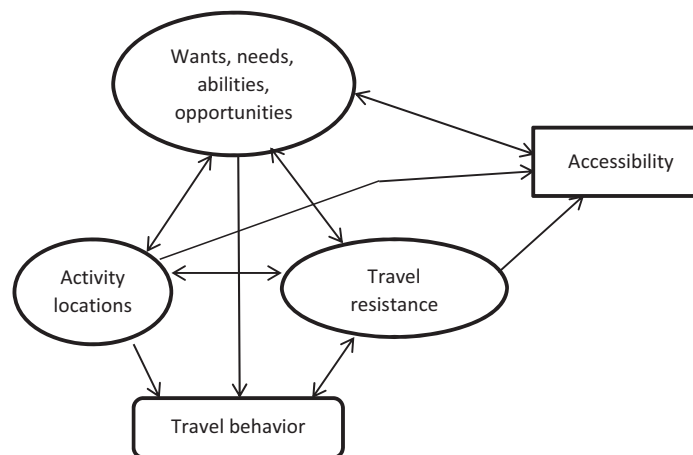


Figure 1 A conceptual model for accessibility, activity locations, and travel behavior. Source: van Wee and Geurs (2016).

- The ease of operationalizing the indicators (availability of data, models, software, etc.)
- The ease of interpreting the indicators and communication results to clients of accessibility research
- The ease of including the indicators in evaluation frameworks like cost–benefit analyses (CBA) or multi-criteria analyses (MCA)

A dilemma exists because generally speaking easy to calculate and next communicate indicators (such as the number of jobs accessible within 45 minutes, or average speeds on the road or rail system) is theoretically less sound and can hardly be included in evaluation frameworks. On the contrary, the utility-based measures generally have theoretically sound underpinnings, but are difficult to calculate, certainly if a logit models-based transport model is not available, and communicate to many clients of research such as policy-makers (Geurs and van Wee, 2004).

ICT and Accessibility

Most of the accessibility literature only considers physical access to destinations. But increasingly information and communication technologies (ICT) provide alternatives, examples being skypeing for personal communication, e-shopping, e-working, and e-learning. So, to fully understand accessibility, in the future researchers and policy-makers should take notice of the relationships between ICT and accessibility. These relationships are quite complex, as visualized by Fig. 2.

Some examples of interactions are: because of the option to e-shops shops might relocate. Or congestion levels decrease if more people can e-work and therefore avoid the rush hours. It can also influence the wants and needs of people: they might consider access to shops or jobs less important if ICT-based alternatives are available. And people might be more inclined to multitask thanks to ICT, performing ICT-based activities while traveling by train being an example, and the option to multitask while traveling might reduce the willingness of train travelers to pay for shorter travel times. Therefore, the valuation of characteristics of the transport system might also change as a consequence. ICT might also relief time constraints, such as due to limited opening hours of shops. Not only does ICT provide alternatives for physical activities, it also can generate activities. For example, people might have met electronically and after some contacts they decide to meet in person. And due to ICT people might reorganize their activities over time and space. Next, ICT can also influence travel mode, departure time, and route choice, for example, because of the availability of real-time travel information.

Challenges for Future Research

Transport Modes, Travel, and Activities-Related Challenges

van Wee (2016) suggests several challenges for future research. The first category of challenges is content related; in other words, they focus on transport modes and activities.

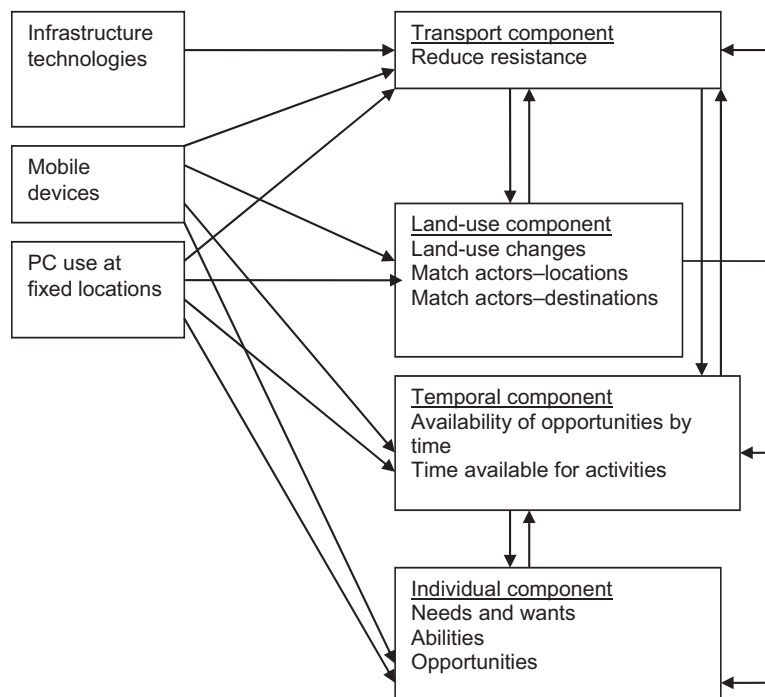


Figure 2 Relationships between ICT and accessibility components. Source: van Wee et al. (2013).

ICT

The importance of ICT for accessibility is explained earlier. ICT can influence accessibility in many complex ways, as illustrated in Fig. 2. The accessibility literature so far has not, in a mature way, discussed how to include these links between ICT and accessibility in accessibility research. Such research should also include the dynamics in ICT options being available. One can think of advanced options to combine modes via Mobility as Service (MAAS) (information, ticketing, etc.), car and bike sharing, e-learning and e-shopping options, etc.

In addition to including ICT in indicators for accessibility, this process will also generate new (big) data for accessibility analyses, via GPS-based data track and log systems, social media, and traffic flow-related data systems. A challenge is the assessment of the representativeness of these data because the use of related systems (e.g., social media) is not equally spread amongst the population. And early adopters of innovations might in some respects differ from the wider audience.

Short Distances and Slow Modes

The emphasis in the accessibility literature has largely been on longer distances and on car and public transport, often related to job accessibility. Active modes (walking and cycling) and nearby destinations have received much less attention, but can be very important for people, especially for having access to grocery shops, medical and other services, kindergartens, etc. Research in this area should include the specific context, for example, with respect to the quality of infrastructure for walking and cycling, and climate conditions.

Multiple Modes

People often combine modes to reach destinations, but the literature on multimodal accessibility is relatively underdeveloped. Furthermore, the fact that people can travel to destinations via multiple unimodal trips (e.g., either drive or cycle, or travel by public transport) is also often ignored in accessibility analyses. Additionally, for international trips people might positively value having available both aircraft and high-speed train options. The level of long-distance accessibility levels via these modes also depends on options to travel to and from airports and HSR station.

Robustness

Since the early 2000s academic literature has recognized the importance of related concepts like robustness, vulnerability, resilience, reliability, flexibility, and travel time variability. Using the term “robustness” as an umbrella term, it is important to recognize the different time scales at which robustness plays a role, ranging from short-term disruptions while traveling to several decades due to the impact of climate change on the transport system. And the relevance also differs strongly ranging from minor disruptions to disasters-related evacuations. A general challenge is to include robustness in accessibility indicators and next analyses.

Perception of Accessibility

Researchers generally calculate accessibility levels based on locations of opportunities, and characteristics of the transport system, and in some cases time constraint and personal characteristics (like car availability). But perceived accessibility, typically used when making choices, is seldom studied. Perceived accessibility can differ from calculated levels, for several reasons, the first example being that people might have misperceptions about the quality of nonused modes or opportunities. It can also be that they include additional characteristics of travel options, such as the attractiveness of the route or perceived social safety, or additional characteristics of opportunities, such as price and service levels of shops. The concept of perceived accessibility so far has not received a lot of attention, but for many people it could be more important than calculated levels of accessibility.

On a side note: in some literature perceived accessibility is sometime labeled as “subjective accessibility,” and calculated accessibility as “objective accessibility.” The terms “objective” and “subjective” are misused here, because measured accessibility does not perfectly reflect all characteristics of travel and opportunity characteristics, and subjective accessibility might include aspects that can objectively be assessed.

Air Transport

Overland transport (car, public transport, and, to a lesser extent, active modes) has dominated the accessibility literature. Long-distance transport, mainly air transport but also high-speed rail, has received less attention. But because of the rapid growth in long distance, transport accessibility by aircraft and high-speed rail is becoming increasingly important. In that area it is not only the direct and indirect (including stopovers) that matter, but also the frequencies, timetables, and prices. And it is important to include that frequencies are subject to diminishing returns: the higher the frequency, the lower the added value of a further increase in the frequency.

Goods Transport

Goods transport is an underexplored area in the accessibility research. It is way less studied in general, and also the importance of the concept, the decomposition between the accessibility components, and the choice of indicators are poorly understood, making it a challenging topic for research. Research in this realm should at least consider that goods do not “experience” travel, unlike people.

Transport Innovations

The impact of innovations in the transport system (and ICT—see earlier) is poorly understood. Examples of innovations in the transport system include autonomous vehicles, MaaS, e-bikes, electric vehicles, and improved information systems such as those forecasting travel times. Note that ICT also contributes to MaaS and forecasting travel times, and other ICT innovations can contribute to improved options to e-learning, e-shopping, and e-working.

Developing Countries

Most applications of accessibility literature focus on developed countries. Many challenges for accessibility analyses of developing countries exist especially due to the lack of available data. Maybe in the future this might change to some extent, because of the availability of big data for travel and activities of people. Another challenge could be that context matters. For example, the social status of modes can differ between countries, as well as preferences for activities (informal jobs, social ties, etc.) and social safety while traveling.

Evaluation-Related Challenges

The next category of challenges relates to the evaluation of accessibility levels.

The Integration of Accessibility in Wider Evaluations

Accessibility is not the only relevant topic for evaluations of (policies for) the transport and land-use system. An important question therefore is: how to include accessibility analyses in wider evaluation frameworks like CBA and MCA? CBA is the most common evaluation framework for at least transport infrastructure projects, but it is also often applied for other topics such as road pricing options. Because a CBA expresses anything that people value in monetary terms, accessibility indicators that express accessibility in monetary terms can relatively easily be included in CBAs. In all other cases, a major question is how to include accessibility analyses in CBAs.

In case of an MCA, results of accessibility can relatively easily be included. The main challenge is to come to adequate weight factors, relative to other important impacts of the transport system, like safety, several environmental impacts, investment, and maintenance costs.

The Logsum as an Accessibility Indicator

The logsum is an indicator that—under specific assumptions—can be used to assess the monetary value of accessibility levels. It is the log of the denominator of a discrete choice model, and that denominator includes travel options to opportunities.

But there are several challenges to further improve the logsum as an accessibility indicator. It ignores the fact that people value having multiple-choice options available, and—related—the love for variety in consumption. And it estimates the value of the decision, not necessarily the experienced utility. And the indicator is difficult to explain to nonexperts. Finally, the so-called IIA assumption of logit models can be violated due to autocorrelation, in case of destination choice.

Distribution Effects, Equity, and Social Exclusion

Accessibility indicators and accessibility studies generally ignore distribution effects and other equity-related issues (e.g., the fact that promises are made to a city or region), and levels of social exclusion. Since about 2010 ([Martens, 2017](#)), these topics have increasingly received attention of researchers, and consequently the inclusion of such topics in accessibility analyses will probably become more important in the future.

The challenge is not so much the development of indicators to study equity aspects of accessibility, because it can make use of indicators that are calculated anyway. And next these can be used to make explicit, for example, the distribution of accessibility levels over income groups or areas. A more daunting challenge is to include the indicators in broader evaluation frameworks, especially CBA. Another challenge is the underpinning of such indicators by ethical theories, although in recent years substantial progress has been made, dominant theories used being sufficientarianism, egalitarianism, and the capabilities approach. Another area for future research is the underpinning of quantitative yardsticks to label people of being socially excluded.

Option Value

People value available options, even if they do not use them. This is referred to in the (economic) literature as the option value ([Geurs et al., 2006](#)). For example, people might value of being able to travel by train, for the rare case of a roadblock, an event that might even never occur for them. The concept of the option value has seldom been included in the accessibility literature.

Freedom of Choice

People value being able to make choices, as expressed by the so-called concept of the “freedom of choice.” So far the accessibility literature has hardly included this concept. It is partly related to the so-called option value (see earlier) but that concept only relates to the availability of options, not to the freedom to choose between options.

Comparisons of Indicators for Real-World Cases

As explained earlier, many accessibility indicators exist, and choosing the “right” indicator for a specific study is not always easy because each indicator has its pros and cons. As explained earlier, there is a trade-off between the ease of communication and scientific quality. Comparing indicators for a specific application can be very relevant, first because it allows the researcher to find out how robust results are for indicator choices. Secondly, if complex indicators would lead to about the same conclusions as simple indicators, it could be an option to communicate the simple indicators to policy-makers and other clients of research, knowing that the results will be about the same if more difficult to interpret indicators would have been used.

The Wishes of the Client

Research studying what the clients of accessibility research need has hardly been done. In other words, we poorly understand the preferences of clients of accessibility analyses. Because a lot of accessibility analyses aims to support decision-making processes, it is an important avenue for future researchers to study those preference.

Concluding Remarks

Despite the fact that it is more than half a century ago that the seminal paper of Hansen about accessibility was published, the concept is still evolving, and there still are many definitions and indicators. In addition, several research challenges remain, not only because of weaknesses in current research, but also because society is changing in many ways, the impact of ICT being one important factor. This makes accessibility research an inspiring category of research, both for academics and for practice.

Biography



Bert van Wee is professor in Transport Policy at Delft University of Technology, the Netherlands, faculty Technology, Policy and Management. In addition, he is scientific director of TRAIL research school. His main interests are in long-term developments in transport, in particular in the areas of accessibility, land-use transport interaction, (evaluation of) large infrastructure projects, the environment, safety, policy analyses, and ethics.

References

- El-Geneidy, A., Levinson, D., 2011. Place rank: valuing spatial interactions. *Netw. Spat. Econ.* 11, 643–659.
- Geurs, K.T., van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. *J. Transp. Geogr.* 12, 127–140.
- Geurs, K., Haaijer, R., van Wee, B., 2006. Option value of public transport: methodology for measurement and case study for regional rail links in the Netherlands. *Transp. Rev.* 26, 613–643.
- Hansen, W.G., 1959. How accessibility shapes land use. *J. Am. Inst. Plan.* 25, 73–76.
- Kwan, M.-P., 1998. Space-time and integral measures of individual accessibility: a comparative analysis using a point-based framework. *Geogr. Anal.* 30, 191–216.
- Martens, K., 2017. *Transport Justice: Designing Fair Transportation Systems*. Routledge, New York/London.
- Neutens, T., 2010. *Space, Time and Accessibility—Analyzing Human Activities and Travel Possibilities from a Time-Geographic Perspective*. Department of Geography, University of Gent, Gent, Belgium.
- van Wee, B., 2016. Accessible accessibility research challenges. *J. Transp. Geogr.* 51, 9–16.
- van Wee, B., Geurs, K., 2016. The role of accessibility in urban and transport planning. In: Bliemer, M.C.J., Mulley, C., Mouton, C.J. (Eds.), *Handbook on Transport and Urban Planning in the Developed World*. Edward Elgar, Cheltenham.
- van Wee, B., Geurs, K., Chorus, C., 2013. Information, communication, travel behavior and accessibility. *J. Transp. Land Use* 6, 1–16.

Further Reading

- Handy, S.L., Niemeier, D.A., 1997. Measuring accessibility: an exploration of issues and alternatives. *Environ. Plan. A* 29, 1175–1194.
- Páez, A., Scott, D.M., Morency, C., 2012. Measuring accessibility: positive and normative implementations of various accessibility indicators. *J. Transp. Geogr.* 25, 141–153.
- Witlox, F., 2015. Beyond the data smog? *Transp. Rev.* 35, 245–249.

Transport Modes and Globalization

Jean-Paul Rodrigue, Department of Global Studies and Geography, Hofstra University, Hempstead, NY, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction: Modes as the Linchpin of the Global Economy	38
The Emergence of a Global Transportation System	38
Connecting Modes: Containerization and Intermodal Transportation	39
Transportation Modes and International Mobility	40
Maritime Transport	40
Air Transport	41
Logistics: Producing, Organizing, and Managing Transportation Modes in a Global Economy	43
Conclusion: Transportation Modes in a Globalized Economy	43
See Also	43
References	43

Introduction: Modes as the Linchpin of the Global Economy

Transportation modes are the conveyances used to support the mobility of passengers and freight (Sultana and Weber, 2017). The substantial diversity, availability, and affordability of goods in the global economy much depend on the capacity to transport them. For instance, a mobile phone manufactured (assembled) in China embarks in a complex journey, involving a multitude of stages with transport modes such as trucks, containerships and trains, and with transport facilities such as ports, railyards, and distribution centers, to ensure that it reaches global markets. This does not even consider the staggering array of movements related to the components of this device. For passengers, increasing international interactions linked with social interactions, business transactions, tourism, sport events and migration has been an ongoing trend. The mobility of passengers and freight within the global economy must thus be supported by transportation modes, terminals, and infrastructures (Black, 2003).

This chapter is organized as follows: The first section provides an historical perspective behind the emergence of major transportation modes and their impacts on globalization. The second section underlines the benefits of intermodalism. The third section looks that two fundamental modes supporting international mobility, maritime, and air transport. The fourth section focuses about how logistics allows for the organization of modes supporting global supply chains.

The Emergence of a Global Transportation System

International transportation predates globalization since for as long that there has been trade, transportation modes have been its conveyances. One is the prerequisite for the other; they are both mutually interdependent. The history of transportation modes is a proxy of the history of globalization with the volume, capacity, speed, and efficiency of transport modes evolving (Bernstein, 2008). As economies and societies emerged, axis of trade and circulation came into existence. One of the first major “international” trade route was the Silk Road, its name taken from the prized Chinese textile that flowed from Asia to the Middle East and Europe. The Silk Road consisted of a succession of trails followed by caravans through Central Asia. Since the transport capacity was limited, taking place over long distance and often unsafe, luxury goods were the only commodities that could be traded. The Silk Road also helped the diffusion of ideas and religions (initially Buddhism and then Islam), enabling civilizations from Europe, the Middle East, and Asia to interact. The road endured from the 2nd century BC to the 15th century when more efficient maritime transportation helped its downfall. A rudimentary system of international trade was thus permitted.

The transport system of the Roman Empire put in place between the 3rd century BC and the 2nd century AD was a fundamental component of a Mediterranean system of trade with two interdependent transport systems; the road and short distance, coastal, maritime shipping. The Mediterranean Ocean provided a central role to support trade between the major cities of Empire, most of them being located along the coast (e.g., Rome, Constantinople, Alexandria, Cartage). These cities were serviced by a road network permitting trade within their respective hinterlands. For instance, the world’s first dual carriageway, Via Portuensis, was built to link Rome and the port of Ostia at the mouth of the Tiber River. At the peak of the Roman Empire around 200 CE, its road system covered about 80,000 kilometers, but once the empire collapsed in the 5th century, the road system fell into disrepair and became fragmented.

It is, however, with long distance maritime transportation that more globally oriented economic systems were permitted, with the creation of commercial empires. Initially, ships were propelled by rowers, and sails were added around 2,500 BC as a complementary form of propulsion. China was one of the first to establish an important fluvial transport network with several

artificial canals connected to form the Grand Canal. At its peak in the 15th century, the canal system totaled 2500 kilometers, with some parts still being used today. By Medieval times, an extensive maritime trade network, the highways of the time, centered along the navigable rivers, canals, and coastal waters of Europe was established. Shipping was extensive and sophisticated using the English Channel, the North Sea, the Baltic and the Mediterranean where the most important cities were coastal or inland ports. Still, transportation was mainly a short distance and very risky endeavor with long distance maritime routes mainly controlled by Arab merchants linking the Middle East, South Asia, and Southeast Asia.

By the 14th century galleys were finally replaced by full-fledged sailships (the caravel and then the galleon) that were faster and required smaller crews. In total, 1431 marked the beginning of the European expansion with the discovery by the Portuguese of the North Atlantic circular wind pattern, better known as the trade winds. A similar pattern was also found on the Indian and Pacific oceans with the monsoon winds (long discovered by the Arabs). The ensuing era of colonialism was mainly the era of the sailship, linking Europe with the rest of the world. Although shipping services were rather sporadic, large colonial empires such as those of the Spanish, British, French, and Dutch were the early expression of a global economy supported by the sailship. For instance, the Dutch East Asian Company, founded in 1602, can be considered as one of the first multinational corporations. By 1750, it employed around 25,000 people and did business in more than 10 countries, mainly from the Dutch colony of Indonesia.

The beginning of the 19th century saw the establishment of the first regular maritime routes linking port cities worldwide, especially over the North Atlantic between Europe and North America. Sailships became increasingly efficient, with some like the Clipper ships mainly designed for speed more than for cargo holding. These “express” cargo ships dominated long distance ocean trade until the late 1850s when they were gradually replaced by steamships. Another significant improvement resided in the elaboration of accurate navigation charts where prevailing winds and sea currents could be used to the advantage of navigation. The improvement of steam engine technology permitted longer and safer voyages. The first regular services for transatlantic passenger transport by steamships were inaugurated in 1838 and until the mid-20th century, Liners accounted for most of international passenger travel. In the 1840s, it took about 12 days for a Liner to cross the Atlantic, a figure that dropped to 4 days in the 1930s.

By the end of the 19th century additional improvements in engine propulsion technology and a gradual shift from coal to oil as a fuel increased the speed and the capacity of maritime transport. Energy consumption by maritime shipping was reduced, and coal refueling stages were bypassed. An equal size oil-powered ship could transport more freight than a coal-powered ship, reducing operation costs considerably and extending its range. Global maritime circulation was also dramatically improved when shortcuts such as the Suez (1869) and the Panama (1914) canals, were constructed. With the Suez Canal, the far reaches of Asia and Australia became more accessible, while the Panama Canal considerably reduced the time it took to link the Atlantic to the Pacific.

The capacity of maritime shipping also increased with the specialization of ships, which enabled to move low-cost bulk commodities such as minerals and grain over long distances. The size of oil tankers grew substantially, especially after World War II when global energy demands surged. Maritime routes were thus expanded to include tanker routes, notably from the Middle East, the dominant global producer of oil. To this day, energy trades are among the foremost forms of modal globalization as they concern large quantities of goods usually carried over long distances.

The airline industry has also played a significant role in supporting the emergence of a global economy. It began with air mail services since they initially proved to be more profitable than transporting passengers. 1919 marked the first commercial air transport service between England and France, but air transport suffered from limitations in terms of capacity and range. The late 1920s and 1930s saw the expansion of regional and national air transport services in Europe and the United States with successful propeller aircrafts such as the Douglas DC-3. The post-World War II period was, however, the turning point for air transportation as the range, capacity, and speed of aircrafts increased. A growing number of people were able to afford the speed and convenience of air transportation. In 1958, the first commercial jet plane, the Boeing 707, entered in service and revolutionized international movements of passengers, marking the end of liner ships, leaving the maritime passenger sector to the niche markets of cruises and ferries. The cruise ship became a new mode of globalization in the tourism industry, offering itineraries in a variety of regional markets. The availability of the Boeing 747 in the early 1970s truly made air transport a global industry. Still, passenger transportation is a marginal, albeit visible, component of globalization.

By the second half of the 20th century, many international transportation systems were in place, forming the basis of a global transport system and reinforced by a global telecommunication network. Fundamental changes were about to take place as the role of transportation evolved. From a situation where transportation was a simple infrastructure permitting and supporting trade and mobility, transportation became a factor shaping global production and markets. The fragmentation of the production and an international division of labor increased the quantity of freight in circulation. Containers have been particularly important in this regard. Introduced in the late 1950s, containers became the main agents of the modern international transport system.

Connecting Modes: Containerization and Intermodal Transportation

While modal transportation is the mean supporting global mobility, intermodalism allows for the connectivity within and between large systems of circulation (Rodrigue, 2017). The container epitomizes this connectivity in the freight sector as a load unit that can

be used by several transport modes. Maritime, railway, and road modes are all able to handle containers, increasing the flexibility of freight transport, mainly by reducing transshipment costs and delays. Containerization underlines a growing relationship between freight transportation modes and a standardization of loads. The container has a reference size of 20 feet long, 8 feet high, and 8 feet wide, corresponding to a 20-foot equivalent unit. The most common containers are 40 feet in length, permitting them to fit well on a ship, truck, or railcar.

By using respective modes in the most productive manner, the line-haul economies of maritime shipping and rail may be used for long distances, with the efficiencies of trucks providing local distribution. The entire transport sequence is now seen as a whole, rather than as a series of stages, each marked by an individual operation with separate sets of documentation and rates. Containers have become the most important component for rail and maritime intermodal transportation, enabling them to interact closely and conferring significant time and cost savings.

Most major ports have been transformed to become container ports with the construction of new container transshipment infrastructures. These infrastructures accommodate containerships, ships built solely for the purpose of carrying stacked containers. For rail, new facilities were also required to handle container traffic as well as specialized railcars. The double stacking of containers on railways has doubled the capacity of trains to haul freight with minimal cost increases, thereby improving the competitive position of railways with regards to trucking for long distance shipments. Road transportation has also adapted to containerization, but the requirements are minimal with chassis able to latch in containers.

Intermodal transport is transforming a growing share of freight distribution across the globe. Large integrated transport carriers provide door-to-door services through a sequence of modes, terminals, and distribution centers. Transportation, in terms of modes and routing, is no longer of much concern for customers, if shipments reach their destinations within an expected cost and time range. The concerns are therefore mainly with cost and level of service. For the customer of intermodal transport services, transportation, and distance appear to be meaningless, but for the intermodal providers routing, costs, and service frequencies are important. Therefore intermodalism and globalization makes specific transport modes less visible to the users but requires move visibility for modal operators to effectively manage.

Transportation Modes and International Mobility

International transportation systems have been under increasing pressures to support the growing demands of international trade and the globalization of production and consumption (Dicken, 2015). This could not have occurred without considerable technical improvements permitting to transport larger quantities of freight and people, and this more quickly and more efficiently. Since containers and intermodal transportation improve the efficiency of global distribution, a growing share of general cargo moving globally is containerized. Transportation is often referred as an enabling factor that is not necessarily the cause of international trade, but a mean over which globalization could not have occurred without.

Because of the geographical scale of the global economy, most international freight flows circulate over several modes. Transport chains must be established to service these requirements, which reinforce the importance of transportation modes and terminals at strategic locations. International trade requires distribution infrastructures that can support its volume and extent. Two transportation modes are specifically supporting globalization and international trade; maritime and air transportation. Road and railways tend to account for a marginal share of international transportation since they are above all modes for national or regional transport services. However, a substantial share of the trade between Canada, United States, and Mexico is supported by trucking, as well as large share of the Western European trade. China with its Belt and Road Initiative has also spearheaded long distance rail transport services across Eurasia.

Maritime Transport

The growth of maritime transportation has been fueled by many issues. First, there is an increase in energy and mineral cargoes derived from a growing demand of developing economies. Second, global value chains have resulted in additional demands for long distance trade. Third, improvements in ship and maritime terminals have facilitated the flows of freight, particularly containerized traffic (Stopford, 2009).

Weight-wise, about 80% of the world trade is carried by maritime transportation, which account for 70% of its value. About half of this trade is handled by large containerships linking producers and consumers along sea lanes. This transportation system is organized around major maritime transport gateways where continental trade converges, namely Shanghai, Hong Kong, Singapore and Busan for Pacific Asia, Rotterdam and Antwerp for Europe and Los Angeles/Long Beach and New York for North America, among the largest container ports in the world (Fig. 1).

Economic development in East Asia has been the dominant factor behind the growth and the changes of international transportation in recent years. Since the distances involved are often considerable, this has resulted in increasing demands on the maritime shipping industry and on port activities. As its manufacturing base developed, China was importing

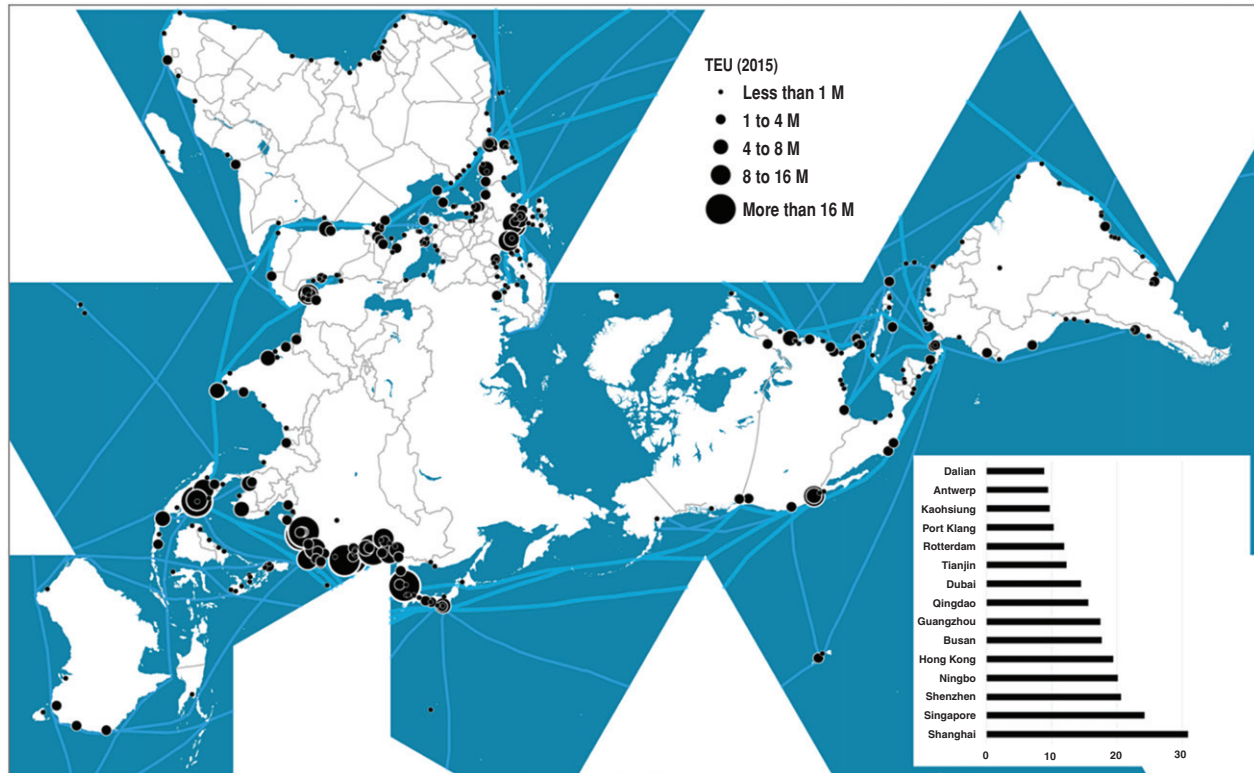


Figure 1 World's major container ports and sea routes, 2015. Source: *Rodrigue, 2017*.

growing quantities of raw materials and energy and exporting growing quantities of manufactured goods. The outcome has been a surge in demands for international transportation, but also problems of imbalanced trade. The ports in the Pearl River delta in Guangdong province now handle almost as many containers as all the ports in the United States combined.

Air Transport

The role of air transportation in the global economy is substantial as a support for passenger and freight mobility (Bowen, 2010). For the mobility of goods, it accounts for about 13% of the value of international trade with goods of a high value to weight ratio such as electronics account for this preponderance. Further, the fast delivery of parcels from e-commerce, which substantially leans on air transport, is a dominant segment of the air freight industry. Traditionally, freight was carried in the bellyhold of passenger airplanes and provided supplementary income for airline companies. However, specialized cargo-only airplanes as well as specialized air freight carriers are becoming dominant. This goes on par with a level of specialization of airports particularly in the United States, where centrally located airports such as Memphis and Louisville are handling a large share of the freight. Pacific Asian airports such as Hong Kong, Singapore and Incheon have also a significant preponderance as air freight gateways for Asian exports (Fig. 2).

Passenger air travel is linked with the level of economic development and the structure of the regional urban system. While most air travel concerns short distances (less than 2 hours of flight time), international passenger movements are dominated by air transport. There are three major concentrations of airports around which the world's air traffic is articulated: North America, Western Europe, and East Asia. The key airports of these platforms, or rather the main airport cities since many metropolitan areas have more than one airport correspond to the world's most prominent cities and the most important financial centers. Yet, this supremacy is being challenged by new hubs of activity such as Beijing and Dubai. There is a direct relationship between the level of air passenger traffic and the primacy of a city in the world urban system. In some cases, the level of passenger activity is related to a pronounced touristic or resort function of an area (e.g., Las Vegas, Orlando, Cancun, Venice, Palma de Mallorca). Air travel is therefore a complex mix of business, personal and leisure derived mobility (Fig. 3).

Large airport terminals are also seeing a substantial concentration of related activities such as distribution centers, just-in-time manufacturers, office parks, hotels, restaurants, and convention centers. These activities have a notable global orientation in terms of their customer base.

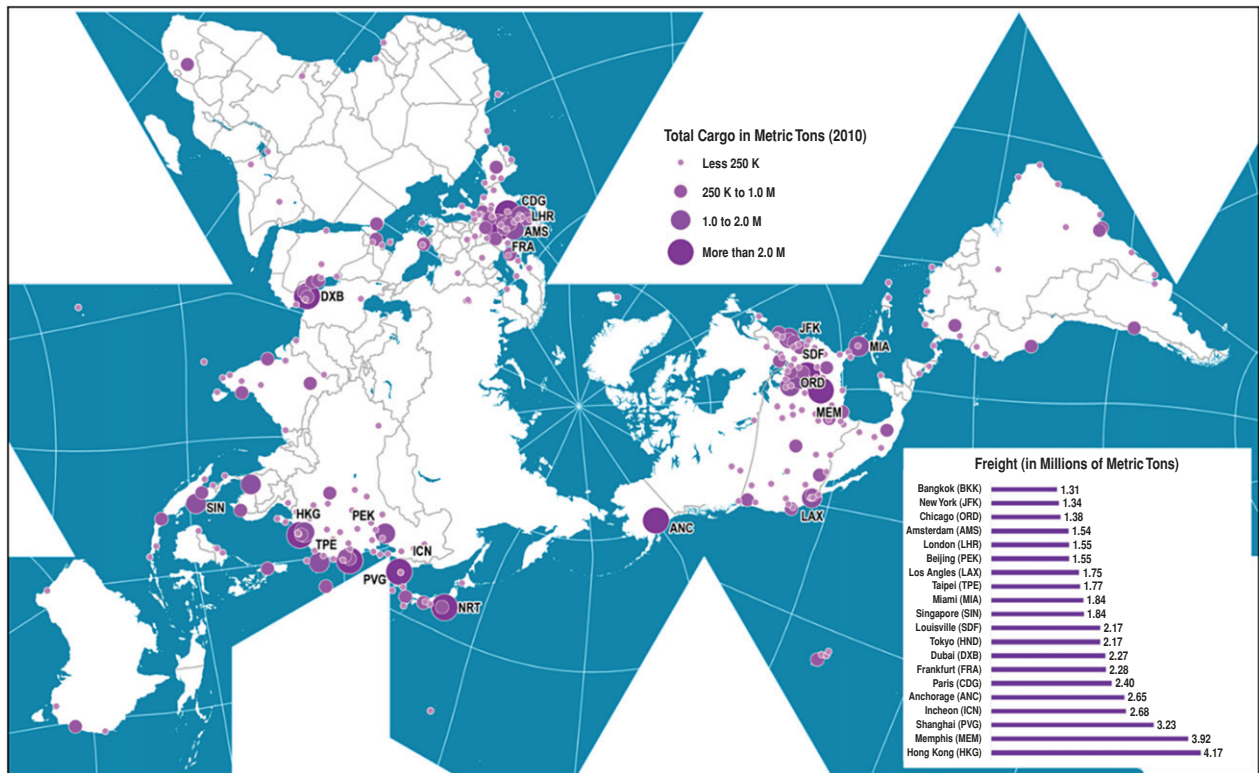


Figure 2 Freight traffic at the world's largest airports, 2010. Source: Rodrigue, 2017.

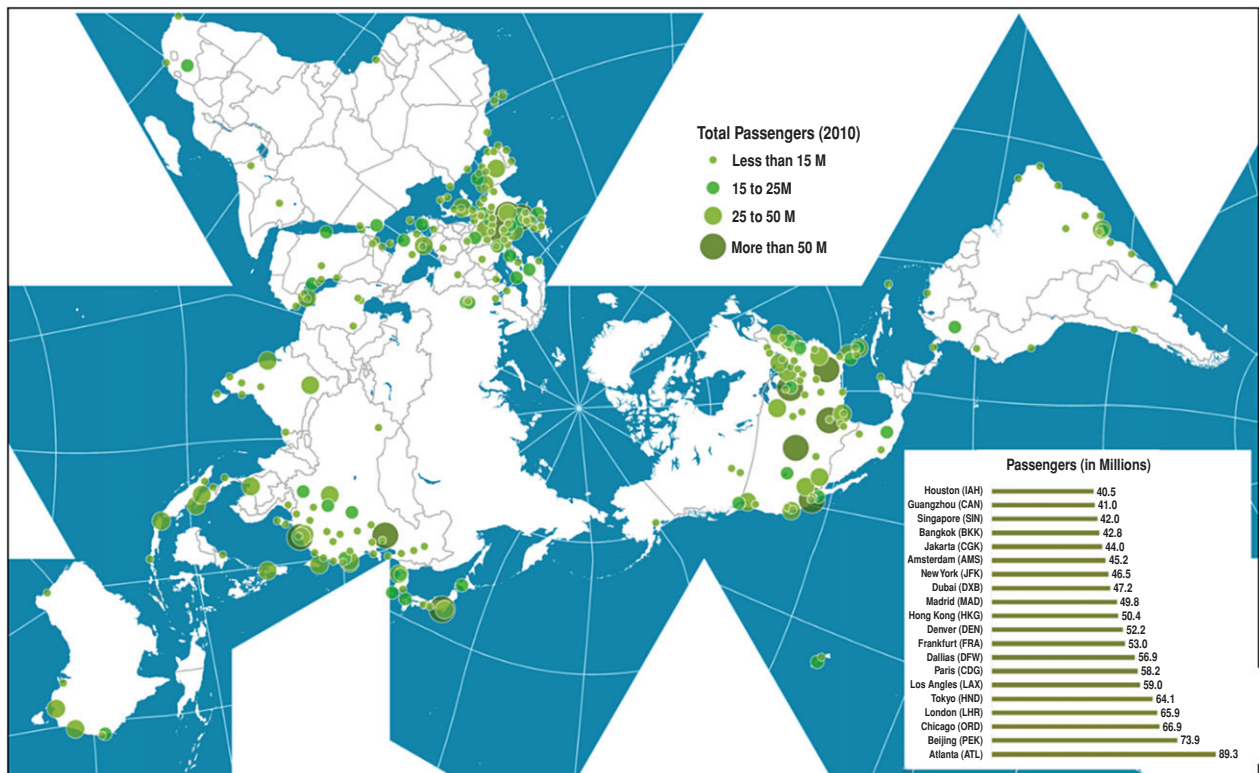


Figure 3 Passenger traffic at the world's largest airports, 2010. Source: Rodrigue, 2017.

Logistics: Producing, Organizing, and Managing Transportation Modes in a Global Economy

The complex web of global production, transportation, and consumption requires managerial efforts. Logistics is the series of activities required for goods to be made available on markets. This mainly includes purchase, orders processing, inventory management, and transportation. As the range of production expanded, transport modes adapted to new demands in freight distribution where the reliability and timely deliveries can be as important as costs. Logistics has consequently taken an increasingly important role in the global economy, supporting a wide array of value chains. First, improvement in transport efficiency expanded the geographical range of value chains. Second, the wide diffusion of information technologies allowed corporations to establish a better level of control over their value chains. Third, technological improvements such as intermodal transportation, enabled an increased continuity between different transport modes and thus within value chains.

The results have been a decrease of the friction of distance and a spatial division of production. This process is strongly imbedded with the capacity and efficiency of international and regional transportation modes, especially maritime and land routes. It is becoming rare for the production stages of a good to occur at the same location, even within a region. Consequently, value chains are integrated with transport modes with logistics as a strategy used to reduce time. For instance, value chains are also becoming increasingly managed by demand, implying that what is being produced is done so to match as closely as possible to what is being consumed. Logistics services are becoming complex and time-sensitive to the point that many firms are subcontracting parts of their distribution activities to specialized logistics providers operating their own modal assets.

An important aspect in the development of transportation modes concerns their manufacturing, which has taken a global dimension. The manufacturing of vehicles such as cars or trucks is mostly for regional use, but the sourcing of parts is commonly global with a propensity to cluster around major assembly plants. Further, exports of assembled vehicles to foreign markets is substantial for vehicles manufactured in Germany, Japan, and Korea, underlining a high level of globalization for both manufacturing and distribution. For aircraft manufacturing, one of the most complex and advanced engineering achievements, centralized final assembly plants, such as in Renton and Everett (Washington, United States) for Boeing commercial jets, are using parts supplied by thousands of manufacturers across the world. The end products have a global market as 70% of Boeing sales concerns customers outside the United States. In a similar fashion, Airbus assembles most of its planes in Toulouse (France) and Hamburg (Germany) from parts mainly provided from across the European Union. For the maritime shipping industry, 90% of the shipbuilding takes place in China, South Korea, and Japan, for conveyances that are used for long distance freight distribution.

Conclusion: Transportation Modes in a Globalized Economy

Globalization has been supported and expended by the development of modern transport modes. From large containerships to cars and delivery trucks, distribution systems have become closely integrated linking manufacturing activities with global markets. However, the 21st century brings many challenges to the role of transportation modes in the global economy. The capacity of many segments of transport system has been stretched by additional demands tying up long distance transportation modes supporting the mobility of passengers and freight. Congestion in many international transport terminals such as ports often causes delays and unreliable deliveries, and there is an acute need for improving inland transportation systems, notably those linked to the major gateways of the global economy. At the urban level, congestion, which is a local phenomenon, is often associated with regional and global distribution such as e-commerce. While developed economies have been able to provide substantial resources in coping with modal demands, many developing economies are facing challenges to provide infrastructures able to support an ongoing demand.

See Also

Wider economic impacts of transport investments; Airline economics; Transport and international trade; Maritime economics; Trade, freight transport, and logistics; Air freight and economic development; Rail freight; Intermodal and synchromodal freight transport; Infrastructural investments in transport modes; Transport modes and an aging society; Transport modes and tourism; Transport modes and accessibility; Introduction to air transport; The geography of road transport; Introduction to rail transport; The geography of rail transport; Introduction to maritime transport; The geography of maritime transport

References

- Bernstein, W.J., 2008. *A Splendid Exchange: How Trade Shaped the World*. Atlantic Monthly Press, New York.
Black, W., 2003. *Transportation: A Geographical Analysis*. Guilford, New York.

- Bowen, J., 2010. *The Economic Geography of Air Transportation: Space, Time, and the Freedom of the Sky*. Routledge, London.
- Dicken, P., 2015. *Global Shift: Mapping the Changing Contours of the World Economy*, seventh ed. The Guilford Press, New York.
- Rodrigue, J-P (Ed.), 2017. *The Geography of Transport Systems*. Fourth ed. Routledge, London.
- Stopford, M., 2009. *Maritime Economics*, Third ed. Routledge, London.
- Sultana, S., Weber, J. (Eds.), 2017. *Minicars, Maglevs, and Mopeds: Modern Modes of Transportation around the World*. ABC-CLIO, Santa Barbara, CA.

Transport Modes and Cities

Erick Guerra*, **Gilles Duranton†**, *University of Pennsylvania, Stuart Weitzman School of Design, Department of City and Regional Planning, Philadelphia, PA United States; †University of Pennsylvania, The Wharton School, Real Estate Department, Philadelphia, PA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	45
A Theory of Mode Choice in Cities	45
Cities and Travel Modes	46
Automobile Cities	46
Transit Cities	47
Motorcycle Cities	47
Bicycle Cities	47
Pedestrian Cities	48
Summary and Conclusions: The Future of Mode Choice and Cities	48
Acknowledgments and Further Reading	49
See Also	49
References	49

Introduction

From sprawling metropolises, where cars dominate, to medieval city centers, where most residents walk, there are big differences in how people choose to travel in different cities and regions. This chapter explores the range in types of urban travel around the world. Although there is often substantial variation in mode choice within metropolitan areas—most people in Manhattan use transit while most people in suburban Westchester drive—we focus our discussion around the most common modes and some of the common features that help to define and differentiate the cities that rely on them.

The remainder of this chapter is organized as follows. First, we present a general theory of where and why transportation mode varies. Next, we provide an overview of the world's various and varied automobile, transit, motorcycle, cycling, and walking cities. Last, we conclude with a summary of the chapter and discussion of the future of mode choice and cities.

A Theory of Mode Choice in Cities

Since Daniel McFadden's seminal work combining utility theory and qualitative choice modeling, random utility modeling has dominated estimations of whether and how people choose to travel to work, recreation, and other activities (McFadden, 2001; 1974). Increases in computing power have helped to contribute to a number of innovations, such as jointly estimating choices of housing location, activity participation, and mode choice (Ben-Akiva and Bowman, 1998). Nevertheless, the basic theory remains the same. People choose the mode that best suits them. Researchers can observe some of many of the factors that help shape individuals' mode choice, such as income and the relative cost of different modes, but others, such as personal preference, often remain obscure. In random utility modeling, the probability of a person choosing a mode depends on the probability that the unobserved utility, or random error, outweighs the estimated utility based on the observed factors about the trip, person, and available modes. Fundamental to the modeling framework is the assumption that people choose the option that provides the highest utility.

In previous work, we have argued that accessibility—quite simply the ease of reaching destinations—is the most important urban commodity from a resource allocation perspective (Duranton and Guerra, 2016). As evidence of this importance, households generally spend half or more than half of their income on housing and transportation. Housing location determines the distance, cost, and convenience to destinations throughout the city by different modes and transportation spending reflects the trips undertaken by different modes to access specific activities. Although defining and measuring accessibility has empirical as well as conceptual challenges, accessibility is high where residents can reach a wide variety of destinations, which are physically close and for which the cost of travel per unit of distance is low. A lack of accessibility is instead characterized by a paucity of destinations, long distances, and high transportation costs per unit of distance. The concept of accessibility also has close ties to utility, with the estimated utility from mode choice models proposed as a measure of individuals' accessibility to specific activities (Ben-Akiva and Lerman, 1979).

Urban form, the population, and public policy interact to shape which modes are likely to provide the best accessibility for the most people. Urban form, for example, directly shapes the relative attractiveness of different modes. Driving and finding parking in downtown Manhattan are generally costly and time-consuming. Transit services are regular, reliable, and inexpensive. In suburban

Westchester, by contrast, driving is relatively fast and inexpensive. Transit, by contrast, only serves specific destinations and is relatively slow and unreliable. Even accounting for costs, trip distances, and travel times, people are likelier to choose to walk, cycle, or take transit in cities and neighborhoods that are denser, more diverse, and more human-scaled (Ewing and Cervero, 2010). However, cities influence far more than the relative attractiveness of a given mode for a given trip. Cities are the physical environments where people live and must travel to interact with others. There are many components to this physical environment and they interact in a complex manner. For example, a new rail system in a densely populated city with good bus service will almost certainly attract more riders than a new rail system in a sparsely populated city with limited bus service.

Cities also change slowly over time. The investments, policies, and technologies of the past often have as much or more influence on accessibility than contemporary policies, households, or firms. For instance, Brooks and Lutz (2019) show that in Los Angeles the influence of the streetcar on patterns of urban density within the city remains conspicuous to this day, even though the streetcar started its decline as the dominant mode of transport after 1920 and the last piece of streetcar track was dismantled in 1963. More generally, contemporary urban form often relates most strongly to the transportation technologies during past periods of growth (Muller, 2004). New York, London, and Paris' impressive transit system's and transit-oriented development patterns were largely put in place a hundred years ago. Even the depopulating effects of US urban highway investments (Baum-Snow, 2007) are a half-century-old legacy in many US cities.

In terms of the population, household wealth is perhaps the most strongly associated with mode choice. As people get wealthier, they are likelier to move from inexpensive but slow modes, like walking or bicycling, to faster modes like transit, motorcycles, or cars. That said, many of the wealthiest cities have high rates of walking and bicycling. Similarly, even the poorest households rely on cars where sprawling cities and suburbs tend to have long trip distances and unreliable transit services (Blumenberg and Thomas, 2014; Boschmann, 2011; Giuliano, 2005). Those without cars tend to move to job-rich areas with better public transit (Glaeser et al., 2008) or borrow cars from neighbors or friends (Blumenberg and Smart, 2014; Lovejoy and Handy, 2011; Rogalsky, 2010). Other factors, like age and gender, are strongly associated with mode choice in some contexts but not in others. For example, although women are less likely to cycle than men in most cities and countries, there is little difference in cycling rates by sex in countries with high overall bicycle use, such as Germany, Denmark, and the Netherlands (Pucher and Buehler, 2010). The cycling gender gap appears to diminish with the quality of infrastructure (Emond et al., 2009; Garrard et al., 2008; Pucher and Buehler, 2008).

Finally, public policy plays an important role in mode choice. Low fuel taxes, minimum parking requirements, and unpriced roads increase car use in the United States (henceforth, US). High fuel taxes, no on-street parking, and substantial transit investments reduce car use in Japan. Public policy may be particularly important for achieving high rates of cycling in the wealthy cities and countries where cycling captures substantial mode share (Forsyth and Krizek, 2010; Pucher and Buehler, 2008).

Cities and Travel Modes

The following section describes some of the key characteristics of the world's automobile, transit, motorcycle, bicycle, and pedestrian cities. Quite simply we define an automobile city as a city where most people drive and, by extension, the automobile provides the best accessibility for the most people. A pedestrian city is a city where walking provides the best accessibility for the most people. Although this simple definition masks much of the complex nature of accessibility and the complex interactions between urban form, people, and public policy, it also gives us the opportunity to quickly summarize the wide range in mode choice across the world's diverse and complex cities and metropolitan areas.

Automobile Cities

Automobile cities tend to have long trip distances, low-density development, substantial amounts of roadway, high travel speeds, and near-ubiquitous free parking. These situations characterize most the US, where cars capture four-fifths of all trips and nearly nine-tenths of work commutes (US Department of Transportation, 2017). Of the 13% of workers, who walk, bike, or take transit to work, nearly all come from the largest, densest cities in the country. Across the hundred largest metropolitan areas in the US, the car continues to capture 88% of typical work commutes. Even in metropolitan New York, 60% of all commuters get to work by car. In the next largest metropolitan areas of Los Angeles and Chicago, the car captures 89% and 83% of all work commutes. Only in the central cores of US cities do transit and walking capture more trips than cars. In large Mexican urban areas, by contrast, the car captures less than a third of all commutes (Guerra et al., 2018). Consistent with our theory, car travel is substantially more convenient in the US than in other countries. For example, average residents of large US metropolitan areas travel at about 40 kilometers per hour (Couture et al., 2018), compared to 15–25 kilometers per hour in Indian cities (Akbar et al., 2019), 20 kilometers per hour in Bogota (Akbar and Duranton, 2017), and 12 kilometers per hour in Mexico City (Guerra, 2014).

Car ownership and car use have also been increasing substantially in other parts of the world, with China surpassing the US as the world's largest car market (The Guardian, 2010). In Mexico, for example, the size of the vehicle fleet in the largest metropolitan areas grew six fold from 1980 to 2010 (Montejano et al., 2019). As in the US and Europe, car ownership and use tend to go down in China and Mexico's densest cities and those with the least roadway and best transit supply (Guerra et al., 2018; Sun et al., 2017).

Transit Cities

There is a common saying in transportation planning that mass transit needs mass. For a public bus to make sense from an economic and environmental perspective, there need to be enough people going along the bus corridor to fill the bus seats. High-capacity systems, such as subway, light rail, and bus rapid transit (BRT), need even more passengers and thus an even greater mass of people living and working around stations. For the most part, cities in the developing world have mass, in terms of density as well as size. In Mexico City, for example, even the most remote suburban neighborhoods are dense enough on average to support high-capacity subways and metros (Guerra, 2014). In fact, many suburban neighborhoods are as dense as or denser than typical downtown neighborhoods.

There is no shortage of large, dense, and fast-growing cities. In fact, some of the fastest global population growth is happening in the largest cities, particularly the lowest-income ones. Nearly 350 million new residents will live in low and middle-income metropolitan with more than 5 million residents by 2030 (United Nations Population Division, 2014). Cities with 1–5 million residents, which are certainly large enough to support high-quality mass transit systems, are projected to add another 250 million residents. In total, 45% of the developing world's urban residents will live in cities with a million or more residents by 2030. These places are generally ideal locations for developing and strengthening transit use. While regular buses are more appropriate for smaller cities that often rely heavily on them, BRT systems offer a greater capacity and are suitable for cities in the 1–5-million range. High land costs—generally only found in the largest and relatively wealthier cities—are needed to justify the expense of going underground with subway systems.

While larger, denser cities tend to have higher transit use, smaller cities often rely heavily on transit as well. In Mexico, for example, workers use transit to get to work 37% of the time in metropolitan areas with between 75,000 and 200,000 residents compared to 51% in those with more than 200,000 residents. As in other parts of the world, many these workers rely primarily on privately provided transit in minivans and minibuses, with varying levels of service quality and official regulation. Researchers often refer to these services collectively as informal transit or paratransit, and locals use local names such as *angkots*, *mutatus*, or *peseros*, which describe specific vehicle types and services. Although frequently criticized for low vehicle standards and high emissions, these informal transit services are often the workhorse of public transportation. For the 1.25 billion residents of low- and middle-income countries with fewer than 300,000 residents, they are frequently the only available form of public transportation and a major source of employment. They also matter in large cities and those with extensive high-capacity transit networks. Across a score of large African cities, 36%–100% of public transit travel is on informal paratransit. In two-thirds of these cities, more than 80% of public transit is on informal paratransit (Behrens et al., 2016). In Mexico City and Bogota, half of motorized trips involve some form of minibus or minivan transit. Even in New York, dollar vans carry more than 100,000 passengers a day (Correal, June 8).

Motorcycle Cities

In many cities, motorcycles, mopeds, and other motorized two-wheelers have come to dominate the transportation system despite their sometimes uneasy and dangerous coexistence with other motorized vehicles. Asia currently has about three-quarters of the global two-wheeler fleet, and motorcycles are the primary mode of transportation in cities, such as Taipei, Ho Chi Minh City, and Hanoi. Small- to medium-sized Asian cities, which tend to have narrow streets, limited parking spaces, poor transit, and short trip distances, are particularly reliant on motorized two-wheelers. Solo, an Indonesian city of 500,000 to 600,000 residents, is an example of one such city. Between 2009 and 2013, the number of registered motorcycles more than doubled from 208,000 to 424,000—nearly one for every man, woman, and child. As in many other Indonesian and Vietnamese cities, rising incomes, inexpensive motorcycles, easy credit, and low fuel prices helped encourage a rapid increase in motorcycle ownership and use. Millions of others live in cities like Solo, where transit service is poor, transit use is low, and the cost and convenience of motorcycles have made them the mode of choice.

Motorcycle ownership and use have also been growing in a number of Latin American and African cities (Law et al., 2015). Motorcycle taxis are an important component of the transportation system in African cities such as Douala, Lagos, and Ouagadougou. In Jakarta, motorcycle owners can use the smartphone application Go-Jek to provide taxi services in a similar way to car owners with Uber or Lyft. Motorcycle use is sometimes thought to follow a Kuznets curve, where use increases with income up until a certain income level before decreasing as more households switch to driving cars. In the 1960s, many poorer European countries had larger motorcycle fleets than car fleets, much like in Asia today (Nishitatenno and Burke, 2014). Some European cities, like Rome, remain famous for high rates of motorcycle use. Despite substantially higher incomes than in Vietnamese or Indonesian cities, Taipei remain a motorcycle city. Given the relatively efficient way that motorcycles use road capacity—comparable to low-capacity transit vehicles in many cases—an income-driven shift from motorcycles to cars would likely produce substantial amounts of roadway congestion in contemporary motorcycle cities.

Bicycle Cities

Bicycles offer another opportunity to increase access to various parts of the city without increasing pollution, infrastructure costs, or traffic fatalities. Many cities in the developing world rely heavily on bicycles. As incomes have increased and the costs of cars and motorcycles decreased, however, bicycle use has decreased in many parts of the world. In Beijing, for example, bicycle mode share decreased sharply, from 54% in 1986 to 23% in 2007 (Zhang et al., 2014). Nevertheless, rising incomes do not necessitate a decrease

in bicycle use. Cities such as Amsterdam and Copenhagen prove that a city can be modern, wealthy, globally connected, and bicycle reliant. The Netherlands, Denmark, and Germany took active steps to make cycling convenient, comfortable, and safe (Pucher and Buehler, 2008). As a result, women and older adults are just as likely to bike as young men, who dominate cycling in the US and Great Britain. Cities such as Bogota and Buenos Aires have seen increased cycling rates as a result of investments in cycle tracks and bicycle paths.

Pedestrian Cities

For all the emphasis on cars and transit, walking remains the most globally important mode of transportation. Nearly everyone walks, even if just to get from the front door to a car parked on the street. Walking rates are particularly high in small, compact cities with low household incomes and walkable historic cores. These cities are common in Africa and South Asia and also in wealthier parts of Latin America and East Asia. Globally, almost 40% of all trips are made by foot, and the figure is close to 90% in many smaller and poorer cities (Un-Habitat, 2013). Even the largest cities can have high rates of walking. In Dhaka, the Bangladeshi capital of 17 million, more than half of trips are by foot, bicycle, or rickshaw. In Bogota, a relatively wealthy and large city, residents make 43% of all trips by foot. Density and a healthy mix of houses, shops, and businesses make walking an attractive alternative even in peripheral neighborhoods of massive metropolitan areas. In 2030, well over a third of the urban residents of low- and middle-income countries will live in cities with fewer than 300,000 people. Most will rely heavily on walking and other nonmotorized modes of transportation.

Because walking produces almost no local or global pollution, creates no traffic fatalities, costs residents only the food needed to power their legs, has proven health benefits, and requires low infrastructure investments relative to highways or transit, maintaining and encouraging high walking rates are important policy goals. The biggest disadvantage of relying on walking, of course, is the slow speeds. As households earn more income, members often shift to motorized forms of transportation such as cars, transit, or motorcycles that allow them to access more of the city. Unfortunately, as a result many cities and countries neglect pedestrian infrastructure under the assumption that residents will switch to motorized modes if they can afford them (Cervero et al., 2017).

Summary and Conclusions: The Future of Mode Choice and Cities

This chapter offered a general theory of where and why transportation mode varies. We then described some of the key characteristics of the world's automobile, transit, motorcycle, bicycle, and pedestrian cities. We now conclude with a brief discussion of the future of mode choice and cities in an era of rapid urbanization and the diffusion of new transportation technologies.

Shifts in settlement patterns and new technologies are likely to influence the relationship between mode choice and cities in the coming years. For example, as the global urban population has increased, particularly in low and middle-income countries, so has the process of suburbanization. Most metropolitan areas are expanding geographically more quickly than the population is growing. Although developing world cities are, on average, twice as dense cities in Europe and around seven times denser than cities in North America (Angel, 2012), densities have been declining by about 1%–2% annually (Angel et al., 2010). Based on experiences in wealthier countries, the rise in GDP per capita, and the rapid increase in vehicle fleets, it is easy to assume that as urban residents are getting wealthier, they are choosing to live in larger homes on larger lots in suburban areas and rely more heavily on cars. Yet the average suburban expansion in a developing city bears little physical or socioeconomic resemblance to the typical US or European suburb. Although there are many examples of high-income gated suburban communities, suburbs are generally poor and densely populated in the fast-growing cities of Latin America, Africa, and Asia. In China, suburban residents live in densely populated uniform tower blocks. In Mexico City—which is wealthier, more suburbanized, and more reliant on private cars than most developing cities—wealthier households opt to own cars and live in more central locations with good accessibility. Poorer households tend to live further from the urban center and rely on transit, particularly for longer trips such as commutes to jobs in the urban center (Guerra, 2015). Mass housing policies in many countries, moreover, have contributed to an increased demand for transportation by encouraging large, single-use housing developments on the periphery (Buckley et al., 2016). Improving transit in suburban neighborhoods is one of the key challenges for cities in the 21st century.

New technologies, such as smartphone-enabled ride-hailing services, have already had significant impacts on cities and transportation systems. Gorbach (2019) shows that ride-sharing services led to more restaurants to open and rents to increase in previously less accessible neighborhoods. Hall et al. (2018) provide evidence that ride-sharing and transit may complement each other. The effects of new technologies will likely strengthen over time. In the future, for example, self-driving cars could fundamentally alter how people travel—whether by car or transit—and have substantial spillover effects on cities. In some circumstance, particularly in wealthy cities with high driving rates, driverless vehicles will likely lead to an increase in the total amount that residents drive and could also eat substantially into transit mode share. In others, automated vehicles could substantially improve the quality of transit services by providing higher frequencies in smaller vehicles and serving lower-demand trip patterns more effectively. In the long run, self-driving cars could also vastly reduce the amount of space dedicated to parking. With more than twice as much parking as vehicles, most parking also sits idle, consuming space that could otherwise be used for housing, offices, or open space. On the other hand, there will be a long, complicated transition where self-driving and human-driven cars will coexist and may combine the inconvenient of both worlds with a lot mileage for users of self-driving cars and the poor driving conditions created by human-driven cars.

Where does that leave us? Rising incomes in many parts of the world strongly push towards an increased use of individual motorized modes of transportation. At the same time, motorized vehicles generate many undesirable effects, such as traffic collisions, pollution, substantial land consumption, congestion, and busy roads. The massive electrification of vehicles could substantially reduce local and global pollution. This will take time, even more so in developing countries where existing vehicles are used for longer and the switch to cleaner sources of electricity lags far behind. The dreamed paradise of little to no congestion and a minimal number of traffic collisions thanks to self-driving cars is even further away. At the same time, many cities of the world experience currently significant changes in their mobility with vehicle-sharing services and new types of vehicles like electric bicycles and scooters. These changes currently appear chaotic, but they are potentially rich in promise.

We would like to draw two broad policy lessons from all this. First, there is no silver bullet, but complementarities abound. In more developed cities, better transit may be appropriately complemented by congestion charging, the expansion of ride-sharing services, and the development of various forms of micromobility. Poorer cities need to think about both their hard transportation infrastructure as well as their soft transportation infrastructure such as the enforcement of traffic rules. Creating high-quality pedestrian environments and better road-based transit services are particularly important. Second, in urban transportation like elsewhere, one size does not fit all. Cities vary a lot in their urban form and levels of developments. The optimal policy mix is unlikely to be the same everywhere. European cities may envision their development centered around transit. The challenge is obviously very different for most US cities where trip distances are long and even the poorest residents mostly rely on private cars.

Acknowledgments and Further Reading

This chapter draws substantially on [Cervero et al. \(2017\)](#) and [Duranton and Guerra \(2016\)](#). Please consult these works for further reading on cities, accessibility, and mode choice.

See Also

Generalized Cost for Transport; Road diets; Latent Demand and Induced Travel; Mode Share/Choice Models; Residential Location Choice Models; Transport Modes and Accessibility; Active/Nonmotorized Transport Modes; Mode Choice and Life Events; Active Travel; Land Use and Transport Planning; Cycling

References

- Akbar, P., Couture, V., Duranton, G., Storeygard, A., 2019. Mobility and Congestion in Urban India. University of Pennsylvania, Mimeo. doi:10.1596/1813-9450-8546.
- Akbar, P., Duranton, G., 2017. Measuring the Cost of Congestion in Highly Congested City: Bogotá. Working Paper, CAF. Available from: <http://scioteca.caf.com/handle/123456789/1028>.
- Angel S.S., 2012. Planet of Cities. Lincoln Institute of Land Policy, Cambridge, MA Available from: http://wagner.nyu.edu/files/faculty/publications/PlanetofCities_Shloimo_Web_Chapter.pdf.
- Angel, S., Parent, J., Civco, D.L., Blei, A., 2010. Atlas of Urban Expansion. Lincoln Institute of Land Policy, Cambridge, MA Available from: <http://www.lincolninstitute.edu/subcenters/atlas-urban-expansion/>.
- Baum-Snow, N., 2007. Did Highways Cause Suburbanization? *Quart. J. Econ.* 122 (2), 775–805, doi:10.1162/qjec.122.2.775.
- Behrens, R., McCormick, D., Mfinanga, D. (Eds.), 2016. *Paratransit in African Cities: Operations, Regulation and Reform*. third ed. Routledge, London; New York.
- Ben-Akiva, M., Bowman, J.L., 1998. Integration of an activity-based model system and a residential location model. *Urban Stud.* 35 (7), 1131–1153, doi:10.1080/0042098984529.
- Ben-Akiva, M., Lerman, S., 1979. Disaggregate travel and mobility choice models and measures of accessibility. *Behavioural Travel Modelling*. In: Hensher, D.A., Sopher, P.R. (Eds.), Croom Helm, Andover, Hants., pp. 654–679.
- Blumenberg, E., Smart, M., 2014. Brother can you spare a ride? Carpooling in immigrant neighbourhoods. *Urban Stud.* 51 (9), 1871–1890.
- Blumenberg, E., Thomas, T., 2014. Travel behavior of the poor after welfare reform. *Trans. Res. Rec.* 2452 (1), 53–61.
- Boschmann, E.E., 2011. Job access, location decision, and the working poor: a qualitative study in the Columbus Ohio metropolitan area. *Geoforum* 42 (6), 671–682.
- Brooks, L., Lutz, B., 2019. Vestiges of transit: urban persistence at a microscale. *Rev. Econ. Stat.* 101 (3), 385–399, doi:10.1162/rest_a_00817.
- Buckley, R.M., Kallergis, A., Wainer, L., 2016. Addressing the housing challenge: avoiding the Ozymandias syndrome. *Env. Urbanizat.* 28 (1), 119–138, doi:10.1177/0956247815627523.
- Cervero, R., Guerra, E., Al, S., 2017. *Beyond Mobility: Planning Cities for People and Places*. Island Press, Washington, DC.
- Correal, A., June 8, 2018. Inside the Dollar Van Wars. *New York Times*, June 8.
- Couture, V., Duranton, G., Turner, M.A., 2018. Speed. *Rev. Econ. Stat.* 100 (4), 725–739, doi:10.1162/rest_a_00744.
- Duranton, G., Erick, G., 2016. *Developing a Common Narrative on Urban Accessibility: An Urban Planning Perspective*. Brookings Institution, Washington, DC.
- Emond, C.R., Tang, W., Handy, S.L., 2009. Explaining gender difference in bicycling behavior. *Trans. Res. Rec.* 2125 (1), 16–25, doi:10.3141/2125-03.
- Ewing, R., Robert, C., 2010. Travel and the built environment: a meta-analysis. *J. Am. Plan. Assoc.* 76 (3), 265–294, doi:10.1080/01944361003766766.
- Forsyth, A., Krizek, K.J., 2010. Promoting walking and bicycling: assessing the evidence to assist planners. *Built Environ.* 36 (4), 429–446, doi:10.2148/benv.36.4.429.
- Garrard, J., Rose, G., Lo, S.K., 2008. Promoting transportation cycling for women: the role of bicycle infrastructure. *Prev. Med.* 46 (1), 55–59, doi:10.1016/j.ypmed.2007.07.010.
- Giuliano, G., 2005. Low income, public transit, and mobility. *Trans. Res. Rec.* 1927 (1), 63–70.
- Glaeser, E.L., Kahn, M.E., Rappaport, J., 2008. Why do the poor live in cities? The role of public transportation. *J. Urban Econ.* 63 (1), 1–24.
- Gorback, C.S., 2019. Your Uber has arrived: ridesharing and the redistribution of economic activity. University of Pennsylvania, Mimeo.
- Guerra, E., 2014. Mexico City's suburban land use and transit connection: the effects of the line b metro expansion. *Transp. Policy* 32 (March), 105–114, doi:10.1016/j.tranpol.2013.12.011.
- Guerra, E., 2015. The geography of car ownership in Mexico City: a joint model of households' residential location and car ownership decisions. *J. Trans. Geograp.* 43 (February), 171–180, doi:10.1016/j.jtrangeo.2015.01.014.
- Guerra, E., Caudillo, C., Monkkenen, P., Montejano, J., 2018. Urban form, transit supply, and travel behavior in Latin America: evidence from Mexico's 100 largest urban areas. *Transp. Policy* 69 (October), 98–105, doi:10.1016/j.tranpol.2018.06.001.

- Hall, J.D., Palsson, C., Price, J., 2018. Is Uber a substitute or complement for public transit? *J. Urban Econ.* 108, 36–50, doi:10.1016/j.jue.2018.09.003, November.
- Law, T.H., Hamid, H., Goh, C.N., 2015. The motorcycle to passenger car ownership ratio and economic growth: a cross-country analysis. *J. Trans. Geogr.* 46 (June), 122–128, doi:10.1016/j.jtrangeo.2015.06.007.
- Lovejoy, K., Susan, H., 2011. Social networks as a source of private-vehicle transportation: the practice of getting rides and borrowing vehicles among Mexican immigrants in California. *Trans. Res. A Policy Pract.* 45 (4), 248–257.
- McFadden, D., 1974. The measurement of urban travel demand. *J. Pub. Econ.* 3 (4), 303–328, doi:10.1016/0047-2727(74)90003-6.
- McFadden, D., 2001. Economic choices. *Am. Econ. Rev.* 91 (3), 351–378.
- Montejano, J., Monkkonen, P., Guerra, E., Caudillo, C., 2019. The Costs and Benefits of Urban Expansion: Evidence from Mexico, 1990–2010. Working Paper. Lincoln Institute of Land Policy, Cambridge, MA. Available from <https://www.lincolnst.edu/publications/working-papers/costs-benefits-urban-expansion>.
- Muller, P., 2004. Transportation and urban form: stages in the spatial evolution of the American metropolis. In: Hanson, S., Giuliano, G. (Eds.), *The Geography of Urban Transportation*, third ed. The Guilford Press, New York, pp. 59–84.
- Nishitaten, S., Burke, P.J., 2014. The motorcycle Kuznets curve. *J. Trans. Geogr.* 36 (April), 116–123, doi:10.1016/j.jtrangeo.2014.03.008.
- Pucher, J., Buehler, R., 2008. Cycling for everyone: lessons from Europe. *Trans. Res. Rec.* 2074 (1), 58–65, doi:10.3141/2074-08.
- Pucher, J., Buehler, R., 2010. Walking and cycling for healthy cities. *Built Environ.* 36 (4), 391–414, doi:10.2148/benv.36.4.391.
- Rogalsky, J., 2010. Bartering for basics: using ethnography and travel diaries to understand transportation constraints and social networks among working-poor women. *Urban Geogr.* 31 (8), 1018–1038.
- Sun, B., Zhang, T., He, Z., Wang, R., 2017. Urban spatial structure and motorization in China. *J. Regional Sci.* 57 (3), 470–486, doi:10.1111/jors.12237.
- The Guardian, 2010. China overtakes US as world's biggest car market. *The Guardian*, January 8, 2010. Available from: <http://www.guardian.co.uk/business/2010/jan/08/china-us-car-sales-overtakes>.
- Un-Habitat, 2013. *Planning and Design for Sustainable Urban Mobility: Global Report on Human Settlements 2013*. Taylor & Francis.
- United Nations Population Division, 2014. *World urbanization prospects, the 2014 Revision*. Available from: <http://esa.un.org/unpd/wup/>.
- US Department of Transportation, 2017. *National household travel survey*. Available from: <http://nhts.ornl.gov/>.
- Zhang, H., Shaheen, S.A., Chen, X., 2014. Bicycle evolution in China: from the 1900s to the present. *Int. J. Sust. Trans.* 8 (5), 317–335, doi:10.1080/15568318.2012.699999.

Transport Modes and Remote Areas

Gina Porter, Department of Anthropology, Durham University, Durham, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	51
Measuring Remoteness and Inaccessibility	52
Pedestrian Travel	52
Cycling Modes	53
Conventional Motorized Transport (Cars, Minibuses, Buses, and Trucks)	53
Motorcycle Transport (and Its Facilitation by Mobile Phones)	53
ICT in the Transport Sector	54
Drones	54
Financing Transport Services in Remote Areas	54
References	55
Further Reading	55

Introduction

Remote areas are usually defined as places far away from cities and major centers of population that are difficult to get to (see, for instance, Collin English dictionary). In geographically remote areas of the world, both in the Global North and the Global South, transportation modes tend to be limited by the deficiencies in infrastructure that commonly define and characterize such places located far from cities and densely populated areas. There are two aspects to this poverty of accessibility: firstly, access to the remote area from more populous regions and, secondly, movement within it. Firstly, considering entry to the remote area, immediate access to remote islands—Tristan da Cunha or Easter Island, for example—will be restricted to sea and air transport modes. Access from populous areas to particularly remote locations in the major land masses of the Global North (such as Alaska, the Scottish highlands, or the Urals) or the Global South (where remote locations are a significant feature of every continent) can encompass more diverse transport modes: road and rail may feature in addition to sea and air transport. However, transport services—both passenger and freight—connecting the remote area to major population centers are likely to be limited as a consequence of the remoteness of the location, even if the area possesses a particularly valuable resource, or is of particular strategic significance.

Within these remote locations, wherever they are sited, there are commonly further challenges to mobility: interconnections linking different places within the area will typically be sparse, being underserved by transport infrastructure such as roads, railways, navigable rivers, and airports. Moreover, it is important to note the low service density—incidence of hospitals, schools, banks, administrative services, shops, market centers, and suchlike—that also commonly characterizes remote areas. These deficiencies in facility provision are inextricably linked to the low population density that helps keep such areas “remote” and to the fact that remote area interests are usually further disadvantaged through having less political clout than urban centers. Low service density, coupled with a shortage of cheap, reliable transport, can impact massively on people’s health, education, livelihoods, and general well-being. The combination of poor services and poor transport is often pernicious, promoting a spiral of poverty with intergenerational implications, evident in such diverse locations as the Amazon basin lowlands, Africa’s Sahelian plains, the Highlands of Scotland, and the forests of Rumania’s Carpathian Mountains.

In both the Global North and South, even very remote areas may be crossed by a few major land transport routeways—perhaps an all-season interregional highway or major railway. Minor settlements in the Global North are often connected to such highways by poor but still motorable single-track all-season roads. In many of the more extensive remote areas of the Global South, a majority of within-area transport connections between settlements are dependent on footpaths and minor, unsurfaced, seasonally impassable roads. This tends to reduce viable transport modes in the latter context to foot traffic, draught animals, bicycles, pushcarts, four-wheel-drive vehicles, motorcycles, light aircraft, and drones. Standard motorized transport that dominates in cities—cars, minibuses, and buses—cannot easily negotiate the poor infrastructure that characterizes remote areas: high maintenance costs and frequent breakdowns limit their use.

This entry focuses on transport modes in remote areas with particular attention to Africa because this is where the problem of limited access to transport infrastructure and consequent unavailability of low-cost, efficient transport modes for travel (over more than a few kilometers) is particularly severe. Regional estimates made in 2006 suggest that only 30% of rural people had access to an all-season road in sub-Saharan Africa, compared to 54% in Latin America and the Caribbean, 58% in South Asia, 75% in Europe and Central Asia, and 86% in East Asia and the Pacific (excluding data for China) (Roberts et al., 2006).

Lack of access to transport has enormous impact on the well-being of resident populations because of the particularly low density of health, education, market, and administrative services (World Bank, 2016). When transport services are sparse, competition

between and within modes is limited and costs per kilometer for motorized transport are inevitably high, unless subsidy is available (a rare occurrence in sub-Saharan Africa). Resident populations—already predominantly living at or below the poverty line—suffer accordingly. Women and children tend to experience such deficiencies particularly strongly for they are usually the least well resourced and thus the least able to afford motor-transport fares. Moreover, in some contexts, especially in low-income countries, the association of women’s improved personal mobility with a perceived increased potential for empowerment—or promiscuity—may be an underlying factor shaping (negative) male attitudes to women’s travel.

The paper is structured as follows. Measuring levels of remoteness and (in)accessibility to transport services is inevitably a major challenge in remote areas and is considered first below, before reviewing, in turn, the principal transport modes currently available to rural residents of remote areas: pedestrian travel, cycling, conventional four-wheel motorized transport (cars, minibuses, and trucks), and motorcycles. In each case, freight transport is considered alongside passenger movements. Final sections reflect on the potential of ICT and drones to alter the mobilities landscape of remote areas and the thorny issue of subsidy in the transport services arena.

Measuring Remoteness and Inaccessibility

Given the scale of the remoteness and accessibility problem in Africa (associated with much lower population densities than those prevailing in other world regions), much recent work has, unsurprisingly, focused on trying to establish detailed data for the continent that will improve on estimates of accessibility produced in 2006 by Roberts et al. (noted earlier). A study combining contemporary population count data with detailed satellite-derived settlement extents to map population distributions showed the average per-person travel time to settlements of more than 50,000 inhabitants to be around 3.5 h, with Central and East Africa displaying the longest average travel times (Linard et al., 2012). Subsequently, a new Rural Access Indicator (RAI)—a key global development indicator in the transport sector—has been developed by the World Bank (2016), utilizing new technologies and high-resolution population distribution data. This demonstrates the high correlation between poverty and areas with low road accessibility in a pilot in eight low-income countries in Africa and Asia. Although the RAI only measures the relationship between population distribution and road condition rather than transport services per se, it shows the share of the population that lives within 2 km of the nearest road “in good condition” in rural areas (i.e., an all-weather road as assessed in transport engineering terms), based on the widespread recognition that the 2 km threshold represents a reasonable distance for people’s normal economic and social purposes. In the eight pilot countries (six in eastern Africa, plus Bangladesh and Nepal), approximately 174 million people were estimated to be without road access: in terms of transport modes, this means they are likely to be dependent principally on pedestrian travel. At most, they may have access to a bicycle or—an increasing trend over the last 2 decades—a motorcycle. If Intermediate Means of Transport (IMT) such as push trucks and bicycles are available, they are often seen as men’s equipment: women’s use may even be culturally prohibited, especially where draught animals are concerned, but more often women simply do not have the financial means to purchase IMTs.

Pedestrian Travel

In remote areas of the Global South, especially parts of sub-Saharan Africa where population densities are particularly low, pedestrian travel dominates: for the most part this is a walking world. Rarely, there may be a paved all-season highway touching the area along which occasional conventional motor transport services travel, but most vehicles will speed through without stopping, on their way to major population centers outside the region. Beyond such a highway there are likely to be only unpaved roads and paths and very little in the way of conventional motor transport services. The impacts on children and women tend to be particularly severe.

The majority of children will walk to primary school here, sometimes over long distances: a factor that contributes to late enrolment, truancy, and early dropout, especially when coupled with household labor demands, including children’s pedestrian transport of goods (e.g., water and firewood for domestic use, agricultural produce to local markets). Secondary schools are sparse in remote areas and, when the only option is a very long walk, girls are generally unable to attend because of parental fears that they may be attacked on journeys between school and home (Porter et al., 2017).

Women’s health also tends to suffer significantly in remote locations, especially maternal health and emergency obstetric care (in this case affecting even women of relatively high socioeconomic status). Babies are still regularly born along the roadside as women about to give birth trek to the nearest health center. Knock-on impacts of poor access to maternal health services in remote areas of Africa not only include some of the world’s highest rates of maternal mortality but also long-term health conditions, notably obstetric fistula which affects all aspects of women’s lives by keeping them virtually isolated from their communities (Ehiri et al., 2018).

Such high dependence on pedestrian transport in Africa’s remote areas has a wider negative influence when it comes to freight, which has to be headloaded to more accessible locations where motor transport is available: culturally, this is often designated women and children’s work. Given the high incidence of poverty, even available motor transport may not be utilized and extremely long distances traversed with loads of up to c. 50 kg or even more. This further impacts negatively on the health status of pedestrian load carriers (Porter et al., 2013).

Cycling Modes

Cycling offers a relatively low-cost means of personal travel and transport of smaller loads for residents of remote areas (using a luggage rack or panniers for the latter purpose). However, cycles are still not within the budget of poorer households. In many cultural contexts in Africa, especially more conservative remote rural areas, moreover, cycling is only an option for men and boys. There is evidence from diverse rural areas in Africa of the constraints around girls' cycling, which range from lack of financial resources to purchase a cycle, to girls never having time available to learn to cycle, arguments that only men's cycles with crossbars are available, that men's cycles are too large for women, difficulties of cycling wearing a wrapper (skirt) or with a baby on the back, dangers of girls showing their legs as they cycle, and the potential to sustain gynecological damage by cycling (Porter et al., 2017).

In some remote regions, bicycles also offer a source of income (albeit only for men) through use as a taxi, when fitted with reinforced forks, stronger brakes, a passenger seat, and footrests. Bicycle taxis known as "boda boda" reportedly originally developed spontaneously in the Kenya–Uganda border area (hence the name) in the 1960s but became widespread across Uganda and western Kenya (where they have, to a considerable extent, been superseded by motorcycle taxis that bear the same name). In Malawi, where incomes are lower, bicycle taxis are still an important source of livelihood in remote rural areas, operating between small market centers and off-road settlements. In areas where there is insufficient demand to support conventional motor services they provide an extremely valuable service but, interestingly, in West Africa bicycle taxi services have always been much less common than in eastern Africa.

Various efforts have been made by external organizations such as the World Bank to promote use of cycle-trailers to aid load carriage in poorer remote areas in Africa and Asia. However, these have generally failed despite their apparent potential. In Ghana, for instance, failure of one such project was blamed on bad workmanship in construction of the trailers (which were required in a hurry) and specific difficulties with the hitch mechanism when the carts were heavily loaded, while twisting and breakage of cartwheel rims was attributed principally to the poor condition of local roads. The fact that the bicycle carts tend to need a track width of at least 1 m is a further factor militating against their widespread adoption in remote areas.

Conventional Motorized Transport (Cars, Minibuses, Buses, and Trucks)

In remote areas, conventional motorized transport tends to be sparse beyond any rare all-season road. Population density and low purchasing power of the remote area together act as a powerful constraint on the potential for bus and lorry hire services to be established (unless public subsidy occurs). In many contexts, it is simply not profitable to operate either from or into remote locations, due to insufficient demand and the fact that ordinary motor vehicles cannot easily navigate unsurfaced tracks, especially in wet conditions. Levels of wear and tear on unsurfaced roads are such that maintenance costs that are extremely high: tyre life, for instance, may be reduced by 25%–50% of normal life. Additionally, there will be reduced fuel efficiency, reduced vehicle utilization due to lower speeds, and an overall reduction in vehicle life. Unsurprisingly, where minibus services operate in such areas, fare prices per kilometer are often double or more those prevailing on good paved roads. Meanwhile in the trucking sector, cartels among those operators working along any international corridors that traverse the remote area may facilitate receipt of substantial profit margins even after high-operating costs are taken into account (Teravaninthorn and Raballand, 2008).

Motorcycle Transport (and Its Facilitation by Mobile Phones)

Motorcycles now form a key mode of transport in remote areas of low-income countries in Asia and Africa, often even in countries (such as Ghana) where they are currently illegal (Bishop and Barber, 2018). Up to about 2 decades ago, motorcycle usage in sub-Saharan Africa was very low in many remote areas and often limited mostly to government services (as exemplified by Ghana Ministry of Health supply of motorcycles to its staff involved in outreach in remote areas). Purchase costs were too high for the majority of rural dwellers and spares were rarely available.

Since that time motorcycle uptake has expanded rapidly in many remoter areas, due to the growing availability of cheap imported machines from China and India and the benefits that motorcycle usage offers for improving passenger services and short-haul of goods, particularly on routes where road conditions are inadequate and/or demand insufficient to support regular conventional motor services. Such benefits, once they are coupled with the ready availability of low-price motorcycles, have promoted a massive uptake of motorcycles as a business venture, that is, as motorcycle taxis, used mostly for passenger transport. The facility of motorcycles in moving through narrow paths and tracks where conventional transport cannot pass, means they can also be used for transporting freight (e.g., crop harvest) from farm to home and/or main consolidation points.

This expansion, which gathered pace first in the 1990s in south Asia, then from the turn of this century in Africa, is still ongoing in remoter areas and has been significantly promoted by the expansion of mobile phone networks and the widespread acquisition of handsets, even in poor rural households, over roughly the same period. It is thus possible to call transport to the house, as opposed to walking or taking a bicycle to the nearest major road junction to try to find transport. For the first time in African rural transport history many—even very poor—residents of remote rural areas have the potential to summon transport when they need it. Even if people cannot afford transport on a regular basis, many can now obtain it in emergency contexts. Despite concomitant growth of motorcycle-related accidents, this improved access to transport is of vital importance to the well-being of residents of remote

locations. These new connectivities associated with the expansion of motorcycle taxis and mobile phone networks look set to dominate inter-settlement travel in many remoter rural locations across Africa, in particular, for the foreseeable future.

Although fares are commonly higher than for other transport services, motorcycle taxis are not only reaching remote villages where there is no alternative service, but services are often stationed within those villages (rather than being limited to base stations at the paved road, as has tended to be the case with conventional motor taxis). Conventional taxis usually do not find it profitable to operate from or into remote locations due to insufficient demand and the fact that conventional four-wheel vehicles cannot easily negotiate unsurfaced tracks, especially in wet conditions. Motorcycle taxis thus commonly act as feeder services, linking off-road villages to other (cheaper) motor transport wherever this is available at the paved road.

Motorcycle taxis have also brought new employment opportunities for many young men (though barely any women) in remote areas, even if the motorcycle is owned by a wealthy urbanite. A common practice is for the rider to pay a daily rental to the owner—many can still make a reasonable level of profit. Unfortunately, however, high accident rates (associated with young men traveling at speed and without adequate helmet or other protection) impact not only on the young men themselves, but also on their (usually female) carers and on already stretched medical services.

Driver training geared specifically to motorcycle riders needs urgent attention. Drivers are supposed to be licensed, but corruption in the licensing authorities, police, etc., is so prevalent, in many low-income countries, that a majority of drivers probably operate without adequate training. Helmet usage is often required by law, for drivers, and their passengers, but rarely implemented or adequately enforced. Moreover, it is important to bear in mind that many helmets are of substandard quality and offer little protection (Bishop and Barber, 2018).

Despite the high incidence of motorcycle usage (whether as a commercial or personal transport), insufficient attention is being paid in road engineering to the question of road access conditions where motorcycles are now the dominant transport. Motorcycles need a firm, cambered running surface but can operate along a width much smaller than that required by conventional four-wheel vehicles. There is growing interest in developing appropriate interventions, including a recent pilot in Liberia (funded by GiZ) upgrading footpaths into motorcycle tracks (Jenkins and Peters, 2016). However, since better road surfaces are likely to encourage drivers to increase their speeds, traffic calming measures may be needed in areas around schools, settlements, etc.

ICT in the Transport Sector

As noted earlier, the mobile phone now plays a significant role in improved transport connectivity in remote rural areas of Africa's low-income countries, notably as a facilitator for accessing motorcycle taxis, but also with reference to calling conventional four-wheel transport. This is increasingly feasible as ever more remote areas come within reach of cell phone towers, gradually reducing the coverage exclusion facing low density rural populations located off-road and uphill (Buys et al., 2009). Although the majority of residents in remote areas are poor, many now have access to a basic mobile phone handset and it is standard practice for people—male and female—to keep a selection of local transport operators' numbers on their phone list whom they can call (with calling generally being preferred to text messaging, especially among illiterate populations) (Porter, 2016). The more sophisticated apps that are now helping to support transport service organization in urban areas as smartphone usage grows in many low-income countries are lacking as yet in remote areas. However, rapid developments in the ICT field are such that the current dependence on basic handsets and associated informal organization of transport by phone through personal networks in remote areas could change dramatically over the next decade.

Drones

Drones (pilotless aircraft) are playing an increasingly important role not only in moving cargo into very remote areas in emergency contexts in low-income countries (where restrictions on drone usage are much fewer than in the Global North) but also in gathering information in remote areas that it is impossible or excessively dangerous to reach by other means. When earthquakes and floods occur, land communication is often severely disrupted and the only means of rapid response to the emergency needs of remote populations (light-weight relief items of vaccines, blood, water purification tablets, and suchlike) may be by drone (Rabta et al., 2018). Drones with landing capabilities would have even greater potential: one novel hybrid, for instance, is reportedly being tested by the government's Medical Stores Department in Tanzania with funding from GiZ. This will deliver medical supplies to and from Ukerewe Island, where deliveries are hampered by the infrequency of ferry services (<https://www.weforum.org/agenda/2018/10/drones-for-development-can-deliver-innovation-for-the-whole-industry-heres-how/>). Although disaster relief is currently the most prominent use of drones outside military contexts, drone delivery services look set to be established widely in the near future across remote rural areas, possibly worldwide.

Financing Transport Services in Remote Areas

In high-income countries (and some middle-income countries), subsidy is widely utilized to support transport services in remote areas—this often applies to buses but also can cover other modes, including air services, as, for instance, in the Scottish highlands

and islands and the Remote Air Services Subsidy (RASS) Scheme that forms part of the Australian Government's Regional Aviation Access Programme. In low-income countries, however, governments and donors tend to focus on road building and maintenance as key to improving access to remote areas and assume that a road in good condition is all that will be needed to bring in service providers. Consequently, transport service provision is left to the informal market. In such remote places, however, since poverty levels tend to be high and demand consequently low, it may be necessary to consider the role of subsidy in transport services, though donor resistance is high (despite continued subsidy of road construction).

There is certainly limited experience of formal transport subsidy in Africa, except in South Africa where the cost and complexity of managing transport subsidy interventions, especially in the informal sector, remain significant issues. However, a 6 month action research study in Malawi involving subsidy of daily minibuses from five villages along a good road, with randomized prices to households, showed that, at any price, a bus operator would never make enough revenue to break even on this road (Raballand et al., 2011). This work suggests that motorized services are often likely to need subsidizing in remoter areas; otherwise a road in good condition will most probably not lead to provision of service at an affordable price for the local population.

Hine et al. (2015) draw together a strong argument as to why there is need to provide ongoing financial support for services in the most remote areas of low-income countries. Where motorized passenger mobility is very low they suggest there is a good case to directly subsidize transport services. Local governments could set basic service standards for frequency, capacity, safety standards, routes and timetabling, loading rates, etc., and only providers adhering to those standards would be supported. As they point out, service subsidies would be very modest in relation to road construction costs and the potential for corruption is far lower than in the case of infrastructure investment. Rural road investment programs funded by donors and governments and road funds could be utilized for the subsidies. They also suggest pilots to test the use of ICT-based applications that could help consolidate goods and passenger loads more efficiently and avoid empty trips.

To conclude, the web of connectivity within which the physical transportation sector sits in remote areas of the world has changed significantly over the last 2 decades. Access to transport, whatever the mode, often now depends significantly on ICT, though for populations in low-income countries this means, above all, connection secured through a basic mobile phone. In such contexts, the direct constraints imposed by physical transport infrastructure—roads, in particular—are being reduced substantially in many locations through the expansion of motorcycle usage and now also, to a small extent, drones. However, careful regulation of both motorcycles and drones will be required by governments in order that their potential to improve service access for remote, poor populations is nurtured while at the same time ensuring that safety considerations are adequately observed. Subsidized motorized public transport services could play a valuable complementary role along selected routes, especially to support the mobility of older or infirm people who cannot easily travel by motorcycle. Further improvement of remote area service access by the expansion of e-services—especially e-health—can be expected and could more fully complement all current modes in future years: it is likely that services will, increasingly, be accessed by a fully joined-up mix of physical transport modes and ICT.

References

- Bishop, T., Barber, C., 2018. Enhancing understanding on safe motorcycle and three-wheeler use for rural transport. Inception Report. ReCAP for DFID, London. Available from: <https://www.gov.uk/dfid-research-outputs/enhancing-understanding-on-safe-motorcycle-and-three-wheeler-use-for-rural-transport-inception-report>.
- Buys, P., Dasgupta, S., Thomas, T.S., Wheeler, D., 2009. Determinants of a digital divide in sub-Saharan Africa: a spatial econometric analysis of cell phone coverage. *World Dev.* 37, 1494–1505.
- Ehiri, J., Alaofè, H., Asaolu, I., Chebet, J., Esu, E., Meremikwu, M., 2018. Emergency transportation interventions for reducing adverse pregnancy outcomes in low- and middle-income countries: a systematic review protocol. *Syst. Rev.* 7, 65, doi:10.1186/s13643-018-0729-2.
- Hine, J., Huizenga, C., Willilo, S., 2015. Financing rural transport services in developing countries: challenges and opportunities. ReCAP Discussion Paper. Available from: <https://assets.publishing.service.gov.uk/media/57a0899340f0b6497400015e/61280-SLOCAT2015-FinancingRuralTransportServicesinDevelopingCountries-GEN2016A-v151216.pdf>.
- Jenkins, J.T., Peters, K., 2016. Rural connectivity in Africa: motorcycle track construction. *Proc. Inst. Civil Eng. Transp.* 9 (6), 378–386.
- Linard, C., Gilbert, M., Snow, R.W., Noor, A.M., Tatem, A.J., 2012. Population distribution, settlement patterns and accessibility across Africa in 2010. *PLOS One* 7, 2, doi:10.1371/journal.pone.0031743.
- Porter, G., 2016. Mobilities in rural Africa: new connections, new challenges. *Ann. Am. Assoc. Geogr.* 106 (2), 434–441, doi:10.1080/00045608.2015.1100056.
- Porter, G., Hampshire, K., Dunn, C., Hall, R., Levesley, M., Burton, K., Robson, S., Abane, A., Blell, M., Panther, J., 2013. Health impacts of pedestrian head-loading: a review of the evidence with particular reference to women and children in sub-Saharan Africa. *Soc. Sci. Med.* 88, 90–97.
- Porter, G., Hampshire, K., Abane, A., Munthali, A., Robson, E., Mashiri, M., 2017. *Young People's Daily Mobilities in Sub-Saharan Africa: Moving Young Lives*. Palgrave Macmillan, London.
- Raballand, G., Thornton, R., et al., 2011. Are rural road investments alone sufficient to generate transport flows? Lessons from a randomized experiment in rural Malawi and policy implications. World Bank Policy Research Working Paper 5535, January 2011. World Bank, Washington, DC.
- Rabta, B., Wankmuller, C., Reiner, G., 2018. A drone fleet model for last-mile distribution in disaster relief operations. *Int. J. Disaster Risk Reduct.* 28, 107–112.
- Roberts, P., Shyam, K.C., Rastogi, C., 2006. Rural Access Index: A Key Development Indicator. World Bank, Washington, DC.
- Teravanihthorn, S., Raballand, H., 2008. Transport prices and costs in Africa: a review of the main international corridors. AICD Working Paper 14, July 2008. World Bank, Washington, DC.
- World Bank, 2016. *Measuring Rural Access: Using New Technologies*. World Bank, Washington, DC.

Further Reading

- Berman, G.S., 2006. Social services and indigenous populations in remote areas: Alaska Natives and Negev Bedouin. *Int. Social Work* 49 (1), 97–106, doi:10.1177/0020872806059406.
- Braathén, S., 2011. Air transport services in remote Regions. International Transport Forum Discussion Papers, No. 2011/13. OECD Publishing, Paris, doi: 10.1787/5kg9mq3cxrxx-en.

- Goodland, A., Kleih, U., 1999. Community access to marketing opportunities: options for remote areas in Sub-Saharan Africa. Literature Review (NRI Report No. 2459). Working Paper. Natural Resources Institute, Chatham, UK.
- Laird, J.J., Mackie, P.J., 2014. Wider economic benefits of transport schemes in remote rural areas. *Res. Transp. Econ.* 47, 92–102, doi:10.1016/j.retrec.2014.09.022.
- Osti, G., 2010. Mobility demands and participation in remote rural areas. *Sociol. Ruralis* 50 (3), 296–310, doi:10.1111/j.1467-9523.2010.00517.x.
- Paneque-Gálvez, J., McCall, M.K., Napoletano, B.M., Wich, S.A., Koh, L.P., 2014. Small drones for community-based forest monitoring: an assessment of their feasibility and potential in tropical areas. *Forests* 5 (6), 1481–1507, doi:10.3390/f5061481.
- Porter, G., 2002. Living in a walking world: rural mobility and social equity issues in sub-Saharan Africa. *World Dev.* 30 (2), 285–300.
- Porter, G., 2011. 'I think a woman who travels a lot is befriending other men and that's why she travels': mobility constraints and their implications for rural women and girl children in sub-Saharan Africa. *Gend. Place Cult.* 18 (1), 65–81, doi:10.1080/0966369X.2011.535304.
- Porter, G., 2014. Transport services and their impact on poverty and growth in rural sub-Saharan Africa: a review of recent research and future research needs. *Transp. Rev.* 34 (1), 25–45, doi:10.1080/01441647.2013.865148.

Modeling Mode Choice in Freight Transport

Lóránt Tavasszy*, Gerard de Jong[†], ^{*}Delft University of Technology, Delft, The Netherlands; [†]Institute for Transport Studies, University of Leeds, United Kingdom and Significance, The Hague, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	57
Theory: Mode Choice of the Firm	57
Random Utility Models for Mode Choice	59
Aggregate Mode Choice Models	60
Alternatives for Random Utility: Prospect Theory, Random Regret, and MCDA Models	60
Extensions to Mode Choice Models	61
Conclusions	62
See Also	62
References	62

Introduction

Mode choice models (sometimes also referred to as modal split models) explain the distribution of a given total freight transport flow over available transport modes: road, rail, waterways, sea, air and pipelines. Mode choice modeling is a vital component of freight transport models and instrumental to predicting future flows and assessing transport policy impacts. In many freight transport models, next to route choice, modal split is a very sensitive demand component, responding more to changes in transport time and cost than transport generation and distribution.

The dependent variable in a mode choice model can be a discrete variable (a choice of one of the modes) or a continuous variable, in the form of a probability or fraction. The latter allows to calculate modal shares for aggregate commodity groups or groups of firms within entire geographic areas. Explanatory variables will most often only include the price of transport by the available modes and the transport time and/or frequency of transports. More recent empirical models will also consider one or more of the following:

- reliability (% of goods delivered on time),
- flexibility (ability to handle short-term change requests),
- risk of damage during transport (% value loss),
- tracking and tracing possibilities of the cargo, and
- emission of noise and harmful gases or particles.

The sensitivity of modal choice to these attributes will differ between different commodity types (e.g., bulk versus general cargo), shipment sizes, industries, firm sizes, transport equipment used (e.g., containers) and geographic distances. The modes available will usually differ, depending on the geographical scale of the model. Different modes of transport have very different properties in terms of their broader economic and environmental impact, requirements on public funding and safety.

A key distinction in freight mode choice modeling is that between aggregate and disaggregate models. By “disaggregate” we mean here that the unit of observation is the individual decision-maker (firms in freight transport) as opposed to “aggregate” models where the units of observation are aggregates of decision-makers, usually geographical zones. The aggregate mode choice model, especially the aggregate logit model, is the most commonly used model for mode choice in freight transport. Disaggregate models, however, are better rooted in theory of individual decision-making behavior.

Mode choice decisions are seldom made in isolation. From logistics management theory and practice, it is well known that, indirectly, decisions about for example, shipment size and distribution channel can play an important role in the mode choice decision. Also for policy makers, it has become important to understand that mode choice cannot be influenced without considering the entire, relevant logistics environment. More and more, therefore, new models are being developed to take these relationships between mode choice and other decisions into account (Combes and Tavasszy, 2016; De Jong et al., 2016; Tavasszy et al., 2012).

In this chapter we first introduce the current theoretical foundation of mode choice models in random utility, discrete choice theory. Subsequently, we present alternative forms for disaggregate (firm level) and aggregate (geographical zone-based) models, and discuss extensions of mode choice models in recent research. We provide lessons from recent examples of recent model applications from the scientific literature and close the chapter with conclusions.

Theory: Mode Choice of the Firm

Below we introduce the theoretical foundations of mode choice models. We refer the reader who is interested in more detailed background material, to De Jong (2014).

We assume that the main decision-maker is the shipper, a firm sending goods to a receiving firm and therefore has a demand for transport services. Shippers have a choice to carry out the transport themselves or to contract it out to a logistics service provider (including carriers and firms that integrate services of different carriers for their customers). The shipper makes mode choice decisions for a flow of shipments. A shipment is defined as a number of units of a product that are ordered, transported and delivered to one receiver; this may be a larger or a smaller volume than one single vehicle load. Alternatives in mode choice can be road transport, rail transport, inland waterway transport, sea transport, air transport, and pipeline transport. An essential characteristic of all these alternatives is that they are discrete (chosen or not chosen, as opposed to the case of continuous variables or ordered choice alternatives).

Discrete choice models at the disaggregate level have first been developed in passenger transport, where the dominant choice paradigm and theoretical foundation is that of random utility maximization, RUM (Ben-Akiva and Lerman, 1985). The basic equation of the RUM model is:

$$U_{ik} = Z_{ik} + \varepsilon_{ik} \quad (1)$$

In which:

U_{ik} is utility that decision-maker k derives from choice alternative i ($k = 1, \dots, K; i = 1, \dots, I$);
 Z_{ik} is observed utility component; and
 ε_{ik} is random utility component.

Utility maximization belongs to the economics of consumer behavior, and seems at first sight inappropriate for explaining the behavior of the firm, where the standard economic paradigm is that of profit maximization or costs minimization. The random utility framework can be applied to freight transport choices by taking the (negative) total generalized transport costs as the systematic part of the utility function and adding a random costs component to this function:

$$U_{ik} = -G_{ik} + e_{ik} \quad (2)$$

In which:

G_{ik} is systematic component of generalized transport cost and
 e_{ik} is random cost component.

In Eq. (2) random costs minimization becomes random utility maximization. For the standard discrete choice models with an independent, zero mean error term and without heteroscedasticity, one might just as well write a plus sign in front of e_{ik} . The mode choice model for freight transport can then be estimated using the same software as for the RUM model in passenger transport. If one assumes negative coefficient signs for the costs variables, one can also write $+G_{ik}$ in Eq. (2), as we will do later.

More specifically, for the choice of mode between three alternative modes i (here road, rail and inland waterways—IWW), one can specify the following linear utility functions:

$$U_{\text{road}} = \beta_0 + \beta_1 \cdot \text{PRICE}_{\text{road}} + \beta_2 \cdot \text{TIME}_{\text{road}} + \beta_3 \cdot \text{REL}_{\text{road}} + e_{\text{road}}, \quad (3a)$$

$$U_{\text{rail}} = \beta_4 + \beta_1 \cdot \text{PRICE}_{\text{rail}} + \beta_5 \cdot \text{TIME}_{\text{rail}} + \beta_6 \cdot \text{REL}_{\text{rail}} + e_{\text{rail}}, \quad (3b)$$

$$U_{\text{IWW}} = \beta_1 \cdot \text{PRICE}_{\text{IWW}} + \beta_7 \cdot \text{TIME}_{\text{IWW}} + \beta_8 \cdot \text{REL}_{\text{IWW}} + e_{\text{IWW}}, \quad (3c)$$

In which:

COST is transport price (monetary units);
 TIME is transport time (hours); and
 REL is transport time reliability (hours).

$\beta_0, \beta_1, \dots, \beta_8$ coefficients to be estimated; we expect negative signs for $\beta_1, \dots, \beta_3$ and $\beta_5, \dots, \beta_8$, the sign for β_0 and β_4 can be positive or negative.

In Eqs. (3a)–(3c), the utility that would be obtained when choosing road transport depends on the transport cost and time for that shipment by road transport and its reliability, and likewise for the other two modes. A key reason for including the alternative specific error components e_{road} , e_{rail} , and e_{IWW} in the utility functions is that these represent important variables that are not observed (well) by the researcher. These may include attributes such as mentioned earlier, for example, concerning risk of damage or availability of tracking and tracing services.

In order to deal with the error components, the researcher assumes that these are random variables. One can derive different discrete choice models with different specific assumptions on the probability distribution of the error components.

Random Utility Models for Mode Choice

The most common assumption for the error components ϵ is that they are independently and identically (i.e., same variance across observations) distributed following the extreme value type I (or: Gumbel) distribution. This leads to the multinomial logit (MNL) model with the choice probabilities:

$$P_{ik} = \frac{e^{G_{ik}}}{\sum_i e^{G_{ik}}} \quad (4)$$

The MNL model can be estimated by Maximum Likelihood methods that do not involve any simulation. Several software packages contain MNL estimation (sometimes called “conditional logit”). The estimated coefficients can be applied on a sample of firms (shipments) to calculate probabilities for all choice alternatives and observations. If this sample is representative of the population studied, one can then simply sum the probabilities (this method is called “sample enumeration”) over all observations in the sample to get the market shares for the alternatives (e.g., the share of road transport in the total for road, rail, and inland waterways) as predicted by the model. For nonrepresentative samples, one can do a weighted summation with the population to sample fractions for each observation as weights. Different zones in a study area or different horizon years might even have different sets of weights. Such applications can give the impact of changing a single variable at a time (which can be expressed in the form of elasticities), but can also predict what would happen in case of an input scenario with changes for several (possibly all) variables in the model. Note that one can correct for the effects of nonrepresentative samples relating to the eventual choice (resulting in underestimation of the share of a mode of transport), by calibrating the alternative specific constants to observed shares in practice.

Discrete choice models estimated on stated preference (SP) data only should not directly be used for forecasting (including the derivation of elasticity values), since the SP data will have a different variance for the error component than real world data, which will affect the choice probabilities. This happens because in the experimental setup of the SP many things that can vary in reality are kept fixed, and vice versa. Therefore, for forecasting it is better to use revealed preference (RP) data or combined SP/RP data where the variance of the SP is scaled to that of the RP. Models estimated on SP data can directly be used to derive ratios of coefficients (such as a value of time or a value of reliability) because in calculating these ratios, the SP error component drops out.

A restriction of the MNL model for policy applications is that the cross-elasticities for competing modes are the same: if the cost of road transport increases, substitution will occur to rail and inland waterways in proportion to their current market shares. Another manifestation of the same phenomenon is the independence from irrelevant alternatives property: the ratio of the choice probabilities between two alternatives does not depend on any other alternative. These properties may be at odds with reality. In practice, there could for instance be more substitution between rail and inland waterways than between any of these alternatives and road transport. One way to avoid this problem is the use of a nested logit model in which alternatives are grouped in a nest, allowing correlation between them. Mathematically, the easiest representation is to distinguish two probabilities, linked to each other by the logsum variable.

$$P_{B_l k} = \frac{e^{G_{ik} + \lambda_l I_{lk}}}{\sum_i e^{G_{ik} + \lambda_l I_{lk}}}, \quad (5a)$$

$$P_{(ik|B_l)} = \frac{e^{H_{ik}/\lambda_l}}{\sum_{i \in B_l} e^{H_{ik}/\lambda_l}}, \quad (5b)$$

$$I_{lk} = \ln \sum_{i \in B_l} e^{H_{ik}/\lambda_l}. \quad (5c)$$

The first probability (5a) gives the chance that decision-maker k chooses an alternative within nest B_l . This depends on the generalized cost G of l plus a coefficient λ_l times the expected cost from the alternatives in the nest, represented by I_{lk} , the so-called “logsum” variable, relative to the same kind of costs for the all alternatives at the nesting level. The second (conditional) probability gives the change of choosing alternative i given that nest B_l has been chosen. This depends on the generalized cost of this alternative relative to those for all alternatives in the nest. The resulting unconditional probability that decision-maker k will select alternative i is:

$$P_{ik} = P_{(ik|B_l)} P_{B_l k} \quad (5d)$$

The coefficient λ_l is the “logsum coefficient” which gives the degree of correlation between the error components of the alternatives in nest B_l : the higher this coefficient, the lower the correlation. In estimation, this is an extra parameter to be estimated. The estimated value must be between 0 and 1 for global consistency (meaning: across the entire range for the exogenous variables) with RUM. If a value above 1 is found, this often is an indication that a different (especially a reversed) nesting structure would work better and be consistent with RUM.

Both MNL and nested logit are members of a family of models, the generalized extreme value family, which is consistent with random utility maximization. Interestingly, most of these have not been applied for freight transport yet, although they offer more flexibility in terms of substitution patterns between alternatives than MNL or nested logit. This includes cross-nested logit, where a single alternative can belong to more than one nest at the same time.

Assuming that the error components follow a multivariate normal distribution leads to the probit model (in the multinomial case abbreviated as MNP). This model is almost as straightforward to estimate as MNL when there are only two alternatives. Probit

models with more than two alternatives lead to multidimensional integrals, and simulation methods can then be used for estimation but this is all (still) relatively cumbersome. On the other hand, the probit model is very flexible, since one can define and estimate specific correlation parameters between the alternatives.

The mixed logit model has two error components $\varepsilon_{ik} + v_{ik}$, where one follows the Gumbel distribution (so that conditional on the other, the model is MNL) and the other can follow any distribution, such as the normal or lognormal. With this technique one can accommodate two ideas, which are both generalisations of MNL: (1) random coefficients, to describe taste variation or heterogeneity in preferences of decision makers and (2) specific correlation structures between alternatives. In practice, for forecasting and transport policy appraisal, the evaluation of mixed logit models is still time consuming. However, several current software packages allow for estimating these models.

An alternative to model heterogeneity is the latent class MNL model, that is related to mixed logit. Within the random coefficients specification, this model uses a discrete number of possible values for a coefficient, instead of one average value. A selection of values remains after estimation, which represents weights of different groups of users. Membership functions link the users to classes. Observable user attributes (e.g., firm size, sector) can be used to analyze class membership patterns.

There have been many applications of the logit model in its different forms, from the early disaggregated SP studies using MNL (Fowkes et al., 1991) to mixed logit work (Arunotayanun and Polak, 2011) and recent latent class models (Duan et al., 2016). Because of the high data acquisition costs, disaggregate mode choice models are often specifically made for submarkets (specific commodity groups, container types, corridors or cities) and are very rarely built to represent an entire national freight market. This is the realm of the aggregate freight models, which we discuss next.

Aggregate Mode Choice Models

The aggregate logit choice model makes us of a different kind of observation than individual choices: the flows of shipments on one and the same origin-destination (OD) zone pair, as distributed across different transport modes. The aggregate modal split model can be explained from the theory of individual utility maximization, assuming that all variations in characteristics of the decision-makers and of the goods are captured by the error component of the utility function. As this is a very questionable assumption, aggregate modal split models can best be seen as pragmatic models with a loose connection to theory of behavior. Nevertheless, they regularly lead to satisfactory empirical results (elasticities, forecasts) at relatively low effort (especially concerning data collection).

Although the model can be expressed with the same mathematical formulation as the disaggregate MNL model, a convenient and classical formulation is the logarithmic or “difference form,” here for the bi-modal case:

$$\log \frac{S_i}{S_j} = \beta_0 + \beta_1 (P_i - P_j) + \sum_w \beta_w (x_{iw} - x_{jw}) \quad (8)$$

In which:

S_i/S_j is the ratio of the market share of mode i to the market share of mode j .

P_i and P_j are the transport costs using these two modes.

$X_{iw} - X_{jw}$ are w ($w = 1, \dots, W$) differences in other characteristics of the two modes.

Aggregate logit models are popular as usually disaggregate data are not available, while national and regional statistics offer good observations of tons by OD zone pair and mode. Aggregate choice models can easily be estimated, producing an intuitively appealing S-shaped market shares curve and market shares between 0 and 1. Depending on the mathematical formulation used, this model can be estimated using the discrete choice estimation software for disaggregate models or by standard regression analysis. Because of these advantages, the aggregate logit model still is the single most used model specification in practical freight mode choice modeling, especially in national freight transport models (De Jong et al., 2016). As the estimation of these models relies on data about large flow volumes between regions, the models will perform less well for the modeling of modes with a very small weight share (say, below 5%) like intermodal transport and air transport. For these markets, disaggregate models are better suited.

Alternatives for Random Utility: Prospect Theory, Random Regret, and MCDA Models

The interpretation of discrete choice models as random utility maximization models provides a foundation in microeconomic theory. For some decisions the economic perspective might be more relevant (here one might think of important well-planned long-run decisions, especially in a business context); for other decisions other perspectives (psychological or sociological) might be more important. In recent years, alternative paradigms have gained popularity in transport modeling: prospect theory, regret minimization, and multicriteria decision analysis (MCDA). We discuss these as follows.

Prospect theory, developed by psychologists, posits the following behavioral traits:

- reference dependence—the valuation of an attribute depends on the current value;
- loss aversion—there is a difference in the valuation of gains and losses in an attribute;

- size dependence—there is a difference in the valuation at different values of an attribute; and
- nonlinear weighting of probabilities by respondents.

Regret minimization is based on the assumption of minimization of anticipated regret by decision-makers. Regret occurs when a nonchosen alternative scores better on some attribute than the chosen alternative. When two regret components are distinguished in the regret function, “observed” or “systematic” regret and random regret, this leads to the random regret minimization model RRM. The distributional assumptions for the random component(s) can be similar to those of the RUM model, and even the same software can be used for RRM and RUM, but the choice probabilities from the two types of models will be different. The key difference is that RRM captures the “compromise effect”: alternatives that perform “in-between” on all attributes, relative to the other alternatives, are generally favored over alternatives with a poor performance on some attributes and a strong performance on others. We refer to [Keya et al. \(2018\)](#) for a recent application for freight transport.

Multiple-criteria decision analysis presents a normative approach to support decisions by calculating the relative performance of alternatives using a compensatory and additive performance function, consisting of several criteria and their weights—exactly as applied in utility theory. Some methods in MCDA allow to infer weights for criteria from survey-based experiments, just as in discrete choice modeling. Although it is not rooted in a descriptive theory of choice behavior, the MCDA approach is powerful because of its mathematical and computational simplicity.

Extensions to Mode Choice Models

It is useful to consider integrative modeling of mode choice with other choices, because of interdependencies with other choices in logistics ([Roorda et al., 2010](#); [Tavasszy et al., 2019](#)). For example, in intermodal transport, route choice and mode choice are done simultaneously in a multimodal network with transshipment centers. A less obvious case is the relation between shipment size and transport mode. Unit costs of shipment will be lower for modes with higher carrying capacity. At the same time, lower unit costs of transport will promote a choice for a smaller shipment size, as transport costs will be less important compared to inventory holding costs. These are two counter-balancing (as opposed to mutually reinforcing) influences, suggesting that, theoretically, an equilibrium could be reached that results in minimal total logistics costs. Modeling these choices together allows to search for a consistent overall solution from a total logistics costs perspective. In practical terms, joint modeling of choices helps to embed mode choice models in comprehensive freight transport models. A recent review can be found in ([Engebrethsen and Dautère-Pères, 2019](#)).

The main streams of research here have combined mode choice with shipment size decision and trading decisions, vehicle type decisions, and route choice decisions. The most commonly used approaches are discrete choice modeling, sometimes embedded in network formulations (see, e.g., [De Jong and Ben-Akiva, 2007](#); [Jourquin and Beuthe, 1996](#)). As combined choices in network models go beyond the conventional route and link choice decisions, they are also termed supernetwork models (in case of mode and route choice combined) and hypernetwork models, in cases where nonnetwork related choices are included such, related for example to transport demand volumes and trade partners. Most multimodal network models assume deterministic choice behavior, but examples can also be found of applications of MNL and MNP.

One criticism of such models is that the simultaneity of decisions is not proven easily as in practice, as even very closely related logistics decisions (such as the shipment size and mode choice decisions) are often taken by different managers, reviewed with different frequencies and evaluated in different and often unconnected processes. One may assume that, over time, the aggregation of experiences in a company provides sufficient feedback to all managers in a company to align their decisions better. As there is only scant empirical evidence of this convergence ([Combes, 2012](#)), the construct of simultaneous optimality in decisions remains largely theoretical.

A connected line of research has emerged that recognizes the influence of different decisions-makers within a company and of related decisions in horizontal or vertical interfirm relationships. One line of approach to modeling interactions is the combination of game theory with experimental economics to collect data ([Holguín-Veras et al., 2011](#)). Such economic experiments are often carried out with students in the role of economic agents that interact with each other in various market settings. The advantage is that the stylized models can lead to simple and easy-to-follow processes. Another line of research tries to integrate the idea of interactions into SP survey design and modeling, using the so-called interactive agency choice experiments, a negotiation variant of SP experiments ([Hensher and Puckett, 2007](#)). As survey costs of such an approach will be relatively high, approaches to keep the amount of interactions within bounds are needed, such as minimum information group inference, MIGI. Building on the multiagent direction of thinking, another new line of research is the use of serious gaming as a way to obtain input for estimating choice models ([Kourouniotti and Tavasszy, 2017](#)). Gaming allows a more immersive experience but, at the same time, a less controlled experiment than surveys. In terms of the number a nature of responses of respondents, the deliberation process followed by agents and the choice situations experienced, the results of a game can vary widely. Yet they allow to record responses of a broad group of interacting players, making it interesting for more complex choice situations.

The emerging broader application frameworks for the aforementioned models are multiagent system models, also referred to sometimes as agent-based models. These include an explicit account of the different agents and their decision-making processes. Recent research has taken steps to include discrete choice models as part of this framework ([Le Pira et al., 2017](#)). For an example of an empirical multiagent model of mode choice, we refer to [Reis \(2014\)](#). Future work is directed toward further completion of the framework of multiple logistics decisions around freight mode choice.

Conclusions

In the above we have introduced the current mainstream framework for empirical mode choice modeling: discrete choice, random utility maximization-based models.

We have described the most often applied model variant, the MNL, including its variants. In everyday practice, the aggregate version of the logit model is the one most often applied for national level freight mode choice studies, while disaggregate models prevail for submarkets of the national freight market. Also we have presented some emerging alternatives in the form of prospect theory, random regret minimization, and (descriptive) MCDA models.

Finally, we discussed increasingly popular extensions to mode choice models, which endeavor to embed the choice of mode in the broader context of multiagent, comprehensive logistics decision making. These provide new frameworks for empirical research into mode choice models.

See Also

Behavioral Research in Freight Transport; Intermodal Freight Transport; National Freight Transport Models; Modeling Freight Transport Networks; Freight Transport Policy; Value of Time for Freight; Demand for Freight Transport

References

- Arunotayanun, K., Polak, J.W., 2011. Taste heterogeneity and market segmentation in freight shippers' mode choice behaviour. *Trans. Res. E Logist. Trans. Rev.* 47 (2), 138–148.
- Ben-Akiva, M.E., Lerman, S.R., 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*, 9. MIT Press, Cambridge, MA, USA.
- Combes, F., 2012. Empirical evaluation of economic order quantity model for choice of shipment size in freight transport. *Trans. Res. Rec.* 2269 (1), 92–98.
- Combes, F., Tavasszy, L.A., 2016. Inventory theory, mode choice and network structure in freight transport. *Eur. J. Trans. Infrastr. Res.* 16 (1), 38–52.
- De Jong, G., Ben-Akiva, M., 2007. A micro-simulation model of shipment size and transport chain choice. *Trans. Res. B Methodol.* 41 (9), 950–965.
- De Jong, G., 2014. Mode choice models. In: Tavasszy, L., De Jong, G. (Eds.), *Modelling Freight Transport*. Elsevier, Cambridge, MA, USA.
- De Jong, G., Tavasszy, L., Bates, J., Grønland, S.E., Huber, S., Kleven, O., et al., 2016. The issues in modelling freight transport at the national level. *Case Stud. Trans. Policy* 4 (1), 13–21.
- Duan, L., Rezaei, J., Tavasszy, L., Chorus, C., 2016. Heterogeneous valuation of quality dimensions of railway freight service by Chinese shippers: choice-based conjoint analysis. *Trans. Res. Rec.* 2546 (1), 9–16.
- Engelbrechtsen, E., Dauzère-Pérès, S., 2019. Transportation mode selection in inventory models: a literature review. *Eur. J. Operat. Res.* 279 (1), 1–25.
- Fowkes, A.S., Nash, C.A., Tweddle, G., 1991. Investigating the market for inter-modal freight technologies. *Trans. Res. A General* 25 (4), 161–172.
- Holguín-Veras, J., Xu, N., De Jong, G., Maurer, H., 2011. An experimental economics investigation of shipper-carrier interactions in the choice of mode and shipment size in freight transport. *Networks Spat. Econ.* 11 (3), 509–532.
- Jourquin, B., Beuthe, M., 1996. Transportation policy analysis with a geographic information system: the virtual network of freight transportation in Europe. *Trans. Res. C Emer. Technol.* 4 (6), 359–371.
- Keya, N., Anowar, S., Eluru, N., 2018. Freight mode choice: a regret minimization and utility maximization based hybrid model. *Trans. Res. Rec.* 2672 (9), 107–119.
- Hensher, D.A., Puckett, S.M., 2007. Theoretical and conceptual frameworks for studying agent interaction and choice revelation in transportation studies. *Int. J. Trans. Econ.* 34 (1), 17–47.
- Kourounioti, I., Tavasszy, L., 2017. Can we apply serious gaming as an alternative to stated preference experiments?—A Freight Modelling Case Study. In: *International Choice Modelling Conference*.
- Le Pira, M., Marcucci, E., Gatta, V., Inturri, G., Ignaccolo, M., Pluchino, A., 2017. Integrating discrete choice models and agent-based models for ex-ante evaluation of stakeholder policy acceptability in urban freight transport. *Res. Trans. Econ.* 64, 13–25.
- Reis, V., 2014. Analysis of mode choice variables in short-distance intermodal freight transport using an agent-based model. *Trans. Res. A Policy Pract.* 61, 100–120.
- Roorda, M.J., Cavalcante, R., McCabe, S., Kwan, H., 2010. A conceptual framework for agent-based modelling of logistics services. *Trans. Res. E Logist. Trans. Rev.* 46 (1), 18–31.
- Tavasszy, L.A., Ruijgrok, K., Davydenko, I., 2012. Incorporating logistics in freight transport demand models: state-of-the-art and research opportunities. *Trans. Rev.* 32 (2), 203–219.
- Tavasszy, L., de Bok, M., Alimoradi, Z., Rezaei, J., 2019. Logistics decisions in descriptive freight transportation models: a review. *J. Suppl. Chain Manag. Sci.* 1–13.

Travel Mode Choice as Reasoned Action

Sebastian Bamberg*, **Icek Ajzen†**, **Peter Schmidt‡**, *Department of Social Sciences, University of Applied Sciences, Bielefeld, North Rhine-Westphalia, Germany; †Department of Psychological and Brain Sciences, University of Massachusetts Amherst, Amherst, MA, United States; ‡Department of Political Science and Centre for Environment and Development (ZEU), University of Giessen, Ludwigstrasse, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	63
Psychological Determinants of Modal Choice Within a Multidisciplinary Approach	64
The Theory of Planned Behavior	64
Empirical Evidence in Support of the TPB in General and in Predicting Individual Car Use	65
Using the TPB for Designing and Evaluating Behavioral Change Interventions	66
Extending the TPB	68
Conclusions	68
See Also	69
References	69

Introduction

Modern, functionally differentiated societies depend on efficient transport systems to bring together their spatially distributed economic, social, cultural, and political activities. Transport science was launched in the 1950s with the aim of providing accurate forecasts of demand for travel (Pas, 1990). It is focused on the following four topics: (1) choice to travel or not, (2) choice of destination, (3) choice of travel mode, and (4) choice of route. These four topics are the pillars of transport policy making and planning referred to as the *four-step model* (McNally, 2007). The present chapter deals with the third step: travel-mode choice. For the other topics see the contributions of xxx in this encyclopedia.

Not least because of the strong political support provided by national governments, the motorized transport system has become the dominant travel mode with particular emphasis on the privately owned automobile. In 2015 worldwide about 1.3 billion cars were used to meet peoples' everyday mobility needs (Wagner, 2018). Offsetting the individual and societal benefits of the motorized transport system is the growing awareness of its serious drawbacks and deleterious side effects (see e.g., van Wee, 2014). At the local level, urban environments have become less livable due to particle emissions, noise, and traffic accidents. Worldwide in 2016 approximately 1.25 million people were killed in road accidents and 20–50 million were injured or disabled (ASIRT, 2019). In addition, road and parking infrastructure infringe on public land and can destroy sites of historical and cultural value with a detrimental effect on overall quality of life. At the global level, the motorized transport system is a major contributor to climate change. Road transportation accounts for about 20% of all carbon dioxide (CO₂) emitted annually into the atmosphere (Bamberg & Rees, 2017). As a result, motorized transit has become a strategic sector for policymakers interested in mitigating climate change. All these effects are societal problems produced by individual behaviors that are reasonable and often well intended from the individual's perspective (de Graaf & Wiertz, 2019).

Technological advances such as electric vehicles could potentially reduce transportation's impact on climate change, yet it seems unlikely that these technologies will be scaled so rapidly that they will have a significant impact on CO₂ emissions in the near term (European Environment Agency, 2018). Moreover, electric cars—like their internal combustion engine counter parts—require an extensive infrastructure of roads and parking facilities. To reduce the negative impacts of motorized transportation, governments and local authorities will therefore have to rely on interventions aimed at motivating people to reduce their reliance on the car and adopt more environmentally friendly means of transportation such as walking, cycling, or use of public transportation. These interventions must go beyond providing alternative means of transportation to change behavior by targeting beliefs, attitudes, and other psychological factors that influence transportation decisions.

Given our present focus on travel-mode choice, we focus on the role of psychological determinants of individual travel-mode decisions, especially car use in comparison to alternative modes of transportation. For this purpose, we introduce in the next section psychological determinants of travel-mode choice within a multidisciplinary framework. The theory of planned behavior (TPB, Ajzen, 1991, 2011) is employed as a general theoretical framework to conceptualize central psychological constructs influencing human decision making, the causal interrelationships among these constructs as well as their intersection with objective socio-demographic, space-related, and journey characteristic as well as contextual attributes like laws and regulations. In the following sections, we summarize the empirical evidence supporting the claim that the TPB provides a useful conceptual framework not only for a better understanding of individual travel-mode choice but also for the development of effective policy interventions (ress) aimed at reducing people's car use. In the final section we discuss limitations of current research and potentially fruitful new theoretical and methodological developments.

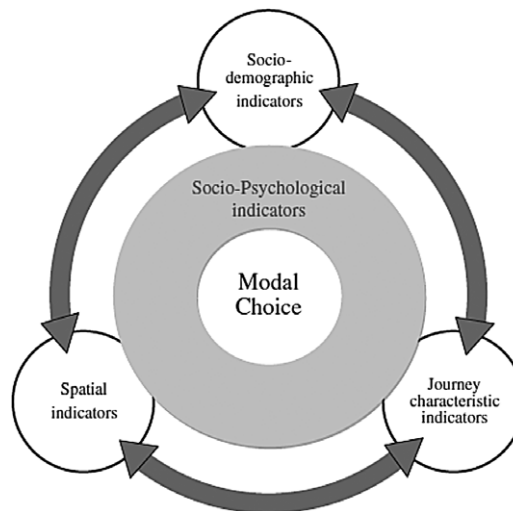


Figure 1 A multi-disciplinary approach for structuring modal choice determinants. Source: De Witte et al. (2013).

Psychological Determinants of Modal Choice Within a Multidisciplinary Approach

Figure 1 depicts the *multidisciplinary approach* proposed by De Witte et al. (2013) for integrating the travel-mode choice determinants discussed in the economic, social-geographic, and psychological literature.

On the outside circle, the framework distinguishes among three types of factors or indicators that can influence choice among available travel modes. These factors can be sociodemographic (e.g., age, gender, income, education, car availability), journey characteristic (e.g., distance, travel time, travel costs, departure time), or space-related (e.g., population density, diversity, proximity to infrastructure and services, frequency of public transport, parking). The connections among these determinant types represent the possible interrelations and dependencies that may occur among them (for a detailed discussion see De Witte et al., 2013). Together they encompass the objective travel-related context for a specific trip. The second circle represents the influence of psychological indicators like perceptions, experiences, attitudes, etc. These factors are assumed to determine how the objective possibilities shaped in the first circle are perceived and acted upon. Choice of travel mode (modal choice) is positioned in the middle as the result of the interactions among the sociodemographic, space, and journey characteristics, mediated by the impact of the social psychological determinants. What is missing however is the important influence of laws and regulations on travel-mode choice.

The Theory of Planned Behavior

To explicate the role ascribed to psychological factors in the multidisciplinary approach, a theoretical framework is needed that will provide answers to the following three questions: (1) Which psychological constructs should be viewed as the most important immediate determinants of observable travel mode choice? (2) How are these constructs causally related to each other? and (3) How is the impact of the objective indicators (outside circle in **Figure 1**) on observable travel mode choice mediated by the more proximal psychological indicators?

Over the past three decades the TPB (Ajzen, 1991, 2012, Fishbein & Ajzen, 2011) has become a frequently used framework for answering these three questions. Concerning the first question the TPB, like other cognition-behavior models (e.g., Rogers, 1983; Triandis, 1977), identifies behavioral intention as the most immediate antecedent of behavior. Behavioral intention refers to a person's readiness to perform an action (e.g., "I intend to drive to work"). In addition, the TPB proposes that perceived behavioral control (PBC), that is, people's appraisals of their ability to perform a behavior (e.g., "I have access to a car") may also be used to predict behavior when perceptions of control accurately reflect actual control over the behavior (Fishbein & Ajzen, 2011). Specifically, behavioral control acts a moderator of the intention-behavior relation (La Barbera & Ajzen, 2019; Yang-Wallentin et al., 2004), such that intention predicts behavior well to the extent that behavioral control is high.

Concerning the second question, the TPB assumes that behavioral intention ("I intend to drive to work") is determined by evaluations of the consequences of this action, that is, the attitude toward the behavior (e.g., "Driving to work is good/bad") and by perceptions of social approval of the behavior, termed subjective norm (e.g., "People who are important to me think that I should drive to work"). In addition, the TPB proposes that PBC also bolsters intention because perceived feasibility ("I have access to a car") is usually a prerequisite for motivation to engage in the behavior. Although in most applications of the TPB, attitude, subjective norms, and perceived behavioral control have been treated as independent predictors of intention, in the original formulation of the TPB (Ajzen & Madden 1986), PBC was postulated to moderate the effects of both attitudes and norms on intention (see also Fishbein

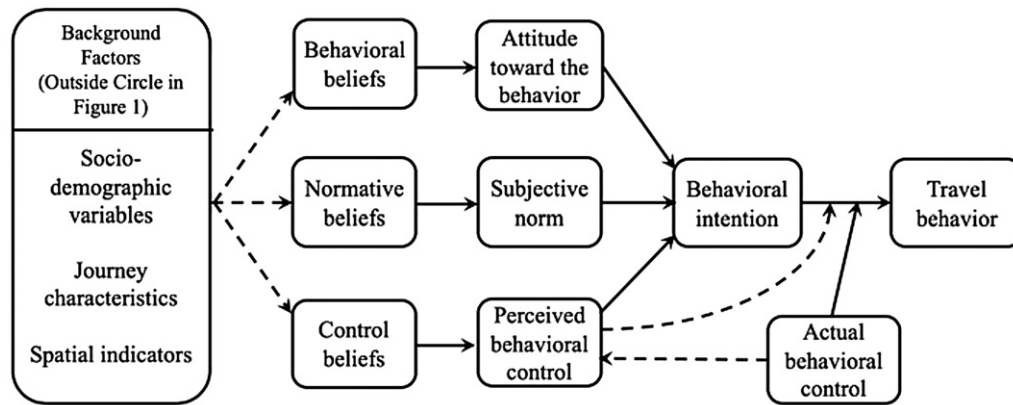


Figure 2 Using the TPB for modeling the psychological factors mediating the impact of objective environmental characteristics on observable travel behavior.

& Ajzen, 2011). Research in recent years has provided empirical support for the proposed interactions (see Castanier et al., 2013; Earle et al., 2019; Hukkelberg et al., 2014; La Barbera & Ajzen, 2019; Yzer & van den Putte, 2014).

Concerning the third question, the TPB assumes that beliefs provide the bases for attitudes, subjective norms, and perceptions of behavioral control. Specifically, attitude toward the behavior is based on beliefs about the likely positive and negative consequences of performing the behavior (*behavioral beliefs*, e.g., “driving to work is fast, reliable, and comfortable”), subjective norm rests on beliefs about the normative expectations of important others (*normative beliefs*, e.g., “my spouse approves of my driving to work”), and PBC is based on beliefs about the presence of factors that may facilitate or impede performance of the behavior (*control beliefs*, e.g., “It is easy to find parking near my place of work”). According to the TPB, behavioral, normative and control beliefs build the theoretical bridge over which objective external sociodemographic, journey, and spatial characteristics, considered *background factors*, impact the more proximal travel-mode choice determinants. An effective transport policy intervention must therefore target the behavioral, normative, and/or control beliefs most people associate with using alternative travel modes. For example, to discourage car use and increase biking to work, an intervention could attempt to raise the accessibility of positive outcomes associated with riding a bike and the accessibility of negative outcomes associated with car use. It could also try to raise the perception that important others approve of riding the bike to work and lower their perceived support for using the car. Finally, the intervention could attempt to increase skills or knowledge related to riding a bike and decrease actual barriers or establish facilitators for bike use, such as establishing dedicated bike lanes. Figure 2 presents graphically how the TPB answers the above raised three questions as a conceptual model. It can be seen that background factors (outer circle in the multi-disciplinary approach) are postulated to influence travel-mode intention and behavior indirectly by their possible effects on behavioral, normative, and control beliefs; and that the effects of behavioral, normative, and control beliefs on intentions are mediated, respectively, by attitude toward the behavior, subjective norm, and perceived behavioral control.¹ Finally, the effect to intentions on travel behavior is moderated by actual behavioral control or, when a measure of actual control is unavailable, by perceived behavioral control under the assumption that perceived control reflects actual control reasonably well.

For the sake of simplicity, we have left out in Figure 2 the additional differentiation of attitudes into instrumental and experiential attitudes, of subjective norms into injunctive and descriptive norms, and of PBC into capacity and autonomy factors, as proposed by Fishbein and Ajzen (2011). In addition, we have also omitted feedback loops from behavior to beliefs. It stands to reason that once a behavior has been performed, e.g., when people ride a bike to work for the first time, the experience may well change some of their existing beliefs and introduce new beliefs. For example, they may discover that riding a bike takes longer than anticipated but that it is less dangerous than feared. Or they may realize that they are not receiving the anticipated approval of supervisor and coworkers for riding their bikes to work. As a result, they may change their intentions with regard to future travel behavior.

In recent years, applications of the TPB in various behavioral domains have relied increasingly on structural equation modeling to test the model’s fit to the data. This requires transformation of the conceptual model shown in Figure 2 into a graphical model with latent and observed variables and analyzed as graphical input or syntax input by such programs as AMOS, LISREL, Mplus, EQS, R-lavaan or STATA (de Leeuw et al., 2015; Ho et al., 2013).

Empirical Evidence in Support of the TPB in General and in Predicting Individual Car Use

Besides its clear and parsimonious structure, the frequent use of the TPB as a theoretical framework for empirical research rests on its ability to predict behavioral intentions and behaviors across a variety of behavioral domains. Over 53 meta-analyses (e.g., Albarracín

¹ Note again that, theoretically, perceived behavioral control moderates the prediction of intentions from attitude and subjective norm, but in most applications, intentions are predicted from a linear combination of the three antecedent variables.

Table 1 Results of three meta-analyses on the relation between TPB variables and individual car use intention and behavior (Gardner & Abraham, 2008; Hoffmann et al., 2017; Lanzini & Khan, 2017)

	<i>Lanzini & Khan (2017)</i>				<i>Hoffman et al. (2017)</i>				<i>Gardner & Abrahams (2008)</i>			
	<i>k</i>	<i>N</i>	<i>r</i> +	95% <i>CI</i>	<i>k</i>	<i>N</i>	<i>r</i> +	95% <i>CI</i>	<i>k</i>	<i>N</i>	<i>r</i> +	95% <i>CI</i>
Car use - Intention	8	3,441	0.83	0.67, 0.91	7	2,375	0.50	0.31, 0.68	4	2,517	0.53	0.35, 0.72
Car use - PBC	7	2,399	0.27	0.09, 0.44	9	1,605	0.39	0.18, 0.60	2	324	0.31	0.03, 0.65
Car use - Habit	17	8,098	0.42	0.29, 0.53	6	2,058	0.47	0.39, 0.56	5	934	0.50	0.21, 0.79
Car use - Moral	7	4,222	-0.26	-0.38, -0.14	5	793	-0.35	-0.42, -0.28	2	563	-0.41	-0.70, -0.11
Intention - Attitude	7	2,906	0.56	0.40, 0.69					6	780	0.44	0.29, 0.58
Intention - Norm	7	2,906	0.42	0.33, 0.51					5	732	0.27	0.21, 0.32
Intention - PBC	7	2,906	0.32	0.12, 0.50					5	732	0.52	0.08, 0.97
Intention - Moral	2	421	-0.51	-0.58, -0.44								

Note. *k* = number of pooled studies, *N* = pooled sample size, *r* + = pooled correlation between car use and psychological construct, 95% *CI* = 95% confidence interval

et al., 2001; Armitage & Conner, 2001; Haus et al., 2013; Overstreet et al., 2013; Schwenk & Möser, 2009) have demonstrated that attitude, subjective norm, and PBC explain 30%–50% of the variance in intention and that intention and PBC explain 20%–40% of the variance in behavior (Armitage & Conner, 2001; Webb & Sheeran, 2006).

In the domain of travel-mode choice, three meta-analysis have been published on the prediction of individual car use intentions and behavior (Gardner & Abraham, 2008; Hoffmann et al., 2017; Lanzini & Khan, 2017). Table 1 summarizes the main results of these meta-analyses.

As proposed by the TPB, in all three meta-analyses behavioral intention and PBC both significantly and substantially predicted individual car use. Two of the three meta-analyses (Gardner & Abraham, 2008; Lanzini and Khan, 2017) also examined the relation between driving intention on one hand and attitude, subjective norm and PBC toward driving on the other. As expected, attitudes, subjective norm and PBC accurately predicted driving intentions. However, it is striking to note the considerable differences in the meta-analytically pooled correlation reported by the Lanzini and Khan (2017) versus Gardner and Abraham (2008) and Hoffmann et al. (2017) meta-analyses: Whereas Lanzini and Khan (2017) reported an average correlation of $r + = 0.83$ between pooled car use intention reported car use, in the other two meta-analyses, the correlations were $r + = 0.53$ and $r + = 0.50$. Because the reported confidence intervals do not cover the large differences, they cannot be attributed to chance. The results reported by Lanzini and Khan (2017) seem to be based on a systematically different data set than the results of the two other two meta-analyses, which were quite similar.

As noted, in the TPB, behavioral, normative and control beliefs not only provide the bases for attitudes, subjective norms, and perceptions of behavioral control, they are also the central targets of effective interventions designed to change behavior. Despite the central role ascribed to beliefs in the TPB, relatively few TPB studies have examined these beliefs (Steinmetz et al., 2016). As a consequence, no meta-analyses on the association between the three belief types and attitude, subjective norm, and PBC are available. We can therefore only report the results of individual studies. In the domain of travel-mode choice, Bamberg and Schmidt (1992, 1998) conducted two of the few studies that systematically analyzed the role of beliefs in individual travel mode choice. For this purpose, Bamberg and Schmidt (1992) asked student to evaluate the importance and probability of ten behavioral beliefs related to different travel modes, such as quick, comfortable, without stress, and cheap. Simultaneously they asked how strongly important reference persons like friends, parents, or professors support the use of the car, bike and public transit for university trips and how motivated participants were to comply with these expectations. Additionally, important control beliefs, including availability of a car, bike, or public transport services, public transport knowledge, and perceived quality of public transport connection were assessed. A composite measure of 10 behavioral beliefs related to car use for university trips correlated $r = 0.75$ with attitude toward this behavior. Similarly, a composite measure of normative beliefs correlated $r = 0.72$ with the subjective norm toward car use, and a composite measures of control beliefs correlated $r = 0.85$ with perceived behavioral control over car use. For bike use the correlation between behavioral beliefs and attitude was $r = 0.69$, for normative beliefs and subjective norm $r = 0.74$, and for control beliefs and PBC $r = 0.74$. In the second study, Bamberg and Schmidt (1998) replicated these results using a much larger sample.

In a study on university employees' travel-mode choices for commuting, Mann and Abrahams (2012) also analyzed the relation between beliefs and attitude, subjective norm, and PBC. They found that five behavioral beliefs predicted 49% of the variance in attitudes toward car use for commuting, two normative beliefs accounted for 22% of the variance in subjective norms, and two control beliefs explained 6% of the variance in PBC. Similar results were obtained with respect to using public transit for commuting.

Using the TPB for Designing and Evaluating Behavioral Change Interventions

As stated in the introduction, one main reason for transport science's interest in better understanding of individual travel-mode choice is growing societal pressure to reduce the negative environmental impacts of motorized transport, especially private car use.

Table 2 Results of TPB interventions (Steinmetz et al., 2016)

	<i>k</i> (ES)	<i>N</i>	δ	95% CI	
Behavioral beliefs	11 (14)	2,902	0.39	−0.08	0.86
Normative beliefs	12 (17)	3,962	0.54*	0.05	1.03
Control beliefs	6 (6)	1,053	0.68*	0.00	1.36
Attitude	70 (101)	21,374	0.24*	0.15	0.34
Subjective norm	61 (85)	15,711	0.14*	0.08	0.19
PBC	80 (113)	24,897	0.26*	0.16	0.35
Intention	72 (108)	22,030	0.34*	0.24	0.45
Behavior	40 (51)	9,395	0.50*	0.24	0.75

Notes. *k* = number of studies; ES = number of effect sizes; *N* = sample size; δ = weighted average effect size; 95% CI = confidence interval.

Because in liberal societies using coercive measures is associated with high political costs, governments and local authorities are interested in developing effective interventions aimed at motivating individuals to voluntarily reduce reliance on their cars and adopt more environmentally friendly means of transportation such as walking, cycling or use of public transportation. Thus, from an applied perspective, the ultimate test of the TPB's usefulness consists in its ability to guide the development of effective behavior-change interventions, or more precisely, interventions that motivate participants to voluntarily reduce their car use. This is described in more detail in Fishbein and Ajzen (2011) and in Ajzen and Schmidt, 2020.

Designing a behavioral change intervention is a time- and labor-intensive process. With the TPB as a conceptual framework, Steinmetz et al. (2016) described this process as follows: Investigators first establish whether individuals fail to perform the desired behavior (e.g., cycling to work on a regular basis) because they are not motivated to do so, because they are motivated but incapable of carrying out the behavior (e.g., distance is too great), or both. To produce an intention to perform the behavior, the intervention must be targeted at behavioral, normative, and control beliefs that ultimately determine the behavior of interest. Once this has been accomplished, it becomes necessary to ensure that the participants have the means to translate their newly formed intentions into action. Such an approach requires the intervention to be closely tailored to the appropriate phase of the process (i.e., motivation vs implementation), accompanied by intermediate measures of the TPB constructs and adaptation of the process as needed. For instance, when prior steps to change motivation were not successful, moving to implementational methods may not be effective. Rather, starting a second cycle of motivational interventions may be more appropriate. The development of such policy measures is not a part of the theory itself but requires formulation of action hypotheses (Chen, 2014) that connect the policy measures with changes in behavioral, normative and control beliefs (see Figure 2).

Evaluating the effects of a newly developed intervention, one perhaps designed to reduce car use, is also costly. Drawing firm causal inferences requires experimental designs with randomized control trials (randomized assignment to a control- and treatment groups, e.g., Angrist & Pischke, 2015; Shadish et al., 2002) or at least quasi-experimental designs that include a control group but without random assignment. As a consequence, compared with the large number of correlational studies that have focused on testing the TPB's ability to predict intentions and behavior, the number of studies using the TPB for systematically designing and evaluating a behavior-change intervention is still relatively small. Recently Steinmetz et al. (2016) provided a meta-analytic summary of the results of these studies. They only included studies that used strong experimental or quasi-experimental designs, that is, studies had to include at least one treatment and one control group. Moreover, the treatment group had to be submitted to one or more behavior change methods that aimed to change one or several of the TPB variables. They were able to identify 82 papers reporting results of 123 interventions that satisfied these two criteria. Due to the limited number of studies that included measures of beliefs, it was not possible to investigate whether the intervention had a significant effect on the targeted behavioral, normative, and control beliefs. Hence, the analysis focused on the effects of the interventions on attitude toward the behavior, subjective norm, PBC, intention, and behavior.

Table 2 presents the main meta-analytical results. On average, interventions were successful in changing TPB variables. Most effect sizes were significant (except for behavioral beliefs) but moderate in magnitude, ranging from 0.14 for subjective norm to 0.68 for control beliefs. Most importantly, the interventions were quite effective in changing intentions ($\delta = 0.34$) and the targeted behavior ($\delta = 0.50$).

The behavior-change methods that could be analyzed in isolation were increasing skills, motivation, persuasion, and planning. Increasing skills was most successful in changing attitudes ($\delta = 0.39$), but surprisingly less effective in changing PBC ($\delta = 0.18$). Persuasion had its largest effect on PBC ($\delta = 0.49$) and intention ($\delta = 0.35$) and motivation was most successful for intention ($\delta = 0.51$). Planning had weak effects on intention ($\delta = 0.10$) and, as the only method, on subjective norm ($\delta = 0.16$). Furthermore, as can be seen in the last column, the assumption of homogeneity (*Q*) was falsified for all relationships, indicating that the effects of the interventions on the different TPB variables varied significantly across studies. Further research is needed to determine the factors responsible for the heterogeneity so that they can be taken into account in the development of policy measures.

In the domain of travel-mode choice the number of methodologically sound evaluation studies is small, and the number of TPB-based evaluation studies even smaller. Möser and Bamberg (2008) synthesized the results of 141 so-called "travel demand

management” interventions (TDM, [Kitamura et al., 1997](#)) from different narrative reviews published on the topic. Generally speaking, this kind of intervention can be characterized as information-centered approaches. [Möser and Bamberg \(2008\)](#) concluded that TDM interventions reduced by 7% the proportion of trips conducted by car (from 39% to 46%), a relatively small effect size that corresponded to a Cohen’s $d = 0.15$. The validity of this conclusion is, however, jeopardized by the weak methodological quality of the included studies, as all 141 studies used a one-group pre-test-post-test design. Two later meta-analyses have addressed this limitation. [Bamberg et al. \(2009\)](#), who assessed the effect of 15 behavioral interventions with experimental designs implemented in Japan, found an effect size of Cohen’s $d = 0.17$ as a result of such programs. Later, [Bamberg and Rees \(2017\)](#) meta-analyzed the results from eleven transport interventions with quasi-experimental and experimental designs, ten of which were conducted in the United Kingdom and one in Germany. Across the studies analyzed, an effect size of Cohen’s $d = 0.12$ was found, which corresponded to a reduction of approximately 5% in the car modal split share.

The above cited research by [Bamberg and Schmidt \(1998\)](#) is one of the rare examples of theory-based evaluation studies in the domain of travel-mode choice. It used the TPB as a theoretical framework for analyzing the effect of a “semester-ticket,” which consisted of the introduction in 1993 of a low-cost permanent regional public transport ticket for 30,000 students. The evaluation study assessed all TPB constructs before and after the introduction of the semester ticket. Results showed a significant increase in the beliefs “cheap” and “knowledge of time tables,” which was reflected in significantly higher means of the TPB constructs attitude, subjective norm, PBC, and intention. Actual use of public transport for university trips increased from 15.7% before introduction of the semester ticket to 30.6% after its introduction. However, a methodological weakness of this study is the problematic simple pre-test post-test design, which does not allow firm causal inferences concerning the intervention effect (see [Heath & Gifford, 2002](#), for a similar study with similar results). A later analysis based on data from all universities that had adopted the semester ticket by the year 2000 showed that, on average, private car use declined by 25%, which can be seen as another confirmation of the effect, this time based on the behavior of 1.6 million students.

Extending the TPB

Misunderstanding the TPB’s reasoned action assumption ([Fishbein & Ajzen, 2011](#)) to mean that in this framework people are assumed to be rational utility maximizers ([Bamberg & Schmidt, 1998](#); [Opp, 2019](#)), investigators have criticized the theory for underestimating the impact that moral beliefs and norms may have on people’s travel-mode choices. To remedy this situation, researchers have distinguished between subjective and moral norms (e.g., [Parker et al., 1995](#)) and have assumed that, in the context of climate change, moral norms may be an additional predictor of people’s intentions to drive their cars rather than use more environmentally friendly modes of transportation. For example, [Abrahamse et al., 2009](#) found that the TPB constructs were more powerful than moral norms in explaining current car use for commuting but that moral consideration were more powerful predictors than the TPB constructs for intentions to reduce car use in the future. In their meta-analysis, [Hoffmann et al. \(2017\)](#) reported a pooled correlation of $r = -0.35$ between perceived personal obligation of not driving (moral norm) and self-reported driving, for the same association [Gardner and Abraham \(2008\)](#) a pooled correlation of $r = -.41$, and [Lanzini and Khan \(2017\)](#) a pooled correlation of $r = -0.26$. Based on these results there is a growing consensus within the research community to view moral norms as an additional travel-mode-choice predictor, especially of the decision to use more environmentally friendly travel modes.

It has also been argued that the TBP does not adequately account for behaviors that may be regulated by automatic processes ([Triandis, 1977](#); but see [Ajzen & Dasgupta, 2015](#)). For instance, when behaviors are practiced in stable environments over time, they can be automatically initiated by environmental prompts with little or no conscious deliberation ([Strack & Deutsch, 2004](#)). Thus, since daily travel tends to occur in stable contexts, transport-mode choice may, over time, become less of a conscious choice and more of a habitual response executed with little reflection ([Gardner, 2009](#); [Gärling & Axhausen, 2003](#)). Habit was first introduced by [Triandis \(1977\)](#) as part of a behavioral model and it has been shown that the formation of habits may change the cognitive mechanisms underpinning travel-mode choice ([Aarts et al., 1998](#)). However, the use of past behavior frequency as an indicator of habit strength is problematic ([Fishbein & Ajzen, 2011](#)). For this reason, in their meta-analysis [Hoffmann et al. \(2017\)](#) included a summary of six studies using an independent measure of habit strength ([Verplanken & Orbell, 2003](#)) for predicting a person’s travel-mode choice. They found a pooled correlation of $r = 0.47$ between the habit measure and travel mode used. [Gardner and Abrahams \(2008\)](#) report a pooled correlation of $r = 0.50$, and [Lanzini and Khan \(2017\)](#) of $r = 0.42$. Thus, there is a growing consensus within the research community that a realistic model of travel-mode choice should also include an independent habit measure.

Conclusions

The research reviewed above provides strong support for the TPB as a useful theoretical framework for studying the role of psychological factors within the travel-mode choice process. Meta-analyses not only confirm the role of intention and PBC in predicting actual mode choice but also the role of attitude, subjective norm, and PBC as predictors of intentions. Still under-researched is the role of behavioral, normative, and control beliefs as determinants, respectively, of attitude, subjective norm, and PBC. In addition, we note two deficits in present research. First, investigators rarely use the framework provided by the TPB to derive systematic recommendations for public policy. Second, we have not seen any research on the effects of using laws or regulations to influence travel-mode choices. An example is provided by the introduction of free rides on public transportation for all employees of

the state of Hessia. It would be very useful to know how the effects of such an intervention, if any, can be explained by changes in the TPB constructs. Strong evidence is available for the effectiveness of TPB-guided interventions in other behavioral domains. Based on the evidence we reviewed in this chapter, we would recommend that researchers interested in understanding and influencing individual travel-mode choice also use the TPB as their conceptual framework for designing and evaluating an effective behavior-change intervention.

See Also

Demand For Passenger Transport; Are Travel Choices Derived From the Rationality Assumption?; Human Factors in Transportation; Behavioral Research in Freight Transport; Mode Share/Choice Models; Mode Choice and Life Events; Travel Behavior; Self-Report Instruments; Behavioral Change; The Theory of Planned Behavior; Habitual Behavior; Social, Personality, and Affective Aspects of Driving; Psychosocial Benefits of Travel; Motivational Messages; Electromobility: Attitudes and Perceptions; Sharing: Attitudes and Perceptions

References

- Aarts, H., Verplanken, B., Van Knippenberg, A., 1998. Predicting behavior from actions in the past: Repeated decision making or a matter of habit? *J. Appl. Soc. Psychol.* 28 (15), 1355–1374.
- Abrahamse, W., Steg, L., Gifford, R., Vlek, C., 2009. Factors influencing car use for commuting and the intention to reduce it: A question of self-interest or morality? *Transp. Res. Part F* 12 (4), 317–324.
- Ajzen, I., 1991. The theory of planned behavior. *Org. Behav. Human Dec. Proc.* 50 (2), 179–211.
- Ajzen, I., 2011. The theory of planned behaviour: Reactions and reflections. *Psychol. Health* 26 (9), 1113–1127.
- Ajzen, I., 2012. Attitudes and persuasion. In: Deaux, K., Snyder, M. (Eds.), *The Oxford Handbook of Personality and Social Psychology*. Oxford University Press, Oxford, pp. 367–393.
- Ajzen, I., Dasgupta, N., 2015. Explicit and implicit beliefs, attitudes, and intentions: The role of conscious and unconscious processes in human behavior. In: Haggard, P., Eitam, B. (Eds.), *The Sense of Agency*. Oxford University Press, New York, doi.org/10.1093/acprof:oso/9780190267278.001.0001, pp. 115–144.
- Ajzen, I., Madden, T.J., 1986. Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. *J. Exp. Soc. Psychol.* 22 (5), 453–474.
- Ajzen, I., Schmidt, P., 2020. Changing behavior using the theory of planned behavior. In: Hagger, M., Cameron, L.D., Hamilton, K., Hankonen, N., Lintunen, T. (Eds.), *Handbook of Behavior Change*. Cambridge, Cambridge University Press.
- Albarracín, D., Johnson, B.T., Fishbein, M., Muellerleile, P.A., 2001. Theories of reasoned action and planned behavior as models of condom use: A meta-analysis. *Psychol. Bull.* 127 (1), 142–161.
- Angrist, J.D., Pischke, J.S., 2015. *Mastering metrics: The path from cause to effect*. Princeton University Press, Princeton.
- Armitage, C.J., Conner, M., 2001. Efficacy of the theory of planned behaviour: A meta-analytic review. *Brit. J. Soc. Psychol.* 40 (4), 471–499.
- Bamberg, S., Fujii, S., Friman, M., Gärling, T., 2011. Behaviour theory and soft transport policy measures. *Transp. Policy* 18 (1), 228–235.
- Bamberg, S., Rees, J., 2017. The impact of voluntary travel behavior change measures: A meta-analytical comparison of quasi-experimental and experimental evidence. *Transp. Res. Part A* 100, 16–26.
- Bamberg, S., Schmidt, P., 1992. Eigentlich sollte ich ja mit dem Rad zur Uni fahren. –Verkehrsmittelbezogene Einstellungen und Verkehrsmittelnutzung Studierender. *Spiegel der Forschung* 2, 22–26.
- Bamberg, S., Schmidt, P., 1998. Changing travel-mode choice as rational choice: Results from a longitudinal intervention study. *Rational. Soc.* 10, 223–252.
- Castanier, C., Deroche, T., Woodman, T., 2013. Theory of planned behaviour and road violations: The moderating influence of perceived behavioural control. *Transp. Res. Part F* 18, 148–158.
- Chen, H.T., 2014. *Practical program evaluation: Theory driven evaluation and the integrated evaluation perspective*. Sage, London.
- de Graaf, N.D., Wiertz, D., 2019. *Societal problems as public bads*. Routledge, London.
- de Leeuw, A., Valois, P., Ajzen, I., Schmidt, P., 2015. Using the theory of planned behavior to identify key beliefs underlying pro-environmental behavior in high school students: Implications for educational interventions. *J. Environ. Psychol.* 42, 128–138.
- De Witte, A., Hollevoet, J., Dobruszkes, F., Hubert, M., Macharis, C., 2013. Linking modal choice to motility: A comprehensive review. *Transp. Res. Part A* 49, 329–341.
- Earle, A.M., Napper, L.E., LaBrie, J.W., Brooks-Russell, A., Smith, D.J., de Rutte, J., 2019. Examining interactions within the theory of planned behavior in the prediction of intentions to engage in cannabis-related driving behaviors. *J. Am. Coll. Health* 1–7. <https://doi.org/10.1080/07448481.2018.1557197>.
- European Environment Agency (2018, June 11). Electric vehicles as a proportion of the total fleet. Retrieved from <https://www.eea.europa.eu/data-and-maps/indicators/proportion-of-vehicle-fleet-meeting-4/assessment-2>.
- Fishbein, M., Ajzen, I., 2011. *Predicting and changing behavior: The reasoned action approach*. Taylor & Francis.
- Gardner, B., 2009. Modelling motivation and habit in stable travel mode contexts. *Transp. Res. Part F* 12 (1), 68–76.
- Gardner, B., Abraham, C., 2008. Psychological correlates of car use: A meta-analysis. *Transp. Res. Part F* 11 (4), 300–311.
- Gärling, T., Axhausen, K.W., 2003. Introduction: Habitual travel choice. *Transportation* 30 (1), 1–11.
- Haus, I., Steinmetz, H., Isidor, R., Kabst, R., 2013. Gender Effects on entrepreneurial intention: A meta-analytical structural equation model. *Int. J. Gender Entrepren.* 5 (2), 130–156.
- Heath, Y., Gifford, R., 2002. Extending the theory of planned behavior: Predicting the use of public transportation. *J. Appl. Soc. Psychol.* 32 (10), 2154–2189.
- Ho, M.R., Stark, S., Chernychenko, O., 2013. Graphical representation of structural equation models using path diagrams. In: Hoyle, R.H. (Ed.), *Handbook of Structural Equation Modeling*. Guilford Press, New York, pp. 43–55.
- Hoffmann, C., Abraham, C., White, M.P., Ball, S., Skippon, S.M., 2017. What cognitive mechanisms predict travel mode choice? A systematic review with meta-analysis. *Transp. Rev.* 37 (5), 631–652.
- Hukkelberg, S.S., Hagtvet, K.A., Kovac, V.B., 2014. Latent interaction effects in the theory of planned behaviour applied to quitting smoking. *Brit. J. Health Psycho.* 19 (1), 83–100.
- Kitamura, R., Mokhtarian, P.L., Laidet, L., 1997. A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. *Transportation* 24 (2), 125–158.
- La Barbera, F., & Ajzen, I. (2019). *Control interactions in the Theory of Planned Behavior: Rethinking the role of subjective norm*. Unpublished manuscript, Department of Psychology, University of Naples Federico II, Naples, Italy.
- Lanzini, P., Khan, S.A., 2017. Shedding light on the psychological and behavioral determinants of travel mode choice: A meta-analysis. *Transp. Res. Part F* 48, 13–27.
- Mann, E., Abraham, C., 2012. Identifying beliefs and cognitions underpinning commuters' travel mode choices. *J. Appl. Soc. Psychol.* 42 (11), 2730–2757.
- McNally, M.G., 2007. The four-step model. In: Hensher, K.J., Button, D.A. (Eds.), *Handbook of Transport Modelling*, 1. Elsevier, Amsterdam, pp. 35–53.
- Möser, G., Bamberg, S., 2008. The effectiveness of soft transport policy measures: A critical assessment and meta-analysis of empirical evidence. *J. Environ. Psychol.* 28 (1), 10–26.

- Opp, K.D., 2019. Can attitude theory improve rational choice theory or vice versa? A comparison and integration of the theory of planned behavior and value expectancy theory. In: Mayerl, J., Krause, T., Wahl, A., Wuketich, M. (Eds.), *Einstellungen und Verhalten in der empirischen Sozialforschung*. Springer, Wiesbaden, [Attitudes and behaviour in empirical social science research], pp. 65–96.
- Overstreet, R.E., Cegielski, C., Hall, D., 2013. Predictors of the intent to adopt preventive innovations: A meta-analysis. *J. Appl. Soc. Psychol.* 43 (5), 936–946.
- Parker, D., Reason, J.T., Manstead, A.S., Stradling, S.G., 1995. Driving errors, driving violations and accident involvement. *Ergonomics* 38 (5), 1036–1048.
- Pas, E., 1990. Is travel demand analysis and modelling in the doldrums? In: Jones, P.M. (Ed.), *Developments of Dynamic and Activity-Based Approaches to Travel Analysis*. Gower, Aldershot, UK, pp. 3–27.
- ASIRT (Association for safe international road travel) (2019, July 1). Road safety facts [webpage content]. Retrieved from <http://www.asirt.org/safe-travel/road-safety-facts>.
- Rogers, R.W., 1983. Cognitive and psychological processes in fear appeals and attitude change: A revised theory of protection motivation. In: Cacioppo, J.T., Petty, R.E. (Eds.), *Social Psychophysiology: A Sourcebook*. Guilford Press, New York, pp. 153–176.
- Schwenk, G., Möser, G., 2009. Intention and behavior: a Bayesian meta-analysis with focus on the Ajzen–Fishbein Model in the field of environmental behavior. *Qual. Quan.* 43 (5), 743–755.
- Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin, Boston, New York.
- Steinmetz, H., Knappstein, M., Ajzen, I., Schmidt, P., Kabst, R., 2016. How effective are behavior change interventions based on the theory of planned behavior? *Zeitschrift für Psychologie* 224 (3), 216–233.
- Strack, F., Deutsch, R., 2004. Reflective and impulsive determinants of social behavior. *Person. Soc. Psychol. Rev.* 8 (3), 220–247.
- Triandis, H.C., 1977. *Interpersonal behavior*. Brooks/ Cole Publishing, Monterey, California.
- van Wee, B., 2014. The unsustainability of car use. In: Gärling, T., Ettema, D., Friman, M. (Eds.), *Handbook of Sustainable Travel*. Springer, Heidelberg, pp. 69–83.
- Verplanken, B., Orbell, S., 2003. Reflections on past behavior: A self-report index of habit strength. *J. Appl. Soc. Psychol.* 33 (6), 1313–1330.
- Wagner, I. (2018, November 26). Number of vehicles in use-worldwide from 2006 to 2015. Retrieved from <https://www.statista.com/statistics/281134/number-of-vehicles-in-use-worldwide>.
- Webb, T.L., Sheeran, P., 2006. Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychol. Bull.* 132 (2), 249–268. <https://doi.org/10.1037/0033-2909.132.2.249>.
- Yang-Wallentin, F., Schmidt, P., Davidov, E., Bamberg, S., 2004. Is there any interaction effect between intention and perceived behavioral control? *Meth. Psychol. Res. Online* 8 (82), 127–154.
- Yzer, M., van den Putte, B., 2014. Control perceptions moderate attitudinal and normative effects on intention to quit smoking. *Psychol. Addict. Behav.* 28 (4), 1153–1161.

Energy Consumption of Transport Modes

Zissis Samaras*, **Ilias Vouitsis[†]**, * Department of Mechanical Engineering, Aristotle University of Thessaloniki, Greece; [†]Department of Mechanical Engineering, Aristotle University of Thessaloniki, Greece

© 2021 Elsevier Ltd. All rights reserved.

Introduction	71
Overview of Global Energy Use in Transport	72
Energy Consumption by Mode	73
Land Transport	74
Air Transport	74
Maritime Transport	74
Fuel Type by Mode	75
GHGs Emissions	76
Road Transport	77
Air Transport	77
Maritime Transport	78
Reducing Transport Energy Consumption	78
Regulations	78
Land Transport	78
Air and Maritime Transport	79
Technology	80
Road Transport	80
Air Transport	80
Maritime Transport	81
Behavioral Changes	81
Summary and Conclusions	81
Road Transport	82
Rail Transport	82
Air Transport	82
Maritime Transport	82
Biography	82
References	83
Further Reading	84

Introduction

Although people have traveled, traded, and shared goods and ideas for thousands of years, over the past 100 years there has been an almost unbroken rise in transport. The continuous mechanization of transport systems, the gradual shift from coal to oil, the improvements in engine propulsion technology and the associated reduction in mobility costs relative to incomes, have allowed people to travel at an unprecedented pace (**Fig. 1**).

Transport activity will continue to increase in the future as economic growth fuels transport demand and the availability of transport drives development. The majority of the world's population still does not have access to personal vehicles, and many do not have access to any form of motorized transport. However, this situation is rapidly changing. Industry analysts predict that there will be two billion motor vehicles on the roads by 2030, as private vehicle ownership spreads to the developing world ([Dargay et al., 2007](#); [Gross, 2016](#)). Both air passenger and air cargo traffic are resilient whereas freight transport has been growing even more rapidly than passenger transport and is expected to continue to do so in the future ([ICAO, 2017a](#)). Urban freight movement is dominated by trucks, while international freight is dominated by ocean shipping ([IEA, 2017a](#); [Mulholland et al., 2018](#); [UNCTAD, 2017](#)). **Fig. 2** summarizes these trends.

The paper is structured as follows: second and third sections provide a brief overview of the global energy end use in transport in general and the energy end use by transport mode; fourth section deals with the type of fuel used by each transport mode while fifth section discusses the transport greenhouse gas (GHG) emission trends; sixth section provides, first, an overview of the main policy approaches adopted by various regions around the world and, second, an overview of currently available best practices and promising developments for the near future to reduce energy-use in each transport mode.

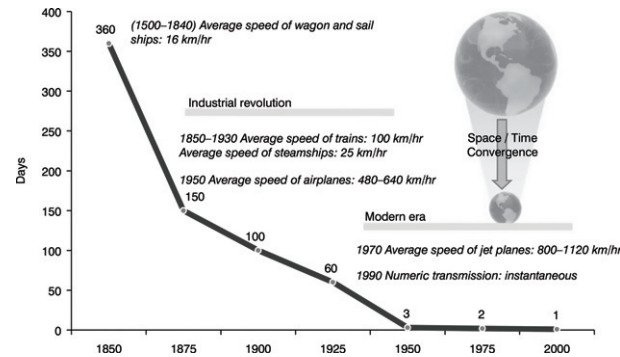


Figure 1 Global space/time convergence: days required to circumnavigate the globe (Rodrigue et al., 2013).

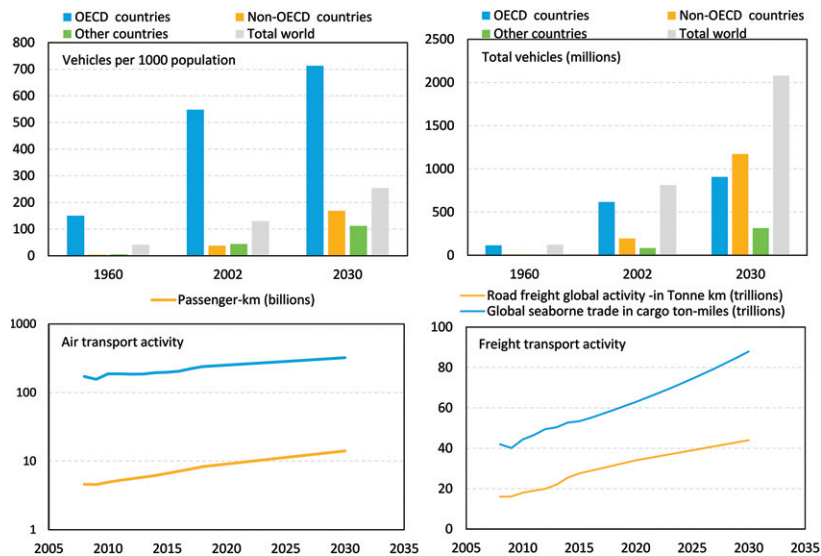


Figure 2 Top left: vehicles per 1000 population all over the world (Dargay et al., 2007); top right: number of vehicles in various countries of the world (Dargay et al., 2007); bottom left: world total air traffic activity (international and domestic) (ICAO, 2017a, 2017c); bottom right: world total freight activity (IEA, 2017c; UNCTAD, 2017; OECD/ITF, 2017).

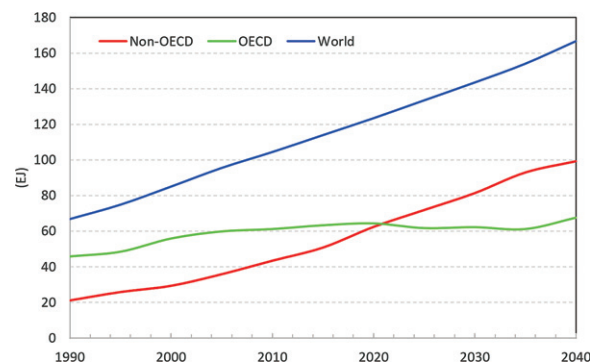


Figure 3 Total transport energy consumption (exajoules) in OECD and non-OECD countries—(projection to 2040) (EIA, 2016).

Overview of Global Energy Use in Transport

According to the Energy Information Administration (EIA, 2016), the energy use of transport sectors has increased substantially in both OECD and non-OECD countries over the last decades (Fig. 3). In 2020, the OECD and non-OECD world transport energy use

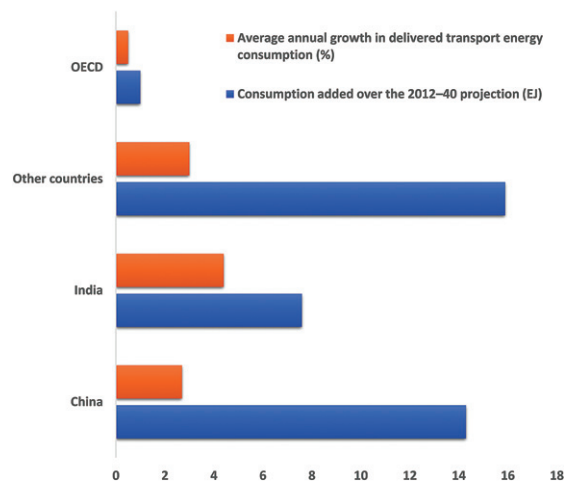


Figure 4 Average annual growth in delivered transportation energy consumption by countries, 2012–40 (percent per year) and energy consumption added over the 2012–40 projection (EIA, 2016).

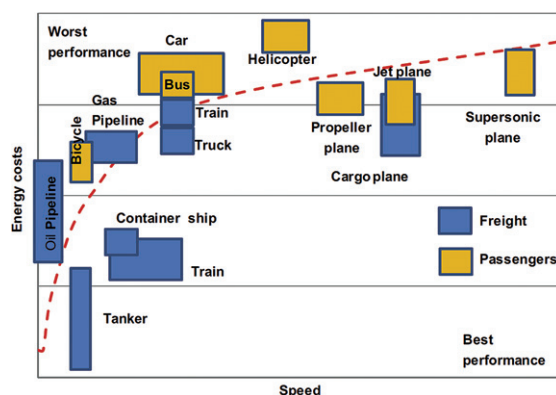


Figure 5 Transportation modes and energy (Samaras and Vouitsis, 2013).

are projected to be equal. The consumption in OECD countries remains relatively flat through the projection period because fuel efficiency improvements keep pace with increases in energy demand.

In non-OECD countries, the energy consumption in transport increased dramatically since 1990 (about 20 EJ) and is projected to reach 100 EJ in 2040. China's consumption of transport fuels is expected to increase on average by 2.7%/year from 2012 to 2040 whereas India also will experience substantial growth, adding eight EJ consumption over the period 2012–40 at an average annual increase of 4.4% (Fig. 4). The other economies of non-OECD also are expected to experience significant growth in transport energy demand from 2012 to 2040, adding 16 EJ consumption at an average annual increase of 3%.

Energy Consumption by Mode

The relationship between transport and energy is a direct one, but subject to different interpretations since it concerns different transport modes, each having a specific performance level. There is often a compromise between speed and energy consumption, related to the desired economic returns (Fig. 5).

Passengers and high-value goods can be transported by rapid but energy intensive modes because the time component of their mobility tends to have a high value. Economies of scale, mainly those achieved by maritime transport, are linked to low levels of energy consumption per unit of mass being transported, but at a lower speed. This fits relatively well with freight transport imperatives, particularly bulk. Comparatively, air freight has high energy consumption levels linked to highspeed services.

Furthermore, energy use in the transport sector has strong modal variations (Rodrigue et al., 2013):

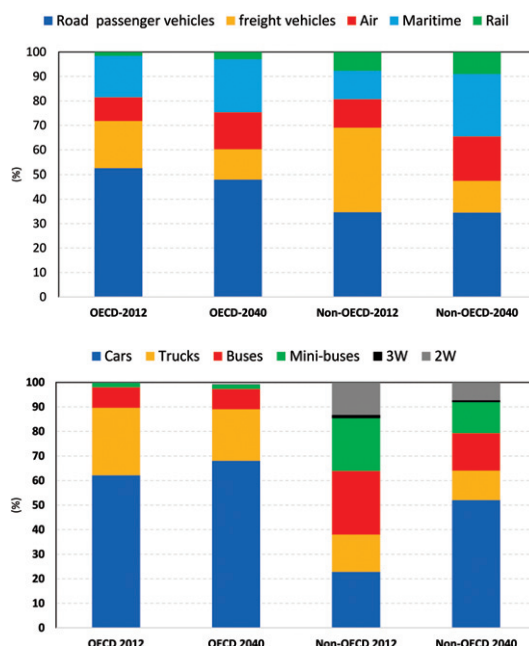


Figure 6 Top: transport energy consumption by mode in OECD and non-OECD countries (2012 and 2040) (EIA, 2016). Bottom: passenger road transport split by vehicle type (2012 and 2040) (GEA, 2012).

Land Transport

Land transport (road, rail, pipelines) accounts for the great majority of energy consumption, using on average 75% of the total transport energy and will continue to weight around 70% up to 2040.

Road transport alone is consuming on average 96% of the total energy used in the land transport. In 2012, road transport accounted for 72% and 60% of the energy consumed in transport in OECD and non-OECD countries, respectively (Fig. 6, top). In OECD countries, road passenger vehicles (cars, light trucks, buses, two- and three-wheel vehicles) accounted for 73% and road freight vehicles for 27% of the land transport energy consumption (Fig. 6, bottom). In non-OECD countries, the corresponding shares were 79% and 21%. In the former, the passenger car dominated, accounting for more than 60% of the total road transport activity. In the latter, road passenger vehicles accounted for 23%, though two, three wheelers and buses also made significant contributions to passenger mobility (15%). Projections for the 2040 suggest that the percentage of energy consumed in road transport will be reduced in both OECD and non-OECD countries (to 60% and 48%, respectively) whereas the contribution of the passenger car in the total road transport activity will be further increased in both country clusters (EIA, 2016).

Rail transport which, on the basis of 1 kg of oil equivalent, is four times more efficient for passenger movement and twice as efficient for freight movement, consumes 2% of the energy used in OECD land transport and 8% of the energy used in non-OECD land transport. There is scope for further improving the energy and emissions performance of rail transport, but most technical measures are only presently cost-effective in new rolling stock (i.e., retrofitting outside of refresh cycles is usually uneconomic). Electric rail is typically 15%–40% more efficient than diesel rail, but only lines that run at least 5 - 10 trains per day are economically suited for electrification. Fig. 7 (top) compares the world railway energy fuel mix (1990 vs. 2015) and Fig. 7 (bottom) shows the sources from which railway-electricity is generated (1990 vs. 2015).

Air Transport

In 2012, the air transport of people and goods accounted for 10% and 12% of the energy consumed in transport in OECD and non-OECD countries, respectively and is expected to grow to 15% and 18%. Thus, in 2040, it will become the second biggest passenger mode after road transport (Fig. 6, top). Passenger air travel demand growth is expected to gain solid momentum in all regions, supported by the improvement in global economic conditions. Underpinned by these conditions that increase import and export orders, air cargo will also increase.

Maritime Transport

In 2012, energy consumption in maritime transport accounted for 17% and 12% of the energy consumed in transport in OECD and non-OECD countries, respectively and is expected to grow to 22% and 24% in 2040 (Fig. 6, top). Maritime transport is essential to



Figure 7 Top: world railway energy fuel mix (%) (1990 vs. 2015); bottom: world railway-electricity by source type (%) (1990 vs. 2015) (IEA, 2017b).

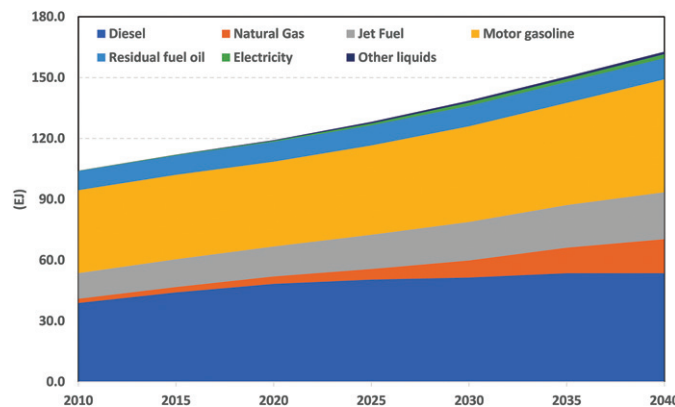


Figure 8 World transport sector delivered energy consumption by fuel type, 2010–40 (exajoules) (EIA, 2016).

the world's economy as over 90% of the world's trade is carried by sea and it is, by far, the most cost-effective way to move goods and raw materials around the world.

Fuel Type by Mode

Worldwide, petroleum and other liquid fuels are the dominant source of transport energy, although their share of total transport energy is projected to decline over the period 2012–40 (from 96% in 2012 to 88% in 2040) (EIA, 2016). Two fuels (motor gasoline (including ethanol blends) and diesel (including biodiesel blends) accounted for 77% of total delivered transport in 2012 and it is projected to account for 70% in 2040. Motor gasoline, which is used primarily for the movement of people remains the largest transport fuel, but its share of total transport energy consumption is expected to decline from 39% in 2012 to 33% in 2040. Diesel (including biodiesel) used primarily for the movement of goods, especially by trucks, shows the largest gain over the period 2012–40 (15 EJ) and jet fuel used in air transport gain more than 10 EJ.

Natural gas share will grow from 3% in 2012 to 11% in 2040 (EIA, 2016). This strong increase is mainly due to natural gas use by large trucks (from 1% in 2012 to 15% in 2040) (OIES, 2016). In addition, 50% of bus energy consumption is projected to be natural gas in 2040, as well as 17% of freight rail, 7% of light-duty vehicles, and 6% of domestic marine vessels. Residual fuel oil (RFO), which is used in marine transport will increase its share from 9% in 2012 to 11% in 2040. Electricity will remain a minor fuel for the world's transport energy use, although its importance in rail transport will remain high - particularly for passenger routes with high traffic volumes. The electricity share of total light-duty vehicle energy consumption grows to 1% in 2040, as new plug-in electric vehicles increasingly penetrate the total light-duty stock.

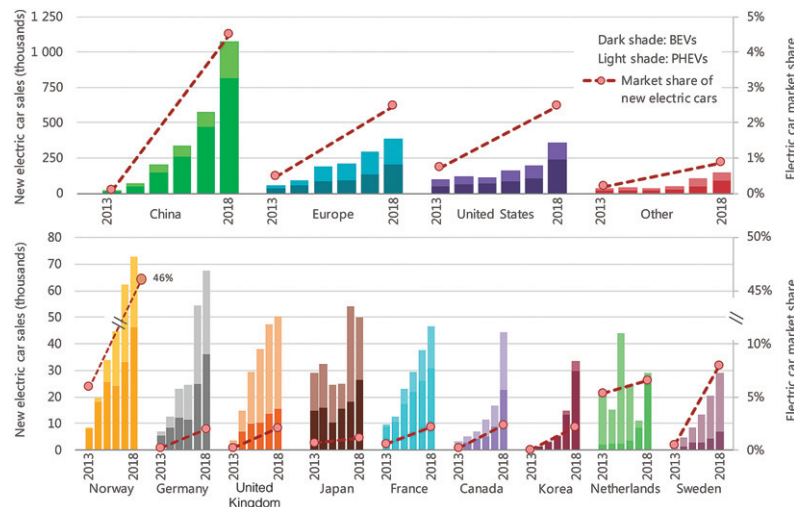


Figure 9 Global electric car sales and market share, 2013–18. (IEA, 2019). Note: the countries in this figure represent the top-ten countries. This ranking closely resembles the 10 leading countries worldwide in term of electric car sales; the only exception is Korea, with 33,000 electric car sales in 2018. Other includes Australia, Brazil, Chile, India, Japan, Korea, Malaysia, Mexico, New Zealand, South Africa, and Thailand. Europe includes Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Netherlands,

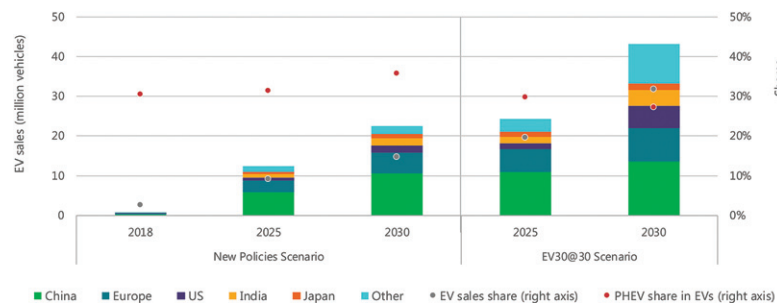


Figure 10 Future global EV sales by scenario, 2018–30 (IEA, 2019). Note: the New Policies Scenario aims to illustrate the impact of announced policy ambitions; the EV30@30 Scenario takes into account the pledges of the Electric Vehicle Initiative's EV30@30 Campaign to reach a 30% market share for EVs in all modes except two-wheelers by 2030 (PLDVs = passenger light-duty vehicles; LCVs = light-commercial vehicles; BEV = battery electric vehicle; PHEV = plug-in hybrid vehicle).

However, EIA' projections for electricity in the light duty vehicle fleets appear rather conservative as indicated in IEA's recent market analysis (Fig. 9) (IEA, 2019). Global electric car sales were close to 2 million in 2018, after having reached the 1 million mark in 2017. This is related with the continuing fall of diesel models' sales as the fuel type falls out of favor with customers. Furthermore, IEA's projections indicate a rising tide of electric vehicles (Fig. 10) (IEA, 2019).

GHGs Emissions

Globally, CO₂ accounts for 95% of transportation GHGs. HFCs, which are used in automobile, aircrafts, truck, and rail air conditioning and refrigeration systems, account for another 3% of transport GHG emissions. N₂O and CH₄, which are both emitted as byproducts of combustion, account for the remainder of the transport GHG emissions inventory. Water emissions at the Earth's surface have no climate effect. If released at cruise altitude by commercial aircraft, however, water vapor emissions can be more significant (Samaras and Vouitsis, 2013).

According to IEA (IEA, 2018), the global transport sector produced 6.62 GtCO₂ emissions in 2016, and it was the second larger emitter after the power generation sector (13.41 GtCO₂), contributing about 21% of the global GHG emissions (Fig. 11, top). The Americas and Asia contribute 38% and 37%, respectively, Europe contributes 19% and Africa/Oceania 7%. North America and Oceania have disproportionately high emissions relative to their population (Fig. 11, bottom).

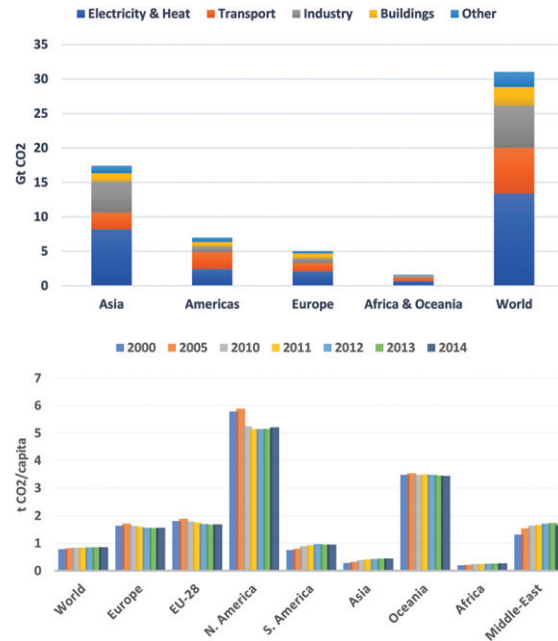


Figure 11 Top: CO₂ emissions by sector for selected regions, 2016 (IEA, 2018). Left: CO₂ emissions of transport per capita for selected regions, 2000–14 (WEC, 2014).

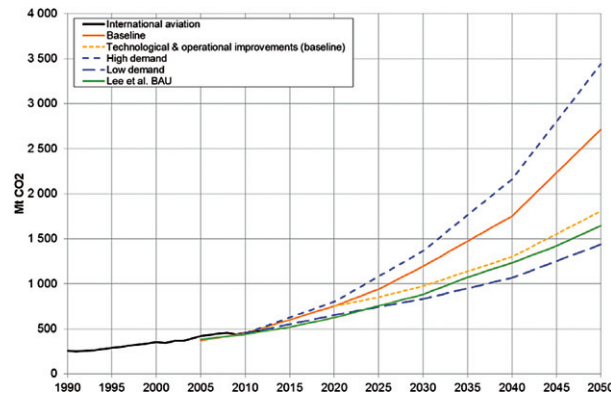


Figure 12 Global CO₂ emissions from aviation under various scenarios, 1990–2050.

Road Transport

Within road transport, automobiles and light trucks produce well over 60% of emissions, but in low- and middle-income developing countries, freight trucks (and in some cases even buses) consume more fuel and emit more CO₂ than the automobiles and light trucks (Samaras and Vouitsis, 2013). Although there is yet no global-gridded rail inventory available, this sector is certainly the lowest CO₂ emitter (in terms of “tailpipe” emissions at least) because it has been to a large extent electrified. Available studies show that rail emissions make up only 1%–3% of the total transport emissions (IEA, 2017b).

Air Transport

Although the average fuel burn of new aircraft fell approximately 45% from 1968 to 2014 (Kharina and Rutherford, 2015), the total amount of fuel burned has still increased due to even higher growth in air traffic. According to a recent study by the International Council on Clean Transportation (Graver et al., 2019), worldwide aviation CO₂ emissions increased by 32% from 2013 to 2018. The total increase over these five years was equivalent to building about 50 coal-fired power plants, according to the study. The implied annual compound growth rate of emissions, 5.7%, is 70% higher than those used to develop International Civil Aviation Organization’s (ICAO) projections that CO₂ emissions from international aviation will triple under business as usual scenario by 2050. Given the sustained growth of the air transport industry, aircraft emissions are expected to increase in the short-term future.

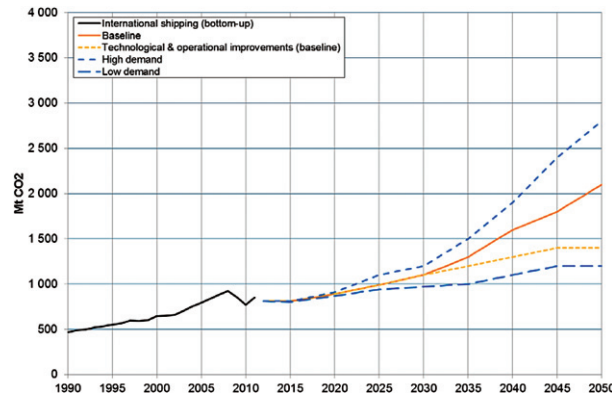


Figure 13 Global CO₂ emission projections for international maritime transport (ENVI, 2015).

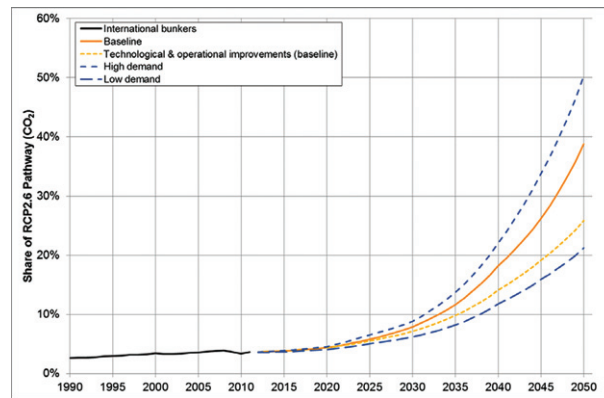


Figure 14 International aviation and maritime transport's share of global GHG emissions under the Representative Concentration Pathway (RCP) 2.6 pathway (ENVI, 2015).

Although technological and operational improvements have a potential to reduce CO₂ emissions by 33% in 2050 compared to the baseline, CO₂ emissions are still projected to be seven times higher in 2050 than in 1990; without the improvements, projections are ten times above 1990 levels in 2050 (Fig. 12).

Maritime Transport

Estimates for the global CO₂ emissions for maritime transport differ substantially depending on the approach (bunker fuel statistics or activity-based figures). Fig. 13 depicts International Maritime Organization (IMO)'s projections under three different scenarios as well as the effect of the most ambitious mitigation option on the reference scenario (ENVI, 2015). As in the case of air transport, the technological and operational improvements reduce CO₂ emissions in 2050 by 33%. Other studies show a similar range of baseline emissions and upper/lower bounds. Despite this, they are still projected to increase by a factor of almost four compared to 1990. Without these improvements, the growth would be sixfold.

It is worth mentioning that initiatives and actions taken by ICAO and IMO to address GHG emissions started late and so far, have been insufficient from an environmental perspective. If the ambition of the two sectors continues to fall behind efforts in other sectors and if action to combat climate change is further postponed, their CO₂ emission shares in global CO₂ emissions may rise substantially to 22% for international aviation and 17% for maritime transport by 2050, or almost 40% of global CO₂ emissions if both sectors are considered together (Fig. 14).

Reducing Transport Energy Consumption

Regulations

Land Transport

Due to various historical, cultural, and political reasons, different countries and regions have chosen to adopt different fuel economy (FE) or emission standards (ES). These standards differ in stringency, by their apparent forms and structures and by how the vehicle

Table 1 Overview of regulation specifications for passenger cars (Yang and Bandivadeka, 2017)

Country/region	Regulated metric	Unadjusted Fleet target/measure (target year)	Form of target curve	Test cycle
EU-28	CO ₂	95 gCO ₂ /km (2021)	Weight-based corporate average ¹	NEDC and WLTC ^a
USA	Fuel economy/GHG	55.2 mpg and 147 gCO ₂ /mi (2025)	Footprint (size)-based corporate average	US combined ^b
China	Fuel consumption	5 L/100 km (2020)	Weight-class based corporate average	NEDC
India	CO ₂	113 g/km (2022)	Weight-based corporate average	NEDC
Japan	Fuel economy	20.3 km/L (2020)	Weight-class based corporate average	JC08 ^c
Saudi Arabia	Fuel economy	17 km/L (2020)	Footprint-based corporate average	US combined
Canada	GHG	217 gCO ₂ /mi ^d (2016) N/A ^e (2025)	Footprint-based corporate average	US combined
Mexico	Fuel economy/GHG	39.3 mpg or 140 g/km (2020)	Footprint-based corporate average	US combined
South Korea	Fuel economy/GHG	24 km/L or 97 gCO ₂ /km (2020)	Footprint-based corporate average	US combined

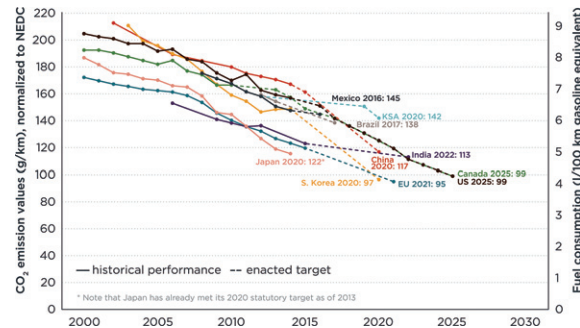
^aNEDC: New European Driving Cycle, WLTC: Worldwide harmonized Light vehicles Test Cycles.

^bFTP-75 cycle and Highway Fuel Economy Test cycle.

^cJapanese JC08 Cycle.

^dIn April 2010, Canada announced a target for its LDV fleet of 246 g/mi for model year 2016.

^eCanada follows the US standards in the proposal, but the final target value would be based on the projected fleet footprints.

**Figure 15** Historical fleet CO₂ emissions performance and current standards (gCO₂/km normalized to NEDC) for passenger cars (Yang and Bandivadeka, 2017).

FE or GHG emission levels are measured—that is, by testing methods. They also differ by implementation requirements, such as mandatory versus voluntary approaches. Thus, policymakers are faced with many choices when they decide GHG emission/fuel economy standards. **Table 1** summarizes the basic policy approaches adopted by various regions for passenger cars and it is an illustrative example of the above-mentioned differences on interpretation and approach. **Fig. 15** compares country standards for passenger vehicles in terms of grams of CO₂-equivalent/km adjusted to the European NEDC test cycle.

The EU-28 has the lowest fleet average target of 95 gCO₂/km by 2021. However, South Korea will match, if not exceed, the EU-28 with a fleet target of 97 gCO₂/km in 2020. With its high hybrid percentage, Japan already reached its 2015 target of 142 g/km in 2011 and 2020 target of 122 g/km in 2013. If Japan continues to reduce CO₂ emissions at the same rate as from 2010 to 2014, Japan's passenger vehicle fleet would achieve 82 g/km in 2020, far below the targets set by other countries. The United States and Canada, long laggard in regulating fuel economy, have evolved into leaders. As the first country with 2025 targets, the US example has encouraged other countries to consider enacting similarly long-term standards. The United States is expected to achieve 45% GHG emission reduction from 2010 to 2025. China announced a fleet average fuel consumption of 4L/100 km by 2025, which would be among the lowest target levels once it is supported by detailed standards. However, a gap between official test results and real-world performance of new cars' CO₂ emissions has grown alarmingly, it dilutes the fuel economy standards and makes real-world fuel consumption an issue that needs to be addressed (Yang and Bandivadeka, 2017).

Air and Maritime Transport

Globally, the majority of GHG emissions from the air and maritime transport are still unregulated. While some countries have enacted domestic policies, most have not. Independently, New Zealand, Australia, and the EU-28 have already taken steps to include aviation in their domestic GHG cap-and-trade programs. The EU-28 Emissions Trading System (EU ETS) is the cornerstone of the EU's policy to reduce, in a cost-effective manner, GHG emissions from aviation among others (EC, 2013). The ETS is a cap-and-trade system, which sets a limit on the number of emissions allowances issued, and thereby constrains the total amount of emissions of the sectors covered by the system. The gradually more limited supply of allowances drives operators in need of additional allowances to

Table 2 Practical limit to energy efficiency improvements by reducing road load

<i>Transport mode</i>	<i>Practical limit (%)</i>
Cars	91
Trucks	54
Rail	57
Air	46
Maritime	63

buy them on the market from other sectors in the system—hence cap-and-trade system. Between 2013 and 2020, an estimated net saving of 193.4 MtCO₂ will be achieved by air transport via the EU ETS through funding of emissions reduction in other sectors (EEA, 2019).

As mentioned above, addressing GHG emissions from international aviation and marine shipping is especially challenging, because they are produced along routes where no single nation has regulatory authority. Unlike other sources of GHG emissions, international emissions from aviation and marine transport are specifically excluded from developed countries' national targets. Instead, limitations or reductions in emissions from these transport modes could be achieved by working through ICAO and IMO. In response to this mandate, both organizations have initiated activities aimed at addressing emissions from their respective sectors, but thus far nothing substantial has been achieved. In 2017, the ICAO Council adopted a new aircraft CO₂ emissions standard which presents the world's first global design certification standard governing CO₂ emissions for any industry sector (ICAO, 2017b). The standard will apply to new aircraft type designs from 2020, and to aircraft type designs already in-production as of 2023. Those in-production aircraft which by 2028 do not meet the standard will no longer be able to be produced unless their designs are sufficiently modified. IMO's "Initial Strategy" (IMO, 2018) envisages for the first time a strategy which aims to reduce the total annual GHG emissions by at least 50% by 2050 compared to 2008.

Technology

Technologically, the reduction of the energy used in transport can be accomplished by two ways: either by engine efficiency and transmission improvements or by reducing road load, that is, the mass of the vehicles and the resistances to motion. Table 2 summarizes the practical limit to energy efficiency improvements for each transport mode (Cullen et al., 2007). The largest savings may occur in passenger cars (91%), where reducing the mass of the car and the rolling resistance are important. Cars are also less optimized because of the diversity of models and the need to compromise efficiency to satisfy the fashion element of car design. Airplane, where the weight of any additional fuel must be supported by the lift force generated by the engine thrust and wings, is the most optimized of the systems with only 46% practical energy savings available. Light weighting strategies for ships and trains are limited because the transported goods make up a large proportion of the total vehicle mass (typically 60% in trucks versus 5%–10% in cars).

More specifically, the technology measures to improve energy efficiency, include:

Road Transport

Road transport is dominated by the internal combustion engine and the technology measures to improve energy efficiency, include (Samaras and Voutsis, 2013):

- Aggressive electrification of the car fleet (hybrid, battery electric and fuel cell vehicles);
- Significant downsizing, including a change of the engine architecture to the next smaller engine family plus turbocharging;
- Stratified part-load operation with key enabler the spray guided stratified combustion, with its best possible fuel economy improvement versus homogeneous operation;
- Natural gas, synthetic diesel, and biofuels;
- Friction reduction and thermal management, as well as high compression ratio enabled by homogeneous direct injection;
- Increasing vehicle loading rates in freight transport.

Air Transport

The need for air transport to be more fuel efficient will be driven in the coming years both by fuel prices and environmental taxes or caps. Past trends in fuel efficiency vary considerably depending on the measure and the period chosen (Singh and Sharma, 2015). For the period 1959–95, the reduction in energy intensity was mainly improvement in engine efficiency (57%), aerodynamic efficiency (22%), aircraft capacity (17%), and other changes such as increased aircraft size (4%). For the period 1970–2006, improvements were mainly due to improvements in load factor (20%), aircraft size (26%), and finally technical and operational improvements to the fleet (24%). Fig. 16 summarizes the modeled trend in average new aircraft fuel burn from 1960 to 2014 on ICAO's Metric Value (green line) and a fuel/passenger-km metric (blue line), normalized to 1968. Beyond technological advances, the fuel efficiency of air transport can be improved by operation measures, and management of air traffic.

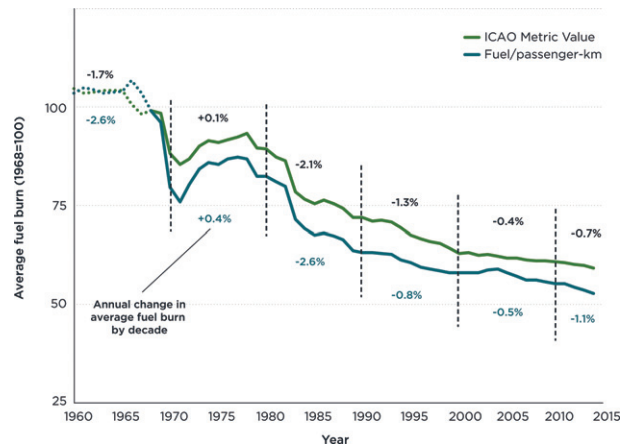


Figure 16 Average fuel burn for new commercial jet aircraft, 1960–2014 (1968 = 100) (Kharina and Rutherford, 2015). ICAO's MV is a proxy of cruise fuel efficiency that is calculated as function of an aircraft's specific air range (SAR) and Reference Geometric Factor (RGF), which is a close approximation of the pressurized floor area of the aircraft.

Maritime Transport

Low speed two-stroke turbocharged diesel engines are the most commonly used for maritime propulsion today. These engines are the most efficient, exhibiting 50% efficiency. Technological measures for increased efficiency may be either those which can be introduced only at the design/building stage or those which can be retrofitted to existing ships (IMO, 2014). A combination of technical measures could reduce total energy use by 4%–20% in older ships and 5%–30% in new ships by applying state-of-the-art knowledge, such as hull and propeller design and maintenance (GEA, 2012). Operational measures, such as speed reduction, offer a large and near-term mitigation option, while improving the energy efficiency of new ships and switching to alternative fuels provide longer-term potential. Reducing the speed at which a ship operates brings significant benefits in terms of lower energy use. For example, cutting a ship's speed from 26 to 23 knots can yield a 30% fuel saving.

Behavioral Changes

The main target is to reduce congestion through encouraging behavioral change towards more sustainable modes of transport. Increasing occupancy rates in passenger transport is an attractive strategy because it can be achieved with the existing public transport or private vehicle stock. Although very difficult to implement in private transport, ride sharing applications would reduce the numbers of vehicles on the road and vehicle-km driven; when differences in occupancy rates are taken into account, public transport modes are usually much more energy efficient than car travel, as can be seen most clearly by comparing bus and car travel, both of which use petroleum fuels (EEA, 2015). Other actions for travel behavior change include, among others, parking policy, limited traffic zones, traffic light timing, job ticket, upgrade of sidewalk infrastructure, development of cycling infrastructure.

Furthermore, it is expected that Cooperative Intelligent Transport Systems (C-ITS) and Connected and Automated Vehicles (C&AV) will have a very important impact on driver behavior. The current phase of development in C-ITS and C&AV has already been considered the biggest innovation in automotive industries since the invention of the car itself. It could help to significantly reduce road fatalities from road accidents. New transport services may also emerge, especially when vehicles are provided with connectivity in addition to automation (e.g., traffic safety related warnings, traffic management, car sharing, new possibilities for elderly or impaired people). Finally, the deployment of driverless mobility—when fully integrated in the whole transport system and accompanied by the right support measures and synergies between driverless mobility and decarbonization measures—is expected to contribute significantly to achieving key societal objectives (EC, 2018)

Summary and Conclusions

Given the huge demand for both vehicle ownership and transport in the non-OECD countries, a strong case can be made for a continuation of high growth in transport energy demand. Current trends in urban developments are a reason of concern in terms of climate, because they are allowing private cars to gain dominance over public transport or nonmotorized travel. No matter if the mode share of automobiles is low, as in emerging economies, or very high, as in the developed world, the trend is clear: people are buying more cars every day and once they have them, they use them. Consequently, the energy use (and GHG emissions) from urban passenger transport continues to increase.

Rail transport, which is the most efficient passenger mode, should be considered as an effective way to make transport more carbon efficient. Air transport has improved its fuel efficiency significantly over the years and technology developments might offer a further

improvement in the future. However, since aviation's growth rate is projected to be high, such improvements may not be enough to keep energy use from increasing. The available maritime transport technology can cost-effectively achieve high efficiency gains.

Achieving sustainability goals in transport within the limited time frame of a few decades requires transformative actions and political commitment from the local communities to the supranational and global level. These include:

Road Transport

- Full electrification of passenger cars—improvement of the range of electric vehicles will be a key success factor the coming years; it will also be important to develop short range, up to 150 km, low cost electric vehicles which will be very effective in city environments; expansion of electrification to urban buses and to heavy duty sector;
- Improving the performance of next generation fuel cell electric vehicles (FCEV);
- Renewable fuels and further energy efficiency improved mono-fuel natural gas engines and direct injection.

Rail Transport

- Full electrification;
- Lighter materials in vehicles and wider use of energy recuperation devices (e.g., regenerative braking or energy storage technologies).

Air Transport

- Innovative aircraft technology and improvements of existing technologies (e.g., better aerodynamics, jet engine efficiency);
- Development and application of sustainable bio-kerosene sourcing production and distribution;
- Hydrogen-powered fuel cells;
- Improving air traffic management technologies, procedures and systems.

Maritime Transport

- Optimization of conventional ship engines, including fuel flexibility, new materials, lifetime performance and near zero emissions engines;
- Development of standardization in order to ease certification of such engines;
- Market based measures.

Biography



Dr Zissis Samaras is a full professor and director of the Lab of Applied Thermodynamics (LAT), Department of Mechanical Engineering, Aristotle University, Thessaloniki, Greece. His research work deals primarily with engine and vehicle emissions testing and modeling and he has carried out a wide range of projects on modeling emissions from internal combustion engines. He has provided expert advice to several organizations and private sector customers, including the European Commission, the European Environment Agency, the World Bank the Greek Ministry of Environment, Rhône-Poulenc, ACEA, Concawe, Toyota Motor Europe. He is elected academic member and vice chairman of the European Road Transport Research Advisory Council (ERTRAC) on "Energy, Environment and Resources." He is the author of more than 160 papers in peer-reviewed journals and seven book chapters, which received according to Scopus more than 4500 citations in peer reviewed articles, reviews, and technical notes.



Dr Ilias Vouitsis is a senior researcher at the Lab of Applied Thermodynamics (LAT), Department of Mechanical Engineering, Aristotle University, Thessaloniki, Greece. His research interests lie in the areas of energy conversion and exhaust emissions with emphasis on particle emissions. He has experience in development of emission factors for light duty, heavy duty vehicles and motorcycles. He has been actively involved in several major research projects funded by both the European Union and the industry on emission related legislation as well as on automotive product development. He has been the author and coauthor of over 30 scientific and technical papers and of three book chapters.

References

- Cullen, J.M., Allwood, J.M., Borgstein, E.H., 2007. Reducing energy demand: what are the practical limits? *Environ Sci. Technol.* 1711–1718.
- Dargay, J., Gately, D., Sommer, M., 2007. Vehicle ownership and income growth, worldwide 1960–2030. *Energy J.* 28 (4), 143–170.
- EC (European Commission), 2003. Directive 2003/87/EC of the European Parliament and of the Council of 13 October 2003 Establishing a Scheme for Greenhouse Gas Emission Allowance Trading within the Community and Amending Council Directive 96/61/EC. Retrieved from: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32003L0087>.
- EC, 2011. White Paper Roadmap to a Single European Transport Area—Towards a Competitive and Resource Efficient Transport System. European Commission. COM 2011, Brussels. Retrieved from: https://ec.europa.eu/transport/sites/transport/files/themes/strategies/doc/2011_white_paper/white-paper-illustrated-brochure_en.pdf.
- EC, 2018. On the Road to Automated Mobility: An EU Strategy for Mobility of the Future. Retrieved from: https://ec.europa.eu/transport/sites/transport/files/3rd-mobility-pack/com20180283_en.pdf.
- EEA (European Environmental Agency), 2015. Is Passenger Transport becoming More Efficient? Retrieved from: <https://www.eea.europa.eu/data-and-maps/indicators/occupancy-rates-of-passenger-vehicles/occupancy-rates-of-passenger-vehicles>.
- EEA, 2019. European Aviation Environmental Report 2019. Retrieved from: <https://ec.europa.eu/transport/sites/transport/files/2019-aviation-environmental-report.pdf>.
- EIA (Energy Information Administration), 2016. Transportation sector energy consumption. In: *International Energy Outlook 2016—With Projections to 2040*. Retrieved from [https://www.eia.gov/outlooks/ieo/pdf/0484\(2016\).pdf](https://www.eia.gov/outlooks/ieo/pdf/0484(2016).pdf) (Chapter 8).
- ENVI (Committee on the Environment, Public Health and Food Safety), 2015. Emission Reduction Targets for International Aviation and Shipping. Retrieved from: [https://www.europarl.europa.eu/RegData/etudes/STUD/2015/569964/IPOL_STU\(2015\)569964_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2015/569964/IPOL_STU(2015)569964_EN.pdf).
- GEA (Global Energy Assessment), 2012. Energy End-use: transport. In: Johansson, Th.B., Patwardhan, A., Nakicenovic, N., Gomez-Echeverri, L. (Eds.), *Global Energy Assessment — Toward a Sustainable Future*. Cambridge University Press, Cambridge, UK and New York, NY, USA and the International Institute for Applied Systems Analysis, Laxenburg, Austria. Retrieved from: http://www.iiasa.ac.at/web/home/research/Flagship-Projects/Global-Energy-Assessment/GEA_Chapter9_transport_hires.pdf (Chapter 9).
- Graver, B., Zhang, K., Dan Rutherford, D., 2019. CO₂ Emissions from Commercial Aviation, 2018. ICCT. Retrieved from: https://theicct.org/sites/default/files/publications/ICCT_CO2-commercl-aviation-2018_20190918.pdf.
- Gross, M., 2016. A planet with two billion cars. *Curr. Biol.* 26 (8), R307–R310.
- ICAO (International Civil Aviation Organization), 2017a. 2017 Air Transport Statistical Results. Retrieved from: https://www.icao.int/annual-report-2017/Documents/Annual-Report2017_Air%20Transport%20Statistics.pdf.
- ICAO, 2017b. ICAO Council Adopts New CO₂ Emissions Standard for Aircraft. Retrieved from: <https://www.icao.int/Newsroom/NewsDoc2017/COM.05.17.EN.pdf>.
- ICAO, 2017c. ICAO — State of Global Air Transport and ICAO Forecasts for Effective Planning, Ms. Sijia Chen, Aviation Data Analysis Officer, Economic Development, ICAO, p. 25 and p. 26. Retrieved from: <https://www.icao.int/Meetings/ICAN2017/Pages/Presentations.aspx>.
- IEA (International Energy Agency), 2017a. The Future of Trucks. Implications for Energy and the Environment. OECD/IEA, 2017. Retrieved from: <https://www.iea.org/publications/freepublications/publication/TheFutureofTrucksImplicationsforEnergyandtheEnvironment.pdf>.
- IEA, 2017b. Railway Handbook 2017. Energy Consumption and CO₂ Emissions. Retrieved from: https://uic.org/IMG/pdf/handbook_iea-uic_2017_web3.pdf.
- IEA, 2017c. The Future of Trucks. Implications for energy and the environment. Retrieved from: <https://www.iea.org/publications/freepublications/publication/TheFutureofTrucksImplicationsforEnergyandtheEnvironment.pdf>.
- IEA, 2018. CO₂ Emissions from Fuel Combustion 2018 Highlights. Retrieved from: <https://www.iea.org/statistics/co2emissions/>.
- IEA, 2019. Global EV Outlook. Retrieved from: http://www.indiaenvironmentportal.org.in/files/file/Global_EV_Outlook_2019.pdf.
- IMO (International Maritime Organization), 2014. Third IMO Greenhouse Gas Study. 2014. Retrieved from: <http://www.imo.org/en/OurWork/Environment/PollutionPrevention/AirPollution/Documents/Third%20Greenhouse%20Gas%20Study/GHG3%20Executive%20Summary%20and%20Report.pdf>.
- IMO, 2018. Initial IMO strategy on reduction of GHG emissions from ships. Retrieved from: http://www.imo.org/en/OurWork/Environment/PollutionPrevention/AirPollution/Documents/Resolution%20MEPC.304%2872%29_E.pdf.
- Kharina, A., Rutherford, R., 2015. Fuel Efficiency Trends for New Commercial Jet Aircraft: 1960 to 2014. ICCT (International Council on Clean Transportation). Retrieved from: https://theicct.org/sites/default/files/publications/ICCT_Aircraft-FE-Trends_20150902.pdf.
- Mulholland, E., Teter, J., Cazzola, P., McDonald, Z., Ó Gallachóir, B.P., 2018. The long haul towards decarbonising road freight—a global assessment to 2050. *Appl. Energy* 216, 678–769.
- OECD/ITF (International Transport Forum), 2017. ITF Transport Outlook 2017, OECD Publishing, Paris. <http://dx.doi.org/10.1787/9789282108000-en>. Retrieved from: https://www.ttm.nl/wp-content/uploads/2017/01/itf_study.pdf.
- OECD/ITF, 2018. Decarbonising Maritime Transport. Pathways to Zero-carbon Shipping by 2035. May 2018. Retrieved from: <https://www.itf-oecd.org/decarbonising-maritime-transport>.
- OIES (Oxford Institute for Energy Studies), 2016. A Review of Prospects for Natural Gas as a Fuel in Road Transport. Retrieved from: <https://www.oxfordenergy.org/wpcms/wp-content/uploads/2019/04/A-review-of-prospects-for-natural-gas-as-a-fuel-in-road-transport-Insight-50.pdf>.
- Rizet, C., Cruz, C., Mbacké, M., 2012. Reducing freight transport CO₂ emissions by increasing the load factor. *Procedia Soc. Behav. Sci.* 48, 184–195.
- Rodrigue, J.-P., Comtois, C., Slack, S., 2013. The Geography of Transport Systems. Routledge, N.Y., p. 16. Retrieved from: https://transportgeography.org/?page_id=322.
- Samaras, Z., Vouitsis, I., 2013. Transportation, energy. In: Pielke, R.A. (Editor-in-Chief), Adegoke, J., Wright, C.Y. (Vol. Eds.), *Climate Vulnerability, Understanding, Addressing Threats to Essential Resources*, Vol. 1: Vulnerability of Human Health to Climate. Elsevier Inc., Academic Press, pp. 183–205.

- Singh, V., Sharma, S.K., 2015. Fuel consumption optimization in air transport: a review, classification, critique, simple meta-analysis, and future research implications.. *Eur. Transp. Res. Rev.* 7–12.
- UNCTAD (United Nations Conference on Trade and Development), 2017. Review of Maritime Transport 2017. Retrieved from: https://unctad.org/en/PublicationsLibrary/rmt2017_en.pdf.
- WEC, 2014. World Energy Council (WEC). Energy Efficiency Indicators. Retrieved from: <https://wec-indicators.enerdata.net/transport-co2-emissions-per-capita.html>.
- Yang, Z., Bandivadeka, A., 2017. 2017 Global update: Light-duty Vehicle Greenhouse Gas and Fuel. 2017. ICCT. Retrieved from: <https://www.theicct.org/publications/2017-global-update-LDV-GHG-FE-standards>.

Further Reading

- D'Agosto, M., 2019. Transportation, Energy Use and Environmental Impacts. Elsevier, Netherlands. Chapter 3: Transportation planning and energy use, pp. 85–122; Chapter 4: Propulsion systems and energy use, pp. 123–176; Chapter 5: Energy sources for transportation, pp. 177–226.
- EC, 2011. White Paper Roadmap to a Single European Transport Area—Towards a Competitive and Resource Efficient Transport System. European Commission. COM 2011, Brussels. Available at: https://ec.europa.eu/transport/sites/transport/files/themes/strategies/doc/2011_white_paper/white-paper-illustrated-brochure_en.pdf.
- Eyring, V., Isaksen, S.A.I., Bernsten, T., et al., 2010. Transport impacts on atmosphere and climate: shipping. *Atm. Environ.* 44 (37), 4735–4771.
- GEA (Global Energy Assessment), 2012. Energy end-use: transport. In: Johansson, Th.B., Patwardhan, A., Nakicenovic, N., Gomez-Echeverri, L. (Eds.), *Global Energy Assessment – Toward a Sustainable Future*. Cambridge University Press, Cambridge, UK and New York, NY, USA and the International Institute for Applied Systems Analysis, Laxenburg, Austria (Chapter 9). Available at: http://www.iiasa.ac.at/web/home/research/Flagship-Projects/Global-Energy-Assessment/Chapters_Home.en.html.
- Lee, D.S., Pitari, G., Grewe, V., et al., 2010. Transport impacts on atmosphere and climate: aviation. *Atm. Environ.* 44 (37), 4678–4734.
- McCollum, D., Gould, G., Greene, D., 2009. Greenhouse Gas Emissions from Aviation and Marine Transportation: Mitigation Potential and Policies. Institute of Transportation Studies, UC Davis. Available at: https://www.researchgate.net/publication/46439904_Greenhouse_Gas_Emissions_from_Aviation_and_Marine_Transportation_Mitigation_Potential_and_Policies.
- Neves, S.A., Marques, A.C., Fuinhas, J.A., 2017. Is energy consumption in the transport sector hampering both economic growth and the reduction of CO₂ emissions? A disaggregated energy consumption analysis. *Transp. Policy* 59, 64–70.
- Rodrigue, J.-P., Comtois, C., Slack, S., 2013. Transport, energy and environment. *The Geography of Transport Systems*. Routledge, N.Y., pp. 255–279 (Chapter 8). Available at: <https://transportgeography.org/>.
- Samaras, Z., Vouitsis, I., 2013. Transportation and energy. In: Pielke, R.A. (Editor-in-Chief), Adegoke, J., Wright, C.Y. (Vol. Eds.), *Climate Vulnerability: Understanding and Addressing Threats to Essential Resources*, Vol. 1: Vulnerability of Human Health to Climate. Elsevier, Netherlands, pp. 183–205.
- Uherek, E., Halenka, T., Borken-Kleefeld, J., et al., 2010. Transport impacts on atmosphere and climate: land transport. *Atm. Environ.* 44 (37), 4772–4816.
- Wang, T., Lin, B., 2019. Fuel consumption in road transport: a comparative study of China and OECD countries. *J. Clean. Prod.* 206, 156–170.
- Yang, Z., Bandivadeka, A., 2017. 2017 Global Update: Light-duty Vehicle Greenhouse Gas and Fuel. ICCT (International Council on Clean Transportation). Available at: <https://www.theicct.org/publications/2017-global-update-LDV-GHG-FE-standards>.
- Yeh, S., Mishra, G.S., Fulton, L., et al., 2017. Detailed assessment of global transport-energy models' structures and projections. *Transport. Res. D-Transp. Environ.* 55, 294–309.

Transport Modes and People With Limited Mobility

Roger L Mackett, Centre for Transport Studies, University College London, London, United Kingdom

© 2019 Elsevier Ltd. All rights reserved.

Limited Mobility	85
Transport Modes Used by Disabled and Elderly People	85
Barriers to Transport Modes Used by Disabled and Elderly People	87
Interventions to Increase Modal Usage by Older and Disabled People	90
Conclusions	90
Biography	91
References	91
Further Reading	91

Limited Mobility

This article focuses on the factors that affect the choice of transport mode by people who have limited mobility, and the barriers that prevent them from using some modes as much as other people. There are several ways in which mobility can be limited, as [Table 1](#) shows.

Limitations on mobility can be permanent or temporary. Any traveler can have a temporary mobility limitation such as heavy baggage. Young children have temporary limitations on their mobility because they cannot walk as far or as fast as an adult and do not have the mental capacity or knowledge to use public transport alone but will acquire these capabilities as they grow up. Escorting young children can place limitations on the mobility of adults.

Elderly people may lose their ability to walk quickly, climb stairs or drive safely because of the gradual deterioration in their faculties. This group overlaps with people with disabilities because many disabilities occur in older age.

[Table 2](#) shows the effects of having a disability on mobility. The range of disabilities indicated in the table are the categories of disability used in surveys carried out for the British Government ([Office for National Statistics, 2017](#)). In Britain, a disability is defined as a health condition expected to last more than 12 months which reduces the ability to carry out day-to-day activities ([Office for Disability Issues, 2011a](#)). On average, the population makes 1012 trips a year. Those with a disability make 853 trips on average, while people without a disability make 1055. There is a range across the various types of disability from 594 for people with social/behavioral conditions to 787 for people with stamina, breathing and fatigue conditions and 930 for other disabilities. This illustrates that nonvisible disabilities may affect mobility. For more information about travel by disabled people see [Article 10750](#) and about travel by older people see [Articles 10163](#) and [10614](#).

There are similar definitions of disability related to mobility used in many countries. For example, in the United States, the National Household Travel Survey asks respondents if they have "a temporary or permanent condition or handicap that makes it difficult to travel outside of the home" ([Federal Highway Administration, 2020](#)).

This article focuses on people with permanent mobility limitations because temporary mobility limitations are unlikely to have an adverse long-term effect on people's lives.

The structure of this article is as follows: the next section describes the modes of transport used by disabled and elderly people. Then the barriers that prevent them from using some modes as much as others are discussed, followed by a brief discussion about interventions that may increase travel by those modes.

Transport Modes Used by Disabled and Elderly People

[Table 3](#) shows the trips made by each mode by disabled and nondisabled people in England. The most popular mode for both groups is the car or van. Combining drivers and passengers, disabled people make 61.5% of their trips by this mode compared with 62.4% by others. More disabled people are car passengers than nondisabled people. Despite this, many more car trips by disabled people are as drivers than as passengers. Combining the public transport modal shares of bus, rail, and taxi, there is little difference between the two groups: 9.9% of trips compared to 10.2%. Disabled people use bus and taxi more than other people, whereas nondisabled people use rail more, particularly the London Underground, a metro system. This difference may reflect the fact that nondisabled people are more likely to be employed and use rail to travel to work. Turning to active travel, disabled people make fewer walking trips than nondisabled people, but walk a greater proportion of their trips. They cycle less than nondisabled people.

[Table 4](#) shows equivalent figures for the United States, for trips by workers and nonworkers up to the age of 64 years. Personal vehicle (car, van, or similar) is the most popular mode for both groups, with more nondisabled people traveling by car, and disabled

Table 1 Types of limitation on mobility

People with disabilities	Permanent mobility limitations
People with age-related mobility limitations	
Elderly people	Temporary mobility limitations
Young children	
People with temporary mobility limitations	

Source: Based on Figure 1 in *Berliner Verkehrsbetriebe* (2003).

Table 2 Total number of trips per person per year by people aged 16+ years in England in 2018

		Number of trips
People with disabilities	Vision	666
	Hearing	769
	Mobility	687
	Dexterity	703
	Learning	703
	Memory	678
	Mental health	733
	Stamina/breathing/fatigue	787
	Social/behavioral	594
	Speech	660
	Other	930
All with a disability		853
All without a disability		1,055
All		1,012

Source: *Department for Transport* (2019).

Table 3 Trips per person per year by each mode in England by people aged 16+ years in 2018

	Disabled		Not disabled	
	Number	%	Number	%
Car or van driver	368	43.1	525	49.8
Car or van passenger	157	18.4	133	12.6
Motorcycle	2	0.2	2	0.2
Other private transport	9	1.1	4	0.4
Bus	53	6.2	49	4.6
Surface Rail	11	1.3	30	2.8
London Underground	3	0.4	16	1.5
Taxi / minicab	15	1.8	10	0.9
Other public transport	2	0.2	4	0.4
Walk	224	26.3	260	24.6
Bicycle	10	1.2	20	1.9
All modes	853	100.0	1,055	100.0

Source: *Department for Transport* (2019).

people more likely to be car passengers and less likely to be drivers than nondisabled people. Disabled people use public transport (local transit) more than nondisabled people. Walking is more popular for disabled people than those without a disability.

A similar picture emerges in the United States for older people, as **Table 5** shows. Car is still the dominant mode with more trips as a car driver than as a car passenger, for disabled people, despite their age and disabilities. Disabled people use local transit and paratransit more than nondisabled people. However, older disabled people walk less than nondisabled people.

Spatial factors such as density and distance from the city center can affect modal choice. **Table 6** shows the percentage of trips by each mode by disabled and nondisabled people in New York in three different areas: the New York Metropolitan Transportation Council (NYMTC) area, which is an area with high densities and high level of provision of public transport, the rest of New York state, which is less dense, but is denser, on average than the rest of the United States, which is the third area shown. The general picture is similar across the three areas, but with higher car use outside the NYMTC area. In all cases, including the NYMTC area, more trips by disabled people in privately owned vehicles are as drivers rather than passengers. One difference between the three areas is that disabled people walk less than other people in the rest of New York State area.

Table 4 Modal share by worker and disability status (age 18-64 years) in the United States in 2017

	Workers		Non-workers	
	Disabled	Not disabled	Disabled	Not disabled
Personal vehicle (driver)	54.5	73.6	42.6	58.3
Personal vehicle (passenger)	23.5	11.5	31.0	21.2
Local transit	4.3	2.7	5.9	3.3
Paratransit	1.2	0.0	1.6	0.1
Walk	13.0	9.2	14.6	14.4
Other modes	3.5	3.0	4.3	2.7
All modes	100.0	100.0	100.0	100.0

Source: Brumbaugh (2018).

Table 5 Modal share by disability status for people aged 65 years and older in the United States in 2017

	Disabled	Not disabled
Personal vehicle (driver)	47.8	69.3
Personal vehicle (passenger)	36.2	16.8
Local transit	2.8	1.8
Paratransit	0.9	0.1
Walk	9.2	10.7
Other modes	3.1	1.9
All modes	100.0	100.0

Source: Brumbaugh (2018).

Table 6 Modal share in New York in 2009

	NYMTC		Rest of NYS		Rest of US	
	Disabled	Not Disabled	Disabled	Not disabled	Disabled	Not disabled
POV-Driver	24.3	40.6	52.9	72.1	50.2	71.3
POV-Passenger	19.2	10.8	28.6	15.0	32.0	15.9
Taxi	4.9	1.5	1.0	0.2	0.6	0.1
Public Transit	17.0	16.1	2.8	0.9	2.7	1.4
Walk	31.1	29.0	8.2	9.4	10.6	9.1
Other	3.3	2.0	5.6	2.3	4.0	1.9
No Response	0.1	0.1	0.9	0.1	0.1	0.2
All modes	100.0	100.0	100.0	100.0	100.0	100.0

NYMTC, New York Metropolitan Transportation Council area; NYS, New York State; POV, Privately owned vehicle.

Source: Hwang et al., (2016).

Turning to older people, **Table 7** shows the modal shares for people aged 50-59 years, 60-69 years, and aged 70 + years. The number of trips made declines with increasing age, as does the volume of car driving. The number of trips as a car passenger increases after the age of 60 years and stays fairly constant after that in terms of the number of trips but increases in terms of modal share. Even over the age of 70 years, more than twice as many car trips are as a driver than as a passenger. The decline in car use is partly compensated for by an increase in bus usage with age, but rail use declines. This may be because of the changes in the nature of trips by older people and the reduction in the distance traveled with increasing age (Mackett, 2017). Walking increases slightly after the age of 60 but declines later in terms of trips, but increases in terms of modal share. The low levels of cycling decline with age.

Table 8 shows equivalent figures for the United States. The shift from being a car driver to being a car passenger with increasing age can be seen, but even over the age of 75 years, car driving is still the dominant mode. The other modes show little change over time because of the dominance of car use.

Barriers to Transport Modes Used by Disabled and Elderly People

Evidence discussed above showed that disabled and older people travel less than other people and use different modes of travel. Nondisabled people may also not use some modes as much as they would like for reasons such as cost or lack of availability. As **Table 9**

Table 7 Trips per person per year by each mode in England by people aged 50 years and older in 2018

	<i>Ages 50-59</i>		<i>Ages 60-69</i>		<i>Ages 70+</i>	
	<i>Number</i>	<i>%</i>	<i>Number</i>	<i>%</i>	<i>Number</i>	<i>%</i>
Car or van driver	593	56.3	541	51.4	369	43.3
Car or van passenger	127	12.0	163	15.5	165	19.3
Motorcycle	4	0.4	3	0.3	1	0.1
Other private transport	3	0.3	6	0.6	6	0.7
Bus	28	2.7	49	4.7	70	8.2
Surface Rail	22	2.1	13	1.2	6	0.7
London Underground	7	0.7	4	0.4	2	0.2
Taxi / minicab	10	0.9	8	0.8	11	1.3
Other public transport	3	0.3	2	0.2	2	0.2
Walk	239	22.7	249	23.7	212	24.9
Bicycle	18	1.7	14	1.3	8	0.9
All modes	1,054	100.0	1,052	100.0	853	100.0

Source: *Department for Transport (2019)*.

Table 8 Percentage of seniors travel by mode in the United States in 2017

	<i>Ages 65-74</i>	<i>Ages 75+</i>
Driver	68	62
Passenger	18	25
Transit	1	1
Walk	10	10
Bike	1	0
All other modes	2	2

Source: *Federal Highway Administration (2019)*.

Table 9 Percentage of adults aged 16+ using modes of travel less than they would like in GB in 2009/11

<i>Mode of transport</i>	<i>Disabled</i>	<i>Not disabled</i>
Motor vehicle	25	12
Local buses	18	12
Long distance buses	15	9
Taxis	10	5
Local trains	17	10
Long distance trains	18	11
Metro	11	7

Note: Motor vehicle includes car, van, motorcycle, or moped.

Source: *Office for Disability Issues (2011b)*.

shows, this is the case. For all modes, more disabled people use the modes less than they would like than non-disabled people. The motor vehicle is the mode that the greatest number of disabled people cannot use. It is also the mode for which the difference between disabled and nondisabled people is greatest at 13%. Local buses and long-distance trains are the two modes after motor vehicles that most disabled people use less than they would like. Taxis are the mode that the fewest disabled people are unable to use.

Table 10 shows the barriers to use of motor vehicles for disabled and nondisabled people. More disabled people than nondisabled people face at least one barrier to using a motor vehicle. For both groups, cost is the top factor, deterring about half of each group. The next two barriers for disabled people are having a health condition, illness, or impairment and having a disability. Some long-term health conditions affect the use of a motor vehicle because it prevents them from obtaining a driving license. The only other issue that affects disabled people more than nondisabled is parking problems, which may reflect the difficulties that wheelchair users can have finding a suitable parking space.

Table 11 shows the barriers to using buses and taxis. For each of the modes shown, more disabled people than nondisabled people face barriers to using the mode. Many people cited having a health condition, illness, or impairment or a disability as a barrier, more so for buses than for taxis. This may be because taxis carry passengers from door-to-door, avoiding the need to walk along the street, which may be a barrier to some people. This is confirmed by the relatively large numbers of people who cited getting

Table 10 Barriers to using a motor vehicle by disabled and nondisabled adults aged 16+ in GB in 2009/11

Barrier	Disabled	Not disabled
At least one barrier to using a motor vehicle	27	14
Cost	51	49
A health condition, illness or impairment	30	3
A disability	18	1
Parking problems	14	12
Vehicle not available when needed	9	13
Too busy/not enough time	9	17
Difficulty getting in or out of the vehicle	8	-
Caring responsibilities	6	5

Note: Motor vehicle includes car, van, motorcycle or moped.

Note: Factors that affect more than 5% of disabled people are shown.

Source: *Office for Disability Issues (2011b)*.

Table 11 Barriers to using buses and taxis by disabled and nondisabled adults aged 16+ in GB in 2009/11

Barrier	Local bus		Long distance bus		Taxis	
	Disabled	Not disabled	Disabled	Not disabled	Disabled	Not disabled
At least one barrier to using the mode	34	21	38	23	24	14
A health condition, illness or impairment	31	2	32	4	13	-
A disability	23	1	20	1	9	-
Transport unavailable	22	37	8	12	3	4
Cost	21	28	31	35	79	90
Difficulty getting in or out of the transport	18	3	11	2	4	1
Difficulty getting to stop or station	17	8	11	8	1	-
Difficulty getting from stop or station to destination	16	7	10	7	1	-
Anxiety/lack of confidence	12	2	11	2	5	1
Delay and disruption to service	11	16	7	12	-	-
Overcrowding	9	10	10	11	-	1
Lack of help or assistance	8	2	5	1	2	-
Too busy/not enough time	7	14	7	13	1	1
Attitudes of passengers	6	7	4	5	-	-
Lack of information	6	9	4	5	1	-
Fear of crime	6	5	5	4	1	2
Lack of space	5	5	8	9	1	-
Other reasons	15	25	19	34	3	4

Note: Factors that affect more than 5% of disabled people for any mode are shown.

Source: *Office for Disability Issues (2011b)*.

to or from the stop or station as a barrier for the bus compared to the numbers mentioning this for taxi. Similar numbers of disabled people mentioned the difficulty of getting in or out of the transport as a barrier to use of the mode. The only other barrier which was cited by many more disabled people than nondisabled people was anxiety/lack of confidence, implying that there are psychological barriers to using these modes. There is evidence of these for people with mental health conditions in [Mackett \(2019\)](#). A factor for many people was cost, especially for taxis. This was a barrier for more nondisabled people than disabled, either because some disabled people are provided with cheap or free travel for these modes or because other factors were such a deterrent to using the modes that they did not consider the cost of using them.

Moving to rail, [Table 12](#) shows the barriers to traveling by rail for disabled people. The key barriers mentioned as well as having a disability, are accessing the train both getting to and from the station and getting on and off the train, and anxiety/lack of confidence.

The possible causes of the anxieties traveling on public transport are illustrated by the figures in [Table 13](#), which shows the percentage of people living in London identifying barriers to greater use of public transport. Barriers cited by more disabled people than non-disabled people were fear of crime both on the bus or train or traveling to it, fear of terrorist attacks, the risk of accidents and graffiti. The much higher figures for the fear of crime in [Table 13](#) compared to those in [Tables 11 and 12](#) probably reflects the fact that [Table 13](#) is for London where street crime is much higher than in other parts of the country.

Table 12 Barriers to using trains by disabled and non-disabled adults aged 16+ in GB in 2009/11

<i>Barrier</i>	<i>Local trains</i>		<i>Long distance trains</i>		<i>Metro</i>	
	<i>Disabled</i>	<i>Not disabled</i>	<i>Disabled</i>	<i>Not disabled</i>	<i>Disabled</i>	<i>Not disabled</i>
At least one barrier to using local trains	31	16	32	18	36	23
Cost	31	37	48	65	8	9
A health condition, illness or impairment	27	2	27	2	14	1
Transport unavailable	19	34	12	9	64	76
A disability	18	1	18	1	9	-
Difficulty getting to stop or station	15	11	11	7	4	2
Difficulty getting from stop or station to destination	12	8	10	6	4	1
Difficulty getting in or out of the transport	10	2	9	2	5	1
Anxiety/lack of confidence	10	2	10	2	8	3
Overcrowding	7	6	8	8	6	7
Other reasons	7	12	8	12	3	3

Note: Factors that affect more than 5% of disabled people for any mode are shown.

Source: *Office for Disability Issues (2011b)*.

Table 13 Percentage of people living in London identifying barriers to using public transport more often

	<i>Disabled</i>	<i>Not disabled</i>
Overcrowded services	52	60
Cost of tickets	26	47
Unreliable services	34	45
Slow journey times	27	45
Concern about antisocial behavior	54	38
Fear of crime getting to the bus/train	40	28
Fear of crime on the bus/train	39	27
Fear about knife crime	40	26
Dirty environment on the bus/train	26	26
Dirty environment getting to the bus/train	15	19
Fear of terrorist attacks	18	12
Lack of info on how to use public transport	11	11
Risk of accidents	17	8
Graffiti	16	8
Don't understand how to buy bus tickets	6	6
Accessibility issues/no lift	5	1
None of these	14	13

Note: Responses are shown if they exceed 4% for disabled people.

Source: *Transport for London (2012)*.

Interventions to Increase Modal Usage by Older and Disabled People

It has been shown that older and disabled people travel less by most modes than other people. Car is the mode that they use most, but less than other people. The development of fully autonomous vehicles which do not require the user to perform any control functions should enable older and disabled people to travel by car as much as they wish, providing that they can afford to do so. For more information about autonomous vehicles see Articles 10093 and 10400.

One area that prevents disabled and older people from using some public transport modes is difficulty accessing the vehicles. This requires better design of both the vehicles and the street environment to facilitate access. The design process should involve disabled and older people. Another barrier to use of some modes was anxiety/lack of confidence. One way to address this is to provide more staff who have been trained to offer appropriate assistance. Clear information about wayfinding, purchasing tickets, and obtaining help are all useful in providing reassurance.

Conclusions

This article has shown that people with long-term mobility limitations travel less than others and they use a different mix of modes. In Britain and the United States, most of their trips, like the rest of the population, is by car, but they are more likely to be car passengers than nondisabled people. Disabled people tend to use buses and taxis more but are less likely to travel by train than

nondisabled people. The reasons for the lower levels of use of some modes by disabled people than the rest of the population include difficulties accessing buses and trains and anxiety and lack of confidence about using them. For car driving, a major factor may be that some disabled and older people are not able to hold a driving license.

Biography

Roger Mackett is Emeritus Professor of Transport Studies at University College London. He has researched into various aspects of transport policy including the barriers to access for older and disabled people, the use of cars for short trips and the effects of car use on children's lives. He is a member of the Disabled Persons' Transport Advisory Committee (DPTAC) and the Standing Committee on Accessible Transportation and Mobility of the US Transportation Research Board (TRB).

References

- Berliner Verkehrsbetriebe, 2003. The accessibility of urban transport to people with reduced mobility. Available from <http://www.eukn.eu/fileadmin/Lib/files/EUKN/2010/accessible-transport.pdf>.
- Brumbaugh, S., 2018. Travel Patterns of American Adults with Disabilities. Available from <https://www.bts.gov/topics/passenger-travel/travel-patterns-american-adults-disabilities>.
- Department for Transport, 2019. National Travel Survey. Available from <https://www.gov.uk/government/collections/national-travel-survey-statistics>.
- Federal Highway Administration, 2019. Travel Trends for Teens and Seniors, 2017 National Household Travel Survey. Available from https://nhts.ornl.gov/assets/FHWA_NHTS_Report_3C_Final_021119.pdf.
- Federal Highway Administration, 2020. National Household Travel Survey, US Department of Transportation. Available from <https://nhts.ornl.gov/>.
- Hwang, H.L., Reuscher, T., Wilson, D., 2016. Characteristics and Travel Patterns of New York Residents: Subpopulations of Persons with a Disability in 2009. Oak Ridge National Laboratory, ORNL/TM-2016/305. Available from <https://info.ornl.gov/sites/publications/files/Pub68412.pdf>.
- Mackett, R.L., 2017. Older people's travel and its relationship to their health and wellbeing, in: Musselwhite C B A (Ed.), Transport, Travel, and Later Life, Emerald, Bingley, Yorkshire, Great Britain, pp. 15-36.
- Mackett, R.L., 2019. Mental health and travel: Survey report, Department of Civil, Environmental and Geomatic Engineering, University College London. Available from <https://www.ucl.ac.uk/civil-environmental-geomatic-engineering/mental-health-and-travel-report>.
- Office for Disability Issues, 2011a. Equality Act 2010: Guidance. Available from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/85010/disability-definition.pdf.
- Office for Disability Issues, 2011b. Life Opportunities Survey: wave 1 results. Available from <https://www.gov.uk/government/statistics/life-opportunities-survey-wave-one-results-2009-to-2011>.
- Office for National Statistics, 2017. Long-lasting Health Conditions and Illnesses; Impairments and Disability, Harmonised Concepts and Questions for Social Data Sources: GSS Harmonised Principle. Available from <https://gss.civilservice.gov.uk/wp-content/uploads/2019/04/LLI-and-Disability-June-17-2.pdf>.
- Transport for London, 2012. Understanding the travel needs of London's diverse communities: Disabled People. Available from <http://content.tfl.gov.uk/disabled-people.pdf>.

Further Reading

- Amadeus IT Group, 2017, Undated. Working towards inclusive and accessible travel for all. Available from <https://amadeus.com/en/insights/research-report/voyage-of-discovery-working-towards-inclusive-and-accessible-travel-for-all>.
- Bekiaris E, Loukea M, Spanidis P, Ewing S, Denninghaus M, Ambrose I, Papamichail K, Castiglioni R, Veitch C, 2018, Research for TRAN Committee: Transport and tourism for persons with disabilities and persons with reduced mobility. Available from [http://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU\(2018\)617465](http://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2018)617465).
- Darcy, S., Burke, P.F., 2018. On the road again: The barriers and benefits of automobility for people with disability. Transp. Res. Part A 107, 229–245.
- Field, M.J., Jette, A.M. (Eds.), 2007. The Future of Disability in America, The National Academies Press, Washington DC. Available from <https://www.nap.edu/download/11898>.

Transport Modes and Commuters

Colin G. Pooley, Lancaster Environment Centre, Lancaster University, Lancaster, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	92
What is Commuting?	92
What is Transport?	92
Unraveling Complexity in Commuting and Transport	93
Ways of Working and Commuting	93
Home, Field, Farm, and Workshop	94
The Concentration of Employment in Factories and Offices	94
The Role of Mass Transit	94
The Desire for Independence and Flexibility	95
Conclusions: The Future of Transport and Commuting	96
Biography	96
See Also	96
References	96
Further Reading	96

Introduction

Some form of paid employment is a necessary part of most people's lives, and travel to and from work can form a significant component of the day. On average, commuters in many parts of the world travel for about 20–30 min each way (irrespective of the travel mode used), but commuting times can be much greater than this (an hour or more) in large metropolitan areas. Perhaps surprisingly, the time spent traveling to and from work has changed relatively little over time, but the distance traveled has increased substantially as commuters have gained access to faster forms of transport. For instance, it has been estimated that in Britain in the period 1920–9 the average commuting distance was 6.8 km with a travel time of 30.3 min (Pooley and Turnbull, 1999), whereas in England in 2017 commuters still traveled on average for 31 min but covered a distance of 14.6 km (DfT, 2018). This article examines the ways in which the journey to work has changed over time, the ways in which it varies in relation to factors such as locality, gender, age, and occupation, and its interaction with deeper structures of cultural, social, economic, technological, and environmental change within societies. First, however, it is necessary to define exactly what is meant by both transport and commuting.

What is Commuting?

The term commuter was first used in the United States in the mid-19th century to define someone who traveled from the expanding suburbs of American cities to work in the city center, but its origins as a word that denotes the exchange of one thing for another can be traced to at least the 17th century. In its most commonly used modern form it simply means exchanging one location for another on a regular basis. However, on closer examination the concept of travel to and from work is rather more complex than this definition suggests. While the term commuter is most commonly associated with travel for paid employment, many people will travel for unpaid work of many kinds. This encompasses much activity that has traditionally been viewed as female “work” such as escorting children or essential travel related to household duties, and it may also include travel for study at school or college. In many parts of the world much labor is undertaken in fields close to the home producing food both for consumption and sale, with men and women traveling daily to tend crops, while others travel for their work or work while traveling. The rhythm of commuting can also be highly varied. Although most people travel daily, others may have a weekly commute between home and workplace while some may travel on a regular basis every month or less often. In some employment such patterns are far from regular and vary according to the availability and location of work. Finally, some work at home most or all of the time and their commute may be from one room to another undertaking both paid and unpaid activities. While the most commonly understood meaning of commuting as regular travel to and from paid employment dominates the literature, it is important to remember the multiplicity of experiences and activities that constitute travel for work in all its possible forms.

What is Transport?

To transport something simply means to move it from one place to another by any means, but in the context of commuting transport is usually viewed as the machine (most commonly a train, bus, car, or bicycle) by which people travel from their home to their

workplace. Thus, the literature is dominated by research on changing transport technologies and their impacts upon individuals and societies. However, both in the past and the present many people travel to and from their places of employment on foot, but walking is rarely viewed as a form of transport and is thus neglected in much transport planning. In some parts of the world walking is the dominant mode of transport for daily activities, and this is especially the case for women and children (Porter, 2002). I argue that walking should be given at least equal status with other forms of transport in studies of commuting and in planning transport networks. The forms of transport used for travel to work have not only varied globally over time and between countries, but also within countries by scale and density. For instance, in large rich countries such as the United States some people may commute long distances by plane on a regular basis, but this is much less common in small or poorer nations. There are also significant differences between urban and rural areas. In high-density urban environments commuting distances for some may be quite short, and are more likely to be by public transport, by bike, or on foot, whereas in rural areas and the low-density suburbs of cities commuting is most often over a longer distance and is mainly undertaken by car or public transport. Finally, with the development of new communications technologies, it is important to consider the role of virtual communications in affecting everyday travel. Many office workers are increasingly able to exchange a daily commute for work at home on some days of the week, with internet-based communication replacing daily travel. As such, new communication technologies must be viewed as another form of transport related to commuting.

Unraveling Complexity in Commuting and Transport

The analysis of transport modes and commuting is intertwined with many other themes, and is embedded within a wide array of societal structures, forming part of a complex system of networks and flows that create distinctive opportunities and resistances. In this section, I outline some of the main factors that influence the transport modes used by commuters and the fields of study relevant to their analysis. Most obviously, the transport modes used by commuters have been structured by the history of transport and communications technologies and their impacts in different places and on different segments of society. There is a tendency to assume that as new technological innovations developed they replaced the old forms of transport, but in many places and especially for more marginalized sections of society, older and traditional transport modes continued and flourished alongside the new. There is also need to consider the evolution of work practices and the ways in which transitions from, for instance, family and home-based work to factory or office-based work totally changed the structure of people's lives and travel behaviors. Working practices, the composition of the work force, and the nature of work itself all impact upon travel modes and commuting practices. Changes in urban structure and the location of both homes and workplaces have also had a major impact on commuting flows and modal choices. For instance, when most daily movement was from suburb to city center the use of radial networks of public transport was easy and normal. This was the case in US cities in the late-19th century when the term commuter was first used in its modern sense. However, as urban structures became more complex and work locations dispersed, more people were faced with complex cross-city journeys that were hard to accomplish by public transport but much easier by car. Where they could, commuters changed their modal choice. An important, but sometimes neglected, factor in analyses of transport modes and commuting is the structure of gender roles within society. In the past many societies (though not all) were strongly male dominated with commuting for work undertaken mainly by males. Women's unpaid work in childcare and house-care was not considered productive and their travel in association with these activities rarely explicitly considered. Increasingly, in at least some societies, women have entered the workforce, but it remains the case that many women continue to carry the main responsibility for family and household duties. They are thus more likely to work part time or flexibly, and to work closer to home so that their complex mix of responsibilities can be managed. Thus, the decisions that women are required to make by the societal and familial structures that exist in many societies strongly influence their commuting patterns and travel modes. Finally, none of these factors can be considered without taking into account global histories of economic development. Persistent inequalities in wealth and associated power within the global economic and political system have meant that the opportunities to adopt the most modern and efficient working practices and transport technologies have been limited. Historical and persisting global inequalities both structure and are reflected in present-day transport and commuting patterns and experiences around the globe.

Ways of Working and Commuting

There are many ways in which the topic of transport and commuting could be tackled, with probably the most conventional being to focus on each of the main transport modes. However, rather than doing this explicitly, I identify distinctly different ways of working with their associated patterns of commuting and transport modes, and consider the individual and societal impacts of each. Although the scenarios I present can be viewed as representing the stages through which many societies have changed, they are explicitly not presented as models of "modernization" or "development." Rather, it is argued that it is possible for different ways of working and traveling to sit perfectly comfortably alongside each other, and with the potential for each to wax or wane over time in a cyclical structure. Each scenario is explained through reference to examples drawn from different parts of the world at various time periods, but these are by no means the only locations or times to which they could apply.

Home, Field, Farm, and Workshop

For much of history, and in many places today, people work and live in very close proximity. For some this is a choice but for others a necessity. Women and men who farm almost anywhere in the world and at any time in the past or present live with their work. For instance on a dairy farm in northwest England accounts and other paper work are typically done in the home, the milking parlor is in the yard together, possibly, with pens for hens or geese, and travel to and from the fields is undertaken mainly on a quad bike or tractor. In poorer countries of the world much of the same pattern occurs, varying only with the agricultural economy and transport available. Much travel to tend field crops is likely to be on foot, as it would have been in the past in most locations. Small-scale industrial activity can also be carried out close to the home. The development of locally based cooperatives both in agriculture and craft industry has enhanced the financial independence of many women in, for instance, sub-Saharan Africa, with productive activities carried out in homes or nearby workshops. This is not dissimilar to the situation in much of preindustrial Europe where work on the land might be combined with small-scale manufacturing in a nearby workshop, possibly as part of a family-based dual economy where all family members contributed to the household income. In all these cases commuting is over very short distances and mostly on foot. The development of new communications technologies in the 21st century have enabled some to exchange a long commute to an office for work at home for at least part of the week, thus in some ways returning to a preindustrial system of employment and commuting. Furthermore, for women or men engaged mainly in domestic duties, whether paid as a housekeeper or *au pair* or unpaid as a housewife/husband, travel to and from work is within the home or immediate locality while carrying out everyday domestic tasks. Both in the past and the present it is important to remember that not all work requires long commutes or complex transport systems, but that much productive activity can be accomplished in and close to the home with most travel on foot. Arguably, if we wish to move toward more sustainable and energy-efficient lifestyles then the growth of employment and commuting practices that minimize transport needs is one possible solution (Banister, 2008).

The Concentration of Employment in Factories and Offices

In 19th-century Europe industrial change led to the rapid growth of larger units of production and associated commercial and office activities. This occurred first in Britain but rapidly spread through most of the richer nations of the world. The development of large-scale factory employment reconstructed the time-space prism through which many workers viewed the world. Rather than working in or close to home, at a pace and time that suited the family, factory or office employment required the presence in a specific place at particular times, with harsh penalties for not conforming to the established work routine. Working hours were usually long and it was thus essential that workers could travel easily and quickly from home to workplace. In an age before mass public transit such as in 19th-century Europe, or in poorer nations today where most cannot afford transport fares, it was and still is essential to live as close as possible to a workplace. A short journey to work is especially important when employment is casual and/or uncertain: failure to be at a workplace on time could lead to a day without income. Thus, in Victorian Britain many families continued to live in poor quality inner-city housing close to new factory employment because of the need to be able to travel easily on foot, even though when in work they may have been able to afford better housing elsewhere in a city. For much of the 19th century only those with higher incomes and more secure employment could countenance a longer commute either on foot or by horse-drawn omnibus (Dennis, 1986). In the past and the present when incomes are limited and working hours are long there is a premium on a short journey to work by the cheapest means possible.

One other consequence of the concentration of employment in larger units is increased differentiation of employment and commuting by gender, age, and skill. In the 19th century office employment was almost entirely male and most factories were dominated by a male workforce, the main exception being in textile districts where female factory labor was important. Thus, while men commuted, albeit a short distance, to work regular hours in a central location most women remained much closer to the home often engaged in work such as sewing or laundry that could be done in the home and combined with childcare and other household tasks. Factories and offices also became more hierarchical, with differentiation of skills and associated income. In general, clerks and other office workers had more regular employment than most factory workers and could thus commit a larger portion of their income to rent, allowing a longer commute to work. However, within the office, status and reward were also increasingly differentiated by age and experience as part of the gradual “professionalization” of employment. Within factories too there was differentiation between those with skills who had served a lengthy apprenticeship and those who relied on manual labor: while the former could afford to travel from the inner suburbs, the latter often were forced to live very close to their workplace. However, only the very rich could afford to live more than a 30-min walk or short bus ride from their workplace. For those on low incomes and with many competing calls on their time a short commute by whatever means is available and can be afforded remains important for many workers, especially women who may work part time and combine paid employment with childcare and household responsibilities.

The Role of Mass Transit

In the 21st century, the large-scale movement of daily commuters into major urban areas from outer suburbs and surrounding communities is a normal phenomenon. For instance, in Europe cities such as Paris, Milan, and London have daily inflows of more than one million people, with megacities such as Beijing and Shanghai in China experiencing much larger inflows. This can lead to very long daily commuting trips: commuters to Beijing and Shanghai have some of the longest commutes with an average one-way journey time of over 50 min, longer than in either New York or London. While city size and high city-center property prices are the

main determinants of long-distance commuting today, this was made possible in the 19th century by technological change. The development of mass transit systems in US cities such as Boston from the 1860s led to the growth of the so-called “street car suburbs,” from which much of the male population commuted to work in the city center on the newly developed network of trams and trolley buses. For the first time commuting by public transit became the norm for a majority of the male population of working age (Warner, 1978). Similar technologies spread rapidly to cities in Europe and elsewhere. For instance in Amsterdam a network of horse-drawn trams developed from the mid-1870s, and after 1900 when the municipality took over operation of the tram network the lines were gradually electrified and extended into growing suburbs. Upgrading of old routes and the development of new lines has continued so that trams continue to be a key element of Amsterdam’s public transit infrastructure. Not all cities that gained a tram network in the 1870s managed to retain it during the 20th century. In Manchester (United Kingdom) horse-drawn trams developed from 1877 with the first electric tram in 1901. However, by 1928 the tram network had reached its peak and the network was gradually replaced by motor buses that were seen to offer a more flexible and modern form of public transport. In 1949 all Manchester trams ceased running and it was to be 43 years before trams reappeared with the opening in 1992 of the first Manchester Metrolink line, a combination of light rail and street tram, partially utilizing existing rail tracks (Pooley and Turnbull, 2000). Today, Manchester has a reasonably dense network of trams that are seen as a sustainable and efficient way of moving large numbers of commuters around the city. A similar history has been repeated in many British cities. Urban metro lines and commuter rail networks also developed at the same time as urban bus and trams routes with most major cities gaining a dense network during the 20th century. Shanghai, Beijing, Guangzhou (all in China), and London all have metro networks of over 400 km in length, with cities such as Moscow, New York, Delhi, and Tokyo not far behind. The mass transit systems that started in the 1860s, and have evolved and become more technologically sophisticated over time, continue to carry the majority of commuters in larger towns and cities. For instance in 2011 some 69.4% of commuters in Paris traveled by public transport compared to a French national average of just 17.8% of commuter trips undertaken on public transit systems.

The Desire for Independence and Flexibility

While mass transit systems can convey large numbers of commuters quickly and conveniently from suburb to city center, they also have a number of disadvantages from the perspective of many individual travelers. They run on fixed routes and at predetermined times that may be inconvenient if they do not coincide with individual time–space schedules. They also require commuters to share space with other travelers, often in crowded compartments that lack privacy and comfort. For many commuters daily travel to work on public transit systems can also be prohibitively expensive. Travel on foot is undoubtedly the cheapest, most flexible, and independent form of transport. You can travel anywhere, at any time for almost no cost and with whoever you choose as a companion. However, expanding residential suburbs and changes in city structure meant that walking became impractical for many commuting journeys. The bicycle offers one alternative that enables the traveler to cover greater distances while retaining independence, flexibility, and relatively low cost. In Europe from the 1920s cycling was increasingly adopted as the commuting mode of choice by many men (Horton et al., 2007). This enabled commuters to travel where and when they wished, especially important for those who worked at times or in locations not well served by public transport, and cycling avoided the need to mix with other travelers. At its peak in the 1930s some 60% of all trips in the city of Berlin were by bike. However, many of the advantages that commuters sought from the bicycle could be fulfilled more easily and comfortably by the use of private motor vehicles, and as cars became more affordable bicycle use in European cities declined to be replaced by driving to and from work. Cars allowed longer journeys to be completed, enabled the easy carrying of passengers and luggage, and protected users from the elements. Their advantages were seductive as most rich countries entered the age of automobility. In the early years of motoring cars were used mainly for leisure and pleasure and not for commuting, but changes in urban structure hastened the rise of the motor car and the relative demise of the bicycle and public transit (Mom, 2014). When workplaces were mostly in the center of cities trams and buses worked well, but as employment nodes became distributed around the urban periphery workers had to make complex cross-town journeys that were much more easily accomplished by car. The dominance of motor travel occurred first in the United States but spread rapidly to Europe and elsewhere by mid-century. This trend has been seriously challenged in only a few cities such as Copenhagen and Amsterdam where today cycling forms a large component of daily commuting.

In poorer counties of the world the timing of change has differed but the trends have been much the same. For instance in much of Africa and Asia, while many continue to walk or cycle to work, or use public transport systems, motor transport has become increasingly important creating high levels of congestion and pollution. For instance, cycling accounted for around 40% of all trips in many Chinese and Indian cities toward the end of the 20th century, but rates were highest in medium-sized towns and cycling has declined markedly in the 21st century with a consequent increase in the use of motor vehicles in all locations, though the use of e-bikes has partially limited the decline in cycling. As in the past, commuting continues to be structured by gender and income. Women continue to be more likely to work closer to home and to travel on foot or by public transport, while those on low incomes seek to minimize travel costs and are often forced to endure long and uncomfortable commuting journeys at unsocial hours. In many cities the characteristics of those who walk, cycle, use public transit, or travel by car are very different. The desire for flexibility and more personalized transport can also be partially met by a variety of *para*-transit systems that exist in various guises in most urban areas. Almost all cities have taxis and the development of digital technologies has revolutionized their use in many large urban areas, but older technologies also persist with, for instance, the cycle rickshaw continuing to provide personalized transport in many Indian cities. Other semiflexible systems include dial-a-bus schemes that can be more convenient for older and less mobile travelers in rural areas and community-based car-sharing clubs. In most cities in many parts of the world there now exist a multitude of travel

options for the commuter, but in rural areas commuting choices are more limited. Most small settlements have poor or nonexistent public transport systems and thus commuters are forced to travel by car if they need to cover a significant distance.

Conclusions: The Future of Transport and Commuting

It has been demonstrated already that older technologies and transport modes can coexist alongside new technologies, but 21st-century awareness of and concern about global climate change and the quality of the urban environment may lead to the further revival of some older forms of transport and commuting behaviors. It is undeniable that the petroleum-powered automobile is a major source of carbon and produces tailpipe emissions harmful to human health. Many countries are now seeking ways to reduce these impacts, especially in urban areas. Travel by air or high-speed train also harms the environment, and although these modes are rarely used for a daily commute some people who travel weekly or monthly to work between one location and another may fly or travel by high-speed train. Commuting, often over increasing distances and by more polluting modes, has become firmly embedded in most societies and will be hard to shift, but changing societal attitudes toward the environment, especially from the younger generations, may mean that change is possible. New automobile technologies such as the use of electric vehicles and autonomous transport systems may reduce some aspects of urban pollution and carbon release, and China has been at the forefront of investment in electric vehicle technology. But these changes also bring other risks. First, they are still cars that can produce associated congestion and risks to pedestrians and cyclists. Second, unless there is a global shift to renewable electricity production the impact on carbon reduction will be limited. Third, the components of batteries, most notably lithium, cobalt, and nickel, have significant environmental costs in terms of their extraction, and disposal. There is a risk that in solving one problem we create another. One solution is simply to travel less: to restructure cities and working practices so that most people can work from home or close to home using new digital technologies, with most travel on foot or by bicycle and with longer trips by energy-efficient and carbon-neutral public transport. In effect, this would be a return to commuting patterns of two centuries ago. Some people do travel in these ways, either by design or necessity, and just as urban structures and working practices have been changed in the past so too there is no reason why creative use of new transport technologies should not enable this to occur again.

Biography

Colin Pooley is emeritus professor of Social and Historical Geography in the Lancaster Environment Centre, Lancaster University, United Kingdom. His research focuses on the social geography of Britain and continental Europe since c. 1800, with recent projects focused on residential migration, travel to work, and other aspects of everyday mobility.

See Also

Sustainable Mobility Paths; Transport Modes and Accessibility; Transport Modes and Cities; Travel Mode Choice as Reasoned Action; Transport Modes and People With Limited Mobility; Active Transport: Heterogeneous Street Users Serving Movement and Place Functions; Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service; Mode Choice and Life Events; Next Generation Travel: Young Adults' Travel Patterns; A Geography of Road Transport in Cities; The Future of Road Transport; Road Modes: Walking; Bus Public Transport Planning and Operations; Street Design for Active Travel; Transit Planning and Management; Car Ownership and Car Use: A Psychological Perspective; Road Transport: E-Scooters; Introduction to Rail Transport; Subway Systems; Women and Transport Modes

References

- Banister, D., 2008. The sustainable mobility paradigm. *Transp. policy* 15 (2), 73–80.
- Dennis, R., 1986. *English Industrial Cities of the Nineteenth Century: A Social Geography*. Cambridge University Press, Cambridge, UK.
- DfT, 2018. *National Travel Survey: England 2017*. DfT, London. Available from: <https://www.gov.uk/government/statistics/national-travel-survey-2017>.
- Horton, D., Rosen, P., Cox, P., 2007. *Cycling and Society*. Routledge, London.
- Mom, G., 2014. *Atlantic Automobility: Emergence and Persistence of the Car, 1895-1940*. Berghahn, New York.
- Pooley, C., Turnbull, J., 1999. The journey to work: a century of change. *Area* 31 (3), 281–292.
- Pooley, C., Turnbull, J., 2000. Commuting, transport and urban form: Manchester and Glasgow in the mid-twentieth century. *Urban Hist.* 27 (3), 360–383.
- Porter, G., 2002. Living in a walking world: rural mobility and social equity issues in sub-Saharan Africa. *World Dev.* 30 (2), 285–300.
- Warner, S.B., 1978. *Streetcar Suburbs*. Harvard University Press, Cambridge, MA.

Further Reading

- Blumen, O., 1994. Gender differences in the journey to work. *Urban Geogr.* 15 (3), 223–245.
- Emanuel, M., Oldenziel, R., Schipper, F. (Eds.), 2020. *Sustainable Urban Mobility since 1850s: Roots of Today's Challenges*. Berghahn, New York.

- Harpaz, I., 2002. Advantages and disadvantages of telecommuting for the individual, organization and society. *Work Study* 51 (2), 74–80.
- Heinen, E., Van Wee, B., Maat, K., 2010. Commuting by bicycle: an overview of the literature. *Transp. Rev.* 30 (1), 59–96.
- Hislop, D. (Ed.), 2008. *Mobility and Technology in the Workplace*. Routledge, London.
- Jones, D., 2010. *Mass Motorization + Mass Transit: An American History and Policy Analysis*. Indiana University Press, Bloomington, IN.
- Lin, X., Wells, P., Sovacool, B., 2017. Benign mobility? Electric bicycles, sustainable transport consumption behaviour and socio-technical transitions in Nanjing, China. *Transp. Res. Part A Policy Pract.* 103, 223–234.
- Lopez-Galviz, C., 2019. *Cities, Railways, Modernities: London, Paris, and the Nineteenth Century*. Routledge, London.
- Oldenziel, R., Emanuel, M., de la Bruheze, A., Veraart, F., 2016. *Cycling Cities: The European Experience: Hundred Years of Policy and Practice*. Foundation for the History of Technology, Eindhoven.
- Reddy, B., Balachandra, P., 2012. Urban mobility: a comparative analysis of megacities of India. *Transp. Policy* 21, 152–164.

Shopping and Transport Modes

Antonio Comi, Department of Enterprise Engineering, University of Rome Tor Vergata, Rome, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	98
Shopping Travel Patterns	99
In-Store Versus On-Line Shopping	99
Travel Pattern	100
Factors Affecting Shopping Mode Choice	101
Empirical Evidence	101
A Probabilistic Behavioral Mode Choice Model	103
Conclusions	103
References	104

Introduction

In 2019, 55% of the world's population lived in urban areas, increasing from 25% in 1950. It is estimated that by the year 2050, 68% of the world's population will be living in towns and cities (UN, 2020). If we consider absolute values (rather than just percentages) and the growth rate of the world's population from 1950 to 2050, the figure becomes even more significant.

Urban population growth and transformation of cities into megalopolises pose formidable problems in all countries. Urban conglomerations are becoming ever more common on a global basis, and there has been a sizeable increase in the number of cities, both those above one million inhabitants and those between 500,000 and 1,000,000. By 2030, the world is projected to have 43 megacities with more than 10 million inhabitants, most of them in developing regions (UN, 2020).

As the world continues to urbanize, sustainable development depends increasingly on the successful management of urban growth. Many countries will face challenges in meeting the needs of their growing urban populations as regards housing, transportation, energy systems and other infrastructure, as well as employment and basic services such as education and health care. Integrated policies are needed to improve the lives of both urban and rural dwellers, while strengthening the linkages between urban and rural areas, and building on their existing economic, social and environmental ties.

To ensure that the benefits of urbanization are fully shared and inclusive, policies to manage urban growth need to ensure access to infrastructures and social services for all, focusing on the needs of urban people for housing, education, health care, and working and living in a sustainable environment.

Therefore, the processes of urban reorganization have often entailed a division of functions, with the separation of residential from commercial/shopping areas. The emergence of out-of-town shopping malls has emphasized this separation. Subsequently, understanding the key trends in urbanization likely to unfold over the coming years is crucial to the implementation of the 2030 Agenda for Sustainable Development, including efforts to forge a new framework of urban development.

Such population concentrations and changes in the social structure, combined with the need for sustainable mobility, heighten problems connected to the purchasing of goods by end users (shoppers) and the restocking of retail outlets. Indeed, this segment of mobility significantly contributes to unsustainability as shown by the extensive literature (e.g., Russo and Comi, 2016; Iwan et al., 2018). Trips for purchases are, in percentage terms, ever closer to those of home-work, exceeding in urban areas systematic home-work trips. In the US in 1983 annual trips per household averaged 537 to/from work and 474 for shopping. By 2009 this had changed to 541 to/from work and 725 for shopping (NHTS, 2009). According to Santos et al. (2011), about 20% of all household travel and 15% of vehicle miles traveled in the US are associated with shopping. In 2014, for instance, shopping was the most common trip purpose accounting for 19% of all trips in England (DfT, 2015). Shopping travel is driven by the dynamics of the retail sector and hence highly influenced by its transformation. For example, average distances traveled for shopping trips have increased significantly in the UK over the past 20 years with the advent of out-of-town retail developments (DfT, 2015; Suel and Polak, 2017).

Besides, given that, as detailed in the sections below, this share of mobility is almost always performed by car, traffic impacts upon cities become significant. As revealed by surveys carried out in the larger European cities (Gonzalez-Feliu et al., 2012; Schoemaker et al., 2006), 69% of urban distances (veh * km) covered each day by motorized vehicles for freight transport consist of shopping trips, 24% of restocking trips (e.g., deliveries to shops, hotels, restaurants or catering activities) and the remaining 7% result from urban management (e.g., building sites, waste collection, network maintenance).

The traditional structure of restocking of neighborhood shops, and hence of buying goods in such shops, is being increasingly modified by new forms of purchase and distribution, resulting from the Internet. In the US, e-commerce sales in the third quarter of 2019 accounted for 11.2% of total sales. In the first quarter of 2010, it was 4.2% (Census, 2020). End users have modified their behavior and purchasing decisions over time, gradually switching from traditional distribution channels. Indeed, innovation in retailing and growth in online shopping infrastructure have changed the nature of shopping in response. Lee et al. (2017) have also

shown that although online shopping has grown significantly over the past few years, traditional stores still dominated the retail market share in the US in 2014.

Given the ever-increasing importance of focusing on behavioral aspects of shopping mobility and the chance to model them, the general objective of this chapter is to point out the behavioral aspects of urban shopping, mapping travel behavior of the end consumer, especially as regards mode choice.

Starting from an analysis of existing studies on shopping processes, the chapter aims:

- to provide an overview of shopping travel patterns, highlighting the shopping choices made by end consumers to purchase in store and on line (Section “Shopping Travel Patterns”);
- to analyze the shopping travel behavioral process, using a modeling framework that allows a linkage between end-consumer choices and restocking choices made by retailers, useful for further urban freight transport analysis (Section “Shopping Travel Patterns”);
- to outline mode choice for shopping and point out the factors that affect it (Section “Factors Affecting Shopping Mode Choice”).

Finally, conclusions and the road ahead are drawn in Section “Conclusions”.

Shopping Travel Patterns

Below, we focus on freight trips due to consumer behavior and not on the overall behavior of consumers as regards shopping (for purchases). This initial restriction is required in order to differentiate the contents of this contribution from studies concerning purchasing decisions which constitute a broad field of research for all those dealing with marketing and related sciences.

First let us recall the main aspects of the structure of the shopping process in order to highlight the phases that entail movements of goods and users. Below we only deal with phases that are directly correlated with goods movements. The decisional sequences in the shopping process have been investigated elsewhere. Here, we refer to the literature revisited by [Mokhtarian \(2004\)](#) and [Russo \(2013\)](#), since its breakdown allows the phases connected with goods movements to be identified. The overall phases are: desire; information gathering and receiving; trial/experience; evaluation; selection; transaction; delivery/possession; display/use; return.

[Mokhtarian \(2004\)](#), moving on from [Couclelis \(2001\)](#), identified three main shopping phases: before, purchase, and after, each of which must be carried out in-store or by e-commerce, obtaining $2^3 = 8$ possible patterns. The eight combinations obtained define as many types of users:

- the traditional shopper who performs the three phases inside the shop (store/store/store),
- the innovative consumer or cybernaut who conducts all the phases by internet (e-commerce/e-commerce/e-commerce),
- the new types like the careful, well-informed user (e-commerce/store/e-commerce),
- the bargain hunter or free rider (store/e-commerce/store).

An even more complex segmentation is proposed by [Cao \(2012\)](#) who considers 16 alternatives to compare e-shopping and store shopping, starting from the basic four indicated by Mokhtarian (2004): Information, Trial, Transaction, and Delivery.

These segmentations are particularly important because they allow consumer behavior to be first fragmented and then reconstructed in respect of the increasingly imposing use of purchase channels connected with the Internet. They also highlight the current and potential impact of the Internet upon all the phases of the shopping process. The above segmentations, and the others found in the literature, do not allow direct specification of trips that occur in relation to the consumption of goods. Although they may be used as a starting point, further initial definitions should be set and specific hypotheses introduced on the patterns of movements and the behavior which generates them. This task lies beyond the scope of this chapter; for more details refer to [Russo \(2013\)](#) and references quoted therein.

In-Store Versus On-Line Shopping

Shopping consists of a heterogeneous set of activities and involves many different behavioral mechanisms with several implications on mobility ([Girard et al., 2003](#); [Mokhtarian, 2004](#); [Visser and Lanzendorf, 2004](#); [Rotem-Mindali and Salomon, 2007](#); [Russo, 2013](#); [Suel and Polak, 2017](#)). In-store shopping as part of daily life enables individuals both to acquire certain items and to fulfill needs for social interaction and entertainment. Individuals also maximize their daily trip chain by including shopping activities in their trips ([Mokhtarian, 2004](#)). Such personal and social needs consist of aspects such as fulfillment of social interaction, or social experiences outside home, communication, attaining status, and getting pleasure from bargaining ([Joewono et al., 2019](#)).

The changes to daily life in the digital era because of the development of ICT have altered the ways in which people carry out activities, including shopping. [Mokhtarian and Salomon \(2001\)](#), [Mokhtarian \(2004\)](#), [Hsiao \(2009\)](#), [Nuzzolo and Comi \(2014a\)](#), and [Schmid and Axhausen \(2017\)](#) have reported the effects of ICT on shopping and travel behavior. E-commerce not only has implications on individual mobility, but also on the complementary side, namely freight restocking ([Russo, 2013](#)). The above studies showed that ICT does not necessarily replace all in-store shopping or other trips, and concluded that: (1) only certain factors related to social and personal motivation can be replaced by ICT services (e.g., the latest trends, fashions, and innovations) while

others cannot (e.g., social interactions); and (2) rates of in-store shopping remain solid, attracting individuals in certain goods sectors, such as groceries (Farag, 2006; Farag et al., 2007; Comi and Nuzzolo, 2016).

As stated above, several authors over the years have explained the attractions of in-store shopping travel and shopping activities. They showed that in-store shopping allows sensory information to be obtained by seeing, trying, feeling, smelling, tasting or manipulating the desired items, providing individuals with full information about and confidence in the product. Besides, it has been found that a lack of trust is a barrier to online shopping. For instance, individuals (end consumers) tend to choose established stores with many years of trading in their town rather than e-shopping retailers, who are unknown and may go out of business without warning. Lastly, the instant possession of items is important; individuals can immediately own the purchased items, rather than waiting for the items to be delivered after e-shopping (Joewono et al., 2019). For the above reasons, some freight types, such as electronic items, attract more e-purchasing than others (Comi, 2020).

In addition to the dimension of competitiveness of in-store shopping as opposed to e-shopping, external components of the in-store shopping process also have substantial influence. Shopping as a physical activity provides opportunities to meet needs for social engagement, since going shopping offers time to enjoy oneself, either alone or with friends and family. In-store shopping can also be a source of entertainment, rather than solely a maintenance activity. The function and facilities provided by modern stores (e.g., restaurants, theatres, cafés, etc.) present positive opportunities for enjoying in-store shopping. Furthermore, some shopping trips form part of trip chains for other purposes (Mokhtarian, 2004; Comi, 2020).

Travel Pattern

Individual behavior aiming to fulfill needs and desires is manifested in certain activity and travel patterns within the constraints of time and space. Daily activities (mandatory or discretionary) require individuals to make decisions about the location, participation, timing, and duration of such activities (Susilo, 2005). Various needs and desires among individuals result in diverse activity arrangements and travel patterns, and this is a topic of interest even if the literature in question is far from extensive (Axhausen et al., 2002). Moreover, the pattern of travel to support shopping activities is subject to considerable variability. Differences across individuals in the variations of their shopping travel patterns across days (e.g. day-of-the-week evolution) as well as the heterogeneity of individual characteristics have also been pointed out (Sun et al., 2017). For example, Kim and Park (1997) found that routine shoppers are identified as having high opportunity and search costs, which are represented by four demographic variables (i.e. gender, employment status, level of education and whether household has children below 6 years of age).

Moreover, land use plays an important role in spatio-temporal constraints on individuals. The structure and characteristics of the built environment can influence the way in which individuals participate in activities, due to their effect on the location of such activities. The link between the built environment and travel behavior has been widely investigated, amongst others by Ewing and Cervero (2001), Reilly and Landis (2003), Schwanen and Mokhtarian (2005), Handy et al. (2005), and Cao et al. (2006).

Based on the above statements and following Comi et al. (2014), and the fact that about 70% of daily shopping trips are home-based (i.e., ring trips) and few are part of trip chains (combining trips undertaken for different purposes), the global demand function for simulating travel for shopping can be decomposed into sub-models, each related to one or more choice dimensions: trip production, shop type and location, mode and path choice. Of course, such a general framework can receive input from upstream modeling frameworks which identify the needs to buy (purchase modeling; Comi, 2020) (Fig. 1).

Trip production or trip frequency estimates the mean number of “relevant” trips undertaken for shopping by a given category of end consumers (e.g., students, employees). Trip generation is mainly affected by socio-economic characteristics and land-use patterns (or the physical characteristics of the area; Cubukcu, 2001; Yao et al., 2008; Nuzzolo et al., 2014), as well as by freight type (e.g., foodstuffs, clothing).

The *shop type and location stage* simulate the choice of a “shop” type (e.g., nearby/small shop, local market, hypermarket) and a destination among possible alternatives where purchasing can be carried out (e.g., shopping areas/zones; Gonzalez-Benito, 2004;

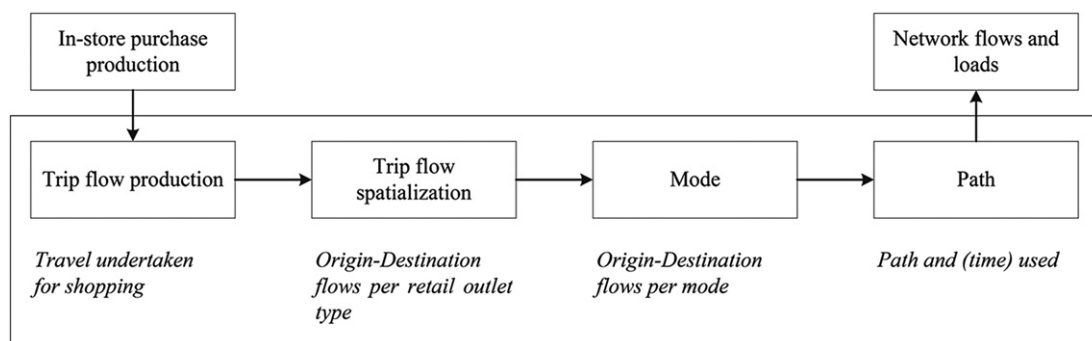


Figure 1 Shopping travel: modeling structure (based on Comi et al., 2014).

Nuzzolo and Comi, 2014a). Such a decision can be influenced both by the built environment and by socio-economic attributes. For example, younger consumers prefer large retail outlets (even if far away) as they may be driven to search for special prices, while the elderly are more devoted to small retail outlets located nearby (Nuzzolo and Comi, 2014a and b).

The *mode choice stage* simulates the fraction (probability) of trips made by end consumers using a given transportation mode. Mode choice is a typical example of a travel choice that can be modified for different trips in which performance or level-of-service attributes have considerable influence as well as shoppers' attitudes, departure time and attractivity of destination (Vrtic et al., 2007; Suel and Polak, 2017; Ramezani et al., 2019; Comi, 2020).

The *path used* should also be investigated. Path choice behavior and the model representing it depend on the type of service offered by the different transport modes. Several models have been proposed both for private or public transport. For an overview we refer to Cascetta (2009) and Nuzzolo and Comi (2018).

Given that shopping trips are mainly performed by car and one of the main policies of urban/regional planners is to push users towards more environment-friendly modes (SUMP 2019), the factors that affect mode choice are pointed out below.

Factors Affecting Shopping Mode Choice

As said, mode choice is a typical example of a travel choice that can be modified for different trips in which performance or level-of-service attributes have considerable influence. Below, the empirical evidence from different urban contexts worldwide is recalled; a behavioral choice model for shopping is then presented.

Empirical Evidence

Some examples of the investigation of shoppers' attitudes to the various transport modes for shopping purposes may be found in Vrtic et al. (2007), Bilková et al. (2016), and Joewono et al. (2019), who analyzed modes for shopping by applying discriminant analysis to explore the profiles of end consumers/shoppers. The study showed that the *private car* is the most typical mode used by shoppers to reach the most frequently visited grocery store. This confirms the attitudes revealed in the past: one of the first studies on shopping mode choice was performed by Marjanen (1995) who pointed out the influence of *socio-economic characteristics* of shoppers. Marjanen (1995) analyzed the shopping behavior of consumers in Turku, Finland, finding that the use of cars for shopping trips increased with income and family size. Marital status was also found to have a significant effect: those who were married or living together used cars more often than those who were single. However, *gender* was not found to be a significant factor. *Education* was also found to be a significantly influencing factor as the better educated were found more likely to use a private vehicle. Handy (2000) also compared mode choice behavior based on socio-demographics and *household* characteristics. Comi and Nuzzolo (2014), through some data collected in Rome (Italy), found that younger users (i.e., between 15 and 29 years old) present the highest share among those using transit (Table 1). The average travel time by transit was about 35 min, while by car it was about 24 min and on foot about 14 min.

In addition, Meena et al. (2019), focusing on shopping activities in Mumbai, revealed that the percentage of car usage increased with the higher *age* group. The age cohort of 18–30 years are physiologically more active; hence the highest percentage of public transport and two wheeler (referring to motorcycles only) usage were reported for this age group as compared to the other age groups.

Income was also shown as one of the main factors influencing the choice of private cars for shopping. Although, in some developed countries the trend is to limit car ownership and financial status cannot be derived from this variable, generally the choice of car use directly indicates the affluence and financial status of users in terms of car ownership.

El-Bany et al. (2014) investigated the effect of bus rapid transit (BRT) as a hypothetical mode with respect to existing modes (car and taxi) by using a stated preference (SP) technique to analyze the effect of socio-demographic characteristics on the mode choice behavior of shopping trips. The results from the study concluded that income is the most dominant variable affecting mode choice behavior in Port-Said city.

Joewono et al. (2019) pointed out the relationships between shopping *expenditure* and mode in developing countries. Shoppers who spend little money are more likely to walk or cycle (i.e., to shops close by) to reach their shopping location than to use a private car. In contrast, shoppers who spend more tend to travel by car. This supports previous findings on the strong relationships between income and car availability (Marjanen, 1995; Meena et al., 2019).

Table 1 Distribution of weekly shopping trips by age and transport mode (Comi and Nuzzolo, 2014)

	<i>Under 29 years</i>	<i>30–64 years</i>	<i>Over 64 years</i>	<i>Average</i>
Private vehicle (e.g., car)	67%	78%	71%	74%
Transit	14%	4%	2%	7%
On foot	19%	18%	27%	19%
Total	100%	100%	100%	100%

Table 2 Mode shares for weekly purchase trips (Nuzzolo and Comi, 2014a)

	<i>Travel distance</i>		
	<i>Less than 1 km</i>	<i>Between 1 and 2 km</i>	<i>More than 2 km</i>
<i>Small retail outlet</i>			
<i>On foot</i>	37.7%	7.1%	6.6%
<i>Private vehicle (e.g. car)</i>	59.6%	79.6%	62.3%
<i>Transit</i>	2.7%	13.3%	31.1%
Total	100.0%	100.0%	100.0%
<i>Medium-size retail outlet</i>			
<i>On foot</i>	32.5%	12.1%	12.5%
<i>Private vehicle (e.g. car)</i>	66.5%	72.7%	68.8%
<i>Transit</i>	1.0%	12.1%	18.8%
Total	100.0%	100.0%	100.0%
<i>Large retail outlet</i>			
<i>On foot</i>	8.8%	2.5%	0.0%
<i>Private vehicle (e.g. car)</i>	87.3%	92.4%	92.9%
<i>Transit</i>	3.9%	5.1%	7.1%
Total	100.0%	100.0%	100.0%

Jiao (2016) showed that end consumers' attitudes toward travel mode, and the network distance between homes and retail outlets exerted the strongest influence on travel frequency while urban form variables only had a modest influence. The study showed that frequent shoppers were more likely to use alternative transportation modes and shopped closer to their homes, and infrequent shoppers tended to drive longer distances to stores and spent more time and money per visit.

According to data collected in Rome (Comi, 2020), the shares among the three main city modes (i.e., on foot, car, and transit) show that, as expected, the main transport mode used is car: its share increases when shop size and trip length rise (Table 2). Younger users account for the highest share of those using transit. The average time spent traveling by car was about 24 min, by transit about 35 min and on foot about 14 min. Besides, it emerged that 74% of users use private vehicles and only 7% use transit (e.g., bus, metro; Tables 3 and 4).

Fun and Enid (2006) analyzed shoppers' perceptions of various public transport modes in Hong Kong, namely bus, west-rail, mini-bus, taxi, and mass-transit railway (MTR and KCRC). The study concluded that shop characteristics and shopping purpose of the trip maker affect respondent's mode choice. Similar results were obtained by Newmark et al. (2004) who analyzed the change in shopping travel behavior before and after the construction of four new shopping malls in the city of Prague, and by Ding et al. (2014) who analyzed the joint choice behavior of shopping destination and travel-to-shop mode for the study area of the

Table 3 Distribution of weekly shopping trips by type of retail outlet and transport mode (Comi and Nuzzolo, 2014)

	<i>Private vehicle (e.g. car)</i>	<i>Transit</i>	<i>On foot</i>	<i>Total</i>
Small retail outlet	66%	11%	23%	100%
Medium retail outlet	68%	4%	28%	100%
Large retail outlet	91%	5%	4%	100%
Total	74%	7%	19%	100%

Table 4 Distribution of weekly shopping trips by activity at origin and transport mode

<i>Activity at origin</i>	<i>Private vehicle (e.g., car)</i>	<i>Transit</i>	<i>On foot</i>	<i>Total</i>
Home	77%	5%	18%	100%
Work/Study	73%	9%	17%	100%
Personal services	63%	13%	25%	100%
Other	73%	13%	13%	100%
Total	74%	7%	19%	100%

Maryland-Washington DC region. They found that high parking cost at the shopping centers significantly impacted the mode choice behavior of car users.

Suel and Polak (2017) indicated that attractiveness of the walking alternative declines with the number of items in the basket (i.e., walking mode becomes less attractive for large basket shopping). Public transport also proved attractive, suggesting that walking and public transport modes are preferred to driving after other effects are accounted for, which might partially be related to lower costs associated with walking and public transport use. Further, the inconvenience associated with the driving mode due to parking availability and fees might be influential. Public transport might also be attractive for certain routes, although time spent on public transport is associated with more discomfort when compared to driving or walking. For public transport mode, increasing accessibility (as measured by the interaction with the variable for the number of stores within a 45 min travel time) reduces the negative effect of choosing public transport. In-depth analyses showed that driving becomes less attractive when one has several options within easy reach of the home location.

A Probabilistic Behavioral Mode Choice Model

According to data collected in the city of Rome (Comi, 2020, to which readers can refer for obtaining survey details), a binomial logit model was estimated. The probability (or fraction) $p[m/od]$ that users select mode m to travel from zone o to zone d for shopping can be calculated as follows:

$$p[m/od] = \exp(V_{m/od}) / \sum_{m'} \exp(V_{m'/od})$$

where $V_{m/od}$ is the systemic utility associated to the alternative m given the origin-destination pair od . In particular, the systemic utilities are expressed as follows:

- car

$$V_{car/od} = -0.122 \cdot T_{car/od} - 0.252 \cdot C_{car/od} + 2.682$$

$\begin{matrix} (-3.21) & (-2.76) & (1.98) \end{matrix}$

- transit

$$V_{transit/od} = -0.032 \cdot T_{transit/od} - 0.786 \cdot C_{transit/od}$$

$\begin{matrix} (-2.45) & (-1.99) \end{matrix}$

where T is the travel time expressed in minutes by *car* or *transit*, while C is the travel cost in Euros. The value in brackets refers to t -stat value. ρ^2 is 0.44. Furthermore, it is important to note the different VOT (Value-Of-Time) values obtained for car and transit, 29€/h and 2.40€/h, respectively, which confirms the findings explained in the earlier section.

Conclusions

Understanding travel mode for shopping becomes crucial for supporting city planners in improving city sustainability and liveability, given that shopping accounts for the second most frequently undertaken trips in most cities and regions. It is widely accepted that there is heterogeneity in shopping travel mode choice due to different factors, taking the influence of telematics into consideration. The literature shows that a number of socio-demographic, psycho-social and travel factors, as well as the built environment, affect shopping trip mode choice. Therefore, there is a need to study diversification in grocery shopping location and frequency that can correctly guide consumer demand and provide some insights for policy making. For example, higher exposure to dense, mixed use and green areas increases the propensity to use more sustainable modes of transportation for shopping trips, especially among older adults.

In addition, travel time was found to have a significant effect on mode choice behavior, as expected. The number of accompanying persons greatly impacted mode choice behavior for private vehicles as well as shopping basket size. Other socio-demographic characteristics like the type of occupation of the individual, age and gender were significant for respective mode utilities.

The indications from the literature can be considered to provide a useful shopping-related sustainability overview for city authorities and experts when designing measures for shifting consumers to more sustainable transport and needing to identify the main factors impacting non-systematic travels. Such an analysis could contribute to a more effective and rational planning process, helping city planners or administrations to select environmental/green measures expected to perform best according to the city's main characteristics. Therefore, the proposed analysis may be a useful preliminary guide when there is the need to ascertain whether a successful approach in one city can be transferred to another.

Indications for further analysis may also be outlined. First, the existing literature mainly considers only grocery shopping, without covering other daily activities. Therefore, future studies should investigate in greater depth grocery shopping as part of the

user's entire daily activities. Second, research should aim to investigate the extent of grocery shopping from day to day for a whole week, distinguishing between durable and nondurable goods as well as daily/weekly consumption purchases. Such extended studies may provide knowledge allowing researchers/planners to suggest how transportation systems should be managed in the digital age with its rapidly occurring changes. Third, suitable measures of travel demand management to support shopping activities should be explored with a view to making transport in question more sustainable.

References

- Axhausen, K.W., Zimmermann, A., Schönfelder, S., Rindsfuser, G., Haupt, T., 2002. Observing the rhythms of daily life: a six-week travel diary. *Transportation* 29, 95–124.
- Bilková, K., Křížan, F., Barlik, P., 2016. Consumers preferences of shopping centers in Bratislava (Slovakia). *Human Geographies–J. Stud. Res. Hum. Geogr.* 10, 23–37.
- Cao, X., 2012. The relationship between e-shopping and store shopping in the shopping process of search goods. *Transport. Res. Part A* 46, 993–1002.
- Cao, X., Handy, S.L., Mokhtarian, P.L., 2006. The influences of the built environment and residential self-selection on pedestrian behavior: Evidence from Austin, TX. *Transportation* 33, 1–20.
- Cascetta, E., 2009. *Transportation Systems Analysis: Models and Applications*. Springer.
- Census. (2020). Estimated quarterly US Retail Sales: Total and e-commerce. US Census Bureau News.
- Comi, A., Nuzzolo, A., 2014. Simulating Urban Freight Flows with Combined Shopping and Restocking Demand Models. In *Procedia–Social and Behavioral Sciences* 125, doi:10.1016/j.sbspro.2014.01.1455, Elsevier, pp. 49–61.
- Comi, A., Nuzzolo, A., 2016. Exploring the Relationships between E-Shopping Attitudes and Urban Freight Transport. In *Transportation Research Procedia* 12, doi:10.1016/j.trpro.2016.02.075, Elsevier, pp. 399–412.
- Comi, A., Donnelly, R., Russo, F., 2014. Urban freight models. In *Modelling Freight Transport*, Tavasszy, L., De Jong, J. (Eds.), chapter 8, doi:10.1016/B978-0-12-410400-6.00008-2, Elsevier, pp. 163–200.
- Comi, A., 2020. A modelling framework to forecast urban goods flows. *Res. Transp. Econ.* DOI: 10.1016/j.retrec.2020.100827, Elsevier Ltd.
- Couclelis, H., 2001. Pizza over the internet: E-commerce, the fragmentation of activity, and the tyranny of the region. Paper presented at the workshop on entrepreneurship, ICT and the region, Amsterdam, June 7–8.
- Cubukcu, K.M., 2001. Factors affecting shopping trip generation rates in metropolitan areas. *Stud. Reg. Urban Plan.* 9, 51–68.
- DfT, 2015. *National Travel Survey: England 2014*.
- Ding, C., Xie, B., Wang, Y., Lin, Y., 2014. Modeling the joint choice decisions on urban shopping destination and travel-to-shop mode: A comparative study of different structures. *Discrete Dyn. Nat. Soc.* 2014, 1–10.
- El-Bany, M., Shahin, M., Hashim, I., Serag, M., 2014. Policy sensitive mode choice analysis of Port-Said City Egypt. *Alexandria Eng. J.* 53 (4), 891–901.
- Ewing, R., Cervero, R., 2001. Travel and the built environment: A synthesis. *Transp. Res. Rec. J. Transp. Res. Board* 1780, 87–114.
- Farag, S., 2006. *E-Shopping and Its Interactions with In-Store Shopping*. Ph.D. Thesis, Faculty of Geosciences, Utrecht University, Utrecht, The Netherlands.
- Farag, S., Schwanen, T., Dijst, M., Faber, J., 2007. Shopping online and/or in-store? A structural equation model of the relationships between e-shopping and in-store shopping. *Transp. Res. Part A Policy Pract.* 2007 41, 125–141.
- Fun, N., Enid, 2006. The effect of transportation options on shopping behavior in Hong Kong. Degree of Master of Housing Management. The University of Hong Kong.
- Girard, T., Korgaonkar, P., Silverblatt, R., 2003. Relationship of type of product, shopping orientations, and demographics with preference for shopping on the Internet. *J. Bus. Psychol.* 18, 101–120.
- Gonzalez-Benito, O., 2004. Random effects choice models: seeking latent predisposition segments in the context of retail store format selection. *Omega* 32, 167–177.
- Gonzalez-Feliu, J., Ambrosini, C., Pluvinet, P., Toilier, F., Routhier, J.L., 2012. A simulation framework for evaluating the impacts of urban goods transport in terms of road occupancy. *J. Comput. Sci.* 3 (4), 206–215, Elsevier.
- Handy, S., 2000. Non-work travel of women: Patterns, perceptions, and preferences. In *Women's Travel Issues Second National Conference*, pp. 317–334.
- Handy, S., Cao, X., Mokhtarian, P., 2009. Correlation or causality between the built environment and travel behaviour? Evidence from Northern California. *Transp. Res. Part D Transp. Environ.* 10, 427–444.
- Hsiao, M.H., 2009. Shopping mode choice: physical store shopping versus e-shopping. *Transp. Res. Part E* 45, 86–95.
- Iwan, S., Kijewska, K., Johansen, B.G., Eidhammer, O., Małeck, K., Konicki, W., Thompson, R.G., 2018. Analysis of the environmental impacts of unloading bays based on cellular automata simulation. *Transport. Res. Part D: Transport Environ.* 61, 104–117.
- Jiao, J., Does urban form influence grocery shopping frequency? A study from Seattle, Washington, USA. In *International journal of retail & distribution management* Vol. 44, 2016, pp. 903–922.
- Joewono, T.B., Tarigan, A.K.M., Rizki, M., 2019. Segmentation, Classification, and Determinants of In-Store Shopping Activity and Travel Behaviour in the Digitalisation Era: The Context of a Developing Country. *Sustainability* 2019, 11, 1591; doi:10.3390/su11061591.
- Kim, B.-D., Park, K., 1997. Studying patterns of consumer's grocery shopping trip. *J. Retail.* 73, 501–517.
- Lee, R.J., Sener, I.N., Mokhtarian, P.L., Handy, S.L., 2017. Relationships between the online and in-store shopping frequency of Davis, California residents. *Transp. Res. Part A Policy Pract.* 2017 100, 40–52.
- Marjanen, H., 1995. Longitudinal study on consumer spatial shopping behaviour with special reference to out-of-town shopping. *J. Retail. Consumer Serv.* 2 (3), 163–174.
- Meena, S., Patil, G.R., Mondal, A., 2019. Understanding mode choice decisions for shopping mall trips in metro cities of developing countries. *Transport. Res. Part F* 64, 133–146.
- Mokhtarian, P.L., 2004. A conceptual analysis of the transportation impacts of B2C e-commerce. *Transportation* 31, 257–284.
- Mokhtarian, P.L., Salomon, I., 2001. How derived is the demand for travel? Some conceptual and measurement considerations. *Transp. Res. Part A Policy Pract.* 35, 695–719.
- Newmark, G., Plaut, P., Garb, Y., 2004. Shopping travel behaviors in an era of rapid economic transition: Evidence from newly built malls in Prague, Czech Republic. *Transport. Res. Rec.: J. Transport. Res. Board* 1898, 165–174.
- NHTS, 2009. *National Household Travel Survey Average Annual PMT (Person Miles of Travel), Person Trips and Trip Length by Trip Purpose*. Table 5.
- Nuzzolo, A., Comi, A., 2014a. A system of models to forecast the effects of demographic changes on urban shop restocking. In *Research in Transportation Business & Management* 11, DOI: 10.1016/j.rtbm.2014.03.001, Elsevier, pp. 142–151.
- Nuzzolo, A., Comi, A., 2014b. Direct effects of city logistics measures and urban freight demand models. In: Gonzalez-Feliu, J., Semet, F., Routhier, J.L. (Eds.), *Sustainable Urban Logistics: Concepts, Methods and Information Systems*, Springer-Verlag, Berlin Heidelberg, pp. 211–226. DOI: 10.1007/978-3-642-31788-0_11.
- Nuzzolo, A., Comi, A., 2018. A Subjective Optimal Strategy for Transit Simulation Models. In *Journal of Advanced Transportation*, vol. 2018, Article ID 8797328, DOI: 10.1155/2018/8797328, 10 pages.
- Nuzzolo, A., Comi, A., Rosati, L., 2014. City logistics long-term planning: simulation of shopping mobility and goods restocking and related support systems. *Int. J. Urban Sci.* 18 (2), DOI:10.1080/12265934.2014.928601, Taylor & Francis, pp. 201–217.
- Ramezani, S., Laatikainen, T., Hasanazadeh, K., Kyttä, M., 2019. Shopping trip mode choice of older adults: an application of activity space and hybrid choice models in understanding the effects of built environment and personal goals. In *Transportation*, DOI: 10.1007/s11116-019-10065-z.

- Reilly, M., Landis, J., 2003. The Influence of Built-Form and Land Use on Mode Choice; University of California Transportation Center Research Paper, Work Conducted at the Institute of Urban and Regional Development, IURD WP 2002-4; University of California: Berkeley, CA, USA.
- Rotem-Mindali, O., Salomon, I., 2007. The impacts of E-retail on the choice of shopping trips and delivery: Some preliminary findings. *Transp. Res. Part A Policy Pract.* 41, 176–189.
- Russo, F., 2013. Modeling Behavioral Aspects of Urban Freight Movements. In *Freight Transport Modelling*, Moshe Ben Akiva, Hilde Meersman, Eddy Van de Voorde (Eds.), Emerald Group Publishing Limited, chapter 18, United Kingdom.
- Russo, F., Comi, A., 2016. Urban freight transport planning towards green goals: synthetic environmental evidence from tested results. *Sustainability* 8 (4), 381.
- Santos, A., McCuckin, N., Nakamoto, H.Y., Gray, D., Liss, S., 2011. Summary of Travel Trends. 2009 National Household Travel Survey; Federal Highway Administration: Washington, DC, USA, 2011.
- Schmid, B., Axhausen, K.W. (2017). In-store vs. online shopping of search and experience goods: A hybrid choice approach. In *Proceedings of the 5th International Choice Modeling Conference (ICMC 2017)*, Cape Town, South Africa, 3-5 April 2017; IVT ETH: Zurich, Switzerland, 2017.
- Schoemaker, J., Allen, J., Huschek, M., Monigl, J., 2006. Quantification of Urban Freight Transport Effects I. BESTUFS Consortium, www.bestufs.net.
- Schwanen, T., Mokhtarian, P.L., 2005. What affects commute mode choice: Neighborhood physical structure or preferences toward neighborhoods? *J. Transp. Geogr.* 13, 83–99.
- Suel, W., Polak, J.W., 2017. Development of joint models for channel, store, and travel mode choice: Grocery shopping in London. *Transport. Res. Part A* 99, 147–162.
- SUMP, 2019. Guidelines. Developing and Implementing a Sustainable Urban Mobility Plan, 2nd ed. European Commission, Brussels, Belgium.
- Sun, Y., Tarigan, A.K.M., Waygood, E.O.D., Wang, D., 2017. Diversity in diversification: An analysis of shopping trips in six-week travel diary data. *J. Zhejiang Univ. Sci. A* 18, 234–244.
- Susilo, Y.O., 2005. The Short Term Variability and the Long Term Changes of Individual Spatial Behaviour in Urban Areas. Ph.D. Thesis, Department of Urban Management Graduate School of Engineering, Kyoto University, Kyoto, Japan.
- UN (2020). 2019 Revision of World Population Prospects. United Nations, Department of Economic and Social Affairs Population Dynamics, www.un.org.
- Visser, E., Lanzendorf, M., 2004. Mobility and accessibility effects of B2C ecommerce: A literature review. *Tijdschrift voor Economische en Sociale Geografie* 2004 95, 189–205.
- Vrtic, M., Fröhlich, P., Schussler, N., Axhausen, K.W., Lohse, D., Schiller, C., Teichert, H., 2007. Two-dimensionally constrained disaggregate trip generation, distribution and mode choice model: Theory and application for a Swiss national model. *Transport. Res. Part A* 41, 857–873.
- Yao, L., Guan, H., Yan, H., 2008. Trip generation model based on destination attractiveness. *Tsinghua Sci. Technol.* 13 (5), 632–635.

Transport Modes and Health

Jennifer S. Mindell*, **Sandra Mandic***, *Health and Social Surveys Research Group, Research Department of Epidemiology & Public Health, UCL, London, United Kingdom; †Active Living Laboratory, School of Physical Education, Sports, and Exercise Sciences, University of Otago, Dunedin, Otago, New Zealand

© 2021 Elsevier Ltd. All rights reserved.

Glossary	107
Introduction	107
Transport Modes	107
Active Travel	107
Walking	107
Cycling	107
Electric Bikes (e-Bikes)	107
Public Transport	108
Private Motor Vehicles	108
Shared Mobility	108
Taxis, Minicabs, and App-Based Private Companies	108
Car-Pooling and Public Hire Bikes and E-Scooters	108
Well-Being	108
What is Well-Being?	108
Green and Blue Environments	109
Physical Activity	109
Travel Modes and Physical Activity	109
Active Travel	109
Walking Versus Cycling	109
Public Transport	109
Private Motor Vehicles	109
Effects of Active Travel on Mental and Physical Health	109
Air Pollution	110
Air Pollutants	110
Greenhouse Gas Emissions	110
Noise	110
Injury	111
Country	111
Gender	112
Age	112
Deprivation	112
Travel Mode	112
Perception	112
Community Severance	112
Barrier Effect	112
Social Networks	112
Pedestrian Crossings	113
Other Benefits of Active Transport	114
Independent Mobility	114
Cognition	114
Improving Access	114
Inequalities in Transport Modes and Health	114
Distribution of Impacts	114
Age	115
Children and Adolescents	115
Older People	115
Gender and Sexuality	115
Socioeconomic Position	115
Ethnicity	115
Impairments	115

The Future	116
Summary and Conclusion	116
Biographies	116
See Also	116
Relevant Websites	116
References	117
Further Reading	117

Glossary

Community severance “Transport-related community severance is the variable and cumulative negative impact of the presence of transport infrastructure or motorized traffic on the perceptions, behavior, and well-being of people who use the surrounding areas or need to make trips along or across that infrastructure or traffic.”

Gig economy A gig economy is a labor market characterized by short-term contracts and self-employed workers, rather than permanent jobs.

Good mental health “A state of well-being in which every individual realizes (their) own potential, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to (their) community.”

Intersectionality The concept that a person or group of people can be affected by a number of disadvantages and discriminations (e.g., due to their gender, race, and age)

Morbidity Having a disease or symptoms of disease (or a state outside normal health and well-being); the amount of disease in a population

Mortality Deaths or death rates

Introduction

Transport has a wide range of impacts on health, and on inequalities (Mindell et al., 2011a; Cohen et al., 2014); these vary by travel mode. This chapter is organized as follows: Transport Modes describes the key elements of different transport modes before the further sections discuss the various key health impacts related to these. Other Benefits of Active Transport outlines how travel modes impact on health inequalities and lack of social justice, before Inequalities in Transport Modes and Health concludes with a brief mention of potential health considerations of future transport options currently being encouraged.

Transport Modes

Active Travel

Active travel is defined as using human-powered modes of transport. The most common modes are walking and cycling. Other forms of active travel include skateboarding or using a non-motorized scooter. More recently, electric-assisted pedal bicycles (e-bikes) are becoming a popular form of active travel in many high-income countries.

Walking

Walking is a familiar low-cost activity that does not require specialized equipment and is a readily available mode of transport for most individuals. Walking is a feasible form of transport for short trips. For longer trips, walking can be combined with motorized forms of transport such as public transport or private car, with walking included at each end of the travel journey.

Cycling

Compared with walking, cycling is faster; covers greater distances; requires greater motor skills; and has more specific infrastructure requirements. In many high income countries, cycling for transport is less prevalent than walking. Cycling rates for transport have declined in most countries globally over the last several decades. Exceptions to those trends are countries with significant investment in extensive cycling-friendly infrastructure, flat topography, and such as Belgium, the Netherlands, and Denmark.

Electric Bikes (e-Bikes)

In recent years, electric bikes, or e-bikes, have emerged as new forms of transport in many high-income countries. Electric bikes are very similar to regular bicycles, with the addition of a battery and motor. Data on prevalence of e-bikes usage in different countries is limited at this time due to some countries, such as The Netherlands, classifying them in the same category as regular bicycles.

According to the European Cyclists' Federation, e-bikes must:

- 1 Have a maximum speed of 25 kph
- 2 Have a maximum power output of 250 watts; and
- 3 Engage the motor only when the rider is pedaling.

In countries with a strong cycling tradition, most users of e-bikes that meet this definition are older adults or patients being introduced to cycling (who lack the strength, ability or confidence to ride a standard bicycle) and commuters with a particularly long or strenuous journey. However, in other countries, e-bikes are being used predominantly by young males as a cheap form of motorized travel, with little regard for their or pedestrians' safety. In many cases, they exceed the speed limits recommended by the European Cyclists' Federation and are driven by the motor with no cycling (and are really electric motorbikes, but mostly ridden by adolescents and young adult males).

Public Transport

Public transport covers a wide range of options that enable members of the public to access assistance to travel to destinations on payment of a fare, usually on fixed routes. The most widely used public transport modes are buses; trains, light rail, and trams vary more in their availability. For longer journeys, public transport that is affordable, accessible, acceptable, and appropriate can bring the benefits of access with fewer adverse effects imposed on others.

Private Motor Vehicles

Private motor vehicles are often said to provide the most freedom to travel for their owners compared with other travel modes. Due to convenience, speed, and independence, travel by private motor vehicles dominates travel preferences of most age groups. For families, travel by private motor vehicles facilitates trip chaining. For example, parents may be choosing to travel by car to drop off and/or pick up children from school on the way to/from work. However, as motorization increases, traffic congestion can increase traveling times to longer than other travel modes. Despite the freedom and convenience that traveling by car offers, high rates of private motor vehicle travel has a number of negative consequences for health and the environment that are discussed later in this chapter.

Shared Mobility

Taxis, Minicabs, and App-Based Private Companies

Taxis (which may be pedal-powered or motorized), minicabs, and app-based private companies such as Uber and Lyft charge passengers for bespoke journeys without fixed routes. In low-income countries, motorcycles are a common form of taxi and are often referred to as public transport. Licensing regulations differ between countries and types of motorized companies, with greater passenger protection against assault for licensed versus unlicensed drivers.

The "gig economy," characterized by being self-employed or having short-term or "zero hours" contracts is also associated with adverse health effects for those workers. This is the typical labor market for these drivers.

Car-Pooling and Public Hire Bikes and E-Scooters

Unlike taxis, these transport modes do not generally provide door-to-door access as the hired vehicle needs to be accessed before it can be used and then parked afterward. Thus there is often some active travel before the main stage of the journey and may be at the end as well. Hire bikes and e-scooters not requiring a specific docking station can be left anywhere, increasing convenience for the user but potentially causing trip hazards for pedestrians and reducing the opportunities for physical activity for e-scooters.

Well-Being

What is Well-Being?

The World Health Organization (WHO) defines good mental health as "a state of well-being in which every individual realizes (their) own potential, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to (their) community." Studies of commuters have consistently found improved mental well-being among those who commute by cycle or foot, with higher ratings for health, confidence and positive mood (affect). However, cycle commuters in some studies scored higher on distress and fear and lower on security.

Public transport commuters vary in how their journey affects well-being. Traffic congestion, poor reliability, overcrowding and other adverse impacts on comfort, and stress each reduce well-being among public transport users. Generally, car commuters report the lowest well-being.

Regardless of travel mode, enhancing the experience of travel could improve well-being. In addition to health benefits, increasing rates of active travel contribute to increasing the liveability of towns and cities and addressing health inequalities.

Green and Blue Environments

“Green streets” encourage walking because they are a more pleasant environment compared with traditional traffic-dominated roads. Being in, or even seeing, green environments improves mental health and well-being and can speed recovery from illness and surgical procedures. There is increasing evidence that “blue environments” (rivers, canals, lakes, coasts) have similar effects. Although these may be considered more relevant in the context of leisure than travel, journeys made through or past green or blue environments will benefit health.

Older people encouraged to take up cycling on e-bikes improved their mental health more than those using conventional bicycles. This may be due to greater stimulation from greater exposure to the natural environment.

Physical Activity

Travel Modes and Physical Activity

Active Travel

Both walking and cycling, the two most common modes of active travel, are associated with increased physical activity. The substantial health benefits of walking and cycling for transport outweigh detrimental effects of air pollution exposure and traffic incidents (Mueller et al., 2015).

Walking for transport is associated with higher levels of physical activity in children, adolescents, and adults. Although some studies suggest that walking to school may also be associated with increased fitness in children and adolescents, the current evidence is inconclusive.

Cycling for transport is also associated with higher levels of physical activity. In addition, children who cycle for transport also have higher levels of fitness compared with their peers who walk or use motorized transport to school.

Despite electrical assistance, e-bikes that meet the European Cyclists’ Federation definition provide moderate-to-vigorous intensity of physical activity, depending on their usage. Some individuals use e-bike as a starting point, or a “gateway” to being more physically active. However, the most common reasons for purchasing and using e-bikes include increasing the cycling speed and/or range/distance; reducing the effort of cycling in general and up hills; environmental benefits (compared with other motorized travel modes); and increased opportunity to cycle for individuals with health issues.

Walking Versus Cycling

Walking and cycling are often considered together under the term “active travel.” Although most of the health benefits associated with these modes are similar with respect to increasing daily physical activity, walking and cycling are influenced by different individual, social, environment and policy factors (Mandic et al., 2017). For example, compared with walking, cycling is generally regarded as less safe. However, the actual risk of an injury or a fatal event as a result of cycling on the road is much lower than the perceived risk. In children and adolescents, rates of cycling to school and other destinations are influenced by their preferences to cycle (or not); parental confidence in their child’s cycling skills; parental and children’s concern about traffic safety; and social support for cycling from parents and peers. Therefore, interventions for promoting active travel need to consider different approaches for encouraging walking or encouraging cycling.

Public Transport

Public transport is a form of motorized transport. However, most journeys by public transport involve walking- or sometimes cycling- at one or both ends of the journey, to and/or from the bus stop or station. Thus, public transport is sometimes also included in the term “active travel.” Public transport users take 30% more steps per day compared with individuals who rely on cars for transport. Walking or cycling to/from public transit can be sufficient to meet adults’ weekly physical activity guidelines.

Private Motor Vehicles

Individuals traveling using private motor vehicles are sedentary during the journey, and often travel door-to-door, with minimal walking at either end. Cars and car-users generally impose adverse health consequences on others, exacerbating health and social inequalities. In addition, high volume of traffic and/or high traffic speed acts as a major barrier to active travel, especially among children and adolescents.

Effects of Active Travel on Mental and Physical Health

Reliance on active travel contributes to maintenance of healthy body weight and better mental and physical health outcomes. Physical activity, including from active travel, reduces risks of developing or dying from obesity, diabetes, circulatory diseases (heart disease, stroke), some cancers, osteoporosis (bone-thinning), and depression, and increases mental well-being. For example, in walkable neighborhoods, adults using active travel have lower prevalence of diabetes and obesity compared with their peers who rely on motorized transport. Adults who change transport mode from driving to active commuting lose weight; those who change from active commuting to car use increase weight; public transport users lie between these two groups (Martin et al., 2015).

Active travel contributes to creating liveable town and cities, designed for people to enjoy, and therefore contributes to increasing social capital. Active travel facilitates social interactions with peers, including casual contacts such as greeting neighbors; it connects individuals with their neighborhood environment; and can both improve neighborhood safety and contribute to better mental health.

Air Pollution

Air Pollutants

There are 4.2 million deaths each year globally due to ambient (outdoor) air pollution, although a recent European study suggests this may be a considerable underestimate. The WHO has estimated that worldwide, ambient air pollution accounts for—29% of cases of deaths from lung cancer; 43% of all deaths and disease from chronic obstructive pulmonary disease (COPD, previously known as chronic bronchitis and emphysema); 17% of deaths and disease from acute lower respiratory infection (e.g., pneumonia); 25% of deaths and disease from ischemic (or coronary) heart disease (e.g., heart attacks), and 24% of all deaths from stroke.

Both short- and long-term exposure to air pollution increases the risks of developing cardio-respiratory diseases and also causes exacerbations, increasing morbidity and mortality from these diseases. For example, air pollution exposure can increase the frequency and severity of acute asthma attacks, increasing the need for medication and the number of hospital admissions. More recently, evidence has also shown that air pollution also increases the risks of obesity, diabetes, neuro-developmental conditions in children, and neuro-degenerative conditions (such as dementia). Exposure to the major air pollutants can reduce lung function. Maternal exposure is also associated with adverse birth outcomes, particularly low-birth weight, with its attendant adverse health impacts for the child ([RCP, 2016](#)).

The pollutants with the greatest effects on the public's health include particulate matter (PM), nitrogen dioxide (NO₂), ozone (O₃), and sulfur dioxide (SO₂). Most evidence is available for PM less than 10 µm (PM₁₀) and less than 2.5 µm diameter (PM_{2.5}). The WHO estimates that 91% of the global population breathe polluted air, with 80% of urban residents being exposed to ambient air pollution that exceeds the WHO limits.

Other air pollutants include volatile organic compounds (VOCs) from fuel. One of these, benzene, is associated with increasing the incidence of leukemia. Particulates may also carry other irritants and allergens into the nose and lungs, increasing asthma and allergic rhinitis (hayfever).

PM, SO₂, and NO₂ are generated by combustion of fossil fuels. In general, concentrations of pollutants are highest inside motor vehicles, compared with the sides of the road where people walk and cycle, especially in congested traffic where a vehicle's air intake is directly behind another vehicle's exhaust (tailpipe). Buses usually have concentrations that are lower than those found within private vehicles but higher than at the side of the road. However, when breathing more deeply or faster, such as when walking or cycling briskly or uphill, pedestrians' and cyclists' dose of pollutant may be higher than in motorized travel. It is thus important to select routes for walking and cycling and for their infrastructure that are as distant as possible from motor traffic flows but without major inclines. Metro systems also have very high levels of PM that exceed WHO limits. The majority of these PM are iron, from friction between the wheels and the rails when braking. It is currently unknown whether these PM have similar or fewer health impacts than anthropogenic PM from fossil fuel combustion.

PM also arise from brakes and tires, and are thus emitted from electric vehicles as well. Electric vehicles use energy generated remotely. They are not non-polluting, unless the energy is from a renewable source, but the pollutants emitted are generally remote from major populations and thus have fewer health impacts than those generated in urban streets.

Greenhouse Gas Emissions

In many countries worldwide, motorized transport is a major or the leading source of greenhouse gas emissions, leading to climate change. In the past decade, there has already been a large increase in extreme weather events, including floods, hurricanes, and tornadoes, leading to widespread destruction of housing and other infrastructure and, in some cases, to loss of life. In general, low-carbon transport policies to mitigate global climate change would have the co-benefits of improving health ([Mindell et al., 2011b](#)).

In addition to the obvious health impacts of these extreme events, climate change has other serious health impacts. Rising temperatures bring tropical diseases to more temperate zones, due to spread of areas suitable for the insect vectors to live. Dengue fever is increasing in the countries where it is endemic. Both drought and floods can destroy or greatly reduce food crops, leading to hunger and malnutrition. [Fig. 1](#) shows some of these direct and indirect effects, and the factors that can magnify or reduce these impacts. There are likely to be global inequalities in the consequences of climate change. Greenhouse gas emissions from motor vehicles predominantly come from high-income countries, but the health consequences will afflict poorer countries more, with their generally larger populations but fewer resources. Many of the poorer countries are low-lying, so flooding will be even more devastating.

Noise

Motor vehicles are the main source of noise in urban areas in most, if not all, countries. Noise is an irritation that reduces well-being. Noise can also interfere with sleep and concentration, of particular concern, if they impact on mental health. They can also impact

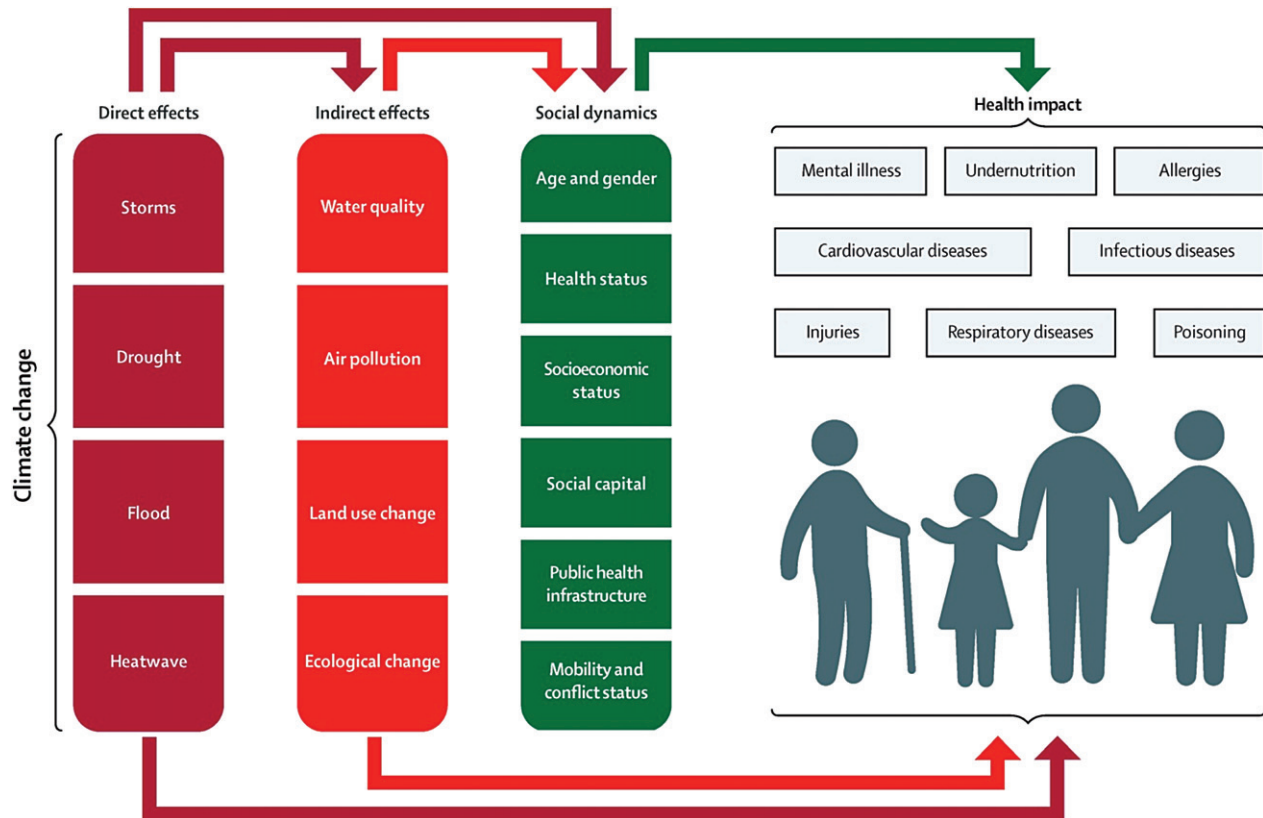


Figure 1 The direct and indirect effects of climate change on health and well-being. Source: Watts et al. (2015)

educational attainment, which is one of the most important social determinants of adult health. Children in classrooms on the noisy side of schools near airports, for example, do less well in tests than their otherwise similar peers in the quieter classrooms. Equally seriously, noise exposure increases blood pressure. Persistently raised blood pressure (hypertension) increases the risks of developing circulatory diseases, particularly heart disease and stroke, and of dying from those diseases.

Noise generated by vehicles has the small health benefit of alerting individuals with adequate hearing to the presence of an approaching vehicle, reducing the likelihood of a collision. The United Kingdom is therefore introducing a requirement for electric vehicles to generate sound at low speeds where the vehicle would otherwise be almost silent. The silence of cycles, scooters and skateboards is also a cause of concern for some pedestrians.

Injury

Road travel injuries are the eighth leading cause of death globally. Worldwide, there are 1.35 million deaths on the road each year in addition to 20–50 million people injured. Many of those injured remain with long-term disabilities, particularly in low-income countries where rehabilitation services are sparse, inaccessible, and/or unaffordable. In many cases, there are no or inadequate welfare payments so if the injured person can no longer work, even subsistence is difficult for them and their dependents, further damaging their own and their family members' health.

Injury rates vary widely by mode but even more by country, gender, age, and deprivation (Feleke et al., 2018). In most low and many middle income countries, numbers (or estimates) are available for deaths and serious injuries but denominators for rates are population figures, at best. In most high-income countries, rates of road travel injuries can be calculated using travel data to provide fatality or injury rates per trip, per billion kilometers (bnkm), or per million hours traveled by a particular mode. This enables better "like-for-like" comparisons.

Country

Low- and middle-income countries have 60% of the world's vehicles but 93% of the world's road travel deaths. The death rate is highest in Africa. Death rates vary from 2.7×10^{-5} population in Norway to 35.9×10^{-5} population in Liberia.

Gender

Males consistently have higher fatality rates than females, in low- and in high-income countries and for every mode. Globally, almost three-quarters of road traffic deaths occur in males under the age of 25 years.

Age

Road travel injuries are the leading cause of death for people aged 5–29 years but older people have much higher injury and death rates than other adults. Although there are reasons for considering older people to be at greater risk of involvement in a collision (impaired hearing, vision, balance, cognition and/or slower reaction times), the main reason for this higher road travel death rate is probably poorer preexisting health. Greater frailty and more preexisting conditions reduce their physiological reserve and their capacity to recover from injury. In Great Britain, pedestrians and cyclists in their 70s and 80s had 10 times the mortality rates bnkm^{-1} than those in their 40s.

Deprivation

People who live or are traveling in more deprived areas have a greater risk of injury and death than their more affluent peers. In Great Britain, this applied to drivers and pedestrians but not to cyclists. In low-income countries, pedestrians, motorcyclists, and cyclists make up the majority of those injured or killed on the roads.

Travel Mode

More than half of the people killed worldwide in road travel collisions are pedestrians, cyclists, or motorcyclists. This is particularly the case in low- and middle-income countries with relatively few journeys made by motor vehicle. The lack of infrastructure, regulation, and enforcement exacerbates the vulnerability of people who have no option but to walk or cycle. Countries with excellent cycling infrastructure have very low-fatality rates among cyclists. Motorcyclists generally have a fatality rate an order of magnitude higher than that of other travel modes, even in high-income countries.

In high income countries, public transport is generally the safest form of transport but elsewhere, overcrowding, poor design and maintenance, and lack of regulation, and enforcement contribute to high risks of collision and of injuries. Car occupants are the main group killed in countries where private vehicles predominate, because of their sheer preponderance of numbers.

The pattern of fatality and injury rates with age is J-shaped for walking and cycling but U-shaped for car occupants. A study in Great Britain found that not only were young male adults safer when cycling than driving, so were all the other road users.

Perception

Perceptions of the risks of injury or death when cycling are greatly exaggerated. As cycling became more popular in London, media coverage of fatalities increased from very few to almost all (while coverage of motorcycle deaths remained at the same very low level), despite no increase in the fatality rate itself. In Great Britain, fatality and injury rates for walking and cycling are very similar, contrary to public perception.

Fear of injury is a major deterrent to active travel and to parents allowing children to travel independently. Fear of “stranger danger” is also cited often. In the United Kingdom, the child abduction rate (which is tiny) has not changed in 50 years, only the extent of media coverage. Furthermore, few cases are perpetrated by strangers; most are family members or otherwise known to the family.

Community Severance

Barrier Effect

“Transport-related community severance is the variable and cumulative negative impact of the presence of transport infrastructure or motorised traffic on the perceptions, behavior, and well-being of people who use the surrounding areas or need to make trips along or across that infrastructure or traffic” (Mindell et al., 2017). Community severance, also known as the barrier effect of busy roads or of transport infrastructure, has a wide range of effects on how and whether people travel and thus on factors affecting health (Fig. 2). A recent estimate in the UK puts the costs of community severance, from journeys not made, money not spent in local businesses, and health consequences of reduced social networks, at around 3% of GDP, similar to the societal costs of road traffic collisions in most countries. Although car users spend more on average in a single trip, they visit local businesses less frequently than other people. Over a period of time, people traveling on foot, by cycle or by public transport spend more in local businesses than individuals traveling by car.

Social Networks

Community severance was first described by Appleyard and Lintell (1972), in their ground-breaking study of San Francisco streets. They found that the heavier the traffic, the fewer the number of friends and acquaintances residents had among the neighbors in their

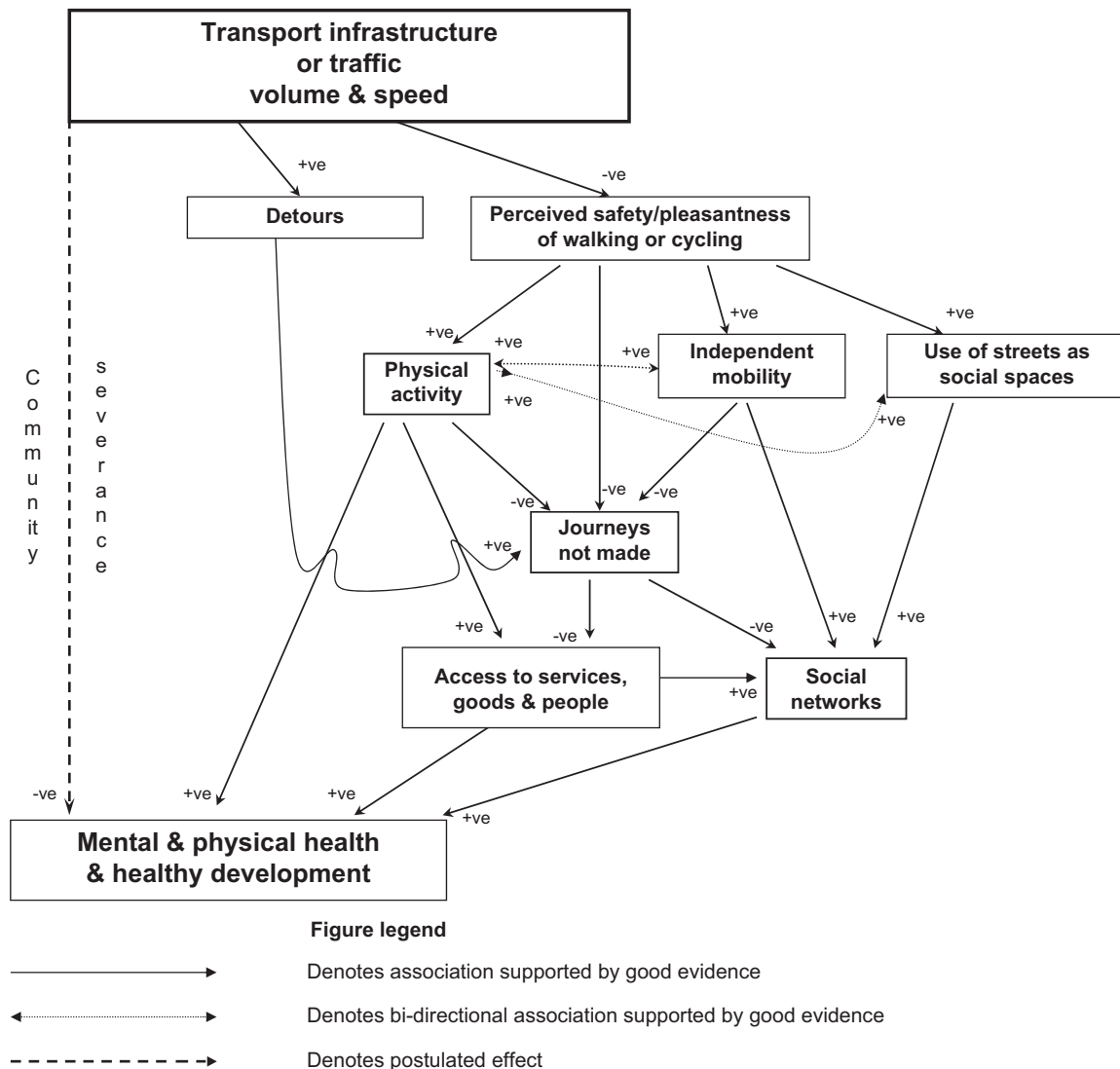


Figure 2 Theoretical model for the health effects of community severance. Source: *Mindell and Karlsen (2012)*

street. They also found that traffic reduced the use of streets as social places to meet, chat, and play, and restricted people's views of their "home territory."

This is important because having an extensive network of social contacts, including nodding acquaintances, has a major impact on health. People with more social contacts are healthier, have fewer hospital admissions, and are less likely to die prematurely. The effects are large, being of similar magnitude to the benefits of stopping smoking (*Holt-Lunstad et al., 2010*).

Pedestrian Crossings

One of the major impacts of community severance is in impeding road crossing. The number, type and location of crossings can affect this considerably. It is particularly a problem for people who walk slowly, due to childhood or old age; have a mobility impairment (whether acute, such as a broken leg, or chronic); have luggage or wheeled baby buggy; or are accompanying children.

Pelican crossings assume a walking speed of at least 1.2 ms^{-1} during the "clearance" time (that allows those who have already started to cross to reach the other side of the road before the motor vehicles resume). However, numerous studies have found that older people, for example, walk much more slowly than this. Typical values are mean walking speeds of 0.9 ms^{-1} for men and 0.8 ms^{-1} for women aged 65 + years (*Asher et al., 2012*). Newer countdown crossings, that give exact information about the time remaining to complete the crossing, and Puffin crossings, that use cameras to keep the signals red for motor traffic while a pedestrian is still crossing, are better solutions. Unsignalised pedestrian "Zebra" crossings can also be a solution, but only where drivers actually stop.

Table 1 Summary of inequalities in transport modes and health

<i>Population group</i>	<i>Health effects of transport</i>					<i>Community severance</i>
	<i>Independence/dependence</i>	<i>Physical activity</i>	<i>Air pollution</i>	<i>Noise pollution</i>	<i>Injury</i>	
Children and adolescents	Limited options	Important but depends on permissions	More susceptible	More susceptible	More susceptible	More exposed
Older people	Limited options	More difficult	More susceptible	More susceptible	More susceptible	More exposed
Gender	Male > female	Cycling mostly male Walking more females			Male > female	Female > male
Sexual orientation	Potential for homophobic attacks on public transport					
Poverty	Limited options	Need to avoid excessively long journeys	More susceptible & more exposed	More susceptible & more exposed	More susceptible & more exposed	More exposed
Ethnicity	Maybe limited options		More exposed	More exposed	More susceptible & more exposed	More exposed
Poor health / impairments	Limited options		More susceptible & more exposed	More susceptible & more exposed	More susceptible	More exposed

Concern about their ability to cross the road safely can prevent older people and others from leaving home if their journey would entail crossing a busy road. This is important because both subjective feelings of loneliness and objective measures of isolation are associated with increased mortality.

Other Benefits of Active Transport

Independent Mobility

In children and youth, active travel to school and other destinations provides an opportunity for independent mobility, which further helps young people to develop their spatial processing skills and skills for navigating safely environments with traffic. Independent mobility and the freedom to play outdoors also enhances musculoskeletal development and self-esteem.

Cognition

Both physical activity and environmental stimulation can improve cognition, as well as well-being, in older people. A recent study demonstrated improved cognitive function in older people who took up cycling (on conventional or e-bikes) for at least half an hour at least three times a week. E-bike users also improved their processing speed (reaction times).

Improving Access

Active travel can also equitably improve access to education, employment, and healthcare. This is particularly important for disadvantaged groups and in areas where public transport is not available, not affordable, or not easily accessible.

Inequalities in Transport Modes and Health

Distribution of Impacts

The health effects described above are inequitably distributed within the population, due to a range of factors. In general, benefits accrue to wealthier car drivers, while adverse effects are experienced mostly by more deprived groups (Mindell et al., 2011a). Car-dependent societies increase inequalities as young people, those with impairments through disease or ageing, and those too poor to afford to own or run a car have difficulties accessing preferred destinations. Intersectionality is also important—some groups of the population have multiple disadvantages. This applies, for example, to those who are particularly vulnerable to adverse health consequences of climate change, which are likely to impact particularly on all the disadvantaged groups described below. **Table 1** summarizes inequalities in transport modes and health related to age, gender and sexuality, socioeconomic position, ethnicity, and impairments are provided in **Table 1**.

Age

Children and Adolescents

The benefits of independent mobility for self-esteem, musculoskeletal, and cognitive development have been mentioned. Without parental permission for active travel, young people are dependent on others to drive them or accompany them on public transport until they are old enough and have sufficient funds to use public transport alone.

Older People

As people age, sensory, cognitive and/or other impairments may lead to cessation of driving. Impairments also restrict the ability to walk or cycle—cycling (using adapted cycle, e.g., tricycle) may be possible when walking is not an option due to musculoskeletal or balance problems. Difficulties with wayfinding is an early symptom of dementia. Public transport can be difficult to use if pedestrian crossing facilities do not allow time to cross roads to access bus stops, and for people with difficulties with steps into buses or stairs at stations.

The injury consequences of collisions and falls are often more serious. In many high-income countries, hospital admissions due to falls when walking are more numerous than those due to collision with a vehicle. Pavements are often uneven, and those with slower reactions or poorer sight and balance will be at greater risk of falling. Even where there is a sidewalk, the quality of the surface may be poor. Falls lead to a loss of confidence and further self-imposed constraints on independent mobility. Dependency on others for transport also leads to loss of self-esteem. Travel difficulties are a cause of social isolation, particularly among older people.

Gender and Sexuality

In many countries, men are more likely than women to hold a driver's licence, to own a car, and to have use of a car where there are more adults than cars in a household. Thus the benefits of driving accrue more to men and the health harms from other people's cars are felt more by women.

There is increasing evidence of sexual harassment on public transport and in the street in many high- and low-income countries. This is predominantly targeted at women in general but can also have a homophobic origin. Such harassment is unpleasant, affects mental well-being, and is a deterrent for travel; it can lead to assault with injury, or even rape or death.

Socioeconomic Position

There are health and inequalities impacts of spatial planning that assumes car use. Car ownership and use is generally expensive but may be necessary, particularly in rural areas without other options (or in cities with no provision except for cars). Injury risk is higher for those living in deprived areas, and is more likely to occur in more deprived locations. Poorer areas may not have public transport or it may be unaffordable. In low income countries and in informal settlements, women may walk 2 hrs each way to employment because no public transport or need to choose between bus fare and food. The adverse mental and physical health effects of inappropriately long journeys outweigh the benefits of this degree of physical activity.

Ethnicity

In many countries, members of minority ethnic groups are poorer than average. Some groups have lower educational attainment; even among those with good education, discrimination often leads to lower level, less well-paid jobs for the same qualifications. Thus, minority ethnic groups often have the same health inequalities as are caused by poorer socioeconomic position. In addition, discrimination or racism can affect people's willingness to use different travel modes. Refugees and other immigrants can experience problems with wayfinding or managing public transport systems due to language problems.

Impairments

Almost any type of disease or medical condition can affect a person's ability to travel or their choice of travel mode(s). Difficulties traveling result in loss of the health benefits of access to employment, education, goods, services, and people. Restrictions on travel mode options can be expensive, particularly if taxis or accompanying persons are required. This can lead to journeys not made, with loss of those benefits, or reduced expenditure on other aspects of life important for health. Impairment refers to a loss or lack of function or ability. Disability is socially defined, when a person with an impairment is restricted in what they want to do by the social, cultural, fiscal and built environment in which they live. People with visual impairment cannot drive or cycle (except on the back of a tandem) but can often walk or use public transport, although they may need aids such as a guide dog or a white stick. Shared spaces can be treacherous, particularly if the curb is removed. However, dropped curbs are important for wheelchair users and others with encumbrances such as buggies or luggage, or who cannot manage steps. Individuals with impaired hearing may find wayfinding harder, if directions or public transport information is aural not visual, and may not hear approaching vehicles, whether motorized or cycles.

Musculoskeletal, cardiovascular, and neurological conditions can restrict the ability to walk or cycle. Cardiovascular and neurological conditions may prevent driving. Restrictions are stricter for professional drivers, who may lose their livelihood following a stroke, heart attack, or epileptic fit.

Mental illnesses also impact on the ability to travel. These impacts relate to difficulties managing the travel mode (e.g. wayfinding, understanding timetables or ticket machines); attitudes and/or behavior of bystanders or other travelers; and expense. Train tickets bought on the day are often much more expensive than tickets bought in advance. People with a fluctuating illness cannot know in advance whether or not they will be able to travel on a given day. Anxiety and depression are common conditions; agoraphobia, claustrophobia, or suffering from panic attacks can also impede travel (Mackett, 2019).

Dementia-friendly cities aim to improve wayfinding for people with cognitive loss. People with learning disability can learn to commute unaided but it takes more time to develop and retain these skills. However, despite the expense of providing trainers, it is very important as employment is a major source of self-esteem that improves mental health.

The Future

When considering health impacts of travel and choice of travel mode, it is always important to specify the alternative. For example, if traveling by private car, electric power will impose fewer negative health consequences on the surrounding population than use of fossil fuel. However, each causes more adverse health impacts for the users and for the population than active travel as an alternative to car use. Autonomous vehicles may improve access for those with limited travel mode options, or those dependent on others to drive, bringing benefits of greater access and reduced isolation. Autonomous vehicles may also reduce the risks of injury while reducing congestion. Shared use rather than private ownership would reduce the number of vehicles and the amount of parking needed, releasing more space for more social uses. However, if shared use schemes are used primarily by those who would otherwise walk or cycle, the overall health impacts will probably be negative.

Summary and Conclusion

Transport has many benefits and harms for health. Benefits include access to education, employment, goods, services, and people as well as opportunities for physical activity and for being in green or blue spaces. Health conditions and illness can restrict travel mode options, reducing the health benefits of travel. Transport-related harms include air and noise pollution and greenhouse gases, sedentary lifestyles, injuries, and community severance. Active modes of transport, such as walking and cycling, contribute to increasing or maintaining physical activity, maintenance of healthy body weight, and better mental and physical health overall. Most public transport journeys involve some active travel. However, in car-dominated societies, adverse effects of transport are experienced mostly by more deprived communities whereas benefits generally accrue to wealthier car users.

Biographies

Professor Mindell is a Public Health Physician. She leads the Health Survey for England team and is the Health lead for UCL's Transport Institute. She conducts research into community severance (the barrier effect of busy roads and transport infrastructure) and also publishes on road fatality rates. She was lead editor of the 2011 report *Health on the Move2*. She is founding Editor-in-chief of the award-winning *Journal of Transport and Health* and is a member of the Executive of the Transport and Health Science Group (founding its Latin American network) and of the International Professional Association for Transport and Health.

Associate Professor Sandra Mandic is an Exercise Scientist. She leads interdisciplinary and multi-sector research projects in physical activity and health in New Zealand with the links to transportation, built environment and sustainability. Her academic training and professional experiences span Europe, Canada, the United States, and New Zealand. She is the academic leader of the Active Living Laboratory, the principal investigator on the Built Environment and Active Transport to School: BEATS Research Programme and the convener of the Transport Research Network at the University of Otago, New Zealand.

See Also

The Value of Life and Health; Bicycle Infrastructure; Weather, Effects of; Transport modes and Ageing Society; Transport Modes and Sustainability; Intermodal Passenger Transport; Active/Nonmotorized Transport Modes; Road Modes: Walking; Road Modes: Cycling; Street Design For Active Travel

Relevant Websites

Active Living Laboratory, University of Otago, New Zealand. Available from: www.otago.ac.nz/active-living.

Essential evidence: Short summaries of the evidence on transport and health for transport planners, by Professor Adrian Davis. Available from: <https://travelwest.info/essentialevidence>.

International Professional Association for Transport and Health / TPH Link evidence summaries. Available from: <https://www.tphlink.com/hot-topics-reference-materials.html>.

Journal of Transport and Health. Available from: <https://www.sciencedirect.com/journal/journal-of-transport-and-health>.

Transport and Health Science Group. Available from: www.transportandhealth.org.uk/.

The Lancet climate change and health infographics. Available from: <https://www.thelancet.com/infographics/climate-and-health>.

World Health Organization. Death on the roads. Available from: https://extranet.who.int/roadsafety/death-on-the-roads/#deaths/per_100k.

World Health Organization. Global ambient air pollution. Available from: <http://maps.who.int/airpollution/>.

References

- Appleyard, D., Lintell, M., 1972. The environmental quality of city streets: the residents' viewpoint. *J. Am. Inst. Plan.* 38, 84–101. 10.1080/01944367208977410.
- Asher, L., Aresu, M., Falaschetti, E., Mindell, J., 2012. Most older pedestrians are unable to cross the road in time: a cross-sectional study. *Age Ageing* 41 (5), 690–694. 10.1093/ageing/afs076.
- Cohen, J.M., Boniface, S., Watkins, S., 2014. Health implications of transport planning, development and operations. *J. Trans. Health* 1, 63–72. 10.1016/j.jth.2013.12.004.
- Feleke, R., Scholes, S., Wardlaw, M., Mindell, J.S., 2018. Comparative fatality risk for different travel modes by age, sex, and deprivation. *J. Trans. Health* 8, 307–320. 10.1016/j.jth.2017.08.007.
- Holt-Lunstad, J., Smith, T.B., Layton, J.B., 2010. Social relationships and mortality risk: a meta-analytic review. *PLOS Med.* 7, e1000316. 10.1371/journal.pmed.1000316.
- Mackett, R., 2019. Mental health and travel. Report on a survey. London: UCL Centre for Transport Studies.
- Mandic, S., Hopkins, D., García Bengoechea, E., et al., 2017. Adolescents' perceptions of cycling versus walking to school: understanding the New Zealand context. *J. Trans. Health* 4, 294–304. 10.1016/j.jth.2016.10.007.
- Martin, A., Panter, J., Suhrcke, M., et al., 2015. Impact of changes in mode of travel to work on changes in body mass index: evidence from the British Household Panel Survey. *J. Epidemiol. Comm. Health* 69, 753–761.
- Mindell, J.S., Anciaes, P.R., Dhanani, A., et al., 2017. Using triangulation to assess a suite of tools to measure community severance. *J. Trans. Geog.* 60, 19–129. 10.1016/j.jtrangeo.2017.02.013.
- Mindell, J., Karlsen, S., 2012. Community severance and health: what do we actually know? *J. Urban Health* 89, 323–346.
- Mindell, J.S., Watkins, S.J., Cohen, J.M. (Eds), 2011a. Health on the move 2. Policies for health-promoting transport. Stockport, UK: Transport and Health Study Group. Available from: http://www.transportandhealth.org.uk/?page_id=32.
- Mindell, J.S., Cohen, J.M., Watkins, S., Tyler, N., 2011b. Synergies between low carbon and healthy transport policies. *Proceedings of the Institution of Civil Engineers – Transport*, 164, 127–39.
- Mueller, N., Rojas-Rueda, D., Cole-Hunter, T., et al., 2015. Health impact assessment of active transportation: a systematic review. *Prev. Med.* 76, 103–114.
- Royal College of Physicians (RCP), 2016. Every breath we take: the lifelong impact of air pollution. London, RCP. Available from: <https://www.rcplondon.ac.uk/projects/outputs/every-breath-we-take-lifelong-impact-air-pollution>.
- Watts, N., Adger, W.N., Agnolucci, P., et al., 2015. Health and climate change: policy responses to protect public health. *The Lancet* 386, 1861–1914. 10.1016/S0140-6736(15)60854-6.

Further Reading

- Badland, H., Mavoa, S., Villanueva, K., et al., 2015. The development of policy-relevant transport indicators to monitor health behaviours and outcomes. *J. Trans. Health* 2, 103–110. 10.1016/j.jth.2014.07.005.
- Larouche, R., Saunders, T.J., Faulkner, G., Colley, R., Tremblay, M., 2014. Associations between active school transport and physical activity, body composition, and cardiovascular fitness: a systematic review of 68 studies. *J. Phys. Act. Health* 11, 206–227.
- Mandic, S., Jackson, A., Lieswyn, J., et al., 2019. Turning the tide - from cars to active transport. Dunedin, New Zealand: Active Living Laboratory, University of Otago. Available from: <https://www.otago.ac.nz/active-living/otago709602.html>.
- Panther, J.R., Jones, A.P., van Sluijs, E.M., 2008. Environmental determinants of active travel in youth: a review and framework for future research. *Int. J. Behav. Nutr. Phys. Act.* 5, 34.
- Sallis, J.F., Cervero, R.B., Ascher, W., et al., 2006. An ecological approach to creating active living communities. *Ann. Rev. Public Health* 27, 297–322.
- Widener, M.J., Hatzopoulou, M., 2016. Contextualizing research on transportation and health: a systems perspective. *J. Trans. Health* 3, 232–239. 10.1016/j.jth.2016.01.008.

Multimodality in Transportation

Sören Groth*, Tobias Kuhnimhof†, *ILS—Research Institute for Regional and Urban Development, Dortmund, Germany; †RWTH Aachen University, Institute for Urban and Transport Planning, Aachen, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction: Different Notions of Multimodality	118
Multimodality as a Property of the Transport System	118
Multimodality as a Facet of Individual Travel Behavior	119
Multimodality as a Transport Policy Strategy	119
Terminological Heterogeneity in Research on Multimodality	119
Contrasting Terminologies	119
Supplementary Terminologies	119
Delimiting Terminologies	120
Measurements of Multimodality	120
Methods of Capturing Transport Mode Use	120
Reference or Observation Periods	121
Types of Transport Modes	121
Variability of Mode Use	121
Reasons and Causes for Multimodal Behaviors	121
Mobility Resources	121
Settlement Structures	122
Sociodemographic Factors	123
Psychographic Characteristics	123
Smart Mobility Based on Interconnected Mobility Services: New Developments in Multimodality?	123
Conclusions	124
See Also	125
References	125
Further reading	126

Introduction: Different Notions of Multimodality

In recent years, the most prevalent use of the term multimodality in transportation referred to the (flexible) use of different transport modes. Within the automobile-oriented societies of the Global North, this notion of multimodality has been postulated as an ecologically sustainable alternative to the exclusive use of private cars since the 1990s at the latest. This is part of general efforts to reduce the use of the car, which is less resource-efficient than alternative modes of ground transportation, specifically when powered with fossil fuels. Multimodal behavior has been observed to go along with less car use and less vehicle-kilometer traveled (VKT) than monomodal car use in everyday travel (Nobis, 2007). Hence, much attention has been paid to multimodality, additionally fueled by the increase in multimodal behaviors among the millennial generation, that is, the generation of young adults born in the period from the early 1980s to the mid-1990s. Compared to previous generations of young adults, millennials exhibited a decline in car ownership and use, suggesting a shift away from the automobile toward more multimodal behavior patterns (Kuhnimhof et al., 2011; 2012) (detailed contribution about Next Generation Travel by Kuhnimhof and Levine, 10429).

While this behavioral dimension of multimodality has mostly been in the focus of transport research, the term multimodality can also be applied to the supply side of the transport system or to summarize a specific set of transport policy strategies.

Multimodality as a Property of the Transport System

On the side of transport supply, multimodality is a property of a transport system that provides travelers with the material option of using different transport modes. Put more generally, a multimodal transport system comprises a nonmonolithic composition of all structural components, that is, the entire set of external and internal mobility resources and the corresponding rules, that enable the movement of people or goods by different transport modes (see e.g., Spickermann et al., 2014). A conventional example in this context is a multimodal corridor which provides travelers with various options to travel from A to B. Key examples for likely increasing multimodality of current transport systems are represented by the dynamic developments in the field of smart mobility based on interconnected mobility services (car sharing, bike sharing, scooter sharing, etc.). These are based on the principle of usership instead of ownership, thereby substantially decreasing the fixed costs while increasing the out-of-pocket costs of using a

specific transport mode. This—in connection with the flexibility of modern ICT—facilitates situationally flexible mode usage and hence fosters multimodal behaviors (detailed contribution about ICT and Transport Modes by Cohen-Blankshtain, 10422).

Multimodality as a Facet of Individual Travel Behavior

On the side of travel demand or travel behavior, multimodality on the level of an individual traveler is defined as the use of more than one mode of transport over the course of time. The prevalence of multimodal behaviors within a society is also an indicator describing the multimodal quality of transport systems and hence the underlying transport policy strategies. By comparing different countries of the Global North, substantial differences can be identified in this regard. For example, representative studies show that multimodal behaviors tend to be less pronounced in North American countries due to strong car centrality (Buehler and Hamre, 2015) than in European countries such as the United Kingdom (Heinen and Chatterjee, 2015), Germany (Nobis, 2007), the Netherlands (Krygsman and Dijst, 2001), or Denmark (Olafsson et al., 2016). However, measurement methods used to quantify multimodal behaviors often vary between the studies, so they do not allow a direct comparison of the prevalence of multimodal behaviors.

Multimodality as a Transport Policy Strategy

Finally, multimodality may also be a transport policy strategy seeking to improve the multimodal qualities of a transport system, hence aiming at increasing multimodal travel behavior and thus achieving more sustainability in the transport sector (e.g., Chlond, 2012; Dacko and Spalteholz, 2014). Such strategies aim at promoting the integration of transport modes through different channels in order to achieve a better utilization of the multimodal supply. This is currently being pursued at various (political) levels. One example is the proclamation of the “Year of Multimodality” for 2018 at the level of the European Union in order to reduce local pollutant and global greenhouse gas emission problems within the EU states and thus to improve the quality of life of the people in accordance with the Paris Agreement. Likewise, states, metropolitan regions, cities, or companies meanwhile have their own strategies to promote multimodal behaviors within their institutions. Unlike other policies to achieve increased sustainability of the transport sector, such as policies to ban cars from city centers, multimodality carries the notion that each transport mode – including the car – fulfills an important function in the transport system. Hence, multimodality as a policy strategy often works as a common denominator across very different stakeholders including automobile advocates and can help building bridges in controversial mobility discourses (Block-Schachter, 2009).

The structure of the paper is as follows: Section [Terminological Heterogeneity in Research on Multimodality](#) presents the existing terminological variety in research on multimodality. Section [Measurements of Multimodality](#) describes different approaches to measuring multimodal behaviors. In Section [Reasons and Causes for Multimodal Behaviors](#), observed causes and reasons for multimodal behaviors are outlined. Then, Section [Smart Mobility based on Interconnected Mobility Services: New developments in Multimodality?](#) sketches emergences in interconnected mobility services, which are associated with increased multimodal behaviors. Finally, a critical conclusion is drawn on the current role of multimodality as an alternative to the private car.

Terminological Heterogeneity in Research on Multimodality

In addition to the different concepts of multimodality as discussed earlier there exists a complex variety of further contrasting, complementary and delimiting terms, which are epistemologically relevant for research or practically relevant for planning:

Contrasting Terminologies

Multimodality is contrasted mainly with monomodality: While multimodal behaviors are described by the (flexible) use of more than one transport mode, monomodal behaviors are characterized by the use of only one mode in a certain time period. These two contrasting concepts play a role within research on everyday mobility in at least two respects: First, in the context of the debate on a possible transition from an automobile to a multimodal society, according to which multimodal behavior is understood as a sustainable alternative to monomodal (“nonsustainable”) car use (e.g., Chlond, 2012; Nobis, 2007). Second, considering the restrictive effects of transport poverty, according to which socially marginalized persons (low income, formally low education, precarious employment, etc.) are often excluded from access to multimodal behaviors due to a lack of mobility resources, and thus practice more often monomodal behaviors (e.g., Blumenberg and Pierce, 2014).

Supplementary Terminologies

In most studies, multimodality is further differentiated in terms of the respective research interests. A prominent example is the use of the terms bimodality, trimodality, etc., to express the additional modes used (e.g., Klinger, 2017). In another example, researchers distinguish between the environmentally sustainable quality of combined transport modes, for example, with regard to “green multimodal behaviors” (i.e., the combination of active and public modes) on the one hand and “car-based multimodal behaviors” (i.e., the use of at least one other mode besides the private car) on the other (e.g., Hunecke et al., 2020; Mehdi-zadeh et al., 2019).

Another emphasis is on modality styles, according to which habitual combinations of transport modes are examined in combination with individual life situations and/or specific psychological values and attitudes. In this respect, [Vij et al. \(2013\)](#), for example, identify three segments of different Modality Styles using economic parameters from German data: (1) "Habitual drivers," who view the world from the steering wheel and imagine distances in terms of driving times and locations in terms of available parking spaces; (2) "Time sensitive multimodals," who switch between many different travel modes, but make a trade-off between the different modes and the estimated travel time; and (3) "Time insensitive multimodals," who flexibly choose their transport options based on other criteria than travel time. Likewise, the segments of [Diana and Mokhtarian \(2009\)](#), which include subjective evaluations of modes and desires to change modes in addition to objective mode use, can be cited. For a French data set they distinguish (1) "Very light travelers, monomodal car users, deprived re other modes," (2) "Light travelers, car dominated but multimodal, deprived re other modes," (3) "Moderate travelers, highly multimodal, mixed satisfaction," und (4) "Heavy travelers, highly multimodal, generally surfeited."

Delimiting Terminologies

Within this conceptual heterogeneity of multimodality, we would like to point at three additional distinct concepts, which are related to the concept of multimodality, but represent independent fields of research and practice:

1. Intermodality describes the interlinking of different transport modes on a single trip. The term originates from freight transport, according to which containers are moved "inter modes," that is, between the transport modes ([Donovan, 2000](#)). In everyday mobility, intermodality plays a role at certain intermodal transfer points, for example, mobility hubs, park-and-ride parking spaces or bike-and-ride spots in public transportation. At the level of regional transport planning, the expansion of such intermodal hubs is considered important, since the opportunities for intermodal travel outside urban areas are limited (e.g., [Oostendorp et al., 2019](#)). In some studies (especially from the Anglo-American context) the term intermodality is often used synonymously with the term multimodality. However, both from a scientific point of view and in planning practice, the inconsistent use of the terms can lead to confusing ambiguities.
2. Intramodality describes the synchronous movement of transport modes along a single trip. Intramodality is practiced, for example, where transport modes may be carried within transport modes. A progressive initiative to promote intramodal transportation is, for example, the possibility to carry a bicycle on public transportation.
3. Multioptionality, which is used primarily in German-language transport and mobility research, shifts the perspective from actual mode choice toward potential mode choice ([Groth, 2019](#)). The concept of multioptionality was introduced to acknowledge the fact that the opportunities for participating in multimodal behaviors are unequally distributed in societies (i.e., "not everyone can be a multimodal"). A distinction can be made between material and mental multioptionality; material multioptionality describes individual access to more than one mode option based on available mobility resources (e.g., driving license and car availability + available bicycle). Mental multioptionality describes individual receptiveness to the use of more than one transport resource (e.g., based on the subjective evaluation processes of individual transport resources).

Measurements of Multimodality

There is a broad variety of multimodal segmentations and part of this heterogeneity is based on different methodological approaches. Typically, the analysis of multimodal behaviors is based on quantitative methods although there are also a few qualitative approaches (e.g., [McLaren, 2016](#)). Among the quantitative approaches, at least four factors can be identified which influence segmentation and vary across studies on multimodality and hence lead to varying proportions of different multimodal groups within the population of travelers: (1) Length of reference or observation period for which transport mode use is captured, (2) methods of capturing transport mode use, (3) types of transport modes to be considered, and (4) intensities of mode variability within the observation period:

Methods of Capturing Transport Mode Use

The conventional method of capturing multimodal behaviors are travel diaries in which survey participants record individual trips during a defined observation period (e.g., 24 hours or 7 days) including relevant trip properties such as the transport mode used. Travel diary surveys are well established since decades and relatively comparable in format across different countries and over time. With the emergence of GPS-tracking as an element of travel surveys there is additional precision and granularity with regard to capturing mode use, for example, it has become easier to capture the use of multiple modes for a single trip ([Kuhnimhof et al. 2018](#)). An alternative to capturing mode use for single trips are survey items inquiring about the typical mode use of individuals. For example, travel surveys may include questions about how often people typically use specific modes with response categories such as "daily," "on one to three times per week" etc. ([von Behren et al. 2018](#)). The use of multiple modes and hence multimodality during a given period of time can hence also be established with such survey items capturing typical mode use. [Nobis \(2015\)](#) found that the findings on multimodality based on such typical mode use questions compare by and large to the results based on trip diary surveys.

Reference or Observation Periods

During any given single day, most travelers use one or few modes only, that is, many are monomodal during short periods of time. During an entire year, however, most travelers use a broad variety of modes and are multimodal. Hence, the longer the reference period for capturing multimodal behaviors, the larger the proportion of multimodal groups as opposed to monomodal groups. Against this background, different reference periods are chosen across the studies, depending on the respective research questions: A reference period of several weeks is suitable, for example, to identify deep-seated mode use habits, within the framework of the above-mentioned Modality Styles (e.g., [Vij et al., 2013](#)). However, a one-week observation period is chosen in most studies to record multimodal behaviors, since during this period cultural processes are cyclically repeated in western societies and the typical variability of mode use can be depicted on this basis (e.g., [Buehler and Hamre, 2015](#)). Also, observation periods of a few days are used. Particularly when project budgets do not allow for extensive travel diaries, just a few days can be observed and still provide helpful indications of different behavior patterns of different social groups (see e.g., [Hunecke et al., 2020](#)).

Types of Transport Modes

With regard to the mode types the following pattern applies: The more types of modes are considered for establishing multimodal behaviors, the larger the share of multimodal groups. In most cases, therefore, only the use of the four main transport modes is recorded, that is, (1) car (as driver), (2) public transportation, (3) bicycle, and (4) Walking. However, some studies emphasize the importance of including as many modes as possible (e.g., car use as passenger, motorcycle use, taxi use, car sharing) (e.g., [Diana and Mokhtarian, 2009](#)). Some others focus on multimodal behavior in the context of a reduced choice of transport modes; for example, [Diana \(2012\)](#), who conducts his multimodality study exclusively involving car and public transport. Sometimes specific transport modes are also grouped together, such as trains, trams and buses for public transportation or cycling and walking for active transportation (e.g., [Heinen and Chatterjee, 2015](#)). With regard to the transport modes to be included in the segmentation process, the need to consider walking is debatable. Each trip begins on feet (i.e., also accessing the car or the train). Therefore, it is not always possible to draw a clear line from which point onward walking should be included as an independent type of transport. Some studies therefore do not include walking.

Variability of Mode Use

Consider the case of two individuals A and B who both use the car and the bicycle during the course of one week. While individual A switches between car and bicycle for the work commute every other day, individual B exclusively uses the car from Monday to Saturday and undertakes a single cycling tour on Sunday. This example visualizes how multimodality can vary in intensity even though both individuals could be equally considered “bimodal.” The intensity of multimodal behaviors hence varies interindividually and can therefore be of importance for a specific research interest. As the example above illustrated, categorical approaches often prove to be insufficient. Some studies that carry out segmentation emphasize proportions of the use of transport modes in order not to overemphasize modes that are rarely used. The best-known distinction between a “multimodal core” and a “multimodal tendency” is provided by [Nobis \(2007\)](#): Only those travelers belong to the multimodal core for whom there is no dominant transport mode; that is, no mode with which they cover over 70% or their trips. With regard to the goal of precisely mapping mode variabilities, however, such categorical approaches are sometimes criticized as heuristic. Alternatively, the methodological application of various indices has been established, originating from welfare economics (e.g., Gini, Dalton and Atkinson indices) or from information theory and ecology (e.g., entropy, Herfindahl-Hirsch index). [Diana and Pirra \(2016\)](#) offer a comprehensive portrait of possible index approaches for measuring multimodality, which in turn have different strengths and weaknesses with regard to specific research questions.

Reasons and Causes for Multimodal Behaviors

The literature on multimodality identifies four classes of factors, which appear to explain individual multimodal behaviors: (1) Availability of mobility resources, (2) settlement structures, (3) sociodemographic, and (4) psychologic characteristics. Among these classes there are factors that facilitate and other that impede multimodal behaviors ([Table 1](#)):

Mobility Resources

Almost all studies on multimodality report similar patterns with regard to the impact of specific mobility resources on travel behavior. For example, studies almost unanimously agree that car-related mobility resources—that is driving license, car availability, car ownership, and an increasing number of cars in the household—bear a strongly negative relationship with multimodality. In contrast to car-related mobility resources, the availability of mobility resources for alternative transport modes to the car (e.g., public transport season tickets or bicycle availability) can be placed in a positive relationship with multimodality. This finding confirms the everyday experience that—in automobile oriented societies—the car is a universal mode of transport enabling a (monomodal) modality style while the alternatives (public and active transportation) need to be combined to satisfy the broad set everyday mobility needs.

Table 1 Beneficial factors and barriers for multimodal behaviors

<i>Determinants</i>	<i>References</i>	<i>Beneficial factors (+) / barriers (-) for multimodality</i>
Mobility resources		
<i>Car</i>	Block-Schachter 2009; Blumenberg & Pierce 2014; Buehler & Hamre 2015; Diana, 2012; Diana & Mokhtarian 2009; Heinen & Chatterjee 2015; Heinen & Mattioli 2017; Hunecke et al. 2020; Kuhnimhof et al. 2006; 2011; 2012; Molin et al., 2016; Nobis 2007; Scheiner et al. 2016; Vij et al. 2013	License holding (-); Car availability (-); Increasing number of cars per household (-)
<i>Public transportation</i>	Block-Schachter 2009; Buehler & Hamre 2015; Heinen & Chatterjee 2015; Hunecke et al., 2020; Kuhnimhof et al. 2006; Nobis 2007; Scheiner et al. 2016; Vij et al. 2013	Season ticket (+); Dense public transportation network (+); Good accessibility of public transportation (+); Short walking distance to rail station (+)
<i>Bicycle</i>	Vij et al. 2013; Heinen & Chatterjee 2015; Heinen & Mattioli 2017	Bike ownership / availability (+)
Settlement structure		
<i>Settlement types</i>	Buehler & Hamre 2015; Heinen & Chatterjee 2015; Heinen & Mattioli 2017	Rural (-); Small urban (-); large urban (+)
<i>Density</i>	Blumenberg & Pierce 2014; Buehler & Hamre 2015; Diana 2012; Heinen & Chatterjee 2015; Kuhnimhof et al., 2006; 2010; 2012; McLaren 2016; Molin et al. 2016	Population density (low -/high +)
<i>Land use patterns</i>	McLaren 2016; Scheiner et al. 2016	Multifunctionality / diversity (low -/ high +); Lack of parking spaces (+)
<i>Housing types</i>	Heinen & Chatterjee 2015; Heinen & Mattioli 2017; McLaren 2016	Single family house settlements (-); mixed urban environments (+)
Socio-demographic factors		
<i>Age</i>	Buehler & Hamre 2015; Groth 2019; Heinen & Chatterjee 2015; Heinen & Mattioli 2017; Kuhnimhof et al. 2006, 2010, 2011, 2012; Molin et al. 2016; Nobis 2007	Young adults (+); medium age group (-); seniors (+)
<i>Educational level</i>	Block-Schachter 2009; Blumenberg & Pierce 2014; Buehler & Hamre, 2015; Diana & Mokhtarian 2009; Groth, 2019; Hunecke et al., 2020; Kuhnimhof et al. 2012; Nobis 2007; Vij et al. 2013; McLaren 2016; Molin et al. 2016	educational attainment (low -/high +)
<i>Gender</i>	Block-Schachter 2009; Blumenberg & Pierce 2014; Buehler & Hamre 2015; Groth 2019; Heinen & Chatterjee 2015; Heinen & Mattioli 2017; Kuhnimhof et al. 2006, 2010, 2011, 2012; Molin et al. 2016; Nobis 2007; Vij et al. 2013	female (-/~/+); male (-/~/+)
<i>Household</i>	Blumenberg & Pierce 2014; Buehler & Hamre 2015; Heinen & Chatterjee 2015; Heinen & Mattioli 2017; McLaren 2016; Molin et al. 2016; Nobis 2007; Scheiner et al. 2016; Vij et al. 2013	Small / single (+); Families with children (-)
<i>Income</i>	Blumenberg & Pierce 2014; Buehler & Hamre 2015; Diana & Mokhtarian 2009; Groth, 2019; Heinen & Mattioli 2017; Hunecke et al. 2020; Kuhnimhof et al. 2012; McLaren 2016; Molin et al. 2016; Nobis 2007; Vij et al. 2013	Low income (-/+); high income (+/-)
<i>Occupation</i>	Blumenberg & Pierce 2014; Buehler & Hamre 2015; Diana & Mokhtarian 2009; Groth, 2019; ; Heinen & Chatterjee 2015; Heinen & Mattioli 2017; Hunecke et al. 2020; Kuhnimhof et al. 2006, 2010, 2012; McLaren 2016; Molin et al. 2016; Scheiner et al. 2016; Vij et al. 2013	Unemployment (-); Employment (-); University Studies (+); Retirement pension (+)
Psychologic factors		
<i>Transport mode related psychographic characteristics</i>	Diana 2012; Diana & Mokhtarian 2009; Hunecke et al., 2020; Molin et al. 2016	Positive attitudes towards car alternatives (+); Single-sides positive attitudes towards the car (-); Environmental awareness (~); Symbolic emotional rejection of the car (+); PT autonomy and enjoyment (+); Bicycle enjoyment (+); Desire for travel mode flexibility (+); Overall traditional 'materialistic' values (-); Progressive 'post-materialistic' values (+)

Settlement Structures

With regard to settlement structures as an explanatory variable, in particular density and land use have been identified as relevant:

First, multimodality studies usually establish connections between multimodality and high population density. On the one hand, there is an explanation that links to the availability of mobility resources as set out above. Public transport and mobility services tend to preferably serve high-density areas as these promise high utilization of the services. On the contrary, sparsely populated suburban and rural areas often do not offer comparable alternatives to the car. On the other hand, car ownership and use in high-density areas is often associated with higher cost (e.g., for parking) or more hassle (e.g., searching for parking) than in less densely populated areas. This factor also restricts car use and gives rise to multimodal behavior.

Second, multimodality studies discover a connection between multifunctional settlement structures and the flexible use of alternatives to the private car. However, since monofunctional and disperse settlement structures are a reality in many instances, the bicycle, for example, represents a typical component of a multimodal transport choice set. It is rarely suitable for reaching all destinations and is therefore combined with other transport modes.

It should be critically noted that the interactions between space and specific travel behavior are much more complex. While settlement-structures clearly correlate with travel behavior, it is unclear, to which degree settlement-structures determine travel behavior or are linked to travel behavior in some other way, for example, through residential self-selection (see contribution, e.g., by van Wee about transportation and space, 10404).

Sociodemographic Factors

Sociodemographic factors are most frequently used to provide explanations for multimodal behaviors. However, it turns out that the individual variables can rarely be considered independently. Two examples of the relationship between sociodemographic variables and multimodal behaviors can be highlighted (see others in the table): First, in the datasets of European countries, women are more often associated with multimodality, while men more often show monomodal car use. In North America it is the opposite. In both cultural contexts, the differences are explained by the researchers as a result of the reproduction of traditional role relationships, according to which women are still more often involved in traditional household commitments, which in turn entails heterogeneous interrelationships of travel. While in the American context the average number of cars per household is higher, European households often have fewer cars available, which in turn tend to be used more often by men (see e.g., [Buehler and Hamre 2015](#); [Heinen and Chatterjee 2015](#)). However, among the generation of young adults there is evidence of equalization of gender differences due to similar life situations (e.g., [Kuhnimhof et al. 2012](#)).

Second, income seems to play an important role with regard to the material preconditions for the realization of multimodal behavior. In this respect, it is assumed that with rising income, the mode options for implementing multimodal behaviors also increase ([Heinen and Mattioli 2017](#)). By contrast, low-income groups are associated with a lack of mobility resources, which makes multimodal behaviors more difficult (e.g., [Blumenberg and Pierce 2014](#)). However, social status has to be taken into account here as well: For example, socially marginalized groups (low income, formally low education in precarious employment, etc.) are particularly affected by transport poverty and have problems in implementing multimodal behaviors, whereas students—who may also have a low income—often benefit from having access to multimodal services (In Germany, e.g., students receive the semester ticket as a low-cost season ticket for public transport) (see also contribution by Mullen about transport modes and inequalities, 10604).

Psychographic Characteristics

The explanations of multimodal behaviors by means of the psychographic characteristics of the traveler are versatile and refer to different psychological dimensions. Three examples:

First, positive and negative attitudes toward the use of different transport modes can correspond to the equivalent combinations of transport modes. For example, [Molin et al. \(2016\)](#) can link the monomodal use of private cars to positive attitudes toward the car and negative attitudes toward the use of public and active transportation. However, they also show that positive attitudes toward the use of modes do not necessarily coincide with the use of transport. This is particularly true for socially marginalized groups (low income, precarious employment, etc.) whose use of transport is more strongly determined by socioeconomic restrictions.

Second, multimodal behaviors can be directly linked to symbolic-emotional evaluations of transport modes. For example, the debate on the transition from the automobile toward a multimodal society is directly linked to the assumption that an emotional departure from the private automobile is taking place within the generation of young adults. However, some initial studies suggest that this is predominantly a milieu specific phenomenon among cosmopolitan intellectual circles (e.g., [Hunecke et al., 2020](#)).

Third, (dis)satisfaction with the use of specific transport modes can be identified and used to make improvements in the structural moments of transport systems. [Diana \(2012\)](#), for example, surveys multimodal car and public transport travelers to get indications on how public transport could be improved to replace car trips by public transport trips.

Smart Mobility Based on Interconnected Mobility Services: New Developments in Multimodality?

The way in which multimodality could be examined in the future is changing with the possible transition toward smart mobility based on interconnected mobility services. Smart mobility is a complex socio-technical composition, whose contours are still emerging. [Figure 1](#) shows two key elements that contribute to a better understanding of how multimodal behaviors work in the context of Smart Mobility: The first crucial key element of smart mobility relates to the increasing options of replacing vehicle “ownership” by vehicle “usership.” In addition to the “traditional” public services (mainly rail, tram, and bus), which are related to services of general interest, there are now also services in business-to-consumer (B2C) or peer-to-peer (P2P) format. B2C services are rental services, in which “users” appear as renters of transport modes offered by the service provider (e.g., station-based and flexible car sharing, bike sharing, scooter sharing). In the case of P2P services, the service providers act as intermediaries in the background, whereby their service consists in matching the providers and demanders of a mobility resource (e.g., carpooling, private car sharing, noncommercial open-source approaches) (see more the detailed contribution about Shared Mobility by Shaheen and Cohen, 10420).



Figure 1 Multioptionality/potential multimodal behaviors based on smart mobility/interconnected mobility services (Similarly published in Groth, 2019).

The second key element of multimodal behaviors based on emerging smart mobility refers to the connectivity of all mobility services. The emergence of smart mobility based on interconnected mobility services today implies a decisive evolutionary step, rooted in the fusion of the virtual world and the physical-material world (see also the more detailed contribution about Mobility as a Service by Mladenovic, 10607). From the user's point of view, modern information and communication technologies allow the user to always flexibly and situationally choose the appropriate transport mode. The smartphone can be seen as the core of multimodal behavior based on smart mobility. For example, smartphone applications are used to obtain real-time information on arrival and departure times or to find available community vehicles. In addition, these apps enable users to buy tickets or reserve a vehicle. Finally, the smartphone guides the way to the station or available vehicle via GPS and serves as a key to open the vehicle lock.

Conclusions

Without doubt, multimodality is one of the major relevant issues in the field of modern everyday mobility and promoting multimodality is a wide-spread paradigm to increase environmental sustainability of transport. However, it should be kept in mind that more multimodality does not always increase environmental sustainability as clarified by the example of a monomodal cyclist who turns multimodal through conducting more trips by car. Hence, multimodality—even though it is often treated as such—is not a

value in itself independent from circumstances. Instead, environmental benefits can only be gained if multimodal behavior replaces behavior that previously was more automobile centered. But even under these circumstances, however, both academic research and transport policy practice cannot automatically conceptualize multimodality as a sustainable alternative to the private car. Three examples:

1. The often emphasized idealization of minor segments of society as drivers of transition processes (e.g. changed mobility in cities or more multimodality among young adults) runs the risk of overlooking the reproduction of dominant automobility. As a consequence, socially and spatially selective trends are being stylized into ubiquitous trends, without further forcing the shift away from the private automobile in terms of transport policy. Consequently, it is important in research to keep the big picture in mind and to examine in detail in which contexts automobility as a hegemonic transport system is actually being replaced and in which it is not.
2. As for the climate impact of passenger transport long distance travel is about as relevant as urban travel. This also extends the view to the boom of global air traffic that has been experienced in recent years. As is known, air travel—facilitating long distances and currently completely dependent on fossil fuels—is at least as worrisome as automobile travel as for its climate impact. However, by nature air travel is intermodal and airports act as intermodal interfaces on transregional travel, whereby passengers use traditional and/or new mobility services after their flights at the departure and destination airports to complete their intermodal route chain. New and traditional mobility service providers enter into new alliances with airlines. Hence, intermodality and multimodality may also be associated with increasing air travel, for example, if those urban dwellers without car more often opt for air travel in the case of long-distance travel. Hence, the concept and analysis of multimodality should not be limited to urban travel but include all relevant facets of travel in order to obtain a comprehensible picture (see, e.g., contribution on geographies of air transport by Lucy Budd and Stephen Ison, 10426).
3. Finally, a dichotomous focus on automobility versus multimodality at the urban-regional level may lead to the neglect of potentially sustainable monomodal transport concepts at the transport policy level. However, in a few countries, such as the Netherlands, bicycle cities have been successfully realized in the last decades, which represent an ecologically sustainable urban transport beyond automobility and multimodality (see detailed contribution about car-free cities by Khreis, 10707).

The methodological and analytical approaches presented in the previous sections help to keep a critical eye on the research on the concept of multimodality and to enrich the discussion on the concept with new observations. Furthermore, from the perspective of transport policy and planning practice, it should always be carefully examined to what extent multimodality can contribute to socially and ecologically sustainable everyday mobility on the ground.

See Also

ICT and Transport Modes; Next Generation Travel: Young Adults' Travel Patterns; Transport Modes and Inequalities; Mobility as a Service; Shared Mobility; Long Distance Travel; Car-Free Cities

References

- Block-Schachter, D., 2009. 'The myth of the single mode man: how the mobility pass better meets actual travel demand. Massachusetts Institute of Technology, Cambridge.
- Blumenberg, E., Pierce, G., 2014. Multimodal travel and the poor: Evidence from the 2009 National Household Travel Survey. *Transp. Lett.* 6 (1), 36–45.
- Buehler, R., Hamre, A., 2015. The multimodal majority?: Driving, walking, cycling, and public transportation use among American adults. *Transportation* 42 (6), 1081–1101.
- Chlond, B., 2012. Making people independent from the car – multimodality as a strategic concept to reduce CO₂-emissions. In: Zachariadis, T.I. (Ed.), *Cars and Carbon*, Dordrecht, Springer Netherlands, pp. 269–293.
- Dacko, S.G., Spalteholz, C., 2014. Upgrading the city: Enabling intermodal travel behaviour. *Tech. Forecast. Soc. Change* 89, 222–235.
- Diana, M., 2012. Measuring the satisfaction of multimodal travelers for local transit services in different urban contexts. *Transp. Res. Part A* 46 (1), 1–11.
- Diana, M., Mokhtarian, P.L., 2009. Desire to change one's multimodality and its relationship to the use of different transport means. *Transp. Res. Part F* 12 (2), 107–119.
- Diana, M., Pirra, M., 2016. A comparative assessment of synthetic indices to measure multimodality behaviours. *Transp. A: Transp. Sci.* 12 (9), 771–793.
- Donovan, A., 2000. Intermodal Transportation in historical perspective. *Transp. Law J.* 27 (3), 317ff.
- Groth, S., 2019. Multimodal divide: Reproduction of transport poverty in smart mobility trends. *Transp. Res. Part A* 125, 56–71.
- Heinen, E., Chatterjee, K., 2015. The same mode again?: An exploration of mode choice variability in Great Britain using the National Travel Survey. *Transp. Res. Part A* 78, 266–282.
- Heinen, E., Mattioli, G., 2017. Does a high level of multimodality mean less car use?: An exploration of multimodality trends in England. *Transportation* 88 (4), 236.
- Heinen, E., Mattioli, G., 2019. Multimodality and CO₂ emissions: A relationship moderated by distance. *Transp. Res. Part D* 75, 179–196.
- Hunecke, M., Groth, S., Wittowsky, D., 2020. Young social milieus and multimodality: Interrelations of travel behaviours and psychographic characteristics. *Mobilities* 85 (115), 1–19.
- Klinger, T., 2017. Moving from monomodality to multimodality?: Changes in mode choice of new residents. *Transp. Res. Part A* 221–237.
- Krygsman, S., Dijst, M., 2001. Multimodal trips in the Netherlands: Microlevel individual attributes and residential context. *Transp. Res. Rec.* 1753, 11–19.
- Kuhnimhof, T., Chlond, B., Ruhren, S., von der, 2006. Users of transport modes and multimodal travel behavior: steps toward understanding travelers options and choices. *Transp. Res. Rec.* 1985, 40–48.
- Kuhnimhof, T., Chlond, B., Huang, P.-C., 2010. Multimodal travel choices of bicyclists. *Transp. Res. Rec.* 2190, 19–27.
- Kuhnimhof, T., Buehler, R., Dargay, J., 2011. A new generation: Travel trends for young Germans and Britons. *Transp. Res. Rec.* 2230, 58–67.
- Kuhnimhof, T., Buehler, R., Wirtz, M., Kalinowska, D., 2012. Travel trends among young adults in Germany: Increasing multimodality and declining car use for men'. *J. Transp. Geogr.* 24, 443–450.
- Kuhnimhof, T., Bradley, M., Straub Anderson, R., 2018. Workshop synthesis: Making the transition to new methods for travel survey sampling and data retrieval. *Transp. Res. Proc.* 32, 301–308.

- McLaren, A.T., 2016. Families and transportation: Moving towards multimodality and altermobility? *J. Transp. Geogr.* 51, 218–225.
- Mehdizadeh, M., Zavareh, M.F., Nordfjaern, T., 2019. Mono- and multimodal green transport use on university trips during winter and summer: Hybrid choice models on the norm-activation theory. *Transp. Res. Part A* 130, 317–332.
- Molin, E., Mokhtarian, P.L., Kroesen, M., 2016. Multimodal travel groups and attitudes: A latent class cluster analysis of Dutch travelers. *Transp. Res. Part A* 83, 14–29.
- Nobis, C., 2007. Multimodality: Facets and causes of sustainable mobility behavior. *Transp. Res. Rec.* 2010, 35–44.
- Nobis, C., 2015. *Multimodale Vielfalt*. Humboldt-Universität zu Berlin, Berlin.
- Olafsson, A.S., Nielsen, T.S., Carstensen, T.A., 2016. Cycling in multimodal transport behaviours: Exploring modality styles in the Danish population. *J. Transp. Geogr.* 52, 123–130.
- Oostendorp, R., Krajewicz, D., Gebhardt, L., Heinrichs, D., 2019. Intermodal mobility in cities and its contribution to accessibility. *Appl. Mobil.* 11 (4), 1–17.
- Scheiner, J., Chatterjee, K., Heinen, E., 2016. Key events and multimodality: A life course approach. *Transp. Res. Part A* 91, 148–165.
- Spickermann, A., Grienitz, V., Gracht, H.A. von der, 2014. Heading towards a multimodal city of the future? *Tech. Forecast. Soc. Change* 89, 201–221.
- Vij, A., Carrel, A., Walker, J.L., 2013. Incorporating the influence of latent modal preferences on travel mode choice behavior. *Transp. Res. Part A* 54, 164–178.
- von Behren, S., Minster, C., Esch, J., Hunecke, M., Vortisch, P., Chlond, B., 2018. Assessing car dependence: Development of a comprehensive survey approach based on the concept of a travel skeleton. *Transp. Res. Proc.* 32, 607–616.

Further reading

- Clifton, K., Muhs, C., 2012. Capturing and representing multimodal trips in travel surveys. *Transp. Res. Rec.* 2285, 74–83.
- Goletz, M., Hausteil, S., Wolking, C., L'Hostis, A., 2020. Intermodality in European metropolises: The current state of the art, and the results of an expert survey covering Berlin, Copenhagen, Hamburg and Paris. *Transport Policy* 94, 109–122.
- Heinen, E., 2018. Are multimodals more likely to change their travel behaviour? A cross-sectional analysis to explore the theoretical link between multimodality and the intention to change mode choice.. *Transportation research part F: traffic psychology and behaviour* 56, 200–214.
- Kroesen, M., van Cranenburgh, S., 2016. Revealing transition patterns between mono-and multimodal travel patterns over time: A mover-stayer model.. *European Journal of Transport and Infrastructure Research* 16 (4).
- Krueger, R., Vij, A., Rashidi, T.H., 2018. Normative beliefs and modality styles: a latent class and latent variable model of travel behaviour. *Transportation* 45 (3), 789–825.
- Litman, T., 2012. *Introduction to Multi-Modal Transportation Planning: Principles and Practices*. Victoria Transport Policy Institute, Victoria.
- Mehdizadeh, M., Ermagun, A., 2012. "I'll never stop driving my child to school": on multimodal and monomodal car users. *Transportation* 47 (3), 1071–1102.
- Ralph, K.M., 2017. Multimodal millennials? The four traveler types of young people in the United States in 2009. *Journal of Planning Education and Research* 32 (2), 150–163.
- Vij, A., Walker, J.L., 2014. Preference endogeneity in discrete choice models. *Transportation Research Part B: Methodological* 64, 98–105.

Transport Modes and Disasters

Brian Wolshon, Department of Civil and Environmental, Louisiana State University, Baton Rouge, LA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	127
Background and Fundamentals	127
Evolution of Emergency and Resilient Transportation	128
Modes	130
Operational Strategies and Approaches	131
Acknowledgment	132
Biography	132
Further Reading	133

Introduction

Transportation and the roles that it plays during disasters are as varied as they are complex. A long history of experience has illustrated the vital functions they play in all phases of emergencies, before, during, and after they occur. In addition, to providing the modes and paths to evacuate people prior to the arrival of a hazard, transportation assets are also vital to prepare and preposition supplies before dangers arrive. During and immediately after a disaster, transportation modes and systems support numerous aspects of disaster response from search and rescue, to transporting the injured, to fighting fires and so forth. Then, after a disaster, transportation supports recovery, both immediate and long term, through the resupply of basic life-sustaining food, medicine, and building materials as well as serving as the conduits of re-entry for the return and re-entry of effected populations back into impacted areas.

Unfortunately, transportation can also become the source of disasters. Airliner crashes, ship sinkings, bridge collapses, and train derailments are a few of many examples of ways in which transportation modes and infrastructure have, themselves, become the source of injuries, loss of life, and damage to property and the environment.

In this entry of the *Encyclopedia of Transportation*, ways in which transportation has been used to prepare for, respond to, and recover from disasters as well as the ways in which it contributes to their occurrence of discussed. In particular, the discussion highlights how important features and characteristics and operational techniques of transportation can and have been be used to counter the detrimental effects of disasters. It also includes a brief summary of the evolution of transportation and its involvement and impact in events and hazards as well as the manner in which different transportation modes have and can be used in times of emergency. From an application standpoint the entry also includes a brief high-level discussion of various operational strategies and approaches to counter the effects of disasters and major events. Then, this entry summarizes recent emerging knowledge and examples of some disaster experiences to discuss the roles played by transportation in them and, conversely, how functionality was impacted.

Background and Fundamentals

While complex, the roles and involvement of transportation in disasters can also be described and analyzed in terms of the spatial and temporal characteristics of any specific hazard; most notably, how big of an area they impact, how long do their effects last, and how much warning time do they give to prepare, react, and respond. For example, some hazards, like hurricanes, impact thousands, if not tens of thousands, of square miles for hours or even days at a time. However, their arrival can be forecasted several days in advance. This gives time for protective action responses regional mass evacuations to be ordered and carried out to move hundreds of thousands of people over tens, if not hundreds, of miles. On the opposite end of this spatio-temporal spectrum are hazards like fires and explosions. While they may give little advanced warning, they typically impact comparatively smaller areas, as such the roles played transportation become limited to response and recovery after the event.

Table 1 includes a list of significant hazards that have, historically, impact transportation modes and systems. Conversely, the impacts of these same hazards have also been lessened though the various means of transportation to people and material goods. Because the contribution of transportation can be impacted by the size, scale, and timing of a hazard as well as its cause the table also includes the sources of each hazard in terms it occurring naturally or some type of intentional and/or unintentional human action. Perhaps most important from a transportation planning and/or action perspective is whether it was expected or unexpected and its levels of threat. The final column describes each of the listed hazards in terms of whether are planned or unplanned and whether or not it would be considered to be an emergency. Not surprisingly, when dealing with disasters, nearly all of these would be considered to be "unplanned emergency events."

Table 1 Hazard, types, sources effecting transportation

<i>Hazard</i>	<i>Cause</i>	<i>Type (transportation perspective)</i>
Earthquakes	Nature	Unplanned/emergency
Floods (rain, snow, hurricanes, surge)	Natural	Unplanned/sometimes emergency
Hurricane/typhoon	Natural	Unplanned/emergency
Ice storms	Natural	Unplanned/sometimes emergency
Landslides	Natural	Unplanned/sometimes emergency
Naturally occurring epidemics	Natural	Unplanned/sometimes emergency
Snowstorms and blizzards	Natural	Unplanned/sometimes emergency
Tornadoes	Natural	Unplanned/emergency
Volcanic eruption	Natural	Unplanned/sometimes emergency
Wildfire	Natural	Unplanned/emergency
Bomb threats and other threats of violence	Human, intentional	Unplanned/emergency
Fire/arson	Human, intentional	Unplanned/emergency
Riot/civil disorder	Human, intentional	Unplanned/emergency
Sabotage: External and internal actors	Human, intentional	Unplanned/emergency
Security breaches	Human, intentional	Unplanned/sometimes emergency
Terrorist (CBRN agents)	Human, intentional	Unplanned/emergency
Terrorist (conv. explosives, weapons)	Human, intentional	Unplanned/emergency
War	Human, intentional	Sometimes somewhat planned/sometimes unplanned/emergency
Workplace violence	Human, intentional	Unplanned/emergency
Accidental contamination/hazmat spills	Human, unintentional	Unplanned/emergency
Accidental damage to or destruction of physical assets (ex. refinery, chemical plant)	Human, unintentional	Unplanned/sometimes emergency
Crashes impacting transportation system	Human, unintentional	Unplanned/emergency
Power outage (widespread, extended time)	Human, unintentional	Unplanned/sometimes emergency

Source: Reproduced from the Institute of Transportation Engineers, *Traffic Engineer's Handbook*, 7th Edition, 2016.

Another important consideration in applying transportation resources, personnel, assets, and systems for disasters understands how likely and often events might occur. One of the most significant and under-addressed, issues of transportation in emergencies and disasters is dealing with the budgetary challenges of low probability-high consequence disaster events. In practical fact, it can be difficult to prioritize resources for conditions that may occur once in a generation, if at all. While the threat of injury and lost lives and impact to infrastructure may be enormous in many of disaster scenarios shown in [Table 1](#), traditional cost–benefit analyses cannot always justify the significant expenditures to counter them. Whether the approach would be to construct new systems, such as the one billion dollar Inner Harbor Navigation Canal Lake Borgne Surge Barrier to block a storm surge from entering Lake Pontchartrain near New Orleans, or reconstruct existing infrastructure to resist failure, nearly all such efforts are costly. Implementing such plans is also further hindered by budgets that are already unable to meet long lists of routine needs.

To counter the challenges of justifying investment to counter low probability-high risk events, newer approaches are being employed to plan, design, operate, and maintain transportation infrastructure and modal assets with a greater consideration of disasters, emergency management, and disruptions of all types. At the broadest level, these philosophies are classified as “resilient engineering.” Emergency management is the field with the responsibility to reduce vulnerability by preparing communities to mitigate, prepare for, respond to, and recover from hazards—from major and minor. By definition, the use of innovative, flexible, adaptable, and cost effective means to achieve overlapping preparedness, response, and rapid recovery needs is the fundamental principles of resilience. Thus, many of the tools and techniques used in emergency management, including hazard/risk/vulnerability assessments, readiness drills and exercises, and interagency coordination, are now being adapted into resilience-oriented transportation.

Evolution of Emergency and Resilient Transportation

In the United States (US), “transportation,” including planning and operation for disaster and emergency recovery and restoration, is designated by the Department of Homeland Security as the first “Emergency Support Function” (ESF) in the National Response Framework (NRF). The NRF, created soon after the terrorist attacks of 2001, guides the US response in disasters and emergencies. ESF are categories used to group emergency preparedness and response capabilities based on their ability to provide support, resources, personnel, etc. to during times of need. Like most current approaches to emergency management, the NRF structure takes a “scalable, flexible, and adaptable perspective” approach to disaster preparedness. These have evolved to be the core principles of resilience-oriented transportation.

In terms of all the disruptive events that can impact or benefit from the involvement of transportation, disasters are the most extreme. In practical terms, transportation disruptions can range from a minor as a stalled vehicle blocking a travel lane, to as a large

as a multistate mass evacuation. Thus, disaster preparedness and responses that seeks to achieve certain fundamental goals, like maintaining full functionality, but that are scalable, flexible, and adaptable to any planned, unplanned, or hazardous condition provide the most cost effective returns on investment and are, thus, more sustainable and achievable over long time horizons. Thus, when discussing transportation in disasters it is also important to first fundamentally understand the nature of the disruption that effects a transportation system, including not only how large of an area they impact and how long their effects last, but also how hazardous they may be to people and property and how often they may occur.

Although the focus of this entry is oriented toward major, large-scale disruptions and disasters, the current perspectives on transportation planning and engineering for such conditions are focused on broader all-disruption conditions. Reviews of practice suggest that the focus of most of the roles for and involvement of transportation disasters is most commonly associated with extreme weather, including most notably, flooding, hurricanes, and winter storms. Often, these events are also framed within the context of the long-term trends of global climate change and sea-level rise. Not surprisingly, the transportation involvement in disasters is also a based on the threats, hazards, and disaster that are faced in an area that is earthquakes in California, volcanoes in Hawaii, and tsunamis in southern Asia. Clearly, security and terrorist threats are also areas of emphasis in transportation; most obviously in air transport, but also in term of specific location and event venues.

Historically in the US, emergency management began as civil defense to protect citizens from military attack. Transportation, particularly the network of federally funded highway was envisioned to move military equipment and personnel. Later, during the Cold War, many of these planned to be used as evacuation routes to support the movement of people out of urban areas. Later, as populations grew in southern coastal areas of the country these same routes were looked upon to serve in a similar capacity, but for hurricanes rather than invasions.

Following the September 11, 2001 terrorist attacks, the newly created Department of Homeland Security developed the National Incident Management System (NIMS) that, with other organizational goals, sought to coordinate the involvement the capabilities of diverse agencies, like transportation. Within the broad frameworks of NIMS and the national Incident Command System (ICS) specific responses can be coordinated and adapted to counter the conditions with any hazard, large or small.

Focusing on transportation specifically, departments of transportation is evolving toward resilience approaches to support response and deal with emergencies and disasters. As part of these resilience philosophies, even broader and more comprehensive “all disruption” approaches are now being used. Functional disruptions expand the traditional hazard-oriented perspectives to also include any event, planned or unplanned, intentional or accidental, natural or man-caused, and catastrophic to routinely occurring. The use of the term functional disruptions is far more than semantics, however. Like the use of all-hazard emergency approaches that promote flexibly apply planning, response, and recovery to adapt and apply a range of options from a broad framework, resilience-oriented management also relies on flexibility and scalability to address an even wider range of conditions.

Table 2 includes a small cross-section of disruptive events to show how the characteristics of each are likely to impact a community and its transportation systems. In keeping with the philosophy of resilience, the table spans spatially-large and -small events, that occur with and without advanced notice, and that occur often and infrequently. In addition, to revealing the varied

Table 2 Event characteristic and relationships to transportation systems

<i>Event</i>	<i>Planning</i>	<i>Advanced notice</i>	<i>Duration</i>	<i>Hazard</i>	<i>Impact area</i>	<i>Frequency</i>
Vehicle Crash	Unplan/some emergency	None	Minutes to Hours	Low	Local to several miles	Frequent
Concert/ Sporting Event	Planned	Months/Years	1+ Days	None	Several Miles	Seasonally Frequent
Playoff Sporting Event	Partially Unplanned	Days	Hours to Days	None	Several Miles	Occasional
Major One-time Events	Planned	Years	1+ Days/Weeks	None	Several Miles	Infrequent
Parades	Planned	Months/Years	Hours	Low (Planned Road Closure)	Few Miles	Occasional
High Security Events	Planned	Days to Weeks	Hours to Days	Low	Several Miles	Occasional
Snow/Ice Storm	Unplanned	Hours to Days	Hours to Days	Medium	Regional	Seasonal, varies by region
Wildfire	Unplan/Emerg.	Minutes to Days	Hours to Weeks	Medium to High	Regional	Seasonal, varies by region
Flooding	Unplan/some emergency	Hours to Days (Usually)	Hours to Weeks to Months	Varies - Low to High	Local to Regional to Multi-State	Frequently, varies widely
Hurricane Evacuation	Unplan/Emerg.	Days to Weeks	Days	High	Regional	Seasonal Infrequent
Hazmat Spill	Unplan/Emerg.	None	Hours	High	Several Miles	Infrequent
Bridge Collapse	Unplan/Emerg.	None	Months	High	Several Miles	Infrequent

Source: Reproduced from the Institute of Transportation Engineers, *Traffic Engineer's Handbook*, 7th Edition, 2016.

nature of these events, the table also suggests commonalities of the event characteristics and opportunities of where transportation planning and operations for one of them may be spatially and/or temporally adapted for us in others.

The benefits of resilient planning goes beyond the ability to more effectively respond to and recover from disruptions. These approaches also use more innovative approaches that while effective are less costly and more importantly can also achieve multiple overlapping benefits that span a larger range potential disruptions and achieve it for a reduced costs. This also gives them a more justifiable as business-need under practical budgetary limitations. A recent review of state-level department of transportation (DOTD) in the US was conducted to determine the what types of disasters, emergencies and other disruptive conditions were focused-on in when integrating resilience into transportation planning, preparedness, design, operations, and maintenance.

Among the findings were that agencies tended to categorize resilience operations in terms of: (1) transport mode (highway, air, rail, maritime, etc.) and their associated roles, needs, and characteristics during periods of disruption and/or by (2) the type of disruption (natural hazard, incident, planned event, etc.) that had occurred or could be anticipated for a location or system. In terms of performance outcomes and metrics to assess the effectiveness of agencies tended to focus on:

- Safety, in various forms that included crash rate and frequency, fatalities, injuries, damage to property, etc.;
- Economics, including the reliability of funding to support resilience;
- Technology, including technologically-oriented data collection surveillance, and communications systems as well as “cyber” issues and data, more broadly; and
- People, including agency employees/staff, system users/travelers, and vulnerable populations.

As societies become more complex and activities become increasingly complex and, in some cases, more potentially hazardous and dangerous, newer and perhaps even more threatening hazards emerge. Natech disasters are events that involve a cascading series of disasters that typically start with natural disaster, like an earthquake or flood, that lead to the occurrence of technological disasters, like chemical spills explosions etc. One of the most significant was the tsunami-nuclear plant natech disaster in Fukushima, Japan. In that incident and earthquake trigger tsunami knocked out control systems on the nuclear power plant resulting in explosions, fires and radiological releases. From a transportation perspective, the tsunami flooded roads, damaged infrastructure, and blocked access with debris slowing the ability to conduct search and rescue operations and move people out of the area. Then, the ensuing contamination made parts of the affected area too hazardous to enter slowing the long-term rebuilding and recovery process.

Increasingly sophisticated transportation modes and systems can also be susceptible to failure. While they are designed with back-up systems and redundant safety features, there have been numerous examples of highway, air, maritime, and train incidents resulting in crashes, spills, explosions, and sinkings which have caused injury, loss of life, and damage to property and the environment. Autonomous control, requiring various levels of human input have also been linked to some of these incidents. Unfortunately, as more and more vehicles become increasingly machine controlled and transportation infrastructure and control systems rely on sophisticated means of communication and operation, the potential for intentional and unintentional disasters exists.

Some other areas worthy of mention that are not “traditional” disasters as they are commonly regarded, but nevertheless involve or impact transportation are have come to attention recently. One of these is the displacement of people and, in some cases, mass movements of entire populations due to war, famine, economics and, increasingly, climate-related causes. Such movement have often taken place on foot or by other vulnerable and illegal means of travel and involved crossing (and prohibition) of national borders. While this remains a complex problem and one in which transportation plays a secondary role it is an important issue and one in which transportation can play a role. Yet another deals with discouraging, restricting, and prohibiting travel and mobility. As seen in the worldwide COVID-19 pandemic of 2020, the implantation of travel restrictions and closure of activity generators significantly decreased all form of travel. While this has resulted in both positive (lowered rates of virus transmission, decreased congestion, delay, crashes, fuel consumption, and emissions) it has also resulted in significant impacts to commercial exchange, employment, and, surprisingly, increases in vehicle fatal crash rates.

Modes

All modes of transportation have roles to play in disasters. However, their effectiveness and impact as either a means of preparedness, recover, and response to a disaster or significant incident or as a contributing source of them has varied. From a wide use perspective, highway transportation is the most commonly used mode for disaster preparedness (evacuation) and recovery (resupply, restoration, and return/repopulation). However, airborne modes are used extensively for similar purposes, particularly when the location is remote, distant, difficult to access, and requires rapid arrival of critical supplies. Airborne modes may also become the sole means of transport into and out of inaccessible areas with damaged highway and rail routes such as may occur during floods and earthquakes. Even during the response period of a disaster, assets from drones to helicopters to heavy aircraft have been used for damage assessment, search and rescue, and wildfire suppression.

Similar to airborne modes, the use of waterborne transport modes have also been used extensively, especially in remote and difficult to access locations like islands. A distinct advantage of maritime modes is the enormous capacity to move cargo making them well-suited to recovery supply roles. Unfortunately, scale comes with the limitations on rapid movement. Between loading, travel, and off-loading, shipments by seas can often take days, if not weeks to arrive. It is also important to note that not every location has port facilities to accommodate every waterborne vessel. In some cases, a combination of water vessels is required and even water-to-air-to-land intermodal transfers. Water-based transportation modes have also varied widely in scale. In addition, to

sea going cargo vessels, military hospital ships have been used to supplement land-based medical care, during the September 11th terrorist attacks a spontaneously organized flotilla of ferry's was used to move evacuees area from the World Trade Towers, and small multiperson airboats and flat bottomed fishing boats are often used for search and rescue in flood-impacted urban areas impacted areas with floating debris.

Rail transport is another key mode in disaster preparedness, response, and recovery. While not used as extensively as other modes for immediate shipping of relief and recovery supplies, rail transport has been an integral part of preparedness, on both heavy and light passenger rail lines. While it has evolved over the years since Hurricane Katrina in 2005, Amtrak passenger rail has been a part of the New Orleans City Assisted Evacuation Plan (CAEP). In fact, the CAEP featured a multimodal transit approach to evacuation; utilizing regional transit busses to circulate in the city picking up evacuees and transferring them to the Amtrak passenger rail hub for transport out of the city.

Rail and bus transit service is planned for use in hurricane-threatened major US urban areas like Houston and New Orleans, particularly for transportation-limited populations like the carless, economically-disadvantaged, and persons with functional needs. Modes such as paratransit, taxis, kneeling and wheelchair-assist busses, and even mobile phone app ride-sharing services are part of mobility assistance plans. Then, ambulances, air medivac craft, and other forms of medically assisted transport have been used and are plan to transport hospital patients, the infirm, and frail elderly person before and after emergencies.

Operational Strategies and Approaches

Historically, transportation systems were largely regarded as passive actors in disaster preparedness, response, and recovery. Transportation agencies often supported first responders and emergency management; however, prior to the late 1990s and early 2000s, these could be described as peripheral and reactionary, rather than proactive. Despite the role served by transportation in the variety of conditions discussed throughout this entry, prior to the formally designated roles within the NRF, most transportation agencies felt their focus should be solely on their traditional responsibilities. To most agencies this meant maintaining transportation systems and assets in as close to a full function as possible at all times, including during disasters and emergencies.

This is not to suggest, however, that transportation had no role in prior disasters and emergencies. As primarily public agencies, departments often supported first responders and emergency management agency's many aspects of preparedness, response, and recovery. When adequate warning time was available transportation agencies commonly supplied personnel, equipment, and materials like barricades for road closures and permitted modal assets to be divert and schedules modified for emergency support. Even private road and bridge authorities would lift toll charges to permit freely flowing traffic. Then, after emergencies, transportation agencies was active in clearing debris, repairing damaged infrastructure like bridges, embankments, pavements, culverts, signal control systems, etc. that impacted the operation of their system. However, these were not viewed upon as core roles of transportation agencies, but rather supporting requests from other agencies who's primarily responsibility was emergency management and coordination.

In the US, the vital roles that could be played by transportation in emergencies became more evident in wake of events like the September 11th terrorist attacks, California wildfires, and catastrophic hurricanes of 2005. Required involvement from transportation agencies and personnel were further formalized by the NRF. These groups are valuable during times of emergency because, among other capabilities, they bring expert knowledge of transportation systems and assets under their jurisdiction. They also tend to be one of the few public sector agencies with field personnel and heavy equipment that can be mobilized on short notice and expedited contracting capability to rapidly bring in additional assistance. They also have expertise and experience in the application of remote sensing and analysis tools like global positioning systems (GPS), drones, geographic information systems (GIS).

Ultimately, the ability to apply the capabilities of transportation to support emergency conditions depends on the characteristics of any particular disaster and disruption and, more specifically, their spatial and temporal characteristics. For example, the application of regional freeway contraflow to support regional mass evacuations needs a lead time to assemble appropriate personnel; position barricades and implement controls; inform travelers. Thus, it tends to be most appropriate for geographically large slow-moving hazards like hurricanes that threaten vast areas, but can also be forecast several days in advanced. Threats like terrorist attacks, explosions, etc. which give little-to-no-notice typically only permit after-the-fact responses. This is why it is critically necessary for transportation agencies to work hand-in-hand with emergency planners to assess potential hazards, threats, and vulnerabilities and create proactive plans and maintain routine coordination, collaboration, and communication, not only after an emergency occurs.

Examples of expertise, activities, methods, and equipment that transportation agencies have brought to past disasters and could bring to those in the future, span several different areas of activity. In general, transportation associated with emergencies is oriented toward maximizing and maintaining safe, consistent, and rapid mobility throughout an event. Traffic planners and engineers also work with emergency managers to suggest strategies to help move traffic in different types of scenarios, situations, and identify resources required. Transportation professionals also coordinate and manage modes; monitor control systems (like traffic signals); and can use tools and technologies to gather data and model scenarios and response strategies to inform emergency managers on what conditions would likely happen under various management strategies. These generally fall into actions that focus on:

- operational methods to maintain, if not increase, the speed, amount, and efficiency of movement, including:
 - coordinating signals and adaptive traffic control using real-time data
 - implementing emergency control strategies (e.g., intersections turn restrictions)

- bus-only routes
- lane closures
- freeway ramp closures
- contraflow
- manual intersection control
- protected pedestrian traffic
- demand management methods to seek to lower surges in demand, often by spreading it over time and/or space to most effectively utilize available network capacity, including:
 - coordinate in-bound response vehicles.
 - provide transportation to mobility limited individuals
 - transportation demand management actions
 - emergency high-occupancy vehicle requirements
 - phased evacuations
 - shelter-in-place options
 - embargo certain vehicles/modes
 - pedestrian and bicycle strategies
- incident response and emergency communications to monitor and communicate information to travelers and emergency decision-makers, including:
 - monitoring roadways and signals
 - applying closed-circuit TV and variable message signing
 - utilizing highway advisory radio
 - aviation/airspace management and control
 - roadway incident management and motorist assistance patrols
 - fuel and personal service facilities
 - Maintenance/construction lanes cleared
 - Tow trucks deployment
- The application of personnel, equipment, and technologies to maintain, and/or restore infrastructure that support the movement of people, vehicles, and material, including damage and impact assessments.

Acknowledgment

The author gratefully acknowledge the United States Department of Transportation (US DOT) for their continued support of the Gulf Coast Center for Evacuation and Transportation Resiliency at Louisiana State University (LSU), a collaborative University Transportation Center (UTC) and a member of the Maritime Transportation Research and Education Center (MarTREC) at the University of Arkansas and the Center for Cooperative Mobility for Competitive Mega regions at the University of Texas. The content of this entry is solely the responsibility of the author and does not necessarily reflect the views of any of these sponsors.

Biography



Brian Wolshon has over 30 years of experience as a practitioner, researcher, and educator in traffic and transportation engineering, particularly related to emergencies, evacuations, and resilience. He is currently appointed as the Edward A. and Karen Wax Schmitt Distinguished Professor of Civil Engineering at Louisiana State University and has served as the founding Director of the *Gulf Coast Research Center for Evacuation and Transportation Resiliency* for more than a decade. In 2001, Dr. Wolshon founded the Transportation Research Committee on Emergency Evacuation and, for 17 years, served as its Chair. He served in visiting professor and researcher positions at Beihang University, China; the University of Hawaii; the University of New South Wales, Australia; and both the Sandia and Los Alamos National Laboratories. He has authored dozens federal reports, book chapters, and more than 80 scientific journal papers, including numerous seminal works related to evacuation planning and engineering. He has also served as an expert consultant to dozens of federal, state, and local government agencies and engineering firms throughout the United States and as appeared in more the 200 media interviews for outlets like *The Discovery Channel*, CNN, CNBC, MSNBC, Fox News, NPR, *The New York Times*, *USA Today*, and *The Times of London*.

Further Reading

- Amdal, J., Ankner, W., Callahan, T., Carnegie, J., MacLachlan, J., Matherly, D., Mobley, J., Peterson, E., Renne, J., Schwab, J., Venner, M., Veraat, N., Whytlaw, R., Wolshon, B., 2018. Improving the Resilience of Transit Systems Threatened by Natural Disasters, Volume 3: Literature Review and Case Study Summaries, Transit Cooperative Highway Research Program, TCRP Web-Only Document 70, Transportation Research Board, National Research Council, Washington DC.
- Department of Homeland Security, 2013. "National Response Framework-Version 2,".
- Federal Highway Administration (FHWA), 2006. "Simplified Guide to the Incident Command System (ICS) for Transportation Professional," Available from: http://ops.fhwa.dot.gov/publications/ics_guide.
- Lindell, M., Murray-Tuite, P., Wolshon, B., Baker, E.J., 2018. In: Large-Scale Evacuation: The Analysis, Modeling, and Management of Emergency Relocation from Hazardous Areas. CRC Press, Taylor and Francis Publishing Group, LLC, ISBN: 978-1-482-25985-8, p. 300.
- Matherly, D., Mobley, J., Wolshon, B., Renne, J., Thomas, R., Nichols, E. Elisa, 2013. A Transportation Guide for All-Hazards Emergency Evacuation, Strategic Highway Research Program, Report 740, ISBN: 978-0-309-25901-9, Transportation Research Board, National Research Council, Washington DC, 193.
- Matherly, D., Mobley, J., et al., 2011. TCRP Report 150: Communication with Vulnerable Populations: A Transportation and Emergency Management Toolkit. Transportation Research Board of the National Academies, Washington, DC.
- Matherly, D., Langdon, N., Kuriger, A., Sahu, I., Wolshon, B., Renne, J., Thomas, R., Murray-Tuite, P., Dixit, V., 2014. A Guide to Regional Transportation Planning for Disasters, Emergencies, and Significant Events, National Cooperative Highway Research Program, Report 777, ISBN: 978-0-309-07023-2, Transportation Research Board, National Research Council, Washington DC, 148 pp.
- Murray-Tuite, P., Wolshon, B., 2013. Assumptions and processes for the development of no-notice evacuation scenarios for transportation simulations. *Int. J. Mass Emerg. Disasters* 31 (1), 78–97.
- Renne, J., Pande, A., Wolshon, B., Murray-Tuite, P., Kim, K., 2020. *Creating Resilient a Transportation System: Policy, Planning and Implementation*. Elsevier Press.
- United States Department of Homeland Security, 2012. "National Response Framework," Available from: www.fema.gov/national-response-framework.
- Wolshon, B., Matherly, D., Murray-Tuite, P., 2016. Traffic Management During Planned and Unplanned Emergency Events Chapter 16. In: *Traffic Engineering Handbook – Seventh Edition*, John Wiley and Sons, Inc., New York, ISBN: 978-1-118-76230-1.
- Wolshon, B., Lambert, L., 2004. Convertible Lanes and Roadways, National Cooperative Highway Research Program, Synthesis 340, ISBN: 978-0-309-07023-2, Transportation Research Board, National Research Council, Washington DC, 92 pp.
- Wolshon, B., 2009. Transportation's Role in Emergency Evacuation and Reentry, National Cooperative Highway Research Program, Synthesis 392, ISBN: 978-0-309-09831-1, Transportation Research Board, National Research Council, Washington DC, 142 pp.

Big Data for Public Transport Planning

Jan-Dirk Schmöcker, Department of Urban Management, Kyoto University, Japan

© 2021 Elsevier Ltd. All rights reserved.

Introduction and Overview	134
Characteristics of Public Transport-Related Big Data	134
Overview of Data Generated by Journey Stage	134
Big Data for Long-Distance Versus Local Public Transport Network Planning	135
Data Types and Analysis Possibilities by Journey Stage	136
Pre-Trip Planning	136
Pre- and Post-Trip Activities	137
At Stops	137
On-Board	137
Boarding and Alighting Points	137
Big Data Generated by Non-PT Sources	138
Conclusions and Further Reading	138
References	139

Introduction and Overview

Characteristics of Public Transport-Related Big Data

The role of “big data” and the changes they start bringing to our transport systems are manifold and significant. Besides offline and real-time planning applications for existing transport services, new forms of transport such as ridesourcing and ridesharing would not exist as they rely on processing a large amount of real-time demand data and utilizing these for providing data on vehicle locations and available rides. The business model of ridesourcing services depends on such big data in that pricing is dynamic and based on demand predictions. Big data used in particular for public transport planning, however, go far beyond directly obtained demand and supply data as this chapter will discuss.

To note at the beginning of this discussion is that there is no unique definition of what makes data to be termed “big.” [Milne and Watling \(2019\)](#) provide a comprehensive discussion on big data for transport planning in general and discuss various definitions including their size, dimensionality, relation to other data, and general complexity in analyzing them. Clearly “big” only partially refers to data volume. Big data are usually nonpurpose driven data collected originally for other purposes and then explored for additional value regarding, for example, transport planning. Therefore, the majority of the data will be meaningless and do not contribute to the secondary question explored but one has to filter out the relevant information. This means that machine learning tools are usually employed together with big data in order to facilitate the exploration. One of the early, widely defined definitions is by [Laney \(2001\)](#) who refer to the 3V characteristics of big data, namely “Volume, Velocity, Variety”. (See also article 10601 on big data for various travel modes.)

Public transport planning relevant big data are generated by various sources and at various stages during a journey as shown in [Fig. 1](#). The clear distinction of the different stages and the multiple, often disintegrated, data sources relating to these stages is a distinguishing point between public transport and car traffic related big data. Another distinguishing point is that public transport inevitably generates big data related to ticket purchase. In the case of car traffic, some roads are priced but area-wide pricing of the whole journey appears to be not yet in place (see also article 10759 on data sources for transport modes and 10047 on transit management).

In the following, the data related to these stages and their importance for planning is discussed. The discussion is first held generally before some differences in importance and availability of data for different modes are discussed. Thereafter the different data generation methods/technologies for these stages are discussed including analysis examples. The overview is then complemented with a discussion on the use of big data generated by other, external systems for service planning before conclusions and further reading are pointed out.

Overview of Data Generated by Journey Stage

[Fig. 1](#) illustrates the general types of data that one might obtain with respect to public transport journeys. A journey will start at an origin that is usually different from the first boarding point. Understanding not only the first and last stops of a journey but travel patterns from the “true” origins and destinations, that is, the places where activities take place, is of importance for public transport planning for a range of reasons. First, it will relate to stop choice of customers so that impact of service changes might be mitigated by passengers adjusting their public transport journey by accessing different stops as their first boarding point. Understanding places

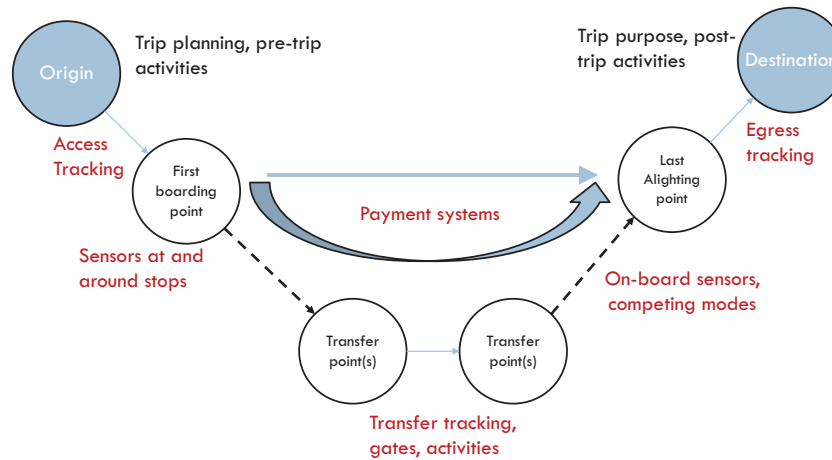


Figure 1 Journey stages and data sources

and types of activities is furthermore important to be able to understand and predict demand fluctuations with respect to external factors such as impacts of weather, opening of shops or other land-use changes (see also articles 10021 and 10367 on transit demand estimation and forecasting). The transport planner might further want to obtain information as to the pre-trip information accessed by travelers. Enquires about available connections could give an idea about latent demand as well as about connections not traveled but considered by (potential) customers. If both pre-trip information and actual journeys made can be observed it will provide information as to the flexibility and sensitivity of travelers in terms of departure time as well as actual routes taken with respect to variations in service attributes (see also article 10020 on value of different journey stages).

Understanding the access and egress modes as well as travel times helps establishing measures of service quality. Clearly longer access times from homes and places of activities will generally lead to more demand sensitivity with respect to competing modes. Observing access and egress modes will also help planners understanding requested facilities at stations such as required taxi stands or bicycle parking places. Finally, it will help planners to understand demand fluctuations due to for example congestion on other modes.

The next stage of a public transport journey is the time spent at a stop, station, port or airport, which in this chapter is generally referred to as stop. The time varies greatly by mode but also by other factors. One goal of (big) data being collected and analyzed is to improve services and provide users information so as to minimize the time spent at stops. This can be achieved either by service improvement such as co-ordination of timetables or by providing information which help users to plan their trip. Installing sensors at stops will hence help observing time spent at stops, as well as factors such as crowding, and queues that determine service quality. Observing behavior at stops can, furthermore, give indications as to route choice strategies of passengers.

Similar information can be collected also at transfer stops and the final alighting stop as indicated in the figure. Observing information about (likely) behavior at transfer points is often easier to obtain than for first stops as the time spent at a transfer stop might be also obtained from on-board sensors. If traveler behavior is inferred from on-board sensors, then, a difficulty is, however, distinguishing pure transfer-related activities such as waiting, walking and time required to purchase and validate tickets for the onward journey from other activities that one might do either to bridge the waiting gap or that were planned and according to which the traveler was willing to spend more time at the stop. This is relevant for both short waits at bus stops as well as long waits at stations or airports. One might choose willingly to wait longer if there are a range of attractions at/near the transfer point.

Vehicle and on-board sensors in their various forms can give the planner further a wide range of information. Vehicle tracking provides travel time and service reliability related information, sensors inside the vehicle can observe crowding and boarding/alighting patterns. This has to be distinguished from origin destination information. In most cases on-vehicle sensors are not person specific so that alighting ratios or likelihoods might be established but not alighting points of specific passengers.

Fig. 1 further includes information obtained from payment systems. These can be paper tickets or electronic tickets in their various forms. Whichever the form of payment, the fare structure will determine the detail of information that can be obtained. In flat fare, open systems often only a single tap-in or tap-out is required so that the planner cannot obtain directly origin destination information.

Big Data for Long-Distance Versus Local Public Transport Network Planning

The above discussion has been kept general and “mode-independent” as it relates to all public transport modes. A useful distinction can be made, however, regarding data needed and obtained for long-distance versus short distance, local public transportation planning.

For long-distance travel in general fare structures tend to be distance-based. Furthermore, for service quality and service control purposes, tickets for long-distance journeys are often personalized and for specific seats. For all of these reasons usually both

boarding and alighting stops are observed as well as transfer points from payment systems including gate checks or on-board checks. As a consequence of this, passengers have far less decision points while on long-distance journeys. Their itinerary is fixed, decisions to be made during the journey are mostly how to spend time waiting as well as what to do in case of disruptions or delays. The pre-journey decisions on which connection (and carrier) is chosen are therefore the main focus of big data analytics. (Not least because different journey schedules usually also mean different operators gain profit). Surveys and big data driven analysis of information accessed by travelers before purchasing tickets have become an important marketing research field. Skyscanner has, for example, analyzed search data from its widely spread application to analyse demand patterns and suggests the potential for AI methods on its (big) data (Skyscanner, 2019).

Access and egress for long distance modes is in many cases by other local transport modes so that also more often data are available, for the access to the station/(air)port, but not necessarily the true origin. Other differences are mode depending. For airplane journeys and high speed rail services (a least those on dedicated tracks such as the Japanese Shinkansen) noteworthy is that the on-board travel time undergoes far less variations as the mode does not suffer from interference with other modes. Therefore, collecting information about link travel times is less required. Instead, especially for aviation, the delay sources at the stops, that is, the nodes of the graph, are manifold, stemming from both demand and supply. For local bus journeys, instead, delays on links connecting nodes are frequent. Delays at bus stops might also occur but can be more easily classified as they are largely due to demand factors.

Data Types and Analysis Possibilities by Journey Stage

Pre-Trip Planning

Pre-trip information as to how a person came to choose a specific public transport itinerary are generally difficult to obtain unless activity patterns are known and a person's preferences can be inferred from other, previous choices. For irregular decisions, as well as to understand choices in complex situations with many alternative options, journey planner engines can be mined. The example of "skyscanner enquiries" was noted above. In case the engine is not connected to ticket sales, as usual for local public transport, the main issue is that the matching of enquiries with actual choices is in most cases impossible.

Exceptions are if the journey planner includes a possibility (and, importantly, user consent) to track the person during the day. An example of this is the "Kyoto Arukumachi" application. Upon installation users are asked if they consent to provide their GPS location around the times of enquiry. The multiple GPS records around the time of enquiry as well as the requested information on possible bus routes to a destination possibly allow one a deeper understanding of the person's desired service and actual route chosen (Fig. 2).



Figure 2 Observed GPS points of a (presumably) tourist with data from the "Arukumachi" application of Kyoto City, Japan. For the same user information on bus route information enquired from the application is available.

Pre- and Post-Trip Activities

Traditional national or regional travel diary surveys that aim to capture all trips usually include questions on trip details including access and egress from stops to places of activities as well as on trip purposes. Nowadays, a common way to replace such costly data collection efforts is instead to create a smart phone application that allows GPS tracking a sample of persons (Calabrese et al., 2014).

To obtain activity related questions the application is often supplemented with short pop-up surveys that enquire the purpose of the journey. Other big data based approaches to obtain pre- and post-trip activities of travelers are based on inferring travel activities from non-travel data. For example, integrated payment systems where the same card can be used across stores can be used to infer the travel activities. Payment systems where the same card can be used to pay for public transport systems, such as the Octopus card in Hong Kong, make such an analysis even easier (though again, privacy issues have to be considered.)

GPS tracking, if accurate enough, allows also inferring access and egress modes, route and time. For example, Marra et al. (2019) describe in a recent contribution a tool to observe person's travel to bus stops and route choices based on low frequency GPS data obtained via a smartphone application. Other possibilities to infer this information for larger public transport stations are by observing the doors/gates through which the person enters the facilities. For large transfer points also the time-stamp data from passing arrival gates can provide information as to which previous connection the traveler might have taken.

At Stops

Next to pre-trip information, behavior at stops is least observed, especially for local public transport trips. Unless there are accurate tracking applications or video cameras the arrival time at stops and the queuing behavior for specific bus lines appears to be rarely observed. Data technologies based on Wifi-request signals or Bluetooth signals of travelers offer, however, new applications (Ben-Moshe et al., 2014). The regularly sent request signals of mobile phones allow one to understand the time spent at a stop (if the Mac-address is not randomized) or at least to obtain some information as to how many Wifi/bluetooth enabled devices are at the vicinity of the stop and hence how many persons are likely waiting for a service. Possibly one is also able to observe if travelers leave the stop in case of delays. Difficult to observe will, however, be "detailed strategies" such as when a person is willing to change to a less preferred bus line leaving from the same stop due to delays to his preferred line. For example, observing that passengers destined for a stop that is served by both lines A and B do not board the slow bus A if the fast bus B arrives soon, will give an indication as to trade-off between waiting and travel time.

On-Board

The main two data sources collected by sensors on-board that have been utilized numerous ways for planning applications are automatic vehicle location (AVL) data and automatic passenger count (APC) data. AVL data are collected either by tracking the vehicle through GPS sensors or by signals exchanged between vehicles and stops so that the location of the vehicle can be updated. AVL data are the foundation for passenger real-time information as well as for control strategies to maintain desired headways between services and to avoid "bunching". AVL data might further be utilized to analyze dwell time information at stops which provide some information about demand patterns. APC data provide information about boarding passengers and alighting passengers, but not OD matrices. Through variances in boarding and alighting demand the latter might be inferred, however. Furthermore, the understanding of on-board passengers from APC data are an important data source for service capacity planning. The possibilities offered by this data source are discussed in more detail in Furth et al. (2006). As a specific application example, Pi et al. (2018) used both AVL and APC data to measure the performance of a public transport system in the Pittsburgh region. (See also article 10445 on transit operations for further reading.)

Boarding and Alighting Points

The data collected via the different payment systems for public transport are one of the most important data sources for public transport planning. For long-distance public transport clearly ticket sales are the most direct information on overall demand (except latent demand due to capacity issues) and demand patterns.

For local transport where tickets are not run specific, electronic ticketing through "smart cards" has brought significant analysis possibilities. To note is that the data fit the notion of "big data generated for different primary purposes." The primary purposes of smart cards are reduced cash handling, convenience for passengers, integrated fare collection and enabling a fair revenue split among several operators. They furthermore allow operators easier implementation of advanced fare structures. For example, "price capping" where users are not charged above a certain limit if they make several journeys would not be feasible without recording of the paid fares. Also loyalty point systems where users obtain discounts or rewards (in some form) for regularly usage of a service are not possible without smart cards.

Exploration of smart card data for additional "secondary" purposes, however, is the main academic interest as discussed in the book "Public Transport Planning with Smart Card Data" (Kurauchi and Schmöcker, 2016). Construction of origin-stop to destination-stop profiles has been a major research area, since in many cases, especially bus systems, passengers do either only tap-in or tap-out. In case of distance-depending fares, as for most metro services, where passengers tap-in and tap-out, a challenge remains to

identify the transfer points and routes taken. Destinations and transfer points are often inferred by considering onward connections as well as long-term patterns of passengers. Routes can be estimated from transfer points (if observed) as well as travel times and considering schedule information or AVL records of runs. A number of contributions have looked into inferring activities of passengers by considering the land-use around origin and destination stops, duration time before records of subsequent journeys as well as regularity patterns of the users. Sensitivity of user behavior with respect to service disruptions is another analysis area. Smart card data can be used to observe immediate changes in network flows, as well as, if the disruption is longer, user adjustment before, during and after the disruption. Similarly, with longer term smart card data, adjustments of users to, for example, schedule, service frequency changes or other policy interventions can be observed, though the literature on long-term data analysis of smart card data appears to be still thin. [Pelletier et al. \(2011\)](#) review the analysis of smart card data (up to that year) and discuss that the data can be used for “strategic (long-term planning), tactical (service adjustments and network development), and operational (ridership statistics and performance indicators)” purposes.

Besides user behavior-related analysis, through fusion of smart card data with other data, among others, service quality indicators can be obtained or spatial and temporal equality and fairness of a public transport service can be assessed. The smart card data can also be used as surrogate for information obtained from on-board sensors, if these are not available. For example, the time of first and last tap-in or tap-out at stops can be used to estimate the arrival and departure time at stops. For buses, where tap-in and tap-outs are occurring on-board, further they can replace on-board counts. In metro systems where the tapping is mostly occurring at gates, for passenger load estimates on specific runs, one needs to estimate the time from and to the gate as an additional source of uncertainty in the estimation process.

To remark is that smart cards tend to become replaced by alternative payment systems. Credit cards with which one can tap-in directly, mobile phone apps that allow payment through close range communication or based on GPS tracking are fast increasing. The data content and hence how these data are explored will be the same or enhanced if the payment data can be combined with other data from the credit cards or mobile phones and if user consent is obtained.

Big Data Generated by Non-PT Sources

Finally, it needs to be stressed that there is a whole range of other non-public transport big data that can be directly utilized for public transport service planning but do not provide journey stage specific information. Big data from other transport modes and data unrelated to transport might here be distinguished. For the former, taxi probe data or electronic toll collection for cars are a good example. These will give a good indication of travel times on links used also by public transport services. An example for the latter group are again data possibilities through mobile phone data. In contrast to the formerly described records obtained through bespoke applications, also non-application based storage of aggregate or disaggregate movement and population density data provide the operator with useful information.

Aggregate data on “mesh basis” can be utilized to obtain information regarding how many persons are in a specific time in a specific mesh. In some cases also mesh movements can be obtained. Both sources can inform public transport operators regarding potential demand and provide some information as to how much demand is moving by public transport if the data are compared to historic ridership data. Disaggregate GPS tracks of individuals are (and should be) restricted in most cases, but obviously, could provide a large range of the above discussed data on access and egress tracking, activity information as well as even regarding lines used. The data can further provide one with information on travel times between parts of the city and can be the basis to evaluate where public transport is competitive or not. The book “Mobile Technologies for Activity-Travel Data Collection and Analysis” by [Rasouli and Timmermanns \(2014\)](#) provides an overview and range of applications on this subject.

Beyond such data, one can imagine a range of “innovative usage” of other big data; following are a few examples: Scanning of social media data can provide information about public transport perceptions and routes (not) recommended. Sales data and shopping customer flows obtained via various sensors can be used to understand and predict patterns in public transport usage behavior. Similarly, mining data on hotel bookings or sightseeing and correlating these to factors such as weather, seasonality, events or even reliability of the public transport service itself, can be a potential source for data prediction. It clearly has to be emphasized once more that for all these applications, in particular the “innovative” ones, ensuring data privacy issues need to be a primary concern.

Conclusions and Further Reading

The purpose of this contribution has been emphasizing the different big data sources and their analysis potential for public transport planners. Less emphasis has been given to the resulting practical implications for operators, travelers as well as modelers/planners. For this we refer to the longer review by [Welch and Widita \(2019\)](#) structured according to applications and methods.

An important overall message of this chapter might be summarized as follows: Big data has led to smart travelers and smart operators. This point is also discussed in [Milne and Watling \(2019\)](#) but not specific to public transport. Travelers are better informed to make time and cost minimizing choices and operators can develop better transport networks and better real time control strategies ([Fonzzone et al., 2016](#); [Wilson and Hemily, 2016](#)). This is leading, however, to numerous research challenges not only in terms of

how the data can be utilized but also in modeling the more complex decision makers of travelers and the interaction with control strategies as is discussed in the book edited by [Gentile and Nökel \(2016\)](#) which is recommended as further read on transit modeling.

Finally, it needs to be noted, that to keep this contribution within the given length limitation, only relatively few references have been given. It is beyond this chapter to give adequate references to all work as the literature is vast. The given references often link to books as well as review and other overview papers to which the reader may be referred to.

References

- Ben-Moshe, B., Hadas, Y., Levi, H., 2014. Energy-efficient framework for indoor and outdoor tracking of public transit passengers using Bluetooth-enabled devices. *Transportation Research Board 93rd Annual Meeting*. No. 14-2485.
- Calabrese, F., Ferrari, L., Blondel, V., 2014. Urban sensing using mobile phone network data: a survey of research. *ACM Computing: Surveys* 1–23.
- Fonzone, A., Schmöcker, J.-D., Viti, F., 2016. New Services, new travelers and new models? Directions to pioneer public transport models in the era of big data. *J. Intel. Transport. Syst.* 20 (4), 311–315.
- Furth, P.G., Hemily, B., Muller, T.H.J., Strathman, J.G., 2006. Using archived AVL-APC data to improve transit performance and management. *TCRP Report 113*. Transportation Research Board of the National Academies. Washington, DC.
- Gentile, G., Nökel, K., 2016. *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*. COST Action TU1004 (TransITS). Springer, Cham. ISBN 9783319250809.
- Kurauchi, F., Schmöcker, J.-D., 2016. *Public Transport Planning with Smart Card Data*. 274 pages, CRC Press (Taylor and Francis Group). ISBN 9781498726580.
- Laney, D., 2001. *3D Data Management: Controlling Data Volume, Velocity and Variety*. META Group Research Report.
- Milne, D., Watling, D., 2019. Big data and understanding change in the context of planning transport systems. *J. Transport Geogr.* 76, 235–244.
- Marra, A.D., Becker, H., Axhausen, K.W., Corman, F., 2019. Developing a passive GPS tracking system to study long-term travel behavior. *Transport. Res. Part C: Emerging Technol.* 104, 348–368.
- Pelletier, P.-M., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. *Transport. Res. Part C: Emerging Technol.* 19 (4), 557–568.
- Pi, X., Egge, M., Whitmore, J., Silbermann, A., Qian, Z.S., 2018. Understanding transit system performance using AVL-APC data: an analytics platform with case studies for the Pittsburgh region. *J. Public Transport.* 21 (2), 19–40.
- Rasouli, S., Timmermanns, H., 2014. *Mobile Technologies for Activity-Travel Data Collection and Analysis*. 325 pages, IGI Global. ISBN 978146661707.
- Skyscanner, 2019. *AI: The Future of Airline Route Strategy?*. Online Report. Available from: <https://partners.skyscanner.net/how-ai-can-better-support-airlines-in-their-fleet-developments/thought-leadership>. [Accessed October 2019].
- Welch, T.F., Widita, A., 2019. Big data in public transportation: a review of sources and methods. *Transport Rev.* 39 (6), 795–818, doi:10.1080/01441647.2019.1616849.
- Wilson, N., Hemily, B., 2016. Opportunities provided to transit organizations by automated data collection systems: Challenges and thoughts for the future.. In: Kurauchi, F., Schmöcker, J.-D. (Eds.), *Public Transport Planning with Smart Card Data*. CRC Press, Taylor & Francis Group, Boca Raton, FL, pp. 245–260.

Active Transport: Heterogeneous Street Users Serving Movement and Place Functions

Regine Gerike, Stefan Hubrich, Caroline Koszowski, Bettina Schröter, Rico Wittwer, Technische Universität Dresden, Faculty of Transport and Traffic Sciences “Friedrich List”, Chair of Integrated Transport Planning and Traffic Engineering, Dresden, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	140
Characteristics of Active Transport	141
Common Characteristics of Walking and Cycling	141
Walking	141
Cycling	141
Drivers and Barriers	142
Overview	142
Person and Household Characteristics	142
Density of Spatial Structures and Diversity of Land-Use	143
Connectivity and Density of Walking and Cycling Networks	143
Comfort and Safety of Walking and Cycling Facilities	143
Quality of Streetscape as Public Space	144
Policies and Strategies for Promoting Active Transport	144
Summary and Conclusions	145
Acknowledgment	145
References	145
Further Reading	146

Introduction

The overarching term “active transport” includes all transport modes which require physical activity to reach a destination. Active transport is a quite recent term; traditionally, walking and cycling were categorized as nonmotorized modes of transport (Gerike and Parkin, 2016). This definition can have negative connotations in that nonmotorized transport modes comprise everything else but motorized vehicles. This more narrow perspective suggests low level of attention and priority given in this direction as compared to motorized transport modes. The term “active transport” has a completely different connotation. It suggests positive dynamics and attracts higher attention; active modes are not on the periphery but are to be treated with equal regard along side the individual and public motorized transport modes both in political discussions and in level of assigned priority. The increased acceptance and usage of the term active transport thus reflects the substantial growth of and support for walking and cycling that can be observed in many cities and countries worldwide.

The term active transport also allows for electric support of the human-powered vehicles as this is the case for electrically assisted bicycles such as pedelecs: These belong to the active transport modes as long as human power is necessary for moving forward. The term “vulnerable road users” is also used for summarizing walking and cycling in the context of safety analysis and management. Further terms such as “slow modes” also are in use but are less present in the academic and non-academic terminology.

The variety of bicycles has increased substantially in recent years with growing numbers of special bicycles such as cargo bikes, bicycles with trailers, folding bikes, tandems, for example, for parents with their kids, or electrically assisted bicycles. In addition, the new solely electrically powered micro-vehicles such as the electric scooters are currently spreading worldwide. These share many similarities in their characteristics and requirements with cyclists. This chapter focuses on walking and human-powered cycling, particularly when it comes to drivers and barriers. Empirical evidence on users and usage of the new electrically powered micro-vehicles is, thus far, fragmented; it documents the high dynamics triggered by these innovative services and vehicles but does not give a consistent picture of users and usage. All traditional and innovative forms of micro-mobility are considered in the discussion of measures for promoting active transport and micro-mobility in the second part of this chapter.

This chapter is structured along three main parts. Following to the introduction, main characteristics of walking and cycling are described. The subsequent section *Drivers and Barriers* summarizes determinants of walking and cycling as identified in the literature: What makes people walking and cycling and what are important barriers? Based on this knowledge on characteristics and determinants of active transport, recommendations for policies and strategies for promoting walking and cycling are developed in the last part of this chapter.

Characteristics of Active Transport

Common Characteristics of Walking and Cycling

Pedestrians and cyclists are vulnerable street users; they do not have any physical shell that protects them in case of an accident or fall. Both modes are far more space efficient than the motorized modes due to their smaller dimensions and slower speeds. Standard pedestrians have a width of mostly 0.80 m and bicyclists of 1.00 m according to international guidelines on urban street design; the design car has a width of around 1.80 m and needs additional space on both sides for rear-view mirrors and lateral movements (Gerike et al., 2019). Lanes for pedestrians and bicyclists are thus narrower compared to lanes for the motorized modes and, in addition, have higher capacities measured as moved users per hour.

Both active modes have low or no emissions in terms of noise, air pollutants and greenhouse gases. They are flexible in space and time. Users can walk or cycle wherever and whenever they want to; they do not depend on any fixed schedule of public transport services or dedicated parking facilities for cars. Walking and cycling require little to no money, presenting low costs for users, operators, and society as a whole. Walking and cycling also contribute to public health objectives as they are one type of moderate physical activity and thus improve quality of life and health (Mueller et al., 2015).

With these common characteristics, active transport is located at the intersection of urban planning, transport planning, and public health. Active transport supports the achievement of objectives in all three of these disciplines. At the same time, stakeholders from all three disciplines can foster active transport through their specific policy measures and expertise. Joint efforts efficiently allow for the maximization of synergies and the attainment of the common objective of increased walking and cycling levels.

Walking

Walking is the slowest of all transport modes. Pedestrians are therefore very distance and detour sensitive. They need direct connections between their origins and destinations and hardly respect any detours, for example, when crossing a street after having left the bus. People only walk for short distance trips; 61% of walking trips in Germany are shorter than 1 km (80% for 2 km) (infafas et al., 2018). Dense spatial structures and mixed land-use are therefore paramount for achieving high walking levels. Another consequence of pedestrians' distance-sensitivity is that pedestrian facilities generally need to be designed for movements in two directions; pedestrians will not accept changing the side of the street for walking on one-directional sidewalks. Walking as a transport mode will never be suitable for satisfying all daily travel needs of any one person; it needs to be combined with cycling or motorized modes and shows particular synergies with high quality public transport provision (Gascon et al., 2019).

Unencumbered walking does not require any specific equipment or skills; it is a human desire from the very early to the latest of life stages. For many, walking represents a form of independence, self-confidence, social participation, and health. These specific characteristics of walking lead to the phenomenon that the pedestrians group is the most heterogeneous one from all transport modes and includes the whole of society with all social and age groups, people with and without mobility restrictions, women and men, etc. Another consequence of this inherent human desire to walk and to move is that people do not reflect on walking. It is a strongly habitual activity and as normal as brushing one's teeth. As a result, people do not remember walking trips and activities as reliably as trips with other transport modes (Aschauer et al., 2018). Pedestrians are flexible in terms of the quality of walking facilities and the physical environment—high pedestrian volumes are found even for narrow sidewalks and at streets with high traffic loads (Kim et al., 2019). At the same time, the literature shows that walking behavior is influenced by the streetscape, the quality of the public space, and the characteristics and usages of the buildings abutting the street and in the neighborhood (Ewing et al., 2016; Gehl, 2010; Mehta and Bosson, 2018; Mehta, 2013).

Due to their low speed, pedestrians can spontaneously stop or change direction which leads to conflicts with bicyclists if both use the same space in the street. Walking not only serves movement but also place functions. When people walk to reach their destinations, streets serve as conduits that should enable all user groups to move safely and comfortably. Pedestrians also use streets as destinations for carrying out 'place activities'. These include: (1) necessary activities such as waiting for the bus; (2) optional activities such as eating or watching the street; and (3) social activities in the form of communication which require the presence of further persons (Gehl, 2010). The objective functions for movement and place functions are completely different: the maximization of moving comfortably and safely vs. the maximization of the number and duration of place activities, respectively.

Pedestrians, that is, those walking as a transport mode, hardly have any lobby support in political discussions. Important reasons for this are that no technology is needed for walking and that no walking industry exists; in addition, all societal groups are (potential) pedestrians. With the exception of safety, almost no data exist on walking, nor do any measurable political goals or monitoring systems, thus significantly reducing the ability to report on the statistics or the successes of walking as a transport mode.

Cycling

Cycling, in contrast to walking, is a conscious activity. Bicyclists and non-bicyclists have a clear opinion about cycling and about why they personally cycle or do not cycle. The level of cycling depends far more than walking on high quality infrastructure. Seamless, comfortable, and safe cycling facilities substantially contribute to higher cycling levels (Mueller et al., 2018). Many municipalities have invested and are still investing high efforts into improving cycling facilities leading to substantial increases in cycling levels in many places in recent years (see <http://www.epomm.eu/tems/for> an overview of modal split developments in various European

cities). The type of cycling facility as well as the rules for deciding which cycling facility is suitable for each setting differ greatly between countries and also between municipalities within one country (Gerike et al., 2019). One reason for this is that cyclists are the only street user group that can be provided either in the carriageway together with or next to the motorized transport modes or on the footway together with pedestrians.

Cycling requires a bicycle, suitable clothes, and safety gear (e.g., a helmet, reflectors, etc.). Cyclists are more affected by varying weather conditions than pedestrians as they are less protected (e.g., they cannot take an umbrella) and more sensitive to slippery or snowy surfaces. In contrast to walking, this leads to substantial differences in cycling levels depending on the season and climate. One of the main barriers for cycling is the inherent higher demand for physical effort (than, e.g., walking), particularly in hilly areas (Handy et al., 2014; Heinen, 2011).

Cycling speed is also increased and is in urban areas almost equal to the motorized modes (when looking at the complete trip, including all stages from the origin to the destination). Such relatively high speeds allow for longer distances than for walking. 31% of all cycling trips range between 2 and 5 km—this appears to be a feasible distance for many cyclists. 11% of all cycling trips are 5–10 km long, and 6% of all cycling trips are even longer (infas et al., 2018). The higher availability of electrically assisted bicycles might further increase the proportion of longer cycling trips. For all transport modes, 58% of trips are shorter than 5 km and 74% are shorter than 10 km; thus, in terms of distance, cycling is well suited for covering daily travel needs. It can also be used for transportation of relatively large and heavy items, enabling cycling as a transport mode for most distances and trip purposes. Synergies with public transport nevertheless also exist for cycling.

The proportion of mandatory trips (work, education and, business) is higher for cycling than for walking. Therefore, once a person decides to cycle, this behavior remains more stable than walking for leisure as the main trip purpose (infas et al., 2018). Another result of the higher speed of cycling is that personal security is less of an issue than for walking, but safety in general and, more particularly, perceived safety of cycling facilities are more important.

Drivers and Barriers

Overview

Various conceptual frameworks concerning drivers as well as barriers for active transport are available in the literature (Götschi et al., 2017; Koszowski et al., 2019). There is consensus that multi-level and multi-disciplinary approaches are required in order to understand and to purposefully shape walking and cycling. One important determinant of active transport is the built environment which includes urban structures, land-use, and transport supply. It has an impact on travel behavior in general, but specifically on active transport on the regional, city, neighborhood and street levels. The whole transport system must be considered in order for active transport to be well understood. The most efficient measures for promoting active transport could potentially be those which address motorized modes, for example, through the improvement of public transport supply or the restriction of car traffic.

Socio-demographic, socio-economic, and socio-psychological variables are relevant at the levels of single persons, households, peers, and the (neighborhood) community. The distinction between objective and perceived quality of the available built environment and transport services is imperative; for example, objectively safe cycling facilities will not be used if they are perceived as risky by the (potential) users. Implemented policy measures and feedback loops for promoting active transport can change the system and bring dynamics into the conceptual frameworks of active transport. Policies can directly influence travel behavior when, for example, new walking and cycling infrastructures are implemented. Policies might also indirectly influence travel behavior through a change in overall attitude and mindset. Policies should also address governance structures including institutions, processes, finance and legal issues.

Person and Household Characteristics

Socio-demographic and socio-economic characteristics are important indicators for the propensity to walk and to cycle (Gascon et al., 2019; Handy et al., 2014). Walking levels are highest in the youngest and oldest age groups with hardly any differences between women and men (infas et al., 2018; Harms et al., 2014). Cycling levels are more evenly distributed across age groups and gender in countries with high modal-split proportions of cycling whereas, in countries with lower cycling levels, mainly young men cycle (Götschi et al., 2015). The age group of 9–17-year-old children and adolescents shows a high affinity toward the active modes (Harms et al., 2014). Through modelling their own travel behavior and setting an example, parents have a primary role in shaping their childrens' travel behavior and for shaping independent travel in early years (Ghekiere et al., 2016). The odds of walking and cycling increase with higher levels of education (Handy et al., 2014); however, findings on the influence of income on active transport are inconclusive (Koszowski et al., 2019).

There is consensus in the literature regarding the influence of the ownership of a driving license and the availability of motorized vehicles. Both factors inhibit active transport and encourage people to adopt car-oriented travel behaviors (Clark and Scott, 2013; Damant-Sirois and El-Geneidy, 2015). Accordingly, better availability of bicycles in a household increases the propensity to cycle (Heinen et al., 2011).

Socio-psychological factors such as perceptions, preferences, attitudes, habits, perceived controls on behavior (e.g., linked with the ability to carry luggage or to ride a bicycle), and social and personal norms are generally found to determine individual mode

choice, more specifically active mode shares. The theory of planned behavior is often applied in this context (Ajzen, 1991). Affirmative opinions about active transport are positively correlated with active transport levels in a bidirectional causal relationship (Kroesen et al., 2017). Motivating factors for cycling include status and lifestyle promotion, mental and physical relaxation, comfort, time savings, privacy and security (Heinen et al., 2011). Heinen (2011) lists the following individually perceived barriers for cycling: time loss compared to motorized transport modes, overly long distances to relevant destinations, bad weather, lack of parking facilities for bicycles, physical overload (e.g., caused by differences in altitude or insufficient fitness and sweating). The influence of subjective motivational factors on walking is less clear than for cycling. For walking, the built environment is of highest importance as described below.

Density of Spatial Structures and Diversity of Land-Use

The “5 Ds” (Density, Diversity, Design, Distance to public transport, and Destination accessibility) are consistently significant and influential particularly for walking (Ewing and Cervero, 2010; Gascon et al., 2019; Lai and Kontokosta, 2018). Ewing et al. (2016) demonstrate the specific relevance of Density, measured in their example as floor area ratio and population density within a quarter mile of the investigated commercial streets. Diversity is often measured by entropy measures describing the number and variety of different land-use types in a given area (Ewing and Cervero, 2010; Lai and Kontokosta, 2018). Lower Distances particularly to rail-based public transport consistently and significantly increase pedestrian volumes (Ewing et al., 2016). Design-variables describe the characteristics and, more specifically, the connectivity of the street network, measured, for example, as intersection density or as a proportion of 4-way intersections (Ewing and Cervero, 2010). Destination accessibility describes the level at which relevant activities can be reached (Ewing and Cervero, 2010). Destinations are operationalized, for example, by the number of nearby stores and amenities weighted by their distance; these are hardly significant and show an overlap with Diversity.

Urban design and land-use are less important for cycling than for walking (Buehler et al., 2017). Thanks to higher speeds, longer distances can be covered with the bicycle resulting in less dependence on the proximity of relevant destinations. The range of bicycle trips further increases with higher availabilities of electrically assisted bicycles in the population and the extension of bicycle facility networks that allow for high speed and cover larger areas, for example, on the regional as well as municipal levels.

Connectivity and Density of Walking and Cycling Networks

Mixed findings exist for the design variables related to the connectivity and density of infrastructure networks as introduced above. These are significant in some studies, in others they are not (Ewing et al., 2016). Hooper et al. (2015) find positive correlations, the higher the street connectivity and the higher the accessibility to commercial land-use within the neighborhood, the higher the odds of walking. Cervero et al. (2009) find that residents in neighborhoods with highly connected street networks tend to have higher active transport levels of a minimum of 30 min/day on average. A high connectivity is mainly related to fine-grained street networks with many intersections [nodes] and requires many route options to get from each starting point to the respective destination. The shorter the distances are between the intersections, the higher the share of pedestrians and cyclists in the local modal split (Ewing and Cervero, 2010; Kaplan et al., 2016). Various studies confirm the importance of the density and connectivity of the network of cycling facilities for achieving high cycling levels (Damant-Sirois and El-Geneidy, 2015; Handy et al., 2014; Mueller et al., 2018).

Comfort and Safety of Walking and Cycling Facilities

A safe and comfortable infrastructure is as important for cycling as urban design and land-use are for walking. This is in direct contrast to pedestrians who also walk on deficient pedestrian facilities if dense and mixed spatial structures and public transport services generate high walking levels.

The perceived safety of provided infrastructure is hence essential for the acceptance of cycling as a transport mode (Handy et al., 2014). Types of cycling facilities and criteria for deciding on their suitability differ greatly between countries and also between municipalities within one country (Schröter et al., submitted). Countries with high levels of cycling such as the Netherlands or Denmark tend to provide facilities for cyclists located off the carriageway and only allow for mixing cyclists with motorized vehicles in the carriageway with low speed limits of maximum 30 km/h. Cycling facilities should be designed in a way that allows for the accommodation of standard and nonstandard bicycles as well as for all the innovative electric microvehicles that have emerged in the last years and months or that might emerge in the future. The so-called safety-in-numbers effect describes the phenomenon of improved safety for cyclists when the number of cyclists in specific regions or at single infrastructure points such as junctions increases (Elvik and Bjørnskau, 2017). The explanation of the safety-in-numbers effect is bi-directional: Better infrastructures for cyclists potentially increase cycling levels which, in turn, lead to higher visibility and safety levels for bicyclists.

For walking, Kang (2015) and Kim et al. (2019) find significant positive impacts of sidewalk widths, crosswalks, trees, and negative impacts of slope on pedestrian volumes. The number of traffic lanes has a significantly positive influence on walking levels but is highly correlated with the distance to public transport. Lai and Kontokosta (2018) compute a composite variable streetscape as the product of sidewalk coverage, pavement quality, and street amenity. This variable significantly increases pedestrian volumes on weekend days but not on work days.

Quality of Streetscape as Public Space

The “5 Ds” as introduced above are of high relevance particularly for walking not only on the neighborhood level but also for the streetscape itself. This holds particularly for Design but also for the other Ds. [Ewing et al. \(2016\)](#) show the significant influence of floor–area ratios of the streets themselves (computed as the total building floor area for parcels abutting the street, divided by the total area of tax lots) and of the proportion of retail frontage along the block face on pedestrian volumes.

For the D-variables on the street level, the streetscape itself also matters for walking. Transparency is defined as the degree to which people can see or perceive what lies beyond the edge of a street. This is measured, for example, by the proportion of first floors with windows and of active uses of adjacent buildings, and significantly increases walking volumes ([Ameli et al., 2015](#); [Ewing et al., 2016](#); [Hamidi and Moazzeni, 2019](#)). [Ewing and Handy \(2009\)](#) define, based on expert ranking, further relevant variables for walking; these are imageability, enclosure, human scale, and complexity. These are significant in some of the few existing studies that empirically analyze their influence on walking volumes ([Ewing et al., 2016](#); [Gehl, 2010](#); [Mehta and Bosson, 2018](#)).

[Ewing et al. \(2016\)](#) view transparency as the most important urban design quality on the street level; they analyze the influence of the following three variables separately: proportion of windows, of street furniture, and of active uses. The latter two are found to significantly increase pedestrian volumes. Overall, the three streetscape design features add significantly to the explanatory power of the statistical models explaining pedestrian volumes compared to models with only the D-variables on the neighborhood and street levels. Street furniture includes signs, benches, parking ticket machines, trash cans, newspaper boxes, bollards, street lights as well as anything at the human scale that increases the complexity of the street. Public seating is found to be of particular importance. The active use includes shops, restaurants, public parks, and other uses that generate significant pedestrian traffic. Inactive uses include blank walls, driveways, parking lots, vacant lots, abandoned buildings, and offices with no apparent activity.

Policies and Strategies for Promoting Active Transport

The aforementioned similarities and differences between walking and cycling directly translate into policies and strategies for their promotion. Both active modes benefit from integrated strategies that aim on the one hand at strengthening walking, cycling and also public transport and on the other hand at restricting car use. Improvements of public transport services are important in this context, particularly when walking levels should be increased. Measures for restricting car use are of highest relevance for promoting active transport; fast access to destinations by car and comfortable parking facilities at the destination are one main barrier for choosing to walk (often in combination with public transport) or to cycle. Examples for these restrictive measures are the re-allocation of street space to active modes or public transport (as in, e.g., London) ([Wittwer et al., 2019](#)) or parking management—when the number of car parking facilities is reduced, their prices are increased or the maximum parking duration is limited.

Both walking and cycling benefit from public health campaigns and activities aimed at stimulating active transport as physical activity. One of the core objectives in the Mayor of London’s Transport Strategy is that by 2041, all Londoners should do “at least the 20 min of active travel they need to stay healthy each day” ([Mayor of London, 2018:22](#)).

The main barrier for walking specifically is distance. Dense and mixed land uses are therefore key for the promotion of walking. On the detailed microlevel of a street segment, the ground-floor usages and the design of the façades of the buildings abutting the street are of high relevance and more important than the street design. Pedestrians move slowly and experience their environment intensively, particularly when they are place users in the street. Active frontages and soft edges of the adjacent building invite pedestrian activities and city life whereas inactive walls without any interesting things to see or do cause people to walk more quickly and as efficient as possible to their destination.

Streetscape also matters for pedestrians. Unobstructed through zones with sufficient widths should be provided at the footway for accommodating the expected pedestrian volumes ([Gerike et al., 2019](#)). Seating significantly increases place activities in all identified studies; shade and shelter as well as all measures increasing the perceived attractiveness of the public street space also invite street life ([Mehta, 2013](#); [Mehta and Bosson, 2018](#)). Safe crossing facilities need to be provided at main crossing points at specific locations or in linear forms. The principles of inclusive design need to be respected, particularly when providing for walking in order to enable all user groups to use the available facilities.

Cycling is less dependent on dense urban structures. Seamless, safe, and comfortable infrastructures that accommodate all forms of cycling vehicles are key for achieving high cycling levels. (Potential) cyclists do not cycle when they do not feel safe even when the provided facilities are judged to be safe by planners and further experts. Suitable cycle parking facilities are one integral component of these infrastructures.

Mobility cultures supporting active transport are more important for cycling than for walking. Particularly in starter cities and countries, the uptake of cycling can be accelerated with campaigns, festivals, and further so-called soft measures ([Koszowski et al., 2019](#)).

Conflicts in space and time between pedestrians and cyclists exist. Both compete for the available street space that is clearly limited, particularly in inner urban areas; both compete for green time at signalized junctions. Such conflicts can only be solved on a case-by-case basis based on clear political priorities given on the city level.

Summary and Conclusions

Active transport currently shows positive dynamics in many urban areas and countries around the world. Walking and cycling levels are increasing which goes hand in hand with growing support and commitment from planners and political decision makers. More and more stakeholders agree that there is no alternative to high levels of walking and cycling in combination with public transport and the emerging new micro-vehicle types when shaping future transport systems.

Societies worldwide are facing major challenges with growing cities, congested transport systems, ambitious goals for greenhouse gas emissions, noise, and air pollutants. Highest attention needs to be paid to safety so that achievements in promoting active transport are not compromised by increased numbers in accidents and injuries. More and more stakeholders commit to the Vision-Zero goal: No person should be killed or severely injured in traffic. This is a challenging goal but one key success factor for strengthening walking and cycling.

The promotion of active transport and the design and management of liveable streets is an interdisciplinary task which supports ambitions in urban planning, transport planning, and public health and can be achieved more successfully and efficiently if all the relevant stakeholders collaborate.

Acknowledgment

Funding for the research project “Active Mobility: Improved quality of life in urban agglomerations” has been received from the German Environment Agency—UBA, and the research benefited from the European Union project “PASTA—Physical Activity Through Sustainable Transport Approaches”. This work was also supported by the European project Multimodal Optimisation of Roadspace in Europe (MORE) (<https://www.roadspace.eu/>) which receives funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No. 769276.

References

- Ajzen, I., 1991. The theory of planned behavior. *Organ. Behav. Human Decision Process*. 50 (2), 179–211.
- Ameli, S.H., Hamidi, S., Garfinkel-Castro, A., Ewing, R., 2015. Do Better Urban Design Qualities Lead to More Walking in Salt Lake City, Utah? *J. Urban Design* 20 (3), 393–410.
- Aschauer, F., Hössinger, R., Schmid, B., Gerike, R., 2018. Reporting quality of travel and non-travel activities: A comparison of three different survey formats. *Transport. Res. Proc.* 32, 309–318.
- Buehler, R., Pucher, J., Gerike, R., Götschi, T., 2017. Reducing car dependence in the heart of Europe: lessons from Germany, Austria, and Switzerland. *Transport Rev.* 37 (1), 4–28.
- Cervero, R., Sarmiento, O.L., Jacoby, E., Gomez, L.F., Neiman, A., 2009. Influences of built environments on walking and cycling: lessons from Bogotá. *Int. J. Sustain. Transport*. 3 (4), 203–226.
- Clark, A.F., Scott, D.M., 2013. Does the social environment influence active travel? An investigation of walking in Hamilton, Canada. *J. Transport Geogr.* 31, 278–285.
- Damant-Sirois, G., El-Geneidy, A.M., 2015. Who cycles more? Determining cycling frequency through a segmentation approach in Montreal, Canada. *Transport. Res. Part A: Policy Pract.* 77, 113–125.
- Elvik, R., Bjørnskau, T., 2017. Safety-in-numbers: A systematic review and meta-analysis of evidence. *Safety Sci.* 92, 274–282.
- Ewing, R., Cervero, R., 2010. Travel and the Built Environment. *J. Am. Plan. Assoc.* 76 (3), 265–294.
- Ewing, R., Hajrasouliha, A., Neckerman, K.M., Purciel-Hill, M., Greene, W., 2016. Streetscape features related to pedestrian activity. *J. Plan. Educ. Res.* 36 (1), 5–15.
- Ewing, R., Handy, S., 2009. Measuring the unmeasurable: urban design qualities related to walkability. *J. Urban Design* 14 (1), 65–84.
- Gascon, M., Götschi, T., Nazelle, A. de, Gracia, E., Ambrós, A., Márquez, S., Marquet, O., Avila-Palencia, I., Brand, C., Iacorossi, F., Raser, E., Gaupp-Berghausen, M., Dons, E., Laeremans, M., Kahlmeier, S., Sánchez, J., Gerike, R., Anaya-Boig, E., Panis, L.L., Nieuwenhuijsen, M., 2019. Correlates of walking for travel in seven EUROPEAN cities: The PASTA Project. *Environ. Health Perspect.* 127 (9), 97003.
- Gehl, J., 2010. *Cities for people*. Island Press, Washington, Covelo, London.
- Gerike, R., Dean, M., Koszowski, C., Schröter, B., Wittwer, R., Weber, J., Jones, P., 2019. “Urban Corridor Road Design: Guides, Objectives and Performance Indicators”, Deliverable D1.2, Multimodal Optimisation of Roadspace in Europe, available at: <https://www.roadspace.eu/>.
- Gerike, R., Parkin, J., 2016. *Cycling futures: From research into practice*. Routledge, London, New York.
- Ghekiere, A., van Cauwenberg, J., Carver, A., Mertens, L., Geus, B. de, Clarys, P., Cardon, G., Bourdeaudhuij, I. de, Deforche, B., 2016. Psychosocial factors associated with children’s cycling for transport: A cross-sectional moderation study. *Prevent. Med.* 86, 141–146.
- Götschi, T., Nazelle, A. de, Brand, C., Gerike, R., 2017. Towards a Comprehensive Conceptual Framework of Active Travel Behavior: a Review and Synthesis of Published Frameworks. *Curr. Environ. Health Rep.* 4 (3), 286–295.
- Götschi, T., Tainio, M., Maizlish, N., Schwanen, T., Goodman, A., Woodcock, J., 2015. Contrasts in active transport behaviour across four countries: how do they translate into public health benefits? *Prevent. Med.* 74, 42–48.
- Hamidi, S., Moazzeni, S., 2019. Examining the Relationship between Urban Design Qualities and Walking Behavior: Empirical Evidence from Dallas, TX. *Sustainability* 11 (10), 2720.
- Handy, S., van Wee, B., Kroesen, M., 2014. Promoting cycling for transport: research needs and challenges. *Transport Rev.* 34 (1), 4–24.
- Harms, L., Bertolini, L., te Brömmelstroet, M., 2014. Spatial and social variations in cycling patterns in a mature cycling country exploring differences and trends. *J. Transport Health* 1 (4), 232–242.
- Heinen, E., 2011. “Bicycle Commuting”, available at: <https://books.bk.tudelft.nl/index.php/press/catalog/book/isbn.9781607507710> (accessed 21 November 2019).
- Heinen, E., Maat, K., van Wee, B., 2011. The role of attitudes toward characteristics of bicycle commuting on the choice to cycle to work over various distances. *Transport. Res. Part D: Transport Environ.* 16 (2), 102–109.
- Hooper, P., Knuiman, M., Foster, S., Giles-Corti, B., 2015. The building blocks of a ‘Liveable Neighbourhood’: Identifying the key performance indicators for walking of an operational planning policy in Perth, Western Australia. *Health Place* 36, 173–183.
- infas, DLR, IVT and infas 360 (2018), “Mobility in Germany: Tabulated Results, available at: <http://www.mobilitaet-in-deutschland.de/> (accessed 20 November 2019).
- Kang, C.-D., 2015. The effects of spatial accessibility and centrality to land use on walking in Seoul, Korea. *Cities* 46, 94–103.
- Kaplan, S., Nielsen, T.A.S., Prato, C.G., 2016. Walking, cycling and the urban form: A Heckman selection model of active travel mode and distance by young adolescents. *Transport. Res. Part D: Transport Environ.* 44, 55–65.

- Kim, S., Park, S., Jang, K., 2019. Spatially-varying effects of built environment determinants on walking. *Transport. Res. Part A: Policy Pract.* 123, 188–199.
- Koszowski, C., Gerike, R., Hubrich, S., Götschi, T., Pohle, M., Wittwer, R., 2019. Active mobility: bringing together transport planning, urban planning, and public health. In: Müller, B., Meyer, G. (Eds.), *Towards User-Centric Transport in Europe: Challenges, Solutions and Collaborations*, Lecture Notes in Mobility, 26. Springer International Publishing, Cham, pp. 149–171.
- Kroesen, M., Handy, S., Chorus, C., 2017. Do attitudes cause behavior or vice versa? An alternative conceptualization of the attitude-behavior relationship in travel behavior modeling. *Transport. Res. Part A: Policy Pract.* 101, 190–202.
- Lai, Y., Kontokosta, C.E., 2018. Quantifying place: Analyzing the drivers of pedestrian activity in dense urban environments. *Landsc. Urban Plan.* 180, 166–178.
- Mayor of London, 2018. "Mayor's Transport Strategy", available from: <https://tfl.gov.uk/corporate/about-tfl/the-mayors-transport-strategy>.
- Mehta, V., 2013. *The Street: A quintessential social public space*. Routledge, London, New York.
- Mehta, V., Bosson, J.K., 2018. Revisiting Lively Streets: Social Interactions in Public Space. *J. Plan. Educ. Res.* 14 (1), 0739456X1878145.
- Mueller, N., Rojas-Rueda, D., Cole-Hunter, T., Nazelle, A. de, Dons, E., Gerike, R., Götschi, T., Int Panis, L., Kahlmeier, S., Nieuwenhuijsen, M., 2015. Health impact assessment of active transportation: A systematic review. *Prevent. Med.* 76, 103–114.
- Mueller, N., Rojas-Rueda, D., Salmon, M., Martinez, D., Ambros, A., Brand, C., Nazelle, A. de, Dons, E., Gaupp-Berghausen, M., Gerike, R., Götschi, T., Iacorossi, F., Int Panis, L., Kahlmeier, S., Raser, E., Nieuwenhuijsen, M., 2018. Health impact assessment of cycling network expansions in European cities. *Prevent. Med.* 109, 62–70.
- Schröter, B., Koszowski, C., Schmotz, M., Weber, J., Wittwer, R. and Gerike, R. (submitted), "Guidance and Practice in Planning Cycling Facilities in Europe—An Overview", Paper submitted for ISHGD 2020.
- Wittwer, R., Gerike, R., Hubrich, S., 2019. Peak-Car Phenomenon Revisited for Urban Areas: Microdata Analysis of Household Travel Surveys from Five European Capital Cities. *Transport. Res. Rec.* 2673 (3), 686–699.

Further Reading

- Buehler, R., Pucher, J., Gerike, R., Götschi, T., 2017. Reducing car dependence in the heart of Europe: lessons from Germany, Austria, and Switzerland. *Transport Rev.* 37 (1), 4–28.
- Ewing, R., Cervero, R., 2010. Travel and the built environment. *J. Am. Plan. Assoc.* 76 (3), 265–294.
- Ewing, R., Hajrasouliha, A., Neckerman, K.M., Purciel-Hill, M., Greene, W., 2016. Streetscape features related to pedestrian activity. *J. Plan. Educ. Res.* 36 (1), 5–15.
- Ewing, R., Handy, S., 2009. Measuring the unmeasurable: urban design qualities related to walkability. *J. Urban Design* 14 (1), 65–84.
- Gehl, J., 2010. *Cities for people*. Island Press, Washington, Covelo, London.
- Gehl, J., Svarre, B., 2013. *How to Study Public Life*. Island Press, Washington, DC.
- Gerike, R., Nazelle, A. de, Wittwer, R., Parkin, J., 2019. Special Issue Walking and Cycling for better Transport, Health and the Environment. *Transport. Res. Part A: Policy Pract.* 123, 1–6.
- Gerike, R., Parkin, J., 2016. *Cycling futures: From research into practice*. Routledge, London, New York.
- Götschi, T., Nazelle, A. de, Brand, C., Gerike, R., 2017. Towards a comprehensive conceptual framework of active travel behavior: a review and synthesis of published frameworks. *Curr. Environ. Health Rep.* 4 (3), 286–295.
- Handy, S., van Wee, B., Kroesen, M., 2014. Promoting Cycling for Transport: Research Needs and Challenges. *Transport Rev.* 34 (1), 4–24.
- Kroesen, M., Handy, S., Chorus, C., 2017. Do attitudes cause behavior or vice versa? An alternative conceptualization of the attitude-behavior relationship in travel behavior modeling. *Transport. Res. Part A: Policy Pract.* 101, 190–202.
- Mehta, V., 2013. *The Street: A quintessential social public space*. Routledge, London, New York.
- Mehta, V., Bosson, J.K., 2018. Revisiting lively streets: social interactions in public space. *J. Plan. Educ. Res.* 14 (1), 0739456X1878145.
- Mueller, N., Rojas-Rueda, D., Salmon, M., Martinez, D., Ambros, A., Brand, C., Nazelle, A. de, Dons, E., Gaupp-Berghausen, M., Gerike, R., Götschi, T., Iacorossi, F., Int Panis, L., Kahlmeier, S., Raser, E., Nieuwenhuijsen, M., 2018. Health impact assessment of cycling network expansions in European cities. *Prevent. Med.* 109, 62–70.

Electric Vehicles

Christine Eisenmann*, **Daniel Görge†**, **Thomas Franke‡**, *German Aerospace Center, Institute of Transport Research, Berlin, Germany; †University of Kaiserslautern, Electromobility Research Group, Kaiserslautern, Germany; ‡Universität zu Lübeck, Institute for Multimedia and Interactive System, Engineering Psychology and Cognitive Ergonomics, Lübeck, Germany

© 2021 Elsevier Ltd. All rights reserved.

Electric Vehicle Concepts	147
Road Electric Vehicle and Charging Technology	147
Definitions	147
Architectures	147
Specifications	148
Charging	148
Assessment of the Electric Vehicle Concept Compared to the Conventional Vehicle Concept	149
Activity Space Effects of Electrified Mobility Tools	151
User Experience and Behavior in Electric Vehicle Usage	152
Summary and Conclusions	153
See Also	153
References	153

Electric Vehicle Concepts

Electric vehicles (EV) are vehicles with at least one electric motor and an electric energy supply system for the purpose of vehicle propulsion. The electric energy supply system can be an on-vehicle electric storage system (e.g., a battery) or an off-vehicle electric power system (e.g., a power grid). Another subset of EVs is vehicles with an additional on-vehicle electric generation system (e.g., fuel cell or solar cell). A further subset of EVs is hybrid electric vehicles (HEV), which contain an additional energy converter (e.g., a combustion engine) and energy supply system (e.g., fuel) for vehicle propulsion.

EVs can be land vehicles (road, rail, offroad, and indoor), water vehicles, air vehicles, and space vehicles. **Figure 1** gives an overview over exemplary electric driven vehicle concepts.

EVs can have various use applications (i.e., such as vehicles with combustion engines). EVs can be used for private purposes (e.g., leisure mobility, commuting) or commercial purposes (e.g., logistics, goods delivery, rental services).

Passenger cars are the most important road vehicle concept in terms of global sales, added value, and global vehicle mileage (and consequently in terms of environmental externalities). Hence, the passenger car market is a major (future) market for EVs. The majority of the passenger car fleet is used for private purposes and owned by private persons. For example, in Germany about 70% of the registered motor vehicle fleet are private cars (DIW Berlin & DLR, 2019). Further applications of passenger cars are rental car fleets, commercial transport (e.g., delivery transport, service transport), company cars, new mobility concepts (e.g., car sharing and ride-sourcing fleets). Given the large field of EV concepts and applications the present article focuses on prevalent EV concepts: EVs for road use having one or more electric motors and a battery, which is charged from the power grid.

Road Electric Vehicle and Charging Technology

Definitions

The subsequent **Table 1** gives an overview of various terms used in the context of EVs. Similar definitions are given, for example, in SAE J1715.

Architectures

An EV can have various architectures. Four typical architectures for two-track vehicles are shown in **Figure 2**.

Architecture I results from the conversion of a conventional vehicle (CV) to an EV either by the manufacturer or a workshop as retrofit. The combustion engine and the fuel tank are replaced by an electric motor and a battery. The clutch and the shift gearbox, which are essentially not required in an EV, remain. Architecture II where the clutch is removed and the shift gearbox is replaced by a fixed gearbox is used in most production vehicles. Architecture III where most transmission components are removed and additional functionalities are enabled such as a distribution of torque to individual wheels for improving vehicle dynamics is used in few production vehicles. Architecture IV where all transmission components are removed and further functionalities are provided, such as a perpendicular alignment of the wheels for parallel parking exists in prototype vehicles. Transmission components induce inertial and friction losses,

Road	Rail	Water
e.g., car van bicycle scooter light vehicle motorbike bus lorry semitruck	e.g., train tram subway cable car	e.g., ferry, trolley boat submarine
	Air	Other
	e.g., helicopter quadrocopter airplane zeppelin	e.g., construction vehicle off-road vehicle space rover

Figure 1 Exemplarily list of vehicle concepts with electric drives.

Table 1 Terms, definitions, and abbreviations used for road electric vehicles

	<i>Label</i>	<i>Description</i>
CV	Conventional vehicle	Vehicle having a combustion engine for vehicle propulsion
EV	Electric vehicle	Vehicle having at least one electric motor and an electric energy supply system for vehicle propulsion (hypernym)
BEV	Battery electric vehicle	EV having a battery as on-vehicle electric storage system
FCEV	Fuel cell electric vehicle	EV having a fuel cell as on-vehicle electric generation system
HEV	Hybrid electric vehicle	EV having another energy converter (e.g., combustion engine) and energy supply system (e.g., fuel) for vehicle propulsion (hypernym)
Micro HEV	Micro hybrid electric vehicle	HEV having an electric storage system which has a small capacity and voltage no permitting electric driving
Mild HEV	Mild hybrid electric vehicle	HEV having an electric storage system which has a medium capacity and voltage permitting electric driving over short distances at low speed
Full HEV	Mild hybrid electric vehicle	HEV having an electric storage system which has a large capacity and voltage permitting electric driving over short distances at medium speed
PHEV	Plug-in hybrid electric vehicle	HEV having an electric energy storage system which has a large capacity and voltage and can be charged from the grid permitting electric driving over medium distances at high speed
REX/ EREV	Range-extender hybrid electric vehicle/ extended range electric vehicle	HEV having an electric energy storage system which has a large capacity for electric driving over large distances

increase production and maintenance costs, raise the weight, require installation space, prevent functionalities and are therefore not desirable. From this perspective the Architectures III and IV are favorable due to fostering new vehicle layouts and functionalities and may thus be increasingly used in the future, particularly in new EV constructions, which are not derived from CV constructions.

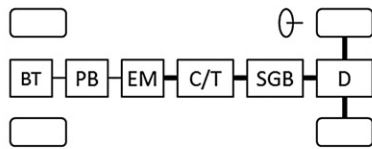
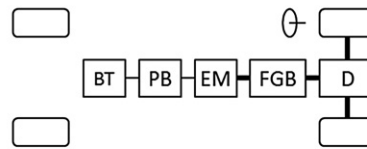
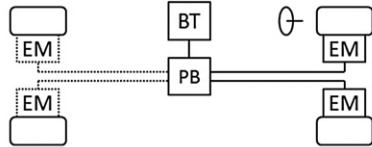
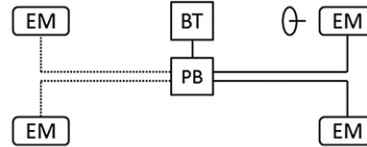
Specifications

EVs can be found in essentially all road vehicle classes with a wide spectrum of driving ranges, battery capacities, energy demands, motor powers and maximum speeds. An overview of the typical specifications of EVs in various road vehicle classes is given in **Table 2**. Note that the range corresponds to the distance, which can be covered with a fully charged battery.

For small and medium vehicles (scooter, bicycles, and motorcycles) the specifications of the electric version and the conventional version are very similar. Partly, the specifications of the electric version are even superior, for example, the range of an electric bicycle in comparison to a conventional bicycle. For medium and large vehicles (vans, trucks, and buses) the specifications of the electric version and the conventional version are very diverse. Particularly, the ranges are significantly smaller for the electric variant, permitting only urban and regional usage.

Charging

Battery EVs (BEVs) can be charged in two ways. The battery is operating with direct current (DC) while the power grid is providing alternating current (AC). Therefore, a charger that converts from DC to AC is required. The charger can be installed in the vehicle or in

I - One Motor, Complex Transmission**II - One Motor, Simple Transmission****III - Two/Four Motors, Simple Transmission****IV - Two/Four Wheel Hub Motors, No Tr.****Definitions**

- BT Battery
 PB Power Link
 EM Electric Motor
 C/T Clutch/Torque Conv.
 SGB Shift Gearbox
 FGB Fixed Gearbox
 D Differential Gearbox
 — Mechanical Coupling
 — Electrical Coupling

Figure 2 Architectures of electric road vehicles.**Table 2** Typical specifications of electric road vehicles (as of end 2019)

Class	Electric scooters	Electric bicycles	Electric motorcycles	Electric cars	Electric vans (up to 3.5 t)	Electric trucks (up to 26 t)	Electric busses (Urban)
Range	24–40 km	30–150 km	50–250 km	100–600 km	80–250 km	50–500 km	100–200 km
Battery capacity	250–750 Wh	250–750 Wh	1.5–12.5 kWh	15–100 kWh	20–80 kWh	60–600 kWh	150–450 kWh
Energy demand	0.5–2 kWh/100 km	0.5–2 kWh/100 km	2–8 kWh/100 km	12–23 kWh/100 km	12–22 kWh/100 km	80–120 kWh/100 km	n/a
Motor power	250–750 W	250–500 W	2.5–100 kW	20–450 kW	45–90 kW	90–450 kW	250 kW
Maximum speed	20–25 km/h	25–45 km/h	45–250 km/h	100–250 km/h	85–130 km/h	80 km/h	85 km/h

the charging station. For low-charging power (up to 22 kW, normal charging) the charger is installed in the vehicle and charging is performed with AC. For high-charging power (above 22 kW, fast charging) the charger is installed in the charging station due to its high weight and cost and charging is performed with DC. Electric cars can support AC charging, DC charging or both. For AC charging a standard connector (e.g., Schuko, NEMA) can generally be used. For safety and security, however, specific connectors which permit a communication between the charging station and the EV to control power flow and prevent unauthorized disconnection are utilized for both AC and DC charging. Other connectors are used in separate regions. Relevant standards are IEC 62196 and SAE J1772. The most common connectors and their specifications are summarized in [Table 3](#). Note that these specifications apply for typical implementations. Note further that for small vehicles (scooters, bicycles, and motorcycles) usually Schuko or NEMA connectors are preferred while for large vehicles (trucks and buses) partly proprietary connectors are used. For all other vehicles Type 1, Type 2, CCS, and CHAdeMO connectors have become standard.

For access and payment various concepts exist, for example, RFID or SMS for access and energy- or time-based tariffs for payment. Various initiatives work towards unified and accessible access and payment solutions.

Uncontrolled charging can lead to a significant load in the power grid, particularly in residential areas and at peak times. Controlled charging can substantially reduce the load in the power grid. The basic idea consists in adjusting the charging power and in turn the charging time such that grid limits are respected and at the same time the mobility demands are fulfilled. For controlled charging standards like ISO 15118 and OCCP 1.6/2.0 have been established. Additionally controlled charging can be used for buffering renewable energy (e.g., solar energy). The EV is charged when renewable energy is available (grid to vehicle or G2V) and discharged when renewable energy is lacking (vehicle to grid or V2G).

Assessment of the Electric Vehicle Concept Compared to the Conventional Vehicle Concept

Early versions of the electric motor were already used in the late 1820s to power tiny cars and carriages with nonrechargeable primary cells, while the first practical implementations came to life in the late 19th century ([Guarnieri, 2012](#)). In the early 1900s, EVs became

Table 3 Overview of charging connectors

Connector	Schuko (With ICCB ^a)	Type 1	Type 2	Combined charging system (CCS)	CHaDeMO
Current (Type)	AC	AC/DC	AC/DC	AC/DC	DC
Voltage/Current (Maximum)	230 V/16 A	230 V/32 A	400 V/32 A	850 V/200 A	500 V/50 A
Power (Maximum)	2.3 kW	7.4 kW	22 kW	170 kW	50 kW
Phases	one	one	one/three		
Charging Time (for 40 kWh)	17.4 h	5.4 h	1.8 h	11.3 min ^b	38.4 min ^b
Region (predominant)	Europe	US, Japan	Europe	Europe	Japan

^aIn cable control box^bTo 80% for battery protection

the primary choice of local motorized transportation. In the 1930s vehicles with gasoline and diesel engine became more and more popular; they managed to overtake the EV leadership through mass production at a reasonable cost and improved performance. Further reasons for a halt in development of EVs were that the need for intercity travel grew and that charging infrastructure and reliable electricity transmission were lacking. It was not until the late 1990s that multiple BEV prototypes were developed and another decade until HEVs were engineered (Situ, 2009). Since then the development has engendered multiple variations of EVs which are being mass-produced and sold today. Their success is not as forthcoming as it was for the EVs in the early 1900s. The share of BEV and HEV in the car fleet in 2020 is still in the single-digit percentage range in most countries. Hence, a market uptake of EVs is required and promoted by various bodies since EVs are seen as one component towards sustainable transportation and they have advantages over CVs. The reasons are manifold:

First, BEVs do not produce greenhouse gases, like CO₂, in the use phase. For an overall assessment of climate impacts of EVs compared to CVs the provision of electricity and other sources of emissions, such as vehicle production and vehicle scrapping, also need to be included. Research shows that the net sustainability effect of EVs is dependent on many factors, such as power generation. The higher the share of renewable energies in electricity generation, the more advantageous the environmental assessment of EVs is (see e.g., Wirges 2016, Stark et. al., 2018).

Second, BEVs show lower noise emissions at low speeds compared to CVs. With low speeds, up to 25 km/h, engine noises are a car's dominant source of noise (e.g., in residential areas, at intersections with traffic lights). In these speed ranges, EVs are more advantageous than CVs as their engine is less noisy. The noise levels of EVs and CVs are similar when driving faster than 25 km/h as the main noise sources at higher velocities are the interaction of tires and road surfaces and aerodynamic noises. However, lower noise emissions might cause negative safety externalities, as it might impair pedestrian safety that is why artificial noise generation are yet increasingly used.

Third, BEVs drive without local exhaust emissions that are hazardous to the environment and health, such as nitrogen oxides and particulate matter. However, EVs like vehicles with other propulsion systems produce emissions by tire abrasion and brakes. However, vehicle production generates local exhaust emissions, both for EVs and CVs. A significant of particular matter is in particular generated in the steel production. As the sources of direct emissions are often taken place in remote areas, those emissions are less relevant to the health of the greater part of the population (BMU 2019).

Fourth, the EV technology is energy efficient. The propulsion system (electric motor) and the transmission system (fixed gearbox) are efficient, moreover idling consumptions (e.g., at traffic lights) are not present.

Fifth, EVs can serve as mobile energy storage in the power grid (i.e., V2G systems) and may therefore provide power grid services. Here, EVs can act either as controllable loads that contribute to leveling the off-peak load by throttling their charging rate (e.g., at night) or they can act as storage devices, for example, when being parked during the day (Guille & Gross 2009).

EVs also have objective disadvantages in comparison to CVs. Those are discussed in the further paragraphs.

First, BEVs are more restricted in their possibilities of car use variation due to shorter driving ranges and longer charging durations compared to CVs.

Second, the purchasing costs of EVs are, due to the current high-battery production costs, in most countries higher than the costs of CVs. However, the recharging respectively refueling costs and maintenance costs of EVs are on average lower than the maintenance costs of CVs as EVs have less and simpler components, less fluids, breaking pads, etc. Hence, if all costs of vehicle usage over the lifespan of the vehicle are considered, EV costs are comparable with CV costs.

Third, an obstacle for the market uptake of EVs is that in the market ramp up phase of a technology, the product is increasingly improving (e.g., in terms of driving range) and also the costs (i.e., production costs and purchase costs) of the product are decreasing with time. In addition, the variation of products available on the market expands gradually over time. Hence, motor vehicles, such as passenger cars are long lasting goods which are not bought on the new car market by many households but on the used car market. Hence in the phase of market uptake, the supply of EVs on the used vehicle market is not sufficient for many potential customers.

Fourth, rare-earth elements (neodymium and dysprosium) and/or critical metals (lithium and cobalt) are needed for the production of the electric motors and batteries, which are predominantly used. With an increased demand in EVs the demand for these rare-earth elements and critical metals will rise as well, and consequently geopolitical issues might impact their availability.

Extensive resource extraction methods (e.g., mining) might lead to air and groundwater contamination in countries with a high-resource density (e.g., China, Chile, Argentina, Bolivia, Congo). Moreover, the resource extraction is emission-intensive. This leads to a global relocation of the EV lifetime emissions to the detriment of countries with a high-resource density (Romare & Dahllöf 2017).

Due to the advantages of the electric motor over the combustion engine, it is the goal of policy makers in various countries to increase the share of EVs in the vehicle stock. In order to support EV diffusion various transport policy incentives can be set. Examples are infrastructure and organizational measures concerning traffic flow (e.g., the permission for EVs to use dedicated bus lanes or high occupancy lanes), infrastructure and organizational measures for stationary traffic (e.g., exemptions from parking fees) and financial incentives on EV costs (e.g., buyer's premium, tax rebate) (Stark et al. 2014).

Activity Space Effects of Electrified Mobility Tools

The electric motor is deployed in two ways for road vehicle concepts: either the electric motor substitutes another motor in the vehicle concepts or vehicle concepts that were originally not motorized are motorized by an electric motor. Both ways have disparate implications to the usage behavior and activity spaces of the users electrify vehicle concepts.

Examples for vehicle concepts with motor substitution are motor vehicles, such as passenger cars, motorcycles, vans, busses and lorries. Those vehicle concepts have, when equipped with a combustion engine, universal possibilities of application. For example, the range of applications includes, but is not limited to, daily commutes, long distance journeys for the summer holiday, shopping, and leisure purposes. The wide range of applications leads to the passenger car not being used the same way every day, but very differently over longer periods of time. If a passenger car has an electric motor instead of a combustion engine, the range of applications might be limited due to range limitations and longer charging durations. However, it depends on the variations in use and applications of the passenger car, the range of the passenger car and the availability and configuration of charging infrastructure whether and to what extent the users are restricted in their travel behavior by the use of EVs compared to CVs.

In terms of passenger transport, diverse population groups have varying and distinct travel needs and activity spaces, which might also differ by transport mode. For example, some elderly might not have the confidence to drive longer distances by passenger car. This also determines whether and under what conditions EVs fit their car usage patterns and mobility needs.

Commercial transport is rather heterogeneous in terms of branches, road vehicle types used and usage patterns. The vehicles used in certain branches, for example, urban distribution and social services, have uniform usage patterns in the longitudinal perspective and are mainly used for short ranges. EVs can be a practicable alternative for these usage patterns. Other branches, for example, logistics, use their vehicles often for longer distances and to transport heavy goods (Frenzel, 2015). Here, BEVs are not applicable. A possible option may be electric trucks powered with overhead lines which are still under investigation.

The electrification of road vehicles for passenger transport, for example, the use of electric busses for public transport or of electric cars for car sharing and ride sourcing, is also an option. A challenge in the deployment planning is that additional charging times and the availability of charging infrastructure must be taken into account (Kley et al., 2011).

Bicycles and scooters are examples of vehicle concepts that were originally not motorized but are now available in configurations with electric motors, for example, pedelecs or e-scooters. Those EV concepts are faster and can be used for longer distances and for the transport of heavier goods than their nonelectric pendants, which are only propelled by muscle power. Hence, they offer new fields of applications, for example, greater trip distances, applications in hilly areas, an interesting vehicle alternative for person groups which do not have the physical condition to use a bicycle over longer distances (e.g., elderly). In commercial transport, cargo pedelecs are suitable to specific branches, for example, delivery services, inner city cargo.

In travel behavior research mobility dynamics, mobility needs and behavior of persons and their vehicle usage are in focus. Research deals, for example, with the market potential of EVs, the substitution potential of CVs by EVs. In order to analyze those issues various methodological approaches are used. A short description of some research methods is given below.

- Stated preference surveys and stated response surveys: survey participants are asked to make decisions in fictive situations, for example, the purchase of a BEV. In order to also account for restrictions that result from the short range of the of EVs, those surveys are combined with other methods, such as vehicle-use diaries. A downside of this approach is that survey participants have often no in-depth experience and knowledge on surveyed topics and products; consequently, they might behave differently in real life situations than indicated in the survey (e.g., Golob, 1990).
- Field trials: EVs are made available to the study participants over a period of several weeks or months and the use of these vehicles is examined in a real life context. A major advantage of these studies is that the use of EVs can be analyzed over longer periods of time. Though, due to the high costs of field trials sample sizes are often small, which might limit the universality and representativeness of the results (e.g., Cocron et al., 2011)
- Cross-sectional travel surveys and data: Cross-sectional travel surveys conducted over one day (e.g., national household travel surveys, available for many countries on a regular basis), are analyzed. The assumption underlying the analysis is that vehicles not used over long distances could be replaced by EVs from a vehicle usage perspective, as EVs do not limit the mobility needs of vehicle users. A drawback of this approach is that the variability of vehicle use is not considered, which leads to an overestimation of the substitution potential by EVs (e.g., Aultman-Hall et al., 2012)

- Car usage surveys and car usage models of CVs over longer periods of time: The underlying assumption for the analysis of longitudinal data is the same as for cross-sectional data; that is, CVs that are not or seldom used over long distances could be replaced by EVs from a usage perspective. However, the variability of vehicle use over longer periods of time is taken into account. For vehicle usage surveys, passive forms of data generation, such as Global Positioning System (GPS) are often used. Another approach to generate vehicle usage information over longer periods of time is the development of vehicle usage models over longer periods of time. Available vehicle usage surveys and other travel surveys serve as input data for those models (e.g., Eisenmann & Plötz 2019).

User Experience and Behavior in Electric Vehicle Usage

EVs come with a number of features that lead to specific, unique dynamics in user experience and behavior. These include the user interaction with range and recharging, the implications of electric drivetrain characteristics for user-system interaction (e.g., bidirectional energy flows and further energy-related dynamics), and the broader implications of EV characteristics for dynamics in user acceptance. The following section focuses particularly on user interaction with BEVs as the most prototypic EV type. Yet, many aspects also transfer to other EV powertrain types that are affected by the dynamics of charging (e.g., PHEVs) or by the partial electric powertrain.

First and foremost, while possibly being a transient phenomenon within the process of mass-market adoption and EV technology evolution, the phenomenon of range anxiety has been commonly discussed as a larger issue of BEVs. Since its first mentioning in scientific literature (Tate et al., 2008) range anxiety, commonly referred to as the fear of becoming stranded with a BEV, has become a widely used term. Yet, empirical user research with actual BEV drivers and psychological theorizing suggests, that the constructs of fear or anxiety do not accurately match the actual range-related user experience in everyday interaction with BEVs. Indeed user experience in situations where the range might not be sufficient to satisfy driver needs is better characterized by constructs such as range discomfort, range workload, or range stress (e.g., Rauh et al., 2015; Franke et al., 2016). Moreover, in everyday usage, drivers actively avoid range stress using available coping strategies like additional recharging, route (re-)planning, or more energy-efficient driving (i.e., eco-driving). In general, drivers thus tend to plan for substantial range safety buffers and do not completely utilize the potentially available range. Hence, psychological theorizing has suggested the concept of three psychological range levels (Franke et al., 2015) which drivers use to manage limited range resources. That is, every limited-range vehicle is characterized by its technical range, meaning the distance it can drive on one charge given certain objectively defined conditions (e.g., driving cycle conditions such as the worldwide harmonized light vehicle test procedure, WLTP). Yet, many psychological factors drive interaction with range. First, competent range reflects the range that users can achieve based on their range management skills (e.g., eco-driving knowledge) and/or based on how easy the vehicle or the environment makes range optimization (e.g., system support for drivers' eco-driving skill acquisition). Second, performant range reflects the range that users usually obtain based on their everyday driving behavior (i.e., eco-driving motivation and habits) and how good the vehicle and environment support and simplify eco-driving (e.g., everyday traffic-related driving conditions). Third, comfortable range is the actual usable range that users utilize with an optimal user experience (i.e., without range stress or range discomfort) based on their perceived coping resources (e.g., available emergency charging opportunities, opportunities and personal skills for extending range) and the perceived accuracy of available remaining range information (e.g., range prediction system). That is, drivers can be expected to develop a preferred range safety buffer over usage, based on their experience with range dynamics (e.g., perceived trustworthiness of range interfaces) and based on personal variables (e.g., technology-related self-efficacy beliefs). Hence, the primary criterion for user-centered optimization of BEV mobility resources is not to maximize technical range but to optimize psychological range.

Structurally similar range management and range experiences dynamics are taking place in drivers of conventional CVs (Philipsen et al., 2019) yet range stress can be avoided much more easily here. Improvements in the total available range and charging speed will obviously limit the likelihood of range stress experiences if users do not heavily change their recharging routines (i.e., negative behavioral adaptation). In any case, designing the range-related users experience (e.g., via optimized user interfaces) is a key task in BEV development.

Charging behavior is closely connected to user interaction with range. Research indicates here that, once drivers have a certain level of abundant charging opportunities at hand (i.e., how many convenient charging opportunities do drivers typically pass before they actually have to recharge to avoid falling below the preferred range safety buffer), they tend to develop diverse charging styles (Franke & Krems, 2013). That is, while some drivers rather tend to recharge whenever possible, other drivers rather tend to recharge only if needed (i.e., based on available remaining range vs expected range demands). Preferred or habituated charging behavior patterns also interact with the issues around the user-centered design of intelligent charging algorithms that aim to increase the efficiency of integrating renewable energy sources into the grid. That is, drivers who charge whenever possible give energy control entities more flexibility to optimize charging processes while drivers who only charge when needed may need a faster recharge providing less time for energy-supply-related charging optimization algorithms.

EVs also come with specific energy dynamics based on the consumption characteristics of electric powertrains and because of the bidirectional energy flow induced by regenerative braking. That is, variations in driver behavior tend to have a particularly pronounced effect on the energy efficiency (and therefore range) achieved in everyday driving. This is the case

for HEVs that have highly complex energy flow dynamics that are difficult to understand for drivers (i.e., resulting in wrong mental models and suboptimal eco-driving strategies in spite of sufficient eco-driving motivation). And it is also the case for BEVs that have relatively simple energy dynamics compared to CVs given their typical single-gear design, yet are still susceptible for wrong mental models because novice BEV drivers tend to neglect the energy conversion losses inherent in regenerative braking thus leading to a biased evaluation of the effectiveness of certain eco-driving strategies (i.e., too positive perception of regenerative braking).

All those specific features of EVs can also have implications for aspects of EV user acceptance. For example, with regard to BEV range preferences, drivers tend to prefer vehicle configuration with much more range than what is needed on an everyday basis. That is, many drivers tend to select range based on more seldom higher range needs unless the BEV is explicitly framed as secondary vehicle, city vehicle or as vehicle for the daily commute which is usually much shorter than the range typical BEVs offer. Further, range preferences can change with practical driving experience leading to dissociations between stated range preferences before vehicle purchase and range satisfaction after more extensive experience.

Summary and Conclusions

This article discusses the role of EVs as a mode of transport. EVs for road use having one or more electric motors and a battery which is charged from the power grid were put into focus. First, the technical aspects of the EV concept were described, such as the various terms and definitions used in the context of EVs, EV architectures and specifications as well as EV charging concepts. Then, the EV concepts were assessed compared to the CV concept and advantages and disadvantages were discussed. Moreover, the activity space effects of electrified mobility tools were described, that is, the fact and subsequent implications that the electric motor either substitutes another motor in the vehicle concepts or vehicle concepts that were originally not motorized are motorized by an electric motor. Subsequently, user experience and behavior in EV usage were briefly discussed, that is, user interaction with range and recharging, the implications of electric drivetrain characteristics for user-system interaction, and the dynamics in users' range preferences.

Given the highly dynamic evolution of electric vehicles, the present article can only provide a snapshot of the currently common technical characteristics, role in the transport system, and usage patterns of electric vehicles. Despite, it seems highly probable that electric vehicles will play a key role in our future transport system.

See Also

Climate Change and Transport Policy; The Impact of Electric Vehicles on Energy Systems; Noise Pollution, from Transport; Transport Noise and Health; E-mobility Device Safety; Road Safety; Electric Vehicles and Health; Fuel Efficiency Rules; The Taxation of Car Use in the Future; Generalized Cost for Transport; Environmental Sustainability in Freight Transportation; Decarbonizing Road Freight Transport; Car Sharing; Car Sharing; Electric Bikes and Health; Bicycles for Urban Freight; Habitual Behavior; Driving Behavior & Skills; Behavioral Adaptation; Behavioral Change; Electromobility: Attitudes and Perceptions

References

- Aultman-Hall, L., Sears, J., Dowds, J., Hines, P., 2012. Travel demand and charging capacity for electric vehicles in rural states: Vermont case study. *Transp. Res. Rec.* 2287, 27–36.
- BMU 2019. *Wie umweltfreundlich sind Elektroautos? Eine ganzheitliche Bilanz*. Berlin: German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety.
- Cocron, P., Bühler, F., Neumann, I., Franke, T., Krems, J.F., Schwalm, M., Keinath, A., 2011. Methods of evaluating electric vehicles from a user's perspective - The MINI E field trial in Berlin. *IET Intel. Transp. Sys.* 5, 127–133.
- DIW Berlin & DLR 2019. *Verkehr in Zahlen*, in: *Infrastructure*, German Federal Ministry of Transport, Digital, Infrastructure, (Ed.). Flensburg.
- Eisenmann, C., Plötz, P., 2019. Two methods of estimating long-distance driving to understand range restrictions on EV use. *Transp. Res. Part D* 74, 294–305.
- Franke, T., Günther, M., Trantow, M., Rauh, N., Krems, J.F., 2015. Range comfort zone of electric vehicle users – concept and assessment. *IET Intel. Transp. Sys.* 9, 740–745.
- Franke, T., Krems, J.F., 2013. Understanding charging behaviour of electric vehicle users. *Transp. Res. Part F* 21, 75–89.
- Franke, T., Rauh, N., Günther, M., Trantow, M., Krems, J.F., 2016. Which factors can protect against range stress in everyday usage of battery electric vehicles? toward enhancing sustainability of electric mobility systems. *Human Fact.* 58, 13–26.
- Frenzel, I., 2015. The commercial early adopters of electric vehicles in Germany – a description of their trip patterns. 2nd Interdisciplinary Conference on Production Logistics and Traffic 2015, Dortmund.
- Golob, T.F., 1990. The dynamics of household travel time expenditures and car ownership decisions. *Transp. Res. Part A* 24, 443–463.
- Guarnieri, M., 2012. Looking back to electric cars. Third IEEE HISTory of Electro-technology Conference (Histelcon), Pavia.
- Guille, C., Gross, G., 2009. A conceptual framework for the vehicle-to-grid (V2G) implementation. *Ener. Policy* 37, 4379–4390.
- Kley, F., Lerch, C., Dallinger, D., 2011. New business models for electric cars—A holistic approach. *Ener. Policy* 39, 3392–3403.
- Philipsen, R., Brell, T., Biermann, H., Ziefle, M., 2019. Under pressure—users' perception of range stress in the context of charging and traditional refueling. *World Elec. Veh. J.* 10, 50.
- Rauh, N., Franke, T., Krems, J.F., 2015. Understanding the impact of electric vehicle driving experience on range anxiety. *Human Fact.* 57, 177–187.
- Romare, M. & Dahlöf, L. 2017. *The Life Cycle Energy Consumption and Greenhouse Gas Emissions from Lithium-Ion Batteries*. Stockholm.
- Situ, L., 2009. Electric Vehicle development: The past, present & future. 2009 3rd International Conference on Power Electronics Systems and Applications (PESA), Hong Kong.

- Stark, J., Klementsitz, R., Link, C., Weiss, C., Chlond, B., Franke, T. & Günther, M., 2014. Future Scenarios of electric vehicles with range extender in Austria, Germany and France. Transport Research Arena, Paris.
- Stark, J., Weiss, C., Trigui, R., Franke, T., Baumann, M., Jochem, P., et al., 2018. Electric vehicles with range extenders: Evaluating the contribution to the sustainable development of metropolitan regions. *J. Urban Plan. Dev.* 144, 04017023.
- Tate, E.D., Harpster, M.O., Savagian, P.J., 2008. The electrification of the automobile: From conventional hybrid, to plug-in hybrids, to extended-range electric vehicles. *SAE Int. J. Passeng. Cars – Electron. Electr. Syst.* 1, 156–166.
- Wirges, J., 2016. Planning the Charging Infrastructure for Electric Vehicles in Cities and Regions. Ph.D. thesis.

Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service

Susan Shaheen PhD, Adam Cohen, McLaughlin Hall, University of California, Berkeley, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction: What is Shared Mobility?	155
Business Models of Shared Mobility	155
Role of Smartphone Apps as an Enabler	156
Growth of Mobility on Demand (MOD) and Mobility as a Service (MaaS)	157
Summary and Conclusion	158
Acknowledgment	158
Biographies	158
See Also	158
References	159
Further Reading	159

Introduction: What is Shared Mobility?

Shared mobility—the shared use of a vehicle, bicycle, or other travel mode—is an innovative transportation strategy that enables users to have short-term access to a transportation mode on an as-needed basis. Shared mobility includes various modes and service models and transportation modes that meet the diverse needs of travelers. Shared mobility encompasses a variety of modes, such as bikesharing; carsharing; for-hire services (bikesharing; carsharing); courier network services (CNS); for-hire services [e.g., taxis and transportation network companies (TNCs), which are also known as ridesourcing and ridehailing]; microtransit; ridesharing (carpooling/vanpooling); scooter sharing; and urban air mobility (UAM) (Shaheen et al., 2016a; Cohen and Shaheen, 2016; Shaheen et al., 2017). Active and low-speed shared modes are sometimes collectively referred to as shared micromobility. Table 1 includes definitions of common and emerging shared modes. Common shared mobility service models include roundtrip services (vehicle, bicycle, or other mode is returned to its origin); one-way station-based services (vehicle, bicycle, or other mode is returned to a different designated station location); and one-way free-floating services (vehicle, bicycle, or other mode can be returned anywhere within a geographic area).

In recent years, shared mobility (both passenger and courier services) have grown rapidly due to technological advancements, changing consumer patterns (mobility and consumption), and shifting perspectives toward transportation and vehicle ownership. Shared mobility can help bridge gaps in the transportation network (e.g., facilitating first- and last-mile connections to public transit, offering late-night transportation services, and services scaled for lower-density communities). As such, shared mobility can help bridge gaps in the public transportation network by extending geographic coverage and the service times of transit services. Additionally, more modal choices and a larger pool of travelers can create a “network effect,” where mobility options in close proximity to one another add collective value. Additionally, a growing number of last-mile delivery and courier services are also changing consumer patterns and disrupting traveler behavior. Last-mile delivery services can serve as a substitute for personal trips to access goods and services as consumers take advantage of “just-in-time delivery.” As such, a household’s decision to have goods delivered rather than pick-up goods on another trip has the potential to contribute to notable changes in travel behavior. This article has four key sections. The first section reviews common business models. The second discusses the role of smartphone apps that enable shared mobility. Next, shared mobility’s relationship to mobility on demand (MOD) and mobility as a service (MaaS) is discussed. The final section provides a summary of the article.

Business Models of Shared Mobility

Shared mobility is enabled through four types of common business models, which are based on the relationship of the service provider and consumer. These common business models include: (1) business to consumer (B2C); (2) business to government (B2G); (3) business to business (B2B); and (4) peer-to-peer (P2P) (Shaheen et al., 2016a).

- **Business-to-Consumer (B2C):** In a B2C model, service providers offer individual consumers access to a business-owned operated transportation services such as a fleet of vehicles, bicycles, scooters, or other modes through memberships, subscriptions, user fees, or a combination of pricing models.
- **Business-to-Government (B2G):** In a B2G model, service providers offer transportation services to a public agency. Pricing may include a fee-for service contract, per-transaction basis, or some other pricing model.

Table 1 Definitions of common and emerging shared mobility and last mile delivery services

<i>Service</i>	<i>Definition</i>
Bikesharing (also known as micromobility)	Provides users with on-demand access to bicycles at a variety of pick-up and drop-off locations for one-way (point-to-point) or roundtrip travel. Bikesharing users access bicycles using one of three bikesharing models: (1) station-based bikesharing (users access bicycles via unattended stations); (2) dockless (users may access (unlock) a bicycle and park it at any location within a predefined geographic region); and (3) hybrid bikesharing systems (users may check out and return bicycles either through a station or non-station location). Bikesharing fleets are commonly deployed in a network within a metropolitan region, city, neighborhood, employment center, and/or university campus.
Carsharing	Individuals gain the benefits of private-vehicle use without the costs and responsibilities of ownership. Individuals typically access vehicles by joining an organization that maintains a fleet of cars and light trucks deployed in lots located within neighborhoods and at public transit stations, employment centers, and colleges and universities. Typically, the carsharing operator provides gasoline, parking, and maintenance. Generally, participants pay a fee each time they use a vehicle.
Courier Network Services (CNS)	Provides for-hire delivery services for monetary compensation via an online application or platform (such as a website or smartphone app) to connect couriers using their personal vehicles; bicycles; or scooters with freight (e.g., packages, food).
Drones [also known as Unmanned Aerial Vehicles (UAVs)]	A short-range unmanned aerial vehicle (or UAV) that can transport small packages, food, or other goods.
Microtransit	Privately or publicly operated, technology-enabled transit services that typically use multi-passenger/pooled shuttles or vans to provide on-demand or fixed-schedule services with either dynamic or fixed routing.
Ridesharing (also known as carpooling and vanpooling)	Facilitates formal or informal shared rides between drivers and passengers with similar origin-destination pairings.
Robotic Delivery	Offer short-range unmanned ground-based delivery of packages, food, or other goods using a small conveyance robot.
Scooter sharing (also known as micromobility)	Users gain the benefits of a private scooter without the costs and responsibilities of ownership. Individuals typically access scooters by joining an organization that maintains a fleet at various locations. The scooter operator usually provides gasoline, parking, and maintenance. Generally, participants pay a fee each time they use a scooter. Scooters can be accessed via unattended stations or accessed (unlocked) and returned (parked) to any location within a predefined geographic region. Scooter sharing includes two types of services: (1) Standing electric scooter sharing using shared scooters with a standing design with a handlebar, deck and wheels that is propelled by an electric motor. The most common scooters today are made of aluminium, titanium, and steel; and (2) Moped-style scooter sharing using shared scooters with a seated-design, either electric or gas powered, that generally having a less stringent licensing requirement than motorcycles designed to travel on public roads.
Taxis	Provide prearranged and on-demand vehicle services for compensation through a negotiated price, zone pricing, or a taximeter. Trips can be made by advance reservations (booked through a phone, website, or smartphone application), street hail (by raising a hand or standing at a taxi stand or specified loading zone), or e-Hail (dispatching a taxi driver using a smartphone application).
Transportation Network Companies (TNCs) (also known as ridesourcing and ridehailing)	Provides prearranged and on-demand transportation services for compensation, which connect drivers of personal vehicles with passengers. Smartphone mobile applications facilitate booking, ratings (for both drivers and passengers), and electronic payment. TNCs also includes “ridesplitting,” in which customers can split a ride and fare in a TNC vehicle (where available).
Urban Air Mobility	The safe and efficient system for air passenger and cargo transportation within an urban area, inclusive of small package delivery and other urban Unmanned Aerial Systems (UAS) services, which supports a mix of onboard/ground-piloted and increasingly autonomous operations.

Source: Adapted from: Shaheen et al., 2016a; Cohen and Shaheen 2016; Shaheen et al., 2018

- **Business-to-Business (B2B):** In a B2B model, service providers sell business customers access to transportation services either through a fee-for-service or usage fees. The service is typically offered to employees to complete work-related trips.
- **Peer-to-Peer (P2P):** In a P2P model, service providers broker mobility and courier transactions among car, bicycle, or other device owners and renters by providing the organizational resources needed to make the exchange possible (i.e., online platform, customer support, driver and motor vehicle safety certification, auto insurance, and technology).

These business models are often enabled through a variety of app-based services (e.g., smartphone apps) that provide single-mode access, multimodal access, trip planning, real-time information, and other traveler services.

Role of Smartphone Apps as an Enabler

While shared mobility can be employed independently of technology enablers, such as smartphone applications (or “mobility apps”), app-based services are common enablers of shared mobility services because they increase consumer convenience and facilitate real-time information (and in some cases vehicle or device access). As such, app-based services paired with other

technologies such as GPS, sensors, wireless systems, information technology, big data, data analytics and management systems, machine learning, artificial intelligence, and augmented reality can be key supporters of the shared mobility ecosystem (Shaheen et al., 2016b).

In particular, mobility apps can include a variety of mode specific applications as well as trip planners and multimodal aggregators. Fundamentally, the primary purpose of mobility apps is to assist users in planning or understanding their transportation choices, and they may increase access to a variety of modal options. Increasingly, some apps are using incentives to reduce congestion and encourage sustainable transportation choices. These apps employ gamification to incentivize more environmentally friendly travel modes. Gamification uses game design elements in a non-game context (e.g., badges, levels, leader boards, etc.) to encourage specific types of user behavior. The growing prevalence of smartphone apps is leading to the growth of connected travelers who are more attuned to real-time transportation information and increasingly dependent on real-time information services throughout their journey.

Growth of Mobility on Demand (MOD) and Mobility as a Service (MaaS)

The growth of shared mobility and enabling technologies is contributing to the commodification and aggregation of transportation services. In North America, consumers are commodifying transportation services by assigning economic values to modes and making mobility decisions (including the decision not to travel and instead have a good or service delivered) based on cost, travel and wait time, number of connections, convenience, and other attributes, a concept known as MOD (Shaheen and Cohen, 2019a, 2019b). MOD enables consumers to access mobility, goods, and services on demand by dispatching or using shared mobility, delivery services, and public transportation strategies through an integrated and connected multi-modal network. The most advanced forms of MOD passenger services incorporate trip planning and booking, real-time information, and fare payment into a single user interface. MOD promotes choice in personal mobility, leverages emerging and existing technologies and big data capabilities, encourages multimodal connectivity and system interoperability, and promotes new business models that enhance traveler experience. MOD has three major guiding principles: (1) traveler centric and consumer driven, (2) data connected and platform independent, and (3) multimodal and mode agnostic. In the United States, the US Department of Transportation and Federal Transit Administration are funding pilots and research as part of its MOD Sandbox program to explore partnerships; develop new business models; integrate public transit and MOD solutions; and investigate enabling technical capabilities such as: integrated payment systems, decision support, incentives for traveler choices, and supply and demand management of the transportation network. With MOD, there is a recognition that digital and goods delivery services can substitute for trips, while simultaneously creating demand for new and different trip types. The convergence of MOD with automation is enabling new opportunities for manned on-demand delivery options such as drones, robotic delivery, and automated delivery vehicles.

In Europe, services that allow travelers to sign up for mobility services in one package are gaining popularity. The concept of MaaS evolved approximately a decade ago with an initial pilot in Gothenburg, Sweden known as UbiGo. UbiGo repackaged existing transportation services (e.g., public transit, taxi, bikesharing, and carsharing) into a one-stop, monthly, paid subscription service ranging from \$185 to \$280 US per month for household (including children) subscriptions (Sochor et al., 2018). MaaS redistributes the mobility chain by integrating the products and services of mobility providers and supplying them to users as a single service. Typically, a digital platform creates and manages trips that users can pay for via a single account. A distinguishing feature of MaaS is giving users the option to purchase MaaS products, such as monthly subscription plans that best fit a user's or household's needs. These subscriptions can include a certain amount of each transportation service (e.g., public transportation, bikesharing, carsharing, taxis, etc.) and are similar to other service bundles, such as mobile phone plans, where the user pays one price for the combination of multiple service elements (e.g., talk, text, data, roaming, long distance, etc.). Brokering travel with suppliers, repackaging, and reselling it as a bundled package is a distinguishing characteristic of MaaS. Generally, there is an emerging consensus that MaaS is about integrating multiple transportation modes into a seamless user experience, often necessitating open data and cooperation by public and private transportation stakeholders.

In Europe, some practitioners maintain MOD is a type of MaaS. In the US, others claim that MaaS is one type of MOD, characterizing MOD as a "lower-level" of MaaS based on a five-level MaaS Integration Index. In contrast, the USDOT MOD concept of operations implies that MaaS is a lower-level of MOD because it is a comprehensive approach to transportation systems management, encompassing transportation systems management and operations (e.g., mobility supply-and-demand management) and goods delivery in addition to passenger mobility. With MOD, there is a recognition that every trip has a purpose, and goods delivery may be able to serve as a substitute for some passenger trips (Shaheen and Cohen, 2019a, 2019b).

Fig. 1 compares similarities and differences between MOD and MaaS, drawing upon USDOT's MOD concept of operations. While there are some similarities between the two, MOD focuses on the commodification of passenger mobility and goods delivery and transportation systems management, whereas MaaS primarily focuses on passenger mobility aggregation and subscription services. With MaaS, there is a strong emphasis on information and fare payment integration through a single digital platform. Digital information and fare payment integration, enabled through smartphone apps that are coupled with commodified mobility services, are making traveler options more convenient and seamless through MOD and MaaS. It is important to note that definitions and understanding of MOD and MaaS are evolving. The comparison in Fig. 1 is drawn from the latest literature.

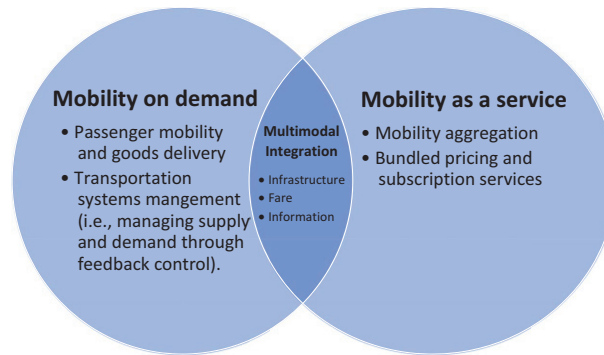


Figure 1 Similarities and Differences of MOD and MaaS.

Summary and Conclusion

Shared mobility is an innovative transportation strategy that enables users to gain short-term access to transportation modes on an as-needed basis for either passenger trips or goods delivery. The advent of bikesharing, carsharing, scooter sharing, TNCs (also known as ridesourcing and ridehailing), and other innovative mobility services is changing how travelers access transportation. The convergence of automation, electrification, and shared mobility will likely transform how people travel and how cities are planned and built in the future. While the impacts of this convergence and other transformative technologies on auto ownership, parking, and travel behavior remain to be seen, policymakers may need to rethink traditional notions of access, mobility, and auto mobility. The advent of shared micromobility and shared automated vehicles enables us to reimagine how we design and interact with cities through street design, curb management, and public policy (e.g., pricing, incentives). The integration of transportation modes, real-time information, and instant communication and dispatch—all possible with the click of a mouse or a smartphone app—have the potential to redefine urban mobility. These changes are contributing to the commodification (i.e., MOD) and aggregation of transportation services (i.e., MaaS). While the impacts of these changes are not fully known, these innovations will likely have a disruptive impact on cities and society in the future. Additional research is needed to understand the longer-term impacts of these innovative and emerging services.

Acknowledgment

The authors would like to thank the American Planning Association, the California Department of Transportation (Caltrans), the Mineta Transportation Institute, and the US Department of Transportation for their generous support of this work. Special thanks also goes to the shared mobility operators, industry experts, and policymakers who make this research possible. The contents of this article reflect the views of the authors and do not necessarily indicate sponsor acceptance.

Biographies

Susan Shaheen is a Professor in the Department of Civil and Environmental Engineering and a Research Engineer with the Institute of Transportation Studies at the University of California, Berkeley. She is also Co-Director of the Transportation Sustainability Research Center (TSRC) at University of California, Berkeley (UC Berkeley). She was among the first to observe, research, and write about the changing dynamics in shared mobility and the likely scenarios through which automated vehicles will gain prominence. She was the policy and behavioral research program leader at California Partners for Advanced Transit and Highways, a Special Assistant to the Director's Office of the California Department of Transportation, and the first Honda Distinguished Scholar in Transportation at the Institute of Transportation Studies at UC Davis, where she served as the endowed chair until 2012. She has authored numerous journal articles, reports and proceedings, book chapters, and co-edited two books. She has a Ph.D. from UC Davis and a M.S. from the University of Rochester.

Adam Cohen is an innovative mobility researcher at TSRC, UC Berkeley. Since joining the group in 2004, his research has focused on shared mobility, mobility on demand, mobility as a service, urban air mobility, smart cities and other innovative and emerging transportation technologies. He has co-authored numerous articles and reports on innovative mobility in peer-reviewed journals and conference proceedings. His academic background is in city and regional planning and international affairs.

See Also

Equity Concerns in Transport; Privatization and Deregulation of Public Transport; Public Private Partnership; How will Autonomous Vehicles Impact Car Ownership and Travel Behavior; Policy Instruments to Reduce Carbon Emissions From Road Transport; TNCs and the Future of Taxi Public Transport; Public Transport in Low Density Areas; Car Sharing Safety and Insurance; Bikesharing/Bicycle Sharing; Carsharing; Car Sharing; Connected and Autonomous Vehicles

References

- Cohen, A., Shaheen, S., 2016. Planning for Shared Mobility. Chicago: American Planning Association. Available from: <https://www.planning.org/publications/report/9107556/>.
- Shaheen, S., Cohen, A., 2019a. Mobility on Demand (MOD) and Mobility as a Service (MaaS) How Are They Similar and Different? Portland: Move Forward. Available from: <https://www.move-forward.com/mobility-on-demand-mod-and-mobility-as-a-service-maas-how-are-they-similar-and-different/>.
- Shaheen, S., Cohen, A., 2019b. Shared Micromobility Policy Toolkit. Retrieved from doi:10.7922/G2TH8JW7.
- Shaheen, S., Cohen, A., Farrar, E., 2018. The Potential Societal Barriers of Urban Air Mobility UAM. National Aeronautics and Space Administration, Washington D.C.
- Shaheen, S., Cohen, A., Yelchuru, B., Sarkhili, S., 2017. Mobility on Demand Operational Concept Report. Washington D.C.: U.S. Department of Transportation. Available from: <https://rosap.ntl.bts.gov/view/dot/34258>.
- Shaheen, S., Cohen, A., Zohdy, I., 2016a. Shared Mobility: Current Practices and Guiding Principles. Washington D.C.: U.S. Department of Transportation. Available from: <https://ops.fhwa.dot.gov/publications/fhwahop16022/fhwahop16022.pdf>.
- Shaheen, S., Cohen, A., Zohdy, I., Kock, B., 2016b. Washington D.C.: U.S. Department of Transportation. Smartphone Applications to Influence Travel Choices: Practices and Policies. Available from: www.ops.fhwa.dot.gov/publications/fhwahop16023/fhwahop16023.pdf.
- Sochor, J., Arby, H., Karlsson, M., Sarasini, S., 2018. A topological approach to mobility as a service: a proposed tool for understanding requirements and effects, and for aiding the integration of societal goals. *Res. Transp. Bus. Manag.* 3–14.

Further Reading

- Ambrosino, G., Nelson, J., Boero, M., Pettinelli, I., 2016. Enabling intermodal urban transport through complementary services: from flexible mobility services to the shared use mobility agency: workshop 4 developing inter-modal transport systems. *Res. Transp. Econ.* 179–184.
- Durand, A., Harms, L., Hoogendoorn-Lanser, S., Zijlstra, T., 2018. Mobility-as-a-Service and changes in travel preferences and travel behaviour: a literature review. Amsterdam: KiM Netherlands Institute for Transport Policy Analysis.
- Hensher, D., 2017. Future bus transport contracts under a mobility as a service (MaaS) regime in the digital age: are they likely to change? *Transp. Res. Part A: Policy and Practice* 86–96.
- Hilgert, T., Kagerbauer, M., Schuster, T., Becker, C., 2016. Optimization of individual travel behavior through customized mobility services and their effects on travel demand and transportation systems. *Transp. Res. Proc.* 58–69.
- Jittrapirom, P., Caiati, V., Feneri, A.-M., Ebrahimigharehbaghi, S., Alonso González, M., Narayan, J., 2017. Mobility as a service: a critical review of definitions, assessments of schemes, and key challenges. *Urban Plan.*
- Kamargianni, M., Li, W., Matyas, M., Schäfer, A., 2016. A critical review of new mobility services for urban transport. *Transp. Res. Proc.* 3294, 3303.
- Pangbourne, K., Mladenovic, M., Stead, D., Milakis, D., 2018. The case of mobility as a service: a critical reflection on challenges for urban transport and mobility governance. In: Marsden, G., Reardon, L. (Eds.), *Governance of the Smart Mobility Transition*, Emerald Group Publishing, Bingley, pp. 33–48.
- SAE International, 2018. J3163: Taxonomy and Definitions for Terms Related to Shared Mobility and Enabling Technologies. SAE, Warrendale, PA. Available from: https://www.sae.org/standards/content/j3163_201809/.
- Shaheen, S., Cohen, A., Dowd, M., Davis, R., 2019. A Framework for Integrating Transportation into Smart Cities: State of the Practice in 20 U.S. Cities. Mineta Transportation Institute, San Jose.

Adoption of New Travel Information Platforms

Sigal Kaplan, Department of Geography, Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel

© 2021 Elsevier Ltd. All rights reserved.

Introduction	160
Existing Information Platforms	160
Adoption Rates	161
Adoption Motivators	162
Adoption and Travel Behavior Change	163
Further Research	163
References	164

Introduction

Travel information is one of the basic needs of transport users, whether they are car users, public transport riders, cyclists or pedestrians. Information is a fundamental element both in decision making and behavioural change. While full information is rarely available, partial information is key to utilitarian-based decisions in terms of choice set awareness, evaluation of expected outcomes, and monitoring the choice results, thus motivating either habit formation or behavioural change. From the demand side at the individual level, travel information helps people to plan, execute and evaluate their travel choices according to their functional, social and emotional needs. At the social level, travel information aids demand management strategies to increase network capacity, encourage safer and efficient use of infrastructures, decrease carbon footprint and provide more inclusive transport environments for physically and socially vulnerable transport users. From the supply side, travel information increases transport resilience by allowing more efficient planning, detecting cavities and service imbalance, system optimization, real-time service adaptation, event detection, management and system recovery.

Starting from the first decade of the new millennium new low-cost information technologies are inducing a rapid and vast change in the way we see the transport world today. Until the end of the nineties, information gathering and dissemination was expensive, time limited and localized in nature, and thus gathered by transport and local authorities for their planning and operation. The introduction of smart-phones with embedded global positioning systems (GPS), wireless broadband reception/transmission and Bluetooth inter-device communication, have caused a major shift in data collection and dissemination because these technologies allow continuous, real-time and instantaneous passive and active information exchange. Moreover, the induced a shift in the transport information business structure as subsidiary bodies such as phone companies and software development companies serve as data brokers both collecting and selling data to authorities and operators. While some information platforms are of local nature and are financed by public authorities (e.g. 'MinRejseplanen' in Copenhagen), other information platforms have been developed by private sector entrepreneurs and are going global (e.g., 'Waze' and 'Moovit'). Both from the user and operator perspective, smart-phone based data platforms offer three main advantages of direct communication. Firstly, they allow better two-way communication between transport users and operators. Using smart-phones to track the space-time prism of the individual can be tracked over time, can be used for the development of agent-based models for prediction, to provide tailor-based solutions to users and to send real-time notifications to users on system changes. Secondly, these technologies have enabled inter-user communication, thus leading to the formation of transport user communities sharing information, travel resources and services. Last, these technologies have created the possibility for intercommunication between user devices, vehicles and road objects such as road signs and traffic signals.

While information technologies are changing the face of ground transportation, the change is inherently dependent on information technology acceptance, adoption rates, information use patterns and compliance. The main research interests focus on demand evaluation and behavioural compliance. From the business perspective, market demand and penetration rates are essential for understanding the potential market value, product cost-effectiveness and interest of private-sector entrepreneurs in developing such platforms either spontaneously or in collaboration with the public sector. From the transport planning and operation perspective, high market penetration is needed for data harvesting, assessing the potential impact on infrastructure use efficiency, travel time and cost saving, and the reducing environmental externalities.

Existing Information Platforms

Early information platforms provided both static and real-time localized information for a specific location and time frame. With the surge of personalized information and location-based services, and the shift from location-based to user-based information approach, recent studies are focusing on system-wide platforms which provide personalized route-planning solutions. The route

planning solutions allow pre-planning and en-route real time decisions. They display time and cost-based origin-destination route alternatives, as well as traffic congestion and incident alerts. Some route planners provide user-based preference options such as excluding toll route and defining preferred departure and arrival times.

While traditional platforms include information as their sole feature new platforms combine information and service packages relevant to the offered transport services (e.g., car sharing, bike-sharing transit). Examples are route-choice and car navigation apps, parking and car-finding apps, transit information apps, multi-modal information apps, and Mobility-as-a-Service (MaaS) apps. Car navigation apps offer price and time-based route information and suggestions, location information for parking and road services, traffic information, incident and weather alerts, and driver behaviour alerts (e.g., calculated speed with reference to the speed limit). Transit navigation apps provide door-to-door travel information consisting of travel route/line suggestions, walking instructions, transit schedules, alerts regarding service delays and changes, transfer suggestions, walking distances, waiting and arrival times. Bicycle information apps provide information bicycle route planning options, recording journeys and sharing user experience. Multi-modal information platforms provide route planning solutions for cities with multiple mobility services including transit, car-sharing, and bicycle options. Most of the research attention is provided to transit apps and recently to multi-modal apps due to the hope of promoting a shift towards shared and sustainable modes, because statistical evidence shows a positive relationship between information use and transit ridership.

The first generation of travel information platforms are developed on basis of information from service-providers such as local transit and road authorities. The second generation increased user participation in the data collection effort by basing the platform performance also on data harvested and traffic incident alerts from registered users. The new generation of travel information platforms heavily relies on both active and passive user engagement and community champions for map enhancements, service evaluation, traffic congestion estimation, system usage rates and increasing data accuracy and reliability at large. The new generation also supports and promote inter-user communication and collaboration across registered users, both on-line and through meetings and gatherings. Thus, the current travel information world moves from a top-down information provision approach to a collaborative community approach of both bottom-up and middle-out information provision from informal sources in addition to the traditional top-down data resources. An example is the transit information app 'Moovit' which amasses four billion anonymous data points a day to add to the world's largest repository of transit data. Moovit's network consists of more than 450,000 local editors called "Moovitors", who support the data gathering. These users help map and maintain local transit information in cities that would otherwise be difficult to serve. Thus the new age of travel information platforms seek to mobilize communities to engage in travel data gathering efforts. Most information platforms are mode-specific, focusing on car, transit and bicycle travel as separate modes.

Voluntary travel behaviour change (VBTC) elements are starting to emerge as an integral part of the next generation travel information apps. Among these elements are health and environmental feedback, tailoring travel options to specific traveller's needs, tracking for self-monitoring, social networking, nudging and gamification elements. From the user perspective, these elements can be translate into functional gain benefits (e.g., free services, discount coupons and upgrade possibilities), normative benefits of acting in accordance with one's life values and hedonic benefits deriving from the competition/game enjoyment

Information platforms for mobility-impaired transport users form a separate category as an important tool empowering safe and autonomous travel and serve as a cost-efficient, independent and effective alternative to enhancing the urban environment with mobility assistance aids. These platforms are mostly directed towards walking as a main mode and are still at their nascent stage. Thus most of the studies are dedicated to finding the data collection needs in pedestrian environments and facilitating user interface. Among the necessary features uncovered in existing studies are map enhancements with street objects, pavement characteristics, traffic flows, and static obstacles, minimum-barrier route selection, turn-by-turn audio instructions, sending location notifications, communication possibilities with trusted sources such as peer and authorities, voice, vibration and gesture interfaces, image analysis and location exploration possibilities, tagging and collaborative user information exchange. These apps are highly dependent on collaborative voluntary mapping efforts for increasing information accuracy and detecting dynamic changes, and thus user and non-user willingness to engage in mapping efforts through information sharing is an integral part of their success.

Adoption Rates

The full extent of travel information platform adoption and transport information consumption cannot be estimated because such data is not elicited in national or regional travel habit surveys. The developers of the two main information platforms for car navigation and transit information, namely 'Waze' and 'Moovit' publish records of 65 million monthly active 'Waze users in 185 countries and 300 million registered 'Moovit' users in over 2,600 cities in 85 countries. The local mobile application for transit information in Melbourne (Australia) is downloaded more than 4000 times a week. Studies from various regions confirm that travel information is continuously and frequently sought by habitual and occasional travellers alike, in routine and non-routine travel, and for both short and long-distances. Moreover, studies show that information search is an important part of travel for route finding, trip planning and en-route solutions regardless of the quality of the transport system in terms of coverage and level-of-service. A study comparing Copenhagen (Denmark) to Recife and Natal (Brazil) in 2016 found that in both cases the most common sources of information are Google Maps (75.35% in Recife and Natal and 81.6% in Copenhagen) and on-line information (72.11% in Recife and Natal and 93.95% in Copenhagen). A study from Innsbruck and Copenhagen in 2019 add that transit information is that 58% of transit users are interested in level-of-service information (e.g. arrival times, delays, crowding, accidents) and that 41% of transit users are seeking information more than twice a week and additional 22% are seeking information once a week. In another survey in

Copenhagen from 2019, a survey found that 32% of the respondents use car navigation apps frequently, compared with 51% adoption rates of transit information apps and 17% adoption rates of bicycle information platforms. These results echo in other countries. In a survey in China, 77% of the respondents sought transit information at least once a week. In Bristol 57% of the respondents sought information very often for leisure and business trips over 50 miles within the UK, and obtained transit information for unfamiliar trips.

Adoption Motivators

Adoption of new information technologies concerns understanding the functional and psychological motivators for hardware and software technology adoption of advanced travel information services (ATIS). The behavioural models are applied to understand the expected market share (who are ATIS users?), the usage rates (how much, when and where ATIS are used?), the reasons for adoption (Why ATIS is used?) and the impact of ATIS use on mode choice, with transit and bicycle use in particular.

Existing studies are often conducted in relation to the development of new region-specific, mode-specific and service-specific travel information platform prototypes. Hence, in accordance with early information platforms most of the existing studies explain platform adoption for receiving information. Experimental qualitative analysis of small-scale user groups and controlled field experiments are designed to explore the usefulness of specific app features (e.g., weather forecasts) or reveal particular user needs (i.e., mobility-impaired). Adoption intentions and app use are investigated by employing empirical surveys and estimating latent-variable models based on motivational psychology. The explored behavioural models are the Technology Acceptance Model (TAM) and the Unified Theory of Technology Acceptance and Use (UTTAU). The models are technology oriented bringing forward system design, facilitating conditions, user characteristics and social-influence. System design includes perceived information usefulness and quality (performance expectancy) and ease-of-use (effort expectancy) and potential risks related to privacy and data security. The facilitating conditions refer to the transport service climate as well as internet and supporting technology availability and are rarely investigated due to the need for cross-regional analysis. User characteristics include observable indicators of socio-economic and trip characteristics, and latent indicators of attitudes, traits and needs. Social influence refers to visibility and social norms. Among the addressed barriers for travel information platform adoption are: travel information availability awareness, perceived technical difficulties involving hardware (i.e., battery consumption, user interface) and software (e.g., privacy concerns for information collection, secondary usage and improper usage as well as data security issues), trust in the provided information, lack of perceived need and habitual behaviour suppressing information seeking needs. Motivational system features are availability of accurate high-quality information, perceived usefulness in case of disruptions and system changes, hands-on experience and comparison with former travel choices. Individual internal motivation factors include innovation affinity, use of other location- and travel-based services, time-saving skills and perceived importance of efficient travel and environmental attitudes. In existing studies system familiarity is positively associated with information consumption and thus is regarded as a motivational factor in the adoption of information platforms.

Considering collaborative information platforms, a sufficiently large user base and the willingness to share information are essential to the future of collaborative exchange, both for passive and active crowd-sourcing. However, information exchange often exhibits the characteristics of a social dilemma. While the best approach from an individual utilitarian perspective is to use data without reciprocity, the notion of 'free riders' impedes the collaborative data gathering effort. Moreover, daily information sharing is a repetitive 'grey' task with benefits being better information quality and small tokens of appreciation. Understanding travellers' willingness to use transit apps in dual mode, also as data providers, is the new research issue in the new age of collaborative travel apps. A new study from Copenhagen and Innsbruck indicate that 35% of transit users are willing to share travel information with other platform users on occasional basis and another 15% are willing to engage in data provision in a frequent manner. The UTTAU has been employed also to uncover the motivators for information sharing. The strongest motivators for information sharing are social in nature: social norms of sharing, self-actualization associated with helping others and social network engagement habits. Thus the sense of community and information sharing mutually enhance each other because collaborative travel information apps have been found to strengthen the sense of community, social trust and group identity through continuous communication among platform users. Motivators with weaker impact are performance expectancy in terms of information quality and trust in user information, as well as network familiarity facilitating the information provision effort. The perceived difficulty in terms of dedicated time, effort and the need of continuous on-line engagement is a strong inhibitor for information sharing in collaborative travel information platforms.

The acceptance and use of VBTC elements including gamification can serve as 'game changers' in raising the attractiveness of some mobility solutions over others, are potentially powerful in motivating behavioural change and have the potential to induce both benefits and externalities. The effectiveness of VBTC is much researched and is discussed in the next section. A new study in Copenhagen explored adoptions intentions of a travel information platform with integrated gamification elements in order to promote 'green' and 'healthy' travel by producing a modal shift towards bicycle and public transport use. The study investigated normative, gain and hedonic framing motivators by applying the Existence Relatedness and Growth (ERG) theory of human needs. Unlike the UTTAU, which is technology-oriented, The ERG model is anthropocentric, focusing on the ability to satisfy individual needs as adoption motivators. The strongest motivators for using the new travel information platform were gain in trip efficiency improvement and normative reasoning of eco-travel promotion. The two concepts were found correlated with beliefs in trip efficiency re-enforcing eco-travel reasoning. Trust in travel information technology is the third strongest motivator for adoption

intentions of the new platform. Social self-concepts related to environmental activism and to trip efficiency improvements constitute the fourth main adoption motivator. Technical difficulties were found to be stronger adoption impediments compared to privacy and data security concerns.

Adoption and Travel Behavior Change

Compliance with travel information and the ability of information to produce behavioural change is directly related to the provision of effective information solutions associated with increased functional utility and psychological needs. At a first glance, the question is seemingly trivial as additional information may contribute to changes in the travel problem structure and clarify utility gains and losses. In fact, providing traffic and weather information was found to be a trigger for behavioural change already in the end of the nineties with information being radio broad-casted. Nevertheless, behaviour is motivated also by habits, need for control, constraints, perceptions and beliefs that may impede behavioural change both at the single trip level and in the long-term. Additionally, in repeated travel experience, where travel information is supplied on per-trip basis, questions related to the perceived information reliability and regret regarding the non-chosen alternatives may guide short- and long-term compliance. While research is scarce regarding the use of information for planning transit trips and the linkage between seeking transit information and transit use, studies from the last five years show a positive relation between transit information and transit ridership. Studies show that transit information reduces waiting time and its associated stress, and increases transit ridership during habitual travel, occasional trips to unfamiliar locations, and during night. A study from Copenhagen shows that the main motivations for travel information platform use are to save travel time and to seek better alternatives during delays or service changes. In the same study, roughly 30% of the respondents strongly agreed that a multi-modal travel information would help them to cycle more and to use transit and car-sharing more often. A study conducted in Atlanta, Sydney and the U.K. over three years from 2014-2016 show that during a cycling challenge registered users on a smart-phone app cycled 2-3 times more frequently than non-app users. A study conducted in Istanbul, provides evidence that 53% of car users 80% of transit users change their route according to the travel information platform suggestions.

Further Research

Technology use - studies are cross-sectional and focus on adoptions intentions rather than on actual use patterns. An important knowledge gap concerns tracking actual use patterns following the travel information platform adoption. Tracking use patterns will allow a better understanding of the app use frequency, manner and purpose as well as user historical search patterns. While adoption is the first step for app use, the actual use frequency and patterns determines the actual extent of the platform penetration rate, which is necessary for product profitability, sustainability and impact. Secondly, cross-product and cross regional comparisons are lacking and research is focused on the use/non-use of a specific product rather than the choice among various alternatives. Thus, knowledge regarding the relative attractiveness and of design and content features as well as technology engagement motivation is lacking. With the proliferation of the travel information platform market and the existence of both private-sector and public-sector solutions, and the global success of a few travel information apps product comparison is important to determine the actual market share of various solutions.

The digital divide - most of the studies on travel information platforms were conducted in cities with well-developed formal transport systems. More research is needed on app use patterns and information exchange in emerging markets and developing economies with higher degree of service proliferation, lack of service information and dominant informal transport markets. In addition, current travel information systems are designed without consideration of vulnerable travellers. Examples are travellers with mobility and visual impairments, learning disabilities (e.g., Dyslexia, Dyscalculia), elderly, immigrants and asylum seekers. Studies addressing vulnerable travellers will be helpful in reducing the digital divide towards a more inclusive transport information services.

Counter-intuitive choices - an interesting further research question concerns information leading to seemingly counter-intuitive choices, such as trading short-term time savings for long-term healthier choices and preferring collective-efficacy for reaching system-optimum over self-efficacy leading to user-equilibrium. A recent study involving computer simulation of repeated games proposes that travellers can learn how to share simplified road networks in a way that maximizes both personal and societal utility through travel information provision. A stated-preference survey also showed that given demand management explanations increases collaboration with system optimum strategies by 10%. An experiment simulating travel information app use and route choice in a controlled environment show a tendency to comply with the travel information platform suggestions as an anchor to decision making, even in conditions of poor information accuracy. Five case-studies entailing transit use, active travel and safe-driving have shown that gamification can lead to significant proportions of modal shifts towards bicycle and transit use as well as reduce road accidents and fatalities. Embedding these strategies in travel information platforms is the next step. Understanding how these strategies generate long-term sustainable behaviour should follow.

Related services - information apps developed in the private sector are now offering the ability to book transport resources from subsidiary providers, as well as peer-to-peer ride sharing possibilities. New multi-modal platforms of mobility services (e.g., 'Whim') include embedded information regarding the transport solutions. New platforms also offer information about activity, retail and entertainment locations. Investigating mobile payment acceptance and the adoption of related non-transport related services within

the transport information platform is important because these services may determine the platform's lifespan through its attractiveness to stakeholders including private sector investors, collaboration partners and consumers.

References

- C. BartleE. AvineriK. ChatterjeeOnline information-sharing: A qualitative analysis of community, trust and social influence amongst commuter cyclists in the UKTransportation Research Part F1620136072.
- E. Ben-EliaR. Di PaceG.N. BifulcoY. ShiftanThe impact of travel information's accuracy on route-choiceTransportation Research Part C262013146159.
- C. BrakewoodS. BarbeauK. WatkinsAn experiment evaluating the impacts of real-time transit information on bus riders in Tampa, FloridaTransportation Research Part A692014409422.
- C.G. ChorusT.A. ArentzeH.J.P. TimmermansE.J.E. MolinB. Van WeeTravelers' need for information in traffic and transit: results from a web surveyJournal of Intelligent Transportation Systems11220075767.
- A.M. DastjerdiS. Sigal KaplanJ. Abreu e SilvaO.A. NielsenF.C. PereiraParticipating in environmental loyalty program with a real-time multimodal travel app: User needs, environmental and privacy motivatorsTransportation Research Part D672019223243.
- S. FaragG. LyonsTo use or not to use? An empirical study of pre-trip public transport Information for business and leisure trips and comparison with car travelTransport Policy2020128292.
- I. GokasarG. BakiogluModeling the effects of real time traffic information on travel behavior: A case study of Istanbul Technical UniversityTransportation Research Part F582018881892.
- X. HouX. ChenAnalysis of urban public transit information requirements in china by web-based surveyProcedia - Social and Behavioral Sciences96201315221527.
- R. Islam SarkerS. KaplanM.K. AndersonS. HausteinM. MailerH.J.P. TimmermansObtaining transit information from users of a collaborative transit app: Platform-based and individual-related motivatorsTransportation Research Part C: Emerging Technologies1022019173188.
- S. KaplanM. Moraes MonteiroM.K. AndersonO.A. NielsenE. Medeiros Dos SantosThe role of information systems in non-routine transit use of university students: Evidence from Brazil and DenmarkTransportation Research Part A: Policy and Practice9520173448.
- I. KleinN. LevyE. Ben-EliaAn agent-based model of the emergence of cooperation and a fair and stable system optimum using ATIS on a simple road networkTransportation Research Part C862018183201.
- A. KramersDesigning next generation multimodal traveler information systemsEnvironmental Modelling & Software5620148393.
- T. MatsumotoK. HidakaEvaluation the effect of mobile information services for public transportation through the empirical research on commuter trainsTechnology in Society432015144158.
- Min, B.C., Saxena, S., Steinfeld, A., Dias M. B., 2015. Incorporating information from trusted sources to enhance urban navigation for blind travellers. 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, Washington, May 26-30.
- M. RinghandM. VollrathMake this detour and be unselfish! Influencing urban route choice by explaining traffic managementTransportation Research Part F53201899116.
- G. VelazquezS. KaplanA. MonzonEx-Ante and Ex-Post evaluation of a new transit information app: modelling use intentions and actual useTransportation Research Record2018110.
- J. WeberM. AzadW. RiggsC.R. CherryThe convergence of smart-phone apps, gamification and competition to increase cyclingTransportation Research Part F: Traffic Psychology and Behaviour562018201833343.
- B.T.H. YenC. MulleyM. BurkeGamification in transport interventions: Another way to improve travel behavioural changeCities852019140149.

ICT and Transport Modes

Galit Cohen-Blankshtain, School of Public Policy, Department of Geography, The Hebrew University, Jerusalem, Israel

© 2021 Elsevier Ltd. All rights reserved.

Introduction: ICT and Transport Modes	165
ICTs as a New Transport Mode	165
ICTs as Modifiers of Existing Transport Modes	167
Direct Impact: Intelligent Transport Systems (ITS)	167
Indirect Impact: Modifying Spatial and Temporal Behavior	168
Temporal and Spatial Modification of Travel	168
Fragmentation of Activities	168
Travel-Based Multitasking	169
Summary and Conclusion	169
Biography	169
See Also	169
References	170
Further Reading	170

Introduction: ICT and Transport Modes

Information and communication technologies (ICTs) are a family of technologies and services used to process, store and disseminate information, facilitating the performance of information-related human activities, provided by and serving both the public at large as well as the institutional and business sectors. Since the 1960s, with the earliest anticipation of the information age, ICTs have attracted the attention of transport experts as the potential new transport mode. In fact, the anticipation of substituting physical movements with virtual movement is even older, as the telegraph and later, the telephone technologies raised similar expectations. As both ICTs and transport technologies facilitate remote activities, there has been much interest in the potential substitution of physical travel by teleactivities. However, with the increasing adoption of ICTs and observation of the penetration style of such technologies it has become clear that, along with its potential to substitute physical travel (which has been lower than expected), ICTs change the transport system and the way we consume it. As such, ICTs are not just a new, virtual transport mode; they have modified the “old” physical transport modes (Mokhtarian and Tal, 2013).

Rapid technological development has exponentially increased the technologies covered under the term ICTs. While until the beginning of the 21st century, ICTs referred mainly to stationary home computers connected to the internet, in the last 20 years, large segments of the population own smartphones, which allow them to be connected to the internet almost anywhere and at any time, and apply a wide variety of software (namely, apps). Therefore it is almost impossible to generalize about the relationships between ICTs and transport behavior, as ICTs are embedded in various devices, activities, and practices and have different relationships with physical activities and travel.

Nevertheless, the extensive and diverse scholarly work regarding ICTs and transportation offers interesting and important insights that go beyond specific technology or application. In the following sections a review of the potential and actual influences of ICT on the transport system is offered, both approaching ICTs as a new transport mode (substitution between physical and virtual activities) and ICTs as modifying the way we use the existing transport modes and perform activities. This paper is organized as follows: the second section considers the potential of ICTs as replacement to the physical transport system, that is, ICTs as a new transport mode. The third section discusses the various modifications of ICTs on the physical system, suggesting that the physical transport system is not replaced but dramatically altered via ICTs, and the last section offers concluding remarks.

ICTs as a New Transport Mode

As the demand for travel is approached as derived demand, that is, derived from the demand for the activities it enables, much of the research regarding substitution between transport and ICTs focuses on different activities and their potential to become virtual, without the costly physical movement. Thus there is large body of research focusing on trips to activities as telecommuting or telework, teleshopping or e-shopping, teleeducation or e-learning, teleconferencing and telebanking. Similarly, organizational activities have also been examined as e-commerce (between firms) or e-governance (between the government and citizens or within the government). The motivation for all “E” and “tele” activities is clear: Instead of spending time, money and other resources traveling to engage in an activity, virtual travel via ICTs is offered.

Since commuting trips are the main contributors to congestion (both of private and public transport), research focusing on substituting telecommuting for commuting, that is, working from home via ICTs, was first suggested in the 1960s. The earliest forecast anticipated significant transportation benefits during the 1980s and early 1990s. During the 1970s it was estimated that in 2000, telecommuting could reduce more than 10% of the total urban VMT (vehicle miles traveled) when fully implemented. However, estimations of the actual impact of telecommuting by 2000 suggest an aggregate impact of less than a 1% VMT reduction (Tal and Cohen-Blankshtain, 2011). While the early expectations for large scale substitution between commuting and telecommuting assumed that most of the workers who *could* telecommute, *would*, and that this would cover all commuting trips, in reality obstacles (both at the workplace and at home) and different personal preferences led to much lower substitution rates and thus to a lower impact on total VMT. Although telecommuting shares have increased over time, telecommuting is mostly partial (i.e., part of the working day) or in addition to commuting trips (i.e., leaving home later, after morning telecommuting). Moreover, since many workers prefer working out of their homes, alternative workspaces have developed, offering shortened commuting trips for all or part of the working week. Thus telecommuting has many facets: working from home part or all of the week, working from home part of the workday, and working from nonhome and nonworkspaces. As a result, the telecommuting impact on the transport system should be measured not only by VMT reduction, but also by its effect on peak hour travel and route choices.

Similar to telecommuting, telescoping or e-commerce was perceived as an activity that would substitute for travel. For shopping purposes, a substitution effect occurs when e-shopping reduces the number of shopping person-trips and/or person-distance traveled. For the most part, when purchases of physical goods are made online, in-store shopping will be replaced by home delivery. Shopping activity is made up of three parts: searching, purchasing, and delivery. Teleshopping can replace all or part of the shopping activity: one can replace physical search with virtual search but continue with physical purchase and delivery while others will replace all three physical activities with virtual activities. Substitution between physical shopping and telescoping can therefore be partial or full, depending on the type of the good, personal preferences and spatial characteristics. Product characteristics affect the potential of online shopping and substitution of different parts of the shopping activity. Books or electronic devices are more suited to purchasing via the Internet than experience goods such as fresh vegetables. Moreover, people tend to buy more expensive products online because of possible larger savings compared with prices in physical shops. Recreational shoppers are less likely to fully substitute shopping compared to functional shoppers, and shoppers with tight time constraints are more likely to replace more physical shopping activities with virtual ones. And finally, spatial characteristics such as accessibility to physical stores by public transport or availability of parking also affect the potential substitution between in store shopping and on-line shopping.

However, various studies have asserted that e-commerce will have a limited or even a neutral impact on the number of shopping trips and distance travelled. This may arise in cases where shopping trips are combined with other activities or when people make multiple purchases during a single shopping trip, and the substitution of e-shopping for only some in-store items does not necessarily decrease the number of shopping trips and will not decrease travel distance.

Nevertheless, e-commerce is growing rapidly. During 2017, one out of five enterprises in the EU-28 made electronic sales, amounting to 17% of the total turnover of enterprises with 10 or more persons employed. In 2017 the percentage of enterprises making e-sales ranged from 8% in Bulgaria to 35% in Ireland, closely followed by Sweden and Denmark (both 32%), Belgium (30%) and the Netherlands (27%).^a In 2019 the total e-commerce worldwide sales are estimated at €621 billion. In the United States, e-commerce sales in the first quarter of 2019 accounted for 10.2% of total sales (\$127.3 billion).^b Against this background it is clear that on-line purchase is increasing steadily and is substituting some in-shop activities. It is less clear, however, *the extent* to which on-line purchases substitute for physical shopping activities and the extent to which it stimulates more consumer activity and thus does not necessarily reduce physical shopping travels.

However, ICTs can affect the location of commercial activities and affect the distance between consumers and suppliers.

Another perspective that is relevant to shopping activities is the supplier side, that is, the extent to which warehouse and commercial chains may replace physical movements with virtual ones (Perego et al., 2011). It was found that ICTs help determine the most time-effective and cost-efficient transportation mode(s) for a given shipment, as well as reduced travel distances as a result of travel optimizations.

Other dominant activities that generate a large share of travel are leisure pursuits, which play an increasingly important role in our daily lives. Their importance has increased steadily over the last 40 years compared with other activities. Part of the growth is due to the increasing numbers of activities that are considered leisure pursuits. In this respect, shopping is not always regarded as a chore; it is frequently perceived as a recreational activity, a pleasant means of physical movement, and as a means of social interaction. Additional leisure activities are social interactions, tourist-related activities and entertainment. Regarding social interactions there are consistent empirical findings indicating that when distance increases, face-to-face contacts are substituted by virtual contacts. Moreover, since time constraints during weekdays are greater than on weekends, virtual social contact substitutes for physical contact during the week, but to a lesser degree during weekends. However, in general, virtual interaction does not replace physical contact, but facilitates it. In other words, ICTs tend to generate physical contacts and do not substitute for face-to-face contacts.

In addition, other leisure activities such as tourist and entertainment related trips do not experienced decline due to virtual activities. The consistent growth in tourist movement suggests that here also, ICTs facilitate and fuel increased demand for tourist activities by presenting places, options, and prices.

^a https://ec.europa.eu/eurostat/statistics-explained/index.php/E-commerce_statistics.

^b https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf.

And finally, maintenance activities such as banking, contacts with governmental bureaucracy or other institutions are strong candidates for substitution with virtual maintenance, since many of these activities depend less on the quality that face-to-face meetings offer. Indeed, the growing number of banking services that are offered online, and the fast penetration of these services is well documented. However, transportation scholars have hardly investigated the transport derivative of these rapid changes in banking services, mainly because, at the personal level, banking activities are not frequent and thus have negligible effect on VMT. Therefore although telebanking has substantial effect on the banking system, the spatial patterns of banks and the accessibility to banking services, transport-related research overlooks the potential and actual effects of telebanking on travel behavior. Similarly, e-government activities and their potential to replace physical maintenance activities are overlooked in transport-related research. More empirical research regarding the substitution between physical and virtual maintenance activity and its potential VMT reduction is called for.

While it is useful to examine different activities and their potential for replacement by teleactivities, it is important to acknowledge that at the personal level, engagement in digital activities is often not limited to one activity, that is, people are often involved in more than one teleactivity. It may be looked upon as a digital lifestyle, including various teleactivities that have replaced physical activities, compared with nondigital life style that makes less use of teleactivities. In addition, research that focuses on just one type of activity may be misleading. As [Salomon \(1985\)](#) stresses, “The analysis of total human activity may lead to different conclusions from the analyses of single trip purpose.” In other words, teleactivity may substitute certain physical activities but induce different activities. The overall impact of teleactivities on the transport system does not envision a replacement of the physical system with a virtual one. While ownership and usage of ICTs has increased dramatically over the years, VMT also continues growing and there is no sign yet of reduction or even saturation in physical movements.

Examining the relationships between ICTs and the transport system can be done not only with respect to various human activities, but also with respect to different transport modes. Such analysis raises the question whether different transport modes have different *replacement* potential with ICTs. However, most of the differences between transport modes are manifested in *changing* the use of this transport mode, not by replacing it, as detailed in the next section. One exception is the general expectation that Millennials who use ICTs applications more frequently will have fewer cars and drive less. Although there are evidences supporting it, analysis found that it is largely a manifestation of economic factors (that is, Millennials have less economic resources compared to their parents in their age) and not necessarily a structural change ([Klein and Smart, 2017](#)).

ICTs as Modifiers of Existing Transport Modes

As the previous section demonstrates, approaching ICTs as new mode of transportation captures just part of its impact on the transportation system. ICTs potential to replace physical transport modes is limited, and empirical evidence suggests that in most activities, substitution between physical and virtual activity is less than expected. However, there is consistent empirical support for the claim that ICTs modify the transport system both through direct impact on the transport system, or indirect impact on spatial and temporal activity patterns.

Direct Impact: Intelligent Transport Systems (ITS)

ICTs are at the heart of the development of an intelligent transportation system (ITS) that supports the physical transport system and enables its more efficient use, coined often as “smart mobility.” ITS can be divided into two categories: intelligent infrastructure systems and intelligent vehicle systems. The ultimate product of ITS is the combination of the two, intelligent vehicles that are connected to intelligent infrastructure that reduce human involvement in driving, leading to automated, autonomous connected vehicles. However, while autonomous vehicles are still not operating in the market, many ITS components are already part of the transport system, with significant impact.

Intelligent infrastructure systems are often concerned with the operational perspective. They offer smart systems for parking management, toll roads, public transport operation, coordinated freight transport systems, firm logistics and road management. The use of ICTs in transport operation increases system capacity, increases its efficiency by means of operation costs and level of service and enhances its reliability. In other words, instead of replacing the physical transport system, ITS enables better use of the system, therefore, enabling more physical movements.

On the other hand, intelligent vehicle systems are personal devices that enable travelers to adapt to road condition, avoid congested routes, park their cars with a higher degree of certainty and use public transport more efficiently. With the substantial development in ICTs and related sensing technologies, new vehicles are equipped with vehicle automation and communication systems (VACS) that can improve individual safety, comfort, and convenience as well as having the potential to develop global traffic efficiency through traffic control.

Using ITS at the personal level reduces travel (in private and public transport) and parking costs, and therefore may have a positive effect on the demand for travel.

The distinction between infrastructure and vehicles is becoming blurred with the emergence of technologies that connect vehicle and infrastructure and the use of big data that have made it easier and cheaper to collect, store, analyze, use, and disseminate multisource data ([Sumalee and Ho, 2018](#)). Vehicle-to-vehicle as well as vehicle-to-infrastructure connections enable more informed and manageable transport systems that can react in real time to expected and unexpected incidents, optimize system usage and better inform its various users. As both intelligent infrastructure systems and intelligent vehicle systems improve the capacity and

usefulness of the transport system, it may be concluded that an increase of physical movements can be accommodated, contradicting the expectation of replacing physical movements with virtual ones.

Freight movements attract special attention within ITS applications. Developments in the field of freight resource management, freight and vehicle tracking and tracing, and front or back-office logistic systems have dramatically improved freight movement and supply. There are numerous types of city logistic schemes that have either already been implemented or are proposed: advanced information systems, cooperative freight transport arrangements, public logistics terminals, load factor controls, and underground freight transport organization. As the share of goods movement in total kilometers increases over time and its presence at the transport system become more dominant, the importance of efficient freight movement increases as well.

Indirect Impact: Modifying Spatial and Temporal Behavior

ICTs not only change our travel mode, they also change our activity patterns. This means that we approach activities differently and perform the activities in different places and at various temporal opportunities. First, ICTs may have temporal and spatial effects on our physical travel, that is, they may facilitate peak-hour avoidance or congested routes by changing departure time or work location. Second, and related to the first, ICTs encourage fragmentation of activities, enabling activity performance at various locations during the day. And third, ICTs intensify multitasking, that is, performing more than one activity simultaneously. As a consequence of activity fragmentation and multitasking, ICTs affect our travel utility by enabling performance of useful, productive activities while traveling, compared with previous conceptions of travel time as waited time.

Temporal and Spatial Modification of Travel

Research has identified several modifications in travel patterns that have resulted from ICT usage as ICTs increase flexibility and open more opportunities for coping with space-time constraints. First, ICTs may affect the length of trips, both shorter and longer trips. Those who choose to work from offices closer to home in places offering temporal working spaces, (all or part of the week) will have shorter commuting trips. On the other hand, those who choose to work from home may move to live further from their workplaces so that, when working from the office, they will have longer commuting trips (though this may be offset by nondriving on days of telecommuting).

Second, ICTs can modify trip routing and destination. This may be a result of real-time estimation of traffic congestion delivered by various applications (like Waze or Google Maps) or ICT facilitation of microcoordination, such as redirection of trips that have already started, by contacting other people during travel and changing the place or the time of a planned meeting.

Third, ICTs may affect the timing of physical travel, especially avoiding peak hour trips, if one chooses to work from home in the early morning, and travel to the workplace after morning peak hours.

Fragmentation of Activities

In contrast with efforts to quantify the total effect of ICT on VMT, scholars have identified the importance of day-to-day management of activities and movements, that is, examining the role of ICTs in the distribution of activities across time and space during the day. As Hubers and her colleagues define it, “activity fragmentation implies that a certain activity becomes divided into several smaller components, which can each be performed at different times and/or locations” (Hubers et al., 2018, 94) and have identified three aspects of fragmentation: the number of fragments, the duration of each fragment and the temporal and spatial distances between the different activity episodes and locations.

ICTs relax the ties between location and activities that are taking place, thus enabling division of activities between locations in a way that better fits personal time needs. For example, working activity that used to be constrained to the workplace (spatially) and during working hours (temporally) can be divided into a morning session at home (e.g., reading and responding to e-mails between 07:00 and 08:00), a traveling session (e.g., a phone discussion with a colleague while driving between 8:30 and 9:00), a workplace session (e.g., meetings between 09:00 and 16:00) and an afternoon session while waiting outside your child’s ballet class (e.g., finalizing a written document between 17:00 and 18:00).

In the same vein, other, nonwork activities can be fragmented more easily with ICTs. Shopping activities can be in-site but also remotized, depending on the place and time of activity, as well as education, maintenance or social relationships. Since fragmentation means that there are episodes of activity that are spread along time, we can see maintenance activities during working hours, or shopping activities during studying hours. The spatial independence of certain activities loosens space-time constraints and compensates for more rigid time or space activities by enabling episodes of other activities when location and time are not flexible.

Fragmentation of activities facilitated by the proliferation of ICTs may have bi-directional consequences for the transportation system. On the one hand, it may encourage movement, as the cost of replacing locations declines when they have less influence on the activities employed. On the other hand, it may reduce the motivation to move and replace locations, since at the same location we can be engaged in various activities, and adjust the activities to time constraints, and that may reduce travel. But the consequences of activity fragmentation go beyond the transport system. Fragmentation and as a result, a mixture of activities and diminishing borders between work, family or leisure activities have wider societal influence, affecting the relative weight of the different activities undertaken and our well-being.

Travel-Based Multitasking

Performing multiple activities simultaneously is referred as multitasking. Specifically, travel-based multitasking is the performance of additional activity while traveling. Scholars distinguish between active and passive activities, that is, travel can be active (as a driver) or passive (not driving) and the other simultaneous activity can be active (reading, talking on the phone) or passive (sleeping, daydreaming) (Circella et al., 2012).

Most empirical studies on travel-based multitasking have focused on public transit passengers, reflecting the fact that transit passengers are the most likely travelers to engage in travel-based multitasking, since they are passive travelers. Bus and rail transit riders, in particular, have benefited from ICT developments that have offered growing number of potential multitasked travel activities. A small (however increasing) number of studies have analyzed travel-based multitasking across multiple transportation modes. Multimodal studies demonstrate that travel-based multitasking is relatively common across all modes, and some activities are more suitable to performing privately and not while using public transport (e.g., talking or eating). Active or passive traveling affects the additional activities that are performed. Car drivers are more likely to be listening to music or other audio. Passive travelers, on the other hand (public transport users or car passengers) tend to read, write, rest, or sleep. In addition, travel-based multitasking appears to increase with travel time.

Multitasking attracts transportation scholars' attention mainly since it may affect the perceived travel time costs, as portable ICTs may influence travel decisions by making travel more productive and therefore, not a complete waste of time. As multitasking during passive travel offers more opportunities for simultaneous activities, multitasking is often considered a tool to promote public transport usage. It suggests straightforward policy implications: public transport should accommodate travelers' additional activities (such as offering wi-fi connection, electrical connection, shelves, or adjusted tables) in order to attract travelers to public transport and increase their utility from the travel (Hong et al., 2019).

Summary and Conclusion

After more than a half century of extensive empirical research regarding possible relationships between ICTs and physical movements, scholars agree that no straightforward, simple relationships can be identified. Those relationships vary across activities, technologies, personal preferences, and life situations. Currently, the expectation that ICTs would be the new mode of transport and replace the physical one has not materialized. Indeed, the use of ICTs has increased dramatically and is expected to continue with its rapid penetration. However, the use of physical transport is not declining; on the contrary, VMT continues to grow and physical movements are still a central part of our everyday lives. However, it may well be that ICTs replace certain activities, at least partially, and by relaxing our travel budget, they enable addition of new physical trips with higher value. Thus although the total VMT is not affected substantially, the composition of our travel has been altered via ICTs-based activities.

Although the expectation for substitution between virtual and physical activities has not taken place, ICT impacts on travel behavior are no less than dramatic. ICTs have modified the transport system both directly via ITSs that are embedded in vehicles, infrastructural personal equipment, and organizations active in the transport system, and indirectly via changing our tempo-spatial activity patterns, simultaneous activities and utility from time spent traveling. ITS increase the efficiency of the transport system and its capacity, thus may encourage more physical travel. Modification in our activity patterns on the other hand may increase physical movements, but in other cases it decreases physical transport usage. The implication of modification of activities pattern goes beyond the transportation system, it affects productivity of the working force, consumption patterns, career and family life, life satisfaction and personal and social relationships, to mention a few. Thus, ICTs should be approached not as a merely new transport mode, but as technologies that modify human perception, activities and human interaction, and therefore, the general urban metabolism.

Biography

Dr. Galit Cohen-Blankshtain is a senior lecturer at the Department of Geography and the School of Public Policy at the Hebrew University, Jerusalem. She earned her PhD from the Free University of Amsterdam and spent a year as a Post doctorate fellow at Harvard University, Kennedy School of Government. Her main research interests regarding public policies in various filed as transport, urban policy, and environmental policy, as well as public involvement in policy processes. She is interested at policy processes and examining different societal, institutional, and political aspects of transport policy and the interaction between policy and community.

See Also

ITS in Road Freight Transport; Virtual and In-person Activity Participation Analysis; The Future of Road Transport; ITS for Transport Planning and Policy; How will Autonomous Vehicles Impact Car Ownership and Travel Behavior

References

- Circella, G., Mokhtarian, P.L., Poff, L.K., 2012. A conceptual typology of multitasking behavior and polychronicity preferences. *Electron. Int. J. Time Use Res.* 9 (1).
- Hong, J., McArthur, D.P., Livingston, M., 2019. Can accessing the internet while travelling encourage commuters to use public transport regardless of their attitude? *Sustainability* 11 (12), 3281.
- Hubers, C., Dijst, M., Schwanen, T., 2018. The fragmented worker? ICTs, coping strategies and gender differences in the temporal and spatial fragmentation of paid labour. *Time Society* 27 (1), 92–130.
- Klein, N.J., Smart, M.J., 2017. Millennials and car ownership: less money, fewer cars. *Trans. Policy* 53, 20–29.
- Mokhtarian, P.L., Tal, G., 2013. Impacts of ICT on travel behavior: a tapestry of relationships. In: Notteboom, J.R., Shaw, J. (Eds.), *The Sage Handbook of Transport Studies*. SAGE Publications, London, pp. 241–260.
- Perego, A., Perotti, S., Mangiaracina, R., 2011. ICT for logistics and freight transportation: a literature review and research agenda. *Internat. J. Phys. Distrib. Logist. Manag.* 41 (5), 457–483.
- Salomon, I., 1985. Telecommunications and travel: substitution or modified mobility? *J. Trans. Econ. Policy* 19 (3), 219–235.
- Sumalee, A., Ho, H.W., 2018. Smarter and more connected: future intelligent transportation system. *IATSS Res.* 42 (2), 67–71.
- Tal, G., Cohen-Blankshtain, G., 2011. Understanding the role of the forecast-maker in overestimation forecasts of policy impacts: the case of Travel Demand Management policies. *Transport. Res. A Policy Pract.* 45 (5), 389–400.

Further Reading

- Ben-Elia, E., Avineri, E., 2015. Response to travel information: a behavioural review. *Trans. Rev.* 35 (3), 352–377.
- Ben-Elia, E., Zhen, F., 2018. ICT, activity space–time and mobility: new insights, new models, new methodologies. *Transportation* 45 (2), 267–272.
- Ettema, D., 2018. Apps, activities and travel: an conceptual exploration based on activity theory. *Transportation* 45 (2), 273–290.
- Keseru, I., Macharis, C., 2018. Travel-based multitasking: review of the empirical evidence. *Trans. Rev.* 38 (2), 162–183.
- Kim, J., Rasouli, S., Timmermans, H.J., 2018. Social networks, social influence and activity-travel behaviour: a review of models and empirical evidence. *Trans. Rev.* 38 (4), 499–523.
- Lyons, G., 2018. Getting smart about urban mobility—aligning the paradigms of smart and sustainable. *Transport. Res. A Policy Pract.* 115, 4–14.
- Lyons, G., Mokhtarian, P., Dijst, M., Böcker, L., 2018. The dynamics of urban metabolism in the face of digitalization and changing lifestyles: understanding and influencing our cities. *Resour. Conserv. Recy.* 132, 246–257.
- Noy, K., Givoni, M., 2018. Is 'smart mobility' sustainable? Examining the views and beliefs of transport's technological entrepreneurs. *Sustainability* 10 (2), 422.
- Tillema, T., Dijst, M., Schwanen, T., 2010. Face-to-face and electronic communications in maintaining social networks: the influence of geographical and relational distance and of information content. *N. Media Soc.* 12 (6), 965–983.
- Vilhelmson, B., Thulin, E., Eldér, E., 2017. Where does time spent on the Internet come from? Tracing the influence of information and communications technology use on daily activities. *Informat. Commun. Soc.* 20 (2), 250–263.

Mode Choice and Life Events

Joachim Scheiner, Technische Universität Dortmund, Faculty of Spatial Planning, Department of Transport Planning, Dortmund, North Rhine-Westphalia, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	171
Mobility Biographies—Basic Ideas	171
Life Events—Empirical Work	172
Understanding Stability and Change in Mobility Biographies	173
Sorting Levels of Change and Stability	173
Resistance to Change	174
Triggering Change	174
The Role of Socialization in Stability and Change	175
Summary and Conclusions	175
Acknowledgment	176
See Also	176
References	176
Further Reading	177

Introduction

Dynamic approaches to travel behavior have emerged over the past decades that aim to investigate mid- to long-term behavioral change over time. Some of this research looks at the effects of voluntary travel behavior change campaigns on mode choice to evaluate policy programs. Other studies shed light on the development of travel behavior throughout people's life courses. They use labels such as "mobility biographies" or the "life course approach to travel." While these two strands are closely related to one another, this chapter is situated in the mobility biographies literature (Lanzendorf, 2003; Chatterjee and Scheiner, 2015; Clark et al., 2016; Rau and Manton, 2016; Scheiner, 2018). The reasoning behind studying travel behavior change under the notion of the life course or biography is because behavioral changes typically end up in the formation of new routines and therefore may have long-term consequences. What is more, they are embedded in wider social, economic, and spatial settings that are stable in the long-term. Thus, the evolution of travel behavior over time can be seen as an interplay between stability and change.

Mobility biographies research to date has strongly focused on estimating the effects of life events on travel and, more specifically, mode choice, although the approach is much broader than this may suggest. This chapter provides an overview on the role of life events for mode choice change and puts this into a wider perspective of theoretical ideas that are relevant for mobility biographies studies.

The chapter is structured as follows. The next section presents basic ideas of the mobility biographies approach. This is followed by an overview of the empirical work done on the effects of life events on travel. The subsequent section theorizes the emergence of stability and change in mobility biographies, including different levels on which change and stability may occur, the factors that contribute to resistance to change and to triggering change, and the role of socialization in this respect. The paper closes with a summary and conclusions.

Mobility Biographies—Basic Ideas

The mobility biographies approach makes use of the basic ideas of travel theory and aims to develop a life course-oriented, dynamic framework. Life course and biography research have a rich tradition in various disciplines. Time geography provides an important point of origin for mobility biographies, as the life course can be seen as a special kind of space-time path, as developed by time geography. Time geography also inspired early work in the 1980s that categorized travel behavior by life cycle stages. The life cycle perspective was later applied in dynamic transport models such as MIDAS or ALBATROSS.

Mobility biographies studies have closer links to life course research than to biography research. The life course is typically conceived as a sequence of events and role transitions that a person lives through from birth to death. In contrast, a biography is understood as a subject's self-reflective, meaningful action within the temporal structure of his or her own life. Accordingly, life course studies attempt to record and analyse stages and sequences in people's lives that are separated by transitions typically defined by discretionary events. This is also what the overwhelming majority of mobility biographies studies to date do (e.g., Scheiner and Holz-Rau, 2013a; Oakil et al., 2014; Scheiner, 2014; Sharmeen et al., 2014; Zhang et al., 2014; Klein and Smart, 2019). On the other hand, biography studies tend to reconstruct the subjective meanings someone associates with his/her own life. Only relatively few

mobility biographies studies fall into this category (e.g., [Rau and Manton, 2016](#); [Plyushteva and Schwanen, 2018](#)). The term mobility biographies itself was only introduced after the turn of the millennium ([Lanzendorf, 2003](#)) when similar ideas were developed in various places around the world, with a nucleus in central Europe.

A number of key concepts in life course studies have been utilized in mobility biographies studies. The central idea is that lives can be understood as temporal paths or trajectories, and any point on such a path is a consequence of past experiences and decisions made, as well as of future expectations, anticipations and aspirations. The paths are structured by life stages (or phases) that are characterized by social roles and statuses, and their combinations (e.g. employed father, student mother). Changes in social role or status are called transitions, and go along with life events (e.g., entry into school, marriage), also called life course events, or key events.

With respect to interrelations between various decisions made in a life at a certain stage, a hierarchical structure between long-term, mid-term and short-term is commonly assumed, in which long-term decisions shape the conditions and options for future decisions. For instance, long-term residential choice is typically assumed to shape mid-term vehicle ownership choices, that in turn shape short-term daily travel choices ([Lanzendorf, 2003](#)). Suffice to say that reverse relationships may occur on all levels.

Five major elements can be identified that play a key theoretical role in the mobility biographies approach ([Scheiner, 2018](#)). Each is discussed in more detail as follows.

1. **Societal embedding:** Though individual in nature, life courses develop within a societal aggregate. Hence, mobility biographies need to be understood in a wider context of historical circumstances and processes in time and space. This is why they may be cohort specific, rather than universal, and may even be specific for certain sub-groups within a cohort, for example, for men or women. This suggests that different levels of development (society, interpersonal relations, individual . . .) need to be distinguished.
2. **Habits:** Habits are reflected in the routine character of daily travel. They are automatic, non-deliberate behaviors that result in strong behavioral stability over time. Habits are a powerful, but not the only, factor that counters change.
3. **Linked-life domains:** Close relationships exist between individual mobility biographies and other domains of the life course, such as housing, family/household, employment, social networks, health, lifestyle, and even religion.
4. **Life events and learning processes.** Significant changes in mobility are motivated by transitions, events and learning processes over an individual's biography, and associated breaks in routines. A large portion of empirical mobility biographies work focuses on the impact of life events on mode choice and car ownership. The events may be deliberate choices, foreseeable but unavoidable events, or unexpected events that are beyond the control of a person or household. They may be directly related to the life course in terms of a person's private or professional career, or not (such as an accident, or an exogenous, targeted or non-targeted intervention into the transport system). Longer-term, gradual processes have been less studied. They may be due to learning processes and experiences made, as well as changes in needs and aspirations induced by gradually changing life situations, such as declining health status in higher age.
5. **Socialization:** There are links between someone's life course and the life courses of others in their social environment. These links may be studied using terms such as "linked lives" or socialization. As the mobility biographies approach has been developed as an individualist approach, this point has attracted relatively little attention to date. It suggests strong links to recent research on the role of personal social networks for travel including both intra-household (or intra-family) and extra-household interactions. For instance, an individual's mode choice may not only change when this individual retires, but also when her partner retires. What is more, socialization also links the individual to his/her wider social environment and is expressed in different levels of change.

Life Events—Empirical Work

The events that have been studied with respect to their relevance for mode choice include a broad range that can be classified into three categories to a large extent:

1. **Household/family biography:** leaving the parental home; formation of a household with the partner; the birth of a child (first- or higher-order); divorce/separation; move-out of a child
2. **Educational and employment biography:** commencement of job training or university entry; entry into the labor market; change of job, workplace location or education; entry into unemployment; retirement
3. **Spatial mobility:** gaining a driver's license; purchase or disposal of a car or other 'mobility tool'; residential move.

This is of course by no means a full list of possible events. For instance, changes in health status or accidents have received relatively little attention to date. Also, daily life in families with children has its turbulences and sometimes seems to be filled with virtually a continuous succession of events. What is more, the definition of an event and its hypothesized effect on travel strongly depends on whose mobility biography is studied when several people are affected by an event. For instance, when a couple with a teenage child separates, this may affect the father's, the mother's and the adolescent's travel in very different ways.

Some existing studies work with secondary data (e.g. national travel panels) and, thus, the choice of life events tends to be driven by data constraints. But even in targeted surveys it may be difficult to capture rare or unexpected, but possibly significant events (e.g., the purchase of a household dog). Another criterion for the selection of events is motivated by the strong representation of planners

and geographers in this research. Both groups are typically interested in changes in accessibility and the built environment. This may be the reason why moving home has been the life event that has probably been researched the most to date (Scheiner and Holz-Rau, 2013b; Zhang et al., 2014; Lee and Goulias, 2018). The reason why residential relocation can be expected to affect travel behavior is because access to opportunities is likely to change. Such opportunities include activity places like the workplace, shopping and leisure facilities, and transport provision. This suggests that two features of relocation are important (Scheiner and Holz-Rau, 2013b; Chatterjee and Scheiner, 2015).

Firstly, the combination of the pre-move home location and the post-move home location is of great importance, as a change in accessibility is as much influenced by neighborhood/location attributes at the former place of residence as those of the new place. Secondly, distance of the move needs to be considered, as it indicates the extent to which existing activity places can be maintained after the move. Long-distance migration might lead to more long-distance travel to visit relatives, whereas the former workplace is not likely to be maintained (in many cases a workplace change will have triggered the long-distance relocation). In double-income households, relocating closer to the workplace of one partner might lead to longer commutes for the other. It should be noted that home moves often coincide with other life events such as the birth of a child or workplace changes.

With respect to outcomes, research has shown that moves from inner cities to suburbia tend to be followed by increases in car use and decreases in public transport use, cycling and walking. The opposite is true for relocations into the city (Scheiner and Holz-Rau, 2013b). The changes may be stronger for car and public transport than for cycling and walking, but are contingent on geographical context. Other studies found less change in mode choice after relocating, particularly in places where people heavily rely on cars irrespective of residential location.

Concerning other life events, results are mixed (e.g., Scheiner and Holz-Rau, 2013a; Scheiner, 2014). Birth of a child tends to be associated with an increase in car ownership, driving and walking (the latter being more true for women than for men). Founding a couple household tends to increase car use as a passenger, suggesting shared trips. Entering the labor market tends to result in more driving and less trip-making as a car passenger, on foot or by bicycle, suggesting tighter time budgets and increasing income. Generally, the effects of life events on general mode choice appear modest, and they vary widely between individuals. They may be somewhat stronger when more specific travel indicators are studied, such as commute mode. Significant effects on changes in mode choice have also been found for baseline variables (such as household composition and urban context) (Scheiner and Holz-Rau, 2013a). This may reflect continuous adaptation to changing circumstances or/and lagged responses to life events. Some direct evidence of anticipated (lead) and delayed (lagged) effects of life events is also available (Oakil et al., 2014).

Two events that have been found to be most influential relate to acquiring the means for mobility (or “mobility tools”): a driving licence and a car. These could be regarded both as life events that affect mode choice (“independent variables”) or dimensions of travel itself that are influenced by (demographic, socioeconomic, or environmental) life events. Both events may be understood as relatively stable pre-decision characteristics of mobility.

Taken together, changes in car ownership and mode choice seem to be more common among those experiencing life events than those who do not. As travel variables are typically used as the outcomes in related research, it is less clear whether car ownership or mode choice decisions have an influence themselves on future life development, such as economic or social integration. An exception are studies on the elderly that tend to find that driving cessation results in negative well-being, although the results are inconclusive (Scheiner, 2006; Chihuri et al., 2015).

Understanding Stability and Change in Mobility Biographies

Sorting Levels of Change and Stability

Distinct levels of analysis can be identified in the literature on theories of change, with authors either focusing on individual (behavior, attitude) change or on systems change (system transformation, macro social change). In between these two extremes there are theories of organizational change that typically look at enterprises or administration. Individual travel behavior change is the main interest of mobility biographies research, though interactions with other individuals, as well as organizations and societies, need to be taken into account. Looking at the span between individual and societal levels, a classification of theories may be as follows.

1. Individual level theories. These theories include “classical” psychological theories that are largely limited to the individual.
2. Interpersonal behavioral theories. These theories include focus on the role of interactions, social embedding, role models, or mentoring.
3. Community theories of behavior. Such theories relate to groups, organizations, social institutions, and communities.
4. Systems transformation theories. They relate to society, politics or/and the economy on the national or global level as a whole.

This chapter cannot discuss all these theories in detail. It is important, however, to be clear that the different levels must be linked. There cannot be communities or organizations without individuals, but on the other hand individual action cannot be adequately understood without taking into account the organization of individuals in social groups and economic, administrative and political institutions. This can best be described as a reciprocal relationship. The links may be shaped in myriads of ways, and they are on multiple levels, not just on two or three.

Resistance to Change

Before turning to change it is important to understand why mode choice appears to be quite resistant to change over time. The most widespread theoretical concept used to explain why behavioral change may not occur is habit. People develop habits based on internal scripts which they automatically retrieve to reproduce action without much effort. Habits work as behavioral “recipes” or behavioral rules that can easily be applied in situations that are experienced as similar to other, previously experienced situations. Habits result in a simple form of path dependency—a good predictor of someone’s behavior tomorrow is his/her behavior today. (Manifest) habits are visible in repeated, routine behavior. Conversely, however, repeated behavior does not in itself provide evidence of habits being at work. People may repeat their behavior because they have found an individual optimum solution and have no reason to change.

Habits resemble the idea of heuristics (or shortcuts) in that they work as behavioral “recipes” to reduce complexity and uncertainty. Heuristics are however not necessarily linked to “automatic” behavior, such as habits, but may be used in a conscious, “controlled” way. Travel behavior is a very complex form of behavior, and heuristics are simplifications of situations. They are stereotypically applied as long as they practically work, irrespective of whether or not the information processed meets reality. As such, they tend to stabilize behavior, rather than contributing to change. When applied in slightly varied situations in which behavior meets unexpected reactions (e.g., following a life event such as residential relocation), precisely the application of heuristics may result in change. On the other hand, in exactly such situations circumstances may prevent people from using “recipes.” This may motivate them to develop new or revised heuristics.

The use of habits and heuristics are relatively universal ways of acting. Resistance to change may also be a personality trait that is specific to some people, but not all. Risk aversion and aversion of regret that may occur in traveling due to choices made under uncertainty (e.g., one does not know whether an alternative mode or route is as beneficial as it appears) may lead to “choice inertia,” that is, resistance to change. Similarly, resistance to change may be caused by threats to an individual’s identity. This is when identity is constructed around symbols (e.g., a car), beliefs or activities associated with a particular behavior (e.g., driving). Also, lack of self-efficacy can cause resistance to change when someone does not believe that (s)he is able to achieve behavioral change or the desired outcomes.

A final, perhaps most important obstacle for mode choice change is lack of motivation. Even in the case of pathological behaviors (e.g. drug abuse) strong patient motivation and incentives are needed to achieve change. Mode choice is not pathological but rather has strong rational and normative foundations. This points to one of the major shortcomings of applying voluntary change policies in the transport or sustainability sector. It is questionable whether strategies such as education, information, reminders or prompts can be incentives as long as there is no suffering on the individual level. The ‘suffering’ in mode choice is mostly on the societal level, while on the individual level travel choices are the outcome of reasonable considerations made under certain circumstances and needs. In terms of protection motivation theory: someone may be convinced of being capable of behavior change towards sustainability (high self-efficacy), but it is unlikely that (s)he is convinced that his/her behavior change substantially affects sustainability in general (low expectancy of threat decrease). The individualist rationality in mode choice also highlights the importance of the quality of the choice alternatives. If people are provided with, say, free public transport tickets, and opt to use them but learn that public transport fails to satisfy their needs, they will hardly become permanent public transport customers.

Last, but not least, it shall only be touched upon here that persistence of behavior can also be due to a higher, system level that prevents change, as multi-level transition theory argues. The concept of regimes has been used to understand the forces that prevent change on these higher levels of political, economic, or sociotechnical systems. This has mostly been discussed in transition theories, but less so in micro-level theories such as the mobility biographies approach.

Triggering Change

The most frequently researched idea in mobility biographies studies is that changes in social role or status and associated life events trigger change in mode choice. This idea suggests that different life domains are linked to travel changes, as life events are typically linked to one or more domains other than mobility. This may result in another form of path dependency. Long-term travel evolution may be the outcome of choices made in another life domain, for example, car dependence based on residential choice. However, these considerations do not answer the question of *why* life events affect travel.

It can be argued that life events (or their anticipation) change people’s life situation, social roles and, sometimes, conditions for mobility. Habits do not work anymore because travel requirements, opportunities and/or abilities (“ROA model”) may have changed (Busch-Geertsema and Lanzendorf, 2015). Travel behavior is reconsidered, and deliberate decision making becomes necessary. An important conceptual trigger in this respect is the emergence of stress (Clark et al., 2016). The theoretical mechanism is that life events produce a discrepancy between current needs or aspirations (e.g., for car ownership) and actual (car ownership) state, and this discrepancy produces stress, imbalance and reassessment. Finally, either attitudes or behavior (or both) will be adjusted.

It may be important to note here that transport researchers tend to highlight the adjustment of behavior as a result of such stress. However, the attitude-behavior link has never seriously been conceptualized as a unidirectional link. Rather, stress may be reduced by adjusting attitudes to existing behavior, and this way of adjustment is probably more common than the adjustment of behavior to pre-defined attitudes (Kroesen et al., 2017).

In order to develop behavioral intentions and, ultimately, change behavior, people need to be motivated. Hence, there is a need to introduce motivation and the question of how motivation changes into theories of behavioral change. On the other hand, motivation alone would not drive action if there was no expectation of success. The term self-efficacy captures whether someone expects to be able to successfully perform certain behavior. Similarly, the theory of planned behavior, that has been used frequently in transport studies, includes perceived behavioral control as a predictor of behavior (mediated by behavioral intention).

The Role of Socialization in Stability and Change

Individual life courses are not isolated from other people's life courses. They are embedded in social structures on the personal level, in family, kin, friendship, and neighborhood networks, as well as in economic, political, or administrative organizations. The term socialization refers to the integration of individuals in the society over the course of their lives by means of learning from significant others who work as socialization agents. Socialization has been proposed as a mechanism through which social norms regarding behaviors are transmitted to someone being socialized via these agents. Socialization may also be understood as mutual interaction that enables group integration. This may refer to a small group, such as a family or a clique, or society as a whole. Typical socialization agents are parents and the family, peer groups, media, or schools (Döring et al., 2019).

Socialization is not limited to childhood and adolescence (although these are the most formative stages), but rather spans the whole life course. The knowledge, pre-dispositions, attitudes or behavior that is transmitted are 'the normal' that needs to be known or done to be part of the group. Mobility socialization is thus the process of transmitting mobility as a norm, and this process makes the individual part of the mobile society.

Socialization thus typically works against change in the aggregate. It might be understood as sort of a "habit on the aggregate (or system) level." This may be a major obstacle to implementing voluntary change. To take an example: Why should someone stop driving while most others around him or her continue to do so? On the other hand, if someone feels that his or her behavior represents only a small minority, (s)he may adjust to the majority and, thus, change. For practical applications this means that if a concept succeeds in changing many people's behavior, it is likely that even those who strongly resist changing may finally adapt.

There is evidence for the impact of parents, spouses, schools, neighbors, collective urban mobility cultures, cohorts, and earlier peer groups (e.g., mobility cultures at previous places of residence) on mode choice. This research generally finds positive behavioral links, suggesting that conformity in behavior predominates over non-conformity (Sunitiyoso et al., 2013). This suggests that socialization is a powerful force that may inhibit mode choice change following life events. However, over and above behavioral correlation, the idea of mobility socialization presupposes that mobility is a relevant phenomenon for group integration; otherwise there would be no pressure to adjust mobility. Experimental results suggest strategic behaviors that follow other group members and conformity to the majority choice, and this suggests socialization has a stabilizing function (Sunitiyoso et al., 2013).

Summary and Conclusions

The mobility biographies approach and, more specifically, research on mode choice changes as a result of life events, has grown into a substantial research field. Psychological and sociological theorizing helps understand why these changes may occur, but also why mode choice may be stable in the long term, even notwithstanding massive changes in someone's life circumstances.

Perhaps the biggest obstacle to applying models that have their roots in psychotherapy to mode choice change is that there is typically little individual suffering in mode choice (as opposed, e.g., to drug abuse). Hence, one may ask where motivation for change should come from. The theoretical mechanism that links life events to mode choice change is stress. Life events may produce a discrepancy between current travel needs and actual state, and this discrepancy produces stress and reassessment and, perhaps, behavior change. However, people need to feel able to change, and they need to have some expectation of benefit in order to start a behavioral change process.

Some broad key issues that may be highlighted as steps forward in the mobility biography literature are as follows.

1. *Develop and test theories.* This chapter has touched on multiple theoretical concepts that may help understand behavioral stability and change over the life course, including habits, motivation, attitudes, expected benefits, self-efficacy, and the interaction of individuals with close others, as well as with the wider social, economic, and spatial environment on various levels. These ideas need to be elaborated upon in more depth to integrate them into a mobility biography framework.
2. *Link quantitative and qualitative approaches.* There is a general focus on quantitative studies in mobility biographies research, and more qualitative work is needed. Linking the two can provide mutual enrichment, which is perhaps even more important. This could also help bridge the gap between life course and biography studies. Recent attempts to add multiple layers of meaning to the somewhat "positivist" empiricist approaches to mobility biographies may be a start.
3. *The social embedding of travel.* Rich research has developed in recent years about personal social networks (including intra-household interactions) and travel. This research has few links as yet to life course and biography perspectives, though the need to understand social networks and travel in a longitudinal sense has clearly been recognised. Future research should also include socialization effects and the social embedding of travel in a wider sense, linking individual behavior over the life course with the behavior of collectives and organizations in policy, the economy or the neighborhood or city of residence, and looking at the role

of social norms over time. The interaction between policy measures and their acceptability and responses on the individual level has been found to change strongly over time, and more research is needed on this topic to support a sustainable transition of travel.

4. *Interactions between life domains, behavioral dimensions, and population groups.* There are also multiple interactions between various life domains, events and developments over the life course. It is not clear yet how multiple events and changes work together to shape mode choice change. This also refers to the sequence in which changes occur, including lead and lagged effects that have clearly been recognized, but rarely studied. Interrelations between various dimensions of travel behavior have been studied in transport research, but hardly in a longitudinal perspective. Previous mobility biographies research has sought to understand direct links between key events and mode choice changes, for instance, but these links may be moderated by changes in activity patterns, destination choice and associated distances, and other variables. Finally, there is a need for dedicated studies of certain population groups that are of special importance to research or policy. For instance, there are few mobility biographies studies on immigrants, the poor, or the elderly.
5. *Develop and test policy approaches.* It is an important task to develop policy approaches that take knowledge from mobility biographies studies into account. Information campaigns, trial tickets and the like that may be linked to life course events such as residential or workplace relocation, the birth of a child, or entry into retirement have proven valuable. On the other hand, policies can make use of socialization effects. Peers and other socialization agents can help target people's travel. "Snowball effects" (effects may be stronger in the long run than in the short run) may increase the effectiveness of concepts. Policies should take path dependencies in people's life courses into account. Early decisions on the locations of residence, education, and workplace can have lifelong consequences for travel behavior, and they can even affect subsequent generations in a family. Policies should also recognize long-term stability in people's travel behavior and recognize that people may adapt their travel slowly and only to a minor extent, even if circumstances in the environment change dramatically.

Acknowledgment

This research was funded by the German Research Foundation (DFG) as part of the project "Veränderungen der Mobilität im Lebenslauf: Die Bedeutung biografischer und erreichbarkeitsbezogener Schlüsselereignisse" (Travel behavior changes over the life course: the role of biographical and accessibility-related key events, 2015–2020).

See Also

Commuting, the Labor Market, and Wages; Choice Models in Transportation; Residential Location Choice Models; Vehicle Ownership Models; Travel Mode Choice as Reasoned Action; Transport Modes and Commuters; Shopping and Transport Modes; Transport Modes and Health; Next Generation Travel: Young Adults' Travel Patterns; Car Ownership and Car Use: A Psychological Perspective; Women and Transport Modes; Transport Modes and Inequalities

References

- Busch-Geertsema, A., Lanzendorf, M., 2015. Mode Decisions and Context Change – What About the Attitudes? A Conceptual Framework. In: Attard, M., Shifan, Y. (Eds.), *Sustainable Urban Transport*. Transport and Sustainability 7. Emerald, Bradford, pp. 23–42.
- Chatterjee, Kiron, Scheiner, Joachim, 2015. Understanding changing travel behaviour over the life course: Contributions from biographical research. A resource paper for the Workshop "Life-Oriented Approach for Transportation Studies". Presented at IATBR (14th International Conference on Travel Behaviour Research), Windsor, UK, 19–23 July 2015.
- Chihuri, S., Mielenz, T.J., DiMaggio, C.J., Betz, M.E., DiGuseppi, C., Jones, V.C., Li, G., 2015. Driving Cessation and Health Outcomes in Older Adults: A LongRoad Study. AAA Foundation for Traffic Safety, Washington, D.C.
- Clark, B., Lyons, G., Chatterjee, K., 2016. Understanding the process that gives rise to household car ownership level changes. *J. Trans. Geog.* 55, 110–120.
- Döring, L., Kroesen, M., Holz-Rau, C., 2019. The role of parents' mobility behavior for dynamics in car availability and commute mode use. *Transportation* 46 (3), 957–994.
- Klein, N., Smart, M., 2019. Life events, poverty and car ownership in the United States: a mobility biography approach. *J. Trans Land Use* 12 (1), 395–418.
- Kroesen, M., Handy, S., Chorus, C., 2017. Do attitudes cause behavior or vice versa? An alternative conceptualization of the attitude-behavior relationship in travel behavior modeling. *Transp. Res. Part A* 101, 190–202.
- Lanzendorf, M. 2003. Mobility Biographies. A New Perspective for Understanding Travel Behaviour. Paper presented at the 10th International Conference on Travel Behaviour Research (IATBR), Lucerne, 10–15 August 2003.
- Lee, J.H., Goulias, K.G., 2018. A decade of dynamics of residential location, car ownership, activity, travel and land use in the Seattle metropolitan region. *Trans. Res. Part A* 114B, 272–287.
- Oakil, A.T.M., Ettema, D., Arentze, T., Timmermans, H., 2014. Changing household car ownership level and life cycle events: an action in anticipation or an action on occurrence. *Transportation* 41 (4), 889–904.
- Plyushteva, A., Schwanen, T., 2018. Care-related journeys over the life course: thinking mobility biographies with gender, care and the household. *Geoforum* 97, 131–141.
- Rau, H., Manton, R., 2016. Life events and mobility milestones: advances in mobility biography theory and research. *J. Trans. Geog.* 52, 51–60.
- Scheiner, J., 2006. Does the car make elderly people happy and mobile? settlement structures, car availability and leisure mobility of the elderly. *Eur. J. Trans. Infrastruct. Res.* 6 (2), 151–172.
- Scheiner, J., 2014. Gendered key events in the life course: effects on changes in travel mode choice over time. *J. Trans. Geog.* 37, 47–60.
- Scheiner, J., 2018. Why is there change in travel behaviour? in search of a theoretical framework for mobility biographies. *Erdkunde* 72 (1), 41–62.

- Scheiner, J., Holz-Rau, C., 2013a. A comprehensive study of life course, cohort, and period effects on changes in travel mode use. *Transp. Res. Part A* 47, 167–181.
- Scheiner, J., Holz-Rau, C., 2013b. Changes in travel mode choice after residential relocation: a contribution to mobility biographies. *Transportation* 40 (2), 431–458.
- Sharmeen, F., Arentze, T., Timmermans, H., 2014. An analysis of the dynamics of activity and travel needs in response to social network evolution and life-cycle events: a structural equation model. *Trans. Res. Part A* 59, 159–171.
- Sunitiyoso, Y., Avineri, E., Chatterjee, K., 2013. Dynamic modelling of travellers' social interactions and social learning. *J. Trans. Geog.* 31, 258–266.
- Zhang, J., Yu, B., Chikaraishi, M., 2014. Interdependences between household residential and car ownership behavior: a life history analysis. *J. Trans. Geog.* 34, 165–174.

Further Reading

- Axhausen, K.W., 2007. Activity spaces, biographies, social networks and their welfare gains and externalities: some hypotheses and empirical results. *Mobilities* 2 (1), 15–36.
- Calastri, C., Crastes dit Sourd, R., Hess, S. (in print): We want it all: experiences from a survey seeking to capture social network structures, lifetime events and short-term travel and activity planning. *Transportation*. (in print, DOI: 10.1007/s11116-018-9858-7).
- Scheiner, J., Rau, H. (eds.) 2020, in preparation. *Mobility Across the Life Course. A Dialogue between Qualitative and Quantitative Research Approaches*. Cheltenham: Edward Elgar.
- Smart, M.J., Klein, N.J., 2018. Remembrance of cars and buses past: how prior life experiences influence travel. *J. Plan. Edu. Res.* 38 (2), 139–151.

Introduction to Air Transport

Milan Janić, Department of Transport & Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	178
The Sub-Systems of Air Transport System	179
Airports	179
Infrastructure	179
Demand/Capacity and Quality of Services	179
Economics	181
Airlines	181
Infrastructure—Air Route Networks and Aircraft Fleet	181
Demand, Capacity, and Quality of Services	184
Economics	184
ATC/ATM Subsystem	185
Infrastructure, Facilities, and Equipment	185
Demand, Capacity, and Quality of Services	186
Economics	187
Sustainability of Air Transport System	188
General	188
Fuel Consumption and Emissions of GHG	188
Land Use	188
Congestion and Delays	189
Noise	189
Air Traffic Incidents/Accidents (Safety)	189
Contribution to the Social-Economic Welfare	190
Conclusions	190
Relevant Websites	190
References	190
Further Reading	191

Introduction

Air transport as the system consists of three main subsystems - airports, airlines, and ATC/ATM (Air Traffic Control / Management). Airports represent the system's ground-based infrastructure consisting of the airside and landside area. Airlines operate the air route networks with flights connecting airports carried out by different aircraft types. The ATC/ATM subsystem embraces the controlled airspace as some kind of the "infrastructure" and the controlling and monitoring facilities and equipment. These all provide capacity to serve demand. For airports, the airline flights, air passengers, and freight/cargo shipments represent demand. For airlines, these are air passengers, and freight/cargo shipments. For the ATC/ATM, the airline flights represent demand. In addition, the air transport system contains the nonphysical components in the format of the operational rules and procedures and information (messages) regulating operations within and among the particular components/subsystems. **Fig. 1** shows the simplified scheme (Janić, 2019). In general, the air passengers and freight/cargo shipments represent the demand for airports and airlines.

The air transport system has been one of the fastest growing sectors of the world's economy. During the period 1991–2007, the air passenger and freight/cargo traffic in terms of Revenue Passenger Kilometers (RPK) and Revenue Ton Kilometer (RTK) (RPK (Revenue Passenger Kilometers) = Number of Transported Passengers · Kilometers Flown; RTK (Revenue Ton Kilometers) = Quantity of Transported Freight · Kilometers Flown) had grown at an average annual rate of 4.5% and 5.7%, respectively. In the year 2007, the volumes had reached about 4.5 trillion RPK and 0.6 trillion RTK. Latter during the period 2008–17, the volumes of RPK and RTK have continued to grow driven by both external and internal driving forces at an average annual rate of 4–5% and 5.5–6.5% and reached about 7.7 and 0.94 trillion, respectively, in the year 2017 (ICAO, 2005/2018). The main external driving forces have been the overall economic growth (gross domestic product (GDP), per capita income (PCI)), globalization of trade, foreign investments, and tourism. The main internal driving forces have included liberalization of the air transport markets driving re-consolidation of the airline industry into two large airline categories (the conventional (legacy) and low cost carrier(s) (LCCs) offering generally lower more attractive airfares. Both types of driving forces have generally resisted to the compromising ones such have been the various large-scale disruptive events (severe weather, September 11, 2001 terrorist attack on the U.S., regional wars—

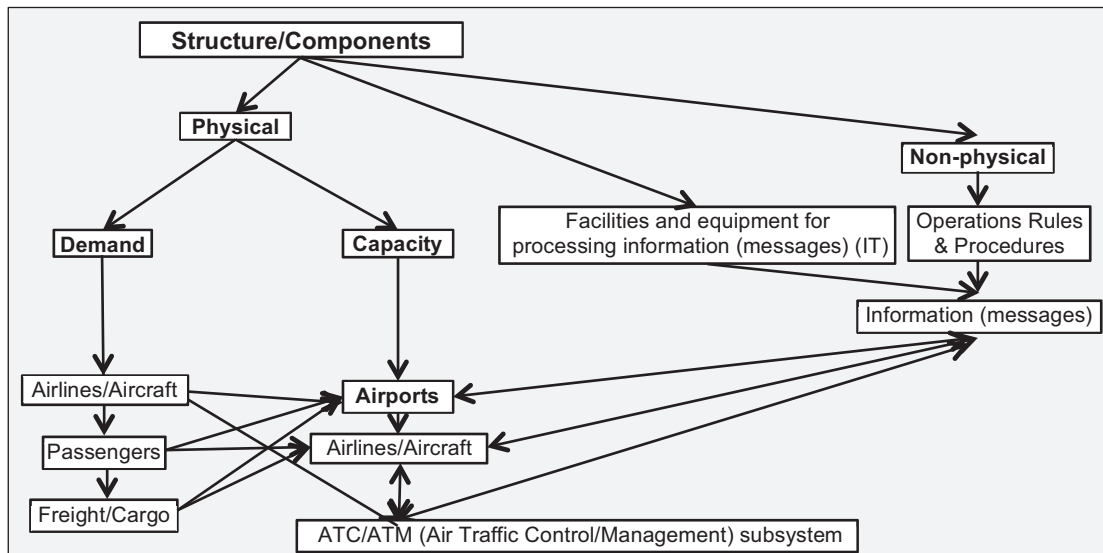


Figure 1 Main components of the air transport system (Janić, 2019).

Iraq, Afghanistan, Syria), the global economic and political crises (2008), and global epidemic diseases (ACI, 2006; AIRBUS, 2006; Boeing, 2007; ICAO, 2005/2008). In addition, the impacts of the air transport system on the society and environment have been increasing, thus requiring undertaking the measures for its more sustainable development in the forthcoming medium- to long-term period (Janić, 2007, 2009).

The Sub-Systems of Air Transport System

Airports

Infrastructure

The Airport Council International (ACI) has reported that 17,678 commercial airports worldwide have handled air passenger, freight/cargo shipments, and business aircraft (ACI, 2018; <https://store.aci.aero/product/annual-world-airport-traffic-report-2018/>). The number of international airports has been 1282 of which 163 in Africa, 311 in South and North America, 319 in Asia, 498 in Europe, and 51 in Oceania (<https://www.bts.gov/content/number-us-airports/>; https://en.wikipedia.org/wiki/List_of_international_airports_by_country#Number_of_International_Airports).

Each airport consists of the airside and the landside area. The airside area includes runways, taxiways, and apron/gate complex. The runways enable the aircraft landing and taking-offs. The taxiways enable the aircraft movements between the runways and the apron/gate complex. The apron/gate complex provides the space-parking stands for the aircraft ground handling, which includes embarking, and disembarking air passengers, their baggage and freight/cargo shipments after landing, and embarking these after the aircraft refueling and before taking-offs. The landside area consists of the passenger and freight/cargo terminal(s) and the landside access modes and their systems (De Neufville and Odoni, 2013). Fig. 2 shows the simplified scheme of an airport layout.

The passenger and freight/cargo terminals can have different layout and size depending on the volumes of corresponding demand to be served during the specified period of time (hour, day, year) under given conditions. The road- and rail-based landside access modes provide the airport landside accessibility for air passengers, aviation employees, greeters, and others (Janić, 2019). The former mode includes the car and van, and bus systems. The latter mode includes the streetcar/tramway, Light Rail Transit (LRT), subway/metro (short distance), regional/conventional rail and High Speed Rail (HSR) (medium and long distance), Trans Rapid Maglev (TRM) (currently short distance), and most recently the rather futuristic Hyperloop (HL) (long distance) system. These all connect airports with their catchment areas (frequently these are the city centers or Central Business District(s) (CDBs).

Demand/Capacity and Quality of Services

The airport demand consists of the atm (aircraft movements) (1 atm = 1 arrival or 1 departure) in the airport airside area, and the air passengers and freight/cargo shipments in the landside area. The airport capacity is usually defined as the maximum number of units of demand, which can be accommodated in the corresponding area(s) during a given period of time under given conditions (hour, day, year). These conditions are either 'the constant demand for service' ("ultimate" capacity) or the 'maximum average delay' ("practical" or "declared" capacity). The total capacity of airside area depends on the capacity of runway system, taxiways, and apron/gate complex (De Neufville and Odoni, 2013). These capacities should be in balance in order to avoid "bottlenecks" and consequent aircraft/flight delays, additional fuel consumption and related emissions of GHG. An example of the relationship between the

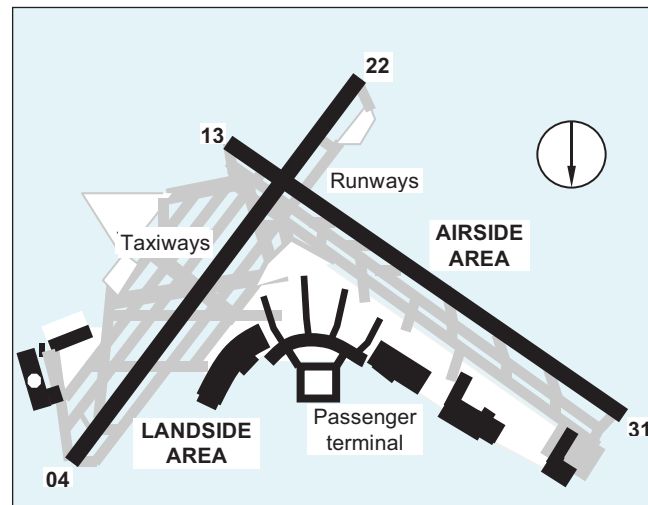


Figure 2 Airport airside and landside area: Case of LGA (La Guardia Airport) (New York, U.S.) (Janić, 2007).

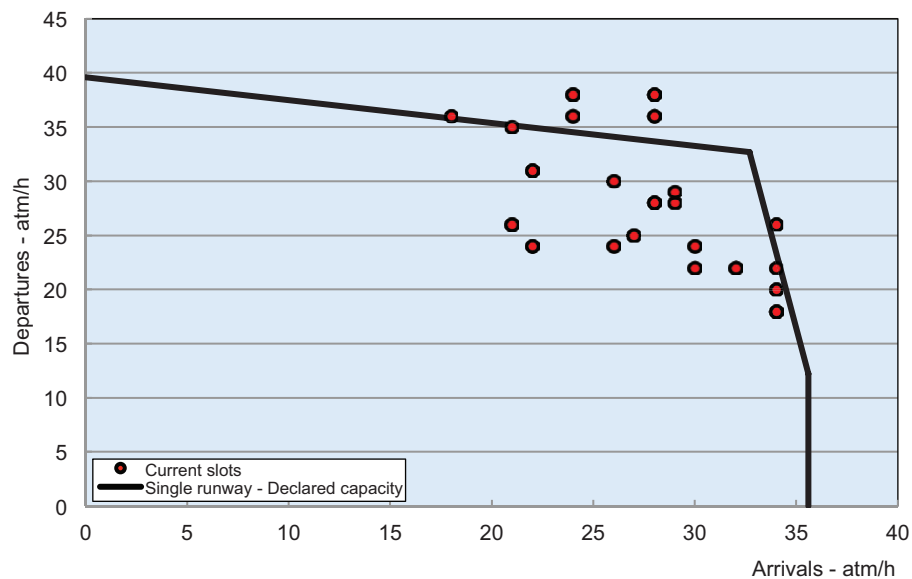


Figure 3 Relationship between the runway systems' "declared" capacity and demand: Case of Dubai International Airport (Dubai, UAE).

runway system "declared" capacity in the form of the "capacity coverage curve" and demand in terms of atms is shown in Fig. 3 (Janić, 2009).

As can be seen, the arrival and departure capacity based on the instrumental flight rules (IFR) are dependent on each other. In some cases, the number of slots exceeds the capacity coverage curve implying that the both arrival and departure aircraft/flight delays can be expected.

The landside area capacity includes that of the passenger and freight/cargo terminals and the landside access modes and their systems. These capacities should also be in balance. They are expressed by the maximum number of served units during the specified period of time under given conditions and similarly as above can be the "ultimate" and "practical." The landside demand appears relevant for operation and planning of the capacity of above-mentioned components of the landside area. Specifically, the annual passenger demand can be used for the medium- to long-term strategic planning of the passenger terminal capacity as shown in Fig. 4 (Janić, 2009).

As can be seen, three solutions for expanding the passenger terminal capacity are proposed to match the prospectively growing demand during the medium- to long-term period.

At the global scale, the quality of services at airports can generally be expressed by their average annual on-time performance, that is, punctuality, defined as the proportion of arriving and departing flights operated within 15 min of their scheduled arrival and departure times (OAG, 2019). In some way, this can generally reflect the quality of services, which airports provide to both airlines

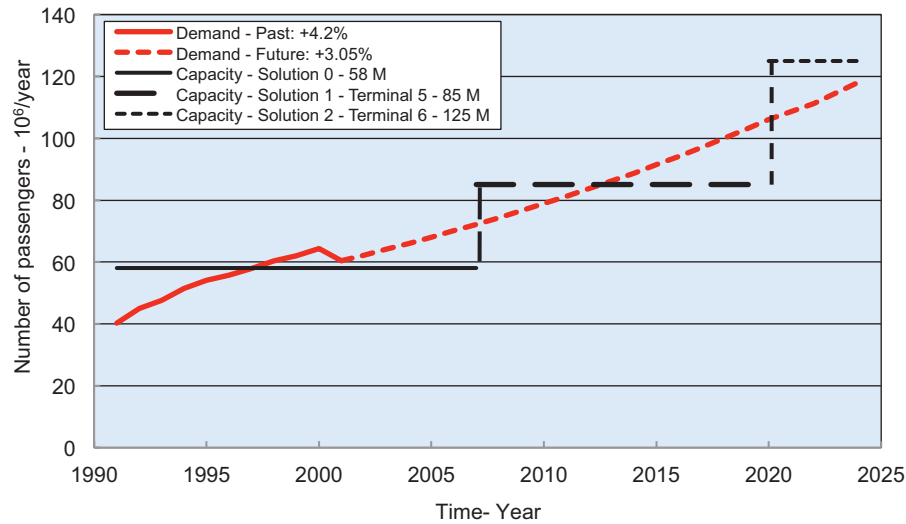


Figure 4 Matching the passenger terminal capacity to demand: Case of Heathrow airport (London, UK) (Janić, 2009).

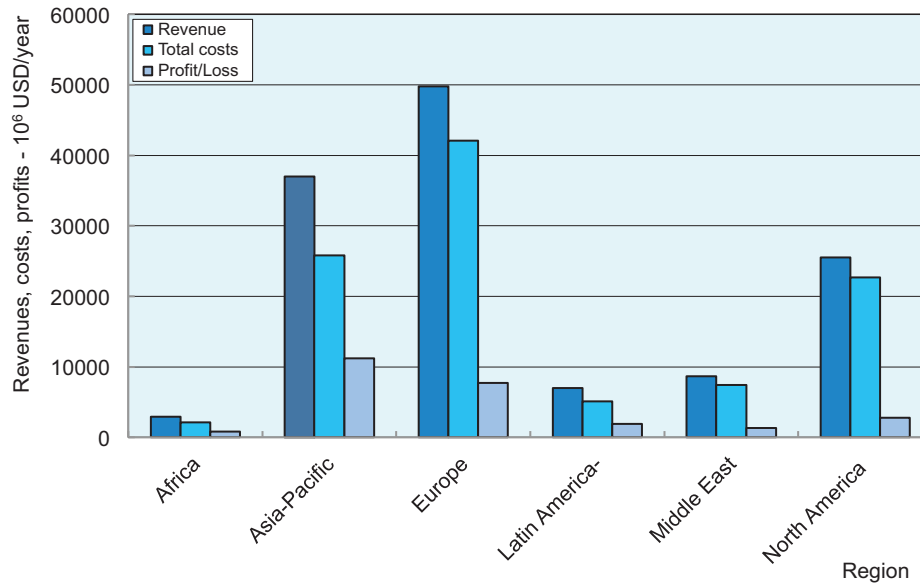


Figure 5 The total revenues, costs, and profits of the world's airports (Period: 2013) (ICAO, 2015).

and air passengers. Developments so far have indicated that the average annual on-time performance has generally decreased with increasing of the volumes of airport traffic.

Economics

The airport economics relate to their revenues, costs, and profits. The revenues can be roughly categorized as the aeronautical obtained from handling airlines, air passengers, and freight/cargo shipments, and the non-aeronautical obtained from some other commercial (non-operating) activities. Fig. 5 shows these for the airports in the particular world's regions (ICAO, 2015). As can be noticed, the profits were higher than the costs in all regions, which indicate that the airports worldwide were generally profitable.

Airlines

Infrastructure–Air Route Networks and Aircraft Fleet

The commercial airlines and their alliances constitute the airline industry. At present, about 5000 airlines operate in the world. However, only 153 national/flag and 146 LCCs are the International Civil Aviation Organization (ICAO) coded. Of these, almost 290 are the members of International Air Transport Association (IATA). During re-consolidation, that is, while adapting to

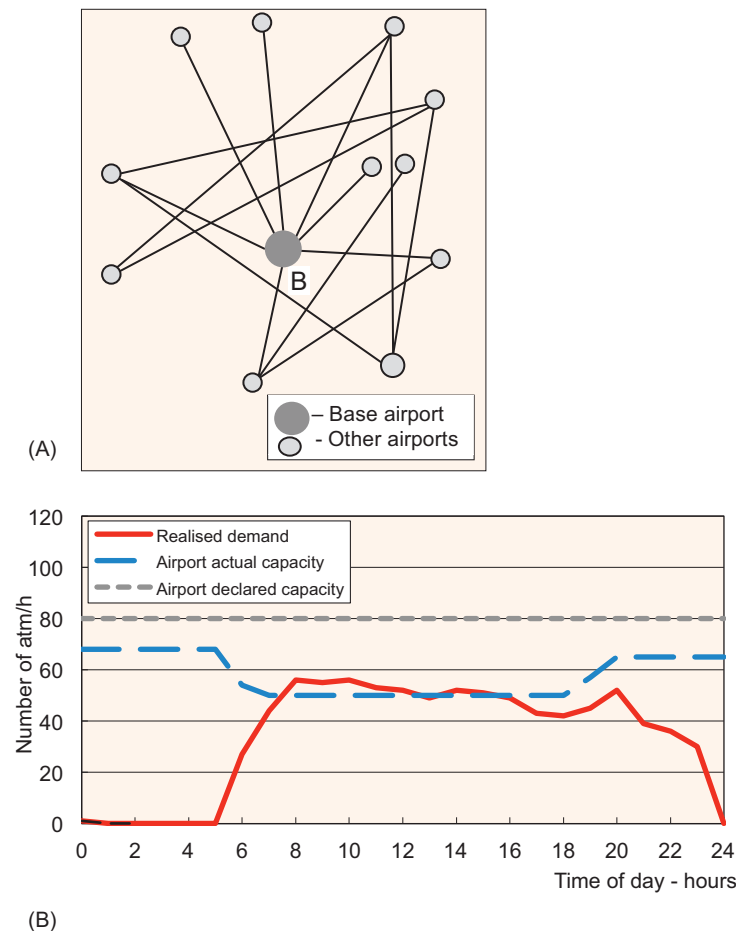


Figure 6 Typical spatial configuration of the airline “point-to-point” network(s) and traffic pattern at one of airports. (A) Spatial configuration; (B) Traffic pattern at base airport.

deregulation and liberalization of the air transport markets in combination with facing to growing demand, most airlines have developed different types of air route networks and business models enabling their categorization. The full cost or the network (legacy) airlines and their alliances have developed the large (global) “hub-and-spoke” networks with one or few centrally located “hub” airports. At these “hubs,” the relatively large number of short-medium, and long-haul incoming and outgoing interconnected flights to/from the other “spoke” airports are scheduled during the relatively short period of time (one or couple of hours). As such, they form one or several flight “waves” at the hub airports during the day enabling transfer of passengers between the particular incoming and outgoing flights in the relatively short time windows. Consequently, the shares of transfer passengers at hub airports have substantively increased. The LCCs have developed the “point-to-point” networks consisting of the short- and the medium-haul routes connecting the regional airports and their base airports, which have also been large. They have mainly operated the fleets consisting of single aircraft type. Fig. 6A, B shows the typical spatial configuration and flight/traffic patterns for the “point-to-point” network(s) (Janić, 2009).

Fig. 6A shows that in the “point-to-point” network(s) the airline base airport is connected by most other airports and that these other airports can be mutually connected. Fig. 6B shows that at the airline can schedule the number of flights also above the airport “declared” capacity implicitly incorporating the expected delays in the scheduling. Fig. 7 (A, B) shows the typical spatial configuration and the flight/traffic patterns of the “hub-and-spoke” network (FAA, 2007).

Fig. 7A shows that the spoke airports are connected to the hub and not between themselves. Fig. 7B shows that the flight scheduling at the hub airport in terms of the relative relationship between the numbers of scheduled departure and arrival flights and the airport capacity during the day. As can be seen, the flights form the arrival and departure “waves” in which the demand/capacity ratio has exceeded the value of one thus implying occurrence of the significant queues and delays, just caused by the airline flight scheduling. Consequently, the airlines have already internalized these delays during design of their schedule.

Operating the above-mentioned networks has inherently required deployment of the larger and more diverse aircraft fleet. Consequently, over the past decade and half, the total number of commercial aircraft has increased from 20,356 in the year 2005 to 29,936 in the year 2017, that is, for 47%, that is, it has grown at an average annual rate of 3.9%/year (AIRBUS, 2018; Boeing, 2017;

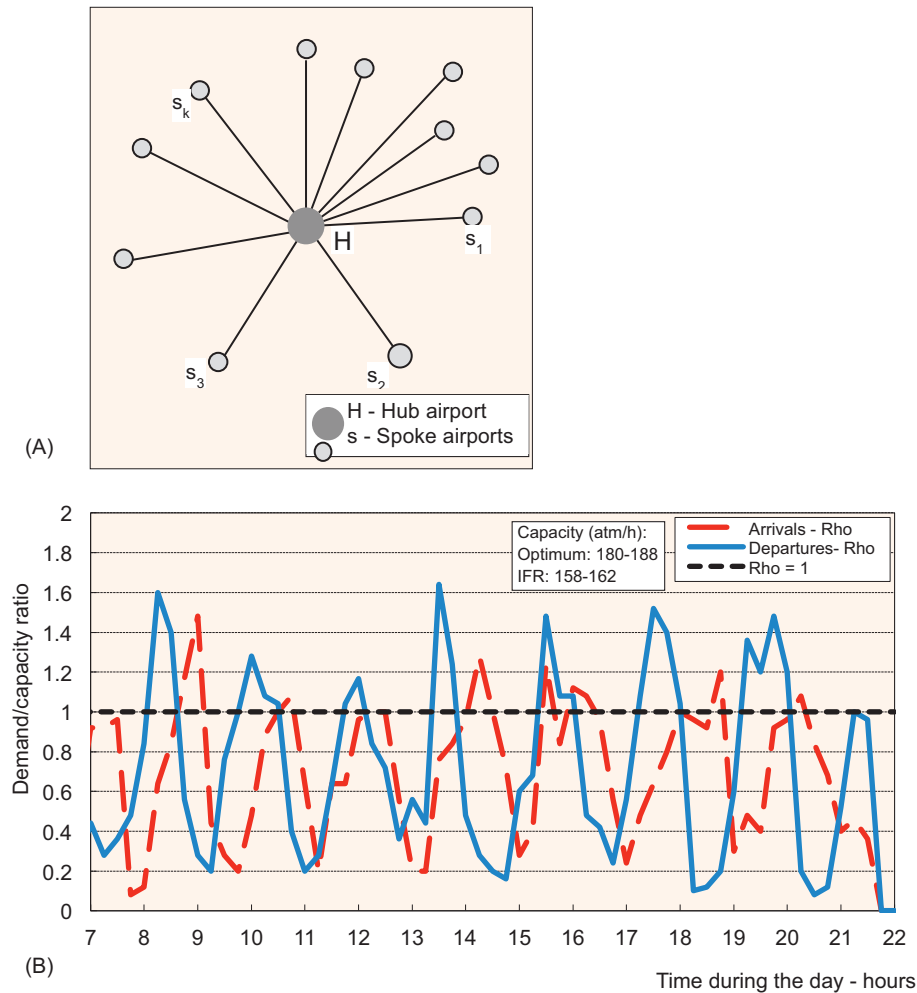


Figure 7 The typical spatial configuration of the airline “hub-and-spoke” network(s) and traffic pattern at the hub airport: Case of Atlanta Hartsfield International airport (Atlanta, U.S.) (FAA, 2007). (A) Spatial configuration; (B) Traffic pattern at base airport.

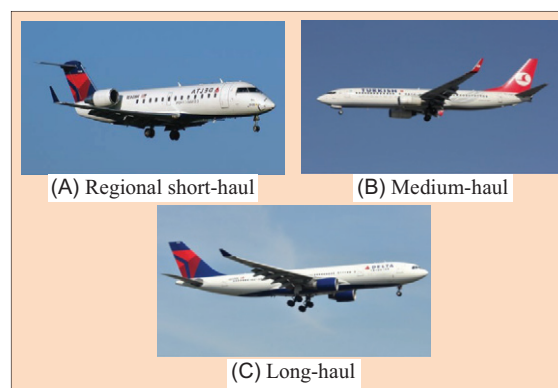


Figure 8 Aircraft categories (DVBBSE, 2018).

ICAO, 2005/2018). A common categorization and related main characteristics of aircraft in the airline fleets of the full cost or the network (legacy) airlines and their alliances are given in Table 1 and Fig. 8 (DVBBSE, 2018; <https://contentzone.eurocontrol.int/aircraftperformance/details.aspx?ICAO=E190/&>).

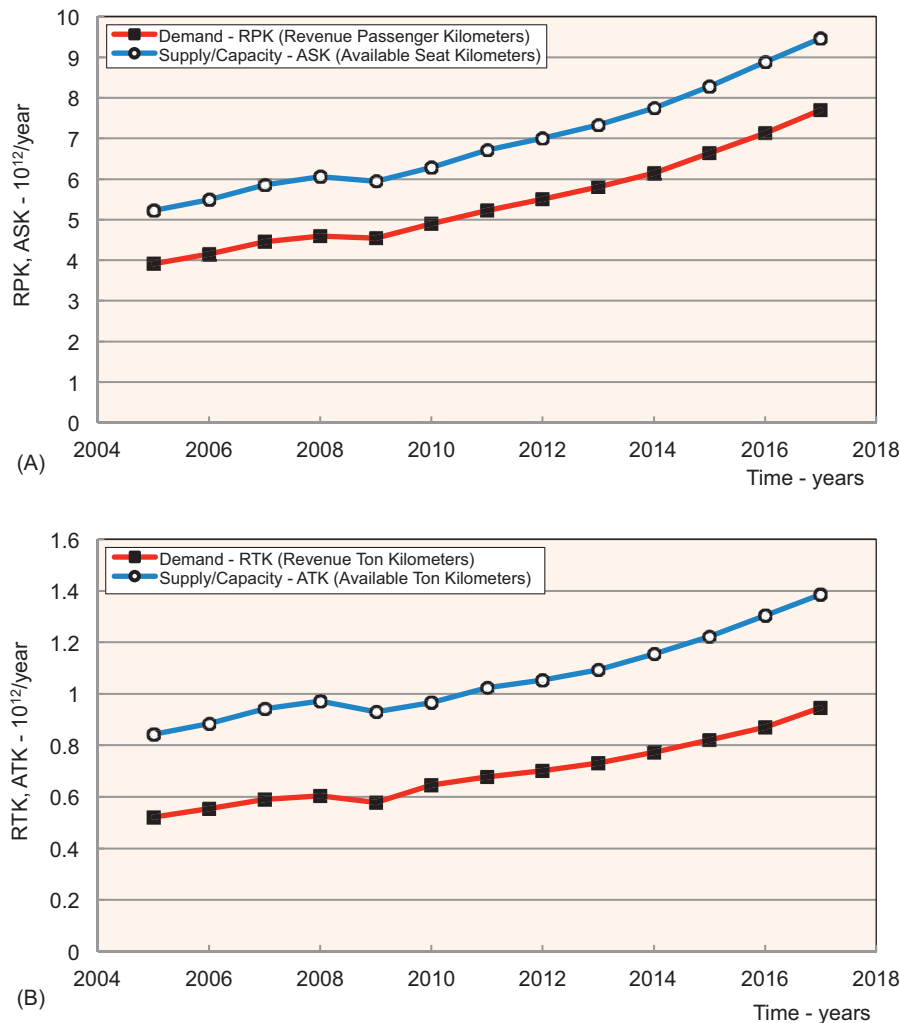


Figure 9 The world scheduled domestic and international air traffic over time (Period: 2005–2017) (ICAO, 2006/2018): (A) Passenger traffic—demand and supply/capacity; (B) Freight/cargo traffic—demand and supply/capacity.

Demand, Capacity, and Quality of Services

During the period 2005–17 the above-mentioned commercial aircraft fleet has supported the growth of the passenger and freight/cargo demand as shown in Fig. 9A, B (ICAO, 2005/2018).

As can be seen, the total passenger and freight/cargo traffic has been growing for about 98% and 81%, respectively, or at an average annual rate of about 5.8% and 5.1%, respectively. The utilization of the available capacity (i.e., load factor) has increased from about 75% to 81% at the passenger and decreased from 68% to 62% at the freight/cargo segment.

The airline quality of services has been expressed, in addition to other measures, by the on-time performance, that is, punctuality, defined again as the proportion of flights arriving and departing within 15 min of their scheduled times. Fig. 10 shows an example for the world's selected airlines. As can be seen, the most punctual have been the mainline airlines, followed by LCCs, and majors (OAG, 2019).

Economics

The airline economics can be considered at different levels such as the individual aircraft and airlines, airline categories—mega, mainline, LCC, regions—continents, and the world airline industry in both aggregate/total and average (per unit of output) term (ASK and PKM). In addition, the particular averages can be expressed over time or depending on the volumes of output during given period. Fig. 11 shows an example of the average revenues and costs of the world's airlines over time (ICAO, 2005/2018).

As can be seen, the average revenues were generally higher than the average costs during the observed period. The exceptions were the years 2008 and 2009 when both were almost equal. After that, they were increasing in parallel until the end of the observed period. This indicates that the airline industry has been overall profitable during the observed period.

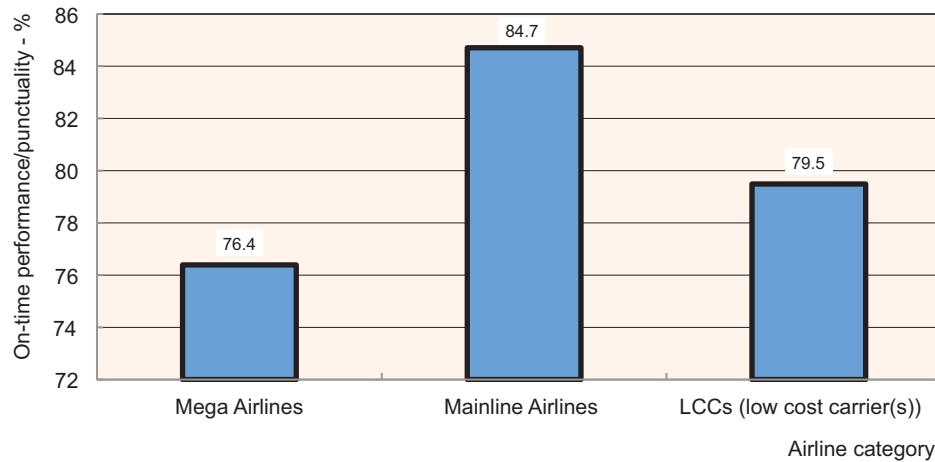


Figure 10 The on-time performance, that is, punctuality, of the selected world's airlines (Period: 2018) (OAG, 2019).

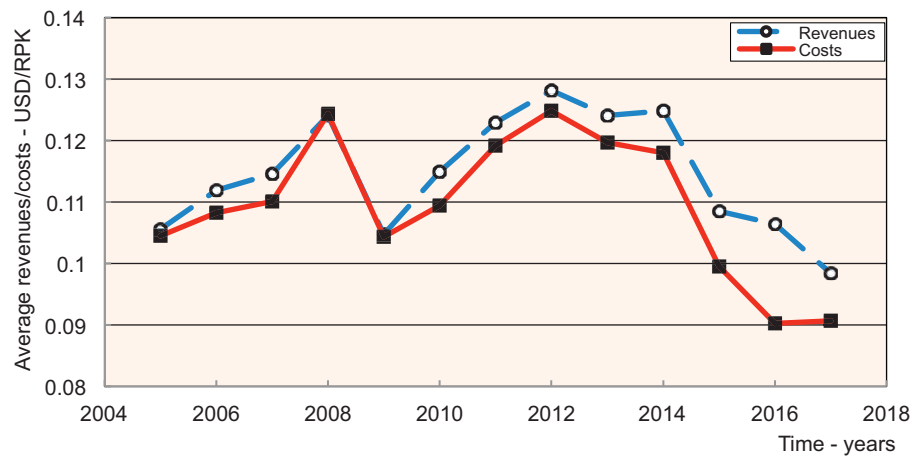


Figure 11 The average revenues and costs of the world's airlines (Period: 2005–2017) (ICAO, 2005/2018).

ATC/ATM Subsystem

Infrastructure, Facilities, and Equipment

The ATC/ATM (Air Traffic Control/Management) subsystem is established to provide safe, effective, and efficient air traffic in the controlled airspace. Its main components are the airspace, technical/technological components, and the ATC/ATM staffs.

1. *The controlled airspace* is divided into smaller parts such as: (i) AZ (Airport Zone) embracing the airport airside area; (ii) TMA (Terminal Airspace) containing the approach and departure trajectories towards a single and/or different runways of a given airport(s); (iii) LAA (Low Altitude Area) enabling the aircraft/flight climbing to and descending from the cruising altitude(s); and (iv) high altitude airspace (HAA) where the cruising phase of flights takes place. **Fig. 12** shows such airspace division (Janić, 2009; <http://www.eurocontrol.int/articles/free-route-airspace/>).
2. *The technical/technological components* are CNS/ATM (Communication Navigation Surveillance/Air Traffic Management) systems including communications, navigation, and surveillance systems using digital technologies, and the satellite systems together with various levels of automation, supporting safe, effective, and effective global ATM system. In addition, these are different tools supporting the tactical and strategic control and management of the individual aircraft/flights and the air traffic flows, respectively (<http://www.sesarju.eu/>; <http://www.faa.gov/nextgen/>).
3. *The ATC/ATM staffs system* include the ATC controllers and ultimately the aircraft pilots. The former are responsible for monitoring, controlling, and managing the aircraft/flights in the airspace of their jurisdiction. At the level of ATC/ATM sectors, they are usually organized as teams of two or three persons with precise division of tasks and responsibilities. The pilots participate in the ATC processes by following, in addition to their own, instructions and guidance from the ATC controllers. These mainly relate to maintaining their aircraft on the assigned trajectories and separation from the other aircraft/flights.

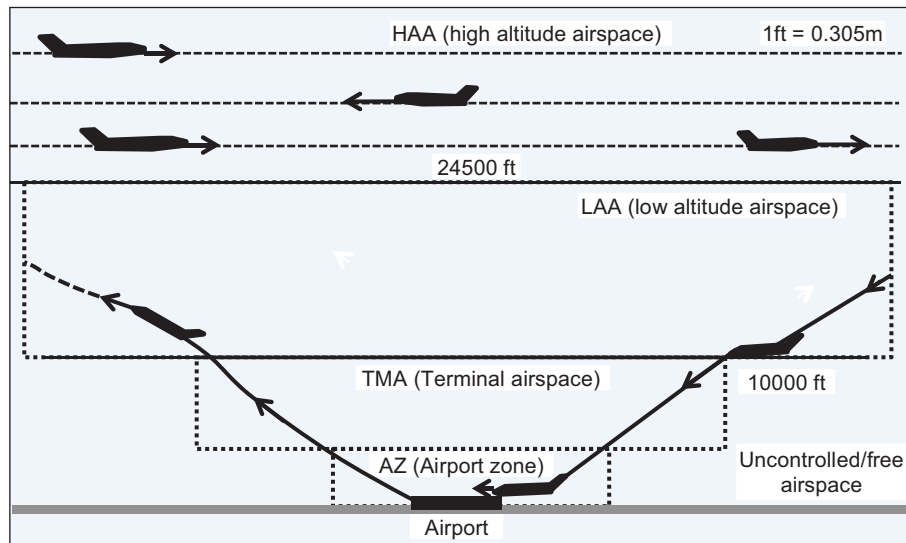


Figure 12 Scheme of division of the ATC/ATM system airspace in the vertical plane.

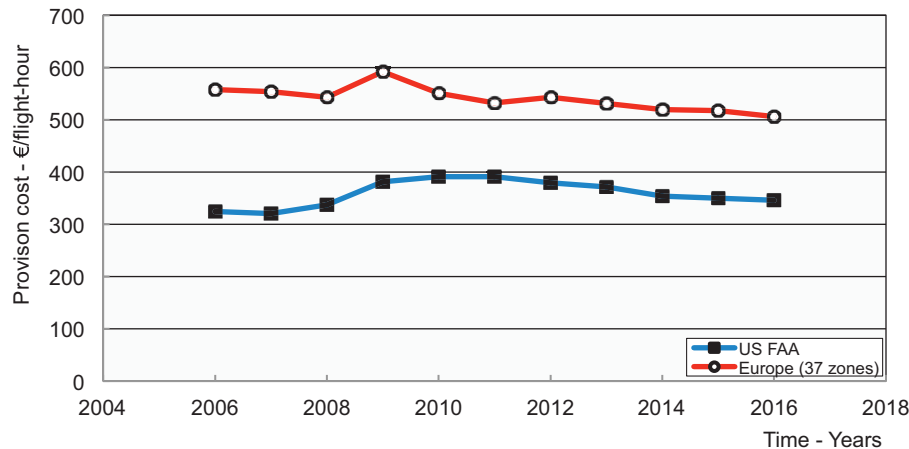


Figure 13 Relationship between the average (unit) cost and the annual number of controlled flight hours in the European and U.S. ATC/ATM system (Period: 2006–2017) (EUROCONTROL, 2019).

Demand, Capacity, and Quality of Services

The ATC/ATM (Air Traffic Control/Management) operates as the sub-system of the ANSP (Air Navigation Service Provider) system, which as being public or private entities provides Air Navigation Services to aircraft/flights during all phases—taking-off, en-route, and landing (ICAO, 2016). These services are as follows: air traffic management (ATM); communication navigation and surveillance (CNS); Meteorological service for air navigation systems (MET); search and rescue (SAR); and AIS/AIM (Aeronautical Information Services/Aeronautical Information Management). At present, 82 ANSP systems in the scope of 166 national Civil Aviation Authority (s) (CAAs) operate in the world. In general, the ATC/ATM subsystems in different regions have possessed different characteristics and performed differently as shown in Table 2 (EUROCONTROL, 2019).

As can be seen, the ANSP is more fragmented in Europe than in U.S. The European area comprises 37 ANSPs with 62 en-route centers and 51 FMPs while the U.S. system controlling for about 10% smaller airspace has comprised only 1 ANSP with 23 en-route centers and 65 TMUs. In addition, the number of controlled airports in Europe has been for about 87% greater than that in U.S. Under such conditions, the U.S. system has handled for about 47% and 47% more IFR and total daily flights, respectively, with about 32% and 43% less ATC controllers and total staff. Despite the higher control time per flight for about 8%, the U.S. system has also performed more effectively regarding the flight arrival punctuality (3%) and the TMA average delays (10%), but worse regarding ATM/TMI flight delays (19%). In addition, Table 3 gives the characteristics of the number of handled flights flight and related delays between airports, that is, in the en-route airspace, of Europe and U.S. (EUROCONTROL, 2019).

As can be seen, during the observed period, an average annual number of 5.1 million IFR flights handled in the European en-route airspace was imposed an average delay of 1.0 min/flight. The proportion of flights delayed more than 15 min was 3.5%/year

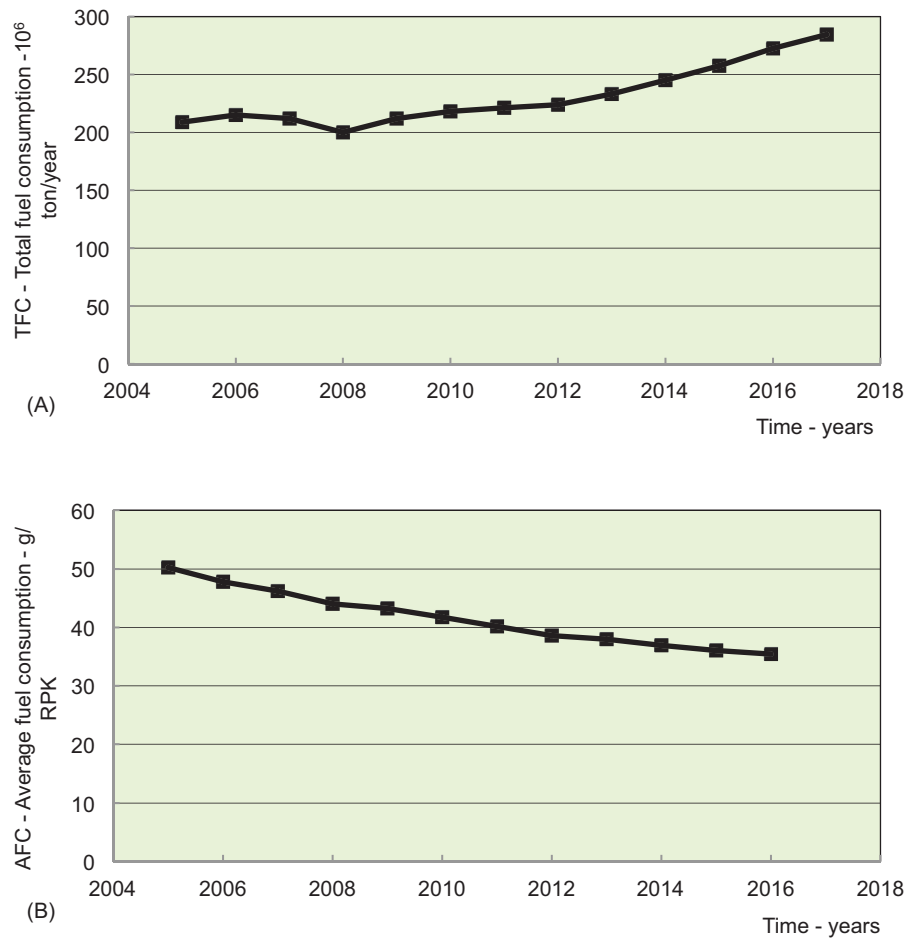


Figure 14 Fuel consumption by the world's commercial air transportation over time (Period: 2005–2018) (ICAO, 2005/2018; <https://www.statista.com/statistics/655057/fuel-consumption-of-airlines-worldwide/>): (A) Total fuel consumption versus time; (B) Average fuel consumption versus time.

with an average delay of 29 min/flight. At the same time, an average annual number of 8.6 million IFR flights handled in the U.S. en-route airspace was imposed an average delay of 0.3 min/flight. The proportion of flights delayed more than 15 min was 0.7%/year with an average delay of 38 min/flight. Similarly as in the above-mentioned case of airports, these figures indicate again that the U.S. airspace have handled for about 70% more flights but on the account of about 3 times shorter and 30% longer average delay per flight and per delayed flight, respectively. At the same time the proportion of these (longer than 15 min) delayed flights was for about five times lower in the U.S. than in the European airspace

Economics

The economics of the ATC/ATM sub-system can be considered through its total and average revenues, costs, and profits/losses. In general, the prices of ATC/ATM services bringing the revenues are usually set up to cover the costs for almost 100%. Therefore, it can be said that the costs reflect the revenues. One of the average economics in the given context is the average (unit) ATC/ATM provision cost, usually expressed as the ratio of total system's provision costs and the volumes of output in terms of the flight-hours controlled (The flight-hours controlled have been estimated as the product between the number of controlled flights and time spend for their controlling). This average (unit) cost includes the ATC controllers' employment cost and the supporting cost (non-ATC controllers' employment and the ATC controllers' training and developments, operating, and depreciation/amortization cost). **Fig. 13** shows some cost characteristics of the European and U.S. ATC/ATM system (EUROCONTROL, 2019).

As can be seen, the average (unit) cost per flight-hour controlled was changing during the observed period in both ATC/ATM systems. It was the highest in the year 2008/2009 when the number of flight-hours was decreased due to decreasing of the number of flights caused by the global economic/financial crisis. In addition, this cost was for about 44% higher at the European than U.S. ATC/ATM system during the observed period.

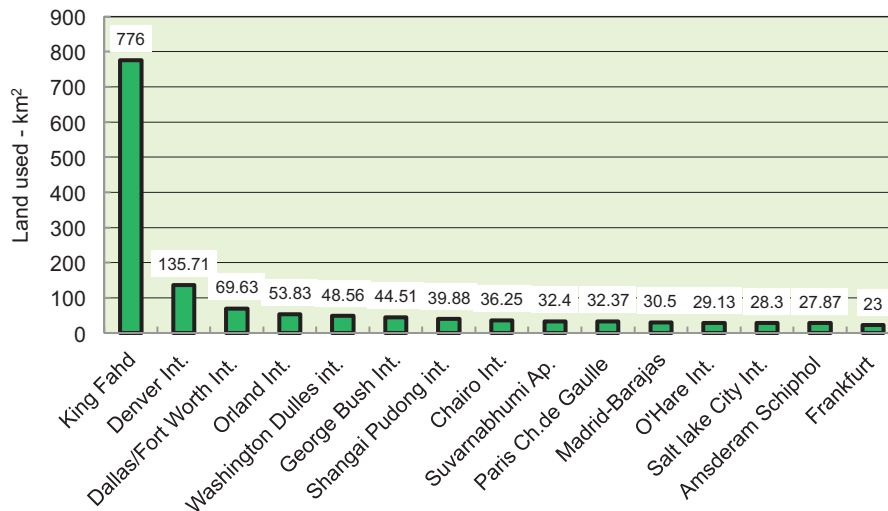


Figure 15 The land used by the world's spacious airports (<https://www.worldatlas.com/articles/the-largest-airports-in-the-united-states.html/>).

Sustainability of Air Transport System

General

The sustainability of air transport system generally relates to its impacts and effects on the society and the environment and their balance in the medium to length period. Its operations create the environmental impacts such as energy/fuel consumption and related emissions of green house gases (GHG) and land use. The social impacts are direct congestion and delays, noise, and safety, that is, traffic incidents/accidents. If internalized, that is, charged, these impacts represent area considered as externalities. The system's effects are contribution to the local, national, and international employment and GDP and consequently the overall social-economic welfare. Both externalities and GDP-contributions can also be considered in the scope of the system's economic performance (Janić, 2007; https://www.icao.int/sustainability/Airport_Economics/State%20of%20Airport%20Economics.pdf/).

Fuel Consumption and Emissions of GHG

The commercial turbojet aircraft mainly consume Jet A fuel or kerosene as a derivative of crude oil. Their fuel consumption over the past decade and half at the global (the world's) scale is shown in Fig. 14 A, B (ICAO, 2005/2018; <https://www.statista.com/statistics/655057/fuel-consumption-of-airlines-worldwide/>).

Fig. 14A shows that after some stagnation at the beginning (2005–07), the total fuel consumption has continuously increased at an average annual rate of 4.7% during the remaining period 2008–17. Fig. 14B shows that the average fuel consumption (g/RPK) has decreased for of about 30%, that is, at an average annual rate of 2.5% during the observed period. Burning Jet A fuel produces GHG of which the most voluminous are CO₂ (Carbon Dioxide) and H₂O (water vapor) emitted at the rate of 3.162 kgCO₂/kg and 1.23 kgH₂O/kg (ICAO, 2014). By multiplying these rates by the above-mentioned fuel quantities, the corresponding total and average emitted quantities of GHG can be estimated. For example, Fig. 13A shows that the total fuel consumption in the year 2016 was 272.552 million tons. This produced 272.552 (3.162 + 1.23) ≈ 1197.048 million tons of CO_{2e} (Carbon Dioxide equivalents). The average external cost of emitted CO_{2e} has estimated to be 212 USD/ton (NAE, 2004; IWG, 2016). In the given case this makes the total costs of emissions of CO_{2e} of 253774.257 million USD or 0.0061 USD/ASK and 0.0075 USD/RPK in the year 2016 (USD - U.S. dollar).

Land Use

Land is used for the airport airside and landside area. Fig. 15 shows the land used by 15 world's most spacious airports (<https://www.worldatlas.com/articles/the-world-s-10-largest-airports-by-size.html/>).

As can be seen, King Fahd (Saudi Arabia) is the most and Frankfurt (Germany) the least spacious airport. In addition, 7 U.S. are among 15 world's most spacious airports. The above-mentioned economics of airports implicitly also reflects economies of their land use. There may be also diseconomies of land use in some cases materialized through changing (diminishing) the value of nearby properties due the impacts such as the aircraft noise, emissions of GHG, and risks of air traffic incidents/accidents potentially damaging properties and threatening lives of the nearby/local population (Janić, 2016).

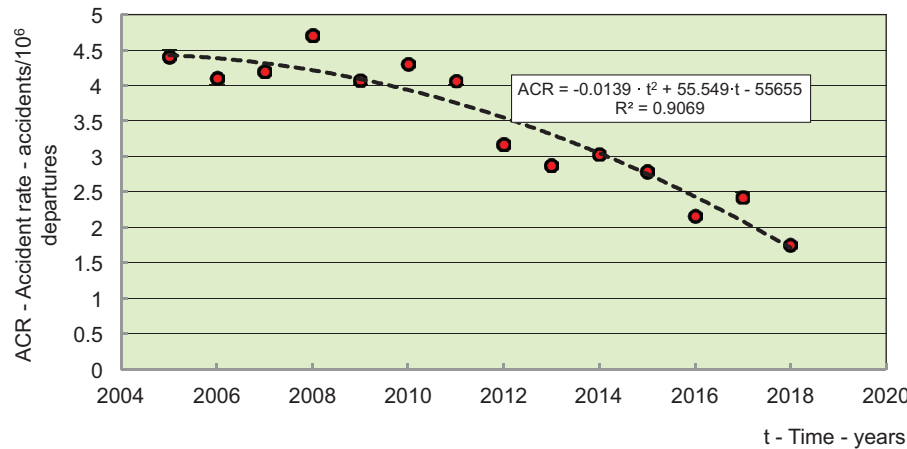


Figure 16 Air traffic accident rate over time (Period: 2005–2018) (<https://www.icao.int/safety/iStars/Pages/Accident-Statistics.aspx/>).

Congestion and Delays

If internalized, the average unit cost of flight delays including the airline operating costs and the cost of passenger time have estimated to be about 81€/min in Europe and 38 min in the U.S. According to Table 3, the cost of en-route delays per delayed flight would be 29 min · 81€/min = 2349 € in Europe and 38 min · 75 USD/min = 2850 USD in the U.S. during the observed period (WU, 2011; <http://airlines.org/dataset/per-minute-cost-of-delays-to-u-s-airlines/>). These costs can also be considered in the scope of the system's economics.

Noise

The relevant aircraft noise is generated near airports by the aircraft taking-offs and landings. In addition to the absolute levels of noise by particular aircraft types, the number of population exposed to the certain levels of noise can be an indicator of the system sustainability respecting to noise impact. In general, regarding the aircraft noise, the US FAA (Federal Aviation Administration) has established the level of day-night level (DNL) of 65 dBA as the threshold or “significant” level. The EEA (European Environmental Agency) has established the threshold level of L_{den} (Level day evening night) of 55 dBA (dBA—decibel). Above these levels, the aircraft noise is considered incompatible with the residential areas. Table 4 gives the population exposed to the aircraft noise around the U.S. airports during the given period (US DOT/FAA, 2015).

As can be seen, despite increasing the number of population, its absolute number and the corresponding proportions exposed to the above-mentioned threshold noise level were decreasing over time, thus indicating the substantive achievements in dealing with the noise burden around the country's airports. This was carried out by the stricter aircraft certification, advanced operational procedures (noise abatement procedures), and improved land use compatibility embracing residential, public, commercial, manufacturing, production, and recreational areas. The excessive noise has been charged at many airports worldwide. These charges have been frequently included into the airport landing fees based on the aircraft noise categories mainly influenced by their size, that is, MTOW (Maximum Take-Off Weight).

Air Traffic Incidents/Accidents (Safety)

The air transport system, like other transport systems, has not been free from the air traffic incidents/accidents. In particular the accidents have resulted in damages of property (aircraft involved), losses of lives and injuries of air passengers and crew on board, destroying freight/cargo shipments, and losses of the third parties involved. Therefore, in the narrower sense (not including the eventual environmental damages) these can be considered as the physical impacts on the society and related costs/externalities rather than on the environment. According the current practice the air accidents have been investigated in order to identify their causes and undertake the appropriate actions aiming at preventing their occurrence due to the already known (investigated) causes. This has resulted in continuous reduction of the number of these accidents despite increasing of the number flights/departures over time, as shown in Fig. 16 (<https://www.icao.int/safety/iStars/Pages/Accident-Statistics.aspx/>).

During the last 10 years (2008–18), the rate of air accidents has decreased for about 63%, that is, from 4.70 to 1.75 events per million annual departures. At least from the statistical point of view this indicates that over time the system has become increasingly safer. Some estimates indicated that the costs of air accidents have been substantive. As an illustration, they were 223 million €/year or 500 €/RPK in the year 2008 in the EU-27 Member States. This amounted 3.33% under low and 0.875% under high scenario of the total external costs.

Contribution to the Social-Economic Welfare

Contribution of the air transport system to the social-economic welfare consists of its employment and created GDP. In the year 2014, the total number of employees was 62.7 million and the created GDP 2.7 trillion USD worldwide. This gives about 43062 GDP USD/employee-year. The highest employment was in Asia-Pacific and the lowest in Middle East region. That in Europe was higher than in North America. At the same time, the highest GDP was created in Europe, followed by that in North America and Asia Pacific region. The lowest GDP was created in Middle East. The highest GDP/employee was in North America followed by Europe and the lowest in Africa region. Some estimates indicate that the total employees by in the year 2034 the air transport system worldwide will increase to about 99 million people and generated GDP to 5.9 trillion USD (ATAG, 2018).

Conclusions

The air transport system worldwide and in particular regions/continents has been growing fast during the past two decades in terms of both air passenger and freight/cargo transportation. The system's external and external driving forces have driven growth of demand served by the supply of adequate system capacity. The former forces have generally included growth of population and GDP, and globalization of international businesses, trade, and tourism. The latter fares have embraced liberalization/deregulation of the air transport markets, consolidation of the airline industry including strengthening of LCCs, and generally decreasing airfares. Consequently, the airlines have increased size and diversity of their aircraft fleets, the airports and ATC/ATM operators have been faced with the shortage of capacity, requesting substantive investments in order to maintain quality and safety of their services at the required level.

The air transport system will continue to grow driven by both already known external and internal prospectively strengthening driving forces. This will require provision of more capacity at all three subsystems—airports, airlines, and ATC/ATM—in combination with further improvements of the efficiency, effectiveness, and safety of operations and related transport services. In parallel, the system will be faced with increasing global and local requirements to be more sustainable, that is, “greener”, regarding all above-mentioned impacts on the environment and society. On one hand, this is expected to be achieved by the technical/technological, procedural, and operational innovations. On the other, these previous would be combined with the policy and economic measures including imposition of the procedural and operational constraints and charging particular impacts as externalities.

Relevant Websites

<https://aci.aero/data-centre/annual-traffic-data/passengers/2017-passenger-summary-annual-traffic-data/>
<http://airlines.org/dataset/per-minute-cost-of-delays-to-u-s-airlines/>
<https://www.bts.gov/content/number-us-airportsa/>
https://en.wikipedia.org/wiki/List_of_international_airports_by_country#Number_of_International_Airports
<https://store.aci.aero/product/annual-world-airport-traffic-report-2018/>
<https://contentzone.eurocontrol.int/aircraftperformance/details.aspx?ICAO=E190&>
<http://www.eurocontrol.int/articles/free-route-airspace/>
https://www.icao.int/sustainability/Airport_Economics/State%20of%20Airport%20Economics.pdf/
<https://www.icao.int/safety/iStars/Pages/Accident-Statistics.aspx>
<https://www.statista.com/statistics/655057/fuel-consumption-of-airlines-worldwide/>
<https://www.worldatlas.com/articles/the-largest-airports-in-the-united-states.html/>
<https://www.worldatlas.com/articles/the-world-s-10-largest-airports-by-size.html/>

References

- ACI, 2006. ACI Worldwide Air Transport Forecasts 2005–2020: Passengers, Freight, Aircraft Movements, Airport Council International, Geneva, Switzerland.
 ACI, 2018. Annual World Airport Traffic Report (WATR), Airports Council International, ACI World, Montréal, Québec, Canada.
 AIRBUS, 2006., Airbus Global Market Forecast, Airbus Industrie AIRBUS S.A.S, Toulouse, France.
 AIRBUS, 2018. Global Market Forecast: Global Networks, Global Citizens 2018–2037, AIRBUS S.A.S., Blagnac Cedex, France.
 ATAG, 2018. Aviation Benefits Beyond Borders, Air Transport Action Group, Geneva, Switzerland.
 Boeing, 2007. Current Market Outlook 2007: How Will You Travel Through Life?, Boeing Commercial Airplanes: Market Analysis, Seattle, WA, USA.
 Boeing, 2017. Current Market Outlook 2017–2036, Boeing Commercial Airplanes Market Analysis, Seattle, WA, USA.
 De Neufville, R., Odoni, A., 2013. Airport System Design, Planning and Management. McGraw Hill, New York, USA.
 DVBSE, 2018. An Overview of Commercial Aircraft 2018–2019, DVB Bank SE, Aviation Research (AR), Schiphol Branch, Schiphol, The Netherlands, London Branch, London, UK.
 EUROCONTROL, 2019. U.S. - Europe Continental Comparison 2006–2016 of ANS Cost-Efficiency Trends, Performance Review Unit on behalf of the European Commission, European Organisation for the Safety of Air Navigation, Brussels, Belgium.
 FAA, 2007. FAA Operations & Performance Data, Department of Transportation, Federal Aviation Administration <http://www.apo.data.faa.gov/>, Washington DC, USA.
 ICAO, 2015. State of Airport Economics, International Civil Aviation Organization, Montreal, Canada.
 ICAO, 2016. Rules of the Air and Air Traffic Services DOC. 4444 - RAC/501/12, International Civil Aviation Organization, Sixteenth Edition), Montreal, Canada.

- ICAO, 2005/2018. Annual Report of the Council 2005-2017, International Civil Aviation Organization, Montreal, Canada.
- IWG, 2016. Technical Support Document: Technical Update of the Social Cost of Carbon for Regulatory Impact Analysis Under Executive Order 12866 (September 2016 Revision), Interagency Working Group on the Social Cost of Greenhouse Gases, Washington, D.C., USA, https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/august_2016_sc_ch4_sc_n2o_addendum_final_8_26_16.pdf/.
- Janić, M., 2007. The Sustainability of Air Transportation: Quantitative Analysis and Assessment. Ashgate Publishing Company, Farnham, UK.
- Janić, M., 2009. The Airport Analysis, Planning, and Design: Demand, Capacity, and Congestion. Nova Science Publishers, Inc., New York, USA.
- Janić, M., 2019. Landside Accessibility of Airports Analysis, Modelling, Planning, and Design, Springer International Publishing, Springer Nature Switzerland AG, <https://doi.org/10.1007/978-3-319-76150-3>.
- NAE, 2004. The Hydrogen Economy: Opportunities, Costs, Barriers, and R&D Needs, National Academy of Engineering, The National Academies Press, <https://doi.org/10.17226/10922>, Washington D.C., USA.
- OAG, 2019. Punctuality League 2019: On-Time Performance for Airlines and Airports Based on Full-Year Data 2018, OAG Connecting the World of Travel, Luton, UK.
- US DOT/FAA, 2015. Statistical Handbook of Aviation (Annual Issues): Departures 1994-95/2014, Bureau of Transportation Statistics/Federal Aviation Administration, Washington D.C., USA.
- WU, 2011. European Airline Delay Cost Reference Values, Final Report (Version 3.2), Department of Transport Studies, University of Westminster, London, UK.

Further Reading

- Holloway, S., 2008. Straight and Level: Practical Airline Economics, 3rd ed. Taylor & Francis Ltd, Imprint: Ashgate Publishing Limited, Aldershot, England, UK.
- IPCC, 2018. Global Warming of 1.5°C: Summary for Policymakers, the Intergovernmental Panel on Climate Change, Geneva, Switzerland.

The History of Air Transportation

Richard P. Hallion, Independent Scholar, Shalimar, FL, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	192
The Dirigible Airship: A False Path	192
The Airplane and Early Air Commerce	193
The Second World War and Reshaping Postwar Civil Aviation	194
The Transition from Piston to Gas Turbine Airliners	195
Leveling the International Airliner Playing Field	195
Summary and Conclusions	196
References	196

Introduction

The invention of the airplane in 1903 by the brothers Wilbur and Orville Wright fulfilled a dream of centuries, giving humanity a practical means of overcoming the vicissitudes of travel across turbulent seas and daunting landscapes. By inventing a means of locomotion that was free of the retarding constraints of geography, moving easily through the air in any direction, the brothers effectively gave to the world three-dimensional mobility. They did so by reviewing and expanding upon the work of European and American researchers such as Sir George Cayley, Otto Lilienthal, and Octave Chanute, adding their own creative insights (particularly in the field of aircraft stability and control), and carefully integrating the four great fields of aeronautical science— aerodynamics, structures, propulsion, and controls—to make the first successful airplane, defined as one capable of powered, sustained, and controlled flight (Gibbs-Smith, 1960; Hallion, 2003; Jakab, 1990).

Less than a century previously, the invention of the steam engine had furnished a motive means of power that revolutionized two-dimensional surface travel. By the early 20th Century, steam-powered vessels were traversing the world's oceans at up to 30 kts (approximately 35 mi/h, or 56 km/h), and steam-powered locomotives were reaching speeds of 60 mi/h (69 km/h) on rail lines that crossed international frontiers, linking the world's major cities. But, being heavy in relation to its power output, the steam engine was not suitable as a practical propulsion system for aircraft. Fortunately, the petroleum revolution, coupled with the advent of the internal combustion engine, enabled the construction of lightweight piston engines capable of furnishing power for flight. The invention of the airplane propeller—essentially a rotating wing generating a horizontal lift vector that pulled the airplane through the air—completed the foundational base for flight propulsion enabling the Wrights to develop a successful aircraft (Hobbs, 1971; Kinney, 2017).

As well as enabling the practical airplane, the petroleum-fuelled engine made possible the practical airship. In contrast to the heavier-than-air airplane, dependent upon aerodynamics—the lifting force of air circulating around its wings (and thus requiring a certain minimum speed, varying from design to design)—to remain aloft, the lighter-than-air airship exploited aerostatics—being buoyed by flammable hydrogen or (in the fortuitous case of America) inflammable helium lifting gases.

This chapter is organized as follows: the first section examines the airship and its failure as a viable air transport vehicle; the second traces the evolution of the airplane and air commerce to the end of the 1930s; the third section discusses the influence and the reshaping of civil aviation after the Second World War; the fourth examines the transition into the jet age and the beginning of mass air travel; the fifth section traces the ending of American air transport hegemony, and the last section offers a summary and conclusion stressing the rise of a “green” consciousness and desire for better environmental stewardship by the global aerospace community.

The Dirigible Airship: A False Path

The prospect of large passenger dirigibles—large rigidly-framed airships containing internal gas cells and multiple engines, most popularly typified by the German Zeppelin airships produced by the firm of their eponymous creator, Ferdinand Graf von Zeppelin—attracted great attention in the first three decades of powered flight. But an international series of dirigible disasters involving American, British, French, and German airships brought the era of the passenger dirigible to a close, most notably the highly publicized loss of the German Zeppelin airship Hindenburg in a landing accident at Lakehurst, New Jersey, in 1937. Although hydrogen conflagrations such as the ill-fated Hindenburg's attracted the greatest attention, the airship's major problem was structural weakness, for many of these large gravity-defying vessels, whether hydrogen-or-helium filled, measured over 750 ft (228.6 m) in length (and sometimes well more), and they were prone to breaking up in storms or from routine air turbulence that would not harm an aircraft. Indeed, considering their fatalities in relation to total flights, dirigible airships hold the dubious distinction of being the most lethal form of mass transportation ever developed. As well, they required immense logistical support, which made them as well the most expensive means of long-range personal travel (Brooks, 1975; Robinson, 1973).

Nevertheless, despite their problems, the credit for the first commercial air transport service goes not to the airplane, but to the airship: in 1909 von Zeppelin, in conjunction with the Hamburg-Amerika shipping line, formed DELAG, the German airship corporation, eventually operating six airships on a sporadic service that, before the First World War, carried 19,100 passengers and flew 65,000 mi (104,650 km) in the course of 881 flights (Davies, 1967).

The Airplane and Early Air Commerce

Experiments in using aircraft for commercial operations took place before the First World War, pioneering carrying cargo, mail, and passengers, subsequently the three major purposes of air transport. On November 7, 2010, Wright pilot Philip Parmelee completed the world's first air freight flight, delivering 100 lb of fine silk to the Morehouse-Martens Company for a cost of \$5,000—equivalent to \$128,000 in 2019—in a flight from Dayton to Columbus, Ohio. On February 18, 1911, Henri Piquet delivered the first air mail parcel on a flight between Allahabad to Naini in the then United Provinces of Agra and Oudh (now the Indian state of Uttar Pradesh). On January 1, 1914, aviator Tony Jannus began the world's first scheduled airplane passenger service on a short flying boat service from St. Petersburg to Tampa, Florida, carrying 1,204 passengers over three months before the service ended, having failed to turn a profit (Davies, 1967; Lane, 1986; Reilly, 1997).

The widespread use of aircraft in the First World War—including using military aircraft to routinely transport dispatches and orders—accelerated the emergence of scheduled air transport immediately after the Armistice (Hallion, 2014, 2018). Even before the end of the war, in 1918, the US Army and the US Post Office had collaborated on establishing the world's first scheduled air postal service, which was privatized after the war. In 1919 three flights crossed the North Atlantic, highlighting the extraordinary technical development of aviation in the decade after 1910: the first by an American flying boat, the second by a converted British bomber, and the third by a British dirigible (Heppenheimer, 1995; Leary, 1985; Linden, 2002; Loftin, 1985).

Though it had invented the airplane, America lagged behind Europe both in the scientific study of flight and in the development of cutting-edge aerodynamic and structural advances. Only by exploitation and elaboration upon German-rooted scientific aerodynamics and structural practices—exemplified by the cantilever (internally braced) thick-wing monoplane and all-metal construction—did the United States catch up, and then surpass, the European nations. By 1930 streamlined “air express” designs by John K. Northrop (working first for Lockheed, then for United Aircraft, and finally on his own) led the world in speed and performance, triggering foreign emulation. In 1933 Boeing and Douglas introduced streamlined twin-engine transports that, attaining flight speeds over 200 mph, revolutionized global air transport design standards. The most advanced of these was the Douglas DC-3, which carried 21 passengers at a maximum speed of 230 mi/h over a range of 2,125 miles, at an operating cost that was just a third that of the earlier 17-passenger Ford 5-AT-C Tri-Motor. The DC-3 became an immediate international success: it was license-produced abroad, and influenced the design of subsequent Russian, Japanese, German, French, and Italian air transport aircraft (Brooks, 1961; Hallion, 2009; Miller and Sawers, 1970).

Over the interwar years, aviation technology advanced rapidly, exemplified by the transformation of the airplane from a wood-and-fabric externally braced biplane to an all-metal monoplane with internally-braced cantilever wings and a streamlined monocoque (shell) fuselage. Accompanying this configuration change were other developments that enhanced aircraft performance including the introduction of retractable landing gears, controllable-pitch propellers, the air-cooled radial engine and streamlined engine cowling, and lift-enhancing wing flaps and leading-edge airflow-controlling wing slots and deployable wing slats. The first application of all-metal and streamlined cantilever design was by German designer-entrepreneurs, notably Hugo Junkers, Adolf Rohrbach, and Claudius Dornier. The appearance of high-performance civil aircraft immediately after the war—beginning with the Junkers F 13 of 1919—reflected the inherently “dual-use” (military-civil) nature of the aeronautics infrastructure underpinning aircraft design (Brooks, 1961; Hallion, 2017; Hirschel et al., 2004; Jarrett, 1997; Loftin, 1985; Miller and Sawers, 1970). Accompanying these developments in airframe design and propulsion were equally significant developments in radio communications, radio navigation and radio beacons, blind-flying instrumentation, radar, instrument landing systems, aviation weather prediction, and airways development and regulation (Volti, 2015; Whitnah, 1966).

Though prewar European air commerce advocates had formed an International Commission for Air Navigation (ICAN) and held periodic meetings, the practical development of an international airline structure only began after the First World War with the formation of the German airline Deutsche Luft Reederei, offering a service between Berlin, Leipzig, and Weimar, using converted military airplanes. Other lines quickly followed across Europe, particularly France, Britain, and Holland, and by the mid-1930s these had coalesced or merged into the great foundational lines of global air transport, Imperial Airways (predecessor to British Overseas Airways Corporation and the modern British Airways), Air France, Lufthansa, Swissair, and KLM (Bluffield, 2009; Davies, 1967; Heppenheimer, 1995; Trimble, 1995).

Having been slower than Europe in supporting aeronautics via government oversight and regulation, America was slower than Europe as well in developing an airline structure, which it did only after passage of key legislation including the Kelly Act of 1925, and the Air Commerce Act of 1926, the latter act marking the birth of what would eventually become the American Federal Aviation Administration (FAA) (Whitnah, 1966). By the mid-1930s, major American airlines such as United, American, and Transcontinental and Western Air (predecessor of TWA), buttressed by smaller regional lines, offered wide-ranging domestic service while Pan American Airways, America's “Chosen Instrument” for international air commerce, offered routine service throughout the Americas and reaching across the Pacific to China (Daley, 1980; Mayo et al., 2009; Van Vleck, 2013). The latter reflected the geography of the United States, separated from its Asian and European nations by vast oceans, which stimulated the development of

what were, in their time, the world's most technically advanced long-range flying boats, designed by the Consolidated, Sikorsky, Martin, and Boeing companies (Smith, 1983). By the end of 1936, Juan Terry Trippe, founder of Pan American Airways, had pioneered air mail and passenger services through the Americas with Consolidated Model 16 and Sikorsky S-38 and S-40 flying boats and across the Pacific to China by island-hopping Martin M-130s and Sikorsky S-42s. In 1939 he extended Pan American's operations to Great Britain and France with the Boeing 314 (Daley, 1980; Davies, 1967; Leary, 1992).

Britain's Imperial Airways and France's Aéropostale and later Air France extensively employed their own flying boats, though typically of lesser range since the Eurasian landmass and the comparatively small area of the Mediterranean permitted shorter hops by mixed-land-and-flying boat services reaching as far as Australia. Holland's KLM made extensive use of new Douglas DC-2 and DC-3 airliners to reach throughout Europe and as far as the then Netherlands East Indies (now Indonesia). Italy developed a comprehensive network through Europe, the Mediterranean and into Africa. The most extensive European airline, however, was Germany's Luft-Hansa (later Lufthansa), which operated landplanes throughout Europe, Russia, and Asia, and an air-mail flying boat service using a mix of packet boats and catapult seaplanes. To Germany goes credit both for the first nonstop East-to-West crossing of the Atlantic, and the first Atlantic-spanning four-engine landplane, the Fw-200 of 1938, anticipating the normative propeller-driven oceanic air transports of the 1940s–1950s. The Soviet Union made extensive use of air transport via a variety of services and partnerships flying a mix of foreign and domestic designs. Japan developed an airline structure through its client states and islands, and its transport design, like that of the Soviet Union, reflected largely German and American landplane and flying boat influence (Bluffield, 2009, 2014; Davies, 1967; Dierikx, 2008; Hirschel et al., 2004).

The Second World War and Reshaping Postwar Civil Aviation

The Second World War (1939–1945) was a war in which air mobility proved of crucial importance, highlighting both the strengths and weaknesses of prewar civil aviation development. In May 1940, for example, Britain's paucity of air transport aircraft necessitated a small-boat seaborne extraction of its soldiers from the beaches at Dunkirk. Similarly, having deemphasized transport development in favor of combat aircraft in the later 1930s, Germany was unable to meaningfully supply its entrapped forces at Stalingrad in 1942–1943. In contrast, the United States became the great supplier of air mobility (and aviation in general) to the Allied cause, with a wide range of airliners-turned-transport, including the ubiquitous DC-3, the newer Douglas DC-4 four-engine transoceanic landplane, the Lockheed 049 Constellation, and other types. By 1945 having mobilized American airline personnel, the US was running an average of 175 transport flights per week across the Atlantic, and, after the end of the war, it launched its first nonstop landplane service across the Atlantic in October 1945 (Bilstein, 2001; Smith, 1969).

In 1943, not wishing to find itself dependent upon American airliners in the postwar era, the British government formed a special wartime committee, the Brabazon Committee, to direct the future fortunes of British civil aviation. While it identified a number of potential new designs, and influentially advocated for both pure turbojet and turbo-propeller future airliners, it could not overcome at war's end both the vast numbers of available surplus American aircraft and newer designs, and crippling decisions by Britain's postwar Labour government (Engel, 2007; Fedden, 1957; Hayward, 1989). By 1951 80% of the world's airliners were American built, and, of these, nearly 60% were from a single firm, Douglas (Brooks, 1961; Miller and Sawers, 1970). Until the advent of Airbus in the 1970s, postwar global air transport was thus dominated by designs from the Douglas, Boeing, and, to a lesser extent, Lockheed companies (Bilstein, 2001; Brooks, 1961; Dienel and Lyth, 1998; Staniland, 2003). The Soviet Union went its own way, its state airline, Aeroflot, making use of war-surplus license-built Lisunov DC-3 derivatives, and subsequent indigenous designs by Antonov, Ilyushin, and Tupolev (Stroud, 1968).

Before war's end, in late 1944, the United States sponsored an international meeting, the Chicago Convention on International Civil Aviation. On 7 December 1944, members formed a Provisional International Civil Aviation Organization (PICAO) to replace the ICAN. The PICAO took effect in June 1945, after the collapse of Nazi Germany, but two months prior to the end of the war in the Pacific. In March 1947 following ratification by 26 nations, it became the ICAO, subsequently being placed under the aegis of the United Nations that October. Formation of the nation-centric ICAO was accompanied by formation of the International Air Transport Association (IATA), an association of 57 of the world's leading airlines, successfully urged by Pan Am's Trippe as a successor to the earlier International Air Traffic Association, formed in 1919; its membership has subsequently grown to 290 airlines. Both worked to steer the fortunes of post-war air transport from the era of the propeller-driven airplane into the era of the jet airliner (Heppenheimer, 1995).

The end of the Second World War marked the end of the great era of commercial flying boats and the onset of the so-called "jet age." By war's end, the global network of airfields rendered the flying boat airliner—compromised by the necessity of having a heavy and aerodynamically inefficient boat hull, together with drag-producing stabilizing sponsons or wingtip floats—obsolete for all but exceptional use. The gas turbine "jet" engine promised routine flight speeds above 500 mi/h (800 km/h) for pure-jet aircraft and cruising speeds well above 300 mi/h (500 km/h) for turbo-propeller ("turboprop") designs. Developed in the late 1930s by German and British pioneers, and transplanted to the United States in 1941, the jet engine was quickly applied to military aircraft, then to civil aircraft. The first civil gas turbine airliner was Britain's Vickers Viscount of 1948, a Brabazon Committee design, which achieved modest international success. Likewise, the first pure-jet airliner was the troubled British de Havilland D.H. 106 Comet of 1949, followed within a fortnight by an experimental Canadian airliner, the Avro of Canada C.102 Jetliner, which introduced the term "jetliner" and was, altogether, an extremely promising design that was unwisely cancelled. America only developed a jet airliner in 1954, with the Boeing 367-80, progenitor of the 707 family; France and Russia followed a year later with the Sud Caravelle and the

Tupolev Tu 104. But these later designs benefited from advances in high-speed aerodynamics, most distinctively the moderately swept-back wing, which enabled these second-generation jet airliners to fly at speeds over 600 mi/h (1000 km/h) (Bright, 1978; Green and Cross, 1957; Stroud, 1994).

The Transition from Piston to Gas Turbine Airliners

Perhaps surprisingly, the introduction of jet airliners did not receive immediate endorsement by airline operators. In part this was due to the catastrophic failure of early Comets due to fatigue-induced explosive decompression at high altitudes, but as well it reflected the inherently high fuel consumption of the pure turbojet that drove up operating costs (overcome by the advent of the so-called “bypass” fanjet engine), and the vast availability of long-range propeller-driven airplanes such as the Douglas DC-4/DC-6/DC-7, the Lockheed 049/749/1049/1649, and the Boeing 377 Stratocruiser operated by airlines including Pan Am, TWA, United, American, Delta, Northwest, BOAC, Lufthansa, Sabena, KLM, Air France, Alitalia, Swissair, and JAL. In the wake of the Comet’s accidents, the Soviet Union was, for over a year, the only country with a civil jetliner in service, the Tu 104, a passenger derivative of the Tupolev Tu 16 bomber (Jarrett, 2000; Stroud, 1968, 1994).

American airlines did not place their first civil jetliners—the Boeing 707 and Douglas DC-8—in service until the end of 1957. But over the year after it did so, tellingly, more than a million passengers flew across the North Atlantic, surpassing for the first time the number of passengers carried across the Atlantic by ship. More fuel-efficient (and hence longer-range) turbofan-powered derivatives of the first 707s and DC-8s continued American market dominance of intercontinental aviation, despite introduction of potential rivals such as the British Vickers VC-10, the Soviet Tupolev Tu 114 (a large sweptwing turboprop airliner derived from the Tu 20 bomber), and the Ilyushin Il 62, the latter closely resembling the VC-10. While the Il 62—which was, essentially, an aerodynamic copy of the Vickers VC-10—was produced in sizable numbers for Soviet and Warsaw Pact airlines, it did not achieve Western market penetration. For its part, the VC-10, though a technically excellent design, likewise proved unable to break the Boeing-Douglas stranglehold on global airline jetliner markets. As jets gradually dominated on domestic and international routes, older legacy propeller-driven designs were employed increasingly for air freight and cargo operations (Brooks, 1961; Dienel and Lyth, 1998; Jarrett, 2000; Miller and Sawers, 1970; Smith, 1969; Stroud, 1994).

From the late 1950s onwards, airlines introduced increasing numbers of short-to-medium range turboprop airliners such as the Fokker F-27, Nihon YS-11, Nord 262, Vickers Viscount, Lockheed L-188 Electra, and, slightly later, a family of Canadian short-take-off-and-landing (STOL) intercity and regional airliners, the de Havilland of Canada DHC-6/DHC-7/DHC-8. Following the example set by France’s Caravelle and the Soviet Union’s Tu-104, over the early 1960s and into the 1970s airlines introduced medium-range jetliners (such as the Hawker Siddeley Trident, British Aircraft Corporation 1-11, Boeing 727, Douglas DC-9, Tupolev Tu 124/134, and Fokker F-28). The end of the decade witnessed the first flights of supersonic transports (SSTs), the Tupolev Tu 144 (December 1968) and the Franco-British Aerospatiale-BAC Concorde (March 1969), and, sandwiched between them, the first widebody “jumbo” jetliner, the Boeing 747 (February 1969) (Jarrett, 2000; Stroud, 1994).

While SSTs proved a false start—the fuel-burn required to overcome high aerodynamic drag rise while passing through the speed of sound made them unattractive to profit-minded airlines, and America abandoned plans to develop its own SST to match the Concorde and Tu 144 largely over environmental concerns—they did enter service with Aeroflot, Air France, and British Airways, largely as symbols of national prestige and technical capability. Capable of cruising at over Mach 2—twice the speed of sound—the Franco-British Concorde remained in service until 2003, though the Tu 144 was quickly withdrawn from service and remained largely in experimental status (Horwitch, 1982).

The emergence of the Boeing 747, however, was a milestone event in commercial aviation, not merely because of its size, but because it signaled the birth of mass commercial air transportation. Overcoming initial sluggish sales and underutilization, the Boeing 747 benefited from airline deregulation to become the greatest long-range people-mover in history, moving hundreds of millions of passengers around the globe over a half-century of continuous flight operations (Sutter and Spenser, 2006). Made possible by the advent of the high bypass ratio turbofan engine, it led to an era of mass air transport that witnessed the development of numerous other wide-body two-three-and four-engine designs by American, European, and Soviet (later Russian Federation) companies and design bureaus, and tremendous expansion of the global airline market, leading to the emergence of a multinational, closely intertwined, and highly regulated air traffic infrastructure. The advent of other technical advances, particularly electronic flight and engine controls; advanced autopilots and stability augmentation; composite structures; computational fluid dynamics; computer-aided design and manufacture; global satellite-based communications, weather, and navigation; and ever-more-powerful, fuel-efficient, and environmentally cleaner fanjet engines; all combined to make air transportation more efficient, safer, and environmentally friendly than other traditional forms of mass transportation such as trains, buses, automobiles, and ships. By the 100th anniversary of the Wrights’ first flights at Kitty Hawk, the world’s airlines had flown nearly 36 billion passengers in roughly eight decades of commercial service (Clark, 2017; Dempsey and Gesell, 1997; Dierikx, 2008; Hayward, 1994; Hirst, 2008; Newhouse, 1982).

Leveling the International Airliner Playing Field

Beginning gradually in the 1970s, the American lock on international civil aeronautics was broken with the advent of foreign-manufactured regional airliners. While aircraft such as the British Viscount and BAC 1-11, and French Caravelle had achieved some

American market penetration, the widespread introduction and acceptance of designs such as the Brazilian Embraer family of turboprop and turbofan regional airliners, the de Havilland of Canada STOL family, the Fokker F-27 and later F-28 and F-100, and the subsequent French ATR-42/ATR-72, British BAe 146, Swedish SAAB 2000, and Canadair (later Bombardier) Regional Jet family signaled that European and other manufacturers could successfully design and market aircraft meeting stringent FAA safety standards and acceptable to American carriers. By 2019 American regional jet fleets were almost entirely equipped with Brazilian and Canadian designs, with a sprinkling of Swedish, French, and British ones as well.

More significant still, the market dominance Boeing and Douglas had over nearly four decades over intercontinental aviation disappeared with the rise of the Airbus in the 1970s. An international consortium established in late 1970 among German, French, British, Spanish, Dutch, and Belgian partners, Airbus flew its first jetliner, the A300B, in October 1972. The A300 gave rise to multiple successors offering great commonality among the types, and which enjoyed widespread success in penetrating the US domestic and international jetliner market beginning with two notable sales orders, the first from Eastern Airlines in 1978 and the second from Pan American in 1984. The advent of Airbus put increasing pressure upon Boeing and McDonnell-Douglas, and contributed to the latter's decline (it eventually merged with Boeing in 1997). Since the Boeing McDonnell-Douglas merger, international air transport fleets have been effectively split between Boeing and Airbus. Arguably, while each rival has stimulated the best from the other, as well, each has at critical points retained or extended its market share based on the other's failures. Examples include Boeing's misstep in attempting to develop a radical canard "sonic cruiser" in 2001, Airbus's misstep in developing the A380 super jumbo, which failed to achieved widespread success, and, most recently, Boeing's troubled 737-Max, which, after two accidents, was grounded worldwide, a grounding still in effect at the time of this writing (Lynn, 1997; Stroud, 1994).

Summary and Conclusions

The growing environmental movement and concerns over good stewardship have led to greatly increased emphasis on reducing even further the fuel burn and emissions of turbofan-and-turboprop-powered aircraft, as well as great interest in finding other means of power generation and motive propulsion via electrical technology. This interest has led as well to revised aerodynamic configurations possibly featuring blended wing-bodies, very broad high-aspect-ratio wings, greater use of lighter composite structures, and advanced flight control systems enabling reductions in the drag-producing size of stabilizing surfaces such as horizontal and vertical tails. In this regard, the earlier derivation of so-called "supercritical" wing and wingtip winglet design has already worked to increase attainable ranges and payload capacity for a given amount of fuel and energy expenditure.

A century after the onset of commercial aviation after the First World War, air transportation has become ubiquitous; by 2010 airplanes carried approximately two billion passengers annually, and 40% of international tourists, with a global economic impact of nearly \$4 trillion USD. If it has lost some of the glamour and exclusivity associated with it during the early jet age, it has become much more commonplace as a means of furnishing humans at every level of society rapid, global-ranging transport, and in bringing together peoples and cultures for the betterment of our shared future. Note to readers: This text was prepared prior to the COVID-19 global pandemic which erupted in early 2020. The effects of the pandemic on air transport have been profound, resulting in numerous aircraft retirements, groundings, and air traffic disruption. The subsequent future of air transport will be critically dependent on how the COVID-19 crisis is addressed by the world's medical and political leadership.

References

- Bilstein, R.E., 2001. *The Enterprise of Flight: The American Aviation and Aerospace Industry*. Smithsonian Institution Press, Washington, DC.
- Bluffield, R., 2009. *Imperial Airways: The Birth of the British Airline Industry, 1914-1940*. Ian Allan Publishing, Hersham, UK.
- Bluffield, R., 2014. *Over Empires and Oceans: Pioneers, Aviators and Adventurers—Forging the International Air Routes 1918-1939*. Tattered Flag Press/Chevron Publishing, Ticehurst, UK.
- Bright, C., 1978. *The Jet Makers: The Aerospace Industry from 1945 to 1972*. Regents Press of Kansas, Lawrence, KS.
- Brooks, P.W., 1961. *The Modern Airliner: Its Origins and Development*. Putnam, London.
- Brooks, P.W., 1975. Why the Airship Failed. *Aeronaut. J.* 79 (778), 439–449.
- Clark, P., 2017. *Buying the Big Jets: Fleet Planning for Airlines*. Routledge, New York.
- Daley, R., 1980. *An American Saga: Juan Trippe and His Pan Am Empire*. Random House, New York.
- Davies, R.E.G., 1967. *A History of the World's Airlines*. Oxford University Press, Oxford.
- Dempsey, P.S., Gesell, L.E., 1997. *Air Transportation: Foundations for the 21st Century*. Coast Aire Publications, Chandler, AZ.
- Dienel, H.-L., Lyth, P. (Eds.), 1998. *Flying the Flag: European Commercial Air Transport Since 1945*. St. Martin's Press, Inc, New York.
- Dierikx, M., 2008. *Clipping the Clouds: How Air Travel Changed the World*. Praeger Publishers, Westport, CN.
- Engel, J.A., 2007. *Cold War at 30,000 Feet: The Anglo-American Fight for Aviation Supremacy*. Harvard University Press, Cambridge, MA.
- Fedden, S.R., 1957. *Britain's Air Survival: An Appraisal and Strategy for Success*. Cassell & Co., Ltd, London.
- Gibbs-Smith, C.H., 1960. *The Aeroplane: An Historical Survey of its Origins and Development*. Science Museum/HMSO, South Kensington.
- Green, W., Cross, R., 1957. *The Jet Aircraft of the World*. Hanover House, Garden City, NY.
- Hallion, R.P., 2003. *Taking Flight: Inventing the Aerial Age from Antiquity to the First World War*. Oxford University Press, Oxford/New York.
- Hallion, R.P., 2009. Streamline Revolution: How Northrop, Boeing, and Douglas Reinvented the Airliner 75 Years Ago. *Aviat. Hist.* 19 (5), 42–47.
- Hallion, R.P., 2014. World War I: An Air War of Consequence. *Endeavour* 38 (2), 77–90.
- Hallion, R.P., 2017. Germany and the Invention of the All-metal Cantilever Airplane, 1915-1925: A Historical Review, SciTech 2017 Forum, AIAA Paper 2017-0112, American Institute of Aeronautics and Astronautics, Reston, VA.

- Hallion, 2018. The First World War's Aviation Legacy.' SciTech 2018 Forum, AIAA Paper 2018-1611, American Institute of Aeronautics and Astronautics, Reston, VA.
- Hayward, K., 1989. The British Aircraft Industry. Manchester University Press, Manchester, UK.
- Hayward, K., 1994. The World Aerospace Industry: Collaboration and Cooperation. Gerald Duckworth Co. Ltd. and the Royal United Services Institution for Defence Studies, London.
- Heppenheimer, T.A., 1995. Turbulent Skies: The History of Commercial Aviation. John Wiley and Sons, Inc, New York.
- Hirschel, E.H., Prem, H., Madelung, G. (Eds.), 2004. Aeronautical Research in Germany: From Lilienthal until Today. Springer Verlag, Berlin-Heidelberg.
- Hirst, M., 2008. The Air Transport System. American Institute of Aeronautics and Astronautics, Reston, VA.
- Hobbs, L.S., 1971. The Wright Brothers' Engines and Their Design, No. 5, The Smithsonian Annals of Flight series, Smithsonian Institution Press, Washington, DC.
- Horwitch, M., 1982. Clipped Wings: The American SST Conflict. The MIT Press, Cambridge, MA.
- Jakab, P.L., 1990. Visions of a Flying Machine: The Wright Brothers and the Process of Invention. Smithsonian Institution Press, Washington, DC.
- Jarrett, P. (Ed.), 1997. Biplane to Monoplane: Aircraft Development 1919-39. Putnam Aeronautical Books, London.
- Jarrett, P. (Ed.), 2000. Modern Air Transport: Worldwide Air Transport from 1945 to the Present. Putnam Aeronautical Books, London.
- Kinney, J.R., 2017. The Power for Flight, NASA SP-2017-631, National Aeronautics and Space Administration, Washington, DC.
- Lane, F.W., 1986. Cargo Flight, No. 2, The Ohio History of Flight Series, Ohio History of Flight Museum/Prop Press Associates, Columbus, OH.
- Leary Jr., W.M., 1985. Aerial Pioneers: The U.S. Air Mail Service, 1918-1927. Smithsonian Institution Press, Washington, DC.
- Leary, Jr., W.M. (Ed.), 1992. The Airline Industry. Bruccoli Clark Layman/Facts on File, New York.
- Linden, F.R. van der, 2002. Airlines and Air Mail: The Post Office and the Birth of the Commercial Aviation Industry. University Press of Kentucky, Lexington, KY.
- Loftin, L.K. Jr., 1985. Quest for Performance: The Evolution of Modern Aircraft, NASA SP-468, National Aeronautics and Space Administration, Washington, DC.
- Lynn, M., 1997. Birds of Prey: Boeing vs. Airbus—A Battle for the Skies. Four Walls Eight Windows, New York.
- Mayo, A.J., Nohria, N., Rennello, M., 2009. Entrepreneurs, Managers, and Leaders: What the Airline Industry Can Teach Us About Leadership. Palgrave Macmillan, New York.
- Miller, R., Sawers, D., 1970. The Technical Development of Modern Aviation. Praeger Publishers, Inc, New York.
- Newhouse, J., 1982. The Sporty Game. Alfred A. Knopf, New York.
- Reilly, T., 1997. BT Jannus: An American Flier. University Press of Florida, Gainesville, FL.
- Robinson, D.H., 1973. Giants in the Sky: A History of the Rigid Airship. University of Washington Press, Seattle, WA.
- Smith, R.K., 1969. Fifty Years of Transatlantic Flight, AIAA Paper 69-1044, American Institute of Aeronautics and Astronautics, Anaheim, CA.
- Smith, R.K., 1983. The Intercontinental Airliner and the Essence of Airplane Performance, 1929-1939. Technology and Culture 24 (3), 428-449.
- Staniland, M., 2003. Government Birds: Air Transport and the State in Western Europe. Rowman and Littlefield Publishers, Inc, Lanham, MD.
- Stroud, J., 1968. Soviet Transport Aircraft Since 1945. Funk & Wagnalls, New York.
- Stroud, J., 1994. Jetliners in Service Since 1952. Putnam Aeronautical Books, London.
- Sutter, J., Spenser, J., 2006. 747: Creating the World's First Jumbo Jet and Other Adventures from a Life in Aviation. Harper Collins Publishers, New York.
- Trimble, W.F. (Ed.), 1995. From Airships to Airbus: The History of Civil and Commercial Aviation, v. 2: Pioneers and Operations. Smithsonian Institution Press, Washington, DC.
- Van Vleck, J., 2013. Empire of the Air: Aviation and the American Ascendancy. Harvard University Press, Cambridge, MA.
- Volti, R., 2015. Technology and Commercial Air Travel. Society for the History of Technology and the American Historical Association, Washington, DC.
- Whitnah, D.R., 1966. Safer Skyways: Federal Control of Aviation 1926-1966. The Iowa State University Press, Ames, IA.

The Geography of Air Transport

Lucy Budd, Stephen Ison, Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	198
The Historical Geographies of Air Transport	198
The Contemporary Geographies of Air Transport	200
See Also	201
References	201
Further Reading	202

Introduction

Air transport is a central component of the functioning of contemporary society and the modern world order. It is also inherently geographical as it concerns the safe, rapid, and routine, but also the spatially uneven, socially inequitable and environmentally degrading, movement of people and goods between prescribed geographic locations. The development and widespread utilization of air transport for commercial civilian purposes during the 20th century has changed not only the patterns and processes of long distance (and often international and intercontinental) transport and human migration but also created new ways of becoming and being aeromobile. In the process, commercial air transport has been responsible for changing the face of the earth as terrain is flattened, mountains sculpted, coastal areas reclaimed, and rivers diverted to satisfy the exacting technical requirements and safety standards associated with new runways and ever larger airports. Air transport's influence has also extended into global culture and the lexicon of everyday conversation with talk of "autopilots", "flying high" and "taking off" becoming metaphors for the challenges and successes of modern life.

As the second decade of the 21st century commences, air transport has grown to the point where in excess of 1300 airlines, operating almost 32,000 aircraft, fly over 4.7 billion passengers each year between 3800 airports worldwide. At the busiest global hubs, an aircraft lands or departs every 45 s while the A380, the largest commercial aircraft currently in service, is certified to carry up to 853 passengers. Driving society's apparently insatiable demand for aeromobility is the need to move goods from their point of manufacture to their point of consumption and people from where they are to where they want to be in the safest, quickest, and most cost-effective manner possible. Air travel is a derived demand designed to overcome (as far as possible) the limitations of gravity and terrestrial geography, and thus it is subject to fluctuations in the health of the global economy, as well as emerging transborder security threats, variations in oil prices and the vagaries of international political relations. In the following sections, we document what we have termed the "historical" geographies of air transport, which roughly correspond to the period between the mid-18th and late 20th centuries, and the "contemporary" geographies of air transport which commenced at the turn of the millennium.

The Historical Geographies of Air Transport

Although it has proved impossible to accurately determine where the chronological development of air transport begins, historical evidence indicates that humans have been trying to "conquer" the air since before the middle of the third century B.C. when Aristotle began to seriously consider the practicalities of flight. Yet the idea of, and desire for, flight has influenced human civilization since ancient times, with many cultures and religions containing accounts of mythical creatures or deities carrying people up into the heavens. Although gravity and human physiology proved an insurmountable barrier for centuries, by the mid-18th centuries, advances in scientific understanding and experimental method led first to the development of aerostatic flight involving tethered balloons filled with hot air and, later, aeromobile flight in free flying hot balloons. Developments with fixed wing gliders followed throughout the 19th century. But while gliders permitted some element of directional control, they obliged aviators to land at a point lower in elevation from that which they departed, and the lack of propulsion ensured that the effective range of the machines was limited.

By the turn of the 20th century, aeronautical efforts were increasingly being focused on creating an engine that was sufficiently powerful and reliable yet simultaneously light enough to be installed on an aircraft. According to most conventional histories of flight, the world's first successful heavier-than-air powered human flight occurred on the morning of 17th December 1903 on the sand dunes of Kitty Hawk, North Carolina when Orville Wright accelerated his aircraft along the ground. Although the flight was short (it could have been accommodated within the economy class section of a B747), it marked the first time that a human had taken off, sustained controlled level flight and landed at a point that was at least equal in elevation to the point of departure. Crucially, the physical location of the flight was deliberately selected on account of its geography-the dunes were located away from human settlement, the lack of vegetation ensured the local flora posed no hazard, the sand provided a supportive surface, and the constant wind would blow over the wings of the aircraft to generate lift. Despite photographic evidence of the flight, many national

politicians and scientists remained sceptical and as late as 1908 the British Government maintained that flight by a heavier than air machine was impossible.

Between the Wright Brothers' first successful flights of December 1903 and the outbreak of the First World War in 1914, aviation developed without any real expertise or experience and often at great personal cost to individual aviators, many of whom were maimed or killed in their pursuit of new speed, endurance or altitude records. Although aircraft only saw limited use during the 1914-18 conflict, by the time the war ended many demobilized pilots sought to operate their superfluous ex-military aircraft for civilian purposes and commercial gain.

Most national Governments now recognized aviation's strategic military and commercial potential and viewed post war flying as a socially and economically progressive force that would foster better international relations by promoting cross-border trade and cultural understanding. This utopian ideal was encouraged through the organization of international air meets, air races and the widely reported successes of pioneering long distance flights. In June 1919, pilots Alcock and Brown became the first aviators to successfully fly across the Atlantic Ocean while in August that year the world's first international scheduled flight took off between London's Hounslow aerodrome for Le Bourget airfield in Paris.

The early commercial flights which followed this inaugural service, while being expensive to operate and vulnerable to the vagaries of the weather, nevertheless demonstrated the utility of aircraft for civilian purposes. In August 1919, the delegates of 26 Allied and Associated Powers signed the Paris Convention of which, among other many notable interventions, formally enshrined the right of individual countries to claim sovereignty over the airspace above their territory. In this way, the skies above nations became highly politicized spaces and international agreements were quickly needed to establish the right of an aircraft registered in one country to access the airspace of another. This fundamental shift in the spatial geographies of human activity highlighted the increasing obsolescence of old geopolitical paradigms which were literally "grounded" in the conventions of Euclidean territorial geometry and forced increased recognition that the 20th century would be increasingly shaped by new geographies of the air.

By the mid-1920s, flying had developed into highly mediated and commercialized activity which attracted considerable prize money, corporate sponsorship, media coverage, and public interest. In addition, aviators including Britain's Alan Cobham performed long distance proving flights to the furthest corners of the Empire. In a bid to capitalize on aviation's popularity and stress its public utility to the development, maintenance and policing of Empire, European colonial powers sought to expand their territorial influence through the medium of the air and many pioneering long-distance flights were staged between European capital cities and the furthest-most points of their nation's respective overseas interests.

In addition to demonstrating the nation's scientific and technological prowess, the flights also sought to identify landing grounds and information on the effect of different climatic zones and weather phenomena on aircraft handling and performance. The earliest geographic studies of aviation that emerged in the 1920s thus addressed practical issues concerning aerial navigation, the load bearing capacity of the geology underlying the landing strips, the impact of different weather phenomenon on flight safety, the prevalence of infectious disease at foreign landing sites and the commercial and political viability of new international air routes.

The development of regular scheduled long-haul international air services in the 1930s also necessitated the development of new legal frameworks and international agreements governing the rights of revenue-earning airlines from one nation state to access the airspace of another. During this time, articles documenting the successes of pioneering long-distance flights and the development of aviation in different world regions began to appear. Physical geography, especially as it pertained to topography and climate, directly influenced both aircraft technology and the operation of flights. In equatorial Africa, for example, where suitable land to construct aerodromes was limited, flying boats were developed to permit operations from lakes and rivers while in the Arabian desert, vast ditches were dug into the sand to aid visual navigation and intermediate (and often fortified) landing grounds were constructed to enable aircraft to be refueled and to offer overnight protection to passengers and crew while they awaited dawn and take off the next morning. Such early morning departures were a requirement of flying in desert regions for although aircraft could not yet fly in hours of darkness the effect of the sun's rays warming the sand created dangerous convective activity and turbulence that pilots and passengers alike were keen to avoid.

The outbreak of the Second World War in 1939 and the need to establish military superiority, stimulated significant advances in aircraft design, production, and technology. In Britain, all civilian flying was banned, and attention turned toward military applications. New aircraft systems, together with new forms of propulsion and more sophisticated navigation, surveillance and communications equipment were rapidly developed and deployed with different nations specializing in the production of types of aircraft. This technological specialization meant that the United States, which had long held the lead in terms of aircraft design, focused on the production of large transport aircraft while Great Britain perfected the technique of mass-producing military machines including Spitfires and Hurricanes. Engineers in Britain and Germany were also independently developing new forms of aircraft propulsion which culminated in the jet engine. These engines enabled jet aircraft to fly faster and higher than the piston and propeller aircraft they replaced and after the war ended, they were quickly installed on commercial airframes.

The late 1940s were an important period in the development of air transport geography and influential works that examined the complex interactions that existed between air travel and geography were published. Some of these studies sought to explore how aircraft were expanding the geographic field of vision by enabling people to look down on the earth from the air and, in so doing, gain a greater appreciation of geological processes and landforms and their impact on patterns of human settlement and land use. In 1947, Balchin discussed the "geometrically static" but "geographically dynamic" nature of distance in an era of global air travel in *"Air Transport and Geography"* while Huntingdon (1947) speculated about the future implications of air transport on effecting the

compression of time and space, on global population distribution, cultural diffusion and the spread of agricultural pests and human infectious disease.

The world's first jet powered commercial airliner - the British designed de Havilland Comet-first flew in 1949 and entered passenger service in 1952 and immediately reduced travel times between London and Johannesburg from several days to a couple of hours. The quantitative revolution in the late 1950s and early 1960s, combined with the post-war expansion of air routes, stimulated the publication of a corpus of work that examined and mapped the network attributes (in terms of time, cost and distance) of airline operations. Edward Taaffe (1959) pioneered work into the description and visualization of the new spatiality and urban connectivity passenger airlines were creating and demonstrated the primacy of certain hub airports. The 1950s saw geographers begin to examine the airport nodes themselves, rather than just the linkages between them. Kenneth Sealy (1957) was among those to explore the geographies of airport location and provided detailed descriptions of the physical, economic and regulatory environments in which air transport operated.

In the 1960s, there was a gradual shift away from network-based studies and a diversification of geographic research into air transport which began to explore the effects of time space convergence, global airport architecture and the apparent feelings of "placelessness" the world's major international airports could engender. The advent of Concorde and progressively larger and heavier commercial aircraft during the 1970s, including the B747, raised the issue of aircraft noise around airports while the vacillations and political controversy surrounding the choice of site for London's third civilian airport ensured that issues of airport location, economics, and regional development were placed firmly on geographic research agendas. Geographers were also becoming aware of the potential network diseconomies that could result from there being insufficient airport or airspace capacity to accommodate growing volumes of passengers and flights. However, in a significant departure from conventional quantitative studies, a few geographers began exploring the corporeal experiences of air travel with Borgstrom (1974) in particular writing about the behavioral, corporeal and affective experiences of flight.

By the end of the 1970s, the economic and political arguments favoring deregulation came to dominate academic agendas. In 1978, President Carter's Administration passed the US Airline Deregulation Act which increased market access, revoked tariff setting legislation and removed capacity constraints at US airports. This new regulatory environment had profound spatial implications which included the consolidation of regional flights at key hub airports, the reorganization of airline networks into hub and spoke systems, and the emergence of a new type of airline operator which competed on price, not service, and flew predominately from smaller less congested secondary and regional airports. Geographers were well placed to discuss the changes this regulatory reform enacted, providing details of the geographical implications of new hub and spoke networks for passengers, airports, airlines, and the regions they served. Concurrently, research into individual airports became common with case studies appearing documenting the development and impact of major airports as well as charting the social geographies of local community protest further expansion.

In the 1980s and 1990s, increased attention was paid to the geographies of air transport-a phenomenon which can be attributed, in part, to the founding of dedicated transport and air transport research journals, the increased use of computers (and, later, the widespread use of the Internet for data collection and processing), and the fact that air travel was becoming a common commodity that was increasingly within the financial reach of a growing number of consumers worldwide. Some of these studies focused on the service and network strategies of individual airlines, others examined the role of air service development in the politics of nation building while others explored the socio-economic and cultural contexts in which individual airports operated. Although much of this work, and the scholars who were undertaking it, was published in English and focused on airline networks in in/from/to the globalized "west" a limited number of studies also began to appear which examined the development and patterns of airline services in other world regions which had hitherto been neglected.

Progressive moves toward deregulation, air service liberalization and airport and airline privatization in Europe and other world markets toward the end of the century stimulated the publication of a large body of research in the changing patterns of air service provision in different world regions. Heightened awareness of the epidemiological and environmental implications of mass aeromobility was also becoming increasingly evident but it was growing awareness of the social dimensions of air travel, forged in part thanks to the "mobilities turn" in the social sciences at the turn of the millennium, that helped to usher in a new era of contemporary air transport geographies.

The Contemporary Geographies of Air Transport

In his discourse on the geographies of mobility, Cresswell (2006) considered the metaphysics of movement in Western social thought and argued that an historical sedentarism suspicion about mobility has been largely superseded by a new global nomadic overview in which hypermobility is the norm. Such flows, whether material or virtual, real or imagined, were seen to circulate around the globe, often spilling over the dams and defences of firewalls, immigration controls and customs inspection posts that were designed to impede their progress. The use of such hydraulic metaphors demanded theoretical perspectives that can comprehend these new "geographies of flow" which are attuned to the high levels of mobility-or hypermobility-which had come to characterize the modern world.

In light of this increased attention to questions of movement, sociologist John Urry (2000) argued that there was a "mobility turn" spreading into and transforming the social sciences which was characterized by a newfound focus on the ways in which the social world is shaped by the global networks which transmit flows of people, goods, money, and knowledge. For geographers, this

enabled new explorations of the airport terminal as a site of spectacle and spectatorship, the aircraft cabin as a place of fear and fascination and the airspace in which aircraft fly as a highly politicized and social realm (Adey et al., 2007).

Although it has been claimed by some that the paradigmatic modern experience is that of routine rapid mobility over long distances, it is important to note that this is only applicable to a relatively small number of the global population and thus it cannot be considered a truly universal experience as not all members of the global community are equal participants. As Massey (1996) has explained, the time-space compression that air transport has affected is highly uneven and exhibits a distinctive "power geometry" in which individuals and particular social groups are either emancipated or marginalized. The contemporary geographies of air transport are thus highly socially variegated, and both reflect and reinforce discourses of political power, monetary wealth and social privilege. High levels of personal mobility by frequent flyers must be juxtaposed against the spatial inertia and inaccessibility of the less affluent. Here, issues of race, ethnicity, citizenship, gender, age, and mental and physical dis/abilities conspire to enable or restrict particular forms of mobility.

Viewing air transport through a lens of social and cultural theory as opposed to economic and technological rationality has alerted geographers to some of the key issues of power and identity that adhere to different spaces of mobility. Indeed, it has been argued that commercial air transport provides one of the most highly visible articulations of global power and the influence of the state. The airport, as a gateway to a nation, has long been a target for terrorist activity and, consequently, the terminal has developed into a place of constant surveillance. Signs and recorded security announcements continually reiterate that passengers are being watched, classified, sorted and screened. These trends are manifest in programs and procedures that employ biometric technologies, artificial intelligence and the cross-border exchange of passenger information which work to facilitate the rapid mobility of "known" and "approved" travelers but purposefully delay or deny mobility to others.

The plight of thousands of airport employees who perform the necessary but repetitive, mundane, poorly paid and often hazardous cleaning, construction, maintenance, and security activities in the terminal has also been brought into sharp relief and their experiences of shift work have been contrasted with the hypermobility of the passengers they serve. The different practices and technologies that create these inequalities of work and access bestow particular emotional attachments for individual airport users. The supposedly "placeless" realm of the international airport may, therefore, be variously experienced as exciting, stimulating, stressful, tedious, disorganized, overwhelming, oppressive, frightening, welcoming, thrilling, frustrating or soporific and trigger a remarkably diversified range of inhabitation.

The debate surrounding the role of "place" in contemporary air transport geographies poses important questions about how "local", regional or national identities are projected onto the global stage. The name and corporate identity of individual airlines, particularly full-service flag carrying airlines, are often intimately tied to expressions of national identity and belonging while major airports are often named after famous local citizens, with politicians, composers and sporting personalities being particular popular choices.

One of the most widely noted characteristics of contemporary air transport is its seemingly inexorable expansion. Despite occasional external shocks temporarily depressing passenger demand, the number of flights and passengers is increasing year-on-year, generating new demand for airports, runways, and air routes. Despite significant concern about the implications of such expansion on the global climate, many aviation futures are being contested at the local level by those concerned by the noise and emissions generated from enlarged and expanded airport operations. Anti-airport expansion campaigns and flight free holiday movements are growing in influence as citizens articulate their opposition to unfettered airport expansion.

Geographers have consistently identified the emergence and expansion of air transport as being one of the key drivers of globalization and a key enabler of social inequity. They have used aerial photographs to map the surface of the earth, data on air traffic flows to delineate the morphologies of air transport networks and identified the social and cultural dimensions of a technology that seamlessly binds important global cities together in an ever-tighter web but leave many other more peripheral spaces only loosely connected.

See Also

Airline Economics; Privatization and Deregulation of the Airline Industry; Demand For Air Travel and Income Elasticity

References

- Adey, P., Budd, L., Hubbard, P., 2007. Flying lessons: exploring the social and cultural geographies of global air travel. *Progr. Human Geogr.* 31 (6), 773–791.
- Balchin, W.G.V., 1947. *Air Transport and Geography*. Royal Geographical Society, London.
- Borgstrom, R.E., 1974. Air travel: towards a behavioural geography of discretionary travel. In: Eliot Hurst, M.E. (Ed.), *Transportation Geography Comments and Readings*, McGraw-Hill, New York, pp. 314–326.
- Cresswell, T., 2006. *On the Move: Mobility in the Modern Western World*. Routledge, London.
- Huntingdon, E., 1947. Geography and Aviation (originally appeared in *Air Affairs* 2(1)) reprinted (1951). In: Taylor, G. (Ed.), *Geography in the Twentieth Century A Study of Growth, Fields, Techniques, Aims and Trends*. Methuen, London, pp. 528–542.
- Massey, D., 1996. A global sense of place. In: Daniels, S., Lee, R. (Eds.), *Exploring Human Geography A Reader*. Arnold, London, pp. 237–245.

Sealy, K., 1957. *The Geography of Air Transport*. Hutchinson University Library, London.

Taafe, E.J., 1959. Trends in airline passenger traffic: A geographic case study. *Ann. Assoc. Am. Geograph.* 49, 93–408.

Urry, J., 2000. *Sociology Beyond Societies: Mobilities for the Twenty-First Century*. Routledge, London.

Further Reading

Goetz, A., Budd, L (Eds.), 2014. *The Geographies of Air Transport*. Routledge, Abington.

Budd, L., Ison, S (Eds.), 2020. *Air Transport Management: An International Perspective*. Routledge, Abington.

The Future of Air Transport

Rico Merkert, James Bushell, Institute of Transport and Logistics Studies, The University of Sydney Business School, Sydney, NSW, Australia

© 2021 Elsevier Ltd. All rights reserved.

More People, More Places—More Integration	203
Improving Footprint, Enhancing Sustainability	204
Big Data, Better Decisions?	205
Policy and the Rules of the Game	205
Air Freight, Slower but Cheaper?	206
The Drone Revolution	206
Beyond the Barriers	206
Final Observations	207
References	207

As we sit here with our crystal ball to write this chapter, we think back to not much more than 100 years ago when the technology for powered flight technology was just taking off. Prior to then, human flight had been a fantasy of Icarus or a risky and unpredictable experience in a balloon. But in that touch more than 100 years, we have flown many times faster than sound, flown into space and have taken control of light like no other species, incorporating it into our way of life. As evidenced over this time, aviation is a rapidly changing and always developing sector, bringing with it substantial impact on many facets of our society. The industry has grown to such an extent that today if measured as a country in terms of generated GDP it would be with 65 million supporting jobs and \$2.7 trillion in global GDP the 21st largest economy in the world (IATA, 2019). While its significant economic, social, and cultural impact has been shown across both city and for rural and regional communities (Baker et al., 2015), the recent debate also includes negative externalities, most notably its carbon footprint (aviation contributes 2%–3% of all carbon emissions globally; European Commission, 2019), as the awareness and urgency of climate change action grows.

So what will the future bring? Notwithstanding the mistiness that grows on the horizon as we look further and further forward, in the future we can see, technology, policy, and new operating models will drive change to air transport, continuing on its trajectory of disruption and growth (demand for air transport has doubled over the last 15 years, see Airbus, 2019). There are various segments of the air transport sector, with some more mature than others. As with Airbus (2019), we argue that air travel will continue to become more and more accessible connecting more and more places and people through new forms of interconnectedness. Environmental policies and agreements such as CORSIA will continue to develop to help manage the carbon costs created by air travel, and big data along with digitization and automation will revolutionize how, what, and when passengers are marketed and serviced to and how their services are delivered. In that context, pure airline operations will remain fundamental yet increasingly smaller functions within large and potentially multinational enterprise that are focused on providing customer centric mobility and logistics services. And finally, new air transport developments will happen outside of the airline space. Drone technologies, which are currently in an embryonic stage of development, will transform how we go about our daily lives in cities, both from a freight and passenger perspective, as we make use of the third dimension in fascinating new and more integrated ways.

The sections of this chapter are structure in a way that aims to provide a comprehensive overview of what we think the future of air transport may look like. Starting with a section on strategies around improved connectivity and deeper integration of the travel value chain, we will then discuss issues around sustainability in the light of an ever-growing aviation industry followed by a section on big data. The chapter moves then on to an evaluation of future governance and regulation scenarios which is followed by more detailed thoughts around air freight and the evolution of drone eco-systems. We conclude the chapter with some innovative thinking around technologies and strategies that will get the air transport industry to the next level.

More People, More Places—More Integration

As we move ourselves to new places around the globe, we reposition ourselves, moving our knowledge to be used for new purposes, exposing ourselves to new cultures and connecting us with new people and places (in the future most likely beyond our planet and into space). Air transport originally formed point-to-point connections between two places, where people were and where they wanted to be. But over time airlines have worked together to make new connections for their customers. Through interline, codeshare and joint venture agreements they have built virtual, shared networks making connections with their peers and strategic alliance partners to deliver more options and seamlessness for consumers both in passenger and freight operations. Initial steps to connect even more destinations into these networks through rail services has had middling success however it has opened discussions about how coordination between operators should be facilitated past the airport departure and arrivals lounge (Merkert and Bushell, 2019a). But pressures on congested airspace will continue to grow, consumers will want more integrated choice (Chowdhury et al., 2018), and

environmental concerns will be privatized into airline cost structures through carbon trading mechanisms (with CORSIA being a first step into that direction). Accordingly, we expect to see refocused effort on these coordination problems to more successfully integrate rail and airline services in the immediate future. This may also require the reformation of rail, bus (and in some ways taxi) operators to reposition them commercially and competitively, to encourage partnerships to develop between different transport operators to develop solutions that are better for everyone, and not just the operator involved. This may be connected with land transport pricing reform (i.e., road user charging—for both road operational and maintenance costs as well as congestion) where all users of the land transport network need to think differently about how they use it most efficiently and will also be impacted by the rise of automated ground vehicles (most likely electric propulsion) in particular for the last mile trip or delivery. We further argue that such integration will further assist airlines and other mobility/logistics providers to overcome the likely disruption due to more frequent and increasingly severe bad weather events (as part of global warming).

From a network strategy perspective, we foresee that the hub and spoke model will continue its decline but will maintain use in select cases, focusing on regional distribution and the major, volume based and cost sensitive trunk routes of bulk carriers of the Middle East and Asia. Connections will continue to be built between origin and destination points and airlines will downsize their routes to match the right aircraft to the right route, which is what customers want (Johnson et al., 2014). Longer and thinner routes will be opened to meet the needs of consumers with Qantas' ultra-long-haul (ULH) project Sunrise being first evidence of that development. This "hub busting" service will seek to bypass competitor hubs (e.g., Singapore Airlines, Emirates and others) and charge a premium to do so, offsetting the costs associated with operating such a long flight. In itself this also represents an interesting market segmentation exercise, with Qantas looking to serve those customers it can do so profitably, instead of fighting a losing battle against the hub airlines with better-cost economics. Aircraft manufacturers will need to match these demands and developments of smaller aircraft that are less risky to fill but more capable in terms of reach will be required.

In a more micro sense, congestion in cities will drive other decisions. Airports are significant generators of urban trips and growing airport usage will increase pressures on urban transport networks. Greater integration into extant public transport networks will be required, through for example better connections of airports into their networks. This may take the form of better terminal design, positioning trains, and buses inside or near airport terminals to link directly with passengers going to or coming from flights. Other integrative measures such as better trip information and linked air and ground transport fares may be investigated. Though ground transport industrial reform may also be required to address competitive issues like the reliance of airports on car parking revenues, and the dominance of taxi and ridesharing services in ground transport services, the entrenchment of which prevents the development of more efficient transport options may need resolving.

For further discussion, see 10021: Demand for Passenger Transport and 10076: Competition of Air and High Speed Rail.

Improving Footprint, Enhancing Sustainability

Integration and providing more service to consumers is a sustainability measure insofar as preserving consumer value leads to financial viability, including in regional contexts. But of course, sustainability is also about the traditional sustainability focus areas, including carbon emissions and noise management. While it has been a significant topic in the early part of this millennium, environmental issues continue to be an area of focus for airlines, if not (at present) the wider political landscape. Recognizing that there does remain (and currently growing) a latent consumer preference for positive environmental activity and that air transport is a significant contributor to environmental damage to the atmosphere, air transport operators are still focusing on the impact they have. Now some of this is in obvious self-interest—reducing fuel consumption not only helps reduce the emission of greenhouse gases but also reduces fuel bills and the overall exposure to the risk of very volatile jet fuel prices (Merkert and Swidan, 2019)—but some of it is definitely a response to consumer preference, for example the recent trend to swap plastics out for other materials for in cabin items. But clearly some of this is driven by airlines wanting to at least appear to be good corporate citizens in the eyes of consumers, no doubt due to the desire by air transport operators to earn social licenses and derive the benefits that these deliver.

As mentioned earlier, future steps to better integrate ground transport into air transport operations are likely and may form a part of sustainability strategies across the industry. However, managing the substitution effect (i.e., increased potential for long-haul flights as a result of increased rail journeys making slots at airports available—e.g., Chiambaretto and Decker, 2012; Janic, 2011) is also needed. Airlines will continue to work as an industry to design their own emissions management program. At an individual level, air transport operators will continue to look to offer voluntary offsets for consumers to purchase as an add on (or in some cases bundled) good.

Though sustainability is not just about the environment. Air transport sustainability has come along way in the last few decades as airlines have been positioned as commercial entities in their own right, and not departments or agencies of governments provided for national interest (though many of these still exist). More and more commercial focus will be required to ensure that airlines can keep reducing the cost curve and keep ahead of declining revenues. Some operators are beginning to better identify consumer demand and are unbundling and bundling products to not only improve margins but also deliver what consumers want in a journey. This will continue to occur with more and more ancillary revenue generated through identifying what consumers want in and related to a flight, and what they want to buy. And the further development of low-cost carriers, into ultra-low-cost carriers, represents a different way of achieving sustainability, by reducing resource consumption further. However, this needs to be balanced with the lower price of flights leading to more overall consumption, and questions are being asked if the debundling of these services and management of them by the consumer is actually leading to lower journey resource consumption (as passengers need to get to far flung LCC airports from the city), or if the disaggregate journeys are more resource consumptive.

Perhaps one of the biggest changes, and for us one of the more opaque and unclear ones, is the future of power generation for the sector. We have had jet engines for close to 80 years now, in various formats, which in reality are extensions of earlier piston engines, which are still used across many aircraft types today. But these appear to be reaching technical limitations, with their further capability expansion limited by both engine and airframe design. We expect that alternative and renewable fuels will see more and more research and deployment as fuel costs continue to rise and more demand for less environmental impact increases. However, there will come a time where brand new ways of propulsion will need to be considered to support air transport growth. What is the successor to the jet engine?

For further discussion, see 10085: The Economics of Reducing Carbon Emissions from Air and Road Transport.

Big Data, Better Decisions?

Air transport operators and other players in the aviation supply chain have been collecting substantial volumes of data for as long as the industry has been in the air, given the relatively advanced state of technology used to collect this data. Aircraft engine manufacturers collect more data than any other industry with for example Pratt and Whitney (McKinsey & Company, 2019) generating 355 Petabytes per day (in comparison, on a daily basis Google processes in total 24 Petabyte only). We argue that in the future, even more data will be collected and analyzed for brand new purposes. Airlines generate so much data that they increasingly wonder why not become better in using it rather than leaving the floor to consultants, aircraft/engine manufacturers, and/or technology giants such as Google or Microsoft.

Big data is being used operationally, measuring climate (especially winds), aircraft performance, and a range of other variables to better design flight paths to reduce fuel consumption and travel time. While big data analysis in airline network planning is becoming somewhat matured now (Hausladen and Schosser, 2020), the intensive use of big data for consumer management has been a more recent development. Right now, airlines are building consumer profiles through their frequent flyer and corporate loyalty programs. Using the attraction of free flights, lounge access and other benefits, these programs are being used to generate nonairline revenue and the data collected is being used (in addition to selling it to or collaborating with data aggregators and retailers) to understand more and more about what and when people buy and at what price. Travel agents are being bypassed in preference for direct sales that airlines can control, and own the data generated by their customers (Fiig et al., 2015; Madalozzo and Fernandes, 2016). In perhaps a more interesting technological development, big data is also being considered to change the way airline pricing is generated, and in the future it may be the case that instead of having the multiple airfare classes we use today, individualized fares (perfect price discrimination) are made available for each passenger as they book, to deliver the flight they want at the best fare they are willing to pay. Through concepts such as blockchain, big data networks are also being used to increase data security, protecting individuals and companies alike from harmful attacks.

However, big data is limited by how it is harnessed. Across all users of big data, including airlines, more needs to be done to link big data sources so that the data can be interpreted (which may be difficult in the context of privacy concerns) and used instead of just collected and stored to gain electronic dust on the shelf. Information technology systems as we know them now need to be reconsidered and newer, more holistic approaches to data will be developed in the future. These systems need to not only reduce the interface issues between data sources and calculations/algorithms that use them (for example risk of human error while migrating information between systems), or subsequent planning, operational and reporting information flows, but also need to be able to highlight new data relationships that can represent efficiency/productivity gains and potential new markets.

Of course, the human element is significant in driving the reengineering these systems and air transport operators need to become more skilled at data management, more so than they are now. Indeed, some transport operators in other sectors are viewing themselves more as technology companies than transport companies, which while may be an extreme, is somewhat reflective of the direction that airlines will find themselves in. Management strategies may in the future need to be more encompassing of incorporating indicators that big data analysis can measure. Whether these are at top line levels like return on capital invested measures, or as subsidiary measures will depend on how significant the measure is to airline success. Overall, we see harvesting data being a strong trend in airline management, which will result in the actual operation of aircraft becoming a less and less important part of commercial viability of successful air mobility/logistics providers.

As with all big data exercises, however, there will be a point, where from a sustainability perspective, airlines will have to decide when a line needs to be drawn, and that some consumer behavior is best left unknown. While consumers are happy to trade their consumption behavior for more flights or more customized products, social responsibility suggests that big data holders, including airlines, will need to recognize the power they hold and manage this appropriately, lest it cause a case for policy intervention.

For further discussion, see 10222: Information Sharing and Business Analytics in Global Supply Chains.

Policy and the Rules of the Game

Air transport policy has seen substantial reformative change, and particularly at the international level, significantly more so than other forms of transport as governments have liberalized air services and routes, and have privatized state-run airlines. This pattern of deregulation is likely to continue where there are clear benefits of doing so, despite current protectionist tendencies. Though there is unlikely to be a rash of government owned airline sales, there will probably be some more collapses or significant restructures to

decrease airline size and activity to reduce government exposure. With airlines increasingly focused on mobility/logistics service provision and data managers there is scope for further regulatory change, particularly around national ownership rules and (bilateral) air service agreements. And just to go crazy for a minute, imagine a world where Uber provides future airline like services similar to those of American and Qantas in the future. Most probably not possible given the governance structure of aviation and the role of defense in all of this but who knows? What is more likely is that future policy decisions in the land transport sector will have more impact on the air transport than direct policy change in the air transport sector. Despite the hub and spoke model being under pressure from more point to point connections, the air–rail experiment may continue if rail operators are liberalized, giving airlines more options to work with on the ground. In Japan and Europe, introducing more competition into rail markets may provide the commercial impetus for partnerships to develop to make air–rail connections more attractive to consumers and also more viable, through encouraging new service pricing and integration (Merkert and Beck, 2020; Merkert et al., 2019). More direct regulatory measures may be taken in other jurisdictions like China.

For further discussion, see 10074: Airline Economics and 10075: Privatization and Deregulation of the Airline Industry and 10436: Airline Regulation

Air Freight, Slower but Cheaper?

In another less than clear vision, we see that air freight could see potential change. Freight has traditionally been more of a by-product of air passenger transport, utilizing available space, though some all freight carriers have arisen and some passenger airlines have established all freight arms (Merkert et al., 2017). However, experiences in other transport sectors show that maybe there are new ways of doing air freight that may see development. Freight, particularly consumer goods, between Asia and Europe has traditionally been carried by ocean going ships, with small, high value and/or time sensitive consignments carried by air (Alexander and Merkert, 2017). Recent rail developments have seen new transport options however with through trains from China to Europe, offering a faster trip than sea freight, but slower than air. Which freight consumers seem to be liking given growth in this sector. This suggests that consumers are willing to sacrifice time to save cost. We may see similar developments in air freight markets. From developing services that take longer to deliver, and perhaps even involving specifically designed aircraft which consume less fuel, air freight could expand its market through lowering costs, particularly in the presence of greater package delivery. We also expect to see air freight transport embrace pilotless technologies before passenger freight services do, arguably due to the public perception of safety and security which is seen to impact passenger far more than freight services.

For further discussion, see 10270: Air Freight Logistics and 10283: Eurasia Rail Freight, Enables and Inhibitors of Future Growth.

The Drone Revolution

The last but potentially greatest future for air transport is the impending revolution that we expect to see brought by drones (Merkert and Bushell, 2019b). Quite unrelated to airlines (other than having the potential to disturb their air space), new air transport markets in urban areas may become possible as the small airborne devices allow for new transport services to be offered (for a detailed literature on this see Merkert and Bushell, 2019b). Dry cleaning pickup, delivery of food, and delivery of that one thing forgot to get that one thing at the shops. Though in many cases this last forgotten delivery may not have been necessary as you may not have even been to the shops in the first place—it may be that you can order all of your needs online and have it all delivered to your door, balcony or other suitable drop off location. This will have obvious consumer benefits though it will no doubt have impacts elsewhere in the economy. What might it mean for retail stores and the retail supply chain? What might it mean for road congestion if air journeys replace road couriers? What might it mean for the social functioning of a society that now does not need to go out as much to shop or run everyday errands? What will people do with the timesavings? Or will cheaper transport options increase total package delivery, much like ridesharing has increased road usage? What if those drones are unmanned? What will happen to today's truck, bus, and train drivers? Will they all be upskilled? It seems clear that this revolution is going to require new management approaches and possible policy intervention. Furthermore, managing so many of these small flights in congested airspace will require new technological solutions and ways of managing the airspace efficiently and safely. Discussion has commenced about how Low Altitude Airspace Management Systems may extend controlled airspace down from current levels to incorporate drones into the extent airspace management system (Merkert and Bushell, 2019b). At this point, it seems fair to say that a future of drone operations will be more complex but also more exciting in terms of opportunities that it will create. As usual innovation should be embraced and supported by public policy wherever possible.

For further discussion, see 10004: The Future of Urban Freight, 10093: How will autonomous vehicles impact car ownership and travel behavior for more discussion on this, and 10272: Drones in freight transport.

Beyond the Barriers

Flight has always been challenging barriers, from the barrier of the third dimension itself to various speed and endurance records. And there are more barriers to be breached, with the speed and gravity barriers being key focus areas. We have been past the airspeed

barrier before, with the Concorde providing faster than sound travel, though we retreated from this in the face of increasing fuel prices. But it will come a time where we push past this barrier again, as we seek to reduce travel times more. ULH flights as operated by Singapore Airlines, Qantas, and likely others can only do so much at their speeds that are just under sound, but eventually consumers will reach a point where they will pay more to arrive at their destination sooner and make more productive use of their time.

Perhaps most interestingly from the perspective of the wider community, it may be that air travel will lose its air component as it moves beyond the atmosphere and take us into space. Now, whether this is in the remit of this chapter or not, this exciting new development does represent an extension of air transport and represents brand new potential for the sector. Currently, space flight is being pursued as a project to provide flights to the wealthy, but no doubt it will move past this as the cost curve of the technology falls and it will provide flights to a broader range of the public and even to freight services. Now this might be a way to provide faster point-to-point services, and a few-hour trip from Sydney to London might be possible. Space flight might be the successor to the jet engine. Though this will also lead to altogether different destinations and into destinations that have to this day been the domain of science fiction only.

Final Observations

Air transport is already a highly developed and advanced industry but as we have shown, it has plenty of latitude and room for growth. The future for some sections of the industry are quite mature and in the absence of any significant disruption, such as transportation beam devices, these segments will continue to mature and refine their current operation. Now, this hypothecation is just one view of a set of possible future events. A great many things may impact this view, especially if our society experiences changes that are not within our foresight. But regardless of this, changes in aviation and air travel and the expanding use of the third dimension will continue to impact on us and our development as a society with many intended benefits but potentially also a number of unintended outcomes. There may also be a future where transnational air mobility and/or logistics providers with aircraft/air borne vehicle operation will become, while continuing to provide services as the safest mode of transport, a secondary function of successful and consumer centric and data driven air transport management.

References

- Airbus, Cities, Airports & Aircraft, Global Market Forecast (GMF) for 2019-2038, 2019, Blagnac Cedex, France.
- Alexander, D., Merkert, R., 2017. Challenges to domestic air freight in Australia: evaluating air traffic markets with gravity modelling. *J. Air Trans. Manag.* 61, 41–52.
- Baker, D., Merkert, R., Kamruzzaman, M., 2015. Regional aviation and economic growth: cointegration and causality analysis in Australia. *J. Trans. Geograp.* 43, 140–150.
- Chiambaretto, P., Decker, C., 2012. Air–rail intermodal agreements: balancing the competition and environmental effects. *J. Air Trans. Manag.* 23, 36–40.
- Chowdhury, S., Hadas, Y., Gonzalez, V.A., Schot, B., 2018. Public transport users' and policy makers' perceptions of integrated public transport systems. *Trans. Policy* 61, 75–83, doi:10.1016/j.tranpol.2017.10.001.
- European Commission. 2019. Reducing emissions from aviation. https://ec.europa.eu/clima/policies/transport/aviation_en. (accessed 07.12.19).
- Fiig, T., Cholak, U., Gauchet, M., Cany, B., 2015. What is the role of distribution in revenue management - Past and future. *J. Rev. Pric. Manag.* 14 (2), 127–133.
- Hausladen, I., Schosser, M., 2020. Towards a maturity model for big data analytics in airline network planning. *J. Air Trans. Manag.* 82, 101721.
- IATA, Annual Review, 2019, 75th Annual General Meeting, Seoul.
- Janic, M., 2011. Assessing some social and environmental effects of transforming an airport into a real multimodal transport node. *Transpor. Res. D Trans. Environ.* 16 (2), 137–149.
- Johnson, D., Hess, S., Matthews, B., 2014. Understanding air travellers' trade-offs between connecting flights and surface access characteristics. *J. Air Trans. Manag.* 34, 70–77.
- Madalozzo, R., Fernandes, P.C., 2016. Do strategic behaviors link travel agencies in Brazil? *BAR* 13 (3), e160018.
- McKinsey & Company, 2019. Data in disruption: travel, transport & logistics. Presentation as part of the Professional Development Workshop on Using Transportation Data for Advancing Management/ Strategy Theories at the 2019 Academy of Management Meeting, Boston.
- Merkert, R., Beck, M., 2020. Can a strategy of integrated air-bus services create a value proposition for regional aviation management? *Transport. Res. A Policy Pract.*, In Press.
- Merkert, R., Bushell, J., 2019a. Long-distance transport service sustainability: management and policy directions from the airline perspective. In *A Research Agenda for Transport Policy*. Edward Elgar Publishing, Cheltenham, pp. 89–98.
- Merkert, R., Bushell, J., 2019b. Revolution or epidemic? A systematic literature review on the effective control of airborne drones, Proceedings of the 2019 Air Transport Research Society 23rd World Conference, Amsterdam.
- Merkert, R., Bushell, J., Beck, M., 2019. Collaboration as a service (CaaS) to fully integrate public transportation – lessons from long distance travel to reimagine mobility as a Service. *Transport. Res. A Policy Pract.*, In Press.
- Merkert, R., Swidan, H., 2019. Flying with(out) a safety net: financial hedging in the airline industry. *Transport. Res. E Logist. Transport. Rev.* 127, 206–219.
- Merkert, R., Van de Voorde, E., de Wit, J., 2017. Making or breaking - key success factors in the air cargo market. *J. Air Trans. Manag.* 61, 1–5.

Air Transport and Its Territorial Implications

Lanfranco Senn, Bocconi University, Milano, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	208
The Relevance of the Territorial Scale in Air Transport	208
The “Spatial” Determinants of Demand for Air Transport	209
The Vertical Integration between Airports and Airlines’ Services	210
The Impacts of Airports on the Surrounding Territory	211
Accessibility to Airports	211
The Management of Airports	211
Airports as “Natural Monopolies”	212
The “Generalized Cost” of Air Transport	212
Conclusions	213
See Also	213
References	213
Further Reading	213

Introduction

Air transport keeps growing and will continue to grow. But it grows in a highly variable way, according to the territorial distribution of demand. The variability of demand depends on several factors. The first is indeed the distance between the origins and the destinations that the demand wants to connect. Distances to be covered depend, in their turn, on the size and the shape of the countries involved and vary of course if mobility is global or intercontinental, continental, or national. A second factor of demand variation is the development level of the places that are origins and destinations of any transport connection: rich and developed areas are likely to generate a higher volume of demand than poor and less developed areas.

A third relevant variable in determining demand for mobility is its motivation: to simplify, as far as passengers are concerned, business or leisure (Graham et al., 2008) represent two very different segments of demand, as well as freight transport does.

A further factor—of a territorial nature—in driving mobility is the so-called “catchment area” of the infrastructure (airport, station, port) that allows the departure of a trip or a travel. It will depend on its accessibility and on its transport function: in the case of air transport—on which our attention is concentrated here—what matters for example are the characteristics of the airport (hub, point-to-point) and the frequency of the flights it generates (regular or episodic, as in the case of touristic charter flights).

This paper is structured in eight short sections. The first deals with the relevance of the territorial scale in air transport. The following section analyzes the “spatial” determinants of demand for air transport. Then, the vertical integration between airports and airlines’ services is discussed. The fourth section underlines the impacts of airports on the surrounding territory, followed by the issue of “accessibility” to airports. The spatial implications of the management of airports are dealt with, followed by the discussion of the “natural monopoly” characteristics of airports. Finally, the specific issue of the “generalized cost” applied to air transport is presented.

The Relevance of the Territorial Scale in Air Transport

The “territorial scale” of air transport depends on these variables, indeed strictly interdependent on each other (Mukkala and Tervo, 2013). It may be noted that in the case of air transport—as well as in the case of other transport modes, such as road, railway, or water—the territorial scale is relevant from a double point of view: the infrastructural one (location, catchment areas) and the services provided by the infrastructure (connections, motivations, frequency). Any approach to air transport analysis implies to keep in mind that air mobility is only allowed by a strict “vertical integration” (see below) between these two phases or aspects that cannot be separated (Fu et al., 2011). The analysis that follows examines in turn the variables that allow to understand the spatial (territorial) implications of air transport.

The first factor that determines the demand for air mobility is the distance that separates any origin and destination “couple” that must be connected. It may be observed that these distances may not be covered necessarily neither by a unique mode of transport nor by a direct connection but through a multimodal (road, rail, air, water) and networked transport system. It is obvious that distances between origins and destinations heavily affect the modal choice.

At the extremes, bicycles will be chosen for short distances, air for very long distances. For medium distances, road and rail will tend to prevail, although there is not a clear cut “market” separation among transport modes as far as distances are concerned.

Modal competition in fact generates overlapping of the spatial market share of different transport modes. In addition, the growing trend of multimodality to satisfy complex mobility demands further complicates the identification of the most frequent distance covered by different modes. However, a sort of rule of thumb states that air is seldom used below the distance of 4–500 miles.

Having said that, it is evident that intercontinental/international mobility, above the threshold of 400–500 miles, is usually met by air, with an increasing competition played by high-speed rail. As far as international mobility is concerned, the relevant variables to be taken into account are the size and the shape of the countries involved. In large countries (such as USA, China, Russia, Brasil, Australia, to cite only a few) air mobility keeps being privileged (always beyond the cited threshold) (Wittman and Swelbar, 2014). But also in countries whose shape is “long” (such as Italy, Germany, Argentina, India) air transport is indeed the preferred transport mode. On the contrary, international flights will not conquer any market share and compete with other transport modes in small countries or in islands (the Baltic countries, the Netherlands, Belgium, Malta, Cyprus, Island, Ireland to cite only European countries).

An almost trivial conclusion is that geography matters in the identification of the role of air transport in meeting spatial demand for mobility.

The development level of origins and destination of air transport is the second relevant variable in ascertaining its importance. Air transport tends in fact to be a concentrated phenomenon, due to the fact that airports—of different size and role—play the fundamental role of being the places for take offs and landings. Therefore, they are located in areas—mainly close to large cities and metropolitan agglomerations—where inward and forward demand is concentrated (Blonigen and Cristea, 2015; Sheard, 2014). The fortune of air transport is therefore strictly linked with sites that show a high level of economic and social development. It is likely that development occurs—more and more frequently—in cities, the places where innovation and a great socioeconomic dynamism take place nowadays. It is well known that urbanization is a process that the world is facing today and will probably face in the future. On the other hand, large cities and metropolitan areas are growing in size and therefore provide a significant market for air transport. The combination of the level of development and of the size of spatial agglomerations drives the location of airports in their proximity as they serve dense “catchment areas.” A potential conflict for these locations is raised by environmental problems, as take-offs and landings of airplanes cause indeed air and acoustic pollution (external diseconomies). In some cases, creative solutions have been achieved, as in the case of the Kansai airport (Osaka) where the airport has been located on an artificial island, some twenty miles in the sea, linking through a bridge the city with the airport. For the same reason, in some large metropolitan areas, more than an airport serves its passengers (and freight), one being located within the area and the other at a certain distance from downtown. This is the case, for instance, of Milan (respectively for Linate and Malpensa airports); of Paris (Orly and Charles de Gaulle); of Tokyo (Haneda and Narita).

The “Spatial” Determinants of Demand for Air Transport

Demand for mobility is very differentiated according to its motivations. As far as passengers are concerned the two main drivers of demand are indeed business and leisure, with different roles played by the time length of planned mobility. Easy examples are represented by the fact that short “business trips” tend to use the quickest modes of transport that help saving time. Combined with the distance and agglomeration issues already commented, the quickest competing modes are air and high-speed rail. On the contrary, longer business trips will tend to favor the most efficient and comfortable modes, as the incidence of travel time is less relevant. For leisure mobility, the arguments are similar but do involve also other transport modes, such as road or even sea (it is the case for cruises). The territorial scale of the catchment areas, in both cases of business and leisure, also varies according to the lengths of the trips. When short trips take place the closest airport to the passenger is preferred and therefore the catchment area tends to be smaller because proximity matters. When longer travels are planned, the passenger will try to reach the airport that presents the highest range of connections: in this case the catchment area of the chosen airport may be very wide and imply a multimodal access (either by road or by rail). The hubbing function of the airport (see below for the concept of hub) may further increase the size of the catchment area of the airport that provides a wider choice of routing to achieve the final destination.

Air transport—for some intermediate and final products, generally speaking of high value, relatively small size and that needs a rapid delivery—serves also the mobility of freight. Transporters of good that will inevitably use the road mode for the first and last mile segment of the overall movement will be confronted with the choice of the closest airport for quick access and the airport which serve a wide network of destinations. The strategy of the logistic operator—that includes time and distances from airports as relevant variables—will therefore affect the role of air transport services (Button and Yuan, 2013).

It is evident also in the case of freight mobility that territorial scale is a relevant variable for the competitiveness of air transport. Again, the areas with a high concentration and development of manufacturing production are likely to be those that present a major role in air freight transport.

The concept of “catchment area” has been frequently referred to as the area from which airports (the infrastructure) and services (the air connections) served by airports attract and capture their market, both passengers and freight (Mellander and Holgersson, 2012). Its width and intensity of attraction depend on several factors. In addition to those already hinted to—demographic and production concentration; level of “local development”; different motivations in demand for mobility—there are other factors that contribute to identify the catchment area of an airport.

The most important is accessibility, which concerns both a spatial and a time dimension. From the spatial point of view, an airport is “accessible” if other infrastructures allow to reach it; that is, if the road and railways systems may convey traffic to it. Airports

that show the widest catchment areas are those that lie in the gravitational center (poles) of capillary networks of roads and railways and act as main “nodes.” Roads and railways may compete between themselves to bring passengers and freight to the airports, but they are only a necessary but not sufficient condition to define their catchment areas. In fact, accessibility is not granted by the sole existence of the “feeding” infrastructure: it will also depend on the services that will be provided. If roads leadings to airports are congested, for example, times to reach the airport risk to be too long and uncertain for private cars and public buses (especially during peak hours). In that case, railways will be favored, but only if the frequency of their service is sufficiently high, punctual, and well scheduled. Sometimes, as far as railways are concerned, these services may be provided by dedicated shuttle services; or being inserted in medium long distance networks with a specific stop in (or underneath) the airport (such as in Paris Charles De Gaulle, Amsterdam, Narita, Rome) to facilitate the multimodality of travels and accessibility.

The spatial distribution of potential users of air infrastructure and services, of course, may not be uniform: it may occur in concentrated areas or along directions/lines (East-West or North-South), to simplify. In this last case, the road or railways infrastructure that give access to the airport will “shape” the catchment area in a much simpler way (such as in the case of Lyon, Bologna, Zurich) and reduce the investments in providing to airports a networked system of accessibility.

Catchment areas of air transport infrastructure and services—due to all just made considerations—may obviously partially or totally overlap. The distances among the locations are thus relevant: the overlapping may be partial if airports are sufficiently distant from each other; or almost total if their location is within the same metropolitan agglomeration: the main example is London, where Heathrow, London City Airport, Gatwick, and Stanstead serve almost the same potential spatial market.

The London case is a very good example for a further consideration. Demand for air mobility may indeed cause the overlapping of catchment areas from the geographical point of view. However, from the functional point of view, the four airports in London play a different role. To simplify, Heathrow is an international and intercontinental hub airport, while other airports play either a role of European hubs or a role of point-to-point connections both at a national and continental scale. A hub airport is an airport that receives feeding flights from different origins and forwards the passengers to different destination, playing a “sorting” function at different territorial scales: the intercontinental one is the most desirable for air companies that are thus able to fill the most expensive long-distance flights and fully exploit their seat capacity. But a minor hubbing role may even be played at the continental (North America, Europe, Asia), or even at the national level (Chicago in the United States, Rio de Janeiro in Brasil, Orly in France, Frankfurt and Munchen in Germany, Fiumicino and Malpensa in Italy).

A point-to-point airport, as the word says, is an airport that connects directly one origin with one destination.

The different role, from the territorial point of view, is self-evident. In the hub case, airports benefit from a very large catchment areas, while in the case of point-to-point connections the catchment area tends to be more local and limited.

The Vertical Integration between Airports and Airlines' Services

Big air companies—due to the “vertical” integration of air transport between infrastructure and services—tend to concentrate their activities in hubs of intercontinental range, because it is convenient from the economies of scale point of view: in hubs they can in fact locate the “maintenance” and operational activities of their fleets of airplanes (Homsombat et al., 2014). To give some examples, Air France, exploits the hub airport in Paris Charles De Gaulle; KLM in Amsterdam; Alitalia in Rome Fiumicino; Lufthansa both in Frankfurt and Munchen; British Airway in London Heathrow; Iberia in Madrid; Some worldwide companies establish their hubs of reference also in different continents, such as Delta, which is based mainly in Atlanta in the United States and in Amsterdam Schiphol and Paris Charles De Gaulle in Europe; or as Qantas in Sidney and Dubai.

In identifying the main hubs, one must take into consideration also the differences existing between regular airlines and low cost companies. These are companies that try to meet air transport demand at very cheap tariffs, to capture the market of low income, potential users (young people above all) that still want to use air for their mobility. They operate mainly in the leisure segment of demand, with an increasing share also in the business segment; provide mainly a point-to-point service; are able to keep costs (and tariffs) low due to the reduction of accessory services to their customers and mainly online booking. They tend to serve only the intracontinental demand for air transport as this allows to increase the efficiency of their operations (e.g., number of flights per day of the same plane).

Low cost companies have rapidly increased their market share in all continents. The largest ones are Southwest Airlines in the United States; Ryanair and easyJet in Europe; the Air Asia Group in Asia with several millions of passengers served every year. Their operational hubs are respectively in Dallas; Dublin; London Gatwick and Malpensa; Kuala Lumpur.

When dealing with the territorial implications of the vertically integrated (infrastructure and services) air transport activities, one must also take into account the locational attraction of economic activities in the proximity to airports (gravitational effect) and the impact that air transport exerts on a more or less wide extension of the territory surrounding airports.

Playing on words, one might say that the “scale” of air transport (size of airports in terms of number of connections offered and volume of passengers and freight moved) is highly correlated with the “scale” (extension) of the territory affected (Redondi et al., 2013). Such a correlation is the outcome both of a centripetal and centrifugal effect. The bigger is an airport the larger is the attraction of firms (Beifert, 2016) desiring to locate in its proximity, either on site or off site, that is within the airport area or nearby the airport area. Headquarters of firms, or even production factories, stores and logistic activities—especially if the airport plays a hub function—are interested in locating as close as possible to the airport to reduce times of access to the airport and its connecting services to a wide number of destinations (markets, for the firm) for managers, employees and freight.

In some cases—one very relevant example is Amsterdam Schiphol—this has given rise to big real estate developments “associated” with the airport. It is obvious that this attraction effect is not so relevant in small or point-to-point airports of small size and providing only few connections: economies of scale, neither from the airport point of view nor from the territory involved, do not play a significant role in these cases (Freestone and Baker, 2011).

The Impacts of Airports on the Surrounding Territory

The “centrifugal effect” of airports on their surrounding territorial scale is captured by impact studies (Allroggen and Malina, 2014; Button, 2010). There are consolidated methodologies—for instance provided by ACI Europe and several case studies, such as the one for the Malpensa airport—that outline how to evaluate this effect. Impacts may concern totally new (Tveter, 2017) or additional investments (for instance, new runways or terminals) (Breidenbach, 2019) as well as management of services impacts: through complex methodological techniques—such as input—output analyses—it is possible to estimate the direct, indirect and induced effects of on site (including the role of air companies operating in the airports) and off site impacts on generated turnover, value added, employment, imports by investments, and current expenditure. Some studies estimate also the fiscal impacts.

Direct impacts measure the effects of the expenditure directly made for purchasing goods and services during the construction or the management phases; indirect impacts are those that are generated by firms directly implied through the procurement of goods and services that they must meet in their turn. Induced impacts are the effects of consumption expenditure made by the employees of the firms directly and indirectly involved in the construction or management phases. These complex estimates allow to measure the “multipliers”—often significant—of each dollar spent for new or additional investments. The range of such multipliers is between 1.4 and 1.8 in most cases. Of course, the territorial impacts may be very concentrated in the area of the airports—if its economic structure is highly diversified—or present some dispersal (leakages) effects if the structural economic features of the area are weaker or less interdependent. The territorial scale of multipliers will therefore depend on the development level of the area in which the air transport activity is high (as in developed metropolitan areas) or low (as in less developed or still developing areas).

Accessibility to Airports

An important effect on the territorial implications of air transport is played by “accessibility” to airports. Accessibility is defined as the easiness with which incoming and outgoing passengers and freight may reach an airport with different modes (road and/or rail).

Accessibility may be evaluated in terms of existence of infrastructure and their level of services (congestion), both of which contribute to the amount of time needed to reach or leave the airport. In particularly dense metropolitan areas, road (with private cars or with public bus transport) and rail (with the regular network or with dedicated shuttle service) compete with each other and provide to air transport users alternative modal choices. Given the competition in trying to increase the modal accessibility of airport to the surrounding territories, it is not easy to define a unique “model of connectivity” of rail compared with road, which maintains a significant prevalence in this particular “market.”

However, examining a certain number of European case studies (Milan Malpensa, Zürich, Geneva, Frankfurt, München, Wien, Oslo), some correlation between the characteristics of rail connectivity to airports and the relative market share in airport accessibility may be found.

For instance, it occurs between the type of travelers and their preferences for mass public transport services, if the travel is long and made by nonresidents.

On the contrary, it is not possible to establish a univocal and solid correlation between the availability of a rail service and its times of access, costs, frequency, and comfort, although these factors are indeed crucial for rail competitiveness (Shujie and Xiuyun, 2012). The catchment area of an airport, in fact, is so variable in terms of preferences and travel needs that it does not seem possible to identify a model of optimal services in all contexts.

Some evidences, however appear, both in the literature and in actual practice.

- Rail is preferred when times of access are significantly lower than those achieved with other modes (road) and if price is aligned with the latter.
- The implementation or the reinforcement of a rail service to airports accessibility modifies passengers’ habits and parking facilities close to the airport.
- Passengers have a reduced preference for train changes.
- The market share to rail accessibility changes if rail services are differentiated in express, suburban. or long haul features.

The Management of Airports

To meet all the characteristics that an efficient air transport activity needs to serve demand and its territorial distribution, a crucial issue concerns its management. It has been pointed out that air transport is a vertical integrated activity, that is, an activity which needs at the same time to provide services (flights, managed by airline companies) and infrastructure (airports), the necessary but not sufficient condition to supply mobility services. In other words “vertical integration” means—in our case—that for the success of air transport in competition with other transport modes and for the sake of the economic returns of the different stakeholders, the management of the whole activity—separately for each of them—has to take into account both the infrastructural and the service

side. In fact, airport managers may not plan the development of the airport without considering what airlines will decide to establish their operational basis in the airport, how to allocate an adequate number of slots (Madas and Zografos, 2006) to the different airline companies. In turn, airline companies, will choose the airports that provide the services they need and whose catchment area is adequate for their desired market.

Focusing the attention on the management of infrastructure and avoiding to deal with nonspatial (i.e., “internal” to the airport) aspects (Brueckener, 2003), it is worth concentrating on the nature of airports and on some issues of their management that are often hinted to as “natural monopolies.”

Airports as “Natural Monopolies”

A natural monopoly is a monopolistic market firm, in which a firm is able to manage the entire supply of a service (or a good) at a lower cost with respect to any other firm. Usually, such a monopolistic form has a large size and benefits of very low average costs, due to economies of scale and growing scale returns. The larger is the production and the supply of the service, the lowest is the average cost of production and therefore its average cost and supply price of the service provided by the monopolistic firm. The potentially competing firms are not able to enter the market, as they are small and/or new comers: in fact, they, inevitably meet higher production costs. The natural monopoly occurs when the firm deals with an exclusive production factor (land, in the case of airports). The market is characterized by natural barriers that prevent other firms to enter that market. Despite its lower production costs, it is not obvious that the monopolistic firm will provide the service at a lower price: it will enable to sell the service at a high price, lowering it only when a new competing firm will try to enter the market. In our case, the firm (the airport) must manage large scale production “plants” serving a significant demand of services (provided by the airlines insisting on that airport). The vertically integrated firm must realize an infrastructure whose catchment area and accessibility are significant and efficient.

In these cases, an antitrust policy favorable to competition is inefficient, as it might cause the disappearance of the service and the growth of the market price. For this reason, public utility services (such as transport activities, air transport in particular) are often managed by public firms or by large private firms, which operate in a State regulated market to guarantee the presence on the territory of at least a firm, able to manage efficiently the plants (airport). The regulator may pretend that some service conditions are respected, such as the application of prices or tariffs at low level and the quality of the service (including a range of connections). To provide efficient air transport services in a particular area the State/Regulator may launch a tender to find a firm that—at those conditions—provides the services as a natural monopolist. In this way, competition is not warranted “in” the market, but “for” the market (Forsyth et al., 2004).

The final aim of the State is to cover the entire national territory with a plurality of airports (natural monopolies), trying to reduce at a minimum the overlapping of their catchment areas.

The natural monopolistic feature of an airport has been mentioned because of its spatial implications: in fact, the spatial subdivision of the airports market due to the accessibility to airports—and the consequent potential modal competition with rail—may deeply affect the economics and the management of air transport.

In general, airport managers will have to deal with all the spatial variables hinted to at the beginning of this note: that is, time and physical distances among origin and the destination of the flights served by the airport; level of development of the served destinations and of the area where the airport itself is located (Hu et al., 2015); the relevant catchment areas, modal, and intermodal accessibility. In addition, of course, to all the organizational on- and off-airport services provided to capture an increasing demand.

Finally, it is worth remembering that the horizons of airport management must include also possible agreements with the stakeholders of other modal infrastructure and services (railways, parking, car sharing). Such agreements help the demand—that air transport managers must interpret and anticipate—to identify its generalized cost of transport to evaluate its modal choice.

The “Generalized Cost” of Air Transport

The “generalized cost of transport” is a typical cost adopted in transport economics that measures not only the monetary cost faced by passengers and freight by using a particular transport service, but also the times implied in the door-to-door travel. Different modes present different generalized costs. In fact, if we take into consideration the different phases of mobility, we must measure the “first mile” phase (to reach a railway station, an airport, or even one’s car parking), the actual travel time, the “last mile” phase to reach the final destination. It is an everyday experience that, driving a personal car, the first and last mile phases imply an almost negligible components of “time costs,” but a longer time of travel. While, if one chooses to move by train, a longer time is needed to get to the station (and arrive with a sufficient advance to make sure to get the scheduled train), to a shorter time of actual travel and a further time from the station of arrival to the final destination. In the case of air transport, the advance with respect to the flight take off, in addition to the time to get the airport, is due to the check-in procedures, that change according to airports and type of flight (national, international, low cost, charter, and so on). On the contrary, the travel time is indeed shorter than by car or even by train. Finally, the “last mile” (exit) time will depend on the luggage delivery time and the distance of the arrival airport to the final destination. From these rough descriptions of the different segments of a travel door-to-door using different modes, it appears that the actual competition among modes is not only determined by the comparison of monetary costs, but it shows that is in fact the outcome of a choice is based on the generalized cost of transport, which is the sum/combination of money and time (and comfort, although difficult to estimate in economic terms).

It may be finally noted that income elasticity to the monetary cost and the value of time is not equal for all types of travelers: business men, tourists, students, and retired people show a different value of time and a different income elasticity to tariffs and therefore will have a different estimate of their generalized cost of transport.

Conclusions

Air transport, due to vertical integration of airports and airline services, has a profound impact on different scales of territory.

At the “local” scale—more or less corresponding to the catchment areas of airports—the impact has several features, essentially linked to the economic and social activities the air mobility implies and enables.

The location of firms is attracted by the existence and the size of an airport and by the connections that it allows. People living within the catchment area of the airport and willing to move by air are also subject to the nature of the airport (hub at point-to-point) and the connections it serves, both for business or for leisure purposes.

The impact at the local scale on passengers and freight mobility is of course stronger if efficient is the accessibility to the airport, made possible by the network of the feeding modes (road and rail). The generalized cost of air transport—in as much as it concerns both monetary and access time—is a relevant driving force in defining the scale impact.

At the wider, “global”, scale—domestic, continental, or intercontinental—effects are provided by the connection that airports allow (also in terms of flight frequency to the different destinations), and the efficiency of their operating management. Two examples are the provision of the relevant services to the different segment of demand and the planning of the slot framework for airline companies.

See Also

Aviation Economics; Introduction to Air Transport; The Future of Air Transport; Transport Modes and Globalization; Transport and International Trade; Transport Modes and Accessibility; Competition of Air and High-Speed Rail; Air Freight and Economic Development; Transport Cost and Location of Firms; Wider Economic Impacts of Transport Investments; The Economics of Airport Slot Allocation; Airport Management; Airport and Airport Network Planning; The Generalized Cost of Transport

References

- Allroggen, F., Malina, R., 2014. Do the regional growth effects of air transport differ among airports? *J. Air Transport Manag.* 37, 1–4.
- Beifert, A., 2016. Regional airports' potential as a driving force for economic and entrepreneurship development – case study from Baltic Sea region. *Entrepren. Sustainab. Issues* 3 (3), 228–243.
- Blonigen, B.A., Cristea, A.D., 2015. Air service and urban growth: evidence from a quasi-natural policy experiment. *J. Urban Econ.* 86, 128–146.
- Breidenbach, P., 2019. Ready for take-off? The economic effects of regional airport expansions in Germany. *Regional Stud.* 2019.
- Brueckner, J., 2003. Airline traffic and urban economic development. *Urban Stud.* 40 (8), 1455–1469.
- Button, K., 2010. Economic aspects of regional airport development, WIT Transactions on State of the Art in Science and Engineering, vol. 38, WIT Press, ISSN 1755-8336 (online).
- Button, K.J., Yuan, J., 2013. Airfreight transport and economic development: an examination of causality. *Urban Stud.* 50 (2), 329–340, February.
- Forsyth, P., et al., 2004. *The Economic Regulation of Airports*. Ashgate.
- Freestone, R., Baker, D., 2011. Spatial planning models of airport-driven urban development. *J. Plann. Liter.* 26 (3), 263–279.
- Fu, X., Homsombat, W., Oum, T.H., 2011. Airport-airline vertical relationships, their effects and regulatory policy implications. *J. Air Transport Manag.* 17, 347–353.
- Graham A., et al., 2008. *Aviation and Tourism. Implications for Leisure Travel*. Ashgate.
- Homsombat, W., Lei, Z., Fu, X., 2014. Competitive effects of the airlines-within-airlines strategy – pricing and route entry patterns. *Transport. Res-E* 63, 1–16.
- Hu, Y., Xiao, J., Deng, Y., Xiao, Y., Wang, S., 2015. Domestic air passenger traffic and economic growth in China: evidence from heterogeneous panel models. *J. Air Transport Manag.* 42, 95–100.
- Madas, M.A., Zografos, K.G., 2006. Airport slot allocation: from instruments to strategies. *J. Air Manag.* 12 (2), 53–62.
- Mellander, C., Holgersson, T., 2012. Up in the air: the role of airports for regional economic development. *Ann. Regional Sci.* 54 (1).
- Mukkala, K., Tervo, H., 2013. Air transportation and regional growth: which way does the causality run? *Environ. Plann. A* 45 (6), 1508–1520.
- Redondi, R., Malighetti, P., Paleari, S., 2013. European connectivity: the role played by small airports. *J. Transport Geogr.* 29, 86–94.
- Sheard, N., 2014. Airports and urban sectoral employment. *J. Urban Econ.* 80, 133–152.
- Shujie, Y., Xiuyun, Y., 2012. Air transport and regional economic growth in China. *Asia-Pacific J. Account. Econ.* 19 (3).
- Tveter, E., 2017. The effect of airports on regional development: evidence from the construction of regional airports in Norway. *Res. Transport. Econ.* 63.
- Wittman, M.D., Swelbar, W., 2014. Capacity discipline and the consolidation of airport connectivity in the United States. *Transport. Res. Rec. J. Transport. Res. Board* 2449 (1), 72–78.

Further Reading

- Barbot, C., 2009. Airport and airline competition: incentives for vertical collusion. *Transport. Res. B* 43 (10), 952–965.
- FAA (U.S. Federal Aviation Administration), 1999. *Airport Business Practises and Their Impact on Airline Competition*. FAA/OST Taskforce Study.
- Fu, X., Zhang, A., Lei, Z., 2012. Will China's airline industry survive the entry of high-speed rail? *Res. Transport. Econ.* 35, 13–25.

Madas, M.A., Zografos, K.G., 2008. Airport capacity vs. demand: mismatch or mismanagement? *Transport. Res. A* 42 (1), 203–226.

Malloom, D., 2003. Auctioning capacity at airports. *Util. Policy* 11 (1), 47–51.

Verhoef, E.T., 2010. Congestion pricing, slot sales and slot trading in aviation. *Transport. Res. B: Methodol.* 44 (3), 320–329.

Wang, K., Gong, Q., Fu, X., Fan, X., 2014. Frequency and Aircraft Size Dynamics in a Concentrated Growth Market: The Case of the Chinese Domestic Market.

Next Generation Travel: Young Adults' Travel Patterns

Tobias Kuhnimhof*, Scott Le Vine†, *Institute of Transport and Urban Planning, RWTH Aachen University, Aachen, Germany; †Department of Geography, SUNY New Paltz, New Paltz, NY, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	215
Young Adulthood: A Phase of Socio-Economic Transitions	216
Advancing to a New Mobility Life Stage: More Options and More Travel	217
Intra- and Intergenerational Changes of Travel	218
Intergenerational Changes of Mobility Patterns Among Young Adults	219
Concluding Remarks	219
References	220

Introduction

This contribution focuses on travel behavior of young adults, defined here as the age group between 15 and 35. In most modern societies, this young adult period of life is the age bracket during which several important life changes take place. This usually includes leaving school, moving out from the parental home, attaining professional or academic education, getting one's first job and starting a family of one's own (Blumenberg et al., 2012). As for mobility, the teenage years in particular are also characterized by increasing mobility options which go along with maturing physical and cognitive skills (Chatterjee et al., 2018; ifmo, 2013; Kuhnimhof et al., 2019). This specifically includes the acquisition of driving licenses for motorcycles and cars. By the age of 18 most young adults around the globe have the right to obtain a driver's license, and in countries where economic conditions permit the majority makes use of this option (Kuhnimhof et al., 2012).

These consequential shifts in individual biographies and associated mobility options result in substantial travel pattern changes occurring during the young adult life stage. On average, in industrialized countries individual travel grows rapidly with increasing age between the ages of 15 and 35. In most cases this is primarily in the form of significant increases in driving (Chatterjee et al., 2018; Clark et al., 2016; Focas and Christidis, 2017).

Because of these fundamental shifts in life circumstances and individual travel, young adult mobility has for a long time been of special interest to researchers, planners and policy makers. In recent years, this topic has drawn additional attention because analysts identified changes in mobility trends among young adults in many industrialized countries. While there has been much heterogeneity across different places and different statistical indicators, the key findings included declines in license holding, car ownership and kilometers driven by young adults (ifmo, 2013; Kuhnimhof et al., 2012; Noble, 2005). These observations have led to the conjecture of declining car orientation among young adults (Grimal, 2018; Hjorthol, 2016; Le Vine et al., 2014).

These observations coincided with the rise of new mobility options and services such as car sharing, ride matching services or transport network companies whose most prominent representative is Uber (Marsden et al., 2018). The concurrence of the behavioral changes among young adults and the emergence of such new services has fueled expectations that a next generation of travelers would establish new mobility patterns with consequences specifically of reduced private car ownership and use (Delbosc and Ralph, 2017).

Against this background, this chapter's objective is to establish the status quo and recent trends in young adults' mobility. The focus of this contribution is on the situation of young adults in industrialized countries and much of the presented information applies to a broad range of high income countries. However, for illustration purposes, that is mostly for figures, this contribution uses data from Germany and England. (England is the focus of this analysis, rather than the United Kingdom as a whole, in respect of the use of England's National Travel Survey, which has no UK-wide equivalent.) These two countries represent the observations of wealthy European regions relatively well. Both countries lie between more car oriented industrialized countries such as the USA and less car oriented ones such as Japan (Ecola et al., 2014). Quantitative figures and specific facets of development vary across locales, however it is not the intent of this analysis to compare between the two countries in great depth, as some of the differences may be due to methodological differences in the German and English surveys.

It is also evident that the situation in emerging economies today is fundamentally different. The spatial and demographic factors linked with China's motorization, for instance, are markedly different than current trends in Europe and North America (Le Vine et al., 2018). Many aspects of future travel demand trends among young adults in emerging economies such as their private car orientation are subject to speculation, in part due to limited data availability. However, it appears reasonable to expect that young adults in emerging economies will in principle follow their wealthier counterparts in that mobility increases as they reach maturity and individual economic independence. Hence, this chapter also allows certain conclusions as to what can be expected regarding future mobility of adults in emerging economies.

This article is structured as follows: first, we present the socio-economic conditions of young adults and how these typically change as individuals transition through this stage of life; second, the article takes a closer look at the implications that this has on mobility options and travel behavior; third, we separate different dimensions of behavioral change that are important to understand aggregate changes in young adults' travel over time; finally, we discuss changes that have been observed with respect to travel of young adults in the last two decades and conclude the paper with final remarks and an outlook.

Young Adulthood: A Phase of Socio-Economic Transitions

To understand the fundamental changes in individual mobility which take place between the ages of 15 and 35 it is important to map relevant external framework conditions influencing travel behavior. A prototypical sequence of life events and stages, experienced by many but certainly not all individuals, involves completing school, moving out of the parents' household, pursuing higher education, getting a first job, moving in with a partner and starting a family. In each of these stages, a specific set of duties (e.g., school or work), desires (e.g., leisure activities or social needs), and possibilities (e.g., travel mode options or economic capability) shape the actual travel behavior. Hence, personal travel in each of these life stages differ and we observe substantial changes in average per-capita mobility indicators as individuals age and move through these different stages of life (Blumenberg et al., 2012; Delbosc and Currie, 2013b; Grimsrud and El-Geneidy, 2013; Oakil et al., 2016; Taylor et al., 2013; Waygood et al., 2020).

Figs. 1 and 2 exemplify the substantial changes in life circumstances taking place between the ages of 15 and 35 with regard to two influential dimensions of every day life, the family situation and the educational or professional situation. While teenagers under 18 live in households with children (because they are the children), **Fig. 1** shows a sharp drop as teenagers age into adulthood themselves. As a consequence, and with most having yet to have become parents themselves, the vast majority of those in their early twenties live in households without children. After this age, however, the proportion with children increases as these young adults start their own families. **Fig. 2** shows a similar sharp transition regarding the occupational status of young adults. While in the early twenties almost half of Germany's young adults pursue higher education, 80% have transitioned into the job market 10 years later. The situation in England is very similar, however, education and transitioning into the job market take place earlier.

Figs. 1 and 2 also illustrate changes that have occurred with respect to the composition of the population between 2002 and 2017 in Germany and England. Shifts regarding the occupational status of young adults (**Fig. 2**) are strongly influenced by economic conditions and specific national conditions of transitioning from school to higher education and economic activity. However, changes observed in **Fig. 1** specifically for Germany reflect a more global and long term trend: The age of family formation continues to shift towards higher ages (OECD, 2011). In many countries this goes along with a higher proportion of young adults pursuing higher education (which can be seen clearly in the 2002 vs. 2017 England data in **Fig. 2**), resulting in a later entry into the job market and age of settling into a stable career (ifmo, 2013). In short: the life stage of maturation is extending into later ages.

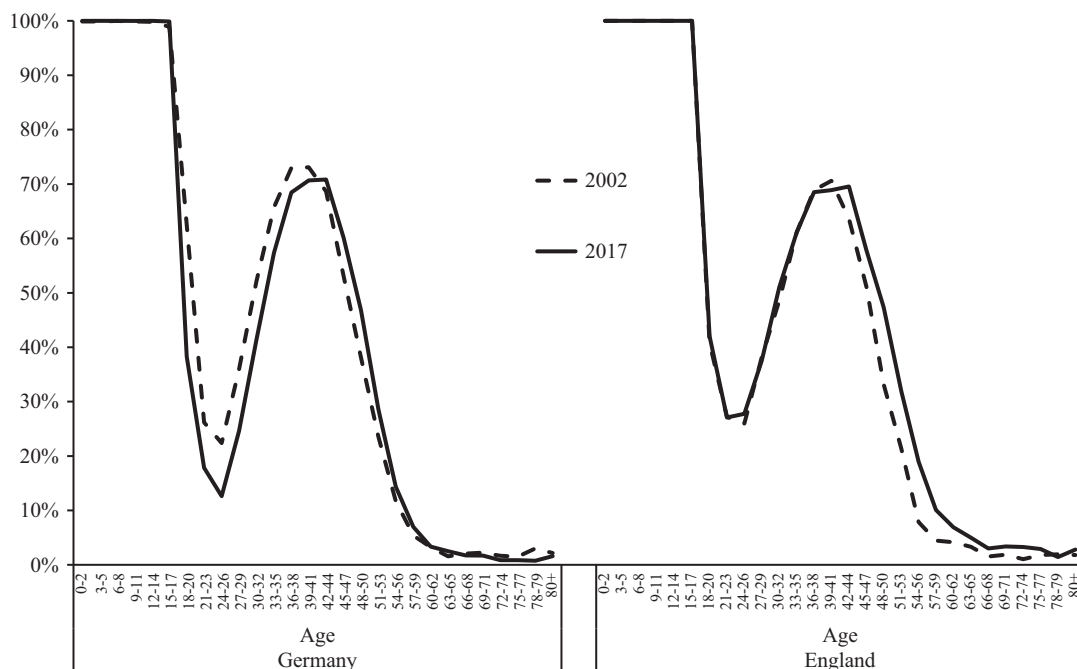


Figure 1 Proportion of population in households with children under 18 by age in 2002 and 2017 in Germany and England (Authors' own representation based on data from Bundesministerium für Verkehr- Bau- und Stadtentwicklung (BMVBS) (2010) and Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) (2019)).

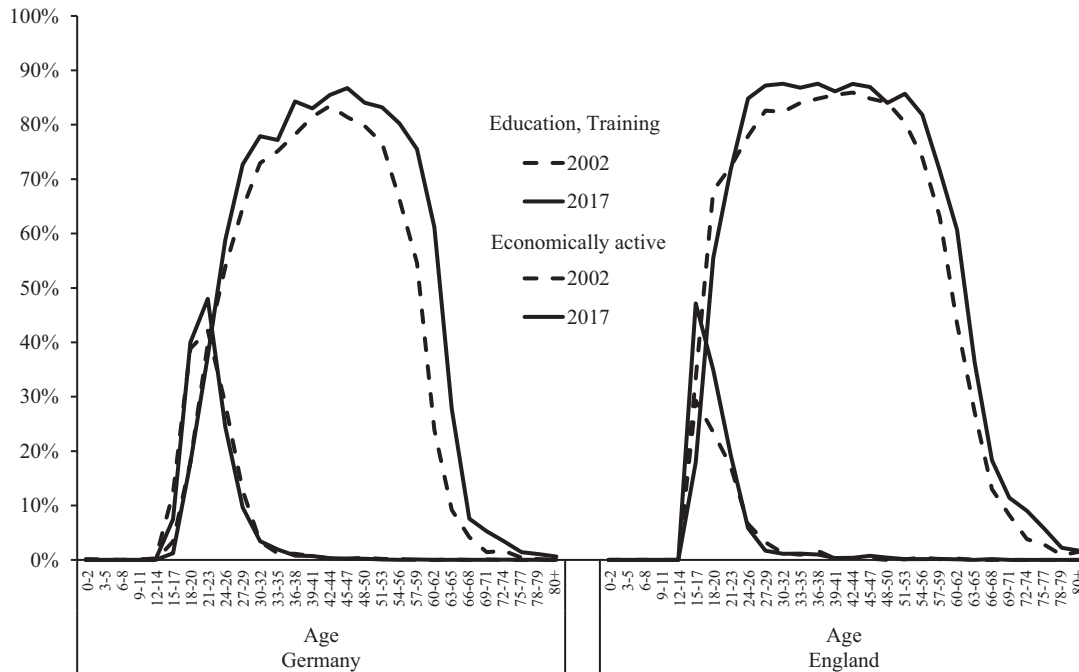


Figure 2 Proportion of population by occupation by age in Germany and England in 2002 and 2017 (Authors' own representation based on data from Bundesministerium für Verkehr- Bau-und Stadtentwicklung (BMVBS) (2010) and Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) (2019)).

Advancing to a New Mobility Life Stage: More Options and More Travel

Between the ages of 15 and 35, the kilometers traveled per person per day increase by about 70% in both Germany and England. There are two main drivers for this development: First, there are new mobility options which specifically go along with driver license acquisition and access to cars. Second, there are stepwise changes from one life stage into the next which are often associated with increased mobility needs, such as commuting, and/or growing economic capabilities to travel via higher cost but more desirable modes (in terms of status or speed) (Oakil et al., 2016).

Figs. 3 and 4 illustrate differences in travel options and travel behavior among different young adult age groups in Germany and England in 2017. Here, the majority of young adults acquire a driving license by the age of 20 and from their mid-20s onward no less

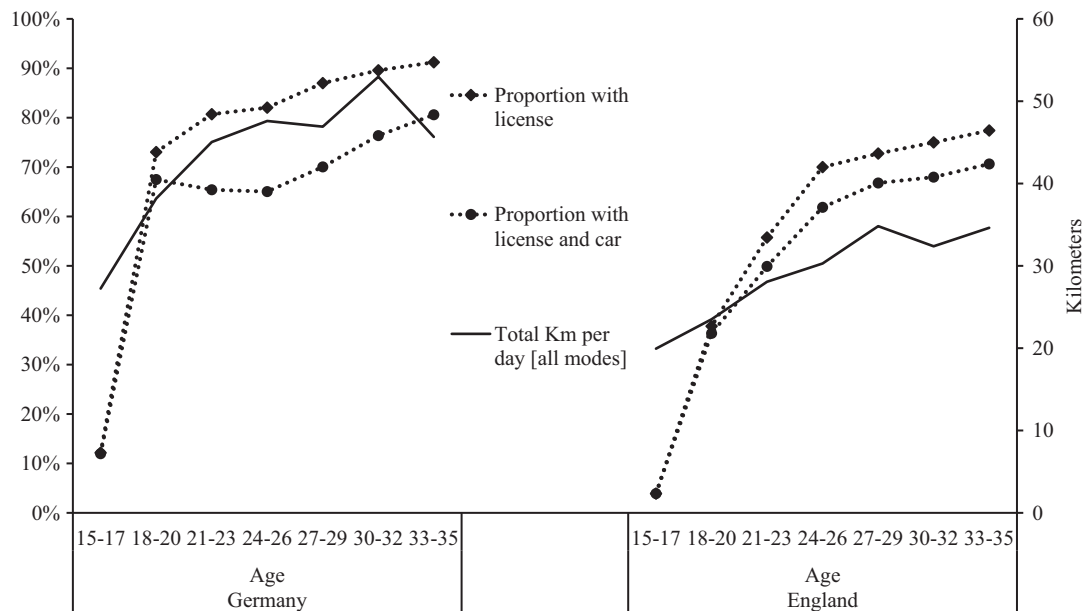


Figure 3 License holding, car availability and kilometers per capita per day by age, Germany and England 2017 (Authors' own representation based on data from Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) (2019)).

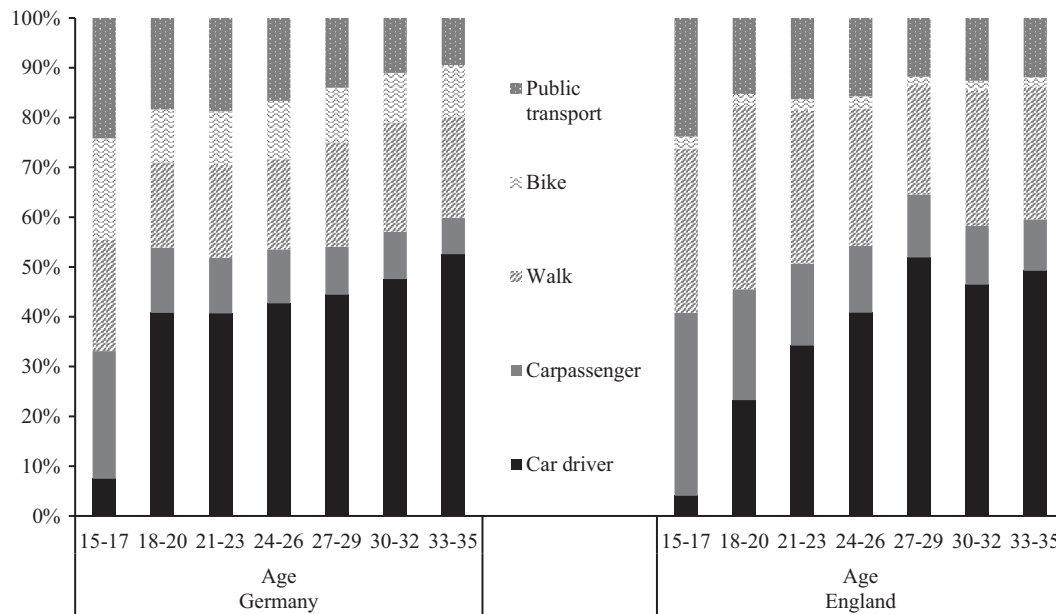


Figure 4 Trip based modal split by age, Germany 2017 (Authors own representation based on data from Bundesministerium für Verkehr und digitale Infrastruktur (BMVI) (2019)).

that two thirds of the population have access to a car within their household in both countries. Beyond the age of the early 20s, license holding and car access keep increasing, but at a much lower rate. The option to drive that a licence provides has an impact on both the total kilometers traveled as well as the modal split (Fig. 4). Hence, much of the growth in mobility during the young adult life stage can be attributed to the few years of age near 20 when acquisition of driving licenses is at its highest rate. There appears to be a somewhat sharper rate of change in Germany at this age with this change spread over a larger number of years among English youth.

From around age 20 onward, there are—on average—still increases of individual travel and modal shifts toward the car, but at a lower rate (particularly in Germany). During these years, significant life stage transitions and socio-economic change continue to take place, as seen above. But by that age the majority of young adults have already settled into a life style of substantial mobility which in many cases is heavily auto-oriented.

Intra- and Intergenerational Changes of Travel

While Figs. 3 and 4 reasonably illustrate the changes in mobility that occur as individuals age, they are actually imprecise representations of these processes. This is because in these snapshots (based on data from only the year 2017) different phenomena overlay, as the following example illustrates: The individuals in the first age bracket (15–17) were born almost two decades after those in the last age bracket (33–35). It is possible, that by the time 2017s teenagers reach 30 years of age, their travel behavior could be quite different from that of 2017s cohort of 30-year olds. They might transition differently through the different stages of young adults' mobility discussed previously. Thus, along with new constraints and opportunities such as from emerging technologies, they may eventually develop distinctive mobility patterns. This phenomenon is known amongst social scientists as a cohort effect.

These considerations lead to the differentiation of two important aspects of change regarding young adult mobility (Bell and Jones, 2013a; Bell and Jones, 2013b; Dargay et al., 2000; Ottmann, 2010):

- *Intra-generational* change of travel behavior describes the change in mobility that goes along with individual aging and transitioning from one life stage to the next. The sum of these individual transitions and the associated changes in life configurations and mobility patterns account for the average changes in travel behavior of a particular cohort or generation of young adults, for example, those born in the 1980s, as this cohort grows older.
- *Inter-generational* change of travel behavior describes the difference in mobility between members of different generations or cohorts, for example, those born in 1980 and those born in 2000, at similar ages or life stages. Inter-generational changes again can be broken down in two components delineating two areas of influential factors:
 - Inter-generational socio-economic shifts (e.g., those depicted in Figs. 1 and 2): Such socio-economic shifts affect how the population in a specific age bracket (e.g., ages 25–30) is composed of different groups, for example, proportion of those with or without children. As these subgroups will exhibit different mobility behavior from each other, socio-economic shifts contribute to different aggregate mobility indicators for different generations of young travelers.

- Intergenerational changes under socio-economic *ceteris paribus* conditions: Members of different year-of-birth cohorts or generations may exhibit different mobility behavior, even if socio-economic life situations resemble. For example, urban students at the age of twenty today may differ from 20-year-old urban students 10 years ago. Reasons for such intergenerational differences may be found in altered preferences or external factors such changes in the transport system.

Thorough analysis of young adults' mobility and its changes over time must take account of these two dimensions of change, as well as the two components of intergenerational change, in order to accurately identify the contributions of hypothesized explanations of changes.

Intergenerational Changes of Mobility Patterns Among Young Adults

Changes in young adults' travel behavior in industrialized countries which began in the late 1990s have drawn much research attention in the recent years. This mostly refers to intergenerational changes. After the turn of the millennium, researchers observed mobility patterns for the millennial generation, that is, those born in the early 1980s, which were unexpected. In short, earlier generations of young adults' exhibited increasingly car oriented mobility patterns relative to yet earlier generations. However, around the turn of the millennium these trends did not continue for the millennial generation. This raised much attention and hopes that the next generation of travelers would organically establish the more sustainable travel patterns that many in the transport and research policy communities seek to deliver (Blumenberg et al., 2012, 2013; Delbosc and Ralph, 2017; ifmo, 2013; Kuhnimhof et al., 2011).

Typical indicators evaluated in this context include license holding, car ownership, mode choice and vehicle kilometers traveled. However, the findings that contributed to this overall picture are extremely heterogeneous across different countries and travel indicators. In some places, license holding among young adults was found to decline, whereas in other places declines in car ownership and car use were found to be more pronounced (ifmo, 2013).

Not only the indicators differ but also the statistical sources for these indicators, which range from administrative databases (driving licence or vehicle registrations) to travel surveys and income/expenditure surveys. Despite this heterogeneity of data resources and findings, the decline of car orientation in young adults' travel in many industrialized countries seems to be sufficiently substantiated. However, the strength of this decline and which indicators are affected is debatable.

As for the reasons of the decline in car orientation, findings are similarly divergent. Many researchers attribute at least part of this trend to socio-economic developments (Chatterjee et al., 2018; Coogan et al., 2017; Delbosc and Currie, 2013a, 2013c). For instance, analyses for Germany and England indicated that between one third (England) and two thirds (Germany) of the decline in young adults' car ownership between 1998 and 2008 can be attributed to changes in income, place of residence (urban/rural), household composition and employment status (ifmo, 2013). There is also ample discussion of possible changes in preferences or psychological factors (Grimal, 2018; Hjorthol, 2016; Kuhnimhof et al., 2019). But the underlying figures are even less standardized across countries than the aforementioned travel indicators, and thus it has not been possible to credibly generalize.

Moreover, much of the research on young adult mobility that continues to frame today's discussion today was carried out in the years 2010–15. However, in some instances trends observed up to 2010 did not continue thereafter. For example, car registrations for young adults in Germany did not decline after 2010. Also in Germany, driver km per person per day has changed little after the 2000s (Kuhnimhof et al., 2019). In England, possible contributory factors that have been raised include rapid increases in motor insurance costs for young drivers, as well as changes to the driving licence acquisition process that have made "learning to drive" (i.e., developing the skills needed to pass the theory and on-road tests) a longer duration and more expensive process.

Concluding Remarks

Changes in young adults travel behavior in industrialized countries have drawn much attention in the past decade. While findings in the literature are not fully in agreement, there is clear indication that driving and the automobile are not as dominant for young adults' mobility as was the case prior to the turn of the millennium. Socio-economic shifts among young adults explain part but not all of this development. However, it is still uncertain which other factors have played a role.

Turning to the crucial question of what can be expected in coming years, it is also still unclear what recent developments will mean regarding the future travel habits of the millennials and successive cohorts as they age, start families and pursue professional careers. Persistence of mobility habits (e.g., modal preferences) acquired during the young adults' life stage is often conjectured. However, as of yet there is no reliable empirical evidence that the less car oriented mobility styles of the millennials carry on through later stages of life. However, it appears plausible that key life decisions made by young adults such as type of job, location of residence and work, starting a family etc. shape life configurations which in turn frame mobility in later stages of life. Hence, transport planners and policy makers should take advantage of the development among young adults, as it opens a window of opportunity to establish new and more sustainable travel patterns among future generations of travelers.

The observed changes in travel behavior for young adults should not be overrated, however. Research on young adults that has been carried out between 2010 and 2015, and is therefore not informed by the latest data, continue to be part of today's discussion,

both in the scholarly discourse as well as mass media, industry and the wider public. Developments after 2010 have yet to receive this same attention and in some instances “urban myths” unsupported by evidence persist in the public discussion. An example is the widespread belief in Germany that licensing rates among young adults are declining while the licensing register confirms that they are not (Böhm, 2015; NDR, 2019). The danger of a mis-informed public debate is the expectation that transport-related challenges will disappear as a new generation of travelers with more sustainable travel habits simply replaces more automobile-oriented predecessor generations.

However, there is no reason to ease policy efforts to guide personal travel in a more sustainable travel direction. The young adult life stage between 15 and 35 years of age continues to be characterized by life's most substantial changes and increases in travel, which nearly doubles during this 20-year period—mostly in the form of increased car use. The singular set of changes at this point in life has seen little change between successive generations of young adults. Hence, young adults' mobility will continue to be of special interest for policy makers and planners, as part of their efforts to ensure sustainable and efficient mobility patterns.

References

- Bell, A., Jones, K., 2013a. Current practice in the modelling of age, period and cohort effects with panel data: a commentary on Tawfik et al. (2012), Clarke et al. (2009), and McCulloch (2012). *Quality & Quantity* 48 (4), 2089–2095, doi:10.1007/s11135-013-9881-x.
- Bell, A., Jones, K., 2013b. The impossibility of separating age, period and cohort effects. *Soc. Sci. Med.* 93, 163–165, doi:10.1016/j.socscimed.2013.04.029.
- Blumenberg, E., Taylor, B.D., Smart, M., Ralph, K., Wander, M., Brumbaugh, S., 2012. What's Youth Got to Do with It? Exploring the Travel Behavior of Teens and Young Adults. UC Los Angeles, Los Angeles.
- Blumenberg, E., Wander, M., Taylor, B.D., Smart, M., 2013. The Times, Are they A-Changing? Youth, Travel Mode, And the Journey to Work. Paper presented at the 92nd Transportation Research Board Annual Meeting, Washington D.C.
- Böhm, R., 2015, 6.1. Führerschein mit 18 ist out. *Der Tagesspiegel*. Retrieved from <https://www.tagesspiegel.de/mobil/smartphone-statt-autofahren-fuehrerschein-mit-18-ist-out/11188710.html>.
- Bundesministerium für Verkehr- Bau-und Stadtentwicklung (BMVBS), 2010. Mobilität in Deutschland 2008.
- Bundesministerium für Verkehr und digitale Infrastruktur (BMVI), 2019. Mobilität in Deutschland 2017.
- Chatterjee, K., Goodwin, P., Schwanen, T., Clark, B., Jain, J., Melia, S., Stokes, G., 2018. Young People's Travel.—What's Changed and Why? Review and Analysis. Report to Department for Transport. UWE Bristol, Bristol, UK.
- Clark, B., Chatterjee, K., Melia, S., 2016. Changes in level of household car ownership: the role of life events and spatial context. *Transportation* 43 (4), 565–599, doi:10.1007/s11116-015-9589-y.
- Coogan, M., Nygaard, N., Weinberger, R., 2017. Understanding Changes in Youth Mobility: AASHTO.
- Dargay, J.M., Madre, J.-L., Berri, A., 2000. Car ownership dynamics seen through the follow-up of cohorts: comparison of France and the United Kingdom. *Transportation Research Record. J. Transport. Res. Board* 1733 (1), 31–38, doi:10.3141/1733-05.
- Delbosc, A., Currie, G., 2013a. Are changed living arrangements influencing youth driver license decline? In B. Transportation Research (Ed.), 92nd Transportation Research Board Annual Meeting.
- Delbosc, A., Currie, G., 2013b. Are changed living arrangements influencing youth driver license decline? Paper presented at the 92nd Transportation Research Board Annual Meeting, Washington D.C.
- Delbosc, A., Currie, G., 2013c. Changing demographics and young adult driver license decline in Melbourne, Australia (1994-2009). *Transportation* 41 (3), 529–542, doi:10.1007/s11116-013-9496-z.
- Delbosc, A., Ralph, K., 2017. A tale of two millennials. *J. Transport Land Use* 10 (1), doi:10.5198/jtlu.2017.1006.
- Ecola, L., Rohr, C., Zmud, J., Kuhnimhof, T., Phleps, P., 2014. The Future of Driving in Developing Countries. RAND Corporation, Santa Monica, CA.
- Focas, C., Christidis, P., 2017. What Drives Car Use in Europe? European Commission, Joint Research Centre, Luxembourg.
- Grimal, R., 2018. Are the millennials less car oriented? Literature review and empirical findings. Paper presented at the European Transport Conference, Dublin, Ireland.
- Grimsrud, M., El-Geneidy, A., 2013. Transit to eternal youth: lifecycle and generational trends in Greater Montreal public transport mode share. *Transportation* 41 (1), 1–19, doi:10.1007/s11116-013-9454-9.
- Hjorthol, R., 2016. Decreasing popularity of the car? Changes in driving licence and access to a car among young adults over a 25-year period in Norway. *J. Transport Geogr.* 51, 140–146, doi:10.1016/j.jtrangeo.2015.12.006.
- Kuhnimhof, T., Armoogum, J., Buehler, R., Dargay, J., Denstadli, J.M., Yamamoto, T., 2012. Men shape a downward trend in car use among young adults—evidence from six industrialized countries. *Transport Rev.* 32 (6), 761–779.
- Kuhnimhof, T., Nobis, C., Hillmann, K., Follmer, R., Eggs, J., 2019. Veränderungen im Mobilitätsverhalten zur Förderung einer nachhaltigen Mobilität. Abschlussbericht. Umweltbundesamt, Dessau-Roßlau.
- Kuhnimhof, T.G., Bühler, R., Dargay, J., 2011. A new generation: travel trends among young Germans and Britons. *Transport. Res. Rec.: J. Transport. Res. Board* 2230, 58–67.
- Le Vine, S., Jones, P., Lee-Gosselin, M., Polak, J., 2014. Is heightened environmental sensitivity responsible for drop in young adults' rates of driver's license acquisition? *Transport. Res. Rec.: J. Transport. Res. Board* 2465 (1), 73–78, doi:10.3141/2465-10.
- Le Vine, S., Wu, C., Polak, J., 2018. A nationwide study of factors associated with household car ownership in China. *IATSS Res.* 42 (3), 128–137, doi:10.1016/j.iatssr.2017.10.001.
- Marsden, G.D., John, Jones, P., Seagriff, E., Spurling, N., 2018. All Change? The future of travel demand and the implications for policy and planning. The First Report of The Commission on Travel Demand. Leeds: The Commission on Travel Demand.
- NDR, 2019. Weniger junge Leute machen den Führerschein. Retrieved from <https://www.ndr.de/nachrichten/niedersachsen/Weniger-junge-Leute-machen-Fuehrerschein,fuehrerschein840.html>.
- Noble, B., 2005. Why are some young people choosing not to drive? Paper presented at the European Transport Conference, Strasbourg.
- Oakil, A.T.M., Manting, D., Nijland, H., 2016. Determinants of car ownership among young households in the Netherlands: the role of urbanisation and demographic and economic characteristics. *J. Transport Geogr.* 51, 229–235, doi:10.1016/j.jtrangeo.2016.01.010.
- OECD, 2011. Doing better for families.
- Ottmann, P., 2010. Abbildung demographischer Prozesse in Verkehrsentscheidungsmodellen mit Hilfe von Längsschnittdaten. (Dissertation). Karlsruhe Institute of Technology, Retrieved from <http://epub.sub.uni-hamburg.de/epub/volltexte/2011/9934> GBV Gemeinsamer Bibliotheksverbund database.
- Taylor, B.D., Ralph, K., Blumenberg, E., Smart, M., 2013. Who Knows About Kids These Days? Analyzing The Determinants Of Youth And Adult Mobility Between 1990 and 2009. In B. Transportation Research (Ed.), 92nd Transportation Research Board Annual Meeting.
- Waygood, E.O.D., Friman, M., Olsson, L.E., Mitra, R., 2020. Chapter One - Introduction to transport and children's wellbeing. In: Waygood, E.O.D., Friman, M., Olsson, L.E., Mitra, R. (Eds.), *Transportation and Children's Well-Being*. Elsevier, pp. 1–17.

- Chatterjee, K., Goodwin, P., Schwanen, T., Clark, B., Jain, J., Melia, S., Stokes, G., 2018. Young People's Travel.—What's Changed and Why? Review and Analysis. Report to Department for Transport. UWE Bristol, Bristol, UK.
- Clark, B., Chatterjee, K., Melia, S., 2016. Changes in level of household car ownership: the role of life events and spatial context. *Transportation* 43 (4), 565–599, doi:10.1007/s11116-015-9589-y.
- Coogan, M., Nygaard, N., Weinberger, R., 2017. Understanding Changes in Youth Mobility: AASHTO.
- Dargay, J.M., Madre, J.-L., Berri, A., 2000. Car ownership dynamics seen through the follow-up of cohorts: comparison of France and the United Kingdom. *Transport. Res. Rec.: J. Transport. Res. Board* 1733 (1), 31–38, doi:10.3141/1733-05.
- Delbosc, A., Currie, G., 2013c. Changing demographics and young adult driver license decline in Melbourne, Australia (1994-2009). *Transportation* 41 (3), 529–542, doi:10.1007/s11116-013-9496-z.
- Grimal, R., 2018. Are the millennials less car oriented? Literature review and empirical findings. Paper presented at the European Transport Conference, Dublin, Ireland.
- Hjorthol, R., 2016. Decreasing popularity of the car? Changes in driving licence and access to a car among young adults over a 25-year period in Norway. *J. Transport Geogr.* 51, 140–146, doi:10.1016/j.jtrangeo.2015.12.006.
- ifmo, 2013. 'Mobility Y'—The Emerging Travel Patterns of Generation Y. ifmo (Institut for Mobility Research), Munich.
- Kuhnimhof, T., Armoogum, J., Buehler, R., Dargay, J., Denstadli, J.M., Yamamoto, T., 2012. Men shape a downward trend in car use among young adults - evidence from six industrialized countries. *Transport Rev.* 32 (6), 761–779.
- Kuhnimhof, T.G., Bühler, R., Dargay, J., 2011. A new generation: travel trends among young Germans and Britons. *Transport. Res. Rec.: J. Transport. Res. Board* 2011 (2230), 58–67.
- Marsden, G.D., John, Jones, P., Seagriff, E., Spurling, N., 2018. All Change? The future of travel demand and the implications for policy and planning. The First Report of The Commission on Travel Demand. Leeds: The Commission on Travel Demand.
- Waygood, E.O.D., Friman, M., Olsson, L.E., Mitra, R., 2020. Chapter One - Introduction to transport and children's wellbeing. In: Waygood, E.O.D., Friman, M., Olsson, L.E., Mitra, R. (Eds.), *Transportation and Children's Well-Being*. Elsevier, pp. 1–17.

Airport Network Planning and Its Integration with the HSR System

Francesca Pagliara*, **Juan Carlos Martín[†]**, **Concepción Román[†]**, ^{*}Department of Civil, Architectural and Environmental Engineering, University of Naples Federico II, Naples, Italy; [†]Faculty of Economics, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	222
Literature Review	223
Theoretical Models	225
Empirical Applications	226
Summary and Conclusions	226
References	227
Further Reading	228

Introduction

The aviation sector is one of the crucial components that facilitates and boosts the world economy, moving passengers and cargo from different origins in countries all over the world to the final destinations. In doing so, it needs the complementarity from other transport modes as it is not possible to imagine a single pure air door-to-door trip, so air transport is in essence multimodal. For this reason, it produces a lot of spill-over effects in other transport modes as well as in other economic activities. [De Neufville and Odoni \(2013\)](#) contend that the aviation sector in the early 21st century is characterized by three important trends: (1) Long-term growth. An increase of 4% per year worldwide has implied that the traffic has been doubled about every 15–20 years. Thus, airports' expansions and new developments have been the norm in the sector. (2) Organizational change. The economic deregulation of the sector as well as the new open-skies agreements have been spread worldwide following the good results observed in the United States and the European Union. The privatization of airlines and airports has created new opportunities in the sector. (3) Technical change. This has been mainly observed in aircraft and air traffic control, but ICT is also expected to disrupt all industries and it is obvious that the aviation sector will not be an exception. The new technology will reshape the way some airport processes are performed, especially for safety, security screenings, and border controls. The new developments will increase airports' efficiency and airports need to be ready to this new era of rapid changes.

The increased demand for flights produces in many circumstances some delays, especially in those congested airports that suffer from lack of adequate capacity and operate near to the maximum capacity during some periods of the day, some weekdays, or certain peak moments. The Airports Council International (ACI) (2018) publishes the Annual World Airport Traffic Report (WATR), and finds that all regions experience a growth with a passenger and air cargo increase of 7.5% and 7.7% in the years 2017 and 2016, respectively. The aircraft traffic movements increased only by 3%. All in all, the airports worldwide accommodate 8277.7 million passengers, 118.6 million metric tons of cargo, and 95.8 million aircraft movements. The center of gravity of the sector shifts eastward as most of the world's fastest growing large airports are located in emerging markets. Sixteen of the fastest growing top 30 airports with over 15 million passengers are located in just two countries, China and India, and 13 airports out of the top 30 busiest passenger hubs are nowadays located in the Asia-Pacific region.

The International Air Transport Association (IATA) (2018) publishes the Annual Review and concludes that in 2017 airlines provided a record number of regular services to more than 20,000 city pairs. This figure doubles the number of existing connections reported in the year 1995. In the period 1995–2017, the flight tickets prices decreased by more than half in real (inflation-adjusted) terms. Some capacity shortages have been detected as airport construction was not able to keep pace with the increase in demand for aircraft movements. In 2017, more than 190 airports worldwide were slot constrained because they had insufficient capacity to meet demand at all hours of the day. The growth in air transport demand is directly transferred into major airport project developments, and the norm is that a number of massive investment projects for airport expansions or new construction are evolving conjointly at any time. Besides expansions and new constructions of airports, the capacity and congestion problems can also be tackled by innovation. According to [De Neufville and Odoni \(2013\)](#), the US aviation sector has led the major innovations affecting the airport planning and design worldwide. Some of the major innovations found by the authors are the following: economic deregulation; Southwest LCC model; US open skies policies; airline alliances; integrated air cargo services; transfer hubs; midfield concourses; automated people movers; and global positioning systems.

Another interesting trend observed in the last 30 years is that governments and public companies are participating less in the aviation sector. Thus, the market forces are nowadays more important than the highly regulated political context of the past, in which the outcomes were better explained by political and national interests. The aviation sector is converging very rapidly to a worldwide global industry. Privatization or corporatization in the airport industry can increase social welfare by the improvement of cost

efficiency, customer service, long-term investments, and innovation. The success of airport privatization or corporatization should be measured in terms of social welfare and not in isolated criteria like for example rates of return to governments or investors.

Similarly to the consolidation of the three global alliances, star, oneworld, and skyteam airports around the world have also explored whether the establishment of airport alliances, like for example, the Galaxy International Cargo Alliance (GICA), Aviation Handling Services (AHS), Pantares, and Aéroports de Paris (ADP)-Schiphol alliance will benefit the competitiveness of each individual airport. Technological spillovers and know-how transfer economic gains have been cited as the main motivations of this trend (Forsyth et al., 2010, 2011; Jiang et al., 2019). The results so far are not promising as most of the airport alliances have failed.

The new context is more oriented than ever to new commercial and economic challenges. Airport planners and designers are nowadays facing a variety of demands beyond the former petitions which were more oriented to standard norms. The business interests do not only include those from airlines but from different types of concessionaires. It is difficult to provide an exhaustive list of the main stakeholders involved in the airport network planning in a country but to cite some of them, it is possible to start with transport governmental policy makers, especially those working for the respective Aviation Administration Bodies, Airport Authorities or Aerospace National Agencies; consultants and practitioners whose professional careers are mainly developed in the field as, for example, CEOs of Airlines, Air Navigation Service Providers, Engineering and Architecture Firms, Oil Companies, Aircraft Manufacturers, Security-Planning-Environmental-Construction Companies, Parking-Airport Lounges-Retail, and Food and Beverages Development Solutions and Information and Communication Companies; and other CEOs of complementary and ancillary services, like for example, Air Traffic Control, Land Transport Companies, Railways, and Hotels. All these main stakeholders have different objectives, and this means that today airport planners need to think in terms of the social benefits taking into account multiple individual interests. De Neufville and Odoni (2013) summarize this by saying that the objectives of airport planners should consequently focus more on performance than on monuments with more low-cost and efficient terminals. Value for money, good service, and functionality are becoming the protagonists.

Airport planning guidance has been largely focused on individual airports, and less attention has been given to the interaction between airports at regional or national level, and to the interaction between the air transport and other transport modes at the time of developing intermodality. Nevertheless, Caves and Gosling (1999) highlight the need to analyze the context of the airport system when individual airport planning is considered. This need is supported by two main causes: (1) the existence of national or regional funding programs that aid airports' development; and (2) the emergence of multi airport city systems that exist in some important world metropolis. The interaction analysis is necessary because the catchment area of the airport developments in one jurisdiction might be extended to hundreds, if not thousands, of kilometers away. Thus, Caves and Gosling (1999) believe that the development of an integral planning of a system of airports, considering at least the main interactions between the airports and other transport modes, will result in a more efficient use of resources for a given provision of air transport services.

Morrel (2010) analyzes the European Commission's policy of encouraging airports to be intermodal hubs, highlighting that the successful cases will be based on something that airports and airlines have known for many years. The modal integration between air and other surface transport modes expands the airport's catchment area and enables airlines to compete in more origin-destination markets without providing additional air legs. High-speed rail (HSR) has been proven to be an excellent candidate for such development, especially as in the cases of Schiphol, Paris Charles de Gaulle, and Frankfurt International, where the HSR station is well connected with the airport and is fully integrated within the railways system. Thus, the passengers can travel from/to more destinations on the first or final leg of the trip without the requirement of making an additional connection. This is not achieved when there is a HSR direct connection between the airport and an important central station like, for example, the London Heathrow Express. The Commission's policy vision is aimed to move EU traffic from air to HSR, with alleged social benefits generated by the time savings, the better comfort provided by the HSR, the reduced congestion in the air approximation movements and the reduction in environmental costs.

The remainder of the chapter is organized as follows: "Literature Review" section reviews the previous studies highlighting mainly the integration between air and HSR. "Theoretical Models" section describes the analytical or theoretical models emphasizing mainly the modal integration between air and HSR. "Empirical Applications" section presents the main empirical applications on the topic. Finally, "Summary and Conclusions" section discusses the future research agenda and presents the main conclusions.

Literature Review

Jiang et al. (2017) highlight that the many of the airlines' hub-and-spoke networks could be used to complement air transport in the hub airports by HSR and other transport modes. Thus, one of the legs in which passengers are nowadays using the air transport will be substituted by HSR or other transport mode if a win-win situation is developed for airlines, HSR, terminals, and especially passengers. This type of integration has been mainly developed for HSR, conventional trains and interurban buses. It seems obvious that if HSR is a very competitive transport mode for trips of around 600 km, then these air legs can be substituted by HSR with an integrated service that can be viewed as a special type of "code sharing"—that is, two airlines cooperate to offer a hub-and-spoke operation with each offering one leg of a trip. Givoni and Banister (2006) analyze the possible relationships between air and HSR transport, finding three interesting states: competition, cooperation, and integration (see Table 1). The introduction of the HSR has given the chance to the railway to become a substitute to the aircraft on routes of about 600 km or a 1 h flight, and it has offered comparable or shorter travel times (from city center to city center). Therefore, transport modes have become substitutes for each

Table 1 Air and HSR relationship

	Market (city)	Relationships between air and rail		Rail link to airport?
		Modes	Operators	
Competition	City center (a) – City center (b)	Substitution	Substitution	No
Cooperation	Airport (a) – City center (a)	Complementary	Complementary	Yes
Integration	Airport (a) – city center (b)	Substitution/Complementary	Complementary	Yes

Source: Givoni and Banister (2006).

other since passengers can either use the aircraft or the railway. Considering that each transport mode is operated by a different operator, competition between them cannot be avoided.

Airports are located at the edge of cities and rail is usually used to access the airport. Therefore the two transport modes can be considered complementary since they both are chosen by users to travel from the origin to the destination and the service providers complement each other. This type of relationship defines cooperation since the two transport modes cooperate given that each mode is contributing and feeding traffic into the other network.

When railway infrastructure is provided at the airport, including the means for a fast and continuous transfer between the aircraft and the train, airlines can use railway services as spokes on their own services. While the airlines provide the service, the train companies can operate it and the two operators complement each other. The railway could serve destinations which were not served by the airline until now or could serve destinations already served. In the first case, the two transport modes work in a complementary way, that is, mode substitution does not exist, while in the second case transport mode substitution takes place and they are used in a complementary way (a flight and a railway journey are used in combination) and are substitutes (on one section either the aircraft or the train can be chosen).

As said, the integrated service should provide a win-win situation for the main stakeholders involved in such development. Jiang et al. (2017) find as the main driving forces of the cooperation the following: (1) Airlines win. Some short-haul flights are unprofitable and only provide value to the airline to maintain the network. Thus, some of the connecting trips can be substituted by HSR (Givoni and Banister, 2006); (2). Hub airports win. Some hub airports, especially in the EU, are very congested and intermodal cooperation can permit that some short-haul slots could be transferred to long-haul routes because passengers in the short-haul route are transported by HSR under this new integrated scenario (Givoni and Banister, 2006; Janic, 2011; Jiang and Zhang, 2014). (3) Society wins. Policy makers usually favor the intermodality because it is usually believed that network integration is more efficient than isolated transport networks as the integration permits to better exploit the network externalities. In addition, in short-haul routes, the substitution of air passengers with HSR passengers is environmentally beneficial (D'Alfonso et al., 2015, 2016). (4) Some foreign airlines win. Some foreign airlines do not have the freedom to enter in some domestic markets, but with the help of the integrated service, the airlines can make a detour and increase the market presence competing with other domestic airlines (Chiambaretto and Decker, 2012; Chiambaretto, 2015). As seen, there are many stakeholders involved, but, so far, airlines have been the more active agents of the integrated services, but it is true that policy makers, HSR operators, the hub airport and the passengers are also actively involved in such development.

Analytical models that analyze Air-HSR modal integration and change the focus from competition to cooperation started in 2010s (Socorro and Vicens, 2013; Jiang and Zhang, 2014; Avenali et al., 2018; Xia and Zhang, 2016, 2017; Jiang et al., 2017). In most of the cases, Air-HSR integration is studied throughout a very basic agreement and network features, in which the costs associated to the integrated terminal seem to come out of the blue. There are some exceptions, Avenali et al. (2018) consider the effects of the sunk costs of the integrated terminal, and Jiang et al. (2017) extend the previous papers considering two important features that were not sufficiently explored: (1) the basics of the network is extended because the HSR operator might cooperate with more than one airline; and (2) the type of modal integration is more heterogeneous, ranging from basic service that only considers a sort of protection agreement that covers major delays to more advanced integrated services with features like seamless tickets, dedicated carriage, coordinated scheduling, and baggage push. Nevertheless, dedicated carriage trains for the luggage are nor common, and there is an opportunity for other tertiary service providers that could take care of the O-D luggage traffic.

The Air-HSR modal integration phenomenon has emerged very intensively because of the existing interests of policy makers who have pushed toward the formation of some Air-HSR partnerships, especially in the EU. The empirical applications that deal with Air-HSR also started in 2010s but are less present in the international literature w.r.t. the theoretical ones (Román and Martín, 2014; Brida et al., 2017; Li and Sheng, 2016; Song et al., 2018). Moreover, it is odd to observe that the few empirical existing studies have not been used in the theoretical models to analyze the diversity of Air-HSR scenarios. The empirical applications have better grounds in the real world than the corresponding theoretical models. Thus, it is fair to suggest using the insights found in the empirical applications and to infer some proxies for the intervening parameters of the theoretical models.

Theoretical Models

Most of the international literature looks at transport mode alternatives in competition with each other proposing theoretical approaches to model this phenomenon. Not many are the contributions on the potential for their integration. [Givoni and Banister \(2006\)](#) argued that in the case of aircraft and HS train services the potential benefits from integration between modes and operators are greater than the benefits from competition. Indeed, the increase of the airport's capacity, through a railway station instead of a new runway, produces social-economic benefits at a lower environmental cost. Moreover, the presence of a railway station in an airport represents a step toward an integrated transportation system. Airlines use railway services as spokes in their network of services from a hub airport to complement and substitute for conventional air services. Railways play an important role in the future of air transport and the [Givoni and Banister's \(2006\)](#) recommendation was to extend the definition of air transport infrastructure in order to integrate the railways. This is fundamental in order to make policy makers thinking about the two transport modes as part of one transport network promoting airline and railway integration. Furthermore, it is necessary to guarantee adequate planning of railway services to and from airports. In order to provide direct, fast, and high-frequency services to many destinations from the airport, the airport itself should be a stop on a conventional railway main line and preferably also on a HSR line, with the result of having a terminal where all services can stop. This can be applied to large airports as well as to smaller ones, with some small modifications.

[Grimme \(2007\)](#) suggested that the decisive prerequisites of successful intermodal products were the availability of the comparable HSR journey time and the integration of airports into the HSR network. Moreover, he reported that very limited slots were freed for alternative use. [Chiambaretto and Decker \(2012\)](#) discussed some competitive issues associated with the AH intermodal agreements. The latter provide a number of potential advantages for airlines, rail operators, intermodal airports, and consumers of transportation services. These agreements involve a form of cooperation between airlines and operators that could, in principle, raise competition concerns. This is the case where agreements involve air and rail services that operate in parallel on a given route, and where the two services are potential substitute forms of transportation.

These contributions have demonstrated the advantages and disadvantages of air-railways integration services. However, they usually adopted a qualitative and descriptive approach and thus failed to quantitatively analyze the effects of the air-railways intermodality on the air transport markets.

[Socorro and Vicens \(2013\)](#) proposed a theoretical model to analyze the social and environmental impacts of airline and HSR integration in two different scenarios, that is, airports with capacity constraints and airports with low airline competition. They considered an economy composed of an oligopolistic sector and a competitive sector summarizing the rest of the economy. The theoretical model developed allows explaining formally why companies can benefit from air-rail integration. When mode substitution takes place due to integration, the companies that integrate can enjoy a higher market power. Moreover, if marginal costs are lower for the HSR, integration implies efficiency gains. It is also shown that integration is always profitable for the airline and for the HSR company when they have access to a new market due to integration. Moreover they showed that airline and HSR integration is more likely to be welfare enhancing if the airport is capacity-constrained. These results would justify, from a social point of view, airline and railway integration in airports that are highly congested.

Integration releases airport slots but the environmental question to promote integration is less straightforward given that more trips take place. In contrast, with no capacity constraints, integration does not necessarily imply a higher covered demand, and, in this case, mode substitution sets some trade-offs, that is, mode substitution is environmentally positive, but on the other hand, it eliminates competition in the market. They found also that in monopoly markets, the main impact of integration is to foster competition. When an airline and the HS train integrates, their combination results in an alternative way of traveling for passengers, thus removing the monopoly status of the airline that was already operating the market. Therefore, integration may introduce competition in markets where competition between airlines is low and particularly in those routes where the HS train is seen as a good substitute for the airline. In this case, integration would mean more competition and therefore lower prices, more variety for consumers, and higher social welfare.

[Jiang and Zhang \(2014\)](#) proposed a different analytical framework to address the welfare effects of air and HSR cooperation under hub airport constraint. They analyzed the economies of traffic density, and vertical differentiation between modes and heterogeneous passenger types. In their model, they assumed that the passenger demand function was specified as linear function. They found that such cooperation reduces traffic in markets where prior modal competition occurs but may increase traffic in other markets of the network. The cooperation improves welfare, independent of whether the hub capacity is constrained, as long as the modal substitutability in the overlapping markets is low. However, if the modal substitutability is high, then hub capacity plays an important role in assessing the welfare impact: if the hub airports are significantly capacity-constrained, the cooperation improves welfare; otherwise, it is likely welfare reducing.

Forecasting passenger demand for the intermodal service is important for stakeholders to evaluate the economic viability of the AH intermodal service. In practice, the passenger demand for such intermodal service is affected by various factors, such as connection time, the luggage through-handling service level, the integrated ticket price, and the competition from other travel modes. [Chiambaretto et al. \(2013\)](#) proposed a conjoint analysis to measure the willingness-to-pay (WTP) of passengers for using the AH intermodal service under different service levels. They showed the differences in the reservation price of passengers according to passengers' socioeconomic characteristics. [Zhi-Chun and Sheng \(2016\)](#) investigated the mode choice behavior of intercity passengers among air transport, HSR, and AH integration services. Stated preference surveys were employed and mode share models were estimated based on the collected data. The models developed were used to identify the key factors affecting users' mode choices and

to estimate the modal share of passenger travel demand for some intercity transportation markets. Sensitivity analysis was carried out as well with the objective of revealing the market potential of the AH integration service. It was then demonstrated that when the intercity travel distance exceeded a threshold, passengers became less sensitive to the connection time of the AH service. The most competitive haul distance for the AH service was between 1200 km and 1600 km and the en route travel time was the most important factor affecting the market share of the AH service.

Empirical Applications

Since the introduction of the first HSR line connecting the cities of Paris and Lyon, many empirical applications have analyzed the competition between HSR and air transport for distances no longer than 1000 km (see e.g., [Chang and Chang, 2004](#); [Román et al., 2007](#); [Jiménez and Betancor, 2012](#); [Dobruszkes et al., 2014](#); [Jung and Yoo, 2014](#); [D'Alfonso et al., 2015](#); [Zhao et al., 2018](#)). In these contributions, the focus was mainly on the identification of the different factors affecting the competition between these two transport modes which are considered substitutes rather than complementaries.

On the contrary, the empirical literature regarding the integration or cooperation between HSR and air transport is less rich, considering that most of these studies focused on passengers' perception of the main attributes that define the level of service of a given transport mode choice context, namely, transfer time, in-vehicle time, luggage transfer services, effort or transfer convenience, degree of fare integration considering different levels of risk split, and access/egress time. As for the area of application, most of the case studies have been conducted in Spain, the leading European country in developing the HSR network, and more recently in China, which has the world's largest HSR network.

The first paper dealing with the analysis of the mode choice behavior in the context of the integration of air and HSR is that of [Román and Martín \(2014\)](#). This work is based on a discrete choice experiment where passengers traveling between Gran Canaria (Canary Islands) and some peripheral cities in mainland Spain are confronted with the choice between the current air-air alternative and an integrated and hypothetical air-HSR option, where several scenarios of integration are considered. Results show that the different components of travel time, specially that used in making the connection between the two modes, as well as the fare integration are highly valued by passengers. In contrast, luggage integration resulted only significant for those who check in luggage and travel for leisure purposes. In a later work using the same data set, [Brida et al. \(2017\)](#) investigate how different market segments, obtained from a previous cluster analysis, have an influence in being more proactive to change to HSR for the second leg in the multimodal trips. Specifically, the results *"confirm the existence of groups of passengers differing in terms of their preferences for integrated multimodal services. A group of individuals appear to gather extra utility for the Air-HSR alternative. The likely factors characterizing this segment are travelling to the central business district as the final destination; higher attitude to work or have a meeting during the trip; interest to use mobile phone and WIFI during the trip; and younger age"* (p. 931).

[Li and Sheng \(2016\)](#) conducted a stated preference survey to passengers travelling in four city pairs located along the Beijing-Guangzhou corridor in China considering air, HSR, and the integrated Air-HSR as the different transport mode alternatives. Different choice models were estimated in order to identify the main drivers affecting mode choice and to determine the potential market share for integrated Air-HSR services in China. They found that haul distances between 1200 and 1600 km make the integrated Air-HSR alternative more competitive, and the in-vehicle travel time is the most important attribute affecting the market share for this option.

In a more recent work, [Song et al. \(2018\)](#) used last generation choice models to analyze how the variety-seeking attitude affects travelers' preferences in the context of Air-HSR intermodality in China. These models are based on the methodological framework proposed by [Ben-Akiva et al. \(2002\)](#) that integrates a latent variable model into the choice model, and are usually referred as integrated choice latent variable models (ICLV) or hybrid choice models. The data used in the analysis were obtained from a survey conducted at Shanghai Pudong International Airport, where travelers responded to different choice tasks and provided information about some attitudinal statements regarding the variety seeking. Thus, in the proposed ICLV, the utility of the different alternatives is explained not only by their characteristics but also by the latent construct of variety seeking which in turn explains the attitudinal statements. The main research findings *"suggest that variety-seeking has different impacts across modes, where variety seekers would be more likely to choose the newly-introduced integrated HSR-air option whereas variety avoiders have a higher propensity to choose car-air or traditional separate HSR-air alternative"* (p. 99).

Summary and Conclusions

[De Neufville and Odoni \(2013\)](#) contend that policy makers need to think strategically and to be flexible in developing tools for airport planning, management, and design. The economic and financial outputs associated with airport operations are nowadays more important than ever due to the observed trend of more commercial and private airports. Technical engineering capability is no longer enough to deliver an optimal social welfare for airports' planning, and a system approach that considers other transport modes and the mobility dynamics of the region and country is compulsory. The standards created in the past for some governments and international agencies are nowadays outdated, as they were mainly based on the planning of an isolated infrastructure without measuring adequately the important interactions that exist with other airports and transport modes. An airport is not independent; it is part of a worldwide transport networks system that is more or less interconnected. Airport planners need to study these

interactions, locally in its hinterland, domestically within the same country, and, internationally, analyzing each of the world regions. Conventional master plans have evolved into dynamic and flexible strategic plans.

De Neufville and Odoni (2013) also observe that many aspects of airport planning, design, and management are diverse across the world, regarding the level of innovation, the legal procedures or simply deep-grounded and traditional ways of doing. The cultural context usually imposes constraints on the main stakeholders involved in airport planning. The authors make the following list: (1) the role of central political power compared to that of the regions; (2) the permissible and desirable level of participation of private business; (3) the relative importance of technical experts and managers; (4) the criteria for excellent performance; and (5) the rights and capabilities of workers. They conclude saying that there is no single right answer—the concept of excellence depends on the context, and the “best practice” of one region may not be transferable to another. The authors find that, for example, in North America, the access to airports is mainly developed putting emphasis in the private cars and automobiles, so the provision of parking spaces is usually generous. Meanwhile, in the rest of the world, the emphasis is given to public ground transport, especially railways and HSTs.

Nowadays, an integrated approach between transport systems that facilitates intermodality such as rail and air services represents an important topic to discuss in the political agenda. It is considered a valid solution to the many transport problems that modern societies face (e.g., rising levels of accidents, emissions, and noise from transport) and plays an important role by enabling better mobility for travelers.

The sustained growth in demand for air travel has led airlines to rethink how they can maximize the effectiveness of their networks. One possibility is to improve the links with other transport modes. Intermodality at airports can involve a combination of: (1) accessibility to airports: local services between the airport and the neighboring city (e.g., via train); (2) complementary feeder services between the airport and the different parts of the surrounding region (mainly provided by buses, conventional trains, and more recently HSTs); (3) competing services between major city centers of neighboring regions; and alternative services that fully replace airline feeder services to airports (in general for those services of less than three hours’ train ride).

Partial or full substitution for air can be considered successful on short- or medium-haul journeys of up to 3 h provided by a HSTs (e.g., between Brussels and Paris). In this case, the train link can also be used to complement air travel, which can be used for the return journey or even at the beginning or end of an intercontinental flight, thus requiring that rail and air schedules, fares, and other transport facilities are carefully coordinated (www.atag.org).

References

- Avenali, A., Bracaglia, V., D’Alfonso, T., Reverberi, P., 2018. Strategic formation and welfare effects of airline-high speed rail agreements. *Transp. Res. Part B: Methodol.* 117, 393–411.
- Ben-Akiva, M., Walker, J., Bernardino, A.T., Gopinath, D.A., Morikawa, T., Polydoropoulou, A., 2002. Integration of choice and latent variable models. In: Mahmassani, H.S. (Ed.), *Perpetual Motion: Travel Behavior Research Opportunities and Challenges*. Elsevier Science, Amsterdam, pp. 431–470.
- Brida, J.G., Martín, J.C., Román, C., Scuderi, R., 2017. Air and HST multimodal products. A segmentation analysis for policy makers. *Netw. Spatial Econ.* 17 (3), 911–934.
- Caves, R.E., Gosling, G.D., 1999. *Strategic Airport Planning*. Pergamon, Oxford.
- Chang, I., Chang, G.L., 2004. A network-based model for estimating the market share of a new high-speed rail system. *Transp. Plan. Technol.* 27 (2), 67–90.
- Chiambaretto, P., 2015. Resource dependence and power-balancing operations in alliances: the role of market redefinition strategies. *Management* 18 (3), 205–233.
- Chiambaretto, P., Decker, C., 2012. Air–rail intermodal agreements: balancing the competition and environmental effects. *J. Air Transp. Manage.* 23, 36–40.
- Chiambaretto, P., Baudelaire, C., Lavril, T., 2013. Measuring the willingness-to-pay of air-rail intermodal passengers. *J. Air Transp. Manage.* 26, 50–54.
- D’Alfonso, T., Jiang, C., Bracaglia, V., 2015. Would competition between air transport and high-speed rail benefit environment and social welfare? *Transp. Res. Part B: Methodol.* 74, 118–137.
- D’Alfonso, T., Jiang, C., Bracaglia, V., 2016. Air transport and high-speed rail competition: environmental implications and mitigation strategies. *Transp. Res. Part A* 92, 261–276.
- De Neufville, R., Odoni, A., 2013. *Airport Systems: Planning, Design, and Management*, 2nd ed. McGraw-Hill, New York (NY).
- Dobruszkes, F., Dehon, C., Givoni, M., 2014. Does European high-speed rail affect the current level of air services? An EU-wide analysis. *Transp. Res. Part A: Policy Pract.* 69, 461–475.
- Forsyth, P., Niemeier, H.M., Wolf, H., 2011. Airport alliances and multi airport companies—implications for competition policy. *J. Airport Manage.* (3), 34–39.
- Forsyth, P., Niemeier, H.M., Wolf, H., 2010. Airport alliances and multi-airport companies: implications for airport competition. In: Forsyth, P., Gillen, D., Muller, J., Niemeier, H.M. (Eds.), *Airport Competition. The European Experience*. Ashgate Publishing Limited, Farnham, pp. 339–352.
- Givoni, M., Banister, D., 2006. Airline and railway integration. *Transp. Policy* 13, 386–397.
- Grimme, W.G., 2007. Air/rail passenger intermodality concepts in Germany. *World Rev. Intermodal Transp. Res.* 1 (3), 251–263.
- Janic, M., 2011. Assessing some social and environmental effects of transforming an airport into a real multimodal transport node. *Transp. Res. Part D: Transp. Environ.* 16 (2), 137–149.
- Jiang, C., Alfonso, T.D., Wan, Y., 2017. Air-rail cooperation: partnership level, market structure and welfare implications. *Transp. Res. Part B* 104, 461–482.
- Jiang, C., Zhang, A., 2014. Effects of high-speed rail and airline cooperation under hub airport capacity constraint. *Transp. Res. Part B* 60, 33–49.
- Jiang, Y., Liao, F., Xu, Q., Yang, Z., 2019. Identification of technology spillover among airport alliance from the perspective of efficiency evaluation: the case of China. *Transp. Policy* 80, 49–58.
- Jiménez, J.L., Betancor, O., 2012. When trains go faster than planes: the strategic reaction of airlines in Spain. *Transp. Policy* 23, 34–41.
- Jung, S.Y., Yoo, K.E., 2014. Passenger airline choice behavior for domestic short-haul travel in South Korea. *J. Air Transp. Manage.* 38, 43–47.
- Li, Z.C., Sheng, D., 2016. Forecasting passenger travel demand for air and high-speed rail integration service: a case study of Beijing-Guangzhou corridor, China. *Transp. Res. Part A: Policy Pract.* 94, 397–410.
- Morrel, P., 2010. Airport competition and network access: a European perspective. In: Forsyth, P., Gillen, D., Muller, J., Niemeier, H.M. (Eds.), *Airport Competition. The European Experience*. Ashgate Publishing Limited, Farnham, pp. 31–46.
- Román, C., Martín, J.C., 2014. Integration of HSR and air transport: understanding passengers’ preferences. *Transp. Res. Part E: Logist. Transp. Rev.* 71, 129–141.
- Román, C., Espino, R., Martín, J.C., 2007. Competition of high-speed train with air transport: the case of Madrid-Barcelona. *J. Air Transp. Manage.* 13 (5), 277–284.
- Socorro, M.P., Vicens, M.F., 2013. The effects of airline and high speed train integration. *Transp. Res. Part A: Policy Pract.* 49, 160–177.
- Song, F., Hess, S., Dekker, T., 2018. Accounting for the impact of variety-seeking: theory and application to HSR-air intermodality in China. *J. Air Transp. Manage.* 69, 99–111.
- Xia, W., Zhang, A., 2016. High-speed rail and air transport competition and cooperation: a vertical differentiation approach. *Transp. Res. Part B: Methodol.* 94, 456–481.
- Xia, W., Zhang, A., 2017. Air and high-speed rail transport integration on profits and welfare: effects of air-rail connecting time. *J. Air Transp. Manage.* 65, 181–190.

Zhao, S., Wu, N., Wang, X., 2018. Impact of feeder accessibility on high-speed rail share: Wuhan–Guangzhou corridor, China. *J. Urban Plan. Dev.* 144 (4), 04018029.

Zhi-Chun, L., Sheng, D., 2016. Forecasting passenger travel demand for air and high speed rail integration service: a case study of Beijing–Guangzhou corridor, China. *Transp. Res. Part A: Policy Pract.* 94, 397–410.

Further Reading

Airports Council International ACI, 2018. *Annual World Airport Traffic Report*, 2018 edition.

International Air Transport Association IATA, 2018. *Annual Review*, 2018 edition.

Airport Management

Peter Forsyth^{*,†}, Hans-Martin Niemeier[‡], ^{*}Department of Economics, Monash University, Clayton, Australia; [†]School of Business and Tourism, Southern Cross University, Australia; [‡]University of Applied Sciences, Bremen, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	229
Governance and the Objectives of Airport Management	229
Airport Costs	230
Pricing Aeronautical Services	230
Non-Aeronautical Services	231
Airport Integration by Alliances and Mergers	231
Noise, Emissions, and Other Environmental Externalities	232
Airport Expansion	232
Performance of Airports	233
Summary	233
Relevant Websites	233
References	233

Introduction

Airport management has become a major area of research as airports turned from an arm of government to businesses. Airports are part of the aviation infrastructure, but offer a variety of commercial services such as shops for travelers. They are regarded as strategically important for the economic development of regions. In the following we provide an overview of the most important issues of airport management. It is naturally selective and [Graham \(2018\)](#) offers a broader overview.

We start with the changes in privatization, regulation, and competition and how this affects the objectives of airport management. In the second section, we discuss airport costs and review the empirical evidence for natural monopoly. This is followed by two sections on pricing which discuss also the role of airport slots for managing scarce capacity and delays. In section five, we discuss forms of horizontal integration with their implications for airport strategy and competition policy. Thereafter, we give an overview of airport externalities. Among them noise is the most important. It is regarded as the most important obstacle for airport expansion in metropolitan areas. Before concluding the effects of changes in governance of the performance of airports are reviewed.

Governance and the Objectives of Airport Management

Airports are today governed by very different institutions than 30 years ago ([Forsyth et al. 2004, 2010](#); [Gillen, 2011](#); [Winston and de Rus, 2008](#)). The downstream market has been liberalized and airlines increasingly pressure airports for better services at lower costs. Ground handling, where many airports had a monopoly, has been partially opened up to competition. Likewise, with commercial services in which airports often have a superior position compared to other local suppliers, the Internet has also given the passengers more choices.

In 1986, UK government started privatizing the British Airport Authority. Privatization was seen as a role model for other countries, but many countries were reluctant to follow. There are substantial differences in the speed, the objectives and the form of privatization. The spectrum ranges from the model of full privatization, which has been adopted in Australia, New Zealand, Portugal, and the UK, to the model of partial privatization of airports with a government majority share, such as Copenhagen, Florence, and Rome to airports with a minority share such as Düsseldorf, Frankfurt, Hamburg, Vienna, Paris. The traditionally public airports have been corporatized, such as Schiphol, Dublin, and Milan. US airports, which are mostly managed at local level, are also public. Their contractual relationships with airlines and with the federal government make them distinct from public airports in Europe.

Privatization very often preserved market power and did not increase competition. This was the case with BAA, which was sold as one entity against the warnings of commentators such as [Starkie and Thompson \(1985\)](#). It was broken up 20 years later. Separation of the Paris Airports (ADP), the Spanish (AENA) and the Portuguese (ANA) airport system was never seriously discussed. When Bratislava Airport was to be privatized, the nearby Vienna airport would have taken over had it not been prevented by competition authorities. Moves toward privatization also failed, as in the case of Berlin Airport or Schiphol airports. In both cases, the public owners also increased the market power as neighboring regional airports were taken over or prevented from market entry. Privatization has generally not been used to increase competition. Cross ownership restrictions are to be found in Australia, where the airports are far away from each other, but not in Europe with a much higher airport density.

While airports see themselves as being in competition with other airports, airlines view them as natural monopolies. The intensity of competition is a controversial issue ([Maertens, 2012](#); [Starkie, 2001](#)). While many issues seem to be difficult to resolve,

there seems to be a consensus that the strength of competition should be assessed on a case by case basis by the competition authorities. In Europe, only the UK Civil Aviation Authority and Dutch Competition Authority did so. This resulted in the de-designation of the airports of Manchester and Stansted. Heathrow and Gatwick are still regulated, though Gatwick nowadays in a light-handed manner. Schiphol's market power was assessed in 2000 and 2009 as being persistent. There is evidence that, in Europe competition is growing, especially for small regional airports. Hubs face competition in the transfer market, but have a dominant position in the origin and destination market. Typically a few large airports in each country have persistent market power. The EU Directive on Airport Charges demands that airports above of 5 million passengers should be regulated. This was regarded as arbitrary, and policy should obviously be more careful to find out which airports should be subject to regulation.

The UK model of an independent regulator controlled by Parliament was also seen as a role model for airport regulation. It has been adopted in Australia and India, but faced resistance in the European Union. The EU Commission demanded that the regulator should be a part of the government which has no ownership share in the regulated airports. This model has been practised by some member states, such as Austria's. It has worked in some cases, but both airlines and airports fear that the other side can politically influence the regulator. Because of such fears, the regulatory agency should be independent from the government. Such a model has not been used much in continental Europe. In spite of all this, in many countries, for example Germany and Spain, the regulator is not independent at all and regulatory capture remains an issue.

Traditionally, airports have been subjected to cost-based regulation. UK-regulation adopted a hybrid price cap, which sets the price cap every 5 years on the regulated asset base (Forsyth et al., 2004). This model of incentive regulation became influential. In Australia, a price cap based was used in the first phase after privatization. Thereafter airports were monitored, a type of light handed regulation. Light-handed regulation does not formally regulate airports, but threatens airports with re-regulation in case of market power abuse. Its effectiveness depends on the credibility of such a threat—an issue which is currently controversial (Littlechild, 2018). A form of light-handed regulation has been used also in New Zealand. In Europe, Copenhagen and more recently Gatwick have been subject to light-handed regulation. UK price cap regulation has been adopted by some countries such as Italy, Portugal, Spain, France, and Hungary, but the incentives have been changed and very often lessened. A traffic risk sharing mechanism was included in the cap, which inversely relates changes in traffic to the level of regulated charges, so that the revenues are more or less stable and incentives weak.

Regulation of the UK airports determined the price cap by including the non-aeronautical commercial revenues (single till). This has been criticized as extending regulation indirectly into the contestable markets of commercial services. Australia opted for a price cap based on a dual till, and in Europe the airports of Hamburg and Malta were the first ones to adopt a dual till price cap. There is a growing trend towards dual till regulation, but unfortunately this is very often not combined with price caps, but with cost-based regulation (e.g., Germany). In such a case charges are relatively high and provide little any incentive for cost and allocative efficiency.

The changes and the diversity of the governance of airports lead to a number of interesting issues. What are airports maximizing? While fully private airports might behave in accordance with the profit maximizing model, it is not clear how the partially privatized airports, in particular those with a minority private share behave. Do they maximize internal rents for managers and employees? Do they try to maximize size and output? Do they try to maximize connectivity for the region or even for the nation? Do they cooperate with and even collude with the dominant airline? What role do private investors play? Are they bringing in entrepreneurial spirit and superior expertise, or are they merely financial investors in a low-risk monopolistic airports with stable rates of return? Are commercialized public airports much different from private or partially privatized airports? Do they take the risks and rewards of incentive regulation as their private counterparts? Many of these questions have not been well researched at all.

Airport Costs

Early research in the 1990s suggested that economies of scale run out about 3–5 million passengers. This would have implied that there were no structural barriers to entry so that competition could work. However, market entry did not occur in regions with excess demand. Recent research based on data from 1991 to 2008 concluded that economies of scale prevail up to 80 million so that most airports are natural monopolies (Martín and Voltes-Dorta, 2011). However, these studies should be taken with some skepticism as the cost and scarcity of land is difficult to capture, and studies estimate only the private costs of airports which do not reflect costs for the users in form of longer taxi times for aircraft or longer paths for passenger to reach the gate. Furthermore, airports are multi-product firms which mean that economies of scope are important, but there are hardly any studies on economies of scope.

Airport investment is not only long term and lumpy by its nature, but also relation specific. As the runway and the terminal cannot be redeployed, the value of these assets depends on how airlines use it. Airports and airlines enter into long run contractual relationships among which regulatory contracts are one form to avoid hold up problems and under investment. This implies that a large part of the fixed costs for airports are sunk, but also some of the airline costs at an airport have the character of sunk costs depending on whether the airline is operating a hub or a node at an airport (Biggar, 2012).

Pricing Aeronautical Services

For many years, until major airports became congested, there was a simple system of pricing aeronautical services, such runways. Aeronautical assets were sunk costs, and marginal and variable costs low. There was a cost recovery problem, and almost all airports

were subject to a weight-based formula, and airports would charge airlines according to the maximum take-off weight (MTOW) of the flight. This might be regarded as a form of charging what the market will bear, since larger aircraft would generate more revenue than smaller aircraft. This approach also has some desirable efficiency properties, since it is an approximation to Ramsey pricing, or pricing to minimize the reduction in traffic as a result of setting a price above marginal cost, which is close to zero, with most of the cost of runways and the like being sunk costs. There was no problem of rationing scarce capacity.

In time, excess demand developed at many airports, especially the major, busy airports. For many airports, land constraints and locational factors make it difficult to add capacity, especially in the short to medium term. This means that there is a problem of rationing scarce capacity. One option which has not been used much at all with airports has been that of peak load pricing, which has been used very extensively for other businesses (such as airlines) subject to peaks. There have been a few cases where airports have set minimum charges for use during the peaks, and this discourages small aircraft from using the airport in busy times. There have been two main approaches to rationing capacity at airports—congestion and queuing, and slots.

The congestion and queuing approach is used in the US extensively, though not in many other countries. It is a system of first come, first served. It means that queues develop, and sometimes the delays from these queues last for some hours. The main problem with a queuing system is that it is wasteful of time and other resources. This has been recognized for many years, though there has been little progress in finding a solution.

The alternative approach has been one of setting up a slot system at the airport (Czerny et al., 2008). In order to use the airport at a specified time, an airline needs to have a slot. These slots will be in short supply in busy periods, and if an airline is unable to obtain a slot for its preferred time, it will need to accept a less preferred time, or not use the airport. The key advantage is the slot system is that queuing time can be reduced, though not eliminated. A critical question is how the valuable slots are allocated. In most countries, this is done by “grandfathering”—if an airline has a slot, it keeps it, unless it is not using it. This may be administratively simple, but it has its drawbacks. In particular, it becomes difficult for new airlines to access the airport—low cost carriers have responded to this by flying to smaller, out-of-town airports and built up their businesses in this way. Airlines also trade slots through “grey” markets, and in the UK, open trading of slots is permitted and used (the current record for a slot pair at London Heathrow airport is \$US75m). More flexible administrative slot rules could improve capacity utilization. There have been several attempts to reform the slot system to allow secondary trading and auctions, but so far these attempts have failed. While airports favor reform, airlines, which receive the slot rent, oppose it.

The move toward privatization and deregulation of airports has meant that aeronautical pricing structures have become more varied. Airports with spare capacity are actively seeking new airlines and new routes, and to this end, they have been offering airlines discounts for new routes. In some cases, they have been offering (confidential) fixed price contracts which last for several years. Another trend has been the move to passenger-based pricing.

Another trend which has been emerging has been a change in the level of charges. It has been argued by competition authorities and airline lobby groups that the greater freedom in pricing has enabled the airports to increase the overall level of charges. In earlier days, airport charges were mainly cost based. Privatization and less strict regulation have given the airports the incentive and freedom to increase their charges. There is some evidence that charges are rising. There may be good reasons for this. Investment in land-constrained airports is costly, and unit costs have been rising. The level of airport charges is becoming a hotly debated issue around the world.

Non-Aeronautical Services

Over the last few decades, non-aeronautical services have become a much more important part of an airport’s product mix, sometimes over 50% of the total, though this trend seems to have moderated in recent years (Graham, 2018). Airports provide a whole range of services, some of which are only distantly related to their core business. These services may include ground handling, retail, car parking, and provision of office and other space. The moves toward privatization and commercialization provided a stimulus to retail, so much so that airports are sometimes described as “shopping malls with runways attached.”

One management decision which airports must take concerns the reliance on contracting out. Some airports rely heavily on contracting out of services, such that a major airport may only have a few hundred direct employees. Others, particularly in Europe, have chosen to go with direct provision of services in-house. For example, some airports in Europe have chosen to provide (labor intensive) ground handling services in-house. This has not always been popular with airlines, who would prefer their own handling or use of their own chosen contractor. In the EU, larger airports are now required to open up ground handling to competition (Forsyth et al., 2010).

Airport Integration by Alliances and Mergers

Traditionally airports were either operated as single entities or as part of national airport systems. This changed with airport privatization. Airport holding companies were created such as Macquarie Airports and airports like ADP, and Frankfurt and Amsterdam Schiphol bid successfully for international contracts. Furthermore, some airports, like ADP and Schiphol, formed alliances. Some industry analysts viewed this process as being similar to the integration process in the airline industry. However the analogy is misleading because airport consolidation is not driven by network economies. In most cases, airports integrate to transfer know how.

Airport ownership across continents was mainly driven by new investment opportunities. Alliances have had mixed success. There are a few cases in which airports tried to actively increase market power through horizontal integration, like in the case of Vienna and Bratislava which was prevented by the competition authorities. On the other hand, given the lack of any cost savings from integration, the economic rationale for not breaking up airport system like ADP, AENA, and ANA at the time of privatization remains weak (Forsyth et al., 2010).

Noise, Emissions, and Other Environmental Externalities

Airports are generators of significant environmental externalities, such as noise, local emissions and greenhouse gas emissions (Lijesen et al., 2010). Of these, noise has been the most problematic until recently. Airport noise has been very unpopular with local communities, who lobby hard to prevent airport expansion or new airports (and governments often choose to locate new airports away from urban areas). The perception of the noise problem does not seem to be reducing, even though aircraft are becoming quieter. Governments have imposed a range of measures to reduce the noise problem, including curfews, restrictions on noisy aircraft, noise budgets and noise pricing. The more blunt measures, such as curfews, are effective, though they can be costly in discouraging economic activity. More nuanced measures, such as noise taxes or prices, do not seem to be particularly effective. Local emissions of harmful gases present similar problems to noise, though they are less obvious or controversial.

The big growing issue is that of airport's contribution to greenhouse gas emissions. These come from the direct emissions of the airport's operation, and the indirect effect of facilitation of emissions from aircraft (made worse if queues to take-off are long). Airports can control their own emissions to some degree, for example, by using power from renewable sources—a number of airports are now claiming to be carbon free. In the EU, they use electricity which is subject to the Emissions Trading Scheme, which puts a price on emissions. The actual construction and expansion of airports can be quite carbon intensive, with their reliance on large amounts of concrete, which is a carbon intensive product.

The role of airports in facilitating air transport emissions is an issue which is developing. Governments are being pressured to limit the expansion of airports, and to stop new airports from going ahead. Such restrictions have not been very carefully analyzed, and it is not clear to what extent they will reduce overall emissions. For example, restrictions placed on one airport may mainly lead to passengers using other, not restricted, airports.

Airport Expansion

With demand for air transport growing rapidly in most markets, airports need to expand their capacity. With small investments, such as a new apron, terminal upgrade or air bridge, this typically does not cause any major problems. The ownership of the airport will determine the criteria for evaluation. A public airport might seek to assess the benefits and costs of expansion to the owner, the airlines and the passengers. A privately owned airport might concentrate on the profitability of the expansion.

With larger investments, such as a major new terminal or runway, it will not be simply up to the airport to do what it wants (ITF, 2014). Many larger airports are subject to regulation, and the regulator has a large say in whether the investment goes ahead. With very large investments, the government may become the decision maker.

Taking the regulation case first, the airport will need to get the approval of the regulator to make the investment, as BAA did when it was developing Terminal 5 at London's Heathrow Airport. The airport may make the proposal to make the investment, but then the regulator takes over. It will make an assessment of whether the investment is worthwhile and investigate funding. Quite possibly, the airport will need to increase charges to fund the investment, and it will need approval from the regulator to increase charges. The regulator will also monitor quality of service, to check that the airport is not downgrading quality.

Airports themselves will have even less of a say when very large investments are being considered, such as when the Third Runway at Heathrow and the new airport in Berlin were proposed. The investment became a major public policy issue, and a separate Commission was set up to investigate options for London's Airports. The investment involved access to land and land use, access roads and climate policy.

A key part of these investments concerns the choice of evaluation technique. The most commonly used are cost benefit analysis (CBA) and economic impact assessment (EIA).

CBA was designed to evaluate investments from a public policy perspective. The effects on all of the stakeholders—the airport, airlines, passengers and the local and wider communities are considered. Externalities such as noise, and increasingly, carbon emissions are estimated and costed. A related development has been that of computable general equilibrium modeling—this enables the decision maker to evaluate effects on the wider economy; such effects on GDP and employment, and measure the effects on income distribution and carbon emissions. The London Airports Commission used both of these techniques.

A more controversial technique is EIA. This has been used in several airport evaluations, such that for the Berlin Airport. It claims to measure the impacts on GDP and employment, amongst other variables. It has several deficiencies, including not taking account of crowding out and effects outside the project itself, ignoring price changes, and it effectively assumes that all resources are free. Not unexpectedly, it always predicts large positive impacts from projects. As a result, it is not to be trusted as a means of evaluating investments.

Performance of Airports

Analyzing the effects of privatization, regulation and competition on the performance has been demanding (Liebert and Niemeier, 2013). This is partly due to the lack of good data. Measuring capital and including land turns out to be challenging, and most studies use some more or less problematic proxies. Comparisons across continents are extremely difficult and prone to distortion. Very often airports are just compared within certain regions for which the quality of data could be controlled. Early studies did not isolate the effects from changes in ownership from changes in regulation and competition. Furthermore, studies just focused on cost efficiency leaving the question whether efficiency gains have been passed, via lower airport charges, to the airlines and passenger unanswered. All this reduces the number of studies to a few reliable studies.

According to these, privatization does improve cost efficiency but partial privatization does not. This makes the dominant form of privatization in Continental Europe problematic. Competition has the strongest effects on efficiency if the catchment areas overlap much. However, this is very often not the case, for geographical or policy reasons, so that regulation must set the incentives (Adler and Liebert, 2014). Generally incentive regulation is superior to cost-based regulation and a dual till price cap is better than a single till. However, there are many types of incentive regulation, including light-handed regulations, and very often the strength of incentives is lessened. As a result, the slow move from cost-based regulation to incentive regulation in Europe brought some efficiency gains (Adler et al., 2015). Many countries adhere to cost-based regulation, even for partially privatized airports. Here it is doubtful as to whether increased competition has changed the airport performance by much.

Summary

The literature on airport management shows that airport management can be improved in many countries. Governance is in particular important as many countries have not developed a coherent system of privatization, regulation and competition. Airport capacity has generally not been managed efficiently at the individual airport and the national level, causing too high cost of underutilized airports in rural regions and too high congestion costs at over utilized airports. Slot allocation has not been reformed. Environmental policy has been similar ineffective to internalize externalities. The expansion of capacity has often not been rigorously assessed and there are cases of white elephant projects as well as delayed welfare increasing projects. Airport management has turned into a complex policy problem offering a fascinating field of research for a political economy approach to airport management.

Relevant Websites

<http://www.atrsworld.org>
www.eac-conference.com
www.garsonline.de
<https://www.itf-oecd.org/>
<https://iteaweb.org/>

References

- Adler, N., Liebert, V., 2014. Joint impact of competition, ownership form and economic regulation airport performance and pricing. *Transport. Res. Part A* 64, 92–109.
- Adler, N., Forsyth, P., Müller, J., Niemeier, H.-M., 2015. An economic assessment of airport incentive regulation. *Transport Policy* 41, 5–15.
- Biggar, D.R., 2012. Why regulate airports? A Re-examination of the rationale for airport regulation. *J. Transport Econ. Policy* 46, 367–380.
- Czerny, A., Forsyth, P., Gillen, D., Niemeier, H.-M. (Eds.), 2008. *Airport Slots. International Experiences and Options for Reform*. Ashgate, Aldershot.
- Forsyth, P., Gillen, D., Knorr, A., Mayer, O., Niemeier, H., Starkie, D. (Eds.), 2004. *The Economic Regulation of Airports*. Ashgate, Burlington.
- Forsyth, P., David Gillen, D., Müller, J., Niemeier, H.-M. (Eds.), 2010. *Airport Competition. The European Experience*. Ashgate, Burlington.
- Gillen, D., 2011. The evolution of airport ownership and governance. *J. Air Transport Manage.* 17, 3–13.
- Graham, A., 2018. *Managing Airports. An international Perspective*, 5th ed. Routledge, Abingdon.
- ITF, 2014. *Expanding Airport Capacity in Large Urban Areas*, ITF Round Tables, No. 153, OECD Publishing, Paris, doi:10.1787/9789282107393-en.
- Liebert, V., Niemeier, H.-M., 2013. A survey of empirical research on the productivity and efficiency measurement of airports. *J. Transport Econ. Policy* 47, 157–189.
- Lijesen, M., van der Straaten, W., Dekkers, J., van Elk, R., Blokdiijk, J., 2010. How much noise reduction at airports? *Transport. Res. Part D* 15, 51–59.
- Littlechild, S., 2018. Economic regulation of privatised airports: some lessons from UK experience. *Transport. Res. Part A: Policy Pract.* 114, 100–114.
- Maertens, S., 2012. Estimating the market power of airports in their catchment areas—a Europe-wide approach. *J. Transport Geogr.* 22, 10–18.
- Marín, J.C., Voltes-Dorta, A., 2011. The econometric estimation of airports' cost function. *Transport. Res. Part B: Methodol.* 45 (1), 112–127.
- Starkie, D., 2001. Reforming UK airport regulation. *J. Transport Econ. Policy* 35, 119–135.
- Starkie, D., Thompson, D., 1985. *Privatising London's Airports*, IFS Report Series No 16, Institute for Fiscal Studies, London.
- Winston, C., de Rus, G. (Eds.), 2008. *Aviation Infrastructure Performance A Study in Comparative Political Economy*. Brookings Institution Press, Washington.

Airport Regulation

Achim I. Czerny, Department of Logistics and Maritime Studies, Hong Kong Polytechnic University, Kowloon, Hong Kong

© 2021 Elsevier Ltd. All rights reserved.

Introduction	234
Airport Pricing of Infrastructure and Concession Businesses	234
Myopic Passengers	235
Foresighted Passengers	235
Airport Ancillary and City-Center Supplies	236
Cost-Based Versus Incentive Regulation	236
Pricing	236
Investments	237
Single-Till Versus Dual-Till Regulation	238
Conclusions	239
References	239

Introduction

A share of around 7.5% of airline operating costs is determined by user charges including airport charges (IATA, 2017). Compared to the share of aircraft fuel and oil costs, which is 22.5% (IATA, 2017), the role of airport charges seems relatively minor. Yet, airlines highlight the importance of regulatory controls for airport charges provided by the government (IATA, 2019). This can be understood given that the absolute value of the total airline payments to airports are substantive despite the fact that the importance of these payments relative to the total sum of operating costs seems minor. Governments are concerned about airport charges because they affect airline profits and air fares, thus, passenger surplus (Van Dender, 2007; Bilotkach et al., 2012).

The present article provides an overview over airport pricing and airport regulatory issues discussed in the aviation industry and in academic research. A major feature of this article is to use highly stylized economic environments to illustrate the most fundamental insights about airport businesses developed in aviation practice and research. The very fact that highly simplified and stylized models are used should, however, be considered as “a feature and not a bug” that helps to provide guidance and clarify fundamental economic relationships without distraction from potentially less relevant practical problems (Rodrik, 2015). For instance, the following discussion will concentrate on passenger businesses, and thus, abstract away from air cargo businesses because the main ideas presented in this article can easily be extended to air cargo businesses while concentrating on passenger traffic simplifies the language.

Airport Pricing of Infrastructure and Concession Businesses

The two-passenger example described in Table 1 is used to illustrate the airlines’ concerns about excessive airport pricing in the absence of regulatory controls provided by the government. Consider two passengers, A and B. The two passengers differ in their willingness to pay (WTP) for a flight. Passenger A’s WTP for a flight is 10, while passenger B’s WTP is only 3. Consider a situation where the airport can control how much they can earn from passengers and also how much passengers pay for their flights, which can be described as a situation where airports have and exercise market power (airlines are left out of this consideration because they are not essential when it comes to demonstrate the possibly detrimental effects of airport market power). The airport chooses between prices 10 and 3. If the airport charges a price of 10, then only passenger A will fly, while a price of 3 would ensure a demand of both passengers A and B. Consider airport operating costs of 2 per passenger. In this case the airport profit would be equal to 8 if a price of 10 is charged, which is high relative to the profit of 2 ($= 2 \times 3 - 2 \times 2$) that would be realized when a price of 3 is charged. Thus, if the airport pursues profit maximization, it would charge a price of 10 (not 3).

A price of 10 is excessive from the viewpoint of a government that cares about airport profits and also the surplus of passengers. To see this, note that the surplus of passengers is zero in the case of a price of 10 because passenger A pays a price that is exactly equal to her WTP while passenger B does not even fly. A price of 3 or less would secure a surplus of 7 or more to passenger A and ensure that passenger B flies. If the price is equal to 3, the sum of profits and passenger surplus is given by 9, which is higher than the corresponding sum if a price of 10 is charged. The price could be further reduced to 2. If a price of 2 is chosen, then the airport profit would be zero and the passengers’ aggregate surplus would be even further increased and given by 9. Many airports in the world follow such a not-for-profit approach. Airports in Canada serve as examples (Tretheway, 2006). Altogether, this illustrates the potentially conflicting goals of profit-maximizing airports and policy makers.

Table 1 Two-passenger example in the absence of non-aeronautical businesses

	<i>Passenger A</i>	<i>Passenger B</i>
Willingness to Pay (WTP)	10	3

Table 2 Two-passenger example in the presence of non-aeronautical businesses and myopic passengers

	<i>Passenger A</i>	<i>Passenger B</i>
WTP	10	3
Concession profit	5	10

Myopic Passengers

Until now, the framework considered airports as a firm that offers a single product or service. In the year 2017, 40% of airport revenues came from non-aeronautical revenue sources, which included revenues from the supply of retail concessions (30%), car parking (20%), property and real estate (15%) as well as other sources (ACI, 2019). Because a large share of airport revenues came from non-aeronautical businesses, the question arises how they affect the infrastructure pricing of a profit-maximizing airport.

Zhang and Zhang (1997, 2003, 2010) were among the first to study the effect of non-aeronautical airport businesses on aeronautical prices. Their research highlighted the role of revenues from concessions for airport cost recovery from the viewpoint of public and welfare-maximizing airports. Starkie (2001), on the other hand, highlighted the relevance of non-aeronautical businesses for profit-maximizing airports. He mentioned that the presence of non-aeronautical airport businesses would make formal price controls of aeronautical airport charges unnecessary. The example in Table 2 is used to illustrate his line of reasoning. In this example, passenger A travels if the ticket price does not exceed 10; however, in the case of flying, passenger A will make use of non-aeronautical airport services which generates an additional airport profit of 5. Passenger B travels if the ticket price does not exceed 3; however, in the case of flying, passenger B will generate an additional airport profit of 10. The fact that, in this example, the WTP for the flight is the same as in Table 1 is used to capture the notion of myopic passengers who do not anticipate a surplus from the consumption of non-aeronautical airport supplies when they decide upon traveling.

The scenario described in Table 2 implies a profit-maximizing price for aeronautical supplies equal to 3. Reducing the price for aeronautical businesses pays off for the profit-maximizing airport because it encourages passenger B to travel and passenger B generates a relatively high profit from non-aeronautical businesses. More specifically, the reduction of the aeronautical charge reduces the aeronautical revenue by 7 and increases the non-aeronautical revenue by 5 from passenger A, while it increases the aeronautical profit by 1 and non-aeronautical profit by 10 from passenger B. This scenario, thus, demonstrates that profit-maximizing airports may reduce the aeronautical charge in order to boost passenger demand and non-aeronautical businesses. This result has also been highlighted by Morrison (2009), Czerny (2013), Gillen and Mantin (2014), and Kidokoro et al. (2016) among others.

Foresighted Passengers

A crucial assumption of the example in Table 2 is whether passengers really do not anticipate the consumption of non-aeronautical airport services in the moment of booking a flight. Consider, for example, car rental and car parking services. Passengers and especially business passengers are likely to be aware of the need to use car rental or car parking services as part of their trips (Czerny et al., 2016b). It therefore seems important to capture the case of foresighted passengers. Czerny (2006) showed that the result of the above-mentioned example, profit-maximizing airports reducing the aeronautical charges to boost concession businesses, can be reversed in the case of foresighted passengers. He shows that, with foresighted passengers, profit-maximizing airports may develop concession contracts designed to reduce the charges for non-aeronautical supplies to boost their aeronautical businesses (Czerny, 2006; Czerny and Lindsey, 2014; Flores-Fillol et al., 2017).

The example in Table 3 illustrates the case of foresighted passengers. In the absence of airport car-rental supplies, the passengers' WTP are equal to the ones mentioned in Table 1. Car rentals do, however, change the WTP of passengers if airport car rental supplies

Table 3 Two-passenger example in the presence of non-aeronautical businesses and foresighted passengers

	<i>Passenger A</i>	<i>Passenger B</i>
WTP without car rent	10	3
WTP with car rent	9	9

are used. Consider the case of passenger A. The use of car rentals reduces this passenger's WTP. The reason may be that passenger A does not possess a driving license and, therefore, would have to pay a car parking fee, if he would have to rent the car. Or, passenger A simply enjoys being picked up by family members and friends. Renting a car would eliminate the joy of being picked up and therefore reduce the overall WTP for the trip. Passenger B is different. Passenger B may represent a business traveler, who needs a car to reduce the travel time needed to reach the final destination and for whom saving travel time is important. Passenger B represents passengers whose WTP for the whole trip would be significantly increased by the use of airport car-rental services.

Consider zero costs of car rental supplies. In this case, the profit-maximizing airport would charge a price equal to -1 for car rental services and a price of 10 for aeronautical services. This means that the airport maximizes profit by providing car rental services below costs to boost passenger demand and charge a high price for aeronautical services. The airport could offer take-it-or-leave-it contracts to concessionaires to impose the desired prices for non-aeronautical supplies. Passenger surplus is zero in this situation, which means that the profit-maximizing prices help the airport to internalize the entire surplus in the market in this specific example. Interestingly, passenger surplus could be raised in this example by imposing a minimum price, that is, a price floor for car rental services. For instance, policy makers could impose a lower limit (not an upper limit) on car rental prices. Consider a lower price limit of zero, which is just enough to ensure to avoid producing a profit loss from car rental supplies. In this situation, the profit-maximizing aeronautical charge is equal to 9, which leaves a passenger surplus of 1 to passenger A.^a

Airport Ancillary and City-Center Supplies

Airports often relate their charges for non-aeronautical ancillary supplies to city-center prices (Czerny and Zhang, 2018). For instance, airports in Atlanta and Vancouver impose an upper limit on prices for ancillary supplies based on the corresponding city-center prices. Hong Kong airport launched an "Urban Price Protection" scheme in 2004, which requires airport retailers and restaurant operators to charge prices that do not exceed the city-center prices. Capturing the rivalry between airport ancillary and city-center supplies, Czerny and Zhang (2018) derived an independence result, which states that, under certain conditions, the assessment of equilibrium prices charged by a profit-maximizing airport and profit-maximizing city-center prices are independent of whether passengers are myopic or foresighted. This is a surprising result given that profit-maximizing airport charges may be low to boost non-aeronautical businesses when passengers are myopic while the opposite could be true in the case of foresighted passengers.

Bracaglia et al. (2014) highlighted that passengers may be more foresighted today than in the past because airports are increasingly involved in e-commerce activities, which allows passengers to easily gain a more complete picture of all travel expenses including, for example, car rental, and car parking payments. That airports may induce congestion to increase the passengers' time spend at the airport and boost non-aeronautical airport businesses has been found by D'Alfonso et al. (2013) and Wan et al. (2015). Most airport studies abstract away from the possibility that people travel to the airport only to demand non-aeronautical airport supplies and not for flight reasons, which is a feature covered by D'Alfonso et al. (2017). Survey papers reviewing the research related to aeronautical and non-aeronautical airport services are provided by Zhang and Czerny (2012) and D'Alfonso and Bracaglia (2017).

Cost-Based Versus Incentive Regulation

The above analysis of profit-maximizing airport charges for aeronautical and non-aeronautical supplies reveals that profit-maximizing aeronautical charges can be excessive from the viewpoint of policy makers despite the important role of non-aeronautical businesses for airports. The present section describes cost-based and incentive-based forms of regulating airport aeronautical charges and their effects on profit-maximizing airport aeronautical charges and airport investments. The consideration of non-aeronautical businesses is not essential to demonstrate the effects of these two regulatory regimes and will therefore be postponed to the subsequent section.

Pricing

Cost-based forms of regulation have in common that they relate revenues with costs in the sense that firms can increase the revenue in the case of a cost increase (Czerny et al., 2016a). Cost-based regulation is, for example, popular in the United States where airports strictly operate as public enterprises (Bilotkach et al., 2012). One way of creating such a regulatory environment is to impose an upper limit on the Return on Investment (ROI), which may be defined as the fraction of profit over investment costs:

$$\text{ROI} = \frac{\text{Profit}}{\text{Investment costs}} \quad (1)$$

Table 4 is used to illustrate the functioning of cost-based regulation and its possible impacts on the airport's regulated prices. Consider an airport that can choose prices of 25, 10, and 5 for aeronautical services. The passenger demand (pax) is equal to 3 when

^aThat price floors on ancillary supplies can increase the surplus of consumers has recently been highlighted by Gomes and Tirole (2018).

Table 4 Cost-based regulation when investments are given

Price	Pax	Revenue	Investment	Profit	ROI
25	3	75	50	25	50%
10	10	100	50	50	100%
5	15	75	50	25	50%

the price is equal to 25, 10 when the price is equal to 10, and 15 when the price is equal to 3. Investments are given by 50. The ROI is maximized when a price of 10 is charged, which leads to a revenue of 100, a profit of 50, and an ROI value of 100%.

The regulator may be unsatisfied with the price of 10 because passenger numbers could be substantially increased by a price reduction from 10 to 5 while maintaining airport cost recovery. The regulator therefore tightens regulation by imposing an upper limit of 50% on the airport's ROI. The airport's response to this policy change is, however, ambiguous in the sense that both a price reduction from 10 to 5 and a price increase from 10 to 15 satisfy the tightened regulatory constraint. This means that there is a risk that the tightened regulation may actually increase and not reduce the price, which is a highly undesired outcome from the policy perspective because this would drastically reduce the passenger number from 10 to 3 instead of increasing the passenger number from 10 to 15.

The ambiguity arises from the fact that the cost-based regulation implicitly imposes a limit on revenues and therefore essentially functions as a revenue-cap regulation; but, revenue is the product of the price and passenger number and therefore the same revenue can be achieved with a low price and high number of passengers and a high price and a low number of passengers. That revenue caps can increase rather than decrease prices has been highlighted by [Crew and Kleindorfer \(1996\)](#).

A simple way to resolve this ambiguity regarding regulated prices is to directly(!) regulate prices instead of ROI (and revenue). An upper limit on prices equal to 5, called price-cap, would ensure that the airport charges a price of 5 rather than 10 or even 15.

Investments

Cost-based regulatory practices can be considered as "traditional" ([Bilotkach et al., 2012](#)), while there has been a gradual trend toward price-cap regulatory regimes in Europe, India, and elsewhere ([Adler et al., 2015](#)). The reason is that price-cap regulatory regimes provide strong incentives for cutting costs relative to cost-based regulation; therefore, price-cap regulation is also called *incentive regulation*.

The example provided in [Table 5](#) is used to illustrate the role of regulatory regimes in the form of cost-based and incentive regulations for investments. It shows passenger numbers, revenues, profits, and ROIs for investments of 50 and 67. Prices and passenger numbers, thus, also revenues are independent of whether investments are equal to 50 or 67. This captures the scenarios where the additional investments of 17 (= 67–50) are *absolutely unnecessary* in the sense that they do not affect capacity, service quality, and passenger number. Expensive architectural designs of buildings or construction materials perhaps appreciated by experts but not necessarily by passengers in general may serve as examples.

Would airports invest the additional amount of 17? This seems a ridiculous strategy because the investment is absolutely unnecessary by construction. A profit-maximizing airport that is unregulated certainly would not do the investment because it increases costs but leaves revenues unchanged, thus, reducing profit defined by the difference between revenues and costs. The picture changes under cost-based regulation.

If the airport's ROI is limited from above and investments are given, it can reduce or increase the price to satisfy the regulatory constraint. The resulting profit would be equal to 25. If investments are determined by the airport, the airport may leave the price unchanged and equal to 10 and respond to the tightening of the ROI requirements by spending the amount of 67 for investments. This is an alternative policy that satisfies the tightened regulatory constraint of an ROI that does not exceed 50%. Increasing the investment is in fact the better strategy for the profit-maximizing airport because it leads to a profit of 33 which is higher than 25. This outcome is a disaster from the policy viewpoint because prices are unchanged by the tightened policy and resources are wasted.

One may wonder whether the outcome where regulated firms waste resources to charge higher prices is realistic because regulators should intervene in the case of unjustified costs and not include them into the "regulatory asset base" used to evaluate regulated prices. The problem here is information asymmetry because regulators are less well informed about industry needs and therefore can have difficulties in assessing the economic value of investment strategies. Price-cap or incentive regulation is used as

Table 5 Cost-based regulation when investments are chosen by the airport

Investment	Price	Pax	Revenue	Profit	ROI
50/67	25	3	75	25/8	50%/14%
50/67	10	10	100	50/33	100%/49%
50/67	5	15	75	25/8	50%/14%

a device that should help deal with information asymmetry by reducing the firms' incentives to waste resources (Beesley and Littlechild, 1989). For instance, consider a price cap of 5. In this case, the airport revenue will be 75 and independent of investments. The result is that the airport has no incentive to waste resources and invest in unnecessary assets. That a move from cost-based to incentive regulation can indeed improve productive efficiency in airport practice has been confirmed by Adler et al. (2015).

Clearly not all investments are unnecessary because investments are needed to, for example, improve service quality. The advantage of cost-based regulation is that it creates conditions that support investments in service quality, while incentive regulation provides strong incentives to cut costs by reducing service quality (Oum et al., 2004; Czerny and Forsyth, 2008; Yang and Zhang, 2012). This problematic feature of incentive regulation will be discussed in more detail in the subsequent section.

Single-Till Versus Dual-Till Regulation

The previous section concentrated on the regulation of aeronautical charges in the absence of non-aeronautical profits. How to deal with profits from non-aeronautical airport businesses is a big question in airport regulation (Lu and Pagliari, 2004). There are two main approaches called single-till and dual-till regulations. With single-till regulation, the profits derived from airport concession services are used to cover the airport infrastructure cost for, for example, runway construction, while the airport is supposed to cover airport infrastructure costs entirely by aeronautical revenues with dual-till regulation (Czerny et al., 2016a).

The functioning of single-till and dual-till regulation is illustrated based on its application to incentive regulation. Under single-till regulation, the price-cap may be written as:

$$\text{Single-till aeronautical charge} = \frac{\text{Investment costs} - \text{Non-aeronautical profit}}{\text{Passenger numbers}} \quad (2)$$

The right-hand side displays the fraction of the difference between investment costs and non-aeronautical profits over passenger numbers. This formula ensures airport cost recovery where a part of the costs of infrastructure investments is covered by the profit generated by non-aeronautical businesses. One advantage of the single-till approach is that it does not require attributing investment costs to aeronautical and non-aeronautical airport businesses, which is arbitrary to a large extent. This is different in the case of dual-till regulation because this regulation regime absolutely requires attributing investment costs to aeronautical and non-aeronautical businesses. Under dual-till regulation, the price-cap may be written as:

$$\text{Dual-till aeronautical charge} = \frac{\text{Infrastructure investment costs}}{\text{Passenger numbers}} \quad (3)$$

Thus, dual-till regulation requires separation of investment costs in investment costs for aeronautical and non-aeronautical businesses. The dual-till price cap is then chosen such that the aeronautical revenues are sufficiently high to cover infrastructure costs. Typically, this means that the single-till price cap is smaller than the dual-till price cap (Kratzsch and Sieg, 2011).

With foresighted passengers, airports can partly compensate the tougher price cap imposed by single-till relative to dual-till regulation by an increase in prices for non-aeronautical supplies (Czerny and Zhang, 2018). Consider the passenger example in Table 3 for illustration. If the dual-till price cap for aeronautical prices is equal to 10 under dual-till regulation, while it is equal to 8 under single-till regulation. In this scenario, the profit-maximizing car rental price equals -1 in the case of dual-till regulation (the price-cap constraint is non-binding in the sense that the profit-maximizing aeronautical charge is independent of the presence of airport regulation), while the profit-maximizing car rental price is given by 2 under single-till regulation (where the price-cap constraint is strictly binding in the sense that the profit-maximizing aeronautical charge is strictly higher in the absence of regulation).

Table 6 illustrates how dual-till regulation can stimulate investments and increase the incentives to improve service quality relative to single-till regulation. The first column shows that the airport can choose between investing 50 or 100 where the higher investment of 100 relative to the lower investment of 50 increases passenger numbers by 5, which is independent of the aeronautical charge in this example. If the aeronautical charge is equal to 25, then the airport profit increases from 25 to 100, that is, by 75. This means that aeronautical charge is so high that the increase in passenger numbers boosts revenue to an extent that makes it easy to cover the higher investment costs. If the aeronautical charge is equal to 10, then the airport profit remains constant and is equal to 50 independent of whether the investment is equal to 50 or 100. This is because the higher service quality boosts revenues to an extent that is just high enough to cover the increase in investment costs. If the aeronautical charge is equal

Table 6 Single-till versus dual-till and service quality

Investment	Aeronautical charge	Pax	Revenue	Profit	ROI
50/100	25	3/8	75/200	25/100	50%/100%
50/100	10	10/15	100/150	50/50	100%/50%
50/100	5	15/20	75/100	25/0	50%/0%

to 5, then the additional investment in service quality reduces airport profit to zero because the revenue increase is too small to cover the higher investment costs.

Altogether, this illustrates how a relaxed price constraint under dual-till regulation can increase the incentives to invest in service quality relative to single-till regulation (Oum et al., 2004; Yang and Fu, 2015). In the specific scenarios discussed here, this is because dual-till regulation may lead to a price cap of 10 or 25 while single-till regulation may lead to a price cap of 5. In these scenarios only dual-till regulation could ensure that the regulated firm will invest in service quality. This further illustrates how incentive regulation can provide strong incentives to cut costs even when this leads to poor service quality, which is especially true under single-till regulation. The result is further in line with the airports' view that dual-till regulations can be associated with more innovations in commercial airport businesses relative to single-till regulation (ACI, 2017). Yang and Zhang (2011) highlighted how dual-till regulation can more effectively manage airport congestion in the absence of airport slot limitations because the higher prices under dual-till relative to single-till regulations reduce airport traffic, thus, congestion.

Conclusions

The discussion in this chapter illustrated that airport regulations can take the form of cost-based, incentive, single-till, and dual-till regulations. The discussion further illustrated that no regulatory regime can solve all pricing and investment problems. Policy makers rather have to choose whether they would like to maintain, for example, a closer control of airport aeronautical charges, which may favor the use of incentive and single-till regulations, or provide an environment good for investments and service quality, which may favor the use of cost-based and dual-till regulations. Policy makers may even reject such forms of economic regulation of airport aeronautical charges to avoid regulatory interferences in airport decisions or because of the presence of airport competition. Examples are airports in Australia, New Zealand, and the United Kingdom. In Australia and New Zealand, the prices and service quality of all major airports are subject to light-hand regulation involving the monitoring of airport behaviors but not directly regulations (Forsyth, 2004, 2008; Yang and Fu, 2015). Airports in the London area of the United Kingdom are freed from regulation because of airport competition and because long-term contracts between airports and airlines are considered an effective tool for airport control (Czerny et al., 2016a).

References

- Adler, N., Forsyth, P., Mueller, J., Niemeier, H.-M., 2015. An economic assessment of airport incentive regulation. *Transp. Policy* 41, 5–15.
- Airports Council International (ACI), 2017. Policy Brief. Airport Ownership, Economic Regulation and Financial Performance 2017/01. ACI World, Canada.
- Airports Council International (ACI), 2019. Airport Economics at a Glance. ACI World, Canada.
- Beesley, M.E., Littlechild, S.C., 1989. The regulation of privatized monopolies in the United Kingdom. *RAND J. Econ.* 20, 454–472.
- Bilotkach, V., Clougherty, J.A., Mueller, J., Zhang, A., 2012. Regulation, privatization, and airport charges: panel data evidence from European airports. *J. Reg. Econ.* 42, 73–94.
- Bracaglia, V., D'Alfonso, T., Nastasi, A., 2014. Competition between multiproduct airports. *Econ. Transp.* 3, 270–281.
- Crew, M., Kleindorfer, P.R., 1996. Price-caps and revenue-caps: incentives and disincentives for efficiency. In: Crew, M.A. (Ed.), *Pricing and Regulatory Innovations Under Increasing Competition*. Topics in Regulatory Economics and Policy Series, vol. 24. Springer, Boston, MA.
- Czerny, A.I., 2006. Price-cap regulation of airports: single-till versus dual-till. *J. Reg. Econ.* 30, 85–976.
- Czerny, A.I., 2013. Public versus private airport behavior when concession revenues exist. *Econ. Transp.* 2, 38–46.
- Czerny, A.I., Forsyth, P., 2008. Airport regulation, lumpy investments and slot limits. *WZB Econ. Politics Semi. Series*.
- Czerny, A.I., Lindsey, R., 2014. Multiproduct pricing with core and side goods. Unpublished.
- Czerny, A.I., Zhang, H., 2018. Rival airport ancillary and city-center supplies. Available from: SSRN: <https://papers.ssrn.com/abstract=3053312>.
- Czerny, A.I., Guilmard, C., Zhang, A., 2016a. Single-till versus dual-till regulation: where do academics and regulators (dis)agree? *J. Transp. Econ. Policy* 50, 350–368.
- Czerny, A.I., Shi, Z., Zhang, A., 2016b. Can market power be controlled by regulation of core prices alone? An empirical analysis of airport demand and car rental price. *Transp. Res. Part A* 91, 260–272.
- D'Alfonso, T., Bracaglia, V., 2017. Two sidedness and welfare neutrality in airport concessions. In: James, H., Peoples, Jr. (Eds.), *Advances in Airline Economics*, vol. 6. Emerald Books, England.
- D'Alfonso, T., Bracaglia, V., Wan, Y., 2017. Airport cities and multiproduct pricing. *J. Transp. Econ. Policy* 51, 290–312.
- D'Alfonso, T., Jiang, C., Wan, Y., 2013. Airport pricing, concession revenues and passenger types. *J. Transp. Econ. Policy* 47, 71–89.
- Flores-Fillol, R., Iozzi, A., Valletti, T., 2017. Platform pricing and consumer foresight: the case of airports. *J. Econ. Manage. Strat.* 27, 705–725.
- Forsyth, P., 2004. Replacing regulation: airport price monitoring in Australia. In: Forsyth, P., Gillen, D., Knorr, A., Mayer, O., Niemeier, H.-M., Starkie, D. (Eds.), *The Economic Regulation of Airports: Recent Developments in Australasia, North America, and Europe*. Ashgate, Aldershot.
- Forsyth, P., 2008. Airport policy in Australia and New Zealand: privatization, light-handed regulation and performance. In: Winston, C., de Rus, G. (Eds.), *Aviation Infrastructure Performance: A Study in Comparative Political Economy*. Brookings Institution Press, Washington, DC.
- Gillen, S., Mantin, B., 2014. The importance of concession revenues in the privatization of airports. *Transp. Res. Part E* 68, 164–177.
- Gomes, R., Tirole, J., 2018. Missed sales and the pricing of ancillary goods. *Quart. J. Econ.* 133, 2097–2169.
- International Air Transport Association (IATA), 2017. The key operating costs for airlines in 2016. IATA Economics' chart of the week.
- International Air Transport Association (IATA), 2019. Aviation Charges, fees and taxes. Fact Sheet.
- Kidokoro, Y., Lin, M.H., Zhang, A., 2016. A general equilibrium analysis of airport pricing, capacity and regulation. *J. Urban Econ.* 96, 142–155.
- Kratzsch, U., Sieg, G., 2011. Non-aviation revenues and their implications for airport regulation. *Transp. Res. Part E* 47, 755–763.
- Lu, C.-C., Pagliari, R.I., 2004. Evaluating the potential impact of alternative airport pricing approaches on social welfare. *Transp. Res. Part E* 40, 1–17.
- Morrison, S.A., 2009. Real estate, factory outlets and bricks: a note on non-aeronautical activities at commercial airports. *J. Air Transp. Manage.* 15, 112–115.
- Oum, T.H., Zhang, A., Zhang, Y., 2004. Alternative forms of economic regulation and their efficiency implication for airports. *J. Transp. Econ. Policy* 38, 217–246.
- Rodrik, D., 2015. *Economic Rules: The Rights and Wrongs of The Dismal Science*. W.W. Norton, New York.
- Starkie, S., 2001. Reforming UK airport regulation. *J. Transp. Econ. Policy* 35, 119–135.

- Tretheway, M.W., 2006. Airport policy in Canada: Limitations of the not-for-profit governance model. InterVISTAS Consulting report.
- Van Dender, K., 2007. Determinants of fares and operating revenues at US airports. *J. Urban Econ.* 62, 317–336.
- Wan, Y., Jiang, C., Zhang, A., 2015. Airport congestion pricing and terminal investment: effects of terminal congestion, passenger types, and concessions. *Transp. Res.* 82, 91–113.
- Wan, Y., Jiang, C., Zhang, A., 2015. Airport congestion pricing and terminal investment: effects of terminal congestion, passenger types, and concessions. *Transp. Res.* 82, 91–113. Yang, H., Fu, X., 2015. A comparison of price-cap and light-handed airport regulation with demand uncertainty. *Transp. Res. Part B* 73, 122–132.
- Yang, H., Zhang, A., 2011. Price-cap regulation of congested airports. *J. Regul. Econ.* 39, 293–312.
- Yang, H., Zhang, A., 2012. Impact of regulation on transport infrastructure capacity and service quality. *J. Transp. Econ. Policy* 46, 415–430.
- Zhang, A., Czerny, A.I., 2012. Airports and airlines economics and policy: an interpretive review of recent research. *Econ. Transp.* 1, 15–34.
- Zhang, A., Zhang, Y., 1997. Concession revenue and optimal airport pricing. *Transp. Res. Part E* 33, 287–296.
- Zhang, A., Zhang, Y., 2003. Airport charges and capacity expansion: effects of concessions and privatization. *J. Urban Econ.* 53, 54–75.
- Zhang, A., Zhang, Y., 2010. Airport capacity and congestion pricing with both aeronautical and commercial operations. *Transp. Res. Part B* 44, 404–413.

Air Route Planning and Development

Renan Peres de Oliveira*, Gui Lohmann[†], *Latin American Center for Transportation Economics, Aeronautics Institute of Technology, São José dos Campos, SP, Brazil; [†]School of Engineering and Built Environment, Griffith University, Nathan, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	241
Air Route Planning (ARP)	241
Demand-Driven Factors	241
Supply-Driven Factors	242
Air Route Development (ARD)	243
Stakeholders' Engagement	243
Destination Management Organizations (DMOs)	244
Airports	244
The Steps of the ARP and ARD Processes and Their Integration	245
Summary and Conclusions	247
See Also	247
Relevant Websites	247
References	247
Further Reading	248

Introduction

Since the turn of the 21st century, the growth in the air transport industry worldwide has been, by all means, impressive. Considering unique commercial city-pair connections, the sector nearly doubled, from approximately 11,000 routes in operation in 2000, to almost 20,000 in the year 2017 (IATA, 2018). As a result, there is an estimated daily average of more than 100,000 commercial passenger flights (IATA, 2018). These figures evidence an ongoing trend in the expansion of services in this industry.

While these numbers show considerable growth in the aviation sector, the establishment of services in unexploited markets continues to represent a significant challenge for airlines. Trial-and-error approaches, with new scheduled services being added according to their ability to have access to capital and assets, may prove to be risky, from both financial as well as reputational perspectives. The establishment of new routes require significant marketing investments to promote them, including the main competitive advantages offered by the new service offered (e.g., frequency, aircraft type, price, etc.).

In this chapter, we discuss two concepts closely related to these ideas, namely the processes of air route planning (ARP) and air route development (ARD). ARP, an activity usually done internally by airlines, is defined as the process of identifying the most efficient way of meeting the demand of a given route by providing adequate flight service. Attributes related to ARP include the number of seats offered on a route, the frequency of flights, the number of connections, the price, and the targeted level of comfort. In contrast, ARD refers to the combined industry effort to make a given route viable, and this is done through the engagement of several aviation and tourism stakeholders.

This chapter is organized as follows. In Section Air Route Planning, ARP is the focus, with a discussion on the determinants from the perspective of demand and supply. In Section Air Route Development, we emphasize the ARD process, presenting the objectives and the roles of the various stakeholders, predominantly airports, and destination management organizations (DMOs). Lastly, Section The Steps of the ARP and ARD Processes and their Integration concludes by bringing together the steps taken by airlines during ARP as well as by the remaining air transportation stakeholders during ARD, highlighting the interfaces between these two processes.

Air Route Planning (ARP)

Demand-Driven Factors

The most critical condition to be met for the establishment and maintenance of route operations is the need for potential passengers to travel between two places. Hence, an understanding of the determinants of air transportation passengers is paramount, and these can be divided into two main groups—geo-economic and airline service-related determinants (Wang and Song, 2010). Geo-economic factors, on one hand, distinguish themselves as they fall outside of the control of airlines. These include characteristics such as level of income, size of the population and demographic factors of the origin and destination regions, the distance between these points, the competition with secondary airports in the surrounding area, the existence/price of alternative transportation modes, the climate of both regions and their seasonality. In addition, macro-economic variables

Air Route Demand**Geo-economic factors**

- Income
- Population
- Demographics
- Distance
- Competing airports
- Price and existence of alternative modes
- Climate
- Seasonality
- Foreign trade
- Exchange rates

Service-related factors

- Level of comfort of the aircraft
- Aircraft capacity
- Aircraft range
- In-flight amenities
- Frequent flyer programs
- Safety record
- Airfares
- Travel time
- Frequency of departures
- Timing of departures
- Day of week
- Service delay

Figure 1 ARP demand-driven factors.

such as the levels of foreign trade, exchange rates and the whole economic situation of the endpoint regions are also taken into account (Lohmann and Vianna, 2016).

Airline service-related factors impacting the demand include the level of comfort of the aircraft, the fleet capacity and range, the presence of in-flight amenities, frequent flyer programs, the airlines' safety records and the passengers' perceptions of safety, and the price of airfares. There are also several other factors directly related to how passengers value their time, including the total travel time, the frequency and timing of departures, the days of the week when the service is available, as well as the level of service delays. Fig. 1 depicts these air route demand factors.

Supply-Driven Factors

While demand may be a necessary factor in the availability of air services, this may not be sufficient to make an air route feasible. Airlines pay close attention to a series of additional circumstances before reaching a final decision, as these may impose severe restrictions on operations. These can be divided into seven major groups (Fig. 2): political, socio-cultural, geographic, touristic, economic, market, and operational factors. The following discussion is based on the findings of Valente and Lohmann (2005), and Lohmann and Vianna (2016):

1. Political factors—relating to the broader geopolitical negotiations between two or more countries, including the social instability of the destination—such as the existence of wars, the threat of terrorism or internal or religious conflicts. In addition, air transport

Air Route Supply**Touristic factors**

- Type
- Tourism infrastructure
- Length of stay
- Destination life cycle
- Seasonality
- Directionality
- Destination appeal
- Destination awareness

Political factors

- Regulations
- Social instability

Socio-cultural factors

- Passenger profile
- Passenger behavior
- Stakeholder relationship

Economic factors

- Economic situation
- Economic forecasts
- Financial incentives
- Level of taxation
- Traffic density
- Airfares
- Load factor
- Taxation
- Costs of operation
- Cargo potential

Market factors

- Competition
- Alternative modes
- Reputation and image
- Publicity
- Advertisement

Operational factors

- Capacity/limitations of aircraft
- Schedule of aircraft
- Technological developments
- Airport infrastructure and constraints (slots, gates, congestion, runway length, curfews)
- Airport location
- Competing airports
- Staff
- Network contribution
- Feeder traffic
- Business strategy

Geographic factors

- Distance
- Location
- Topography
- Geomorphology
- Climate

Figure 2 ARP supply-driven factors.

regulatory frameworks fostering air traffic rights can be included here. For example, multi/bilateral air service agreements, open skies and/or liberalization of the air transportation market.

2. Socio-cultural factors—encompassing travelers' profiles (such as age or income) and behaviors (e.g., preference for certain type of aircraft); the airline's relationship with further stakeholders in the region.
3. Geographic factors—comprehending the route distance in terms of the locations of the destinations, the topographies (accessibility and ease of mobility) and geomorphologies (urban environments, islands and archipelagos, mountainous or rural areas), and climate, influencing not only aircraft operations but also peoples' desire for travel.
4. Tourism-related aspects—consisting of the type of tourism segment offered, including the competitiveness of the destination (e.g., the quality of lodging, catering, recreation and leisure facilities), expected length of stay, seasonality, directionality of passenger traffic between the two ends of the route, appeal and awareness of the destination, the current stage of the destination's life cycle (emerging vs. mature); and the existence of the appropriate tourism infrastructure and features which are attractive to visitors and influence their length of stay.
5. Economic factors—associated with the economic situation of the endpoint cities of the routes, the economic forecasts, the existence of airport financial incentives, the overall level of taxation, the expected traffic density on the route, as well as possible ranges in price of airfares to be charged, aircraft load factors, estimated costs of operating the route, and the potential for cargo freight.
6. Market characteristics—comprising competition from other airlines or other modes of transportation, the reputation and image of the airline (which can be affected by the failure of a route development endeavor), advertisement costs and potential spillover gains in publicity for the company.
7. Operational factors—comprehending the appropriateness of available aircraft capacity/limitations to operate the route, the current schedules of aircraft and staff, the costs/benefits of using new technology, airport infrastructure and constraints, including the length of the runway, slot/gate availability, existence of curfews and levels of congestion, location of the airport and the presence of competing airports in the same city or in its surroundings, the network contribution of a new route and its potential for providing feeder traffic at an airline's hub (allowing cross-subsidization of routes), and, not least, the company's business strategy.

Air Route Development (ARD)

Turning our attention to the ARD process, particular emphasis on the role air transportation stakeholders play is fundamental. Traditionally an initiative led by airlines, this process has undergone some significant changes over the years, as other stakeholders have become more and more involved, in particular in the early stage of demand exploration. Especially since the 1990s, airports have become progressively more commercially oriented. This has followed a series of global privatization efforts, seeing airports take a central role in the responsibility for and interest in promoting new routes, improving airlines' choice of aircraft, and increasing the frequency of services.

However, it is important to note that the ARD process may yet assume a range of broader objectives than the ones cited—ultimately depending on the nature of the airport operator and its relationship with the other parties involved. These may include improving air access, the capacity and the economic development of a community or region, or even retaining air services on current routes that may be at risk of suspension. While ARD is undoubtedly a growing commercial activity across the globe, individual countries like Singapore and the UAE have been particularly successful, notably in terms of cooperation and collaboration amongst various key stakeholders (Lohmann et al., 2009, Spasojevic and Lohmann, 2019).

Stakeholders' Engagement

Evidence from the literature has suggested that ARD initiatives are associated with the successful financial performance of airports, in recent times becoming increasingly more important for their profitability (Halpern and Graham, 2016). ARD may have a prominent role on the significant increase of non-aviation revenues from the increase number of passengers' consumption at airports, as well as additional revenues derived from airport charges. Moreover, these benefits may not be confined only to the airport, as spillover effects such as employment, tourism, trade and business opportunities may manifest throughout their surrounding regions, making airports key influential players in regional development.

In light of this, the growing involvement and influence of additional stakeholders in ARD can be easily understood. Furthermore, it is apparent that this change in paradigm has been well received by airlines, with evidence from recent studies indicating that the most important factor for an airline in operating a given route (after its profitability) is a good relationship with the relevant stakeholders (Stephenson et al., 2018).

Although airports are recognized as the leading stakeholders in ARD, the literature has suggested that tourism organizations are their most reliable partners (Stephenson et al., 2018). This is to be expected, as both airports and destination management organizations (DMOs) present similar objectives, rendering the whole partnership mutually beneficial. Nevertheless, other secondary stakeholders can also be mentioned which although not essential to the process, can impact the level of success and the efficiency of the desired outcome. These secondary stakeholders include local authorities, chambers of commerce, regional development organizations, other airports, and the various organizations comprising the tourism production chain (e.g., travel agencies, tourism

Stakeholders

Primary

- Airlines
- Airports
- Destination management organizations (DMOs)

Secondary

- Local authorities
- Chambers of commerce
- Other airports
- Tourism production chain
- Regional development organizations

Figure 3 ARP primary and secondary stakeholders.

operators, hotels, event organizers, tourist transport providers, among others). All these stakeholders, primary and secondary, are presented in [Fig. 3](#).

So far, this chapter has predominantly focused on how airlines engage with ARP. We now focus more specifically on the unique roles that DMOs and airports play in ARD.

Destination Management Organizations (DMOs)

Destination management organizations (DMOs) are specific bodies responsible for the development of tourism in a particular destination or area, and they are commonly established by local, regional, state, or national governments.

Given these characteristics, DMOs are provided with considerable political power, which they deploy by leading and coordinating the promotion of destinations along with airlines (employing marketing strategies, for example). Besides, DMOs may also be in a position to engage local businesses to influence the demand of a region, reducing the uncertainties associated with the operation of routes to or from it. In some cases, an airline may go as far as to seek collaboration when a route is not performing well, but they still may want to invest in it. In this way, these stakeholders also have a role in avoiding route suspensions ([Lohmann and Vianna, 2016](#)).

We note that such DMO initiatives will have their greatest impacts in the cases of smaller airports, where the chance of success is expected to be lower, given that these airports have limited budgets along with smaller markets and potential demands, making the additional engagement of stakeholders crucial. However, convincing airlines to establish new routes to regional destinations has its challenges and risks, with airlines favoring the main capital cities and metropolises.

Airports

Of all stakeholders, airports provide the strongest link to airlines, as symbiotic relations between them commonly emerge, reflected through several types of agreements which have come into being in the industry over the years. These include an airport having a signatory airline; the airline having ownership or control over airport facilities or signing a contract for long-term use; employment of airport revenue bonds to airlines; as well as revenue sharing between these two actors.

Airport managers' in-depth knowledge of the local infrastructure and the travel habits of their catchment area population confers a central role upon them, enabling them to engage a wide range of stakeholders and facilitate the creation, promotion, and amplification of ARD strategies to attract airline services.

[Halpern and Graham \(2015\)](#) indicate that the vast majority of airports are actively involved in route development initiatives, with the level of involvement being higher for larger airports, predominantly those located in the United States and Europe. Moreover, concerning ownership, results indicate that about 47% of publicly funded airports retain a route development team, with this proportion increasing to about 86% in airports operated by the private sector.

As airports compete for more and more limited airline capacity, they may resort to the use of incentives to attract or retain airlines, such as offering discounted charges for the use of their infrastructure. Notably, this appears to be widespread in Europe and the United States and at airports of all sizes, no matter whether publicly or privately owned, as indicated by research in this field.

Two main objectives commonly underlie the use of incentives for ARD: (1) route growth initiatives, associated with the addition of new destinations, and (2) volume growth initiatives, related to the development of passenger throughput, higher flight frequencies and/or greater aircraft capacities. Furthermore, the types of incentives that can be offered are broadly classified into financial and non-financial. Financial incentives can include direct payments made on a flight or passenger basis, direct payments made through a marketing budget where joint advertising and promotional campaigns for a new air service may be offered, discounts on airport charges, property rental and/or taxation, and risk sharing arrangements such as revenue guarantees, travel banks, and/or joint ventures. Non-financial incentives may involve regulatory and/or market information, the provision of contacts to feeder airlines, service providers, agents, tourism organizations, DMOs and the media, and aiding other stakeholders

in better serving these airlines. Additionally, airports may still offer support through lobbying activities, particularly concerning traffic rights and capacity obstacles.

Incentives can be a particularly relevant measure for attracting low-cost airlines, as charges may account for approximately 10% of their total operating costs. To reduce an airline's financial risks, in this way, the presence of incentives can be crucial for the economic viability of some routes, at least during their start-up phases. In addition, these incentives may even be justified, provided that carriers continue their air services after the incentives expire, as airlines may make considerable investments in a particular airport.

The Steps of the ARP and ARD Processes and Their Integration

After describing the significant factors considered by airlines in their planning of an air route and the primary stakeholders involved in the ARD process, as well as their roles and objectives, this section considers the steps taken through both processes. The discussion is based on Valente and Lohmann (2005), Halpern and Graham (2016), Stephenson et al. (2018), and an ASM Global Route Development (n.d.) White Paper devoted to the ARD process. While not particularly emphasized in this section, it is worth considering that the ways in which stakeholders will engage depends on the importance of the airport/city in terms of attracting new routes, the regulatory framework in place and the governance and leadership style of the organizations involved. Spasojevic et al. (2019) provide a thorough analysis of the leadership and governance attributes influencing ARD across several regions throughout the world.

From the airline's perspective, ARP can be seen as an on-going process, manifested through the addition and/or removal of routes or changes in the frequency of services, ultimately creating the 'structure' of the airline's network. This whole process is depicted in the middle column of Fig. 4. It begins with a route solicitation for the airline's route planning department, which may

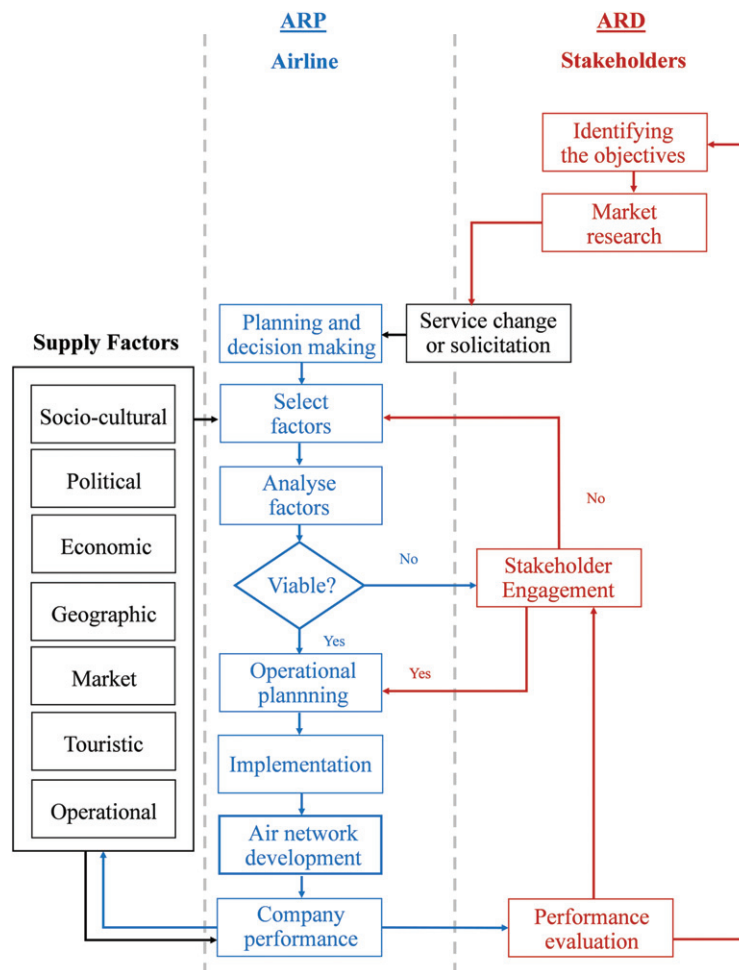


Figure 4 The steps of the ARP and ARD processes.

be internal (usually advanced by the airline's marketing and sales department) or external (promoted by air transportation stakeholders). After that, as described in Section Air Route Planning, the relevant supply factors to be considered are selected and carefully analyzed. In the case of an opportunity being verified, the planning department will then conduct preliminary surveys about the feasibility of operating the market. Having technical and operational responsibilities, it will provide analyses of the best fleet to be used, how the route fits into the airline's network, the costs to be incurred, expected yield, number of staff required, and the necessity (or possibility) of intermediate stops. The planning department (or, depending on the airline, its board of directors), may decide for the operation of the route or against it, possibly being persuaded by stakeholders who are regularly involved in the whole process. Regardless of the decision to start or not a route, airlines will be in possession of critical intelligence data, eventually only waiting for the right moment for developing a particular route. Having decided on its operation, the process is advanced to the phase of operational planning, consisting of activities related to network planning, fleet assignment, aircraft routing and crew rostering.

New routes are implemented after being thoroughly planned, and they become part of the set of markets operated by the airline, contributing to the achievement of the network's development. However, the control and feedback stage of the company's performance is still crucial. If the targeted market share and profit do not match expected values, the company may need to revisit its strategy. Moreover, the company's performance may influence a set of the air route supply factors, more importantly, operational and/or economic factors which are necessary for the operation of the new route and even of the other routes in its network. Likewise, supply factors may also influence a company's performance, with the market, political, and social factors being particularly impactful.

Turning our attention to the ARD process, this also constitutes a constant process from the perspective of the various stakeholders, and it occurs in parallel with the ARP process. It is depicted in [Figure 4](#), in the right column. This process begins with the identification of the objectives of the route development, be it by the achievement of connectivity or traffic targets and many more of the broader objectives related to regional impacts, as listed in Section Air Route Development.

Having identified the objectives, a route development team will take matters of the process. This team will be responsible for three critical tasks—that of market research and forecast; negotiation and relationship management (i.e., the attraction of new airlines); and marketing (the creation of awareness and building of passenger numbers). The route development team will begin by undertaking a primary assessment of the size of the region's catchment area and hence quantify the potential demand for air travel. The necessary information will include the population size, residents' characteristics, their propensity for travel (or in other words, the factors influencing outbound traffic) and the region's tourism and business appeal (the factors affecting inbound traffic). Similar to the route planning department of an airline, the same set of factors will be selected and analyzed, and the resulting information will be compared with current operations to identify unserved or underserved routes and the potential market share to be achieved. At this stage, it is crucial to be realistic about the whole process, as overambitious ARD on ultimately unsustainable routes can have appalling effects on traffic development at an airport in the long run. It is important to remember the existence of a path-dependency in the development of air services, as preceding route development actions may limit the capacity of change of subsequent ones. In other words, there may be multiple optimal actions at any given time, but not all of those will lead to an optimal solution in the long-run—implying the possibility of increased levels of service by the appropriate selection of routes to be developed at each period ([Kita et al., 2005](#)). As a follow-up, an analysis of the competition will be conducted. An understanding of other airports in the analyzed catchment area (e.g., their market segments, relative performances or relative airfares) will be essential. Additionally, an assessment of other transport options; competing stakeholders; and even of communications technology possibilities (e.g., video chat and voice call applications) can be implemented, in order to better quantify the market potential (accounting for all possible factors which may reduce it). Finally, the team will analyze which airlines may be suitable to operate the new route and will produce a financial viability assessment, taking their business models into consideration.

With all this information in place, the route development team will communicate it to the target airlines, making service solicitations (in the case of a new market to be operated by that carrier) or a solicitation for change in existing services. At this stage, ARD opportunities will be marketed, usually through the presentation of a business case at the airline's offices. As previously mentioned, this can be interpreted as an external route solicitation. Price incentives and marketing support may also be established at this stage. A good relation with DMOs pays off at this stage, as the promotion of the destination (as opposed to the airport infrastructure) will be key for a favorable outcome of the negotiation.

Having presented the opportunities, stakeholders will stay engaged in the process, trying to influence airlines to begin or retain operations on a given route. To get the most out of this stage, it is essential for stakeholders to determine processes and procedures for dealing with airlines at all stages of ARD from their first contact, from the launch of a new or a changed route, to on-going care necessary to ensure that route opportunities are maximized and a long-term relationship is established and maintained. At this stage, stakeholders may also be involved in actions focused on avoiding route suspensions, through their influence over many of the supply factors presented in Section Air Route Planning (the touristic, political, and social factors being the most obvious ones). Particularly in the case of an airport, the offer of additional financial and/or non-financial incentives may also take place at this stage.

Lastly, it is vital that stakeholders be aware of how well airlines are performing within the entire portfolio of the markets in which they operate—and especially in those influenced by their initiatives. The establishment, maintenance, and growth of new services all require ongoing evaluations. Although airports may assume a leading role at this stage, the fact is that stakeholders must control and evaluate their ARD efforts, while also assessing the performance of individual routes on a regular basis to understand whether there is

indeed progress, identifying issues or problems and addressing them before they become worse. For continued success, ongoing route-performance analysis is crucial, as is a long-term relationship with the airlines. The results provided by this control and evaluation stage will then be fed into the setting of new or revised objectives, the first stage of the stakeholders' ARD process, thus closing the cycle. Likewise, such results will also offer relevant information about the possible avoidance of route suspensions (also associated with the stakeholder engagement stage).

Summary and Conclusions

The establishment of new routes, particularly on previously unexplored markets, offers some important challenges for airlines. This requires thorough planning with a careful estimation of the potential demand as well as a precise understanding of the conditions which must be met to provide a sustained market development. To the same extent, it also requires a critical analysis of the factors, from both the demand and supply sides, that influence market growth and are beyond the control of the airline. Such collection of intelligence data is an important step in the airline internal process of ARP. However, external stakeholders have been playing an increasingly important role in such endeavors on what has been called the ARD process. Airports, progressively more commercially-oriented, especially after a worldwide privatization trend around the world, have had a central role in such activities, as these have acquired a greater role in the profitability of their operations. In partnership with DMOs, airports have enabled the reduction of financial risks for many airlines, providing many spillover effects for regional development. ARP and ARD, nevertheless, do not function in isolation or are finished after the initiation of a target route. On the contrary, these are ongoing, cyclic and parallel processes. For continued success, ongoing route-performance analysis is crucial, irrespective of the process being carried out internally or externally to the airline, as is a long-term relationship between the airlines and the stakeholders.

See Also

What Drives Transport Demand Trends? The Chicken or Egg Problem; Demand for Passenger Transport; Elasticities For Travel Demand; Demand Management and Capacity Planning of Airports; The consumer Surplus, the Rule of the Half Approximation: Existing and Generated Demand; Airline Economics; Privatization and Deregulation of the Airline Industry; Latent Demand and Induced Travel; Route Choice and Network Modeling; Travel Demand Forecasting; Where are We and What are the Emerging Issues; Travel Model Calibration and Validation; Introduction to Air Transport; The History of Air Transport; The Geography of Air Transport; The Future of Air Transport; Air Transport and Territorial Scale; Airport and Airport Network Planning; Airport Management; Airport Regulation; Airline Management; Airline Regulation; Place Perception and Travel Behavior; Demand For Air Travel and Income Elasticity; Air Transport; Planning Tourism Travel

Relevant Websites

Airline Network News and Analysis: www.anna.aero.
Aviation Economics: www.aviationeconomics.com.
International Air Transport Association: www.iata.org.

References

- ASM Global Route Development (no date). The fundamentals of route development: The essential guide to air service development and airport strategy [White paper]. Retrieved on October 4, 2019.
- Halpern, N., Graham, A., 2015. Airport route development: a survey of current practice. *Tour. Manag.* 46, 213–221.
- Halpern, N., Graham, A., 2016. Factors affecting airport route development activity and performance. *J. Air Trans. Manag.* 56, 69–78.
- IATA, 2018. IATA Annual Review 2018. Montreal: International Air Transport Association.
- Kita, H., Koike, A., Tanimoto, K., 2005. Air service development of local airports and its influence on the formation of aviation networks. *Res. Transp. Econ.* 13, 233–249.
- Lohmann, G., Albers, S., Koch, B., Pavlovich, K., 2009. From hub to tourist destination – an explorative study of Singapore and Dubai's aviation-based transformation. *J. Air Trans. Manag.* 15 (5), 205–211.
- Lohmann, G., Vianna, C., 2016. Air route suspension: the role of stakeholder engagement and aviation and non-aviation factors. *J. Air Trans. Manag.* 53, 199–210.
- Spasojevic, B., Lohmann, G., 2019. Air Route Development and Transit Tourism in the Middle East. In: Timothy, D. (Ed.), *Routledge Handbook on Tourism in the Middle East and North Africa*. Routledge, Oxon & New York, pp. 290–305.
- Spasojevic, B., Lohmann, G., Scott, N., 2019. Leadership and governance in air route development. *Ann. Tour. Res.* 78.
- Stephenson, C., Lohmann, G., Spasojevic, B., 2018. Stakeholder engagement in the development of international air services: a case study on Adelaide Airport. *J. Air Trans. Manag.* 71, 45–54.
- Valente, F.J., Lohmann, G., 2005. Tourism as a decision factor in air route planning among Brazilian regional airlines. *Proceedings of the IX Air Transport Research Society World Conference*. Rio de Janeiro: Air Transport Research Society.
- Wang, M., Song, H., 2010. Air travel demand studies: a review. *J. China Tour. Res.* 6 (1), 29–49.

Further Reading

- Allroggen, F., Malina, R., Lenz, A.K., 2013. Which factors impact on the presence of incentives for route and traffic development? Econometric evidence from European airports. *Transp. Res. Part E: Logist. Transp. Rev.* 60, 49–61.
- Aviation Economics, 2014. What does good route development look like? [Web log post].
- Lohmann, G., Duval, D.T., 2014. Destination morphology: a new framework to understand tourism–transport issues? *J. Destin. Market. Manag.* 3 (3), 133–136.
- Goodovitch, T., 1996. A theory of air transport development. *Transp. Plann. Technol.* 20 (1), 1–13.
- Swan, W.M., 2002. Airline route developments: A review of history. *J. Air Trans. Manag.* 8 (5), 349–353.
- Weber, M., Williams, G., 2001. Drivers of long-haul air transport route development. *J. Trans. Geog.* 9 (4), 243–254.

Airline Management

Sveinn Vidar Gudmundsson, 2 Rue des Pyrenees, Escalquens, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	249
Internal Environment	249
Strategy	250
Product	250
Distribution and Promotion	250
Operations and Network	251
Operations	251
Network	251
Pricing and Markets	251
Elasticity of Demand	251
Market Selection	252
Revenue Management	252
Organization and Human Resources	252
Human Resources	252
Organization	252
External Environment	252
Regulation	253
The Economy	253
Political Climate	253
Taxation	254
Value Added Tax (VAT)	254
Fuel Tax	254
Environmental Taxes	255
Societal Changes	255
Economic Wellbeing	255
Price Elasticity of Demand	255
Leisure	255
Migration	256
Education	256
Conclusion	256
References	256
Further Reading	257

Introduction

Airline management is the organization and coordination of all airline activities to achieve some defined objectives and goals. Strategies are employed by managers to help the airline to get to the objective within the timeframe and resources allocated. The management of airlines involves the integration of many functions through policy, organizing, planning, controlling, and directing of resources toward an objective. Thus, airline managers deal with both internal and external variables, some of which can be manipulated, but in other cases the managers can only anticipate changes and react. This paper is divided into two parts the *internal environment* and the *external environment*. Part 1, internal environment, contains the following: strategy, product, distribution and promotion, operations and network, pricing and markets, and organization and human resources. Part 2, external environment, contains the following: regulation, the economy, political climate, taxation, and societal changes.

Internal Environment

The key internal variables for airline management are strategy, product, distribution and promotion, operations and network, pricing and markets, as well as organization and human resources.

Strategy

The world's air transport liberalization trend started with the advent of the Airline Deregulation Act of 1978 in the United States. Not only was this the start of a worldwide regulatory reformation in air transport but also a new era of segmented business models. Before deregulation airlines were segmented into full-service carriers, regional carriers, cargo carriers and charters. However, after deregulation the low-cost business model took off, usually cloned on Southwest Airlines, a carrier that had thrived in the liberalized Texas intrastate market before deregulation took place. A large part of air transport growth and new services in deregulated domestic markets can be traced to these carriers.

There are two generic strategies that polarize the airlines and their business models today, namely cost leadership and differentiation (Porter, 1980). The former strategy tries to achieve the lowest possible costs in order to offer very low fares. What defined this strategy for the airlines was the attempt by Southwest Airlines to compete with cars and buses on short-haul secondary markets in the Texas intrastate market. In order to generate such new demand costs had to be very low to offer low enough fares to pull people out of their cars (Gudmundsson, 1998). This has been termed as the *Southwest Effect* (Bennett and Craun, 1993) and denotes the price level where a rapid increase occurs in demand on a particular route. The Southwest Effect also explains why most low-cost airlines do not charge monopoly prices on monopoly routes, unlike the full service carriers.

To achieve very low-cost in comparison with the differentiators, low-cost airlines offer fewer service features than the full service airlines, and operate predominantly point-to-point services to secondary cities. Simplification of operations keeps costs down by avoiding both the direct and indirect costs of performing "unnecessary" activities. In this way a low-cost airline only offers services that matter the most to the customer, a seat between A and B, that parts on time. However, over time both low-cost and full service airlines have focused on increasing revenues through unbundling of services. For example, by charging fees for: priority boarding, emergency exits seats with more legroom, food and beverages on board, and hold baggage. Low-cost airlines will normally offer extra services as well, so far as the associated activity does not cause complexity in operations incurring added direct and indirect costs. For example, most low-cost airlines will not offer connection services, or business class seats. In this way, unbundling causes convergence of low-cost and network carrier services in the mind of the customer. To come to the point, low-cost airlines will strive to maintain high fit between all activities performed to keep the cost structure low.

Full service carriers differ by offering a wider range of service features and more extensive customer care. What is more, such carriers usually operate long-haul routes and hub and spoke networks to generate economies of scope and density that boost aircraft load factors. Hub and spoke means hauling passengers and cargo from many points of origin to many destinations through a single centre, the hub airport (Oum and Tretheway, 1990). Thus, full service carriers use differentiation strategy based on the assumption that customers are willing to pay a premium price for more service features and high connectivity. This is especially true for business travelers that fly on behalf of their companies. Today's full service carriers offer an extensive global network through branded alliances, such as skyteam, oneworld, and Star, to help meet the total travel needs of the customers and assure seamless connections between the alliance partners.

The low-cost business model was an important innovation enabling new carriers to enter liberalized air transport markets and mount effective competition with the full service carriers in domestic markets. However, the low-cost model works well on point to point short haul routes, but has not enjoyed the same success on long-haul routes, a market that still remains the key competitive advantage of the full service carriers. In other words, there appears to be a mobility barrier traversing from a low-cost to a full service network carrier. A mobility barrier (Mascarenhas and Aaker, 1989) refers to obstacles in moving from one strategic group of firms (low-cost carriers) within an industry to another group (network carriers) and may include barriers to entry, barriers to exit, barriers to changes in market position, and barriers embedded in cohesive business models.

Product

Whereas, a low-cost airline may offer only the most important product features sought by the customer, a full service carrier will offer a wide range of features fitting a more expansive network of both short-haul and long-haul flights, such as food on board, airport lounges, frequent flyer programs, network connectivity, and ticket flexibility. Thus, the product is strongly influenced by the generic strategy of the airline.

The elemental product attributes of a low-cost airline are safety, reliability, punctuality point-to-point service, low fare, flexible baggage allowance, hold bag fee, one-way fares, a variety of food and beverages for sale, priority boarding for a fee, and selling insurance to reduce customer's perceived risk if flights are late or cancelled. For full service network carriers we can add some important service differences such as a global network through alliances, seamless connections throughout the world, frequent flyer programs that reward the most valuable customers with personal benefits such as free flights, assistance and rerouting in case of missed flights, and so forth. Defining all the different product features of airlines to place them in either the low-cost or the full service categories is becoming somewhat blurred as low-cost and full service airlines have unbundled the product and new business models have emerged that can be considered hybrid (Mason and Morrison, 2008). In other words, airline business models are now more blurred, but the key difference between airlines of different origins, low-cost or full service, remains the hub-and-spoke versus point-to-point and long-haul versus short-haul.

Distribution and Promotion

The largest global distribution systems (GDS) are Amadeus, Sabre and Travelport (Galileo, Worldspan and Apollo), but there are also a regional GDSs like Travelsky in China, and KIU System in Latin America (Williams and Rhoades, 2017). In the past airlines

extended their own internal computer reservations systems (CRS) to travel agencies, but with time these became independent businesses that access the internal seat inventories of practically all airlines. Although airlines are increasingly emphasizing booking through their own websites, the GDS still plays a role as a “one-stop-shop” for complex travel needs, such as multiple-itineraries involving airline interlining as well as hotels and other travel services, and for travel agencies specializing in corporate travel management. Today’s GDS systems are designed to interoperate with all major airline systems, and provide total IT system solutions to the airlines. Thus, GDS will exist in one form or another but the airline distribution industry has become much more segmented than in the past when the GDSs operated in an oligopolistic market.

Today, a larger and larger proportion of flights is booked by customers themselves through internet booking sites either the website of the airline itself or non-GDS virtual booking sites. Disintermediation takes place when airlines try to circumvent intermediaries like travel agencies and GDS. However, despite disintermediation in terms of GDS sales and travel agents, another type of intermediation appeared in the form of websites that indexed airline offers through the search engines (Kracht and Wang, 2010). Such virtual intermediaries provide customers with a comparison of a broad selection of flights and fares that is otherwise only available to travel agents using the GDS systems. Thus, virtual intermediaries provide information transparency about fares and services that would otherwise be tedious to unravel for customers using the internet for flight bookings.

Operations and Network

Operations

Maximizing operating profits through optimum use of resources is the global objective of airline operations management. Thus, airline operations strive to match aircraft capacity to demand, and maximize productivity of the resources used, especially the aircraft and the employees. US FAA regulation Part 135 specifies maximum pilot duty period as 1200 flight hours in any calendar year, 120 flight hours in any calendar month, and 34 flight hours in any 7 consecutive days, putting major constraints on how pilots are allocated, in addition to aircraft type ratings they possess. Thus, airline operations are occupied with employing the right aircrafts, scheduling to maximize daily utilization, fulfill maintenance requirements and mandatory checks, assign the right crew (duty times and qualifications), and to assure smooth turnarounds at airports (ground handling). Airline operations managers strive to optimize schedules so that slack (aircraft on the ground) is minimized. However, the lower the slack in the schedule the greater the effect of delays (weather delays, aircraft breakdowns, etc.) in the network, affecting aircraft availability, crews pairing (duty times and pilot qualifications) and passengers (missed connecting flights, etc.). A major disruption of airline schedules can take days to correct due to domino effects (Kohl et al., 2007).

Aircraft have four levels of mandatory inspections (A, B, C, D) that are accounted for in the annual scheduling of aircraft (Lee et al., 2008). The A Check is roughly once a week or approximately every 60 flight hours; the B Check is once a month or every 300–500 flight hours; the C Check is performed after approximately 1800 flight hours or 2000 cycles and takes between 3 and 5 weeks; and the D Check is a major check that involves dismantling parts of the aircraft for detail inspection of its metal skin, structure and systems. The D Check can involve up to 50 thousand man-hours over 8–12 weeks and is performed every 6–12 years depending on aircraft type and age.

Network

Airline networks are at three levels: airline route networks, alliance route networks, and global route networks (Hsu and Shih, 2008). Airline route networks are shaped by airline strategy, demand, available fleet, marketing and sales. Whereas low-cost airlines tend to operate only point-to-point networks with limited if any network extension through alliances, the network carriers usually have extensive alliance relations and intercontinental long-haul routes. When designing alliance route networks much emphasis is on global network coverage and seamless connections in major airports. This is one key difference between low-cost and network airlines that remains difficult for low-cost airline to duplicate, for two principal reasons: First, low-cost airlines rarely operate hub and spoke networks as these can affect negatively on aircraft utilization. As the arriving and departing banks of aircraft become larger in hub and spoke airports, the ground time of aircraft tends to increase, increasing operating costs because of lower aircraft utilization. Second, low-cost airlines tend to keep costs down by operating a single type of aircraft aimed at short to medium haul routes and quick turnarounds, avoiding long-haul aircraft that are more costly to operate. In other words, network carriers seek economies of density and scope through hubs, while low-cost airlines seek *Southwest Effect* through very low fares in secondary point-to-point markets.

Over time the global airline route network has become dense with more and more communities being served, mostly by low-cost and regional airlines operating into secondary airports. What is more, with increased congestion in major airports, frequencies and new services have spilled over to other less congested major airports and secondary airports (Gudmundsson et al., 2014; Redondi and Gudmundsson, 2016). Thus, air transport services are more accessible for larger and larger part of the global population, but increased network density means also that more and more communities resist airport infrastructure expansion due to environmental and health concerns.

Pricing and Markets

Elasticity of Demand

Price elasticity of demand for air transport differs because passengers fly for different reasons and have different income levels and are often flying on behalf of employers rather than for own funds. Thus, we broadly distinguish between business and leisure travelers.

Whereas, leisure travelers strive to maximize income constrained utility and holiday experiences, business travel is an input in the firm's operations, and time and service levels usually play a greater role than price. Thus, the marginal utility of air travel differs between the two segments causing different demand sensitivities to two relationships, namely time over price and price over time. Thus, business travel, on average, tends to be less price elastic than demand for leisure travel (Castelli et al., 2003). For these reasons, understanding the price elasticity of demand for different markets is essential for pricing and supply decisions in the airlines.

Market Selection

Market selection, the destinations the airline serves, is one of the key decision variables in airline management. No market is the same in how it reacts to airline offerings and prices. For this reason, airlines analyse markets carefully by assessing demographic information such as income levels (see Calzada and Fageda, 2019), type of businesses operating in the area, population size and growth, political composition of the population, minorities and proportion of foreign origin residents, average and disposable income by customer segments, regional economic growth, and so forth. With this information combined with historical data the airline can predict market reaction to different fare classes, such as leisure, visiting friends and relatives, and business travel. For a low-cost airline, the most important aspect of analyzing markets is to determine at what price the *Southwest Effect* kicks in, namely a kink in the demand curve that represents price-market stimulation.

Revenue Management

Airlines have since the 1980s utilized sophisticated information systems to determine fare offerings, the so-called yield management system that sits on top of the internal computer reservation system (see Smith et al., 1992). Such systems allow the airline to maximize average revenue per flight by varying fares based on seat demand and the propensity to pay by different customer segments. Some segments are time sensitive while other segments are price sensitive. Based on this information the yield management system attempts to separate the segments as much as possible to prevent sales cannibalization between the various customer segments. For these reasons, most network carriers will retain as many high fare seats as possible close to departure, as business passengers are most likely to book flights late and are not as price sensitive. However, if too many high fare seats are blocked and demand does not materialize, airlines use overbookings to avoid spill (fly empty seats) (see Swan, 2002). When overbooking the airlines seek to maintain a balance where the gains from overbookings are larger than the costs and penalties incurred due to passenger rights regulation existing in some countries.

Just like full service carriers, low-cost carriers also operate yield management systems, but these tend to be simpler and with some notable strategic differences (see Dunleavy and Westermann, 2005). For example, low-cost carriers may try to capture last minute price sensitive customers from the full service carriers by offering much lower last minute fares. Such carriers also offer one-way fares that are not much different from the round trip fares, that is, the roundtrip fare is simply a combination of two one-way fares. Whereas full service carriers, tend to charge much higher one-way fares as proportion of the round trip fare. On the other hand, low-cost carriers do not offer multi-segment fares, whereas full service carriers frequently offer attractive multi-segment fares.

Organization and Human Resources

Human Resources

Airlines are labor intensive and there is a strong linear relationship between the number of aircraft operated and the number of employees. Despite this, average salaries per employee can vary greatly between airlines, even in the same country. What is more, productivity per employee can also vary substantially, explaining dramatic differences between airline cost structures and performance. Some low-cost airlines may have similar average salaries, but large differences in productivity, thus remaining low-cost by making much better use of human resources, despite higher salaries. One element that contributes to such a difference is simplicity of operations, by only performing activities that are necessary to offer the service features the passenger values the most. It is for this reason that low-cost airlines avoid offering business class seats, connecting flights, long-haul flights, and to operate multitype aircraft fleets.

Organization

Full service airlines are organized and structured to offer a full line of services and to innovate to enhance customer value at a price premium. The low-cost airlines, however, organize to keep costs in check and innovate to reduce costs, yet they strive to command high price-value perception on the service features offered. Despite low-cost airlines focusing so much on costs, they increasingly innovate to generate ancillary revenues, through unbundling of the product and acting as intermediaries for associated services such as insurance, car rentals, hotels and even train or bus tickets to and from airports.

External Environment

The key external variables influencing airline management are regulation, the economy, political climate, taxation, and societal changes. Airlines also face extraordinary events that are difficult to foresee in terms of timing and impact. Examples of such global impact events are terrorism such as 9/11 and the 2008 Credit Crises. There are also external events that are specific to the airline

industry such as long delivery delays of new aircraft (e.g., A380 and B787) as well as grounding of an entire fleet of aircraft types for safety reasons (e.g., DC10 and B737max) that can cause major financial hardship for the airlines and customer relations issues due to schedule disruptions.

Regulation

The State of California, as early as 1946 promoted the liberalization of its air transport intrastate market by allowing free entry and price competition, leading to the entry of 18 airlines in the period between 1946 and 1965 (Entry was free until the Cal. Stat. Act of June 17, 1965, ch. 736, 1965), many with different business models. The State of Texas followed by allowing price competition in its intrastate market. Several efficient airlines emerged from the pre-deregulation period such as PSA (1949–88), Air California (founded 1967–87, merged with American Airlines) and Southwest Airlines (founded 1967). The two States, California and Texas, therefore became the showcase for the impact of liberalized air transport on fares and used as examples by the proponents of full deregulation (Levine, 1965; Jordan, 1970). There is no doubt that liberalization in these two states provided consumers with the lowest airfares in the world, stimulating a major increase in passenger volumes. Liberalization, according to the proponents, increased consumer welfare, increased airline efficiency, and was in the public interest.

The 1978 Deregulation Act in the US domestic market spearheaded the liberalization trend in air transport. However, over the several decades since, other regulatory reforms in air transport have taken place: liberalized air services agreements in associated services such as ground handling, and regional liberalization spanning two or more countries, the so-called *open aviation areas*. Observing air transport deregulation in the United States, the EU followed suit through three packages of liberalization, implemented between 1987 and 1997 (Gudmundsson, 2011, 2019).

The best examples of liberalization in international air transport are the *Open Skies* air services agreements between the EU countries and the United States, the ground-handling liberalization in the EU is an example of liberalization of associated services, while the European Common Aviation Area (ECAA) and the Open Aviation Area (OAA) are examples of regional liberalization.

The Economy

Airline managers take into account potential fluctuations in the economy in order to manage the negative influence on the airline's profitability. There are two major influencing factors, oil prices and economic downturns. Although these two correlate, changes in fuel prices tend to have an immediate effect on the airlines finances, whereas downturns, causing reduction in demand, may involve some lag. The influence of the economic factors tends to differ on the two main strategic groups. In some cases, the low-cost airlines may benefit from economic downturns, as individuals and firms reduce costs by shifting to low fare airlines. At the same time, the leading low-cost airlines enjoy higher average margins, creating a buffer against cost increases and demand shifts. Full service carriers, as a strategic group, are more likely to experience losses during economic recessions because of thinner margins. Among these, full service carriers that are most resilient against economic shifts tend to be the most productive in terms of low input costs and high revenues per output unit or employee.

Political Climate

In the early 1970s the world economy was hit by a hike in oil prices and decline in consumer confidence. These events helped shift politics to consumer advocacy and away from the protectionist climate that had ruled many industries, including the airlines. It was in this backdrop that deregulation gained momentum in the US. Advocates of deregulation pointed out that the airlines had never been profitable (PBS, 2008) because regulation prevented price competition causing the airlines to fly half-empty planes in price sensitive markets. Even though the airlines were generally not keen on deregulation they suffered from overcapacity in the early 1970s that could be alleviated by stimulating extra demand through attractive discount fares. However, the CAB had stood in the way of such moves by the airlines. One of the key proponents of deregulation Alfred Kahn, stated that the government had to get out of the way of the market forces. Politicians at the time saw such change as a way to counter rising inflation by increasing fare competition. In other, words deregulating the airlines were seen as counter-inflationary as the US economy battled with the inflationary effects of the oil crisis (Gudmundsson, 2011).

The political event that rolled the ball in this direction was The Kennedy Hearings in 1975. The Civil Aeronautics Board (CAB) created the so called "route moratorium" by not granting any new routes since the late 1960s due to over capacity. Thus, within its mandate of maintaining industry stability, it did not grant new route rights unless public need was proven, rarely the case. The Kennedy Hearings, contrary to the CAB view, argued that the CAB's policies were preventing public access to affordable air fares. The view that emerged was that protectionism affected costs and prices because of lack of competition and innovation in the industry hindering its progression (Gudmundsson, 2011).

The CAB had stabilized the industry and market-shares of the major players, but it inhibited the ability of the airlines to benefit larger segments of the public through competitive prices. Society had changed and new expectations emerged that called for a more competitive industry. By deregulating pricing decisions, fares would approach the marginal costs at the same time that stronger incentives to improve resource utilization would push down costs. Thus, service, costs and price should be left to the market economy (Kahn, 1990). Although regulation was often seen as a cure to market failures (Levine, 1981) and instability (Keeler, 1984),

increasing evidence showed the opposite (Levine, 1965; Jordan, 1970). Thus, theories of deregulated markets were put to a test in a climate of political support and the airline industry was a high profile testbed for these ideas.

The EU followed suit, but much later and through a markedly different political process. The European Coal and Steel Community (est. 1951) was the first step in a long process to integrate Europe with the aim to stabilize the volatile political forces post WWII. The European Economic Community (EEC) was the second step, formally established through the Treaty of Rome in 1957 with the aim of dismantling the economic borders in Europe. The third step was to dismantle the political and social borders by establishing the European Union (EU) in 1993 through the Maastricht Treaty (sign. 1992).

Many European countries saw the model of big government and heavy control as ineffective in meeting changing public needs, especially when compared to Germany that was growing its industry at a faster phase than most other EU nations. In this climate the role of government was seen as hindering value creation in a more complex society where regulating pricing and supply of industries was questioned. Although the political climate and theories behind these changes went through reevaluation in the post-Thatcher-Reagan era, few argue for the comeback of economic reregulation of the air transport industry.

There are several important events that enabled air transport liberalization in the EU (Gudmundsson, 2011): (1) The *Nouvelles Frontières* case (Ministre Public v Asjes, Cases 209-13/84) that led to the 1985 decision of the European Court of Justice to rule that Article 81 was applicable, rendering approval of airfares by a Member State unlawful; (2) the 1985 Schengen Agreement (implemented in 1995); (3) the Commission's first White Paper on future development of common transport policy; (4) the Bosman ruling on free movement of people in 1995; and (5) the "open skies" judgments in 2002, stating that Member States cannot act in isolation when negotiating international air services agreements. All of these events, especially the *Nouvelles Frontières* ruling, facilitated liberalization and growth of air transport in the EU by opening up markets, enable new entry low-cost airlines, and enable free flow of people between the member states.

The European Union is the amalgamation of very different countries with different legal systems and law, making integration through economic deregulation the most attractive approach to unification. Although air transport was exempt from Article 81 of the Treaty, it entered the jurisdiction of the European Commission in the 1980s, after the ruling in the *Nouvelle Frontières* case. The ruling of the European Court of Justice enabled the Commission to attack price fixing in air transport, to uphold the principle of free movement of people, and apply the anti-trust law to the industry. Consequently the governmental approval of IATA tariffs became anticompetitive and in conflict with both Articles 10 and 81 of the Treaty (Gudmundsson, 2011). These developments forced, the Member States to embark on gradual liberalization of air transport between 1987 and 1997, in three steps.

Taxation

Article 15 of the Chicago Convention states that no fees, dues or other charges shall be imposed by any contracting State in respect solely of the right of transit over or entry into or exit from its territory of any aircraft of a contracting State or persons or property thereon (Abeyratne, 1993).

There are two aspects to taxes in air transport, state taxes levied by the authorities on airlines, and taxes that the airlines add directly to fares. Airlines in some cases advertised fares that excluded taxes and fees levied on top of the basic fare. To address this issue the US DoT regulated airline advertising so that taxes listed separately are included in the advertisement on fares.

The tax situation on domestic flights is relatively straightforward. In the US an "excise" tax of 7.5% is levied on all domestic tickets and this tax must be included in advertised fares. However, in the US airlines exclude other government charges from published fares. On top of this airports can impose passenger facility charges (PFCs) up to a maximum of four different fees on a round-trip ticket.

On international tickets the US government levies a departure fee, an arrival fee on international tickets, an immigration and customs fee, as well as an animal and plant inspections fee. All of these fees are collected when tickets are sold.

In Europe a regulation was issued in 2008 laying down rules to improve transparency about taxes and charges disclosed in published ticket prices. Consequently, airlines are obliged to disclose final prices in advertising and in the reservations process.

State-aid is defined according to article 87 of the Treaty, as state grants, interest relief, tax relief (or relief of airport charges), state guarantee or holding, and provision by the state of goods and services on preferential terms. It was in the background of this article, that the European Commission, in 2004, decided that Ryanair had received concessions on airport taxes and fees, over fifteen years, in return for a certain throughput of passengers at Charleroi (Gröteke and Kerber, 2004). The arrangement was new in the sense that a low-cost airline and a regional airport worked together to mount effective competition with larger airports and airlines. In other words, the collaboration provided public benefit by offering competing services in an airport that was less convenient than the main airport. The case went in favour of Ryanair, although the European Court of Justice ruled that such offers had to be available to all airlines that served the airport.

Value Added Tax (VAT)

Airline tickets are generally exempt from VAT with the exception of domestic aviation in some countries (Keen and Strand, 2007). Aircraft are also VAT exempt in Europe if used by airlines on international routes. There is increasing pressure that air transport should incorporate negative externalities by eliminating the VAT exemptions of the industry (Graham and Shaw, 2008).

Fuel Tax

Aviation is generally exempt from fuel taxes in bilateral air services agreements since the expanding post-WWII years. Although a common practice in the industry it is a nonbinding resolution of the ICAO to include fuel tax exemptions in air services agreements.

Under the 2003 Energy Taxation Directive most fuels and energy products in the EU are subject to tax. However, the Directive permits member states the discretion to exempt aviation fuel. Most EU member states do not tax aviation fuel with the exception of the Netherlands, Norway and Switzerland that tax domestic aviation fuel.

State or local taxes on aviation fuel in the United States go to airports and are subject to revenue-use requirements. These can be used for the operating costs of the airport, the local airport system, or any other local facilities owned or operated by the airport. In addition to this, state taxes on fuel may be used to support state aviation programs. However, state sales taxes are not collected on airline tickets.

Environmental Taxes

According to ICAO Assembly Resolution A39-3 an environmental global market based measure (MBM) is one measure to achieve carbon-neutral growth for aviation from 2020 onwards. A global MBM scheme, the so called Carbon Offsetting and Reduction Scheme for International Aviation (CORSIA), will address CO₂ emissions from international civil aviation that exceeds the 2020 levels. The CO₂ emissions from international aviation between 2019 and 2020 constitute the baseline that emissions in future years will be compared against. If in any year from 2021 international aviation CO₂ emissions exceed the baseline the difference must be offset for that year. However, there are several implementation phases of CORSIA (see [Scheelhaase et al., 2018](#)), and the first phase is voluntary, and some states are exempted from offsetting requirements: (1) pilot phase (from 2021 to 2023); first phase (from 2024 to 2026), and second phase (from 2027 to 2035). The second phase applies to all states that have international aviation output above 0.5% of total revenue tone kilometers (RTK) in 2018. Exemptions are least developed countries (LDCs), small island developing states (SIDS) and landlocked developing countries (LLDCs). However, any country can volunteer to participate in this phase.

An international route covered by the scheme must involve participating nations, otherwise it is not covered. Once participation of states and routes covered by the CORSIA is defined in a given year, and offsetting requirements (i.e., increased emissions beyond the average baseline emissions of 2019 and 2020) are set, the requirements are distributed among aircraft operators participating, according to a formula in the ICAO Assembly Resolution.

Other national schemes worth mentioning are environmental taxes that have been introduced in Sweden and Germany that levy a flat fee on every ticket sold with an aim to reduce the impact of air travel on the climate. The assumptions behind such taxes is twofold: (1) to reduce air travel by making it more expensive, a controversial approach as it affects the lower income groups in society proportionally more than higher income groups ([Shaw and Thomas, 2006](#)); and (2) to generate revenues that are used in state projects or to make tax refunds to individuals or firms that reduce their environmental footprint (electric cars, housing isolation, solar heating, sustainable energy, etc.) (see [Krenek and Schratzenstaller, 2016](#)). In Germany the green tax is distance related so tickets of more than 6000 km incur an extra charge. France has followed suit and a green tax on all tickets for outbound flights will take effect in 2020, to fund environmentally friendly transportation.

Societal Changes

Economic Wellbeing

Countries such as India and China have the benefit of rapid air transport growth over a long period reflecting policy changes that increase competition and lower air fares in combination with strong economic growth ([Wang et al., 2018](#)). Economic growth and market liberalization has impact on air transport demand in two ways: (1) increase in disposable income; and (2) the reduction in average air fares, making air travel affordable for more people. In the early days of air travel it was a transport mode for the elites, but as more and more capacity was added through larger aircraft, the airlines reduced air fares to fill more seats. In liberalized air transport markets, especially the US, EU and India, low-cost airlines have had a major impact in making air travel more affordable.

Price Elasticity of Demand

One other possible use of price elasticities of demand is for policy makers to predict policy effectiveness when formulating tax and environmental policies concerning aviation. Whereas, in past decades stimulating demand through increased competition and lower fares was seen as increasing social welfare, recently, environmental concerns and congestion has raised the policy makers' interest in demand reducing measures. Understanding the price elasticity of demand helps authorities to assess if air transport environmental taxes will reduce demand or simply increase the state's revenue from the industry ([Brons et al., 2002](#)).

Leisure

With increased consumer disposable income and travel experience, leisure travel has changed over time. For example in Europe in the 1970s and 1980s, a large proportion of the leisure traffic was between Northern and Southern Europe, especially Spain, Italy and Greece. This market then matured and in the 1990s and 2000s many charter carriers serving this market either went out of business or reoriented to meet their customers' wants for long-haul services to more exotic locations. Leisure travelers also took advantage of more attractive fares on the scheduled carriers that offered attractive package deals on city breaks all over the world. In this way the leisure market has become more diverse as travelers have become more accustomed to different cultures and more affordable air fares, more variety in hotel offers and higher quality tourism services around the world.

Migration

Air transport plays an important role in emigration by moving countries closer in time enabling people to keep contact with their origin and enjoy the benefits of living and working in another country. The origins of emigration vary, ranging from economic, social, political and conflicts reasons. There is also increased flow of knowledge workers that seek work where their competences are most valuable, no matter where in the world. Emigration also stimulates trade between an adopted country and country of origin. In this way emigration and related trade have major impact on air transport. In Germany for instance there are 2.8 million people of Turkish origin (Destatis, 2018), this is reflected in the fact that Germany is the number one destination of Turkish Airlines (Statista, 2019). Similarly, over two million Indian migrants live in the United Arab Emirates, an estimated 27% of the UAE population. Emirates Airlines of Dubai, serves a total of 9 Indian destinations with 170 weekly flights. What is more, the Emirates Group employs almost 14 thousand Indian nationals mostly in high-skill jobs, counting for 21% of its total workforce around the world (Emirates, 2019).

Education

Education has become highly internationalized with rising proportion of students in many countries being educated at distant national and foreign universities. This segment travels within own country or to and from country of origin. Students with more disposable income travel not only for their studies but also for study breaks, and also attract VFR traffic, as friends and relatives visit in vacation periods and attend graduations. A proportion of these students emigrates to the destination country and join the labor force as high-skill knowledge workers and reinforce in that way the VFR connection with the country of origin. As many as 1.1 million international students attend colleges and universities in the United States, and 334 thousand US citizens study abroad (IIE, 2019). In this way, air transport plays a role in improving accessibility and attracting high-skill labor force from other regions and countries. Academic institutions and air transport accessibility, in combination, play a crucial role in maintaining the competitiveness of regions and cities (Varga, 2000). In fact, research has shown that the air transport network stimulates more university location alternatives for students over longer distances and boosts regional human capital (Cattaneo et al., 2015).

Conclusion

Airlines are probably the most difficult business to manage, demonstrated in its challenging financial performance in comparison with most other industries. Airline managers must deal with many constraints and understand the industry practices and strategies that do not transfer readily from conventional economic theory or other industries. Due to these constraints, the job of an airline manager is probably one of the most difficult jobs to master. Warren Buffet, the investor stated once that despite all the investment in the airline industry the return in the long term is less than zero, thus, it is the most difficult industry to be in. Airline managers cannot rely on prescriptive rules of thumb on how to run their business but must adopt a dynamic and most importantly a deep understanding of the principles that guide effective strategic decisions in the airlines.

Airline managers must keep costs in check, no matter what the generic strategy of the airline, by assuring good fit among activities performed within the existing business model. In other words, an airline can undertake any activity or service feature so far as it does not cause major deterioration in fit. For example, fast turnaround of aircraft in airports helps increase aircraft utilization that in turn means fewer aircraft to serve the network, fewer pilots and therefore lower financing and crew costs. Thus, to maintain high fit among activities is the key to lowering costs and increasing productivity of the resources employed in an airline.

This being said the airline industry is one of the most exciting industries to work in and popular among young people looking for high profile entry level jobs with excellent career development opportunities.

References

- Abeyaratne, R.I., 1993. Air transport tax and its consequences on tourism. *Ann. Tour. Res.* 20 (3), 450–460.
- Bennett, R.D., Craun, J.M., 1993. The airline deregulation evolution continues: the southwest effect. Office of Aviation Analysis, US Department of Transportation.
- Brons, M., Pels, E., Nijkamp, P., Rietveld, P., 2002. Price elasticities of demand for passenger air travel: a meta-analysis. *J. Air Transport Manage.* 8 (3), 165–175.
- Calzada, J., Fageda, X., 2019. Route expansion in the European air transport market. *Regional Stud.* 53 (8), 1149–1160.
- Castelli, L., Ukovich, W., & Pesenti, R. (2003). An airline-based multilevel analysis of airfare elasticity for passenger demand. *The Conference Proceedings of the 2003 Air Transport Research Society (ATRS) World Conference, Volume 2*; UNOAI-03-6-Vol-2. Available from: <https://ntrs.nasa.gov/search.jsp?R=20050147587>.
- Cattaneo, M., Malighetti, P., Pleari, S., Redondi, R., 2015. Evolution of long distance students' mobility, the role of transport infrastructures in Italy. In: 55th ERSA Congress. World Renaissance, Changing roles for people, places.
- Destatis, 2018. Bevölkerung und Erwerbstätigkeit Bevölkerung mit Migrationshintergrund: Ergebnisse des Mikrozensus 2017. Statistisches Bundesamt, August.
- Dunleavy, H., Westermann, D., 2005. Future of revenue management: future of airline revenue management. *J. Rev. Pricing Manage.* 3 (4), 380–383.
- Emirates, 2019. Emirates to expand reach in India with SpiceJet codeshare partnership. Media Center. Available from: <https://www.emirates.com/media-centre/emirates-to-expand-reach-in-india-with-spicejet-codeshare-partnership>.
- Graham, B., Shaw, J., 2008. Low-cost airlines in Europe: reconciling liberalization and sustainability. *Geoforum* 39 (3), 1439–1451.
- Gröteke, F., Kerber, W., 2004. The case of Ryanair-EU state aid policy on the wrong runway. *Ordo* 55 (1), 313–332.
- Gudmundsson, S.V., 1998. Flying too close to the sun: the success and failure of the new entrant airlines. Ashgate Pub, Aldershot.
- Gudmundsson, S.V., 2011. Air transport liberalization. In: Finger, M., Künneke, R. (Eds.), *International Handbook on the Liberalization of Infrastructures*. Edward Elgar Pub., Cheltenham, UK.
- Gudmundsson, S.V., 2019. European air transport regulation: achievements and future challenges. In: Cullinane, K. (Ed.), *Airline Economics in Europe*. Emerald Pub., Bingley, UK.

- Gudmundsson, S.V., Paleari, S., Redondi, R., 2014. Spillover effects of the development constraints in London Heathrow. *J. Transport Geogr.* 35, 64–74.
- Hsu, C.I., Shih, H.H., 2008. Small-world network theory in the study of network connectivity and efficiency of complementary international airline alliances. *J. Air Transport Manage.* 14 (3), 123–129.
- IIE, 2019. Open Doors 2018. The Power of International Education. Available from: <https://www.iie.org/Research-and-Insights/Open-Doors/Data>.
- Jordan, W., 1970. *Airline regulation in America: effects and imperfections*. Johns Hopkins Press, Baltimore.
- Kahn, A.E., 1990. Deregulation: looking backward and looking forward. *Yale J. Reg.* 7, 325.
- Keeler, T.E., 1984. Theories of regulation and the deregulation movement. *Public Choice* 44 (1), 103–145.
- Keen, M., Strand, J., 2007. Indirect taxes on international aviation. *Fiscal Stud.* 28 (1), 1–41.
- Kohl, N., Larsen, A., Larsen, J., Ross, A., Tiourine, S., 2007. Airline disruption management—perspectives, experiences and outlook. *J. Air Transport Manage.* 13 (3), 149–162.
- Kracht, J., Wang, Y., 2010. Examining the tourism distribution channel: evolution and transformation. *Int. J. Contemp. Hosp. Manage.* 22 (5), 736–757.
- Krenek, A., & Schratzenstaller, M. (2016). Sustainability-oriented EU taxes: the example of a European carbon-based flight ticket tax. Working Paper. Umeå universitet, Umeå, p. 39. Available from: <http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A930270&dswid=4475>.
- Lee, S.G., Ma, Y.S., Thimm, G.L., Verstraeten, J., 2008. Product lifecycle management in aviation maintenance, repair and overhaul. *Comput. Ind.* 59 (2–3), 296–303.
- Levine, M.E., 1965. Is regulation necessary? California air transportation and national regulatory policy. *Yale Law J.* 74, 1416–1447.
- Levine M.E., 1981. Revisionism revised? Airline deregulation and the public interest. *LawContemp. Prob.* 44, 179–195. Available from: <https://scholarship.law.duke.edu/lcp/vol44/iss1/7>.
- Mascarenhas, B., Aaker, D.A., 1989. Mobility barriers and strategic groups. *Strateg. Manage. J.* 10 (5), 475–485.
- Mason, K.J., Morrison, W.G., 2008. Towards a means of consistently comparing airline business models with an application to the 'low-cost'airline sector. *Res. Transport. Econ.* 24 (1), 75–84.
- Oum, T., Tretheway, M., 1990. Airline hub and spoke system. *J. Transport. Res. Forum* 30, 380–393.
- Porter, M.E., 1980. *Competitive strategy*. Free Press, New York.
- Public Broadcasting Service (PBS), 2008. Alfred E. Kahn interview. Retrieved from <http://www.pbs.org/fmc/interviews/kahn.htm>.
- Redondi, R., Gudmundsson, S.V., 2016. Congestion spill effects of Heathrow and Frankfurt airports on connection traffic in European and Gulf hub airports. *Transport. Res. A Policy Pract.* 92, 287–297.
- Statista, 2019. Plane passengers in Germany in 2019, by country of origin or destination. Statista. Available from: <https://www.statista.com/statistics/590295/germany-number-plane-passengers-by-country-of-origin-or-destination/>.
- Scheelhaase, J., Maertens, S., Grimme, W., Jung, M., 2018. EU ETS versus CORSIA—a critical assessment of two approaches to limit air transport's CO2 emissions by market-based measures. *J. Air Transport Manage.* 67, 55–62.
- Shaw, S., Thomas, C., 2006. Discussion note: Social and cultural dimensions of air travel demand: Hyper-mobility in the UK? *J. Sustain. Tour.* 14 (2), 209–215.
- Smith, B.C., Leimkuhler, J.F., Darrow, R.M., 1992. Yield management at American airlines. *Interfaces* 22 (1), 8–31.
- Swan, W.M., 2002. Airline demand distributions: passenger revenue management and spill. *Transport. Res. E Log. Transport. Rev.* 38 (3–4), 253–263.
- Varga, A., 2000. Local academic knowledge spillovers and the concentration of economic activity. *J. Reg. Sci.* 40, 289–309.
- Wang, K., Zhang, A., Zhang, Y., 2018. Key determinants of airline pricing and air travel demand in China and India: policy, ownership, and LCC competition. *Transport Policy* 63, 80–89.
- Williams, M.J., Rhoades, D.L., 2017. Airline distribution systems: History, challenges and solutions. In: Ahmed, A. (Ed.), *World Sustainable Development Outlook 2007: Knowledge Management and Sustainable Development in the 21st Century*. Routledge, p. 163.

Further Reading

- Hares, A., Dickinson, J., Wilkes, K., 2010. Climate change and the air travel decisions of UK tourists. *J. Transport Geogr.* 18 (3), 466–473.
- Kahn, A., 1971. *The economics of regulation, II: Institutional Issues*. Wiley and Sons, New York.
- Barnhart, C., Belobaba, P., Odoni, A.R., 2003. Applications of operations research in the air transport industry. *Transport. Sci.* 37 (4), 368–391.
- Button, K., Taylor, S., 2000. International air transportation and economic development. *J. Air Transport Manage.* 6 (4), 209–222.
- Doganis, R., 2013. *Flying off course: the economics of international airlines*. Routledge.
- Fu, X., Oum, T.H., Zhang, A., 2010. Air transport liberalization and its impacts on airline competition and air passenger traffic. *Transport. J.* 49 (4), 24.
- Helmreich, R.L., Merritt, A.C., Wilhelm, J.A., 1999. The evolution of crew resource management training in commercial aviation. *Int. J. Aviat. Psychol.* 9 (1), 19–32.
- O'Connell, J.F., Williams, G., 2005. Passengers' perceptions of low-cost airlines and full service carriers: a case study involving Ryanair, Aer Lingus, Air Asia and Malaysia Airlines. *J. Air Transport Manage.* 11 (4), 259–272.
- Schipper, Y., Rietveld, P., Nijkamp, P., 2001. Environmental externalities in air transport markets. *J. Air Transport Manage.* 7 (3), 169–179.
- Shaw, S., 2016. *Airline marketing and management*. Routledge.

Air Cargo

Volodymyr Bilotkach, Singapore Institute of Technology, Singapore, Singapore

© 2021 Elsevier Ltd. All rights reserved.

Introduction	258
Specifics of Air Cargo	258
What Goods Travel by Air	259
Air Cargo Airline Business Models	259
Key Cargo Airports	261
Concluding Remarks	262
References	262
Further Reading	262

Introduction

Delivering mail has historically been one of the first uses of aviation. Air cargo business has developed in parallel to the commercial passenger aviation, and currently represents a crucial part of some global value chains. Air cargo is also an important part of international trade. The share of internationally traded goods traveling by air is estimated at about 35% by value. Furthermore, about 87% of all air cargo shipments worldwide are international rather than domestic.

Conventional metrics for air freight (I will use air cargo and air freight interchangeably in this article) are freight tons carried; and freight ton-kilometers (or ton-miles) performed. The latter is calculated simply as product of tons carried and kilometers flown. Thus, a flight carrying ten tons of cargo over 5000 km will generate 50,000 ton-kilometers of cargo. Interestingly, cargo is recorded and reported by the airlines separately from mail carried by the aircraft. Note also that in commercial aviation the notion of freight excludes passengers' checked luggage, even if the passengers have separately paid the airline for it. Only freight accompanied by air freight airway bill and air courier shipments are counted as air cargo.

According to the International Air Transport Association (IATA), in 2018 airlines worldwide have carried 63.3 million freight tons of cargo, collecting over \$110 billion in revenue for it (total airline revenues in 2018 are estimated at \$812 billion). This makes 2018 the first year in which air cargo revenues globally exceeded \$100 billion per year (IATA, 2019).

Air cargo flows generally reflect the global economic activity. Developed European/North American nations account for about half of the total global air freight, with Asia-Pacific region (including India and China) responsible for further 36% of this industry. As the economic activity is expected to grow faster in emerging than in mature markets over the next two decades; the shares of air cargo market are forecast to adjust accordingly. Airbus forecasts Asia-Pacific and European/North American segments to represent 42% and 45% of the global air cargo market, respectively, by 2035 (Airbus, 2018).

Specifics of Air Cargo

Air cargo business is different from the commercial passenger airline business, in a number of ways. Most importantly, perhaps, unlike passengers, cargo cannot be a significant source of nonaeronautical revenue for the airports. An airport can make money off passengers buying food, drinks, and duty free; parking their cars; and using other airport facilities. Few such opportunities exist in cargo business.

On the global scale, air cargo demand tends to be considerably more volatile than passenger demand. It is also not growing as fast. Looking at the comparison between the growth in passenger and air cargo demand from 2005, as reported by IATA, two things stand out. First, cumulatively passenger demand (as measured by revenue passenger miles, or the sum of the products of passengers and distance across all the flights performed globally over a year) have increased by about 75%. At the same time, air cargo demand only grew by 25%. Second, demand for air freight has been hit much harder during the Great Recession as compared to the passenger demand, and is generally subject to more severe fluctuations.

The next difference between air cargo and passenger businesses is in the extent of competition between the airports for cargo. We know that passengers can travel to a more remote airport if the fare to their destination is considerably lower and/or if a more remote airport offers a nonstop service that is not available from the nearby gateway (see Copenhagen Economics, 2012, for an extensive discussion on this issue). The extent of such competition between airports for the passengers is subject to some debate; however, we know that competition between the airports for air freight tends to be much more intense. Freight can travel hundreds if not thousands of miles by surface transport before being transported by air.

On the other hand, airports specializing in freight transport sometimes need to be located further away from the population centers. For instance, in Europe night time flight restrictions (mostly related to local noise pollution) mean that airlines wishing to handle cargo at night will need to choose more remote airports.

Last but not least, the nature of intermodal competition is different in air cargo market. Specifically, while for some products on some markets there may be no viable alternatives to air freight transportation (e.g., flowers grown in Africa for sale in Europe); in many cases intermodal competition for air cargo is present even on long-haul markets. This can be easily contrasted to the commercial passenger air travel, where intermodal competition is typically feasible only on short-haul routes; and on long-haul routes the only viable alternative to air travel is not taking the trip in the first place.

What Goods Travel by Air

Compared to surface transportation, the key advantage of air transport is of course time. It will take weeks for a shipment to reach Europe from China by sea; whereas shipping by air is a matter of days. According to the World Bank, air freight is priced at 4-5 times the cost of road transport (if we consider weight-based pricing). The price difference between air and sea freight can be even more dramatic, with air shipments being priced at up to 16 times higher than the equivalent weight shipped by sea.

Therefore, benefits from shipping by air rather than by sea have to be sufficient to justify the extra costs involved. Consequently, goods shipped by air tend to be high-value-added and/or perishable products. Air cargo shipments are also used by the large companies worldwide to enable their just-in-time supply chains, whereby components of complex products are delivered in smaller batches when they are required in production process. This kind of supply chain is an alternative to shipping a large order to the plant, where the components are then stored for some time.

In addition to the obvious example of perishable goods traveling long distance from the place of production to where they are sold (such as flowers grown in Africa traveling to Europe or asparagus shipped from Latin America to the United States); the following industries tend to be heavy users of air freight services.

- Consumer electronics, and other high-tech goods. Most likely, your smartphone or laptop has been flown from where it was manufactured to where you bought it. These products, in addition to being a classical example of high value-added goods, tend to be compact, and can therefore be shipped by air at a reasonable cost relative to their sale price.
- Pharmaceutical industry. This is yet another classical example of high value-added and mostly compact products (some of which are also perishable).
- Fashion industry. Certainly, some products here are sold at high margins. However, even lower margin (so-called "fast fashion") brands tend to ship their products by air rather than surface transport. The following factors play a role in this decision. First, air shipment allows the products to arrive in better condition, as air freight provides for better humidity and dust control while the goods are in transit. Second, air shipments allow the companies operating stores worldwide to better control their inventory (the same goods, produced in Bangladesh, can appear in both American, Japanese, Middle Eastern, and European stores at the same time), and quickly respond to changes in consumer demand.
- Auto and aircraft manufacturing industries. These industries are increasingly using complex just-in-time supply chains, enabled by air freight.

As an example, the German Federal Statistical Office reported that in 2017, 61% of electronic and optical goods (by value) exported from Germany were shipped to their destination countries by air. The corresponding figure for the pharmaceutical and chemical industry was 25%; and for the vehicle, aircraft, and ship manufacturing it was 14%.

Air Cargo Airline Business Models

The following key facts about the air cargo market are of note:

- On the global scale, about half of air cargo by volume is shipped by freighter aircraft. The other half of air freight travels in the belly of commercial passenger flights, generating additional revenue for the airlines.
- Air cargo market tends to be more fragmented than the passenger market—there are many more small-scale cargo airlines than small passenger carriers in operation.
- There are three air cargo business models currently in operation, broadly speaking.

According to Boeing's estimates, over the next 20 years or so we can expect the share of cargo carried by full freighter aircraft to decrease to about 37% from the current 50-50 split ([Boeing, 2019](#)). This of course does not mean that there will be fewer cargo planes in the airlines' fleet—both full freighter and belly cargo segments will be growing, but the latter will increase at a faster rate than the former. This expected trend can be explained by the following factors. First, the airlines' fleet will generally continue growing. Second, passenger carriers, facing steep competition and considerable aircraft ownership costs, will continue looking for ways to increase revenue from their aircraft, and belly cargo is exactly one of such ways. This will be facilitated by the modern information

technology, which will simplify the logistics related to the belly cargo. Third, continued liberalization of international aviation will make it easier for the airlines operating on international routes to adopt new business approaches.

The cargo aircraft fleet consists of purpose-built cargo planes, and the converted cargo aircraft. A number of passenger planes find their second careers as cargo aircraft. The largest purpose-built cargo aircraft currently in service is Antonov-225 Mriya (Ukrainian for "the Dream"). Designed by Antonov Design Bureau and built in 1985 in Kyiv, the single plane of this type is operated by the Ukrainian company Antonov Airlines. Mriya's maximum take-off weight is 640 tons, which is about 17% more than the same figure for Airbus A380—the largest commercial passenger jet built to date.

Both Airbus and Boeing manufacture purpose-built cargo planes, on their popular passenger platforms. One can say that iconic Boeing-747 aircraft found its new life in freight transport: while the manufacturer stopped making the passenger version of the plane, and the airlines are phasing them out of their fleet; the freighter version of the newest 747-8 plane is still in production. A number of combi passenger-cargo airplanes have also been built over the years. These planes feature partition of the cabin into passenger and cargo compartments. A number of air carriers have historically operated combi planes. At present time, however, these aircraft are not very popular, and many of those remaining in operation have been converted for either all-passenger or all-cargo use.

Airbus estimates that by 2035 global fleet of freighter aircraft will number about 3000, nearly double from the current figure. Interestingly, the manufacturer projects that about two-thirds of the new freighter planes (1860) will be conversions from the passenger fleet. The remaining third of new freighters (870 according to Airbus estimates) will be factory-built new planes.

Belly cargo transportation (where the freight is carried in the belly of a passenger jet operating its scheduled flight) has increasingly gained popularity in the age of higher oil prices and fiercer competition between the airlines. Many carriers realized the revenue potential of the space below the passenger deck, which otherwise remains empty on a flight that is operating anyway. This means that belly cargo can in principle be priced at the marginal cost—this cargo only needs to pay for the additional fuel and other costs, whereas in a dedicated freighter aircraft cargo must pay for the entire cost of the flight. With wide-body aircraft being produced at record rates, a lot of belly cargo capacity is added to the aviation system. For instance, Boeing's wide-body 777 aircraft, in its 777-300ER modification, has belly cargo capacity of 202 m³, or approximately six standard 20-foot containers used in sea shipping. By contract, cargo capacity of the Boeing-777 freighter is about 630 m³, or 19 standard 20-foot containers.

We can generally speak about three business models employed by the airlines carrying cargo (Morrell and Klein, 2018). First, traditional network carriers offer cargo transportation as one of their services, in addition to the passenger transportation. These airlines transport cargo both in the belly of their passenger flights, and by full freighter aircraft. Sometimes cargo is transferred at the airline's hub between passenger and full freighter flights. The "traditional" airlines differ in terms of how they incorporate air cargo into their business models. Some carriers have dedicated freight divisions or subsidiaries that operate a fleet of full freighter aircraft (e.g., Emirates Sky Cargo, or Martinair-cargo airline owned by Air France-KLM). Other airlines focus solely on belly cargo—for instance, the three major US network carriers (American, United, and Delta) do not have full freighters in their fleet, but they accept shipments for transportation in bellies of their scheduled passenger services. These shipments are handled by the airlines' cargo divisions. Many of the low-cost carriers, such as Ryanair-Europe's largest airline in this class—stay away from air cargo business altogether.

The second air cargo business model is the freight integrator company. In this category, we have such recognizable companies as FedEx, UPS, and DHL. As the term "integrator" suggests, these companies arrange the shipment from start to finish, using the assets that they own. When you ship something via, say, FedEx, your parcel is usually dropped off at a FedEx facility. Then it is loaded onto a FedEx truck, which delivers it to a FedEx sorting facility, from where it may be loaded onto a FedEx aircraft (and could be transferred from one FedEx aircraft to the next at one of the company's hubs). Once the shipment arrives to the FedEx sorting facility at the other end of the journey, it is then again loaded onto a FedEx truck, and the final delivery to the door is made by FedEx as well. On some "thinner" markets (i.e., more remote origins or destinations) the first and/or last steps might be performed by local delivery companies working on a contract with FedEx. Integrators tend to focus on the smaller parcel segment of the cargo industry, specializing in express delivery. These companies tend to be large, as scale here is required to achieve global coverage and efficiency of operations. In fact, FedEx and UPS are the largest and third largest air cargo carriers in the world, respectively, as measured by freight-kilometers flown (the second being Emirates Sky Cargo).

Last but not least, most of the cargo airlines follow what is known as the freight forwarder business model. These airlines own (or lease) their aircraft and fly cargo around the world; however, the freight comes to them from third parties (oftentimes transportation is arranged via agents, and forwarding airlines deal with the handling agents rather than directly with the customers shipping the freight). These airlines tend to be mid-sized or small, and only have full freighter aircraft in their fleet. The largest freight forwarded airline in the world by freight-kilometres flown is Luxembourg-based Cargolux, which flies cargo around the world using its fleet of over 20 Boeing 747-8 freighters. Cargolux is at the bottom of top 10 cargo carriers in the world, as measured by freight-kilometers flown. One peculiarity of business models these carriers tend to follow is that their routes and schedules are generally different from those of passenger and cargo network carriers. Rather than operating in a familiar hub-and-spoke pattern, whereby an aircraft leaves the airline's hub, reaches its destination, and returns to the hub; freight forwarders tend to operate multi-stop routes, which sometimes take the aircraft around the world before it returns to its home base.

Of the top ten cargo airlines in the world, two are integrators (FedEx and UPS, as noted above), one is a freight forwarder (Cargolux), and remaining seven are cargo divisions or subsidiaries of passenger network carriers. Interestingly, in the latter category six out of seven carriers are based in the Middle East or Far East, with Lufthansa Cargo being the only European cargo subsidiary of a passenger airline to make this top-ten list.

Table 1 Top-20 cargo and passenger airports worldwide, 2009-18

Airport	Position in top-20 list by cargo volume		Position in top-20 list by passenger traffic	
	2010	2018	2010	2018
Hong Kong	1	1	11	8
Memphis	2	2	-	-
Shanghai Pudong	3	3	20	9
Seoul Incheon	4	4	-	16
Anchorage	5	5	-	-
Dubai	8	6	13	3
Louisville, Kentucky	7	7	-	-
Taipei	14	8	-	-
Tokyo Narita	9	9	-	-
Los Angeles	13	10	6	4
Doha	-	11	-	-
Singapore Changi	11	12	18	19
Frankfurt	6	13	9	14
Paris Charles de Gaulle	7	14	7	10
Miami	12	15	-	-
Beijing Capital	16	16	2	2
Guangzhou	-	17	19	13
Chicago O'Hare	18	18	3	6
London Heathrow	15	19	4	7
Amsterdam Schiphol	17	20	15	11
Atlanta	-	-	1	1
Tokyo Haneda	-	-	5	5
New Delhi	-	-	-	12
Dallas-Ft Worth	-	-	8	15
Jakarta	-	-	16	18
Denver	-	-	10	20
New York JFK	19	-	14	-
Bangkok Suvarnabhumi	20	-	17	-
Istanbul Ataturk	-	-	-	17
Madrid	-	-	12	-

This list has been compiled based on the data from Airports Council International (ACI).

Naturally, the three business models described above represent "pure types" of ways cargo airlines arrange their business. In reality, many carriers operate more or less hybrid models.

Key Cargo Airports

If we compare the lists of top 20 passenger and cargo airports in 2018 by volume of traffic handled (number of passengers and tons of cargo, respectively), we will see the following discrepancies:

- Top five lists of airports by passenger and cargo traffic are entirely different. None of the top five passenger airports makes the list of top five air cargo gateways, and vice versa.
- Only three of the top 10 passenger airports are in the corresponding list of top cargo airports. These airports are Hong Kong (number eight by passenger traffic and top cargo airport), Shanghai Pudong (number nine by passenger traffic and number three by cargo volume), and Los Angeles (number four by passenger traffic, number ten by cargo).
- Three US airports in the top 10 cargo list (Memphis, Anchorage, and Louisville, Kentucky) are not even in the list of top-50 passenger airports worldwide.
- Of the top-20 cargo airports, ten are in the Middle East and Asia, six are in the USA, and four are in Europe.

Most of the top cargo airports gain their position by being either key regional gateways (Hong Kong, Seoul Incheon, Tokyo Narita, Miami), key hubs (Singapore, Dubai, Doha, Anchorage), or both (Frankfurt, Chicago O'Hare). Two largest US cargo airports (Memphis and Louisville) gained their position by being hubs of the largest cargo integrators (FedEx and UPS, respectively).

Interestingly, top cargo airports in different regions tend to be driven by different kinds of cargo airline business models. In the USA, three of the four cargo airports in the global top-20 are hubs of the integrator carriers. The top Middle Eastern cargo airports are

also hubs of the network carriers based in the region. The key cargo airports in China (as well as Hong Kong and Taipei) are the main gateways for the region's trade with the world.

Volatility of demand for air cargo is known to be higher than that for passenger travel. Despite this, the list of top-20 cargo airports has remained rather stable over the last 10 years. Out of 20 airports in this list in 2010, 18 are still in it in 2018. The same is true for the list of top-20 passenger airports. If anything, there has been more movement in the top-10 list of airports by passenger volume as compared to the same list for cargo airports. **Table 1** illustrates this point.

Concluding Remarks

Air cargo industry is a dynamic and complex business. In 2018, air cargo revenues have exceeded \$100 billion mark for the first time in history. This means that in revenue terms cargo constitutes about 12% of the airline industry globally. Air cargo business plays an indispensable role in enabling both international trade and the global supply chains. According to Boeing, about 35% of internationally traded goods by value travel by air.

Compared to the passenger airline business, the air cargo industry is more competitive and volatile. In addition to competition between the airlines, intermodal competition for air cargo is often present even on long-haul routes. Competition for air cargo between the airports is also said to be quite keen, as freight forwarders use surface transport to move cargo for hundreds or thousands of miles from the point of shipment origin before it is loaded onto an aircraft.

Air cargo is used to ship a variety of products. Most products shipped by air are high-value-added or perishable. However, some industries use air cargo to ship less than high-value-added products for logistics reasons, such as enabling just-in-time supply chains.

Air cargo is travels in both full freighters and the belly of passenger airliners. Currently, there is about 50-50 split between these two modes, but the share of belly cargo is projected to increase in the future as more wide-body aircraft capacity becomes available, fierce competition between passenger airlines incentivises the carriers to look for additional revenue sources, and information technology enables the relevant logistics. The COVID-19 crisis may change the structure of the air cargo industry. At the time of this writing, the future of the commercial aviation remains subject to much speculation. However, most analysts believe that the industry will eventually return to its growth path. Air cargo demand has remained robust throughout the crisis, with some airlines temporarily transforming their passenger aircraft for cargo use.

Intermediaries (cargo handling agents) play a more important role in air cargo business as compared to the passenger side of airline industry. In the air passenger business, travel agent's role is often limited to selling the tickets. Whereas cargo handling agents are closely involved with handling the goods being shipped. Freight integrator companies essentially integrate freight forwarding agents into their structure, dealing with shipments door-to-door. Unlike in the passenger business, where the airlines are trying to divert bookings from the agents to their own channel; the role of the agents in the cargo business will likely persist.

References

- Airbus, 2018. Global market forecast, Available from www.airbus.com.
 Boeing, 2019. Commercial market outlook, Available from www.boeing.com.
 Copenhagen Economics, 2012. Airport competition in Europe, Available from www.copenhageneconomics.com.
 IATA, 2019. Air cargo and e-commerce: enabling global trade, White Paper, March 2019, Available from <https://www.iata.org/whatwedo/cargo/stb/Documents/StB-Cargo-White-Paper-e-commerce.pdf>.
 Morrell, P.S., Klein, T., 2018. *Moving Boxes by Air: The Economics of International Air Cargo*. Routledge, London.

Further Reading

- Feng, B., Li, Y., Shen, Z.-J., 2015. Air cargo operations: Literature review and comparison with practices. *Transport. Res. Part C: Emerging Technol.* 56, 263–280.
 Kupfer, F., Meersman, H., Onghena, E., Van de Voorde, E., 2017. The underlying drivers and future development of air cargo. *J. Air Transport Manage.* 61, 6–14.
 Mayer, R., 2016. Airport classification based on cargo characteristics. *J. Transport Geogr.* 54, 53–65.
 Merkert, R., Van de Voorde, E., de Wit, J., 2017. Making or breaking-key success factors in the air cargo market. *J. Air Transport Manage.* 61, 1–5.
 Morrell, P.S., Klein, T., 2018. *Moving Boxes by Air: The Economics of International Air Cargo*. Routledge, London.
 Reis, V., Silva, J., 2016. Assessing the air cargo business models of combination airlines. *J. Air Transport Manage.* 57, 250–259.
 Sales, M., 2016. *Aviation Logistics*. KoganPage, London.
 Wensveen, J., 2011. *Air Transportation: a Management Perspective*. Routledge, London.
 Yuen, A., Zhang, A., Hui, Y.V., Leung, L.C., Michael Fung, 2017. Is developing air cargo airports in the hinterland the way of the future? *J. Air Transport Manage.* 61, 15–25.

Airline Regulation

Andrew R. Goetz, Department of Geography & the Environment, University of Denver, Denver, CO, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	263
Historical Background	263
Major Issues in Contemporary Airline Regulation	265
Aviation Security	265
Safety	266
Consumer Protection	266
Environmental Impacts	266
Health	267
Summary and Conclusion	267
References	267
Further Reading	268

Introduction

Air transport is a technological marvel that has experienced an astonishing rate of progress. From the Wright Brothers' initial heavier-than-air powered flight in 1903 to the introduction of jumbo jet aircraft in the 1960s and global networks of airliners in the 2000s, commercial air transport has come a long way in a relatively short period of time. An important part of the story behind this progress has been the role of commercial airline regulation which created an operating environment that allowed air transport to grow into a successful and indispensable component of global transportation and continues to facilitate its development.

This article will highlight the significance of airline regulation by first discussing its importance as part of the historical evolution of commercial air transport. Concerns with safety, national security, and economic stability were the principal justifications for instituting airline regulations historically, and continue to be important aspects of regulation today. Both safety and economic stability have improved greatly since the early years of commercial air transport, due to technological advancements but also to the role of regulation in mandating safer operations and in creation of a more stable industry. Yet both safety and economic stability remain important considerations, and underscore the need for continued vigilance from national and international regulatory agencies. Concerns with air transport security have grown over time due to the rise of terrorism, while other issues such as environmental impacts (especially greenhouse gas emissions), consumer protection, and health have been added more recently to the regulatory agenda. Recent developments in each of these areas will be discussed.

This paper is organized as follows: Section 1 provides some historical background to airline regulation while Section 2 discusses contemporary issues including aviation security, safety, consumer protection, environmental impacts, and health, followed by a brief summary and conclusions.

Historical Background

Commercial air transport began with the establishment of air mail and limited passenger services in the U.S. and several European countries in the late 1910s and 1920s. The expansion of aviation technology developed rapidly as a result of its application to military purposes during World War I. The U.S. Post Office began air mail service between Washington and New York in 1918, and the U.S. Contract Air Mail Act of 1925 (also known as the Kelly Act) transferred air mail and early passenger service to private sector companies by way of competitive bids. The first regularly scheduled daily international air passenger service had commenced in 1919 between London and Paris. This led to the need to establish some initial national security regulations for international air service, codified by the Paris Convention of 1919. The signatories to this agreement recognized not only the "complete and exclusive sovereignty" of states to the air space above their territories, but also the "freedom of innocent passage" through the airspace of other contracting states (Budd, 2014; Goetz, 2017) (see also articles 10425 The history of air transport, 10426 The geography of air transport, and 10428 Air transport and its territorial implications).

Because aeronautical technology was still in a developmental stage, early aviation was beset by numerous aircraft accidents, fatalities and injuries. As an example, pilots for the U.S. Post Office air mail service had an average life expectancy of only 4 years; 31 of the first 40 pilots were killed in action (Perrow, 1984). In response, the U.S. created safety standards and a regulatory framework through the Air Commerce Act of 1926 that vested jurisdiction over safety and the maintenance of navigation facilities and communication networks to the Aeronautics Branch of the Department of Commerce. It introduced initial regulations involving aircraft registrations, pilot certifications, and aircraft traffic safety rules. The US pioneered air navigation through the development of

lighted airways using high-intensity beacons to guide pilots along prescribed routes at night, as well as a national system of airfields. These and other supportive activities underscore a well-recognized principle that “ninety percent of aviation is on the ground” (Davies, 2011: 7). Similar regulations and safety measures were instituted in other countries with newly developing airline services.

Economic regulation of airlines became a more pressing concern with the onset of the Great Depression in the 1930s. Concerns about the economic viability of the fledgling private sector airline industry in the U.S. led to the Civil Aeronautics Act of 1938 (The Civil Aeronautics Act also transferred safety regulation from the Bureau of Air Commerce (formerly Aeronautics Branch) to the Civil Aeronautics Authority. In 1940, President Roosevelt split the Civil Aeronautics Authority into the Civil Aeronautics Administration (for air traffic control, pilot and aircraft certification, safety enforcement, and airway development) and the Civil Aeronautics Board (for economic regulation and some safety regulation, including accident investigations)) that vested the newly-created Civil Aeronautics Authority (later the Civil Aeronautics Board) with the power to control entry and exit of carriers (both interstate and foreign commerce) on new or existing routes, regulate fares, award subsidies, control mergers and intercarrier agreements, and investigate deceptive trade practices and unfair methods of competition. This type of economic regulation was similar to what had been established for the U.S. railroad industry through the creation of the Interstate Commerce Commission in 1887 to limit the monopoly power of railroads through regulation over pricing, entry, exit, mergers, and acquisitions (see also article 10078: Regulation and Competition in Railways). Economic and/or antitrust regulation was extended to airlines, motor carriers, telephone, banking, steel, petroleum, and electric power in the U.S. during the 1880s–1930s period because these industries possessed characteristics of public utilities subject to monopoly or oligopoly control and served as important “public interest” industries (Goetz, 2017). In a newly emerging industry, such as the airline industry of the 1930s, government regulation of private carriers was considered a necessity to avoid the “destructive competition” that would have been deleterious to economic stability and the longer-term viability of a key industry with national security implications. In other countries, airlines were already either state-owned or were awarded exclusive franchises as “chosen instruments” to develop air service as national “flag” carriers, so regulatory oversight was already embedded in the institutional structure of these industries.

Aviation technology advanced greatly as a result of military applications during World War II which laid the groundwork for a massive expansion in the postwar period. Anticipating the need to organize the postwar commercial aviation environment, representatives from 52 states met in Chicago in 1944 to create a new international regulatory framework. The Chicago Convention of 1944 recognized the so-called “five freedoms” of the air, thus leading to a framework which individual states could use to negotiate air service agreements. While the first two freedoms (right to fly over territory of another country and right to land for refueling and maintenance) were derived from the 1919 Paris Convention, the third and fourth freedoms refer to the right to transport passengers or cargo between the home country and a foreign country, while the fifth freedom refers to carrying passengers or freight between two foreign countries by a flight that originates in the home country. The United States was eager to pursue a more open multilateral regime that would allow all five freedoms for all signatory states. European nations, most notably the United Kingdom, preferred a more restrictive approach that would require negotiated bilateral agreements concerning the relative amount of air service provided between two countries based on the principle of reciprocity. The ensuing Bermuda I bilateral air service agreement in 1946 between the US and the UK included which city-pairs could be served, the specific airlines designated to provide service, as well as rules concerning capacity levels, fares, and freight rates, and became the basis for other international bilateral agreements (Bowen, 2010; Debbage, 2014). The International Air Transport Association (IATA) was established in 1945 as the fare-setting and global trade organization for the international airline industry, while the International Civil Aviation Organization (ICAO) was founded in 1947 as an agency of the United Nations principally concerned with rules and procedures of international air navigation. The Chicago Convention and the formation of IATA and ICAO set the stage for postwar commercial air transport expansion (Goetz, 2017).

Commercial air transport expanded greatly in the 1950s and 1960s with the onset of the “jet age.” Concerns over safety, however, continued as a result of a number of major air accidents. The first commercial passenger jet, the British-built de Havilland Comet 1, was entered into service in 1952, but was grounded in 1954 due to several high-profile accidents caused by the previously unknown problem of “metal fatigue.” A mid-air collision over the Grand Canyon in the U.S. in 1956 uncovered problems with communication and air traffic control systems. These and other accidents led to additional safety regulations and procedures, and in the U.S., the passage of the Federal Aviation Act of 1958 that increased funding for airline safety and transferred most safety responsibilities and air traffic control to a newly-created Federal Aviation Agency (later renamed the Federal Aviation Administration [FAA]). Authority over aircraft accident investigations was transferred from the CAB to the National Transportation Safety Board (NTSB), a new independent agency.

By the 1970s, several economic studies (Caves 1962; Cherrington 1958; Douglas and Miller 1974; Jordan 1970) had begun to question the desirability of continued economic regulation in the U.S. airline industry. It was felt that the airline industry had become more mature and competitively balanced since its early development in the 1930s, and that by removing regulatory control from the Civil Aeronautics Board (CAB), markets would operate more efficiently and carriers could offer consumers a wider range of price/service options. Competition would flourish and contribute to enhanced economic productivity. After more studies and much policy debate, the US promulgated the Airline Deregulation Act of 1978 which allowed US airlines to make decisions on entry, exit, and fares, and transferred authority over mergers to the US Department of Justice upon termination of the CAB in 1985 (Goetz, 2017) (see also article 10075: Privatization and deregulation of the airline industry).

At the same time, the US also adopted an “open skies” policy for international air service and, as part of a wider neoliberal policy agenda, encouraged other countries to follow suit. The European Union liberalized its air transport industry in a succession of three packages in 1987, 1990, and 1993 with full liberalization starting in 1997. Privatization and liberalization policies were adapted to

varying degrees in Asia, Latin America, the Middle East, and Africa. What ensued was a change in the market structure of the airline industry around the world whereby many state-owned airlines were privatized, many new entrants including low-cost carriers started airline service, mergers resulted in a smaller number of very large airlines, and global strategic alliances created mega-airline groupings.

It has been over 40 years since the US deregulated its airline industry, and studies have found US deregulation to be successful in lowering average fares, providing more flights, increasing carrier efficiency, and encouraging new entrant competition, while maintaining a good safety record (Transportation Research Board, 1991, 1999; US Department of Transportation, 2000; US Government Accountability Office, 2006; Morrison and Winston, 2008). On the negative side, many airlines have encountered wide swings in profitability, with losses being much larger than gains. This financial turbulence has resulted in increasing instability in industry structure and employment, while service quality has declined. Fares and service for smaller cities, shorter-haul routes, and more concentrated markets (where single-carrier domination persists) have also been negatively affected (Goetz and Vowles, 2009). Deregulation and liberalization have also led to increased concerns about labor relations and consumer protection as a result of increased power and control exercised by the airline companies (see also article 10606 Transport Policy and Planning: Air Transport).

Major Issues in Contemporary Airline Regulation

Aviation Security

Among the issues of most concern to airline regulators today is security. Concerns with aviation security date back to the 1960s and 1970s when airline hijackings became a problematic trend followed by an increasing number of terrorist incidents on airlines, including the infamous September 11, 2001 attacks in the US. Because of its high-profile role in the global economy and the symbolic importance of national flag carriers, air transport has become a target for terrorists thus necessitating increased security regulations and more technologically enhanced security systems.

Prior to the 1960s, aircraft hijackings and other criminal incidents were relatively rare, so security regulations were minimal. Hijackings increased in the 1960s involving persons attempting to escape from persecutors and/or to seek asylum, most notably on flights to and from Cuba. Responses to airline hijackings in the 1960s included regulations prohibiting weapons onboard and refusing to transport threatening passengers, as well as the introduction of trained air marshals on random flights. The Tokyo Convention, an international treaty signed in 1963 and effective from 1969, specified that the state in which an aircraft is registered has jurisdiction over crimes committed onboard, and that countries in which a hijacked airliner lands have an obligation to allow the crew and passengers to continue their journey and for the aircraft to be returned to its owner. The Hague Convention of 1970 enabled states to prosecute hijackers from other countries and to deprive them from asylum (Dempsey and Gesell, 1997).

Despite these efforts, civil aviation became an increasingly recognized target of international terrorism from the late 1960s onward. In 1968, terrorists from the Popular Front for the Liberation of Palestine hijacked an Israeli aircraft and held its passengers hostage, which led to more terrorist incidents (Enerstvedt 2017). In response, the Convention for the Suppression of Unlawful Acts against the Safety of Civil Aviation, also known as the Sabotage Convention or the Montreal Convention, was signed in 1971 and called for signatory states to prohibit and punish behavior such as acts of violence onboard an aircraft and destroying or damaging aircraft. Aided by reports and initiatives from ICAO in the 1970s, security at airports was increased through greater electronic surveillance, rudimentary passenger and baggage screening, and information sharing among national security agencies. The US FAA emphasized a “layered” system of defense including intelligence, pre-screening and checkpoint screening (contracted to private security companies), and onboard security.

By the 1980s, while the number of airline hijackings declined, terrorist bombings and sabotage started to increase, including Air India Flight 182 in 1985 which killed 329 people, and Pan Am Flight 103 in 1988 that resulted in 270 deaths. The number of aircraft hijackings increased again in the 1990s, most notably in China and East Asia, as well as in Europe. Additional security measures, such as manual searches, use of metal detectors, and passenger interviewing and profiling were implemented in many countries.

The September 11, 2001 (9/11) terrorist attacks in the United States involved hijacking of four airliners by 19 Al-Qaeda terrorists, and the use of these airliners as weapons to destroy the World Trade Center in New York City, and damage the Pentagon building in Washington DC. Altogether 2996 people died in these attacks, the deadliest aviation terrorist incident in history. The 9/11 attacks represent a critical watershed moment in national and aviation security when the threat level dramatically increased, and new regulations and countermeasures were enacted in the US and throughout the world. In the wake of 9/11, the US Transportation Security Administration (TSA) was created and given authority over all US aviation security, including enhanced passenger and baggage screening, federal air marshalls, explosive detection, and other airport and airline security activities. The TSA introduced numerous regulations and procedures, such as: (1) more stringent traveler identification requirements including no-fly lists and passenger personal data transfer, (2) more intrusive passenger and baggage screening involving explosive trace detection, X-ray and backscatter X-ray machines, CTX machines, canine units, camera surveillance, full-body scans and physical pat-downs, (3) restrictions for carry-on baggage including volume limits on liquids and gels and prohibitions on a range of other items that could be used as weapons, and (4) restricted access to gates and terminals only for ticketed and screened passengers and employees. ICAO adopted resolutions that called for locking cockpit and flight deck doors, sharing of threat information between national aviation security

agencies, and instituting a comprehensive aviation security audit program (Enerstvedt, 2017) (see also article 10104 Airport Security and article 10432 Airport Regulation).

While the number and severity of aviation terrorist incidents declined after 9/11 due in part to the enhanced security regulations and procedures, terrorist attacks have continued in different places and settings. In 2015, for example, an inflight explosion on a Russian passenger aircraft departing Egypt resulted in 224 deaths, and was attributed to ISIL terrorists.

Safety

During the early years of aviation as a newly developing technology, regulations were enacted to ensure safe operations of civilian aircraft. While the safety of commercial air transport has improved greatly since those early years, the need for safety regulation continues nevertheless as new technologies and equipment require vigilance on the part of ICAO and national aviation safety agencies. ICAO (2019d) considers safety to be at the core of their fundamental objectives. ICAO continues to develop and maintain standards, recommend practices and procedures for air navigation services, and develop global strategies to improve safety (see also article 10113 Aviation Safety–Commercial Airlines).

Two recent accidents involving the Boeing 737-MAX aircraft highlight the need for continued safety regulation. In October 2018, a Lion Air flight crashed shortly after takeoff resulting in 189 fatalities, and in March 2019, an Ethiopian Airlines flight also crashed shortly after takeoff, resulting in 157 deaths. In response, ICAO, the US FAA, and other national aviation safety agencies grounded the 737 MAX aircraft until further notice. The US Department of Transportation has requested an audit of the regulatory process that allowed the aircraft to be certified for airworthiness in 2017. As of this writing, the 737 MAX has not yet been cleared to resume flight operations.

The introduction of unmanned aerial vehicles (UAVs), also known as drones, presents additional aviation safety challenges. Drones were first used in military applications, but have been used increasingly in civilian applications, including aerial photography, data collection, delivery of goods, and by recreational hobbyists. The expanding use of drones has required aviation regulatory agencies to establish safety guidelines related to their use. As part of its Next Generation Air Transportation System (NextGen), the US FAA is developing regulations and procedures to accommodate the increased demand for drone use while ensuring safe operations of all aircraft. A special airworthiness certificate is required for drone operation, and an automated sense-and-avoid system will be required for drones to be certified as part of the NextGen system. There are also significant security and privacy issues associated with the introduction of drones (ALTawy and Youssef 2016).

Consumer Protection

In the wake of economic deregulation and liberalization across the airline industry, concerns about consumer protection have grown. As airlines have amassed more market power and operational control, some of their policies and practices have worked to the disadvantage of consumers. Issues such as overbooking flights and bumping passengers, tarmac delays and flight cancellations, false and misleading advertising, lost or damaged baggage, reductions in seat pitch and width, and transparency in pricing and ancillary fees have caused airline passengers to seek redress through enhanced consumer protection regulations, such as the proposed Airline Passengers' Bill of Rights in the US (Correia and Rouissi, 2015; US Congress, 2017).

Tarmac delays have been especially noteworthy. From May to September 2009 in the US, there were 535 instances of fully-loaded passenger aircraft delayed on airport tarmacs for more than three hours. In some extreme cases, passengers were kept on planes anywhere from 6 to 12 h without access to the airport terminal. In 2010, the US Department of Transportation issued a regulation that passenger-laden aircraft cannot idle on the tarmac for more than 3 h or the airline will be fined. This regulation resulted in a drastic reduction in long tarmac delays (Potter, 2017).

Environmental Impacts

As a mode of transport, aviation possesses many advantages, especially its ability to cross land or water at a higher speed of travel than any other commercial mode. But air transport also creates significant negative environmental impacts, especially in terms of noise, local air pollution, and global greenhouse gas emissions (see also article 10720 Environmental Issues of Airports and Aviation).

Noise has been recognized as a problem in commercial aviation since at least the 1960s. The US FAA issued noise standards in 1969, which were enhanced by the Noise Pollution and Abatement Act of 1970 and the Noise Control Act of 1972. Noise standards were amended in 1977 that created three stages of aircraft based on noise emission levels, together with plans for systematically phasing out the noisiest aircraft. The US Aviation Safety and Noise Abatement Act of 1979 provided assistance to airport operators to create noise compatibility programs and noise exposure maps for the purpose of reducing noise impacts. Regulations restricted noise-sensitive land uses, such as residences, schools, and hospitals, in airport areas (Ward et al., 2010). The US Airport Noise and Capacity Act of 1990 required that airlines phase out Stage 2 aircraft by 2000, while the European Union adopted a similar program phasing out noisy aircraft (Dempsey and Gesell, 1997). ICAO (2019c) and other national agencies have also created noise standards and developed programs to reduce noise exposure through reduction of noise at the source, land use planning and management, noise abatement operational procedures, and operating restrictions (see also article 10723 Noise Pollution from Transport and article 10731 Transport Noise and Health).

Because aircraft are powered by engines that utilize petroleum as their main fuel source, and because air transport activity continues to increase, air pollution emissions that directly affect human health are a growing concern. The principal pollutants that impact local areas near airports include nitrogen oxides (NO_x), carbon monoxide (CO), and non-methane volatile organic compounds (NMVOCs) (Upham et al., 2003). These and other pollutants were targeted in the U.S. Clean Air Act Amendments of 1970 through the establishment of National Ambient Air Quality Standards (NAAQS) to monitor and assess their emission levels. The US Environmental Protection Agency (EPA) and FAA established aircraft production standards to limit emissions of these pollutants (Dempsey and Gesell, 1997). ICAO also established engine emission certification requirements for these pollutants (Upham et al., 2003). Airport-related ground transportation contributes to local air pollution mainly through emissions from internal combustion engine motor vehicles. Many countries and cities have established standards and regulations that seek to control levels of pollutant emissions in airport areas (see also article 10712 Transportation Contribution to Air Quality).

Air transport is a contributor to global atmospheric emissions, most notably greenhouse gases such as carbon dioxide (CO₂). The Intergovernmental Panel on Climate Change (IPCC) has estimated that aviation emissions of CO₂ are 2–3% of total global greenhouse emissions, but that the amount of CO₂ emissions from aviation is expected to grow around 3–4% per year (ICAO 2019b). The Carbon Offsetting and Reduction Scheme for International Aviation (CORSIA), the EU Emissions Trading Scheme (EU ETS) for air transport, and others in New Zealand, South Korea, and China are aimed at limiting air transport's CO₂ emissions (Scheelhaase 2019). The Paris Agreement of 2016 as part of the United Nations Framework Convention on Climate Change (UNFCCC) calls for signatory countries to reduce GHG emissions so as to limit temperature increase to 1.5°C above pre-industrial levels. Many countries have made individual carbon reduction commitments, such as the UK with a target of 80% reduction from 1990 levels by 2050 (Palling et al., 2014). As a UN agency, ICAO (2019a) supports initiatives aimed at reducing GHG emissions from aviation including coordination of and assistance for state action planning on CO₂ emission reduction activities (see also article 10620 Climate Change and Transport Policy and article 10713 Transportation Contribution to Greenhouse Gases).

The IPCC (1999) has found that other aircraft emissions such as nitrogen oxides (NO_x), methane (CH₄), sulfur oxides (SO_x), water vapor, as well as contrail and cirrus cloud formation, also contribute to climate change especially in the upper atmosphere (Palling et al., 2014). Lee et al (2009) assessed aviation's CO₂ and non-CO₂ contribution to total radiative forcing to be 4.9%, with CO₂ accounting for 1.6% and non-CO₂ emissions at 3.3%. The EU Council is considering how to integrate regulation of non-CO₂ emissions into its existing EU-Emission Trading Scheme (Scheelhaase, 2019).

Health

Air transport has facilitated unprecedented levels of intercontinental travel, thus contributing to globalization and increasing economic and social interaction. While these trends have resulted in many positive effects, one concern has been increased risk of rapidly spreading communicable diseases and viruses that can affect human health. An outbreak of Severe Acute Respiratory Syndrome (SARS) in Asia in 2003 spread to other cities via air transport passengers, and this outbreak was implicated in declining levels of air passenger traffic that year. In 2009–10, an outbreak of the H1N1 flu virus led to the European Union health commission warning that nonessential travel to the US or Mexico should be postponed. More recently, outbreaks of the Ebola virus in Africa in 2014 and the Coronavirus (Covid-19) in China and throughout the world in 2020 have raised alarms for international air travel. Due to coronavirus-related restrictions on air travel starting in January 2020, worldwide air passenger travel declined precipitously resulting in over 90% reductions in year-over-year passenger totals during the first half of 2020. The World Health Organization (WHO) in coordination with ICAO has established regulations and procedures to limit the spread of communicable diseases including the provision of accurate and timely information to all persons involved in international air travel. Furthermore, ongoing health concerns related to the enclosed aircraft cabin environment, especially exposure to recirculated air on long flights, have prompted the WHO and international aviation agencies to issue guidelines for reducing risk of airborne illness transmission (see also article 10701 Sustainable Development Goals and Health and article 10708 Planetary Health).

Summary and Conclusion

Airline regulation continues to be an important activity of national and international aviation agencies to ensure well-functioning air transportation systems. While the historical focus of airline regulation has been on safety and economic stability, more recent challenges in aviation security, environmental impacts, consumer protection, and health have expanded the regulatory gaze to include these and other issues as well (see also article 10427 The future of air transport).

References

- AlTawy, R., Youssef, A.M., 2016. Security, Privacy, and Safety Aspects of Civilian Drones: A Survey. *ACM Trans. Cyber-Phys. Syst.* 1,2, Article 7: 1-25.
- Bowen, John., 2010. *The Economic Geography of Air Transportation: Space, time and the freedom of the sky*. Routledge, London and New York.
- Budd, L., 2014. The historical geographies of air transport. In: Goetz, A.R., Budd, L. (Eds.), *The Geographies of Air Transport*. Surrey. Ashgate, England.
- Caves, R.E., 1962. *Air Transport and Its Regulators: An Industry Study*. Harvard University Press, Cambridge, MA.
- Cherrington, P.W., 1958. *Airline Price Policy: A Study of Domestic Airline Passenger Fares*. Harvard University, Graduate School of Business Administration, Cambridge.

- Correia, V., Rouissi, N., 2015. Global, Regional and National Air Passenger Rights—Does the Patchwork Work? *Air & Space Law* 40 (2), 123–145.
- Davies, R.E.G., 2011. *Airlines of the Jet Age: A History*. Smithsonian Institution Scholarly Press, Washington, DC.
- Debbage Keith, 2014. The geopolitics of air transport. In: Goetz, A.R., Budd, L. (Eds.), *The Geographies of Air Transport*. Surrey. Ashgate, England.
- Dempsey, P.S., Gesell, L.E., 1997. *Air Transportation: Foundations for the 21st Century*. Coast Aire Publications, Chandler, AZ.
- Douglas, G.W., Miller III, J.C., 1974. *Economic Regulation of Domestic Air Transport: Theory and Policy*. The Brookings Institution, Washington, DC.
- Enerstvedt Olga Mironenko, 2017. *Aviation Security, Privacy, Data Protection and Other Human Rights: Technologies and Legal Principles*. Law, Governance and Technology Series. Springer, Cham, Switzerland.
- Goetz, Andrew R., 2017. Air Transport: Speed, Global Connectivity and Time-Space Convergence. In: Warf, B. (Ed.), *Handbook on Geographies of Technology*. Edward Elgar, pp. 211–230.
- Goetz, Andrew R., Vowles, Timothy M., 2009. The Good, the Bad, and the Ugly: 30 Years of U.S. Airline Deregulation. *J. Transport Geogr.* 17 (4), 251–263.
- ICAO [International Civil Aviation Organization]. 2019a. Climate Change: State Action Plans and Assistance. https://www.icao.int/environmental-protection/Pages/ClimateChange_ActionPlan.aspx. (Accessed July 27, 2019).
- ICAO [International Civil Aviation Organization]. 2019b. Environment: Aircraft Engine Emissions. <https://www.icao.int/environmental-protection/Pages/aircraft-engine-emissions.aspx> (Accessed July 27, 2019).
- ICAO [International Civil Aviation Organization]. 2019c. Environment: Noise. <https://www.icao.int/environmental-protection/Pages/noise.aspx> (Accessed July 26, 2019).
- ICAO [International Civil Aviation Organization]. 2019d. Safety. <https://www.icao.int/safety/Pages/default.aspx> (Accessed July 26, 2019).
- IPCC [Intergovernmental Panel on Climate Change]. 1999. *Aviation and the Global Atmosphere: A Special Report of IPCC Working Groups I and III*. Cambridge, UK: Cambridge University Press.
- Jordan, W., 1970. *Airline Regulation in America*. Johns Hopkins Press, Baltimore.
- Lee, D., Fahey, D., Forster, P., et al., 2009. Aviation and global climate change in the 21st century. *Atmos. Environ.* 43, 3520–3537.
- Morrison, S., Winston, C., 2008. The State of Airline Competition and Prospective Mergers. Statement for “Competition in the Airline Industry” Hearing before the Judiciary Committee Antitrust Task Force, United States House of Representatives, Washington, DC, April 24.
- Palling, C., Paul, Hooper., Callum, Thomas, 2014. The sustainability of air transport. In: Goetz, A.R., Budd, L. (Eds.), *The Geographies of Air Transport*. Ashgate, Farnham, UK, pp. 125–140.
- Perrow, C., 1984. *Normal Accidents: Living with High-Risk Technologies*. Basic Books, New York.
- Potter, Everett. 2017. America's Worst Tarmac Delays. *Travel and Leisure*, March 6, 2017. <https://www.travelandleisure.com/slideshows/worst-tarmac-delays?> (Accessed July 26, 2019).
- Scheelhaase, J.D., 2019. How to regulate aviation's full climate impacts as intended by the EU council from 2020 onwards. *J. Air Transport Manage.* 75, 68–74.
- Transportation Research Board, 1991. Special Report 230: *Winds of Change: Domestic Air Transport Since Deregulation*. National Research Council, Washington, DC.
- Transportation Research Board, 1999. Special Report 255: *Entry and Competition in the U.S. Airline Industry—Issues and Opportunities*. National Research Council, Washington, DC.
- U.S. Department of Transportation, 2000. *Changes in Pricing since Deregulation*. Office of Aviation Analysis, Washington, DC.
- U.S. Congress. 2017. S. 1418—Airline Passengers' Bill of Rights. <https://www.congress.gov/bill/115th-congress/senate-bill/1418/text> (Accessed July 26, 2019).
- U.S. Government Accountability Office. 2006. *Airline Deregulation: Reregulating the Airline Industry Would Likely Reverse Consumer Benefits and Not Save Airline Pensions*. GAO-06-630, Washington, DC, June.
- Upham Paul, Janet Maughan, David Raper, Callum Thomas, (Eds.), 2003. *Towards Sustainable Aviation*. Earthscan, London.
- Ward, Stephanie, A.D., Regan A. Massey, Adam E. Feldpauch, Zachary Puchacz, Christopher J. Duerksen, Erica Heller, Nicholas P. Miller, Robin C. Gardner, Geoffrey D. Gosling, Sharon Sarmiento, Richard W. Lee. 2010. *Enhancing Airport Land Use Compatibility, Volume 1: Land Use Fundamentals and Implementation Resources*. Airport Cooperative Research Program Report 27, Transportation Research Board. Washington, DC: National Academy of Sciences.

Further Reading

- Bowen, John., 2014. The economic geography of air transport. In: Goetz, A.R., Budd, L. (Eds.), *The Geographies of Air Transport*. Surrey. Ashgate, England.
- Haanappel, P.P.C., 2018. Air passenger rights in the electronic age. *Air & Space Law* 43 (1), 3–20.

Air Vehicles Classification

Vincenzo Torre, Center for Near Space, Italian Institute for the Future, Naples, Italy

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	269
Introduction and Definitions	269
Lighter-Than-Air Aircraft (Aerostats)	270
Heavier-Than-Air Aircraft (Aerodynes)	272
Air Vehicles (Aircraft) With Aerodynamic Lift Devices	273
Aircraft With Fixed-Wing Surfaces, With a Propulsion System	273
Aircraft With Fixed-Wing Surfaces, Without a Propulsion System	275
Aircraft With Mobile-Wing Surfaces (Rotary Wings)	276
Helicopter (Rotorcraft)	276
Autogyro	277
Air Vehicles Categories by Types and Roles	278
Civil Aircraft	278
Tourism (General Aviation) Aircraft	278
Aerobatic Aircraft	279
Ultralight Aircraft	279
Executive-Business Aircraft	280
Commercial Air Transport Aircraft (Airliners)	282
Cargo Air Transport Aircraft (Freighters and Outsize Cargo Freight Transporters)	283
Summary and Conclusions	288
Acknowledgments	288
See Also	288
Relevant Websites	288
References	289
Further Reading	289

Nomenclature

CAA Civil Aviation Authority
DGAC Direction General de l'Aviation Civile (French authority for Civil Aviation)
EASA European Aviation Safety Agency
EUROCONTROL European Intergovernmental Organization for Air Traffic Management
FAA Federal Aviation Administration
FAI Fédération Aéronautique Internationale
ICAO International Civil Aviation Organization
LSA Light Sport Aircraft
MTOW Max Take-Off Weight
NTSB National Transportation Safety Board
RA Recreational Aircraft
RPAS Remotely Piloted Aircraft Systems
STOL Short Take-Off and Landing
STOVL Short Take-Off and Vertical Landing
UAS Unmanned Air/Aircraft Systems
UV Ultralight Aircraft
VTOL Vertical Take-Off and Landing

Introduction and Definitions

Air vehicles—aircraft—are fundamentally conceived and built as transportation means, for civil and commercial use, for military and special applications, as well as experimental purposes. The aim of this article is to present and discuss the general, widely renowned, classification of aircraft largely employed for civil and/or commercial transportation, for both civil passengers and cargo.

Table 1 Air vehicles classification.

	Type of lift force	Lifting devices		Nomenclature
Aerostats (Lighter-than-air)	Static lift	Lifting envelopes with lighter-than-air gas	with propulsion system	Airships
			without propulsion system	Free balloons, kite balloons
Aerodynes (Heavier-than-air)	Aerodynamic lift	Fixed wing surfaces	with propulsion system	Aircraft, Seaplanes, Amphibious, UAS
			without propulsion system	Gliders, Tethered flying cerfs
		Mobile wing surfaces	Rotary wings (rotorcraft)	Helicopters, Autogyros, Quadrotors (quadcopters)
	Flapping wings		Ornithopters	
		Direct reaction (reaction engines)	Directed/steerable jet thrust (turbojets, rockets)	
	Mixed lift	Rotary/Fixed wing surfaces + directed lift jets		VTOL, STOL aircraft, Convertiplanes (Tiltrotor)

Air vehicles are transportation means able at moving through the airflow deriving their lift—the sustaining force in flight—by fundamentally two distinct sustaining phenomena: static lift or dynamic lift (Losito, 1982). Hence, all air vehicles, lighter-than-air and heavier-than-air, can be classified in the following two primary categories:

1. Aerostats: lighter-than-air aircraft with static sustaining force, where the lift is produced by Archimede's physical law (no relative motion body-airflow);
2. Aerodynes: heavier-than-air aircraft with dynamic sustaining force, where the lift is produced by means of the known Newton's law of dynamics (relative motion body-airflow).

Based on these two primary categories, a general, exhaustive classification of air vehicles is reported in **Table 1**. It has to be specified that, due to limitations in terms of space given for this editorial contribution, the categories of aircraft not purely associated with civil transportation, namely: military aircraft, Unmanned Air Systems (UAS), Remotely-Piloted Aircraft Systems (RPAS), drones, quadrotors, flapping-wing devices, direct-reaction aerodynes (missiles), and mixed-lift aircraft (VTOL/STOL, convertiplanes) reported in **Table 1** for completeness, will not be herein discussed. Notwithstanding, a complete list of additional sources in the section "Further reading" is provided for further investigations and details on these nomenclature and topics.

This paper is structured as follows: First section describes the first type of air vehicles, those which are lighter-than-air, called aerostats; Next section presents and describes air vehicles which are categorized as heavier-than-air, the larger category in the aeronautics nomenclature; Further section describes air vehicles conceived with aerodynamic lift devices; and the last section presents air vehicles categories classified by aircraft types and specific roles.

Lighter-Than-Air Aircraft (Aerostats)

Aerostats are classified into three main categories:

- Free balloons, having a spherical (hot-air balloon) or non-spherical shape, are not anchored to the ground so that they are free to move in the air and they are not equipped with a propulsion system. The uplift for a free balloon is given by hot air, provided by burners in a cabin attached with iron cables and karabiners to the balloon or by a lighter-than-air gas such as helium or hydrogen;
- Kite balloons, with a streamlined shaped and tethered with cables to the ground, without a propulsion system but with trimming mechanisms to derive stability from the air winds. Kites were destined primarily to military reconnaissance uses during the first and the Second World War;
- Airships, characterized as a lifting craft—the main sustaining device for lift—containing lighter-than-air gas (helium or hydrogen) that provides the adequate shape to the craft, with controls for steering and maneuvering the craft in flight. Airships are aerodynamically shaped, with a flexible, semi-rigid or rigid structure, and are equipped with a propulsion system and an underbelly for accommodating crew and passengers. In contrast to balloons, that can only follow the wind direction and can change altitude, airships have flight controls for maneuvering.

The balloons filled with hot air or gas (**Fig. 1**), receive an upward lift equal to the weight of the fluid which is equal to the volume of the balloon (Archimede's principle). The balloon can move upwards in the air when the lift received is greater than the weight of the balloon itself. The main difference between a hot-air balloon and one filled with gas is the means used to change altitude: the former utilizes burners to warm or cool the air in their shape, the latter uses to release ballast to climb and the venting of gas through



Figure 1 Hot air balloon at Holloman US Air Force Base. Source: <https://www.holloman.af.mil/>. Credits: Airman 1st Class Michael Means

appropriate valves for the descent. Modern free balloons are today used mainly for sightseeing tours, able to carry between 3 and 5 passengers, up to 10–12 passengers for larger sizes. Free balloons can fly at extremely high altitudes, up to 20,000 m.

Typically, hot-air or gas balloons for manned flight are made by a single-layered, fabric bag (lifting envelope), having an aperture at the bottom to release air or gas, with an attached basket or gondola for carrying passengers. Transport applications for balloons are for entertaining purposes, tourist rides on cultural and historical heritage sites, aerial photography, and sport events worldwide. Hot-air ballooning is recognized by the Fédération Aéronautique Internationale (FAI) as the safest flying sport in aviation, with rare fatalities occurred in flight, according to the NTSB.

The Rozier balloon (Fig. 2) is a hybrid balloon, deriving the name by the French scientist Pilatre de Rozier, combining helium and hot air in its envelope. For the take-off, the gas chamber is partially filled with helium, and as the craft is lifted, the diminishing



Figure 2 Rozier balloon Breitling "Orbiter 3." Source: https://en.wikipedia.org/wiki/Rozier_balloon#/media/File:Breitling_Orbiter_3_aloft.jpg. License: <https://creativecommons.org/licenses/by-sa/3.0/deed.en>.



Figure 3 Zeppelin Airship D-LZZF Baden-Württemberg shortly after the takeoff in Friedrichshafen, district Bodenseekreis, Baden-Württemberg, Germany. *Source:* [https://commons.wikimedia.org/wiki/Category:D-LZZF_\(aircraft\)#/media/File:Zeppelin_NT-DSC_9550.jpg](https://commons.wikimedia.org/wiki/Category:D-LZZF_(aircraft)#/media/File:Zeppelin_NT-DSC_9550.jpg). *License:* <https://creativecommons.org/licenses/by-sa/3.0/deed.en>. *Credits:* Dietrich Krieger.

atmospheric pressure makes the helium to expand. To stabilize at a certain altitude or descend, the helium is spilled through valves on top of the balloon. The Rozier balloons are technological sophisticated aerostats, employed in oceanic flights or circumnavigation of the Earth, and their operations are managed by a team of scientists, engineers, meteorologists, and technicians.

Airships are flyable thanks to a specific “torpedo”-shaped profile, made as a rubberized, fabric bag, inflated with lighter-than-air gas (hydrogen or helium), equipped with a propulsion system and flight controls for maneuvering. Historically, the first airship was built with a vapor engine in 1852, followed by an electrical propulsion airship and in 1898 by the first one with an internal combustion engine. In the same year Count Ferdinand von Zeppelin built its first Zeppelin LZ-1 airship in Southern Germany. The larger Zeppelin LZ-129 Hindenburg had a 245 m length, a crew of 40 flight officers plus 10–12 stewards, carrying max 72 passengers at a cruising speed of 125 km/h.

The use of hydrogen as an inflated gas was common since the beginning of the airship era, but due to its easy flammability it was abandoned after the Hindenburg airship disaster in 1937 followed by many other airship crashes. Helium gas, for its non-flammable properties, resulted the most practical lifting gas for manned airship operations and passengers transport.

A large utilization as a transport mean, at the beginning of the 20th century and at the onset of the second World War, made the airship the speediest air vehicle of the time over long transatlantic routes, able to be employed for commercial transport from Europe and the United States, offering passengers comfort and a luxury on-board service (Croccon, 1923).

The modern Zeppelin NT (New Technology) (Fig. 3), in operation since 1997, is a semi-rigid airship, with a high-strength multilayer laminate envelope, with structures made in carbon-fiber and aluminum, helium-inflated, driven by three propeller Textron Lycoming piston engines, capable to accommodate 14 passengers and 2 pilots, with a maximum speed of 125 km/h. Usually the airship flies at half this speed for sightseeing and tourism purposes.

Heavier-Than-Air Aircraft (Aerodynes)

Aerodynes are the heavier-than-air flying machines most commonly and widely known, categorized on the basis of the different way Newton’s law of dynamics is applied to generate the lift for the flight. Depending on the lift force generated for the flight, aerodynes are classified in three main categories:

1. Aerodynes with aerodynamic lift devices: where the lift is generated by the relative motion of a body through the airflow. The body-lift devices, moving in the oncoming airstream, encounter a certain amount of air masses due to the body-fluid relative motion at a certain speed. Because the body lift devices are characterized by appropriate geometrical configurations, capable to exert momentum to the airflow, air masses are accelerated downwards generating a field of pressure forces and a resultant upward lift on the body (Losito, 1982; von Karman, 2004).

Air vehicles lift devices are mainly platforms or wing surfaces, having several configuration types:

- a. Fixed wing surfaces (aircraft)
- b. Rotary wings, engine-powered (helicopters, quadcopters)
- c. Inclined platforms (stag beetle kites)
- d. Flapping wing surfaces (ornithopters)
- e. Self-rotary wings due to translational motion, with or without engine (autogyros)

2. Aerodynes with lift jet devices: where the lifting devices are mechanical systems (lift jets) accelerating air and gas with no body-airflow relative motion. Lift jets can be characterized as:
 - a. Cold jets released by ventilated, engine-actuated, fans
 - b. Hot jets released by turbojet engines or rockets
 - c. Mixed jets, released by turbojets which, through a turbine, provide the motion to ventilated fans accelerating cold air from outside downwards

Missiles are rocket-propelled vehicles but will not be treated in this article, because of their specific propulsion dynamics and utilizations and are not conceived as transportation means.

3. Aerodynes with mixed lift devices: using independently or concurrently aerodynamic lift as well as lift jets. VTOL/STOL air vehicles belong to this group.

In the general classification a major distinction has to be made between fixed-wing surfaces aerodynes (aircraft), which provide small accelerations to a large airflow mass, and rotary-wing aerodynes (rotorcraft), where the momentum variation is given to a small airflow mass, so that these flight vehicles are characterized by larger vertical accelerations.

As we have done for aerostats, we will follow the classification of [Table 1](#) for describing aerodynes and the corresponding nomenclature, and we will focus mainly on two major groups, with fixed-wing surfaces and with mobile (rotary) wing surfaces.

Air Vehicles (Aircraft) With Aerodynamic Lift Devices

As [Table 1](#) shows, these kinds of air vehicles—usually called *aircraft* in aeronautical terms—are classified in two main clusters:

- With fixed-wing surfaces, with or without a propulsion system
- With mobile-wing surfaces (rotary wings) with a propulsion system

Aircraft With Fixed-Wing Surfaces, With a Propulsion System

These types of aircraft—which is the largest group—are defined as manned air vehicles having a standard architecture characterized by ([Losito, 1982; Roskam, 2015](#)):

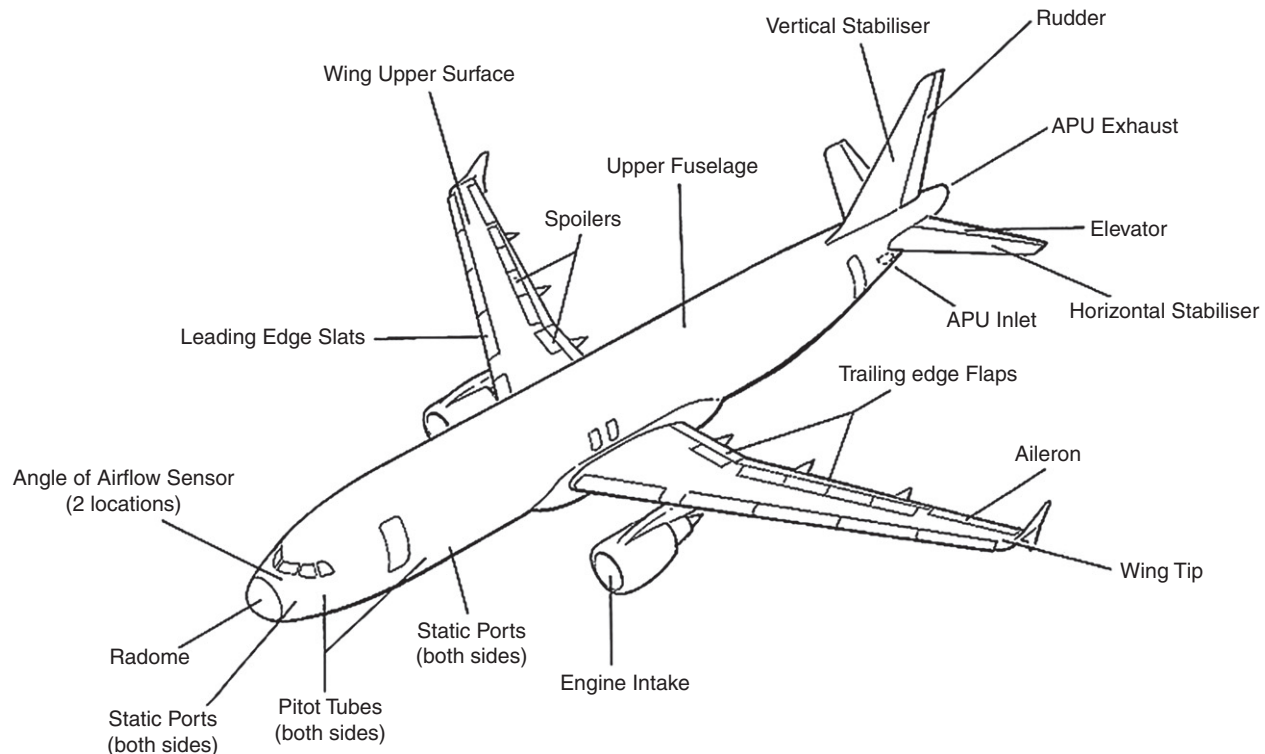


Figure 4 Aircraft general description and parts. Source: https://it.wikipedia.org/wiki/File:Aircraft_Parts. Credits: Dtom.



Figure 5 Canadair CL-415 C-GOX Ontario. Source: <https://commons.wikimedia.org/wiki/>. Credits: Saleen219



Figure 6 De Havilland DHC-6 Twin Otter floatplane. Source: <http://www.flickr.com/>. License: <https://creativecommons.org/licenses/by/2.0/deed.it>. Credits: Tony Hisgett, Birmingham, UK.

- the fixed-wing surfaces that generate the aerodynamic forces to fly;
- the nose cockpit for the crew and the fuselage, a central, fuse-shaped body, made in aluminum alloys and/or composite materials for mobile surfaces, having the role to transport payloads (passengers or cargo), acting as the main structural joint between the wing surfaces and the tail surfaces group (empennage); and
- the horizontal and vertical empennages, made by a fixed surface and a movable surface, or by a full movable horizontal and vertical surfaces (taileron), a very common solution for modern commercial and/or military aircraft, which are responsible for stabilizing the aircraft and make it highly maneuverable in flight.

The aircraft architecture (Fig. 4) is completed by some other movable surfaces, part of the wing—flaps, spoilers, ailerons, trims—as well as landing carriages and wheels and a propulsion system, being it a propeller-driven engine (turboprop), or a jet engine (turbojet or turbofan).

A standard aircraft is mainly operated on solid, prepared or unprepared, strips for take-off and landing operations. The Air Vehicle nomenclature in Table 1 also counts for other two special aircraft types equipped with propulsion systems: seaplanes and amphibians.

A seaplane aircraft is conceived for utilization on liquid surfaces, and it is capable of taking off and landing on water. Amphibian aircraft are a sub-class of seaplanes that can be operated on solid airfields as well as liquid surfaces (Fig. 5). Seaplanes and amphibians are further divided into two categories, float seaplanes and flying boats, the former are equipped with a pair of floats supporting the craft on water, the latter are larger aircraft with a capable hull, as its main bodies, that can carry more payload (Figs. 6 and 7).

Utilization is usually limited to aerial surveying over liquid surfaces (lakes and rivers), transport toward areas with limited access and airfields or for firefighting purposes. The development of seaplanes ended soon after 1940s, because of their poor aerodynamic and structural efficiency compared to land aircraft, and also due to the V-shaped configuration of floats and hulls that contributes to the increase in the body's aerodynamic drag (Vagianos and Thurston, 1970). Futuristic design proposals with a completely new aerodynamic configuration suggest that large seaplanes could be used for long-range airline operations, accommodating between 200 and 2000 passengers from new offshore airports (Levis and Serghides, 2014).



Figure 7 Grumman HU-16E Albatross. Source: U.S. Coast Guard photo <http://www.uscg.mil/history/im/>. Credits: USCG.



Figure 8 DG-1000 Glider. Source: <http://en.wikipedia.org/wiki/>. Image: [DG1000_glider.jpg](#). Credits: Paul Hailday.



Figure 9 PIK-20E, finnish glider. Used by NASA Dryden Flight Research Centre 1981-1991. Sailplane Research Aircraft. Source: <http://www1.dfrc.nasa.gov/gallery/photo/PIK-20/HTML/EC91-504-1.html>. Credits: NASA Dryden Flight Research Centre.

Aircraft With Fixed-Wing Surfaces, Without a Propulsion System

The most common and well-known manned aircraft without a propulsion system are gliders, or sailplanes. A glider is designed without engine and it is incapable of taking-off autonomously from an airfield, so it is towed by another engined aircraft to take-off, released at a certain altitude and from there it can only hover, using airstream currents (thermals) for climbing and fly appreciably long distances, and descend in gliding flight. Gliders have usually a small weight, wings with a high aspect ratio (more than 20–30) and a very streamlined aerodynamic shape for minimizing drag in flight (Fig. 8). Designed with a cockpit accommodating one or two pilots, at the most, gliders are generally employed for recreational flights or tourism. Also, the US National Aeronautics and Space Administration (NASA) utilize sailplanes for research purposes for investigating the airflow behavior over lifting surfaces (Fig. 9).

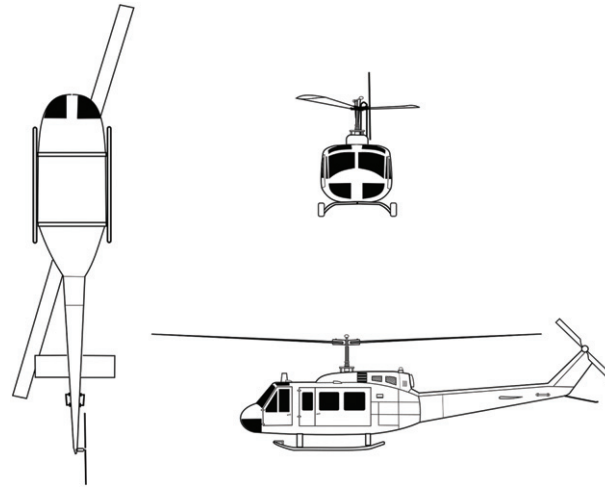


Figure 10 Helicopter general description and parts. Source: https://en.wikipedia.org/wiki/Bell_UH-1_Iroquois.



Figure 11 Boeing 234 - N239CH Commercial Helicopter. Source: https://commons.wikimedia.org/w/index.php?title=File:Boeing_234_-_N239CH.jpg&oldid=341136457. Credits: Tequask.

Aircraft With Mobile-Wing Surfaces (Rotary Wings)

Helicopter (Rotorcraft)

A helicopter is a manned flying vehicle with a main rotor (or rotors) mechanism with mobile aerodynamic surfaces (blades), and a smaller antitorque tail rotor, whose rotary motion develops the necessary aerodynamic forces for the lift, translation of the aircraft in the air and hovering (fixed point). The power the rotor needs for developing the drive, balancing the helicopter weight, enables to achieve vertical take-off and landing, ensures adequate evolutionary flight, in forward, backward, lateral directions and in hovering, providing the necessary thrust to overcome the resistance forces. The rotor is actuated by a turbine power plant. The first modern helicopter was the German Focke 61 that successfully flew in 1937, achieving a 3427 m altitude, flying at 122 km/h. A confirmation of these outstanding performances was the flight, years later, of the American V.S.3000, designed by Igor Sikorsky, which was the first of an interesting and long series of helicopters, reaching full-scale production in 1942, up to now.

The helicopter configuration consists of an aluminum alloy and/or composite material fuselage, with the nose cockpit for the crew separated by the passenger cabin, in most cases and for larger vehicles, or a whole smaller cabin with the seats for the pilots, a single 2–6 blades rotor actuated by the turbine engine, placed along the axes of the center of gravity of the machine. The helicopter tail, joint with a boom to the cabin, is equipped with a small antitorque rotor with few smaller blades, which enables the directional control of the vehicle (Fig. 10). Some helicopters of larger size and weight for heavy transport and cargo can have a tandem counter-rotating rotor, one mounted behind the other (Fig. 11).

Helicopters of different size, small or larger, are not comparable with fixed-wing aircraft, in terms of speed, altitude and performances, but due to their ability to take-off and landing very quickly from very small airfields or congested city centers, or even from platforms on top of residential buildings, they are employed in a variety of civilian roles, such as search-and-rescue, air ambulance for emergency medical assistance, aerial surveying, tourism, passengers, and cargo transport (Figs. 12 and 13).



Figure 12 Coast Guard Air Station Cape Cod's MH-60 Jayhawk aircrew prepares to take off in Massachusetts, Jan. 19, 2017. Source: <https://coastguard.dodlive.mil/2017/02/below-zero-new-england-aircrews-take-on-frigid-winter-weather/>. Credits: U.S. Coast Guard photo by Petty Officer 3rd Class Nicole J. Groll.



Figure 13 EUROCOPTER EG135P2, EC-KDA/05338 Generalitat de Catalunya, Dept de Salut. Source: Credits: Brian G. Nichols.



Figure 14 EC-MCZ, AG8A5246 ELA Aviacion Auto-Gyro. Source: Credits: Brian G. Nichols.

Autogyro

An autogyro is the simplest *gyrodine* with a single rotor and flapping hinges, but with conventional aircraft controls for roll, yaw and pitch. The first autogyro prototype was designed by the Spanish Juan Codornio de la Cierva in 1923, and achieved success and acceptance, followed by other innovations. An autogyro is capable to sustain itself in flight at very modest speed and can also land exactly on a vertical axis, using a very limited surface to take-off because the rotor with flapping hinges can develop high air masses at low speed. Driving the rotor before take-off was achieved by a coiled rope system activating the rotor at a speed to 50% of that required. With the C.19 Mk.4 model a direct drive from the engine to the rotor was fitted, so the rotor could be accelerated up to the necessary inflight speed, and then disengaging the rotor clutch before the take-off run (Fig. 14).

Table 2 Civil aircraft categories.

Civil aircraft	Gliders Tourism (General aviation)	Sailplanes Tourism Trainers Ultralight aircraft (RA, LSA, UV) Aerobatic aircraft	
	Executive (Business aviation) Amphibious, flying planes, flying boats Commercial airliners	Passenger air transport Cargo air transport	Small regionals (20–60 seat segment) Large regionals (60–100 seat segment) Small single-aisle (100–150 seat segment) Large single-aisle (150–220 seat segment) Wide Body (220+ seat segment) Freighters + Outsize Cargo Freight Transporters



Figure 15 Cessna F172G (UK registration G-BGMP) at Kemble Airfield, Gloucestershire, England. Source: https://it.m.wikipedia.org/wiki/File:G-BGMP_Reims_F172_@Cotswold_Airport,_July_2005.jpg. Credits: Adrian Pingstone.

Air Vehicles Categories by Types and Roles

According to the classification criteria of **Table 1**, air vehicles (aircraft) with fixed-wing surfaces and a propulsion system are subsequently sub-divided into two major categories: Civil Aircraft and Military Aircraft. These two major groups depend on the specific utilization types and roles the aircraft has been designed and built for, according to technical and customer specifications. Only Civil Aircraft will be discussed herein.

Civil Aircraft

Table 2 shows the categories in which civil aircraft are grouped. In the table, gliders, seaplanes, and amphibious have been already described in the section of Aircraft with fixed-wing surfaces and with a propulsion system. In the following, the three categories of aircraft for tourism (General Aviation), executives-business and commercial transports will be briefly discussed.

Tourism (General Aviation) Aircraft

This category includes aircraft for private, tourism and recreational flight, flying school and training, aerobatics and ultralight aircraft. As per ICAO international standard classification, aircraft in this category are divided in:

- General aviation (GA) aircraft
- Aerial work (AW)
- Commercial air transport (CAT)

A tourism aircraft is generally twin-seats, side-by-side, or in tandem, aircraft with a nose propeller engine, usually with a high wing aerodynamic configuration and fixed landing gear (**Fig. 15**). Aircraft for aerial work usually are employed for aerial survey and photography, parachute dropping, crop spraying, and banner towing (**Fig. 16**).



Figure 16 An Air Tractor Model 502 aircraft flies eight to 10 feet off the ground to spray local fields. Source: <https://www.altus.af.mil/News/Article-Display/Article/352299/crop-dusters-to-grace-altus-skies-soon/>. Credits: Airman 1st Class Kenneth W. Norman/ 97th Air Mobility Wing Public Affairs, USAF.



Figure 17 Otis Air Base, Cape Code, MA - Maj. John Klatt, of the Air National Guard's aerobatic team in his Staudacher S-300D. Source: <https://www.nationalguard.mil/Resources/Image-Gallery/News-Images/igphoto/2000014062/>. Credits: Senior Master Sgt. Thomas Meneguín.



Figure 18 S-301 Sorvanets ultralight aircraft, Kiev, Svyatoshin airfield, October 1999. Source: https://commons.wikimedia.org/wiki/File:S-301_ultralight_aircraft. Credits: Zinnsoldat.

Aerobatic Aircraft

Aerobatic aircraft are single or twin-seater aircraft, with low wings and outstanding maneuverability and strength characteristics because of the high accelerations achieved in maneuvering during aerobatic air as shows in Fig. 17.

Ultralight Aircraft

Single-seaters with weight of up to 300 kg or twin-seaters with weight of up to 650 kg. These aircraft are similar to a general aviation aircraft but with a very light material structure and limited weight and stall speeds, due to regulations limitations. (Figs. 18 and 19).



Figure 19 SHARK AERO SHARK UL, 83-APS/029/2014 Private. *Source: Credits: Brian G. Nichols.*



Figure 20 Cessna 560XL Citation Excel, N2/560-5333 (FAA). *Source: Credits: Brian G. Nichols.*



Figure 21 Learjet 35, N880Z/591, Potomac Corp. *Source: Credits: Brian G. Nichols.*

Executive-Business Aircraft

Business aviation aircraft have intermediate size between general aviation aircraft and larger commercial airliners. Aircraft in this category have configurations with two or three turbojet engines, a pressurized cabin to operate at high altitudes (up to 15,000 m) and relevant cruising speeds (over 800 km/h). Business aircraft can accommodate up to 10–12 passengers, having cabins with elegant and outstanding interiors for business or corporate executives transportation and relative baggage.

The Cessna 560XL (Fig. 20) is 9-pax business jet with a cruising speed of 802 km/h, able to cover 3907 km range, and a maximum altitude of 13715 m. Another high-performance jet is the Learjet 35A (Fig. 21), 8-pax accommodation and a cruising speed of 852 km/h, it can cover a range of 4067 km and a similar max altitude, 12495 m. The Falcon 7X (Fig. 22), a great success of the Dassault



Figure 22 Dassault Falcon 7x Armée de l'Aire (CTM) F-RAFB - MSN 86. Source: License: <https://creativecommons.org/licenses/by-sa/2.0/deed.fr>. Credits: Laurent ERRERA from L'Union (french press), France.



Figure 23 Partenavia P.68C, N79N/305, Flying Fisherman LLC. Source: Credits: Brian G. Nichols.



Figure 24 PIAGGIO P-180 AVANTI, D-IZZY/1034 (AIRGO FLUG SERVICE GMBH). Source: Credits: Brian G. Nichols.

business jet family in operation, is able to carry 12–19 passengers, depending on the configuration, a cruising speed of 904 km/h, a max altitude of 15,545 m and a range of 11,020 km.

Some small firms also use general aviation piston-engine or turboprop airplanes for their business and for commuting executives and personnel, due to the affordable operating and maintenance costs. The Partenavia (now Vulcan Air) P-68 (Fig. 23) is a 1990 kg twin piston-engine aircraft, hosting a weather radar in the nose, capable to carry six passengers with a cruising speed of 311 km/h. One of the most innovative project of executive aircraft is the Piaggio Aerospace P180 Avanti (Fig. 24), sporting a company-registered “Three-Lifting-Surfaces” configuration.



Figure 25 Gulfstream C-20A. UAVSAR underbelly pod is mounted on NASA's C-20A environmental research aircraft to conduct data collection for a variety of geophysical research missions. Source: <https://www.nasa.gov/centers/armstrong/news/FactSheets/FS-089-DFRC.html>. Credits: NASA Photo/Lori Losey.



Figure 26 City Airline Embraer ERJ 135 (SE-RAA) landing at Birmingham International Airport, Birmingham, England. Source: https://it.m.wikipedia.org/wiki/File:City_airline_embraer_erj135_se-raa_ar.jpg. Credits: Adrian Pingstone

Equipped with a powerful two pusher-turbo-propeller engines, this aircraft exhibit a max cruise speed of 732 km/h, accommodates max 6–9 passengers and can carry 5 passengers over a 3240 km route. Some business aircraft types are also used for technological experimentation with special electronic equipment installation (Fig. 25).

Commercial Air Transport Aircraft (Airliners)

Used by world airlines for revenue air transport on assigned regional and international networks, these aircraft are designed on major specific requirements as range, speed, take-off and landing distances, comfort, and payload capacity. As shown in Table 2, commercial airliners are divided in passenger air transport and cargo, the former usually sub-divided in five large subcategories according to the fleet segmentation and to the specific market segment of utilization:

- Small regionals (20–60 seats), with range in the window 1482–2780 km (Figs. 26 and 27)
- Large regionals (60–100 seats), with range in the window 2410–5560 km (Figs. 28 and 29)
- Small single-aisle (100–150 seats), with range in the window 4080–7410 km (Fig. 30)



Figure 27 Crossair Saab 2000; HB-IZR@ZRH. Source: [https://commons.wikimedia.org/wiki/File:38cb_-_Crossair_Saab_2000;_HB-IZR@ZRH;23.08.1998_\(6115679325\).jpg](https://commons.wikimedia.org/wiki/File:38cb_-_Crossair_Saab_2000;_HB-IZR@ZRH;23.08.1998_(6115679325).jpg). Credits: Aero Icarus from Zürich, Switzerland.



Figure 28 Flybe (Stobart Air) ATR 72-600 aircraft taxiing at Manchester Airport. Source: <https://www.flickr.com/photos/levien66/27321615950/>. Credits: Transport Pixels

- Large single-aisle (150–220 seats), with range in the window 3700–7780 km (Fig. 31)
- Wide body (220+ seats), with range in the window 6950–16780 km (Figs. 32–34)

Regional aircraft can be equipped with two turboprop or jet engines, whilst the small and large single-aisles, usually called narrow bodies, are equipped with a couple of powerful turbofan engines. Wide body airliners have two or four, high by-pass ratio, turbofan engines.

Cargo Air Transport Aircraft (Freighters and Outsize Cargo Freight Transporters)

Because more than one-third of world trade value is carried by air, air cargo has become even more critical for serving specific markets demanding speed and reliability for the transport of goods. Freighter versions of large aircraft have, over the years, been developed offering outstanding capacity in terms of volume, weight, cabin dimensions, controlled transport, direct routing, and reliability. These aircraft are designed with very large cross-section cabins, some with space available for passengers, and cargo compartments to host containers and/or pallets where to house heavy loads, cabins offered with side doors and cargo loading systems for small as well as large loads.



Figure 29 SKYWEST CRJ-700 N609SK. Source: [https://commons.wikimedia.org/wiki/File:SKYWEST_CRJ-700_N609SK_\(2750765607\).jpg](https://commons.wikimedia.org/wiki/File:SKYWEST_CRJ-700_N609SK_(2750765607).jpg). Credits: Eddie Maloney from North Las Vegas, USA.



Figure 30 KLM Boeing 737-700; PH-BGH@ZRH;26.02.2013/693au. Source: KLM Boeing 737-700; PH-BGH@ZRH;26.02.2013/693au, Credits: Aero Icarus from Zürich, Switzerland.



Figure 31 Airbus A320-214, D-ABFP/4606, AIR BERLIN. Source: Credits: Brian G. Nichols.



Figure 32 Boeing B787-8 Dreamliner B788 QATAR. Source: <https://www.flickr.com/photos/78631472@N03/27642199744/>. Credits: flybyeigenheer



Figure 33 Airbus A340: wide-bodied airliner, flying in the company livery. Source: Author's personal archive.



Figure 34 Airbus A380 Emirates Airlines over New York JFK airport. Source: [https://commons.wikimedia.org/wiki/File:A6-EDY_A380_Emirates_31_jan_2013_jfk_\(8442269364\).jpg](https://commons.wikimedia.org/wiki/File:A6-EDY_A380_Emirates_31_jan_2013_jfk_(8442269364).jpg). Credits: Maarten Visser from Capelle aan den IJssel, Nederland.

The Antonov An-124 (**Fig. 35**) is a typical outsize freight transporter, it can carry a flight crew of 6 and 88 persons in the aft of wing carry-through cabin, with max payload between 120 and 150 tons, over 3700–4800 km, while the An-225 *Mriya* (**Fig. 36**)—the world's largest aircraft—established a world record of flying 253,82 tons load to 2000 m. Commercial payload attains 188 tons load.



Figure 35 Volga-Dnepr Antonov An-124 Ruslan taking off from Helsinki-Vantaa airport. Source: https://it.wikipedia.org/wiki/GNU_Free_Documentation_License. Credits: Antti Havukainen



Figure 36 Antonov An-225 Mriya outside cargo freight transporter. Source: <http://www.airliners.net/photo/Antonov-Design-Bureau/Antonov-An-225-Mriya/1858115/L>. Credits: Oleg V. Belyakov - AirTeamImages.



Figure 37 Boeing 747-200F of Evergreen International Airlines after landing at the Airport Hahn. Source: License: <https://creativecommons.org/licenses/by-sa/3.0/de/deed.en>. Credits: Wo st 01 (https://de.wikipedia.org/wiki/Benutzer:Wo_st_01).

Derived from the standard passenger version the Boeing B747-200F Freighter (**Fig. 37**) can deliver 90.720 kg of containerized or palletized load in its main cargo deck over more than 8340 km range. A very special conversion of the B747 is the B747 LCF Dreamlifter (**Fig. 38**), used to transfer B787 Dreamliner aircraft major sub-assemblies from one manufacturing site to another, with a maximum revenue freight load of 113 tons and a 8149 km range.



Figure 38 Boeing 747-4H6(LCF) Dreamlifter (N718BA) at Chubu International Airport. Source: <https://flickr.com/photos/143081994@N02/28155819474>. License: <https://creativecommons.org/licenses/by-sa/2.0/deed.en>. Credits: Muroi 8210.



Figure 39 Airbus A300-600F FEDERAL EXPRESS CDG AIRPORT. Source: [https://commons.wikimedia.org/wiki/File:N724FD_CDG_\(23123663860\).jpg](https://commons.wikimedia.org/wiki/File:N724FD_CDG_(23123663860).jpg). License: <https://creativecommons.org/licenses/by-sa/2.0/deed.en>. Credits: Eric Salard, France.



Figure 40 Airbus A300-600ST Beluga. Source: [https://commons.wikimedia.org/wiki/File:DSC_2183-F-GSTC_\(10496737936\).jpg](https://commons.wikimedia.org/wiki/File:DSC_2183-F-GSTC_(10496737936).jpg). Photo prise à l'aéroport de Toulouse-Blagnac (LFBO) en France. License: <https://creativecommons.org/licenses/by-sa/2.0/deed.en>. Credits: Laurent ERRERA from L'Union, France.

The Airbus A300-600 Freighter (A300-600F) (**Fig. 39**) offers a pure freight capability up to 54 tons over 4900 km range. A special outsize cargo freight transport version is the Airbus A300-600ST Super Transporter *Beluga* (**Fig. 40**), a modified version of the A300-600F, used for carrying complete sections of Airbus aircraft from the major European partners' production sites around Europe to the final assembly lines in Toulouse, France and Hamburg. The *Beluga* has a MTOW of 155 tons and is capable of carrying max 40 tons payload over 2779 km, or 26 tons payload over 4632 km. To support the company aircraft production rate increases Airbus



Figure 41 Maiden flight of the Airbus “Beluga XL” on 2018, July 19th at around 10:31a.m. Source: [https://commons.wikimedia.org/wiki/File:%22Beluga_XL%22_A330-743L_\(cropped\).jpg](https://commons.wikimedia.org/wiki/File:%22Beluga_XL%22_A330-743L_(cropped).jpg). License: <https://creativecommons.org/licenses/by-sa/4.0/deed.en>. Credits: Julien Jeany.

A300-600ST *Beluga* will be gradually replaced by the A330 *BelugaXL* (Fig. 41), which performed its maiden flight in July 2018, and is derived from the A330 wide-body airliner. The *BelugaXL*, expected to enter service in the second half of 2019, will incorporate a highly enlarged cargo bay structure with modified rear and tail section.

Summary and Conclusions

Air vehicles are technological objects so diverse in terms of aerodynamic configurations, categories, and roles that a definitive classification cannot be exhaustive of the matter. The paper aims at giving the fundamental, widely renowned, and worldwide-accepted classification of aircraft utilized in civil commercial transportation, for both passengers and cargo purposes. The most important, almost unique, element chosen in the classification process has been the physical dynamic sustaining phenomenon by which an air vehicle derives its motion through the airflow, being the static lift for lighter-than-air flying vehicles—as aerostats—and the dynamic lift for heavier-than-air aircraft, classified as aerodynes. This latter cluster of air vehicles has been given the largest emphasis due to the technological advancements embedded in the particular aerodynamic configuration, with aerodynamic lift devices—for fixed wing or rotary wing aircraft—with or without a propulsion system, or specific flying solutions, as for the gyroclines. Also, a categorization of aircraft by types and specific roles of utilization has been presented, with emphasis on transport airliners and freighters, due to the importance this type of transportation plays in several world markets. Specific attention could be given in the future to manned-unmanned air vehicles objects, a category which will play a relevant role in the imminent future for urban and autonomous transportation over short urban routes and mostly for the technological advancements to be adopted by aircraft designers.

Acknowledgments

The author conveys his sincere acknowledgments to Carlo De Nicola and Francesco Marulo, professors, respectively, of Aircraft Aerodynamics and Aerospace Structures & Aeroelasticity in the University of Naples “Federico II.” A special acknowledgment for the valuable inherited knowledge is conveyed to the memory of our Master, Valentino Losito (1941–2002), professor of Applied Aerodynamics from 1967–92 in the same University and of Aeronautics in the Italian Air Force Flight Academy. To these beautiful minds of my university years, I owe the rigor and the disciplined approach toward the intricacy and the riveting secrets of the Science of Flight. Specific aircraft images in this article are courtesy of Mr. Brian G. Nichols, photojournalist, who is gratefully acknowledged.

See Also

Aviation Safety—Commercial Airlines; Helicopters in Emergency Medical Response; Airfreight and Economic Development; Introduction to Air Transport; The History of Air Transport; The Future of Air Transport; Air Cargo Carriers; Aircraft Manufacturing; Air Transport

Relevant Websites

<https://www.airships.net/zeppelin-nt/>
<http://lisa-airplanes.com/>

<http://www.seaplaneinternational.com>
<http://lancet.mit.edu/>
<https://altigator.com/drone-uav-uas-rpa-or-rpas/>
<https://caa.co.uk/>
<https://commons.erau.edu/publication/72>
<https://www.hybridairvehicles.com/>
<https://www.enac.gov.it/>
<https://www.festo.com/>
<https://www.easa.europa.eu/>
<https://www.eurocontrol.int/>
<https://www.boeing.com/>
<https://www.airbus.com/>
<https://www.dassault-aviation.com/fr/>
<https://cessna.txtav.com/>
<https://www.bombardier.com/>
<http://www.gulfstream.com/>
<https://www.dvbbank.com/>
<http://www.faa.gov/>
<http://www.tecnam.com/>
<http://piaggioaerospace.it/>

References

Crocco, A.G., 1923. Commercial airships. *Aeronautical Digest*, III, n. 6, New York, pp.12, 44–45.
 Levis, E., Serghides, V.C., 2014. The potential of seaplanes as future large airliners. *Applied Aerodynamics Conference: Advanced Aero Concepts, Design and Operations*, Bristol (UK).
 Losito, V., 1982. *Fondamenti di Aeronautica Generale*. Italian Air Force Flight Academy, Pozzuoli, Naples.
 Roskam, J., 2015. *Airplane Design Part 1, Vol. 1. Design, Analysis & Research (DARcorporation)*. Kansas, USA.
 Vagianos, N.J., Thurston D.B., 1970. Hydrofoil seaplane design. Report N.6912. Thurston Aircraft Corporation, Sanford, Maine.
 von Karman, T., 2004. *Aerodynamics: Selected topics in the light of their historical development*. Dover Publications, New York.

Further Reading

Doganis, R., 2002. *Flying Off Course: The Economics of International Airlines*. Routledge, London.
 EASA, EUROCONTROL, 2018. *UAS ATM Integration: Operational Concept*. Brussels.
 Eriksson, S., Steenhuis, H.J., 2015. *The Global Commercial Aviation Industry*. Routledge Publishing, London.
 Everaerts, J., 2008. The use of unmanned aerial vehicles (UAVs) for remote sensing and mapping. *ISPRS Proceedings, XXXVII Congress*.
 Federal Aviation Administration (FAA), 2013. *Integration of civil unmanned aircraft systems (UAS) in the National Airspace System (NAS) Roadmap*. US Department of Transportation.
 Festo AG & Co. KG., 2011. *Smartbird. Aerodynamic lightweight design with active torsion*. Esslingen (Germany).
 Jane's All The World Aircraft – In Service 2017–2018. Jane's By HIS Markit. Coulsdon, Surrey (UK).
 Korchenko, A.G., Il'yash, O.S., 2013. The generalised classification of Unmanned Air Vehicles. *IEEE Second International Conference Proceedings*.
 Leary, W.M., 1995. *From Airships to Airbus: The History of Civil and Commercial Aviation*. Smithsonian Institution Printing, Washington, DC.
 Leishman, G.J., 2016. *Principles of Helicopter Aerodynamics*. Cambridge University Press, Cambridge, MA.
 Robb, Raymond, L., 2006. *Hybrid Helicopters: Compounding the Quest for Speed*. American Helicopter Society, Vertiflite.
 Torenbeek, E., 1982. *Synthesis of Subsonic Airplane Design*. Kluwer Academic Publication, Norwell, MA, USA.
 von Mises, R., 1945, 1959. *Theory of Flight*. Dover Publications Inc., New York.
 Wilson, S., 1999. *Airliners of the World*. Australian Aviation.
 Wilson, T., Schneider, J., Pierpoint, P., 2009. *Unmanned Aircraft System Propulsion Systems Technology Survey*. Embry-Riddle Aeronautical University, Washington, DC DOT/FAA/AR-09/11.

Aircraft Manufacturing

Antonio Sollo, Piaggio Aerospace, Villanova d'Albenga (SV), Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	290
The Start of Aircraft Manufacturing	290
The Riveting for Aircraft Assembly	292
First Attempts for Alternative Materials in Aircraft Manufacturing	292
First Born of Jigs and Tools in Aircraft Manufacturing	292
Aircraft Manufacturing from Blocks	292
The Use of Lightweight Materials in Aircraft Manufacturing	293
The Evolution of Aluminum Alloys for Aircraft Manufacturing Application	293
The Carbon Fiber Development for Aircraft Manufacturing	293
Fiber Laminates	296
Rivetless Structures for Aircraft Manufacturing and Assembly	297
Aircraft Final Assembly Line	297
Future Trends	298
Conclusions	298
Biography	298
References	299
Further Reading	299

Introduction

When Wilbur and Orville Wright made their first successful flight in 1903, aircraft manufacturing was an homework practise developed in the small shops or garages of experimenters and adventurers. The First World War and the significant superiority given by aviation pushed in a significant way the evolution helped of airplanes and their manufacturing out of the workshop and into mass production. Second-generation aircraft helped post-war operators to make start commercial use of them, particularly as carriers of mail and express cargo. Airlines, however, remained unpressurized, poorly heated with spartan interiors and accommodation for passengers and unable to fly above the bad weather. Despite these drawbacks, passenger travel increased by 600% from 1936 to 1941, but was still a luxury that relatively few experienced. The huge advances in aeronautical technology and the concomitant use of air power during the Second World War boosted the explosive growth of aircraft manufacturing capacity after the war in the United States, the United Kingdom and the Soviet Union. The availability of turbojet-powered civilian aircraft in the late 1950s made air travel faster and more comfortable and began a faster growth in commercial air travel. In the 1970–80s Europe has enormously developed the capability of designing and manufacturing passenger transport aircraft by the Airbus consortium who has taken advantage of the huge technological effort spent for the Concorde development, then invested in the A300/A320 development as well as in the turboprop ATR 42/72 family, the most successful commuter in the world. At the same time the development of military airplanes as Tornado, Eurofighter and Mirage/Rafale has in parallel made progressing the military aircraft manufacturing capabilities that has then been strengthened by the development of the large military transport airplane A400M, the first military program managed by the Airbus consortium. By 1993 over 1.25 trillion passenger miles were flown worldwide annually and this figure nowadays has almost quadrupled thanks to the introduction of new generation airplanes powered by very efficient engines and manufactured with the most advanced manufacturing technologies and materials. Thanks to this development the costs, weight and pollution have been drastically reduced, and the simultaneous development of airport infrastructure has contributed to the fast growth of the air transportation during last decades.

The Start of Aircraft Manufacturing

Aircraft may not seem to have changed much in the past few decades, but within the last half-century advances in how they are manufactured have been as great as the evolution in aircraft production over the first 50 years of aviation. Even tough from outside the airplanes may look like always the same, the technology hidden inside continues to evolve as industry drives down costs and weight. The wood frame, wire bracing and fabric skin of the Wright 1903 Flyer was setting the standard of the airframe architecture and manufacturing in the first decades of powered flight. Weight was the main concern for the first airplanes design and manufacturing and wood was the only light material that was strong enough. It was available and affordable, easy to work, resilient and repairable. The Flyer's long wing spars were made from spruce, and the airfoil-shaped ribs from ash. Cotton fabric was applied



Figure 1 1903 Wright Flyer—National Air and Space Museum—Smithsonian Institution. *Source: Wikipedia.*

provide strength and sealed with canvas paint. But wood-and-fabric structures were not that strong, so the wire-braced, strut-supported biplane became the first “standard” configuration (**Fig. 1**).

Wood remained the principal material, but structures rapidly evolved through World War I. Cantilevered wings with thicker airfoils and stronger box spars replaced draggy bracing wires, notably with the Fokker Dr1 triplane in 1918. In 1912 Deperdussin racer introduced the monocoque fuselage, formed of thin plywood layers over a circular frame—light, strong and streamlined. After the war, the Loughhead (later Lockheed) brothers and Jack Northrop developed a method of forming fuselage half shells from laminated spruce in concrete molds and produced the Vega, Orion and other successful monoplanes. The monocoque fuselage was lighter and offered lower drag, but was more costly to manufacture and more complicate to repair. This led to semi-monocoque construction, used in German Albatros fighters, for which load-bearing plywood skin panels were glued to longitudinal longerons and internal bulkheads. As metal replaced wood after the war, the semi-monocoque concept was replaced by the stressed skin concept, but to this day—along with the cantilevered wing—it still remains the prevalent aircraft structural configuration.

Flown in December 1915, Hugo Junker’s J1 was revolutionary—an all-metal, cantilever-wing, stressed-skin monoplane. Wood was light, but subjected to be damaged and deteriorated along the years. The J1 was made of steel; a welded fuselage frame covered with thin sheet and wing skins internally reinforced by panels with spanwise corrugations. But steel is heavy, and development of the lightweight aluminum alloy led in 1919 to the Junkers F13, the first all-metal transport aircraft. In the meantime aircraft such as the Ford Trimotor and Junkers Ju52 also used corrugated skins for strength, but these increased drag (**Fig. 2**).

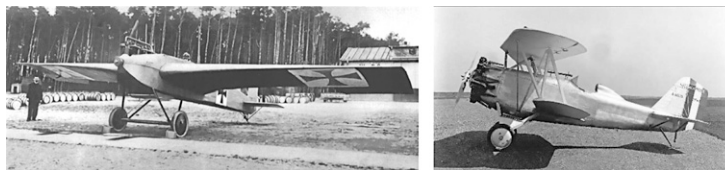


Figure 2 Junkers J1 (left) XFH naval fighter prototype (right). *Source: Wikipedia.*

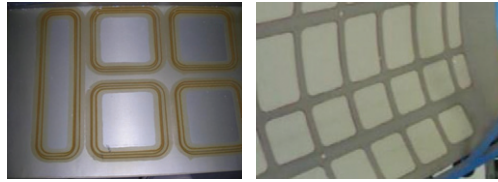


Figure 3 Prototype bonded fuselage skin panels with reinforcement pads (multilayers left, two layers right). *Source: Courtesy of Piaggio Aerospace.*



Figure 4 Piaggio P180 Avanti Vacuum jigs for central fuselage assembly. *Source: Courtesy by Piaggio Aerospace.*

The Riveting for Aircraft Assembly

The first riveted aluminum structure—standard for aircraft manufacturing for decades—was Hall Aluminum Aircraft Co.’s XFH naval fighter prototype flown in 1929. Shortly after, Hall built perhaps the first flush-riveted aircraft, the PH-1 flying boat. Flush riveting and butt joints between skin panels, rather than the dome rivets and lap joints then used, reduced drag and were famously featured in the Hughes H-1 racer, the streamlined, all-metal, retractable-gear monoplane in which Howard Hughes in 1935 set a world landplane speed record of 352 mph (Fig. 3).

First Attempts for Alternative Materials in Aircraft Manufacturing

By the 1930s, riveted sheet-metal construction dominated aircraft manufacture, but lack of aluminum availability during wartime brought a revamping in wood—notably with the de Havilland Mosquito. Plywood facings, bonded to a balsawood core and formed using molds, produced monocoque structures. Experience led de Havilland to use metal-to-metal bonding in the Comet jet airliner. This and other U.K. post-war aircraft used Redux, a strong and durable adhesive invented in 1942. In the 1950s and 1960s, Fokker made extensive use of metal bonding in the F27 turboprop and F28 twinjet airliners, as well as Saab on 340 and 2000 aircraft fuselage skin with bonded stringers or other business aviation aircraft (Cessna family) that have embodied this technology for reinforced skin panels (fuselage and wing) manufacturing.

First Born of Jigs and Tools in Aircraft Manufacturing

Through the 1940s, aircraft were assembled by building skeletal frameworks and attaching the skins. Fairey Aviation developed the technique of building an aircraft as a series of subassemblies in jigs. Fairey’s method involved installing the skin in a jig that provided accurate, repeatable part location, then attaching the substructure as individual parts or as a subassembly produced in another fixture. This technique became an industry standard and later on this concept was revolutionized by the Piaggio P180 fuselage manufacturing. This airplane is characterized by an evolutive shape of the fuselage with tight tolerances to keep the airflow as much as possible with lower drag and the fuselage is manufactured from outside to the inside where the formed skins are maintained in place by vacuum tools and then frames and stiffeners are riveted on (Gavazzi, 2001) (Fig. 4).

Aircraft Manufacturing from Blocks

One of the biggest “hidden” advances in aircraft manufacturing was the development of numerical control (NC) machining. This enabled complex structures such as bulkheads and integrally stiffened wing skins to be manufactured from solid blocks of alloy, rather than assembled from sheet metal—improving quality, reducing weight and saving time and cost. NC machining was slow to get interest and being diffused along with manufacturers, until in the 1950s the U.S. Army purchased 120 machines and leased them to industry. Today five-axis high-speed precision machining is standard for metal structures, mainly for wing panels manufacturing where integrally stiffened skins are often used to reduce eliminate the rivets especially for laminar flow wing design application.

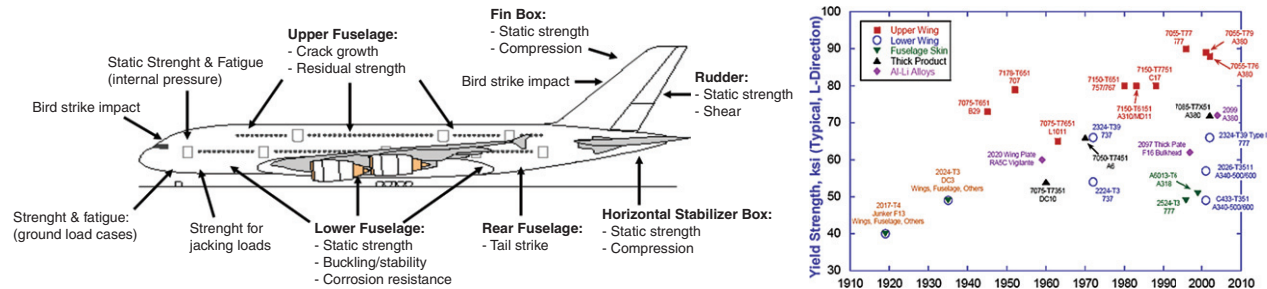


Figure 5 Design drivers for aircraft zones design (left) and Aluminum Alloy evolution over last 80 years (right).

The Use of Lightweight Materials in Aircraft Manufacturing

Titanium's low weight, high strength and heat resistance made it ideal for high-speed aircraft of the 1950s and 1960s and the first titanium aircraft was the Douglas X-3 Stiletto, flown in 1952 and designed to cruise at Mach 2, where skin friction required the heat resistance of titanium. Capable of Mach 3.2, Lockheed's A-12 and SR-71 were also mainly titanium, and the material was to be used for the canceled Boeing 2707 supersonic transport, designed to cruise at Mach 2.7—faster and hotter than the conventionally constructed Concorde.

At the end of 1970s aircraft manufacturers started to look at the use of Carbon Fibers for aircraft component manufacturing and military aircraft manufacturer as usual were the forerunner of this technology. U.S. moved to carbon-fiber composite for wing skins on Boeing AV-8B, F/A-18 and Northrop B-2 and the airframe of the Bell Boeing V-22 tiltrotor. Later also civil aircraft manufacturers started to look at the use of composite for airplane manufacturing, in a first step for secondary parts manufacturing and the for primary structures as well.

The Evolution of Aluminum Alloys for Aircraft Manufacturing Application

Aircraft designer and manufacturers looked at aluminium with great interest. Infact aluminum has the best weight/mechanical strength/thickness ratio. Steel has a density of 7.85 kg/dm^3 and a mechanical resistance (R_m) indicative of 750 N/mm^2 , while aluminum alloys (in the annealed state) have a density of 2.7 kg/dm^3 and an R_m of about 240 N/mm^2 (or less) respectively. As can be seen, the weight/strength ratios are similar and this means that a steel sheet with the same strength would be 1/3 thicker than that of an equivalent aluminum alloy sheet, but this will have implied problems and concerns about the mechanical stability of the fuselage and wing skin panels (Sollo, 2010)

In the continuous infinite race aimed at increasingly reducing the weight of airplanes, over the years aluminum alloys with increasingly high strength characteristics have been developed. Moreover different alloys have been specifically developed and tailored for the different applications on different areas of the airplane where the design drivers are different, depending by the dimensioning loads of the different aircraft components (Niu, 1995) (Fig. 5).

The Carbon Fiber Development for Aircraft Manufacturing

The first carbon-fiber primary structure in a production commercial aircraft was the vertical stabilizer of the Airbus A310-300, first flown in 1985. It marked the beginning of a progression of increasing composites use in the European manufacturer's airliners that added the horizontal stabilizer with the A320 in 1987 and A330/A340 in 1994 and the center wing-box and aft fuselage of the A380 in 2005.

However the first-generation carbon structures were labeled “black aluminum” as their designs were carried over from metallic airframes and did not make full use of carbon fiber’s benefits, as well as, being the composites a completely new and unexplored technology for the Certification Authorities, stringent conservative safety factors were applied and the airplane resulted in a significant overweight. The most significant example of this “black aluminium” approach was the Beechcraft Starship, a canard business aviation airplane that was designed during 1980s and was conceived as a full composite airplane. At the end the airplane was in a significant overweight and this caused the airplane to fail and only 53 airplanes were built (Fig. 6).

A different approach was undertaken in the same period by the Piaggio P180 AVANTI (three lifting surface airplane developed in the same period) where the use of carbon fiber laminate composites was limited to the tail, to the forward wing and to engine nacelles (Gavazzi, 2001) (Fig. 7).

Today the P180 airplane is still an innovative high performance airplane after more than 30 years from its design start and has been celebrated also for the Aviation centenary in 2003 at NBAA (National Business Aviation Association) Exhibition together with Wright brothers airplane (Fig. 8).



Figure 6 Beech Starship - Source Ken Mist, <https://secure.flickr.com/people/eyeno/5999381564/>.

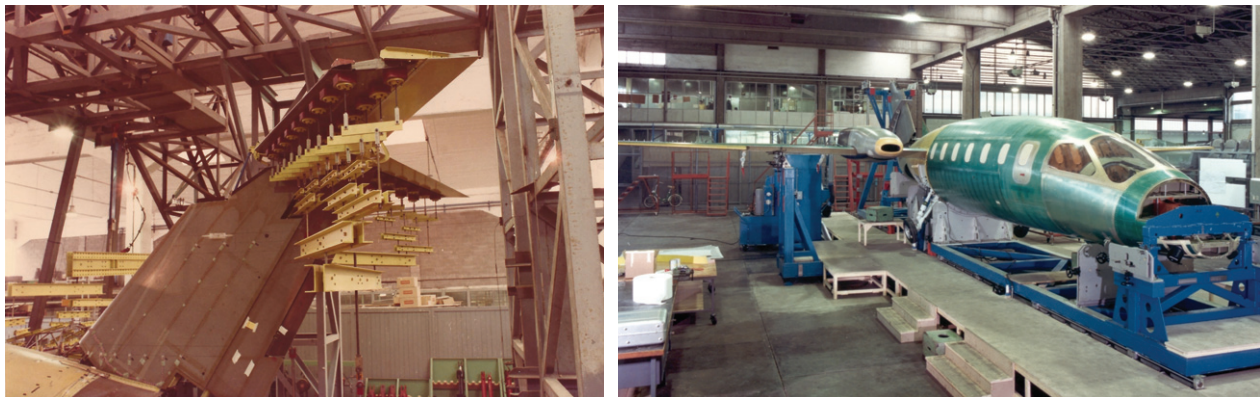


Figure 7 P180 Avanti Carbon Fiber Composite Tail during static tests (left) and Proto airplane assembly (right). Source: Courtesy of Piaggio Aerospace.

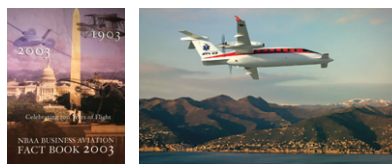


Figure 8 P180 Celebration of 100 Years of Flight (left) Avanti EVO Airplane (right). Source: Courtesy of Piaggio Aerospace.

The lightness, stiffness, and corrosion resistance of carbon-fiber composites led to their use for 50% or more of the structural weight of the Boeing 787 and Airbus A350. But carbon-fiber airframes imposed a return to costly manual layup and assembly (Fig. 9).

The aerospace industry is nowadays entered a period of unparalleled growth in the use of advanced composites (which currently comprise predominantly epoxy resins), with the advent of the Airbus A380, the Boeing 787 'Dreamliner' and the Airbus A350 XWB that has been the first Airbus constructed with both fuselage and wing structures from composite. It is difficult to predict where this



Figure 9 Material map for B787 manufacturing. *Source: Wikipedia.*

growth in composite usage in civil airliners will end, as at least half the weight of both 'Dreamliner' and A350 XWB comprises composites (50% and 53%, respectively), with large carbon fiber panels and fuselage frames made from this material (Denkenaa et al., 2016).

The composite industry's aim is therefore to develop processes whereby large sections, such as the entire fuselage or wing can be made from one piece of composite and also done so in an automated process. This process is now possible thanks to the advent of complex machinery able to impregnate carbon fiber and automatically lay it on a mold according to a predefined pattern that reproduces the design lay-up. One of the leader in this sector is Coriolis, that has developed the first robot Automated Fiber Placement (AFP) solution to comply with aerospace standards for both Thermoset, thermoplastic, and dry fiber manufacturing ability and that is capable of unique convex & concave geometry realization (Fig. 10).

This class of AFP machinery characteristics are the following: capable to manage single fiber from " to 1½", with cut and feed repeatability* ± 2.5 mm at 1 m/s and compaction force range 150–1000 N. The maximum lay-up speed is up to 1.2 m/s that provides very large efficiency combined with a tolerance between course laid up on separate tapes $+2.5/-0$ mm (Lan et al., 2015).

The result of this extensive use of carbon fiber composite structures is a move to more integrated structures to reduce parts count and more automation to drive down costs. The early-2000s Beechcraft Premier and Hawker 4000 had filament-wound fuselages, but automated fiber placement became the industry standard, coupled with new nondestructive inspection techniques. Moreover today the use of thermosetting material is widely diffused versus thermoplastic that have still limited applications in the aerospace field (Fig. 11).

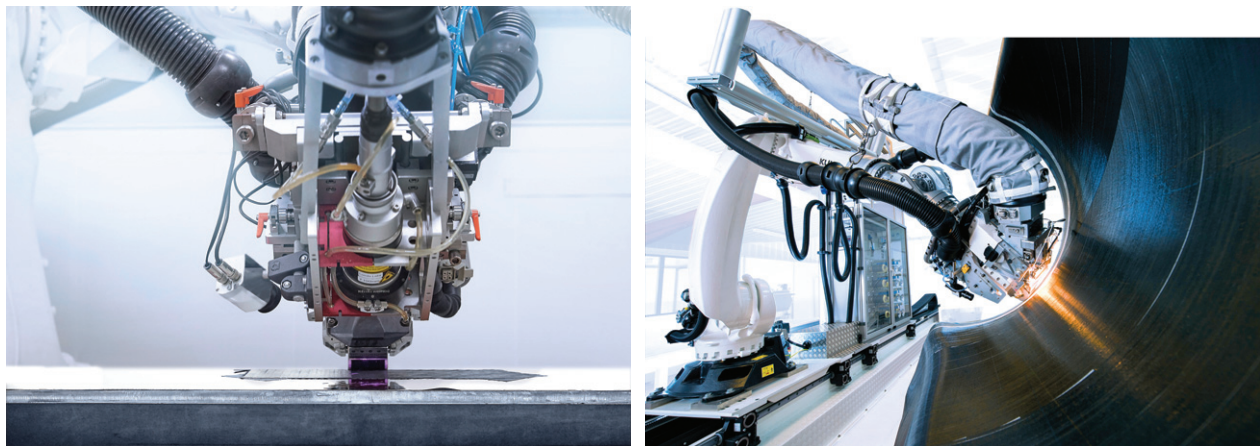


Figure 10 AFP Equipment. *Source: Courtesy of Coriolis Composites.*

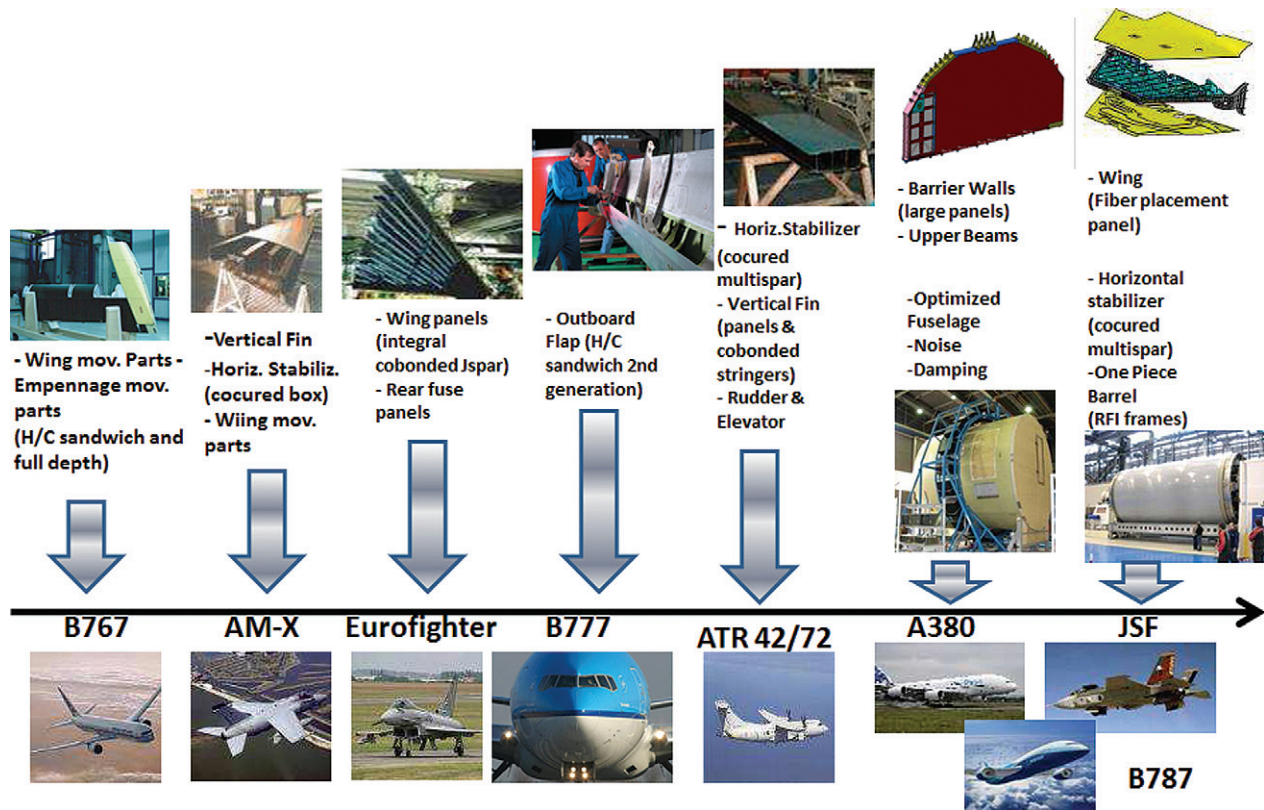


Figure 11 Extension of Composite Carbon Fiber application on most relevant civil and military airplanes of last 50 years.

Fiber Laminates

Between aluminum and carbon fiber is a family of materials, fiber metal laminates, that have found limited but important applications in aircraft manufacturing. Fatigue concerns with aluminum led in the late 1970s to development of an aramid-fiber-reinforced aluminum laminate, Arall, by TU Delft and Alcoa. However flat-sheet Arall had cost and manufacturing issues and this led to a second-generation glass-fiber-reinforced aluminum laminate, Glare, which is resistant to fatigue, impact damage, lightning strikes and fire burn-through. Suitable for double-curved panels, high-strength Glare is used in the Airbus A380 fuselage.

The Glare panels are conformed for alternate layers of aluminum alloys (approximately 0.3 mm thick) and a polymeric compound (plus or minus 0.2 mm). The polymeric compound is made from long fibers (50%–60% by weight) and epoxy resin as a matrix. It is called the configuration formula for the number of aluminum sheets/number of sheets of polymeric compound.

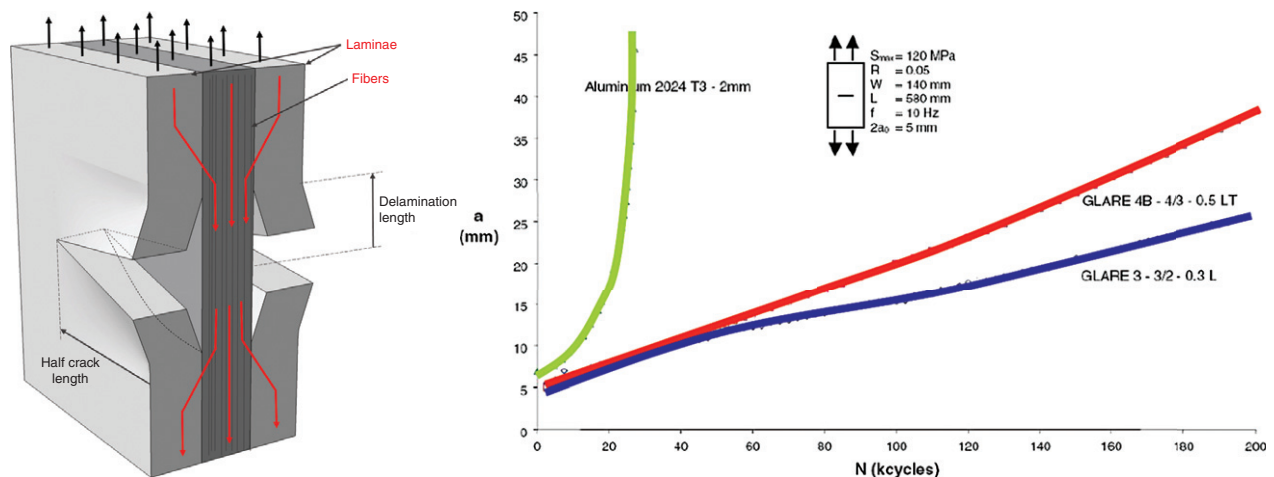


Figure 12 Crack stop mechanism in GLARE panels (left) and crack growth comparison (right) of bare Aluminum 2024 (green curve) vs Glare (blue and red curve).

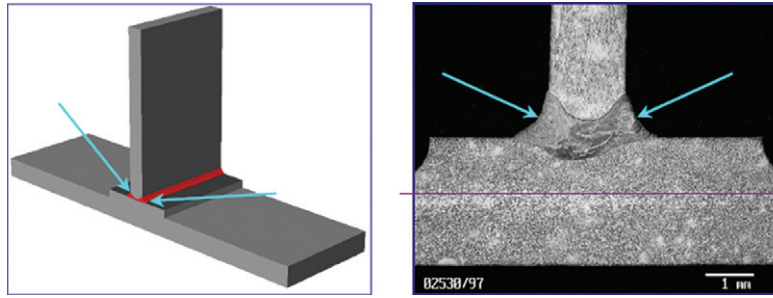


Figure 13 Laser beam welding concept (left) and X-Ray photo of typical welded stiffener (right). *Source: courtesy Piaggio Aerospace.*

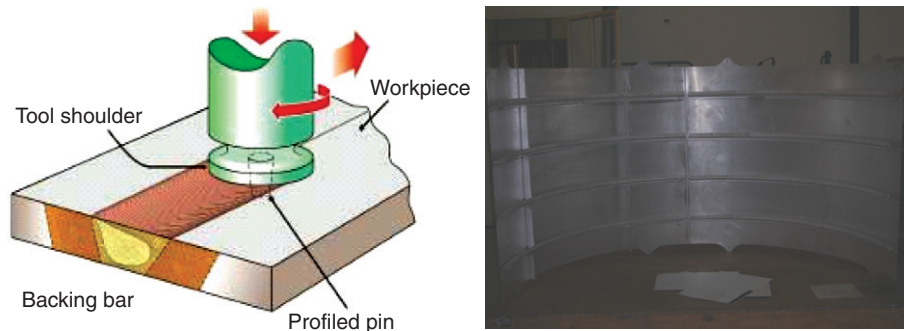


Figure 14 FSW concept (left) – Demonstrator aircraft fuselage panels with welded frames and stiffeners (right). *Source: courtesy Piaggio Aerospace.*

Very directional properties, imposed by the fibers. Its main advantage is the exceptional fatigue behavior in the longitudinal direction. The crack is triggered and propagated only for aluminum sheets. The fibers contain the load therefore the concentration factor of on-board loads is smaller, and also the speed of growth is less. Second, the crack grows, the speed decreases, even stopping (Fig. 12).

Rivetless Structures for Aircraft Manufacturing and Assembly

In an attempt to reduce as much as possible the manufacturing costs, huge effort have been allocated in the development of affordable and automated process aimed to eliminate rivets from the airframe by means of welding technique to join structural components without riveting. Just as reference the time needed for the installation of one rivet is about 1.5 min of two workers that multiplied by several hundred thousand of rivets distributed of the whole airframe of a modern commercial airplanes justify the effort of finding alternative solutions to riveting even if automated (SAE, 1994), to further speed-up the manufacturing and assembly of complex airplane structures. Two welding techniques have been developed in aerospace manufacturing: laser welding and friction stir welding. Laser beam welding (LBW) is a welding technique used to join pieces of metal through the use of a laser. The beam provides a concentrated heat source, allowing for narrow, deep welds combined with filler material use. The process is particularly suited for single curvature stiffened panels where no complex welding pattern had to be realized and the first industrial application of this technique has been done in 2000 on Airbus A318 lower fuselage panels where stringers have been welded to skin (Fig. 13).

The Friction Stir Welding (FSW) is a type of solid-state friction welding in which the melting temperature is not reached in the material to be welded and there is no supporting material. FSW is mainly used to join aluminum alloys (Akshansh et al., 2019) that would otherwise be difficult to weld with other techniques. It was invented in 1991 by The Welding Institute (TWI) of Cambridge (Fig. 14).

This technology has been for the first time implemented in the manufacturing of the airframe of the Eclipse 500, a Very Light Jet (VLJ) aircraft designed by Eclipse Aviation that has been the first certified aircraft with welded structure in 2006.

Aircraft Final Assembly Line

The current trend in aircraft assembly, is to have available equipped aircraft segments on the final assembly line (FAL) that are joined at different station of the FAL and more often these segments are manufactured by different partners and/or subcontractors.

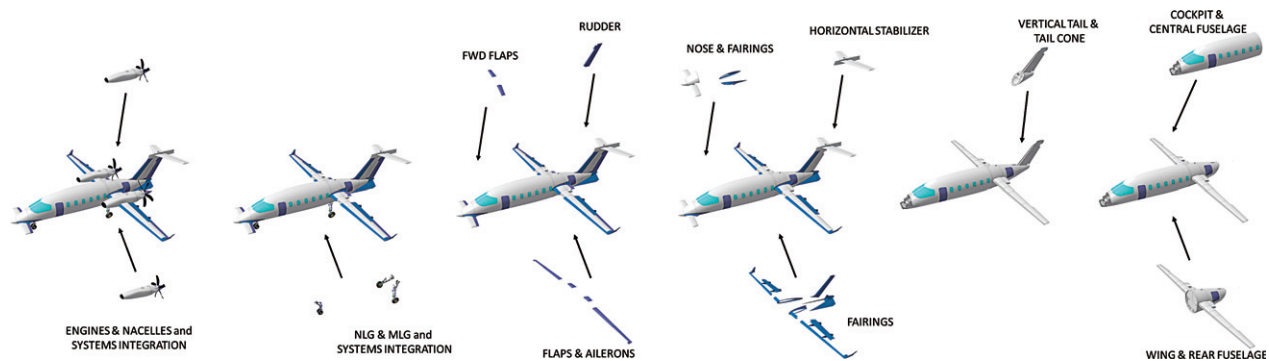


Figure 15 Typical Final Assembly Line Sequence of a Business Aviation Aircraft. *Source: Courtesy Piaggio Aerospace.*

Generally the FAL is shortened to a limited number of station (4-5) and this approach allows to concentrate the integration of high value equipment at the end of the aircraft assembly, thus shifting closer to the aircraft delivery the most significant cash flow thus allowing a faster recovery of the invested cash (Fig. 15).

Future Trends

3D printing, also known as additive manufacturing, is making its way into the large-scale manufacturing process. Recently, Boeing teamed up with Oak Ridge National Laboratory to 3D print the largest single piece ever made using this technology—a 777x wing trim tool. The tool was printed in just 30 h, something that usually takes three months with traditional manufacturing methods. In addition, Airbus and Rolls Royce have revealed plans to begin 3D printing parts for the Trent XWB-97 engine used in the A350-1000. Continued integration of this technology could turn the shipment of aircraft parts into an electronic exchange of 3D printing blueprints. While this technology is still in the initial phases, adoption on a major scale has incredible implications for the industry, including vast efficiency increases, large-scale reductions in the amount of manufacturing waste and tremendous changes to aircraft manufacturing and the supply chain (Werkheiser, 2014).

Conclusions

Over the past few years, the aircraft manufacturing industry has seen countless innovations coming to fruition with many more on the horizon. The advancement made are enormous and more and more innovation are expected in the next decades to achieve the target set for the aircraft of the future: the zero-fuel aircraft. The recent surge in interest has put pressure on global aerospace and defense industries to create a long-term development strategy for the zero-fuel aircraft concept and drive market growth. This pressure is put not only on new propulsion system development, but also aircraft design and manufacturing to reduce weight and then reduce the power needed for the propulsion. The emerging technologies as 3D printing will be more and more developed and may be that in a short a full 3D printed aircraft airframe will not be a dream but will become reality.

Biography

Antonio Sollo is the Chief Technical Officer and Head of Design Organization of Piaggio Aerospace managing the whole Design Office activities for all airplanes of Piaggio Aerospace. Accountable Manager for DOA obligations according to EASA PART 21 requirements. He has a deep knowledge of the aeronautical product, having covered various roles in several aerospace companies from the civil and military transport sector (Alenia/Leonardo 1987–2001), Business Aviation and UAV (PIAGGIO AEROSPACE 2001–13 and 2018 up today), General Aviation (OMASUD 2015–17). Multi-year experience in Aircraft Structural design (both Metallic and Composite Structures), Flight Technologies disciplines, Certification disciplines, Interior & Exterior Noise disciplines, Innovative Manufacturing Technologies (Composite Structures, Metallic Build-up and Machined Structures, Welded Structures). As Manager of Piaggio Aerospace Structure Department as well as Piaggio Aerospace Technical Manager of Pozzuoli Design Office he has managed complex projects such as P180 Avanti II and III, P1XX aircraft development and has managed the whole aerostructure supply chain for the P1XX project and all the research on the next-generation composite primary structures (Fiber Placement, Infusion, RTM, Out of Autoclaves). He also defined the whole airframe preliminary architecture of the PIAGGIO AEROSPACE UAV family (P1HH and P2HH) as well as being the Chief Engineer of the MPA aircraft during pre-development and development start. During OMASUD experience, he has managed the whole company and defined the entire industrialization phase and launch of serial production of the SKYCAR Airplane (first two serial airplanes delivered to launch customer) and of a fully carbon fiber airplane REDBIRD. He also worked as consultant in various fields (rail, naval and automotive) and held teaching in Aeronautical Engineering Technology Manufacturing at the Aerospace Engineering Faculty of Federico II University of Naples. He is the author of many papers presented at national and international congresses, chairman of various sessions at congresses of the AIAA Aeroacoustic Association, Evaluator of research projects of V, VI and VII FQ of the European Community, EUREKA and IWT.

References

- Akshansh Mishra, Devarrishi Dixit, Friction Stir Welding of Aerospace Alloys - International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Volume 7 Issue III, Mar 2019 - Available from: www.ijraset.com.
- Denkena, B., Schmidt, C., Weber, P., Automated Fiber Placement Head for Manufacturing of Innovative Aerospace Stiffening Structures—16th Machining Innovations Conference for Aerospace Industry - MIC 2016.
- P. Gavazzi, Volare AVANTI—Storia degli aerei Piaggio—Proedi Editore 2001.
- Lan, M., Cartié, D., Davies, P., Baley, C., Microstructure and tensile properties of carbon-epoxy laminates produced by automated fibre placement: Influence of a caul plate on the effects of gap and overlap embedded defects. ELSEVIER Composites Part A: Applied Science and Manufacturing Volume 78, November 2015, pp. 124–134.
- Niu, M., Airframe Structural Design: Practical Design Information and Data on Aircraft Structures.
- SAE Technical Paper—1994-10-01 Automatic Riveting Cell for Commercial Airplane Floor Grid Assembly 941837.
- Sollo, A., Technologies for Aircraft Manufacturing—Course Notes—University of Naples—Aerospace Engineering Faculty 2010.
- Werkheiser, N., Overview of NASA Initiatives in 3D Printing and Additive Manufacturing—2014 DoD Maintenance Symposium Birmingham, AL—November 17-20, 2014.

Further Reading

- Composite World - The first composite fuselage section for the first composite commercial jet—7/1/2018.
- Mallik, P.K., Fiber-Reinforced Composites: Materials, Manufacturing, and Design.
- Manufacturing Technology Leaps Benefit Aircraft Evolution Sep 4, 2018 Graham Warwick—Aviation Week & Space Technology.
- Thumbnail Look At History of Aircraft Construction Mar 25, 2016 Graham Warwick—Aviation Week & Space Technology.

A Geography of Road Transport in Cities

Cristian Domarchi*, **Juan de Dios Ortúzar†**, *Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, Santiago, Chile; †Department of Transport Engineering and Logistics, Instituto Sistemas Complejos de Ingeniería (ISCI), BRT+ Centre of Excellence, Pontificia Universidad Católica de Chile, Santiago, Chile

© 2021 Elsevier Ltd. All rights reserved.

Introduction	300
Urban Mobility Hierarchy	300
The Role of Paratransit in Modern Transport Systems	302
The Shared Taxi: A Chilean Paratransit Mode	303
Mode Characteristics	303
Some Basic Indicators	303
What Motivates Shared Taxi Users? An Experiment in Santiago	303
Perceptual Indicators	304
Hybrid Mode Choice Model	304
Summary and Conclusions	304
Acknowledgments	305
References	305
Further Reading	305

Introduction

The transport system is a fundamental element of cities. It provides opportunities for the development of cities' economic activities, allowing for their growth and well-being, and also gives their inhabitants access to jobs, education, health, and mutual interactions. A key challenge for cities is how to implement efficient transport systems, both from the operational and environmental viewpoints, to allow a harmonic and sustainable urban development.

Urban transport networks must offer mobility opportunities for a multiplicity of movements, both of passengers and freight, between spatial locations that can be far apart, and for a large diversity of trip purposes, hours of the day, and days of the week. The modes used vary with the socioeconomic level and the characteristics of the activity system of each city.

Using the Millennium Cities database, [Aguilera and Grébert \(2014\)](#) compared modal shares among different regions in the world. They found that 47% of trips used motorised private transport (cars and motorbikes), 37% corresponded to non-motorised transport (walking, bicycles), and only 16% used public transport. The cultural, socioeconomic and physical differences among the cities compared help to explain the significantly different modal shares observed ([Table 1](#)).

Note that the proportion of trips by car is especially high in North America. Public transport use reaches 45% in Eurasia, but only 19% in Latin America. Non-motorised transport accounts for nearly 50% of urban trips in Asia Pacific and Africa, 30% in Europe, and less than 10% in North America. Note also that in spite of the strong socioeconomic and cultural differences between Latin America and Europe, the aggregate modal shares in both regions are remarkably similar.

These aggregate figures show that, sadly, the main urban transport mode is still the private car, who has played a historic role in the economic development of countries and has become an integral part of the movement of people in all cultures. Its use is not only explained by its practical advantages (flexibility, comfort) but also by deeper psychological aspects such as status, social norm and affective variables that are behind habit formation; these aspects are a barrier for car users changing to other modes.

However, and considering its reduced carrying capacity, the automobile is a highly inefficient mode in terms of road space use and also produces strong negative externalities in the form of congestion, accidents and environmental pollution. It is just not possible to conceive a sustainable urban transport system based on the private car.

The rest of the paper is structured as follows. Section Urban Mobility Hierarchy presents a recommended hierarchy for urban mobility, under the principle that road space—scarce by definition—should be prioritised for its use by the more efficient and sustainable modes. Section The Role of Paratransit in Modern Transport Systems discusses the role of paratransit in modern transport systems and section The Shared taxi: A Chilean Paratransit mode considers, as an example, the case of the Chilean shared taxi, which could be seen as a modern paratransit mode. Finally, the last section summarises our conclusions.

Urban Mobility Hierarchy

Among urban planners and transport specialists there is consensus about the need to prioritise road space for the use of more efficient and less contaminating modes. Multiple alternatives may complement each other to generate feasible (i.e. efficient and

Table 1 Aggregate modal shares in different world regions

Region	Public transport (%)	Non-motorised transport (%)	Motorised private transport (%)
Eurasia	45	20	35
Middle East	10	35	55
Asia Pacific	18	50	32
Africa	5	50	45
Europe	15	30	55
Latin America	19	38	43
North America	3	8	89
Global	16	37	47

sustainable) transport options, and several countries have recognised the need to focus transport planning and road space assignment into these.

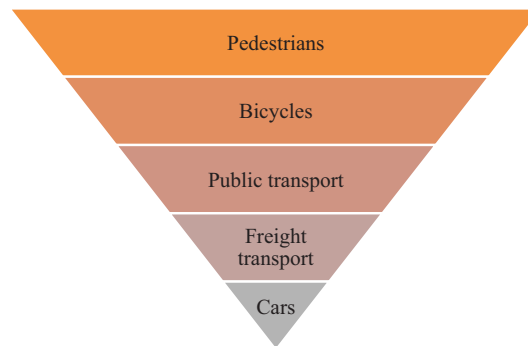
We classified the alternatives to motorised private transport in just two categories above (non-motorised transport and public transport), ignoring the ample variety of vehicles and services available to provide feasible movement alternatives, as well as the importance and potential role of each. Now, we consider a mode classification that emphasises the need to guarantee the mobility of people (and goods) rather than vehicles. This urban mobility hierarchy can be represented as an inverted pyramid where modes allowing for movements with more positive effects for the urban environment receive a higher priority (**Fig. 1**).

The top two tiers of the hierarchy contain alternatives that, together, are now known as *active modes* (i.e. movement is propelled by energy provided by the traveller). Walking does not only interact with practically all remaining modes (as most require access and egress walks), but also constitutes a relevant option for short distance trips. Thus, it is the fundamental link between land uses and the transport network.

The bicycle also presents important benefits in terms of cost, health and the environment. While it allows users to cover longer distances (in Santiago, Chile, it has been shown to be unbeatable for trips up to 7 km), it has problems in terms of comfort and convenience that renders its use lower than expected given its undeniable advantages. The provision of dedicated infrastructure (i.e. cycleways) and complementary facilities (like secure parking and showers) should help to increment its attractiveness.

The next tier in the hierarchy contains public transport. The formal transit systems' supply consists of regularly programmed vehicles with the capacity to move users with a variety of origins, destinations and trip purposes. Usually, the formal public transport modes are classified by vehicle type as follows:

- Buses: Reduced size vehicles (up to 150–200 passengers/vehicle) with routes that may be adapted in time following demand changes. Due to their cost efficiency, buses can serve areas with different population densities, offering shorter access/egress distances than other transit modes. Notwithstanding, the absence of dedicated infrastructure (i.e. bus lanes and bus only corridors) implies irregularities in travel and waiting times that affect their reliability. BRT implementations tend to resolve this problem through adequate interchange infrastructure and road separation.
- Metro (underground) and suburban trains, are high-capacity modes, operating at higher speeds through completely dedicated infrastructure, but with stations that are more separated than bus stops. These tend to be the preferred mode of public transport users, because of their comfort, lower travel times, higher regularity and network simplicity. However, their much higher investments costs only justify their presence in networks with very high demands.
- Light rail (LRT) and trams operate with mid capacity vehicles at relatively low speeds, generally on the surface, either sharing the way with other modes or through exclusive lanes. Although they have lower investment and operational costs than heavy rail, the level-of- service offered to passengers tends to be lower, especially in terms of travel time, regularity and reliability.

**Figure 1** Urban mobility hierarchy. Source: Adapted from Bradshaw (1992)

We will not discuss the next tier, freight transport, but it is obvious that a city cannot live without a well-designed freight and logistics system, so it is of utmost importance that planning considers the movement of freight in harmony with the rest of the system.

Finally, at the bottom of the hierarchy are cars and motorbikes, normally known as private transport. These are the least efficient and less sustainable urban transport modes. Although there is scant published information about the role of motorbikes in urban areas, in several countries of Asia and Latin America they have become a plague, leading to far more fatal accidents and almost equal congestion than cars. Cars, on the other hand, have been amply studied and it has been shown that their marginal external cost/passenger^a is about 10 times higher than that of buses in the peak period, and nearly 2.5 times higher in the off-peak (Rizzi and de la Maza, 2017).

Making people change from cars to more sustainable modes is difficult, among other things, because car users have strong attachments to their mode (status, habit) over and above its intrinsic qualities. To make car users consider other alternatives, it is first necessary to give priority to public transport and to charge car users for their externalities (i.e. congestion pricing). This has been done successfully in several cities and is certainly a way forward. Another good news is that younger people are more concerned about the environment and consider almost unhealthy to be a habitual car user; this has led to a strong increase in the share of active mobility, particularly in Europe, but also in other regions. Finally, there is scope to consider the potential benefits of modal integration. As mentioned, walking is an essential component of any public transport trip; therefore, we need to guarantee that access, egress and interchanges have an adequate quality and be comfortable to users. Likewise, several cities in Europe and North America have been successful in terms of integrating the bicycle with public transport modes, either by providing adequate cycle parking at bus stops or stations, or by allowing the bikes inside public transport vehicles.

Finally, public transport services must be planned such that their various modes satisfy specific roles better adjusted to their design. For example, heavy rail and LRT should cover high demand origin-destination pairs, complemented by BRT, while normal buses and less formal modes (i.e. *paratransit*) may be used to cover zones with lower density, providing greater capillarity to the network.

The Role of Paratransit in Modern Transport Systems

The public transport modes discussed earlier are formal, in the sense of being regulated by the respective authority. However, in many regions of the world the difficulty of providing formal public transport services at times and locations where not enough demand exists, prompts the spontaneous emergence of complementary or alternative informal services. Such *paratransit* services help to fill the wide space between private transport, on the one hand, and conventional, regulated public transport on the other.

While a consensus definition of *paratransit* does not exist, we will take as reference the one provided by Gwilliam (2002) “... any publicly available passenger transport service that is outside the traditional public transport regulatory system.” This definition^b is ample enough to include a myriad of transport options, including motorised vehicles (from taxi-like services provided by motorbikes, cars and minivans) to non-motorised vehicles (bicycles, bici-taxis, rikshaws, etc.), but excludes vehicles reserved for private use and those forming part of the conventional transit system.

Paratransit services become more relevant for the population when conventional transit systems are less developed. They are perceived as versatile and flexible, providing mobility alternatives in areas with difficult access or low coverage by traditional services. In certain cases, *paratransit* services may become an important alternative and even outclass the established transit system. In such cases, the transport authorities may be forced to recognise them as part of the established system, turning them legal without losing their informal flexible/adaptive essence.

This flexibility is one of the main advantages associated with *paratransit* services. These are *gap-filler* services complementing the major transit network in terms of coverage, frequency and adaptability to market changes. By operating with small vehicles, they are able to provide lower waiting times and a greater feeling of comfort and security. On the other hand, as the services are essentially private initiatives, they tend to be the only public transport mode in cities with low availability of economic resources to finance formal public transport systems.

However, the operation of *paratransit* also brings significant societal costs. For example, the use of small vehicles increases congestion, as larger fleets are required to satisfy demand. Likewise, it is not uncommon to observe operative malpractices favoured by the lack of regulation, such as service over-supply in peak periods and insufficient frequencies in the off-peak, lack of coverage in routes with smaller demand, discretionary fares and “street competition” for passengers by different services. Finally, they also generate negative economic effects derived from their market informality, such as inappropriate vehicle maintenance and social security problems for their staff.

In summary, it is not clear whether *paratransit* has effectively a global social benefit. For this reason, the behaviour of authorities has tended to evolve from ignoring (or even fighting against) them to making them formal and regulate their operation. The next step should be helping them to modernise and become an integral part of the formal system, recognising their supporting role to urban sustainability, both as a complementary mode to the formal system and as an alternative to the indiscriminate use of the car.

^aThat is, considering congestion, pollution, and accidents.

^bThis definition is valid for developing countries. In first world countries, *paratransit* tends to concentrate in providing regular services for special demands, such as old age pensioners, people with reduced mobility, and so on. These services are not considered in this chapter.

The Shared Taxi: A Chilean Paratransit Mode

To illustrate the role of *paratransit* options in middle income countries, we present a case study about the *shared taxi*, a typically Chilean mode that—although regulated and constrained in certain aspects—shares some characteristics with *paratransit* services operating in other parts of the world.

Mode Characteristics

The *shared taxi* is minor transit mode that operates in Chilean cities using cars with a maximum capacity of four seats, a predefined coverage area and a route which is, in principle fixed (but that can be adapted to the passengers' needs, especially at its ends). Historically, this flexibility has allowed its operation as a complementary mode to the major transit network (mainly composed by buses), in zones and periods with low coverage or level-of-service below par. The services started spontaneously as private initiatives that the Ministry of Transport has mildly attempted to regulate for quite some time^c. The degrees of freedom set by current legislation make it close to the *paratransit* standard. However, there is not absolute freedom to choose routes and both vehicles and drivers must satisfy certain norms.

Shared taxi operators have low levels of professionalism (services are usually managed by the vehicle owners or by third parties that hire them). The lines are not proper firms; rather, they represent associations of owners/operators belonging to a certain coverage area, who administer basic facilities but do not coordinate departures, level-of-service or fare collection, as all these depend directly of the driver. This precarious operation level, without economies of scale, is a main obstacle for regulation by the authorities.

Some Basic Indicators

Although the participation of *shared taxi* in the transport market of Santiago, Chile's capital and main conurbation is relatively low (Table 2), it increases at night and in the periphery of the city. In the rest of the country, however, the panorama changes; in fact, in most mid-size cities (between 200 and 500 thousand inhabitants) the mode's market share has been under constant increase, in some cases (Arica, Iquique-Alto Hospicio) overtaking the share of the formal transit system.

This increasing demand poses an important challenge in terms of regulation by the Chilean authorities, especially because it has affected the buses and has had perceptible congestion effects in the urban centres of the affected cities.

What Motivates Shared Taxi Users? An Experiment in Santiago

A revealed preference (RP) survey was designed and applied to a sample of current users of bus, metro, *shared taxi*, and combinations. Apart from data about their current trip and socioeconomic characteristics, we gathered information about relevant attributes (travel cost and times, in-vehicle, walking and waiting) for both the chosen alternative and for the others in each individual's choice set. Using a 5-point Likert scale, we also collected a set of perceptual indicators determined on the basis of focus groups with regular and occasional *shared taxi* users, with the aim of understanding their motivations to choose the mode.

Table 2 Shared taxi modal shares in Chilean cities over 200,000 inhabitants

City type	City	Population	Shared taxi proportion of total trips (%)	Shared taxi proportion of public transport trips (%)	Survey year
Conurbation	Great Santiago	6,254,000	4	13	2012
	Great Concepción	971,000	4	12	1999
	Great Valparaíso	944,000	7	20	2014
Middle size city	Coquimbo-La Serena	415,000	14	46	2010
	Antofagasta	354,000	10	26	2010
	Temuco-Padre Las Casas	309,000	8	22	2013
	Iquique-Alto Hospicio	295,000	16	55	2010
	Rancagua-Machalí	294,000	10	35	2006
	Puerto Montt	290,000	15	47	2004
	Arica	221,000	13	55	2010
	Talca	220,000	10	30	2003
	Chillán	216,000	8	31	2003
	Los Ángeles	202,000	13	45	2004

^c Regulatory measures, such as routes' norms and firm requirements, definition of maximum fleet size and prohibition to circulate through certain avenues, were set in the 90s and are still applicable through special laws that prolonged their validity.

Table 3 Perceptual indicators and CFA results

Perceptual indicators	Mean valuation	Confirmatory factorial analysis		
		Convenience	Reliability	Safety
1. Lower travel times	4.20			
2. Travel time reliability	3.60		0.350 (4.0)	
3. Possibility of travelling seated	4.32	0.325 (6.6)		
4. Possibility of carrying cargo	3.54			
5. Less transfers	4.18	0.386 (8.0)		
6. Lower waiting times	3.65	0.180 (3.0)	0.326 (4.0)	
7. Waiting time reliability	3.20		0.866 (5.2)	
8. Leaves me in a convenient place	4.38	0.539 (10.9)		
9. Less accidents	3.29			0.446 (4.3)
10. Less delinquency	3.60			0.863 (4.3)
11. Can make comments/suggestions	3.90			
12. The driver helps the passengers	3.61			

Perceptual Indicators

The analysis of these indicators revealed interesting aspects about the motivations of users for choosing *shared taxi* in Santiago (Table 3). The highest evaluations were associated with the comfort and flexibility offered by the service—Leaves me in a convenient place (4.38), Possibility of travelling seated (4.32), Lower travel times (4.20) and Less transfers (4.18). The worst evaluated aspects were: Waiting time reliability (3.20) and Travel time reliability (3.60), and the perception of safety in case of accidents (3.29).

The last three columns of the table show the results of a confirmatory factorial analysis (CFA) to test the existence of correlations among some of the factors and group them into potential latent variables. We found three such constructs: *Convenience*, *Reliability* and *Safety* which were used, together with the measured alternative attributes and individuals' socioeconomic characteristics, to estimate a hybrid choice model (Ben-Akiva et al., 2012; Bahamonde-Birke et al., 2015).

Hybrid Mode Choice Model

The RP survey allowed us to build a set of available routes, by metro, bus and *shared taxi* for each individual in the sample; these were used to model choices using a hybrid nested logit model incorporating the above-mentioned latent variables (Domarchi et al., 2019). A single *Bus* nest allowed us to treat the potential correlation between alternative bus routes for each individual.

The model's utility function incorporated several level-of-service variables—the ratio of the fare and the individual's wage rate (Gaudry et al., 1989), walking, waiting and in-vehicle travel time, and number of transfers. Some had interactions with the individual's socioeconomic or trip characteristics (i.e. systematic taste variations, Ortúzar and Willumsen, 2011, p. 279) and some also interacted with the latent variables. In particular, *walking time* was found to interact with *gender*, *Convenience* and having made a transfer; *waiting time* interacted with *age*, *Reliability* and with being a student. The variables *fare/wage rate*, *in-vehicle time*, *number of transfers* and *safety* had no interactions.

The results reveal that the marginal utility of walking and waiting time is similar and higher than that of in-vehicle time. Old age pensioners value waiting time significantly higher than the rest, and men value walking time less negatively than women. Also, users that had answered the survey after having transferred, value walking time significantly higher than those who started their trips near the contact point (for whom one minute waiting is equivalent to four minutes inside the vehicle).

High-negative weights assigned to walking time and number of transfers, confirm that users prefer more direct and convenient alternatives. The *shared taxi* is an attractive alternative precisely for these characteristics, in spite of not always offering the best travel or waiting times (including that some *shared taxi* users pay a higher fare). In terms of the latent variables, users who value the mode's *Convenience* have a significantly higher valuation of walking time than other users; also, those that find *shared taxi* more reliable tend to value waiting time (an attribute strongly correlated with reliability) less negatively. Finally, passengers that have a positive evaluation of the mode's security are more likely to choose it.

Summarising *shared taxi*'s flexibility is the best perceived attribute and can, in many cases, compensate for other characteristics perceived negatively or better provided by other modes. In Santiago, users value positively the convenience of having to walk less and require less transfers. They also value the possibility to reduce the egress walk time (through the mode's adaptability in its routes' ends) and to travel with higher safety. This way, *shared taxi* is perceived both as a complementary mode to the formal transit network (when there are no alternative options) and as a feasible and competitive option when there are other alternatives, even potentially less expensive and/or quicker.

Summary and Conclusions

Transport systems are suffering vertiginous changes in the world. The increasing economic power of many developing countries has implied a sustained increase in car ownership that, in turn, has brought in strong increases in congestion and pollution. On the other

hand, the climate change menace and the need to reduce transport emissions, have encouraged governments to design policies to disincentive private car use and to create feasible alternatives to achieve greater urban sustainability.

In this sense, the urban mobility hierarchy offers a valid orientation for planning in terms of promoting more sustainable modes. Active modes, such as walking and bicycle, are key in helping to reduce the obesity pandemic that is affecting developing and developed countries alike. Public transport plays a fundamental role in using urban roads efficiently to move large trip volumes (often for relatively long distance). This is done better when alternative modes in the network exist and act in a coordinated and complementary fashion, fulfilling diverse roles depending on the type of trip and user.

For these reasons, *paratransit*-like modes may play a relevant role, not only complementing the formal transit networks in certain zones or in periods of low coverage, but also by becoming a genuine alternative, valued by its users on account of its flexibility, coverage and comfort. In this sense, the Chilean *shared taxi* is paradigmatic. Our analysis reveals that the attributes valued more highly by its users are precisely those that allow a better adaptation to their needs; in its current status, this is associated with a relatively informal and scantily regulated operation scheme that, unfortunately, could represent a harm for society as a whole.

So, regulating *shared taxi* and *paratransit* modes in general—is a challenge for developing countries. The advent of *ridesourcing* platforms, such as Cabify or Uber, has brought a strong resistance worldwide by the taxi unions (and also by *paratransit* operators), who perceive them as an unjust threat, in lieu of their weak organisation and technological precariousness. If the Chilean authorities, for example, wish to maintain *shared taxi* as an alternative available to users, there is a need for clearly determining the scope and role of each mode—a subject still under discussion. In principle, one would expect the promotion of its complementary role to the formal public transport system (rather than as unregulated competition), especially in areas badly covered by the network, when it makes sense to maintain flexibility in routes and fares.

The need to plan for sustainable transport systems requires decisively implemented policies, under a solid institutional framework, appropriate legal instruments and enough audit capacity. Any modernisation process requires considering both operators and users. Also, the establishment of trade-offs between the need to maintain the attributes inherent to each mode and the welfare of the society as a whole. In this sense, regulation could be the doorway to the modernisation and integration of all public transport modes (formal and informal), but it also must have—as final goal—delivering better levels-of-service to users. Only in this way the public transport system will represent a feasible alternative to the private car, and will serve to building sustainable urban transport networks.

Acknowledgments

We wish to thank the support of the Instituto Sistemas Complejos de Ingeniería (CONICYT PIA/BASAL AFB180003) and BRT+ Centre of Excellence (www.brt.cl) funded by the Volvo Research and Educational Foundations.

References

- Aguilera, A., Grébert, J., 2014. Passenger transport mode share in cities: exploration of actual and future trends with a worldwide survey. *Int. J. Auto. Tech. Manag.* 14, 203–216.
- Ben-Akiva, M., Walker, J., Bernardino, A.T., Gopinath, D.A., Morikawa, T., Polydoropoulou, A., 2002. Integration of choice and latent variable models. In: Mamassani, H. (Ed.), *In Perpetual Motion: Travel Behaviour Research Opportunities and Application Challenges*. Emerald Group Publishing, Bingley.
- Bahamonde-Birke, F., Kunert, U., Link, H., Ortúzar, J. de D., 2016. About attitudes and perceptions: finding the proper way to consider latent variables in discrete choice models. *Transportation* 44, 475–493.
- Bradshaw, C., 1992. Green transportation hierarchy: a guide for personal and public decision-making. Prepared for Ottawalk and the Transportation Working Committee of the Ottawa-Carleton Round-table on the Environment (Greenprint), Heart Health. Available from: <https://hearthealth.wordpress.com/>, Ottawa.
- Domarchi, C., Coeymans, J.E., Ortúzar, J. de D., 2018. Shared taxis: modelling the choice of a paratransit mode in Santiago de Chile. *Transportation* (doi.org/10.1007/s11116-018-9926-z).
- Gwilliam, K., 2002. *Cities on the Move: A World Bank Urban Transport Strategy Review*. The World Bank, Washington, D.C.
- Ortúzar, J. de D., Willumsen, L.G., 2011. *Modelling Transport*. John Wiley and Sons, Chichester.
- Rizzi, L.I., de la Maza, C., 2017. The external costs of private versus public road transport in the Metropolitan Area of Santiago, Chile. *Transp. Res. Part A: Policy Prac.* 98, 123–140.

Further Reading

- Booz Allen, 2012. *Integrating Australia's Transport Systems: A Strategy for An Efficient Transport Future*. Infrastructure Partnership Australia, Sydney.
- Cervero, R., Golub, A., 2007. Informal transport: a global perspective. *Trans. Policy* 14, 445–457.
- Haas, A.R.N., 2018. Key considerations for integrated multi-modal transport planning. Policy Paper, Cities that Work, International Growth Centre, UN Habitat, Nairobi.
- Litman, T., 2017. *Introduction to Multi-Modal Transport Planning, Principles and Practices*. Victoria Transport Policy Institute, Victoria.
- Pojani, D., Stead, D., 2015. Sustainable urban transport in the developing world: beyond megacities. *Sustainability* 7, 7784–7805.
- Rodrigue, J.P., Comtois, C., Slack, B., 2016. *The Geography of Transport Systems*. Routledge, Abingdon-on-Thames.

The Future of Road Transport

Preston L. Schiller, Department of Civil and Environmental Engineering, University of Washington (Seattle), Bellingham, WA, <United States.

© 2021 Elsevier Ltd. All rights reserved.

Introduction and Overview	306
The Future of Road Transport has an Interesting Past	306
What do Roads and Conveyances Mean to People? The Road as Metaphor and Symbol	307
The Road and Travel in Literature and Popular Culture	307
The Roads Frequently Taken	308
Road Transport Safety and Health	308
Major Modes, Energy, and Space Consumption	308
High-tech Proposals, Planning Proposals and Possible Scenarios	309
CAV Stampede or Creep?	310
Promising CAV Applications for Buses and Trucks	310
Could 3D Printing and Drones Replace Most Trucking?	311
No More Need for Roads as We Know Them: Is Diminution of Motor Vehicle Travel Possible?	311
Conclusion: Roads and Road Transport Face an Uncertain Future on a Planet Facing a Most Uncertain Future	312
Biography	313
See Also	313
Relevant Websites	313
References	314
Further Reading	314

Introduction and Overview

This chapter will examine a variety of aspects of roads and forms of transport they have sustained and might sustain in the future. It will explore what the history of road transport might tell us about its future as well as some of the cultural meanings that surround road transport. The term “road” has often been used interchangeably with “street” and “path.” This has been historically true and somewhat at present, this chapter will consider “road” to refer to infrastructure that is of sufficient size and construction quality to be able to accommodate large numbers of persons and/or vehicles, generally at a velocity higher than that of streets or paths. “Street” can be generally thought of as an improved passageways in towns and cities that may or may not accommodate large numbers of persons or vehicles. It may be an extension of a road connecting more than one city or region or it may be an urban road that is designed or designated to handle less traffic than those known as roads. “Path” generally refers to a narrower passage that may or not be improved by humans, probably would not accommodate the mass movement of people and may or may not accommodate vehicles. Many paths, some in use for millennia, have their origins in animal tracks and trails.

After examining the history and cultural meanings of roads and road transport the chapter moves to an exploration of the present situation of road transport. The expanse of transport infrastructure as well as the extent of travel are presented. Types of modes constituting the major forms of road transport, from feet to the emerging automated vehicle, are discussed. The chapter then examines the types of modes and technologies currently animating discussions about the direction in which road transport might go in the future. Special attention is paid to that mode and technology most animating these discussions, the connected and automated vehicle (CAV). Questions are raised whether certain rapidly developing technologies, including drones and 3D printing, might render much road transport redundant. Three scenarios are then presented, each representing very different speculations about our future mobility on and off the road.

The Future of Road Transport has an Interesting Past

Roads were a relative latecomer to human and freight mobility. The earliest ways that humans traveled were by water, since many early settlements were along shores, or as hunters and gatherers moving in pursuit of food, traversing open terrain and savannahs, or following paths created by animals. When animal paths were created by large herds, such as buffalo in North America, the resultant paths were wide enough for wagon travel, as discovered by early settlers. Roads generally did not emerge until more humans began to live in cities and fixed communities. Road development was pushed along by the rise of larger political entities; regional principalities, kingdoms and early empires, that had an interest in moving large numbers of persons from one part to another either for military or commerce purposes. An early form of the dynamic, which exists today began; roads connect towns and regions and facilitate their growth, and their growth contributes to increasing the use of roads.

Roads were important for military and commercial purposes. The Great Wall of China was as much of an elevated roadway wide enough for substantial troop movement as it was a wall to keep out the herds of sheep, which accompanied and sustained invaders from the north. The vast and elaborate Inca (sometimes Inka) Road system was designed wide enough in most parts to allow substantial troop movement, but also controlled against invaders through a series of complicated suspended rope bridges which could be raised when necessary. The extensive roads and bridges of Imperial Rome served the purpose of moving large numbers of troops at rapid paces, often built over difficult terrain. It is worth noting that parts of these, and other well-built roads of centuries long past, are still in use today. (Hindley, 1971; Hyslop, 1984; Lay, 1992; Derry and Williams, 1993; Schiller and Kenworthy, 2018)

The terms transportation and communication were used interchangeably in previous eras since the relay of important messages and documents, whether by scrolls or the *chasquis* (language knotted into ropes) of the Inca, were transmitted by messengers along these routes. The Inca messengers ran at their top speed between relay stations spaced about 1.4 km apart throughout the kingdom. Even after the discovery and spread of the wheeled vehicle and the use of animals for human mobility the majority of persons on roads and paths were on foot and not always easy or safe. The relatively short trips between nearby towns could be difficult and hazardous. Much more so for the longer trips undertaken on foot for commercial or religious purposes. Often long distance religious pilgrimages involved some buying and selling of goods as well. The logistics of road travel were complicated and shaped by climate, especially seasonal factors, terrain, the existence of water and nourishment—for both humans and the animals accompanying them. Political considerations such as areas of hostility or taxation (tribute paying) led to the development of alternate routes skirting these areas.

What do Roads and Conveyances Mean to People? The Road as Metaphor and Symbol

A road is not just a road and a vehicle is not just a vehicle. Each has symbolic and metaphorical meanings that reach deeply into the human psyche and shape perception and experience for individuals and society. In order to understand the road transport of the future it is necessary to examine these meanings. Railroads as a major form of transport and its infrastructure are not addressed in this chapter as they are beyond its scope.

The road has often been portrayed, intentionally or not, as a way to a better and freer future. This has been true in painting, where the road is often the part of a composition denoting a journey or the future. In literature the road and its conveyances might have multiple layers of meaning. Mechanical vehicles, including bicycles, automobiles and trucks have often been seen as the means to achieve a desired future. A good deal of the treatment of vehicles and mobility in popular media reflects popular notions of why and how humans should move; quickly, comfortably, and with priority over other modes. Forms of mobility also shape how individuals relate to others and society; whether or how they interact with others and whether they accept regulations and fees as part of a motorist's responsibility—or whether they reject such as infringements on their "freedom to drive."

The Road and Travel in Literature and Popular Culture

From Walt Whitman extolling the pleasures of the wanderer on foot in his poem "Song of the Open Road" to Kerouac's automobile-based odyssey *On the Road*, to Neal Cassaday (the Dean Moriarty of Kerouac's novel) driving Ken Kesey and his Merry Pranksters across the country in the bus named "Further," as well as *The Motorcycle Diaries* by Che Guevara, the road and its modes of travel have been portrayed as a means to self-discovery and personal freedom.

The road and its forms of transport, including walking, have often symbolized life choices and changes. Robert Frost's popular poem "The Road Not Taken" (1915) as well as the Beatles' crossing of Abbey Road upon the completion of their last album together (1969), signify that change lies ahead. Roads have been important historically, and at present, for long distance travel by persons dedicated to a religious practice or to solitary searching. Pilgrimages, whether for religious or secular motives, involve many millions of persons on foot around the world each year, solo, or as part of groups, for short or long distances. These may be carried out under physically or climatically grueling conditions. Individuals who have logged thousands of miles on foot include the 1960s activist Peace Pilgrim (Pilgrim, 1991) and the folksinger Art Garfunkel's walks across America as well as other parts of the world (Schiller and Kenworthy, 2018; Garfunkel, undated).

John Steinbeck's *The Grapes of Wrath* (1939) chronicles the plight of a family displaced by the Oklahoma dust storms of the 1930s struggling to get to a better future. Their old car, overloaded with kin and furniture, sputters along, breaking down frequently, while their entry into the "promised land" of California is beset by formidable setbacks, symbolizing the societal break-down and obstacles to getting out of the Great Depression. The cover photo of *The Road Ahead* by Bill Gates, Jr. (1995) places the author in the middle of a rural road that stretches far into the distance as a metaphor for the rosy future that awaits us as internet technologies advance. The imagery and symbolism of road transport figure prominently in popular music ranging from "Daisy Bell" ("Bicycle Built for Two") of 1892 and "In my Merry Oldsmobile" of 1905 to Woody Guthrie's 1940s and 1950s folksongs, to the more modern car-rock songs of Bruce Springsteen, the Beach Boys, to the paeans to trucking and truckers such as "Convoy," as well as many contemporary rap songs glorifying fast cars and racing.

Road travel by modes deemed to be "fast," such as automobiles are not always so fast when subjected to a more holistic analysis, as that done by Ivan Illich in *Energy and Equity* (1976) (Illich, 1976). Fast modes of travel, from cars to planes and trains, have also been seen as "destroyers of time and space" (Schivelbusch, 1986; Sachs, 1992). Whether destructive or not, modern modes,

beginning in the 19th century, have definitively shaped social and psychological notions of time and space. While the speed of travel has greatly increased during the past two hundred years, the average of amount of time most humans are willing to commit to travel has remained relatively constant at about one hour per day on an average while the distances and the range of travel have greatly expanded (Marchetti, 1994; Whitelegg, 2016). The language and terminology of road transport have also been adapted, literally or metaphorically, into ordinary usage. Karl Marx wrote of revolutions as “the locomotives of history,” and terms such as “arterial,” “interchange,” and “hard drive” demonstrate such borrowing.

The Roads Frequently Taken

In order to discuss road transport in the future it is necessary to understand it at present. Road transport and travel is a vital aspect of everyday life for most of the planet’s people. Even for persons who do not travel much, road transport is crucial for the delivery of goods and services to them. The expanse of transport infrastructure, travel, and vehicle ownership is huge. But there is great variation in the extent of infrastructure provision, vehicle ownership and type of vehicle from one continent to another and, within continents, from one country or city to another.

The extent of roads and road transport worldwide and the disproportionate share of the United States, which constitutes only a little over 4% of the world’s population, is highly instructive (Schiller and Kenworthy, 2018). Worldwide there are over 64 million km of roads, approximately two-thirds of which are paved. The United States accounts for a little over 10% of these—although that does not take into account the fact that there are many multilane roads in the United States. When these are calculated there are almost 14 million lane miles of roads. The world motor vehicle population is approximately 1.3 billion and growing rapidly; personal motor vehicles (PMVs) constitute about 90% of these. The U.S. accounts for approximately 20% of these. Worldwide, on average—with significant variation between wealthier and less wealthy countries, there are 160 PMVs per 1000 persons; in the United States, there are 810 PMVs per 1000 persons. This motorization of everyday travel, sometimes termed “hypermobility,” or “hyperautomobility” (Adams, 2005; Freund & Martin, 2009; Whitelegg, 2016) has implications that are far reaching and affect many domains of human existence. These include human health and safety; environmental factors such as energy consumption, air pollution and greenhouse gases (GHGs), and urban heat entrapment.

Road Transport Safety and Health

Worldwide over a million persons are killed and as many as fifty million are seriously injured each year as victims of road crashes. In the United States over 37,000 persons are killed and 2,746,000 are seriously injured each year as victims of road crashes. Pedestrians killed by motor vehicles account for 15% of these deaths and bicyclists killed by motor vehicles account for 2% of the total deaths. The rates of pedestrian and bicyclist deaths and injuries from vehicle crashes are increasing faster than other categories reported (Dora et al., 2011; ITF, 2017; Smart Growth America 2019b).

Air pollution from road transport contributes significantly to human mortality. According to the World Health Organization (WHO, 2019a) each year some 4.2 million persons die prematurely from the effects of air pollution. The major pollutants include particulate matter (PM), ozone (O₃), nitrogen dioxide (NO₂), and sulfur dioxide (SO₂), all closely associated with road transport emissions and effects such as particulation from tires, brakes, and road surfaces—especially asphalt. Those living close to heavy traffic, especially the young and the elderly, are among those most affected and suffer high rates of asthma.

Particulates from tire and brake wear as well as tailpipes, and fluid drippings from motor vehicles also pollute bodies of water through road runoff. Emissions, such as the nitrates of oxygen (NOX) can be absorbed by bodies of water and can lead to acidification of these waters. Motor vehicle emissions, especially when smog-forming, also contribute significantly to acid deposition, or acid rain, which is very harmful of vegetation; especially forested and agricultural areas.

The transport sector, worldwide, is 95% dependent on fossil fuels for energy and the sector and its demands are growing rapidly. Motor vehicle emissions from road transport are a significant contributor to climate change and GHGs accumulation in the atmosphere. Transport constitutes about one-quarter of worldwide GHGs, with road transport as the source of about three-fourths of these; urban transport accounts for about 40% of these and about 40% of end-use energy consumption (WHO, 2019b).

In addition to pollution from emissions, transportation infrastructure itself has impacts which affect climate change such as the albedo and urban heat island effects which refer to the degree to which solar radiation reaching the earth’s surface and its structures is absorbed (and stored as heat in pavements, building facades, and rooftops) or reflected back and not absorbed. (Kelbaugh, 2019)

Major Modes, Energy, and Space Consumption

The major modes of road transport worldwide at present are: walking; cycling (including a growing number of electric bicycles or e-bikes), motorcycles, PMVs which includes standard automobiles, light pick-up trucks and sports utility vehicles (SUVs); transit vehicles; commercial and freight vehicles ranging in size from delivery vans to 18-wheeled tractor-trailer combination trucks. External energy consumption per kilometer by mode ranges from none for walking and cycling; to the relatively small amount of energy consumed by e-bikes and electric scooters, to a fair amount consumed by even the most energy efficient standard



Figure 1 Massive freeway interchange near Toronto, but could be almost anywhere in a North American metropolitan area. Note the separation of uses, lack of pedestrian and cycling connections between quadrants and the great distances between uses. *Source: Preston L. Schiller.*

automobiles, to large amounts for small trucks and SUVs, to very large amounts for large trucks (Vanek et al., 2014). When per capita energy consumption by mode is examined the personal vehicle modes are far less efficient than even transit modes at less than capacity. When energy consumption per unit of weight-distance is examined it is found that freight rail modes are greatly more efficient than road freight modes. The least energy efficient freight mode is air transport. (Schiller and Kenworthy, 2018)

The over-commitment to automobility and the support for the ever-expanding infrastructure devoted to sustaining it, also has significant impacts on the ways in which urban and suburban space is allocated and developed. Motor vehicles take up a fixed amount of area when parked—as they are for 90%–95% of the time, and an even greater amount while in motion. A technique for assessing the extent to which a specific road transport mode consumes urban space has been developed; the time-area concept, which demonstrates the great advantage that transit, walking and cycling have for preserving urban space, strengthening commercial districts, and fostering compact development (Bruun 2014) (Fig. 1).

“The future, like everything else, is no longer quite what it used to be.”

(Paul Valery, 1937)

Hands, feet, and attention-free driving; high velocity and safe spacing remotely controlled from a distant central authority; highways and interchanges engineered for maximum speed, separation of overlapping, or interconnecting roads from each other, and quick and easy connections with other roads; the relegation of pedestrians to separated way: Vanished are traffic frustrations and worries over getting to a destination on time—is this a current depiction of the near future of road transport we are entering? No, this is the world envisioned for 1960 in General Motors Futurama, Highways and Horizons exhibition viewed by thirty million persons at the 1939 World’s Fair at New York City’s Flushing Meadows (General Motors Corporation, 1940). Thanks to General Motors’ Motorama (1956) and its heavy influence over the creation of the national interstate highway system, begun in 1956, there were now plenty of suburbs for sleeping, modernist office towers for working in, and a huge network of separated multilane expressways crisscrossing the country side and traversing cities of all sizes along the way. Futurama II did nod at freight and passenger rail improvements giving them a little space in the median of expressways (Mayersohn 2014), especially those exploiting the resources of natural areas such as the Amazon (nywf64.com). But was this the bright future that had been expected? And will a number of modes undergoing development and refinement at present live up to their rosy expectations?

High-tech Proposals, Planning Proposals and Possible Scenarios

The future of road transport depends upon the outcome of the contest over which modes might prevail, the types of infrastructure that might be available to support these modes, the types of land uses that will dominate towns and cities, as well as the types of freight movement that emerge in reaction to these factors. The patterns of industrial and consumer demands that emerge in the future will also play a key role. All of these together may shape a future that is quite different from that which seems likely at present. The modes that are dominant or emerging now need to be examined in order to ascertain the likelihood that they will shape the future of road transport.

Issues around these factors include: How likely is the technology to emerge as promised or currently described? Who will likely use this technology and to what purposes? What is the likelihood of the technology’s deployment? What might some of the new

technology's unintended consequences be? How much public resistance will be generated in reaction to these changes? At present there exists considerable tension between promoters of new types of vehicles, often ones that greatly increase speed and reduce human control, and social movements, such as "complete streets," and "active transportation" which aim to reduce motor vehicle dependence through increasing the shareability and safety of urban streets ([Smart Growth America 2019a](#)).

CAV Stampede or Creep?

The mode that is dominating discussions about the future of road transport is the autonomous vehicle (AV) or the CAV. In its most complete form the AV would not be dependent on a human driver but to achieve this it would need to be telecommunications connected so as to recognize and interact with other AVs, the infrastructure it encounters or uses, as well as detecting or interacting with pedestrians, cyclists, scooters, and dogs and cats using the roads upon which it is traveling. For purposes of our discussion we will use the term CAV. Most depictions of CAVs indicate that they will be fully electrified which would simplify their engines and drive systems considerably, significantly reduce air pollution—at least locally, if not regionally, where power generation is not based on fossil fuels, and possibly reduce energy consumption on a per unit basis. Whether a transition to a 90% or 95% electrified CAV fleet would reduce overall energy consumption would depend on the extent to which it is used. Most analyses indicate that CAVs would lead to more driving overall ([Martin, 2019](#)).

The slow creep toward automating cars has been occurring since cruise control was first introduced in 1958 and came into wider use in the 1960s. Several semiautomated and semiconnected features have since been introduced including braking assistance, anti-swerve assistance, lane roaming alerts, parking assist, and computers under the hood—that can be hacked by mischievous experts ([Greenberg, 2015](#)).

The CAV, whether fully implemented as its promoters project, or whether only partially implemented through "AV creep" has significant implications for personal and freight mobility. There is wide-ranging discussion about whether the projected benefits of the CAV are mostly "hype" ([Schiller, 2016](#)) or of value if done carefully with new forms of transit applications ([Grush & Niles, 2018](#)) or likely but uncertain about its impacts ([Martin, 2019](#)). Some are already accepting the CAV as inevitable and suggesting ways in which such vehicles need to be designed to be least harmful and how our cities need to be redesigned to accommodate CAVs while protecting other street users and helping transit ([NACTO, 2019](#)). Unabashed promoters of the CAV for personal mobility include the automobile and high tech industries as well as several marketing interests. An internationally prominent marketing and opinion research firm has made grandiose claims for its future; promising that it will not only be more safe but that it will create more "me time," save money, reduce stress, and improve users' health. ([Ipsos MORI, 2016](#)). The promoters conveniently overlook the sobering implications of the origins of AV research; it was initiated and supported by the Defense Advanced Research Projects Administration (DARPA) beginning in 2004 as part of its interest in fostering the robotic battlefield of the future ([Schiller 2016](#)). Given the challenges facing CAVs, including whether or not they will or can be as fully implemented as their promoters claim, the reality may be one of creeping connectivity and autonomy.

Promising CAV Applications for Buses and Trucks

The use of automated controls for group transport vehicles in separated guide ways has been well-established in rail transit and people movers for several decades. There are some CAV applications that probably can be implemented in the near term which do not face as many of the issues raised by their wide-scale use for personal mobility. CAVs could probably be introduced into some road systems that are separated from general traffic and would not have to share their road space with other vehicles, including CAVs used for personal mobility. Some separated and access-controlled routes could be automated as is being done with the automated shuttle between the Gare d'Austerlitz and Gare de Lyon train stations in Paris. A number of other experimental AV bus uses have been introduced into relatively easy to control environments such as college and corporate campuses. One issue that has come to the fore is the very slow velocity, about the same as brisk walking, of the vehicles, which is a manifestation of the relatively underdeveloped state of the art of the sensor systems. In other locations, such as a separated busway, a full-sized bus could be both electric powered and automated. This is already happening in some cities in China and could be implemented in a number of North American busways.

The automation and electrification of the large truck fleet could have major implications for energy consumption, truck transport logistics and costs, and air quality ([Chottani et al., 2018](#)). Some of these, such as fuel and pollution reductions, could be beneficial, especially through electrification ([Martin 2019](#)). The automation of transit and freight vehicles would have serious implications for job displacement, since there are millions of persons employed as drivers, warehouse and others servicing the transit and trucking industries. While large trucks and buses are a relatively small portion of the total motor vehicle fleet, well under 10%, they consume fuel at a disproportionately high rate, as much as 20% of transportation fuel and are a major contributor to air pollution, especially particulate pollution, since almost all large trucks and most buses are diesel-fueled.

Large trucks that operate on controlled access highways could be a target for connected automation and electrification. Autonomous trucks (ATs) are currently in experimental trials, with human monitors on board in several locales, including 650 miles of I-10 between Texas and Southern California. Although the state of battery technology is advanced enough to project the possibility that all PMVs could be electrically powered in the future, it will likely remain insufficient for powering large trucks for many years. In controlled rights of way, and perhaps certain urban freight corridors, it might be possible to power such vehicles in the same way that most streetcars and light rail transit vehicles are powered; through an overhead wire or catenary system or possibly,

though less likely, a third rail—as some trains use in urban and airport transit. Hybrid and hydrogen fuel cell technology might be able to act as a bridge between present fossil energy sources for large trucks and their future use of battery power.

Trucking fleet automation could have significant impacts on freight logistics; it may accelerate the pace of “Just In Time” (JIT) delivery as well as displace a great deal of freight from rails onto rubber-tired vehicles due to its probable lower costs—perhaps as much as 40% lower (Tryggestad et al., 2017). This is a future projected from current realities and foreseeable technological improvements. But what might transpire if the need for commodity transport by road were to diminish considerably?

Could 3D Printing and Drones Replace Most Trucking?

If the technology of 3D printing were to continue to develop rapidly could most products, from small widgets to large buildings, be produced onsite? What might happen if most packaging, a major source of the bulk of transported commodities, were eliminated as a consequence of 3D printing? What might happen if the combination of 3D printing and drones replaced transfer warehouses, delivery vans and medium-sized trucks for local and regional deliveries? Would cargo containers, on land or sea, no longer be needed? What might happen to freight transport by road if a strong “3-Rs” ethic took hold such that citizens truly reduced consumption dramatically, reused to a much greater extent than at present, and only recycled when necessary? What might happen if 3D printing were developed to be able to use raw materials close at hand rather than relying on long-distance or even regional-local bulk commodity truck transport? (Sauerwein and Doubrovski 2018)

No More Need for Roads as We Know Them: Is Diminution of Motor Vehicle Travel Possible?

Road and street development in the United States in the post-World War II era, stressed maximum accommodation of PMVs and minimal or no accommodations for bicyclists and pedestrians. More recently concerns over the environmental and personal and population health effects of over-dependence on motor vehicles for personal mobility have been increasing. This has prompted the development of modifications or new forms of road transport infrastructure, also termed “complete streets,” (Smart Growth America, 2019a) as well as the greater use of smaller forms of mobility for some travel, such as various types of bicycles, cargo bicycles, and scooters.

When streets and roads are designed as “complete streets” with safe pedestrian and cycling features, the travel share of these “micromobility” modes often grows sizeably (Shaheen et al., 2019). When traffic calming, road diets, and pedestrianization are introduced commercial and retail zones often thrive more than before and downsizing the urban freight delivery fleet becomes a possibility. Various types of smaller right-sized delivery vehicles appear, from very small vans towing small trailers to various types of cargo bikes. Many are electric powered and some delivery services are shared by businesses so that only one service delivers to as many as several hundred enterprises in the same area, as is the situation for Gothenburg, Sweden (Eriksen, 2015). There are ways of addressing mobility, local, and long distance, including the conversion of existing excess road space to public transport, that would reduce environmental and social impacts without imprisoning us within the CAV or other questionable or harmful modes (Bruun & Givoni, 2015) (Figs 2–4).



Figure 2 San Jacinto Street, Trianna District, Seville, Spain, 2009: Traffic chaos before street redesign, man in center is an informal parking director. *Source:* Preston L. Schiller



Figure 3 San Jacinto Street, Trianna District, Seville, Spain 2015: Traffic separations after street redesign; separated sidewalk and cycle path, separated motorcycle and automobile parking, lanes reduced to one. *Source: Preston L. Schiller.*



Figure 4 San Jacinto Street, Trianna District, Seville, Spain: Several blocks approaching the Puente de Isabel II bridge are pedestrianized, traffic is diverted to parallel streets *Source: Preston L. Schiller*

Conclusion: Roads and Road Transport Face an Uncertain Future on a Planet Facing a Most Uncertain Future

Each of the three scenarios described earlier; full on CAV, CAV creep, or diminution of road transport, is at least partially rooted in existing trends. Whether any one of these prevails, or a hybrid of these, or, as Monty Python would have it, might it be “time for something completely different,” will only be revealed in years and decades to come. Ultimately, the future of road transport will depend on the shape in which the perennial struggle between the faster-further-riskier (personally and environmentally) wishes of those yoked to PMVs and CAV romance and those of aspirants to a slower-nearer-safer active transportation urging is resolved. Faster-further-riskier aspirants would likely welcome a CAV dominated future. Slower-nearer-safer advocates would likely favor complete streets, lane reductions and pedestrianization scenarios that encourage walking, cycling, scootering, and transit as well as vastly improved intercity public transport options. In a time of pandemics and increasing climate disruption, travel and JIT logistics might lose some of their necessity or allure and localization could become much more attractive, desirable, and necessary.

Biography



Preston L. Schiller, PhD, is a faculty member in the Master's of Sustainable Transportation program at the University of Washington, Seattle. He has taught at several universities, including the planning program of Queen's University, Kingston, ON. His interests have ranged from public health to autonomous vehicles. He has been an advocate for transportation and transit policy and planning reforms at local, state and federal levels in the United States and Canada.

See Also

The Future of Urban Freight; How will Autonomous Vehicles Impact Car Ownership and Travel Behavior; Road Diets; Shared Space; Just-in-Time; Autonomous Goods Transport; Bicycles for Urban Freight; Transport Modes and Globalization; The History of Road Transport; Active Travel; Ageing Infrastructure; Car-Free Cities

Relevant Websites

www.transportenvironment.org
www.its.dot.gov/automated_vehicle/
<https://smartgrowthamerica.org/>
www.nacto.org
<http://www.eco-logica.co.uk/worldtransport.html>
www.streetsblog.org
www.itdp.org
www.bts.gov
www.cascadiacabs.com

References

- Adams, J., 2005. Hypermobility: a challenge to governance. In: Lyall, C., Tait, J. (Eds.), *New Modes of Governance: Developing an Integrated Policy Approach to Science, Technology Risk and the Environment*, Ashgate, Aldershot.
- Bruun, E., 2014. *Better Public Transit Systems: Analyzing Investments and Performance*. Routledge, London and New York.
- Bruun, E., Givoni, M., 2015. Sustainable mobility: Six research routes to steer transport policy. *Nature* 523 (7558), 29–31.
- Chottani, A., Hastings, G., Murnane, J., Neuhaus, F., 2018. Autonomous trucks disrupt US logistics. McKinsey & Company; Travel, Transport & Logistics. <https://www.mckinsey.com/industries/travel-transport-and-logistics/our-insights/distraction-or-disruption-autonomous-trucks-gain-ground-in-us-logistics#December>.
- Derry, T., Williams, T., 1993. *A Short History of Technology: From the Earliest Times to A.D. 1900*, Dover, New York, NY.

- Dora, C., Hosking, J., & Mudu, P., 2011. Urban Transport and Health: Module 5g; Sustainable Transport: A source book for policy-makers in developing cities. Federal Ministry for Economic Cooperation and Development (BMZ). Available from: https://www.who.int/hia/green_economy/giz_transport.pdf.
- Eriksen, L., 2015. The innovative delivery system transforming Gothenburg's roads. *The Guardian*, Nov. 18. Available from: <https://www.theguardian.com/cities/2015/nov/18/innovative-delivery-system-transforming-gothenburg-roads>.
- Freund, P., Martin, G., 2009. The social and material culture of hyperautomobility: Hyperauto. *Bullet. Sci. Technol. Soc.* 29 (6), 476–482.
- Garfunkel, A., (undated) Walks. Available from: <https://www.artgarfunkel.com/walks.html>.
- General Motors Corporation, 1940. Futurama (exhibition pamphlet), New York (State).
- Greenberg, A., 2015. Hackers remotely kill a jeep on the highway—with me in it. <https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/>
- Grush, B., Niles, J., 2018. *The end of driving: Transportation systems and public policy planning for autonomous vehicles*, 1st. Elsevier.
- Hindley, G., 1971. *A history of roads*. Peter Davies, London.
- Hyslop, J., 1984. *The Inka Road System*. Academic Press, Orlando, FL.
- Illich, I., 1976. *Energy and Equity*. Marion Boyars, London.
- I.T.F., 2017. *Road Safety Annual Report 2017*. OECD Publishing, Paris, doi:10.1787/irtad-2017-en.
- Ipsos MORI, 2016. The future of driving: Five ways connected cars will change your life. <https://www.ipsos-mori.com/newsevents/ca/1776/The-future-of-driving-Five-ways-connected-cars-will-change-your-life.aspx>
- Kelbaugh, D., 2019. *The urban fix : Resilient cities in the war against climate change, heat islands and overpopulation*. Routledge, New York.
- Lay, M., 1992. *Ways of the World: A history of the world's roads and of the vehicles that used them*. Rutgers University Press, New Brunswick, NJ.
- Marchetti, C., 1994. Anthropological invariants in travel behavior. *Technol. Forecast. Soc. Change* 47 (1), 75–88.
- Martin, G., 2019. *Sustainability prospects for autonomous vehicles: Environmental, social and urban*. Routledge.
- Mayersohn, N., 2014. Auto Nirvana (slide show of New York World's Fair, 1964-1965). Available from: <https://www.nytimes.com/slideshow/2014/04/20/automobiles/13fair-slides/s/13fair-slides-slide-91Y1.html>.
- NACTO, 2019. *Blueprint for autonomous urbanism*, 2nd edition. Available from: www.nacto.org.
- Pilgrim, P., 1991. *Peace Pilgrim: Her Life and Work in Her Own Words*. Ocean Tree, Santa Fe, NM.
- Sauerwein, M., Doubrovski, E.L., 2018. Local and recyclable materials for additive manufacturing: 3D printing with mussel shells. *Mater. Today Commun.* 15, 214–217, <http://www.sciencedirect.com/science/article/pii/S2352492817301046>.
- Schiller, P.L., 2016. Automated and connected vehicles: High tech hope or hype? *World Trans. Policy Prac.* 22 (3), 28–44.
- Schiller, P.L., Kenworthy, J.R., 2018. *An introduction to sustainable transportation : Policy, planning and implementation*, 2nd Ed. Routledge.
- Schivelbusch, W., 1986. *The Railway Journey*. University of California Press, Berkeley, CA.
- Shaheen, S., Cohen, A., Chan, N., Bansal, A., 2019. Sharing strategies: carsharing, shared micromobility (bikesharing and scooter sharing), transportation network companies, microtransit, and other innovative mobility modes. In: Deakin, E. (Ed.), *Transportation, Land Use, and Environmental Planning*, Elsevier, pp. 237–262.
- Smart Growth America, 2019a. *National Complete Streets Coalition*. Available from: <https://smartgrowthamerica.org/program/national-complete-streets-coalition/>.
- Smart Growth America, 2019b. *Dangerous by design*. Available from: <https://smartgrowthamerica.org/app/uploads/2019/01/Dangerous-by-Design-2019-FINAL.pdf>.
- Tryggstad, C., Sharma, N., van de Staaij, J., & Keizer, A., 2017. New reality: electric trucks and their implications on energy demand. *McKinsey Energy Insights*. September, Available from: <https://www.mckinseyenergyinsights.com/insights/new-reality-electric-trucks-and-their-implications-on-energy-demand/>.
- Vanek, F.M., Angenent, L.T., Banks, J.H., Daziano, R.A., Turnquist, M.A., 2014. *Sustainable Transportation Engineering Systems*. McGraw-Hill Education, New York.
- Whitelegg, J. 2016. *Mobility: A new urban design and transport planning philosophy for a sustainable future*. Createspace.
- WHO, 2019a. *Ambient air pollution: Health impacts*. <https://www.who.int/airpollution/ambient/health-impacts/en/>
- WHO, 2019b. *Climate Impacts*. Available from: <https://www.who.int/sustainable-development/transport/health-risks/climate-impacts/en/>.

Further Reading

- Bruun, E., 2014. *Better Public Transit Systems: Analyzing Investments and Performance*. Routledge, London and New York.
- Dora, C., Hosking, J., Mudu, P., 2011. Urban Transport and Health: Module 5g; Sustainable Transport: A source book for policy-makers in developing cities. Federal Ministry for Economic Cooperation and Development (BMZ). Available from: https://www.who.int/hia/green_economy/giz_transport.pdf.
- Grush, B., Niles, J., 2018. *The end of driving: Transportation systems and public policy planning for autonomous vehicles*, 1st Ed. Elsevier.
- Illich, I., 1976. *Energy and Equity*. Marion Boyars, London.
- Marchetti, C., 1994. Anthropological invariants in travel behavior. *Technol. Forecast. Soc. Change* 47 (1), 75–88.
- Mumford, L., 1963. *The Highway and the City*. Harcourt, Brace. New York.
- Newman, P., Kenworthy, J., 2015. *The end of automobile dependence: How cities are moving*. In: *Away from Car-Based Planning*, Island Press, Washington, DC.
- Sachs, W., 1992. *For Love of the Automobile: Looking Back into the History of Our Desires*. University of California Press, Berkeley, CA.

Road Transportation and Territorial Scale

Ana M. Condeço-Melhorado, Complutense University of Madrid Geography Department, Madrid, Spain

© 2021 Elsevier Ltd. All rights reserved.

Glossary	315
Accessibility Impacts	315
Land Use Changes	316
Territorial Cohesion	316
Road Network Vulnerability	318
Further Considerations on the Magnitude of Impacts and Spillover Effects of Road Infrastructure	318
Relevant Websites	319
References	319
Further Reading	319

Glossary

Cost-benefit analysis: Cost/Benefit Analysis is a technique for deciding whether to make a change. As its name suggests, it compares the values of all benefits from the action under consideration and the costs associated with it.

Generalized transport cost: It is the sum of all costs involved in a trip. For a business trip, this can involve the value per hour and the labor cost to the firm, as well as costs associated with the trip itself such as tickets, tolls, and vehicle maintenance.

Urban sprawl: Decentralization and suburbanization, also known as sprawl, has been associated with lower residential densities that cannot be efficiently serviced by mass transit systems outside specific corridors. Under such circumstances, transit is limited to a part of the city and citywide services become prohibitive.

Hub-and-spoke networks: These are networks serving a range of geographically dispersed locations through a central hub. This is generally more efficient, rather than operate direct services between locations.

Resilience: For a transportation system, resilience is the capability to recover from a disruption to an operational level similar to prior to the disruption in a timely manner. The longer and deeper the impact of the disruption on operations, the less resilient a transport system is.

Network effect: For transport networks, network effects are effects in other parts of the transport system as a result of a transport improvement.

Production functions: Production function estimates the maximum set of output(s) that can be produced with a given set of inputs. Use of a production function implies technical efficiency. Some of these functions include the transport capital stock as an input variable.

Trans-European transport networks (TEN-T): European Union (EU) program for the construction and upgrade of transport infrastructure across the EU, for all transport modes plus logistics and intelligent transport systems.

Accessibility Impacts

Accessibility can be defined as the ease of access to economic and social opportunities from a given location, by means of a transport system. The level of accessibility of a given place depends mainly on two factors: (1) its location, considering the spatial distribution of social-economic activities and (2) the quality of existing transport infrastructures. Places that have a central place in the economic system are more accessible than remote and peripheral areas. On the other hand, new and efficient road infrastructures, such as motorways, increase the accessibility of places by reducing travel times, transport costs, and increasing connectivity.

Improvements in road networks entail much more than a reduction of transport costs. They are associated with increased accessibility to labor markets, knowledge transfers between firms, and higher productivity levels. They are also related to regional development and the appearance of specialization and economies of scale. From a social perspective, road transport is essential to ensure access to basic services such as healthcare, education centers, or job opportunities. Consequently, increasing the accessibility of places is one of the main objectives of many transport projects/plans.

Accessibility can be measured with a wide variety of indicators (**Table 1**), which usually include two components:

- The cost of transport, expressed as distance, time or generalized transport costs.

Table 1 Most commonly used accessibility indicators.

Weighted average travel times	Shows the average travel time to a set of relevant destinations. Where A_i is the accessibility of node i , t_{ij} is the travel time by the minimum-time route in the network between node i and destination j (in minutes) and m_j is the mass (activity volume) of destination j .
$A_i = \frac{\sum_{j=1}^n t_{ij} m_j}{\sum_{j=1}^n m_j}$	
Economic potential	Shows the weight of accessible markets, discounted by the distance necessary to reach them. The closer to bigger markets the higher the economic potential of a given location. Where P_i is the economic potential of node i , m_j and t_{ij} are already known and α reflects the distance decay function usually calibrated with data from mobility surveys.
$P_i = \sum_{j=1}^n \frac{m_j}{t_{ij}^\alpha}$	
Cumulative opportunities	Counts the number of opportunities available to a certain location, considering a specific distance threshold. Where CO_i is the cumulative opportunities of node i , m_j is already known and δ_{ij} takes the form of 1 if t_{ij} is below a certain time threshold and 0 otherwise.
$CO_i = \sum_{j=1}^n m_j \delta_{ij}$	

- The attraction capacity of destination places that will depend on the study objective. Studies looking at accessibility to services may include the location of, for instance, schools, healthcare centers, and supermarkets. If the interest is to measure access to jobs or consumer markets, employment or population variables could be used.

A common practice for evaluating the accessibility impacts of road projects/plans is to compute the accessibility indicators for at least two scenarios: (1) a scenario showing the network after the road improvements and (2) a baseline scenario showing the road network before the investments. The comparison between both scenarios will uncover the areas where accessibility changes and impacts will be higher.

Fig. 1 shows the accessibility impacts of a hypothetical scenario of implementing tolls in the Spanish highway system. Notice that accessibility impacts are negative in this case, since the introduction of tolls will increase transport costs.

Land Use Changes

A change in road transport infrastructure may change the spatial distribution of land uses (Fig. 2). Increased accessibility levels brought about by improvements in road infrastructures may attract economic activities to the areas that benefit the most from these improvements. Some paradigmatic examples are shopping malls and logistic firms, usually located at highly accessible places such as road network intersections.

However, some studies point out potential negative effects associated with the construction of new roads, since some of the activities attracted to the area may be in fact relocations from less accessible regions.

As for the effects of road infrastructure in residential areas, some studies conclude that the decrease of transport costs and the promotion of private cars, linked to the extension of the road network in cities and suburbs, are related to urban sprawl. The land use changes caused by lower travel costs and greater accessibility can be controversial. On the one hand, roads are essential for economic and social activities to flourish, while on the other hand, they may lead to a relocation of economic activities and the dispersion of residential areas. In both cases, the construction of new roads may result in changes to activity patterns that can cause higher congestion levels due to the concentration of activities in certain places and to the increase of trip length because of urban sprawl.

Territorial Cohesion

Transport infrastructure and especially roads are essential to reducing regional inequalities and promoting territorial cohesion. Road networks are the most extended, connecting places that are not easily accessed by other modes. Increasing the accessibility of peripheral and rural areas to markets and essential services partly compensates for their remoteness and peripheral location.

The concept of territorial cohesion has been introduced in political agendas and in the impact assessments of transport policies, mainly since 2007 when it was included in the EU's Lisbon Treaty. The spatial distribution of accessibility is one of the variables used to measure the impact of transport plans on regional disparities and territorial cohesion. Dispersion and inequality distribution measures, such as the variation coefficient or the Gini coefficient, are some of the indicators used to analyze the level of territorial cohesion (López et al., 2008). When analyzing the impact of road network improvements on territorial cohesion, several studies compare the distribution of accessibility for the scenarios before and after a project. If the work has an impact on a limited area, and especially if improvements benefit places that are already very accessible, disparities between the most and the least accessible places will increase and the territorial cohesion impacts of such projects will be negative. If, however, road network improvements are spread out and benefit remote areas, the accessibility differences will be reduced and territorial cohesion will improve.

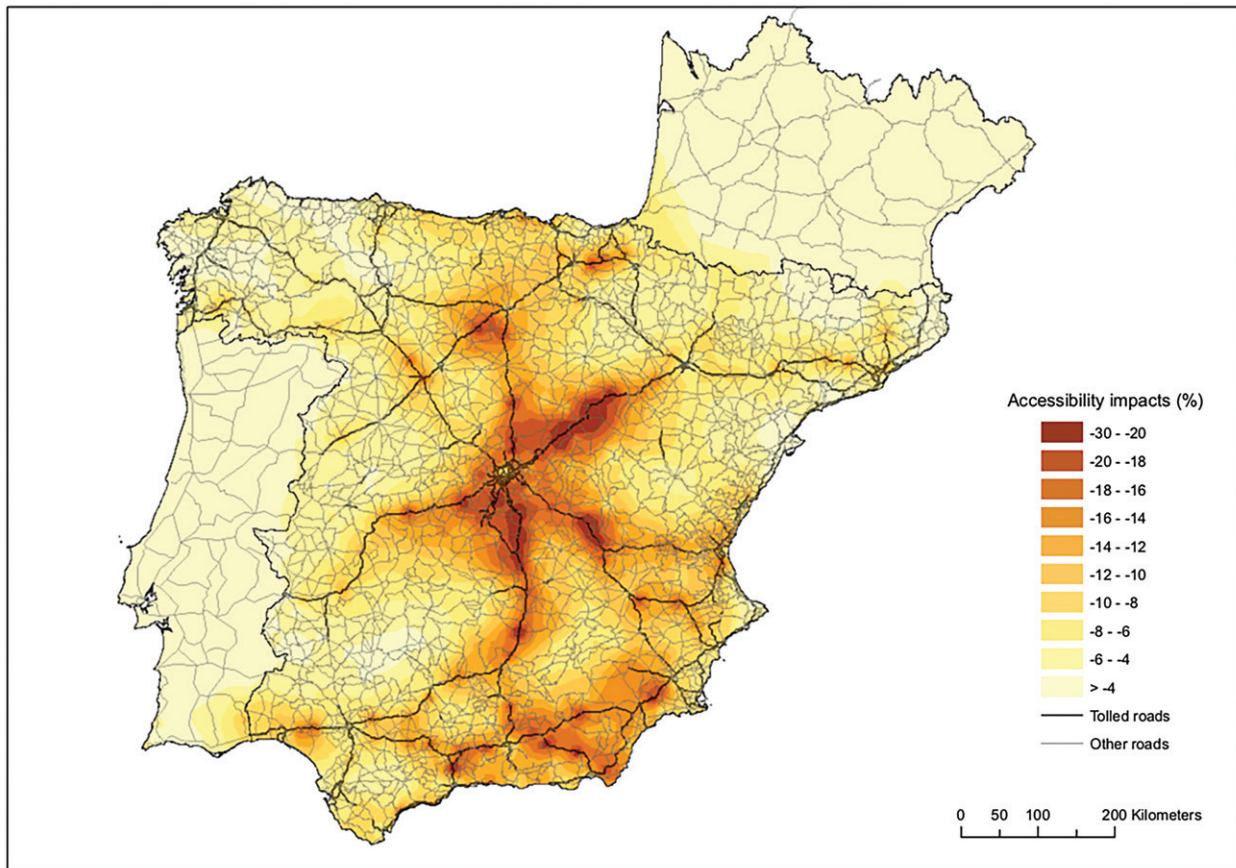


Figure 1 Accessibility impacts of introducing road charges in the Spanish highway system. Source: *Condeço-Melhorado A et al. (2011)*.

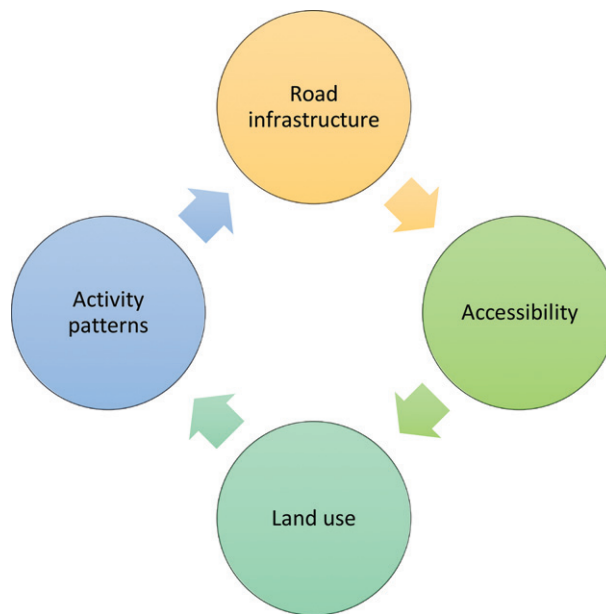


Figure 2 Interaction between road infrastructure and land-use.

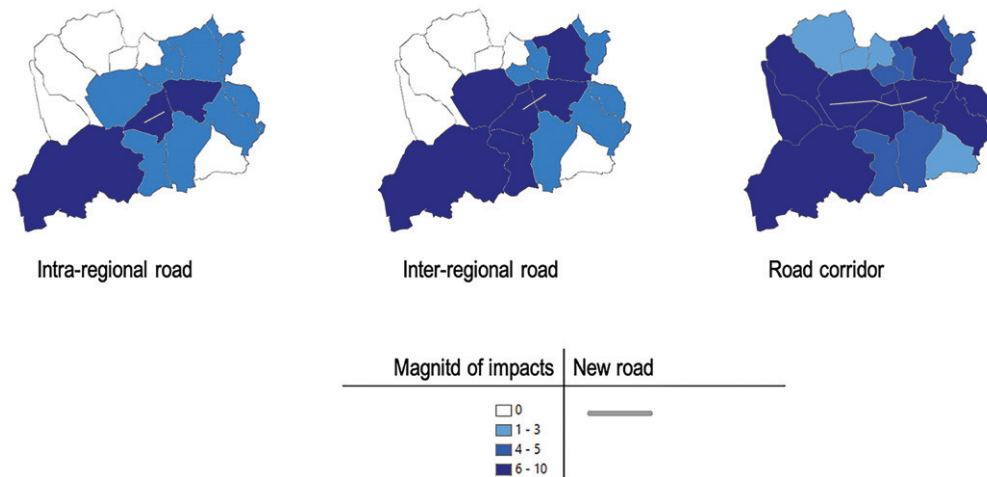


Figure 3 Territorial variation of spillover effects. Source: Own elaboration, based on [Laird et al. \(2005\)](#).

Road Network Vulnerability

The vulnerability of roads can be defined as the impacts associated with system failures (i.e., congestion, car accidents, earthquakes). The road structure contributes to different network vulnerability levels. Distributed, mesh-like networks are more resistant to random failures of nodes and links than centralized, hub-and-spoke networks. Furthermore, road links will be more or less vulnerable to traffic disruptions depending on their traffic flows and on the availability of good alternative routes ([Jenelius, 2009](#)). Studies show that consequences of road closures are higher in densely populated areas, where traffic flows are greater. At the same time, regions with denser road networks will be more resilient to any potential traffic disruption, as there will be more alternatives for traffic diversions. Accessibility indicators, such as those seen above, have been used to assess network vulnerability since they can explicitly consider the interplay between a system failure and the travel behavior of road users.

Vulnerability analyses are used to show the most vulnerable and critical links in a network. Critical links are identified as those causing the highest traffic delays/costs in the eventual case of a disruption. These analyses provide interesting insights into transport agencies regarding the improvement of network robustness and resilience to failure, through better network design and emergency plans.

Further Considerations on the Magnitude of Impacts and Spillover Effects of Road Infrastructure

The level of road infrastructure impacts depends on the quality and quantity of the existing infrastructure. Accessibility, land use, and territorial cohesion impacts will be particularly high in regions with less network density or where road network improvements lead to the elimination of existing bottlenecks. On the other hand, the effects of investments in regions with dense and efficient road network will be marginal. A similar situation occurs regarding the effects on the vulnerability of road networks since the impact caused by adding extra links to an already dense network will be residual.

Another consideration when measuring the territorial impacts of road infrastructure are spillover effects. Roads generate impacts that go beyond the regional borders where those roads are located. Improvements on the road network of a certain region may benefit distant regions, due to the network effect ([Laird et al., 2005](#)). Closely linked to the concept of network effect is the concept of spillover effect, understood as the benefits received by a region due to the infrastructure built in a different region. Spillover effects have traditionally been measured with production functions, based on the work of Aschauer ([Aschauer, 1989](#)), who found that the transport infrastructure plays a determinant role in GDP. More recently, some studies have applied accessibility measures to analyze spillover effects, arguing that transport capital stock figures fail to capture the usage of transport networks made by existing traffic flows ([Gutiérrez et al., 2010](#),). Accessibility measures have the advantage of considering the characteristics of road networks and their role in reducing distances and bringing economic agents closer. These studies show that spillover effects depend on the characteristics of the existing road infrastructure as displayed in [Fig. 3](#) and summarized further.

- A road located in the interior of a country/region will primarily benefit national/regional economic centers, and the impacts will be higher inside the country/region. However, accessibility benefits may spill over to neighboring countries/regions, as they use this infrastructure in their commercial relations.
- In the case of a new road located between two countries/regions, spillovers will be higher, given that more international/interregional connections will benefit.

- Spillover will be even higher in the case of road corridors crossing different countries/regions, since they will benefit a greater number of international/interregional relations.

Spillover effects have been measured to assess Transport Master Plans, the construction of new road corridors such as the Trans-European transport networks (TEN-T), or to evaluate the introduction of road charges. While spillovers are expected to be positive in the case of improved accessibility levels, they can also have a negative sign. If, for example, tolls are applied, transport cost will increase. Another example of negative spillovers lies in the increased congestion levels that may cross regional borders and cause delays in neighboring regions. Spillover effects have also been measured through the lens of territorial cohesion, as they will potentially change the distribution of opportunities across the space, benefiting regions that depart from a poorer situation (progressive spillovers) or, on the contrary, benefiting already well-served regions (regressive spillovers). Given the spatial nature of spillover effects, they can also be measured in terms of land use changes or network vulnerability, although studies in this respect are very limited.

Relevant Websites

Accessibility and transport appraisal: <https://www.itf-oecd.org/accessibility-and-transport-appraisal-roundtable>
 The Geography of Transport Systems Book: <https://transportgeography.org/>
 ESPON – European Territorial Observatory Network: <https://www.espon.eu/>
 Accessibility Observatory – University of Minnesota: <http://ao.umn.edu/>
 Online Transportation Demand Management Encyclopedia: <https://www.vtpi.org/tdm/tdm12.htm>
 Transportist blog: <https://transportist.org/>

References

- Aschauer, D.A., 1989. Is public expenditure productive? *J. Monet. Econ.* 23 (2), 177–200, doi:10.1016/0304-3932(89)90047-0.
 Condeço-Melhorado, A., Gutiérrez, J., García-Palomares, J.C., 2011. Spatial impacts of road pricing: accessibility, regional spillovers and territorial cohesion. *Transport. Res. A* 45 (3), 185–203.
 Gutiérrez, J., Condeço-Melhorado, A., Martín, J.C., 2010. Using accessibility indicators and GIS to assess spatial spillovers of transport infrastructure investment. *J. Transp. Geogr.* 18 (1), 141–152, doi:10.1016/j.jtrangeo.2008.12.003.
 Jenelius, E., 2009. Network structure and travel patterns: explaining the geographical disparities of road network vulnerability. *J. Transp. Geogr.* 17 (3), 234–244, doi:10.1016/j.jtrangeo.2008.06.002.
 Laird, J.J., Nellthorpe, J., Mackie, P.J., 2005. Network effects and total economic impact in transport appraisal. *Transp. Policy* 12 (6), 537–544, doi:10.1016/j.tranpol.2005.07.003.
 López, E., Gutiérrez, J., Gómez, G., 2008. Measuring regional cohesion effects of large-scale transport infrastructure investments: An accessibility approach. *Eur. Plan. Studies* 16 (2), 277–301, doi:10.1080/09654310701814629.

Further Reading

- Berdica, K., Mattsson, L.G., 2007. Vulnerability: a model-based case study of the road network in Stockholm. In: *Advances in Spatial Science*. Springer International Publishing, pp. 81–106. doi: 10.1007/978-3-540-68056-7_5.
 Boisjoly, G., El-Geneidy, A.M., 2017. How to get there? A critical assessment of accessibility objectives and indicators in metropolitan transportation plans. *Transp. Policy* 55, pp. 38–50. doi: 10.1016/j.tranpol.2016.12.011.
 Geurs, K.T., van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. *J. Transp. Geogr.* 12 (2), doi:10.1016/j.jtrangeo.2003.10.005.
 Gutiérrez, J., et al., 2011. Evaluating the European added value of TEN-T projects: a methodological proposal based on spatial spillovers, accessibility and GIS. *J. Transp. Geogr.* 19 (4), doi:10.1016/j.jtrangeo.2010.10.011.
 Holl, A., 2007. Twenty years of accessibility improvements. The case of the Spanish motorway building programme. *J. Transp. Geogr.* 15 (4), 286–297, doi:10.1016/j.jtrangeo.2006.09.003.
 Jenelius, E., Petersen, T., Mattsson, L.-G., 2006. Importance and exposure in road network vulnerability analysis. *Transp. Res. A* 40 (7), 537–560, doi:10.1016/J.TRA. 2005.11.003. first ed. Lucas, K., et al. (Eds.), *Measuring Transport Equity*, first ed., Elsevier.
 Mattsson, L.G., Jenelius, E., 2015. Vulnerability and resilience of transport systems—a discussion of recent research. *Transportation Research Part A: Policy and Practice*, 81. Elsevier Ltd, pp. 16–34, doi:10.1016/j.tr.2015.06.002.
 Páez, A., Scott, D.M., Morency, C., 2012. Measuring accessibility: positive and normative implementations of various accessibility indicators. *J. Transp. Geogr.* 25, 141–153, doi:10.1016/j.jtrangeo.2012.03.016.
 Reggiani, A., 2019. *Accessibility, Trade and Locational Behaviour*. Routledge. doi: 10.4324/9780429435300.
 Thomopoulos, N., Grant-Muller, S., 2013. Incorporating equity as part of the wider impacts in transport infrastructure assessment: an application of the SUMINI approach. *Transportation* 40, 315–345, doi:10.1007/s11116-012-9418-5.
 van Wee, B., 2016. Accessible accessibility research challenges. *J. Transp. Geogr.* 51, 9–16, doi:10.1016/j.jtrangeo.2015.10.018.

Road Modes: Walking

Kevin Manaugh, PhD Associate Professor, Department of Geography & McGill School of Environment McGill University, Montreal, QC, Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction	320
Background	320
Bringing Pedestrian Issues Into Policy	321
Measuring Walkable Pedestrian Space	321
Issues and Challenges	322
Opportunities for Change	323
References	323
Further Reading	325

Introduction

Encouraging walking as a safe, convenient, and practical mode of transport for everyday utilitarian trips is increasingly seen as an important step as cities around the world are grappling with how to move toward more sustainable transport systems based on low-carbon modes of transport. At the same time, efforts to make cities more livable and happy as well economically vibrant, and equitable places to live have become common policy goals at various levels of regional and city planning. Walking is perhaps uniquely situated to address a wide range of societal goals including reducing GHG emissions and traffic congestion; increasing social interaction, and improving air quality and population health.

Automobile-centric design, infrastructure, and policy of many urban areas worldwide have enormous environmental, economic, health, and safety impacts. The widespread use of automobiles is associated with greenhouse gas emissions and other pollutants with local, regional, and global impacts. In much of the world, privately owned cars are used for the majority of daily travel. In the United States, for example, it is estimated that over 80% of both work and shopping trips are by car. Transportation represents the largest single source of GHG emission in the United States (29%) (US EPA, 2015), of which the majority is from cars and light trucks. As well, the amount of space devoted to automobile movement and parking is associated with ecosystem disruption and issues related to storm water runoff and the urban heat island effect, and takes space that could be devoted to housing, green space, and other uses. Automobile-centric design of urban areas also has further impacts on human health and well-being including links to obesity and onset of chronic disease, traffic injuries and fatalities, and social exclusion. While some European and Asian cities have much lower car-ownership than what is typically observed in North America, the majority of daily trips are still taken by car in much of the world and most cities are actively attempting to curtail the dominance of the automobile for these environmental, economic, social, and aesthetic reasons.

Automobiles pose risks to people both inside and outside the vehicle. The World Health Organization estimates that of the roughly 1.4 million people killed in traffic crashes each year, 310,000 (23%) are pedestrians (National Highway Traffic Safety Administration, 2018). While technological and design advances have made cars safer for their occupants, pedestrian fatalities are also increasing more rapidly than automobile driver and passenger fatalities and now account for approximately 16% of all traffic-related deaths (National Highway Traffic Safety Administration, 2017). As speed and mass are the main determinants of injury severity in crashes between vehicles and people, the trend towards heavier vehicles and the fact the SUVs and light trucks represent an increasing percentage of automobile sales is troubling.

Background

Walking plays a foundational role in transport systems—all trips, regardless of primary mode, contain some amount of walking (whether it is traversing a parking lot, walking to a bus stop, transferring from one subway to another, walking after locking up or docking a bicycle). While not everybody is physically capable of walking on two legs, people of all abilities travel as pedestrians, sometimes using wheelchairs or other assistive devices. In addition, walking is available to almost all ages (except the very young and very old).

As the sole means of human mobility for millennia (until the domestication of pack animals, the development of boats, and the, much later, invention of motorized transport) walking plays a central and fundamental role in human history and the development of settlement patterns. Walking speed determined the size and location of settlements, villages, and cities. Until the late 1800s, the vast majority of major world cities could be traversed on foot in roughly 30 minutes. While “The Walkable City” (Speck, 2013) is often touted as a novel idea, it is important to remember that walkable urbanism was in fact simply the way cities were organized

until transport technologies such as streetcars, trains, subway systems and later cars stretched urban areas horizontally by opening up what could be accessed within reasonable travel times.

However, several decades of transport planning have prioritized the fast and efficient movement of cars over pedestrian safety and comfort. This prioritization has impacted the design of intersections, crosswalks, and pedestrian signal times meaning urban space is often designed in such a way as to discourage travel by foot. Pedestrian-focused safety improvements often face backlash from residents and other stakeholders if projects are perceived to slow down (car) traffic. Which is, in itself somewhat a historical if not ironic, as urban roads were not originally designed or intended to carry car traffic (as the surviving street grid of cities like New York and Boston, which have remained essentially unchanged since the 1800s can attest. The shift away from walkable urbanism, particularly in North America, has a myriad of social (Norton, 2011; Urry, 2004), legal (Shill, 2020) and political causes including post World War II mortgage reforms, the construction of Interstate Highway (Rose, 1979), and suburbanization of jobs (Brown et al., 2009). Other more recent trends include increasing amounts of “big box” retail located in suburban locations, which often are inconvenient, difficult, and dangerous to access on foot. While the evidence is not yet clear, online shopping is also likely to continue the trend away from residents walking to local shops for daily utilitarian needs. In addition, social norms, liability laws, and even how the media reports on traffic crashes (Ralph et al., 2019; Ralph and Girardeau, 2020) help reinforce the dominance of travel by car.

Bringing Pedestrian Issues Into Policy

Along with the increasingly ubiquitous term “walkability,” concepts or policy preoccupations such as the “8 to 80 City” (a city designed to accommodate anyone from the age of 8–80) and the “Fifteen Minute City” (all daily needs accessible within a 15 walk along safe and inviting paths), for example, have entered city-builders’ vocabulary when speaking about how cities can reallocate street space, prioritize active modes, and invest in intersection improvements in order to increase the attractiveness of walking. Subsequently, along with these new discourses in urban design, new policy approaches have entered city plans and budgets, (e.g., Complete Streets, Vision Zero, Safe Systems). In contrast to previous decades, pedestrian needs are now considered by a number of professions with a range of job titles, including planners, engineers, designers, and traffic monitoring specialists. Pedestrian issues are also addressed by other related disciplines such as economic development and public health, as the links to both the economic value and health benefits of pedestrian space is consistently highlighted by researchers (Sohn et al., 2012).

Simultaneously, pedestrian research has dramatically increased since the early 1990s, this goes along with the increased formal recognition of pedestrian issues in policy and legislative frameworks. More globally, the Vision Zero movement, started in Sweden in 1997 and based on the idea that traffic crashes are preventable and that the only socially acceptable number of traffic fatalities is zero, offers a system-focused look at road safety and moves beyond individual approaches to road safety. “Complete Streets” policies and initiatives have also blossomed in recent years, which attempt to bring in multi-modal approaches to street design and space allocation.

Measuring Walkable Pedestrian Space

Measuring and describing the elements of built form that encourage walking has been an important aspect of research on active travel for several decades. “Walkability” has become a common term used in both research and practice in a wide range of fields including urban planning, architecture, geography, public health, and increasingly in the media and real estate world. The term has been used in the academic literature since the 1970s (Abrams, 1972) and saw a resurgence in the 1990s and has since been used in a variety of settings to examine how neighborhood designs were associated with travel behavior (Owen et al., 2007), physical activity (Brennan Ramirez et al., 2006), and health outcomes, such as obesity and hypertension (Creatore et al., 2016; Hoehner et al., 2011; Ross et al., 2017; Rundle et al., 2016) as well as happiness, travel satisfaction, and emotional well-being (St-Louis et al., 2014). Influential work (Appleyard, 1980) also connected walkability to aspects of social interaction and social capital.

While some degree of confusion and lack of clarity around the term “walkability” exists (Forsyth, 2015; Shashank and Schuurman, 2019), most scholars and decision makers understand the term to refer to a quality of the built, natural, and social environment which allows people to access daily utilitarian needs (such as groceries, health care, transit stops, restaurants, cafes, retail stores, and green space) in an easy, safe, and convenient manner. It is therefore a quality of both the built form (street grid, quality of sidewalks) and the “content” (i.e., what is available to walk to, a function of land use mix and the location of various amenities) (Southworth, 2005). However, it is also, in many ways, a social construct, one that depends not only on the physical built environment, but on prevailing social attitudes and norms about what behaviors are expected or accepted in a particular location (be it on a scale of a country, city, neighborhood, or individual city block). Seniors and people with mobility issues or disabilities are forced to deal with crosswalk signalization that are often timed for people who move faster. In addition, some people (people of color and other marginalized people) face increased harassment by law enforcement while walking (Concepcion, 2014). In many locations, women encounter harassment and may feel uncomfortable and unsafe walking at certain times of the day. Therefore, the built environment is only one aspect of what characterizes a walkable space (Adkins et al., 2012). A truly walkable area would allow for anyone (regardless of race, gender, age, and ability) to safely and conveniently walk to desired destinations without fear of traffic danger, harassment by law enforcement or other street users, and along pleasant pedestrian paths.

Measurements of walkability rely on insights from Cervero and Kockelman's seminal 1997 work on the "3Ds" (Density, Diversity, and Design), referring to population and retail density, a diversity of land uses in close proximity, and a well-connected street grid design that makes walking trips feasible and comfortable (Cervero and Kockelman, 1997). Additional "D's" have been added to this conception (adding destination accessibility, and distance to transit to the original) (Ewing and Cervero, 2001), and later demand management and demographics (as a confounding variable in travel behavior) (Ewing and Cervero, 2010). Other scholars have also more explicitly incorporated safety and comfort (shade trees, street furniture, etc.), however, the 3D conception remains foundational in terms of how researchers conceive of walkable environments. Approaches of collecting data vary from audits (often using trained research assistants and systematic checklists, (Boarnet et al., 2011), surveys, remote sensing and GIS-based approaches to measure street grid and land use patterns (Manaugh and Kreider, 2013).

Researchers in the field have made distinctions around a hierarchy of walkable attributes in terms of their importance (Alfonzo, 2005). For example, Spoon (2005) identified four essential characteristics (mixed land use; accessibility/convenience; presence of pedestrian facilities; high connectivity), and a further five "encouraged" characteristics (street pattern; density; aesthetics; presence of parks, plazas and open space; traffic calming/street speeds). Similarly, Forsyth (2015) identified nine components of walkability of which four are considered necessary to achieve walkability (an environment that is traversable, compact or close, safe, and physically enticing).

Scholars have described a variety of additional factors at different scales (such as continuous, well-maintained sidewalks, crosswalk quality, separation of pedestrians from traffic, and landscape features (Lo, 2009). Another important area of research examines residents' perception of neighborhood walkability relative to objective measurements (Adkins et al., 2012; Leslie et al., 2005; van Lenthe and Kamphuis, 2011). Some focus on the needs and preferences of seniors (Negron-Poblete and Lord, 2014), or children (Buck et al., 2015).

Many of these measures and indices have been validated with travel behavior data and can help predict and explain the prevalence of walking (Adams et al., 2009; Frank et al., 2005; Owen et al., 2007), this is true even when controlling for residential self-selection and other confounding variables (Cao et al., 2009). However, there is compelling evidence that socio-economic status can modify the impact of the built and natural environment with, for example, members of poor, carless households walking even when conditions might not be seen to objectively support this behavior (Adkins et al., 2017; Manaugh and El-Geneidy, 2011; Riggs and Sethi, 2020; Steinmetz-Wood and Kestens, 2015).

Scholars have also suggested alternatives to the term "walkability" to encompass the wider health, social, environmental, and equity-related aspects of built, natural, and social environments that encourage active travel. Much of the confusion in the literature regarding measurement likely comes from different priorities and different behavioral outcomes of interest, for example, a street that promotes dog-walking or exercise might not be the same as one that allows for utilitarian shopping trips, "Walkable Suburbia" (Leinberger, 2010; Moudon et al., 2002; Southworth, 2007) looks different than a dense walkable downtown area. In recent years a variety of new approaches, such as walkalong interviews (Batista and Manaugh, 2017; Middleton, 2010), and Google street view analysis, sometimes using AI and machine learning techniques (Lu, 2019) have also deepened the conception of how to think about a walkable area.

Issues and Challenges

While walking can be a healthy, energizing, efficient mode of transport, the opportunity to access safe comfortable places to walk vary across socio-demographic factors. This has been explored by a number of researchers (Day, 2006; Ingram et al., 2017; Wolch et al., 2014). In addition to the enormous toll caused by automobile crashes, the risks of pedestrian injuries and fatalities are higher in low-income and racialized communities where residents often rely on walking as a daily mode of transport but where the local environment is not necessarily inviting and safe (Campos-Outcalt et al., 2003; Chandra et al., 2017; Morency et al., 2012; Rebentisch et al., 2019). In addition to evidence pointing toward large discrepancies in the provision of walkable urban space across income and racial lines, concern has been raised by scholars with regard to the possible gentrification and displacement impacts of improved pedestrian infrastructure (Bereitschaft, 2019). Large-scale pedestrian-oriented infrastructure projects (such as Manhattan's "High Line" and Philadelphia's Rail Park project) have been criticized for these impacts (Rigolon and Németh, 2018).

As well, in addition to discrepancies in provision of safe walkable environments, evidence points to differences in police enforcement (Concepcion, 2014) and driver yielding behavior (Goddard et al., 2015) making walking more dangerous and uncomfortable for people of color, and gender-based harassment that may, for example, limit the areas and times of day that women feel comfortable walking.

As perhaps a remnant of traffic engineering approaches, pedestrian planning and research has also struggled with how to approach design and analysis. For example, Pedestrian Level of Service (PLOS) measures rely on assumptions around the fact that maximizing pedestrian flow should be prioritized, as opposed to, for example, safety, comfort, and experience. In fact, planning for pedestrians raises important questions about the role and place of streets.

Many new transportation technologies have been introduced in recent years (and many lie just over the horizon). These technologies are often touted for their potential safety improvements as well as environmental benefits. Electric vehicles, for example, may reduce pedestrian exposure to vehicle emissions but may also introduce unintended negative consequences, such as being more difficult to hear, and maybe particularly so for blind pedestrians who are used to perceiving cars by sound.

Likewise, autonomous vehicles (AVs) present both opportunities and potential concerns for pedestrians. The emergence of highly automated driving systems has the potential to improve pedestrian safety, however, these vehicles may also have a variety of negative impacts from the transformation of urban form, travel behavior, roadway and vehicle design, and interactions between roadway users in unexpected ways. Many unanswered questions concerning liability and social and legal norms remain in terms of how these new technologies will impact pedestrian behavior and experiences. For example, will pedestrians be expected to change behavior, only walk in designated areas, or wear sensors in order to be detected by AVs?

The adoption of new forms of e-bikes, e-scooters, and expanding bikeshare systems have the potential to present challenges for pedestrians competing for limited amounts of street space. The needs of pedestrians and the social, environmental, health, and economic benefits of pedestrian travel should be fully considered in research and policy that incorporates these new technologies.

Planning for pedestrian space is also highly intersectional transport issue in that it is difficult to separate the many roles that the pedestrian realm plays (from a practical way to access desired destinations, but also a site of formal and informal meetings, a way to experience the social life of a city, as a place of protest, etc.) in a way that does not come to play to the same extent with other modes of transport. In this way, the many stakeholders who are engaged with the measurement, planning, and analysis of the pedestrian realm bring a variety of contrasting views, priorities, and ideas to the subject.

Opportunities for Change

Walking is a healthy, environmentally friendly mode of transport that is increasingly seen as an important part of a move towards sustainable urbanism. However, existing land use patterns, in addition to prevailing social and cultural norms are slowing the shift back towards walking being a common way to access daily travel needs. As Janette Sadik-Khan describes in “Street Fight: Handbook for an Urban Revolution,” (Sadik-Khan and Solomonov, 2016) cities can overcome resistance (be it from stakeholders in government and residents) to create healthy, livable and inclusive walking environments. New York City’s successful pilot projects closing streets to car traffic to allow safer pedestrian travel have pointed the way for other car-dependent areas to rethink the role of streets as solely for automobiles. Much of the history of transport planning in North America over the past century can be characterized by conflicts around the allocation of street space. These battles over the relative importance of different modes and mode users have been ongoing since the arrival of the commercially viable automobile in the 1920s and have dimensions related to environmental, social, and economic tradeoffs that have repercussions throughout society.

Writing this in the time of the 2020 COVID-19 lockdown of course opens the perspective on the importance of “walkability” as a way to ensure health, social inclusion, and safety for residents. Understanding where local residents can access food, health care, and green space while being able to practice physical distancing, is a vital public health and social justice issue. Somewhat optimistically, we can imagine that cities might learn from this experience how urban space has been devoted to cars instead of people. In early 2020, over 300 cities have significantly devoted space and budget to increasing the amount of pedestrian (and cyclist) infrastructure. What cities will learn medium and long-term remains to be seen. Tragically, the COVID-19 pandemic has also shone a light on the discrepancies in access to walkable space that scholars have been writing about for years, in terms of areas that lack quality sidewalks, green space, local amenities and how this is correlated to poverty and racialized communities. On a cautiously optimistic note, the speed at which municipalities were able to, at least temporarily, reallocate street space for pedestrians and cyclists in the time of a crisis illustrates that changes are possible and that a low-carbon transport system based on the comfort, safety, and convenience of pedestrians is possible and done in an equitable manner might be possible.

References

- Abrams, C., 1972. The pedestrian oriented neighborhood. *Ekistics* 33 (197), 299–301.
- Adams, M.A., Ryan, S., Kerr, J., Sallis, J.F., Patrick, K., Frank, L.D., Norman, G.J., 2009. Validation of the neighborhood environment walkability scale (NEWS) items using geographic information systems. *J. Phys. Activity Health* 6 (s1), S113–S123. doi:10.1123/jpah.6.s1.s113.
- Adkins, A., Dill, J., Luhr, G., Neal, M., 2012. Unpacking walkability: testing the influence of urban design features on perceptions of walking environment attractiveness. *J. Urban Des.* 17 (4), 499–510. doi:10.1080/13574809.2012.706365.
- Adkins, A., Makarewicz, C., Scanze, M., Ingram, M., Luhr, G., 2017. Contextualizing Walkability: do relationships between built environments and walking vary by socioeconomic context? *J. Am. Plan. Assoc.* 83 (3), 296–314.
- Alfonzo, M., 2005. To Walk or Not to Walk? The Hierarchy of Walking Needs. Available from: <https://journals.sagepub.com/doi/abs/10.1177/0013916504274016>.
- Appleyard, D., 1980. Livable streets: protected neighborhoods? *Ann. Am. Acad. Political Soc. Sci.* 451 (1). <https://journals.sagepub.com/doi/abs/10.1177/000271628045100111>.
- Batista, G., Manaugh, K., 2017. Using embodied videos of walking interviews in walkability assessment. *Transp. Res. Rec.* 2661 (1). https://journals.sagepub.com/doi/abs/10.3141/2661-02?casa_token=hr_WDwVXR0AAAAA:HdjnfBdRPavZC5PXBLch77TU5LkvzdfW2J_kQ42_70XFIFePeAAPE4PBHQs71v0fZED6GBjDbqD.
- Bereitschaft, B., 2019. Neighborhood walkability and housing affordability among US urban areas. *Urban Sci.* 3 (1). <https://www.mdpi.com/2413-8851/3/1/11>.
- Boarnet, M., Forsyth, A., Day, K., Oakes, M., 2011. The street level built environment and physical activity and walking: results of a predictive validity study for the irvine minnesota inventory. *Environ. Behav.* 43 (6). <https://journals.sagepub.com/doi/10.1177/0013916510379760?icid=int.sj-full-text.similar-articles>.
- Brennan Ramirez, L.K., Hoehner, C.M., Brownson, R.C., Cook, R., Orleans, C.T., Hollander, M., Barker, D.C., Bors, P., Ewing, R., Killingsworth, R., Petersmarck, K., Schmid, T., Wilkinson, W., 2006. Indicators of activity-friendly communities: an evidence-based consensus process. *Am. J. Prevent. Med.* 31 (6), 515–524. <https://doi.org/10.1016/j.amepre.2006.07.026>.
- Brown, J., Morris, E., Taylor, B., 2009. Planning for cars in cities: planners, engineers, and freeways in the 20th century. *J. Am. Plan. Assoc.* 75 (2). <https://www.tandfonline.com/doi/full/10.1080/01944360802640016>.
- Buck, C., Tkaczick, T., Pitsiladis, Y., De Bourdehaudhuij, I., Reisch, L., Ahrens, W., Pigeot, I., 2015. Objective measures of the built environment and physical activity in children: from walkability to moveability. *J. Urban Health* 92 (1), 24–38. doi:10.1007/s11524-014-9915-2.

- Campos-Outcalt, D., Bay, C., Dellapena, A., Cota, M., 2003. Motor vehicle crash fatalities by race/ethnicity in Arizona 1990-96. *Injury Prevent.* 9 (3), 251–256., doi:10.1136/ip.9.3.251.
- Cao, J., Mokhtarian, P., Handy, S., 2009. Examining the Impacts of Residential Self-Selection on Travel Behaviour: A Focus on Empirical Findings. *Transp. Res.* 29 (3). <https://www.tandfonline.com/doi/full/10.1080/01441640802539195>.
- Cervero, R., Kockelman, K., 1997. Travel demand and the 3Ds: density, diversity, and design. *Transport. Res. Part D* 2 (3), 199–219.
- Chandra, S., Jimenez, J., Radhakrishnan, R., 2017. Accessibility evaluations for nighttime walking and bicycling for low-income shift workers. *J. Transp. Geogr.* 64 (Suppl. C), 97–108, doi:10.1016/j.jtrangeo.2017.08.010.
- Concepcion Jr., R., 2014. The untapped potential of the fair housing act in addressing aggressive enforcement of walking while black or brown University of Pennsylvania. *J. Law Soc. Change* 17 (4), 383–400.
- Creatore, M.I., Glazier, R.H., Moineddin, R., Fazli, G.S., Johns, A., Gozdyra, P., Matheson, F.I., Kaufman-Shriqui, V., Rosella, L.C., Manuel, D.G., Booth, G.L., 2016. Association of neighborhood walkability with change in overweight, obesity, and diabetes. *JAMA* 315 (20), 2211–2220, doi:10.1001/jama.2016.5898.
- Day, K., 2006. Active living and social justice: planning for physical activity in low-income, black, and latino communities. *J. Am. Plan. Assoc.* 72 (1), 88–99.
- Ewing, R., Cervero, R., 2001. Travel and the built environment: a synthesis. *Transp. Res. Rec.* 780, 87–114.
- Ewing, R., Cervero, R., 2010. Travel and the built environment. *J. Am. Plan. Assoc.* 76 (3), 265–294, doi:10.1080/01944361003766766.
- Forsyth, A., 2015. What is a walkable place? The walkability debate in urban design SpringerLink. Available from: <https://link.springer.com/article/10.1057/udi.2015.22>.
- Frank, L.D., Schmid, T.L., Sallis, J.F., Chapman, J., Saelens, B.E., 2005. Linking objectively measured physical activity with objectively measured urban form: Findings from SMARTRAQ. *Am. J. Prevent. Med.* 28 (2, Suppl. 2), 117–125, doi:10.1016/j.amepre.2004.11.001.
- Goddard, T., Kahn, K.B., Adkins, A., 2015. Racial bias in driver yielding behavior at crosswalks. *Transport. Res.* 33, 1–6, doi:10.1016/j.trf.2015.06.002.
- Hoehner, C.M., Handy, S.L., Yan, Y., Blair, S.N., Berrigan, D., 2011. Association between neighborhood walkability, cardiorespiratory fitness and body-mass index. *Soc. Sci. Med.* 73 (12), 1707–1716, doi:10.1016/j.socscimed.2011.09.032.
- Ingram, M., Adkins, A., Hansen, K., Cascio, V., Somnez, E., 2017. Sociocultural perceptions of walkability in Mexican American neighborhoods: implications for policy and practice. *J. Transp. Health* 7, 172–180.
- Leinberger, C.B., 2010. The Option of Urbanism: Investing in a New American Dream. Island Press.
- Leslie, E., Saelens, B., Frank, L., Owen, N., Bauman, A., Coffee, N., Hugo, G., 2005. Residents' perceptions of walkability attributes in objectively different neighbourhoods: a pilot study. *Health Place* 11 (3), 227–236, doi:10.1016/j.healthplace.2004.05.005.
- Lo, R.H., 2009. Walkability: what is it? *J. Urban.* 2 (2). https://rsa.tandfonline.com/doi/full/10.1080/17549170903092867?casa_token=3f2VwpMBbKwAAAAA%3Ai4iFvJvAfI-W2tD_pia_8-jlHMSHFJKAL3iSFpUwLFwt-vQXHfIQ1Z5j7J1jBO2Hbomu-jRzC.
- Lu, Y., 2019. Using Google street view to investigate the association between street greenery and physical activity. *Landscape Urban Plan.* 191, 103435, doi:10.1016/j.landurbplan.2018.08.029.
- Manauth, K., El-Geneidy, A., 2011. Validating walkability indices: How do different households respond to the walkability of their neighborhood? *Transport. Res.* 16 (4), 309–315, doi:10.1016/j.trd.2011.01.009.
- Manauth, K., Kreider, T., 2013. What is mixed use? Presenting an interaction method for measuring land use mix. Available from: https://www.jstor.org/stable/26202648?seq=1#metadata_info_tab_contents.
- Middleton, J., 2010. Full article: sense and the city: exploring the embodied geographies of urban walking. *Soc. Cult. Geogr.* 11 (6). https://www.tandfonline.com/doi/full/10.1080/14649365.2010.497913?casa_token=085Ti0iP_RUAAAAA%3ANi7hCzmSDb2pw-Z3ThD9lY99cOD4bBEHbBpGKKb6Pgev_MgE-LeVFVtpzdgMcOH-DiQ6aVuyZN.
- Morency, P., Gauvin, L., Plante, C., Fournier, M., & Morency. (2012). Neighborhood Social Inequalities in Road Traffic Injuries: The Influence of Traffic Volume and Road Design. *American Journal of Public Health*. <https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2011.300528>.
- Moudon, A.V., Hess, P.M., Matlick, J., Pergakes, N., 2002. Pedestrian Location Identification Tools: Identifying Suburban Areas with Potentially High Latent Demand for Pedestrian Travel. <https://journals.sagepub.com/doi/abs/10.3141/1818-15>
- National Highway Traffic Safety Administration, 2017. Traffic Safety Facts. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812681>.
- Negron-Poblete, P., Lord, S., 2014. Marchabilité des environnements urbains autour des résidences pour personnes âgées de la région de Montréal: Application de l'audit MAPPA. *Cahiers de géographie du Québec* 58 (164), 233–257. <https://doi.org/10.7202/1031168ar>.
- Norton, P.D., 2011. Fighting Traffic: The Dawn of the Motor Age in the American City. MIT Press.
- Owen, N., Cerin, E., Leslie, E., duToit, L., Coffee, N., Frank, L.D., Bauman, A.E., Hugo, G., Saelens, B.E., Sallis, J.F., 2007. Neighborhood walkability and the walking behavior of Australian adults. *Am. J. Prevent. Med.* 33 (5), 387–395, doi:10.1016/j.amepre.2007.07.025.
- Ralph, K., Girardeau, I., 2020. Distracted by "distracted pedestrians"? *Transp. Res. Interdiscip. Persp.* 5, 100118, doi:10.1016/j.trip.2020.100118.
- Ralph, K., Iacobucci, E., Thigpen, C.G., Goddard, T., 2019. Editorial Patterns in Bicyclist and Pedestrian Crash Reporting. Available from: <https://journals.sagepub.com/doi/10.1177/0361198119825637>
- Rebentisch, H., Wasfi, R., Piatkowski, D., Manauth, K., 2019. Safe streets for all? Analyzing infrastructural response to pedestrian and cyclist crashes in New York City, 2009–2018. *Transport. Res. Rec.* <https://journals.sagepub.com/doi/full/10.1177/0361198118821672>.
- Riggs, W., Sethi, S.A., 2020. Multimodal travel behaviour, walkability indices, and social mobility: How neighbourhood walkability, income and household characteristics guide walking, biking & transit decisions. *Local Environ.* 25 (1), 57–68, doi:10.1080/13549839.2019.1698529.
- Rigolon, A., Németh, J., 2018. We're not in the business of housing." Environmental gentrification and the nonproliferation of green infrastructure projects. *Cities* 81, 71–80, doi:10.1016/j.cities.2018.03.016.
- Rose, M.H., 1979. Interstate: express highway politics 1941–1956. Available from: <https://trid.trb.org/view/150164>.
- Ross, N., Dasgupta, K., Sanmartin, C., Wasfi, R., Hajna, S., Mah, S., 2017. Using linked data to explore the relationship between walking-friendliness of neighbourhoods physical activity and body mass. *Int. J. Population Data Sci.* 1 (1). Article 1 <https://doi.org/10.23889/ijpds.v1i1.343>.
- Rundle, A.G., Sheehan, D.M., Quinn, J.W., Bartley, K., Eisenhower, D., Bader, M.M.D., Lovasi, G.S., Neckerman, K.M., 2016. Using GPS data to study neighborhood walkability and physical activity. *Am. J. Prevent. Med.* 50 (3), e65–e72, doi:10.1016/j.amepre.2015.07.033.
- Sadik-Khan, J., Solomonov, S., 2016. Street Fight: Handbook for an Urban Revolution. Penguin Books.
- Shashank, A., Schuurman, N., 2019. Unpacking walkability indices and their inherent assumptions. *Health Place* 55, 145–154, doi:10.1016/j.healthplace.2018.12.005.
- Shill, G.H., 2020. Should law subsidize driving? (SSRN Scholarly Paper ID 3345366). Social Science Research Network. Available from: <https://doi.org/10.2139/ssrn.3345366>.
- Sohn, D.W., Moudon, A.V., Lee, J., 2012. The economic value of walkable neighborhoods. *Urban Des. Int.* 17 (2), 115–128., doi:10.1057/udi.2012.1.
- Southworth, M., 2005. Designing the walkable city journal of urban planning and development. *J. Urban Plan. Dev.* 131 (4.), <https://ascilibrary.org/doi/full/10.1061/%28ASCE%290733-9488%282005%29131%3A4%28246%29>.
- Southworth, M., 2007. Walkable Suburbs?: An Evaluation of Neotraditional Communities at the Urban Edge. *J. Am. Plann. Assoc.* 63 (1), https://www.tandfonline.com/doi/abs/10.1080/01944369708975722?casa_token=8aQ7NZxWM3cAAAAA:L4RkZLfQdX1PwaSx5YleAFTU0sW374JfRz8fmHoN7fEZ19F9GMat_OY0KsXLZBeegmDmXEBE1FC.
- Speck, J., 2013. The Walkable City: How Downtown can Save America. In: One Step at a Time, North Point Press.
- Spoon, S.C., 2005. What defines walkability: Walking behavior correlates. Masters Thesis. University of North Carolina at Chapel Hill.
- Steinmetz-Wood, M., Kestens, Y., 2015. Does the effect of walkable built environments vary by neighborhood socioeconomic status? *Prevent. Med.* 81, 262–267, doi:10.1016/j.ypmed.2015.09.008.
- St-Louis, E., Manauth, K., van Lierop, D., El-Geneidy, A., 2014. The happy commuter: A comparison of commuter satisfaction across modes. *Transport. Res.* 26, 160–170, doi:10.1016/j.trf.2014.07.004.
- Urry, J., 2004. The 'System' of Automobility. Available from: <https://journals.sagepub.com/doi/abs/10.1177/0263276404046059>.

- US EPA, O., 2015. Fast Facts on Transportation Greenhouse Gas Emissions [Overviews and Factsheets]. US EPA. Available from: <https://www.epa.gov/greenvehicles/fast-facts-transportation-greenhouse-gas-emissions>.
- van Lenthe, F.J., Kamphuis, C.B.M., 2011. Mismatched perceptions of neighbourhood walkability: Need for interventions? *Health Place* 17 (6), 1294–1295, doi:10.1016/j.healthplace.2011.07.001.
- Wolch, J.R., Byrne, J., Newell, J.P., 2014. Urban green space, public health, and environmental justice: The challenge of making cities 'just green enough'. *Landscape Urban Plan.* 125, 234–244, doi:10.1016/j.landurbplan.2014.01.017.

Further Reading

- Cerin, E., Conway, T.L., Saelens, B.E., Frank, L.D., Sallis, J.F., 2009. Cross-validation of the factorial structure of the Neighborhood Environment Walkability Scale (NEWS) and its abbreviated form (NEWS-A). *Int. J. Behavior. Nutr. Phys. Activity* 6 (1), 32, doi:10.1186/1479-5868-6-32.
- Ewing, R., Handy, S., 2009. Measuring the unmeasurable: urban design qualities related to walkability. *J. Urban Des.* 14 (1), 65–84, doi:10.1080/13574800802451155.
- Leslie, E., Coffee, N., Frank, L., Owen, N., Bauman, A., Hugo, G., 2007. Walkability of local communities: Using geographic information systems to objectively assess relevant environmental attributes. *Health Place* 13 (1), 111–122, doi:10.1016/j.healthplace.2005.11.001.
- Marshall Julian, D., Brauer, Michael, Frank Lawrence, D., 2009. Healthy neighborhoods: walkability and air pollution. *Environ. Health Perspect.* 117 (11), 1752–1759, doi:10.1289/ehp.0900595.
- Middleton, J., 2018. The socialities of everyday urban walking and the 'right to the city'. *Urban Studies* 55 (2), 296–315, doi:10.1177/0042098016649325.

Bus Public Transport Planning and Operations

Ehab Diab*, **Ahmed El-Geneidy†**, *University of Saskatchewan, Department of Geography and Planning, Saskatoon, Saskatchewan, Canada;
†McGill University, School of Urban Planning, Montréal, Québec, Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction	326
Bus Transit Operations: Concepts and Terminologies	326
Bus Transit Users' Perception of Service Quality	327
Bus Transit Service Improvement Strategies	328
Reserved Bus Lanes	328
Transit Signal Priority (TSP)	329
Limited-Stop Service (Express)	330
Articulated Buses	330
All-Door Boarding	330
Service Coordination	331
Summary and Conclusion	331
See Also	332
References	332
Further Reading	332

Introduction

Over the past decades, public transport has been introduced as a solution to solve many environmental and social goals in our cities. Bus transit service (or fixed-route bus transit systems) is the dominant form of public transport in most cities around the world and is the backbone of any public transport system. It comprises more than half of ridership compared to all other modes (e.g., commuter rail, light rail, paratransit, vanpool) in North America, where other modes are present.

While bus public transport system is a central element in the mobility in our cities, it is one of the most challenging systems to operate in an efficient and reliable manner due to the variety of internal and external factors that impacts its performance. External factors include weather conditions, traffic congestions and incidents, and demand variations, while internal factors include operators (or drivers) absences, scheduling and transferring problems, equipment failures, and other public transport modes (e.g., metro or streetcar systems) breakdowns. These factors present a persistent challenge for public transport agencies, hindering their ability to achieve the required level of service needed to enhance users' experience. Therefore, public transport agencies are in continuous search for new ways to enhance service quality, to retain existing riders, and to attract new ones.

Public transport users consider a system to be reliable only when they have easy access to the service (at both origin and destination) and with low and consistent waiting and travel times. To achieve these goals, over the past few decades, many strategies have been widely introduced to improve the bus service quality and users' experience.

This chapter starts with presenting an overview of several concepts and terminologies that was normally used in bus service operations. This is followed by a description of transit users' trip components and how they are weighing the importance of different components. This is, then, followed by a critical discussion of several bus transit service improvement strategies, while highlighting these strategies pros and cons. Finally, the chapter wraps up with a reiteration of the main conclusions.

Bus Transit Operations: Concepts and Terminologies

Good transport planning and efficient transit operations are required to balance the trade-offs among the goals, constraints, and values of transit agencies and passengers. *Transit service operations* refers to a wide range of activities done by a transit agency, which covers system management, scheduling, and functioning. *Reliability* is a key concept in the planning and operations of the public transport service. It refers to invariability of transit service attributes at certain locations. In the transportation setting, *service planning* involves the evaluation, assessment, and design and siting of transportation facilities. Other important operational concepts that will be used in this chapter includes *service headway and frequency*, *service capacity*, *bus stopping*, *running and travel times*, and *cycle time*.

The concepts of *headway* and *frequency* mean the same thing; however, they are often confused. *Headway* (h) usually refers to the time interval (in minutes) between two successive transit vehicles passing a fixed point (a stop) on a transit line in the same direction, while offering service for users. Headway is calculated based on departure times of successive transit vehicles. *Frequency* (f) refers to the number of transit vehicles that pass a point on a line (vehicles per hour). That said, it is the reverse of headway—a route with 20-min headway has a frequency of three vehicles/h. The mathematical relationship between both can be expressed as follows:

$$Frequency(f)_{veh/h} = 60/Headway(h)_{min} \quad (1)$$

There is always an important trade-off between users and transit agencies related to service frequency. Users are interested in shorter headway service (more frequent service) to minimize their waiting time. In contrast, transit agencies are interested in reducing their operational costs by operating fewer vehicles with large capacity (e.g., articulated buses, discussed later) rather than a large number of vehicles with smaller capacity.

Service line capacity (C) refers to the maximum number of users that can be transported past a fixed point in one direction during 1 h (or in a given time period). Therefore, the capacity of bus transit line is a function of its frequency and the used vehicle crush capacity (C_v). *Crash capacity* (or crash load) refers to the maximum number of passengers per bus targeted by the transit agency. Capacity is an important concept used by transit agencies to compare different vehicle types.

$$Capacity(C)_{passengers/h} = Vehicle\ capacity(C_v)_{Max.\ passengers\ per\ bus} * Frequency(f)_{veh/h} \dots \quad (2)$$

Running time (r_t) refers to the time that a bus takes to start travel from one station until it arrives at the next station in one direction along a route. It equals to the bus arrival time at the next stop minus bus departure time from the previous stop. *Stop time* (s_t) refers to the passage of time a bus spends at stops serving passenger movements. It equals to vehicle departure time minus vehicle arrival time at the same stop. The largest component of stop time is *dwell time* (or *passenger service time*). Dwell time is influenced by a considerable number of factors. These factors include passenger demand (more passengers getting on and off buses cause longer dwells), fare payment type (time required for fare validation), and vehicle and stop configurations (platform level, and vehicle number of doors and the number of channels per door). They also include vehicle passenger load (number of standing passengers on the bus, more standing passengers lead to longer dwells), door usage policy (alighting passengers exiting through the front door delay the start of the boarding process), and other issues related to bus lift (lifts are used to help seniors and people on wheelchairs to get on and off buses) and bike racks usage.

Travel time (T_r) is the time between departure from one terminal and arrival at the other terminal. It equals to the sum of running times and dwell times between departure from one terminal and arrival at the other terminal. *Terminal time* (t_t) is the time spent by a bus at a line terminal (Fig. 1). Terminal time (also called *Layover time*) is an important part of the bus operations practice. It is used to provide time to turnaround the vehicle, change driver, driver break for rest and/or meal, as a cushion to absorb and recover from delays. Terminal time is also used for while building bus schedules for schedule adjustment. *Cycle time*, C_t is the round trip time for a vehicle from the moment it leaves one terminal until the moment it leaves the same terminal after completing one cycle. *Deadhead time* refers to the travel time during which the bus is not in revenue service (e.g., travel from depot to start of line or travel between lines) (Fig. 2).

Bus Transit Users' Perception of Service Quality

A typical trip (door-to-door trip) for a passenger includes walking to transit stop (access time), waiting time at transit stop, on-board travel time in a vehicle, transferring time (to transfer between lines), on-board travel time in the vehicle, and, finally, walking to their final destinations (egress time). Passengers perceive the passage of time differently for each portion of their trip. Also, the above components can include driving to a park and ride rather than walking and not all trips include transfers.

Waiting time is usually perceived differently compared to the actual time (Psarros et al., 2011). During this time portion, users are exposed to irregular weather conditions as well as the adverse surrounding traffic, noise, and pollution. They also can be stressed by not knowing if the bus is arriving as indicated on schedules or not (Hess et al., 2004). Transferring between different routes has also an exponential effect on users' perceptions, leading to a considerable overestimation of their actual spent times (Iseki and Taylor, 2009). In contrast, typically, users perceived travel times an accurate manner compared to the actual travel time spent in the vehicle.

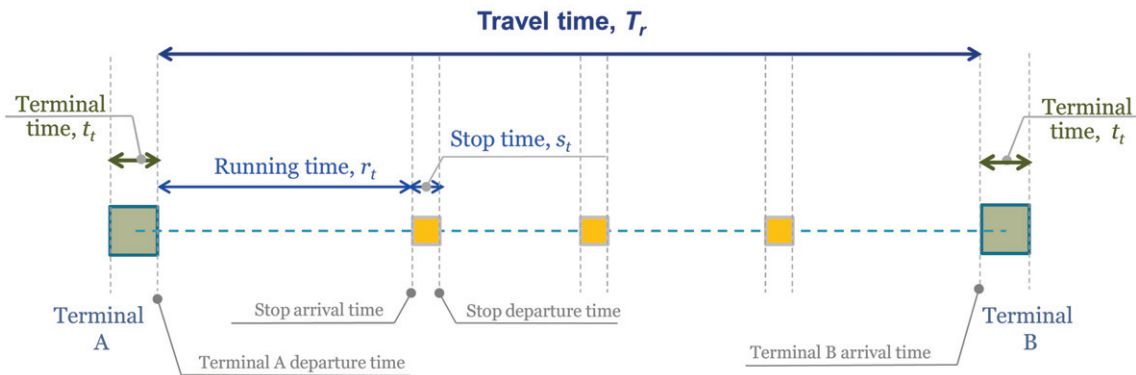


Figure 1 Bus running time, stop time, and travel time.

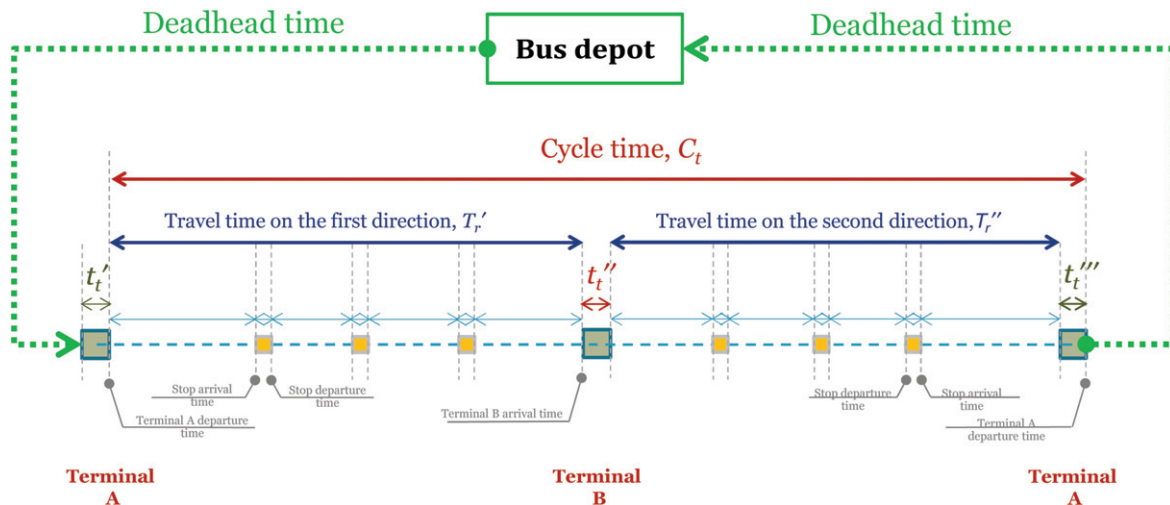


Figure 2 Bus cycle time and deadhead time.

Poor (or unreliable) service operations have undesirable consequences on both transit users and transit agencies (Evans, 2004). Transit agencies will have to provide an excessive buffer time (or layover time) between trips to absorb service delays and variation in travel times to make sure that the following trip starts on time. This means increasing the service cycle time, which will be translated into less frequent service, decreasing the quantity of the offered service to the public. Furthermore, travel time variability (or uncertainty) represents a problem for users, which they must account for during their trip planning process by employing *buffer time*. *Buffer time* is the additional travel time that users add to their trips to make sure they arrive to their destinations on time, demonstrating a direct impact of unreliable service on users' time.

In addition, variations in travel time between stops can further lead to bus bunching, particularly for frequent service of 15 min or less. *Bus bunching* (or *bus platooning*) occurs when two subsequent vehicles (or more) are not able to maintain the even-spaced schedule (in terms of time), and ending up traveling in close proximity to each other (as a platoon) and serving the same stops at the same time. Leading vehicles, in this case, get more delayed by picking up passengers, who are waiting for the next bus, and exacerbate the delays of the leading vehicle further and further. This is converted to a gap in the service that increases users' waiting time and vehicle overcrowding due to the uneven distributions of the vehicles.

For the previous reasons, passengers usually rate reliability as the second most important transit attribute after arriving safely at destinations. Transit users' behavioral surveys show that passengers consider service reliability twice as important as bus frequency (number of trips per hour). Furthermore, service frequency and shorter waiting time are more important than greater accessibility, because passengers prefer to walk more than to wait longer at stops.

Bus Transit Service Improvement Strategies

In the face of the previously discussed challenges and issues, transit agencies use several public transport preferential treatments and improvement strategies. *Improvement strategies* refer to any tactic used to improve the service operations and the perceived quality of the service. While, *preferential treatments* refer to a subcategory of measures that gives public transport vehicles advantage over other mode vehicles (i.e., private cars) at intersections or along the streets. Transit agencies employ a wide range of operational and physical improvement strategies that differ according to the required (or desired) level of improvement, capital and operational costs, and political acceptance.

The most commonly used strategies include transit signal priority (TSP), new buses (articulated buses and low-floor buses), reserved bus lanes, limited-stop services (express services), and service coordination and intelligent transportation system (ITS) (Diab et al., 2015). These strategies are typically combined to improve the conventional bus system's capacity, productivity, and reliability in addition to users' experience, and called *BRT/BRT-like* systems. BRT systems also come with special promotion campaigns and advertisings to separate them from local bus services. They are considered as one of the most economically effective tools to increase transit ridership that influences users' behavior.

Reserved Bus Lanes

Overview: Reserved bus lane (or dedicated bus lane) is one of the most effective strategies used by transit agencies along urban corridors. It refers to dedicating a traffic lane for bus service operations while prohibiting the general traffic from using this lane. There are different types of reserved bus lanes according to the infrastructure and traffic flow. Reserved bus lanes can be

semiexclusive, without any physical separation, allowing the general traffic to use the lanes at certain locations and segments (e.g., for turning movement at intersections).

Reserved bus lane can be operated only during peak periods or all day. They can be also fully exclusive, reserved of transit use at all the times, with a physical separation. Typically, these lanes are used for BRT systems, with some interference only at intersection from the perpendicular traffic. Therefore, they are combined with TSP systems at intersections (discussed later). Reserved lanes can be located adjacent to the curbside, offset to the lane adjacent to the curbside, or adjacent to the median.

Excepted impacts: Reserved bus lanes have considerable benefits for public transport service operations. They also have been associated with positive perceptions of quality of service by the users. More specifically, compared to bus operations in mixed-traffic lanes, reserved bus lanes improve transit service travel times, while decreasing service variations in most of the cases. This means a faster service observed by users as well as more predictable waiting and travel times. Transit users tend to overestimate their travel time savings associated with the implementation of such a measure.

For transit agencies, reserved bus lanes help in providing more accurate and efficient schedules due to the increases in service reliability. Nevertheless, these benefits can differ according to the level of delays experienced in mixed-traffic situation before the implementation of such a measure and according to the type of reserved lanes (semiexclusive vs. exclusive). In some cases, semiexclusive lanes slightly improve bus service travel times, with minimal or even negative impacts on bus service variations, due to turning movements at intersections and no-turn on red policy (i.e., right turns are not allowed on red) that used in some cities such as Montréal (Surprenant-Legault and El-Geneidy, 2011).

Ease of implementation: As reserved bus lanes have a physical impact on the built environment and transportation networks, the planning process for introducing reserved bus lanes needs to involve many stakeholders while assessing the potential impacts on other road users, adjacent properties, and the needs for law enforcement. Therefore, the first step in the planning process is to identify different stakeholders, which may include different department within the city (e.g., city's planning, public work, and law enforcement departments) and government agencies, in addition to local community members, private industries, businesses owners, and elected officials.

Due to the removal of traffic lane from car users, sometimes obtaining the appropriate political acceptance for the implementation of such measure can be challenging. Additionally, financial constraints in terms of providing the required capital cost for constructing such lanes can represent a barrier for their implementation.

Transit Signal Priority (TSP)

Overview: TSP is an operational strategy that gives priority to public transport vehicles at traffic-signalized intersections, facilitating their movement in a reliable manner. Transit vehicles typically experience substantial delays at traffic signals in busy urban areas. Therefore, one of the most popular strategies commonly used is TSP. TSP systems can be "passive" or "active." Passive systems rely on fixing the signal timing plans based on the expected transit vehicles' running and dwell times. They do not require any special technology for transit vehicle detection. However, passive TSPs provide minor benefits to no benefits, particularly if transit operations are unreliable and unpredictable. In contrast, active TSPs work on modifying signal timing in real time to provide priority treatment to a specific vehicle at a specific location (i.e., intersection) following its detection and succeeding priority request. Therefore, active TSP systems include hardware and software technologies: detection system, priority request generator and server, and communication system.

Typically, modifying the signal timing is done by activating "green extension" or "early green." Green extension strategy works upon the detection of a transit vehicle upstream of an intersection (within a specific window of time) while a green phase is active--this strategy extends the green phase. The extension maximum value of time (varies between 16 and 30 s according to the type of intersection, using Toronto as a case study). This value of time can determine the effectiveness of the strategy and minimize the number of failed extensions. Early green (or red truncation) works upon detection of a transit vehicle upstream of an intersection (within a specific window of time) while a red phase is active--this strategy shortens the red phase and returns of the next green phase. Generally, both approaches, green extension and early green, can be available together along a corridor; however, they cannot be applied at the same time.

Excepted impacts: TSP systems can significantly improve the transit vehicles' travel times by reducing their delay by an entire red interval (or around 90 s, using downtown Toronto as an example). However, in most of the cases, only fewer buses that arrive at the TSP equipped intersection within a short window of time benefit from this large saving. This means that these systems while typically having considerable impacts on service operating times may have a mixed impact on transit service variations.

Users experience travel and waiting times savings due to the implementation of TSPs. However, they tend not to overestimate these time savings. In other words, TSP systems have a limited impact on users' perceptions of changes in the quality of service in contrast to reserved bus lanes and other strategies that have a physical component that can be observed by users. Furthermore, potential negative impacts of TSPs consist primarily of delays to nonpriority traffic and road users and possible delays at neighboring intersections.

Ease of implementation: TSP is one of the most common and preferred strategies that have been used by transit agencies around the world. Since TSP is an operational strategy that has no physical impact on the built environment and road space, it is normally implemented with a higher degree of ease compared to other strategies (i.e., reserved bus lanes).

Regarding the cost of TSP implementation, in addition to a relatively high capital cost of system hardware and software components, considerable operating and maintenance costs are required to keep the system functioning in a reliable manner.

However, typically, monetary benefits of TSP systems' implementation, with regard to improved travel times and reduced delays, outweigh costs.

Limited-Stop Service (Express)

Overview: Limited-stop service is an effective strategy used by transit agencies along urban corridors. In contrast to local bus services, which serves all stops along the route, limited-stop service refers to offering that transport service only at major stops (in terms of ridership volumes) and transfer locations along a bus corridor. Limited-stop bus services could be offered in combination with local bus services. A good example is Montréal Route 467, which serves only 40% of the stops of its local route, Route 67 (Diab and El-Geneidy, 2013).

BRT routes are typically limited-stop routes. It should be noted that the concept of limited-stop service is often mixed with the concept of "express service." Express service is a limited-stop service in essence; however, it is used for long distance trips and for providing a direct connection between suburban areas and central locations (e.g., central business districts (CBDs)).

Expected impacts: Limited-stop services provide a faster service compared to serving only fewer stops and usually are associated with superior performance in terms of providing more predictable and reliable service. They have also an overall positive impact on users' experience and perceptions due to the addition of more buses that are fast and reliable.

For transit agencies, limited-stop services offer lower operating costs as vehicles can complete a cycle trip faster than the local services, leading to fewer buses and the possibility of adding more service (by increasing the frequency of service). The only trade-off, from the passengers' perspective, is related to the need to walk further to access destinations than otherwise needed with local services (Murray and Wu, 2003).

Ease of implementation: Limited-stop service is considerably simple and cheap to implement because it usually uses the existing infrastructure (e.g., bus stops and routes) compared to the previous two strategies. However, due to the previously discussed trade-offs, a delicate planning process that considers "local passengers" needs to access local destinations at stops and "through passengers" prospect of a faster service should be done.

From transit agencies' perspective, operating a local and an express service along the same streets can represent an operational and scheduling challenge. Transit agencies should make sure that local services are not delaying the express service or the other way.

Articulated Buses

Overview: Different types of vehicles can be used to deliver public transport services in urban areas. These vehicles include minibuses, regular buses (12-m length), and articulated buses (18-m length). Regular buses are the most commonly used bus type in North America. However, recently more transit agencies are incorporating articulated buses into the service. Articulated buses are mostly low-floor buses and normally used along the busy corridors with high ridership volumes. They provide between 30% and 60% more seats and standing capacity compared to regular buses.

Expected impacts: Articulated buses offer more capacity than regular buses, helping agencies in providing more capacity without changing the number of trips along corridors by replacing regular buses with articulated buses. Transit agencies also use articulated buses to reduce the number of buses (and operational costs) along corridors, while maintaining the same offered capacity. This reduction in frequency while may have a negative impact on user perceptions due to increases in waiting times, it helps agencies to maintain better service operations by reducing bus bunching.

Research has shown that articulated buses are slightly slower than regular buses; however, they are more predictable in terms of less service variations. Therefore, implementing these buses will improve transit agencies scheduling and reliability in comparison with regular buses. Along busy corridors, articulated buses have shown positive impacts on users' waiting time and travel time perceptions due to the increased offered capacity. This increased capacity raises the probability of finding a seat for users and decreases their anxiety for the need to wait for the following buses at transit stops, which occurs along busy corridors during peak periods.

Ease of implementation: Articulated buses are an operational strategy, which can be planned and implemented by transit agencies, with a minimum need for involving different stakeholders. However, if the agency is planning to reduce the frequency to introduce those buses, this will require consulting local passengers about the required changes. Reducing service frequency while maintaining the same capacity, for transit agencies, will decrease transit agencies' operating costs significantly (i.e., fewer buses equal fewer operators and less operational and maintenance costs). Nevertheless, for transit agencies, replacing regular buses with articulated buses while maintaining the existing frequency is associated with higher capital costs and potential higher operational costs that are associated with modifying maintenance facilities to accommodate the new vehicles and the need for more fuel (articulated buses are typically less fuel-efficient).

All-Door Boarding

Overview: Traditional fare collection method requires fare inspection and validation upon boarding a vehicle add a considerable amount of time to public transport service operations. Using data from Vancouver and Montréal in Canada, several studies showed that using exact change can take up to 4–5 s per passenger. In addition, while it takes around 3–4 and 2–3 s for using smart cards and for the visual inspection of flash passes, it takes only 1–2 s for entering the vehicle without presenting any fare. This shows a

considerable difference per passenger associated with the type of the presented fare versus no fare. Therefore, one of the most effective strategies to decrease dwell time at stops is to use all-door boarding.

All-door boarding (or prepaid fare, proof-of-payment) refers to moving fare collection off the vehicles and allowing passengers to use all the bus doors to board at stops. This strategy can be used only at a few locations along the system (major activity nodes and transfer location) or it can be used system wide.

Expected impacts: While there are many factors affecting bus dwell time such as the number of doors, number channels per door, alighting policy, nonlevel boarding, and bus crowding, all-door boarding strategies have proven to have considerable benefits on bus operations and user perceptions. All-door boarding reduces service activity time per passenger, resulting in dwell time saving that can reach up to 40%–50% compared to the base case of using exact cash. According to bus route location and structure, dwell times can encompass between 10% and 40% of the total trip travel time.

This previous dwell time and travel time savings are typically combined with significant decreases in dwell time and travel time variations, making the service more predictable, for transit agencies. From the users' perspective, all-door boarding has a positive impact on their perception of the quality of service, making their transfer seamless between vehicles at location with common all-door boarding areas.

Ease of implementation: All-door boarding strategies have an important physical component, which is related to the need of providing an extra space at stops for fare inspection and validation. This space is not always easy to acquire and requires changes in street design and sidewalk configuration. Therefore, a wide range of stakeholders should be involved in the planning and the implementation of such a strategy.

As a physical strategy, a considerable capital cost will be associated with the construction of the new stops and providing fare validation machines and areas. To cut capital costs, some transit agencies use "honor system," which allows transit users to get on and off the vehicles without presenting fares, while trusting that users have proof of payment. However, this system is still associated with a substantial loss in revenue due to fare fraud and the cost of hiring a large number of inspectors.

Service Coordination

Overview: Intramodal and intermodal transfers are routinely done by more than 40% of passengers in North America during their trips. Nevertheless, for low frequent services, transferring between lines can represent a serious problem for users due to excessive transferring time. Therefore, transit agencies tend to tackle this problem by coordinating schedules at transferring locations, which is combined by employing intelligent transportation systems (ITSs). Automatic vehicle location (AVL) and computer-aided dispatching (CAD) systems are used to protect transfers by holding vehicles to ensure adequate transferring times for users.

Timed Transfer System (TTS), known as Pulse System, is also a common approach used when different routes are connected at a single terminal. In a typical TTS, vehicles from all routes arrive shortly before a specific departure time and then depart together, leaving a small window of time for transferring between routes.

Expected impacts: Service coordination has a considerable impact on users' perceptions. Users usually weight transferring time 5–10 times higher than in-vehicle travel time and 2–5 times more than waiting time. In the transportation literature, this weighting is based on the difference between the actual time and the perceived time during different activities, which represents the actual value of time for users. Therefore, improving and securing transfer times will have a strong positive effect on users' perceptions of the quality of service and their future travel behavior. The main disadvantage of service coordination is related to the need for holding and delaying some vehicles (to protect transfers), which may result in extra operational costs.

Ease of implementation: Service coordination is an operational strategy, which can be implemented by transit agencies, with almost no need for involving different stakeholders. The main capital costs are related to obtaining and operating AVL/CAD systems and purchasing software that helps transit agencies in protecting their transfer. Nevertheless, most of the world's modern transit agencies already have AVL/CAD systems to monitor and adjust bus operations. Therefore, the implementation of such operational strategy can be relatively inexpensive; however, it includes extensive scheduling and planning exercises to make sure that buses can arrive at locations around the same time.

Summary and Conclusion

Planning and design for improving transit service is not an easy task that requires trade-offs between different objectives and values. Bus transit service improvement strategies are typically introduced to improve both the transit operations and users' perceptions. This chapter discussed the benefits and potential impacts of several operational and physical improvement strategies that are commonly used by transit agencies across the world. These strategies included TSP, limited-stop service, articulated buses, all-door boarding, and service coordination.

Dwell and travel times and service reliability are the key determinants of service quality, speed, and capacity. In other words, strategies that decrease travel time and improve transit service reliability (i.e., all-door boarding, reserved bus lanes, TSP) reduce operating time, recovery time, and, thereby, service cycle time. This can be translated into a reduced fleet size and/or improved headway. For existing and new passengers, this means shorter actual and budgeted travel times, with improvement in waiting times and overall comfort, which will be resulted from the increase in service frequency and reliability.

Some improvement strategies can improve only one aspect of operations that is related to travel time (TSP), service variation (articulated buses), or transferring time (service coordination), enhancing user perceptions of the service. As a role, operational strategies are usually easier to implement and acquire the required political acceptance compared to physical strategies. However, both types of strategies are required to be reviewed while assessing the local context. Trade-off between capital and operating costs is essential to be considered too while estimating each strategy cost and benefits. Overall, these strategies will help cities in achieving their broader sustainability and equity goals.

See Also

Operation Costs For Public Transport; What Drives Transport Demand Trends? The Chicken or Egg Problem; The Value of Security, Access Time, Waiting Time and Transfers in Public Transport; Road Modes: Walking; Transit Planning and Management; Planning For Bus Rapid Transit (BRT); Bus Route Planning

References

- Diab, E., Badami, M., El-Geneidy, A., 2015. Bus transit service reliability and improvement strategies: integrating the perspectives of passengers and transit agencies in North America. *Transport Rev.* 35, 292–328.
- Diab, E., El-Geneidy, A., 2013. Variation in bus transit service: understanding the impacts of various improvement strategies on transit service reliability. *Public Transport* 4, 209–231.
- Evans, J., 2004. Transit Scheduling and Frequency: In Traveler Response to Transportation System Changes. Transportation Research Board, Washington D.C.
- Hess, D., Brown, J., Shoup, D., 2004. Waiting for the bus. *J. Public Transport.* 7, 67–84.
- Iseki, H., Taylor, B., 2009. Not all transfers are created equal: towards a framework relating transfer connectivity to travel behaviour. *Transport Rev.* 29, 777–800.
- Murray, A., Wu, X., 2003. Accessibility tradeoffs in public transit planning. *J. Geograph. Syst.* 5, 93–107.
- Psarros, I., Kepaptsoglou, K., Karlaftis, M., 2011. An empirical investigation of passenger wait time perceptions using hazard-based duration models. *J. Public Transport.* 14, 6.
- Surprenant-Legault, J., El-Geneidy, A., 2011. Introduction of reserved bus lane: impact on bus running time and on-time performance. *Transport. Res. Rec.* 2218, 10–18.

Further Reading

- Boisjoly, G., Grisé, E., Maguire, M., Veillette, M., Deboosere, R., Berrebi, E., El-Geneidy, A., 2018. Invest in the ride: a longitudinal analysis of the determinants of public transport ridership in 25 North America cities. *Transport. Res. Part A: Policy Pract.* 116, 434–445.
- Ceder, A., 2015. *Public Transit Planning and Operation: Modeling Practice and Behavior*, 2nd ed. CRC Press, Boca Raton, FL.
- Chapman, B., Iseki, H., Taylor, B., Miller, M., 2006. The Effects of Out-of-Vehicle Time on Travel Behavior: Implications for Transit Transfers. California Department of Transportation, Sacramento, CA.
- Feigon, S., Murphy, C., 2016. Shared mobility and the transformation of public transit (TCRP report 188). Transit Cooperative Research Program (TCRP). Transportation Research Board (TRB), Washington, D.C.
- Furth, P., Hemily, B., Muller, T., Strathman, J., 2006. Using archived AVL-APC data to improve transit performance and management (TCRP report 113). Transportation Research Board (TRB), Washington, D.C.
- Furth, P., Muller, T., 2006. Service reliability and hidden waiting time: insights from automatic vehicle location data. *Transport. Res. Rec.* 1955, 79–87.
- Hinebaugh, D., 2009. Characteristics of Bus Rapid Transit for Decision-making. National Bus Rapid Transit Institute, Tampa, Florida.
- Kittelson & Associates, I., Parsons Brinckerhoff, KFH Group, I., Texas A&M Transportation Institute & Arup 2013. Transit capacity and quality of service manual, third edition. Transportation Research Board (TRB), Washington, D.C.
- Smith, H., Hemily, B., Miomir Ivanovic, G.F., Inc., 2005. Transit signal priority (TSP): A planning and implementation handbook. U.S. Department of Transportation, Washington, D.C.
- Transsystems, Planners Collaborative Inc. & Tom Crikelair Associates 2007. Elements needed to create high ridership transit systems (TCRP report 111). Transit Cooperative Research Program (TCRP). Transportation Research Board (TRB), Washington, D.C.
- Vuchic, V., 2005. *Urban Transit: Operations, Planning and Economics*. John Wiley & Sons, Hoboken, New Jersey.

Street Design for Active Travel

Bruce Appleyard, San Diego State University, San Diego, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	333
Designing for the Needs of Pedestrians and Bicyclists	333
Street Safety and Livability	334
Need for Humane Speeds	334
Speed and Vehicle Throughput Efficiency	334
Streetscapes That Encourage Walking	335
Choosing to Walk or Bike	335
Choosing to Walk: Overcoming Distance and Time	335
A Note on Choosing to Walk: Overcoming Perceptions of Distance by Keeping Trips Interesting	336
Choosing to Bicycle: Ensuring Safety and Comfort	336
Choosing to Bicycle: Ensuring Safety and Comfort	337
Creating Livable and Complete Streets and Communities	337
Getting People Along the Street, Safely and Comfortably	338
Livable Walking Environments	339
The Needs of Bicyclists Along the Street	339
Level of Traffic Stress: Bicyclists Need Space and Protection	340
Bicycle Treatments: Improving Safety and Comfort for Cyclists Along the Street	341
Getting People Across the Street, Safely and Comfortably	341
Providing Access, Connection, and Convenience	342
Creating Better Crossings	345
Design Principles for Better Crossings	345
Importance of Human Scale	347
Summary and Conclusions	347
RESOURCES: FACILITY DESIGN	347
References	347
Further Reading	348

Introduction

This article describes the main points of creating livable and complete streets for active travel. It begins by discussing the key principles and knowledge of the street ecology such as the need for humane speeds, vehicle throughput efficiency at lower speeds, and the broader frameworks about walkability, bikability, and street livability. It then builds on this discussion to provide practical guidance on the design and creation of livable and complete street for active travel.

Designing for the Needs of Pedestrians and Bicyclists

For well over half a century, transportation and land use planning in the United States has been driven by automobile needs.¹ Therefore, creating walkable and bikeable communities requires a coordinated, consistent, and comprehensive approach. The “Es” needed to effectively promote walking and bicycling are *engineering* (of planning and building facilities), *enforcement* (of laws related to safe driving), *education/encouragement* (of all roadway users, including motorists, pedestrians, and bicyclists), and *encouragement*. To these should be added *evaluation*, or the measurement of conditions for pedestrian and bicycle travel, and *equity* to ensure that all actions are not only fair, but also try to overcome past harms. Thoughtful urban design and land use planning—including zoning, subdivision ordinances, design guidelines, and project review—are also essential to creating walkable, bikeable, and ultimately more livable streets and communities.²

¹ Some of the most vocal initial advocates for improving U.S. roads were bicycling organizations, active around the 1890s. A good reference for how city streets in the United States were transformed from being truly multimodal to being dominated by automobiles, or “motordom,” as it was called by Peter Norton, *Fighting Traffic: The Dawn of the Motor Age in the American City* (Cambridge: MIT Press, 2008).

² During a 2006 expert panel discussion on what makes a community “great,” there was unanimous agreement with University of Michigan professor Douglas Kelbaugh’s suggestion that they be safe and comfortable for all nondrivers.

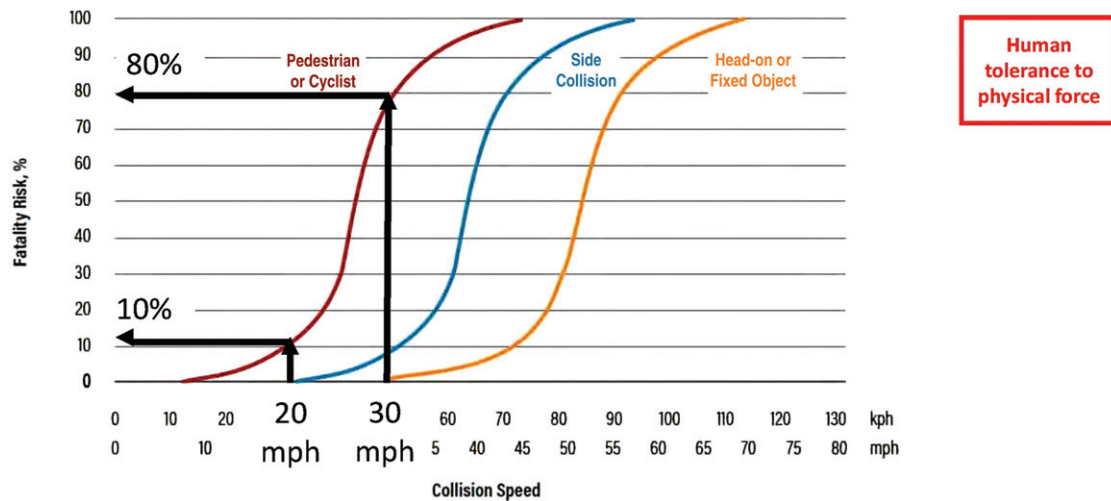


Figure 1 Fatality risk for collision speed, by crash type. Source: Offer Grembeck based on Wramborg, P. 2005. "A New Approach to a Safe and Sustainable Road Structure and Street Design for Urban Areas." Paper presented at 13th International Conference on Road Safety on Four Continents, Warsaw, Poland, October 5-7.

Street Safety and Livability

Starting in the late 1960s, Donald Appleyard championed the idea of street livability, finding that decreasing the volume and speed of traffic enhances residents' sense of comfort and actually encourages them to spend time in front of their homes, socializing and building stronger communities (Appleyard, 1981; Appleyard and Appleyard, 2020; Appleyard and Smith, 1981; Appleyard, 2006) (Fig. 1).³

Need for Humane Speeds

We should pause to ponder why fatalities increase dramatically going from 20 to 30 mph (about 5-8 times according to some studies). After presenting this research at a number of workshops, a theory posed by a neurosurgeon sticks out as a plausible explanation—we humans, as a species have likely evolved to withstand head trauma at the fastest speed at which we can run—about 20-25 mph. If so, why not argue for speed limits on the basis of our own physio-biological limits, rather than some preconceived notion of the requirements of automobile machinery and their drivers? For example, we should not let things like the 85th percentile rule drive speed limits to be determined by the highest speeds of the 85th percentile of drivers. Instead we should allow people to set and enforce lower speeds that support a desired level of street livability.

In the future we should consider limiting the speeds of autonomous vehicles in urban areas wherever pedestrians are present, and desired to be, to 12 mph, and corridors accessing them to 20 mph, optimizing street comfort and livability for all. On shared streets and woonerfs, vehicles should be governed to travel no faster than an adult or child casually walks, which can be between 2 and 3 mph.

Speed and Vehicle Throughput Efficiency

An additional advantage to keeping speeds low is to more efficiently accommodate the number of vehicles that can pass through an area (AKA "vehicle throughput"). According to a study by Michael Li, "The Role of Speed-flow Relationship in Congestion Pricing Implementation with an Application to Singapore", found that the greatest amount of throughput can occur at relatively moderate traffic speeds at about 48 kph (or about 30 mph). In short, there is little capacity advantage to designing our streets for higher speeds. So for reasons of street livability and traffic efficiency, we should keep speeds low in urban areas (Fig. 2).

Moreover, as speeds decrease, drivers are not only better able to see pedestrians and cyclists who are immediately in front of them, but are also more readily able to stop in time, and within a shorter distance, to avoid a collision. And safer streets make it easier and more attractive for people to engage in activities that will also increase their physical and creative health.

³ See Donald Appleyard, *Livable Streets* (Berkeley: University of California Press, 1981); Donald Appleyard and Daniel T. Smith, *Improving the Residential Street Environment* (Washington, D.C.: Federal Highway Administration, 1981); Bruce Appleyard, "Home in the Zone: Creating Livable Streets in the US," *Planning* (October 2006): 30-35.

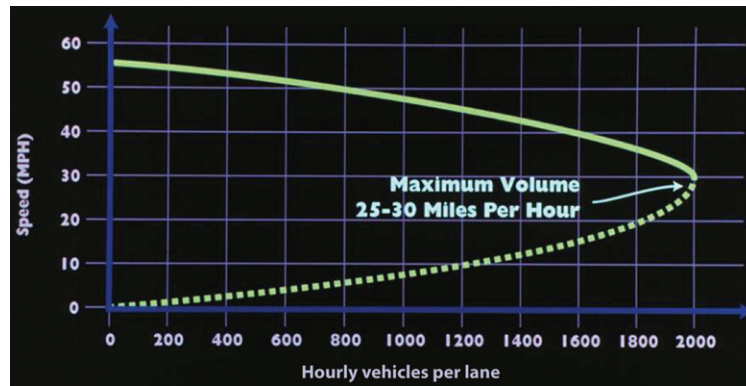


Figure 2 The relationship between mean journey speed of vehicles and total vehicle mileage on a network, suggesting that about 30 mph is a speed at which optimal vehicle throughput could be achieved. Courtesy: Walter Kulash. Source: Based on research in the 1985 *Highway Capacity Manual*, and a more recent 2002 study by Li, Michael Z.F. “The Role of Speed-flow Relationship in Congestion Pricing Implementation with an Application to Singapore.” *Transportation Research Part B: Methodological* 36, no. 8 (September 1, 2002): 731–54. [https://doi.org/10.1016/S0191-2615\(01\)00026-1](https://doi.org/10.1016/S0191-2615(01)00026-1).

Streetscapes That Encourage Walking

The kinds of streetscapes that encourage walking also appear to contribute to traffic safety: Eric Dumbaugh found that beautifying streets with “livable streetscape” elements, such as buildings with visually complex façades (often historic buildings) built close to the street and trees planted along the street (which, to many traffic engineers, decreases *automobile* safety), actually cause drivers to drive more carefully, thereby increasing the street’s safety for nondrivers (Dumbaugh, 2005).⁴ And through his research, Peter Jacobsen concluded that there is “safety in numbers”: collision rates decline as the number of pedestrians and cyclists present increases (Jacobsen, 2003).⁵ Drivers apparently travel with more care when they expect people to be on and around the street, so streets with street life and activity are safer than those devoid of it, leading to a virtuous cycle of street livability.

Choosing to Walk or Bike

To successfully support walking and bicycling, a community needs to provide both physical space and connectivity. Adequate walkways, dedicated bicycle and pedestrian spaces can help encourage these alternatives to driving.

Choosing to Walk: Overcoming Distance and Time

While many factors other than recreation enter into a person’s decision to walk (e.g., car availability and parking convenience; retail, housing, and residential land use mix; urban design), the ultimate decision is determined by the traveler’s perception of distance, time, safety, and comfort (“livability”), along with the inconvenience and cost of other modes of travel (automobiles, transit service, etc.).

Important work in this regard has been conducted in by Robert Schneider in his dissertation and subsequent works developing “a Theory of routine mode choice decisions” (Schneider, 2013). In this work, Schneider breaks down mode choice decision in this way:

- 1) Awareness and availability,
- 2) Basic safety and security
- 3) Convenience and cost
- 4) Enjoyment (personal physical, mental, and emotional benefits as well as social and environmental benefits from using the mode), and
- 5) Habit.

Furthermore, Schneider notes that each of the above are moderated by an individual’s socioeconomic characteristics. For this discussion, focus is being placed on those decisions that can be most influenced by the physical design of the public realm: #2 (Basic safety and security), #3 (Convenience and cost), and #4 (Enjoyment). Of course #1, Awareness and Availability can be influenced by

⁴ Eric Dumbaugh, “Safe Streets, Livable Streets,” *Journal of the American Planning Association* 71, no.3 (2005): 283–300.

⁵ Peter Lyndon Jacobsen, “Safety in Numbers: More Walkers and Bicyclists, Safer Walking and Bicycling,” *Journal of Injury Prevention* 9 (2003): 205–209, tsc.berkeley.edu/newsletter/Spring04/JacobsenPaper.pdf (accessed September 11, 2008).



Figure 3 Equipping buses with bike racks and allowing bicycles on railcars are effective ways to make cycling competitive, in terms of time, with regional auto trips, even if the trip begins and ends in a low-density suburban area with poor transit service. *Source: Bruce Appleyard.*

wayfinding strategies, public service announcements, and other Educational and Encouragement Campaigns. For more information, see discussion of the use of the “Es”, or in Appleyard, B. (2003).

A Note on Choosing to Walk: Overcoming Perceptions of Distance by Keeping Trips Interesting

According to a 2008 study, people are willing to walk farther than previously believed—about a half mile—to reach a transit station (Weinstein et al., 2008).⁶ The average adult walks three to four feet per second, which translates to about 2–3 mph (although children, seniors, and people with mobility limitations tend to move more slowly).⁷ Moreover, studies have shown that a pedestrian’s perception of time—and thereby his or her perception of distance—can be influenced by the aesthetic quality of the experience: a street alive with activity, human-scaled buildings with interesting façades, and a sense of enclosure tends to shorten a person’s perception of time, which is likely to extend the distance he or she is willing to walk.⁸

Choosing to Bicycle: Ensuring Safety and Comfort

More research is needed to fully determine how far bicyclists are willing to travel. Until now, studies have been hampered by small sample sizes, wide variances in trip lengths, and a failure to take into account both individual attitudes and the distinct community characteristics of the built environment, although a small sample size as early as the 2001 National Household Travel Survey it was found that the average bicycle commute-to-work distance was about three miles.⁹ According to a 2014 US Census report, the average bicycle commute time is 19.3 min, and most bicycle commutes were between 10 and 14 min long.¹⁰ If anyone is bicycling about 15 mph, then this is about 2.5–3.5 miles. But with the provision of more, and better protected bicycle lanes, greenway/bicycle boulevards, and better end-of-the-trip facilities (e.g., lockers, showers, and secure bicycle-locking facilities),¹¹ in addition to the wider availability of transit-carrying capacity for bicycles, the distances that cyclists may be able and willing to travel for work, school, and other purposes may well increase (Fig. 3).¹²

⁶ Asha Weinstein, Marc Schlossberg, and Katja Irvin, “How Far, by Which Route, and Why? A Spatial Analysis of Pedestrian Preference,” *Journal of Urban Design* 13, no. 1 (2008): 81–98.

⁷ Richard L. Knoblauch, Martin T. Pietrucha, Marsha Nitzburg, “Field Studies of Pedestrian Walking Speed and Start—Up Time,” *Transportation Research Record* 1538 (1996): 27–38, [enhancements.org/download/trb/1538-004.PDF](https://www.enhancements.org/download/trb/1538-004.PDF) (accessed September 11, 2008); John LaPlante and Thomas P. Kaeser, “A History of Pedestrian Signal Walking Speed Assumptions,” in *3rd Urban Street Symposium: Uptown, Downtown, or Small Town: Designing Urban Streets That Work*, Seattle, Washington, June 24–27, 2007.

⁸ For more on the relationship between time and distance traveled on foot, see Peter Bosselmann, *Urban Transformation: Understanding City Design and Form* (Washington, D.C.: Island Press, 2008); Peter Bosselmann, *Representation of Places: Reality and Realism in City Design* (Berkeley and Los Angeles: University of California Press, 1998), 61; and Raymond Isaacs, “The Subjective Duration of Time in the Experience of Urban Places,” *Journal of Urban Design* 6, no. 2 (2001): 109–127.

⁹ According to the National Household Travel Survey 2001, the average bicycle commute to work distance was about 3 miles (based on a sample of only 71 bicycle work commute trips reported by respondents). Of those surveyed, the average distance for all bicycle trips was about two miles (sample of 1,851 total trips). See Federal Highway Administration, *National Household Travel Survey 2001* (Washington, D.C.: U.S. Department of Transportation, 2001), nhts.ornl.gov/download.shtml.

¹⁰ <https://bikeleague.org/content/new-census-data-bike-commuting>

¹¹ Awarding zoning bonuses to developers who install such amenities in their buildings is one way to encourage bicycle use.

¹² For more information, see TCRP Synthesis 62: *Integration of Bicycles and Transit*, onlinepubs.trb.org/Onlinepubs/tcrp/tcrp_syn_62.pdf (accessed September 11, 2008).

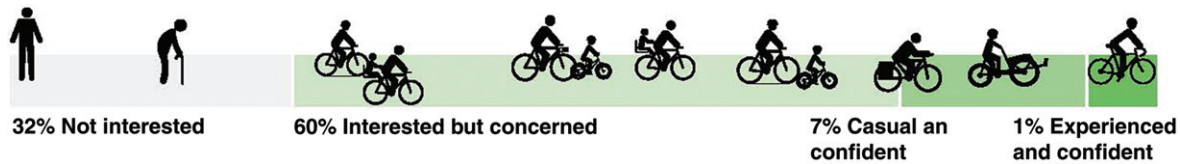


Figure 4 The four identified groups of cyclists (Geller, 2009; Dill and McNeil, 2016) Source: NACTO.

Choosing to Bicycle: Ensuring Safety and Comfort

While bicyclists and pedestrians have many similar needs, the greater levels of speed, momentum, and inertia characteristic of bicycle travel make it more critical for bicycle plans to recognize that cyclists have a broader range of comfort and skill levels than pedestrians. In 2005 Roger Geller with the city of Portland, Oregon, posited that there was a potentially large demand for bicycle commuting, provided that the right encouragement is given through both infrastructure improvements and relevant programs (e.g., safety education, active living, energy conservation). The study identified almost two-thirds of all commuters as “interested, but concerned” regarding bicycle commuting; these are commuters who would likely “ride if they felt safer on the roadways—if cars were slower and less frequent (Geller, 2009).”¹³

A 2016 survey validating Roger Geller’s theory (Geller, 2009) that there is a large group (about 60%) who are “Interested but Concerned” in bicycling (Dill and McNeil, 2016) (Fig. 4). This is in alignment with other studies that have shown that perceived safety presents a more significant barrier for those with the least experience in traffic; precisely the group with the greatest potential to adopt cycling for transport.

This finding is consistent with other research suggesting that perhaps the best way to get noncyclists to start cycling is to create bicycle greenways/boulevards (see Fig. 9). The bicycle boulevard/greenway concept is a local street and street network that has been modified with traffic calming devices and other controls to a comfortable street network to encourage even the most cautious bicyclist while maintaining local access for automobiles.

Creating Livable and Complete Streets and Communities

As described in more detail in *Livable Streets 2.0* by Appleyard, B. and Appleyard, D. (2020), to create livable and complete streets that equitably prioritize people (pedestrian, bicyclists, scooter/skateboarders, etc.) and those who live, work, and play along streets, planners, designers and engineers should strive to achieve the Strategies, Objectives, and Tactics outlined in the **Charter for Humane and Equitable Streets**. These guiding design principles can be used for (1) understanding the nature of the problem faced by your street ecology of interests, and (2) help determine the best collection of design solutions and policies going forward. This charter, specifically the Tactical Areas, such as *getting people safely and comfortably across the street*, are used to organize the discussion below—the understandings and lessons of which have been developed and refined throughout the years of working with communities (Fig. 5).

In short, livable and complete streets and communities are created with the following components:

- Safety & Comfort that is Peaceful
- Access, Connection, and Convenience
- Aesthetics and a Positive Sense of Place that is Enjoyable
- Be Ethical, Equitable, and Just (which should be monitored on all steps, and at all times, with work to overcome past inequities).

There are then several key tactical areas to achieve the above, as follows:

- Getting people along the street, *safely and comfortably*
- Getting people across the street, *safely and comfortably*
- Slow speeds and reduce traffic volumes for *safety and comfort*
- Ecosystem thinking: design and plan from the human to regional scale to provide *access, connection, and convenience*

Some of the main components needed for overcoming distance for both walkers and cyclists, and for creating a safe, inviting, and livable walking and bicycling experience, are as follows:¹⁴

¹³ Roger Geller. The City of Portland, Office of Transportation, “Four Types of Transportation Cyclists,” portlandonline.com/TRANSPORTATION/index.cfm?a=158497&c=44597 (accessed August 5, 2008).

¹⁴ For more information, see American Association of State Highway and Transportation Officials (AASHTO), AASHTO Guide for the Design of Pedestrian Facilities (Washington, D.C., AASHTO, 2004); AASHTO Guide for the Development of Bicycle Facilities (Washington, D.C., AASHTO, 1999; updated 2009); Institute of Traffic Engineers (ITE), ite.org/traffic/; and Context-Sensitive Solutions in Designing Major Urban Thoroughfares for Walkable Communities (Washington, D.C.: ITE, 2006), ite.org/bookstore/RP036.pdf (accessed September 3, 2008).

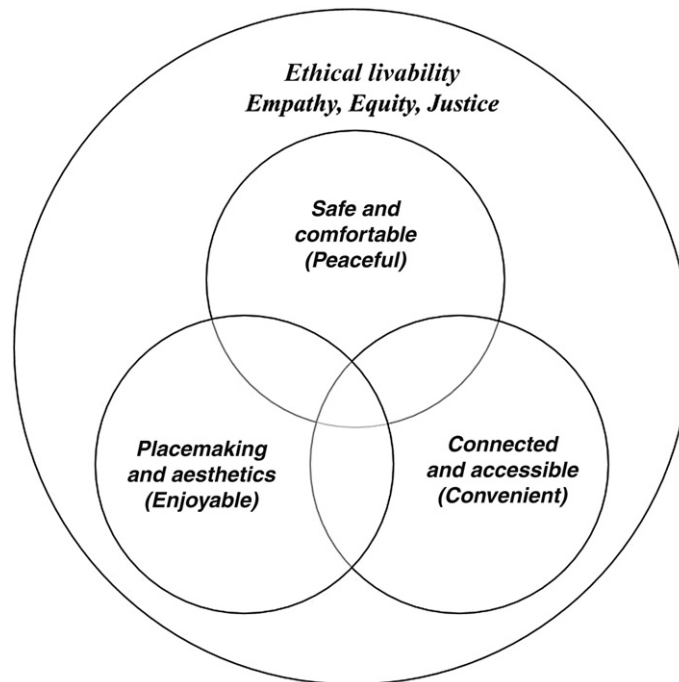


Figure 5 The four components of ethically livable and complete streets *Source: Bruce Appleyard.*

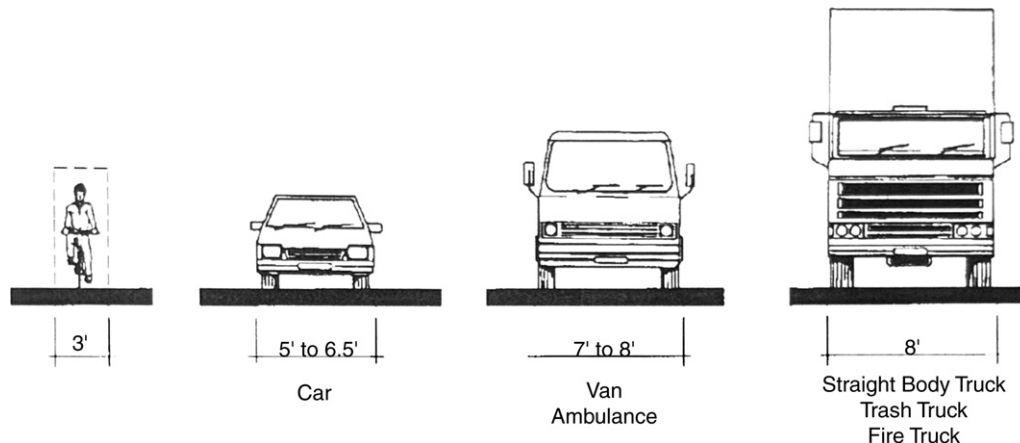


Figure 6 How wide does a street really need to be? We should design streets for what is needed, not oversized for preconceived notions. We should also use vehicles that are as small as possible—fire trucks need particular attention. *Source: Eran Ben-Joseph after Devon Engineering Department.*

Getting People Along the Street, Safely and Comfortably

Have you ever thought about how wide a lane for a car or truck really needs to be? when I ask my students how wide a car is, they usually say it's much larger than the average size of 6-6.5 feet. In fact, the widest a vehicle can be without a "wide load" sign is 8.5 feet (or 102 in.) (Fig. 6). We need to right size your lane widths and the numbers needed.

So why are lanes in the US 12, 13 feet, or even wider, when narrower lanes (10 foot) would work to balance both safety and livability, while still allowing cars to travel through.

In the 2000s, traffic engineer Peter Swift looked at the relationship between the physical characteristics of streets and the number of accidents. He found that the typical 48-foot-wide street had a crash rate that was 18 times higher than that of a 24-foot-wide street.

According to Speck:

"A typical American urban lane has historically been 10 feet wide, which comfortably supports speeds of 35 mph. A typical American highway lane is 12 feet wide, which comfortably supports speeds of 70 mph. Drivers instinctively understand the connection between lane width and driving speed, and speed up when presented with wider lanes, even in urban locations. For this reason, any urban lane more than 10 feet wide encourages speeds that increase risk to people walking."

So a suggestion is to use 10 foot lanes where you can—it is always good to proceed with these changes with consideration of the upstream and downstream story you are telling drivers.

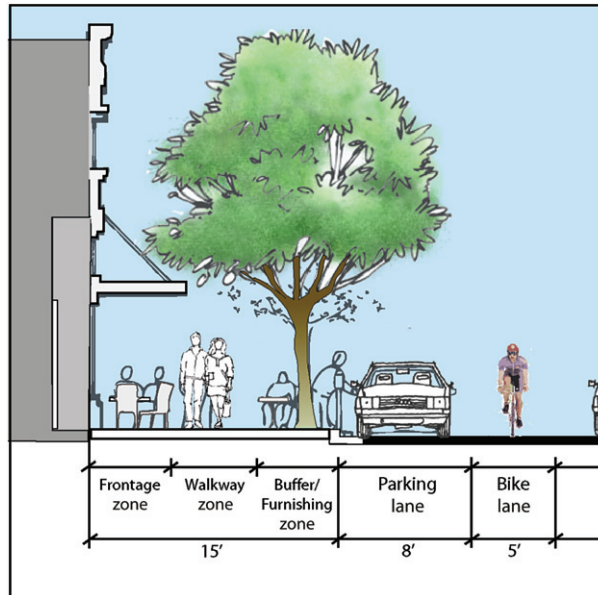


Figure 7 To permit social activity, sidewalks need to be wide enough for two people to walk side by side and a third person and/or a couple to pass comfortably. In addition, sidewalks along commercial streets should allow space for seating, displays, and traffic buffers/furniture zones. It is suggested to have at least 6–8 feet so two people can walk side by side and one person or a couple can comfortably pass. In commercial zones, at least 15 feet is needed, but places can make 10 feet work, if needed. In these places, consider parklets for curbside dining. Bike lanes should be at least 5 feet, but 6 feet is preferred. Parking should be 8 feet. Avoid placing a narrower parking lane (less than 8 feet) next to a narrow (less than 5 feet) bike lane, or else you may create a “dooring” problem. *Source: Bruce Appleyard.*

Livable Walking Environments

- *How wide should the sidewalk be?* As walking should be viewed as a social activity, paths should be at least six to eight feet wide (seven to nine feet, if the walkway has a wall on one side) to provide enough room for two people to walk side by side and a third person or a couple to pass comfortably. Sidewalks along commercial streets should accommodate the interaction between a building’s activity and street life by allowing space for seating, displays, etc., as well as walkway space and traffic buffers/furniture zones, as described above. Fifteen feet appears to be a workable width; sidewalks may be even wider in areas with high levels of pedestrian activity (Fig. 7).

A note about traffic buffers: Sidewalks should be buffered from traffic annoyances (threats to personal safety, noise, etc.). Buffers can be provided by on-street parking, bike lanes, and a “furniture zone” that might include lights, signs, benches, transit shelters, planters, and/or trees.

As Speck (2018) points out, we need to recognize that “The walkable city is the rollable city, and when a city works well for people in wheelchairs, it works well for everyone”. Therefore we need to make sure wheelchairs at all times have at least 36 in. (about a meter) clearance width and directional curb-ramps at every intersection that direct people into the line of travel (not into the middle of the intersection) and that they have yellow tactile bumps for the sight impaired.

The Needs of Bicyclists Along the Street

A cyclist in motion requires width to maintain balance and to weave to the extent necessary to move forward while keeping the bicycle upright; “shy distance” is also necessary to separate the bicyclist from curbs, posts, and other potential hazards. Combining these allowances with the width of an average bicycle means that a bicyclist will need at least a five-foot-wide space to ride comfortably.¹⁵

In cases where the road is narrow, wide enough perhaps for only one bike lane, a “climbing lane” could be added on the uphill side and an in-pavement “sharrow” painted in the downhill lane to remind drivers to share the lane with cyclists.¹⁶ A “Share the Road” sign could be posted as well.

¹⁵ The space occupied by a bicycle and its rider is relatively modest. Generally, bicycles are between 24 and 30 in. wide from one end of the handlebars to the other. An adult tricycle or a bicycle trailer, on the other hand, is approximately 32–40 in. wide.

¹⁶ The incorporation of pedestrian and bicycle facilities into design and construction, originally referred to as “routine accommodation” is now commonly known as “completing the streets,” which includes a movement toward retrofitting existing streets; see completethestreets.org.

LEVEL OF TRAFFIC STRESS

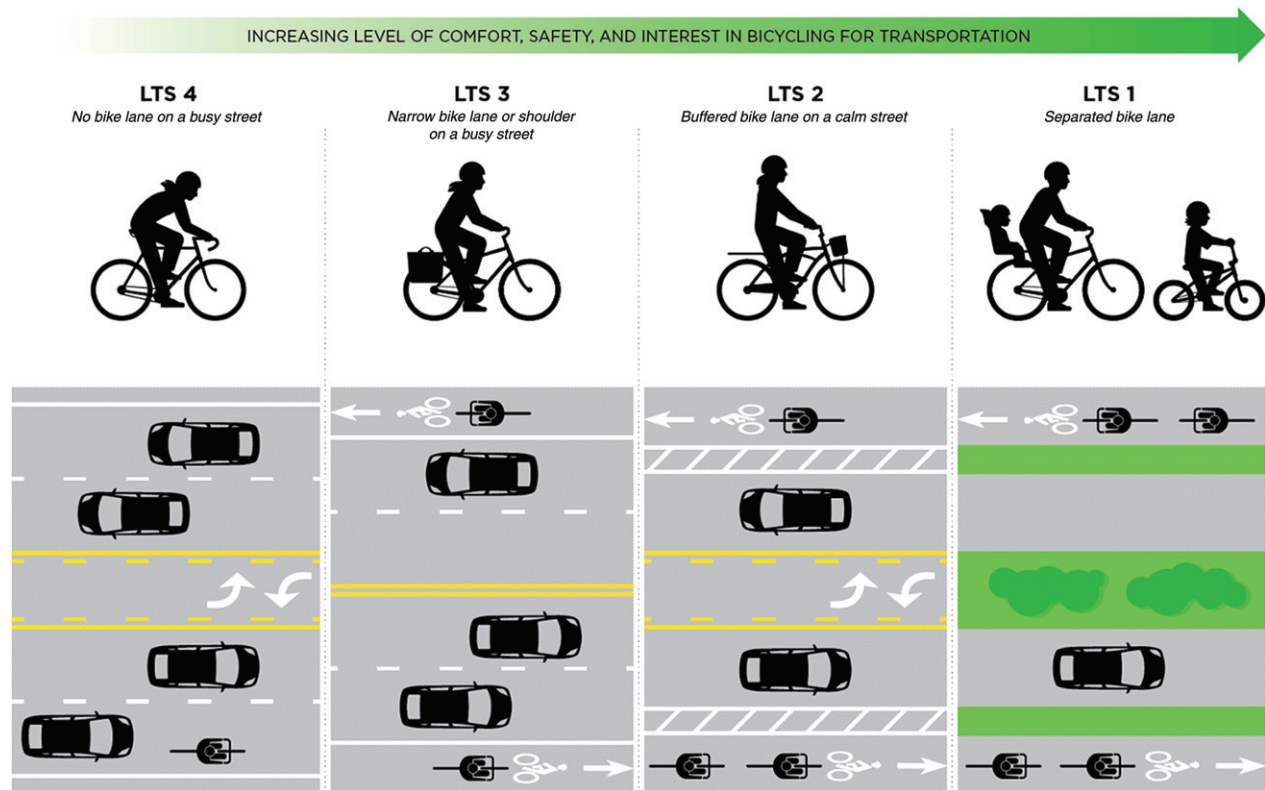


Figure 8 The street design of bicycle traffic stress. Source: Alta Planning <https://blog.altaplanning.com/level-of-traffic-stress-what-it-means-for-building-better-bike-networks-c4af9800b4ee>.

Sharrows, however, should be used with caution, as research is showing they can actually be more dangerous (Ferenchak et al., 2019).¹⁷ In all cases, where possible, build separated and protected bike lanes, as research shows these can lead to overall street safety, even for drivers (Marshall and Ferenchak, 2019).¹⁸

Level of Traffic Stress: Bicyclists Need Space and Protection¹⁹

Along these lines, there have been several efforts to measure cyclist stress levels along sections of roadways (Harkey et al., 1998; Mekuria et al., 2012; Mekuria et al., 2017). Harkey et al. (1998) included such things as the presence of negative factors such as vehicle volumes, speed (as indicated by posted speed), driveways, and then positive factors such as wider bicycle lanes or paved shoulders, as shown in both the table and figure below.²⁰ Basically, bicyclists need space and protection!

LTS scores range from 1 to 4, with LTS 1 being the most comfortable for all riders, including children, and LTS 4 appealing to only the most confident bicyclists (commonly described as “strong and fearless,” who are comfortable and sometimes prefer to travel in mixed traffic) (Fig. 8).

In response, we need to develop cycling infrastructure that lowers the actual and perceived risk, and frankly stress, from traffic to attract the large number of potential riders who make up the “Interested but Concerned” group discussed above to meaningfully invite all people, regardless, gender, age, income or race bicycling abilities.

¹⁷ Ferenchak, Nicholas N., and Wesley E. Marshall. “Advancing Healthy Cities through Safer Cycling: An Examination of Shared Lane Markings.” *International Journal of Transportation Science and Technology* 8, no. 2 (June 1, 2019): 136-45. <https://doi.org/10.1016/j.ijtst.2018.12.003>.

¹⁸ Marshall, and Ferenchak. “Why Cities with High Bicycling Rates Are Safer for All Road Users - ScienceDirect.” Accessed May 30, 2019. <https://www.sciencedirect.com/science/article/pii/S2214140518301488?via%3Dihub>.

¹⁹ This section had parts adapted from Caltrans’ Smart Mobility Framework How-To Guide

²⁰ Handy and Xing identified high monthly parking costs, the social environment of the workplace and a residential preference for a good environment for cycling, a measure of self selection, as positively impacting bicycling travel behaviors (Handy and Xing, 2011).

Methods for conducting a Level of Stress analysis can be found in the paper by Mekuria, Furth, and Nixon. *Low-Stress Bicycling and Network Connectivity* published by the Mineta Transportation Institute²¹. Fig. 8 illustrates some of the LTS scores for streets with a variety of speeds and lanes. Planners can use these tables to perform a basic LTS analysis for facilities.

Bicycle Treatments: Improving Safety and Comfort for Cyclists Along the Street

Drawing on suggestions from Pucher and Dijksma (2003), who compared bicycle infrastructure improvements U.S. and Europe, the following are some design suggestions to make streets work better for bicyclists.

1. Special bicycle streets (bike boulevards or greenways) Fig. 9 which permit car traffic but give bicyclists right-of-way priority over the entire breadth of the street, with cars prohibited from rushing bicyclists or otherwise interfering with them.

A refinement of the shared roadway concept, the bicycle greenway/boulevard (as shown in the figures above and below) is a local street that has been modified with traffic calming devices and other controls to function as a through street for bicycles while placing a priority on local (over through) access for automobiles.

While some suggest 3000 ADT is OK for bicycle greenway/boulevards, I suggest carrying as little as possible and no more than 1500 vehicles a day—which roughly translates into about 150 vehicles during the afternoon peak hour, or about an average of one vehicle passing roughly every 30 seconds, which seems reasonably comfortable (greater gaps in time may occur as cars tend to platoon).

Prototypical bicycle boulevards function as through streets for bicycles while maintaining only local access for automobiles through the use of traffic calming devices that reduce car speeds and discourage through traffic. Traffic controls should be designed to limit conflicts between motorists and bicyclists and give priority to all bicycle movements. Manage traffic on bike boulevards so there are no more than 1500 vehicles a day. Some other considerations are as follows:

1. If streets are one-way for cars, make them two-way for bicyclists.
2. Create reserved bus lanes that can be used by bicyclists but not by cars.
3. Street networks with deliberate dead ends and circuitous routing for cars but direct, fast routing for bikes, including special cut-through short-cuts (Fig. 10).
4. Allow bicyclists to make left and right turns where prohibited for motor vehicles. In addition, allow bicyclists to make right turns on red, while motorists cannot (Fig. 10).

Getting People Across the Street, Safely and Comfortably

Street crossings and intersections: Safety at crossings can be cost-effectively improved through the use of signs, in-pavement markings, and clearly visible crosswalks so that drivers will proceed cautiously.²² Planners might consider ways to minimize pedestrians' and cyclists' crossing distances and exposure to traffic, and to make it easier for these travelers to see and be seen by motorists. Such improvements can be accomplished by "breaking up the task" of crossing a street with medians and islands, while "shortening the distance of that task" with curb extensions (bulb-outs) and smaller, tighter corner radii. For cyclists, a recent innovation are the protected intersections (Fig. 11).

Separated and protected intersections (see figure below) further facilitate crossing streets and making left turns. These designs greatly enhance the corners with island-separated, amply-sized refuge areas for bicyclists and pedestrians that enable them to wait far forward of the motor vehicle limit line. The refuges are shifted outward substantially so a turning vehicle can pause while rounding the corner to yield to through bicyclists and pedestrians without being rear-ended by through traffic. Bicyclists have a separate "cross-bike" area adjacent to the pedestrian crosswalk (Figs. 11 and 12).

And intersection treatments such as pavement markings, warning signs, raised crossings, and merging cyclists in advance of an intersection onto an on-street (non-buffered) bike lane have been shown to improve safety (Frank et al., 2019) (Buehler and Dill, 2016; DiGioia et al., 2017; Thomas and DeRobertis, 2013) (Fig. 13).

If you cannot create a complete protected intersection, a partial, but complementary strategy is to put in bicycle boxes—a waiting area that is clearly marked for cyclists at signalized intersections in front of waiting cars (see Fig. 14).²³

²¹ Maaza C. Mekuria, Peter G. Furth, and Hilary Nixon, *Low-Stress Bicycling and NetworkConnectivity*, MTI Project 1005, May 2012 Retrieved from: <http://transweb.sjsu.edu/sites/default/files/1005-low-stress-bicycling-network-connectivity.pdf>

²² Crosswalks in commercial areas should be at least 12 feet wide to allow people to flow in both directions. On wide streets, crossing islands with a median nose provide added protection. The clear path for pedestrians through the median island should be six feet. Countdown signals that let pedestrians know how much time they have left to cross reduce stress and accidents. See Paul Zykofsky and Dan Burden, "Walkability," in *Planning and Urban Design Standards* (Hoboken, N.J.: John Wiley & Sons, 2006), 478–480.

²³ For more information, see City of Portland, Office of Transportation, "What Is a Bike Box," portlandonline.com/shared/cfm/image.cfm?id=185112 (accessed September 11, 2008).

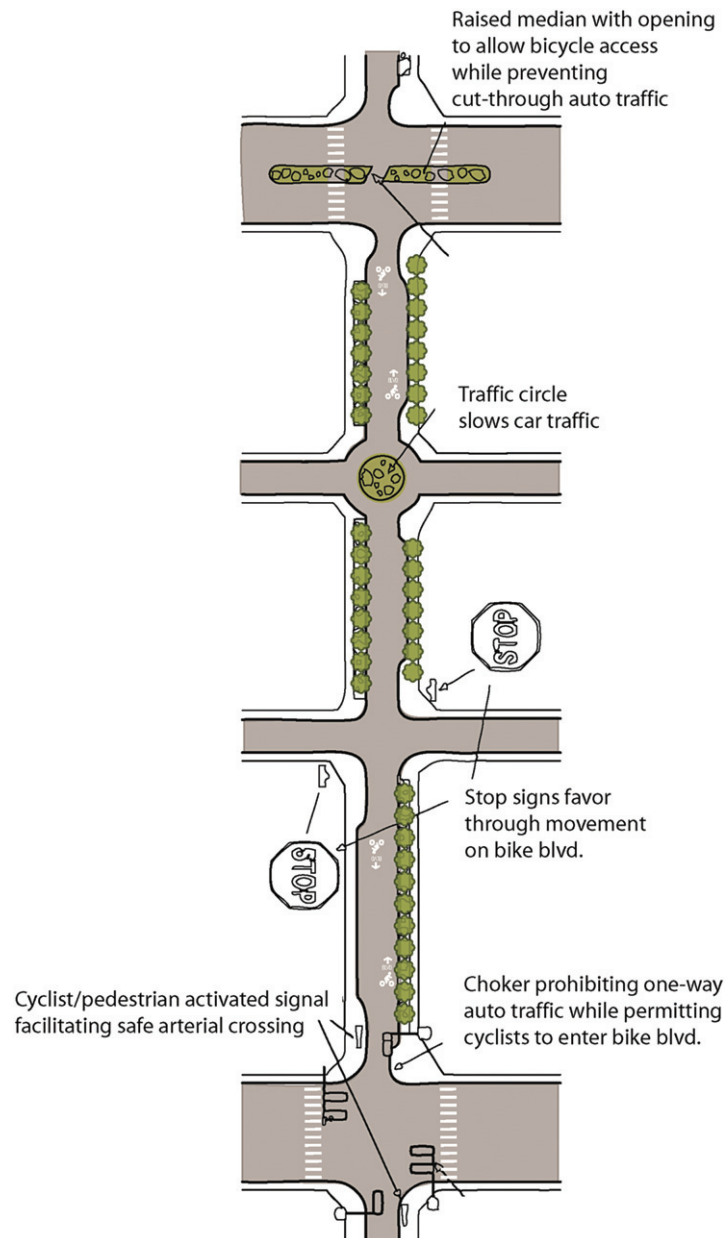


Figure 9 Overview of a bike blvd/greenway which gives bicyclists priority over drivers through the use of traffic and barriers that are permeable to only bicyclists and pedestrians. *Source: Bruce Appleyard.*

Providing Access, Connection, and Convenience

Providing access, connection, and convenience has as much to do with how land use activities are planned out and street networks are designed. Here are some key considerations.

- *A network of safe, direct, and comfortable routes and facilities:* A 2004 Planning Advisory Service report recommends that pedestrian (and bicycle) path connections be every 300-500 feet; for motor vehicles, they authors recommend 500-1000 feet.²⁴ For new development, these standards can be implemented through subdivision ordinances.²⁵

²⁴ Susan Handy, Robert G. Paterson, and Kent Butler, *Planning for Street Connectivity: Getting from Here to There*, PAS Report #515 (Chicago: APA Planners Press, 2004).

²⁵ The regional government of Portland (Oregon) Metro requires street connectivity in its regional transportation plan and in the development codes and design standards of its constituent local governments as follows: local and arterial streets must be spaced no more than 530 feet apart (except where barriers exist); bicycle and pedestrian connections must be made (via pathways or on road rights-of-way) every 330 feet; and cul de sacs (or dead-end streets), which are discouraged, can be no longer than 200 feet and have no more than 25 dwelling units.



Figure 10 Russell St. Berkeley, CA. By making street closures permeable for pedestrians and bicyclists, pedestrians and bicyclists can pass through these while vehicle traffic is diverted and disincentivized. *Source: Bruce Appleyard.*

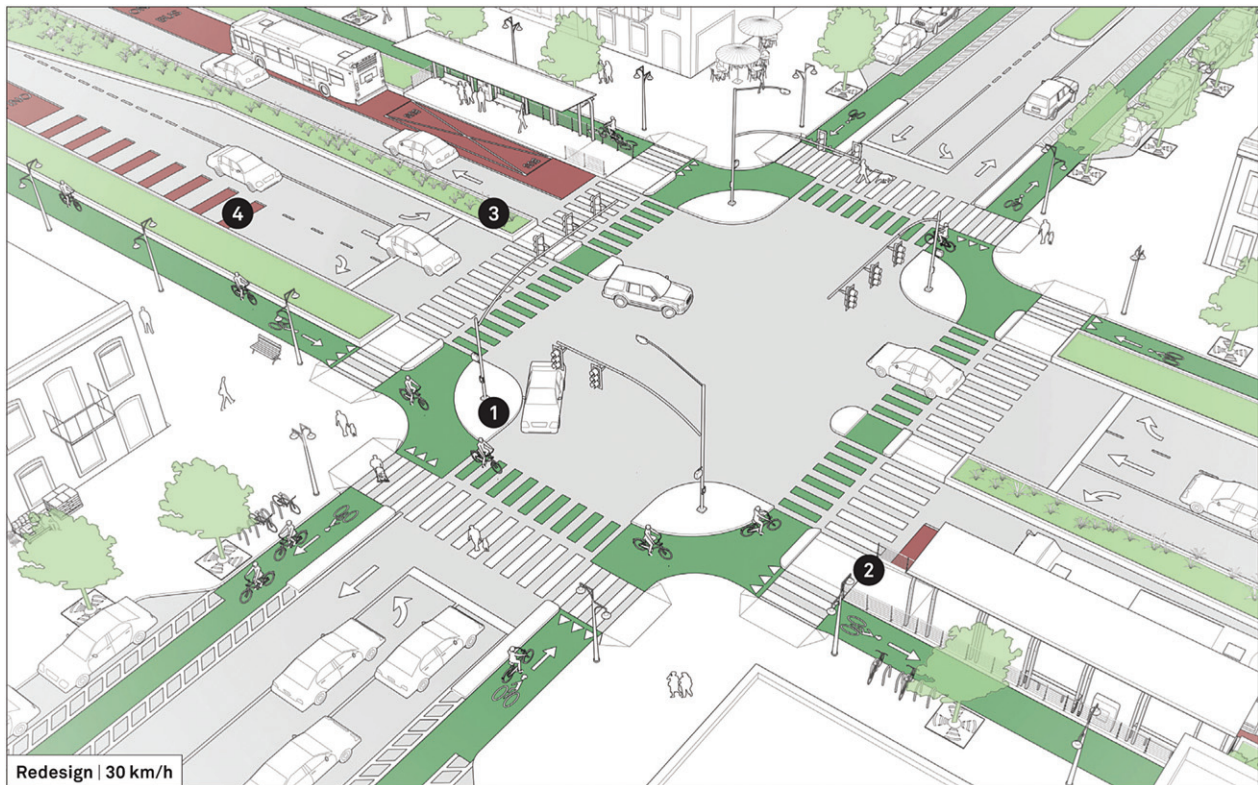


Figure 11 Separate and protected intersection for bicyclists, also known as a Dutch intersection, provides safe refuge spaces for cyclists; and cyclists are given priority position using advanced stop boxes, leading signal priority, and smaller curb radii to slow vehicles turning across the cycle path. *Source: NACTO.*

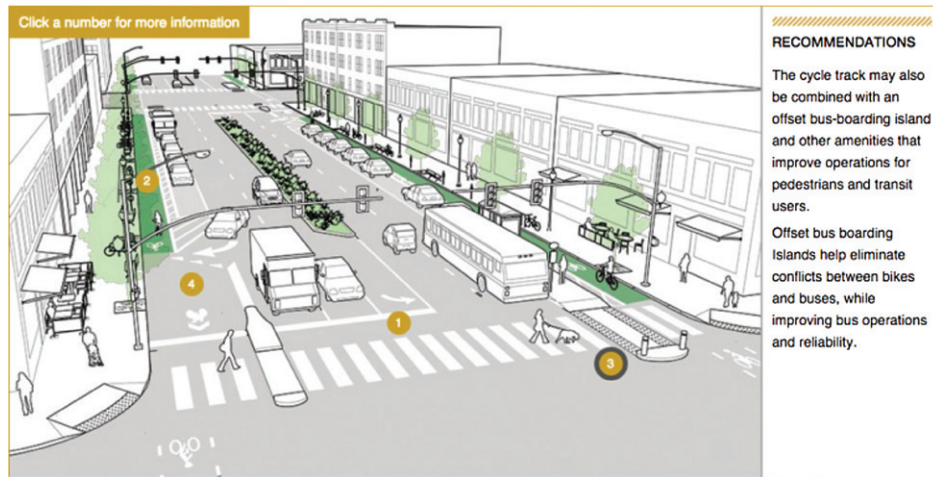


Figure 12 One can also create separated and protected bike lanes alongside a bus boarding island. *Source: NACTO.*



Figure 13 You can even put bikelanes down the middle of a road, as is done on Pennsylvania Avenue in Washington DC. *Source: Bruce Appleyard.*



Figure 14 A bicycle box in Portland, OR which gives bicyclists priority over cars at intersections. *Source: Bruce Appleyard.*

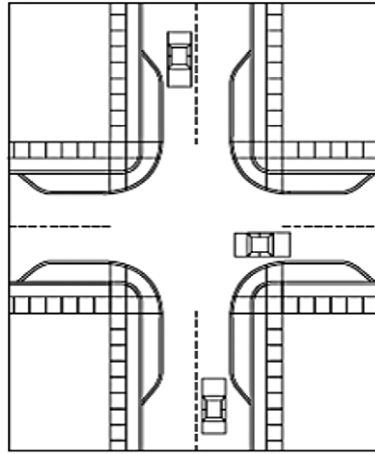


Figure 15 Pedestrian bulb-outs shorten the task of crossing while empowering pedestrians to be better seen, helping them cross the street with more safety and comfort. Source: Donald Appleyard.

Creating Better Crossings

To create better crossings, consider some of the following principles. First, there are two mainly physical design principles that should be applied, as follows:

1. Shorten the Distance
 - a. Use curb bulb-outs and/make curb radii smaller;
 - b. Narrow streets by either narrowing lanes and/or lowering number of lanes;
2. Break up the Task (of crossing the street) with such things as medians and refuge islands.

Two more functional/operational principles are needed to address the informal negotiation and choreography between pedestrians and drivers:

1. Empower pedestrians to more accurately assess their own risk, AND allow them to *assert* their presence.
2. Enhance a driver's ability to anticipate and be ready to stop for pedestrians in the street.

Design Principles for Better Crossings

1. Shorten the Distance and Improve Visibility of Pedestrians by the Drivers
 - a. First, minimize pedestrians' and cyclists' exposure to traffic by *shortening the crossing distance and breaking up the task*. Some of the key measures that help achieve this are medians and islands, curb extensions (bulb-outs) and tighter corner radii.

Pedestrian Bulb-outs or Neckdowns ("Chokers")

Pedestrian Bulb-outs (Neckdowns) are devices that physically narrow the street width at either intersections or midblock. These are useful at pedestrian crossings where pedestrians wishing to cross can be better seen by motorists. They also shorten the task of crossing the street (Fig. 15).

The great advantage of bulb-outs for the task of crossing the street is they improve visual contact between drivers and human travelers (pedestrians/bicyclists)-an important component to the safety of the negotiated choreography of crossing the street. Bulb-outs essentially increase the sidewalk area at an intersection, and can also provide a location for traffic signs closer to the drivers' normal line of sight, where obstruction is also less likely to occur.

Consideration should be given to how bulb-outs can act as a barrier to a bicyclist's freedom of movement at intersections.

Importance of Urban Design and Placemaking: Qualities Believed to be Highly Correlated with Walkability

Urban design experts identified the following qualities believed to be highly correlated with walkability (Ewing and Handy, 2009).²⁶ Appleyard (2012, 2016) finds empirical evidence suggesting that these elements are indeed associated with the increased likelihood one may decide to walk or bicycle (in these cases, to access transit stations). Furthermore, research efforts testing measures of aesthetics found 25 and even 41% positive differentials in pedestrian and bicycle activity in prescribed circumstances (Heath et al., 2006).²⁷

²⁶ These measures are currently being evaluated in the field, but focus only on commercial streets.

²⁷ Heath et al. (2006) systematically explored a number of aesthetic interventions to see which factors increased physical activity. Community-scale and street-scale design factors, such as improved street lighting, adding center islands, and landscaping, were found to be the most effective.

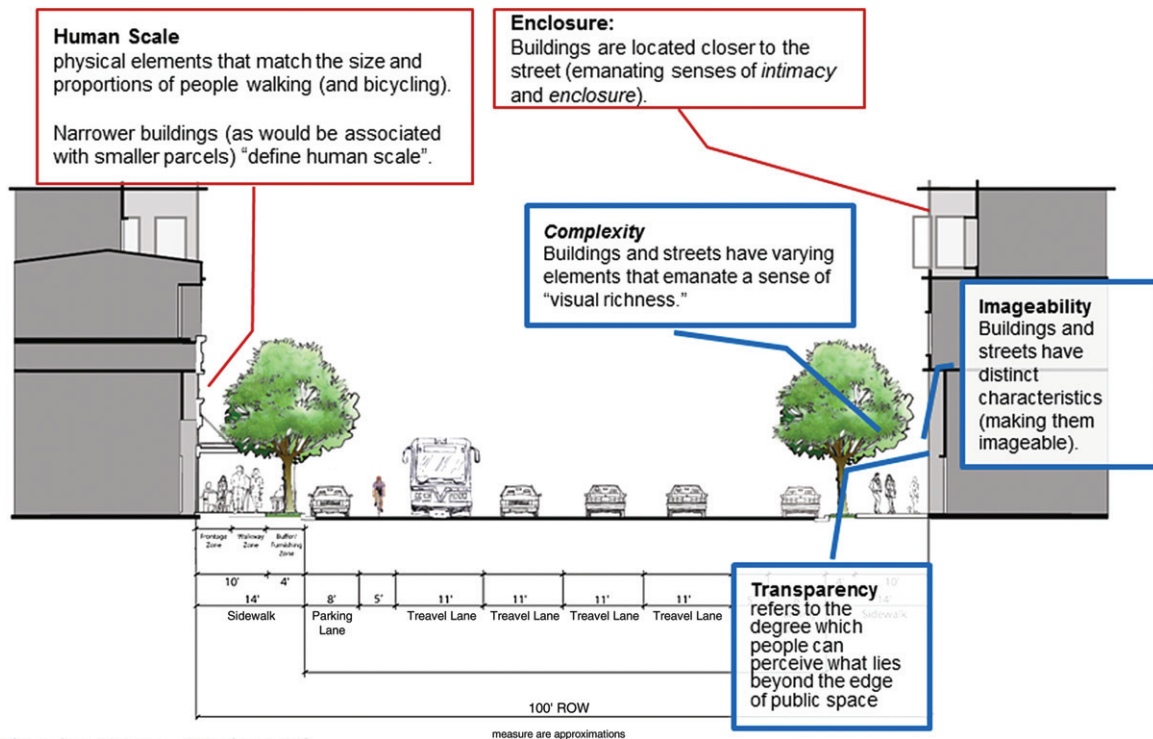


Illustration by Bruce Appleyard

Figure 16 This figure illustrates findings from Appleyard (2012, 2016) in support of conclusions drawn from a survey of urban design experts by Ewing and Handy (2009) regarding the key urban design qualities of the street environment that encourage walking, including (1) provide a "human scale"; (2) provide a stronger sense of enclosure; (3) emanate imageability; and (4) have windows on the ground floor (transparency) and (5) detailed fenestration, emanating a visually rich sense of complexity. Source: Bruce Appleyard.

The urban form qualities important to walking are as follows:

- **Imageability** is the quality of a place that makes it distinct, recognizable, and memorable.
- **Enclosure** is the degree to which streets and other public spaces are visually defined by buildings, walls, trees, and other vertical elements.
 - For this, I suggest all buildings along a street be designed to have a relationship and connection to the street. They could have a driveway or even parking to the side—but they all need to be designed intentionally with the street in mind. In terms of enclosure, Speck (2018) points out that this enclosure is part of our desire for refuge and why good places need good edges and streetwalls to create "outdoor living rooms." He goes on to explain that "the location of those walls, and their proper height relative to the width of the space, is a subject that has been under discussion for centuries." He turns to the *Lexicon of the New Urbanism* and how "many consider 1:1—the Renaissance ideal—to be the best for streets, while, beyond 1:6 (with 1 being the building and 6 being the road), the sense of spatial definition may be lost. But he points out "even 5:1 can be lovely; think medieval Salzburg"
- Human scale is about size, texture, and articulation of physical elements that match the size and proportions of humans and, equally important, correspond to the speed at which humans walk.
- Transparency is the degree to which people can see or perceive what lies beyond the edge of a street or other public space, particularly human activity.
- Tidiness is the importance of operations and maintenance of a place and making sure that community members feel responsible for their property and community.
- Complexity is the visual richness of a place.

This list could also include measures recognizing the importance of perceptions of distance (see Isaacs, 2001) and the other measures of vitality, livability, and sense of place (see Bosselmann, 2012, Article 3).

Importance of Human Scale

Human scale urban design elements are also shown to matter to encourage walking and bicycling rates. Appleyard (2010, 2012, 2016), finds highly likelihoods of people choosing more environmentally beneficial and active human transport options (walking and bicycling) when:

- Streets are composed of buildings that are at a human scale, are located closer to the street (emanating a stronger sense of *enclosure*), have distinct characteristics such as front porches (emanating a sense of *imageability*), and providing a “visual richness” (sense of *complexity*).
- Communities are designed with narrower, well-connected streets and/or more direct walking and bicycling paths.
- When small, personal service retail opportunities are provided.

A recent definition of livable places from the US Transportation Research Board states that they are “Places where transportation, housing, and commercial development investments are coordinated to better serve the people living in those communities.”

We should recognize, however, that this is largely achieved through land use planning and urban design, primarily at a regional or subregional level of coordination. Of course this land use planning needs to be coordinated with transport investments.

Nevertheless, we should recognize that we need to start solving transportation problems through land use as well as transportation solutions—oftentimes transport professionals focus only on transportation solutions.

In his 2012 report for the German Marshall Fund,²⁸ Denver Igarta concludes that there are three key lessons to create livable streets:

1. Emphasize sojourning (opportunities to interact or linger).
2. Prioritize active transportation
3. Ensure a human scale and pace

Summary and Conclusions

This article reviewed a broad range of aspects involved with creating livable and complete streets for active travel. It began by discussing important issues facing active travelers, such the need for humane speeds, vehicle throughput efficiency at lower speeds, and factors influencing the choices to walk and bicycle. This article presented a series of overarching goals for livable and complete streets such as the need for feelings of safety & comfort; access, connection, and convenience; a positive sense of place (importance of aesthetics); and the importance of people feeling they are in an ethical, equitable, just and caring street ecology. It then built on this framework to provide a practical guide for designing and creating livable and complete streets for active travel.

RESOURCES: FACILITY DESIGN

A variety of resources for facility design are available. A selection of leading technical practices are listed below.

1. NACTO Urban Street Design Guide
<https://nacto.org/publication/urban-street-design-guide/>
2. NACTO Urban Bikeway Design Guide
<https://nacto.org/publication/urban-bikeway-design-guide/>
3. NACTO Transit Street Design Guide <https://nacto.org/publication/transit-street-design-guide/>
4. FHWA Achieving Multimodal Networks: Applying Design Flexibility and Reducing Conflicts
https://www.fhwa.dot.gov/environment/bicycle_pedestrian/publications/multimodal_networks/fhwahep16055.pdf
5. FHWA Separated Bike Lane Planning and Design Guide- https://www.fhwa.dot.gov/environment/bicycle_pedestrian/publications/separated_bikelane_pdg/page00.cfm
6. FHWA Small Town and Multimodal Networks https://www.fhwa.dot.gov/environment/bicycle_pedestrian/publications/small_towns/fhwahep17024_lg.pdf
7. FHWA PedSafe/Bike Safe Countermeasures
http://www.pedbikesafe.org/bikesafe/guide_analysis.cfm
8. APTA Design of On-street Transit Stops and Access from Surrounding Areas
<https://www.apta.com/resources/hottopics/sustainability/Documents/APTA%20SUDS-RP-UD-005-12%20On%20Street%20Transit%20Stops.pdf>

References

- Appleyard, D., 1981. *Livable Streets*. UC Press.
 Appleyard, B., Appleyard, D., 2020. *Livable Streets 2.0*. Elsevier.
 Appleyard, B., 2012. Sustainable and healthy travel choices and the built environment. *Transport. Res. Rec.* 1, 38–45, doi:10.3141/2303-05.

²⁸ Igarta, Denver. “Livable Streets Where People Live: Policy Lessons on Broadening the Civic Role of Residential Streets from Munich, Rotterdam, Copenhagen, and Malmo,” December 12, 2012. <http://www.gmfus.org/publications/livable-streets-where-people-live-policy-lessons-broadening-civic-role-residential>.

- Appleyard, B., 2016. New methods to measure the built environment for human-scale travel research: individual access corridor (IAC) analytics to better understand sustainable active travel choices. *J. Transport Land Use* 9 (2), 121–145.
- Bosselmann, P., 2012. *Urban Transformation: Understanding City Form and Design*. Island Press.
- Buehler, R., Dill, J., 2016. Bikeway networks: a review of effects on cycling. *Transport Rev.* 36 (1), 9–12.
- DiGioia, Jonathan, Kari Edison, Watkins, Yanzhi, Xu, Michael, Rodgers, Randall, Guensler, 2017. Safety impacts of bicycle infrastructure: a critical review. *J. Safety Res.* 61, 105–119, doi:10.1016/j.jsr.2017.02.015.
- Dill, J., McNeil, N., 2016. Revisiting the four types of cyclists: findings from a national survey. *Transp. Res. Rec.* 2587 (1), 90–99.
- Dumbaugh, E., 2005. Safe streets, livable streets. *J. Am. Plann. Assoc.* 71 (3), 283–300.
- Ewing, R., Handy, S., 2009. Measuring the unmeasurable: urban design qualities related to walkability. *J. Urban Design* 14 (1), 65–84.
- Ferenchak, Nicholas, N., Marshall, W.E., 2019. Advancing healthy cities through safer cycling: an examination of shared lane markings. *Int. J. Transport. Sci. Technol.* 8 (2), 136–145, doi:10.1016/j.ijtst.2018.12.003.
- Frank, Ulmer, Appleyard, Bigazzi, 2019. *Urban Design Companion*. In: Loukaitou-Sideris & Banerjee eds, Routledge.
- Geller, R. 2009. Four types of cyclists. Portland Office of Transportation. www.portlandoregon.gov/transportation/44597?a=237507.
- Harkey, D.L., Reinfurt, D.W., Knuiman, M., 1998. Development of the bicycle compatibility index. *Transport. Res. Rec.* 1636 (1), 13–20.
- Heath, G.W., Ross, C., Brownson, Judy, Kruger, Rebecca, Miles, Kenneth, E.Powell, Leigh, T. Ramsey, 2006. The effectiveness of urban design and land use and transport policies and practices to increase physical activity: a systematic review. *J. Phys. Activity Health* 3 (2006), S55.
- Isaacs, R., 2001. The subjective duration of time in the experience of urban places. *J. Urban Design* 6 (2), 109–127, doi:10.1080/13574800120057809.
- Jacobsen, P.L., 2003. Safety in numbers: more walkers and bicyclists, safer walking and bicycling. *J. Inj. Prev.* 9, 205–209.
- Maaza C. Mekuria, Peter G. Furth, and Hilary Nixon 2012. Low-Stress Bicycling and Network. Connectivity, MTI Project 1005.
- Mekuria, M.C., Appleyard, B., Nixon, H., 2017. Improving Livability Using Green and Active Modes: A Traffic Stress Level Analysis of Transit, Bicycle, and Pedestrian Access and Mobility.
- Marshall, Ferencak. "Why Cities with High Bicycling Rates Are Safer for All Road Users - ScienceDirect." Accessed May 30, 2019. <https://www.sciencedirect.com/science/article/pii/S2214140518301488?via%3Dihub>.
- Pucher, J., Dijkstra, L., 2003. Promoting Safe Walking and Cycling to Improve Public Health: Lessons From The Netherlands and Germany. *Am. J. Public Health* 93 (9), 1509–1516.
- Thomas, Beth, DeRobertis, Michelle, 2013. The safety of urban cycle tracks: a review of the literature. *Accid. Anal. Prevent.* 52, 219–227.
- Schneider, R.J., 2013. Theory of routine mode choice decisions: an operational framework to increase sustainable transportation. *Transport Policy* 25, 128–137.

Further Reading

- Livable Streets 2.0. Bruce & Donald Appleyard. 2020. Elsevier.
- Walkable City Rules. Speck, Jeff. Island Press, 2018. <https://islandpress.org/books/walkable-city-rules>.
- NACTO Urban Street Design Guide <https://nacto.org/publication/urban-street-design-guide/>.
- NACTO 2000 Urban Street Bikeway Design Guide [ps://nacto.org/publication/urban-bikeway-design-guide/](https://nacto.org/publication/urban-bikeway-design-guide/).
- Complete Streets: Best Policy and Implementation Practices PAS Report 559. By Barbara McCann, Suzanne Rynne, American Planning Association.

Transit Planning and Management

Zakhary Mallett, Marlon G Boarnet, University of Southern California, Los Angeles, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	349
Transit in the Context of Transportation History	349
Modes of Transit and Mode Shares	350
Transit Mode Shares	350
Capacity Across Transit Modes	350
Contemporary Public Transit Governance	350
Transit Management	353
Transit Operations	353
Capital Project Planning	354
Transit Finance	354
Public Financing	355
Other Considerations	356
The Future	356
Summary and Conclusions	357
Biographies	357
See Also	357
References	358
Further Reading	358

Introduction

Passenger transit is shared passenger transport in which the provider of the transport is independent of its consumers. Transit is most often discussed or studied with regard to *mass transit*, or the transport of very large scales of people using high-capacity vehicles, and is usually operated in the form of *fixed route service*, in which the path of travel of the vehicle is fixed with only service frequency being scalable to demand patterns. Finally, transit in most societies is provided as *public transit*, making its operations, planning, and management subject to much public discord.

The remainder of this chapter begins with a section that discusses how transportation patterns have evolved over time with a particular focus on transit. In subsequent sections, we define different transit modes and their mode shares, transit governance practices, the management and operations of transit, transit finance, and then close with a discussion about contemporary and future challenges facing transit in light of new mobility.

Transit in the Context of Transportation History

The earliest form of transit was the horse-drawn carriage, which debuted in seventeenth century France, but did not become popularized until the 1800s. Because travel distance during this era was constrained by speed, much travel was local and central cities were especially dense. It was not until the 1890s that motorized travel, in the form of streetcar railroads, began to emerge as a serious travel mode. Like horse-drawn carriages, streetcars were operated by private companies, so fares needed to cover costs. But unlike horse-drawn carriages that used public rights of way and competed through differentiation, streetcar railroads were monopolies that owned and operated on their own tracks. As a result, while streetcar railroads provided higher travel speed and hence an ability to live further from central cities, such options were exclusive to the well-off due to the cost of the service, particularly as streetcars expanded outward from city centers. Streetcars introduced the lifestyle of suburbia, but because of transportation costs, only the wealthy typically had access to this lifestyle (e.g., Warner, 1962).

However, the combination of over-investment, limited demand, inefficient financial structures, and evolving regulation made the longevity of private streetcar railroads short (Jones, 1985; Warner, 1962). The invention of streetcars in the 1890s was a major leap in transportation technology. Investors wanted a hand in the invention and the perceived financial gains it offered, so many streetcar operations were subsidiaries of agglomerated investment enterprises that also invested in public utilities and real estate. In fact, enterprise companies began to rely on the sale of real estate along their railroads to finance their operations and maintenance. But as more and more private streetcar railroads laid tracks to get “in” on this newfound venture and revenues from tangential sources dried up, demand became spread while maintenance needs and reliance on fares increased—all at the dismay of workers who bought into the “streetcar suburb” concept. This resulted in regulatory responses, including caps on fares and demands for expanded or minimum service levels that could not realistically be fulfilled. By 1912, investor interest in streetcars was effectively

over and the subsidiary operators were left to fend for themselves—which proved infeasible (Jones, 1985). These effects were compounded by labor law changes from the 1918 National War Labor Board (NWLB) that brought about the 8-to-5 workday, thereby creating a concentrated peak period of travel that capital and labor resources had to be built to serve, but which mostly went unutilized during off-peak hours. Eventually, the only way these services could be provided to the levels demanded (by regulation) was through public ownership (Jones, 1985).

Almost simultaneous with railroads' finances turning negative, mass production of the automobile was introduced with the Ford Model T in 1913 (Adler, 1991; Wachs, 1984). Mass production of the automobile and the introduction of credit financing meant that the marginal costs of production and time to produce vehicles went down, while mass ownership became financially feasible. This compromised the monopoly that railroads had for providing long-distance travel. If that was not enough, the luxury of traveling in one's own space was alluring, so many countries constructed roads to support car ownership on the basis of that being a public interest. This especially compounded the challenge transit faced of having just a peak period of travel; what is now called off-peak travel become increasingly done by car instead of transit (Jones, 1985). Today, the automobile is not only the dominant form of travel in many western countries, but much travel depends on the automobile since transit cannot effectively serve widely dispersed trips (e.g., Pushkarev and Zupan, 1977).

Yet, public transit provides a lifeline of travel for those who are financially, physically, or otherwise impaired from accessing a car (Sanchez, 1999), as well as a policy response to increased awareness of the social and environmental impacts of dispersed travel done via the automobile (e.g., Ewing et al., 2003; Johnson, 2001; Newman and Kenworthy, 1988). Today, investment in public transit is often lauded as a way to mitigate these impacts. It, for example, provides capacity that can support density at levels that the automobile cannot (see, e.g., Table 3).

Modes of Transit and Mode Shares

Transit Mode Shares

Transit mode share is usually measured either as the share of all trips taken by transit (bus and/or rail) or the share of commute trips taken by transit (an exception is in the European Union, which tracks mode share statistics in some instances as a share of total distance by mode. See, e.g., The European Commission European Commission, 2009, *Transport in the European Union Current Trends and Issues*, March 2019, pp. 25–162. Because shares as a fraction of trips and distances are not comparable, our attention here is restricted to mode shares as fractions of trips). As can be inferred from the above historic overview, transit's mode share in the United States peaked in the early 1900s and quickly declined thereafter to become negligible today. Fig. 1 shows travel and commute mode shares in the United Kingdom and United States, respectively, from 1960 to 2010. Clearly, travel via car has increased as travel via transit has decreased. This trend represents the experience in many countries and metropolitan areas of western societies, though there are exceptions.

A focus on national mode shares for all trips or commute trips will obscure the important role of transit in larger cities. Table 1 shows transit mode share, as a percent of commute trips only, for the top ten transit commuting metropolitan areas (by mode share) in the United States.

Transit mode share varies substantially across world cities. Table 2 shows transit mode share, as a share of all trips, in metropolitan areas in selected cities. Note that major Asian and European cities in some cases move over half of all trips by transit (Seoul, Paris, Tokyo), and the share of total trips in the most transit-oriented metropolitan area in the United States, New York City, at 23%, is somewhat below the level in major Asian or European cities.

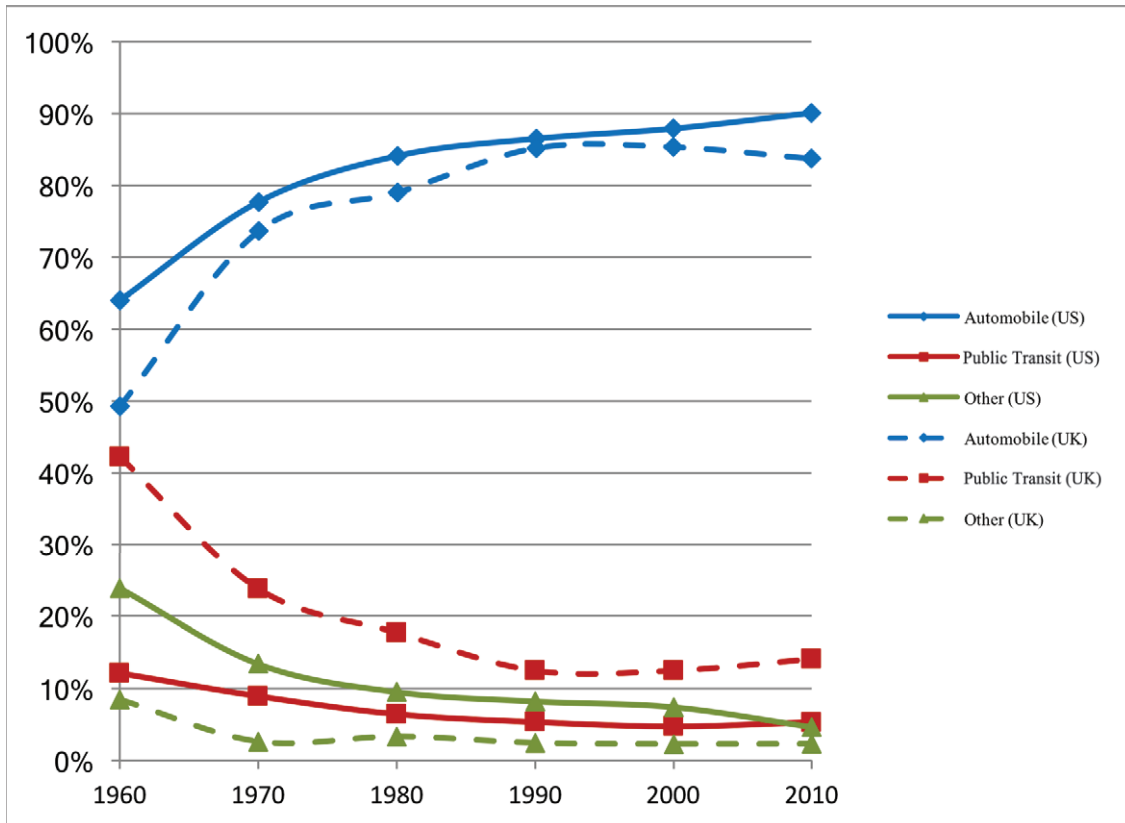
Capacity Across Transit Modes

Passenger capacity and throughput (passengers per hour) varies across different modes. Consider the general analysis from Vukan Vuchic's "Urban Transit: Systems and Technology," (2007). Table 3 reproduces information from Table 2.4 of Vuchic's book (2007, p. 76).

The peak person-per-hour capacity for a freeway lane, from 1800 to 2600 persons, agrees well with traditional highway capacity standards of 2200 vehicles/h and an average vehicle occupancy of 1.2 (for U.S. FHWA highway capacity guidelines, see p. 4 of https://www.fhwa.dot.gov/policyinformation/pubs/pl18003/hpms_cap.pdf). For data on solo and carpool commuters in Los Angeles, Schweitzer (2017) cites 2011 U.S. Census (American Community Survey) data that shows in Los Angeles 74% of workers commute to work in single occupant vehicles and 11% commute in carpool vehicles, consistent with a 1.2 vehicle occupancy rate if all carpools are two-person). The bus rapid transit (BRT) and light rail transit (LRT) frequencies from Vuchic (2007) assume frequencies that have been realized in, for example, Latin America but that go beyond what is common in many cities in the United States. The assumed BRT frequency in Vuchic is 60–300 vehicles/h—feasible, with dedicated lanes and multiple lanes of BRT. The LRT frequency is 40–60 per hour—indicative of some systems outside the United States but uncommon within the United States. The HRT frequency, 20–40 trains/h, is consistent with the most frequent service internationally, but is higher than the current peak HRT frequencies of many United States systems.

Contemporary Public Transit Governance

In the United States, the Constitution defers land use and transportation decisions to states, and states have discretion in determining how much control to delegate to municipalities. States are classified as either *local-control* or *home rule states*, in which local jurisdictions are assumed to have freedom from state government in the planning of land use and transportation, or *Dillon's Rule*



Sources:

1990 Census of Population, STF3C. United States Census.

1980 Census of Population, General Social and Economic Characteristics, United States Summary. United States Census.

1970 Census of Population, Characteristics of the Population, United States Summary. United States Census.

1960 Census of Population, Characteristics of the Population, United States Summary. United States Census.

2010 American Community Survey, Means of Transportation to Work. United States Census.

Passenger Transport by Mode from 1952 (TSGB0101). United Kingdom Department of Transport.

Notes: Travel via taxi is included in the public transit category in the United States; automobile category in the United Kingdom. “Other” includes all other recorded modes of travel, such as air, walking, and pedalcycles.

Figure 1 United Kingdom travel mode shares and United States commute mode shares (1960–2010).

Table 1 Top ten transit commuting metropolitan areas^a

Consolidated metropolitan area statistical area (CMSA)	% Commute trips by public transportation
New York-Northern New Jersey-Long Island, NY-NJ-PA	30.50%
San Francisco-Oakland-Fremont, CA	14.60%
Washington-Arlington-Alexandria, DC-VA-MD-WV	14.10%
Boston-Cambridge-Quincy, MA-NH	12.20%
Chicago-Naperville-Joliet, IL-IN-WI	11.50%
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	9.30%
Seattle-Tacoma-Bellevue, WA	8.70%
Baltimore-Towson, MD	6.20%
Los Angeles-Long Beach-Santa Ana, CA	6.20%
Portland-Vancouver-Beaverton, OR-WA	6.10%

Note: The top ten transit commuting metropolitan areas are from among the 50 largest metropolitan areas. It is unlikely that smaller metropolitan areas would have higher transit commute mode splits.

^aAmong the largest 50 metropolitan statistical areas.

Source: 2009 American Community Survey as summarized in McKenzie, 2010, Table 2, pp. 5-6, Table 4, p. 10, and Table 3, p. 10, as adapted from Marlon Boarnet, *Land Use, Travel Behavior, and Disaggregate Travel Data*, In: Susan Hanson, Genevieve Giuliano (Eds.), *The Geography of Urban Transportation*, 4th edition, The Guilford Press, New York, Table 7.1, pp. 163–164.

Table 2 Transit mode share in selected world cities, all trips

City/Metropolitan area	Transit mode share, all trips
Seoul	66%
Paris	59%
Tokyo	51%
Singapore	44%
London	27%
Beijing	23%
New York city	23%

Source: Lisa Schweitzer, *Mass Transit*, in: Susan Hanson, Genevieve Giuliano (Eds.), *The Geography of Urban Transportation*, 4th edition, The Guilford Press, New York, Table 8.2, pp. 187–188.

Table 3 Capacity for freeway lane, bus rapid transit, light rail transit, and heavy rail transit, from Vuchic (2007)

	Vehicle occupancy	Vehicles per hour	Person capacity per hour ^a
Freeway lane	1.2–2.0	1500–2000	1800–2600
Bus rapid transit	40–150	60–300 ^b	4000–8000 (low end) to 20,000 (with multiple BRT lanes)
Light rail transit	100–750 ^c	40–60 ^d	6000–20,000
Heavy rail transit	140–2400 ^e	20–40	10,000–70,000

^aPer Vuchic (2007), low and high extremes are not products of all low and high values. Vuchic argues the extremes rarely coincide. Values here are from Vuchic (2007).

^bBRT frequency high end assumes multiple parallel lanes and overtaking at stations.

^cLRT assumes 1–4 cars.

^dLRT frequency is much higher than common design capacity in U.S. systems. As an example, in Los Angeles light rail, for the Expo and Blue LRT lines, 10 trains per hour is close to design capacity.

^eHRT assumes 1–10 cars.

states, in which the state is assumed to have authority in planning land use and transportation unless explicitly delegated to a locality by law (Briffault, 1991; Russell and Bostrom, 2016). Depending on the type of state, transit boards are generally established as either an agency of the state or a local agency with corresponding authority. As a result, some transit agencies tend to have an inherent regional or state orientation toward their planning and management, while others are more locally focused.

An additional consideration is the board or governance structure of a transit agency. Many transit agencies are departments of a *general-purpose government*, like a city, county, or state (e.g., Leland and Smirnova, 2008). These are sometimes referred to as a *government department form* of government enterprise. In other instances, transit agencies are an interjurisdictional *special-purpose district*—that is, an independent agency created to fulfill a specific purpose (e.g., Leland and Smirnova, 2008). Special-purpose transit districts are usually governed (and financed) by a *joint-powers authority* (JPA), which is a shared-governance arrangement among the local jurisdictions—called member agencies—that receive service. Both types of transit agencies are traditionally governed either by all or some proportional subset of the area's elected leaders. However, transit agencies can also be overseen by a citizen appointed board, in which nonelected persons are chosen—by either local officials (e.g., Long Beach Transit (“About Us.” *Long Beach Transit*, ridelbt.com/about-us/)) or state officials (e.g., Massachusetts Bay Transportation Authority (“Leadership | MBTA” *Massachusetts Bay Transportation Authority*, <https://www.mbta.com/leadership>))—to oversee the transit agency. Less common are special-purpose transit districts that are governed by a directly elected board; to our knowledge, only three such transit agencies—the Alameda-Contra Costa Transit District (AC Transit), San Francisco Bay Area Rapid Transit District (BART), and Denver Regional Transportation District (RTD)—exist in the United States. Special purpose districts governed by specially appointed or directly elected officials are sometimes referred to as *public corporations*.

Conceptually, these different board structures subject transit agencies to different levels of politicization and constituent responsiveness (e.g., Gerber and Gibson, 2009). On one extreme, while special-purpose elected boards are more directly accountable to the public they serve, by their members being locally elected for a narrowly defined purpose, they are also more susceptible to being especially focused on parochial and special interests in transit decision-making—sometimes at the expense of at-large considerations. For JPAs and general-purpose government agencies overseen by elected officials, while board members are directly accountable to the public through the election process, they are responsible and elected for overseeing many other government functions, making their attentiveness to transit agency oversight possibly less focused and less informed compared to directly elected special-purpose boards (e.g., Hamilton et al., 1983). When they do attend to transit oversight, funding and service delivery decisions tend to be geopolitically driven (e.g., Gerber and Gibson, 2009; Wachs, 1985). In consideration of this, many suggest that a public corporation form of governance with appointed representatives is ideal because appointees can be attentive to, and well-informed

about, transit decision-making, while being several degrees removed from politics (e.g., [Hamilton et al., 1983](#)). On the other hand, by being several degrees removed from public oversight, appointed boards may be less responsive to and representative of the public (e.g., [Mitchell, 1997](#)).

There are many derivatives from these different combinations of transit agency governance formats. Among the most evident is the number of transit agencies an urban area has and their approaches to financing projects. In the State of California, for example, the nine-county San Francisco Bay Area region has 32 different transit operators (“Transit Agencies.” [511 SF BAY](#), Metropolitan Transportation Commission, [511.org/transit/agencies](#)). This creates fragmentation for a transit user seeking seamless travel, but is a product of the local control nature of the state; transit provision is locally controlled and, unless localities voluntarily form a JPA or the state legislatively creates a special-purpose district, localities fend for themselves. By comparison, in the Dillon’s Law State of New York, differently branded transit services in the New York City area—including the Long Island Railroad, Metro North Railroad, and New York City Transit—are all governed under a unified agency, the Metropolitan Transportation Authority (MTA). In addition, as of 2018, 24 counties that account for 88% of California’s population rely on local option sales taxes (LOSTs) to fund transportation, and transit receives a disproportionate share of these funds relative to its mode share ([Lederman et al., 2018](#)). However, transit funds from LOSTs are disproportionately used for capital projects, the distribution of funds are largely influenced by geographic equity interests rather than demonstrated need, and voters’ support for LOSTs rely on substantively invalid arguments that are dissociated with their potential to use transit ([Crabbe et al., 2005](#); [Lederman et al., 2018](#); [Manville and Cummins, 2015](#); [Manville and Levine, 2018](#)). These realities surrounding LOSTs can largely be attributed to the authority and geopolitical influence that different transit board structures have in a local control state. New York’s MTA, on the other hand, is governed by a board of appointees, and the MTA is not permitted by law to pursue LOST measures to finance operations or capital projects. As a result, the MTA cannot so easily turn to system expansion as a solution to demand and must find alternative ways to address such challenges.

Transit Management

Transit Operations

Transit operations emphasize either coverage or performance. Given that a principal function of public transit is to provide a transportation lifeline to those who cannot afford an automobile, many transit providers operate using the *coverage-based model*, in which routes and frequencies are determined in order to achieve maximum coverage area. Under a strict coverage-based model, transit routes tend to be circuitous and route frequencies are standardized across all routes. This contrasts with the *performance-based model*, in which routes and frequencies are defined to concentrate resources where there is demand and create a grid-based network. These models are mainly used to differentiate types of bus operations since both routes and frequencies can be easily changed in a bus environment. For rail, once the network is built, the only adjustable aspects of service operations are generally vehicle size and service frequency. Thus, from an operation’s standpoint, rail is strictly a performance-based operation.

Among the most common performance measurements in the industry are *ridership*, which is the total number of riders; *passenger-miles*, which is the sum of the miles that each individual rider travels; and *passenger-hours*, which is the sum of the hours that each individual rider travels (see e.g., [Schweitzer, 2017](#)). These measurements are often then compared to the *vehicle-miles*, *vehicle-hours*, *seat-miles*, *seat-hours*, and *labor-hours* of the transit providers to assess per unit efficiencies.

Some constraints that impact transit operations are running times, labor work rules, and system capacity. *Capacity* refers to the throughput of a system, can vary by location and time, and includes both vehicle and passenger components (i.e., *vehicle capacity* and *passenger capacity*). To identify how to allocate rolling stock and labor resources, transit operators use *load factors*, which are a measurement of how full a transit vehicle is relative to the number of seats on the vehicle ([Schweitzer, 2017](#)). A load factor of 1 means that the number of passengers on the vehicle is equal to the number of seats on the vehicle; less than 1, that there are more seats than passengers; more than 1, that there are more passengers than seats.

Different transit operators have different ways of measuring their load factors. Most bus operators use automated people counters on bus doors to derive load factors at every stop. Some rail operators assume passengers make round trips and use a common fare payment media, so use the time and station where they enter the system on two different occasions to associate them with trains, times, and locations. More sophisticated rail systems require passengers to tag in and out of the system and use this information to assign them to specific trains to determine the trains’ load factors.

When multiple routes in a system experience load factors that exceed capacity, transit operators must decide how to allocate scarce resources, such as buses and rail cars. The objective is to either mitigate or equalize the problem across all routes. Most often, this is done by calibrating system demand to identify a standard load factor that is then used to allocate vehicles. For example, if the standardized load factor is determined to be 1.2, this may result in one route having a frequency of five buses per hour and another having a frequency of twelve buses per hour, but these frequencies help ensure that both routes experience similar crush loads, given capacity and resource constraints.

A unique capacity constraint of rail is the track network. There are many rail systems where two or more rail lines converge into a single rail line, which creates a capacity constraint for all train routes that use the shared segment. While, for example, the shared rail segment may be able to serve a train every two minutes, if five branch lines converge onto the shared line, they are each limited to an average throughput of one train per ten minutes. The most common way this is mitigated is by building multiple sets of track in the respective area(s), which requires extensive real estate.

Negotiated labor contracts and route run times have overlapping impacts on operations. The *run time* of a route is the amount of time it takes to travel from one end to the other. Most labor contracts have multiple provisions that limit the amount of work that frontline workers can do without a break or lunch. At the same time, transit operations managers need to ensure that routes run on a standard frequency and also have an interest in defining routes that are optimal for transit users. It is not uncommon for work rules, route runtimes, and route frequencies to not align, resulting either in workers receiving extended breaks or optimal routes not being implemented. A common method of overcoming this dilemma is through *interlining* routes, wherein multiple routes terminate at a common terminus, and a vehicle and/or its operator arrive under one route but depart as a different route. This allows a transit provider to maintain a standardized frequency on the different routes without having to provide longer than reasonable break periods—however long “reasonable” is. However, a drawback of this practice, particularly with bus operations, is that it can divide routes up at the expense of seamless travel for riders.

In addition, the transit sector is well-documented as a sector where frontline worker earnings have outpaced other industries of similarly skilled workers. In the United States, although federal and state transit operating subsidies have tended to be justified on the basis of maintaining or increasing service quality to compete against the automobile, operating costs have increased while both labor productivity and service quality have decreased—suggesting subsidies have not gone to maintain or improve service, but to increase labor earnings. Wachs (1989), referencing Pickrell (1986), shows that, between 1965 and 1983, after controlling for inflation, transit operating costs per vehicle-mile increased 80%. This increase paralleled the transition of transit operations from private to public; whereas the share of transit vehicle-miles of service that were publicly operated in 1965 was 48%, it was 96% in 1985 (Wachs, 1989). More recently, Morales Sarriera and Salvucci (2016) similarly found that transit workers in the United States received pay and benefit raises that greatly exceeded the rate of earnings growth in similar sectors, even while productivity has remained relatively stagnant. Some (e.g., Jones, 1985) suggest that this is caused by outdated labor management practices that lead to few promotional opportunities, leaving only wage increases as a viable way to increase workers’ annual earnings.

Finally, transit operators must consider fares. In the United States, effectively no transit operator sets fares that cover even half the cost of providing service (Schweitzer, 2017). While private transit companies used differential fares to price different lengths and travel times based on demand and costs, public transit services, particularly in the United States, generally do not differentiate fares. The migration to flat, undifferentiated fare structures increased sharply in the 1970s and many studies have shown that this facilitates cross-subsidies from short-distance riders to long-distance riders, off-peak riders to on-peak riders, and lower income populations to higher income populations (e.g., Brown, 2018; Cervero, 1981; Oram, 1979). These practices and outcomes are particularly true of bus operators. One reason for this is that flat fares are simpler to implement and manage. Most transit operators that have differential fares operate rail and either have computerized systems to account for the fare differentials and charge accordingly, or on-board conductors who validate fares. The implications these practices have on financing are discussed later.

Capital Project Planning

Another element of transit management is the planning and implementation of capital projects. A *capital project* is any project that involves the building of new infrastructure, including the replacement or upgrade of existing infrastructure.

Capital project benefits are generally classified into one of two categories: transportation effects and expenditure effects (Taylor, 2017). A *transportation effect* is usually long-term and pertains to the benefits the infrastructure project will have on travel and its residuals. Congestion relief, reduction in air pollution, faster commute times, or a more pleasant travel experience are all transportation effects of transit capital projects. *Expenditure effects*, on the other hand, are generally short-term and pertain to economic or other localized gains, not from the infrastructure itself but from the expenditure of money to build the project. Common examples of expenditure effects include the creation of local jobs for construction and, relatedly, increased local economic spending resulting from the increase in workers, and income in the project area. Both benefits are leaned on to politically advance capital projects, though only transportation effects are an economically effective reason to pursue them because they are the only benefits that are directly germane to the investment (Taylor, 2017).

Even so, transit capital projects are often oversold on their benefits, underdelivered in time, and underestimated in their costs (Flyvbjerg et al., 2003). A review of the 1970s’ resurgence of rail in the United States offers a good illustration of this and why this occurs (Pickrell, 1992). Eight rail projects of the era—in Atlanta, Buffalo, Miami, Pittsburgh, Portland, Sacramento, and Washington, DC—averaged being eight years behind schedule, which contributed to each project’s capital costs being much higher than projected due to inflation. In addition, operating expenses of projects averaged 38% higher than projected, and ridership modeling was positively biased during planning phases. These deficiencies in the performance of system expansions can generate or compound *core system impacts*, which are any adverse impacts posed upon pre-existing services that is caused by an expansion project. Among other examples, trains and buses may experience crowding earlier along the route than before, resources may be redistributed away from existing services to accommodate the new service, and added wear and maintenance demands are placed onto the system.

Transit Finance

Transit budgets generally fall into two categories: an operating budget and a capital budget. A transit agency’s *operating budget* covers day-to-day, recurring costs of operations, such as the labor costs of mechanics, bus and rail operators, and administrative staff. *Capital budgets*, on the other hand, cover the short-term costs of technical contract staff and construction labor associated with the development and implementation of capital projects.

Table 4 Comparative Farebox recovery ratios

<i>Transit agency</i>	<i>Country</i>	<i>Type of service</i>	<i>Fare structure</i>	<i>Farebox recovery ratio</i>
Mass Transit Railway (MTR)	Hong Kong	Heavy rail, light rail, bus	Distance-based	124% ^a
San Francisco Bay Area Rapid Transit (BART)	United States	Heavy rail	Distance-based	70% ^b
Transport for London	United Kingdom	Heavy rail, bus	Zonal	72% ^c
Calgary Transit	Canada	Light rail, bus	Flat	45% ^d
Metropolitan Atlanta Rapid Transit Authority (MARTA)	United States	Heavy rail	Flat	33% ^b
Los Angeles County Metropolitan Transportation Authority (LA Metro)	United States	Light rail, bus	Flat	19% ^b
Santa Clara Valley Transportation Authority (VTA)	United States	Light rail, bus	Flat	10% ^b

^a"2016 Annual Report – Notes to the Consolidated Accounts" (PDF). MTR Corporation. Retrieved 31 Jan 2020.

^bNational Transit Database 2018 Transit Agency Profiles.

^cTfL Budget 2019/20 – Transport for London" (PDF). Transport for London. Retrieved 31 Jan 2020.

^dAttach 3 – Calgary Transit Fare and Revenue Framework – TT2018-0617" (DOCX). City of Calgary. Retrieved 31 Jan 2020.

Because transit services cannot be self-sufficient in most areas of the globe today, their operations are inevitably financed through sources other than fares. Transit operators use *farebox recovery ratios*—the quotient of the revenue generated from fares divided by the operating costs of the service—to measure their cost efficiency. A ratio of 1 (100%) means that the operating costs are equal to the agency's farebox revenue; less than 1, that the agency's operating costs exceed the revenue generated through fares; greater than 1, that the agency is generating more through fares than it is expending on operations. **Table 4** illustrates that the vast majority of transit providers across the globe operate at a loss when only farebox revenues are considered. Asian-based rail providers stand out as the main exception. While some of the differences between operating costs and farebox revenues are made up through nonpassenger sources of revenue, such as the rent or sale of advertisement space or real property, these sources are historically not enough to balance the budget. The differences between operating costs and operating revenues are most often funded by local, state, and federal taxes.

Unlike operating budgets whose sources and costs are recurrent and predictable, and whose costs can be scaled to match total funding, capital costs are none of these. When a transit agency decides to invest in a capital project, it must both determine the costs of the project and identify one-time sources to finance the projects. Very often, capital projects experience cost overruns and delays (Flyvbjerg et al., 2003; Pickrell, 1992). When this occurs, transit agencies must identify additional one-time sources of revenue to bridge the gap in capital financing. One-time sources of capital funds usually come in the form of competitive grants from state and federal governments, as well as through local taxes, like sales tax or bonds financed through property taxes. Oftentimes, the sunk cost of how much has already been invested in a project is used to bargain for additional funds from financiers. Dubbing a *commitment trap*, rather than ending the construction of a halfway built project, the tendency is to give the money necessary to complete the project (Schweitzer, 2017). This reality carries on into operations, even if the ridership volume falls well short of what was estimated, thereby requiring additional operating subsidies to cover the costs.

Public Financing

Funding is the primary mechanism through which higher levels of government that defer authority to lower levels of government nevertheless influence transit operations and planning. By setting conditions for which funds will be allotted, higher levels of government are able to impose how transit is operated and planned. In the United States, for example, transit operations are partly influenced by Title VI of the Civil Rights Act of 1964. Under this act, any transit operator who receives federal funding must conduct disparate impact and disproportionate burden studies for any fare or operating change in order to ensure that protected populations are not adversely affected by the changes compared to other groups. A *disparate impact* occurs when a face-neutral change results in a community of color being adversely affected relative to other racial groups, while a *disproportionate burden* occurs when low-income populations are negatively affected relative to other income groups. Central funding agents that receive and allocate federal funds to capital projects, like metropolitan planning organizations, must also demonstrate that their method of funding capital projects does not disproportionately short change low-income or racial minority populations. There are also many other federal rules and acts that transit operators must abide by as a condition of their receipt of federal funds, such as how they provide paratransit service in accordance with the Americans with Disabilities Act (ADA). While operating funds come with many conditions attached to them and their distribution is subject to political discourse, they are mostly dispersed based on defined and persistent formulae.

Because capital project costs generally far exceed what can be financed through local resources, transit agencies rely on state and federal funding to complete local capital projects (Taylor, 2017). While this is also true for operations, finance for operations is able to be established by a recurring formula, so is not as frequently subject to political discord. By comparison, each time that a new capital project seeks financing, it is subject to political debate about how to divide up a limited amount of funds across multiple projects. As a result, rather than capital project financing being allocated based on some measurement of need—for example, where

the transit investment would generate the most use—they are allocated based on political bargaining. The often-parochial focus of transit financing exacerbates this. As an example, even though the New York MTA's ridership accounts for over a third of transit ridership in the United States and this magnitude in share has not changed for decades, transit capital investments over the past several decades have focused on cities with markedly less transit ridership.

To be sure, there are federal, state, and regional financiers of capital projects whom have competitive review procedures or build conditions into their financing programs—for example, that the “requesting agency” demonstrates strong transportation effects of their project using approved methods. However, general fund allocations are not subject to this, are most often a part of the financing of any mega capital project, and are much more geopolitically determined. In addition, the methods that requesting agencies are expected to use to legitimate a project are able to be undermined using biased inputs that generate positive results (Flyvbjerg et al., 2003; Pickrell, 1992; Wachs, 1990). Once funds are allocated, whether competitively or politically, it creates the aforementioned commitment trap for higher levels of government. In many respects, transit agencies in local-control states, when they have passed a local financing measure for a capital project, have an advantage because they have already committed so much to the project themselves and can use this in their bargaining.

Thus, while higher levels of government influence projects and operations through financial conditions, they tend to also bear the risk of any capital or operating cost overruns associated with a project. Some suggest that, if this risk was placed on the local agency advancing the project instead, it would improve the quality of projects that seek funding from higher levels of government (Flyvbjerg et al., 2003; Pickrell, 1992). In other words, the commitment trap enables poor capital project planning.

Other Considerations

While transit in the United States is financed almost exclusively through a combination of farebox revenues and local or state taxes (for operations) and a combination of federal, state, and local taxes (for capital construction), other models are possible. Most notable is the use of land value capture in some Asian countries.

Hong Kong is possibly the best example of land value capture in transit. The Hong Kong Mass Transit Railway (MTR) was formed in 1975 and is the mass transit rail authority for the Hong Kong special administrative region (SAR). Due to its British colonial heritage, the government of the Hong Kong SAR owns all lands within the region. The Hong Kong SAR government is the primary stockholder in the MTR, which is a private company. Beginning in the 1990s, the Hong Kong MTR and the Hong Kong SAR government developed the “rail + property” model of rail transit capital finance. Lands above and near future stations are granted to the MTR by the Hong Kong government, and MTR then plans, constructs, and operates the rail line and engages private developers to build developments near stations. The difference between the pre-rail land value and the after-rail land value accrues to MTR via development agreements, and is used to finance MTR rail line construction (Suzuki et al., 2015).

This approach is commonly called land value capture. Rail transit, by producing a land value premium (through improved access), increases land value near stations. Using that increased land value to finance rail transit was not uncommon in the United States before World War II (Bernick and Cervero, 1997) but is rare in the United States now. Land value capture is more common in Asia, and the method described in Hong Kong is similar to land value capture methods used in, for example, Tokyo (Bernick and Cervero, 1997).

The Future

Transit is in a unique predicament today that makes its future uncertain. The rise of ridehail services, like Lyft, Uber and Didi Chuxing (China), and dockless shared bicycle and scooter services, like Bird and Lime, have shaken the transportation industry. These services have come to be known collectively as *disruptive transportation technologies*. It is clear that overall transit ridership has fallen in parallel with the rise of disruptive transportation technologies. After experiencing ridership and mode share growth for several years following the 2008 recession, both measurements of transit utilization have been on the decline since 2015 in the United States (Graehler et al., 2019; Manville et al., 2018; Taylor et al., 2019). These changes coincide with for-hire transportation—inclusive of traditional taxicabs and limousines, in addition to ridehailing—doubling in use since 2009, which is reasonably attributable to the onset of ridehail services (Conway et al., 2018).

A lot of current research is seeking to explain this decrease in transit ridership with a particular focus on disruptive transportation technologies. While there is a clear correlation between the introduction of disruptive transportation technologies and the fall in transit ridership, the causal relationship, if any, is unknown, as of yet. For example, whereas Hall et al. (2018) suggest that ridehail services complement rather than substitute for transit trips, which would suggest that the decline in transit ridership is explained by other factors; Doppelt (2018) and Graehler et al. (2019) suggest that ridehail services are a substitute for transit, but that their substitution effect is a function of time since the services' penetration into the given city. Other research suggests that there are mixed results, including that ridehail services are initially complementary of transit but become substitutive as they penetrate the local market (Nelson and Sadowsky, 2017), and that the services are substitutive of local transit services but complementary of heavy and commuter rail services (Clewlow and Mishra, 2017; Grahm et al., 2019; Graehler et al., 2019). Dockless mobility services are still too new for much to be known about their impacts.

At the same time, it is well documented that none of the disruptive transportation technology companies have made a profit since their inception. In fact, they have all relied on heavy venture capital investments to subsidize their services. Thus, all of the aforementioned research findings on the use of disruptive transportation technologies, as well as any associated effect it has on transit use, are artificially inflated; if users paid the full costs of the services, their use and effect would be less.

If ridehail services find a path to becoming financially solvent, the market of use will likely look much different than it does today, and it will have a different effect on transit than the latest estimates suggest. Under one scenario, these services are an impetus to a future of autonomous, all electric, shared mobility, where the costs of paying a driver are removed from the equation and the flexible service model of transit becomes more broadly adopted (e.g., [Sperling, 2018](#)). In fact, a number of transit service providers have responded by converting fixed route services into flex services. Under this concept, the unsustainable costs of ridehailing services are merely a short-term reality until autonomous and connected vehicles are fully developed. To date, however, this “three revolution” concept of a transportation future has not developed as expeditiously as promised, and many practical constraints to the vision remain, including the many engineering constraints to automating dispersed trips in multiple vehicles across multiple traffic intersections. Instead, some suggest that fixed route transit service—principally by their being fixed routes—are the best candidates of any “three revolution” future ([Currie, 2018](#)).

On the other hand, if disruptive transportation technologies do not find a financially sustainable model to exist, they will no longer exist as private companies. Theoretically, this could return public transit service and ridership back to its 2015 levels. Alternatively, if the public accepts disruptive transportation technologies as a public interest, much like how unsustainably financed private railroads and suburbanization became accepted as a public interest to finance in the early 1900s, ridehail services could conceivably be incorporated into public transit financing and service provision.

Summary and Conclusions

In this chapter, we have summarized the evolution of transit, surveyed ways that it is governed and operated, and some of the contemporary and foreseeable challenges the industry faces.

In the United States, early forms of transit, including the horse-drawn carriage and early streetcars, were privately financed and operated. However, the unsustainable financial structure of private transit providers eventually caught up to the industry. With a sizable share of the population building their life-work locations and travel around dependable transit, the financial fallout of the industry led to an increase in public oversight. Production of the car and adoption of automobile travel by the wealthy compounded the financial insolvency of the industry. In time, mass production of the automobile, the invention of credit-based financing for automotive purchases, and extensive construction of roadways made the automobile more accessible and desirable to the broader population. Today, in much of the developed world, transit mainly operates to provide a lifeline mode of travel to those who are unable to use a car or as a policy instrument for addressing harms attributable to automobile reliance.

These contemporary goals for providing transit service inform how transit is financed and operated. For example, many transit operators seek to maximize coverage area rather than service performance. In addition, in most areas of the developed world, transit’s mode share is less than 10% and fares generally cover less than 25% of the costs of providing service. As a result, public transit providers generally have to rely on alternative sources of revenue to finance both the operating and capital costs of transit services. How locally or regionally focused a transit agency is, the amount of independent authority it has, and the degree to which its board structure exposes it to geopolitical factors, can greatly influence how some of these decisions are made at a microlevel.

Among the most live challenges facing the transit industry at present is the rise of disruptive transportation technologies and the impact these have had on transit ridership, mode share, and revenues. Ridehail services, in particular, have made car-quality access without the car ownership cost, widely available. The degree and direction of impact these services have had on transit is unclear, however, primarily due to data not being available for analysis. Some studies suggest that ridehail services complement transit; some suggest that they compete or substitute for transit; and some suggest that the services complement rail transit but substitute for local transit. Notwithstanding these findings, all indications, including ridehail companies’ public financial records, suggest that disruptive transportation technology companies have yet to generate a profit, and their path to making a profit is uncertain. In other words, like private transit service of the 1890s and 1900s, disruptive transportation technology companies are so far financially unsustainable as private entities. So far, research has only illuminated the short-term impacts of these services. We are at an early stage of transportation disruption, and as in the previous era of the 1890s and 1900s, it will take time for the impact on public transit to become clear.

Biographies

Zakhary Mallett is a Ph.D. Student in Urban Planning and Development at the Sol Price School of Public Policy of the University of Southern California. He previously served as an elected member of the governing board of the San Francisco Bay Area Rapid Transit District (BART).

Dr. Marlon Boarnet is a Professor of Public Policy at the Sol Price School of Public Policy of the University of Southern California.

See Also

Operation Costs for Public Transport; Public Transport Fare and Subsidy Optimization; Transportation Network Companies (TNCs) and the Future of Public Transportation; Public transport in low density areas; Demand-Responsive Transit, Evaluation

Studies; Public Transit Ridership Forecasting Models; Transport Modes: Data, Methods and Analysis; Modeling Passenger Transport Mode Choice; Transport Modes and Commuters; Big Data for Public Transport Planning; Bus Public Transit Operations; Bus Route Planning; Policy and Governance; Demand Responsive Transport; Land use and transport planning; Politics of Mobility Policy; Transport Megaprojects; Planning for Rail Transport; Public Transport Fares; Public Transport Subsidy and Regulation; Public Transport Network Planning

References

- Adler, Sy., 1991. The transformation of the pacific electric railway: Bradford Snell, Roger Rabbit, and the Politics of Transportation in Los Angeles. *Urban Affairs Quarterly* 27 (1), 51–86.
- Bernick, M., Cervero, R., 1997. *Transit Villages in the 21st Century*. McGraw-Hill, New York.
- Briffault, R., 1991. The role of local control in school finance reform. *Conn. L. Rev.* 24, 773.
- Brown, A.E., 2018. Fair fares? How flat and variable fares affect transit equity in Los Angeles. *Case Stud. Transp. Policy* 6 (4), 765–773.
- Conway, M., Salon, D., King, D., 2018. Trends in taxi use and the advent of ridehailing 1995–2017: evidence from the US National Household Travel Survey. *Urban Sci.* 2 (3), 79.
- Cervero, R., 1981. Flat versus differentiated transit pricing: what's a fair fare? *Transportation* 10 (3), 211–232.
- Clewlow, R.R., Mishra, G.S., Disruptive transportation: The adoption, utilization, and impacts of ride-hailing in the United States. (2017).
- Crabbe Amber, E., et al., 2005. Local transportation sales taxes: California's experiment in transportation finance. *Public Budget. Fin.* 25 (3), 91–121.
- Currie, G., 2018. Lies, damned lies, AVs, shared mobility, and urban transit futures. *J. Public Transport.* 21 (1), 3.
- Doppelt, L. Need A Ride? Uber Can Take You (Away From Public Transportation). Unpublished Dissertation. Georgetown University, 2018.
- European Commission, 2019. Transport in the European Union Current Trends and Issues. March, 25–162, available at <https://ec.europa.eu/transport/sites/transport/files/2019-transport-in-the-eu-current-trends-and-issues.pdf> (accessed 16.01.20).
- Ewing, R., Rolf, P., Don, C., 2003. Measuring sprawl and its transportation impacts. *Transport. Res. Rec.* 1831.1, 175–183.
- Flyvbjerg, B., Nils, B., Werner, R., 2003. *Megaprojects and Risk: An Anatomy of Ambition*. Cambridge University Press.
- Gerber, E.R., Gibson, C.C., 2009. Balancing regionalism and localism: How institutions and incentives shape American transportation policy. *Am. J. Polit. Sci.* 53 (3), 633–648.
- Graehler, M., Mucci, A., Erhardt, G.D., 2019. Understanding the Recent Transit Ridership Decline in Major US Cities: Service Cuts or Emerging Modes?. *Transportation Research Board 98th Annual Meeting*, Washington, DC, January.
- Grahn, R., et al., 2019. Socioeconomic and usage characteristics of transportation network company (TNC) riders. *Transportation* 1–21.
- Hall, J.D., Palsson, C., Price, J., 2018. Is Uber a substitute or complement for public transit? *J. Urban Econ.* 108, 36–50.
- Hamilton, N., Feldhusen, B., Crisp, H., 1983. A comparison of governance of publicly- owned mass transit. *Can.-USLJ* 6, 1.
- Johnson, M.P., 2001. Environmental impacts of urban sprawl: a survey of the literature and proposed research agenda. *Environ. Plan. A* 33 (4), 717–735.
- Jones, D.W., Jr., 1985. Urban transit policy: An economic and political history.
- Lederman, J., et al., 2018. Lessons learned from 40 years of local option transportation sales taxes in California. *Transport. Res. Rec.* 2672 (4), 13–22.
- Leland, S., Smirnova, O., 2008. Does government structure matter? A comparative analysis of urban bus transit efficiency. *J. Public Transport.* 11 (1), 4.
- Manville, M., Cummins, B., 2015. Why do voters support public transportation? Public choices and private behavior. *Transportation* 42 (2), 303–332.
- Manville, M., Taylor, B.D., Blumenberg, E., 2018. Falling Transit Ridership: California and Southern California.
- Manville, M., Levine, A.S., 2018. What motivates public support for public transit? *Transport. Res. Part A: Policy Pract.* 118, 567–580.
- Mitchell, J., 1997. Representation in government boards and commissions. *Public Admin. Rev.* 57 (2), 160–167.
- Morales Sarriera, J., Salvucci, F.P., 2016. Rising costs of transit and Baumol's cost disease. *Transp. Res. Rec.* 2541 (1), 1–9.
- Nelson, E., Sadowsky, N., 2017. Estimating the Impact of Ride-Hailing App Services on Public Transportation Use in Major US Urban Areas. Bowdoin College. Department of Economics. Brunswick.
- Newman, P.W.G., Kenworthy, J.R., 1988. The transport energy trade-off: fuel-efficient traffic versus fuel-efficient cities. *Transport. Res. Part A: Gen.* 22 (3), 163–174.
- Oram, R.L., 1979. Peak-period supplements: the contemporary economics of urban bus transport in the UK and USA. *Progr. Plan.* 12, 81–154.
- Pickrell, D.H. 1986. *Federal Operating Subsidies for Urban Transit: Their Origin, Consequences, and Reform*. Transportation Systems Center, U.S. Department of Transportation, Cambridge, Mass.
- Pickrell, D.H., 1992. A desire named streetcar fantasy and fact in rail transit planning. *J. Am. Plan. Assoc.* 58 (2), 158–176.
- Pushkarev, B., Zupan, J.M., 1977. *Public transportation and land use policy*. Indiana Univ Pr.
- Russell, J.D., Bostrom, A., 2016. Federalism, Dillon rule and home rule. American City County Exchange. White paper: A publication of the American city county exchange. Arlington, VA: author. Retrieved from <https://www.alec.org/app/uploads/2016/01/2016-ACCE-White-Paper-Dillon-House-Rule-Final.pdf>.
- Sanchez, T.W., 1999. The connection between public transit and employment: the cases of Portland and Atlanta. *J. Am. Plan. Assoc.* 65 (3), 284–296.
- Schweitzer, L., 2017. *Mass Transit. Geography of Urban Transportation* 184–217.
- Sperling, D., 2018. *Three Revolutions: Steering Automated, Shared, and Electric Vehicles to a Better Future*. Island Press.
- Suzuki, H., Murakami, J., Hong, Y.-H., Tamayose, B., 2015. *Financing Transit-Oriented Development with Land Values*. The World Bank, Washington, D.C.
- Taylor, B.D., 2017. The geography of urban transportation finance. *Geogr. Urban Transport.* 247–272.
- Taylor, B.D., Manville, M., Blumenberg, E., 2019. Why is public transit use declining? Evidence from California. No. 19-03394.
- Vuchic, V., 2007. *Urban Transit: Systems and Technology*. Wiley, Hoboken, NJ.
- Wachs, M., 1984. Autos, transit, and the sprawl of Los Angeles: The 1920s. *J. Am. Plan. Assoc.* 50 (3), 297–310.
- Wachs, M., 1985. Management vs. political perspectives on transit policymaking. *J. Plan. Educ. Res.* 4 (3), 139–147.
- Wachs, M., 1989. US transit subsidy policy: In need of reform. *Science* 244 (1), 1545–1549.
- Wachs, M., 1990. Ethics and advocacy in forecasting for public policy. *Bus. Prof. Ethics J.* 141–157.
- Warner Jr, Sam Bass, 1968. "Streetcar Suburbs: The Process of Growth in Boston 1870–1900 (Cambridge, MA, 1962)." Sam Bass Warner, Jr., *The Private City: Philadelphia in Three Periods of Its Growth*, Philadelphia, PA.

Further Reading

- Boarnet, Marlon, G., 2017. *The Geography of Urban Transportation*. In: Susan Hanson, Genevieve Giuliano, (Eds.), *Land use, travel behavior, and disaggregate travel data*. 4th edition. The Guilford Press, New York.
- Eidlin, Eric, 2005. The worst of all worlds: Los Angeles, California, and the emerging reality of dense sprawl. *Transport. Res. Rec.* 1902 (1), 1–9.
- Hamilton, N., Feldhusen, B., Crisp, H., 1983. A comparison of governance of publicly- owned mass transit. *Can.-USLJ* 6, 1.

- Hensher, D.A., Button, K.J., 2000. Handbook of transport modelling. No. 1.
- King, D., 2011. Developing densely: estimating the effect of subway growth on New York City land uses. *J. Transport Land Use* 4 (2), 19–32.
- Levy, B., Spiller, P.T., 1996. Regulations, Institutions, and Commitment: Comparative Studies of Telecommunications. Cambridge Univ Pr.
- Sanchez, T.W., Shen, Q., Peng, Z.-R., 2004. Transit mobility, jobs access and low income labour participation in US metropolitan areas. *Urban Stud.* 41 (7), 1313–1331.
- Snell, B.C., 1974. American ground transport: A proposal for restructuring the automobile, truck, bus, and rail industries. US Government Printing Office.
- Taylor, B.D., et al., 2008. Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas. *Transp. Res. A* 43 (1), 60–77.
- Tiebout, C.M., 1956. A pure theory of local expenditures. *J. Polit. Econ.* 64 (5), 416–424.
- Watkins, K., et al., 2019. Analysis of Recent Public Transit Ridership Trends. TCRP Research Report 209.

Road Infrastructure: Planning, Impact and Management

Jos Arts*, Wim Leendertse*, Taede Tillema[†], *Department of Planning, Faculty of Spatial Sciences, University of Groningen, The Netherlands; [†]Department of Economic Geography, Faculty of Spatial Sciences University of Groningen, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction: Characteristics of Road Infrastructure	360
Planning for Road Infrastructure	363
Management of Road Infrastructure	365
Impacts of Road Infrastructure	367
Outlook: Future Trends and Challenges	369
Biographies	370
Relevant Websites	370
References	371
Further Reading	372

Introduction: Characteristics of Road Infrastructure

The word *infrastructure* is a literal combination of the Latin “infra”—what means below—and “structure.” Infrastructure is therefore an underlying structure that supports activities such as transport. Generally, infrastructure refers to the physical structure, however some authors include in their definition also the organizational structure, the institutions, and the management, linked to that hardware. Infrastructure can be defined as “all physical assets, equipment and facilities of interrelated systems and the necessary service providers, together with the underlying structures, organizations, business models and rules and regulations, which are used to offer certain sector specific commodities and services to individual economic entities or the wider public to enable, sustain or enhance social living conditions” (Weber and Alfen, 2011, p.9). Roads are by far the *most dominant form of transport infrastructure*, which is related to the diversity of road-related transport. Motorized transport—such as cars, busses, trucks, motors, mopeds, e-bikes—as well as nonmotorized modes—such as horses, camels, bicycles, pedestrians—all make use of the road network. Also, roads belong with waterways to the *oldest forms of transport infrastructure*. Already 2000 years ago in e.g. China and in the Roman empire huge long-distance road networks were built to connect the various parts of the empire—see Fig. 1. These old major roads are still visible in the current landscape and on maps [10440]. For instance, the old Roman road in Northern Italy, now an autostrada, is connecting the original resting places for changing horses that over time developed into the current major cities. Another example is the old silk road connecting Eastern and Western civilizations already for thousands of years, now revived as a new silk road as part of the ‘One-Belt-One-Road’ initiative. Here, we see a *paradox of inertia and innovation*. Existing road connections prove to *sustain for long periods* and tend to extend over time, creating fixation of socio-economic development along roads (resulting in agglomeration, urban sprawl, scarcity of space, environmental problems). Moreover, once a road is created it is difficult to “erase” it and the developments related to it. However, roads prove to be also strong *axes of innovation* and road transportation is innovating rapidly concerning road technology (smart roads, ITS), transportation (electrification, automation – CAD, 2020) and concerning the services of mobility (MaaS)—(see ERTRAC, 2019; KIM, 2018) [10420, 10422].

Another characteristic of road infrastructure is the very *dense, extensive, and interconnected network* that consists of nested subnetworks at various geographical scales such as international corridors (e.g., TEN-T and E-road networks, Pan-American Highway), national highways (e.g., US Interstate system; see Fig. 2), regional roads, and local streets, and paths up to individual buildings and houses—see also [10443, 10403, 10404]. The term ‘network’ is used in the discourse on infrastructure development as a designation for both the physical network (of e.g. roads) and the social network (of actors involved in planning and management). In this chapter focus is on the first (if referred to the latter this is made explicit). Road infrastructure serves both *long-distance and last-mile transportation* and its network density is unique, which causes car mobility to be by far dominant in the modal split globally (Banister, 2008). Road infrastructure facilitates diverse forms of transportation: *private and public transport*, individual and shared transport, as well as *persons and freight transport*, which all compete for the available traffic capacity of the road network. Because of all of this, road infrastructure is seen as an *important prerequisite for socio-economic development* of places and regions. Here the concept of *land use transport integration (LUTI)* (Bertolini 2012; Kelly, 1994; Wegener and Fürst, 1999—see Fig. 3) is instrumental: roads create better accessibility and connectivity [10402], which is a driver for (new) land uses (housing, working, recreation, facilities), which creates mobility, and the need for more road infrastructure. As a consequence, planning for roads is often closely linked to planning of area development resulting from policies for economic growth [10396]. Therefore, one can often observe that new administrations propose large road infrastructure development programs as they not only stimulate economic growth of regions and countries, but also provide directly extra work and jobs for the construction sector and indirectly stimulate the automotive sector – both important economic sectors in most countries.



Figure 1 Network of main road in the Roman Empire under emperor Hadrianus (Wikipedia, Andrein)

As a consequence, road infrastructure itself and its use (especially by car traffic) create *positive and negative impacts* that affect fundamental interests such as welfare, health, and property, which usually call for involvement of government. As these impacts are distributed unevenly amongst societal groups, issues of equity, and justice become important in road infrastructure planning (Martens, 2017; Banister, 2018; see also [10604]). Moreover, roads and their use create many *externalities*—land take, fragmentation, barrier impacts, noise, air pollution, safety issues etc.—which ask for a careful involvement of “passive users” of roads—the people who live and companies that are located nearby. The functional interrelatedness between road infrastructure and spatial regional development has given rise to more *integrated planning approaches*—so-called context-sensitive, place-based, or area-oriented approaches (Heeres et al., 2012). Such approaches aim to limit negative impacts by an integrated design and to create synergies (“win-wins”) between infrastructure and spatial development.

The road infrastructure network also has a *multilevel characteristic*. Roads connect localities and link the local scale with the regional scale, and at a higher spatial scale, roads link regions with each other thereby involving the (inter)national scale level. The benefits of better connectivity are usually felt by society as a whole at the network level and the higher scale of the cities, regions or countries—and of course at those places that have good access to the network (especially *hubs*, being the intersections of different modes of transportation). Especially the urban-regional scale is important for the socio-economic advantages of road infrastructure, this is the level of the Daily Urban System (DUS – closely related to the term Functional Urban Area, Dijkstra et al. 2019), the area around a city in which the daily commuting occurs to work (but also to education, facilities etc.); a potential outcome of a DUS is urban sprawl. However, the negative effects of the existence of road infrastructure and its usage is mainly with local individuals that live nearby. Road infrastructure can be characterized as a *common good*. To build a highway—let alone to develop a road infrastructure network—is a very expensive and time-consuming endeavor and little market parties can afford this. Also, roads are part of a network which developed after long time frames and which is used by many diverse parties, while its management is done at the network level by a single or few dominant infrastructure provider(s). Moreover, roads and their use create many *externalities*—land

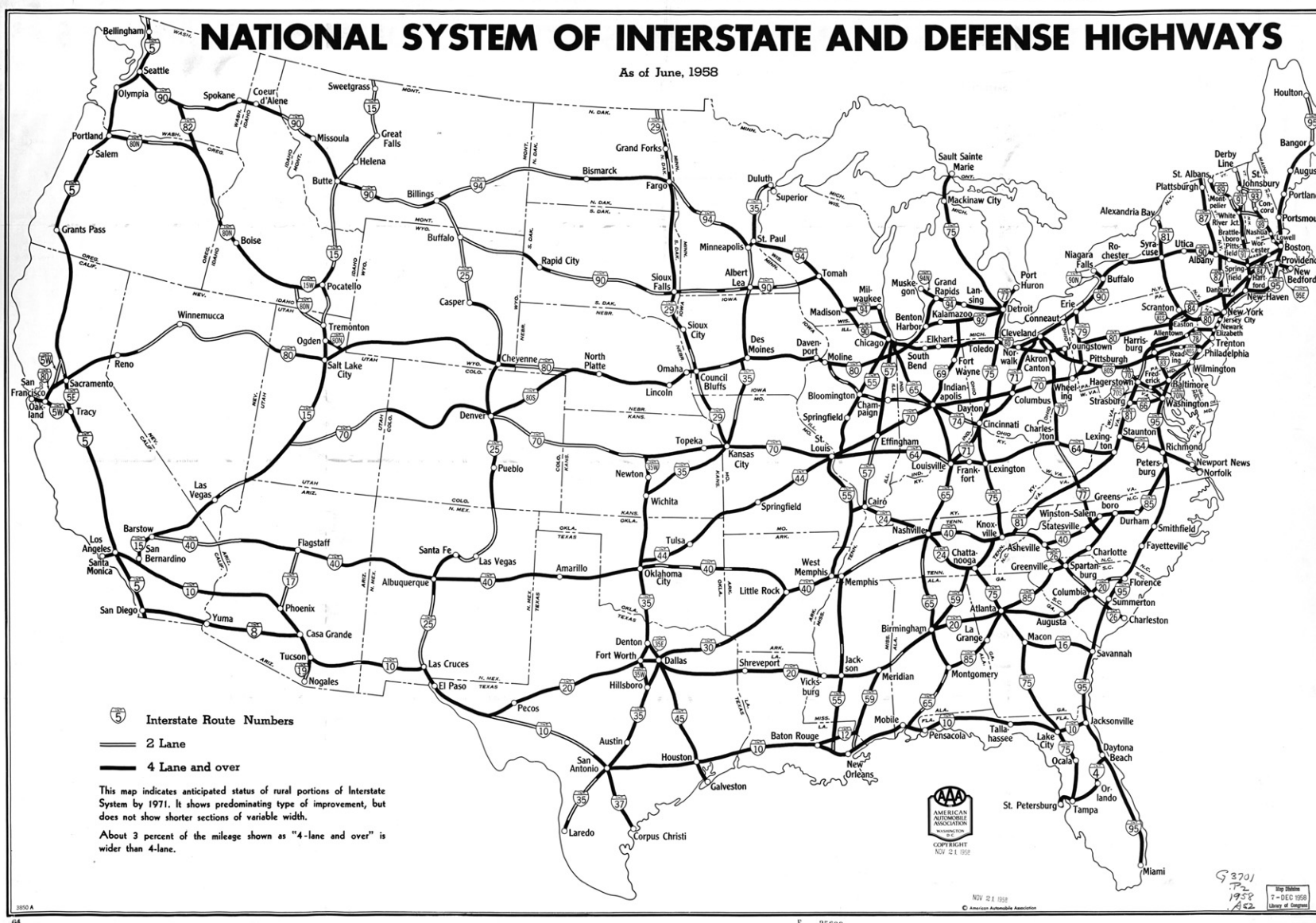


Figure 2 Original map of the Dwight D. Eisenhower System of Interstate and Defense Highways, USA. Source: FHWA.

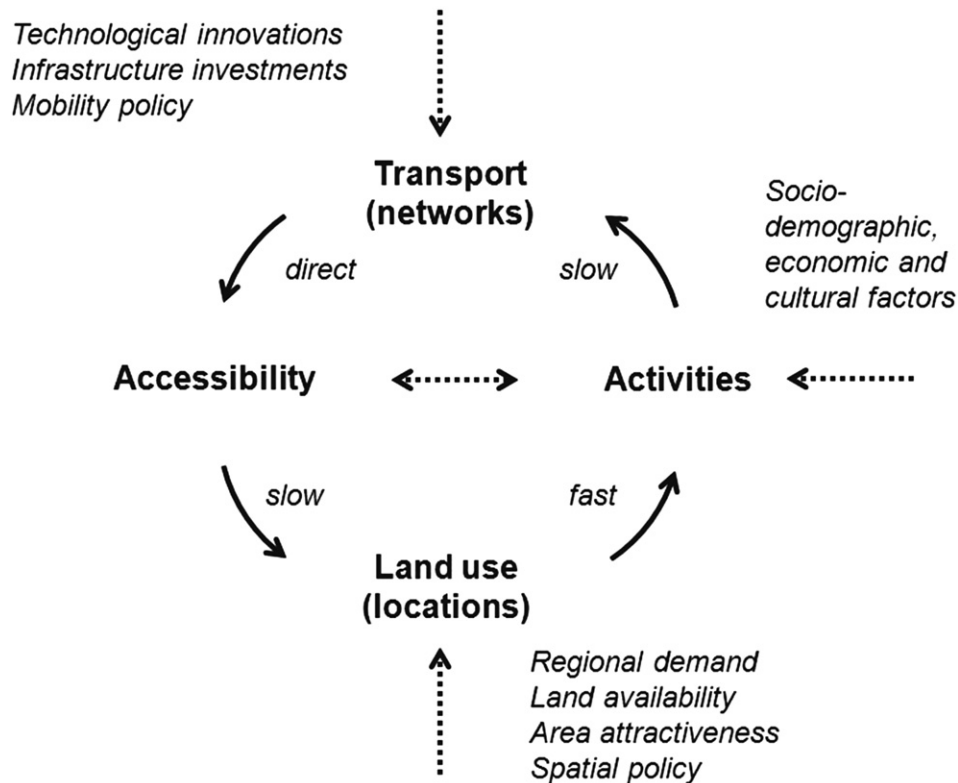


Figure 3 The land use transport interaction cycle (after Bertolini, 2012).

take, fragmentation, barrier impacts, noise, air pollution, safety issues etc. —which ask for a careful involvement of “passive users” of roads—the people live and companies that are located nearby.

All of this results in extensive *planning and management processes* for road development in order to carefully allocate (the huge) budgets—that is taxpayer’s money—needed to balance in a sustainable manner socio-economic growth and environmental impacts in the various places involved, and to weigh the various interests of a diversity of societal groups. Moreover, roads are planned and delivered in in a dynamic context, which means that throughout the life-cycle of road infrastructure—plan preparation, construction, operation, and renewal—*adaptive management* is needed to deliver societal desired outcomes.

After this discussion of the main characteristics of road infrastructure, the remainder of this chapter discusses: how road planning (section “Planning for road infrastructure”) and management (section “Management of road infrastructure”) is done, the impacts of road infrastructure and their evaluation (section “Impacts of Road Infrastructure”), and the future trends and challenges in road infrastructure planning and management (section “Outlook: Future Trends and Challenges”).

Planning for Road Infrastructure

Planning can be described as the “process of preparing a set of decisions for action in the future directed at achieving goals by preferable means” (Dror, 1963). Worldwide, one can see that *national governments play an important role* in road planning. This regards not only national highway networks but also (co)financing and -developing of regional and local roads, because of the huge budgets and the planning complexities involved that often go beyond the capacity of decentral authorities. The planning of road infrastructure is *complex*, there are many fields of expertise needed (e.g. civil, traffic and environmental engineering, spatial design, ecological, economic, sociological, legal, and management knowledge), many conflicting interests, many interacting actors involved (both public authorities, private companies, civil society), at different levels, at different stages of the development and management process (see **Introduction**). Content and process have to be carefully dealt with addressing the dimensions of space, level, time, and actors in order to achieve sustainable outcomes—see **Fig. 4**. Regarding this, *different planning approaches* evolved over time, which can be seen next to each other in current planning practice.

Traditionally, road infrastructure planning has been mainly a *technical-rational* exercise, in which engineers and economists were dominant, looking for the optimal road network for the “the greater good”—sometimes coined with the term “*blue-print planning*” in which “predict and provide” and projects are central (Banister, 2008). Such planning approach—related to *Traditional Public Administration*—has been dominant in the Global North for most of the 20th century. However, the development of road infrastructure and (auto)mobility led also to more environmental impacts and awareness, and consequently to *resistance against*

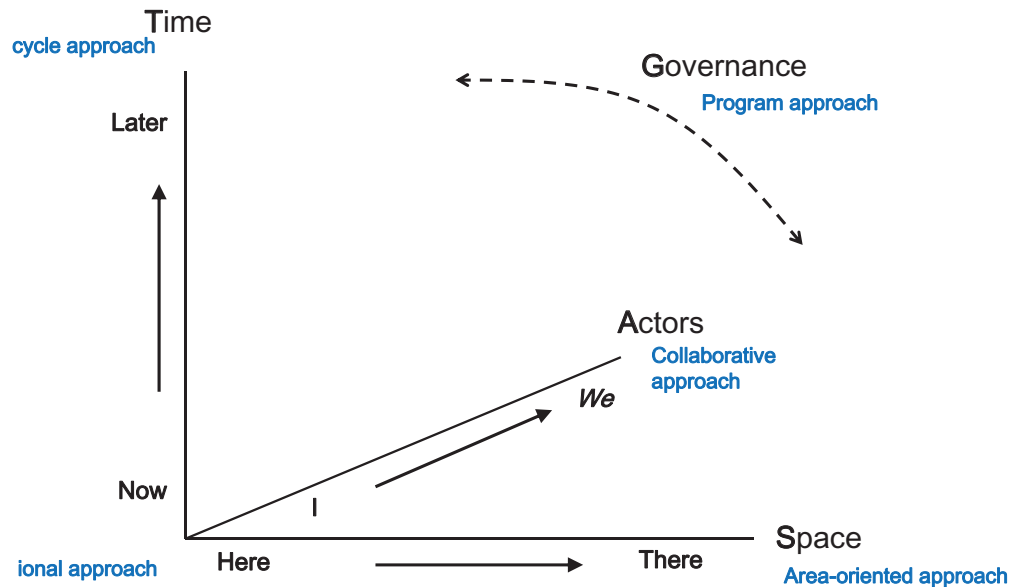


Figure 4 Different dimensions of inclusive road planning and management and related approaches.

the negative impacts of road development. As a reaction to this, since the 1960s the planning process for road infrastructure has become more and more formalized (Arts et al., 2016) resulting in a structured planning process of forecasting, agenda-setting, defining goals, developing alternatives, assessing impacts, elaborating mitigation measures, decision-making, implementation, and operation. To support this process, tools and approaches were introduced in road planning such as traffic modelling, cost benefit analyses, Environmental Impact Assessment (see Impacts), which often became part of (legal) regulations—a process of *institutionalization*. As a result, information about goals and impacts were disclosed to external parties—such as environmentalists and residents—who became more and more successful in voicing their critique in the media and asking for an explicit role in the previously closed planning process. This fueled the introduction of formal and informal participation in the planning process from the 1970s onward. As a consequence, stakeholders could not only better voice their interests (e.g., environmental, local) but also were given the tools to effectuate their resistance before court—enlarging their *hindrance power*, NIMBY (Not In My BackYard). As a result, planned road projects were delayed or halted (Arts et al., 2016). In reaction to this, often new, more refined, procedures were introduced in an attempt to balance interests, enhance public participation (*participatory planning*) as well as giving (central) government instruments for final decision-making and implementation—that is to enhance government's *development power*. The latter fueling even more a process intense societal discussion, with government trying to enforce a final decision for “iconic” projects, but at higher costs, longer planning periods and often general discontent amongst all parties (Cantarelli and Flyvbjerg, 2013; Priemus et al., 2008).

To overcome this, *collaborative planning approaches* have been introduced that acknowledge that the development power for societal sensitive issues as road infrastructure is not with only one party—for example, the national government—but is distributed amongst many governmental levels (national, regional, local) as well as nongovernmental parties—interest groups, residents, and also private companies. These *institutional interdependencies* drove a shift from government to governance and from hierarchic governance to network governance and market governance (Kooiman, 2003; Meuleman 2008). During the 1980s–1990s, the role of national government was criticized, giving rise to *New Public Management* (Osborne and Gaebler, 1992) resulting in transferring tasks and responsibilities from the central government to private companies (resulting in e.g. *public–private partnerships*, see **Management**), decentral government (giving rise to *multilevel governance* approaches) as well as to the public itself (self-governance; engaged civic society). However, these new approaches introduced new (institutional) complexities—strategic uncertainty about other stakeholders’ objectives and behavior. In combination with societal challenges of climate change, energy transition, biodiversity, globalization, and social inclusion this results in deep uncertainties (see Outlook), that call for more *adaptive planning and management approaches* that address well the public values and various societal interests involved (see **Introduction**)—*Public Value Management*.

As discussed above, in many countries an intricate *system of planning procedures and processes* has been developed, which usually involves an overall arrangement for planning programming budgeting (PPB) regarding the road network development, and a more specific planning cycle for road projects—plan preparation, construction, operation, and renewal. As road infrastructure development involves huge public budgets, see **Introduction** and to allocate “tax payers’ money” carefully, many countries have a system of *planning programming budgeting (PPB)* in place that closely interacts with the (yearly) budget and account cycle of national government (Klakegg et al., 2016; Lee et al., 2013; van Geet, 2020). Usually, on basis of traffic forecasts, bottlenecks, and missing links in the infrastructure networks are detected resulting in a visioning process about what new road development is needed (see **Impact**). This *agenda-setting* process involves usually intensive negotiation between national, regional, and local government levels

and is heavily politicized—what is included in the scope and who is (co)financing future projects. In many countries the tax base of local and regional authorities is too limited to finance fully the development of roads. Usually national government is not only responsible for the national road network but also (co-)funding local and regional road networks. To implement the results of the agenda, *operational programming* is needed of projects, their budgets and timing—especially as long-time frames are involved in infrastructure development. For providing the budgets for the program, financing by other parties is often important: tolling or other forms of road pricing for road users; cofinancing by decentral authorities; loans by international development banks (e.g., the WorldBank); or private financing via public–private partnerships (see **Management**).

For developing specific sections of the road infrastructure network, a *planning cycle* is relevant that is usually laid down in regulations and procedures. The PPB process results in proposed projects that are elaborated in a *project study* involving such elements as: traffic forecasting; baseline monitoring; development of design alternatives (alignment of the route, road lay-out, preliminary engineering); assessment of environmental and social impacts; analysis of (societal and actual) costs and benefits; development of mitigation and compensation measures; reporting; public participation; and formal decision-making. Usually, government is initiator of a project study—unless a concession has been given to a market party—but the study itself is often outsourced to private consultancy firms. In most countries, (larger) road projects are (legally) subject to a formal Environmental Impact Assessment (EIA, see **Impacts**) and forms of public participation—also international development banks usually require this for providing loans. In addition to formal *participation*, open planning approaches are increasingly employed to disclose information and to involve local residents, companies, NGOs, and other stakeholders, requiring careful process management (see **Management**). The *formal decision*, giving consent to build the road, is with authorities, and for the national network often centralized to national government. In most countries such decision is open to appeal to court—where consent decisions are regularly annulled because of not sufficiently taking into account other (stakeholders’) interests. The social cost–benefit analysis (see **Impacts**) is often important to substantiate the need for the budgets, thereby providing a linkage between project decision-making and strategic planning and programming. To improve this linkage, sometimes an explorative study is done for defining the project’s preliminary scope, timing and budget, and for linking PPB-programming to individual project elaboration as discussed above. To *implement* a road project, construction activities are usually contracted out to building companies by way of procurement—see **Management**. If a road is delivered, it becomes part of the wider road network and its *operational management and maintenance*—which involves daily maintenance of the road, traffic management, and incident management. During operational stages also monitoring of traffic and environmental issues is important in order to enable ex post evaluation—although this is still not common practice (see **Impact**). Over time road capacity can become too limited, which after addressing in the PPB-process may result in a *new planning cycle of road (re)development*. As infrastructure is aging and functionality is decreasing, renovation and renewal can also become a reason for starting a new planning cycle (see **Outlook**).

Management of Road Infrastructure

Infrastructure management is about the organizing, conditioning and steering of actors, activities, and processes in such a way that policy is effectively and efficiently implemented. It may concern the management of the functioning of the road infrastructure network as well as the management of the development of a road section (see **Planning**). Tasks of a public or private infrastructure network manager are manifold and typically comprise: *network management*, i.e. keeping the mobility network up-to-date and future proof; *asset management*, i.e. keeping the physical assets (pavement, bridges, viaducts, noise barriers, information panels etc) functioning; *traffic management*, i.e. managing the traffic flow over the infrastructure (relevant regarding congestion, safety, environmental issues etc.); and *incident management*, i.e. quickly responding on unexpected disturbances of the network (such as traffic accidents, spoilage of hazardous goods, flooding because of heavy rainfall). Asset management, traffic management, incident management—as well as mobility and logistics management—need to be carefully coordinated by the network manager in order to make optimal use of the existing road capacity.

The development of the road infrastructure network itself—a new section, extension, renewal, renovation, solving bottlenecks—is typically performed by project management and often involves process management or (portfolio) program management (see also **Planning**). A project can be defined as a temporary organization designed to deliver a specific result by a subsequently phased process (Turner and Müller, 2003). In *project management* focus is on achieving goals and meeting success criteria, such as scope, time, budget and quality (see e.g. Meredith and Mantel, 2011; Kerzner, 2013). A specified scope and focus on risk management should prevent that changing circumstances will affect project performance – resulting in ‘being in control’ (Bourne and Walker, 2005). *Process management* is often used in situations with many stakeholders involved and where careful interaction is needed (Glasbergen and Driessen, 2005). It is focused on guiding the planning process by reacting flexibly to changes, by bringing different stakeholders together (authorities, market parties as well as civic society), and gaining acceptance (Edelenbos and Klijn, 2009; de Bruijn and ten Heuvelhof, 2002). The increasing influence of the environment on projects (see **Planning**) and the increasing influence of projects on their environment (see **Impact**) makes that projects do not stand alone and should be managed in a more interactive way. The traditional (closed) way of project management, focusing on the efficient realization of project objectives, gradually shifts to process management, interactively developing projects with their environment (Heeres, 2017). To provide a framework for strategic direction to a set of projects, portfolio programs are used so that in combination projects can deliver higher order objectives or win-win’s (Pellegrielli, 1997). In such *program management* various projects, project stakeholders and related activities are coordinated in order to realize benefits that could not be obtained where projects are implemented separately

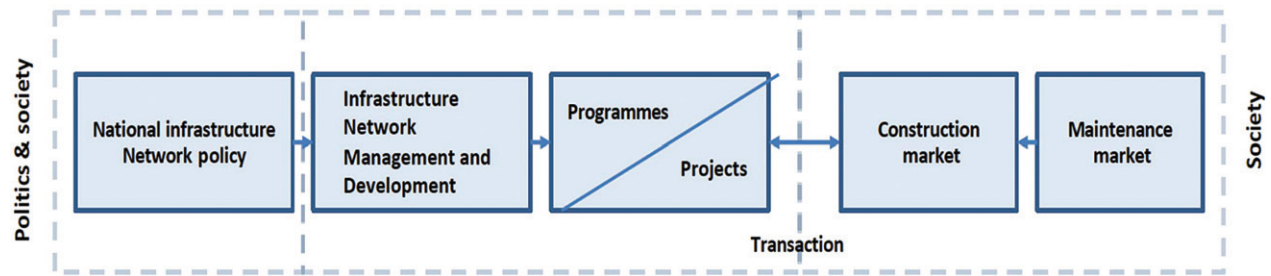


Figure 5 The public road infrastructure value chain (after Leendertse and Arts, 2020).

(Busscher, 2014; Pellegrinelli, 1997). Program management can be seen as a response to the fragmentation caused by ‘projectification’ (Buijs and Edelenbos, 2012). Sometimes related projects are combined and managed on two levels, the program level to gain synergy and address overproject objectives and the project level for efficient output realization (Turner and Müller, 2003).

Following New Public Management (see **Planning**), road infrastructure management has been “projectified” (Buijs and Edelenbos, 2012; Busscher, 2014; Maylor et al. 2006) to a large extent, with a strong focus on efficiency in the realization of projects and involvement of market parties thereby. Projects are linearly (and technocratically) organized (Leendertse and Arts, 2020—see Fig. 5), which means that the individual planning stages are successively run through following strict protocols, with each phase building on the outcome of the previous phase. Management is thus strongly focused on control. Central for control and an important tool of project management is *risk management*. Here a risk can be seen as a probable event with a possible negative effect on time, cost, scope or quality. Typical risk management involves the identification and quantification (probability * effect), the establishment of risk management measures, the implementation of these measures and the regular evaluation and updating of the measures (see **Impact**).

In infrastructure development private companies are usually involved in preparing project studies, construction and maintenance of the infrastructure, which has to be done carefully considering the major societal interests and budgets involved (see **Planning**). As a consequence, an important part in project and program management is the management of market parties’ involvement—that is *procurement and contracting*. Procurement and contracting are part of private law, while in most countries planning procedures are part of public law. A careful link between both proves to be often complicated but also essential for delivering societal ambitions (see e.g. Lenferink et al., 2012). Public authorities are bound by (international and national) procurement regulations, based on equal treatment of potential market suppliers (“level playing field”). Procurement is thus very strictly regulated in most countries (Leendertse and Arts, 2020). In an *open procurement procedure* any interested market party can, as reaction on an announcement, register. The registrations are then tested for exclusion criteria, suitability requirements, terms of reference and awarding criteria. The tender ends with a winner who is awarded the contract. In a *restricted procedure* a preselection of interested candidates is added at the beginning of the procurement procedure. In order to stimulate a bidding market party—a private company or consortium of private companies, as projects can be very large and involve many expertise fields—to think creatively about the societal challenges involved, new forms of procurement procedures have been developed (Lenferink, 2013). These procurement procedures can only be used in special cases, such as the *negotiated procedure*, in which after tendering the contracting authority can negotiate with one or more tenderers and/or the *competitive dialogue*, in which a dialogue is held with the candidates on the framework of the tender. A special procedure is the *prize competition*, which may be an open or restricted procedure with award by means of judging by a jury. The procurement process ends with the awarding of a contract to the winning bidding consortium.

In order to accommodate the complexity of a project and dealing carefully with the interests at hand, various **contract forms** can be seen in the road infrastructure construction sector. The **build contract** is the traditional contract model for construction projects—see Leendertse and Arts (2020). The responsibility of a contractor is solely to build in accordance with the specifications of the request—provided by the authorities—called the client. A variant is the *engineer & build contract*, where the contractor has to do basic engineering of a by the client provided design. In a *design & build contract* (or design & construct) no specified design is provided by the client, only a desired outcome is presented. The essence of design & build is that a “functional space” is provided, within which market parties are allowed to search for solutions themselves. The solutions must meet the minimum quality level and must deliver the requested functions as specified by the functional specifications. Often within the functional space, direction is given through an incentive scheme about prize/quality such as *Most Economical Advantageous Tender (MEAT)*. This tender criterion enables the contracting authority to take account of criteria that reflect qualitative, technical, and sustainable aspects of the tender submission as well as price when reaching an award decision.

In line with the *New Public Management*, since the 1980s, more and earlier involvement of market parties has been stressed in order to enhance efficiency of road infrastructure (project) delivery, resulting especially in the introduction of *public–private cooperation (PPP)*. Starting with the so-called Private Finance Initiative in the UK many countries followed introducing *Design, Build, Finance, and Maintain (and Operate) contracts*. The essence of these contracts is that private capital is used to (pre)finance public infrastructure. The private consortium delivers infrastructure (or even operates it) in return for a fee over a relatively long period

(25–50 years), which fee is related to the quality delivered. The fee enables the private consortium to pay back the loan to the lenders (banks and other financial organizations). Through the “shadow of the banks” the private consortium is stimulated to deliver the required quality. In these type of contracts not only tasks—to enlarge the functional space for optimizing the (development of the) road infrastructure—but also risks are transferred to the private consortium. Since risks are rather high in infrastructure development leading to a high-risk exposure of the construction sector, this has lately led to an intense discussion about the desirability of these kind of contracts (Leendertse and Arts, 2020; Verhees and Arts, 2016). More recently, in many countries a trend can be seen to go back to the more traditional way of procurement through design & build contracts separated from maintenance, operation and finance, where the public authorities (re)take the majority of risks and effects of uncertainty. This provides also room for addressing (broader) public interests and values of the area adjacent to the road—*Public Value Management*.

This trend is reinforced as road projects and their contexts are becoming more *complex* through the increasing interaction with, and dependency, on the project’s environment and stakeholders. As a consequence, uncertainties are larger; complex systems are *less predictable and are sensitive for changes*. Moreover, the world of infrastructure and spatial development is increasingly dynamic (see **Outlook**). For instance, smart mobility and data and information technology may alter the role of the traditional road infrastructure manager and the traditional road. Smart Mobility concerns a wide range of ‘smart’ initiatives and measures to address control of travel speed, accessibility and mobility in general. It includes specialized applications in vehicles, mobile apps and mobile services (Mobility as a Service, MaaS), behavioral influencing measures, organizational measures, smart infrastructure and measures to increase the flow and capacity of infrastructure (traffic and mobility management). New market parties offering new mobility products require new relationships between market, user, and road infrastructure manager. The different Infrastructure networks and different forms of mobility will be increasingly connected by integrated efficient transfer hubs which makes them mutually dependent. Climate change, sustainability, and the energy transition might have enormous impact on our road infrastructure. However, there is little no certainty about the magnitude and even the direction of development. As a response to this increasing uncertainty and dynamism, enhancing *adaptivity* in planning and management is seen as a way forward in both practice and scientific literature (Rauws and de Roo, 2016; Skrimizea et al., 2019; Zandvoort and Willems, 2017). Adaptivity can be described as the capacity of a system to manage adjustment to changed situations, whereby the core function, the ultimate goal of the system, remains intact (Folke et al., 2010; Walker et al. 2004). *Adaptive management and planning* of road infrastructure have the ability to generate alternative solutions when the context changes (variation) (see also Holling, 1978; Axelrod and Cohen, 2000).

Impacts of Road Infrastructure

Road infrastructure has brought economic development in many cities and regions. At the same time ecological and social impacts are involved. Only a balance between social (“people”), ecological (“planet”) and economic (“profit”) impacts will result in truly sustainable infrastructure (and its usage). In general, economic and social impacts concentrate on human beings, while ecological effects merely focus on ecosystems (Geurs et al., 2009). Furthermore, economic impacts are often analyzed at the macro level of countries and regions, whereas social impacts are examined at the individual and group level on a local scale (Geurs et al., 2009). Regarding impacts we make a distinction here between the effects of infrastructure construction and use (Fig. 6).

Infrastructure has often been seen as a necessary condition for (regional) *economic growth*—see also [10396]. Historical examples include the construction of railroads (e.g., Wang and Latham, 2001) and the highway system—for example the Interstate Highway System in the USA, initiated by the Eisenhower Administration in 1956 (see Fig. 3)—which led to enormous travel time gains and connected new areas. Costs of transporting goods for instance fell by as much as 95% during the 20th century (Glaeser and Kohlhase, 2004).

The *construction of infrastructure* has direct temporary effects on labor in the construction sector. However, structural labor effects can also be discerned and relate to the operation and maintenance of roads. *Infrastructure usage effects* on the short run relate to travel and environmental hindrance due to construction works. After completion, however, travel time gains occur, which convert into

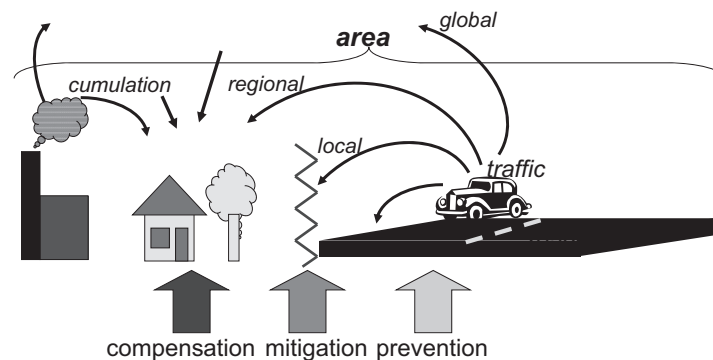


Figure 6 Impacts of road infrastructure and its usage.

economic impacts by increased productivity, employment, firm relocations, and increasing office prices. For people, these accessibility benefits result in higher incomes, a higher productivity and also in increasing house prices.

Theoretical microeconomic models from the 1960s link transport infrastructure to consumer and producer location behavior (e.g. Alonso, 1964; Muth, 1969; Wingo, 1961) and suggest that *location decisions* of firms and individuals are largely based on transport. However, a growing body of empirical studies (see Melo et al., 2013) in the last decades indicate that the relationship between infrastructure and economic development may not be as high as traditionally expected (i.e. “the law of decreasing returns to scale”; Bruinsma, 1995). Generally speaking, at a reasonable infrastructure level, transport development serves as a growth supporter rather than a growth generator (Banister and Berechman, 2000).

Infrastructure creation and usage also have *social impacts* (Geurs et al., 2009). Examples of the former include relocation (from land), destruction of buildings and cultural heritage artefacts, barrier effects, noise nuisance due to construction works, traffic hindrance, and related safety issues. In a structural way, road infrastructure influences visual quality and may result in community fragmentation. Infrastructure usage has both positive and negative impacts. From a positive side infrastructure has a direct impact on accessibility of activity locations, which gives the opportunity to develop and increase personal wellbeing (better accessibility to work, education, health, shopping, and other facilities, to family, friends, social networks). However, more indirectly transport causes *external effects*, which road users often do not fully take into account in their decision processes. On a national and global scale, these include climate related greenhouse gas (GHG) emissions such as carbon dioxide (CO₂). On a local level road use influences traffic safety and leads to noise, vibrance, and air pollution (van Wee et al., 2013), all of which result in stress, livability, and health issues. These local impacts are often subject to intensive debate during the planning stages, resulting in *mitigation measures* such as different routing (bypasses), tunnels, (noise reducing) porous pavements, noise barriers, speed limits (see Planning).

At the same time, many effects mentioned influence *ecosystems* (van Bohemen 2005). Roads act as barriers which fragment nature and influence wild life safety. Moreover, noise, light, soil, water, and air pollution, such as small particles, sulfur dioxide and nitrogen oxides, negatively impact the flora and fauna. The latter relates to impacts on biodiversity which go beyond the local scale. Moreover, indirectly CO₂ and other GHG-emissions cause climate change and thereby fundamentally affect ecosystems at a global scale. Measures relating to ecosystems impacts may vary from porous pavements, noise barriers, to culverts, under- and overpassages (“ecoducts”) to remediate habitat fragmentation. If ecological impacts cannot be prevented, nature compensation can be implemented by developing elsewhere comparable nature—ecological offsets (Cuperus, 2005).

Reducing traffic congestion, and thereby striving for shorter travel times, is an important reason for extending road capacity. The *societal costs* due to traffic congestion in many countries are large. However, these costs are generally speaking substantially lower than the costs related to traffic safety and the environment. For instance, the societal costs due to traffic jams and delays on the Dutch main road network, for instance, amounted to approximately €3- 4 billion in 2016. This is comparable to half a percent of the gross domestic product (GDP) of the Netherlands. However, the costs for noise and air pollution from traffic amounted to no less than €8 billion in comparison (KiM, 2016). And the total costs of road safety in the same year were even rounded up to between €13-16 billion. This is more than four times the congestion costs.

The *distribution of costs and benefits* of transport infrastructure are uneven across space and people. By improving regional connections, regional differences (e.g. in employment) may decrease but may also increase as a consequence of agglomeration. In the latter case economic activity further increases in already stronger economic regions at the cost of the “weaker” ones (i.e., “core-periphery model” by Krugman, 1998). Not only the most populous cities in such strong regions may then benefit, but also smaller towns which are part of the *Daily Urban System* (see also Planning). This relates to the concept of *borrowed size*, which occurs according Meijers and Burger (2017) “when a city possesses urban functions and/or performance levels normally associated with larger cities. This is enabled through interactions in networks of cities across multiple spatial scales”.

Moreover, effects may be *distributed unevenly* from a regional compared to a local perspective—between the central city and its surrounding rural area. Decisions about transport infrastructure are often dominated by macroscale accessibility issues and economic development of a region (Banister and Berechman, 2000; Tillema et al., 2012). However, undesirable impacts such as noise, air pollution and local traffic are particularly important at the local level (Bateman et al., 2001)—see Planning. This results in uneven social impacts for people living close to transport infrastructure without using it regularly. This relates to issues *equity and fairness* (coined as *transport justice* by Martens, 2017; see also Banister, 2018). Another example of unequal distributions revolves around public funds that are used for highway extensions. Such investments are particularly beneficial for those already using the highway, but far less for nonregular car users (Martens, 2017).

The impacts of new infrastructure projects can be evaluated at different stages of the planning cycle—before (*ex ante*) and after (*ex post*) the infrastructure realization. In planning practice, *ex ante evaluations* are often applied in road planning, whereas *ex post evaluations* are rare. Some important *ex ante* evaluation tools and initiatives can be discerned. *Social cost benefit analysis* (sCBA) has become an important decision support tool for infrastructure projects in many countries (Berechman, 2009; Heeres, 2017; SACTRA, 1999). The aim is to derive a summary indicator of all costs and benefits for actors affected. (Estimated) travel time gains of road infrastructure are the most important and quantifiable benefit component. Although sCBA also includes environmental impacts, such effects are harder to measure in monetary terms (van Wee et al., 2013). Therefore, to safeguard sustainable development also from a social, environmental, and ecological perspective *Environmental Impact Assessments* (EIA) and *Social Impact Assessments* (SIA) for large infrastructure plans or projects are applied (see IAIA, 1999; Vancley et al., 2015). According to IAIA (1999) EIA is “the process of identifying, predicting, evaluating and mitigating the biophysical, social, and other relevant effects of development proposals prior to major decisions being taken and commitments made”. It comprises the planning, design, decision making and implementation (follow-up) stages of that action (Morrison Saunders and Arts, 2004). SIA focuses specifically on the social

perspective – social inclusion, social inequality – and can be defined as the “process of identifying and managing the social issues, including effective engagement of affected communities in participatory processes of identification, assessment and management of social impacts” (Vancly et al., 2015). These assessments often make use of multicriteria analysis to weigh and compare different alternatives using both qualitative and quantitative data. EIA is regulated in almost all countries (NCEA, 2020) and large road projects are usually subject to a formal EIA procedure, in which environmental impacts of road development are assessed and measures are discussed to prevent, mitigate, or compensate the impacts. As impacts are difficult to predict because of the many uncertainties, *ex post monitoring, and evaluation*—that is after the consent decision during construction and operation—is essential to safeguard that the impacts stay within the boundaries of the formal planning consent and to facilitate *adaptive management* to implement remedial measures (see **Management**). Examples of this are EIA follow-up and Environmental Management Systems, which however are not (yet) standard practice in road planning and management (Arts and Faith-Ell, 2012; Morrison-Saunders and Arts, 2004).

Outlook: Future Trends and Challenges

Many trends and challenges are likely to impact road infrastructure planning and management in the coming decade(s). Without having the intention of being complete or exhaustive, three important and related (clusters of) trends can be distinguished: sustainability, smart (technological) innovation and spatial integration—see also [10442].

Important *sustainability* challenges of road infrastructure relate to climate change and broader environmental aspects—such as depletion of nonrenewable resources, biodiversity, environmental pollution, health, social inclusion—see also [10775]. To limit greenhouse gas emissions there is an increasing focus on the use of clean and renewable energy—such as electricity [10419] and hydrogen—which may impact refueling locations and processes. At the same time road infrastructure could be used for (local) energy generation, transportation and storage. Here one might think of for instance solar panels in asphalt, overhead electricity lines for electric trucks, windmills near roads. Such development might open possibilities for road infrastructure managers to link up with initiatives of civic society regarding local energy production (Spijkerboer, et al. 2019). Moreover, an increasing attention can be seen for mitigating air pollutants and noise hindrance such as: more strict regulations for car engines, tyres, establishing Low Emission Zones (LEZ), specific time periods in which certain vehicles allowed (or not), road pricing [10450], or stimulating a shift from car to other modes such as public transport, biking, walking. This is important as in densely populated areas road projects are increasingly difficult to realize because of growing environmental concerns. Future trends, however, do not only focus on mitigation but also on adaptation and on circular processes (i.e. reuse of materials—such as asphalt, concrete, steel—which becomes more relevant as infrastructure objects are ageing). Climate change will lead to more extreme weather conditions (e.g. drought, extreme rainfall) and to higher flood risks, which influence the availability and robustness of infrastructures. Planners will have to increasingly take account of all such impacts to maintain livability in localities and regions (see **Impacts**).

The use of *smart technologies*—such as Cooperative-Intelligent Transport Systems C-ITS—focuses on making better use of road capacity, increasing sustainability, and traffic safety—see **Management**. Examples include autonomous vehicles—see also [10400], smart infrastructures, and mobility-as-a-service (MaaS). These developments mutually interact. Autonomous vehicles are expected to communicate with each other and with the physical and digital road infrastructure (vehicle-to-vehicle, vehicle-to-infrastructure, and vehicle-to-cloud communication). The other way around, smart vehicle sensors may give information about the road condition which can be used for improving maintenance planning but may also be relevant for road pricing. In practice, however, privacy of information issues proves to be important institutional barrier for sharing such information between public and private parties. Another important trend is the gradual shift towards more sustainable multimodal travel—intermodal travel, or from “modality” to integrated “mobility” (from A to B) [10414]. This closely links to the concept of *mobility-as-a-service* (MaaS), which focuses on demand driven travel, fully facilitated via digital platforms, and real-time information [10420]. As a consequence of these innovations, new (market) parties enter the field of road planning and management (see **Management**). Growing interest in MaaS-concepts is stimulated by the development of (multimodal) *hubs*, both for persons transport (mobility management) and for freight transport (logistics).

Finally, a growing attention can be seen to *spatial integration*, with roads being an integral part of their surrounding area. This includes the integration of different transport networks via (multimodal) hubs and corridors (internal integration), but also relates to a better integration of infrastructure in its spatial environment (external integration; Heeres et al., 2012) – see **Planning**. For instance, multi-modal hub development relates closely to *Transit Oriented Development* (TOD; see Curtis et al., 2009) in which nodal and spatial quality are enhanced by employing land use transport interaction (LUTI – Wegener and Fürst, 1999; Bertolini et al. 2005, Bertolini, 2012). By an integrated planning of infrastructure and spatial development, negative environmental impacts may be prevented or undone (by more integrated design), social inequalities addressed (by carefully engaging communities) and economic synergies achieved (by optimizing infrastructure development with spatial development). Such so-called area-oriented approach may enhance the value of road infrastructure at multiple scales: accessibility at corridor level, socio-economic development at regional level, and spatial and environmental quality at local level (Arts et al. 2014; Heeres 2017; see also www.nuvit.eu, www.vitalnodes.eu, www.cedr.eu). This more inclusive focus asks for integrated area-oriented planning and management approaches with innovative governance arrangements, for collaboration between public authorities, private companies with civic society (van Geet, 2020). Such approach is not only relevant to the planning of new road infrastructure but also to enhance adaptiveness of existing road infrastructure (operation, maintenance, management) and its renewal—creating a new cycle of sustainable (re) development in road infrastructure planning and management.

Biographies



Jos Arts is a professor of Environment & Infrastructure Planning at the Faculty Spatial Sciences, University of Groningen, the Netherlands. He is also extraordinary professor at Northwest University, Unit Environmental Sciences & Management, Potchefstroom, South Africa. His research focuses on impact assessment and on institutional analysis and design for integrated planning approaches for sustainable infrastructure networks (Orcid: 0000-0002-6896-3992).



Wim Leendertse is a professor of Management in Infrastructure Planning at the Faculty Spatial Sciences, University of Groningen, the Netherlands. He is also top advisor for Rijkswaterstaat, the executive agency of the Ministry of Infrastructure and Water Management in the Netherlands. His work focuses on area-oriented infrastructure development, adaptive infrastructure management, and sustainable infrastructure (Orcid: 0000-0001-7706-1272).



Taede Tillema is a professor of Transport Geography at the Faculty Spatial Sciences, University of Groningen, the Netherlands. He is also senior researcher at KiM Netherlands Institute for Transport Policy Analysis. His research focuses on transport mobility patterns, transport innovation and transport planning. He also focuses specifically on the accessibility of, and the transport mobility patterns in, rural areas (Orcid: 0000-0001-7158-9893).

Relevant Websites

Information on planning and management of road infrastructure, see e.g.

<https://highways.dot.gov>

https://ec.europa.eu/transport/modes/road_en

<https://www.itf-oecd.org/road>

Information on integrated road planning, planning approaches, tools and examples see on: integrated planning for road infrastructure and spatial development 'networking for urban vitality'

- <https://www.nuvit.eu>
- the integration of urban nodes at TEN-T corridors www.vitalnodes.eu; <https://www.vitalnodes.eu>; collaborative planning of infrastructure and spatial development, and on freight and logistics in a multimodal context <https://www.cedr.eu/home/publications/>

Information impact assessment (e.g. EIA, SIA), see:

<https://www.iaia.org/publications.php>

Information about innovations:

<http://connectedautomateddriving.eu>

References

- Alonso, W., 1964. Location and Land Use. Harvard University Press, Cambridge, MA.
- Arts, J., Faith-Elli, Ch., 2012. New governance approaches for sustainable project delivery. *Procedia - Soc. Behav. Sci.* 48, 3239–3250.
- Arts, J., Hanekamp, T., Dijkstra, A., 2014. Integrating land-use and transport infrastructure planning: towards adaptive and sustainable transport infrastructure. *Transport Research Arena Proceedings 2014*, Paris.
- Arts, J., Filarski, R., Jeekel, H., Toussaint, B. (Eds.), 2016. Builders and Planners—A History of Land-use and Infrastructure Planning in the Netherlands. Eburon, Delft.
- Axelrod, R., Cohen, M., 2000. Harnessing Complexity: Organizational Implications of a Scientific Frontier. The Free Press, New York.
- Banister, D., Berechman, Y., 2000. Transport Investment and Economic Development. UCL Press, London.
- Banister, D., 2008. The sustainable mobility paradigm. *Transport Policy* 15 (2), 73–80.
- Banister, D., 2018. Inequality in Transport. Alexandrine Press, Marcham.
- Bateman, I., Day, B., Lake, I., Lovett, A., 2001. The Effect of Road Traffic on Residential Property Values: A Literature Review and Hedonic Pricing Study. Scottish Executive & Stationery Office, Edinburgh.
- Berechman, Y., 2009. The Evaluation of Transportation Investment Projects. Routledge, London.
- Bertolini, L., 2012. Integrating mobility and urban development agendas: a manifesto. *disP—Plann. Rev.* 48 (1), 16–26.
- Bertolini, L., Le Clercq, F., Kapoen, L., 2005. Sustainable accessibility: a conceptual framework to integrate transport and land use plan-making. *Transport Policy* 12, 207–220.
- van Bohemen, H., 2005. Ecological Engineering—Bridging between Ecology and Civil Engineering. Aeneas Technical Publishers, Delft.
- Bourne, L., Walker, D.H.T., 2005. The paradox of project control. *Team Perform. Manage.* 11 (5/6), 649–660.
- de Bruijn, H., Ten Heuvelhof, E., 2002. Policy analysis and decision making in a network: how to improve the quality of analysis and the impact on decision making. *Impact Assess. Project Appraisal* 20 (4), 232–242.
- Bruinsma, 1995. The Impact of New Infrastructure on the Spatial Patterns of Economic Activities. Vrije Universiteit, Amsterdam Research Memorandum 1995-12.
- Buijs, J.-M., Edelenbos, J., 2012. Connective capacity of programme management: responding to fragmented project management. *Public Admin. Q.* 36 (1), 3–41.
- Busscher, T., 2014. Towards a Programme-oriented Planning Approach—Linking Strategies and Projects for Adaptive Infrastructure Planning. University of Groningen, Groningen.
- CAD, 2020. Connected and Automated Driving. Retrieved 20200830 from <http://connectedautomateddriving.eu>
- Cantarelli, C.C., Flyvbjerg, B., 2013. Mega-projects' cost performance and lock-in: problems and solutions. In: Priemus, H., van Wee, B. (Eds.), *International Handbook on Mega-Projects*. Edward Elgar, Cheltenham, pp. 333–355.
- Cuperus, R., 2005. Ecological Compensation of Highway Impacts: Negotiated Trade-off or No-net-loss? Leiden University, Leiden.
- Curtis, C., Renne, J., Bertolini, L., 2009. Transit-Oriented Development: Making it Happen. Ashgate, Aldershot.
- Dijkstra, L., Poelman, H., Veneri, P., 2019. The EU-OECD Definition of a Functional Urban Area. OECD, Paris OECD Regional Development Working Papers.
- Dror, Y., 1963. Planning process: a facet design. Reprinted in: Faludi, A. (Ed.), 1973. *A Reader in Planning Theory*. Pergamon Press, Oxford, pp. 323–344.
- Edelenbos, J., Klijn, E.-H., 2009. Project versus process management in public-private partnership: relation between management style and outcomes. *Int. Public Manage. J.* 12 (3), 310–331.
- ERTRAC, 2019. Connected Automated Driving Roadmap. ERTRAC, Brussel.
- Folke, C., Carpenter, S., Walker, B., Scheffer, M., Chapin, T., Rockström, J., 2010. Resilience thinking: integrating resilience, adaptability and transformability. *Ecol. Soc.* 15. (4).
- van Geet, M., 2020. Policy Design and Infrastructure Planning—Finding Tools to Promote Land Use Transport Integration. University of Groningen, Groningen.
- Geurs, K.T., Boon, W., van Wee, B., 2009. Social impacts of transport: literature review and the state of the practice of transport appraisal in the Netherlands and the United Kingdom. *Transport Rev.* 29 (1), 69–90.
- Glaeser, E.L., Kohlhase, J.E., 2004. Cities regions and the decline of transport costs. *Pap. Regional Sci.* 83 (1), 197–228.
- Glasbergen, P., Driessen, P., 2005. Interactive planning of infrastructure: the changing role of Dutch project management. *Environ. Plann. C: Government Policy* 23, 263–277.
- Heeres, N., 2017. Toward Area-oriented Approaches in Infrastructure Planning: Development of National Highway Networks in a Local Spatial Context. University of Groningen, Groningen.
- Heeres, N., Tillema, T., Arts, J., 2012. Integration in Dutch planning of motorways: From 'line' towards 'area-oriented' approaches. *Transport Policy* 24, 148–158.
- Holling, C.S., 1978. *Adaptive Environmental Assessment and Management*. John Wiley, New York.
- IAIA, 1999. Principles of Environmental Impact Assessment best practice. Retrieved 20200720 from <https://www.iaia.org/uploads/pdf/Principles%20of%20IA%2019.pdf>.
- Kelly, E.D., 1994. The transportation land-use link. *J. Plann. Liter.* 9 (2), 128–145.
- Kerzner, H., 2013. *Project Management; A Systems Approach to Planning, Scheduling and Controlling*. John Wiley & Sons, Hoboken, NJ.
- KIM, 2016. Mobility Report 2016. KiM Netherlands Institute for Transport Policy Analysis. Retrieved 20200830 from <https://english.kimnet.nl/mobility-report/publications/documents-research-publications/2016/10/24/mobility-report-2016>.
- KIM, 2018. Mobility-as-a-Service and changes in travel preferences and travel behavior: a literature review. Retrieved 20200830 from <https://english.kimnet.nl/latest-news/feature/2018/09/17/mobility-as-a-service-and-changes-in-travel-preferences-and-travel-behaviour>.
- Klakegg, O.J., Williams, T., Shiferaw, A.T., 2016. Taming the 'trolls': major public projects in the making. *Int. J. Project Manage.* 34, 282–296.
- Kooiman, J., 2003. *Governing as Governance*. Sage Publications.
- Krugman, P., 1998. What's new about the new economic geography. *Oxf. Rev. Econ. Policy* 14. (2).
- Lee, R.D., Johnson, R.W., Joyce, P.G., 2013. *Public Budgeting Systems*, 9th ed. Jones & Bartlett Publishers, Burlington.
- Leendertse, W., Arts, J., 2020. Public-Private Interaction in Infrastructure Networks—Towards Valuable Market Involvement in the Planning and Management of Public Infrastructure Networks. In *Planning*, Groningen.
- Lenterink, S., 2013. Market Involvement throughout the Planning Life-cycle—Public and Private Experiences with Evolving Approaches Integrating the Road Infrastructure Planning Process. University of Groningen, Groningen.

- Lenferink, S., Arts, J., Tillema, T., van Valkenburg, M., Nijsten, R., 2012. Early contractor involvement in Dutch infrastructure: initial experiences with parallel procedures for planning and procurement. *J. Public Procurement* 12 (1), 1–42.
- Martens, K., 2017. *Transport Justice: Designing Fair Transportation Systems*. Routledge, London.
- Maylor, H., Brady, T., Cooke-Davies, T., Hodgson, D., 2006. From projectification to programmification. *Int. J. Project Manage.* 24 (8), 663–674.
- Meijers, E.J., Burger, M.J., 2017. Stretching the concept of 'borrowed size'. *Urban Stud.* 54 (1), 269–291.
- Melo, P., Graham, D.J., Brage-Ardao, R., 2013. The productivity of transport infrastructure investment: a meta-analysis of empirical evidence. *Regional Sci. Urban Econ.* 43 (5), 695–706.
- Meredith, J.R., Mantel, S.J., 2011. *Project Management: A Managerial Approach*. Wiley & Sons, Hoboken, NJ.
- Meuleman, L., 2008. *Public Management and the Metagovernance of Hierarchies, Networks and Markets*. Physica Verlag, Heidelberg.
- Morrison-Saunders, A., Arts, J. (Eds.), 2004. *Assessing Impact, Handbook of EIA and SEA Follow-up*. Earthscan, London.
- Muth, R.F., 1969. *Cities and housing, the spatial pattern of urban residential land use*. University of Chicago Press, Chicago.
- NCEA, Netherlands Commission for Environmental Assessment, 2020. Why ESIA/SEA, Utrecht. Retrieved 2020 0720 <https://www.eia.nl/en/our-work/why-esiasea>.
- Osborne, D.E., Gaebler, T.A., 1992. *Reinventing Government: How the Entrepreneurial Spirit is Transforming the Public Sector*. Addison-Wesley, Reading.
- Pellegrinelli, S., 1997. Programme management: organising project-based change. *Int. J. Project Manage.* 15 (3), 141–149.
- Priemus, H., Flyvbjerg, B., van Wee, B. (Eds.), 2008. *Decision-making on Mega-projects: Cost-Benefit Analysis, Planning and Innovation*. Edward Elgar, Cheltenham.
- Rauws, W., de Roo, G., 2016. Adaptive planning: generating conditions for urban adaptability: lessons from Dutch organic development strategies. *Environ. Plann. B – Plann. Des.* 43 (6), 1052–1074.
- Rayner, P., Reiss, G., 2013. *Portfolio and programme management demystified*. In: *Managing Multiple Projects Successfully*, Routledge, London.
- SACTRA, 1999. *Transport and the Economy*. SACTRA: Standing Advisory Committee on Trunk Road Assessment, London.
- Skrimiza, E., Haniotou, H., Parra, C., 2019. On the 'complexity turn' in planning: an adaptive rationale to navigate spaces and times of uncertainty. *Plann. Theory* 18 (1), 122–142.
- Spijkerboer, R.C., Zuidema, C., Busscher, T., Arts, J., 2019. Spatial integration of renewable energy and transport infrastructure networks: an institutional analysis. *J. Cleaner Prod.* 209, 1593–1603.
- Tillema, T., Hamersma, M., Sussman, J.M., Arts, J., 2012. Extending the scope of highway planning: accessibility, negative externalities and the residential context. *Transport Rev.* 32 (6), 745–759.
- Turner, R., Müller, R., 2003. On the nature of the project as a temporary organization. *Int. J. Project Manage.* 21, 1–8.
- Vancly, F., Esteves, A.M., Aucamp, I., Franks, D., 2015. *Social Impact Assessment: Guidance for Assessing and Managing the Social Impacts of Projects*. International Association for Impact Assessment (IAIA), Fargo, ND.
- Verhees, F., Arts, J., 2016. Public private partnerships—pursuing adaptive qualities in spatial projects. In: de Roo, G., Boelens, L. (Eds.), *Spatial Planning in a Complex Unpredictable World of Change—Towards a of Change*. InPlanning, Groningen, pp. 211–225.
- Walker, B., Holling, C., Carpenter, S., Kinzig, A., 2004. Resilience, adaptability and transformability in social-ecological systems. *Ecol. Soc.* 9 (2), 5.
- Wang, L., Latham, M., 2001. The role of railroad in the development of the American west-railroad-migration and urban growth. *Chinese Geogr. Sci.* 11 (3), 223–232.
- Weber, B., Alfen, H., 2011. *Infrastructure as an asset class. Investment Strategies, Project Finance and PPP*. Wiley, Hoboken, NJ.
- van Wee, B., Annema, J.A., Banister, D. (Eds.), 2013. *The Transport System and Transport Policy—An Introduction*. Edward Elgar, Cheltenham, ie.
- Wegener, M., Fürst, F., 1999. *Land-use Transport Interaction: State of the Art*. IRPUD, Dortmund.
- Wingo, L., 1961. *Transportation and urban land – Resources for the Future*. Washington, DC.
- Zandvoort, M., Willems, J., 2017. What's adaptive to what? Defining and using the concept of adaptiveness in planning. In: Zandvoort, M. (Ed.), *Planning Amid Uncertainty: Adaptiveness for Spatial Interventions in Delta Areas*. Wageningen University, Wageningen, pp. 49–71.

Further Reading

- Banister, D., 2008. The sustainable mobility paradigm. *Transport Policy* 15 (2), 73–80.
- Banister, D., Berechman, Y., 2000. *Transport Investment and Economic Development*. UCL Press, London.
- Bertolini, L., Le Clercq, F., Kapoen, L., 2005. Sustainable accessibility: a conceptual framework to integrate transport and land use plan-making. *Transport Policy* 12, 207–220.
- Busscher, T., *Towards a Programme-oriented Planning Approach—Linking Strategies and Projects for Adaptive Infrastructure Planning*, 2014, University of Groningen, Groningen.
- van Geet, M., 2020. *Policy Design and Infrastructure Planning—Finding Tools to Promote Land Use Transport Integration*. University of Groningen, Groningen.
- Geurs, K.T., Boon, W., van Wee, B., 2009. Social impacts of transport: literature review and the state of the practice of transport appraisal in the Netherlands and the United Kingdom. *Transport Rev.* 29 (1), 69–90.
- Heeres, N., Tillema, T., Arts, J., 2012. Integration in Dutch planning of motorways: from 'line' towards 'area-oriented' approaches. *Transport Policy* 24, 148–158.
- Krugman, P., 1998. What's new about the new economic geography. *Oxf. Rev. Econ. Policy* 14, (2).
- Leendertse, W., Arts, J., 2020. Public–private interaction in infrastructure networks. In: *Towards Valuable Market Involvement in the Planning and Management of Public Infrastructure Networks*, InPlanning, Groningen.
- Marshall, T., 2013. *Planning Major Infrastructure*. Routledge, Milton Park.
- Martens, K., 2017. *Transport Justice: Designing Fair Transportation Systems*. Routledge, London.
- Melo, P., Graham, D.J., Brage-Ardao, R., 2013. The productivity of transport infrastructure investment: a meta-analysis of empirical evidence. *Regional Sci. Urban Econ.* 43 (5), 695–706.
- Priemus, H., Flyvbjerg, B., van Wee, B. (Eds.), 2008. *Decision-making on Mega-projects: Cost-Benefit Analysis, Planning and Innovation*. Edward Elgar, Cheltenham.
- Tillema, T., Hamersma, M., Sussman, J.M., Arts, J., 2012. Extending the scope of highway planning: accessibility, negative externalities and the residential context. *Transport Rev.* 32 (6), 745–759.
- Van Wee, B., Annema, J.A., Banister, D. (Eds.), 2013. *The Transport System and Transport Policy—An Introduction*. Edward Elgar, Cheltenham.

Road Transport Planning at the Urban Scale

David A. King*, Kevin J. Krizek†, *Arizona State University, Tempe, AZ, United States; †University of Colorado Boulder Boulder, CO, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	373
Evolving Transport Modes	373
From Where and How?	374
Progressions	375
Performance Measures	375
Human-scaled Access	375
Re-boot	376
Concluding Thoughts	377
References	377
Further Reading	377

Introduction

Roads have always been organizing principles for cities. In the oldest settlements, roads grew organically based on local needs and desire lines. Over time, civilizations developed institutional capacities to plan and manage their character. While the primary function of roads, regardless of their ancestry, has always been to ease how people reach activities, be it the ability to exchange commercial goods, reach services or interact with others, advances in technology have affected changed their nature. What was once space reserved for the pedestrian morphed to space for the oxcart, then to the carriage, bicycle, omnibus, road car, and eventually the automobile. Each successive innovation created a different use and function.

In responding to quick advances in mobility innovation, planners and public officials are embarking on new era of road planning. They are being called on to more fully consider their function, which modes should be prioritized and supporting regulations. The outcomes of these decisions—how to manage and plan for roads—are rising in importance as cities tackle challenges of climate change, livability, and accessibility in cities.

The importance of understanding how roads are managed and planned is underscored by the fact that roads operate in rights-of-way which are largely fixed; expansion, reduction, or relocation is unlikely. For users of roads, they respond to policies from all levels of government. These policies send signals to favor various types of modes and how they are used (e.g., the speed of travel). As advances in technology play out and new travel modes prompt more claimants to limited street space, important questions emerge. How to move traffic, park cars, let kids play, encourage cycling, and anticipate whatever comes next yields an issue any planning textbook would easily depict as being truly “wicked.”

Consider a few of the contemporary issues with which city leaders are wrestling. Proponents of emerging technologies argue that tech-based gadgets provide revolutionary answers to urban traffic problems. Mobile phones and app-based inventions promise to deliver responsive mobility services, whatever shape those services may assume. Hive-minded autonomous or connected automobiles will inevitably increase the efficiencies of automobiles. Doing so, however, will continue to diminish the human qualities of streets and sidewalks. As long as road planning, inadvertently or not, favors personal auto mobility (Manville, 2017)—as it has in most cities for the past decades—cities will provide more of the same.

The right approach to road planning is to first recognize them as a public asset. Estimates suggest road space consumes between one-fifth and one-half of all land in cities. Any approach to design this space hinges on one’s perspective. Users of the transport system have preferences that are affected by their main mode of travel. Infrastructure providers and other interests, no doubt, have varying answers. But the transport industry is not charged with the responsibility to legislate how the road should function and what modes should or should not be encouraged to flourish. For many roads in a city, that responsibility goes to the political leaders.

Political leaders are the ones to legislate how local roads are used and the degree to which one mode is prized over others in decisions about the character of road space and how to appropriate its use. This chapter examines the complex and sometimes contradictory characteristics of mobility innovations and the rise of planning processes that shape mindsets and constrain change. We identify these processes as a major challenge in efforts to help cities move forward with, based on first principles, a future generation of public road space that prizes, above all, human-scaled accessibility.

Evolving Transport Modes

When a new transport innovation arrives, an opportunity emerges to envision how roads could or should evolve. Road designers have an obligation to consider the impact on the character of movement and rhythms within this public space. More fundamentally,

it allows the broader planning community to consider how cities regulate that space. Urban areas change incrementally as to most of their roads; they respond to demands and services that evolve with time. Regulations follow suit, accompanied by design standards.

The standards and regulations that have, for the past century, guided how roads should serve society firmly reinforce one form of mobility: vehicular travel. That was a policy decision. Changing these standards and regulations is a formidable challenge, no doubt, because planning processes and institutions are so firmly entrenched. In many respects, city leaders have lost an ability to envision roads as anything other than serving automobile. In cases where roads are redesigned, the scale of change is so small—such as a single bike lane or pedestrianized road—that the automotive status quo remains unchallenged. Losing sight of how other functions could be served restricts an ability to move on new and creative initiatives.

A main struggle currently facing city leaders is that a prevailing mindset about how road space is managed has permeated every aspect of city planning, urban design, and transport engineering. Roads are exclusively viewed as a space for car traffic. Consider, for example, a road that is temporarily re-appropriated by a city to support a neighborhood event or a *cyclovía*. City propaganda refers to these roads as closed even though during such events they may accrue more use than when they are open exclusively for cars. Residents have difficulty conceiving anything else. Politicians see little advantage to challenge the status quo. Even innovative transport professionals are quickly brought back in the fold, being reminded that, legally, it is best for cities to subscribe to current standards and regulations. The automotive mindset is pervasive. Efforts to prioritizing car movement has had profound impact in defining the role of road space and held hostage interests furthering anything else. This attention therefore affects the purpose and function of roads, their design characteristics, and role in connecting cities.

From Where and How?

We can trace the origin of how these ideas were legitimated: the development of bureaucratic management of city roads in early 20th century. Before 1900, roads were public places that served general purposes: buying and selling things, socializing—in addition to being grounds for passage. To move around, residents used whatever means of transport they considered best, including horses. Bicycles then emerged as a transformative transport innovation. Special interest groups, led by cyclists, quickly assembled themselves to improve both the quality of the roads (e.g., the Good Roads movement) and to pioneer institutional reforms for their use (Reid, 2015).

Then, the quick onset of the internal combustion engine and automobiles then monumentally changed the calculus of road users. Mixing horses with bikes, streetcars and eventually, automobiles, produced chaos. Cars were far faster and more powerful than anything prior road users had access to. Cars mitigated the need for horses which offered their own type of pollution (e.g., dung and carcasses) and were quickly viewed as the superior form of mobility. It became apparent that the interference between cars and other modes needed to be tamed.

To the rescue, new city departments of transportation were formed to address new traffic concerns. In the spirit of all forms of governmental reform that were sweeping initiatives during this period, powerful institutional structures were created. These new institutions developed strict rules about how roads can be used and who had priority. Their mission was to take an asset of the public, city roads, and do what was necessary to facilitate travel by car. The efforts of this institutionalization sparked traffic signals, speed limits, and designated spaces for parking. All seem selfevident in hindsight. But in the early decades of last century, these innovations dramatically expanded how public space was regulated. The populous accepted these changes because motorized transport provided advantages that were previously, inconceivable.

Pursuing the aim of auto mobility for everyone invoked a cascade of effects, including but not limited to measures for their performance, ways to classify them, and standards to reinforce each. To this day, one of the auto industry's most celebrated victories came when, in 1924, the National Conference on Road and Highway Safety, convened by President Hoover, agreed to adopt the National Automobile Chamber of Commerce's preferred safety metric: fatalities per vehicle.

Strict classification systems emerged. In urban areas (at least in the US), roads were assigned to one of handful of categories: interstates, freeways and expressways, principal/minor arterials, primary/secondary collectors, and locals. What differentiated one category from the others had to do exclusively with how easy it would be, by automobile, to actually get to a destination. For example, principal arterials existed to move cars along the network and little else. Local roads, similarly responsible for moving cars through the network, also carried an additional burden. They actually allowed travelers to get within striking distance of a destination by facilitating turns and stops through curb cuts, driveways or allowing parking. The system created artificial inverse relationship between mobility function and access function of roads. Scant consideration was given to any other mode as this modern road hierarchy took form. Exclusive was concern to moving motorized traffic.

Two central criteria were enacted to measure how these roads functioned: volume and speed of motorized vehicles. Which roads were to carry how much traffic was an outgrowth of forecasts that were based on models in which it was assumed that an automobile was taken for every trip. Given that the auto needed a place to park at the end of the trip, minimum parking standards were installed in land use codes, thereby reinforcing the demand for other infrastructure requirements.

Given vehicular throughput was heralded as the central criteria for success, there was (and still is) little tolerance for any delay caused from traffic. Congestion jeopardized the free-flow of vehicles and such instances were viewed as weaknesses in the system. Given that the proverbial (infrastructure) tap was already open, more space was added to the network to satisfy the increase vehicular demand. The mobility system was deemed to be functioning well as long as cars are moving through the network of roads, preferably at fast speeds to minimize travel times. Even additional standards were introduced to ease free-flow speeds; these included regiments

prescribing the radii of curbs at intersections, the width of travel lanes, and aggressive signage. After all, public welfare was at issue and all roads needed to be able to easily accommodate quick response movement from a 35 foot long fire truck. With each mounting year, an unwavering mindset was reinforced. Additional legislation was added and even more minute standards were installed. Habits were formed and more roads with the same feel were stamped out by cities. Even older roads, those built in the first half of the 21st century, eventually became beholden to the standards aimed to make driving easier, such as through automatic road widening (Manville, 2016). This happened so slowly, the majority of the people didn't even notice.

Progressions

Such practices were not completely absent of critics. City intellectuals such as Lewis Mumford and Jane Jacobs, in the early 1960s, warned against continued proliferation of such auto centric ways. The plight of their claims, however, was largely steamrolled over, literally. Machines won in the end because transport and road building practices had already gained an overwhelming inertia. Cultural expectations that fundamentally allowed most roads to be hostage to a single mode were reinforced. Combined, these factors fueled a sense of path dependency that continues to this day. From these origins, cities have largely accepted that when they manage vehicular traffic they have been managing the right thing. The rise of auto mobility was so sudden, and so quietly supported by many public decisions (traffic engineering, parking requirements, paved roads, etc.) that decades of undisputed dominance for the motor vehicle has gained its own sense of inertia and path dependency.

The sirens forecasted by the likes of Mumford and Jacobs, however, did not completely fall on deaf ears. They warned that it was necessary to guard against the perils of everything cars and their thoughts helped sew a few seeds. Their thinking helped draw attention to how, by making travel easier for cars, it made travel by any other mode more difficult. They argued how by prioritizing mechanized travel, cities relegated human powered travel to a novelty act that eventually, would eventually have detrimental consequences. Bringing to light these arguments helped, at least, provide seeds that they sewed to grow slightly alternative thinking. When one thing is prioritized, another is diminished. This is a simple function of allocating scarce space.

Performance Measures

One focus of the assault targeted traffic engineers. Specifically, performance measures that guided their day-to-day work and Level of Service (LOS) standards were low hanging fruit. Not only did LOS measures aim to employ an objective formula to answer a subjective question—how much congestion people are willing to accept—these measures only gauged the success of roads through the perspective—or windshield—of car use (Dumbaugh et al., 2014). Some progressive engineers helped city leaders understand how a LOS score of C or D could strike a balance between over and underinvestment of the infrastructure needed. They, however, were in the minority. A few communities are slowly moving away from LOS guidelines, after decades arguing against their overall suitability, but what is replacing them is still far from perfect and still prioritizes vehicular travel.

One option specifies the success of a road by measuring how many vehicle miles it carries. A different alternative measures success based on the throughput of persons, regardless of the mode travel. Prioritizing the flow of people, by measuring person-trips (and not vehicular trips) is a step in a new direction. By definition, it represents a more humane goal because it accounts for the flow of individuals, regardless of their means of travel. And, by facilitating the flow of person trips regardless of mode, it has concomitantly helped support new road design approaches, such as complete roads (Hui et al., 2018) and processes of managing curb space as a scarce resource (King and Saldarriaga, 2018).

Following years of struggle, new guidelines are now available to design roads—guidelines which do not have the automobile at center of everything. Some progressive cities are adopting them. These guidelines have been undergoing underground transformative updates to better accommodate more diversity of transport modes and new design guidelines for the roads. For example, the American Association of State Highway and Transportation Officials (e.g., AASHTO and National Association of City Transportation Officials (Vega-Barachowitz, 2013)) for roads continue to offer tweaks that grant a stronger role to other human-scaled vehicles such as bicycles. However, these guidelines—progressive they might be—still subscribe to measuring throughput. They run up against an inherent limitation in they still principally gauge the success of a road based on how much activity is going through the corridor. They fall down in offering a sought ending place to value roads as anything *other* than a conduit for movement (Jacobs, 1993). Approaches to primarily legislate space in future roads based on particular types of modes are misguided; they fail to capture the essence of what many future roads in cities could or should be used for.

Owing in large part to the histories traced thus far in this chapter, changing the Overton window of what is possible to project new performance measure is difficult, to put it mildly. This difficulty is compounded because residents have a strong bias toward the status quo, even when the use of the status quo helps kill nearly 40,000 Americans each year. People also experience the city in perception and reality as motorists. Taking something away from drivers seems a personal attack for many.

Human-scaled Access

While throughput of anything will inevitably be one important element to consider for roads of the future cities, throughput by itself misses the mark. It fails to help understand other and more important dimensions. We contend that in deliberating future

performance criteria for roads, one fundamental characteristic stands out: how well are modes that support human-scaled access supported. The best way to do that is to moderate how fast things are moving, whatever that thing is. And, if the goal of preserving human-scaled access of roads through moderating speed can be overcome, other benefits will fall out.

The safety of movement on any road is predominantly affected by the size and speed by which things are moving. Both interact through laws of physics. Things that are moving faster gain more force and can do more damage than things moving more slowly. The size part of the equation comes in the form of cage to protect the user, usually of the metal armor variety, characteristic of cars, trucks, buses and the sort. Bigger and heavier objects, as opposed to lighter objects, compound the force, therefore increasing the differential in both the speed and size of transport users on the road. This is owed to the fact that if a collision happens, properties of kinetic energy take over.

The City of London rolled some of this logic out when Transport for London implemented new road guidelines. They described, for example, how a cyclist's fate is severely compromised when competing with vehicles that are faster or are superior in size and weight. Worse yet, both. A cyclist hit by a car hurtling at 40 miles/h results in a 70% chance that the cyclist will be killed. Should the cyclist be struck by a car moving 30 miles/h, there is an 80% chance they survive. A mere 10 miles/h can separate life from death.

Given the Cambrian explosion of vehicular types that are increasingly littering public rights-of-ways in cities, future roads will be well served to support varied transport modes. This raises fundamental questions around which decision makers need to come to grips. What defines a vehicle? Size, power, or shape? Some road users will inevitably be more vulnerable than others and it is a difficult problem to balance the needs of different types of users. The most vulnerable ones will be those lacking speed and size but that does not mean they don't belong. Moderating speed, regardless of mode provides a straightforward and robust target. Lowering speeds is the best, easiest, and fastest way to quickly radically improve the composition for roads, for both drivers and anyone in front of them.

Re-boot

A complete re-boot is needed to measure the success of roads such as locals, collectors, and likely, some arterials. Efforts to recalibrate an old needle will not cut it as they are, old. They fail to directly account for inevitable conflict among the increasing modal heterogeneity that will populate roads. They fail to capture the quality and experience of the movement. Safety and speed, working together, do.

Road planning processes from Northern Europe have been tending to such issues for decades. Localities from these contexts don't have all the answers, but they are further along than US counterparts in thinking through issues and identifying trouble spots. Take, for instance, Vision Zero, the road safety campaign that is sweeping the globe. Hailing from Sweden, Vision Zero aims for no deaths or serious injuries involving road traffic and use. It doesn't care what type of vehicle is at issue ([Johansson, 2009](#)). For US cities, where adopted, Vision Zero is perhaps the most compelling initiative to likely affect the actions of municipal transport departments and design of roads since that original 1924, performance measure was introduced.

In another example, in select corridors in some communities in Netherlands, long considered a leader in progressive road design, road planning efforts are no longer resorting to the traditional strategy of separating modes. Instead, they convert the whole road to all forms of movement based on a design speed. In so doing, they are removing barriers that help protect vulnerable road users, not adding them. In corridors with substantial bicycle traffic, other vehicles such as cars are permitted, but they do so on the terms delineated by the bike. Bringing multiple types of vehicles that are using road space into harmony with one another is the most important criteria defining successful streets.

A subtheme of roads that are built around the concept of speed, and not mode, is that it engenders a sense of social policing, where drivers are naturally cautious and people walking or biking feel safe. Standards for behavior are more firmly established and behavior that deviates from these standards adds a selfpolicing effect. The second effective strategy relies on enforcement, most often through technology. But focusing on enforcement can introduce new concerns about bias and profiteering in policing, which undermines political support but can also have devastating effects on individual freedom and liberty ([Seo, 2019](#)).

Based on the above rationale, we see a new breed of road emerging to fill a void in future transport planning. This new breed would be one that allows a road to be viewed both as a passageway and also as a form of connective tissue to socially unify residents. Both are advanced by using speed—as paramount considerations. This breed of roads falls outside the realms of traditional shopping roads, classic high roads or commercial corridors. It is also outside the bounds of most residential roads as they too separate play from movement. It is the type of road that is a neighborhood asset and amenity, supportive of a variety of users engaging in low-speed activities. It is premised on an ideology of using roads as public spaces for social interaction, commerce, play and also, movement.

There is risk in being over-prescriptive. Many cities installed pedestrianized roads in the 1960s and 70s which have never lived up to their promise (see Fulton Mall in Fresno, California as one example) ([Levinson and Krizek, 2018](#)). Other examples have worked. A one-size-fits-all recipe is too much to ask for. This admittedly makes it hard for planners and action-oriented officials. Trial and error is not a convincing policy platform and it certainly is not in the DNA of government. While private firms fail all the time (see recent bike share companies, for instance), it is important to safeguard against the public doing so. A city that tries and fails will be branded as wasteful, not enterprising. Futures that are inevitably different from the current are more difficult to plan for, especially under short time horizons (e.g., next year) and murky legal standing. While the legal legitimacy of road changes in beyond the scope of this chapter, any changes will have to be validated and upheld in court.

Concluding Thoughts

Motivated by the needs of planners and local officials to more quickly respond to a rapidly changing transport landscape, this chapter first helps understand why change is so hard. It provides context to allow readers to allow them of the formidable challenges facing those who shape roads—the performance measures, standards, and rules.

A troubling issue is that cultural and institutional practices regarding the nature of roads have become so sclerotic. Entrenched bureaucratic interests are geared toward preserving the status quo, thereby locking out society's ability to change them. Based on the evolving climate of micromobility, we argue that a new foundation is needed on which to build current-day and future policy. This provides future cities the opportunity to evolve to a more durable approach for future infrastructure that helps protect the public realm.

Innumerable obstacles to nonvehicular mobility exist in how roads are assessed, the standards guiding their design and the laws governing their use; each uphold car supremacy (Shill, 2019). This means that other social imperatives for cities such as the preservation of life, livability, and economic vitality fall to the wayside. One goal here was to identify the problem: sclerotic road planning. Leaders lack a way to intelligibly talk about this regime and knowing what needs to change is a first hurdle.

A second goal was to offer something that could or should replace it. Measuring performance by the relative speed and volume of cars is not only outdated, it is also counter-productive. Human-scaled accesses, regardless of mode, are two important characteristics for future roads to shoot for. The institutions that govern and manage roads are lacking criteria to reliably reinforce these metrics. This will inevitably involve new governance and policies that affect road use. Elected officials and the city planners that advise them deserve a more reliable north star to work towards. Managing speed helps provide such. Overcoming the automobile-centric nature of roads requires brutal realism—about what they could or should stand for and more pointedly, the sclerotic regulations, processes and mindsets that have guided their use. What can work for the future urban roads centers itself on what is important for cities to thrive: human-scaled access to services.

References

- Dumbaugh, E., Tumlin, J., Marshall, W.E., 2014. Decisions, values, and data: understanding bias in transportation performance measures. *ITE J.* 84 (8), 20.
- Hui, N., Saxe, S., Roorda, M., Hess, P., Miller, E.J., 2018. Measuring the completeness of complete roads. *Transp. Rev.* 38 (1), 73–95.
- Jacobs, A., 1993. *Great Roads*. MIT Press.
- Jacobs, Jane, 1961. *The Death and Life of Great American Cities*. Random House.
- Johansson, R., 2009. Vision Zero—Implementing a policy for traffic safety. *Saf. Sci.* 47 (6), 826–831.
- King, D.A., Saldarriaga, J.F., 2018. Measuring changes in taxi trips near infill development and issues for curbside management of for-hire vehicles. *Res. Transp. Bus. Manag.* 29, 93–100.
- Levinson, D.M., Krizek, K.J., 2018. *Metropolitan Transport and Land Use: Planning for Place and Plexus*. Routledge.
- Manville, M., 2016. Automatic road widening: Evidence from a highway dedication law. *J. Transp. Land Use* 10. (1).
- Mumford, L., 1964. *The Highway and the City..* New American Library.
- Reid, C., 2015. *Roads Were Not Built for Cars: How Cyclists Were the First to Push For Good Roads and Became the Pioneers of Motoring*. Island Press.
- Seo, S.A., 2019. *Policing the Open Road: How Cars Transformed American Freedom*. Harvard University Press.
- Shill, G.H., Should Law Subsidize Driving? (2019). 95 *New York University Law Review* (2020 Forthcoming); U Iowa Legal Studies Research Paper No. 2019-03.
- Vega-Barachowitz, D., 2013. Changing the DNA of City Roads: NACTO's Urban Road Design Guide and the New City Road Design Paradigm. *ITE J.* 83 (12), 36.

Further Reading

- Hoehne, C.G., Chester, M.V., Fraser, A.M., King, D.A., 2019. Valley of the sun-drenched parking space: The growth, extent, and implications of parking infrastructure in Phoenix. *Cities* 89, 186–198.
- Jones, P., Boujenko, N., Marshall, S., 2007. *Link & Place-A guide to road planning and design*.
- Marshall, S., 2005. *Roads and Patterns*. Spon Press, London and New York.

Road Traffic Regulation: Road Pricing and Environmental Quality

Marco Percoco, Department of Social and Political Sciences and GREEN Università Bocconi Milano, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	378
The London Congestion Charge as a Prototype	378
How to Evaluate the Impact of Road Pricing on Pollution	379
The Impact of Road Pricing on Environmental Quality: The Case of London	380
Conclusion	382
Acknowledgment	383
References	383
Further Reading	383

Introduction

Climate change, the health burden of pollution and economic cost of congestion are pushing policy makers to adopt measures to contain such externalities. Road traffic is widely considered as one of the main sources of externalities in urban areas, especially in developed countries, where manufacturing activities have re-located in the outskirts of cities. In 2013, congestion cost \$8.5 billion and 82 hours per driver in London. These figures are expected to rise by 63% in 2030 (Gordon and Pickard, 2014). Similar costs are faced by most of the world cities. According to Cohen et al. (2004), urban pollution causes up to 6.4 million premature deaths every year. Given this empirical evidence, policy makers are implementing measures at local level to decrease the concentration of some pollutants; in particular, through transport policy actions (OECD, 2010; Greater London Authority, 2006).

The aforementioned evidence is the empirical background for a vast series of measures designed to contain traffic and the use of private cars more specifically. Parking restrictions and pricing, expansion of the supply of public services, road pricing are all examples of traffic restrictions. To cope with the external costs of transport, several cities have introduced or are considering to introduce road pricing schemes, as in the case of: London (Banister, 2003), Milan (Percoco, 2013; 2014a; Rotaris et al., 2010), Hong Kong (Ison and Rye, 2005), Singapore (Santos, 2005), Stockholm (Eliasson et al., 2009), and several Norwegian cities (Ieromonachou et al., 2006).

In the case of London, pollution and congestion were considered to be the reasons that led to the then Mayor of London, Ken Livingstone, overseeing the implementation of the London Congestion Charge (henceforth denoted as LCC). The LCC, introduced in 2003 and then modified to extend the treated area, is probably the most known and studied example (Banister, 2003; Givoni, 2012; Ison and Rye, 2005; Prud'homme and Bocarejo, 2005; Quddus et al., 2007; Santos and Bhakar, 2006; Santos and Fraser, 2006; Santos and Shaffer, 2004). However, literature has not reached a consensus on the socio-economic convenience of such measures, since infrastructure and administrative costs seem to exceed the benefits in terms of a reduction in external costs (Mackie, 2005; Prud'homme and Bocarejo, 2005; Raux, 2005).

In this chapter, we will devote particular attention to road pricing schemes in terms of cordon pricing as a growing number of cities have introduced (or are planning to implement) such policy to curb externalities generated by urban traffic. Furthermore, we will devote special attention to the London Congestion Charge (henceforth, LCC), one of the most notable attempts to restrict traffic in a given area by using a pricing scheme. It should be noted that the effects of the LCC on atmospheric pollution is not clear a priori and it highly depends on the behavioral response of road users. On one hand, road pricing should reduce traffic volumes, given demand elasticity to changes in transport cost and this should result in an overall improvement in pollution concentration. On the other hand, the reduction in the number of cars reduces congestion, by increasing travel speed and this, in its turn, will result in an increase in fuel consumption and in a deterioration of air quality. Givoni (2012) has in fact argued in favor of a more robust statistical analysis of the effects of road pricing experiences. This is because figures used in ex post evaluations are, in general, unreliable and biased by other phenomena (confounding factors) not considered in the analysis.

The London Congestion Charge as a Prototype

London's fight against pollution has its origins in the second half of the XIX century; although it was only after the Great Smog of 1952 that the first policies were introduced, aiming to improve air quality. This event was disastrous for Londoners: a huge blanket of smog covered the entire city for four days and, by some estimates, caused the deaths of 4000 people (Mayor of London, 2002). Since then, air quality has much improved; although it remains one of the European cities with the highest levels of pollution. Road transport is a major cause of the high concentration of pollutants, accounting for about 40% of emissions of nitrogen oxides and more than 60% of particulates (Greater London Authority, 2006; Kelly et al., 2011).

In an attempt to reduce traffic flow, the LCC was introduced on February 17, 2003. The objective was to reduce congestion in the central area of London, covering an area of 22 sq. km, or 1.4% of the territory of Greater London. LCC consists of a daily payment to obtain permission to move freely in the area. The policy's enforcement is achieved through the use of cameras and the automatic recognition of cars' license plates. The charge is in operation from Monday to Friday, from 7:00 a.m. to 6:00 p.m. Initially, the scheme provided for a daily fee of £5. Subsequently, this increased to the current rate of £10. Exemptions are provided for the means of public utility, such as buses, vehicles of law enforcement, and for all vehicles powered by alternative sources of fuel. Finally, for the vehicles of the residents in the area, the price is discounted by 90%.

The stated objective of implementing the congestion charge was a 15% reduction in traffic in the central areas of London, with a simultaneous maintenance of traffic levels in the surrounding affected area. Transport for London (TfL, 2006) reports that in the first years of the scheme's operation, the number of cars entering the central area of London significantly decreased by 21%—with respect to the pre-treatment period, although with no significant changes in travel time. However, when considering the effect of the policy at the level of congestion, the results are less positive. In regards to congestion considered as excess delay, above conditions not congestion; in the first years of the policy's operation, there was a substantial reduction in the level of congestion in the order of 30%. However, since 2006; despite the previously mentioned reduction in the number of vehicles, the congestion level has increased, compared to the levels prior to the policy. The aim of the congestion charge was largely two-fold (TfL, 2004). First, to reduce congestion and second, to use the funds raised to improve transport infrastructure. In so far as private consumption of motor transport can be seen to impose negative externalities, for example, increased congestion, noise, pollution, etc., LCC can be thought of as a form of Pigouvian taxation. It can better equate the marginal private and social costs of transport; that is, to make individual agents incorporate external costs of their consumption into their private costs.

However, it should be noted that the marginal cost of congestion was not used as a basis for the charge. Instead, to find the optimal pricing scheme, simulations and models of household behavior were used to predict changes in traffic. Nonetheless, Santos and Shaffer (2004) state that the £5 per day charge is a reasonable approximation of the marginal congestion costs for an agent driving through the congestion charge zone.

How to Evaluate the Impact of Road Pricing on Pollution

The most promising empirical approach to evaluate the effects of road pricing is based on fairly recent literature using the Regression Discontinuity Design (henceforth denoted as RDD) to examine the impact of policies related to transport and air quality (Chen and Whalley, 2012; Davis, 2008; Percoco, 2013).

RDD is a non-experimental approach that uses ex post to evaluate a program's impact on a situation in which units are considered treated or not, according to a certain threshold in a reference variable (forcing variable). In our case, the date in which the traffic restriction was implemented is used as a threshold that introduces an exogenous variation in the access of polluting vehicles in the city. Thus, the expected outcome is a reduction in the level of pollutants. It should be stated that RDD identifies the impact of traffic restriction under mild assumptions; hence, it excludes the bias imposed by confounding factors.

Let y_0 and y_1 denote the counterfactual outcomes before and after the treatment T (the introduction of the policy), let x be the forcing variable (in our case, the time) and consider the following assumptions (Angrist and Pischke, 2009):

$$A1. E(y_g | T, x) = E(y_g | x), g = 0, 1$$

$$A2. E(y_g | x), g = 0, 1 \text{ is continuous at } x = x_0$$

$$A3. P(T = 1 | x) \equiv F(x) \text{ is discontinuous at } x = x_0, \text{ that is, the propensity score of the treatment has a discrete jump at } x = x_0.$$

Following Imbens and Lemieux (2008) the goal is to estimate the parameter ρ on treatment of this form (for the moment, we do not specify the model across locations):

$$y_{i,T} = \theta + \rho \text{Restriction}_i + f(\tilde{x}_{i,T}) + \eta_i \quad (1)$$

where $y_{i,T}$ in our case is the concentration of a given pollutant in day t whose treatment status is T (i.e., before or after the introduction of the restriction), θ is a constant, $\tilde{x}_{i,T}$ is the forcing variable properly normalized (a time trend centered at the date of the introduction of the policy, i.e., 17 February 2003). Consequently, ρ expresses the impact of the treatment at $x_{i,T} = x_0$. The $f(\tilde{x}_{i,T})$ term is a p -th order parametric polynomial to account for non linearity of the relationship between the time trend and pollution and thus to control that the eventual break in $x_{i,T} = x_0$ is not due to unaccounted non-linearity. Lastly η_i is an error term. *Restriction* is our treatment variable taking the value of 1 after the introduction of the congestion charge and zero before.

Seasonal and climatic factors are crucial in explaining the level of pollutants in the air. To deal with these problems in the reference model Eq. (1), seasonality is accounted for with day of the week, month and year dummies.

To make Eq. (1) operational for the analysis of the impact of the LCC and to account for the heterogeneity of monitoring stations, we have estimated an equation in the form:

$$y_{it} = \alpha_i + \rho LCC_t + \gamma LCC_t \cdot Treated_i + f(x_i) + \varepsilon_i \quad (2)$$

where α_i is a full set of station-specific fixed effects and $Treated_i$ takes the value of 1 if station i is located in the charged area and

Table 1 Descriptive statistics.

	<i>Whole sample</i>	<i>Treated area</i>		
	(1)	(2)	(3)	(4)
	Before	After	Before	After
PM10	21.49 (9.874)	40.24 (16.48)	22.26 (7.914)	40.95 (15.94)
O ₃	29.32 (15.16)	34.94 (19.86)	22.74 (12.37)	29.16 (18.10)
CO	0.387 (0.268)	0.390 (0.250)	0.468 (0.188)	0.495 (0.240)
NOX	74.82 (57.06)	88.11 (59.68)	76.30 (46.67)	99.03 (64.21)
SO ₂	7.033 (5.372)	9.670 (8.209)	4.553 (3.406)	7.211 (5.418)

zero otherwise. It should be stated that station-specific fixed effects are of particular relevance in Eq. (2) to identify the policy parameter γ as *Treated_i* is time-invariant. In all the specifications we make use of a 5th order trend polynomial and control for weather conditions and standard errors were clustered by month in order to account for possible spatial correlation. Seven days of temporal lags of the dependent variable are also used to account for the temporal persistence of pollutants, especially in the case of particulate matter. Finally, model Eq. (2) is estimated in logarithms, so that parameter estimates can be interpreted as percentage changes.

In Eq. (2), parameter γ identifies the effect of the policy in the charged area (the Average Treatment on the Treated), whereas ρ indicates the average effect of the introduction of the London Congestion Charge across stations, regardless of the their location.

The Impact of Road Pricing on Environmental Quality: The Case of London

To evaluate the case of London by using the empirical methodology described in the previous section, it is possible to use the data used in the analysis were made available by the LAQN (London Air Quality Network). They comprise daily observations of pollution from several monitoring stations in London, over period January–March 2003. The dataset contains information on the concentration of five pollutants: PM10, O₃, CO, NOX, SO₂. Of those pollutants, only SO₂ is less related to transportation. The concentrations of all the others are widely considered to be indicators of transport-related pollution (although not exclusively). Furthermore, information on: temperature, wind speed, rain, and humidity is also available^a. Table 1 shows a mean concentration before and after the introduction of the LCC. Interestingly, the results show a sizeable increase in the emission levels for almost all of the pollutants considered, also within the area where the congestion charge was implemented.

Results of the descriptive analysis in Table 1 are admittedly puzzling and need to be scrutinized in a more systematic way, through a parametric analysis. In Table 2, baseline models for the evaluation of the LCC are reported. They have been all estimated over the period January–March 2003 with a 5th order polynomial trend; weather controls and seven days of temporal lags of the dependent variable. In Panel A, we report results of the difference-in-discontinuity model in Eq. (2). A slight but only marginally significant increase in the concentration of PM10 is detected for the whole city and a significant reduction in the treated area is estimated in -0.059% . As for ozone, a reduction of an order of magnitude of 0.394% is found, with no significant spatial differentiation. Interestingly, as for CO, NOX, and SO₂ a significant increase in the concentration of pollution is detected, with a contemporary reduction in the treated area, although this decrease is significant only in the case of NOX. In Panel B, we consider only monitoring stations out of the treated area but located within 5 km from the boundary of the charged area. Therefore, the estimated model assumes $\gamma = 0$. Results confirm the overall reduction in the concentration of ozone, but also significant increases in PM10, CO, NOX, SO₂ with parameters ranging from 0.106% as in the case of PM10 to 0.513% for SO₂ (although, only marginally significant).

Finally, in Panel C, we have excluded monitoring stations located in the treated area and in the surrounding area within a 5 km arc distance. Therefore, we have considered only monitoring stations in the outer circle within a 10 km from the treated area. Also in this case results are confirmed: an overall decrease in O₃ equal to -0.277% and an increase in PM10, CO, NOX, and SO₂.

Overall, econometric results show unclear effects of the congestion charge in London. For some pollutants, such as PM10, CO, NOX, and SO₂, an increase in the concentration was found out of the treated area, with only small reductions in the charged area. This can be due to the diversion of traffic from the city center to external areas, with a subsequent increase in the kilometers traveled and the potential of polluting emissions.

^a It should be mentioned that the complete dataset covers the years 2000–2013. However, in this paper, we have restricted the sample to only three months (January–March 2003) to make our regression analysis comply with general practice of RDD consisting in restricting the interval around the threshold. An analysis over longer period was conducted with results qualitatively similar, although less reliable. Results are available upon request.

Table 2 The effect of the London Congestion Charge on air quality.

	(1)	(2)	(3)	(4)	(5)
	PM10	O3	CO	NOX	SO2
<i>Panel A: Whole sample</i>					
LCC	0.0669*	−0.394***	0.463***	0.423***	0.472***
	(0.0337)	(0.0773)	(0.0875)	(0.0633)	(0.117)
LCC x Treated	−0.0592***	0.00397	−0.0219	−0.0451***	−0.0186
	(0.0172)	(0.0145)	(0.0344)	(0.0129)	(0.0406)
Observations	4,149	1,827	1,691	5,126	2,139
R-squared	0.620	0.329	0.306	0.397	0.366
Number of st.	59	27	28	72	35
<i>Panel B: Excluding treated area, within 5 km</i>					
LCC	0.106**	−0.491**	0.424***	0.416***	0.513*
	(0.0344)	(0.0690)	(0.0296)	(0.0314)	(0.187)
Observations	1,016	340	487	1,242	604
R-squared	0.622	0.380	0.377	0.402	0.418
Number of st.	14	4	6	16	9
<i>Panel C: Excluding treated area, more than 5 km</i>					
LCC	0.0768*	−0.277**	0.646**	0.484***	0.540***
	(0.0336)	(0.101)	(0.218)	(0.0985)	(0.134)
Observations	2,896	1,183	1,029	3,672	1,262
R-squared	0.623	0.313	0.311	0.397	0.365
Number of st.	39	15	14	48	18

Notes: Standard errors are clustered by month. Significance levels:

*** $P < 0.01$

** $P < 0.05$

* $P < 0.1$.

Table 3 Difference-in-differences in traffic within London (dependent variable is the number of vehicles; in thousands).

	(1) Descriptive statistics	(2) Least squares	(3) Least squares
Treated	−128.538 (123.444)	−12.358 (90.738)	77.973 (83.302)
Surrounding	111.325*** (23.332)		279.596*** (89.434)
Control	−207.296 (231.211)		
Observations		12,846	12,846
R-squared		0.094	0.094

Notes: Standard errors in models 2 and 3 are clustered by local authority. Significance levels:

*** $P < 0.01$, ** $P < 0.05$, * $P < 0.1$.

In particular, the Department for Transport makes traffic counts for the period 2000–13 available, with annual observations for 2141 count points. Count points have been geolocalized and hence assigned to three groups: *Treated* (if located in the congestion charge area), *Surrounding* (if located in a borough partially treated by the congestion charge or neighboring the charged area^b), or in the control group. In column 1 in **Table 3**, descriptive statistics of differences-in-mean is reported. In particular, statistics refer to the change in the mean of traffic counts before the treatment and after the introduction of the charge. It is noted that the pre-treatment mean is computed over the years 2000–02, whilst the post-treatment mean is calculated over the years 2003–05. Descriptive statistics show a decrease in the number of vehicles by 128,538 and 207,296 in the treated and controlled areas, although both estimates are not significantly different from zero. Interestingly, count points surrounding the treated area registered an average increase by 111,325 vehicles. This estimate is significant at a 99% statistical level. Together, these statistics imply a diversion of traffic from the treated to the surrounding area. However, a compelling parametric analysis is needed to account for unobserved heterogeneity and

^bThe following boroughs have been considered to be in the Surrounding group: Wandsworth, Lewisham, Greenwich, Newham, Waltham, Haringey, Barnet, Brent, Kensington, Hammersmith.

Table 4 The effect of the London Congestion Charge on traffic composition (dependent variable is the number of vehicles by type; in thousands).

	(1) <i>Heavy goods vehicles</i>	(2) <i>Light goods vans</i>	(3) <i>Cars</i>	(4) <i>Motorbikes</i>	(5) <i>Bikes</i>
Treated	-13.434* (7.121)	7.002 (15.401)	41.967 (77.960)	34.664*** (11.123)	30.335*** (8.338)
Surrounding	-3.952 (9.541)	31.497 (20.818)	242.441** (101.360)	3.592 (8.988)	-.479 (12.804)
Observations	12,846	12,846	12,846	12,846	12,846
R-squared	0.012	0.012	0.013	0.016	0.014

Notes: Standard errors are clustered by local authority. Significance levels:

*** $P < 0.01$

** $P < 0.05$

* $P < 0.1$.

time trend. To this end, the following difference-in-difference model can be estimated over the years 2000-2005:

$$traffic_{it} = \alpha_i + \beta trend_t + \gamma post_t + \delta_1 Treated_i \cdot LCC_t + \delta_2 Surround_i + ng_i \cdot LCC_t + \varepsilon_{it} \quad (3)$$

where the dependent variable is the number of traffic count in year t at count point i , α_i are count point specific fixed effects, $trend$ indicates a temporal trend, $post$ is a dummy variable taking the value of one after 2002 and zero otherwise, $Treated_i$ and $Surround_i$ are indicator variables for the treatment and surrounding group to which count point i belongs to. In model (2) in [Table 3](#) we consider $\delta_2 = 0$ and it emerges a decrease by 12,358 vehicles in the treated area with respect to the rest of the city, although this coefficient is not significant. In model (3) the full model is reported and it emerges that the introduction of the congestion charge has resulted in an increase in traffic counts by 279,596 vehicles. This figure is statistically significant, whereas no significant (at conventional level) change is found in the treated area.

Estimates in [Table 3](#) corroborate the hypothesis advanced previously to justify the finding of a decrease in pollution concentration in the treated area and an increase outside of this area. However, results may still hide some heterogeneity in terms of traffic composition. In [Table 4](#), the sample is split into four categories (heavy goods vehicles, light goods vehicles, cars, motorbikes) and the number of bikes is added. Interestingly, the introduction of the LCC has decreased the number of heavy goods vehicles by 13,434 units in the treated area, although this estimate is only marginally significant. Additionally, no significant effect is found in the case of light goods vans; whilst, in model 3, a clear increase in the surrounding area by 242,411 cars is estimated. According to estimates in models 4 and 5, an increase by 34,664 and 30,335 occurred in the case of motorbikes and bikes. Hence, from an environmental perspective, a shift towards uncharged vehicles (and in this case, motorbikes) is not efficient. Another possible transmission channel might make a change in the kilometers traveled; drivers may be willing to avoid the charged area by traveling longer routes around the area. Finally, it should be noted that these simple results are vastly consistent with the results obtained in the case of Milan ([Percoco, 2013](#)).

Conclusion

In this chapter, the environmental effects of road pricing schemes have been discussed with specific reference to the London case. The exogenous variation in traffic flows after the introduction of the policies was used to estimate a variation into the concentration of: PM10, O3, CO, NOX, SO₂. A negligible effect of the policy was found, along with a spatial displacement effect, since a reduction in the concentration of several pollutants in the treated area and an increase in the surrounding areas was found. In particular, a significant decrease in the concentration of O3 was found across the whole city of an order of magnitude of -0.4% (on average). CO, NOX, and SO₂ present significant increases by 0.4%–0.5% (on average) in the area surrounding the charged area. A reduction in the concentration of PM10 and NOX was also found in the charged area. Overall, in terms of pollution concentration of the whole city, it seems that congestion charges have limited or even adverse impact, possibly because of the spatial diversion of traffic. This result calls for careful consideration of the spatial extent and of the cross elasticity of traffic flows when implementing road pricing schemes.

An important factor for the interpretation of the results is the relationship between emissions of a given pollutant and average speed. In fact, the relationship between the production of nitrogen oxides and traveling speed is directly proportional to the vehicles with diesel engine ([EEA, 2010](#)). Thus, an increase of the vehicle's average speed is due to a greater production of nitrogen oxides. This might be relevant because of the reduction of congestion caused by the introduction of the charge, especially in the early years after the policy was introduced ([Greater London Authority, 2006](#)). However, the relationship between pollutants and average speed of vehicles is still not very clear, especially for ozone and particulates, and further studies are needed to shed light on this issue ([EEA, 2010](#)).

Acknowledgment

The author wishes to thank Sergej Gubins, Jos Van Ommeren, participants in the NECTAR Conference in Ann Arbor and in seminars at Università Bocconi and Università di Bergamo for useful comments on an earlier draft.

References

- Angrist, J., Pischke, J.S., 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton.
- Banister, D., 2003. Critical pragmatism and congestion charging in London. *Int. Soc. Sci. J.* 176, 248–264.
- Chen, Y., Whalley, A., 2012. Green infrastructure: the effects of urban rail transit on air quality. *Am. Econ. J. Econ. Policy* 4 (1), 58–97.
- Cohen, A., Andersen, R., Ostra, B., Pandey, K., Kryzanowsky, M., Kunzli, N., Gutschmidt, K., Pope, A., Romieu, I., Samet, J., Smith, K., 2004. The global burden of disease due to outdoor air pollution. *J. Toxicol. Environ. Health* 68, 1–7.
- Davis, L.W., 2008. The effect of driving restrictions on air quality in Mexico city. *J. Pol. Econ.* 116 (1), 38–81.
- EEA, 2010. *Do Lower Speed Limits on Motorways Reduce Fuel Consumption and Pollutant Emissions?* European Environment Agency, Copenhagen.
- Eliasson, J., Hultkrantz, L., Nerhagen, L., Smidfelt Rosqvist, L., 2009. The Stockholm congestion charging trial 2006: overview of the effects. *Transport. Res. A* 43 (3), 240–250.
- Givoni, M., 2012. Re-assessing the results of the London congestion charging scheme. *Urban Studies* 49 (5), 1089–1105.
- Gordon, S., Pickard, J., 2014. London's \$8.5bn traffic jam slows down growth, *Financial Times*, October 13, online edition.
- Greater London Authority, 2006. *50 Years On: The Struggle for Air Quality in London Since the Great Smog of December 1952*. Greater London Authority, London.
- Ieromonachou, P., Potter, S., Warren, J.P., 2006. Norway's urban toll rings: evolving towards congestion charging. *Transp. Policy* 13 (5), 29–40.
- Imbens, G.W., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. *J. Econ.* 142 (2), 615–635.
- Ison, S., Rye, T., 2005. Implementing road user charging: the lessons learnt from Hong Kong, Cambridge and Central London. *Transp. Rev.* 25 (4), 451–465.
- Kelly, F., Anderson, H.R., Armstrong, B., Atkinson, R., Barratt, B., Beevers, S., Derwent, D., Green, D., Mudway, I., Wilkinson, P., 2011. *The Impact of Congestion Charging Scheme on Air Quality in London*. Health Effects Institute, Research Report no. 155.
- Mackie, P., 2005. The London congestion charge: a tentative economic appraisal – a comment on the paper by Prud'homme and Bocarejo. *Transp. Policy* 12 (3), 288–290.
- Mayor of London, 2002. Statement by the Mayor concerning his decision to confirm the Central London Congestion Charging scheme order with modifications. Office of the Mayor of London, London.
- OECD, 2010. *The transport system can contribute to better economic and environmental outcomes*. OECD Economic Survey, The Netherlands, Paris.
- Percoco, M., 2013. Is road pricing effective in abating pollution? Evidence from Milan. *Transp. Res. D*, 25(December), 112–118.
- Prud'homme, R., Bocarejo, J.P., 2005. The London congestion charge: a tentative economic appraisal. *Transp. Policy* 12 (3), 279–287.
- Quddus, M.A., Carmel, A., Bell, M.G.H., 2007. The impact of congestion charge on retail: the London experience. *J. Transp. Econ. Policy* 41 (1), 113–133.
- Raux, C., 2005. Comments on "The London Congestion Charge: A Tentative Economic Appraisal". *Transp. Policy* 12 (4), 368–371.
- Rotaris, L., Danielis, R., Marcucci, E., Massiani, J., 2010. The urban road pricing scheme to curb pollution in Milan, Italy: description, impacts and preliminary cost-benefit analysis assessment. *Transp. Res. A* 44, 359–375.
- Santos, G., 2005. Urban congestion charging: a comparison between London and Singapore. *Transp. Rev.* 25 (5), 511–534.
- Santos, G., Bhakar, J., 2006. The impact of London congestion charging scheme on the generalised cost of car commuters to the city of London from a value of travel time savings perspective. *Transp. Policy* 13 (1), 22–33.
- Santos, G., Fraser, G., 2006. Road pricing: lessons from London. *Econ. Policy* 21 (46), 263–310.
- Santos, G., Shaffer, B., 2004. Preliminary results of the London congestion charging scheme. *Public Works Manage. Policy* 9, 164–181.
- TfL, 2004. *Annual Report 2004*. Transport for London, London.
- TfL, 2006. *Central London Congestion Charging: Impacts monitoring. Fourth Annual Report*, Transport for London, London.

Further Reading

- Lee, D., Lemieux, T., 2008. Regression discontinuity design in economics. *J. Econ. Literat.* 48, 281–355.
- Lee, D.S., Lemieux, T., 2010. Regression discontinuity designs in economics. *J. Econ. Literat.* 48 (2), 281–355.
- Percoco, M., 2014a. The effect of road pricing on traffic composition: evidence from a natural experiment in Milan, Italy. *Transp. Policy* 31(January), 55–60.
- Percoco, M., 2014b. *Regional Perspectives on Policy Evaluation*. Springer, Heidelberg.
- TfL, 2008. *Impacts Monitoring—Sixth Annual Report*. Transport for London, London.
- TfL, 2009. *Travel in London Report number 1*. Transport for London, London.
- TfL, 2010. *Annual Report 2010*. Transport for London, London.

Car Ownership and Car Use: A Psychological Perspective

J.L. Veldstra^{*,#}, A.B. Ünal^{*,†,#}, E.M. Steg^{*}, ^{*}Department of Psychology, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, Netherlands; [†]Campus Fryslan, University of Groningen, Leeuwarden, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	384
Motives for Car Ownership and Car Use	384
Personal Values, Beliefs, and Norms	385
Social Norms	386
The Influence of Past Personal Experiences and Habits	387
Perceived Costs	388
Transport Policies	388
Summary and Conclusions	389
See Also	389
References	389
Further Reading	390

Introduction

The general consensus is that car use needs to be reduced given its serious negative impacts on the environment (Garling and Steg, 2007) and on human well-being, including the growing health problems caused by car-related sedentary lifestyles (Dora and Phillips, 2000). Despite these negative effects, why do many people continue to own and drive a car?

This article gives an overview of psychological factors that are related to car ownership and car use. The relationship between car use and car ownership is a matter of debate and the view on this relationship differs between research disciplines. In psychology car use is often studied in context of the broader topic of traveling behavior. Psychological studies mostly use general concepts from psychological theory, and apply them to traveling behavior in general or car use more specifically. Car ownership is, however, hardly studied. In this article, we will therefore discuss psychological factors that influence travel behavior with a focus on private car use. The premise being that people who use a private car are mostly car owners. Although, in recent years other forms of car use without car ownership are becoming more popular the market share of car-sharing is still small as compared to car-ownership (Loose et al., 2006), indicating that car users are probably mostly car owners.

We will start this article by discussing motives for owning a private car and the influence of the social context on private car use. We zoom in to how personal values, beliefs, and norms are related to car use and discuss the influence of past personal experiences with other modes of transport on car use and the difficulty of changing car use habits. Next, we discuss how knowledge on psychological factors influencing car ownership and use can be used to design interventions to reduce car use, for example by promoting car sharing. Last, we will discuss the influence of pricing policies on car use.

Motives for Car Ownership and Car Use

Generally, when people are asked why they own a car, they will come up with all kinds of practical arguments, such as my car gives me flexibility to go to work at any time; the car enables me to bring my kids to school. This illustrates that car ownership is generally regarded as mostly utilitarian and practical. However, besides the convenience a car brings it is also a means to express one's identity. For example, individuals can express that they have high status by owning an expensive car, or show that they care for the environment by driving an electric vehicle, meaning that besides the often first-mentioned instrumental motives for owning a car, there are also symbolic motives that influence car ownership. Also, for many people driving is in itself a positive and enjoyable activity (Gatersleben, 2007; Mokhtarian, 2005; Steg et al., 2001; Steg, 2005). In other words, people have several motivations for owning a car. Steg (2005) studied motivations of private car use and in line with Dittmar (1992), proposed that the motivations to use the private car fall into three categories: Instrumental, Symbolic, and Affective motivations.

Instrumental motives are defined as practical motives such as the convenience or inconvenience a car brings its owner (Crossreference: 10677: Car use benefits). Symbolic motives reflect the extent to which one can express one's identity and enhance one's status by driving a certain car. Affective motives refer to various emotions that can be evoked by driving a car, such as pleasure, excitement, or relaxation but also stress or anxiety. In line with this, Steg and Tertoolen (1999) proposed a theoretical model in which they illustrated how these motivations influence each other and private car use (see Fig. 1). As shown, the instrumental and

[#]Veldstra and A.B. Ünal contributed equally and have shared first authorship of the paper.

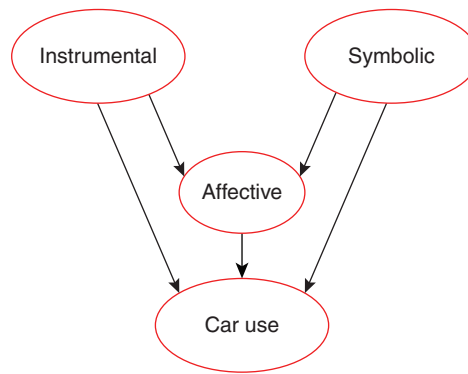


Figure 1 Motivational model for private car use. Source: [Steg and Tertoolen \(1999\)](#).

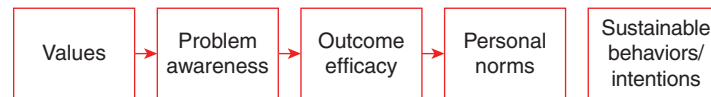


Figure 2 Value-belief-norm theory. Source: Adapted from [Stern \(2000\)](#).

symbolic motives influence the affective motives meaning that if one regard driving a car as convenient, low cost, and safe, for example, and one also feels one can express who they are with the car, they are likely to appraise driving it as giving them a good feeling, which in turn makes using the car more likely.

[Lois and López-Saez \(2009\)](#) tested this model empirically to predict the frequency of car use from the motives for car use and found that the affective link one has with one's private car explains 12% of the frequency of car use. They also found that the symbolic and instrumental appraisals predict this affective link, confirming the hypothetical relationship between affective, instrumental, and symbolic motives as proposed by [Steg and Tertoolen \(1999\)](#). These findings also indicate why it is difficult to convince people to switch to other transport modes with similar instrumental qualities. People are less willing to give up on driving because they seem to enjoy it and they seem to associate it with prestige and high-status. Can we then also make environmentally friendly alternatives attractive to people?

In a recent study, researchers replicated the study of [Steg \(2005\)](#) by applying it to the purchase of electric vehicles ([Noppers et al., 2014](#)). The findings revealed a similar pattern. Participants indicated that instrumental aspects of electric vehicles were more important to them in their decision to buy an electric vehicle, but purchasing intentions were actually better predicted by the evaluation of the symbolic and environmental aspects of the electric vehicle. More specifically, the intention to purchase an electric vehicle was mostly associated with gaining status. This finding is in line with [Egbue and Long \(2012\)](#), who proposed that early adoption of innovations is associated with high-status amongst people. Hence, people might prefer certain transport modes, including relatively environmentally friendly modes, if they are associated with high-status and prestige ([Noppers et al., 2014](#)). In other words, mobility choices might have a signaling function that is used to convey one's identity and status.

Given the rise in new technologies, trends, and innovations in the domain of mobility (such as automated vehicles or transport apps), the potential signaling function of the adoption of green mobility innovations seems to be very relevant to promote sustainable options.

Personal Values, Beliefs, and Norms

Sustainable modes of transport do not always score high on the three main motivations to use and own a car that we discussed above. For instance, using a bike would score low on instrumental aspects because it takes longer to travel from A to B with a bike as compared to a car. Similarly, when it is cold or rainy, riding the bike would be rather inconvenient and costly for the individual as compared to the comfort of private car. But does this mean that efforts to reduce car ownership and car use are doomed to fail? As we will discuss below, moral considerations could play a role in achieving a sustainable mobility transition despite the seeming disadvantages such as inconvenience. So the question is to what extent moral considerations play a role to change to sustainable mobility behavior and what kind of process would fuel the activation of moral norms among car users?

According to the Value-Belief-Norm (VBN) theory ([Stern et al., 1999](#); [Stern, 2000](#)), a value-triggered norm-activation process predicts moral actions, such as prosocial and proenvironmental behaviors. Notably, the VBN theory proposes a causal chain of relationships (see [Fig. 2](#)) whereby personal norms that are defined as one's feeling of moral obligation to act in a particular way ([Schwartz, 1977](#)) are the direct predictors of behavior. For example, when people have strong personal norms to reduce car use, they are more likely to accept car use reduction policies ([De Groot et al., 2008](#); [Jakovcovic and Steg, 2013](#); [Ünal et al., 2018](#)).

Such activation of personal norms, in turn, depends on whether one is actually aware of the problems resulting from car use, such as increased air pollution or reduced space for other road users like pedestrians (i.e., problem awareness) and whether one

recognizes that she/he contributes to these problems by using the car (i.e., ascription of responsibility). According to the VBN theory, when one feels such responsibility, one is also more likely to feel a personal moral obligation to change behavior.

The VBN theory further posits that problem awareness is triggered by values, which are defined as goals that transcend time and situations and that act as guiding principles for people (Schwartz, 1992). Four values seem to be of particular relevance when it comes to proenvironmental behaviors like car use reduction: biospheric, altruistic, egoistic, and hedonic values (Steg et al., 2014; De Groot and Steg, 2008; Stern, 2000; Stern and Dietz, 1994), which we will elaborate below.

People who endorse biospheric values have a strong concern to reduce their environmental impact and to protect the environment and natural resources. People with strong biospheric values would therefore act proenvironmentally even though it is somewhat difficult or costly to do so. They would, for example, not mind giving up their car despite some inconvenience it would bring if this behavior would benefit the environment. People who strongly endorse altruistic values have a strong concern to help improve the well-being of others in general. They would act pro-environmentally when it is good for the community and the well-being of others. They would for example give up owning a car to help others, maybe to create green space for social activities instead of holding it for parking cars. Since biospheric and altruistic values are both not directed at gains for the self but rather put emphasis on goals that are beyond the self, they are called selftranscendence values (Schwartz, 1992).

Egoistic and hedonic values, on the other hand are called selfenhancement values since they do put emphasis on personal gains and pleasure (Schwartz, 1992). The difference in these two value types lies in the type of benefits sought. People who endorse egoistic values have a strong concern to improve benefits, including financial benefits or status. As such, people with strong egoistic values will act proenvironmentally particularly when it brings those financial gains or when it is status-enhancing to do so. People who strongly endorse hedonic values have a key concern to maximize pleasure and fun. This means that people with strong hedonic values will act proenvironmentally particularly when it is pleasurable right now. Reducing car use is oftentimes perceived to be not very pleasurable meaning that it might be difficult to move toward sustainable behavior, especially for people with strong hedonic values.

Research largely supports this reasoning. Strong biospheric and/or altruistic values are found to trigger a process of norm activation by having a positive impact on problem awareness (De Groot and Steg, 2008; De Groot et al., 2012; Jakovcevic and Steg, 2013; Nordlund and Garvill, 2003; Steg et al., 2005; Stern, 2000; Ünal et al., 2018; Ünal et al., 2019), which in turn was found to be related to recognizing one's own contribution to these problems (ascription of responsibility) and feeling a moral obligation to act sustainably (personal norm). On the other hand, strong hedonic and egoistic values were found to be either not related to problem awareness or negatively related (De Groot and Steg, 2008; De Groot et al., 2007; Jakovcevic and Steg, 2013; Stern, 2000; Stern et al., 1995; Thøgersen and Ölander, 2002; Ünal et al., 2018; Ünal et al., 2019). These findings indicate that values could act as a motivational source to trigger sustainable mobility decisions and behaviors. Yet, sometimes, even people with strong biospheric values might find it hard to drive less. For instance, when it is pouring rain and one need to bring children to school and everyone around you is driving to school with their private car, even those with strong biospheric values may decide to drive to school as well. In the following sections, we will elaborate on these other psychological factors that might have an influence on our decisions and sustainable behaviors.

Social Norms

Determining one's mobility behavior and decisions are not only influenced by individual characteristics, such as values, beliefs, and personal norms, but also by the perception of what others do (descriptive social norms) or what we think that others expect us to do (injunctive social norms) (Cialdini et al., 1990). For instance, if a person thinks that the majority of the people in their neighborhood use public transport instead of owning a car, then using the bus becomes a descriptive norm in that environment. Descriptive norm can influence individual actions as they reflect what is the appropriate and sensible thing to do, and we tend to think that the majority will make the most sensible choice. That is, if the majority is doing a particular behavior, that must be the right behavior.

Communicating the descriptive norm can be effective in influencing one's behavior by for instance, emphasizing the extent to which others engage in the desired behavior in information campaigns, and providing information to people about what others do when it comes to this particular behavior. For example, if a municipality of a city wants to motivate car owners to commute by bike instead of by private car, they could inform their citizens that the majority of people living in their city travel to work by bike. However, in some instances the undesired behavior can be the descriptive norm. For instance, at the moment, driving is still a strong descriptive norm, as the majority prefers to drive a vehicle, which might make other people to adopt a private vehicle or ignore car use reduction campaigns. As such, descriptive norm can influence behavior in both directions: it not only reinforces the desired behavior but also the undesired behavior. In that respect, information regarding the descriptive norm should be credible to be effective.

Injunctive norms, on the other hand reflect unwritten rules about what a person thinks others expect him or her to do in a given situation (Cialdini et al., 1990). People follow injunctive norms to gain social approval, or avoid social disapproval. Different than descriptive norms, injunctive norms do not only tell what is the right thing to do at a certain moment, but they also signal that norm-violations would be associated with social sanctions while normative behavior would be associated with social rewards. Hence, when an injunctive norm is violated, this could have social consequences for the individual, such as being negatively judged by one's immediate community. As such, communicating the injunctive norms can also be used to change behavior. But what happens when there is a mismatch between descriptive and injunctive norms? How would an individual behave when the rule is not to drive your

kids to school and no cars are allowed in front of the school but almost everyone is driving their kids to school and parking in front of the school?

In such situations where there is a conflict between the injunctive and descriptive norm, the most salient norm becomes most influential (Reno and Kallgren, 1990). Following from the above example, an injunctive norm, such as a no-parking sign in front of the school, might make it more salient that other drivers are not obeying the rules (descriptive norm), increasing the likelihood that one follows the descriptive norm of disregarding the no parking sign. Situational cues (e.g., signs) are crucial in that respect in underlining whether other people are following the norm or not; thereby increasing the chances of norm-abiding or norm-violating behavior. [Sparkman and Walton \(2017\)](#) argues that particularly for behaviors like car use indeed, which are clearly carried out by the majority, the descriptive norm as a static and seemingly unchanging norm might work against sustainability.

They reason that for such behaviors it would be more effective to inform people about the so-called dynamic norm, which would emphasize that there is a change in the behavior of some large group of minorities. Following their reasoning, it would be an effective strategy to tell that “30% of the people last year reduced their car use and switched to other modes of transport” because such information would make it salient that there is a trend of car use reduction in the society, which would increase even further in the future. Indeed, anticipation of future behavior (referred to as pre conformity by the authors) seems to explain how dynamic norms could affect behavioral intentions as well as real behavior in the desired direction. When the idea of “changing norms” is salient in people’s minds, they seem to be more open to considering changing individual behavior.

The Influence of Past Personal Experiences and Habits

[Weinberger and Goetzke \(2011\)](#) proposed that the choice to use the private car is influenced by what people have learned from past experiences. They propose that when people have no experience in using public transport, for example, it may not be part of their choice pallet when making travel decisions. Hence when faced with travel decisions, people might choose to drive a private car even if that may not be the best choice given the circumstances. [Weinberger and Goetzke \(2011\)](#) illustrate this by comparing two hypothetical women with the same demographic characteristics moving to a transit-rich city. One woman moving from a rural area where one needs a car to get out and about and one woman from another big city who has experience using the tube, bus and even shared cars. They propose that the first woman, having no experience with public transport, will be more likely to continue using a car since she is used to driving to get around. She may either not even think of using the tube instead of the car, or the fact that she does not know how the system works may be an obstacle for considering it as a viable option since she has to invest time and energy into learning it, and thus the high nonmonetary cost. In addition, she might have some misperceptions regarding public transport due to lack of experience with it, such as public transport being crowded all the time or tickets being expensive. For the second woman the differences in public transport systems of two cities is a small challenge that will not stop her from using it as she has experience with public transport and it is a natural part of her choice pallet. She will continue to decide on what mode serves her best given the trip.

This is in line with the proposition that mobility behavior is often habitual (Crossreference to 10656: Habitual behavior) ([Verplanken et al., 2008](#)), meaning that once you are used to behaving in a certain way, or in this case choose a particular form of transport, you will most likely keep doing that as long as the consequences are satisfactory. As a result promoting car use reduction and promoting the use of more sustainable ways of transport instead might be difficult because people might find it easier to follow what comes habitual to them.

Habits are indeed a barrier to safe and sustainable transport ([Verplanken et al., 1997](#); [Verplanken and Wood, 2006](#)). More specifically, people with a strong habit of car use were found to seek less information regarding alternative ways of transport, and they required less travel-related information (such as on whether conditions) to decide whether they would pick a car or not ([Verplanken et al., 1997](#), also see [Aarts et al., 1998](#)). These findings suggest that habit strength leads to a biased way of thinking and information seeking, or even a shallow way of processing new information regarding alternative modes of transport. As a consequence, the habitual behavior of car use is likely to remain unchanged.

It appears that people reconsider habitual behaviors only when they experience drastic or natural changes in their lives. For instance, there has been evidence suggesting that habits can be countered, especially for those individuals who are in a transition in their lives, such as moving to a new neighborhood or changing offices ([Fujii and Garling, 2003](#); [Verplanken and Wood, 2006](#); [Walker et al., 2014](#)). Such natural changes in one’s life seem to provide a good opportunity to inform people about alternative ways of commuting, and getting them try out more sustainable modes of transport. Enriching their perceived choice pallet is therefore best tried when circumstances change because habits are strongest in the context that they had been initially developed: the context serves as a cue, eliciting the automated behavior ([Verplanken and Wood, 2006](#)). When such contextual cues are not present in the environment (which is provided in the example of moving to a new neighborhood), then habit strength is weakened, leaving room for considering alternatives and behavior change toward the desired direction.

In addition, a new context might also make people think about possible ways of commuting in that neighborhood in a more systematic way, which might make people, reconsider the automatic behavior of driving. It should be noted that even in new circumstances where habit strength is weakened, it might still be very difficult for people to cease car use, particularly if people do not have much experience with other transport modes ([Weinberger and Goetzke, 2011](#)) but also if alternative modes of transport are unavailable or somewhat difficult to use. Hence, interventions to break habits could best be supported by making the alternatives

available and attractive to people. When the alternative behavior is (cognitively) costly and effortful, it might indeed inhibit us from following it, which we will elaborate on in the next section.

Perceived Costs

Earlier we discussed that moral considerations (i.e., personal norms) have a strong motivational power to stimulate proenvironmental behavior such as reducing car use. But do moral considerations predict any behavior or are there some preconditions for moral considerations to be most effective? We will elaborate on this question in the light of two competing theories that both focus on perceived cost of executing a behavior as a key determinant: the low-cost hypothesis (Diekmann and Preisendörfer, 2003) and the A-B-C Model (Guagnano et al., 1995).

According to the low-cost hypothesis (Diekmann and Preisendörfer, 2003), motivational factors such as attitudes or personal norms would predict behavior only if the behavior is a low-cost behavior, meaning it is not an effortful or largely inconvenient behavior. That is because when the costs are too high, it is assumed that no one will execute the behavior irrespective of the level of their motivation. Hence, the low-cost hypothesis assumes an interaction between perceived cost and motivational factors in which behavior is predicted by motivational factors such as moral considerations when perceived costs are lower. Evidence for the low-cost hypothesis comes from studies that found that moral considerations were strong predictors of intention but not actual behavior (Abrahamse et al., 2009). Similarly, it was found that personal norms were more predictive of car use reduction for short-distance trips, which can be considered rather feasible than for long distance trips (Keizer, 2014). Leaving the comfort of one's private vehicle to travel by public transport, when distances are somewhat bigger would probably be regarded as rather costly and very unattractive (Anable and Gatersleben, 2005).

The A-B-C model (Guagnano et al., 1995) gives support to the low-cost hypothesis that moral considerations would indeed not be related to executing behavior when costs are very high. In a similar way, the A-B-C model posits that when costs are too high, motivation would not be sufficient to carry out behavior. Interestingly and as opposing to the low-cost hypothesis, the A-B-C model further argues that moral considerations would not be related to behavior when costs are too low either. Notably, the model predicts that moral considerations would affect behavior only when perceived costs are moderate, thereby pointing out to a curvilinear relationship between perceived costs and motivational factors in affecting behavior. In other words, when behavior is somewhat costly but still doable, then motivational factors come at play to trigger the right behavior, which is depictive of a U-shaped relationship between motivational factors and perceived costs.

A recent comparison of the low-cost hypothesis and the A-B-C model in explaining policy acceptance for car use reduction measures revealed that moral considerations were not related to the acceptance of a low-cost policy measure that will have no negative impact on one's driving (i.e. improvements in public transport), supporting the A-B-C model, and opposing the low-cost hypothesis (M. Keizer et al., 2019). In addition, moral considerations were found to be related to the acceptance of a high-cost policy measure on car use reduction (i.e., paying tax for car use), when people felt more able to reduce their car use, indicating a somewhat moderate perceived cost, supporting the A-B-C model.

In short, perceived barriers and costs to reduce car use or stop car-ownership might inhibit sustainable mobility behaviors. An important step to stimulate or sometimes even mandate the right behavior is to develop transport policies that are effective. But what kinds of transport policies are most effective to reduce car use or ownership? And to what extent are they acceptable?

Transport Policies

Policies aimed to encourage car use reduction can be broadly defined as pull and push policy measures (Steg and Vlek, 1997). Push policy measures are those that aim at changing behavior by "forcing" the individual to behave in a certain way. Examples are implementing transport pricing measures, congestion taxes or charges, or reducing parking spaces. Pull policy measures aim at changing behavior by making the desired alternative more attractive. Campaigns on raising awareness on public transport, offering daily commuters a free travel card for a couple of weeks or changing infrastructure to include e-bike charging stations are examples of pull policy measures.

Pricing strategies, which are commonly implemented by policy-makers often aim at reducing the financial costs of using environmentally-friendly means of transport (e.g., pull measures) or increasing the financial costs of environmentally-harmful modes (e.g., push measures; Thøgersen, 2009; Schuitema and Steg, 2008). In one particular study, Danish car-owners were given a voucher that allowed them to travel with public-transport for a full month for free (Thøgersen, 2009). The idea was to break their car use habit, and to help them develop positive associations regarding the use of public transport, such as regarding the quality of service and the convenience of using public transport. Findings revealed a positive effect of the intervention on the frequency of using public transport. Notably, the price promotion led to an increase in the use of public transport, which was twice as much than the use of public transport before the intervention. The positive influence of the price promotion lingered for a period of 5 months after the removal of the price promotion. This suggests that promoting the use of public transport by making it financially appealing to car-owners might lead to a long-term change in transport behavior, even after the financial incentive is removed (Thøgersen, 2009). Importantly, a positive experience with public transport might help change the misperceptions about the quality of public transport

(Fujii et al., 2001), and a financial incentive might be useful to make people be open to such an experience at the first place (Fujii and Kitamura, 2003).

Similar findings have been reported when it comes to the acceptability of pricing policies in the form of a congestion charge, meaning a tax that needs to be paid when you enter or leave the city center (Nilsson et al., 2016; Schuitema et al., 2010). For instance, before the introduction of a congestion charge in Stockholm, acceptability of this pricing policy was low because people expected an increase in financial costs and a restriction on one's car use. Interestingly, acceptability increased after the implementation of the congestion charge. Findings suggest that acceptability was probably higher because the charge had more positive effects than people anticipated, such as reduction of environmental and parking problems. Furthermore, the pricing policy was supplemented with infrastructural improvements such as an increase in the frequency of public transport service in Stockholm, which could have made it easier for car users to reduce their car trips and consider other more sustainable transport modes.

Yet, incentives do not always work as intended, and might only be effective in the short-term. For example, in one study, drivers were given a discount on their insurance premium if they adopted an environmentally friendly and safe driving style by driving according to the speed limit (Bolderdijk et al., 2011). Their driving behavior was observed via GPS devices. It was found that drivers indeed reduced their speed when incentivized, and engaged in less speeding violations as compared to another group of drivers who were not given a discount on their insurance. Interestingly though, when the financial incentive was removed, drivers increased their speeding violations again, and eventually, the difference between those who received a financial incentive, and those who were not given the financial incentive vanished. This finding indicates that it might be very difficult to have long-term behavioral changes with short-lasting financial incentives, especially when the targeted behavior has no positive individual rewards in the long-run.

Summary and Conclusions

In this article, we aimed to give an overview of psychological factors that influence car ownership and car use. In view of the general consensus that car use needs to be reduced in the face of negative impacts on the human environment, behaviors such as less car use or buying an electric vehicle instead of a fossil-fuelled car can be promoted. Achieving a sustainable mobility transition depends for a large part on the extent to which people are willing and feel that they are able to change to sustainable mobility behavior though. As we discussed, feelings of moral obligation to use a sustainable mode, awareness regarding problems resulting from car use, and a sense of personal responsibility seem to predict sustainable mobility intentions and behaviors. Notably, this process of norm activation can be triggered by personal values, and particularly biospheric values.

Mobility behavior is not only influenced by individual characteristics but also by the perception of what others do, what individuals think others expect them to do as well as what behavior is becoming more common in the society. Such knowledge of these social norm mechanisms can help to change behavior by for instance, reminding people of the trending sustainable mobility behaviors that other people are doing.

Habitual behavior and past experiences, such as a strong car use habit, might challenge the transition to sustainable mobility behaviors. That is because changing habits can be cognitively effortful and require the processing of new information, meaning that behavior change is usually perceived as very costly and effortful for people with strong car use habits. As discussed, it is still possible to challenge perceived barriers and costs to promote sustainable mobility behaviors. For instance, people with a car use habit could be approached in moments when they are going through a transition in their lives, such as moving to a new city, as research shows that they will be more open to new information in such times.

To summarize, several psychological factors including individual, motivational, perceptual, and habitual factors could play a role in affecting people's car-ownership and cars-use behavior. By designing evidence-based interventions that stem from psychological theories, we can stimulate car use reduction and sustainable mobility behaviors.

See Also

Car Use Benefits; Habitual Behavior

References

- Aarts, H., Verplanken, B., Van Knippenberg, A., 1998. Predicting behavior from actions in the past: Repeated decision making or a matter of habit? *J. Appl. Soc. Psychol.* 28 (15), 1355–1374.
- Abrahamse, W., Steg, L., Gifford, R., Vlek, C., 2009. Factors influencing car use for commuting and the intention to reduce it: a question of self-interest or morality? *Transport. Res. Part F: Traffic Psychol. Behav.* 12 (4), 317–3248.
- Anable, J., Gatersleben, B., 2005. All work and no play? The role of instrumental and affective factors in work and leisure journeys by different travel modes. *Transport. Res. Part A Policy Pract.* 39, 163–181.
- Bolderdijk, J.W., Knockaert, J., Steg, E.M., Verhoef, E.T., 2011. Effects of pay-as-you-drive vehicle insurance on young drivers' speed choice: results of a Dutch field experiment. *Accid. Anal. Prev.* 43 (3), 1181–1186.
- Cialdini, R.B., Reno, R.R., Kalgren, C.A., 1990. A focus theory of normative conduct. *J. Pers. Soc. Psychol.* 58 (6), 1015–1026.

- De Groot, J.I.M., Steg, L., Dicke, M., 2008. Transportation trends from a moral perspective: value orientations, norms and reducing car use. In: Gustavsson, F.N. (Ed.), *New Transportation Research Progress*. Nova Science Publisher, Hauppauge, NY.
- Diekmann, A., Preisendörfer, P., 2003. Green and greenback: the behavioral effects of environmental attitudes in low-cost and high-cost situations. *Rational. Soc.* 15 (4), 441–472.
- Dittmar, H., 1992. *The Social Psychology of Material Possessions. To Have Is to Be* (Hemel Hempstead 1992)
- Dora, C., Phillips, M., 2000. Transport, environment and health. WHO Regional Office for Europe, Copenhagen.
- Egbue, O., Long, S., 2012. Barriers to widespread adoption of electric vehicles: an analysis of consumer attitudes and perceptions. *Energy Policy* 48, 717–729.
- Fujii, S., Garling, T., 2003. Development of script-based travel mode choice after forced change. *Transp. Res. F* 6, 117–124.
- Fujii, S., Garling, T., Kitamura, R., 2001. Changes in drivers' perceptions and use of public transport during a freeway closure: effects of temporary structural change on cooperation in a real-life social dilemma. *Environ. Behav.* 33, 796–808.
- Gatersleben, B., 2007. Affective and symbolic aspects of car use. In: Garling, T., Steg, L. (Eds.), *Threats to the Quality of Urban Life from Car Traffic: Problems, Causes, and Solutions*. Elsevier, Amsterdam, pp. 219–233.
- Guagnano, G.A., Stern, P.C., Dietz, T., 1995. Influences on attitude-behavior relationships: a natural experiment with curbside recycling. *Environ. Behav.* 27 (5), 699–718.
- Jakovcevic, A., Steg, L., 2013. Sustainable transportation in Argentina: values, beliefs, norms and car use reduction. *Transport. Res. Part F* 20, 70–79.
- Keizer, M., Sargisson, R.J., van Zomeren, M., Steg, L., 2019. When personal norms predict the acceptability of push and pull car-reduction policies: Testing the ABC model and low-cost hypothesis. *Transport. Res. Part F Traffic Psychol. Behav.* 64, 413–423.
- Lois, D., López-Saez, M., 2009. The relationship between instrumental, symbolic and affective factors as predictors of car use: a structural equation modeling approach. *Transport. Res. Part A* 43, 790–799.
- Loose, W., Mohr, M., Nobis, C., 2006. Assessment of the future development of car sharing in Germany and related opportunities. *Transport Rev.* 26 (3), 365–382.
- Mokhtarian, P.L., 2005. Travel as a desired end, not just a means.. *Transport. Res. Part A Policy Pract* 39 (2–3), 93–96.
- Noppers, E.H., Keizer, K., Bolderdijk, J.W., Steg, L., 2014. The adoption of sustainable innovations: driven by symbolic and environmental motives.. *Global Environ. Change* 25, 52–62.
- Schuitema, G., Steg, L., 2008. The role of revenue use in the acceptability of transport pricing policies. *TransportRes. Part F* 11, 221–231.
- Schuitema, G., Steg, L., Forward, S., 2010. Explaining differences in acceptability before and acceptance after the implementation of a congestion charge in Stockholm. *Transport. Res. Part A Policy Pract.* 44 (2), 99–109.
- Schwartz, S.H., 1977. Normative influences on altruism. *Advances in Experimental Social Psychology*, 10. Academic Press, pp. 221–279.
- Schwartz, S.H., 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries.. In: Zanna, M.P. (Ed.), *Advances in Experimental Social Psychology*, vol. 25. Academic Press, San Diego, pp. 1–65.
- Sparkman, G., Walton, G.M., 2017. Dynamic norms promote sustainable behavior, even if it is counter normative. *Psychol. Sci.* 28 (11), 1663–1674.
- Steg, L., Vlek, C., 1997. The role of problem awareness in willingness-to-change car use and in evaluating relevant policy measures.. In: Rothengatter, J.A., Carbonell Vaya, E. (Eds.), *Traffic and Transport Psychology. Theory and Application*. Pergamon, Elmsford, NY, pp. 465–475.
- Steg, L., Tertoolen, G., 1999. Sustainable transport policy: the contribution from behavioral scientists. *Public Money Manag.* 19 (1), 63–69.
- Steg, L., 2005. Car use: lust and must instrumental, symbolic, and affective motives for car use. *Transport. Res. Part A Policy Pract.* 39, 147–162.
- Steg, L., Perlaviciute, G., van der Werff, E., Lurvink, J., 2014. The significance of hedonic values for environmentally relevant attitudes, preferences, and actions. *Environ. Behav.* 46 (2), 163–192.
- Steg, L., Vlek, C., Slotegraaf, G., 2001. Instrumental-reasoned and symbolic-affective motives for using a motor car. *Transport. Res.-F: Psychol. Behav.* 4 (3), 151–169.
- Stern, P.C., Dietz, T., 1994. The value basis of environmental concern. *J. Soc. Issues* 50, 65–84.
- Stern, P.C., Dietz, T., Abel, T., Guagnano, G.A., Kalof, L., 1999. A value belief norm theory of support for social movements: The case of environmentalism. *Human Ecol. Rev.* 6, 81–95.
- Stern, P.C., 2000. Toward a coherent theory of environmentally significant behavior. *J. Soc. Issues* 56 (3), 407–424.
- Thøgersen, 2009. Promoting public transport as a subscription service: Effects of a free month travel card. *Transport Policy* 16 (6), 335–343.
- Ünal, A.B., Steg, L., Gorsira, M., 2018. Values versus environmental knowledge as triggers of a process of activation of personal norms for eco-driving. *Environ. Behav.* 50 (10), 1092–1118.
- Ünal, A.B., Steg, L., Granskaya, J., 2019. To support or not to support that is the question testing the VBN theory in predicting support for car use reduction policies in Russia. *Transport. Res. Part A Policy Pract.* 119, 73–81.
- Verplanken, B., Aart, H., Van Knippenberg, A., 1997. Habit, information-acquisition, and the process of making travel mode choice. *Eur.J. Soc. Psychol.* 27, 539–560.
- Verplanken, B., Wood, W., 2006. Interventions to break and create consumer habits. *J. Public Policy Mark.* 25 (1), 90–103.
- Verplanken, B., Walker, I., Davis, A., Jurasek, M., 2008. Context change and travel mode choice: combining the habit discontinuity and self-activation hypotheses. *J. Environ. Psychol.* 28 (2), 1210127.
- Walker, I., Thomas, G.O., Verplanken, B., 2014. Old habits die hard: Travel habit formation and decay during office relocation. *Environ. Behav.* 1–18.
- Weinberger, R., Goetzke, F., 2011. Drivers of auto ownership: the role of past experience and peer pressure. In: Lucas, K., Blumenberg, E., Weinberger, R. (Eds.), *Auto Motives*. Emerald Group Publishing Limited, pp. 121–135.

Further Reading

- Cialdini, R.B., Reno, R.R., Kalgren, C.A., 1990a. A focus theory of normative conduct. *J. Pers. Soc. Psychol.* 58 (6), 1015–1026.
- Ünal, A.B., Steg, L., Granskaya, J., 2019. To support or not to support, that is the question" Testing the VBN theory in predicting support for car use reduction policies in Russia. *Transport. Res. Part A Policy Pract.* 119, 73–81.

Road Transport: E-Scooters

Gysele Lima Ricci, Klaus Bogenberger, Department of Civil, GEO and Environmental Engineering at Technical University of Munich, Munich, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	391
Background and Context	391
Benefits	393
Challenges and Opportunities	394
Anticipated Future Impact	396
Conclusions and Recommendations	397
See Also	398
References	398
Further Reading	398

Introduction

Micromobility has become a catch-all term for several modes of transportation and is becoming popular for its convenience and environmental friendliness. It will play a huge role in the transition toward Smart Cities. The term includes fully or partially human-powered of small and lightweight electric-driven vehicles, for example, bicycles, e-bikes, e-scooters, human pods, e-skateboard, and other single-person vehicles as shown in the figures. In this concept, the main idea is that all people should have the right to get access to more affordable modes of transportation. Shared mobility has emerged to be one of the cheapest and fastest ways of moving from point A to point B. These systems are an increasingly important part of city transit and mobility systems, as they help people move around cities more seamlessly and efficiently (Fig. 1A–D).

Recent years, since 2018 in particular in the United States and since 2019 in Germany, have been marked by a race toward micromobility, where bikes and e-scooters provide a new way for residents to move throughout their communities. While there is a great deal of promise with these innovations, the emergence of micromobility comes with its own set of challenges and considerations for planners, residents, and local decision makers. At the same time, many communities still have vast surface transportation needs, which must be addressed for micromobility to take shape.

E-scooters are an urban trend in transportation as a way to get around. This class of mobility has truly taken off. From the different categories of electric vehicles, e-scooters were found to be the most disruptive due to their ride-share growth. E-scooters are convenient, cheap and comprise first and last mile options. They have been deployed by start-up companies quite often without any communication between cities and the start-ups. Because of not communicating a plan regarding regulations with the local government, many scooter programs failed or were rolled back due to temporary bans.

To grasp a better understanding about e-scooters, key facts are shown in section, Background and Context and the benefits of the use of e-scooters are described in section, Benefits. Section, Challenges and Opportunities displays challenges and opportunities for these modes of urban mobility transportation. To conclude, section, Anticipated Future Impact describes the anticipated future impact, while section, Conclusions and Recommendations summarizes a set of recommendations that cities can consider as they work to regulate the use of e-scooters.

Background and Context

Since the first scooter sharing system was launched in Santa Monica, California in 2017, the e-scooter market is growing day after day. Numerous e-scooter operators have received vast amounts of venture capital funding and rapidly spread to cities worldwide.

The global electric scooters market size was estimated at USD 18.6 billion in 2019. Only the US micromobility market is expected in 2030 to be worth between 200 and 300 billion US dollars. In Europe alone the market for shared e-scooter services is expected to reach at least 12 billion US dollars by 2025. There are more than 150.000 scooters in cities in the United States and Europe available for sharing (Mobility Foresights Transportation, 2020).

In September 2019, there were more than 10.000 units of e-scooters in Berlin, Germany and 2.865 units of e-scooters in Portland, United States (from April 26 to November 30 of 2019, rental e-scooters in Portland traversed 1.01 million miles during 954,000 trips). As of March 2020, e-scooter sharing services are available in 177 US cities and European cities, down from 223 cities in December 2019. In addition, in March 2020, cities such as Paris, Madrid, and Berlin were ranking as the three main European cities in terms of the size and use of fleet of e-scooter sharing (Mobility Foresights Transportation, 2020).

How can one explain the varying performance of the major scooter providers across different cities?



Figure 1 (A) Different vehicles currently implemented in the micromobility market: electric bikes. (B) Different vehicles currently implemented in the micromobility market: electric scooters. (C) Different vehicles currently implemented in the micromobility market: human pods. (D) Different vehicles currently implemented in the micromobility market: electric skateboards.

The year 2018 was the decline of many ride-hailing and bike-sharing operators. Many new start-ups foraying the e-scooter sharing market gain more strength in the urban mobility. According to the National Association of City Transportation Officials—NACTO's Guidelines for Regulating Shared Micromobility (2019), the addition of e-scooters boosted the total number of micromobility trips from 35 million in 2017 to 84 million in 2018. In fact, e-scooter operators have already pumped in \$5.7 billion into the industry and looking to capture new cities and urban spaces around the world.

The number of companies is still growing in several cities on several continents and has positive and negative impacts on the society. Lime and Bird, for example, two of the largest companies, are currently valued at \$ 1 billion and \$ 2 billion, respectively. **Table 1** shows the profile of some major e-scooter operators.

There are several models for how these systems are managed. In the United States, the competitive landscape is extremely consolidated. In Europe, on the other hand, it is comparatively fragmented with local players like Tier in Germany that is giving a tough fight to the American duo—Lime and Bird. The fleet of e-scooter vehicles might be owned and maintained privately, like the e-scooter provider VOI and Lime, or owned and maintained publicly, like Tier in Germany, which uses a hybrid model which is publicly owned but privately maintained.

E-scooters (sharing) are now located in all major US and European cities except London where it is prohibited (some cities have banned scooters or are taking plenty of time to think about appropriate introduction and rollout). The Asian market of e-scooter sharing industry is currently less than 4% of the US market size. They are not only available for rental in major cities but also in medium-sized cities with around 200,000 inhabitants or less.

Regulations and fleet size vary from country to country, even from city to city, an ever-growing number of companies is placing as many devices as they can in as many cities as possible. Many cities don't have a balance of size and fleet distribution (Madrid in Spain alone has authorized 18 different operators) in an attempt to achieve market dominance. Due to higher upfront licensing costs per e-scooter and reduction in the max fleet size allowed per operator in many cities, some e-scooter operators have started scaling down their fleet size. According to Mobility Foresights Transportation (2020), Lime, the global market leader in e-scooter sharing, has significantly scaled down its US presence, that is, around from 90 cities in December 2019, to 54 cities in March 2020.

To understand more about e-scooters, the following paragraph describes the benefits, challenges and opportunities for the sector.

Table 1 E-scooter operator's profile

Operator	Profile
LIME	Founded in 2017, LIME is an electric scooter startup base in San Francisco raised over \$455 million of capital. Lime has a fleet of 120,000 e-scooters.
BIRD	Founded in 2017, Bird is a micromobility company based in Santa Monica-California is a dockless scooter sharing startup. Bird raised over \$455 million of capital. Operated shared e-scooters in over 100 cities in Europe, the Middle East, and North America.
GRIN and YELLOW	Founded in 2018, is a Brazilian startup and is the largest e-scooter service in South America. This startup raised over \$270 million of capital.
FLASH	Founded in 2019, is a Switzerland micromobility startup from Delivery Hero raised over \$63 million of capital.
VOI	Founded in 2018, Europe's leading home-grown e-scooter operator from Sweden, VOI hat operates an e-scooter service in a 38 cities across 10 European countries, has raised an \$85 million of capital.
CIRC	Founded in 2018 as Flash, is an electric scooter startup in Berlin-based raised over \$55 million of capital. Operates in more than 40 cities across 14 European countries, in addition to United Arab Emirates.
TIER	Founded in 2018, is a Germany electric scooter startup raised over \$55 million of capital. Tier currently operates in over 40 cities across 12 markets worldwide.
SKIP	Founded in 2017, SKIP is a San Francisco-based dockless electric scooter startup raised over \$31 million of capital.
DOTT	Founded in 2018, is a Netherlands electric scooter startup raised over \$30 million of capital.
WIND MOBILITY	Founded in 2017, WIND is an electric scooter startup base in Berlin and Barcelona. WIND raised over \$22 million of capital.
SPIN	Founded in 2016, SPIN is an electric scooter startup that has raised a total funding amount of \$8 Million. They is one of the popular electric scooter service providers that is currently available in more than 19 cities in U.S.
BEAM	Founded in 2018, is Asian electric scooter startup raised over \$6,4 million of capital. Beam is focused on expanding transportation options in Asia (Singapore).

Benefits

E-scooters offer a variety of advantages for users in urban spaces. It's not hard to understand the benefits of electric scooters. Are e-scooters only fun? For a commuter arriving by train from the suburbs, facing a half or one-mile slog on foot to the office, or a tourist with only a day in a city, the devices present an easy solution: download the app, add credit card information, unlock one of millions of scooters parked nearby, and off you go. It is a particularly easy mode of transportation to book and pay for. Also, e-scooters present an opportunity to support local tourism, in both urban and more rural locations.

"Are they attractive enough that e-scooters will be considered an alternative to urban mobility transport?"

One of the biggest benefits is that e-scooters are dockless, as they can be left anywhere. And because they run on electric batteries, they allow users to feel good about making a climate-friendly choice. A survey conducted in 2019 of attitudes to use e-scooters in New Zealand cities with 591 persons showed 57% of e-scooter trips replaced trips that would have been made by active or sustainable modes (on foot, by bicycle, skateboard, or e-bike), 28% of e-scooter trips replaced a trip by private car or van, motorcycle, ride source vehicle, or taxi (Fitt and Curl, 2019). Also, the benefits of e-scooters can differ widely between geographic areas that are only a few blocks apart due to the differential access of these areas to transit lines and bus routes. The emergence of e-scooter options has inspired many cities to rethink the ways in which their transportation infrastructure might accommodate alternative modes.

Another benefit of e-scooters is the cheaper mode of travel. Also, riding an e-scooter doesn't need any special skills or license. It's also much faster to drive an e-scooter than a car on congested roads. The most-common pricing structure in Europe so far has been 1 EUR for unlocking the scooter and 15 cents per minute as rental time. In cities with higher income levels some operators like Lime and Tier increase prices up to 20 cents per minute.

"How can e-scooters be integrated into overall mobility transportation system?"

E-scooters offer cities another tool in fighting mobility deserts. As a flexible means of transport, e-scooters can play a key role in solving the "first and last mile gaps for the transit systems, filling the gap between a rider's home or destination and public transport stops, opening access to underserved populations." More generally, they also add more options to multi-modal mobility systems and provide better connectivity (Fig. 2).

Additionally, e-scooters are extremely efficient. E-scooters are an efficient way of traveling short distances using efficient and clean energy. Just for comparison, a fuel-powered car can travel a maximum of a mile on 1-kilowatt-hour of energy. However, 1-kilowatt-hour of energy is sufficient for an electric scooter to travel 80 miles. So, e-scooter options provide huge efficiency in terms of energy usage and reduce the dependency on fossil fuels (Hollingsworth et al., 2019). Compared to cars e-scooters are much cheaper. For example, the BMW Model i3, an electric car comes for around €34,950,00. You can buy 100 e-scooters high-quality ones for the same price. It is not yet known about the future of e-scooters. However, it is possible to see the great economic difference compared to cars. So, e-scooters are not only energy-efficient but also have lower costs of production. Also, e-scooters have brought that advantage of

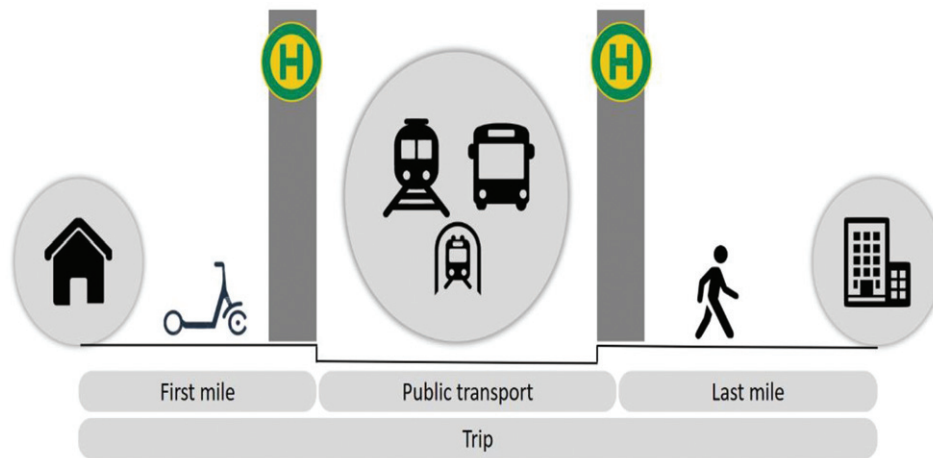


Figure 2 First and last mile of e-scooters.

no noise and the vehicles are very quiet. They reduce the weight of CO₂ emissions not only while traveling but also during production. As a result, they are ideal for an economically and environmentally viable transport option.

Table 2 shows e-scooters cost compared to other transportation modes.

Challenges and Opportunities

In the past years, we have seen an avalanche of e-scooters in several cities, and it is apparent that these new means of transport have become the favorite of millions of people, and consequently the “cash cow” of micromobility. Even though e-scooters are a growing trend, it has its share of challenges.

“How should e-scooters be designed and controlled with regard to service concept, price, vehicle and infrastructure, in order to achieve the highest possible customer acceptance on the one hand and to achieve the municipal goals on the other hand?”

Some of the challenges are related to the infrastructure of the city, while others deal with regulations and safety or sustainability (Bekhit et al., 2019; Fitt and Curl, 2019; Todd et al., 2019). A lack of data and scholarly research on e-scooters also compounded these issues (NACTO, 2019a). The data ensure that municipalities can make informed decisions about what is happening on the public streets and how it might impact safety, health, equity, environmental outcomes, and the distribution of people and resources.

One of the main challenges for e-scooters is about safety (Bekhit et al., 2019). Perhaps the most controversial, and greatest pain point for city leaders is e-scooter operation on the streets and sidewalk. Collisions with both cars and pedestrians have sent thousands of e-scooter users and their unwitting victims to hospitals. Some of the misuse of the dockless vehicles can be chalked up to users’ unfamiliarity with the vehicles and the city’s regulation of their operation.

Sometimes injuries are minor, other times, riders face serious injury or even death. Every city has different rules about where e-scooters can be operated, and ultimately, it is up to the user to educate him or herself. E-scooter operators have also been engaging in various efforts to educate the public about local regulations and the dangers of riding on sidewalks with guidelines, like safety equipment and velocity.

Another important safety challenge is the helmet usage. Many e-scooter related injuries are directly tied to riders not wearing helmets. However, many users claim that the not use a helmet provides a lot of freedom but also leaves riders potentially unprepared and vulnerable. Representative surveys regarding the use of helmets, determined the risk of an accident. Examples of these violations include operating a scooter on the sidewalk, traveling opposite to the flow of traffic, being less than 18 years old, there being more than one rider at a time on a given scooter, and not wearing a helmet. The growth rates of injuries demonstrate the need to ensure that helmets remain available for users (Fig. 3A,B).

Another challenge inherent to e-scooter usage is that many communities lack the infrastructure for alternative modes. Each city needs proper infrastructure to introduce e-scooters. Many regulatory systems surrounding these new modes are not yet settled. If a city doesn’t have requirements like routes or parking spaces, the adoption of e-scooters becomes difficult. That’s the main reason the adoption of e-scooters is low in poor countries like Africa and developing well in countries like United States and France.

“How can we operate and prevent e-scooter vehicles from being parked where we don’t want them to? How do we modify our lanes to keep riders safe while also keeping enough sidewalk space for pedestrians?”

The proper requirements for e-scooter use is necessary so that people can travel safely without fear of accidents or injuries and see this mobility mode favorable. In Munich, Germany, the requirements of the use of e-scooters include: any person riding in prohibited locations (e. g., on the sidewalk) is subject to a €15 fine, which may increase if the behavior hinders (€20) or endangers

Table 2 Comparative transportation costs

Transportation system	Price ^a
Shared e-scooter	1€ + 0,15 min
Bikeshare	1 €/ 30 min
E-bikeshare	1 €/0,15 min
Personal car	€ 0,50/km
Car share	€ 0,30/min
Metro	€ 3,30 (short trip 1,50)
Bus	€ 3,30 (short trip 1,50)
Taxi	€ 5 up to 2 km

^aMunich Price (Germany).**Figure 3** Examples of restrictions to riders of e-scooters.

€25) others, or if it causes damage (€30) (AGORA, 2019). These fines are relatively low compared to other locations internationally like in France where the violations start from 35 euros and range to 135 euros. For example, in Washington DC, e-scooter requirements comprise: up to 600 vehicles to be re-evaluated quarterly; cannot travel more than 10 mph; must provide users with a free helmet within 14 days of request; if parked incorrectly, providers must move vehicles within 2 hours of notification.

These challenges have inspired many cities to commit to designated infrastructure that can accommodate alternative modes. Some cities like Atlanta, Canada, Portland and Washington DC, United States, and Sydney, Australia have begun to paint bike lanes in spaces previously dedicated to curbside parking spots or even created road barriers between e-scooters and bike lanes and vehicle lanes (Fig. 4A,B). These sorts of policies and actions create a more robust micromobility culture, by making e-scooters an alternative mode of transportation to be easier, safer and more efficient, reducing crash risk for all road users (PBOT, 2018; NACTO, 2019b).

Regarding the parking zones for e-scooters, if a city doesn't have requirements like lanes or parking spaces, the adoption of e-scooters becomes difficult. Many American cities like San Francisco, Washington, DC, Kansas, New York, Norfolk, and Los Angeles launched dockless pilot programs in order to improve the regulation system e-scooters (NLC, 2019; NY City Transportation Department, 2019). Most cities use short-term pilots and time-limited permits to explore options for shared bikes and e-scooters in their city in a controlled manner.

Examples can see in several cities. A pilot study launched in Washington, DC found that 8% of 181 observed shared dockless e-scooters and bikes were parked undesirably (DDOT, 2019). Another study in Portland, OR, 3% of 357 found e-scooters impeded ramps, curb cuts, or handrails, 5% completely blocked pedestrian movement, and 8% partially blocked pedestrian movement (Portland Bureau of Transportation, 2018). Last, a study in Rosslyn, Virginia found that 16% of 606 observed e-scooters not parked properly and 6% (36 e-scooters) were blocking pedestrian right-of-way (Owain et al., 2019) (Fig. 5A,B).

E-scooters have also proven to have a unique ability to create public anger and perturbation. Many cities experience negative feedback from residents and pedestrians about e-scooters being discarded carelessly in public spaces, such as sidewalks calling them "e-waste on the side-walk." Cities and providers require users to leave their vehicles in locations that do not block foot traffic or access points.

A challenge not less important concerns the equity in urban environment. In recent years, the urban mobility is challenged by innovative changes and new opportunities of transportation integrating into urban landscape with challenging operating conditions for the cities. However, most of the population and users don't have information about these transportation modes and its impacts. There is a lack of access to these items, which is one of the reasons why many people do not use an e-scooter. Several cities, such as Columbia, Ohio, and Washington, DC, are requiring e-scooters to deploy in underserved areas to ensure, these new pilots and programs align with their goals around equity. Additionally, the distribution of the fleet throughout the city should be balanced, so that it serves all communities equitably.



Figure 4 E-scooters and bike lanes.

Anticipated Future Impact

Since e-scooter vehicles arrived in several cities, local governments have struggled and been challenged to develop new methods to manage this transportation system. The rapid unexpected deployments of local government and the public elicited a range of responses from the different parties. Many cities crafted amenable regulations and pilot programs; others were less welcoming.

"How does the future of e-scooters look like?"

Without doubt, the future is promising. There's more than enough demand, and the promise is admirable. Thus, the question that really interests me is *how can e-scooters grow beyond its problematic introduction to truly become a harmonic service to a city and its residents?*

To start, e-scooters need to be dropped off in specific areas that represent a tidy solution to the city landscape. This means that they aren't in the sidewalk and not a danger to pedestrians and those with disabilities.

Economically thinking, we believe that improving people's mobility and optimizing economic and business cost (total cost, user cost) cities can take different approaches to cost recovery like using the revenue from e-scooters to fund a separate account dedicated to expanding alternative infrastructure. This requires a bond to cover the cost of removing broken or improperly parked scooters. Washington, DC, for example, is requiring a \$10,000 bond to cover the costs of removing broken or improperly parked scooters (NLC, 2019).

When it comes to regulation, we believe that improving regulated market access of the services providing high performance for users and operators. The municipalities should create a comprehensive permitting process for providers. This allows the city to regulate and monitor the deployment of e-scooters while also allowing providers to quickly roll them out. Providers have an amount of quality data on vehicle locations and trips, which can be critical to city governance decisions. Not only this data can help bolster safety and accountability efforts, but we will be also able to identify who is using these services, where they're going and when, and how well their current transportation infrastructure maps that information.

Regarding social aspects, we believe there could be maximizing participation, accessibility, and education of the population. An approach to provide these services to low-income residents, the municipalities should include diverse payment options and fare

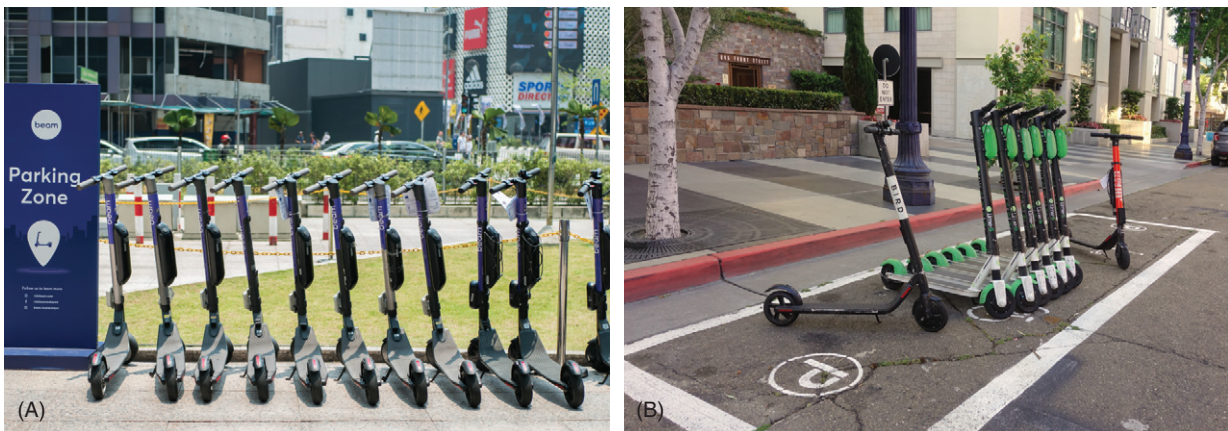


Figure 5 Parking zones.

discounts and should reduce barriers to participation. E-scooters are also poised to promote equity by improving services to low-income and underserved communities. Additionally, there should be constant emphasis on balancing fleets, so that they serve all communities equitably.

The last point is improving safety and health. One of the major lessons learned from the short history of e-scooters that companies should encourage (but most of these not enforce) safety standards. That responsibility falls squarely on the city's shoulders. Understanding how to keep residents safe while allowing them to utilize these new services is one of the biggest challenges that cities will face. Safety means more than requiring riders to use helmets or imposing speed limits. It means re-evaluating the city's entire transportation system, especially for the non-users too.

Many citizens expect that municipalities to develop an antidote against congestion and pollution and e-scooters may be a way to provide a real, everyday alternative. However, in addition to boosting their business, companies in the industry can put more influence on management, taking a pro-active role in creating a real industry in major urban areas. Because of this, we consider e-scooter operators' participation along of the claims of the local communities and government. E-scooter operators will also engage in various efforts to educate the public about local regulations and the dangers of riding on sidewalks.

Conclusions and Recommendations

We conclude a set of recommendations cities can consider as they work to regulate the use of e-scooters. As e-scooters are relatively new to us, many municipalities are still at the early stage of discovering how e-scooter trips are distributed and how urban environments might relate to them.

So the question remains: "What can key players do to exploit the great opportunities of e-scooters in order to achieve the municipal goals of transport planning and to improve the overall system?"

Naturally, urban environment will become more crowded and the e-scooters can go a good solution long way to reduce the stress on existing transportation networks and offer an affordable and viable means of moving around. As cities grapple with increasing pollution and congestion, e-scooters provide a solution that can move more people in a cleaner way and makes efficient use of valuable public road space. Also, the e-scooters have also accelerated a discussion about road safety, and how cities must redesign streets to help keeping all road users safe. As a result, e-scooter operators have emerged as partners for cities looking to meet important sustainability and safety aspects.

The integration of public authorities with operators of e-scooters as a partner ensures that the legal framework, public welfare, social acceptance and networking between cities and countries. The integration will provide Governments with more information in how to conduct the regulations and disseminate information to the population and the countries opening the way for the mobility of the future. There is a huge lack of scientific data and body of knowledge about e-scooters. In order to build on this subject, there is a necessity for more scientific research and data, which will help understanding this industry better and provide better analyses. A combination of digital and physical infrastructure can help aid the conformity of users. This can be made compulsory by regulation. Hence, the authorities should in a position to build up expertise and contribute to the development of new mobility forms through funding together with partners from industry and science.

The e-scooter data sharing will eventually provide more information about rides and this will aid further development in the field. Only infrastructure like lanes and parking does not warrant many changes. But this is hard to accomplish without supervision, regulation and more personnel. Steps like use of helmets and indicators like velocity should be made compulsory for using e-scooters and can have a positive impact on safety. Additionally, licenses for e-scooters can be created to add an additional condition of accountability and conformity in all cities. Municipalities should regulate the number of operators, types of e-scooters implemented, and fleet size released on the street.

To conclude, many of the practices include dynamic recommendation, but much remains to be done.



See Also

Powered Two- and Three-Wheeler Safety; Transport Modes and Cities; Traffic Flow Volume and Safety

References

- AGORA – Verkehrswende, 2019. E-Tretroller im Stadtverkehr Handlungsempfehlungen für deutsche Städte und Gemeinden zum Umgang mit stationslosen Verleihsystemen. Available from: https://www.agora-verkehrswende.de/fileadmin/Projekte/2019/E-Tretroller_im_Stadtverkehr/Agora-Verkehrswende_e-Tretroller_im_Stadtverkehr_WEB.pdf.
- Bekhit, M.N.Z., Le Fevre, J., Bergin, C.J., 2019. Regional healthcare costs and burden of injury associated with electric scooters. *Int. J. Care Injured* 51(2), 271–277. Available from: <https://doi.org/10.1016/j.injury.2019.10.026>.
- District Department of Transportation, DDOT, 2019. Dockless Vehicle Sharing Demonstration. Available from: https://www.researchgate.net/publication/336453702_Pedestrians_and_E-Scooters_An_Initial_Look_at_E-Scooter_Parking_and_Perceptions_by_Riders_and_Non-Riders.
- Fitt, H., Curl, A., 2019. E-scooter use in New Zealand: insights around some frequently asked questions. Available from: <https://ir.canterbury.ac.nz/handle/10092/16336>.
- Hollingsworth, J., Copeland, B., Johnson, J. X., 2019. Are e-scooters polluters? The environmental impacts of shared dockless electric scooters. *Environ. Res. Lett.* 14(8). Available from: <https://iopscience.iop.org/article/10.1088/1748-9326/ab2da8>.
- Mobility Foresights Transportation, 2019. Electric scooter sharing market in US and Europe 2019–2024. Available from: <https://mobilityforesights.com/product/scooter-sharing-market-report/>.
- National Association of City Transportation Officials, NACTO, 2019a. Policy 2019 managing mobility data. Available from: https://nacto.org/wp-content/uploads/2019/05/NACTO_IMLA_Managing-Mobility-Data.pdf.
- National Association of City Transportation Officials, NACTO, 2019. NACTO's guidelines for regulating shared micromobility. Version 2, September 2019. Available from: https://nacto.org/wpcontent/uploads/2019/09/NACTO_Shared_Micromobility_Guidelines_Web.pdf.
- NY City Transportation Department, 2019. A local law to amend the administrative code of the city of New York, in relation to a pilot program for shared electric scooters. Available from: <https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3763648&GUID=05DFE6C0-4E1E-4E49-A95C-BDD1F9C44555>.
- Owain, J. et al., 2019. Pedestrians and e-scooters: an initial look at e-scooter parking and perceptions by riders and non-riders, sustainability. *MDPI* 11(20), 1–13. Available from: <https://doi.org/10.3390/su11205591>.
- Portland Bureau of Transportation, PBOT, 2018. E-scooter findings report. Available from: <https://www.portlandoregon.gov/transportation/article/709719>.
- The National League of Cities, NLC, 2019. Micromobility in cities: a history and policy overview. Available from: https://www.nlc.org/sites/default/files/2019-04/CSAR_MicromobilityReport_FINAL.pdf.
- Todd, J., Krauss, D., Zimmermann, J., Dunning, A., 2019. Behavior of electric scooter operators in naturalistic environments. *SAE Technical Paper* 2019-01-1007. doi: 10.4271/2019-01-1007. Available from: https://www.researchgate.net/publication/332157397_Behavior_of_Electric_Scooter_Operators_in_Naturalistic_Environments.

Further Reading

- Holley, P., 2019. Pedestrians and e-scooters are clashing in the struggle for sidewalk space. *The Washington Post*. Available from: https://www.washingtonpost.com/business/economy/pedestrians-and-e-scooters-are-clashing-in-the-struggle-for-sidewalk-space/2019/01/11/4ccc60b0-0ebe-11e9-831f-3aa2c2be4cbd_story.html?noredirect=on&utm_term=.cba986fdc618.
- Morano, L.H. et al., 2019. Characterization of dockless electric scooter related injury incidents—Austin, Texas, September–November, 2018. *CDC Centers for Disease Control and Prevention. 68th Annual Epidemic Intelligence Service (EIS) Conference, 2018*. Available from: https://www.austintexas.gov/sites/default/files/files/Health/Epidemiology/APH_Dockless_Electric_Scooter_Study_5-2-19.pdf.
- Moreau, H. et al., 2020. Dockless e-scooter: a green solution for mobility? Comparative case study between dockless e-scooters, displaced transport, and personal e-scooters. *Sustainability* 12, 1803. doi:10.3390/su12051803. Available from: https://www.researchgate.net/publication/339648067_Dockless_E-scooter_A_Green_Solution_for_Mobility_Comparative_Case_Study_between_Dockless_E-Scooters_Displaced_Transport_and_Personal_E-Scooters.
- Populus Platform, 2018. The micromobility revolution. Available from: https://research.populus.ai/reports/Populus_MicroMobility_2018_Jul.pdf.
- Shaheen, S., Cohen, A., 2019. Shared micromobility policy toolkit: docked and dockless bike and scooter sharing. UC Berkeley Transportation Sustainability Research Center: Richmond, CA, USA. Available from: <https://escholarship.org/uc/item/00k897b5>.

Railway Station and Network Planning

Ingo Arne Hansen, Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	399
Analysis	399
System Specification	401
Geometric Design	402
Timetabling	402
Simulation	403
Optimization	404
Conclusions	404
References	405

Introduction

Railway stations are places, where trains are designated to stop. The location and design of railway stations depend on the:

- Kind and volume of transport demand, either for passengers or goods,
- Links with other stations,
- Kind of trains and scheduled transport services,
- Topography, land-use and environment,
- Population, economy, and transport policy.

At stations passengers may board and alight trains or goods can be loaded and unloaded. Station facilities collect and distribute passengers, who get on/off the trains via platforms, and may transship goods, respectively. As the classification, formation and blocking of freight trains is much more complicate than the (de-)coupling of passenger trains, passenger railway stations are segregated from freight railway yards. The design of freight railway service networks is not considered here and treated by [Crainic \(2020\)](#) in this work. For the geography of rail transport see [Dobruszkes and Moyano \(2020\)](#), while rail transport and territorial scale is further dealt with by [Monzon and Lopez \(2020\)](#) in this work. Railway traffic management approaches are discussed by Corman in this work.

Currently, the planning and choice of the place and design of railway stations is mostly done in an evolutionary process governed by political decisions. Transport, spatial, urban, and/or environmental planners analyze the performance of existing mobility patterns, traffic, and transport modes in order to develop alternatives to cope with growing demand, traffic congestion, and consumption of natural resources. Then, alternatives for station capacity extension at existing locations or new stations are evaluated by means of system performance criteria, costs, and benefits. Finally, local, regional, or national governments decide, which alternative to choose based on political preferences, budget, and constraints.

The planning process for railway stations consists is explained in the following sections:

- Analysis,
- System specification
- Geometric design,
- Timetabling,
- Simulation,
- Optimization.

Analysis

The analysis of railway stations describes their functions within the transport network, location in the city, transport flows, platform and track geometry, train operations, information, customer service, safety, and security. The network function of stations can be classified according to their network positions, rank and kind of train services. The station's strategic network position is generally divided into ranks according to the kind of trains calling there:

- Top level stations (Inter-/National trains; High-speed/Intercity trains),
- Middle level stations (Inter-/Regional trains; Express trains),
- Low level stations (Sub-/Urban trains; Metro/Subway trains).

Furthermore, railway stations are disaggregated according to their positions along the train line into:

- Terminal stations (Start/End of train lines)
- Intermediate stations.

The station's location in the city determines its accessibility, attractiveness, and design. Accessibility is measured as connection (in distance or time) with other modes of transport (walk, bicycle, taxi, bus, tram, ferry, air), infrastructure (road, river, sea port, airport), and whether it is barrier-free (step-less) for mobility impaired people. Accessibility to/from railway stations is calculated as generalized journey time from/to different (sub-) urban areas (Papa and Bertolini, 2015).

The attractiveness of railway stations depends on different factors, such as:

- Number, density and distance of residents', commuters' and visitors' population, jobs and schools in the catchment area,
- Transport access and egress modes, times, and costs,
- Connectivity, frequency, travel time and costs of train services to/from other destinations and origins,
- Existence of travel facilities and vicinity of commercial and recreation opportunities,
- Image, safety and security of the station and its neighborhood area.

The geometry of platforms and tracks respectively at a railway station can be classified according to its arrangement, use and levels into:

- Lateral platform,
- Central platform,
- Operation by line,
- Operation by direction,
- At-grade,
- Grade-separated,
- Stub-end tracks,
- Through tracks,
- Merging, diverging or crossing tracks,
- Tangent tracks.

Lateral platforms are simple to design and build, but need more space and equipment than central platforms, which require increased track spacing and track bending. Platform operation line by line is easier to design, less expensive to build and used independently, but less comfortable and efficient for transfer of passengers and train circulation than operation by direction (Figure 1). The latter one enables (almost) simultaneous cross-platform transfer between different lines and minimum train interval times.

Stations with grade-separated underground or elevated platforms and through tracks offer maximum capacity at high costs, can be situated close to city centers, generate high transport demand, and are indispensable for multiple long distance train operations of bigger railway networks and convenient transfer of passengers from/to urban and regional public transport lines.

Stub-end tracks are standard for terminal stations, but require many switches, complicated train routing, occupy much space and restrict track capacity. Through tracks are typical of intermediate stations, easier to design, less costly to build and enable maximum throughput. Terminal stations cannot be accommodated easily in densely built-up urban areas, require high cost of construction and are situated mostly at-grade, except underground or elevated metro terminal stations which are extremely costly and therefore restricted to few platform and stabling tracks.

The train movements in stations with multiple tracks are performed according to the timetable via route set-up and released by interlocking machines, as well as by fixed block signals and automatic train protection (ATP), automatic traffic control (ATC) or automatic train operation (ATO) systems. So far, ATO systems have only been successfully implemented on a growing number of driverless (sub-)urban metro lines, whose platforms are usually equipped with screen doors to protect waiting passengers from falling on the track. ATO is still neither available for long distance passenger trains operated on dedicated high-speed railway lines, nor on lines operated by both passenger and freight trains.

Timetables describe the scheduled arrival and departure times, as well as the platform track numbers of all trains at each station. The times between arrivals and departures of the same train indicate the scheduled dwell times at the platform. The interval time between the departure (and the arrival respectively) of the preceding and the following train on the same track or junction is called headway time. The frequency of trains per line is given by the total number of trains in the same direction divided by the length of the period considered (mostly per day or per hour). The scheduled mean headway time at a platform track is equal to the inverse of the train frequency.

The standard information provided at railway stations consists of the (planned/actual) arrival and departure times and platform track number of every train per kind of day and season, signage of entry, exit, platform number, other public transport platforms, taxi, ticketing counters or machines, toilets and other facilities. Timetable information is locally transmitted via printed tables, central and platform (video and/or digital) displays, as well as public address. Frequent updating of actual customer information and advice by station service personnel concerning train delays, canceled trains, platform track changes, quick public alarm and professional guidance in case of incidents, fire, accidents, attacks are essential.

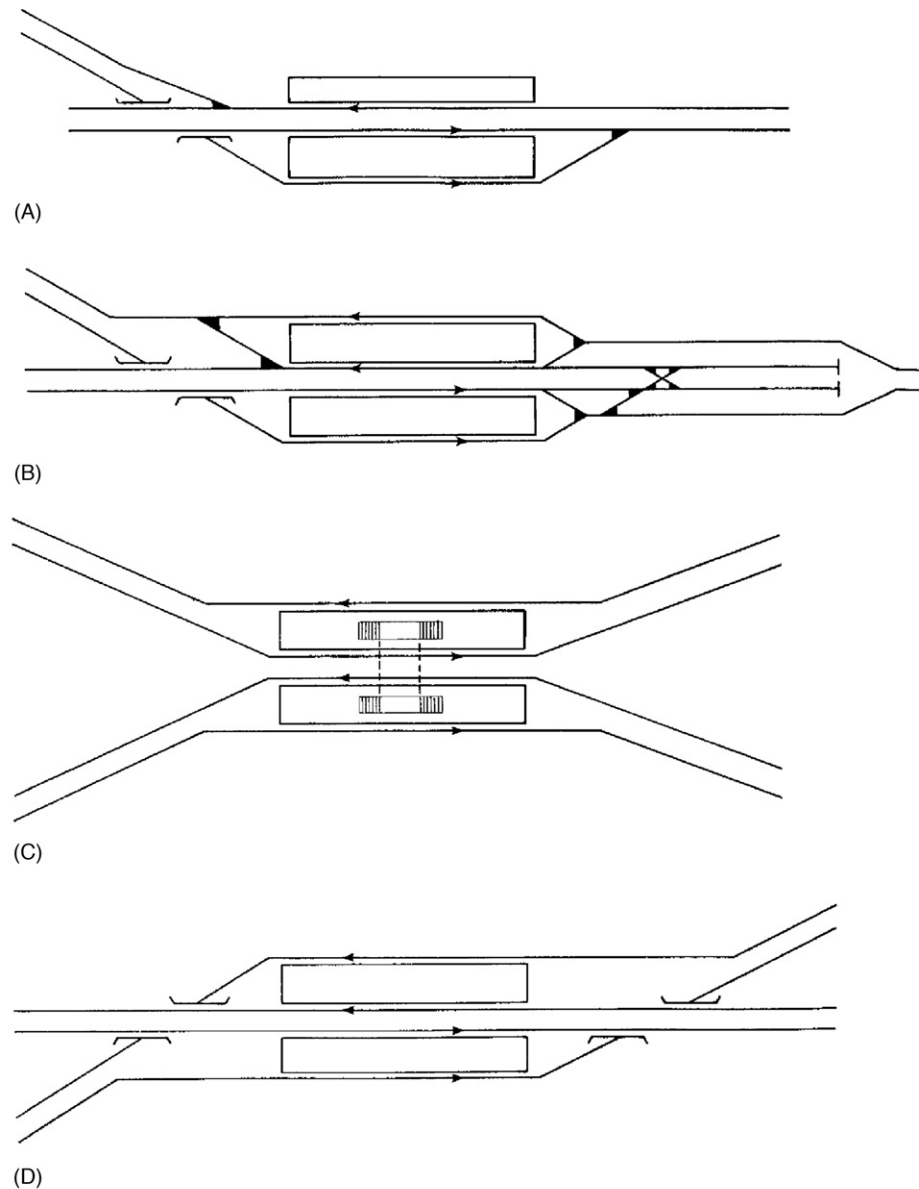


Figure 1 Platform arrangement and track layout of railway stations operated by line or direction. (A) Three-track station at a branching point, (B) four-track station at a branching point; storage tracks are shown, (C) joint station of two lines, and (D) "Weaving station" of two lines with simultaneous two-way transfers. *Source: Vuchic, 1981, p. 420*

System Specification

The functional specification of passenger railway stations should contain at least a description of the following principal elements:

- Location, level (rank) and position in the railway network and line,
- Passenger transport design volume and capacity,
- Kind, train frequency and speed of line services per day in each direction,
- Maximum length, loading gauge (profile) and axle weight of rolling stock,
- Number, length, width, and level above top of rail of the platforms,
- Track layout with number, gauge, spacing, design speed, maximum gradient of through, platform, shunting, stabling, maintenance and workshop tracks, crossings, and switches,
- Level, spacing and voltage of overhead wires, catenary or third rail,
- Track capacity,
- Maximum evacuation time from platforms in case of fire, blasts, and disaster.

The passenger volume of railway stations mainly depends on transport capacity, frequency, occupancy of the kind of train services and the number of lines operated.

The track capacity is defined as the maximum number of trains that may in theory be operated over one track section per period, usually per hour or per day. The theoretical track capacity is unlikely to be reached in practice due to timetable and safety margins, synchronization of train arrivals and departures, and waiting times during operations. Track capacity of single tracks is measured in both directions, while for double tracks it is measured per direction. The minimum headway time between a pair of trains on a platform track depends on the:

- Sequence of train arrival, passing and departure or vice versa,
- Speed, acceleration, deceleration, and length of trains,
- Signaling and safety systems,
- Volume, density, spatial distribution, and speed of passengers waiting, alighting and boarding.

Geometric Design

The design of railway stations is governed by their functional requirements, available space, urban master planning, environmental, economic, and architectural considerations. The geometric design of tracks, stairs and platforms in stations is closely interdependent with the (vertical) levels of the top of rail, platform, horizontal alignment, gradient and cross-sectional spacing.

Timetabling

The operational program determines the design of the station's infrastructure, including the arrangement and number of tracks and platforms. Timetables can be either periodic or nonperiodic. Integrated periodic timetables for railway networks are characterized by almost simultaneous arrival, and departure times respectively, of the different trains at the principal nodes of the network at regular intervals to offer synchronized smooth passenger transfer between lines. A comprehensive review of timetabling and capacity estimation methods for stations with multiple tracks such as Stuttgart 21 is given in Hansen (2017). Queuing models have been applied to estimate the waiting probability and queue length of the planned high-speed and regional express trains on the approaching routes and platform tracks, whereas microscopic simulation runs of the scheduled train services have been performed in order to estimate the track occupation times, number of platform tracks required and operations quality.

Track capacity consumption is estimated analytically by the mean of the track occupation times divided by the mean of the scheduled inter-arrival (headway) times per period for each (platform) track (Eq. (1)). The estimated theoretical track capacity (performance) is given by the length of the considered period divided by the mean minimum headway time (Eq. (2)). The practical track capacity is estimated by adding the mean buffer time to the denominator (Eq. (3)). The buffer time is a time margin on top of the minimum headway time to reduce the propagation of train delays. The distribution of buffer times in railway timetables and its impact on track capacity and quality of operations was first described comprehensively in Schwanhäuber (1974).

$$\text{Track capacity consumption } \rho = \frac{tB_m}{tA_m} \quad (1)$$

tB_m : mean track occupation time
 tA_m : mean inter – arrival time

$$\text{Theoretical track capacity } C = \frac{T}{tH_m} \quad (2)$$

T : Time period [min]
 tH_m : mean minimum headway time

$$\text{Practical track capacity } C_p = \frac{T}{tH_m + tR_m} \quad (3)$$

tR_m : mean buffer time

The Union of International Railways (UIC) recommends certain levels of track capacity consumption per kind of train traffic and period (Table 1). The corresponding values for junctions in station yards are reported in Table 2.

Table 1 Recommended levels of track capacity consumption by UIC (2013)

Type of line	Peak hour	Daily period
Dedicated suburban passenger traffic	85%	70%
Dedicated high-speed lines	75%	60%
Mixed traffic lines	75%	60%

Table 2 Recommended levels of track occupancy in station areas by UIC (2013)

Type of node area	Concatenated occupancy rate	Additional time rate
Switch area	60%–80%	67%–25%
Track area	40%–50%	150%...100%

The minimum headway time is equal to the minimum inter-arrival time between two train paths considering their requested speed profiles. The minimum headway time is determined in the first block section of that part of the infrastructure which is used in common by both trains. The minimum headway time between a pair of trains following each other on the same track section is equal to the maximum difference between the end of the blocking time of the preceding train and the start of the blocking time of the following train (Besinovic et al., 2016).

The blocking time is defined as the time interval during which a section of the track is allocated for the exclusive use of one train and is therefore blocked for other trains (Hansen and Pachel, 2014). The blocking time of a track section depends on the operations program, timetable, routing of trains, signaling and safety systems, and is computed as the sum of the switch time for setting up the route and corresponding signals, the train's running time over the sight distance (usually 300 m), the approach time, the running time over the nominal braking distance, block length, train length, safety distance, and the time to clear the route (Eq. (4)).

$$\text{Blocking time } tB = t_{\text{switch}} + t_{\text{sight}} + t_{\text{approach}} + \frac{l_{\text{block}} + l_{\text{train}} + l_{\text{safe}}}{v} + t_{\text{clear}} \quad (4)$$

t_{switch} : switch time for setup of the route and signals, t_{sight} : running time over the sight distance, t_{approach} : approach time $\geq \frac{v}{2a}$, t_{clear} : clearing time, v : speed, a : nominal braking rate

The number of platform tracks needed is equal to the mean track occupation time divided by the mean minimum headway time per peak hour (Eq. (5)). The timetable quality can be measured either by the standard deviation of the scheduled inter-arrival times divided by the mean of the interval times (Eq. (6)), or the standard deviation of the track occupation times divided by the mean (platform) track occupation time per track per period (Eq. (7)). The latter values correspond to the variation coefficient of the scheduled train interval V_A , and track occupation times V_B respectively.

$$\text{Number of platform tracks } N = \frac{tB_m}{tH_m} \quad (5)$$

$$\text{Timetable quality } q_A = \frac{sdev tA}{tA_m} = V_A \quad (6)$$

$$\text{Timetable quality } q_B = \frac{sdev tB}{tB_m} = V_B \quad (7)$$

Track occupation of individual track sections in multi-channel stations:

$$\psi = \frac{n \cdot tB_m}{s \cdot T} = \frac{\rho}{s} \quad (8)$$

n : number of train paths per period, s : number of service channels

Simulation

Simulation of passenger flows in railway stations is used to study the behavior of people when getting on and off the trains in order to evaluate the performance and safety of design options for access/egress and passenger processing (to/from gates, main ticketing facilities, corridors, stairs, escalators, lifts, and on platforms (Zou et al., 2019).

Railway timetables are simulated to calculate the running time of planned train services and analyze the impact of train performance characteristics, timetable options, infrastructure alternatives, and train operations on station and network performance. Railway simulation approaches can be classified according to their kind, scale, scope and data processing into:

- deterministic or stochastic,
- continuous or discrete,
- macroscopic or microscopic,
- synchronous or asynchronous

models (Watson and Medeossi, 2014).

Simulation is deterministic if all parameters are defined by the user and do not contain any random components. Stochastic simulations are based on probability distributions of design variables and are used to investigate the system behavior if the design parameters vary like the fluctuation of transport demand, train running, dwell and headway times. Continuous railway simulation

models are applied to compute the train motion in time and space, while discrete timetable simulation models are used to calculate the change of state variables at fixed time intervals (arrival and departure) or events (signal at danger/proceed/start braking, route set-up/release, train order).

Macroscopic simulation models apply a simplified railway infrastructure model, where the stations are represented as nodes and the links in between as arcs. Microscopic railway simulation models describe the detailed infrastructure, train movements, state of signaling and safety systems and timetable as directed graph. Macroscopic infrastructure and timetable simulation models use a scale of kilometers in space and minutes in time, whereas microscopic simulation models represent the infrastructure topology in centimeters and time in seconds. Based on rules of operation, timetable designers determine minimum headway times in macroscopic simulation models as input, while the required safety margins are automatically calculated in microscopic simulation models according to blocking time theory.

Synchronous timetable simulation models continuously estimate and monitor the state of the infrastructure and position/speed/acceleration/deceleration of every scheduled train in the network simultaneously at short time steps (usually one second) in order to test timetable feasibility, robustness against incidents and safety of train operations at microscopic scale. Asynchronous (discrete) timetable simulation models are used during timetable design and construction to evaluate the impact of different train priorities, speed levels, train orders, routing, platform allocation on track capacity and system performance.

Optimization

Optimization models aim to maximize or minimize an objective in the presence of complicating constraints. Exact search methods include linear programming (LP) and some forms of integer (IP) and mixed integer programming (MIP) which allow only little flexibility and are based on simplifying assumptions. Therefore, many optimization models apply heuristics for problem solving. Optimization methods can effectively support the search for finding (near) optimal solutions of a variety of planning problems concerning design of railway station and network infrastructure, timetabling and train operations (Corman, 2020). In contrast, iterative station location and network planning is very time-consuming and may lead to suboptimal solutions.

Borndörfer et al. (2018) give a comprehensive overview of current modeling and basic algorithmic solution approaches for the optimization of strategic railway network assessment, infrastructure management, timetabling, vehicle, crew and traffic management. Various optimization models are proposed for strategic railway network line planning, tactical timetabling problems, and operational railway traffic management (train routing, rescheduling). A detailed review of transit network design and scheduling literature and terminology is done in Guihare and Hao (2008).

Public transportation and railway line planning approaches can be divided into cost-oriented models, passenger-oriented, game-theoretic, location-based models, and heuristics (Schöbel, 2011). In general, the network and number of travelers between each pair of origin and destination station (OD), a set of potential lines, lower and upper frequencies are assumed to be given, while the objectives, constraints and solution methods vary a lot. Instead of choosing a subset of lines from a given pool, the lines can be newly constructed respecting rules or satisfying certain criteria. A generic, bi-objective model for integrating line planning, timetabling, and vehicle scheduling is presented in Schöbel (2017). The problem of simultaneous design of rail rapid transit network and line planning is modeled by Canca et al. (2017).

Train timetabling problem (TTP) approaches for stations with multiple tracks and differ with regard to the granularity of the infrastructure (microscopic and macroscopic models of junctions, signaling and interlocking routes) and/or discretization of running, dwell and track occupation times, as well as the choice of objectives, constraints and solvers corridors (Burggraefe and Vansteenwegen, 2017; Cacchiani et al., 2016; Feng et al., 2018; Ghaemi et al., 2017; Pellegrini et al. 2014; Qi et al., 2016; Sama et al., 2017; Sels et al., 2014). Besinovic et al. (2016) propose an integrated hierarchical approach for the TTP, which combines microscopic and macroscopic models of the network infrastructure and train schedules.

Conclusions

The knowledge of planning, design, analysis, and timetabling of railway stations and networks has grown considerably. Innovative analytical, simulation, and mathematical programming approaches have been developed quickly and in practice during the past fifty years. Powerful and fast electronic data processing, transmission and computation technologies are used to increase the accuracy, effectiveness, reliability, efficiency and quality of railway planning and train operations. Moreover, a lot of intelligent mathematical algorithms and programming methods can support rational decision making on optimized railway network infrastructure investment, line planning, timetabling, train operations, and traffic management.

The main research challenge in this area is now to bridge the gap between academics, railway professionals and politicians by more customer-friendly, transparent, accurate and integrated modeling of passenger flows and train traffic in railway stations and networks. Easy-to-understand decision support tools for planners, timetable designers, railway passengers, operators, governments, and other people affected are required to find best solutions that balance common and diverging interests of stakeholders. Modeling the impact of multimodal transport flows in large networks on the design of sustainable railway transport infrastructure and mobility service networks which achieve the desired reduction of fossil energy consumption, carbon and noise emissions remains a big task.

References

- Besinovic, N., Goverde, R.M.P., Quaglietta, E., Roberti, R., 2016. An integrated micro–macro approach to robust railway timetabling. *Transport. Res. Part B* 87, 14–32.
- Borndörfer, R., Klug, T., Lamorgese, L., Mannino, C., Reuther, M., Schlechte, T. (Eds.), 2018. *Handbook of Optimization in the Railway Industry*. In: *International Series in Operations Research & Management Science* 268. Springer.
- Burggraeve, S., Vansteenwegen, P., 2017. Robust routing and timetabling in complex railway stations. *Transport. Res. Part B* 101, 228–244.
- Cacchiani, V., Furini, F., Kidd, M.P., 2016. Approaches to a real-world Train Timetabling Problem in a railway node. *Omega* 58, 97–110.
- Canca, D., De-Los-Santos, A., Laporte, G., Mesa, J.A., 2017. An adaptive neighborhood search metaheuristic for the integrated railway rapid transit network design and line planning problem. *Comput. Oper. Res.* 78, 1–14.
- Corman, F., 2020. *Railway Traffic Management*, in: *International Transportation Encyclopedia*.
- Crainic, T.G., 2020. *Freight railway service networks*, in: *International Transportation Encyclopedia*.
- Dobruszkes, F., and Moyano, A., 2020. *Geography of rail transport*, in: *International Transportation Encyclopedia*.
- Feng, Z., Cao, C., Liu, Y., Zhou, Y., 2018. A multiobjective optimization for train routing at the high-speed railway station based on TabuSearch Algorithm. *Math. Probl. Eng.* 22, doi:10.1155/2018/8394397.
- Ghaemi, N., Cats, O., Goverde, R.M.P., 2017. A microscopic model for optimal train short-turnings during complete blockages. *Transport. Res. Part B* 105, 423–437.
- Guihare, V., Hao, J.-K., 2008. Transit network design and scheduling: a global review. *Transport. Res. Part A* 42, 1251–1273.
- Hansen, I.A., 2017. Review of planning and capacity analysis for stations with multiple platforms – Case Stuttgart 21. *J. Rail Transport Plan. Manage.* 6, 313–330.
- Hansen, I.A., Pacht, J. (Eds.), 2014. *Railway Timetabling & Operations Analysis Modelling Optimisation Simulation Performance Evaluation*. Eurailpress, Hamburg.
- Monzon, A., and Lopez, E., 2020. *Rail transport and territorial scale*, in: *International Transportation Encyclopedia*.
- Papa, E., Bertolini, L., 2015. Accessibility and transit-oriented development in European metropolitan areas. *J. Transport Geogr.* 47, 70–83.
- Pellegrini, P., Marlière, Rodriguez, J., 2014. Optimal train routing and scheduling for managing traffic perturbations in complex junctions. *Transport Res. Part B* 59, 58–80.
- Qi, J., Yang, L., Gao, Y., Li, S., Gao, Z., 2016. Integrated multi-track station layout design and train scheduling models on railway corridors. *Transport. Res. Part C* 69, 91–119.
- Samà, M., Pellegrini, P., D'Ariano, A., Rodriguez, J., Pacciarelli, D., 2017. On the tactical and operational train routing selection problem. *Transport Res. Part C* 1–15.
- Schöbel, A., 2011. Line planning in public transportation: models and methods. *OR Spectrum* 34 (3), 491–510.
- Schöbel, A., 2017. An eigenmodel for iterative line planning, timetabling and vehicle scheduling in public transportation. *Transport. Res. Part C* 74, 348–365.
- Schwanhäußer, W. 1974. *Die Bemessung der Pufferzeiten im Fahrplangefüge der Eisenbahn*. Veröffentlichungen des Verkehrswissenschaftlichen Instituts der RWTH Aachen 20.
- Sels, P., Vansteenwegen, P., Dewilde, T., Cattrysse, D., Waquet, B., Joubert, A., 2014. The train platforming problem: the infrastructure management company perspective. *Transport. Res. Part B* 61, 55–72.
- Union Internationale des Chemins de Fer (UIC), 2013. 2nd edition. *UIC Code 406 - Capacity*, 2nd edition, Paris, pp. 1–51.
- Vuchic, V.R., 1981. In: *Urban Public Transportation Systems and Technology*. Prentice-Hall, Englewood Cliffs N.J., pp. 416–421.
- Watson, R., Medeossi, G., 2014. Simulation. In: Hansen, I.A., Pacht, J. (Eds.), *Railway Timetabling & Operations* Chapter 6. Eurailpress, Hamburg, pp. 191–215.
- Zou, Q., Santos Fernandez, D., Chen, S., 2019. Agent-based evacuation simulation from subway train and platform. *J. Transport Saf. Security*, doi:10.1080/19439962.2019.1634661.

Introduction to Rail Transport

Chris Nash, Tony Fowkes, Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	406
Rail Technology	406
Long Distance Passenger Trains	408
Other Passenger Traffic	409
Freight	410
Organization	411
Conclusion	412
References	412
Further Reading	412

Introduction

Rail transport in the world today plays a varying role according to the characteristics of the country (Table 1). In the EU, it plays a specialist role in the transport market, while road dominates. However, in some of the world's largest countries, rail plays a much greater role. The rail share of the passenger market is much higher in China and Russia than in average sized countries. Similarly the shares of the freight market in the United States and Russia are very high. Note the exceptions to this rule. Passenger mode share in the United States is very low due to competition from air and high levels of car ownership, and the freight share would be higher in China if not for the importance of coastal shipping.

The principal aim of this chapter is to establish what determines the role that rail plays in the transport market. In the next section we provide a brief and highly simplified overview of rail technology and its capabilities. We then examine the different markets for rail. We consider in turn long distance passenger, other passenger traffic, and freight. We then briefly consider railway organization before concluding.

Rail Technology

Rail transport comprises trains of vehicles running on a track, which guides them. Usually, the wheels and track are steel, although on some metro systems, particularly in France, they comprise rubber-tired wheels running on concrete. The advantage of steel is low friction; its main disadvantage as a result is that braking takes a relatively long distance, and thus except on low speed tramways some form of signaling is needed. Railways normally run on the principle that two trains should never be closer than the braking distance of the rear train (plus a safety margin, or "overlap," to cope with factors such as reduced adhesion due to rail surface condition, or worn brakes), and it is the function of the signaling system to achieve this. Traditionally, rail routes have been divided into sections called blocks, and the signaling ensures that trains are always at least one block apart and that the rear train is not traveling at a speed such that stopping within one block is not possible. This was achieved by lineside signals, which told the driver whether it was safe to proceed at full speed, safe to proceed at reduced speed, or whether he/she should immediately stop. More recently, such systems have been backed up (or replaced) by electronic signals transmitted to the cab, which would automatically apply the brakes if the train was deemed to be traveling too quickly to stop in good order before the start of the safety overlap.

On a system with lineside signals, the capacity of a line is determined by the distance signals are apart. For instance, on a line with 3 signaling aspects (clear, caution, and stop) trains traveling at full speed must be at least two block sections apart and must be able to stop within those two sections. The minimum headway between trains in minutes is thus the time a train takes to travel those two sections. The capacity of a section of line per hour is 60 divided by the headway in minutes. When calculating the capacity of a line we need to consider differences in travel times in different sections, and the stopping patterns of trains at any intermediate stations (Hansen and Pacht, 2014).

High speed lines (i.e., lines designed for speeds of 300 kph and above) invariably have signaling systems where electronic instructions are transmitted to the driver's cab rather than relying on sighting lineside signals. It is the wish of the European Union to see such a system extended, at least to all main lines, using the European Rail Traffic Management System (ERTMS). In its more advanced form, ERTMS will have moving block signaling, under which signals are transmitted continuously to the train, rather than at the beginning of fixed blocks. These signals show the speed at which it is safe for the train to proceed given the locations of it and other trains on the system. From this it is a small step to complete automation of train driving and many urban metros are already completely automated. However, unless the railway is completely segregated (e.g., in tunnel or elevated) it is necessary to have some way of detecting obstacles on an automated system.

Table 1 Rail mode share (%) 2015

	<i>Passenger km</i>	<i>Freight ton km</i>
EU	8.3	11.3
US	0.1	33.1
China	40.2	12.8
Russia	33.4	45.3

Notes: Passenger km are the sum of those by all motorized modes of transport, including for the EU all intra EU trips; other than this, international traffic is excluded. Rail includes metro, tram, and light rail.

Freight ton km are the sum of those by road, rail, pipeline, inland waterways, air, and sea including for the EU all intra EU traffic; otherwise international traffic is excluded.

Source: Derived from EU Pocketbook. EU Transport in Figures, 2018.

Having considered the capacity of a line in terms of numbers of trains per hour, we need to consider the capacity (for passengers, freight, or whatever) of each of those trains. First, we shall consider train length. Given suitable terminals, sufficient rolling stock, and haulage power, the principal determinant of permitted train length is the need for trains to pass each other, either in the same or opposing directions. For single-track lines, passing points will be created at which one train can be parked while another passes. Where, given its degree of importance, a train is deemed likely to be looped there for another to pass, the length of that loop limits the length of the train. The lowest of all such limits for the loops on the line gives the overall permitted length limit. Similarly, on a multiple track section, if a slow train has to be parked (in a loop or siding) so as not to delay a following faster train, the length of that loop or siding also feeds into the length limit calculation. For example, on a line where the only overtaking points are at stations, the published length limit for the line will be the length available at the “shortest” station. Trains longer than the published length limit can be permitted to run if they will never need to be overtaken (or passed on a single track line).

Second, the maximum height and width of trains is determined by the height of over-bridges and tunnels and the gaps to other structures including station platforms. Subject to this, it is feasible to run double deck passenger trains and double stack container trains (i.e., with a second layer of containers on top of the first). Passenger trains of up to 30 coaches run in countries such as China and freight trains of up to 20,000 tons in various places. France runs many high speed trains using two double deck units coupled together with a total capacity of around 1000 passengers, while high density suburban rolling stock with provision for standing passengers may accommodate many more. So a total capacity for a double track railway of 20–30,000 passengers per hour in each direction is perfectly feasible and some metros regularly exceed this.

Thirdly, we consider how much can be loaded onto each vehicle. The maximum weight of a rail vehicle is determined by the permitted axle weight of the route, which in turn depends on the strength of the rail used and the structures crossed (such as “under-bridges”). For example, in the EU the standard maximum axle weight is 22.5 tons. So a 4-wheeled wagon with tare weight 22 tons evenly loaded with stone has a capacity of $90 - 22 = 68$ tons. Railways built solely to move heavy freight use much heavier rail and more robust construction and so have much higher axle weight limits. Conversely, as the name implies, Light rail systems are much more lightly engineered and have lower axle weight limits.

Naturally, the capacity of any railway depends on the investment made and how well its operation is planned. Obviously railways require installation of track and (usually) signaling, and that is expensive in terms of first cost, maintenance, and renewals. On the other hand, they have a large potential capacity. For instance, a two-track main line signaled for 3-min headways has a potential capacity of 20 (similar speed) trains each way per hour. Normally a margin of capacity is left free for recovery from delays, but with 3-min headways 14 trains per hour is common and up to 18 is practical as long as the system runs reliably. Given their relatively low speed, metros often have lower headways of the order of a minute and correspondingly higher frequencies than this.

Very high frequencies of service are generally only possible where all trains have identical characteristics in terms of speed (inclusive of station stops), acceleration, and decelerations. Otherwise a fast train following closely behind a slow train would very soon catch it up and be delayed, unless there were facilities for overtaking. It is clear that a large gap must be left before a fast train may follow a slow train if it is not to be delayed. For instance, suppose in the example of a two track main line signaled for 3 min headways fast and slow trains follow each other for 10 km (the slow trains may be freight, or passenger trains making intermediate stops); fast trains take 6 min; slow 20 min. If fast and slow trains alternate, capacity is reduced from 20 trains per hour to only 6 trains per hour, with departure and arrival times as follows: [0, 6], [3, 23], [20, 26], [23, 43], [40, 46], [43, 63] then [60, 66], etc. Capacity is also reduced by junctions (unless these are grade separated to avoid conflicting movements on the level).

On lower volume routes, the cost of track and signaling may be reduced by using a single track, but this has the disadvantage that trains in one direction have to enter a passing loop for trains in the opposite direction to pass. It is thus the travel time between passing loops that dictates the capacity of the system. Because of the delays to trains at passing loops and the risk of late running trains in one direction delaying those in the other, single track is rarely used for routes with more than 1 or 2 trains per hour in each direction.

Trains themselves may comprise a locomotive pulling passenger coaches or freight wagons or (mainly in the case of passenger trains) self propelled vehicles, which may be coupled to form longer trains (multiple units). In either case, they may be powered by steam, diesel, or electric traction. Except for short distances, where batteries may be used, electric traction requires provision of

current by means of overhead wires or third rails, incurring a capital cost roughly proportional to the length of track electrified. On the other hand, electric traction provides a clean source of traction at lower weight relative to its power and with lower maintenance and typically fuel cost than diesel. These benefits are roughly proportional to the number of trains run. So the case for electrification depends on the density of traffic; electric traction is the norm on more densely used railways.

Given the high capital cost, railways generally require high volumes of traffic to be economically viable. Stations are also expensive and stopping at a station slows the train and reduces capacity of the line. Therefore typically there will be a limited number of stations and freight terminals, and a need for passengers and goods to be taken to them by another mode, compared to the door to door convenience of road transport (although some freight transport does move direct between private sidings). As a consequence, rail transport will only be chosen when there is enough advantage (in terms of cost or speed) in using it for the trunk part of the journey to justify the cost and inconvenience of the feeder journeys and transfers between modes. This tends to mean that rail is more competitive with road for longer journeys.

Long Distance Passenger Trains

Traditional long distance rail services operate at up to 200 kph, giving a commercial speed usually not more than 150 kph allowing for starting and stopping. Many are slower due to constraints of the infrastructure, particularly curvature. But there is also time involved in getting to and from the station and waiting for the train. Even if that is as little as 1 hour, it still means that a 150 km journey by rail will take at least 2 h. If door to door road transport on high quality roads is available, such that car transport can average 100 kph door to door, this journey may be undertaken more quickly and conveniently by car; it is only for journeys above 300 km that rail will be quicker (Fig. 1 F). In this situation, unless road transport is more expensive than rail, it will only be those without access to a car or who value rail for other reasons (such as comfort or the ability to work on the train) that will consider rail. Where road infrastructure is good, even bus transport may be able to compete with rail on journey time, especially in cases where rail speeds are substantially below 200 kph. Of course, it is not usually possible to average 100 kph by car and certainly not by bus. Usually, there will be a part of the journey on lower quality or congested roads and sometimes the main roads themselves will be congested or of poor quality. Thus the ability of rail to attract passengers who could use car or bus transport depends very much on not just the quality of the rail service but also that of the roads.

Use of air transport usually involves a longer journey to and from the airport than for rail, plus much longer at the airport going through security and waiting for boarding. Suppose that the sum of these times is 2 h and that the flight time is typically half an hour, increasing little with distance. Then traditional rail will be faster than air door to door for journeys of up to 225 km, but beyond that air will start to compete on journey time, although many passengers seem to regard rail as more comfortable and convenient than air so it may be only above around 300 km that air starts to become important.

On some routes new high speed lines have been built with a maximum speed of up to 320 kph. Suppose that the rail service is provided by a new high-speed line with a commercial speed of 250 kph. This will enable rail to dominate other modes in terms of journey time over a wide range of medium distances. Even on the assumption that car can average 100 kph, rail will be faster than car for distances above around 170 km. More importantly, it will be faster than air door to door for distances of up to 375 km. In fact, evidence (Table 2) suggests that high speed rail can compete with air for greater distances than this; typically where rail station to station journey time is 3 h, rail has more than half the combined air-rail market and a substantial share at 4 h. This may relate to a distance approaching 1000 km. However, this evidence comes from Europe and Asia, where cities are dense, well served by public transport and therefore the central railway station is more convenient than the airport for most passengers. It might not apply to the less dense cities of North America and Australia.

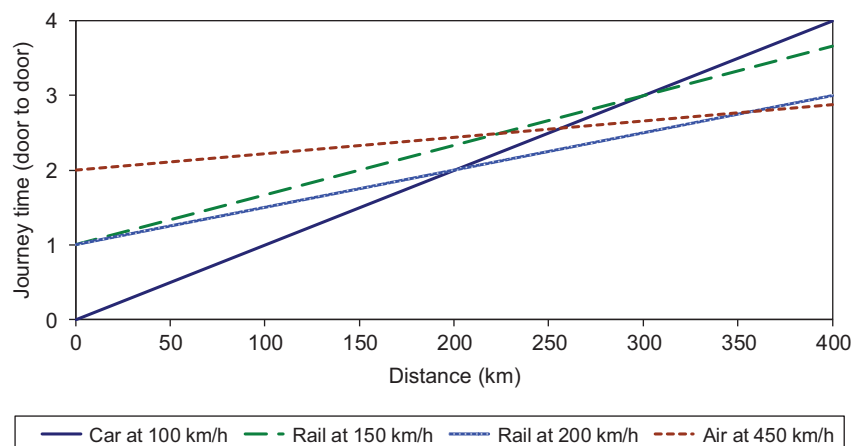


Figure 1 Breakeven distances for journey time by intercity rail.

Table 2 Rail share of rail/air market and rail station to station journey times

Corridor	Year	Rail Travel time	Rail share (%)
Paris–Brussels	2006	1 h 25 min	100
Paris–Lyons	1985	2 h 15 min	91
Madrid–Seville	2003	2 h 20 min	83
Brussels–London	2005	2 h 20 min	60
Tokyo–Osaka	2005	2 h 30 min	81
Madrid–Barcelona	2009	2 h 38 min	47
Paris–London	2005	2 h 40 min	66
Tokyo–Okayama	2005	3 h 16 min	57
Paris–Geneva	2003	3 h 30 min	35
Tokyo–Hiroshima	2005	3 h 51 min	47
Paris–Amsterdam	2004	4 h 10 min	45
Paris–Marseilles	2000	4 h 20 min	45
London–Edinburgh	1999	4 h 25 min	29
London–Edinburgh	2004	4 h 30 min	18
Tokyo–Fukuoka	2005	4 h 59 min	9

Source: Nash, 2013.

Table 3 CO₂ emission by mode (kg per 100 pass km)

Car (fleet average) with occupancy of 2	0.075
Car (best) with occupancy of 2	0.057
Double deck motorway coach at 60% load factor	0.030
Air (500 km flight) at 75% load	0.100
Electric high speed train at 70% load factor (British mean electricity mix)	0.050

Source: Derived from CILT, 2011.

Of course, the above are only examples. Circumstances differ, and there will always be some individuals for whom the rail station or the airport are more or less convenient than assumed above. But these simple examples do illustrate that conventional rail struggles to compete when faced with air services and high quality uncongested roads. It may play a much greater role in the long distance travel market where roads are congested or poor quality, or where car ownership is low. High-speed rail is likely to be competitive for medium distance journeys even where roads are high quality and uncongested and air services exist. Rail may also offer benefits in terms of reduced congestion and environmental benefits including reduced global warming (Table 3 gives estimates of emissions for Britain, at a time when coal and gas were still extensively used for electricity generation; as electricity production is decarbonized so the advantage of rail will rise). It should be noted that, with the sort of load factors achieved on the French TGV network and Eurostar services between London and Paris, high speed rail still has an advantage over car, and more so air. But of course, high-speed rail is very expensive to build (typically earlier routes in Europe cost 5 m–30 m euros per route km, with more recent routes tending to cost more as environmental standards tighten) and requires a substantial volume of traffic to be economically justified.

Other Passenger Traffic

Here we look at commuter services into large cities and cross-country services linking smaller towns to cities and to each other. They offer advantages to users over buses in terms of comfort and speed, especially where buses are on poor quality or congested roads, and similarly may attract traffic from car, relieving congestion, and pollution. But they may be very expensive per passenger where volumes are low. Thus they are typically only an economically viable option, even in social cost benefit terms, where volumes are relatively high, particularly into large cities.

Metros are rail systems found in large cities, usually run underground in the city center and operate at high frequencies with frequent stops. By contrast, trams generally operate on street in mixed traffic. Light rail is somewhere between the two, able to run on street, though usually with reserved lanes or priority over other traffic, but also with sections on segregated track or in tunnels. This may make it substantially cheaper to install than metros or conventional railways (Grimaldi et al., 2019).

Table 4 summarizes the capabilities of the alternative modes. Again, circumstances differ and it is possible to argue with the specific figures given for speed and capacity, but the pattern is clear: rail tends to have advantages where the desire is for high speeds (usually only needed where distances are longer) and high capacity. Again these factors dictate a particularly important role for rail in

Table 4 Characteristics of different urban public transport modes

Mode	Capacity of vehicle/train	Headway (mins)	Hourly capacity (000s)	Commercial speed incl stops (kmph)
Ordinary bus	75–100	1	4.5–6	20–30
Busway	60–75	1	3.6–4.5	60
Articulated bus	75–150	1	4.5–9	60
Bus in convoy	70–105	8 s	16–27	20
Tram	150–180	1	September 22	20
Light rail	180–210	1.5	July 25	30–40
Metro	400–800	1.5	16–32	40–50
Heavy rail	1000	March 2	20–30	50–60

Source: Adapted from *CILT*, 1996.

large cities. But segregated busways may rival rail where there is space to fit them in, and do have the advantage that buses can leave them in the suburbs to serve places to which it would not be economically justified to build a fixed track.

Rail services other than inter city almost invariably need subsidies to survive (though Japan, with its exceptionally high population density, and high costs of motoring and parking, has extensive regional and commuter services that are unsubsidized; these do often benefit from the involvement of the parent company in property markets). Governments often regard such subsidies as worthwhile because of the resulting reductions in congestion and pollution from cars, particularly as only a few cities worldwide have appropriate systems for charging cars for these costs. More recently there has also been a great deal of interest in the wider economic benefits of urban and suburban rail, with a belief that they raise productivity and attract economic development, leading to a further argument for subsidizing rail.

Freight

Traditionally, a large proportion of rail freight traffic has been handled by wagonload services, whereby individual wagons containing consignments for separate customers are loaded at terminals, taken to marshalling yards to be shunted into trains for a yard near their destination, and remarshalled into trains for the destination terminal. Such services were typically slow and unreliable, particularly if freight had to be transhipped from and to road transport for the collection and delivery at each end. While such services may be economically worthwhile if the speed and cost of rail for the trunk haul is well below road, these services have been in severe decline in Europe in recent decades. Usually they are only competitive for long distance traffic, such as can be found in North America.

With the growth of containers and swap bodies, a more efficient way of transferring freight between modes has come to dominate unit load traffic, and if the containers are transferred at large terminals then a more extensive range of through services without intermediate marshalling becomes possible. In the case of maritime traffic, much is already in containers and has to be transhipped between ship and another mode at the port in any case, so rail is more competitive for this than for purely domestic traffic.

Where volumes permit, there has also been a concentration on running complete freight trains for a single customer without intermediate marshalling. Typically this approach is used for bulk commodities such as coal, ores, stone, and oil. Where the train is running between private sidings at the premises of the customer and their supplier (for instance coal from a coal mine or a port to a power station), rail is competitive for this traffic over any distance.

A typical rail journey consists of some or all of the components illustrated in [Fig. 2](#).

By comparing road and rail costs for journeys of differing distances we can establish the breakeven distance ([Fig. 3](#)). For any distance shorter than the breakeven distance, road haulage charge is likely to be cheaper, but for distances over the breakeven distance rail should offer the best price. Breakeven distances are affected by the type of rail journey considered and the types of goods being hauled. Breakeven distances do not take into account the quality of service offered by each mode, with road generally being favored all else equal.

[Fig. 3](#) shows distances between 200 and 425 km, and compares road costs with 3 rail scenarios. The bottom line shows a private siding to private siding movement, with no marshalling costs or collection and delivery costs. It is always cheaper than road, but may not be so convenient. The top line shows a traditional collection and delivery movement. There are four marshalling operations, two lifts, two rail trip legs and road legs at both ends of the journey. For all distances shown it is much more expensive than road, with breakeven not occurring until around 1000 km. In the middle of the diagram is shown the total road cost line intersecting at around 300 km with the cost line for a movement from private siding to delivery point. This journey involves two marshalling operations and one lift, and a rail trip leg and road leg at the collection end of the journey. Rail is then seen to be competitive on price above this breakeven distance of 300 km. In each case we assume the trunk rail leg to be 20 km shorter than the overall road journey distance, and where applicable, the road collection/delivery distances to be 10 km and each rail tripping movement to be 30 km.

Relative costs will vary over time and from country to country so, although based on real UK data, this diagram can be taken to be nothing more than illustrative.

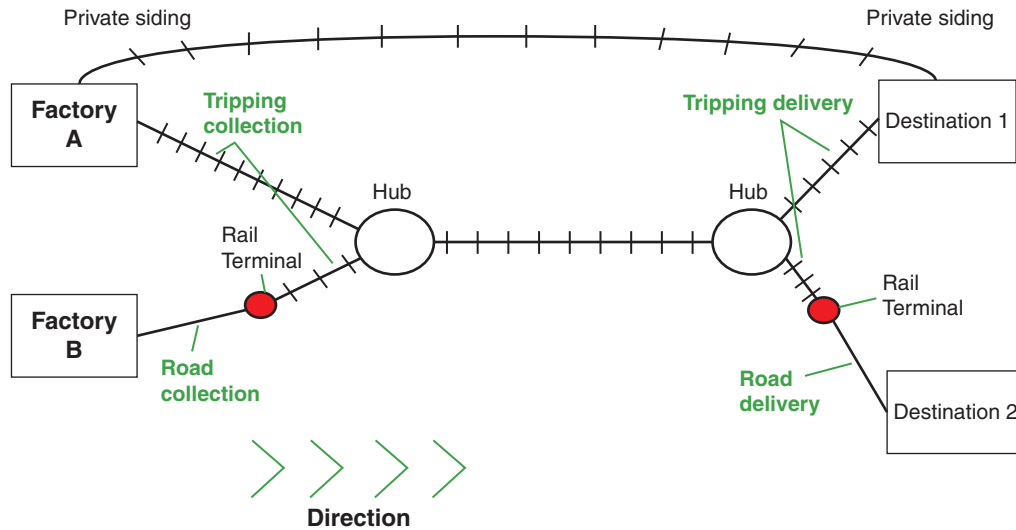


Figure 2 Components of a rail freight journey. Source: Fowkes et al., 2006.

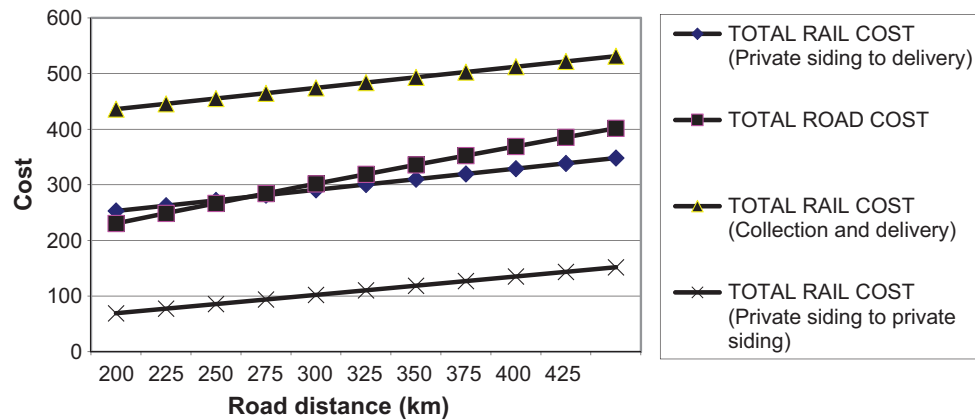


Figure 3 Illustration of rail and road costs by distance for a given payload. Source: Fowkes et al., 2006.

In many countries, bulk traffic is in decline, and the growth areas are services and high value manufactured goods, which need to be delivered on a just-in-time basis. This is a challenge for railways, but intermodal services are a growth area with the greatest potential for rail freight. Rail freight is typically not subsidized, although in some countries it pays only its marginal cost to use lines mainly dedicated to passenger services. Obviously an important factor determining its profitability is the strength of the competition from road freight, which depends both on the quality of the road system and on the system for charging for the use of roads. While the European Union favors a distance based charging system differentiated according to both the impact of the vehicle on the roads and on the environment, few countries have so far introduced such a system.

Organization

In many countries, railways were originally built by private companies but there are now few countries where rail infrastructure is privately owned. It does remain the case that most railways are privately owned in North America, where railways are predominantly used for freight, and in Japan, where the three largest passenger companies have been privatized after many years of public ownership, and where there are also many privately owned smaller commuter railways. These privately owned railways run largely without government subsidy, freight traffic in North America and passenger traffic in Japan being dense enough to support this. However, in both cases services for the minority form of transport—passenger in North America and freight in Japan—are operated by government owned companies over the tracks of the private companies.

Some of the largest railways in the world—in India, China, and Russia—remain government owned vertically integrated companies. But in Europe, while infrastructure remains a government owned monopoly, it is the policy of the European Union

to encourage competition in the provision of train services, either by open access entry (typically some 30% of freight traffic is handled by new entrants while a handful of countries—notably Italy, the Czech Republic, and Sweden—have significant open access entry in inter city passenger services) or by franchising by means of competitive tenders (here Britain, Sweden, and Germany are the main users of this approach) (Thompson, 2003). There is also experience of franchising freight services in South America, where long vertically integrated franchises have been used as a way of revitalizing collapsing government owned freight railways.

Conclusion

Rail transport still plays a very important role in freight and/or passenger transport in some of the largest and most important countries in the world. In many other countries, it plays an important role in specific markets, such as passenger traffic into and between the largest cities and the inland collection and distribution of maritime containers. Rail transport is particularly suitable for longer distances and for larger volumes. While in the past, railways were almost invariably government owned outside North America and Japan, there is now a greater variety of organizations, with private sector open access operators or franchisees playing a bigger role in the provision of rail services.

References

- Chartered Institute of Transport, 1996. *Better Public Transport for Cities*, CILT, Corby.
Chartered Institute of Transport, 2011. *Transport Use of Carbon Report Appendix C*, CILT, Corby.
Fowkes, A.S., Johnson, D.H., Whiteing, A.E., Nash, C.A., 2006. *Freight Modeling Final Report*, Rail Research U.K. Report 3/5, ITS. University of Leeds, UK.
Grimaldi, R., Beria, B., Laurin, A., 2014. A stylized cost-benefit model for the choice between bus and light rail. *J. Trans. Econ. Policy* 48 (Part 2), 219–239.
Hansen, I.A., Pachi J., 2014. *Railway Timetabling and Operations*. Hamburg: Eurail Press.
Nash, C., 2013. When to invest in high speed rail. Discussion paper 25, International Transport Forum, OECD, Paris.
Thompson, L.S., 2003. Changing railway structure and ownership: is anything working? *Transp. Rev.* 23 (3), 311–355.

Further Reading

- Ledbury, M., Schänzler, E., Ludewig, J., Dimas, S., Barrot, J., 2008. *Rail Transport and the Environment: Meeting the Challenge*. Community of European Railways Infrastructure Companies: Local Transport Today, Brussels.
Nash, C., Wardman, M., Button, K., Nijkamp, P., 2002. *Classics in Transport Analysis-3*. Edward Elgar, Cheltenham.
Harris, N.G., Godward, E.W., 1992. *Planning Passenger Railways*. Transport Publishing Company, Glossop.
Profillidis, V.A., 2014. *Railway Management and Engineering*. fourth ed., Ashgate, Farnham.

The History of Rail Transport

Carlo Ciccarelli*, **Andrea Giuntini†**, **Peter Groote‡**, *Department of Economics and Finance, University of Rome Tor Vergata, Rome, Italy; †Department of Economics, University of Modena and Reggio Emilia, Modena, Italy; ‡Department of Cultural Geography, University of Groningen, Groningen, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Revolution or Continuity?	413
The Debut: Stockton-Darlington	413
The First Developments	414
Technological and Labor Issues	414
The Local and the National	414
The Maturity Phase	415
Regulation, Standardization, and International Agreements	415
The Journey by Rail	416
The 20th Century Decline	416
A Resurrection of Rail Transport?	417
Railways and Economic Development: A Review of the Literature	417
References	425
Further Reading	425

Revolution or Continuity?

Questions concerning continuity and discontinuity are often asked in social sciences. The field of transportation economics makes no exception. For some historians the railways are part of a continuous process of improvement of transport and communications that accelerated with the first industrial revolution (see [Atack, 2019](#) for a discussion; [Fogel, 1964](#)). After all, railways were a combination of already known principles and techniques, which competed with already existing transport modes. Horse-drawn railways over wooden tracks to transport coal from coal deposits are clear predecessors of modern-day railways. The improvements that Richard Trevithick developed at the beginning of the 19th century, among which the use of high-pressure engines and iron tracks, gradually led to the first modern locomotive.

For other historians instead, the development of railways was more than that, and in fact a true transport revolution (see [Atack, 2019](#) for a discussion; [Rostow, 1960](#)). The quantitative and qualitative level of the changes and the intensity in which they were integrated and institutionalized was truly revolutionary. From a technological point of view the extraordinary novelties brought by the railways are evident, in particular if we refer to locomotives. The commitment of capital to build the railways was unprecedented ([Chandler, 1977](#); [Mitchell, 1964](#)). And the engineering challenges to build railways in mountainous areas dwarfed even those linked to the construction of canals, which represented an important ingredient of the first industrial revolution. From the physical point of view, not just the speeds, but also the strict scheduling that railways offered created true networks of transport and communication on the national scale. Due to the railways, many countries had to switch from many regional time zones to one national system ([Smith, 1976](#)). The railways also were the first large-scale technical system that triggered a total turnaround of economic interests, stimulating enormous financial investments. They influenced the structure of the early mature capital markets, inducing modern stock exchange practices and favoring the invention of new financial products, although also creating new stimuli for speculation and the creation of stock market bubbles. And, obviously, railways changed the geographies of the world. They influenced population density patterns and the concentration of entrepreneurial capital as well as the opening of new territories to cultivation and residential living ([Alvarez et al., 2013](#)).

The railway became a symbol of progress in itself. Few objects visualized progress in the way the railway did, standing for the complete disruption of the relationship between man and space, developing a new concept of mobility. Steam, iron, coal, and the genius of man that pushed the boundaries of technical knowledge ever further: all this finds its ideal synthesis in the picture of a running train, as in celebrated paintings such as William Turner's *Rain, Steam and Speed* (1844), and John Gast's *American Progress* (1872), where trains almost literally push civilization forward. What used to be far away and uninhabitable was no longer so thanks to the train; the formerly impossible was brought into the sphere of everyday opportunities.

The Debut: Stockton-Darlington

In the English case, the introduction of the railway linked with the economic needs of an already industrializing country. In other countries, the reasons for railway building often included political and cultural reasons, alongside economic ones. These could be linked to desires for prestige, for creating unity and a national identity (railways as an instrument for nation-building), direct military purposes or general sociocultural modernization.

In England, the opening of the Stockton–Darlington line on September 27, 1825 is commonly considered the first act of the railway era, although on part of the line stationary steam engines were still used to pull wagons. The locomotive used in this occasion

was built by George Stephenson, reputedly the father of railways, and had the simple name of *Locomotion*. It traveled the route at an average speed of 20 kilometers per hour. On the inaugural train, as reported by enthusiastic chronicles, at least 500 people mounted the train, partly on board and partly clinging outside the 33 wagons. Four years later, in 1829, the other major event of the first railway era took place: the Rainhill race. A speed race among locomotives was organized a few kilometers outside of Manchester. Stephenson, now at the helm of his famous *Rocket*, won the 500 Pound Sterling at stake. The *Rocket* towed a convoy of twenty wagons at an average speed of 25 kilometers per hour; in some places the train touched almost 50. Stephenson won by virtue of the introduction of a 2-meter long tubular boiler built by the French engineer Seguin. The following year the entire line between Manchester and Liverpool was opened, including a viaduct with 9 arches with a height of almost 22 meters. Around the same time, Stephenson founded his first locomotive factory. Although euphoria was first mixed with mistrust, the success of the railways was enormous and in a few years the country was crossed by numerous lines, built by competing companies. The lines were not yet integrated into a proper network, but remained rather fragmented.

The railways quickly conquered continental Europe as well, despite resistance of those with opposing interests, the huge effort to raise enough capital and technical difficulties. More or less everywhere, the first lines were short and still had a suburban character. By 1845, however, already more than 5,000 kilometers were constructed on the continent and networks were gradually developing.

The First Developments

In Great Britain ownership of the railways was private (Mitchell, 1964). The initiative was left entirely in the hands of private individuals. In 1837 there were already around a hundred railway companies active in Great Britain. In 1850, just 25 years after the opening of the Stockton–Darlington line, no less than 10,000 kilometers of railway lines had been built.

In continental Europe, private capital was certainly important, but it was often accompanied by state planning, control and often financing (Fremdling, 2003). In some countries, such as in France and in Belgium, this was true almost from the beginning. In others, such as The Netherlands, the state stepped in when railway construction seemed to lag behind other countries. In Germany and Italy the railways were quickly assigned an even more crucial role by the state, namely as a vehicle for national unification (Schram, 1997). Normally, a myriad of small companies played an initial role in the construction and management of the lines. Quickly, a process of consolidation and concentration developed, however, resulting in more robust corporate realities.

The various national networks developed in different formats, depending both on geographical and orographic reasons, on the size of the spaces, on the territorial distribution of the population (population density), on the kind of political and institutional system. Sometimes, reasons of centrality of the city capital prevailed. In Spain this led to the formation of a radial network focusing on Madrid.

Technological and Labor Issues

In the first phase, engineering techniques made giant strides. Civil engineering, locomotives and rolling stock all improved. Britain and Belgium were forerunners. Engineers and contractors of British railways stood out in particular. They traveled around Europe with their skilled workers, exporting knowledge of which they were the first depositories. George Stephenson and Isambard Kingdom Brunel were the most requested names, but Mackenzie, Samuel Morton Peto and Thomas Brassey were big names as well. Railway created thousands of new jobs, related to new constructions and maintenance of both the network and rolling stock.

The diffusion of the railways represented in many ways the definitive consecration of technology. From a technological point of view the locomotive was an extremely sophisticated product. The question of the most efficient and effective traction system kept generations of engineers busy. Great progress was achieved with speed, even though fuel efficiency was as important as speed: in 1890 a French locomotive reached 144 kilometers per hour. In 1879 the first electric locomotive was presented in Berlin.

In 1869 Westinghouse developed another breakthrough innovation with the air brake. The driver could now operate simultaneously the brakes placed on all the carriages by means of an air duct that originated from the locomotive. The automation of signaling really started off in 1872 with the system of automatic closure of the signals along the railway sections. Equally crucial was the remote control of signaling, based on the electro-magnetic telegraph. Its trajectory of development is strongly linked to that of the railways.

Way less visible, but allegedly as important, was a breakthrough in the materials used. Steel rails, more durable and consequently cost effective, replaced gradually iron ones. Toward the end of the 19th century carriages were made of iron and steel as well, leading to the demise of wooden ones.

The introduction of refrigerated wagons made transport of perishable goods, such as meat, possible. This allowed revolutionary economies of scale to be reaped in the meat packing industry, and Chicago to become the “hog butcher to the world” (Cronon, 1991). Lighting and heating in the train were taken care of right away: first based on vegetable oil, then gas and finally electricity. This allowed comfortable passenger travel. In the same direction went the adoption of the sleeping car, originally created by George Pullman, also in Chicago.

The Local and the National

The railways allowed a different spatial organization of economic activities, both on the national and the local or regional level. The railways were crucial in the process of spatial market integration (Donaldson and Hornbeck, 2016), and contributed to reversal of economic imbalances, as natural protection of less efficiently producing areas fell (Enflo et al., 2018).

The influence of the railways on urban development is diverse (Duranton and Turner, 2012; Roth and Polino, 2003; Schwartz et al., 2011). It ranges from the unhinging of traditional urban networks, to the development of railway stations as primary objects of architecture and urban planning. The development of railway networks in the 1880s and 1890s in urban areas often coincided with the demolition of traditional protective walls. This gave space to build huge, now monumental, railway stations on the outskirts of traditionally built-up areas, including related railway quarters with, for example, hotels (*hotel terminus*) and residential quarters for railway employees. In many cities, the 19th-century railway station is still a symbol of the power of progress, a temple of technological advancement, and a privileged field of experimentation by generations of architects and engineers called upon to design a radically new building.

The Maturity Phase

What is commonly called the second industrial revolution (conventionally, 1870–1914) also represents the stage of railway maturity. At the end of the 19th century, with an endowment of some 35,000 kilometers of railway lines, Great Britain was still among the leading countries. On the continent, however, France (40,000 kilometers) and in particular Germany and Russia (50,000 kilometers each) had surpassed Great Britain. Even British India had a larger network than the colonial power itself (39,000 kilometers). These countries were dwarfed, however, by the United States.

The dominant railroad country in the 19th century clearly was the United States. It had been an early starter as well with, already in 1840, twice as many railroads as Great Britain (4500 kilometers). In 1860 the network reached the Mississippi and in 1869 the connection between the Atlantic and the Pacific coast was completed (“from the sea to the sining sea,” as the adage goes), when in Promontory (Utah) the tracks of the Union Pacific Railroad were linked to those of the Central Pacific Railroad. In 1900 the network had grown to an almost unimaginable 311,000 kilometers. This was more than all European countries together. Although the expansion took place in more or less the same time as in Great Britain, the latter still supplied a lot of the necessary financial capital (Fenoaltea, 1988). Railways were first built to meet the demand for transport of agricultural products, such as grain and livestock, but the railways also created patterns of industrial growth by opening up the country.

The same would play a role in Russia, another country characterized by huge territories (Haywood, 1969, 1998). In 1855, four years after the opening of the Moscow-Saint Petersburg line, the network consisted of almost 1,200 kilometers. The railways represented for a strong stimulus for the industrialization of Russia, and, more generally, an opportunity for its modernization. The Russian metallurgical sector responded to the demand for iron and steel for rails and equipment. However, the country needed to look abroad to obtain the necessary financial capital. Between 1866 and 1890 the Russian network increased tenfold. Territories of more recent acquisition such as the in the Caucasus and the regions toward the Caspian Sea were integrated in the national domain by means of railway building. The line that ignited the imagination of contemporaries more than any other, however, was the Trans-Siberian. It was built between 1891 and 1903. Russia used it to effectively conquer its eastern territories, significantly consolidating imperial political power and with the purpose to settle the Siberian plains with Russians. At the same time it responded to an economic perspective for the transport of cereals and the potential exploitation of oil reserves.

Technological progress continued in the maturity phase of railways. The application of compounding and super-heating to steam locomotives reduced fuel and water consumption, and lowered the weight-to-power ratio, a standard index of technological progress (Ciccarelli and Nuvolari, 2015).

Regulation, Standardization, and International Agreements

The most significant change in the maturity phase concerned corporate finance and investment. In the second half of the 19th century new ways of supplying money came forward. The needs of large infrastructural projects innovated the financial system: from the adventurism of the early years, the world of railways had moved toward growing stability. If the commitments were too costly and the risks too large, the state often got involved (Bogart, 2013). In Germany, Romania, Serbia, and Bulgaria the state nationalized railway services from the second half of the 1870s. In Austria, Denmark, the Netherlands, Russia, and Switzerland the government started buying or constructing the main railway lines. Even where the railway sector was managed as essentially a private business governmental control and regulation became standard practice. Slowly, the idea that railways were a necessary social service and consequently a prime responsibility of the state was making its way.

As railway networks became more integrated, also at the international level, problems arose concerning the standardization of tracks, equipment, signaling, and organization (Carreras et al., 1995). From the 1870s an international regime of regulation of the various technical and commercial issues made its way. Technical integration in Europe, as in other continents, took many forms; overall it was an expensive and not at all easy process, hampered also by persistent political fragmentation (Federico and Sharp, 2013).

The first major problem that needed standardization was that of gauge, first between companies and later between countries. The increase in trade flows also pushed toward simplified customs procedures and in the use of wagons in international service. In 1872 the *Conférence Européenne des Horaires des Trains des Voyageurs et Services Directs* was created in Bern, Switzerland. It was an international organization with the aim of coordinating international connections. Shortly thereafter a complementary technical unit was created, the *Conférence Internationale pour l'Unité Technique des Chemins de Fer*. In 1922 the *Union Internationale de Chemins de fer* was erected in Paris (the website <https://uic.org/>, last accessed December 2019, provides a wide set of publications and data covering the more recent decades).

The Journey by Rail

The diffusion of railways introduced a fundamental change in the idea of travel and allowed the birth of railway tourism (Schivelbusch, 1986; Smith, 1998). This would experience its greatest expansion during the 20th century. The train offered itself as the ideal environment for pleasure trips. It facilitated traveling to reputed places of pleasure, while at the same time constituting itself as a tourist attraction as well. The rise of tourism as an organized economic activity took shape with the advent of the railway, roughly between 1830 and 1914. Everywhere in Europe railway companies were equipped for Sunday trips and inaugurated tourism policies. Since 1848, international pleasure trips were planned. The concept of holiday abroad is contemporaneous to the early diffusion of trains and the first travel agencies—the most famous of which was Thomas Cook—made their appearance and organized trips by rail. The tourism factor ultimately influenced the creation and subsequent exploitation of an integrated international rail network in terms of increasing the number of trains, choosing routes, coordinating schedules, increasing travel comfort, and speeding up transfers. Being comfortably seated in the compartment for a long time led to the spread of the habit of reading, which in turn caused a thriving travel literature to develop, among which the railway guides themselves. The most famous were Baedeker, Murray, Bradshaw, and the French Guide Chaix. French publishing house Hachette even inaugurated a specific series, the *Bibliothèque des chemins de fer*, specifically designed for rail travel. In 1889 the circulating library for travelers was tested on the Austro-Hungarian Empire railways, thanks to which it was possible to borrow books at each station and return them at a different one.

In 1883 the first trip took place of a train that would soon become legendary, the Orient Express. It was conceived by the *Compagnie Internationale des Wagons-Lits* and started in Paris, to go through Munich Vienna, Giurgiu (Romania), and the port of Varna on the Black Sea. From there, the final bit to Constantinople was done by boat.

The Semmering was the first true Alpine railway. It was part of the link between Vienna, the capital city of the Habsburg Empire, and Trieste, with its port on the Adriatic Sea. The line was completed in 1854. There were 15 tunnels and 16 long viaducts on the line. In 1998 it was declared a Unesco World Heritage site.

Many were also the failures, however. In particular in Africa, a number of projects did grab the imagination, but not the necessary financial capital to be successfully completed. The Trans-Saharan railway, for example, was long studied by the French colonial powers, but never realized. The same happened with the famous Cape to Cairo line, which was conceived to link and integrate a series of major British colonies in Africa, such as South Africa, Kenya, and Egypt. In the 21st century, China has taken up the challenge again of building railways in Africa, also to strengthen its economic and political influence.

In the 1870s, the newly unified Germany started considering the building of a railway line connecting Berlin to the (then) Ottoman Empire city of Baghdad. It was not just geopolitical motivation that drove Berlin to this project, but also the oil fields of the region until then mainly exploited by British and American companies. When the railway line finally opened in 1940, after endless political and financial difficulties, the world of transport had profoundly changed, however.

The 20th Century Decline

Rail transport maintained its prevalence over other forms of transport in the early decades of the 20th century. Between 1900 and 1915 the world network grew by almost 40%, so that at the start of the First World War more than one million kilometers of railway lines were in operation. Growth continued until the Second World War, but at a much slower pace: from 1920 to 1940 the increase was only 10%. The global network kept growing until in 1960 a total of more than 1.2 million kilometers was in existence. In many of the early adopting countries, however, the decline had already started at an earlier time: in Germany and the United States in 1915, in Great Britain in 1925, and in most other Western European countries around the 1930s. The rise of automotive transport (private cars, lorries, coaches) was the prime cause. It led to a clear decline in the use of the railways for public transport. Thousands of kilometers of lines, often the ones that were built as the last extensions of the network, were abandoned for a lack of passengers and consequent low profitability. Lately, abandoned railway lines often gained a second life as cycling or walking trails, both for domestic use and for tourism. The German Ruhr area, for example, markets its extensive network of cycling routes on abandoned (industrial) railway lines.

In Africa and Asia, differently from Europe and North America, the railway network kept growing until the end of the 20th century (Jedwab and Moradi, 2016; Jedwab et al., 2017; Okoye et al., 2019). The temporal trends and the role played by railway to sustain economic development was, as should not be surprisingly, ultimately different depending on the position of a given country on the (political and socio-economic) world map.

The most important technical innovation of the 20th century was the transition from steam engines to diesel and electricity. From a management point of view, the role played by the state increased enormously. The latter went hand in hand with the creation of the welfare state, and indeed the provision of *public* transport was often considered as a standard task for governments. Switzerland opted for nationalization of its railways in 1902, Italy in 1905, Japan in the following year, and Belgium in 1913. France and Great Britain, in, respectively, 1937 and 1946.

At the end of the 20th century the role of governments as the dominant players in the railway sector was brought into question again (Fremdling and Knieps, 1993). Great Britain, under Prime Minister Margaret Thatcher, took the lead and reintroduced competition between privatized companies in the exploitation of railways. Other countries switched to mixed systems, with The Netherlands, for example, allowing regional monopolies to commercial companies on (peripheral) parts of the network.

A Resurrection of Rail Transport?

In the last decades of the 20th century railways started to gain momentum again on both ends of the continuum of spatial scales. Globalization, in the first place, created an impetus for the construction of an integrated international network of high-speed railway lines. Japan was first with its famous Shinkansen (bullet) trains in 1964. From the 1970s high-speed trains began to be developed in Europe as well. France was the forerunner with its *Train à Grande Vitesse*. Germany followed with the InterCity Express, and Italy with the *Pendolino*. Also in Germany, Siemens developed and tested a Maglev (train with magnetic levitation), but this did not lead to large-scale adoption of the system. Another development on the global scale is intermodal transport, mainly with containers. Railway companies have secured some position in the ship-road-rail operational and management integration.

On the local level, urban tramways are making something of a comeback. A considerable number of cities all over the world have (re)introduced trams not merely to solve traffic congestion problems, but also to create an appealing atmosphere for the creative classes that is seen as the backbone of the current urban economy. A telling example is the reintroduction in 2018 of the tramway (or streetcar) in Detroit, the Motor City and the very symbol of the era of the automobile.

Railways and Economic Development: A Review of the Literature

History represents an important factor conditioning railways and more generally transport systems. About 70% of the railway lines operating today in Europe already existed in 1900 (Martí-Henneberg, 2013). A vast literature by economists, economic historians, geographers, and scholars of related fields includes various attempts to measure or quantify the impact of railway diffusion on various socioeconomic outcomes. Indeed, with the development of railways for the first time transportation costs over land were competitive to those over water.

In their seminal publications of the mid-1960s, Fogel (1962, 1964) and Fishlow (1965) introduced the *Social saving* (SS) methodology. It relates to the aggregate cost saving that the new technology (railway) allows in comparison with the next best alternative. SS represents a possible measure of the benefit to society of technological improvements. Omitting here some technical yet important details (such as the elasticity of the demand curve and the consumer surplus dear to economists), if P_{water} represents the price of quantity transported by water and P_{rails} the price of quantity transported by rail, and Q_{rails} the quantity (freight or passenger) transported by rail, then SSR (SS of Railways) can be defined, as $SSR = (P_{\text{water}} - P_{\text{rails}}) \times Q_{\text{rails}}$. SSR was considered by Fogel mostly as an upper bound of the gains induced by railways. The contributions by Fogel and Fishlow stimulated a vast new literature labeled cliometrics (for a review of the debate see O'Brien, 1977; Fogel, 1979; Leunig, 2010; Atack, 2019; Bogart, 2019).

Railway developments in the 19th century have been also analyzed in the more general context of rising international capital movements during the age of first globalization (1870–1914). The basic idea is that when in a given country net imports of capital from abroad were abundant (scanty), public investments on the construction sectors, including the railways, were relatively high (scanty) (Fenoaltea, 1988, Roth and Dinshob, 2008). Hence new construction is inherently of a cyclical nature. Various scholars have analyzed the backward linkages of railway construction and maintenance: the induced demand for manufacturing (metalmaking and engineering) and in particular of rolling stock (locomotives, freight and passenger cars; see for instance Ciccarelli and Fenoaltea, 2012). Backward linkages often had a limited, often local, spatial context. The development and diffusion of electricity as a source of power since the late 19th century induced a gradual replacement of steam locomotives with electric locomotives, as the average life-span of a steam locomotive was of about 40–50 years (Ciccarelli and Nuvolari, 2015) and consequently a new round of backward linkage effects.

In more recent decades, scholars have begun to focus on the spatiality of the impact of the railway, with analysis carried out at the regional level. This literature takes advantage of new developments in geographic information systems (Atack, 2013). This proved to be a powerful tool for displaying spatial phenomena, and to analyze the geographical diffusion of railways and its possible socioeconomic impacts, often at a much lower level of geographical aggregation than before (municipalities or parishes). The building of a geodatabase is often based on both historical sources and maps (for instance, the Library and Archive section of the website of the Italian Fondazione FS—<http://www.fondazionefs.it/content/fondazionefs/en.html>, last accessed December 2019—provides a rich set of digitized historical sources). The railway sector is generally well documented, as it was often regulated by public authorities, typically producing annual reports to Parliament. Information on the exact day of opening and length of each track, as well as the locations of stations and stops are often available. Historical maps represent another precious ingredient in the production of geodatabases, especially so when they are homogenous over time (Ciccarelli and Groote, 2017, 2018). Retrospective maps were often published in the occasion of international meetings such as the various International Exhibitions of the 19th century.

In most (western) countries the early diffusion of railway occurred in concomitance with the early phases of the first demographic transition. This is the a secular process of the shift from a preindustrial demographic system characterized by high and fluctuating birth and death rates to a modern regime of low and more stable death rates (first) and birth rates (later). As a result, the population was young and expanding rapidly during the first phases of railway construction. The concurrence between rapid railway diffusion and population growth stimulated various contributions attempting to identify a possible causal relation between them. Indeed, railway development might both follow and induce population and economic growth. Reverse causality is also possible representing a threat to the identification of causal relations (the endogeneity problem, to use economists' jargon). Emblematic of this approach is Atack et al., 2010 "Did railroads induce or follow economic growth? Urbanization and population growth in the American midwest, 1850-1860." It stimulated various contributions concerning different countries (among others, Gregory and Henneberg, 2010 for England and Wales; Groote et al., 2009 for The Netherlands; Berger and Enlfo, 2017 for Sweden).

The remaining of the chapter illustrates the geographical and temporal distribution of railway at the global and continental level, and for selected countries (Figs. 1–9). Tables 1–6 quantify the diffusion of railway during the "golden age" of the late 19th century.

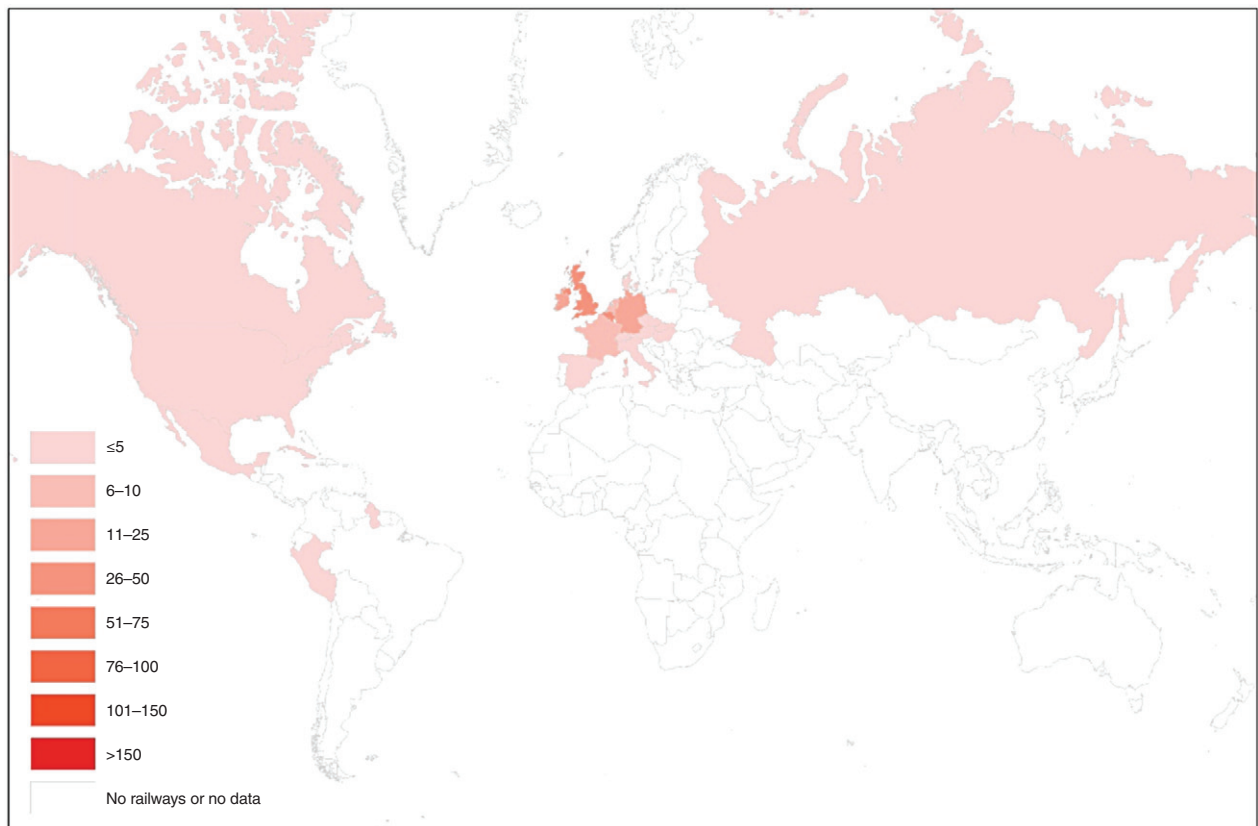


Figure 1 World—railway density in 1850 (km of lines per 1000 km²). *Source: Authors' elaborations on Mitchell (2013)*

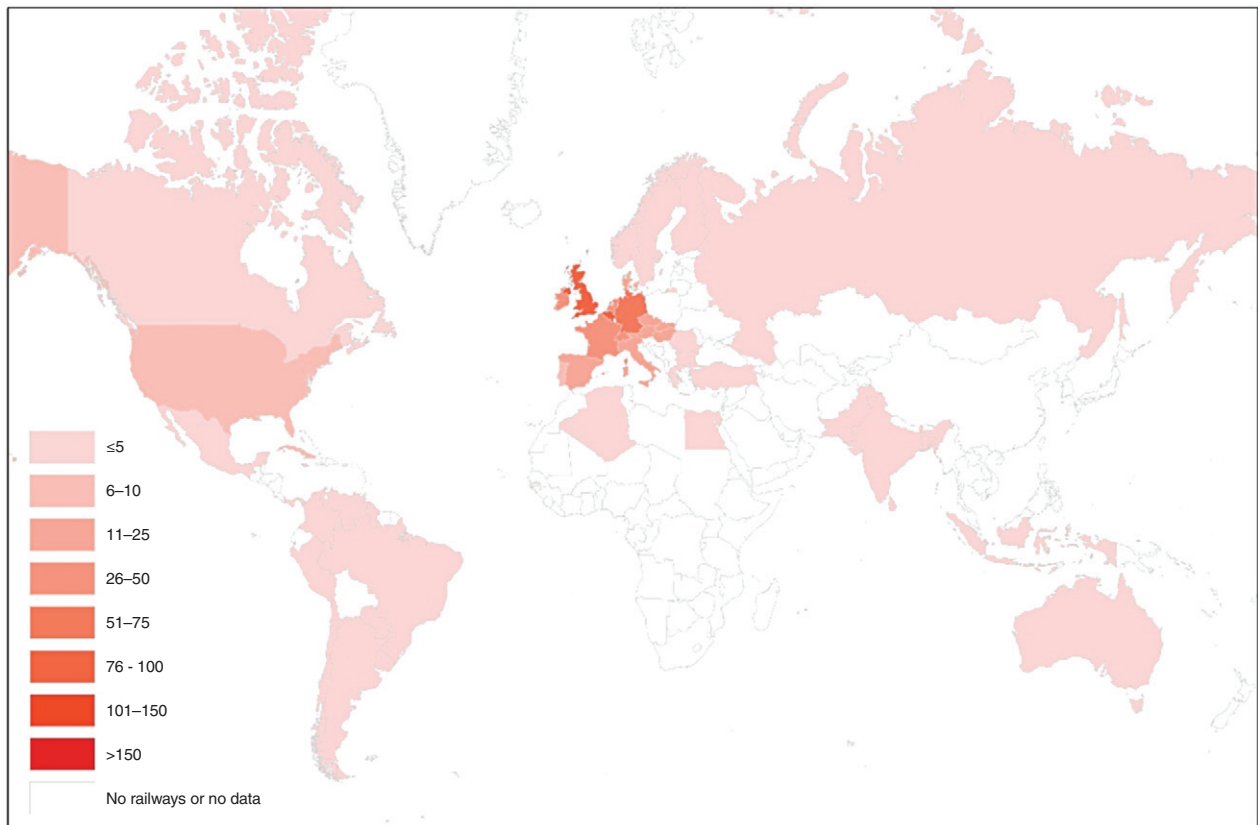


Figure 2 World—railway density in 1870 (km of lines per 1000 km²). *Source: Authors' elaborations on Mitchell (2013)*

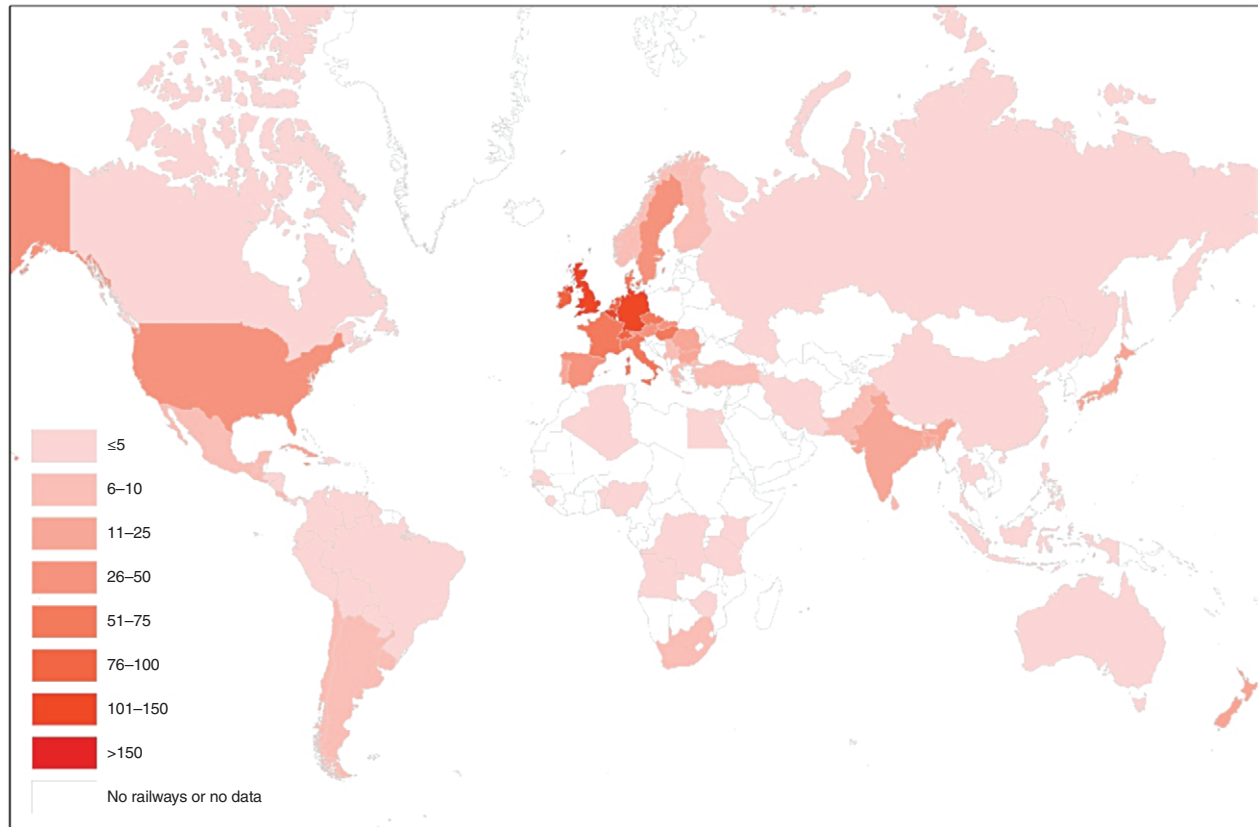


Figure 3 World—railway density in 1900 (km of lines per 1000 km²). Source: Authors' elaborations on [Mitchell \(2013\)](#)

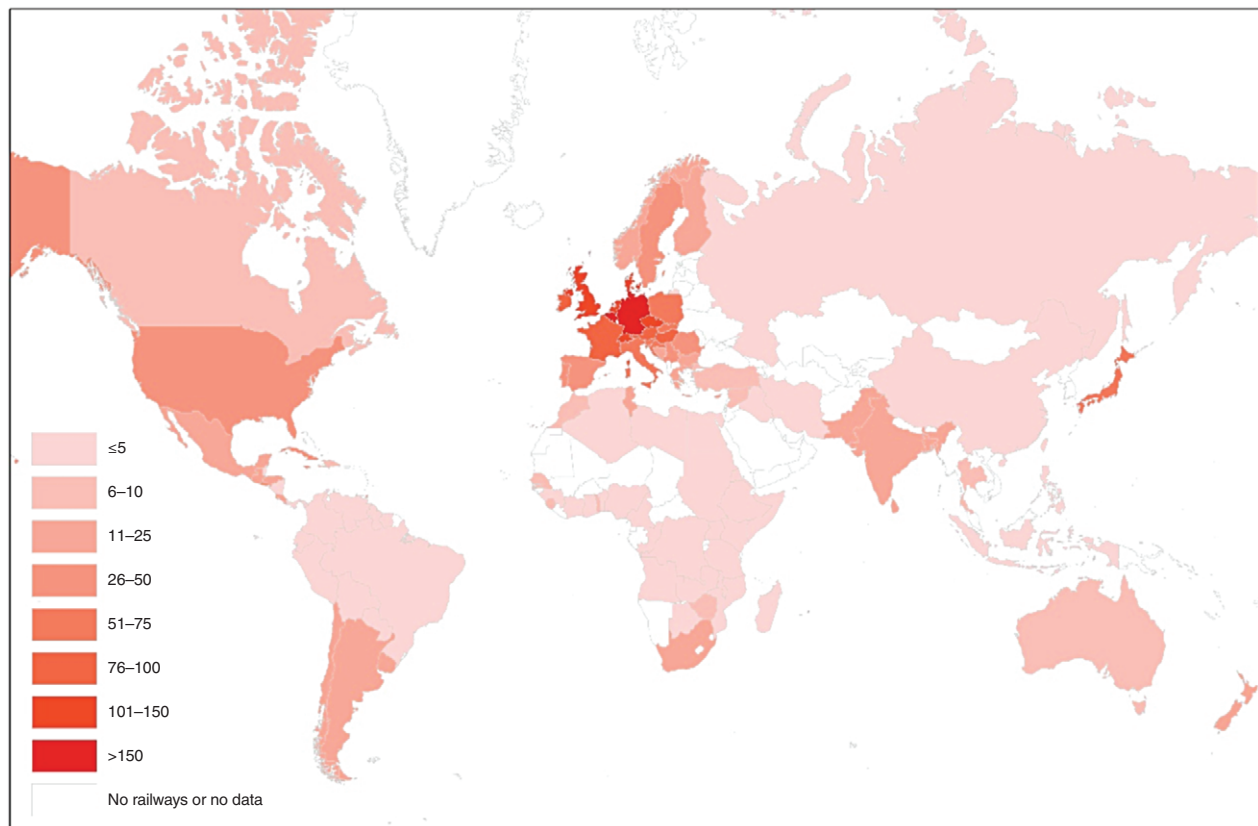


Figure 4 World—railway density in 1930 (km of lines per 1000 km²). Source: Authors' elaborations on [Mitchell \(2013\)](#)

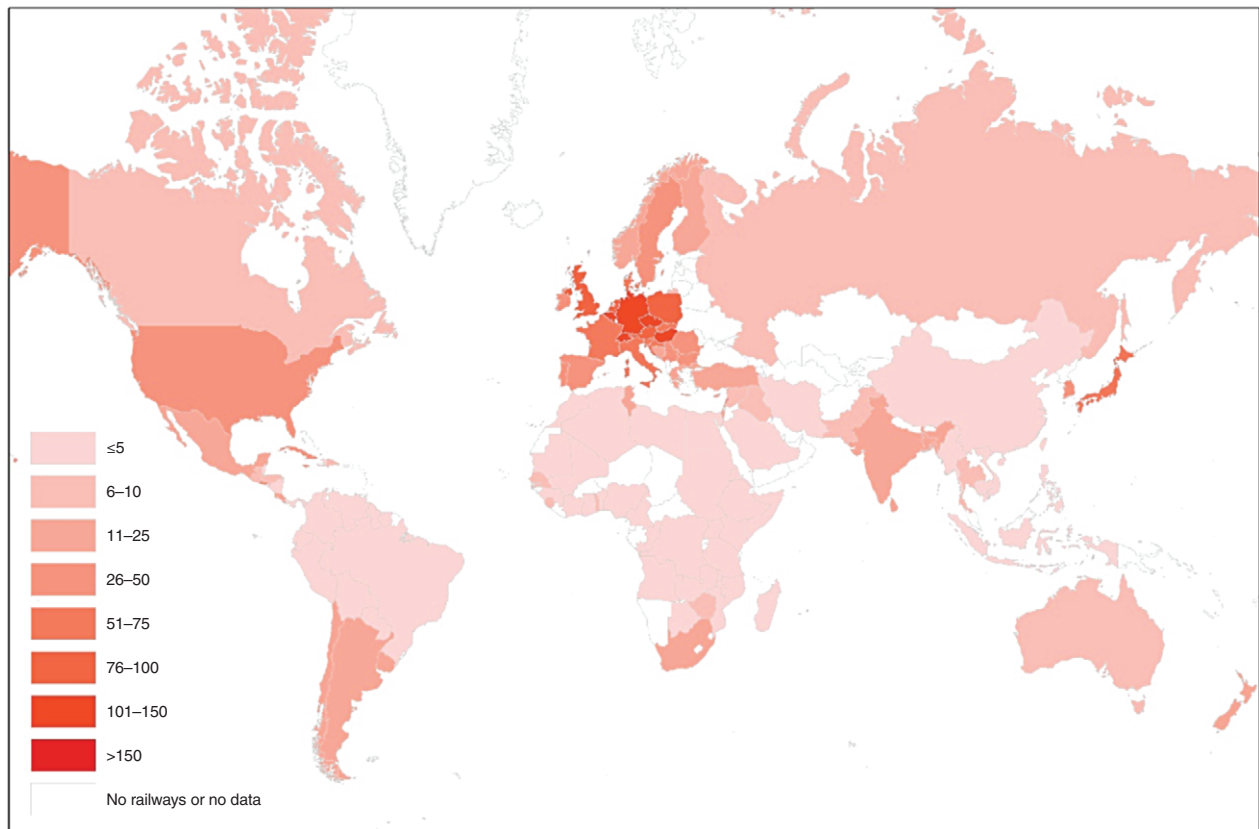


Figure 5 World—railway density in 1970 (km of lines per 1000 km²). *Source: Authors' elaborations on Mitchell (2013)*

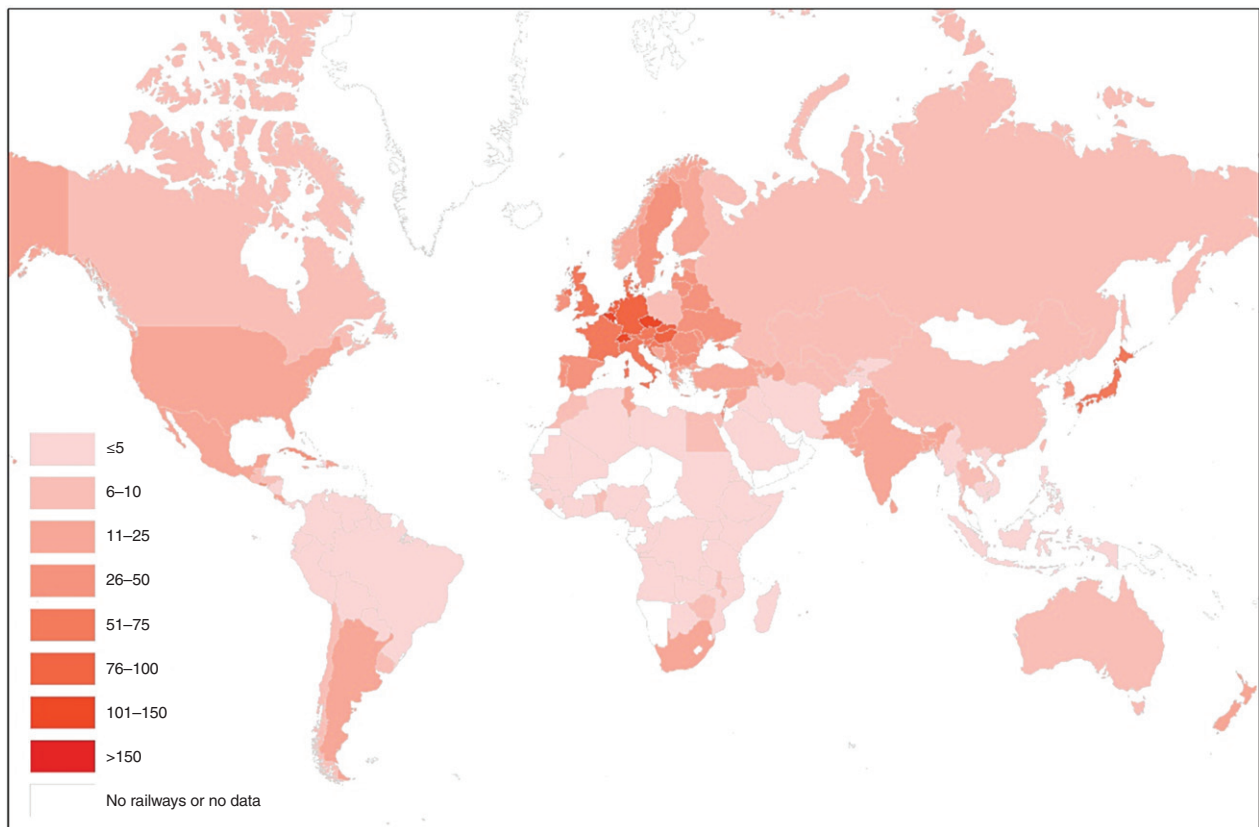


Figure 6 World—railway density in 2010 (km of lines per 1000 km²). *Source: Authors' elaborations on Mitchell (2013)*

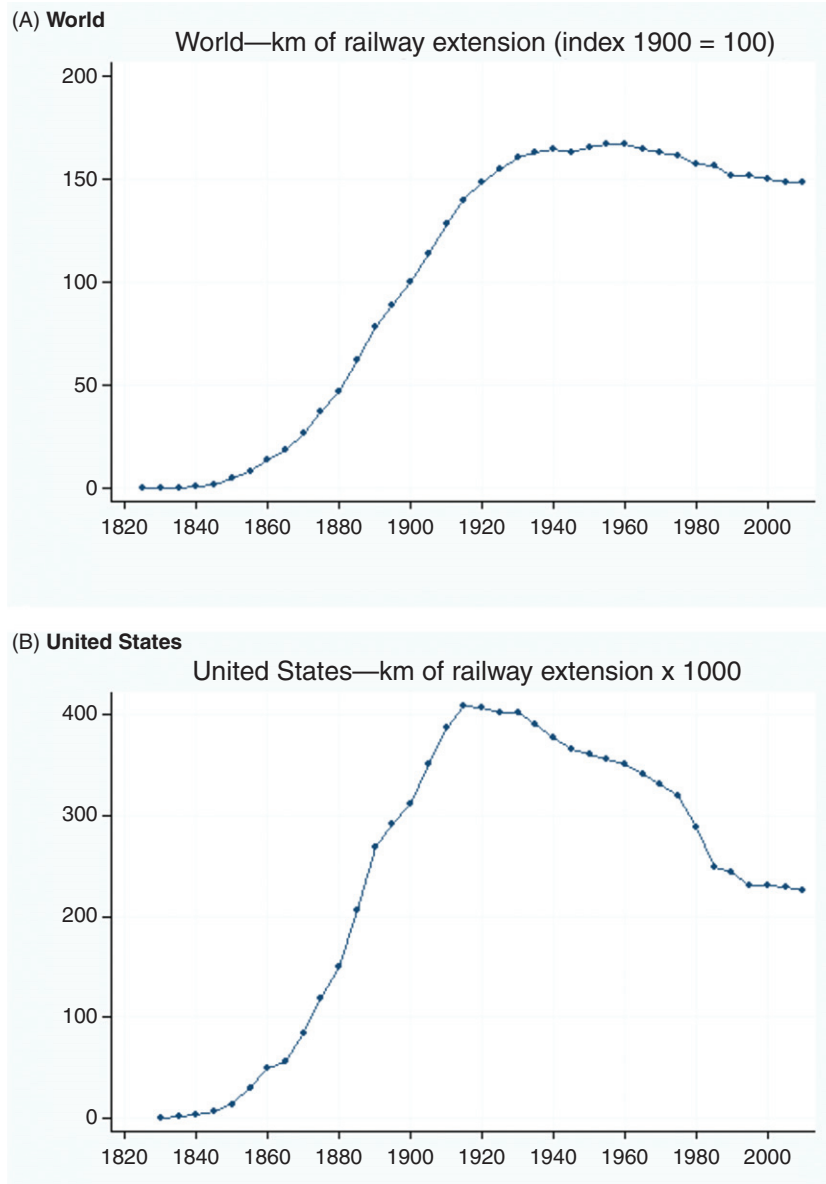


Figure 7 Railway extension, 1820–2010. World and selected countries. (A) World, (B) United States, (C) United Kingdom, (D) China Source: Authors' elaborations on Mitchell (2013)

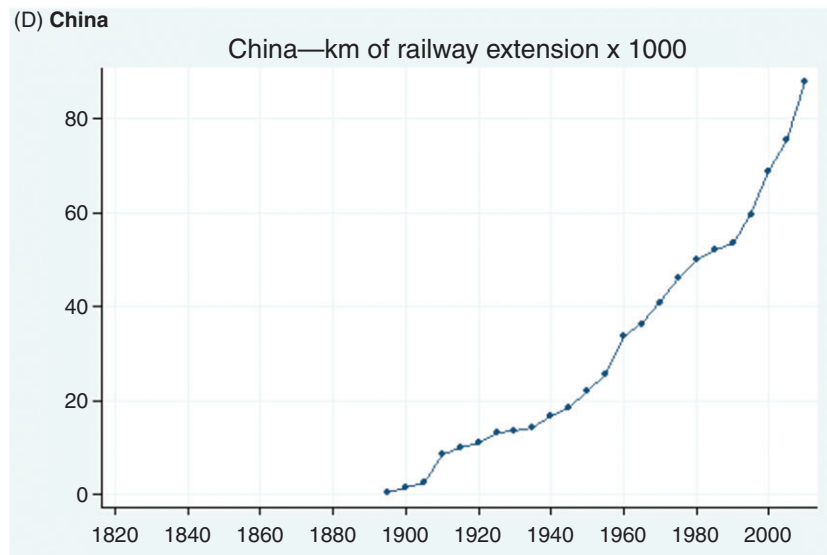
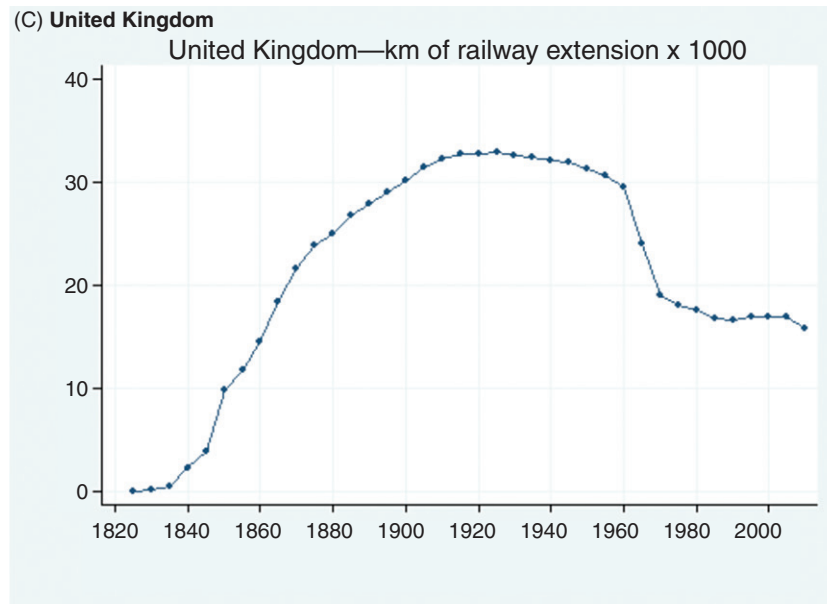


Figure 7 (Continued)

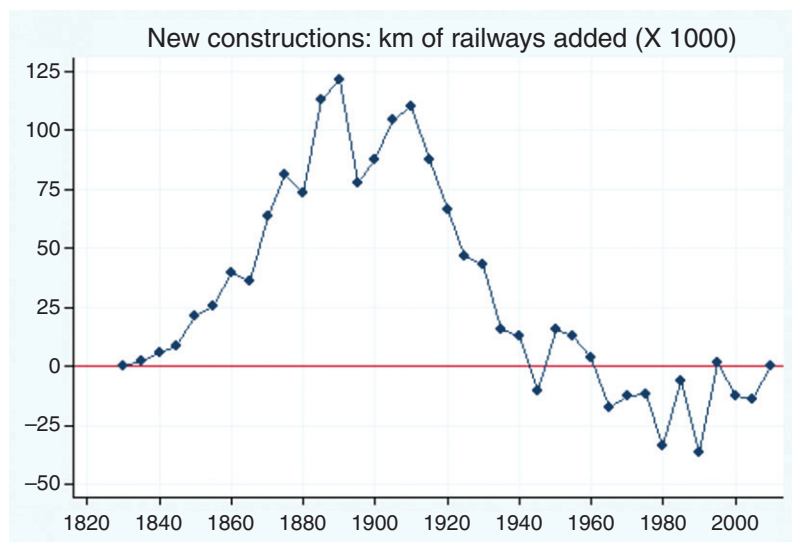


Figure 8 World: new constructions, 1820–2010. Source: Authors' elaborations on [Mitchell \(2013\)](#)

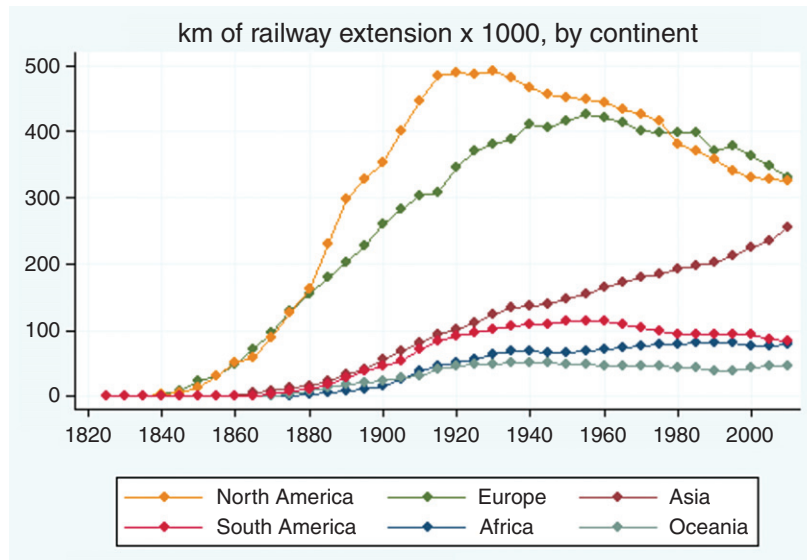


Figure 9 Railway extension, 1820–2010, by continent. Source: Authors' elaborations on Mitchell (2013)

Table 1 World—railway extension (km), by continent 1840–1913

	1840	1850	1860	1870	1880	1890	1900	1913
Europe	2,925	23,504	51,862	104,914	168,983	223,869	283,535	346,235
America	4,754	15,064	53,935	93,139	174,666	331,417	403,171	570,108
Asia	0	0	1,393	8 185	16,287	33,724	60,301	108,147
Africa	0	0	455	1,786	4,646	9,386	20,114	44,309
Oceania	0	0	367	1,765	7,847	18,889	24,014	35,418
Total	7,679	38,568	108,012	209,789	372,429	617,285	791,135	1,104,217

Source: Authors' elaborations on Treccani (1932)

Table 2 Europe—railway extension (km), by countries 1840–1913

	1840	1850	1860	1870	1880	1890	1900	1913
GER (1835)	549	6,044	11,633	19,575	33,838	42,869	51,391	63,730
AUH (1838)	144	1,579	4,543	9,589	18,512	27,113	36,883	46,195
GBI (1825)	1,348	10,653	16,787	24,999	28,854	32,297	35,186	37,717
FRA (1832)	497	3,083	9,528	17,931	26,189	36,895	42,827	51,188
BEL (1835)	336	854	1,729	2,997	4,120	5,263	6,345	8,814
RUS (1838)	26	601	1,589	11,243	23,857	30,957	48,107	62,198
NET (1839)	17	176	335	1,419	2,300	3,060	3,209	3,256
ITA (1839)	8	427	1,800	6,134	8,715	12,907	15,787	17,634
SWI (1844)	0	27	1,096	1,449	2,571	3,190	3,783	4,863
DEN (1847)	0	32	111	764	1,579	1,986	3,001	3,771
SPA (1848)	0	28	1,918	5,475	7,481	9,878	13,357	15,350
SWE (1851)	0	0	522	1,708	5,906	8,018	11,320	14,491
NOR (1854)	0	0	68	359	1,059	1,562	2,053	3,092
POR (1854)	0	0	137	714	1,150	2,149	2,376	2,983
ROM (1870)	0	0	0	245	1,387	2,543	3,098	3,763
Other	0	0	66	313	1,465	3,182	4,812	7,190
Total	2,925	23,504	51,862	104,914	168,983	223,869	283,535	346,235

AUH, Austria and Hungary; BEL, Belgium; DEN, Denmark; FRA, France; GBI, Great Britain and Ireland; GER, Germany; ITA, Italy; NET, Netherlands; NOR, Norway; POR, Portugal; ROM, Romania; RUS, Russia; SPA, Spain; SWI, Switzerland. The year of opening of first railway line is in parenthesis.

Source: Authors' elaborations on Treccani (1932)

Table 3 America—railway extension (km), by countries 1840–1913

	1840	1850	1860	1870	1880	1890	1900	1913
United States (1830)	4,534	14,515	49,292	85,139	150,717	268,409	311,094	410,918
Canada (1840)	26	114	3,359	4,018	11,087	22,533	29,697	47,150
Mexico (1850)	0	11	32	349	1,120	9,800	14,573	25,492
Perù (1851)	0	0	89	411	1,852	1,667	1,667	2,766
Chile (1852)	0	0	195	732	1,800	3,100	4,586	6,370
Brasil (1854)	0	0	129	691	3,200	9,500	14,798	24,985
Argentina (1857)	0	0	39	732	2,273	9,800	16,369	33,215
Uruguay (1869)	0	0	0	98	370	1,127	1,841	2,638
Bolivia (1873)	0	0	0	0	56	209	1,000	2,418
Other countries	194	424	800	969	2,191	5,272	7,546	14,156
Total	4,754	15,064	53,935	93,139	174,666	331,417	403,171	570,108

The year of opening of the first railway line is in parenthesis.

Source: Authors' elaborations on [Treccani \(1932\)](#)

Table 4 Asia—railway extension (km), by countries 1840–1913

	1840	1850	1860	1870	1880	1890	1900	1913
British India (1853)	0	0	1,350	7,683	14,977	27,000	38,235	55,761
Turkey (1860)	0	0	43	243	372	800	2,760	5,468
Dutch India (1867)	0	0	0	150	450	1,361	2,094	2,854
Japan (1872)	0	0	0	0	121	2,333	5,934	10,986
China (1871)	0	0	0	0	11	200	646	9,854
Russia (1880)	0	0	0	0	125	1,433	8,869	15,910
Siam (1893)	0	0	0	0	0	0	327	1,130
Other countries	0	0	0	109	231	597	1,436	6,184
Total	0	0	1,393	8,185	16,287	33,724	60,301	108,147

The year of opening of the first railway line is in parenthesis.

Source: Authors' elaborations on [Treccani \(1932\)](#)

Table 5 Africa—railway extension (km), by countries 1840–1913

	1840	1850	1860	1870	1880	1890	1900	1913
Egypt (1856)	0	0	443	1,056	1,500	1,547	3,358	5,946
Algeria and Tunisia (1862)	0	0	0	517	1,379	3,104	4,251	6,382
Southern Africa (1860)	0	0	12	105	1,617	3,825	8,807	17,628
Other countries	0	0	0	108	150	910	3,698	14,353
Total	0	0	455	1,786	4,646	9,386	20,114	44,309

The year of opening of the first railway line is in parenthesis.

Source: Authors' elaborations on [Treccani \(1932\)](#)

Table 6 Oceania—railway extension (km), by countries 1840–1913

	1840	1850	1860	1870	1880	1890	1900	1913
Victoria (1854)	0	0	151	443	1,930	4,325	5,178	5,910
New South Wales (1855)	0	0	113	545	1,368	3,641	4,523	6,594
Queensland (1865)	0	0	0	331	1,019	3,435	4,507	7,753
West Australia (1873)	0	0	0	0	116	825	2,194	5,519
Other countries	0	0	103	446	3,414	6,663	7,612	9,642
Total	0	0	367	1,765	7,847	18,889	24,014	35,418

The year of opening of the first railway line is in parenthesis.

Source: Authors' elaborations on [Treccani \(1932\)](#)

References

- Alvarez, E., Franch, X., Martí-Henneberg, J., 2013. Evolution of the territorial coverage of the railway network and its influence on population growth: the case of England and Wales, 1871–1931. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 46 (3), 175–191.
- Atack, J., 2013. On the use of geographic information systems in economic history: the American transportation revolution revisited. *J. Econ. History* 73 (2), 313–338.
- Atack, J., 2019. Railroads. In: Diebolt, C., Hauptert, M. (Eds.), *Handbook of Cliometrics*. Springer, Berlin, Heidelberg.
- Atack, J., Bateman, F., Haines, M., Margo, R.A., 2010. Did railroads induce or follow economic growth? Urbanization and population growth in the American Midwest, 1850–1860. *Soc. Sci. Hist.* 34 (2), 171–197.
- Berger, T., Enflo, K., 2017. Locomotives of local growth: the short-and long-term impact of railroads in Sweden. *J. Urban Econ.* 98, 124–138.
- Bogart, D., 2013. The economic history of transportation. In: Whaples, R., Parker, R.E. (Eds.), *Routledge Handbook of Modern Economic History*. Routledge, New York, pp. 167–176.
- Bogart, D., 2019. Clio on speed. In: Diebolt, C., Hauptert, M. (Eds.), *Handbook of Cliometrics*. Springer, Berlin, Heidelberg.
- Carreras, A., Giuntini, A., Merger, M., 1995. *Réseaux européens transnationaux, XIXe-XXe siècles: quels enjeux?* Ouest Editions, Nantes.
- Ciccarelli, C., Fenoaltea, S., 2012. The Rail-Guided Vehicles Industry in Italy, 1861–1913: the burden of the evidence. *Res. Econ. History* 28, 43–115.
- Ciccarelli, C., Groote, P., 2017. Railway endowment in Italy's provinces, 1839–1913. *Rivista di Storia Economica* 33 (1), 45–88.
- Ciccarelli, C., Groote, P., 2018. The spread of railroads in Italian provinces: a GIS approach. *Sci. Regionali: Italian J. Reg. Sci.* 17 (2), 189–224. Geodata Available from: <http://bit.do/italianrailways>.
- Ciccarelli, C., Nuvolari, A., 2015. Technical change, non-tariff barriers, and the development of the Italian locomotive industry, 1850–1913. *J. Econ. Hist.* 75 (3), 860–888.
- Chandler Jr., A.D., 1977. *The Visible Hand: The Managerial Revolution in American Business*. The Belknap Press of Harvard University Press, Massachusetts and London, England.
- Cronon, W., 1991. *Nature's Metropolis. Chicago and the Great West*. W.W. Norton & Company, New York.
- Donaldson, D., Hornbeck, R., 2016. Railroads and American economic growth: a “market access” approach. *Quart. J. Econ.* 131 (2), 799–858.
- Duranton, G., Turner, M.A., 2012. Urban Growth and Transportation. *Review of Economic Studies* 1, 1–36.
- Enflo, K., Alvarez-Palau, E., Martí-Henneberg, 2018. Transportation and regional inequality: the impact of railways in the Nordic countries, 1860–1960. *J. Hist. Geogr.* 62, 51–70.
- Federico, G., Sharp, P., 2013. The cost of railroad regulation: the disintegration of American agricultural markets in the interwar period. *Econ. Hist. Rev.* 66 (4), 1017–1038.
- Fenoaltea, S., 1988. International resource flows and construction movements in the atlantic economy: the Kuznets cycle in Italy, 1861–1913. *J. Econ. Hist.* 48 (3), 605–637.
- Fishlow, A., 1965. *American Railroads and the Transformation of the Ante-Bellum Economy*. Harvard University Press, Cambridge, MA.
- Fogel, R.W., 1962. A quantitative approach to the study of railroads in American economic growth: a report of some preliminary findings. *J. Econ. Hist.* 22 (2), 163–197.
- Fogel, R.W., 1964. *Railroads and American Economic Growth*. Johns Hopkins Press, Baltimore.
- Fogel, R.W., 1979. Notes on the Social Saving Controversy. *J. Econ. Hist.* 39 (1), 1–54.
- Fremdling, R., 2003. European Railways 1825–2001, an overview. *Jahrbuch für Wirtschaftsgeschichte* 44 (1), 209–221.
- Fremdling, R., Knieps, G., 1993. Competition, regulation and nationalization: The Prussian railway system in the nineteenth century. *Scandinavian Econ. Hist. Rev.* 41 (2), 129–154.
- Gregory, I., Martí-Henneberg, J., 2010. The railways, urbanization, and local demography in England and Wales, 1825–1911. *Soc. Sci. Hist.* 34 (2), 199–228, doi:10.1017/S014553200011214.
- Groote, P., Elhorst, J.P., Tassenaar, P.G., 2009. Standard of living effects due to infrastructure improvements in the 19th century. *Soc. Sci. Comp. Rev.* 27 (3), 380–389.
- Haywood, R.M., 1969. *The Beginnings of Railway Development in Russia in the Reign of Nicholas 1, 1835–1842*. Duke University Press, Durham, NC.
- Haywood, R.M., 1998. *Russia Enters the Railway Age, 1842–1855*. East European monographs, Boulder.
- Jedwab, R., Moradi, A., 2016. The permanent effects of transportation revolutions in poor countries: evidence from Africa. *Rev. Econ. Stat.* 98 (2), 268–284.
- Jedwab, R., Kerby, E., Moradi, A., 2017. History, path dependence and development: evidence from colonial railroads, settlers and cities in Kenya. *Econ. J.* 127, 1467–1494.
- Leunig, T., 2010. Social saving. *J. Eco. Surv.* 24 (5), 775–800.
- Martí-Henneberg, J., 2013. European integration and national models for railway networks (1840–2010). *J. Trans. Geog.* 26, 126–138.
- Mitchell, B.R., 1964. The coming of the railway and United Kingdom economic growth. *J. Econ. Hist.* 24, 315–336.
- Mitchell, B.R., 2013. *International Statistics, 1750–2010*. Palgrave Macmillan, Basingstoke, Hampshire.
- O'Brien, P., 1977. *The New Economic History of the Railways*. St. Martin's Press, New York.
- Okoye, D., Pongou, R., Yokossi, T., 2019. New technology, better economy? The heterogeneous impact of colonial railroads in Nigeria. *J. Dev. Econ.* 140, 324–354.
- Rostow, W.W., 1960. *The Stages of Economic Growth; a Non-communist Manifesto*. Cambridge University Press, London.
- Roth, R., Dinholz, G., 2008. *Across the Borders: Financing the World's Railways in the Nineteenth and Twentieth Centuries*. Ashgate, Aldershot, UK.
- Roth, R., Polino, M.N., 2003. *The City and the Railway in Europe*. Ashgate, Aldershot.
- Schivelbusch, W., 1986. *The Railway Journey: The Industrialization and Perception of Time and Space*. The University of California Press, Los Angeles.
- Schram, A., 1997. *Railways and the Formation of the Italian State in the Nineteenth Century*. Cambridge University Press, Cambridge.
- Schwartz, R., Gregory, I.N., Thévenin, T., 2011. Spatial history: railways, uneven development, and population change in France and Great Britain, 1850–1914. *J. Interdiscip. Hist.* 42 (1), 53–88.
- Smith, H.M., 1976. Greenwich time and the prime meridian. *Vistas Astron.* (20), 219–229.
- Smith, P. (Ed.), 1998. *The History of Tourism: Thomas Cook and the Origins of Leisure Travel*. Routledge/Thoemmes Press, London.
- Treccani, 1932. *Ferrovie*. In: *Enciclopedia Italiana di scienze, lettere ed arti*, vol. 15, Istituto della Enciclopedia Italiana, Rome, pp. 123–159.

Further Reading

- Atack, J., Margo, R.A., 2011. The impact of access to rail transportation on agricultural improvement: the American Midwest as a test case, 1850–1860. *J. Trans. Land Use* 4 (2), 5–18.
- Atack, J., Historical Geographic Information Systems (GIS) Database of U.S. Railroads, 2016, Available from: <https://my.vanderbilt.edu/jeremyatack/data-downloads/>. Accessed December 2019.
- Banister, D., Berechman, Y., 2001. Transport investment and the promotion of economic growth. *J. Trans. Geog.* 9 (3), 209–218.
- Bogart, D., 2009. Nationalizations and the development of transport systems: cross-country evidence from railroad networks, 1860–1912. *J. Econ. Hist.* 69 (1), 202–237.
- Bogart, D., 2014. The transport revolution in industrializing Britain. In: Floud, R., Humphries, J., Johnson, P. (Eds.), *The Cambridge Economic History of Modern Britain: Volume I, 1700–1870* 4th ed, Cambridge University Press, Cambridge, pp. 8, pp. 53–8.
- Cambridge Group for the History of Population and Social Structure, 2019. Transport, Urbanization and Economic Development in England and Wales, c. 1670–1911. Available from: <https://www.campop.geog.cam.ac.uk/research/projects/transport/data/railwaystationsandnetwork.html>, Accessed December 2019.
- Coatsworth, J.H., 1979. Indispensable railroads in a backward economy: the case of Mexico. *J. Econ. Hist.* 39 (4), 939–960.
- Crafts, N., Mills, T.C., Mulatu, A., 2007. Total factor productivity growth on Britain's railways, 1852–1912: a reappraisal of the evidence. *Explorat. Econ. Hist.* 44 (4), 608–634.
- Donaldson, D., 2018. Railroads of the Raj: estimating the impact of transportation infrastructure. *Am. Econ. Rev.* 108 (4–5), 899–934.
- Free, D., 2008. *Early Japanese Railways, 1853–1914: Engineering Triumphs that Transformed Meiji-Era Japan*. Tuttle Publishing, Tokyo.
- Gourvish, T.R., 1986. *British Railways 1948–73. A Business History*. Cambridge University Press, Cambridge.

- Groote, P., Jacobs, J., Sturm, J., 1999. Infrastructure and economic development in the Netherlands, 1853-1913. *Eur. Rev. Econ. Hist.* 3 (2), 233–251.
- Herranz-Loncán, A., 2006. Railroad impact in backward economies: Spain, 1850-1913. *J. Econ. Hist.* 66 (4), 853–881.
- Hornung, E., 2015. Railroad and growth in Prussia. *J. Eur. Econ. Associat.* 13, 699–736.
- Köll, E., 2019. Railroads and the Transformation of China. Harvard University Press, Cambridge, MA.
- O'Brien, P., 1983. Railways and the Economic Development of Western Europe, 1830-1914. Macmillan, London.
- Quataert, D., 1977. Limited Revolution: The Impact of the Anatolian Railway on Turkish Transportation and the Provisioning of Istanbul, 1890-1908. *Bus. Hist. Rev.* 51 (2), 139–160.
- Redding, S.J., Turner, M.A., Transportation costs and the spatial organization of economic activity, 2014, NBER Working Paper No. 20235.
- Roth, R., Divall, C. (Eds.), 2015. From rail to road and back again? A century of transport competition and interdependency. Ashgate, Farnham.
- Roth, R., Jacolin, H., 2013. Eastern European Railways in Transition (19th-21st Centuries). Ashgate, Farnham Surrey.
- Rusanen, J., Kotavaara, O., Antikainen, H., 2011. Urbanization and transportation in Finland, 1880-1970. *J. Interdiscip. His.* 42, 89–109.
- Sawai, M. (Ed.) The Development of Railway Technology in East Asia in Comparative Perspective, 2017, Springer, Singapore.
- Silveira, L.E., Alves, D., Lima, N.M., Alcantara, A., Puig, J., 2011. Population and Railways in Portugal, 1801-1930. *J. Interdiscip. His.* 42 (1), 29–52.
- Stanev, K., 2013. A historical GIS approach to studying the evolution of the railway and urban networks: the Balkans, 1870-2001. *Hist. Methods Journal Quant. Interdiscip. Hist.* 46 (3), 192–201.
- Tang, J., 2014. Railroad Expansion and Industrialization: evidence from Meiji Japan. *J. Econ. Hist.* 74 (3), 863–886.

The Geography of Rail Transport

Frédéric Dobruszkes*, Amparo Moyano†, *Université libre de Bruxelles (ULB), Brussels, Belgium; †Universidad de Castilla La Mancha (UCLM), Ciudad Real, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	427
The Overall Geography of Rail Transport	427
Rail Geography as a Production of Multiple Factors	429
Geo-Economic Factors Versus Economic History	429
Physical Geography	431
Urban Systems	432
Distances, Travel Times, and Intermodal Competition	432
Railway Logics	433
Political Factors	433
Conclusions	435
Acknowledgment	435
Biography	435
References	435

Introduction

Railways as a system began with the introduction of the steam-powered locomotive in the early 19th century and the opening of the first public intercity railway: the Liverpool and Manchester Railway in the United Kingdom in 1830. This was the first guided transport mode that combined steel rails/steel wheels, steam engine locomotives, and a convoy. It was also the first mode of transport to be fully timetabled. Similarly, in many other countries, contemporary railways were built during the 19th century and, in specific contexts, were associated mainly with the development of different economic situations.

Surprisingly, the geography of railways systems has not been extensively investigated. In several countries, railways are part of the landscape, so scholars generally have not questioned their presence. In other countries, there are no railways anymore or only some residual rail operations, which are not considered.

In this context, the chapter presents the current overall geography of rail transport and offers an in-depth examination of the different factors that influence the construction, operation, and/or maintenance of railway systems around the world. The chapter is organized as follows: section “The Overall Geography of Rail Transport” analyzes the current situation of railway systems and their development, considering both infrastructure and use-related variables; section “Rail Geography as a Production of Multiple Factors” discusses the main factors affecting the development of railway systems; and last section concludes the chapter.

The Overall Geography of Rail Transport

While railways are commonplace in many countries worldwide, their geography varies significantly in terms of network length, density, and utilization (traffic intensity and passengers versus freight markets). In an attempt to introduce overall railway geography, country-level statistics supplied by the International Union of Railways (UIC) have been considered to build a typology of railways systems. Four main variables have been considered: the density of rail lines (DensityLines) and the density of high-speed lines (at least 250 kph) both in km/km² (DensityHSL); passengers’ traffic (Pkm_km) and freight traffic (Tkm_km), both in relation to the total length of rail lines.

This led to six main groups (Fig. 1):

- Medium-sized countries with a high density of rail lines (cluster 1)*: In this group, mainly countries in Eastern Europe and the United Kingdom are included. The density of railways is the highest of the sample and traffic per kilometer is limited, slightly over the median value (see Fig. 1: boxplot analysis).
- Medium-sized countries with both conventional and high-speed lines (clusters 2 and 3)*: These two clusters encompass most of the countries in Western Europe and South Korea (cluster 2), in addition to Japan and Taiwan (cluster 3), where the highest density of high-speed rail networks can be found. The main difference is that those countries in cluster 3 present high population densities along major rail corridors, and the use of railways is more consolidated, which implies a higher Pkm_km value.
- Large or very large countries with high freight traffic (clusters 4 and 5)*: In this case, due to the large size of the countries included in this group, the density of rail lines is very low compared to the above-mentioned countries. Of course, this does not prevent dense railways networks in densely populated areas, like Eastern China (see below). However, this group is characterized by high volumes of freight

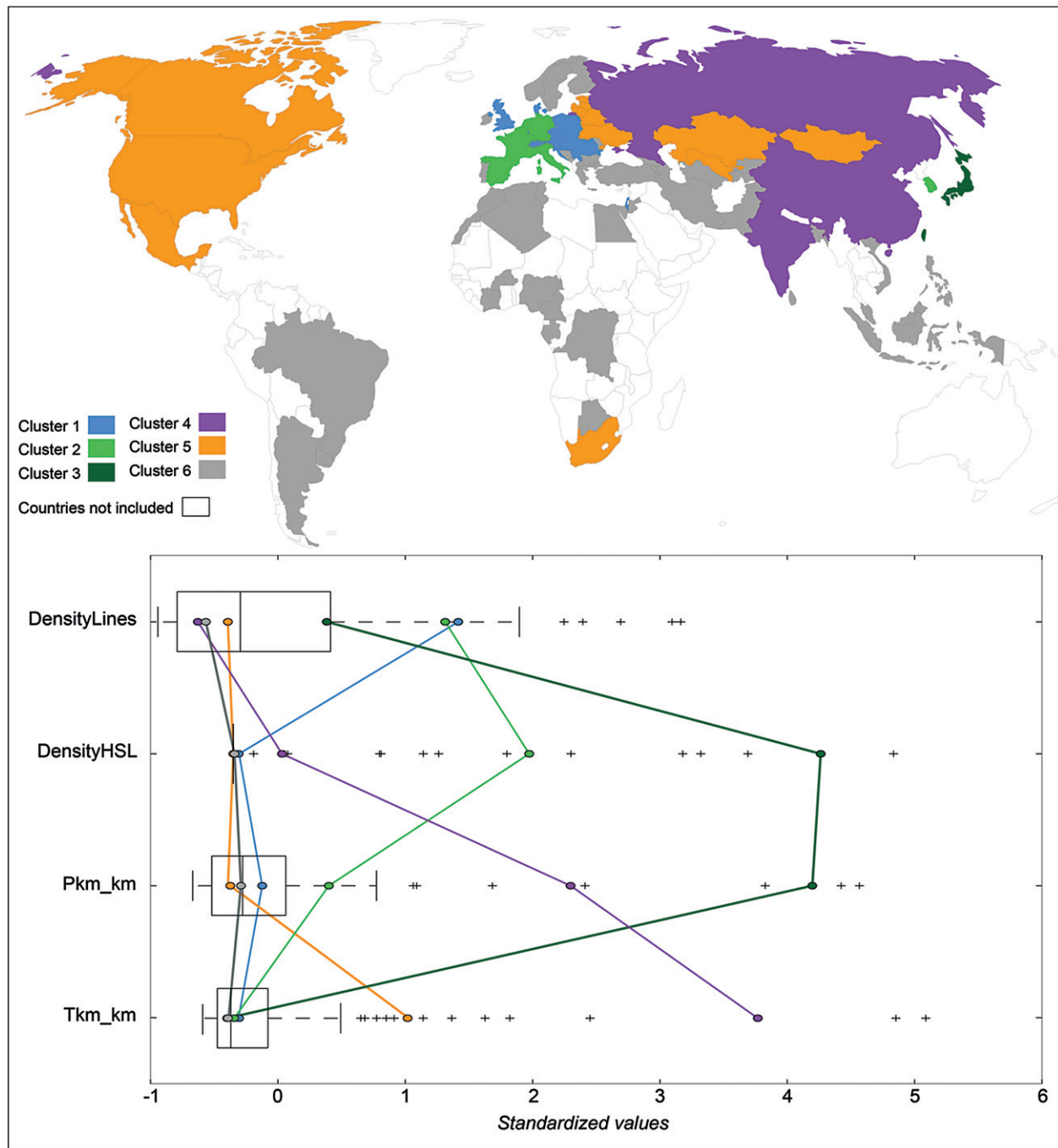


Figure 1 Classification of the world's countries: clustering and boxplot analysis (based on using the *k*-means clustering technique and data from UIC 2016 as a compromise between actualized data and the sample size).

traffic (in relation to their length of rail networks). The main difference between clusters 4 and 5 is passenger traffic, which is much higher in cluster 4 (namely, Russia, China, and India) and follows similar patterns to those found in clusters 2 and 3.

- d. *Large countries with reduced railway use (cluster 6)*: In this group, countries present a very low density of rail infrastructure (even nonexistent in some cases) and reduced use of the railway system, both for passengers and freight.

In addition to the overall geography of the world's countries, close attention must be paid to the current situation of high-speed rail systems (Fig. 2):

Currently, there are 13 countries with high-speed rail infrastructure for speeds over 250 km/h (mostly included in clusters 2 and 3). In this sense, China is the world's leader in HSR extension (21,782 km length), followed by Japan (2848 km), France (2734 km), and Spain (2482 km). However, the situation concerning the density of lines (*DensityHSL* variable) is quite different: while Japan,

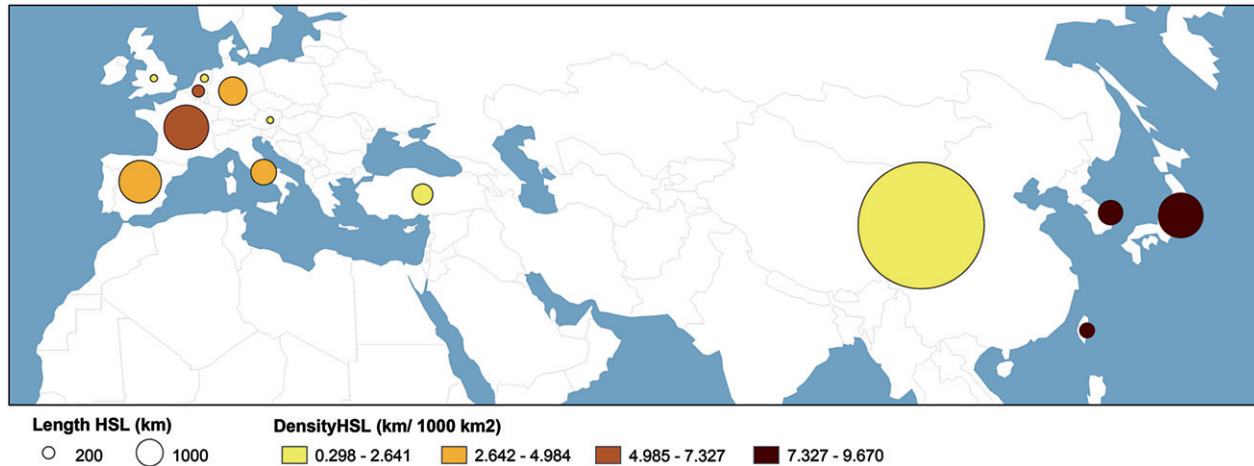


Figure 2 World's countries with high-speed rail infrastructure (at least 250 km/h) in 2018. *Source: UIC.*

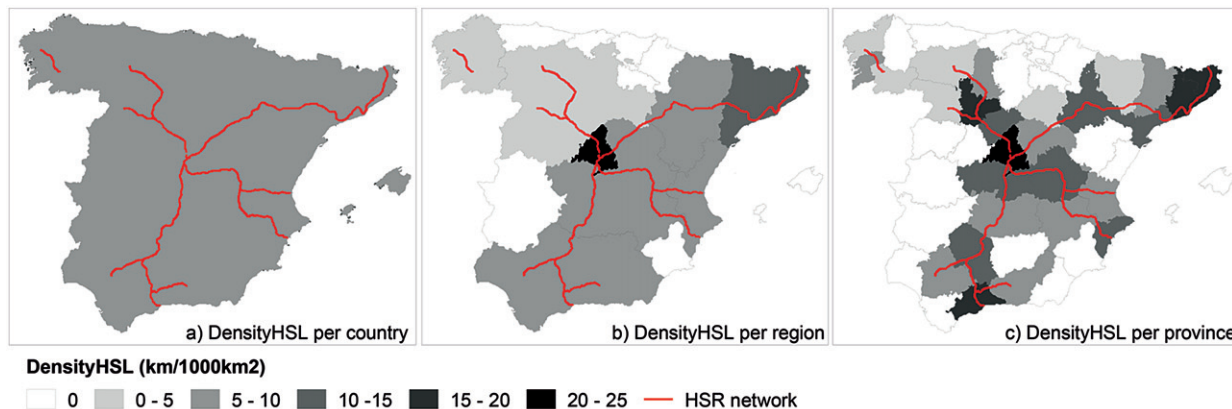


Figure 3 The impact of spatial unit on indicators: high-speed lines in Spain (at least 250 km/h). *Sources: Data from ADIF (Spanish Infrastructure Administrator) and FFE (Spanish Railway Foundation), 2019.*

South Korea, and Taiwan present the highest density values of this kind of infrastructure (over 7.327×10^{-3} km/km²), China, Turkey, Austria, The Netherlands, and the United Kingdom have less than 2.641×10^{-3} km/km².

Of course, any analysis of the geography of rail systems depends on spatial units. This chapter is not the place to elaborate on the so-called Modifiable Areal Unit Problem (MAUP) (Arbia, 1989). But for sure, the picture one gets changes dramatically, subject to spatial units, in line with urbanization patterns that are better rendered at subnational level (Fig. 3).

It is also worth noting that in countries with more than one single rail line, traffic density can be quite diverse within a single network. In France, typically, the lowest-density lines account for 46% of network length but only 6% of the supply. Conversely, 78% of the supply is concentrated in only 30% of the network (Table 1).

Rail Geography as a Production of Multiple Factors

The geography of rail transport unveiled in the previous section obviously does not fall from the sky. It is the product of numerous factors that are discussed below, with a special emphasis on rail networks.

Geo-Economic Factors Versus Economic History

It would be tempting at first to expect that countries' demographic weight and surface would favor the development of railways, notwithstanding internal heterogeneity. However, developing and operating railways is expensive. All other things being equal, economic development is thus also a key factor. Economic development would directly increase passenger and freight demand and would suggest that public authorities or private interests would have enough means to build the infrastructure.

Table 1 Network vs. supply concentration in France

Line group	Length	Supply
Main lines (UIC 1-4 and HSLs)	30% (8900 km)	78%
Intermediate lines (UIC 5-6)	24% (7000 km)	16%
Thin lines (UIC 7-8)	46% (13,600 km)	6%

Supply is given in gross ton-kilometers hauled (GTKM).

Source: Rivier and Putallaz (2005).

Table 2 Regression models for Ln railway length at the country level ($n = 82$)

	Model I	Model II	Model III
Intercept	-2.034*	-6.570***	-6.762***
Ln POPULATION	0.617***	0.689***	0.401***
Ln GDP/CAPITA		0.707***	0.768***
Ln AREA			0.379***
Adjusted R ²	0.498	0.692	0.794

Significant at *** 99%, ** 95% or * 90% level of confidence

For the subset of 82 countries considered above, we set up a multiple regression model where the length of railways is a function of population, area, and GDP/capita. Following the logarithmic transformation of all variables, the parameter estimates should be interpreted as elasticities. The first factor is population (proxy for market size); then comes GDP/capita. Once controlling also by surface area, railway length has the highest elasticity with GDP/capita (a 1% increase in GDP/capita would increase railway length by 0.77%, all other things being equal) (Table 2).

However, such a multiple regression does not really make sense. Indeed, it is worth noting that many of today's rail lines were actually built a long time ago and would arguably not be built from scratch today, given the development of other transport modes in the intervening time.

Following Kondratiev's theory, the economy experiences long-term waves. Although this theory has been challenged, each cycle involves major innovations that will drive the next period of high growth (A phase). The second wave, more or less in the second part of the 19th century, was based largely on railways, which were both a major innovation and a main industrial activity and were intensively utilized by the then heavy industries. As a result, countries that took part in the industrial revolution at this time were very quickly covered by a high grid of railways, of which many are still operated today (Martí-Henneberg, 2013). A typical example is Belgium (Fig. 4). The network has been built mostly over the second Kondratiev wave, then consolidated over the third wave, and significantly dismantled during the fourth wave, when car ownership and motorways became standard. This process was only balanced by the opening of some high-speed and freight lines over the past 25 years. Given the cost of maintaining and operating railways and intermodal competition (see below), it is likely many existing lines would not still exist if they had to be built today from scratch.

In another context, many railways in Africa were designed by the colonizer. British explorer Stanley, who was recruited by King Léopold II of Belgium to explore and acquire lands in Congo, stated that "without the railroad, the Congo is not worth a penny" (Deckard, 2019: 246). Railways were essential to overcome non-navigable parts of the Congo River. More generally, colonizers in Africa built railways to control lands and to transfer huge amounts of minerals and agricultural resources to metropolitan areas. This ensured penetration lines with low density and scarce interconnections (Taaffe et al., 1963), often until today (Oliete Josa and Magrinyà, 2018). Several lines have ended (even though they are still part of local landscapes), but surviving lines mostly date back to colonial times.

During the 20th century, especially in the second half, the evolution of railways was oriented mainly toward high-speed rail systems. The first high-speed line was the Tokaido Shinkansen line, built in 1964 between Tokyo and Osaka in Japan. Since then, high-speed lines have been developed in many other countries, mainly in Western European and Asian countries. High-speed lines (HSL) were designed initially as a transport alternative to intermetropolitan connections, which were working close to maximum capacity and facing growing traffic demand (Vickerman, 2009). Therefore, in the beginning they were oriented to discharge conventional lines servicing the main corridors in different countries. However, it quickly appeared that HSR could compete with airlines mainly for business purposes, securing higher market shares and contributing to a decrease in their absolute level of use (see Dobruszkes et al., 2017, for a review). In recent decades, the development of high-speed rail systems, transitioning from single lines to a complex mode of transportation, and their coexistence with conventional railways, require a global analysis to understand the different factors affecting the evolution of this transport system.

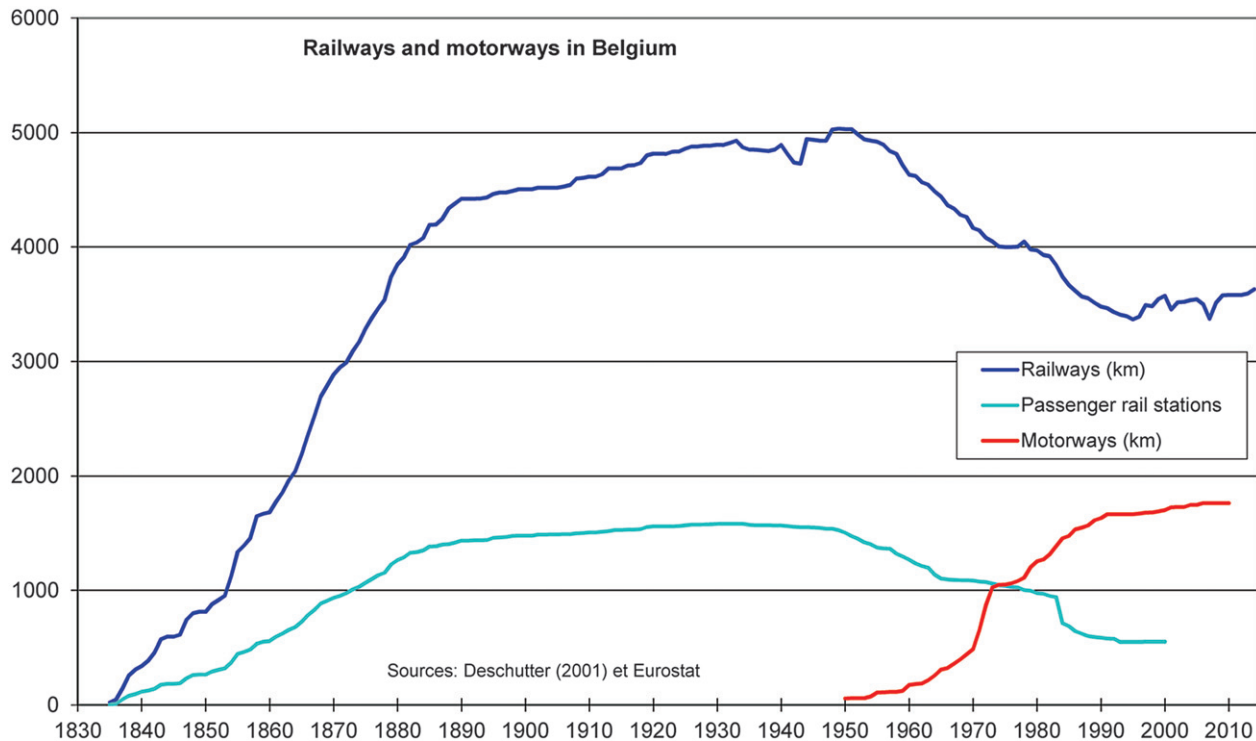


Figure 4 Belgium's railway over time.

Physical Geography

Anyone who would like to develop a railway would first face the constraint of physical geography. Major adverse constraints are water and relief. Rivers, lakes, and seas impose bridges or tunnels, which can only increase the cost of new infrastructure, as well as maintenance costs. Of course, bridges spanning seas and undersea tunnels are limited in length. In 2019, the tunnel with the longest underseabed section was the Channel Tunnel between France and England (37.9 km, out of 50.5 km). In northern Japan, the Seikan Tunnel is 53.9 km, of which 23.3 km are under the seabed. Beyond these extreme figures, it is much more common for over- or undersea rail tunnels to range from a few hundred meters to a few kilometers. There are actually more long bridges over the water than long tunnels. The longest bridge is currently the Danyang–Kunshan Grand Bridge (164.8 km long in total, with a 9 km section across the open waters of the Yangcheng Lake in Suzhou). The bridge opened in 2010 as part of the Beijing–Shanghai high-speed line. While long undersea tunnels are recent, long bridges over the water date back to the 19th century. For instance, the 9.3-km Norfolk Southern Lake Pontchartrain Bridge in the United States was opened in 1884.

Relief is not friendlier to railways. “Heavy” rolling stocks can only face gentle slopes, about 1% (or 2.5%, which is the recommended maximal high-speed rail slope). Lighter rail systems accept higher slopes (with a slope of 7.9%, Zurich’s S-Bahn 10 to Uetliberg is reported to be a record in Europe), but this may restrict operations, especially speed. Rack systems may help to overcome relief, but in most cases relief is overcome by bridges over valleys and tunnels inside mountains. Rail tunnels through mountains were first established during the late 19th century, and at rather high altitudes to make them shorter. For instance, the first Alpine rail tunnel, the Gotthard Tunnel, which opened in 1882, was already 15 km long. Its highest point is 1151 m. In addition to high construction costs, its altitude involves rather inefficient access and egress ways, with curves (if not serpentes) and bridges to maintain acceptable slopes. Length has been limited by technology and the use of steam engines, which were difficult to manage in urban underground systems (Dennis, 2013). More recently, some countries have engaged in building so-called base tunnels. In short, these flat tunnels are built at much lower altitudes, but are thus much longer and more expensive. The new Gotthard base tunnel is 57 km and is currently the longest and deepest tunnel in the world. Its highest point is some 602 m below the older tunnel’s one. Its cost was estimated at CHF 9.56 billion in late 2015 (€8.75 billion euros at that time).

Soils may also make the building of railways more complex. In Scotland (UK), for instance, the construction of the West Highland Line had to confront the problem of peat bogs, which eventually involved the application of specific techniques (Le Vay, 2009).

All this significantly increases the cost of developing railways. In the first instance, the rationale for building railways, despite adverse physical geography, would be balanced by the wealth of countries (public authorities and/or private investors), potential market size, and modal alternatives, which can all change over time. An alternative is to avoid or skirt around obstacles at the cost of detours of all scales. Some stakeholders have proposed restricting HSLs to single-track lines in case of very high infrastructure costs and moderate traffic (Castillo et al., 2015).

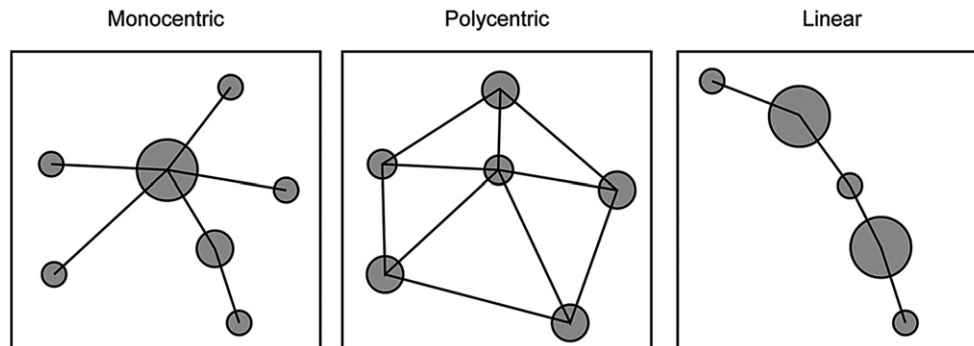


Figure 5 The impact of an urban system on a high-speed rail network.

Urban Systems

The previous modeling exercise, although apparently successful, does not reveal the subtle impact of urbanization patterns on the geography of rail systems. For instance, China is a large country by surface and population, and its GDP per capita is close to the world's average. But its rail network is apparently less developed than expected because the Chinese population is highly concentrated. In other words, a large part of China is not heavily populated, so railways are scarce in these areas. Actually, the geography of railways could be better understood under the framework of urban systems. An urban system is a combination of city size and the intensity of interaction between them. All other things being equal, this significantly shapes the need for travel, and thus the geography of flows and possibly the rationale for railways. For instance, the national urban system's shape has affected the design of high-speed lines in recent decades. **Fig. 5** clearly shows that in a country where one city dominates the urban system (typically Paris in France and, to a lesser extent, Madrid in Spain), the HSR would be star-shaped. At best, radial lines would be interconnected through the city center (Madrid) or its suburb (Paris), but they are built primarily to link the dominant city to other cities. In contrast, a polycentric urban system would spread the demand across more links, which calls for more rail lines, although it may also make them less economically or even socially profitable, given the cost of the infrastructure. Germany is a typical case here. A much easier case is when the urban system is mostly linear, as in Japan or Italy. In such cases, one single, high-speed line makes it possible to serve a significant share of the national population. In extreme cases, however, the demand along this axis may become so large that a second line is needed. This is the case in Japan, where traffic forecasts had the effect of justifying the construction of the Chūō Shinkansen maglev line between Tokyo and Osaka (which is expected to be partially opened in 2027).

Obviously, the shape of urban systems not only influences the spatial patterns of infrastructures, but also services and passenger flows. The territorial situations of HSR cities (Ureña et al., 2012) in relation to other cities in the network suggested the implementation of different kinds of services adapted to each connection's needs. For instance, in Spain, the existence of regional HSR services between large metropolitan areas and close smaller cities is the cause/consequence of interactions between cities, mainly related to labor markets (in many cases fostered by the new infrastructure itself); or even new low-cost HSR services are oriented to cover the increasing demand for leisure mobility.

Distances, Travel Times, and Intermodal Competition

In addition to urban systems (and linked to them), the geography of railway lines and their use is also affected by distances—and thus travel time—between places and related intermodal competition. Once travel time/distance becomes too long, airlines start to gain market share. Considering 161 city-pairs across Europe, Dobruszkes et al. (2014) found the ability of HSR to compete with airlines decreases significantly between 2.0 and 2.5-h of HSR travel time. Conversely, the Brussels case shows that rail use to commute to Brussels increases with distance. Commuters who live in close suburbs underuse trains, usually because the service is poor, but likely also because of lifestyles connected to social attributes (Strale, 2019). Of course, the terms of competition between railways, road transport, and air travel also depend on factors other than travel time (Zhang et al., 2019). At a time when medium- and long-distance accessibility is considered a key element of the attractiveness of cities and regions, intermodal competition goes beyond infrastructure and travel times to also consider services (routes, frequencies, and timetables) (Moyano and Dobruszkes, 2017; Moyano et al., 2018). In addition, potential travelers' income vs. fares is key. There are countries in which trains have remained a transportation mode for the masses. In other cases, especially in the case of HSR, premium fares apply. All other things being equal, this favors competition from other transportation modes.

Railway development and operations have become increasingly challenged by the rapid expansion of air services and decreasing airfares. In many areas, airlines have opened multiple new routes, which increases pressure on railways. Competition is direct (for those passengers whose destination is fixed). But it is also indirect as cheap flying options can divert passengers from trains because they offer new destination options. For instance, a cheap flight from London to Lisbon may divert passengers from British resorts. Especially in Europe, lots of long-distance rail services, both domestic and international, and both daytime and overnight, have disappeared. International rail services from Brussels (**Fig. 6**) are a typical example. At best, some of them have become HSR services. Note that the advent of HSR may also have created links that did not exist before, typically to London via Lille. Precisely, in contrast

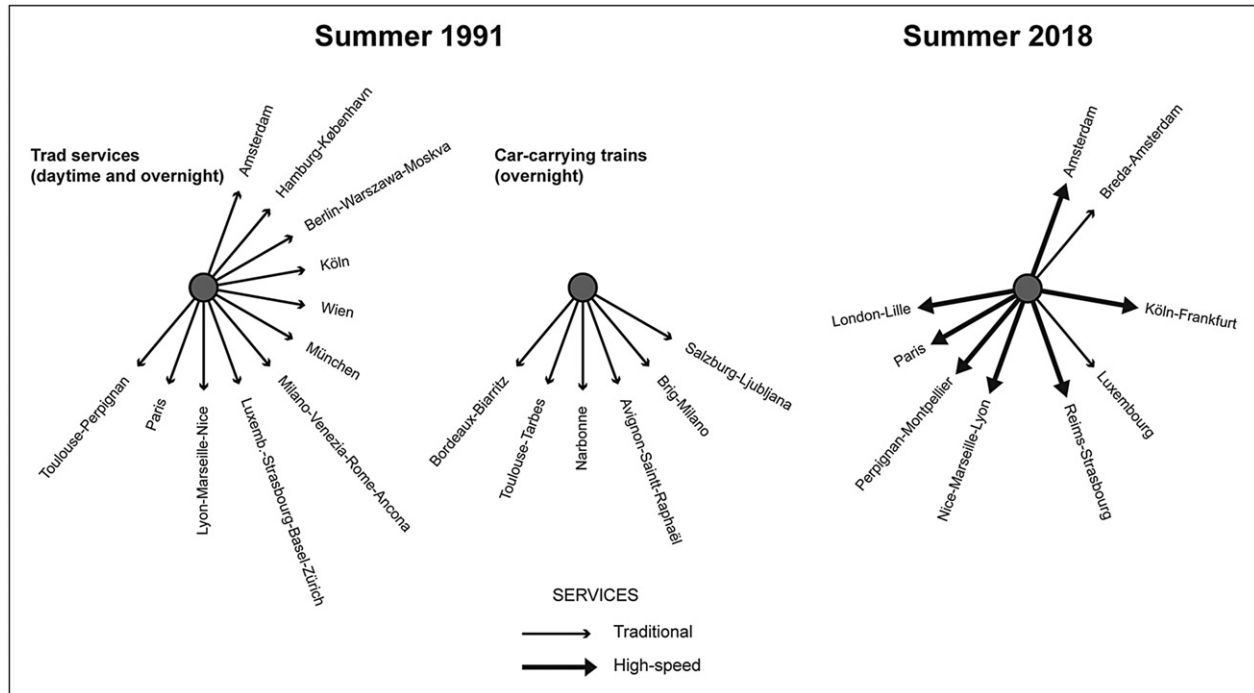


Figure 6 Changes in international services from Brussels.

with air transportation, where the flexibility to fit supply and demand is higher because connections are city-to-city, HSR runs through the territory and potentially generates impacts on the in-between regions. This implies that, in time, HSR acquired a social and political compromise for servicing smaller intermediate cities (Moyano and Coronado, 2018), guaranteeing that smaller communities would remain accessible and (or) that their long-distance accessibility would improve.

It is worth noting, however, that the terms of competition can change over time. Today's domination of many markets by air travel could later be challenged in the event of a sharp increase in energy costs or if the implementation of strong regulation forced passengers to reconsider trains in the name of environmental issues (Dobruszkes et al., 2017).

Railway Logics

The current shape of railways is also influenced by the essentials of this transport infrastructure itself. Many of the existing lines are the consequence of different steps of the development process: the railway system was not built in one go but is the result of consecutive expansions. Therefore, preexisting lines condition future corridors and branches. In the case of HSR, the logics of infrastructure growth are strongly related to preexisting conditions, mainly because the costs of construction are expensive. For instance, in Spain, the new connection to Granada takes advantage of the existing line between Madrid and Málaga (which is still below its capacity of exploitation), and only an HSR branch from Antequera is newly built. For sure, if the Andalusian corridor did not exist, the connection between Madrid and Granada would be different.

Political Factors

All the previous factors can be considered natural constraints and market forces, which likely prevailed in earlier times, when private investors took the risk of developing railways and rail travel was nearly the only efficient option. But times have changed. In many countries, the development and maintenance of railways are largely (if not exclusively) funded by public authorities, which also compensate companies for the lack of financial profitability of many services. Public authorities have many rationales for funding railways; these range from social goals (favoring the masses' right to mobility) to environmental concerns (rail operations are "greener" than other transportation modes if traffic is dense enough) to economical concerns (e.g., supporting one's national seaport against competitors) to land planning considerations (e.g., serving remote areas and increasing national or international integration). But of course, the one who pays also decides. As a result, the rationality of natural constraints and market forces is then partially replaced by political considerations. Obviously, in a democracy, governments have the power to fund facilities or services to achieve various goals, whatever the views of many transportation experts, especially liberal economists. In one sense, economical rationality is partially replaced by political rationalities. Of course, the border between dreams and the actual impact of such

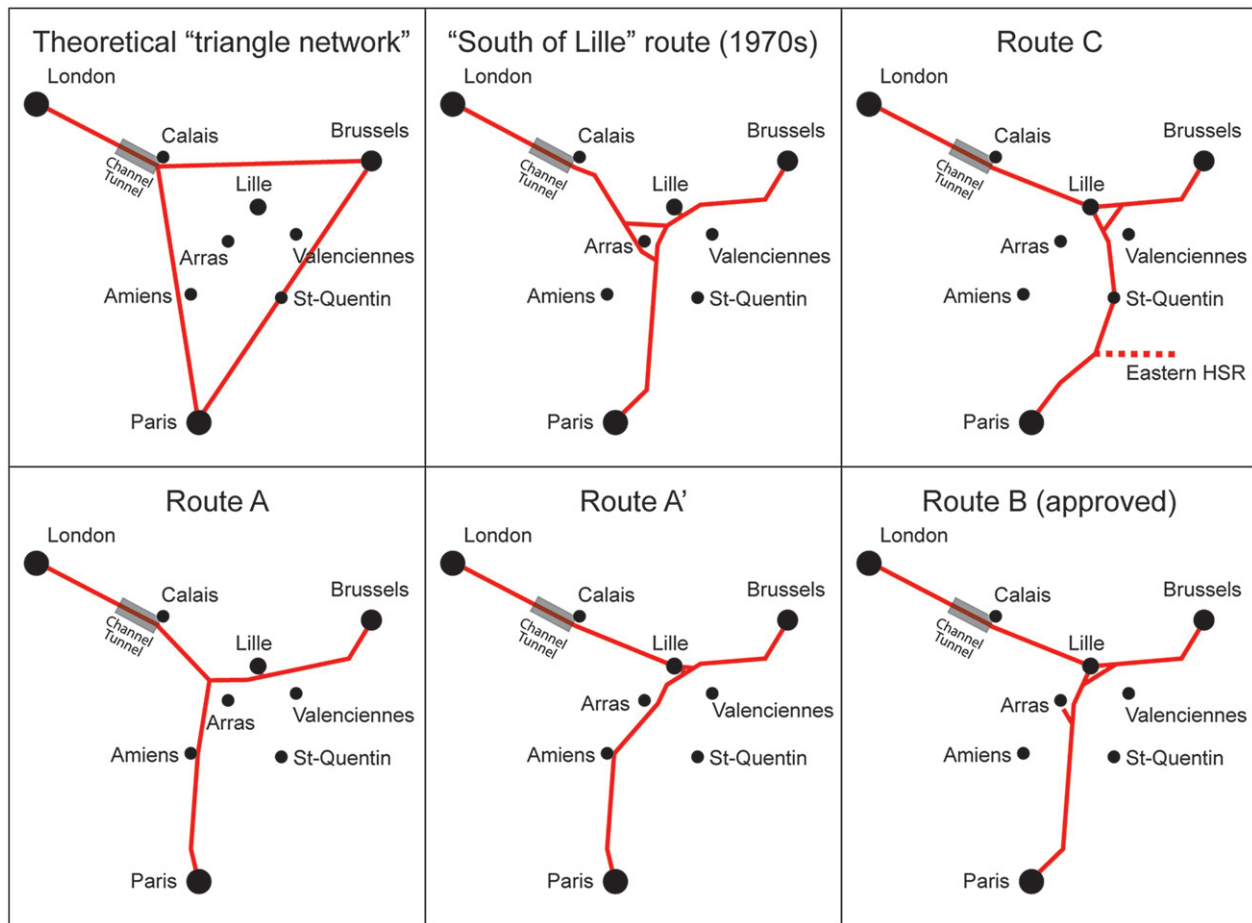


Figure 7 Potential itineraries for the Northern European HSR Source: Adapted from Moyano and Dobruszkes (2017), after SNCF (1998).

investments is somewhat unclear. It is thus not surprising that economists have largely opposed investments they deem socially unprofitable (see Albalade and Bel, 2012).

In the meantime, there are many examples of public authorities diverging from market forces, for the best and the worst. For instance, Spain first built an HSL between Madrid and Seville in anticipation of the 1992 Universal Exposition of Seville. The authorities argued that the less-developed south needed to be boosted. As a result, the Madrid-Barcelona HSL, which links Spain's two largest cities by any metric, was opened more than a decade later. There's no need to be an expert in spatial economics to understand that the largest market (and higher potential for the modal shift from planes) was between Madrid and Barcelona. Germany is also an interesting case. As long as the country was divided, HSLs were developed only within Western Germany, following a north-south logic. Then, in 1989, the Berlin Wall fell and Eastern Germany's communist regime collapsed in the aftermath. Germany was reunified in 1990, with Berlin (located in the middle of former Eastern Germany) as the capital. As a result, it became necessary to link Berlin with the western cities by HSR, so the priority became the construction of an east-west HSL, which opened in 1998.

The weight of political factors can play out at various levels. For instance, once it has been decided to link two main cities by HSR, the public authorities have to agree on an itinerary. The choice is often a mix of all the factors highlighted in this section. But because HSR has become an appealing facility that is supposed to boost the economy (Delaplace, 2017; Vickerman, 2018; Zhang et al., 2019), influential elected persons located en route would want to secure some services too and thus influence the HSL's shape (Moyano and Dobruszkes, 2017). Fig. 7 clarifies this in the case of the so-called Northern European HSR, designed primarily to connect Paris to the Channel Tunnel but also to Lille, Brussels, and beyond (Amsterdam and Cologne). Without any cost constraints, a triangle going straight from Paris to the Channel Tunnel and to Brussels may have been considered. But this would have maximized the length of the HSL, and thus the cost. Furthermore, Lille, which is one of the largest of France's cities, would have been bypassed. Various HSL designs via Lille were analyzed (Fig. 7). The final compromise mixes attractive travel times between the main cities and reasonable costs, on the one hand (market force), and Lille's location in a central position, if not an HSR hub, on the other hand (notwithstanding the real services that would later serve the city, see Moyano and Dobruszkes, 2017).

The Northern European HSR was partly justified by its contribution to European integration. But railways are like the Janus face, and they can also serve much darker designs. Existing railways were intensively used by the Nazis to send people to concentration and extermination camps. And under the Apartheid regime in South Africa, railways were a concrete tool the government used to manage spatial-racial

segregation (Baffi, 2014; Pirie, 1987). Non-white persons were forced to live in townships. Such suburban townships were either built along railways, or suburban railways were built to serve townships. The result of this design was a deliberate separation of nonwhite and white residents that still allowed for the exploitation of the segregated workforce via suburban line transport. In contrast to Jefferson's idea of "civilizing rails," Pirie (1982) was right to conclude that "decivilizing rails" existed in Southern Africa.

Conclusions

This chapter presents an overall geography of railways worldwide, identifying a typology of countries, which depends on the characteristics of their national rail systems, and evaluating the main factors affecting the spatial patterns of both rail infrastructures and traffic. These main factors are related to natural/physical constraints, market forces, and political interests and, although they are discussed separately in this chapter, all of them are interconnected: some factors influence others in one way or another. In addition, not only railways are influenced by the presented factors; transport systems can also reinforce or even create certain logics: for instance, while urban systems condition railway infrastructure and services, a good railway connection can influence inter-urban relationships (a new HSR connection with regional services can foster new mobility patterns). In this sense, this chapter exposes the relevance of these factors in the development of railways, and how their importance and relationship have been changing depending on the context (geographical, economic, political, etc.) during decades of evolution.

Acknowledgment

We would like to thank the Paris-based UIC team for letting us access the data utilized in Section 2.

Biography



Frédéric Dobruszkes holds a PhD in geography (2007) from the Free University of Brussels (ULB). He is currently an FNRS research associate and lecturer at the same institution. His main research interests are transport geography (especially long-distance mobilities, urban transport, and social conflicts around aircraft noise) and transport policy, including the design of public transport networks and accessibility issues faced by persons with reduced mobility.



Amparo Moyano holds a PhD in civil engineering (2018) from the University of Castilla La Mancha (UCLM), Spain. She is currently an assistant professor and researcher in the School of Civil Engineering of the same. Her main lines of research relate to transport and urban planning, especially focused on (1) territorial effects of high-speed rail systems and their influence on mobility patterns and (2) urban mobility and accessibility analysis using new data sources (Big Data), mainly oriented to the integration of rail stations into urban structure and local transport systems.

References

- Albalade, D., Bel, G., 2012. *The Economics and Politics of High-Speed Rail. Lessons from Experiences Abroad*. Lexington Books, Plymouth.
- Arbia, G., 1989. *Spatial Data Configuration in the Statistical Analysis of Regional Economics and Related Problems*. Kluwer, Dordrecht.
- Baffi, S., 2014. Chemins de civilisation? Le rail dans les politiques territoriales en Afrique du Sud. *L'Espace géographique* 43 (4), 338–355.

- Castillo, E., Gallego, I., Sánchez-Cambronero, S., Menéndez, J.M., Rivas, A., Nogal, M., Grande, Z., 2015. An Alternate double-single track proposal for high-speed peripheral railway lines. *Comput. Aided Civil Infrastruct. Eng.* 30 (3), 181–201.
- Deckard, S., 2019. Trains, stone, and energetics: african resource culture and the neoliberal world-ecology. In: Sharae Deckard, Stephen Shapiro, (Eds.), *World Literature, Neoliberalism, and the Culture of Discontent*. Palgrave Macmillan, Cham, pp. 239–262.
- Delaplace, M., 2017. Assessing ex-ante the wider effects of high-speed rail service in cities? The lessons drawn from a service innovation-based analysis. *Eur. Rev. Serv. Econ. Manage.* 3, 105–130.
- Dennis, R., 2013. Making the underground underground. *Lond. J.* 38 (3), 203–225.
- Dobruszkes, F., Dehon, C., Givoni, M., 2014. Does European high-speed rail affect the current level of air services? An EU-wide analysis. *Transport. Res. A: Policy Pract.* 69, 461–475.
- Dobruszkes, F., Givoni, M., Dehon, C., 2017. Assessing the competition between high-speed rail and airlines: a critical perspective. In: Daniel Albalade, Germà Bel, (Eds.), *Evaluating High-Speed Rail: Interdisciplinary Perspectives*. Routledge, pp. 159–174.
- Le Vay, B., 2009. *Britain from the Rails*. Bradt Travel Guides, Bucks.
- Marti-Henneberg, J., 2013. European integration and national models for railway networks (1840–2010). *J. Transport Geogr.* 26, 126–138.
- Moyano, A., Dobruszkes, F., 2017. Mind the services! High-speed rail cities bypassed by high-speed trains. *Case Stud. Transport Policy* 5 (4), 537–548.
- Moyano, A., Coronado, J.M., 2018. Typology of high-speed rail city-to-city links. *ICE-Transport* 171 (5), 264–274, doi:10.1680/jtran.16.00093.
- Moyano, A., Martínez, H.S., Coronado, J.M., 2018. From network to services: a comparative accessibility analysis of the Spanish high-speed rail system. *Transport Policy* 63, 51–60, doi:10.1016/j.tranpol.2017.11.007.
- Oliete Josa, S., Magrinyà, F., 2018. Patchwork in an interconnected world: the challenges of transport networks in sub-Saharan Africa. *Transport Rev.* 38 (6), 710–736.
- Pirie, G., 1987. African township railways and the South African state, 1902–1963. *J. Historical Geogr.* 13 (3), 283–295.
- Pirie, G., 1982. The decivilizing rails: railways and underdevelopment in Southern Africa. *Tijdsch. Econ. Soc. Geogr.* 73 (4), 221–228.
- Rivier, R., Putallaz, Y. (Eds.), 2005. *Audit sur l'état du réseau ferré national français*. Rapport, version 1.2. SNCF and RFF.
- SNCF, 1998. *Projet de desserte du Nord de la France par trains à grande vitesse*. T.G.V. NORD. Dossier d'enquête préalable à la déclaration d'utilité publique. Partie E: Évaluation socio-économique. Document kept at the French National Archives, Pierrefitte-Sur-Seine (France), mark 19980601/6, and consulted there.
- Strale, M., 2019. Travel between Brussels and its outskirts: contrasting situations. *Brussels Stud.* 137, <https://journals.openedition.org/brussels/2831>.
- Taafe, E.J., Morrill, R.L., Gould, P.R., 1963. Transport expansion in underdeveloped countries: a comparative analysis. *Geogr. Rev.* 53 (4), 503–529.
- Ureña, J.M., Coronado, J.M., Garmendia, M., Romero, V., 2012. Territorial implications at national and regional scales of high-speed rail. In: Ureña, J.M. (Ed.), *Territorial Implications of High Speed Rail: A Spanish Perspective*. Ashgate, London, pp. 129–162.
- Vickerman, R., 2009. Indirect and wider economic impacts of HSR. In: de Rus, G. (Ed.), *Economic Analysis of High Speed Rail in Europe*. Fundación BBVA, Madrid, pp. 89–118.
- Vickerman, R., 2018. Can high-speed rail have a transformative effect on the economy? *Transport Policy* 62, 31–37.
- Zhang, A., Wan, Y., Yang, H., 2019. Impacts of high-speed rail on airlines, airports and regional economies: a survey of recent research. *Transport Policy* 81, A1–A19.

Rail Transport and Territorial Scale

Prof. Andrés Monzón, Dr. Elena López, Transport Research Centre, TRANSyT-Universidad Politécnica de Madrid, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	437
Territorial Effects of Rail Transport	437
Territorial Equity Effects	439
Micro-Level Analysis: First/Last-Mile Links	441
Summary and Conclusion	442
References	442
Further Reading	443

Introduction

Transport infrastructure networks are the main structuring element in a territory. In order to serve as such, it is of prime importance to guarantee access to the network, which varies with each transport mode. Rail, maritime, and air transport networks have only discontinuous access through their nodes, in the form of stations. This creates a special relation with the territory, known as the tunnel effect: only stations allow entry to and exit from that “tunnel.” Although railway lines run along a territorial corridor, accessibility is only possible through stations, which become the access nodes. All the connections with the surrounding territories are fully dependent on the station location and its access by other modes: walking, cycling, public transport, or private car. Most are dependent on the road/street network that serves the rail node.

Rail transport should therefore be studied on two different territorial scales:

1. macroscale: connection with other stations (nodes in the rail network); travel time, frequency of services, etc.
2. microscale: potential accessibility of people and activities to/from stations; access and egress times, modes, congestion, reliability, etc.

Rail issues are most important at the macroscale. Rail networks are usually planned at national or international strategic levels, and must take into account a large number of territorial scale issues in regard to rail network planning, such as the location of the main origin/destination points and distances versus travel time through the rail network.

The scope of the following sections focuses on long and medium distance rail transport, as urban rail transport is covered in the urban transport section of this Encyclopedia, such as in 10610 Light Rail. Other relevant and interrelated readings in this Encyclopedia are 10693 High Speed Rail, 10635 Planning for Rail Transport, and 10462 The geography of rail transport.

• Stages in a rail trip

In long and medium distance travel, a trip using the rail mode can in most cases be disaggregated into at least three stages, as shown in Fig. 1: T1 (the first-mile trip) includes the connecting trip from the origin (i) to the nearest train station (A_i , entry station), T2 includes the trip on the rail service, and T3 (the last-mile trip) includes the final trip from the arrival train station (B_j) to the final destination (j).

The rail station thus acts as an intermediate—transfer—node, which is not highly attractive in itself as a final destination, but where some utility is obtained. Hence, accessibility to the nearest railway station via the local network plays an important role as a provider of “connectivity” to the long-distance network.

Railway lines therefore produce territorial impacts that are relevant for regional and local development and cohesion policies, and which must be measured to design adequate investment and management policies as part of a fully integrated intermodal vision.

This chapter is organized as follows: after Introduction, Territorial Effects of Rail Transport Section includes an overview of the main territorial impacts of rail transport, focusing on describing accessibility indicators as a valuable tool to assess these impacts. Then, Territorial Equity Effects Section discusses on territorial equity effects and measures them in the case study of the Spanish HSR network development. A micro-level analysis is then included in Microlevel Analysis: First/Last-Mile Links Section, where territorial impacts present in first/last mile links are also assessed through a case study; in this case a HSR corridor. The chapter ends with a “Summary and Conclusions” section.

Territorial Effects of Rail Transport

The territorial policy of the European Union has established a specific action on territorial cohesion with the goal of promoting a more cohesive and balanced territory, seeking greater socio-economic equity and reinforcing territorial cooperation (EC, 2008). One

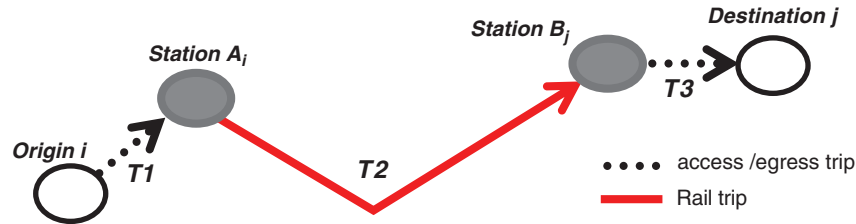


Figure 1 Travel stages including access, rail service, and egress connection to destination.

aim of the EU cohesion policy is to provide better connectivity through the so-called TEN-T (Trans-European Networks-Transport), with the additional purpose of fostering competitiveness and employment. At the core of this cohesion policy is the development of an EU-wide rail network.

However, new rail transport developments may induce spatially polarizing impacts, a phenomenon that could apply to any territory in any region. There are ad hoc methodologies to assess these impacts, such as those designed to measure changes in the spatial distribution of accessibility caused by transport infrastructure networks, and high-speed rail (HSR) lines in particular. These methods generally target strategic international or national macro levels (Banister and Berechman, 2003; López et al., 2008) and require the calculation of accessibility indicators—mainly using Geographical Information Systems (GIS)—to analyze changes in accessibility levels impacted by new rail lines (Bröcker et al., 2010).

The impacts of the territorial accessibility provided by rail networks are not limited to locations in the immediate vicinity of railway stations, and this effect is even more significant in the case of HSR networks. Travel patterns in cities outside rail corridors may also change in response to the new services (Garmendia et al., 2012). For these cities, the nearest station functions as an interchange node to connect to the whole rail network, so cities located outside the rail corridor may also obtain accessibility benefits. These territorial impacts are known as “spillover effects” (Gutiérrez et al., 2010). Cities with railway stations became core locations for access to the rail network, and thus gain accessibility to the whole region/country (Monzón et al., 2013).

- **Accessibility indicators to measure the territorial impacts of rail transport**

The analysis of how rail networks influence the territorial distribution of quality of access to desired destinations is usually supported by calculating accessibility indicators. From a spatial perspective, accessibility is a feature of a given location, related to the ease with which “desired” destinations can be reached from it. There are many formulations for measuring the accessibility provided by transport networks, all of which seek to provide an output to different study targets. An extensive review of accessibility indicators and their applications in different territorial contexts can be found in Geurs et al. (2012).

The accessibility enabled by rail networks is usually studied at national or international macro strategic planning levels. This broader scope means that accessibility indicators must be calculated taking data availability restrictions into account (basically detailed travel and mobility data), and the affordability of the corresponding computation efforts. On this coarser scale, travel times are calculated using long-distance transport networks, and major urban agglomerations are usually selected as “desired” destinations. One general formulation of an accessibility indicator for measuring macro rail territorial impacts is the “potential accessibility indicator” (Monzón et al., 2019) shown in Eq. (1):

$$PA_i = \sum_j \frac{P_j}{I_{ij}} \quad (1)$$

PA_i is the potential accessibility for each origin i to all destinations j . P_j is a variable that measures the activity level of each destination (population/GDP), and I_{ij} is the generalized travel time/cost from i to j . The formulation for the travel time is as follows (Eq. (2)):

$$I_{ij} = t_{\text{road}}(i, S_i) + t_{\text{rail}}(S_i, S_j) + t_{\text{road}}(S_j, j) + \theta_{\text{rail}} \quad (2)$$

where $t_{\text{road}}(i, S_i)$ is the access time by road (default mode) from the trip origin to the nearest station; $t_{\text{rail}}(S_i, S_j)$ is the travel time using the railway network to the station nearest the destination, and $t_{\text{road}}(S_j, j)$ is the egress time by road from this station to the destination itself. It also considers a set of penalties (θ_{rail}): frequency of service, railway line change, road to railway mode change.

The PA_i value is a measure of the absolute rail accessibility of location i , and can be used to measure the impacts of new lines or improvements in existing ones by measuring the change in the accessibility values between two horizon years. The difference in accessibility is calculated as a percentage of the change in potential accessibility. The changes between two consecutive horizon years are as follows (Eq. (3)):

$$PAC_i^{H_x-H_y} = \frac{PA_i^{H_x} - PA_i^{H_y}}{PA_i^{H_x}} \times 100 \quad (3)$$

$PAC_i^{H_x-H_y}$ is the change in the values of the potential accessibility indicator between two horizon years H_x and H_{x+1} .

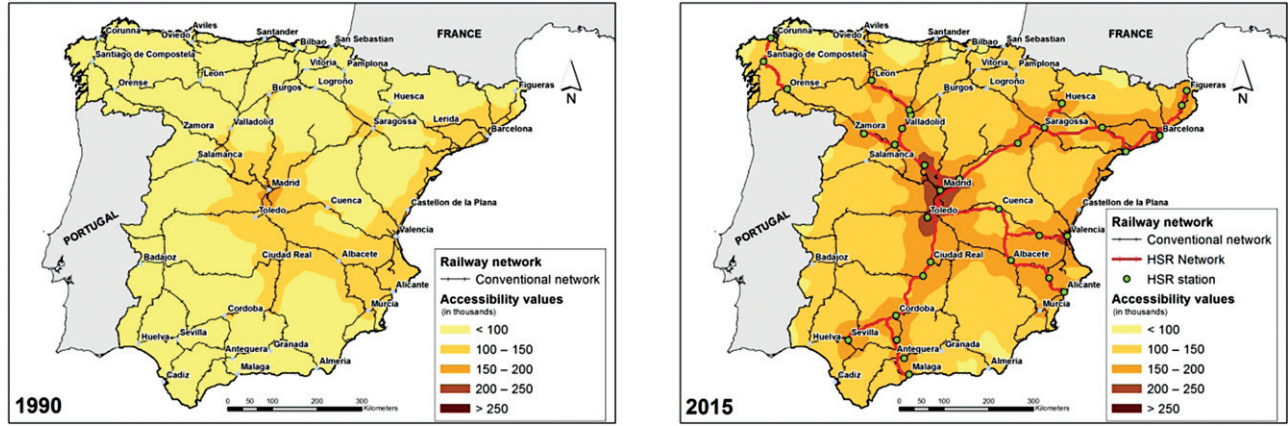


Figure 2 Absolute potential accessibility (PA) due to the rail network in the horizon years 1990 and 2015 (before and after the development of the HSR network).

A positive PAC_i value indicates an increase in accessibility in year H_y compared to year H_x . The higher PAC_i values therefore mean greater improvements in accessibility in a specific territory.

Territorial Equity Effects

The degree of territorial equity can be measured in terms of the spatial distribution of accessibility levels across a given territory. These spatial distributive effects are highly significant for the assessment of rail developments, as it is important to evaluate whether new rail corridors contribute to a more balanced distribution of accessibility. If the accessibility improvements are located in places that already have good accessibility levels, there is a risk of polarizing the territory: rich locations became richer and poor ones even poorer.

• Accessibility indicators to measure territorial equity effects

Once the differences in accessibility have been determined in the different periods of building the rail network, it is advisable to calculate whether these accessibility gains are equally distributed throughout the territory. The dispersion of accessibility values is a measure of accessibility-based territorial equity, which must be analyzed in conjunction with the overall increase in accessibility. It may be that greater accessibility for the country as a whole could be caused by very high increases in certain locations and much lower increases in others. The final result may exacerbate the differences between regions and lead to losses in territorial cohesion. This effect is measured by calculating the “potential accessibility dispersion” (PAD) index (Ortega et al., 2012) as formulated in Eq. (4):

$$PAD^{H_x} = \frac{\sigma^{H_x}}{\frac{\sum PA_i^{H_x} \cdot p_i^{H_x}}{\sum p_i^{H_x}}} \quad (4)$$

where PAD^{H_x} is the coefficient of variation in the horizon year H_x , and σ^{H_x} is the standard deviation of $PA_i^{H_x}$ values, weighted by the population $p_i^{H_x}$. High PAD values mean more polarized accessibility distributions, and less territorial cohesion. Finally, the change in territorial cohesion between horizon years can be measured by the territorial cohesion index (TC), as expressed in Eq. 5:

$$TC^{H_x-H_y} = \frac{PAD^{H_x} - PAD^{H_y}}{PAD^{H_x}} \times 100 \quad (5)$$

where TC is the change in territorial cohesion between two horizon years: H_x and H_{x+1} . A positive TC value represents an increase in territorial cohesion between the two horizon years. This value can be normalized by assigning a value of 1 to the average PA value. Once the values are mapped, they reveal which cities have higher relative accessibility values, and whether the improvements are uniformly distributed (approximate average values) between two horizon years.

• Case study: Territorial effects of 25 years of Spanish HSR network development (1990–2015)

Let us now analyze how to apply the indicators defined above to measure territorial impacts and equity. They have been used to assess the impacts of developments in the Spanish HSR network in the period between 1990 and 2015 (Monzón et al., 2019). It is currently the second longest HSR network in the world, with over 3000 km. This major network development has significantly improved rail accessibility in absolute values (PA) in the country, along with a wholesale transformation of the spatial distribution of rail accessibility values (PAC). For this analysis, the potential accessibility formulation was computed in GIS for each municipality (NUTS-5 level).

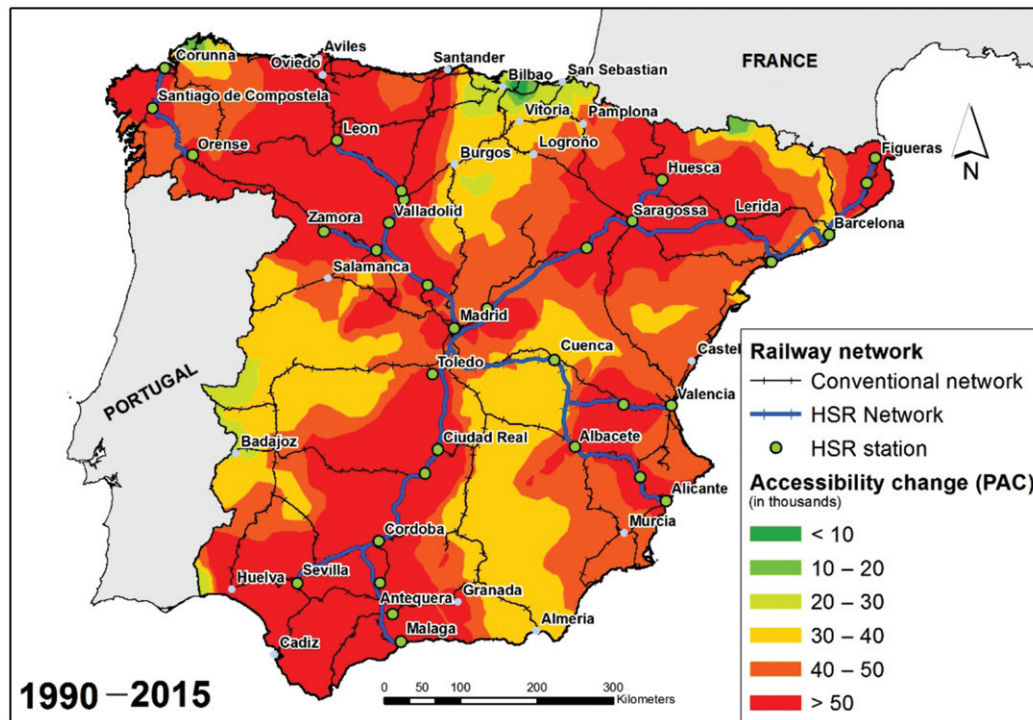


Figure 3 Potential accessibility changes (%) due to rail developments in Spain during 1990–2015.

Table 1 Accessibility values (in thousands) 1990–2015 and their changes (%)

Area/year	1990	2015	1990–2015
<i>Accessibility indicator</i>	<i>PA (000)</i>		<i>PAC (%)</i>
National average	128	190	48.6
Barcelona	210	263	25.5
Bilbao	117	134	14.5
Malaga	105	198	88.3
Madrid	282	393	39.3
Seville	123	215	74.5
Santander	90	128	47.2
Valladolid	122	236	93.9
<i>Territorial cohesion</i>	<i>PAD</i>		<i>TC index (%)</i>
National average	0.46	0.35	14.5

PA, potential accessibility; PAC, potential accessibility change; PAD, potential accessibility dispersion; TC, territorial cohesion.

The accessibility values for the PA indicator are shown in Fig. 2 for both the horizon year 1990, before the start of the HSR network, and for 2015. The figure on the right shows the values for the same indicator after the construction of some 3000 km of lines operating at commercial speeds between 250 and 350 km/h. Quantitatively, average accessibility values on the Spanish mainland have increased by 48.6% since 1990, although as these improvements are associated to the new HSR corridors, they are unevenly distributed.

The 2015 accessibility values are not only higher than in 1990, but also more balanced, with a 15% reduction in the dispersion of accessibility values: PAC (%). Fig. 3 shows a graphic representation of the percentage of change for each municipality in the 1990–2015 period.

Table 1 shows the evolution of accessibility for the whole country and for certain specific nodes. Cities with low PA values in 1990 such as Seville, Malaga and Valladolid—due to their travel time to the main centers of economic activity—received a significant impact when they became HSR stations on the new lines. Cities like Bilbao and Santander in the north of the country, where no new lines were built, received only spillover effects, which are much weaker. Large cities like Madrid and Barcelona are less dependent on their connection to other parts of the country so their accessibility improvements are not as high, even though they became nodal

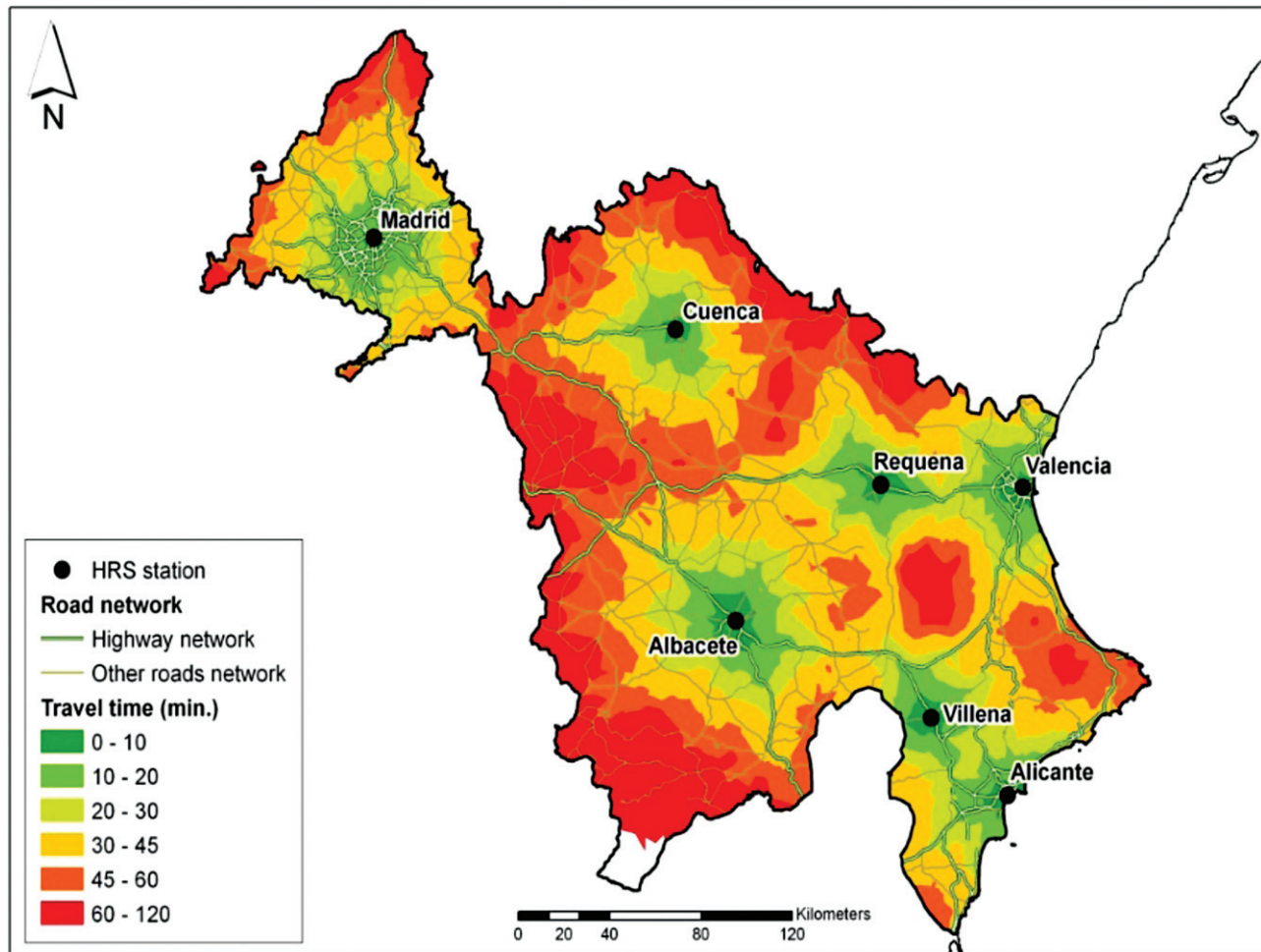


Figure 4 Isochrones to access stations in the Madrid-Valencia HSR corridor (Spain).

stations on the new HSR network. The lower part of [Table 1](#) also shows the positive territorial cohesion: less dispersion of accessibility values (PAD), representing a net increase of 14.5% in the territorial cohesion index.

Micro-Level Analysis: First/Last-Mile Links

Many studies on accessibility and spatial equity issues have been done at micro/local levels, where “desired” destinations usually refer to urban facilities such as healthcare services, playgrounds, or urban parks. This requires transport planners to efficiently interconnect railway stations with secondary transport networks and local public transport in order to reduce access/egress travel times. This also highlights the importance of quality of access to the railway station, since “most of the travel time (and effort) is spent on getting to and from the station, and this constitutes the bulk of the journey travel time” ([Givoni and Banister, 2012](#), p. 306). It is, therefore, necessary to consider the entire integrated transportation chain—the door-to-door trip—in the study of rail services. Demand for rail services can be assumed to be higher if railway stations are situated in central locations that are well connected to rail and transit systems, and the issue of first/last-mile trips is indeed one of the current challenges facing rail services ([Lane, 2012](#)). As most rail accessibility studies have been done at the macro level, accessibility derived from first- and last-mile trips has been unevenly and scarcely researched. Investments in quality local transit services to stations can improve both access to and from stations and interconnections between short- and long-distance transport networks.

Another important issue to ensure access from local transport (including walking and cycling) is the location of stations. The best transportation solution is to locate stations in dense, centralized, and highly accessible locations ([Guirao and Campa, 2014](#)). However, in already developed metropolitan areas, land availability and building restraints make it difficult to locate new stations in central business districts (CBDs). Indeed, the type of access available may pose difficulties in car-oriented cities affected by sprawl, as may occur in many US cities according to [Lane \(2012\)](#).

- Case study: Measuring the spatial distribution of access times to a railway station

Accessibility indicators are frequently used to assess access to the nearest railway station. The accessibility indicator in this case is simply travel time from each municipality i to the nearest station in the Madrid-Valencia HSR corridor in Spain (Monzón et al., 2016). The minimum time path itinerary to access the HSR station is calculated for each municipality, making possible to graphically represent the isochrones of the territory served by each station. Isochrones are often taken as accessibility indicators as they are easy to understand and show the areas that can “reach” certain destinations within a given time threshold in graphic form. In Fig. 4, the destinations are the HSR stations in the corridor, and the time thresholds were 10, 20, 30, 45, 60, and 120 min. The use of isochrones helps identify the territories with higher accessibility needs in order to decide on the best solution to improve their situation.

Summary and Conclusion

In summary, this chapter has highlighted the importance of territorial equity effects of rail transport, focusing on the important role of accessibility indicators as supporting spatial analysis tools. In addition, evidence from two selected case studies is included, providing sound assessment methodologies based on the computation of accessibility indicators. These include the case of HSR on both macro- and micro-assessment levels.

It is therefore necessary, when planning rail lines, to approach territorial scale issues from two different perspectives: macro, considering the whole corridor/region; and micro, focusing on the location of the station and its connectivity through local public and private transport networks, that is, addressing first/last mile links.

Assessment at the macroscale should aim to identify solutions that ensure improvements in accessibility levels in the region/corridor that is directly affected by the new investments, as well as in nearby regions/corridors in order to achieve wider territorial spillover effects. This requires planning large-scale rail projects following strategic assessment frameworks that consider not only rail accessibility, but also connectivity with road and air networks. A comprehensive assessment of the territorial effects of rail projects at the macroscale is thus a challenge for transport planners. Accessibility indicators are widely applied to measure this kind of territorial effect and, together with GIS tools, to map their impacts. However, new or improved railway lines and services should seek to achieve a better distribution of accessibility to promote territorial equity. Rail networks have positive structural effects when improving cohesion among different locations within the region. Every town and city is entitled to accessibility potential in order to foster their development and ensure their access to services.

The lower scale consists of assuring accessibility to stations from the surrounding territory, implying actions in two dimensions: good station location and its access by local transport modes (Ureña et al., 2009). In principle, all local transport—public transport, private cars, and nonmotorized means—should facilitate access to railway stations, which requires a joint land-use transport strategy, city design, and supporting services for travelers in the vicinity of the stations. It is therefore necessary to implement a multimodal and LUTI (land use & transport interaction) approach to transform railway stations into core elements of the urban fabric (Monzón and Di Ciommo, 2016).

References

- Banister, D., Berechman, J., 2003. The economic development effects of transport investments. In: Pearman, A., Mackie, P., Nellthorp, J. (Eds.), *Transport Projects Programmes, and Policies: Evaluation Needs and Capabilities*. Ashgate, Aldershot.
- Bröcker, J., Korzhenevych, A., Schürmann, C., 2010. Assessing spatial equity and efficiency impacts of transport infrastructure projects. *Transport. Res. Part B: Methodol.* 44 (7), 795–811.
- EC, 2008. Green Paper on Territorial Cohesion. Turning territorial diversity into strength. The Committee of the Regions and the European Economic and Social Committee, October 2008, Brussels.
- Garmendia, M., Ribalaygua, C., Ureña, J., 2012. High speed rail: implication for cities. *Cities*, 29, pp. 26–31.
- Geurs, K., Krizek, K., Reggiani, A. (Eds.), 2012. *Accessibility and transport planning*. NECTAR series on Transportation and Communications Networks Research. Edward Elgar, Cheltenham.
- Givoni, M., Banister, D., 2012. Speed: the less important element of the High-Speed Train'. *J. Transport Geogr.* 22, 306–307, doi:10.1016/j.jtrangeo.2012.01.024.
- Guirao, B., Campa, J., 2014. The construction of a HSR network using a ranking methodology to prioritise corridors. *Land Use Policy* 38, 290–299.
- Gutiérrez, J., Condeço-Melhorado, A., Martín, J., 2010. Using accessibility indicators and GIS to assess spatial spillovers of transport infrastructure investment. *J. Transport Geogr.* 18 (1), 141–152.
- Lane, B., 2012. On the utility and challenges of high-speed rail in the United States. *J. Transport Geogr.* 22, 282–284.
- López, E., Gutiérrez, J., Gómez, G., 2008. Measuring regional cohesion effects of large-scale transport infrastructure investments: An Accessibility Approach. *Eur. Plan. Stud.* 16 (2), 277–301.
- Monzón, A., Di Ciommo, F., 2016. *City-Hubs. Sustainable and Efficient Urban Transport Interchanges*. CRC Press, Taylor & Francis, Florida, USA.
- Monzón, A., López, E., Ortega, E., 2019. Has HSR improved territorial cohesion in Spain? An accessibility analysis of the first 25 years: 1990–2015. *Eur. Plan. Stud.* 27 (3), 513–532.
- Monzón, A., Ortega, E., López, E., 2013. Efficiency and spatial equity impacts of high-speed rail extensions in urban areas. *Cities*, 30, 18–30.
- Monzón, A., Ortega, E., López, E., 2016. Influence of the first/last mile on HSR accessibility levels. In: Geurs, K.T., Dentihno, T., Patuelli, R. (Eds.), *Accessibility, Equity and Efficiency: Challenges for Transport and Public Services*. Edward Elgar, Aldershot, pp. 125–143.
- Ortega, E., López, E., Monzón, A., 2012. Territorial cohesion impacts of high-speed rail at different planning levels'. *J. Transport Geogr.* 24, 130–141.
- Ureña, J., Menerault, P., Garmendia, M., 2009. The high-speed rail challenge for big intermediate cities: A national, regional and local perspective'. *Cities*, 26(5), 266–279.

Further Reading

- Chen, C.L., Loukaitou-Sideris, A., de Ureña, J.M., Vickerman, R., 2019. Spatial short and long-term implications and planning challenges of high-speed rail: a literature review framework for the special issue.
- Moyano, A., Moya-Gómez, B., Gutiérrez, J., 2018. Access and egress times to high-speed rail stations: a spatiotemporal accessibility analysis. *J. Transport Geogr.* 73, 84–93.
- Urena, J. (Ed.), 2016. *Territorial Implications of High Speed Rail: A Spanish Perspective*. Routledge, 308 p.
- Wang, L., Duan, X., 2018. High-speed rail network development and winner and loser cities in megaregions: The case study of Yangtze River Delta, China. *Cities*, 83, 71–82.

Railway Management

Vassilios A. Profillidis, Democritus University of Thrace, School of Civil Engineering, Section of Transportation, Xanthi, Greece

© 2021 Elsevier Ltd. All rights reserved.

Introduction	444
Defining Railway Management and Differentiating it From Management (in its General Sense)	445
A Definition of Management	445
Managing Railways: Dealing With a Heavy and Rigid Industry	445
Railways: Engineering Oriented or Customer Oriented	445
Understanding a Constantly Changing World	446
The Internal and External Environment of Railways	446
A Systems Approach for Railways	446
Lifestyle, Environmental Sensitivity, Artificial Intelligence, and Management of Railways	446
Understanding the Transport and Rail Market	447
An Oligopolistic Transport Market	447
The Railway Market: From Monopoly to Competition	447
A Unified or Separated Railway	447
Public or Private Railways	448
Managing Finances of the Rail Activity	448
The Revenues/Expenses Ratio	448
A Good Understanding of Costs and Yields	448
Increase of Productivity	448
A Segmentation of Traffic	449
Planning the Railway Activity	449
Planning As an Essential Prerequisite for Management	449
Business Plans and Master Plans	449
Project Management: a Necessary Tool for Most Railway Projects	449
Management of Railway Infrastructure	450
Deciding Efficient Values of Infrastructure Charges	450
Level of Outsourcing	450
Condition-Based Maintenance	450
Management of Passenger Traffic	451
Competition From Other Modes	451
Monitoring of Passengers' Exigencies	451
Increase of Traffic or Profit	452
Optimization of Finances	452
Advantages for Frequent Travelers and Targeted Segments of the Market	452
Effects of Elasticities	452
Considering Options for Door-To-Door Transport	452
Management of Freight Traffic	452
Focus on Products Suitable for Railways	452
Increase of Average Freight Speed	453
Incorporation of Rail Freight in the Logistics Chain	453
The Issue of Truck Subsidiaries of Rail Freight Companies	453
Summary and Conclusion	453
References	453
Further Reading	453

Introduction

Railway management differs from management of other technological or commercial activities. The present analysis focuses on the particularities of railway management in relation to the characteristics of the industry. Therefore, it unveils and discusses all the factors related to the internal and external environment of railways that are essential for an efficient management. The choice of a particular method of management is based on (1) the organizational structure of railways, performing in an oligopolistic transport market, (2) the decision to split the whole railway activity in homogeneous units, and (3) an eventual involvement of the private

sector. It is explained how the fundamental financial indices of a company (in terms of costs, yields, revenues-expenses, productivity) affect managerial decisions. Management is usually correlated with planning and particularly with Business Plans and Project Management. The various issues related to management of railway infrastructure, such as level of infrastructure charges, outsourcing, reduction of maintenance expenses, and attraction of funds for modernization are discussed. Management of rail passenger considers passengers' exigencies (through regular marketing and advertising campaigns) and must make an important decision whether to increase traffic or make a profit. Traditional pricing strategies can be combined with yield management techniques in order to achieve both a higher load factor and increased revenues. However, no managerial decision can ignore the various elasticities in relation to other transport modes. Management of freight should overcome weaknesses such as low average speed and difficulty of door-to-door transport and focus on products suitable for rail transport. Rail freight should be incorporated as a component of the logistics chain and adopt flexible commercial and pricing strategies, similar to those adopted by competitors of railways.

Defining Railway Management and Differentiating it From Management (in its General Sense)

A Definition of Management

Many definitions of management have been suggested over the years so as to respond to the demands of various sectors of the economy and to the needs that management is called to address. However, there are certain underlying characteristics that management encompasses: (1) recognizing and understanding a more or less human activity, (2) defining goals and objectives, (3) setting a plan of successive and interactive steps to be followed, (4) deciding and selecting the most appropriate financial and human resources, and (5) following, controlling, measuring as well as modifying procedures at each step of the activity.

It seems that there is no consensus over whether management constitutes a science or whether it is just an art. These different views stem from the fact that management requires not only a good understanding of a complex problem or system but also other essential qualities such as imagination, intuition, and forecasting abilities of future (and difficult to predict) events and situations, and in some cases even luck.

Managing Railways: Dealing With a Heavy and Rigid Industry

Railways made their appearance in the early 1800s as revolutionary technology at the time, promising to increase traveling speeds and reduce travel times. Before the era of railways, land transport of both people and goods was made possible by walking short and long distances and by horse power, with an average traveling speed at 5–10 km/h worldwide. Railways are based on the guided movement of the wheels of successive wagons on two rails through a metal-to-metal contact. This fact permitted railways to achieve speeds of 125–200 km/h during the 19th century, of 300 km/h in 1989, of 350 km/h in 2020.

The necessary energy for the movement of trains was provided during the 19th century by steam-powered engines and was replaced by diesel-powered engines during the 20th century. Energy can be provided alternatively by electric traction, which was introduced in the early 1900s and proved to be an effective solution from an economic point of view for railway lines which had heavy traffic load.

The railway system has one degree of freedom, in contrast to the road system which has two degrees of freedom; thus once the direction of the train movement changes, it becomes more complicated than in the case of road vehicles, as special installations, called switches and crossings, are needed. Furthermore, the braking distance, or stopping distance, is greater for trains than it is for road vehicles, and normally needs around 1300–1400 m for a speed of 200 km/h, 7500–9000 m for a speed of 320 km/h. In order for railways to offer safety of movement, they are operated by a complicated system of signaling, which nowadays is facilitated by application of the Global Positioning System (GPS) and Intelligent Systems.

As is the case with any technical system, railways require regular maintenance to run properly and accurately. However, if railways are to offer a non-stop service, the maintenance of railway infrastructure is supposed to be conducted either during the intervals between successive trains or during night hours, under extremely difficult conditions.

Effective management of railways is based on an optimal synergy of all components of the railway system, track electrification (or other power providing technique)—signaling, a really heavy and rigid system. However, as for many decades, the operation of railways was under state control, it was ascertained that railway legislation and particularly labor legislation protected the rights of all working parties within the sector, rendering the whole situation rather inflexible and difficult to change.

As railways have been developed within the context of national interests and regulations, they present numerous differences regarding track gauge, electrification and signaling systems, loading gauge, axle load, etc. A remedy to tackle this problem is interoperability, that is, the ability of the rail system to allow safe and uninterrupted operation of trains and assure a specific level of performance. Interoperability refers to technical, commercial, and legal issues and should be among the first priorities of railway managers.

Railways: Engineering Oriented or Customer Oriented

Transport is not an end in itself, but only a means to satisfy other human needs such as work, leisure, provision of goods at the right place and time. Railways have been operating for decades under quasi-monopolistic conditions and put more emphasis on technological and engineering aspects and less on commercial ones. Thus, when competition was introduced in the transport

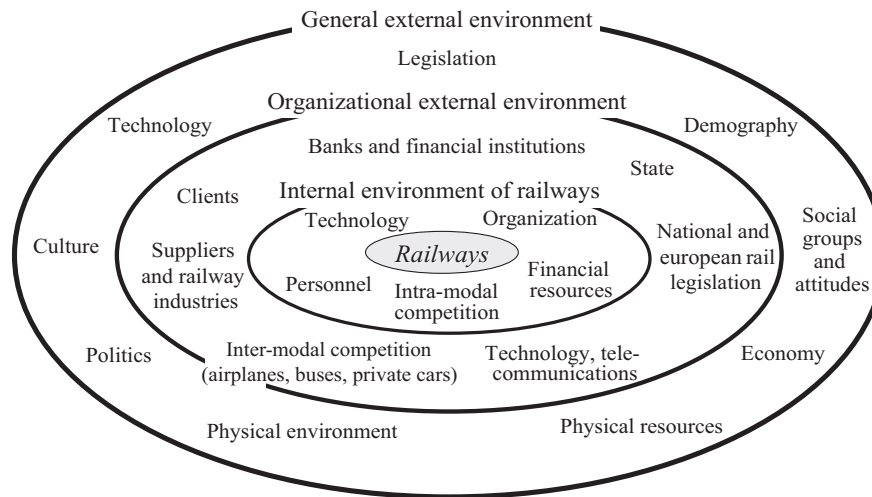


Figure 1 The internal and external environment of railways. *Source: Modified by Profillidis (2016).*

and recently within the rail sector, it proved difficult to convince railway staff that engineering was not their priority but rather the satisfaction of consumers' needs by offering lower tariffs, reduced travel times, increased quality of service, while claiming less or no funding from the state (in the form of subsidies), so as to survive in an extremely competitive environment.

Understanding a Constantly Changing World

The Internal and External Environment of Railways

A fundamental prerequisite for any form of management is to understand anything surrounding the activity under study (i.e., in our case, railways), that is all the entities that can affect directly or indirectly the specific activity. All these entities constitute the environment of railways. We often distinguish between the internal and the external environment of a company or an entity. The internal environment comprises all the components of a company (e.g., employees, climate within the company, management, etc.) as well as all interactions with its competitors, whereas the external environment refers to the outside factors or even conditions (e.g., societal changes, shift in regulatory issues, etc.) that could somehow have an impact on it. The external environment can be further subdivided into: the micro-external (or organizational external) one, which includes all the factors that impact directly the operation of a company and the macro- external (or general external) one, which includes all the factors that a company finds quite impossible to control.

Fig. 1 illustrates the various factors of the internal and external environment of railways. The success of any form of organization and management of railways depends on the ability to adapt to the changing situations of the internal and external environment.

A Systems Approach for Railways

Effective management of railways aims to provide the best synergy of all railway components described previously, on the one hand, and on the other hand to react to all factors that are related to the environment of railways. Such a synergy cannot be ensured, unless railways are considered as a system; for that reason, the management of any railway entity should be considered within a systems approach, which usually has the following successive steps: definition of the problem, targets to be met, evaluation criteria, alternative solutions, legislation and external constraints, environmental constraints, priorities and criteria of evaluation, assessment of alternative solutions and choice of the best one, application, problems during application, and finally evaluation of results.

Lifestyle, Environmental Sensitivity, Artificial Intelligence, and Management of Railways

Lifestyle as a synthesis of concepts, interests, opinions, and behavioral orientations is always a crucial factor that should be carefully studied before any managerial decision is made. Among those characteristics that permeate lifestyle in the 2020s are the following: increasing individualism, mimesis and expanding domination of the image over thinking and rationality.

Both the increase of environmental sensitivity and the urgent need and call for reduction of CO₂ emissions can prove powerful tools in the hands of railway managers, particularly in commercial and tariff policies, to attract new demand or stabilize existing demand.

Our way of thinking and acting is revolutionized by artificial intelligence (AI), which is the science of making machines or systems doing things that would otherwise require human intelligence. Some essential features of AI systems are that they are programmed to think like people, to act like people, and to think rationally and reasonably (to the extent that it is possible based on

the programming performed). Applications of AI in railway management can simplify procedures, facilitate decision-making aspects and replace routine staff (Engelbrecht, 2007).

Understanding the Transport and Rail Market

An Oligopolistic Transport Market

Railways coexist and compete in the transport market with other transport modes: private cars (for short distances), buses (for short and medium distances), airplanes (for distances longer than 500 km). The transport market has clearly the characteristics of oligopoly: a small number of large companies tend to dominate the market by controlling most of the transportation of people and goods and thus gaining the profits rendered from this activity. It is noteworthy that these companies are strongly interdependent in terms of their pricing and commercial policies. For example, if for any reason airlines serving the London-Paris route modify their airfare prices, the operating company of high-speed trains (i.e., Eurostar) has to modify inevitably its ticket pricing so as to be quite competitive. Interdependence in the transport market is a key characteristic for railway management and this form of competition is often called inter-modal competition.

The Railway Market: From Monopoly to Competition

With regard to the railway market, depending on the orientations of the economic policy of each state, we can witness many forms of organization which can range from monopoly to full competition, often referred to as intra-modal competition.

In a monopolistic railway market, such as the one operated in China, there is only one provider of railway services who dominates the market and sets tariffs. As in most cases, the only provider of railway services is owned by the state and thus the only form of control is the state.

In a fully competitive railway market, such as the one in the United States, there are many providers of railway services who are quite competitive in the market and offer tariffs and conditions of services without any kind of control exercised by the state. Regulation is at a low level and refers particularly to issues related to safety and financial solvency of railway companies operating in the market.

Within the range between monopoly and full competition, there are many forms of partial competition, such as in the European Union; full competition in some sectors, for example, international railway freight, monopoly in other sectors, for example, domestic railway passenger transport (also known as “cabotage” rights).

A Unified or Separated Railway

Before introducing any kind of competition within the railway market, railway managers should respond to a crucial question regarding the structure of railways: will they continue to keep the old structure of railways unchanged, by controlling both infrastructure and operation, or will they decide to separate the infrastructure from the operation of railways? In the early 1990s particularly in Europe, there was a strong belief that it would be difficult to introduce competition in the railway sector if the old organization of railways was kept unchanged, that is if the same company was responsible for the railway infrastructure and operation. In this line of thinking, the separation of railway infrastructure from operation was seen as a first (and inevitable) step towards introducing competition in the railway sector. In the separated form of railways, many rail operators can run on the same infrastructure by paying appropriate charges in relation to the quality of railway infrastructure (values of maximum speed, axle load, electrification or not, etc.), the time (slot) of the journey, the distance traveled, etc. However, there should not be any kind of discrimination in favor of existing state-owned operators and against new entrants.

This European conception of separating infrastructure from operation as a prerequisite for competition is not considered universally acceptable. Thus, in the United States, Japan, Canada, and elsewhere, one rail operator owns the infrastructure and operates services, while permitting other rail operators to run on that infrastructure by paying appropriate charges and competing with the owner of that operating company.

Separation of infrastructure from operation can take place in many shapes and forms, the most common being the following:

- Accounting separation: Both infrastructure and operation are part of the same company, but accounts of infrastructure are separated from those of operation. Accounting separation can be more detailed if the whole railway activity is separated in more business units such as infrastructure, electrification, signaling, operation of passengers, operation of freight, maintenance of rolling stock.
- Institutional separation: Infrastructure and operation are legally different companies without any form of legal interdependence.
- Within a holding company: Infrastructure and operation are different companies, but not totally independent, as they are part of a holding company with a common board and chairman.

The unified model of railways is in congruence with a more interventionist state policy, whereas the separated one is closer to state policies favoring competition. Separation or not is not a purpose in itself, but only a tool to achieve the most effective railway, offering a high level of services (such as reduced travel times, appealing prices, easiness, etc.) to clients, higher productivity and revenues to the company, increased levels of safety and security, lower state subsidies, satisfactory salaries for the personnel. The decision of railway managers to separate or not the unified railway should be based on such criteria and not on ideological or political prejudices.

Public or Private Railways

It is noteworthy that the ownership of a company (depending on the management policy) does not directly affect neither the market nor the consumer (or client). What matters is the existence of healthy competition or its absence. As the state usually fails to establish a real competitive market, it often presents partial privatization or full sale of railways as a condition for competition; this usually turns out to be a policy error or an illusion.

The privatization of railways often follows the path set by the privatization in other sectors such as telecommunications, airways, etc. This approach suggests that railways should be treated in the same way as profit making businesses by overlooking the complex character of the railway system and that most activities of railways are loss making. One would expect that a privatized railway would increase productivity and quality of services offered to the clients and attract private funding for the necessary modernization. However, it seems that common practice fails to meet the expectations set.

Failure to undertake the necessary changes, once privatization of a railway company has been realized, seems to have as origin the fact that railways have low margins of profitability and only in some sectors of freight and long-distance (particular high-speed ones) passenger transport. Introduction of modern management in a private railway can improve indices of finances and productivity but cannot overturn the situation from a heavily loss making activity to a profit making business. Indeed this is the case for the privatization of infrastructure, if safety continues to be a top priority upon privatization.

Throughout the world, successful and unsuccessful examples of privatizing railways have come to light. British railways tried to privatize their railway infrastructure in 1995 but were obliged to renationalize it in 2003. Passenger and freight activities of railways in the United Kingdom were allocated to a number of private companies, each one monopolizing specific routes. Government support per passenger-kilometer was (according to the British Office for Rail and Road) 0.04€–0.08€ between 1990 and 1995 and increased to 0.08€–0.17€ between 2005 and 2018 (all values of year 2018). Countries such as New Zealand and Estonia were obliged to renationalize their railways after an unsuccessful privatization. On the other hand, privatization of railways was more successful in the United States, Canada, Australia, and Japan, to mention a few.

To overcome any eventualities, railway managers should proceed with caution once considering privatization and focus on some categories of freight (bulk volumes, combined transport), some local passenger services (for which sufficient public service obligations are provided) and high-speed services.

Managing Finances of the Rail Activity

The Revenues/Expenses Ratio

Revenues of unified railways have as principal sources the tariffs paid by clients (passengers and/or freight) and state subsidies (provided usually only for passenger services). In the separated form of railways, revenues of the infrastructure manager are fees paid by operating companies for the use of infrastructure and state subsidies. A marginal part of revenues can come from fees of advertising companies. The expenses made by railways are principally related to paying personnel, energy and materials, and interests that concern the service of past loans.

The revenues-expenses ratio clarifies the ability of a rail company to survive (if it is higher than 1.0) or not. Successive years with revenues/expenses ratios lower than 1.0 will oblige managers either to borrow (if the market considers they possess some financial solvency) or be absorbed by another company (if some comparative advantages of the company in difficulty are essential for the purchasing company) or to declare bankruptcy.

A Good Understanding of Costs and Yields

Any managerial activity is an attempt to establish equilibrium between costs and revenues. Railways can reduce their costs by replacing labor intensive activities with computers and machines, by increasing productivity and outsourcing some railway activities (e.g., cleaning of trains).

Parallel to controlling costs, railway managers should try to increase yields, that is, unit revenues. This can be only achieved if railways have some kind of monopolistic or competitive activity; otherwise, any increase of yields risks shifting railway traffic to competitors, in relation to the values of price and cross elasticities. Railway yields were in the mid-2010s as follows:

- For passenger transport, in international US\$ per p-km: 0.18–0.25 for Western Europe, 0.10 for Eastern Europe, 0.3 for the United States, 0.15 for Japan.
- For freight transport, in international US\$ per t-km: 0.06 for Western Europe, 0.08 for Eastern Europe, 0.025 for the United States.

Increase of Productivity

Productivity reflects the rates of a production activity and is the outcome of a combination of several factors: labor, equipment, etc. We usually use as index of railway productivity the labor productivity, that is, the output volume of traffic to the total number of employees; this is expressed in traffic units of passenger-kilometers + ton-kilometers per employee. Indicatively in 2018, on a worldwide scale, average railway productivity was 2 million passenger-kilometers + ton-kilometers per employee.

High railway productivity is the result of a number of factors, and as such we could isolate high traveled distance, appropriate number of employees, use of modern equipment, effective management, strong competition in a specific market, and motivation of staff.

A Segmentation of Traffic

As railway traffic is extremely heterogeneous, it is practically impossible for managers to know which segments contribute profits and which ones make losses (and to what degree). For this reason, railway traffic is usually segmented as follows:

- Passenger traffic: Commuting (it serves the suburbs of a city and is strongly subsidized), regional (it serves regional centers), intercity (it serves major population centers), high-speed,
- Freight traffic: Small isolated parcels, bulk freight (minerals, grains, coal), general merchandise (products sold in supermarkets), express and urgent freight.

Financial indicators should be provided for each of the above segments of the market. Thus, railway managers can decide in which segments of the market have an interest to remain and which segments to discard.

Planning the Railway Activity

Planning As an Essential Prerequisite for Management

The only way to ensure consistency and compatibility between goals and results to a complex activity is to have a plan for this activity, that is, set goals, define actions to be performed, estimate and allocate resources, determine stages of application and deadlines, identify responsibilities of actions, define mechanisms of monitoring and evaluation. Planning is the management of change and should be characterized by consistency, flexibility, and adaptability.

Business Plans and Master Plans

Everyday management should be integrated in a general context of planning so as to avoid improvised and unprepared actions. To achieve this, essential tools for any railway manager are Business Plans and Master Plans. Whereas a Business Plan emphasizes more on organizational, economic, financial and investment aspects, a Master Plan refers to the whole of the railway activity and includes more details on technical equipment and projects. A Business Plan is a guide and a frame of action and not an implementation program and it contains analysis of the following (UIC, 2008):

- position and share of railways in the transport market,
- financial analysis, revenues and expenses, costs and tariffs, comparative analysis with other transport modes,
- weaknesses of the present organization and management,
- proposal of a new strategy with clear targets, quantification and forecast for the evolution of the fundamental indices of the railway activity, such as passenger and freight traffic, revenues, expenses, cash flow, productivity,
- new investment required, sources of financing, expected return of investment, volume of loans,
- sensitivity analysis for all forecasts, risk analysis.

Project Management: a Necessary Tool for Most Railway Projects

Railway projects are complex, include many heterogeneous components (e.g., signaling, rails, ballast, etc.), and take time to be materialized. In order to execute a railway project successfully, that is to follow up and respect the initial budget, partial and final deadlines and quality of work, effective management is necessary and is known as project management.

Project management consists in directing and administrating a complex project (by breaking it down in smaller components) and is in principle a science but to a certain degree an art. Project management can be executed by in-house staff or by outside consultants. It usually comprises four discrete stages: organization (requirements, accurate definition and description of the project, constraints, costs, programming of works, allocation of resources, funding), development (standards, studies, alternative strategies, more accurate programming of schedules), setting up (quality assurance and procurement plan, organization of site management, calculation of quantities of materials and man-hours of staff, invitation to tenders), execution (control of materials and quality of work, security measures and medical facilities, progress reports, measurement of works done and payments, test runs, recording of data of the project, certification to deliver the project for use).

Some cases of rail projects that require project management are: construction of a new high-speed line, track upgrading, electrification of a line, signaling and safety installations, a new tunnel or bridge, a marshalling yard, a railway station.

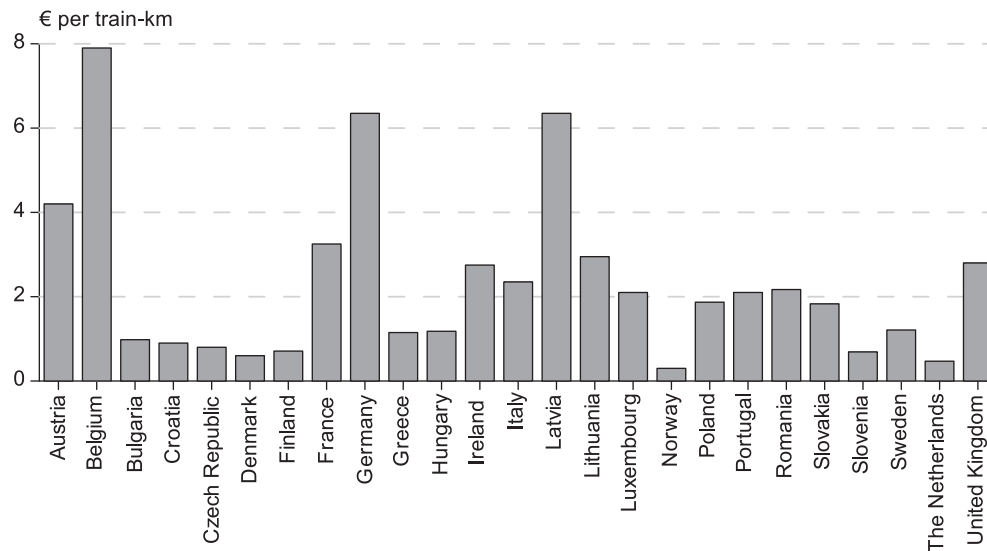


Figure 2 Values of infrastructure charges for the operation of passenger trains in European countries in 2016. Source: *European Commission (2019)*.

Management of Railway Infrastructure

Deciding Efficient Values of Infrastructure Charges

Management of infrastructure aims at ensuring safe operation of passenger and freight trains at the scheduled speed, with a high quality of transport, at the lowest possible cost. Costs of infrastructure refer to costs of maintenance and operation of track, electrification equipment and signaling equipment, and also to management of rail traffic. In the monolithic organization of railways (when a company owns both infrastructure and operation) expenses of infrastructure are covered from charges paid by other operators running on the track, eventual public subsidies and a part of revenues of the company from the operation of passenger and freight trains (cross subsidies). In the separated organization, expenses of infrastructure are covered from charges paid by operators and eventual public subsidies.

Thus, the level of infrastructure charges is a critical decision of infrastructure management: low charges favor operators' finances but burden the infrastructure's finances, whereas high charges favor the infrastructure's finances but burden the finances of operators. The following figures illustrate certain values of infrastructure charges for the operation of passenger and freight trains in European countries. Average values of charges are 1–3 €/train-km for passenger trains and 2–4 €/train-km for freight trains (Figs. 2 and 3).

Level of Outsourcing

In order to balance revenues and costs, railway managers are obliged to reduce their costs. As 40%–60% of infrastructure charges are personnel salaries, it is often preferable to outsource a number of infrastructure activities (principally maintenance works), usually through a bidding procedure.

Condition-Based Maintenance

In the past, infrastructure managers had to opt for maintenance strategies of the track and the rolling stock among the following alternatives:

- planned maintenance: It follows predefined scheduled intervals with no consideration about the best time for maintenance,
- preventive maintenance: Based on regular monitoring of each component of the track and the rolling stock but also on previous experience, a maintenance schedule is defined. It does not ensure that the specific railway component under maintenance or replacement has attained its ultimate strength or operational possibilities,
- corrective maintenance: When failure of a railway component is eminent, maintenance works are undertaken. It is expensive, not optimal and a quite risky method.

Contrary to the previously mentioned strategies, condition-based maintenance aims at conducting the maintenance at the moment that is needed, that is, neither earlier nor later than the appropriate time. It is based on a continuous monitoring of the status of a railway component thanks to monitoring devices appropriately located and an immediate transmission of data to decision making centers. Thus, condition-based maintenance can be facilitated with the help of the internet of things and artificial

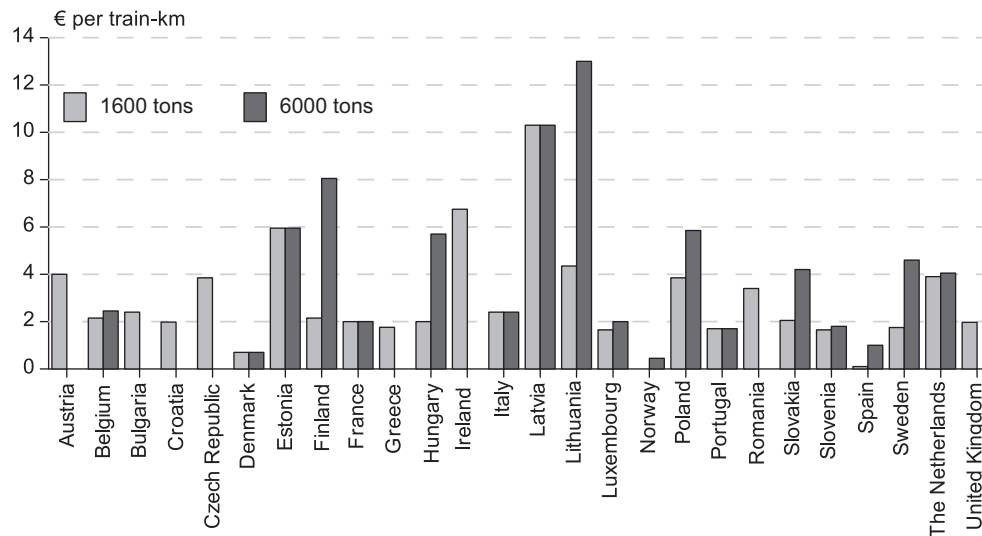


Figure 3 Values of infrastructure charges for the operation of freight trains in European countries in 2016. Source: *European Commission (2019)*.

intelligence: cloud data with the help of mobile mapping systems permit to transpose field records in accurate figures and design of the accurate position of the component under study, compare with the situation considered as critical and optimize the preventive maintenance plan.

Management of Passenger Traffic

Competition From Other Modes

Rail passenger traffic faces competition from other modes as follows (*Profillidis and Botzoris, 2018*):

- *Urban and commuting* traffic. It serves the suburbs of a city. Principal competitors of railways are buses and private cars. Comparative advantages of railways are reliability (as they are not affected by road traffic congestion), better travel times, lower values of ticket prices, easiness and possibility to exploit productively travel time. However, in some cases low quality of transport has been noted among the passengers' complaints,
- *Regional* traffic. It serves regional centers. As with commuting traffic, principal competitors of railways are buses and private cars. Due to a number of stops that regional trains make, private cars may provide lower travel times, in addition to an uninterrupted door-to-door travel,
- *Intercity* traffic. It serves major population and economic centers distanced some hundreds of kilometers among them. Due to high travel times, there is less competition in this case from buses and private cars, but railways may face competition from airplanes. Quality of rail services is in this case more critical than in the previous ones. A specific case of intercity traffic is *high-speed* traffic, offering rail speeds higher than 200 km/h and up to 350 km/h.

Monitoring of Passengers' Exigencies

The only way for railway managers to have an accurate estimate of their customers' characteristics, preferences, degree of satisfaction or annoyance, desired changes etc. is to conduct regularly (e.g., all 6 months) market surveys, which are usually realized with the completion of questionnaires either by the passengers themselves or by the staff of the company (in some cases by outside interviewers). Beside the determination of the profile and characteristics of passengers, market surveys can monitor assessment of passengers' choices for the services already rendered to them (revealed preference surveys) or record intentions of passengers regarding changes in rail services to be offered to them (stated preference surveys).

A survey is an experimentation of the trends, attitudes or opinions of a population and is achieved by studying a very small portion (called the sample) of that population. The sample reflects just part of the truth of the population and should represent it as accurately as possible. In most rail transport surveys, a confidence level of 95% for a margin of error of 5% can be considered as satisfactory. In order to quantify attitudes, views or desires of the respondents of a survey, some form of scale can be used (like the 5-point or 7-point Likert scale) and thus the relative degree of satisfaction of respondents can be approached. Questions of a survey refer to the following: purpose of traveling, degree of satisfaction, assessment of services and facilities in the train and the station, reasons that deter passengers from choosing traveling by train, ease of preparing the trip, transport mode to access the railway station, and frequency of traveling by train.

Railway managers should carefully examine respondents' intentions about future attitudes related to new railway services or changes in existing services. Not all positive positions should be taken for granted. It has been found that in many cases respondents were very enthusiastic about the idea of a new railway service; nevertheless when this was realized, they changed their mind and only a percentage of them realized their initial intention to use the new service. In a number of cases, market surveys are followed by advertising campaigns, as part of promoting new services or products to the public in general or to specific segments of the market.

Increase of Traffic or Profit

The old dilemma of all railway managers is whether they should aim at the increase of traffic or profit. Strategies aiming at the increase of traffic favor more socioeconomic factors and less finances of the railway company. Such strategies lead very often to deficits, which are covered by state subsidies. Strategies aiming at the increase of profit lead to higher unit tariffs and inevitably to the abandonment of a number of secondary lines in order to balance revenues and expenses.

Optimization of Finances

As rail operating costs are a function of distance, rail tariffs have been also calculated for many decades as a function of the distance traveled. Such an approach ignores competition from other transport modes (and in some cases from other rail operators). Indeed, the choice of a customer for a specific transport mode is based on the generalized cost of transport, which is the sum of the tariff paid by the customer and the monetary value of its travel time. Thus, the key for higher rail tariffs is to achieve competitive travel times and quality of service, in general.

For most itineraries, railways still apply a flat tariff, that is, the same tariff no matter the time of the day or the time of procurement of the ticket. A characteristic of all transport modes is that if the offered capacity (i.e., number of available seats) is not used, it is lost. In order to make the maximum profit from the capacity offered, railways introduce for some of their routes yield management techniques, which suggest a differentiation of the tariff paid as follows:

- in relation to the period of the day, the day of the week and the season of the year, by offering lower tariffs for off-peak hours,
- in relation to the time of procurement of the ticket: the earlier the ticket is purchased, the higher the offered tariff reduction.

Advantages for Frequent Travelers and Targeted Segments of the Market

Many companies tend to reward loyal clients. Railways should also introduce similar measures. For itineraries with a low load factor, instead of having vacant seats, railways could offer these seats at zero cost or at the marginal cost for passengers using frequently the railway services, at very reduced cost for the seniors or young, at advantageous costs for families and groups.

Effects of Elasticities

No managerial decision should be made without a thorough consideration of the values of the various elasticities, such as price elasticity (with a mean value for railways of around -0.6), income elasticity (with a mean value for railways of around 0.8), and cross elasticities of railways in relation to other transport modes (e.g., cross elasticity of rail transport with regard to the cost of use of private car has in Western Europe values around 0.20 – 0.25).

Considering Options for Door-To-Door Transport

A principal handicap of rail transport is that it does not provide door-to-door accessibility. A railway trip is just a component of a more complex procedure to assure door-to-door mobility and requires the services of bus and metro systems or of private cars and taxis or of bicycles. Some railway companies include in the price of a rail ticket the possibility to use certain free of charge services such as bus, metro, shuttle buses or bicycles at the beginning and the end of a rail trip. Thus, railways can integrate rail transport within the concept of seamless transport.

Management of Freight Traffic

Focus on Products Suitable for Railways

In the past, railways undertook any kind of freight transport, from small isolated parcels and general merchandise (for which railways are totally unsuitable compared to road transport) to bulk products and specialized freight (such as containers and chemicals) for which railways have comparative advantages, particularly regarding costs of transport. This cannot continue in a market which is becoming more and more competitive. Railway managers have strong interest in abandoning loss-making freight activities and focus essentially on bulk and specialized freight.

Increase of Average Freight Speed

Average speed of consignment of freight trains in Europe hardly exceeds 20 km/h and is far lower than an average speed of road trucks at the order of 50 km/h. This difference in average speeds is due to several reasons: delays in marshalling yards, long controls when crossing borders, a lower priority in traffic regulation for freight trains compared to passenger trains. A remedy to overcome this inherent handicap of railway freight is the creation of tracks dedicated only for freight traffic on routes with heavy freight flows. Such tracks are often called railway freight corridors and their essential technical characteristics should have a maximum speed of 120 km/h, a length of platforms of 750 m, axle loads of at least 22.5 tons, ERTMS signaling system of level 3, horizontal loading in marshaling yards. With the creation of rail freight corridors, rail freight trains can increase average speeds from the actual level (in Europe) of 20 km/h to 50 km/h or even more and ensure a punctuality of 95% (with tolerable delays at the order of 1 h).

Incorporation of Rail Freight in the Logistics Chain

Rail freight is just a component of freight transport that could be used for the transport of goods from their origin to the expected destination and should be integrated within the frame of the logistics chain, which includes in addition to the simple freight rail transport some of the following: loading, unloading, delivery, distribution, storage, flexible organization with immediate response to any kind of freight demand. Regain of rail freight competitiveness can be obtained, in addition to improvements in track and equipment, with a better organization of successive steps and a use of smart techniques in order to ensure just-in-time delivery.

The Issue of Truck Subsidiaries of Rail Freight Companies

The central inherent handicap of railways is that they cannot provide door-to-door transport, unless railway tracks reach directly the heart of industrial or production centers. But this is feasible from an economic point of view only in the case of huge industries. Thus, connection of the origin or destination of rail freight to the railway station is achieved with the use of road trucks. Railway managers should decide whether to possess a truck subsidiary (but many similar attempts like the truck subsidiary Sernam of French Railways turned to a financial disaster) or to outsource truck activities. The other alternative is to entrust road transport of freight from origin or destination to the railway station to freight forwarders, but there is the risk that rail freight managers will lose control of the market.

Summary and Conclusion

For decades, railways have been managed within a very regulatory, interventionist and state policy spirit, ignoring the situation in the market, expectancies and exigencies of clients and the amounts of money spent for them. The present analysis explains how railways can be run as a modern, open, friendly to the client, less subsidized, more commercially oriented activity. The methods presented are not uniform and have the necessary flexibility so as to adapt to the contrasting differences of the various societies and economies around the world. Competition and contestability are strong incentives for railways to modernize, together with a gradual reduction of political power upon them. Inspired from other sectors of the economy, railways can adopt some aggressive commercial and pricing strategies and revalorize their human resources.

References

- Engelbrecht, A.P., 2007. *Computational Intelligence: An Introduction*, second ed. John Wiley and Sons.
- European Commission, 2019. 6th Report on Monitoring Development of the Rail Market. Retrieved from: https://ec.europa.eu/transport/sites/transport/files/6th_rmms_report.pdf
- Profillidis, V., Botzoris, G., 2018. *Modeling of Transport Demand—Analyzing, Calculating, and Forecasting Transport Demand*, 2018, Elsevier.
- Profillidis, V., 2016. *Railway Management and Engineering*, fourth ed. Routledge.
- UIC, 2008. *EURAIL 2025: Vision Stratégique des Chemin de Fer Européens ver 2025*.

Further Reading

- Crevier, B., Cordeau, J.F., Savard, G., 2012. Integrated operations planning and revenue management for rail freight transportation. *Transp. Res. B* 46 (1), 100–119.
- International Transport Forum, 2011. Meeting Society's Transport Needs under Tight Budgets, Discussion Paper No. 2011-10. Available from: <https://www.itf-oecd.org/sites/default/files/docs/dp201110.pdf>.
- Nash, C., Nilsson, J.E., Link, H., 2013. Comparing three models for introduction of competition into railways. *J. Transp. Econ. Policy* 47 (2), 191–206.
- Thompson, L., 2010. A Vision for Railways in 2050, International Transport Forum, Forum Paper No. 2010-4. Available from: <https://www.itf-oecd.org/sites/default/files/docs/10fp04.pdf>.
- Thompson, L.S., Bente, H., 2014. What is rail efficiency and how can it be changed?, International Transport Forum, Discussion Paper 2014-23. Available from: <https://www.itf-oecd.org/sites/default/files/docs/dp201423.pdf>.

Railway Terminal Regulation

Nacima Baron, Université Paris Est, Laboratoire Ville Mobilité Transport - ENPC, France

© 2021 Elsevier Ltd. All rights reserved.

Terminal Regulation Definition and Scope	454
Railway Terminal Regulation Goals	454
Historical Background of Terminal Regulation	455
Railway Diversification Reframes Terminal Regulation	455
Railway Market Opening in Europe and its Consequences on Terminal Regulation	456
Railway Terminal Pricing Principles in Europe	457
Terminal Management Charging Models Across European Countries	457
Charging Patterns for Railway Terminal Use Across European Countries	460
Hot Topics and Future Trends in Terminal Regulation	460
Summary and Conclusions	462
Biography	462
References	462
Further Reading	463

Terminal Regulation Definition and Scope

Railway terminal regulation covers following areas:

- Physical management of terminal station (construction, maintenance, security, etc.);
- Definition terminal station access and exploitation, service pricing;
- Definition of legal status of owners and operators (state, collectivity, public, or private bodies);
- Competition rules between operators;
- Public service obligations (quantity, quality, and price of services delivered);
- Conflict prevention and solving;
- Price and access conditions to essential facilities and resources necessary for railway terminal activity (energy, land);
- Network interconnexion (technical compatibility, flow capacity).

In this broad perspective, terminal regulation is connected with railway terminal governance (who decides what), with railway terminal financing models (who pays for what), with railway terminal merchandising (how to locate railway terminal and how to manage them so as to attract passengers and clients and extract value). Railway terminal regulation also include technical regulation such as network regulation as well as security and safety norms appliance [A link to 10473: Rail network regulation can be done here], and is part of public transport regulation laws (international, national, metropolitan levels) when railway terminal is connected to intermodal nodes (buses, metro). When railway networks cross international borders, a bi- or multilateral regulation system can regulate the management of an international station and the right and duties of governments and railway companies.

This chapter is organized as follows: section “Railway Terminal Regulation Goals” introduces railway terminal regulation goals and section “Historical Background of Terminal Regulation” provides the historical background of terminal regulation. The section “Railway Diversification Reframes Terminal Regulation” explores the way terminal functional diversification goes along with a multiplication of stakeholders and a reframing of regulation conditions. Then section “Railway Market Opening in Europe and its Consequences on Terminal Regulation” details the specific but variegated conditions of terminal opening to market in European countries and section “Railway Terminal Pricing Principles in Europe” the zoning and pricing principles attached to terminal space and use. Sections “Terminal Management Charging Models Across European Countries” and “Charging Patterns for Railway Terminal Use Across European Countries” develop the terminal management charging models and their different patterns (the way terminal services are aggregated into blocks of activity across European countries), before reviewing hot topics and trends (section “Hot Topics and Future Trends in Terminal Regulation”).

Railway Terminal Regulation Goals

Following cases, railway terminal can be either railway stations related to passenger travel, either cargo terminals interesting freight or depots, interesting both. All these terminals are part of network and provide services which exploitation may lead to scarcity of resources and overexploitation. For example, traffic caused by a busy railway crossing creates lack of capacity in a freight bottleneck. Hence railway terminals are essential facilities. Cargo terminals, marshalling yards, and train depots provide the infrastructure, building, and machines for flexible and efficient railway activity and maintenance (for the notion of efficiency, see Cantos and

[Maudos, 2001](#)). Having access to a network of strategically located terminals with good connection to seaports, and offering modern rolling stock and good maintenance conditions help to reduce the time that vehicles are out of service and avoid empty trains. It secures and makes easier transshipment between rail and road or rail and sea or waterways and contribute to better flow management, which in turn leads to a reliable timetable.

Passenger railway stations are also essential facilities for transit and interurban transport. If railway activity is opened to market, stations turn into an essential facility. Railway undertakings need to have open information concerning the technical and financial conditions to access these stations, they must know how they can put a staff basis and what let services will be proposed to their final clients, the railway passengers. If railway activity is not opened to market, train flow regulation and passenger movement management are also very important to optimize the quality of railway service (train punctuality, comfort) [A link to the article 10014: Value of crowding in this Encyclopedia can be done here]. This needs also a sound regulation of the contractual relationships between station manager teams and train operator departments. Flow regulation is therefore necessary to assure fluidity in getting on and off a train, as well as to organize multimodal transport services from station to end destination. Terminal regulation is also needed in order to clarify the policy making conditions offered to nonrailway stakeholders (national governments, metropolitan governments, estate agents, data firms, etc.) in station development projects and to integrate side business activities such as shops, hotels, and other activities in stations ([Dirhaug, 2013](#)).

Historical Background of Terminal Regulation

Railway terminal regulation was necessary till the time railway lines were constructed. Even if it has changed a great deal since the 19th century, when train boomed in Europe and in the Americas, the fundamental question is still the capacity of envisioning a railway terminal as an essential facility. The genealogy of this notion is to be found at the time when various companies were obliged to find agreements in order to share infrastructural nodes, platforms, depots, carbon storage, and water for steam machines. In 1911, a conflict aroused between competing American railroad companies (respectively the Wiggins Ferry Company, the Eads Railroad bridge terminal company, and the Merchant's bridge terminal company) that were converging at Saint Louis station, one of the largest railroad centers in the world at this time. Organizing the access of such competing railroads was not a basic thing. The solution proposed was that these three independent terminal systems merged into a single system owned by 14 independent railway companies, thereby completely controlling all interconnection facilities between both sides of the river Mississippi. According to the Supreme Court, mergers of terminals into one single system avoided unnecessary duplication of facilities. However, since not all railroad companies were owners of the terminal, the Sherman Act of 1912 did require nondiscriminatory access to the merged railroad terminals. Mandatory third-party access to complementary monopolistic infrastructure led to the birth of the crucial notion of essential facility ([Dobbin, 2004](#)) ([A link with article 10006: Natural monopoly in transport can be done here]).

Railway Diversification Reframes Terminal Regulation

Most railway companies were nationalized before or after world war two in Europe, Asia, and in a majority of decolonizing countries, so terminal regulation changed a lot. Railway terminals turned to be driven by monopolistic companies and access to infrastructure was not so a problem, but large-scale railway programs (including high speed lines) were developed, along with railway modernization (e. g., line electrification). State or regional governments investments were needed for such great works. Rebuilding attractive stations and efficient terminals with public money was a part of this intent to foster rail development, seen as sustainable transport. So, terminal regulation was at this time included in the broader development of freight and passenger railway national regulation systems. Some terminals were specialized to serve port development or inland industrial plants, while urban terminals (including old depots and technical infrastructure) could be extended and redesigned into urban multifunctional programs (sometimes on an important scale). Railway terminal development schemes concerned progressively a broader array of stakeholders, including railway or metro companies, state administrations, public and private banks, municipal and metropolitan bodies, and industries. All these entities were eager to know precisely the conditions for intervening on railway land or the possibilities to buy pieces of land in terminals, and wanted to know the types of contracts (concessions, leasing, etc.), which could be signed with railway authorities for different kind of projects. Terminal regulation had to adapt and integrated new specifications. Depending on the countries, side businesses activities took place in terminal regulation schemes or were integrated in other type of regulation (land regulation, urban planning, etc.).

Japanese railway companies vertical separation and functional diversification occurred throughout the 1990s, after 1987 railway business act (which integrated «the construction of facilities around a station to promote connectivity with other transportation modes», quoted by [Fumio 2018](#), p. 395) and then after 2005 Act on Enhancement of Convenience of Urban Railways. Terminals were located in the eastern metropolitan corridor and their urban environments were exposed to the steady rise of land values, hence the railway terminal premises, tracks, technical installations, and former passenger stations were converted into dense, high rise, intermodal, and multifunctional urban centers. Many Japanese companies engaged in a wide range of side businesses, such as land development, real estate development, department stores, and mobility services. Other project interested railway corridor at the peripheries of big cities [A link to 10454: Railway station (and network) planning and management article of the Encyclopedia can be done here]. Modest train station and old tracks could be turned into new transit-oriented development schemes integrating the reconstruction of a commuting railway line, a renovated commuting station and adjacent housing programs. Regulated benefits

coming from train operation and unregulated side businesses deriving from other activities were the two dimensions of a new terminal regulation model. Each railway company was required to keep separate accounts for railway and other businesses and railway regulation was imposed only upon the railway side. Yet, according to economic cycles and to the long-term trajectory of megaprojects (in city centers) and miniprojects (in peripheral areas), railway activity, or side business activities could be more profitable, and the benefits taken from one activity could sustain the sustainability and long-term vision of the whole group.

Railway Market Opening in Europe and its Consequences on Terminal Regulation

"In theory, mixed-used developments around centrally located rail stations offer a perfect answer to the challenges of a future-oriented, post-peak oil sustainable development agenda focused on transit-accessible urban cores. In practice, however, the implementation of such megaprojects is highly complex, and costs and benefits are unevenly distributed." As Peters quote in 2012 (p. 163), Japanese terminal station megaprojects have been copied and globalized (Peters, 2009). The opening up to railway competition and the introduction of private capital in railway projects is of actuality worldwide, even in countries where a monopolistic company still dominates the national railway market. Technological companies, media and entertainment firms, real estate businesses, and hedge funds are newcomers in station (re)development projects or in railway infrastructure building or renovation. An example comes from China. Since 2014, the Chinese government launched a China railway development fund, scheduled to last for 15–20 years, with the aim of promoting private capital investments into rail projects through public–private partnerships (PPPs), in order to alleviate the debt carried by public authorities. In 2016, the Zhejiang government signed the first of such PPP agreements (34 years contract, the first four years of which for the construction, followed by 30 years of management) with the conglomerate Shanghai Fosun High Technology Group Co. Ltd. The project included land development around and above the eight new stations of the Hangzhou to Taizhou 269 km high speed line, which generated revenues could be used toward paying back railway development.

Russia and United States as well as in China and in European Union, which published, till the 1990s, various railway law packages encouraging railway market opening, historical railway companies rethink the place of terminals in their business model, and make proposition for adapting terminal regulation. Gómez-Ibáñez and De Rus (2006), and then Laurino et al. (2015) state that worldwide surveys are not numerous, the information very difficult to gather, and, moreover, both delineate with difficulty terminal economic regulation from overall railway regulation. Yet, a turning point seems to be attained. All these companies and incumbent railway undertakings understand that they have to add to the traditional tasks of terminal manager new expertises and skills, with new rules. Fig. 1 explains the value-oriented station regulation model, which associates traditional technical missions (passenger welcoming and train operation) with news targets such as passenger flow commercial fertilization and asset development. The basic

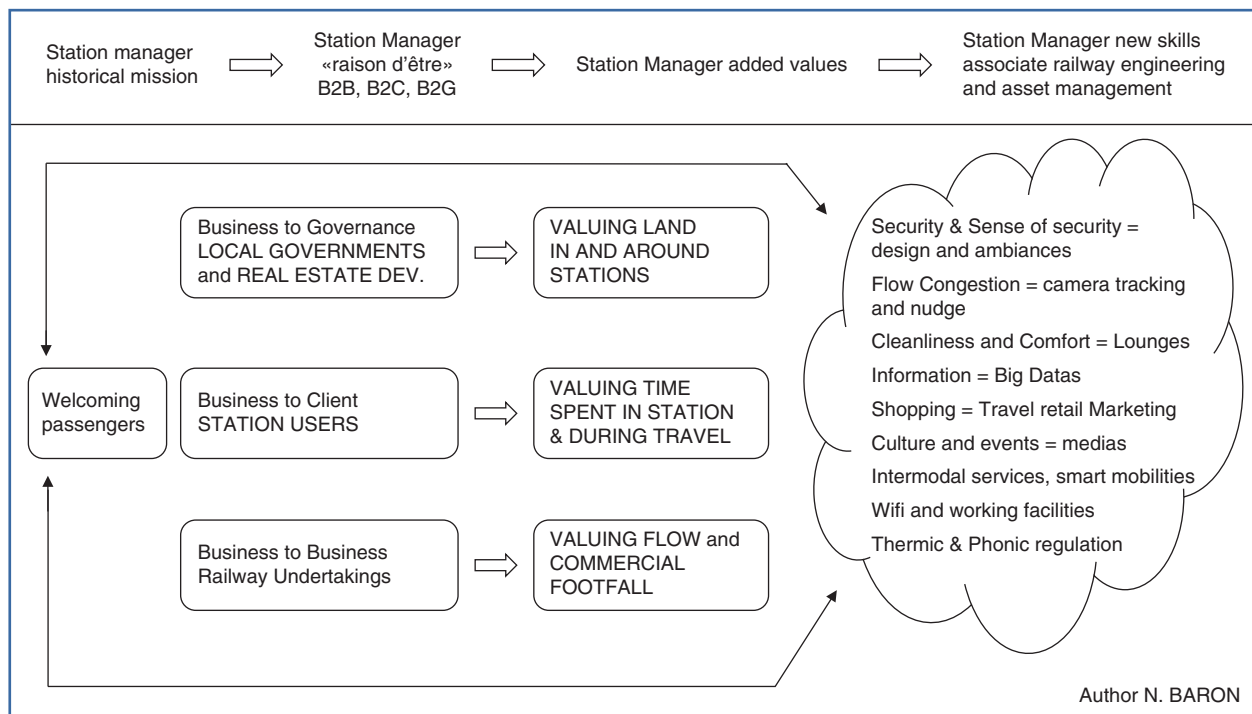


Figure 1 Value-oriented station regulation associate traditional technical missions (passenger welcoming and train operation) with news targets: flow fertilization and asset development.

role of station management bodies is still related to the maintenance of the terminal buildings, its efficient management, its development in collaboration with the urban public authority, and transport systems. But, with liberalization trend, new challenges appear when they consider their strategic role being related to the maximization of three kinds of values: land value and long-term money return to be assured in shared programs signed with public authorities (in a business to governance perspective); flow value and careful asset management guaranteed with estate, retail and digital firms (in a business to business perspective); and time value, as an element of good travel experience of railway passengers (in a client oriented approach) (Beria et al., 2012). Such a commercial repositioning includes the redefinition of terminal station owner and manager legal duties and professional capacities. Providing comfort and security, assuring cleanliness and multilingual signage, offering first and last mile innovative mobility services as well as business lounges and WIFI network is now a must in main passenger stations and it changes regulation frames, especially in European countries.

Railway Terminal Pricing Principles in Europe

Terminal regulation is affected by many ways in the general context of the opening up to competition in the European Union. The main goal of EU railroad regulation is to increase competition on the railway markets by improving the independence of stakeholders, increasing the power of regulatory bodies and guaranteeing not only regulated access to monopolistic bottleneck components, but also regulating access to rail-related services, such as maintenance facilities, terminals, stations for freight and passenger trains (Finger and Messulam, 2015).

The introduction of fair and nondiscriminating access to passenger stations, maintenance depots, and cargo terminals is presented from the 1990s onward as a necessity in order to save railway industry, which is threatened by car transport for passenger as for freight competition. Vertical separation, autonomy, and neutrality are keywords. The autonomy and responsibilities of a station manager entity must be clearly delineated in the transformation or historical companies into vertically separated entities (infrastructure manager and railway undertaking). The station manager entity has to present service neutrality. The conditions of access and use have to be published and legally guaranteed by the independent Authority in order to avoid barriers to entry for newcomers in the market. Fig. 2 shows the terminal regulation system in European Union and how public stakeholders' interactions are reorganized.

The station manager must present price neutrality toward former monopolistic railway companies and incumbent undertakings for stations, depots, maintenance units, and freight terminals access and use. Terminal zoning and terminal access charges must be defined. This means that the stations and terminals are subject to a pricing process, and a value is assigned to the use of each element of a terminal [A link with 10016: Pricing Principles in the Transport Sector of this Encyclopedia can be done here]. The price of each elements and the final price is constructed from five elements:

- The terminal position on the rail network;
- The singular nature of the location and the scarcity and value of space (urban or industrial);
- The services available at the terminal (especially intermodal services, transshipment facilities);
- The environment and amenities around the terminal;
- The possibility of capturing large flows of passengers with varying abilities to pay (in the context of a passenger station).

The pricing of a terminal is based on regulatory engineering, but also generates discussion and dialogue, with differences of opinion between the players regarding the value they ascribe to the constraints that affect the station, so the independence of regulation authority is compulsory to stabilize the criteria and make consensus on the creation of price. The notion of "reasonable profit" is included in the main charging principles laid down in Directive 2012/34/EU in this quotation: "The charge imposed for track access within service facilities referred to in point 2 of Annex II, and the supply of services in such facilities, shall not exceed the cost of providing it, plus a reasonable profit." (Article 31.7.) (...) "Where services listed in points 3 and 4 of Annex II, as additional and ancillary services, are offered by only one supplier, the charge imposed for such a service shall not exceed the cost of providing it, plus a reasonable profit." (Article 31.8.). Reasonable profit means a rate of return on own capital that takes account of the risk taken by the terminal manager in the production of the facility or service, including that to revenue, or the absence of such risk, incurred by the operator of the service facility. Such return and is in line with the average rate for the sector concerned in recent years. An external rail regulator verifies the specifications for station access, defines the implementation of the safety measures, flow conditions, intermodal services and retail development possibilities. If several operators use a terminal, the rail regulator validates the access charging system and the station pricing schemes.

Terminal Management Charging Models Across European Countries

The idea of the terminal as a natural monopoly, separated from nonnatural monopoly components, is supposed to be applied in the European countries in a similar way, with the outcome being similar. In reality, the national interpretations vary substantially and these differences are deepened by the different starting point of the national rail companies (importance of historical background and national context). Two dimensions must be taken into account.

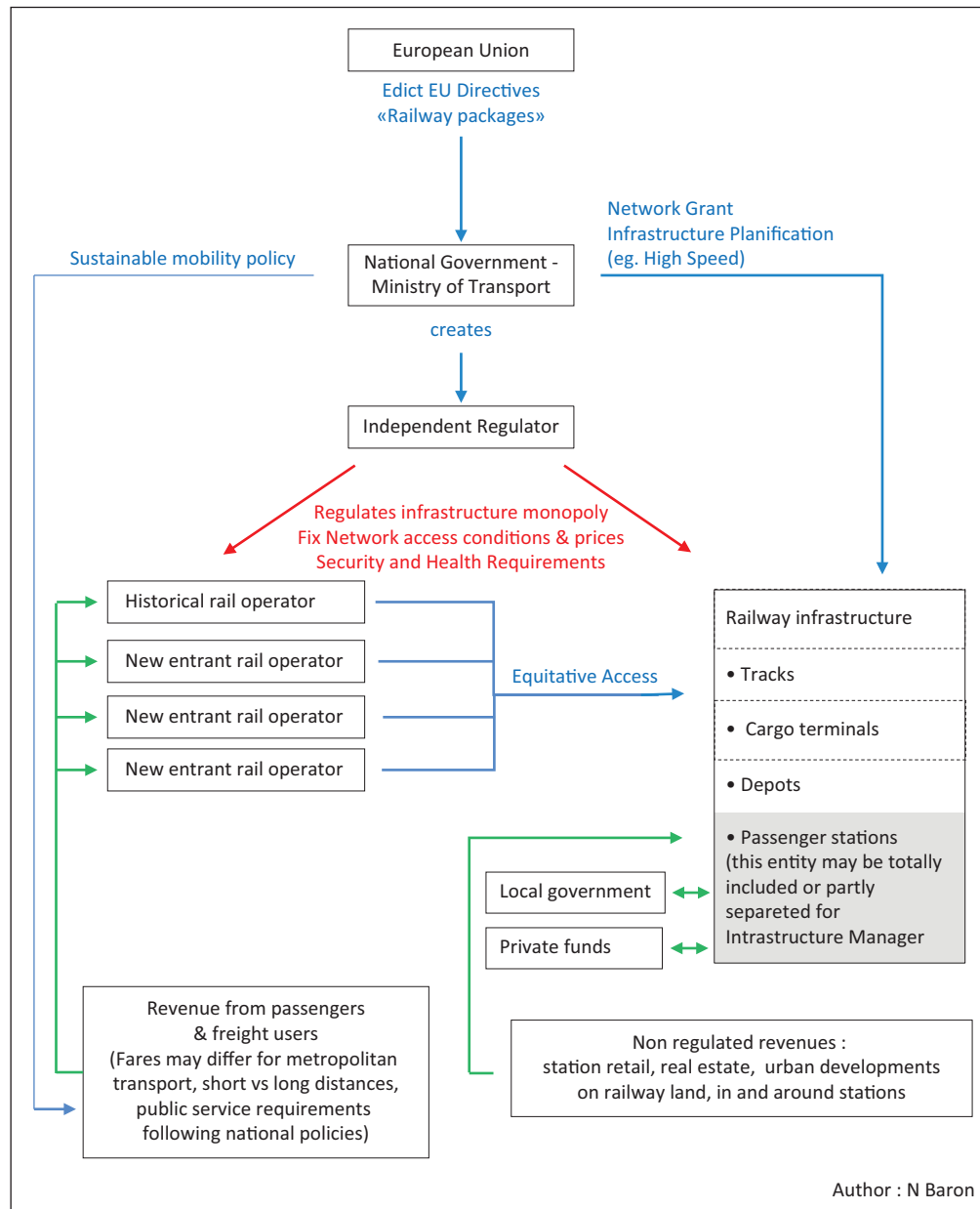


Figure 2 Terminal regulation system in European Union stakeholders interactions.

First, the way of separating railway premises ownership varies. In some countries, the ownership of depots and maintenance yards is of the incumbent, in other countries in the possession of the infrastructure manager. Thus, in Spain such facilities are not open to competitors and, in Italy, access of these facilities to competitors is subject to the incumbent's approval.

Second, the legal nature and the position of station manager entity in each national post market opening railway ecosystem differ a lot from one country to another. **Table 1** offers an overview of European countries terminal regulation approach: the very heterogeneous importance of railway terminal infrastructure in number of stations and traffic activity, the existence or inexistence of a neutral entity in charge of railway terminal management, the types of charges and revenues (regulated and unregulated) that this entity manages (Teixeira et al., 2013). In Sweden, autonomy is at its maximum (Alexandersson and Rigas, 2013). Terminal ownership and management is carried out solely, by an independent company named Jernhusen, which is fully public owned (state and local governments are involved). The main task of Jernhusen is to manage land, technical and commercial assets as well as offices and housing buildings in the vicinity of terminals and stations. All other European countries include terminal management entities either as an element of the infrastructure manager, or, more rarely, in a department, a branch or a subsidiary with closer contact to former monopolistic train operator. In Germany, Deutsche Bahn Station&Service AG is a legally independent holding company created during the reorganization of Deutsche Bahn. This is equivalent in Spain, where stations are conceived and operated

Table 1 Overview of European countries terminal regulation approach : railway terminal infrastructure and activity, existence of a neutral entity in charge of railway terminal management, types of charges and revenues.

	Number of stations	Number of train stops (in Millions)	Annual number of passengers (in Millions)	Entity responsible for the management of railway station	Number of railway undertakings using passenger stations	Station segmentation	Number of application charges (main service package) and modulations	Publication or negotiation and regularity of tariff renovation	Unregulated revenues (eg retail)
Austria	1438	20	278	Infrastructure Manager	Limited number	yes	66 tariffs	Yearly published	Yes
Belgium	550		225	Incumbent railway undertaking	Limited number	no	1	Yearly published	Yes
Bulgaria	299		24,6	Infrastructure Manager	Single railway undertaking				
Croatia	507		18	Infrastructure Manager	Single railway undertaking	yes	differs for each station	Yearly published	
Finland	192		68,3		Single railway undertaking			Negotiated	
France	3029	40	1600	Incumbent railway undertaking (holding)	Limited number (Thello & Thalys)	yes	173 tariffs x 9 modulations	Yearly published	Yes
Germany	6547	152	2700	Incumbent railway undertaking	Multiple railway undertakings (opened market)	yes	196 tariffs x2 modulations	Yearly published	Yes
Greece	376		7,4	Infrastructure Manager	Limited number	yes	included in the charges for the use of infrastructure	Yearly published	
Hungary	1367	13,5	146	Infrastructure Manager	Limited number	yes	4 station segments x 2 modulations	Yearly published	
Italy	2260		855	Infrastructure Manager for most Stations + Grandi Stazioni	Limited number	no	track access included in the station charges for minimum access package		Yes
Luxembourg	68	129	21,5	Infrastructure Manager	Single railway undertaking	no	1 tariff + Extra fees for long trains	Yearly published	
Norway	336			Infrastructure Manager	Limited number	no	2 (one is zero, 1 for the city -airport line)	Yearly published	Yes
Poland	2730	21	269	Infrastructure Manager	Multiple railway undertakings (opened market)	yes	15 tariffs	Yearly published	
Slovenia	273		15,5	Infrastructure manager	Single railway undertaking	yes	4 tariffs		
Spain	1939		470	Infrastructure manager	Limited number	yes	3 tariffs x 4 modulations		Yes
Sweden	136		205	Specific entity Jernhusen	Multiple railway undertakings (opened market)	yes	6 tariffs x modulations		Separated account
United Kingdom	2537	89	1640	Infrastructure Manager in most cases + specific entities related to HS	Multiple railway undertakings (opened market)	yes	differs for each station	Partly published and partly negotiated every 5 years	Yes

Source : Teixeira P., Prodan A., Lopez Pita A., 2013 *Railway stations and auxiliary charges in 27 European countries, final report*, International Railway union UIC and IRG-Rail, 2015, *Overview on charging principles for passenger stations in Europe, technical report*, vol 15, n° 8, 23 p.

by infrastructure entity ADIF; as well as in Austria where ÖBB-Immobilienmanagement GmbH is 100% owned by ÖBB-Infrastruktur AG; and in France, where passenger terminal management and development Gares&Connexions, which was initially created as an autonomous holding within the historic company SNCF, is recently reintegrated to infrastructure manager SNCF Réseau.

The situation is on the opposite in Netherlands, where NS Stations is a specific body integrated in Dutch public owned train company NS; as it is in Switzerland, where stations are managed as part of CFF passenger department; as in Belgium, where SNCB has the control of stations buildings, platforms, car parks and retail. Italy and Great Britain presents specific cases, in the way that a horizontal division is created, that is to say that different kinds of stations obey to different regulation systems. In Italy ([Arrigo and Di Foggia 2013](#)), railway infrastructure manager RFI operates the great majority of stations, save the 14 biggest, under the responsibility of Grandi Stazioni, a body more oriented toward station asset management and opened to corporate groups and private funding. In Great Britain, only the main larger stations are managed and operated by the mainline Infrastructure manager, Network Rail, and the rest of stations are being a part of a franchise and managed by the franchised railway undertaking. High Speed stations are regulated apart. In the case of High Speed 1 line, Saint Pancras station and its five-star hotel were purchased and developed as a property asset in the context of the high-speed regulation system.

Charging Patterns for Railway Terminal Use Across European Countries

In most of European countries, the charging models obey to a common structure and are based on the principle of full cost. The methods used to set the charges are quite similar since the charging system aims at covering all the costs of the passenger station managers. But the situation is also very different from one country to another ([IRG-Rail, 2015](#)). To illustrate this, let us consider that the document Irish rail access charging system is a text of 12 pages where terminal access conditions take no more than three pages, whereas French “Document de référence des gares de voyageurs” has 66 pages. In few cases, such as Ireland, there is no specific charge for passenger stations since such charge is included in the charge for the use of infrastructure. In the other cases, the structure of charging systems is the following. In the majority of cases, a station manager develops a station segmentation using station characteristics and variables (passenger frequentation, traffic intensity, platforms number) and service characteristics. Moreover, according to the European norm, station charges are included into three possible categories: minimum access package fees, additional services, and ancillary services. The first group includes some criteria defining the infrastructure and some other elements describing the services provided in the station by station manager or by other providers. Among them, the use of common spaces and equipment (furniture, lifts, escalators, toilets); the reception, information, and orientation system (audio and display); the reception of persons with reduced mobility; other services such as heating, surveillance, lost & found objects. The second group includes train parking facilities in adjacent depots, energy (preheating and precooling of trains) and water sources and maintenance facilities. Auxiliary or ancillary services refer to other services, such as access to information network, specific conditions for the announcement of trains, maintenance and inspection of vehicles, train dispatching services. Each criteria presence and weigh in the final pricing method is different from one country to another, and external charges (that is to say charges exterior to the three groups) are applied for the access and use of terminal premises (e.g., offices and desks, spaces for ticket sales, railway undertaking personnel) is applied in Austria, Bulgaria, France, Italy, and Luxembourg.

Table 2 shows railway station and terminal charging structure and most used criteria for pricing methods. Station tariff systems differ greatly. Some countries have tariff systems for line charges, but not for stations. Other countries have relatively simple systems (that is to say that stations are categorized by only few variables). However, some countries have built highly complex station pricing systems, whose structure depends on numerous variables. According to a 2015 UIC Study, 16 on the 27 European countries have some type of stations charges or auxiliary charges, 11 have only station charges, nine countries defines their station charges only by categorizing the stations by importance. Concerning the importance of station prices on the global price of access to network, the situation differs greatly between countries, types of network (for example, access to high speed station are much more expensive than access to conventional station) and between different pair of cities (origin–destination). France, Belgium, and Spain have the highest price levels (regularly between 8% and 15%, sometimes 20% of total access price to network) and also the greatest difference between prices for small or medium sized station and prices for bigger stations. Concerning routes, the maximum stations charges observed account for 40% of total infrastructure charges, but remains as an exception.

Hot Topics and Future Trends in Terminal Regulation

The implementation of neutral access conditions and transparent charging systems remains a problem in many countries of Europe. Some of the convergence opportunities expected by Crozet in station charging did not occur ([Crozet 2004](#)). [Tomeš and Jandová \(2018\)](#) have evaluated the international lines between Austria, Czech Republic, and Slovakia. They state that new entrants have caused major increases in train frequencies and cut prices. Yet, given the diverse pathways and levels of liberalization of rail activity in each country, significant problems remain. For example, when Austrian newcomer WESTbahn tried to access infrastructure and other essential facilities, it had difficulties in obtaining attractive timetable slots and its passengers could not easily use ticket selling facilities in terminals. «Further complaints included discriminatory path allocations and the exclusion of WESTbahn services from the timetable published by the incumbent» (p. 76). The construction of fees for using station and terminal facilities remains also very heterogeneous and continues to evolve as a consequence of the European Directives transposition slow rhythm in national laws. But

Table 2 Railway station and terminal charging structure and most used criterias in pricing methods

		Main Variables used in the pricing method
Station segmentation	Station characteristics	- Station category and size - Train traffic - Passenger traffic - Platform number and size
	Service characteristics	- Service type, - Stop type, - Trip length - Stopping type - Peak vs off peak pricing
Minimum access package	Station physical infrastructure descriptors	- Signage - Information area - Seating - Loudspeaker - Platform displays, - Escalator, lift - Security cameras - Garbage bins, - Trolley base - Ticket machine + ticket gate - Weather canopy - Bike parking
	Minimum Proposed Services	- Cleaning service - Security service - Info point - Announcement - PRM service
Additional Services		- Train parking facilities, - Water, wastewater disposal - Preheating /cooling facilities - Maintenance facility
Auxiliary / Ancillary Services		- access to telecommunication networks - provision of supplementary information (eg intermodal information) - technical inspection of rolling stock - Specific ticketing services (eg smart card, intermodal pass ...) - Heavy maintenance

Source : EU Directives 2001 and 2012, IGRail, UIC

even if it is an uncertain and a long-term process, some trends can be observed. First, the “Europeanization” of stations is a strange process in the fact that adapting to the principles of competition leads simultaneously to the standardization of a general regulation pattern and to a specific situation in each country: that is one of the odd consequences of a long term movement of “europeanization of infrastructure” (Schipper 2011). Second, Beria et al. (2012) consider that terminal regulation (and in a broader approach of railway regulation) in biggest European countries (Germany, France, Spain, Italy) shows a balance between liberalization trends and a protectionist attitude from incumbent railways against liberalization, “with the states backing this behavior” (p. 110). The hypothesis is the same in Finger and Messulam, who consider that “the complexity of terminal access charges are not a market tool for pricing the use of rail infrastructures, but rather should be seen as public policy tools” (Finger and Messulam, 2015, p. 12). Third, if station management companies still play a dominant role as railway terminal infrastructure technical operators, but some of them consider to be as well service providers in vertically integrated structures and experts in travel retail and urban real estate. In this objective, some terminal management bodies progressively distinguish subdepartments (some with highly technological approach of flow management, some with a more strategic position of asset managers). These entities use all the possibilities of railway regulation to multiply and diversify the contractual relationships with other businesses, gradually evolving from a framework based on concessions (e.g., restaurant or hotel concessions) to another based on licenses, concessions, contracts, PPPs. Such terminal owners and/or managers consider their mission to be based not only on infrastructure providing, but on infrastructure plus service providing. Each national regulation system can limit or support this orientation. In some countries, unregulated

revenues can either be completely excluded from the regulated charging system (when there is one, as in some European countries), or fully or partially taken into account.

An optimistic vision would state that, in Europe, the terminal manager has the capacity to sell services, to earn the revenues and to reinvest part of benefits in infrastructure projects, or in the train operating budgets, thus decreasing the price of train ticket and, as a final consequence, fostering train competitiveness. A pessimistic vision would consider that terminal managers specialization in service activities make them prefer managing big and rentable stations to small ones, thus threatening the integrity of network. Small stations and terminals may be less rentable but their openness may be considered as a public service obligation, and state or local governments have, in the European regulation system, few tools to support such policy. Station managers that are focused in priority on the production of economic results and are completely dissociated from other railway activities can also be invited to turn station into retail places and threaten overall railway sustainability, performance, and attractiveness in the long term. Another unresolved issue concerns the way station manager will regulate themselves and/or follow overall regulation concerning big data production, openness, treatment, and commodification. These are crucial questions, giving the importance of good and passenger flows in railway terminals and the possible use of such information for a great quantity of service providers. Anyway, such transformations deeply change the governance and the nature of interactions between terminal entities, railway authorities and other stakeholders. The provision of terminal-bound services has shifted from being primarily a railway intraorganizational to an interorganizational task building on the idea of negotiation across a variety of economic sectors, among different levels of government, and between public and private actors, owners and contractors, consultants, and the end-users (consumers and passengers) themselves.

Summary and Conclusions

This chapter has presented an overview of terminal railway regulation principles and implementation. It has described the constitution of terminal infrastructure. It has defined the concept of essential facility since its birth in 19th century US railway race to its regulatory implementation throughout the world, but especially in Japan, Europe and other Asian countries. The results and principal conclusions of the chapter are the following ones.

1. Terminal regulation transformations are caused by from two mains factors: (1) the transformative relationships between State, rail infrastructures, transport companies, and other stakeholders; and (2) the growing diversification of terminal functions, especially service, retail, land development, etc.
2. European institutions has framed specific regulation conditions and access conditions for terminals in the whole Union, but regulatory and pricing schemes remain different in the 27 countries analyzed.
3. Consequently, there is slow and heterogeneous pathways to free and indiscriminate access to terminals in Europe and only a few countries take the opportunity of their progressive autonomy from railway incumbents to develop side businesses in the context of regulated or in nonregulated railway models.

Biography



Nacima Baron holds a PhD in urban planning from Sorbonne University. She is full professor at Paris School of Urbanism at University Paris Est and also teaches transportation planning and urban projects at TU Delft and Rabat University. She holds a chair of railway station development and collaborates to various research projects on station retail, station and railway redevelopment schemes, and smart mobilities.

References

- Alexandersson, G., Rigas, K., 2013. Rail liberalisation in Sweden. Policy development in a European context. *Res. Transport. Bus. Manage.* 6, 88–98.
- Arrigo, U., Di Foggia, G., 2013. Competition and pricing of essential inputs: the case of access charges for the use of the Italian rail infrastructure. *UTMS J. Econ.* 4 (3), 295–307.
- Beria, P., Quinet, E., de Rus, G., Schulz, C., 2012. A comparison of rail liberalization levels across four European countries. *Res. Transport. Econ.* 36, 110–120.
- Cantos, P., Maudos, J., 2001. Regulation and efficiency, the case of European railways. *Transport. Res. A* 35, 459–472.
- Crozet, Y., 2004. European railway infrastructure: towards a convergence of infrastructure charging? *Int. J. Transport Manage.* 1, 5–15.
- Dirhaug, H., 2013. *Railway Policy-making: On Track?* Palgrave Macmillan 239.

- Dobbin, F., 2004. How institutions create ideas: notions of public and private efficiency from early French and American railroading. *L'année de la régulation* 8, 15–50.
- Finger, M., Messulam, P., 2015. *Rail Economics, Policy and Regulation in Europe*. Edward Elgar 345.
- Fumio, K., 2018. A study of vertical separation in Japanese passenger railways. *Case Stud. Transport Policy* 6(3), 391–399.
- Gómez-Ibáñez, J., De Rus, G., 2006. *Competition in the Railway Industry: An International Comparative Analysis*. Cheltenham. Edward Elgar.
- IRG-Rail, 2015. Overview on charging principles for passenger stations in Europe. Technical report, 15, 8, 23 p.
- Laurino, A., Ramella, F., Beria, P., 2015. The economic regulation of railway networks: a worldwide survey. *Transport. Res. A: Policy Pract.* 77, 202–212.
- Peters, D., 2009. The renaissance of inner-city rail station areas: a key element in contemporary urban restructuring dynamics. *Crit. Plann.* 16, 163–185.
- Schipper, F., 2011. Infrastructural Europeanism, or the project of building Europe on infrastructures: an introduction. *History Technol.* 27 (3), 245–264.
- Teixeira, P., Prodan, A., Lopes Pita, A., 2013. *Railway Station and Auxiliary Charges in Europe*. UIC, Paris, 30 p.
- Tomeš, Z., Jandová, M., 2018. Open access passenger rail services in Central Europe. *Res. Transport. Econ.* 72, 74–81.

Further Reading

- ARAFER Autorité de régulation des activités ferroviaires et routières, 2017. Etude thématique sur la gestion des gares ferroviaires de voyageurs en France, 32 p.
- Gares&Connexions SNCF Document de référence des gares de voyageurs pour l'horaire de service 2018 2019 2020, 66 p.
- Iarnród Éireann/Irish Rail, (s.d.) Access Charging System & Performance Regime, 12 p.

Service Network Design for Freight Railroads

Teodor Gabriel Crainic^{*,†}, ^{*}Analytics, Operations and Information Technology Department, School of Management, Université du Québec à Montréal, Montréal, QC, Canada; [†]Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation (CIRRELT), Montréal, QC, Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction	464
Rail Freight Transportation Structure and Planning	465
Rail Freight Transportation System	465
Freight Railroad Tactical Planning	465
Service Network Design for Railroad Planning	466
Basic SND and Service-Selection Planning	467
Static SND for Integrated Planning	468
Time-Dependent SND for Integrated Planning	468
Using Service Network Design Models	469
Conclusions	469
Acknowledgments	469
See Also	470
References	470

Introduction

The literature on railroads and railroad planning goes many years back. Several survey papers synthesize the story and contributions of operations research to railroad planning, for example, Assad (1980b), Dejax and Crainic (1987), Crainic (1988), Cordeau et al. (1998), Crainic (2000, 2003, 2009), Newman et al. (2002), Ahuja et al. (2005), Crainic and Kim (2007), Bektaş and Crainic (2008), Yaghini and Akhavan (2012), Piu and Speranza (2014), and Chouman and Crainic (2020). This chapter follows the last contribution. See Crainic and Hewitt (2020) and Crainic et al. (2020) for service network design and applications to consolidation-based transportation.

Railroads move large quantities of products, for example, bulk materials (e.g., grains, minerals, petroleum, and chemical products), general packaged freight, containers, trailers, automobiles, and machinery. Railroads are particularly efficient for long-haul movements in terms of per ton-km monetary, energy-consumption, and pollutant-emission costs. The transcontinental North-American land bridge between the West and East coasts, and the Eurasian land bridge between China and Western Europe bear witness to this efficiency. Freight rail transport is thus an essential link of intermodal transportation and supply chains, supporting national and international trade.

To achieve this efficiency, railroads operate mostly as consolidation-based carriers, similar to less-than-truckload trucking and liner shipping, for example. Consolidation offers the means to take advantage of economies of scale and reduced handling in terminals, by grouping freight from different shippers, with possibly different origins and destinations, and loading them into the same vehicles for efficient long-haul transportation. Railroad consolidation policy generally involves two levels, however, as cars are grouped into blocks, which are then grouped into trains. Loaded and empty cars, with different origins and destinations but being present simultaneously in the same terminal, are thus sorted and grouped into a block, which is then moved as a single unit by a series of trains until its destination, where it is broken down, the cars being either delivered to their final consignees or sorted for inclusion into new blocks.

The performance and profitability of the carrier then depends, for a large part, on efficient and coordinated terminal and long-haul transport operations, supported by an efficient management of resources, providing an offer of service meeting the cost/profit objectives of the railroad and the cost and quality criteria of its customers. Tactical planning aims to address this challenge through a network-wide transportation plan stating the train services to operate, the routing of the customer demand by those services, and the associated yard and line activities. The problem is complex and difficult to address, involving many decisions linked in a web of economic, resource utilization, and time-performance objectives, limitations, and trade-offs.

Service network design (SND) is the methodology of choice to build tactical plans for consolidation-based railroads, making the most efficient use of resources to achieve economic and customer-service performance objectives. This chapter provides a high-level overview of freight railroad planning and the service network design models and methods proposed to address it. Space is limited. Consequently, fundamental modeling concepts are discussed only, with pointers for further reading and more in-depth study. Chouman and Crainic (2020) provide detailed models for various cases, together with solution-method discussions and a rich literature review.

The paper is structured as follows: section “Rail Freight Transportation Structure and Planning” presents brief descriptions of rail freight transport and associated tactical planning issues; section “Service Network Design for Railroad Planning” is dedicated to the main SND problem settings and models: basic concepts and problem structure (section “Basic SND and Service-Selection Planning”), followed by integrated-planning models for the static (section “Static SND for Integrated Planning”) and time-dependent (section “Time-dependent SND for integrated planning”) cases; section “Using Service Network Design Models” discusses the utilization of SND models for railroad planning and the chapter concludes in section “Conclusion.”

Rail Freight Transportation Structure and Planning

We present brief descriptions of rail freight transportation and tactical planning.

Rail Freight Transportation System

Railroads operate on an infrastructure made up of a large number of terminals and rail tracks linking them. Most terminals are stations where demand originates and terminates, while a few are the yards specially equipped to handle large quantities of cars, sorting and grouping them for long-haul transportation, as well as to make up and disassemble blocks and trains. The backbone component of this network is made up of main lines connecting the yards of the system. The network is completed by secondary lines used to move loaded and empty cars between stations and their designated yards. Tactical planning targets operations on this main physical network.

Customer demand concerns a number of cars (containers to be loaded on special cars in the case of intermodal transportation), of a type appropriate to the commodity to be moved, to be shipped from an origin station, yard at the tactical level of planning, to a destination one. Empty cars, which need to be repositioned for the next cycle of operations, are also part of the demand. Each origin–destination (OD) demand is characterized by a volume, in terms of number of cars or containers, an availability time at the origin, and a due time at destination. For tactical planning, demand is an estimation of the “regular” traffic the railroad expects to satisfy over the planning horizon.

Freight is moved by train services, each composed of one or more locomotives providing power and a series of cars. A *service* is made up in an origin yard with a final destination yard, where it delivers the cars currently hauled and liberates the locomotives. En route, it may stop at a number of intermediary yards to deliver or pick up cars, eventually grouped into blocks as described below. The service is also characterized by time-related information, which may take the form of a frequency of service, that is, the number of times the “same” train is run during the schedule length (the length of time the railroad uses to define its recurring operations, for example, one week), or a schedule indicating the departure time from the origin yard, arrival and departure times at each intermediary yard, and the arrival time at destination.

Railroads aim to maximize revenue, which often translates into achieving the best balance between the cost of operating resources and services, on the one hand, and the quality of the service measured according to customer expectations in terms of timely and reliable delivery, on the other hand. Railroads therefore operate direct trains for important customers only or when the demand volume between two stations is significant (e.g., the equivalent of at least a full train). Most of the time, railroads rather aim for economies of scale through consolidation of cars from different demands into blocks and of blocks into trains (blocking is less performed when trains are short, e.g., in Europe). Schematically, cars at their origin yard are sorted, classified is the term generally used, to be then grouped into particular blocks with other cars, with potentially different origins and destinations. The block is then handled as a single unit from its origin yard, where it is formed, to its destination yard, where it is taken apart, its cars at their final destination being delivered to the respective consignees, the others being reclassified and blocked with other cars present at the yard for the next part of their trips. Services are thus made up of blocks and, when appropriate, it is blocks that are picked up and delivered at intermediate stops. Blocks may thus be transferred from one service to another.

Notice that intermodal traffic (containers and trailers) is often handled separately from the general one, both in yards and on the tracks, being moved on dedicated services. The most important difference, however, concerns the loading and unloading of intermodal traffic, which is performed in particular zones of the railroad’s yards. Numerous types of containers and cars exist, yielding a large number of loading alternatives subject to many rules and limitations, particularly when double-stack loading is permitted as in, for example, North America (Mantovani et al., 2017). Accounting for this complex issue at tactical planning level is particularly challenging.

The double (triple for intermodal) consolidation organization of freight railroads provides the sought-after economies of scale in operation costs and resource utilization, reduces car handling activities at yards, and fosters a timely service for pairs of cities or regions with low traffic volumes. It also implies more complex operations in terminals, with potentially higher possibilities for delays and incidents, which translate into more complex decision-making problem settings. Service network design models and methods are proposed to address these challenges.

Freight Railroad Tactical Planning

The planning activities of rail operations may be broadly classified into three levels: (1) strategic, involving long-term decisions on, for example, system design and acquisition of major resources (e.g., buy or rent locomotives or cars); (2) tactical, medium-term

planning process dedicated to building for the next “season,” an efficient service and resource-utilization network and schedule, specifying the train services to operate, the blocks that will make up each train, the blocks to be built in each terminal and the sequence of trains moving them, and the routing of the cars, empty and loaded, using these services, blocks, and terminal operations; (3) operational, concerning the short-term planning, monitoring, and adjustment of operations.

We focus on tactical planning in this chapter, as it yields the plans and policies of which depends the efficiency of the railroad. The service network design models discussed may also be used at the other levels of planning (Chouman and Crainic, 2020).

The tactical plan is built for a rather short schedule length (e.g., a week), to be repeatedly applied over the season (e.g., six months). Planning occurs some time before the season starts. Thus, even though some part of demand, for example, long-term contracts with customers, may be known at planning time, most is forecast using history, customer-relation representatives’ knowledge, and customer input, among other data sources. The goal is to satisfy the forecast regular demand in the most efficient way possible with respect to costs (profits) and resource-utilization, while satisfying the service-quality levels set by the carrier to answer customer requirements.

A number of main decisions/issues make up the tactical planning process and are addressed through SND models and methods, namely:

- *Service selection* selects among a set of possible services, which ones to operate the next season to move demand efficiently, profitably, and on time. The set of alternatives may represent a complete yard-to-yard network with intermediate stops for all service types, or the last-season network enriched with new potential services to address changes in demand, economic environment, railroad policies, etc. The resulting service network specifies the movements through space, and time for scheduled railroads, of trains and cars.
- *Blocking and classification* focus on how cars are grouped in yards yielding the blocks to be moved by trains. A *block* is characterized by its origin and destination yards, sequence of yards on the infrastructure network where it is transferred from one service to another, the sequence of train services performing the transportation. Blocking selects the blocks to build at each yard. Classification defines how cars are sorted and assigned to blocks at yards. For intermodal services, the problem also includes the assignment and consolidation of containers to multiplatform cars.
- *Train makeup* yields the list of blocks, and cars, each train service hauls out of its origin yard, drops and picks up at intermediary stops, and delivers at its destination yard.
- *Freight routing* determines the *itinerary*, or itineraries when splitting of demand is allowed, to transport each demand from its origin to its destination through the selected service network.

Resources, for example, locomotives and cars, are required to operate services. Resource management assigns the appropriate resources to services, and determines the economically, maintenance, and operationally efficient resource movements over the schedule length. Resource management is generally considered an operational-level managerial activity (Cordeau et al., 1998; Dejax and Crainic, 1987; Piu and Speranza, 2014) and its impact on tactical-level decisions was often limited to the somewhat simple case of accounting for the need to reposition empty cars for the next cycle of operations. The situation is evolving, however, and resource-management concerns are increasingly part of SND models for tactical planning (Andersen et al., 2009; Chouman and Crainic, 2020; Kienzie et al., 2020; Pedersen et al., 2009).

These problems may be, and have often been, addressed individually reflecting a planning process decomposed into a sequence of decisions. The strong and complex interconnections among these decisions increasingly lead, however, to integrated approaches addressing several problems jointly. Such approaches do not make problems easier, however, as planning must be performed network-wide aiming for the best trade off among the not necessarily convergent operational, service-quality, and economic characteristics of the individual problems and decisions. Thus, for example, one could try to increase the level of service by increasing the frequency of services, but this could result in higher levels of congestion in yards and on the tracks, resulting in increased costs and delays and, thus, possibly, a lower quality of service (increased costs as well, of course).

Service network design provides the methodology to model the rail freight transportation dynamics at a network-wide level, integrating the different operations, decisions/issues, objectives, and trade-offs, yielding an optimized operation plan for the season.

Service Network Design for Railroad Planning

Service network design (Crainic and Hewitt, 2020) models are built on networks representing the railroad’s structure, activities, and decision-to-be-made, and take the form of multicommodity, capacitated fixed-charge network design formulations (Crainic et al., 2020).

The problem and SND model are called *static* when one assumes that neither demand nor the other problem characteristics vary during the schedule length considered. The time dimension of the service network is then implicitly considered through the definition of services and the inter-service operations at terminals. The main decision in such cases is whether a service is selected to be included in the plan or not (its frequency, if selected, over the schedule length may be part of the decision). *Time-dependent* problem settings and formulations address the cases when the moments demand becomes available and is due at destinations are explicitly considered, which implies an explicit representation of demand and activities in time. Time-dependent formulations, called *scheduled service network design* (SSND), target the selection of services together with their respective schedules indicating arrival and departure times at yards.

Reflecting the traditional goals of railroads and expectations of customers to “get at there at the lowest possible cost,” the primary objective of these models is the minimization of the total operating cost. This cost is computed mainly as the sum of two terms, which may take different forms according to the problem and model considered (more on this shortly). The first represents the total cost of setting up the service network based on the fixed costs of the services. The second computes the cost of moving the freight given the service network, based on the commodity volumes moved by each service leg and the corresponding unit transportation costs.

Service-performance measures are increasingly included in tactical planning and SND models, reflecting rising customer expectations for high-quality service and environmental concerns. They are generally modeled through delays incurred by freight and resources or the amplitude of violation of predefined performance targets (e.g., delivery within a given time length). Constraints may be imposed on these service-performance measures, or one may add them as costs and penalties to the objective function of the SND formulation. The resulting generalized cost function then captures the trade-offs between operating costs and service quality.

All SND models addressing rail tactical-planning issues share the same structure. The basic problem setting and SND formulation are presented first, followed by more general static and time-dependent cases. See [Chouman and Crainic \(2020\)](#) for detailed presentations and literature review.

Basic SND and Service-Selection Planning

SND modeling starts with a network representation of the railroad physical infrastructure, where nodes represent yards, while arcs stand for the physical track paths between adjacent yards. Capacity measures are attached to nodes defining, for a given time period, the yard capacity to classify (number of cars that can be sorted and assigned to blocks), to block (number of blocks which can be built), to transfer blocks (number of blocks), and to make up train services (relative to the number of available tracks or resources). Capacity measures may be also defined for arcs reflecting the physical or operational limits of each track on, for example, the number of trains that may operate “simultaneously” (following, meeting, or overtaking), or the total weight or length of train on the track (the latter translates lower and upper limits on the block length and weight).

The demand side of the system is represented by a set of OD commodities, each commodity standing for the request to move a given volume of freight (number of loaded cars or containers, generally) from its origin yard to its destination yard. Demand is moved along itineraries, each specifying the sequence of blocks, services, and yard classification and transfer activities, between the demand origin and destination. These itineraries are defined on the network supplied by the train services the railroad runs on the physical network.

Services to include in the operation plan are selected from a set of potential services, each one being characterized by a path in the physical network between its origin and destination yards, as well as by a number of characteristics in terms of, for example, hauling power, speed, and priority. Single-leg services operate nonstop between these yards, while multileg services stop at one or several yards on the route to drop and pick up blocks and cars. Several cost and capacity measures may be associated to services depending on the particular problem addressed. In almost all cases, however, one finds the fixed cost to select the service (including, depending on the application, the costs for making up, dismantling, inspection, power, crews, etc.), the leg and commodity-specific unit transportation costs (including weight and handling-related costs), and the leg-specific capacity standing for the volume of demand the service may haul on each leg (determined by the physical and operational characteristics of the tracks).

Each SND model is then defined on a network based on the physical one and the set of potential services. We present the general SND-modeling framework in the context of the service selection and train makeup problem setting (e.g., [Assad, 1980a](#)). This problem arises when no blocking is performed (e.g., most European railroads), or blocking is performed once the main service network is decided. Given the set of possible services, the problem aims to (1) select the services to run and their frequencies over the planning period, and (2) assign cars to trains and determine the associated freight routing to accommodate all demand at minimum cost. As in all problem settings considered in this chapter, freight routing may involve single-service demand itineraries, and itineraries with service-to-service transfers at intermediary yards. The latter require time and resources, generating costs and possible delays. These concerns must be accounted for in the cost representation.

The nodes of the SND network are those of the physical one. The service legs of the potential services make up the arcs. Two main sets of decision variables are defined on this network. Binary *service design* variables indicate for each service if it is selected (variable will take value 1; when frequencies are included, this value becomes a non-negative integer) or not (value equals 0). Non-negative continuous *flow* variables are defined for each commodity on each service leg and stand for the demand volumes assigned and transported on each leg. Auxiliary binary decision variables model decisions to transfer commodity-specific flows at yards. Unit handling and transfer costs are given for each commodity and yard. The SND model yields an optimal operation plan, given by the values of the design and flow variables in the optimal solution, by minimizing the total generalized cost of the plan, subject to flow conservation and capacity/linking constraints.

The generalized-cost objective function is the sum of service selection and operation (total fixed cost of selected services), transporting the demand on the selected services (the sum over all commodities and service legs of the volume moved multiplied by the corresponding unit cost), plus the cost of car transfer at yards (a fixed cost associated to the transfer decisions or, a nonlinear penalty function of the decision to transfer and the quantity transferred).

The first set of constraints is typical of all network-flow optimization problems, enforcing the demand-satisfaction requirements, and the conservation of commodity flows at the nodes of the network. The linking constraints ensure that the total flow on any service leg cannot exceed the capacity of the service on that leg, provided the corresponding service is selected. Constraints identifying

the commodities and yards involved in transfers, as well as enforcing the integrality and non-negativity of the decision variables complete the model.

Static SND for Integrated Planning

The previous model may be easily adapted to address other tactical-planning issues, blocking in particular (e.g., Ahuja et al. 2007; Barnhart et al., 2000). We rather discuss an integrated SND static model (e.g., Crainic et al., 1984) aiming simultaneously to select the service and the block networks and, thus, define the classification strategy, as well as to determine the demand itineraries, establishing how freight is to be routed through the service and block network.

The integrated SND model is built on the same network of potential services defined in the previous subsection, and the yard and service definitions and characteristics are the same as well. The car classification and blocking set of issues and decisions need to be added.

Recall that cars, with possibly different origins and destinations, are classified and grouped into blocks, which are moved by trains and are handled as a single unit from their origins to their destinations. Recall also that cars at their origin yard may follow itineraries where they are classified and assigned to a block which brings them to their destination yard, from where they are to be delivered to their consignees. Alternative itineraries may also be used where the initial block brings cars to an intermediate yard, where they are reclassified into new blocks, and continue their trip toward the final destination or another intermediate yard and reclassification. More than one reclassification may make up the itinerary. The yard characteristics are enriched to account for these activities by defining for each yard, the unit car classification cost and the maximum numbers of cars which may be classified, blocks one may build or transfer, and trains one may make up.

Similarly to the service definition, one defines the set of potential blocks linking the yards of the network, each block being defined by (1) the sequence of yards where it is formed (origin), dismantled (destination), and transferred from one service to a different one; and (2) the sequence of service legs transporting the block from its origin to its destination, through the transfer yards, when relevant. A fixed cost, representing the building, transferring, and dismantling costs, and a capacity (the same units used for services) further characterize blocks. Commodity-specific unit transportation costs are associated to each block (based on corresponding leg-specific costs).

Sets of design and flow decision variables are again defined. To the integer service design variables, one adds the binary block design variables, which indicate for each block if it is selected or not. Non-negative continuous flow variables are now defined for each commodity on each block (instead as service legs as previously). The objective is still to minimize the generalized cost of the operation plan. The total block selection and operation cost (accumulation of the fixed costs of selected blocks) is added to the total service selection cost. The total cost of transporting the demand on the selected blocks (and services) and the total car classification cost at yards complete the objective function. The flow conservation constraints are part of the model, as well as linking constraints that state that a block cannot be used unless all the services moving it are selected. Linking/capacity constraints play a similar role for car assignment to blocks and services (cannot assign to nonselected blocks and trains). Constraints enforcing the yard capacity measures, and enforcing the integrality and non-negativity of the decision variables complete the model.

Chouman and Crainic (2020) detail this model and its path-based version, as well as more advanced formulations addressing congestion-induced delays, nonadditive itinerary attributes (e.g., tariffs), and penalties for violation of dues dates or other service-quality measures.

Time-Dependent SND for Integrated Planning

Scheduled service network design, SSND, targets time-dependent problem settings, where time and service schedules are explicitly considered. SSND formulations are built on *time-space* networks, using a discrete (more rarely, continuous) representation of the schedule length. A discrete representation of the schedule length is obtained by defining a sequence of time instances, from the start (time 0) to the end of the schedule, a time period in this representation corresponding to the length of time between two consecutive time instances.

The set of nodes of the time-space network includes copies of the yards in the physical network at all the time periods defined (this definition may be made specific to each yard). Each demand is defined on the time-space network and, thus, further characterized by an availability time at origin and a due time at destination. Each potential service in the SSND context is similarly defined on the time-space network. The sets of yards and service legs identifying it now being defined by nodes with a physical identify and time-related attributes, namely, arrival and departure times at the yards where the service originates, stops, and terminates. These define the service *schedule*.

The SSND network is completed by a set of arcs, which includes moving and holding arcs. The former correspond to the legs of the potential services, an arc being defined for each service leg. The latter are arcs connecting two time-consecutive representations of each yard, and represent the possibility for equipment and freight to wait at a terminal for a time period.

Recalling that the scheduled service plan is to be applied repeatedly over the season, the time-space network is cyclic to reflect the fact that activities and decisions do not stop with the end of the schedule length, but rather involve the next application of the plan. Thus, for example, when building a week-long schedule, a service may start on Friday and arrive at destination the following Tuesday. These situations are modeled through “wrap-around” arcs (when the destination time of an arc is later than the end of the schedule, the destination time is set to the corresponding time in the current schedule through a modulo computation).

The general SSND integrated-planning modeling framework addresses the same main tactical-planning issues identified before, service selection and scheduling, blocking and classification, train makeup, and freight routing. The goal is again to minimize the total generalized cost of the system, while satisfying demand with the available resources. The notation, definitions, and structure of the mixed-integer network design formulation introduced for the SND case above applies to the time-dependent SSND case, but on the time-space network. The resulting formulations may be very large and require advanced solution methods, but provide the advanced decision-support instruments for the problem.

Chouman and Crainic (2020) detail this methodology in the general case (see also Zhu et al., 2014) and discuss extensions to intermodal settings with continuous time representation (Morganti et al., 2020) and SSND models integrating resource management concerns (Andersen et al., 2009; Kienzle et al., 2020; Pedersen et al., 2009; see also Crainic and Hewitt, 2020).

Using Service Network Design Models

The service network design methodology developed for tactical planning may be used rather broadly to address other railroad planning problems.

SND and SSND models may be expanded to include strategic-level decisions, for example, investments to enhance infrastructure, acquisition or renting resources, and outsourcing service in particular regions. An extra design layer is added to the formulation to represent the strategic alternatives to choose among, the tactical-planning layers providing the evaluation of those decisions within the global optimized plan (Crainic et al., 2018; Hewitt et al., 2019).

The tactical formulations may also be directly used as a simulation tool to perform cost-benefit analyzes and evaluate the impact of strategic scenarios on operations, resources, and system performance. Typical scenarios may concern variations in, for example, economy (e.g., fuel prices or changing production and consumption levels of certain goods), laws (e.g., emission charges or trade restrictions), or regulations (e.g., speed and weight limits when carrying hazardous goods).

The service network design formulations may also be adapted to review, weekly for example, the tactical plan built for the season. One would then re-optimize and adjust the plan and operations to current conditions. What may be adjusted depends strongly on the railroad application context. Canceling or adding services on a short notice is not easily performed by railroads and SND models may assist in selecting the best alternative and determining the network-wide impacts. Updating the actual demands or resources, or both, assigned to blocks and trains is taking place quite often within railroad management and, again, SND models are appropriate.

Conclusions

The chapter presented a high-level overview of freight railroad planning and the service network design models and methods proposed to address its challenges, as well as pointers to further reading and more in-depth study.

It is noteworthy that this is a field still evolving, both in how freight railroad transport, and the world around, evolves and in the modeling and algorithmic challenges these application raise. Chouman and Crainic (2020), and the references cited therein, identify several important research perspectives. We do not have the space here for an extensive list. We emphasize only the need for significant research efforts in (1) the integration of the main components of railroad planning at the tactical level, from scheduled service selection to resource management; (2) the development of efficient solution methods for large problem instances, including the dynamic generation of the sets of potential services and blocks, decomposition methods, and parallel optimization; (3) the study and explicit integration of uncertainty, in demand and travel and yard-activity times, into SND planning models and the development of appropriate solution methods; (4) the definition of revenue management mechanisms for rail transport and the study of the interactions between these operational-level mechanisms and tactical planning processes and service network design models and methods.

Acknowledgments

The author wishes to address the warmest thanks to the collaborators and students who worked with him over the years to develop models and methods for network design and railroad planning. The research benefited from the environment offered by the CIRRELT, its staff and professionals. The author thanks these individuals, as well as the Fonds de recherche du Québec, the Université du Québec à Montréal (UQAM) and the Université de Montréal (UdeM), for the financial contributions, the infrastructures and the resources they offer through the CIRRELT. The author also gratefully acknowledges the research-funding support of the Government of Canada, through its NSERC programs, of the Government of Québec, through its FRQ programs, and of public and private partners. The author holds the Chair on Intelligent Logistics and Transportation Systems Planning of UQAM. The financial contributions of ClearDestination and NSERC to the Chair are gratefully acknowledged. He is Adjunct Professor, Computer Science and Operations Research Department, UdeM, and Co-Director, Intelligent Transportation Systems Laboratory, CIRRELT.

See Also

Environmental Sustainability in Freight Transportation; Freight Network Modeling; The Value of Time in Freight Transport; Container (Liner) Shipping; Scheduling of Liner Container Shipping Services; The Physical Internet and Logistics; Modeling Mode Choice in Freight Transport; Railway Station and Network Planning; Introduction to Rail Transport; The History of Rail Transport; The Geography of Rail Transport; Railway Management; Maritime Route Planning; Planning for Rail Transport

References

- Ahuja, R.K., Cunha, C.B., & Bahin, G., 2005. Network scheduling. In: *Tutorials in Operations Research INFORMS 2005*, INFORMS, pp. 54–101.
- Ahuja, R.K., Jha, K.C., Liu, J., 2007. Solving real-life railroad blocking problems. *Interfaces* 37, 404–419.
- Andersen, J., Crainic, T.G., Christiansen, M., 2009. Service network design with asset management: formulations and comparative analyses. *Transport. Res. C: Emerg. Technol.* 17 (2), 197–207.
- Assad, A.A., 1980a. Modelling of rail networks: toward a routing/makeup model. *Transport. Res. B: Methodol.* 14, 101–114.
- Assad, A.A., 1980b. Modelling of rail networks: models for rail transportation. *Transport. Res. A: Policy Pract.* 14, 205–220.
- Barnhart, C., Jin, H., Vance, P.H., 2000. Railroad blocking: a network design application. *Oper. Res.* 48 (4), 603–614.
- Bektaş, T., Crainic, T.G., 2008. A brief overview of intermodal transportation. In: Taylor, G.D. (Ed.), *Logistics Engineering Handbook*, 28. Taylor and Francis Group, Boca Raton, FL, USA, pp. 1–16.
- Chouman, M., Crainic, T.G., 2020. Freight railroad service network design. In: Crainic, T.G., Gendreau, M., Gendron, B. (Eds.), *Network Design with Applications in Transportation and Logistics*, Springer, Boston, forthcoming.
- Cordeau, J.F., Toth, P., Vigo, D., 1998. A survey of optimization models for train routing and scheduling. *Transport. Sci.* 32 (4), 380–404.
- Crainic, T.G., 1988. Rail tactical planning: issues Models and Tools. In: Bianco, L., La Bella, A. (Eds.), *Freight Transport Planning and Logistics*. Springer-Verlag, Berlin, pp. 463–509.
- Crainic, T.G., 2000. Network design in freight transportation. *Eur. J. Oper. Res.* 122 (2), 272–288.
- Crainic, T.G., 2003. Long-Haul freight transportation. In: Hall, R.W. (Ed.), *Handbook of Transportation Science*. 2nd ed. Kluwer Academic Publishers, Norwell, MA, pp. 451–516.
- Crainic, T.G., 2009. Service design models for rail intermodal transportation. In: Bertazzi, L., Speranza, M.G., van Nunen, J.A.E.E. (Eds.), *Lecture Notes in Economics and Mathematical Systems*, 619. Springer-Verlag, Berlin, Heidelberg, pp. 53–67.
- Crainic, T.G., Hewitt, M., 2020. Service network design. In: Crainic, T.G., Gendreau, M., Gendron, B. (Eds.), *Network Design with Applications in Transportation and Logistics*. Springer, Boston, forthcoming.
- Crainic, T.G., Kim, K.H., 2007. Intermodal transportation. In: Barnhart, C., Laporte, G. (Eds.), *Transportation, Handbooks in Operations Research and Management Science*, vol. 14 North-Holland, Amsterdam, pp. 467–537 (chapter 8).
- Crainic, T.G., Gendreau, M., Gendron, B. (Eds.), 2020. *Network Design with Applications in Transportation and Logistics*. Springer, Boston, forthcoming.
- Crainic, T.G., Ferland, J.A., Rousseau, J.M., 1984. A tactical planning model for rail freight transportation. *Transport. Sci.* 18 (2), 165–184.
- Crainic, T.G., Hewitt, M., Toulouse, M., Vu, D.M., 2018. Scheduled service network design with resource acquisition and management. *EURO J. Transport. Logist.* 7 (3), 277–309.
- Dejax, P.J., Crainic, T.G., 1987. A review of empty flows and fleet management models in freight transportation. *Transport. Sci.* 21 (4), 227–247.
- Hewitt, M., Crainic, T.G., Nowak, M., Rei, W., 2019. Scheduled service network design with resource acquisition and management under uncertainty. *Transport. Res. B: Methodol.* 128, 324–343.
- Kienzie, J., Crainic, T.G., Frejinger, E., Bisailon, S., 2020. The Intermodal Railroad Blocking & Car Fleet Management Planning Problem. Tech. Rep. CIRRELT-2020, Interuniversity Research Centre on Enterprise Networks, Logistics and Transportation, Université de Montréal, Montréal (QC), Canada.
- Mantovani, S., Morganti, G., Umang, N., Crainic, T.G., Frejinger, E., Larsen, E., 2017. The load planning problem for double-stack intermodal trains. *Eur. J. Oper. Res.* 267 (1), 107–119.
- Morganti, G., Crainic, T.G., Frejinger, E., Ricciardi, N., 2020. Block planning for intermodal rail: methodology and case study. *Transport. Res. Procedia* 47, 19–26.
- Newman, A.M., Nozick, L.K., Yano, C.A., 2002. Optimization in the rail industry. In: Pardalos, P.M., Resende, M.G.C. (Eds.), *Handbook of Applied Optimization*. New York, NY, Oxford University Press, pp. 704–718.
- Pedersen, M.B., Crainic, T.G., Madsen, O.B.G., 2009. Models and Tabu search meta-heuristics for service network design with asset-balance requirements. *Transport. Sci.* 43 (2), 158–177.
- Piu, F., Speranza, M.G., 2014. The locomotive assignment problem: a survey on optimization models. *Int. Trans. Oper. Res.* 21 (3), 327–352.
- Yaghini, M., Akhavan, R., 2012. Multicommodity network design problem in rail freight transportation planning. *Procedia Soc. Behav. Sci.* 43, 728–739.
- Zhu, E., Crainic, T.G., Gendreau, M., 2014. Scheduled service network design for freight rail transportations. *Oper. Res.* 62 (2), 383–400.

Subway Systems

Guillaume Monchambert*, **Daniel Hörcher†**, **Alejandro Tirachini‡**, **Nicolas Coulombel§**, *Univ Lyon, Université Lyon 2, LAET, Lyon, France; †Transport Strategy Centre, Imperial College London, London United Kingdom; ‡Civil Engineering Department, Universidad de Chile and Instituto Sistemas Complejos de Ingeniería (Complex Engineering Systems Institute), Chile; §LVMT, Ecole des Ponts, Université Gustave Eiffel, Champs-sur-Marne, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	471
A Brief History	471
Early Days and Expansion	471
Deeper and Deeper	472
Electrification	472
Automation	472
Supply	472
Network	472
Rolling Stock	474
Operator Costs	474
Demand	474
Total Ridership	474
Spatial and Temporal Demand Patterns	475
Pricing	475
Subsidies and Value Capture	475
Distance-Based Versus Flat Pricing	476
Externalities	476
Mohring Effect	476
Crowding	477
Reliability	477
Conclusion	477
See Also	477
Acknowledgment	477
References	478
Further Reading	478

Introduction

Subway (also called metro, underground, and tube) is an electric railway on devoted rights-of-way, fully segregated from other means of motorized and unmotorized transport. Designed for dense cities, subway systems have a high capacity, able to handle more than 30,000 passengers per hour and per direction.

This chapter relies on several occasions on data from the CoMET and Nova subway benchmarking groups facilitated by the Transport Strategy Centre at Imperial College London. These groups, established in the early 1990s, enable the largest metro operators in the world to share performance data and best practices in a systematic and scientifically rigorous way. Their membership includes Berlin, London, Madrid, Moscow, and Paris in Europe; Beijing, Guangzhou, Hong Kong, Seoul, Shanghai, Singapore, Taipei, and Tokyo in Asia; and New York, Mexico City, Santiago de Chile, and Sao Paulo in the Americas¹.

The chapter has five main sections. Section [A Brief History](#) describes the history of subway systems with an emphasis on major innovations. In Section [Supply](#), we characterize the subway supply through its physical and financial aspects. Section [Demand](#) provides quantitative and spatial information on demand. In Section [Pricing](#), we discuss usual pricing and subsidies schemes. Section [Externalities](#) zooms in on a number of subway-related externalities. Conclusions are provided in the final section.

A Brief History

Early Days and Expansion

The first subway system opened in London in 1863. This 6-kilometer-long line was a great success, with 38,000 passengers on the opening day and 9.5 million passengers in the first year of service. Other cities followed London and built their own subway systems:

¹The full list of benchmarking group members is available online: cometandnova.org.

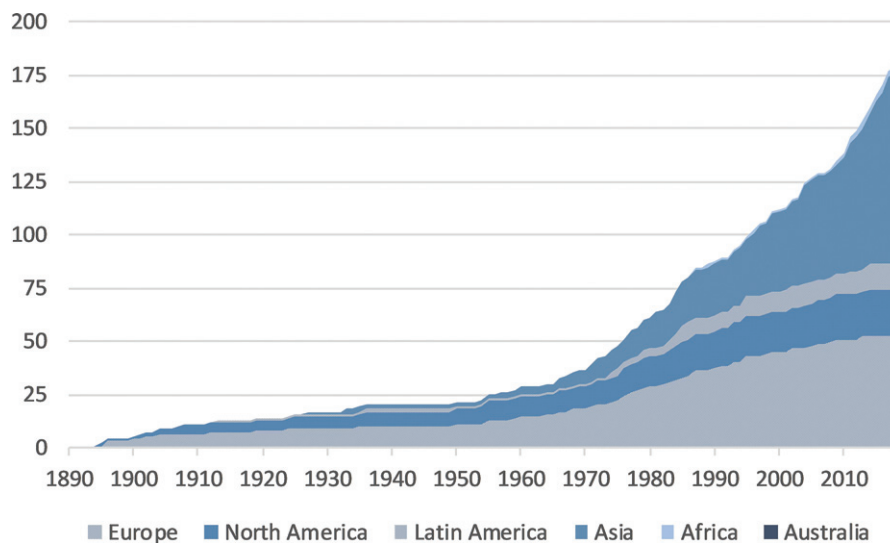


Fig. 1 Number of subway systems in operation by continent. Source: own elaboration.

Budapest and Glasgow in 1896, Boston in 1897, Paris in 1900, New York in 1904, Buenos Aires in 1913, Tokyo in 1913, and Moscow in 1935. Today, there are almost 200 subway systems in operation around the world (see Fig. 1). The significant increase in the number of systems since the 1970s is mainly driven by Asia, particularly by the opening of several networks in China.

Deeper and Deeper

Over time, subway lines have been built deeper below ground, from subsurface to deep level. The first subway lines were built using the cut-and-cover method: this means that a trench is excavated from the surface, which is then roofed over with an overhead support system. In Paris, the rails are sometimes only 4.5 m below the level of the roadway. For construction convenience, most of the first subway lines follow the layout of the streets above.

The deep-level lines are built with a tunneling shield. These lines are often at a depth of more than 10 meters below ground. Some stations may be more than 100 meters below the ground. The Arsenalna station of the Kiev subway in Ukraine is the deepest in the world: it is 105.5 meters deep. It is followed by Admiralteyskaya station in St. Petersburg (102 m) and Puhung station in Pyongyang (100 m). Grand Paris Express lines are planned between 15 and 55 meters below ground.

Electrification

The first trains were pulled by steam locomotives, which required efficient ventilation from the surface. Ventilation ducts at various points along the route allowed steam locomotives to expel smoke and bring fresh air into the tunnels. However, the use of steam locomotives led to smoke-filled stations and carriages that were unpopular with passengers. As a result, the electrification of the London Underground began in 1890. From then on, all new lines had become electrically powered. Electric power is usually supplied via third (and fourth) rails, in order to save the space that overhead wires and pantographs would require in the underground environment.

Automation

Automation of operations is the latest major technological improvement of subway systems. A subway line is said to be automated when the responsibility of train operations is transferred from the driver to a train control system. The UITP defines five degrees of automation, from 0, which corresponds to on-sight operation, to 4, in which trains run without an operator on board.

Automation may increase the capacity of subway systems, due to a reduction in the safety headway (i.e., the minimum headway between two consecutive trains). The coordination of the acceleration and braking phases also saves energy. Finally, the supply is more flexible because the frequency can be increased without the need for additional driver. The staff savings need to be qualified because the savings made on drivers are partly offset by the creation of new positions to manage the line.

Supply

Network

The length of subway systems varies considerably between cities (the total commercial line length goes from a few kilometers to almost 700 km, see Fig. 2). Some systems serve the suburbs, others only the city center. The network is then complemented by other

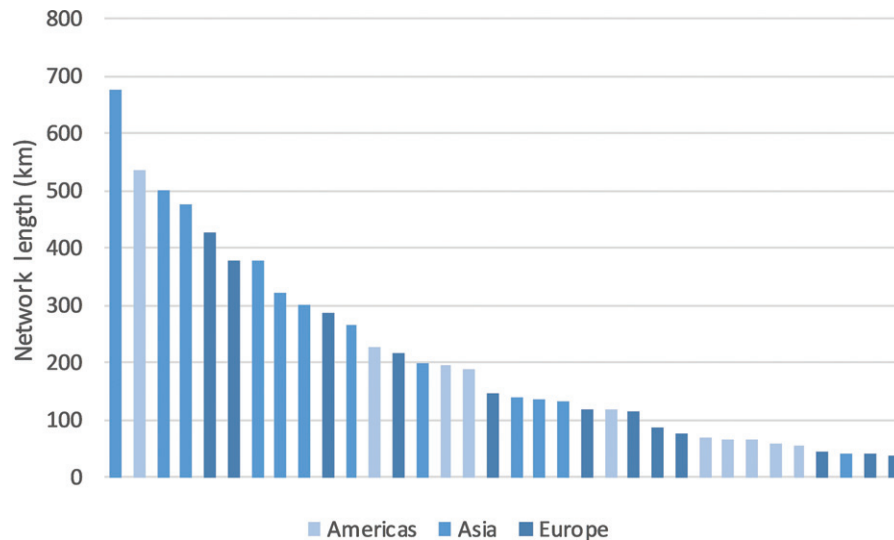


Fig. 2 Network length in kilometers of commercial lines of 34 networks in the CoMET and Nova subway benchmarking groups.

means of public transport such as feeder bus services or suburban trains. The topology of the networks is also varied, with networks taking different shapes depending on geographical and historical constraints, travel patterns and construction costs. Roth et al. (2012) studied the 15 largest networks in the world and identified a convergence of network forms towards a form composed of a dense core, fed by branches radiating from the core.

Each subway system consists of one or more lines, distinguished by colors, names, or numbering and served by at least one specific route. Trains stop at all or some of the stations on the line. Lines running through the city center can be separated into two or more branches in the suburbs, allowing for higher frequency service where demand is highest.

The hourly capacity of a line is obtained by multiplying the car capacity (passengers/car), the number of cars per train and the service frequency (vehicles/h). A typical metro line can accommodate 150 passengers per car, 8 cars per train, and 30 trains per hour, giving 36,000 people per hour and direction. Actual maximum car occupancy depends on seating and standing configurations and on the level of passenger crowding that is socially accepted in each city. Fig. 3 provides an overview of the annual supply density for some subway systems in the CoMET and Nova subway benchmarking groups, where supply density is defined as standard vehicle capacity kilometers normalized by the commercial network length. The highest capacity is sufficient for almost 400 million passengers per network kilometer per year, but most networks offer between 50 and 200 million passengers.

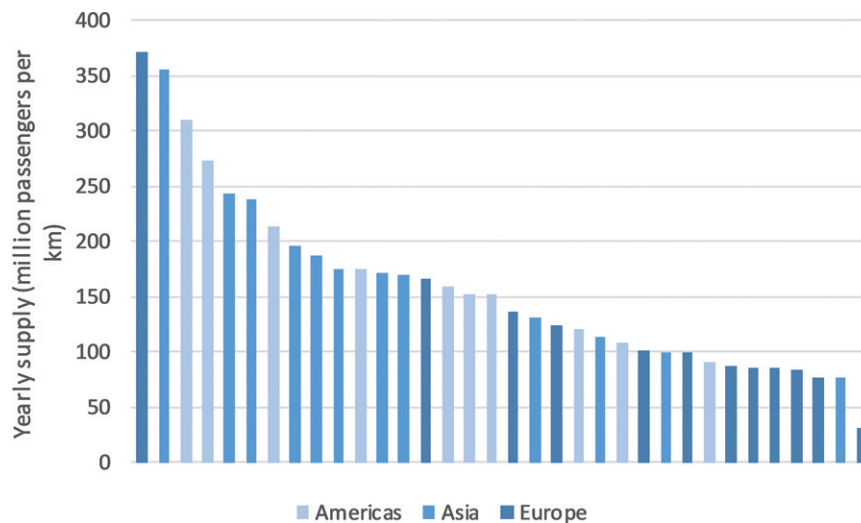


Fig. 3 Yearly supply density in million passengers per kilometer of 34 networks in the CoMET and Nova subway benchmarking groups.

Rolling Stock

Subway vehicles are high-performing assets designed for highly standardized, uniform railway operations. The vehicles are almost exclusively electric multiple units with a driving cab at both ends of the train (with the exception of some of the fully automated vehicles). Given that platform length and, in general, station capacity is a critical determinant of the cost of subway construction, vehicles are designed to maximize the proportion of the train length that can be utilized by passengers. Other dimensions of subway vehicles are mainly defined by the diameter of underground tunnels, which is again subject to a trade-off between construction costs and capacity.

Several unique features of subway vehicles are meant to increase capacity and travel speed, as experienced by passengers. High capacity can be achieved by (1) short headways, (2) reduced dwell times by ensuring fast boarding and alighting, (3) effective acceleration and deceleration, and (4) targeting maximum interior capacity utilization within vehicles. The upper bound of service frequency is the minimum distance between consecutive trains, where advanced communication systems are deployed to reduce this gap without compromising safety standards. The ultimate lower bound of the headway is the dwell time itself: the dwell time at the busiest station determines service frequency on the entire line, and therefore speedy boarding and alighting at this bottleneck is critical to overall line performance. To ensure fast acceleration and deceleration, electric power units are distributed along the vehicle, often to each wheel set. Rubber-tire technology, which is unique to the urban rail industry, further enhances traction. Finally, subway vehicles feature lowered seated to standing ratios and large doors to increase total capacity and speed up interior passenger flows for boarding and alighting.

Operator Costs

Among 31 members of the CoMET and Nova subway benchmarking groups, the average breakdown of major expenditure components in 2018 is as follows:

- Operational and service costs: 28.8% (train service costs 13.3%, station service costs 9.7%, service management, and support 5.7%);
- Maintenance costs: 24.1% (rolling stock maintenance 9.1%, infrastructure maintenance 8.5%, and station facilities maintenance 6.5%);
- Administrative costs: 12.7%;
- Renewals and investments: 34.4% (nonexpansion investments 19.9%, expansions completed in the year of observation 3.1%, and ongoing investments 11.5%).

From this breakdown we highlight that the sum of maintenance and nonexpansion investment costs constitute around 44% of all monetary expenditures of subway operators, while 29% is allocated to direct operational costs. This ratio illustrates the presence of substantial fixed costs in the industry, and a relatively small share of operational costs directly related to the firm's final output, that is, the number of passengers carried. Another important inference from this data is that in a representative sample of the global subway industry, almost 15% of all financial expenses are invested into service expansion, which showcases the prevalent growth rate in urban rail transport.

More insights can be gained on the production process and the costs of subway service via statistical modeling techniques. The empirical literature of cost function modeling concentrates on the identification of returns to scale (RTS) and returns to density (RTD) on the level of individual operators and the industry as a whole. RTS measures the impact of firm (network) size on the average operational cost, while RTD relates infrastructure utilization (i.e., service frequency on a network of fixed size) to the average cost. A major challenge in the unbiased estimation of these indices is the potential endogeneity of subway output, which is likely to be affected by the unobserved, subway-specific level of productivity. [Anupriya et al. \(2020\)](#) find strong scale economies with respect to both network size and service density: their RTS estimate is 1.223 while RTD is 1.562, which are higher than the majority of previous estimates that do not control for output and factor price endogeneity. Scale and density economies imply that (1) network expansions lead to external benefits in the form of network-wide reduction in operational costs, and (2) subsidies are justified from an operational cost point of view as well.

Demand

Total Ridership

Subway is a mode of mass transportation that allows to accommodate high passenger density along a transport corridor. A total of 178 metro networks identified by UITP (International Association of Public Transport) provided more than 53 billion journeys in 2017 ([UITP, 2018](#)), compared with only 44 billion in 2013, which represents an increase of over 21%.

This growth in global patronage dissimulates quite different trends across the continents. [Fig. 4](#) shows the annual usage of CoMET and NOVA networks by continent. Asia's leading role in number of networks (see [Fig. 4](#)) is paralleled by the increase in ridership as well, with a 83% increase between 2008 and 2018. Growth is more moderate in Europe (+100 million between 2008 and 2018), whereas ridership increases on average by 13.5% between 2008 and 2018 in Americas subways networks.

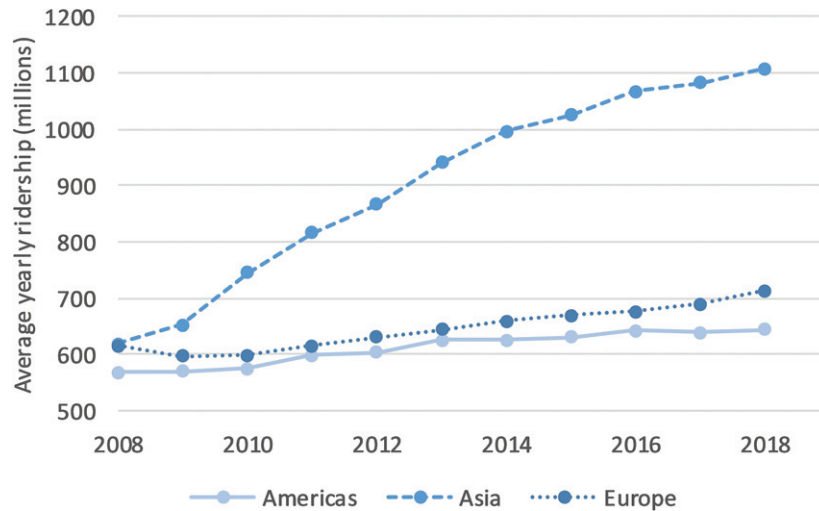


Fig. 4 Average network yearly ridership by continent in million trips, based on 34 networks in the CoMET and Nova subway benchmarking groups.

Spatial and Temporal Demand Patterns

The actual route density and direction of lines determine the temporal and spatial characteristics of subway demand. Radial lines are usually characterized by a high demand toward the Central Business District (CBD) in the morning peak and a relatively low demand in the opposite direction. This type of demand unbalances sometimes leads to the application of strategies to increase supply in the peak direction, for example through station skipping and short-turning. Circular lines may have more balanced demand patterns both temporally and spatially, depending on the land use around the stations. [Hörcher and Graham \(2018\)](#) show that the magnitude of demand imbalances has crucial consequences on the financial and economic performance of public transport operations.

Pricing

Subsidies and Value Capture

The amount of subsidy required by a subway system to finance its operations is a subject of intensive policy debates worldwide. Given the presence of strong scale and density economies, competition is rare in the urban rail industry and subways may have substantial monopoly power in the travel market. It is a consensual standpoint in public policy, however, that transport providers should not pursue a profit maximizing objective, and therefore subways are normally under public ownership or subject to strict price regulation.

Economic theory does not offer direct rules to determine the welfare maximizing subsidy. The optimal financial deficit (or surplus) is a derived quantity: an outcome of the optimization of fares and capacity supply, as the latter determines the magnitude of operational costs. In principle, capacity shortages, crowding externalities and costly public funds reduce the optimal subsidy, while the degree of scale and density economies, substitution with car use, the severity of car externalities, and potential agglomeration benefits are typical drivers of increased subsidies ([Hörcher et al., 2020](#)). These parameters may vary on a wide range in large metropolitan areas of the world. However, we observe a tendency of discontent among metro operators who claim that current subsidies are inadequate to maintain a good level of service, cope with increasing demand and customer expectations, and keep expensive assets in good state of repair. One potential explanation is an inconsistency in how long-run operational, maintenance and renewal costs are taken into account by funding bodies, including local and higher level governments.

Theory suggests that pricing policies should be driven by the short-run marginal social cost of subway usage. In practice, prices are set with several different social and economic objectives, subject to an externally defined subsidy, while investments and long-run decisions are evaluated on the basis of separate cost-benefit analyses. If governments do not evaluate the funding need of costly renewals and service upgrades on a regular basis, the financial health of subway operations may become uncertain. Under such circumstances, it has become a popular belief among subway operators that they should achieve financial independence from governments in terms of renewals and investments by covering these costs from commercial revenues. While this belief is a natural response to the uncertainty of government funding, the resulting cost recovery ratios may diverge from their welfare maximizing levels, in the form of fare hikes or sub-optimal capacity provision. Another major trend in the industry is that transport operators intend to increase their commercial revenues from non-transport activities such as real estate development. Once again, this turns out to be an efficient tool to reduce reliance on government funding and ensure financial stability, but if the additional revenues are generated using the operator's monopoly power in the related nontransport markets, then the resulting distortions may offset the perceived benefits from a social welfare point of view.

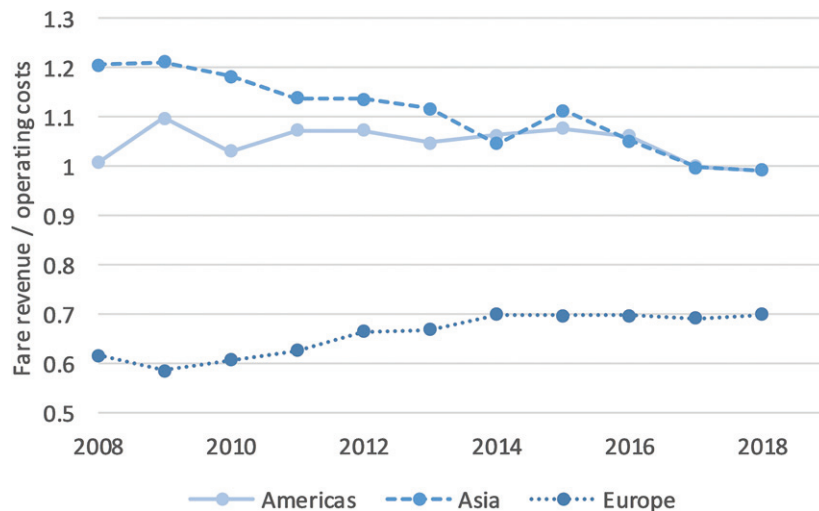


Fig. 5 Cost recovery ratios by continent, based on 34 networks in the CoMET and Nova subway benchmarking groups.

Fig. 5 plots the evolution of cost recovery ratios among 24 members of the CoMET and Nova metro benchmarking groups between 2008 and 2018. The cost recovery (or self-financing) ratio is the fraction of operational costs covered by fare revenues. Renewals and investment expenditures are excluded from operating costs, while the revenue side includes ticket sales only. We observe that subways in Asia and the Americas are close to full cost recovery. European subways cover a substantially lower share of their operating cost from fare revenues. The plot shows a marginal increase in the ratio from the range of 60% around 2010, which has gradually increased to a stable level of 70%. The density of the urban environment in which subways operate in Asia and Latin America, and the resulting large occupancy levels and capacity shortages, might be a key reason behind the systematic difference in this metric. However, it is worth noting that the standard deviation of self-financial ratios is 0.3 among European metros, and substantially higher, 0.4, for cities in both Asia and the Americas.

Distance-Based Versus Flat Pricing

The way in which the intensity of use, for example, trip length, should be considered in the determination of fares has been a matter of discussion and research in the academic literature (see, e.g., [Kerin, 1992](#), [Jara-Díaz and Gschwender, 2005](#)) and there is no definitive answer on whether or not fares should increase with travel distance. [Turvey and Mohring \(1975\)](#) show that first best public transport fare can decrease with travel distance in radial corridors, if in the peak direction (toward the CBD), passengers that travel shorter distances get on trains when more passengers are already on board, which represents a greater social cost. This inverse relation between fare and travel distance might not hold if we consider that long-distance passengers yield greater discomfort externalities than short-distance passengers due to crowding and standing externalities ([Kraus, 1991](#), [Hörcher et al., 2018](#)). Because in metro systems, dwell times are much more standardized than in buses service, the delay effect is usually negligible compared to the crowding externality, so distance-based fares are likely to be more efficient in the case of metro lines.

In practice, subway systems that have zonal or distance-based fares reflect that longer trips also have a larger marginal operator cost, as commonly done in Europe. At the same time, there are plenty of subway systems with flat fares for single trips, regardless of trip length, which can be based on simplicity or on equity grounds, in those cities where lower income households are concentrated in the outskirts, pushing them to travel longer distances for work. The latter is an argument put forward in some Latin American cities to have flat fares in subway systems.

Externalities

Mohring Effect

Subways systems are designed to accommodate a large number of users, which allows benefiting from substantial economies of scale when facing sufficient demand. This includes economies of scale on the supply side, as well as on the user costs. To illustrate, consider that if an increase in demand is met with an increase in supply (service frequency), average waiting time cost is reduced for each user. The average user cost therefore decreases with patronage, a phenomenon often called the Mohring effect. Empirically, the Mohring effect is found to justify subsidizing public transport. Yet, as patronage of the subway line increases the various forms of public transportation congestion (in-vehicle, on platform for boardings and alightings), this leads to a partial or full mitigation of the Mohring effect (see [Section Subsidies and Values Capture](#)).

Crowding

Crowding becomes an increasing concern in several subway networks across the world. As public transit use intensifies as a result of population growth or transit oriented policies, the number of users within each train (in-vehicle density) and on the platform increase. On the most congested lines, large queues may develop during the rush hour within or even outside the subway station.

These phenomena negatively affect the user experience in several ways ([Tirachini et al., 2013](#)). First, crowding implies discomfort as some users become unable to sit. Users, seated or standing, may also be hindered in their activities (reading, using their mobile phone, etc.) if other users are too close. Excessive proximity may also generate anxiety, or accentuate the perception of risk to personal safety and security. These cause the value of travel time savings (VITS) to increase with the crowding level. For instance, [Whelan and Crockett \(2009\)](#) find that the VITS of a sitting (standing) passenger increases by a factor 1.00–1.63 (1.53–2.04) as in-vehicle density rises from 0 to 6 pax per square meter. Additionally, crowding also increases travel time as it leads to longer boarding and alighting times, or even waiting times when users must wait for the next train.

The economic cost of crowding is accordingly substantial for heavily used subway networks. In the case of Paris, [Haywood et al. \(2018\)](#) find a marginal external cost of 0.123€ per passenger – kilometer at rush hour, resulting in a welfare loss of approximately 65 M€/year. To mitigate this cost, subway network managers may resort to several strategies. First improving frequency, including through the automation of the line which allows higher frequencies and makes it less costly to do so. The vehicle design may also be enhanced to increase the available floor space and the number of seats through communicating cars or double-decker vehicles. Finally, travel demand management policies such as crowding-dependent pricing ([de Palma et al., 2017](#)), alternative work schedules or passenger flow control may be used to regulate demand.

Reliability

Unreliable travel times cause travelers to arrive earlier or later than expected at their destination, which is a major source of discomfort as shown by the travel behavior literature. Thus, modal reliability is a key factor that travelers consider when making basic travel choices, such as decisions regarding mode, route, and departure time. How predictable travel times are make a mode more or less reliable. The literature on travel time variability usually shows that subway systems are more reliable than motorized road-based modes such as buses and cars, when these are subject to congestion ([Durán-Hormazábal and Tirachini, 2016](#)).

Conclusion

Subway systems deliver unparalleled performance in high-capacity transport provision. Their efficiency might further increase due to numerous efforts aiming at standardizing subway constructions and operations. Under regular conditions, it is anticipated that the global growth of subway ridership and the expansion of the number and coverage of subway networks will continue, fueled by urbanization. The ongoing global health crisis imposes new challenges on the industry, however. With social distancing rules, the full potential of mass rail transit cannot be exploited, and a general shift in urban lifestyle, including more remote working, may jeopardize subway demand anyway. This imposes a short to medium term financial threat on operators, especially on the ones operating mature networks that require substantial renewal expenditures. Eventually, the future of subways is closely linked to the prevalence of the densely populated city model of the early 21st century. Assuming that physical proximity will remain a major requirement for the productivity and competitiveness of cities, it is unlikely that other transport modes will ever be able to reach the efficiency of underground rail technology.

See Also

Operation Costs for Public Transport; The Concept of External Cost: Marginal Versus Total Cost and Internalization; Value of Crowding; Demand for Passenger Transport; The Mohring Effect; Public Transport Fare and Subsidy Optimization

Acknowledgment

We thank the CoMET and Nova metro benchmarking groups facilitated by the Transport Strategy Centre at Imperial College London for providing us unique data on subway systems. Partial funding from ANID Chile (Grant PIA/BASAL AFB180003) is also acknowledged.

References

- Anupriya, Graham, D.J., Carbo, J.M., Anderson, R.J., Bansal, P., 2020. Understanding the costs of urban rail transport operations. *Transp. Res. Part B: Method.* 138, 292–316.
- de Palma, A., Lindsey, R., Monchambert, G., 2017. The economics of crowding in rail transit. *J. Urban Econ.* 101, 106–122.
- Durán-Hormazábal, E., Tirachini, A., 2016. Estimation of travel time variability for cars, buses, metro and door-to-door public transport trips in Santiago, Chile. *Res. Transp. Econ.* 59, 26–39.
- Haywood, L., Koning, M., Prud'Homme, R., 2018. The economic cost of subway congestion: Estimates from Paris. *Econ. Transp.* 14, 1–8.
- Hörcher, D., De Borger, B., Seifu, W., Graham, D.J., 2020. Public transport provision under agglomeration economies. *Region. Sci. Urban Econ.* 81, 103503.
- Hörcher, D., Graham, D.J., Anderson, R.J., 2018. The economics of seat provision in public transport. *Transp. Res. Part E: Logist. Transp. Rev.* 109, 277–292.
- Hörcher, D., Graham, D.J., 2018. Demand imbalances and multi-period public transport supply. *Transp. Res. Method.* 108, 106–126.
- Jara-Díaz, S.R., Gschwender, A., 2005. Making pricing work in public transport provision. *Handbook Transp. Strat. Policy Inst.* 6, 447–459.
- Kerin, P.D., 1992. Efficient bus fares. *Transp. Rev.* 12 (1), 33–47.
- Kraus, M., 1991. Discomfort externalities and marginal cost transit fares. *J. Urban Econ.* 29 (2), 249–259.
- Roth, C., Kang, S.M., Batty, M., Barthelemy, M., 2012. A long-time limit for world subway networks. *J. Royal Soc. Interf.* 9 (75), 2540–2550.
- Tirachini, A., Hensher, D.A., Rose, J.M., 2013. Crowding in public transport systems: Effects on users, operation and implications for the estimation of demand. *Transp. Res. Part A: Policy Pract.* 53, 36–52.
- Turvey, R., Mohring, H., 1975. Optimal bus fares. *J. Transp. Econ. Policy* 9, 280–286. Available from: <https://www.jstor.org/stable/20052415?seq=1..>
- UITP, 2018. World Metro Figures 2018. International Association of Public Transport (UITP), available from: <https://www.uitp.org/publications/world-metro-figures/>.
- Whelan, G., & Crockett, J., 2009. An investigation of the willingness to pay to reduce rail overcrowding. In *Proceedings of the first International Conference on Choice Modelling*, Harrogate, England (Vol. 30).

Further Reading

- Anderson, M.L., 2014. Subways, strikes, and slowdowns: The impacts of public transit on traffic congestion. *Am. Econ. Rev.* 104 (9), 2763–2796.
- Parry, I.W., Small, K.A., 2009. Should urban transit subsidies be reduced? *Am. Econ. Rev.* 99 (3), 700–724.
- Small, K.A., & Verhoef, E.T. (2007). 3.1 Cost functions for public transit. In Routledge, (Ed.). *The Economics of Urban Transportation*.
- Small, K.A., & Verhoef, E.T. (2007). 4.1 Pricing of public transit. In Routledge, (Ed.). *The Economics of Urban Transportation*.
- Subway, 2017. In Britannica editors, (Eds.), *Encyclopaedia Britannica*. Available from: <https://www.britannica.com/technology/subway>.
- Tirachini, A., Hurtubia, R., Dekker, T., Daziano, R.A., 2017. Estimation of crowding discomfort in public transport: Results from Santiago de Chile. *Transp. Res. Part A: Policy Pract.* 103, 311–326.
- Vuchic, V.R., 2017. *Urban transit: operations, planning, and economics*. John Wiley & Sons, United States.
- Wu, J., Gao, Z., Sun, H., Huang, H., 2004. Urban transit system as a scale-free network. *Modern Phys. Lett. B* 18 (19–20), 1043–1049.
- Xu, X.Y., Liu, J., Li, H.Y., Jiang, M., 2016. Capacity-oriented passenger flow control under uncertain demand: Algorithm development and real-world case study. *Transp. Res. Part E: Logist Transp. Rev.* 87, 130–148.

Railway Company Management

Vilius Nikitinas, Skaiste Miliauskaite, Law firm NIKITINAS LEGAL, Vilnius, Lithuania

© 2021 Elsevier Ltd. All rights reserved.

Introduction	479
Management Models	480
Separated Management Model (Vertical Separation)	480
Integrated Management Model (Vertical Integration)	481
Comparison of the Models: Advantages and Disadvantages	482
Summary and Conclusions	484
See Also	484
References	484
Further Reading	484

Introduction

In every state, the need and importance of rail transport differ. Nowadays the rail transport sector faces competition from air, road, and water transport. However, rail transport remains relevant and attractive today. One of the objectives set out in the White Paper adopted by the European Commission in 2011 is to increase the volume of rail freight and passengers ([White Paper, 2011](#)). Seeking to achieve the results in the rail transport sector the management of the railway undertakings has an essential meaning. It should be noted that each country has a substantially unique rail transport practices, for example in one country are more widely developed passenger transport services and in other—freight transport services. In view of these characteristics, a specific management strategy for railway undertakings should be selected and implemented accordingly.

The rail sector has been undergoing reforms in the world for decades, which concern the management of railway undertakings. However, it should be borne in mind that the experience of reform varies, both in terms of the reform models chosen and the consequences for the appropriate implementation of the model. In the United States, for example, there has been a shift from a heavily regulated environment to a less regulated one, while the emphasis in the EU is not on deregulation, but rather on greater harmonization between the Member States and on the goal of creating more competition—this is a trend that can be seen in other countries around the world ([Finger and Künneke, 2011](#)).

The EU requirements in this area are of particular importance when examining the management of railway companies in the Member States. One of the cornerstones of the EU is to open up the markets of the Member States and ensure the development of fair competition between them. On this basis, common guidelines, rules to ensure that these objectives are achieved, are being developed in the rail transport sector. Member States are complying with their obligations when implementing the rail transport packages adopted by the EU:

1. The first railway package was adopted in 1998. The package, consisting of three directives, provided all railway undertakings licensed on the basis of community criteria with equal and nondiscriminatory access to the railway infrastructure in order to offer services in all Europe. The package has led to a major overhaul of the sector's legislation to ensure its integration into the European area and to allow it to compete with other modes on the most favorable terms.
2. The second railway package, covering three directives and a regulation, adopted in 2004, provided for a wider opening of the freight market, promoting liberalization and harmonization of technical standards.
3. The third railway package, consisting of two directives and two regulations, was adopted in 2007. It provided for the full opening of international passenger services as well as for the licensing of railway drivers, rail passengers' rights regulation, etc.
4. The fourth railway package, consisting of three directives and three regulations, was adopted in 2016. One of the main aims was to create a single market for railway services. This package is relevant to the subject under consideration as it lays down requirements for the management of railway undertakings.

Prior to the adoption of the last fourth railway package, it was considered that it would require the complete separation of the infrastructure manager from the provision of services. However, the fourth railway package allowed for a vertically integrated management structure, allowing the infrastructure manager and rail transport services to be combined, but only within the strictly so-called "Chinese walls", ensuring legal, financial, and operational separation ([House of Commons Transport Committee, 2012](#)). The purpose of such EU regulation is to ensure that the infrastructure manager is independent and financially separated from the railway undertakings by performing essential functions (such as allocating public railway infrastructure capacity).

This chapter is structured as follows: the second section describes the management models, the third section analyzes the comparison of the models and the last section presents our conclusions.

Management Models

Models of railway management are essentially singled out according to the relation between the management of public railway infrastructure and the provision of rail transport services. On that basis, two models can be singled out ([Gomez-Ibanez, 2006](#)):

1. Separated management model (vertical separation) where the infrastructure manager is separated from railway services (Sweden, Great Britain).
2. Integrated management model (vertical integration), based on the holding principle, where the infrastructure manager and the railway undertakings are parts of the same organization (Germany, Italy).

A fully integrated management model can be distinguished as a form of integrated management where the infrastructure manager and the railway service provider (carrier) are one and the same undertaking. Such a model existed in many European countries before the reforms of railway companies (Lithuania) and is based on the monopoly principle.

It is noteworthy that the EU is unique in that it seeks to separate infrastructure manager and rail transport services in order to ensure nondiscriminatory conditions for railway undertakings. In other parts of the world, different models prevail, such as:

- In Japan railways have a vertically integrated structure. The companies generally own their rolling stock as well as the infrastructure on which they operate.
- In the United States, most rail networks are owned by private freight carriers. However, long-distance passenger services operated by Amtrak (effectively owned by the government) are mainly carried on lines owned by freight carriers or government agencies ([Williams Rail Review, 2019](#)).

Separated Management Model (Vertical Separation)

The management of British and Swedish railway companies are the most frequently analyzed examples of vertical separation. Models from these countries are a frequent subject of evaluation to see if a real separation between infrastructure and services can achieve greater efficiency in the rail sector.

The British Rail industry has undergone a number of reforms. The privatization process in Britain between 1994 and 1997 saw the dismantling of “British Rail” and the separation of public rail infrastructure and rail transport services. Ownership has largely gone to the private sector. At the time of privatization, the private company Railtrack became the owner of the rail infrastructure network. Later Railtrack was transformed into the railway administration and “Network Rail” (NR). NR was founded at the time as a private company, but in 2014 NR was retrained into a public sector body. At the time of privatization, “British Rail” passenger train fleet was transferred to private rolling stock operating companies (ROSCO). ROSCO owns most of the rolling stock and leases it to the companies operating the trains. British Rail freight trains have been sold to freight carriers—English Welsh and Scottish, freightliner and others.

Currently, the rail sector in Great Britain includes the infrastructure manager, NR, the independent economic and security regulator, the Office of Rail and Road, and passenger and freight services provided by private railway companies. Most of them are subsidiaries of railway companies such as Deutsche Bahn (German national rail operator), SNCF (French national rail operator), and ND (Dutch national rail operator).

The main features of today's railway management in Great Britain are as follows:

- The network infrastructure is state-owned and operated by NR, which has one shareholder, the Secretary of State for Transport. The network is operated and maintained through the setting of access charges and the subsidies to NR.
- The majority of passenger train services are provided through publicly established franchise agreements. In addition to franchising, there are few private passenger carriers on the network. These operators do not have a contract with public authority and provide a very small proportion of passenger transport services.
- Freight carriers are private sector companies. Access charges are payable to NR.
- Most of the British fleet is owned by the private sector ROSCO and leased to train operators for the duration of the franchise agreements.
- The Office of Rail and Road plays an important role in regulating NR and train operators. The role of the bureau includes independent safety regulation, ensuring access rights that NR offers to its customers (operators of passenger and freight services), etc. ([Williams Rail Review, 2019](#)).

One of the foundations of the reforms in Great Britain was the conviction that private ownership and management is better than public ownership, as private companies are better able to ensure that market needs are met for maximum profits. For example, this explains why even public rail infrastructure has been privatized ([Alexandersson and Hulten, 2006](#)). It should be noted, however, that infrastructure has subsequently returned to the hands of the state, and that privatization of infrastructure has often been seen as a mistake of the reforms in Great Britain and an example to be learned by other countries.

Another country where the separation between infrastructure manager and services was also implemented is Sweden. A reform of the railway sector in Sweden was initiated in 1960 with the aim of liberalizing the market and creating conditions for competition. The Swedish State Railways (SJ) was a state business administration with a monopoly on freight and passenger rail transport, completely isolated from competition. However, SJ was in difficulties and it was necessary to reform the railway sector. The state has

faced problems such as the reduction in passenger numbers due to the use of private cars, the closure of passenger lines, etc. In 1988 The Transport Policy Act separated rail infrastructure from rail transport services. The new state-owned company Banverket became fully responsible for infrastructure and maintenance investments, and SJ was transformed into a service provider paying access fees. In 2000 the Transport Policy Act changed SJ organizational structure into several companies focusing on different sectors of the railway: SJ (Passenger Division), Green Cargo (Cargo Division), Jernhusen (Real Estate Division), EuroMaint and SweMaint (Service Department), TrafficCare (Cleaning Services), and Unigrid (Computer Information System Division). In 2007 TrafficCare, Unigrid, EuroMaint, and SweMaint have been privatized. Legislation was subsequently amended to bring the legal framework into line with the EU requirements. In order to improve infrastructure management, Banverket was merged with Vagverket Road Administration and was formed Trafikverket to improve infrastructure management.

The main features of today's railway management in Sweden are:

- The national authority Trafikverket owns state railway infrastructure (80% of the railway lines).
- Most railway stations are owned and maintained by Trafikverket.
- Jernhusen, state-owned real estate firm, owns station buildings.
- EuroMaint and SweMaint provide maintenance services.
- Train Traffic control, a traffic management unit, in conjunction with Trafikverket monitors all train movements on the Swedish rail network.
- The Swedish Transport Agency is the regulatory body responsible for safety issues, permits, etc.
- State-owned Green Cargo is the largest freight carrier, accounting for approximately 70% of the market (Alexandersson and Hultén, 2011).

In Sweden recurring problems were key drivers of reform seeking to make SJ profitable. SJ problems were due to the high level of competition from other modes of transport, and the main aim of the reforms was to improve rail's ability to compete with other modes of transport. This was a key objective of the 1988 reform. Another important factor was the maintenance of unprofitable lines for social use. However, the reform process in Sweden was gradual compared to the more radical measures implemented in Great Britain (Alexandersson and Hultén, 2006).

Integrated Management Model (Vertical Integration)

Germany has implemented integrated railway management model. In Germany the infrastructure manager and the railway services providing companies are separated—they have separate accounts and boards, but are accountable to the holding company—Deutsche Bahn AG (DB AG). Alongside the vertically integrated companies of DB AG, the market is also served by private freight and passenger transport companies. Reform in Germany began in 1980 and DB AG was transformed into a holding company in 1999. Both DB AG and private operators have the right to negotiate access to the infrastructure and to require access to other facilities owned by DB AG companies such as DB Station Service, DB Energie, etc.

The main features of today's railway management in Germany are:

- Infrastructure in Germany is managed by subsidiaries of DB AG: DB Netze Energy, DB Netze Stations, and DB Netze Track. DB Netze Energy is responsible for supplying all types of energy to DB AG and other private companies. DB Netze Stations' activities include the management of passenger stations, transport stops, as well as the development and marketing of train stations. DB Netze Track is a subsidiary of DB AG, which includes other companies:

DB Netz AG is responsible for railway infrastructure. Its activities include the ongoing maintenance and development of infrastructure, as well as cooperation with railway undertakings in the preparation of train schedules, maintenance and repair of the railway network.

DUSS GmbH is responsible for transshipment terminals owned by the infrastructure manager DB Netz AG and third parties.

DB Fahrwegdienste GmbH is responsible for risk management and logistics.

DB RegioNetz Infrastruktur GmbH is responsible for the organizational management of its railway lines and stations. It owns regional railways with some 13,000 railway tracks.

- Cargo transportation is provided by DB Cargo AG (with a Europe-wide freight presence with 16 subsidiaries in different countries) and DB Schenker (integrated logistics services)
- Passenger transportation is provided by DB Long-Distance (domestic and international transportation), DB Regional (domestic passenger services), and DB Arriva (international passenger services).
- DB Services is a service segment that combines six different sectors: DB Vehicle Maintenance (inspection and maintenance, modernization and train maintenance), DB Systel (IT and telecommunication service provider), DB Services (facility management and industrial services), DB Fleet Management, DB Communication Technology (maintenance and repair of IT and network technologies), and DB Security (security function).
- Bundesnetzagentur (BNA, a regulatory body for services such as electricity, gas, telecommunications and postal services), since 2006, became responsible for overseeing the rail market. The mission of BNA is to ensure nondiscriminatory access to rail infrastructure.

It is to be appreciated that the implementation of an integrated management model may facilitate the planning of infrastructure investments and railway operations in the country. In addition, it is possible to have a generally compatible tariff policy throughout the network and eliminate or minimize any contracts with other companies. This means that vertical integration helps optimize the entire system, what is quite hard to do with a vertically separated management model (Pittman, 2007). However, it is noticeable that the German model faces certain problems due to the close relationship between infrastructure management and service provision. In the case of Germany, the DB holding structure is quite confusing, making it difficult, for example, to trace cash flows between holding companies, which may lead to an under-enforced EU requirement for financial transparency (Nigrin, 2014).

It should be noted that as a form of integrated management a fully integrated model can be singled out. It is a traditional model of the organization of railway management in which a single public entity controls all or most of the infrastructure, as well as the freight and passenger services sectors, and performs administrative functions. Several countries are still using this mode—it is more common outside Europe, for example, clearly visible in China and India. There are also remaining countries in Europe that use fully integrated model. For example, Lithuania has a fully integrated management model where the infrastructure manager and the service provider are structural units of the same company. However, it was decided to reform the Lithuanian railway sector and move to a holding structure. This choice was determined by the fact that Lithuania is one of the few states with the only state carrier operating in the railway transport market. In order to comply with EU requirements, it was decided to reform the management of railways. In Lithuania it was decided to follow the example of Germany and to entrust the functions of infrastructure manager, freight, and passengers to the newly established subsidiaries of AB Lietuvos Geležinkeliai. In this way, AB LG Infrastruktūra will perform the function of the infrastructure manager, AB LG Cargo will carry the freight and AB LG Passengers will carry the passengers.

The main argument why full integration (state monopoly) can be beneficial is that the services provided, the condition of the railway infrastructure and rolling stock and other technical aspects are planned and organized together. Instead of distributing planning, operations to multiple organizations, all decisions can be made within a single company (Van de Velde et al., 2012). On the other hand, such a monolithic structure leads to a lack of competition, making it difficult for such a monopolistic company to improve the efficiency and productivity of the industry. In addition, as with most government-controlled industries, profitability goals are often complicated by set social goals (Finger and Künneke, 2011).

As mentioned earlier, the main goal of the reforms is to create a single railway market in the EU. According to the European Commission's Sixth report on monitoring development of the rail market, in 2016, new operators competing with incumbents were operating in all Member States except Greece, Ireland, Lithuania, and Luxembourg. This means that these countries probably did not properly implement the requirements of EU, as their markets are still closed. It should be noted that the European Commission monitors the transposition of the Directives by the Member States. In the event that a Member State is suspected of not transposing or transposing incorrectly the Directive, the European Commission shall send a letter of formal notice to the Member State. If the European Commission decides that the Member State is not fulfilling its obligations correctly, the European Commission shall send a reasoned opinion requesting the implementation of its obligations. If, in this case too, the State does not comply with the requirements of the European Commission, the European Commission has the right to apply to the Court of Justice. For example, in January 2019, the European Commission sent a letter of formal notice to Greece and Ireland requesting information on the transposition of Directive 2016/2370/EU (Fourth Railway Package). The Member States did not reply to the European Commission within the deadline of two months and were therefore given a reasoned opinion on July 25, 2019. If the Member States are found to have failed to comply with the requirements of the Directive, the European Commission will be entitled to apply to the Court of Justice.

Comparison of the Models: Advantages and Disadvantages

Europe has faced a number of challenges in implementing rail reforms. Problems had to be solved, such as encouraging changes in the major historical operators, improving the supply of rail transport, the creation of cooperation and competition between market players. The rail systems of the European countries are very different in terms of services provided and in terms of technical parameters. For example, freight transport is more developed in Eastern Europe, while Western Europe is more specialized in passenger services. Among other things, technical parameters such as track gauge, signaling systems are different in the countries and this makes it difficult to reconcile national railway systems. For these reasons, the EU is seeking coherence and harmonization between the different systems of Member States.

The European Member States have an opportunity to choose which management model they would like to implement. The Community of European Railway has expressed the view that it is unrealistic for a single model to be applicable to all European countries, so every country can choose the most appropriate management model (Bodiroga-Vukobrat et al., 2016).

The appropriate choice of model is influenced by national characteristics, such as the well-established management practices of the national railway undertaking, the need for rail services and the level of development of the railway sector. Reforms also take place depending on the separation of infrastructure from services provided and the level of privatization. For example, in countries such as the United States, Japan or New Zealand, companies remain vertically integrated. In Europe, the general trend is to separate infrastructure management from passenger and freight services (Gomez-Ibanez, 2004).

The primary objective pursued by the states in the implementation of their chosen railway management model is efficiency, that is, to maximize the potential available and to derive maximum benefit from it. A number of studies have been carried out to determine which of these models is more effective and could help to improve competition in the market, to reduce costs or to meet

the needs of customers using rail transport. As a result, the various studies in this field are quite difficult to interpret, since different criteria are used to evaluate efficiency. In addition, the same model is implemented differently in each country, so only the main features of the models can be singled out.

As a general rule, states begin to consider the implementation of a separated management model that could encourage private sector involvement in the provision of train services. However, this choice is perceived as complex and “bold” in the sense that it is necessary to define precisely the responsibilities and accountability between the railway infrastructure manager and the train service operators. States that consider the implementation of a separated management model often reject it as too complex or endangering some of the integration benefits, such as being closer to the end customer, coordinating public rail infrastructure, managing emergencies, etc. (World Bank, 2017). The choice of a separated management model poses greater challenges for the state as it loses control over the provision of services that could be maintained by an integrated management model.

However, in countries that have opted for a separated management model, can be singled out some prevailing trends and benefits of the model. From a competition standpoint, one of the major advantages of vertical separation is that it ensures a level playing field between different operators when no operator has more control over the infrastructure than others (Haylen and Butcher, 2017). A separated management model can be considered when it is seeking to protect nondiscriminatory access to rail infrastructure as the charging or capacity allocation body of an integrated railway undertaking seeks to favor its own carrier, while an independent infrastructure manager will normally seek commercial benefits and has an incentive to satisfy applications from both new and established carriers (Directorate General for Internal Policies, 2011). It is appreciating that the clearest way to avoid discrimination in access to rail infrastructure is the institutional separation of rail activities from rail infrastructure (Merkert et al., 2008). Thus it is noticeable that the development of competition is more established in countries that have opted for a separated management model than in countries where reforms have been limited. This confirms that a separate management model, combined with a strong regulatory body, can be an effective measure of ensuring nondiscriminatory access to rail network.

The advantages of a separated management model include greater separation between private and public interests, clear and unambiguous conditions for the use of railway infrastructure and economic clarity in the rail transport sector. An important consideration here is the ability to take greater account of customer needs, as business is better able to respond to market demand than monopolistic railway companies (Profillidis, 2006).

From the point of view of the provision of specific services, the transfer of services from the state to the private sector improves market and commercial performance, especially in the freight sector. However, privatization of the rail network has proved less attractive in countries where national railways have a strong passenger transportation base (World Bank, 2017). It is appreciated that the separation of passenger services from infrastructure management activities is less beneficial, as passenger services are generally heavily dependent on public funding (Directorate General for Internal Policies, 2011). Studies have even shown that a separated management model can significantly weaken railways in passenger transport sector rather than strengthen them (Laabsch and Sanner, 2012). This implies that a separated management model can be more useful and efficient in countries where rail transport is more used in freight transport. However, other studies and a broader analysis of the impact of vertical separation on the market have shown that there is no substantial link between vertical separation and rail freight or increased rail freight traffic overall. If the primary objective is to promote the efficiency and growth of rail freight, vertical separation may even undermine rail development in certain circumstances, especially in some Central and Eastern European countries that do not have adequate infrastructure funding (Drew and Nash, 2011). This could, for example, be the result of high infrastructure charges, which would prevent new railway undertakings from entering the market and launching their services, thus even a separated management model would not help to occur competition in the market.

The argument against a separated management model and the choice of an integrated model is that in some parts of the world (e.g., United States) railway may be the owner of the infrastructure and this in itself does not prevent other operators from entering the market after paying appropriate fees. Competition can exist without separation. But separation can be a catalyst for competition and for facilitating the entry of other rail operators (Profillidis, 2006).

One of the great advantages of integration is the cost reduction. For example, empirical studies of railway companies from 27 European countries between 2000 and 2004 found that integrated companies can achieve greater productivity through economies of scale, whereas with a separated management model, states may face increased costs (Growitsch and Wetzel, 2009). However, the integrated management model has received considerable criticism for restricting competition. For example, incumbent freight operators can exploit their market position and this may lead to conflicts of interest in areas such as access to terminals and other facilities, train path allocation, maintenance of rolling stock, etc. (The European Court of Auditors, 2016).

Politicians and rail administrations in various countries are therefore weighing the advantages and disadvantages of vertical separation to determine whether this type of reform is an appropriate way to improve the performance of their rail sector. Despite various structural and property reforms carried out around the world, there is no clear model for achieving this goal of increasing competition and efficiency (OECD, 2006). It is difficult to evaluate different models between countries, as the results may vary depending on different factors. However, it is important to understand that separation may be more beneficial in some industries than in others (Pittman, 2005). For example, it was found that full separation between infrastructure and services is not a prerequisite for improving rail efficiency (Friebel et al., 2003). Among other things, later studies concluded that railways are quite sensitive to reforms and that applying a single model to all countries is certainly not a good option seeking rail efficiency (Friebel et al., 2008). Consequently, the main step that all countries must take implementing the relevant model is to evaluate the priority areas for rail transport activities and define the results to be analyzed on the basis of examples from other countries.

Summary and Conclusions

Thus now it is possible to single out the prevailing tendency in the practice—the states are at the stage of implementation of the chosen management model. Following the example of these countries, it is possible to analyze how they are achieving their goals by implementing the management model chosen, what are the advantages and disadvantages of a particular management model. It should be noted that the fundamental objective of the reforms is to achieve the highest possible level of efficiency in the railway sector, therefore the assessment of the efficiency of the chosen model is based on calculating the costs incurred, analyzing changes in competition in the markets, etc. Although the various studies conducted do not provide a firm answer to the question which model is more effective, the main features and trends dictated by the implementation of the model in the selected countries can be singled out. A separated management model can help the state improve competition in the market, while an integrated management model may expose new market players to discriminatory conditions. Among other things, a separated management model enables the market itself to refine what services it needs, for example, freight services or long-distance passenger services. The integrated management model has more advantages in terms of control of public railway infrastructure and gives more guarantees to the users of passenger services.

See Also

Railway Management; Regulation of Rail Infrastructure and Services; The History of Rail Transport; The Geography of Rail Transport; The Future of Rail Transport

References

- Alexandersson, G., Hultén, S., 2006. Competitive tenders in passenger railway services: Looking into the theory and practice of different approaches in Europe. *European Transport \Trasporti Europei*, 33, 6–28.
- Alexandersson, G., Hultén, S., 2011. Reforming Railways: Learning from Experience, Community of European Railway and Infrastructure Companies, Brussels, 115–125.
- Bodiroga-Vukobrat, N., Rodin, S., Sander, G.G., 2016. *New Europe – Old Values? Reform and Perseverance*. Springer, Switzerland.
- Directorate General for Internal Policies. Policy department B: structural and cohesion policies, 2011. The impact of separation between infrastructure management and transport operations on the EU railway sector. European Parliament, Brussels
- Drew, J., Nash, C.A., 2011. Vertical Separation of Railway Infrastructure – Does It Always Make Sense? Institute for Transport Studies, University of Leeds, Working Paper 594.
- Gomez-Ibanez, J.A., 2004. Railroad reform: an overview of the options, paper presented at the Proceedings of the Railway Reform Conference, Madrid, 8 September.
- Gomez-Ibanez, J.A., 2006. Competition in the Railway Industry: An International Comparative Analysis. Transport Economics, Management, and Policy. Edward Elgar.
- Growitsch, C., Wetzel, H., 2009. Testing for economies of scope in European Railways: an efficiency analysis. *J. Trans. Econ. Policy* 43 (1), 1–24.
- Finger, M., Künneke, R.W., 2011. *International Handbook of Network Industries*, Books, Edward Elgar Publishing. United Kingdom of Great Britain and Northern Ireland.
- Friebel, G., Ivaldi M., Vibes, C., 2003. Railway (de) regulation: a European efficiency comparison, IDEI report no 3 on passenger rail transport, University of Toulouse.
- Friebel, G., Ivaldi, M., Vibes, C., 2008. Railway (de) regulation: a European efficiency comparison. *Economica* 75, 1–15.
- Haylen, A., Butcher, L., 2017. Rail Structures, Ownership and Reform. Briefing paper Number CBP 7992.
- House of Commons Transport Committee, 2012. The European Commission's 4th Railway Package. Twelfth Report of Session 2012—13, Vol. I.
- Laabsch, C., Sanner, H., 2012. The impact of vertical separation on the success of the railways, *Intereconomics* 47(2), 120–128.
- Merkert, N., Nash, C., Smith, A., 2008. Looking beyond separation – a comparative analysis of British, German and Swedish railways from a new institutional perspective. European Transport Conference, Institute for Transport Studies (ITS) University of Leeds, United Kingdom.
- Nigrin, T., 2014. Open competition or discrimination on tracks? Examples of anti-competitive behaviour of The Deutsche Bahn. *Rev. Econ. Perspect.* 14 (1), 1–18.
- OECD, 2006. Structural reform in the rail industry. *OECD J* 8(2), 67–175.
- Pittman, R., 2005. Structural separation to create competition? The case of freight railways. *Rev. Net. Econ.* 4, 181–196.
- Pittman, R., 2007. Options for restructuring the state-owned monopoly railway. *Res. Transport. Econ.* 20 (1), 179–198.
- Profillidis, V.A., 2006. *Railway Management and Engineering*, 3rd ed. Section of Transportation, Democritus Thrace University, Greece.
- The European Court of Auditors, 2016. Rail freight transport in the EU: still not on the right track, Special Report No. 08.
- Van de Velde, D., Nash, C., Smith, A., Mizutani, F., Uranishi, S., Lijesen, M., Zschoche, F. for CER, EVES-Rail, 2012. Economic effects of Vertical Separation in the Railway Sector. Full technical report for CER Community of European Railways and Infrastructure Companies; by inno - V (Amsterdam) in cooperation with University of Leeds - ITS, Kobe University, VU Amsterdam University and civity management consultants.
- White Paper - Roadmap to a single European transport area - Towards a competitive and resource-efficient transport system (COM(2011) 144 final of 28 March 2011).
- Williams Rail Review, 2019. Current railway models: Great Britain and overseas. Evidence paper.
- World Bank, 2017. *Railway Reform: Toolkit for Improving Rail Sector Performance*. World Bank Group, Washington, DC.

Further Reading

- Directive 2012/34/EU of the European Parliament and of the Council of 21 November 2012 establishing a single European railway area *OJ L 343, 14.12.2012, p. 32–77 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)*. *Special edition in Croatian: Chapter 07 Volume 025 P. 136 - 181*
- Directive 2016/2370/EU of the European Parliament and of the Council of 14 December 2016 amending Directive 2012/34/EU as regards the opening of the market for domestic passenger transport services by rail and the governance of the rail infrastructure. *OJ L 352, 23.12.2016, p. 1–17 (BG, ES, CS, DA, DE, ET, EL, EN, FR, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)*
- Domergue, P., Quinet, E., 2001. Situation and problems of railway industry in Europe. *JRTR* 26, 4–7.

Regulation of Rail Infrastructure and Services

Javier Campos, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction: Current Regulatory Scenarios in the Rail Industry	485
Relevant Regulatory Issues in the Rail Industry After the Reforms	486
When (and How) Should be Regulated Rail Services?	486
How to Guarantee Cost Recovery?	487
How to Promote Competition?	488
What About Quality Regulation?	488
Summary and Conclusion	489
See Also	489
References	489
Further Reading	489

Introduction: Current Regulatory Scenarios in the Rail Industry

As in other transport modes, the regulation of the provision of infrastructure and rail services is closely determined by the degree of vertical integration and the overall opportunities for competition “on the tracks” and “for the tracks” that are feasible in each particular market. However, some additional features, such as the multiproduct nature of this activity, the presence of indivisibilities in inputs and outputs, the role of rail transport within the public service network, or the existence of externalities, make regulatory tasks in this sector particularly interesting for the transport system as a whole (Bitzan, 2003).

With few exceptions around the world, from its birth in the mid-19th century to the last decades of the 20th century, the most widespread market structure in the railway sector has been that of a single state-owned company, responsible for the integrated management of both infrastructure and services. Although there were some differences in their degree of commercial autonomy, the regulation and control of these companies were relatively homogeneous: traditional price and service regulations were widely assumed to protect the general interest, and there was an explicit obligation on the part of the companies to meet any demand at those prices: closing existing lines or opening new services required government support. Thus, competition was perceived almost as an oxymoron and even discouraged, while the preservation of the public nature of industry was seen as the key regulatory principle.

Under this protective environment, these national companies progressively lost any connection with market forces and were characterized by increasing losses (usually financed by public subsidies) and business decisions mainly oriented toward production targets rather than commercial objectives. As a result, most railways were in a very weak position to face fierce intermodal competition in their provision of freight and passenger services, and it was clear that a profound restructuring process was necessary to guarantee the survival of the sector. The changes started in the 1980s and were aimed at improving the sector performance by introducing or strengthening competition. The most salient characteristic of the restructuring processes was the consolidation of alternative organizational structures for the industry which, in turn, defined new regulatory scenarios. These structures differed along two main dimensions represented in Fig. 1: the extent of vertical separation introduced after the changes and the amount of private participation allowed in the industry after the reforms.

With regard to the first dimension, the existence of substantial (and sub-additive) fixed costs, as well as the presence of notable indivisibilities in their operating assets, has led economists to consider the provision of rail infrastructure as one of the few remaining textbook examples of a natural monopoly. Unlike the provision of services—where different forms of competition are more feasible—regulation of rail infrastructure remains more necessary, specifically requiring efficient pricing rules, clear access rights, and sound investment criteria for its long-term development. Indeed, after the restructuring, each country addressed this relationship between infrastructure and services in a different way: many chose to keep infrastructure in public hands, creating dedicated agencies to regulate (public or private) train operators; others established nominally independent (but state-owned) companies to manage stations and tracks, whereas only a few completely privatized infrastructure and/or operations, with or without unbundling them first.

The vertical axis in Fig. 1 summarizes the three main options for the vertical restructuring of the railway industry and provides several examples in different countries. The first option (*vertical integration*) corresponds to the traditional rail model (which still exists, e.g., in India), where a single (public) operator controls all infrastructure facilities as well as the provision of freight and passenger services. The *competitive* (or open) *access* model is characterized by the coexistence of an infrastructure manager (usually in the hands of the government), which is required to make it available to other operators (including public firms) on a fair access basis. This model—common in several European countries with minor differences across them—retains the advantages of integration, but its overall effectiveness may be compromised in the absence of supervision by competition authorities to avoid discriminatory practices. Under the *complete vertical separation* model, the management (and, possibly, the ownership, as in the United Kingdom) of

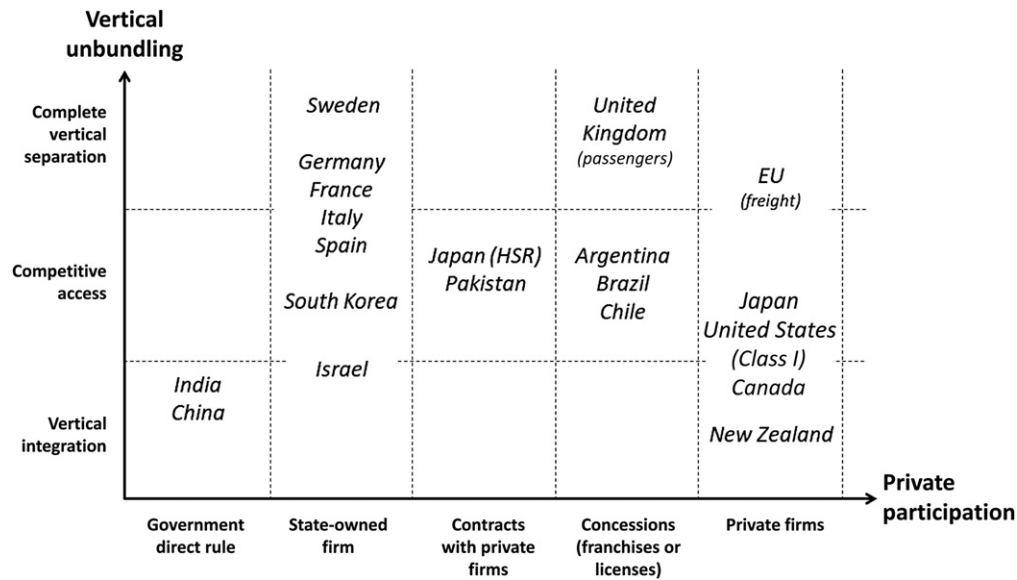


Figure 1 The rail restructuring process: examples. Source: Own elaboration, adapted from Campos (2008).

the infrastructure is fully separated from services, and the manager is not allowed to provide transport services. This option promotes the entry of private operators and competition “on the tracks” or “for the tracks” (through tendering mechanisms) may emerge. However, the unbundling may also entail a number of disadvantages for the rail industry. The main one is the loss of economies of scope that are obtained from the joint operation of infrastructure and services. There is also a greater risk that the unbundled network will be less attractive for the users as compared to alternative transport modes and it should be noted that vertical separation also requires more complex institutional arrangements, with higher transaction costs and, sometimes, less investment incentives, due to higher regulatory risks. In some countries, the potential conflict between sectoral regulators, competition authorities, infrastructure managers, and rail services operators is unescapable. Its resolution often requires a clear definition of the competences of each party and a reinforcement of the independence of the supervisory bodies.

Regarding the privatization dimension, the horizontal axis in Fig. 1 provides an orderly list of the main models found on railways around the world after their restructuring processes. The first one corresponds to *government’s direct rule*, where the industry is still controlled by one or more Cabinet departments (e.g., the Ministry of Railways in India). The second case refers to the existence of one or more *state-owned firms* in the sector (as in China, Israel, or several European countries), with some degree of commercial autonomy, but still requiring government approval for most investment and pricing policies, as in the traditional model. In some countries, this ownership model is often completed with other forms of public–private partnership, including *contracts* for outsourced activities, for the *management* of infrastructure or services, or even the *leasing* of specialized assets, such as locomotives or wagons. The use of tendering mechanisms, such as *concessions* (franchises or licenses), constitutes one of the most popular methods to introduce private participation in railways, particularly in some European and South-American countries, and, finally, *total privatization* is less frequent in this industry, although private operators have existed for many years in countries such as the United States or Japan.

After this introduction, next section provides a review of the most relevant regulatory issues after the reforms, particularly focusing on the need of regulation itself and how to achieve its objectives. This paper concludes with a brief summary and conclusion.

Relevant Regulatory Issues in the Rail Industry After the Reforms

The two-dimensional post-restructuring space presented in Fig. 1 defines a new institutional environment for the rail industry whose most relevant feature is the diversity of regulatory scenarios that have emerged in different countries around the world (see Nash et al. (2013) or Laurino et al. (2015) for details). Most of these scenarios, however, share a number of similar concerns about prices, cost recovery, access to infrastructure, and quality regulation.

When (and How) Should be Regulated Rail Services?

In countries where there exists private participation in freight or passenger markets, the nature of price regulation has changed as compared to the traditional rail model: instead of simple price-setting rules for rail services, regulation has progressively become a price-supervision process intended to avoid market dominance or other uncompetitive practices. This process is generally designed

according to two factors: the degree of market power effectively enjoyed by the operators and the existence of other limiting factors, such as intermodal competition. This latter element is relevant in freight services (intermodal competition from trucking) and high-speed services (in competition with air transport), but in the case of commuter and regional routes, social pressure to keep fares low may affect price interventions. In practice, the most common instruments for price control in these cases adopt the form of a rate of return regulation or a price cap mechanism.

Rate of return regulation is commonly used in Canada, Japan, and the United States. Its objective is to set prices so that the regulated operator only earns a fair rate of return on its capital investment. The regulator typically determines a revenue requirement based on the estimated total costs during a test year, and an estimate of the cost of capital for the firm, given by a “reasonable” rate level multiplied by an asset base. In spite of this simplicity, this type of price regulation is often criticized in railways on the grounds that it offers little efficiency incentives, since firms can pass production costs on to final users in the form of higher prices. In addition, it often leads to overinvestment in capital (Averch–Johnson effect).

As a price-control alternative, many European countries have opted for cost-plus incentives whose main aim is the achievement of dynamic efficiency by sharing its gains between the producer and the regulator. In its most standard form, it consists in setting a maximum price scheme based on long-run marginal costs in order to provide the train operator with an incentive to minimize costs by keeping all or part of the gains associated with the firm's increases in efficiency. This mechanism emerged as a consequence of the criticism directed at the lack of cost minimization incentives embedded in the rate of return regulation and other traditional price regulation mechanisms. However, its efficiency gains have to be balanced with the higher informational requirements that it implies.

There are a number of minor variations of the price cap system currently applied in the rail industry. One of the most widely used is the RPI-X formula. In this setup, the price for a basket of the firm's prices may increase in any one year by no more than the increase in the retail price index (RPI) for that year, minus some (efficiency related) X-parameter. In the case of multiproduct activities, such as railways services, this expression can be easily adapted by requiring that a certain weighted average of percentage price increases do not exceed the rate of growth of the RPI, less X percent. In the United Kingdom, for example, this mechanism has been applied to passenger traffic franchises. Commuter fares are regulated with respect to a basket containing all relevant fares, broadly weighed by the income that the operator derives from each of them.

How to Guarantee Cost Recovery?

A second regulatory issue is related to the recovering of infrastructure costs. Apart from a high ratio of fixed to marginal costs, the provision and operation of tracks, stations, and other rail infrastructure is characterized by the existence of avoidable and unavoidable costs. Avoidable costs are uniquely associated with a particular output: if it is not produced, no cost is incurred; therefore, avoidable costs may be considered as a floor to regulated prices (if any), since charging less than these costs is equivalent to operating at an economic loss. This makes standard pricing rules inoperable in this sector, since first-best or marginal cost pricing rules result in large deficits that jeopardize the long-run survival of the operator.

Another complexity factor is the existence of cross-subsidization problems when common costs are relevant. This can be illustrated with the case of a profit-regulated railroad connecting two large cities and also providing rail services to a smaller town along the route between the two cities. The fares charged from the small town generate revenues exceeding the additional cost of serving it, such as ticketing and station costs, but not enough to cover an equal or proportionate (however, defined) share of the common costs, such as trackage, signaling or trainyard costs. The issue is how to allocate common costs among customers and services. In many cases, cost sub-additivity and efficiency require joint production and allocation of fixed costs among all services, without cross-subsidization (accounting for externalities whenever present).

This is not only an equity problem, as in this example, but also a relevant issue for efficient pricing of infrastructure like railbeds, signals, or stations. The standard procedure in these cases is to rely on the so-called fully distributed costs method, under which common costs are allocated on the basis of some common measure of utilization, such as gross tons/km, or other measure of relative output or gross revenue. Alternatively, common costs can also be allocated in proportion to costs that can be directly assigned to the various services. The arbitrary nature of fully distributed cost methods and its lack of a conceptual foundation have been criticized, but they remain a useful measure for recovering common costs.

However, the treatment of the cross-subsidization problem should not be based on excessively rigid criteria, particularly in markets with few alternative finance mechanisms. The analysis should be made on a case-by-case basis, since, for example, stand-alone cost tests do not apply if railroads are not allowed to abandon nonprofitable facilities or services. If that freedom is denied, a railroad cannot earn adequate revenues if its rates on potentially remunerative activities are constrained by stand-alone cost ceilings.

The cost recovery principle should be a central issue in the design of any rail infrastructure pricing procedure. The theoretical and political debate focuses on two options. Many public firms still advocate the use of efficient pricing mechanisms and propose marginal cost rules with the simultaneous use of public subsidies to cover fixed costs. Alternatively, a growing literature suggests the use of full cost recovery prices, including price discrimination, multiple part tariffs, or cross-subsidization schemes, if needed. Although it is thought that it might yield inefficient outcomes, it constitutes the second-best available alternative in most cases.

With respect to infrastructure pricing, it is clear that in principle, track access charges should be based on marginal cost pricing rules. In practice, however, the use of these criteria is difficult due to at least three reasons: the cost structure of the rail network, which cannot always be recovered with simple pricing rules; the asymmetric information problem faced by the regulator with respect to

these costs; and the subsidy levels that can be sustained in the long run. Several econometric studies have shown that the marginal cost of rail operators that are still vertically integrated lies in the range of 60%–70% of average cost, whereas when services are separated from infrastructure, the marginal social cost of rail infrastructure alone often is well below the same range. Price discrimination, if feasible and politically acceptable, may help to raise cost recovery to around 60% of total cost without driving demand off the market. Thus, full cost recovery would require a further price markup of more than 60% above the efficient price.

Alternative proposals, in terms of the so-called Ramsey pricing principle, have been defended for infrastructures with high fixed costs and low marginal costs. However, they rarely work in practice, since they increase users' suspicions of unfair treatment and undue discrimination. Moreover, under Ramsey pricing rules, all unattributable fixed and common costs are apportioned on the basis of the services' demand characteristics. For these reasons, a reasonable conclusion is to advocate a balance between cost-recovery and efficient pricing rules, giving preferential treatment to one over the other according to the case. The issue remains unsolved and depends on how different countries have faced their access pricing problem.

How to Promote Competition?

Apart from cost recovery issues, access pricing may create additional regulatory concerns in vertically unbundled scenarios or under competitive or open access. The problem is common in network industries where a single, vertically integrated firm (either private or public) controls the supply of an essential facility (in this case, rail tracks or stations), restricting "downstream" competition. It is obvious that in these cases there are incentives for infrastructure manager to set prices high to raise rivals' costs, but it could also be the case in which the regulator sets access prices too low in order to favor the entrants.

Depending on the discretion allowed to the infrastructure manager, potentially distortive effects on access prices can be determined in several ways. Firstly, when infrastructure is still state-owned, the regulator can determine the price as an integral part of the access terms defined in a contract with one of several private train operators. Secondly, the regulator may allow the operators to choose from a menu of alternative regulatory schemes, usually related to incentive-based regulation mechanisms. Thirdly, the firm may have discretion over aspects of access pricing subject to some overall regulatory constraint, and, finally, the firm may have full discretion over the price and is only restricted by the competition authorities.

In all of these cases, there are two main approaches to setting access prices when the principles of cost recovery plus the normal rate of return are required. Firstly, some countries use cost-related charges, which are based on the optimal first-best principle of pricing according to marginal cost (considered the forward-looking long-run incremental cost). The higher the proportion of common costs, the more complex the principle. It is based on the so-called efficient-component rule, which determines that optimal access charge is equal to the direct cost plus the opportunity cost of providing access (given by the reduction in the dominant firm's profit). To compute these costs, the regulator has to consider economic depreciation (physical depreciation plus technological progress) and forecast future usage. The first problem to be solved is that of the actual value of capital assets: nominal value versus potential to generate cash. While the latter is clearly a function of the privatization and regulation methods and the extent of competition envisaged in bidding for the right to operate concessional infrastructure services, the former is more likely to reflect a past situation that domestic reforms are trying to overcome.

The second method of setting access prices consists of developing usage-related charges. Once-avoidable costs are covered by increasing prices that are inversely related to demand elasticity. Another option (less controversial) is the use of a two-part tariff to avoid service cuts by train operators to save charges even when the network has no cost saving. In any case, excessive infrastructure access charges may discourage new railway operators from entering the market. European Union rules, for example, require track access charges to be set on the basis of direct costs, namely the cost directly incurred as a result of operating a train service.

What About Quality Regulation?

Quality (including safety) performance is not neutral for the economic contribution of the rail transport sector to the social welfare. The particular level of quality achieved by train operators and particular features in regard to three main dimensions that broadly define quality in the rail industry (service, externalities, and investment) critically determine the value added by this transport mode. The first question that naturally arises is why quality regulation is needed at all in this industry, and to what extent this regulation relates to the standard price regulation mechanisms described in the previous section. Economic theory provides a well-known argument to answer these questions: real-world transport activities are characterized by market failures due to information problems. Ideally, with a large number of competitive rail transport operators and well-informed users, quality regulation requirements would be minimal, and low quality and less reliable rail companies would be driven out the market. However, since full information does not exist, markets cannot exert this disciplinary role and competition may result in unsafe, unreliable, or unpleasant services. In the traditional model of the rail industry some years ago, price-quality adjustment problems were frequent, since the monopoly's optimal level of quality may not have coincided with social standards. Simple price regulation is seldom a solution. Any regulated, multiproduct monopolist in an environment of asymmetric information tends to degrade quality in order to achieve higher profits once it enters the market. Railway firms are not immune to this temptation, for example, in terms of punctuality and cancellation standards. The quality outcome of any monopolist, not only in the rail sector, heavily depends on the specific regulation mechanism. For example, with rate of return regulation, overinvesting in non-required technological quality may accentuate the Averch–Johnson effect. Alternatively, with price caps, a subtle quality reduction can be a very tempting way to cut costs. Therefore, the price regulation mechanisms analyzed earlier are considered incomplete if they do not include quality provisions. This is not always easy, since

adjusting price mechanisms by quality may render them inoperative or excessively difficult for the operators to manage or the regulator to monitor. Therefore, most regulators set quality standards or targets for train operators instead of correcting price control mechanisms. The use of yardstick competition or regulation by comparison is common in many countries (such as in Australia).

Summary and Conclusion

Declining traffic volumes, increasing intermodal competition and historical inefficiencies of the sector have entailed the need for rail market reforms over the last decades. With the aim of guaranteeing the long-term survival of the sector by making it face the forces of competition, the rail restructuring process is still very much in progress in many countries. Two key dimensions—perceived as major obstacles before the changes—have been addressed: the vertical integration of the industry, where only infrastructure may now retain a justifiable natural monopoly role, and excessive public sector intervention, which has been reduced through the progressive introduction of private rail operators. In this new context, the main roles of railways regulation should be to focus on setting charges level and structure to ensure cost recovery, manage conflicts, and avoid unfair practices over infrastructure access and, in general, to create a level playing field which ensures fair competition “on the tracks” and “for the tracks.”

See Also

Natural Monopoly in Transport; Principles of Pricing in the Transport Sector; Pricing Versus Regulation; Privatization and Deregulation of Public Transport; How to Finance Transport Infrastructure? Marginal Cost Versus Cost Coverage. Investment and Pricing of Infrastructure: Financing of Transport Infrastructure with Negative, Positive, or Zero Scale Economies; Deregulation and Vertical Separation in the Railway Sector; Rail Cost Functions; Vertical and Horizontal Separation in the European Railway Sector and its Effects on Productivity; Introduction to Rail Transport

References

- Bitzan, J.D., 2003. Railroad costs and competition: the implications of introducing competition to railroad networks. *J. Transp. Econ. Policy* 37, 201–225.
- Campos, J., 2008. Recent changes in the Spanish rail model: the role of competition. *Rev. Netw. Econ.* 7 (1), 1–17.
- Laurino, A., Ramella, F., Beria, P., 2015. The economic regulation of railway networks: a worldwide survey. *Transp. Res. Part A Policy Pract.* 77, 202–212.
- Nash, C.A., Nilsson, J.E., Link, H., 2013. Comparing three models for introduction of competition into railways. *J. Transp. Econ. Policy* 47 (2), 191–206.

Further Reading

- Beria, P., Quinet, E., de Rus, G., Schulz, C., 2012. Comparison of rail liberalisation levels across four European countries. *Res. Transp. Econ.* 36 (1), 110–120.
- Friebel, G., Ivaldi, M., Vibes, C., 2010. Railway (de)regulation: a European efficiency comparison. *Economica* 77 (305), 77–91.
- Gómez-Ibáñez, J.A., de Rus, G. (Eds.), 2006. *Competition in the Railway Industry: An International Comparative Analysis*. Edward-Elgar Publishing, Cheltenham.
- Nash, C., 2008. Passenger railway reform in the last 20 years—European experience reconsidered. *Res. Transp. Econ.* 22 (1), 61–70.
- OECD, 2013. *Recent Developments in Rail Transportation Services*. OECD Policy Roundtables. Organisation for Economic Co-operation and Development, Paris.

Rail Vehicle Classification

Christos Pyrgidis, Alexandros Dolianitis, Aristotle University of Thessaloniki, Thessaloniki, Greece

© 2021 Elsevier Ltd. All rights reserved.

Introduction	490
Individual Rail Vehicle Classification	491
Elementary Classification	491
Secondary Classification	491
Power Vehicles	492
Trailer Passenger Vehicles	493
Trailer Freight Vehicles	493
Engineering Vehicles	494
Classifications by International Union of Railways (UIC)	495
Definition and Classification of Trainsets	495
Railway Transport Systems Classification	497
Elementary Classification of Railway Systems	497
Secondary Classification of Railway Systems	497
Typical Rolling Stock Characteristics per Railway System	499
Summary and Conclusions	499
Annex	501
References	506
Further Reading	507

Introduction

The term “railway” or “rail” encompasses terrestrial mass transport systems, on which trains move, either on their own or with the use of remotely transmitted power, on a dedicated steel way corridor (for a small number of metro lines and for many cases of driverless railway systems (cable propelled and self-propelled) of low/medium transport capacity, rubber-tyred wheels are also used.) defined by two parallel rails (Pyrgidis, 2016). From a transport system point of view, it is by default considered to be comprised of three (3) constituents:

1. railway infrastructure
2. rolling stock
3. railway operation

The term “rolling stock” refers to all rail vehicles moving on the dedicated corridor of the system and which either directly or indirectly facilitate railway transportation (Pyrgidis, 2016).

This article aims to classify rail vehicles based on three levels of integration within a wider railway system. Specifically:

- as an individual vehicle
- as part of a trainset
- as rolling stock for a specific category of railway system

This approach was deemed appropriate, since the choice of technical and operational characteristics of vehicles is often made on the basis of the trainset of which they are a part or of the system in which they are called to operate.

Alternatively, considering the classification of “rolling stock” (as previously defined) rather than a narrowest sense of the “rail vehicle,” the following three criteria for rolling stock high-level classification could be adopted:

- Based on function, namely power, trailer, and self-propelled (railcars and trainsets) rolling stock.
- Based on type of use, namely passenger rolling stock, freight rolling stock, and engineering vehicles.
- Based on the type of services for which the rolling stock is made, namely intercity trains (high or conventional speed), suburban trains, urban trains, rolling stock for steep gradients, and freight trains.

The aforementioned categories of rolling stock could be further classified on the basis of their technical and operational characteristics at a secondary level, in manner similar to the one proposed for categorization adopted for rail vehicles.

Based on the above, this article is structured as follows. Individual Rail Vehicle Classification Section provides an overview of individual rail vehicle classifications. Specifically, Elementary Classification Section provides a high-level or elementary classification, Secondary Classification Section provides a secondary level classification, and Classifications by International Union of

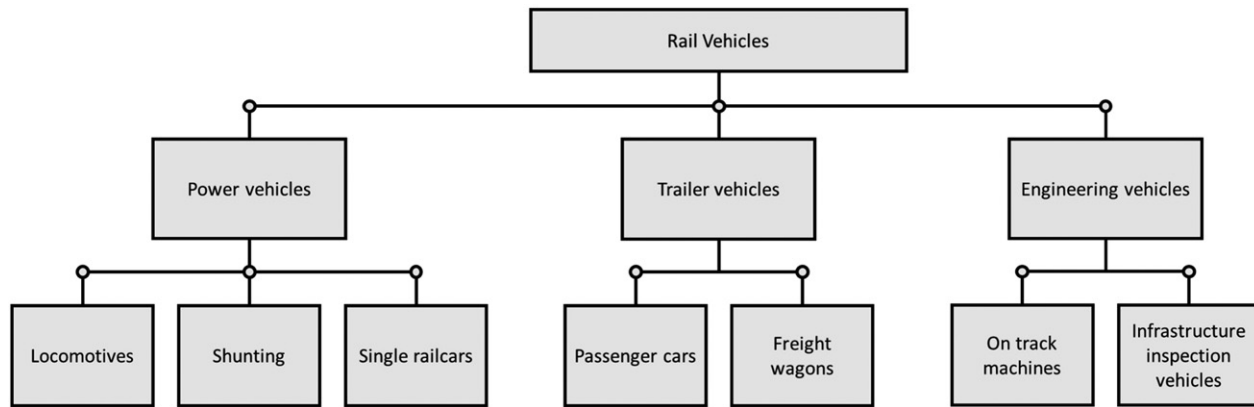


Figure 1 Rail vehicle elementary classification. Source: Modified from Pyrgidis, C., 2016. *Railway Transportation Systems - Design, Construction, and Operation*. CRC Press, Boca Raton.

Railways (UIC) Section presents a classification as proposed by the International Union of Railways. Definition and Classification of Trainsets Section provides a classification for “trainsets”. Railway Transport Systems Classification Section includes, in 1 Elementary Classification of Railway Systems Section, a high-level classification of railway systems and in Secondary Classification of Railway Systems Section a lower level classification. Finally, this article concludes, in Typical Rolling Stock Characteristics per Railway System Section, by summarizing typical rolling stock characteristics for each of the identified railway transport systems. Supporting figures depicting identified categories of individual railway vehicles, trainsets, and railway systems are included in the form of an Annex.

Individual Rail Vehicle Classification

Elementary Classification

At the highest level rail vehicles are distinguished based on their main function into powered and trailer vehicles (Fig. 1).

Power vehicles are self-propelled (i.e., they are equipped with traction motors). Such vehicles may:

- serve the sole purpose of hauling the trailer vehicles and are then called “locomotives” (or traction units, engines)
- transport a number of passengers and are then called either “single railcars” or “railbus” (when they have a driver’s cab at one or both ends) or “motor cars” (when they are remote controlled from another vehicle).
- be used for shunting hence they are called “shunting locomotives,” “shunters,” or “switchers”

Trailer vehicles are not self-propelled. They serve the purpose of transporting passengers or goods. Such vehicles may be classified into three (3) basic categories depending on their type of use (Metzler, 1981).

- Passenger vehicles (or passenger cars or coaches or carriages), which are intended to transport passengers.
- General—use freight vehicles (or freight cars or goods wagons or trucks), which are intended to transport goods.
- Specific—use freight wagons, which are intended for the transportation of certain types of freight only.

The term “rail vehicles” also encompasses the various engineering vehicles that are used to carry out track panel installation works as well as the various track inspection and maintenance works. These types of vehicles, due to their special function, may on their own be considered as a high-level special category (Fig. 1). These vehicles are further distinguished into two main categories, specifically into (Biasin et al., 2008):

- On track machines (OTMs), which include vehicles specifically designed for the construction and maintenance of the track. OTMs may be used in different modes, namely working mode and transport mode (either as a self-propelled vehicle or as a hauled vehicle).
- Infrastructure inspection vehicles or track recording or diagnostic cars, which are utilized to monitor the condition of the infrastructure. For such vehicles, there is no distinction between transport and working modes, instead they are operated in the same way as freight or passenger vehicles.

Secondary Classification

This section provides secondary level classifications of individual rail vehicles. Specifically, a secondary level classification is provided for power vehicles, trailer passenger vehicles, trailer freight vehicles, and engineering vehicles based on their technical and operational features.

Every rail vehicle, whether powered or trailer, consists of three basic parts as shown in Fig. 2. These are the car body or body shell, the bogies or trucks, and the wheelsets.

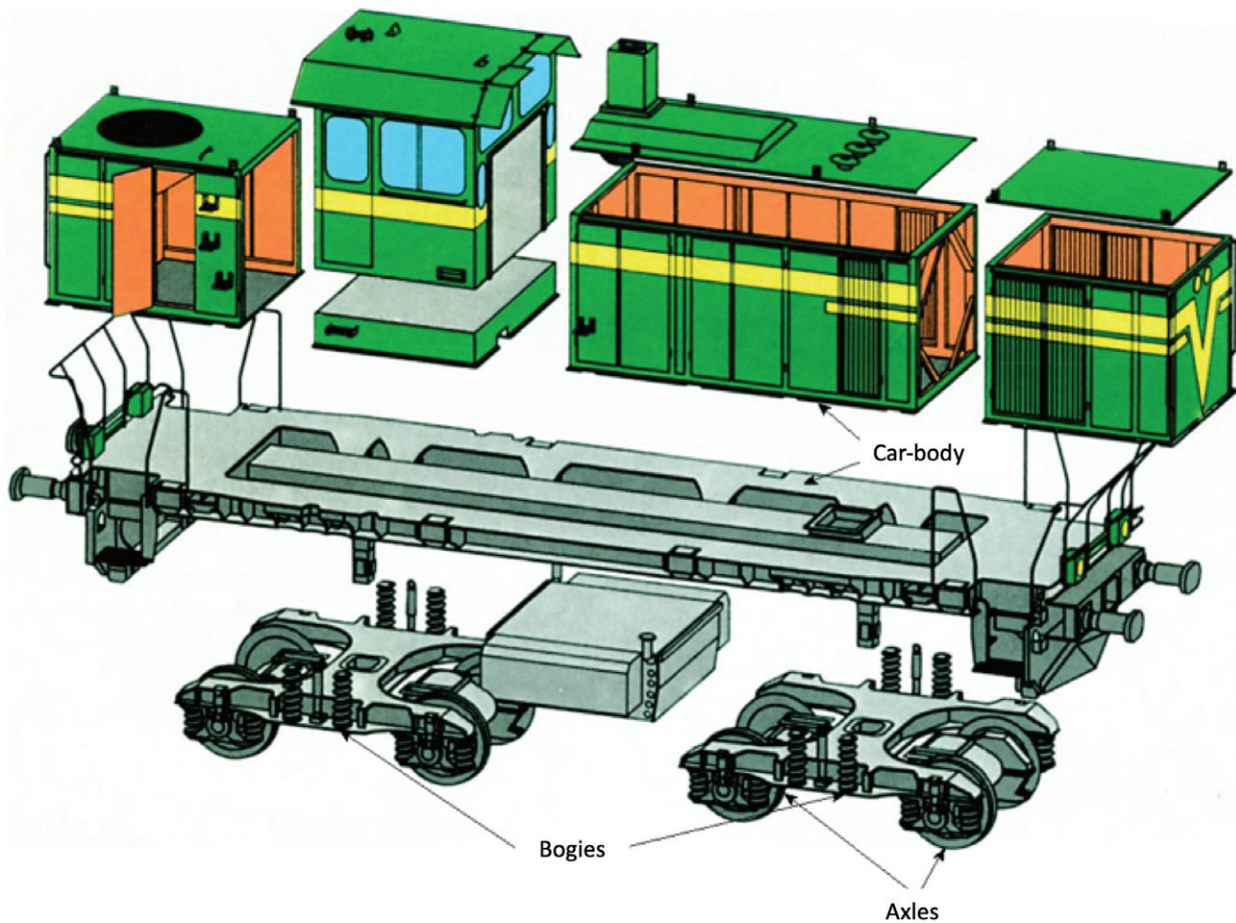


Figure 2 Main parts of a rail vehicle (Diesel locomotive). Source: Courtesy Siemens; Pyrgidis, C., 2016. *Railway Transportation Systems - Design, Construction, and Operation*. CRC Press, Boca Raton.

Several specific characteristics of these parts may be used to classify rail vehicles further. Specifically, the width of the axles may be used to classify rail vehicles based on the track gauge on which they may travel. Based on this criterion they may be classified into: (1) broad gauge vehicles, (2) standard gauge vehicles, (3) meter gauge vehicles, and (4) narrow gauge vehicles.

Moreover, the construction material of the car body may be used to classify modern rail vehicles into those possessing: (1) aluminum shells, (2) steel shells, or (3) stainless steel shells.

Power Vehicles

Locomotives depending on the type of traction power they utilize are classified into: (1) steam locomotives, (2) diesel locomotives, (3) electric locomotives, (4) hybrid locomotives, (5) gas turbine locomotives, and (6) fuel cell locomotives.

All these categories may be further classified based on a series of criteria specific to each category. For instance, steam locomotives may be distinguished based on the type of utilized steam (liquid/saturated or dry/superheated), their transmission method and whether the tender is part of or separated from the locomotive. Similarly, **Fig. 3** depicts a lower level classification of traction units based on power distribution. Diesel and electric locomotives are further classified based on the number of bogies and axles they have, as follows:

- BB (Locomotive with two twin-axle bogies)
- CC (Locomotive with two triple-axle bogies)
- CCC (Locomotive with three triple-axle bogies)

Electric locomotives are further classified based on (Frey, 2012; Sheed and Allen, n.d.; Railsystem, n.d.):

- The operating voltage/current under which they are designed to operate: AC locomotives, DC locomotives, multirole locomotives.
- The type of electric motor used: three-phase asynchronous motors, DC brush motors, synchronous AC motors.
- The solution used for power control and transmission between the electric motor and the bogies: mechanic, reostatic with series/parallel motor commuting and shunting, chopper, GTO/IGBT.

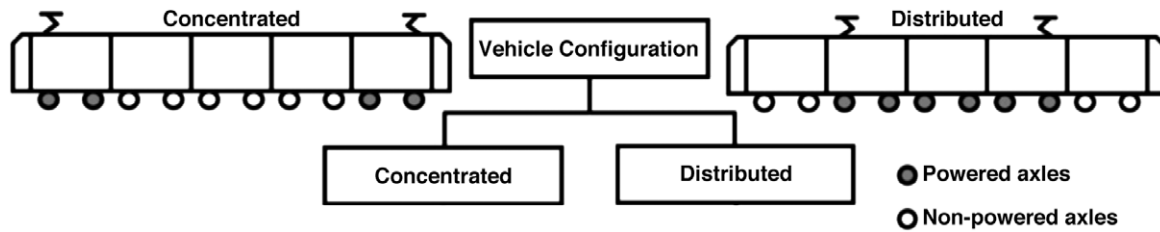


Figure 3 Classification of rail vehicles based on power distribution. Source: Modified from Ronanki, D., Singh, S., Williamson, S., 2017. Comprehensive topological overview of rolling stock architectures and recent trends in electric railway traction systems. *IEEE Trans. Transp. Electr.* 3(3), 724–738.

Hybrid powered vehicles utilize on-board rechargeable energy storage systems placed between the power-source and the transmission (Konarzewski et al., 2013). They may be classified based on:

- The nature of the power source, with the majority of hybrid vehicles utilizing an energy storage system in conjunction with Diesel, while other types of power source also exist (e.g., electric) (Hyatt, 2018).
- The type of storage system, for instance whether they use Ni-MH or Li-On batteries (Jeong et al. 2011).
- Whether the storage system is placed in series or in parallel.

Diesel locomotives can be further classified based on the solution adopted to transfer power from the prime engine (diesel generator) to the wheels. Thus, they are classified into:

- Diesel-mechanical locomotives, in which the power generated by the diesel engine is transferred to the wheels through a mechanical powertrain. This solution is usually adopted in vehicles with limited power (usually <300 kW).
- Diesel-hydraulic locomotives, in which the power generated by the diesel engine is transferred to the wheels through a torque converter. This solution is usually adopted in vehicles with power of middle-range (<1500 kW) and is very common in railcars.
- Diesel-electric locomotives, in which the power generated by the diesel engine is used to generate electric power, which is then transferred to electric motors in a way similar to that in electric locomotives. This solution is as of today the most widely adopted for mainline locomotives, and also the one that allows highest power, even exceeding 2000 kW.

Other types of power vehicles, such as railcars, may distinguished into types in a similar manner as presented for locomotives.

Trailer Passenger Vehicles

Classification of trailer passenger vehicles may be based on: (1) their internal arrangement, (2) the number of decks, (3) the maximum allowed running speed ($V_{\max}$), (4) the floor height, (5) the level of service provided, (6) their specific use, (7) the technology of equipped bogies/axles, and (8) the capability or not for tilting. Table 1 provides a list of the various types of trailer passenger vehicles based on each of the aforementioned classification criteria.

Trailer Freight Vehicles

General—use freight vehicles may be classified based on the type of covering and height of sides, into:

- covered wagons (or boxcars)
- open high sided wagons (or gondolas)
- open low sided wagons (or gondolas)
- open vehicles with no sides (or flat cars)

Specific—use freight wagons, may be classified based on the nature of products they transport into (indicatively):

- tank cars (for petroleum products, chemical or food products, for liquefied gas or cryogenic transport)
- refrigerated wagons
- hopper cars (with central, bilateral or pneumatic unloading)
- road vehicles carriers (double deck or single deck)
- cars for road/rail combined traffic (piggyback cars)
- cement wagons
- special flat cars for containers and other specific goods
- center partition cars

Table 1 Classification of trailer passenger vehicles based on various criteria

<i>Classification criteria</i>	<i>Trailer passenger vehicle type</i>
Internal arrangement	• Vehicles with compartments • Vehicles with individual passenger seats
Number of decks	• Single deck • Double deck
Maximum running speed	• Conventional speed vehicles ($V_{\max} < 200 \text{ km}\cdot\text{h}^{-1}$) • High speed vehicles ($V_{\max} \geq 200 \text{ km}\cdot\text{h}^{-1}$)
Floor height	• High floor • Floor at platform level • Combined solutions
Level of service	• Exclusively A Class vehicles • Exclusively B Class vehicles • A and B Class vehicles (C Class vehicles have typically been discontinued)
Specific use (indicatively)	• Passenger coach • Sleeping car • Dining car • Lounge car • Luggage car • Observation car • Specialized types (hospital car, driving trailer, dome car, private car, etc.)
Technology of equipped bogies/axles	• Vehicles with conventional bogies • Vehicles with Jakobs-type bogies • Vehicles with bogies with independently rotating wheels • Vehicles with self-steering axles • Vehicles with mixed behavior bogies (conventional and independently rotating wheels)
Tilting capability	• Active tilting vehicles • Passive tilting vehicles • Nontilting vehicles
Several criteria are applicable to powered vehicles as well	

Engineering Vehicles

As already stated in Elementary Classification Section, engineering vehicles are mainly distinguished into OTMs and infrastructure inspection vehicles. Within the first category, the following vehicle types may be identified based on the function they fulfil:

- tamping machines
- stabilizing machines
- grinding machines
- track renewal and track laying machines
- ballast regulating machines
- ballast cleaning machines
- drainage system cleaning machines
- formation rehabilitation machines
- catenary maintenance machines
- multifunction machines
- staff vehicles (facilities, offices, etc.)
- vehicles carrying cranes/lifts/telescopic ladders

While infrastructure inspection vehicles are initially distinguished into self-propelled, towed, or hi-rail vehicles, while further categorization is based on the measurement or imaging systems with which they are equipped. Based on said equipment, they may measure (ENSCO, 2015; Railway Technology, 2017) track geometry, rail profile and rail wear, equivalent conicity, third rail geometry, gauge restraint, ride quality, rail surface defects (corrugation), vertical axle box acceleration, tunnel clearance, ballast profile, track center distance, catenary parameters (voltage, pantograph pressure, wire position).

Table 2 Trailer passenger vehicles as classified by UIC

Type of vehicle	Code	Comment
Passenger coach with 1st class seating	A	
Passenger coach with 2nd class seating	B	
Composite coach with 1st and 2nd class seating	AB	
Coach with 1st class seating, kitchen and dining area	AR	
Coach with 2nd class seating, kitchen and dining area	BR	
Composite coach with 1st or 2nd class seating and a luggage compartment	AD, BD	
Luggage van	D	
Composite coach with 2nd class and a mail section	BPost	
Luggage van with a mail section	DPost	
Club car	SR	
Saloon coach	Salon	
Restaurant car	WR	
Sleeper	WL	(only in combination with A or B)
Narrow gauge coach	K	(only in combination with A, B or D)

Source: Adapted from Leo, n.d. *Types of coaches (in German)*. <https://web.archive.org/web/20081228112927/http://www.leo.org/information/freizeit/bahn/gattungszeichen.html>.

Classifications by International Union of Railways (UIC)

This section provides classifications for trailer vehicles (both passenger and freight) as proposed by UIC. **Table 2** provides a sample of their classification for trailer passenger vehicles, while **Table 3** does so for trailer freight vehicles. Individual rail vehicles need to be uniquely identifiable (Hamm, n.d.), but also need to be registered in a way that clearly communicates such information across various stakeholders. Therefore, codes and categorizations like the ones provided in this section are of use, when vehicles are marked and numbered according to national or international rules.

Definition and Classification of Trainsets

A “train” or “trainset” or “consist” is an operational formation consisting of one or more units, while a unit may in turn be composed of several vehicles. A trainset may be either of a fixed or predefined formation. A fixed formation is by definition not intended to be reconfigured, except within a workshop environment. It may be composed of only motorized or motorized and nonmotorized vehicles. A predefined formation consists of several vehicles coupled together, in a manner which is defined at the design stage, and can be reconfigured during operation (Biasin et al., 2008).

The combination of locomotives and trailer vehicles forms the loco—hailed passenger or freight trains. When two traction units are included in the same train formation then the operation is called “2-loco operation” (Pyrgidis, 2016).

The combination of single railcars, motor cars, and/or trailer vehicles forms railcars. The railcars can move in both directions without the need of a shunting locomotive in contrast with loco-hauled trains which need a shunting locomotive (Pyrgidis, 2016).

Push–pull trains are trainsets with a locomotive at the front (pull–push) or at the rear (push–pull), in which a trailer vehicle at the rear or at the front respectively is equipped with a driving cab (also called control car, driving trailer) allowing the train to be driven from either end. It may have a number of intermediate trailer vehicles, and can move in both directions without the need for a shunting vehicle. **Fig. 4** depicts both a push–pull and pull–push train formations.

The term “multiple unit (MU),” either electric (EMU) or diesel (DMU), refers to a trainset in which all units are capable of carrying a payload (passengers, luggage/mail, or freight) (Biasin et al., 2008). Specifically, they exhibit the following characteristics (Pyrgidis, 2016):

- Units are made of single railcars, motor cars and/or trailer vehicles semipermanently coupled.
- Driving cabs are provided at each end of the unit. The driver simply changes ends at terminus.
- Train length can be modified by adding or subtracting units.
- Power equipment is distributed along the whole train (only motor cars and single railcars have power equipment).

Assuming that PR refers to a single railcar, TC refers to a trailer vehicle and MC refers to a motorcar, example formations of multiple units would be: PR + TC + PR + PR + TC + PR, PR + MC + MC + PR, PR + PR + PR + PR.

Table 3 Freight trailer vehicles as classified by UIC

Type	Class	Description	Further classification
Ordinary open high-sided wagons	E	These have at least 85 cm high walls, side door and do not possess self-discharging equipment. They are mainly used for the transportation of coal, scrap, wood, or paper	• 2-axle vehicles • 4-axle vehicles
Special open high-sided wagons	F	These are mainly self-discharging hoppers, which use gravity based unloading. Typical such wagons are loaded with bulk goods, such as coal, ore, sand or gravel	• Hopper wagons • Saddle-bottomed wagons • Side-tipping wagons • Modalohr road trailer carriers
Ordinary covered wagons	G	These have a fixed roof and fixed rigid walls with sliding doors on each side. There are closable openings of various types along the upper third of the side walls. This class of wagons have been largely superseded by other classes and they are mainly used for the transportation of part-load goods or parcels	• 2-axle vehicles • 4-axle vehicles
Special covered wagons	H	These types of wagons are often distinguished by their large loading volumes	• Livestock wagons • Wagons with end doors • Leig-Einheit units • Ferry wagons • Sliding-wall wagons
Refrigerated vans or refrigerated wagons	I	Insulated covered vehicles, which are either cooled by a cooling medium such as water or dry ice like conventional refrigerated vans, or are machine-cooled wagons with their own cooling system	• 2-axle refrigerated vans • 4-axle refrigerated vans
Ordinary flat wagons with separated axles	K	Vehicles with no walls or low walls that are not higher than 60 cm	
Special flat wagons with separate axles	L		
Open multi-purpose wagons	O		
Ordinary flat wagons with bogies	R		
Special flat wagons with bogies	S		
Wagons with opening roof	T	This class of vehicles is mainly used for the transport of hygroscopic bulk commodities such as cement, plaster, lime, potash, and grain	• Lidded wagons with level floors • Sliding-roof wagons and sliding-roof/sliding-side wagons • Swing-roof and rolling roof wagons with level floors
Special wagons	U		• Bulk goods wagons for transporting powders, etc. • Dual coupling wagons for joining wagons with different coupling systems • Well wagons including Schnabel wagons • Self-discharging hoppers with loading hatches • Trials vehicles for RoadRailer and Kombirail systems for intermodal transport
Tank wagons	Z		• Tank cars to carry pressurized gases • Milk cars • Liquid hydrogen tanks • Pickle cars • Tank containers • Torpedo wagons or bottle cars • Vinegar cars • “Whale belly” cars

Source: Based on Schiffer, V., 2005. Generic types of freight cars (in German). <http://home.wtal.de/gueterwagen/gatt-zde.htm>.

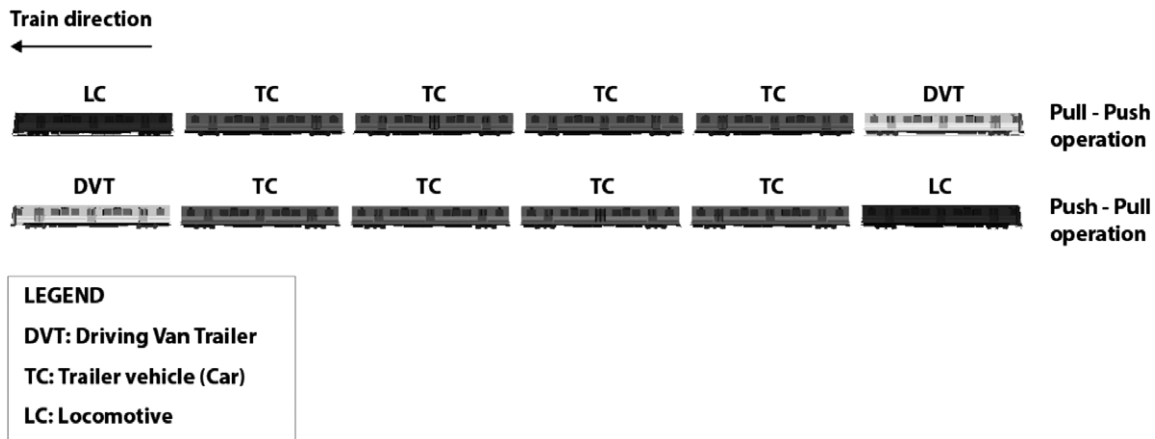


Figure 4 Push-pull/pull push train formations. Source: Modified from Pyrgidis, C., 2016. *Railway Transportation Systems - Design, Construction, and Operation*. CRC Press, Boca Raton.

Railway Transport Systems Classification

Elementary Classification of Railway Systems

Railway systems are classified based on the geographic scope in which they operate as well as on their functionality / provided services. They are divided into: (a) Intercity systems, (b) Suburban systems, (c) Urban systems, (d) Steep gradient railway systems, and (e) Freight railway systems. These classes may be analyzed further as shown in Fig. 5.

Intercity railway systems serve trips greater than 150 km and usually link major urban centers. Assuming that $V_{\max}$ refers to the maximum running speed, V_c to the commercial speed, and L to the route length, then these type of systems include conventional speed railway systems ($V_{\max} < 200 \text{ km} \cdot \text{h}^{-1}$), high-speed railway systems ($200 \text{ km} \cdot \text{h}^{-1} \leq V_{\max} < 250 \text{ km} \cdot \text{h}^{-1}$, $V_c \geq 150 \text{ km} \cdot \text{h}^{-1}$, $L \geq 150 \text{ km}$) and very high speed railway systems ($V_{\max} \geq 250 \text{ km} \cdot \text{h}^{-1}$, $V_c \geq 200 \text{ km} \cdot \text{h}^{-1}$, $L > 150 \text{ km}$) (Pyrgidis, 2017).

The suburban railway is a means of transport usually running on electrified lines with characteristics adapted to commuter services within the impact limits of major urban areas (suburbs and satellite regional centers). Its range can exceed 100 km and may extend up to 150 km. The nomenclature varies. When covering lengths of 30–50 km it is designated as urban rail, while when it covers longer lengths it is sometimes dubbed as “regional” railway.

Urban railway systems include: (a) metros, (b) light metros, (c) tramways, (d) monorail systems, and (e) driverless railway systems of low/medium transport capacity.

“Steep gradient railway” systems serve small distance connections with a significant difference of altitude between the two edges of the railway line. They are separated into rack railways and cable propelled railway systems.

Freight railway systems may be either dedicated or share tracks with intercity or suburban railway systems. Freight routes have a range far longer than passenger routes, which may reach 10,000 km. Depending on their load per axle and the nature of goods transported they may be classified as follows:

- trains of conventional loads (axle load $Q \leq 25 \text{ t}$)
- trains of heavy loads (axle load $Q > 25 \text{ t}$)
- trains transporting hazardous goods
- trains transporting small parcels

Secondary Classification of Railway Systems

As already stated, metros may be classified into heavy or light, based on the passenger volume they serve. Moreover, metros being closed systems with generally uniform rolling stock are prime candidates for higher levels of automation. Therefore, they may be distinguished based on the level of automation they achieve from grade 1 (Automatic Train Protection with Driver) to Grade 4 (Unattended Train Operation) (UITP, n.d.). Moreover, metro trains may be distinguished based on the guidance systems they utilized namely into trains with steel wheel and trains with rubber-tyred wheels.

Tramway systems may be divided into four categories based on their technical/operational aspects, and more specifically on the nature and extent of services they provide. These categories are urban tramways, long distance tramways (tram-trains), tourist tramways or cultural heritage trams, and transport of freight.

Typically, tramway systems are also categorized based on the distance between the floor of the vehicle and the top of rail, into low floor (30–40 cm), very low floor ($< 30 \text{ cm}$), moderately high floor (40–70 cm), and high floor (70–100 cm). Moreover, depending on the percentage of their length which is low-floor, trams are distinguished in the following categories:

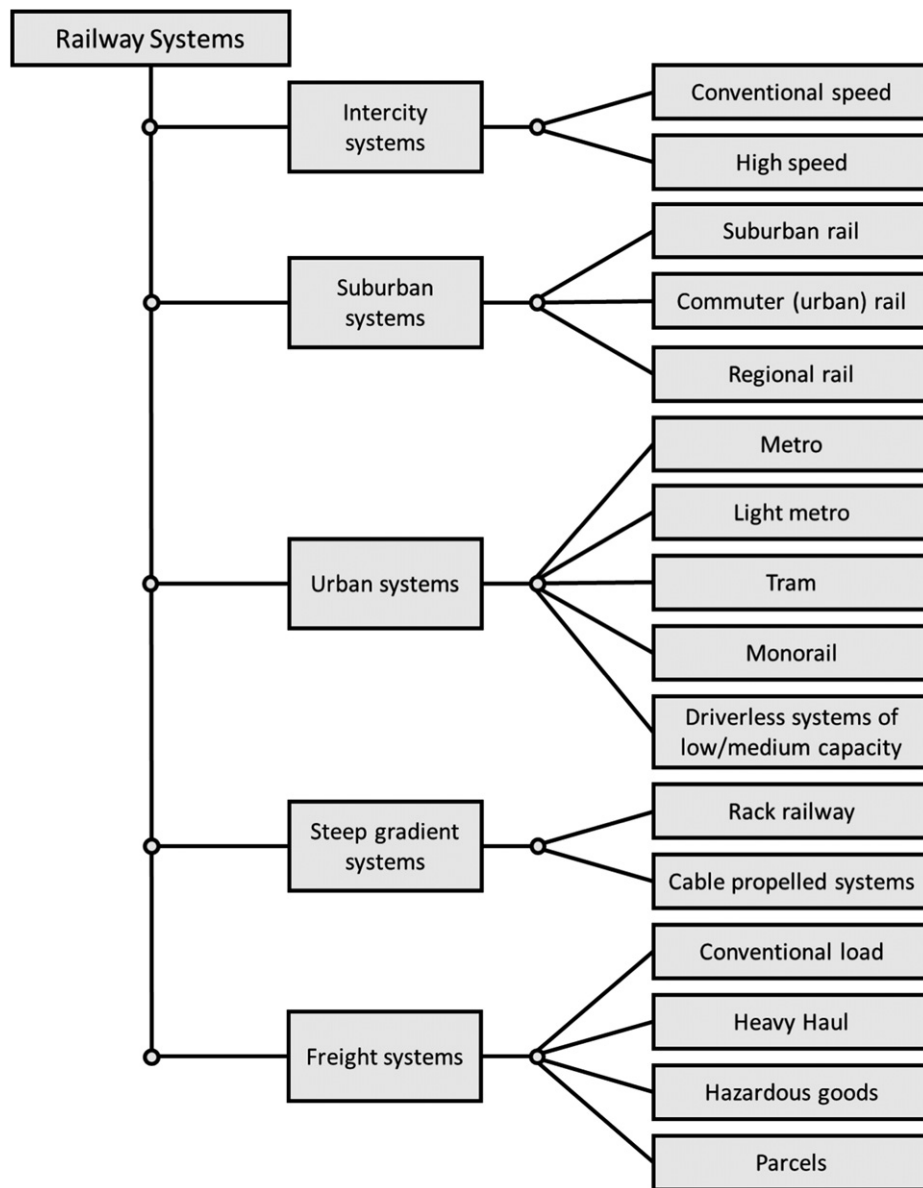


Figure 5 Classification of railway transportation systems. Source: Modified from Pyrgidis, C., 2016. *Railway Transportation Systems - Design, Construction, and Operation*. CRC Press, Boca Raton.

- totally (100%) low floor (low floor throughout the length of the vehicle)
- partially low floor (e.g., for 70% of the vehicle length)

Based on their power supply, they are distinguished into electrically powered via overhead wires, and electrically powered via supply at ground level. The latter may be conventional systems with a third rail, conventional systems with fourth and fifth rail, or possess more novel technologies such as reliance on inductive (contactless) transfer of energy to the vehicle. The electrical power supply may also be provided via energy storage devices, namely supercapacitors charged at stops via overhead wires). Finally mixed supply systems (e.g., overhead wires and storage devices) are also encountered.

Based on the formation of trains, trams may be divided into:

- simple
- articulated
- coupled (articulated or simple)

Cable propelled railway systems are divided based on the technique that is used for their acquisition of adequate traction for operating in steep gradients, namely into funicular (nondetachable cable), cable railway (detachable cable), and inclined elevator. Railway systems that utilize a rack/cog to achieve adequate traction for operating in steep gradients, are split into two categories based on the type of adhesion, namely into those with purely toothed (racked) tracks, and those exhibiting mixed adhesion operation (dual rack and adhesion systems).

Monorail systems are divided based on the placement of the trains in regards to the guidance beam and in regards to their transport capacity. Based on the former they are divided into straddled, suspended, and cantilevered systems. Based on the latter they are divided into small, large, and standard (compact) monorails (Kato et al., 2004).

In freight railway transportation, there are differences in the provided services which can be attributed to the variety of customer demands and, most importantly, to the volume and type of the transported cargo. The provided services fall into the following categories (Pyrgidis, 2016):

- less-than-wagonload services
- wagonload services
- services of combined transportation
- services of heavy haul transport
- special services

Less-than-wagonload services, also termed as part-load traffic, LCL (less-than-carload), or LTL (less-than-truckload) concern the transport of small packages under very strict time conditions and can be characterized as high speed freight services, e.g., TGV Postal high-speed trains and more recently the Mercitalia Fast service of the Italian State Railway (Troche, 2005; Zanuy, 2013).

Wagonload services are distinguished in Single WagonLoad (SWL) services and in TrainLoad (TL) services. SWL services involve cargo movements using single wagon trains. In such traditional freight services, the wagons (each wagon containing loads of single or multiple shipper(s)/consignee(s)) start their journey from different origin points or end their journey at different destinations, a fact which necessitates that they pass through intermediary marshaling yards. Regarding TrainLoad services cargo movement is performed by:

- Unit trains also termed full trains. These trains are formed of full wagons, which start from the same origin and travel to the same destination. These trains make full use of the tractive power of the locomotives. They do not stop at marshaling yards and they run between two specific terminal stations. They have one shipper, one consignee, one bill of lading and serve more than one final receivers (Zanuy, 2013).
- Block trains. These, like unit trains, run between two specific terminal stations without stopping at any marshaling yard. Their difference lies in the fact that they do not use all the power of their traction units while hauling.
- Exclusive trains. These are unit trains or block trains which serve a single final receiver.

Special services involve dedicated connections between mines and ports or mines and factories. The railway operates on a continuous basis ensuring a constant load flow, while special wagons are often used for transportation. A “door to door” delivery can also be considered as a special service, since the railway company undertakes receiving and delivering products from/to the user’s “door.”

In several cases (e.g., Block or Exclusive trains) the nomenclature varies. Block trains and exclusive trains are usually termed also unit trains

Typical Rolling Stock Characteristics per Railway System

This section provides in Table 4 typical rolling stock technical and operational characteristics for each of the aforementioned types of railway systems. These are indicative of the majority of railway systems currently in operation.

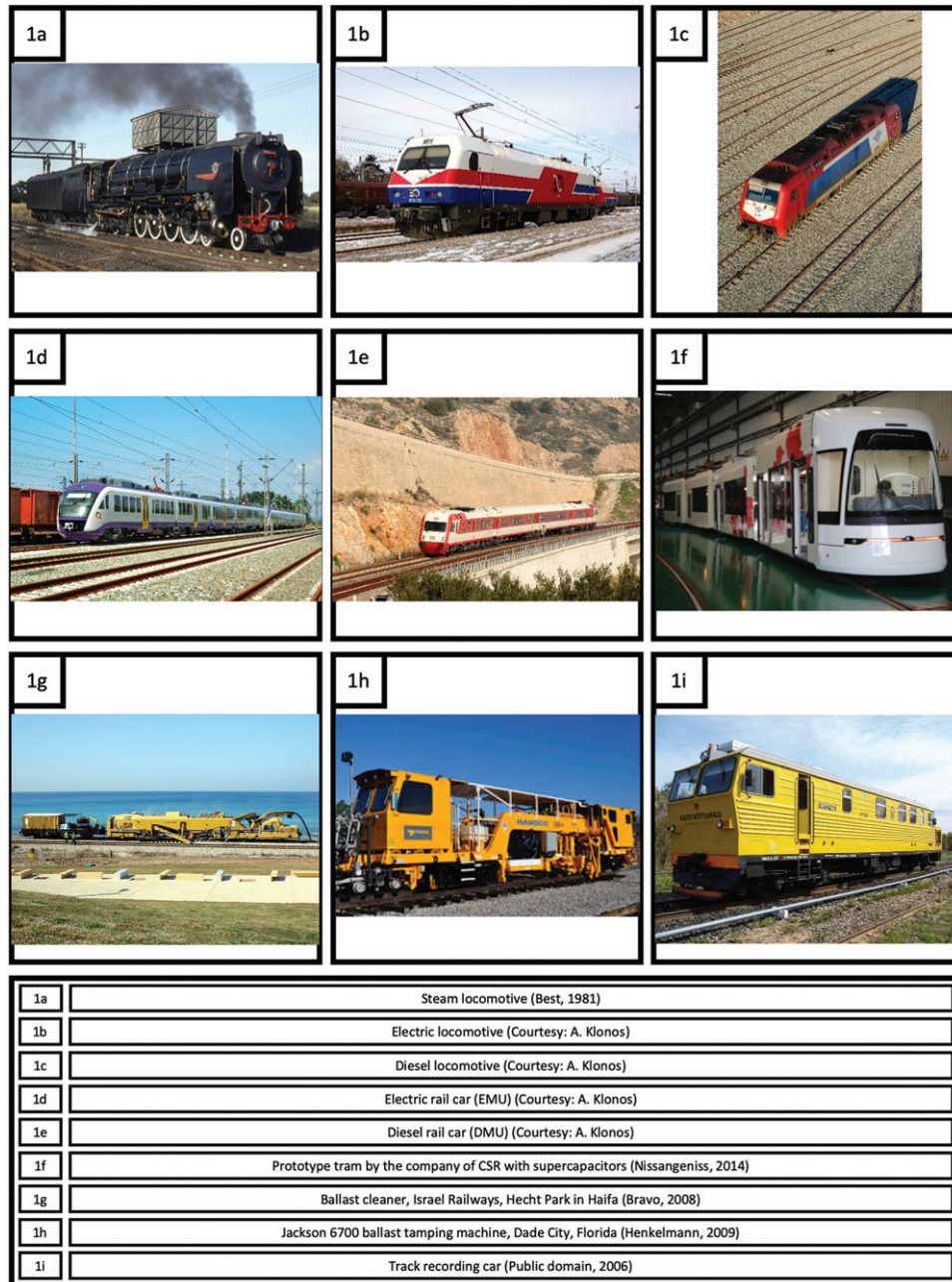
Summary and Conclusions

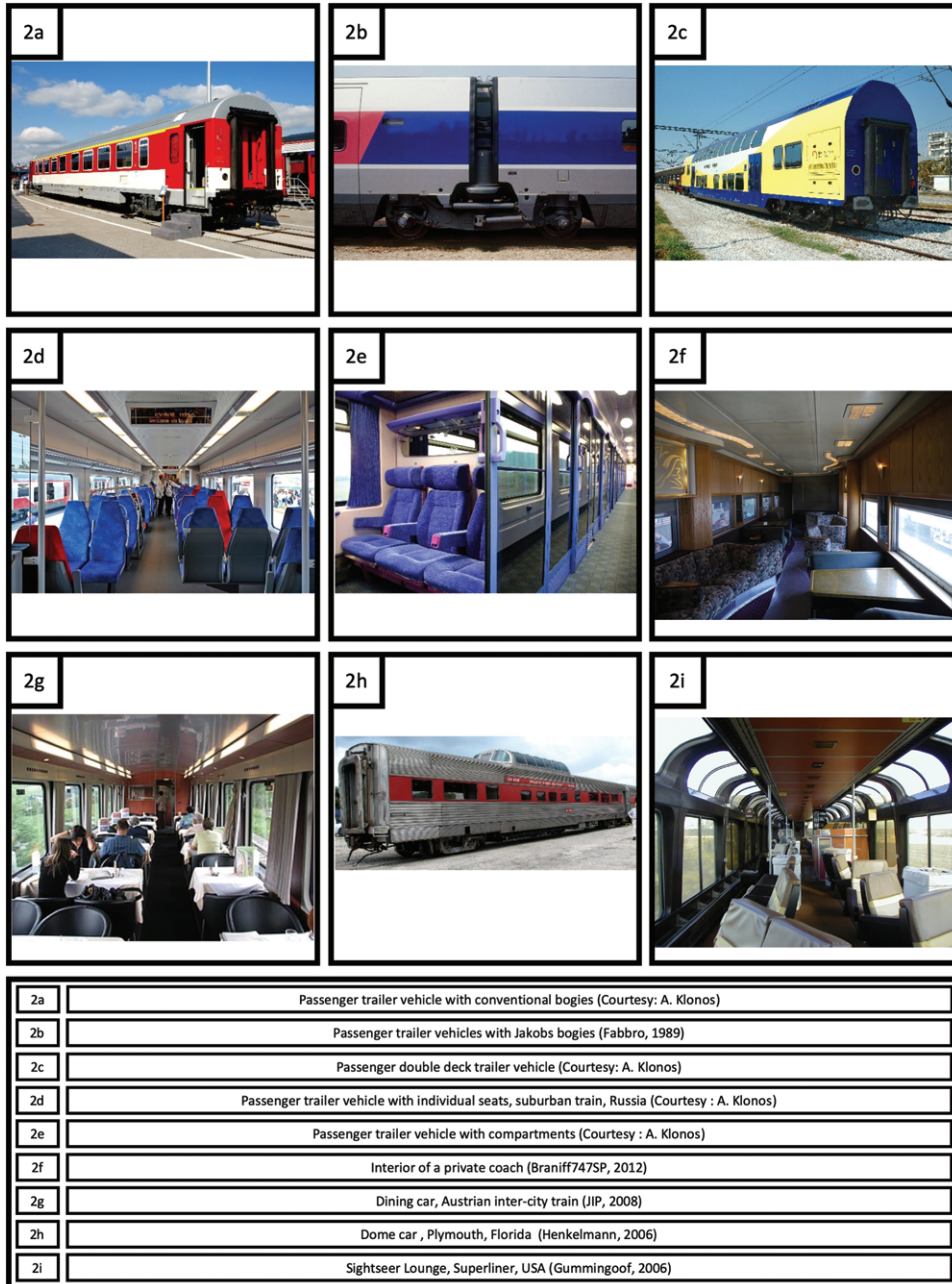
Railway vehicles may be classified based on various criteria and at different levels based on their integration in the wider railway system. Not all classification criteria are applicable for all types of vehicles and it is paramount to keep in mind the category of railway system in which the vehicles are called to operate, as this distinction shapes the vehicles characteristics to a great extent. Various organizations, from individual companies to international entities, use their own versions of existing classification and/or numbering systems, though guidelines exist from official sources.

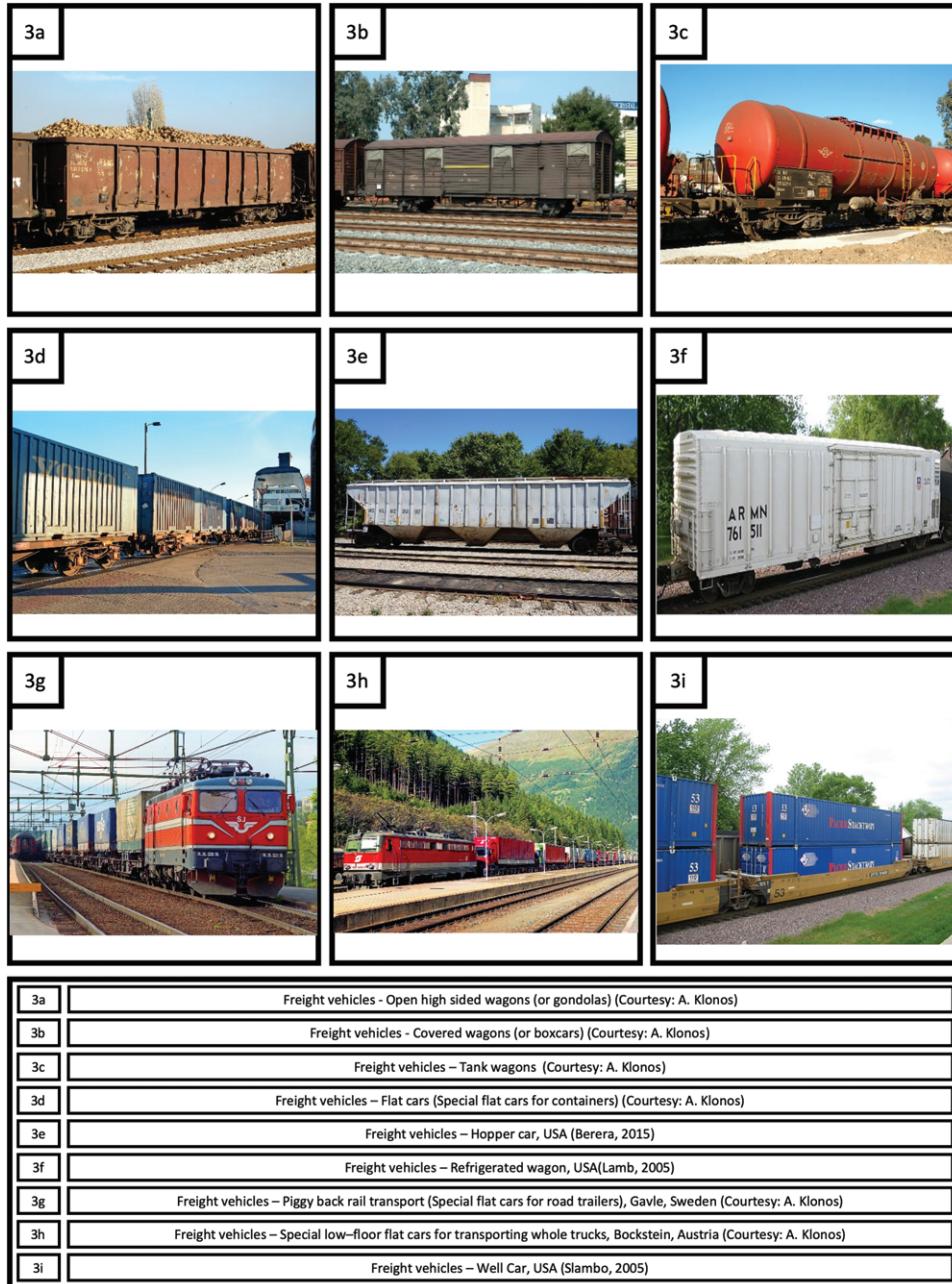
Table 4 Typical rolling stock technical and operational characteristics per category of railway system

Railway system	Rolling stock characteristics			
	Traction type	Max running speed	Train formation	Other
High-speed railway	EMU or electric loco-hauled trains	$V_{\max} \geq 200 \text{ km} \cdot \text{h}^{-1}$	4–10 vehicles	- Conventional car body or car body equipped with tilting technology - Vehicles of normal or broad track gauge - Vehicles with Jakobs-type bogies or conventional bogies or bogies with independently rotating wheels - Single deck or double deck vehicles - Trains with air-tight car-body structure - Vehicles with compartments or with individual passenger seats - A and B Class vehicles
Conventional speed intercity railway	EMU or DMU or diesel or electric loco-hauled trains	$V_{\max} < 200 \text{ km} \cdot \text{h}^{-1}$	4–10 vehicles	- Conventional car body or car body equipped with tilting technology - Vehicles of all track gauges - Vehicles with conventional bogies (usually) - Single deck vehicles (usually) - Vehicles with compartments or with individual passenger seats - A and B Class vehicles
Suburban/urban/regional railway	EMU or DMU or diesel or electric loco-hauled trains	$V_{\max} \leq 160 \text{ km} \cdot \text{h}^{-1}$	2–8 vehicles/push–pull trains	- Conventional car body - Vehicles of all gauges - Vehicles with conventional bogies (usually) - Single or double deck vehicles - Vehicles with individual passenger seats (relatively high number of seats and low number of standing passengers)
Metro	EMU	$V_{\max}$: $90\text{--}120 \text{ km} \cdot \text{h}^{-1}$	4–10 vehicles	- Special rolling stock - Floor at platform level - Aluminum, steel, or stainless steel carbodies - Vehicles of normal or metric track gauge - Vehicles with conventional bogies - Vehicles with individual passenger seats (relatively low number of seats and high number of standing passengers)
Tramway	Electric traction (overhead wires, through the ground, with energy storage systems)	$V_{\max}$: $80\text{--}90 \text{ km} \cdot \text{h}^{-1}$	Articulated trains	- Special rolling stock - Aluminum carbodies - Vehicles of normal or metric track gauge - Low floor or very low floor vehicles - Vehicles with bogies with independently rotating wheels or bogies of mixed behavior - Vehicles with individual passenger seats (relatively low number of seats and high number of standing passengers)
Monorail	Electric motors setting a system of rubber tyred wheels to roll	$V_{\max}$: $60\text{--}100 \text{ km} \cdot \text{h}^{-1}$	Articulated trains (2–6 and rarely 8 vehicles)	- Special rolling stock - Suspended or straddle design - Rubber or metal wheels
Cog railway (pure rack)	Electric and diesel railcars	$V_{\max}$: $40 \text{ km} \cdot \text{h}^{-1}$	Usually a single railcar	- Special rolling stock - Vehicles of all track gauges (usually metric)
Funicular	Via pulled electric cables placed on the track superstructure	$V_{\max}$: $50 \text{ km} \cdot \text{h}^{-1}$ (usually $< 20 \text{ km} \cdot \text{h}^{-1}$)	Cabin vehicles (one ascending, one descending)	- Special rolling stock - Nondetachable vehicles—counter balance system - Vehicles of all gauges (usually metric)
Freight	Diesel or electric loco-hauled trains	$V_{\max} = 120 \text{ km} \cdot \text{h}^{-1}$	Very long trains (750 m–2 km for conventional loads and longer for heavy loads)	- Vehicles of all track gauges - Various types of wagons

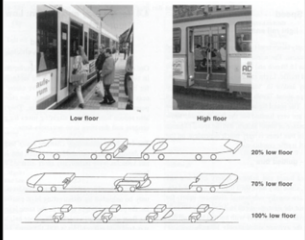
Annex

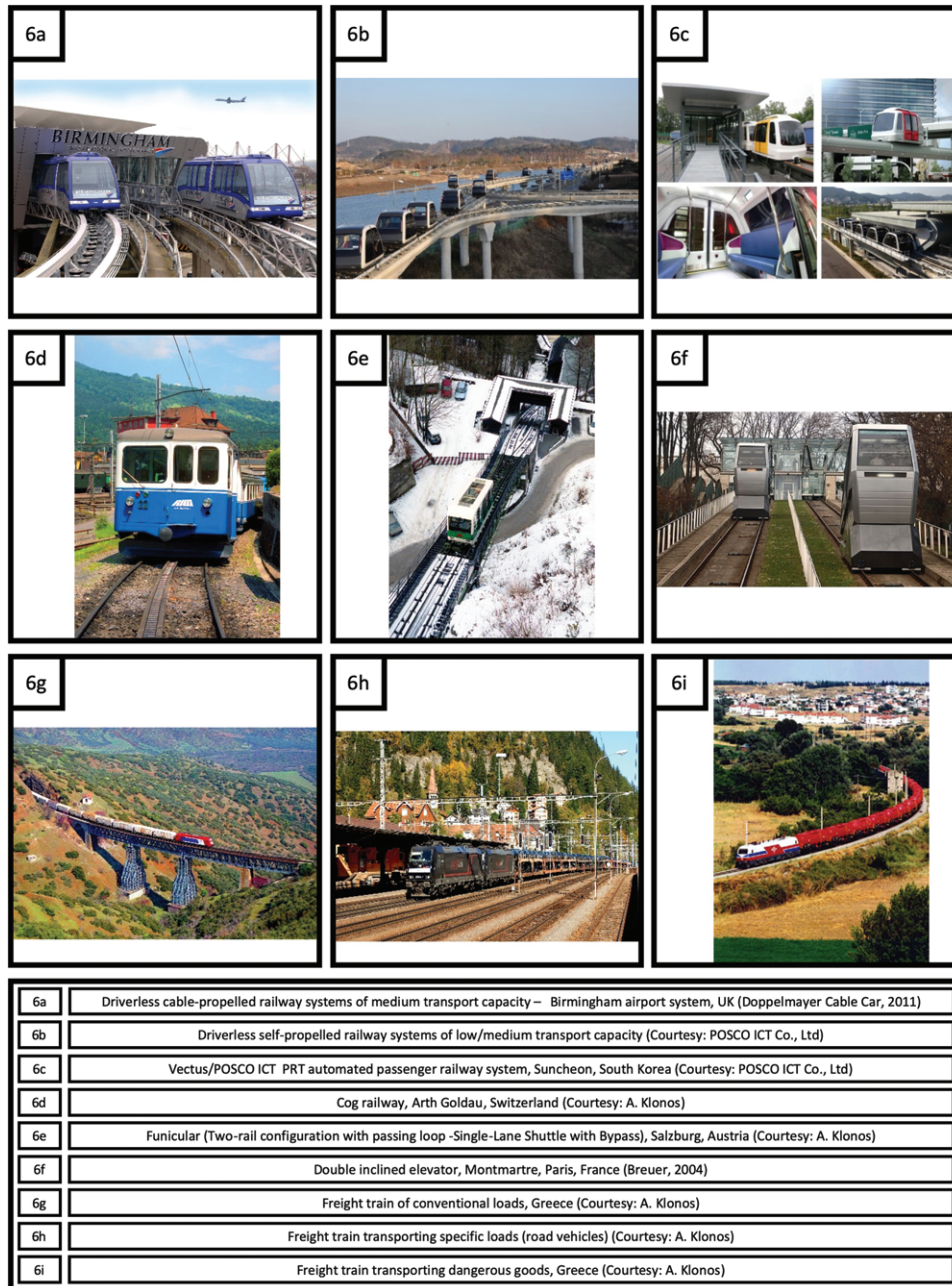






4a		4b		4c	
4d		4e		4f	
4g		4h		4i	
4a	Conventional speed intercity railway system, Helsinki, Finland (Electric loco-hauled passenger train) (Courtesy: A. Klonos)				
4b	High speed railway system (THALYS) (Electric loco-hauled passenger train) (Courtesy : A. Klonos)				
4c	Tilting train, Silenen, Switzerland (Courtesy: A. Klonos)				
4d	Suburban railway system, Greece (diesel rail car) (Courtesy : A. Klonos)				
4e	Suburban railway system, Germany (Double-deck electric railcar) (Courtesy : A. Klonos)				
4f	Heavy metro (with driver), Athens, Greece (Courtesy : A. Klonos)				
4g	Driverless metro, Paris, France (Courtesy: A. Klonos)				
4h	Light metro (Driverless), Copenhagen, Denmark (Courtesy: A. Panagiotopoulos)				
4i	Metro with rubber tyred wheels(driverless), Lausanne, Switzerland (Joam, 2006)				

5a		5b		5c	
5d		5e		5f	
5g		5h		5i	
5a	Tramway system, Montpellier, France (Courtesy: A. Klonos)				
5b	Free catenary tramway system (APS), Bordeaux, France (Pline, 2008)				
5c	Tram -train, Hague, The Netherlands (Courtesy: A. Klonos)				
5d	Freight tram, Dresden, Germany (Kaffeeinstein, 2008)				
5e	Low and high floor tram (Hass-Klau, et al, 2003)				
5f	Double articulated tram, Hague, The Netherlands (Courtesy: A. Klonos)				
5g	Straddled monorail system, Dubai, UAE (Own work)				
5h	Suspended monorail system, Memphis, Tennessee, USA (Nightryder84 ,2005)				
5i	Cantilevered monorail system (The Monorail Society, n.d.)				



References

- Biasin, D., Lavogiez, H., & Pichant, J.C., 2008. Trans-European Conventional Rail System - Subsystem Rolling Stock. European Railway Agency.(ERA).
- ENSCO, 2015. Track Inspection Vehicles. Retrieved from: <https://www.ensco.com/rail/track-inspection-vehicles>.
- Frey, S., 2012. Railway Electrification Systems & Engineering. White Word Publications, Delhi.
- Hamm, D., n.d. Rail vehicle numbering explained. RSSB. Retrieved from: <https://www.rssb.co.uk/Pages/improving-industry-performance/rail-vehicle-numbering-explained.aspx>.
- Hyatt, K., 2018. Bombardier introduces a battery-electric hybrid passenger train. CNET. Retrieved from: <https://www.cnet.com/roadshow/news/bombardier-germany-electric-hybrid-train/>.
- Jeong, W., Kwon, S.B., Park, D. & Jung, W.S., 2011. Efficient Energy Management for Onboard Battery-driven Light Railway Vehicle. 9th World Congress on Railway Research, Lille.
- Kato, M., Yamazaki, K., Amazawa, T., Tamotsu, T., 2004. Straddle – type monorail systems with driverless train operation system. Hitachi Review 53 (1), 25–29, 2004.
- Konarzewski, M., Niezgoda, T., Stankiewicz, M., Szurgott, P., 2013. Hybrid locomotives overview of construction solutions. Journal of KONES Powertrain and Transport 20 .
- Leo., n.d. Types of Coaches (In German) URL: <https://web.archive.org/web/20081228112927/http://www.leo.org/information/freizeit/bahn/gattungszeichen.html>.
- Metzler, J.M., 1981. Le matériel à voyageurs. Lecture Notes, ENPC, Paris.
- Pyrgidis, C., 2016. Railway Transportation Systems - Design, Construction, and Operation. CRC Press, Boca Raton.
- Pyrgidis, C., 2017. Defining high-speed rail: a proposal that endeavors to meet the requirements of both rail infrastructure manager and railway operator. Rail Engineering International (2), 14–16.

Railsystem, n.d. Electric Traction Systems. Retrieved from: <http://www.railsystem.net/electric-traction-systems/>.

Railway technology, 2017. Track Inspection Technologies. Retrieved from: <https://www.railway-technology.com/products/track-inspection-technologies/>.

Ronanki, D., Singh, S. & Williamson, S., 2017. Comprehensive Topological Overview of Rolling Stock Architectures and Recent Trends in Electric Railway Traction Systems. IEEE Transactions on Transportation Electrification (Volume: 3, Issue: 3), pp. 724-738.

Schiffer, V., 2005. Generic Types of Freight Cars (In German). URL: <http://home.wtal.de/gueterwagen/gatt-zde.htm>.

Sheed, T.C., Allen, G.F., n.d. Locomotive. Encyclopedia Britannica. Retrieved from: <https://www.britannica.com/technology/locomotive-vehicle>.

Troche, G., 2005. High-speed rail freight Sub-report in Efficient train systems for freight transport, KTH Railway Group Report 0512, Stockholm.

UITP, n.d. Press Kit Metro Automation Facts, Figures and Trends. Retrieved from: <https://www.uitp.org/sites/default/files/Metro%20automation%20-%20facts%20and%20figures.pdf>.

Zanuy, A.C. ,2013. Future Prospects on Railway Freight Transportation A Particular View of the Weight Issue on Intermodal Trains, PhD, Technische Universität Berlin Retrieved from: https://www.railways.tu-erlin.de/fileadmin/tg98/papers/2013/PhD_Armando_Carrillo_Zanuy.pdf.

Further Reading

Berera, M. (2015). A hopper car at the Pilgrim's Pride feed mill in Pittsburg, Texas (United States). URL: [https://en.wikipedia.org/wiki/Hopper_car#/media/File:Pittsburg_August_2015_17_\(hopper_car\).jpg](https://en.wikipedia.org/wiki/Hopper_car#/media/File:Pittsburg_August_2015_17_(hopper_car).jpg).

Best, M. (1981). South African Class 25 4-8-4. URL: https://en.wikipedia.org/wiki/South_African_Class_25_4-8-4.

Braniff747SP. (2012). Interior of a private coach. URL: https://en.wikipedia.org/wiki/Private_railroad_car#/media/File:National_Train_Day_-_Private_Varnish_Interior.JPG.

Bravo, G. (2008). Israel Railways ballast cleaner in action at Hecht Park in Haifa, Israel. URL: https://ru.wikipedia.org/wiki/%D0%A4%D0%B0%D0%B9%D0%BB:Ballast_Cleaner171.jpg.

Breuer, R., (2004). BreMontmartre Funicular. URL: https://en.wikipedia.org/wiki/Montmartre_Funicular.

DCC Doppelmayr Cable Car. (2011). AirRail Link. URL: https://en.wikipedia.org/wiki/AirRail_Link.

Fabbro, J.M. 1989, SNCF Médiathèque.

Gummingoof, M. N. (2006). Upper level of Superliner I Sightseer Lounge (formerly Lounge Café) car. URL: https://en.wikipedia.org/wiki/File:Superliner_I_Lounge_upper.jpg.

Hass-Klau, C. et al. 2003, Bus and Light rail: Making the right choice, ETP, Brighton.

Henkelmann, H. (2006). California Zephyr dome car in 2008. URL: https://en.wikipedia.org/wiki/Dome_car#/media/File:InlandLakesDome.jpg.

Henkelmann, H. (2009). Jackson 6700 switch tamping machine. URL: https://en.wikipedia.org/wiki/Tamping_machine#/media/File:Ballast_Tamper.JPG.

JIP. (2008). A dining car on an Austrian inter-city train in 2008 URL: https://en.wikipedia.org/wiki/Dining_car#/media/File:Restaurant_car_in_Austria.jpg.

Joam, I. (2006). Rubber-tyred metro. URL: https://en.wikipedia.org/wiki/Rubber-tyred_metro.

Kaffeeinstein. (2005). Dresden CarGoTram. URL: <https://www.flickr.com/photos/kaffeeinstein/161221383/>.

Lamb, S. (2005). ARMN 761511 20050529 IL Rochelle. URL: https://commons.wikimedia.org/wiki/File:ARMN_761511_20050529_IL_Rochelle.jpg.

Nightryder84. (2005). Memphis Suspension Railway. URL: https://en.wikipedia.org/wiki/Memphis_Suspension_Railway.

Nissangenis. (2014) Guangzhou Trams URL: https://en.wikipedia.org/wiki/Guangzhou_Trams#/media/File:Guangzhou_Haizhu_District_CSR-Zhuzhou_Tram_For_No.05.jpg.

Pline. (2008). citadis of the Line B near the square "place des Quinconces", Bordeaux, France. URL: https://commons.wikimedia.org/wiki/File:XDSC_7576-tramway-Bordeaux-ligne-B-place-des-Quinconces.jpg.

POSCO ICT Co., Ltd. (2014)., Korea's First Personal Rapid Transit (PRT), SkyCube, URL: <http://globalblog.posco.com/koreas-first-personal-rapid-transit-prt-skycube/>.

POSCO ICT Co., Ltd. (2015). SunCheon SkyCube PRT Project Overview v1.1. URL: <http://en.minimetro.com/Application-area/For-Traffic-Planners>. (accessed 7 April 2015).

Public Domain. (2006). Track geometry car in Russia. URL: https://en.wikipedia.org/wiki/Track_geometry_car#/media/File:MTKP.JPG.

Slambo. (2005). Well car. https://en.wikipedia.org/wiki/Well_car.

The Monorail Society. (n.d.). Cantilevered. URL: <http://www.monorails.org/tMspages/TPindex.html>.

Introduction to Maritime Shipping

Christa Sys, Thierry Vanelslander, University of Antwerpen, Antwerpen, Belgium

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	508
Introduction	508
The Relationship Between International Maritime Trade and Maritime Transport	508
Market Concentration	511
Supply, Demand, and Freight Rates	513
Tramping	513
Container Liner Shipping	514
Acknowledgment	515
Biographies	515
Permissions	516
See Also	516
References	516
Further Reading	516

Nomenclature

BCTI Baltic Clean Tanker index

BDTI Baltic Dirty Tanker index

DWT Deadweight

GT Gross tonnage

HHI Herfindahl-Hirschman Index

TEU Twenty feet equivalent unit

VLCC Very large crude carriers

If appropriate, list and define any unusual symbols used in your article.

Divided into appropriate sections, covering the contents as agreed with the Editors.

Introduction

Maritime transport is a powerful and diversified industry with international dimensions. The international business environment in which maritime transport operates has changed significantly in recent years and makes it subject to a high degree of uncertainty (e.g., regulation). To understand this industry, the aim of this chapter is to introduce commonly used maritime shipping terminology (section, The Relationship Between International Maritime Trade and Maritime Transport), to discuss the market structure and setting that determines shipping companies' behavior (section, Market Concentration) and to look at another key issue affecting the shipping operator's performance, namely, the freight rate as an indicator (section, Supply, Demand, and Freight Rates).

The Relationship Between International Maritime Trade and Maritime Transport

International maritime trade is inextricably linked to maritime transport (i.e., concerns the movement of passengers and freight). The importance of international maritime trade for total international trade is considerable. The developments of each of them influence the other. It is a continuous interaction in which an economically sustainable response must be formulated to key trends such as industrialization and urbanization in emerging economies, changing demographics, disruptive technologies (a.o. digitalization, (semi-)autonomous ships, artificial intelligence), sustainability and climate-related disruptions, and trade and geopolitical tensions (Fig. 1).

If the trade forecasts of the International Monetary Fund (IMF), the World Trade Organization (WTO) and the World Bank come true, world total trade is expected to grow; and consequently world seaborne trade too. In what follows, world (seaborne) trade will first be considered before going deeper into the evolution of the world fleet.

In 2018, the world total trade (all modes) equalled 14.16 billion tons, of which 11.85 billion tons total seaborne trade. Since 2000, about 5.5 billion tons total seaborne trade was added. Seaborne trade expressed as a percentage of total trade corresponds to 83%, before the 1990s highly correlated to GDP; while put in global trade by value carried by sea, it amounts to about 70% (UNCTAD, 2019). To study world seaborne trade trends, figures for trade are often expressed in tons-miles or a ton of cargo moved one mile, a very accurate indicator of ship demand. Over the period 2000–09, total seaborne trade expressed in billion tons-miles

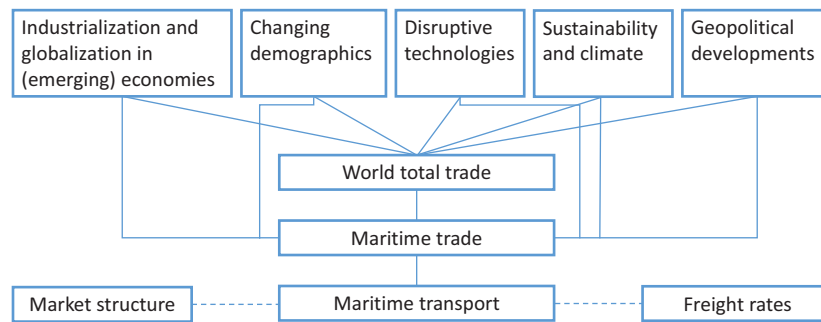


Figure 1 Relationship world total trade and maritime trade. Source: Own compilation based on Meersman (2009).

increased from 30,718 to 39,713. It is expected that in 2019 this indicator will exceed 60,000 tons-miles (2020: 62,000 tons-miles); thus nearly doubling due to more seaborne trade and increased number of long-haul trades (Clarkson Research Service, 2019; UNCTAD, 2019).

The choice was made to go back two decades, hence 1999–2019, so as to observe sufficiently long evolutions in the shipping business, which is marked by exceptional shorter-run spikes on the one hand, which have specific causes, and longer-run slowly evolving patterns on the other hand. Not only recapturing the past/present but also looking toward the future, it is particularly challenging for international organizations (International Monetary Fund, World Bank, International Transport Forum, UNCTAD), renowned industry analysts, classification societies, port authorities, and carriers to come up with accurate forecasts. A number of disruptive factors including economic (population, (de-)globalization, re-shoring), political (e.g., tariffs and trade tensions), climate, advanced technologies, infrastructure development, new regulations, digitalization and the growth of e-commerce are influencing the long-term outlook of the maritime sector. Forecasts (2030, 2050) cannot take away the growing uncertainty; at most isolate it. The uncertainty in the maritime transport environment can then be filtered by working with scenarios.

Commonly, three scenarios (low, base, and high scenario) are put forward; however given that the maritime sector faces a period of transformation, here, a min/max scenario is put forward. Numerical data below is based on the Shipping Review and Outlook tables published by Clarkson Research Service (2019). Based on different forecasts, the minimum scenario likely corresponds to an average growth of 0.5%–1% per annum, hence varying between 13.1 and 13.8 billions tons total sea trade; while the maximum scenario estimates a steady average annual growth rate of 2.4%–3.2% world sea trade volume (15.8 – 17.1 billions tons) over the period 2019–30. For the subsequent 2 decades, DNVGL (2018) assumes “a minimal 0.1% average world seaborne trade volume annual growth rate”; a very cautious view in comparison to other forecasts. Assuming 16 billions tons in 2030 and a growth rate of 1.5% per annum might result in 22.7 billion tons, just not doubling in comparison to 2019. In contrast, there is a greater consensus with regard to the expectations for the tons-miles indicator. Due to the fact that some maritime supply chains are expected to become more regionally oriented, the outlooks for the tons-miles indicator might level off as off 2030, given that the variable ‘distance’ (expressed in miles) may decrease.

Fig. 2 shows data about the world seaborne trade, with further classification by commodities, for 2000, 2009–19. First of all, all commodity categories are growing, but each has a different growth pattern. Next, the biggest shipping segments are liquid and dry bulk tonnage. Regarding the latter, commonly, a distinction is made between the major and minor bulk commodities. Under major bulk commodities is understood: iron ore, grain, coal, phosphates, and bauxite; while minor bulk commodities are agribulks, metals and minerals, and manufactures. Expressed in estimated billion tons-miles, the share of dry bulk amounted to 42.28% in 2000 and 49.55% in 2019. This evolution is due to the increase of the three most important raw material trades (iron ore, coal, and grain) and a considerable growth of the minor bulk during the last decade. In 2000, iron ore provides 38.43% of the volume for the dry bulk tonnage and nearly half of the volume in 2019 due to an increase of its related tons-mile. The share of liquid bulk decreased from 31% to 23% over the same period; while the share of containers increased from 9.49% toward 14.78%. In the category “gas,” LNG featured the biggest growth. Other dry cargo has no marked growth, rather only recovered from the economic crisis.

For the next 2 decades, it is expected that growth in the world seaborne trade will be driven by growth in containerized (however, at a lower space) and gas and chemical cargoes (i.e., other tanker trade); while the outlook for the raw materials of the steel industry (coal, iron ore) and the power generation (crude oil trade) are challenged by geopolitical tensions and the climate debate. The impact of this evolution on the maritime sector is still unclear. In total, maritime cargo might stabilize. However, given that the world population continues to expand, ensuring supply of food (grain) and energy will have to remain ensured and/or new supply chains (water, alternative energy products) may arise.

In addition to cargo, in 2018, passenger traffic accounts for 28.2 million passenger numbers, and a continued rise (e.g., Chinese tourist market) is expected (UNWTO, 2019). However, the magnitude of the growth rates might be suppressed due to the mega trends (more demanding customers, transforming passenger experience due to advanced technology like AI, growing public awareness of industry’s impact on the environment, overtourism, and so on).

Not surprisingly, the evolution of the seaborne trade is reflected in the development of the world fleet. Dry bulk cargo refers to cargo that is typically not carried in tankers. Following the supply structure proposed by Stopford (2009), Fig. 3 presents the classification of the world fleet over two decades.

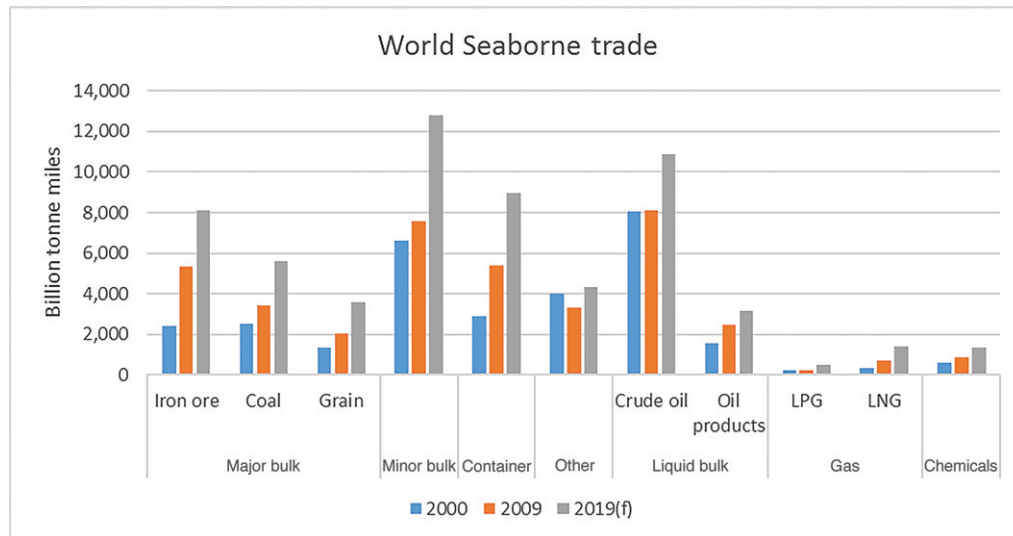


Figure 2 World seaborne trade. Source: Own compilation based on *Clarkson Research Service (2019)*.

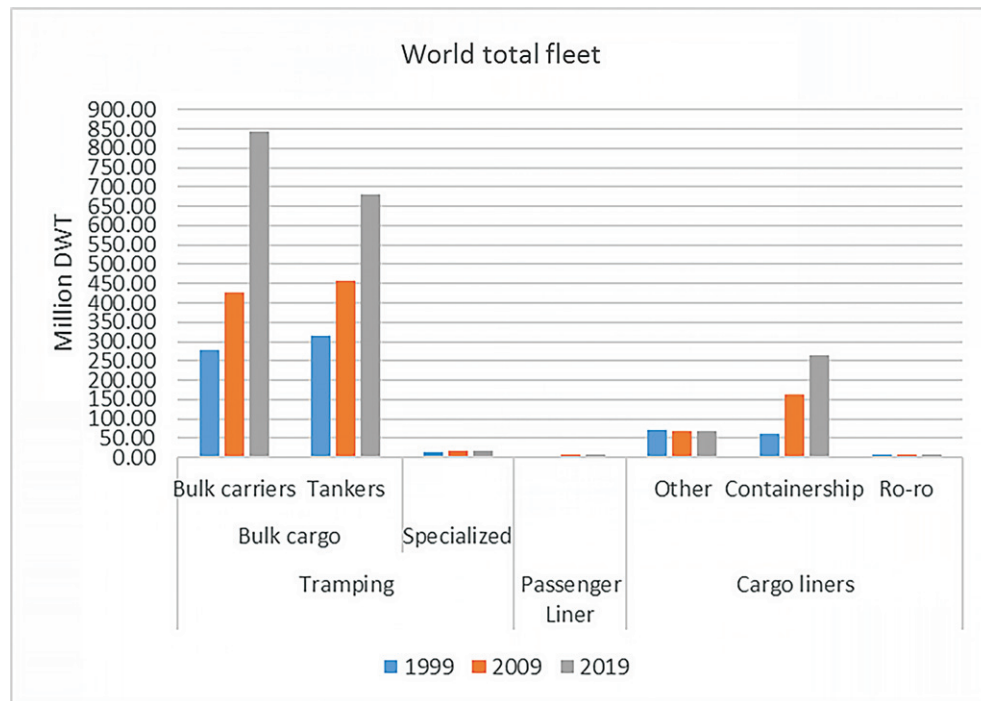


Figure 3 World total fleet. Source: Own compilation based on *Clarkson Research Service (2019)*.

The shipping market consists of two segments, tramping (no fixed sailing schedule) and (passengers, cargo) liners (fixed schedule). Homogeneous bulk cargoes are transported in conventional dry bulk carriers or tankers; while specialized or non-homogenous parcels (liquefied gas, chemicals, refrigerated, wood products, motorized vehicles) are transported each with their own fleet of especially designed ships (e.g., chemical parcel tanker, product tanker, reefer ship). Liner shipping relates to containership, Roll-on/Roll-off (RoRo), and other ship types.

In 1999, the world's total fleet (including offshore, dredgers, and other non-cargo) amounted to 61,770 ships, accounting for 776.6 million deadweight tons (DWT) of capacity. Deadweight tonnage or the cargo carrying capacity of a vessel, excluding the weight of the ship itself, is a way to measure the size of ships. In 2009, the world total fleet stood at 77,898 ships (1,206.5 million DWT). A decade later, the number of ships deployed corresponded to 98,027 (1,981.2 million DWT) or a growth factor of about 1.5 over the 1999–2019 period. In 2019, bulk carriers and oil tankers maintained the largest market shares of ships in the world fleet

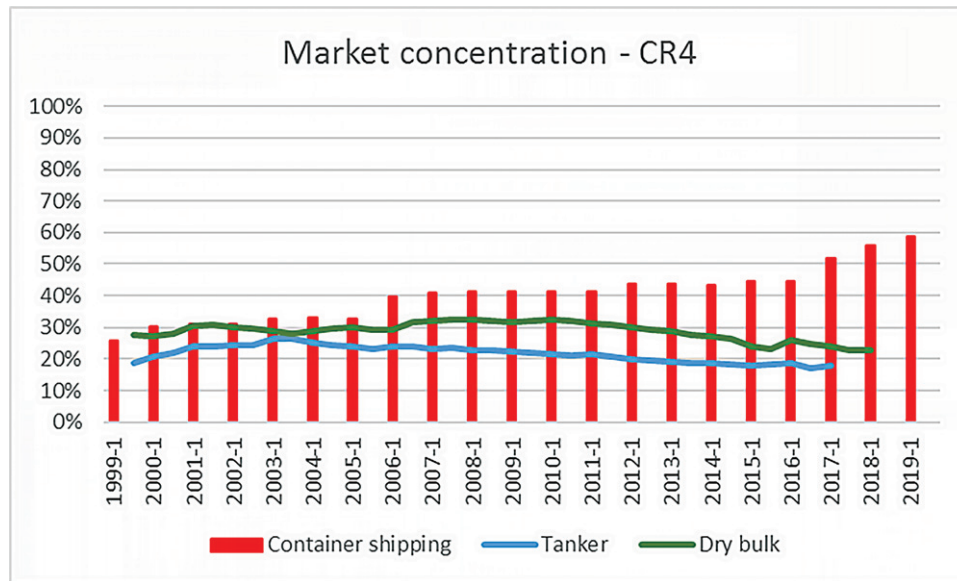


Figure 4 CR4 for the tanker, bulk and container shipping market. *Source: Own calculations based on Alphaliner (various editions) and Clarkson Research Service (2019).*

(dwt), at 42.7% (1999: 35.9%, 2009: 35.4%) and 30.9% (1999: 38.0%, 2009: 34.8%), respectively. Within the segment “cargo liner,” the number of containerships doubled (1999: 2,539 ships, 2019: 5,295 ships); while the capacity increased by a factor 5 (1999: 62,3 million DWT (8%), 2009: 162,2 million DWT (13.4%), 2019: 265,9 million DWT (13.4%)). The majority of the containers are transported with fully cellular ships (expressed in 20 feet equivalent—TEU), accounting for 98% of the total container capacity. In 1999, the carrying capacity of the fully cellular ships stood at 4,260, 870 TEUs. A decade later, the container carrying fleet capacity amounted to 12,351,410 TEUs. In 2019, the world fleet of fully cellular containerships stood at 5,295 cellular ships for 22,324,652 TEUs.

As the unit of Fig. 3 is million dwt, the share of passenger fleet is very small (1999: 4.7 million DWT, 2019: 7.1 million DWT). Expressed in number of ships, this segment grew from 6,055 (1999) to 8,252 ships in 2019.

Gross tonnage (GT) is an alternative way to represent the size of the vessel (based on volume). In 2019, the total world fleet equalled 98,027 ships, including 41,984 non-cargo ships (21.5% offshore, 4.9% dredgers, 47.2% tugs, 1.0% cruise, 19.0% ferries and 6.1% other) or corresponds to 1,384 million GT, operated by 25,547 owners (4,071 tanker owners, 3,345 bulk owners and 923 container ship owners). Dividing the number of ships by the total of owners results in an average of 3.8 ships per owner; hence suggesting many small entrepreneurial shipping companies, particularly in the tramp business. Based on forecasts of Clarkson Research Service (2019), the world fleet is expected to increase to about 1,800 m GT (2.5% per annum) in 2030. From 2030 to 2050, the world fleet will continue to grow, but at a slower pace. There will be notable difference in growth pattern per ship type linked to growth prospects of the commodities (see above).

Having an idea about the demand (cargo side) and supply (ship side), the next section analyses the market structure more in-depth. Hereafter, the analysis is limited to the largest segments in shipping: bulk carriers, tankers and containerships. These are three different markets; Commonalities among them will allow identifying global trends.

Market Concentration

In order to better understand the behavior of shipping companies, insight in the market structure (i.e., ranging from perfect competition at one extreme to monopoly at the other) is needed. This section looks at two measures of market concentration, namely, the four-firm concentration ratio (CR4) and the Herfindahl-Hirshman Index (HHI) and the instability index to determine the magnitude of market share instability. The first measure computes the combined market share of the top four ship owners in the respective industries (Fig. 4 and Fig. 5). Summing the squared market shares of each owner in the respective industry and multiplying by 10,000 allows calculating the HHI. (Fig. 6) (Lipczynski and Wilson, 2017; Sys, 2010) The instability index is the sum of the absolute value of the change between two points in time in the market share of each firm (Hymer and Pashigian, 1962; Sys, 2009)

Fig. 4 shows the evolution of market concentration over the 1999–2019 period, on the basis of data from the AXS-Alphaliner top 100 (www.alphaliner.com) and Clarkson Research Service (2019). Clarkson Research Service publishes the top 50 tanker owners and top 50 bulk owners twice a year (spring and autumn); while Alphaliner publishes monthly the Top 100. Here, the CR4 is measured for the three selected shipping markets as the share of the four leading ship owners against the top 50. A high CR4 should be interpreted as a highly concentrated market. What conclusions can be derived from this?

Year	Tanker			Bulk carrier			Container liner shipping		
	1999	2009	2019	1999	2009	2019	1999	2009	2019
1	MOL	Frederiksen Group	China Merchant	COSCO	COSCO	COSCO	Maersk Line	APM-Maersk	APM-Maersk
2	VELA International	MOL	COSCO	MOL	NYK	NYK	Evergreen/Uniglor	MSC	MSC
3	Frederiksen Group	NYK	Euronav	NYK	MOL	K-Line	P&O Nedlloyd	CMA CGM	COSCO
4	NYK	Teekay Corporation	Bahri	P&O Group	K-Line	Starbulk carriers	Hanjin/DST-Senator	Evergreen Line	CMA CGM

Figure 5 Top four carriers. Source: Own calculations based on Alphaliner (various editions) and Clarkson Research Service (2019).

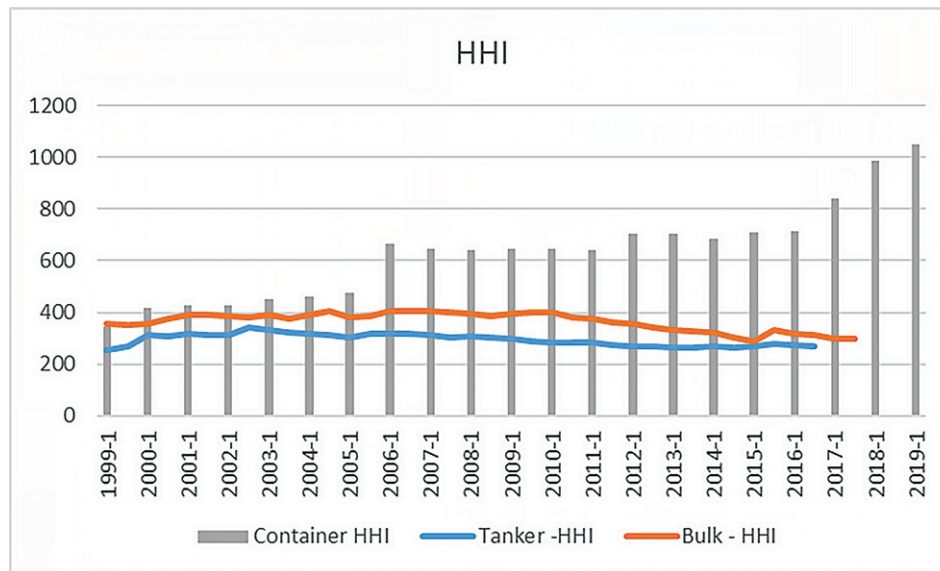


Figure 6 HHI. Source: Own calculations based on Alphaliner (various editions) and Clarkson Research Service (2019).

According to Lipczynski and Wilson (2017), any CR4 ratio of more than 40% corresponds to an oligopolistic market; thus all players recognize their interdependence. With respect to the container liner shipping, Fig. 4 shows that the CR4 has exceeded the 40% limit since the year 2007. In 2019, the top four container carriers nearly accounted for 60% of industry's carrying capacity. Moreover, the three consolidations waves are clearly observable: 1999/2000: Maersk Line + Sealand; 2005/2006: Maersk Sealand acquired RPONL; and the merger domino effect covering the 2015–18 period (2015: Hapag Lloyd + CSAV, Hamburg Süd + CCNI, CMA CGM + ODPR, 2016: COSCO + CSCL, CMA CGM + APL, 2017: Hapag Lloyd + UASC, Maersk Line + Hamburg Süd, 2018: COSCO + OOCL), resulting in increased concentration. On the other hand, both bulk-shipping markets are characterized by a less concentrated market. The top four carriers have a combined market share below 40%; the others are essentially fighting for the minor places. Clearly, the main characteristics of the tramp market are close to the perfect competition model ($CR4 < 5\%$). By economic theory, the perfect competition model is characterized by many buyers and sellers ("pricetaker"), perfect knowledge, identical products (homogeneous), free entry or exit, and firms that maximize profits. The last aspect is quite a challenge in the shipping industry.

In practice, there is no such thing as a maritime shipping market. Rather, it is a sum of different sub-markets, organized along the type of commodities/customers needs/ship types in bulk shipping, contributing to competition between the sub-markets; and the trade lane in the container liner shipping. The top three trade lanes, in terms of percentage of the global trade deployment by cellular TEU capacity, published by Alphaliner (2019), are Far East-Europe: 20%, Far East-North America (Transpacific) and 16%, Latin America related: 13%.

The comparison of the degree of concentration measure also clearly shows that the market structure is different. Both bulk segments are evolving toward perfect competition (i.e., less concentration, more competitive), while container shipping shows an increasing concentration trend.

Fig. 5 lists the four largest ship owners (capacity in service) compiled with data from AXS-Alphaliner and Clarkson Research Service. The restructuring of the COSCO group clearly results in a rise in its rankings in the tanker and container-shipping segment.

Fig. 6 shows the evolution of the HHI, an alternative measure of market concentration. The HHI can be as high as 10,000 (monopoly). The lower the HHI, the more competitive the market. The value of the HHI is meaningless, unless viewed against a benchmark: a value between 1,000 and 1,800 generally indicates moderate concentration, while any value over 1,800 points to a highly concentrated market (Shepherd, 1999; Sys, 2010).

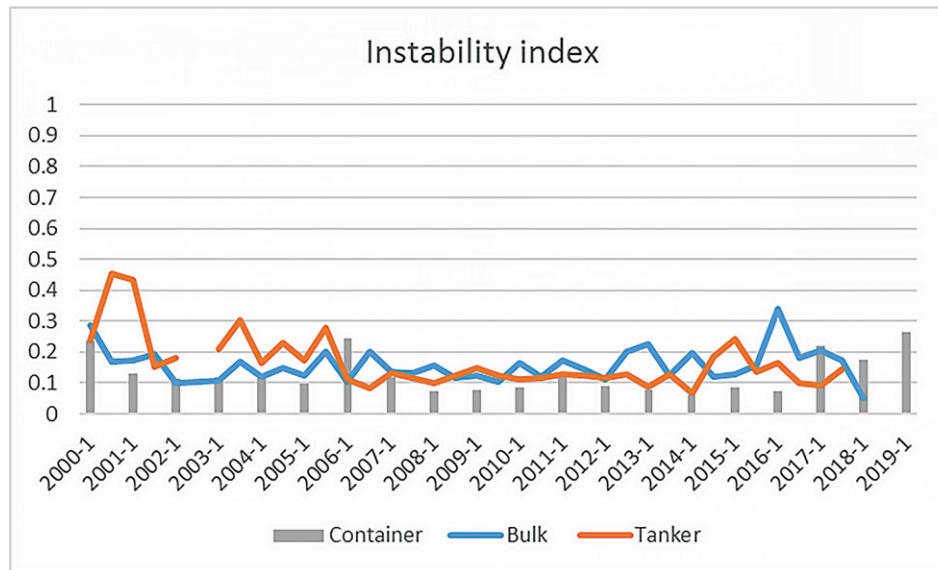


Figure 7 Instability index. Source: Own calculations based on *Alphaliner* (various editions) and *Clarkson Research Service* (2019).

Given the earlier-mentioned guidelines, the bulk shipping industry must still be considered as unconcentrated. The HHI is never higher than the 1,000 limit. In general, a decrease in the HHI points toward a loss of market power and an increase in competition. Comparing with the container-shipping industry, in 2019, the HHI exceeded for the first time the 1,000 limit.

The fact that the tramp is operating in a more competitive market than container liner shipping is also clear from the instability index. This index is calculated with the Hymer-Pashigan or instability index. It is used to indicate the degree of competitiveness. The upper limit of the index is 1, corresponding with fierce competition; while the lower limit equals 0 or no variation in market shares (Fig. 7). Intensified competition among carriers corresponds with mergers and acquisitions (certainly in container liner shipping), the acquisition of Maersk's VLCC fleet by Euronav (ranked 14th in 2014, 6th in 2015) or the new division of tasks within the COSCO group (clearly visible for the bulk segment in 2016).

Actors in the tramp industry operating in a highly competitive environment will behave differently from ship owners/shipping lines in the container liner shipping industry, which faces less competition. From economic theory, companies in a competitive environment will keep their prices down and be as efficient as possible, simply to survive.

Knowing the market structure is important as it is central in strategic decision making and determines the carrier's behavior (profit, freight rate, etc.) (Baumol, 2004; Sys, 2010; Meersman et al., 2015; Lipczynski and Wilson, 2017).

In the next section, the focus is on the freight market, as fluctuations in freight rates (revenue) directly impacts cash flow, not to be underestimated in a particularly capital-intensive industry.

Supply, Demand, and Freight Rates

This section looks at the supply and demand dynamics. By the intersection of demand and supply, the freight rate is determined. Stopford (2009) defines the term freight rate as follows: "amount of money paid to a ship owner or shipping line for the carriage of each unit of cargo (tons, cubic meter, or container load) between named ports," hence covering the transport cost by sea. Chartering narrates to the chartering and operation of the ship.

Tramping

The demand for the most important commodity in the tanker trade (crude oil—left panel) and the bulk trade (iron ore—right panel) is placed against the supply of the shipping fleet. The year-on-year growth rates show a market that is often being affected by excess capacity. Maintaining some balance in the market in the demand-supply is more challenging in the dry bulk segment (Fig. 8).

Shipping indexes are the best way to evaluate how well the shipping market is doing. Different indexes are published in the maritime sector. The two main indexes, discussed hereafter below, are published by the Baltic Exchange and Clarkson Research Service. For the tanker sector, the BDTI/BCTI is the index of the applicable freight rate for the most important trade routes for crude (D or Dirty) and C or Clean (BCTI), which comprises the transport of refined petroleum products. For the dry bulk sector, the Baltic Dry Index (BDI)—a leading economic indicator of industrial production and economic activity, measures the supply and demand for bulk cargo, which is often just one kind of raw material per shipment (based on a weighted average of 11 different trade routes: grain 4, coal, 3, iron ore 3). The Clarksons Average Earnings reflects "a weighted average index of earnings for the main vessel types where the weighting is based on the number of vessels in each fleet sector" (Clarkson Research, 2019) (Fig. 9). Over the period

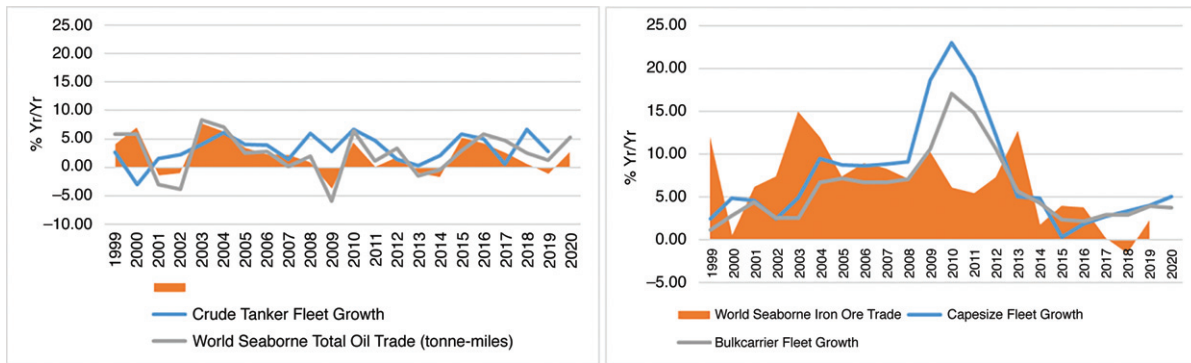


Figure 8 Supply and demand. Source: Own calculations based on *Clarkson Research Service (2019)*.

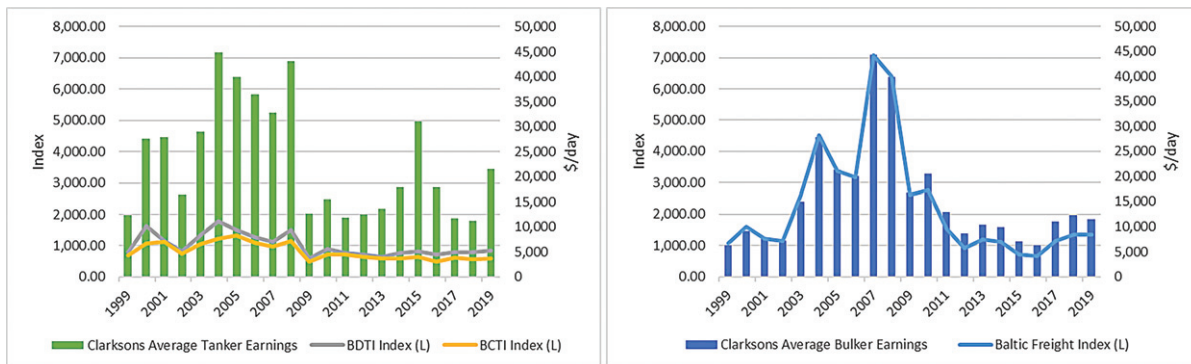


Figure 9 Baltic freight index and Clarksons Average Earnings. Source: Own calculations based on *Clarkson Research Service (2019)*.

1999–2019, the average value of the BFI (daily published) is 2,266 (with a standard deviation of 2,028 points). The Baltic Dry Index is more volatile compared to the indices commonly used for the tanker industry. In 2008, the BFI noted a historical record followed by a significant fall. Back to the new normal?

The variability of the freight rates for the tankers can be expressed in timecharter rates (US\$ per day) or in worldscale (W100–US\$ metric tons). The latter is used for voyage charter. The term “worldscale” refers to the cost of transporting a ton of cargo using the standard vessel on a round voyage (Stopford, 2009). A worldscale above 100 (spikes in 2000, 2004, and 2008) suggests a greater willingness to pay by the charterer, while a worldscale below 100 is rather the question of how much a ship owner is prepared to lose. Fig. 10 shows the evolution of the timecharters for different ship types (line), and the evolution of a voyage charter for VLCC (right panel), and iron ore (left panel)(column bar). In contrast to the VLCC and capesize, respectively, the freight rates and timecharter rates for the handysize, supramax, and panama fluctuate evenly.

Referring back to the megatrends (Fig. 1), geopolitical disruption might be the biggest concern for the capesize and VLCC.

Container Liner Shipping

In parallel to the tramp, here, Fig. 11 gives an overview of the supply-demand relationship (see left hand). Two observations impose themselves. Often supply is outpacing demand, resulting in a downward pressure on freight rates (see right hand) represented by the China/Shanghai Containerised Freight Index (CCFI/SCFI). Moreover, a steady increase in both the CCFI/SCFI and the timecharter rate index (Alpha liner) in 2009–10 can be observed; while the freight indices hold steady and timecharter rates increased in the 2016–18 period.

The findings of the earlier sections, in combination with economic theory, make clear that shipping companies active in the tramp market and therefore operating in a rather high competitive environment (see previous section) will behave quite differently than container shipping companies, who face a lower degree of competition. However, in practice, shipping companies facing competition from many other carriers will have a hard time raising their prices, hence, they need to operate as efficiently as possible, if they are to survive in this sector. To cope with economic, political, technological, environmental disruptions, carriers focus on further cost reduction via innovation.

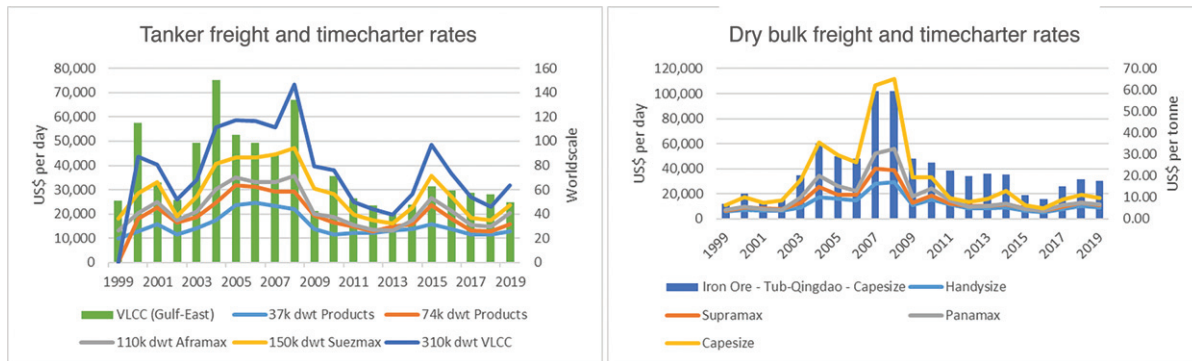


Figure 10 Freight and timecharter rates. Source: Own calculations based on *Clarkson Research Service (2019)*.

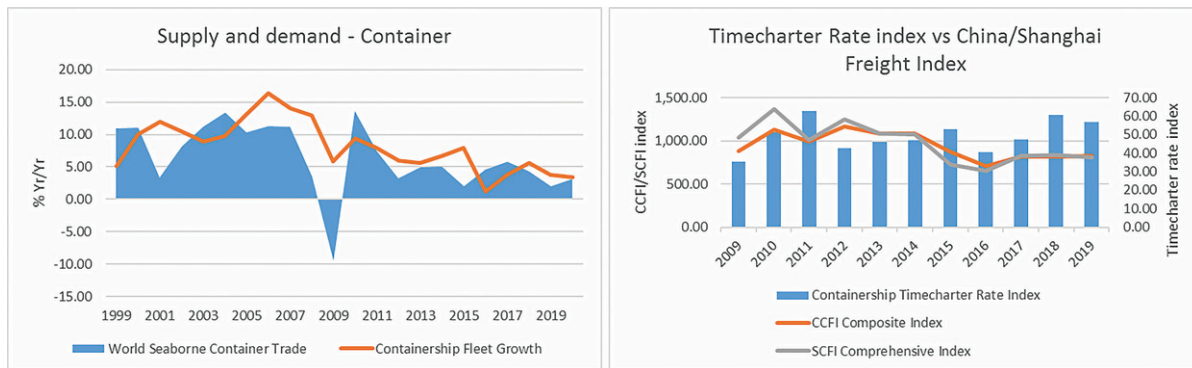


Figure 11 Supply and demand dynamics, impacting freight rates. Source: Own calculations based on *Alphaliner (various editions)* and *Clarkson Research Service (2019)*.

Acknowledgment

The authors want to pay their respects to Prof. Dr. Eddy Van de Voorde and Mr. Honoré Paelinck for their continuous encouragement to study the maritime sector more and more in-depth. Next to this, the authors want to thank Ruben Van Deuren for his support with making the available data congruent.

Biographies

Associate Professor, Christa Sys currently is holder of the BNP Paribas Fortis chair on transport, logistics, and ports at the Department of Transport and Regional Economics. Until October 2013, she was scientific director of the Research Centre on Freight and passengers flows. Next, she is course coordinator for the courses “Business Environment,” “Maritime Economics and Businesses” and “Maritime Supply Chain” at the Centre for Maritime and Air Transport (C-MAT). She also teaches at the Faculty of Business and Economics, University of Antwerp. Her educational activities focus on (operational aspects of) maritime transport, maritime supply chain, transport economics and inland transportation. Her research centers on maritime economics and co-operation and competition in shipping. She jointly graduated as a doctor in Applied Economics at the Ghent University and the University of Antwerp (December 2010). Her doctoral research dealt with the competitive conditions, the concentration and the market structure of the container liner shipping industry (link). This research was awarded by the Section of Technical Sciences of the Royal Academy for Overseas Sciences with the General Manager Fernand Suykens Prize for Port Studies.

Thierry Vanelslander currently is Associate Professor at the Department of Transport and Regional Economics. He graduated as a doctor in Applied Economics at the University of Antwerp. Until 2013, he was holder of the BNP Paribas Fortis chair on transport, logistics, and ports. Until halfway 2009, he was director of the Research Centre on Freight and Passenger Transport, hosted by the Department of Transport and Regional Economics. He is currently course coordinator for the courses “Management of Innovation and Technology” and “Port Economics and Business” at C-MAT, and “Transport Economics” at the Faculty of Applied Economics. His research focuses on business economics in the port and maritime sector, and in land transport and urban logistics. His PhD dealt with co-operation and competition in seaport container handling. He is also the chair of the World Conference on Transport Research (WCTR) Special Interest Group A2 (Ports and Maritime), and topic area manager for WCTR’s track A (Transport Modes). Equally, he is the chair of the Freight & Logistics group within the European Transport Conference.

For photo see

- Christa Sys: <https://www.uantwerpen.be/en/staff/christa-sys/>
- Thierry Vanelander: <https://www.uantwerpen.be/en/staff/thierry-vanelander/>

Permissions

Every author has a responsibility to gain permissions for any figures/tables they wish to include that have already been published. This also includes artwork that have previously been published by Elsevier, as although there will be no charge permission still needs to be gained. To request permission please go to: <http://www.copyright.com/>

Please be aware that we will require your manuscript in an editable format (e.g. MS Word or LaTeX/ PDF) not a PDF file alone. If you encounter any problems please contact Elsevier at MRW-TRNS@elsevier.com

See Also

Maritime Economics: Organizational Structures; The history of Maritime Transport; The Geography of Maritime Transport; The Future of Maritime Transport; Maritime Route Planning (Pax/Freight); Maritime Company Freight Management/Bulk Industry

References

- Alphaliner, 2019. (Various editions). Alphaliner Monthly Newsletter.
- Baumol, W.J., 2004. Welfare economics and the theory of the state. In: *The Encyclopedia of Public Choice*. Springer, Boston, MA, pp. 937–940.
- Clarkson Research, 2019. Shipping Intelligence Network.
- Clarkson Research Service, 2019. Shipping outlook. Available from: <https://sin.clarksons.net/>.
- DNV-GL, 2018. Seaborne trade outlook: The energy transition. Available from: <https://www.dnvgl.com/expert-story/maritime-impact/Seaborne-trade-outlook-the-energy-transition.html>.
- Hymer, S., Pashigian, P., 1962. Turnover of firms as a measure of market behaviour. *Rev. Econ. Statist.* (44), 82–87.
- Lipczynski, J., Wilson, J., 2017. *Industrial Organization Competition, Strategy and Policy*. Pearson Education Limited, Harlow.
- Meersman, H., 2009. Maritime traffic and the world economy. In: Meersman, H., Van De Voorde, E., Vanelander, T. (Eds.), *Future Challenges for the Port and Shipping Sector*. Informa, London, pp. 1–25.
- Meersman, H., Sys, C., Van de Voorde, E., Vanelander, T., 2015. L'histoire se répète? Current Competition Issues in Container Liner Shipping. OECD Working Party No. 2 on Competition and Regulation, 11.
- Shepherd, W.G., 1999. *The Economics of Industrial Organization* (4th), fourth ed. Waveland Press, Inc, Prospect Heights, Illinois.
- Stopford, M., 2009. *Maritime Economics*. Routledge, Oxon.
- Sys, C., 2009. Is the container liner shipping industry an oligopoly? *Transp. Policy* 16 (5), 259–270. <https://doi.org/10.1016/j.tranpol.2009.08.003>.
- Sys, C., 2010. Inside the box: Assessing the competitive conditions, the concentration and the market structure of the container liner shipping industry (Dissertation, Ghent University). Available from: <http://hdl.handle.net/1854/LU-4399458>.
- UNCTAD, 2019. Review of maritime transport 2019 (p. 129). Available from: Unctad website: https://unctad.org/en/PublicationsLibrary/rmt2019_en.pdf.
- UNWTO (Ed.), 2019. *International Tourism Highlights, 2019 Edition*. Available from: <https://doi.org/10.18111/9789284421152>.

Further Reading

- Branch, A.E., Robarts, M., 2014. *Branch's Elements of Shipping*. Routledge.
- Stopford, M., 2009. *Maritime Economics*. Routledge, Oxon.
- Panayides, P.M. (Ed.), 2019. *The Routledge Handbook of Maritime Management*, Routledge.

The Geography of Maritime Transport

César Ducruet, Justin Berli, Centre National de la Recherche Scientifique (CNRS), Paris, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	517
Mapping Maritime Transport	518
The Evolution of Classic Maritime Cartography	518
The Digital Era	518
Maritime Transport as a (Spatial) Network	520
From Regional to Global Maritime Networks	524
The Maritime Region as a Framework	524
Substructures in the Global Maritime Network	527
Socio-Economic and Territorial Determinants	529
Cities, Regions, and Maritime Flows	529
The Interconnection with Other Networks	530
Conclusion	532
Acknowledgments	533
References	533
Further Reading	534

Introduction

Dealing with the geography of maritime transport has two directions of which only one can be chosen in this article: (1) doing a geography of maritime transport or (2) analyzing the geography of maritime transport itself. Far from proposing an exhaustive review and comprehensive history like in a bibliometric study, we propose in this article to ask one central question: what makes maritime geography different from—or linked with—maritime/transport studies and geography/spatial sciences in general? Although such a question seems to point to a very specialized research arena, it has a wide range of possible answers. We thus selected two major issues that concern many other (if not all) arenas: (1) data visualization and spatial shift as well as (2) network science and complex systems. Such an angle of attack has two merits.

First, it avoids repeating the usual contents of maritime textbooks and official reports, where the evolution of technology, cost, and traffic is already well covered, and of dedicated geography handbooks showing major shipping routes, the ranking of world's top ports, and other features such as traffic distribution by fleet, continent, flag, etc. The majority of maritime geography long remained a descriptive collection of information constantly updated but with no ambition to connect other domains. Despite its pioneering role in mapping world shipping routes based on real data, maritime geography soon became closer to maritime economics and management with a growing emphasis on actors and their strategies, cf. the behavioral turn. But the spatial turn converged with a growing interest for the sea from other disciplines, so that numerous visualizations of maritime transport emerged in the past decade or so. In recent years, several international and/or nongovernmental organizations produced huge reports about maritime transport using cartography, such as the United Nations mapping piracy acts since the mid-1990s (UNITAR) and calculating a Liner Shipping Connectivity Index (UNCTAD), the OECD measuring the competitiveness and visualizing the maritime forelands of global port cities, the World Bank displaying the connectivity, efficiency, and economic development of Mediterranean maritime networks, and the European Commission Joint Research Center analyzing vessel position data for practical purposes and calculating a composite index of global urban accessibility including vessel flows (see a recent review in [Ducruet, 2017](#)). After presenting a critical review of existing visualizations of maritime transport, we propose a new methodology based on the new software Geoseastems.

Second, it is interesting to compare how maritime geography had been adopting dedicated analytical tools such as network analysis, from the early times up to nowadays. Is maritime transport a specific network? What is the influence of its spatial nature on its structure, morphology, and topology? In this approach, the network concept will be extended to other dimensions such as firms networks in the maritime sector as well. This approach thus places the geography of maritime transport, should it be performed by geographers or not, at the center of multiple arenas such as corporate studies, urban studies, network studies. To reach such an objective, recurrent topics such as port–city relationships, technological evolutions, traffic distribution, hinterland connectivity, port governance and supply chain integration will be inevitably dealt with, to name but a few. Case studies as well as general models are selected to further illustrate what is the geography of maritime transport.

Mapping Maritime Transport

The Evolution of Classic Maritime Cartography

It is a fact that only a dozen maps of world shipping flows have been produced until the years 2000s (Ducruet, 2015). Historically, maritime maps serve overseas explorations and are strongly impregnated by human imagination on sea monsters and mysterious places. In the 19th and early 20th centuries, such maps gained in precision, indicating major and minor sea lanes, and sometimes the services of specific carriers, notwithstanding information on natural conditions. The emergence of flowmaps only occurred in the 1940s from French and US geographers representing global maritime flows based on historical archives and customs data, employing, at the time, state-of-art technique of flow visualization.

In the mid-1970s, US authorities provided a more systematic cartography of shipping flow density based on a world square grid. But filling cells with the number of vessel trips was not the most efficiency method compared with nowadays tools. Another exception is a PhD thesis from Le Havre University mapping the main interport container flows in the early 1990s. All other maps before the 2000s remained broadly estimated drawings of major maritime routes, published in related textbooks, but not based on actual data. Exceptions are traffic maps showing maritime trade volume per continent and region but not between them. Old data such as ship logs were used to produce similar maps but with a very different purpose: the reconstitution of past weather conditions across the globe (the CLIWOC project). Text-based sources and archeological records helped to reconstitute Roman and Greek routes of the Antiquity, while local port registers allowed to map 18th century French trading routes (the Navigocorpus project), Venetian trade (1283–1453), slave voyages from Africa to the Americas (1514–1868), and an animation of British Royal Navy ship movements (1913–1925) using CartoDB. Other known sources on past maritime transport remain to be systematically exploited, such as the Suez Canal archives (1865–today), the Sound Toll Registers (1497–1857), and the Dutch East Indian Trading Corporation with 25 million ship records (1595–1795).

The discovery of Lloyd's List publications on global and daily vessel movements, covering approximately 80% of the world fleet and published since 1880 (1696 for the British fleet only) opened the path to plenty of previously unthinkable analyses, long-term and at the global scale. For instance, mapping the global maritime network at the continent level between 1890 and 2010 clearly confirmed the shift from Atlantic to Pacific, with the drawback of using vessel calls instead of vessel tonnage in recent years as ships grew in size and rationalized their services (Fig. 1).

About maritime transport as a firm or alliance, mainly French and Canadian geographers offered, from the mid-2000s onwards, maps of individual ocean carriers. Such a shift to more managerial aspects was motivated by the growing bargaining power of large companies over the entire supply chain including port and terminal operations as well as logistics. Data were usually obtained directly from websites or from yearbooks such as Containerization International, detailed at a given point in time the scheduled services of each company. One of the first maps of the kind showed the service geography of Evergreen and COSCON in 1990 and 2000 with the main ports of call. Another study compared the global network of the largest shipping lines in 2002 based on no less than 601 services and 440 ports, but the maps were aggregated by continent. This helped to categorize firms and better understand their mainstream or niche strategies according to their geographic coverage and their nationality. Later on, parent studies followed, such about COSCON in 1990 and 2000, mapping its services related with China, to discuss competition and integration trends, and about the historical development of CMA-CGM and Maersk, questioning the respective roles of public and private forces conducting to routing choices, with a growing emphasis on market power and efficiency.

Finally, some types of studies looked at vertical integration, cooperation, and interfirm agreements while others looked at maritime firms through the localization of their headquarters and establishments, the local determinants of maritime advanced producer services (APS; e.g., finance, brokerage, insurance), should it be globally or at the intracity level, also questioning their colocation with other APS or with commodity traders as well as their concentration in global cities. The modal specialization of cities was measured in Europe based on the portfolio of about 8000 transport companies (Kompass database), discussing the varying importance of vertical integration and the strong national influence. Such studies looked more at the sectoral disaggregation rather than individual companies.

Other disaggregated works relevant to the geography of maritime transport also include the analysis of ship-to-ship interactions, should it be for military purposes (cf. tactical and telecommunication networks) or understanding a particular type of transfer of goods, people, or information as specific points in the sea, considering the ship “as a vehicle for knowing and understanding colonialism, commerce, trade and conflict; embodied and resistant performances, and residual materiality, demonstrating the place of the ship in geography”. Certain geographers got even closer when interviewing captains and mariners to map their mental representation of the ocean, having in common with the former that maritime issues are peripheral in geography.

The Digital Era

Tremendous progress has been made since the early 2000s in the visualization of maritime transport. Numerous online tools serving the calculation of interport shipping distances, (e.g., Sea-distances, Searates), replacing the tedious port-to-port printed matrices. Other tools were designed to measuring intermodal shipping costs as well as the environmental, energy, and economic impact of freight transportation (e.g., GIFT model). Radar and satellite data had been the common material to be used for a wide variety of analyses and visualizations about ship emissions, environmental impacts of shipping, accidentology, navigability, security, vessel trajectories, and coastal management. The latter works remained somewhat separated from the ones presented in the previous

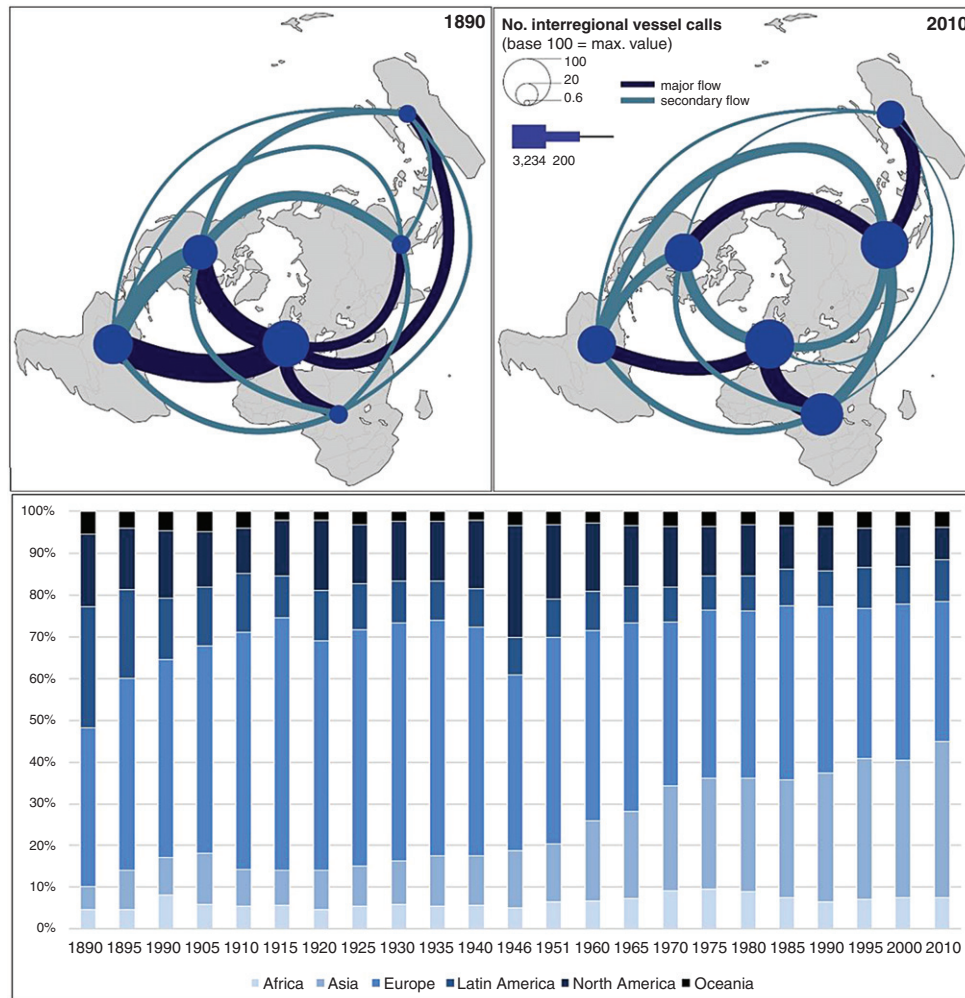


Figure 1 World distribution of vessels calls in 1890 and 2010.

section (social sciences), being more on the side of traffic engineering and operations research. Yet, this digital data could not go back prior the mid-2000s, but gained in popularity with the emergence of numerous online resources of which the now famous marinetraffic.com providing the real time position of vessels, freely accessible.

Numerous maps using such real-time data were thus proposed, for the sake of obtaining a more precise vision of corridors by type of ship for instance (cf. cargo and tanker ships by ShipMix), a multilayered cartography of the Anthropocene (Globoïa), global heatmaps of ship densities (Marinetraffic and ExactEarth). Dynamic visualizations have also been proposed by FleetMon Satellite AIS and FleetMon Explorer, of which the Shipmap project looking at environmental impacts. Other explored issues include the analysis of ships' traces across oceans using Advanced Very High Resolution Radiometer (AVHRR) data from space, noise data from waves and currents of which anthropogenic (e.g., <https://www.oceannoise.com>) using for instance NOAA AIS data. The rise of online mapping solutions also includes the European Atlas of the Seas (European Commission), route-planning algorithms to model maritime paths, the upcoming Google Oceans application, and the OpenSeaMap project with complementary information on shipping lanes.

Despite the mentioned advances, there was little if no possibility to map past maritime flows and thus to understand the long-term evolution of maritime transport. Fig. 2 provides a recent example based on 2008 data of a flowmap and heatmap based on the same information. Each map shows a different dimension; the heatmap is somewhat biased by extreme values (port traffic) in East Asia while the flowmap is more precise in terms of global coverage and the distribution of routes, with lesser concentration. Such visualizations were made possible using a virtual maritime grid based on distance and cost of supporting links, after the world map had been divided into squares through eight iterations and their centroids linked with the impossibility to cross land surfaces.

Maritime Transport as a (Spatial) Network

The application of graph theory to maritime transport started in 1968, in a Canadian PhD thesis about Vancouver (Robinson, 1968), but had only a handful of followers until the late 1990s and early 2000s in isolated theses and articles not even mentioning each other. A rapid count using Google Scholar revealed two facts: the very low number of works mentioning explicitly “maritime network” in French and English languages (i.e., less than 1% of all transport network studies) and the majority of such works using the concept of network in its common sense but without engaging in graph theory and the emerging network science. The main value of applying graph theory to maritime transport had been to provide local or global centrality and connectivity indicators, thereby creating more links with other modes and domains as well as complementing the usual description of tonnage. In this section, we discuss the merits and demerits of such a conceptual and methodological adoption through two main themes: regionalization and globalization; cities and regional development.

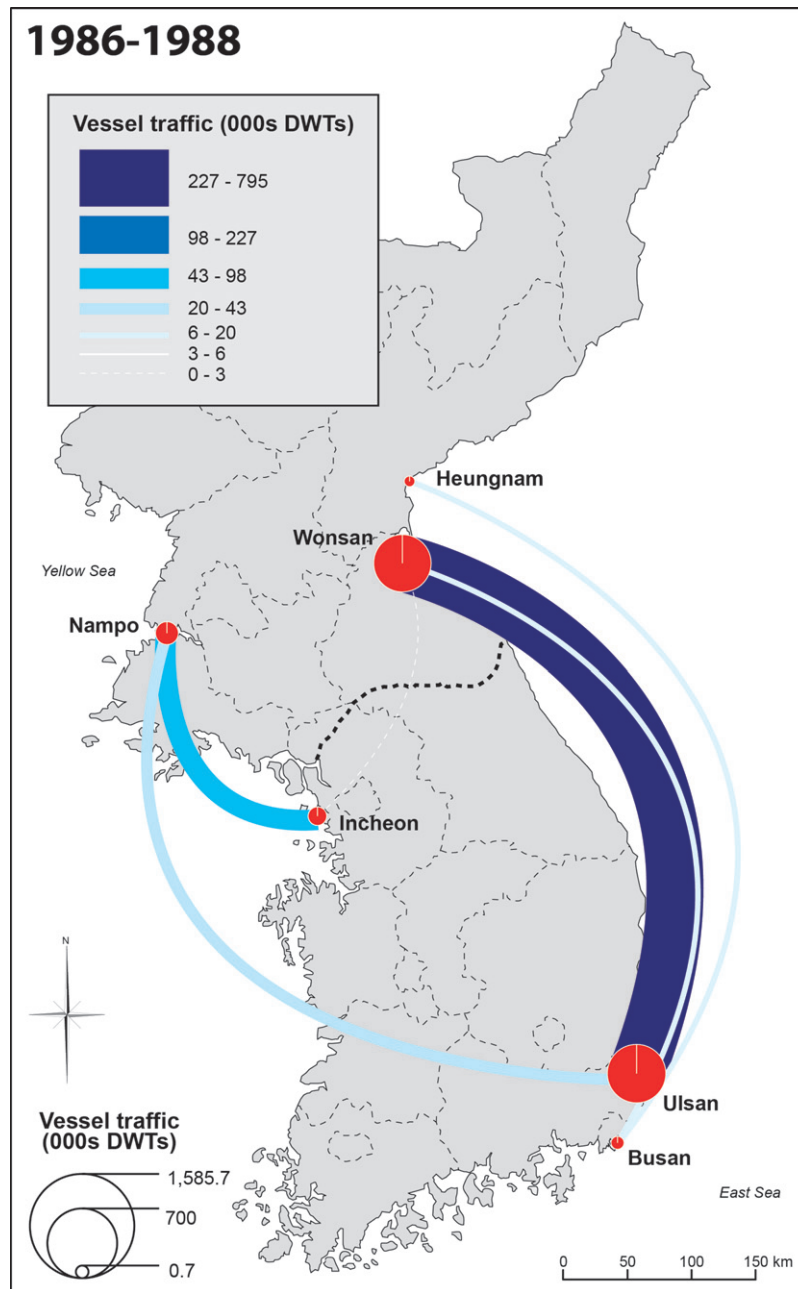


Figure 2 Inter-Korean maritime flows, 1986–2006.

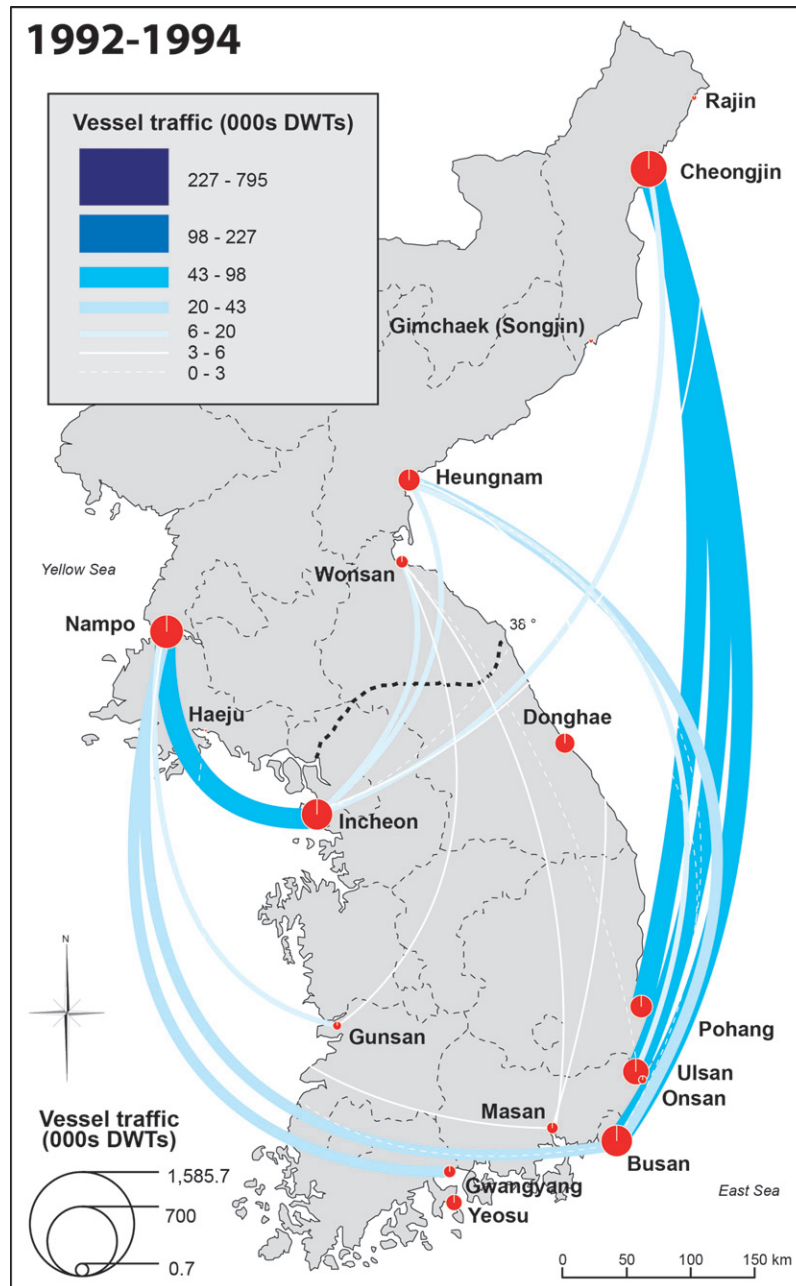


Figure 2 (Continued)

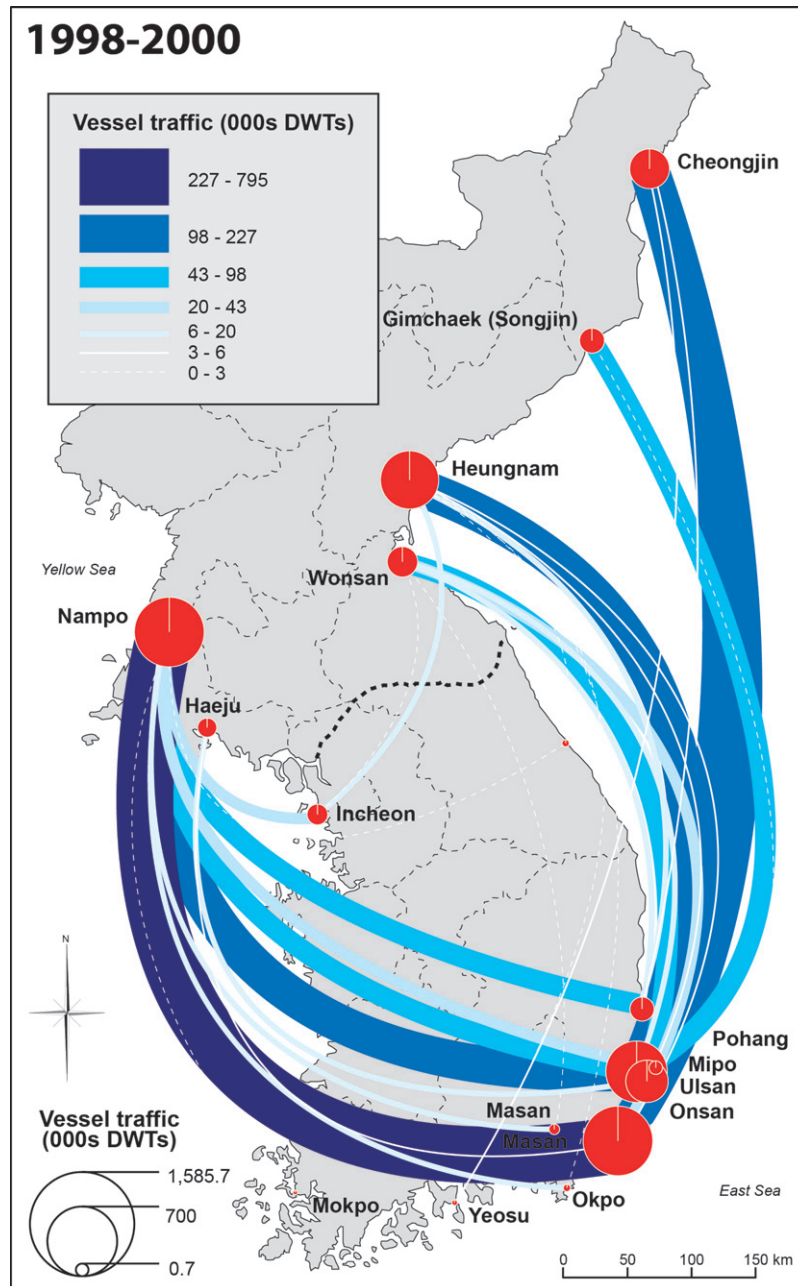


Figure 2 (Continued)

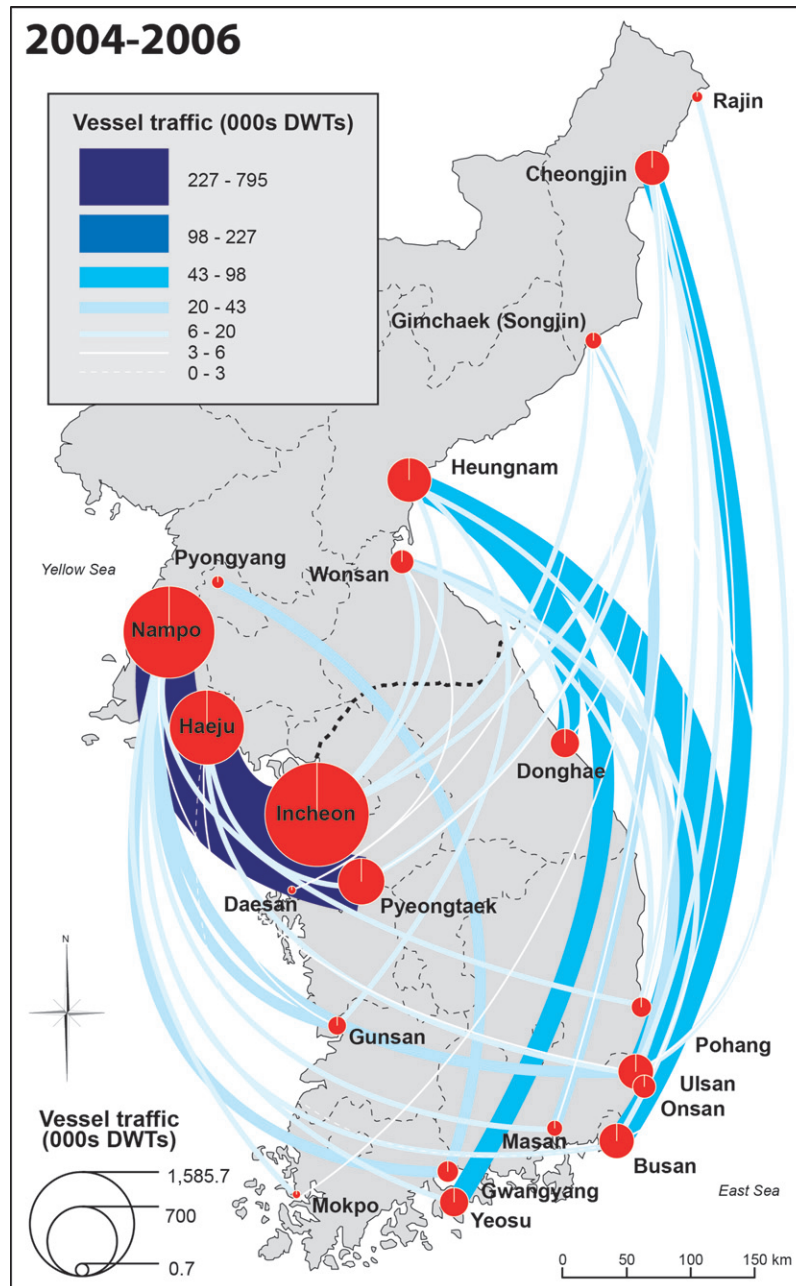


Figure 2 (Continued)

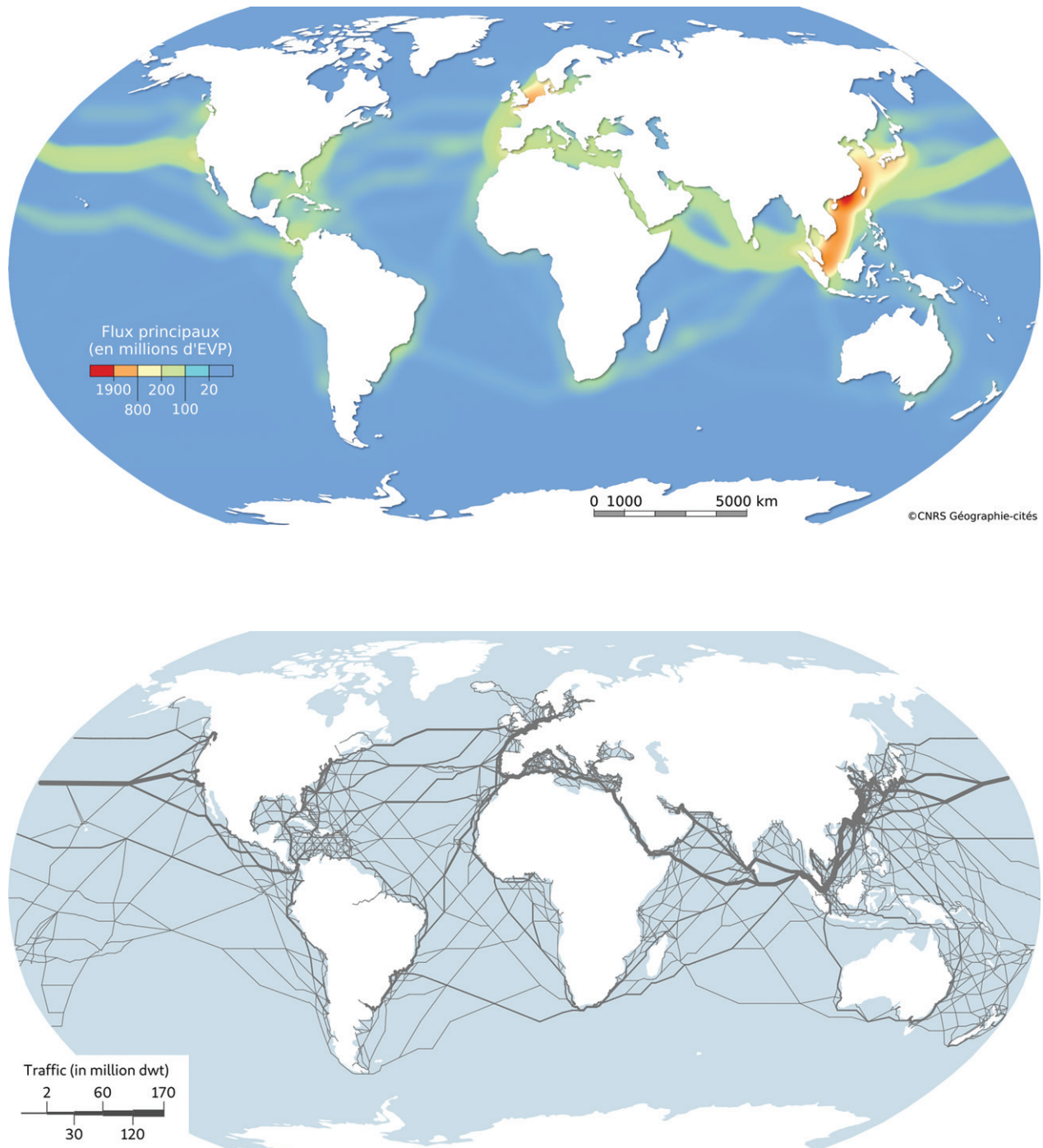


Figure 3 Global pattern of maritime container shipping routes in 2015.

From Regional to Global Maritime Networks

The Maritime Region as a Framework

Choosing a specific part of the world to study connections between ports made the study of maritime transport more geographic. Several studies of the kind emerged in the mid-2000s, about arbitrarily defined regions. This led to the questioning of comparability across regions and of the concept of “maritime region,” often badly defined and corresponding to various realities such as facade, range, seaboard, basin, port system (Lewis and Wigen, 1999). The regional scale is also a way to discuss wider issues of which regional integration, transnational and cross-border studies, and the importance of proximity in the pattern of flows. Yet, such

studies oscillate between understanding the network (actors), the region (territories), or both. A study of port geography papers (Ng et al., 2014) concluded that among them, 48% focus on one single port, 32% on a single country, 14% on a transnational area, and 6% on the world. The national scale has been relevant for countries made of numerous islands, as seen with analyses of the Finnish, Indonesian, and Greek maritime networks, mainly from a passenger and regional planning perspective.

Transnational studies mainly focused on container shipping, with many works on the Caribbean and Mediterranean basins, due to their situation as crossroads and the proximity to Panama and Suez canals, respectively. Network analysis was useful to compare the connectivity of intra- and extra-regional linkages, the role of main shipping lines versus alliances, and the emergence of hub ports. More recent studies explicitly questioned the integration level of specific systems, such as the North European port range from Le Havre to Hamburg, southern Africa, the Canary Islands, the East-West circumterrestrial corridor, and increasingly, the “new” Maritime Silk Road deployed by China through its vision “One Belt, One Road” (OBOR). Comparative studies also emerged, also including North Europe and East Asia, but mainly for the sake of further understanding transport actors, with little citations about regional studies.

1996

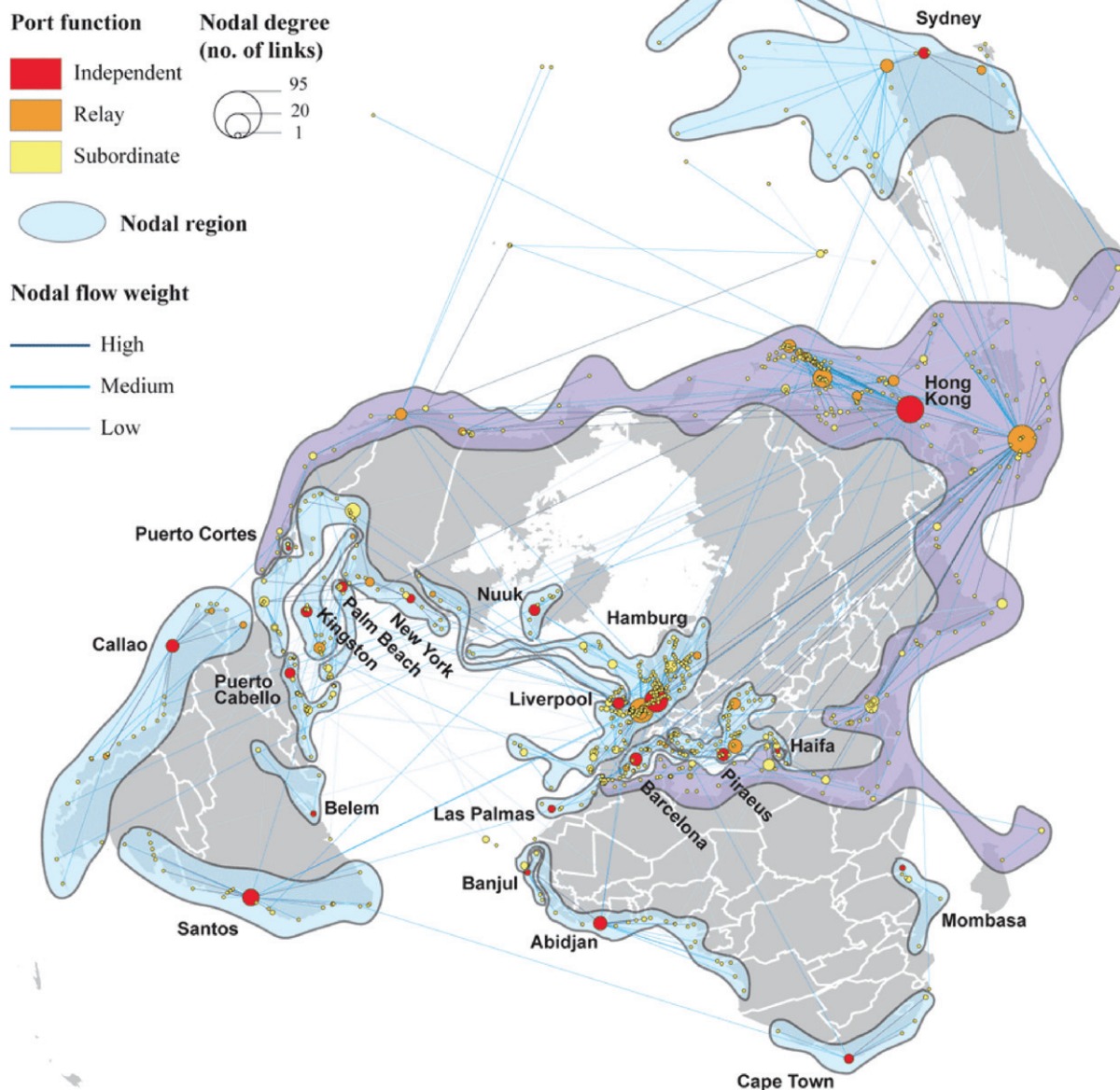


Figure 4 Nodal regions in the global container shipping network in 1996 and 2006.

2006

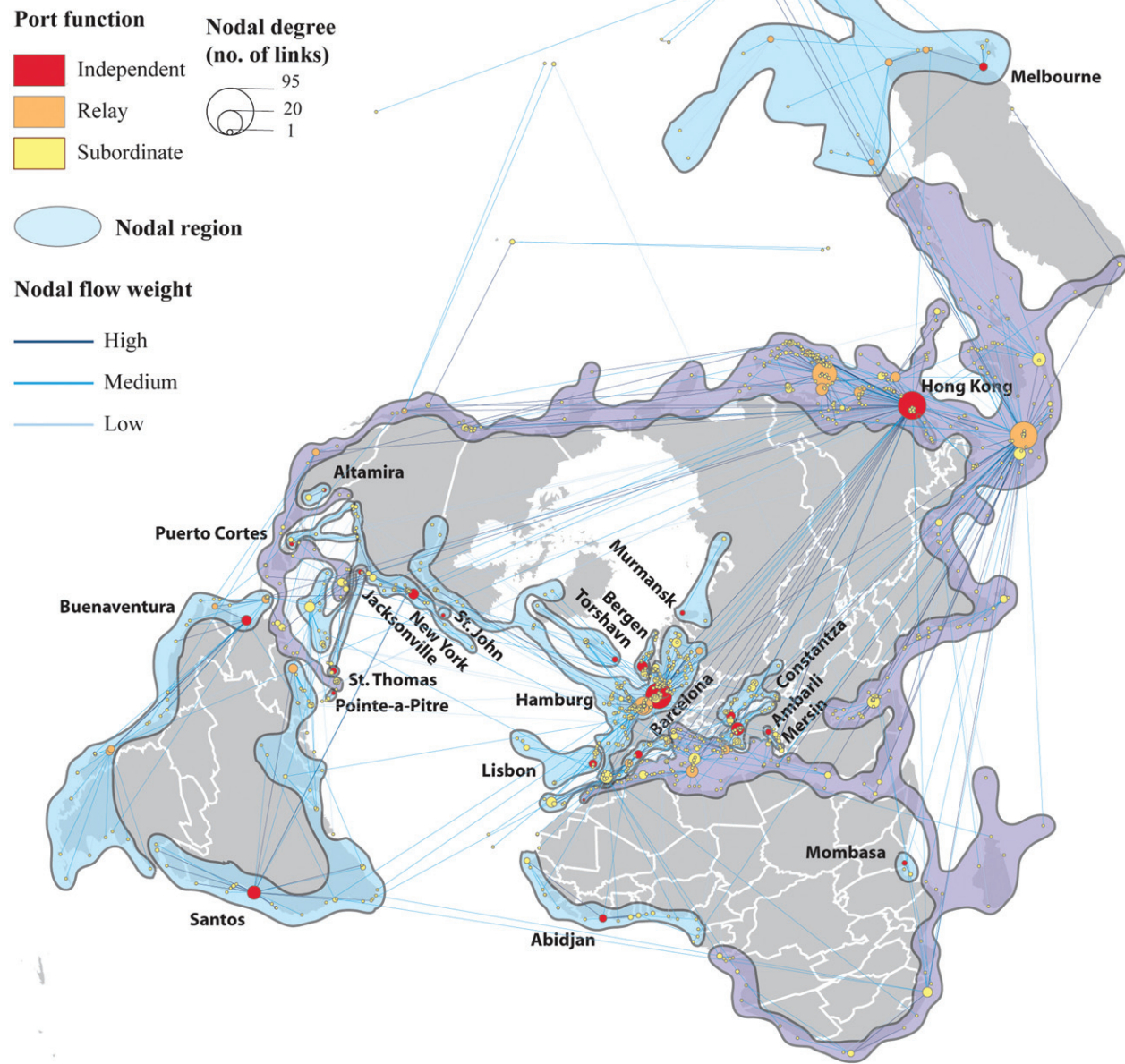


Figure 4 (Continued)

The main outcome had been to observe variations in the level of regional network centralization around hubs, from homogeneous to polarized or polycentric maritime regions; some of them being more or less under the influence of alliances, hubs, and related dynamics such as the rationalization of services. Another result was that the multiplication of nodes and links could mean both regional integration and transshipment hub centralization, depending on the context. Nevertheless, such regional studies often faced the difficulty of disentangling flows of various scale (e.g., short-sea shipping, trunk lines, and feeder) as well as territorial vs supply chain effects.

Another outcome had been to reveal the influence of political borders and other “barrier effects” on the changing distribution of regional maritime networks. First came a series of works on North Korea, revealing its shrinking maritime foreland due to internal and external crisis following the end of the USSR and socialist block; the multiplication of inter-Korean shipping linkages following the 2000 presidential summit and the 2004 maritime agreement, and even the concentration of most North Korea’s traffic between the main ports of the respective capital cities, Nampo and Incheon, before North Korea’s connections became, in the late 2000s,

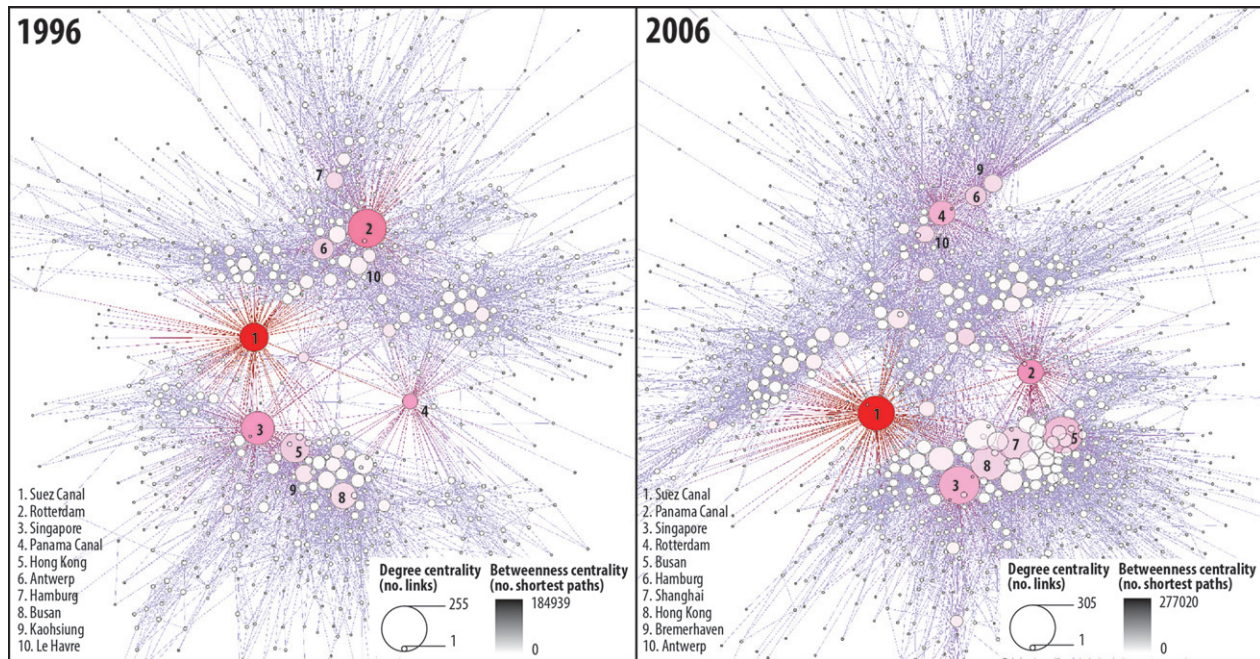


Figure 5 Graph visualization of canal and port centrality in 1996 and 2006.

dominantly Chinese (Fig. 3). Similar maps were proposed to study the interplay between border effects and shipping patterns in Maghreb, Northwest Africa, South Asia, and the Soviet Union, at times of crisis and independence, showing important shifts overtime. In central Maghreb, ports of the three countries (i.e., Morocco, Algeria, and Tunisia) remained disconnected in the mid-1990s. Ten years later, Algiers somewhat increased its dominance, albeit mainly domestic, until the opening of the new hub Tangier-Med (2008) attracted all the regions' main flows for transshipment, competing with the other external hubs of Algeiras and Marsaxlokk. At the contrary, the analysis of the Northeast Asian liner shipping network since the mid-1990s questioned the rise of Shanghai as Asia's new hub. The application of single linkage analysis demonstrated that South Korea, and in particular Busan port, remained dominant in the region despite rapid Chinese port growth and surpassing Hong Kong, due to its logistics hub strategy and the fact that China is still hampered by customs procedures and other political factors that stand in contrast with its enormous potential.

Substructures in the Global Maritime Network

The search for substructures (also coined communities, clusters, or subgraphs) is one major pillar that had been applied successfully to maritime transport. Numerous algorithms dissect the network into groups of nodes, this time without an arbitrary definition of the "region," but still trying to delineate meaningful elements, where geography, politics, and economics also create barrier effects. Physicists were the first to apply modularity to the global networks of tankers and container ships, confirming a strong influence of spatial proximity on the delineation of clusters, their initial goal being to further understand marine bioinvasions. Chinese geographers were the first to reveal the world's container hubs and their respective "regions" where they dominate; they had many followers. Single linkage analysis applied to the global container shipping network in 1996 and 2006 (Fig. 4) confirmed the importance of geographic proximity as most subnetworks were in fact ports in proximity of one local hub. However, one exception to this was the giant "nodal region" centered upon Hong Kong, extending towards both North America and Europe, and including the whole African continent in 2006, mainly driven by the "China effect." Another way to see such changes was to map the entire network and the centrality of port nodes (Fig. 5). The major change observed was the growing centrality of Asian ports and the growing bypass of the interoceanic canals to connect Asia with Africa and Latin America more directly, with increasing transversal linkages between the world's two main poles, Asia-Pacific and Euromed-Atlantic.

With the prime goal to check the multiplex (or multilayered) structure of the global shipping network, the application of single linkage analysis to five layers of maritime flows (containers, passengers and vehicles, solid bulks, liquid bulks, and general cargo) confirmed that not only more diversified ports are larger and more central more faraway connected than more specialized ports, final results also concluded that shorter linkages are more diversified; mostly container ports dominate their respective nodal region, and the latter regions tend to specialize in one main traffic type (Fig. 6). Such invariants largely confirmed stylized facts found in urban studies for instance. Other clustering algorithms helped revealing the other side of the coin: the permanency of "weak links", outside main hubs, for cultural and historical reasons. The Bisecting K-means algorithm applied to the Atlantic network revealed tight but invisible links between the Iberian Peninsula ports and Latin American ports, while the topological decomposition algorithm demonstrated that in the worldwide network, the intercontinental linkages between small and medium-sized ports always

contained one or more European ports in addition a port of the former colonial world, showing the traces of past empires. These nonhierarchical subnetworks carry a lot of questions about the volatility of the supposedly economically driven containerization. Links could be made with key notions of resilience, robustness, and proximity in geography and other disciplines.

The long-term analysis of the global maritime network (1890–2010) confirmed the shift from a polycentric and dense structure (colonial times) to a bipolar and centralized structure (post-1950s), together with a growing simplification or rationalization of the network (Fig. 7). One of the main conclusions was that this centralization upon fewer large hubs overtime had not been caused, but reinforced by, containerization, which is often seen as the turning point in shipping pattern distribution, although similar competition fostered by technological evolution also occurred in the past such as from sail to steam and combustion, long before the current era of mega-ships. The most striking is that the basic material used for such studies had never been used before, although it had been cited two or three times in the works of . . . geographers! The *Lloyd's Shipping Index* allowed to build a global interport matrix comprising nearly 9000 ports, 200,000 linkages, and 850,000 vessel movements. The main conclusion was that the global maritime network shifted from efficiency/robustness to centralization/vulnerability.

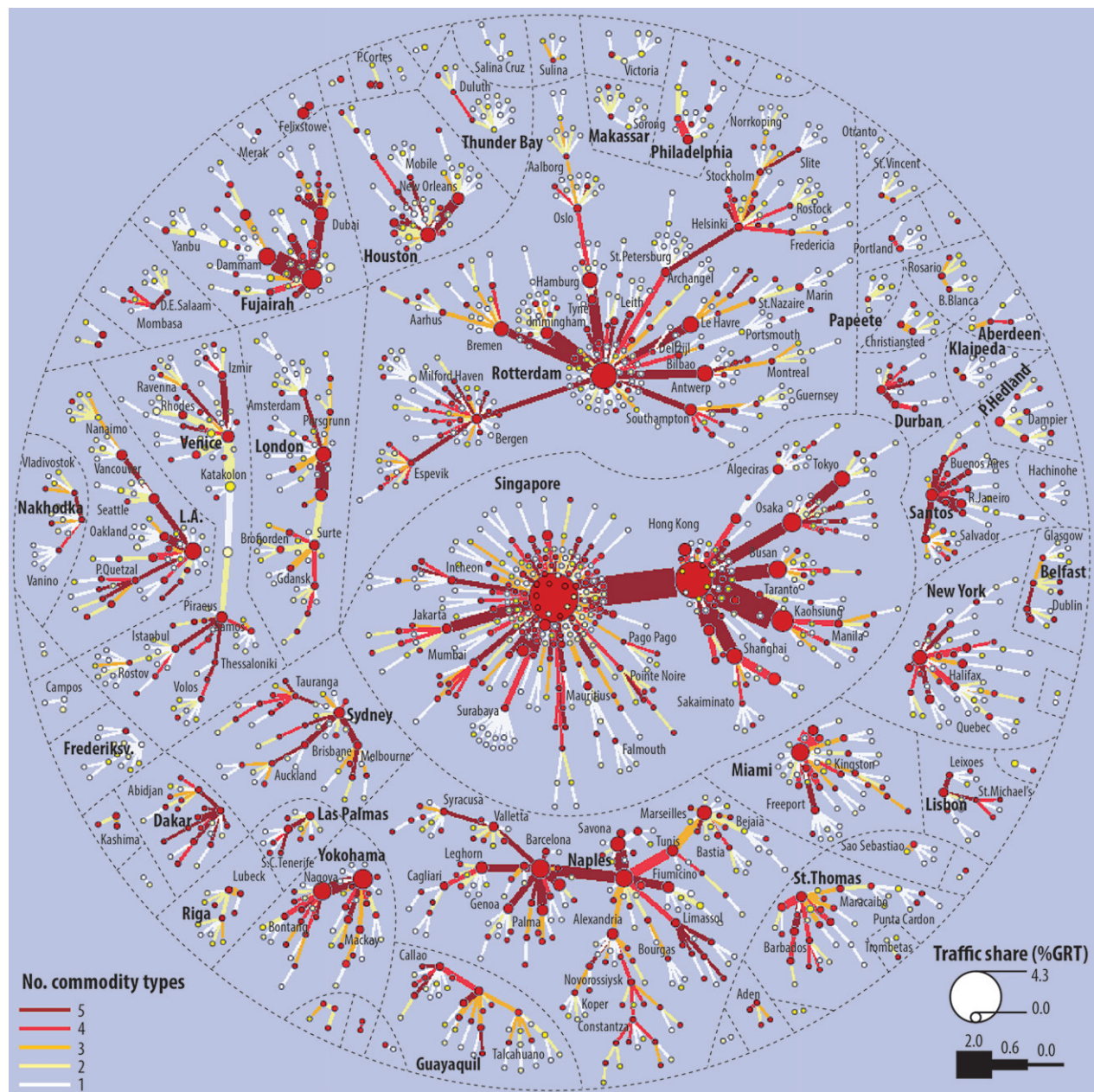


Figure 6 Network diversity and centralization of global maritime flows in 2004

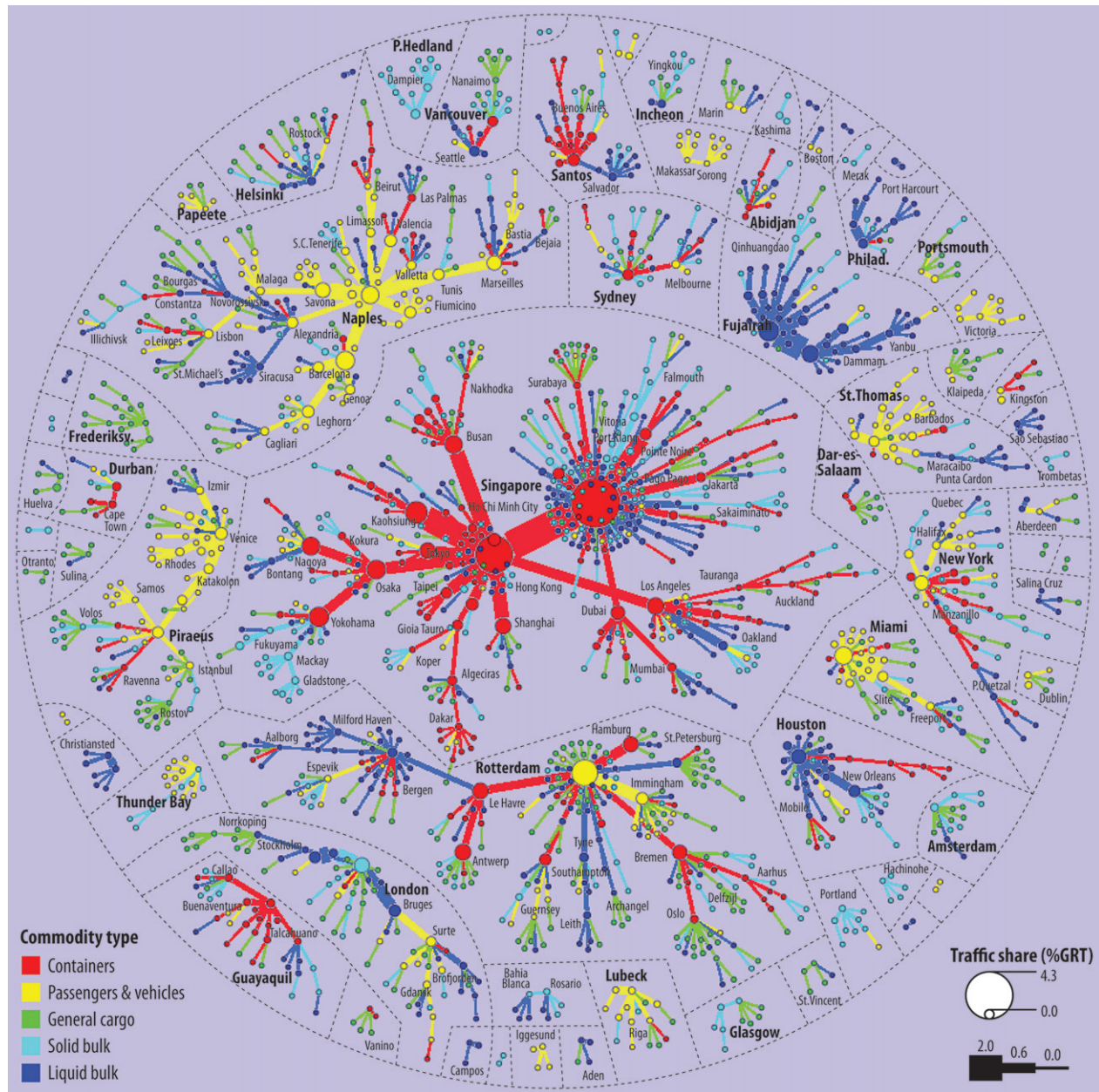


Figure 6 (Continued)

Socio-Economic and Territorial Determinants

The “location” of ports is often given little importance in maritime network studies, as their vast majority only considers the topological space, and the plethora of works about port-city relationships is often theoretical, monographic, without a relational perspective. This section investigates the interplay between maritime transport and two main components of geographic space: (1) cities and (subnational) regions and (2) other transport networks of which the road network principally.

Cities, Regions, and Maritime Flows

Cities and regions have been the focus of many works on commodity chains, production networks, innovation clusters, and so on (Hall and Hesse, 2012), but mostly from the perspective of multinational firms’ networks, airlines, etc. Never have cities been confronted in a systematical manner to maritime flows on a global scale and in the long-term. More likely are models and case studies of port-city separation and waterfront redevelopment, transforming former port areas for new urban uses. A minority of works still demonstrate the positive (economic) impact of ports on their belonged territory, especially in emerging economies (cf. BRICS countries), but most scholars ignore the city or see it as a constraint to port development which should be bypassed due to

lack of space and congestion. In addition, new ports and transit hubs have barely any effect on their outlying environment. Port-related benefits (value, employment) tend to concentrate in locations without necessarily having a port, like inland or coastal nonport cities.

However, the emergence of large-scale quantitative analyses of port cities and regions in the 2000s allowed detecting important linkages, such as the fact that in Europe, North America, and Japan, larger, richer, denser, and service-oriented cities concentrate the most diversified and valued flows whereas regions handling bulks remain often specialized in the primary sector. Such findings motivated a relational study although most of the time, transport and communication networks were studied without proper urban or regional attributes. The first-ever analysis of the world's cities connected by water, based on theories of urban systems, was fruitful in many ways. Mains results (1890–2010) confirmed that the immense majority of world maritime traffic kept concentrating in larger cities, while the latter always exhibited, on average, superior centralities of all kinds, higher rich-club coefficient, higher commodity and ship type diversity, and farther connectivity than smaller cities. Interestingly, the fading linear correlation between urban population and vessel calls observed at port cities per se had, at the contrary, constantly increased at extended urban regions, i. e., cities hosting or not ports but being near the coast, thereby confirming the dilution of hinterlands and the changing meaning and nature of the study unit. This study also confirmed that current global cities like London, New York, and Tokyo had been the world's largest maritime hubs in the 1950s, but lost their dominance to more efficient hubs in the following decades (Fig. 8), such as Rotterdam and Singapore, which are large cities but not the largest.

The Interconnection with Other Networks

Although the idea of studying how different (transport) networks combine dates back to the 1960s in geography, it is only very recently that such a task had been achieved. A growing number of theories and empirical studies of so-called coupled or multiplex networks emerged in natural and social sciences in the late 2000s. Transport studies have increasingly seen maritime transport as one segment only of a wider supply chain comprising other modes, while land transport as a whole often represents about 60% of the

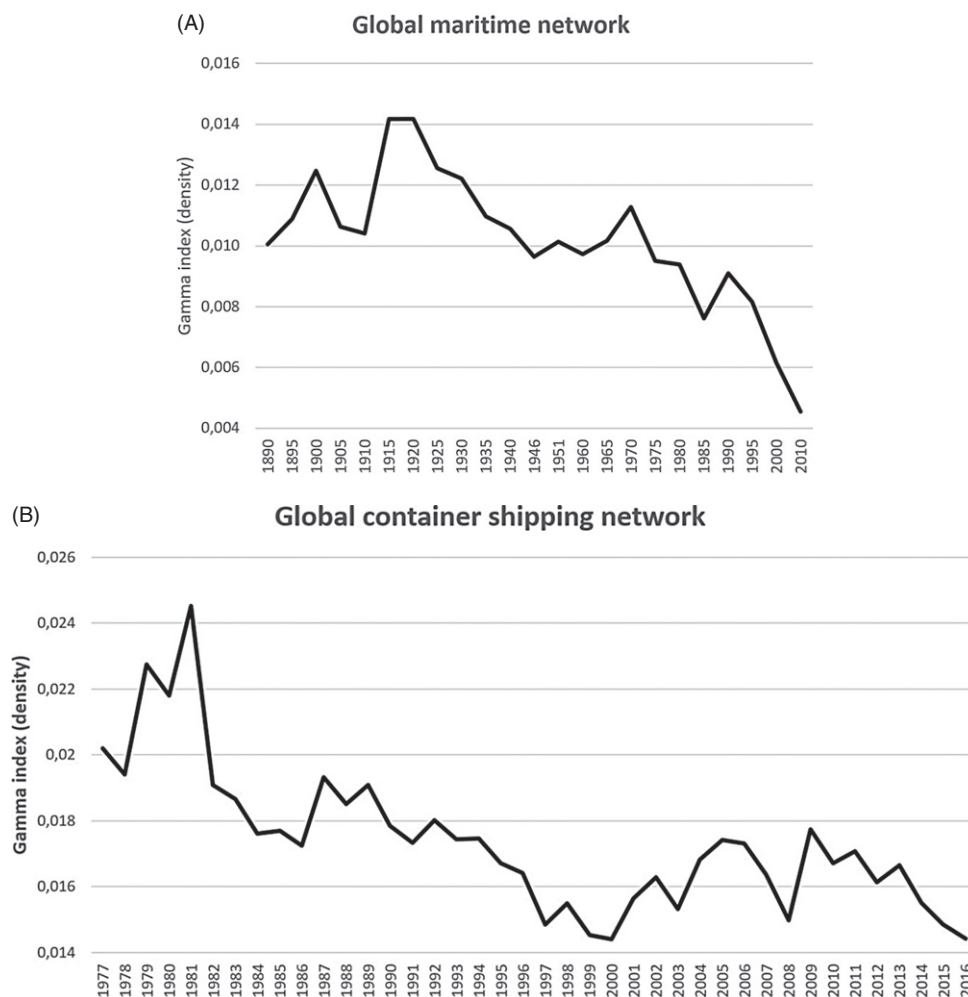


Figure 7 Density evolution of the global maritime network, all vessels (A) and containerships (B).

total cost of moving a container from origin to destination. Physicists and geographers first analyzed the coupling or “intersimilarity” between global airline and maritime networks. The online GIS ORBIS allows calculating the time and cost of travelling through the sea-land network in Roman antiquity. All other studies of the kind did not include the maritime mode, such as airline/Internet, airline/multinational firms, and airline/road/rail to name but a few.

We present in more detail recent efforts to push further the analysis of sea-road networks. One major obstacle being the absence of data on road (freight) flows, except from certain customs data or commodity flow surveys in developed countries. Such an approach having a global ambition, it relies on maritime flows on the one side and road infrastructure on the other. Raw data from OpenStreetMap was modelled and thus simplified taking into account the quality, rank, and speed limits of street segments to construct a global road network connecting both ports, port cities, and nonport cities inland and along coastlines. Applications were made on Australia and Europe (Fig. 9). In both cases, it revealed that combined land-sea centrality is always more correlated with urban population than single centrality.

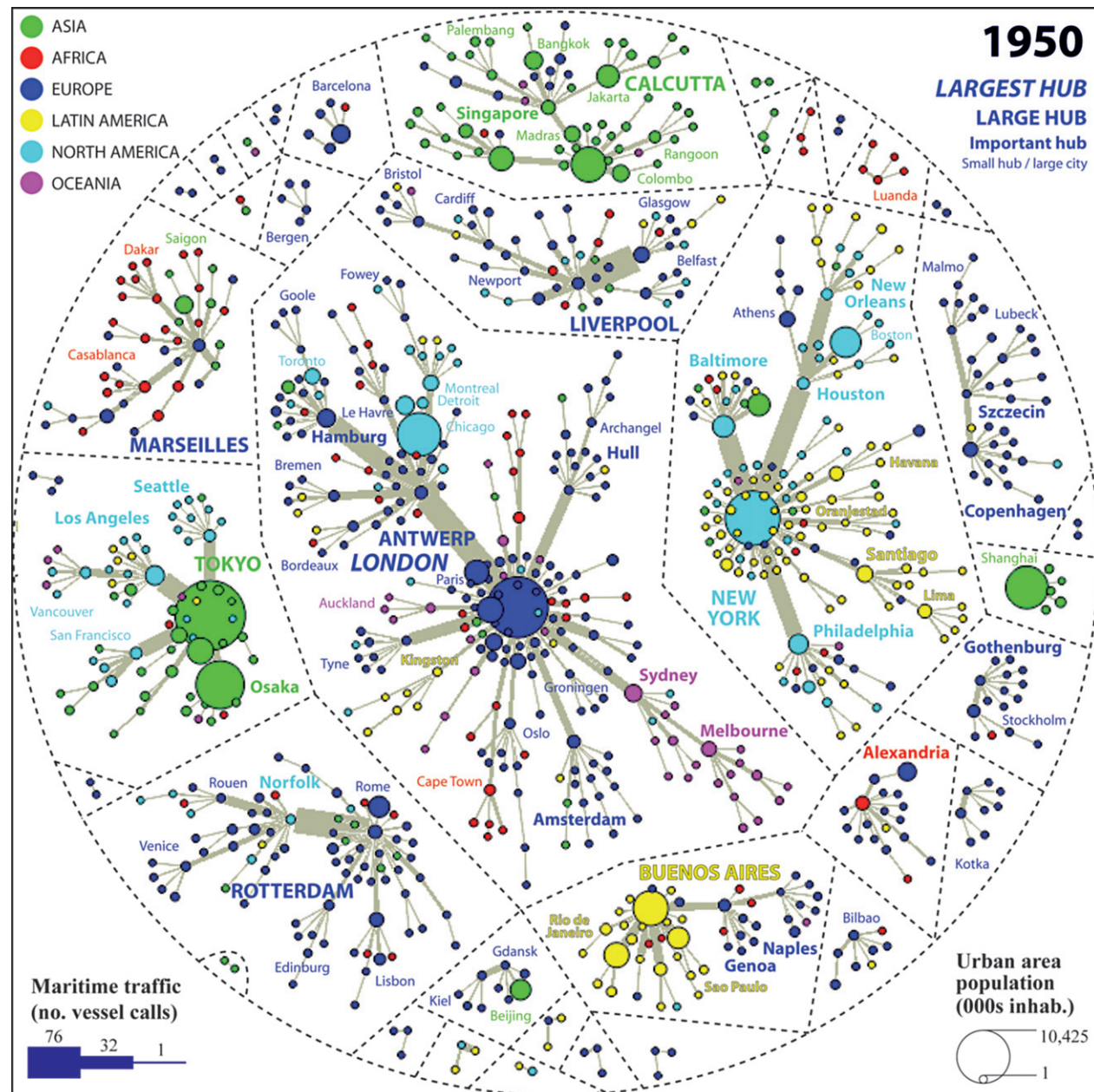
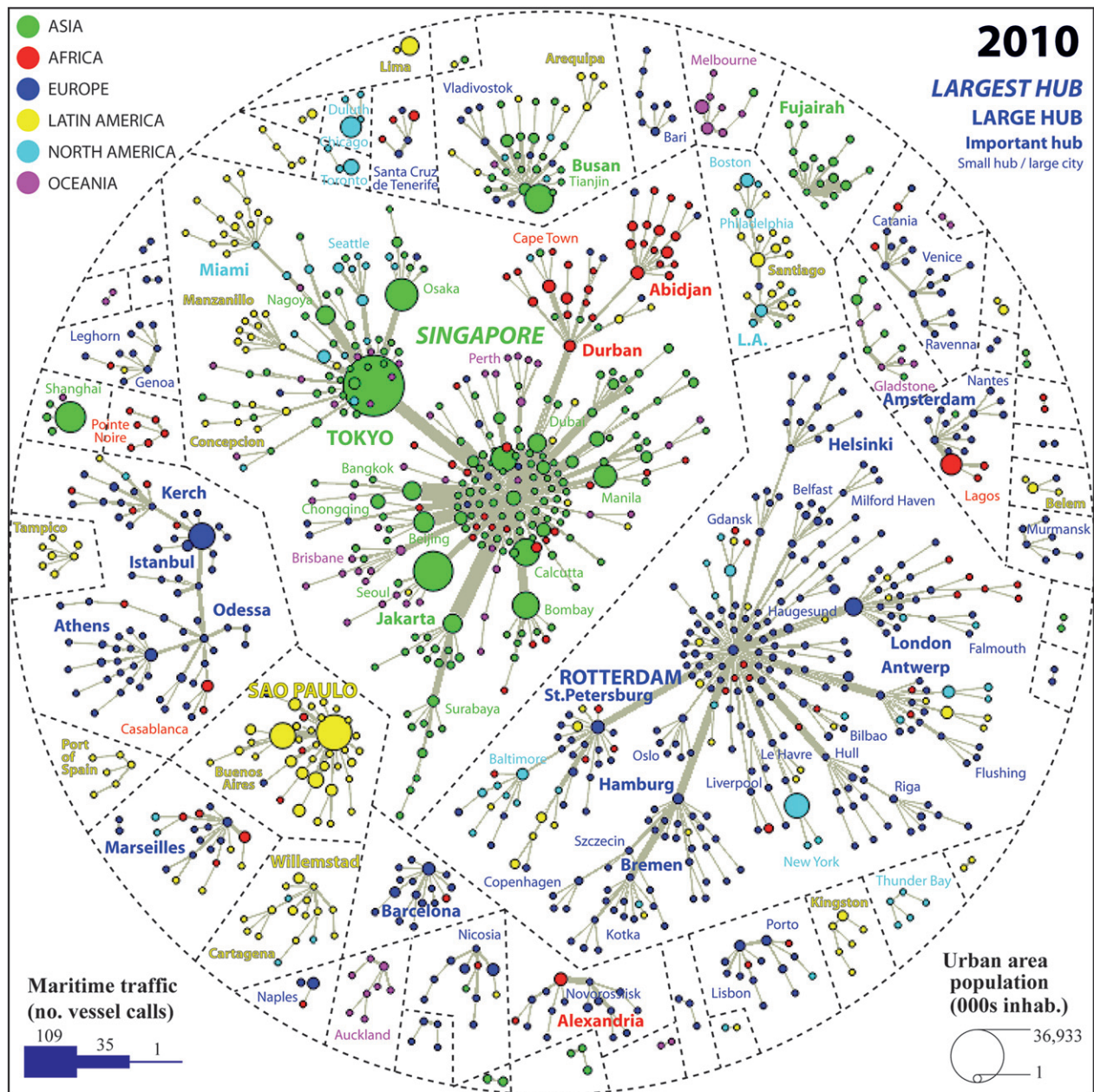


Figure 8 Urban centralization of the global maritime network in 1950 and 2010.



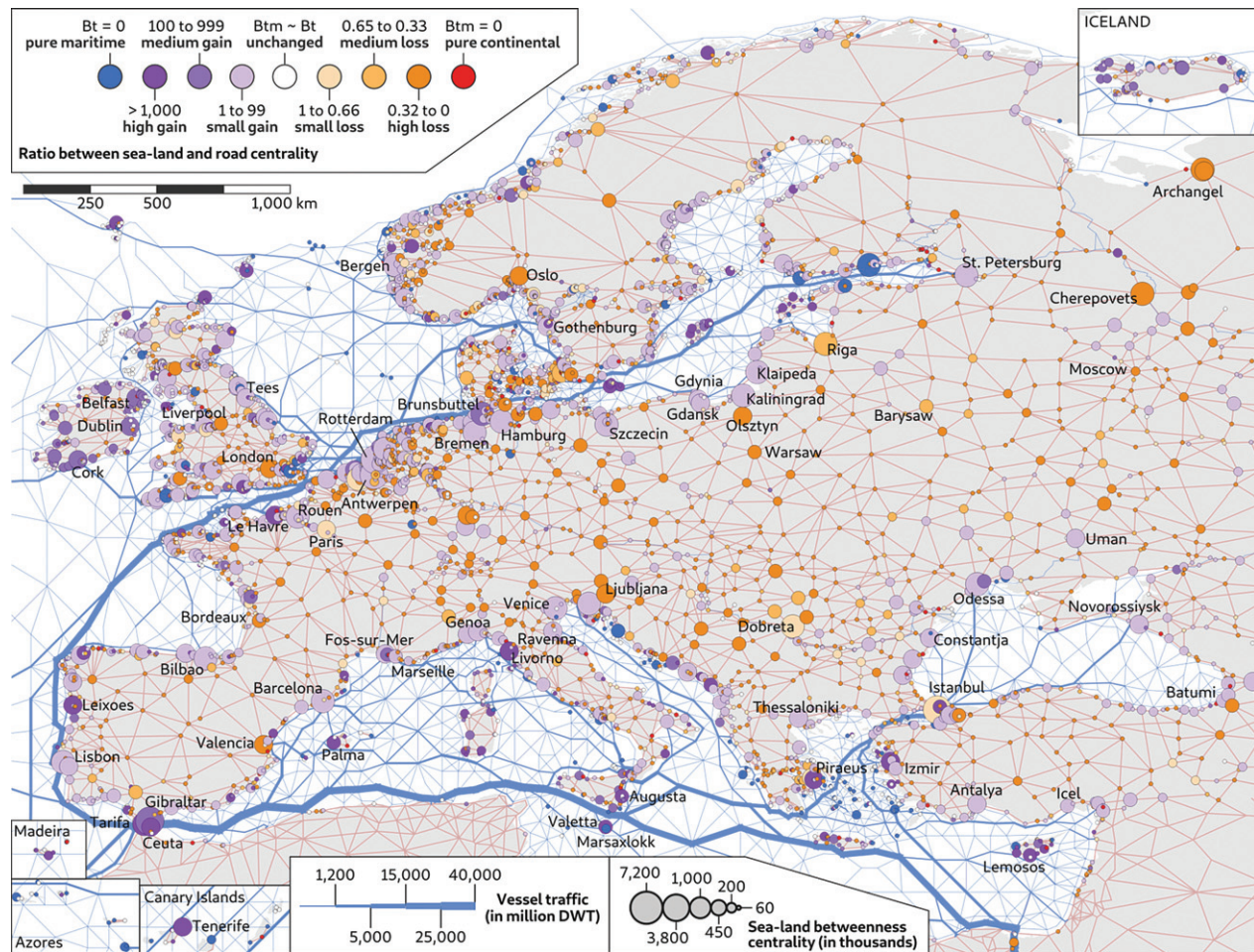


Figure 9 Sea-land connectivity of European ports and cities in 2008.

of space is increasingly taken into account after a period more focused on actors and strategies, the two being not mutually exclusive. Maritime transport has its specificities in terms of industry but it clearly remains, today, a useful way to understand the world from the local to the global.

Acknowledgments

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. [313847] "World Seastems".

References

- Ducruet, C., 2015. *Maritime Networks. Spatial Structures and Time Dynamics*. Routledge Studies in Transport Analysis. Routledge, London & New York.
- Ducruet, C., 2017. *Advances in Shipping Data Analysis and Modeling. Tracking and Mapping Maritime Flows in the Age of Big Data*. Routledge Studies in Transport Analysis. Routledge, London & New York.
- Hall, P.V., Hesse, M., 2012. *Cities, Regions and Flows*. Routledge, London & New York.
- Lewis, M.W., Wigen, K., 1999. A maritime response to the crisis in area studies. *Geogr. Rev.* 89 (2), 161–168.
- Ng, A.K.Y., Ducruet, C., Jacobs, W., Monios, J., Notteboom, T.E., Rodrigue, J.P., Slack, B., Tam, K.C., Wilmsmeier, G., 2014. Port geography at the crossroads with human geography: between flows and spaces. *J. Transp. Geogr.* 41, 84–96.
- Robinson, R., 1968. *Spatial Structuring of Port-Linked Flows: The Port of Vancouver, Canada, 1965*. Université de Colombie Britannique (Thèse de Géographie).

Further Reading

- Berli, J., Bunel, M., Ducruet, C., 2019. Sea-land interdependence in the global maritime network: the case of Australian port cities. *Netw. Spat. Econ.* 18 (3), 467–471.
- Birtchnell, T., Savitzky, S., Urry, J., 2015. *Cargomobilities. Moving Materials in a Global Age*. Routledge, London & New York.
- Frémont, A., 2007. Global maritime networks: the case of Maersk. *J. Trans. Geogr.* 15 (6), 431–442.
- Jacobs, W., Koster, H.R.A., Hall, P.V., 2011. The location and global network structure of maritime advanced producer services. *Urban Stud.* 48 (13), 2749–2769.
- Leymarie, P., Rekacewicz, P., Stienne, A., 2014. UNOSAT Global Report on Maritime Piracy. A Geospatial Analysis 1995-2013. United Nations Institute for Training and Research (UNITAR).
- Rodrigue, J.P., Notteboom, T.E., Shaw, J., 2013. *The SAGE Handbook of Transport Studies*. SAGE Publications.
- Rodrigue, J.P., Comtois, C., Slack, B., 2013. *The Geography of Transport Systems*. Routledge, New York.
- Schwanen, T., 2017. Geographies of transport II: reconciling the general and the particular. *Progress Human Geogr.* 41 (3), 355–364.
- Shaw, J., Sidaway, J.D., 2011. Making links: On (re)engaging with transport and transport geography. *Progress Human Geogr.* 35 (4), 502–520.

The Future of Maritime Transport

Harilaos N. Psaraftis, Department of Technology, Management and Economics, Technical University of Denmark Lyngby, Denmark

© 2021 Elsevier Ltd. All rights reserved.

Introduction	535
The Drive to Decarbonize Shipping	535
General	535
Natural Gas (NG) and Liquefied Natural Gas (LNG)	536
Hydrogen	537
Ammonia	537
The Way Ahead	537
References	539

Introduction

No other mode of transport is as important for international trade than maritime transport. According to [UNCTAD \(2019\)](#), maritime transport carries around 80% of total trade. The strong growth of the global economy, at least after the end of WWII, can be significantly attributed to, and linked with, a similarly strong growth in international trade, which in turn would be impossible without important and ground breaking advances in ship design and propulsion, cargo handling equipment, ports and terminals technologies, and intermodal supply chains, among other factors. Such advances were caused, at least in part, by the ever-lasting drive of the shipping industry to innovate and offer services of competitive cost and quality to shippers.

Maritime transport is very diverse, spanning a very broad variety of different ship types and several segments of shipping markets and services, ranging from tankers to containerships, from passenger to cargo services, from deep sea to short sea routes, and from bulk commodities to unitized cargoes. All this diversity has evolved over a long history that has paralleled the evolution of human civilization, and the ever-lasting drive to explore new worlds, engage in trade, connect remote regions and peoples together, and even wage wars among nations. To state an example, the port of Piraeus container terminal, operated by Cosco Pacific and which moved some 5,650,000 TEUs in 2019, is in exactly the same stretch of water where, some 2500 years ago, the fleet of Themistocles demolished the fleet of Xerxes in the naval battle of Salamis.

Clearly there is as much resemblance between ancient times and today as between an Athenian trireme and a 20,000+ TEU containership. So, making a long-term prediction what maritime transport will look like in thousands or even hundreds of years from now is impossible. A more reasonable question is, can we at least predict what it will look like in the next 30 to 50 years? It turns out that the answer is just as difficult. The purpose of this chapter is to attempt to present an exposition of some of the main factors that come into play as regards the future of shipping, by focusing on a 30–50 year time horizon. It is of course realized that the exposition for such an important topic is to a large degree subjective and incomplete, especially also given the size limitations of this chapter.

There is no doubt that autonomous shipping and digitalization are technological mega-trends that may radically transform maritime transport in the years to come. Autonomous shipping in particular has the potential to radically transform the face of shipping in the future. However, and in spite of much progress, there remain many issues mainly as regards maintenance, resilience, and legal and insurance matters, before autonomous shipping can become the norm. In this chapter, we shall leave these specific trends aside and focus on the *drive to decarbonize shipping*, which is a goal that is very pressing and very much connected to the global quest to combat climate change. So, in the rest of this chapter, we shall discuss the drive to decarbonize shipping and then we conclude with some thoughts on what may lie ahead.

The Drive to Decarbonize Shipping

General

Perhaps no other regulatory development in the history of maritime transport is likely to have a higher impact on its future than the decision of the 72nd session of the Marine Environment Protection Committee (MEPC 72) of the International Maritime Organization (IMO) in April 2018 to achieve by 2050 a reduction of at least 50% in green house gas (GHG) emissions from international shipping, vis-à-vis 2008 levels. There is also an intermediate target to achieve by 2030 at least a 40% reduction of CO₂ emissions per transport work, against vis-à-vis 2008 levels. The so called Initial IMO Strategy ([IMO, 2018](#)) is in the form of Resolution MEPC.304(72), and includes, among others, the following elements: (1) the vision, (2) the levels of ambition, (3) the guiding principles, (4) a list of short-term, medium-term and long-term candidate measures with a timeline, and (5) miscellaneous other elements, such as follow up actions and others.

The so-called 3rd IMO GHG study (Smith et al., 2014) provided updated estimates of CO₂ emissions from international shipping from 2007 to 2012. The 2012 figure, estimated by a “bottom up” method, was 796 million tons, down from 885 million tons in 2007, or 2.2% of global CO₂ emissions. It should be noted that the equivalent percentage reported in the 2nd IMO GHG study (Buhaug et al., 2009) was 2.7% for 2007 fleet data. CO₂ from all shipping was estimated at 940 million tons in 2012, down from 1100 tons in 2007. The reduction from 2007 to 2012 was mainly attributed to slow steaming due to depressed market conditions after 2008. The 4th IMO GHG study, going to 2018 fleet data, is expected to be released in late 2020.

As expected, much of the discussion after the adoption of the strategy in 2018 has focused on short-term measures, and specifically on how such measures can help achieve the intermediate 2030 target, reduce CO₂ emissions per transport work, also known as “carbon intensity,” as an average across international shipping, by at least 40% by 2030, compared to 2008. As this chapter was being finalized, no IMO decisions on such measures had been made. Whatever these measures might be, none of them are expected to change the picture of maritime transport in any significant sense. However, if one stands any reasonable chance to reach the 2050 GHG reduction target, it is clear that some drastic changes should be expected. Among such changes, the use of alternative (low-carbon and zero-carbon) fuels seems the most important instrument.

In fact, it is very clear from the Initial IMO Strategy that such alternative fuels are centrally placed within the IMO decision document, and reference to such fuels is made in many instances in that document. If anything, this reflects the fact that the development and eventual use of such fuels is a matter of high priority for the maritime industry, particularly if the targets set in the IMO Initial IMO Strategy stand any reasonable chance to be reached. Put another way, if fossil fuels continue to be used in a “business as usual” (BAU) fashion, the chances of reaching the GHG reduction targets look slim. In that sense, it is reasonable to expect that low-carbon and zero-carbon fuels could provide the necessary “quantum leap” that might give the quest for decarbonization enough momentum to reach the stated targets.

The subject of alternative fuels is vast, with many studies and projects examining the subject from various twists and angles, and also with specific real-life demonstration projects whose purpose has been to test such fuels in actual ships under specific scenarios in terms of technical and economic viability. See for instance, among others, the work of DNV GL (2018) in the area. For recent surveys on the decarbonization possibilities of such fuels and other technologies and measures, see Bouman et al. (2017), OECD (2018), (de Kat and Mouawad, 2019; DNV GL, 2015, 2018) and LR and UMAS (2020), among others. In this section and due to size limitations we provide only a flavor of some of the pertinent issues, by focusing on alternative fuels. We do this by drawing heavily from Psaraftis and Zachariadis (2019).

Natural Gas (NG) and Liquefied Natural Gas (LNG)

Natural Gas (NG) is about 90% methane (CH₄) and Liquefied Natural Gas (LNG) is about 95% CH₄. CH₄ is a GHG much worse than CO₂ in terms of effect on global warming. The Global Warming Potential (GWP) of CH₄ is not constant, but depends on the time horizon contemplated. Thus, for a time horizon of 5, 10, 20, and 100 years, CH₄ is respectively 116, 110, 86, and 34 times worse than CO₂ (Myhre et al., 2013; Howarth, 2014; Allen, 2014). It is clear that a time horizon of 100 years, being the lifetime of CO₂ in the atmosphere, is not very relevant in the climate change discussion, given the urgency of the situation as regards 2050. The above is an important point because using the proper GWP factor of 86 (instead of 25) immediately reverses all claims of LNG having a better GHG footprint than conventional liquid fuels. We refer to the GHG effect when LNG/ CH₄ escapes unburned to the atmosphere.

Indeed, when LNG is burned in a ship’s engine, it produces about 20%–25% less CO₂ than a conventional liquid fossil fuel, due to its lower carbon content. However, this assumes perfect combustion, which exists only in chemistry textbooks. In real life, a quantity of NG remains unburned and is emitted to the atmosphere together with the combustion exhausts. This escape, also known as methane slip, occurs in sufficient amounts in the vast majority of marine engines (4-stroke and 2-stroke dual fuel or Otto cycle-spark plug-engines). For all these engines, the methane slip—even according to the engine manufacturers’ stated numbers—renders them immediately worse than conventional liquid fuel engines, even when a mild GWP factor of 25 is used (Corbett et al., 2015; EU, 2016). For some latest high pressure combustion engines, methane slip is stated to be very small, but this refers more to the engine laboratory shop-test conditions and not necessarily to measured values in actual ship operation. In any case, multiplying even very small quantities of methane slip by 86, brings these state-of-the-art engines also at par to existing conventional ones, as is also discussed further.

The vast majority of gas engines on ships operate using the Otto cycle (as opposed to the Diesel cycle). Such engines have a published methane leak rate of 3–6 gr CH₄/kwh corresponding to a fuel loss of 2%–3%. Actual measurements invariably show a larger fuel leak rate (e.g., 6–8 gr/kwh in optimal operation and much higher multiples at low loads) (Corbett et al., 2015). These are LPDF (Low Pressure Dual Fuel) engines with pilot fuel and Lean Burn Spark Ignition (LBSI) engines. SINTEF (2017) has measured the methane leak of these engines at 6.9 gr/ kwh and 4.1 gr/ kwh, respectively with much higher values reported throughout the literature especially for LBSI engines. At these methane slip levels, either published by engine makers or independently measured, the GHG footprint of combustion is quite higher than conventional marine diesel engines, even using a 100-year GWP of 25. When a GWP of 86 is used, the GHG performance of NG combustion in marine engines becomes substantially worse than diesel or heavy fuel combustion. Furthermore, there is a limiting trade-off in that, if the engines are operated at a lower air to fuel ratio to reduce methane slip, then NO_x emissions increase exceeding Tier III limits. Already, exhaust after-treatment options are being discussed among regulators, involving methane oxidation/capture devices at the stack.

The latest type newly introduced gas engines (High Pressure Dual Fuel- HPDF) are promoted as zero slip (0.2% of consumption) by their makers (e.g., the ME-GI DF engine by MAN Energy Solutions) although independent references (Corbett et al., 2015) give

0.7 gr/kwh as more common (i.e., 0.4% of consumption). Although this is definitely a big improvement over the Otto cycle engines, we should remember that even just a 1 gr/kwh methane slip represents 86 gr/kwh of equivalent CO₂, and thus any GHG benefit over standard diesel engines is mostly lost.

The above discussion concerns only the LNG methane slip during combustion *at the engine*, which is only one of the areas of methane slip, perhaps the most minor one. Of course, its *lifecycle* methane slip should be considered as well. That involves possible leaks during transportation to consumer and, most of all, leaks during its extraction from the ground, which all currently are either ignored or severely understated from a proper GWP assessment. These leaks invariably occur and they can be considerable.

Hydrogen

What can be a more attractive fuel to burn than hydrogen? Pure hydrogen when burned emits no CO₂, and no SO_x. Thus, it is potentially a viable alternative fuel, provided the CO₂ intensity of its production could be addressed and provided that several technological and safety hurdles could be overcome. About 95% of the world's hydrogen production comes from reforming fossil fuels, mostly NG (Milne et al., 2006; Collodi, 2010). Obviously then the GHG issues described in the previous section on NG and LNG are also applicable to hydrogen, not including the CO₂ emissions in producing it from NG. The most common method to produce hydrogen is by steam CH₄ reforming whereby steam under high pressure, and in the presence of a catalyst, produces hydrogen and carbon monoxide and, with further "water-gas shift reaction," CO₂ and more hydrogen are produced. One ton of hydrogen produced in this process releases 9–12 tons of CO₂ (Collodi, 2010).

The CO₂ intensity of hydrogen production could be reduced by capturing and storing the released CO₂. Nevertheless, its origin, being mostly CH₄, results in the potent GHG effects related to methane's lifecycle (methane slip, etc). Clearly, the carbon footprint of hydrogen produced from NG is higher than that of HFO or MDO (DNV GL, 2018).

An alternative way to produce hydrogen is from electrolysis. Only about 4% of the world's hydrogen production uses that method. However, even this hydrogen cannot be considered as "green," since electrolysis requires large amounts of electricity, which usually comes from the grid. Only if this electricity originates from renewable sources (solar, wind, hydro) or even nuclear, could the hydrogen produced be considered carbon-free. About 55 kwh are required to produce 1 kg of hydrogen at an assumed efficiency of more than 60%. One kwh of electricity, when produced from a coal burning power plant, generates about one kg of CO₂. The US average is about 0.69 kg of CO₂ per kwh (DNV GL, 2015), while China and Russia produce most of their electricity from coal and most other countries from diesel oil. At an assumed worldwide average of 0.80 kg CO₂ per kwh, 55 kwh to produce 1 kg of hydrogen from electrolysis will emit 44 kg of CO₂.

As things stand, the technology required to enable the use of hydrogen as a marine fuel is still under development. Hydrogen is not an easy fuel to handle, transport and store. The boiling point of liquid hydrogen is –253°C (DNV GL, 2018) thus super-insulated (cryogenic) pressure vessels are needed for storage. Depending on the pressure, the size of the tanks needs to be 10–15 times larger than those of conventional liquid fuels (DNV GL, 2015). Also safety issues due to the volatility of hydrogen need to be resolved.

Ammonia

Ammonia (NH₃) is produced from hydrogen by adding nitrogen. Nitrogen is obtained from the air through liquid air distillation or an oxidative process where air is burned and the residual nitrogen is recovered. As such, in addition to the GHG effects of hydrogen and NG (being the primary source of hydrogen), the CO₂/GHG effects of nitrogen synthesis must be added to ammonia's lifetime GHG effects. The advantage of ammonia over hydrogen is that it can be stored as liquid in an easier temperature to maintain (–34°C) while, being a hydrogen carrier, its liquid form allows more hydrogen storage per cubic meter than liquid hydrogen itself (OECD, 2018). Ammonia used as fuel could offer GHG reductions to the extent that it is processed using renewable energy and is sourced from hydrogen made from electrolysis using renewable sources.

The Way Ahead

The picture of maritime transport in the future will be significantly shaped by the extent to which alternative low or zero carbon fuels that would provide the necessary "quantum leap" vis-à-vis BAU would be adopted. These fuels will have to become technically and (most important) economically viable if they stand any reasonable chance to be adopted in the future. This section addresses, among other things, how economic viability would be possible. As things stand, one thing is clear, to this author at least: Economic viability will not happen by itself and will depend upon a combination of factors, many of which are in the political arena. The exposition below borrows from Psaraftis and Zachariadis (2019) and Psaraftis and Kontovas (2020).

Nine years after the adoption of the Energy Efficiency Design Index (EEDI) in 2011, which is still (together with the Ship Energy Efficiency Management Plan- SEEMP) the only mandatory GHG emissions reduction measure, there is no doubt that the adoption of the Initial IMO Strategy in 2018 was a landmark decision. Achieving GHG emissions in 2050, which are at least 50% lower than they were in 2008, is a substantial and ambitious target that has to be taken very seriously by all involved.

At the same time, any hope that substantial GHG reductions can be achieved via improvements on EEDI is in our opinion grossly unsubstantiated. In a study conducted for Danish Shipping, Smith et al. (2016) showed, among other things, that the existence of EEDI versus a scenario in which there is no EEDI as we move to 2050 amounts to a GHG emissions difference of about 3%.

It remains to be seen whether the short-term measures currently on the table at the IMO will achieve a credible dent in GHG emissions. At the time this chapter was being finalized, no decisions on these or any other measures had been made. The COVID-19 outbreak in the spring of 2020 put a break on the speed of IMO regulatory decision making, and it is unclear whether the pandemic will have a real (positive or negative) impact on prospects for future GHG reduction. Our default assumption is that the IMO GHG reduction discussion will resume as soon as the COVID-19 situation is under control.

As stated earlier, there is still no sense of priority among the wide array of candidate measures, all of which are on the table. Market Based Measures (MBMs) have been put into the medium-term class (to be agreed upon between 2023 and 2030) *but only as a possibility*, even though the Damocles sword of a European Union (EU) Emissions Trading System (ETS) is looming, as will be explained further. There appears to be no sense of urgency for any MBM to be discussed at the IMO. Some industry circles seem to favor a bunker levy, but no one dares propose it explicitly at this point in time.

In December of 2019, the new President of the European Commission revealed that in the context of the “European Green Deal” (EU, 2019) shipping would be included in the EU ETS. To be clear, ETS is clearly an MBM. Thus far the position of the EU has been to align itself with the IMO process on decarbonization, and essentially refrain from acting on a possible inclusion of shipping into the EU ETS before seeing what the IMO intends to do on GHGs. The EU ETS is a major instrument in EU energy policy, covering electricity production and several other major industries (but not shipping). The European Commission has been closely monitoring the IMO process, starting from what is agreed on the initial strategy in 2018 and all the way to 2023. Until recently, the Commission had refused to take the ETS option off the table or even to specify what would trigger action on its part. However, as of December 2019 this has changed, and the European Green Deal clearly points to the ETS path for shipping.

One week after the European Green Deal was announced, as many as eight shipping industry organizations to IMO (ICS, BIMCO, WSC, Intertanko, Intercargo, Interferry, CLIA, and IPTA) publicly announced to IMO a to be jointly funded project for a 5 billion USD fund that would be used to finance R&D for technologies and fuels toward decarbonization. This would be raised by a 2 USD/ton mandatory surcharge on bunker fuel (ICS, 2019).

This proposal, which had actually been prepared by the shipping industry for some time now but essentially kept in the drawer until its recent public release, is not labeled as an MBM and in fact it is hardly an MBM. It is clear that two dollars per ton will not reduce GHG emissions, either in the short run or in the long run. *A fortiori*, any link between the R&D that could be conducted going forward and any actual decarbonization that might occur as a result of this R&D is sketchy at best. The proposers state that the R&D fund would only cover applied research and not commercial development of fuels and other technologies, which would be left to other stakeholders, such as shipbuilders, engine manufacturers, energy producers, ports, etc. At the same time, the proposers state that their proposal is not intended to frustrate or delay the development of an MBM should there be consensus for this among Member States, and that the architecture of their proposal could be used should the IMO proceed with a levy (implying that this would be their preferred MBM).

Whatever the case, the European Green Deal and its link to shipping decarbonization via the EU ETS has suddenly upgraded the role of the EU as an important player in the IMO process. After the news was announced, the IMO Secretary General rushed to defend the overall IMO approach and indicated his willingness to talk to the European Commission and Parliament. How these discussions will evolve remains to be seen. In the current IMO agenda on GHGs, the European Green Deal is still below the horizon, at least for the time being.

Substance-wise, our position is that the main plus of the European Green Deal, as regards shipping of course, is the increased pressure it can exert on the IMO to move faster, and that a levy on bunker fuel or on carbon emissions would be a far better MBM than ETS in the quest to decarbonize shipping. For a discussion of the relevant issues see Psaraftis (2012), Psaraftis and Lagouvardou (2019), and Lagouvardou et al. (2020).

Irrespective of the fact that MBMs are not up for discussion at the IMO in the foreseeable future, one idea that might be worth considering would be to impose a significant bunker levy at a global level. By significant we mean not 10 or 20 USD per ton, as is being occasionally contemplated by industry, but at least one order of magnitude higher. To put it very simply, if society truly does not like fossil fuels, or any other fuel that produces GHGs (and this includes LNG), and cannot, for obvious reasons, mandate their outright ban, society should at least try to implement the “polluter pays” principle by internalizing (even partially) the external costs of GHGs. The only way to do so is by putting a significant price on the fuels that produce these GHGs. Conversely, and so long as these fuels are affordable, there is no doubt that they will be used. All the debate on LNG, hydrogen, and other alternative fuels (see section, The Drive to Decarbonize Shipping above) critically hinges upon the economic dimension: we would like to know not only how much GHGs these alternative fuels would avert, but also what is the cost of producing and using them. Conversely, and barring a technological quantum leap, for as long as these alternative fuels are not viable economically, they will not be used. In turn, and as long as fossil fuels are cheap, they will be used.

We basically think that a substantial bunker levy would be the best (or maybe the only) way to induce technological changes in the long run and logistical measures (such as slow steaming) in the short run. In the long run, it would lead to changes in the global fleet toward vessels and technologies that are more energy efficient, more economically viable and less dependent on fossil fuels than those today. In that sense, it would have the potential to drastically alter the face of maritime transport. However, as things stand, and mainly for political reasons, the adoption of such a measure is considered as rather unlikely. If proposed, it will likely encounter the resistance of many IMO member states, with the US leading the pack, at least as things stand.

So to conclude, it would seem that the future of maritime transport, and specifically the direction it will take in the years ahead, critically depends on whether or not maritime policy makers can find a way to make the world shipping community adopt the bold measures that are necessary to incentivize the viable development and use of alternative low or zero carbon fuels.

References

- Allen, D.T., 2014. Methane emissions from natural gas production and use: reconciling bottom-up and top-down measurements. *Curr. Opin. Chem. Eng.* 5 (1), 78–83.
- Bouman, E.I., Lindstad, E., Rialland, A.I., Strømman, A.H., 2017. State-of-the-art technologies, measures, and potential for reducing GHG emissions from shipping—a review. *Transp. Res.* 52, 408–421.
- Buhaug, Ø., Corbett, J.J., Endresen, Ø., Eyring, V., Faber, J., Hanayama, S., Lee, D.S., Lee, D., Lindstad, H., Markowska, A.Z., Mjelde, A., Nelissen, D., Nilsen, J., Pålsson, C., Winebrake, J.J., Wu, W.Q., Yoshida, K., 2009. Second IMO GHG study 2009, IMO document MEPC59/INF.10.
- Collodi, G., 2010. Hydrogen Production via Steam Reforming with CO₂ Capture, Foster Wheeler.
- Corbett, J.J., Thomson, H., Winebrake, J.J., 2015. Methane Emissions from Natural Gas Bunkering Operations in the Marine Sector: A Total Fuel Cycle Approach, Report for US Maritime Administration.
- de Kat, J., Mouawad, J., 2019. Green ship technologies. In: Psaraftis, H.N. (Ed.), *Sustainable Shipping: A Cross Disciplinary View*. Springer.
- DNV GL, 2015. The Fuel Trilemma: Next generation of marine fuels. Strategic Research and Innovation position paper 03-2015.
- DNV GL 2018. Assessment of Selected Alternative fuels and Technologies, Guidance Paper 2018-05.
- EU, 2016. Methane emissions from LNG-powered ships higher than current marine fuel oils, European Commission DG Environment News Alert Service, http://ec.europa.eu/environment/integration/research/research_alert_en.htm.
- EU, 2019, Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions "the European Green Deal" (Com/2019/640 Final).
- Howarth, R.W., 2014. A bridge to nowhere: methane emissions and the greenhouse gas footprint of natural gas. *Energy Sci. Eng.* 2 (2), 47–60.
- ICS, 2019. Proposal to establish an International Maritime Research and Development Board (IMRB) Submitted to IMO/MEPC 75 by BIMCO, CLIA, ICS, INTERCARGO, INTERFERRY, INTERTANKO, IPTA, and WSC.
- IMO, 2018. RESOLUTION MEPC.304(72) (adopted on 13 April 2018) Initial IMO strategy on reduction of GHG emissions from ships, IMO doc. MEPC 72/17/Add.1, Annex 11.
- Lagouvardou, S., Psaraftis, H.N., Zis, T., 2020. A literature survey on market-based measures for the decarbonization of shipping. *Sustainability* 12 (10), 3953, doi:10.3390/su12103953.
- LR and UMAS, 2020, Techno-economic assessment of zero-carbon fuels, Report by Lloyds Register and UMAS, London, UK, March.
- Milne, T.A., Elam, C.C., Evans, R.J., 2006, Task 16, Hydrogen from Carbon-Containing Materials, Report for the International Energy Agency, Agreement on the Production and Utilization of Hydrogen, National Renewable Energy Laboratory, Golden, CO USA.
- Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestad, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., Zhang, H., 2013. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), *Anthropogenic and Natural Radiative Forcing. Climate Change 2013 The Physical Science Basis Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- OECD, 2018. Decarbonising Maritime Transport. Pathways to zero-carbon shipping by 2035, Report of the Organisation for Economic Cooperation and Development, International Transport Forum, Paris, France, March.
- Psaraftis, H.N., 2012. Market based measures for green house gas emissions from ships: a review, *WMU. J. Maritime Affairs* 11, 211–232.
- Psaraftis, H.N., Kontovas, C.A., 2020, Influence and Transparency at the IMO: the Name of the Game. *Maritime Economics and Logistics*, <https://doi.org/10.1057/s41278-020-00149-4>.
- Psaraftis, H.N., Lagouvardou, S., 2019. Market based measures for the reduction of green house gas emissions from ships: a possible way forward. *Samfundskøkonomi* 4/19, 60–70.
- Psaraftis, H.N., Zachariadis, P., 2019. In: Psaraftis, H.N. (Ed.), *The way ahead, chapter in Sustainable Shipping: A Cross-Disciplinary View*, Springer.
- SINTEF, 2017, GHG and NO_x emissions from gas fueled engines, SINTEF Ocean AS Report.
- Smith, T.W.P., Jalkanen, J.P., Anderson, B.A., Corbett, J.J., Faber, J., Hanayama, S., O'Keeffe, E., Parker, S., Johansson, L., Aldous, L., Raucci, C., Traut, M., Ettinger, S., Nelissen, D., Lee, D.S., Ng, S., Agrawal, A., Winebrake, J.J., Hoen, M., Chesworth, S., Pandey, A., 2014, Third IMO GHG Study 2014, International Maritime Organization (IMO) London, UK, June.
- Smith, T., Raucci, C., Haji Hosseinloo S., Rojon I., Calleya J., Suárez de la Fuente S., Wu P., Palmer, K., 2016. CO₂ emissions from international shipping. Possible reduction targets and their associated pathways. Report prepared by UMAS, London, UK, October.
- UNCTAD, 2019. Review of Maritime Transport 2019, United Nations Conference on Trade and Development, Geneva.

Maritime Transport and Territorial Scale

Brian Slack, Department of Geography, Planning and Environment, Concordia University, Montreal, QC, Canada

© 2021 Elsevier Ltd. All rights reserved.

The Port as a Land Use	540
The Changing Character of Port Sites	541
The Port as a Regional Hub	542
Future Challenges to Ports and Their Hinterlands	542
The Energy Transition	543
Automation and Digitalization	543
Cybersecurity and Ports	543
Ports and Climate Change	543
References	544
Further Reading	544

While maritime transport is a mode that obviously depends upon seas and oceans, it possesses distinctive and critically important territorial dimensions. Ships have to exchange their cargoes and passengers at sites on land, and it is with terrestrial-bound customers that their revenue-generating business depends. The territorial dimensions of maritime transport are therefore of fundamental importance.

Ports are the points of contact between ships and land. Ships require sites that they can access and where they can dock safely. Accessibility and safety are the most fundamental requirements of a port. Ships must also be able to load and/or discharge their cargoes efficiently. This requires port infrastructures involving quay walls, terminal space, and equipment that are capable of achieving ship turnarounds with the minimum of delays. From a shipping perspective ports are land-based facilities that are absolutely essential to maritime trade.

Ports also serve wider functions on land. They are essential components of supply chains that link producers and consumers. For many supply chains the maritime link may be the longest spatially or timewise, but the landward connections between the port and the original producer or final consumer are the connections that account for the highest costs and present challenges for on-time performance. Landward links from ports have become a critical concern for customers and represent one of the most difficult challenges in global supply chains.

This article examines these two territorial dimensions of maritime transport. The first considers the port as a land site that carries out specific functions for shipping. It is an urban land use with a very distinctive character, one that has undergone significant change. Because it occupies waterfront sites it frequently enters into conflict with other land uses. The second dimension is its regional importance. Ports serve markets, and these are becoming more extensive and competitive in many cases. Ports, and the actors who comprise the port community, seek to extend and strengthen the linkages, so that the port region, or hinterland, becomes a means of securing traffic and strengthening competitiveness.

The Port as a Land Use

For maritime transport a key territorial requirement has always been that of a site where ships can shelter and load and discharge their cargoes with safety and efficiency. Favored port sites were where natural features provided protection from storms and where there was sufficient depth of water to allow access. Over time, and dependent upon traffic growth, physical infrastructures such as wharves were built to facilitate loading and unloading from ship to shore and artificial barriers were constructed to provide protection. The port area became a distinct zone but was integrated within the general urbanized area. The trade passing through generated economic multipliers due to the manpower employed in the port itself, in the manufacturing industry that processed goods shipped through the port, and in the transport related businesses that managed its trade and commerce. It is not surprising that many port cities became major urban centers, such as London, Hamburg, Marseilles, Genoa, New York, New Orleans, Shanghai, and Tokyo.

Over time ships changed from wooden craft powered by sail to steel ships powered by fossil fuels. The repercussions of these technological developments were significant particularly with regard to speed, but their impacts on ports were not substantial because the capacities of ships remained constrained by the ability of ports to load and unload cargoes efficiently. Port operations were labor intensive. Freight was brought from storage sheds to dockside using animal and human labor; the goods lowered by sling and pulleys into ships' holds, wherein gangs of men had to secure the freight, an extremely dangerous procedure. All kinds of cargoes were involved, from raw materials to packages and barrels, of different shapes and sizes. Loading and unloading took time, and

vessels were laid up in port for weeks. Thus there were diseconomies of scale in shipping since increasing the size of the ship would result in longer port calls and still fewer potential revenue-generating trips per year.

The most important factors precipitating the changes were market conditions and the mechanization of port operations. The earliest market sectors to change were the bulk trades, because the global economic and population growth that occurred after the end of World War II generated increased demands for liquid bulk (oil) and dry bulk (coal, iron ore, and grain). Because these were products that could be handled more easily than others, using pumps and conveyer belts, they permitted scale economies to be realized, and in both sectors vessel size increased dramatically. In 1951 the largest tanker had a capacity of 20,000 DWT (deadweight tons), while the typical size of a dry bulk ship was 10,000 DWT. By the late 1970s maximum tanker size had increased to over 300,000 DWT. The growth of dry bulk carriers has been slower but the maximum size by the 2010s has reached 400,000 DWT.

The transformation of general cargo handling in the ports of the world remained largely unchanged until the mid 1960s after the introduction of the container. Its concept may have been simple, but the economic consequences of the container have been profound (Levinson, 2006). Standardizing the unit load allowed rapid transfers of cargo across quay walls by cranes that slot the containers one on top of other in the holds of ships. The storage of containers and their repositioning in the terminal could now be mechanized too. Because of the mechanization, vessel turnarounds could be achieved far more rapidly than before, so that instead of ships spending weeks in port their turnarounds today are measured in hours. As with the bulk trades the size of container ships increased as a result the increased efficiency of handling. The first cellular ships of the early 1970s had capacities of 1000 TEUs (20 foot equivalent units). Size increased to 5,000 TEUs by the late 1980s, 10,000 TEUs by the end of the 20th century, and 20 years later the largest container vessels exceed 21,000 TEUs.

As much as the increasing size of ships was made possible by growth in international trade and economies of scale, it must not be forgotten that ports have played a significant role. It has been their ability to exchange cargoes between land and ships more efficiently that made the growth in vessel size possible. In turn ports were transformed. Larger capacity vessels require deeper channels, longer berths, and more storage space as well as more expensive handling equipment. Existing docks sites were rarely suitable, and so new larger sites had to be developed and equipped. In most cases new terminals were built on greenfield sites some distance from the port city. The effects of the changes have been to change the character and functioning of ports themselves and to alter the dynamics between ports and the cities.

The Changing Character of Port Sites

The ability of ports to respond to the growth of traffic and the increase in the size of ships have required significant capital investments. While the shipping industry itself has had to invest heavily in new vessels, a great deal of the costs of has been borne by the ports themselves. Satisfying this demand represents significant capital costs for port authorities, either through relocation to extensive sites adjacent to deep water or for dredging. There are differences between bulk and general cargo concerning the capital costs of developing new terminals. This is because they possess different operational, ownership, and organizational characteristics. Bulk shipping involves freight that is handled in large quantities at terminals owned or leased by corporations whose shipments are components of industrial supply chains controlled by the corporations. They use ships that are owned or leased by the same corporations. Fluctuations in traffic are shaped by broad market conditions. In contrast, general cargo (largely containerized today) is transported by independent carriers that operate between ports on regular liner services, who charge rates according to demand, and serve a wide range of customers from small shippers to giant retailers. The terminals are built by port authorities who, in some cases operated the terminals, but more frequently lease them to terminal operators. There is a high degree of uncertainty in this business: the shipping lines are footloose and have freedom shift services to competing ports, and the terminal operating companies have little control over traffic volumes. Port authorities have to attract carriers in order to secure business and have to invest heavily in terminals with no guarantee of success.

The capital requirements for port development raise a number of questions that have led to different solutions. The first set of questions arises out of port governance (Ng and Pallis, 2010). Up until World War II, ports in most countries were owned by central governments and operated as a public service. The costs of providing the services and the ability of a public authority to manage a resource in the competitive field of containerization resulted demands for reform in port governance. The UK and New Zealand opted for total privatization; many others adopted a system of commercialization while retaining nominal public control. Others, under the control of municipalities such as those in Northern Europe, remained largely unchanged. Academic studies of the success of port reforms are numerous, but the results are inconclusive. The major issue is that as much as port reform has led to greater efficiencies, the capital requirements still require financial support from national governments, many of which are reluctant to undertake either because of the demands other more popular projects, or because of uncertainties of outcomes. In some countries, such as Canada, where port reform took place 20 years ago, the issue of privatization or other models of governance are presently under discussion.

A second issue concerning port expansion is that of public acceptability (Hoyle, 2000). Port activities now take place in relative isolation from their host cities. Their sites are extensive and located beyond the daily routine movements of most residents. In addition, they are gated and fenced off, because of international regulations on port security. Many residents see ports in a negative light because they are perceived to cause pollution by the refineries and smelters in the port zones, and they generate traffic that clogs urban highways. An important additional factor is that the direct benefits of jobs that tied the port to the city in the past have diminished. The number of dockworkers has fallen by over 90% in some ports, and although the remaining unionized workers are well-paid, there is evidence that other workers dependent on the ports, such as warehouse workers and truckers, have seen their

salaries decline in constant dollars over the years (Hall, 2009). It is no surprise therefore that the port is no longer seen by many citizens as a positive contributor to urban life. They see governments being called upon to invest in new infrastructures to accommodate the longer ships that require deeper channels, when there are other more pressing urban issues.

A result of these concerns is that port expansion and development especially in Europe and North America has become a time-extensive process. Not only have the commercial justifications for port projects are subject to greater scrutiny, but are also exposed to rigorous environmental and social acceptability assessments. Research on recent projects reveal that several have been abandoned because of negative assessments, such as Dibden Bay in the case of Southampton, while others have required 10 years of evaluation in the case of Deurganck Dock of the Port of Antwerp, while that of Maasvlakte II of the Port of Rotterdam took 20 years. If this trend continues, it may result in port congestion as ports lose their abilities to cope with trade growth. It also means that by the time projects are realized market and technology may have changed, thus threatening the relevance of the project. The situation is very different in China and many other developing economies, where projects are completed without comparable delays.

The Port as a Regional Hub

The container has not only allowed ports to achieve high levels of productivity, but has greatly facilitated its transport beyond the port gates. Intermodality has been one of the great beneficiaries of containerization. Other modes of transport can combine to extend services across ever-wider market areas, because the container, as a standard unit of freight, enables relative seamless interchanges between the modes. In this way container ports have opportunities to extend their market areas inland and compete with other ports for business. Ports that can generate more volumes enhance their attractiveness for the ocean carriers, which in turn offer customers a wider range of services that then becomes a competitive advantage over other ports. Ensuring market dominance over a productive hinterland has become a key commercial objective for ports, and this has led academics to explore the process and challenges of what is referred to as “port regionalization” (Notteboom and Rodrigue, 2017).

Port regionalization implies not only the attempt to manage and extend the market coverage of ports, but is also tied in with a problem of managing the growth of container traffic in the ports themselves. In many ports traffic growth is causing congestion. This becomes evident at the gates of entry and exit for trucks delivering and picking up containers, in the roads adjacent to the terminals and along corridors through cities. A wide range of solutions are in place to try to resolve congestion at the port gates. These include requiring trucks to make appointments, reducing charges at off-peak times, and extending hours of operation. The results reveal varying degrees of success. One problem with using time as a means of managing truck flows is that truckers are also dependent upon the hours of operation of the warehouses and logistics centers whose schedules may be more restrictive.

More general and extensive solutions to port and urban congestion revolve around promoting other modes of transport, especially rail, and where possible, inland shipping. Effecting a modal transfer is not always simple, since may give rise to operational and issues and involve commercial viability. Even where rivers are navigable the port has to find space for berths required for transferring containers to barges. In the case of rail transfers there are issues of exchanging containers between the port terminals and rail freight yards as well as the interest and ability of the railways to offer efficient services. The result is that solutions are varied and involve different sets of actors. In the case of European ports the lead has been taken by some ocean carriers who operate dedicated rail services from Rotterdam, Le Havre, and Hamburg to inland markets. In the case of barge transfers the ports have played a role, either by subsidizing barge services, as in the case of the Port of New York–New Jersey (which failed after the subsidy was terminated), or where port authorities form alliances with smaller inland ports to organize barge terminals and coordinate information exchanges as is the case of the port of Rotterdam.

A feature and concept closely related to port regionalization rationalization is that of dry ports (Witte et al., 2019). These are inland centers linked by dedicated rail service to ports that carry out many of the functions of ports, including customs clearance. Because they have more space they can relieve congested ports, while still ensuring that containers can arrive on time in ports ready for loading, and can store containers after being off-loaded from the ship. They also serve as a means of diverting traffic in hinterland regions from competing ports. Examples include the Dutch inland port of Venlo on the border of the Ruhr, the inland port of Zaragoza established by the port of Barcelona to divert potential traffic from Valencia as well as capture transborder traffic from southern France, and Front Royal, referred to as Virginia Inland Port, established by Virginia Ports Authority to divert Mid West traffic from Baltimore to Norfolk. Dry ports tend to be established by port authorities or private terminal operators.

Academic research indicates that port regionalization rationalization combines three levels of activity. First is that of the physical transport infrastructure involving all modes between the seaport and interior markets. Second are the service networks that refer to the direction, capacity, and frequency of transport connections and the actors that operate and manage the transport systems. Third are the flows of freight along the logistics chains that connect the port with customers across the hinterland, which requires understanding of the actors, their roles and activities that manage the flow of goods.

Future Challenges to Ports and Their Hinterlands

As they enter the third decade of the 21st century ports are confronting challenges that have the potential to modify and even destroy certain features that are taken as granted today. Four basic issues are considered. They are not unique to ports but they are likely to require adjustments or provide opportunities that are distinct. Detailed exploration of these topics is impossible, rather a brief

indication of what they are and how they might impact ports is undertaken. The order in which they are presented in no way reflects their importance.

The Energy Transition

There are growing concerns world-wide about burning of fossil fuels and contributions to the increasing levels of CO₂ in the atmosphere. There are policies in place internationally that are calling for a reduction in their use. For ports, this issue has two dimensions. First is the fact that liquid bulk, especially crude oil and refined products, represents a very important traffic. Many ports regard this as a cash cow, since the revenues generated are realized without significant costs to the port authority. Coal is another product that is threatened, especially since coal-fired thermal stations are being closed down. A decline in traffic would affect the financial health of ports, and in the long run present challenges and opportunities for the possible re-use of liquid bulk terminals. Alternate energy products might be considered, liquified natural gas and liquified petroleum gas are possible substitutes, but there are major concerns about their social acceptability close to urban centers.

The second dimension of fossil fuels is their use within the port area. Already many ports offer “cold ironing,” where ships in port can turn off their diesel power units by being attached to port electric outlets. There is also a growing trend for the replacement of diesel-powered equipment within the port, such as cranes and vehicles, by electric motors. For the port authority and terminal operators this represents a considerable financial commitment.

Automation and Digitalization

The modern port achieved its present importance by facilitating world trade through mechanization. Today it is on the cusp of further transformation as a result of automation. Indeed, evidence is already accruing that the digital revolution has begun. One of the trends in container operations is the automation of terminals. Mechanization of ports reduced port labor requirements very significantly, but the tasks of many aspects of port terminal are already digitized, with tracing and tracking of containers from the moment they enter the terminal, as well as their ultimate stowage position on board the vessel. Further automation is already evident in some ports. This includes the use of autonomous machines that physical remove arriving containers from trucks and rail wagons that move them and lift them to storage lines, and, when required, load them on board ship by gantry cranes. Automated terminals now exist in some ports, such as Rotterdam, Singapore, Shanghai, and Los Angeles. Evidence suggests that full automation may not be appropriate at present because of costs for smaller ports or those where traffic is variable with peaks and valleys. On the other hand, the present high capital costs may be expected to be lowered over time. The issue with dock labor remains a challenge.

The container port is a nexus of information. Documentation is essential, but it has many sources (shipper, consignee, insurance company, transport providers, customs and excise, homeland security or equivalent, terminal operators, port authorities). What information is required differs. The result has been the gradual replacement of paper documents and faxes with electronic data interchange. The process may have been speeded up but it is still cumbersome. Estimates suggest the cost of information handling represents 15% of all transport costs. Blockchain technology is being proposed as technology well suited to large networks of disparate partners. As a distributed ledger technology, blockchain establishes a shared, immutable record of all the transactions that take place within a network. It then enables permitted parties access to trusted data in real time. In late 2018 IBM and A.P. Moeller-Maersk formed a partnership to exploit blockchain with several ports.

Cybersecurity and Ports

Ports, their customers, and related partners all use information systems and communication technologies for equipment operation, cargo movement and tracking, business operations, and security. These not only control the operation of terminals but are also used to monitor security, handle business functions such as invoice preparation and billing. There are different systems in place among the actors. This makes the port community prone to malicious activities and software errors because there are so many points of entry into the systems. In 2013 Maher terminals, a major terminal operator in New York, experienced difficulties in switching between old and new computer systems and the result brought the terminal to a standstill and had impacts on other terminals in the port as well as deliveries throughout the north-east United States. In 2017 Maersk shipping line was attacked by hackers that interrupted operations at 73 ports around the world. It cost the company \$300 million to recover from the attack. Cyber attacks have also been reported in the ports of Barcelona, Long Beach, and San Diego.

There are growing calls for better training, regular vulnerability inspections, and in particular, contingency plans. The International Maritime Organization is preparing regulations on cyber security.

Ports and Climate Change

Climate change is claimed as one of the greatest challenges to humanity. Occupying sites at the interface of land and water, ports are vulnerable to environmental impacts produced by changing climate (Ng et al., 2015). Higher sea level, increased storminess, longer periods of high temperatures all have the potential to influence ports and cargo handling. The difficulty, however, is that the changes are not uniform across the world. In some locations it is land that is rising that is resulting in local changes of sea level, in others it is increasing droughts that is the major climate change impact. Consequently, climate change must be addressed at the local level,

with the risks assessed and mitigation measures applied case by case. Rather, maritime transport as the greenest of transport modes has the opportunity to improve its position in the global transport industry if the existing black marks of using heavy oil as fuel is changed.

References

- Hall, P.V., 2009. Container ports, local benefits and transportation worker earnings. *GeoJournal* 74, 67–83.
- Hoyle, B., 2000. Global and local change on the port-city waterfront. *Geograp. Rev.* 90 (3), 395–417.
- Levinson, M., 2006. *The Box: How the Shipping Container Made the World Smaller and the World Economy Bigger*. Princeton University Press, Princeton.
- Ng, A.K., Becker, A., Cahoon, S., Chen, S.L., Earl, P., Yang, Z. (Eds.), 2015. *Climate Change and Adaptation Planning for Ports*. Routledge, New York.
- Ng, A.K., Pallis, A.A., 2010. Port governance reforms in diversified institutional frameworks: generic solutions, implementation asymmetries. *Environ. Plan. A* 42 (9), 2147–2167.
- Notteboom, T., Rodrigue J.-P. Re-assessing port-hinterland relationships in the context of global commodity chains. in: Wang, J., Olivier, D., Notteboom, T., Slack, B. (Eds.), *Ports, Cities and Global Supply Chains*, Routledge, New York.
- Witte, P., Weigmans, B., Ng, A., 2019. A critical review on the evolution and development of inland port research. *J. Trans. Geogr.* 74, 53–61.

Further Reading

- Hayuth, Y., 2007. Globalisation and the port-urban interface: conflicts and opportunities. In: *International Workshop on Ports, Cities and Global Supply Chains*, 2005, Hong Kong, China.
- Rodrigue J.P., Comtois C., Slack B., 2016. *The Geography of Transport Systems*, Routledge.

Containerization and the Port Industry

Hercules Haralambides, Paris School of Economics, Pantheon-Sorbonne University, Paris, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction: What is a Port?	545
The City-Port	546
Port Planning	547
Port Terminal Operations	547
Inland Terminals (Dry Ports)	548
Ports Before Containerization	548
Port Labor	549
Port Competition and the Need for Port Reforms	549
Development of Infrastructure	550
Port Reforms	550
Financing and Pricing of Port Infrastructure	550
The Impact of Containerization on Port Planning, Development, and Operations	551
Containerization	551
Containerization's Impact on Ports	551
Containerization and Labor Productivity	552
Containerization's Impact on Maximum Containership Sizes	552
Mega-Ships and Diseconomies of Scale in Port	552
Hub and Spoke Systems	554
Summary and Conclusions	555
Biography	556
References	556
Further Reading	556

Introduction: What is a Port?

A port is generally understood as the interface between sea and land, where goods change *mode of transport*, for example, from ship to truck or rail and vice versa (Fig. 1). To operate safely, a port needs to be protected, or *harbored*, from the elements of nature (waves, winds, currents) by such things as breakwaters and windbreakers, but also by essential port services like towage, pilotage and line-handling which assist the ship to berth and start in safety its cargo-handling operations. The latter operations require a wide spectrum of port equipment such as ship-to-shore cranes, straddle carriers, reach-stackers, forklifts, and more.

Fig. 1 shows a part of the historic Port of Brindisi, South Italy, at the turn of the previous century (c. 1920). The train would arrive right at the waterfront—something very uncommon at that time—in the true sense of *intermodality* (see below), 100 years before the term was invented (The train, coming from London and Paris, would bring to Brindisi “la valigia delle Indie,” that is, the diplomatic bag (actually a trunk) and other mail and documents heading for India. Crossing Suez was great fun and an adventure, but the ladies had to be careful with the scorching sun. Thus, cabins had different prices with the most expensive ones being the *posh* (i.e., Port-Out-Starboard-Home). At the very same spot, Phileas Fogg and Passepartout (Jules Verne: *Around the World in 80 Days*) were waiting to board P&O's m/s Mongolia. Fogg was worried of a delay but Passepartout was reassuring him that, opposite to sailing ships, steamships were never late!).

Although the general public may understand the importance of ports (and shipping) for trade, growth and prosperity, most people have never visited or seen a port at close distance. Rather, the impression people have of a port is usually one of an exotic spot, of romanticism, adventure or organized crime. Cinematography has helped a lot in creating this impression: A beautiful account of the “waterfront” can be enjoyed in Elia Kazan's 1956 masterpiece “on the waterfront,” with Marlon Brando, or Mike Newell's 1997 drama “Donnie Brasco,” with Al Pacino and Johnnie Depp.

To put things in perspective, a port can be anything from a sheltered stretch of sea, protecting a handful of fishing boats somewhere in the South Pacific; a block of cement in a small Greek island, on which a passenger ferry would lower its ramp to disembark passengers; a buoy onto which a tanker would moor to offload its oil through a pipeline; a finger-pier alongside which a bulk carrier would unload its coal on a conveyor belt; a *cool port* (i.e., a refrigerated facility) in Latin America exporting fruit to Europe; a mega-yacht marina in Monaco or Nice, or just a water-taxi that would disembark passengers from a cruise ship

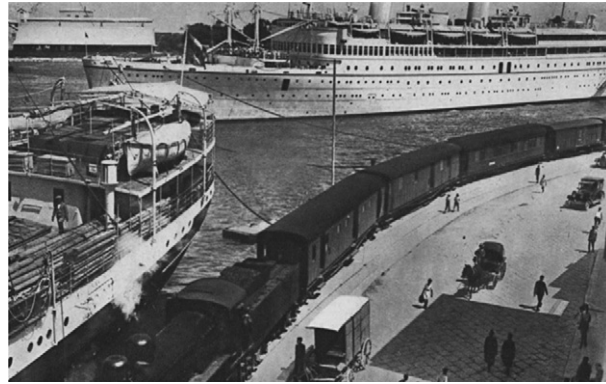


Figure 1 Port: the interface between land and sea.

anchored in the middle of the sea, outside Amalfi, in the absence of any berthing infrastructure at the picturesque village of South Italy. At the other end, there is the mindboggling Yangshan Deep Water Port (of Shanghai), handling 40 million containers a year, or the equally impressive industrial complex of the Port of Rotterdam, running for 40 km along the river Meuse to the North Sea, comprising in its domain a *cluster* of thousands of companies, from the large refineries of the oil majors, to the small paint shop, inconspicuously hidden under an abandoned bridge, next to a pub.

The chapter is organized as follows: The following section discusses the important city–port relationship and the way ports have developed and relocated over the years, in parallel with increases in trade and ship sizes. The complexities in port management are also discussed here, as well as the new phenomenon of inland terminals (dry ports), aiming to relieve seaports of congestion pressures and improve on seaport efficiency. The section thereafter presents the port landscape before the arrival of containerization. Port labor is discussed here, and the need for economic reforms as a result of the new, intensified, competition among ports. The section concludes with a note on port finance and the increasing participation of the private sector. The chapter continues in its next section with the *container revolution* and the impacts and radical changes this development brought about on the emergence of *hub-and-spoke* systems in global distribution; containership sizes and port labor productivity. Diseconomies of scale in ports, due to the emergence of mega-ships, are also discussed here. The final section concludes.

The City-Port

In the past, and whenever possible, cities were built along the banks of rivers and so were their ports. The “city-port” was the epicenter of commercial activity and trade. Historical examples are those of Antwerp, Hamburg, London, Manchester, Paris, Ghent, Bruges, Amsterdam, Rotterdam, Bordeaux, Nantes, Rouen, and many others. The economics of the city-port concept were simple: The port is an essential element of trade; trade requires transport; and transport costs money; sometimes a lot, particularly in the earlier days and in the absence of “transport and logistics” as we understand them today. Thus, minimizing transport costs was essential and what better way to achieve this than building ports close to cities.

As trade was increasing—and together with it the size of ships—, ports needed bigger and deeper *turning basins* and *approach channels*, so that ships could approach and maneuver safely. In addition, increasing trade necessitated more storage areas (warehouses) in the port, and these either did not exist, or their (opportunity) cost was very high given the alternative uses of port land in a city (housing, commercial, etc.). Thus, ports started to relocate downstream to river estuaries. Whenever this was problematic, artificial canals were dredged, like the Nieuwe Waterweg (New Waterway) of Rotterdam.

Downstream, space was ample and water depths sufficient to accommodate the larger and larger ships without the need for too much dredging which is expensive, often technically difficult, and above all environmentally impactful: the siltation of rivers brings with it soil which is polluted from agricultural fertilizers (of cultivations irrigated from the river), or cooling-water sewage from industrial activity. Disposing of dredged materials is subject to very strict laws in most countries and this makes dredging an even more difficult decision. Notable downstream examples are Bremerhaven (Bremen), Le Havre (Rouen), Maasvlakte (Rotterdam), Zeebrugge (Bruges), Le Verdon (Bordeaux), Tilbury (London), Liverpool (Manchester), and Nantes St. Nazaire (Nantes) among many others.

Thus, old city-ports were abandoned and many of them still remain so. Others have been beautifully rehabilitated and have become the most exclusive residential; commercial; business; and entertainment areas of the city. Notable examples include the Docklands of London, the Koop van Zuid in Rotterdam, Hamburg’s HafenCity, the Porto Antico of Genoa, or the Vieux Port of Marseilles (Fig. 2).



Figure 2 Old port areas rehabilitated (clockwise from upper left: Rotterdam, Hamburg, Amsterdam, Genoa).

Port Planning

In most cases, moving downstream has allowed ports to overcome the problem, albeit only temporarily. Up to the economic crisis of 2009, trade was increasing at an average annual rate of 8%, while the infrastructure of ports and terminals, once built, would remain fixed for many years, if not decades. Thus *port planning* has always been one of the greatest challenges of a port administration.

If one asks an ocean carrier how big a port should be, they would immediately reply “as big as possible.” Apparently, a carrier would not want to wait even for a minute if they could help it, and it has been established (see below) that, at a port capacity utilization of around 75%, congestion starts to set in.

If instead one would ask the same question to a port manager, particularly one responsible for his “bottom line,” the answer would be “as small as possible.” Obviously, he would love having ships queuing up outside his port, like the good old times, if he could help it.

As usual, both need to put some water in their wine and compromise: too little port capacity is as bad as too much: The first would cause congestion and loss of custom to competitors; the second would be wasting scarce taxpayer money. In other words, the ship cannot wait, nor, however, can the port continue spending taxpayer money so that the ship does not do so. This is particularly true in increasingly commercial activities, such as those of transshipment container terminals, whose impacts are not localized at the economies that have borne the brunt of the investment, but are dissipated throughout the supply chain, from the country of origin to the country of final destination of the transported goods.

Port Terminal Operations

Port land in downstream ports is available but not unlimited. Moreover, land at seafronts has also a lot of alternative uses—recreation; fishing; residential—and thus high opportunity cost. And although with the arrival of the container demand for conventional warehousing space may have declined, the number of containers arriving at the port—both for export and import—has been increasing steeply. In most ports with scarce port land the solution has been to stack them high, that is, stack containers on top of each other. This, combined with a need for highest efficiency in terminal operations—the result of intensified terminal competition amongst neighboring ports—has led to a number of interesting optimization or, rather, minimization, problems in *operations research* or *management science*. Some of them are briefly discussed in what follows.

As usual, the thing to “minimize,” that is, the problem’s *objective function*, is costs. In our case, such costs relate to “unproductive moves” of containers and of cargo-handling and transport equipment in the port. These movements, obviously, cost time, fuel for cargo-handling equipment (e.g., straddle carriers) and other energy costs; labor; and environmental impacts. Listed below are some of the optimization problems that need to be addressed, piecemeal or jointly, by the management of a port terminal:

- Containers that arrive at the port, either to be exported or imported, cannot be stowed anywhere at will. To minimize distances travelled by handling equipment (straddle carriers and trucks, for example), export containers are usually placed in designated areas closer to the ship and import containers closer to the gate.

- An export container, to depart shortly, should be placed as high up as possible in the stack, preferably with no other container on top of it. In the opposite, *reshuffles* would be necessary, if the top containers depart later. Moreover, reshuffles increase dramatically with the height of the container stack.
- Special containers, such as reefers and containers carrying dangerous goods, should be stowed in designated areas; this is also the case when stowing containers in ships.
- In large busy terminals, hundreds of trucks arrive, often simultaneously, to pick up containers. This may create severe congestion at the gate, on the motorway, or at parking areas outside the port.
- With regard to export containers, these must be loaded in a certain way, according to a *stowage plan*. Containers heading for the same destination port should be stowed close together.
- Filling up the bigger ship in Asia is easier said than done. To do so, the ship has to call at more Asian ports than what her size would warrant, often picking up containers at random and at short notice, without due consideration to the importance of proper stowage planning. As a result, ship- and terminal stowage planning at the other end (Europe/North America) often becomes a nightmare and this is one of the reasons carriers are investing in their own *dedicated terminals* (Haralambides et al., 2002).
- To ensure the right vessel stability, weight distribution should be observed and heavy containers should be stowed below deck as much as possible.
- Dangerous containers should be segregated from the rest.
- Shipping alliance members have their own bays on the ship which adds inflexibility and higher complexity to the stowage problem.
- Same as at the terminal, stowing containers onboard ships should minimize *rehandles*.

Inland Terminals (Dry Ports)

In spite of all this, and of terminal automation advances which often approach the limits of science fiction, the problem remains: container throughput increases much faster than what ports can handle, at least at the level of efficiency expected by their clients, whose ships become bigger and bigger. Concentration of cargo at main hubs does not make things easier either. A reversal of trends can thus be seen as of recent. From the earlier era when ports were obliged to move downstream to find space, ports now look back at the hinterland to find the additional space they require. Inland terminals (or dry ports) are therefore mushrooming, connected to the seaport by rail, road or inland waterways (Haralambides and Gujar, 2011, 2012). As such, inland terminals are usually developed close to railway and motorway junctions, to facilitate the transfer of containers between modes of transport, favoring, to the extent possible, the more environmentally friendly modes like rail and inland waterways transport.

Depending on population density and level of economic activity in the seaport's hinterland, inland terminals have often been described as *satellites* (or extensions) of the seaport: These terminals are located at close distance to the seaport and are often managed by it, at least in their initial stages of development. Such arrangements are typical in densely populated Europe. Oppositely, where distance to the seaport is the main issue (e.g., North America or China), *long-distance intermodal rail hubs* are the preferred solution.

An inland terminal is not much different than a seaport apart from the lack of a seafront! Stacking yards, cargo-handling equipment and gates are all similar to those of the seaport. As the purpose of the inland terminal is to *feed* the seaport with containers at the right time, thus by-passing the need for storage and extra handling at the seaport, the presence of customs authorities at inland terminals is of importance. The same is true for import containers which should speedily leave the seaport for the inland terminal without spending time in the seaport's main stack. In this way, the inland terminal serves another important function, that of an *inland distribution center*.

In all cases the role of the *Port Authority* (PA)—the autonomous body which manages the seaport—is crucial. As it is further discussed below, PAs are assuming an increasingly enterprising role as “stakeholder managers,” extending their “gates” to the hinterland, as further out as possible, so as to also widen their catchment areas. To do this, the “*intelligent port authority 4.0*” needs to smooth out supply chain bottlenecks among ships, terminals, customs authorities, and hinterland transport, storage, and distribution. The seaport of today understands, in its supply chain manager role, that it is now supply chains and not ports that compete for custom.

Such a role requires modern management practices and the transformation of the PA from a government dependency to an entity operating under increasingly commercial terms. Often, “stakeholder management” is an extremely frustrating exercise in view of conflicting interests, particularly when such “stakeholders” sit in the governing or supervising boards of port authorities.

Ports Before Containerization

In the earlier days (up to the beginning of the 1960s) *general cargo*, that is, manufactured and semimanufactured goods, carried by liner shipping, was transported, in various forms of packaging (pallets, boxes, barrels, crates, slings), by relatively small vessels, known as general cargo ships. These were *twin-deckers* and *multideckers*, that is, ships with *holds* (cargo compartments) in a shelf-like arrangement where goods were stowed in small prepackaged consignments (parcels) according to destination (Fig. 3).

This was a very labor-intensive process and, often, ships were known to spend most of their operational time in port, waiting to load or discharge. Extensive warehousing facilities were in this way also necessary to protect goods-in-waiting from the elements of nature. This required lots of space which often did not exist in bustling ports, making port work even more labor-intensive.



Figure 3 Multipurpose, general cargo vessels.

Unpredictability of ship arrivals made *congestion* a chronic problem in most ports, raising the cost of transport and hindering the growth of trade. Equally importantly, erratic and unpredictable cargo movements (and consequently port work), obliged manufacturers, wholesalers and retailers to keep large stocks. As a consequence, warehousing and carrying costs were adding up to the cost of transport, making final goods more expensive and, again, hindering the growth of international trade.

Port Labor

Seafaring was great fun in those days [sic]—seafarers were exploring foreign ports and cities in a multitude of different ways—, but the same cannot be said for the *casual* labor ports were employing, which was rather ill-considered and looked down by society. And this is why: As a result of the unpredictability of port work, port management could not possibly employ permanent staff, paying them while idle, and waiting for the next ship to arrive, if it ever did. Yet, once in port, cargo had to be unloaded quickly, so that produce would not spoil. Labor was thus *casual*, that is, employed for as long; as much; and whenever required. This is how it worked: Each morning, a number of *dockers* would present themselves to a union foreman, often a mobster, and he, on the basis of certain “criteria” that had more to do with *natural selection* rather than anything else, would thumb-in the youngest, the strongest, the favorites of the union, or those prepared to return a kickback to the union (to indicate this “preparedness,” a docker would put a toothpick behind his right ear). Those dockers not selected would return to their “locales” [sic] and indulge in whatever it was they were indulging in. As a result of this type of (casual) employment, ports were ill-famed and often centers of organized crime.

Port Competition and the Need for Port Reforms

Ports had no incentive to change; on the contrary: in the former times of import substitution policies (protectionism), inefficient ports were seen by many as effective nontariff barriers to trade. It has sometimes been argued that import-competing domestic producers had strong vested interests in the continuing existence of inefficient ports, as this offered them effective protection. Such producers could also be effective lobbyists and influential members of pressure groups that resisted port reform.

Moreover, ports were run by a public body—a municipality or a government department—while their *hinterland* (i.e., their catchment area) was *captive*: In other words, due to tariffs and other trade restrictions, national borders and lacking infrastructure, ports enjoyed considerable monopoly power; that is, cargo destined for, say, Italy would in all likelihood land at an Italian port, as close as possible to the final consignee. Thirty years ago, and as a result of “captive” port demand, forecasting port capacity was a fairly straightforward exercise, and a popular one at that, among students: population, GDP and trade data, together with a simple regression model, would easily do the trick with some excellent results. **Fig. 4**—still used by modern day port consultants—tells us that a GDP multiplier of 2.2 means that a 1% increase in national GDP would on average lead to a 2.2% increase in the country’s containerized imports. The same graph shows the dramatic decline in trade since the 2008–9 economic crisis. Localization vs globalization of manufacturing; protectionist policies; populist politicians; the US–China trade standoff; contempt for cheap consumerism and similar trends have all contributed to this.

Two things have changed this picture since: the development of extensive land transport infrastructure on the one hand, and containerization and the “through transport” concept on the other.

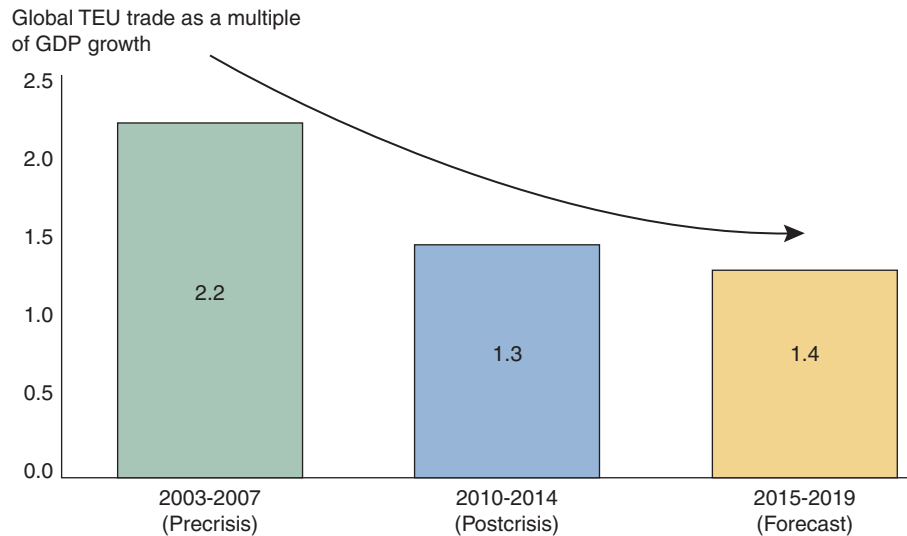


Figure 4 The GDP multiplier. *Source: IMF; BCG.*

Development of Infrastructure

Europe, for instance, dilapidated by a ruinous WWII, was being rebuilt, and what better way of doing this than developing your roads and railroads before anything else. If one were to take a cursory look at Europe's map today, what one would see would not differ much from a nice dish of Italian spaghetti. Roads and railroads, together with European economic integration, free trade, and the abolition of national borders, opened up the market for port services. Before people knew it, any Asian cargo could in principle reach any European destination, passing through any gateway port around Europe. Port demand was no longer captive but *stochastic*, and ports started to compete fiercely for custom, particularly for *transshipment*—i.e., other peoples'—cargo.

Port Reforms

Around the world, the port industry has invested heavily in order to cope with the intensifying competition. Modern container terminals—and corresponding cargo-handling equipment—have been built and new, more efficient, organizational forms (including privatization) have been adopted in an effort to speed up port operations. Private interest manifested in *dedicated container terminals* (Haralambides et al., 2002), operated/owned by the carriers themselves, and in the emergence of private terminal operators with extensive terminal investment portfolios around the world (Some notable ones at the time of writing are China Shipping Ports; Hutchison Port Holdings (Hong Kong); PSA International (Singapore); APM Terminals (Denmark)). Moreover, operational practices have been streamlined; the element of uncertainty in cargo flows largely removed; forward planning has been facilitated; port labor regularized; and customs procedures simplified. These developments took place under the firm understanding of governments and local authorities that ports, now, constitute the most important link (node) in the overall door-to-door supply chain and thus inefficiencies (bottlenecks) in the port sector can easily whither all benefits derived from economies of scale in liner shipping and in global supply chain management.

To modernize and face competition, ports have often invested far ahead of demand. A combination of factors can explain this. Excess port capacity, over and above national demand requirements, is often created in the belief of the infamous economic law of Jean-Baptiste Say according to which “supply creates its own demand.” In other words, once the port is there, the customer is bound to arrive. Unfortunately, this is not always so, particularly in cases where those who run ports have no competence in modern port marketing. Next, adequate berthing capacity is necessary in order to turn ships around as quickly as possible. Otherwise, a “footloose” container, with no loyalty to any port, might move easily to an adjacent competitor. Queuing theory easily shows that once a port reaches 70% capacity utilization, congestion starts to set in and this is something no modern port can afford. In this light, excess port capacity should be seen as an “operational necessity,” or cost, rather than as inefficiency or a waste of scarce resources. Carriers, having the upper hand as a result of their higher negotiating power (conferences, alliances, consortia) are able to play one port against another, forcing them to invest excessively in infrastructure. Finally, managerial egos, easy and cheap public finance, and port managers' heartburning desire to leave their footprint in history would all play a role, together with *economies of scale* in port construction: apparently, it is much cheaper to build 2 km of quay-wall rather than one.

Financing and Pricing of Port Infrastructure

Investment in container terminals, however, is not without risks, particularly if the investment aims at capturing container transshipment traffic which, as just said, is “footloose.” The situation is not helped much by the increasing competition among

neighboring ports, aimed at “stealing” transshipment traffic from each other. To do this, ports usually discriminate against (captive) domestic cargo in their pricing policies, while underpricing (footloose) transshipment cargo. (Under-pricing here should be taken to mean the charging of port dues and concession fees below the opportunity cost of port land.) This is particularly worrying in publicly funded port investments. Clearly, there is a need for public policy intervention, at least in the Europe Union, which needs to harmonize public spending among its member states, as well as to ensure free and fair competition among its ports. Solutions to this problem include transparency in port accounts; a common glossary of port finance terms; and pricing aiming at cost recovery (Haralambides, 2002).

Moreover, the automated nature of container terminals creates comparatively limited employment and this has been an additional reason governments are often seen reluctant to fund transshipment facilities whose benefits are not localized but diffused from the land of (Asian) origin to the country of (European) destination of the goods. In the form of a witticism, it is often said in the Netherlands that “the Port of Rotterdam is Germany’s biggest port.” Also, arguments are not rare in Germany regarding the usefulness of spending taxpayer money to maintain motorway A1: a north-south, motorway which facilitates Asian exports to reach European countries, rather than having the regional and developmental impacts infrastructure is supposed to have.

In comparison, the port of Antwerp, as a result of its multipurpose, more labor-intensive, character, creates four times more value added per square meter of port land than Rotterdam, the latter port being a highly automated and labor-saving. Thus, the Belgian taxpayer is much happier to pay for port development than his Dutch counterpart, who has often questioned the social utility of developing more port capacity at Rotterdam.

Exceptions to the above do of course exist, and Rotterdam is a good one: the port’s value added is not created simply by cargo-handling, storage and concession leases, but by its *port cluster*, encompassing 50% of Europe’s Asian and North American European Distribution Centers (EDC), in a city 50% of whose inhabitants are holders of a foreign passport, just because of the port. Unfortunately, not all ports can realistically aspire to such an enviable situation, developed not today but over a period of 70 years of hard work.

The Impact of Containerization on Port Planning, Development, and Operations

Containerization

It is many times said that *containerization* has revolutionized the transport and port industries. Such an emphatic characterization could be quite acceptable if one considers the enormous impact this originally purely technical solution in cargo-handling methods has had on the design and sizes of general cargo ships; the lay-out, equipment, development, operations and employment in ports; on inland transport requirements; land use; human skills; and shippers’ perceptions regarding the functioning of the overall supply chain.

The innovation entailed in the concept of containerization is credited to Malcolm Mclean: an American trucker who thought of separating the tractor from the trailer part of his trucks, standardizing (unitizing) the latter (trailer) so as to be able to be transported—with its contents intact—by various transport means, handled at ports by standardized cargo-handling equipment (quay cranes), and stacked uniformly one on top of the other, both in ships and at terminals. The container, or the maritime container, as it is often called, is a fairly robust and sturdy structure, manufactured at very high standards, intended to withstand the harshest of conditions, such as those often prevailing on the high seas. Containers are waterproof, vandal-proof, and adequately ventilated, to avoid possible accumulation of condensation, and, if treated well, they could give their owner many years of problem-free service (the average economic life of a container is 15 years). There are many types of containers (high cube, flat rack, open side, open top, tank, reefer (refrigerated), etc.), depending on the intended use.

Containerization’s Impact on Ports

General cargo goods are now increasingly carried in containers of standardized dimensions; most common is the $8 \times 8 \times 20$ feet container known as TEU—Twenty (feet) Equivalent Unit—, although containers of double this size (FEU—40 feet) are increasing in importance. Perhaps one of the most important effects of containerization is that, now, containers can be packed (stuffed) and unpacked (stripped) away from the busy waterfront, either at the premises of the exporter (consignor) and/or the importer (consignee), or at Inland Container Depots (ICD), warehouses, and distribution centers (dry ports).

By-passing the waterfront in the stuffing and stripping of containers, and thus having them ready in port to be handled by automated equipment, has increased immensely the punctuality, predictability and reliability of cargo movements and transport systems, enabling manufacturers and traders to reduce high inventory costs through the adoption of flexible Just-in-Time and Make-to-Order production technologies. Inter alia, such technologies have helped manufacturers to cope with the vagaries and unpredictability of the business cycle and plan business development in a more cost effective way. Indisputably, containerization has been the kindle wood under global logistics and supply chain management. And more: The concept of logistics does not stop at cargo systems, but it permeates every aspect of our everyday lives. For instance, in a reliable transport system, I know precisely what time I need to leave home to make it to the airport. But if taxis are frequently on strike; rail under continuous maintenance; or security controls at airport a mess, I need to leave home an hour earlier. And this hour is “my” inventory cost.



Figure 5 Cargo-handling before and after containerization.

Containerization and Labor Productivity

Containerization has had a number of significant advantages over the conventional, labor-intensive methods of handling general cargo. Apart from the remarkable improvements in port safety and the limitation of pilferage, damages and cargo claims, the system's major breakthroughs—particularly in the United States where it was first introduced—were in cutting down on expensive and often abusively organized labor, as well as reducing ship turnaround times.

Fig. 5 shows cargo-handling operations before and after containerization. Labor productivity in the earlier days (left side of Fig. 5) was roughly 1 ton per man-hour; with containerization (right part of Fig. 5), this went up hundredfold. In the first case, a docker would climb up the gangway 10 times an hour, with a sack of rice on his shoulder. In the second, a crane-driver would load 20 containers of rice onboard the containership, comfortably seated and handling a “joystick” from the cubicle of a ship-to-shore gantry-crane, or from the terminal office, or even from his home! The first docker would be paid peanuts (if he was lucky) while the second has a salary every worker in the world would envy today.

Containerization's Impact on Maximum Containership Sizes

The dramatic improvements in cargo-handling operations, brought about by the introduction of containerization, have enabled general cargo ships to spend hours or days now in ports, rather than weeks or months that was customary before. The reduction of port time, and the corresponding increase in time at sea, have eventually led to the substitution of the previous multipurpose general cargo ships by specialized high-speed container vessels of substantially (and ever increasing) larger dimensions (Fig. 6) which can take advantage of the *economies of scale* afforded by the shorter port turnaround times. Port efficiency and productivity have thus been the main drivers behind the increase in containership sizes, particularly since 1988 when American President Lines introduced the postpanamax vessels (Fig. 6). Up to that year, the maximum size of containerships (and of course of other types of vessels) was determined by the width of the Miraflores locks of the Panama Canal and it had remained constant, at around 4500 TEU, for almost a decade.

Mega-Ships and Diseconomies of Scale in Port

Undoubtedly, *Economies of Scale* in shipping have led to cargo consolidation, as well as to storage and distribution requirements, and thus to the emergence of global hubs and global logistics as we know them. There are, however, limits to the growth in ship sizes and these depend on freight demand; port capacity and technology; land infrastructure; other logistical costs; the future of global shipping alliances; and the attractiveness and future of the hub-and-spoke system in container transportation.

Gigantism in container shipping (Haralambides, 2019) has been the combined result of competition and concentration in this industry, as well as of shipbuilding technology which has made possible the construction of 400-m long ships, able to carry in excess of 22,000 containers (at the time of writing). The objective of the mega-ship trend has been to enjoy *economies of scale* leading to reductions in unit costs. Up to a point, this has been an economically legitimate objective, until diseconomies of scale start to set in, particularly in ports, that is, the place where start the “supply chain diseconomies of scale” of larger ships, including their impact on shippers' inventory costs.

To enjoy economies of scale and low costs, the ship needs to sail full and spend as little time in port as possible. In other words, carriers often require from the port that the ship be turned around at a given fixed time, for example, within 48 h, irrespective of its size. For most ports, this is a challenge. Whenever this is the case, it can be shown (Haralambides, 2017) that it costs more to handle a container arriving on a large ship than one arriving on a smaller one. In other words, port time per TEU is an increasing function of ship size, and this is a distinct “port diseconomy of scale” (Fig. 7).

		Nominal TEU tdw	LOA m	Breadth m	Depth m	Draft m
MSC GÜLSÜN 6 units in series from July 2019		22,960 teu 228,149 tdw	399.9	61.5	33.2	16.5
					Operated by MSC Built by Samsung H.I. MSC also has 5 units to be built at DSME	
OOCL HONG KONG 6 units in series from May 2017		21,413 teu 191,317 tdw	399.9	58.8	32.5	16.0
					Operated by OOCL(COSCO) Built by Samsung H.I.	
COSCO SHIPPING UNIVERSE 6 units in series from Jun 2018		21,237 teu 198,485 tdw	399.9	58.6	33.5	16.0
					Operated by COSCO Built by CSSC COSCO also has in addition 11 units of 19,200-20,100 teu built in CSSC/CSC/COSCO shipyards	
MADRID MAERSK 11 units in series from Apr 2017		20,568 teu 210,019 tdw	399.9	58.6	33.2	16.5
					Operated by Maersk Built by Daewoo (DSME)	
EVER GOLDEN 11 units in series from Mar 2018		20,388 teu 199,692 tdw	400.0	58.8	32.9	16.0
					Operated by Evergreen Built by Imabari	
MOL TRIUMPH 6 units in series from Mar 2017		20,170 teu 192,672 tdw	400.0	58.8	32.8	16.0
					Operated by MOL Built by Samsung H.I.	
BARZAN 6 units in series from Apr 2015		19,870 teu 199,744 tdw	400.0	58.6	30.6	16.0
					Operated by UASC Built by Hyundai Samho/Hyundai H.I.	
MSC OSCAR 12 units in series from Jan 2015		19,224/19,437 teu 197,362 tdw	395.4	59.0	30.3	16.0
					Operated by MSC Built by Daewoo (DSME) MSC also has in addition 6 units of 19,462 teu built in Samsung and 2 units of 19,362 teu at Hyundai H.I.	
CSCL GLOBE 5 units in series from Nov 2014		18,982 teu 184,320 tdw	399.7	58.6	30.5	16.0
					Operated by COSCO Built by Hyundai H.I.	
Maersk 'EEE' 20 units in series from Jun 2013		18,340 teu 194,153 tdw	399.2	59.0	30.3	16.0
	ALPHALINER				Operated by Maersk Built by Daewoo (DSME)	

Figure 6 Development of maximum size of containerships. Source: Alphaliner.

It is not so difficult to understand why: As crane productivity cannot be stretched much beyond 30 moves/h (it actually declines after a certain *crane density*), the only way to serve a larger ship in the same time (48 h) is by adding more and bigger (in terms of *air draft* and *outreach*) cranes. However, increasing crane density—that is, the number of cranes per 300 m of quay length—reduces crane productivity, among others nullifying the advantages of having bigger hatches.

Fig. 7 deserves a few more words. The downward-sloping blue line describes the economies of scale in shipping. One would observe that, after a certain ship size, the line becomes flatter, that is, economies of scale are actually lost. The red lines expresses diseconomies in port as a function of ship size. At smaller sizes, one would observe, the line is fairly flat, meaning that the time required to handle a ship increases linearly to its size. After a certain size, however, the line starts sloping upwards, meaning, as said already above, that it takes more time to handle a container arriving on a bigger ship. The u-shaped black line above is the sum of the

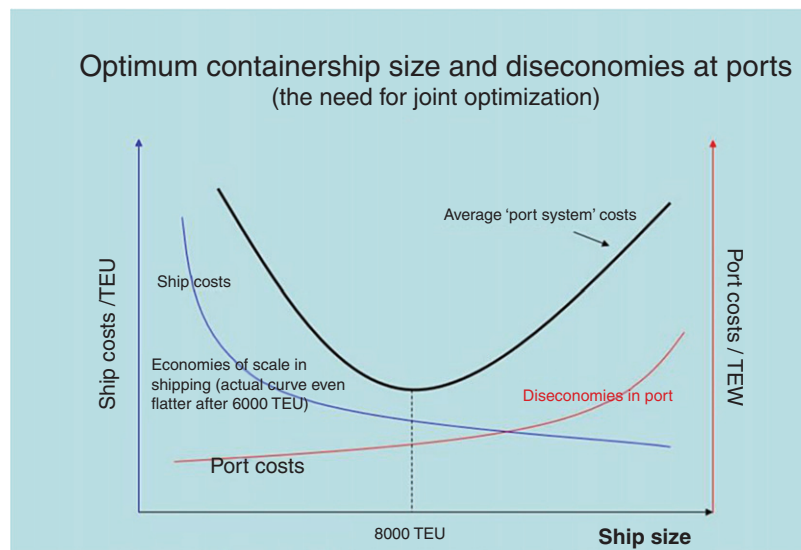


Figure 7 Optimum containership size and diseconomies in port.

other two lines, representing the “usual” compromise between two opposite developments: If one were to take into account, together with the benefits of bigger ships, the costs these impose on ports, the optimum ship size should be much smaller. In practice, port costs that should be taken into account go beyond cargo-handling time and into the significant infrastructure investment costs ports are obliged to undertake in their “anxiety” to attract the increasingly bigger ships from their competitors. At some point in time, the taxpayer will have something to say on this, particularly in view of the limited impacts of highly automated terminal operations.

Big ships impose substantial demands on port capacity, without however paying commensurately for this demand. For instance, when it comes to berth utilization, whereas before we could accommodate three Panamax vessels (i.e., three berths) along 1 km of quay-wall, today we can host there only two mega-vessels of the latest generation (about 400 m long each). Berth utilization obviously goes down and so does the utilization of Ship-to-Shore (StS) cranes, for bigger ships mean lower call frequency. A port manager responsible for his bottom line might have a problem in explaining this to his shareholders or supervisory board, unless carriers were bringing more traffic to the port with their larger vessels. But this does not happen either. As it can be shown (Haralambides, 2019), *call size*—the amount of containers a ship drops at a certain port out of the total number of containers it carries in its itinerary—is only moderately correlated with vessel size.

Finally, one needs fewer bigger ships, and fewer port calls, to serve a given amount of yearly demand. Thus, *connectivity* goes down and, with it, the contribution of shipping to trade and development. Here too, berth and crane utilization decline and this impacts on the capital costs of the port and of the terminal operator. In addition, a reduction in the frequency of carrier itineraries (i.e., number of services), caused also by slow-steaming, impacts the inventory costs of traders, thus defying the very principles of supply-chain optimization; and this is a clear *diseconomy* along the supply chain.

Hub and Spoke Systems

The capital intensity of mega-ships obliges them to limit their ports of call at each end of the voyage to just a few *hub ports* or load centers such as Shanghai, Singapore, Hong Kong, Rotterdam, or Hamburg. Here, thousands of containers are consolidated each day for export, and import containers are distributed with smaller vessels (feeder), rail, road, or river barge, to regional and local ports and final destinations. Complex *hub-and-spoke networks* have thus evolved whose logistical fine-tuning and optimization bears directly on consumer pockets.

Such a hub-and-spoke network is depicted in the infographic of Fig. 8. Ports 1 and 3 are both gateway and transshipment hubs in continents A and B, respectively, while Port 2 is a pure transshipment hub along the way. Ports 4 and 5 are other ports of call in the proximity of the two hubs, while ports 6–12 are regional facilities.

A pure transshipment hub such as Port 2 is a facility with comparatively limited domestic traffic. Examples are Damietta, Malta, Singapore, Gioia Tauro, Colombo, and Kingston. Without some captive (domestic) cargo base, transshipment hubs are risky investments due to the “footloose” nature of the container. Two roles need to be present in a successful hub like Hamburg or Rotterdam: A “gateway” role (i.e., serving domestic cargo) and a transshipment role, in fairly balanced proportions. Also, regardless of the typology of a hub, i.e., foreland consolidation hub like Singapore or hinterland distribution hub like Rotterdam, two elements

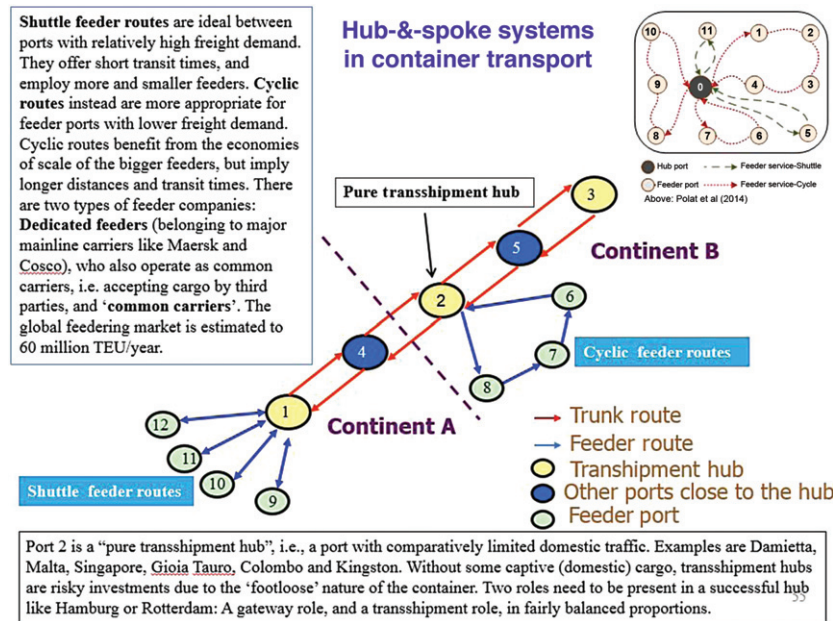


Figure 8 Hub and spoke systems and feeder networks.

(I have decided to leave this figure as an infographic for the benefit of hundreds of colleagues around the world who use it in their course materials).

should be present for successful operations: *centrality* (to main consumer markets) and *connectivity* (in terms of frequency and number of shipping companies calling at the port).

Cargo distributed from a hub port is also *feedered* on smaller ships to regional ports. This is done in two ways: shuttle feeder routes or cyclic routes. Shuttle feeder routes are ideal between ports with relatively high freight demand. They offer short transit times, and employ more and smaller feeders. Cyclic routes instead are more appropriate for feeder ports with lower freight demand. Cyclic routes benefit from the economies of scale of the bigger feeders, but imply longer distances and transit times. There are two types of feeder companies: Dedicated feeders (belonging to major mainline carriers like Maersk and Cosco), who also operate as common carriers, that is accepting cargo by third parties, and "common carriers." The global feeder market is estimated to 60 million TEU/year (Polat et al., 2014).

Summary and Conclusions

A bicycle manufactured in Wuhan (China) can reach Vienna (Austria) in 147 different ways; which one is the best? During the past 20 years, ports have come to realize that it is actually supply chains themselves that compete while they, ports, are important, not to say crucial, *nodes* along those supply chains. Ports have thus also realized that inefficiencies of whatever nature in their operations can jeopardize the efficiency of the overall supply chain.

At the same time, the development of land infrastructure around the world has expanded the catchment areas (hinterlands) of ports and has thus intensified competition amongst them.

These trends, amidst the container revolution, have forced ports to modernize from the slow, expensive, inefficient and bureaucratic public facilities of the past, to modern enterprising entities, increasingly operating under principles of commercial management.

Port infrastructure has thus become not only the pivotal facilitator of international trade, but also a lucrative asset for private investment, as evidenced by the many *global terminal operators*, developing impressive port terminal portfolios around the world.

Infrastructure investments around the world will amount to trillions of dollars in the course of the present century. The financing of infrastructure is increasingly undertaken with heightened attention, in view the mounting public debt of countries, and of the property of infrastructure to require long gestation periods, while the servicing of its debt cannot wait. In this light, ports and port terminals could be seen as "success stories" in the financing and development of infrastructure, setting a most useful example for others to follow.

Biography

Since 1992, Hercules Haralambides is Professor in Maritime Economics and Logistics, having taught at eight universities (and in six different countries), most prominent of which being Erasmus University Rotterdam and National University of Singapore. Currently he is Distinguished Chair Professor at Dalian Maritime University (China) and Adjunct Professor at Texas A&M University (USA). Hercules is the founder of the Erasmus Center for Maritime Economics and Logistics (MEL) (<https://www.maritimeeconomics.com>) and also the founding Editor-in-Chief of the quarterly journal Maritime Economics & Logistics (MEL), published by Palgrave-Macmillan (<https://www.palgrave.com/41278>). He has written and published over 300 scientific papers, books, reports, and articles in the wider area of ports, maritime transport and logistics and has consulted governments, international organizations and private companies all over the world including, for a series of years, the European Commission. In the period 2011–5 Professor Haralambides was President of the Italian port of Brindisi and at the end of that period (2015) he established “Haralambides & Associates”: a global maritime think-tank engaged in executive education and strategic policy analysis. In 2008, Professor Haralambides was decorated with the Golden Cross of the Order of the Phoenix, by the President of the Greek Republic.

References

- Haralambides, H.E., Cariou, P., Benacchio, M., 2002. Costs, benefits and pricing of dedicated container terminals. *Int. J. Marit. Econ.* IV (1), 21–34.
- Haralambides, H.E., 2002. Competition, excess capacity and the pricing of port infrastructure. *Int. J. Marit. Econ.* 4, 323–347.
- Haralambides, H.E., Gujar, G., 2011. The Indian dry ports sector: pricing policies and opportunities for public-private partnerships. *Res. Transport. Econ.* 33, 51–58, doi:10.1016/j.retrec.2011.08.006.
- Haralambides, H.E., Gujar, G., 2012. On balancing supply chain efficiency and environmental impacts: an eco-DEA model applied to the dry port sector of India. *Marit. Econ. Logist.* 14 (1), 122–137.
- Haralambides, H.E., 2017. Globalization, public sector reform, and the role of ports in international supply chains. *Marit. Econ. Logist.* 19 (1), 1–51.
- Haralambides, H.E., 2019. Gigantism in container shipping, ports and global logistics: a time-lapse into the future. *Marit. Econ. Logist.* 21 (1), 1–60.
- Polat, O., Günther, H.-O., Kulaka, O., 2014. The feeder network design problem: application to container services in the Black Sea region. *Marit. Econ. Logist.* 16, 343–369.

Further Reading

- Alderton, P.M., 2008. *Port Management and Operations*. Taylor & Francis.
- Haezendonck, E., Verbeke, A. (Eds.), 2018. *Sustainable port clusters and economic development: Building competitiveness through clustering of spatially dispersed supply chains*. Palgrave Studies in Maritime Economics, Palgrave Macmillan.
- Haralambides, H.E. (Ed.), 2015. *Port Management*. Palgrave Readers in Economics, Palgrave Macmillan.
- Talley, W.K., 2009. *Port Economics*. Routledge.

Port Management

Lourdes Trujillo*, **Daniel Castillo Hidalgo[†]**, **Manuel Herrera[‡]**, * Department of Applied Economics, University of Las Palmas de Gran Canaria (Campus Tafira), Faculty of Economics, Las Palmas de Gran Canaria, Spain; [†]Department of History, University of Las Palmas de Gran Canaria (Campus Tafira), Faculty of Economics, Las Palmas de Gran Canaria, Spain; [‡]Louvain School of Management (LSM), Université Catholique de Louvain, Louvain-la-Neuve, Belgium

© 2021 Elsevier Ltd. All rights reserved.

Introduction	557
A Historical Overview of Port Management Evolution before the 1980 Decade	557
Main Port Management Models	558
Port Structure	558
Port Management Models	559
Public Service Ports	559
Tool Ports	559
Landlord Port	560
Private Service Ports	560
Ports Insertion into the Global Networks: Governance Implications	560
Summary and Conclusion	561
References	561
Further Reading	562

Introduction

This paper is organized as follows: The first section presents a historical perspective of the main changes developed in port management models as a result of the transformations of the shipping industry since the late 19th century. Then, we present the main port management models observing their main characteristics as well as their administrative and institutional structure. The last section includes a general approach to the governance structure in ports and its decisive influence on the insertion of ports in global economic networks.

A Historical Overview of Port Management Evolution before the 1980 Decade

Before the arrival and development of steamship during the second half of the 19th century, port management structure was relatively limited to a number of civil servants engaged on the collection of taxes, health inspection, and other military issues. The low-intensity schemes of maritime traffics and the scarce development of port infrastructures during the preindustrial age did not demand complex administrative structures for the most of seaports in spite of the existence of specific major trade centers where administrative organization was relatively more complex. Nevertheless, the rapid expansion of steamship throughout the late 19th century fostered major changes at both, institutional and entrepreneurial levels (Harlaftis et al., 2012). Advances on maritime shipping developed during the industrial revolution promoted the sustained international trade growth where seaports had played a key role as nodes for the spread of economic globalization.

Moreover, maritime networks were reinforced in a global scale and ports rapidly received the massive attention of policy makers as well as national merchant fleets expanded and international trade flourished. Thus, seaports were conceived as essential tools for the wealth of the nations and reforms should be made in order to enlarge commercial traffics and connect them to economic progressive regions. However, reforms would include institutional and management issues. The increased complexity of port communities and the variety of new activities born in the heat of globalization demanded administrative reforms. Hence, major port reforms (both at the institutional and physical levels) were led—in overall terms—by the state as a consolidation of public liberal reforms developed throughout the 19th century.

Since the 1880, in most of industrialized countries, the government assumed the task to modernize ports and waterways as a mechanism to foster the economic performance of their respective countries. Public reforms on port management were then led by the central government to introduce policies of relative coordination between ports into the same country. This coordination was building on similar taxes and more broadly on common regulatory frameworks in terms of international trade in order to avoid interport competitiveness. Reinforcing port activity required a trained staff heavily composed by engineers and marine officers. The administrative staffs also included economists and lawyers whom role was to create an appropriate environment for the investment of private port companies and the attraction of maritime traffics.

On the other hand, there were port managements models built on historical paths such as them dependent on public local bodies where interport competitiveness was encouraged even into the same country. In addition to that, private actors operated private port

such as ore or bulk-trade terminals as the result of public concessions made by the public institutions throughout the 19th century. Through the establishment of agreements with the public institutions, private operators exploited the port terminal and contributed with tax-collection, annual fees, and other revenues to the public expenditures.

In overall terms, port management structures since the late 19th century up to the 1980 decade were characterized by slow developments and the progressive adoption of managerial structures dominated by the weight of the state and public bodies in their management and decision making.

Nevertheless, considering port management evolution in the long run, it could be argued that the nature and specificities of the different models were deeply affected by historical, economic, and political endowments. Impressive transformations on informatics, the renaissance of liberal policies, and the further effects of the second wave of globalization since the 1980 decade fostered structural changes on the way how port management was conceived (Kaukianen, 2014). Since that decade, port authorities emerged as local-regional key actors. These institutions received increased competences from the national governments and state agencies in terms of planning and economic strategy. Those authorities tended to collaborate with the private sector in a process known as “devolution” that represented the increased economic and political importance of regions.

Main Port Management Models

The trade globalization implies a heavy integration between supply chains and productive structures being all of them heavily dependent on the efficiency of the transport system. The ports, as part of both the supply chains and the transport system, play an important role in the integration of the different transport modes into the supply chain structures. Consequently, port management turns out to be a crucial factor in the trade globalization. It involves the performance of several agents jointly operating at different levels (local, regional, transregional, and global) and additionally, the participation and interaction in a broad sense of public and private partners: port authorities, port users and associations, international institutions, global investors, and so on. The contemporary port management theoretical frame involves a number of formal and informal institutions and the common recognition of rules and managerial structures ultimately oriented to foster traffic growth, in a free trade framework promoting socioeconomic welfare and environmental protection. Interaction between private and public actors requires a comprehensive legal structure compatible with global trade networks. The way the public and private actors are intertwined within the ports, defines on the other hand the different management models. All of them, involves the use of strong analytical tools to help their implementation at the different levels of the ports structure extending their scope from the financial aspects (investments, profits, assets, cash, solvency) to operations (ship handling, cargo handling, service provision, environment protection). These elements also include all the different aspects of a modern port day to day management, from human resources to market orientation.

Broadly speaking, port management includes all the aspects of the day-to-day modern port performance but one of them has come to be especially relevant: port governance. Involving most of the components of port management, port governance models are focused in the institutional aspects of the port performance defining the institutional framework that relates and intertwine all the stakeholders involved in the port life. Given the crucial role of the port authority (PA) in the modern port life, the governance of the ports can be approached in two different ways: port governance and PA governance. Port governance regulates the interactions between the different port services and companies, how they are managed and how they interact with the port (World Bank, 2007). PA governance on the other hand is referred to the way that the PA is internally managed and how it interacts with the public administration, the port companies, and the rest of port stakeholders (Caldeirinha et al., 2018). Because of the extended implementation of the devolution concept, the port authorities play more and more the role of a business company. Acting with a very high degree of autonomy, they are managed as any other corporation involving not only day-to-day administration but strategical issues as regional development, supply chain insertion, port–city interactions, and maritime and land connectivity. As in any corporation, is governance at the end what defines both the internal relationships of the PAs and the external interconnection with their stakeholders.

Port Structure

Since there are many aspects involved, it is useful to divide seaport activities between: (1) infrastructure, (2) services provided by the port, which require the use of the former, and (3) coordination between the different activities performed at ports. The main characteristics of these three elements are analyzed below (Fig. 1).

Therefore, there are many different activities being performed simultaneously within the limited space of port areas, with ships constantly entering, being serviced and exiting. Therefore, there is a need for an agent to act as a coordinator to ensure the proper use of common facilities, and to take care of safety and the general design of port facilities. In most seaports, this function is played by an organization called the *port authority*. These are generally public institutions, where local interests are represented, but this configuration is not unique, and it is possible to find examples of purely private port authorities (Trujillo and Nombela, 2000). There are several organizational modes for seaports, depending on the role that port authorities assume. These are usually labeled as *landlord port*, *tool port* and *services port* (public or private).

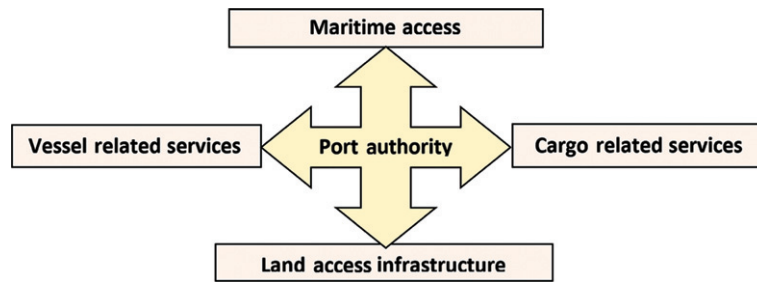


Figure 1 Port structure.

Port Management Models

Port management models evolved hand in hand with the 1970s port reform processes that gained momentum in the late 1980s with the public-sector reform in the United Kingdom followed in the 1990s all over the world. The British experience had a great impact on the consequent reform processes since it meant a radical change in the way the ports were managed. Part of the British port system was fully privatized including land property, complete absence of regulatory bodies and the deregulation of port labor market. In the subsequent port reform processes boarded from the 1990s until now all over the world, this approach has been considered radical in excess, not being extensively followed in its wholeness (Brooks and Pallis, 2011). In 2005, the World Bank in a highly cited work, set up the basis of the modern port governance models in its World Bank Port Reform Tool Kit (WBPRTK) in an attempt to systematize port governance models. Widely used all over the world especially as a guide for port reforms in the developing countries, this tool kit highlights the role played for the PA as a crucial actor in the management and governance of the ports, classifying the ports according to this role in public service ports, tool ports, landlord, and private service ports. Defining in this way the role of the PA, this classification gives a clear picture of the balance of functions and responsibilities between the private and the public sectors (Trujillo and Nombela, 2000).

There is an ample variety of port governance models what makes very difficult a detailed analysis of all of them. It is possible however to find common characteristics—variables—in order to identify the different tendencies followed around the world in the port reform processes conducting to the main port governance models. Thus, in Europe we can find three main tendencies defining the corresponding models: the Hanseatic model, mainly based in governance performed by local administrations, the Latin model with a more centralistic approach and the Anglo-Saxon model based in a completely independent management most of the time in private hands (Suykens and Van De Voorde, 1998).

Many others types and subtypes of governance models are applied all over the world, depending on the political environment, the size of the ports, their integration in the supply chains, the public administration traditions, and in general, any aspect of the port management influenced by the governments, local and PA, and the rest of stakeholders involved in the port life. Given the PA preeminence, the activities, performed by these institutions, constitute an important variable of port governance. It must be remarked that the already mentioned World Bank classification of the ports—according to the PA role—as public service ports, landlord ports, tool ports, or private service ports is not rigid, presenting an institutional plasticity and normative adaptation to the local characteristics of each region, country, or even ports. It is also interesting to note that any of these models can be considered as potentially more efficient than the others.

Public Service Ports

Public service port management models are mainly defined by the predominance of the public sphere on the governance structures. In these ports, the port authority is the main actor, performing all the services and owning the whole infrastructure. Being normally part of a higher-level public institution as a ministry or government department, the port authority is responsible for the construction, management, and operation of the whole port's infrastructure. Thus, is a public institution who supervises, guides, coordinates, and control port activities. Consequently, with all these, port funding is almost entirely dependent on public finances. However, even under full-public port management models, the private sector also could benefit from concessions and specific agreements for the provision of services. More and more, these types of ports enhance the participation of private actors providing these services through cooperative entrepreneurship strategies, approaching these ports to the tool port model.

Tool Ports

This kind of model is an intermediate step between the public service port and the landlord model. As in the service ports, the port authority is still preeminent, owning the infrastructure and performing most of the services. Cargo operations, however, are performed by private companies operating in terminals and docks owned by the port who owns most of the time the handling equipment too. As a transitional model, the tool ports are of especial interest for developing countries pursuing the improvement of the port system efficiency but unable to transfer the total control of the port operations to private hands either for political reasons, market deficiencies, or the lack of administrative resources enough as to enhance and control an extensive devolution process. It often happens in addition that the size of the ports in these countries is not large enough to justify a private company investing in a cargo terminal or any similar infrastructure, so it is the public administration that must face the investment.

Landlord Port

Landlord ports are the clearest exponent of the devolution process initiated in the 1980s to transfer services in ports from the public sector to private corporations. As an evolution of the tool port toward the fully privatized port, it is an intermediate solution that protecting the public interest. It enhances the efficiency and productivity of the ports, being nowadays the predominant model in the developed countries. However, we could remark some exceptions as the United Kingdom that embraced the fully private model in some of its ports after the reform of the 1980s.

Forced by the challenges of the globalization, public powers looked for a model that evolving from public ownership and control to public–private partnership. These governance models were able to avoid the current inefficiencies derived from heavy bureaucracy, lack of market orientation, political interference, lack of financial capacity to board the huge investments associated to the industry, and even the distortive effect of the trade unions power, traditionally very strong in the ports all over the world.

The model involved strong devolution processes transferring most of the services and responsibilities from public institutions to private entities. Under this collaborative schedule, the public institutions remain as the owner and provider of the major infrastructures, transferring to private investors the economic exploitation of ports services, what results most of the time, in the concession and transference of public dominion to private agents over port infrastructures. Through this transference, the public administration holds the property of the land leasing it to a private corporation—normally via a concession agreement—for a determined period. The lessee acquires the commitment to make the necessary investments in restructure and equipment for the normal operation of the facilities. The amount to be invested is normally is specified in the concession agreement as well as the rent to be paid to the port authority—normally an amount per square meter leased—and any other charges and taxes of application. There is an ample variety of concession agreements but most of the time, the lessee compromises a minimum volume of traffic to be operated on the facility on an annual base and during the whole period of the leasing contract. Discounts and exemptions are normally applied on a progressive scale by the port authority if this minimum volume is exceeded. For the case that the compromised investment during the whole leasing period is exceeded, is a common practice too to contemplate in the concession agreement, discounts, exemptions, or even an extension the leasing period according to the exceeded investment.

In what respect to the port services (anchoring, port pilotage, berth occupation, etc.), it is normally the port authority who determines the maximum tariffs for them, as well as the different taxes and the necessary basic regulations such as environmental or labor and safety protection policies. Most of the time, the PAs are at the same time the economic regulator that warrants the fulfillment of the economic conditions of the concession agreements and services licenses. This scheme present regulatory problems since the proximity of the PAs to the operators can lead to the capture of the regulator.

Moreover, the role the port authority also includes the political and economic promotion of the port as well as negotiations with public–private national and international partners fostering and promoting port competitiveness in order to attract seaborne traffics and investments from the private sector. The public sphere is also responsible for infrastructure planning and investing, funding expansion and improvement works of the basic infrastructure.

Private Service Ports

This management model is the extreme case of the application of the devolution concept. In fact, implies the full privatization of the port's infrastructure conditioned to the maintenance of their original function: to attend and handle the maritime traffic. It implies the full-private ownership of port infrastructures as well as full-private management of the services. Consequently, are private investors after obtaining of the corresponding permissions, who entirely build, exploit, and manage the port infrastructures. This model is based on a full autonomy in terms of funding and planning strategy. Private stakeholders entirely face the risks and benefits of their investments. Hence, under this administrative scheme, competitiveness and efficiency are privileged. Thus, the port is conceived as an integrate private company providing the complete range of services. The role of the state is quite limited, maintaining in any case the regulatory functions under the national legal framework in issues like environmental regulation, labor conditions, maritime safety, etc. In some cases, public entities can be integrated as shareholders of the port corporation in to keep a certain degree of influence in the port policies, orientating them to promote and safeguard the public interest

Ports Insertion into the Global Networks: Governance Implications

The integration of productive processes in the global supply chains has forced the ports—as nodes of interchange—to integrate themselves into these chains. This integration is carried out through the port operators that suffer a double direction integrative process: upstream, into the supply chains they serve and downstream, into the ports in which they operate. This process has also a double character: operational upward and territorial downward.

Operationally, the port operators are integrated into the systems and procedures of the supply chain, including every step from the production process to the final distribution of the product. In short, production, transportation, administrative processes, and information flow and all this, in a continuous process incorporating at every moment. Thus, the technological improvements that are taking place in each one of the links of the chain affected the process as a whole.

This operational integration is what explains the position taken by many of the global shipping groups in the port terminals business through processes of vertical integration that allow the insertion of handling operations in the maritime transport link of the supply chains and consequently, in the whole operation of them.

Territorially on the other hand, the port operators are integrated into the port spaces, immersing themselves into the administrative, legal, and operational systems of each port. This territorial integration, would explain to a large extent the processes of horizontal integration that have led to the creation of port operators participated by shipping alliances that, through the application of economies of scale, can extend their penetration into ports on a global scale, conducting in many cases to situations of monopoly in some ports.

From the point of view of both the public interest and port governance, the previous scenario poses a double problem. On one hand, the strong concentration in the sector of terminal operators gives them a very high negotiating capacity against the PAs, that in many cases are also the economic regulators of this industry. This capacity is reflected not only in terms of negotiating administrative concessions contracts, but demanding and imposing special working conditions, operational priority or preferment tariff rates, what can lead in many ports to situations of predominance even with the capture in some cases of the regulator. The special characteristics of the container traffic increases this negotiation capacity in that, the port operator—most of the time part of a shipping group—not only provides both, transport and handling capacity, but the possibility of increasing the volume of the port operations bringing to the port transshipment traffic. On the other hand, the position of the APs is strongly conditioned by the fact that if they invest public money in new infrastructures, such infrastructures must be used exclusively for the purpose that is designed, what requires a specialized operator willing to invest in the auxiliary infrastructure and running the risk of the business during the concession period.

The result of all this is a balance situation in which the PA has already made the investment and needs to use it and the operator on the other hand, controls the traffic and has the know-how but needs the physical infrastructure. This balance situation has its most clear manifestation in the process of granting concessions, where the bargaining power of each of the parties is revealed. It should be borne in mind that the strong concentration of port operators has given rise to large groups of global terminal operators (GTOs), what increases substantially their bargaining power. On the other hand, and especially in the landlord-type ports where the devolution concept is widely extended, the PAs have no other option that negotiation with the operators in order to provide the port with efficient infrastructures.

The governance implications of the above are relevant but suffice to say that the PAs have the normative and legal capacity to favor what it would come to constitute the best policy of safeguarding the public interest: the stimulation of the intraport competition. The AP is who has the capacity to favor the reduction of entry barriers that impede the access of new actors to the port activity, acting on the three barriers that traditionally are considered relevant in the port activity: economic, regulatory, and positional barriers. The economic entry barriers are derived from the fact that established operators can compete with lower costs, mainly for three reasons: better location in the port area, lower operating costs due to economies of scale, and benefits derived from cumulative public investment (De Langen and Pallis, 2007). It is the case in many ports that an already established operator -especially if it has reached an important volume of operations—opts to a terminal concession tender from a position of clear advantage. This predominant position would be also fostered both by its location and by its operating costs—product of the magnitude of its turnover-, benefiting also from accumulated public investment (as land connections to the terminal that it is already operating).

Regulatory and positional barriers to entry are less important from the point of view of intraport competition, but they are interested from the point of view of interport competition. Thus, regulatory barriers may in general be the same for all operators in a particular port, but between competing ports those barriers can be different. The same would apply to positional barriers that are derived from the fact that land in ports is always limited, which would affect all operators in a port, but could favor the establishment of new players in nearby competing ports with available land.

It can be concluded that the devolution processes lived in the ports all over the world in the last decades give a great prominence to the GTOs that have gained a significant bargaining power along the years. In order to cope with this situation, keeping the advantages of the landlord scheme in terms of efficiency, investment capacity, and operational flexibility, protecting at the same time the public interest, the PAs have the normative and legal capacity to promote intraport competition what seems to be the best way to balance this situation.

Summary and Conclusion

This article has pointed out the main differences between the existing port governance models. Each of these models also offers other institutional specificities as a result of adaptation to the local environment. Therefore, the level of efficiency of each of these management models must be observed in their context, both institutional, but also economic and social. Finally, this article has remarked the relevance of port management models in the context of the global supply chain. Ports act as essential nodes for the operation of international commercial exchanges, adapting their administrative structures to these global and interconnected dynamics.

References

- Brooks, M.R. and Pallis, A., 2011. Port Governance. In :W.K., Talley, (Ed.), *Maritime Economics - A Blackwell Companion*.
 Caldeirinha, V.R., Felício, J.A., Da Cunha, S.F., Machado, L., 2018. The nexus between port governance and performance. *Maritime Policy Manage.* 45 (7), 877–892.
 Harlaftis, G., Tenold, S., Valdaliso, J.M. (Eds.), 2012. *The World's Key Industry. History and Economics of International Shipping*. Palgrave Macmillan, London.
 De Langen, P.W., Pallis, A.A., 2007. Entry barriers in seaports. *Maritime Policy Manage.* 34 (5), 427–440.

- Kaukianen, Y., 2014. The role of shipping in the “second stage of globalisation”. *Int. J. Maritime History* 26 (1), 64–81.
- Suykens, F., Van De Voorde, E., 1998. A quarter a century of port management in Europe: objectives and tools. *Maritime Policy Manage.* 25 (3), 251–261.
- Trujillo, L., Nombela, G., 2000. Privatization and regulation of the seaport industry. In: Estache, A., de Rus, G. (Eds.), *Privatization and Regulation of Transport Infrastructures: Guidelines for Policymakers and Regulators*. World Bank Institute Publications, Studies in Development Series, Washington D.C.
- World Bank, 2007. *Port Reform Toolkit*, second edition, Washington D.C.

Further Reading

- Baltazar, R., Brooks, M.R., 2007. Port governance, devolution and the matching framework: a configuration theory approach. *Res. Transport Econ.* 17, 379–404.
- Brooks, M., 2016. Port governance as a tool of economic development: revisiting the question. In: Lee, P., et al. (Eds.), *Dynamic Shipping and Port Development in the Globalized Economy*. Palgrave Macmillan, London, pp. 128–157.
- Brooks, M.R., Cullinane, K., 2007. Governance models defined. *Res. Transport Econ.* 17, 405–436.
- de Langen, P.W., 2004. Governance in seaport clusters. *J. Maritime Econ. Logist.* 6, 141–156.
- Debie, J., Lavaud-Letilleul, V., Parola, F., 2013. Shaping port governance: the territorial trajectories of reform. *J. Transport Geogr.* 27, 56–65.
- Estache, A., Trujillo, L., 2009. Changes in port authorities. In: Meersman, H., Vande Voorde, E., Vanellander, T. (Eds.), *Future Challenges for the Port and Shipping Sector*. Informa Law from Routledge, London.
- González, M.M., Trujillo, L., 2009. Efficiency measurement in the port industry: a survey of the empirical evidence. *J. Transport Econ. Policy* 43 (2), 157–192.
- Pallis, A., Vitsounis, T.K., de Langen, P.W., 2009. Port economics, policy and management: review of an emerging research field. *Transport Rev.* 30 (1), 115–161.
- Trujillo, L., González, M., 2011. Maritime ports. In: Finger, M., Künneke, R. (Eds.), *International Handbook for the Liberalization of Infrastructures*. Edward Elgar, Cheltenham.
- Trujillo, L., Nombela, G., 2003. Puertos. In: Estache, A., de Rus, G. (Eds.), *Privatización y regulación de infraestructuras de transporte*. Alfaomega Grupo, Mexico D.F.

Economic and Environmental Regulation in the Port Sector

Beatriz Tovar*, **Alan Wall†**, *Tourism and Transport Research Unit, Institute of Tourism and Sustainable Economic Development, University Las Palmas, Las Palmas University, Las Palmas de Gran Canaria, Spain; †Oviedo Efficiency Group, Department of Economics, University of Oviedo, Oviedo, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	563
Ports: What are They and What are Their Economic Characteristics?	563
Governance	564
Economic Characteristics of Ports	564
Economic Regulation	565
Price and Access Regulation	565
Port Concessions as a Regulatory Tool	566
Environmental Regulation	566
Conclusions	567
Biography	568
See Also	568
References	569

Introduction

As argued by [Goss \(1990\)](#), the economic function of seaports is to benefit producers and consumers from the trade that passes through them. This implies maximizing consumers' and producers' surpluses, which in turn requires competitive conditions. However, the economic characteristics of ports may result in them enjoying considerable market power, and when this translates into anti-competitive practices, the sheer scale of port activity and the maritime industry more generally means that this will have serious adverse effects on end-users and the economy as a whole. This is underscored by the fact that maritime transport concentrates over 80% of the volume of international trade and over 70% of its value, making it essential for the global economy ([UNCTAD, 2018](#)). In cases where competition cannot be introduced to address this market failure through competition law, economic regulation will be required ([OECD, 2011](#)). Another form of market failure is the negative environmental externalities generated by port activity, and increasing concerns over climate change and the effects of pollution from port activity are driving moves toward more stringent environmental regulation. In this chapter on port regulation we will therefore discuss both economic and environmental regulation.

The chapter proceeds as follows. In the first section we define what ports are, their governance structure, and the economic characteristics that make them a candidate for regulation. The next section is dedicated to economic regulation, where we discuss price and access regulation and the use of port concessions as a regulatory tool. Environmental regulation is addressed in the following section and covers both internationally binding regulatory initiatives as well as voluntary initiatives that have been implemented by port management bodies to mitigate adverse environmental impacts. The last section concludes.

Ports: What are They and What are Their Economic Characteristics?

A discussion of port regulation requires a definition with some precision of what a port actually is. While this may seem obvious, a study by the [European Parliament \(1999\)](#) highlighted that the definition of "port" has changed considerably over time due to technological advances in maritime transport and the increasing role of maritime transport in national and world economies. Thus, the original meaning of port as a harbor, or refuge for ships, has given way to more comprehensive definitions that underline the role of ports in the transportation logistics chain and the different forms of management to which they are subject. This evolution is clearly described in the academic port literature, which classifies ports into five generations, with recent literature even referring to a sixth generation ([Kaliszewski, 2018](#); [UNCTAD, 2000](#)).

The [European Parliament \(1999\)](#) study also suggests a definition of ports as "commercial areas located beside water deep enough for sea-going vessels and having several or many port undertakings [and which] have special maritime as well as conventional road and rail infrastructure, and are supervised or administered by a public or private port authority." Thus, ports can be characterized as providing a wide range of different services using infrastructure, installations and mobile equipment by different agents. Moreover, ports can be owned and managed according to widely different models. The ownership and management structure of ports, the economic characteristics of individual port services and the possibility of whether competition can be realistically introduced will all determine the need for regulation and the type of regulation required. A useful working definition of economic regulation which we

use here is provided by [Decker \(2015\)](#), who defines it as “interventions that impact on the structure . . . or which attempt to guide or control the behavior of firms on terms of their decisions in respect of pricing, investment, quality, and coverage of service, as well as the terms on which access is provided to other firms, including competitors.”

Governance

Port management was traditionally characterized by a central public authority that was responsible for port development and that managed or operated the lion's share of port services. In recent decades, however, technological advances in maritime transport have increased pressure on ports to be able to provide service to larger and more specialized vessels and led to increased competition among ports to attract these vessels. Simultaneously, financial pressure on public sectors had led States to actively seek private sector involvement in the port sector in both the provision of services as well as investment in the construction of port infrastructures. This in turn has led to a need for regulation of private firms that provide port services in contexts of limited competition.

As summarized by the [World Bank \(2007\)](#), four main governance models of ports have emerged over time which the State can use to regulate the sector. These models are the public service port, the tool port, the landlord port, and the fully privatized port or private service port. The models differ with regard to characteristics: who provides the service (public, private, or mixed provision); orientation (local, regional, or global); ownership of infrastructure/port land and superstructure and equipment; and status of dock labor and management. As noted by [Ferrari et al. \(2015\)](#), the key difference between the models has to do with the possibility for private companies to manage port operations and be involved in port activities. Thus, public service ports and tool ports mainly focus on the realization of public interests, whereas at the other extreme, fully privatized ports—mainly found in the United Kingdom and New Zealand—focus exclusively on private (shareholder) interests. Unlike pure public ports, tool ports allow private operators to carry out activities within ports and in port terminals, although they cannot have their own infrastructure. According to the [World Bank \(2007\)](#), this model serves as a useful stepping stone to the landlord model, especially when the confidence of the private sector is not fully established and the investment risk is considered high. Finally, landlord ports aim to balance public (port authority) and private (port industry) interests. In this model, which is the dominant model in medium and large ports, the port authority acts as regulatory body and landlord, while private companies carry out port operations, especially cargo handling. While in principle there is no ideal model, [Monios \(2019\)](#) notes that there seems to be a tacit consensus in the literature on port governance that some version of the landlord model combining public ownership and private operation is the most desirable. [Monios \(2019\)](#) also notes that the terminal itself sits within its own chain of governance as it often forms part of a global portfolio of terminal operations and/or may be vertically integrated with shipping lines, such as APM and Maersk, that have a great deal of market power over individual ports. Hence, there are elements of both vertical (different levels in the transport chain, such as between carrier and terminal and sometimes hinterland transport) and horizontal (with other terminals) integration, that should be taken into account when it comes to port governance.

Economic Characteristics of Ports

Port infrastructure possesses a series of economic characteristics which have a bearing on competition in the sector and the possible need for public intervention and regulation ([OECD, 2011](#); [World Bank, 2007](#)). Among these are the fact that the demand for port services is derived (with the exception of passenger services) from the overall demand for transport between origin and destination. The choice of port will therefore depend on the generalized transportation cost, which comprises not only port service fees *per se* but also the cost of queueing and the transport costs from the port to the final destination of goods. The sensitivity of customers to changes in port services prices will therefore be smaller than their sensitivity to changes in generalized transport costs, raising the likelihood of excess pricing for port services. Another important characteristic is that port services may have few or no substitutes, which leads to low elasticities of demand with respect to prices and quality of service. On top of this, interport competition may not exist where there are captive hinterlands, comprising regions where a particular port may have a substantial advantage over other ports in terms of the generalized cost of transport to these regions. Ports on relatively small islands are a case in point.

Indivisibility of port infrastructure and superstructure, which require minimum sizes and which cannot easily be adjusted to the flow of traffic, is another characteristic and may lead to overcapacity or congestion in the port which can only be adapted to over relatively long periods of time and great cost. The long duration of the useful life of port infrastructure and its high cost generate uncertainty for investors in the sense that the returns from the large investments needed are generated over long time periods in a context of uncertainty with regard to demand for port services. Finally, ports generate externalities, both positive and negative. Positive externalities include spillover effects on local or regional development, whereas negative externalities include the contamination of local waters and beaches as well as air pollution and road congestion in port cities.

These characteristics of ports may lead to market power in some port services and lead to anticompetitive practices, including excessive pricing and refusal to grant access to port services. Also, if external costs are not internalized, ports may generate socially sub-optimal levels of contamination and pollution, leading to environmental degradation. These economic concerns over anti-competitive or “natural monopoly” behavior and the environmental concerns due to negative externalities have led to economic and environmental regulation in ports.

Economic Regulation

Standard microeconomic theory holds that competitive markets lead to Pareto-optimal allocation of resources. Market failures may hinder the achievement of competitive outcomes, and as we have seen, market power arising from natural monopolies may exist in ports, thereby providing a case for economic regulation.

However, even when competition cannot be introduced, care must be taken to ensure that regulation is in the public interest. The justification for public interest regulation is a market failure test based on three components (Church and Ware, 2000): (1) a determination of the existence and magnitude of inefficiencies if the market is not regulated; (2) determination of the feasibility of intervention to correct market inefficiencies; and (3) the benefits of regulation justify the costs associated with regulation, where these costs can be direct (costs of establishing a regulatory body, compliance costs for firms) or indirect (regulation-induced inefficiencies due to lack of incentives for firms to minimize costs).

If regulation is deemed feasible and realistic regulatory tools can be identified, care must be taken to prevent regulation from being overly restrictive or controlling in the sense that it should not deter private companies from providing services nor stifle their ability to introduce innovative and efficient practices. To avoid heavy-handedness in regulation, the World Bank (2007) suggest a series of guidelines to aid in the design of a successful economic regulatory policy for ports which fit in well with the market failure test components above. These guidelines include, for example, that governments should have a clear understanding of the competitive environment of the port sector and that regulatory decisions should be based on the risk of anticompetitive behavior.

The nature of the port service provided will determine the need for regulation. Following Trujillo and Tovar (2007), a distinction can be made between services requiring port infrastructure as a basic input (goods terminals, warehouses and storage facilities, etc.) and those that do not require infrastructure (pilotage, tugboats, etc.). When infrastructure is required, competition will be limited or nonexistent for the corresponding port services and regulation is needed to avoid abusive practices and inefficient behavior. When infrastructure is not required, competition can generally be introduced to the market provided the market is sufficiently large. If the market is small and the number of service providers should be limited, some sort of licensing scheme will generally have to be introduced.

The form of economic regulation (e.g., rate of return or price caps) and the circumstances where it applies should be clearly defined. Also, to avoid the risk of regulatory capture and potential conflict of interest between port operations regulation and competition policy, it is recommended that responsibility for these should be separated and assigned to two different entities. Finally, in terms of the costs of regulation, it is recognized that regulators will have to weigh the trade-offs between the need for information and the burden of the reporting requirements on the operators.

Price and Access Regulation

With the aim of eliminating impediments to the functioning of competitive markets or providing conditions which mimic competitive outcomes, economic regulation comes in two basic forms: price regulation and access regulation. The principal aims of price regulation in ports are to ensure that customers are not exploited through excessive prices and that the regulated firms receive a reasonable return on their investment. For example, the recent “EU Regulation 2017/352 of the European Parliament and of the Council establishing a framework for the provision of port services and common rules on the financial transparency of ports” states that the charges for services provided under public service obligation and services provided that are not exposed to effective competition (such as pilotage services) should be transparent, objective and non-discriminatory and “shall be proportionate to the cost of the service provided” (art. 12). The idea here is to avoid potential abuse of monopoly power, and elsewhere the regulation explicitly states that Member States are permitted “to regulate charges in order to avoid overcharging of port services in cases where the situation of the market in port services is such that effective competition cannot be achieved” (Recital 40, available at: <http://data.europa.eu/eli/reg/2017/352/oj>). Price regulation is usually appropriate when a port or port service is a natural monopoly. More generally, rate-of-return or price-cap methods may be used. Rate-of-return regulation involved setting prices in such a way that firms are permitted to earn a “fair” return on their capital investments. This has the advantage that firms will cover investment costs, reducing uncertainty over investment, but has the disadvantage that it provides little incentive for firms to reduce costs as they can pass increased costs on to customers. It also suffers from the well-known Averch–Johnson effect whereby firms overinvest in capital. Under price-cap regulation, on the other hand, maximum prices for services are set according to some variant of the RPI-X methodology, where the Retail Price Index (RPI) captures the inflation rate and ‘X’ is the ‘offset’ reflecting the firm’s expected efficiency improvements. Under this methodology, firms are permitted to keep cost savings until the review period finishes. The differences between the two methods of regulation have often been described in terms of the different incentives they provide. Thus, rate-of-return regulation is good at extracting excess rents from firms for customers, as prices are set to reflect the cost of service, but provides no incentives for firms to try to reduce their costs. In the terminology of Laffont and Tirole (1993), this scheme provides ‘low-powered incentives’. Price-caps do provide incentives for firms to reduce costs (‘high-powered incentives’) as they are the residual claimants for cost reductions, but this implies that excess rents are not transferred to customers, at least until the following review of the price cap.

Access regulation can be used to address problems of refusal of supply as well as reducing discriminatory practices when ports are capacity-constrained or where the port is vertically integrated with downstream interests. Access is commonly regulated under the “Essential Facilities Doctrine” under which a monopoly that owns a facility essential to competitors must provide reasonable access to this facility (Defilippi and Flor, 2008). Ports have been the principal vehicle for the development of the essential facilities doctrine

by the European Commission—see [OECD \(1996\)](#) for references to some legal cases illustrating the use of the doctrine. Regulators have different tools at their disposal to ensure access is granted, including transparency obligations, accounting separation obligations, publishing of access codes, and specification of equivalence standards governing the principle of nondiscrimination ([OECD, 2011](#)).

Port Concessions as a Regulatory Tool

Under the landlord model, concession agreements can be considered as a policy tool that enables Port Authorities to regulate the port without a direct involvement in the commercial activities of the port. In fact, for Europe it has been argued that port concessions appear to be the only common instrument to regulate the port market. Moreover, they appear to be the only effective means through which Port Authorities can achieve public service goals in the face of increasingly powerful shipping lines and port terminal operators ([Ferrari et al., 2015](#)).

Given their long duration and the existence of unknown future contingencies, concession contracts are problematic in the sense that they are incomplete, that is, the parties cannot express all contractual terms in detail. This in turn raises the issue of which party to the agreement has the right to make decisions over future contingencies and therefore enjoy greater bargaining power in possible renegotiation of terms. This may lead to opportunistic behavior, benefitting one or some of the parties at the expense of social welfare. In terms of concession agreements for port infrastructure, incompleteness of contracts can lead to inefficient or suboptimal investment, anticompetitive behavior and collusion, lack of transparency and clarity in the management of public–private partnerships, regulatory capture, vertical integration, and risk of exclusion of access ([Sánchez and Chauvet, 2019](#)). To ensure that infrastructural concession agreements align with social objectives, a solid institutional framework is therefore needed with efficient regulation and a strong competition policy. The issue is particularly problematic if an infrastructural services concession holder vertically integrates with one or more clients, as happens, for example, when a port terminal operator integrates with a shipping company or large logistics operator leading to the exclusion of other shipping lines from access to port services.

Environmental Regulation

Aside from greater economic efficiency and the attraction of private investment as goals of regulatory policy, there is an increasing emphasis on environmental sustainability in the port sector due to the negative environmental externalities produced by ports as well as the effects of climate change on ports through phenomena such as, for example, rising sea levels. Seaports, as major centers of economic activity that are usually located near highly populated areas, generate significant air pollution for coastal areas and port cities. As generators of greenhouse gas emissions, they also contribute to global warming. It should be noted that the legal framework for measures to combat greenhouse gas emissions extends beyond the maritime industry and any particular geographical area, with relevant international regulatory developments including the implementation of the 2030 Agenda for Sustainable Development and the Paris Agreement under the United Nations Framework Convention on Climate Change.

Emissions by ports and in shipping more generally have been the subject of international regulation ([Tichavska et al., 2019](#); [Tovar and Tichavska, 2019](#)). Particularly important here is the VI Annex of the MARPOL Convention of the International Maritime Organization (IMO) on “Regulations for the Prevention of Air Pollution from Ships.” This was adopted in 1997 and entered into force in 2005 and directly regulates shipping sector emissions of the shipping sector globally. Controls have tightened over time, with limits of sulfur emissions of fuel being reduced to 0.5% from the beginning of 2020. An important innovation of the IMO’s (2019) MARPOL legislation that indirectly affects ports is the creation of Emission Control Areas (ECAs), which are located in areas with high concentrations of coastal population and shipping activity and which apply stricter requirements to fuel content. Thus, whereas sulfur content of fuel is limited to 0.5% of fuel from 2020, in ECAs this limit is 0.1%. An example of regulation more directly affecting ports is the 2014 EU Directive 2014/94/EU that all major ports should seek to install Onshore Power Supply (OPS) (also known as “cold ironing.”), which is the delivery of shore side electrical power to a ship at berth while its main and auxiliary engines are turned off, and LNG facilities.

Ports themselves have been actively adopting strategies to reduce air emissions. In April 2008, the International Association of Ports and Harbors (IAPH) requested its Environmental Committee in consultation with the Regional Organization of Ports for a mechanism to assist ports to abate air pollution and climate change. As a result, 55 ports around the globe adopted the C40 World Ports Climate Declaration, initiating studies, strategies and actions to reduce exhaust emissions and to monitor and improve air quality. In 2014, a new website (www.lngbunkering.org), also promoted by the IAPH, was launched to address the use and bunker of LNG in shipping as an alternative to pursue the reduction of air pollution and negative effects on human health. Moreover, the latest theme for the annual award promoted by European Sea Ports Organization (ESPO) also addresses innovative port projects that lead to environmental improvement for the benefit of the wider port and local community. In Europe, the ESPO provides a platform for port authorities to address environmental management of five selected issues: air quality, energy conservation, noise management, waste management, and water management.

In an effort to go further than IMO regulation, a series of mainly voluntary initiatives have been implemented by port management bodies, such as the Environmental Ship Index (ESI) and the Green Award (GA). The Green Award Foundation was established in collaboration with the Port of Rotterdam in 1994 and GA certificates have been introduced in several ports which provide incentives to encourage compliance with environmental standards beyond existing regulation. Among these incentives are

discounts in port dues and discounts on services. The ESI project, initiated through the World Port Climate Initiative of the IAPH in 2010, identifies seagoing ships that perform better in reducing air emissions than required by the current emission standards of the IMO and is intended to be used by ports to promote clean ships. The ESI formula evaluates emissions from ships based on the amounts of nitrogen oxide (NO_x), sulfur (SO_x), particulate matter (PM), and carbon dioxide (CO₂), and incorporates a reporting scheme on the greenhouse gas emissions of the ship. It also rewards ships that can use Onshore Power Supply (OPS) while at berth. While the majority of ports that have adopted this scheme are European, ports from the United States, Japan, Australia and Korea have also signed up (see <http://www.ppmc-transport.org/environmental-ship-index-part-of-the-international-association-of-ports-harbors-world-ports-climate-initiative>).

In general, environmental charging and differentiation in port and fairway dues according to environmental performance has become an increasingly popular policy instrument in recent years to address negative environmental impacts. Swedish ports, for instance, have adapted to regulation and technical standards by implementing a fairway due rebate system since 1998 (revised in 2005), where dues are differentiated according to the NO_x and sulfur content in fuels. By 2011, 33 major ports in Sweden had introduced the system based on certificates issued by the Swedish Maritime Administration (SMA), and the Finnish ports of Mariehamn and Turku have also joined this initiative. Spain is the only other country in Europe aside from Sweden that regulates environmental charging at national level: in the other countries, these are self-regulated at the level of the port authority. Most of these self-regulated schemes offer rebates on port dues ranging from 0.5% to 20% to vessels that are certified under well-acknowledged indexes and/or certification programs (COGEA, 2017).

The role of differential charging for environmental aims has been also recognized by the recent EU Regulation. Thus, Article 13 of this Regulation states that port infrastructure charges should be set in line with port's commercial strategy and investment plans and comply with competition rules, but permits these charges to be varied in certain circumstances, including more efficient use of port infrastructure and sustainability criteria such high environmental performance, energy efficiency, and carbon efficiency.

A drawback of voluntary schemes is that they may lead to reduced business as a result of increased prices, especially if competing ports act as free riders and do not implement the charging scheme (COGEA, 2017). In order for differentiated port charges to be feasible - in the sense that they do not distort competition and lead to shifts in interport traffic toward free riders - there is a need for a coordinated regulatory framework of a larger area as opposed to a single port authority (Bergqvist and Egels-Zander, 2012). This potential loss of business is in line with the "pollution haven hypothesis" in environmental economics which holds that firms tend to move operations from regions or countries with relatively stringent environmental regulations to others with relatively lax regulations. This issue has been discussed in the context of ECAs by Chang et al. (2018). Coordinated regulation would avoid this problem and permit widespread gains. For example, the study by COGEA (2017) concludes that if external environmental costs are factored in, savings from reduced emissions in ports may outweigh the costs connected with the revenue foregone from reduced port dues.

Emission Trading Schemes (ETS) are also under discussion within the IMO and the EU strategy toward sustainability in shipping. At present, however, policy actions to abate air emissions relate mainly to the quality of fuel and the technological alternatives available. One of the technological options available that directly affects port infrastructure is OPS, which improves the port-city environment in terms of noise, vibration and air emissions. Since 2002, OPS has been implemented in ports in Europe, North America and Asia. The European Union has taken active measures to foster the introduction of the technology, including the co-financing of pilot schemes to facilitate OPS in several Spanish ports (OPS Master Plan: <http://poweratberth.eu/?lang=en>). Despite the technical challenges involved, it appears that this will form part of future regulation in Europe. Thus, a recent document (*The European Green Deal (COM(2019) 640 final)*) published by the European Commission that includes a reference to ports states that "transport should become drastically less polluting, especially in cities . . . the Commission will take action in relation to maritime transport, including to regulate access of the most polluting ships to EU ports and to oblige docked ships to use shore-side electricity."

Finally, the enforcement of environmental regulation depends on where breaches occur. The legally-binding United Nations Convention on the Law of the Sea (UNCLOS) applies at sea beyond national jurisdictions. In ports, on the other hand, vessel compliance with regulation is implemented under the mechanism of Port State Control (PSC), which involves the inspection of foreign ships in national ports to verify that the conditions of the ship and its equipment comply with the requirements of international regulations and that the ship is manned and operated in compliance with these rules (see the IMO website for details: <http://www.imo.org/en/OurWork/MSAS/Pages/PortStateControl.aspx>).

Conclusions

There is an increasing trend in the maritime and port sector toward greater control of logistics chains through mergers, acquisitions and strategic alliances. This has manifested itself in greater vertical integration along the logistics chain as well as greater horizontal integration among terminal operators. This raises the clear possibility of anticompetitive behavior. Moreover, the creation of portfolios of terminal operations at global level and the integration of terminals with shipping lines increase the need for and importance of regulation that can be enforced across wide geographical areas. The recent EU port services regulation is an example of this. As always, regulation must also be careful not to discourage investment, especially in a context where ships are becoming ever larger in size and ports have to compete to be able to provide the infrastructure to cater for these so-called megaships. However, growing environmental concerns at global level and a scenario with powerful global terminal operators that are horizontally and vertically integrated point to a need for coordinated regulation that takes into account the interests of all stakeholders and multiple

levels of governance. Under these circumstances, it is a major challenge for regulators to maintain competitive conditions and ensure environmental sustainability.

Biography

Dr Beatriz Tovar is Full Professor at Las Palmas University and a member of the University Institute for Tourism and Sustainable Economic Development, a leading research team at international level in the field of innovative and sustainable tourist development and a member of its Tourism and Transport Research Unit, a leading research team at international level in the field of Transport. She teaches in undergraduate and graduate (including MBA and PhD) courses regarding the productivity and efficiency measurement for utilities and Transport. She has been actively involved in contracts and research projects with the World Bank, CEPAL, the European Commission or the Spanish Ministry of Transport. She has published in several respected international journals including *Applied Economics*, *Energy Economics*, *Energy Policy*, *European Journal of Transport and Infrastructure Research*, *Fisheries Research*, *Journal of Air Transport and Management*, *Journal of Energy in Southern Africa*, *Journal of Production Economics*, *Journal of Transport Economic and Policy*, *Maritime Economics & Logistics*, *Maritime Policy and Management*, *Research in Transportation Business & Management*, *Tourism Economics*, *Transport Policy*, *Transport Review*, *Transportation*, *Transportation Research Part A*, *Transportation Research Part D*, *Transportation Research Part E*, *Transportation Science*, and *Utility Policy*. She is also a member of the editorial board of *Transportation Research Part E*. She does also act as a reviewer for several international journals such as *Case Studies on Transport Policy*, *Economía Mundial*, *Energies*, *Energy Economics*, *Energy policy*, *European Journal of Transport and Infrastructure Research*, *International Journal of Transport Economics*, *International Journal of Sustainable Transportation*, *Journal of Air Transport and Management*, *Journal of Environmental Management*, *Journal of Shipping and Trade*, *Journal of Transport Economic and Policy*, *Journal of Transport Geography*, *Maritime Policy and Management*, *Research in Transport Economics*, *Research in Transportation Business & Management*, *Revista de Economía Aplicada*, *Science of the Total Environment*, *Sustainability*, *Transport Policy*, *Transportation Research Part A*, *Part D*, and *Part E*, *Utility Policy*, *WU Journal Maritime affairs* and also for international congress such as IAME 2014, IAME 2015, IAME 2016, IAME 2018, IAME 2019. Finally, she is a member of the Editorial advisory board (EAB) for the journal *Transportation Research Part E: Logistics and Transportation Review (TRPE)* and also a member of the Editorial advisory board (EAB) for the journal *Sustainability (Sustainable Transportation Section)*.



Dr. Alan Wall is Associate Professor of Economics at the Department of Economics of the University of Oviedo and a member of the Oviedo Efficiency Group, a leading research team in the field of productivity and efficiency analysis. Alan's research in efficiency and productivity analysis covers several fields including transport economics, health economics and agricultural economics. He has published in several respected international transport journals including the *Transportation Research Part A*, *Transportation Research Part D*, *Journal of Transport Economics and Policy*, *Maritime Economics and Logistics*, *European Journal of Transport and Infrastructure Research*, *Transportation Science*, and *Transport Policy*. He teaches a graduate course in regulation and competition at the University of Oviedo. In recent years, he has held visiting research positions at the University of Connecticut (USA) and the University of Aberystwyth (UK).

See Also

Introduction Transport Economics: Market Failures and Public Decision Making in the Transport Sector; Natural Monopoly in Transport; Economic Regulation/Deregulation and Nationalization/Privatization in Freight Transportation; Environmental Sustainability in Freight Transportation; Shipping and the Environment; Seaports; Port Hinterlands; Seaports as Clusters of Economic Activities; Port Efficiency and Effectiveness; Containerization and the Port Industry; Ports Management; Port Performance; Port Management; Climate Change and Transport Policy; Environmental Issues of Harbors and Shipping

References

- Bergqvist, R., Egels-Zander, N., 2012. Green port dues: the case of hinterland transport. *Res. Transport. Bus. Manag.* 5, 85–91.
- COGEA, 2017. Study on differentiated port infrastructure charges to promote environmentally friendly maritime transport activities and sustainable transportation. Report prepared for the European Commission. <https://ec.europa.eu/./2017-06-differentiated-port-infrastructure-charges-report.pdf>.
- Chang, Y.-T., Park, H., Lee, S., Kim, E., 2018. Have emission control areas (ECAs) harmed port efficiency in Europe? *Transport. Res. D* 58, 39–53.
- Church, J.R., Ware, R., 2000. *Industrial Organization: A Strategic Approach*. McGraw-Hill, New York.
- Decker, C., 2015. Modern Economic Regulation. An Introduction to Theory and Practice. Cambridge University Press, Cambridge.
- Defilippi, E., Flor, L., 2008. Regulation in a context of limited competition: a port case. *Transport. Res. A* 42, 762–773.
- European Parliament, 1999. Sea Port Policy. Available at: [http://www.europarl.europa.eu/RegData/etudes/etudes/join/1999/168252/DG-4-TRAN_ET\(1999\)168252_EN.pdf](http://www.europarl.europa.eu/RegData/etudes/etudes/join/1999/168252/DG-4-TRAN_ET(1999)168252_EN.pdf).
- Ferrari, C., Parola, F., Tei, A., 2015. Governance models and port concessions in Europe: commonalities, critical issues and policy perspectives. *Transport Policy* 41, 60–67.
- Goss, R.O., 1990. Economic policies and seaports. 1. The economic functions of seaports. *Maritime Policy Manag.* 17 (3), 207–219.
- Kaliszewski, A., 2018. Fifth and sixth generation ports. Evolution of Economic and Social Roles of Ports. https://www.researchgate.net/publication/324497972_FIFTH_AND_SIXTH_GENERATION_PORTS_SGP_6GP_-_EVOLUTION_OF_ECONOMIC_AND_SOCIAL_ROLES_OF_PORTS.

- Laffont, J.J., Tirole, J., 2019. *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA.
- IMO, 2019. MARPOL ANNEX VI AN DNTC 2008 with Guidelines for Implementation 2017 Edition. http://www.imo.org/en/Publications/Documents/Supplements%20and%20CDs/English/QQC664E_022019.pdf.
- Monios, J., 2019. Polycentric port governance. *Transport Policy* 83, 26–36.
- OECD, 1996. The Essential Facilities Concept. Roundtable in Competition Policy Series. OCDE/GD(96)113. Available at: <http://www.oecd.org/competition/abuse/1920021.pdf>.
- OECD, 2011. Competition in Ports and Port Services. Organisation for Economic Co-operation and Development – Directorate for Financial and Enterprise Affairs (Competition Committee).
- Sánchez, R.J., Chauvet, P., 2019. Contratos de concesión de infraestructura: incompletitud, obstáculos y efectos sobre la competencia, serie Comercio Internacional, N° 150(LC/TS. 2019/104), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL).
- Tichavska, M., Tovar, B., Gritsenko, D., Johansson, L., Jalkanen, J.-P., 2019. Air emissions derived from passenger vessels in port: does regulation make a difference? *Transport Policy* 75, 128–140.
- Tovar, B., Tichavska, M., 2019. Environmental cost and eco-efficiency from vessel emissions under diverse SO_x regulatory frameworks: a special focus on passenger port hubs. *Transport. Res. D* 69, 1–12.
- Trujillo, T., Tovar, B., 2007. The European port industry: an analysis of its economics efficiency. *Maritime Econ. Logist.* 9, 148–171.
- UNCTAD, 2000. TD/BC4/Ac.7/14.
- UNCTAD, 2018. Challenges faced by developing countries in competition and regulation in the maritime transport sector. UNCTAD Secretariat. https://unctad.org/meetings/en/SessionalDocuments/ciclpd49_en.pdf.
- World Bank, 2007. *World Bank Port Reform Toolkit*, 2nd ed. Washington, DC.

Maritime Route Planning

Johan Woxenius, Department of Business Administration, University of Gothenburg, Gothenburg, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Glossary	570
Introduction	570
Commercial Route Planning	571
Market Planning	571
Service Planning	571
Voyage Planning	572
Regulatory Route Planning	572
Navigational Route Planning	573
Route Selection	573
Operational Routeing	573
Weather and Current Routeing	573
Tramp Shipping	573
Tanker Shipping	574
Crude Oil Tanker Shipping	574
Product Tanker Shipping	574
Dry Bulk Shipping	574
Liner-Shipping	574
Container Shipping	575
Deep Sea Container Shipping	575
Short Sea Container Feeder Shipping	575
RoRo and RoPax Shipping	575
Cruise	575
Summary and Conclusion	575
See Also	576
References	576

Glossary

Liner shipping A regular shipping service offered to different shippers adhering to a fixed and published itinerary
Parcel A shipload/consignment (in shipping)
RoPax Roll-on/passenger (car ferry)
RoRo Roll-on-Roll-off – handling of rolling cargo
Shipper Transport buyer
Shipping line Maritime transport provider
Tramp shipping A shipping service offered on a voyage to voyage basis, often on the spot market

Introduction

The path between origin and destination is obviously at the heart of any transport service. Among the traffic modes, shipping is more restricted than air when deciding on the route but less restricted than road and rail that must stick to paths made up by infrastructure. Navigation on inland waterways, however, follows the principles of the land modes. Equally important and intrinsically connected is the scheduling of services (Christiansen et al., 2013; Norstad et al., 2011) and voyages allocating vessels and deciding upon departure and arrival times, and thus also speed through water and over ground, but also the departure frequency for regular services. When deciding on speed, shipping is less restricted than air that is rather well aligned with the aircrafts' design speed and road with trucks following speed limits set by the infrastructure provider but also temporary road conditions. Rail can theoretically and technically vary the speed considerably, but the mode is hampered by the rigidity of time tables, coordination between trains, and propagating delays in the network.

The first step of maritime route planning is to find and select the customer and the goods (or passengers) to transport. The information about the character of the goods flow in terms of loading and discharging locations, type and amount of goods/

passengers, regularity and the shippers' preferences for transport time, timing, frequency and other logistics parameters is then used for deciding on vessel type and size, potential mix with other cargo or passengers, operating speed, scheduling and, of course, the appropriate coarse route between ports. Using the tools of logistics, this activity is generally performed by the shipping lines' office staff and is here denoted *commercial route planning*.

International shipping is an industry that can rather easily move its registration, management, administration, operations and assets to where the business opportunities are considered best. The main reason for moving the firms' registration is to utilize a competitive tax or subsidy system, and registering vessels in flags of convenience is generally motivated by a more relaxed regulatory framework. Registration of firms and vessels, however, is rarely decisive for vessel routing but *regulatory route planning* considers parameters such as regional and local regulation of fuel, emissions, vessels' technical standard, crew size, and working conditions. In addition, various political conflicts such as trade wars and boycotts restrict how ships can be routed.

The general idea of how to transport the goods between ports stemming from commercial and regulatory route planning is then fine-tuned by vessel crews weighing in information and forecasts of currents, weather, waves, ice, congestion in ports and canals, risks of piracy and political conflicts and other short-term operational parameters for creating a more detailed voyage plan. In this encyclopedia entry, it is referred to as *navigational route planning*.

Some aspects of maritime route planning are generic but the discussion must be broken down into shipping segments for further understanding. The main differences between tramp and liner shipping relate to the number of shippers and connected ports, parcel sizes and the regularity and rigidity of the service. By definition, tramp shipping is offered on a voyage-by-voyage basis whereas liner shipping depends on a network for offering a regular multiuser service. Furthermore, the conditions for route planning are quite different for the tramp shipping segments tank and dry bulk; and the liner shipping segments container, RoRo, RoPax, and cruise. This text is mainly about freight transport, which is the main task for the maritime mode, but passenger transport is included in the RoPax and cruise segments.

The text is structured with a general discussion in three sections on commercial, regulatory and navigational route planning. The following section deepens the rendering in the context of tramp shipping divided into tanker and dry bulk shipping. More specifics on route planning in liner shipping is then provided in a section further divided into the market segments of container, RoRo/RoPax and cruise shipping. The text is finally concluded including some thoughts on how maritime route planning is likely to evolve in the coming years.

Commercial Route Planning

This section deepens the discussion on logistics performance preparing for the more detailed rendering for each shipping market segment. Focus is on the transport service and the goods/passengers rather than the vessel operations.

Market Planning

Before narrowing into the route of a particular voyage, a number of scoping decisions must be taken. A shipping line must first decide upon the general maritime transport market it wants to serve including main shipping segment, port geography, type of cargo or passengers, whether to serve a customer at a time or finding a compromise addressing several customer demand patterns and whether to offer a premium service or cutting corners to offer lowest possible price. A shipping line with wider ambitions might also consider offering door-to-door services rather than confining the market offer to port-to-port. Setting these main parameters of the offer is a strategic activity with a long time horizon that can be denoted market planning.

Service Planning

When the market planning has set the wider boundaries, the shipping line needs to decide upon further details of the offer including which principles to guide the operation of the vessel or fleet to satisfy the shipper or shippers and be competitive and yet securing a positive economic return. Service planning addresses the character of the transport offer at a medium time horizon. Liner shipping is generally more complex than tramp, but not necessarily as planning the ballast voyage repositioning a tanker to a port range with a positive market sentiment when arriving there requires much experience and significant analytic capacity, whereas operating a regular RoPax bridge substitute is fairly straight forward.

Network operation principles can be used to reduce complexity in transport networks. **Fig. 1** shows six network configurations for connecting an origin to a destination directly or via other nodes to facilitate consolidation of people or goods. The configurations assume that nodes can be connected as the crow flies, but in reality the detours taken through nodes are often over-layered by detours taken for other reasons as elaborated on by [Woxenius \(2012\)](#).

Direct link is obviously the simplest configuration and the evident example is a taxi service but much of tramp shipping follows the same principle as do bridge substitute RoPax services. The role model for the corridor configuration is a bus or inter-city rail service with stops along the route. Coastal shipping, container feeder services and barge services on inland waterways apply a similar configuration and so does container shipping between East Asia and Europe with a number of port calls en route. Hub-and-spoke is mostly applied in air transport but there are a few maritime examples found in archipelago ferry services for commuters. Connected hubs is modeled on road-based general cargo consolidation networks, but also applied by inter-continental services between second

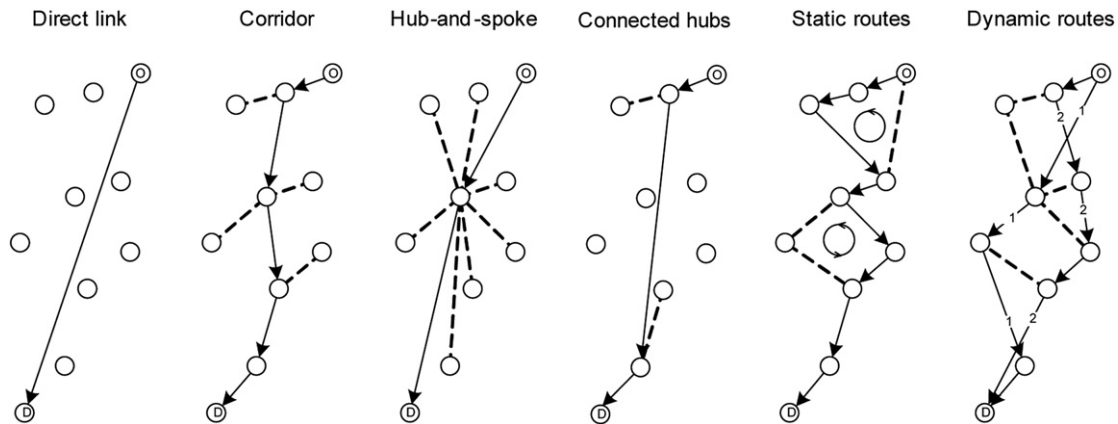


Figure 1 Six options for transport from an origin (O) to a destination (D) in a network of ten nodes. Dotted lines show operationally related links in the network designs. In “Dynamic routes”, two alternative routes are shown; in all other designs, the routing is predefined. Source: Woxenius (2007).

tier airports. It also resembles the way transpacific and transatlantic container services are operated with deep sea vessels connecting a few hub ports, from which short sea feeder vessels connect to smaller ports. The container feeder services themselves often operate along the principle of static routes with a fixed itinerary combining a number of ports, of which some are hub ports, like in South-East Asia, the Mediterranean or the Baltic Sea. The most evident example, however, is a public transport service in a bigger city, in which passengers can find their way forward by using different tram, subway and bus lines making up a network. In dynamic routes, finally, transport planners start every new cycle with the current transport demand and optimize the routes. An airport hotel shuttle constitutes a passenger transport application and a typical shipping example is product tankers distributing petroleum product parcels from a refinery to distribution depots along the coast.

Voyage Planning

Unless it is a regular liner shipping departure that follows a pre-set plan, each voyage is coarsely planned in the shipping line office including selecting the particular vessel to use, the main route with order of port calls, speed, deciding on whether to use canals and waters requiring pilots or icebreakers as well as avoiding pirate ridden or war-torn regions. This is a major activity in tramp shipping and it is also performed for ships in ballast repositioning for the next laden voyage. Liner shipping stowage planning is also affected if there are ports along the route that have draft limitations prohibiting the vessel to call if it carries a too heavy load.

Regulatory Route Planning

The general right to navigate on the high seas has a long history. The idea was not new but it was formulated by Hugo Grotius in his work *Mare Liberum* in 1609 but it took considerable time to get the principle generally adopted. By the end of World War 1, when German submarines had launched unrestricted submarine warfare as a response to the British naval blockade of Germany, both actions challenging the Freedom of the Seas, Woodrow Wilson emphasized the right to navigate as the second out of 14 points for reaching a lasting peace:

“Absolute freedom of navigation upon the seas, outside territorial waters, alike in peace and in war, except as the seas may be closed in whole or in part by international action for the enforcement of international covenants” (Wilson, 1918).

Although the general rule is to navigate freely on the oceans, shipping, like all businesses, has to adhere to regulations and some of the constraints are relevant for route planning. Some regulation strictly limits routing, whereas other increases costs of passing certain waters or calling ports and is thus seen as an option fed into commercial route planning.

One example is the Sulfur Emission Control Areas (SECAs) stipulating the maximum sulfur content of the fuel, other types of fuels or use of cleaning technologies to get an equivalent result. Sulfur emission rules differ between seas, which can motivate shipping lines to navigate around an area with stiff regulation. Another example is the slowly stiffening regulation on CO₂ emissions. In 2018, the European Union (EU) implemented the Monitoring, Reporting, and Verifying (MRV) system, which requires shipowners and operators to report CO₂ emissions from their ships both per traffic work (in g CO₂ per nautical mile) and per transport work (in g CO₂ per ton-NM). It includes emissions from the last port call before and first port call after calling a EU port. The MRV system is likely to be the basis when including shipping in the EU Emission Trading Scheme (EU-ETS) and there is hence an incentive to add port calls just outside the EU during longer voyages to reduce the share of emissions that has to be traded under EU-ETS.

A US example of regulatory route planning relates to cabotage. For a hundred years, Merchant Marine Act of 1920, better known as Jones Act, has required that transport between two US ports is carried out by a ship built and flagged in the US as well as owned and crewed by US citizens. Jones Act thus restricts routes combining international and domestic shipments.

Yet another example is that routeing in the Eastern Mediterranean is hampered by difficulties to include port calls in Israel and Arabic states in the same route considering the political tension in the Middle East. Also technical and working condition requirements can differ in different waters and thus affect route planning.

Navigational Route Planning

This section narrows the focus to the day-to-day operations of vessels, which is often done by the crew but increasingly assisted by shipping line offices, information sharing between vessels and by new tools for decision support.

Route Selection

The main route is defined before departure by updating the result from the voyage planning considering more short-term conditions in canals, congestion and labor market conflicts in ports, forecasts for weather, currents, ice conditions and other temporary conditions.

Operational Routeing

During the voyage, the crew takes logistics parameters into account and puts it into an operational reality fine-tuning the route and speed to save fuel but also for operating safely or to catch up lost time (Wang and Meng, 2012). Information is gathered from shipping line offices, from other vessels in the fleet and from other organizations involved in the transport chain as analyzed by Andersson and Ivehammar (2016). Examples of actions are to cancel or change order of port calls, negotiate alternatives with shippers and ports, book slots with canal authorities, pilots, tug and icebreaker operators. During the corona pandemic, for instance, crew replacement was severely hampered forcing staff to stay onboard and finally required ships to reroute to call ports allowing for crew changes.

Operational routeing is also affected by the position and trajectories of other vessels. Radar systems have assisted navigation for decades and transparency is improved by transponders feeding vessel identity, position, velocity and some operational data into the Automatic Identification System (AIS). In addition, traffic management authorities and onshore pilots assist and restrict crews navigating in ports and narrow straits with heavy traffic.

Weather and Current Routeing

In early days of shipping, weather, currents and ice were strict conditions for navigation and mother nature also defined the shipping seasons. Cyclic currents like tides and El Niño and seasonal weather patterns like monsoons in the Indian Ocean and harsh winter storms in the North Atlantic are weighed in already during commercial route planning, but with a short time-horizon crews must align the route and speed over ground to more temporary storms, shifting currents and ice conditions. The water level of inland waterways varies over time, which is obviously closely monitored by barge operators.

Weather routing is the aspect of route planning with most focus on decision support tools (Zis et al., 2020). Weather forecasts are easily accessible, but shipping lines often subscribe to specialized forecasts that feed into the navigation system or to dedicated weather routing applications. The precision of forecasts has increased tremendously with better data capture by satellites, buoys and radar stations on land but also through better meteorology algorithms. Furthermore, data on weather, waves, ice and currents but also operational data is increasingly distributed within fleets to improve accuracy (Andersson and Ivehammar, 2016) but the data can also be distributed by authorities or specialized firms on a commercial basis. This is particularly important to smaller shipping lines.

Weather routing is not only used for reducing bunker consumption, improving safety and crew comfort but also to prolong technical lifetime of ships. One example is that Atlantic Container Line sponsored research on weather routing and used the results to keep operating safely in the North Atlantic with its G3 container-RoRo vessels, which had a reduced structural strength after sailing many years in tough conditions.

If weather or other conditions are particularly harsh prohibiting the crew to keep up the schedule, the new conditions are fed into commercial route planning deciding on options for rerouting or rescheduling.

Tramp Shipping

The majority of tramp shipping constitute an integrated part of large-scale industrial systems adapted to maritime transport with a rail link from a mine or pipeline from a well to the nearest port and the next step of refining the raw material, for instance a steel mill, coal-fuelled power plant or refinery located close to water with an industrial port. Although the industrial system including the

maritime part is comparatively stable over long periods of time, shipping is highly volatile as tanker and dry bulk carriers are generally traded on the spot market.

This section deepens the discussion in the setting of tramp shipping highlighting some specifics for routing of tankers and dry bulk vessels and some characteristic decision situations that occur. The route planning horizon is often relatively short for tramp shipping and scheduling simple when it is provided as just one voyage between two ports for a single shipper. The detailed plan and deviations from it can be negotiated with the customer and the consequences of each option is rather easy to estimate.

Tanker Shipping

For petroleum tanker shipping, the big divider in logistics as well as chemical terms is the refinery. Large single-load tankers bring crude oil to the refinery and smaller product tankers distribute a different oil products to depots closer to the final market. The segments follow different commercial route planning logics. One example of regulatory route planning particular for tankers is that the US Government, deterred by the disastrous effects of the Exxon Valdez oil spill in 1989, required all new tankers calling US ports to be equipped with a double hull. Until the international regulation kept up, it was common for single hull crude oil tankers to call Caribbean ports with refineries and then let newer double hull tankers feed into US ports rather than have a direct route from the oil well to the US port requiring double hull vessels the full distance.

Crude Oil Tanker Shipping

A peculiarity of crude oil shipping is that the load is subject to speculation and can shift owner several times during the long voyage. Following the logistics strategy postponement, the crude oil tanker is often sent in roughly the right direction and then rerouted on the way as the final destination is set by speculators. For crude oil from the Middle East, for instance, final routing decisions for East Asia are typically taken when ships pass the Strait of Malacca and when passing Bay of Biscay for Northwestern Europe.

Voyage in ballast also requires route planning and it is often possible to find a shorter route, for instance very large crude carriers (VLCCs) passing the Suez Canal, which is too shallow for fully loaded VLCCs.

Product Tanker Shipping

Product tanker shipping involves an intricate cargo planning. The hold of the vessel is divided into several tanks that can carry different petroleum products and often also chemicals. Cleaning the tanks is very expensive and the frequency is minimized by planning routes allowing for a sequence of loads from cleaner to dirtier fractions and then cleaning and starting over with clean products. The planning is done for each tank and requires much experience and knowledge of the market to get an efficient sequence of loads.

Dry Bulk Shipping

Dry bulk shipping lines trade in a mix of industrial shipping with long contracts of affreightment for a set transport task and the highly volatile spot market. Dry bulk shipping often has a high portion of ballast voyages but it is more flexible to find return loads than crude oil shipping. Some vessels are specialized for specific cargoes such as iron ore, coal or grain and some are built as flexible combination carriers for oil and dry bulk. The benefits of a specialized or very flexible ship is, however, difficult to reap (Stopford, 2009) and most dry bulkers are built for fitting the requirements of the major bulk commodities to relax the restrictions in route planning. An exception is when shipping lines get contracts of affreightment spanning over the full vessel life cycle such as the one between the shipping line COSCO and the mining company Vale. It covers annual transport of 16 million tons of iron ore from Brazil to China for 27 years from 2018. To put it into perspective, the transport work (in ton-kms) equals 5 years' of transport by all trucks in the EU. A transport task with such a time horizon implies that route planning is more or less confined to navigation. Dry bulk shipping is further elaborated on in the entries 10255: Bulk Shipping Markets: An Overview of Market Structure and Dynamics and 10488: Maritime company freight management/Bulk industry.

Liner-Shipping

Liner shipping is characterized by the use of a fixed and published itinerary (Stopford, 2009), multiple shippers and a strong dependency between vessels serving the same route. It can also be regarded as an art of finding a good compromise between multiple shippers' preferences, considering a wide set of restrictions and yet operating at a profit. Liner vessels sail with a certain regularity and depart although they are not fully loaded. Each ship simply has to fulfill its task in the itinerary. Some cargo like cars and trucks is transshipped using their own wheels, but most cargo is stowed into load units with a unitized interface toward vessels, port equipment and land vehicles. There is hence less focus on the particular commodities but rather the load units in terms of containers, cassettes, semi-trailers and accompanied trucks.

Liner shipping has a stronger connection to hinterland transport compared to tramp shipping. It means that logistics service providers can route the cargo via different ports and shipping services combined with rail shuttles and road haulage to serve a door-to-door market (Tongzon, 2009). Depending on the system level analyzed, the route planning can be very complex including port

choice (Slack, 1985), positioning of empty containers (Song and Dong, 2012) and optimization is a very popular research topic (Christiansen et al., 2013; Wang and Meng, 2012).

Container Shipping

When the container was widely introduced in the 1960s, it was nothing less than a revolution for international trade. It affected routing in the sense that the traffic was concentrated to fewer ports and although many ports have since invested in container cranes, there are prevailing economies of scale stemming from the use of large container vessels connecting a few mega-hubs connected to smaller ports with feeder vessels as in the connected hubs route option in Fig. 1. Container shipping is further described in the entries 10254: Container (liner) shipping, 10482: Containerization and the Port Industry and 10266: Scheduling of Liner Container Shipping Services.

Deep Sea Container Shipping

For routing in intercontinental container shipping, geography and demography affect the tradelanes differently. The transpacific and transatlantic tradelanes are not very restricted by geography but there are, on the other hand, few options for port calls en route so the connected hubs option of Fig. 1 dominates. At the East Asia-Europe tradelane, however, geography funnels traffic through the Strait of Malacca, the Suez Canal and Gibraltar Channel and vessels sail close to shore also when doubling the Indian subcontinent, North Africa and the Iberian Peninsula. It implies that the route can include several port calls in densely populated areas without deterring detours (Woxenius, 2012). The intermediate port calls, in turn, can connect to regional feeder systems serving Southeast Asia, the Indian subcontinent, East Africa and the Mediterranean. Container shipping between East Asia and Europe hence follows the principles of the corridor routing option of Fig. 1 with some elements of hub-and-spoke. As containers are not only moved between East Asia and Northwestern Europe but also between ports along the route, stowage planning easily gets complex. Scheduling is also affected as just a smaller share of the containers are transhipped at each port, comparing to transpacific and transatlantic services with just a few ports at each coast leading to long time in port.

Short Sea Container Feeder Shipping

Feeder shipping operates under rather flexible schedules and feeder routes are frequently changed in terms of which small ports to call and in what order. If too few containers are booked to or from a port, the call can be cancelled at short notice and containers either have to wait for the next departure or moved by land modes to another port. Short sea feeder shipping thus resembles the routing option dynamic routes in Fig. 1 and if it is used for fully intra-regional transport it can use the hub port for hub-and-spoke routing. It should be noted though that also RoRo vessels move containers and increasingly so within Europe.

RoRo and RoPax Shipping

Roll-on-Roll-off (RoRo) move unaccompanied rolling cargo between and within continents and Roll-on-Passenger (RoPax) combine accompanied cargo and passengers on shorter routes. The main commercial route planning issue is whether to minimize the distance over water, often referred to as bridge substitutes, or trying to call ports as close as possible to the main origin and destination of freight and passengers fully using the benefits of maritime transport. This is a multifaceted problem for RoPax services with a truly wide variety of customer demands to consider. Route or speed changes affect voyage time, which in turn can require adaptation of turnaround and departure times and vessel reconfiguration with trade-offs in terms of lane meters versus passenger space, seats versus cabins, and the dimensioning of restaurants, bars and shops. RoPax bridge substitute offer high frequency and operate with short turnaround times whereas RoRo services often are strictly coordinated with the main shippers' industrial system acting as a moving belt.

The main routes are stable over decades and are almost regarded as infrastructure whereas shipping lines can challenge the incumbents with new routes or capture a market opportunity for a season. In some instances, ferry shipping also make detours to ports that are not motivated by the amount of cargo or number of passengers stepping on or off, but to grant the legal right to sell tax-free items. Examples are Åland Islands in the Baltic and Norway in the North Sea. Short crossings often reduce weather routing to the decision whether to cancel a departure for safety and comfort for passenger and freight.

Cruise

Perhaps not downright liner shipping, but the route of cruise ships, finally, is definitely planned ahead. The route planning is based on port visits rather than moving items between ports and often the passengers depart and arrive in the same cruise hub port. Hence, the routing is not necessarily directly between ports, but detours are routinely taken for beautiful scenery like Norwegian fjords.

Summary and Conclusion

This entry discusses how the character of the transport demand, regulation and mother nature affect maritime route planning in general but it has also brought up special routing conditions for individual shipping segments. The route options change over time

as canals are built or extended, rivers and straits silt and, unfortunately, global warming opens up for new routes in the Arctic like the Northern Sea Route between East Asia and Europe (Liu and Kronbak, 2010).

As route and schedule is so decisive for customer satisfaction and costs, voyage optimization systems increasingly support routing decisions by integrating and analyzing a truly wide set of data from sensors onboard, from other vessels in the fleet and from public and private data providers. The improved access to data and the complexity of routing have led to that decisions are increasingly moved to the shipping line offices as master mariners seem to speed up too much not risking delays, but the offices have more comprehensive knowledge needed for deciding on the trade-offs between revenues and costs. But some decisions stay onboard as the master mariner is responsible for the ship, the cargo/passengers and the crew's safety and well-being. Unmanned and automatically navigated ships will change maritime route planning dramatically, but that lies a bit into the future.

See Also

Transport and International Trade; Freight Network Modelling; Freight Vehicle Routing & Scheduling; Freight Mode Choice; The Value of Time in Freight Transport; Container (liner) Shipping; Bulk Shipping Markets: An Overview of Market Structure and Dynamics; Ferries and Short-Sea Shipping; Seaports; Port Hinterlands; Scheduling of Liner Container Shipping Services; Route Choice and Network Modeling; Air Route Planning & Development; Introduction to Maritime Transport; The History of Maritime Transport; The Geography of Maritime Transport; The Future of Maritime Transport; Maritime Transport and Territorial Scale; Containerization and the Port Industry; Maritime Company Passenger Management/Liner Industry; Maritime Company Freight Management/Bulk Industry; Cruise Industry; Bus Route Planning

References

- Andersson, P., Ivehammar, P., 2016. Cost benefit analysis of dynamic route planning at sea. *Transport. Res. Proc.* 14, 193–202.
- Christiansen, M., Fagerholt, K., Nygreen, B., Ronen, D., 2013. Ship routing and scheduling in the new millennium. *Eur. J. Oper. Res.* 228, 467–483.
- Liu, M., Kronbak, J., 2010. The potential economic viability of using the Northern Sea Route (NSR) as an alternative route between Asia and Europe. *J. Transport Geogr.* 18, 434–444.
- Norstad, I., Fagerholt, K., Laporte, G., 2011. Tramp ship routing and scheduling with speed optimization. *Transport. Res. Part C: Emerging Technol.* 19, 853–865.
- Slack, B., 1985. Containerization, inter-port competition, and port selection. *Maritime Policy Manage.* 12, 293–303.
- Song, D.P., Dong, J.X., 2012. Cargo routing and empty container repositioning in multiple shipping service routes. *Transport. Res. Part B: Methodol.* 46, 1556–1575.
- Stopford, M., 2009. *Maritime Economics*, 3rd Edition. Routledge/Taylor and Francis, London.
- Tongzon, J.L., 2009. Port choice and freight forwarders. *Transport. Res. Part E: Logistics Transport Rev.* 45, 186–195.
- Wang, S., Meng, Q., 2012. Sailing speed optimization for container ships in a liner shipping network. *Transport. Res. Part E: Logistics Transport Rev.* 48.
- Wilson, W., 1918. Address of the President of the United States, delivered at a joint session of the two houses of Congress, January 8, 1918. [Govt. print. off.], Washington.
- Woxenius, J., 2007. Generic framework for transport network designs: Applications and treatment in intermodal freight transport literature. *Transport Rev.* 27, 733–749.
- Woxenius, J., 2012. Directness as a key performance indicator for freight transport chains. *Res. Transport. Econ.* 36, 63–72.
- Zis, T.P.V., Psaraftis, H.N., Ding, L., 2020. Ship weather routing: a taxonomy and survey. *Ocean Eng.* 213, 1–18.

Inland Waterway Transport and Inland Ports: An Overview of Synchronomodal Concepts, Drivers, and Success Cases in the IWW Sector

Behzad Behdani*, Bart Wiegmans†, Yun Fan*, *Operations Research and Logistics Group, Wageningen University, Wageningen, The Netherlands; †Department of Transport and Planning, Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	577
Intermodal Freight Transport: Definitions, Types, and Challenges	577
Definition of IFT	577
Different Types and Classification of IFT	578
The Position of Inland Ports and Terminals in Inland Waterway IFT	579
Challenges for Intermodal Transport	579
Synchronomodality Defined	580
Challenges and Drivers for Synchronomodal Transport	581
Cases on Synchronomodal Transport in Inland Waterway Sector	585
Summary and Conclusion	585
References	585

Introduction

With the introduction of the container in the 1960s and high growth numbers since then, also the transport parts preceding and following sea container transport have gained in volumes and importance. Initially, these pre- and end-haulage parts of the total transport supply chain took place by unimodal transport, which was most times truck. Since these early days, container transport has grown enormously, leading to also enormous growth in the fore- and hinterland and the use of the inland ports. Truck transport is very suitable to handle the transport in the fore- and hinterland given its good quality, high speed, reliability, and low price. However, its success has also caused certain problems associated with truck transportation to grow to unsustainable levels. Especially external effects such as emissions, noise nuisance, and congestion have grown to unsustainable levels. In response, policymakers have sought to develop policies aiming for a modal shift from single-modal truck to multimodal rail and inland waterway transport. In these multimodal transport solutions, the major part to and from the seaport takes place via rail or inland waterway, next the containers pass an inland port, while the first and last part of the trip to and from the customer usually takes place by truck. The idea is that trucks are disappearing from roads and are being replaced by more sustainable transport solutions. Although this policy has not been very successful in improving the market share of intermodal freight transport (IFT), growth has caused all transport systems to be quite overloaded. Road transport experiences high congestion, rail transport has only limited space for additional rail freight transports and inland waterway transport shows increasingly congestion at locks and bridges. Inland ports and inland container terminals form important elements of the IFT solution and also these ports and terminals do need to cope with limited space and seek efficiencies in their existing operations in order to accommodate further growth of freight flows or expand further, sometimes even by adding different locations. So all in all, transport modes are overloaded and also inland ports and terminals show signs of congestion and often limited expansion possibilities at current sites. This is exactly the moment when synchronomodality starts to play a role in IFT. Synchronomodality seeks a balanced and integrated approach toward all three transport modes (inland waterway, rail, and truck) where the aim is to use rail and inland waterway transport including inland ports to the maximum extent possible and only use the truck when no other options are available. In addition, planning plays a crucial role and this planning takes place at different levels such as freight flows, modes, transport means, capacities, terminals, etc. Therefore the aim of the chapter is to analyze the characteristics of IFT and inland ports and terminals. Next, synchronomodality will be defined and challenges and drivers for IFT will be discussed. Finally, a number of successful cases of synchronomodality will be presented in order to show the potential of synchronomodality for the successful further development of inland waterway transport, inland ports, and terminals.

This article is organized as follows: the first section gives the introduction. The second section discusses the definitions, types and challenges regarding IFT. The third section introduces the new concept of synchronomodality. The fourth section analyses the challenges and drivers for synchronomodal transport, while the fifth section analyses several cases regarding these drivers and challenges. The article summarizes and concludes with the last section.

Intermodal Freight Transport: Definitions, Types, and Challenges

Definition of IFT

IFT (usually involving the transport of intermodal load units such as containers, trailers and continental load units) can be defined in several different ways. First, intermodal transport can be defined as “The movement of goods in one and the same loading unit or road vehicle, which uses successively two or more modes of transport without handling the goods themselves in changing modes”

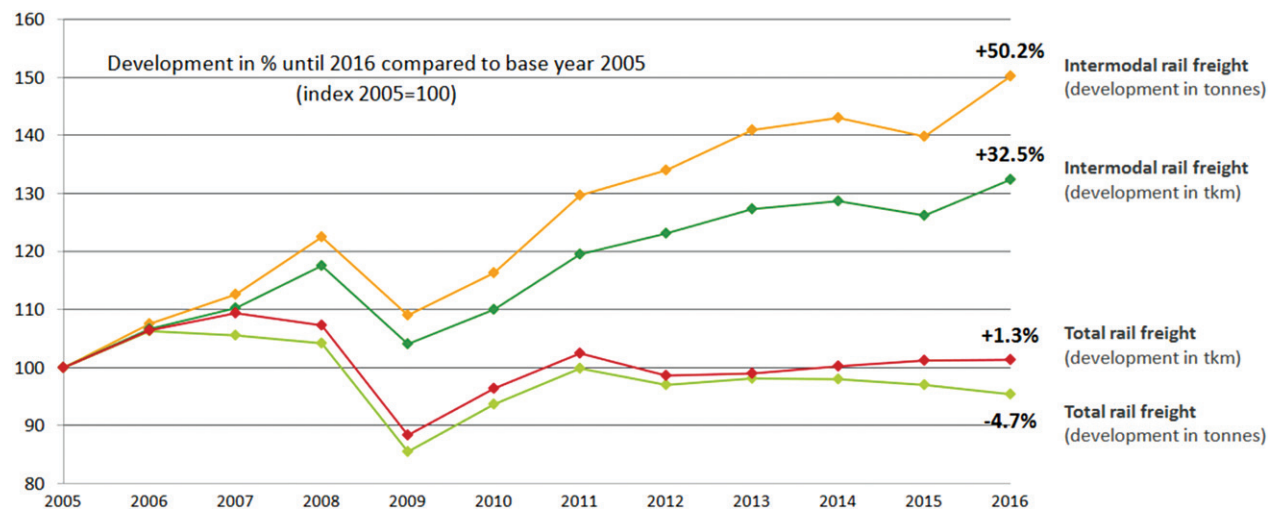


Figure 1 Main important aspects of intermodal freight transport. Source: Based on Wiegmans and Konings (2015).

(<https://uic.org/combined-transport>). A shorter version of this definition is “door-to-door transport that uses two or more transport modes.” Second, intermodal transport enables the transfer of unit loads from the place of origin to the final destination by combining two or more different transport modes (road, rail, inland waterway, sea, and air). Road transport is then often used for local collection and/or distribution only (www.hupac.com). Morlok and Spasovic (1994) define intermodal transport as “the movement of highway trailers or containers by rail in line-haul between rail terminals and by tractor-trailers from the terminal to receivers (consignees) and from shippers to the terminal in the service area.” In these definitions, it can be observed that transport modes and load units are important elements of the definition of IFT. Jones et al. (2000) define intermodal transport as “the shipment of cargo involving more than one mode of transportation during a single seamless journey.” Hayuth (1987) defines intermodal transport as “the movement of cargo from shipper to consignee using two or more different modes under a single rate, with through billing and through liability.” Especially interesting about these last two definitions is the aspect of single journey and single rate, which connects well with the characteristics of synchromodal freight transport (SFT). In the Fig. 1, the main aspects of IFT are depicted. Incoming freight flows arrive at a maritime container terminal. Next the main transport takes place (via inland waterway) and next the inland container terminal handles the containers. Finally, the end haulage takes place towards the final destination. See also the chapter by Macharis et al. (2020), “10285: Intermodal and synchromodal freight transport.”

Different Types and Classification of IFT

In theory, two types of IFT exist: rail-oriented IFT, where the main haul takes place by rail and inland waterway-oriented IFT, where the main haul takes place by inland waterway. In Europe intermodal rail freight transport is the most important transport mode after truck transport.

The developments for rail freight transport are not that positive in an area of high growth of overall freight transport and in the light of the aimed for market share increase of rail freight transport (https://uic.org/IMG/pdf/2018_report_on_combined_transport_in_europe.pdf). Over the period, tons transported by rail have decreased by almost 5% and the ton-km only slightly increased. Intermodal rail freight showed impressive growth over the period, however, this part of the market carries only a little importance in terms of tons. Important parts of the transported freight by rail are formed by dry bulk, liquid bulk, intermodal transport units, and wagonloads. Intermodal transport units can be further subdivided into swap bodies, containers, trailers on a train, and palletized wagonloads. Strong points of rail IFT are its competitive price if door-to-door solutions are possible and if distances are relatively long, its high level of safety, its fuel efficiency, its ability to carry large volumes, and rail IFT is less harmful to the environment than truck transport. Weak points of rail IFT are its lack of accessibility and flexibility, its bad image, its unreliability, its relatively low speed, and its lower suitability for smaller shipment and smaller load units. Improving the competitive position of rail IFT proves to be very challenging over the last decades. Research by among others Ferreira (1997), and FitzRoy and Smith (1995) show that crucial elements are faster transit times, higher frequencies, and more reliability. In order to realize this, large increases in both the number of connections and the operating costs are required for which the funding and the political approval lack. For a more detailed discussion on rail freight see the chapters “10281: Rail freight” by Woodburn (2020) and “10283: Eurasia rail freight: enablers and inhibitors of future growth” by Templar and Rodemann (2020).

After trucking and intermodal rail freight transport, intermodal inland waterway transport is the third important transport mode. The inland waterway infrastructure network is 40,000 km long (rivers, canals, and lakes), of which almost 20,000 km is concentrated in Germany, Belgium, France, Austria, and the Netherlands (<http://www.inlandnavigation.eu/what-we-do/facts-figures/>). Also 20,000 km of the network is accessible for 1,000 ton vessels while 75% of all inland waterway traffic crosses borders. On a yearly

Table 1 Inland waterway traffic and volumes 2015–2016

	Thousands of tons			Millions of ton-km		
	2015	2016	%	2015	2016	%
EU28	548,953	554,022	+0.92	147,471	147,319	−0.10
Austria	8,599	9,071	+5.49	1,806	1,962	+8.64
Belgium	188,158	192,938	+2.54	10,426	10,331	−0.91
Bulgaria	17,201	17,467	+1.55	5,595	5,477	−2.11
Croatia	6,642	6,409	−3.51	879	836	−4.89
Czech Rep.	850	832	−2.12	33	36	+9.09
France	63,094	65,162	+3.28	8,516	8,307	−2.45
Germany	221,369	221,349	−0.01	55,315	54,347	−1.75
Hungary	8,163	8,224	+0.75	1,824	1,975	+8.28
Italy	379	407	+7.39	62	67	+8.06
Luxembourg	7,106	6,075	−14.51	235	190	−19.15
Netherlands	359,898	365,510	+1.56	48,535	49,398	+1.78
Poland	5,036	3,911	−22.34	88	108	+22.73
Portugal	31	31	0			
Romania	30,020	30,484	+1.55	13,168	13,153	−0.11
Slovakia	5,721	6,758	+18.13	741	903	+21.86
United Kingdom	4,065	3,804	−6.42	120	108	−10.00

Source: <http://www.inlandnavigation.eu/>

basis, the barge fleet transports around 550 million tons of freight. Important sectors for inland waterway transport are dry bulk, liquid bulk, containers, general cargo, Ro/Ro, special transport, and push barges. For the Netherlands, inland waterway transport is very important and represents the second mode of “national” transport (in terms of ton-km) after road (Table 1).

Strong points of inland waterway transport are sufficient availability of infrastructure capacity although lock capacity increasingly poses problems. Furthermore, it is the most environmentally friendly transport mode, it is fuel-efficient, it is reliable, it is able to carry large volumes, it is not too expensive if distances are relatively long and door-to-door transport is possible, and its safety is regarded high. Weak points are high investments needed for new barges combined with a long life-time expectancy, lack of accessibility, tendency toward overcapacity, seaport congestion, reliability, low speed, lack of qualified employees, limited role in transport organization, natural constraints (ice, water level changes), and limited flexibility.

The Position of Inland Ports and Terminals in Inland Waterway IFT

Besides the mail haulage by rail or inland waterways, the inland ports and their terminals are important. The role, function, and operation of inland container ports in the total IFT solution is quite important and gets even more important as a closer integration between maritime and inland freight transport systems is occurring where deep-sea ports and also container carriers seek more control over the hinterland. A variety of different scales can be observed as some inland container ports are just simple container terminals while others are much more complex areas that include multiple container terminals, additional logistics zones and a governance structure, such as a port authority that governs the complete area for one or more municipalities. In these inland container ports, besides the core function of container handling other functions such as storage, consolidation and deconsolidation of freight, acting as depot, maintenance of containers, and customs clearance can also be made available. In the freight transport network, the position of the inland ports can vary from a “simple” extended gate of a major seaports that helps accommodate freight flows from large deep-sea ports further inland because in the deep-sea port there is insufficient capacity to large inland ports with own strategies and own networks that not solely depend on major deep-sea ports (such as Venlo and Duisburg).

Challenges for Intermodal Transport

Different challenges exist for IFT posing threats to the performance and operation of IFT. An important challenge is infrastructure capacity. In rail IFT, there is a lack of additional infrastructure capacity (lack of new train paths) and the priority of freight trains also influences the capacity for freight trains. This lack of capacity severely hampers the growth opportunities for IFT rail. The capacity issue also holds for IFT inland waterway but concentrates more on the locks and bridges that not always offer sufficient capacity. The capacity on the waterways (the links) usually is sufficient. A challenge linked to capacity is the availability of funds for the growth of capacity. Large investments are needed in capacity while the funds for these investments often lack (Wiegman and Behdani, 2018). A third challenge is the difficulty of finding qualified personnel which especially in IFT inland waterways is a serious issue. The last challenge is the increasing efforts of truck transport to reduce their emissions and work toward less polluting trucks. This means that also IFT rail and inland waterway have to increase their efforts in order to stay more environmental friendly compared to truck transport. Synchromodality might offer possibilities to work on these challenges and give IFT the next boost.

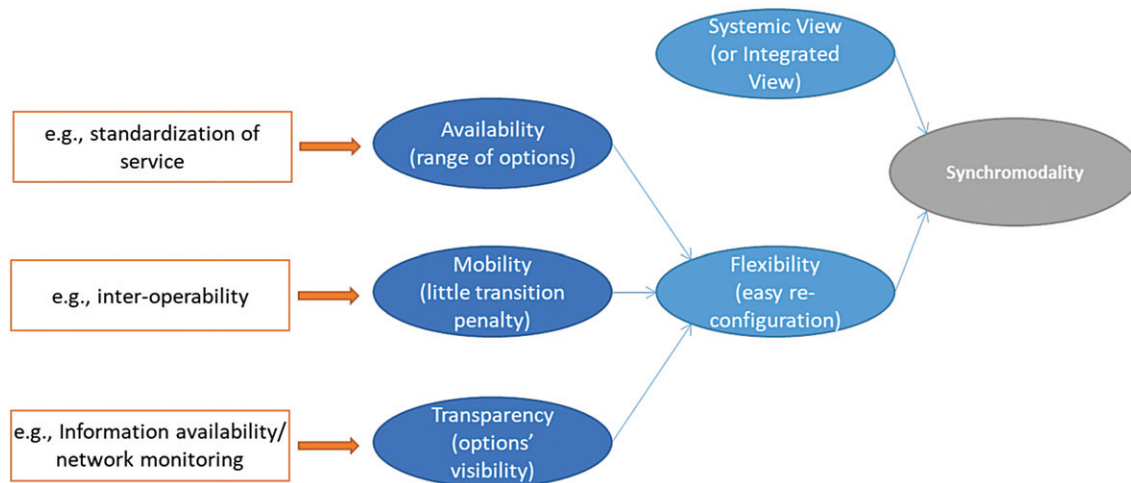


Figure 2 The main elements of synchromodality. *Source: Authors' own illustration.*

Synchromodality Defined

Establishing an agreed definition of synchromodality has not proved to be an easy task since there is no consensus on a precise definition of this concept. There is, however, a general agreement that synchromodality is about an integrated view in planning and using different transport modes to provide the flexibility in handling freight transport demand in the real time. Therefore synchromodality is mostly defined based on two main concepts (Fig. 2). The first concept is the integrated or systemic view on intermodal network and service design. This implies that—instead of looking at different intermodal services by different transport operators (either barge, rail or truck operators)—we must look into the options from an integrated view and based on the complementary nature of these different modalities (StadieSeifi et al., 2014; Van Riessen et al., 2013).

The second aspect of synchromodality is flexibility (or easy switching). In a SFT service, a contract is made between shippers and a synchromodal transport service provider in which shippers “only” book the transportation service with the agreed cost and quality requirements. Synchromodal transport service providers then decide on which transport modes to use according to the specifications of the customer. Depending on the actual situation and mode availability, the cargo can be distributed over several modalities, enabling the shipper to benefit from the advantages of low cost and sustainable modes and transport provider to benefit from higher asset utilization (Behdani et al., 2016). As a result, synchromodality offers greater flexibility in mode choice. It also improves the reliability, makes the lead time in the transport chains shorten, and increases the utilization of road, rail, and inland waterway.

The essential flexibility in a synchromodal transport system can be achieved only when three other conditions are met. In other words, *flexibility* itself is a multidimensional construct and is composed of three other elements. *Availability* is about the presence of options. The premise of a synchromodal system is an easy and real-time change the modality from one mode to another or from one path to another one based on the network situation and with the aim of optimizing the network utilization. To achieve that, there need to be multiple options available for transporting a container from point A to point B. Without existing those options—which can be primarily caused by lack of infrastructures like rail or inland waterways—a synchromodal system service cannot be defined. The second concept is *Mobility*, which is about the possibility of switching from one modality to the other one. In some case, multiple options for moving a container from point A to point B are available, but the transport planner cannot switch to the best transport modality because there is a high penalty or cost associated with that. A common example of this is the concept of A-modal booking (or Mode-free Booking) in which a shipper gives the right of choosing a modality or even a route for the container to the transport operator (Behdani et al., 2016). In other words, booking transport service is done without specifying the explicit mode or path for transporting the cargo. As a result, the transport operator can freely (i.e., without penalty) choose the appropriate route or mode based on monitoring the whole transport network. As discussed in the next section there might be additional barriers that can make it extremely difficult or even impossible to flexibly switch from one modality to another one.

The third aspect (or driver) of flexibility is *Transparency*. There can be situations in which multiple options or modalities are existing and the transport operator can switch among modalities; yet, because of no visibility or shared awareness about the whole transport network, the alternative best options cannot be found (De Borger and De Bruyne, 2011).

To conclude, the elements of the framework of Fig. 2 describe the necessary aspects or drivers of a SFT system. More integrated view in the design of intermodal network or services will provide more opportunity for synchromodality in IFT. Similarly, providing more flexibility—by focusing on each of the aforementioned three aspects—will facilitate to transform an IFT system into a synchromodal one. It must be emphasized that the concepts defining the synchromodality can be applied to design a synchromodal infrastructure network. In other words, an integrated view is applied in the development planning of transport/transshipment infrastructures (Fig. 3). This includes a long-term strategic view of synchromodal design and is called network synchromodality. A more frequently discussed business model of synchromodality is, however, synchromodal service design which is about designing

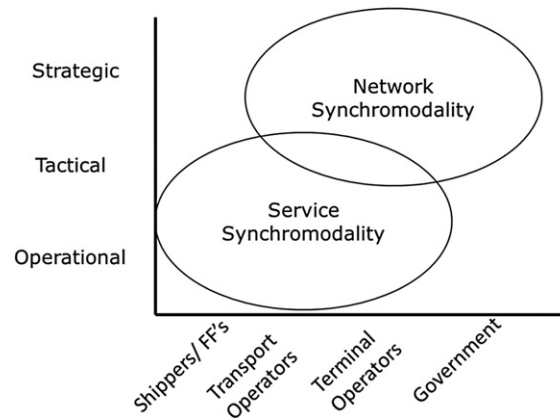


Figure 3 Network synchronomodality and service synchronomodality. Source: Authors' own illustration.

transport services (e.g., defining the timetables or contracting with service providers) from an integrated systemic view and providing the flexibility in the design of service.

Challenges and Drivers for Synchronodal Transport

While interest in synchronomodality is growing, there are several challenges in and barriers realizing a synchronodal transport system. These challenges are categorized into four groups as shown in Fig. 4.

Cultural or social barriers are about the view, values, and beliefs of individual actors in a transport chain. Moving toward SFT calls for a mind-shift in all actors in the chain. The primary mind-shift is for shippers by accepting a mode-fee booking business model (Lucassen and Dogger, 2012). Several studies show that—especially for cargos with the high value of time—the shippers are not ready for such a move (Behdani et al., 2019). The other cultural aspect is the mind-shift in transport operators. Emerging a SFT system calls for collaboration among operators working with different modes of transport (Franc and Van der Horst, 2010). These operators (in different modalities) are currently seen as competing service providers. For synchronomodality this competition mind-set needs to be replaced by one that emphasizes on cooperation among different transport operators (looking at the complementary nature of different modes of transport).

Technical barriers are about developing technical platforms for synchronodal planning. As discussed in the next section, there have been several projects in the last few years focusing on developing (IT) platforms. Yet, a trusted platform that supports actors to

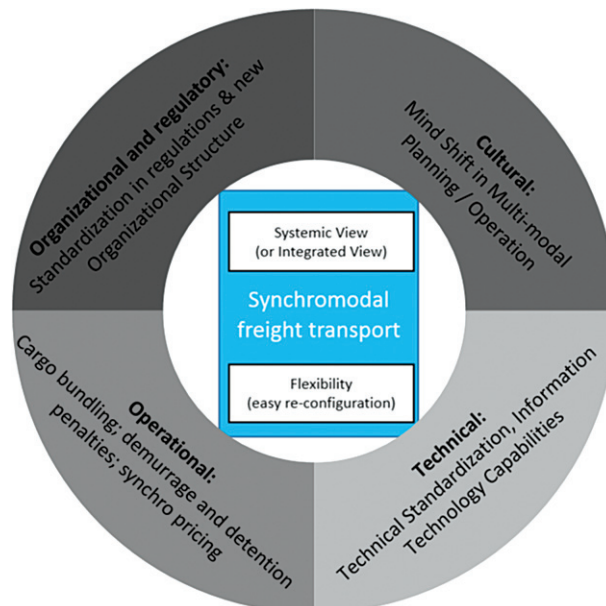


Figure 4 Categories of synchronomodality barriers/drivers. Source: Authors' own illustration.

5	IT Services for Synchronmodality (SynchronodalIT)	Dutch logistics chain	(1) To fulfill the need for an integrated multi-modal logistics network (2) To increase the efficiency, reliability and sustainability of logistics services (3) To promote a mental shift to bring about shippers and logistics service providers, so that their working methods are more in line with synchronodal ideas	(1) Real time and Big Data: Concerning logistic infrastructure data management (2) Planning and Gaming: Concerning optimal synchronodal planning techniques (3) Architecture and Services: With respect to a platform for smart pluggable services for Synchronodal Logistics Service Provisioning networks	A number of decision making scenarios will be developed, leading to innovative pilots and implementation of new logistic services. The outcomes are expected to have long-term effects for all participants on their actual decision making in freight transportation	^aThales Nederlands BV ^aCombi Terminal Twente (CTT) ^aPost-Kogeko Logistics ^aCAPE Group ^aNexusZ Hengelo ^aARCADIS ^aSimacan ^aOV Software ^aPort of Amsterdam ^aTata Steel Ijmuiden ^aCombi- Terminal Twente ^aAir Cargo Netherlands ^aTNO
6	Complexity Methods for Predictive Synchronmodality (COMET-PS)	-	To bridge the gap from the non-transparent and inefficient current transport state to a streamlined logistic system with improved transport efficiency, higher loading rate of vehicles, less emissions and costs	The concept of predictive synchronmodality is proposed. Three real-life use cases are investigated, which are predictive synchronmodality for container logistics, dry bulk logistics, and air cargo logistics	To develop models, methods and tools based on predictive data analysis and stochastic decision making in distributed control environments, for exploiting the great potential of synchronmodality, addressing the question what to transport, how and when	(1) Customer reactions on synchronodal pricing (2) The impact of risk factors on synchronodal transport prices (3) Cargo revenue management for synchronodal transportation (4) Synchro pilot evaluation (5) (Dynamic) pricing of synchronodal services (6) Analysis of planning and booking processes at EGS (7) Conceptual model for pricing of synchronodal services In addition to the development of the game and the gaming workshops, the project will also demonstrate the lessons learned in real-life. Daily operations on the network of the partners will be improved in a Living Lab environment based on the outcomes of gaming. This will give insight in development of new operational procedures and their fit in current way of working and also in an increased number of a-modal bookings from a specific target group.
7	Pricing of maritime and continental synchronodal services	-	To develop methods for a-modal pricing of synchronodal services in maritime and continental hinterland transport	(1) Assessing and reducing the risks of not using the cheapest transport solution (2) The customer acceptance of a-modal pricing (3) The possibilities for dynamic pricing (including yield management) of synchronodal services	(1) Customer reactions on synchronodal pricing (2) The impact of risk factors on synchronodal transport prices (3) Cargo revenue management for synchronodal transportation (4) Synchro pilot evaluation (5) (Dynamic) pricing of synchronodal services (6) Analysis of planning and booking processes at EGS (7) Conceptual model for pricing of synchronodal services In addition to the development of the game and the gaming workshops, the project will also demonstrate the lessons learned in real-life. Daily operations on the network of the partners will be improved in a Living Lab environment based on the outcomes of gaming. This will give insight in development of new operational procedures and their fit in current way of working and also in an increased number of a-modal bookings from a specific target group.	^aKLG Europe ^aDen Hartogh Liquid Logistics ^aEGS Fontys
8	Serious Synchronodal Gaming	The hinterland network between Rotterdam, Venlo, Nijmegen and Duisburg	To develop a synchronodal game and gaming workshops which are suitable for experiencing synchronodal transport planning by operational staff	Development of the game and the gaming workshops	(1) Customer reactions on synchronodal pricing (2) The impact of risk factors on synchronodal transport prices (3) Cargo revenue management for synchronodal transportation (4) Synchro pilot evaluation (5) (Dynamic) pricing of synchronodal services (6) Analysis of planning and booking processes at EGS (7) Conceptual model for pricing of synchronodal services In addition to the development of the game and the gaming workshops, the project will also demonstrate the lessons learned in real-life. Daily operations on the network of the partners will be improved in a Living Lab environment based on the outcomes of gaming. This will give insight in development of new operational procedures and their fit in current way of working and also in an increased number of a-modal bookings from a specific target group.	^aECT ^aDanser ^aTCT Venlo ^aDeCeTe ^aTNO

(Continued)

Table 2 An overview of synchromodality projects including inland waterway (cont.)

No.	Project	Scope	Goal	Focus	(Expected) Results	Project partners
9	Ketendigitalisering 2.0	-	To strengthen the management of information flows of seaport forwarders. To improve the control of the safety of information flows by making good agreements about the accessibility of information flows.	(1) Integral improvement of one-to-one data communication connections (2) Digital exchange of sea freight rates (3) Improving payment process forwarder - shipping company (4) Improving information exchange for demurrage and detention (5) Developing a planning tool for modality choice	(1) To develop digital applications for improvement of information flow that can be introduced industry-wide (2) To develop the technical standards for applications (3) To identify management options (4) To draw up an action plan for the industry-wide rollout of an application TEUbooker offers a synchromodal solution: it makes use of the unused capacity of all modalities, so that the exchange options are maximized and transport costs can be reduced.	^a FENEX
10	TEUbooker	Port of Rotterdam, Maasvlakte II	To facilitate booking container shipments	-		^a Danser ^a DistriRail ^a Pro-Log ^a Port Shuttle

^aDifferent project partners.

securely share their information (e.g., the existing capacity in their network in real time) needs further attention. The other technical aspect is physical similarity of different modalities; for example, for shifting from a rail service to a barge service for a reefer container, the barge should have the power plugs in order to cool down the reefer containers on board (Behdani et al., 2019). Without such a technical arrangement, there is no possibility to change the modality in a synchromodal service.

The organizational and regulatory barriers are the third category of barriers for SFT. Conflicting or inconsistent requirements for different modalities—for example in terms of insurance or customs documentation/process—and lack of standard regulations (or similar insurance rates) for different modalities can be potential legal barriers for synchromodality (Veenstra et al., 2012). Additionally, involving multiple stakeholders—both private and public organizations—in the design and operation of different modalities may obstruct the full potential of synchromodality. This calls for new types of contract settings to share the gains/pains fairly among all actors in a synchromodal network (Veenstra et al., 2012).

Operational barriers are embedded in the common business practices in the IFT. For example, active regional bundling of container flows may provide more opportunities for synchromodality in a specific hinterland region (Lucassen and Dogger, 2012). Yet, some operational constraints like detention and demurrage costs or different cut-off times for modalities may hinder this bundling as well as the possibility to switch from trucking to rail or barge. Additionally, similarity between the administration processes involved in managing different modes of transport can facilitate the transition toward a synchromodal freight transport system.

It should be noted that these four categories of barriers are interrelated. For instance, a business environment in which shippers are hesitant toward synchromodality will not provide enough incentives for governmental bodies to develop new regulations or adapt the existing regulation and processes. This means that cultural barriers can induce legal barriers, which induce further cultural barriers. The interrelatedness of the four categories of synchromodality barriers can result in a chain reaction toward failure, with the IFT system then remaining in its current business-as-usual.

Cases on Synchromodal Transport in Inland Waterway Sector

Table 2 gives an overview of the projects and practical cases on synchromodality. The projects in this table especially include the cases in which IWW was included as a part of synchro-modal service. As one of the first studies on this topic, TNO performed a pilot study on SFT between ECT Euromax and Delta terminals in Rotterdam and two inland terminals (i.e., Moerdijk and Tilburg) using a trimodal service with barge, rail, and road connections (Lucassen and Dogger, 2012). Based on this pilot study, Lucassen and Dogger (2012) presented a framework for synchromodality in five areas: (1) role of shippers (the necessity of mode-free booking as a prerequisite for synchromodality); (2) dynamics (real-time planning and switching among modalities); (3) cooperation and control (the need for cooperation within transport chains and between transport chains); (4) network and organization design (the need for network view instead of point-to-point connections and new business models to stimulate cooperation and network operation); (5) infrastructure use (more efficient use of existing infrastructure and removing bottlenecks—like opening hours of locks).

The other practical case of synchromodality is the European Gateway Services (EGS) by ECT in which a synchromodal transport system delivers containers between the ports of Rotterdam and Antwerp and several inland terminals in the Netherlands, Germany, and Belgium by means of more than 30 weekly inland barge services and 40 weekly rail services (European Gateway Services, 2019). Additionally, the handling of customs formalities is done in some inland terminals (i.e., an Extended Gate Service). EGS is well known as the best practice of defining and implementing synchromodal solutions.

Table 2 also gives an overview and the details of several other projects (and research studies) on synchromodality involving inland waterways. The majority of these projects are focused on the technical barriers for synchromodality. The legal and organizational aspects are still less thoroughly studied domains and call for further attention.

Summary and Conclusion

This chapter discusses the concept of SFT and how it helps in meeting the challenges in using inland waterways as an alternative modality. A framework for defining the SFT is presented. The framework describes two concepts of “Systemic view” and “Flexibility” as the main pillars of synchromodality. Flexibility itself is a multidimensional construct and is supported by “Availability,” “Mobility,” and “Transparency.” The challenges in developing a SFT system and an overview of the projects and practical cases on synchromodality in inland waterway sector are presented. Although the technical barriers for synchromodality are generally addressed, the legal and organizational aspects are still less thoroughly explored and should be centralized in future studies.

References

- Behdani, B., Fan, Y., Bloemhof, J.M., 2019. Cool chain and temperature-controlled transport: an overview of concepts, challenges, and technologies. In: *Sustain. Food Suppl Chain.*, Academic Press, pp. 167–183.
- Behdani, B., Fan, Y., Wiegman, B., Zuidwijk, R., 2016. Multimodal schedule design for synchromodal freight transport systems. *Eur. J. Trans. Infrast. Res.* 16 (3), 424–444.

- De Borger, B., De Bruyne, D., 2011. Port activities, hinterland congestion, and optimal government policies: the role of vertical integration in logistic operations. *J. Trans. Econ. Policy* 45 (2), 247–275.
- European Gateway Services, 2019. Available from: www.europeangatewayservices.com.
- Ferreira, L., 1997. Planning Australian freight rail operations: an overview. *Transport. Res. A Policy Pract.* 31 (4), 335–348.
- FitzRoy, F., Smith, I., 1995. The demand for rail transport in European countries. *Trans. Policy* 2 (3), 153–158.
- Franc, P., Van der Horst, M., 2010. Understanding hinterland service integration by shipping lines and terminal operators: a theoretical and empirical analysis. *J. Trans. Geograp.* 18 (4), 557–566.
- Hayuth, Y., 1987. *Intermodality: concept and practice*. Lloyd's of London Press, London.
- Jones, B.W., Cassady, R.C., Bowden R.O., 2000. Developing a standard definition of intermodal transportation, symposium on intermodal transportation, University of Denver.
- Lucassen, I.M.P.J., Dogger, T., 2012. *Synchromodality Pilot Study. Identification of Bottlenecks and Possibilities for a Network between Rotterdam. TNO, Moerdijk and Tilburg*.
- Macharis, C., Ambra, T., Mommens, K., 2020. Intermodal and synchromodal freight transport
- Morlok, E.K., Spasovic, L.N., 1994. Redesigning rail-truck intermodal drayage operations for enhanced service and cost performance. *J. Transport. Res. Forum* 34 (1), 16–31.
- StadieSeifi, M., Dellaert, N.P., Nuijten, W., Van Woensel, T., Raoufi, R., 2014. Multimodal freight transportation planning: a literature review. *Eur. J. Operat. Res.* 233 (1), 1–15.
- Templar, S., Rodemann, H., 2020. Eurasia rail freight: enablers and inhibitors of future growth
- Van Riessen, B., Negenborn, R.R., Dekker, R., Lodewijks, G., 2013. Service network design for an intermodal container network with flexible due dates/times and the possibility of using subcontracted transport. In: *Proceedings of the International Forum on Shipping, Ports and Airports, Hong Kong*.
- Veenstra, A., Zuidwijk, R., van Asperen, E., 2012. The extended gate concept for container terminals: expanding the notion of dry ports. *Marit. Econ. Logist.* 14 (1), 14–32.
- Wiegman, B., Konings, J.W., 2015. Intermodal inland waterway transport: modelling conditions influencing its cost competitiveness. *Asian J. Ship. Logist.* 31 (2), 273–294.
- Wiegman, B., Behdani, B., 2018. A review and analysis of the investment in, and cost structure of, intermodal rail terminals. *Trans. Rev.* 38 (1), 33–51.
- Woodburn, A., 2020. Rail freight

Maritime Company Passenger Management/Liner Industry

Claudio Ferrari*, Alessio Tei†, *Department of Economics and Business, University of Genoa, Genoa, Liguria, Italy; †School of Engineering, Newcastle University, Newcastle upon Tyne, Tyne and Wear, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction: The Transport of Passengers	587
The Fleet	587
The Service Organization	589
The Market Organization	590
Summary and Conclusions	592
References	592

Introduction: The Transport of Passengers

The maritime transport of passengers includes two distinct markets—cruise and ferry shipping. These two markets differ for two main characteristics—the typology of vessels used and the nature of the demand served. In fact, cruise shipping uses vessels that can accommodate only passengers (and the crew) while maritime transport service serves a final demand (i.e., the cruise itself is connected to a touristic need rather than a need for transportation). The ferry industry serves a derived demand (i.e., passengers are interested in reaching the arrival destination for their own purposes) and it often uses vessels that can accommodate both passengers and vehicles (Stopford, 2009). These two separate purposes affect the characteristics of the ships as well as their size. International statistics show a global average size for cruise ships of about 52,000 DWT (deadweight tonnage) against an average size of the ferries of about 1500 DWT. Similarly, services on board are substantially different (e.g., hoteling and entertainment services for cruises against essential welcoming services for ferry passengers), making the two markets completely separated not only from a demand point of view but also from the ship supply one. Eventually, the two markets largely differ in terms of international exposure: on the one hand, with only few exceptions, cruise services are organized to connect international destinations; on the other hand, ferry services are mainly provided to connect regional or national destinations (Stopford, 2009). Nevertheless, this latter element does not mean that international connections are excluded from the service market, given that certain world regions are currently experiencing a growing trend in international services as well (e.g., Mediterranean, South East Asia). Because of these general characteristics, ferries often compete against either air transport—especially in case of connections with islands—or rail and road services—in case of coastline connections—but they hardly compete with other maritime passenger services. A notable exemption is the organization of the mini-cruise services that often are substitute of ferry services, promoted with ships that can—at least partially—offer a short cruise experience.

The ferry sector is characterized by specific elements, mainly in terms of ship used, market organization and regional coverage, regulation, and operator strategies. These elements characterize the maritime passenger transport sector in several ways.

This paper is structured as follows: after this brief introduction, Section The Fleet assesses the fleet composition, highlighting the heterogeneity of the ships used for providing ferry services. Section The Service Organization describes current service organization with the aim of showing main areas in which ferry services are developed. Section The Market Organization focuses on the market organization, with the statistics connected to the main shipping companies as well as the different business challenges. Lastly, Section Summary and Conclusions contains a brief recap of main characteristics of the market as well as some conclusive remarks.

The Fleet

Under the kind of vessels deployed perspective, ferries are heavily used in the so-called “Short Sea Shipping” market (Paixão and Marlow, 2002; Wang and Lo, 2008) in which passengers are often transported together either with their own vehicles or with cargo (e.g., trucks, trailers, etc.). Because of this, often ships used in this market could be classified as multipurpose vessels, that is, ships able to provide both basic services for passengers (and sometime even essential hoteling services for specific passenger routes) as well as some cargo services. Thus, ferry ships are often classified together with the Ro-Pax vessels (i.e., ships able to transport passengers as well as trucks and containers), Pax/Car carriers (i.e., ships able to transport both passengers and vehicles), and with some Ro-Ro vessels (i.e., ships that mainly carry cargo through the loading of vehicles and trucks but that from time to time could be used for passenger services as well).

From a port point of view, traditional ferries as well as Ro-Pax are characterized by a low need of infrastructures in order to be productive: in fact, they need only a wharf to drop a ramp for the embarkation and disembarkation of passengers and vehicles. These conditions make ferries a mode of transport very popular, in some cases the primary transport mode, in developing coastal countries as well as in connecting remote regions. Moreover, the absence for a need for technological handling or welcoming facilities, has

promoted the use of such kind of transport service as a basic way to promote marine transport among different regional areas (Baird, 2007; Medda and Trujillo, 2010). This is reflected in the diversified size and characteristics of the vessels (Lai and Lo, 2004)—in order to serve popular markets, ferries, and Ro-Pax can have a capacity of more than 1500 people, despite this the average vessel size is of about 200 passengers (with some small ferries with a capacity of just 20 passengers), showing a great variety of vessel typologies that need to co-exist in order to serve the different market segments (i.e., touristic destinations, remote areas, islands).

The flexibility of this maritime service has been often highlighted by international agencies (e.g., UNCTAD) and local national governments (e.g., Indonesia, Philippines, Greece) as the main way to interconnect island communities. Moreover, several countries promoted mixed passenger/cargo transport services through ferries in an attempt to provide alternatives to congested coastal highways, as happens in certain European and South American countries (Albanese et al., 2013).

This segment of the maritime transport is extremely diversified as testified by the characteristics of the fleet both in terms of the capacity of the vessels deployed—from very small vessels to very big ones and in terms of the kind of vessels (i.e., Ro-Pax, Ferries, Car/Pax carriers, Ro-Ro). Currently, about 4000 ferries are in service, dedicated to exclusive passenger transport only. Moreover, there are about 2300 ships specialized in Car/Pax transport, about 1000 ships classified as Ro-Pax, and 600 ships that regularly belong to the Ro-Ro market and that potentially could be used into the ferry market as well. Therefore, the overall “ferry” segment could count on about 8000 of ships, representing about 10%–15% of the overall commercial ships [these potential overlaps in the utilization of the ships in multiple market segments often generate issues in having precise market statistics on specific businesses, as highlighted by Albanese and Cusano (2012)].

Given that the service provided can sometime be performed with really small ships connecting several couples of ports, the market is currently not concentrated, registering more than 3000 companies worldwide that own and operate either ferries, Ro-Pax, or other similar passenger ships. Among these companies, almost 70% of them owns less than 5 ships, with operations that span only at regional and national level and no presence in international markets. Thus, in terms of economic characteristics, the disposability of a ferry fleet doesn't represent a significant entry barrier to the market since the ships may be easily moved from a market to another. Moreover, there are scope economies in the industry but they are not so relevant, as demonstrated by the existence of several small companies, while the existence of a second-hand ferry market contributes to reduce the time lag between the decision to enter a particular market and the moment the fleet is ready to operate.

Overall ship building trends are presented in Fig. 1, highlighting another peculiarity of the ferry industry in respect to other shipping markets: the capacity of the ferries did not improve in the last years—in both terms of average passengers carrying capacity and DWT—as it happened for all other main ship types (e.g., cruise and container ships are still registering an increase in size)—this peculiarity is mainly linked to the maturity of the market as well as for the stable demand of transportation that reduces the need (and the benefits) of further pursuing economies of scale. On the one hand, the characteristics of many markets (i.e., small island ports) played also a big role on this limited increase in size, given the infrastructural limitations and the low level of demand. On the other hand, the growing need of providing differentiated services (e.g., mix cargo/passenger services) or multi-call solutions, affected the introduction of bigger ship in terms of DWT for accommodating trucks or for the provision of extra-services without affecting the passenger carrying capacity.

Fig. 1 shows the overall trend in terms of average ship capacity of newbuilt ships in terms of both passengers and DWT and also the number of ships entered into the service in the referenced year—one notices that many fluctuations affect the market, a peculiarity of all shipping sectors.

Fig. 2 reports the age of the ferries currently in service, disaggregating the ships per regions in connection to the nationality of the vessels. Most of the vessels are deployed in Europe (29% of the total) followed by Asia. In terms of age of the fleet, the vessels deployed in the market are relatively old (more than a third is older than 30 years) with half of the old ships registered by EU companies. Similarly, all the companies belonging to the main regional markets (i.e., Europe, Far East, South-East Asia) have half of their ships built more than 30 years ago. “Young” ferries are relatively uncommon, with about 15% of the ships having less than 10 years.

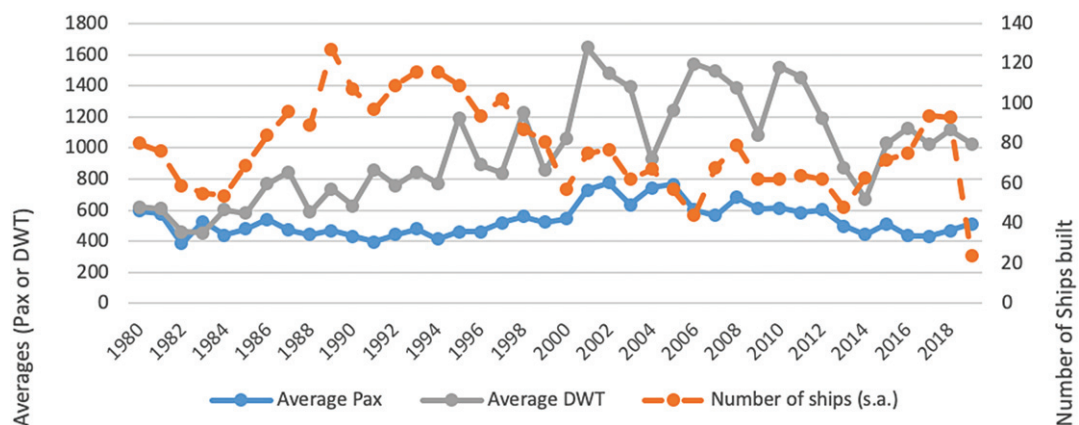


Figure 1 Trends in liner passenger fleet. Source: Authors' elaboration from Clarkson-SIN (2019).

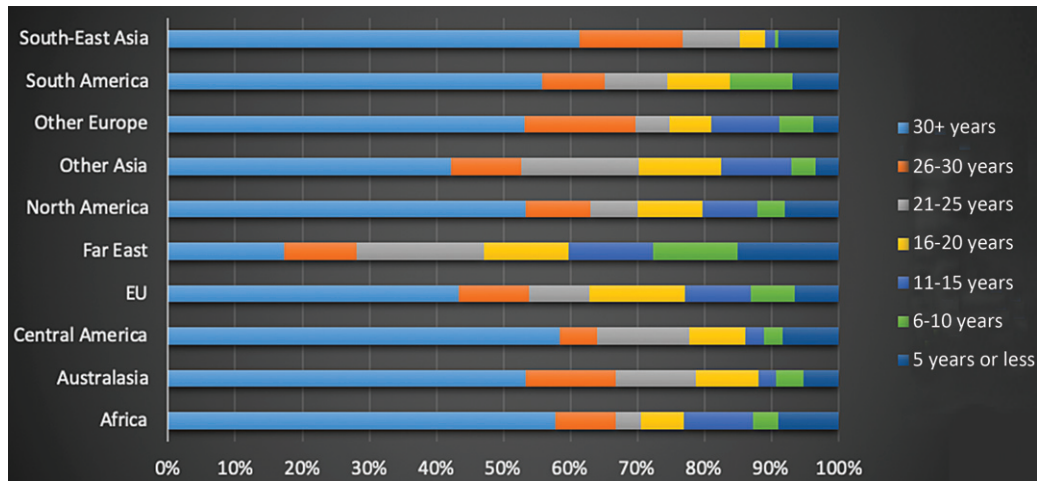


Figure 2 Distribution of ship age per registered flag of the ferry. Source: Authors' elaboration from Clarkson-SIN (2019).

The Service Organization

Overall, ferry transport services are organized as a liner service connecting two or few ports behaving as floating bridges. The ferry industry is particularly developed in geographical contexts with long coastlines, islands, and archipelagos as in Europe (e.g., in the Mediterranean Sea and in the Northern Sea), in Southeast Asia, in Japan, and in the African Great Lakes region, just to mention a few. Fig. 3 shows the position of the vessels classified as "Ferry" worldwide. As it is possible to see, most of the ships are deployed either in Europe—in the Baltic and North Sea regions and in the Mediterranean Sea as well—or in Asia—North-East Asia and South-East Asia. These two regions account for most of the ship registered worldwide as well as of their movements (in the map they are represented by the *white-dotted circles*). Other important areas for regional ferry traffic are North America and the Middle East (*light grey-dotted circles* in Fig. 3).

Given the distribution of the services, worldwide fleet appears quite concentrated in Europe and Asia, as shown in Fig. 4.



Figure 3 Density of ferries worldwide. Source: Authors' elaboration from Clarkson-SIN (2019).

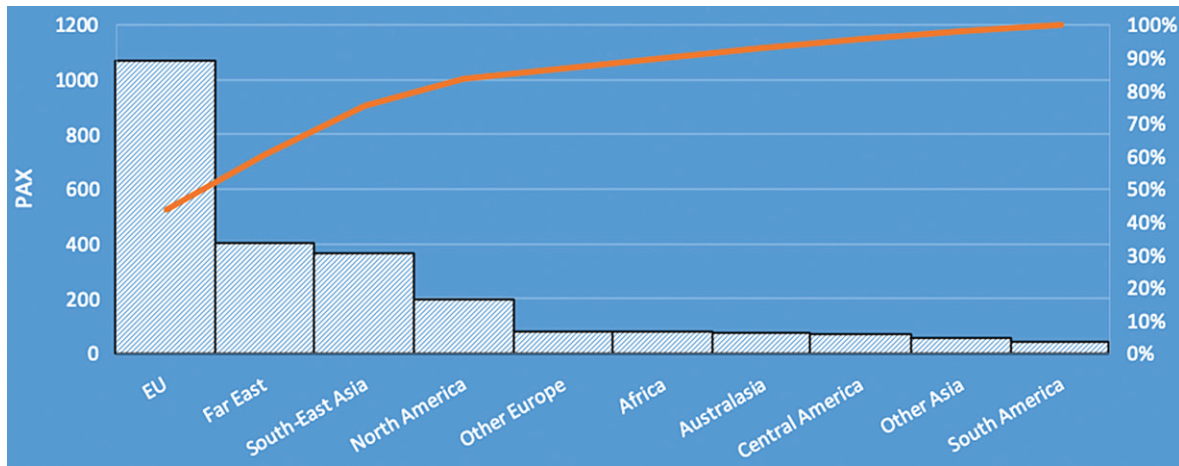


Figure 4 Worldwide fleet distribution. Source: Authors' elaboration from Clarkson-SIN (2019).

Conventionally, ferry direct connections last for no more than 48 hours (longer journeys are considered as cruise services). The importance of this kind of transport service is given by the volume of passengers served; according to the Interferry's website—the association that groups over 200 ferry operators worldwide—every year almost 2.1 billion of passengers are transported by ferries.

The Market Organization

Considering the top 100 companies per number of ships owned, Fig. 5 shows the vessel distribution in terms of number of ships as well as in terms of passenger capacity. Considering the overall market, the sector appears quite disperse, with low-market concentration (the HHI index is lower than 200, considering all main competitive variables, that is, number of ships, DWT, pax capacity) that highlights the dynamics of the sector as well as the possibility for many operators to be competitive into the market. In terms of

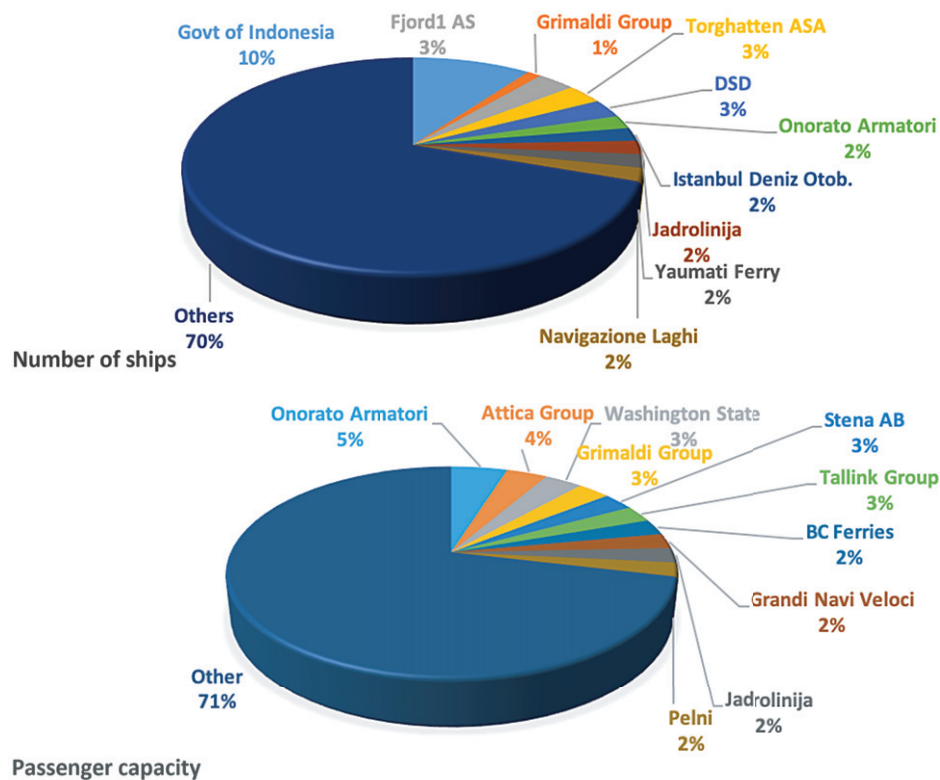


Figure 5 Fleet distribution in terms of ship number and passenger capacity. Source: Authors' elaboration from Clarkson-SIN (2019).

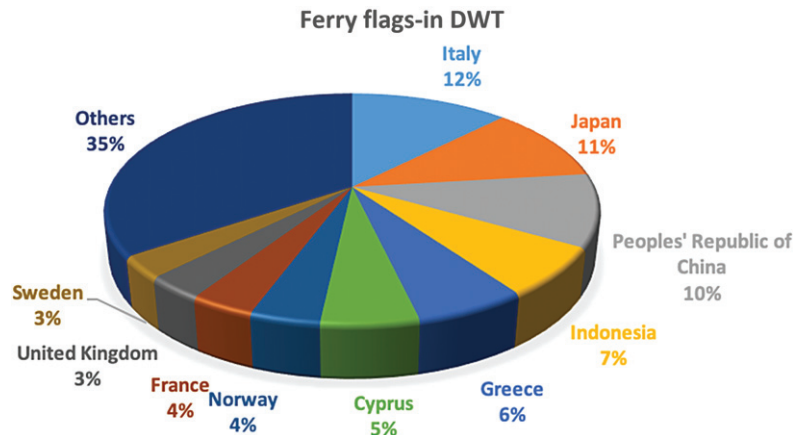


Figure 6 Ships' flags in DWT. Source: Authors' elaboration from Clarkson-SIN (2019).

competition in providing alternative services, most of the point-to-point routes are traditionally served by only few operators (often just by one) and this actually reduces the level of actual competition on the market, despite the dispersions of the sector. The presence of few operators on many routes is often linked to either regulatory regimes that reduce the possibility for several operators to be present into certain markets (e.g., limits to assure service quality) or to the level of demand that reduces the attractiveness of providing perfectly competitive services (You and Park, 2017; Kizielewicz et al., 2017).

Fig. 3 highlights how, in terms of ship numbers, Asian operators are market leaders, with ships that mainly operate in South East Asian markets. Despite this, most of the vessels in operation in South East Asia are relatively small. In fact, the ranking in terms of ship capacity is quite different: most of the top 10 companies operate in Europe, especially in the Mediterranean Sea.

In terms of ships' flag, Fig. 6 shows the main flags for the ferry market, in relation to the size of the ship deployed. As it is possible to see, main flags are related to either European Union Countries (more than a third of the market) or main insular Asian countries (i.e., Indonesia and Japan).

In terms of services offered, the passengers related markets are characterized by a unique capability to diversify the offer of transport, not only through the supply of mixed services (i.e., Ro-Pax, Pax, Car/Pax) but also through the use of different speed, ships, and complimentary services. Given that the main purpose for using a ferry is the arrival to a certain destination, as per other modes of transport, the possibility to offer different trip durations generate diversified markets, in terms of both price and frequency. Often destinations are offered in competition among few different operators—or even by the same operator—but the service is perceived as different due to the different time spent on the ship, allowing for the creation of competing markets. This element constitutes a quite peculiar characteristic within the maritime transport sector, given that traditional freight services are not often diversified in terms of speed. The speed diversification is often linked to the possibility of finding different kind of ferries deployed on the same route, as in the case of the catamarans (that allow operators to travel faster than through traditional ships).

Moreover, the economic and social role of the ferry in satisfying transport needs has often been linked to specific national interests, such as the connection of remote regions that pushed national authorities to protect local industries as well as to influence the planning of the service schedule and characteristics. Because of this, many countries apply strict legislations to enter into their local markets, i.e. a reserve based on the national flag or the nationality of the shipowner. These restrictions, that represent a peculiarity in the international nature of the shipping industry, are often linked to the need to establish strict service standards (e.g., service frequency, fixed price) in order to allow the use of ferries as primary mobility solution for the remote communities or the island inhabitants.

Another important characteristic that differentiates the ferry market from other shipping sectors is that—given that most of the services are offered regionally—most of the ships are designed to be optimized on the basis of specific local needs. This affects both the size and the overall design of the ships. Thus, specific characteristics of the ship may change from market to market. This peculiar element helps the operators to offer different services that do not only allow passengers to reach the designated destination but also to sell extra activities on-board (e.g. night trips with hotel rooms, sightseeing). The possibility of offering alternative services allows the operators to provide loop services with multiple calls (e.g., Fjord service in Norway) or mini-cruise experiences (e.g., connecting only touristic destinations and offering extra-services onboard) used by passengers not only for moving from a place to another but also as an essential part of their own holiday (Albanese et al., 2013; Mathisen and Jørgensen, 2012).

While the point-to-point service is still the main organizational scheme for this shipping market (e.g., direct connections to an island from a mainland port), the multi-calls services are often linked to specific regional needs or—as mentioned above—they might act as a potential selling strategy for operators. This peculiarity is particularly relevant for serving specific touristic markets (e.g., mainly in Europe and South East Asia) or whenever the demand for certain locations would not allow an economic viable solution for providing the direct service all the year round. In fact, in case of strong seasonality or unbalanced transport flows, multi-calls services are often used as a solution to constitute an economic viable market. In case of un-balanced flows, an alternative to the multi-calls service is to provide mixed services, such as in the case of Ro-Pax ships and other not dedicated ships.

Among the industry trends, two are especially worth mentioning: the introduction of innovations and the improvement of safety, namely in some regions. The drivers of innovation in the ferry market are the increasing awareness of the environmental footprint of the sector together with the need to be more efficient as possible. In this sector innovations mainly refer to the design and building techniques together with the materials used and the propulsion system. Also international legislation (MARPOL—Annex VI) pushed the sector to be more green and sustainable lowering the limits of nitrogen oxides (NOx) and sulfur oxides (SOx) due to the combustion of fuel initially in four areas—the so-called Emissions Control Areas (ECAs)—and in the future in other regions. It is noteworthy that these emission control areas particularly involve coastal regions and then the ferry sector is greatly impacted. This contributes to explain why the ferry sector may be considered the early adopter of relevant innovations as the adoption of LNG (liquefied natural gas) as ships' propeller, but some companies are testing also the use of methanol in their fuel mix, as well as dual fuel systems and the use of batteries for port maneuvers.

The second trend refers to the enforcement of a stricter regulation of the sector, mainly in the developing countries, in order to reduce, and hopefully avoid incidents and casualties (You and Park, 2017). In particular, the efforts go toward an improvement of the skills of seafarers together with an improvement of national legislations and controlling procedures in order to raise the safety and security standards of the world ferry industry.

This introduces to another feature of the industry that is the role (and type) of regulation in the ferry sector. As above mentioned since in most cases ferry services are the solely transport means to connect islands with mainland some regulation is needed if the production of the transport service is not sufficiently profitable, moreover the cabotage services traditionally were restricted to the national flag. Things are changing at least in some regions, as in the European Union, toward the opening of the market to competition of foreign ferry companies and the flagging out of the fleet in order to comply with a less strict rules (with reference to the nationality of seafarers, for instance). Moreover, the sector needs some forms of regulation because in most cases the degree of competition is often relatively low. In fact, a ferry service competes with another ferry service operating on the same shipping route, or a similar (i.e., connecting couples of ports close to each other) shipping route, and competes with road and air transport where this competition is feasible. This means that in most cases the number of maritime operators competing each other is extremely low, thus the need for relevant Authorities controlling for the abuse of dominant positions in the market.

Summary and Conclusions

Marine transport reports and studies are often focusing on either the cargo sector or the cruise market—nevertheless, the passenger transport sector represents an important development factor for many world regions. Often used to limit the environmental impact of the road transport as well as to provide essential links to remote regions. The heterogeneity of the locations served makes the fleet composition quite peculiar: with ships that fails to achieve big sizes and relatively old vessels. The heterogeneity is not only guaranteed by the geographical location of different served destinations, but it also depends on the purpose of the travel (i.e., commuting or tourism) as well as on the possibility to offer mixed services (e.g., as in the case of the Ro-Pax).

Given its strategic role, the ferry market is often highly regulated and characterized by national champions, with a concentration of interests in key countries. This concentration in terms of flag nationality does not prevent the market to avoid concentration in terms of ship capacity, even if few alternatives are provided in relation to specific routes (i.e., the market concentration is only possible on specific point-to-point services).

Eventually, as anticipated above, the key ferry market characteristics can be summarised by the presence of several local markets that affect the ships used (in terms of number, size and technology) as well as the number of operators competing among each other. The presence of strict regulations and the “social” role of the ferries, make this market peculiar within the shipping sector with a relevant regional role that is always linked to assure a “mobility” need of passengers rather than trade (e.g., cargo)/other economic activity (e.g., leisure), as it happens for most of other shipping sectors.

References

- Albanese M., Cusano M.I., 2012. The Italian Roll-On Roll-Off Sector: From the Puzzle of Statistics to a Realistic Scenario, Proceedings of the 2012 International research conference on short sea shipping, 1-2 April 2012, Lisbon: Portugal.
- Albanese, M., Cusano, M.I., Ferrari, C., Tei, A., 2013. West Med Short Sea Shipping: From Point to Point to Multi-Call Services. In: Forte, E. (Ed.), *Economics and Logistics in Short and Deep Sea Market*. Franco Angeli, Milan, Italy, pp. 25–44.
- Baird, A.J., 2007. The economics of motorways of the sea. *Marit. Policy Manag.* 34 (4), 287–310.
- Clarkson-SIN (2019). Clarkson—Shipping Intelligence Network. Available from: <https://sin.clarksons.net>.
- Kizielewicz, J., Haahti, A., Luković, T., Gračan, D., 2017. The segmentation of the demand for ferry travel – a case study of Stena line. *Econ. Res.* 30 (1), 1003–1020.
- Lai, M.F., Lo, H.K., 2004. Ferry service network design: optimal fleet size, routing, and scheduling. *Transp. Res. Part A* 38 (4), 305–328.
- Mathisen, T.A., Jørgensen, F., 2012. Panel data analysis of operating costs in the Norwegian car ferry industry. *Marit. Econ. Logist.* 14 (2), 249–263.
- Medda, F., Trujillo, L., 2010. Short-sea shipping: an analysis of its determinants. *Marit. Policy Manag.* 37 (3), 285–303.
- Paixão, A.C., Marlow, P.B., 2002. Strengths and weaknesses of short sea shipping. *Mar. Policy* 26 (3), 167–178.
- Stopford, M., 2009. *Maritime Economics*. Routledge, London, UK.
- Wang, D., Lo, H.K., 2008. Multi-fleet ferry service network design with passenger preferences for differential services. *Transp. Res. Part B* 42 (9), 798–822.
- You, J., Park, Y.M., 2017. Capturing Collusion: The Industry and the Government in Ferry Safety Regulation. In: Suh, J.J., Kim, M. (Eds.), *Challenges of Modernization and Governance in South Korea*, 95–120. Palgrave Macmillan, Singapore.

Cruise Industry

Athanasios A. Pallis, Aimilia A. Papachristou, Department of Shipping, Trade and Transport, University of the Aegean, Chios, Greece

© 2021 Elsevier Ltd. All rights reserved.

Introduction	593
What is Cruising?	593
Uninterrupted Growth and Globalization	594
The Geography of Cruise Shipping	594
Market Concentration	595
Cruise Fleet: Economies of Scale	595
Market Segmentation	595
Source Markets	595
Cruise Ports and Destinations	596
The Sustainability Challenge	596
Reversing Social Perceptions	597
Limitation of Environmental Externalities	598
Conclusions	598
See Also	599
Relevant Websites	599
References	599
Further Reading	599

Introduction

Cruise shipping has witnessed an uninterrupted growth and unstoppable globalization trends over the last decades. The globalization of cruise activities is highly supported by profitable cruise lines who have endorsed economies of scale and advanced segmentation, in order to offer innovation and schedule new types of itineraries, and not least, by the desires and strategies of cruise ports to host more cruise activities. It is also advanced by changes in the demand side, in particular the growing internationalization of cruise passengers' source markets. With the positive direct and indirect impacts diffused to the port cities or nearby touristic destinations, cruise seaports are gaining importance. The interest in hosting more cruise calls and cruise passenger movements has been supported, in general, by broader communities and decision makers. Still growing cruise business, like any other economic activity, is also associated with externalities raising social, economic, and environmental questions and challenges for cruise port and the surrounding areas.

The reader is being introduced to the chapter with a short definition of the term "cruising," followed by an analysis of the evolution of a globalized industry. Then the chapter reviews the key challenges and strategic issues of the cruise industry today: the geography of cruise shipping, market concentration, the presence of economies of scale, market segmentation, passengers source markets, and the cruise ports and the destinations that have developed an interest in advancing their cruise activities. Finally, it highlights the importance of sustainability, in particular the socially and environmentally responsible growth, as a strategic issue for the cruise world.

What is Cruising?

Cruise is a mixture of maritime transport, travel and tourism services, facilitating the leisure activity of passengers paying for an itinerary and, potentially, other services on board, and including at least one night on board on a seagoing vessel having a capacity of at least 100 passengers. Transportation (the cruise ship) is the core element of the experience instead of a simple conveyance.

Cruise shipping has first established as the transportation of pleasure-seeking upper class travelers on seagoing vessels offering one or more ports of call in the United States and the Caribbean. Today this is a highly efficient global business. Modern specialized ships which are radically different from cargo vessels, the use of an increasing number of cruise ports of call and turnaround ports, so as to provide their customers excellent in-port and destination experiences, and convenient departures from proximal embarkation cities, are fundamental tenets of the modern cruise industry.

This is an activity-taking place in specific markets, each having its own regional characteristics, with the Caribbean and the Mediterranean being the most important. As such, the cruise industry must address multiple considerations related to on-board amenities, itineraries, ports of call, and shore excursions.

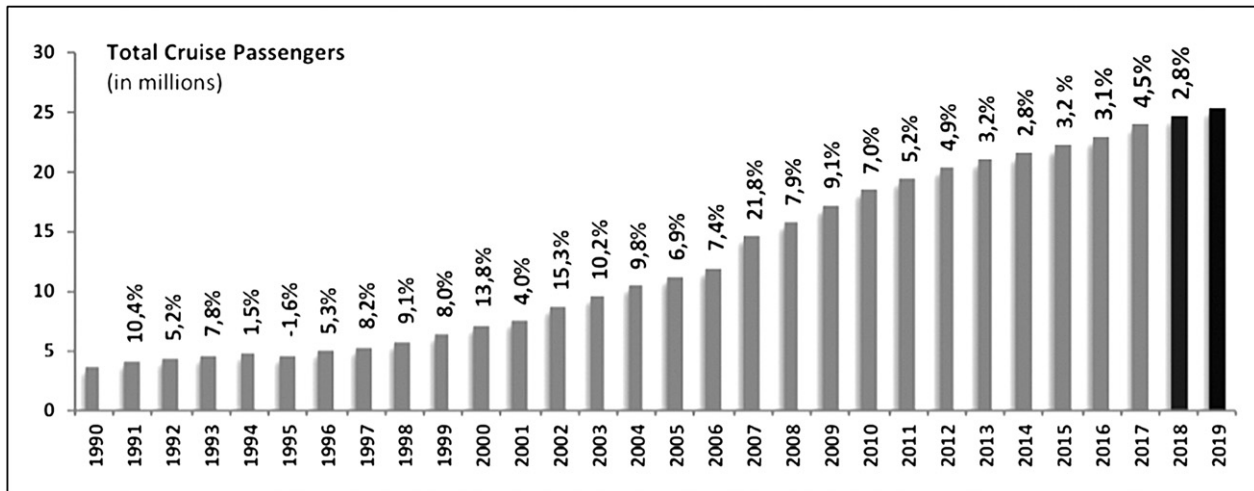


Figure 1 Global cruise passengers growth (1990–2019). Source: Cruise Market Watch, CLIA

Uninterrupted Growth and Globalization

Since the late 1960s, when specialized vessels of speed and comfort replaced the last liners, cruise shipping has witnessed uninterrupted growth. This growth is the result of a remarkable resilience in the face of economic, social, political, or any other crises that regularly challenge both shipping and tourism sectors. While economic recessions had a major impact over maritime cargo shipping, cruise lines and cruise ports managed to “cruise through the perfect storm” (Peisley, 2012), the global financial crisis of 2008–09 and the Costa Concordia loss. In 2018 a total of 28.2 million people took cruise vacations in one of the cruise regions of the world (North America, Caribbean, South America, Mediterranean, North Europe, Australia, Asia, Africa). In total, 10 years before the number of people that had embarked on a cruise were just 11.8 millions (Fig. 1).

This rise of cruising passengers has been the outcome of the renewals and rising capacity scale of cruise vessels, the improvements in shipbuilding, ports, and the growing interest of destinations that allow for planning of more complex itineraries, the sophistication and specialization of the product offered. Thus the 2019 cruise industry outlook of Cruise Lines International Association (CLIA, 2019), is projecting that more than 30 million single persons will cruise worldwide in 2019. Five years earlier, forecasts had projected this milestone to be reached in 2024.

The starting point of cruise geographical expansion dates back in the early 1970s; to a certain extent modern cruise shipping was invented in Greece by the pioneer cruise line Epirotiki Lines and the aggressive expansion strategy endorsed by its owner Mr. Potamianos who identified the limits of the East Med and decided to plan itineraries at the other side of the Atlantic. Proximity to the major passenger source market (United States) led rapidly to a focus on the United States and the Caribbean that was destined to provide the basis for a remarkable growth and expansion of cruising.

A seemingly unstoppable globalization is highly supported by (1) the profitability of cruise lines; (2) economies of scale; (3) the growing segmentation of cruise shipping; and not least, and (4) the desires and strategies of cruise ports and destinations to host more cruise activities.

Globalization of cruise shipping is based on three types of itineraries structured by cruise lines: (1) perennial, responding to a region that is served throughout the year due to the resilient demand (with high/low periods) and stable weather conditions; (2) seasonal, to serve periodical market potential in periods with good weather conditions; and (3) repositioning, between perennial or seasonal markets (Rodrigue and Notteboom, 2013). Deployed vessels are involved in these different types of itineraries, so as to allow maximum capacity utilization in combination with the maximization of earnings per passenger and lowering of expenses (mostly via lower fuel consumption). Beyond the geographical distribution of passenger source markets, the factors affecting the deployment patterns include the need to match brands and ships to source market demographics; the links with airlift, and landside transportation; the opportunity to balance marquee destinations with lesser known gems, or that of developing new routes and generating itineraries for new markets, the potential shore excursions revenues comparing to on-board spending; fuel sourcing, availability and cost consideration. All these are examined in parallel with the time, speed and distance formula in order to decide where and how to deploy cruise ships.

The Geography of Cruise Shipping

Cruise lines restructure their deployment strategies trying to better understand what fits best in each case, and what are the strategies responding to the global presence of cruising. Differences among the markets and regions exist, the most obvious being the variance of profitability. The emerging multiregion industry development is directly linked with a changing geography.

The deployment shares of each region are a combination of the existing demand, and the willingness of cruise lines to develop new markets. Looking at the capacity of the cruise vessels deployed in each region, Caribbean stands as the dominant, with 34% of the global capacity. The Mediterranean Sea is the second biggest region, with a share of 17%. The nonMediterranean European market reaches 11%. While it has been for long standing as the third major region of all, today it is the fourth in the respective ranking, given that in the very recent years the Asian market—in particular China—experienced a record breaking growth (11%). The deployment of increased capacity, combined with the opening of sales offices by many cruise lines in China, Hong Kong, Korea, Singapore, and Taiwan resulted in a growth of cruise activities in Asia from 1.3 million in 2012 to almost four million in 2020. The growth is mostly based on the number of Chinese cruise passengers, which is as big as all other Asian markets combined.

Market Concentration

The three largest cruise lines, Carnival, Royal Caribbean, and Norwegian represent 80% of a highly concentrated cruise shipping industry. Carnival Corporation, with its nine different brands controls 47% of the global cruise market, Royal Caribbean Cruise Line and its three different brands control approximately one quarter, Norwegian Cruise Line Holdings is the owner of three different brands representing 9.5%, while the MSC owned vessels represent a 7% market share. Approximately 30 more cruise lines are currently in operation.

The top two corporations consolidate different brands aiming to cover a variety of market segments. Beyond different and potentially more effective corporate entities, this allows for differentiated services to be offered and the marketing of each brand as a unique of its kind, captivating a larger part of the demand.

Cruise Fleet: Economies of Scale

Economies of scale have been a key future of modern cruising. Today, the biggest cruise ships have a capacity of 6687 passengers, with the capacity of each of the 50 biggest cruise vessels in operation having exceeding 3000 passengers. The average dimension of a cruise ship is over 200 meters long, 26 meters beam, and 3220 passengers capacity, though standard deviation is big and such numbers need to be treated with caution. The average capacity of cruise fleet exceeded 2000 passengers for the first time in 2002, and 3000 passengers in 2006. The current order book suggests a clear trend of stabilization, though this does not include potential withdraws from the market: out of the 24 new vessels to be delivered in 2020, 6 will have a capacity that of more than 4000 passengers, and 3 of more than 2000. In total, 4 of the 24 vessels to be delivered in 2021 have a capacity of more than 5000 passengers. In 2019 the 407 ships provide 600,000 berths capacity, while in 2027 the number it is expected to increase to 472 ships and 785,000 berths.

Market Segmentation

The growth of cruising has led to market segmentation. Different types of vessels, associated with different amenities offered on board and ashore, define types of cruises offered, in turn having as target different groups of potential cruisers. Attempting to further penetrate the market, cruise lines, or specific brands of the bigger corporations, are present in one or more, of the major segments, aiming to expand the social and age groups they target, but also to generate repeaters (past cruisers that decided to return).

Contemporary cruises are popular amenity-packed cruises for people looking for lots of activities and a great value. These mainstream cruises rival land based vacation by offering a comprehensive and amenity filled vacation, inclusive of accommodations, meals, and entertainment, in a casual environment, with newer (or extensively renovated) ships offer modern design and comforts, and many more activities. Premium cruises are more upscale cruises also offering many amenities, with increased focus on refined service and more space. Priced inclusive of accommodations, meals, and entertainment, premium cruises value still exceeds or rivals the best packages offered by upscale hotels and resorts. Luxury cruises are defined by the highest levels of quality and personalized service offered on luxury cruise ships and ashore to exotic ports. Expensive when compared to the rest of the industry, luxury lines deliver value by offering more inclusive pricing than other cruise lines and opportunities to travel to exotic destinations. A fourth market segment is specialty cruises. These focus on a destination niche or a special style of cruising including expedition-style cruises, sailing ships, and a growing number of river cruises. They visit some of the world's most remote and unspoiled places to offer a unique experience that guests find educational and adventurous (on the structures of the cruise industry: Pallis, 2015; also contributions in Pallis et al., 2014).

Source Markets

The North American market has always been, and continues to be, the largest and most stable market since the beginning of modern cruise. The actual number of Americans that cruise per annum, mostly with American brands, is estimated to be 14 million passengers. There are also Americans who cruise on European brands in the Caribbean and Europe. Yet, while growth has slowed

in recent years, the low levels of cruise penetration to the local market support projections for further expansion of cruise in the region.

Beyond North America demand for cruising is mostly expressed in Europe. According to 2018 data, about 7 million European decided to cruise, with almost one out of four living in United Kingdom and Ireland. A total of 5 million of passengers came from Australasia, Asia, South America, and Middle East and Africa. Even though these higher percentages are to a certain extent the result of the small penetration levels of cruising in the particular markets in past years, it is also indicative of the advancement of stakeholders' strategies aiming to expand cruising in the particular areas.

Cruise Ports and Destinations

In all regions, ports and destinations have developed an interest in advancing their cruise activities. This is not least to the association of cruising with considerable financial contribution to the port cities or nearby touristic destinations. With the significance of the societal integration of ports with the port-cities rising, cruise is part of respective agendas of port authorities and other port managing organizations. In several parts of the world they have moved from multipurpose terminals or temporary docking facilities toward specialized terminals, in order to act as ports-of-call, and whenever possible as home-ports hosting the, financially profitable, departure, and conclusion of a cruise (Niavis and Vaggelas, 2016). A growing interest by third parties, including cruise lines, to invest in port facilities has followed (Pallis et al., 2018).

Among the core issues to be addressed by cruise ports are the availability of adequate infrastructure and the organization of the cruise terminal operations in an efficient way. Cruise ports need deeper and lengthier docks in order to facilitate the new generation of cruise ships efficiently. The necessity of new infrastructures poses significant challenges especially to those ports that are facing land scarcity or the need for regular dredging of their basin. In the latter case there is a need to minimize potential impact on the sea flora and fauna (through land reclamation) and find the best way for processing and use of the dredging operation products.

Beyond matters worth consideration in all ports, such as number of berths, water and sewage facilities, customs, agents, pilots, security and immigration processes, gangways usage etc., additional ones are present in the case of homeports. The development of transport and tourism capacities is also critical. These concerns need to be addressed following a better understanding of the exact implications that the enduring increase of the size and capacity of cruise vessels and the resulting scale of operations produce.

The optimal planning of cruise ports and their terminals enables the most sustainable operations possible. Inevitably this has brought in the agenda of cruise lines and port managers the issue of long-term arrangements. The parameters that would enable a location, or a port, to secure a long-term commitment of cruise lines that would provide the motive to proceed to product and process adjustments need to be defined. Destinations and ports seek ways to reach ways of collaboration with cruise lines toward this end and broader expertise on what needs to be done toward this direction is essential.

Berth allocation is a long-term planning issue for ports with key social implications. The practice refers to the advance planning of which cruise vessels will visit the given port a specific day for a specific timespan. Given the limitations imposed by the geographical distances between ports included in an itinerary and the lengths of cruises, the phenomenon of many operators berthing for the same hours is not rare. The problem of berth allocation is even more important in the case of smaller, secondary, cruise ports. In small picturesque destinations cruise calls might mean the relatively unpleasant situations of a crowded location at certain days or hours, even distortion of other tourist activities. In bigger ports this might take the form of congestion at the time of arrival of bigger ships on which thousands of people are cruising. The arrival of two average size cruise vessels at a given port means more than 6000 passengers disembarking at the same time. Hosted passengers might increase without an increase of the number of cruise calls. Yet, congestion in small and medium destinations is in some cases produced simply because of small additional number of calls. Without effective planning, during some days these destinations are subject to the pressure and the negative effect of too many passengers that can hardly be accommodated in a way allowing for a positive experience. The fact that in several cases the presence of cruises is marked by seasonality deteriorates the problem that smaller tourist attractive destinations face.

Local communities also endorsed at large, if not universally, the idea of hosting more cruise activities. Admittedly, passenger and crew spending contribute significantly to the cruise hosting economies. Cruise lines expenditures for goods and services in support of their operations, along with a number of other indirect and/or induced positive effects, provide additional motives for hosting more cruise activities. According to the Cruise Lines International Association, in 2013, a total of 114.9 million onshore visits by passengers and crew generated \$52.3 billion in direct expenditures at destinations and source markets around the world. These expenditures generated a total (direct, indirect, and induced) global output of \$117.1 billion. The production of this output required the employment of 891,000 full-time equivalent employees who earned \$38.5 billion in income. In the United States alone, the total impact of cruising grew by 26% since 2009, comparing to a rise of 14% of the GDP over the same period. In Europe, a number of economies (i.e., United Kingdom and Ireland, Germany, Italy, Spain, Scandinavian countries, and France) enjoyed cruise spending that in total (i.e., direct, indirect, plus induced) exceeds one billion Euros annually.

The Sustainability Challenge

The aspirations to host more cruise activities are increasingly combined with the strategic importance of sustainable growth; a combination that, ironically, is to a great extent founded on the achievement of the desirable growth.

The gigantism of cruise ships and the multiplication of cruise itineraries, implies a growing number of passengers arriving at a destination with one call alone, posing significant challenges on cruise ports and destinations hosting them. The “contemporary cruise” segment of the sector, that represents more than 75% of cruises, is offered by vessels that exceed in capacity the 3000 passengers/ 1000 crew threshold. Beyond the need for further infrastructure, this implies mass arrivals of tourists at local destinations, concerns for overcrowding, congestion, but also needs for massive operations, considerable footprint, and needs for receiving quantities of waste (for the Mediterranean case: [Pallis et al., 2017](#)). As a result—justifiably or not—local communities have started questioning the unqualified growth of cruising, which had for long been taken as a beneficial development. While the benefits, in term of spending at destinations, are profound this growth might be associated with congestion and a number of related externalities to be addressed.

A condition to enable local communities to extract the most benefits from the rapid cruise market growth is the definition and endorsement of the best practices and policy options at the disposal of cities, ports, and related stakeholders in order to mitigate the externalities produced by cruise shipping, such as traffic or environmental ones. The existing challenges to be addressed are of operational, social, and environmental nature.

Overall, for the cruise world sustainability implies a holistic management addressing issues related to three main pillars: (1) the economy, (2) the society, and (3) the environment. Since many different stakeholders are involved in this process, a critical factor in supporting sustainable shipping is the understanding of all parties concerns, needs and expectations. It is increasingly acknowledged that a strategic approach—involving constructive dialogues, partnerships, synergies, even joint Research and Development (R&D) initiatives—are some of the key instruments toward this end.

Securing most of the potential local gains via an expansion of the cruise activities requests a consensus of how this might be done and how it might be integrated with the local economic development strategies. Destinations and cruise ports have no unlimited spaces for the development of all different activities that they might wish to advance. There is frequently a restriction of space, either at the city or its waterfront, or at the port. The antagonism of different actors and industries to maintain or increase the shares of the space they use is not rare.

The development of cruise activities, or just cruise terminals might be associated with restrictions to residents’ access at the waterfront. On the one hand, the growth of the industry is based not only on the modernization of existing infrastructures, but also on the presence of new facilities and the spatial expansion of the terminals and the cruise related activities. On the other hand, waterfront development is appreciated in many cases as the way to preserve alternative uses and respect traditions of the cities. Given this appreciation, citizens’ objections to develop tourist activities or related infrastructures (parking slots, restricted access zones etc.) are not rare. The definition of the principles to be adopted and the parameters to be examined would secure a balanced approach.

The same stands true for port development. The limited port zone marks imply choices as regards a wide spectrum of activities that might develop. As multipurpose terminals are not anymore a viable option of cruise activities development, the growth of cruise sustaining results in a potential interference with other maritime transport markets. Cargo ports also seek to spatially and functionally expand so as to improve the level of their own integration in supply chains. Stakeholders involved in these markets would like to see biggest parts of the port devoted to their activities, rather than the expansion of cruise ports. A balanced port planning needs to take into account contradicting potentials of different segments development. How this might be done is a question deserving attention.

Cruise terminal site location is the question that follows. Which site, in which way, should be planned cannot be decided apart from broader destination or regional planning. When several potential port sites can be considered, broader goals such as spreading tourism to new areas, help strengthening infrastructure, create tourism routes, provide investment for key facilities are all part of the equation. Plans for other sector development and for other forms of tourism create cumulative effects. The forms of involvement of stakeholders in location choice and, when essential, spatial and traffic planning, is increasingly important.

The planning of the port and the hosting of cruise activities does not, however, end at the gate of the terminal. The transfer of passengers from the terminal to city and transportations related to shore excursions are conversely important. Efficient location choices are conditioned by the efficiency of private, or public, transport strategies (i.e., services provision to the terminal but also avoid interference with urban road traffic when a cruise call arrives or departs). Location choices are also linked with the capacity of several tourism related industries to connect and the urban tourism strategies that tourism organizations and other relevant decision makers have underway.

Thus resolving most of the above problems demands more than an agreement on some technical issues (i.e., berth allocation details). It also demands the development of two types of coordination. The first one is the coordination between cruise ports and cruise lines in order to synchronize the system at the port and the operations taking place at the port terminal. The second one is the coordination of tourist destinations, including local public authorities, museums, retailers and, foremost transport service providers (coaches, buses, taxis) and travel related industries, so as to create smooth embarkation and disembarkation processes and passengers flow in the destination. Even port arrangements such as the berthing planning cannot be efficiently implemented if there are no means to involve other actors at the visited destination so as to orchestrate the entire cruise supply chain.

Reversing Social Perceptions

The expansion of cruise activities has not left unaffected the image of cruising. The elite activity of some passengers per year has been replaced by the mass transportation of thousands of cruise passengers at once. A community’s decision to seek cruise ship visits

requires a number of industries to be involved but will also affect many, either directly or not. Besides, cruise is not dissimilar with the impacts generated by any other tourism development on the milieu and services of visited communities and sites. It might displace current activities by other tourists or by local residents, causing changes in costs, access, and variety. These changes can be positive or negative, that is overloading dock facilities or causing improved ones to be built; creating new services, or pricing the locals out of existing ones. The same change may be viewed as positive by those who benefit and negative by those who may not benefit. All these lead to societal approaches that conceive cruise growth being associated with the deterioration of the quality of life.

Aesthetics have turn to a major issue, as did overcrowding of marquee destinations. A marquee destination that has experienced the negative effect of such approaches is Venice. A ban on large cruise ships passing through the center of Venice was imposed in late 2014, preventing all ships over 96,000 gross tons from sailing to the city's main cruise terminal, and limit the number of bigger ships over doing so to five per day. The debate had gained momentum as citizens and local protest groups were discontent with the presence of the bigger vessels, arguing that they produce pollution and result in substantial levels of emissions, with the acid nature of the pollution thought to be potentially speeding up the erosion of the city's medieval buildings, which are already sinking into the lagoon surrounding Venice, itself an UNESCO heritage site. Even with the real environmental assessment pending, the image of giant vessels next to traditional buildings led to an overheated reaction by local communities; aesthetics has come to play, and cruise liners had to limit number of big ships to Venice.

Other marquee destinations including Barcelona, Dubrovnik, Santorini, French Riviera have experienced similar issues. Barcelona port authority had to substantiate the benefits of cruising via several studies in order to gain the "justification" of its cruise operation, Dubrovnik decided to limit the number of arriving tourists (8000 max) per day at the historical city, Santorini put the same threshold (8000 max per day) to cruisers arriving per day, while French Riviera ports limit the arrivals to 5000 cruisers per day.

Whether these challenges are unsubstantiated or not, these are conceptions and perceptions that have to be addressed in order to achieve sustainable growth. Cruise is inextricably linked with the carrying capacity of a destination, a concept applicable in tourism development strategies, that is, the ability of the destination to absorb tourism before negative impacts are felt from host country, or the maximum number of people who can use a site without causing unacceptable changes to physical environment or affect negatively the quality of visitors' experience. Of course, answering how many cruise passengers can be hosted or how many can be wanted are questions associated with several parameters (i.e., who decides, which are the management objectives, which are the attributes and capabilities of the destination) all of them associated with what local communities perceive.

Limitation of Environmental Externalities

In recent times the emerging question is how cruise shipping, ports and the related economic chain can operate efficiently, within a socially responsible and acceptable framework (Pallis and Vaggelas, 2018). The various environmental externalities refer to the handling of waste produced, water quality, air emissions, noise, and soil, whereas other issues (i.e., constructions that alters natural or build environment, fauna, energy resources etc.) are also part of the relevant agendas.

Addressing two key externalities produced by the provision of cruise shipping and the hosting of vessels and cruise passengers at cruise ports stand today as priorities. The same externalities are illustrative of the need for discussion and conclusions on measures to take place at international level. These externalities are waste management, and the various forms of emissions, including air and noise emissions.

Overall environmental issues with reference to cruise shipping and those with reference to cruise ports are strongly interconnected. Thus the tools, measures, and policies to combat the externalities caused by these two industries require a holistic study and approach developed internationally. Reports, studies, and essential inventories of measures for internalizing external costs are another field, beyond regulations that the international level discussions might advance facilitating growth of cruise activities.

Conclusions

An uninterrupted growth of cruise activities has been recorded for every single year, for more than three decades, supported by an unstoppable globalization trend. What started as a local industry, in the Caribbean and North America, has eventually turned to a global one. Today globalization refers to both supply (offer of cruises) and demand (internationalization of passenger source markets) sides of cruising.

Regional variants exist and the fundamentals of growth are not similar around the globe, these revealed when comparing the details of the trends in the traditional North American market, with the maturing European market(s), or the booming Asian one. The latter stands currently as the major force toward further growth. The same details also reveal that intraregional dynamics go hand-in-hand with substantial intraregional shifts of cruise activities.

Specific business strategies work in support of growth and globalization—such as taking advances of economies of scale in shipbuilding, modernization of product and fleets, market segmentation, and innovative itinerary planning. The adjustment of ports to serve these strategies and the desires of port-cities and linked tourist destinations to enjoy the benefits produced work also in favor of these trends.

The aforementioned developments fuel a growing interest in further understanding the structures and implications of the tremendous growth of cruise activities, the sophistication and the geographical span of the sector is present. The externalities of cruise activity, in the light of the increased social awareness, are a field that the cruise port community needs to develop “green” strategies in order to mitigate the key externalities produced in and outside the port area due to the continuous cruise activities growth. Such strategies are beneficial not only from an environmental and a social, but also from an economic perspective, improving the potential and sustainability of such development.

See Also

Seaports; Seaports as Clusters of Economic Activities; Port Management; Transport Modes and Tourism; Transport Modes and Globalization; Transport Modes and Cities; Transport Modes and Remote Areas; The Geography of Maritime Transport; Port Regulation; Harbours and Shipping, Environmental Issues of

Relevant Websites

Cruise Lines International Association: www.cruising.org
 Florida-Caribbean Cruise Association, www.f-cca.com
 MedCruise: www.medcruise.com
 Cruise Market Watch: www.cruisemarketwatch.com
 Seatrade Insider: www.seatrade-insider.com
 PortEconomics cruise studies database: www.porteconomics.eu/cruise

References

- CLIA, 2019. 2019 Cruise Trends & Industry Outlook. CLIA, Washington DC.
- Niavis, S., Vaggelas, G.K., 2016. An empirical model for assessing the effect of ports' and hinterlands' characteristics on homeports' potential. The case of Mediterranean ports. *Marit. Busin. Rev.* 1 (3), 186–207.
- Pallis, A.A., 2015. Cruise shipping and Ports. Cruise Shipping and Urban Development State of the Art of the Industry and Cruise Ports. OECD-ITF Discussion Paper 2015-14, OECD, Paris.
- Pallis, A.A., Papachristou, A.A., Platias, C., 2017. Environmental policies and practices in Cruise Ports: waste reception facilities in the Med. *Spoudai Quart. Econ. J.* 67 (1), 54–70.
- Pallis, A.A., Parola, F., Satta, G., Notteboom, T., 2018. Private entry and emerging partnerships in cruise terminal operations in the mediterranean sea. *Marit. Econ. Logist.* 20 (1), 1–28.
- Pallis, A.A., Rodrigue, J.-P., Notteboom, T.E., 2014. Cruises and cruise ports: Structures and strategies. *Res. Transport. Busin. Manag.* 13 (1), 1–5 (special issue).
- Pallis, A.A., Vaggelas, G.K., 2018. Cruise shipping and green ports: a strategic challenge. In: Monios, J., Bergqvist, R. (Eds.), *Green Ports: Inland and Seaside Sustainable Transportation Strategies*. Edward Elgar, Cheltenham, pp. 255–273.
- Peisley, T., 2012. Cruising Through the Perfect Storm: Will Draconian New Fuel Regulations in 2015 Change the Cruise Industry's Business Model Forever? Seatrade Communications Ltd., Colchester.
- Rodrigue, J.-P., Notteboom, T., 2013. The geography of cruises: itineraries, not destinations. *Appl. Geograp.* 38, 31–42.

Further Reading

- CLIA, 2018. European economic contribution report, CLIA, Washington DC.
- Esteve-Perez, J., Garcia-Sanchez, A., 2015. Cruise market: stakeholders and the role of ports and tourist hinterlands. *Marit. Econ. Logist.* 17 (3), 371–388.
- Lamers, M., Eijgelaar, E., Amelung, B., 2015. The environmental challenges of cruise tourism: impacts and governance. In: Hall, C.M., Goessling, S., Scott, D. (Eds.), *The Routledge Handbook of Tourism and Sustainability*. Routledge, London, pp. 430–439.
- Pallis, A.A., Arapi, K.P., 2016. A multi-port cruise region: dynamics and hierarchies in the med. *Tourismos* 11 (2), 168–201.
- Pallis, A.A., Arapi, K.P., Papachristou, A.A., 2019. Models of cruise ports governance. *Marit. Policy Manag.* 46 (5), 630–651.
- Pallis, A.A., Vaggelas, G.K., 2019. The Changing Geography of Cruise Shipping. In: Wilmsmeier, G., Monios, J., Browne, M., Woxenius, J. (Eds.), *Geographies of Waterborne Transport: Transitions from Transport to Mobilities*. Edward Elgar, Cheltenham.
- Stefanidaki, E., Lekakou, M., 2014. Cruise carrying capacity: a conceptual approach. *Res. Transport. Busin. Manag.* 13, 43–52.
- Vayá, E., Garcia, J.R., Murillo, J., Romaní, J., Suriñach, J., 2018. Economic impact of cruise activity: the case of Barcelona. *J. Travel Tourism Market.* 35 (4), 479–492.

International Maritime Regulation: Closing the Gaps Between Successful Achievements and Persistent Insufficiencies

Laurent Fedi, KEDGE BS, CESIT, Maritime Governance, Trade and Logistics Lab, Marseille, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction: IMO as the Global Leader of Maritime Regulation but Challenged by Unilateral Initiatives	600
The Key Features and Drivers of Maritime Regulation	600
The Key Features of Maritime Regulation	600
The Key Drivers of Maritime Regulation	601
The Three Pillars of Maritime Regulation: Safety, Security and Environmental Protection	602
The Traditional Era: "Safety First"	602
The Contemporary Period: Environmental and Security Concerns	603
Environmental Protection	603
Maritime Security	604
Fragmentation Versus Harmonization of Maritime Regulatory Framework? A Polycentric Response	604
Closing the Gaps Between Successful Achievements and Persistent Insufficiencies	605
Summary and Conclusion	605
References	606

Introduction: IMO as the Global Leader of Maritime Regulation but Challenged by Unilateral Initiatives

Shipping is certainly one of the most globalized industries with an intrinsic international dimension at different levels. Vessels evolve through different seas and oceans, move between different countries and therefore jurisdictions. Linking continents, shipping underpins the global economy and from commodities to finished products, sea carriage enables countries to satisfy their needs in raw materials and to export their production all around the world (read 10476: Introduction to Maritime Transport). According to the 2018 Maritime Review (UNCTAD, 2018), around 11 billion tons were carried by sea in 2017 and sea traffic is expected to grow in coming decades.

Considering the international context of maritime activities whether for shipbuilding or ownership, operation, manning, inspection and survey, an international regulatory framework has quickly been necessary. In 1948 the United Nations (UN) adopted a convention establishing the first specialized agency devoted to maritime matters: the International Maritime Organization (IMO). Entered into force in 1958, the IMO has progressively become the worldwide leader of maritime regulation involving the largest shipping nations and accounting for 172 member states and three associate members. Approximately 80 nongovernmental organizations (NGOs) and 65 intergovernmental organizations representing a wide variety of interests have consultative status and can attend IMO meetings. It is generally admitted that their contribution to IMO work is valuable notwithstanding some critics against the harmful effects of lobbying.

IMO is now a sexagenarian institution having adopted more than 50 conventions and protocols, more than 1000 codes and recommendations mainly focused on safety, security, prevention, and control of pollution from ships, compensation and related legal matters. Whereas successful achievements to date, especially as regard maritime safety, some persistent weaknesses remain and more particularly on ships' environmental footprint. In reaction to the IMO's slow pace on this topical issue, national policymakers have followed suit with unilateral legal frameworks calling into question the universal and uniform application of maritime regulation as promoted by IMO.

This study aims to take stock of IMO progress since its creation and to identify key weaknesses where action is necessary. It suggests which possible orientations maritime regulation has to follow in order to meet the current challenges of shipping confronted to the fundamental stakes of the 21st century. This paper is organized as follows: after this introduction, the second section deals with the key features and drivers of maritime regulation while the third section contemplates its main pillars. The fourth section highlights the current fragmentation of maritime regulatory framework versus the stakes of harmonization. An assessment of the IMO's successful achievements and persistent insufficiencies is carried out in the fifth section. Finally, some concluding remarks are provided in the last section.

The Key Features and Drivers of Maritime Regulation

The Key Features of Maritime Regulation

One often ignores the key features and drivers of maritime regulation. Maritime regulation is a large concept made up of different legal components. It implicitly includes maritime law and law of the sea. Maritime law on the one hand, is simply defined by all rules

dealing with navigation undertaken by sea. It is focused on vessels, maritime operators and related contracts such as contract of sea carriage, contract of chartering, insurance, or assistance. Maritime law mainly governs private business and solves the fundamental issues of torts. Maritime law comes from different sources beyond the sole IMO instruments: the Comité Maritime International or the United Nations Commission on the International Trade Law have contributed to the adoption of significant rules often influenced by English law. On the other hand, law of the sea is a body of rules on the use of the seas and oceans. Codified by the famous 1982 UN Convention on the Law of the Sea, it defines the state territorial waters and other key areas (e.g., internal waters or economic exclusive zones), it determines the right of states to control navigation, to fish or to mine oceans. Compared to maritime law, law of the sea shows a greater public nature insofar as it deals with sovereignty of states. While their rules are complementary, maritime law mainly addresses private relationships and law of the sea frames both relationships between public entities and between public and private stakeholders. Maritime law is recognized as a field of expertise, old traditions and specific provisions notably with regard to the limitation of ship-owners' liability. In case of marine casualty (collision or pollution for instance), in case of claims for property damage (e.g., cargo loss), in case of claims for loss of life or personal injury, the ship-owner is legally entitled to limit his liability. Numerous international conventions set out this strong principle that features the ship-owners' regime of responsibility and that confirms the singularity of maritime regulation. The 1976 Convention on Limitation of Liability for Maritime Claims and its 1996 Protocol or the International Convention for the unification of certain rules of law relating to Bills of Lading (1924) are famous instruments with a wide scope of application that lay down the key principles of this specific liability regime not always understood and often criticized by cargo interests. Accordingly, maritime regulation taken in the wide sense, encompasses maritime law and law of the sea.

The Key Drivers of Maritime Regulation

Concerning the drivers of maritime regulation, they are numerous and show different natures (Fig. 1). For decades, there is no doubt that the first driver has been repeated accidents. Regulators have indeed adopted new rules after the occurrence of such accidents and depending on their root causes. One can observe that IMO's strategy has mainly been reactive and barely proactive. Nevertheless, new drivers have progressively emerged. Environmental awareness is certainly the second driver but still raised by accidents. A long series of oil pollutions and tanker groundings have made the headlines from the 1970s to 2000s: *Torrey Canyon*, *Amoco Cadiz*, *Exxon Valdez*, *Agean Sea*, *Sea Empress*, *Erika* or *Prestige* which are the well-known, have constrained the IMO to adopt fundamental regulatory instruments (Wijnolst and Wergeland, 2009). In 1973 IMO launched the International Convention for the Prevention of Pollution from Ships (MARPOL) amended by Protocols in 1978 and 1997. At the same time, the first unilateral regulations were implemented by key players such as the United States. The famous 1990 "Oil Pollution Act" (OPA) for instance pointed out that IMO rules were not strict enough to enhance maritime safety and to prevent oil pollution. Even though the IMO could not adopt the same stringent US standards, the UN Agency has maintained its efforts to address safety and environmental concerns by enacting several major instruments, as we shall see in the following section. However, the third driver of maritime regulation is closely linked to unilateral initiatives. When a large country or a group of countries such as the European Union (EU) decides to adopt a regulation with transboundary effects, that is to say applicable to foreign ships, it can contribute to a positive evolution of international rules even though it calls into question the IMO's leading role. This has been verified for double-hull requirement imposed by the EU after the *Erika* case (Carpenter, 2012) and the recent regulation on CO₂ emissions (Fedi, 2017). These two initiatives have constrained the IMO to modify its own rules and to set higher standards.

The last and the most controversial driver is certainly the influence of certain lobbies on maritime regulation. This influence has recently been highlighted especially against IMO environmental policies. Notwithstanding their consultative status in the decision-making process, a 2017 study demonstrated that some NGOs representing industry sectors such as the International Chamber of

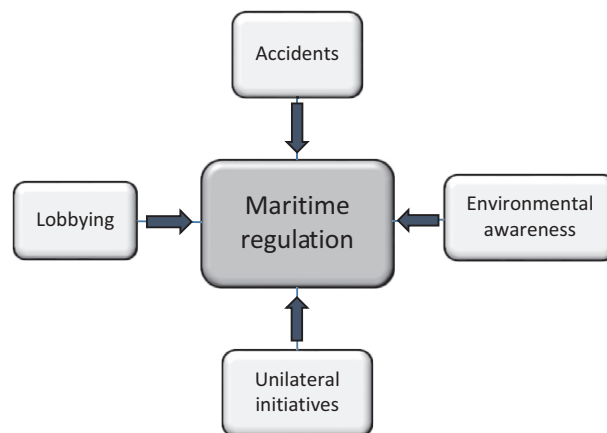


Figure 1 The four drivers of maritime regulation. Source: From the author, 2019

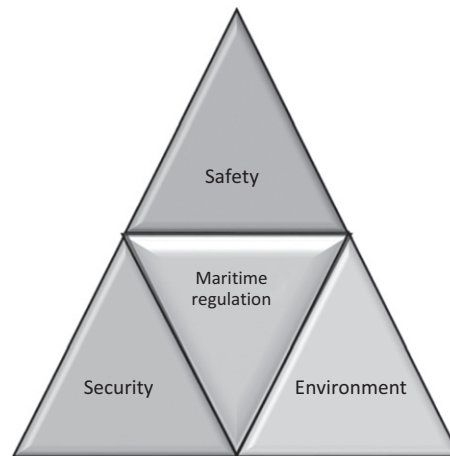


Figure 2 The three pillars of maritime regulation. *Source: From the author, 2019*

Shipping, the World Shipping Council, and the Baltic and International Maritime Council have hindered the adoption of climate change measures (InfluenceMap, 2017).

The Three Pillars of Maritime Regulation: Safety, Security and Environmental Protection

The IMO's slogan "Safe, secure and efficient shipping on clean oceans" summarizes the three pillars of maritime regulation: safety, security, and environmental protection (Fig. 2). These three complementary elements do not belong to the same history in the IMO development and one assumes that two stages of maritime regulation have existed—a "traditional" era as well as a more "contemporary" period—where a number of instruments have been developed. We will merely recall the most important regulatory tools.

The Traditional Era: "Safety First"

The first "traditional" stage has indubitably been ship safety that concentrated the greatest attention over the second half of the 20th century until the 1980s. While safety still represents a significant concern in the 21st century, other issues have constrained the IMO to diversify its policy measures.

Maritime safety encompasses a wide number of topics related to the ship herself, her typology (bulk carrier, oil tanker, container ship and passenger ship), her crew, equipment, cargoes, and her external environment. The key objectives are to continuously improve the operational conditions of navigation both in normal or emergency situations, to reduce pollution and more generally, to prevent casualties leading to detrimental consequences for crew, passengers and properties (cargoes and vessel itself).

Adopted in 1914 in response to the *Titanic* disaster, the famous International Convention for the Safety of Life at Sea (SOLAS) was revised in 1974 and many times after with the aim of enhancing new stricter safety rules on board. The binding International Safety Management Code that governs the management of ship operation safety (Chap. IX of SOLAS), the automatic identification systems (AIS) capable of providing information on the ship to other vessels and to coastal authorities automatically or the recent International Code for Ships Operating in Polar Waters (Fedi et al., 2018) are relevant illustrations of the constant evolution of SOLAS in line with new challenges in maritime safety. Thanks to the tacit acceptance procedure, an amendment to SOLAS enters into force within 18 to 24 months after its adoption unless, before that date, objections to the amendment are received from an agreed number of Parties. It enables rapid implementation of new maritime rules. It has recently been the case with the Verified Gross Mass rules on packed containers adopted in 2014 and entered into force 1 July 2016 (Fedi et al., 2019). SOLAS currently represents around 99% of the world tonnage. One cannot forget the 1966 Load Lines Convention that took into account the potential hazards present in different zones and different seasons with the aim to ensure adequate stability and avoid excessive stress on ship hulls as a result of overloading. Moreover, it has to be mentioned the 1972 Convention on the International Regulation for Preventing Collisions at Sea that provided guidance in determining safe speed and the conduct of vessels operating in or near traffic separation schemes. A Search and Rescue (SAR) Convention, adopted in 1979, completed the previous instruments. This major convention established an international system that covers SAR operations providing assistance for the rescue of persons and vessels in distress.

Regarding the human element, the 1978 Convention on Standards of Training, Certification and Watchkeeping for Seafarers set out standards for the seafarers' competence. This strategic instrument has regularly been revised by amendments with the purpose of enhancing requirements on the professionalism of seafarers. Considering that the human factor generally represents more than 90% of accident causes, training and experience of crew members are two strategic components of maritime safety.

To conclude these synthetic developments, we can consider that IMO has framed a robust and comprehensive regulatory framework covering the key aspects of maritime safety that classification societies closely monitor throughout the ship's life (IACS, 2015; read 10491: Ships Classification). Numerous rules are in force, whatever the vessel's position and situation. The designation of separation traffic schemes, vessel traffic services, ship's routeing, and places of refuge have allowed significant positive outcomes in terms of accident prevention. While casualties still occur, one observes a long-term downward trend of total losses for the period 2008–2018. 94 losses occurred in 2017 and 46 in 2018 compared to 151 in 2008 (ALLIANZ, 2019). However, some gaps remain. For instance, few international rules govern safety issues of ports and oil-platforms. Their specific status and geographical position still impede the development of international maritime regulation. In addition, with the development of IT systems, new safety challenges arise especially with regard to autonomous ships that will profoundly disrupt safety systems and patterns.

The Contemporary Period: Environmental and Security Concerns

Environmental Protection

The “contemporary” stage of maritime regulation started with the environmental awareness in the 1970s even if the 1954 International Convention for the Prevention of Pollution of the Sea by Oil initiated a greener path (read 10257: Shipping and the Environment). Nevertheless, the environmental shift has been forced by bad circumstances: oil spills. From 1970 to 1980, 25 oil spills over 700 tons occurred and IMO had to respond. In 1973 the IMO established the Marine Environmental Protection Committee (MEPC) and adopted the well-known International Convention for the Prevention of Pollution from Ships (MARPOL) and its Protocol in 1978. As SOLAS, MARPOL is governed by the tacit acceptance principle that allows a rapid entry into force of new amendments. Updated several times, MARPOL initially focused on the Prevention of Oil Pollution (Annex I) and Noxious Liquid Substances in Bulk (Annex II). The scope of application of MARPOL has been enlarged to different sources of pollution less media hype but also with significant negative environmental impacts: Harmful Substances Carried by Sea in Packaged Form (Annex III), Sewage (Annex IV), and Garbage (Annex V). Notwithstanding the global warming concern, regulatory provisions on atmospheric pollution from ships were only implemented at the end of the twentieth century with the 1997 Annex VI—Prevention of Air Pollution from Ships that entered into force in 2005. This latest text has concentrated a hot topic of debate in recent years. Modified regularly since its adoption, the Annex VI mainly imposes limits on different greenhouse gases (GHG)—in particular sulphur oxide (SOx) and nitrogen oxide (NOx), emissions from ship exhausts as well as particulate matter (PM)—and prohibits deliberate emissions of ozone depleting substances such as halons and chlorofluorocarbons. More stringent provisions on SOx and NOx emissions in specific areas with fragile ecosystems called “Emission Control Areas” (ECAs) have been established (e.g., Baltic Sea, the North Sea, North America, the United States Caribbean Sea, the Virgin Islands, Puerto Rico and Saint Pierre and Miquelon). The Mediterranean Sea, the Black Sea, the Norwegian and Barents Seas are potential ECAs. In addition, the Revised Annex VI implemented a Technical Code, which set limits on NOx emissions from diesel engines depending on their date of construction. Regarding CO₂ emissions, the IMO introduced mandatory energy efficiency rules in 2013 representing the first major step against the reduction of the most prevalent GHG. This regulatory framework is based on two main pillars: the “Energy Efficiency Design Index” aiming at requiring for different ship types a minimum energy efficiency level as regards CO₂ emissions per capacity mile and the “Ship Energy Efficiency Management Plan” for managing, improving and monitoring the energy efficiency of vessel operation (read 10258: Energy Efficiency of Ships). However, experts and IMO itself, recognize that these operational measures will not be sufficient to “satisfactorily” reduce GHG emissions in light of maritime growth. A latest amendment of Annex VI on data collection system (DCS) for fuel consumption entered into force in 2018 (Fedi, 2017). The purpose of this mandatory DCS is to collect ships' annual consumption by fuel type. In accordance with the values declared, the IMO shall produce an annual report to the MEPC summarizing the data collected. Despite this important progress, no decisions have been made as regards market-based measures such as bunker fuel levy or emission trading schemes and consequently, shipping is still without CO₂ restriction targets while decarbonization is more than ever a pressing issue (Psarafitis, 2018). However, MARPOL can be considered as a successful instrument representing around 99% of world tonnage for Annex I, II, III, and around 97% for Annex VI (October 2019 status).

Other key instruments have been adopted focusing on significant environmental concerns for marine preservation. Some of them are particularly important notably the 2001 International Convention on the Control of Harmful Antifouling Systems on Ships, which prohibits the use of harmful organotin in antifouling marine paints entered into force in 2008 and the 2004 International Convention for the Control and Management of Ships' Ballast Water and Sediments (BWM). This instrument aims to prevent and minimize the introduction of harmful aquatic organisms and pathogens from ships' ballast water tanks that is considered as one of the four greatest threats to the world's oceans. Once established, invasive marine species get out of control and can drastically modify marine ecosystems.

There is no doubt that the numerous regulatory measures on vessel-sourced pollution illustrate the commitment of the IMO toward protecting environment. In total, 21 IMO instruments out of 53 are indeed focused on environmental issues. One observes significant positive outcomes regarding oil pollution for instance and mechanisms designed to compensate damages such as the International Oil Pollution Compensation Funds. Nevertheless, one can be skeptical on many points and particularly on the slowness of certain topical concerns (Monios, 2019). The BWM Convention that governs prevention of aquatic invasive species only entered into force 13 years after its adoption. The Hong Kong International Convention for the Safe and Environmentally Sound Recycling of Ships adopted in 2009 is still not enforced (only 13 ratifications up to October 2019) while environmental and

working conditions are problematic in numerous ship recycling yards notably in India and Bangladesh. We will examine more closely these aspects in the fifth section.

Maritime Security

Maritime security is the second key aspect of the contemporary era of maritime regulation while piracy is an age-old issue. Two key instruments deal with security. The first is the 1988 Convention on the Suppression of Unlawful Acts against the safety of maritime navigation (SUA Convention) entered into force in 1992. The aim of this text is ensuring that appropriate action is taken against persons committing unlawful acts against ships such as the seizure of ships by force, and acts of violence against persons. The SUA Convention is completed by the Protocol for the Suppression of Unlawful Acts against the Safety of Fixed Platforms located on the Continental shelf. The second key regulatory framework is the famous 2002 International Ship and Port Facility Security (ISPS) Code. As a comprehensive security regime for international shipping and port facilities, the ISPS Code is a response to the perceived threats to ships and seaports in the wake of the 9/11 terrorist attacks in the United States and other specific events against ships. In the early 2000s, Al Quaida considered the shipping industry as a priority target and several terrorist attacks occurred notably against an American destroyer the *USS Cole* in Aden killing 17 crew members in September 2000. Two tankers endured the same scenario in oil terminals, the *Silk Pride* in Sri Lanka, seven people died (October 2001), and the *Limburg* in Yemen, killing one seafarer (October 2002). The ISPS Code has profoundly modified ports management. It has had numerous positive impacts such as an overall better control of the port area and a safer working environment. No major terrorist attacks have been observed since the implementation of the ISPS Code.

Nevertheless, with the development of IT systems, ports and ships are more and more connected in the cyberspace. Vessels are equipped with increasingly sophisticated IT systems such as AIS, e-navigation tools, and ports are also using numerous IT tools notably port or cargo community system (PCS-CCS) and dematerialized customs procedures. A new significant security threat has consequently appeared: the cyber risk. Stakeholders are particularly exposed insofar they are ever more interconnected. In 2013 the Port of Antwerp was targeted by mafia that used hackers to take control of the PCS in order to hide their drug smuggling. In June 2017 the world's largest shipping company Maersk was the victim of a cyber-attack. It caused computer breakdowns paralyzing online bookings and heavily disrupted the operation of several large container terminals especially Rotterdam, New York-New Jersey, and Jawaharlal Nehru (near Mumbai) for two weeks. Maersk decided to transparently communicate on the consequences of the attack and reported around USD 300 million of loss. Accordingly, shipping companies and port authorities are now aware of the cyber risk and take its management very seriously. In July 2017 IMO thereby promulgated the Guidelines on Maritime Cyber Risk Management (MSC-FAL.1/Circ.3) providing high-level recommendations on cyber risk management to safeguard shipping from cyber threats and vulnerabilities.

Regarding piracy, whereas one observes a downward trend off the Horn of Africa, West Africa, and Asia still face piracy and armed robbery acts. In 2018 201 incidents were reported showing a marked rise in attacks against ships and crews compared to 2017. In total, 143 vessels were boarded, 141 crew members taken hostage, 83 kidnapped, and 8 injured (ICC-IMB, 2018). The Gulf of Guinea concentrated the highest number of attacks. While oil tankers are still a privileged target, a wide range of vessels are also concerned by these attacks. Whereas several IMO instruments aiming at preventing piracy and illegal acts of violence against ships are in force, for example, Codes of Conduct (Djibouti in 2009 and Yaoundé in 2013) or the Best Management Practices, supported by dedicated international institutions such as the International Maritime Bureau, piracy remains a recurring phenomenon. Solutions are not only at sea but, above all, ashore. Unfortunately, the concerned states do not have the appropriate means or do not want to prevent and suppress acts of piracy and armed robbery at sea.

Fragmentation Versus Harmonization of Maritime Regulatory Framework? A Polycentric Response

According to the IMO "because of the international nature of the shipping industry, it has long been recognized that action to improve safety in maritime operations is more effective if carried out at the international level rather than by individual countries acting unilaterally and without co-ordination" (IMO, 2013). Consequently, the IMO promotes harmonized rules intended to be recognized and implemented by all. This principle is contradictory to the principle of Common but Differentiated Responsibilities and Respective Capabilities in accordance with the United Nations Framework Convention on Climate Change, Kyoto Protocol, and World Trade Organization rules. Nevertheless, the principle of uniform and global application of IMO rules is called into question and not only with regard to other UN governing principles.

Unilateral maritime regulation and private instruments have progressively emerged in different countries and regions. The United States have led the way over a period of time in particular as regard oil pollution from ships. With the 1990 OPA enacted after the *MV Exxon Valdez* accident, the United States demonstrated that unilateral regulations could have a crossborder impact since foreign tanker vessels had to comply with OPA provisions especially in terms of insurability in case of damage to the marine environment. Defining a global regime of liability both for shipowners and terminal operators, the holistic dimension of the OPA has never been reached at IMO level. The EU followed suit in the aftermath of the *MV Erika* oil slick in 1999 phasing out single-hull tankers that compelled IMO to accelerate its own timetable. More recently, the EU decided to set up a system for monitoring, reporting, and verification (MRV) of CO₂ emissions (Regulation 2015/757) that implements a close control of ships' fuel consumption and CO₂ emissions (EP, 2015; Fedi, 2017). In 2016 the IMO reacted in launching its own DCS (Resolution MEPC 278(70)), although diametrically different and less ambitious. Beyond the abovementioned rules, private instruments have also flourished both through

port authorities such as the World Ports Climate Initiative or from shipowners with the Clean Cargo Working Group with the same aim to improve ships' environmental footprint.

However, it seems that this fragmented patchwork splits up maritime regulation and weakens its coherence. While our ambition is not to give to our readers an unequivocal answer, one observes that the centre of "legal gravity" has greatly changed over the last few decades and IMO is no longer the unique custodian of shipping rules. A polycentric governance of maritime regulation is underway, particularly at an environmental level (Gritensko, 2017; Ostrom, 2010; Van Leeuwen, 2015), and this process cannot be stopped since local, national, or regional legislations have clearly demonstrated their efficiency in terms of rapid implementation and complementarity compared to international ones (Gilbert and Bows, 2012). Even though we are convinced that harmonized rules can guarantee a level playing field and avoid distortion of competition, universality has turned into a traditional myth of maritime regulation rather than a reality shaping the coming years of the 21st century. The adoption process of maritime regulation at the international level is indeed too slow to address topical issues that require appropriate and short-run responses.

Closing the Gaps Between Successful Achievements and Persistent Insufficiencies

Notwithstanding successful achievements of decades of incremental maritime regulation, the key question is how to enhance them and how to correct the persistent insufficiencies. First, with regard to positive outcomes, it is noteworthy that IMO instruments have contributed to a significant decrease in pollution from ships and marine casualties. Recent statistics clearly highlight these trends. According to 2019 Allianz Safety and Shipping Review, 46 losses occurred in 2018 compared to more than 200 in 2000 representing the lowest number of losses ever reached. Considering that the IMO is the smallest UN agency with a modest budget (€38 million in 2017), one can consider that these results are better than satisfactory. Moreover, it should be recalled that the enforcement of IMO conventions depends upon the Governments of Member Parties and experience has shown that some states have been negligent or reluctant to adopt and ratify certain instruments. This has significantly delayed the implementation of valuable regulatory tools and constrained states to adopt policy responses. The establishment of Port State Control (PCS) in different regions has illustrated the necessity to respond to shipowners' negligence and their substandard ships (Cariou and Wolff, 2015). Whereas the number of substandard ships has significantly shrunk, the recourse to flags of convenience remains a strong tendency of the shipping industry. Shipowners still want to operate their vessels with low taxes and low social costs in order to maximize their profits.

Considering the current situation of maritime regulation, a plea in favor of a more efficient, more transparent and more holistic legal framework is required. First, it is clear that IMO's slow pace on different key issues is no longer acceptable. Concerning ballast water management, more than 40 years of "wait-and-see" policy have dominated this topic. It is obvious that the same strategy is pursued for ships' CO₂ emissions (Bows-Larkin, 2015) that constituted a certain legal loophole of the scope of application of MARPOL until 2013. The same criticism is applicable for ships' recycling since the Hong Kong Convention is still not in force and most shipowners take advantage of this situation in flagging-out their vessels prior to releasing them to dismantling yards located in low-cost countries with little concern for labor conditions and environment. In the same vein, while the IMO adopted the famous Polar Code (Fedi, 2019; Fedi et al., 2018), it is also acknowledged that the Arctic Ocean requires stronger protective measure such as the designation of Particularly Sensitive Sea Areas and ECAs (read 10268: Arctic Shipping). Second, greater transparency could be encouraged through maritime regulation. A governance by disclosure could contribute to a better image of maritime transportation. While the EU has already moved in that direction with MRV Regulation, the IMO has still not adopted this path so far (Fedi, 2017). The IMO's slowness and incapacity to take decisions on topical concerns discredit its judicial status. Prof. Monios considers that "there are fewer places to hide for lobbyists and those arguing in bad faith for solutions known to be sufficient" (Monios, 2019). Third, a more holistic legal framework would be useful. The best example is the situation of seaports. These strategic facilities are still not fully incorporated within the IMO regulatory framework while a recent interest by the organization, which hosted the first-ever dedicated event in 2018 recognized that "port industry faced many challenges that mirrored those faced by the shipping community" (IMO, 2018). Few international conventions setting out provisions for port facilities are in force except the ISPS Code, MARPOL or the Blue Code for the loading operations performed in bulk terminals (Fedi, 2008). To sum up, ports are not fully integrated and recognized throughout a universal legal framework, whereas they are nodal points of maritime supply chains engaged in greener policies (Bergqvist and Monios, 2019). More rules should already have been set out at a global level in order to ensure safety and environmental standards in ports. This justifies why an International Port Safety and Environmental Management Code has been developed by Bureau VERITAS, a French classification society, with the purpose of preventing injury or loss of life, and damage to environment through management systems (Chauvel, 2008). One can regret that such a regulation comparable to ISM Code for ships does not exist at IMO level since it would provide both a harmonized legal framework and avoid distortions of competition between ports (read 10484: Port Regulation).

Summary and Conclusion

Obviously, maritime regulation is facing more complex challenges at different scales. While the IMO affirms that "shipping is the most efficient, safe and environmentally friendly mode of transporting goods around the world", strong criticisms are addressed, and the IMO is now under severe environmental pressure (read 10479: The Future of Maritime Transport). That is not necessarily "black or white". The first challenge lies in the governance of regulation. We have underlined the flourishing of unilateral

legislations due to the IMO's slow pace. Finally, IMO continues to be seen as a "laggard" instead of a leader on environment. The IMO has to restore its slightly tarnished image and to reestablish its leadership. In 2016 the adoption of a Code of Ethics aiming to enhance the trust in, and the credibility of the IMO was a good sign in terms of independence, impartiality, integrity, and accountability of IMO personnel. Nevertheless, it is not enough. We are convinced that the IMO has to set up a radical change in its environmental policy if the Organization wishes to reaffirm its long-standing position while it requires a strategic shift. On the one hand, more ambitious policies are needed and concerning air pollution from ships above all. On the other hand, greater transparency and accountability are needed.

Today, more than ever, it belongs to IMO to regain—or not—leadership in maritime regulation. The EU has clearly shown the path forward for a more sustainable shipping in terms of safety or environmental protection and other countries will follow the same strategies especially as regards air pollution. The IMO Strategic Plan 2018-2023 demonstrates that the organization is aware of the main current challenges on efficient implementation of regulations, new technologies, and climate change while one regrets that is not its top priority. Then, the main strategic stake is obviously to turn words into tangible actions as of today.

References

- ALLIANZ, 2019. Safety and Shipping Review: An Annual Review of Trends and Developments in Shipping Losses and Safety. Available from: <https://www.agcs.allianz.com/content/dam/omemarketing/agcs/agcs/reports/AGCS-Safety-Shipping-Review-2019.pdf>.
- Bergqvist, R., Monios, J., 2019. Green ports in theory and practice. In: Bergqvist, R., Monios, J. (Eds.), *Green ports Inland and Seaside Sustainable Transportation Strategies*. Elsevier, Cambridge, MA, p. 284.
- Bows-Larkin, A., 2015. All adrift: aviation, shipping and climate policy change. *Clim. Policy* 15 (6), 681–702.
- Cariou, P., Wolff, F.-C., 2015. Identifying substandard vessels through Port State Control inspections: a new methodology for concentrated inspection campaigns. *Mar. Policy* 60, 27–39.
- Carpenter, A., 2012. The EU and marine environmental policy: a leader in protecting the marine environment. *J. Contem. Eur. Res.* 8 (2), 249–267.
- Chauvel, A.-M., 2008. The IPSEM Code. Proceedings of the 24th International Port Conference on International Trade and Port Logistics, 17–19 February 2008, Alexandria, Egypt.
- EP, 2015. European parliament. Regulation (EU) 2015/757 of the European Parliament and of the Council of 29 April 2015 on the monitoring, reporting and verification of carbon dioxide emissions from maritime transport, and amending Directive 2009/16/EC, Official Journal L 123, 19 May 2015.
- Fedi, L., 2008. The international legal framework of port facilities: from current atomized legislation to future systemic approach. Proceedings of the 24th International Port Conference on International Trade and Port Logistics, 17–19 February 2008, Alexandria, Egypt.
- Fedi, L., 2017. The monitoring, reporting and verification (MRV) of ships' CO₂ emissions: a European substantial policy measure towards accurate and transparent CO₂ quantification. *Ocean Yearbook* 31, 381–417.
- Fedi, L., 2019. Arctic shipping law from atomised legislations to integrated regulatory framework: the polar code (R)Evolution? In: Lasserre, F., Faury, O. (Eds.), *Arctic Shipping, Climate Change, Commercial Traffic and Port Development*, Routledge, New York, NY, USA, pp. 117–136.
- Fedi, L., Faury, O., Gritensko, D., 2018. The impact of the Polar Code on risk mitigation in Arctic Waters: a 'Toolbox' for underwriters? *Marit. Pol. Manag.* 45 (4), 478–494.
- Fedi, L., Lavissière, A., Russell, D., Swanson, D., 2019. The facilitating role of IT systems for legal compliance: the case of port community systems and container Verified Gross Mass (VGM). *Supply Chain Forum: An International Journal* 20 (1), 29–42.
- Gilbert, P., Bows, A., 2012. Exploring the scope for complementary sub-global policy to mitigate CO₂ from shipping. *Ener. Policy* 50, 613–622.
- Gritensko, D., 2017. Regulating GHG emissions from shipping: local, global, or polycentric approach? *Mar. Policy* 84, 130–133.
- IACS, 2015. Classification Societies: what, why and how? International Association of Classification Societies. Available from: <http://www.iacs.org.uk/media/3785/iacs-class-what-why-how.pdf>.
- ICC-IMB, 2018. Piracy and armed robbery against ships report for the period 1 January–31 December 2018. International Maritime Bureau.
- IMO, 2013. Brochure IMO What it is OMI Ce qu'elle est OMI Qué es. Available from: http://www.imo.org/en/About/Documents/What%20it%20is%200ct%202013_Web.pdf.
- IMO, 2018. IMO News - The Magazine of the International Maritime Organization, Autumn 2018, 36p.
- InfluenceMap, 2017. Corporate capture of the International Maritime Organization How the shipping sector lobbies to stay out of the Paris Agreement. Available from: <https://influencemap.org/report/Corporate-capture-of-the-IMO-902bf81c05a0591c551f965020623fda>.
- Monios, J., 2019. Environmental governance in shipping and ports: sustainability and scale challenges. In: Ng, A.K.Y., Monios, J., Jiang, C. (Eds.), *Maritime Transport and Regional Sustainability*. Elsevier, Cambridge, MA, pp. 13–29.
- Ostrom, E., 2010. Polycentric systems for coping with collective action and global environmental change. *Glob. Environ. Change* 20 (4), 550–557.
- Psarafitis, H.N., 2018. Decarbonization of maritime transport: to be or not to be? *Marit. Econ. Logist.* 21 (3), 53–371.
- UNCTAD, 2018. Review of Maritime Transport, 116p. Available from: https://unctad.org/en/PublicationsLibrary/rmt2018_en.pdf.
- Van Leeuwen, J., 2015. The regionalization of maritime governance: towards a polycentric governance system for sustainable shipping in the European Union. *Ocean & Coastal Management* 117, 23–31.
- Wijnolst, N., Wergeland, T., 2009. Shipping innovation. IOS Press Amsterdam, The Netherlands, 831p.

Ship Classification

Gareth C. Burton, Mimosa T. Miller, American Bureau of Shipping, Spring, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	607
Marine Classification	607
Cargo Transportation	607
Bulk Carrier	608
Tanker	609
Container Carrier	609
Gas Carrier	609
General Cargo Carrier	610
Ro-Ro Carrier	611
Passenger Transportation	611
Ocean Liner	611
Cruise Ship	612
Ferry	612
Yacht	612
Service Vessels	612
Military Ships	613
Summary and Conclusion	615
Biographies	615
See Also	616
References	616
Further Reading	616

Introduction

The categorization of ships into different types is not to be confused with the process of vessel classification in the maritime industry. Under the vessel classification process, classification societies, which are nongovernmental organizations, establish and maintain technical standards addressing the construction and maintenance of ships and offshore structures. Through the process, classification societies verify the design of a ship against established rules and international regulations with the overall objective of promoting maritime safety and environmental protection.

This chapter provides an overview of the classification process before exploring the main categories and subcategories of ships.

Marine Classification

The process of classification in maritime industry provides a means to verify the integrity of the structure and machinery including the propulsion, steering, and other essential components of a ship. Classification societies act as independent arbiters of standards, establishing rules and verifying compliance during the design, construction, and operational phases of a ship's life cycle. Classification societies also act as Recognized Organizations, performing statutory inspections for vessels on behalf of flag states. This role of classification has been recognized by the International Maritime Organization within the International Convention on the Safety of Life at Sea (SOLAS) and the International Convention on Load Lines (1988). The classification status of a ship has important implications on insurance coverage.

Today, more than 50 organizations provide ship classification services worldwide. Among them, 12 classification societies are part of the International Association of Classification Societies (IACS), which collectively classes 90% of the commercial tonnage in the world (iacs.org.uk).

Cargo Transportation

As technology evolves, modern ships, especially commercial ships, are designed to efficiently serve very specific purposes. The majority of maritime transport comprises cargo transportation—among which oil tankers and bulk dry cargo carriers are the main

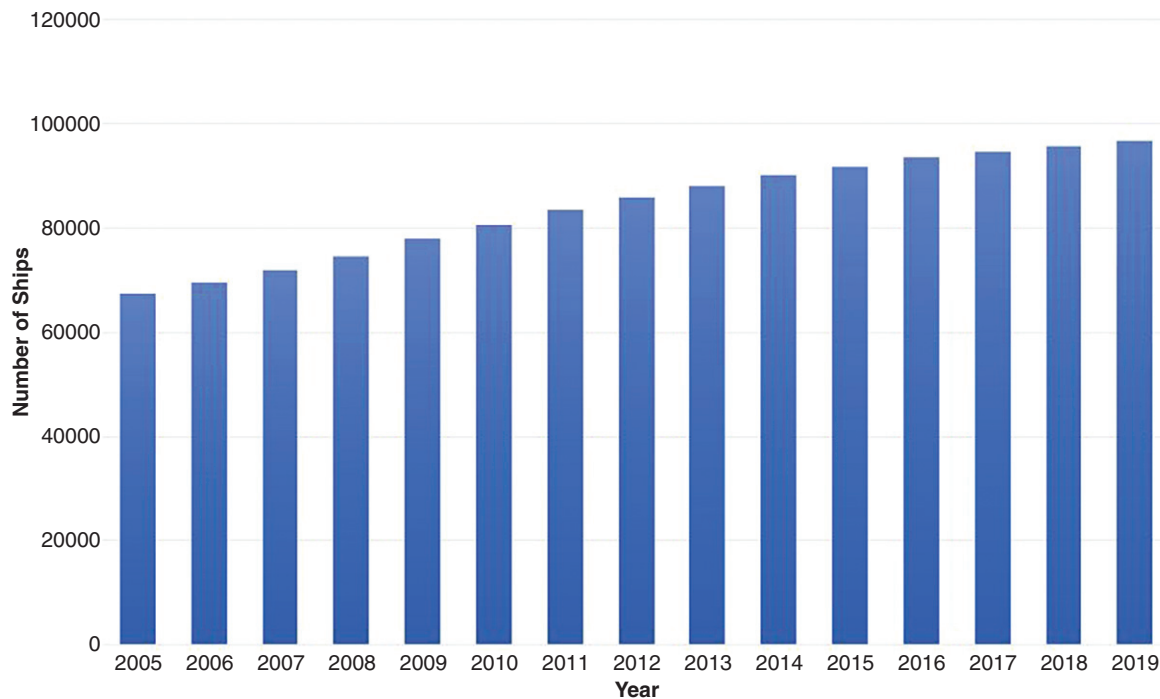


Figure 1 Total number of vessels over 100 gross ton.

players. The world fleet is growing each year. In 2019, the total number of vessels over 100 gross ton in the world was exceeding 96,000¹ (ABS database) **Fig. 1**.

Vessels have evolved significantly over the years. Utilizing human power and wind to propel vessels, our ancestors left footprints all over the world traveling by ships. Ships were designed and built to carry passengers, transport cargo, and when needed were equipped with weapons for use in attacking others or for protection from pirates. In a way, the vessels could be considered as multipurpose. When global trade boomed in the 16th century, businesses began to experience a demand for a large amount of trading goods, marking the beginning of specialized vessels. Vessels carrying large quantities of rice, wheat, spices, barreled oil, and wine were in demand, with new designs optimized for maximum carrying capacity.

The invention of steam engines boosted transatlantic traffic in the 19th century. However, wooden-hulled vessels were no longer sustainable for the newly installed steam engines, especially traveling transatlantic. Iron hull and later steel hull vessels became the new standard, which allowed vessels to be built substantially larger (Sager, 1996).

Bulk Carrier

A bulk carrier or bulker generally is designed to carry dry cargo in bulk, such as ore, grain, or coal; it is one of the most popular vessel types in modern time. Bulk carriers come in a variety of sizes, measured in Deadweight Tons (DWT), which represents the sum of the weights of cargo, fuel, fresh water, ballast water, provisions, crew, and if applicable passengers that a ship can carry **Table 1**.

Although vessels can be built in any size and shape, for the purpose of maximizing the economic value, they are generally built based on the trading route. The Handysize and Handymax vessels are the most popular sizes for bulk carriers. Their convenient size lets them into almost any port and terminal in the world. Panamax and Suezmax are sizes of vessel with maximum length and draft to go through the Panama and Suez Canals. Post Panamax are vessels built to pass the newly expanded Panama Canal which was completed in 2016 (pancanal.com, 2019). Port Kamsar in the Republic of Guinea is known for exporting large deposits of bauxite, which is used for producing aluminum. The sole pier in Port Kamsar, however, has vessel length restriction of 229 m. Thus, a Kamsarmax vessel is usually a Panamax sized vessel with an overall length less or equal to 229 m. Capesize Vessels are too large for Panama and Suez Canals, they trip around the Cape Horn in South America or Cape of Good Hope in South Africa instead. Very Large Ore Carriers (VLOC) and Very Large Bulk Carriers (VLBC) are also restricted to only trade between these “Cape” terminals around the world **Fig. 2**.

¹The number accounted for propelled sea going merchant vessels in excess of 100 Gross Tonnage
The number accounted for propelled sea going merchant vessels in excess of 100 Gross Tonnage

Table 1 Categories of bulk carriers

	Category of bulk carrier
≥ 200 K	VLOC and VLBC
120–200 K	Capesize
83–120 K	Post Panamax
80–83 K	Kamsarmax
65–80 K	Panamax
40–65 K	Handymax
10–40 K	Handysize
< 10 K	Small bulker

VLBC, very large bulk carrier; VLOC, very large ore carrier.

**Figure 2** Bulk carrier.

Tanker

Originally, liquid products such as oil and wine were transported in containers, primarily barrels for easy offloading and subsequent transportation. In 1886, in Newcastle Upon Tyne, United Kingdom, the world's first oil tanker, *Gluckauf*, was built with cargo compartments to hold large quantities of oil instead of barrels (Nielsen, 1958). *Gluckauf* was designed with longitudinal and transverse bulkheads to subdivide the cargo space into eight oil tanks. By removing the need for barrels, more oil could be stored on the ship. Today, tankers are equipped with high- and low-pressure piping systems to facilitate the loading and offloading of the liquid cargo at port facilities Table 2 and Fig. 3.

Container Carrier

Container carriers are easily recognizable with their neatly stacked containers. The truck-sized intermodal containers are standardized to be either 20 feet (6 m) or 40 feet (12 m) long. As a result, the capacity of container carriers is measured in how many containers they can carry, known as twenty-foot equivalent unit (TEU) Table 3 and Fig. 4.

Today, container carriers are an integral part of the intermodal transport network. The containers can be seamlessly transported between truck, rail, and vessels with minimum disturbance to the cargo within. The container terminals at ports are equipped with large cranes to load and offload containers, allowing minimum turnover time.

Gas Carrier

Natural gas and petroleum gas have gained popularity in recent years as alternative energy sources to traditional coal and petroleum. These gases were once considered as wasted by-products of petroleum drilling and production until extraction and storage technologies developed. Natural gas is a hydrocarbon gas consisting predominately of methane. Petroleum gas has a more complex profile, often referred to as propane and butane isobutane, while it can be a mixture of these two along with other hydrocarbon gases. Since the gases have different chemical composition, they are separated during transportation to avoid contamination. Piping, gas storage, and venting arrangements along with loading and off loading operations need to follow established requirements to ensure there is no leakage of the gas into the atmosphere.

Table 2 Categories of tankers

<i>Vessel DWT</i>	<i>Category of tanker</i>
≥200 K	VLCC
125–200 K	Suezmax
85–125 K	Aframax
55–85 K	Panamax
25–55 K	MR
<25 K	SR
≥10 K	Handysize
<10 K	Small tanker

DWT, deadweight; *MR*, Medium range; *SR*, short range; *VLCC*, very large crude carrier.

**Figure 3** Tanker.**Table 3** Categories of container carriers

<i>Vessel container capacity (TEU)</i>	<i>Category of container carrier</i>
>15 K	ULCS
8–15 K	Neo-Panamax
3–8 K	Intermediate
<3 K	Feeder

ULCS, ultra large container ship.

Trading and transporting gas in a gaseous state can be difficult to justify economically. As a result, a liquefaction procedure is commonly used to change the gases into their liquid state, which can result in up to 600 times the volume being carried in the same space. Although gas liquefaction technology has been around since early 1900s, the first Liquefied Natural Gas (LNG) carrier was retrofitted in the late 1950s and traveled in 1964 for long-distance transportation (Tusiani and Shearer, 2007). The liquefied gases are stored in insulated cargo spaces with temperature and pressure control that is unique to their chemical characteristics. For example, natural gas needs to be cooled to approximately -162°C to remain in a liquid state. The loading and offloading piping system at the port terminal allows the gases to return to gaseous state and transported to local distribution networks. The capacity of gas carriers is measured in the volume (in cubic meter) of the liquified gas they carry **Table 4**.

General Cargo Carrier

General cargo carriers are designed to carry homogeneous or inhomogeneous general cargos that are manufactured and stored in any shape or form. General cargo carrier designs include single-deck, tween-deck, or triple-deck configurations. The tween-deck cargo carrier was a popular vessel type before the invention of the container carrier. The 9000 DWT tween-decker Liberty-type vessels, built in the United States during World War II, were used to transport a variety of war supplies. After World War II, Liberty vessels were



Figure 4 Container vessel.

Table 4 Categories of gas carriers

<i>Vessel capacity (cubic meters, cbm)</i>	<i>Category of gas carrier</i>
≥80 K	LNG—Large
<80 K	LNG—Small & medium
≥50 K	LPG—Large
22.5–50 K	LPG—Mid
<22.5 K	LPG—Small

bought by Greek ship owners and became prevalent in the maritime trading market. Other general cargo carriers were soon built following Liberty's success. A more modern multipurpose general cargo carrier design was added to this type that can carry regular or irregular shaped and sized cargo (Spirou, 2011).

Ro-Ro Carrier

Roll on/Roll off (Ro-Ro) carriers allow wheeled cargos such as vehicles or trains to roll from pier to the vessel through ramps and vice versa. This unique type of vessel can be tracked back to the 19th century when the Ro-Ro carriers were used to carry steam trains across the river; the Firth of Forth ferry in Scotland dated back to 1851 was an example of an early Ro-Ro carrier (Hilton, 2003). The functionality of Ro-Ro carrier expanded to carrying war vehicles such as tank during World War II, and subsequently commercial cargos after the war. Nowadays, Ro-Ro carriers often operate as ferries with passengers driving their vehicles on to the vessel and driving off when the vessel reaches port.

Passenger Transportation

While passenger transportation was traditionally one of the main functions of ships, its popularity declined in the 20th century for long transits, as air transportation became a more efficient means of transportation. However, ferries and cruise ships remain popular.

The International Maritime Organization (IMO) defines a vessel carrying more than 12 passengers as a passenger ship. When traveling on international voyages, these passenger ships must comply with relevant IMO regulations, including those in the SOLAS and Load Line Conventions. Safety measures incorporated into these requirements include emergency escape routes, life-saving appliances, and lifeboats.

Ocean Liner

Ocean liners were once the only transportation mode available for transoceanic travel, operating on a regular schedule of ports. In the early 20th century, many of the liners were able to travel between continents and were built with exceptional luxury, offering a range of passenger accommodations. Many of the ocean liners were retrofitted to troopships during World War II to transfer soldiers. *RMS Queen Mary* held the record of most passengers transported on one vessel with a total of 16,683 soldiers and crews (Miller and Miller, 2012). After the war, many of the surviving liners converted back to ocean liners and resumed transatlantic voyages. Later, some of the liners were converted to cruise ships upon the rise of air travel while others eventually retired. Today, the *RMS Queen Mary 2*, remains as the only ocean liner, traveling between Southampton, England and New York City, United States.



Figure 5 Cruise ship.

Cruise Ship

Modern cruise ships often have a capacity of multiple thousand passengers, providing a wide range of onboard entertainment, including fine dining, theater, and sports. Cruise ships are often designed with passenger cabin arrangements to maximize the number of ocean-viewing rooms with private verandas. Although speed was not originally a priority for cruise ships, recently more cruise ships are designed to travel at a high-speed range, comparable to the ocean liners, to serve more seasonal business around the world.

Special destination cruises are designed to travel in different sea conditions. Popular destinations include the tropical waters in the Caribbean. This route requires cruise ships to have a shallow draft that allows them to enter ports. Cruise ships visiting polar regions are referred to as polar cruises and are designed to meet the International Code for Ships Operating in Polar Waters by the IMO [Fig. 5](#).

Ferry

Ferries come in a variety of sizes and typically travel between designated ports over relatively short distances on a regular schedule with fixed fares. The route can be point to point or with multiple stops in between as a water taxi or bus. They can be an integral part of local public transportation, or offered as an entertainment service for tourism. Some of the larger ferries travel longer distance and offer overnight cabins for passengers; the accommodations on the ferries can be very similar to those found on cruise ships.

There are many different designs of ferries offering diverse functionalities for passengers and their cargos. One of the most common types is the catamaran, a high-speed ferry that typically carries passengers only. Catamarans have a distinctive twin-hull design with very slender hull forms. The hull surface in contact with water is reduced in this design and thus has less drag resistance, allowing the catamaran to operate more efficiently at a high-speed range. Another unique design of ferries is the double-ended ferry. This type of ferry does not need to turn around at port because it has the same propulsion capability when operating in either direction.

Yacht

Though once used to pursue pirates by the Dutch Republic Navy ([Clark, 2015](#)), yachts today are mostly used for leisure in forms of racing and cruising. A yacht hull can be constructed with steel, fiberglass-reinforced plastic (FRP), or aluminum; the latter two are lighter in weight and are especially suitable for large yachts. The propulsion for modern yachts typically comes from internal combustion engines while sails are often installed as well. While typically monohull, yachts can also have twin or triple hulls for high-speed options, similar to a catamaran.

Service Vessels

Besides the vessels mentioned above, there are many other types of vessels, which exist to serve various purposes. These include service vessels such as tugboats and towing vessels, icebreakers in polar regions, industrial fishing vessels, research vessels, and service vessels to the oil and gas industry. Some of them are designed to carry passengers and special cargos while their range of activity is usually limited to certain regions depending on their service application.



Figure 6 Floating production storage and offloading unit (FPSO).

Vessels such as *tugboats* and *towing vessels* usually work close to harbors or inland waterways to provide assistance to larger vessels by providing additional propulsion power and maneuverability. Tugboats are often small in size but equipped with high power engines and strong hulls. They can help large vessels in narrow and busy ports or maneuver immobile craft such as disabled ships, barges, or oil platforms. Towing vessels function similar as tugboats, being used for emergency towing for disabled vessels and maneuvering large loads.

Icebreakers are specifically designed to travel through polar regions to break ice layers without damaging their hull. Such ships are designed to specific regulations, including the International Code for Ships Operating in Polar Waters, with a stronger and thicker hull and a sloping bow profile to cut through ice.

Fishing vessels can be found operating in lakes, rivers, or deep sea. Industrial fishing vessels are commercially operated to catch large quantities of fish and other seafood. A trawler is a popular type of fishing vessel that uses a trawl net to catch seafood. They often have large refrigerated cargo spaces to store the catch. Factory vessels can also be part of a fishing fleet, providing services to clean, process, and store seafood before transporting it to land to sell. Smaller privately owned fishing vessels can be considered supplemental, serving local or regional markets.

Research vessels are designed for exploration and data collection. There are many different types of research they can conduct, including oceanographic surveys, hydrographic surveys, polar research, and other industrial research such as fisheries and oil exploration.

Ships are also engaged in offshore and gas exploration and production. While some offshore installations are fixed platforms installed at one location for their service life, others are ship-shaped vessels used in mobile or semimobile forms.

Drillships are used for exploratory drilling. Such vessels typically have a moonpool opening in the middle of the vessel through which drilling operations into the subsea structure can be completed. Dynamic positioning systems installed on the vessels allow the vessel to remain at a fixed location, countering the impact of wind, waves, and current while the drilling operations are completed.

Ship-shaped Floating Production, Storage and Offloading (FPSO) vessels are another type of vessel used in offshore applications. These vessels, which are typically permanently moored, process the oil from the well, separating the oil and gas, storing the product onboard before offloading to a shuttle tanker **Fig. 6**.

Military Ships

Military ships include navy, coast guard, and constabulary vessels, and may be both combatants and noncombatants. They serve very different purposes compared to commercial or civilian ships and are typically national government assets. In many cases—particularly navy and coast guard ships—they are designed to go in harm's way; this could be for both coastal defense and combat roles with an aggressor's naval forces, as well as power projection in action far from the home nation. Consequently, these ships are typically designed to different requirements with much higher standards of structural, mechanical, and electrical integrity and redundancy, with greater focus on damage control and survivability in the event of an incident.

"Survivability" is a critical consideration for a military ship. For commercial vessels, this entails the ability to withstand loads due to inclement environments and operate even if critical components or systems fail. Military ships have the additional burden of surviving in combat environments. Both offensive and defensive measures can be taken to improve the survivability of the vessel. Survivability for naval vessels is composed of four basic properties:

Susceptibility: This is principally a ship's ability to avoid detection from other offensive ships, which depends greatly on its signature. Each naval ship has unique signature characteristics in various regions of the electromagnetic spectrum. The spectrums most commonly used for identification and detection purposes are visual, microwave, and infrared. Reducing or modifying the various signatures such that they match the operating environment can enhance the survivability of the vessel. Of course, sound and visual identification remain equally important to signature.



Figure 7 French Navy Aquitaine-class FREMM multipurpose frigate FS Provence (D652).



Figure 8 The United Kingdom's Royal Navy aircraft carrier *HMS Queen Elizabeth* (RO 8) (The Queen Elizabeth-class aircraft carrier HMS Queen Elizabeth (RO 8) sails the Atlantic Ocean).

Vulnerability: The ship's ability to continue to conduct its mission, either fully or partially, and to protect its crew, after an offensive weapon has struck the ship. The design is to attempt to minimize damage as best as practicable; this can either be done through thwarting the ability of an enemy to successfully acquire targets on the ship (such as through electronic countermeasures, or even through design by providing protected accesses for crew transit onboard the ship to minimize open air movement); providing means to reduce the energy of the weapon effect (such as through the use of outer tanks to avoid penetration into vital ship spaces); or structural hardening to prevent weapon penetration into the ship's hull or superstructure.

Recoverability: The ability of a ship to quickly recover post enemy engagement; this could be either to withdraw from engagement and seek sheltered waters, or to continue to engage the enemy while engaged in casualty repair (through either active damage control measures or via inherent system redundancies).

Lethality: This is the ability of a ship to successfully engage an enemy using its offensive weapons or other assets such as aircraft, helicopters, or unmanned underwater, surface, or aerial vehicles. This is typically considered the prime mission of military ships **Fig. 7.**

The modern navy fleet can be divided into two major groups: (1) Combatants and (2) Auxiliaries, such as Repair, Scientific Support, and Supply and Transport (also referred to as Sealift) vessels. The combatant vessels make up the battle fleet, with each vessel designated a specific role for military and combat operations. In lower intensity conflict operations, such as antipiracy or coastal patrol operations, some combatant vessels can and are expected to operate on solo missions, such as medium endurance patrol ships or offshore patrol vessels (OPV); some highly capable military ships may work alone in significant conflict environments such as the destroyer or frigate type ship. However, in most cases, navy ships are designed to operate as part of a larger task force or battle group; for those with helicopter carriers or aircraft carriers, these are sometimes referred to as carrier groups. Typical combatant ship types found in navies range from relatively small armored gunboats, minesweepers, and missile patrol boats to medium-sized offshore patrol vessels, corvettes, and frigates; to larger destroyers and cruisers; and finally to the largest classes of multirole capital ships referred to as helicopter carriers, amphibious warfare vessels, and aircraft carriers **Fig. 8.**

In addition to the combatant vessels, auxiliary support vessels are used to perform a vast array of fleet support functions such as repair (ship and submarine tenders), resupply (stores ships, ammunition carriers, and tankers known as "oilers"), transport (amphibious support vessels or sealift ships), and special mission ships. Special mission ships can incorporate a large variety of



Figure 9 United States Navy dry cargo ship *USNS Robert E. Peary* (Lewis and Clark-class dry cargo ship *USNS Robert E. Peary* (T-AKE 5) transits the Atlantic Ocean after successfully conducting a replenishment-at-sea).

vessels such as for oceanographic survey and mapping, electronic and oceanographic surveillance, and hospital and medical services. The auxiliary fleet provides vital logistics and support to the combatant force and has the added capability in many navies of resupplying a task force while underway, referred to as underway replenishment **Fig. 9**.

Summary and Conclusion

The maritime industry is made of vastly diverse ship types that are designed to perform different and specific tasks. Most of the transportation ship types are dictated by the cargo they intend to carry, may it be passengers, dry or wet cargo. Classification societies continuously serve an integral role in the industry as a third-party verification with the state-of-the-art development to ensure the safety standards for these ships.

Biographies



Dr. Gareth Burton is Vice President Technology in ABS (American Bureau of Shipping). He began his career with a consulting engineering company before joining ABS in 2001. During his time with the organization, he has held various roles in engineering, business management, and product development in the United States, Mexico, and Singapore. In his current role, he is responsible for the development and execution of the ABS research program.

Gareth holds a Bachelor's Degree from the University of Manchester, England and a Master's and Doctorate of Engineering from the University of Ulster, Belfast, Northern Ireland. In addition, he has completed the Executive MBA program through Texas A&M University.



Mimosa Miller is an engineer in ABS (American Bureau of Shipping). She began her career in 2016 in the "Aspire" rotational program and engaged various activities in different departments. Currently, she is part of the Technology department in Houston, and her fields include structural analysis, hydrodynamic analysis and emission control. Mimosa holds a Bachelor's Degree for Civil Engineering from the Cooper Union for the Advancement of Science and Art in New York, United States and a Master's Degree for Ocean Engineering from Texas A&M University.

See Also

Shipbuilding; Maritime Regulation; The History of Maritime Transport

References

- Clark, A.H., 2015. *The History of Yachting*. Fb&c Limited, United States.
- Hilton, G.W., 2003. *The Great Lakes Car Ferries*. Montevallo Historical Press, Incorporated, United States.
- Miller, W.H.J., Miller, W.H., 2012. *Picture History of the Queen Mary and Queen Elizabeth*(n.p.). Dover Publications.
- Nielsen, R.S.M., 1958. *Ownership of Tankers: The Independents Versus the Integrated Oil Companies*. University of California, Berkeley, United States.
- Sager, E.W., 1996. In: *Seafaring Labour: The Merchant Marine of Atlantic Canada, 1820–1914*. McGill-Queen's University Press, United Kingdom.
- Spirou, A.G., 2011. *From T-2 to Supertanker: Development of the Oil Tanker, 1940–2000*, Revised. iUniverse.
- Tusiani, M.D., Shearer, G., 2007. *LNG: A Nontechnical Guide*. PennWell, United Kingdom.

Further Reading

- ABS Guide for Building and Classing Yachts, 2019.
- ABS Rules for Building and Classing Marine Vessel, 2019.
- Alter, J., 2008. *Passenger Ships*. Cherry Lake Publishing, United States.
- Day, T., 2014. *Ecosystems: Oceans*. Taylor & Francis, United Kingdom.
- Delpizzo, R.D., Bhineka, K., 2019. *Addressing Military Survivability and Safety Aspects in Classification of Naval Design*. Pacific 2019, Sydney, Australia.
- Goebel, F., 2018. *Classification Societies: Competition and Regulation of Maritime Information Intermediaries*. LIT Verlag, Switzerland.
- International Association of Classification Societies, www.iacs.org.
- International Code for Ships Operating in Polar Waters, International Maritime Organization.
- International Convention on the Prevention of Pollution from Ships, MARPOL, International Maritime Organization.
- International Convention on the Safety of Life at Sea, SOLAS, International Maritime Organization.
- Liquefied Natural Gas (LNG). Office of Fossil Energy. Department of Energy. [Energy.gov](https://www.energy.gov/fe/science-innovation/oil-gas/liquefied-natural-gas). Retrieved December 2, 2019, from <https://www.energy.gov/fe/science-innovation/oil-gas/liquefied-natural-gas>.
- Paine, L., 2013. *The Sea and Civilization: A Maritime History of the World*. Knopf Doubleday Publishing Group, United States.

The Shipbuilding Industry and its Interactions With Shipping

Paul William Stott, Newcastle University, Newcastle, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Abbreviations	617
Introduction	617
Demand and Cycles	617
Competition and Competitors	619
Price and Backlog	622
The Interrelation Between Shipping and Shipbuilding Cycles	623
References	624
Further Reading	624

Abbreviations

GT: Gross tonnage

IMO: International Maritime Organization

SOLAS: Safety of life at sea

UK: The United Kingdom of Great Britain and Northern Ireland

US: The United States of America

WWII: World war two

Introduction

The function of the shipbuilding industry in the context of maritime transport is rather obvious—it creates the shipping capacity that moves seaborne trade. How it does, is a question of engineering and industry, and is of limited relevance in an encyclopedia of transport. Of more significant interest in this context is an exploration of how the shipbuilding and shipping sectors interact and this is the focus of the following chapter.

The mandatory use of classification societies in assuring the quality of new ships (IMO's SOLAS convention dictates that a ship cannot trade without a valid Cargo Ship Safety Construction Certificate) means that there is limited room for shipbuilders to differentiate themselves from competitors on the basis of product quality or features. From the ship owners' perspective, therefore, the significant questions relating to shipbuilding are fundamentally limited to: where am I likely to order a new ship; how much will it cost and; how long will it take to deliver?

Further to this, shipping and shipbuilding are both highly cyclical and the interaction of the cycles tends to be periodically antagonistic. There are two fundamental problem periods. In the trough phase of the shipping cycle the shipping industry generates too low demand to utilize shipyard capacity, with resulting low prices, bankruptcies of shipyards, and the use of government subsidy to keep some suppliers afloat. Conversely, at the peak of the cycle, the shipbuilding industry has a tendency to oversupply shipping tonnage at a high price, due to overinvestment in new ships, with resulting overcapacity in the fleet and low freight rates. For reasons that will be postulated later, the nature of the industry is to oversupply and this leads to cyclical troughs in shipping that are both longer and deeper than they would need to be if overinvestment and overproduction could be avoided (Oecd, 2019; Stott, 2018a).

These matters are explored by reviewing the following fundamentals that form the structure of this article:

1. Development of demand and cycles
2. Development of competition and competitors
3. Development of backlog and price
4. Interaction between shipping and shipbuilding cycles

Demand and Cycles

Output of commercial shipbuilding between 1890 and 2018 is presented in Fig. 1, showing gross tons (GT) of shipping delivered.

The first significant feature of this time series that might be noticed is four significant peaks of output that can be seen in 1919, 1943, 1975, and 2011. The average period between peaks has been 30.6 years, but with the pitch of the cycle increasing over time: 24 years in the period to 1943, 32 years up to 1975, and 36 years for the most recent cycle that peaked around 2011.

The second feature that might be noticed is the exponential increase in the scale of the industry in particular in the era of globalization following WWII. Fig. 2 shows output in the peak years and its exponential development.

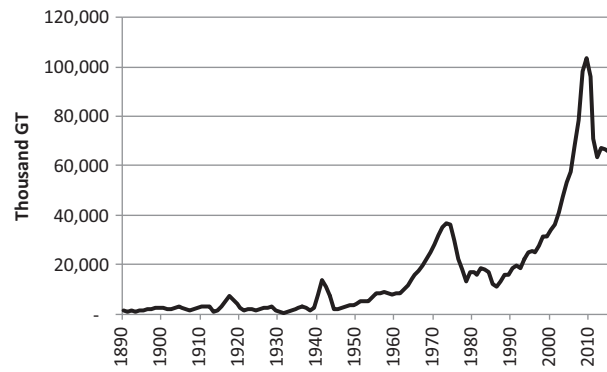


Figure 1 Global output of commercial shipping in Gross Tons (GT). WWII figures are difficult to obtain, with information on output from 'axis powers' in particular being scarce. The figures for the period 1939–44 are therefore estimates. *Source: Compiled from various archive sources available at the Marine Technology Special Collection, Newcastle University.*

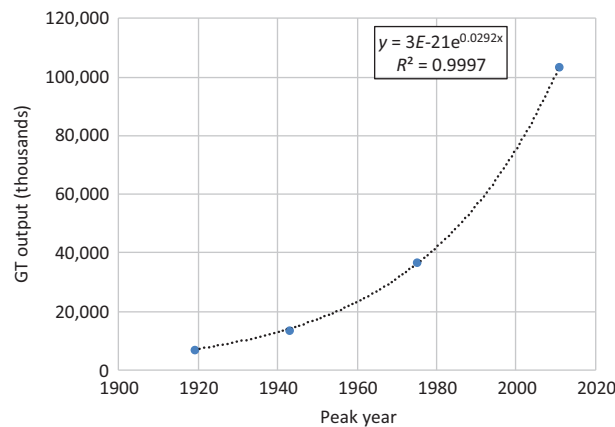


Figure 2 The exponential development of peak outputs.

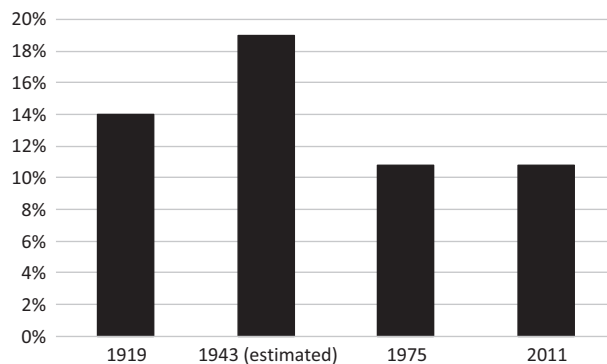


Figure 3 Peak shipbuilding output (GT) as a percentage of total fleet tonnage extant at the time. *Source: Compiled from various archive sources available at the Marine Technology Special Collection, Newcastle University.*

Each peak is exponentially larger than the preceding one and this has a tendency to “mask” the significance of previous peak values. Earlier peaks were equally significant at their time, however, as demonstrated by benchmarking the peak values against the total tonnage operating in the fleet at the time of the peak, as illustrated in Fig. 3.

The first two peaks of the 20th century were more significant, at 14% and 19% of total fleet tonnage, probably due to uplift from production to replace war losses. The latter two peaks were remarkably similar at 11% of fleet size. The significance of these statistics is that while past peaks look relatively small in retrospect, they would have been experienced as relatively similar in terms of size at the time they happened.

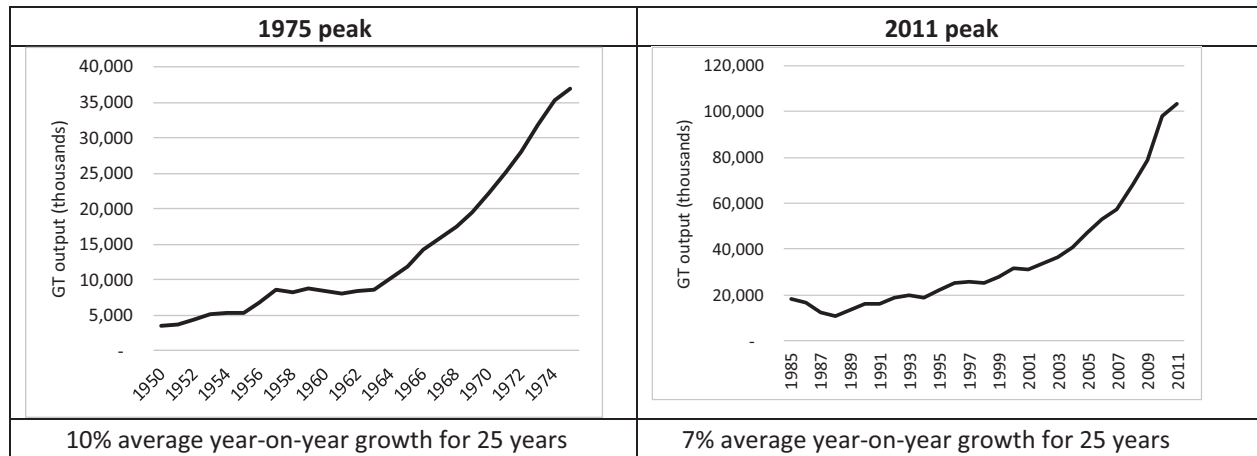


Figure 4 The appearance of market growth in the 25 years leading to peak levels.

Such an exponential increase in output would be anticipated in response to the exponential increase in seaborne trade that has also occurred in the period examined: quite simply, more trade requires more ships. What may not be anticipated, however, is the highly cyclical nature of this output and the extent of the dramatic and sudden fall in demand following the peak years—no such dramatic cycles are seen in the development of the seaborne trade that ultimately drives the demand for merchant ships.

Historical evidence reveals that this pattern of boom and bust long pre-dates the time series presented here, going back to at least the 18th century. There is clearly much more going on here than a rational investment in shipping tonnage to carry seaborne trade. In fact, demand for new ships is created at least in part by the demand to invest capital in shipping tonnage, rather than purely demand to fulfill the needs of world trade. This recognizes that the ship itself, possibly uniquely as a means of production, is a tradable commodity whose value rises and falls, presenting the opportunity for investors to make significant gains in what shipowners term “trading in steel,” that is to say the buying and selling of ships (Thanopoulou, 2002). Speculation and overinvestment are key drivers of the exaggeration of the shipbuilding cycles.

One final aspect of the pattern of output shown in Fig. 1 that is worthy of note is the consequence of the lengthening cycle in the postwar period and how the market would have looked to investors at the time these peaks were occurring. Fig. 4 shows how demand development would have looked to investors in the early 1970s and in the early 2000s.

Extended periods of almost unbroken strong market growth in shipbuilding in these two periods may have given the illusion that the underlying pattern of boom and bust in shipbuilding had not only disappeared but also that shipbuilding presented a reliable high-growth prospect for investment, with significant booms in shipyard construction accompanying these periods. The length of the cycles therefore carries with it something of a risk. The average of 30 years is of the order of magnitude of the length of a career and undoubtedly the senior management that had experienced the previous peak would predominantly have retired by the time the next peak manifests. Lack of memory in the system is therefore a concern and previous errors are prone to be repeated in future cycles.

Competition and Competitors

As with the development of demand, the availability of long time series enables trends to be identified that have a significant effect on the market. Fig. 5 shows the development of share of the market leader in global commercial shipbuilding in the period between 1892 and 2018 (Stott, 2018b).

It is easy to see from Fig. 5 why some commentators have tended to characterize competition in shipbuilding as a hegemony (Bruno and Tenold, 2011), that is to say dominated by a major supplier, with the market lead moving from United Kingdom to Japan following WWII, to South Korea in the 1990s and most recently to China. In the true sense, however, hegemony is probably an appropriate label only for United Kingdom in the late 19th century, with a peak of 82% market share in 1892. A noticeable trend in Fig. 5 is that the ultimate share achieved by each successive leader has reduced over time and that the period for which dominance has been maintained has also reduced with each leader: 50 years for United Kingdom, 40 years for Japan, but only 10 years for South Korea before China challenged the supremacy. Most recently, the three leading suppliers have seen something of a convergence of shares, rather than perpetuation of dominance, as was the pattern in the 20th century.

It is also relatively common for commentators to speculate where the dominating lead will move next, with both Vietnam and India, for example, having been mentioned as possible successors to China. Objectively, however, there is no reason to conclude that the pattern of “shifting hegemony” will propagate into the future and both these countries have not successfully pursued a market lead, at least so far. Barriers to success include very high capital cost and high risk of that capital in the face of the extreme market cycles previously described (Bruno and Tenold conclude that it is unlikely to be possible to establish a modern large scale industry without very significant government financial help). On the other hand, shipbuilding has proven to be a strong economic generator

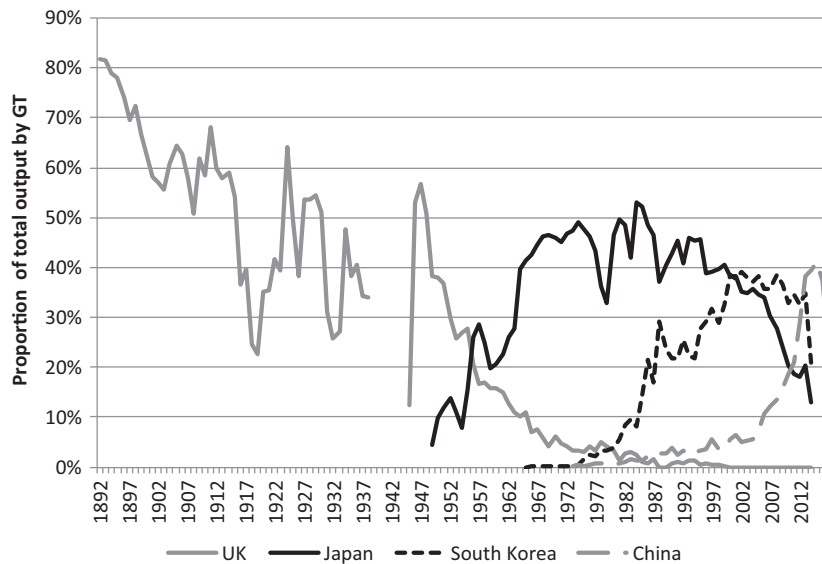


Figure 5 Development of share of market leaders between 1892 and 2018. Source: Compiled from various archive sources available at the Marine Technology Special Collection, Newcastle University.

in the development of all three post-WWII leaders and, probably, in United Kingdom that preceded them. Japan specifically chose shipbuilding, among other industries, as part of the vehicle for re-building of its decimated economy post-WWII, leading eventually to Japan's "economic miracle" of the 1960s. Shipbuilding's strong economic multiplier, requiring skills and supply chain industries, was subsequently used strategically in South Korea to help to progress that country from an agrarian to an industrial economy (commencing with the construction of the Hyundai Heavy Industries Ulsan shipyard in 1973) and shipbuilding has subsequently played a similar role in China's economic emergence (Stott, 2017a).

The analysis of market share alone is to some degree misleading because it masks the effects of the exponential expansion in demand discussed in the previous section and a more nuanced view of competition is obtained if the time series is represented as tonnage delivered, rather than share. This is presented in Fig. 6.

The United Kingdom's extreme dominance in the late 19th century can be seen to have been dominance of a different historic industry to that which developed post-WWII in the era of globalization. United Kingdom's typical output of around 1 million GT per annum has been dwarfed by successors. WWII was a watershed for the shipbuilding industry, which almost changed beyond recognition in the postwar period. Not only the scale of the industry and its products changed but the processes and methods also changed, moving from riveting and erection of the ship piece by piece on a slipway, to welding and construction of prefabricated

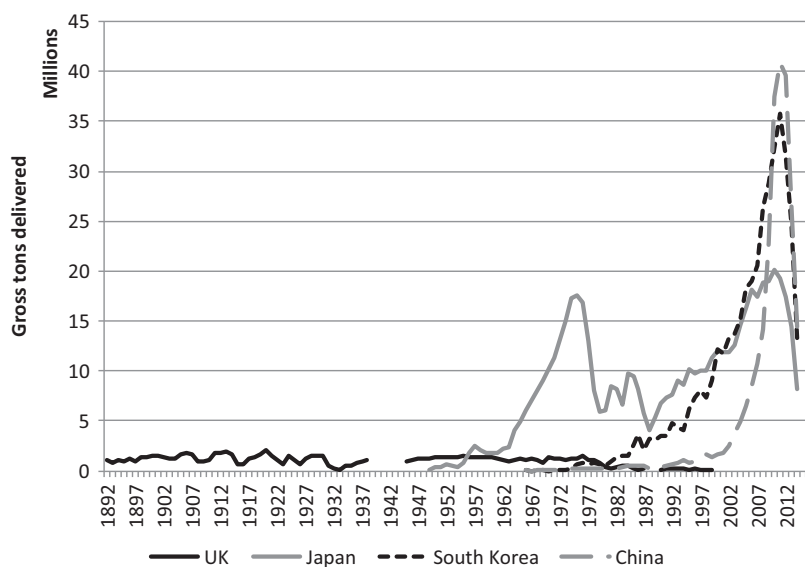


Figure 6 Tonnage delivered by market leaders.

blocks, propagating the technologies and methods developed in the United States “emergency shipyards” during the war period, to mass produce “Liberty Ships” and “T2 Tankers.” The United States seeded this technology into Japan along with capital for shipyard development and that industry set the pattern for the industries that followed.

What has also changed is the way that shipbuilders seek to gain competitive advantage, with economy of scale having become the predominant mode in recent decades. Both United Kingdom and Japan achieved competitive dominance through technical leadership and the availability of strong home markets. South Korea, on the other hand, not having the benefit of the strong home market, has focused on competitiveness through economy of scale, which in turn has been facilitated by the exponential expansion of demand. The use of economy of scale in shipbuilding relates not only to the shipyard itself but also to the supply chain, bearing in mind that typically around 70% of the cost of construction of a merchant ship is constituted by steel and equipment purchased for fabrication and assembly. The factors of competitiveness developed by South Korea include strong domestic steel and equipment supply industries and the use of shipbuilding contracts that direct material purchases to those suppliers, giving buyers very little choice. A shipbuilding contract typically includes an attachment normally termed the “makers list,” which details the manufacturer of the main equipment items, such as main engine, generators, pumps, and so on. In Korean shipbuilding, the list directs the equipment purchases to domestic companies with established places in the supply chain, minimizing equipment cost through volume of purchases and by minimizing delivery and transaction costs. Control of the makers list is therefore a key to making economy of scale work in shipbuilding. To give an idea of the scale of volume achieved, the leading South Korean shipyards, at the peak of 2011, were fabricating around three million tones of steel per annum each, and purchasing, for example, around 80 main engines each per annum. Such volumes lead to significant economic advantages through sheer buying power and the incentive to set up supplier industries as close as possible to the shipyard. On the other hand, the supply chain suffers with the shipyards in periodic downturns, multiplying the economic problems that result (Stott, 2017b).

The level of scale in shipbuilding can be judged by reviewing the output of market leaders at the start of each decade over the past 60 years, as shown in Fig. 7.

It can be seen that the market leading shipyard in 2010 produced over thirty times as much tonnage and almost seven times as many ships as the market leader in 1960, with scale particularly taking off since 1990 as South Korea assumed the market lead. As a consequence of the pursuit of scale, the shipbuilding industry has become increasingly concentrated. In 2018, one quarter of the total commercial shipbuilding orderbook, measured by work content, was found in just five shipyards and one third in the leading ten shipyards.

China’s place at the top of shipbuilding’s “hegemony” is still playing out at the time of writing this article. China has the advantage that United Kingdom and Japan had before it, with a significant home market. Beyond that the pursuit of competitiveness through economy of scale and the development of ultra-large “mega yards” has to be questioned. The strategy works perfectly at the peaks but as seen earlier, the shipbuilding market is highly cyclical and the effectiveness of scale is reduced as utilization of capital investment, human capital, and the supply chain reduces rapidly in the downturns that have tended to follow the peaks. A significant question, therefore, is to ask whether shipbuilding can find a new operating model to move the industry on from the pattern of boom and bust, and the intermittent dependency on government subsidy that currently characterize it and that may be an inevitable consequence of the pursuit of scale and the development of “mega-yards.” There has to be a better way!

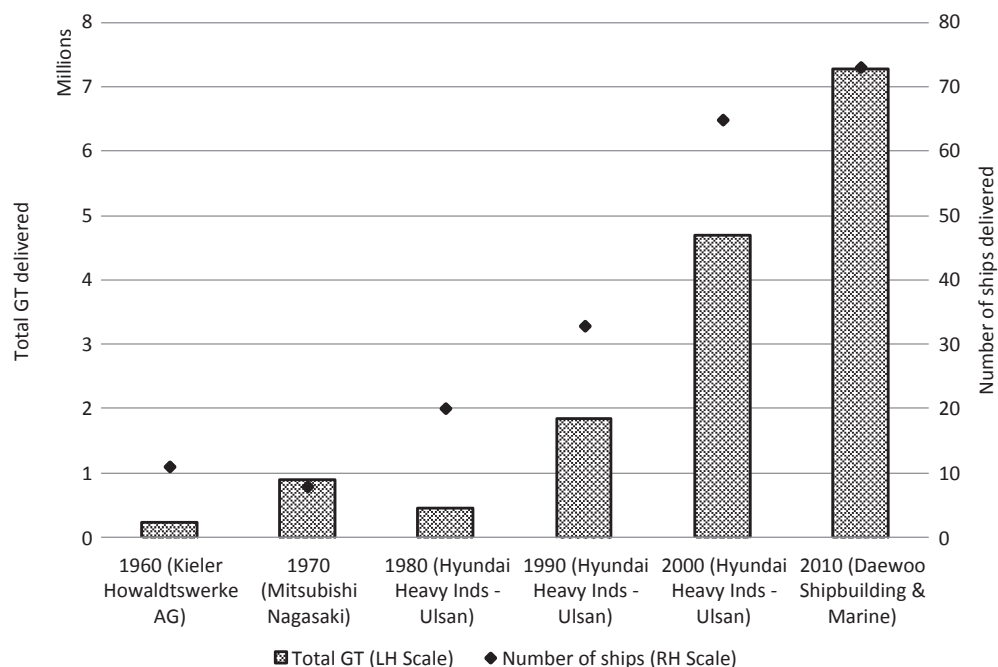


Figure 7 Output from the market leading shipyard at the start of each decade since 1960.

Price and Backlog

The economics of shipbuilding are substantially under-researched, with the profession of maritime economics concentrating primarily on shipping, trade, and port economics. It may come as a surprise, therefore, to learn that the definition of the global shipbuilding market in economic terms has not yet been achieved and the boundaries of the market and the econometrics of price formation remain elusive (Oecd, 2019).

From a pragmatic perspective it has long been assumed by those working in the industry that a single market exists, with prices determined by forces of supply and demand that are common across all ship types that compete for the same units of capacity. This is reflected in reality by a remarkable coincidence of price movement across almost all ship types over time. This tendency was first noticed a century ago (Ballard, 1921) and more recent studies have confirmed the parallel movement of prices of different ship types over time (Wijnolst and Wergeland, 2009). This pattern of behavior suggests that different ship types are part of the same market and that, from the perspective of the shipbuilder, the main ship types (tankers, dry bulk, and container ships) are substitutes in competing for available capacity. Shipyard capacity is broadly flexible across most ship types within the appropriate size range, with few exceptions. The exceptions are principally LNG tankers, for which technical, capital investment, and regulatory barriers to entry may exist, and passenger ships, for which technical, capital investment, supply chain, and customer factor barriers may exist.

The strong correlation of prices gives validity to the use of composite indices as a proxy for evaluating ship prices. The derivation of such an index over the long term is difficult due to shifts in ship types involved in world trade and because prior to the most recent period data on ship prices was relatively scarce. The most commonly used index is that of Clarkson research, presented in Fig. 8, covering the period of the most recent shipbuilding cycle.

Probably, the most striking feature of this graph is the extent of fluctuation of prices. The price of a VLCC, for example, rose from \$64 million in the beginning of 2003 to the peak at \$162 million 5 years later in 2008, only to fall back again to below \$100 million by 2010. The opportunity to make significant sums on the speculative purchase of ships is clear from this example, as is the opportunity to lose significant sums with poorly timed investments.

Research has generally concluded that there are two primary determinants of newbuilding price: the rational value of the ship to the buyer based on their expected earnings through freight, in other words what the buyer can afford to pay, and the cost of construction, or the minimum price a yard could accept without making a loss (Dikos, 2004; Haralambides et al., 2005; Jiang and Lauridsen, 2012). The relative influence of these factors appears to shift over the cycle, with rational value predominating in strong (sellers') markets and cost of construction predominating in weak (buyers') markets.

A significant predictor of price is backlog, that is to say the period between placing an order and receiving a ship (Oecd, 2019). The development of backlog (measured in years) over the most recent cycle is presented in Fig. 9. It can be seen that at the peak of the market in 2008, backlog was approaching 6 years, presenting a significant risk for buyers and a very comfortable forward workload for builders.

Backlog is of significance to both the buyer and the builder of a new ship. For the buyer, it represents scarcity of capacity and introduces an element of risk into the purchase: what will shipping market conditions be like when the vessel is delivered? For the builder, it represents scarcity of orders and ultimately the time up to the point when the shipyard will run out of work if no more contracts are won. It is calculated by dividing the current orderbook by the rate of production over the past year, with rate of production providing a proxy for active capacity in commercial shipbuilding.

The strong relationship between backlog and price is illustrated in Fig. 10, where the two factors are plotted together.

Two conclusions in particular have been drawn from the analysis of capacity and price. First, buyers become more price sensitive once the backlog exceeds about 4 years, becoming resistant to further price rises even if demand continues to increase. Second, backlog tends to stay above about 2 years, being a minimum level for confidence in the shipyard that it is not running out of work.



Figure 8 Clarkson Research Newbuilding Price Index.

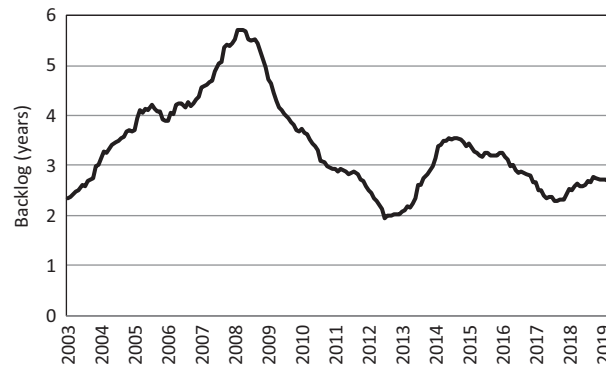


Figure 9 Development of backlog over the most recent cycle.

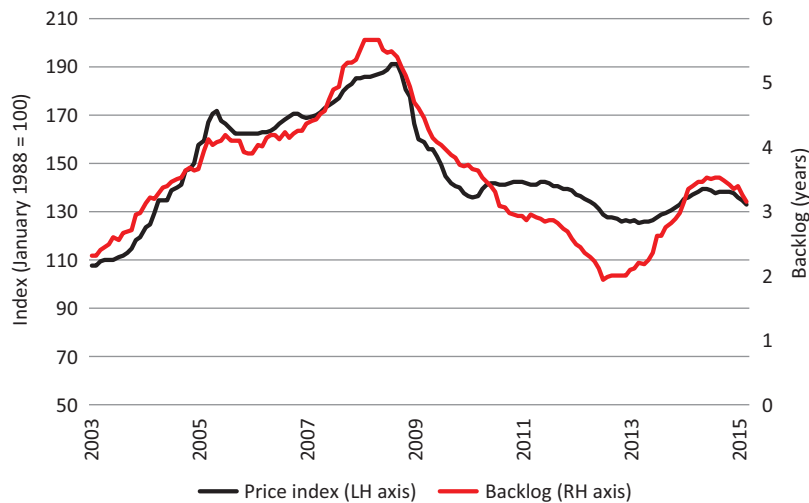


Figure 10 Backlog and price plotted together in the most recent cycle.

The Interrelation Between Shipping and Shipbuilding Cycles

Freight rates in commercial shipbuilding are determined by the interrelation between supply and demand, supply being the capacity of the fleet and demand being the volume of seaborne trade that needs to be moved to keep the global economy functioning. The economics of the supply and demand functions, particularly for bulk trades, are well established in the literature of maritime economics and the shapes of the functions tend to promote a particular pattern of freight rate development (Stopford, 2009). For most of the time the shipping markets function at relatively low freight rate levels, with an overcapacity of ships looking for cargoes. Periodically a balance is achieved, however, and at that point freight rates tend to rise spectacularly and make shipping a highly attractive business. Demand continues to grow as world trade expands but the backlog means that it takes shipbuilding some years to expand the fleet and a shortage of tonnage results. The resulting pattern is illustrated in Fig. 11, showing average daily earnings for a capsize dry bulk carrier (a bulk carrier of about 170,000 dwt that typically carries iron ore or coal).

It can be seen that earnings fluctuated around \$15,000–\$20,000 per day between 1990 and 2003, and had been at that level for most of the 1980s since the collapse of the previous boom. In 2004, however, the market picked up and rates rose to around \$90,000 per day and then in 2007/8 up to stratospheric levels of \$190,000 per day. At these rates, bulk shipping becomes highly profitable and investors and capital providers become very interested in the prospects of shipping. It had been 30 years since such conditions had been last seen and the memory in the system was insufficient to engender any caution in the lending of money and placing of orders to participate in this boom market.

What had happened was that, with freight rates persisting at a low level for so long, investment in new ships had not been sufficient to keep up with the expansion of trade and the fleet surplus generated by the previous peak in the 1970s was eventually eliminated, with the consequent upswing in freight rates. The rate of investment in ships that followed was far beyond that needed to sustain trade growth, however, and by 2008 it had become clear that the fleet was being overinvested and the adverse balance was being re-established. At that point, freight rates plummeted back to their long-term low point, this time well below \$20,000 per day, but backlog had grown to almost 6 years. In other words, there are no brakes in the system. Once it has been realized that

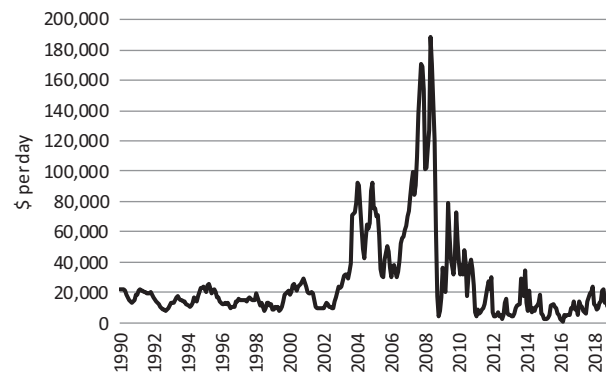


Figure 11 Average capesize long run historical earnings. *Source: Clarkson Research.*

overcapacity has become a problem, it took another 4–5 years before volume orders of new ships stopped flooding into the market. The result is that shipping switches back to the trough phase. The capital providers and investors exit the industry, which by this time has developed a poor reputation for investment, and the fleet will then proceed through another extended period whilst the expansion of world trade absorbs the overcapacity that has been created. At which point, the vicious cycle will reassert itself, unless we learn the lessons and try to find a way to stop this damaging pattern.

For those that know how to play the system, the exaggeration of the shipping and shipbuilding cycles is an opportunity to make large amounts of money in a short period. Over the long term, however, few really win. Shipyard capacity is left without sufficient orders and yards face either closure or the need for subsidy and shipping capacity is left facing a long period of suppressed freight rates and difficult business conditions.

It is both unlikely and undesirable that the cycles in these systems could be eliminated. What might be eliminated, however, is the damaging exaggeration of the cycles, rendering the businesses on a more sustainable footing. Part of the solution is to introduce memory into the system, we have enough time series now to understand what is happening. It might also be possible to introduce a leading indicator, to warn inexperienced capital when the rate of inflow of money is exceeding the rational demand for new ships that seaborne trade can sustain.

References

- Ballard, M., 1921. The Relation Between Shipbuilding Production, Prices, and the Freight Market. Transactions of the North East Coast, Institution of Engineers and Shipbuilders, California.
- Bruno, L., Tenold, S., 1970. The basis for South Korea's ascent in the shipbuilding industry. *Mar.'s Mirror* 97 (3), 201–217.
- Dikos, G., 2004. New building prices: demand inelastic or perfectly competitive? *Marit. Econ. Logist.* 6 (4), 312–321.
- Haralambides, H.E., et al., 2005. Econometric modelling of newbuilding and secondhand ship prices. *Res. Transp. Econ.* 12, 313.
- Jiang, L., Lauridsen, J.T., 2012. Price formation of dry bulk carriers in the Chinese shipbuilding industry. *Mar. Policy Manag.* 39 (3), 331–343.
- OECD, 2019. An analysis of market-distorting factors in shipbuilding. In: Policy Papers OECD Directorate for Science Technology and Innovation, OECD publishing, Paris.
- Stopford, M., 2009. *Maritime Economics*, third ed. Routledge, London.
- Stott, P.W., 2017a. Shipbuilding innovation: enabling technologies and economic imperatives. *J. Ship Des.d Prod.* 33 (3), 1–11.
- Stott, P.W., 2017b. Competition and Subsidy in Commercial Shipbuilding. PhD, Newcastle University, England.
- Stott, P.W., 2018a. Towards a better understanding of the commercial shipbuilding market. P. Session 5 of the OECD Working Party on Shipbuilding, 15 May 2018 Paris, OECD Working Party on Shipbuilding, Session 5, 15th May.
- Stott, P.W., 2018b. Competition and subsidy in commercial shipbuilding (in Japanese), Japan Ship Centre, London.
- Thanopoulu, H., 2002. Investing in ships: an essay on constraints, risk and attitudes. In: Grammenos, C. (Ed.), *The Handbook of Maritime Economics and Business*. Informa, London, pp. 623–641.
- Wijnolst, N., Wergeland, T., 2009. *Shipping innovation*. Amsterdam, IOS Press.

Further Reading

- M. Beenstock A. Vergottis *Econometric Modelling of World Shipping* 1993 Chapman & Hall New York; London.
- A. Burton *The Rise and Fall of British Shipbuilding* 1994 Constable London.
- W. Charemza M. Gronicki *An econometric model of world shipping and shipbuilding* Mar. Policy Manage. 81 1981 2130.
- T. Jiang P.N. Davies *The Japanese Shipping and Shipbuilding Industries: A History of Their Modern Growth* 1990 Athlone Press Atlantic Highlands, NJ; London.
- Cho, D.S., Porter, M.E., 1986. Changing global industry leadership: the case of shipbuilding. In: *Competition in Global Industries*, Harvard Business School Press, Boston, Massachusetts, pp. 539–568.
- L. Jiang *The international competitiveness of China's shipbuilding industry* Transp. Res. Part E: Logist. Transp. Rev. 60 2013 948.
- T. Jiang S.P. Strandenes *Assessing the cost competitiveness of China's shipbuilding industry* Mar. Econ. Logist. 144 2012 480 497.
- Lamb, T., Society of Naval Architects and Marine Engineers, 2003. *Ship Design and Construction*. Society of Naval Architects and Marine Engineers, Jersey City, NJ.
- Newcastle University Marine Technology Special Collection, British Shipbuilding Database. Newcastle University, UK.

Methods for Designing Public Transport Networks

Zain Ul Abedin*, Avishai (Avi) Ceder[†], *TUMCREATE Ltd. 1 CREATE Way, Singapore; [†]Transportation Research Institute, Technion-Israel Institute of Technology, Haifa, Israel

© 2021 Elsevier Ltd. All rights reserved.

Introduction	625
Literature Review	626
Heuristic and Metaheuristic Approaches	626
Exact Method Approaches	626
Typical Methodological Procedures	626
Limitations of Existing Methodologies	627
A Novel Route Overlap-Based Approach to Solving TNDFSP	627
Objectives	627
Route and Network Creation	630
Demand Assignment	631
Holistic Network Evaluation	632
Example	633
Summary and Conclusion	634
Acknowledgment	636
References	636
Further Reading	637

Introduction

For a public transport system to be successful, the network must be designed properly. Like any other system, the users of the system define the attributes that characterize, in their view, a good system. In the case of public transport, the users are the passengers. According to Walker (2012), there are seven broad expectations from passengers: (1) spatial accessibility, (2) temporal accessibility, (3) speed, (4) affordability, (5) safety and comfort, (6) reliability, and (7) simplicity. These seven expectations can be assigned to different planning levels when designing a public transport system. As identified by Ceder and Wilson (1986), there are three levels of planning: strategic, tactical, and operational. While spatial and temporal accessibility are covered in strategic and tactical planning, reliability, safety, and comfort fall under operational planning.

Since the focus of this chapter is network design, only strategic planning will be covered. This includes network design, frequency setting, and holistic evaluation of the designed network. Still, other passenger expectations are important and merit consideration. There is a common belief that if the strategic planning is done properly, other expectations are much easier to manage. Public transport network design involves creating a set of routes and their frequencies by incorporating objectives such as minimization of passenger costs or minimization of operator costs. These objectives are subjected to a number of different constraints, such as fleet size and unsatisfied demand percentage. Formally, it is known as the transport network design problem (TNDP) and if frequency setting problem is also involved then TNDFSP.

When considering the interests of both operators and passengers, it is quite clear that many of these are in conflict with each other. Adhering to the interests of one party is likely to lead to an unrealistic network for the other. In order to have a realistic public transport network that provides a suitable balance between passenger and operator viewpoints, transport authorities must play a role. They can do this by setting service-standard levels and policy guidelines.

While most public transport networks have evolved over the past few decades, the networks themselves have become more and more redundant and in some cases extremely inefficient. Therefore, restructuring or realignment of such networks is necessary for network improvement. There are different approaches to designing public transport networks. According to Kepaptsoglou and Karlaftis (2009), these approaches can be categorized into two major classes: conventional and heuristics. Conventional methods include the analytical models and mathematical programming formulations, whereas the heuristic approaches include heuristic and metaheuristic algorithms.

This chapter is structured as follows: first section will begin by presenting the literature review, which includes different approaches to tackling TNDP/TNDFSP. In second section, a novel approach will be used to explain how different types of public transport networks can be created, how demand is assigned, and finally how to holistically evaluate the public transport network. Finally, in third section we offer comments on how public transport authorities in North America have dealt with network redesigns and how an algorithmic approach can benefit public transport authorities to design better networks.

Literature Review

There are two main approaches to solving TNDFSP, namely heuristic and exact methods. Heuristics are the simplest procedures used in a systematic way to solve the TNDFSP. Occasionally, heuristics are used along with metaheuristic methods (stochastic optimization algorithms) to solve TNDFSP. These models are not problem specific and are extremely flexible, making it easy to adopt for any objective and constraint. In exact methods, the TNDFSP needs to be formulated into known mathematical models. Such models can find the optimal solution, but they are less flexible and can only be implemented on generalized small networks.

Heuristic and Metaheuristic Approaches

The use of heuristic and metaheuristic techniques are the most common types of approaches used in solving the TNDFSP; they are widespread and have been used in public transport design for many years. The basic versions are mostly domain specific and depend upon the knowledge and experience of public transport planners. In contrast, more advanced versions—called metaheuristics—do not require such knowledge.

Mandl (1980) use a rule-based heuristics to derive optimal routes with fixed headways. Ceder and Wilson (1986) adopt two-level approach for bus route network design and the frequencies. Baaj and Mahmassani (1991, 1995) propose an artificial intelligence-based approach to solve the TNDFSP problem. Carrese and Gori (2002) propose a hierarchical urban transport-network design methodology, composed of express and feeder routes. Lee and Vuchic (2005) use an iterative approach to solve the TNDFSP with varied demand. Fan and Mumford (2008) propose a metaheuristic-based approach for solving the TNDP. Bagloee and Ceder (2011) propose a hierarchical network-design methodology for actual-size networks. This include categorization of stops, manipulated shortest-path algorithm and metaheuristics.

Szeto and Wu (2011) use genetic algorithm (GA) and neighborhood-search heuristics to solve TNDFSP. Nayeem et al. (2014) use a greedy algorithm and GA with elitism to solve TNDP. Kiliç and Gök (2014) employ heuristics, hill climbing and tabu search to solve TNDFSP. Arbex and da Cunha (2015) use heuristics and alternating objective GA to find the best public transport network. Kim et al. (2016) solve the TNDP, by simultaneously deciding the mode type, route configuration, and its frequency. Huang et al. (2018) introduce a three-stage method to solve TNDFSP, including network hubs identification, route creation and optimization.

Exact Method Approaches

The mathematical approaches for the TNDFSP (TNDP + TNFSP) tend to focus on specific aspects of the problem, mainly due to its NP-hard nature and their tendency to get more inefficient as the problem instance grows.

Furth and Wilson (1981) propose a model to solve the problem of bus allocation to the public transport routes to maximize net social benefits. Constantin and Florian (1995) formulate a bi-level min-min problem to optimize the route frequencies, with the main objective being minimizing the total travel time. Bussieck (1998) solve the TNDP via relaxation and the branch-and-bound method. Wan and Lo (2003) solve the TNDFSP by modifying the existing transport system with the objective of minimizing the sum of the operating costs. Guan et al. (2006) solve the TNDP by a standard branch-and-bound method. Cancela et al. (2015) formulate TNDFSP as mixed-integer linear programming (MILP) formulation with the objective of minimizing passenger costs while adhering to the limited fleet-size constraint. Chu (2018) proposes an innovative mixed-integer programming (MIP) model to solve TNDFSP for minimizing sum of operating cost, passenger cost and unsatisfied demand.

The ability of exact methods to find the optimal solution makes it an obvious choice in many domains. However, in TNDP and TNDFSP, these methods do not scale well to actual-sized public transport networks. Most of the mathematical models are tested either on small test networks or simplified real transport networks.

Typical Methodological Procedures

According to Kepaptsoglou and Karlaftis (2009), the methodological procedures used to solve TNDFSP via heuristic and metaheuristic approaches can be grouped into four general categories: (1) route creation, (2) route selection, (3) route configuration, and (4) network improvement. Different approaches, methodologies, and procedures are used to solve TNDP/TNDFSP. To provide an overview and summarize the key features, selected studies are classified according to eight features, listed in Table 1. For the sake of simplicity and organization of the information, the following terms are used in Table 1:

AB	Available budget
AON	All or nothing
BM	Benchmark
CAM	Custom assignment model
CCHWAM	Capacity-constrained headway-based assignment model
CDAM	Combined-distribution assignment model
DS	Demand satisfaction
DS 0/1/2	Demand satisfaction through 0, 1, 2 transfers
DT	Direct trips

FIC	Fictitious
FS	Fleet size
H	Heuristics
HWAM	Headway-based assignment model
LF	Load factor
M	Mathematical
MH	Metaheuristics
ND	Network directness
NI	Network improvement
NoR	Number of routes
NoS	Number of stops
NoT	Number of transfers
OC	Operator cost
P&OC	Passenger and operator cost
PAM	Probabilistic assignment model
PC	Passenger cost
PD	Path deviation
RBT	Route backtracking
RC	Route creation
RCO	Route configuration
RE	Route efficiency
RH	Route headway
RL	Route length
RS	Route set
RSL	Route selection
RsC	Route set connectivity
SPAM	Shortest path-based assignment model
TNL	Total network length
TTT	Total travel time
USD	Unsatisfied demand
VC	Vehicle capacity
WT	Waiting time

Limitations of Existing Methodologies

Extensive literature is available on TNDFSP, and a number of reviews—such as those done by [Farahani et al. \(2013\)](#), [Guihaire and Hao \(2008\)](#), [Ibarra-Rojas et al. \(2015\)](#), [Kepaptsoglou and Karlaftis \(2009\)](#), and [Schöbel \(2012\)](#)—have been done in recent years. These reviews discuss the methodology type, inputs, outputs, constraints, and general aspects of the TNDFSP. Although they summarize different aspects of TNDFSP, none of them discuss the quality of results produced by these methodologies. More recently, [Ul Abedin et al. \(2018\)](#) conducted a study on the evaluation of the solution quality of the 27 studies. They performed the evaluation by creating a quality-evaluation platform that enables a comparison of their solution qualities. It revealed the following:

- The proposed solutions could perform according to their specific objective functions for smaller networks, but failed to show the same quality for larger networks.
- The quality of the solutions was not consistent with the variation in the network size and other inputs.
- Realistic demand assignment procedures should be used instead of simple procedures.
- Mode type should be considered as an integral part of TNDFSP, especially for large networks.
- For smaller networks, there were many comprehensive solutions, meaning that they performed well on all related indicators, but for larger networks, only a few solutions showed partial comprehensiveness.

A Novel Route Overlap-Based Approach to Solving TNDFSP

A novel heuristic approach is proposed to showcase how the TNDFSP can be solved and to design a network that is acceptable to relevant stakeholders, including passengers, operators, and authorities. The developed methodology includes three major components: (1) route and network creation; (2) demand assignment; (3) network evaluation. For network optimization, such approach can be integrated into any existing stochastic optimization algorithms such as GA (see [Ul Abedin 2019](#)), simulated annealing and particle swarm optimization.

Objectives

In this study, two objective functions are selected. For passengers, total travel costs are considered, including origin waiting time, in-vehicle time, transfer waiting time, and transfer penalties. For operators, only the operational costs are considered; these include

Table 1 Classification of studies related to TNDFSP

Study	Objective "To minimize"	Decision variables	Constraints	Approach	Procedures				Demand assignment	Evaluation indicators	Network
					RC	RSL	RCO	NI			
Lampkin and Saalmans (1967)	PC	RS, RH	FS	H			●	●	-	Operating deficit	Real
Rea (1971)	-	RS, RH	-	H				●	AON	-	-
Silman et al. (1974)	PC	RS, RH	FS	H	●			●	-	-	Real
Mandl (1980)	P&OC	RS, RH	FS	H				●	AON	DS 0 / 1 / 2, TTT	BM
Schéele (1980)	PC	RH	FS	M	-	-	-	-	CDAM	TTT, DS 0 / 1 / 2, RH	Real
Furth & Wilson (1981)	PC	RH	FS, RH, LF	M	-	-	-	-	HWAM	WT	Real
Ceder & Wilson (1986)	P&OC	RS, RH	RH, NoR, PD, FS	H	●	●			-	-	FiC
van Nes et al. (1988)	PC	SR, RH	AB, FS	H	●	●			-	RS, FS, DT	Real
Israeli & Ceder (1989)	P&OC	RS, RH	RH, FS	H	●	●			PAM	-	-
Baaj & Mahmassani (1991)	P&OC	RS, RH	RH, FS, ND	H	●	●		●	HWAM	DS 0/1/2, TTT, FS	Real
Shih & Mahmassani (1994)	P&OC	RS, RH	RH, FS, NoS	H	●	●		●	HWAM, CAM	DS 0/1/2, TTT, FS	Real
Constantin & Florian (1995)	PC	RH	FS	M	-	-	-	-	HWAM	TTT	Real
Baaj & Mahmassani (1995)	P&OC	RS, RH	RH, FS, ND	H	●	●		●	HWAM	DS 0/1/2, TTT, FS	Real
Pattnaik et al. (1998)	P&OC	RS, RH	LF, RH	MH	●	●			HWAM	DS 0/1/2, TTT, FS, USD	Real
Carrese & Gori (2002)	P&OC	RS, RH	DS, NoS, FS	H	●		●	●	HWAM	RS, TTT, FS, TNL	Real
Chakroborty & Wivedi (2002)	PC	RS	-	MH	●		●		AON	DS 0/1/2, TTT	FiC
Tom & Mohan (2003)	P&OC	RS, RH	LF, RH	MH	●	●			HWAM	DS 0/1/2, TTT, FS, USD	Real
Ngamchai & Lovell (2003)	P&OC	RS, RH	RH, LF	MH	●		●		HWAM	-	FiC
Petrelli (2004)	P&OC	RS, RH	-	MH	●		●		HWAM	-	-
Schbel and Scholl (2005)	PC	RS, RH	VC	M	-	-	-	-	CCHWAM	TTT, NoT	FiC
Lee & Vuchic (2005)	P&OC	RS, RH	-	H	●		●	●	HWAM	RS, TTT, TNL	FiC
Zhao & Zeng (2006)	PC	RS	-	MH				●	-	DS 0/1/2, RS	Real
Guan et al. (2006)	P&OC	RS	RL, NoT	M	●	●			CAM	TNL, NoT, TTT	Real
Fan & Machemehl (2006)	P&OC	RS, RH	RH, LF, FS, RL	MH	●	●			HWAM	-	FiC
Fan & Mumford (2008)	PC	RS	RBT, NoS, RsC	MH	●			●	SPAM	TTT, USD 0/1/2	-
Antoniou et al. (2019)	P&OC	RS, RH	-	H	●	●		●	HWAM	DS 0/1/2, TTT, FS	Real
Fan et al. (2009)	P&OC	RS, RSL	RsC, DS, RL	MH	-	-	-	-	SPAM	TTT, TNL, DS 0/1/2	BM
Zhang et al. (2010)	P&OC	RS, RSL	RBT, RsC, NoS	MH	-	-	-	-	-	TTT, TNL, DS 0/1/2	BM
Bagloee & Ceder (2011)	PC	RS, RH	PD,		●	●			CCHWAM		Real
Szeto & Wu (2011)	PC	RS, RH	RBT, FS, RH	MH			●	●	HWAM	TTT, NoT	Real
Cipriani et al. (2012)	P&OC	RS, RH	RL, RH	MH	●	●			HWAM	TTT, NoT, USD	Real
Mumford (2013)	P&OC	RS, RSL	RsC, DS, RL, RBT	MH				●	SPAM	TTT, TNL, DS 0/1/2, USD	BM
Hosapujari & Verma (2013)	P&OC	RS, RH	FS, LF	MH	●	●		●	-	TTT, DS 0/1/2, FS	BM
Nikolić and Teodorović (2013)	PC	RS, RH	-	MH	●	●			SPAM	TTT, DS 0/1/2, UDS	BM
Chew et al. (2013)	P&OC	RS	NoS, RBT, RsC	MH	●			●	SPAM	TTT, DS 0/1/2, UDS	BM
Kuo (2013)	PC	RS	RL, NoR	MH	●			●	-	TTT, DS 0/1/2, UDS	BM
Kechagiopoulos & Beligiannis (2014)	PC	RS	RL, RBT, NoR	MH	●		●		-	TTT, DS 0/1/2, UDS	BM
Nayeem et al. (2014)	PC	RS	-	MH	●		●		SPAM	TTT, DS 0/1/2, UDS	BM
Kılıç and Gök (2014)	P&OC	RS, RH	RBT, NoS	MH	●			●	SPAM	TTT, DS 0/1/2, RSL	BM

Majima et al. (2014)	P&OC	RS, RH	-	MH	●			●	SPAM	TTT, DS 0/1/2, FS	BM
Majima et al. (2015)	P&OC	RS, RH	-	MH	●			●	SPAM	TTT, DS 0/1/2, FS	BM
Zhao and Jiang (2015)	PC	RS, RH	FS, RH, RS	MH	●			●	SPAM	TTT, DS 0/1/2, FS	BM
Cancela et al. (2015)	P&OC	RS, RH	FS, VC	M	●	●			CCHWAM	TTT, DS 0/1, RH	Real
Arbex and da Cunha (2015)	P&OC	RS, RH	RsC, RBT, NoR, RH, RL, FS	MH	●	●			HWAM	TTT, DS 0/1/2, FS	BM
Kim et al. (2016)	P&OC	RS, RH	RH	H	●		●	●	-	-	BM
Buba & Lee (2016)	PC	RS	RL, NoR, RsC,	MH	●			●	-	TTT, DS 0/1/2, USD	BM
Huang et al. (2018)	P&OC	RS, RH	LF, RL, FS, RH	MH	●			●	HWAM	TTT, DS 0/1/2, RE	Real
Chu (2018)	P&OC	RS, RH	RH, FS	M	-	-	-	-	CCHWAM	TTT, DS 0/1, UDS, RH, NoR	BM

on-vehicle crew cost, vehicle operating cost, infrastructure maintenance cost, and overhead costs. These operational cost components are adopted from the [Australian Transport Council \(2006\)](#).

$$\min Z_1 = OwT + InVehT + TwT + tp.NoT \quad (1)$$

$$\min Z_2 = VehCrew + VehOpt + InfOptMain + Ovrh \quad (2)$$

Where

<i>OwT</i>	Origin waiting time
<i>InVehT</i>	In-vehicle travel time
<i>TwT</i>	Transfer waiting time
<i>tp</i>	Transfer penalty
<i>NoT</i>	Number of transfers
<i>VehCrew</i>	Hourly on-vehicle crew cost
<i>VehOpt</i>	Direct vehicle operating cost
<i>InfOptMain</i>	Infrastructure operations and maintenance cost
<i>Ovrh</i>	Overhead cost

The first term minimizes Z_1 , which is the total travel costs for all passengers. The second term minimizes Z_2 , which is the operating cost borne by the public transport operators. The selected constraints are as follows:

- The minimum number of nodes in a route must be in the predefined range (e.g., 2).
- All nodes in the network must be served by the public transport network.
- Each route must share at least one node with another route.
- Two routes cannot have the same node sequences.
- The maximum allowed transfers per trip must be below the prespecified range (e.g., 2).

Route and Network Creation

The feasible routes are created by employing the k-shortest path algorithm ([Yen, 1971](#)). There is, however, one condition that is applied when creating these feasible routes and that is a route must not deviate more than the specified detour limit (e.g., 1.4 times)

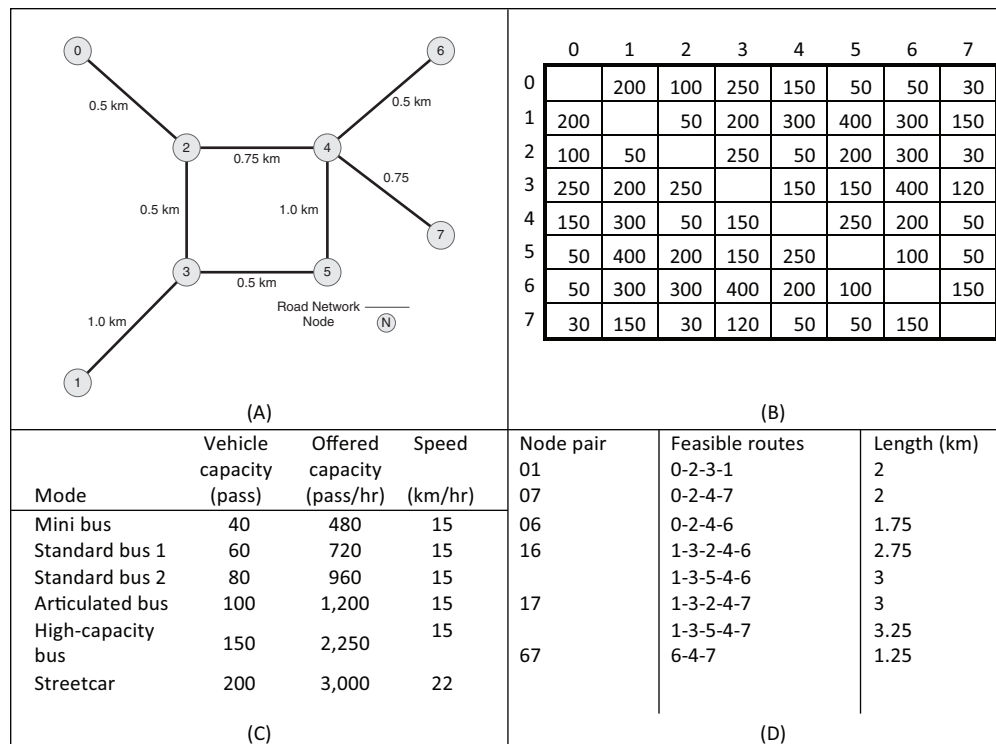


Figure 1 Example toy network, demand matrix, and k-shortest path-based feasible routes. (A) Network with stops and link lengths, (B) Demand matrix, (C) Vehicle selection, and (D) Feasible routes.

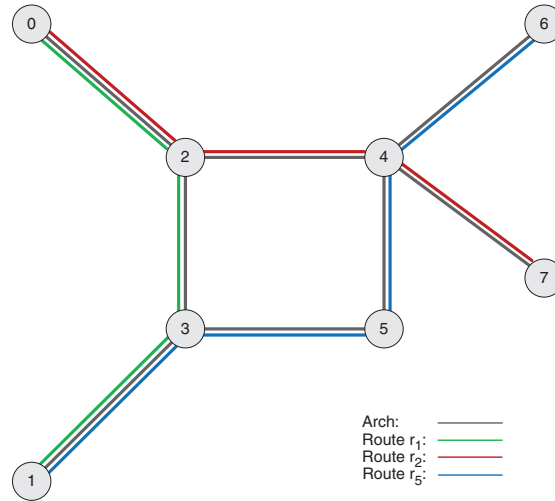


Figure 2 Public transport network created by route overlap concept.

from the shortest path. Let's consider a toy network depicted in Fig. 1A for calculating the feasible paths between selected node pairs. A total of eight feasible routes are created (Fig. 1D) among the selected node pairs.

Once all the feasible routes are created, the next step is to select a subset of routes to create a public transport network. This is done by using a route overlap ratio (ROR) concept. A ROR is the common route arcs that a potential route shares with the rest of the routes inside the public transport network. The term $A_{pr,r}^c$ represents the common arcs of the potential route pr that are shared with route r . The term $ROR_{A_{pr}}^{A_{pr,r}^c}$ is the ratio of the common arcs to the total arcs of the potential route pr . The pr is checked against all the routes R in the network for ROR. The pr will only be accepted if its value is less than the β for all public transport routes.

$$N_{pr,r}^c = N_{pr} \cap N_r; \quad \forall r \in R, \forall pr \in PR \quad (3)$$

$$ROR_{A_{pr}}^{A_{pr,r}^c} = \frac{A_{pr,r}^c}{A_{pr}}; \quad ROR \leq \beta, \forall r \in R, \forall pr \in PR \quad (4)$$

To understand the concept, let's consider the feasible routes from Fig. 1D. There are eight feasible routes and the β value is set to 0.35. The first route is inserted in the network, because it is the first route to be inserted, no comparison is required. For the second route it will become a potential route, and it will be checked against the first route. There is only one common arc between route1 and route2, with the ROR value of 0.33. This is acceptable; therefore, this route is inserted into the network.

$$N_{pr,r1}^c = ((n_0, n_2)) \quad (5)$$

$$ROR_{A_{pr}}^{A_{pr,r1}^c} = 0.33 \quad (6)$$

The third route becomes the potential route. However, this time there are two routes in the network to compare to. The ROR for first two routes are 0.33 and 0.66, among which the overlap with route2 is higher than the threshold value of β . Therefore, route3 will be discarded.

$$N_{pr,r1}^c = ((n_0, n_2)), \quad N_{pr,r2}^c = ((n_0, n_2), (n_2, n_4)) \quad (7)$$

$$ROR_{A_{pr}}^{A_{pr,r1}^c} = 0.33, \quad ROR_{A_{pr}}^{A_{pr,r2}^c} = 0.66 \quad (8)$$

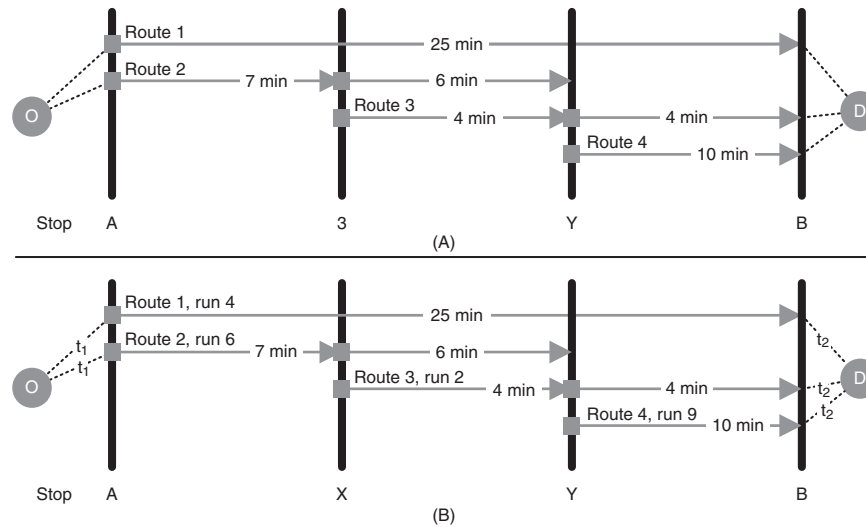
In similar fashion, other routes are evaluated and the final network includes route1, route2 and route5, and these routes are depicted in Fig. 2.

Demand Assignment

The demand assignment is an important component in the design of a public transport network. It helps in estimating: (1) the passenger behavior for route choice in the network, (2) the accumulated demand loads on different routes within the network, and (3) appropriate headways to offer adequate supply for required demand. According to de Dios Ortuzar and Willumsen (2011), there are two basic assignment methods: (1) headway based and (2) schedule based. A headway-based assignment is ideal for strategic

Table 2 Stakeholders perspectives with related performance indicators

Perspectives of stakeholders	Performance indicators								
	Route overlap index	Empty seat hours	Avg num of transfers	Required seating capacity	Unsatisfied demand percentage	Total network length	Travel time	Operating cost	Transport network directness
Passenger	■		■		■		■		■
Operator	■	■		■		■		■	■
Authority	■	■	■	■	■	■	■	■	■

**Figure 3** Route choice for headway and schedule-based assignment *Source: Spiess and Florian (1989).*

planning, as it involves relatively lower computation costs and requires less information for assignment. A timetable-based assignment is ideal when detailed analysis about the service being offered is required, such as actual arrival, departure time, and individual vehicle loads.

Based on the selected assignment method, the next step is to select the most suitable route choice model, in which passenger behavior is modeled based on the path or set of paths the passenger will choose to complete the trip. For the headway-based approach, the concept of optimal strategy is used, meaning that a passenger does not select an exact path, but rather considers a set of attractive paths. **Fig. 3A** shows a small example composed of four routes and four stops (A, X, Y, and B) for the given origin--destination (OD) pair. A typical strategy for a passenger would be that the passenger arrives at stop A from origin O, he/she will consider whichever vehicle comes first from either Route1 or Route2. If a vehicle from Route1 comes first, he/she will exit at stop B to destination D. If a Route2 vehicle comes first, he/she will transfer at stop Y to either Route3 or Route4 and then exit at B to the destination D.

In the schedule-based approach, the passenger has prior spatio-temporal information about the individual service runs. Based on the passenger's selected departure time, different access stops and service runs are available to him/her. In **Fig. 3B**, two different runs for two routes are available from stop A from origin O to the destination D. In this case, based on the access time t_1 to stop A, a passenger can consider taking either Route1 – Run4 or Route2 – Run6. In Route1 – Run4, he/she will exit at B through t_2 to reach destination D. If Route2 – Run6 is selected. Then, based on arrival time of Run6 at stop X and departure time of Route3 – Run2, the passenger can consider Route3 – Run2 at stop X or Y and exit at stop B to reach destination D through t_2 .

Holistic Network Evaluation

The evaluation of public transport networks usually provides a set of metrics for judging their quality from different perspectives. As previously mentioned, there are three main stakeholders in the system: passengers, operators, and authorities. The proposed transport network evaluation component considers all these perspectives and tries to perform quantitative and qualitative analysis of a public transport network. A total of nine indicators representative of all these three stakeholders are used for the evaluation. These three viewpoints with their associated indicators are listed in **Table 2**.

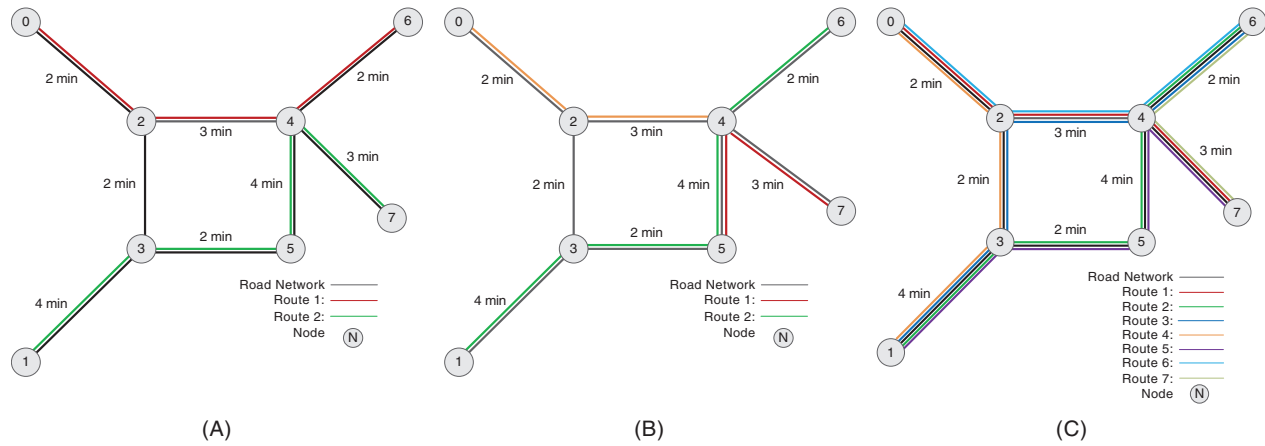


Figure 4 Three different public transport concepts. (A) Transfer-based network. (B) Trunk-&-feeder network. (C) Direct network.

- The route overlap index is adopted from [Vuchic \(2005\)](#), which measures the redundancy in the public transport network.
- Empty-seat hours is adopted from [Ceder \(2016\)](#) and measures the unutilized available seat capacity in the public transport network over a given time period.
- The average number of transfers measures, on average, the number of transfers required to complete the trip.
- The required seating capacity measures the total number of seats required to satisfy the demand for a given time period.
- The unsatisfied demand percentage measures the percentage of demand that is unable to complete the trips.
- The total network length measures the combined length of all the routes in the public transport network.
- The travel time measures the total travel time that all the passengers spend in the public transport network.
- The operating cost measures the cost borne by the operator for offering public transport services.
- The transport network directness measures the directness of the public transport network for all routes compared to the shortest paths.

Example

By using the route overlap concept, different types of public transport networks can be designed. Let's consider the network and the demand matrix depicted in [Fig. 1](#). Three different types of networks (see [Fig. 4](#)) can be created: (1) a transfer-based network, (2) a trunk-&-feeder network, (3) a direct network. A headway-based assignment along with optimal strategy for route choice is used for this example. Moreover, the headways and the number of vehicles required to provide the adequate supply are listed in [Fig. 1C](#).

All three networks have different attributes and functions. The transfer-based network require transfers for most OD pairs to complete the trip. The maximum allowed ROR is set to 10%. The trunk-&-feeder network utilizes the concept of a trunk route with feeder routes, with the maximum allowed ROR set to 25%. The direct network offers direct trips to most OD pairs, with maximum allowed ROR value of 100%. Let's evaluate these three networks based on the selected performance measures mentioned in the previous subsection. The travel-time components of these three networks is depicted in [Fig. 5A](#). One can see that the all three networks depict comparable travel times, with the direct network offering the least travel time. However, if both objectives are considered (see [Fig. 5B](#)), direct network cost almost 50% more than trunk feeder and transfer-based network.

The results from [Fig. 5](#) show that these networks offer trade-offs among selected objectives. Therefore, in order to see a more holistic picture, results from other related performance measures should also be analyzed. The results from selected performance measures are listed in [Fig. 6](#).

- The route overlap index (see [Fig. 6A](#)) shows that the transfer-based network and trunk-&-feeder network offer the least overlap between the routes, whereas the direct network offers three times the overlap of the transfer-based network.
- The empty-seat hours show that the maximum utilization of the offered capacity is attained in the trunk-&-feeder network and the transfer-based network, whereas in the direct network, we have the largest the unutilized capacity.
- The average number of transfers shows that the direct network offers the least number of transfers compared to other two networks (see [Fig. 6C](#)).
- The required seating capacity shows (see [Fig. 6D](#)) that the trunk-&-feeder network and transfer-based networks need the lowest number of seats for required service. This is due to better utilization of routes.
- The unsatisfied demand percentage for all three networks shows that these networks fully satisfy the demand with no OD pair requiring more than two transfers (see [Fig. 6E](#)).
- The total network length suggests that the transfer-based and the trunk-&-feeder networks require one-third of the length of the direct network (see [Fig. 6F](#)).

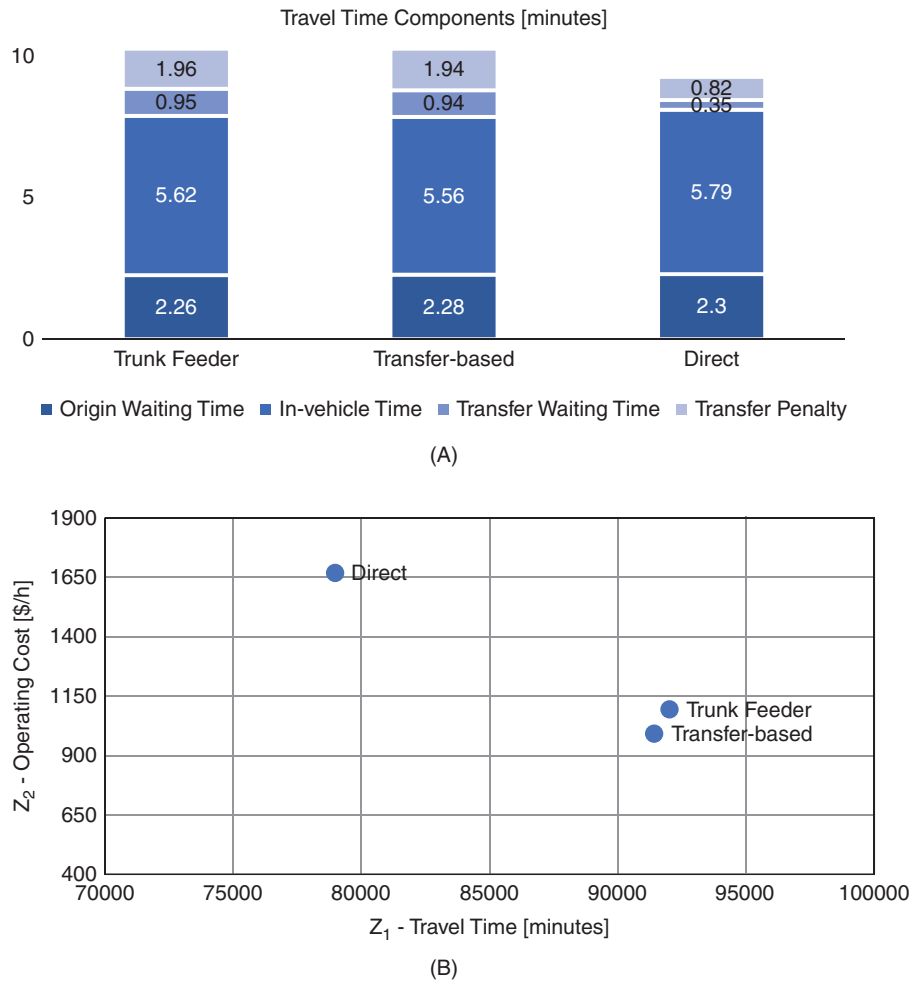


Figure 5 Travel time components and objective values for different network concepts. (A) Travel time components for different public transport network concepts. (B) Objective Z_1 and Z_2 comparison.

- The operating cost shows that the costs of transfer-based and trunk-&-feeder networks are two-third of the direct network (see Fig. 6G).
- The transport network directness shows that all three networks have similar directness values (see Fig. 6H).
- In terms of the required fleet size (Fig. 6I), the transfer-based and trunk-&-feeder networks requires lesser, but higher-capacity models compared to the direct network, which requires a much higher number of low-capacity vehicles.

One can see from the results of these performance measures that the transfer-based and trunk-&-feeder networks offer slightly higher travel times compared to a direct network. However, they are operationally viable, cheaper to operate, simple and easier to manage.

Summary and Conclusion

Public transport networks around the world are regularly tweaked or realigned slightly to attain route- and/or corridor-level enhancements. However, until very recently, it was not common to do major redesigns/restructuring of a whole public transport network. The availability of data and the demand for efficient, effective, operationally viable, and, most importantly, reliable public transport service convinced many public transport authorities to restructure their whole network. This helps in improving network realignment with changes in demand patterns, simplification of the network with high-frequency services, enhancing priority measures, and service reliability.

A recent study conducted by the National Academies of Sciences, Engineering, and Medicine (2019) about bus-network redesign in North America involved a survey about bus-network redesigns among 36 public transport authorities. The survey revealed that the main objectives for bus-network redesigns are: (1) to restructure the network, (2) to increase efficiency and effectiveness, (3) to

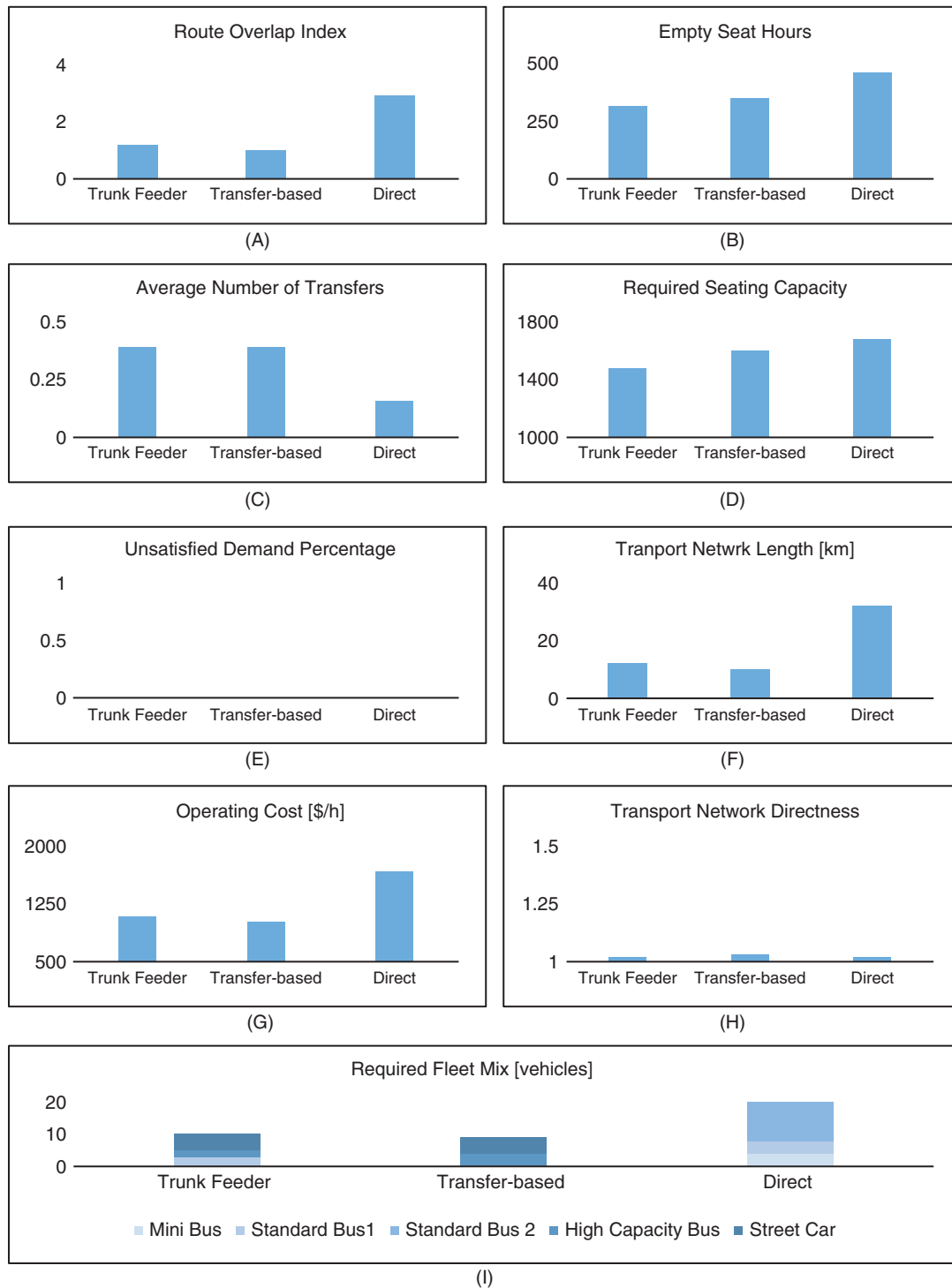


Figure 6 Evaluation of three different public transport network concepts. (A) Route overlap index. (B) Empty-seat hours. (C) Average number of transfers. (D) Required seating capacity. (E) Unsatisfied demand percentage. (F) Total network length in km. (G) Operating cost in \$/hour. (H) Transport network directness. (I) Required fleet size.

introduce the concept of a high-frequency network, (4) to simplify the existing network for ease of use, (5) to improve the reliability of the service, and (6) to increase the ridership of the system.

Among these 36 public transport authorities, one-third of them have completed a bus-network redesign, while more than half of them are either engaged in or currently involved in different phases of bus-network redesign implementation. More than 85% of these agencies have either performed or intend to perform a full redesign of the system. Roughly more than 50% of these agencies

redesign the network from scratch compared to 46%, which redesign the network based on the existing network. In terms of timeframes, for 80% of these authorities, it took at most 2 years for bus-network redesign from planning to implementation.

After the implementation, 70% of these authorities believe that the operating cost increases up to 10%. The evaluation criteria used by these agencies involves area coverage and impact on travel time, cost, transfers, accessibility, and on-time performance. Overall, such a redesign exercise has a positive impact, including increased ridership, improved operational management, and an improved level of service.

The National Academies survey revealed two things: first, that the redesign of public transport networks is inevitable and must therefore be carried out to improve the offered service quality, its efficiency, and reliability; and second, that its holistic evaluation must include perspectives from all three major stakeholders—that is, passengers, authorities, and operators.

At present, most of the work done in redesigning networks is carried out using expert opinion along with geographic information system tools. Network redesign can benefit greatly from adopting an algorithmic approach to assist the transport authorities in charge. Such an approach aids in designing different network configurations together with their evaluations, making the whole process of redesigning transparent, effective, and less time consuming.

Technological innovations such as electric vehicles and autonomous driving, along with the availability of static and dynamic data sources coupled with high computation power, provide new insights into mobility, thus generating new ideas such as mobility as a service, on-demand micro-transit, dynamic routing, and scheduling. These new ideas should be incorporated during the redesign process and/or enhancement of existing public transport networks to see how public transport can benefit with regard to service standards, efficiency, cost reduction, and reliability.

The public transport planning is an extremely complex task with route planning being one of its sub-tasks. It is extremely difficult to encompass every aspect of its planning procedures. Therefore, for a further understating of how the public transport planning process relates to future perspectives and implications of mobility automation some literature is suggested in the list of further reading.

Acknowledgment

We are grateful to TUMCREATE in Singapore to allow us for using some of their data for the analysis of this chapter.

Zain Ul Abedin is a research fellow at TUMCREATE Ltd. His main research area includes public transport network analysis, new service design, and network evaluation. He is a member of the rapid road transport group within TUMCREATE, which is responsible for designing a new public transport concept called dynamic autonomous road transit (DART), an efficient, flexible, modular, road-bound autonomous vehicle-based system. He designed the strategic network concept for DART and is involved in the identification of a suitable corridor for such service design in real networks.

Avishai (Avi) Ceder is a Professor at the Technion—Israel Institute of Technology, and Honorary Professor at the University of Auckland, see <http://ceder.net.technion.ac.il>. He is the founder and director of the Transportation Research Centre (TRC) in Auckland, the head of the Transportation Engineering and Geo-Information Department at the Technion, and the chief scientist of the Israel Ministry of Transport. He is a member of various international symposia (e. g., ISTTT, CASPT). In 2007, he released a book *Public Transit Planning and Operation*, published by Elsevier, Oxford, UK, with its second edition appearing in 2016 by CRC Press, Boca Raton, US, followed by translations into Chinese (2017) and Korean (2018).

References

- Arbex, R.O., da Cunha, C.B., 2015. Efficient transit network design and frequencies setting multi-objective optimization by alternating objective genetic algorithm. *Transport. Res. Part B: Methodol.* 81, 355–376.
- Baaj, M.H., Mahmassani, H.S., 1991. An AI-based approach for transit route system planning and design. *J. Adv. Transport.* 25 (2), 187–209.
- Baaj, M.H., Mahmassani, H.S., 1995. Hybrid route generation heuristic algorithm for the design of transit networks. *Transport. Res. Part C: Emerging Technol.* 3 (1), 31–50.
- Bagloee, S.A., Ceder, A.A., 2011. Transit-network design methodology for actual-size road networks. *Transport. Res. Part B: Methodol.* 45 (10), 1787–1804.
- Board, T.R., National Academies of Sciences and Medicine E., 2019. *Comprehensive Bus Network Redesigns*. The National Academies Press, Washington, DC.
- Buba, A.T., Lee, L.S., 2016. Differential evolution for urban transit routing problem. *J. Comput. Commun.* 4 (14), 11.
- Bussieck, M., 1998. *Optimal lines in public rail transport* (Doctoral dissertation, Verlag nicht ermittelbar).
- Cancela, H., Mauttone, A., Urquhart, M.E., 2015. Mathematical programming formulations for transit network design. *Transport. Res. Part B: Methodol.* 77, 17–37.
- Carrese, S., Gori, S., 2002. An urban bus network design procedure. In: *Transportation planning*. Springer, Boston, M.A., pp. 177–195.
- Ceder, A., 2016. *Public Transit Planning and Operation: Modeling, Practice and Behavior*. CRC Press.
- Ceder, A., Wilson, N.H., 1986. Bus network design. *Transport. Res. Part B: Methodol.* 20 (4), 331–344.
- Chakroborty, P., Wivedi, T., 2002. Optimal route network design for transit systems using genetic algorithms. *Eng. Optimiz.* 34 (1), 83–100.
- Chew, J.S.C., Lee, L.S., Seow, H.V., 2013. Genetic algorithm for biobjective urban transit routing problem. *J. Appl. Math.* 2013.
- Chu, J.C., 2018. Mixed-integer programming model and branch-and-price-and-cut algorithm for urban bus network design and timetabling. *Transport. Res. Part B: Methodol.* 108, 188–216.
- Cipriani, E., Gori, S., Petrelli, M., 2012. Transit network design: A procedure and an application to a large urban area. *Transport. Res. Part C: Emerging Technol.* 20 (1), 3–14.
- Constantin, I., Florian, M., 1995. Optimizing frequencies in a transit network: a nonlinear bi-level programming approach. *Int. Trans. Oper. Res.* 2 (2), 149–164.
- Council, A.T., 2006. *National guidelines for transport system management in Australia*. Background material 5.
- de Dios Ortuzar, J., Willumsen, L.G., 2011. *Modelling Transport*. John Wiley & Sons.
- Fan, W., Machemehl, R.B., 2006. Optimal transit route network design problem with variable transit demand: genetic algorithm approach. *J. Transport. Eng.* 132 (1), 40–51.

- Fan, L., Mumford, C.L., 2008. A metaheuristic approach to the urban transit routing problem. *J. Heuristics* 16 (3), 353–372.
- Fan, L., Mumford, C.L., Evans, D., 2009, May. A simple multi-objective optimization algorithm for the urban transit routing problem. In *2009 IEEE Congress on Evolutionary Computation* (pp. 1–7). IEEE.
- Farahani, R.Z., Miandoabchi, E., Szeto, W.Y., Rashidi, H., 2013. A review of urban transportation network design problems. *Eur. J. Oper. Res.* 229 (2), 281–302.
- Furth, P.G., Wilson, N.H., 1981. Setting frequencies on bus routes: Theory and practice. *Transport. Res. Rec.* 818 (1981), 1–7.
- Guan, J.F., Yang, H., Wirasinghe, S.C., 2006. Simultaneous optimization of transit line configuration and passenger line assignment. *Transport. Res. Part B: Methodol.* 40 (10), 885–902.
- Guihaire, V., Hao, J.K., 2008. Transit network design and scheduling: A global review. *Transport. Res. Part A: Policy Pract.* 42 (10), 1251–1273.
- Hosapujari, A.B., Verma, A., 2013. Development of a hub and spoke model for bus transit route network design. *Procedia-Social Behav. Sci.* 104, 835–844.
- Huang, D., Liu, Z., Fu, X., Blythe, P.T., 2018. Multimodal transit network design in a hub-and-spoke network framework. *Transport. A: Transport Sci.* 14 (8), 706–735.
- Ibarra-Rojas, O.J., Delgado, F., Giesen, R., Muñoz, J.C., 2015. Planning, operation, and control of bus transport systems: A literature review. *Transport. Res. Part B: Methodol.* 77, 38–75.
- Israeli, Y., Ceder, A., 1989, September. Designing transit routes at the network level. In *Conference Record of papers presented at the First Vehicle Navigation and Information Systems Conference (VNIS'89)* (pp. 310–316). IEEE.
- Kechagiopoulos, P.N., Beligiannis, G.N., 2014. Solving the urban transit routing problem using a particle swarm optimization based algorithm. *Appl. Soft Comput.* 21, 654–676.
- Kepaptsoglou, K., Karlaftis, M., 2009. Transit route network design problem: review. *J. Transp. Eng.* 135 (8), 491–505.
- Kılıç, F., Gök, M., 2014. A demand based route generation algorithm for public transit network design. *Comput. Oper. Res.* 51, 21–29.
- Kim, H.S., Kim, D.K., Kho, S.Y., Lee, Y.G., 2016. Integrated decision model of mode, line, and frequency for a new transit line to improve the performance of the transportation network. *KSCE J. Civil Eng.* 20 (1), 393–400.
- Kuo, Y., 2013. Design method using hybrid of line-type and circular-type routes for transit network system optimization. *Top* 22 (2), 600–613.
- Lampkin, W., Saalmans, P.D., 1967. The design of routes, service frequencies, and schedules for a municipal bus undertaking: A case study. *J. Oper. Res. Soc.* 18 (4), 375–397.
- Lee, Y.J., Vuchic, V.R., 2005. Transit network design with variable demand. *J. Transport. Eng.* 131 (1), 1–10.
- Majima, T., Takadama, K., Watanabe, D., Katuhara, M., 2014, December. Application of community detection method to generating public transport network. In *Proceedings of the 8th International Conference on Bioinspired Information and Communications Technologies* (pp. 243–250). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).
- Majima, T., Takadama, K., Watanabe, D., Katuhara, M., 2015. Characteristic of passenger's route selection and generation of public transport network. *SICE J. Control Measure Syst. Integr.* 8 (1), 67–73.
- Mandl, C.E., 1980. Evaluation and optimization of urban public transportation networks. *Eur. J. Oper. Res.* 5 (6), 396–404.
- Mumford, C.L., 2013, June. New heuristic and evolutionary operators for the multi-objective urban transit routing problem. In *2013 IEEE Congress on Evolutionary Computation* (pp. 939–946). IEEE.
- Nayem, M.A., Rahman, M.K., Rahman, M.S., 2014. Transit network design by genetic algorithm with elitism. *Transport. Res. Part C: Emerg. Technol.* 46, 30–45.
- Ngamchai, S., Lovell, D.J., 2003. Optimal time transfer in bus transit route network design using a genetic algorithm. *J. Transport. Eng.* 129 (5), 510–521.
- Nikolić, M., Teodorović, D., 2013. Transit network design by bee colony optimization. *Expert Syst. Appl.* 40 (15), 5945–5955.
- Pattnaik, S.B., Mohan, S., Tom, V.M., 1998. Urban bus transit route network design using genetic algorithm. *J. Transport. Eng.* 124 (4), 368–375.
- Petrelli, M., 2004. A transit network design model for urban areas. *WIT Transactions on The Built Environment*, 75.
- Rea, J.C., 1971. Designing Urban Transit System: An Approach to the Route Technology Selection Problem, Research Report no 6.
- Schbel, A., Scholl, S., 2005. Line planning with minimal traveling time. Kroon LG, Mhring RH (eds).
- Schéele, S., 1980. A supply model for public transit services. *Transport. Res. Part B: Methodol.* 14 (1–2), 133–146.
- Schöbel, A., 2012. Line planning in public transportation: models and methods. *OR Spectrum* 34 (3), 491–510.
- Shih, M.C., Mahmassani, H.S., 1994. A design methodology for bus transit networks with coordinated operations (No. SWUTC/94/60016-1).
- Silman, L.A., Barzily, Z., Passy, U., 1974. Planning the route system for urban buses. *Comput. Oper. Res.* 1 (2), 201–211.
- Spies, H., Florian, M., 1989. Optimal strategies: a new assignment model for transit networks. *Transport. Res. Part B: Methodol.* 23 (2), 83–102.
- Szeto, W.Y., Wu, Y., 2011. A simultaneous bus route design and frequency setting problem for Tin Shui Wai, Hong Kong. *Eur. J. Oper. Res.* 209 (2), 141–155.
- Tom, V.M., Mohan, S., 2003. Transit route network design using frequency coded genetic algorithm. *J. Transport. Eng.* 129 (2), 186–195.
- Ul Abedin, Z., 2019. A Methodology to Design Multimodal Public Transit Networks: Procedures and Applications, Technische Universität München.
- Ul Abedin, Z., Busch, F., Wang, D.Z., Rau, A., Du, B., 2018. Comparison of public transport network design methodologies using solution-quality evaluation. *J. Transport. Eng. Part A: Syst.* 144 (8), 04018036–4018113.
- van Nes, R., Hamerslag, R., Immers, L.H., 1988. The design of public transport networks (Vol. 1202). National Research Council, Transportation Research Board.
- Vuchic, V.R., 2005. *Urban Transit: Operations, Planning, and Economics*. John Wiley & Sons.
- Walker, J., 2012. *Human Transit: How Clearer Thinking about Public Transit can Enrich Our Communities and Our Lives*. Island Press.
- Wan, K.K., Lo, H.K., 2003. A mixed integer formulation for multiple-route transit network design. *J. Math. Model. Algorithms* 2 (4), 299–308.
- Yen, J.Y., 1971. Finding the k shortest loopless paths in a network. *Manage. Sci.* 17 (11), 712–716.
- Zhang, J., Lu, H., Fan, L., 2010, August. The multi-objective optimization algorithm to a simple model of urban transit routing problem. In *2010 Sixth International Conference on Natural Computation* (Vol. 6, pp. 2812–2815). IEEE.
- Zhao, H., Jiang, R., 2015. The Memetic algorithm for the optimization of urban transit network. *Expert Syst. Appl.* 42 (7), 3760–3773.
- Zhao, F., Zeng, X., 2006. Optimization of transit network layout and headway with a combined genetic algorithm and simulated annealing method. *Eng. Optimiz.* 38 (6), 701–722.

Further Reading

- Antoniou, C., Dimitriou, L., Pereira, F.C., 2019. Big Data and Transport Analytics: An Introduction. In: *Mobility Patterns, Big Data and Transport Analytics*. Elsevier.
- Cohen, T., Cavoli, C., 2019. Automated vehicles: exploring possible consequences of government (non) intervention for congestion and accessibility. *Transport Rev.* 39 (1), 129–151.
- Feigon, S., Murphy, C., 2018. Broadening understanding of the interplay between public transit, shared mobility, and personal automobiles. National Academies Press.
- Haboucha, C.J., Ishaq, R., Shifan, Y., 2017. User preferences regarding autonomous vehicles. *Transport. Res. Part C: Emerging Technol.* 78, 37–49.
- Hensher, D.A., Button, K.J. (Eds.), 2007. *Handbook of Transport Modelling*. Emerald Group Publishing Limited.
- Lam, W.H., Bell, M.G. (Eds.), 2002. *Advanced Modeling for Transit Operations and Service Planning*. Emerald Group Publishing Limited.
- Mees, P., 2009. *Transport for Suburbia: Beyond the Automobile Age*. Routledge.
- Neilsen, G., Nelson, J.D., Mulley, C., Tegner, G., Lind, G., Lange, T., 2005. *HiTrans Best Practice Guide 2: Public Transport—Planning the Networks*. HiTrans, Stavanger.
- Singleton, P.A., 2019. Discussing the “positive utilities” of autonomous vehicles: will travellers really use their time productively? *Transport Rev.* 39 (1), 50–65.
- Strömberg, H., Karlsson, I.M., Sochor, J., 2018. Inviting travelers to the smorgasbord of sustainable urban transport: evidence from a MaaS field trial. *Transportation* 45 (6), 1655–1670.
- Vazifeh, M.M., Santi, P., Resta, G., Strogatz, S.H., Ratti, C., 2018. Addressing the minimum fleet problem in on-demand urban mobility. *Nature* 557 (7706), 534–538.
- Vuchic, V.R., 2007. *Urban Transit Systems and Technology*. John Wiley & Sons.

Space Transportation

Mark Hemsell, Hemsell Astronautics Limited, Herts, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	638
Orbits and Sub-Orbits	638
Modern Launchers	640
Impact of Small Satellites	641
Economics	642
Reusable Systems	642
Inter-Orbit Transport	644
Summary and Conclusions	645
Biography	645
References	645
Further Reading	645

Introduction

All transportation is about the controlled movement of a vehicle so it can travel to an intended destination. This means applying forces on that vehicle to achieve the controlled motion and that in turn means applying Newton's third law of motion. In other words, the vehicle must push against something in order to move. All terrestrial transport, such as cars and trains, push against the solid earth, at sea, propeller and paddle driven ships push against the water, and aircraft engines push against the air. The problem for Space vehicles is that for practical purposes space is empty—there is nothing to push against. The only solution to this problem is the vehicle must carry on board the mass to react against, and any vehicle that carries its own reaction mass (called propellant) is a rocket. So Space transportation is fundamentally about rockets.

To apply force, a rocket vehicle expels some of its propellant with a velocity c , called the exhaust velocity. When this happens the mass of the rocket vehicle, M , is reduced by the propellant expelled, dm . Then, by conservation of momentum, the velocity change to the vehicle dV can be found from:

$$dV = \frac{c \, dm}{M}$$

If many of these dm -sized packages of propellant are expelled, then the total velocity V can be achieved by integrating the following equation:

$$\int_0^V dV = c \int_{M_s}^{M_e} \frac{dm}{m}$$

Where M_s is the mass, then the rocket starts operating, and M_e is the mass when it finishes. Solving this integral gives:

$$V = c \ln \left(\frac{M_s}{M_e} \right)$$

This is known as the “rocket equation” and it shows the two principle drivers of astronautical engineering, the rocket engine's exhaust velocity and the fraction of the vehicle that can be propellant. For chemical rockets, the exhaust velocity varies around 3 km/s for the best solid propellant rockets where the chemicals are stored and then burnt in a tube (like a firework rocket) to 4.5 km/s for an engine using liquid hydrogen and liquid oxygen. The Space Shuttle (**Fig. 1**) had both types of engine—two solid fuel boosters that are thrown away after 2 minutes of flight, and a cluster of three main liquid engines that powered the Shuttle's Orbiter into Space.

This chapter first considers the history of rockets moving from sub-orbital to orbital flight and then outlines modern launch systems and the missions they undertake. It then discusses the impact on satellite design and then a section on the economic drivers leading to a section on the introduction of reusable vehicles. Finally, the paper considers the systems used for inter-orbit transportation.

Orbits and Sub-Orbits

Although there is no clearly defined place where the atmosphere stops and Space begins (called the Karman line), it is generally taken to be 100 km altitude. To reach it with a rocket will require a velocity of about 1.4 km/s, a feat first achieved by the German A4, a



Figure 1 The Space Shuttle at launch with three liquid propellant engines on the core and two side mounted solid propellant boosters. *Source: NASA.*



Figure 2 A V2 rocket take off. *Source: https://commons.wikimedia.org/wiki/File:The_V2_Rocket_BU11149.jpg*

prototype of the V2 missile, during its test program. This had a mass of 12.8 tons of which 8.7 tons was propellant, the alcohol/oxygen engines had an exhaust velocity a little over 2 km/s poor by today's standards (Riedel, 2005). The rocket equation predicts that this should give V2 a speed of 2.3 km/s but in practice, air resistance and counteracting gravity reduced its actual speed to 1.6 km/s. But this is enough to reach a maximum altitude of 175 km—well into Space (Fig. 2).

Gaining enough speed to get over the Karman line is a class of flight called “sub-orbital.” Any missile with a range of a few hundred kilometres can do this, but it has also been used by many scientific rockets (called sounding rockets) and is the objective of several companies who are offering tourist space flights, such as Virgin Galactic and Blue Origin. The problem with such trips is that they are brief. To quote an old adage “what goes up, must come down,” and anything on a sub-orbital trajectory will have only a few

minutes in Space before it re-enters the atmosphere, and if it does not wish to crash into the ground, it must now take out over 1 km/s velocity within the height of the atmosphere; leading to high deceleration forces of several gravities.

However, that old adage about the inevitable fall is not quite right. If an object can both reach Space, which we have seen requires over 1 km/s vertically upward, AND attain a velocity of over 7.8 km/s in the horizontal direction, then it enters the orbit and does not fall back to the ground. That means the rocket has to provide about 5 times the velocity and 25 times the energy than required for sub-orbital flight. So, it is much more difficult to achieve but if you can achieve it then you are permanently in Space. This is the basic goal of a Space transportation system—to reach orbit and then move between orbits.

The first rocket to achieve this feat was the Russian R-7 Semyorka, which launched Sputnik 1 on 4th October 1957. Like the V2, it was born out of a military program. The R-7 was developed as an intercontinental missile with a nuclear warhead and the ability to reach orbit was a serendipitous consequence of this requirement. Sixty years later this rocket, upgraded and renamed “Soyuz,” remains one of the world’s main launch systems, capable of putting around seven tons into low Earth orbit.

Modern Launchers

The Soyuz launcher (**Fig. 3**) follows the basic approach used by all launch systems in that it has stages; two or three rockets are stacked on top of each other and then fired in sequence. Once air resistance and supporting the rocket against gravity is accounted for, the velocity that should be put in the rocket equation for an Earth orbit launcher is around 9.4 km/s. Even if the propellants are hydrogen and oxygen, this means that 89% of the take-off weight of the launch vehicle needs to be propellant. With today’s technology, it is not possible to make practical rockets with meaningful payloads that can achieve this mass ratio. However, if we have two-staged rockets, they only need to reach 4.7 km/s each and then these stages only need to be 75% propellant, which can be easily achieved with rocket technology from the 1950s.

Currently, there are 30 operational launchers around the world and all are staged systems ([Stakem, 2017](#)). However, unlike the very early launch systems, which were mostly based on military missiles, today’s launchers are nearly all specialized developments to reach Space. There are several drivers creating this divergence. As technology progressed in the 1960s, the differing requirements for a



Figure 3 The Soyuz launcher which is a modernized version of the rocket that carried the world's first satellite, Sputnik 1. *Source: Arianespace.*

military missile and a launch system meant the optimum technical approach for each significantly changed. Another difference was the need for larger launch systems to meet 1960s space race ambitions, while at the same time missiles could be made smaller as warhead masses reduced.

There was another more economic factor that drove the size of launch systems up. The dominant factor that determines the cost per kilogram into orbit with expendable rockets is their size. If you double the size of the rocket, you can get a better ratio of propellant to dry mass (so more kilograms of payload) and while the costs such as development, build and operation rise as you would expect, none of them double. The end result is that the cost per kilogram drops dramatically, as launch systems get bigger. This is the economic reasoning behind “launch share” where two or more satellites are launched together. An approach that has been most successfully employed by the European launch system operator, Arianespace. Their launch systems Ariane 4 and Ariane 5 were both sized to carry two satellites at once which gave them a significant competitive advantage and they became the dominant system for launching satellites intended for geostationary orbit, managing to capture half the available market.

Geostationary orbit is a circular orbit in the Earth’s equatorial plane at a high altitude of 35,786 km. At this height, the period of the orbit exactly matches the 24-hour spin of the Earth and so the satellite appears to hover in the sky, which is why domestic TV satellite dishes can be pointed at them. The value of this location for communications was first pointed out by Arthur C Clarke in 1945, and today there are several hundred satellites typically with a mass of a ton or two. This orbit is so crowded that some satellites are separated by less than 100 km.

Another popular class of orbit are polar orbits; that is where the orbit flies over the Earth’s poles. These are valuable for observation of the Earth’s surface, as any satellite in these orbits will fly over the Earth’s entire surface. Many applications use Sun-synchronous orbits; these orbits precess to match the Sun’s passage over the year so that they always fly over a region at the same time of day, so images taken over time can more easily detect changes on the ground.

Third important orbit location are circular Earth orbits around 20,000 km. These are used by navigation satellites transmitting the signals that enable our phones and cars to know precisely where we are on the Earth’s surface.

So most of the major applications we have found for Space require special orbits and all are significantly more difficult (i.e., require more velocity from the rocket) than simply getting into any low altitude orbit a few hundred kilometers above the Earth’s surface. This is another point of divergence between launchers and military missiles; not only are the payloads typically heavier but they need to be accelerated to much higher velocities. It is this specialization trend that is exemplified by the Ariane system being optimized for a particular orbit, the one that is most attractive from a commercial perspective.

Impact of Small Satellites

An interesting development in recent times has been the impact of the miniaturization of electronics on satellites. In geostationary orbit, the size of satellites has remained roughly the same because they are driven by the large antennas and solar arrays needed to communicate with the Earth, so the technology improvements have enabled an enormous increase in the number of communications links and the amount of on-board switching between the channels. However in other orbits this technical trend has dramatically reduced the size of satellites that are needed to do a particular job.

This “small satellite” approach became a definable movement, which took off in the 1990s as a way to reduce the cost of Space activity. The use standard commercial components, not waiting for the years it takes to get Space qualified versions, was successfully pioneered by companies like Surrey Satellite Technologies. At the end of the 1990s this trend led to a specific form of small satellite—the “Cubesat” format (Cal Poly, 2014). The Cubesat (Fig. 4) was the invention of an academic team led by the California Polytechnic State University. It laid out a specification for a standard 10 cm cube weighing no more than 1.33 kg, with double sized and triple size options. Originally, Cubesats were intended as a route for academic institutions to explore Space technology at low cost, but as the potential of modern electronics housed in this standard packaging became apparent they became commonly used for commercial applications. This standard has been so successful at making Space access to small users that currently the majority of satellites launched are Cubesats of one form or another.

Originally, the economics of small satellites were based on the principle that they could share a launch with a larger payload; using any spare mass capacity that was available. This meant that the satellite was not paying anything like the true launch cost, just something to pay for the trouble of being integrated on to the rocket, and in earlier times these launches were often free. The downside was that the small satellite had to accept the same orbit and launch schedule as the primary payload, often restricting their application to technology demonstrations.

As the importance of the small satellite increased, the compromises of the shared launched approach have become increasingly problematic. So, there has been increasing attention on developing a commercially viable small satellite launch system with a payload capability that matches the satellites that want launching. Many early launch systems, such as America’s Scout and the original Delta, the French Diamant, the Japanese Lambda 4S, and the British Black Arrow, had payloads in the tens to low hundreds of kilograms—perfect for the small satellite market. But all proved unsustainable, as scale economies caused much higher cost per kilogram than large systems, and by the mid-1970s all of these had been abandoned except Delta, which had grown to launch much larger payloads of over a ton.

So the economics of satellites and their launchers pull in opposite directions, satellites want to get smaller to be more economic and launchers want to get bigger to be more cost effective. This is not a problem that has yet been completely resolved. As an example, SpaceX developed Falcon 1 which had a payload of a few hundred kilograms but this was retired after only one commercial launch,

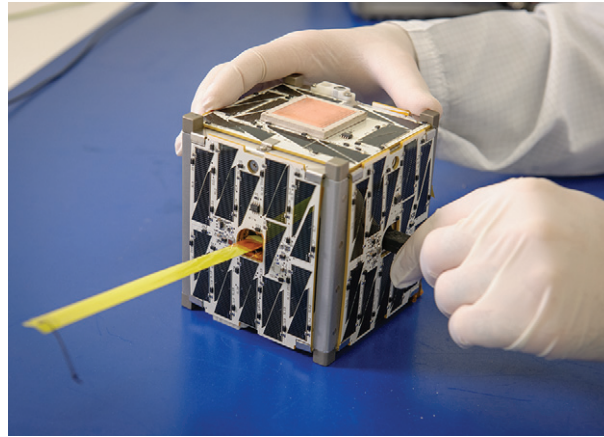


Figure 4 PhoneSat an example of a Cubesat. *Source: NASA Ames.*

as the price moved to reflect the real costs and customer interest dried up, so the company concentrated on the much larger Falcon 9. But the allure of an economically viable service for small satellites is still strong and several ventures persist in attempts to achieve this “holy grail.”

Economics

The apparent dominance of commercial considerations in launch systems can be misleading. With the divergence of military missiles and launch systems, launch systems could be solely civilian enterprises, with the proviso that the military also want to launch satellites with similar requirements to civil satellites and so they become a significant part of the user base. So in legal terms launch systems are “dual use” in that they have both civil and military roles. As a result, while many launch systems are the result of commercial developments even these are largely dominated by their interaction with governments to the point that there is effective market failure in the launch business.

There are about a hundred launches annually, so on average each launch system conducts around three launches a year. So, most of these systems are greatly underutilized, and owe their existence to government funding for both development and operations. The governments’ investments are for strategic reasons, including a desire to be able to launch military satellites, a need to secure guaranteed access to Space for a countries’ civilian satellites, and as a demonstration of the nation’s technical prowess. However, as a consequence there is never an intention of recovering all the national investment money through the fees charged on the open market for launches. Indeed most launchers are operated with some sort of subsidy, so not even their operating cost are covered by the fees charged for commercial launches.

In this context “value for money” for the governments footing the bill means the smallest development cost, and expendable launch systems are much cheaper to develop. So almost all current launchers are still like missiles, one shot systems that are used just once and thrown away. Given that they cost almost as much as a civil airliner to manufacture, it is hardly a surprise that the cost of reaching even the easiest orbits is around \$10,000 per kg. However, high costs are not the only penalty when using expendable rockets. Because launchers cannot be flight-tested, they have a failure rate that would be unacceptably appalling in any other context. Even the best launch systems fail between one in 70 and 1 in 90 flights, and every failure is a crash and the payload is lost. Another major problem with expendable launchers is availability. As each launch vehicle as to be made especially for each payload, it has to be ordered several years in advance.

Reusable Systems

Cost, unreliability, and unavailability are the trio of constraints holding back the development of space activity and they are inherent in basing the launch infrastructure on expendable rockets. The obvious resolution of all three is to use reusable launch systems. A reusable vehicle should operate like an aircraft, it waits in a hanger, ready for launches at short notice, and because it is flown many times the cost of each flight comes down by an order of magnitude, and the vehicle is more reliable.

The reason that expendables still dominate the launch system market place is the high acquisition cost (development and build) of a reusable system which has to be recovered from the restricted market for launches, which is rather a “chicken and egg” problem as the market is restricted because of the problems with expendables. There are two ventures, SpaceX (**Fig. 5**) and Blue Origin, who are developing and operating systems with reusable first stages, with plans to evolve to fully reusable systems later. However, it is significant that both are privately funded by very high net worth individuals with a very long-term outlook and are not at the mercy of government funding corridors, or a commercial demand for rapid return on investment ([Davenport, 2018](#)).



Figure 5 Two SpaceX Falcon Heavy booster rockets return to their launch site. *Source: SpaceX.*

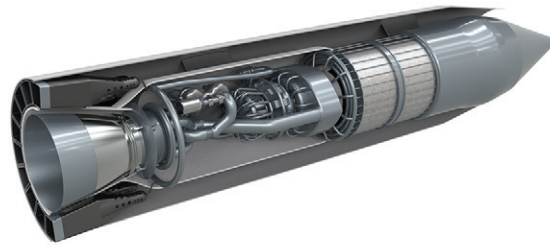


Figure 6 SABRE air-breathing rocket engine. *Source: Reaction Engines Limited.*

Reusable launchers cost more to develop, in part because they require additional systems like heat shielding, a means to fly back and landing gear, but also because there is a much bigger development testing program. For a two-stage vehicle, the cost is not affected by the technology employed, as the requirements are not very exacting and the technology readiness level achieved by aircraft and expendable launchers in the 1960s is more than adequate.

However, ideally what is needed is a single stage to orbit reusable launcher. As we have seen, this is on the very boarder line of possibility for expendable rockets let alone a vehicle with all the extras needed for reusability. A few ventures have seriously tried this approach but none has succeeded. However, the ideal of a reusable single stage to orbit may be possible if an air breathing engine can be used while the system is rising through the atmosphere. Air-breathing engines consume much less propellant to achieve a given thrust and could add enough performance to get a rocket from the ground to orbit.

But air breathing is not magic; the air the engine uses creates a drag as it matches the speed of the engine. This is called “momentum drag” and for practical purposes as the speed reaches somewhere between Mach 5 (5 times the speed of sound) and Mach 6 the force required to collect the air is comparable to the thrust of the engine and the system cannot accelerate further. The system is likely to be flying at over 20 km altitude when this speed is reached so in terms of reaching orbit it has done about 20% of the job, leaving the remaining 80% to be done using pure rockets. While this makes a significant difference, the weight of the air breathing machinery now needs to be added to the dry mass, and generally air-breathing engines are heavy for the thrust they produce. This means the engines used by aircraft do not work in the context of single stage to orbit vehicles and it requires a special lightweight engine which can provide the performance boost while in the atmosphere, but at the same time not cripple the mass ratio for the rocket powered phase of the flight.

The most mature example of this approach is the SABRE engine (Fig. 6) being developed by Reaction Engines in the United Kingdom (Varvill and Bond, 2003). This uses a pre-cooler to take heat from the air so that it can be compressed and fed into a high-pressure rocket combustion chamber; the heat is transferred into a helium loop that uses the energy to power the pumps and compressors. One elegant feature of this engine is that once the air-breathing limits have been reached, the engine can become a pure rocket engine, so the vehicle does not require two separate engines.

An alternative air-breathing approach is to reduce the momentum drag by not stopping the air as it travels through the engine. These are called scramjets—a contraction of supersonic combustion ramjet—because the air stays supersonic while the fuel is burned and the exhaust accelerated. They have the attraction that they should be able to travel way beyond Mach 6, but their problem is they

do not work below Mach 5 meaning another engine is needed to reach that speed. A further problem is they are heavy and large and so they need to be able to reach very high speeds (Mach 16 or more) to compensate for this performance disadvantage. There have been some experimental test vehicles to explore this engine type but they have only had limited success and currently the technology falls way short of what is required for a single stage to orbit vehicle.

Inter-Orbit Transport

Although it was over 60 years ago that mankind first acquired the ability to reach orbit, the fact that the transport system that achieved it is still a viable launch system, indeed arguably the best launch system, suggests the field has not made much progress since. Reusable systems once operational have the promise to greatly improve the way we reach orbit, and then the main area where Space transportation technology will need to advance, are the systems that operate from low Earth orbit to higher orbits, the Moon, and the planets beyond. These missions require velocities approaching what is required to reach low Earth orbit in the first place, for example, to deliver a satellite to geostationary orbit or to a lunar orbit requires around 4 km/s and to reach Mars orbit requires 6.6 km/s. At the moment around 3 km/s of this is supplied by the launch system, which puts the payload directly into a transfer orbit, bypassing low Earth orbit. In terms of energy, this is slightly more efficient but it still requires a typical satellite to provide between 1 and 2 km/s with their internal propulsion for the final orbit insertion maneuvers.

However, once in Space new forms of propulsion are possible with much higher exhaust velocities. These technologies do not use chemical energy sources and could offer a more efficient way of traveling in space, a game changer that could be as significant as the introduction of reusable launchers. But, if the chemical energy in the propellants is not providing the energy, the key problem becomes how the energy to be generated is?

The power required by a rocket engine is determined by its exhaust velocity and thrust by the simple relationship:

$$\text{Power} = \frac{1}{2} \times \text{Thrust} \times \text{Exhaust velocity}$$

Each of the Space Shuttle's main engines has a power of 5 GWatts and, when combined with the solid propellant boosters, the total power of the Shuttle on launch was around 50 GWatts, which is considerably more than the average power consumption of the United Kingdom. This is the key reason non-chemical propulsion is most suited for use in space. To take off from the Earth, the rocket needs to have more thrust than the weight of the launch system, but if we raise the exhaust velocity to get the same thrust, the power required also rises and that must come from an on-board system that delivers many power stations worth of watts. However, once in orbit a rocket no longer needs to overcome gravity and can move around with the tiniest of thrusts. So, here it is possible to have very high exhaust velocities with realistic power levels, but with miniscule thrusts.

The most mature advanced rocket engines are in a class generically known as electric propulsion. The most widely used of these engines are "ion thrusters" which ionize the propellant and then electrically or electromagnetically accelerate the ions. These engines have been successfully used on both planetary missions and geostationary satellites. An example is the QinetiQ T6 ([Fig. 7](#)), which has an exhaust velocity around 40 km/s, but a thrust of only a fifth of a Newton ([Lewis et al., 2015](#)). It needs around 4 kW of power, which is consistent with the solar electric arrays used by satellites. This engine was used in the propulsion stage of the BepiColumbo mission to the planet Mercury, and similar engines are increasingly common on geostationary satellites to perform the final orbit insertion and then keep it on its orbital station.

There are downsides to the very low thrust. The transfer trajectories they follow typically double the velocity requirements in the rocket equation, but this of course is more than compensated for by the order of magnitude improvement in the exhaust velocity. A more important downside is the increase in travel time, journeys that take hours or days with high thrust chemical rockets, take months or years with electric propulsion. While this can be tolerated for robotic satellites and probes, it is a serious problem for human spaceflight. To get travel times down to something appropriate to human mission to geostationary orbit, or the Moon or

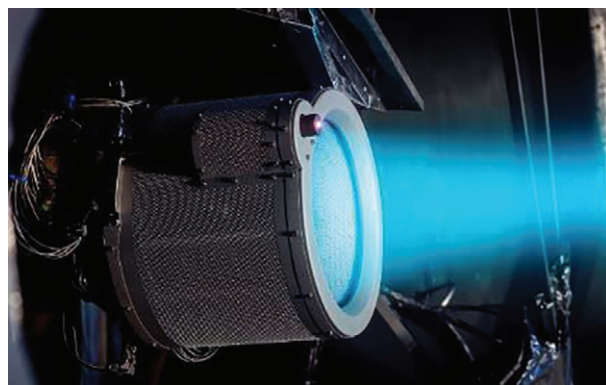


Figure 7 A QinetiQ T6 ion engine firing. *Source: QinetiQ Plc.*

Mars will require high thrust and that means high power, GWatts, coming from generators that weigh only a few tons and, with our current knowledge of physics that means nuclear power.

In the 1970s, NASA came very close to completing the development of a nuclear fission engine called NERVA (Robbins and Finger, 1991). This had a thrust of around 30 tons and an exhaust velocity of 8.3 km/s (a power level of 1.3 GWatts) and was intended to be the primary means by which American astronauts would reach the Moon and Mars. This engine had a simple cycle; hydrogen propellant was heated by passing it over the hot nuclear core. Using existing technology in more complex cycles involving further heating of the propellant by electric arc-jets (also powered by the nuclear reactor) could raise the exhaust velocity to over 12 km/s (Bond, 1972). Beyond that, engines using nuclear fusion power, while outside our current abilities, could in theory provide high thrust with exhaust velocities up to 100,000 km/s (Sutton, 1975).

Summary and Conclusions

Currently, Space transportation is still at its most primitive stage; dominated by expendable chemical rockets giving a restricted, high cost and unreliable means to reach orbit. Despite these issues, a viable Space industry has evolved to explore and exploit the Space environment. However, with technology we have had for many decades, we can improve the Space transportation infrastructure, through reusability and advanced non-chemical rocket engines. These technologies are slowly being introduced and, when they take their full effect, the impact on the size and scope of the Space industry is likely to be considerable.

Biography



Mark Hemsell worked for 13 years as a spacecraft systems engineer at British Aerospace in Stevenage, then for 17 years at the University of Bristol lecturing in Astronautics. In 2009, he joined Reaction Engines Limited as the Future Programs Director responsible for the SKYLON airframe. In 2013, he left Reaction Engines to form Hemsell Astronautics Limited to further his interest in space systems engineering. He is a Fellow and past President of the British Interplanetary Society twice, and a former editor of the Society's Journal.

References

- Bond, A., 1972. A nuclear rocket for the space tug. *J. Br. Interplan. Soc.* 25, 625–641.
- Cal Poly, 2014. CubeSat Design Specification REV 13.
- Davenport, C., 2018. *The Space Barons*. PublicAffairs, New York.
- Lewis, R.A., Pérez-Luna, J., Coombs N., 2015. Qualification of the T6 Thruster for BepiColombo.
- Riedel, W.H.J., 2005. *Rocket Development with Liquid Propellants*. Rolls Royce Heritage Trust.
- Robbins W.H., Finger, H.B., 1991. An Historical Perspective of the NERVA Nuclear Rocket Engine Technology Programme, NASA Contractor Report 187152.
- Stakem, P., 2017. *Space Launch Vehicles*, Amazon.
- Sutton, G.P., Ross, D.M., 1975. *Rocket Propulsion Elements*, fourth ed. John Wiley, New York.
- Varvill, R., Bond, A., 2003. A comparison of propulsion concepts for SSTO reusable launchers. *J. Br. Interplan. Soc.* 56, 108–117.

Further Reading

- Baker, D., 1978. *The Rocket*. New Cavendish Books, London.
- Harland, D.M., 1998. *The Space Shuttle; Roles Missions and Accomplishments*. John Wiley and Sons with Praxis Publishing, Chichester.
- McElyea, T., 2003. *A Vision of Future Space Transportation*. Apogee Books, Ontario.
- Pappalardo, J., 2017. *Spaceport Earth*. Abrams Press, New York.
- Lewis, R.A., Pérez-Luna, J., Coombs N., 2015. Presented at Joint Conference of 30th International Symposium on Space Technology and Science 34th International Electric Propulsion Conference and 6th Nano-satellite Symposium, Japan.

Pipelines

Franco Cotana^{*,†}, Mattia Manni[†], ^{*}Department of Engineering, University of Perugia, Via Goffredo Duranti, Perugia, Italy; [†]Interuniversity Research Center on Pollution and Environment “Mauro Felli”, Via Goffredo Duranti, Perugia, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	646
History and Development	646
Pneumatic Capsule Pipelines during the Victorian Age	646
The Cold War of Pneumatic Capsule Pipelines	647
From the Occident to the Orient	649
Hydraulic Capsule Pipelines in the 20th Century	650
Magnetic Levitation Transport Systems	650
Low-pressure Pipelines and Hypersonic Transportation	650
Taxonomy of Pipeline Transport	651
Future Trends	652
Conclusions	653
Acknowledgment	654
Biographies	654
Relevant Websites	654
References	654
Further Reading	655

Introduction

A pipeline consists of a line of pipes that is equipped with control devices for the movement of solids and fluids. Tubes can be made of metal, concrete, clay products, and plastics. The different sections are welded together and usually placed underground. Pipelines have represented the preferred mode of transportation for fluids since the Bronze Age when they were used for water supply. Regarding the transportation of solid particles, coal and other minerals can be moved over long distances, while mail, envelopes, grain, rocks, cement, concrete, and solid wastes are conveyed over short distances. The transportation of freight through pipes brings potential advantages over currently used methods: transiting of consumer goods takes place underground, within an automated transport infrastructure that is separated from passenger traffic. Managing traffic flows in such a way would be beneficial not only in congested centers but also for decongesting traffic jams on highways (Turkowski and Szudarek, 2019).

The freight pipeline transportation was conceptualized in the 19th century for conveying mail and small packets. According to this concept, goods were charged into a capsule (unit of carriage) that was propelled by a fluid along the pipeline network. Two different technologies have been defined depending on the fluid's nature: when the propeller is a gas, the system is termed “pneumatic capsule pipeline,” conversely when it is a liquid, the system is called “hydraulic capsule pipeline.” The fluid can be blown down or extracted from the pipeline, the difference of pressure between the two ends of the cylindrical container pushes this along the tube. The system layout was completed by the loading terminal (where cargo entered the system) and the unloading terminal (where cargo left the system), as depicted in Fig. 1 near here.

When it came to moving cargo units through pipes with a diameter longer than one meter, the aforementioned propulsion technologies show lower effectiveness as well as excessive energy consumption. Hence, an optimized configuration combining linear electric motors with the advantages of evacuated tubes (minimized drag) was implemented. Such freight transport systems could be a valid alternative for, but also an addition to existing transport modes. This would also contribute to solving issues related to noise, air pollution, and global overwarming phenomena.

The chapter is structured as follows: an introductory section that describes the pipeline transportation technology and which also explains limits and advantages; a *History and development* section where milestones in pipeline evolution are summarized; a section about taxonomy which focuses on defining a cluster of parameters for pipeline transportation systems' classification; a *Future trends* section describing the upcoming innovation; a conclusive section which recapitulates the potentials in pipeline transportation.

History and Development

Pneumatic Capsule Pipelines during the Victorian Age

Pneumatic transportation was firstly theorized and developed by William Murdoch around 1799. Back then, the British engineers were the only ones investigating the potentials of pneumatic capsule pipeline in transmitting telegrams. Small envelopes were

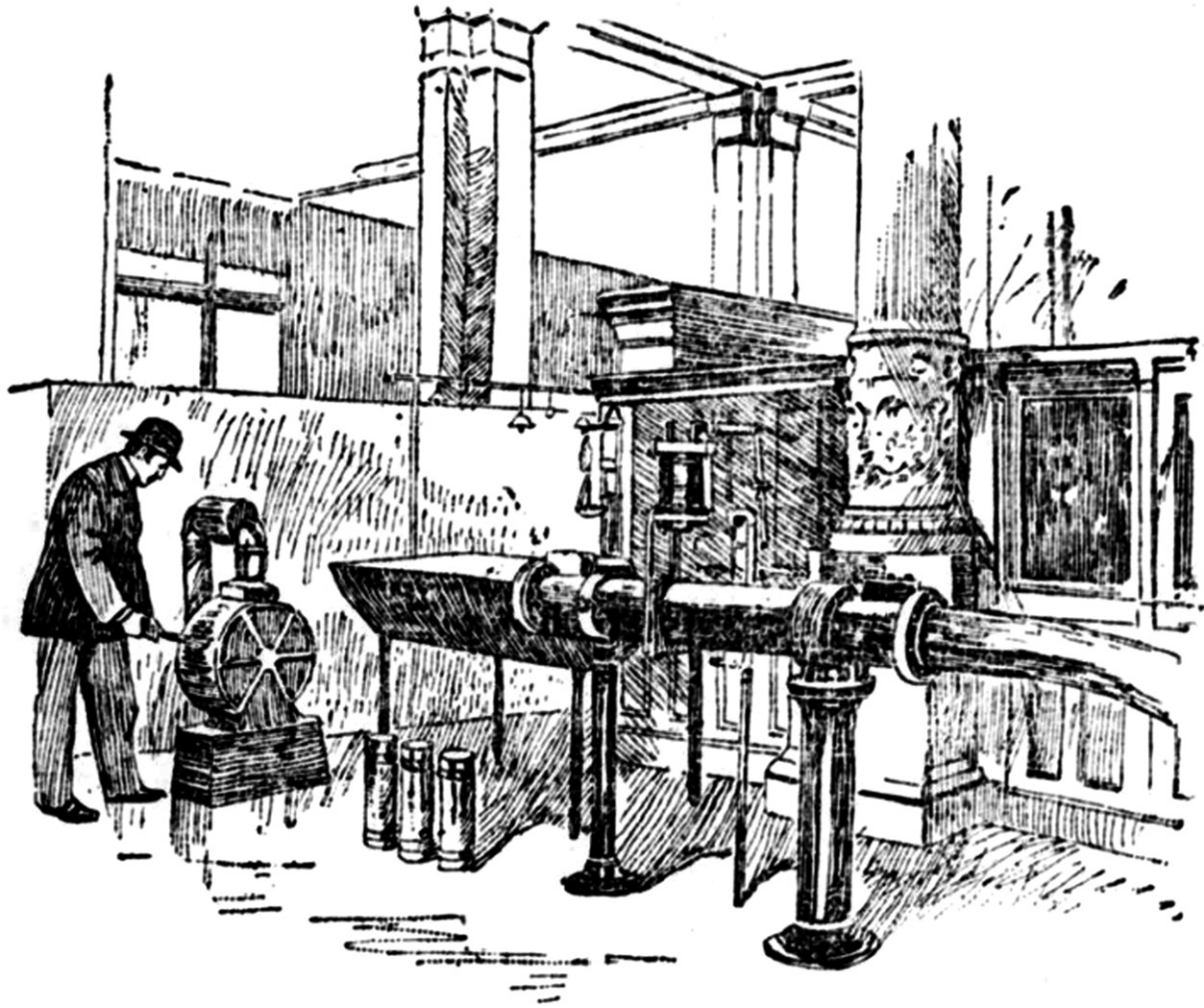


Figure 1 Drawing from the database *Chronicling America* at the Library of Congress that represents the terminals of the pneumatic capsule pipeline.

sealed into capsules and conveyed between two places by the pressure of air and vacuum ([Fig. 2](#)), as reported in Josiah Latimer Clark's patent (1854). The first line was installed in 1853 between the offices of the Electric and International Telegraph Co. and the London Stock Exchange. The pneumatic tube negated the need to employ runners to carry messages between the two buildings, while guaranteeing speed and independence from weather conditions. Next, the Electric and International Telegraph Co. implemented similar systems in Liverpool, Birmingham, and Manchester. By 1880, there were over 30 km of tubes in London, which increased to 64 km in 1909. From the mid-1860s to the late-1890s, such systems became operational in the rest of the world: Berlin, Munich, and Hamburg in Germany; Vienna in Austria; Prague in Czech Republic; Rio de Janeiro in Brazil; Dublin in Ireland; Rome, Naples, and Milan in Italy; Paris and Marseille in France; New York City, Boston, Philadelphia, St. Louis, and Chicago in the United States; Melbourne in Australia; Tokyo, Osaka, Nagoya, and Kobe in Japan ([Bush, 2017](#)). From 1859, wheeled capsules were also tested by the London Pneumatic Dispatch for the movement of heavier cargos (up to three tons). A prototype of the Pneumatic Passenger Railway was realized in New York for the movement of people, as shown in [Fig. 3](#). However, the evolution of the telecommunication methods into the modern ones caused the pneumatic tube to lose its importance. In 1874, the London Post office considered it was no longer convenient using the system, and it was closed.

The Cold War of Pneumatic Capsule Pipelines

In the 1960s, renewed concern in large-diameter pneumatic tubes started as a consequence of the growth in interest in high-speed ground transportation. The invention of the "jet pump injector" (or booster pump) for pneumatic capsule pipelines permitted moving larger volumes ([Carstens, 1970](#); [Howgego and Roe, 1998](#)). Two prototypes were realized in Stockbridge (Georgia, 1971) and Houston (USA, 1973) for transporting heavy cargo over relatively long distances ([John A Volpe National Transportation Systems Centre, 1994](#); [Vance and Mills, 1994](#)). Both the systems were named "Tubexpress" and they differ for the diameter length: the first



Figure 2 Miss Helen Ringwald works with the pneumatic tubes in Washington, D.C. in 1943.

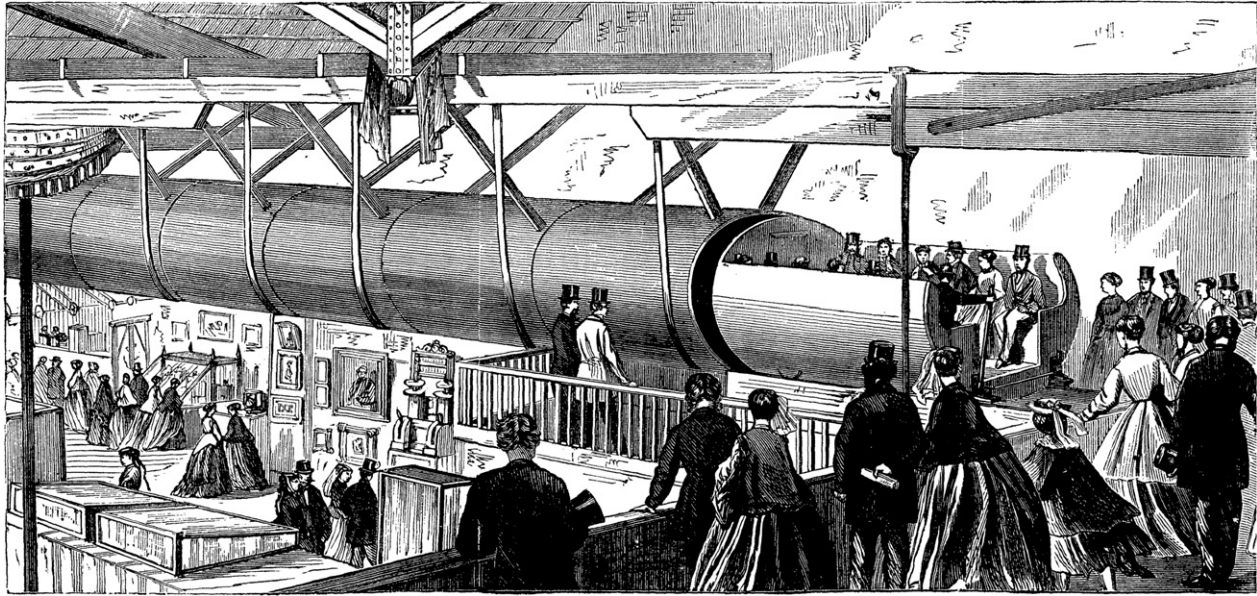


Figure 3 The Pneumatic Passenger Railway, as Erected at the American Institute, Fourteenth Street, New York, 1867.

prototype's diameter was equal to 0.91 m, and it doubled the second's. Following this, wheeled capsules were put into operation also in the United Kingdom and USSR. A 550 m long ring was installed in Cranfield (UK) in 1976, and tests were conducted there until 1979 (Baker and Hawrych, 1982; Fidler, 1973; Simper and Baker, 1975). The experimental facility permitted collecting performance data while optimizing the design of containers and evaluating the suitability of different materials for making pipes. Concerning Russian projects, several "Transprogress" systems were built and became operational between 1971 and 1983 (Jvarsheishvili, 1981). The first was designed by SKB Transnefteavtomatika for the movement of crushed rock at Tbilisi (Georgia, 1971) along a 2.1 km long pipeline. Cargo was transported into a train formed of up to six capsules, having a maximum load capacity of 15 ton. Such a train was capable of achieving a peak speed of 50 km h^{-1} when unloaded. A similar pneumatic pipeline having a smaller diameter (0.18 m) was installed within a manufactory plant in Moscow (Jvarsheishvili, 1981). This freight transport system called "Start" was used for moving unspecified products from the working zones to the warehouses. Soviets planned to build 20 pneumatic tubes across the former USSR, and most of these were realized and used for the movement of gravel, sand, minerals, and waste. In 1980, the "LILO-2," an 18 km long mineral transport system with a diameter larger than one meter (1.22 m) was realized in Georgia, and its length was more than doubled within the following four years (until 49 km) (Alexandrov, 1978). The system, that is still the most extensive on the planet, used to move around 2 million tons per year. Also, a pneumatic tube for waste disposal became operational in the former Leningrad in 1983, but the system was dismissed before the 1990s (Bahke, 1982; Simper and Baker, 1975).

From the Occident to the Orient

Between 1975 and 1985, the major developers of pipeline transportation (USA, USSR, and the UK) attempted to find markets for their technologies (Butler et al., 1989; Institute Research Office, 1979). Although the Soviets were the most successful, the implementation of the system within the USSR was suddenly interrupted after 1985. No commercial application was found for the British technology, while the Tubexpress unsuccessfully tried to promote the construction of a pneumatic tube for the transportation of grains in Montana. After this phase, both the Soviet and American technologies were sold to Japan that had started investigating pneumatic capsule pipelines back in the 1960s (Yanaida, 1982). On the one hand, the Transprogress concept was applied to a system for the movement of limestone between the mine and the cement plant of Sumitomo Cement Company, in Tochigi Prefecture. This 1-m diameter pipeline was realized on a disused railway line in the early-1980s. Capsules were arranged into trains formed of three elements, and each capsule was loaded with up to 1.6 ton. On the other hand, Nippon Steel Corporation and Daifuku further enhanced Tubexpress technology (Yanaida, 1982). Four pneumatic tubes with different diameters (ranging from 0.2 m to 0.9 m) were constructed for testing the suitability of the system for conveying automobile parts, reliability, and durability of materials, and also operational with heavier cargos. Following this, Airapid technology was implemented. Twin tubes having a 0.61 m diameter were installed for transporting burnt lime over a distance of 1.5 km between the different areas of the steelworks (Nippon Steel Corporation, n.d.). While Sumimoto's technology was mainly exploited in temporary applications (construction of tunnels), the Airapid was exclusively used in permanent installations.

Hydraulic Capsule Pipelines in the 20th Century

In contrast to the long history of pneumatic systems, the hydraulic concept is still in its infant stage. Such technology was implemented and used by the British military for transporting war materials during World War II. Between 1958 and 1975, engineers from Canada and United States worked on the further enhancement of this method, resulting in a new type of hydraulic tube for moving compressed coal logs. This system did not require capsules; the coal was submerged in the water and propelled by the flow.

Magnetic Levitation Transport Systems

Another milestone in the history of pipeline transportation is represented by the utilization of linear electric motors for the propulsion of capsules. Such prime mover technology, also known as magnetic levitation (maglev), was originally exploited on the railway and applied in place of traditional tracks ([Fig. 4](#)). Therefore, from now on, the evolution of pipeline systems cannot be investigated without taking into account some advances in railway transportation. Progresses concerning track-based methods, which also contributed to the enhancement of pipelines will be reported in brief in this chapter. Nonpneumatic propulsion technology was subject to a US patent granted to William Vandersteel in 1984 ([Vance and Mills, 1994](#)). Such a concept was implemented on the "SUBTRANS" project, a capsule pipeline developed in the United States in 1992 and exploiting linear induction motors. The system was intended to operate at a constant speed of 100 km h^{-1} and capsules were capable of accepting pallets. In 1994, researchers at Kawasaki and NKK tested a similar prototype where capsules were pushed along a 45 m long tube by linear synchronic motors. Following this, Japanese engineers undertook test runs on "Shinkansen" trains (1997). Shinkansen trains consist of maglev vehicles, running at speed of over 500 km h^{-1} . Also, a demonstration project from Magplane Technology Co. using the same propulsion became operational in a phosphate mining company (IMC-Agrico Company) in Lakeland, Florida (2001) ([Montgomery et al., 2000](#)). Cargos were moved at a peak speed of 65 km h^{-1} through a 200 m long pipelines. At the beginning of the 21st century, an automated transport system for people called MicroRail was realized in the Texas region. According to this concept, a maglev train is supposed to run on a guideway above the ground. During 2013, Mole Solutions company realized a pipeline demonstrator near Cambridge. Two capsules, that were propelled by linear induction motors, were put into operation to convey bulk materials along a track length of 105 m. Such technology was also exploited in the CargoCap plant which was conceptualized in Germany in the late-1990s. A test track prototype was built in Bochum and it is still operating. When fully developed, this system would allow transporting consumer goods along an underground network, at a peak speed of 40 km h^{-1} .

Low-pressure Pipelines and Hypersonic Transportation

The modern age of pipeline transportation began in 1991 when Gerard O'Neill submitted a patent application describing the "vactrain" hypersonic technology ([Dyson, 1993](#)). It described a vacuum-tube train, capable to exploit magnetic levitation tracks coupled with the advantages of partly evacuated tubes. O'Neill estimated such trains could reach a speed as high as 4000 km h^{-1} if



Figure 4 A maglev train is coming out of the Pudong International Airport.

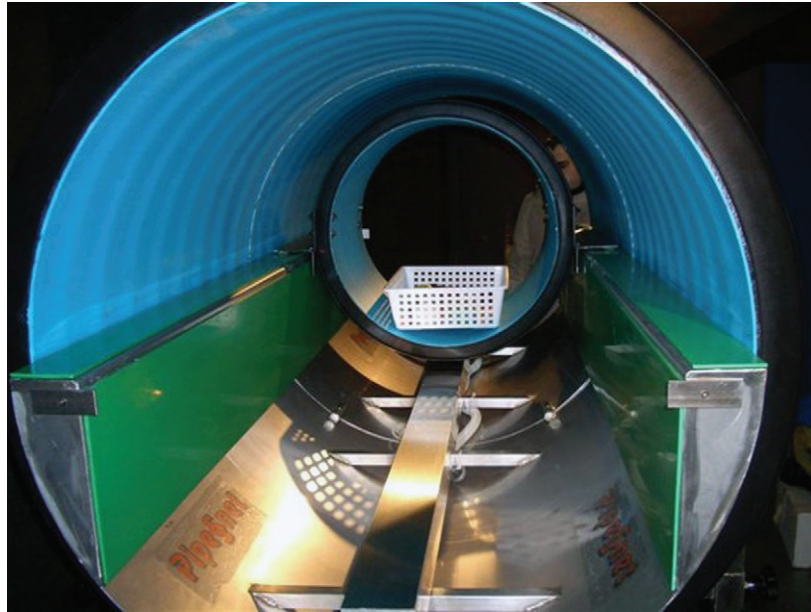


Figure 5 A test running on PipeNet prototype for testing the installed linear electric motors.

the air was completely blown out from the tunnels. During 1992, the company Swissmetro was created to develop a futuristic project for Swiss national transportation based on hypersonic vehicles (Jufer et al., 2006). Maglev trains were supposed to travel in low-pressure tunnels, connecting the Swiss network to the major pathways of the German Transrapid monorail (built in 1991). However, the project has never been realized and the company went into liquidation in 2009 because of the lack of economic support. During the same decade, the PipeNet system was implemented in Italy (Cotana et al., 2008). Research activities coordinated by professor Franco Cotana (University of Perugia) resulted in a patent that was granted in 2005. This innovative concept firstly proposed applying the vactrain technology to capsule pipelines for the movement of commodities between two hubs (Fig. 5). The fluid dynamics of a running capsule was analyzed in a 100 m long tube in Terni (Fig. 6), highlighting that such containers can theoretically achieve a peak speed of 1500 km h^{-1} . In the United States, the ET3 Global Alliance, founded by Daryl Oster, focused on the global implementation of evacuated tube transport technologies. The main findings were used for filling a US patent that was granted in 2009. The ET3 experience has significantly influenced the Hyperloop concept presented by Elon Musk in 2013. Following this, several companies such as Virgin Hyperloop One (Fig. 7), Hyperloop Transportation Technologies, and TransPod have been founded to develop such maglev system for either people or freight transportation through evacuated tubes (Janzen, 2017). Test runs have been successfully undertaken, and full-scale systems are expected to be realized soon in Texas, Colorado, Florida, Chicago, Dubai, and the United Kingdom.

Taxonomy of Pipeline Transport

In this section, a list of criteria for classifying pipeline transport systems is presented. Firstly, water and sewer lines, oil pipelines, slurry tubes, pipe networks for gases, and other fluids as liquid fertilizers, and also freight transport lines (either for light-weight or heavy cargos) can be distinguished. Furthermore, plants for people transportation have been added to these as a new category, during the 21st century.

Secondly, pipelines are classified basing on the way commodities are moved along their network. Consumer goods can be conveyed either into capsules or submerged in the fluid. Regarding capsules, these can be eventually arranged into a train.

Thirdly, the nature of the fluids filling the tubes represents a parameter for determining different clusters. The presence of fluid generally characterizes pipeline inner volume, although the movement of freights can also happen through the evacuated environment (vacuum tube). Such fluid can be the transported one (i.e., oil, slurry, natural gas) or the one pushing cargos (air or water).

Fourthly, the exploited prime mover represents another criterion for the classification. Pneumatic pumps or hydraulic pumps can be exploited for blowing in or extracting the fluid from pipes, generating thus the difference of pressure between the two capsule's ends, which is needed for the propulsion. If no pump is installed, linear electric motors are used for the movement of cargos.

Finally, pipeline transportation systems can be classified depending on the speed as well as on the covered distance. The speed categories range from low-velocity to hyper-velocity, while the ones based on the distance vary from building-scale to urban-scale, although some potential applications of the existing prototypes could even cover a whole region.

For example, two pipeline transport technologies from different ages are here defined in accordance with the criteria described in the lines above. Therefore, the system used for moving mail and telegrams between the offices of the Electric and International



Figure 6 Professor Franco Cotana and the 100 m long prototype of PipeNet system built in Terni.

Telegraph Co. and the London Stock Exchange (1853) was a “low-velocity pneumatic capsule pipeline for light-weight freight transportation at district scale,” while the PipeNet technology (2005) consists of a “hypersonic maglev capsule evacuated pipeline for heavy freight transportation at regional scale.”

Future Trends

The review of pipeline transport systems presented in the present chapter highlighted that the technology required for moving capsules and trains at hypersonic speed has been already developed, and its reliability demonstrated. During the last decades, different small-scale prototypes were realized to gain knowledge about the effectiveness of linear electric motors coupled to evacuated tubes. Therefore, the upcoming step would imply the construction of the first pilot project capable of reaching the hypothesized peak speed (higher than 1000 km h^{-1}) within a partly evacuated environment, and research groups working on such technology are rapidly approaching this goal. Switzerland is planning to construct a comprehensive system named “Cargo Sous Terrain,” that would be fully automated and designed for transporting freights and people. The first phase of this system is expected to be launched by 2030. Also, numerous companies such as PipeNet, Virgin Hyperloop One, Hyperloop Transportation Technology, and TransPod have promoted fundraising campaigns in order to support the realization of their systems. The lack of adequate safety systems to reduce the risk for the transported freights or persons currently represents a barrier to the development of hypersonic vehicles. Thus, it can be reasonably assumed that consumer goods will be traveling at such speed earlier than people due to the lower level of safety needed.

Regarding the transportation of commodities through pipelines, various diameter lengths had been tested for capsules and tubes to simulate different utilizations. In the future, these values should be leveled out and standardized. A specific regulation should assign a diameter length to the pipeline depending on the characteristics of moved freight. Sizing process of capsules should also take into account the potential interconnection with the existing logistics network. Following this, the contributions from Modular Logistics Units in Shared Co-modal Networks (Modulushca) project and the work done by Alliance for Logistics Innovation through



Figure 7 Front view of Virgin Hyperloop One XP-1 test pod on display at Rockefeller Center on September 2019.

Collaboration in Europe (ALICE) European Technology Platform would be fundamental (in Europe). The former aims at enabling operations with iso-modular cargo units, providing a basis for an interconnected logistics system for 2030. The latter has as a goal the development of a comprehensive strategy for research, innovation and market deployment of logistics and supply chain management innovation in Europe.

Finally, the further enhancement of pneumatic and hydraulic tubes should be limited to small diameter systems (and light-weight cargos) since linear electric motors have been demonstrated being more efficient when it comes to larger pipes (and heavier cargos). Improvements should consist of integrating these to the hypersonic network as last-mile transport solutions, hence delivering packs to end-users. Determining a range of potential solutions capable to solve issues concerning last-mile logistics represents another aspect that is worth investigating.

Conclusions

The increase in freight transportation and the consequent congestion of the highway and rail systems have provoked a renewed interest in alternative solutions like capsule pipelines. The implementation of the linear electric motor technology has permitted to reduce the energy consumption for propelling cargos over long distances. Also, the conceptualization of vactrain vehicles has encouraged the development of hypersonic systems capable of achieving peak speeds never experienced. Using very high-speed automated-guided capsules for the movement of freights and people would lower traffic jams, the amount of pollutants released in the atmosphere, and the dependency on fossil fuels, while increasing the level of transport security. However, such plants require the construction of dedicated infrastructure and considerable funds, and also they are still considered as too futuristic by society. Lack of social acceptance along with a shortage of policy support is widely recognized as two of the main barriers to the technology application.

Despite this, private investors are getting more and more interested in funding such projects as the cases of PipeNet in Italy, Cargo Sous Terrain in Switzerland, Hyperloop in the United States, and TransPod in Canada. The next stages and challenges have been

already identified; and they consist of realizing the first pilot systems, defining the standards regulating capsules' sizing, and investigating last-mile delivery systems. Within this framework, the contribution of governments with respect to decision making would be essential. Developing proper and integrated planning of the activities (on surface and subsurface) with adequate standards, which are established in advance, is fundamental for introducing tube transportation in congested areas.

In conclusion, the beginning of the 21st century can be reasonably considered as the turning point in the history of pipeline systems as well as transportation in general. Hypersonic technologies have been developed and previous limits concerning peak speed have been overcome, and also people are getting familiar with the "just in time" concept. Promoting an adequate regulatory framework or private initiatives is expected to represent the definitive push towards the realization of such projects.

Acknowledgment

Acknowledge any kind of support and/or financial assistance here when applicable.

Biographies

Franco Cotana was born in Marsciano (Perugia) in 1957. He graduated in electronic engineering in 1983 at the University of Rome "La Sapienza". He was a full professor in "Industrial Applied Physics" at the University of Perugia; member of the Board of Administration of the University of Perugia from 2014 till 2019; director of CIRIAF (Interuniversity Research Centre for Pollution and Environment–Mauro Felli), based at the University of Perugia, from 2013 till 2015; founder and director of CRB (National Biomass Research Centre), a research center of the University of Perugia sponsored by the Italian Ministry of Environment, Sea, and Land Protection, from 2003 till 2013; and president and coordinator of the PhD in "Energy and Sustainable Development" at the University of Perugia. He is an author of more than 400 scientific papers and 25 patents in the fields of energetics, sustainable transports, applied acoustics, heat transfer, applied thermodynamics, and lighting design; coordinator of many national and international research projects such as the "GreenPost" and "PostalZEV" projects; national coordinator of the "FISR Vettore Idrogeno" project; member of the New York Academy of Sciences since 2010; representative of the Italian Government in Brussels for the EIBI SET PLAN Bioenergy and Renewable Fuels for Sustainable Transports; and member of the Coordination Board of the AIA-IPPC Commission at the Italian Ministry for the Environment, Sea, and Land Protection.

Mattia Manni was born in Foligno (Perugia) in 1991. He graduated in Architecture and Construction in 2015 at the University of Perugia. He worked as an urban planner on the enhancement of the suburb areas around the Perugia train station. He is doctor of philosophy in "Energy and Sustainable Development" since 2020; author of publications in the fields of energetics, sustainable development, solar analyses, smart materials, responsive building design, and low-carbon technologies. He is involved in national cluster technology projects funded by the Italian government on third-generation biofuel solutions and cooperates with the temporary working group of the SET Plan Renewable Fuels and Bioenergy IP.

Relevant Websites

'MicroRail Personal Automated Transport (PAT) System' www.faculty.washington.edu
 'MOLE – Shaping the future of freight' www.molesolutions.co.uk
 'CargoCap–Freight transportation in congested urban areas' www.cargocap.com
 'PipeNet' www.pipenet.it
 'Hyperloop' www.tesla.com/blog/hyperloop
 'Virgin Hyperloop One' www.nytimes.com/2019/02/18/technology/hyperloop-virgin-vacuum-tubes.html
 'Hyperloop Transportation Technologies' www.hyperlooptt.com
 'Modular Logistics Units in Shared Co-modal Networks' cordis.europa.eu/project/id/314468
 'ETP–Alice' www.etp-logistics.eu

References

- Alexandrov, A., 1978. Pneumatic pipeline container transportation of goods. In: *Proceedings Pneumotransport 4*. Carmel-by-the-sea, USA, pp. 51–59.
- Bahke, E., 1982. Pipelines were only the beginning—modern structural design for freight pipelines. *J. Pipelines* 2, 133–147.
- Baker, P.J., Hawrych, J.R., 1982. Pneumatic capsule pipeline system research and development in the UK. *J. Pipelines* 2, 237–249.
- Bush, C., 2017. Letters in a tube: the rise and demise of pneumatic mail. *Hist. Mag.* 31–34.
- Butler, T., Jacobs, B., Veasey, D., 1989. Commercial viability and potential market for pneumatic capsule pipelines. In: *ProcEedings of the 6th International SympOsium on Freight Pipelines*, p. 129.
- Carstens, M., 1970. Analysis of a low-speed capsule-transport pipeline. In: *Hydrotransport 1*. BHRA, Warwick.
- Cotana, F., Rossi, F., Marri, A., Amantini, M., 2008. Pipe\$net: innovation in transport through high rate small volume payloads. In: *The University of Texas (Ed.), Proceedings from the 5th International Symposium on Underground Freight Transportation by Capsule Pipelines and Other Tube/Tunnel Systems*. Arlington.
- Dyson, F.J., 1993. Gerard Kitchen O'Neill. *Phys. Today* 46, 97.

- Fidler, J., 1973. Pneumatics can be used to dispose of your firm's waste. *Eng.* 68–71.
- Howgego, T., Roe, M., 1998. The use of pipelines for the urban distribution of goods. *Transp. Policy* 5, 61–72. [https://doi.org/10.1016/S0967-070X\(98\)00012-2](https://doi.org/10.1016/S0967-070X(98)00012-2).
- Institute Research Office, 1979. Transportation of grain by pipeline and other methods; a preliminary response to 1979 HJR 61, an official document presented to the 1979 Montana Legislature. Bozeman.
- Janzen, R., 2017. TransPod ultra-high-speed tube transportation: dynamics of vehicles and infrastructure. *Procedia Eng.* 199, 8–17. <https://doi.org/10.1016/J.PROENG.2017.09.142>.
- John A Volpe National Transportation Systems Centre, 1994. *Tube Transportation*. Cambridge, MA.
- Jufer, M., Bourquin, V., Sawley, M.L., 2006. Global modelisation of the Swissmetro maglev using a numerical platform. In: *ProcEedings of the 19th International Conference on Magnetic Levitated Systems for Linear Drives*.
- Jvarsheishvili, A.G., 1981. Pneumo-capsule pipelines in USSR. *J. Pipelines* 1, 109–110.
- Montgomery, B., Fairfax, S., Beals, D., Taylor, E., Whitley, J., Smith, B., 2000. Electromagnetic pipeline demonstration project. In: *TRAIL Research School (Ed.), Proceedings 2nd International Symposium Underground Freight Transportation by Capsule Pipelines and Other Tube/Tunnel Systems (ISUFT 2000)*. Delft.
- Nippon Steel Corporation, n.d. Nippon Steel Corporation. Daifuku Machinery Works Ltd. Airapid, Capsule-Tube Transport System. Promotional Information.
- Simper, J., Baker, P.J., 1975. Pneumatic pipeline capsule systems — the future potential. In: *Proceedings Pneumotransport 2, BHRA Conference*. Guildford, UK, pp. 31–39.
- Turkowski, M., Szudarek, M., 2019. Pipeline system for transporting consumer goods, parcels and mail in capsules. *Tunn. Undergr. Sp. Technol.* 93, 53–66. <https://doi.org/10.1016/j.tust.2019.103057>.
- Vance, L., Mills, M.K., 1994. Tube freight transportation. *Public Roads* 58, 21–27.
- Yanada, K., 1982. Capsule pipeline research in Japan. *J. Pipelines* 2, 117–131.

Further Reading

- Hayhurst, J.D., 1974. *The Pneumatic Post of Paris*. France & Colonies Philatelic Society, Blairgowrie.
- Janić, M., 2014. *Advanced Transport Systems*. Springer, London.
- Kieve, J.L., 1974. *The Electric Telegraph in the U.K.: A Social and Economic History*. David & Charles Publishers, Newton Abbot.
- Konings, R., Priemus, H., Nijkamp, P., 2005. *The Future of Underground Freight Transport*. Edward Elgon Publishing, Cheltenham.
- Kulwiec, R.A., 1985. *Materials Handling Handbook*. Wiley, Hoboken.
- Liu, H., 2003. *Pipeline Engineering*. Lewis Publishers, Boca Raton.
- Standage, T., 1998. *The Victorian Internet: The Remarkable Story of the Telegraph and the Nineteenth Century's Online Pioneers*. The Berkeley, Phoenix.
- Woodcock, C.R., Mason, J.S., 1987. *Bulk Solid Handling*. Blackie Academic & Professional, Glasgow.
- Egbunike, O.N., Potter, A.T., 2011. Are freight pipelines a pipe dream? A critical review of the UK and European perspective. *J. Transport Geogr.* 19, 499–508.
- Visser, J.G.S.N., 2018. The development of underground freight transport: an overview. *Tunnel. Underground Space Technol.* 80, 123–127.
- Shahooei, S., Farooqhi, F., Zahedzadani, S.E., Shahandashti, M., Ardekani, S., 2018. Application of underground short-haul freight pipelines to large airports. *J. Air Transport Manage.* 71, 64–72.
- Oster, D., Kumada, M., Zhang, Y., 2011. Evacuated tube transport technologies (ET3)™: a maximum value global transportation network for passengers and cargo. *J. Modern Transport.* 19, 42–50.
- Cotana, F., 2000. Br. Pat. 01319584.
- Cotana, F., 2005. Br. Pat. 0001365803.

Women and Transport Modes

Priya UtengPhD*, **Yusak SusiloPhD†**, *Department of Sustainable Urban Development and Mobility, Institute of Transport Economics (TOI) Gaustadalléen Oslo, Norway; †Department of Landscape, Spatial and Infrastructure Sciences, Institute for Transport Studies (IVe), University of Natural Resources and Life Sciences, Gregor-Mendel-Straße Vienna, Austria

© 2021 Elsevier Ltd. All rights reserved.

Introduction and Background	656
Trip Purposes	656
Trip Chains	656
Daytime Distribution and Concentration of Trips	656
Licensing and Car Availability	657
Trip Distances and Trip Duration	657
Transport Mode Choice	657
Smart and New Modes of Transport Provision and Women	660
Summary and Conclusion	661
References	662
Further Reading	664

Introduction and Background

The socialization of role assignments over time, with women being the care-giver and male being the breadwinner, gets reflected in access to space-time dynamics, opportunities/activities and resources. Whilst in the last couple of decades there has been a significant rise in women's participation in employed workforce, particularly in the Global North, women are still more likely than men to be the primary carer at home and responsible for various domestic tasks, and less likely to be in full-time employment. This leaves women with complex and intricate scheduling activities in both time and space (Kwan, 1999; Susilo and Dijst, 2009; Hanson, 2010). Furthermore, low mobility levels and constrained access to resources among women reduce their possibilities of solving this complex scheduling problem. This has also led to a markedly different travel behavior of women which gets expressed in the following variations:

Trip Purposes

Previous studies (e.g., Priya Uteng, 2006; Susilo and Maat, 2007) show that low mobility levels and time-space constraints, such as childcare obligation, are two important factors explaining why women tend to work closer to home and use more public transport, compared with their male counterparts. Gender differences regarding resources of time, money, skills, and technology lead to differences in the travel and transport patterns (Law, 1999), which in turn create a differentiated labor market and personalized space-time opportunities.

Since women are either unemployed, working part time, or working in close vicinity of their homes along with balancing their household responsibilities, it is noted that women make fewer job and business trips, but more shopping and escort (children and elderly) trips (e.g., Best and Lanzendorf, 2005). This has led to more varied/complex activity patterns of women (e.g., Scheiner and Holz-Rau, 2017).

Trip Chains

Following on point one, relatively higher incidences of complex trip chaining are found among women (e.g., Paleti et al. 2011; Scheiner and Holz-Rau 2017). Since women's trip chains are more complex than those of their male counterparts, they are more dependent on a comprehensive public transport network. King (1980) argued that while travel choices for men primary depend on travel time, availability and affordability are the crucial issues for women. Further, since women's travel destinations are closer to home, they are more likely to use public transport twice as often as men. However, it has been noted that the gender gap in public transport use might be reducing over time (Susilo and Maat, 2007; Hjorthol, 2008).

Daytime Distribution and Concentration of Trips

Both due to personal security concerns and as a matter of preference, women go out by night less often (e.g., Scheiner, 2013) and given their employment status, women travel less often in rush hour.

Licensing and Car Availability

Traditionally, license and car availability has been higher for men (e.g., [Hjorthol, 2008](#)) but in recent works undertaken in the Global North, it is noticed that both licensing and car availability are converging ([Scheiner 2018](#)).

Trip Distances and Trip Duration

Typically, both trip distance and trip duration have been shorter for women. This is especially the case for work trips but holds true for other purposes as well (e.g., [Hjorthol, 2008](#); [Scheiner et al., 2011](#)). It is also noted that women participate less in long distance work- or business-related travel. On average, women have shorter travel time and more non-work activities than their male counterparts (e.g., [Transport for London, 2012](#)). Recent studies, however, point toward a convergence, taking place in this arena as well. There is even a mention of “gender turnaround” in distances traveled by young people ([Tilley and Houston, 2016](#)).

Discussions expanding on one or more of these themes can be found in the following contributions of this volume:

Gendered mobility. *Sheila Mitra-Sarkar*

Women's & gender's issues in transportation

Women in Transportation

Gender Issues

The value of security, access time, waiting time and transfers in public transport. *Juan de Dios Ortúzar, Raquel Espino and Luis I Rizzi*

The economics of transportation safety. *Ian Savage*

Age and Gender as factors in road safety. *Marion Sinclair*

Pedestrian safety, general. *Muhammad Zaly Shah*

Safety culture. *Tor-Olav Nvestad and Ross Phillips*

Traffic safety and security of taxis and ride-hailing vehicles. *Helai Huang*

Transportation Safety and Security. *Marcy - Maria Burns*

This paper is organized as follows: after an introduction to the topic in the first, the second section expands on the thematic area of women and transport modes. Third section further builds on the topic in light of the smart cities and smart mobilities agenda. Last section presents a brief summary and concludes the chapter.

Transport Mode Choice

Traditionally, women have been car passengers, while men have been the drivers. And across the world, women use public transport and walk more often than men ([Priya Uteng 2012](#)). Given the high incidence of trip chaining, women are more multimodal (e.g., [Heinen and Chatterjee, 2015](#)). When it comes to the case of women drivers, it is noted that women typically drive smaller cars ([Choo and Mokhtarian, 2004](#)) ([Fig. 1 and 2](#)).

In the case of car driving as well, studies indicate that a convergence might be taking place in the Global North (e.g., [Hjorthol 2008](#); [Scheiner et al., 2011](#)).

To illustrate this case further, we draw on the case of Sweden where though the gender gap in terms of driving has reduced and a convergence is taking place, the gap still remains ([Susilo et al., 2018](#)). The decline in gender difference over the generations (both in total effect and in direct effect) appears most substantially in the group of partnered/married families without children (a drop from 48% in the 1930s to 18% in the 1970s). The women of the 1980s and 1990s generations still have on average 10% less car driving share than men, although that number is 30% for the 1940s generation. These trends seem to show that although more women have started driving, the gender gap in terms of driving still remains.

Regarding bicycling, previous studies have found that women's perception of danger on the road prevents them from commuting by bicycle ([METPEX, 2013](#); [Susilo and Cats, 2014](#)). [Lam \(2019\)](#) presents a synthesis highlighting that women prefer protected cycle lanes away from motorized traffic because it heightens their perception of safety ([Garrard et al., 2008](#); [Pucher et al., 2010](#); [Aldred and Woodcock, 2008](#); [Lam and Cosgrave, 2019](#)).

Lack of knowledge about bicycle maintenance and the need to do multiple trip patterns further deters women from cycling. In terms of trip purposes, [Krizek et al., \(2005\)](#) showed that women were more likely than men to cycle for shopping, making other errands and visiting friends, while they were less likely to cycle on their commute. Furthermore, they also demonstrated that women

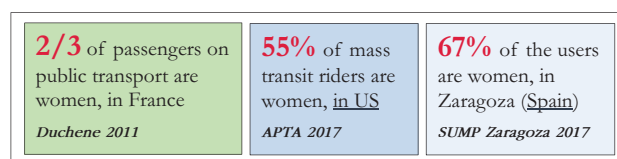


Figure 1 Women rely more on Public Transport than men. Source: [Delatte \(2018\)](#).

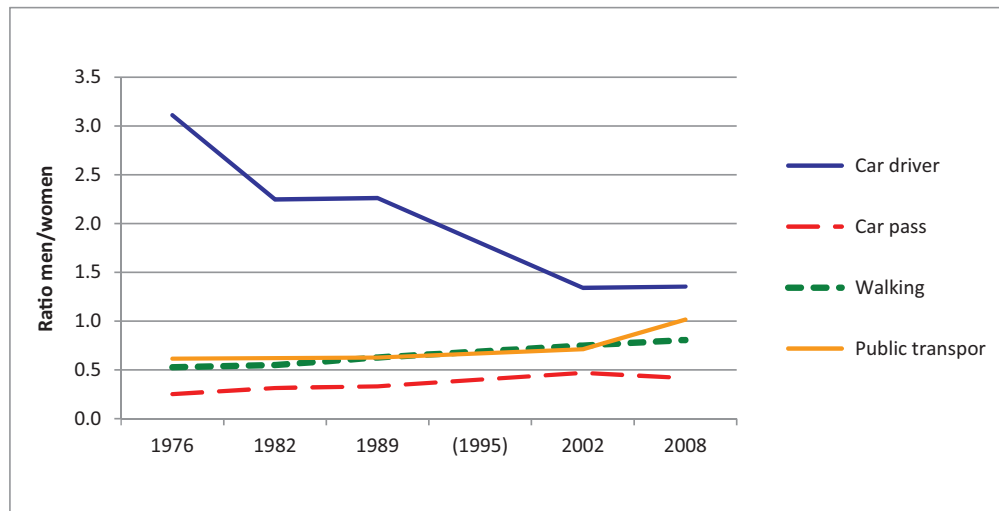


Figure 2 Gender ratios—mode choice over time in Germany. Source: *Scheiner et al. (2011)*, Reproduced in *Scheiner (2018)*.

are more risk averse and show a stronger preference for safer forms of cycling infrastructure. Previous research (e.g., [Law, 1999](#); [Waygood and Susilo, 2011](#)) found that the built environment such as land use characteristics, physical layout and design of transport networks, and facilities influence the travel behavior of men and women differently. It has also been found that wherever cycling is a part of the travel culture, women bicycle longer distances).

Bicycling for women is a direct outcome of spatial planning covering issues related to safety, slow cycling and preference for dense neighborhoods. For example, a 2018 study from Stockholm, based on time-series data from 1985 to 2015, highlighted that while the gender gap in car use and bicycling “closed completely” in the inner city of Stockholm; the gap continues to exist the outer and suburban areas ([Bastian and Börjesson, 2018](#)). Similar findings have emerged from other parts of the world as well ([Abasahl et al., 2018](#); [European Parliament, 2012](#); [Sovacool et al., 2018](#))([Fig. 3](#)).

Women are also more likely than men to travel with buggies and small children for all kinds of purposes ([METPEX, 2013](#)). In major cities, this can be a stressful challenge, and some women therefore develop strategies to avoid traveling with luggage and children ([Hine and Mitchell, 2001](#)).

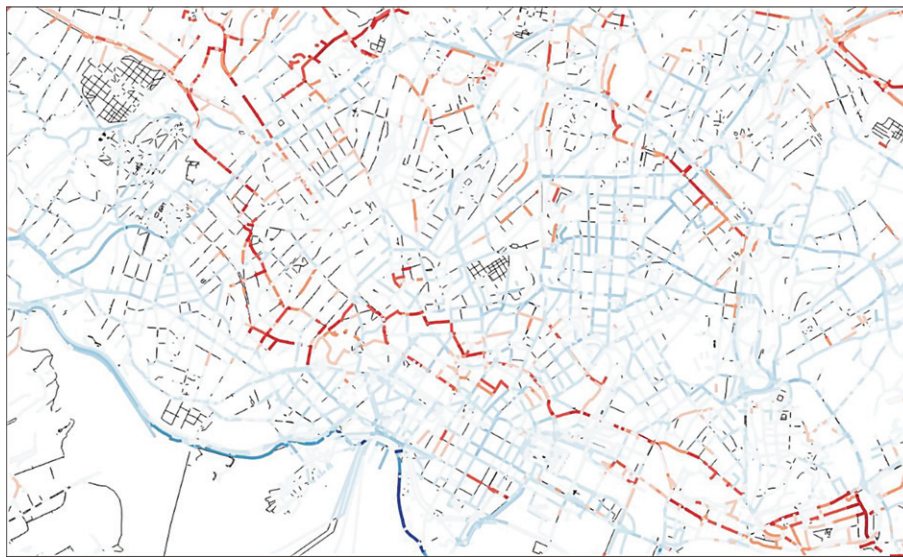


Figure 3 Difference in bicycling route preference between men and women in Oslo. The colors illustrate the gendered distribution of preferences for particular routes and roads for bicycling. The color *blue* represents a relatively higher share of men bicyclists while strong *red* color highlights female bicycling shares and preferences. Source: *de Jong et al. (2018)*.



Figure 4 Moving from a car-centric to people-centric approach in urban and transport planning. Source: *Svanfelt (2019)*.

Furthermore, women are thought to be more cautious in their travel behaviors than men (*Transport for London, 2009*). The customer typology suggests that women are more likely to fit into the categories of “travel shy,” “reassurance seeker,” and “cautious planner.” For all three categories, levels of confidence using the public transport network are relatively low (particularly so for those who are “travel shy”). As a result, women may choose to complete familiar journeys where possible or will seek advice and information to help them plan and complete their journeys (*Transport for London, 2009*). Among public transport users, multiple studies have argued that a safe, reliable and frequent service is extremely important for women. *Law (1999)* revealed that self-imposed precautionary measures adopted by women due to their safety concern limit their mobility significantly, especially in developing countries (*Priya Uteng, 2012; Hossain and Susilo, 2011; Tarigan et al., 2014*).

Women are more likely than men to walk the children to school (e.g., in London, 22% compared with 12% of men, respectively), and to walk to work/school/college (in London 53% of the women walk to work/school/college but only 49% of the men) (*Transport for London, 2012*).

Summing up, women drive less, walk more and use public transport more in comparison to men. But the field of transport planning has operated primarily on the principles of providing for cars, affecting negatively not only women but also the world economy at large. In Morocco, for example, 80% of women surveyed in 2010 affirmed that lack of transport provision limited their autonomy (*CoMun 2014*). In Jordan, 40% of women reported having to turn down employment opportunities solely due to access to an affordable and acceptable mode of transport. A McKinsey report from 2015 concludes that women’s employment is the driver for global economic growth and closing the gender gap could deliver \$12 to 28 trillion of additional GDP in 2025.

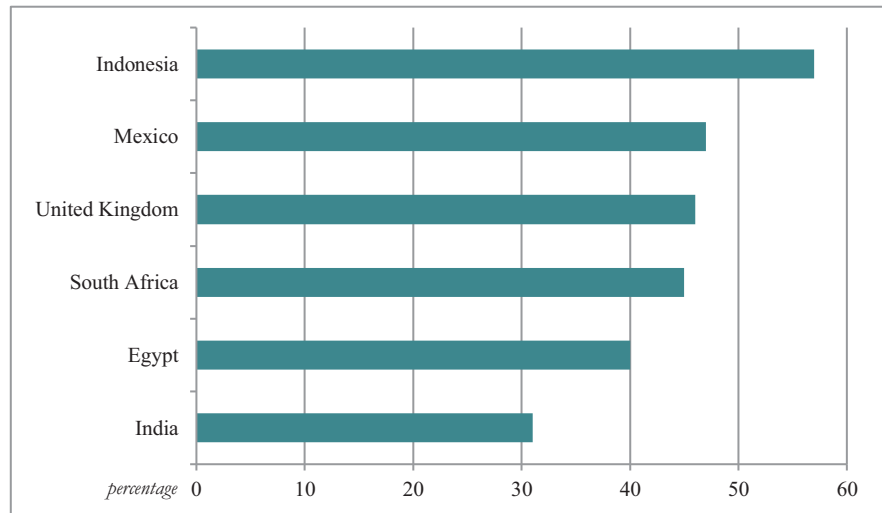
In light of Agenda 2030 and Sustainable Development Goals (SDGs), women as a category are the best candidates to build our planning systems around, as they both exhibit more sustainable travel behavior as compared to men and have a preference to continue doing so given supportive conditions. To put the trends in perspective, a simple calculation was done for the city of Malmö, which highlighted that if women were to adopt the current travel patterns of men, the following changes will occur (*Svanfelt, 2019*):

- the modal share by car would increase by 17 %
- CO₂ emissions from car traffic would increase by 31%
- Additional demand for driving and parking space would add up to 190 Möllvångstorget (standard town square).

Table 1 Gender split among private customers of various car-sharing providers in Europe around 2010.

<i>Car-sharing provider and/or location</i>	<i>Share of male customers</i>	<i>Share of female customers</i>	<i>Source</i>
Cambio, Brussels (Belgium)	58%	42%	Taxistop, cambio (2009)
Several providers	58%	42%	Italian Ministry of Environment (2009)
Three providers in London (UK)	69%	31%	Synovate (2006)
Mobility, Switzerland	53%	47%	Bundesamt für Energie (Swiss Federal Office of Energy) (2006)
Two providers in Frankfurt (Germany)	63%	37%	traffiQ (2007)
Ten providers in Germany	58%	42%	Wuppertal Institute (2007)

Source: Loose (2010), p. 54, quoted in Lenz (2017).

**Figure 5** Share of female Uber service riders 2017. Source: IFC and Accenture (2018), p. 33, quoted in Lenz (2017).

In essence, both from the sustainability and gender equity perspectives, we are looking for solutions, which can cater to multi-modal trips with high priority assigned to safe walking/bicycling conditions and creation of areas which caters to multi-purpose trips and the task of trip-chaining (Fig. 4).

Smart and New Modes of Transport Provision and Women

Given this background, it becomes important to discuss the general direction which the field of urban and transport planning is taking in light of both smart solutions and new modes of mobility.

Recent Studies highlight that the uptake of the new mobility trends is higher among men in comparison to women, for example, use of BEVs in carsharing (Kawgan-Kagan, 2015), car sharing and Uber services (Lenz, 2019). To illustrate this case further, 95% users of BEVs in Sweden are men (Haustein and Jensen, 2018) while the figure for male car sharing users in Berlin remains at a staggering 84% (Kawgan-Kagan, 2015).

The main reasons cited for the adoption of electric vehicles (EVs) and carsharing include environmental awareness, affinity towards technology, and innovation as lifestyle and it is a consistent finding that women exhibit a relatively higher flexibility to adapt their mode choice in light of the cited reasons (Li et al., 2017; Rezvani et al., 2015; Shaheen et al., 2015). Studies have found and positioned gender as an important variable, but this topic has yet to be fully explored in discussions surrounding the upcoming variations of the private car. For example, Li et al. (2017) literature review of over 40 studies advertently points towards gender being a major socio-demographic factor influencing the intention to adopt an EV. Rezvani et al. (2015) found similar results while reviewing studies on the attitudes and factors influencing the acceptance of plug-in EVs (Table 1) (Fig. 5).

Gender differences can also be found in the uptake of bike sharing schemes, and provision and designing of new biking infrastructure in general. A recent study from Oslo (Priya Uteng et al., 2020) highlights that men and women demonstrate clear differences in both preference and usage of shared bikes in Oslo. Women are more concentrated on the outer fringes of the city in



Figure 6 Route mapping of shared bikes vis-à-vis share of female employment. Source: Priya Uteng et al. (2020).

terms of spatial distribution of both their residential and employment locations but the provision of docking stations for shared bikes is concentrated in the city center. Further, a high employment/residential density of women correlates strongly with high bike sharing so it seems that the mismatch between provisions of shared bikes in the peripheral areas might be leading to underutilization of the bike-sharing scheme by women in Oslo. While men primarily use shared bikes for the purpose of access and egress, women use shared bikes for multi-purpose trips (Fig. 6).

Discussing the case of London, Lam (2019) states that despite London's increased investment in cycling infrastructure and overall growth in cycling over the past two decades, the gender gap in cycling persists. The current cycling shares highlight that men make 63% of cycle journeys in the city. Lam (ibid) further compares the case of cycle superhighways versus the quiet ways to highlight the existence of multiple kinds of systemic and structural inequalities that interact in complex ways to produce uneven urban cycling experiences. She states "Specifically, implicit male bias is present in London's treatment of cycling infrastructure, representations of cycling and cyclists and 'smart' cycling innovations. Consequently, London's cycling interventions raise the profile of already-visible, privileged cyclists (predominantly white, middle-class, able-bodied men) for whom cycling is a lifestyle choice, while erasing those for whom cycling is an economic or spatial necessity" (Fig. 7).

Summary and Conclusion

Women have been the traditional users of sustainable transport modes like walking, bicycling and public transport. But given the consistent focus on cars in both urban and transport planning domains, there has been a convergence in women's usage and access to cars. But even then, differences remain which could be partially ascribed to a continued preference for using other modes than cars and the multiple roles undertaken by women, which fits well with multi-modality. Given this backdrop, women stand to gain from new, "smart" mobility services if the solutions provide for complex activity scheduling, trip chaining and varied mode usage within an affordable time and pricing range. Services that may contribute to this are, for instance, networked public and semi-private systems, such as conventional public transport linked with last-mile bike sharing, or stand-alone sharing systems, such as car sharing, bike sharing and ridesharing (Scheiner, 2018).

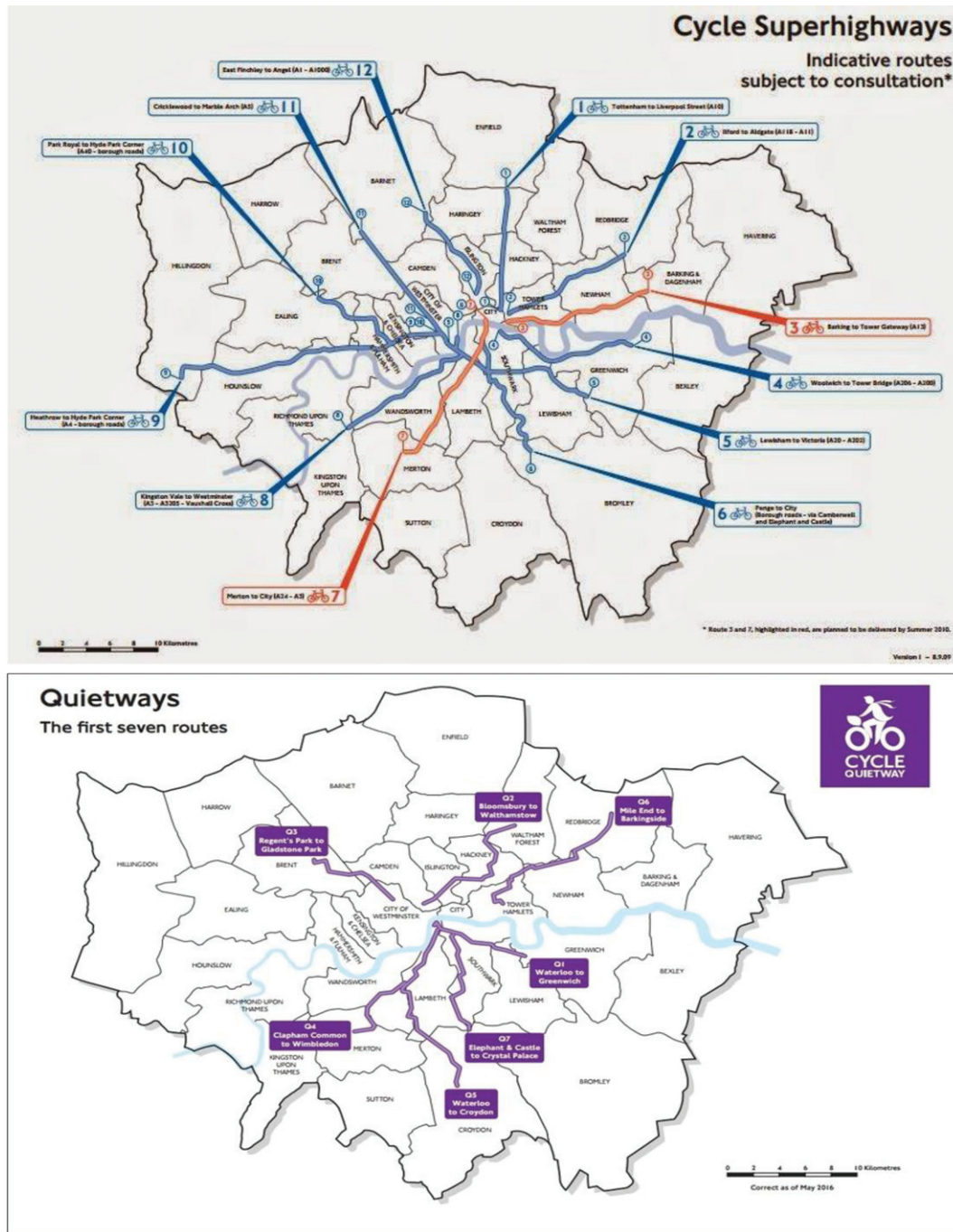


Figure 7 Cycle superhighways and quiet ways of London. Source: Transport for London, <https://tfl.gov.uk/modes/cycling/routes-and-maps>

References

- Abasahl, F., Kelarestaghi, K.B., Ermagun, A., 2018. Gender gap generators for bicycle mode choice in Baltimore college campuses. *Travel Behav. Soc.* 11, 78–85, doi:10.1016/j.tbs.2018.01.002.
- Aldred, R., Woodcock, J., 2008. Transport: challenging disabling environments. *Local Environ.* 13 (6), 485–496.
- Bastian, A., Börjesson, M., 2018. The city as a driver of new mobility patterns, cycling and gender equality: travel behaviour trends in Stockholm 1985–2015. *Travel Behav. Soc.* 13, 71–87, doi:10.1016/j.tbs.2018.06.003.
- Best, H., Lanzendorf, M., 2005. Division of labour and gender differences in metropolitan car use: An empirical study in Cologne, Germany. *J. Transp. Geogr.* 13 (2), 109–121, doi:10.1016/j.jtrangeo.2004.04.007.
- Choo, S., Mokhtarian, P.L., 2004. What type of vehicle do people drive? The role of attitude and lifestyle in influencing vehicle type choice. *Transport. Res. A* 38, 201–222.

- CoMun, 2014. Réseau Marocain de Transport Public - Etat des lieux synthétique du transport public dans les villes membres du REMA-TP. Available from: <http://www.comun.net/maroc/les-reseaux-thematiques/rema-tp-transportpublic/etat-des-lieux-et-benchmarking>.
- Delatte, A., 2018. Women in Transport International outlook, Workshop 3: Smart Mobilities, Planning and policy processes - Gender and diversity mainstreaming, Nos-Sh workshop series, 17–18 September, VTI, Stockholm, Sweden.
- de Jong, T., Fyhri, A., Priya Uteng, T., 2018. Gender gap—Perception of safety and cycling use, Workshop 2: Smart Mobilities—Walking and Biking, Nos-Sh workshop series, 21 April, University of Copenhagen, Copenhagen, Denmark.
- European Parliament, P.D. C., 2012. The Role of Women in the Green Economy—the Issue of Mobility.
- Garrard, J., Rose, G., Lo SK, 2008. Promoting transportation cycling for women: the role of bicycle infrastructure. *Prevent. Med.* 46 (1), 55–59.
- Hanson, S., 2010. Gender and mobility: new approaches for informing sustainability gender, place and culture. *J. Femin. Geogr.* 17 (1), 5–23.
- Haustein, S., Jensen, A.F., 2018. Factors of electric vehicle adoption: a comparison of conventional and electric car users based on an extended theory of planned behavior. *Int. J. Sustain. Transport.* 12 (7), 484–496, doi:10.1080/15568318.2017.1398790.
- Heinen, E., Chatterjee, K., 2015. The same mode again? An exploration of mode choice variability in Great Britain using the National Travel Survey. *Transport. Res. A Policy. Pract.* 78, 266–282.
- Hine, J., Mitchell, F., 2001. The Role of Transport in Social Exclusion in Urban Scotland. Scottish Executive Central Research Unit. Available from: <http://www.scotland.gov.uk/Resource/Doc/156591/0042062.pdf>.
- Hjorthol, R.J., 2008. Daily mobility of men and women—a barometer of gender equality? In: Uteng, T.P., Cresswell, T. (Eds.), *Gendered Mobilities*, Ashgate.
- Hossain, M., Susilo, Y.O., 2011. The exploration of rickshaw usage pattern and its social impacts in Dhaka city Bangladesh. *Transport. Res. Rec.* 2239, 74–83.
- International Finance Corporation (IFC) and Accenture, 2018. *Driving Toward Equality: Women, Ride-Hailing, and the Sharing Economy*. IFC, Washington D.C.
- Kawgan-Kagan, I., 2015. Early adopters of carsharing with and without BEVs with respect to gender preferences. *Eur. Transp. Res. Rev.* 7 (4), 1–11, doi:10.1007/s12544-015-0183-3.
- King, R., 1980. Residential location, travel mode and income: the likely effects of changes in housing cost and transport cost in Melbourne. Australian Road Research Board Internal Report 311-1. Available from: <http://trid.trb.org/view.aspx?id=1207067>.
- Krizek, K.J., Johnson, P.J., Tilahun, N., 2005. Gender differences in bicycling behavior and facility preferences. In: *Research on Women's Issues in Transportation*, Rosenbloom, S. (Ed.), Transportation Research Board, Washington, DC, pp. 31–40.
- Kwan, M., 1999. Gender and individual access to urban opportunities: a study using space-time measures. *Profes. Geogr.* 51 (2), 210–227.
- Lam, T., 2019. Cycling London: an intersectional feminist perspective. In: Priya Uteng, T., Levin, L., Rømer Christensen, H. (Eds.), *Gendering Smart Mobilities*, Routledge, Abingdon and New York.
- Lam, T., Cosgrave, E., 2019. *Gender Inclusive Climate Action in Cities: A report for C40 Cities Women4Climate initiative*.
- Law, R., 1999. Beyond 'women and transport': towards new geographies of gender and daily mobility. *Progr. Hum. Geogr.* 23 (4), 567–588.
- Lenz, B., 2019. Smart mobility—for all? Gender issues in the context of new mobility concepts, paper presented in Workshop 1: Gendering Smart Mobilities, Nos-Sh workshop series, 24–25 August, Institute of Transport Economics TØI, Oslo, Norway.
- Li, X., Chen, P., Wang, X., 2017. Impacts of renewables and socioeconomic factors on electric vehicle demands—Panel data studies across 14 countries. *Energy Policy* 109, 473–478, doi:10.1016/j.enpol.2017.07.021.
- Loose, W., 2010. Aktueller Stand des Car-Sharing in Europa. Endbericht D 2.4 Arbeitspaket 2; Bundesverband CarSharing e.V. Available from: http://www.carsharing.info/images/stories/pdf_dateien/wp2_endbericht_deutsch_final_4.pdf.
- METPEX, 2013. Identification of User Requirements Concerning the Definition of Variables to be Measured by the METPEX Tool (Deliverable 2.3 of the METPEX project), EU Commissions.
- Paleti, R., Pendyala, R.M., Bhat, C.R., Konduri, K.C., 2011. A joint tour-based model of tour complexity, passenger accompaniment, vehicle type choice, and tour length. Technical paper, Arizona State University, Tempe, AZ.
- Priya Uteng, T., 2006. Mobility: discourses from the non-western immigrant groups in Norway. *Mobilities* 1 (3), 435–462.
- Priya Uteng, T., 2012. Gender and mobility in the developing world: Gendered bargains of daily mobility—Citing cases from both urban and rural settings. *World Development Report*, World Bank.
- Priya Uteng, T., Espegren, H.M., Throndsen, T.S., Bocker, L., 2020. The gendered dimension of multi-modality: exploring the bike sharing scheme of Oslo. In: Priya Uteng, T., Levin, L., Rømer Christensen, H. (Eds.), *Gendering Smart Mobilities*. Routledge, Abingdon and New York.
- Pucher, J., Dill, J., Handy S, 2010. Infrastructure, programs, and policies to increase bicycling: an international review. *Prevent. Med.* 50, S106–S125.
- Rezvani, Z., Jansson, J., Bodin, J., 2015. Advances in consumer electric vehicle adoption research: a review and research agenda. *Transport. Res D Transp. Environ.* 34, 122–136, doi:10.1016/j.trd.2014.10.010.
- Scheiner, J., Sicks, K., Holz-Rau C, 2011. Gendered activity spaces: trends over three decades in Germany. *Erdkunde* 65 (4), 371–387.
- Scheiner J, 2013. Gender roles between traditionalism and change: time use for out-of-home activities and trips in Germany, 1994–2008. In: Gerike, R., Hülsmann, F., Roller, K. (Eds.), *Strategies for Sustainable Mobilities: Opportunities and Challenges*, Farnham, Ashgate, pp. 79–102.
- Scheiner, J., Holz-Rau, C., 2017. Women's complex daily lives: a gendered look at trip chaining and activity pattern entropy in Germany. *Transportation* 44 (1), 117–138.
- Scheiner, J., 2018. Gender, travel and the life course. Thoughts about recent research, Workshop 3: Smart Mobilities, Planning and policy processes—Gender and diversity mainstreaming, Nos-Sh workshop series, 17–18 September, VTI, Stockholm, Sweden.
- Shaheen, S., Chan, N.D., Micheaux, H., 2015. One-way carsharing's evolution and operator perspectives from the Americas. *Transportation* 42 (3), 519–536, doi:10.1007/s11116-015-9607-0.
- Sovacool, B.K., Kester, J., Noel, L., de Rubens, G.Z., 2018. The demographics of decarbonizing transport: the influence of gender, education, occupation, age, and household size on electric mobility preferences in the Nordic region. *Global Environ. Change* 52, 86–100, doi:10.1016/j.gloenvcha.2018.06.008.
- Susilo, Y.O., Maat, K., 2007. The influence of built environment to the trends in commuting journeys in the Netherlands. *Transportation* 34, 589–609.
- Susilo, Y.O., Dijst, M., 2009. How far is too far? Travel time ratios for activity participations in the Netherlands. *J. Transport. Res. Rec.* 2134, 89–98.
- Susilo, Y.O., Cats, O., 2014. Exploring key determinants of travel satisfaction for multi-modal trips by different travellers' groups. *Transport. Res. A* 67, 366–380.
- Susilo, Y.O., Liu, C., Börjesson, M., 2018. Examination across gender, life-cycle, and generations: the changes of activity-travel participation in Sweden over 30 years, Workshop 3: Smart Mobilities, Planning and policy processes—Gender and diversity mainstreaming, Nos-Sh workshop series, 17–18 September, VTI, Stockholm, Sweden.
- Svanfelt, D., 2019. Some gender equality and equity planning cases in urban planning in Malmö, or How I became a transport feminist, Workshop 3: Smart Mobilities, Planning and policy processes—Gender and diversity mainstreaming, Nos-Sh workshop series, 17–18 September, VTI, Stockholm, Sweden.
- Tarigan, A.K., Susilo, Y.O., Joewono, T.B., 2014. Segmentation of paratransit users based on service quality and travel behaviour in Bandung, Indonesia. *J. Transport. Plan. Technol.* 37 (2), 200–218.
- Tilley, S., Houston, D, 2016. The gender turnaround: young women now travelling more than young men. *J. Transp. Geogr.* 54, 349–358.
- Transport for London, 2009. Customer touch points needs and gaps, Part 1: overview. Available from: <http://www.tfl.gov.uk/cdn/static/cms/documents/Travel-in-London-report-1.pdf>.
- Transport for London, 2012. Annual report and statement of accounts. Available from: <https://www.tfl.gov.uk/cdn/static/cms/documents/annual-report-and-statement-of-accounts-2013.pdf>.
- Waygood, E.O.D., Susilo, Y.O., 2011. Distinguishing universal truth from country-specific influences on children's sustainable travel: a comparison of children in the UK and the Osaka regional area. *J. EastN. aSoc. Transp. Stud.* 9, 271–286.

Further Reading

Duchène, C., 2011. Gender and Transport. Discussion Paper No. 2011-11. International Transport Forum, Leipzig.

McKinsey and Company, 2015. The power of Parity: How advancing women's equality can add \$12 trillion to global growth. Available from: www.mckinsey.com/mgi.

Tranter, P., 1995. Behind Closed Doors: Women, Girls and Mobility. Paper presented at the Women and Transport Conference, Adelaide, Australia.

Transport Modes and Big Data

Hannah D Budnitz^{*}, Emmanouil Tranos[†], Lee Chapman[‡], ^{*}Transport Studies Unit, School of Geography and the Environment, University of Oxford, Oxford, Oxfordshire, United Kingdom; [†]School of Geographical Sciences, University of Bristol, Bristol, United Kingdom; [‡]School of Geography, Earth & Environmental Sciences, University of Birmingham, Birmingham, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	665
The V's of Big Data	665
Transport Big Data	666
Public Transport	666
Private Transport	666
Freight	667
New Modes and Systems of Mobility	667
Big Data from ICT Sources	668
Mobile Phones	668
Social Media and Location-Based Services	668
Summary and Conclusion: Implications for Uncertainties	669
Acknowledgment	669
See Also	669
References	670
Further Reading	670

Introduction

Transport has long been a data-driven discipline. The requirements for transport to operate efficiently in time and space in order to fulfil its economic and societal purposes have propelled the collection of ever-growing quantities of data, and often its regulation and standardization. The standardization of time zones to regulate rail timetables is a well-known example. The transport industry also historically stores large amounts of data, from the geographical coordinates of every bus stop in a network to the outputs of large traffic counts and surveys. However, the size of a database or dataset is not the defining characteristic of what has come to be known as “big data.” Rather, this entry defines big data as:

Data that are generated automatically and collected digitally, rather than collected manually and then compiled in digital form. Data that can be described and assessed using a number of terms beginning with “V”: Volume, Velocity, Variety, Veracity, and Value.

In recent years, more transport-related data meet this “big data” definition, from electronic ticketing transactions to Automatic Number Plate Recognition (ANPR) cameras. The purposes of such data are usually operational, but data can also provide insights for transport planning and strategy. Meanwhile, the Information and Communication Technology (ICT) sector also generates big data that can be applied to the study of mobility, from the Call Detail Records (CDRs) that mobile phone companies catalogue in order to bill their customers to geo-tagged social media posts.

This paper is structured as follows: Section VThe V's of Big Data describes the defining characteristics of big data; Section Transport Big Data reviews the many sources of big data from the transport sector, including public transport, private transport, the freight industry, and newly emerging modes, and business models; Section Big Data from ICT Sources explores big data from ICT sources, such as mobile phones or applications on mobile and fixed internet services; and Section Summary and Conclusion: Implications for Uncertainties concludes this entry.

The V's of Big Data

The first “V,” Volume is the most obvious characteristic of big data. Big datasets are so large that they can be uncomfortable to conceive of and require extensive computing memory to manage. In the past, such volumes of data would have been difficult and expensive to collect and store, and retention of data over time was limited. The exponentially growing capacity of computers to file information on ever smaller drives, and the introduction of cloud computing make it possible to store big data, often through automated processes. Big data can then be retained until assessed and analyzed, which large increases in computing power and reductions in the cost of this power make possible. This inexpensive capacity is necessary not only to handle the volume of data, but also the second “V” characteristic, its Velocity. Big data are being generated ever faster through automated means, with unimaginable quantities added each second. Translating the data into use-able information in real time is a growing challenge, involving complex analytical methodologies.

The third “V” is Variety. Big data are not only numbers. Big data include sound, images, text, sensor readings of weight, speed, temperature, or environmental parameters such as pollution, or detections of an object or event. Big data can be qualitative or quantitative, official or crowd-sourced, and are generated to fulfil a wide array of functions prior to any use for research or other secondary purposes. The challenges such variety presents for analysis raise the specter of the fourth “V,” Veracity. Veracity refers to how transparent different datasets are in their source, independence, and biases, and how accurately any inconsistencies can be overcome when the data are applied or analyzed. Extracting useful, trustworthy information can be challenging when working with big data.

Which leads to the fifth and final “V,” Value, which is not always included, but raises a number of questions: Can big data be transformed into not only information, but also intelligence? Are particular big data sources worth keeping, processing, and analyzing, in financial and in ethical terms? Will the use of big data lead to more sustainable outcomes, for the economy, environment, and society? Can big data improve the practice of transport planning and operation, and the infrastructure and accessibility delivered? The rest of this entry will consider the types and uses of big data in the field of transport and aim to provide some brief, qualified answers to the above questions.

Transport Big Data

Transport big data are often generated and/or owned by transport organizations, which employ the data in order to fulfil their statutory, regulatory, or commercial functions, including delivering the basic business of mobility and logistics services and infrastructure. The organizations involved include public transport operators, haulage companies, those delivering public sector contracts on the highway network, vehicle manufacturers, and technology entrepreneurs aiming to capture a proportion of the transport market.

Public Transport

The growth of big data in the public transport sector is one of the most noticeable, as it has a public-facing aspect. While train, bus, and tram operators still publish timetables and register bus stops or map stations and other assets, many also collect data on the precise location of their services at regular intervals or at a continuous velocity. These real time information systems use Global Positioning Systems (GPS) fixed to the vehicles and to sensors at stops and stations, and often feed into screens and applications to provide passengers with more accurate information for journey planning, travel alerts, and wayfinding. The veracity of this information is high, as such systems are often designed for communicating with passengers, and research suggests that the provision of reliable dynamic information, particularly during disruptions, improves passengers’ perceptions of service quality (Pender et al., 2014). Some public transport networks have additional technology to monitor vehicle or carriage occupancy, ambient comfort, and driver behavior. As well as providing information to the customers, the variety and volume of data such systems create can be used to improve end-to-end journey times of services, smooth out headways, reduce bunching, track any disruptions or diversions, train drivers in fuel efficiency, and inform performance monitoring of delays and punctuality. Closed Circuit Television (CCTV) is also commonly used on public transport to increase security, and although such systems are often still monitored manually, new data science methodologies can be applied to image-based big data.

Electronic ticketing systems are also a major source of big data in public transport. These transactions are captured and retained for commercial purposes. However, such data have the potential to inform analyses of public transport demand across a network by time of day, season, and particular events. The data usually can be analyzed by location of boarding. Sometimes the location of alighting is also recorded or can be deduced from fare structures and timestamps. Electronic ticketing data also offer insights into the travel patterns of different passenger groups depending upon the variety of ticketing options—commuters using season tickets, youth or senior concessions, the use of a local smartcard versus an internationally recognized bankcard or mobile payment. These data inform the accounts of a public transport operating company, but can also provide intelligence on the integration of public transport networks with different land uses, which can improve the efficiency and commercial success of services through adjusting stops, frequency, and vehicle capacity. Such adjustments could be targeted to increase social equity and inclusivity, market new services, or respond more effectively to planned or unplanned events.

Private Transport

Highways authorities have a responsibility to manage their road networks and infrastructure, and thus have automated and digitized data collection both to enforce any restrictions, such as on speed or parking, and also to levy charges, such as in low emission or congestion zones. These data come in not only large volumes, but also a wide variety of forms. ANPR cameras identify individual vehicles and their movements and are often placed at junctions and key points on the network, whether as a cordon or at nodes along particular links. Other nodal data are produced by environmental sensors that automate the illumination of streetlights, monitor air pollution, or inform winter gritting rotas. As in public transport settings, CCTV may be installed for security purposes, but captures images of all road users, and thus has the potential to measure categories of transport demand at a particular location using visual recognition and machine learning methods. In urban settings, CCTV and other types of sensors identify both individuals and vehicles. Bluetooth sensors on transport infrastructure are device-based, and generate big data including timestamps and

geolocations on almost every enabled vehicle or individual device that passes. Although this does create sampling biases toward newer vehicles and younger segments of the population who hold the relevant devices and keep such connections active, such sensors can capture data that are limited in more traditional cross-sectional surveys. For example, one scheme introduced Bluetooth sensors to record travel times along particular routes, but the data were also used in research to identify temporal variability of travelers along those routes (Crawford et al., 2018).

Many of these road network technologies have become part of what is known as the Internet of Things (IoT) where cameras and sensors monitor conditions and send data back to a central control system at high velocity, often in real time. Private cars have also become part of the IoT, using vehicle telematics that communicate online to enable the operation and regular updating of features such as satellite navigation. Complex on-board computers record not just mileage, but operational aspects like speed, fuel efficiency, engine, and gear optimization, as well as offering assistance in vehicle maneuvers, using sensors and cameras for obstacle detection, lane changes, and parking. They also provide additional security and remote access to a vehicle. Many of these electronic features on private vehicles are not currently linked to the IoT nor do they necessarily save the data they generate about driving behavior over time, but the option for them to do so is available. For example, insurance schemes for younger drivers offer discounts to those drivers (or their parents) who are willing to have data on their driving behavior monitored remotely as a means to reduce accident rates and improve road safety among this vulnerable demographic.

Freight

Remote monitoring of driver activities is of particular importance to the freight industry. Freight vehicles are usually equipped with GPS and other intelligent transport systems (ITS) or vehicle telematics that track the movements and journey times of the vehicle. The freight industry also has automated means of collecting data on where inventory is loaded and unloaded, vehicle speeds, routing, weight, and driver behavior, including fuel consumption and length of time between rests. This volume, velocity, and variety of data are essential in increasing the efficiency of an industry that is dependent upon just-in-time deliveries and small margins. Thus, for operational and commercial reasons, goods vehicles are often at the forefront of Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) technology, creating connected networks that enable more precise fleet management, improved driver and road safety, and the development of platooning, remote and automated control of vehicles.

Indeed, due to its innovative nature and the investment required, V2I systems tend to be installed on strategic road networks serving freight and long distance traffic before being implemented on local roads. On these networks, traffic count loops may be accompanied by sensors able to classify vehicles by weight, length and other characteristics, which are important to those responsible for managing the road network, for example, to enforce weight restrictions. Such systems have also informed early uses of big data in network management, providing infrastructure authorities with intermediate data on loads/flows on their network between fixed monitoring points. Some data are even used to determine the accessibility of major destinations or journey times to important junctions, which is then broadcast through the highway network's real time information system, that is, electronic, variable message signing (VMS). However, with the spread of vehicle telematics in not just the logistics, but also the private mobility sector, this information has begun to be relayed directly into vehicles' GPS and the satellite navigation displays.

New Modes and Systems of Mobility

Although there are some biases in the analytical methodologies and resources employed to sort through the volume, velocity and variety of big data and derive value, the sources of data described above are rarely of questionable veracity. The IoT that are built into transport networks and services are usually installed by operators or responsible parties and validated by official sources to inform network performance or service demand. Their proliferation can be seen as part of the current pace of change in the transport sector as it accommodates the increasing automation of mobility as well as logistics, and potentially a future of connected autonomous vehicles (CAVs). Even if fully driverless cars are not imminent, by creating a connected infrastructure network, additional channels of more personalized communication and advice are available to both those managing the roads and those using them, which research suggests would be welcome and potentially more effective than existing road network channels, such as VMS (Chorus et al., 2006).

The wealth of big data has also driven the creation of new transport services, such as public transport that can vary routes and timetables according to demand, ride-hailing services, car clubs, shared bicycles, and electric scooters. Most of these vehicles produce a wide variety of big data from in-built systems that track the vehicles either continuously or at key nodes such as pick-up and drop-off points. They utilize digital platforms that enable customers to journey plan, book, pay for, and track their trips. The business models of these new transport services are often founded upon the volume and velocity of big data, which allows them to provide an efficient service even where demand is constantly shifting in space and time. The personalization of transport services is at the other end of the spectrum from the regularization and standardization that previously drove the collection of transport data and traditional transport service and infrastructure design. Thus, big data are the foundation for any "mobility as a service" (MaaS) business model.

Likewise, other big data sources in transport, such as journey planners, customer feedback channels, and transport applications designed to promote loyalty, rewards, fitness, social interaction, or status-sharing potentially have value in the development of future, personalized transport services. New transport services depend upon using tracking and feedback data from their customers to build their businesses. However, such data are more difficult to verify. Applications generate "crowd-sourced" big data from anonymous members of the public, who may plan a journey but not make one, share news of congestion or crowding that others

would consider a normal level of service for that route, or claim a journey made by a mode they did not take. Therefore, such data will have sampling and other biases that require careful analysis before being used to describe wider transport trends or issues or being applied to planning and strategy purposes. For example, commercial applications with a fitness market will not collect data on a representative sample of those who use active travel modes for utilitarian purposes, and privacy policies render contextual information such as journey purpose and socio-demographic profile unavailable (Romanillos et al., 2016). Still, big data with time-stamps and geolocations may offer useful insights into mobility, accessibility, and travel demand, particularly when linked to more randomized or validated locational data or where the volume of data points enable statistical methodologies to correct for certain biases.

Big Data from ICT Sources

Greater volumes of high velocity, volunteered geographic information are available from ICT sources, many of which have lower sampling biases than big data collected by the transport sector, yet have tremendous utility to generate insights about mobility. For example, mobile phones are now almost ubiquitous among the working adult population in many countries. Likewise, electronic financial transactions are more common than cash in some geographies. ICT sector data are not collected for transport purposes and cannot necessarily be linked to flows on the transport networks, and there are also various interpretive challenges in terms of spatial and temporal granularity. However, such data does have value for transport modeling and appraisal and for various strands of research into travel behavior, event management, accessibility/equity issues, and urban service delivery. Both transport and ICT data tend to be collected for operational or commercial purposes, not modeling and research, but there is speculation that ICT big data could offer similar or even more comprehensive intelligence at a lower cost. Issues around biases, however, and also conflicts between privacy/commercial sensitivity and the openness of data, mean this is far from certain.

Mobile Phones

Mobile phones have generated the most interest among modelers and researchers seeking big data from non-transport sources. Coverage can be close to population level, and CDRs are created automatically whenever a mobile phone is used to make or receive a voice call or text or, for smartphones, to download data. Further details automatically collected by the mobile operators include when a phone is switched on or regains signal, when it moves between clusters of cell towers, when it switches between signal generations, for example, 3G to 4G, and periodic checks on a device's location to ensure service. The sum of such logs are sometimes known as Mobile phone Network Data (MND). These records have timestamps, but provide spatial information at a scale dependent upon the density of the cell network, which is more granular in urban areas, particularly if any records on WiFi use are incorporated into the dataset. Although the phone may not be in use during travel, the data do provide a clear record of users' locations when not moving, from which movement can be derived. Places that are regularly visited for long periods of time, such as home and work are fairly straightforward to identify and validate. From such data, researchers derive aggregate commuting distances and patterns, CO₂ emissions, and the density of central business district workers drawn from particular neighborhoods in a metropolitan area (Becker et al., 2013).

However, there are substantial gaps in the data and purposes for which such data may not be suited. There may be no records made during shorter trips, particularly on foot, or during stops made as part of linked trips. The routes taken for rural trips will be less precise due to the less dense network of cells. In addition, there may be inaccuracies in matching trips to routes and speeds to modes. Although separating road from heavy rail can be accomplished with some confidence, it is much more difficult to separate and validate different modes used on local, urban roads, such as bicycle, bus, car, and taxi. Aspects like vehicle type and occupancy are also inferred, rather than surveyed directly, and assumptions about mobile network users must be made in order to scale up the data to population level. Finally, as data owners, mobile operators recognize the value of their big data and the importance of privacy to their customers, and therefore may charge for any use, and insist on anonymizing and aggregating or pre-processing their data at their discretion. Despite those caveats, the benefits of covering a larger geographical area and sample size, as well as capturing longitudinal rather than cross-sectional data, should not be underestimated. Other datasets, including socio-demographics, can both improve the analysis and reduce the gaps in this key ICT data source.

Social Media and Location-Based Services

While some datasets that can be matched to mobile phone data are derived from the transport big data, or from surveys and censuses, other ICT data also provide more spatially granular records of movements, through volunteered geographic information. Some users choose to have their social media posts geotagged, or they may agree to their movements being tracked in order to benefit from location-based services that advertise nearby amenities. Some applications provide both social interaction and location-based services through gaming or arranging meet-ups. Many of these applications incorporate Application Programming Interfaces or APIs, which provide open access to the big data generated in real time. The variety of this data matches its volume and velocity, as text, images, and various other content accompanies the timestamp and geolocation. Veracity is also in question, as GPS may be switched on and off, and such applications tend to be favored by younger, more urban, and more educated populations.

However, these data provide valuable channels for analyzing activity patterns and augmenting other data sources, including traditional surveys, which consistently underestimate the number of short trips made on foot and struggle to account fully for variable travel patterns. For example, using geolocation, rating, and “check-in” data on food-based activity places combined with transport big data on traffic congestion, one study was able to measure the temporal variability of the attractiveness and accessibility of such locations (Wang et al., 2018). Personalized ICT datasets also add value in terms of economic and social behaviors, as they can be used to provide newsfeeds and alerts, inform market research for the delivery of online access, nudge changes in travel behavior, and offer channels for feedback. Indeed, if the purpose of transport modes is to offer choices in accessibility instead of mobility, then online access should be included in the mix, and ICT data can be used for the direct measurements of Internet access, activity, and quality of service. Yet transport practitioners are rarely well-versed in the geography and geometry of ICT networks, and how, similar to road or public transport networks, they offer varying levels of service based upon where and when access is required over what type of connection and for what purpose. The overarching purpose is transfer of information, but ICT access is affected by the volume, velocity, and variety of data being transmitted, and understanding the structure and operation of ICT networks better is essential to effectively using either the datasets generated by ICT and online services or those generated by transport modes and their users.

Summary and Conclusion: Implications for Uncertainties

The study and sector of transport are subject to more change at the present time than they have seen for decades, due to new digital technologies, new transport technologies, an aging population, more flexible working patterns, and more extreme weather, to name a few. All these factors affect not only the supply of transport infrastructure and services, but also the structure of demand. In many cases, their impact on the future of transport is still uncertain. Previous assumptions regarding averaged travel patterns or routine maintenance regimes, based often on annual surveys, offer limited insight into a variety of potential futures and limited ability to respond robustly to systemic shocks. Yet big data have the potential to provide the spatial and temporal granularity to fill knowledge gaps, more accurately depict changing societies and environments, and prepare for an uncertain future. Big data might offer more insight into individual behavior, but using data science techniques can also provide insight into the resilience of human mobility more broadly, for example, during natural disasters (Wang and Taylor, 2016).

Big data from various transport modes and other sources can provide transport operators and managers with the means to improve the efficiency or safety of their networks. Transport modelers and policy-makers can achieve more variable and locally customized outputs from big data that offer more realistic representations of the transport supply or demand being assessed. These outputs can be translated into knowledge that can be applied to solve problems, make policy, or predict the likelihood of future scenarios. However, big data can only be successfully applied if transport researchers and practitioners have the resources and skills to manage the volume, velocity and variety of big data, and the understanding to appropriately question the veracity and improve the value of big data through linking with other traditional and emerging data sources. Regulation and statute has struggled to keep pace with the issues involved in the collection, storage, access, and analysis of big data. Data generators, both infrastructure and service operators as well as individuals using new transport and ICT services, have legitimate privacy and security concerns, while data holders or owners are concerned about commercialization and competition. Even as new digital technologies and methodologies make the use of big data possible, data do not become information until analyzed, nor intelligence until the messages derived can be communicated to an audience beyond the transport discipline or sector. Thus, this article aims to provide an introduction to transport modes and big data, and the risks and benefits of using big data to contribute to the future of transport study and practice.

Acknowledgment

This work was supported by the Natural Environment Research Council (grant number NE/M009009/1), and the Economic and Social Research Council, as part of the centre for doctoral training on Data, Risk, and Environmental Analytical Methods which the authors were affiliated with through the University of Birmingham.

See Also

Automobile Accidents and Passive Prevention Systems; Logistics Information Systems; Electronic Toll Collection; Road Pricing – Theory and Applications; Road Pricing 2: Short- and Long-Run Equilibrium of Road Transportation; Road Pricing 3: The Implications for Pricing Public Transportation; Traffic Incident Detection; Advanced Travelers Information Systems (ATIS); Parking Information Systems; Transit Information Systems; Computational Methods and Data Analytics; Advanced Traveler Information Systems; The Use and Value of Geographic Information Systems in Transportation Modeling; Transportation Statistics and Databases; Travel Demand Forecasting; Where Are We and What Are the Emerging Issues; Transportation Data Sources and Modes; Big Data for Public Transport Planning; Adoption of New Travel Information Platforms; ICT and Transport Modes; Safety Data Quality Management; Mobility as a Service; Technology Enabled Data for Sustainable Transport Policy

References

- Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J.M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., Volinsky, C., 2013. Human mobility characterization from cellular network data. *Comm. ACM* 56, 74–82, doi:10.1145/2398356.2398375.
- Chorus, C.G., Molin, E.J.E., Van Wee, B., 2006. Use and effects of advanced traveller information services (ATIS): a review of the literature. *Trans. Rev.* 26, 127–149, doi:10.1080/01441640500333677.
- Crawford, F., Watling, D.P., Connors, R.D., 2018. Identifying road user classes based on repeated trip behaviour using bluetooth data. *Trans. Res. Part A Policy Prac.* 113, 55–74, doi:10.1016/j.tra.2018.03.027.
- Pender, B., Currie, G., Delbosc, A., Shiwakoti, N., 2014. Social media use during unplanned transit network disruptions: a review of literature. *Trans. Rev.* 34, 501–521, doi:10.1080/01441647.2014.915442.
- Romanillos, G., Austwick, M.Z., Ettema, D., De Kruijff, J., 2016. Big data and cycling. *Trans. Rev.* 36, 114–133, doi:10.1080/01441647.2015.1084067.
- Wang, Q., Taylor, J.E., 2016. Patterns and limitations of urban human mobility resilience under the influence of multiple types of natural disaster. *Plos One* 11, 1–14, doi:10.1371/journal.pone.0147299.
- Wang, Y., Chen, B.Y., Yuan, H., Wang, D., Lam, W.H.K., Li, Q., 2018. Measuring temporal variation of location-based accessibility using space-time utility perspective. *J. Trans. Geogr.* 73, 13–24, doi:10.1016/j.jtrangeo.2018.10.002.

Further Reading

- Arribas-Bel, D., 2014. Accidental, open and everywhere: emerging data sources for the understanding of cities. *Appl. Geogr.* 49, 45–53, doi:10.1016/j.apgeog.2013.09.012.
- Blazquez, D., Domenech, J., 2018. Big Data sources and methods for social and economic analyses. *Technol. Forecast. Soc. Change* 130, 99–113, doi:10.1016/j.techfore.2017.07.027.
- Corbett, C., 2018. How sustainable is big data? *Prod. Operat. Manag.* 27, 1685–1695, doi:10.1111/poms.12837.
- Cottrill, C., Derrible, S., 2015. Leveraging big data for the development of transport sustainability indicators. *J. Urban Technol.* 22, 45–64.
- Cottrill C., 2018. Data and digital systems for UK transport: change and its implications. In, Government Office for Science, (ed). *Future of Mobility: Evidence Review*.
- Cuauhtemoc, A., Erath, A., Fourie, P.J., 2017. Transport modelling in the age of big data. *Int. J. Urban Sci.* 21, 19–42, doi:10.1080/12265934.2017.1281150.
- Graham, M., 2013. Geography/internet: ethereal alternate dimensions of cyberspace or grounded augmented realities? *Geogr. J.* 179, 177–182, doi:10.1111/geoj.12009.
- Lovelace, R., Birkin, M., Cross, P., Clarke, M., 2016. From big noise to big data: toward the verification of large data sets for understanding regional retail flows. *Geogr. Anal.* 48, 59–81, doi:10.1111/gean.12081.
- Ricci, F., Pearce, C.T., Reeves, V., Yee, V., 2012. *Intelligent Transport Systems (ITS) for Sustainable Mobility*. United Nations Economic Commission for Europe, Geneva.
- Speranza, M., 2018. Trends in transportation and logistics. *Eur. J. Operat. Res.* 264, 830–836, doi:10.1016/j.ejor.2016.08.032.
- Steenbruggen, J., Tranos, E., Nijkamp, P., 2015. Data from mobile phone operators: a tool for smarter cities? *J. Telecommun. Policy* 39, 335–346, doi:10.1016/j.telpol.2014.04.001.
- Wang, Z., He, S., Leung, Y., 2017. Applying mobile phone data to travel behaviour research: a literature review. *Travel Behav. Soc.*, doi:10.1016/j.tbs.2017.02.005.

Transport Modes and Inequalities

Caroline Mullen, Institute for Transport Studies, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	671
Why Transport Matters for Equality	671
Why Does Equality Matter?	671
Equality and the Benefits and Values of Different Transport Modes	672
Risks and Harms of Different Transport Modes	672
Road Traffic Collisions	672
Pollution and Land Use	673
Climate Change	673
Assessing Inequalities Associated with Transport Mode	674
Tackling Inequalities Related to Transport Modes	674
Summary and Conclusions	675
Biographies	676
See Also	676
References	676
Further Reading	677

Introduction

Transport matters for most aspects of life, and so tackling transport inequalities can be vital for a just or fair society. There are more forms of transport mode inequality than could possibly be listed. Yet perhaps not all inequalities matter, and even where they do matter, they cannot all be tackled. So priorities need to be determined. To make sense of inequality and transport we need some conception of why equality matters at all. This allows assessment of what transport inequalities are relevant, which trivial or unimportant, or even which would be required within a just mobility system. This in turn enables development of means of assessing whether there are relevant transport related inequalities in a given context, and if there are, what if anything, could or should be done to tackle them. Given this, the article is structured as follows. The next section begins by sketching a high-level conception of equality, which is sufficient to guide assessment of transport mode equalities, and then gives a narrative account of major forms of inequalities associated with transport mode. Following this, the second section considers approaches planners and researchers can take in assessment of inequalities related to transport mode in different contexts. The section outlines the rationale for assessment measurement and indicator. It discusses the challenges of formulating indicators and other assessment in ways which are feasible given available data and limited resources, but which do not lead to perverse outcomes in which relevant inequalities are overlooked, or where some inequalities risk being exacerbated. Finally, the third section presents considerations for tackling the complex, and interrelated forms of inequality, which can be created by transport use. As we see through this article, transport mode inequalities are frequently context specific, but there are some broad considerations, which can help guide efforts to reduce transport mode inequalities.

Why Transport Matters for Equality

Why Does Equality Matter?

A good starting point for assessment of equality is to ask “equality of what?” (Cohen, 1990). It is sometimes said that equality is about treating people the same. If there were no more to it than this, then it would not be a very helpful, or plausible concept, for assessing social and economic organization, or indeed for assessing other aspects of life and the way in which we treat one another. In the numerous theories of justice or fairness which exist and in which equality is central, the idea of equality is never given such a superficial meaning. A conception of equality common to many, although not all, theories of justice is that equality fundamentally matters because:

Each person matters, simply in virtue of being a person. This means that:

- their life matters and
- their quality of life and welfare matters.

This conception is compatible with most theories of distributive justice, including consequentialist theories, communitarian theories, and rights-based theories. As we would expect, interpretation of this conception differs within different theories, however, the conception by itself is sufficient to guide discussion of equality and transport modes. In this article we will use the conception rather than adopt a full theory, since prospects for consensus about the conception are far greater than those for consensus over any given theory.

Equality and the Benefits and Values of Different Transport Modes

We all benefit from the mobility system, whether or not we are actually traveling, since it is a vital component of systems of production, and distribution of goods and services. Yet, as is well recognized, there are enormous inequalities in distribution of those benefits, and this is one sense in which we see inequalities associated with use of transport. If each person matters, then of course we might say we are better off with the transport system than without it, however, this leaves a major question about whether and how changing the way in which we use transport could mean those benefits are more evenly distributed.

For individuals, families, and social groups, the ability to move around—and the ease and safety with which this can be done—can be crucial for possibilities of social and economic activities, including education work, organizing family and social life, for communities, for participation in society and politics, and for numerous other activities which people find valuable. It is worth keeping in mind that barriers to activities found in one part of life could have long far reaching implications: for instance, consider the implications if someone does not access to the transport they need for education. To assess how transport use can have a role in equalities in relation to these benefits of transport, we should keep in mind the number of points:

- First that mobility is not simply a matter of traveling in a vehicle, it can also be movement on foot, and this can be an important part of some social activities. In such cases motorized modes can actually be a barrier to mobility. We see this in [Appleyard's \(1981\)](#) famous study on social interactions between neighbors on lightly trafficked streets and those on heavily trafficked streets.
- Second that the transport modes required for similar types of activities depends on the built and natural environment in which people live and on other differences between people's priorities. So, people in densely populated areas with a safe environment for walking cycling, may rely far less on motorized transport for everyday activities and those in more sparsely populated areas. Recognizing this can be important for assessing equalities and transport, however the relationships are complex. We may need to question whether, or the extent to which, people can choose where they live and so can choose to have everyday lives which rely on differing transport modes ([Mattioli, 2017](#)). Further, as noted earlier, people are not the same and the sorts of activities in which they engage in are unlikely to be the same either: for some people organization of large families may be the priority, whereas for others it may be participation in public services or business, and for others something different again.
- Third that existence of transport is necessary but not sufficient to ensure that people can make use of it. To be usable, transport needs to be responsive to the variations in people's needs and circumstances. This means being physically accessible, available, and affordable. Physically inaccessible transport can create acute inequality with some people becoming socially and economically excluded. Partially accessible transport can also exacerbate inequalities. For instance, if people with disabilities need to make additional arrangements in order to be able to use transport, the time and planning required can place them at a disadvantage. An example is requirements that people with disabilities make arrangements in advance for assistance to use public transport ([Lawson and Matthews, 2005](#)). Severance, which can be caused by vehicular transport infrastructure with inadequate crossings for those using nonvehicular mobility can also be understood as a barrier to accessibility as it can leave people unable to participate in activities involving nonvehicular mobility, or can prevent people reaching and so using available public transport. Partial inaccessibility can also be caused where the rules for using public transport are not easily understood by people. Examples of this would be timetables that cannot be easily read, rules about ticketing which are complex, but which come with stark warnings about penalties for getting anything wrong, and poor information about public transport routes and about where to get on and off.

Availability of transport can be quite a straightforward matter of either owning or having access to private transport, having access to safe environments for nonmotorized transport, or the existence of public transport. Where people rely on nonmotorized transport or public transport, the most visible forms of inequality are associated with unsafe conditions and a lack of services. Even where some services exist, inequalities in people's ability to organize their lives and engage in social and economic activities can depend on the comprehensiveness of those services. Where relatively short journeys take a very long time because public transport services are infrequent or not well integrated then people's ability to engage in activities is reduced. Affordability of transport can also seem a fairly straightforward matter, however, in considering how transport affects equality we might need to take account of the possibility that people can pay for the transport they need but doing so can create other hardships. There is evidence of this in relation to car dependence, including a study indicating that some of those people living with what the European Commission call material deprivation nevertheless maintain ownership of a car. There is a correlation between those in material deprivation who own a car and those living in areas with poor public transport ([Mattioli, 2017](#)).

Risks and Harms of Different Transport Modes

Road Traffic Collisions

The [World Health Organization \(2018\)](#) estimate road deaths at around 1.35 million worldwide, a number which has increased since 2001, although there has been a small reduction in rates of death per 100,000 people, from 18.8 to 18.2 over that time (as the population has increased). As the WHO point out there are apparent inequalities in levels of exposure to risk of road traffic collision between countries, and within countries: In low-income countries the average rate of death is 27.5 per 100,000 population, whereas in high income countries it is 8.3 per 100,000 population. Further, there are variations in proportion of road deaths by mode used by the victim as a percentage of all road deaths. The WHO report that worldwide, on average, "pedestrians and cyclists represent 26% of all deaths, while those using motorized two- and three-wheelers comprise another 28%" and 29% are "drivers or passengers of 4-

wheeled vehicles" (2018, p. 6). The remaining 17% are described as "others" or "unspecified" because "data is not available or is incomplete" (World Health Organization, 2018).

Road traffic deaths and injuries are clearly a matter of concern, however, there is a more complicated question of how they raise particular concerns for inequality associated with transport modes. The high rates of deaths in road traffic collisions, and particularly the far higher rates in low-income than in high-income countries, suggest that the sorts of transport systems which exist are disadvantaging some people more than others. We might consider whether those at risk from road traffic deaths also benefit from the transport that they are using, but given the differences in rates of death associated with low and high income countries it seems difficult to sustain any argument that benefits are in anyway proportionate to the risks faced. The differences in proportions of road deaths by transport mode adds to this concern of inequalities. The proportion of road deaths for pedestrians and cyclists is actually (on average) slightly lower than drivers and passengers of four wheeled vehicles. However, to assess whether this really means there are not disproportionate risks faced by pedestrians and cyclists, we may need to consider what proportion of travel is done by different modes, and we may also want to consider what choice people have about the transport mode that they use. The reasons for suggesting these considerations follow from recognition that people need to travel in order to get on with their lives, and that there are differences in the choices or control that people have over the transport which they use (see above). Yet while we can quite easily, assuming we have the necessary data, compare rates of death on the roads as proportion of a population, the metric for assessing inequalities in risk by transport mode is not straightforward. Should we judge rates of risk according to distance traveled, or according to the number of trips made, by some other method such as length of time spent on the road (Mullen et al., 2014)? If we look at distance traveled as the metric for risk, then we are arguably taking account of people's need to travel certain distances, and of how people may have different levels of choice over the transport they can use. However, we may be overlooking the possibility that some people need to travel further distances in order to go about their activities (again see above). If we look at number of trips, we may be taking account of differences in the distance that people need to travel, but we may still miss the possibility that some people need to make more trips in order to get on with their own lives and contribute to society. Finally, if we look at length of time spent on the road, we might consider whether we are that we are not giving too much account to the risks faced by people in just one part of their life, that is, the part when they are traveling. However, if we use this metric to inform policy interventions seeking reduce inequalities and risks from collision associated with transport mode, it will imply that the policy should focus on safety of those who traveled at high speed rather than on pedestrians or cyclists. The reason for this is that the level of risk/time spent traveling is much higher for a person facing a 1 in x risk of death while traveling 1000 km at 90 km/h on average, compared to one facing the same 1 in x risk of death while traveling 1000 km at 15 km/h. We can question whether this policy focus would be justified by a conception of equality motivated by recognition that each person matters equally. There are further, instrumental reasons, for considering policy responses to different metrics for assessing inequalities in exposure to risks on the roads, and these will be considered in the following sections.

Pollution and Land Use

Poor local outdoor air quality from pollution including nitrogen dioxide, particulate matter, a large proportion of it coming from transport (a proportion which differs in different countries depending a range of factors such as other industrial practices) is held responsible for over a million deaths each year worldwide (World Health Organisation, 2016). Understanding of the relationship between such pollution and health is something which is increasing over time. As more evidence emerges, estimates of deaths attributed may change, as may understanding of the forms of ill health associated with this pollution. The mortality and morbidity attributed to this pollution include excess deaths from cardiovascular problems; it is also found to be a cause of lung cancer (International Agency for Research on Cancer, 2016). Transport pollution can create inequalities as some people's health means they are at greater risk than others risk of suffering from transport pollution. There are also, in some places, correlations between exposure to transport pollution and the levels of households living in poverty (Barnes and Chatterton, 2016).

Along with poor air quality, certain transport modes are associated with noise pollution (Kaltenbach et al., 2016), and there will be inequalities in exposure to this pollution which are associated with proximity to the use of those transport modes. So those living close to busy roads, or near to airports maybe particularly affected. Finally, the land taken for different types of transport infrastructure can have implications for people's lives, and for inequalities. We have already outlined how infrastructure favoring some, especially motorized, modes can have implications for the ability of people to use other, especially nonmotorized modes. Beyond this we may need to consider inequalities associated with compulsory land purchase, or the changing use of what was previously common, or unowned, land for transport infrastructure.

Climate Change

Fuel combustion for transport is responsible for just under one quarter of carbon dioxide emissions worldwide (International Energy Agency, 2018). The evidence of the need to rapidly decarbonize is widely recognized (Intergovernmental Panel on Climate Change, 2018), and there is increasing attention on transport as one of the sectors with a particularly poor record in contributing to decarbonization, with failures to sufficiently improve emissions from, and to reduce demand for, motorized vehicles. In some countries, such as Britain, where there are statutory requirements to decarbonize, levels of carbon from road traffic actually increased slightly over the period 2013 to 2018 (Committee on Climate Change, 2019). There are acute concerns about the implications of climate change for the lives and livelihoods of people across the world now and in future. There is a wide body of work, which indicates that the harm of climate change is unevenly distributed, with less wealthy countries and people being particularly vulnerable.

Assessing Inequalities Associated with Transport Mode

The last section considered how transport modes can be associated with a range of inequalities. What that section also indicates is that understanding those inequalities can be quite a complex matter, requiring us to simultaneously take account of a range of factors (such as the level of choice people have, or benefit that they gain from different transport modes), and where there is scope for discussion about exactly how these different factors should be considered. In that discussion, numerous correlations were used to indicate where we might find inequalities, such as between the levels of road deaths by mode as a proportion of all road deaths. Those correlations do not by themselves tell us the nature of the inequality, and may actually not be revealing a relevant inequality at all, but what they do is suggest where to focus attention. We can add that the correlations can only be identified if the relevant data are collected, so we should keep in mind that we may be missing relevant inequalities due to lack of data. Further, if we are to use an understanding of inequalities associated with transport mode to inform policy, we may need, in addition to a narrative accounts of inequalities, indicators which can be used to illustrate levels and types of inequalities which currently exist, and to inform debate and development of interventions to reduce those inequalities. In other words, there may be a need to make some simplifying assumptions for indicators, but to do this in a way which does not lead to perverse consequences such as missing important inequalities, or creating new inequalities through the use of the indicators to inform policy.

In the rest of this section we briefly outline three examples of methods for assessing transport mode inequalities, and indicate questions which each of these assessments might raise for policy interventions to reduce inequality. One type of indication of transport mode inequalities comes from looking at either the absolute number of trips made by different social groups, or differences in trips made by different mode, particularly by private motor vehicle or by aviation. Given the discussion earlier, we can see that care needs to be taken in drawing inferences about inequalities from the number of journeys made, or the modes used since people have different travel needs in order to fulfill similar sorts of activities. Nevertheless, where we see large variances between journeys or modes used by different social groups, we might investigate whether there are underlying inequalities in opportunities to engage in a diverse range of activities. For instance, are there some groups who are, and some who are not, able to experience international leisure travel or to benefit from international a long-distance business travel ([Banister, 2018](#))?

Second, we can consider correlations between households in poverty and exposure to transport pollution. If we accept that each person's life is equally important then there is little doubt that this is a concern for equality. It has been argued that this form of inequality may be further exacerbated because of evidence that those who faced the higher levels of exposure to pollution are not the ones creating that pollution and so are not directly benefiting from the travel, which creates the pollution. The further question this raises is about who is responsible for this additional inequality—does the responsibility rest with the transport users creating the unequal levels of pollution or does it rest with transport planners and society more widely? If the former, then it might be used to justify measures which restrict or impose additional costs on transport users. Yet if it is the latter, then imposing burdens on the transport users may create further inequalities if those transport users have little option but to use the transport they do.

Accessibility, understood as a measure of the time taken to get to key services or employment from different residential areas and by different transport modes is a prominent method of assessing the extent of transport related social exclusion and inequality ([Martens, 2016](#)). This method recognizes importance of transport for social and economic activity and so for people's lives and welfare. It also takes account of the opportunity cost of having to make time-consuming journeys for everyday activities and it recognizes that not everybody has the same ability to use different transport modes. Yet, looking at this form of accessibility alone leaves open a number of questions for equality. One set of questions surrounds the nuance of the measurement, for instance, does it take account of the frequency of public transport, or the quality and safety of the environment for walking and cycling? Further, if it is used to inform policy we may also need to consider the quality of the services or employment at issue (cf. [Guimarães et al. 2019](#)), and whether actions which focus on access to these services could lock in relative inequality with certain groups of people being excluded from opportunities because of where they live or the transport modes to which they have access.

Tackling Inequalities Related to Transport Modes

The discussion of the previous two sections emphasizes that transport mode inequalities can be highly context dependent, involving questions such as the extent to which the local transport system enables travel by different modes, the extent to which people have the option to live in areas which are accessible by different modes, and questions about proximity to transport infrastructure and associated pollution. We can however make some broad points about approach ([Table 1](#)).

One approach to tackling inequalities is to prioritize those which are most acute. This could mean, for instance, a focus on removing severance which prevents people from participation in social and economic and other activities, and which therefore creates more severe inequalities than those which occur when people can access activities but where doing so is awkward. Beyond this is a case for policy responses, which recognize that there are multiple forms of inequalities related to transport modes, with some relating to the benefits of transport and others to the harms of transport. Tackling transport mode inequalities will tend to require attention to these multiple forms. [Table 1](#) offers a short guide to this assessment, and shows that while there is complexity, there are also common questions which can be asked in relation to measures intended to tackle different types of inequality (a more detailed guide is given in [Mullen et al. \(2014\)](#)). This can already be seen in the discussion above, for instance in relation to the question of whether equality would be served or harmed by measures which impose burdens on motor transport users in order to reduce pollution. A similar concern can arise if policy focuses on tackling pollution primarily through promotion of cleaner vehicles. Even if

Table 1 Overview of guide for assessment and policy intervention to tackle transport mode inequalities

Assessment of inequalities		
	Benefits of transport	Harms of transport
Considerations: identify inequalities	Differences in activities involving transport which are im/possible. Include time taken to travel Physical accessibility and affordability of different modes: time taken to travel Severity of exclusion: for example severance	Risks of death and injury in road traffic collision by mode/trip and mode/distance traveled Levels of pollution and difference in exposure
Responsibility: collective or individual?	Degree of choice about modes used: include safety	Carbon dioxide emissions Degree of choice about modes used: include safety: includes questions about what a reasonable level of welfare consists in
Assessment of policy intervention to reduce inequalities in availability of transport		
	Benefits of transport	Harms of transport
Considerations	Implications for availability and affordability and accessibility of transport modes: include consideration of severance	Absolute levels and inequalities in risks of death and injury in road traffic collision by mode/trip and mode/distance traveled Implications for pollution including carbon dioxide emissions, and distribution of exposure to transport pollution
	Degree of choice about modes if the measure imposes costs on individuals	
Assessment of policy intervention to reduce inequalities in harms of transport		
	Benefits of transport	Harms of transport
Considerations	Implications for availability and affordability and accessibility of transport modes: include consideration of severance	Absolute levels and inequalities in risks of death and injury in road traffic collision by mode/trip and mode/distance traveled Implications for pollution including carbon dioxide emissions, and distribution of exposure to transport pollution
	Degree of choice about modes if the measure imposes costs on individuals	

we set aside matters of affordability of cleaner vehicles (perhaps through subsidy), such an approach can still leave inequalities which are created by high levels of traffic—these include difficulties for pedestrians and cyclists, and therefore difficulties in undertaking activities involving walking or cycling such as neighborhood social activities (Mullen and Marsden, 2016).

Addressing multiple forms of inequalities at the same time might be supported by consideration of the extent to which we can prioritize transport modes associated with creating fewer inequalities. This is the case for walking and cycling, which impose relatively few risks from road traffic collisions, and which create little pollution, and take up relatively little space and so have relatively little burden on land use and congestion. Taking this approach helps to guide policy responses to inequalities. For example, as noted in the first section, while pedestrians and cyclists impose fewer risks in collisions they do face proportionately high risks. One type of policy approach for improving pedestrian or cyclist safety could be through measures which restrict walking or cycling (perhaps through prohibitions on walking or cycling on some routes, or insisting that pedestrians and cyclists use road crossings which increase journey length). But this approach might be counterproductive in tackling wider inequalities including those associated with transport pollution, climate change, and with high levels of traffic creating barriers or severance for use of other modes.

Summary and Conclusions

Assessment of inequalities relating to transport modes, and planning aimed at reducing those inequalities, should have embedded within them a conception of why equality matters. Without such a conception which can guide responses to the question “equality of what” we have little basis for determining whether we should be concerned about any (or all) of the multiple differences in how people use or are affected by different transport modes. Perhaps more importantly we risk arbitrarily privileging some inequalities—and the interests of some people—over others. For instance, we can risk focusing on accessibility or affordability at the expense of addressing health inequalities associated with exposure to transport pollution. If transport inequalities matter because each person

matters, then it follows that we should be concerned about inequalities associated with the ability to travel and so engage in social and economic activities, and simultaneously have concern for inequalities associated with transport pollution, carbon emissions, and collisions. Yet while assessment of transport mode inequalities might be unfeasible without a conception of equality and why it matters, it remains difficult even with such a conception. The nuance and complexity of many of these inequalities raises finely balanced questions for assessment: for instance, we may have concerns that users of active modes do not face higher risks of harm from road traffic collisions than do car occupants, but it is more difficult to determine what metrics should be used to compare those risks. The picture is further complicated when we try to take account of the extent to which people can be said to exercise choice over forms of travel and mobility that they use. While questions about the nuance of assessment are a work in progress, we can make some broad comments about the sorts of approach, which might be required to tackle transport mode related inequalities. The need to simultaneously consider multiple aspects of inequality makes the case for looking to measures which will bring cobenefits across those different aspects. Much of this is the bread and butter of sustainable transport, including a focus on active travel and affordable and accessible public transport. Yet concern with inequalities points to particular ways of making decisions concerning these measures, for instance emphasizing the case for prioritizing the reduction of severance since this can be a cause of acute inequality, and focusing on implementation of active travel or public transport measures which prioritize inclusion.

Biographies

Dr Caroline Mullen is a Senior Research Fellow at the Institute for Transport studies, University of Leeds. Her research and publications focus on sustainability and mobility justice, especially on the relations between mobility justice and wider equalities and social and environmental sustainability. She also researches transport governance, policy and politics of mobility planning and carbon reduction in transport, and active travel. Her research uses theoretical analysis and primary and secondary social research methods, including innovative methods such as mobile interviews and citizen and expert deliberation. Much of her research has included collaboration and engagement with public and practitioners including transport planners from across Europe, third and private sector actors. Please see <https://environment.leeds.ac.uk/transport/staff/966/dr-caroline-mullen>.

See Also

Equity Concerns in Transport; Inequality and Traffic Safety; Pedestrian Safety, Children; Pedestrian Safety, Older People; Pedestrian Safety, Visually Impaired; Safety Culture; Energy Consumption of Transport Modes; Transport Modes and Disabled People; Transport Modes and Commuters; Transport Modes and Health; Climate Change and Transport Policy; Sustainable Transport; Social Exclusion and Health, Related to Transport; Community Severance and Health, Related to Transport

References

- Appleyard, D., 1981. *Liveable Streets*. University of California Press, Los Angeles.
- Banister, D., 2018. *Inequality in Transport*. Alexandrine Press, Marcham.
- Barnes, J., Chatterton, T., 2016. An environmental justice analysis of exposure to traffic-related pollutants in England and Wales. In: *Air Pollution XXIV, Proceedings of the 24th International Conference on Modelling, Monitoring and Management of Air Pollution, Greece, 20–22 June 2016, Crete, Greece, 20–22 June 2016*. Southampton, UK, Wessex Institute of Technology Press.
- Cohen, G.A., 1990. Equality of What? On Welfare, Goods and Capabilities. *Louv. Econ. Rev.* 56 (3/4), 357–382.
- Committee on Climate Change, 2019. *Reducing UK emissions 2019 Progress Report to Parliament*.
- Intergovernmental Panel on Climate Change, 2018. *Global Warming of 1.5 °C Summary for Policymakers*. Available from: <http://www.ipcc.ch/report/sr15>.
- International Agency for Research on Cancer, 2016. *IARC monographs on the evaluation of carcinogenic risks to humans. Vol 109. Outdoor air pollution*, IARC.
- International Energy Agency, 2018. *CO₂ Emissions from fuel combustion 2018 Highlights*. Available from: <https://webstore.iea.org/co2-emissions-from-fuel-combustion-2018-highlights>.
- Guimarães, T., Lucas, K., Timms, P., 2019. Understanding how low-income communities gain access to healthcare services: a qualitative study in São Paulo, Brazil. *J. Trans. Health* 15. doi:10.1016/j.jth.2019.100658.
- Kaltenbach, M., Maschke, C., Heß, F., Niemann, H., Führ, M., 2016. Health impairments, annoyance and learning disorders caused by aircraft noise synopsis of the state of current noise research. *Internatl J. Environ. Protect.* 6 (1), 15–46.
- Lawson, A., Matthews, B., 2005. Dismantling barriers to transport by law: the European journey. In: Colin Barnes, Geof Mercer, (Eds.), *The Social Model of Disability: Europe and the Majority World*. The Disability Press, Leeds.
- Martens, K., 2016. *Transport justice: designing fair transportation systems*. Routledge, New York.
- Mattioli, G., 2017. Forced car ownership' in the UK and Germany: socio-spatial patterns and potential economic stress impacts. *Social Inclus.* 5, 147–160, <http://dx.doi.org/10.17645/si.v5i4.1081>.
- Mullen, C., Marsden, G., 2016. Mobility justice in low carbon energy transitions. *Ener. Res. Soc. Sci.* 18, 109–117.
- Mullen, C., Tight, M., Whiteing, A., Jopson, A., 2014. Knowing their place on the roads: What would equality mean for walking and cycling? *Transport. Res. A Policy Pract.* 61, 238–248.
- World Health Organisation 2016. *Ambient air pollution: a global assessment of exposure and burden of disease*. Available from: <http://apps.who.int/iris/bitstream/10665/250141/1/9789241511353-eng.pdf?ua=1>.
- World Health Organization, 2018. *Global status report on road safety 2018: summary*. (WHO/NMH/NVI/18.20). Licence: CC BY-NC-SA 3.0 IGO). World Health Organization, Geneva.

Further Reading

- Beyazit, E., 2011. Evaluating social justice in transport: lessons to be learned from the capability approach. *Trans. Rev.* 31 (1), 117–134.
- Bostock, L., 2001. Pathways of disadvantage? Walking as a mode of transport among low-income mothers. *Health Soc. Care Commun.* 9 (1), 11–18.
- Fedtke, J., 2003. Strict liability for car drivers in accidents involving "Bicycle Guerrillas"? Some comments on the proposed fifth motor directive of the European Commission. *Am. J. Comparat. Law* 51 (4), 941–995.
- Lucas, K., Mattioli, G., Verlinghieri, E., Guzman, A., 2016. Transport poverty and its adverse social consequences. *Proce. Institut. Civil Eng. Trans.* 169 (6), 353–365.
- Mullen, C.A., Marsden, G., Philips, I., 2019. Seeking protection from precarity? Relationships between transport needs and insecurity in housing and employment. *Geoforum*.
- Mattioli, G., 2016. Transport needs in a climate-constrained world. A novel framework to reconcile social and environmental sustainability in transport. *Ener. Res. Soc. Sci.* 18, 118–128.
- Preston, J., Rajé, F., 2007. Accessibility, mobility and transport-related social exclusion. *J. Trans. Geogr.* 15 (3), 151–160.
- Schwanen, T., Lucas, K., Akyelkena, N., Cisternas Solsonac, D., Carrascoc, J.-A., Neutensd, T., 2015. Rethinking the links between social exclusion and transport disadvantage through the lens of social capital. *Transport. Res. A Policy Pract.* 74, 123–135.
- Tang, C.K., 2016. Traffic externalities and housing prices: evidence from the London Congestion Charge, Spatial Economics Research Centre Discussion Paper, 205. Available from: <http://www.spatial-economics.ac.uk/textonly/SERC/publications/download/sercdp0205.pdf>.

Railway Traffic Management

Francesco Corman, Stefano Franscini Platz, Zürich, Switzerland

© 2021 Elsevier Ltd. All rights reserved.

Introduction	678
Background	678
General Description	678
Time Scales and Degrees of Freedom of the Different Problems	679
Examples of Flexibility in Planning Which Can be Exploited by Real Time Railway Traffic Management	680
Mathematical Models for Railway Traffic Management	681
State of Industry and Application of Railway Traffic Management Concepts	681
Future Outlooks	682
Summary and Conclusion	682
Acknowledgment	682
References	682
Further Reading	683

Introduction

Railway traffic management pertains the most efficient usage of railway resources (infrastructure, vehicles, crew) in order to deliver the best possible service for operators and passengers, despite unavoidable disturbances from planned conditions. Its relevance relates to the need for better exploiting the limited, expensive resources needed to run railway services. The key idea is the exploitation of some flexibility in planning, by suitable online control and optimization, when a given evolution of the system in the future can be estimated. Therefore, railway traffic management is applied at different time scales. This contribution describes, in order: background on railway traffic management; general description of needs and constraints of railway traffic; the time scales of flexibility, and degrees of control, which could be exploited; examples identified from science and industry; some mathematical foundations and pointers for the most relevant problems; and an outlook and research needs.

Background

Railway traffic has been a key enabler of economic growth and transport performance for more than hundred years. Its advantages against private traffic are still evident in terms of transport capacity (passengers or freight transported in unit of time), speed, dedicated infrastructure connecting point to point cities and places of interests, space capacity (passenger or freight transported per vehicle length, infrastructure length, infrastructure width), energy efficiency, high reliability, very high safety, as well as costs. Typically, railway traffic is organized in lines providing transport services to passengers, interconnected over a specified network to fulfill transport demand by people or goods (Monzon and Lopez, 2020).

A railway network relies on an effective coordinated usage of a set of resources, including infrastructure asset (tracks, stations), vehicles (trains, locomotives), human resources (drivers, staff), energy and control/monitoring (Hansen, 2020). The broad concept of railway planning scheduling and management targets efficient allocations of such resources, at different time scales, and involving different stakeholders. Currently, it is accepted that the challenging capacity increase targets for railway systems identified by policy makers can only be achieved through intelligent use of the existing rail infrastructure and efficient use of the available transport capacity. Moreover, the possibility to adjust to very fast-changing situations (software) is a peculiarity of the management, in contrast to building new infrastructure and purchasing new vehicle (hardware).

General Description

The mobility needs of people and goods are represented as a **transport demand**, whose precise knowledge is a major challenge in transportation studies (Patrizia Franco, 2020). Trains are operated to fulfill this demand, bundled in lines. For a detailed description of infrastructure planning, stop planning, and line planning, see (Nielsen et al., 2005) and (Profillidis, 2020).

Train vehicles operate on railway networks. Differently from car or buses traffic, unsafe movements of vehicles are intrinsically forbidden, by means of a signaling and safety system that detects train positions, and determines safe zones where any vehicle can move (movement authority). Those movement authorities relate to a minimum safety separation between trains; a key concept is the block section, an infrastructure part that is assigned exclusively to at most one train at any one time.

Train movements can be modeled by a set of **characteristic times**, as follows. The running time of a train through a block section starts when its head (the first axle) enters the block section and ends when the train passes the end of the block section. Typically, railway operations specify heterogeneous traffic sharing the available capacity; this leads to challenges as the different services have very different speed, and stopping policies. Safety regulations impose a minimum separation between consecutive trains traveling through the same block section, which translates into a minimum headway time between the start of the running times of two consecutive trains on the same block section. This time depends on the length of the block section and (among others) the speed and length of the trains, including the time the entire length of the train gets out of the previous block sections, plus time margins for dissemination of information (signal) to the driver. The interaction between two successive trains moving safely over the infrastructure can be studied based on the blocking time theory (Hansen and Pacht, 2014), determining how much time must pass between two successive trains, depending on their characteristics, to have safe and smooth operations.

The effective usage of railway systems is normally related to a **timetable** (a schedule of operations), which can be computed (months) earlier than operations and/or possibly updated very shortly before operations. The timetable specifies the time at which trains are arriving at departing from any planned station and relevant point. A train is not allowed to depart from a stop before its scheduled departure time and is considered late when arriving later than scheduled. At a platform stop, the scheduled stopping time of each train is called dwell time. Generally, the frequency between successive services of the same type is relatively large, which reflects on demand satisfaction. Transfer connections with a minimum synchronization time enable a smooth continuation of a trip of the same passenger across two different services.

Disturbances affect rail traffic. Traffic perturbations have small magnitude and extent; (e.g., longer dwell times, headway conflicts, passenger connections, or technical failures); and disruptions are more substantial modifications of scheduled train speeds and routes, requiring strong changes to service. In both cases keeping the traffic running to fulfill the transport demand requires changes to the plan. In general, very little can be done against exogenous delays, triggered outside the railway systems (primary delays: i.e., failures, interaction with persons at stations or along tracks, longer dwell time, bad weather).

Railway infrastructure has normally very limited possibility for overtaking, typically only at major stations, for cost reasons, and infrastructure efficiency. This results in delays being propagated between successive trains as consequences to route conflicts. A route conflict occurs when the movement authority of a train is restricted by a preceding train, that is, a train meets a restricted signal and must prepare to stop before a stop signal protecting an occupied block section, increasing further its delay. **Knock on (or secondary) delays** represent this domino effect of increasing disturbances, which reduces seriously the quality of the service offered.

The process of updating the schedule when new information is available defines the railway traffic management. The timetable is updated into a **disposition timetable**, where train distance headways are respected, and the current delays are considered and reduced.

Time Scales and Degrees of Freedom of the Different Problems

A **service intention** is a concept specifying the frequency, the amount of train runs on a line to fulfill the customers' needs, and a description of departure, arrival running times but without specifying too precisely what operations should be (Caimi et al., 2011). The term "intention" refers to the envisaged operation, while not being yet sure if the offer is feasible or happening in a very precise manner. Service intentions are systematized into periodic offers running almost regularly throughout the day, to facilitate the use of public transport.

The **timetable** consists of the "production plan" of the service intentions, scheduling the detailed resource assignments, that is, departure and arrival of all services, and time supplements for delay avoidance and/or recovery (Caprara et al., 2002), in order to:

- maximize performance indicators attractive to customers, and maximize the trains load;
- match travel demand patterns including peaks;
- secure headways between train services to make them resistant to delays; and
- synchronize the arrival and departures of different lines, for smooth interchange at major stations.

The determination of a timetable requires the interaction of many stakeholders (authorities, train operators) currently done by human planners; automated timetabling computation showed great potential, especially for incremental adjustments or computation of variants (Kroon et al., 2009).

Most railway networks are operated based on a **periodic timetable**, that is, the services repeat themselves after a certain period of time (e.g., an hour). This allows passengers to remember better the available service. Full periodicity deals badly with peaks of demand in rush hour, often requiring additional services. To determine a periodic timetable, the time horizon is restricted to the period only; but the complexity is increased by the enforced periodicity. Specific types of timetables have symmetry axes, or symmetry axes for some train line along a specific time point, which occurs at stations. When this is the case of multiple train lines, it is possible to structure a pulsed timetable, where most services arrive at the station just before a critical pulse moment, they are simultaneously at the station, and leave all shortly after the pulse time. This enables smooth transfer connections within a large set of trains, thus enabling practically faster transport chains, at the expense of large infrastructure requirement at stations needed for the pulse moment.

A key characteristic of a timetable is to remain performant also against delays. To this end, **robust timetables** include running time supplements and buffer times to absorb minor deviations from nominal characteristic times. Some capacity is traded off to time

reserves increasing reliability. Larger disturbances that a timetable can cope with without major adjustments require larger time reserves, thereby decreasing capacity available for traffic. A higher volume of traffic results in very high sensitivity of train operations to perturbations, resulting in instability with even minor disturbances hindering the adherence of the actual traffic to the expected plan.

The determination of platforms at stations, (**platforming**) and routes (**routing**: sequence of block sections, that trains follow to reach a destination station starting from an origin station) is sometimes included in the timetabling stage, in an integrated manner (computationally hard), or in a sequential manner (prone to sub-optimality). Rostering is the tactical process (horizon of one or few weeks) of **assigning specific vehicle and personnel** to the running trains and services, based for instance on forecasted passenger flows.

Real time traffic management strategies (other names used in the literature including: dynamic traffic management, dispatching online, railway traffic control; train conflict detection and resolution) exploit buffers and margins in the timetable to steer the system towards a desired state. Once a perturbation is manifest, control actions such as changing and adjusting running times, dwell times, departure times, train speeds as well as train orders and routes are implemented by traffic control managers within seconds or minutes. The goal is to keep good performance and reduce the propagation of delays in the network (Cacchiani et al., 2014; Corman and Meng, 2014). The real time railway traffic management is the ideal required complement of a service intention, that is, improves performance by exploiting the details not fixed in the plan (or, available for change in the real time operations) by control decisions which are based on the precise ever changing situation, in real time. This approach requires effective and fast traffic control procedures by either human dispatchers, or automated systems based on mathematical optimization.

The running time of trains can be updated (**retiming**) by resulting in updated arrival departure time at stations, while keeping fixed train orders and timetable structure. Trains can follow the updated advice by route setting and signals prescriptions, or more sophisticated driver advisory systems, including speed-space-time goals (so called Train Path Envelopes); or even automated train operations directly implementing the computed prescriptions (Wang and Goverde, 2016).

Reordering or rescheduling has instead a view over multiple trains, and updates orders of trains. Similarly, **rerouting** changes their route. Both actions have a greater potential for reducing delay propagation, allowing to exploit the available infrastructure capacity in the space and time where/when it is available. Some of those choices might have impact to passengers (i.e., cancelled stops, different running time, different platform at stations). Most often, when changing orders or routes, those decisions must be communicated opportunely to passengers, so that they can board the most convenient train in the updated disposition timetable.

When disturbances are of particularly serious intensity, train services and lines need to be substantially changed (**reservicing**), resource plans updated, new train routes have to be followed, short-turning or even cancellation of services applied. In such cases, the entire structure of the offer is modified, partially dropping the planned timetable, but still with the goal to match the service intention, as much as possible (Binder et al., 2017). The challenge is the evaluation of such situations, which require specific modeling of very limited infrastructure capacity; multiple objective functions, representing the many stakeholders involved, including the passengers; dissemination of the right information to train managers and passengers.

Examples of Flexibility in Planning Which Can be Exploited by Real Time Railway Traffic Management

The most common objective function for railway production is punctuality, defined as trains reaching their destinations (or control points) according to the planned timetable plus a given threshold (say 3 minutes). While punctuality is specific and measurable, more comprehensive indicators relate to precise delays of trains. Both punctuality and delays do not measure exactly what customers want, that is, reaching their destination on time and with a high reliability, and not caring per se about just train delays. To this extent, planning flexibility can be exploited, to match the service intention to a new plan specifically addressing the contingent situation.

A timetable with strict arrival/departure times can be replaced by **flexible time windows** of [minimum, maximum] arrival/departure times at each platform, which enable some flexibility to be later exploited in a real time stage. The operational timetable (for railway operators, with explicit flexibility) therefore differs from the public timetable (for passengers, only the maximum arrival time and the minimum departure time). The longer travel times would be compensated by a higher reliability of train services. Similarly, a service intention can result into an imprecise timetable, with provisional, possibly infeasible train separation and/or orders resolved every day according to the precise conditions in real time. This flexibility allows to reduce delays if (possibly different) good schedules can be found within strict time limits of computation. Finally, a service intention can specify **imprecise description of stopping platform**, typically limited to the two sides of the same platform island; the final choice is left to real time traffic management. Similar options can describe routes outside stations.

The **delay management** problem focuses specifically at keeping or dropping passenger connections in presence of delays. If a connection is kept (i.e., train is held at a station waiting for a feeder train) there can be substantial passenger travel timesavings against the case in which a connection is dropped, reducing delays for trains, but increasing it for the passengers. Often, demand is considered known, without elasticity, sometimes described as Origin-Destination pairs. Due to the combinatorial complexity of the problem, limited details of the infrastructure are kept. Recently the inclusion of precise management of infrastructure and description of passenger-oriented indicators has been shown possible (Corman et al., 2017).

Mathematical Models for Railway Traffic Management

The mathematical models used for railway traffic management can be classified into few broad families. Time continuous approaches report time variables in a continuous domain. Time indexed approaches allow time variables to have only a few possible values, that is, an integer (often considering a minimum time resolution, say 1 minute). A different classification regards the **time resolution** and **space resolution** considered: macroscopic models consider link and stations as infinite capacity resources. Microscopic models represent block section as resources with unitary capacity, used in sequence, resulting in much larger sets of constraints and variables. Mesoscopic models somehow provide hybrid solutions between the two cases. Regarding degrees of freedom, the **inclusion of speed control** relates to continuous variables, and quadratic relations between speed space-time, resulting in non-linear programming models. The inclusion of train orders and traffic requires binary constraints due to the determination of which train must occupy in which sequence the infrastructure elements. The determination of which connection to drop or keep also relates into similar binary constraints. This is solved by means of integer programming approach or mixed integer linear programming approach, MILP (see for instance, [Cacchiani et al., 2014](#); [Corman and Meng, 2014](#); [Pellegrini et al., 2014](#); [Meng and Zhou, 2011](#)). Finally, **inclusion of passengers** desires and routing might take the simpler way of schedule based deterministic assignment, or stochastic assignment. Competition between different passenger groups complicates further the problem and requires behavioral assumptions.

Most railway traffic management problems can be **modeled** and formulated along a **timed event graph**, in which nodes represent events in time and space (possibly with associated services/ trains, resources) and arcs represent relations between them (such as precedence relations, durations). Models based on job shop scheduling with additional constraints (no-store and no-swap) have been relatively popular ([Mascis and Pacciarelli, 2002](#)), and termed **alternative graphs**, as the key decision is the alternative possible orders of trains over shared resources (block sections). Most microscopic approaches combine the blocking time theory for the recognition of routing conflicts, and a general discrete optimization model to evaluate possible solutions of train retiming, reordering, and rerouting against a specified objective function.

The **passenger assignment and guidance** is the problem of computing and disseminating a route advice to passengers (guidance), and evaluating the results of the actual route they took. In practice it can happen that the advice can range from complete and correct, to incomplete, imprecise, or impossible, especially under disturbances. More research on the interaction between information and route choice has been done in the general transit literature, identifying frequency-based transit assignment, and hyperpaths concepts. Currently, route advices are made available to passengers by online applications and websites, or by personal construction of a plan when looking at printed timetables at stations.

In typical railway networks, passengers have **free route choice** (within some boundaries given by a fare system) ([Freijnger and Zimmermann, 2020](#)). Thus, passengers might react to reordering decisions taken in the interest of minimizing their delay, by finding the path of minimum travel time. In theory, a system optimum (minimizing globally passenger travel time) might not correspond to a user equilibrium (a situation where nobody has incentive to take a different decision); the incomplete information of future events makes the situation even more complex.

Few approaches have been tried successfully in real life, able to **connect railway traffic management decisions with expected passenger reactions** and outcomes. For instances transfer connections should include the reaction of passengers to dynamic railway traffic management actions, even though those are still unknown at the planning stage. In such perspectives, static passenger flows are used, sometimes including realistic behavioral assumptions. Some data on past operations is utilized to compute distributions of delays and travel times. Even though more advanced data-driven deterministic models are able to explain a large percentage of process time variability using the values of explanatory variables, a certain degree of uncertainty, especially for dwell times, remains unresolved. The usage of such models in a real time perspective is still in its infancy.

State of Industry and Application of Railway Traffic Management Concepts

Basic dispatching systems have been developed to aid traffic controllers, whose computer support is often limited to graphical interfaces and simple automatic route setting systems, triggered automatically if a train approaches a signal according to the plan. The usual policy still consists of scheduling trains following their order in the timetable, or according to pre-determined dispatching rules, with few manual modifications actually considered. Experienced dispatchers have developed strategies which allow them to foresee simple forms of possible disturbance and their effects in advance, and to take compensatory control actions based the information they might have. Existing support system for dispatchers rarely go beyond monitoring the situation, as they provide acceptable solutions only for small-scale problems or for simple perturbations. Some simple implementations exploit flexibility in orders or platforms choice by first-come-first-served rules on those resources. They found their application in bottlenecks (e.g., underground stations such as Schiphol Airport in the Netherlands); moreover they are sometimes complemented by more sophisticated approaches able to determine a minimum look ahead in the future results of current decisions.

For larger delays, priority rules for different classes of train types are used, but cannot quantify the gap from an optimal network solution. For very large disruptions, operators often have prepared on beforehand a set of possible alternative timetables, which can then be adapted and implemented for the specific situation encountered. The path for implementation of automated real time railway traffic management systems has been started with a few implementations ([Borndörfer et al., 2018](#)); and demos of closed loop control systems ([Corman and Quaglietta, 2015](#)).

One of the best examples of the potential for automated and optimized approaches is showcased in the Dutch example (Kroon et al., 2009). The use of advanced operations research techniques for timetable, vehicles, and crew scheduling, lead to large benefits to the main railway operating company of the Netherlands, NS. The resulting yearly profit could be quantified in millions of Euro, achieved directly by cutting expenses or by increased passenger demand and punctuality bonuses for improved reliability of operations.

Future Outlooks

In the forthcoming decades **railway transport is expected to face a significant growth** of traffic flows, which will mostly have to be accommodated over the existing infrastructures. An increase of capacity utilization is needed to avoid reduction of reliability and punctuality of transport services. To remain competitive in a market with increased competition due to disruptive technologies and uncertainties, continuous economic and performance improvements are required. The last years have seen in general an increase in transport performance (passenger kilometers and ton kilometers) against a generally increased transport demand, by which ultimately the relative mode share of railways is often reducing.

More comprehensive **inclusion of passengers** into the problem will finally enable to directly target the passenger wishes, and not only the operator requirements. The interplay between assigning the passenger flows and generation of the schedule can arise at different levels, which has not been studied in detail so far.

A reason for online control also includes **major disruptions**, i.e. blockades, trains or facilities out of control. These situations require major changes in the train service provided, including canceling partial or full train services. The approaches should consider explicitly information availability, and the impact of very large delays towards the users, rather than immediate small delays of a transport link.

Technological developments currently allow **higher autonomy and controllability for most train operations**, by means of pervasive automation, depending on conditions of infrastructure and traffic. Having the possibility to control trains with very small margins and high reliability leaves further space to increase traffic at network level; and include other relevant objectives, for instance minimize energy usage (De Martinis and Corman, 2018).

The computational complexity of automated decision support for (real time) railway traffic management typically restricts the application to small cases and few possible degrees of freedom. On the other hand, the largest benefits are in large cases and simultaneous inclusion of multiple degrees of freedom. To this end, **decomposition approaches** in sequential, integrated, or iterative approaches are promising techniques.

The widespread use of **data analytics** enables opportunities for improvements of railway operations. This certainly includes forecast the future operations, by means of black box approaches, or explicit mathematical models. Stochastic prediction and optimization models attribute each event with a probability distribution in order to describe the uncertainty of its realization. Most delay analysis is now done a posteriori, or to estimate delay distributions when designing a timetable. Instead, dynamic models for real time prediction or management would be updated in real time as new information becomes available.

Summary and Conclusion

In summary, railway traffic management is a promising way to complement by resource optimization (software) the (hardware) potential of a railway system, which is typically resource-intensive and infrastructure-heavy. The current trend in railway operations goes decisively towards a widespread usage of such technologies. Their application has been estimated in academic studies, or already proven in industrial applications, to improve cost efficiency and quality of service over a range of time scopes and problems, for passengers and operators.

Acknowledgment

The author acknowledges the support of the Swiss National Science Foundation, with Project 181210/DADA.

References

- Binder, S., Maknoon, Y., Bierlaire, M., 2017. The multi-objective railway timetable rescheduling problem. *Transport. Res. C* 78, 78–94.
- Borndörfer, R., Klug, T., Lamorgese, L., Mannino, C., Reuther, M., Schlechte, Th., 2018. Handbook of Optimization in the Railway Industry. *International Series in Operations Research & Management Science*, vol. 268, Springer International. doi: 10.1007/978-3-319-72153-8.
- Cacchiani, V., Huisman, D., Kidd, M., Kroon, L., Toth, P., Veelenturf, L., Wagenaar, J., 2014. An overview of recovery models and algorithms for real-time railway rescheduling. *Transport. Res. B* 63, 15–37.
- Cairni, G., Fuchsberger, M., Laumanns, M., Schüpbach, K., 2011. Periodic railway timetabling with event flexibility. *Networks* 57, 3–18.
- Caprara, A., Fischetti, M., Toth, P., 2002. Modeling and solving the train timetabling problem. *Operat. Res.* 50 (5), 851–861.
- Corman, F., Meng, L., 2014. A review of online dynamic models and algorithms for railway traffic control. *IEEE Transact. ITS* 16 (3), 1274–1284.

- Corman, F., Quaglietta, E., 2015. Closing the loop in railway traffic control: framework design and impacts on operations. *Transport. Res. C* 54, 15–39.
- Corman, F., D'Ariano, A., Marra, A., Pacciarelli, D., Samà, M., 2017. Integrating train scheduling and delay management in real-time railway traffic. *Transport. Res. E* 105, 213–239.
- De Martinis, V., Corman, F., 2018. Data-driven perspectives for energy efficient operations in railway systems: current practices and future opportunities. *Transport. Res. C* 95, 679–697.
- Freijnger and Zimmermann, 2020. Route Choice and Network Modeling, in this encyclopedia.
- Hansen, 2020, Railway station and network planning in this encyclopedia.
- Hansen, I.A., Pachl, J., 2014. Railway timetable and operations. Eurailpress, ISBN 978-3777104621.
- Kroon, L.G., Huisman, D., Abbink, E., Fioole, P.J., Fischetti, M., Maróti, G., Schrijver, A., Steenbeek, A., 2009. The new Dutch timetable: The OR revolution. *Interfaces* 39, 6–17.
- Mascis, A., Pacciarelli, D., 2002. Job shop scheduling with blocking and no-wait constraints. *Eur. J. Operat. Res.* 143, 498–517.
- Meng, L., Zhou, X., 2011. Robust single-track train dispatching model under a dynamic and stochastic environment: A scenario-based rolling horizon approach. *Transport. Res. B* 45 (7), 1080–1102.
- Monzon and Lopez, 2020, Rail transport and territorial scale, this encyclopedia.
- Nielsen, G., et al., 2005. Public Transport—Planning the networks, HiTrans Best Practice Guide 2. HiTrans international steering group, Stavanger.
- Patrizia Franco, 2020, Multimodality and transport demand modelling, in this encyclopedia.
- Pellegrini, P., Marlière, G., Rodriguez, J., 2014. Optimal train routing and scheduling for managing traffic perturbations in complex junctions. *Transport. Res. B* 59, 58–80.
- Profillidis, 2020, Railway management, in this encyclopedia.
- Wang, P., Goverde, R.M.P., 2016. Multiple-phase train trajectory optimization with signalling and operational constraints. *Transport. Res. C* 69, 255–275.

Further Reading

Nash, The history of rail transport, in this encyclopedia.

Indoor Transportation

Lutfi Al-Sharif^{*,†,‡}, ^{*}Mechatronics Engineering Department, The University of Jordan, Amman, Jordan; [†]Al-Hussein Technical University, Amman, Jordan; [‡]Peters Research Ltd., Great Missenden, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	684
Historical Background	684
Introduction to Indoor Transportation	685
General Overview of a Traction Elevator	685
Elevator Traffic Design	687
Space Planning	688
Escalators	690
Passenger Conveyors	690
Recent Developments	690
Advances in Motor and Variable Speed Drive Technologies	690
Intelligent and Sophisticated Control Systems	691
Optimized Exploitation of Shaft Space in Buildings	693
Future Developments	693
Summary and Conclusions	694
Biographical Notes	695
See Also	695
References	695
Further Reading	696

Introduction

The main components of indoor transportation (sometimes also referred to as Building Transportation) are elevators (also referred to as Lifts in the United Kingdom), escalators and moving walkways (also referred to as passenger conveyors). These components address both vertical as well as horizontal transportation.

This article presents a concise introduction to indoor transportation. It looks at the main five disciplines that comprise the science of indoor transportation. It then looks at the components and principle of operation of the most widely used elevator, the roped traction elevator. A high-level overview of the topic of elevator traffic engineering is then presented. The components of elevator space planning in buildings are presented, highlighting its close dependence on the capacity and rated speed of the elevator. An overview of escalators is given in the following section as well as moving walkways in the following section. The article concludes by discussing the modern developments that have been introduced in the last 30 years and that have transformed the sector of indoor transportation, as well as future developments in this area. The article concludes with a summary and conclusions section, that includes recommendations as to the most pressing areas of research.

For more detailed information on indoor transportation, readers are referred to the following references: [Hirscher \(1987\)](#), [Strakosch \(1998\)](#) and [CIBSE \(2000\)](#).

Historical Background

In the year 1853, Elisha Graves Otis (1811–61) demonstrated the safe operation of the roped elevator system in the show in the New York World's Fair. This invention made the safe use of roped traction elevators possible in buildings, by protecting the passengers from the risk of a falling elevator if the suspension ropes are cut. The demonstration gave the public the confidence to use roped-traction elevator. Elisha Otis went on to form what today is known as the Otis elevator company.

The basic concept of using suspension ropes to move the elevator car up and down in order to move passengers has remained unchanged for the last 150 years, despite many technological developments in terms of electrical motors, drive systems, control systems, and safety features.

Elevators have basically transformed the modern city landscape by making the concept of a high-rise building or a skyscraper possible. It inverted the concept of the office hierarchy whereby the higher management started to occupy the top floors in the building because they became the more desirable floors, with the lower ranks in the company employing the lower floors ([Harford, 2018](#)). By making the concept of the skyscraper feasible, it made it feasible to have high density populations concentrated in city centers.

Introduction to Indoor Transportation

When considering indoor transportation systems, a number of disciplines are relevant. It is thus important to distinguish between the following five disciplines:

Passenger Flow: This discipline involves the analysis of passenger (pedestrian) movements within the building mainly in the vertical dimension (although horizontal movement is sometimes considered simultaneously as part of integrated passenger flow studies in buildings). Readers who are interested in gaining a deeper insight into the area of human pedestrian flow in general are referred to the important book by J. J. Fruin (Fruin, 1987), as well as Chapter 2 of the CIBSE Guide D, 2015 edition (CIBSE, 2000).

Traffic Engineering: This discipline is concerned with the analysis of the expected vertical passenger flows in the building (usually referred to as Demand) and the sizing of suitable elevator systems that can meet such demand ((Barney and Al-Sharif, 2016), (Markon et al., 2006)). The selection process entails the specification of the number of elevators, their rated speeds and rated capacities.

Electrical Engineering A, Control Systems and Safety Devices: This discipline involves the use of electrical technology for the logic control of elevators, such as the registration and allocation of landing and car calls, and the direction of movement of the car, as well as the signals for opening and closing the doors. *Electrical Engineering B, Drives and Propulsion Systems:* This discipline is concerned with the means of movement of the elevator car. The main components are the electrical motor and the variable speed drives controlling it, as well as the speed feedback devices and position feedback devices. An overview of the electrical energy consumption in elevators is given in Al-Sharif (1996).

Mechanical Engineering: This area involves the mechanical components that comprise the elevator system, such as: the steel ropes, the guide rails, the gearbox, the car frame, the counterweight, the buffers in the pit, and the over speed protection system (Janovsky, 1993).

Structural Issues and Space Planning: A structural support system is required in the building that would support the drive motor and provide guidance to the movement of elevator car, as well as supporting the forces in the elevator shaft pit. Space planning is necessary to ensure that the elevator system and its ancillary component can be installed and safely operated inside the building.

The collective knowledge and expertise in all the disciplines listed above contributes to the successful design and operation of indoor building transportation system. This article will provide a high-level overview of the area of indoor transportation.

General Overview of a Traction Elevator

Elevators are either roped traction elevators or hydraulic elevators. The use of hydraulic elevators is continuously declining, and thus most new elevators are of the roped traction elevators. With reference to the figure shown below, the operation of the roped elevator can be summarized in the following points:

1. An electric motor is used as the hoisting machine. In the majority of cases, the motor is a three-phase squirrel-cage-induction-motor, usually driven by a variable frequency drive system.
2. The electric motor is connected to a sheave, either directly (gearless) or indirectly through a gearbox (geared). Where provided, the gearbox achieves the required reduction in speed and the required increase in torque. A sheave is a pulley that has grooves on its surface. The steel ropes fit inside the grooves on the sheave's surface in order to achieve the required traction (hence the term roped-traction-elevator).
3. The ropes are connected to the elevator car on one side of the sheave and to the counterweight on the other side. The counterweight is basically dead-weights used to balance the car and is made of metal blocks slotted inside a frame. The counterweight is usually equal to the mass of the empty car plus half the passenger rated load. The use of the counterweight allows a small motor to be selected and reduces the energy consumption as it balances all of the dead-load in the system (the mass of the car) as well as half the live-load (the passengers).
4. The elevator car is fitted inside a car frame (informally referred to as a sling). Resilient mountings are fitted between car and the car frame in order to isolate the passengers from vibrations.
5. The travelling path of both the elevator car and the counterweight are guided using guide rails. There are two guide rails for the elevator car and two guide rails of the counterweight. The guide rails only provide lateral guidance but will not prevent or obstruct the vertical movement of the elevator car and the counterweight.
6. Power and signals are connected to the elevator car from the controller using the travelling cable (sometime referred to as the trailing cable).
7. The door fitted to the elevator car is called the car door. This is powered by an electric motor to open it and close it when the car has stopped at one of the landings.
8. The landing doors are not powered but have automatic closers for safety purposes. The closers are either spring operated or weight operated.
9. Once the elevator car stops at a landing and the elevator car doors and the landing doors are correctly aligned, the car door motor opens the elevator car doors, which in turn, engage with the landing door opening mechanism and open the landing doors as well. A similar sequence takes place when the elevator car doors are closed.
10. If for some unexpected reason the elevator car moves away from the landing doors when they are open, the automatic closers will ensure that the landing doors automatically close.
11. The elevator doors and the landing doors have electrical inter-locking safety contacts that prevent movement of the elevator car if any of the doors are not locked.
12. In order to protect against the possibility of the elevator car passing the lowest landing at rated speed (and below the over speed), buffers are installed in the pit of the elevator shaft. The buffers will safety arrest and bring to a standstill the elevator car that is travelling at rated speed. A similar buffer system is also installed under the counterweight.

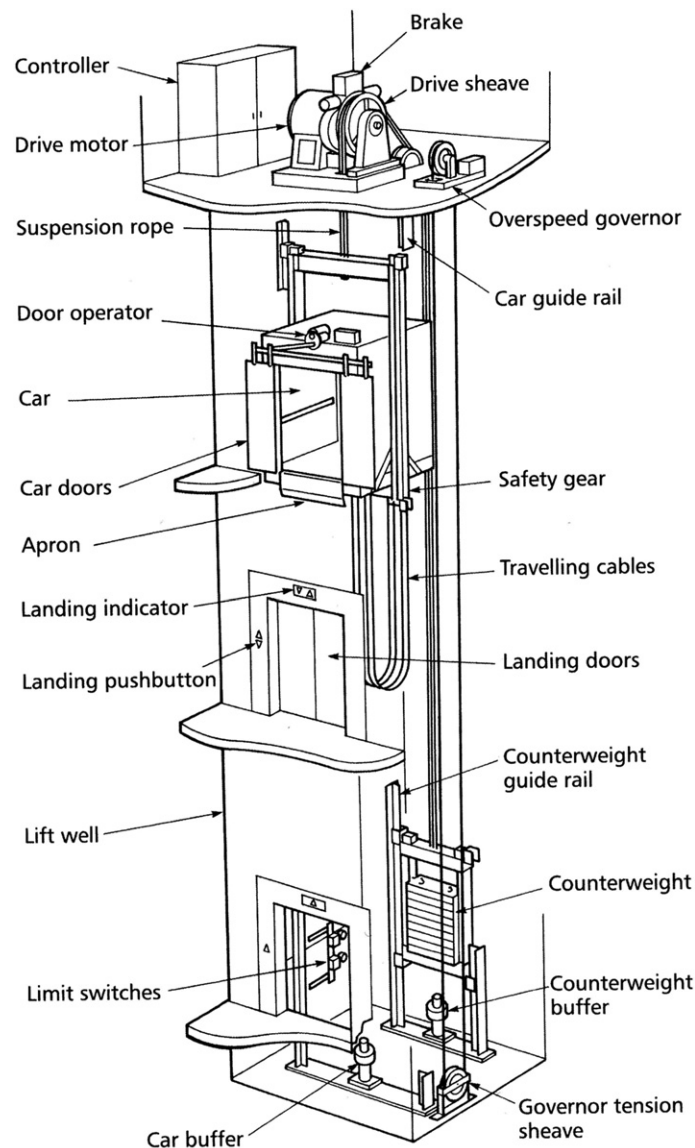


Figure 1 General Overview of a roped traction elevator Source: From CIBSE Guide D, 2015 with permission).

13. The buffers are generally of two types: Energy accumulation buffers and energy dissipation buffers.
14. In order to protect against the risk of freefall if the ropes are severed, an over speed protection system is installed. The over speed protection system comprises the following:
 - a. A single smaller rope that is connected to the elevator car and causes the rotation of the speed governor wheel that is located in the machine room.
 - b. The speed governor wheel contains a centrifugal mechanism that locks the wheel if the speed of the elevator (as detected by the single small rope) exceeds the over speed value.
 - c. Once the speed governor wheel locks, it pulls on the small rope.
 - d. The small rope activates a mechanical lever on the elevator car.
 - e. The mechanical lever then operates a braking system that is installed on the elevator car, which in turns locks onto the elevator car guide rails

An enlarged view of the roped-traction machine is shown in [Fig. 2](#). The drive motor is usually a squirrel cage induction motor which is mechanically coupled to the gearbox. Sizing of elevator motors can be carried out using the methodology outlined in [Al-Sharif \(1999\)](#). The purpose of the gearbox is to reduce the speed and increase the torque. The output shaft of the gearbox drives a sheave. The sheave is a pulley with grooves on its surface. The traction steel ropes sit in the grooves. The sheave provides traction with

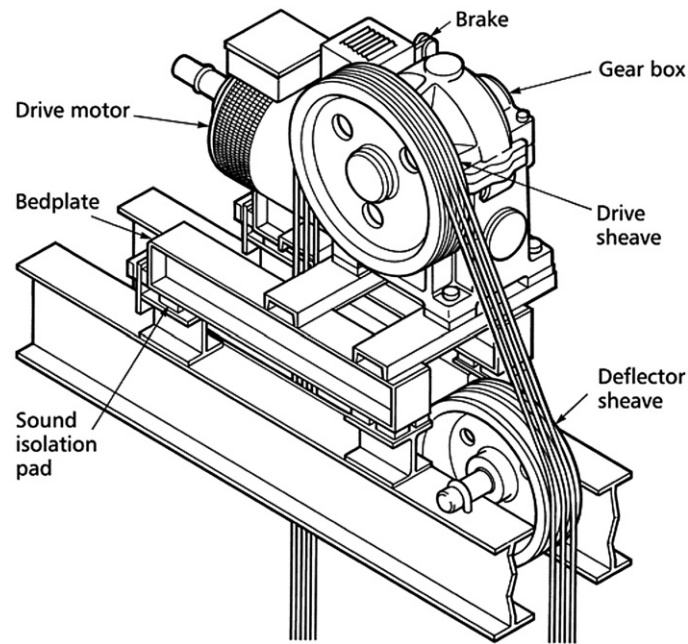


Figure 2 Typical geared traction elevator drive system *Source: From CIBSE Guide D, 2015 with permission).*

the ropes in order to raise and lower the elevator car. A counterweight is connected to the other end of the ropes on the other side. A spring-applied electromagnetically released brake operates on the shaft between the motor and the gearbox. The brake is necessary from a safety point of view in order to ensure the stopping of the elevator car in case of an electrical system failure or when a safety device is operated. In [Fig. 2](#), the bedplate on which the hoisting machine (i.e., the motor, gearbox and its frame) can be seen. Sound isolation pads (usually made of a resilient material such rubber or polyurethane) isolate the hoisting machine from the building structure in order to prevent structural-borne-vibration from being transmitted.

Elevator Traffic Design

The first most important step in sizing elevator traffic systems is assessing the expected passenger demand. Assessing demand depends on usage of the building (e.g., offices, residential, hotel) and on the building population.

Elevator systems have been traditionally sized based on the percentage of the building population arriving in the worst five minutes. This has usually been taken as 15% or 12% of the population arriving during the busiest 5-min period. Modern practices show that this figure is nearer to 10%–12%, due to the prevalence of flexible starting and finishing times ([Stanhope, 2004](#)). More details about assessing demand can be found in [Al-Sharif \(2014\)](#). A typical arrival-departure pattern can be seen in [Fig. 3](#). The curve above the x -axis represents the passenger traffic entering the building which is referred to as incoming traffic or up-peak traffic. The curve below the x -axis represents the traffic departing from the building which is referred to as out-going traffic or down peak traffic. It is the peaks in these two curves that are used as the basis for the design of the system, especially the morning incoming traffic peak.

Once the demand has been estimated, it is then possible to size the main passenger elevator system. This is only the part of the indoor transportation system that caters for the bulk movement of passengers during non-emergency situations. Other aspects of the building transportation needs are addressed later (e.g., goods movement, emergency evacuation, firefighting, kitchen, and waste movement). The number, grouping, and method of control of the elevator system are decided at this stage. Two criteria exist for the selection of the number, speed, capacity, grouping, and method of group control of the elevators: The quality of service and the quantity of service.

The quality of service is best reflected in the interval between the arrivals of the elevators at the main entrance and is called the “interval” and affects quite strongly the passenger waiting time. The quantity of service is best reflected by the number of passengers moved to their destination during a period of 5 min expressed as a percentage of the building population (called handling capacity).

The traffic analysis can then be either carried out using calculation or simulation. Under calculation, the value of the round-trip time (RTT) is calculated. This is then used to find the required number of elevators to achieve the specified value of the interval (usually 30 s), which is representative of the quality of service. This is then followed by another calculation to decide on the required car capacity. For these figures, the handling capacity of the system can be found and matched to the expected passenger arrival rate. More details about the RTT calculations and elevator traffic engineering design can be found in [Al-Sharif \(2014\)](#), [Al-Sharif \(2015\)](#), [Al-Sharif et al. \(2015\)](#).

Simulation can then be carried out to better understand the performance of the system under real life conditions and under the random nature of the passenger arrivals (([Hakonen and Siikonen, 2008](#); [Finschi and State-of-the-art traffic, 2010](#); [Al-Sharif and Al-](#)

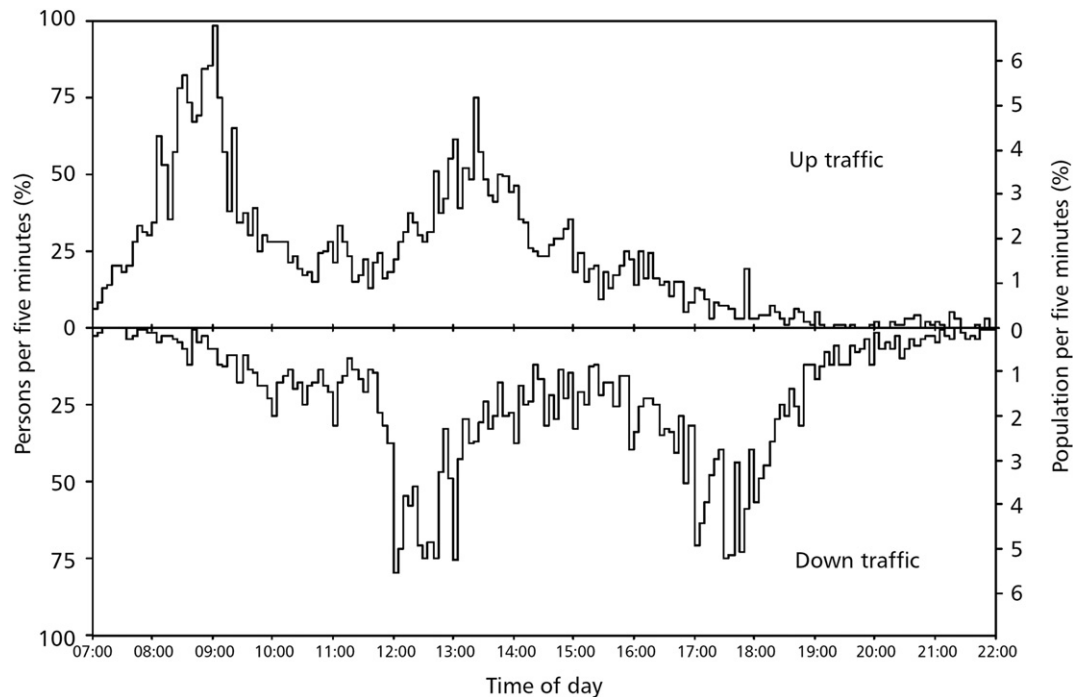


Figure 3 Passenger demand for simple traffic survey (page 4-4 CIBSE Guide D, 2015)

Adhem, 2014; Siikonen, 1993; Siikonen, 2000; Siikonen et al., 2000; Peters, 2013)). Simulation can provide more detailed results, such as passenger waiting time, passenger travelling time, and actual car loading. More importantly, simulation can show the effect of the elevator group control system on the performance.

Both calculation and simulation require a number of parameters to be specified such as: Door opening time and closing time; passenger/wheelchair transfer time (affected by factors such as door width, car depth and width, nature of passengers, familiar/unfamiliar); number of passengers boarding the elevator car; acceleration and deceleration times; jerk; inter-floor distances; advance door opening time if applicable. More details about the elevator traffic design for high rise buildings can be found in Wit (2017), To and Yip (2004), Al-Sharif et al. (2017).

Passenger traffic surveys present a powerful tool for verifying existing designs or for understanding the reasons for poor performance (Peters and Evans, 2008). In its simplest form, passenger surveys, as related to elevator traffic engineering, involve the counting of the intensity of passenger arrivals during the busiest 5 minutes during the day. More specifically, human surveyors are often employed to stand in elevator lobbies and count the number of passengers arriving for elevator service in each minute. In cases where there are many elevators in the lobby (e.g., four elevators facing four elevators, or in the case of multiple entrances at different levels), then a number of surveyors are employed. When this is done, it is important to synchronize the time reference in order to ensure consistent data from the different surveyors.

With the advent of modern surveying systems, it is now possible to automate the process and gather much larger amounts of data. Readers who require more details on this topic are referred to chapter 18 of the Elevator Traffic Handbook (Barney and Al-Sharif, 2016).

Space Planning

When planning the installation of elevators in a building, it is necessary to carefully plan the spaces needed. The requirements for the space are closely linked to the capacity of the elevator car, the rated speed of the elevator and the type of drive system.

The elevator rated speed is quoted in units of m s^{-1} and follows preferred values within a Renard series (R-series). The elevator rated capacity is usually cited in units of kg, with the equivalent number of passengers stated as well. The rated capacity in kg also follows preferred values within a R-series. In order to convert from the elevator rated capacity in kg to persons and *vice versa*, it is customary to use a standard passenger mass of either 75 kg or 80 kg.

Internal elevator car width and depth determine whether the elevator car is wide (width larger than depth) or deep (depth larger than width). In the case of passenger elevators, it is preferable to use wide cars, as this speeds up the transfer of passengers into and out of the elevator car and hence contributes to the improvement of the traffic performance of the elevator traffic system. Deep cars are discouraged for passenger elevators but are widely used for goods/service elevators and for bed-lifts in hospitals.

An important standard for the preferred speeds and capacities of elevators and for shaft dimensions is the following ISO standard: “BS ISO 8100-30: 2019 entitled: Lifts for the transport of persons and goods. Part 30: Class I, II, III, and VI lifts installation” (ISO, 2019).

The elevator preferred speeds and loads follow an R10 series (Renard 10). Below are some typical capacities in kg and persons for passenger elevators:

- 630 kg (8 person)
- 800 kg (10 person)
- 1000 kg (13 person)
- 1275 kg (17 person)
- 1600 kg (20 person)
- 2000 kg (26 person)

The numbers shown above in kg follow a Renard 10 (R10) series, and hence are vital in facilitating standardization between different manufacturers, improving competition and allowing planners and architects to provide the required space in the building prior to tendering.

In terms of shaft space planning, the following are the four most critical dimensions in the well space (also referred to as the shaft space):

- The shaft width
- The shaft depth
- The headroom (defined as the vertical distance from the finished floor level (FFL) of the topmost floor to the top of the inside of the shaft)
- The pit depth (defined as the vertical distance from the FFL to the floor at the bottom of the shaft)

The headroom and the pit depth requirement incorporate within them the safety distance necessary to protect the maintenance personnel working on top of the car or in the pit. The main parameters that need to be decided for each elevator are:

1. The shaft width: The shaft width is the clear internal width of the shaft. The width is the dimension that is parallel to the elevator doors direction of opening and closing. It is necessary to add around 150–200 mm for the divider beam between elevators in shaft to allow for supporting structures.
2. The shaft depth: The shaft depth is the clear internal depth of the shaft. The depth is the dimension that is perpendicular to the direction of door opening and closing.
3. The pit depth: The pit depth is the vertical distance between the FFL of the lowest-most floor and the floor of the pit.
4. The headroom: The headroom is the vertical distance between the FFL of the top-most served floor and the underside of the slab at the top of the shaft.
5. The machine room headroom: The headroom inside the machine room is the distance from the FFL on which the motor is placed to the underside of the slab at the roof of the machine room. This ranges from 2000 mm to 4000 mm depending on the size of the motor.

A general overview of the pit is shown in Fig. 4. In this case, the elevator car has two buffers underneath it, and the counterweight has one buffer. At the sides of the pit, the guide rails can also be seen for both the elevator car and the counterweight.

The concept of the “human rectangular block” is an important concept in the protection of personnel working inside the elevator shaft, either in the pit or on top of the car. In the pit, when the car is sitting on the fully compressed buffer(s), the space should be able to accommodate a rectangular block of dimensions $0.5 \times 0.6 \times 1$ m sitting on one of its sides. In a similar way, when the counterweight is sitting on its fully compressed buffer(s), a space should exist above the car that can accommodate a rectangular block of dimensions $0.5 \times 0.6 \times 0.8$ m, sitting on one of its sides.



Figure 4 A general overview of an elevator pit, showing the two elevator car buffers, the counterweight (at the back of the pit) and the guide rails at the sides.

Table 1 Escalator handling capacity for different step widths and different speeds in units of passengers/hour (based on EN115).

Speed (m/s)	Step width (m)		
	0.6	0.8	1
0.5	4500	6750	9000
0.65	5850	8775	11,700
0.75	6750	10,125	13,500

More about the detailed requirements for firefighting elevators can be found in [BSI \(2008\)](#) and Chapter 6 of CIBSE Guide D 2015 ([CIBSE, 2000](#)).

Escalators

Escalators are also an important component of indoor transportation. Escalators have a much higher handling capacity than elevators. They are generally used in cases where a significant flow of passengers is expected over short distances (e.g., one or two floor journeys). This is typical of shopping centers. A table of the handling capacities of escalators in units of passengers per hour is shown in [Table 1](#), based on [EN \(2009\)](#). Real life data of handling capacity on escalators can be found in [Mayo \(1966\)](#).

Most escalators have an angle of incline of 30 degrees. An angle of incline of 35 degrees is allowed by the European norm (EN115-1:2008) if the speed does not exceed 0.5 m/s and the vertical rise is no more than 6 m. An angle of incline of 35 degrees would reduce the space footprint taken up by the escalator. However, there is evidence to suggest that the risk of passenger falls is higher on 35-degree-inclined escalators.

From a mechanical point of view, an escalator is basically an endless chain of consecutive cleated steps that is made to move along an inclined tracking system. A geared motor is installed inside the upper machine space below the upper landing of the escalator. The geared motor is used to provide the movement to the steps. Two endless handrails are also installed and run at the same speed as the step-band. A general overview of an escalator is shown in [Fig. 5](#).

Passenger Conveyors

A moving walkway, moving walk or a passenger conveyor is a device that moves passenger horizontally indoors. It is a very effective and practical means of indoor horizontal transportation, which makes it widely used in airports, railway terminals, and even shopping centers.

In most applications, the moving walkway is horizontal, although it can have an incline of up to 12 degrees (based on the European standard EN115-1:2008). In some cases, it is used to move passengers between different floors. However, even when used in an inclined configuration, steps do not form on the moving walkway and it remains a flat walkway. At an inclination of 12 degrees, this requires extra care by travelling passengers, especially those passengers using shopping trolleys or baggage carts.

The moving walkway usually has rated speeds of 0.75 m/s, although a speed of 0.9 m/s is allowed under certain conditions.

The tread width of the moving walkway can range from 800 mm up to 1400 mm. A width of 1400 mm is ideal as it allows passengers using baggage carts and shopping trolleys to overtake other passengers. The tread way on the moving walkway is either made from metallic pallets (pallets are flat steps similar to those used in escalators) or from belts.

One of the early examples of this device were two inclined moving walkways installed inside a tunnel at Bank Underground station in London. They were opened to the public in September 1960. They are inclined at around 8 degrees to the horizontal and have an approximate inline length of 90 m.

Recent Developments

A number of important developments have taken place in elevator technology in the last 30 years. Although the basic concept of the elevator has not changed since its invention more than 150 years ago, which is basically described as follows:

An electrical motor is used to rotate a sheave (usually through a gearbox that reduces the speed and increases the torque). The sheave (which is a pulley that has a number of grooves on its surface) has steel ropes in the grooves. When enough traction is present between the sheave and the steel ropes, it is possible to raise and lower the elevator car.

The most important of these developments are discussed below.

Advances in Motor and Variable Speed Drive Technologies

Recent advances in motor technology have led to more efficient and more compact motors. The progress of motor technology has progressed through a number of states historically, starting with brushed dc motors, then to geared induction motors driven by variable frequency drives and ending with permanent magnet synchronous motors (PMSM) driven by variable frequency drives.

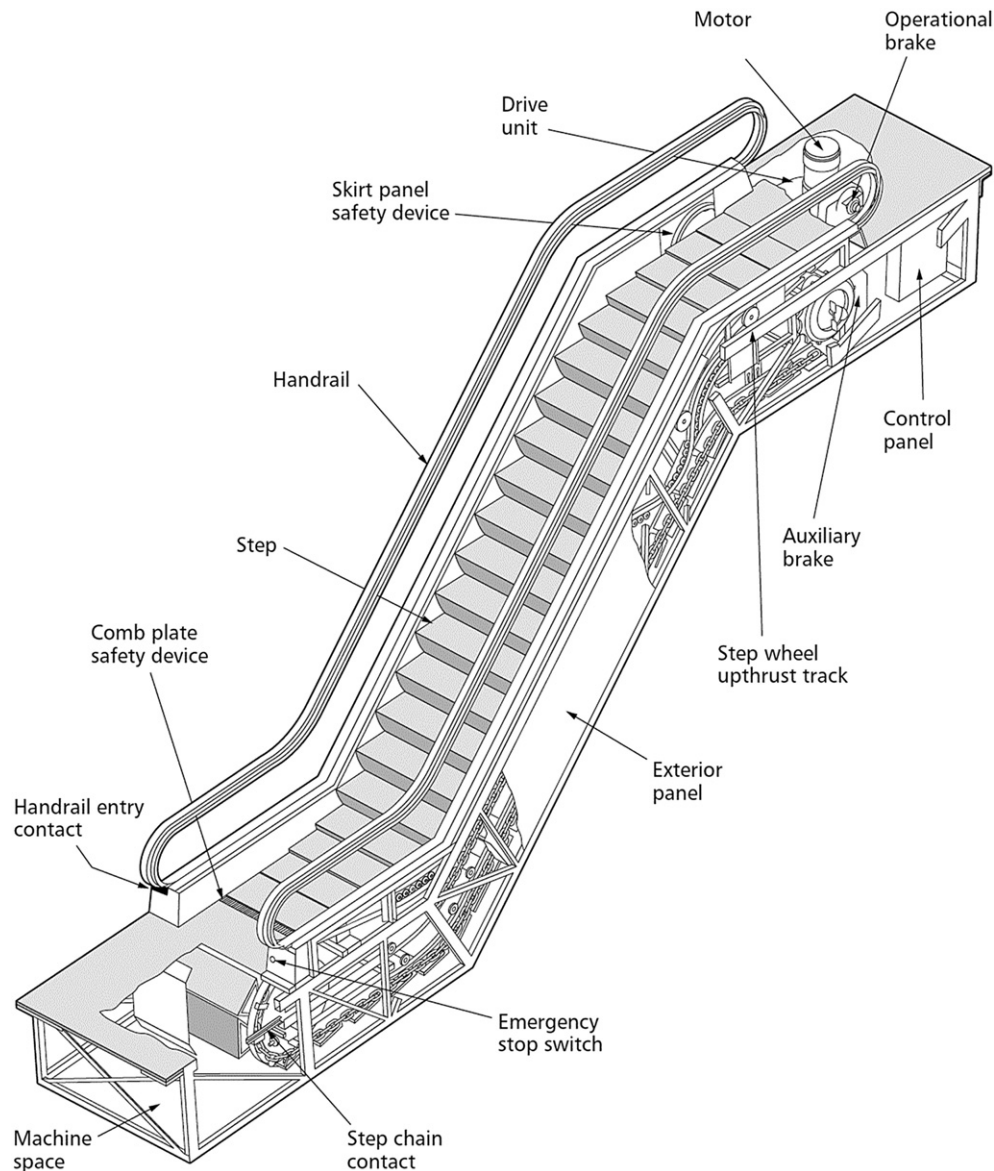


Figure 5 General Arrangement of a Typical Escalator *Source: courtesy of CIBSE, from CIBSE Guide D, 2015).*

Moreover, the advances in power electronics has revolutionized variable speed drive technology, leading to unprecedented levels of riding comfort for passengers as well as excellent leveling accuracies.

With modern motors and drive technologies and advanced electronic feedback systems, higher speeds are now possible where there are currently elevators in high rise building running at rated speeds of 17 m s^{-1} (17 m/s is equivalent to around 60 km/h), thus reducing the required time to arrive the floor destination in high rise buildings.

Intelligent and Sophisticated Control Systems

There has been a revolution in the processing power, speed and miniaturization of elevator control systems. These systems are employed in the elevator group control systems, allowing better allocation of landing calls to elevator cars in group control, thus reducing waiting times and journey times and avoiding queues. In most cases, these systems have relied on advanced passenger interfaces as well as the use of artificial intelligence tools (such as genetic algorithms, fuzzy logic and neural networks).

Most of the developments in this area have relied on the concept of “passenger grouping” (Christy, 2014). The group control system attempts to group the passengers with common destinations into the same elevator car, in order to reduce the number of stops in a round trip and thus reduce the value of the round trip time and boost the handling capacity.

One such control system that was introduced by the Otis Elevator Company is called channeling (an example is shown in Fig. 6). In this system, each elevator serves a specific set of floors in each journey, when it returns to the main entrance to pick up more



Figure 6 Example of a Channeling system installed in a building in Sydney, Australia.

passengers (Powell, 1992; Fortune, 2002). The floors that it will serve in the next journey are announced using a dynamic electronic display. Thus, passengers will have to board the specific elevator that is serving their destination floor in that journey. Passengers cannot simply board the next elevator car that arrived in the lobby; they must board the elevator that serves their destination floor.

Another group control system that employs passenger grouping is the destination group control. This technology was introduced in the early 1990s by the elevator company Schindler. Details about destination group control systems can be found in Peters (2006), Smith and Peters (2002), Schroeder (1990), Sorsa et al. (2005), Smith and Peters (2004), Lauener (2007), Al-Sharif et al. (2015). It works by asking the passenger to enter his/her destination prior to entering the elevator. The group controller would evaluate the most appropriate elevator car and allocate it to that landing call. It will ask the passenger to wait for that specific allocated elevator in order to board it when it arrives. Once the passenger boards the allocated elevator, his/her car call is already registered, and the elevator car will display it inside as one of the floors at which it will stop. An example of a landing call registration panel is shown in Fig. 7.



Figure 7 Destination Group Control landing call panel.

As mentioned earlier, the effect of passenger grouping in general is that it boosts the handling capacity at the expense of some increase in passenger waiting time.

Optimized Exploitation of Shaft Space in Buildings

A number of innovations have been introduced that aim to optimally exploit the elevator shaft space in buildings by introducing more compact elevator arrangements such as double decker systems and multiple elevators in the same shaft.

The machine room-less elevator was introduced by KONE Elevators in the late 1990s. During the company's search for a linear motor drive elevator, they came across the concept of using a radial (as opposed to an axial) permanent magnet synchronous motor. The advantages of using a permanent magnet synchronous motor, driven by a variable frequency drive system, offered the following advantages:

1. The radial motor (sometimes referred to as a "disk" motor or even informally as the "pancake" motor) was effectively two disks allowed to rotate relative to each other, one with coils fitted to it, and the other with permanent magnets fitted to it. By exciting the adjacent coils with the correct excitation currents, the resultant interaction between the magnetic flux densities of the coils and the permanent magnets allowed the precise control of the speed and torque. This allowed the motor to be run at much lower speeds than conventional induction motors while still providing the required torque.
2. This obviated the need for a gearbox. By eliminating the gearbox, the efficiency of the system improved and valuable space was released, enabling the motor to be fitted inside the elevator shaft, without the need for a machine room.

Since then, all the other manufacturers have introduced their own versions of machine room-less elevators, employing gearless permanent-magnet-synchronous-motors. However, all of them were axial motors rather than radial motors.

Another relevant development was introduced by ThyssenKrupp. It introduced the Twin elevator system that comprises two roped traction elevators sharing the same shaft (Fig. 8). Such a system provides a boost in the handling capacity of the elevator shaft, saving valuable net-lettable area in the building. A number of safety systems provide successive means of protection to prevent the collision of the two elevators cars.

Future Developments

The most important recent development which is expected to transform the landscape of indoor transportation is the new development of the MULTI, a system of elevators electromagnetically propelled vertically and horizontally in a rectangular shaft. Multiple elevator cars can move in the same shaft (hence the name MULTI). The typical arrangement would be a rectangular loop, whereby one vertical shaft is dedicated to the up direction and the other vertical shaft for the down direction. A prototype of this system is shown in Fig. 9.



Figure 8 Two Roped Elevators in the Same Shaft Source: courtesy of ThyssenKrupp Elevators, TKE.

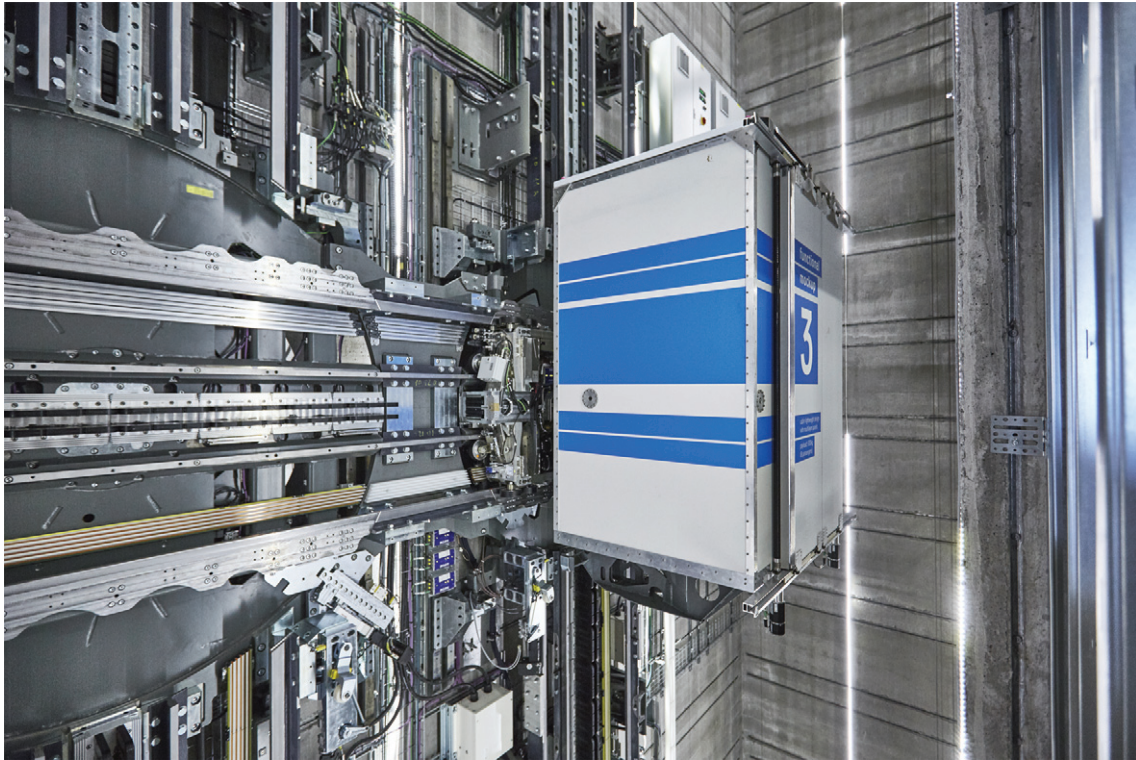


Figure 9 The prototype of a the first MULTI system *Source: courtesy of ThyssenKrupp Elevator.*

Such a system would potentially create savings in the required shaft space for the elevators and allow more flexibility in the movement of passenger within and between buildings. More details about two-dimensional elevator system in buildings can be found in [Gerstenmeyer and Peters \(2014\)](#), [Gerstenmeyer and Peters \(2016\)](#) and [So et al. \(2015\)](#).

A futuristic vision of these system includes elevators moving vertically and horizontally within a “mega-building”, delivering passengers to their exact destination, and even, moving between different buildings. This technology could eventually evolve into a personal “pod” for each passenger that transports him/her to their selected destination in other buildings or even to/from transport terminals such as railway stations and airport terminals. The issues of safety, privacy and security would be some of the challenges that need to be addressed.

Another area that is gaining momentum at the moment is the use of elevator systems in passenger evacuation in buildings. More details about this topic can be found in [CTBUH \(2004\)](#), [Wit \(2010\)](#), [Siikonen and Hakonen \(2002\)](#), [Fortune \(2010\)](#), [Klote \(1993\)](#).

Summary and Conclusions

This chapter has described the main components of indoor transportation, namely elevators, escalators, and moving walkways. Most of the chapter has concentrated on elevators, as they move the majority of passengers inside buildings. Specifically, the article has concentrated on the analysis of the roped traction elevator (the most widely used means of vertical transportation inside buildings), the traffic design process, space planning issues including preferred capacities and speeds, escalators and moving walkways. The chapter has concluded by discussing the most important recent technological developments in this area and the future technological developments. Based on these points, the following is a list of the major areas that require more research:

1. Energy consumption: There will be more focus in the future on the energy consumed in the elevator as well as whole life cycle analysis of its components.
2. There is a need to improve the traffic performance parameters in building transportation, in order to allow more transparency and common terminology.
3. With the advent of rope-less electromagnetically elevators, more research is needed in the area of rope-less propulsion systems and electromagnetically driven elevators in both vertical and horizontal shafts.
4. One of the most important future developments in indoor transportation is the integration of building transportation with intra-building transportation and urban transportation leading to truly integrated transportation systems.
5. There is currently significant demand to include elevator systems in the process of emergency evacuation of buildings. Research in this area is needed on two fronts: evacuation control systems as well as hardware components.

Biographical Notes



Lutfi Al-Sharif is currently Dean of Engineering and Professor of Electrical Engineering at Al-Hussein Technical University in Amman/Jordan and jointly Professor of Building Transportation Systems at the Department of Mechatronics Engineering, The University of Jordan. He received his Ph.D. in elevator traffic analysis in 1992, from the University of Manchester, U.K. He worked for 10 years for London Underground, London, United Kingdom in the area of elevators and escalators.

In 2002, he formed Al-Sharif VTC Ltd, a vertical transportation consultancy based in London, United Kingdom. He has over 30 papers published in peer reviewed journals the area of vertical transportation systems and is co-inventor of four patents and co-author of the 2nd edition of the Elevator Traffic Handbook.

He is also a visiting professor at the University of Northampton (UK), member of the scientific committee of the annual Symposium on Lift and Escalator Technologies and a member of the editorial board of the journal Transportation Systems in Buildings.

He is a passionate believer in making higher education simple and accessible for engineering students and has a YouTube channel on engineering that has more than 55,000 subscribers and around 8.5 million views. He has also been working as a member of the METHODS Erasmus+ Project that aims to improve teaching methods in higher education in Jordan and Palestine. He is also the author of the Mechatronics Engineering Module on Saylor.org.¶

See Also

As mentioned in the section on elevator traffic design, simulation is one of the main tools currently used in. Interested readers are also referred to the general use of simulation in transportation in the following two sections: 10186: Simulators and 10638: Modeling and Simulation for Transport Planning.

Surveys are widely used in indoor transportation to verify new installations and to plan for the upgrade of existing installation. More can be found about surveys in the section in the chapter entitled: 10387: Travel Surveys.

This section has been concerned with indoor transportation. For more on passenger outdoors transportation, refer to the chapter in this publication entitled: 10495: Cable Transport.

In indoor transportation, elevators are the means of transport that can be used by mobility impaired passengers. For more information on disability planning in transportation systems in general, readers are referred to the chapter in this publication: 10774: Disability in Transport Planning.

References

- Al-Sharif, L., 1996. Lift and escalator energy consumption. In: Proceedings of the CIBSE/ASHRAE Joint National Conference, Harrogate 1. pp. 231–239.
- Al-Sharif, L., 1999. Lift and escalator motor sizing with calculations & examples. Lift Report 52 (1.).
- Al-Sharif, L., 2014. Introduction and assessing demand in elevator traffic systems (METE I). Lift Report 40 (4), 16–24.
- Al-Sharif, L., Sep/Oct 2014. Calculating the elevator round trip time for the most basic of cases (METE II). Lift Report 40 (5), 18–30.
- Al-Sharif, L., May/Jun 2015. Introductory elevator traffic system design (METE VI). Lift Report 41 (3), 32–44.
- Al-Sharif, L., Al-Adhem, M.D., 2014. The current practice of lift traffic design using calculation and simulation. Build. Serv. Eng. Res. Technol. 35 (4), 438–445, doi:10.1177/0143624413504422, Published online 26 September 2013.
- Al-Sharif, L., Abdel-Aal, O.F., Abu-Alqumsan, A.M., Abuzayyad, M.A., 2015. The HARint Space: a methodology for compliant elevator traffic designs. Build. Serv. Eng. Res. Technol. 36 (1), 34–50, 0143624414539968.
- Al-Sharif, L., Hamdan, J., Hussein, M., Jaber, Z., Malak, M., Riyal, A., AlShawabkeh, M., Tuffaha, D., September 2015. Establishing the upper performance limit of destination elevator group control using idealised optimal benchmarks. Build. Serv. Eng. Res. Technol. 36 (5), 546–566.
- Al-Sharif, L., Al Sukkar, G., Hakouz, A., Al-Shamayleh, N.A., 2017. Rule-based calculation and simulation design of elevator traffic systems for high-rise office buildings. Build. Serv. Eng. Res. Technol. 38 (5), 536–562, doi:10.1177/0143624417705070.
- So A.T.P., Al-Sharif L., Hammoudeh A.T. Traffic analysis of a simplified two-dimensional elevator system. January 2015. Building Service Engineering 36(5).
- Barney, G.C., Al-Sharif, L., 2016. In: Elevator Traffic Handbook: Theory and Practice. 2nd ed. Routledge, Abingdon-on-Thames, United Kingdom.
- BSI, 2008. BS 9999:2008: Code of practice for fire safety in the design, management and use of buildings
- Christy, T., July 2014. An Evolution of Lift Passenger Grouping. Elevator Technology 20 2014; 20: 41-50, The International Association of Elevator Engineers (Brussels, Belgium), the proceedings of Elevcon 2014, Paris, France, pp. 8–10.
- CIBSE, 2000. In: CIBSE Guide D: Transportation systems in buildings. Fifth Edition. 2015. Published by the Chartered Institute of Building Services Engineers, Balham, London, United Kingdom.
- CTBUH. Emergency evacuation elevator systems guideline. Council on Tall Buildings and Urban Habitat, 2004. 47 pages
- B S EN 115-1:2008 + A1:2010: Safety of escalators and moving walks. Construction and installation. January 2009
- Finschi, L., State-of-the-art traffic, analyses., 2010. Elevator Technology 18. The International Association of Elevator Engineers (Brussels, Belgium), the proceedings of Elevcon 2010, Lucerne, Switzerland, pp. 106–115.
- Fortune, J.W., June 2002. Electronic Up-Peak Booster Options. In: Elevator Technology 12 2002; 12: 84-90 Proceedings of Elevcon 2002, Milan, Italy. The International Association of Elevator Engineers, Brussels, Belgium.
- Fortune, J.W., 3rd to 5th February 2010. Emergency Building Evacuations via Elevators, CTBUH 2010., Remaking Sustainable Cities in the Vertical Age, Mumbai, India.
- Fruin, J.J., 1987. In: Pedestrian Planning and Design. Elevator World.
- Gerstenmeyer, S., Peters, R., 25th to 26th September 2014. Reverse journeys and destination control. In: 4th Symposium on Lift and Escalator Technologies, University of Northampton. pp. 61–70.
- Gerstenmeyer, S., Peters, R., 2016. Safety distance control for multi-car lifts. Build. Ser. Eng. Res. Technol. 37 (6), 730–754.

- Hakonen, H., Siikonen, M.L., 2008. Elevator Traffic Simulation Procedure. Elevator Technology 17. The International Association of Elevator Engineers (Brussels, Belgium), in Proceedings of Elevcon 2008., 17. Thessaloniki, Greece, pp. 131–141.
- Harford, T., 2018. In: Fifty Things that Made the Modern Economy. 2nd ed. Abacus.
- Hirscher, P., 1987. Elevators: Technology, Planning, Operation. Thyssen Aufzüge. La Nouvelle Librairie 50–51.
- ISO, B.S., 2019. "BS ISO 8100-30: 2019 entitled: Lifts for the transport of persons and goods. Part 30: Class I, II, III and VI lifts installation".
- Janovsky, L., 1993. In: Elevator Mechanical Design. 2nd ed. IAEE, Ellis Horwood, N.Y.
- Klote, J.H., 1993. A method for calculation of elevator evacuation time. J. Fire Prot. Eng., 5 (3), 83–95.
- Lauener, J., 2007. Traffic performance of elevators with destination control. Elevator World (Mobile, AL, USA) 55 (9), 86–94.
- Markon, S., Kita, H., Kise, H., Bartz-Beielstein, T., 2006. In: Control of Traffic Systems in Buildings. Springer, NY, USA.
- Mayo, A.J., September 1966. A study of Escalators and Associated Flow Systems. In: M. Sc. Thesis. Imperial College of Science and Technology, University of London.
- Peters, R., June 2006. Understanding the benefits and limitations of destination dispatch. Elevator Technology 2006 ; 16 : 258-269, The International Association of Elevator Engineers (Brussels, Belgium), the proceedings of Elevcon 2006, Helsinki, Finland.
- Peters, R., September 2013. The Application of Simulation to Traffic Design and Dispatcher Testing. In: Proceedings of the 3rd Symposium on Lift and Escalator Technologies 2013; 3. The University of Northampton, Northampton, United Kingdom, pp. 128–139.
- Peters, R.D., Evans, E., 2008. In: Measuring and Simulating Elevator Passengers in Existing Buildings., Elevator Technology 17, proceedings of Elevcon 2008, Thessaloniki, Greece, pp. 289–300.
- Powell, B., May 1992. Important issues in up peak traffic handling. Elevator Technology 4 1992 ; 4 : 207-218, The International Association of Elevator Engineers (Brussels, Belgium), Proceedings of Elevcon, Amsterdam, The Netherlands, pp. 92.
- Schroeder, J., March 1990. Elevator calculation, probable stops and reversal floor, "M10" destination halls calls + instant call assignments. Elevator Technology 3 1990 ; 3 : 199-204, Proceedings of Elevcon '90, The International Association of Elevator Engineers, (Brussels, Belgium), Rome, Italy, .
- Siikonen, M.L., 1993. Elevator traffic simulation. Simulation 61 (4), 257–267, doi:10.1177/003754979306100409.
- Siikonen M L. On traffic planning methodology. Elevator Technology 10 2000; 10:267-274. The International Association of Elevator Engineers (Brussels, Belgium), the proceedings of Elevcon May 2000, Berlin/Germany
- Siikonen, M.L., Hakonen H., Efficient Evacuation Methods in Tall Buildings. Elevator Technology 12, 202; 12: 237-246. Proceedings of Elevcon 2002, Milan, Italy, June 2002, IAEE.
- Siikonen, M.L., Susi, T., Hakonen, H., November 2000. Passenger Traffic Flow Simulation in Tall Buildings. IFHS. In: International Conference on Multi-Purpose High-Rise Towers and Tall Buildings.
- Smith, R., Peters, R., June 2002. ETD algorithm with destination dispatch and booster options. In: Elevator Technology 12 2002 ; 12 : 247-257. The International Association of Elevator Engineers (Brussels, Belgium), proceedings of Elevcon 2002, Milan, Italy.
- Smith, R., Peters, R., April 2004. Enhancements to the ETD dispatcher algorithm. In: Elevator Technology 2004 ; 14 : 234-243, The International Association of Elevator Engineers (Brussels, Belgium), Proceedings of Elevcon 2004, Istanbul, Turkey.
- Sorsa, J., Hakonen, H., Siikonen, M.L., 7th to 9th June 2005. Elevator selection with destination control system. In: Elevator Technology 15 2005 ; volume 15, The International Association of Elevator Engineers (Brussels, Belgium), proceedings of Elevcon, Peking, China.
- Stanhope, February 2004. In: Lift Performance Specification. Position Paper, London, United Kingdom.
- Strakosch, G., 1998. In: The Vertical Transportation Handbook. John Wiley & Sons.
- To, A., Yip, P.K., 11th to 13th October 2004. The challenge of vertical transportation system for two IFC. In: First International Conference on Building Electrical Technology (BETNET) 2004; 1. The Institute of Engineering and Technology, Hong Kong, pp. 152–157.
- Wit, J., June 2010. Evacuating Using Elevators, Enhancing the CTBUH Approach for the Dutch High-Rise Covenant. In: Elevator Technology 18, 2010; Proceedings of Elevcon Lucerne 2010, IAEE 18. pp. 419–430.
- Wit, J., 30th July 2017. In: Elevator Planning for High-Rise Buildings. Deerns Consulting Engineers Publication, The Netherlands, Amsterdam, www.deerns.com. Access on.

Further Reading

- Strakosch, G., 1998. The Vertical Transportation Handbook. Third edition. Wiley.
- Barney, G.C., Al-Sharif, L., 2016. Elevator Traffic Handbook: Theory and Practice, 2nd ed. Routledge, Abingdon-on-Thames, United Kingdom.
- CIBSE. CIBSE Guide D: Transportation systems in buildings. Published by the Chartered Institute of Building Services Engineers, Fifth Edition, 2015.
- Janovsky, L., 1999. Elevator Mechanical Design, 3rd ed. Elevator World Inc.
- Markon, S., Kita, H., Kise, H., Bartz-Bielstein, T., 2006. Control of Traffic Systems in Buildings. Springer.
- Andrew, J.P., Kaczmarczk, S., 2011. Systems Engineering of Elevators. Elevator World.
- Hirscher, P., 1987. Elevators: Technology, Planning, Operation. La Nouvelle Librairie.
- Proceedings of the Symposium on Lift and Escalator Technologies: <https://liftsymposium.org/>. (10th, Sep. 2019 Northampton; 9th, Sep. 2018 Northampton; 8th, May 2018, Hong Kong; 7th, Sep. 2017 Northampton; 6th, Sep. 2016 Northampton; 5th, Sep. 2015 Northampton; 4th, Sep. 2014 Northampton; 3rd, Sep. 2013 Northampton; 2nd, Sep. 2012 Northampton; 1st, Sep. 2011, Northampton, United Kingdom).2000.
- Proceedings of the International Congress on Elevator Technologies (Elevcon), 1-22 (1986 to 2019).

Bicycle as a Transportation Mode

Raktim Mitra*, Paul M. Hess[†], *School of Urban and Regional planning, Ryerson University, Canada; [†]Department of Geography and Planning, University of Toronto, Canada

© 2021 Elsevier Ltd. All rights reserved.

Introduction: Bicycle as a Transportation Option	697
A Social-Political History of Bicycles	697
Factors that Influence Cycling	698
Bicycle Infrastructure and Active Transportation Planning	699
Conclusion: Cycling to the Future?	700
References	701
Further Reading	701

Introduction: Bicycle as a Transportation Option

Bicycles, or simply-bikes, are single-track (i.e., with two wheels) pedal-driven vehicles, and are used across the world for both transportation and recreational purposes. As a mode of transportation, bicycles are inexpensive to purchase and operate, take little road space and have zero emissions. Bicycling (or just “cycling” or “biking”) also provides an opportunity for endurance-type physical activity when used frequently for transportation. For these reasons, advocates, professionals, and some policy makers have highlighted the bicycle as an environmentally sustainable and healthy alternative to automobiles as a transportation choice, and a potential solution to many of the problems posed by the current automobile-dominant transportation systems. Bicycles are also seen as an important component of urban transportations system to ensure “basic mobility”, in other words, to provide basic access to destinations by large population groups regardless of their socioeconomic conditions.

Many of us learned how to bicycle as children and across much of the world, bicycles are popularly used by children and youth to explore their neighborhoods and to be independently mobile. In some societies, particularly in parts of Asia and Europe, bicycles are also seen as an important mode of transportation for the entire population including adults for going to work, school, shopping, and other destinations (e.g., accessing transit stations). In other societies, including much of North America, bicycles are still largely seen as a toy for children, as an exercise equipment or as an undesirable transportation option used mostly by low income people. While these perceptions are changing, they remain a key barrier to wider adoption of bicycles as a transportation mode for working adults, as well as to the development of cycling related policy and infrastructure. In this entry, we focus on bicycles for their potential role as a transportation mode.

For the most part, cycling is a human powered transportation option, although electrically powered bicycles, called e-bikes, are becoming more common. Thus, bicycles are considered most appropriate for short- to medium-distance trips, with North American transportation policies often considering 2–5 km as a typical bikable distance. This bears out in practice. For example, 74% of all bicycle trips in the Toronto Region (the Greater Toronto and Hamilton Area- GTHA) in Canada in 2011 were less than 5 km in length. In some European countries, longer distance bicycle trips are not uncommon and have gained policy attention particularly in the last decade. Countries including the Netherlands, Germany and England have invested in what is called “bicycle highways” to enable longer distance (10–15 km) bicycle trips for transportation purposes.

Pucher and Buehler’s (2012) international comparison of cycling rates further confirms large national differences in bicycle use. While some European countries reported very high-cycling rates, including the Netherlands (26% in 2008), Denmark (18%), and Germany (10%), the mode share of bicycle for daily trips remained very low in other parts of the western world, including 2% in the United Kingdom (UK) (2008), and 1% in the United States (US) (2009). In Canada, 1.5% of population commuted via bicycles in 2016. However, the use of a bicycle is on the rise across these countries compared to even two decades ago. For example, Le et al. (2019) examined bicycle counts at 1319 locations in 13 US metropolitan areas, between 2004 and 2016. Their analysis indicates that cycling volume has increased between 2% and 6% per year at these count locations, controlling for other variations.

The chapter is organized as follows. The next section discusses the social-political history of the bicycle, including national differences in cycling adoption, particularly in Europe and North America. The following section briefly summarizes factors that influence cycling rates as explored in the existing literature, including demographic factors, built environment conditions, and personal attitudes. The paper then turns to the role of bicycle infrastructure development and policy in promoting cycling, and finally concludes by discussing the trajectory of cycling as an increasingly important mode of urban transportation.

A Social-Political History of Bicycles

The early versions of the modern bicycle, featuring two equal sized wheels and a chain-drive, was developed in the late 1800s. These early bicycles were a symbol of the modernization of transportation and were a faster and more practical alternative to horses, which

were the dominant mode of personal transportation at that time (Oosterhuis, 2016). While this new form of mobility was introduced to postal services, the police, fire departments and the army, the general public overwhelmingly adopted bicycles for recreation and sport rather than daily transportation.

Some scholars also suggest that bicycles have played an important role in women's rights movements across the western world (Bopp et al., 2018), when woman had little presence in public life. Gendered expectations for woman included restricted mobilities, with the majority of women in the late 1800s walking to their destinations (primarily relating to household shopping) or using short trolley trips. In this context, bicycles offered an opportunity for longer, more independent travel, and the popularity of bicycles among women grew rapidly in this period. However, female cyclists were criticized for their presence outside of the domestic sphere, and for what was seen as their "liberated attitude" and inappropriate attire (Bopp et al., 2018; Oosterhuis, 2016). As a result, many women cyclists would ride along with men to avoid the "male gaze" and to avoid being painted as symbols of militant feminism (Oosterhuis, 2016), and the overall proportion of women among cyclists was low. Even to this date, gender equity in cycling has yet to be established in many societies. While in some parts of Europe including the Netherlands, Denmark, Germany, and Sweden, there is no significant gendered difference in cycling rates, less than a third of all cyclists in North America are women (Garrard et al. 2012).

Early advocates for bicycle use and better road infrastructure to support cycling in the late 1800s were often wealthy, politically influential people. Before massive road construction to accommodate motor-vehicles, their efforts to establish the right to free mobility on public roads for bicycles was instrumental in the institutionalization of the development of transportation infrastructure as an active government responsibility (Bopp et al., 2018). However, when mass production and lower prices made bicycles more accessible to a larger population, bicycle use began to lose support from elites. Then, in the 20th century, motorized personal vehicles (i.e., cars) quickly came to be recognized as the *avant garde* mode of transportation in many western societies including the US, Canada, England and Australia (Oosterhuis, 2016). In this period, Transportation Engineering and Urban Planning were emerging professions that recognized automobiles as the mobility option for the future and sought to rebuild cities to accommodate motorized traffic. With the exception of some short-lived periods of popularity since the 1940s, the bicycle has continued to fight for road space against automobiles (Wilson and Mitra, 2020).

Some western societies outside the English-speaking world, however, have experienced different trajectories of social and political support for bicycles, and the bicycle has remained much more central to their transportation cultures. In the Netherlands and Denmark, for example, bicycle had already become a national symbol by the 1920s (Oosterhuis, 2016). In the Netherlands, civic organizations emphasized bicycles in terms of "Dutch qualities and civil virtues" and the use of bicycles as a transportation mode was actively promoted. In contrast to the English-speaking world, when bicycles become more accessible to a wider population, they were seen as a "democratic" mode of transportation. These cultures, too, promoted automobile ownership and use after World War II, and there were no significant post-war policy focusing on bicycles until the 1970s oil-shocks. However, bicycles always enjoyed more support and use from all levels of the society. This continues to be the fact to this day, with bicycles seen as an integral element of their urban everyday life.

In contrast, countries such as France, Belgium and Italy embraced cycling as a serious sport, which resulted in the development of world-famous racing/ touring events including the *Tour de France* and *Giro d'Italia*. The commercialization of professional cycling created wide popularity of this sport among both rich and the poor, and the cycling culture in these countries continue to be strongly associated with cycling-related sports.

Still, across Europe today, promoting cycling as a transportation mode and developing extensive systems of bicycle infrastructure is seen as key element to urban sustainability. Some Latin American cities, too, are building extensive cycling systems. Bogota, Columbia is considered a leader in investing heavily in bicycle infrastructure starting in the 1990s to promote bicycle uses as a low-cost transportation mode accessible to the majority of the population. More recently, cycling has received much serious attention as a transportation mode in some North American cities, with New York City, Portland (Oregon), Montreal, Vancouver, and other cities seen as being leaders with accelerated infrastructure investment since the 2000s.

Factors that Influence Cycling

Scholarly literature on cycling behavior is relatively new and smaller compared to the literatures on driving and public transportation use. However, over the past two decades, this literature has grown substantially, and more extensive summaries of current research on cycling behavior can be found elsewhere (e.g., Bopp et al., 2018; Pucher and Buehler, 2012).

Sociodemographic factors such as gender, income, age, and automobile ownership may influence cycling, but the evidence remains mixed. Ironically, while many people view cycling as a transportation option for the poor, some researchers have found that higher income individuals are more likely to bicycle than lower income individuals, and also that cycling is more common among Caucasian sub-population compared to people of color. In addition, younger people tend to favor bicycles more than the older population.

Evidence on gendered differences in cycling is unclear. North American research has consistently reported lower cycling rates among women. For example in the GTHA, Canada, only 29% of all cyclists were women. In the United States, 24% of all cycling trips were taken by women. Further, between 2000 and 2010, the number of female cyclists in the United States actually went down by 13% (www.bicycle-guider.com). But such gender difference is almost non-existent in communities and countries with stronger cycling culture and relatively higher cycling rates. In some inner urban neighborhoods in Toronto, Canada, for example, where more

than 10% adults commute to work/school by bicycle, the cycling rates are similar for both men and women. While the causal relationship is unclear, advocates often argue that encouraging women to bicycle is an important way to promote and increase bicycling rates at the population level. Chapter 10637 of this volume offers a more broader discussion of gendered mobility.

In terms of the urban built environments, researchers have provided evidence of correlations between cycling rates and measures of population/ development density, land use mix and traffic and personal safety (Bopp et al., 2018). Communities with higher densities have higher rates of walking/cycling than low-density neighborhoods. However, other researchers point to the roles of land use mix and street design characteristics as being key factors beyond population density. As discussed in the next section, cycling advocates and professionals have always emphasized bicycle infrastructure as a critical factor in promoting the bicycle and advancing cycling at a population level.

Previous studies have also emphasized the importance of travel distance on the likelihood of cycling. For example, Dill & Carr (2003) reported that trip distance was the most frequently cited reason for avoiding cycling as a means to commute to work, although this conclusion may be particularly relevant to North America. In Europe, the use of bicycles for long distance commutes and as a means to complete the first/last mile trips (e.g., trip to train station) is more common.

In addition, the decision to bicycle is clearly influenced by personal attitudes, social norms, and self-efficacy or confidence, compared to some other major modes of transportation such as driving or taking transit where it is believed that the generalized travel cost is the key determinant of travel mode choice (see Chapter 10513 for a detailed discussion of the generalized cost of travel). Although limited, existing research has consistently reported that those who enjoy cycling are more likely to commute by bicycle. Positive attitudes toward cycling and being environmental conscious, also may encourage cycling for transportation (Bopp et al., 2018). Chapter 10675 of this volume provides a detailed discussion of the behavioral psychology of cyclists. Because cycling is generally considered to be a less safe mode of transportation compared driving or taking transit, confidence in cycling has been identified as a key indicator of whether and how frequently an individual bicycles for transportation. An emerging literature on cyclist typology suggests that only very confident cyclists are willing to bicycle in traffic with no or little dedicated infrastructure (Dill and McNeil, 2013). Most of the population who are interested in bicycling, however, report they desire high-quality dedicated infrastructure before they would consider regularly cycling for transportation due to safety-concerns

Bicycle Infrastructure and Active Transportation Planning

Chapter 10116 of this volume discussed bicycle infrastructure. Bicycle infrastructure is arguably a key enabling factor to improve cycling rate. Very high cycling rates in some European countries, such as the Netherlands, is often credited to the existence of expansive and well-designed bike-path network and other bicycle-related infrastructure and facilities. Previous research has consistently reported an association between the availability of bicycle parking and commuting by bicycle (Bopp et al., 2018). While bicycle parking and storage is an important component of many active transportation plans, most policy and plans focus on providing painted bicycle lanes, on-street separated cycle tracks or off-street bicycle paths/trails. For example, between 2014 and 2016, we examined nine Complete Street projects and 13 active transportation improvement projects in the Greater Golden Horseshoe Region of Canada (i.e., the large urban region around the City of Toronto). Every project included an improvement to bicycle infrastructure.

The association between higher rates of cycling for commute purposes and the presence of bicycle infrastructure is consistently shown in the literature. For example in US cities, higher amounts of bicycle infrastructure has been shown to be related to higher rates of bicycle commuting regardless of socioeconomic factors, land use, cycling safety, transit supply, gasoline price, and climate (Buehler and Pucher, 2012). Bicycle infrastructure improves accessibility between locations and creates a safer bicycling environment, thereby may encourage more cycling by both current and “new” cyclists. However, it is difficult to isolate the impacts of a particular infrastructure investment, such as one new bicycle lane on a street segment especially as bicycle infrastructure is easier to implement (and expand) in locations where there is stronger community support and political leadership (Wilson and Mitra, 2020). Thus, the relationship between infrastructure supply and demand may be reciprocal, with cycling demand and a culture contributing to the development of a robust bicycle infrastructure.

In this context, an emerging body of research has attempted to systematically examine bicycle infrastructure, but with only mixed results. Several studies have surveyed cyclists who were using a newly built cycling facility (e.g., cycle track, multi-use trail) in the US and Canada, and found that at least a quarter of current users would have chosen another mode for a similar trip in the absence of a bicycle infrastructure (Mitra et al., 2016; Rowangould & Tayarani, 2016). In Australia and the UK, studies using case control and/or longitudinal data did not find a statistically significant mode substitution effect (e.g., driving to cycling) as a result of a new bicycle infrastructure (Rissel et al., 2015; Song et al., 2017). Instead, an increase in bicycle counts on streets with bicycle infrastructure were attributed to a potential route substitution, where those who were already cycling on other streets were using the newer and potentially safer infrastructure.

More recently, our study of 17 neighborhoods in Toronto, Canada and the surrounding urban region (11 with a bicycle infrastructure; six without) produced more in-depth insights. After two-to-three years of construction, those living near a new bicycle infrastructure were more likely to commute by bicycle regularly (at least once a week) compared to those living in neighborhoods without a cycling facility (Odds Ratio = 1.5), but this difference was not statistically significant. Instead, those who were already cycling regularly before 2017, were cycling more frequently in 2019 when they lived near a newly built bicycle infrastructure, compared to those regular cyclists who lived in neighborhoods with no new bicycle infrastructure (Odds Ratio = 2.85). The results are not entirely unexpected. In North America in particular, bicycle lane/ tracks are often planned and

built in isolation without proper connectivity to the broader network of bicycle infrastructure. Creating better connected networks of on-street bicycle infrastructure would produce meaningful improvements to accessibility and cyclist safety and may have a more measurable impact. Besides, creating safer transportation conditions for existing cyclists is an important policy rationale.

Perhaps the biggest limitations of advancing bicycle infrastructure are public opposition and as a result, the lack of political support toward on street bicycle lanes and cycle tracks (Wilson and Mitra, 2020; Wild et al., 2018). Building off-street bike trails that do not compete for roadway space is more popular and is less politically charged than on-street facilities, but are still challenging to finance. Wilson and Mitra (2020) explain this opposition by situating bicycle planning within a broader cultural context where automobiles are not just seen as a means for transportation, but are considered an extension of the private sphere and something that protects families from “dangers” of the outside environment. Not surprisingly, maintaining the dominance of automobiles on roads is a key popular and political priority.

Even in cities with relatively high cycling rates, such as Copenhagen, Denmark; Seville, Spain; New York, US and Vancouver, Canada, imposing limits on cars, or taking away a driving lane from an urban road is a difficult undertaking (Wilson and Mitra, 2020). Building off-street facilities also requires political support and substantial funding. This is seen, for example, with the recent European experience of building long-distance “bicycle highways.” In the Netherlands, with its long history of public support for cycling, the central government supported communities in building 30+ of these facilities. In comparison, in Germany, the central government left financing in the hands of regional and municipal governments, which significantly slowed down the implementation due to lack of resources. Competition for road space and funding between motor vehicles and bicycles remains and may continue to be a major policy and professional challenge in improving cycling infrastructure.

Still, the bicycle as a mode of transportation is becoming more popular across the western world, influencing policy and active transportation plans that are creating new opportunities to promote cycling and create bicycle infrastructure. The European Commission, for example, advocates that Sustainable Urban Mobility Plans (SUMP) to be developed by all towns and cities, promoting modes with low-environmental impacts such as walking and cycling (Consult, 2019). Research networks such as Physical Activity Through Sustainable Transport Approaches (PASTA; www.pastaproject.eu) have also been established that aim to further demonstrate the health benefits of cycling. In the US, over 1400 agencies at all levels of government have adopted Complete Streets Policies that often focus on improving cycling safety, and 31 States now recognize a bicycle as a vehicle, allowing laws and regulations focusing on the safety of the cyclists (Bopp et al., 2018). In 2010, the US Department of Transportation called for the collection of more active transportation data, resulting in a range of tools and approaches to measure bicycling in US communities (Bopp et al., 2018). Since 2005, Federal Government commitment to the Safe Routes to School (SRTS) programs has resulted in the construction of significant bicycle infrastructure. In Canada, Montreal, Quebec has long been a leader in promoting both urban and long distance cycling, but cycling planning and infrastructure development is also becoming institutionalized in the rest of the country, with provincial funding being provided for local projects, and all major cities actively developing new facilities.

Conclusion: Cycling to the Future?

Bicycles predate automobiles as a form of modern, urban transportation mode. In much of the western world, this role was displaced by motor-vehicles, as 20th century cities were increasingly planned for and became dominated by automotive travel, particularly so in North America. Bicycles not only became marginalized on increasingly hostile roadways, but, outside of childhood and recreational contexts, were also associated with marginalized populations. Bicycles were perceived as a transportation option for people who did not have “better” options.

More recently, however, northern European cities such as Amsterdam, the Netherlands and Copenhagen, Denmark have become models for what cycling can be, with some 35% or more of work trips made by bicycle. This achievement required decades of activism, government policy, and infrastructure investment to realize, but also demonstrate what could be possible in other cities. Indeed, almost all European cities have now been influenced by these lessons and see the bicycle as a key to improving urban sustainability, public health and social equity. Even large cities, such as London and Paris, are reallocating road space away from automobiles, and dedicating it to cycling facilities. While political and financial support can be hard fought, bicycles are, once again, seen as a modern and progressive form of transportation, suitable for adults as well as children, for women as well as men and, indeed, as a low-cost, zero-carbon, healthy form of transportation for all classes of society.

Cycling rates are much lower in North America, but cycling and bicycle infrastructure are also increasing in the US and Canadian cities. Whereas cycling is still seen as being low status by some parts of the population, research shows that cycling is actually positively associated with income and it is, in fact, in desirable, gentrifying central-city neighborhoods where cycling rates are the highest. Not only are these built environments more supportive for cycling than in neighborhoods developed after World War II, but their residents with many transportation options are choosing cycling as a desirable mode of travel that is representative of modern, urban lifestyle. Indeed, some marginalized neighborhoods have resisted the building of new bicycle lanes, seeing them as a harbinger of gentrification. Still, Portland, Oregon, and New York City have become their own models for bicycle planning. In the last decades, for example, New York has added over 2000 kilometers of bicycle lanes and New York’s expertise has been incorporated into guidelines published by the National Association of City Transportation Officials (NACTO, 2012), itself having been established only 25 years ago. That expertise is now being exported across North American, and often to other cities around the world.

There are major opportunities and challenges for continuing this trajectory. In particular, the planning response to COVID-19 pandemic has accelerated existing cycling plans and new cycling infrastructure has been rapidly created in scores of cities. People avoiding public transit due to health concerns has led to dramatic increases in cycling rates, but others may be turning to more auto use, potentially creating increased competition for roadway space. It is not yet clear where new bicycle infrastructure (that are being built as part of the post-COVID19 recovery strategies) will be made permanent, and where it will be pared back, nor what the “new normal” in travel patterns may be as public health concerns from COVID-19 become less pronounced. This will clearly vary geographically. In North America, for example, the built environment conditions outside of city cores is often difficult for cycling, with large, multi-lane arterial streets designed for cars, low densities, and long distances between destinations, all features that research suggests will limit bicycle use. While there is more space in these areas and some suburban municipalities are building off-street paths, reallocating space or capital funding away from automobiles faces political opposition and as a result, cycling facilities are often fragmented, rarely forming continuous networks of bicycle infrastructure. Thus, it is clear that tensions will remain between environments planned and designed for automobiles, and those that accommodate safe and convenient cycling.

It is also clear, however, that cycling, even in the suburban contexts, was on the rise both in areas with and without cycling facilities even before COVID-19 pandemic. This will continue to create additional demand for infrastructure and continue to build a cycling culture in a virtuous circle. In a world where environmental sustainability, health, and social equity are seen as key issues, bicycles are being seen as contributing to solutions. The emergence of new micro-mobility options such as electrically powered bicycles (e-bikes) and electric pedal scooters (e-scooters) and their recent popularity are something to be acknowledged in this context. In the past few years, e-bikes and e-scooters have become common features in hundreds of cities across Asia, Europe and North America, primarily in the form of pay-per-use (e.g., bike-share, scooter-share) services. Whether or not these new mobility options will impact the rates of cycling in the coming decade remains to be seen. But the future of cities increasingly looks like one in which bicycles will play a central role.

References

- Bopp, M., Sims, D., Piatkowski, D., 2018. *Bicycle for Transportation: An Evidence-based for Communities*. Elsevier, Amsterdam.
- Buehler, R., Pucher, J., 2012. Cycling to work in 90 large American cities: new evidence on the role of bike paths and lanes. *Transportation* 39, 409–432.
- Consult, R. (Ed.), 2019. *Guidelines for Developing and Implementing a Sustainable Mobility Plan*. 2nd ed.. https://www.eltis.org/sites/default/files/sump_guidelines_2019_interactive_document_1.pdf.
- Dill, J., Carr, T., 2003. Bicycle commuting and facilities in major U. S. cities: If you build them, commuters will use them – Another look. *Transp. Res. Rec.* 1828, 1–9.
- Dill, J., McNeil, N., 2013. Four types of cyclists? Examination of typology for better understanding of bicycling behavior and potential. *Transp. Res. Rec.* 2387, 129–138.
- Garrard, J., Handy, S., Dill, J., 2012. Women and cycling. In: Pucher, J., Buehler, R. (Eds.), *City Cycling*, MIT Press, Cambridge, MA.
- Le, H.T.K., Buehler, R., Hankey, S., 2019. Have walking and bicycling increased in the US? A 13-year longitudinal analysis of traffic counts from 13 metropolitan areas. *Transp. Res. Part D* 69, 329–345.
- Mitra, R., Ziemba, R.A., Hess, P.M., 2016. Mode substitution effect of urban cycle tracks: Case study of a downtown street in Toronto, Canada. *Int. J. Sustain. Transp.* 11 (4), 248–256.
- National Association of City Transportation Officials (NACTO), 2012. *Urban Bikeway Design Guide*, 2nd edition. Available from: <https://nacto.org/publication/urban-bikeway-design-guide/>.
- Oosterhuis, H., 2016. Cycling, modernity and national culture. *Soc. Hist.* 41 (3), 233–248.
- Pucher, J., Buehler, R., 2012. *City Cycling*. MIT Press.
- Rissel, C., Greaves, S., Wen, L.M., Crane, M., Standen, C., 2015. Use of and short-term impacts of new cycling infrastructure in inner-Sydney, Australia: A quasi-experimental design. *Int. J. Behav. Nutr. Phys. Act.* 12 (129), doi:10.1186/s12966-015-0294-1.
- Rowangould, G.M., Tayarani, M., 2016. Effect of bicycle facilities on travel mode choice decisions. *J. Urban Plan. Develop.* 142 (4), [https://doi.org/10.1061/\(ASCE\)UP.1943-5444.0000341](https://doi.org/10.1061/(ASCE)UP.1943-5444.0000341).
- Song, Y., Preston, J., Ogilvie, D., 2017. New walking and cycling infrastructure and modal shift in the UK: A quasi-experimental panel study. *Transp. Res. Part A Policy Pract.* 95, 320–333.
- Wild, K., Woodward, A., Field, A., Macmillan, A., 2018. Beyond ‘bikelash’: Engaging with community opposition to cycle lanes. *Mobilities* 13 (4), 505–519.
- Wilson, A., Mitra, R., 2020. Implementing cycling infrastructure in a politicized space: Lessons from Toronto, Canada. *J. Transp. Geogr.* 86, <https://doi.org/10.1016/j.jtrangeo.2020.102760>.

Further Reading

- Aldred, R., Watson, T., Lovelace, R., Woodcock, J., 2019. Barriers to investing in cycling: Stakeholder views from England. *Transp. Res. Part A* 128, 149–159.
- Aldred, R., Jungnickel, K., 2014. Why culture matters for transport policy: The case of cycling in the UK. *J. Transp. Geogr.* 34, 78–87.
- Dill, J., 2009. Bicycling for transportation and health: The role of infrastructure. *J. Pub. Health Policy* 30 (1), S95–S110.
- Furness, Z., 2010. *One Less Car: Bicycling and the Politics of Automobility*. Temple, Philadelphia.
- Golub, A., Hoffmann, M.L., Lugo, A.E., Sandoval, G.F. (Eds.), 2016. *Bicycle Justice and Urban Transformation. Biking For All?* first ed.. Routledge, New York.
- Piatkowski, D.P., Marshall, W.E., Krizek, K.J., 2019. Carrots versus sticks: Assessing intervention effectiveness and implementation challenges for active transportation. *J. Plan. Edu. Res.* 39 (1), 50–64.
- Saelens, B.E., Sallis, J., Frank, L.D., 2003. Environmental correlates of walking and cycling: Findings from the transportation, urban design, and planning literatures. *Ann. Behav. Med.* 25 (2), 80–91.
- Stehlin, J.G., 2019. *Cyclescapes of the Unequal City: Bicycle Infrastructure and Uneven Development*. University of Minnesota Press, Minneapolis, MN.

Urban Air Mobility: Opportunities and Obstacles

Adam Cohen, Susan Shaheen PhD, University of California, Berkeley, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Nomenclature	702
Introduction: What is Urban Air Mobility and On-Demand Aviation?	702
Potential Barriers and Challenges	703
Potential UAM Use Cases	703
The Role of Research, Pilot Demonstrations, and Regulation	703
Conclusion: Leveraging COVID-19 Medical and Sanitation Use Cases to Build Community Awareness	708
Acknowledgment	708
Biographies	708
See Also	709
References	709
Further Reading	709

Nomenclature

eVTOL electric vertical take-off and land
UAM urban air mobility
UAS unmanned aerial systems
UTM unmanned aircraft systems traffic management
AAM advanced air mobility
VTOL vertical take-off and land

Introduction: What is Urban Air Mobility and On-Demand Aviation?

Urban Air Mobility (UAM) (also known as advanced air mobility) is an emerging concept that envisions a safe, sustainable, affordable, and accessible air transportation system for emergency management, cargo delivery, and passenger mobility within or traversing a metropolitan area. UAM is part of a broader ecosystem of on-demand mobility where consumers can access mobility and goods delivery services on demand by dispatching or using urban aviation services, courier services, shared automated vehicles, shared mobility, public transportation, and other innovative and emerging transportation technologies.

Across the globe, a number of converging innovations, such as advancements (e.g., vertical lift, electrification, and automation) and the growth of app-based services are contributing to the concept of on-demand aviation, including UAM and rural air mobility (for less urbanized areas). Although the concept of UAM is not new - several helicopter services began providing early UAM services in Los Angeles, New York City, San Francisco, and other metropolitan areas in the 1950s, 1960s, and 1970s. In recent years, on-demand aviation services similar to transportation network companies (TNCs) dispatched through a smartphone app have entered the marketplace (Cohen et al., 2020). For example, in the United States, BLADE provides helicopter services booked through a smartphone app in numerous cities (e.g., Los Angeles, Miami, New York City, and San Francisco). BLADE employs third-party operators that own, manage, and maintain their aircraft under federal regulations that govern intrastate air taxis and commuter services (BLADE, 2019). Similar helicopter services, such as Airbus' Voom are operational in Mexico City, Sao Paulo, and the San Francisco Bay Area. Travelers can use Voom to request on-demand flights from partner air taxi companies. Trips can be booked anytime between one hour to 90 days in advance through the Voom app or website (Voom, 2020).

In recent years, several companies are developing and testing enabling elements of a contemporary UAM concept including prototypes of vertical take-off and landing (VTOL) capable aircraft, operational concepts, and market studies to understand potential business models. Original equipment manufacturers are developing a variety of piloted, remote piloted, partially automated, and fully autonomous aircraft for a variety of applications. Common terms and definitions are summarized in Table 1.

In the United States, The National Aeronautics and Space Administration's (NASA) Advanced Air Mobility (AAM) National Campaign aims to improve UAM safety and accelerate scalability through demonstrations beginning in 2020. In addition to public sector research, a number of TNCs, such as Uber and Grab, are also working to develop their own UAM delivery and passenger services in partnership with electric VTOL (eVTOL) aircraft and small unmanned aerial systems (also known as sUAS and drone) manufacturers. A number of planned delivery services (for consumer goods and medical supplies) using sUAS/drones are being explored. Planned passenger services using eVTOL include Dallas, Los Angeles, Melbourne, and Singapore (Etherington, 2020; Hawkins, 2017).

Table 1 Common terms and definitions

Small Unmanned Aircraft	Aircraft that weigh less than 55 pounds (or 25 kg) on take-off, including everything that is on board or otherwise attached.
Small Unmanned Aircraft System (sUAS)	Small unmanned aircraft and its associated elements (including communication links and the components that control the small unmanned aircraft) that are required for the safe and efficient operation of the small unmanned aircraft in the national airspace system.
Vertical Take-off and Land (VTOL)	An aircraft that can take off, hover, and land vertically.
Rural Air Mobility	Envisions a safe, sustainable, affordable, and accessible air transportation system for passenger mobility, goods delivery, and emergency services within or traversing rural and exurban areas. Rural air mobility could include travel between a rural and urban area.
Urban Air Mobility (UAM)	Envisions a safe, sustainable, affordable, and accessible air transportation system for passenger mobility, goods delivery, and emergency services within or traversing metropolitan areas.
Unmanned Aircraft Systems Traffic Management (UTM)	A traffic management system that provides airspace integration requirements, enabling safe low-altitude operations. UTM provides services such as: airspace design, corridors, dynamic geofencing, weather avoidance, and route planning.

While UAM may be enabled by the convergence of several factors, several challenges such as: community acceptance, safety, social equity, issues around planning and implementation, airspace, and operations, could create barriers to mainstreaming. While numerous societal concerns have been raised about these approaches (e.g., privacy, safety, security, social equity), on-demand aviation has the potential to offer additional options for emergency services, goods delivery, and passenger mobility in urban and rural areas using small piloted, remote piloted, and autonomous aircraft (Shaheen et al., 2020).

This chapter is organized into four sections. The first section discusses potential barriers and challenges to implementation. Next, potential use cases of UAM are discussed. In the third section, the role of research, regulation, and demonstrations in developing policies and programs to guide the use of on-demand aviation are discussed. The chapter concludes with a discussion of how COVID-19 is presenting opportunities to demonstrate medical, emergency response, and sanitation use cases to build community acceptance of this novel aviation concept.

Potential Barriers and Challenges

Renewed interest in on-demand aviation coupled with innovative and emerging technologies has increased awareness about potential UAM societal concerns. In spite of an array of potential use cases and applications for on-demand aviation, UAM could face notable safety, financial, and community acceptance challenges, among others. Potential challenges to UAM deployment and adoption are summarized in Table 2.

Potential UAM Use Cases

UAM may serve a variety of use cases, such as emergency response, goods delivery, and passenger mobility and has the potential to:

- Support emergency management missions including: air ambulance, emergency supply delivery, organ transport, and search and rescue operations;
- Expand access to goods delivery, particularly in remote locations; and
- Enable additional mobility and delivery options.

A few potential UAM market segments and service models are described in Table 3.

There is a lack of research and understanding about the: (1) environmental, (2) travel behavior, (3) lifecycle, and (4) surface transportation network effects of implementing UAM. Research, pilot projects, and public policies could help to address many UAM concerns.

The Role of Research, Pilot Demonstrations, and Regulation

Due to the exploratory nature of UAM, more research is needed to better understand the potential impacts of UAM. Numerous efforts are underway across the globe to explore potential UAM opportunities and challenges to guide the use of on-demand aviation.

A few research gaps include:

- How will travelers' access vertiports, both from their origin and to their destination?
- How will UAM impact modal shift? Will UAM remove enough vehicles from the surface transportation network to make a noticeable impact on congestion? Could UAM induce travel demand due to reduced travel times or encourage more people to drive (due to reduced congestion from travelers switching to UAM)?
- How many gas-powered vehicles could UAM remove from the road?
- What are the lifecycle emission impacts associated with UAM compared to other transportation modes?
- What are the broad land use and societal impacts of UAM on communities?

Table 2 Potential barriers and challenges to UAM deployment and adoption

Potential barriers and challenges	Description	Use case		
		Emergency Management and Response	Goods	Delivery
Passenger Mobility				
Affordability and Social Equity	With respect to passenger mobility, current UAM passenger services are premium offerings that have, in recent years, typically averaged \$149–300 US per seat using traditional helicopters. At present, there are concerns that UAM may not be an affordable transportation option by lower- and middle-income households, and UAM may be used by upper income households to avoid congestion. Electrification and automation of UAM could reduce costs; however, uncertainty exists about whether UAM can obtain mass-market affordability. Additionally, it is important for UAM to be accessible for people with disabilities and other users with special needs. Similar concerns about affordability could also be associated with some emergency response use cases, such as medical transport (e.g., people without any or sufficient levels of medical insurance coverage may not be able to afford aeromedical use cases and/or be left with unaffordable medical transportation bills after using the service).	X		X
Aesthetics and Visual Pollution	An overcrowding of low-altitude aircraft in urbanized areas could create unwanted visual disturbances, particularly for non-essential use cases, such as mobility that could be accomplished with surface modes and drone delivery for retail packages.		X	X
Noise Pollution	Noise pollution is a potential problem that could arise with multiple low-altitude aircraft in urban areas. In the future, the use of conventional helicopters for early UAM passenger mobility service could increase community concerns about noise and limit adoption, particularly for non-essential use cases, such as passenger mobility and retail goods delivery. UAM may need to meet a stricter noise standard due to multiple aircraft, rotorcraft, and drones operating a low altitude over populated urban areas.		X	X
Increased Aircraft Activity Over Residential Areas	Residential communities may be concerned with low-altitude aircraft flying over homes and yards due to safety, privacy, noise, aesthetic, and other concerns. For example, residential communities could be particularly concerned about during the evenings, late night, and weekends due to concerns about privacy and noise disturbances at times when most residents are sleeping.		X	X
Electrification and Range Anxiety	While electrification has the potential to offer reduced emissions, scaled UAM operations could have an impact on the electric grid. Additionally, there could be performance and safety concerns associated with battery weight and energy storage in different climates, as well as consumer concerns associated with range anxiety. For emergency response use cases employing larger aircraft (e.g., air ambulance), electrification and range limitations could be notable barriers to adoption. Emergency response use cases with electric aircraft create notable operational challenges that must be considered, such as increased downtime due to charging and reduced mission flexibility due to limited range (requiring an aircraft to return to its base for recharging). This can result in greater downtime for electric aircraft compared to hybrid and gas-powered alternatives, including existing air ambulance services using helicopters. Technological innovations, such as improved battery capacity, reduced charging times, and battery swapping could help overcome some of these limitations.	X	X	X
Remotely Piloted and Autonomous Operations	There are a variety of technical and operational challenges that must be overcome before UAM can be deployed in urban areas at scale. In particular, there could be community concerns associated with remotely piloted and autonomous aircraft (both from the perspective of users and non-users). Machine learning and other algorithms used for automation are non-deterministic, which means that even for the same input, the algorithm may exhibit different behaviors on different runs. This could present challenges for certifying that these systems are safe for autonomous flight. Finally, there are a number of technical and security concerns (i.e., unmanned traffic management and cybersecurity) that will need to be addressed to accommodate remotely piloted and autonomous operations.	X	X	X
Airspace and Unmanned Traffic Management (UTM)	Integrating innovative and emerging aviation technologies into an existing, antiquated, and analog system of air traffic management can pose notable challenges for UAM, particularly as the number of new users of low-altitude airspace is projected to increase. A core challenge is enabling an automated air traffic management system that can handle increasing flight activity into a single ecosystem that accommodates an array of airspace users and communication methods. Designing a system that leverages the latest innovations, while ensuring cybersecurity and backward compatibility with existing systems, could be key.	X	X	X
Safety and Certification		X	X	X

Business Model and Financial Sustainability	Concerns about safety, particularly with new aircraft types coupled with scaled operations over highly urbanized areas could present challenges for UAM. Additionally, most aviation authorities lack defined certification categories for UAM aircraft that in many cases may contain novel features and combinations of features that are not typically found in other aircraft (distributed electric propulsion/tilt-wing propulsion, VTOL, autonomy hardware and software, and others). For goods delivery use cases, in particular, potential interactions between untrained individuals and drones could present significant safety challenges.			
	Each category of use cases presents their own business model and financial challenges. For the emergency response use case, a NASA market study examining the viability of air ambulance vehicles found that electric VTOL aircraft, in the best case scenarios, could only yield nominal cost savings over helicopters due to the high fixed costs of flight crews and medical personnel, limited range, and increased operational complexities. Drone delivery presents related challenges. A ratio of one drone to one package is likely an inefficient method of delivering a network of packages to a community. While early launches using eVTOL aircraft are expected in the early to mid-2020s, profitable passenger services are not anticipated until the late-2020s and early-2030s, based on exploratory market studies.	X	X	X

Source: (Shaheen et al. 2018; Reiche et al. 2018; Serrao et al. 2018; McKinsey and Company 2018)

Table 3 Potential UAM market segments and service models

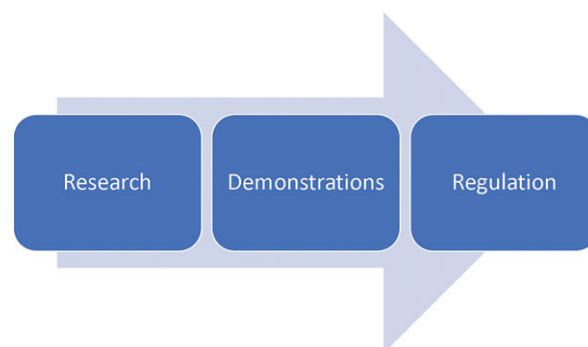
Market segment	Service models
Emergency Management and Response	<i>Medial Response, Evacuation, and Transport</i>—Providing medical care, evacuating patients from an incident scene, and transporting them to medical facilities <i>Disaster Response</i>—An array of missions that could include warning/evacuation, search and rescue, firefighting, assessing damage, providing immediate assistance, and restoring critical infrastructure <i>Disaster Relief</i>—An array of missions to help provide relief to return a community back to normal following an incident or disaster (e.g., distributing food and supplies)
Goods Delivery	<i>Small Unmanned Aircraft Systems (sUAS)</i>—A small aircraft (and enabling elements such as communication links) that weighs less than 55 pounds on take-off, including everything that is on board or otherwise attached (i.e., a package)
Passenger Mobility	<i>Personal Aircraft</i>—An aircraft that serves an individual or party for a length of time greater than the duration of a single flight <i>Air Pooling</i>—An on-demand service where multiple individual users share an aircraft <i>Air Taxi</i>—An on-demand service for a single user or group of users that reserve an entire aircraft for a flight and determine the flight's origin, destination, and timing <i>Semi-Scheduled Service</i>—A model with semi-flexible scheduling that can be modified to meet a customer's preferences (e.g., a flight scheduled to depart within a 2-h window) <i>Scheduled Service</i>—Offering near on-demand by offering frequent flights along a route with regularly scheduled service

Source: (Reiche et al. 2018; Patterson et al. 2018; McKinsey and Company 2018)

More research is needed to:

- Study the environmental impacts of UAM implementation and policies to support sustainability;
- Understand how to integrate UAM and small UAS (i.e., drones) into the same airspace and traffic management system;
- Research the safety and health impacts of UAM (including personal safety onboard autonomous aircraft, such as crime);
- Identify data needs, including data metrics, data formats, and standards for sharing;
- Model the potential traffic and land use impacts of UAM on the community;
- Identify flight path profiles for innovative aircraft and if traditional helipad approach paths should be adapted or changed;
- Understand public perception of aviation technologies, such as issues associated with electric range anxiety, willingness to fly on autonomous aircraft, etc.;
- Identify best practices for multimodal integration and vertiport design; and
- Study the social equity and economic impacts of UAM on communities (e.g., opportunities for increased employment, reduced ground vehicle traffic, accessibility for UAM by disadvantaged communities and users with special needs) (Fig. 1).

A number of public sector initiatives have begun to explore these unanswered questions and challenges. Key research, demonstration, and regulatory developments from the National Aeronautics and Space Administration (NASA), the US Federal Aviation Administration (FAA), the European Union Aviation Safety Agency (EASA), and the International Civil Aviation Organization (ICAO) are summarized further.

**Figure 1** Understanding and guiding the potential UAM impacts.

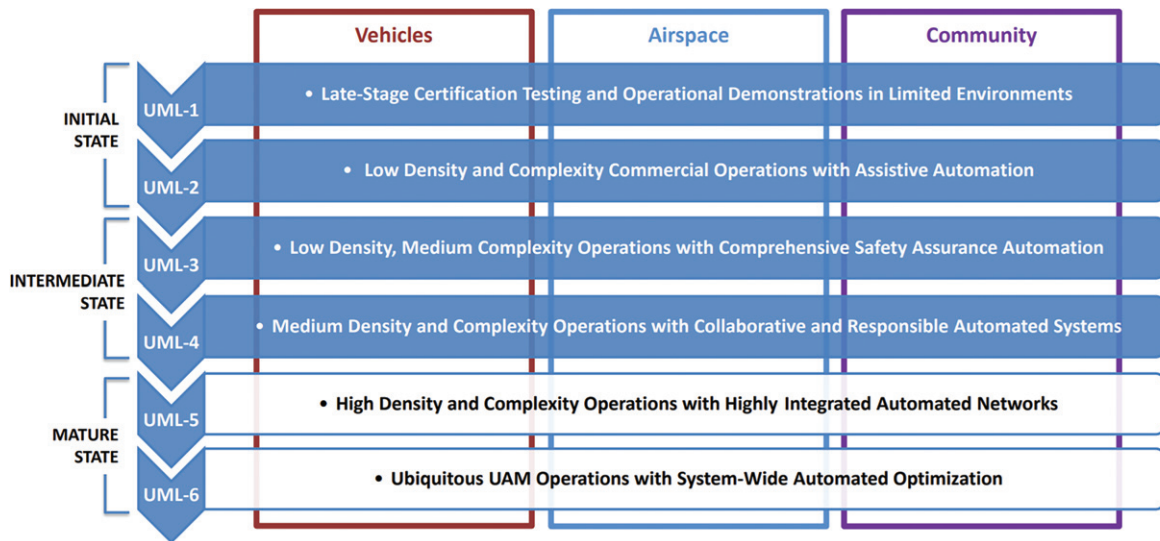


Figure 2 Urban air mobility maturity levels. Source: NASA.

NASA's Advanced Air Mobility National Campaign

In the United States, NASA's Aeronautics Research Mission Directorate (ARMD) launched the advanced air mobility working groups in Spring 2020 focused on enabling the UAM ecosystem through collaboration with public, private, and academic stakeholders. NASA has also developed a series of urban air mobility maturity levels intended to overcome vehicle, airspace, and community integration barriers from an initial to mature state (Fig. 2). NASA's *Advanced Air Mobility National Campaign* aims to improve UAM safety and accelerate scalability through integrated demonstrations through a series of ecosystem-wide challenges beginning in 2020. It is envisioned that the AAM National Campaign will support the United States. FAA in developing an approval process for UAM aircraft certification, develop: (1) flight procedure guidelines, (2) evaluate communication, (3) navigation and surveillance requirements, (4) define airspace operations management activities, and (5) characterize vehicle noise levels. The first testing opportunity in the National Campaign will focus on the testing of US developed aircraft and will include airspace operations management services to explore architectures and technologies needed to support future UAM safety and scalability of operations (Garrett-Glaser, 2019).

FAA's Unmanned Aircraft Systems (UAS) Integration Pilot Program (IPP)

In addition to the AAM National Campaign, the FAA's UAS Integration Pilot Program is helping the US Department of Transportation (DOT) develop new rules for unmanned aircraft (e.g., drones) that support more complex low-altitude operations by: (1) identifying ways to balance local and national interests related to drone integration; (2) improving communications with local, state, and tribal jurisdictions; (3) addressing security and privacy risks; and (4) accelerating the approval of operations that currently require special authorizations. Key goals of this initiative include establishing a dialog between local and national interests related to drone integration and providing actionable information to the US DOT on expanded and universal integration of drones into the National Airspace System. The IPP has funded nine lead participants that are evaluating a host of operational concepts including: package delivery, flights over people and beyond the pilot's line of sight, night operations, detect-and-avoid technologies, and the reliability and security of data links between pilot and aircraft (FAA, 2019). While the IPP has started this conversation, the industry still has not perfected the integration of unmanned systems for small unmanned aircraft (aircraft that weigh less than 55 pounds on take-off) nor large UAS into routine operations. Leveraging emerging lessons learned from the IPP, in Summer 2020 the FAA released the UAM Concept of Operations (ConOps) v1.0 which presents an air traffic management vision to support initial UAM operations in urban and suburban environments. The ConOps provides an initial framework for evolving from initial low complexity and low frequency operations to more mature operations with increasing frequency and complexity.

EASA Regulatory Developments

Similar to the FAA, the EASA certifies, regulates, standardizes, investigates, and monitors air transportation for the European Union. In May 2017, EASA published UAS regulations. In January 2019, the European Union, outside of EASA, passed UAS regulations that guide UAS operations and licensing. In Summer 2019, EASA published a special condition for certification of small electric and hybrid VTOL aircraft. The regulation establishes two categories of aircraft: basic and enhanced. Basic aircraft certification standards will be used for personal and rural aircraft. Enhanced aircraft certification standards apply to aircraft used in commercial operations and flight over urban areas. The regulation allows for aircraft with a pilot on board, remotely piloted, or various degrees of autonomy (EASA, 2019). In Summer 2020, EASA released a means of compliance (MOC) publication with draft guidance for manufacturers of

multirotor, lift-plus-thrust, and vectored thrust aircraft configurations. The document is intended to provide manufacturers guidance on how to comply with certification requirements.

ICAO Regulatory Developments

Internationally, ICAO regulates aviation safety, security, efficiency, regularity, and environmental protection for international air travel. In 2011, ICAO published a report that set standards for UAS airworthiness and licensing. However, laws regulating UAS vary by country. In an ICAO executive meeting in July 2019, UAM was presented by the United Arab Emirates as a high-level policy issue that required ICAO's input. ICAO determined that UAM poses a variety of challenges including lack of certification definitions and operational rules, lack of guidance to develop regulatory frameworks, gaps in existing information on unmanned systems, and a need for further research (ICAO, 2019). At present, international regulation on UAM is limited. Organizations including EASA and US-based ASTM International are working on developing airworthiness certifications for UAM and VTOL.

Air traffic management and public policy may need to evolve quickly to safely manage a growing number of new airspace users. Additionally, while many aviation authorities have emphasized national pre-emption of regulatory authority of the airspace, local and regional governments may argue for more local control over when and where urban air taxis fly. While many of these regulatory challenges have yet to be addressed, the policy environment will likely need to adapt quickly to emerging aviation technologies (e.g., electrification and automation) and new users of the airspace, such as UAM and small UAS package delivery.

Conclusion: Leveraging COVID-19 Medical and Sanitation Use Cases to Build Community Awareness

COVID-19 has presented a meaningful moment to demonstrate the potential of UAM and UAS, particularly in providing critical goods delivery and medical response. In May 2020, United Parcel Service began to expand its drone delivery program with CVS Health Corporation to delivery prescription medications to more than 135,000 residents in a Florida retirement community. In Ghana, California-based drone start-up, Zipline, began delivering critical medical supplies and COVID-19 test samples in Accra and Kumasi. In the United Arab Emirates and Indonesia, communities are using drones to spray disinfectant as part of a program to sanitize urban streets. In China, drones have been used to remind people to wear a facemask and to maintain social/physical distancing. These represent a handful of examples of how the COVID-19 pandemic has helped push UAS and UAM from the fringe to a potential strategy for emergency evacuation, response, and delivery of life saving equipment, tests, and medications. Although these examples are relatively limited, COVID-19 could be the moment that helps raise community awareness of this nascent aviation sector. Moving forward research on the potential impacts of UAM, coupled with prudent planning and implementation, are needed to balance commercial interests, technology innovation, and the public good.

Acknowledgment

The authors would like to thank the National Aeronautics and Space Administration and Toyota Motor Company for their generous support of this research.

Biographies

Adam Cohen is a mobility futures researcher at Innovative Mobility Research Group at the Transportation Sustainability Research Center (TSRC) at the University of California, Berkeley. Since joining the group in 2004, his research has focused on innovative mobility strategies, including urban air mobility, vehicle automation, smart cities, last mile delivery, shared mobility, smartphone apps, and other emerging technologies. Adam has published numerous peer-reviewed journal articles and reports on innovative and emerging transportation technologies and the future of mobility. Adam also served three combat tours in support of Operation Enduring Freedom as a rated aviator. Adam's unique multidisciplinary background gives him unique insight into automation, electrification, landside and airside aviation issues, and the potential impacts of innovative and disruptive technologies. Previously, Adam worked for the US Department of Homeland Security (DHS) and the Information Technology and Telecommunications Laboratory (ITTL) at the Georgia Tech Research Institute (GTRI). His academic background is in city and regional planning and international affairs.

Susan Shaheen is a professor in the Department of Civil and Environmental Engineering and a research engineer with the Institute of Transportation Studies at the University of California, Berkeley. She is also Co-Director of the Transportation Sustainability Research Center at UC Berkeley. She was among the first to observe, research, and write about the changing dynamics in shared mobility and the likely scenarios through which automated vehicles could gain prominence. She was the policy and behavioral research program leader at California Partners for Advanced Transit and Highways, a Special Assistant to the Director's Office of the California Department of Transportation, and the first Honda Distinguished Scholar in Transportation at the Institute of Transportation Studies at UC Davis, where she served as the endowed chair until 2012. She has authored 73 journal articles, over 125 reports and proceedings, 18 book chapters, and co-edited two books. She has a PhD from UC Davis and an MS from the University of Rochester.

See Also

Equity Concerns in Transport; Mobility as a Service (MaaS); The Future of Air Transport; Shared Mobility: An Overview of Definitions, Current Practices, and its Relationship to Mobility on Demand and Mobility as a Service; TNCs and the Future of Taxi Public Transport; Sharing: Attitudes and Perceptions; Future Mobility: Attitudes and Perceptions

References

- BLADE, 2019. BLADE Operating Standards and Flight Safety FAQs. Available from: <https://blade.flyblade.com>.
- Cohen, A., Guan, J., Beamer, M., Dittoe, R., Mokhtarmousavi, S., 2020. Reimagining the Future of Transportation with Personal Flight: Preparing and Planning for Urban Air Mobility. UC Berkeley: Transportation Sustainability Research Center, Berkeley. Available from: <https://escholarship.org/uc/item/9hs209r2>.
- EASA, 2019. Special Condition for small-category VTOL aircraft, SC-VTOL-01. Available from: <https://easa.europa.eu>.
- Etherington, D., 2020. Volocopter and Grab to study the feasibility of deploying air taxi services in Southeast Asia, TechCrunch. Available from: <https://techcrunch.com>.
- FAA, 2019. UAS Integration Pilot Program. Available from: <https://www.faa.gov>.
- Garrett-Glaser, B., 2019. NASA Launches Urban Air Mobility Grand Challenge Program, Aviation Today. Available from: <https://www.aviationtoday.com>.
- Hawkins, A.J., 2017. Uber's 'flying cars' could arrive in LA by 2020 — and here's what it'll be like to ride one', The Verge. Available from: <https://www.theverge.com>.
- ICAO, 2019. Urban air mobility, paper presented to the ICAO Assembly – 40th Session Executive Committee. Available from: https://www.icao.int/Meetings/a40/Documents/WP/wp_292_en.pdf.
- McKinsey and Company, 2018. Urban Air Mobility (UAM) Market Study. Washington D.C.: National Aeronautics and Space Administration.
- Patterson, M.D., Antcliff, K.R., Kohlman, L.W., 2018. A Proposed Approach to Studying Urban Air Mobility Missions Including an Initial Exploration of Mission Requirements, paper presented to the AHS International 74th Annual Forum & Technology Display, Phoenix, 14–17 May.
- Reiche, C., Goyal, R., Cohen, A., Serrao, J., Kimmel, S., Fernando, C., Shaheen, S., 2018. Executive Briefing: Urban Air Mobility Market Study, 5 Oct 2018. Washington D.C.: National Aeronautics and Space Administration. Available from: <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20190000519.pdf>
- Serrao, J., Nilsson, S., Kimmel, S., 2018. A legal and regulatory assessment for the potential of urban air mobility (UAM). In Urban Air Mobility Market Study. Washington D.C.: National Aeronautics and Space Administration. Available from: <https://escholarship.org/uc/item/49b8b9w0>.
- Shaheen, S., Cohen, A., Broader, J., Davis, R., Brown, L., Neelakantan, R., Gopalakrishna, D., 2020. Mobility on Demand Planning and Implementation: Current Practices, Innovations, and Emerging Mobility Futures. Washington D.C.: U.S. Department of Transportation Intelligent Transportation Systems (ITS) Joint Program Office.
- Shaheen, S., Cohen, A., Farrar, E., 2018. The Potential Societal Barriers of Urban Air Mobility (UAM). In Urban Air Mobility Market Study. Washington D.C.: National Aeronautics and Space Administration. Available from: <https://escholarship.org/uc/item/7p69d2bg>
- Voom, 2020. Helicopter Flights for Everyone. Available from: <https://www.voom.flights/en>.

Further Reading

- Duffy, M.J., Wakayama, S., Hupp, R., Lacy, R., Stauffer, M., 2017. A Study in Reducing the Cost of Vertical Flight with Electric Propulsion, paper presented to the AIAA Aviation Technology, Integration, and Operations Conference, Denver, 5–9 June.
- Johnson, W., Silva, C., 2018. Observations from Exploration of VTOL Urban Air Mobility Designs, Proceedings of the Seventh Asian/Australian Rotorcraft Forum, Jeju Island, Korea.
- Kasliwal, A., Furbush, N.J., Gawron, J.H., McBride, J.R., Wallington, T.J., De Kleine, R.D., Kim, H.C., Keoleian, G.A., 2019. Role of flying cars in sustainable mobility. *Nat. Commun.* 10 (1555), 1–9.
- Porsche Consulting, 2018. The Future of Vertical Mobility: Sizing the market for passenger, inspection, and goods services until 2035, Available from: Porsche Consulting.
- Reiche, C., Brody, F., McGillen, C., Siegel, J., Cohen, A., 2018. An Assessment of the Potential Weather Barriers of Urban Air Mobility. In Urban Air Mobility Market Study. Washington D.C.: National Aeronautics and Space Administration. Available from: <https://escholarship.org/uc/item/2pc8b4wt>.
- Silva, C., Johnson, W.R., Solis, E., Patterson, M.D., Antcliff, K.R., 2018. VTOL Urban Air Mobility Concept Vehicles for Technology Development. In 2018 Aviation Technology, Integration, and Operations Conference, p. 3847. DOI:10.2514/6.2018-3847.
- Yedavalli, P., Mooberry, J., 2019. An Assessment of Public Perception of Urban Air Mobility (UAM). Available from: Airbus.

Transport Modes and Sustainability

Long Cheng*, **Jonas De Vos***, **Frank Witlox***, *Department of Geography, Ghent University, Ghent, Belgium; †Bartlett School of Planning, University College London, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	710
Sustainability in Transport	710
Strategies of Transport Mode to Achieve Sustainability	711
Transport Technological Advancements	711
Clean and Energy Efficient Cars	711
Alternative Fuels and Vehicles	711
Intelligent Transport Systems	711
Eco-Driving and Eco-Flying	712
Modal Shift to Alternate Transport Modes	712
Urban Public Transport	712
Walking and Cycling	712
Train	712
Determinants of Transport Mode Choice	713
Urban Mobility	713
Intercity Mobility	713
Summary	713
References	713
Further Reading	714

Introduction

Transport modes enable mobility which drives social and economic developments and allows people to access services, jobs, education, and the interactions that help create fulfilled lives. However, the transport sector, as one of the top consumers of fossil fuels, is a major contributor to air pollution and generates a variety of emissions which impact the climate. This chapter presents how transport modes relate to sustainability, particularly sustainable paradigms of transport mode developments.

The rest of the chapter is structured as follows. Section 2 summarizes the concept of sustainability in transport. This is followed by Section 3 where strategies of transport mode to achieve sustainability are articulated. Section 4 presents the determinants of transport mode choice in intra- and intercity mobility. Section 5 concludes with a summary of findings.

Sustainability in Transport

The Brundtland Report of 1987 famously defined sustainable development as “development that meets the needs of the present without compromising the ability of future generations to meet their own needs” (Brundtland Commission, 1987, p. 8). The European Council (2001, pp. 15–16) defined a sustainable transport system as one that: (1) allows the basic access and development needs of individuals, companies and societies to be met safely and in a manner consistent with human and ecosystem health, and promotes equity within and between successive generations; (2) is affordable, operates fairly and efficiently, offers choice of transport mode, and supports a competitive economy, as well as balanced regional development; (3) limits emissions and waste within the planet’s ability to absorb them, uses renewable resources at or below their rates of generation, and uses nonrenewable resources at or below the rates of development of renewable substitutes while minimizing the impact on the use of land and the generation of noise.

The goals of transport sustainability have been summarized into aspects of environmental, social, and economic quality (Gudmundsson et al., 2016). Transport influences the environment both directly and indirectly. Direct influences include energy consumption, land-take for infrastructure, visual intrusion and noise from construction and use of infrastructure. Indirect influences include air and water pollution, and climate change. In terms of social impacts, some are direct (e.g., time waste, safety hazards) while others are displayed in complicated manners interacting with broader societal trends (e.g., transport inequity, social exclusion). Despite the numerous economic benefits brought by transport, it also generates economic costs. For example, traffic accidents and air pollution would cause the loss of life or reduced quality of life, which impose direct costs to the economy regarding lost productivity and lower welfare. Traffic congestion also brings significant costs on the economic developments, for example, through waste of time.

Different modes of transport play a different role in improving the sustainability of mobility patterns. Pomykala (2018) compared different modes of urban transport—bus, rail, car, and motorcycle, concluding that the most sustainable mode in terms of environmental performance (i.e., CO₂ emissions and energy consumption) is public transport. Therefore, there is a clear need to develop transport policies which encourage people to ride the public transport. In the long-distance travel market, aviation is increasingly conflicting the sustainable goals to limit climate change and challenges concerning noise, air pollution, and infrastructure (Gössling et al., 2019). Although aviation supporters have always emphasized the significance of air traffic for employment and economy growth, it is apparent that the increasing dependence on air travel contradicts principles of the United Nations' Sustainable Development Goals, including "Responsible Production and Consumption" and "Climate Action," and as well undermines the Paris Agreement (Scott et al., 2016). The EU for instance is taking actions to reduce aviation emissions in Europe. CO₂ emissions from aviation have been included in the EU emissions trading system (EU ETS) since 2012. Under the EU ETS, all airlines operating in Europe are required to monitor, report and verify their emissions, and to surrender allowances against those emissions. When considering sustainable transport mode for long-distance journeys, railway network is widely recognised (Givoni, 2006).

Strategies of Transport Mode to Achieve Sustainability

Sustainable mobility regarding transport mode falls into two paradigms. First, transport technology advances current patterns of modes by consuming less resources and generating less pollution. Second, sustainable travel behavior targets more sustainable travel mode choices. The latter forms a more holistic approach by recognizing that the sustainable movement of people and goods requires policy interventions to facilitate the use of sustainable transport modes (Sultana et al., 2019).

Transport Technological Advancements

Black (1997) pointed out that the most straightforward approach to enhance transport sustainability is to make the same trips using the same modes more sustainable. Technology-reliant paradigm recognizes that in some countries (the United States, for example) and cities lifestyles are so much revolving around the automobile that this dependency is extremely difficult to alter in the short and medium run (Sperling and Gordon, 2010). These transport technological advancements may include energy efficiency, pollution reduction, alternative fuels, innovated infrastructure, intelligent transport systems, and efficient driving and flying techniques (Kamga and Yazici, 2014). Some transport policies are also made to accelerate technological developments, such as the implementation of low emission zones or the ban of diesel cars by 2030–35 in various cities.

Clean and Energy Efficient Cars

Providing cleaner fossil fuel-based cars to the driving public is important to transport sustainability. In the United States, the Corporate Average Fuel Economy (CAFE) standards aim to increase fuel economy. The European Union also encourages its member countries to pursue taxation policies to promote the purchase of fuel-efficient vehicles. Many other countries similarly take the fuel taxes as leverage to increase the market share of efficient cars. These regulatory mechanisms—requiring pollution control equipment—have considerably reduced the harmful pollutant emissions from vehicles. Hybrid electric power cars, for instance, have the capability of fuel savings by using an electric motor in replacement of combustion engine during multiple stages of driving. Nonetheless, these fuel savings brought by hybrid vehicle operation need to trade-off the increased vehicle costs associated with the increased complexity of the vehicle (Aftabuzzaman and Mazloumi, 2011; Manzie et al., 2007).

Alternative Fuels and Vehicles

Under the Paris Agreement, worldwide nations have set the long-term goal of the reduction of greenhouse gas emissions from 2005 levels by 2050, for example, the United States and Canada 80%, Mexico 50%. In that context, motor vehicles require greener alternatives to fossil fuel. Relevant industries are developing fast on electricity, natural gas, biofuel, and hydrogen fuel in terms of vehicle and fuel production, repair, insurance, and the deployment of refuel stations. However, it is noted that these alternatives to fossil fuel are basically not equally sustainable (Edwards et al., 2004). In addition, the cost of some fuels is quite economically burdensome, either directly or indirectly by the required associative technologies. Wright (2006) investigated the hydrogen economy and pointed out that hydrogen has the potential to be an alternate transport fuel in Australia in the long run. At the same time, Rand and Dell (2005) argued that the development of hydrogen economy poses scientific and technological challenges in the process of production, delivery, storage, conversion and end-use (Aftabuzzaman and Mazloumi, 2011).

Intelligent Transport Systems

An intelligent transport system (ITS) is an advanced technology which aims to provide services relevant to different transport modes and traffic users and enable users to be better informed and make safer, more coordinated, and smarter use of transport networks. For example, vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) applications will be tools for improved connectivity, and safety for all modes of transport. They enable vehicles to act like nodes on a network and communicate with the surrounding environment. With the evolution of digital technologies—robotics, artificial intelligence, internet of things, high-performance computers and powerful communication networks—vehicles are eventually expected to guide themselves without human interventions. Connected and automated vehicles (CAVs) are among the most heavily researched automotive technologies, offering mobility

solutions which are cleaner, safer and more consumer-focused than ever. New mobility and smart transport business models are also emerging worldwide. Car sharing, bike sharing, and scooter sharing schemes are continuing to gain popularity; electric vehicle charging schemes is a growing market segment, while novel and smart parking solutions are being employed by shoppers and commuters. All these new models may provide opportunities for achieving sustainability in transport.

Eco-Driving and Eco-Flying

Promoting more efficient driving and flying techniques is helpful to improve transport sustainability. A set of driving behaviors and practices—known as “eco-driving”—includes techniques of shifting, braking, and maintaining a steady speed as well as habits such as ensuring proper tire pressure, regular vehicle maintenance, and limiting vehicle cargo (Kamga and Yazici, 2014). To assist these efforts, some newer cars have dashboard displays which show fuel use to help drivers adjust behaviors. Electric and hybrid vehicles are equipped with regenerative braking which gets energy back to the car. In terms of aviation, airlines are working actively to implement operational techniques which maximize fuel efficiency. Taking as “eco-flying,” these strategies include reducing cruising speeds, optimizing flight and descent paths, and employing on-the-ground procedures which reduce fuel waste, local pollution, delays, and traffic congestion.

Modal Shift to Alternate Transport Modes

Another important paradigm for improving sustainability is to promote the use of alternative transport modes, such as active travel (walking and cycling) and public transport (bus, metro, train, etc.). These modes of transport have comparably low or zero fuel consumption with private vehicles. Bouhot (2007) analyzed how efficient public transport consumed fossil fuel in France and concluded that the energy efficiency of public transport is clearly better than other modes. A global study of cities on transport energy by Newman and Kenworthy (2001) evidenced the energy, environmental, social benefits, as well as economic values of sustainable transport, that is, walking, cycling, and public transport. Cities with sustainable transport systems have shown reduced costs on road construction and maintenance, fewer road accidents, better operating cost recovery and fuel-efficiency, and fewer pollutant emissions. Some transport policies have also been developed to result in such a shift, such the Sustainable Urban Mobility Plans (SUMP) in Europe or road pricing schemes in different cities (e.g., London and Singapore).

Urban Public Transport

With growing concerns for transport sustainability, effective and well-organized public transport is regarded as a primary approach which will have an important influence on future urban mobility. In this context, transport demand management policies are thought to be a significant role using a range of services and incentives to facilitate the use of public transport. In many regions worldwide, public transport ensures that non-drivers (e.g., cyclists and pedestrians) have the minimum level of access to daily necessities, in particular given the need for some form of motorized travel in many western countries (Aftabuzzaman and Mazloumi, 2011). The increase of public transport use also brings about societal benefits, resulting in an increase of physical activity and a reduction in traffic congestion and travel times. The impacts of stress produced by traffic congestion have been found to spill over to workplace, leading to decreased work productivity, reduced employment satisfaction, and even absenteeism (Ma and Ye, 2019).

Walking and Cycling

Walking and cycling are two transport modes which provide the opportunity to form a sustainable transport system in the urban areas. They do not generate pollutants, produce no noise, require limited space and result in safety benefits if a critical mass of walking and cycling is achieved. In addition, walking and cycling offer a good form of regular exercise. The association between active travel and health has been largely investigated and is generally recognized (Hinde and Dixon, 2005). Key elements have been variously defined to increase the use of active travel (Pucher and Buehler, 2008). They include land use change, provision of dedicated infrastructure, demand management of motorized transport, development of supportive public transport, and measures to change attitudes (Tight, 2016).

Train

For the intercity mobility, train is generally accepted to be “greener” than its competing modes. An efficient train network is expected to relieve air traffic and highway congestion, make positive intermodal shifts and resilient transport systems (direct effects), and enhance integrated regional economies, environmental friendliness, and urban industrial competitiveness (indirect effects). In recent years, high speed rail (HSR) development has become a global trend and has ushered in “the second railway age” (Hall and Banister, 1993). HSR—which is more energy efficient given it is operated on electricity, and more sustainable than fossil fuel dependent transport systems—has the potential to reduce the amount of short- and medium-distance flights. Due to its speed (250 km/h or higher), HSR can offer shorter or comparable travel times than the flight on some routes and may therefore substitute it (Givoni, 2006). Modern HSR came into operations in Japan in 1964, and in Europe in the 1970s, and are now being planned and built in many countries internationally. As of 2019, 18 countries have developed HSR to connect major cities, with the network amounting to over 52,000 km.

Determinants of Transport Mode Choice

Urban Mobility

Existing studies show that urban travel mode choice is influenced by a range of elements including socio-demographics, built environment and attitudes. Behavioral heterogeneity in travel mode choices is observed, varying by gender, age, household income, driving license availability, education level, private vehicle ownership, and household structure (Buehler, 2011; Scheiner and Holz-Rau, 2007). Li et al. (2012)—investigating travel mode choices in the UK—revealed that the share of car use decreases among older people after their 1970s. In terms of gender, Cheng et al. (2019) found that females depend more on public transport than males in Nanjing, China. Bhat and Srinivasan (2005) moreover noted that travelers with higher income are more apt to travel by car. Using the 2001 American Housing Survey, Plaut (2005) identified that highly educated people make more trips (recreation-related trips in particular) by public transport. Car availability is a decisive predictor of car trips. Finally, travelers from larger households have a lower inclination to use non-motorized modes than those from smaller households (Ryley, 2006).

Furthermore, multiple key attributes of the built environment have been identified to play a significant role in travel mode choice, such as land-use mixture, building density, dedicated infrastructure for cyclists and pedestrians, proximity to public facilities, and transport provisions (Cervero, 2002). Generally, high density and mixed land use encourage people to walk or cycle or use public transport facilities. Dedicated neighborhood design toward walking and cycling facilitates the use of active modes. In addition, the enhanced accessibility tends to positively impact walking. In order to improve accessibility, it is essential to decrease the travel distance to public facilities as well as increase transport network connectivity.

Since the 1990s, the influences of attitudes on travel mode choice have gained considerable attention. Studies show that attitudes may be more powerful predictors of travel mode choice than the conventionally used objective measures (e.g., travel time, distance, or generalized costs). Analyzing the commuting data from Sweden, Johansson et al. (2016) found that attitudes towards comfort, flexibility, and environmental preferences significantly impact modal choice. Heinen et al. (2011) analyzed the impact of commuters' attitudes toward the benefits of cycling (e.g., low cost, convenience, and health benefits) on commuting mode choice. Their findings suggested that attitudinal factors provide additional explanations for commuters' cycling choices.

Intercity Mobility

As with studies on urban mobility, socio-demographics and spatial attributes have been found to influence mode choice for intercity travel. Income is one of the most important socio-economic factors. Private car is in general the mostly used transport mode for all income groups. However, low-income people are more likely to take inter-city train and coach services (Arbués et al., 2016), while high-income people use faster modes more frequently, particularly flights (Llorca et al., 2018). The income effect could also partly relate to the differences in employment status. Students and employees are thought to make more trips and also use a variety of modes, with students travelling more by train and coach and less by air and car than employees. Car availability is also an important factor, negatively impacting the tendency to use train for long-distance commuting, business and recreational trips (Reichert and Holz-Rau, 2015). Other research also documented how spatial attributes impact mode choice for long-distance travel. Access to inter-city transport services is important, where Reichert and Holz-Rau (2015) observed that high accessibility to railway station contributes to the likelihood of performing long-distance trips, in particular by train and flights. Urban population is also found to travel more by train and flights, compared to their rural counterparts. It is noted that both the spatial characteristics at the origin and destination account for the difference in mode choice (Limtanakool et al., 2006). van Acker et al. (2020) indicate that coach travelers are mostly female and students, live in low-income households and in urban environments, and travel mostly for private purposes (not commuting or business).

Summary

This chapter has introduced that transport sustainability encompasses environmental, social, and economic aspects. Different modes of transport have different impacts on sustainability and public transport and active modes are generally thought to be well performed. Two main paradigms are identified to achieve transport mode sustainability: transport technological advancements and modal shift to alternate transport modes. The latter forms a more holistic approach which requires policy interventions to alter individuals' travel mode choice. Furthermore, socio-demographics, built environment (or spatial attributes) and attitudes are discussed to be important factors affecting mode choice in intra- and intercity mobility.

References

- Aftabuzzaman, M., Mazloumi, E., 2011. Achieving sustainable urban transport mobility in post peak oil era. *Trans. Policy* 18 (5), 695–702.
- Arbués, P., Baños, J.F., Mayor, M., Suárez, P., 2016. Determinants of ground transport modal choice in long-distance trips in Spain. *Trans. Res. A Policy Pract.* 84, 131–143.
- Bhat, C., Srinivasan, S., 2005. A multidimensional mixed ordered-response model for analyzing weekend activity participation. *Transport. Res. B: Meth.* 39 (3), 255–278.
- Black, W.R., 1997. North American transportation: perspectives on research needs and sustainable transportation. *J. Transport Geogr.* 5 (1), 12–19.

- Bouhot, C., 2007. Greenhouse gas emissions and urban public transport. In: Proceedings of the 14th International Union of Air Pollution Prevention and Environmental Protection Associations (IUAPPA), Clean Air Society of Australian and New Zealand, Brisbane, Australia.
- Brundtland Commission, 1987. Our common future: report of the world commission on environment and development. Oxford University Press, Oxford, United Kingdom.
- Buehler, R., 2011. Determinants of transport mode choice: a comparison of Germany and the USA. *J. Transport Geogr.* 19 (4), 644–657.
- Cervero, R., 2002. Built environments and mode choice: toward a normative framework. *Transport. Res. D Transport Environ.* 7 (4), 265–284.
- Cheng, L., Chen, X., Yang, S., Wu, J., Yang, M., 2019. Structural equation models to analyze activity participation, trip generation, and mode choice of low-income commuters. *Transport. Lett.* 11 (6), 341–349.
- Edwards, R., Mahieu, V., Griesemann, J.C., Larivé, J.F., Rickeard, D.J., 2004. Well-to-wheels analysis of future automotive fuels and powertrains in the European context. *SAE Trans.* 113, 1072–1084.
- European Council, 2001. 2340th Council meeting—Transport/telecommunications—Luxembourg. Council of the European Union, Belgium.
- Givoni, M., 2006. Development and impact of the modern high-speed train: a review. *Transport Rev.* 26 (5), 593–611.
- Gössling, S., Hanna, P., Higham, J., Cohen, S., Hopkins, D., 2019. Can we fly less? Evaluating the ‘necessity’ of air travel. *J. Air Transport Manage.* 81, 101722.
- Gudmundsson, H., Marsden, G., Josias, Z., 2016. Sustainable transportation: indicators, frameworks, and performance management. Springer, Heidelberg, Germany.
- Hall, P., Banister, D., 1993. The second railway age. *Built Environ.* 19 (3), 157–162.
- Heinen, E., Maat, K., van Wee, B., 2011. The role of attitudes toward characteristics of bicycle commuting on the choice to cycle to work over various distances. *Transport. Res. D Transport Environ.* 16 (2), 102–109.
- Hinde, S., Dixon, J., 2005. Changing the obesogenic environment: insights from a cultural economy of car reliance. *Transport. Res. D Transport Environ.* 10 (1), 31–53.
- Johansson, M.V., Heldt, T., Johansson, P., 2006. The effects of attitudes and personality traits on mode choice. *Transport. Res. A: Policy Pract.* 40 (6), 507–525.
- Kanga, C., Yazici, M.A., 2014. Achieving environmental sustainability beyond technological improvements: potential role of high-speed rail in the United States of America. *Transport. Res. D: Transport Environ.* 31, 148–164.
- Li, H., Raeside, R., Chen, T., McQuaid, R.W., 2012. Population ageing, gender and the transportation system. *Res. Transport. Econ.* 34 (1), 39–47.
- Limtanakool, N., Dijst, M., Schwanen, T., 2006. The influence of socioeconomic characteristics, land use and travel time considerations on mode choice for medium-and longer-distance trips. *J. Transport Geogr.* 14 (5), 327–341.
- Llorca, C., Ji, J., Molloy, J., Moeckel, R., 2018. The usage of location based big data and trip planning services for the estimation of a long-distance travel demand model. Predicting the impacts of a new high speed rail corridor. *Res. Transport. Econ.* 72, 27–36.
- Ma, L., Ye, R., 2019. Does daily commuting behavior matter to employee productivity? *J. Transport Geogr.* 76, 130–141.
- Manzie, C., Watson, H., Halgamuge, S., 2007. Fuel economy improvements for urban driving: hybrid vs. intelligent vehicles. *Transport. Res. C: Emerg. Technol.* 15 (1), 1–16.
- Newman, P., Kenworthy, J., 2001. Transportation energy in global cities: sustainable transportation comes in from the cold? *Nat. Resour. Forum* 25 (2), 91–107.
- Plaut, P.O., 2005. Non-motorized commuting in the US. *Transport. Res. D Transport Environ.* 10 (5), 347–356.
- Pomykala, A., 2018. Effectiveness of urban transport modes. In: MATEC web of conferences. 03003, p. 180.
- Pucher, J., Buehler, R., 2008. Making cycling irresistible: lessons from the Netherlands, Denmark and Germany. *Transport Rev.* 28 (4), 495–528.
- Rand, D.A.J., Dell, R.M., 2005. The hydrogen economy: a threat or an opportunity for lead-acid batteries? *J. Power Sour.* 144 (2), 568–578.
- Reichert, A., Holz-Rau, C., 2015. Mode use in long-distance travel. *J. Transport Land Use* 8 (2), 87–105.
- Ryley, T., 2006. Use of non-motorised modes and life stage in Edinburgh. *J. Transport Geogr.* 14 (5), 367–375.
- Scheiner, J., Holz-Rau, C., 2007. Travel mode choice: affected by objective or subjective determinants? *Transportation* 34 (4), 487–511.
- Scott, D., Hall, C.M., Gössling, S., 2016. A report on the Paris Climate Change Agreement and its implications for tourism: why we will always have Paris. *J. Sustain. Tourism* 24 (7), 933–948.
- Sperling, D., Gordon, D., 2010. Two billion cars: driving toward sustainability. Oxford University Press, Oxford, United Kingdom.
- Sultana, S., Salon, D., Kuby, M., 2019. Transportation sustainability in the urban context: a comprehensive review. *Urban Geogr.* 40 (3), 279–308.
- Tight, M., 2016. Sustainable urban transport—the role of walking and cycling. *Proc. Inst. Civil Eng. Eng. Sustain.* 169 (3), 87–91.
- van Acker, V., Kessels, R., Cuervo, D.P., Lannoo, S., Witlox, F., 2020. Preferences for long-distance coach transport: Evidence from a discrete choice experiment. *Transport. Res. A Policy Pract.* 132, 759–779.
- Wright, J., 2006. Towards an Australian hydrogen economy. *Int. J. Environ. Stud.* 63 (6), 837–843.

Further Reading

- Aparicio, A., 2016. Exploring the sustainability challenges of long-distance passenger trends in Europe. *Transport. Res. Proc.* 13, 90–99.
- Banister, D., 2018. Policy on sustainable transport in England: the case of High Speed 2. *Eur. J. Transport Infrastruct. Res.* 18 (3), 262–275.
- Buehler, R., 2011. Determinants of transport mode choice: a comparison of Germany and the USA. *J. Transport Geogr.* 19 (4), 644–657.
- Cheng, L., Chen, X., Yang, S., Cao, Z., De Vos, J., Witlox, F., 2019. Active travel for active ageing in China: the role of built environment. *J. Transport Geogr.* 76, 142–152.
- Cherry, C., Cervero, R., 2007. Use characteristics and mode choice behavior of electric bike users in China. *Transport Policy* 14 (3), 247–257.
- De Vos, J., Mokhtarian, P.L., Schwanen, T., Van Acker, V., Witlox, F., 2016. Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility. *Transportation* 43 (5), 771–796.
- Edge, S., Dean, J., Cuomo, M., Keshav, S., 2018. Exploring e-bikes as a mode of sustainable transport: a temporal qualitative study of the perspectives of a sample of novice riders in a Canadian city. *Can. Geogr.* 62 (3), 384–397.
- Gössling, S., 2018. ICT and transport behavior: a conceptual review. *Int. J. Sustain. Transport.* 12 (3), 153–164.
- Herrmann-Lunecke, M.G., Mora, R., Sagaris, L., 2020. Persistence of walking in Chile: lessons for urban sustainability. *Transport Rev.* 40 (2), 135–159.
- López, C., Ruiz-Benítez, R., Vargas-Machuca, C., 2019. On the environmental and social sustainability of technological innovations in urban bus transport: the EU case. *Sustainability* 11 (5), 1413.
- Rahul, T.M., Verma, A., 2018. Sustainability analysis of pedestrian and cycling infrastructure—a case study for Bangalore. *Case Stud. Transport Policy* 6 (4), 483–493.
- Reis, V., Meier, J.F., Pace, G., Palacin, R., 2013. Rail and multi-modal transport. *Res. Transport. Econ.* 41 (1), 17–30.
- Suatmadi, A.Y., Creutzig, F., Otto, I.M., 2019. On-demand motorcycle taxis improve mobility, not sustainability. *Case Stud. Transport Policy* 7 (2), 218–229.
- van Acker, V., Witlox, F., 2011. Commuting trips within tours: how is commuting related to land use? *Transportation* 38 (3), 465–486.

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

Page left intentionally blank

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

EDITOR-IN-CHIEF

Roger Vickerman

*School of Economics, University of Kent, Canterbury, United Kingdom
and*

Transport Strategy Centre, Imperial College, London, United Kingdom

VOLUME 6

Transport Policy and Planning

SECTION EDITORS

Prof. Maria Attard

Department of Geography, Faculty of Arts, University of Malta, Msida, Malta



Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2021 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-08-102671-7

For information on all publications visit our
website at <http://store.elsevier.com>



Publisher: Oliver Walter
Acquisitions Editor: Oliver Walter
Content Project Manager: Natalie Lovell
Associate Content Project Manager: Manisha K and Ramalakshmi Boobalan
Designer: Matthew Limbert

EDITORIAL BOARD

Editor in Chief

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Section Editors

Maria Börjesson

Professor of Economics, VTI Swedish National Road and Transport Research Institute; Affiliated Professor at Linköping University, Sweden

Per Gårder

Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States

Prof. Kevin P.B. Cullinane

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

Prof. Edward C.S. Chung

Department of Electrical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

Chandra R. Bhat

Center for Transportation Research (CTR), The University of Texas at Austin, TX, United States

Edoardo Marcucci

Department of Political Sciences, University of Roma Tre, Rome, Italy; Department of Logistics, Molde University College, Molde, Norway

Prof. Maria Attard

Department of Geography, Faculty of Arts, University of Malta, Msida, Malta

Prof. Carlo G. Prato

School of Civil Engineering, The University of Queensland, Brisbane, Australia

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Page left intentionally blank

INTRODUCTION



Roger Vickerman

In an increasingly globalized world, despite reductions in costs and time, transportation has become even more important as a facilitator of economic and human interaction; this is reflected in technical advances in transportation systems, increasing interest in how transportation interacts with society and the need to provide novel approaches to understand its impacts. This has become particularly acute with the impact that Covid-19 has had on transportation across the world, at local, national, and international levels. This Encyclopedia brings a cross-cutting and integrated approach to all aspects of transportation from work in many disciplinary fields, engineering, operations research, economics, geography, and sociology to understand the changes taking place.

Transportation is both influenced by, and an influencer of, changes in the economy and society. Increasing speeds have reduced journey times and made the world a smaller place as globalization has affected both where people live and work and from where they source their goods and materials. Increasing volumes of traffic, often on old and life-expired infrastructure, lead to congestion and delays. Constraints on public budgets have led to increasing pressure on the private sector to fund improvements requiring new and innovative financial solutions. While there are clear differences in the nature of the pressures felt in the developed and less developed economies, there is an increasing recognition that in all societies, there is an accessibility problem such that certain groups become disadvantaged by the lack of access to reliable and cost-effective transport. While the problems are clearly multidimensional, research on transportation is often constrained by single disciplinary approaches and this carries over into the practice of transport planning and policy. The Encyclopedia cuts across these artificial boundaries by taking an approach that emphasizes the interaction between the different aspects of research and aims to offer new solutions to understand these problems. Each of the nine sections is based around a familiar dimension of work on transportation, but brings together the views of experts from different disciplinary perspectives. Each section is edited by an expert in the relevant field who has sought chapters from a range of authors representing different disciplines, different parts of the world, and different social perspectives. In this way, the work is not just a reflection of the state of the art that serves as a starting point for researchers and practitioners, but also a pointer toward new approaches, new ways of thinking, and novel solutions to problems.

The nine sections are structured around the following themes: Transport Modes; Freight Transport and Logistics; Transport Safety and Security; Transport Economics; Traffic Management; Transport Modeling and Data Management; Transport Policy and Planning; Transport Psychology; Sustainability and Health Issues in Transportation. Some of the chapters provide a technical introduction to a topic while others provide a bridge between topics or a more future-oriented view of new research areas or challenges. While there is guidance to cross-referencing between chapters, readers are encouraged to explore the tables of contents of all the sections to get a full understanding of the issues. Much of the Encyclopedia was completed before the Covid-19 pandemic and clearly this will have changed the situation in many areas covered by this work; the advantage of this type of reference work is that relevant updates will be possible in future editions.

The Encyclopedia has only been possible because of the cooperation of a large number of people. Robert Noland, Georgina Santos, Xiaowen Fu, and Dick Ettema served as an Editorial Advisory Board identifying possible editors of sections and advising on the overall structure. The Section Editors, Edoardo Marcucci, Kevin Cullinane, Per Garder, Maria Börjesson, Edward Chung, Chandra Bhat, Maria Attard, and Carlo Prato,

carried out the work of identifying potential authors of individual chapters, commissioning these, encouraging authors and reviewing drafts. More than 600 authors and co-authors of chapters are, however, ultimately responsible for the success of this venture. Thanks are also due to the key people at Elsevier, particularly the Publishers, Andre Wolff and Oliver Walter, and the Project Managers, Sophie Harrison and Natalie Bentahar; they have shown exemplary patience over more than 3 years in bringing this work to fruition.

LIST OF CONTRIBUTORS TO VOLUME 6

Maria Attard

Department of Geography, Faculty of Arts, University of Malta, Msida, Malta

Stephen Ison

Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

Lucy Budd

Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

Miloš N. Mladenović

Otakaari 4, Espoo, Finland

Cyrille Médard de Chardon

University of Hull, Hull, United Kingdom; Luxembourg Institute of Socio-Economic Research (LISER), Maison des Sciences Humaines, Esch-sur-Alzette/Belval, Luxembourg

Fiona Ferbrache

Keble College, Oxford, United Kingdom

Luca Zamparini

Department of Law, University of Salento, Lecce, Italy

Mengqiu Cao

School of Architecture and Cities, University of Westminster, London, United Kingdom; Bartlett School of Planning, University College London, London, United Kingdom

George Yannis

National Technical University of Athens, Department of Transportation Planning & Engineering, Athens, Greece

Eleonora Papadimitriou

Delft University of Technology, Delft, The Netherlands

Charles B A Musselwhite

Centre for Innovative Ageing, Swansea University, Swansea, United Kingdom

Graham Parkhurst

University of the West of England, Bristol, United Kingdom

Lisa Hansson

Faculty of Logistics, Molde University College - Specialized University in Logistics, Norway

Paul Martin Timms

Institute for Transport Studies University of Leeds, United Kingdom

Benjamin Büttner

Technical University of Munich, München, Germany

Marcus Enoch

School of Architecture, Building and Civil Engineering, Loughborough University, Leicestershire, United Kingdom

Robin Hickman

Bartlett School of Planning, University College, London

Christine Hannigan

Bartlett School of Planning, University College, London

David A. King

Arizona State University, United States

Luis A. Guzman

Universidad de los Andes, Department of Civil and Environmental Engineering, Bogotá, Colombia

Daniel Oviedo

Development Planning Unit, University College London, London, United Kingdom

Mariajose Nieto Combariza

Development Planning Unit, University College London, London, United Kingdom

E. Owen D. Waygood

Transport Engineering, Department of Civil, Geotechnical, and Mining Engineering Polytechnique Montréal, Montreal, QC, Canada

Raktim Mitra

School of Urban and Regional Planning, Ryerson University, Toronto, ON, Canada

Geoff Vigar

School of Architecture Planning and Landscape, Newcastle University, Newcastle, United Kingdom

Susan M. Grant-Muller
University of Leeds, Leeds, West Yorkshire, United Kingdom

Mahmoud Abdelrazek
Department of Civil, Environmental and Geomatic Engineering, University College London, London, United Kingdom

Hannah Budnitz
University of Birmingham, Edgbaston, Birmingham, United Kingdom

Caitlin D. Cottrill
Transport Studies Unit, University of Oxford, Oxford, United Kingdom

Fiona Crawford
University of the West of England, Bristol, United Kingdom

Charisma F. Choudhury
University of Leeds, Leeds, West Yorkshire, United Kingdom

Teddy Cunningham
University of Warwick, Warwick, United Kingdom

Gillian Harrison
University of Leeds, Leeds, West Yorkshire, United Kingdom

Frances C. Hodgson
University of Leeds, Leeds, West Yorkshire, United Kingdom

Jinhyun Hong
Department of Urban Studies, University of Glasgow, Glasgow, Scotland, United Kingdom

Adam Martin
University of Leeds, Leeds, West Yorkshire, United Kingdom

Oliver O'Brien
Department of Geography, University College London, London, United Kingdom

Claire Papaix
Systems Management Strategy Department, University of Greenwich, London, United Kingdom

Panagiotis Tsoleridis
University of Leeds, Leeds, West Yorkshire, United Kingdom

Cyriac George
Department of Mobility and Organisation, Institute of Transport Economics (TOI), Oslo, Norway

Tanu Priya Uteng
Department of Mobility and Organisation, Institute of Transport Economics (TOI), Oslo, Norway

John Ward
University College London, United Kingdom

Karel Martens
Technion - Israel Institute of Technology, Israel Institute of Technology, Haifa, Israel

Simon P. Blainey
Transportation Research Group, Faculty of Engineering and Physical Sciences, University of Southampton, Boldrewood Campus, Southampton, United Kingdom

Dr. Alexandros Nikitas
Huddersfield Business School, University of Huddersfield, Huddersfield, United Kingdom

Sheila Mitra-Sarkar
Institute of Public Urban Affairs, San Diego State University and Principal, FutureTrans, San Diego, CA, United States

Michela Le Pira
Department of Civil Engineering and Architecture, University of Catania, Catania, Italy

Giuseppe Inturri
Department of Electric, Electronic and Computer Engineering, University of Catania, Catania, Italy

Matteo Ignaccolo
Department of Civil Engineering and Architecture, University of Catania, Catania, Italy

Silvio Nocera
Università IUAV di Venezia - Department of Architecture and Arts, Venice, Italy

Gonçalo Homem de Almeida Rodriguez Correia
TU Delft, Faculty of Civil Engineering and Geosciences (CiTG), Transport & Planning Department Delft, The Netherlands

Ida Kristoffersson
VTI Swedish National Road and Transport Research Institute, Stockholm, Sweden

Maria Börjesson
VTI Swedish National Road and Transport Research Institute, Stockholm, Sweden

Debbie Hopkins
Transport Studies Unit, School of Geography and the Environment, University of Oxford, Oxford, United Kingdom

Christian Brand
Transport Studies Unit and Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, United Kingdom

Laura Eboli
University of Calabria, Department of Civil Engineering, Rende, Calabria, Italy

Gabriella Mazzulla
*University of Calabria, Department of Civil Engineering,
 Rende, Calabria, Italy*

Mario Intini
*Department of Economics, Management and Business
 Law, University of Bari Aldo Moro, Bari, Italy*

Marco Percoco
*Department of Social and Political Sciences and GREEN,
 Università Bocconi, Milan, Italy*

Niek Mouter
Delft University of Technology, The Netherlands

Dr Fabio Galatioto
*Connected Places Catapult, Milton Keynes, United
 Kingdom*

Esther Anaya-Boig
*Centre for Environmental Policy, Imperial College
 London, London, United Kingdom*

Paulo Ancaes
*Centre for Transport Studies, Civil Environmental and
 Geomatic Engineering, UCL (University College
 London), United Kingdom*

Jennifer S. Mindell
*Health and Social Surveys Research Group,
 Research Department of Epidemiology and Public
 Health, UCL (University College London),
 United Kingdom*

Claus H. Sørensen
K2/VTI, Lund, Sweden

Fredrik Pettersson
Lund University, Lund, Sweden

Joel Hansson
Lund University, Lund, Sweden

Marie Delaplace
Gustave Eiffel University, Champs sur Marne, France

Miriam Ricci
*Department of Geography and Environmental
 Management, University of the West of England, Bristol,
 United Kingdom*

Giulio Mattioli
*Technische Universität Dortmund, Faculty of Spatial
 Planning, Department of Transport Planning, Dortmund,
 Germany*

Muhammad Adeel
*University of Leeds, Institute for Transport Studies, Leeds,
 United Kingdom*

Laetitia Dablanc
University Gustave Eiffel, Marne-la-Vallée, France

Chia-Lin Chen
*Department of Geography and Planning, University of
 Liverpool, United Kingdom*

Romeo Danielis
*Dipartimento di Scienze Economiche, Aziendali,
 Matematiche e Statistiche (DEAMS), Università degli
 Studi di Trieste, Italy*

Lucia Rotaris
*Dipartimento di Scienze Economiche, Aziendali,
 Matematiche e Statistiche (DEAMS), Università degli
 Studi di Trieste, Italy*

Bruno Dalla Chiara
*POLITECNICO DI TORINO (I), Engineering, Dept.
 DIATI - Transport Systems, Italy*

Kate Pangbourne
*Institute for Transport Studies, University of Leeds, Leeds,
 United Kingdom*

Ian Shergold
*University of the West of England, Bristol, United
 Kingdom*

Sandra Melo
*CEiiA - Center of Engineering and Development, Porto,
 Portugal*

David Metz
*Centre for Transport Studies, University College London,
 London, United Kingdom*

Patrizia Franco
Connected Places Catapult, Milton Keynes, United Kingdom

Carlos Cañas Sanz
University of Malta, Msida, Malta

Jonathan Cowie
*Edinburgh Napier University, School of Engineering and
 the Built Environment, Edinburgh, United Kingdom*

Thomas Vanoutrive
University of Antwerp, Antwerp, Belgium

Laura Walker
*Social and Equalities Impact Assessment Consultant
 (AECOM), Birmingham, United Kingdom*

Angela Curl
University of Otago, Christchurch, New Zealand

Ersilia Verlinghieri
*University of Oxford, Oxford, United Kingdom;
 University of Westminster, London, United Kingdom*

Begoña Guirao
*Department of Transport Engineering, Regional and
 Urban Studies. Universidad Politécnica de Madrid,
 Madrid, Spain*

Christopher Clott
State University of New York Maritime College, United States

Chris Petrocelli
State University of New York Maritime College, United States

Corinne Mulley
The University of Sydney, Sydney, NSW, Australia

John D. Nelson
The University of Sydney, Sydney, NSW, Australia

Anthony Perl
*Urban Studies Program, Simon Fraser University,
Vancouver, BC, Canada*

Özgül Ardiç
Turkish State Railways (TCDD), Ankara, Turkey

J.A. Annema
TU Delft, Jaffalaan 5, Delft, The Netherlands

Stephen Potter
*School of Engineering and Innovation, Open University,
Milton Keynes, Buckinghamshire, England*

Rosário Macário
Av Rovisco Pais 1, Lisboa, Portugal

Per Erik Gårder
University of Maine, Orono, ME, United States

Mohammad Sadrani
*Department of Civil, Geo and Environmental
Engineering, Technical University of Munich, Munich,
Germany*

Constantinos Antoniou
*Department of Civil, Geo and Environmental
Engineering, Technical University of Munich, Munich,
Germany*

Fahimeh Khalaj
The University of Queensland, QLD, Australia

Dorina Pojani
The University of Queensland, Australia

Sara Alidoust
The University of Queensland, Australia

CONTENTS OF ALL VOLUMES

<i>Editorial Board</i>	<i>v</i>
<i>Introduction</i>	<i>vii</i>
<i>Contributors to Volume 6</i>	<i>ix</i>

VOLUME 1

Introduction to Transportation Economics	1
Market Failures and Public Decision Making in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	2
Demand for Freight Transport <i>José Holguín-Veras and Diana G. Ramírez-Ríos</i>	7
Cost Functions for Road Transport <i>José Manuel Vassallo</i>	13
Future of Urban Freight <i>Russell G. Thompson</i>	20
Operation Costs for Public Transport <i>Marco Batarce</i>	26
Natural Monopoly in Transport <i>André de Palma and Julien Monardo</i>	30
Freight Costs: Air and Sea <i>Yulai Wan and Dong Yang</i>	36
Transport Production and Cost Structure <i>Ricardo Giesen and Darío Farren</i>	46
The Concept of External Cost: Marginal versus Total Cost and Internalization <i>Sofia F. Franco</i>	56
Value of Time <i>John J. Bates</i>	67
Valuation of Carbon Emissions <i>Svante Mandell</i>	72
Valuation of Travel Time Variability Using Scheduling Models <i>Katrine Hjorth</i>	76

Value of Crowding <i>Daniel Hörcher</i>	84
What Drives Transport and Mobility Trends? The Chicken-and-Egg Problem <i>Nathalie Picard</i>	89
Pricing Principles in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	95
Long-Run Versus Short-Run Valuations <i>Stefanie Peer</i>	102
Value of Noise <i>Jon P. Nelson</i>	106
The Value of Life and Health <i>Henrik Andersson</i>	114
The Value of Security, Access Time, Waiting Time, and Transfers in Public Transport <i>Raquel Espino, Juan de Dios Ortúzar, and Luis I. Rizzi</i>	122
Demand for Passenger Transportation <i>Kenneth A Small and Robin Lindsey</i>	127
Real-World Experiences of Congestion Pricing <i>Charles Raux</i>	134
Distributional Effects of Congestion Charges and Fuel Taxes <i>Jonas Eliasson</i>	139
The Bottleneck Model <i>Dereje Abegaz and Yili Tang</i>	146
Dynamic Congestion Pricing and User Heterogeneity <i>Kathrin Goldmann and Gernot Sieg</i>	150
Economics of Parking <i>Daniel Albalade and Albert Gragera</i>	159
Loss Aversion and Size and Sign Effects in Value of Time Studies <i>Andrew Daly</i>	165
Intertemporal Variation of Valuations <i>James Fox</i>	170
The Rebound Effect for Car Transport <i>Bruno De Borger, Ismir Mulalic and Jan Rouwendal</i>	174
Elasticities for Travel Demand: Recent Evidence <i>Fay Dunkerley, Charlene Rohr and Mark Wardman</i>	179
Parking Price Elasticities <i>Stephan Lehner</i>	185
The Pareto Criterion and the Kaldor Hicks Criterion <i>Harald Minken</i>	190
Social Discount Rates <i>Lars Hultkrantz</i>	195
Cross-Elasticities between Modes <i>Mark Wardman and Jeremy Toner</i>	201
First-Best Congestion Pricing <i>Moez Kilani</i>	209

Ethical Aspects-Can We Value Life, Health, and Environment in Money Terms? <i>Jose M. Grisolia and Ken Willis</i>	216
Car tolls, Transit Subsidies for Commuting, and Distortions on the Labor Market <i>Ioannis Tikoudis¹ and Kurt Van Dender¹</i>	221
Demand Management and Capacity Planning of Airports <i>Stephen Ison and Lucy Budd</i>	227
Dealing With Negative Externalities: Low Emission Zones Versus Congestion Tolls <i>Valeria Bernardo, Xavier Fageda, and Ricardo Flores-Fillol</i>	231
The Rule-of-a-Half and Interpreting the Consumer Surplus as Accessibility <i>Mogens Fosgerau and Ninette Pilegaard</i>	237
Producer Surplus <i>Chau Man Fung</i>	242
The Robustness of Cost-Benefit Analyses <i>Morten Welde and James Odeck</i>	249
The GDP Effects of Transport Investments: The Macroeconomic Approach <i>James Laird and Daniel Johnson</i>	256
The Mohring Effect <i>Hugo E. Silva</i>	263
Public Transport Fare and Subsidy Optimization <i>Qianwen Guo and Zhongfei Li</i>	267
Transportation Equity <i>Rafael H. M. Pereira, and Alex Karner</i>	271
Impact of Transport Cost-Benefit Analysis on Public Decision-Making <i>Niek Mouter</i>	278
Causal Inference for Ex Post Evaluation of Transport Interventions <i>Daniel J. Graham</i>	283
Transport Cost and Location of Firms <i>Adelheid Holl</i>	291
Commuting, the Labor Market, and Wages <i>Jan Rouwendal, and Ismir Mulalic</i>	297
How to Buy Transport Infrastructure <i>Johan Nyström</i>	302
Procurement of Public Transport: Contractual Regimes <i>Andrew Smith, and Chris Nash</i>	308
The Mono-Centric City Model and Commuting Cost <i>Zhi-Chun Li, and Ya-Juan Chen</i>	315
Value of Time in Freight Transport <i>Gerard de Jong</i>	321
The Economics and Planning of Urban Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	326
Incentives in Public Transport Contracts <i>Andreas Vigren</i>	332
Transportation Improvements and Property Prices <i>Alex Anas</i>	337

Transport Infrastructure Effects on Economic Output: The Microeconomic Approach <i>Patricia C. Melo</i>	347
Wider Economic Impacts of Transport Investments <i>Anthony J. Venables</i>	355
Employment Effects of Transport Infrastructure <i>Anna Matas, and Javier Asensio</i>	360
Regulatory Reforms and Competition in Public Transport <i>Didier van de Velde</i>	365
How to Finance Transport Infrastructure? <i>Tiziana D'Alfonso, and Giuseppe Catalano</i>	371
Congestion, Allocation and Competition on the Railway Tracks <i>John Armstrong, and John Preston</i>	378
Public Private Partnership <i>David Meunier, and Emile Quinet</i>	385
Airline Economics <i>Anming Zhang, Yahua Zhang, and Zhibin Huang</i>	392
Privatization and Deregulation of the Airline Industry <i>Achim I. Czerny, and Hao Lang</i>	397
Price Discrimination and Yield Management in the Airline Industry <i>Guoquan Zhang, Colin C.H. Law, Yahua Zhang, and Hangjun Yang</i>	404
Regulation and Competition in Railways <i>Chris Nash, and Andrew Smith</i>	409
Cycling Economics <i>Bert van Wee</i>	414
The Economic Rationale for High-Speed Rail <i>Ginés de Rus</i>	419
Rail Cost Functions <i>Kristofer Odolinski, and Phill Wheat</i>	425
Transport and International Trade <i>Siri Pettersen Strandenes</i>	431
Maritime Economics: Organizational Structures <i>Kevin P.B. Cullinane</i>	436
Port Planning and Investment <i>Jasmine Siu Lee Lam</i>	443
Estimating the Capital Stock of Transport Infrastructure <i>Heike Link</i>	449
The Economics of Reducing Carbon Emissions From Air and Road Transport <i>Olga Ivanova</i>	457
Regulation and Financing of Toll Roads <i>Marco Ponti</i>	464
Are Megaprojects too Transformational for Cost-Benefit Analysis? <i>Tom Worsley</i>	470
Economics of Transportation Safety <i>Ian Savage</i>	476

Cost Overruns of Transportation Infrastructure Projects <i>James Odeck, and Morten Welde</i>	483
The Downs-Thomson Paradox <i>Joel P. Franklin</i>	490
Policy Instruments for Plug-In Electric Vehicles: An Overview and Discussion <i>Jake Whitehead, Patrick Plötz, Patrick Jochem, Frances Sprei, and Elisabeth Dütschke</i>	496
Vertical and Horizontal Separation in the European Railway Sector and Its Effects on Productivity <i>Pedro Cantos-Sánchez</i>	503
How will Autonomous Vehicles Impact Car Ownership and Travel Behavior <i>Patrick M. Bösch, Felix Becker, Henrik Becker, and Kay W. Axhausen</i>	508
Policy Instruments to Reduce Carbon Emissions from Road Transport Computable General Equilibrium Analysis in Transportation Economics <i>Johannes Bröcker</i>	520
Estimation of Value of Time <i>Stefan Flügel, and Askill H. Halse</i>	527
The Taxation of Car Use in the Future <i>Griet De Ceuster, and Inge Mayeres</i>	534
Cost-Benefit Analysis and Other Assessment Techniques: Contrasts and Synergies <i>Paolo Beria</i>	540
Demand for Air Travel and Income Elasticity <i>Jing Lu, Yucan Meng, Changmin Jiang, and Cheng Lv</i>	547
Generalized Cost for Transport <i>Jeppé Rich</i>	555
The Impact of Electric Vehicles on Energy Systems <i>Patrick E.P. Jochem, Jake Whitehead, and Elisabeth Dütschke</i>	560
Uber versus Taxis <i>Georgina Santos</i>	566
Contract Efficiency in Public Transport Services <i>Philippe Gagnepain, and Marc Ivaldi</i>	572
Company Cars <i>Stefan Gössling</i>	580
Transportation Network Companies (TNCs) and the Future of Public Transportation <i>Susan Shaheen, and Adam Cohen</i>	584
Public Transport in Low Density Areas <i>Jani-Pekka Jokinen, Leif Sörensen, and Jan Schlüter</i>	589
Market Failures in Transport: Direct and Indirect Public Intervention <i>Federico Boffa, and Alberto Iozzi</i>	596
The Braess Paradox <i>Anna Nagurney, and Ladimer S. Nagurney</i>	601
VOLUME 2	
Introduction to Transportation Safety and Security <i>Per Gårder</i>	1

The Concept of “Acceptable Risk” Applied to Road Safety Risk Level <i>Claes Tingvall</i>	2
Crash Not Accident <i>Robert A. Scopatz</i>	6
Age and Gender as Factors in Road Safety <i>Marion Sinclair</i>	11
Aggressive Driving and Road Rage <i>James E.W. Roseborough, Christine M. Wickens, and David L. Wiesenthal</i>	17
Aircraft Maintenance and Inspection <i>Alan Hobbs</i>	25
Airport Security <i>Richard W. Bloom</i>	34
Transport Safety and Security: Alcohol <i>James C. Fell</i>	40
Animal Crashes <i>Michal Bil</i>	53
Attenuators <i>Simonetta Boria</i>	63
ATV, Snowmobile, and Terrain Vehicle Safety <i>David P. Gilkey, and William Brazile</i>	77
Automobile Safety Inspection <i>Subasish Das</i>	85
Aviation Safety: Commercial Airlines <i>Clarence C. Rodrigues</i>	90
Aviation Safety, Freight, and Dangerous Goods Transport by Air <i>Glenn S. Baxter and Graham Wild</i>	98
Bicycle Collision Avoidance Systems: Can Cyclist Safety be Improved with Intelligent Transport Systems? <i>Lars Leden</i>	108
Bicycle Infrastructure <i>Rock E. Miller</i>	115
Bicycles, E-Bikes and Micromobility, A Traffic Safety Overview <i>Höskuldur Kröyer</i>	125
Bicycles: The Safety of Shared Systems Versus Traditional Ownership <i>Mercedes Castro-Nuño and José I. Castillo-Manzano</i>	139
Bridge Safety <i>Per Erik Garder</i>	144
Carsharing Safety and Insurance <i>Elliot Martin and Susan Shaheen</i>	150
Carjacking <i>Terance D. Mieth, Christopher Forepaugh, Tanya Dudinskaya</i>	157
Collision Avoidance Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	164
Collision Avoidance Systems, Automobiles <i>Erick J. Rodríguez-Seda</i>	173

Connected Automated Vehicles: Technologies, Developments, and Trends <i>Azra Habibovic and Lei Chen</i>	180
Construction Zones <i>Jalil Kianfar</i>	189
Costs of Accidents <i>Ulf Persson</i>	196
Critical Issues for Large Truck Safety <i>Matthew C. Camden, Jeffrey S. Hickman, Richard J. Hanowski, and Martin Walker</i>	200
Demerit Points and Similar Sanction Programs <i>Matúš ťucha and Kristýna Josrová</i>	210
Driver State and Mental Workload <i>Dick de Waard and Nicole van Nes</i>	216
Drugs, Illicit, and Prescription <i>Rune Elvik</i>	221
Education, Training, and Licensing <i>Matúš ťucha and Kristýna Josrová</i>	228
Elderly Driver Safety Issues <i>Mark J King</i>	233
Emergency Response Systems <i>Frances L. Edwards</i>	240
Emergency Vehicles and Traffic Safety <i>Shamsunnahar Yasmin, Sabreena Anowar, and Richard Tay</i>	247
Encouragement: Awards and Incentives <i>Fred Wegman</i>	255
Enforcement and Fines <i>Matúš ťucha and Ralf Risser</i>	263
Epidemiology of Road Traffic Crashes <i>Sherrie-Anne Kaye, Judy Fleiter, and Md Mazharul Haque</i>	269
Evacuation Planning and Transportation Resilience <i>Karl Kim</i>	276
Exposure: A Critical Factor in Risk Analysis <i>Frank Gross, PhD, PE</i>	282
Passenger Ferry Vessels and Cruise Ships: Safety and Security <i>Wayne K. Talley</i>	290
Fuel Economy Standards: Impacts on Safety <i>Kenneth T. Gillingham and Stephanie M. Weber</i>	296
Hazardous Materials Transport <i>Dr. Arjan Vincent van der Vlies</i>	304
Head-on Crashes <i>John N. Ivan</i>	311
Helicopters in Emergency Medical Response <i>Stephen J.M. Sollid, M.D., PhD</i>	316
Horizontal and Vertical Geometry <i>Victoria Gitelman</i>	322

Human Factors in Transportation <i>Alison Smiley, Christina (Missy) Rudin-Brown</i>	331
In-Depth Crash Analysis and Accident Investigation <i>Yong Peng, Helai Huang, and Xinghua Wang</i>	346
Incident Detection Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	351
Inequality and Traffic Safety <i>Miles Tight</i>	358
Lighting <i>John D. Bullough</i>	361
Macroscopic Safety Analysis <i>Mohamed Abdel-Aty and Jaeyoung Lee</i>	367
Motor Vehicle Crash Reportability <i>John J. McDonough</i>	380
Nominal Safety <i>Per Erik Garder</i>	386
Parking Lots <i>Maxim A. Dulebenets</i>	392
Passenger Van Safety <i>Saksith Chalermpong and Apiwat Ratanawaraha</i>	401
Passive Prevention Systems in Automobile Safety <i>B. Serpil Acar</i>	406
Pedestrian Safety, Children <i>Mette Møller</i>	415
Pedestrian Safety, General <i>Muhammad Z. Shah, Mehdi Moeinaddini, and Mahdi Aghaabbasi</i>	420
Pedestrian Safety, Older People <i>Carlo Lui</i>	429
Visually Impaired Pedestrian Safety <i>Robert S. Wall Emerson</i>	435
Photo/Video Traffic Enforcement <i>Charles M. Farmer</i>	439
Powered Two- and Three-Wheeler Safety <i>Fangrong Chang, Helai Huang, and Md. Mazharul Haque</i>	443
Railroad Safety <i>Xiang Liu and Zhipeng Zhang</i>	451
Railroad Safety: Grade Crossings and Trespassing <i>Rahim F. Benekohal and Jacob Mathew</i>	466
Recreational Boating Safety: Usage, Risk Factors, and the Prevention of Injury and Death <i>Amy E. Peden, Stacey Willcox-Pidgeon, and Kyra Hamilton</i>	477
Refuge Islands <i>Christer Hydén</i>	487
Risk Perception and Risk Behavior in the Context of Transportation <i>Martina Raue and Eva Lerner</i>	494

Road Diets	500
<i>Robert B. Noland</i>	
Road Safety Audits	508
<i>Xiao Qin</i>	
Roadside Safety Barriers	513
<i>Dean C. Alberson</i>	
Road Safety Management in Selected Countries	519
<i>Paul Boase and Brian Jonah</i>	
Roadway Pavement Conditions and Various Summer and Winter Maintenance Strategies	529
<i>Kamal Hossain, PhD P. Eng</i>	
Safety of Roundabouts	539
<i>Khaled Shaaban</i>	
Rumble Strips, Continuous Shoulder, and Centerline	549
<i>AnnaAnund and AnnaVadeby</i>	
Safety Culture	554
<i>Tor-Olav Nvestad</i>	
Safety Data Quality Management	560
<i>Robert A. Scopatz</i>	
School Bus Safety	564
<i>Yousif A. Abulhassan</i>	
School Campus Traffic Circulation	568
<i>Dimitrios Nalmpantis</i>	
Sexual Violence in Public Transportation	576
<i>Vania Ceccato</i>	
Shared Space	584
<i>Allison B. Duncan</i>	
Side Area Safety and Side Slopes	593
<i>José María Pardillo-Mayora</i>	
Simulators	602
<i>Nichole L. Morris, Curtis M. Craig, Jacob D. Achtemeier, and Peter A. Easterlund</i>	
Sleep-Related Issues and Fatigue	611
<i>Prof Marion Sinclair and Estelle Swart</i>	
Speed Governors and Limiters	617
<i>Christer Hydén</i>	
Speed Limits on Rural Highways	622
<i>Peter Tarmo Savolainen and Timothy Jordan Gates</i>	
Speed Limits on Urban Streets	632
<i>Anna Bray Sharpin, Claudia Adriazola-Steil, Ben Welle, and Natalia Lleras</i>	
Speed-Reducing Measures	641
<i>Vinod Vasudevan</i>	
Striping, Signs, and Other Forms of Information	648
<i>Kasem Choocharukul and Kerkritt Sriroongvikrai</i>	
Suicides	656
<i>Brendan Ryan</i>	

Surrogate Measures of Safety <i>Nicolas Saunier and Aliaksei Laureshyn</i>	662
Targeting Transit: the Terrorist Threat and the Challenges to Security <i>Brian Michael Jenkins</i>	668
The Swedish Vision Zero: A Policy Innovation <i>Matts-Åke Belin</i>	675
Tire Safety <i>Saied Taheri</i>	681
Town Gates: Section on Transport Safety and Security <i>Charles Tijus</i>	685
Traffic Flow Volume and Safety <i>Athanasios Theofilatos and Apostolos Ziakopoulos</i>	692
Traffic Safety and Security of Taxis and Ride-Hailing Vehicles <i>Zhe Wang, Helai Huang, and Ye Li</i>	699
Traffic Signals and Safety <i>Andrew P. Tarko</i>	706
Tunnels, Safety and Security Issues-Risk Assessment for Road Tunnels: State-of-the-Art Practices and Challenges <i>Konstantinos Kirytopoulos, Panagiotis Ntzeremes, and Konstantinos Kazaras</i>	713
Understanding, Managing, and Learning from Disruption <i>Karl Kim</i>	719
Use/Analysis of Crash Data and Underreporting of Crashes <i>Mohammadali Shirazi and Dominique Lord</i>	726
Utility Poles <i>Lai Zheng and Tarek Sayed</i>	731
Value of Life and Injuries <i>David A. Hensher</i>	737
Effects of Weather <i>Maria Pregnolato, Amirhassan Kermanshah, and Wisinee Wisetjindawat</i>	742
Wrong-Way Driving on Motorways <i>Huaguo Zhou and Md Atiquzzaman</i>	751

VOLUME 3

Freight Transport and Logistics <i>Sharon Cullinane and Kevin Cullinane</i>	1
Expanding the Perspective of Logistics and Supply Chain Management <i>David J. Closs</i>	8
Logistics and Supply Chain Management Performance Measures <i>David B. Grant and Sarah Shaw</i>	16
Economic Regulation/Deregulation and Nationalization/Privatization in Freight Transportation <i>Wayne K. Talley</i>	24
Freight Transport Policy <i>Luca Zamparini and Aura Reggiani</i>	29

Planning and Financing Logistics Spaces <i>Nicolas Raimbault</i>	35
Supply Chain Risk Management: Creating the Resilient Supply Chain <i>Richard Wilding</i>	41
Transportation Safety and Security <i>Maria G. Burns</i>	47
Resilience in Freight Transport Networks <i>Zhuohua Qu, Chengpeng Wan, and Zaili Yang</i>	53
Environmental Sustainability in Freight Transportation <i>Lisa M. Ellram</i>	58
Sustainable Logistics, CSR in Logistics, and Sustainable Supply Chain Management <i>Maria Björklund and Maja Piecyk-Ouellet</i>	64
Information Sharing and Business Analytics in Global Supply Chains <i>Prof Usha Ramanathan and Prof Ramakrishnan Ramanathan</i>	71
Logistics Information Systems <i>Petri Helo and Javad Rouzafzoon</i>	76
Factors Affecting the Selection of Logistics Service Providers <i>Aicha AGUEZZOUL</i>	85
Logistics Service Performance <i>Kee-hung Lai, Jinan Shao, and Yongyi Shou</i>	89
The World Bank's Logistics Performance Index <i>Christina K. Wiederer cwiederer, Jean-François Arvis, Lauri M. Ojala, and Tuomas M. M. Kiiski</i>	94
Outsourcing Logistics Functions <i>Evi Hartmann, Hendrik Birkel, and Matthias Kopyto</i>	102
Freight Transport and Logistics in JIT Systems <i>James H. Bookbinder and M. Ali AËelkü</i>	107
Supply Chain Finance <i>Erik Hofmann</i>	113
Packaging Logistics <i>Jesús García-Arca, Alicia Trinidad González-Portela Garrido, and J. Carlos Prado-Prado</i>	119
The Bullwhip Effect <i>Jan C. Fransoo and Maximiliano Udenio</i>	130
Blockchain Applications in Logistics <i>Yingli Wang</i>	136
Logistics in Asia <i>Shong-lee Ivan Su</i>	143
Logistics in the Developing World <i>Charles Kunaka</i>	150
Freight Network Modeling <i>Lóránt Tavasszy and Yousef Maknoon</i>	157
National Freight Transport Models <i>Gerard de Jong</i>	162
Freight Flows in Cities <i>Genevieve Giuliano</i>	168

Urban Logistics and Freight Transport <i>Michael Browne, José Holguín-Veras, and Julian Allen</i>	178
Omni-Channel Logistics <i>Tom Van Woensel</i>	184
Humanitarian Logistics <i>Gyöngyi Kovács and Diego Vega</i>	190
Optimization of Humanitarian Logistics <i>M. Teresa Ortuño, Jose M. Ferrer, Inmaculada Flores, and Gregorio Tirado</i>	195
Event Logistics <i>Rev. Ruth Dowson and Dan Lomax</i>	201
Reverse Logistics <i>Dale S. Rogers and Ronald S. Lembke</i>	208
E-Tailing and Reverse Logistics <i>Sharon Cullinane</i>	219
Green Routing of Freight Vehicles <i>Tolga Bektaş</i>	224
Freight Mode Choice <i>Hyun Chan Kim and Alan Nicholson</i>	231
The Value of Time in Freight Transport <i>María Feo-Valero, Amaya Vega, and Bárbara Vázquez-Paja</i>	236
Behavioral Research in Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	242
Container (Liner) Shipping <i>Theo Notteboom</i>	247
Bulk Shipping Markets: An Overview of Market Structure and Dynamics <i>Manolis G. Kavussanos and Stella A. Moysiadou</i>	257
Ferries and Short Sea Shipping <i>Lourdes Trujillo and Alba Martínez-López</i>	280
Shipping and the Environment <i>Karin Andersson, Selma Brynolf, Lena Granhag, and J. Fredrik Lindgren</i>	286
Energy Efficiency of Ships <i>Harilaos N. Psaraftis</i>	294
Seaports <i>Mary R. Brooks and Geraldine Knatz</i>	299
Port Hinterlands <i>Francesco Parola, Giovanni Satta, and Francesco Vitellaro</i>	305
Seaports as Clusters of Economic Activities <i>Peter W. de Langen</i>	310
Port Efficiency and Effectiveness <i>Lourdes Trujillo, María Manuela González, Casiano Manrique-De-Lara-Peñate, and Ivone Pérez</i>	316
Container Port Automation <i>Michael G.H. Bell</i>	323
Optimizing Crane Operations in Ports Scheduling of Liner Container Shipping Services	335

<i>Yuquan Du and Qiang Meng</i>	
Dry Ports	344
<i>Gordon Wilmsmeier and Jason Monios</i>	
Arctic Shipping	349
<i>Yufeng Lin, David G. Babb, and Adolf K.Y. Ng</i>	
Airfreight and Economic Development	355
<i>Kenneth Button</i>	
Air Freight Logistics	361
<i>Keith Debbage and Neil Debbage</i>	
Air Freight Marketing	369
<i>Lucy Budd and Stephen Ison</i>	
Drones in Freight Transport	374
<i>Oliver Kunze</i>	
Duty of Care in the Selection of Motor Carriers	382
<i>Thomas M. Corsi</i>	
Carrier Selection for Less-Than-Truckload (LTL) Shipments	388
<i>Dinçer Konur, Gonca Yildirim, and Bahriye Cesaret</i>	
Decarbonizing Road Freight Transport	395
<i>Heikki Liimatainen</i>	
The Rebound Effect in Road Freight Transport	402
<i>Tooraj Jamasb and Manuel Llorca</i>	
Autonomous Goods Transport	407
<i>Heike Flømig</i>	
Rail Freight	413
<i>Dr Allan Woodburn</i>	
Rail Freight Vehicles	423
<i>Maksym Spiryagin, Qing Wu, Peter Wolfs, Colin Cole, Valentyn Spiryagin, and Tim McSweeney</i>	
Eurasia Rail Freight: Enablers and Inhibitors of Future Growth	436
<i>Hendrik Rodemann and Simon Templar</i>	
Intermodal and Synchromodal Freight Transport	456
<i>Tomas Ambra, Koen Mommens, and Cathy Macharis</i>	
Pipelines	463
<i>Matthew E. Oliver</i>	
3D Printers and Transport	471
<i>Wouter P.C. Boon and Bert van Wee</i>	
The Physical Internet and Logistics	479
<i>Eric Ballot and Shenle PAN</i>	
Bicycles for Urban Freight	488
<i>Barbara Lenz and Johannes Gruber</i>	
China's Belt and Road Initiative	495
<i>Paul Tae-Woo Lee</i>	

VOLUME 4

Introduction to Traffic Management <i>Edward C.S. Chung</i>	1
Urban Motorway Management <i>John Gaffney and Hendrik Zurlinden</i>	2
Ramp Metering Application <i>John Gaffney and Hendrik Zurlinden</i>	10
City Wide Coordinated Ramp Meters <i>John Gaffney and Hendrik Zurlinden</i>	21
Variable Speed Limits for Traffic Efficiency Improvement <i>José Ramón D. Frejo and Bart De Schutter</i>	33
Hard Shoulder Running <i>Justin Geistefeldt</i>	41
High-Occupancy Vehicle (HOV) and High-Occupancy Toll (HOT) Lanes <i>Roxana J. Javid, Jiani Xie, Lijiao Wang, Wenruifan Yang, Ramina Jahanbakhsh Javid, and Mahmoud Salari</i>	45
Reversible Lanes: Guidelines, Operation and Control, Research Directions <i>Gowri Asaithambi, Venkatesan Kanagaraj, and Madhuri Kashyap</i>	52
Electronic Toll Collection <i>Azusa Toriumi</i>	60
Road Pricing-Theory and Applications <i>Kian Keong Chin</i>	68
Road Pricing 1: The Theory of Congestion Pricing <i>Timothy D. Hau</i>	74
Road Pricing 2: Short- and Long-Run Equilibrium of Road Transportation <i>Timothy D. Hau</i>	83
Road Pricing 3: The Implications for Pricing Public Transportation <i>Timothy D. Hau</i>	90
Road Pricing 4: Case Study-The Implementation of Electronic Road Pricing in Hong Kong <i>Timothy D.</i>	103
Advanced Travelers Information Systems (ATIS) <i>Chintan Advani and Ashish Bhaskar</i>	106
Travel Time Reliability <i>Sharmili Banik, Anil Kumar, and Lelitha Vanajakshi</i>	109
Traffic Incident Management <i>Ruimin Li</i>	122
Traffic Incident Detection <i>Shuyan Chen and Yingjiu Pan</i>	128
Bottleneck <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	134
Flow Breakdown <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	143
Recurrent Congestion <i>Takahiro Tsubota</i>	154

Nonrecurrent Congestion <i>Takahiro Tsubota</i>	158
Freeway to Arterial Interfaces <i>Abolfazl Karimpour and Yao-Jan Wu</i>	162
Arterial Road Management <i>Jiaqi Ma, Yi Guo and Adekunle Adebisi</i>	169
Signalized Intersections <i>Chaitrali Shirke</i>	178
Protected Phase <i>Jiarong Yao, Yumin Cao, and Keshuang Tang</i>	185
Permitted Phase <i>Yumin Cao, Jiarong Yao, and Keshuang Tang</i>	196
Hook Turns: Implementation, Benefits, and Limitations <i>Sara Moridpour and Amir Falamarzi</i>	203
Traffic Signal Coordination <i>Rahim F. Benekohal</i>	213
Emergency Vehicle Priority (Preemption): Concept and Advancements <i>Chaitrali Shirke</i>	221
Signalized Roundabouts <i>Yetis Sazi Murat and Rui-jun Guo</i>	227
Turbo Roundabouts: Design, Capacity and Comparison With Alternative Types of Roundabouts <i>Marco Guerrieri and Raffaele Mauro</i>	238
Capacity of an Intersection <i>Ashish Verma and Milan Mathew Thomas</i>	247
Delay <i>Shinji Tanaka</i>	258
Queue Length <i>Shinji Tanaka</i>	263
Local Area Traffic Management <i>Michael A.P. Taylor</i>	268
On-Street Parking <i>Jun Chen and Guang Yang</i>	278
Off-Street Parking <i>Jun Chen and Guang Yang</i>	285
Parking Information Systems <i>Behrang Assemi and Douglas Baker</i>	289
Performance-Based Parking Management <i>Douglas Baker and Behrang Assemi</i>	299
Transit Priority <i>Wanjing Ma, Qiheng Lin, and Ling Wang</i>	307
Adaptive Bus Control <i>Monica Menendez</i>	315
Transit Fare Collection <i>Mahmoud Mesbah and Kamal Khanali</i>	325

Transit Information Systems <i>Ankit Kumar Yadav and Nagendra R Velaga</i>	331
Tram Lane Configurations and Driving Rules <i>Farhana Naznin</i>	338
Railway Crossing <i>Chunliang Wu and Inhi Kim</i>	341
Pedestrian Crossing (Crosswalk) <i>Miho Iryo-Asano, Wael K.M. Alhajyaseen, and Koji Suzuki</i>	346
Bike-Sharing System: Uncovering the “1Success Factors” <i>S.K. Jason Chang and Amanda Fernandes Ferreira</i>	355
Airspace Systems Technologies-Overview and Opportunities <i>Banavar Sridhar and Gano B. Chatterji</i>	363
Air Traffic Flow and Capacity Management <i>Li Weigang and Cristiano P Garcia</i>	380
Port Management <i>Maria G. Burns</i>	390
Port Performance Measurement from a Multistakeholder Perspective <i>Min-Ho Ha, Zaili Yang, and Young-Joon Seo</i>	396
Transport Modeling and Data Management <i>Chandra Bhat</i>	407
Computational Methods and Data Analytics <i>Bilal Farooq and David López</i>	408
Activity-Based Models <i>Renato Guadamuz and Rajesh Paleti</i>	414
Advanced Traveler Information Systems <i>Stephen D. Boyles</i>	418
Demand-Responsive Transit, Evaluation Studies <i>Sebastián Raveau</i>	423
Location Choice Models <i>Adam Wilkinson Davis</i>	428
Full Feedback and Equilibrium Modeling in Urban Travel Forecasting <i>Yu (Marco) Nie, Jun Xie, and David Boyce</i>	432
The Use and Value of Geographic Information Systems in Transportation Modeling <i>Ming Zhang</i>	440
Latent Demand and Induced Travel <i>Charisma F. Choudhury</i>	448
ICT, Virtual and In-Person Activity Participation, and Travel Choice Analysis <i>Jacek Pawlak and Giovanni Circella</i>	452
Public Transit Ridership Forecasting Models <i>Ipsita Banerjee, Deepa L, and Abdul Rawoof Pinjari</i>	459
Spatial Mismatch, Job Access, and Reverse Commuting <i>Gian-Claudia Sciarra</i>	468

Microsimulation and Agent-Based Models in Transportation <i>Milos Balac</i>	473
Choice Models in Transportation <i>Naveen Chandra Iraganaboina and Naveen Eluru</i>	477
Multi-Criteria Decision Analysis <i>Zhanmin Zhang and Srijith Balakrishnan</i>	485
The National Household Travel Survey Data Series (NPTS/NHTS) <i>Nancy McGuckin</i>	493
Route Choice and Network Modeling <i>Emma Frejinger and Maëlle Zimmermann</i>	496
Departure Time Choice Modeling <i>Khandker Nurul Habib</i>	504
Traveler Responses to Congestion <i>David T. Ory and Gayathri Shivaraman</i>	509
Origin-Destination Demand Estimation Models <i>William H.K. Lam, Hu Shao, Shuhan Cao, and Hai Yang</i>	515
Parking Demand Models <i>S.C. Wong, Zhi-Chun Li, and William H.K. Lam</i>	519
Pavement Management Systems <i>Senthilmurugan Thyagarajan</i>	524
Residential Location Choice Models <i>Shlomo Bekhor and Sigal Kaplan</i>	531
Transport Demand Management <i>Feiyang Zhang and Becky P.Y. Loo</i>	537
Traffic Flow Analysis <i>H. Michael Zhang and Jia Li</i>	544
Autonomous Vehicles and Transportation Modeling <i>Annesha Enam, Felipe de Souza, Omer Verbas, Monique Stinson, and Joshua Auld</i>	557
Ride-Hailing and Travel Demand Implications <i>Felipe F. Dias</i>	564
Transportation Modeling and Planning Software <i>Joel Freedman</i>	569
Transportation Statistics and Databases <i>Taha Hossein Rashidi</i>	574
Travel Surveys <i>Stacey G. Bricka</i>	587
Travel Demand Forecasting: Where Are We and What Are the Emerging Issues <i>Thomas F. Rossi</i>	590
Travel Model Calibration and Validation <i>Ram M. Pendyala</i>	596
Trip Chaining Analysis <i>Cynthia Chen and Yusak Susilo</i>	606
Vehicle Ownership Models <i>Dr. Sören Groth and Prof. Dr. Dirk Wittowsky</i>	612

Bicycle Sharing/Bikesharing <i>Catherine Morency and Jean-Simon Bourdeau</i>	617
Carsharing <i>Shiva Habibi and Frances Sprei</i>	623
Urban Recreational Travel <i>Long Cheng and Frank Witlox</i>	629
Place Perception and Travel Behavior <i>Kathleen Deutsch-Burgner and Konstadinos G. Goulias</i>	635

VOLUME 5

Transport Modes <i>Edoardo Marcucci</i>	1
Infrastructure Transport Investments, Economic Growth and Regional Convergence <i>Xavier Fageda and Cecilia Olivieri</i>	2
Transport Modes and an Aging Society <i>Charles B.A. Musselwhite and Theresa Scott</i>	6
Sustainable Mobility Paths <i>Erling Holden and Geoffrey Gilpin</i>	13
Vehicles that Drive Themselves: What to Expect with Autonomous Vehicles <i>Michele D. Simoni and Kara M. Kockelman</i>	19
Transport Modes and Tourism <i>Ila Maltese and Luca Zamparini</i>	26
Transport Modes and Accessibility <i>Bert van Wee</i>	32
Transport Modes and Globalization <i>Jean-Paul Rodrigue</i>	38
Transport Modes and Cities <i>Erick Guerra and Gilles Duranton</i>	45
Transport Modes and Remote Areas	51
Modeling Mode Choice in Freight Transport <i>Lóránt Tavasszy and Gerard de Jongb</i>	57
Travel Mode Choice as Reasoned Action <i>Sebastian Bamberg, Icek Ajzen, and Peter Schmidt</i>	63
Energy Consumption of Transport Modes <i>Zissis Samaras and Ilias Vouitsis</i>	71
Transport Modes and People With Limited Mobility <i>Roger L Mackett</i>	85
Transport Modes and Commuters <i>Colin G. Pooley</i>	92
Shopping and Transport Modes <i>Antonio Comi</i>	98
Transport Modes and Health <i>Jennifer S. Mindell and Sandra Mandic</i>	106

Multimodality in Transportation <i>Sören Groth and Tobias Kuhnimhof</i>	118
Transport Modes and Disasters <i>Brian Wolshon</i>	127
Big Data for Public Transport Planning <i>Jan-Dirk Schmöcker</i>	134
Active Transport: Heterogeneous Street Users Serving Movement and Place Functions <i>Regine Gerike, Stefan Hubrich, Caroline Koszowski, Bettina Schröter, and Rico Wittwer</i>	140
Electric Vehicles <i>Christine Eisenmann, Daniel Görges, and Thomas Franke</i>	147
Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service <i>Susan Shaheen, PhD and Adam Cohen</i>	155
Adoption of new travel information platforms <i>Sigal Kaplan</i>	160
ICT and Transport Modes <i>Galit Cohen-Blankshtain</i>	165
Mode Choice and Life Events <i>Joachim Scheiner</i>	171
Introduction to Air Transport <i>Milan Janić</i>	178
The History of Air Transportation <i>Richard P. Hallion</i>	192
The Geography of Air Transport <i>Lucy Budd and Stephen Ison</i>	198
The Future of Air Transport <i>Rico Merkert and James Bushell</i>	203
Air Transport and Its Territorial Implications <i>Lanfranco Senn</i>	208
Next Generation Travel: Young Adults' Travel Patterns <i>Tobias Kuhnimhof and Scott Le Vine</i>	215
Airport Network Planning and Its Integration with the HSR System <i>Francesca Pagliara, Juan Carlos Martín, and Concepción Román</i>	222
Airport Management <i>Peter Forsyth and Hans-Martin Niemeier</i>	229
Airport Regulation <i>Achim I. Czerny</i>	234
Air Route Planning and Development <i>Renan Peres de Oliveira and Gui Lohmann</i>	241
Airline Management <i>Sveinn Vidar Gudmundsson</i>	249
Air Cargo <i>Volodymyr Bilotkach</i>	258

Airline Regulation <i>Andrew R. Goetz</i>	263
Air Vehicles Classification <i>Vincenzo Torre</i>	269
Aircraft Manufacturing <i>Antonio Sollo</i>	290
A Geography of Road Transport in Cities <i>Cristian Domarchi and Juan de Dios Ortúzar</i>	300
The Future of Road Transport <i>Preston L. Schiller</i>	306
Road Transportation and Territorial Scale <i>Ana M. Condeço-Melhorado</i>	315
Road Modes: Walking <i>Kevin Manaugh, PhD Associate Professor</i>	320
Bus Public Transport Planning and Operations <i>Ehab Diab and Ahmed El-Geneidy</i>	326
Street Design for Active Travel <i>Bruce Appleyard</i>	333
Transit Planning and Management <i>Zakhary Mallett and Marlon G Boarnet</i>	349
Road Infrastructure: Planning, Impact and Management <i>Jos Arts, Wim Leendertse, and Taede Tillema</i>	360
Road Transport Planning at the Urban Scale <i>David A. King and Kevin J. Krizek</i>	373
Road Traffic Regulation: Road Pricing and Environmental Quality <i>Marco Percoco</i>	378
Car Ownership and Car Use: A Psychological Perspective <i>J.L. Veldstra, A.B. Ünal, E.M. Steg</i>	384
Road Transport: E-Scooters <i>Gysele Lima Ricci and Klaus Bogenberger</i>	391
Railway Station and Network Planning <i>Ingo Arne Hansen</i>	399
Introduction to Rail Transport <i>Chris Nash and Tony Fowkes</i>	406
The History of Rail Transport <i>Carlo Ciccarelli, Andrea Giuntini, and Peter Groote</i>	413
The Geography of Rail Transport <i>Frédéric Dobruszkes and Amparo Moyano</i>	427
Rail Transport and Territorial Scale <i>Prof. Andrés Monzón and Dr. Elena López</i>	437
Railway Management <i>Vassilios A. Profillidis</i>	444
Railway Terminal Regulation <i>Nacima Baron</i>	454

Service Network Design for Freight Railroads <i>Teodor Gabriel Crainic</i>	464
Subway Systems <i>Guillaume Monchambert, Daniel Hörcher, Alejandro Tirachini, and Nicolas Coulombel</i>	471
Railway Company Management <i>Vilius Nikitinas, Skaistė Miliauskaitė</i>	479
Regulation of Rail Infrastructure and Services <i>Javier Campos</i>	485
Rail Vehicle Classification <i>Christos Pyrgidis and Alexandros Dolianitis</i>	490
Introduction to Maritime Shipping <i>Christa Sys and Thierry Vanelslander</i>	508
The Geography of Maritime Transport <i>César Ducruet and Justin Berli</i>	517
The Future of Maritime Transport <i>Harilaos N. Psaraftis</i>	535
Maritime Transport and Territorial Scale <i>Brian Slack</i>	540
Containerization and the Port Industry <i>Hercules Haralambides</i>	545
Port Management <i>Lourdes Trujillo, Daniel Castillo Hidalgo, and Manuel Herrera</i>	557
Economic and Environmental Regulation in the Port Sector <i>Beatriz Tovar and Alan Wall</i>	563
Maritime Route Planning <i>Johan Woxenius</i>	570
Inland Waterway Transport and Inland Ports: An Overview of Synchromodal Concepts, Drivers, and Success Cases in the IWW Sector <i>Behzad Behdani, Bart Wiegmans, and Yun Fan</i>	577
Maritime Company Passenger Management/Liner Industry <i>Claudio Ferrari and Alessio Tei</i>	587
Cruise Industry <i>Athanasios A. Pallis and Aimilia A. Papachristou</i>	593
International Maritime Regulation: Closing the Gaps Between Successful Achievements and Persistent Insufficiencies <i>Laurent Fedi</i>	600
Ship Classification <i>Gareth C. Burton and Mimosa T. Miller</i>	607
The Shipbuilding Industry and its Interactions With Shipping <i>Paul William Stott</i>	617
Methods for Designing Public Transport Networks <i>Zain Ul Abedin and Avishai (Avi) Ceder</i>	625
Space Transportation <i>Mark Hempsell</i>	638

Pipelines	646
<i>Franco Cotana and Mattia Manni</i>	
Women and Transport Modes	656
<i>Priya Uteng, PhD and Yusak Susilo, PhD</i>	
Transport Modes and Big Data	665
<i>Hannah D Budnitz, Emmanouil Tranos, and Lee Chapman</i>	
Transport Modes and Inequalities	671
<i>Caroline Mullen</i>	
Railway Traffic Management	678
<i>Francesco Corman</i>	
Indoor Transportation	684
<i>Lutfi Al-Sharif</i>	
Bicycle as a Transportation Mode	697
<i>Raktim Mitra and Paul M. Hess</i>	
Urban Air Mobility: Opportunities and Obstacles	702
<i>Adam Cohen and Susan Shaheen, PhD</i>	
Transport Modes and Sustainability	710
<i>Long Cheng, Jonas De Vos, and Frank Witlox</i>	

VOLUME 6

Introduction to Transport Policy and Planning	1
<i>Maria Attard</i>	
Workplace Parking Levy	2
<i>Stephen Ison and Lucy Budd</i>	
Air Transport	7
<i>Lucy Budd and Stephen Ison</i>	
Mobility as a Service	12
<i>Milo N. Mladenović</i>	
Bicycle Sharing	19
<i>Cyrille Médard de Chardon</i>	
Light Rail	31
<i>Fiona Ferbrache</i>	
Planning Tourism Travel	39
<i>Luca Zamparini</i>	
Transport Planning and Management and its Implications in Chinese Cities	44
<i>Mengqiu Cao</i>	
Road Safety	51
<i>George Yannis and Eleonora Papadimitriou</i>	
Mobility Planning and Policies for Older People	59
<i>Charles B A Musselwhite</i>	
Electric Mobility	64
<i>Graham Parkhurst</i>	
Transport Policy and Governance	73
<i>Lisa Hansson</i>	

Transferability of Urban Policy Measures <i>Paul Martin Timms</i>	77
Accessibility Tools for Transport Policy and Planning <i>Benjamin Büttner</i>	83
Demand Responsive Transport <i>Marcus Enoch</i>	87
Transport and Climate Change <i>Robin Hickman and Christine Hannigan</i>	94
Taxicabs and Microtransit <i>David A. King</i>	101
Land-Use and Transport Planning <i>Luis A. Guzman</i>	107
Parking <i>Stephen Ison and Lucy Budd</i>	113
Transport Planning in the Global South <i>Daniel Oviedo and Mariajose Nieto-Combariza</i>	118
Planning for Children's Independent Mobility <i>E. Owen D. Waygood and Raktim Mitra</i>	125
The Politics of Mobility Policy <i>Geoff Vigar</i>	131
Technology Enabled Data for Sustainable Transport Policy <i>Susan M. Grant-Muller, Mahmoud Abdelrazek, Hannah Budnitz, Caitlin D. Cottrill, Fiona Crawford, Charisma F. Choudhury, Teddy Cunningham, Gillian Harrison, Frances C. Hodgson, Jinhyun Hong, Adam Martin, Oliver O'Brien, Claire Papaix, and Panagiotis Tsoleridis</i>	135
Car Sharing <i>Cyriac George and Tanu Priya Uteng</i>	142
Toward a More Holistic Understanding of Mega Transport Project (MTP) Success <i>John Ward</i>	147
Equity Considerations in Transport Planning <i>Karel Martens</i>	154
Planning for Rail Transport <i>Simon P. Blainey</i>	161
Connected and Autonomous Vehicles: Priorities for Policy and Planning <i>Dr. Alexandros Nikitas</i>	167
Gendered Mobility <i>Sheila Mitra-Sarkar</i>	173
Modeling and Simulation for Transport Planning <i>Michela Le Pira, Giuseppe Inturri, and Matteo Ignaccolo</i>	184
Externalities and External Costs in Transport Planning <i>Silvio Nocera</i>	191
Planning for Public Transport with Automated Vehicles <i>Gonçalo Homem de Almeida Rodriguez Correia</i>	198
Urban Congestion Charging in Transport Planning Practice <i>Ida Kristoffersson and Maria Börjesson</i>	206

Energy and Transport Planning <i>Debbie Hopkins and Christian Brand</i>	214
Customer Satisfaction as a Measure of Service Quality in Public Transport Planning <i>Laura Eboli and Gabriella Mazzulla</i>	220
Car Sharing and the Impact on New Car Registration <i>Mario Intini and Marco Percoco</i>	225
Evaluation Methods in Transport Policy and Planning <i>Niek Mouter</i>	230
Transport and Air Quality Planning and Policy <i>Dr Fabio Galatioto</i>	236
Cycling Policies <i>Esther Anaya-Boig</i>	241
Community Severance <i>Paulo Anciaes and Jennifer S. Mindell</i>	246
Planning for Bus Priority <i>Claus H. Sørensen, Fredrik Pettersson, and Joel Hansson</i>	254
High-Speed Rail and the City <i>Marie Delaplace</i>	261
Public Engagement in Transport Planning <i>Miriam Ricci</i>	266
Long-Distance Travel <i>Giulio Mattioli and Muhammad Adeel</i>	272
Urban Freight Policy <i>Laetitia Dablanç</i>	278
Regional Transport Planning <i>Chia-Lin Chen</i>	286
Transport Project Financing <i>Romeo Danielis and Lucia Rotaris</i>	292
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	298
Transitions and Disruptive Technologies in Transport Planning <i>Kate Pangbourne and Maria Attard</i>	309
Community Transport: Filling the Gaps for Those in Need of Mobility <i>Ian Shergold</i>	314
Fundamental Emerging Concepts and Trends for Environmental Friendly Urban Goods Distribution Systems <i>Sandra Melo</i>	320
Plateau Car <i>David Metz</i>	324
Emerging Trends in Transport Demand Modeling in the Transition Toward Shared Mobility and Autonomy <i>Patrizia Franco</i>	331
Policy and Planning for Walkability <i>Carlos Cañas Sanz and Maria Attard</i>	340

Public Transport Subsidy and Regulation <i>Jonathan Cowie</i>	349
Urban Regeneration and Transportation Planning <i>Thomas Vanoutrive</i>	356
Social and Distributional Impact Assessment in Transport Policy <i>Laura Walker and Angela Curl</i>	361
Mobility Planning for Healthy Cities <i>Ersilia Verlinghieri</i>	368
Transport Demand Management <i>Begoña Guirao</i>	374
Planning and the Global Movement of Goods and Commodities <i>Christopher Clott and Chris Petrocelli</i>	380
Public Transport Network Planning <i>Corinne Mulley and John D. Nelson</i>	388
A Timely Perspective on Planning for Ageing Infrastructure <i>Anthony Perl</i>	395
The role of media in transport planning and the transport policy process <i>Özgül Ardiç and J.A. Annema</i>	400
Travel Plans <i>Stephen Potter and Marcus Enoch</i>	408
Home Deliveries and their Impact on Planning and Policy <i>Rosário Macário</i>	413
Planning for Safe and Secure Transport Infrastructure <i>Per Erik Gårder</i>	418
Sensors and Data Driven Approaches in Transport <i>Mohammad Sadrani and Constantinos Antoniou</i>	426
Planning Active Travel and School Transport <i>Fahimeh Khalaj, Dorina Pojani, and Sara Alidoust</i>	432
VOLUME 7	
Introduction to Transport Psychology <i>Carlo Prato</i>	1
From Self-reports to Auto-Tech-Detect (ATD)-based Self-reports in Traffic Research <i>Türker Özkan and Timo Lajunen</i>	2
Observational Field Studies in Traffic Psychology <i>Tova Rosenbloom and Hodaya Levy</i>	8
Driving Simulators <i>Karel A. Brookhuis</i>	14
Naturalistic Driving Studies: An Overview and International Perspective <i>Johnathon P. Ehsani, Joanne L. Harbluk, Jonas Bärghman, Ann Williamson, Jeffrey P. Michael, Raphael Grzebieta, Jake Olivier, Jan Eusebio, Judith Charlton, Sjaanie Koppel, Kristie Young, Mike Lenné, Narelle Haworth, Andry Rakotonirainy, Mohammed Elhenawy, Gregoire Larue, Teresa Senserrick, Jeremy Woolley, Mario Mongiardini, Christopher Stokes, Paul Boase, John Pearson, and Feng Guo</i>	20

A Detailed Approach to Qualitative Research Methods <i>Sonja Forward and Lena Levin</i>	39
Behavioral Change <i>Sonja Haustein</i>	46
Habitual Behavior <i>Carlo G. Prato</i>	54
Driving Behavior and Skills <i>Timo Lajunen and Türker Özkan</i>	59
The Multidimensional Driving Style Inventory <i>Orit Taubman - Ben-Ari</i>	65
Risk Perception in Transport: A Review of the State of the Art <i>Trond Nordfjrn, An-Magritt Kummeneje, Mohsen F. Zavareh, Milad Mehdizadeh, and Torbjørn Rundmo</i>	74
Social-Symbolic and Affective Aspects of Car Ownership and Use <i>Birgitta Gatersleben</i>	81
Pitfalls of Statistical Methods in Traffic Psychology <i>J.C.F. de Winter and D. Dodou</i>	87
Data Analysis: Structural Equation Models <i>Marco Diana</i>	96
Data Analysis: Integrated Choice and Latent Variable Models <i>Carlo G Prato</i>	102
Explaining Data Analysis Using Qualitative Methods <i>Lena Levin and Sonja Forward</i>	107
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	113
Driver Aggression and Anger <i>Mark J.M. Sullman and Amanda N. Stephens</i>	121
Speeding: A "Tragedy of the Commons" Behavior <i>Bryan E. Porter, Thomas D. Berry, and Kristie L. Johnson</i>	130
Cycling as a Mode Choice: Motivational Psychology <i>Sigal Kaplan</i>	136
Motorcyclists <i>Narelle Haworth</i>	144
Drivers' Hazard Perception Skill <i>Mark S. Horswill and Andrew Hill</i>	151
Driver Education and Training for New Drivers: Moving beyond Current 'Wisdom' to New Directions <i>Teresa Senserrick, Oscar Oviedo-Trespalacios, David Rodwell, and Sherrie-Anne Kaye</i>	158
Road Safety Advertising: What We Currently Know and Where to From Here <i>Ioni Lewis, Barry Watson, Katherine M. White, and Sonali Nandavar</i>	165
Traffic Law Enforcement Theories and Models <i>Richard Tay</i>	171
Satisfaction with Travel and the Relationship to Well-Being <i>Tommy Gørling and Filip Fors Connolly</i>	177
	182

Electromobility: History, Definitions and an Overview of Psychological Research on a Sustainable Mobility System <i>Josef F. Krems and Isabel KreiÃŸig</i>	
Sharing: Attitudes and Perceptions <i>Yi Wen and Christopher R Cherry</i>	187
Women's Travel Patterns, Attitudes, and Constraints Around the World <i>Sandra Rosenbloom</i>	193
Driver Stress and Driving Performance <i>Lisa Dorn</i>	203
Introduction to Sustainability and Health in Transportation <i>Roger Vickerman</i>	225
Sustainable Development Goals and Health <i>Rosa Surinach</i>	226
Human Ecology <i>Roderick J. Lawrence</i>	234
Car- Free Cities <i>Haneen Khreis and Mark J. Nieuwenhuijsen</i>	240
Superblocks Base of a New Model of Mobility and Public Space. Barcelona as an Example <i>Salvador Rueda Palenzuela</i>	249
Resilience of Transport Systems <i>ErikJenelius and Lars-GöranMattsson</i>	258
Impact of Shipping to Atmospheric Pollutants: State-of-the-Art and Perspectives <i>Daniele Contini and Eva Merico</i>	268
Noise Pollution From Transport <i>Marianna Jacyna, Emilian Szczepański, Konrad Lewczuk, Mariusz Izdebski, Ilona Jacyna-Gotda, Michał, Kłodawski, Paweł, Gotda, Piotr Got, and Ebiowski</i>	277
Visual Impacts From Transport <i>Paulo Rui Anciaes</i>	285
Light Pollution <i>John D. Bullough</i>	292
Wildlife Crossings and Barriers <i>Scott D. Jackson</i>	297
Environmental Justice, Transport Justice, and Mobility Justice <i>Devajyoti Deku</i>	305
Transport Noise and Health <i>Elisabete F. Freitas, Emanuel A. Sousa, and Carlos C. Silva</i>	311
Climate Change and Health, Related to Transport <i>Ersilia Verlinghieri</i>	320
Urban Greenspace, Transportation, and Health <i>Payam Dadvand and Mark J. Nieuwenhuijsen</i>	327
Transport Access and Health <i>Alireza Ermagun</i>	335
Social Exclusion and Health, Related to Transport <i>Roger L. Mackett</i>	341

Burden of Disease Assessment <i>David Rojas-Rueda</i>	347
Achieving a Near-Zero CO <i>Lewis M. Fulton</i>	353
Disabled Travelers <i>Bryan Matthews</i>	359
Health Impacts of Connected and Autonomous Vehicles <i>Soheil Sohrabi</i>	364
Electric Vehicles and Health <i>Kanok Boriboonsomsin</i>	372
Shared Mobility Opportunities and their Computational Challenges for Improving Health-Related Quality of Life <i>Cristiano Martins Monteiro, Cláudia Aparecida Soares Machado, Adelaide Cassia Nardocci, Fernando Tobal Berssaneti, José Alberto Quintanilha, and Clodoveu Augusto Davis</i>	376
Bike Sharing and Health <i>David Rojas-Rueda and Mark J. Nieuwenhuijsen</i>	384
E-Bikes and Health <i>Aslak Fyhri and Hanne Beate Sundfør</i>	393
<i>Index</i>	399

Introduction to Transport Policy and Planning

Maria Attard, Department of Geography, Faculty of Arts, University of Malta, Msida, Malta

© 2021 Elsevier Ltd. All rights reserved.

In a world dominated by growing cities, increasing population, global environmental change at a scale which is unprecedented, the issues surrounding transport and mobility come to the forefront to support cities and functioning urban areas, mobilize and provide opportunity for populations around the world and unfortunately contribute significantly to the climate crisis. The COVID-19 pandemic which has spread across the world at the time of writing this introduction has also shown the key role of transport in both spreading the contagion as well as supporting efforts to control the virus amidst fears of economic collapse. The changes affected to transport systems during the year 2020 across a number of cities in Europe and beyond, will provide ample opportunities for urban and transport planners in the future to seek better ways of connecting people and places, with less negative impact on the environment and more positive impact on community well being.

The importance of transport policy and planning has never been so crucial, and research in this field is necessary because of the dynamic nature of transport systems and the speed of change experienced in the last decades. Changes that are being driven by technology but not only. The necessity to reduce emissions and the need for a better quality of life, with well being becoming a growing priority for many policy makers around the world, is changing the attitudes and perspectives of urban and transport planners who are moving away from the classic planning models of the 20th century into uncharted territories of liberalized and less regulated transport markets, higher levels of mobilities, greater access to information (big data) and customers, and different travel behavior needs in cities worldwide. Policy making is becoming more reactive to population demands, some choosing more populist approaches, other driven by increasing media and social media polarization, and some moving forward and leading the way toward sustainability and net-zero carbon transport systems.

This section of the Encyclopedia of Transportation deals with transport policy and planning. It provides for a very broad spectrum of topics and subjects which span the policy and planning fields. The number and variety in the articles and topics underline the extensive nature of transport policy and planning, and its crosscutting character with and across other disciplines, both in transport and beyond. The articles include the classical and more basic foundation topics of land use and transport planning, evaluation, project financing, transport demand management, planning for public transport, rail and air transport, transport externalities, safety and security, and goods and freight transport. Others are more focused on elements of transport policy such as parking, with an article dealing specifically with the most recent concepts of the workplace parking levy, taxi services and microtransit in cities, tourism travel, road pricing, and climate change. There are articles which provide for an understanding of planning tools such as the latest developments in accessibility tools, the use of sensors and data in transport planning, multimodality and transport demand modeling tools, and modeling and simulation tools. The need to focus on specific population groups is provided in chapters dealing with elderly mobility, childrens' independent mobility and active travel, community transport, and a look at gendered mobilities.

Furthermore, this section of the Encyclopedia provides for a discussion of topics related to the politics of transport, the role of media, as well as public engagement. All important topics when considering how transport policy is formulated and how transport planning works. The growing importance of transport justice and equity in transport is covered by articles dealing with equity, social and distributional impact assessment in transport, policy and governance, transport for healthy cities, severance, and the importance of walkability, especially in urban areas. The section also includes articles dealing with transport planning in China and the Global South, both considered important in the overall discussions and debates about transport planning in large, but potentially poor cities.

And finally the section would not have been complete without a look at the emerging technologies that are supporting the development of new transport systems and services. These include articles on Mobility as a Service (MaaS), the rise of bicycle and car sharing, electric mobility, autonomous and connected vehicles and the technologies that will support such future modes through Intelligent Transport Systems and infrastructures. Counter to these more future oriented discussions, a chapter dealing with ageing infrastructure is included to highlight the imminent need of transport infrastructures which, like any other infrastructure, fall victim to financial and structural sustainability issues.

There are many other articles not mentioned in this introduction which play a significant role in the future of transport policy and planning. I invite readers to browse through the subjects and also crossreference with other sections of this encyclopedia. In that manner readers will learn to appreciate the breadth of topics and disciplines that make up transport policy and planning.

Workplace Parking Levy

Stephen Ison, Lucy Budd, Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	2
Background	3
The Nottingham WPL	4
Public Acceptance of the WPL	4
The Structure of the WPL Scheme	5
Impact of the Nottingham WPL	5
Limitations of the WPL and Lessons Learned	5
Conclusion	5
References	5
Further Reading	6

Introduction

Transport Demand Management (TDM) is a catch-all term used to describe a range of strategies and interventions which are designed to influence individual travel behavior and result in the more efficient use of increasingly scarce road space (Ison and Rye, 2008). More specifically, TDM covers a range of measures which aim to address urban traffic-related issues associated with the prevalence of the private motor car as the dominant mode of urban mobility. The negative externality effects of private car use, including congestion, environmental degradation, the inefficient use of land for parking stationary vehicles, skewed economic development, reduced public transport provision and patronage, and increased risk of road accidents, impact all users of the urban space, irrespective of whether or not they use a car. For example, pedestrians are subjected to the impacts of air pollution. TDM measures include not only economic and regulatory interventions but also support for modal substitution and enhanced public transport provision (Table 1).

Imposing a direct charge on users of private cars provides a financial disincentive to car use. Charges can be levied both on mobile private vehicles, through road user charges or tolls, and/or on stationary vehicles that are parked. Regulatory controls on the road space that private vehicles are (or are not) permitted to use when in motion (for example, dedicated bus corridors, high vehicle occupancy lanes, and pedestrianized zones) and when parked are also employed. The regulation of parking policy and the charging regimes that are employed thus form an important component of any urban TDM strategy. Significant areas of land are devoted to car parking spaces in cities world-wide with many of the spaces free of charge to motorists (Burchell et al., 2019). In fact, Shoup (2005) stated that “parking is the single biggest land use in cities”. There is a high cost involved in this land use and since it is often free for office car park users motoring can be seen as more attractive than it would be if charged for so as to meet the full economic cost of providing a space. It is clearly the case that providing car parking spaces free of charge can encourage the use of the car leading to traffic congestion and environmental degradation. The level of traffic congestion can indeed be more severe since commuters will be traveling at morning- and evening-peak periods. While urban authorities are generally able to control on-street and off-street parking provision in multistorey car parks or surface facilities that falls under their direct jurisdiction, their ability to control private nonresidential parking spaces (including those associated with office developments and major employment centers such as hospitals and universities), is far more limited.

The boom in new office development in the UK in the 1950s and 1960s exacerbated this challenge as, at the time, the provision of parking spaces, commensurate with the size of the proposed building, was required to obtain planning permission for the development. For reasons of spatial efficiency and cost, this parking was often provided underground in the basement of the building or, less commonly, on the roof. The perceived wisdom was that if cars could be removed from the public roadway then additional road space would be available for through traffic. However, congestion, directly caused by rising levels of commuter traffic, increased. One of the responses to this was to limit the number of parking spaces permitted at new build offices from the 1970s onwards. However, a large- volume of private nonresidential parking spaces remained over which the local authority had little, if any, control once they had been created.

The purpose of this chapter is to review the factors which led to the formation of legislation to enable the introduction of a Workplace Parking Levy (WPL), discuss the implementation of the scheme in Nottingham and provide insights into the lessons that can be learned from the Nottingham experience.

Table 1 Example TDM measures

Type of measure	Examples
Economic	Road user charges Charges for on-street and off-street parking Subsidization of public transport
Regulatory	Bus Lanes Controls on parking Pedestrianized zones
Travel information	Real time passenger information Public transport website provision
Land use policy	Land use and transportation strategy Park and ride facilities
Travel substitution	Home working Video conferencing

Source: Based on Ison and Rye (2008)

Background

The legislation relating to the WPL can be traced back to the [UK Government White Paper “New Deal for Transport: Better for Everyone” \(1998\)](#). This document advanced the notion that local authorities could introduce a road user charge or a workplace parking levy (or both) to address problems resulting from urban traffic congestion and poor local air quality providing that the revenue raised was hypothecated for improvements in the local transport network. The 1998 White Paper led to legislation in the form of the Transport Act (2000) which detailed the provisions and scope of the WPL and granted powers to local authorities to implement and enforce a WPL (and road user charge) if they so desired. The [Workplace Parking Levy \(England\) Regulations, 2009](#) No 2085 added to the existing legislation with respect to the issuing of penalty charges for noncompliance and the management of this process.

In the aftermath of the Transport Act (2000) 35 local authorities in the UK were interested in introducing either the WPL or a road user charge. However, only one major city in the UK, namely Nottingham, ultimately sought to address the issue of congestion and environmental externalities through introducing a WPL. This imposed a charge on companies who provide more than 10 private parking spaces at the workplace. 20 years later and following Nottingham’s lead, other UK cities including Bristol, Cambridge, Edinburgh, Glasgow, Leicester and Reading plus the London Boroughs of Camden, and Hounslow are reportedly considering introducing similar schemes.

Other than the Nottingham scheme there are a number of schemes, most notably in Australia, of a similar type. [Table 2](#) provides a comparison of the schemes operating in Perth, Sydney, Melbourne, and Nottingham. The Nottingham scheme focuses solely on workplace employer provided parking. The aim of all four schemes is to address the issue of traffic congestion be it as a TDM pricing tool and/or as a means of raising revenue in order to invest in improving public transport infrastructure ([Dale et al., 2017b](#)). [Legorreta and Newmark \(2015\)](#) undertook a review of worldwide parking space levies (PSL). Of the 11 schemes they examined it

Table 2 Parking levy schemes

City	Area covered	General Description	Liable for a charge				Introduced
			On Street Parking	Public Car Parks	Unoccupied Spaces	Small Business	
Perth Parking License Fee	Central Business District (CBD)	All nonresidential parking bays that are in use	YES	YES	NO	NO	1999
Sydney Parking Space Levy (PSL)	CBD + five other outlying business areas	Off street private nonresidential parking, occupied or un- occupied, does not apply to public car parks.	NO	NO	YES	YES	1992
Melbourne Congestion Levy	CBD	All public and private long stay nonresidential car parking spaces currently in use	NO	YES	NO	YES	2006
Nottingham Workplace Parking Levy	City of Nottingham	Occupied private nonresidential off-street workplace parking	NO	NO	NO	NO	2011

Source: Adapted from [Dale et al. \(2017b\)](#)

would appear that only Perth, Sydney, and Melbourne impose a scheme that resembles that of the Nottingham Scheme. As can be seen in [Table 2](#) however, unlike Nottingham it would seem that the Australian schemes include other types of off-street parking spaces. In fact, the Nottingham City Scheme is the only one which is focused exclusively on workplace parking provided by employers. All the schemes primarily focus on addressing traffic congestion be that as a price-based tool or via the revenue raised being used for public transport improvements. Whilst there are cultural, legislative, and geographical differences between Nottingham and the Australian cities, the Nottingham Scheme is similar in respects reflecting elements of policy transfer in developing the WPL in the UK.

The Nottingham WPL

In 2007, the City Council in charge of transport planning and provision in Nottingham, a city of almost 330,000 people in the English East Midlands, stated its intention to introduce a WPL on car parking spaces that are provided by major employers within the city boundary ([Bishop, 2018](#); [Dale et al., 2014](#)). An employer would become liable for the charge if they provided more than 10 parking spaces for use by employees, students/pupils, or regular business visitors. To take account of the fact that one employer's car parking spaces could be located at different sites around the city, the total number of spaces an employer has is considered to be the sum of all liable parking spaces that the employer provides within the Nottingham City Unitary Authority area. Any employer with 10 or fewer spaces is given a 100% discount meaning there is no charge as are emergency services employers (namely Ambulance, Police, Fire and, National Crime Agency), qualifying NHS premises and any space that is used by a disabled motorist displaying a valid Blue Badge. Parking spaces that are used by customers (for example, shoppers visiting retail parks or retail premises) are not defined as Workplace Parking Places in the enabling legislation and are therefore also exempt.

Nottingham City Council began issuing licenses on 1st October 2011, six months before the WPL scheme commenced. The WPL aimed to:

1. Reduce congestion in the morning and evening peak period (which was estimated to cost the city economy £160 million every year in the morning peak period alone) by encouraging employers to reduce the number of parking spaces they provide and limiting the number of parking destination points for employees;
2. Use the hypothecated revenue to fund major public transport improvements, including two additional tram lines linking the suburbs of Clifton and Beeston/Chilwell with the city center, redeveloping the city's mainline railway station by refurbishing its concourse and platforms; and investing in local bus services;
3. Help businesses adopt workplace travel plans and develop 'smarter travel choices';
4. Enhance Nottingham's attractiveness as a location for conducting and investing in business;
5. Be managed in such a way as to not impose any displaced parking issues for communities located immediately outside the WPL charging zone.

Overall, the intention of the WPL was to stimulate regeneration of Nottingham City, provide environmental improvements, support business in developing their travel plans and improving land use, in addition to impacting directly and indirectly on traffic congestion ([Burchell et al., 2015](#)). The WPL was introduced from April 1st 2012 and, from an initial capital outlay of £4 million for development and implementation, generated over £44 million in revenue in the first 5 years of its operation ([Bishop, 2018](#)). The main reason the City Council in Nottingham adopted a WPL as opposed to a road user charging scheme was the speed with which income could be generated for public transport improvements and the relatively low cost of implementation. The WPL scheme and the associated improvements to public transport can be identified as the "WPL package" with each element complementing the others to provide an overall enhanced TDM provision.

Public Acceptance of the WPL

Public acceptance of the proposal was identified early in the development phase of the scheme as representing a potentially significant barrier to implementation ([Frost and Ison, 2009](#)). The scheme was initially opposed by Nottingham-based businesses and sections of the local media for imposing additional costs and bureaucracy on businesses which may damage the city's economy, for being unfair on private motorists who already paid road tax, fuel duty, and insurance for their vehicles, and for the fact that some claimed it would have little impact on congestion.

In recognition of the need to gain widespread public acceptance of the scheme, Nottingham City Council engaged in an extensive public consultation process and sought to promote the health and productivity benefits of better public transport provision. This consultation period involved numerous meetings and focus groups with business leaders and firms within the city as well as commuters, residents, and neighboring local authorities. Extensive discussions about the practicalities of the proposal, its administration, management, costs, and possible charging regimes were conducted. By July 2007, the WPL proposal was sufficiently well developed to permit the Council to hold a week-long nonstatutory "public examination" of the proposals which was led by an independent chairperson and received contributions from 28 stakeholders ([Dodd, 2007](#)). The resulting report, "The Proposed Nottingham WPL, Report of the Public Examination" offered various recommendations, including the need for ongoing business support to offset the cost of the WPL to employers.

The Structure of the WPL Scheme

Every owner or occupier of premises with 11 or more workplace parking spaces within the city boundary has to apply to the traffic authority for a license. Over 3000 smaller businesses, each with fewer than 10 spaces within the city boundary, were exempted but over 500 larger employers were required to apply for licenses. The WPL enforced a principle of maximum vehicle occupancy not exceeding the license. An implementation charge of £288 per space per year was initially levied (a minimum of £3168 per liable firm) with enforcement enacted by a dedicated team of 5 WPL Officers with penalties for noncompliance. Subsequent annual increases in the charge were linked to inflation and the charge for each liable space stood at £424 for the year 2020–21.

By 2017, 5 years after it was implemented; the WPL was raising £9 million a year (at an operating cost of £500,000) and achieving 99% compliance. It was reported that half of liable firms absorbed the charge directly whilst the rest passed it onto their employees (Clayton et al., 2017). The monies raised were invested in local public transport improvements. As a consequence, Nottingham witnessed considerable mode shift from private cars to more sustainable public transport modes. Other impacts have been noted in terms of improvements to the urban environment and public realm as land which was previously earmarked for parking has been converted into alternative residential and commercial use.

Impact of the Nottingham WPL

While the introduction of a WPL does not necessarily result in a reduction in congestion, in the case of Nottingham, Dale et al. (2017b) state that the number of liable workplace parking spaces is positively related to the level of delay. It is, however, somewhat difficult to identify the overall effect of such a scheme given that such schemes tend to be implemented as part of a package of measures, with the revenue ring fenced for the provision of a package of TDM measures. In this respect the Nottingham WPL is no different. Dale et al. (2017a) conclude that “there is no observable negative effect on overall macroeconomic performance associated with the introduction of the WPL”. Overall, the WPL would appear to be of relatively small consideration as part of a business investment decision.

Limitations of the WPL and Lessons Learned

One key issue with any WPL is the equity of the charge. One limitation of the present scheme is that it does not (and cannot) distinguish between commuters who drive within the city boundary in peak periods and those who do not. In addition, it cannot differentiate between commuters who have a suitable viable public transport alternative to the private car and those who do not.

In recognition of this, it could be argued that a conventional road user charge, which levies higher fees on drivers for driving in congested peak periods, could be perceived as being a more equitable way of tackling congestion and poor local air quality. The introduction of real time monitoring and “smart” vehicle density and speed responsive traffic management systems could offer a potential way to address the WPL’s limitations.

Conclusion

In terms of the UK, Nottingham was the first city to introduce a WPL. Its introduction provides important lessons for those authorities thinking of implementing such a scheme, most notably in terms of public acceptance. This is in fact the case with all market-based approaches to addressing the issue of traffic congestion. One thing is clear namely that the increased awareness in recent years with respect to environmental issues both local and global not least in terms of air quality, and climate change, plus the financial constraint public authorities are working under, means that the potential for implementing a WPL will continue to be on the political agenda of many cities. Evidence would suggest that the introduction of the WPL has led to a reduction in the level of congestion over the period 2011/12 and intuitively one would expect this to be the case namely as a result of (1) increasing the cost of commuting to work, if the employer passes on the levy to employees, (2) the number of car parking spaces being reduced, hence the number of termination points, and (3) the enhanced public transport provision. In addition, the enhanced public transport network can be seen to improve the environment within which the business community operates offsetting what can be seen as the cost of the WPL.

References

- Bishop, D., 2018. *Idea Exchange: How Nottingham created the UK's first workplace parking levy*, [online]. Local Government Chronicle (accessed 28.12.2018) <https://www.lgcplus.com/idea-exchange/how-nottingham-created-the-uks-first-workplaceparking-levy/7022862.article>.
- Burchell, J., Ison, S.G., Enoch, M.E., 2015. The smeed report fifty years on: A role for the Workplace Parking Levy? *Transport. Plan. Technol.* 38 (1), 62–77.
- Burchell, J., Ison, S.G., Enoch, M., Budd, L., 2019. Assessing the likely take up of the Workplace Parking Levy as a Transport Policy Instrument. *J. Transport Geogr.* 80, 102543.
- Clayton N., Jeffrey, S., Breach, A., 2017, Funding and financing inclusive growth in cities How cities can unlock more investment for inclusive growth policies. Centre for Cities Available online at <https://www.centreforcities.org/publication/funding-and-financing-inclusive-growth-in-cities/>.

- Dale, S., Frost, M., Gooding, J., Ison, S.G., Warren, P., 2014. In: A Case Study of the Introduction of a Workplace Parking Levy in Nottingham chapter 15 Parking: Issues and Policies. Emerald Publishing Limited, Bingley, pp. 335–360.
- Dale, S., Frost, M., Ison, S.G., Nettleship, K., Warren, P., 2017a. An evaluation of the economic and business investment impact of an integrated package of public transport improvements funded by a Workplace Parking Levy. *Transport. Res. A* 101, 149–162.
- Dale, S., Frost, M., Ison, S.G., Quddus, M.A., Warren, P., 2017b. Evaluating the impact of a workplace parking levy on local traffic congestion: The Case of Nottingham, UK. *Transport Policy* 59, 153–164.
- Department of the Environment, Transport and the Regions, 1998. A New Deal for Transport: Better for Everyone, The Government's White paper on the Future of Transport, Cm 3950. The Stationery Office, London.
- Dodd, B. 2007. The Proposed Nottingham Workplace Parking Levy, Report of the Public Examination. The UK Planning Inspectorate.
- Frost, M.W., Ison, S.G., 2009. Implementation of a Workplace Parking Levy, Lessons from the UK, 88th Annual Meeting of the Transportation Research Board, TRB, 09-0249, Washington D.C, USA, 15th January 2009, pp 1–17.
- Ison, S.G., Rye, T., 2008. In: Ison, S.G., Rye, T., Ashgate, (Eds.), TDM Measures and their Implementation, in The Implementation Effectiveness of Transport Demand Management Measures: An International Perspective, Publishing Limited, Aldershot.
- Legorreta, A.M., Newmark, G.L., 2015. See Spot Taxed: The Global Experience with Parking Levies. 95th Annual Meeting of the Transportation Research Board, Washington DC, U.S.
- Shoup, D., 2005. The High Cost of Free Parking, Chicago, IL: American Planning Association Transport Act 2000, (c.38 Part III Chapter II). HMSO, London. Retrieved from <http://www.legislation.gov.uk/ukpga/2000/38/part/III/chapter/II> [Accessed 03/12/2019].
- Workplace Parking Levy (England) Regulations 2009, SI 2009 No 2085. HMSO, London. Retrieved from <http://www.legislation.gov.uk/uksi/2009/2085/contents/made> [Accessed 03/12/19].

Further Reading

- Dale, S., Ison, S.G., Frost, M., Budd, L., 2019. The impact of the Nottingham Workplace Parking Levy on travel to work mode share. *Case Stud. Transport Policy* 7, 749–760.

Air Transport

Lucy Budd, Stephen Ison, Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	7
Commercial Air Transport—Past to Present	7
Airports	8
Aircraft	8
Airlines	9
Airspace	10
Challenges	10
Future Trends	10
Conclusion	11
References	11

Introduction

Of all the principal modes of transport, aviation is the most recent as it has only existed for a little over 100 years. In that timeframe, aircraft have transformed international relations and military engagements, permitted new forms of aerial recreation, international travel, and tourism and created new ways of becoming and being (aero)mobile (see [Goetz and Budd, 2014](#)). Air transport has had a profound impact on the structure and functioning of global society and aeronautical expressions including “autopilot”, “flying high” and “taking off” have entered the cultural lexicon of everyday commerce and conversation.

Air transport describes the use of aircraft for a range of different purposes which can be categorized into commercial, military, and general aviation activities. Commercial air transport describes the use of civilian aircraft, which are available for public use, to fly people and goods between defined origin and destination airports at agreed standards of safety and service in exchange for revenue. Military air transport, in contrast, describes the use of dedicated military airframes (which are not available for public use) to transport military personnel and equipment between airbases and to/from/within theatres of operation. The third category, general aviation, covers the use of smaller and lighter aircraft for recreational, agricultural, medical, and civilian surveillance purposes. In each category of air transport activity, the aircraft that are deployed can be fixed or rotary wing. In the case of general aviation, the aircraft can be both heavier and lighter-than-air (hot air balloons). All three types of activity make use of unmanned aerial vehicles (drones) as well as other specialized aircraft which have been developed to deliver particular performance characteristics in terms of range, capacity, payload, operating costs, speed, maneuverability, aerodynamics, and environmental standards.

With the exception of consumer drones (which can be operated from any location providing the operator acts in accordance with national regulations and local restrictions), all other aircraft are flown between fixed ground-based nodes which have been specifically designed and designated for the purpose of flight. These sites are variously described as being airports (which denote commercial facilities for passengers and/or goods which feature customs and immigration inspection posts), airbases (military facilities at which squadrons of aircraft and/or personnel are based), and airfields (smaller civilian facilities for general aviation use which lack the landside and airside infrastructure of airports).

Irrespective of its size and geographic location, every airport, airbase or airfield must, as a minimum, feature a dedicated runway (or lift off area for helicopters) that has been demarcated and prepared for the use of aircraft landing and taking off. They will also feature an area adjacent to the runway on which aircraft can be parked and serviced between flights. Given the larger size and weight of commercial and military airframes vis-à-vis general aviation aircraft, the runway/s, apron areas and the taxiways which connect them are usually hard surfaces constructed from asphalt or concrete. General aviation airfields in contrast may have grass, crushed cinder or compacted earth runways and maneuvering areas. Most, though not all, will also have an air traffic control facility and tower from which air traffic movements on the ground and within the proximate airspace are monitored and controlled. As such, they are vital nodes in a global air network as they facilitate the interface between ground transport modes and aerial ones and between ground and sky. Given the social and economic importance of commercial air transport to the functioning of the modern world economy (on which see [Bowen, 2013](#)), the remainder of this article focuses on the aerial conveyance of passengers and freight on aircraft which are available for public use.

Commercial Air Transport—Past to Present

Prior to the COVID-19 pandemic, over 1400 commercial airlines transported more than 4.6 billion passengers and over 55 million tons of airfreight between 3800 airports around the world ([ATAG, 2018](#)). The scale of this mobility is far removed from the first international passenger flight which occurred between London and Paris in August 1919 and transported a single passenger. Over

the intervening 100 years, innovations in aircraft design and scientific understanding have enabled a species, which is physiologically unable to fly without recourse to mechanical intervention, to safely and routinely take to the skies.

Rapid advances in aircraft design, aerodynamics, and flight control systems have enabled the construction of progressively larger and more reliable civil aircraft. This has improved the standard of safety and passenger comfort and lowered aircraft operating costs leading to cheaper airfares. In the political and policy arena, regulatory reform from the late 1970s onwards has reconfigured and/or removed many of the restrictive bi- and multilateral air service agreements which structured the early development of air transport operations and opened up more markets to competition (Budd, 2018). This has stimulated innovation in airline business models. The emergence, expansion and evolution of the low-cost airline model which began in the US in the 1970s before being emulated in other world markets (see Budd and Ison, 2014) has transformed the patterns and processes of flight. The availability of lower airfares stimulated demand for air services between underserved airports while the low-cost carriers' "no frills" business philosophy and use of the internet to lower distribution costs has changed consumer purchasing behavior and passenger expectations and experiences of in-flight service. The low-cost airline era intensified already growing consumer demand for flight and air transport has acted both as an enabler and a driver of globalization.

Yet the intensity of current levels of aeromobility have come at a high cost to the environment and local communities whose lives are impacted by high levels of aircraft noise and pollution or the "over tourism" experienced by cities including cities such as for example Venice, Rome, and Barcelona. In order to discuss air transport's societal contributions and challenges, the remainder of this article considers commercial air transport in terms of airports, aircraft, airlines, and airspace. It ends with a discussion of the contemporary challenges and future trends that are likely to occur within the sector.

Airports

It is believed that the word *airport* first entered common use in English during the 1920s. The date and description are important for it was during this decade that the first long-haul routes to destinations in what was then the British Empire were being inaugurated and new inland "ports of the air", featuring customs and immigration inspection posts were developed.

Modern commercial airports are still characterized by their landside and airside features, a distinction that was first established in the early period of flying in the 1910s when it became necessary to separate aircraft maintenance activities from the enthusiastic crowds who flocked to early air shows. Convention dictates that the landside areas of an airport are those which occur between entering an airport and security. These areas are accessible to passengers and nonflying members of the public alike and typically feature ground access interchanges and car parks as well as publicly accessible shops and retail concessions. The airside areas after security control are characterized by duty free shops and departure lounges. The airfield beyond it (here defined as the airside aircraft maneuvering areas) features runways, taxiways, apron areas, and the expanses of grass between them. For reasons of planning, public safety, and customer service, the design of commercial airports is informed by the design parameters and considerations contained within the International Air Transport Association's (IATA) Airport Development Reference Manual (ADRM). The ADRM, which is now in its 9th edition, provides global guidance for airports, architects and developers and stipulates every aspect of airport design from the configuration of taxiways to the allocation of inter-personal space in different subsections of the passenger terminal ecosystem (IATA, 2011).

Historically, most commercial airports were state-owned enterprises which derived most of their revenue from aeronautical sources including aircraft landing fees, parking charges and passenger handling payments. However, the last 30 years or so, particularly though not exclusively in Europe, have seen moves towards greater airport privatization and commercialization. This has not only resulted in many major airports transferring into the private sector but also airport operators deriving ever greater proportions of their revenues from non-aeronautical sources (particularly car parking, retail, and consultancy services). Attracting and maintaining high levels of passenger throughput is thus key to improving economic efficiency, yet as demand begins to reach capacity, bottlenecks and delays can occur which damage an airport's reputation. The incremental development of London Heathrow airport since 1946 is a case in point, as a fourth and fifth terminal were added as demand increased, and proposals to construct a third runway continue to be debated (on which see Le Blond, 2018). Already, a growing number of major airports (including London Heathrow) are operating close to or beyond their original design capacity and demand management interventions including airport slots (see article on airport slots) have been introduced to try and deal with the mismatch between supply and demand. Any form of slot control, however, has its drawbacks and can in certain circumstances effectively prevent or limit market entry by new operators. For this reason, a proportion of slots at the most capacity constrained airports are typically reserved for new entrant operators. At the other end of the spectrum are numerous under-utilized loss-making facilities that cannot attract and retain sufficient custom year-round to make them viable financial propositions. These airports often serve smaller settlements with limited demand and may oscillate between public and private sector owners. A number of commercial airports, including Coventry, Sheffield City, Plymouth, and Blackpool airports in the UK have closed due to lack of demand and competition from neighboring airports in Birmingham, Manchester, Exeter, and Liverpool. Elsewhere, purpose built new airports at Montreal Mirabel, Canada, and Ciudad Real, Spain, did not attract anticipated levels of custom and closed.

Aircraft

Air transport is a derived demand but one that is highly cyclical in nature and one which rapidly responds to changes in global economic health and prosperity. The two principal advantages air travel has over ground based transport modes—superior speed

and the ability to overfly topographical obstacles such as mountain ranges, oceans and deserts with ease—were recognized in the early days of commercial flight and the desire for ever-increased speed and range has resulted in the development of faster aircraft (which culminated in the supersonic Concorde of the late 1970s) and the inauguration of so-called “ultra-long haul” routes which have effectively compressed time and space by enabling long distance journeys to be accomplished more quickly than ever before. As of 2019, the Australian national airline Qantas’s direct B787-9 service between London Heathrow and Perth, Western Australia, offered the shortest connection between the two cities, covering the 14,498 km distance in around 17 h. At the other end of the spectrum, the commercial service between the Scottish islands of Westray and Papa Westray, which can cover the 2.7 km distance in under 1 min, remains the world’s shortest commercial flight.

Although air transport operates on all 7 continents and airlines fly over 45,000 routes between some 3800 airports (ATAG, 2018), the global geographic distribution of flights is unequal and some locations are either only partly connected or are wholly unconnected to the global web of air traffic flows (see Bowen, 2014, 2019). The different demand characteristics and geographical constraints imposed by particular locations has resulted in the production of different types of aircraft from high-capacity long range A380 “Super Jumbos” to narrow-body workhorses including the A320 and B737 and single seat propeller driven aircraft that are capable of operating from unprepared landing strips. Most passenger aircraft can also carry freight, either in the hold or at the rear of the main passenger deck. Specialist quick change (QC) aircraft can be quickly converted from flying passengers during the day to freight at night while fleets of hundreds of dedicated cargo-only aircraft are equipped with strengthened main decks and enlarged side cargo doors (Budd and Ison, 2017).

Airlines

The commercial operators of aircraft, namely the airline companies, can be categorized according to the type of business model they adopt—whether full service, low cost, charter, regional, cargo, hybrid, or specialist—as well as the geographic size and scale of their operations—intercontinental, international, national, regional—their revenue and their customer service and safety reputation (see Table 1).

Irrespective of the business model a carrier adopts, safety is the most important consideration for if an airline is not considered to be safe then passenger demand evaporates, investor confidence disappears, and the airline will not be able to sustain its operations. In a relatively short space of time, aviation has developed from a hazardous mode of travel to the safest form of long-distance transport. Improvements in safety have been swift and welcome and new technological and procedural interventions, including more sophisticated flight controls, heightened training, awareness of the importance of a safety culture, and more rigorous accident investigations have collectively helped to further reduce the risk of flying (see Quddus, 2020).

Given that aircraft can cross international borders with ease, it was recognized that the legal implications of any accident or incident could be significant as there would be a potential conflict of laws. The problem of legal jurisdiction was recognized in the 1920s and in 1929 leading aeronautical nations convened in the Polish capital Warsaw and signed a Convention on air carrier liability. The resulting Warsaw Convention established the international legal mechanism for paying compensation to passengers in the event of injury or loss of luggage. This international convention was followed in 1944 by the Chicago Convention which set in place the framework for the post-1945 development of international commercial aviation. The Convention also established a series of Standards and Recommended Practices (SARPs) which covered issues including carrier licensing, airport facilities, air navigation, and security (ICAO, 1944).

Given the importance of air transport to global society, commercial airlines and airports vulnerable to terrorism and illegal activity. As a consequence, the aviation security regime has been progressively tightened and new screening technologies have been introduced to identify the presence of prohibited items in passengers’ hold and hand luggage as well as on their person and in consignments of cargo (see Trembaczowski-Ryder, 2020). These interventions have imposed financial costs on the sector and have led to delays and consumer dissatisfaction.

As passenger numbers have grown and airlines have sought to lower their costs and improve the efficiency of their operations, many of the functions associated with reservations, check in and flight transfers have been automated and consumers encouraged (or obliged) to use selfservice technologies including online check-in, automated bag drops and digital boarding pass readers. The rise of automated processes has had significant implications for employees and passengers (not all of whom are able to use or wish to use the automated equipment) as well as the airlines themselves who must ensure the security, integrity, and reliability of their systems.

Table 1 The principal types of commercial airline operator

Full service carriers (FSCs) are characterized by their international network, use of major airports, mixed aircraft fleet, provision of different cabin classes, frequent flyer programs and their range of inflight products. They typically have high levels of customer service and high levels of fixed and operating costs.
Low-cost carriers (LCCs) in contrast only provide what is necessary for safe and efficient flight. Costs are kept low by only operating one type of aircraft, flying to cheaper and less congested secondary or regional airports and providing lower levels of customer service.
Charter operators sell their seats to tour operators and fly to popular seasonal “sun and ski” destinations from regional airports.
Regional operators typically fly smaller aircraft on lower demand routes between smaller airports. Their cost base is higher than an LCC and the passenger demand is more modest.
Hybrid operators combine elements of FSC, LCC, Charter and regional operations.

Airspace

One of the areas of air transport that is often overlooked in conventional academic studies of the sector is the (air)space in which aircraft fly. Unlike road and rail vehicles which have to follow a defined track and ships which can only sail between ports, aircraft are theoretically free to operate in three dimensions. However, in reality, the airspace that is available to commercial air traffic is limited by the provision of military or otherwise restricted sectors of sky into which commercial aircraft cannot enter. In the years leading up to 1900, the aerial realm was not subject to any degree of ownership or control but as soon as aerial vehicles from one state began accessing the skies above another, debates about airspace ownership and control commenced in earnest.

The British Government was the first to declare that all airspace above the United Kingdom and British overseas colonies, dominions and mandates, was sovereign territory. Taking its cue from maritime law, it was agreed that in the case of coastal states, airspace sovereignty would extend for 12 nautical miles offshore and vertically upwards to a predefined altitude. This meant that foreign aircraft had to obtain permission to access another country's airspace. Until the late 1930s, all aircraft navigated visually with reference to the ground and pilots followed clear visual markers such as roads, railway lines, and canals.

As air traffic volumes grew over time, increasingly complex regulations and air traffic management systems were introduced to safely and efficiently manage the flow of air traffic (on which see [Budd, 2020](#)). The most visible manifestation of this aerial network are the air traffic control towers which act as the command and control posts of the aerial transport network. In some instances, the towers have been designed with architectural flourishes that attest to the traditions and/or aspirations of the city or nation that hosts them. Other visual manifestations of the air traffic network are high-altitude contrails created by aircraft in the cruise, circling stacks of aircraft waiting to land at capacity constrained airports and the low altitude approach and departure tracks flown by arriving and departing aircraft. A key challenge of airspace management thus now concerns the environmental impacts of aircraft noise and emissions.

Challenges

As a consequence of burning nonrenewable fossil fuel in the form of kerosene in their engines, the world's commercial aircraft fleet (which now numbers almost 32,000 units) generate in excess of 600 million tons of carbon dioxide on international flights as well as large quantities of climate warming water vapor, nitrous oxides, and particulate matter (see [ICAO, 2019](#)). At a local level, aircraft operations generate noise and degrade local air quality around airports, creating odors and emissions which impact on human health and wellbeing. Although individual aircraft are quieter and more fuel efficient than previous generations, technological improvements are not keeping pace with growing consumer demand. Despite innovations in sustainable aviation fuel, including biofuels derived from biological feedstocks, advanced biofuels derived from residues and waste, and electrofuels (also called power-to-liquid or synthetic fuels) which are produced by capturing carbon dioxide from the atmosphere, the high cost and lack of production capacity mean that attempts to decarbonize aviation through fuel innovation are currently limited.

In recognition of this, market-based measures including the EU emissions trading system and ICAO's CORSIA scheme (which aims to limit future aviation CO₂ emissions to 2020 levels), have been introduced as a way of making the polluter pay for their cost of their carbon emissions. Such interventions are, however, unlikely to have a major impact on their own and more dramatic demand management measures may be required. Evidence suggests that growing numbers of consumers are voluntarily choosing to limit the number of flights they undertake, actively seek out more sustainable ground-based public transport modes such as high-speed rail, or cease flying altogether (see [Higham and Font, 2020](#)).

Future Trends

Predicting the future of a sector as dynamic as commercial air transport is fraught with difficulty. Nevertheless, some trends are becoming evident and are likely to continue into the near future. Continued regulatory reform is leading to greater freedom of market entry and exit. Market deregulation is characterized by a wave of market entry as new airlines seek to build a presence in the market. This rapid growth phase is quickly followed by a period of consolidation as airlines are bought up, merge with rivals, or leave the market. This process is leading to the emergence of a relatively small number of carriers who are dominant in their respective markets and who have developed a particular operating niche. The global airport sector too is also undergoing a period of transformation. Commercialization is driving development of nonaeronautical revenue streams while privatization is leading to greater levels of investment by overseas airport operators, foreign pension plans and overseas sovereign wealth funds (see [Graham, 2018](#)).

As far as aircraft technology is concerned, it would appear that passenger demand is driving the development of longer-range and more fuel efficient twin-engine widebody aircraft including the Airbus 350XWB and Boeing 787. The closure of the A380 assembly line, with final deliveries scheduled for 2021 ([Topham, 2019](#)), indicates at the likely direction of the market for aircraft. At the other end of the spectrum, drones and unmanned aerial vehicles are being explored as potential forms of future urban air mobility and such machines have already been used to test deliveries of medicines and small packages. Although talk of the widespread future use of urban air taxis seating one or two passengers is at present arguably premature, but if the cost and safety case can be made then there is every possibility that some form of aerial taxi service may be introduced in time. More far-fetched, perhaps, is the possibility of commercial space travel and the development of vertiports (airports for vertical take-offs) for reusable space vehicles.

Conclusion

For the passenger, the journey of the future is likely to be characterized by even greater levels of automation and customization. Indeed, the impact of the COVID-19 pandemic notwithstanding, the drive for greater personalization and customization of the travel experience (driven by more sophisticated processing power and big data) continues. At the same time, growing consumer awareness of the environmental impact of flying may alter individual consumer's purchasing and travel behavior and, in the absence of a step change in technology, some form of transport demand management (TDM) incorporating the pricing mechanism may be required to suppress consumer demand to more environmentally sustainable levels. This will have significant implications for future air transport policy and planning as the need to "right fit" airport infrastructure and airline business models to a rapidly changing commercial environment will become more acute.

References

- ATAG, 2018. Aviation Benefits Beyond Borders. Geneva, Air Transport Action Group. Available online at <https://www.atag.org/our-publications/latest-publications.html>.
- Bowen, J.T., 2013. *The Economic Geography of Air Transportation: Space, Time and Freedom of the Sky*. Routledge, Abingdon.
- Bowen, J.T., 2019. *Low-Cost Carriers in Emerging Countries*. Elsevier, Amsterdam.
- Budd, L., 2018. Airline deregulation. In: Cowie, J., Ison, S. (Eds.), *The Routledge Handbook of Transport Economics*. pp. 141–154.
- Budd, L., 2020. Airspace and air traffic management. In: Budd, L., Ison, S. (Eds.), *Air Transport Management: An International Perspective*. 2nd ed. Routledge, Abingdon.
- Budd, L., Ison, S., 2014. *Low Cost Carriers: Emergence, Evolution and Expansion*. Ashgate, Farnham.
- Budd, L., Ison, S., 2017. The role of dedicated freighter aircraft in the provision of global airfreight services. *J. Air Transport Manage.* 61, 34–40.
- Goetz, A., Budd, L. (Eds.), 2014. *The Geographies of Air Transport*. Ashgate, Farnham.
- Graham, A., 2018. *Managing Airports: An International Perspective*, 5th ed. Routledge, Abingdon.
- Higham, J., Font, X., 2020. Decarbonising academia: confronting our climate hypocrisy. *J. Sustain. Tourism* 28 (1), 1–9.
- IATA, 2011. *Airport Development Reference Manual (ADRM)*. 9th ed. IATA, Montreal.
- ICAO, 1944. *Convention on International Civil Aviation - Doc 7300*. ICAO, Montreal.
- ICAO, 2019. *Destination Green: The Next Chapter 2019 Environmental Report, Aviation and the Environment*. Available at <https://www.icao.int/environmental-protection/Pages/envrep2019.aspx>.
- Le Blond, P., 2018. *Inside London's Airports Policy: Indecision, Decision and Counter-Decision*. ICE Publishing, London.
- Quddus M., 2020. Aviation Safety. In: Budd, L., Ison, S. (Eds.), *Air Transport Management An International Perspective*. 2nd ed. Routledge, Abingdon.
- Topham, G., 2019. A380: Airbus to stop making superjumbo as orders dry up *The Guardian* 14/02/2019 Retrieved from <https://www.theguardian.com/business/2019/feb/14/a380-airbus-to-end-production-of-superjumbo> on 26/10/2020.
- Trembaczowski-Ryder, D., 2020. Aviation security. In: Budd, L., Ison, S. (Eds.), *Air Transport Management: An International Perspective*. 2nd ed. Routledge, Abingdon.

Mobility as a Service

Miloš N. Mladenović, Otakaari 4, Espoo, Finland

© 2021 Elsevier Ltd. All rights reserved.

Emerging Technology as a Lens for understanding MaaS	12
MaaS as Evolution of Transport System Integration	12
Anticipated Positive Implications for the Current MaaS Vision	13
Socio-Technical Convergence around the Current MaaS Vision	14
Networked Multi-Actor MaaS Innovation Processes	14
Unanticipated Negative Implications from the Current MaaS Vision	15
Toward Responsible MaaS Innovation and Institutional Transition	16
Acknowledgment	16
See Also	17
Biography	17
References	17
Further Reading	17

Emerging Technology as a Lens for understanding MaaS

Mobility as a Service (MaaS) is as a powerful and value-laden rhetorical vision of complex socio-technical system for mobility transition that is currently redrawing responsibility and innovation boundaries across societal sectors. Over the following sections, we will attempt to define MaaS through the lens of emerging technology (ET), which as opposed to a variety of societally embedded technologies, has a set of specific properties listed below (Collingridge, 1980; Jasanoff, 2016; Rotolo et al., 2015).

- During the foundational stage, ET is open to further development, having associated uncertainties, prominent (re)distributive impacts, fast development pace, and coherence of the assemblage.
- ET development faces a dilemma, as on the one hand, we are unable to estimate the changes from ET until it is fully formed and embedded in society, while on the other hand, changing a technological development trajectory is very difficult once the technology is fully formed.
- In reflection about desired ET outcomes in the future, we tend to discount harms as more speculative, rather than equally uncertain benefits.
- ET is coconstructed through (ideological) visions and rhetoric, usually only by the few, simultaneously framing the societal challenge and technological solution, with intention to persuade, as well as align the activities of different actors and, crucially, attract funding and publicity.
- ET has interpretative flexibility, where technology working is the result and not the cause of it becoming a successful artifact, and the success of a technology is often assessed through the lens of particular social groups, even if the use is intended for a wider society.
- ET development is often happening through experimentation with the actual artifacts, helping in building a transition pathway for ET from niche to regime.
- Partially influenced by path dependence, ET development trajectory is often significantly influenced by previous lock-in decisions around synergetic technologies, while it generally develops through nonlinear pathways, aided by convergence and assemblage with a multitude of socio-technical aspects.
- Inter- and intraorganizational learning processes depend on strong and loose social networks with multiple types of actors, coconstructing understanding, vision, and evaluation of the performance of the innovation as well as the needs for improvement.
- ET often challenges institutional landscape, structures, and patterns of interaction among actors in unanticipated ways, resulting in redistribution of roles, responsibilities, and power, while often facing an institutional void as well as distributed (ir)responsibility in hybrid institutional networks.

MaaS as Evolution of Transport System Integration

Even if many of the same or similar subconcepts have existed before in the mobility domain, the discourse around MaaS was first introduced in 2013–2014, based on the Go:smart project in Gothenburg, Sweden, and Sonja Heikkilä's masters thesis at Aalto University, Finland (Heikkilä, 2014). That thesis defined MaaS as "a system, in which a comprehensive range of mobility services are provided to customers by mobility operators." Here, we can already see that framing MaaS depends on integration, an important historical concept for user centrality in mobility systems. In fact, the pursuit of increased integration in mobility systems by relying on complementarity of different transport modes has been going on for decades, previously also driven by

Intelligent Transport Systems discourse and deployment. With the socio-technical changes in the recent years (see “Socio-technical Convergence around the Current MaaS Vision” Section), the integrative vision of MaaS relies not only on infrastructural and operational integration but also on informational and transactional integration of mobility services. The user is expected to receive information, book, and pay for a choice of different mobility services by accessing such “one-stop-shop” or “mobility platform” via mobile phone applications or other digital interfaces. Avoiding a binary understanding of MaaS implementation, Lyons et al. (2019) define six levels of MaaS integration, in addition to Sochor et al. (2018), who also mention policy integration:

1. No integration: no operational, informational, or transactional integration across modes
2. Basic integration: informational integration across some modes
3. Limited integration: informational integration across some modes with some operational and/or transactional integration
4. Partial integration: some journeys have full informational, transactional, and operational integration
5. Full integration under certain conditions: some but not all available modal combinations offer a fully integrated experience
6. Full integration under all conditions: full operational, informational, and transactional integration across all modes and all journeys

At the center of MaaS vision is not only the concept of integration but also a dialectic with the contrasting concepts of personalization and customization. Besides the integrated choice of different transport options, MaaS vision also includes an increasing reliance on new transport modes, vehicle sharing, on-demand ordering, and other customization aspects. As opposed to “pay as you go” approach, which could include dynamic pricing, an emerging alternative are bundles/packages including a range of transport modes that users can subscribe to. With sunk cost of paying up front to save later, as with mobile phone contracts, the packaging idea is that such pricing alternative would have lower unit costs and play a role in behavior change. As mentioned in the Heikkilä’s definition, MaaS architecture would have to rely on a mobility intermediary (i.e., operator or broker) layer, in addition to operational/infrastructural (e.g., vehicles, terminals, right-of-way), transactional (e.g., booking, payment, ticketing), and informational (e.g., journey planning, en-route info) layers of interoperability. The initial idea of the mobility intermediary is that such company would be responsible for mobility service distribution, by brokering between users and service producers from whom it buys services on behalf of current and future users (Hensher, 2017). In principle, there could be multiple mobility intermediaries in the same open market that could be established beyond existing administrative borders, similar to mobile phone roaming, and positioned as “optimizers” from both end-user and transport system perspective.

Anticipated Positive Implications for the Current MaaS Vision

Overall, MaaS vision includes an assemblage of interdependent promises to citizens, cities, as well as the private sector, centered on the transition toward a more sustainable “postcar” mobility system. With its interpretative flexibility, this assemblage of promises enables fluidity of MaaS conceptualizations for interpretation depending on its context. One could say that MaaS promises everything, for everyone, by sounding intuitive, understandable, and attractive. From the user perspective, MaaS is supposed to provide a convenient alternative to the private car, relying upon availability of well-integrated transport modes, useful and usable information, and easier payment and billing transactions. Such a shift from ownership toward access to an attractive combination of transport modes includes reduced cognitive efforts for users, enabling convenient door-to-door trip-chains with transfers between modes, as well as departure, mode choice and routing suggestions based on personal preferences (e.g., safety, time constraints, different capabilities). Further customization of mobility services for different travel needs (incl. persons with reduced mobility) would go hand in hand with changing preconceptions about transport modes through positive experiences, and related long-term change in mobility habits and societal norms (e.g., acceptance of electric vehicles).

The expectation of seamless multimodal travel across modes, time and space, combining different trip categories currently in existence, is expected to lead to increased efficiency of the entire transport system. Simultaneously, the expectation is that transport planning and policy institutions will also undergo a transition toward more experimental and agile processes, reconsidering the role of incentives for transition in mobility habits as boundaries of different transport modes eventually fade away. Moreover, the pursuit of system efficiency is assumed to be easier if underpinned with data bonanza about user preferences and behavior. Some of these expectations about system performance could be general across the globe, but could also have particular instances (e.g., increasing the use of mass transit). In addition to the overall improvement in the distribution of accessibility, current MaaS vision is anticipated to have positive implications on environmental sustainability by encouraging higher use of noncar and shared-vehicle modes. Similarly, reduction in private car ownership and use would result in better utilization of land use, such as reduced parking areas, as well as indirect effects on safety and liveability of urban streets, which also positively contributes to city image. Furthermore, the interdependent set of promises focuses on creating a digitally based business opportunity for the private sector, by opening new market domains and being a testing ground for new service models aspiring for international markets. Ultimately, for some, the current MaaS vision might be even a concept beyond mobility, to include booking payments for other everyday activities (e.g., through personal calendar integration), thus aiding to a truly society-wide disruption.

Socio-Technical Convergence around the Current MaaS Vision

Beyond the above aspects of the current vision, MaaS is at the convergence among several socio-technological trajectories, involving a multitude of factors in dynamic interaction. The backbone of technological convergence are long-term technical developments and standardizations, such as development of information and payment interfaces, opening of data (e.g., routes, timetables), interoperability of back-end systems and application programming interfaces (APIs), development of new vehicle features, operational design for near-real-time demand-responsiveness and trip combining for some modes or defined areas (Haglund et al., 2019), as well as strategic changes in transport network structures (e.g., trunk-feeder) and associated land use (Weckström et al., 2019). In the background of these technical developments are general technological trends in sensing, communication, and processing technology, Internet and smartphones, unlocking of data access and improving data quality, providing real-time information, and more secure booking and payment systems. Ultimately, these technical developments are continuing the blurring of boundaries between alternative transport modes.

Besides the technical developments, there is a range of societal trends and evolving ways of life in and beyond the mobility domain, coevolving together with the MaaS vision. Several of these trends are specific for urban environments, such as the population increase, reduction or delay in car ownership, and changing patterns of travel both spatially and related to trip purpose or mode (e.g., reduced number of shopping trips due to online purchases, fewer trips by car, more cycling trips). These urban trends are accompanied with wider societal trends such as an increase in smartphone ownership and digital literacy, political importance of the climate crisis (contrasted with lagging transport decarbonization), uncertainty of long-term employment and income, trust in digital transactions, and proliferation of the so-called sharing economy. The sharing economy aspect is particularly important for the arguments around suboptimal vehicle and urban asset utilization, underlined with social movements for peer-to-peer asset sharing and crowd-sourcing of data, exemplified with higher availability, convenience and affordability of ride hailing services. These transitions are occurring in parallel with more general shifts in urban and regional governance, including more public-private collaborative initiatives; redefining role of the nation-state and strengthening of lower levels of self-government; increasing diversity, variation, and even asymmetry of governance; increasing marketization of the public domain; and shifting rationales for intervention. In particular, the transport sector is experiencing higher emphasis on servitization and more flexible regulation, aiming for promoting openness that should provide testing grounds for new business models. Finally, wider ranges of decision makers and market actors are highly favorable of technological smartness and innovation policy focus on transport technologies, as evident in relation to connected and automated vehicles.

Networked Multi-Actor MaaS Innovation Processes

MaaS is not solely an app or the value concept of service packages, or even the revenue streams that are defining the emerging business model (e.g., current MaaS companies rely on large budgetary losses), but it is a set of organizations, legislation, and other aspects, which collectively serve to lock a technology into society. Thus, the multitude of factors in dynamic interaction outlined in the previous sections develop in interactions between a multitude of networked actors (Wong et al., 2020), engaged in semistructured interorganizational (un)learning. Some of these actors are traditional transport sector actors (e.g., transport operators, data providers, technical solutions providers, national and EU authorities, consulting companies, unions, marketing companies), while some previously fringe actors are changing their roles (e.g., automotive, oil, insurance, or tourism companies), and new actors are also entering the sector (e.g., telecommunication, ICT and big tech companies, retail companies, real-estate developers). Simultaneously, the role of users is changing toward a higher degree of presumption, while new public-private partnerships are being established (e.g., MaaS Alliance in Europe). As previous business models are being challenged, there is an ongoing attempt to shift away from being a supplier of technology to being a provider of mobility services, following the logic of financial viability and profitability for mobile phone markets. Certainly, academic communities have their role in the ecosystem too, exemplified with the fact that several Transportation Research Board committees have mentioned MaaS in their recent centennial papers. These changes in actor constellations are also underlined with expectations from the public sector to be enabler of change and provider of favorable conditions, while responsibility for service development is supposed to be with the private sector. These changes are contrasted with a premise of an open ecosystem, where all transport operators are supposed to offer the same pricing schemes, data, and APIs access to all willing mobility intermediaries. In turn, mobility intermediaries are supposed to be an organizer that cooperates, uses data, and manages the ecosystem.

Ongoing changes in actor constellations and their associated roles do not only mean entrance of new actors, but also redistribution of power and responsibility among the multiactor network, where some organizations are even supposed to disappear, along with associated jobs. In contrast to the MaaS vision, the reality of ecosystem formation in the field remains different. For example, current transport regulation remains highly mode specific and operator-focused rather than user-focused. Similarly, there is a very limited number of MaaS pilots around the world, often with limited scale implementation. Despite the assumption of cooperative and interconnected ecosystems, bringing together organizations providing the means of travel, cooperation is not necessarily advancing in practice (e.g., reluctance of transit operators to engage with MaaS companies). Such lack of cooperation is underpinned with an institutional void and organized irresponsibility that emerging technologies often face. This means that none of the current

institutions has a full understanding or control of undesirable consequences associated with technological constellations. In turn, the resulting distributed responsibility for transition management and technological development limits individual and institutional accountability. An example of such distributed irresponsibility is a full commercial model in small and disconnected markets where everybody agrees with everybody, without there being a single organization with a mandate and interest to take a role of promoting interoperability and cooperation.

Effective cooperation among such established actors with a variety of institutional cultures and often-diverging objectives faces traditional challenges around resistance and inertia of multi-actor innovation processes. Incumbent regime change through disruption relies on the expectation of changes to governance, regulation, and procurement procedures of transport system supply-side developments, as well as changing demand-side developments in terms of the preferences and behaviors of individuals and organizations. In addition to the questions of service procurement rules and revenue flows, cooperation depends on agreements on adequate service pricing, structures of funding programs, policy coordination, as well as possible reduction in some responsibilities (e.g., advertising). Research from Sweden has been proposing a specific solution for the institutional void challenge, by adding a layer of mobility integrators between MaaS operator and transport service provider, who would be responsible for technical, business, and process facilitation, acting as impartial clearinghouse platforms (Smith et al., 2018). Nonetheless, successful cooperation still highly depends on negotiation processes, where trust building and transparency of decision processes remain central ingredients for deliberating between specific wishes and unified solutions.

Unanticipated Negative Implications from the Current MaaS Vision

In contrast to many anticipated positive implications from the current MaaS vision, there is a potential danger that if MaaS comes to fruition, the only fulfilled promise is the one for private sector actors (Pangbourne et al., 2020). At its core, there is challenge with the simultaneous promise of freedom for the users and efficiency on the system level in the context of simultaneous demand in transport networks with finite capacity. Even if the idea of user centrality sounds intuitively positive, there are many currently unanticipated but possible negative implications for users themselves. For example, users will be responsible to judge if MaaS offers provide value for money, which might be more cognitively challenging if the number of choices increases significantly while aiming for customization. These cognitive aspects might be further underlined with the choices around routing, accounts based on social networks, feedback provided on mobility plan execution, or service add ons based on usage trends. Similarly, there could be perverse incentives (e.g., loss aversion) leading people to generate additional discretionary trips to match the package size and avoid the feeling of regret from not maximizing usage. Moreover, discrepancies between travel expectations and experienced level of service might be negatively affecting meanings of trip activities, and other everyday practices beyond mobility, especially if people start seeing through the intentional nudging of their behavior. Thus, there might be a dangerous redefinition of (mobility) commons that could degrade norms of civility. Such changes in expectations can also relate to changes in daily activity spaces, and norms around walking and cycling. For example, given the fact that individuals should undertake a minimum level of daily physical activity to maintain their well-being, current MaaS vision has little reflection about potential replacement with sedentary, shared vehicle modes.

Beyond the above effects on the users themselves, there could be wider questions of equity and technological gentrification. From the current examples, we can see that MaaS concepts are mainly focusing on commuting and business trips, not accounting for existing obligations in transport legislation, such as special user groups. If MaaS becomes a sole access point to geographically bounded transport networks, it is important to ask what happens to those excluded from that newly enclosed system. Such exclusion can be due to affordability, technological availability, technological aversion, or even dissent. For example, MaaS's reliance on registration and use of digital accounts might further exclude social groups experiencing difficulties in handling new technologies or having access to banking, especially older groups or those in the Global South, associated with poverty or lower education level. Furthermore, tying all mobility services to a specific pricing scheme could lead to further transference of existing monetary inequalities, causing further distributive injustices. Finally, an essential set of negative implications revolves around the traditional equity questions of digital technologies, such as cybersecurity, privacy (Cottrill, 2020), and explainability, especially if underlined with aspects of data monetization and biased profiling (e.g., strong identity verification needed for service customization).

An additional aspect to underline is that current visions of MaaS establish user relations mostly on the individual level, failing to acknowledge that current problems are large-scale emergent phenomena arising from the aggregate of our individual activities and preferences. Thus, emergence of systemic effects could also include various undesired consequences, especially in relation to pollution and energy consumption. In fact, it is plausible that a reduction in personal car ownership could be accompanied with an increase in the on-demand vehicle fleet or vehicle kilometers traveled, in order to meet demand. Here, we quickly run into an old question of market monopolies in multiactor systems. With the loss of institutional capacity coupled with resource constraints in urban governance, it might be tempting for many cities to let MaaS "take care of it all," resulting in divestment and erosion of public transport services, especially in smaller markets. Going back to resulting user travel experiences in such context, there is a related question of legal liability and consumer protection, and who is responsible to provide redress if problems arise. Furthermore, we must particularly note the reliance of MaaS on the new ability to generate and utilize Big Data, and thus establish power position

through data enclosure and monetization schemes. Essentially, MaaS has the potential to create a new market by selling data analysis to many different actors, including other private companies, such as retailers. Thus, geographic MaaS monopolies might have to be guarded against, as the first-mover MaaS providers could block new ones from entering the marketplace through developing new entrance exclusion mechanisms related to data ownership or exclusive actor relationships. In this situation, in the case of lack of alternatives, single MaaS operators could constantly raise mobility prices to the end users, with little recourse from those users, and with the possibility of the raised income being taken as profit by the MaaS brokers rather than distributed among the transport operators.

Above challenges could be further mutated by a significant legal uncertainty around contracting, overambitious implementation schedules, missing intraorganizational procedures (e.g., on behalf purchasing), and significant investment efforts for API interoperability without sharing efforts. In this context of organizational change and consequent power and resource redistribution, there are associated risks for implementation, especially related to procurement of services, and impacts on labor market. Beyond these systemic market effects, there could be negative implications from power imbalances on degrading trust, both by citizens toward organizations and in those social networks involved in interorganizational innovation processes. In these situations, legislation that does not fully understand the complexities of everyday innovation activities could easily push different sector organizations into conflicting positions. Such conflicts, even if challenging in themselves, are even more dangerous if they end up degrading interorganizational affects, thus essentially limiting the boundaries of anticipation and willingness to dedicate necessary time for interorganizational learning among those individuals participating in innovation processes.

Toward Responsible MaaS Innovation and Institutional Transition

In addition to all the above possibilities and challenges, the ultimate question that MaaS brings to the table is reflectivity about innovation processes and institutional transition around emerging mobility technologies. From the implications above, it should be evident that systemic change needs systemic understanding of a wide range of factors and their interdependencies. With such uncertainty and complexity, we need to consider new pathways for speculative thinking concerning mobility futures in more responsible innovation processes. Wider system redesign for transition as opposed to cannibalization and enclosure of the current system forces us to ask questions about our innovation policies, investment flows, and objectives of wider societal governance. Simultaneously, we will have to keep on remembering that human ends are not static, but heterogeneous and dynamic set of capabilities, behaviors, and values. Here, we cannot try to streamline innovation processes by narrowing the interpretation for everyday mobility technology (Mladenović et al., 2019). Furthermore, innovation processes around MaaS will have to face a question of Universal Basic Mobility, aiming to include procedures for defining fair distribution of access and experiences, or at least a minimum safety net for those at risk of exclusion. In general, such value-sensitive innovation processes cannot overemphasize speed for regulative development, as it might be counterproductive for implementation activities themselves.

At the center of responsible MaaS innovation debate is a need to avoid depolitized normative foundations in deciding about technological emergence. Potential repolitization of technological development for enabling wider organizational and societal learning requires that debate moves away from narrowly defined expert circles, avoiding the Kehoe problem of building path dependence based on limited knowledge. For innovation to flourish, citizens need to be established as full participants in the envisioning process, alongside networks of actors from different societal sectors. Transparency and open participation of decision processes can also rely on developing open-source standards on the functional requirements for MaaS operation. If tailor-made solutions for interoperability are to be based on developing trust across social networks of innovation, hiding key details behind commercial confidentiality cannot be the power lever for private sector to control interorganizational learning. Moreover, such innovation processes cannot have dominant focus on technical development and anticipated positive effects, while avoiding questions of wider (re)distribution of consequences or solidifying existing societal inequalities. Here, we will have to recognize that debate might lead us to conclude that we need to leave behind some institutions and values, which can only be decided in a democratic process. Such processes for continuously responsible innovation will have to rely on governing piloting and experimentation efforts, and considering new governance levers, such as data governance. Stemming from the recent General Data Protection Regulation, aspects of data collection, storage, and multiway sharing across actors, as well as potential commercial transactions, have to be considered as part of a governance toolkit. Ultimately, it will remain important to continue understanding how does evolving web of promises around MaaS capture and divert (or not) strategic attention of the policy community.

Acknowledgment

This work has been partially supported by the Academy of Finland PROFI 4 initiative, as well as EU funding, through Finest Twins Center of Excellence. The author acknowledges the discussions with Kate Pangbourne, Dominic Stead, Dimitris Milakis, Moshe Givoni, Christoffer Weckström, and members of the Philosophy of the City group.

See Also

Operation Costs for Public Transport; Natural monopoly in transport; Pricing Principles in the Transport Sector; Demand for Passenger Transport; Equity concerns in transport; Procurement of Public Transport: Contractual regimes in public transport (competitive tenders, negotiations); Privatization and deregulation of public transport; Uber versus taxis; Efficiency of different types of contracts in public transport; TNCs and the Future of Taxi Public Transport; Public transport in low density areas; Advanced Traveler Information Systems; Demand-Responsive Transit, Evaluation Studies; Bikes sharing / Bicycle sharing; Carsharing; Transport modes and ageing society; Transport modes and sustainability; Transport modes and accessibility; Transport modes and cities; Transport modes and remote areas; Energy consumption of transport modes; Transport modes and health; Intermodal passenger transport; Vehicle sharing; ICT and Transport modes; The future of road transport; Transit planning and management; Transport modes and inequalities; Policy and Governance; Demand Responsive Transport; Politics of Mobility Policy

Biography

Miloš N. Mladenović is an assistant professor at the Spatial Planning and Transportation Engineering Group, Department of Built Environment, Aalto University. Miloš has obtained his BSc in Transport Engineering from the University of Belgrade, Serbia, and his MSc and PhD in Civil Engineering from Virginia Tech, USA. In addition, he has held a visiting research position at the Spatial Planning and Strategy chair, Delft University of Technology, the Netherlands. His current research interests include assessment of emerging mobility technologies, and development of decision-support methods. His previous and current teaching responsibilities include a range of courses in transport systems policy, planning, modeling, management, and design. Finally, he has been the organizing chair for Aalto University Summer School on Transportation since 2015.

References

- Collingridge, D., 1980. *The Social Control of Technology*. Frances Pinter, London, UK.
- Cottrill, C.D., 2020. MaaS surveillance: Privacy considerations in mobility as a service. *Transport. Res. Part A: Policy Pract.* 131, 50–57.
- Jasanoff, S., 2016. *The Ethics of Invention: Technology and the Human Future*. WW Norton & Company, New York, US.
- Haglund, N., Mladenović, M.N., Kujala, R., Weckström, C., Saramäki, J., 2019. Where did Kutsuplus drive us? Ex post evaluation of on-demand micro-transit pilot in the Helsinki capital region. *Res. Transport. Bus. Manage.* 32, 1–14.
- Heikkilä, S., 2014. *Mobility as a Service - A Proposal for Action for the Public Administration*. Aalto University, Case Helsinki.
- Hensher, D.A., 2017. Future bus transport contracts under a mobility as a service (MaaS) regime in the digital age: Are they likely to change? *Transport. Res. Part A: Policy Pract.* 98, 86–96.
- Lyons, G., Hammond, P., Mackay, K., 2019. The importance of user perspective in the evolution of MaaS. *Transport. Res. Part A: Policy Pract.* 121, 22–36.
- Mladenović, M.N., Lehtinen, S., Soh, E., Martens, K., 2019. Emerging urban mobility technologies through the lens of everyday urban aesthetics: case of self-driving vehicle. *Essays Philos.* 20 (2), 3.
- Pangbourne, K., Mladenović, M.N., Stead, D., Milakis, D., 2020. Questioning mobility as a service: unanticipated implications for society and governance. *Transport. Res. Part A: Policy Pract.* 131, 35–49.
- Rotolo, D., Hicks, D., Martin, B.R., 2015. What is an emerging technology? *Res. Policy* 44 (10), 1827–1843.
- Sochor, J., Arby, H., Karlsson, M., Sarasini, S., 2018. A topological approach to Mobility as a Service: a proposed tool for understanding requirements and effects, and for aiding the integration of societal goals. *Res. Transport. Bus. Manage.* 27, 3–14.
- Smith, G., Sochor, J., Karlsson, I.M., 2018. Mobility as a Service: development scenarios and implications for public transport. *Res. Transport. Econ.* 69, 592–599.
- Weckström, C., Kujala, R., Mladenović, M.N., Saramäki, J., 2019. Assessment of large-scale transitions in public transport networks using open timetable data: case of Helsinki metro extension. *J. Transport Geogr.* 79, 102470.
- Wong, Y.Z., Hensher, D.A., Mulley, C., 2020. Mobility as a service (MaaS): Charting a future context. *Transportation Research Part A: Policy and Practice*. 131, 5–19.

Further Reading

- Durand, A., Harms, L., Hoogendoorn-Lanser, S., Zijlstra, T., 2018. *Mobility-as-a-Service and changes in travel preferences and travel behaviour: a literature review*. Netherlands Institute for Transport Policy Analysis.
- Hensher, D., and Mulley, C., 2019. *Forthcoming. Editorial. Transportation Research Part A: Policy and Practice* In Hensher, D.A., Mulley, C., (Eds.), *Special issue on developments in Mobility as a Service (MaaS) and intelligent mobility*.
- Hirschhorn, F., Paulsson, A., Sørensen, C.H., Veeneman, W., 2019. Public transport regimes and mobility as a service: governance approaches in Amsterdam, Birmingham, and Helsinki. *Transport. Res. Part A: Policy Pract.* 130, 178–191.
- Jittrapirom, P., Caiati, V., Feneri, A.M., Ebrahimigharehbaghi, S., Alonso González, M.J., Narayan, J., 2017. Mobility as a service: a critical review of definitions, assessments of schemes, and key challenges. *Urban Plan.* 2 (2), 13–25.
- Jittrapirom, P., Marchau, V., van der Heijden, R., Meurs, H., 2018. Future implementation of Mobility as a Service (MaaS): Results of an international Delphi study. *Travel Behav. Soc.* (In press).
- Kamargianni, M., Li, W., Matyas, M., Schäfer, A., 2016. A critical review of new mobility services for urban transport. *Transport. Res. Proc.* 14, 3294–3303.

- Liimatainen, H., Mladenović, M.N., 2018. Special Issue: Understanding the complexity of mobility as a service. *Res. Transport. Bus. Manage.* 27, 1–2.
- Lyons, G., Hammond, P., Mackay, K., 2019. The importance of user perspective in the evolution of MaaS. *Transport. Res. Part A: Policy Pract.* 121, 22–36.
- Mulley, C., 2017. Mobility as a Services (MaaS)—does it have critical mass? *Transport Rev.* 37 (3), 247–251.
- Pangbourne, K., Mladenović, M.N., Stead, D., Milakis, D., 2019. Questioning mobility as a service: unanticipated implications for society and governance. *Transport. Res. Part A: Policy Pract.* 131, 35–49.
- Smith, G., Sochor, J., Karlsson, I.M., 2019. Intermediary MaaS Integrators: a case study on hopes and fears. *Transport. Res. Part A: Policy Pract.* 131, 163–177.

Bicycle Sharing

Cyrille Médard de Chardon, University of Hull, Hull, United Kingdom; Luxembourg Institute of Socio-Economic Research (LISER), Maison des Sciences Humaines, Esch-sur-Alzette/Belval, Luxembourg

© 2021 Elsevier Ltd. All rights reserved.

Bicycle Sharing	19
Variations	19
Stations	20
Operators	21
Ownership	21
Revenue	23
Pricing	23
Access and Affordances	24
Contracts	24
Integration	24
Actors	24
Governance	25
History	25
Free Bicycles	25
Coin-Deposit Bicycles	25
Marketing Innovation	25
Dockless and Electric	26
Outcomes	27
Critique	28
Concepts for Further Research	29
See Also	29
Relevant Websites	29
References	29
Further Reading	30

Bicycle Sharing

Bicycle sharing is an automated service providing shared-use of a bicycle fleet, typically within a dense urban setting, allowing short duration one-way trips from an origin to another location. Since 2005 technological innovation of an older concept combined with public, political, and corporate appeal made these systems common in larger cities, especially within the global North. As a new form of public transport typically implemented within public space, bicycle sharing has been a visible and relatively uncontentious public policy at a time of health, environmental, and equity challenges. This article describes the origin of bicycle sharing, their variations, recent growth and diversification, outcomes, and critiques.

Bicycle sharing is referred to using a variety of names. Bicycle sharing systems (BSS), schemes, services, or programs and public bike share (PBS) or hire, or similar variants, are some of the most common, but misnomers to some degree. Most BSS are not publicly owned or operated, nor is sharing strictly taking place (Belk, 2010). The “sharing” term likely came from an adaptation of car sharing (Shaheen et al., 1998), already using the term since the 70s explicitly, such as Ipswich’s 1977 Share-a-Car service. The use of “sharing” links BSS with the sharing and social economies’ positive connotations of social justice and community solidarity (Moulaert and Ailenei, 2005). Debate regarding the use of the term, or what is part of the sharing economy (Acquier et al., 2017), replicates debates about the contested use of shared space by BSS (Chan et al., 2018). More adequately descriptive terms exist, such as “self-serve bicycles” (France) and “hire bikes” (UK). While the term bikeshare is commonly used colloquially, within the literature BSS is common and will be used throughout this article.

Variations

What constitutes a BSS has become increasingly broad as new business models have appeared. Most commonly, BSS refers to automated bike-access through kiosks, docks, or the bikes themselves, allowing personal bicycle use for, depending on the case, relatively short trips between stations or within an area. The BSS principle relies on multiple users of the bicycles and is priced accordingly to encourage short trips. Punitive pricing schemes exist to discourage longer duration trips (Figure 5). Membership, which is integral to the BSS concept in order to link users’ identity or credit cards with their accounts and the bicycles they are



Figure 1 Clockwise, from the top left, examples of station-based, dockless, hybrid and electric bicycle sharing systems in Vancouver (CA), Manchester (UK), Cardiff (UK), and Luxembourg (LU), respectively.

entrusted with, have a range of costs, free durations of use, and affordances. Most BSS are unique combinations of complex relationships between diverse companies, actors, business models, local laws, system and cycling infrastructure, city and membership contracts, and transportation services and cultures. As these unique systems make generalizing difficult, statements refer to the more prevalent. The most common division between BSS types are those systems commonly referred to generically as bikeshare, which are station-based and dockless.

Stations

Station-based (SBSS) schemes constrain user trips to start and end at stations but generally allow unlimited free trips of a duration, typically, inferior to 30 or 45 minutes. While a smartphone may be practical to find stations, if unfamiliar with the city, a more likely use is to search for a station with available bicycles or docks, near the trip origin and destination respectively. The popularity of SBSS predate smartphone adoption, therefore is not required to use the system. These systems typically rely on bicycles (Figure 1) which can mate securely with docks (Figure 2), preventing, to a certain degree, theft and vandalism. Some systems use more conventional, but flexible, locks that can enable “piggybacking” locked bikes at full stations (See Cardiff image in Figure 1). Additional measures, such as component incompatibility with commercial bicycles, discourage theft. Stations consist of groups of docks, typically, controlled by a kiosk (Figure 3), that can offer the possibility to register, dispense RFID keys or cards, report damage or necessary maintenance, and display account information. Kiosks vary from providing additional services and information points, to requiring interaction in order to check out a bicycle. Importantly, kiosk backs can serve as advertising billboards. Checking out a SBSS bicycle typically requires interaction with the kiosk, dock, bicycle, a combination of these, or all three.

Dockless BSS (DBSS) users access bicycles almost exclusively through personal smartphone applications, allowing the start of a trip wherever a bicycle is available, and end trips anywhere within a demarcated, geofenced, zone (See Figure 4). Geofences are restrictive, limiting deposit of bicycles for legal, operational, and cost-effectiveness reasons, among others. Unlike SBSS, membership costs are typically trivial but users usually pay to unlock the bicycle and for each additional minute or longer time-segments (e.g., 20–30 min, see Figure 5). Some DBSS, such as in Cologne, Germany, could be checked out by calling operators from payphones for lock combination and to report the deposit location of the bicycle after use (before GPS was common).



Figure 2 Examples of station-based and dockless bicycle sharing systems components used to lock and secure bicycles.

Although, BSS are typically discretized into a dichotomy, a variety of hybrid systems exist, with overlaps and differences to station-based and dockless systems. Hybrid (HBSS) systems usually have bikes similar to DBSS but the requirement that bikes be used between stations (also called hubs) that may be no more than paint on the ground (sometimes called havens) or an operator provided bicycle rack. Some systems allow leaving bicycles outside hubs for an added marginal but nontrivial cost (e.g., some SOBI schemes).

Operators

The operation or provision of a BSS consists of an organization, often private, but potentially a nonprofit, private advertiser, municipal government, university, or transportation authority (DeMaio, 2009), maintaining, rebalancing, repairing, providing call-centers, and other aspects, such as marketing and administration, of the service. Rebalancing of bicycle distributions, which goes by many other names as well, is the process of dispatching field workers, and their carrier vehicle, effectively, potentially based on complex optimization algorithms, to achieve desired spatiotemporal distributions of bicycles. This process can be contentious (Médard de Chardon et al., 2017) in terms of desired outcomes, but has other under-recognized aspects, such as creating BSS spatial usage patterns that are not representative of demand. In line with the shared economy aims of unlocking trapped capital constrained in personal ownership (Acquier et al., 2017), rebalancing of BSS bicycles maximizes revenue by relocating bicycles to increase intensity of use. This process has familiar outcomes of capital accumulation: the inability of workers (who may live outside BSS coverage area) accessing the service, having precarious part-time or seasonal contracts without access to health care, low wages, and increased service provisionment in areas of higher urban density (and income).

Ownership

Bicycle and station BSS infrastructure is mainly owned by operators, especially in Europe where advertiser operators are more common, or host cities. Despite the appeal for cities to own and control infrastructure, the complexities of technology patents, and integration can make municipal ownership a burden. The inability of a city to expand a privately owned BSS by a company that has no such desire can be as constraining (e.g., Washington's 2008 SmartBike DC, see Klein, 2015) as the inability of enlarging a municipally-owned BSS with discontinued patented technology (e.g., Seattle's Pronto).



Figure 3 Kiosks are used exclusively by station-based systems to, potentially, register or dispense new memberships, check-out bicycles, report faults, and complete other queries.

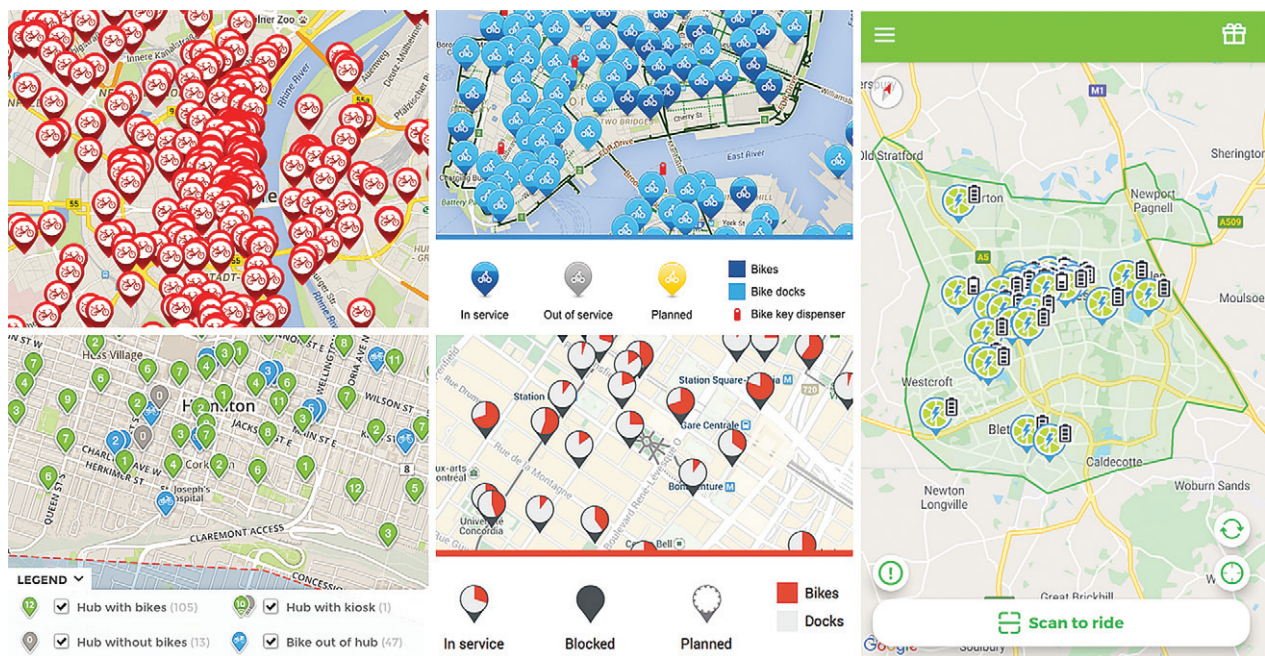


Figure 4 Examples of station-based and dockless bicycle sharing information systems.

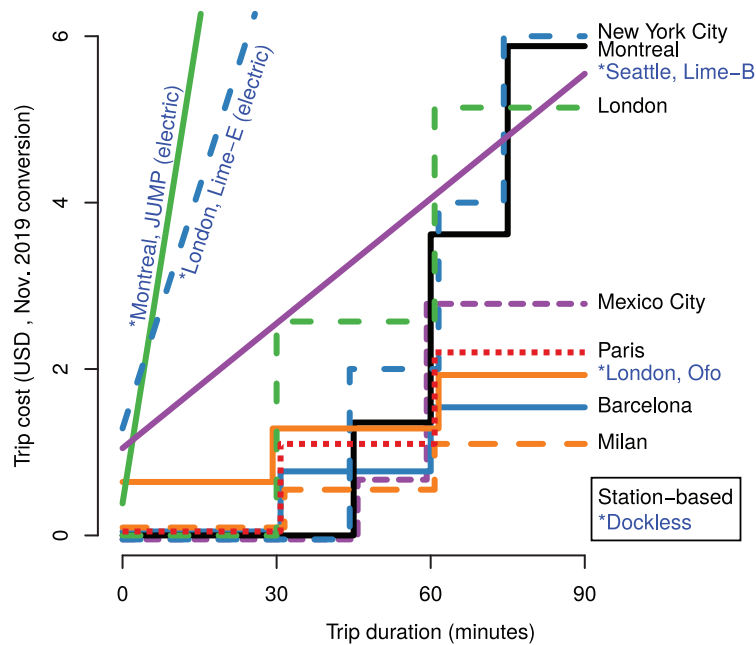


Figure 5 Cost structures for a selection of larger systems. Note most trips last less than 30 minutes. Membership prices are not included. *Source: Data source: bikeshare-research.org.*

Through technological or contractual leverage, BSS operators can determine or strongly influence BSS outcomes. Growth of this new mobility technology has been largely constrained within a relatively small diversity of companies providing BSS operations and technology (e.g., Uber, Lyft, Tencent, Alibaba, and JCDecaux owning, to some degree, Jump/Lime, Motivate, Mobike, Hellobike, and Cyclocity, respectively).

There exists a large number of smaller companies providing diversified services beyond the classic BSS models. Ownership of most BSS, like any specialized and patented technologies, limit owner independence, repair potential, and cost-effectiveness. Station-based systems are particularly vulnerable with the coupling of docks and bicycles requiring one system for “network effect” of potential trips to be realized.

Revenue

Advertising revenue has been the motivation, purpose, and cause of early BSS growth in Europe since 1998. Facilitated by technological innovation, a resurgence in the popularity of cycling, concerns for environmental sustainability, health, equity, and road congestion, BSS promoters regularly promote the systems as solutions. Despite claims of benefits, a clear BSS goal is often absent (Fishman et al., 2015; Ricci, 2015), preventing evaluation or, as is the case of advertiser provisioned systems, displaying the hypocrisy of sustainability narratives contradicted by consumerism promotion. Such schemes allowed outdoor advertising companies, such as Clear Channel and JCDecaux, to expand their offerings and markets by advertising to, and servicing, the urban privileged (Nitschke, 2015). Separating potential cycling benefits of BSS from the detriments of associated advertising have been challenging to publicly communicate (Tironi, 2014).

Aside from advertising billboards (Figure 3), advertising is also common on bicycles, through station branding, smartphone applications, and system sponsorships, targeted to the desired demographic. Additional revenue sources are municipal and federal grants and loans, membership fees, penalty fees, and the sale of user travel data (Duarte, 2016; Spinney and Lin, 2018; Waes et al., 2018). Penalty or extended-use fees are incurred by members exceeding trip lengths beyond the 30 minutes or so free period. In some situations, such as in Minneapolis, visitors, and recreational users, with short-term memberships and fees, have generated roughly two-thirds of revenue. Despite common claims of being self-funded, public subsidy for BSS is common in North America.

Branding has become an integral part of positioning BSS as sophisticated, fun, healthy, and green, to be promoted locally and externally (McCann, 2013). Conjointly, sponsors or service provisioners have also associated their own brand implicitly, through association, or explicitly, to the naming and styling of the systems (e.g., colors). Further, advertisers offer the opportunity to potential clients to green their brands by placing ads on BSS kiosks or adjacent while targeting a desirable demographic.

Pricing

Membership types have diversified from being strictly of various durations (e.g., day, week, month, year) to also providing different fee structures depending on bicycle type (e.g., manual or electric) and desired “free” period length (e.g., 30, 45, or 60 min). Memberships for DBSS, which were present in the 2017 wave of systems, are now typically free, trivial, or credit toward services.

However, DBSS typically charge a fixed fee at the start of trips and per additional minute. Electric DBSS, (eDBSS) typically have more expensive rates (Figure 5).

In addition to temporal fees, there exist spatial fees. Ending a trip away from a station and performing “out of hub locking” for HBSS typically has added costs (e.g., Hamilton, Canada, and Berlin, Germany). Leaving a bicycle outside of the geofenced service areas, for HBSS and DBSS, incurs high economic or social credit penalties.

Access and Affordances

Rather than sharing, BSS replicate renting, especially for DBSS, and membership models. While this paper focuses mainly on the mainstream systems of this strict definition, there are other bicycle models providing alternative affordances. Alternative cycling as a service (CaaS) models exist (Petzer et al., 2019) based on peer-to-peer renting (e.g., Spinlister), leasing (e.g., Swapfiets), and long-term BSS-like service (e.g., OV-Fiets), but only in a few countries which also have high-cycling modal share (e.g., Netherlands, Germany, parts of Belgium) and where mainstream BSS tend to have less of a presence. These services, just like BSS, vary between (implicit) goals of profit maximizing and facilitating cycling policies, and are priced accordingly. Long-term regional BSS, such as OV-Fiets, are integrated with public transport to allow rapid multimodal trips at trivial cost but the added constraint and responsibility of returning the bicycle to the origin.

Aside from the different systems providing variance in access and affordances, there are concrete BSS spatial, temporal, and economic accessibility factors. While all BSS constrain access spatially, based on availability of stations or a service zone, many also do so temporally, closing during nights and winter periods. Rather than strictly based on decreased demand, seasonal closures are preventative measures due to the negative effects of salt and snow plows, degrading or damaging the bicycles, performance issues with locking mechanisms, and the need to shovel free the bicycles from the snow.

Membership prices, for SBSS vary largely for annual subscriptions, from the trivial to the equivalent of the purchase of a low-end bicycle. Municipal governance decisions impact whether a BSS maximizes private profit generation or access to a bicycle. Many North American systems, but not most, provide discounted memberships to low-income residents. It is unclear to what extent this demographic group takes advantage of such offers in terms of membership or number of trips.

The usability of BSS at the scale of the bicycle and checking it out can alter access time from seconds to over a minute. Some systems require multistep procedures of interaction through the kiosk, such as presenting your card, pin, and designating the bicycle desired followed by unlocking the dock and adjusting impractical bicycle seat-height mechanisms. Necessary interactions also vary according to membership or short-term users.

Finally, additional accessibility constraints exist due to the availability of lights, for night use, range of seat heights, fenders for use in wet conditions, and alternative bicycle designs for people with disabilities.

Contracts

Contracts between municipalities and operators, who may also be the technology or patent owners, can have large impacts on BSS outcomes. Generally, contracts formalize the cost and period of service, number of bicycles and stations, and service levels in terms of rebalancing, and call centers, but also stipulate the provision of open-trip data, monthly reporting, and annual user surveys. Contractual details can have large impacts (and costs for the operator), such as the number of working and active bicycles on the street and the frequency, duration, and location of outages, where stations are full or empty. In some situations the ability or cost of future expansion has not been defined, resulting in constraints. Advertiser provisioned BSS agree to quantities of advertising panels of specified sizes. Barriers can arise, due to station kiosks sometimes doubling as advertising panels (Figure 3), if a municipality desires BSS expansion into neighborhoods where provisioners have little interest advertising.

Service cost contracts can fluctuate heavily according to municipality negotiations. While many advertiser provisioned BSS in Europe were provided without cost, in some cases, such as in Aix-en-provence, France, in addition to advertisements the municipality paid high rates for the services. Contractual opacity and constraints between conjoined advertising, BSS, and even bus shelters, have encouraged a disassociation of advertiser-provisioned services.

Integration

The integration of a BSS, spatially and in terms of quality of service, facilitates its use. While some cities facilitate integration with rail service, others avoid it due to the difficulty in satisfying demand or desire for BSS use for purposes other than daily commutes. At the street network scale some cities may more prominently allocate space for BSS, such as at intersections, while others relegate them to side-streets. While stations may replace on-street parking or lanes, many systems' stations are located on sidewalks. Either option can cause dissatisfaction with sidewalk and road users and local residents and businesses. In selecting either option municipalities communicate underlying values, such as the importance of cycling over other modes, that in some cases contradict suggested BSS outcomes while also creating tension between cyclists and other mode users (e.g., pedestrians or drivers). Municipal mediation strategies include charging operators rent to utilize parking space for stations (for a service which the city may be paying—indicating a performative practice).

Actors

The planning and deployment of BSS creates interaction between a variety of actors with complimentary and competing interests. While heavily influenced by the urban regimes of mayors, decision makers, technocrats, and large companies, wishing to attract the skilled or “creative” classes, some BSS deployments have also had to deal with public scrutiny. Within these actors are project

promoters seeking political gain, economic profit, or belief in BSS as a means of providing equity, health benefits, environmental sustainability, or improved quality of life. While well-placed promoters, such as mayors, can shape the narrative of BSS outcomes by championing their benefits, they can also facilitate and expedite their deployment. As BSS occupy public space physically, where many other public policies or private initiatives do not, they provide platforms for criticism and discontent. Some cities, particularly in North America, provide public participation in the BSS planning processes. Whether these initiatives result in pacification rather than democratizing project planning is unclear. Regardless, the diversity of actors, their interactions and tensions, shape BSS deployment, and outcomes.

Governance

Station-based systems, which largely popularized BSS, were often championed by mayors (DeMaio, 2009) but also bound by municipal requirements to collaborate in order to gain advertising rights and access to public land and, depending on the system, utilities. Dockless systems however (e.g., Mobike and Ofo) operate with much less constraint. Governance, the application of state power, of DBSS has only recently been enforced in some cities, resulting in bicycles being impounded or requiring permits and payment of fees. In Seattle, for example, operators pay a \$50 annual fee per bicycle, among other requirements.

There has been little governance surrounding where BSS are and are not being provided, in terms of stations but also the bounds of DBSS service zones, which results in unequal access to the services. Municipalities, who frequently champion BSS, desire the systems to be self-funding, have high use, and be labeled as a success, facilitating justification of the service being located in denser more exclusive urban cores.

The less formal relationships between companies and municipalities can result in reduced collaboration and unpredictability of future system presence, making planning a challenge. Rather than municipalities integrating DBSS for a desired urban mobility framework, these innovations are shaping transportation, with governance, adaptation and regulation being carried out in a reactionary manner. Bicycle sharing has largely been dictated by private and advertising interests shaping transport decision-making toward capital-maximizing and individualized travel with little public input.

Cycling governance has sometimes been adjusted to accommodate BSS. As BSS riders have lower rates of helmet use than personal bicycle riders (Zanotto and Winters, 2017), helmet laws were recognized as a barrier to use. Some cities have repealed their helmet laws proceeding BSS deployment (Fishman, 2015).

History

The evolution of BSS is regularly discretized into three generations. The social and technological innovation of new system types has been driven by various factors: protest to air pollution and consumerism, the affordances of an accessible and worry-free bicycle, the battle for advertising market share, and most recently, access to market niches. Efforts to define the attributes of a fourth generation BSS regularly overstate the importance of incremental changes. Rather than a linear evolution, a diversification of system technologies and service types has occurred.

Free Bicycles

Bicycle sharing, as a concept, was founded in 1965 in Amsterdam where the counter-culture movement *Provo*, within a context of urban renewal facilitating increased automobility and decreasing cycling (Ploeger and Oldenziel, 2020), protested against air pollution and consumerism by creating the witfietsen (White Bikes) (Beatley, 2000; Teun Voeten, 1990). Activists, in an initiative led by Luud Schimmelpennink, gathered bicycles, painted them white, and placed them on the street for free public use. The initiative successfully provoked police, who confiscated many of the unlocked bicycles through the use of a bylaw. Despite this the initiative demonstrated that an alternative model of bicycle access was possible. Schimmelpennink, believing in the concept, went on to develop and influence, among other things (Ploeger and Oldenziel, 2020), BSS growth for many other systems of note (e.g., Copenhagen and JCDecaux's systems) facilitating the rise and popularity of BSS in the mid 2000s.

Other cities mimicked the White Bikes, such as in Portland (US), but struggled in achieving a durable presence. La Rochelle, in France, due to strong mayoral support, managed to deploy and maintain a system for a longer period and of greater scale than others.

Coin-Deposit Bicycles

In 1995, a coin-operated system was developed in Copenhagen. Users could place a 20 Kroner coin (equivalent to \$4.50 today) to check out a bicycle, which would be returned upon locking at their destination. The Copenhagen system is the most well-known of this generation due to its size (2000 bicycles at 150 stations) and length of operation, until 2012 (17 years).

Neither the first nor second BSS generations became more common than oddities. While the vulnerability to theft and vandalism is often named as limiting factor to their growth, and as the following generations have to some degree similar problems, constraints where likely also the (un)popularity of cycling and lack of economic incentives and a viable business model.

Marketing Innovation

The technological development of automating the association of users to checked-out bicycles using smart-card technology in 1996 in Portsmouth, United Kingdom, enabled conventional BSS. However, it is market innovation that popularized BSS through advertising competition in Europe. In 1998 in Rennes, France, outdoor advertising furniture giant JCDecaux lost its bus-shelter contract to Clear Channel due to the latter offering a BSS in addition to bus shelters. JCDecaux rapidly developed its own BSS offering

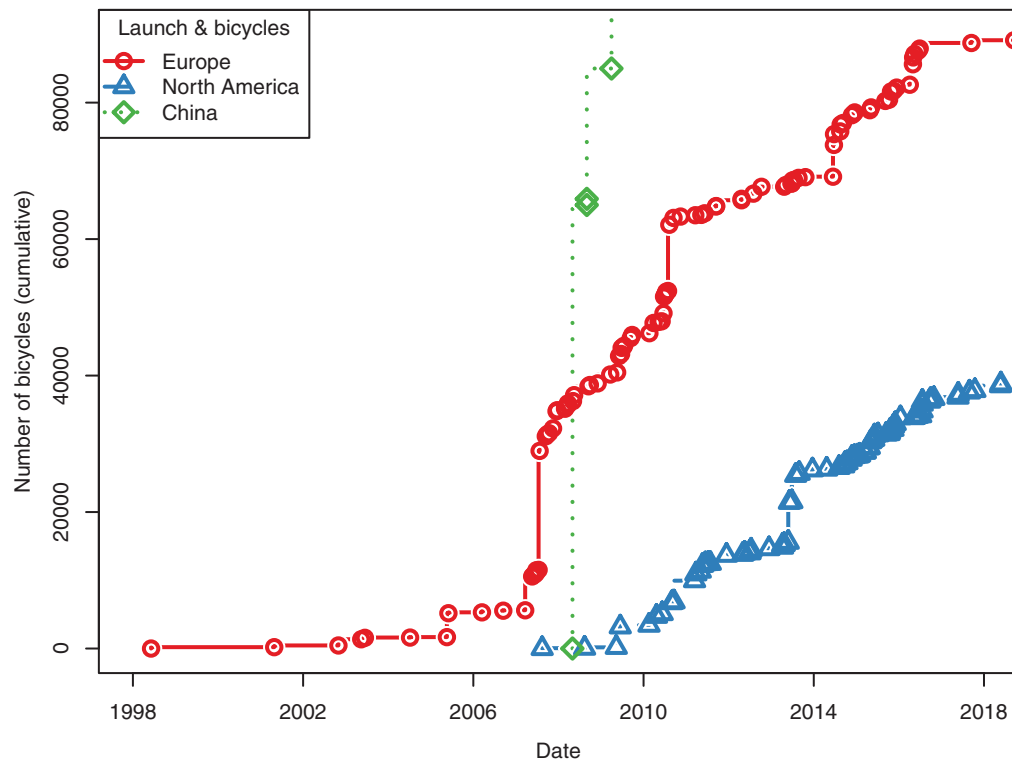


Figure 6 Third generation station-based bicycle sharing systems growth in terms of bicycles at launch (or earliest available statistic) in Europe, North America and China. Source: Data source: *bikeshare-research.org* and, for China, *Zhang et al., 2013*.

in order to remain competitive. As a result Lyon (2005) and Paris' (2007) requests for offers were contested. Driven by desire to maintain market share and facilitated by technical innovation, Lyon and Paris became lighthouse projects popularizing the BSS concept while also creating new markets for out-of-home advertising.

While growth in Europe precedes that of China and North America, the scale of Chinese system sizes fast outpaced the rest of the world combined. **Figure 6** shows the growth of third generation SBSS in terms of number of bicycles at launch. The figure highlights the large number of small systems that are present in Europe and North America, relative to China.

Europe's BSS growth, while having some stability in system continuity, was impacted by the 2007–2008 economic crisis resulting in some smaller systems being poorly funded or closed, particularly in Spain and Italy. Germany was very different than other countries. Students initiated a DBSS in 1998, requiring the use of pay phones. It was purchased by the German railway company and expanded nationally. This system grew to have a large presence in some cities while also being only present at railway stations in many other cities, and in some case limiting the deployment of larger systems (Nitschke, 2015).

While advertiser-provisioned BSS dominated the deployment of larger systems in Europe, in North America SBSS were largely provided by various combinations of technology providers, infrastructure ownership, and operators. Opposing desires between the municipality and operator (Clear Channel) for North America's (Washington DC) first BSS trial in 2008 (Klein, 2015) provided a cautionary demonstration of advertisement-provisioned limitations, making them an unpopular choice.

While many systems have been deployed in North America, mainly in the United States, a very large proportion are small. **Figure 7** depicts SBSS growth in number of systems and bicycles. The figure contains limitations due to some small BSS missing. Data for DBSS is more difficult to summarize due to their ephemeral nature, barriers to data accessibility, and reduced public collaboration.

China's BSS growth, 2008–2011, was largely station-based, operated by nonprofit enterprises working in collaboration and with government support. Prior policies of "progress" toward the automobile and to decrease bicycle use in the 90s, were being reversed with plans for extensive bike lane development. The growth of DBSS in 2015 contrasts with previous systems, by being privately owned and operating without municipal support. Similarly to, for example, the US and UK, the barriers to BSS use in China are the same as the barriers to cycling, mainly safety concerns and lack of adequate infrastructure. There were 12 BSS in China in 2012 (Zhang et al., 2013) while North America had 26 (Shaheen et al., 2014). China however had over 180,000 bicycles while North America had fewer than 14,000.

Dockless and Electric

The 2014–2017 period was turbulent for BSS. Rather than an evolution to a fourth generation of BSS, a diversification of service and technology types appeared that are not superior but filling market niches. In Europe some early service contracts expired and were

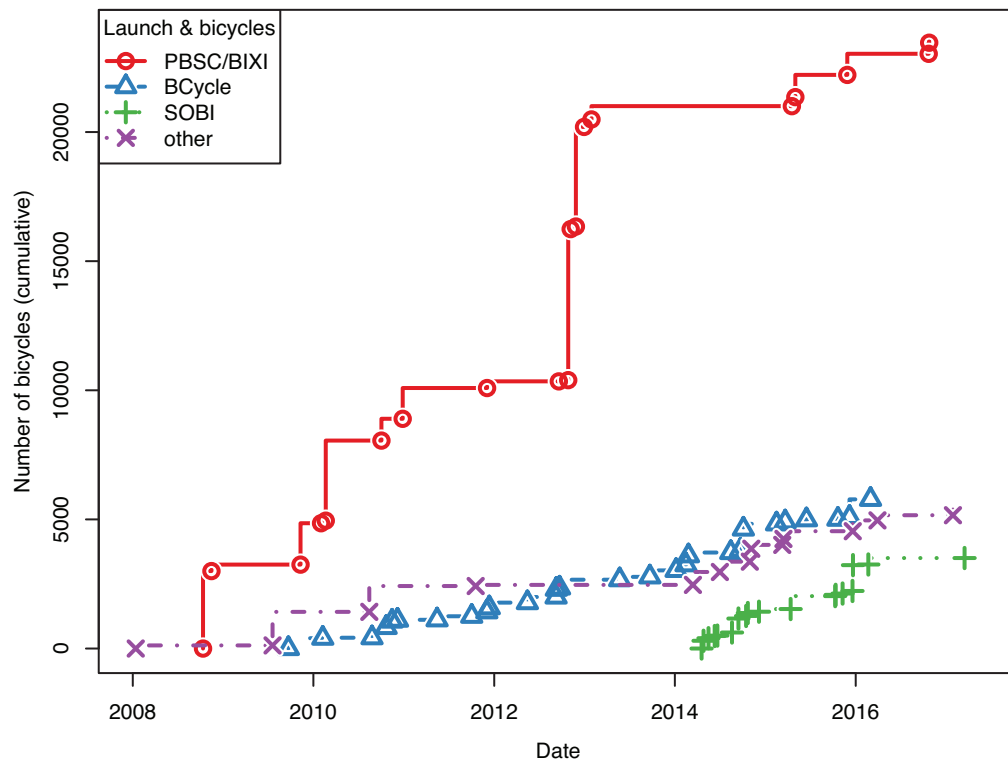


Figure 7 North American third-generation station and hybrid bicycle sharing system growth in terms of bicycles at launch (or earliest available statistic), grouped by technology provider. Source: Data source: bikeshare-research.org.

not renewed, such as in Paris, JCDecaux's showcase city. While advertiser provisioned BSS, more common in Europe, were always in the marketing business, in North America large BSS-related companies transitioned from the transport sector to ownership by investment and technology companies. Monetary constraints, and related conflicts, impacted the largest North American BSS technology provider and operator, PBSC and Motivate, respectively, creating uncertainty (e.g., Seattle) and opportunity (e.g., SOBI) (Médard de Chardon, 2019). In China, in 2016, the rapid growth and expansion of DBSS, such as Mobike and Ofo, flooded Chinese streets, rapidly transforming positive associations to problematic (Gu et al., 2019). In 2017, large numbers of these bicycles appeared in North America and Europe with little local government collaboration or adequate preparation. Vandalism was once again blamed as a cause of the exit of these DBSS, but declining funding, competition, supply chain problems, lack of local knowledge and municipal responses, through fees or impounding of bicycles, challenged their permanence. While DBSS may have greater potential to scale (Waes et al., 2018) due to reduced costs, other costs appear to be prohibitive to long-term viability (Gu et al., 2019).

During this period the hybrid BSS appeared in North America, mainly through SOBI, but other technology providers appeared, such as the French Smoove, promoting its "smart-bikes" which provided no additional affordances over regular SBSS. Riding on the popularity of smart cities, reference to smart-bikes has been used by a variety of BSS. Additionally, the functionality of DBSS, in some cases, are resembling more SBSS due to increasing requirements that free-floating bicycles be placed within delineated spaces due to conflict with sidewalk users. Earlier fears of SBSS being supplanted by DBSS have been unrealized.

Bicycle sharing novelty has been a large driver of its growth, explaining, since 2015 roughly, the rising number of electric BSS (eBSS) that provide a more luxurious service (and image). This investment in eBSS has impacts in terms of coverage potential and number of concurrent users relative to a mechanical BSS, given a fixed investment, mimicking similar debates between bus versus tram or light-rail services (Walker, 2012).

Outcomes

Generalizing the outcomes of BSS or evaluating their potential is challenging due to the large number of variables (see Variations above). There exists a large variability of usage rates (Médard de Chardon et al., 2017) determined by many factors (Faghih-Imani et al., 2017; Sun et al., 2018; Wang and Akar, 2019). The most consistent across BSS studies is that riders are more likely male, white, wealthier, younger, more educated, owning a bicycle, and employed, than the general population (Bauman et al., 2017; Fishman, 2015; McNeil et al., 2017; Nickkar et al., 2019; Ricci, 2015; Wang and Akar, 2019). The fundamental cause of this is BSS being located in exclusive spaces, such as university campuses, urban cores (Hosford and Winters, 2018; Ursaki and Aultman-

Table 1 Modal share replacement of bicycle sharing systems from member surveys. Variable member usage rates prevent modal shift impact prediction from total trips counts or distances. Sources: Fishman et al., 2013; Fishman et al., 2014; Fuller et al. (2011), LDA Consulting (2016); Murphy and Usher (2014); and Winters et al. (2017).

Case study	Driving	Bus	Rail	Walking	Mode shift (%)		New trip
						Private cycling	
Dublin	20	26	9	46	-	-	-
London	2	20	37	26	8	8	4
Lyon	7	50*	*	31	6	6	4
Melbourne	19	41*	*	27	9	9	1
Minneapolis	19	20*	*	40	8	8	9
Montreal	35	44*	*	13	5	5	-
Vancouver	6	19	4	54	9	9	2
Washington	5	14	21	39	3	3	2

* Bus and rail combined

Hall, 2016), and the roads themselves, facilitating access for those with economic means and willingness to cycle within hostile spaces.

Who uses BSS determines where beneficial outcomes are located. At the individual user level, BSS provide convenience but with limitations. While granting use of a bicycle, potentially electric, for a one-way trip without fear of theft and responsibility for maintenance and storage, there are trip or membership costs, the added weight of a durable bicycle, potentially incompatibility in terms of size and (dis) ability, requiring travel to or from the bicycle, and the uncertainty of having access to a bicycle nearby and a free dock at the desired destination. Additionally, convenience or limitations are relative to individually-accessible alternative modes for the same trip and societal barriers to cycling (Golub et al., 2016).

Usage rates of BSS are frequently used to determine the potential impacts on health, environmental sustainability, equity, and even congestion mitigation. The modal shift rates from other transport to BSS is helpful in evaluating outcomes (Table 1). More than simply replacing modes, BSS can complement or substitute modes depending on other factors.

Findings of outcomes are mixed in regards to reducing carbon emissions (Audikana et al., 2017; Fishman et al., 2014), improving health (Murphy and Usher, 2014; Zhang et al., 2013), gender and social equity (Garrard et al., 2012; Hoffmann, 2016; Ricci, 2015; Wang and Akar, 2019), and facilitating a larger cycling modal shift (Bauman et al., 2017; Hosford et al., 2019; Pucher et al., 2011; Ravalet and Bussiere, 2013). Despite the abundance of new transportation data BSS provide, abstracting clear causes and outcomes is still challenging, with small groups of BSS users potentially biasing overall outcome perceptions (Winters et al., 2019).

The benefits of BSS are often confounded with those of cycling, causing statements by urban regime actors, BSS technology and service providers, and even academic literature as normative opening statements to be largely unfounded (Médard de Chardon, 2019). Despite this, BSS policies are innovative for being pro-cycling, where few existed before. Additionally, they help to raise awareness and acceptance of bicycles, normalizing casual and utility cycling, due to their presence (Goodman et al., 2014).

Critique

While bicycle sharing systems are convenient and enjoyable to use for a privileged subset of the population, overstated benefits, usually the result of conflating cycling and BSS, are heavily promoted by those associated with the systems. Specifically, health, environmental sustainability, congestion reduction, and equity benefits are exaggerated. Most BSS are of such a size that they have little impact at the urban scale of transport, making them tokenistic performances. Similarly to the BSS deployments accompanying Republican (Cleveland, 2016) and Democratic (Charlotte, 2012 and Denver, 2008) National Conventions in the United States (Marshall et al., 2015), the Olympics and World Expo in China also motivated BSS launches (Zhang et al., 2015).

While BSS serve some commuters and recreational users, alternative, more cost-effective, efficient, and established practices are more proven to yield greater cycling modal share and all their associated benefits (Buehler et al., 2016). While it is unlikely BSS will bring about a societal transition, it may exacerbate existing inequalities by providing uneven accessibility improvements, resulting in increased tensions and self-policing (Spinney and Lin, 2018).

The recent growth of eBSS, in replacing mechanical BSS, supports perspectives of these systems being urban symbols of sophistication for self-promotion (McCann, 2013), interlocal competition (Peck and Tickell, 2002), and tools of gentrification (Hoffmann, 2016) due to their providing a luxurious experience at greater cost to fewer people than their predecessors.

Finally, the rapid growth of BSS can be explained by a confluence of self-serving interests, facilitated by a metaphorical and literal neoliberal vehicle for entering and expanding markets (Duarte, 2016), like other smart-city initiatives (Martin et al., 2018), utilizing unfounded narratives of sustainability, health, and social equity (Médard de Chardon, 2019). The evaluation of BSS have been too often using hard to evaluate metrics rather than opportunity costs, revealing the lack of clear BSS objectives, and greater benefits of alternative policies.

Concepts for Further Research

Bicycle sharing research frequently provides case-specific insight. While some trends are apparent, such as the dominant type of BSS user, generalizations are challenging due to the many variations between BSS (see Variations above). In the growing sectors of the sharing economy, micro-mobility, and mobility and cycling-as-a-service, BSS are one of the most widely adopted. Despite this, its efficacy at providing clear outcomes is unproven.

The ease of BSS data access and aggregation has made the domain attractive for quantitative analysis in an effort to describe urban mobility. This explains the richness of statistical and regression work on the topic. Dockless BSS conversely due to lacking data access are much more opaque. Regardless, there is a clear need for more nuanced and qualitative analysis of BSS' governance and interactions between its actors. The technical aspects of BSS, relating to technology patents, contracts, and their lock-ins, have had large impacts on the direction and outcomes of key BSS projects (e.g., Rennes, Paris, Seattle) that help reveal the motivations and procedure within urban regimes and markets.

While there has been extensive study of theoretical bicycle rebalancing optimization, very little research has explored this process in practice that has important control over system outcomes. Additionally, little work has explored why BSS are not frequent or having high usage in countries and cities with high-cycling modal shares, raising the question of whether they serve as necessary tools for a cycling transition. If they are inadequate tools for modal behavioral change, it reinforces their outcome as being an additional, largely privatized, accessibility option for the urban hyper-mobile.

Finally, in a world of institutional automobility, little work has studied how BSS governance relates to challenging this problematic state. Despite their promotion and sometimes dubious goals, BSS still appear to be "fit" in urban spaces, without challenging the pervasive facilitation of cars.

See Also

Mobility as a Service; Active Travel; Cycling; Transitions and disruptive technologies in transport planning; Sustainable Urban Mobility Plans; Transport Demand Management

Relevant Websites

<http://bike-sharing.blogspot.com/>, <https://bikeshare-research.org>

References

- Acquier, Aurélien, Daudigeos, Thibault, Pinkse, Jonatan, 2017. Promises and paradoxes of the sharing economy: An organizing framework. *Technol. Forecast. Soci. Change* 125, 1–10, doi:10.1016/j.techfore.2017.07.006.
- Audikana, Ander, et al., 2017. Implementing bikesharing systems in small cities: Evidence from the Swiss experience. *Trans. Policy* 55, 18–28, doi:10.1016/j.tranpol.2017.01.005.
- Bauman, Adrian, et al., 2017. The unrealised potential of bike share schemes to influence population physical activity levels – A narrative review. *Prev. Med.* 103S, S7–S14, doi:10.1016/j.ypmed.2017.02.015.
- Beatley, Timothy, 2000. *Green urbanism: Learning from European cities*. Island Press, Washington, DC.
- Belk, Russell, 2010. Sharing. *J. Cons. Res.* 36.5, 715–734, doi:10.1086/612649.
- Buehler, Ralph, et al., 2016. Reducing car dependence in the heart of Europe: lessons from Germany, Austria, and Switzerland. *Trans. Rev.* 37.1, 4–28, doi:10.1080/01441647.2016.1177799.
- Chan, Jeffrey Kok Hui, Zhang, Ye, 2018. Sharing space: Urban sharing, sharing a living space, and shared social spaces. *Space Cult.* 1–13, doi:10.1177/1206331218806160, 120633121880616.
- DeMaio, Paul, 2009. Bike-sharing: History, impacts, models of provision, and future. *J. Public Transp.* 12.4, 41–56, doi:10.5038/2375-0901.12.4.3.
- Duarte, F., 2016. Disassembling bike-sharing systems: surveillance, advertising, and the social inequalities of a global technological assemblage. *J. Urban Tech.* 23.2, 1–13, doi:10.1080/10630732.2015.1102421.
- Faghihi-Imani, Ahmadsreza, Eluru, Naveen, Paleti, Rajesh, 2017. How bicycling sharing system usage is affected by land use and urban form: analysis from system and user perspectives. *Eur. J. Trans. Infrastruct. Res.* 17, 425–441.
- Fishman, Elliot, 2015. Bikeshare: A review of recent literature. *Transp. Rev.* 1–22, doi:10.1080/01441647.2015.1033036.
- Fishman, Elliot, Washington, S., Haworth, N., 2013. Bike share: A synthesis of the literature. *Trans. Rev.* 33, 148–165, doi:10.1080/01441647.2013.775612.
- Fishman, Elliot, Washington, Simon, Haworth, Narelle, 2014. Bike share's impact on car use: Evidence from the United States, Great Britain, and Australia. *Transp. Res. Part D* 31, 13–20, doi:10.1016/j.trd.2014.05.013.
- Fishman, Elliot, et al., 2015. Factors influencing bike share membership: An analysis of Melbourne and Brisbane. *Transp. Res. Policy Prac.* 71, 17–30, 10.1016/j.tra.2014.10.021.
- Fuller, D., et al., 2011. Use of a new public bicycle share program in Montreal, Canada. *Am. J. Prev. Med.* 41.1, 80–83, doi:10.1016/j.amepre.2011.03.002.
- Garrard, J., Handy, S., Dill, J., 2012. Women and cycling, in: John R. Pucher, J.R., Buehler, R. (Eds.), *City Cycling, Urban and Industrial Environments*. MIT Press, Cambridge, MA, pp. 211–234.
- Golub, A., et al., 2016. *Bicycle Justice and Urban Transformation: Biking for all*. Routledge, Taylor & Francis Group, London, UK 294 pp. url: http://www.ebook.de/de/product/24364200/bicycle_justice_and_urban_transformation.html.
- Goodman, A., Green, J., Woodcock, J., 2014. The role of bicycle sharing systems in normalising the image of cycling: An observational study of London cyclists. *J. Trans. Health* 1.1, 5–8, doi:10.1016/j.jth.2013.07.001.
- Gu, Tianqi, Kim, Inhi, Currie, Graham, 2019. To be or not to be dock-less: Empirical analysis of dockless bikeshare development in China. *Transp. Res. Part A* 119, 122–147, doi:10.1016/j.tra.2018.11.007.
- Hoffmann, M.L., 2016. *Bike Lanes Are White Lanes: Bicycle Advocacy and Urban Planning*. University of Nebraska Press, Lincoln & London pp. 210.

- Hosford, Kate, Winters, Meghan, 2018. Who are public bicycle share programs serving? An evaluation of the equity of spatial access to bicycle share service areas in Canadian cities. *Transp. Res. Rec.* 2672, 42–50, doi:10.1177/0361198118783107.
- Hosford, Kate, et al., 2019. Evaluating the impact of implementing public bicycle share programs on cycling: the International Bikeshare Impacts on Cycling and Collisions Study (IBICCS). *Int. J. Behav. Nutr. Phys. Act.* 16, 1–11, doi:10.1186/s12966-019-0871-9.
- Klein, Gabe, 2015. *Start-up city: Inspiring private and public entrepreneurship, getting projects done, and having fun.* Island Press, Washington, DC.
- LDA Consulting (2016). 2016 Capital Bikeshare Member Survey Report. url: <https://www.capitalbikeshare.com/system-data> (visited on 2019-11-05).
- Marshall, Wesley E., Duvall, Andrew L., Main, Deborah S., 2015. Large-scale tactical urbanism: the Denver bike share system. *J. Urban. Int. Res. Placemak. Urban Sustain.* 9, 135–147, doi:10.1080/17549175.2015.1029510.
- Martin, Chris J., Evans, James, Karvonen, Andrew, 2018. Smart and sustainable? Five tensions in the visions and practices of the smart-sustainable city in Europe and North America. *Technol. Forecast. Soc. Change* 133, 269–278, doi:10.1016/j.techfore.2018.01.005.
- McCann, Eugene, 2013. Policy boosterism, policy mobilities, and the extrospective city. *Urban Geog.* 34.1, 5–29, doi:10.1080/02723638.2013.778627.
- McNeil, Nathan, et al., 2017. Breaking barriers to bike share: Insights from residents of traditionally underserved neighborhoods. *Tech. Rep. Trans. Res. Edu. Center*, doi:10.15760/trec.176.
- Médard de Chardon, C., 2019. The contradictions of bike-share benefits, purposes and outcomes. *Transp. Res. Part A* 121, 401–419, doi:10.1016/j.tra.2019.01.031.
- Médard de Chardon, C., Caruso, G.R., Thomas, I., 2017. Bicycle sharing system 'success' determinants. *Transp. Res. Part A* 100, 202–214, doi:10.1016/j.tra.2017.04.020.
- Moulaert, Frank, Ailenei, Dana, 2005. Social economy, third sector and solidarity relations: A conceptual synthesis from history to present. *Urban Stud.* 42.11, 2037–2053, doi:10.1080/00420980500279794.
- Murphy, Enda, Usher, Joe, 2014. The role of bicycle-sharing in the city: Analysis of the Irish experience. *Int. J. Sustain. Transp.* 9.2, 116–125, doi:10.1080/15568318.2012.748855.
- Nickkar, Amirreza, et al., 2019. A spatial-temporal gender and land use analysis of bikeshare ridership: The case study of Baltimore City. *City Cul. Soc.* 18, p100291, doi:10.1016/j.ccs.2019.100291.
- Nitschke, L., 2015. Public bike sharing in Munich: A critical view on bike sharing and redistribution of urban space. MA thesis. Aalborg and Denmark: Aalborg University. url: http://projekter.aau.dk/projekter/files/213479008/Master_thesis_LucaNitschke_text.pdf (visited on 2015-09-14).
- Peck, J., Tickell, A., 2002. Neoliberalizing space. *Antipode* 34.3, 380–404, doi:10.1111/1467-8330.00247.
- Petzer, B.J.M., Wiecek, A.J., Verbong, G.P.J., 2019. Cycling as a service assessed from a combined business-model and transitions perspective. *Environ. Innov. Soc. Transit.* doi:10.1016/j.eist.2019.09.001.
- Ploeger, Jan, Oldenziel, Ruth, 2020. The sociotechnical roots of smart mobility: Bike sharing since 1965. *J. Trans. His.* 41, 134–159, doi:10.1177/0022526620908264.
- Pucher, John, Buehler, Ralph, Seinen, Mark, 2011. Bicycling renaissance in North America? An update and re-appraisal of cycling trends and policies. *Transp. Res. Part A* 45.6, 451–475, doi:10.1016/j.tra.2011.03.001.
- Ravalet, E., Bussiére, Y., 2013. Les systèmes de vélos en libre-service expliquent-ils le retour du vélo en ville? *Recher. Transp. Sec.* 2012, 15–24, doi:10.1007/s13547-011-0020-6.
- Ricci, Miriam, 2015. Bike sharing: A review of evidence on impacts and processes of implementation and operation. *Res. Trans. Bus. Manag.* 15, 28–38, doi:10.1016/j.rtbm.2015.03.003.
- Shaheen, Susan, Sperling, Daniel, Wagner, Conrad, 1998. Carsharing in Europe and North America: Past, present, and future. *Transp. Quart.* 52.3, 35–52.
- Shaheen, Susan A., et al., 2014. Public Bikesharing in North America During a Period of Rapid Expansion: Understanding Businessmodels, Industry Trends and User Impacts. Mineta Transportation Institute, San Jose, CA url: <https://transweb.sjsu.edu/sites/default/files/1131-public-bikesharing-business-models-trends-impacts.pdf> (visited on 2019-01-11).
- Spinney, Justin, Lin, Wen-l, 2018. Are you being shared? Mobility, data and social relations in Shanghai's Public Bike Sharing 2.0 sector. *Appl. Mobil.* 3.1, 66–83, doi:10.1080/23800127.2018.1437656.
- Sun, Feiyang, Chen, Peng, Jiao, Junfeng, 2018. Promoting public bike-sharing: A lesson from the unsuccessful Pronto system. *Transp. Res. Part D* 63, 533–547, doi:10.1016/j.trd.2018.06.021.
- Teun, Voeten, 1990. Dutch Provos. *High Times* 32–36, 64–66, 73. url: www.marijuanalibrary.org/HT_provos_0190.html (visited on 2016-03-24).
- Tironi, Martin, 2014. (De) politicising and ecologising bicycles. *J. Cult. Econ.* 8.2, 1–18, doi:10.1080/17530350.2013.838600.
- Ursaki, Julia, Aultman-Hall, Lisa, 2016. Quantifying the Equity of Bike-share Access in U.S. Cities. Transportation Research Board, Washington, D.C.
- Waes, Arnoud van, et al., 2018. Business model innovation and socio-technical transitions. A new prospective framework with an application to bike sharing. *J. Clean. Prod.* 195, 1300–1312, doi:10.1016/j.jclepro.2018.05.223.
- Walker, Jarret, 2012. *Human Transit.* Island Press, Washington, DC pp. 256.
- Wang, Kailai, Akar, Gulsah, 2019. Gender gap generators for bike share ridership: Evidence from Citi Bike system in New York City. *J. Trans. Geogr.* 76, 1–9, doi:10.1016/j.jtrangeo.2019.02.003.
- Winters, Meghan, Hosford, Kate, Javaheri, Sana, 2019. Who are the 'super-users' of public bike share? An analysis of public bike share members in Vancouver, BC. *Prev. Med. Rep.* 15, p100946, doi:10.1016/j.pmedr.2019.100946.
- Winters, M., Therrien, S., Zanotto, M., 2017. Understanding a New Bike Share Program in Vancouver. Report Prepared for the City of Vancouver & Mobi by Shaw Go Partners.
- Zanotto, Moreno, Winters, Meghan L., 2017. Helmet use among per-sonal bicycle riders and bike share users in Vancouver, BC. *Am. J. Prev. Med.* 53, 465–472, doi:10.1016/j.amepre.2017.04.013.
- Zhang, H., Shaheen, Susan A., Chen, Xingpeng, 2013. Bicycle evolution in China: From the 1900s to the present. *Int. J. Sust. Trans.* 8.5, 317–335, doi:10.1080/15568318.2012.699999.
- Zhang, L., et al., 2015. Sustainable bike-sharing systems: Characteristics and commonalities across cases in urban China. *J. Clean. Prod.* 97, 124–133, doi:10.1016/j.jclepro.2014.04.006.

Further Reading

- Beroud, B., Anaya, E., 2012. Private interventions in a public service: An analysis of public bicycle schemes. *Cycling and Sustainability*, in: Parkin, J. (Ed.), Vol. 1. Transport and Sustainability. Emerald Group Publishing Limited, pp. 269–301. doi: 10.1108/s2044-9941(2012)000001013.
- Hannig, J., 2016. Community disengagement, in: Golub, A., et al. (Eds.), *Bicycle Justice and Urban Transformation*, Taylor & Francis Group, London, UK.
- Marsden, Greg, Reardon, Louise, 2018. *Governance of the Smart Mobility Transition.* Emerald Publishing Limited, >Bingley, UK pp. 192.
- Parkes, Stephen D., et al., 2013. Understanding the discussion of public bike-sharing systems: Evidence from Europe and North America. *J. Trans. Geogr.* 31, 94–103, doi:10.1016/j.jtrangeo.2013.06.003.

Light Rail

Fiona Ferbrache, Keble College, Oxford, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Light Rail	31
History	32
Renaissance	33
The Promise of Light Rail	34
User Benefits	34
Non-User Benefits	34
Light Rail and Urban Development	35
Land and Property Values	36
Sustainability	36
Social Exclusion and Gentrification	37
References	37
Further Reading	38

Light Rail

Light rail transit (LRT) describes a range of urban rail transit systems running largely on exclusive rights of way within city centers and between city centers and their suburbs. Light rail operates mainly at street level but may include elevated and underground sections as well as tunnels, and can operate on converted rail routes. Light rail is electrically powered, usually from overhead cables, but since APS (*alimentation par le sol*) ground system was pioneered in Bordeaux (which opened 2003/2004), ground system supply has been used in other cities, particularly in central locations where heritage or aesthetics demand protection of buildings or urban features such as bridges. Light rail usually features level-boarding and pre-paid fares and provides an enhanced quality of transit in terms of service speed, frequency, safety and reliability, as well as a modern design (see Figs. 1 and 2). However, light rail is varied and includes forms more commonly referred to as, for example, modern streetcars (e.g., Portland Streetcar, Oregon, US), driverless light metros (e.g., Copenhagen Metro, Denmark; Dockland's Light Railway (DLR), London, UK), supertrams (e.g., Sheffield Supertram, UK) and tramways (e.g., Toulouse, France). There is no well-established boundary between these different forms and sometimes the terms are used interchangeably.

More broadly, light rail is often distinguished from other transit modes such as buses and heavy rail, though they may share certain features. Commonly noted points of distinction include light rail's fixed infrastructure versus buses' flexibility (often expressed as permanence versus temporality), its image and "the assumption that trains are sexy and buses are boring" (Hensher, 2016:289), and light rail's reputation as better value for money than bus rapid transit (Hensher, 1999). In terms of rail, light rail is usually a cheaper alternative to heavier rail modes such as suburban and commuter railways and (underground) metros, quicker to implement than heavy rail, and while light rail may share road space with other vehicles, stations, stops and services are more frequent in the city center, thus providing a fast and accessible service. Table 1 provides a summary of light rail characteristics.

Geographically, light rail investment has spread to cities across the globe since its 1970s renaissance (see below). UITP, the International Association of Public Transport, identified 411 cities with light rail in 2018 though this is not definitive given the broad definition of light rail (UITP 2019). Europe and North America are regarded as traditional regions of light rail development and Europe records the highest number of cities featuring light rail (see Fig. 3). As shown in Fig. 4, Asia-Pacific was identified as the leading region for light rail development in 2017 and 2018 (as measured in kilometers (km) of track) with 110.3 km constructed in 2018. Light rail in the Middle East and North Africa (MENA) is also growing rapidly (and ahead of North America). In Europe and North America, while light rail investment continues, it tends to be more focused on maintaining and replacing existing infrastructure rather than new developments. The expansion of light rail investment—new development, extension of lines, or conversions to light rail—has been boosted by its implementation in medium-sized cities (100,000–300,000 inhabitants) as well as larger conurbations. Not all countries that invested in light rail have continued to develop their systems. Some older systems have been closed during this period and some countries, including the UK and the Netherlands, have slowed investment.

Light rail investment has become popular in the wake of concerns over urban traffic congestion and has become part of broader sustainable mobility and sustainable urban agendas (see below). In addition, it is widely recognized that light rail can act as a catalyst or enhancer of urban (re)development and investment that can help to boost a city or region's economic growth and raise its competitive standing. Light rail literature, which includes academic work from the social sciences, engineering, environmental science and urban planning, as well as a range of professional impact studies and reports, provides substantial insight to ex ante and ex post assessments of light rail. This literature tends to adhere to neoliberal urbanism, competitive urbanism and, more recently, sustainable urbanism, leaving a significant gap on more environmental, social and humanistic approaches to light rail.



Figure 1 An example of light rail: Bordeaux tram, France. *Source: Marc Ryckaert/Wikimedia Commons/Creative Commons Attribution 3.0 Unported license.*



Figure 2 Tramway stop with level boarding, Toulouse, France. *Source: Author's own.*

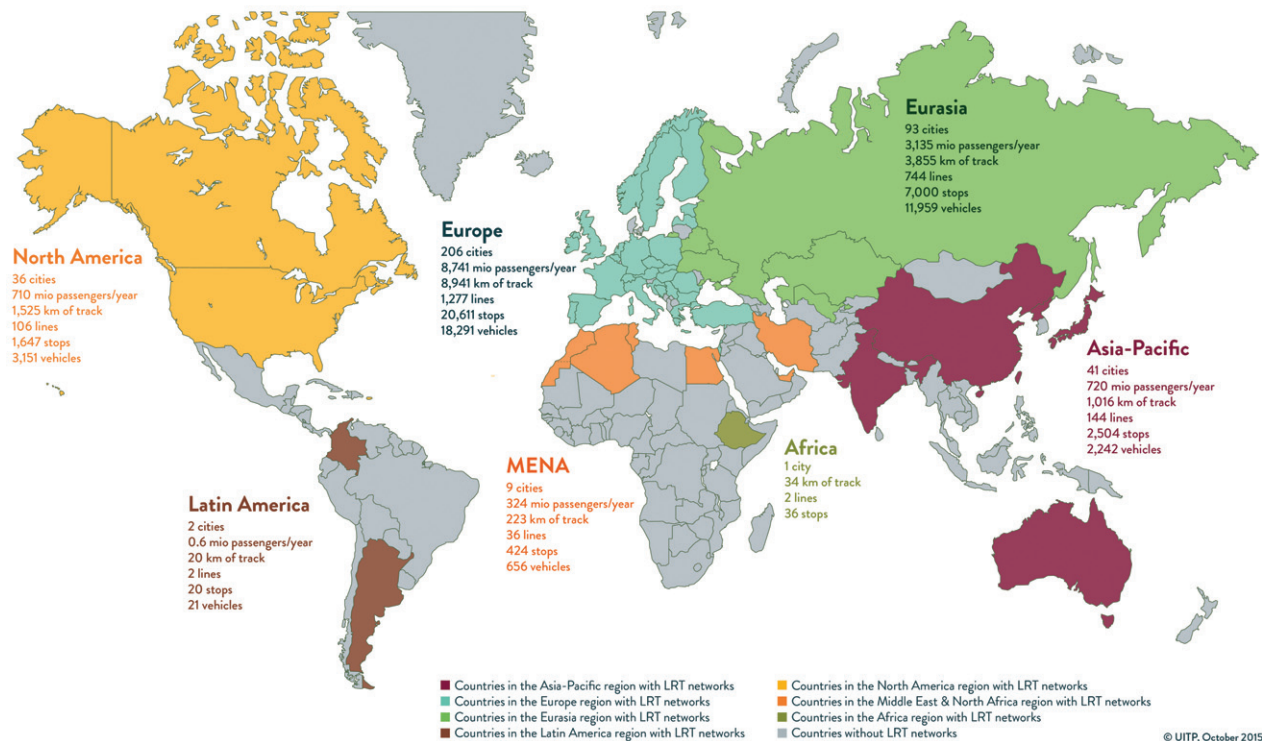
History

The origins of light rail are typically grounded in the development of streetcars in North America and Europe during the nineteenth century (in New York City 1832, Paris 1855, London 1861, Copenhagen 1863). In France, tram routes totaled more than 3400 km in length in their heyday during the early 20th century (Bouquet, 2017) and 55,150 streetcars were listed in the inventory operated by North American transit companies in 1930 (Culver, 2017). These early streetcars enabled city residents to move out of the center and live in the suburbs as well as opening new land for development outside of the city and providing vital connections between places. However, streetcars declined through the 1920–1960s as the burden of maintaining and replacing rail track contributed to a rise in bus services as a key means of public transport, while the rise of automobile sales, particularly post World War II, further diminished

Table 1 Common characteristics of light rail.

Running ways	Exclusive rail transit-ways, surface, underground or overhead; maximum gradients 8%–10%; priority signaling where rails cross/share road space.
Stations/stops	Large stations with security and passenger information systems; sometimes transit interchanges; sometimes space for cafes and shops; park-and-ride facilities. Medium distance between stops 250–1200 m.
Service design	Frequent services (max. 40–60 trains/h); often integrated with other transport services; high capacity (8000–15,000 passengers/h); average speed 20–40 km/h.
Fare collection	Off-board prepayment; smart/travel cards.
Additional features	Passenger information systems; multi-door and level-boarding; branding; modern design.

Source: compiled by the author with data from Kotoś and Taczanowski (2016).

**Figure 3** Light rail transit around the world. Source: UITP 2015. Reproduced with permission.

the need for and provision of streetcars. With urban investment directed towards the automobile and buses, many streetcar lines were abandoned. By 1960, only 15 systems continued to operate in France, while in North America the number of streetcars had declined to 2856. Germany is often given as an example where streetcar infrastructure continued.

Renaissance

The term “light rail” emerged with a renaissance of streetcars in North America and European countries in the 1970s. Employed in 1972 by the US Mass Transportation Administration (now the US Federal Transit Administration (FTA)) to describe streetcar transformations, light rail was more formally defined in 1989 in the Transportation Research Board’s Urban Public Transportation Glossary. By that time, the well known pioneers of the streetcar renaissance—Edmonton, Canada, 1978; San Diego, US, 1981, Nantes and Grenoble, France, 1985 and 1987 respectively—had been joined by other cities, and investment in light rail was growing quickly (UITP, 2015). Reinvestment in this renaissance was largely a response to concerns over urban traffic congestion and pressures for policy-makers to invest in new and improved public transport for environmental reasons. Light rail became popular; however, it is recognized as more than just a transport mode and its potential to accompany or facilitate urban

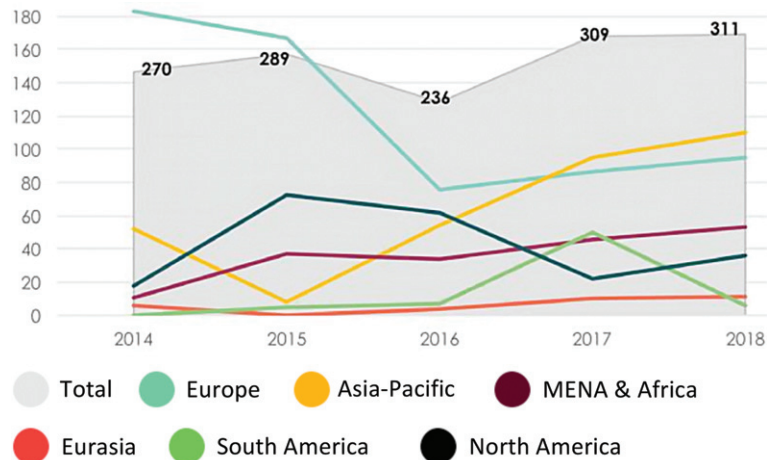


Figure 4 Evolution of LRT development by regions (km). Source: UITP 2019. Reproduced with permission.

development has made it very attractive for cities pursuing economic growth. This economic logic is a key reason for investment alongside improving public transport and alleviating traffic congestion.

The Promise of Light Rail

Light rail's potential and actual impacts are key concerns for professionals and academics. Commonly expressed as direct impacts or user benefits on the one hand, and indirect impacts or non-user benefits on the other, assessments of what light rail could and has achieved form important assessment criteria in the planning, decision-making, and evaluation of light rail investment.

User Benefits

Direct benefits acknowledge that light rail usually provides an attractive high-capacity and high-quality service to and through the city center in terms of frequency, speed, and reliability, particularly at peak hours (Bhattacharjee and Goetz, 2012; McAslan, 2019). Light rail can improve accessibility through new connections and reduced journey times. Light rail can also increase use of public transport (via newly generated trips as well as capture of users from other modes) though not all cities report this success (Engerbretsen et al., 2017; Senior 2009). Common performance measures include farebox revenues, journey times and reliability, ridership and travel behavior including individuals' access to light rail stops and stations, and modal shift of passengers, particularly from cars (Ingvardson and Nielsen, 2018). Some researchers have found that light rail can help to bring about modal shift with positive evidence from light rail lines in Nantes (17%–37%), San Diego (30%–50%) (Lee and Senior, 2013), and Manchester (21%) (Knowles, 1996). Other researchers find that modal shift has not occurred (Olesen, 2014).

Multimodal transport integration is another user benefit where light rail provides trunk and feeder services to help provide a seamless journey that includes first- and last-mile interconnections. Research has shown that other transit modes (e.g., buses, cars and bike-sharing) can be integrated effectively but there is also evidence that alternative modes act as competitors, particularly where they are in direct competition for fare revenues (Lee et al., 2017) (see Table 2).

Light rail user benefits also contribute to the perception of light rail as a sustainable transport mode (see below).

Non-User Benefits

Non-user benefits of light rail follow the broader classification of economic, environmental or social impacts though studies of economic impacts (wider economic impacts (WEIs)) significantly outnumber those analyzing environmental and social impacts and most focus on light rail's impact on development (land-use) and land and property values. Studies are based on the premise that transport facilitates the shape and extent of urban development and can lead to increased land values around stations and stops. A seminal report by Knowles and Ferbrache (2014) on the economic impacts on cities of investment in light rail provides a synthesized review of academic and non-academic literature and recommendations for boosting light rail's wider economic potential. Findings demonstrate that light rail can facilitate and enhance urban development through stimulating new growth, regeneration and redevelopment, and city center revitalization and boosterism as explained below.

Table 2 Strengths and limitations of light rail success in seven cities.

	<i>Strengths</i>	<i>Limitations</i>
St Louis Metrolink, US	Location of line and stations Provision of access over the river Integration with bus services Free city center journeys (off-peak) Security staff on board and at stations Provision of car parks at station sites	Weak and declining CBD Lack of comprehensive redevelopment project Over-allocation of station area space for car parks
San Diego Trolley, US	Location of the initial line Integration with bus services City center development project Other joint development projects Transit oriented development schemes	Weak integration of local plans with the trolley in some municipalities
Sacramento Light Rail, US	Strong and economically vital CBD Integration with bus services including free transfers between systems Pedestrianization of a city center street	Inconvenient and low density urban form Inconvenient route selection High income neighborhoods contra TOD Lack of funds to improve buses Poor service levels (low frequency) Provision of only one park and ride facility
Vancouver SkyTrain, Canada	Medium-high density urban area Substantial levels of public transport usage High frequency service Integration with bus services Integration with regional and local plans Redevelopment of old industrial sites Development bonuses (TOD incentives) Joint development schemes Relocation of government buildings to stations Early opening of a section for demonstration	
Tyne and Wear Metro, UK	High density urban area with radial corridors Substantial levels of public transport usage Location and length of lines Part of a city center redevelopment project Integration with bus services initially	Over time, lack of integration with buses Poor co-ordination with local plans and more recent urban projects
Manchester Metrolink, UK	High density urban area with radial corridors Substantial levels of public transport usage Provision of access through the city center High frequency Location of lines: replacing rail commuter service and providing city center rail link Renewal of some city center buildings Pedestrianization of a city center street	Lack of integration with bus services Poor integration with local municipal plans
Sheffield Supertram, UK	Medium density urban area with radial corridors Substantial levels of public transport usage Ticket sales on board Signaling priority improvement	Small and weak CBD System serving low income neighborhoods combined with competition from buses Low segregation and low signaling priority Demolition of high density residential areas Poor co-ordination with urban renewal project

Source: Modified from Knowles and Ferbrache (2014).

Light Rail and Urban Development

Light rail can help to stimulate new investment by raising the appeal and attractiveness of city locations as well as providing access to new sites. It was widely reported that the relocation of many of the British Broadcasting Corporation's (BBC) activities from London to the MediaCityUK development at Salford Quays rested on a £20 million public sector investment in an extension of the Manchester Metro link to MediaCityUK (Knowles and Binder, 2017). Several methodological caveats accompany evidence of new economic investment and some of the other potential benefits of light rail including: (1) linking specific economic investments solely to light rail remains problematic for it does not operate in isolation; (2) analytical scale is important for new investment may be based on relocation and thus investment decline elsewhere.

Light rail can facilitate and enhance urban regeneration and redevelopment including former derelict docklands (for example the London Docklands, UK), brownfield sites and industrial areas (e.g., Sunderland, UK). In this respect, light rail can open access to areas previously difficult to reach. Land was unlocked in this way in Copenhagen where Ørestad New Town (mixed-use



Figure 5 The pedestrianized and landscaped Place Garibaldi, Nice, France. Source: Myrabella/Wikimedia Commons/CC BY-SA 3.0 & GFDL.

development for educational, leisure, office, residential and retail use) was built around Copenhagen's new light rail mini Metro system on reclaimed land (Knowles, 2012).

Light rail has also been a catalyst for revitalization, notably in terms of car-oriented city centers transformed into public transit hubs with pedestrianized areas and landscaping to help stimulate interest and investment in the heart of the city (see Fig. 5). Grenoble is a leading example of this transformation and has become a model for other cities to follow under the popular label "the Grenoble effect." Revitalization effects have also been captured through the term "city boosterism" that reflects a competitive urbanism where light rail is used to boost a city's image (Ferbrache and Knowles, 2017). Through branding light rail vehicles (the 'M' logo of Newcastle and Tyne's Metro has become an emblem of the city), creating iconic infrastructure (such as light rail stations, bridges, tunnels, and grassed tracks), and designing light rail as an integral part of the city (trams in the French port city of Marseille are designed to look like the bow of a vessel, while burgundy-colored trams in Dijon are designed to represent the wider region's wine heritage), both the transit infrastructure and the city can become symbols of urban modernity. Other authors have expressed such design features in terms of place-making (Ferbrache and Knowles, 2017; Olesen, 2014; Olesen and Lassen, 2016).

Land and Property Values

By improving accessibility, light rail usually increases land and property values close to stations and stops. For example, the DLR triggered an increase in accessibility and land values in the Isle of Dogs, rising from £70,000 an acre in 1981 to £4.9 million an acre in 1988. A relatively significant number of studies examine light rail impacts on property values for residential and commercial buildings (see reviews in Knowles and Ferbrache, 2015; Ingvardson and Nielsen, 2018; Mohammad et al., 2013; Song et al., 2019) and while results vary, they tend to demonstrate an upward trend. Property and land premiums are usually expressed as positives under a neoliberal framework where they are valued for boosting economies or as potential mechanisms to help fund the very infrastructure that lies behind their value-rise. In this way, land and property value increases can be captured and reinvested although this is not always the case in practice (Cervero and Murakami 2009; Knowles and Ferbrache, 2014). Ørestad Metro in Copenhagen is one example that was partly funded by sale of public land at enhanced values to private developers. Reflecting a primary economic stance, and one which celebrates the benefits of light rail, analysis of property prices often neglects issues of displacement and gentrification (Bardaka et al., 2018).

While the above categories are not exclusive (new investment may be an essential part of redevelopment), they indicate significant ways that light rail has been valued. Given the number of light rail systems operating, available evidence and robust studies are relatively thin. Most literature examining economic impacts aim to tell success stories, suggesting that negative impacts may be overlooked or under-reported. There are also methodological issues involved with identifying potential impacts and a risk of assuming that light rail impacts can be isolated from other influential factors. Cervero (1984:146), for example, states that "For LRT to have an extensive impact on urban form, a strong and growing regional economy is an important prerequisite." Furthermore, light rail requires heavy capital investment, political and public commitment, and some systems are implemented cheaply and without adequate planning to bring about their potential benefits when tailored to the needs and design of specific cities. Table 2 provides examples of factors that determined the strengths and limitations of different light rail systems and this type of analysis is another important area of light rail research (Babalik-Sutcliffe, 2002; Mackett and Edwards, 1998; Mackett and Sutcliffe, 2003). Common within the literature is the argument that the most successful light rail schemes are those that are integrated with urban planning initiatives, particularly with transit oriented development (TOD) (Arrington 2009; Goetz 2019; Lavery and Kanaroglou, 2012).

Sustainability

Light rail is valued in terms of sustainability and "smart growth" strategies with emphasis on sustainable transport and mobility (infrastructural efficiency and resilience), sustainable (urban/economic) growth and environmental sustainability. Sustainability in terms of social equity has been largely neglected in comparison (Ferbrache and Knowles, 2016).

From an environmental perspective, light rail is electrically powered and therefore relatively low energy compared to other transit modes. It is expected to support reductions in energy use, noise and air pollution, including carbon emissions. However, while Park

and Sener (2019) observed lower levels of carbon monoxide and other emissions in Houston, US, little empirical evidence currently exists on air quality effects of light rail (FTA, 2016).

As a sustainable transport mode light rail can move relatively high passenger capacities at higher speeds in city center locations and its very characteristics and user-benefits enhance its reputation as one of the leading sustainable modes. Moreover, light rail is the transit mode at the center of many TOD projects which combine transport investment and urban planning to create resilient cities into the future (Arrington, 2009; Goetz, 2019; Knowles and Ferbrache, 2019). While TOD has often been justified in terms of economic growth potential, a more critical approach that pays attention to social mobility and liveability in this urban-transport development context holds potential for drawing attention to the possible social impacts of light rail.

An emergent area of research and practice, Mobility-as-a-Service (MaaS) has important implications for light rail in terms of enhancing passenger information systems, fare/payment systems, and multi-modal integration, including innovative modes such as aerial cableways and autonomous vehicles (Knowles et al., 2020). Research around this concept has the potential to stimulate further interest in light rail among researchers, and help to improve the efficiency, usability and sustainability of light rail going forward.

Social Exclusion and Gentrification

Grengs' (2005) claim that planners have lost sight of potential social benefits of mass transit resonates with literature and experience of light rail (Cuthill et al., 2016; Jones and Lee, 2016). Few studies have considered the social, health and wellbeing impacts of light rail. Exceptions include Farmer (2011) who demonstrated how light rail contributed to fragmentation and exacerbation of social divisions in Los Angeles. Nilsson and Delmelle (2018) found that impoverished neighborhoods were more likely to experience gentrification with transit station opening. Other researchers demonstrate light rail's socioeconomic achievements: Sari (2015) demonstrates that Bordeaux's tram has helped to reduce socio-economic inequalities through improving access to jobs. Cuthill et al. (2019) found that Croydon's Tramlink has also been able to help alleviate some social equity issues, particularly for young people. Adopting a different approach, Nolte and Yacobi (2015) demonstrate the powerful representations of light rail promoted in Jerusalem to unite divided Israeli and Palestinian populations. Such studies bring an important humanistic perspective to light rail that connects it with literature examining quality of life, liveability and the creation of healthier cities (Appleyard et al. 2019; Diemer et al. 2018), while at the same broadening what it means for light rail to be sustainable. A potential gap remains, however, between research of this nature and practical approaches to light rail planning and investment, which remain largely reliant on cost benefit analyses. In this way, the economic case for light rail remains pivotal in decision-making (Knowles and Ferbrache, 2014).

References

- Appleyard, B.S., Frost, A.R., Allen, C., 2019. Are all transit stations equal and equitable? Calculating sustainability, livability, health, & equity performance of smart growth & transit-oriented-development (TOD). *J. Transport Health* 14, 1–15.
- Arrington, G.B., 2009. Portland's TOD evolution: from planning to lifestyle. In: Curtis, C., Renne, J.L., Bertolini, L. (Eds.), *Transit Oriented Development: making it happen*. Ashgate Publishing, Farnham, UK and Burlington, Vermont, USA, pp. 109–124.
- Babalik-Sutcliffe, E., 2002. Urban rail systems: analysis of the factors behind success. *Transport Rev.* 22, 415–447.
- Bardaka, E., Delgado, M.S., Florax, R.J.G.M., 2018. Causal identification of transit-induced gentrification and spatial spillover effects: the case of the Denver light rail. *J. Transport Geogr.* 71, 15–31.
- Bhattacharjee, S., Goetz, A.R., 2012. Impact of light rail on traffic congestion in Denver. *J. Transport Geogr.* 22, 262–270.
- Bouquet, Y., 2017. The renaissance of tramways and urban development in France. *Miscellanea Geograph.* 21, 5–18.
- Cervero, R., 1984. Light rail transit and urban development. *J. Am. Plan. Assoc.* 50, 133–147.
- Cervero, R., Murakami, J., 2009. Rail and property development in Hong Kong: experiences and extensions. *Urban Stud.* 46, 2019–2043.
- Culver, G., 2017. Mobility and the making of the neoliberal "creative city": The streetcar as a creative city project? *J. Transport Geogr.* 58, 22–30.
- Cuthill, N., Cao, M., Lin, Y., Gao, X., Zhang, Y., 2019. The association between urban public transport infrastructure and social equity and spatial accessibility within the urban environment: an investigation of Tramlink in London. *Sustainability* 11, 1229.
- Diemer, M.J., Currie, G., de Gruyter, C., Hopkins, I., 2018. Filling the space between trams and places: adapting the 'movement and place' framework to Melbourne's tram network. *J. Transport Geogr.* 70, 215–227.
- Engerbretsen, Ø., Christiansen, P., Strand, A., 2017. Bergen light rail—effects on travel behaviour. *J. Transport Geogr.* 62, 111–121.
- Farmer, S., 2011. Uneven public transport development in neoliberalizing Chicago, USA. *Environ. Plan. A* 43, 1154–1172.
- Ferbrache, F., Knowles, R.D., 2016. Generating opportunities for city sustainability through investments in light rail systems: Introduction to the Special Section on light rail and urban sustainability. *J. Transport Geogr.* 54, 369–372.
- Ferbrache, F., Knowles, R.D., 2017. City Boosterism and Place-making with Light Rail Transit: a critical review of light rail impacts on city image and quality. *Geoforum* 80, 103–113.
- Federal Transit Administration (FTA), 2016. Transit's Role in Environmental Sustainability. <https://www.transit.dot.gov/regulations-and-guidance/environmental-programs/transit-environmental-sustainability/transit-role> (accessed 24.09.2019).
- Goetz, A., 2019. Effects of transit oriented development in Denver, Colorado, USA. In: Knowles, R.D., Ferbrache, F. (Eds.), *Transit oriented development and sustainable cities: economics, community and methods*. Edward Elgar, Cheltenham, pp. 96–117.
- Grengs, J., 2005. The abandoned social goals of public transit in the neoliberal city of the USA. *City* 9, 51–66.
- Hensher, D., 1999. A bus-based transitway or light rail? Continuing the saga on choice versus blind commitment. *Road Transport Res.* 8, 3–19.
- Hensher, D.A., 2016. Why is light rail starting to dominate bus rapid transit again? *Transport Rev.* 36, 289–292.
- Ingvardson, J.B., Nielsen, O.A., 2018. Effect of new bus and rail rapid transit systems—an international review. *Transport Rev.* 38, 96–116.
- Jones, C.E., Lee, D., 2016. Transit-oriented development and gentrification along Metro Vancouver's low-income SkyTrain corridor. *Canadian Geographer* 60, 9–22.
- Knowles, R.D., 1996. Transport impacts of Greater Manchester's Metrolink light rail system. *J. Transport Geogr.* 4, 1–14.

- Knowles, R.D., 2012. Transit oriented development in Copenhagen, Denmark: from the Finger Plan. *J. Transport Geogr.* 22, 251–261.
- Knowles, R.D., Binder, A., 2017. MediaCityUK: A Sustainable Transit-Oriented Development, Chapter 1, 3–12. In: Theakstone, W. (Ed.), *Manchester Geographies*. Manchester Geographical Society, Manchester, UK.
- Knowles, R.D., Ferbrache, F., 2014. An investigation into the economic impacts on cities of investment in light rail systems. A Report for UK Tram, Centro, Birmingham, UK.
- Knowles, R.D., Ferbrache, F., 2015. Evaluation of wider economic impacts of light rail investment on cities. *J. Transport Geogr.* 54, 430–439.
- Knowles, R.D., Ferbrache, F. (Eds.), 2019. *Transit oriented development and sustainable cities: economics, community and methods*. Edward Elgar, Cheltenham, UK.
- Knowles, R.D., Ferbrache, F., Nikitas, A., 2020. Transport's historical, contemporary and future role in shaping urbandevelopment: Re-evaluating transit oriented development. *Cities* 99, 1–11.
- Kołos, A., Taczanowski, J., 2016. The feasibility of introducing light rail systems in medium-sized towns in Central Europe. *J. Transport Geogr.* 54, 400–413.
- Lavery, T.A., Kanaroglou, P., 2012. Rediscovering light rail: assessing the potential impacts of a light rail transit line on transit oriented development and transit ridership. *Transport. Lett.* 4, 211–226.
- Lee, J., Boarnet, M., Houston, D., Nixon, H., Spears, S., 2017. Changes in service and associated ridership impacts near a new light rail transit line. *Sustainability* 9, 1827.
- Lee, S.S., Senior, M.L., 2013. Do light rail services discourage car ownership and use? Evidence from Census data for four English cities. *J. Transport Geogr.* 29, 11–23.
- Mackett, R., Edwards, M., 1998. The impact of new urban public transport systems: will the expectations be met? *Transport. Res. A* 32, 231–245.
- Mackett, R., Sutcliffe, E.B., 2003. New urban rail systems: a policy-based technique to make them more successful. *J. Transport Geogr.* 11, 151–164.
- McAslan, D., 2019. Planning an effective transport system: learning from resident transit use behaviour and perspectives. In: Knowles, R.D., Ferbrache, F. (Eds.), *Transit Oriented Development and Sustainable Cities: Economics, Community and Methods*. Edward Elgar, Cheltenham, pp. 136–150.
- Mohammad, S.I., Graham, D.J., Melo, P.C., Anderson, R.J., 2013. A meta-analysis of the impact of rail projects on land and property values. *Transport. Res. Part A* 50, 158–170.
- Nilsson, I., Delmelle, E., 2018. Transit investments and neighbourhood change: on the likelihood of change. *J. Transport Geogr.* 66, 167–179.
- Nolte, A., Yacobi, H., 2015. Politics, infrastructure and representation: the case of Jerusalem's light rail. *Cities* 43, 28–36.
- Olesen, M., 2014. Framing light rail projects—case studies from Bergen, Angers and Bern. *Case Stud. Transport Pol.* 2, 10–19.
- Olesen, M., Lassen, C., 2016. Rationalities and materialities of Light Rail Scapes. *J. Transport Geogr.* 54, 373–382.
- Park, E.S., Sener, I.N., 2019. Traffic-related air emissions in Houston: effects of light-rail transit. *Sci Total Environ.* 651, 154–161.
- Sari, F., 2015. Public transit and labor market outcomes: analysis of the connections in the French agglomeration of Bordeaux. *Transport. Res. A* 78, 231–251.
- Senior, M.L., 2009. Impacts on travel behaviour of Greater Manchester's light rail investment (Metrolink Phase 1): evidence from household surveys and Census data. *J. Transport Geogr.* 17, 187–197.
- Song, Z., Cao, M., Han, T., Hickman, R., 2019. Public transport accessibility and housing value uplift: evidence from the Docklands light railway in London. *Case Stud. Transport Pol.* 7, 607–616.
- UITP, 2015. Light rail in figures. Statistics Brief. Worldwide outlook. http://www.uitp.org/sites/default/files/cck-focus-papers-files/UITP_Statistic_Brief_4p-Light%20rail-Web.pdf (accessed 28.08.2019).
- UITP, 2019. New Urban Rail Infrastructure 2018 <https://www.uitp.org/sites/default/files/cck-focus-papers-files/Statistics%20Brief%20-%20NEW%20URBAN%20RAIL%20INFRASTRUCTURE%202018.pdf> (accessed 28.08.2019).

Further Reading

- Knowles, R.D., 2007. What future for light rail in the UK after the Ten Year Transport Plan targets are scrapped? *Transport Policy* 14, 81–93.

Planning Tourism Travel

Luca Zamparini, Department of Law, University of Salento, Lecce, Italy

© 2021 Elsevier Ltd. All rights reserved.

Traditional Planning Model for Tourism Travel	39
The Role of Transport in Planning Tourism Travel	39
Factors Affecting Accessibility and Planning in Tourism Travel	40
The Current Trends in Planning Tourism Travel	41
The Role of ICT for Tourism Travel Planning	41
The Relevance of Social Networks and User-Generated Content for Tourism Travel Planning	42
Conclusions	42
References	42
Further Reading	43

Traditional Planning Model for Tourism Travel

Every tourism experience can analytically be segmented into its five component phases (anticipation, outward journey, experience phase, return journey, and memory). The planning of tourism travel represents the main activity that characterizes the anticipation phase. A model proposed by [Bansal and Eiselt \(2004\)](#) considers travel planning as the interaction of motivations and constraints that leads first to the selection of a determined region and consequently to the transport means, accommodation, and all other complementary goods and services that compose the tourism experience.

Several studies have tried to analyze the motivations that may lead a tourist to plan a determined holiday or vacation (for a review, see [Yousaf et al., 2018](#)). A seminal paper by [Maslow \(1943\)](#) has proposed a general motivation theory based on five levels of needs. The first level is constituted by physiological needs (e.g., food, warm, and rest) that in the case of tourism determine the basic expectations of people traveling toward a determined destination. The second level is related to safety and security needs. It is evident that countries or destinations which do not offer adequate levels of security will experience drastic decreases in tourist demand, unless they propose some *enclaves* with physical and technological barriers between visitors and residents. These first two levels constitute the basic needs that destinations have to supply in order to be the possible target of any tourism travel planning. The third level attains the belongingness and love needs (as intimate relationships and friendships). In the case of tourism travel, this relates to the attitude, especially of neophilic tourists, to try and become familiar with the local communities of the regions that they visit. This specific need determines also the cultural tourism demand where the main aim of the travel is the desire to acquire knowledge about other cultures and countries. The fourth level relates to esteem needs. In the case of tourism, travel is performed in order to impress the group of peers or family and friends. Lastly, the fifth need is the feeling to have accomplished one's objectives (self-actualization or achieving one's potential, by also engaging in creative activities and performances). This motivates some tourists to travel to challenging destinations and to perform demanding activities. Two variants of Maslow's theory have been presented by [Pearce \(1988\)](#) and by [Pearce and Lee \(2005\)](#), listing all possible motivational factors that drive tourist travel planning.

A different theory was presented by [Dann \(1997\)](#) who made a division between push factors that are internal to the tourist and are related to the need to rest and escape from the everyday setting and pull factors that emerge from destinations, which present a series of attractions, amenities, activities, and ancillary services that motivate the tourists toward them.

Several authors have proposed a taxonomy of possible constraints to tourism travel planning (see, among others, [Nyaupane and Andereck, 2008](#)). It is possible to consider three possible categories of constraints: intrapersonal, interpersonal, and structural constraints. Intrapersonal constraints are related to the psychological conditions of individuals, which may change even in very short periods of time and that may lead to the willingness to avoid the participation to a travel activity. Possible examples are represented by anxiety, depression, lack of interest, or stress. Interpersonal constraints consider the situations in which an individual is part of a traveling group (family, friends, or colleagues), and consequently all travel decisions must be bargained within the group with the result that individual preferences are only very seldom met. In other cases, it may happen that a potential traveler is not able to find a traveling companion or group for her desired tourism experience. Lastly, structural constraints refer to all the external factors that are interposed between preferences and participation to a travel. The most cited examples of external constraints are the scarcity of time and money. However, many others should be considered such as adverse weather conditions (which determines a remarkable seasonality of demand for many tourism destinations), the lack of effective marketing of some countries or regions, and the low degree of accessibility that in some cases is due to the inefficient transport services and planning.

The Role of Transport in Planning Tourism Travel

One of the most important drivers of the evolution of tourism activities in the past decades is represented by the increased availability and affordability of transport services (see, among others, [Schiefelbusch et al., 2007](#)). This has led to a marked upward trend in

tourism travel, both in terms of volumes and distances. As an example, international tourists have passed from 25 million in 1950 to 1.4 billion in 2018. The drastic increases in tourism travel and mobility have determined momentous consequences for the natural and human environments where the transit regions have borne the burden of transport externalities without any sensible economic advantage. It would then be important to plan tourism travel by taking into consideration all possible measures that may lead to sustainable management of transport activities. Such measures would likely fall in one or more of the following six categories: alternative fuels/energy sources, improved energy efficiency and technology, investments in infrastructure, operational measures, behavioral measures, and taxes and subsidies (Peeters et al., 2019). However, sustainable transport measures should also take into account the effects of a reduction in the number of trips, given that tourism represents an important source of revenues, jobs, and gross domestic product growth. More sustainable tourism travel (which may imply slower journeys and longer periods of staying once at destination) may be feasible only for a limited, privileged, share of citizens. Most of the population may deem these sustainable tourism travel activities either as undesirable or unaffordable in terms of time and/or of economic budget. It is then evident that the development of tourism travel depends on the degree of accessibility of a determined area and of the planning activities that have been undertaken to increase it. Litman (2008) has proposed a taxonomy of all the factors that affect accessibility and, in the case of tourism, determine the likelihood of people to choose a determined destination in their planning activities. The next subsection will offer a general description of each of these factors and will then contextualize them to the case of tourism travel planning.

Factors Affecting Accessibility and Planning in Tourism Travel

The following is a description of the factors affecting accessibility and that are of paramount importance when planning tourism travel. One of the factors considered in Litman's taxonomy, mobility substitutes or the telecommunications and mobility services that substitute for physical travel, will not be considered given the lack of relevance in tourism travel. The general descriptions will be based on Litman (2008), while the discussions of their relevance for tourism travel are the author's elaborations.

- *Transport demand* refers to the amount of mobility and access that people would choose. In the case of tourism travel, it considers the amount of tourism demand for a determined destination. For the large majority of destinations, it is important to consider the seasonal pattern of demand, characterized by peak periods, shoulder periods, and low demand periods. For planning purposes, it is important to forecast tourism demand. This can be accomplished by using time series analysis, autoregressive moving average methods, support vector machines, or neural networks models.
- *Mobility* considers the travel speed and the distance to be covered. In general transport literature, this topic has been considered as travel time and travel reliability. In the case of tourism travel, this issue has two possible meanings. On the one hand, it may be related to the amount of time that is necessary to reach the desired tourism destination (mobility toward tourism destinations). On the other hand, it may be directed at accounting for the amount of time required to move within the destination in order to perform the activities that characterize a tourism experience.
- *Transport options (modes)* indicates the quality of the various transport options, including walking, cycling, public transport, and private motorized transport modes. In the case of tourism travel, the availability of various transport modes has to be carefully considered given that at the tourism destinations there may be options that are not normally part of the possible choice set in everyday life. Moreover, during holiday periods, the amount of discretionary travel increases substantially with respect to required or forced travel activities.
- *User information* represents the availability of reliable information on mobility and accessibility options. The availability of information is a particularly important topic to consider when analyzing the planning of tourism travel. In this context, both information and communication technologies (ICT) and smart technologies and social networks have assumed pivotal roles. Their relevance will constitute the object of the analysis in the following section "The Current Trends in Planning Tourism Travel."
- *Integration* denotes the degree of connectivity among transport system links and nodes. In the case of tourism travel, it is very important to create an integrated transport system that allows for tourists to reach their desired destination in a reasonable amount of overall traveling time. This is because the travel from the place of residence to destination requires the use of multiple transport modes (e.g., plane and bus, plane and train, and so on), unless people decide to travel with their own private motorized vehicle. In this context, the coordination and the partnerships between private stakeholders and public administrations can determine positive synergies and enhance the degree of integration among transport modes.
- *Affordability* refers to the cost of mobility for individuals as a share of their relative incomes. This increased economic affordability has most probably been the main determinant to move tourism from a luxury activity available only for a very limited share of the overall population to a possibility for the vast majority of people in developed countries with the ensuing development of mass tourism travels.
- *Land use factors* consider how land use density and mix affect travel distances and costs. Within tourism, the land use affects the degree of satisfaction of tourists once they are at destination. Moreover, it is an important element of the relationships between tourists and residents. In some cases, these two categories share the same places, with the emerging positive and negative externalities. In other destinations, according to the social exchange theory, land use planning determines a clear distinction between the areas where tourists interact with each other and the zones where residents go to in order to avoid tourists.
- *Transport network connectivity* concerns the density of roads and path connections and then the directness of travel between destinations. This is an issue where tourism travel is only part of a more general planning activity by the public administrations

that need to take into account the needs and requirements of a large number of stakeholders. This is a long-term concern where it is possible to observe a remarkable path dependency. In the case of tourism, accurate investments may favor the growth of complementary forms of tourism with sensible externalities for all the interested adjoining areas.

- *Transport management* designates how the rules and regulations issued at the various administrative levels impact on the accessibility of a determined destination. This is a complementary issue with respect to land use factors and transport network connectivity. An appropriate transport management ensures that the most efficient (and possibly environmentally and socially sustainable) alternative is available for tourists at the right place and at the right time. Tourist transport management is, in general, connected to public administration activities, but it can also stem from private initiative.
- *Prioritization* implies strategies that favor more efficient travel activities. This constitutes an important element in the overall tourism development strategies, given that the prioritization of specific infrastructures and services can determine the development path of the tourism phenomenon in a destination.
- *Inaccessibility* denotes the value of inaccessibility and isolation, which in some cases requires to limit the access possibilities. For some niches of the tourism travel, such as adventures and elite forms of tourism, the inaccessibility is considered as a positive asset. In all other cases, the degree of inaccessibility should be reduced as much as possible so that the destination can increase its market share and the satisfaction of tourists.

A review proposed by [Van Truong and Shimizu \(2017\)](#) has ascertained that, despite the relevance of all these factors for the development of tourism travel, very few studies have taken them into consideration when estimating the effect of transport on tourism travel. Some studies have considered infrastructure improvements, information provision, and travel costs while most of the others described factors have been neglected and would deserve to be included in the analysis of tourism travel planning.

The Current Trends in Planning Tourism Travel

The passage of most of the goods and services that compose the tourism activities from experience goods to search goods is one of the most relevant methodological shifts that have characterized the tourist sector in the past decades. An experience good is a consumption good (or service) whose quality and peculiar characteristics can only be asserted by the tourist only after its consumption. On the other hand, a search good is a consumption good (or service) whose quality and peculiar characteristics can be asserted by the tourist before its purchase and consumption. This shift has momentous consequences on the planning of tourist travel. Two of the most important determinants of this trend are represented by the diffusion of ICT and by the increased relevance of social networks and user-generated content, originating from the use of these new technologies. The following two subsections will present a discussion of these two factors, emphasizing their role in the planning activities of tourists.

The Role of ICT for Tourism Travel Planning

A study by [Money and Crofts \(2003\)](#) has stated the principle that the amount of activities devoted to the acquisition of information is limited to the point in which, coherently with general economic theory, the marginal benefit of a further act of search is higher than the marginal cost. Moreover, search is partitioned in internal search (which identifies the retrieval of previously acquired information) and in external search. In the latter case, the sources of information can either be: (1) personal (e.g., previous experiences of friends and relatives), (2) marketing or advertisements, (3) neutral by third parties who are not directly interested in the act of purchase (see section “The Relevance of Social Networks and User-Generated Content for Tourism Travel Planning”), or (4) experiential through direct contacts with the retailer. Important sources of external search that may in some cases be linked to marketing and advertisement practices and in others to the neutral information sources and that have taken a leading role in tourism travel planning are the ICT and the smart technologies ([Steen Jacobsen and Munar, 2012](#)). Their main benefit for the tourist who has to plan the travel is represented by the reduction of search costs and the consequent possibility to perform a larger number of searches in a determined period of time. Moreover, they allow the option to double-check the information obtained through traditional channels, such as travel agents and guidebooks. [Xiang et al. \(2015\)](#) have characterized the use of ICTs for travel planning as an adaptive behavior that tries to take advantage of the new opportunities provided by these new search tools. They have then identified various trends that have emerged from the use of ICTs for travel planning. The first one considers that the use of ICTs for travel planning is saturated and represents the main source of information. Within all products that compose the tourism experience, ICTs are particularly used for the choice and purchase of travel tickets, accommodation, and car rental. The problems that had marked the early phases of ICT adoption for tourism travel planning (e.g., poor usability and lack of personalized services) appear to have been overcome by their evolution. The second trend is related to the use of ICT for planning purposes by all age cohorts, although the younger ones appear to be more engaged, and to choose from a wider range of options. A further trend is constituted by the adoption of ICT tools also for the search and purchase of secondary products (e.g., museum tickets and tourism shopping). [Derrick Huang et al. \(2017\)](#) have clarified that the main attributes of ICT for travel planning are informativeness (quality and trust of information), accessibility (easiness to access appropriate information), interactivity (real time feedbacks and active communications), and personalization (possibility to obtain information that is coherent with personal travel planning needs). They then conclude that the use of ICT allows to consider a much wider range of alternatives and to enhance the satisfaction of tourists. The issue of tourists’ satisfaction linked to the use of ICTs has also been carried out by [Woo Yoo et al. \(2017\)](#). They conclude that the level

of satisfaction and the quality of information obtained by using ICTs for travel planning is positively correlated with the skills and confidence of ICTs possessed by the tourist who use them.

The Relevance of Social Networks and User-Generated Content for Tourism Travel Planning

The transformation of the informal word of mouth channels to retrieve efficient information for the tourism travel planning to a formal set of websites, apps, and social networks has been one of the effects of the widespread use of Internet (Cox et al., 2009). One of the main issues originating from the diffusion of this new tool to acquire information was related to the degree of trustworthiness that the provided information may have. This is evident for both consumers and suppliers of tourism goods and services. The latter must take into account the increasing popularity of these new planning tools and must consequently monitor the information provided about their firms, services, and establishments. A study by Fotis et al. (2012) has shown that social media are used throughout the entire process of travel planning and that their influence is higher on the destination and accommodation choices. The degree of trustworthiness is second only to the information gathered through friends and relatives. Lastly, it appears evident that there is a high rate of heterogeneity depending on the various considered national markets. Another source of heterogeneity has been proposed by Simms (2012) who has highlighted that some of the trip characteristics (i.e., location of the destination, familiarity with the destination, and travel party composition) influence the engagement of tourists in user-generated content for travel planning. Ayeh et al. (2013) have suggested that ease of use, perceived enjoyment, and positive attitude play an important role in the use of social media and user-generated content for travel planning. On the other hand, complexity and the perceived effort have negative effects on the likelihood of travel information searches through social media (Chung and Koo, 2015). A last important factor is represented by the perceived similarity of interests between the generator and the user of relevant information. A general theory has been proposed by Mendes-Filho et al. (2018) by jointly taking into account psychological empowerment and technology acceptance. The former considers the capacity of people to develop efficient decision making and the discretion to carry out a specific behavior. The latter comprises perceived usefulness and ease of use of the technological tools.

A different tourism travel planning issue has been studied by Cenamor et al. (2017) who have proposed a possible tourism tool that mediates between social networks and smart travel technologies. The system would create personalized excursions and tourism routes, based on a series of parameters introduced by the tourist (e.g., length of stay and specific cultural interests), by using the consumer-generated information gathered in a traveling social network. This interaction of technology and social networks may represent an important evolution that may greatly enhance the tourism travel planning options based on the specific preferences of the individuals.

Conclusions

The article has shown that tourism travel planning is based on the interactions between motivations and constraints, which lead to the selection of a determined destination and of the activities to perform once it is reached. Motivations may be categorized either as a multilayered series of needs or as the interaction of push and pull factors. On the other hand, constraints can either be intrapersonal, interpersonal, or structural. Within the latter category, transport constitutes a pivotal element in planning tourism travel. The past decades have been characterized by a remarkable increase in the availability and affordability of transport services that have allowed a wide diffusion of tourism activities and that have called for an effort to enhance the share of sustainable tourism travel activities. A related important issue is the degree of accessibility of tourist regions, in terms of travel both toward and within the destination. All factors affecting accessibility (transport demand, mobility, transport options, user information, integration, affordability, land use, transport network connectivity, transport management, prioritization, and inaccessibility) represent topic elements in tourism travel planning for both private actors and public administrations. They are particularly important elements to consider in periods marked by disruptive events, such as the COVID-19 pandemic whose effects on transport and tourism have started to be manifest in 2020. In this context, ICTs, social networks, user-generated content, and their interactions may demonstrate to be more and more essential, given the degree of informativeness and of interactivity that they allow to achieve. Moreover, they permit to reduce the search costs for the best possible alternative, and they are characterized by a high degree of trustworthiness among users. All the aforementioned factors of accessibility would deserve to be jointly considered in the future analyses of tourism travel planning in order to propose robust models and sound evaluations of current practices and activities.

References

- Ayeh, J.K., Au, N., Law, R., 2013. Predicting the intention to use consumer-generated media for travel planning. *Tour. Manag.* 35, 132–143.
- Bansal, H., Eiselt, H.A., 2004. Exploratory research of tourist motivations and planning. *Tour. Manag.* 25, 387–396.
- Cenamor, I., de la Rosa, T., Nunez, S., Borrajo, D., 2017. Planning for tourism routes using social networks. *Expert Syst. Appl.* 69 (1), 1–9.
- Chung, N., Koo, C., 2015. The use of social media in travel information search. *Telemat. Inform.* 32, 215–229.
- Cox, C., Burgess, S., Sellitto, C., Buultjens, J., 2009. The role of user-generated content in tourists' travel planning behavior. *J. Hosp. Mark. Manag.* 18, 743–764.
- Dann, G.M.S., 1977. Anomie, ego-enhancement and tourism. *Ann. Tour. Res.* 4 (4), 184–194.
- Derrick Huang, C., Goo, J., Nam, K., Woo Yoo, C., 2017. Smart tourism technologies in travel planning: the role of exploration and exploitation. *Inform. Manag.* 54, 757–770.

- Fotis, J.N., Buhalis, D., Rossides, N., 2012. Social media use and impact during the holiday travel planning process. In: Fuchs, M., Ricci, F., Cantoni, L. (Eds.), *Information and Communication Technologies in Tourism*. Springer-Verlag, Vienna, pp. 13–24.
- Litman, T., 2008. Evaluating accessibility for transportation planning. Victoria Transport Policy Institute. Available from: <http://www.vtpi.org/access.pdf> (accessed 28.07.20.).
- Maslow, A.H., 1943. A theory of human motivation. *Psychol. Rev.* 50, 370–396.
- Mendes-Filho, L., Mills, A.M., Tan, F.X., Milne, S., 2018. Empowering the traveler: an examination of the impact of user generated content on travel planning. *J. Travel Tour. Mark.* 35 (4), 425–436.
- Money, R.B., Crofts, J.C., 2003. The effect of uncertainty avoidance on information search, planning and purchases of international travel vacations. *Tour. Manage.* 24 (2), 191–202.
- Nyaupane, G.P., Andereck, K.L., 2008. Understanding travel constraints: application and extension of a leisure constraints model. *J. Travel Res.* 46 (4), 433–439.
- Pearce, P.L., 1988. The Olysses factor: evaluating tourists in visitor's settings. *Ann. Tour. Res.* 15 (1), 1–28.
- Pearce, P.L., Lee, U.I., 2005. Developing the travel career approach to tourist motivation. *J. Travel Res.* 43 (3), 226–237.
- Peeters, P., Higham, J., Cohen, S., Eijgelaar, E., Gossling, S., 2019. Desirable tourism transport futures. *J. Sustain. Tour.* 27 (2), 173–188.
- Schiefelbusch, M., Jain, A., Schäfer, T., Müller, D., 2007. Transport and tourism: roadmap to integrated planning developing and assessing integrated travel chains. *J. Transp. Geogr.* 15, 94–103.
- Simms, A., 2012. Online user generated content for travel planning—different for different kinds of trips? *e-Rev. Tour. Res.* 10 (3), 76–85.
- Steen Jacobsen, J.K., Munar, A.M., 2012. Tourist information search and destination choice in a digital age. *Tour. Manag. Perspect.* 1, 39–47.
- Van Truong, N., Shimizu, T., 2017. The effect of transportation on tourism promotion: literature review on application of the computable general equilibrium model. *Transp. Res. Procedia* 25, 3096–3115.
- Woo Yoo, C., Goo, J., Derrick Huang, C., Nam, K., Woo, M., 2017. Improving travel decision support satisfaction with smart tourism technologies: a framework of tourist elaboration likelihood and self-efficacy. *Technol. Forecast. Soc. Change* 123, 330–341.
- Xiang, Z., Magnini, V.P., Fesenmaier, D.R., 2015. Information technology and consumer behavior in travel and tourism: insights from travel planning using the internet. *J. Retail. Consum. Serv.* 22, 244–249.
- Yousaf, A., Amin, I., Santos, J.A.C., 2018. Tourists' motivations to travel: a theoretical perspective on the existing literature. *Tour. Hosp. Manag.* 24 (1), 197–211.

Further Reading

- Dickinson, J.E., Ghali, K., Cherrett, T., Speed, C., Davies, N., Norgate, S., 2014. Tourism and the smartphone app: capabilities, emerging practice and scope in the travel domain. *Curr. Issues Tour.* 17 (1), 84–101.
- Edgell, D.L., DelMastro Allen, M., Smith, G., Swanson, J.R., 2008. *Tourism policy and planning*. In: *Yesterday, Today and Tomorrow*, Routledge, London.
- Michael Hall, C., 2008. *Tourism Planning. Policies, Processes and Relationships*. Pearson, Essex.
- Vassakis, K., Petrakis, E., Kopanakis, I., Makridis, J., Mastorakis, G., 2019. Location-based social network data for tourism destinations. In: Sigala, M., Rahimi, R., Thelwall, M. (Eds.), *Big Data and Innovation in Tourism, Travel and Hospitality. Managerial Approaches, Techniques and Applications*. Springer Nature, Singapore, pp. 105–114.
- Wu, X., Guan, H., Han, Y., Ma, J., 2017. A tour route planning model for tourism experience utility maximization. *Adv. Mech. Engineer.* 9 (10), 1–8.

Transport Planning and Management and its Implications in Chinese Cities

Mengqiu Cao^{*,†}, ^{*}School of Architecture and Cities, University of Westminster, London, United Kingdom; [†]Bartlett School of Planning, University College London, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	44
The History of Transport Development in China	44
Current Transport Planning and Management Problems and Strategies	45
Current Situation	45
Discussion of Current Transport Planning and Management Strategies and Their Implications	46
Conclusions	48
Acknowledgments	49
References	49
Further Reading	50

Introduction

In general, transport planning over the past 50 years has lacked a clear theoretical foundation ([Banister, 2002](#)). Transport-related problems, such as climate change, air pollution, energy consumption, traffic congestion, accidents and safety issues, and transport inequalities have become apparent due to the increased demand for unsustainable transport and inefficient transport planning and management, particularly in most developing countries. As a growing number of people are living in urban areas, one of the key factors in terms of creating successful and sustainable cities involves developing a sustainable transport system which is able to meet the needs of rapid economic growth, as well as caring for the environment, and improving the quality of people's lives ([Attard and Shiftan, 2015](#); [Hickman and Banister, 2014](#)). As the world's largest developing country, China has been significantly affected by the aforementioned issues. This chapter aims to enrich our general understanding of the status quo and the current strategies that have been implemented in terms of transport planning and management in Chinese cities, as well as offering new insights into the debate about how to improve transport users' perceptions, an aspect which has long been overlooked.

The next section describes the history of transport development in China, while the third section provides an overview of the status quo regarding transport planning and management issues, as well as critically discussing some of the strategies that have been implemented, and their implications for Chinese cities. The conclusion summarizes the key findings and contributions, as well as offering reflections.

The History of Transport Development in China

Since the Chinese revolution of 1949, and particularly since the reform and opening up policy, instigated by Xiaoping Deng, was introduced in December 1978, transport has undergone historic changes, which have made a significant difference to people's lives in terms of economic and social development, travel safety and increased mobility. Since the beginning of the 21st century, the Chinese government has comprehensively extended the reform of transport, accelerated the construction of a modern and comprehensive transport system, and continuously promoted the modernization of the governance system and governance capacity of the transport industry.

In the mid-twentieth century, the development of the Chinese transport system lagged far behind that of the rest of the world. For example, the total mileage covered by the national railway amounted to only 21,800 km, half of which was virtually unusable. Furthermore, there were only 80,800 km of roads and 51,000 private vehicles in the country at that time. Inland waterways were in need of improvement, while a mere 12 civil aviation lines served the whole country. There were fewer postal service points due to the poor quality of freight transport. The main means of transport were rickshaws and wooden sailboats ([State Council Information Office, 2016](#)).

Since 1953, the construction of transport infrastructure has been carried out in a planned manner. In 1956, the General Administration of Highways, a department of the Ministry of Communications, appointed experts to study the first draft of highway planning in China. During the first and second 5-year plans and the adjustment of the national economy from 1953 to 1965, the state started investing in the transport infrastructure, renovating it and building a number of railways, roads, ports and civil airports, thus increasing the coverage of the transport infrastructure in western China and remote areas. Dredging of the main waterways also took place, as well as the opening up of new international and domestic waterways and air routes, expansion of the postal network, and a substantial increase in the amount of transport equipment. During the Cultural Revolution (The Cultural Revolution was a

socio-political movement in China) between 1966 and 1976, the development of transport was significantly disrupted, but the scale of facilities, equipment and transport lines was still increased overall (ibid.).

In 1978 the Chinese economic reform (known as the “Opening of China”) heralded a new chapter, particularly in terms of economic and social development. During this period, the Chinese government prioritized the development of transport, strengthened policy support, undertook groundbreaking work in liberalizing the transport market and established socialized financing mechanisms. In 1982, it was pointed out that, “if people want to be rich, they need to build the road first.” This was first done in Meishan county, in Sichuan Province, and then across the whole of China. In 2007, the China railway speed-up campaign was implemented in six rounds. Comprehensive highway and water transport planning was carried out in relation to the main skeleton of the road network, the key water transport routes, and the main ports and station hubs and their support systems. In 2008, the Ministry of Transport was set up, and substantive steps were taken toward reforming the system of transport departments. In the same year, the Beijing–Tianjin inter-city railway became fully operational, and China entered the “era of high-speed rail” (ibid.).

In 2013, the construction of comprehensive transport, intelligent transport, green transport, and safe transport were put forward as key themes to inform the development plans for the construction of the Belt and Road Initiative (The Belt and Road Initiative (also known as One Belt One Road) is a global development strategy implemented by the Chinese central government involving infrastructure development and investment in the Americas, Middle East, Africa, Europe and Asia), the coordinated development of Beijing, Tianjin, and Hebei (Jing-Jin-Ji project), and the construction of the Yangtze River economic belt (ibid.).

Over the past few years, with China’s rapid urbanization (Guan et al., 2018; Wu and Ma, 2006), expanding the public transport network and improving its efficiency have become the main priority for Chinese cities (General Office of the State Council, 2018). The following section discusses the current situation and some of the main strategies used in relation to transport planning and management in China.

Current Transport Planning and Management Problems and Strategies

Current Situation

At present, transport-related economic (cost–benefit analysis and/or wider impacts of transport investment) and environmental (low-carbon transport and/or green transport) concerns are a priority for transport planning and management in China (Cao et al., 2017; Guo and Meng, 2019; He et al., 2019; Loo, 2018; Wang et al., 2015). There has also been a growing awareness of the importance of other pillars, such as transport-related social impacts (Attard, 2020; Cao and Hickman, 2019b, 2020; Li and Zhao, 2018; Zhao and Bai, 2019). In other words, boosting local and regional economic development remains the key aim for most Chinese cities, as well as accommodating the rapid growth in private vehicles, particularly in some of the megacities, where the transport system is overcrowded but public transport services remain limited. In addition, similarly to other cities around the world, increased car usage and urban traffic congestion, especially in tier 1 and tier 2 cities (e.g., Beijing – Tier 1 city; Xi’an – Tier 2 city), present problems that are becoming increasingly difficult to resolve in relation to urban development and operation (Li et al., 2019). Consequently, people have to allow a lot of time for their commuting (Zhu et al., 2017). Although transport planners and policymakers recognize that reducing the heavy congestion, and shifting away from extending roads and overpasses to prioritizing the development of public transport within Chinese cities is paramount (State Council, 2012), there are still a number of issues that need addressing. Three key points, namely the quality and demand/supply capacity of the public transport system; transport pricing; and traffic rules, are discussed below.

First, the quality and adequacy of the existing public transport system have not met most transport users’ expectations or kept pace with demand. For instance, Wu et al. (2019) pointed out that, apart from commuters who use the shuttle bus, statistics have shown that few people felt significantly greater satisfaction with commuting by public transport than by private car. If public transport fails to offer any potential benefits and/or greater satisfaction, it is difficult to persuade private car users to switch from their current travel mode to using public transport. Furthermore, an increase in commuting has caused the public transport network to become increasingly overloaded, particularly at peak times due to the limited capacity on buses and on the underground. Therefore, in order to avoid using crowded and uncomfortable modes of public transport, many people choose to commute by private vehicles, as their non-instrumental—symbolic and affective—motives may not be fulfilled by public transport (Steg, 2005). People perceive that car ownership can improve their travel satisfaction and enhance the quality of life (Gärling and Schuitema, 2007), in a way that public transport, in its current state, cannot. Thus, transport planners and policymakers should also consider improving the quality (e.g., providing Wi-Fi, air-conditioning, and women-only carriages) and capacity of public transport rather than focusing solely on increasing its speed and extending the transport network.

Furthermore, public transport pricing is one of the most important factors affecting transport-related social inequity, and also influences people’s travel behavior (El-Geneidy et al., 2016; Low et al., 2020; Zhao and Zhang, 2019). Undoubtedly, a growing number of people are commuting by public transport in Chinese cities. For instance, in Beijing, although private car ownership is still very high, 72% of people used public transport to travel to work in the central urban area in 2017 (BITI, 2018). More broadly, the total operating mileage of China’s metro system reached 5376.89 km in December 2018, while there were 41 cities served by the metro as of September 2019. However, it is argued that a distance-based fare may not be sufficient to reduce the travel demand for people who do not need to commute at peak times (Monday to Friday between 6:30 a.m. and 9:30 a.m., and between 4:00 p.m. and 7:00 p.m.). In other words, although some public transport users can travel at off-peak times, they may still choose to travel at peak-times, because they do not have to pay higher fares at the busiest times of the day; therefore, the current distance-based fare system is

not achieving the aim of reducing travel demand at peak times. In the case of London, passengers are charged higher fares if they travel at peak times, which may deter people from traveling during the rush hour if they do not really need to. At present, in most Chinese cities, public transport fares are still relatively cheap, in addition to being substantially subsidized by local governments, unlike in most cities in the global north. In addition, although some local governments have recommended that companies and institutions offer flexible working schedules, this has proved difficult for them to implement in reality, meaning that many people still need to travel at peak times.

Last but not least, in many areas of public life, nothing can be accomplished without norms or standards, and transport is no exception. In order to reduce the heavy traffic congestion and traffic-related injuries, traffic laws and regulations are essential to ensure that the overall traffic flow moves smoothly and people can travel safely. However, a proportion of pedestrians, cyclists, and drivers continue to violate the traffic rules. For example, statistically, the following megacities: Beijing, Shanghai, Washington, New York, London, Manchester, Tokyo, New Delhi and Mumbai, Beijing and New Delhi, witness the most traffic violations by motor vehicle drivers and pedestrians, as well as the highest numbers of traffic accidents. The main reason is that an individual who violates the traffic rules knows they are guilty but thinks that they will be fortunate enough to do so without incurring punishment. Indeed, those violations that are spotted are only punished mildly or in some cases not at all, showing that rules and regulations are not particularly effective in these cities. It was also found that most of the violations occurred in the peripheral areas in the case of Beijing and New Delhi (Baluja, 2008; Zheng et al., 2012). Hence, it can be argued that the level of punishment is currently far from sufficient to deter people. For instance, in Beijing, since the introduction of the new traffic law on 1st January 2013, the rules stipulate that drivers who go through a red light after that date will incur a fine of 200 RMB (1 GBP $\approx$ 9 RMB (Chinese Yuan)) and get a 6 point deduction (drivers with a total deduction of 12 points are banned) instead of the previous fine of 200 RMB with a 3 point deduction (MPS, 2016). In many places, drivers who break the rules are only fined 200 RMB if caught on camera, and do not have any points deducted. Clearly, traffic regulations and penalties are not an appropriate means with which to compel all commuters to obey traffic laws. Although these are still helpful at the current stage in most Chinese cities, educating people so that they understand the importance of obeying traffic laws is the ultimate goal.

In this respect, Tokyo offers one of the best examples that others can learn from. The “no jaywalking” regulation is largely adhered to by pedestrians, and designated pedestrian lanes are used by most pedestrians in Japan. In general, most Japanese people wait obediently at stoplights, although car drivers are held responsible even if an accident is primarily caused by pedestrians (HRDB, 2020). Top priority is given to the pedestrian in Japan, but most pedestrians and cyclists still comply with the traffic rules in any case. There are also fewer cameras used in Japan to capture violations committed by vehicles, cyclists or pedestrians. Additionally, in London, in order to reduce surface transport casualties, Camden council has produced a policy to educate young people about how to cross the road and stay safe when walking along the streets (Camden Council, 2020). Therefore, it can be argued that educating commuters to obey the traffic regulations and recognize the importance of sustainable transport development is vital. However, there are some questions about how this can be replicated in Chinese cities.

Discussion of Current Transport Planning and Management Strategies and Their Implications

In order to alleviate traffic congestion, reduce car usage and mitigate air pollution and CO₂ emissions, governments have adopted various efficient transport policy packages and strategies aimed at achieving sustainable transport development.

In terms of reducing car usage, the Beijing Municipal Government (BMG) has pursued a car registration plate restriction policy since 2008 that prevents vehicles whose last digit is 1 or 6 from driving on a Monday, vehicles whose registration ends in 2 or 7 from driving on a Tuesday, and so on, following a system of rotation. Furthermore, all private vehicles driving within the sixth ring road (urban areas, excluding the sixth ring road) and including Tongzhou District (excluding the main highways) which do not have Beijing registration plates must apply for a Beijing temporal pass (i.e. a Beijing entry permit) in advance, as well as limiting the maximum stay (including both driving and parking) in the urban area of Beijing and the Tongzhou District to 84 days per year from 1st November 2019 (China Daily, 2019). In addition, cars without Beijing registration plates are banned from driving within the fifth ring road at peak times even if they have a Beijing temporal pass. Moreover, a licence plate lottery system has been in force since 2011 in order to curb the rapid growth in purchases of new vehicles in Beijing. Under this system, a purchase permit draw usually takes place once every two months (the lottery was held every month before 2014) for potential vehicle buyers. Similar licence plate lottery schemes have also been rolled out in other cities, such as Shenzhen, Shijiazhuang, Guangzhou, Hangzhou, Tianjin, Guiyang, and Hainan. Moreover, a congestion charge scheme represents one of the most efficient approaches to managing transport demand, which has been successfully applied to reduce urban traffic jams within congestion charge zones in several international cities (Börjesson and Kristofferson, 2018; TfL, 2008, 2018; Wu et al., 2017). Beijing is currently planning to introduce a congestion charge (Liu et al., 2019), following in the footsteps of Singapore, Oslo, London and Stockholm. However, because the urban fabric and design rule does not permit CCTV cameras to be hung on a frame over or next to a road, in an unsightly manner, this has now been delayed.

In order to reduce traffic jams and car usage in Shanghai, an auction system has been used to maintain a set quota of new vehicles since 1994. In September 2019, it was announced that 149,507 candidates took part in Shanghai’s auction system, but only approximately 6% of them were successful. The average price of a car licence plate in Shanghai was 89,637 RMB in September 2019. Similar car registration plate auction policies have also been rolled out in other cities, such as Tianjin, Guangzhou, Hangzhou, and Shenzhen. Such schemes may also raise other issues in relation to transport-related social inequity (Cao and Hickman, 2019a; Cao et al., 2019; Cuthill et al., 2019; Knotts, 2019), although this point is not discussed in this chapter. However, it is still worth



Figure 1 Beijing cycle expressway.

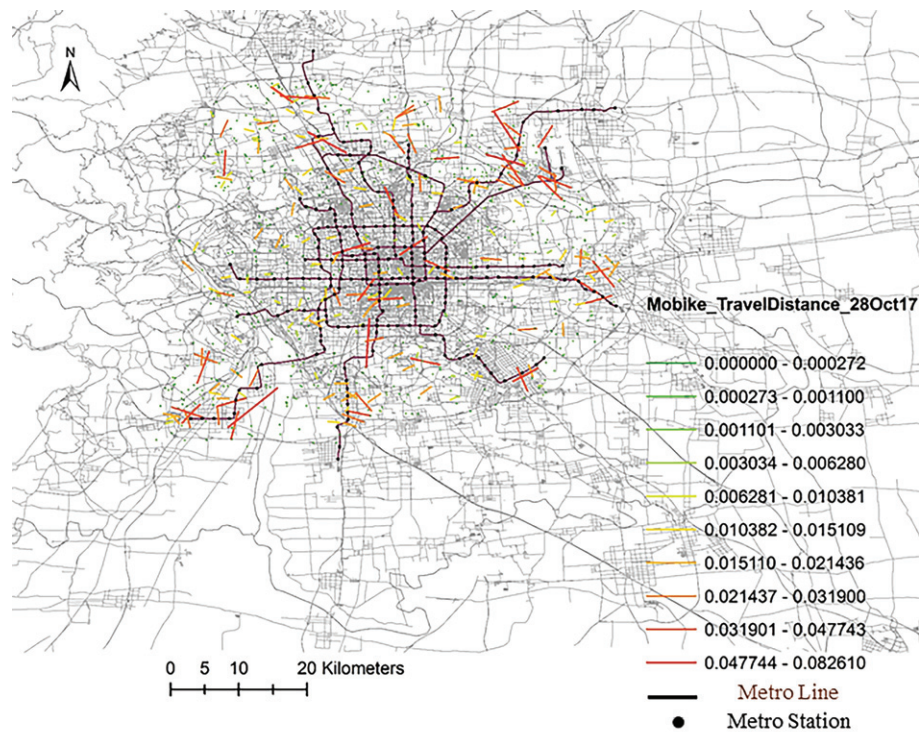


Figure 2 Origin-destination commuting trips by dockless bikes (sample: "Mobike") in Beijing on 28 October 2017 between 8:00 a.m. and 8:30 a.m.

thinking carefully about who the winners and losers are in the car usage reduction process. It is likely that the better-off and higher-income cohorts will be the main beneficiaries.

Active travel promotion, such as dockless-bicycle-sharing programs and/or segregated cycle lane development, is one of the most successful approaches that has been implemented in terms of achieving sustainable transport development in many Chinese cities (Cheng et al., 2019; Gu et al., 2019; Jia and Fu, 2019; Zhang and Mi, 2018). The wider positive impacts, such as reducing traffic jams, air pollution, and CO₂ emissions, and improving people's health (e.g., reducing obesity), can clearly be attributed to encouraging active travel via cycling and walking (Aldred et al., 2019; De Vos et al., 2016). For example, the first Beijing cycle expressway opened on 31st May 2019, covering 6.5 km, and connects Huilongguan to Shangdi in Northern Beijing (see Figs. 1 and 2). It offers safe, convenient and fast travel to over 8000 commuters daily. However, it has been argued that active travel may be more effective in addressing the "first mile" and/or "last mile" of Chinese commuters' journeys, particularly in the megacities, rather than making the whole trip by bicycle or on foot. For example, Fig. 2 shows that most of the origins or destinations that people travel to or from using dockless bicycles (e.g., Mobike) corresponded to the subway stations in Beijing. A possible explanation for this is that Beijing is a

large city covering an area of 16,410 km². Therefore, it has different mobility challenges in terms of active travel, compared to London for example, which covers an area of only 1577.3 km², meaning that people can complete their entire commute by bicycle and/or on foot. In 2018, the average commuting time for Beijing commuters was 56 min within the sixth ring road, and the average travel distance was around 12.4 km (BTI, 2019). Therefore, it is difficult for cyclists or pedestrians to commute from their home to their workplaces, especially if they live in any of the Chinese megacities such as Beijing, Shanghai, Tianjin, and Shenzhen. In some tier 3 or 4 cities, or relatively small-scale Chinese cities, pro-active travel would be more feasible. However, it is essential to provide the necessary support facilities, because it may be too expensive for these smaller cities to invest the required amount in cycling infrastructure. For example, in the UK city of Nottingham, it would cost around £1 million to build a one-mile segregated cycle path (1 GBP ≈ 9.2 RMB (Chinese Yuan)). This means that the local government and/or transport authority are more likely to reject the transport appraisal proposal to build a cycle highway, particularly in a tier 3 or 4 city, as the conventional cost–benefit analysis model is still widely applied to evaluate mega transport projects in China, and even in the UK (Hickman and Dean, 2018; Spurling et al., 2019; Vickerman, 2017). Last but not least, the management of dockless bicycles also constitutes an issue for most Chinese cities, primarily because of oversupply (Lyu et al., 2020), arbitrary parking (Zhang et al., 2019), and repair and recycling (Lyu et al., 2020). It is evident that the advantages and disadvantages need to be carefully considered and balanced when implementing strategies to tackle transport issues.

Finally, but no less importantly, land-use planning and urban structure can significantly influence travel demand and behavior (Hu et al., 2019). A higher range of densities, pedestrian-oriented design, mixed land-use, improved travel demand management, and an increased level of accessibility always play a key role in shaping sustainable transport development (Cervero and Kockelman, 1997; Ewing and Cervero, 2001, 2010; Hickman et al., 2010). In China, sustainable urban planning and land-use policies have already been taken into consideration by decision-makers and urban planners in order to meet sustainable urbanization goals (Xu et al., 2019). For instance, the collaborative development strategies in Beijing, Tianjin and Hebei (also known as the Jing-Jin-Ji project) will not only boost local and regional economic growth, but also dramatically improve the levels of accessibility between each city (Sheng et al., 2018). However, it should also be noted that a lesson can be learnt from the case of Tongzhou, a dormitory satellite town of Beijing. In order to help reduce traffic congestion within Beijing, the Beijing Municipal Government (BMG) tried to develop several new satellite towns with good quality links to the existing city centre, with the aim of dispersing traffic and reducing population overcrowding (see Kong et al., 2005, for more details of the Beijing Master Plan 1991–2010). In theory, a satellite town should be able to be selfsupporting, selfresourcing, and selfsufficient. However, in reality, Tongzhou has been transformed from a satellite town into a dormitory town due to the job–housing imbalance issue, partly caused by unaffordable house prices (Housing affordability has already been identified as a social issue (Edwards, 2016)); however, it has not been recognized and/or analysed together with the joint effects of transport and housing costs (Cao and Hickman, 2018; Guerra and Kirschen, 2016; Liu et al., 2020)) in the central area of Beijing, as well as generating a high “tidal traffic flow” at peak times between Tongzhou and Guomao (Central Business District in Beijing). In an attempt to mitigate this problem and relieve some of the pressure on Beijing, the BMG has moved a number of state-owned companies, schools, hospitals and industries to Tongzhou. It is expected that approximately 400,000 people will have moved from the central area of Beijing to Tongzhou by 2020. It is hoped that this policy-oriented action will help turn Tongzhou into a proper satellite town and subcenter of Beijing, thereby reducing some of the overload on central Beijing.

Conclusions

This chapter has reviewed the history of transport development in China and critically discussed the status quo and some current transport planning and management strategies and their implications. The key findings show that, although investment has been made in the transport infrastructure, with the network being extended, and some effective transport strategies being rolled out, traffic issues persist in many Chinese cities. Although significant efforts have been made to address some of the main traffic problems, the goal of sustainability remains unfulfilled. In addition, mitigating strategies and transport planning approaches may not always be directly replicable from developed countries to developing countries, at least in the Chinese context.

This chapter makes two main contributions to both academia and practice. In the academic context, the review and critical discussion of the status quo and current strategies applied to tackle traffic issues can contribute to the international literature in order to enrich our understanding of transport planning and management, particularly through the lessons learnt from Chinese cities. In practice, it can make transport planners and policymakers aware of the importance of understanding and implementing both government policy- and people-oriented strategies to achieve sustainable transport development, as well as taking the unexpected and/or potential side-effects into consideration when making transport policy decisions in the future.

It is argued that wherever there are cities in China, there will be an indefinite demand for road capacity over the next couple of years following the current trajectory of the “business as usual” model. Therefore, it is imperative to implement sustainable transport strategies, manage travel demand, and integrate land-use planning and transport via a top-down strategy. In addition, I would argue that implementing top-down-oriented transport planning strategies (e.g., transport demand enforcement and management) may only temporarily reduce car usage and mitigate traffic congestion in the short term. In order to achieve the ultimate goal of sustainable transport development in the long run, the most important thing is to change the perceptions and awareness of people who use motorized vehicles, and enable them to realize the fundamental significance of using sustainable transport modes (e.g., public transport, and cycling and walking) and in turn help to create sustainable and liveable cities in China.

Acknowledgments

This research is partly funded by the NSFC (Project No. 51808392), the EPSRC (EPSRC Reference: EP/R035148/1), the SCUE Research Funding, and School Funding from the University of Westminster. The author would like to extend his appreciation to Maria Attard (editor) from University of Malta (Malta), Yuerong Zhang from UCL (UK), Yue Qiu from Beijing International Studies University (China), Zijia Wang from Beijing Jiaotong University (China), Cheng Shi from Tongji University (China), Chen Liu and Zhexi Zhao from Beijing Transport Institute (China) and Shuang Liu from Zero Zero Family (Japan) for providing their comments on the initial draft. The author would also like to thank Qihao Liu from UCL (UK), and Yajing Chen and Tianze Liu from Beijing University of Agriculture (China) for assisting in collecting relevant data and materials.

References

- Aldred, R., Croft, J., Goodman, A., 2019. Impacts of an active travel intervention with a cycling focus in a suburban context: One-year findings from an evaluation of London's in-progress mini-Hollands programme. *Transport. Res. Part A: Policy Pract.* 123, 147–169.
- Attard, M., 2020. Mobility justice in urban transport – the case of Malta. *Transport. Res. Proc.* 45, 352–359.
- Attard, M., Shiftan, Y., 2015. *Sustainable Urban Transport*. Emerald Group Publishing Limited, Bingley.
- Baluja, R., 2008. Assessing road safety through road traffic violations: A case study of New Delhi, India. *UNECE Transport Review*, United Nations Commission for Europe, pp. 98–99.
- Banister, D., 2002. *Transport Planning*, 2nd ed. Spon Press, London.
- Beijing Transport Institute (BTI), 2018. *Beijing Transport Annual Report 2018*. BTI (in Chinese), Beijing.
- Beijing Transport Institute (BTI), 2019. *Analysis of Commuting Characteristics and Typical Areas in Beijing*. Beijing: BTI (in Chinese).
- Börjesson, M., Kristoffersson, I., 2018. The Swedish congestion charges: ten years on. *Transport. Res. Part A: Policy Pract.* 107, 35–51.
- Camden Council, 2020. *School Travel and Child Road Safety*. Available from: <https://www.camden.gov.uk/school-travel-and-child-road-safety#kxax> (accessed 2nd July 2020).
- Cao, M., Hickman, R., 2018. Car dependence and housing affordability: an emerging social deprivation issue in London. *Urban Stud.* 55 (10), 2088–2105.
- Cao, M., Hickman, R., 2019a. Urban transport and social inequities in neighbourhoods near underground stations in Greater London. *Transport. Plan. Technol.* 42 (5), 419–441.
- Cao, M., Hickman, R., 2019b. Understanding travel and differential capabilities and functionings in Beijing. *Transport Policy* 83, 46–56.
- Cao, M., Hickman, R., 2020. Transport, social equity and capabilities in east Beijing. In: Chen, C.-L., Pan, H., Shen, Q., Wang, J. (Eds.), *Handbook on Transport and Urban Transformation in China*. Edward Elgar, London.
- Cao, M., Chen, C.-L., Hickman, R., 2017. Transport emissions in Beijing: a scenario planning approach. *Proc. Inst. Civil Eng. Transport* 170 (2), 65–75.
- Cao, M., Zhang, Y., Zhang, Y., Li, S., Hickman, R., 2019. Using different approaches to evaluate individual social equity in transport. In: Hickman, R., Mella-Lira, B., Givoni, M., Geurs, K. (Eds.), *A Companion to Transport, Space and Equity*. Edward Elgar, London, pp. 209–228.
- Cervero, R., Kockelman, K., 1997. Traffic demand and the 3Ds: density, diversity, and design. *Transport. Res. Part D: Transport Environ.* 2 (3), 199–219.
- Cheng, L., Chen, X., Yang, S., Cao, Z., De Vos, J., Witlox, F., 2019. Active travel for active ageing in China: the role of built environment. *J. Transport Geogr.* 76, 142–152.
- China Daily, 2019. *Beijing Restricts Non-locally Registered Autos to Ease Congestion, Reduce Emission*. Available from: <https://www.chinadaily.com.cn/a/201911/02/WS5dbd26eea310cf3e355750a3.html> (accessed 1st July 2020).
- Cuthill, N., Cao, M., Liu, Y., Gao, X., Zhang, Y., 2019. The association between urban public transport infrastructure and social equity and spatial accessibility within the urban environment: an investigation of Tramlink in London. *Sustainability* 11 (5), 1229.
- De Vos, J., Mokhtarian, P.L., Schwanen, T., Van Acker, V., Witlox, F., 2016. Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility. *Transportation* 43 (5), 771–796.
- Edwards, M., 2016. The housing crisis and London. *City* 20 (2), 222–237.
- El-Geneidy, A., Levinson, D., Diab, E., Boisjoly, G., Verbich, D., Loong, C., 2016. The cost of equity: assessing transit accessibility and social disparity using total travel cost. *Transport. Res. Part A: Policy Pract.* 91, 302–316.
- Ewing, R., Cervero, R., 2001. Travel and the built environment: a synthesis. *Transport Res. Rec.* 1780 (1), 87–114.
- Ewing, R., Cervero, R., 2010. Travel and the built environment: a meta-analysis. *J. Am. Plan. Assoc.* 76 (3), 265–294.
- Gärling, T., Schuiterna, G., 2007. Travel demand management targeting reduced private car use. *J. Social Issues* 63 (1), 139–153.
- General Office of the State Council (GOSC), 2018. *Opinions on Further Strengthening Urban Rail Transit Planning and Construction Management*. GOSC, Beijing.
- Gu, T., Kim, I., Currie, G., 2019. To be or not to be dockless: empirical analysis of dockless bikeshare development in China. *Transport. Res. Part A: Policy Pract.* 119, 122–147.
- Guan, X., Wei, H., Lu, S., Dai, Q., Su, H., 2018. Assessment on the urbanization strategy in China: achievements, challenges and reflections. *Habitat Int.* 71, 97–109.
- Guerra, E., Kirschen, M., 2016. Housing plus transportation affordability indices: uses, opportunities, and challenges. *OECD round-table on income inequality, social inclusion, and mobility*, Paris, April 2016.
- Guo, M., Meng, J., 2019. Exploring the driving factors of carbon dioxide emission from transport sector in Beijing-Tianjin-Hebei region. *J. Cleaner Prod.* 226, 692–705.
- He, D., Yin, Q., Zheng, M., Gao, P., 2019. Transport and regional economic integration: evidence from the Chang-Zhu-Tan region in China. *Transport Policy* 79, 193–203.
- Hokkaido Regional Development Bureau (HRDB), 2020. *Japanese Road Traffic Rules and Recommendations*. Available from: <https://www.hkd.mlit.go.jp/ky/ki/renkei/ud49g70000003k5h-att/04CHNb-chap3-1.pdf> (accessed 10th July 2020) (in Chinese).
- Hickman, R., Banister, R., 2014. *Transport, Climate Change and the City*. Routledge, Abingdon.
- Hickman, R., Dean, M., 2018. Incomplete cost – incomplete benefit analysis in transport appraisal. *Transport Res.* 38 (6), 689–709.
- Hickman, R., Seaborn, C., Headicar, P., Banister, D., Swain, C., 2010. Spatial planning for sustainable travel? *Town Country Plan.* 77–82.
- Hu, L., Yang, J., Yang, T., Tu, Y., Zhu, J., 2019. Urban spatial structure and travel in China. *J. Plan. Lit.* 1–19.
- Jia, Y., Fu, H., 2019. Association between innovative dockless bicycle sharing programs and adopting cycling in commuting and non-commuting trips. *Transport. Res. Part A: Policy Pract.* 121, 12–21.
- Knotts, J., 2019. *Urban Development: Mapping the Unique Challenges That Face Beijing*. Available from: <https://www.thebeijinger.com/blog/2019/10/23/urban-development-unique-challenges-face-beijing> (accessed 23rd October 2019).
- Kong, X., Chen, Y., Xin, Y., Gu, H., 2005. Current situation, problems and suggestions for the development of satellite towns in the city of Beijing. *Beijing Acad. Soc. Sci.* 3, 9–16.
- Li, S., Zhao, P., 2018. Restrained mobility in a high-accessible and migrant-rich area in downtown Beijing. *Eur. Transport Res. Rev.* 10 (4), 1–17.
- Li, Y., Xiong, W., Wang, X., 2019. Does polycentric and compact development alleviate urban traffic congestion? A case study of 98 Chinese cities. *Cities* 88, 100–111.
- Liu, C., Cao, M., Yang, T., Ma, L., Wu, M., Cheng, L., Ye, R., 2020. Inequalities in the commuting burden: Institutional constraints and job-housing relationships in Tianjin, China. *Res. Transport. Business Manage.*, <https://doi.org/10.1016/j.rtbm.2020.100545>.
- Liu, Q., Lucas, K., Marsden, G., Liu, Y., 2019. Egalitarianism and public perception of social inequities: a case study of Beijing congestion charge. *Transport Policy* 74, 47–62.
- Loo, B.P.Y., 2018. *Unsustainable Transport and Transition in China*. Routledge, New York.
- Low, W.-Y., Cao, M., De Vos, J., Hickman, R., 2020. The journey experience of visually impaired people on public transport in London. *Transport Policy* 97, 137–148.

- Lyu, Y., Cao, M., Zhang, Y., Yang, T., Shi, C. 2020. Investigating users' perspectives on the development of bike-sharing in Shanghai. *Res. Transport. Business Manage.*, <https://doi.org/10.1016/j.rtbm.2020.100543>.
- Ministry of Public Security (MPS), 2016. Regulations for the Application and Use of Motor Vehicle Driving Licence. Law Press (in Chinese), Beijing.
- Sheng, X., Cao, Y., Zhou, W., Zhang, H., Song, L., 2018. Multiple scenario simulations of land use changes and countermeasures for collaborative development mode in Chaobai River region of Jing-Jin-Ji, China. *Habitat Int.* 82, 38–47.
- Spurling, D., Spurling, J., Cao, M., 2019. *Transport Economics Matters: Applying Economic Principles to Transportation in Great Britain*. Brown Walker Press, Boca Raton.
- State Council, 2012. Guiding Opinions of the State Council on Giving Priority to the Development of Public Transport in Chinese Cities. State Council, Beijing (in Chinese).
- State Council Information Office (SCIO), 2016. The Development of Transportation in China (White Paper). SCIO, Beijing.
- Steg, L., 2005. Car use: lust and must. Instrumental, symbolic and affective motives for car use. *Transport. Res. Part A: Policy Pract.* 39, 147–162.
- Transport for London (TfL), 2008. Central London: Congestion Charging Impacts Monitoring – Sixth Annual Report. TfL, London.
- Transport for London (TfL), 2018. Annual Report and Statement of Accounts. TfL, London.
- Vickerman, R., 2017. Beyond cost-benefit analysis: the search for a comprehensive evaluation of transport investment. *Res. Transport. Econ.* 63, 5–12.
- Wang, Z., Chen, F., Fujiyama, T., 2015. Carbon emission from urban passenger transportation in Beijing. *Transport. Res. Part D: Transport Environ.* 41, 217–227.
- Wu, F., Ma, L.J.C., 2006. Transforming China's globalizing cities. *Habitat Int.* 30 (2), 191–198.
- Wu, K., Chen, Y., Ma, J., Bai, S., Tang, X., 2017. Traffic and emissions impact of congestion charging in the central Beijing urban area: a simulation analysis. *Transport. Res. Part D: Transport Environ.* 51, 203–215.
- Wu, W., Wang, M., Zhang, F., 2019. Commuting behavior and congestion satisfaction: Evidence from Beijing, China. *Transport. Res. Part D: Transport Environ.* 67, 553–564.
- Xu, Q., Zheng, X., Zheng, M., 2019. Do urban planning policies meet sustainable urbanization goals? A scenario-based study in Beijing, China. *Sci. Total Environ.* 670, 498–507.
- Zhang, Y., Mi, Z., 2018. Environmental benefits of bike sharing: a big data-based analysis. *Appl. Energy* 220, 296–301.
- Zhang, Y., Lin, D., Mi, Z., 2019. Electric fence planning for dockless bike-sharing services. *J. Clean. Prod.* 206, 383–393.
- Zhao, P., Bai, Y., 2019. The gap between determinants of growth in car ownership in urban and rural areas of China: a longitudinal data case study. *J. Transport Geogr.* 79, 102487.
- Zhao, P., Zhang, Y., 2019. The effects of metro fare increase on transport equity: new evidence from Beijing. *Transport Policy* 74, 73–83.
- Zheng, L., Tang, Q., Liu, X., Tao, R., 2012. A List of Beijing's Top Traffic Violation Places. Available from: http://www.xinhuanet.com/politics/2012-12/03/c_124035762.htm (accessed 1st July 2020). (in Chinese).
- Zhu, Z., Li, Z., Liu, Y., Chen, H., Zeng, J., 2017. The impact of urban characteristics and residents' income on commuting in China. *Transport. Res. Part D: Transport Environ.* 57, 474–483.

Further Reading

- Asian Development Bank (ADB), 2015. *Improving Interchanges: Introducing Best Practices on Multimodal Interchange Hub Development in the People's Republic of China*. ADB, Manila.
- Attard, M., Ison, S.G., Emberger, G., 2018. Case studies in transport policy special issue transport planning and policy. *Case Stud. Transport Policy* 6 (3), 309–310.
- Chen, C.-L., 2018. Tram development and urban transport integration in Chinese cities: a case study of Suzhou. *Econ. Transport.* 15, 16–31.
- Chen, C.-L., Pan, H., Shen, Q., Wang, J., 2020. *Handbook on Transport and Urban Transformation in China*. Edward Elgar, Cheltenham.
- Chen, Z., Haynes, K., Zhou, Y., Dai, Z., 2019. *High Speed Rail and China's New Economic Geography: Impact Assessment from the Regional Science Perspective*. Edward Elgar, Cheltenham.
- Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. *Travel Behav. Soc.* 14, 1–10.
- Hickman, R., Givoni, M., Bonilla, D., Banister, D., 2015. *Handbook on Transport and Development*. Edward Elgar, Cheltenham.
- Hu, L., Iseki, H., 2019. Land use and transportation planning in a diverse world. *Transport Policy* 81, 282–283.
- Knowles, R.D., Ferbrache, F., 2019. *Transit Oriented Development and Sustainable Cities: Economics, Community and Methods*, Edward Elgar, Cheltenham.
- Shammut, M., Cao, M., Zhang, Y., Papaix, C., Liu, Y., Gao, X., 2019. Banning diesel vehicles in London: Is 2040 too late? *Energies* 12 (18), 3495.
- Van Wee, B., Annema, J.A., Banister, D., 2013. *The Transport System and Transport Policy: An Introduction*. Edward Elgar, Cheltenham.
- Yang, L., Hu, L., Wang, Z., 2019. The built environment and trip chaining behaviour revisited: the joint effects of the modifiable areal unit problem and tour purpose. *Urban Stud.* 56 (4), 795–817.
- Ye, R., Ma, L., 2019. Does daily commuting behavior matter to employee productivity? *J. Transport Geogr.* 76, 130–141.
- Zhang, Y., Marshall, S., Manley, E., 2019. Network criticality and the node-place-design model: classifying metro station areas in Greater London. *J. Transport Geogr.* 79, 102485.
- Zhao, P., Hu, H., 2019. Geographical patterns of traffic congestion in growing megacities: big data analytics from Beijing. *Cities* 92, 164–174.

Road Safety

George Yannis*, Eleonora Papadimitriou†, *National Technical University of Athens, Department of Transportation Planning & Engineering, Athens, Greece; †Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

The Road Safety Problem	51
Key Global Statistics	51
Road Safety and Macroscopic Developments	51
The Road Safety System	52
Road Safety Risk	53
Definitions of Risk and Exposure	53
Confounding Factors in Risk Assessment	54
Risk Assessment Through Traffic Conflicts and Surrogate Safety Measures	54
Road Safety Management	54
Structure of Road Safety Management Systems	54
Safe System Approach	54
Safety Impact Assessment	55
Road User	55
The Road Environment	56
The Vehicle	57
Future Road Safety Perspectives	57
Relevant Websites	58
References	58

The Road Safety Problem

Key Global Statistics

Road traffic crashes constitute a major global public health problem, since each year 1.35 million people are killed in road crashes worldwide. According to the World Health Organization, road traffic injuries are the 8th leading cause of death for people of all ages and the 1st cause of death among children and young adults (5–29 years old).

Over the last 15 years, the number of road fatalities has increased from 1.15 million to 1.35 million, while the fatality rate per population has remained stable at almost 18 fatalities per 100,000 population. The respective rate per vehicles has reduced more than 50% (from 135 fatalities per 100,000 vehicles in 2001 to 64 vehicles per 100,000 vehicles in 2015), while the number of motorized vehicles increased during the same period (WHO, 2018).

However, the burden of road traffic injuries is not distributed proportionately among the countries. As shown in Fig. 1, the low and middle income countries account for 85% of global population and 60% of registered motorized vehicles, however, the vast majority of road fatalities (93%) have been recorded in these countries.

From a geographical perspective, the highest rates of road traffic deaths were recorded in Africa (26.6/100,000 people) and in Asia-Pacific group (19/100,000 people). On the contrary, Western European and other countries had the lowest fatality rate with 8.2 fatalities per 100,000 population. In terms of road safety progress, the road fatalities per population have increased in all regions during the last three years, except the Eastern European and the Western European and other countries (Fig. 2).

Finally, great variations are also found in the types of road users killed in road crashes. At a global level, more than half of road fatalities concern Vulnerable Road Users (VRUs), and more specifically, 28% were Powered Two Wheelers (PTWs) riders, 23% were pedestrians and 3% were cyclists. Moreover, the highest percentage of pedestrian fatalities was recorded in African and Eastern Mediterranean countries, while in South-East Asia and Western Pacific the majority of road fatalities were PTWs (WHO, 2018).

Road Safety and Macroscopic Developments

Historically, the level of motorization has been proven to have a nonlinear correlation with road crashes (Kopits and Cropper, 2005). Early experiences from industrialized countries suggested that, while the level of motorization increased, the number of road fatalities also increased, due to the increased traffic and related speeds enabled by developments in vehicle technology and road infrastructure. However, a “breakpoint” was observed, roughly situated during the 1970s in most countries, after which the level of motorization still increased but the number of fatalities showed a decreasing trend; this was attributed on the one hand to the increase in congestion in industrialized countries (hence lower speeds) and on the other hand to the awareness created among stakeholders in these countries on the need to address the road safety problem, leading to the first wave of coordinated efforts. Several countries exhibited a similar pattern, although slightly shifted to later in time, depending on their patterns of economic growth (Yannis et al., 2011).

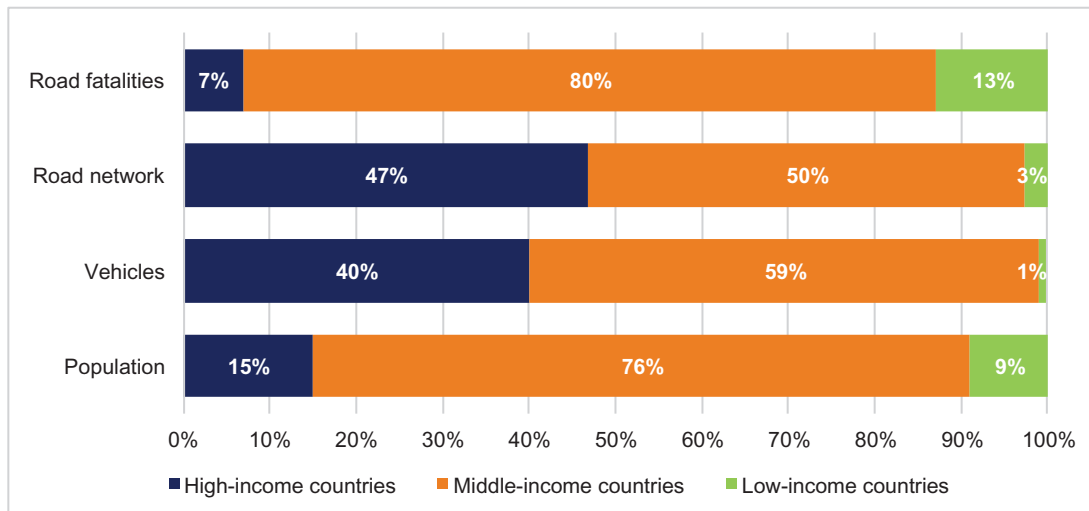


Figure 1 Proportion of population, registered motor vehicles, road network and road fatalities by country income category, 2016. Source: IRF-WRS (2018), WHO (2018) and World Bank (2018).

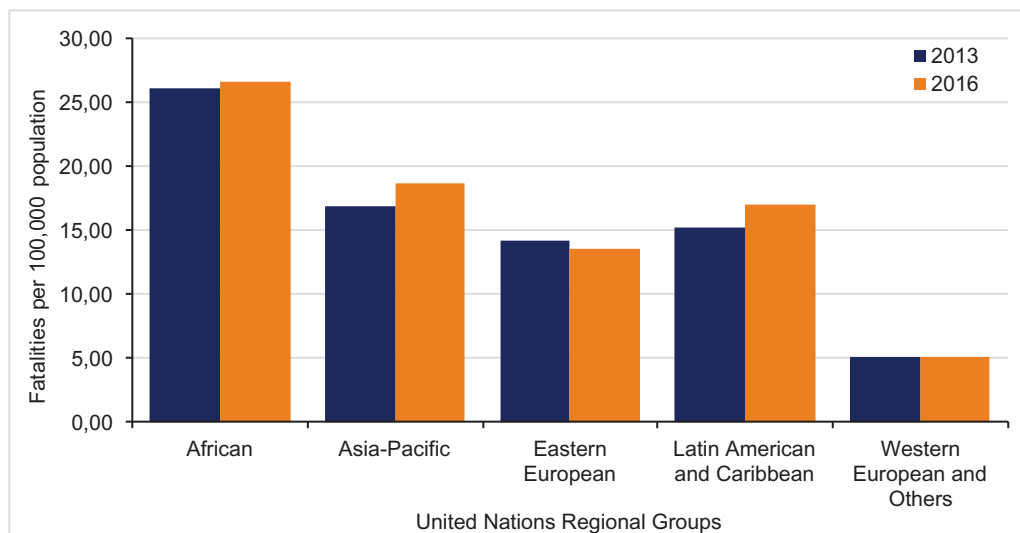


Figure 2 Fatality rates per population in the UN Regional Groups in 2013 and 2016. Source: WHO (2018); <https://www.un.org/depts/DGACM/RegionalGroups.shtml>.

A similar correlation is observed with respect to economic indicators such as GDP per capita and unemployment—all reflecting economic growth, social welfare, as well as cultural developments.

It should be noted, however, that a short-term relationship exists between road safety and economic recessions. When GDP per capita decreases, the number of road fatalities also decreases, as a result of reduced traffic exposure of all road users. Once economic growth resumes, the number of fatalities soon returns into an increasing trend. This pattern was recently validated in several studies dealing with the economic recession of 2008–15 in Europe and other industrialized countries (Yannis et al., 2014), as a short-term fluctuation within the macroscopic trend described previously.

The understanding of the macroscopic relationship between road crashes and economic growth is particularly important when considering the economic development expected in the next decades in the low- and middle-income countries. Many of these countries are a long way from reaching the “breakpoint” of the level of motorization at which road fatalities start to decrease, therefore the global road safety problem is expected to dramatically worsen, and there is urgent need for proactive action.

The Road Safety System

The road safety system includes three distinct components: the road user, the vehicle, and the road environment. Road crash investigations have shown that human factors are the main contributory factor in more than 65% of road crashes, and one of the contributory factors in more than 95% of road crashes (See Fig. 3).

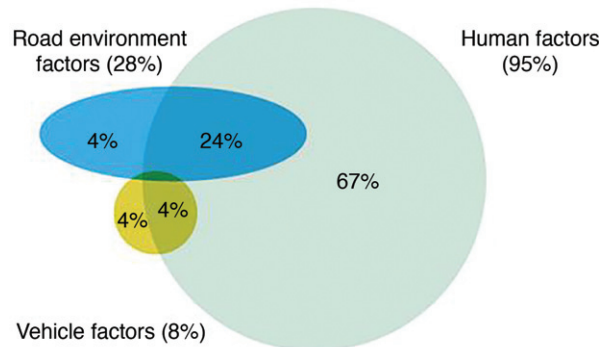


Figure 3 Overview of crash contributory factors. Source: NSW Roads and traffic Authority (1996).

A fourth component concerns the structural and operational characteristics of the system (including legislation and rules, cultures, and general socioeconomic elements), which affect the human-vehicle-road interaction; this is referred to as road safety management. In the following sections, each of the components of the road safety system is analyzed, following an introduction to the basic definitions and theories of road crash risk.

Road Safety Risk

Definitions of Risk and Exposure

The number of fatalities and injuries alone are seldom sufficient, and can even be misleading when assessing road safety problems; risk estimates are needed. Road risk can be generally defined as the probability of crash/fatality occurrence.

This definition draws from the probabilistic background of road crash occurrence; road crashes can be considered as the result of binomial (Bernoulli) trials, with two possible outcome (success or failure), and the Poisson theorem of crash counts consolidates the sum of these trials as a Poisson-distributed random variable. This is an imperfect consideration, although in several cases for practical purposes a Poisson distribution can be assumed. From a statistical viewpoint, a Poisson distribution has a basic property of the equality of its variance to its mean. When crash counts are over- or under-dispersed, it has been shown that a mixture of Poisson and Gamma distributions may best describe the process of crash occurrence, resulting in a Negative Binomial distribution (Hauer, 2001).

In the recent years, developments in statistical and econometric modeling have allowed to overcome many of the limitations of conventional models. For instance, the zero-inflated Poisson model may account for the presence of many zero counts in the data (Lord et al., 2005). Hierarchical, or multilevel, or mixed effects models may account for unobserved heterogeneity and other dependences among crash observations (Huang and Abdel-Aty, 2010).

The role of exposure in the estimation of road risk can be understood when considering the nature of Bernoulli trials. At an aggregate level, however, the number of crashes depends on the number of trials (and the trials cannot be considered as being independent).

If continuous measurements of exposure are available, risk can be defined through a function mapping road safety outcomes onto exposure. Within this “continuous” risk approach, crash occurrence will be modeled as a Negative Binomial model with the following general form, also known as Accident Prediction Model or Safety Performance Function:

$$N_i = L \times e^{[a+b \times \ln(c \times \text{exposure})]}$$

N_i is the predicted average crash frequency crashes in road segment (i); L is the segment length (in km); exposure is most often expressed through the annual average daily traffic [AADT - vehicles/day]; and a , b , c are coefficients to be estimated.

Continuous exposure measurements are not always possible, and therefore a “discrete” risk approach may be opted for, in which risk is defined as an (incidence) rate: the number of crashes divided by the amount of exposure over a given time period (Papadimitriou et al., 2013).

In practice, there are different measures of the amount of exposure, all with different advantages and limitations, but can be broadly classified in two groups:

- Traffic estimates: vehicle-kilometers of travel, road length, fuel consumption, and vehicle fleet.
- Persons-at-risk estimates: person-kilometers of travel, number of trips, time spent in traffic, driver population, and population.

Although any of the above indicators could be a pertinent approximation of exposure in certain contexts (e.g., fuel consumption for macroscopic analysis, population for epidemiological analysis, and road length for developing countries), the number of vehicle- and person-kilometers of travel is widely considered as the optimal measure.

Despite important efforts at national and international level, the lack of exposure data remains one of the main limitations of the potential for reliable and evidence-based road safety analysis. While exposure data (in the form of traffic flow/AADT) are largely available for the main interurban road networks and the main vehicle types, the measurement of exposure in urban areas and for road user characteristics is much more scarce.

Confounding Factors in Risk Assessment

Road crashes are rare probabilistic events. Due to the randomness involved in the crash process, it is often observed that crash counts tend to fluctuate around an “expected” value. This is called regression-to-the-mean and unless taken into account it may lead to misunderstanding of safety developments, especially over a short time period. Moreover, there are time dependences in the occurrence of crashes, both in the macroscopic level (long-term developments) and in the form of serial correlations. Also, the occurrence of crashes is not independent from the amount of exposure and other external factors.

The Empirical Bayes method (Hauer, 2001) is widely used in order to obtain the best estimate of the “expected” number of crashes. The Empirical Bayes method uses on the one hand a model-based estimate of crashes through a Safety Performance Function (SPF) developed on the basis of historical data from a large number of similar locations (taking thus into account long term trends and external factors), and on the other hand the observed value of crashes in a given location. A weighted estimate is made, and the weighting factor can be calculated for the specific sample, taking into account the size of the sample and the dispersion of the crash counts.

Risk Assessment Through Traffic Conflicts and Surrogate Safety Measures

A traffic conflict is defined as an event involving at least two road users, where one road user makes a nontypical maneuver (e.g., harsh maneuver, traffic rules violation) resulting in a crash avoidance maneuver from the other road user. Several studies have shown that the number and types of traffic conflicts correlate very well with actual crashes in many contexts (Polders and Brijs, 2018). Moreover, traffic conflicts are directly observable (although not always measurable in an easy and straightforward way), and therefore analyses can be proactive (i.e., identify risk scenarios before actual crashes occur) and/or allow a faster appraisal of the safety impact of a treatment, without waiting for several years of crash data to become available) (Ismail et al., 2010).

Several traffic conflict techniques are available, all based on criteria for observing and classifying traffic conflicts based on their criticality/severity. The indicators used are referred to as “surrogate safety measures” and typically include (but are not limited to): TTC—time to collision; PET—post-encroachment time, deceleration rates, and so on.

In the recent years, the availability of traffic observations through cameras, and the developments in video image processing have boosted the potential of traffic conflicts analyses, with numerous applications on intersection safety, interactions between vehicles and VRUs, and so on.

Road Safety Management

Structure of Road Safety Management Systems

The road safety management system is often described as a “pyramid” with five layers hierarchically structured as follows:

- Structural and cultural characteristics (i.e., policy input) at the bottom level, which reflects the “results focus” of stakeholders.
- Resulting safety measures and programs (i.e., policy output), at level 2.
- An intermediate layer specifying the operational level of road safety, through key performance indicators (KPIs) on issues like speeding, drinking, and driving, as well as state of the road network and the vehicle fleet.
- Final outcomes expressed in terms of road casualties are then observed (level 4).
- The top of the pyramid includes an estimate of the social costs of road crashes.

This concept was initially presented and further developed by different stakeholders within the decade 2000–10, and is still widely used for the planning of interventions for road safety improvements, although it concerns only a “snapshot” in time, and the time component has not been integrated (see Fig. 4).

Therefore, road safety management has a structural/institutional component, as well as an operational component. The structural component involves the establishment of a strong Lead Agency and an efficient intersectoral coordination (as road safety is a typically multisectoral problem). In this framework, a vision and strategy for road safety are the backbone of road safety policies (policy input). At the operational level (policy output), these are translated into specific programs and measures.

Effective road safety management systems need to be evidence-based. This implies the availability of data and tools for safety analysis. Evidence-based approaches can be quantitative or qualitative, but should follow the hierarchical process of road safety management. In this framework, KPIs are the key component and possibly the weakest link along the various stages of the evaluation—partly due to the difficulty in systematically collecting the necessary data.

Safe System Approach

While historically the emphasis of road safety policies was to tackle the “human error” as the main cause of crashes, in the last couple of decades a new approach is developed. The “safe system” approach shifts the traditional blame from the human, and introduces

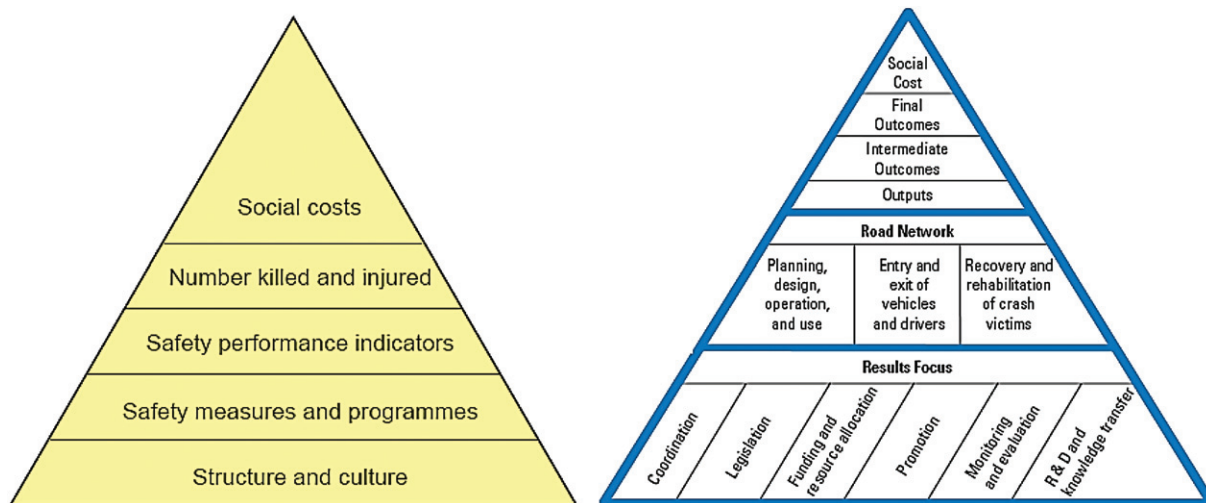


Figure 4 Generic structure of road safety management systems. Source: Koornstra et al. (2002) – left panel, Bliss and Breen (2009) – right panel.

the concept of “shared responsibility”: (1) system designers to accept responsibility for road users safety, (2) responsibility taken by vehicle manufacturers for investments in safety research, and (3) those that use the system accept responsibility for complying with the rules and constraints of the system.

The safe system approach underpins several road safety visions, namely the Swedish “Vision Zero” which, in addition to the shared responsibility concept, introduced the moral unacceptability of loss of human life, and the adaptability of roads, streets, and vehicles to human capacity and injury tolerance. The vision, although adopted by many countries, has also been criticized for socio-economic inefficiency (due to very high costs) and ethically questionable restrictions of freedom, privacy, and autonomy.

A milder version of the safe system approach has been expressed in the Dutch Sustainable Safety vision, in which the road infrastructure design by itself should inherently and drastically reduce crash risk (see Section 5). In addition, a “social” forgiveness dimension is introduced, including a willingness to anticipate and “forgive” a potentially unsafe action of other (less competent) road users.

Safety Impact Assessment

The road safety policy-making cycle includes a number of stages: (1) risk assessment/risk identification, (2) identification of causes of crashes, (3) selection and prioritization of measures/programs, (4) implementation of measures/programs, and (5) evaluation of the impact of measures/programs. The evaluation stage should also inform on new or pending risks, re-initiating thus the cycle.

The aim of road safety impact assessment is to yield quantitative estimates of crash/risk reduction. The most common relevant quasi-experiments are observational before-and-after studies (Hauer, 1997), in which the development of crashes in a treatment site/area is compared to that of a number of sites/a wider area with similar characteristic—in order to control for regression to the mean and other confounding factors. The estimated odds-ratio has as nominator the odds of crashes in the treatment site after the treatment to before the treatment, and as denominator the respective odds in the control group. The odds-ratio is also known as Crash Modification Factor (CMF); a CMF > 1 indicates an increase in crashes, while a CMF < 1 reflects a reduction in crashes. The percentage crash reduction is then calculated as $(1 - \text{CMF}) \times 100$. CMFs are also used (as multiplicative factors) in order to adjust a SPF on the basis of site-specific geometric and operational characteristics.

Once the safety impact of a measure is known, the socio-economic impacts are often estimated in monetary terms by means of cost-benefit analysis (CBA). CBA of safety treatments may be complicated due to lack of robust data on measures’ costs and benefits, and is generally criticized for ignoring the distribution of costs and benefits among different groups, the monetization of human life for statistical purposes, and other ethical issues. However, it remains a key tool for evidence-based approaches aiming to assess and compare policy alternatives.

Road User

Road users participating in traffic are exposed to risk in an asymmetric way. By nature, the difference in mass, protection and speed of the different traffic participants induces a different “baseline” fatality risk (Dupont et al., 2010): nonmotorized road users such as pedestrians and cyclists are far less protected than passenger car occupants. PTW riders combine the lack of protection with increased speed, exhibiting thus increased risk. On top of that asymmetric “baseline” risk, human factors, including personal characteristics (age, gender, experience, etc.), intrinsic characteristics (perceptions, attitudes, motivations, etc.) and behavioral characteristics (e.g., risk-taking, compliance, behavioral adaptation) significantly affect risk.

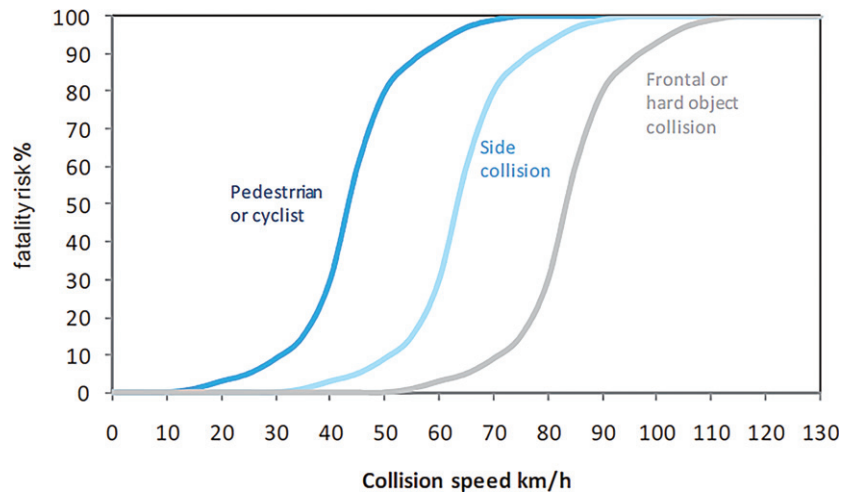


Figure 5 Human body tolerance to physical force. Source: Based on [Tingval and Haworth \(1999\)](#).

In general, males have a higher risk of crash involvement. Young people are more likely to be involved in crashes; this is due to a combination of inexperience (novice drivers) and age-related personality elements (e.g., low risk perception, underestimation of risk, over-estimation of skills, peer pressure, increased likelihood to drive under the influence of alcohol/drugs, and so on). Driver training and licensing procedures aim at reducing the risk of novice drivers, namely through “graduate licensing” schemes; an initial “probation” period is set for novice drivers, in which driving is allowed under specific condition for example, daytime only, accompanied driving, and so on.

Risk also increases for older people (>65 years old), as a combination of physical vulnerability (and thus increased risk of injury once involved in a crash) and physical or cognitive impairments that may be due to natural consequences of ageing, or medical/neurological conditions that are more prevalent among the elderly. The increased risk of elderly people is to some extent an artifact created by their reduced traffic participation (i.e., low exposure). Periodic medical testing is in place for older drivers, in most cases together with fitness-to-drive assessment procedures (regular or upon medical referral) with on-road tests at dedicated facilities. These fitness-to-drive assessments are often criticized for lack of objectivity in the evaluation, due to a lack of common framework, and concrete quantitative measures of driving performance.

Regarding behavioral factors, speeding is commonly known as the major risk factor affecting not only crash risk but also crash severity. The “safe system” approach stresses the human body tolerance to a collision as a function of impact speed (see [Fig. 5](#)). Driving under the influence of alcohol or drugs remains a major issue and crash contributory factor in most countries. The nonuse of restraint systems (seat belt, child seats) and protective devices (helmets, reflective clothing) are also major causes of road fatality.

Distraction and inattention have been consistently recognized as important crash contributory factors. Sources of distraction may be in-vehicle (operating navigation or entertainment systems, eating, or drinking etc.) or external (from observing landscape to simply daydreaming). However, the major source of in-vehicle distraction is the use of mobile phone while driving, which may increase crash risk by four times—and the use of hands-free devices does not appear to bring any benefit in most cases ([Caird et al., 2008](#)). Moreover, behavioral adaptation (a.k.a. risk compensation) through speed reduction is not sufficient to counterbalance the distraction-related impairments.

Human factors and road user behavior are mainly studied through three methods: (1) simulator studies, which allow for a controlled and ethically acceptable experimental environment, but with decreased physical validity, (2) naturalistic driving experiments, based on monitoring real driving behavior through instrumented vehicles, which allow for real driving patterns to be explored in detail, however with a high implementation cost and huge data processing needs, and (3) survey instruments aiming to understand intrinsic factors underlying the observed behavior, which are cost-effective but suffer from the known limitations of self-reported information.

The Road Environment

In general, in interurban/rural areas there are fewer number of crashes, but higher numbers of fatalities, whereas the opposite is the case in urban area crashes. There are well-established relationships between geometric design features, such as horizontal and vertical curvature, cross-section design (lane/shoulder width, superelevation, etc.), roadside treatments and pavement conditions, all taken into account in national road design guidelines.

Within the “safe system” approach (see Section 3.2), the concept of predictability of roads has emerged; this is defined as the consistency and continuity in road alignment, avoiding abrupt changes and thereby allowing drivers to properly align their expectations and behavior (“self-explaining roads”). This is combined with a concept of forgiveness in road design; “forgiving”

roads (and roadsides) are designed in a way that human error is anticipated and measures are taken so that this error does not result in a fatality or serious injury (e.g., obstacle-free zones, crash cushions, median separation).

The implementation of road design standards does not by itself guarantee safety (ITF, 2015). A continuous management of infrastructure safety is needed, through actions such as road safety audits, road safety inspections, identification and treatment of high risk sites, network safety management, and so on.

At the operational level, traffic flow and speed are the key factors that affect the level of safety. As shown in Section 2, the typical SPF implies a positive correlation between traffic and crashes. Moreover, speed management is one of the most important elements of infrastructure safety management. Speed limits should be credible, consistent, and relevant to the functional class of the road. Within a “safe system” approach, speed limits should prevent the exposure of humans in collision forces that would be physically intolerable. Especially in urban areas, there is a strong tendency to develop 30 km/h zones around areas with increased VRU traffic; this is often combined with traffic calming engineering treatments such as speed humps, raised crossings, woonerfs, and so on.

Environmental factors may also play a role: weather conditions have an impact on crash risk, although not always in a straightforward manner. Most studies suggest a risk reduction with precipitation, and a risk increase with extremely high or low temperatures. It is suggested that the decrease in crashes during rainy conditions is a result of reduced exposure.

The Vehicle

Although the vehicle has a minor role as crash contributory factor, the developments in active and passive safety have contributed to the substantial improvement of road safety outcomes. Active Safety Systems (also commonly known as Advanced Drivers Assistance Systems) play a preventive role in mitigating crashes by providing advance warning or additional assistance in steering/controlling the vehicle. Head-Up Display (HUD), Anti-Lock Braking Systems (ABS), Electronic Stability Control (ESC), Lane Departure Warning System (LDWS), Adaptive Cruise Control (ACC), Blind Spot Detection (BSD) and Night Vision System (NVS) are common Active Safety Systems – and most have been proved very effective in safety impact evaluations.

Passive Safety plays a role in limiting the injuries caused to road users in the event of a crash: airbags, seatbelts, Whiplash Protection System, and so on. In this context, vehicle crashworthiness, that is the ability of the vehicle to protect its occupants during an impact, has received particular focus through crash tests of typical scenarios. The New Car Assessment Programme (NCAP) and its star rating of vehicles provide a concise depiction of the level of safety of the current vehicle fleet.

In recent years, the vehicle is found at the center of road safety developments, through the important research and pilot testing of connected and automated vehicles (CAV). Six levels of automation are considered, ranging from no automation (Level 0), to driver assistance (Level 1), to partial automation (Level 2), to conditional automation (Level 3), to high and full automation (Levels 4 and 5). In all except Levels 4 and 5, the human driver maintains a strong supervising role over the autonomous system (SAE, 2018).

Safety has been a key “banner” value for the deployment of autonomous vehicles, but it is proved very difficult to responsibly claim it, as current evidence is insufficient (ITF, 2018). The challenges of mixed traffic conditions (conventional vehicles sharing the road with automated vehicles) remain to be addressed, as well as aspects related to the interaction of CAVs with pedestrians, cyclists, disabled people, and so on.

Most importantly, the interaction between human and machine, especially in intermediate levels of automation, is expected to create new safety issues. As humans are not good at passively monitoring systems, the so-called “out-of-the-loop” (OOTL) problem is highly likely. During OOTL, humans become inattentive and lose situational awareness of the driving task. This becomes critical when an authority transition is triggered by the system (e.g., due to an emergency situation, a sensor deficiency, or simply because the system reaches the boundaries of its Operational Design Domain). In this case, the takeover performance of human drivers appears to be compromised by several personal and external factors, with uncertain safety consequences.

Therefore, further research is required in order to prove the improved safety of CAVs and the conditions for their safe deployment. It is expected that an iterative process of increased penetration rate leading to increased trust, leading to further increase in penetration rate, and so on, will be required, until a critical mass of CAVs is reached—a point at which safety benefits are estimated to be maximized. On the other hand, increased trust to automation should be also understood as “properly aligned” trust; over-trust (over-reliance) to automation may lead to mis-use of the system, while the lack of trust may lead to dis-use.

Future Road Safety Perspectives

Despite the wealth of scientific knowledge available, road safety remains to some extent poorly understood: a typical example is the often misunderstood relationship of traffic exposure and speed with crash risk. Other key road crash contributory factors (i.e., alcohol, distraction) have not been fully addressed, and VRUs are not safely integrated in many urban traffic contexts.

From a policy perspective, despite important progress, there is need for further strengthening of road safety cultures and the “safe system” approach. Safety monitoring and evaluation are only fragmentarily implemented, often resulting in important investments not bringing results. Finally, road safety remains poorly integrated within other sectoral policies, namely mobility, energy/environmental and health. The above challenges are still relevant in all countries, and especially in LMIC.

New technologies and digitalization in transport have the potential of bringing dramatic safety improvements, however more research and evidence is required to responsibly manage the transition phase to higher levels of automation properly integrated

within new models of transport operations in urban and interurban areas. Regarding the safety of automated vehicles, it should be kept in mind that zero crashes cannot be envisaged - on the contrary, certain new crash mechanisms are to be anticipated.

Relevant Websites

WHO Road traffic injuries: https://www.who.int/violence_injury_prevention/road_traffic/en/ .
 The European Road Safety Observatory: https://ec.europa.eu/transport/road_safety/specialist_en.
 The European Road Safety Decision Support System: <https://www.roadsafety-dss.eu>.
 African Road Safety Observatory: <http://www.africanroadsafetyobservatory.org/>.
 AASHTO Highway Safety Manual: <http://www.highwaysafetymanual.org/Pages/default.aspx>.

References

- Bliss, T., Breen, J., 2009. Implementing the Recommendations of the World Report on Road Traffic Injury Prevention. Country Guidelines for the Conduct of Road Safety Capacity Reviews and the Related Specification of Lead Agency Reforms, Investment Strategies and Safety Projects. World Bank Global Road Safety Facility, Washington, DC.
- Caird, J.K., Willness, C.R., Steel, P., Scialfa, C., 2008. A meta-analysis of the effects of cell phones on driver performance. *Accid. Anal. Prev.* 40, 1282–1293.
- Dupont, E., Martensen, H., Papadimitriou, E., Yannis, G., 2010. Risk and protection factors in fatal accidents. *Accid. Anal. Prev.* 42 (2), 645–653.
- Hauer, E., 1997. Observational Before/After Studies in Road Safety. In: *Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*, Pergamon Press, Oxford.
- Hauer, E., 2001. Over dispersion in modeling accidents on road sections and in Empirical Bayes estimation. *Accid. Anal. Prev.* 33, 799–808.
- Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. *Accid. Anal. Prev.* 42, 1556–1565.
- Ismail, K., Sayed, T., Saunier, N., 2010. Automated analysis of pedestrian-vehicle conflicts: context for before-and-after studies. *Transp. Res. Rec.* 2198, 52–64, doi:10.3141/2198-07.
- ITF/OECD, 2015. Road Infrastructure Safety Management. ITF, Paris.
- Kopits, E., Cropper, M., 2005. Traffic fatalities and economic growth. *Accid. Anal. Preven.* 37, 169–178.
- Koornstra, M., Lynam, D., Nilsson, G., Noordzij, P., Pettersson, H-E., Wegman, F., Wouters, P., 2002. SUNflower: a comparative study of the development of road safety in Sweden, the United Kingdom, and the Netherlands. SWOV Institute for Road Safety Research, Leidschendam.
- Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accid. Anal. Preven.* 37 (1), 35–46.
- Papadimitriou, E., Yannis, G., Bijleveld, F., Cardoso, J.L., 2013. Exposure data and risk indicators for safety performance assessment in Europe. *Accid. Anal. Preven.* 60, 371–383.
- Polders, E., Brijs, T., 2018. How to analyse accident causation? A handbook with focus on vulnerable road users (Deliverable 6.3). Horizon 2020 EC Project, InDeV. Hasselt University, Hasselt, Belgium.
- SAE, 2018. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.
- Tingvall, C., Haworth, N., 1999. 'Vision zero: an ethical approach to safety and mobility. Road safety and traffic enforcement, Institute of Transportation Engineers international conference, 6th, 1999, Melbourne, Victoria, Victoria Police, Melbourne, Vic, 7 pp.
- WHO, 2018. Global Status Report on Road Safety. WHO, Geneva, 2018.
- Yannis, G., Antoniou, C., Papadimitriou, E., Katsochis, D., 2011. When may road fatalities start to decrease? *J. Safety Res.* 42, 17–25.
- Yannis, G., Papadimitriou, E., Folla, K., 2014. Effects of GDP changes on road traffic fatalities. *Safety Sci.* 63, 42–49.

Mobility Planning and Policies for Older People

Charles B A Musselwhite, Centre for Innovative Ageing, Swansea University, Swansea, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Importance of Mobility in Later Life	59
An Example of Positive Transport Policy for Older People—Free Bus Pass in the United Kingdom	60
Use of Life Stages in Transport Planning and Policy	60
Involving Older People in Policy-Making	60
Cross-Sector Working to Develop Shared Planning and Policy	61
Changing the Vision of Transport Policy	61
Transport Innovation, Older People and Policy	62
Conclusion	62
References	62

Importance of Mobility in Later Life

In almost all higher income countries across the globe, along with an increasing number of lower to middle income countries, a lower birth rate coupled with people living longer is creating an ageing society. In an ageing society, the population is getting older both in absolute numbers and in the higher proportion of older people within their total population. In 1950, there were only 384.7 million people aged over 60 years, representing only 8.6% of the global population [(UN (United Nations), 2015)]. There are now around 840 million people over 60 across the World, representing 11.7% of the population. By 2050, the world's population of people aged over 60 years will nearly double to 22%.

The way we live our lives as we age is also changing. Older people are more active and more mobile than ever before. They are more likely to be working and playing important roles in society than recent generations. Hence, transport is more important to older people than ever before. People become more dispersed from one another, having moved around for work or leisure throughout the life course, with connections all over the country and even internationally. As people age they want to stay connected to these different communities and also to family and friends that are widespread geographically and hence high levels of mobility are needed to achieve this. Coupled with this is mobility needed for older people to work or to support others who may have working or caring responsibilities. In high-income countries, some older people are able to have high amounts of leisure time and enough money to be highly active and hence highly mobile. Years of policies aimed at supporting high mobility, mean society is often geared supporting long distance travel, for example by car, high speed train or domestic airlines. Policies have led to shops and services agglomerating at the edge of city centres, along major trunk roads, freeways, or motorways, designed around access by car.

In many high-income countries life without a vehicle can become difficult and people can face social exclusion. In addition, the wider negative externalities of car dominant societies, such as air and noise pollution, severance of communities and road traffic collisions, injuries and deaths tend to be under acknowledged. For nondrivers, these are faced of course, with fewer benefits of using the car itself. Older people are more likely to want to reduce or give-up or some people end up having to give up their driving altogether (Musselwhite and Haddad, 2010; 2018). Hence, older people can result in being less mobile as they choose or have to drive less, which transport policy should be highly supportive of. One of the main aims of transport policy is to balance the negative externalities of the car with the benefits of hypermobility. Yet, older people tend to be forgotten in such policy, and transport policy has traditionally championed economic efficiency above other concerns. Policy has often assumed that space and distance are an inefficient inconvenience to be overcome, as quickly and efficiently as possible (Parkhurst et al., 2014). Concentration of transport policy and subsequent provision of resource is often geared around the working adult and inter- and intra-urban mobility with a focus on provision in core work hours 9–5 (Musselwhite, 2018a; Musselwhite and Curl, 2018). Over concentration on economic factors within transport (and associated) policy, means those viewed to be less economically active have their needs given lower priority. People in the United Kingdom aged 65–74 years are those most likely to be involved in voluntary activities (Parkhurst et al., 2014). Older people's work patterns may be different to those younger in age, whether in paid work or volunteering or caring for others, the need to travel often involve avoiding rush hour, if they can. Transport policy also prioritizes utility trips, for example, work, shopping, or visiting the doctors or hospital and space between these formal activities tends to be devalued (Cresswell, 2010; Cresswell and Merrimen, 2011). However, research suggests older people really value mobility for its own sake, to get out and about and see the world passing by, to pass the time of day in the company or being copresent with others, or of being in the world with others (Musselwhite, 2017). To incorporate older people's needs and behaviors within transport policy, a radical shift and a different approach is needed. This paper examines some different approaches that might have more success, including using a life stages approach to transport policy, a greater need to involve older people in transport planning, changing the vision, and understanding future mobilities.

An Example of Positive Transport Policy for Older People—Free Bus Pass in the United Kingdom

People aged over 65 years of age, or those of any age with a disability, have been entitled to travel free of charge on any off-peak local public bus service in England since 2007, thanks to the English National Concessionary Travel Scheme. Around 76% of all women and 79% of men take up the free bus pass in England, compared to 61% and 50% in 2005—the year before free local travel. [Humphrey and Scott \(2012\)](#) suggest that ownership of a free bus pass is higher (around 80%–82%) among those on lower income (less than £15,000). This group are also more likely to use the bus once a week than those on higher incomes who use it less frequently.

Unsurprisingly, the free bus pass increases bus use ([Mackett 2013a, 2013b](#)). The most commonly reported activity older people cite as their destination across all these surveys is shopping, followed by social and leisure, day trips, visiting friends then medical, meaning people are socially connected and hence reduce isolation ([Mackett 2013b](#)). [Mackett \(2013b\)](#) also notes how such journeys support volunteering and caring work older people undertake which would otherwise not take place. [Webb et al. \(2011\)](#) examined three waves of English Longitudinal Study of Ageing data (2004, 2006, and 2008) and found a link between eligibility of a free bus pass and increased use of buses. They also found those who used the bus more often also had a reduced chance of becoming obese, suggesting that using the bus is associated with physical activity such as walking to and from bus stops and allowing people to engage with more activity. [Dargay et al. \(2010\)](#) modeled bus use against what would have happened if no free bus pass had been introduced. They found journeys made are more numerous and also often longer in duration and distance ([Dargay et al., 2010](#)). The findings suggest the number of bus stages (groups of bus stops) traveled through by older people increased by 45.4% in rural areas and 26.5% in urban areas. Additionally, [Andrews \(2011\)](#) found 74% of his respondents stating that the free bus pass improved their quality of life and general well-being.

Use of Life Stages in Transport Planning and Policy

It has been noted that transport planning or policy should be placed around life stages (e.g., retirement from work, becoming a grandparent, and becoming a carer), but even within this context age per se is very rarely mentioned in policy discourse ([Avineri and Goodwin, 2010](#)). Significant events or transition points in people's lives present an important opportunity for intervening at some or all of the levels, because it is then that people often review their own situation and potentially change their behavior. Typical transition points include: leaving school, entering the workforce, becoming a parent, becoming unemployed, retirement, or bereavement ([NICE, 2007](#)).

Research has identified that transition points such as retiring from work can offer opportunity for changes in mobility patterns and even modal choice including the chance to try out alternative modes including walking or cycling ([Chatterjee and Scheiner, 2015](#); [Goodwin, 1989](#); [Jones et al., 2015](#)). [Burningham and Venn \(2017\)](#) describe this as a “moments of change” theory where disruption of norms of behavior can create opportunities for change of behavior and where life-course transitions such as retirement can trigger changes in a variety of aspects of everyday life such as energy use, travel and purchase of consumer goods. They also note that life-course transitions may be a point at which individuals pause and consciously reflect on the life style they want and make changes to move toward that life style.

Involving Older People in Policy-Making

[Phillips and McGee \(2018\)](#) note, “transport policies have traditionally been developed based on government priorities rather than community needs, resulting in a lack of services” (pg 232–233). They relate this in particular to two examples of indigenous populations being marginalized by current transport provision, citing older Maori people being excluded through infrastructure and design issues and land use and transport planning and older people from the indigenous populations of Australia being isolated through lack of culturally appropriate transport in rural areas. However, it is equally true of other places across the globe. In the United Kingdom, as in many other high-income countries, older people are marginalized from society by poor transport planning not meeting their needs. This is compounded further for older people in rural areas ([Mollenkopf et al., 2004](#); [Parkhurst et al., 2014](#); [Shergold et al., 2012](#)), those in deprived communities ([Buffel et al., 2012](#)) and those from black, Asian, and minority ethnic groups in urban areas of the United Kingdom ([Higgs and Gilleard, 2015](#); [Owen, 2013](#); [Rees et al., 2012](#)) all whose needs are even less considered in transport planning. To get more community involvement there should be not just a shift in attitude from transport providers but a whole change in philosophy, to recognize the value of more community involvement; to recognize that transport is about people and community first and foremost, and technologies, economics and engineering second. It also requires both the time and skill of individuals to get involved, along with the desire to want to get involved. For this to happen a number of different ways to get involved are needed along with reciprocation of getting something back from being involved; people must be able to see their input resulting in change.

Older people also want more involvement in transport planning and decision making ([Musselwhite, 2018a](#)). There is an increasing emphasis on involving people in policy making thereby democratizing decision-making. Public involvement in services, through collaborative codesign are beginning to take shape in urban design, urban regeneration, and in some aspects of town and country planning, for example, collaborative, redistributive urbanism and intergenerational urbanism ([Buffel et al., 2015](#);

Buffel, 2018). It is highly recommended that transport follows these models. In addition, other stakeholder groups from third sectors and charities need more involvement in transport planning to champion the needs of older people.

Many of the issues older people face in the transport system would benefit from engaging more closely with older people in policy making. This could be done in a variety of different ways. It is suggested that local groups could be mobilized to collect information on their local area and mapping key issues that can then be used by local policy and practice when redesigning and planning the local area. First this could be done through individual or group audits of local streets. Older people could audit their local streets identifying the general age friendliness of the area using tools such as OPERAT (Older People's External Residential Assessment Tool; Burholt et al., 2016) which identify need for crossings, speed control, such as traffic calming, speed cameras, etc., need for improvements and decluttering of pavements, and identification of the need for cycle paths. To complement auditing, local groups could organize interactive mapping sessions to plot and note and discuss key issues in a local community in relation to transport. These could be carried out annually or be in response to a new initiative or event. They can be bottom-up and driven by the community or top-down and introduced as a form of consultation from local authorities or transport service providers. To support this, new technology can help with the capture of transport issues in real time, helping to geolocate issues and capture them and aggregate them easily and efficiently. The problem with involvement includes representation. Certain groups of older people are much more likely to take part in such exercises and there is a need to mobilize and involve other sectors of the community who do not normally take part in such exercises. Hence, different techniques aimed at getting different people involved are suggested, for example some people enjoy group work, others would rather monitor the local area from home or at distance. Monitoring who has taken part and who has not against known local community figures, shows who is engaging and who is not and who needs to be targeted. Barriers to involvement need to be examined and where possible managed or reduced.

In addition, there is a need for the involvement process to feel less tokenistic. As well as information coming from older people themselves, information must be fed back from the decision makers, the local policy, or transport service providers themselves. That is not to say that every issue needs a comment but there needs to be some two-way dialogue taking place with some significant feedback coming from the authorities. Obviously managing expectations is important. It cannot be given that people will have all of their issues satisfied but a feeling that they are being listened to and a dialogue opened up to be honest about why certain issues are priorities and not others. It is suggested that the local authority dedicate some web space to such community groups to allow space for discussion and dialogue.

Cross-Sector Working to Develop Shared Planning and Policy

Transport does not sit in isolation from the rest of an individual's life. It is entwined with other societal influences and barriers. A particular issue that would help older people would be better scheduling of healthcare appointments, so they do not take place when older people cannot access them. An examples of this is reflected in hospital appointments before 9 am when the concessionary bus pass is not applicable and travel is not free. In addition, there is often much waiting around for appointments then waiting for hospital or community transport to be coordinated on the way from appointments. Hence, it is suggested that healthcare helps by asking about and coordinating patients travel arrangements as best as possible. There is potential here for new technologies to help with the scheduling of transport and potentially sharing of transport services for people accessing similar healthcare appointments and travel to neighboring residential locations. Where there are known events taking place that are put on for older people, transport information, and even transport provision needs to be well-thought about and brought together.

In 2014, the United Kingdom Transport Select Committee made a series of recommendations to improve transport services to isolated communities, including the idea of "total transport" which involved better use of existing transport resources, including hospital and school transport. For example, it was suggested that at times when vehicles are not in use, they could be used to fill gaps in local provision by taking a more coordinated logistical approach. Devon County Council was an initial pilot for Total Transport, working to improve nonemergency patient transport services in partnership with the local Clinical Commissioning Group. As well as making better use of local resources there is a central information service for both those who qualify for patient transport and those who do not—who are directed to community transport options.

Changing the Vision of Transport Policy

Another aspect to consider for transport planners is to start to vision the world differently. Too much transport planning sees the future landscape as business as usual, with only marginal changes. Huge changes are ahead due to an ageing population. It is possible that there will be a significant increase in the amount of mobility that older people are undertaking, both in total, as there are more older people within society, but also more mileage per person, as older people end up being fitter and healthier. The current 50–64 year olds in the United Kingdom, for example, spend the highest number of miles driving per person of any age group (DfT, 2019) and so are likely to want to carry on this mobility into later life, resulting in huge increases in the number of older drivers or the number of people giving up driving and using other modes. Older people in high-income countries are also likely to be working longer and later on in life than currently, putting additional strain on the transport infrastructure. How to manage this demand will require significant resources. Older people are likely to be more heterogeneous than ever, there will be a variety of people working

full-time and those who have retired from work, those very fit to those with long-term chronic conditions affecting health and mobility. Hence, there is more reason here to involve older people in making decisions about transport.

There is often an assumption that interventions aimed at older people need to offer a like for like solution to mobility. For example, the need to provide literal mobility to replace lost literal mobility as someone gives-up driving. But as Parkhurst et al. (2014) note, actual literal mobility may be met nowadays through virtual mobility by electronic means, online shopping, staying connected to family and friends via computer technologies, and mobile and telehealth solutions (Marston et al., 2019). In addition, Musselwhite (2018c) notes that older people may not want to be as mobile as they have been, and may replace literal mobility with imaginative mobility, thinking, and reminiscing about mobility or observing others' mobility from their home. Underpinning this, though, is a need to feel that they can be mobile at any point, the potential for mobility is important to older people (Metz, 2000). This level of perceived control over mobility helps older people feel more secure if something changed, or they faced an emergency, often related to health issues of themselves, or a friend or family member (Musselwhite, 2018b).

Transport Innovation, Older People and Policy

Looking ahead there are potential changes in transport innovation that may have huge effect on older people (Shergold et al., 2015). These may have different effects on different groups of older people. Autonomous vehicles are quite popular among older people wanting to maintain their hypermobile life, especially among those facing the prospect of giving-up driving (Musselwhite, 2019). However, this could further increase the distances needed for travel, especially if take up of autonomous vehicles is large. This could have a detrimental effect on local neighborhood services and shops. Thus, autonomous vehicles could change interpersonal spaces; the spaces for human exchange outside of home environments which is vital for community life (Musselwhite, 2019). Older people spend greater amounts of time in their local neighborhood closer to home and community and social cohesion which may be decimated if hypermobility is maintained or even enhanced by autonomous vehicles. There is a need to more closely involve older people in future transport scenario development, to examine how such changes as autonomous vehicles might impact on their lives, both positively and negatively. At present the discussions are largely technology focused and there is a need for a more nuanced discussion involving the impact of social aspects of such technology.

Conclusion

Overall, older people's needs and issues are not well represented in transport policy. This is because transport policy has traditionally championed economic productivity associated with transport, coupled with a mistaken belief that older people are less productive and their activity therefore is less essential to support. There is also the idea that older people might not need to have their transport needs looked after. Older people are now more likely than ever to be involved directly in work, continuing to work later on in life than previous generations. Also, older people's mobility supports other economic activity or reduces the economic burden of society, for example supporting the work of others by looking after grandchildren or volunteering and caring for others. Older people also want to be active. They want to stay connected to communities and to family, and to leisure pursuits and transport has to help them fulfill these needs, especially in a hypermobile society geared around the needs of highly mobile, car drivers. However, it is important to be wary of introducing planning and policy that simply fosters growth in hypermobility. In a world of balancing health and environmental sustainability, policy, and planning need to be more social in nature, more nuanced and hence involve end users more.

It is therefore suggested that older people's transport needs must be addressed through policy. As we live in an ever-ageing population, it is necessary to get an understanding of older people's transport issues and needs, and involve older people in policy formation aimed at resolving some of the issues and planning for a more age-friendly transport system.

References

- Andrews, G. (2011). *Just the Ticket? Exploring the Contribution of Free Bus Fares Policy to Quality of Later Life*. A thesis submitted in partial fulfilment of the requirements of the University of the West of England, Bristol, for the degree of Doctor of Philosophy.
- Avineri, E. and Goodwin, P. (2010) *Individual Behaviour Change: Evidence in Transport and Public Health*. London: Department for Transport.
- Buffel, T., 2018. Social research and co-production with older people: Developing age-friendly communities. *Journal of Aging Studies* 52–60.
- Buffel, T., Asumu, J., Bromley, R., Bysouth, R., Downing, A., Gem, J., Goulding, T., Greenmantle, F., Hibberd, H., Kaur, R., O'Mahony, M., Page, R., Ricard, C., Singh, D., Unegbu, E., Williams, B., 2015. Researching Age-Friendly Communities: Stories from older people as co-investigators. The University of Manchester Library, Manchester.
- Buffel, T., Verté, D., De Donder, L., De Witte, N., Dury, S., Vanwing, T., Bolsenbroek, A., 2012. Theorising the relationship between older people and their immediate social living environment. *Int. J. Lifelong Edu.* 31 (1), 13–32., doi:10.1080/02601370.2012.636577.
- Burholt, V., Roberts, M.S., Musselwhite, C.B.A., 2016. Older People's External Residential Assessment Tool (OPERAT): A complementary participatory and metric approach to the development of an observational environmental measure. *BMC Public Health* 16, 1022.
- Burningham, K and Venn, S. 2017. Moments of Change — Are lifecourse transitions opportunities for moving to more sustainable consumption? CUSP Working Paper No 7. Guildford: University of Surrey. Available from: www.cusp.ac.uk/publications.
- Chatterjee, K., & Scheiner, J., 2015. Understanding changing travel behaviour over the life course: Contributions from biographical research. Paper presented at 14th International Conference on Travel Behaviour Research.

- Cresswell, T., 2010. Towards a politics of mobility. *Environ. Plan. D: Soc. Space* 28 (1), 17–31., doi:10.1068/d11407.
- Cresswell, T., Merriman, P., 2011. *Geographies of Mobilities: Practices, Spaces, Subjects*. Routledge, London.
- Dargay, J., Ronghui, L. and Toner, J., 2010. Concessionary bus fares in England: The impact on bus travel by the over-60s. Presented at *European Transport Conference*, Glasgow, Scotland.
- Department for Transport (DfT), 2019. Transport statistics 2018. Department for Transport, London. Available from: <https://www.gov.uk/government/statistics/transport-statistics-great-britain-2018>. (Last Accessed 26 June 2020)
- Goodwin, P.B., 1989. Family changes and public transport use 1984–1987. *Transportation* 16 (2), 121–154.
- Higgs, P., Gilleard, C., 2015. Rethinking old age: Theorising the fourth age. Palgrave Macmillan, London.
- Humphrey, A. and Scott, A., 2012. Older people's use of concessionary bus travel, report by NatCen Social Research for Age, UK, London, England.
- Jones, H., Chatterjee, K., Gray, S., 2015. Understanding change and continuity in walking and cycling over the life course: A first look at gender and cohort differences. In: Scheiner, J., Holz-Rau, C. (Eds.), *Räumliche Mobilität und Lebenslauf - Studien zu Mobilitätsbiografien und Mobilitätssozialisation*. Springer-VS, Wiesbaden., pp. 115–132.
- Mackett, R., 2013a. The impact of concessionary bus travel on the wellbeing of older and disabled people. Paper presented at the Transportation Research Board 92nd Annual Meeting, Washington DC, 13–17 January. *Transp. Res. Rec.* 2352, 114–119.
- Mackett, R., 2013b. The benefits of a policy of free bus travel for older people. *Proceedings of 13th world conference on Transport Research*, Rio de Janeiro, 15–18 July.
- Marston, H.R., Genoe, R., Freeman, S., Kulczycki, C., Musselwhite, C., 2019. Older adults' perceptions of ICT: Main findings from the technology in later life (TILL) study. *Healthcare* 7, 86.
- Metz, D.H., 2000. Mobility of older people and their quality of life. *Transp. Policy* 7 (2), 149–152., doi:10.1016/S0967-070X(00)00004-4.
- Mollenkopf, H., Marcellini, F., Ruoppila, I., Széman, Z., Tacken, M., Wahl, H.-W., 2004. Social and behavioural science perspectives on out-of-home mobility in later life: Findings from the European project MOBILATE. *Eur. J. Age.* 1 (1), 45–53., doi:10.1007/s10433-004-0004-3.
- Musselwhite, C.B.A., 2017. Exploring the importance of discretionary mobility in later life. *Work. Older People* 21 (1), 49–58.
- Musselwhite, C.B.A., 2018a. Community connections and independence in later life. In: Peel, E., Holland, C., Murray, M. (Eds.), *Psychologies of Ageing: Theory Research and Practice*. Palgrave Macmillan, Cham, Switzerland, pp. 221–252.
- Musselwhite, C.B.A., 2018b. Mobility in later life and wellbeing. In: Friman, M., Ettema, D., Olsson, L. (Eds.), *Quality of Life and Daily Travel. Applying Quality of Life Research (Best Practices)*, Springer, Cham, Switzerland, pp. 235–251.
- Musselwhite, C.B.A., 2018c. Virtual and imaginative mobility: how do we bring the outside indoors and what happens when mobility is no longer available? In: Musselwhite, C. (Ed.), *Transport, Travel and Later Life (Transport and Sustainability, Volume 10)*, Emerald Publishing Limited, Bingley., pp. 197–205.
- Musselwhite, C.B.A., 2019. Older people's mobility, new transport technologies and user-centred innovation. In: Müller, B., Meyer, G. (Eds.), *Towards User-Centric Transport in Europe – Challenges, Solutions and Collaborations*, Springer, Lecture Notes Series, Switzerland, pp. 87–103.
- Musselwhite, C.B.A., Curl, A., 2018. Geographical perspectives on transport and ageing. In: Curl, A., Musselwhite, C. (Eds.), *Geographies of Transport and Ageing*, Palgrave Macmillan, Cham, Switzerland, pp. 3–24.
- Musselwhite, C., Haddad, H., 2010. Mobility, accessibility and quality of later life. *Qual. Age. Older Adults* 11 (1), 25–37.
- Musselwhite, C.B.A., Haddad, H., 2018. Older people's travel and mobility needs a reflection of a hierarchical model 10 years on. *Qual. Age. Older Adults* 19 (2), 87–105.
- NICE (National Institute of Health and Clinical Excellence) (2007). Behaviour change at population, community and individual levels (Public Health Guidance 6).
- Owen, D., 2013. DR18: Is there evidence that national, and other, identities change as the ethnic composition of the UK changes? If so, how? Are there differences between people from different ethnic groups? And of different ages/social classes? *Future Identities: Changing identities in the UK – the next 10 years*. London: Government Office for Science.
- Parkhurst, G., Galvin, K., Musselwhite, C., Phillips, J., Shergold, I., Todres, L., 2014. Beyond transport: understanding the role of mobilities in connecting rural elders in civic society. In: Hennessey, C., Means, R., Burholt, V. (Eds.), *Countryside Connections: Older people Community and Place in Rural Britain*, Policy Press, Bristol, pp. 125–175.
- Phillips, J., McGee, S., 2018. Future ageing populations and policy. I. In: Curl, A., Musselwhite, C. (Eds.), *Geographies of Transport and Ageing*, Palgrave Macmillan, Cham, Switzerland, pp. 257–287.
- Rees, P., Wohland, P., Norman, P., Boden, P., 2012. Ethnic population projections for the UK, 2001–2051. *J. Populat. Res.* 29 (1), 45–89., doi:10.1007/s12546-011-9076-z.
- Shergold, I., Parkhurst, G., Musselwhite, C., 2012. Rural car dependence: an emerging barrier to community activity for older people? *Transp. Plan. Technol.* 35 (1), 69–85.
- Shergold, I., Lyons, G., Hubers, C., 2015. Future mobility in an ageing society – where are we heading? *J. Transp. Health* 2 (1), 86–94.
- UN (United Nations), 2015. *World Population Ageing*. United Nations, New York. Available from: http://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf. (Last Accessed 7 April 2020).
- Webb, E., Netuveli, G., Millett, C., 2011. Free bus passes, use of public transport and obesity among older people in England. *J. Epidemiol. Comm. Health* 66 (2), 176–180.

Electric Mobility

Graham Parkhurst, University of the West of England, Bristol, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Electric Mobility: From Niche Solution to Emergent New Regime?	64
Emergence of Electric Mobility Niches	64
Innovation and Diversification to Circumvent the Storage Deficit	65
The Re-Emergence and Rise of the Battery–Electric Vehicle	67
A Sustainable Mobility Policy Perspective on Electric Mobility	69
Biography	71
References	71

Electric Mobility: From Niche Solution to Emergent New Regime?

Electric mobility (EM) technologies have been part of the mechanized and motorized transport system for some 130 years, but over the last quarter century EM has been propelled from a position of niche applications, albeit including some critical ones, to becoming the central tenet of a technology-led strategy for the transport sector to solve its growing contribution to the paradigmatic problem of global heating (Morgan, 2020; Parkhurst and Seedhouse, 2019).

The article frames an analysis of the changing role of EM with concepts from the “sociotechnical transitions” literature, and in particular that fundamental societal change occurs only when one “sociotechnical regime” (Geels, 2004) replaces another, the regime encompassing sociocultural, technological, scientific, market and policy facets and being the dominating set of practices and technologies at a given time. The regime can also be understood as part of a “multi-level perspective” with “niches” of alternative activity existing as “protected” from the stabilizing influence of the regime, and allowing for innovation, while the regime interacts with a “landscape” of features, which are resistant to influence by individual actors, such as the configuration of the built environment and infrastructure, or the dominant belief systems of society. The growing priority given to global heating is a contemporary example of shifting beliefs. Drawing on these concepts, the article evaluates the new importance given to EM to establish how far this change is part of a fundamental revision in the conduct of transport policy—a regime change—or simply the substitution of one technology for another, with the effect of stabilizing the current regime.

The following section considers how EM technologies emerged in a 19th century context in which innovations were also occurring in respect of power from steam and from internal combustion engines (ICE). From a position of technical competition, electrification retreats into specific niches as a regime emerges around the ICE, which in turn influences the landscape, reinforcing its stability. The third section then examines these niches in terms of their technical solutions, but concludes that the most significant challenge to the ICE regime by EM will rely on the rechargeable battery. Finally, the last section examines the case that EM is likely to displace the ICE regime, focusing on the key role of the private, owned, passenger car, as the dominant mode within a landscape which is physically car dependent and features strongly-engrained socioeconomic practices which maintain the car’s centrality.

Emergence of Electric Mobility Niches

Transport-sector applications were amongst the first practical implementations of the direct-current electric motor which emerged as an industrial technology in the 1820s and 1830s. From the start, the history of electric mobility cut across transport modes but was immediately in competition with other powertrain technologies.

The first demonstrations of electric rail traction were undertaken by Robert Davidson from 1837, little more than a decade after the Stephenson brothers inaugurated steam locomotion on the pioneering Stockton–Darlington railway in northern England (Day and McNeil, 1996). Although the battery technology deployed by Davidson would eventually become a rarity, electric rail transport would in the 20th century become the most important set of EM niches.

The development trajectory of small electric watercraft shares commonality with the rail systems. Experimentation began in the 1830s into battery power for small boats via paddlewheel or propeller, and they subsequently became commonplace until ICE power dominated after 1920 and electric power retreated to a few small niches, such as service on inland bodies of water where environmental regulations prohibited ICE or steam power.

Although early battery technologies created significant limitations on the application of EM, in the road domain steam power faced very significant safety and practicality constraints on networks prioritized for horse-drawn traffic, while ICE development was subject to a number of technical barriers. Indeed, the technologies underpinning battery–electric (BE) and ICE road vehicles—both niche modes in this period—show similar development trajectories through the middle half of the 19th century.

In the case of the electric vehicle (EV), the invention of the rechargeable lead-acid battery by Gaston Planté in 1865 was a critical step, with Thomas Parker's first BE car of 1884 thought to have preceded Karl Benz's first ICE car by a year, with Parker's first electric bus following in 1886 (Guarnieri, 2012). A photograph of Parker's car shows a four-wheel "horseless carriage" with three passengers aboard and an altogether more solid appearance than Benz's tricycle; a difference which presages the next 20 years of development. The first vehicle of any kind on road or rail to exceed 100 km h^{-1} , in 1899, was an electric car. More battery cars were in circulation in 1900 than ICE cars, with performance both in range (30–50 km) and speed $15\text{--}25 \text{ km h}^{-1}$ being competitive with contemporary ICEs. Electric vehicles provided taxi services in New York and postal deliveries in France, where electric car production was most developed. Guarnieri (2012: 6) identifies electricity as the "high tech" fashion of the time, with the electric car as a key symbol of the success of positivist science, offering a "silent, odorless, reliable, simple to drive, and easy to start" alternative to both ICE and steam road vehicles. Ferdinand Porsche is primarily remembered for his design of the Volkswagen Beetle and the eponymous sportscar company, but his first car, in 1899, was electric.

However, during the 1900s the "balance of power" between the two technologies began to change. Notable, given that it remains a competitive disadvantage a century later, electric cars in the 1910s were costly not only in absolute terms, but in the United States at least, bore a two-and-a-half-to-five times price premium over ICE cars, and were therefore a luxurious, exclusive option (Guarnieri, 2012). On the other side of the balance sheet, discoveries of plentiful crude oil led to falling gasoline prices while technological improvements to ICE vehicles improved their reliability and comfort, whereas further enhancements to battery technology to enhance EV range were not forthcoming. Guarnieri places particular emphasis on a specific feature during the 1910s—ironically an ancillary electrical device—the electric starter, which replaced the manual starting procedure of hand-cranking.

Another contextual factor was the improvement of interurban roads. During the second half of the 19th century, the railways offered by far the most efficient means of interurban freight and passenger movement and demand for long-distance road traffic was low and their development and maintenance only a local, not national, concern or responsibility (Plowden, 1971). The passenger car and the freight delivery truck were primarily local, urban vehicles; a situation which suited the capabilities, and in particular the range, of the BE vehicle. As road vehicle performance improved, however, the demand for paving highway networks to ease the passage of pneumatic tires, rather than surfaces which facilitated horses' hooves, strengthened (Plowden, 1971) and from the 1920s countries such as Italy, the United States and Germany pioneered high-speed, high-capacity highways, for long-range traffic, with access restricted to motor vehicles, a technology which would be a growing source of competition with the railways.

The rise of Fordist mass production methods from 1908 seems only to have enhanced the affordability of the ICE vehicle, and set in train intense technological development of the automotive industry around the ICE to become one of the most significant global industrial sectors and spawning a coevolution of society and car known variously as car culture, automobility (Urry, 2004), or car dependence (Mattioli et al., 2020). In a similar way to which tramways and suburban railways had facilitated urbanization around "transit corridors," ICE road vehicles encouraged suburbanization and commuter dormitory towns; development forms which locked-in dependence on medium-range motor transport by road. EVs "retreated" to niches typically characterized by specific environments and duty cycles, such as the before-dawn, door-to-door, stop-start, local deliveries of milk traditional in British residential neighborhoods, for which quiet operation was essential.

Hence, the ICE, and private car in particular, came to dominate the regime of mobility, after 1920 even influencing the wider landscape. While several factors favored the ICE in the road sector, a key competitive disadvantage for EM was that vehicles in motion need a mobile power source. Like solid coal burnt in an external combustion engine, or petroleum-derived liquid fuels exploded in an ICE, electricity is an energy transfer "vector." However, unlike coal or gasoline, its storage needs are technologically complex and their development has lagged that of the other facets of the electric powertrain, such as the electric motor. Davidson's early demonstrations involved battery power from expensive non-rechargeable batteries (Day and McNeil, 1996). Rechargeable batteries followed, but their development has represented a struggle for greater "energy density": to increase the amount of energy that can be carried for the weight and size of the battery, in order to increase the range that BE vehicles can travel between charges. The following section considers technical solutions which have emerged to address the storage problem and the niches they created.

Innovation and Diversification to Circumvent the Storage Deficit

Given some clear advantages with electrification - the absence of exhaust emissions, low noise, and the simplicity and reliability of electric motor design - it was natural that solutions would be sought to combine electric traction with an alternative to complete reliance on battery storage.

The first of these solutions was to provide an intermittent or permanent electrical contact between the mobile vehicle and its fixed infrastructure. Werner von Siemens demonstrated the first externally-alimented electric locomotive in 1879. By this time, steam traction had spread across the industrializing world, but a range of technologies for tramways and railways followed, initially relying on electrified rails, later mostly deploying overhead pantographs. While underground steam railways powered by coke or smokeless coal had been introduced in London from the 1860s, the smoke-free basis of electric traction made it a clear preference for this form of railway, with the first electric line in London opening in 1890. Electrification would also become the power source of choice for surface suburban commuter rail markets, where it has been preferred for technical performance in a frequent-stop service regime as well as its environmental credentials. Similarly, whereas steam-powered street trams had occasionally been deployed, it was the electric tram and street car networks that replaced the horse tram and enabled the late 19th and early 20th century urban expansions.

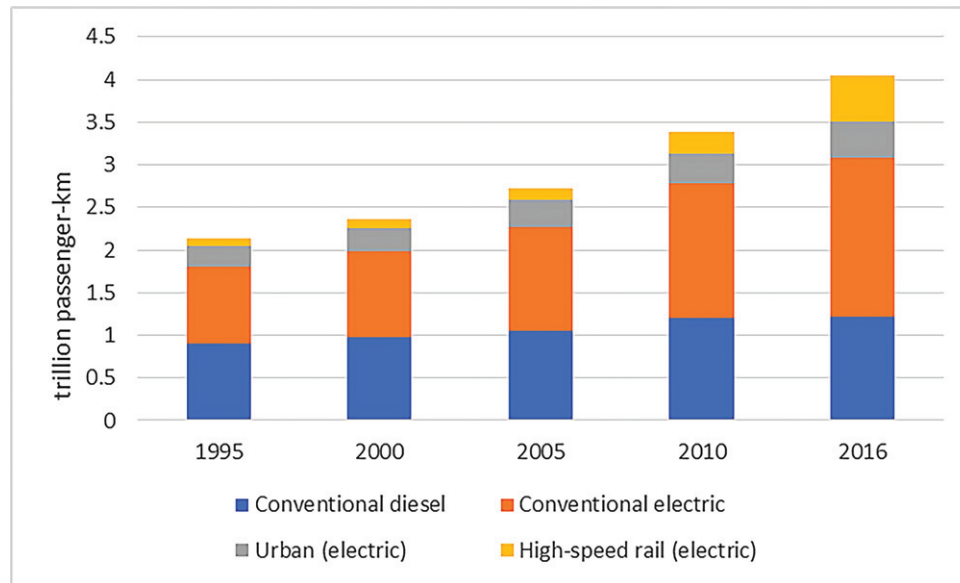


Figure 1 Rail passenger-km by electric and diesel-powered services for selected years. *Source: Based on data from the International Energy Agency (IEA) (2019). The future of rail: opportunities for energy and the environment, IEA, www.iea.org/statistics. All rights reserved; as modified by the author.*

The first interurban electrified line was opened in Italy in 1902. By 2016, 70% of rail passenger-km globally were on electric services (Fig. 1), although the share for freight ton-km was closer to 50% and the extent of adoption of electrification for long-distance railways shows a much more varied position globally, depending on national context and policy. In some countries, such as the United States and United Kingdom, with domestic coal and oil reserves and powerful associated industries, electrification had low priority, with the high capital investment in pantograph infrastructure judged not to be worth the long-term benefits in terms of energy savings and maintenance costs, and in those jurisdictions steam was mainly replaced by diesel. In Japan and much of continental Europe, in contrast, where there has been greater dependence on fossil fuel imports, and sometimes significant domestic nuclear or hydroelectric generation, electric power is near-universal.

Some states without full electrification are now seeking to retrofit as part of their decarbonization plans, and some form of EM will be the motive power of choice for all new-line rail investments. Indeed, since the 1960s the development of high-speed rail, defined as lines achieving maximum speeds in excess of 225 km h^{-1} , is with few exceptions (Such as the UK's "High Speed Train" developed in the 1970s to run on hybrid diesel-electric power) a project intimately related to advances in the electronics of power supply and control. This relationship has been further cemented by the realization of a limited number of "maglev" guided transport systems, such as the Shanghai-Shanghai Airport shuttle in China, which exploit the long-established principles of the electric linear induction motor. If realized, Elon Musk's "Hyperloop" would place maglev systems in partial-vacuum tubes to reduce air resistance in order to reach theoretical speeds of 1200 km h^{-1} , creating competition with air travel.

Away from railways there are few examples of electric supply via fixed contact, and they share a "corridor"-basis, often have exclusive alignments, and have networks of modest total length and complexity. They include urban "trolleybuses" operating in a similar service niche to trams, but making use of roads rather than rails, and a few very rare circumstances of ferry or tug operations on confined water passages such as canals. However, there is renewed interest in the last few years in electrifying roadway infrastructure. Experiments have taken place with the powering of heavy freight vehicles via pantograph on a highway in Sweden (Mendelsohn, 2016), which amounts to the transfer of a tested rail technology, as well as more radical proposals for the inductive charging of EVs on the move (Jang, 2018), although significant technical and capacity barriers would need to be addressed to realize such a system at wide scale.

The second approach to the problem is to provide a source of onboard power generation to avoid the need for energy storage. The highest profile application has been the use of diesel ICEs or gas turbines in railway locomotives to produce electricity for traction. The most extreme, but also one of the most long-standing solutions has been, uniquely in the marine sector, applications of nuclear reactor technology, requiring the substitution of the radioactive fuel rods only a few times in a vessel's lifespan. The submarine grew in significance as a weapon during the 20th century, with its importance fundamentally linked to it being hard to detect while underwater. However, while a snorkel device can be used to supply air to and extract exhaust gases from an ICE for immediate subsurface operation, this is not practical for deep immersion. Hence, the limitations of battery power on submerged range led, from 1955, to the application of nuclear power generation. Subsequently the technology was extended to a small number of large oceangoing electric surface vessels, mainly warships, but with a handful of civilian "icebreakers" developed for use in the Arctic by the Soviet, then Russian, states.

A far more affordable and applicable mobile power generation source than nuclear technology is the hydrogen fuel cell (FC), producing electricity by electrolysis rather than fission, although requiring frequent refueling with a source of hydrogen. FCs have a

technical heritage to rival batteries but their first high-profile application was in providing electrical power onboard spacecraft. Experimentation into terrestrial transport applications mainly emerged in the current century, with a number of limited-production fuel-cell cars having been demonstrated or offered to a small number of consumers under lease arrangements. The first model with a significant production run and being available to purchase outright (at least in some markets) was the Toyota Mirai, from 2014, offered at a premium over ICE vehicles similar to that of BE vehicles. By 2020, sales amounted to around 5000 globally. The advantage of the FC EV is that it largely replicates the ICE vehicle user-experience; the refuel time of hydrogen pressured to 70 MPa is around 4 min and gives a range of 500 km. However, the sales figures reflect the limited availability of hydrogen refueling infrastructure globally; an infrastructure the development of which is in competition with BEV charging facilities, and in second place for investment funds.

The FC solution is also dependent on the provisioning of carbon-neutral hydrogen. While in principle this can be achieved using renewable electricity to electrolyze water, established industrial hydrogen production is based on reforming natural gas, producing carbon dioxide in the process. Efficiency losses occur in the production process, and also from compressing hydrogen to facilitate the transport of practical quantities. Advances in the storage of hydrogen in nanoporous membranes could reduce both compression costs and the need to carry hydrogen in heavy tanks (Noguera-Díaz et al., 2016), but for now these practical factors combined with lingering concerns about the safety of hydrogen mean battery technology remains the more attractive option for most light road vehicle applications.

However, if BE technology is a constraint for light-duty vehicles, it is a far greater one for heavy-duty vehicles such as buses and trucks, where the weight of the batteries necessary to provide for even short-range operation cuts into the payload and comes at very high cost. This has not prevented over half a million EV buses entering service in China, but only a few thousand in total elsewhere in the world (IEA, 2020), sometimes with in-service top-up charging at stops and termini. Indeed, the case for FCs remains potentially strong for niches which involve fleet operation from designated depots, and particularly for long-haul heavy goods vehicles. FC applications for rail vehicles are also emerging where permanent way electrification is not financially or technically viable, and are under development even for aircraft, where weight is particularly critical, and there are safety concerns about the lithium batteries. Here, electrification is not only a pathway to decarbonization, but could also avoid the release of water vapor at high altitude, as this is an additional source of the “greenhouse” effect.

The third approach has been to adopt hybrid technologies whereby a vehicle draws upon a battery for part of its energy needs, and this is combined, either in concert or in alternation with, another power source. In 1997, recognizing that battery technology was not commercially viable for a saloon-sized, high-specification vehicle, Toyota launched the first mass-produced hybrid-electric car, the Prius, using electric power to replace or assist the most inefficient operating phases of the ICE, that is, running at slow speed, accelerating and decelerating. Initially hybrid EVs ultimately obtained all their energy from the liquid fuel and had negligible zero-emissions operation, but by 2020 most hybrids had batteries which could be pre-charged from the electric grid. By the time, the fifth-generation Prius model was launched it had a battery-only range of 40 km and an increasing share of primarily ICE vehicles had some hybrid capability. The Toyota Mirai is in fact an FC-BEV hybrid.

However, while hybridization has been a useful transitional technology, and will remain so in locations without adequate BE infrastructure, it comes with the higher cost and vehicle weight of duplicating the powertrain, and increases complexity over the ICE, rather than reducing it as a pure EV does. There are also concerns about the extent to which consumers actually make use of the pre-charging capability of “plug-in” hybrids (PH) or mainly run them on fossil liquid fuels. Hence, for many commentators concerned with road transport, hybrid technology, and for that matter FCs or electrifying the infrastructure, remain less effective and deliverable solutions than enhancing BE performance. The following section considers the “emergence of a trajectory of electric mobility” (Dijk et al., 2013: 135) led by BE solutions, focusing on the electric car.

The Re-Emergence and Rise of the Battery-Electric Vehicle

Similarly to the way in which the original electric vehicle technology shadowed the rise of electrical science, the re-emergence of the EV from the 1990s reflected advances in electronics including semiconductors and microprocessors used in power electronics from the 1950s and the lithium-ion battery in the 1980s, which increased capabilities and performance. The first half of the 1990s also brought political salience to the matter of urban air pollution from ICEs, and by the end of the decade tackling “climate change” was a dominant idea in policy rhetoric, if not matched in intensity by policy action.

However, while battery car performance was improved in this early period of the re-emergence, the available vehicles were not “substitute good” competitors for ICEs. Consumers could essentially choose between small production-run electric adaptations of ICE models, mostly vans intended for urban delivery use, or microcars such as the Reva/G-Wiz; a spartan design with a range of 80 km and maximum speed of 80 km h⁻¹, and not legally defined as a car in some jurisdictions. A breakthrough for the BE vehicle came in 2010, when Nissan launched its Leaf model; the first purpose-designed, battery-only, compact car from a mainstream manufacturer and intended for mass production. From an initial specification providing a range of around 120 km, by 2020 the capability of second-generation vehicles was up to 270 km and cumulative sales were reaching half-a-million cars.

While Japan has been an important locus of the development and commercialization of EV technologies, certain states in the United States, notably California, have also been an important market both for development and consumption, where the adoption of noxious emissions controls gave an important regulatory emphasis to low and zero-emission vehicle sales. As well as being important as a launch-market for vehicles such as the Prius and Leaf, California has seen the founding of Tesla, an exceptional case of

a private-sector start-up managing to break into a crowded automotive sector to become a significant “disrupter” of its business model.

Formed in just 2003, Tesla had its limited-production Roadster derived from an existing ICE model available around the same time as the Leaf, followed by its fully purpose-designed Model S in 2012. On most metrics (market capitalization, EV production) Tesla is the most successful EV producer to date, and in 2020 was the most valuable automotive company of all power-types, although had only recently reached consistent profitability. Tesla’s disruptive credentials arise conceptually from confounding the belief that it was impossible for a start-up to break into an expensive and competitive industry, and in technical terms from being a vertically-integrated manufacturer *and* retailer. Alongside these were radical design choices such as the use of carbon-fiber polymer bodies and a new approach to battery construction achieving higher energy-density. From the consumer perspective, product design and marketing choices emphasized performance and significant battery range (with even the standard Model S having a theoretical range of 335 km), despite the implications for pricing. Combined with features such as “automation-ready” control systems and the establishment of a bespoke network of exclusive recharging facilities, much higher speed than the typical public-sector backed facilities, the result has been the development of a premium brand with strong social desirability and customer loyalty, even if the commercial strategy is to target future models to the more affordable mainstream as technology costs fall.

Alongside Tesla and the Japanese manufacturers, China has emerged as a third force in EV production, notably through manufacturers such as BYD Auto, which has specialized on EV production ranging from city cars through to trucks. China has also become the largest market for EVs in the world. Fig. 2 summaries the global stock of electrified cars in 2019. It should be noted, however, that despite a recent 40% year-on-year growth rate, plug-in vehicles represent just 1% of the global fleet.

Although the sales of Chinese and South Korean EVs had focused on price competition and had been directed towards the emerging and developing economies, imports to Europe were of growing significance from 2020. In Europe itself, industry resistance to change combined with a policy focus on reducing carbon emissions through the adoption of more energy-intensive diesel ICEs has left the continent’s manufacturers seeking to catch up, even if a crisis in air quality, ironically exacerbated by a “dieselization” policy that limited EV development, is now stimulating demand.

The passenger car and the light commercial vehicle are the critical vehicles in transport policy in the advanced industrialized states. Even in countries associated with cycling, such as The Netherlands, the modal share of travel measured in passenger-km is dominated to the order of 80% by cars (Parkhurst et al., 2012). And linked with changes including the rise of digital commerce, light goods traffic has shown particularly sharp increases. Nonetheless, electrification is also associated with the rise of new, or re-envisioned modes of transport, facilitated by the enhanced power-to-weight ratio of improved and affordable lithium-ion batteries and digital control technologies which render EV technologies efficient and can provide extra value via consumer interfaces, such as theft warning and optimization advice.

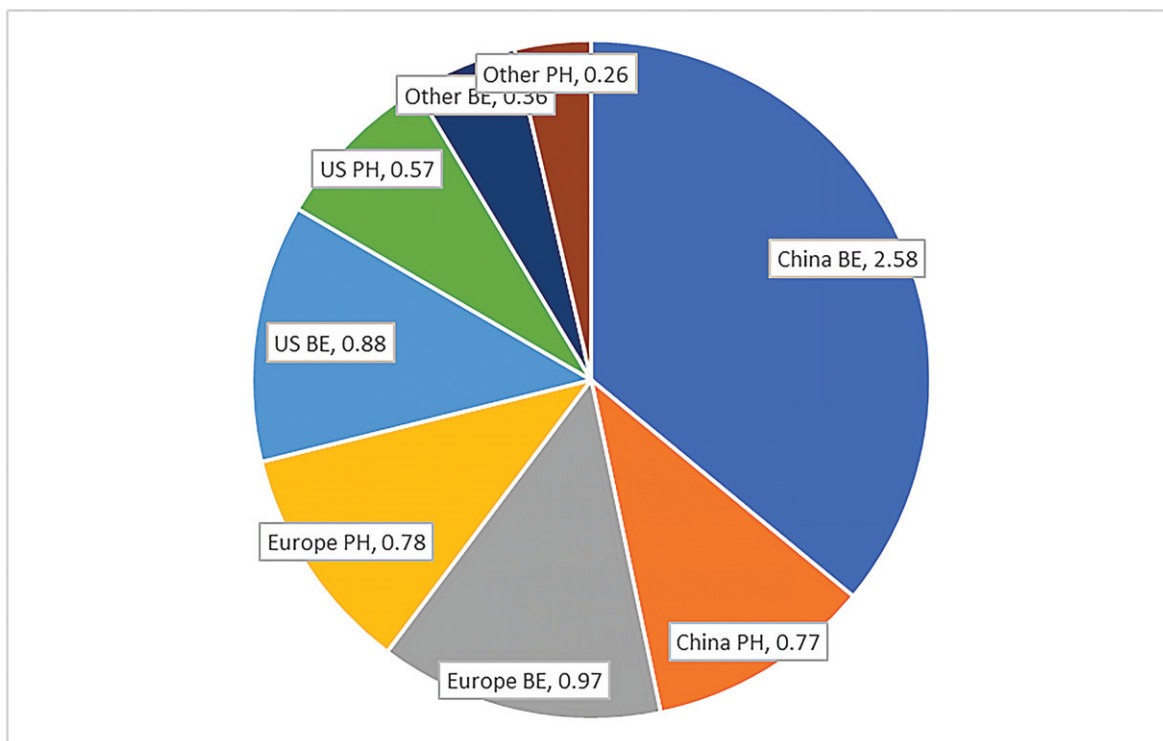


Figure 2 Global stock of BE and PH cars in 2019 (millions). Source: Based on data from IEA (2020). Global EV Outlook 2020, IEA, www.iea.org/statistics. All rights reserved; as modified by the author.

One successful area of application of electrification has been in respect of existing two-wheeled vehicles. Bicycles and motorcycles, particularly the lower-powered urban “mopeds” and scooters, have typically had a modest speed of operation over relative short distances. Combined with lithium battery technology, and low vehicle weight compared to cars, these EVs can easily achieve their duty-cycles, while the batteries can be detached from the vehicle for charging at a standard power outlet indoors, or swapping for charged units at a service center. (An attempt was made by US–Israeli firm “Better Place” to establish a similar battery-swap scheme for passenger cars. Pilots began in 2012 in Israel and Denmark but the business model did not prove viable due to the scale of investment required to establish a swap network at wide geographical scale and the difficulty of attracting car manufacturers willing to accept battery standardisation to the specification of a small start-up (Gunther, 2013)).

Motorcycles constitute a significant share—up to two-thirds—of vehicle traffic in some Asian cities so substituting that vehicle class with EVs offers significant air and noise pollution reduction benefits as well as offering the potential for decarbonization. The electric bicycle, instead, is usually a hybrid of human and electric-power and has great potential to open up the mode of cycling for people for who find unassisted cycle journeys too hard work, or too long, but at the same time encourages cycling as a form of physical exercise, for transportation or leisure. The latter is an important non-transport policy consideration given the crisis in the wealthy countries around the “diseases of affluence” and inactivity, such as obesity, heart disease and diabetes. Even in countries in which cycling is an important transport mode, such as the Netherlands, e-bike sales now have significant market share, as a means of cycling longer journeys.

A second area of application is in respect of “personal” mobility or “micromobility”. It has long been argued that the passenger car is over-specified for most journeys, made by sole travelers for short distances, and that smaller single-person vehicles, typically electrically powered, could offer a more efficient alternative. Indeed, light-weighting is a particularly effective strategy in the case of EM, as vehicle weight has a greater influence on efficiency than for ICE vehicles, for which engine power is more important (Weiss et al., 2020). However, high-profile failures (such as the Sinclair C5) and highly-qualified successes (the Segway) underline that the established two-wheel vehicle options already offer an alternative to the private car that is hard to improve on for many traveler-journey options. Nonetheless, the recent emergence of the “e-scooter” (confusingly the same term scooter has been applied to rather different technologies; in this case is meant the vehicle with two small in-line wheels and a standing platform with vertical handlebar) is deserving of attention. Facilitated by automated short-term hire systems, the vehicles have become a feature of the central areas of many major cities in recent years, where they offer an alternative to walking and other transport modes for short journeys. While promoted for their energy efficiency compared to other powered modes, concerns have emerged about their safe use either when mixing with pedestrians or road traffic, the longevity of the vehicles when offered in public sharing schemes, and problems with inconsiderate parking at the end of hire. However, given their lower capital cost compared with e-cycles and greater ease of carriage and storage, e-scooters have strong advocates and the mode benefitted from the emphasis placed on personal rather than collective mobility triggered by the Covid-19 pandemic outbreak in 2020.

Freight equivalents of personal mobility have also been beneficiaries of electrification, as well as automation technologies. The need for zero-emissions local freight has boosted interest in cargo-bikes in various formats in recent years, and three-quarters of these purchased in Europe now have electric assistance (cyclelogistics.eu, 2020). (The same technology is also highly appropriate for cycle taxis or rickshaws, where “pullers” in low average income states often suffer from excessively physically demanding work.) However, due to the labor-intensity of urban deliveries, combined with consignment sizes which are often small, the company Starship is developing delivery-bot services, using a secure wheeled box with a maximum payload of around 50 kg and a maximum range of 6 km, achieved at walking speed and using automated navigation and control systems.

A third area of innovation is essentially to take these surface EV developments to the skies, in the form of “urban aviation,” to take advantage of the vertical dimension of space as a means to reduce congestion. According to this vision, short-range, low-altitude, light BE aircraft, typically powered by multiple rotors, could offer both passenger and freight services. Remote-controlled “drone” operation is long-established in military applications and has seen diversification into specific applications such as emergency medical logistics in rural areas. Transferring this technology to the urban consumer sphere, and potentially automating its operation, raises a number of safety, privacy, noise, and regulatory concerns, including the fundamental issue of overflight rights, all of which remain to be solved. Additionally, the energy efficiency and commercial viability of such services remains to be proven.

A Sustainable Mobility Policy Perspective on Electric Mobility

So far in the article, EM has been established as a bundle of transport modes and services that many commentators hope will transform the impact of our mobility on the environment while ensuring high accessibility for all. It has been seen how ingenious technological development is enabling the possibility of motive power which is efficient and emissions-free at the point of use for the entire global transport system. But a number of questions remain to be answered and challenges overcome before EM can be confirmed as being the core technology for a more sustainable future, and indeed one that contributes to a transport regime change.

The first of these concerns just how much more efficient EM will be compared to existing options? This is important as decarbonization must involve two related changes in order to be deliverable in time to meet global climate management needs: all energy must be sourced from carbon neutral sources, and to make that achievable in a context of growing populations and economies, energy must be used more sparingly. For EVs the answer to this question is reasonably clear in theoretical terms: based on current technologies a pure EV car is around 60%–66% more efficient than its ICE equivalent in use (UK Department for Transport, 2018), with those benefits arising, despite relatively heavy like-for-like weight, from the relative efficiency of the motor,

the elimination of “idling” and features such as the regeneration of energy under braking. But even with this significant benefit, for a wealthy, highly-industrialized economy the additional demand on the electric supply infrastructure will be very significant; in the case of the UK completely electrifying the transport system will require a doubling in demand (Eyre and Killip, 2019), as well as major challenges to manage peaking in demand across the key sectors: transport, domestic, and industrial. Here “smartgrid” techniques are emerging to ensure BE vehicles charge at periods of low domestic and industrial demand, with power potentially being offered back to the grid from individual vehicle batteries at peak times if the charge state is higher than necessary for expected immediate travel needs. Enthusiastic, expert consumers can already take part in such a market, but it remains to be seen how far the typical consumer is willing and able either to plan needs in a tariff-led system or simply leave it to the algorithms to decide.

The key uncertainty regarding the other part of the efficiency-neutrality tandem concerns the significant variation globally in the dependence of regions and nation states on high-carbon electricity. Including the emissions which derive from producing an ICE vehicle or an EV and disposing of it at the end of its life, and assuming the current global mix of fuels used in electricity generation, then the whole lifecycle greenhouse gas (GHG) impact of an EV is around 70% of an ICE equivalent (Elgowainy et al., 2020). A 30% saving is certainly worth having but is not in itself a transformative change. Shifting to carbon-neutral electricity could improve that saving to at least 60% of an ICE equivalent (assuming the ICE continues to operate on fossil fuel). EV technology is not the only “vehicle-fuel pathway,” which can theoretically deliver this magnitude of reduction, but it is currently the most deliverable option for replacing the fossil-fuel ICE regime, and within the options for electrification, if the challenge of providing sufficient battery-electric recharging infrastructure is significant, at least a well-developed electric supply grid exists, whereas the supply of hydrogen for FCs requires a refueling system to be provided almost from nothing.

Even a 60% saving in GHG emissions must be qualified by forecasts globally for very significant expansions in both vehicle fleets and distance travelled, which will to some extent offset the benefits of electrification (Greene and Parkhurst, 2017; Morgan, 2020). Indeed, EVs themselves are a factor predicted to increase traffic for both economic and psychological reasons. EVs are relatively expensive to buy but relatively cheap to use. The high investment in an EV is likely to “lock-in” owners to car use, while the negative elasticity of travel demand with respect to cost of car use implies traffic growth due both to their efficiency and the low cost of (often low-tax) electricity. “Rebound effects” of this nature can be expected to be further encouraged by EVs being perceived as an environmentally-benign technology; arguments in favor of public transport or active travel may carry less weight with the owners of EVs.

Even if there is high global success in decarbonizing electricity for motive power, emissions from vehicle provisioning will remain important (Morgan, 2020). For these reasons, it is argued that decarbonization must not become overly-reliant on technological substitution, and that an evolving sociotechnical regime of sustainable mobility should also involve changes in individual behaviors and social practices to moderate the rising demand for vehicles and travel. But herein lies a political and commercial conundrum which undermines the principles of the smart mobility revolution. Purchasers of “early modern” BE cars understood they were making a lifestyle change, often for overt reasons of environmental politics. They would adopt energy-conserving driving styles to maximize battery range and did not expect the EV to provide for all the journeys they expected to make, relying instead on the EV as part a diverse multimodal “mobility style.” Precisely in order to widen and deepen the appeal of EVs, though, significant research and development has been invested in both larger and more energy intense battery packs and faster charging facilities both to reduce the “range anxiety” of users and facilitate long-range journeys through 20-min recharging capability. As well as providing business-as-usual for the consumer, such a strategy preserves as much of the 20th century business model of the ICE automotive industry as possible. Alongside extended range, the “business-as-near-usual” powertrain continues to promote the “owner-driver” model of car use as well as perpetuating the equipping of private cars with performance capabilities in terms of acceleration and maximum speed which exceed practical requirements and most legal constraints. Although Tesla seeks to emphasize differences in design and production, it is through this same overarching business model that Tesla has become the most successful producer of electric cars globally to date.

Greater range also removes the incentive for energy-efficient driving; indeed, the relatively high torque of electric motors may result in a rebound effect whereby part of the efficiency gain is reinvested in more “sporting” driving. As exhaust emissions are reduced, non-exhaust sources of pollution become relatively more important, but as vehicle fleets grow in number and weight they also increase in absolute terms (Air Quality Expert Group, 2019). A more aggressive driving style also accelerates the rate of tire and friction brake wear that arises from EVs as well as ICE vehicles, in turn increasing the significant contribution of automotive sources of particulates and microplastics to environmental pollution (Evangelidou et al., 2020).

Another facet of policies to promote EVs as a direct substitute for ICE technology is that governments around the world have introduced fiscal and regulatory incentives to reduce the perceived barriers to becoming an owner-driver of a BE. Norway is the prime example (Figenbaum, 2017), where 42% of passenger cars newly registered in 2019 were zero-emission, and another 26% hybrids (Norwegian Road Traffic Information Council, 2020). The range of incentives applied around the globe has variously included grants towards the vehicle purchase price; income-related tax incentives; grants for home, workplace, and public-domain charging; exemptions from tolls, emissions charges, and parking fees; even the permission to use bus lanes, despite the fact congestion caused by an EV is just as delaying as that caused by an ICE vehicle, and that the decarbonization case for public transport solutions is strong considering both emissions in use (Logan et al., 2020) and across the lifecycle (Elmers et al., 2020). However, rather than re-envisioning the passenger car as just one modal contributor to a “Mobility-as-a-Service” environment, with vehicles hired for specific trips, and even ride-shared with strangers (Parkhurst and Seedhouse, 2019), providing subsidies and incentives for EV use continues a policy tradition of not expecting private road users to meet the costs they externalize, and prioritizes supporting volume of production in the automotive sector. Looking to a potential future with vehicles which are automated as well as electric, a similar

prospect beckons: while it is conceivable that these could be an efficiently-used collective fleet, a more plausible outcome is that market forces and social practices will perpetuate overwhelming dependence on the private car. Indeed, there is already evidence that the innovative niches of more sustainable mobility, such as shared e-bikes, will be short-lived (Lin et al., 2018).

In summary, the article has reviewed how electric propulsion emerged in the first half of the 19th century, and by the latter part was serving specific niches in the transport sector, some of them important enablers of urbanization. EM niches have been characterized by particular operating environments, or specific kinds of delivery service, or duties in industrial sites, or in submarines, or where the conditions have been particularly suited to major electric infrastructure provision, notably on the railways. Recently, a resurgence in EM has also supported the emergence of entirely new modal niches, such as personal transporters, and adapted niches, such as the case of civilian drone applications close to the functions of helicopters. However, much of the growing emphasis placed on EM represents technology substitution; substitution into a regime, that is to say a set of social, spatial, and industrial-economic conditions, which has coevolved to optimize synchronicity with other technologies, mainly the liquid-fueled ICE. Hence, multiple barriers need to be overcome both through developing the technology, but also, it must be argued, reshaping the regime.

To conclude, it is just possible that further rounds of technological innovation and infrastructure investment could result in effectively limitless carbon-neutral energy. Similarly, this plentiful energy might be combined with much enhanced storage, for example if long-life, stable and recyclable metal-air batteries can be achieved then the theoretical energy density is ten times higher than current lithium-ion batteries (Imanishi and Yamamoto, 2019; Liu et al., 2017), and concerns about the supply of sufficient critical raw materials are resolved (Narins, 2017). However, the precautionary principle urges that we cannot risk that these circumstances do not prevail. We need to use our motorized vehicle fleets more efficiently (higher load-factor per km), more intensively (fewer vehicles per capita), more sparingly (greater modal share to human-powered transport) and over shorter distances (spatial and economic planning to reduce the need to move people and goods), as well as electrifying our mobility.

Biography



Dr Graham Parkhurst is Professor of Sustainable Mobility and Director of the Centre for Transport & Society at UWE Bristol. He holds degrees in Psychology and Philosophy (BA University of Warwick), Human Biology (MSc University of Oxford) and Transport Geography (DPhil University of Oxford), and has researched and taught academic transport and mobility studies since 1991. Past research interests include urban and subregional transport policy, modal interchange policy, air quality policy, mobility of the ageing population, transport policy instruments and the evaluation of urban transport policy implementations (specific infrastructure interventions, mobility services, and vehicle technologies). His current research is examining the wider implications of the trends to greater automation, electrification, flexibility, and use of digital technologies in the transport sector, taking a critical lens to the discourse and practices of "smart mobility" and "smart cities", through projects such as *Driverless Futures?* (Economic and Social Research Council) and *Replicate* (European Commission). Graham has provided social and behavioral research contributions in respect of UK-Government-funded (Innovate UK) research consortia examining connected and automated vehicles (*Venturer*, *Flourish*, *CAPRI*, *MultiCAV*, *CAVForth*) and flexible collective transport solutions (*Mobility on Demand Laboratory Environment*). Graham also teaches Transport Policy at Masters' level and is a Fellow of the Royal Geographical Society.

References

- Air Quality Expert Group, 2019. Non-Exhaust Emissions from Road Traffic. UK Department for Environment, Food and Rural Affairs, London. https://uk-air.defra.gov.uk/library/reports.php?report_id=992 (accessed 27 October 2020).
- cyclogistics.eu, 2020. First European Cargo Bike Industry Survey. CityChangerCargoBike, Brussels/Berlin. <http://www.cyclelogistics.eu/market-size> (accessed 27 October 2020).
- Day, L., McNeil, I., 1996. In: *Biographical Dictionary of the History of Technology*. Routledge, London, p. 195.
- Dijk, M., Orsato, R.J., Kemp, R., 2013. The emergence of an electric mobility trajectory. *Energy Policy* 52, 135–145, doi:10.1016/j.enpol.2012.04.024.
- Elgowainy, A., Han, J., Ward, J., Joseck, F., Gohlke, D., Lindauer, A., Ramsden, T., Biddy, M., Alexander, M., Barnhart, S., Sutherland, I., Verdusco, L., Wallington, T., 2016. Cradle-to-grave lifecycle analysis of U.S. light-duty vehicle-fuel pathways: a greenhouse gas emissions and economic assessment of current (2015) and future (2025–2030) technologies. *ANL/ESD-16/7*. Argonne National Laboratory, Argonne, Illinois. <https://greet.es.anl.gov/publication-c2g-2016-report> (accessed 27 October 2020).
- Elmers, E., Dietz, J., Weiss, M., 2020. Sensitivity analysis in the life-cycle assessment of electric vs combustion engine cars under approximate real-world conditions. *Sustainability* 12, 1241, doi:10.3390/su12031241.
- Evangelou, N., Grythe, H., Klimont, Z., et al., 2020. Atmospheric transport is a major pathway of microplastics to remote regions. *Nat. Commun.* 11, 3381, doi:10.1038/s41467-020-17201-9.
- Eyre, N., Killip, G. (Eds.), 2019. *Shifting the focus: energy demand in a net-zero carbon UK*. Centre for Research into Energy Demand Solutions. Oxford, UK. <https://www.creds.ac.uk/publications/shifting-the-focus-energy-demand-in-a-net-zero-carbon-uk/> (accessed 27 October 2020).
- Figenbaum, E., 2017. Perspectives on Norway's supercharged electric vehicle policy. *Environ. Innov. Soc. Transit.* 25, 14–34, doi:10.1016/j.eist.2016.11.002.
- Geels, F.W., 2004. From sectoral systems of innovation to socio-technical systems: insights about dynamics and change from sociology and institutional theory. *Res. Policy* 33 (6–7), 897–920, doi:10.1016/j.respol.2004.01.015.
- Greene, D., Parkhurst, G., 2017. Decarbonizing transport for a sustainable future - mitigating impacts of the changing climate: a White Paper. Conference Proceedings 54, Transportation Research Board, Washington, DC. <http://www.trb.org/Main/Blurbs/177088.aspx> (accessed 27 October 2020).
- Guarnieri, M., 2012. Looking back to electric cars. In: *Proceedings of Third IEEE HISTORY of ELECTRO-technology CONFERENCE (HISTELCON)*, Pavia, Italy, pp. 1–6, <https://doi:10.1109/HISTELCON.2012.6487583>.
- Gunther, M., 2013. Better place: what went wrong for the electric car startup? *The Guardian*. <https://www.theguardian.com/environment/2013/mar/05/better-place-wrong-electric-car-startup> (accessed 27 October 2020).
- Imanishi, N., Yamamoto, O., 2019. Perspectives and challenges of rechargeable lithium-air batteries. *Mater. Today Adv.* 4, 1–12, doi:10.1016/j.mtadv.2019.100031.

- IEA, 2019. The Future of Rail: Opportunities for energy and the environment. IEA, Paris. <https://www.iea.org/reports/the-future-of-rail> (accessed 27 October 2020).
- IEA, 2020. Global EV Outlook 2020: Entering the decade of electric drive? IEA, Paris. <https://www.iea.org/reports/global-ev-outlook-2020> (accessed 27 October 2020).
- Jang, Y.J., 2018. Survey of the operation and system study on wireless charging electric vehicle systems. *Transport. Res. C: Emerg. Technol.* 95, 844–866, doi:10.1016/j.trc.2018.04.006.
- Lin, X., Wells, P., Sovacool, B.K., 2018. The death of a transport regime? The future of electric bicycles and transportation pathways for sustainable mobility in China. *Technol. Forecast. Social Change* 132, 255–267, doi:10.1016/j.techfore.2018.02.008.
- Liu, Y., Sun, Q., Li, W., Adair, K.R., Li, J., Sun, X., 2017. A comprehensive review on recent progress in aluminum–air batteries. *Green Energy Environ.* 2 (3), 246–277, doi:10.1016/j.gee.2017.06.006.
- Logan, K.G., Nelson, J.D., Hastings, A., 2020. Electric and hydrogen buses: shifting from conventionally fuelled cars in the UK. *Transport. Res. D: Transport Environ.* 85, doi:10.1016/j.trd.2020.102350.
- Mattioli, G., Roberts, C., Steinberger, J.K., Brown, A., 2020. The political economy of car dependence: a systems of provision approach. *Energy Res. Social Sci.* 66, doi:10.1016/j.erss.2020.101486.
- Morgan, J., 2020. Electric vehicles: the future we made and the problem of unmaking it. *Cambr. J. Econ.* 44 (4), 953–977, doi:10.1093/cje/beaa022.
- Mendelsohn, T., 2016. Sweden trials electrified highway for trucks. <https://arstechnica.com/cars/2016/06/sweden-trials-electrified-highway-for-trucks/> (accessed 27 October 2020).
- Narins, T., 2017. The battery business: lithium availability and the growth of the global electric car industry. *Extractive Ind. Soc.* 4, 321–328, doi:10.1016/j.exis.2017.01.013.
- Noguera-Díaz, A., Bimbo, N., Holyfield, L., Ahmet, I., Ting, V., Mays, T., 2016. Structure-property relationships in metal-organic frameworks for hydrogen storage. *Colloids Surf. A. Physicochem. Eng. Aspects* 496, 77–85, doi:10.1016/j.colsurfa.2015.11.061.
- Norwegian Road Traffic Information Council, 2020. Car sales in 2019. <https://otv.no/bilsalget/bilsalget-i-2019> (accessed 27 October 2020).
- Parkhurst, G., Kemp, R., Dijk, M., Sherwin, H., 2012. Intermodal personal mobility: a niche caught between two regimes. In: Geels, F., Kemp, R., Dudley, G., Lyons, G. (Eds.), *Automobility in Transition?: A Socio-Technical Analysis of Sustainable Transport*. Routledge.
- Parkhurst, G., Seedhouse, A., 2019. Will the 'smart mobility' revolution matter? In: Docherty, I., Shaw, J. (Eds.), *Transport Matters*, Policy Press, Bristol, pp. 359–390 (Chapter 15).
- Plowden, W., 1971. *The Motor Car and Politics 1896–1970* Bodley Head London.
- UK Department for Transport, 2018. The Road to Zero. UK Department for Transport, London. <https://www.gov.uk/government/publications/reducing-emissions-from-road-transport-road-to-zero-strategy> (accessed 27 October 2020).
- Urry, J., 2004. The 'system' of automobility. *Theory Culture Soc.* 21 (4–5), 25–39, doi:10.1177/0263276404046059.
- Weiss, M., Cloos, K.C., Helmers, E., 2020. Energy efficiency trade-offs in small to large electric vehicles. *Environ. Sci. Eur.* 32, 46, doi:10.1186/s12302-020-00307-8.

Transport Policy and Governance

Lisa Hansson, Faculty of Logistics, Molde University College – Specialized University in Logistics, Norway

© 2021 Elsevier Ltd. All rights reserved.

The Governance Turn in Transport Policy	73
Understanding Transport Governance—Key Features	73
Future Challenges	75
References	76

The Governance Turn in Transport Policy

This chapter aims to describe the development of governance in transport policy studies and show central aspects that have emerged within this field. The paper draws from studies in transport as well as general policy theories on policy making and governance.

Governance research in the transport sector can be traced back to the privatization reforms that were introduced to this sector in the 1980s. During this period, many countries gradually introduced different types of deregulation elements. The regulatory reforms initiated in the United Kingdom, where competition was introduced into the public transport sector in London in 1984 and the rest of the country in 1985 (Gwilliam, 2008), can be seen as the starting point of changes in policy making within the transport sector. In this discussion, the new role of the government was set in focus; it emphasized its role in deregulating the sector or withholding authority in terms of having a strategic position within contracting regimes (Beesley & Glaister, 1985; Gwilliam et al., 1985). Hence, in the 1980s transport policy studies were strongly focused on aspects related to market regulation and competition (Gwilliam, 2008). In the 1990s, studies discussing the forms and mechanisms of competition emerged, linking organizational reforms to output (Hensher & Wallis, 2005; Gwilliam, 2008). In parallel with privatization reforms, there was an increased focus on sustainable transport policies, which was driven by both international organizations, such as the EU, and bottom-up initiatives. In these policies, coordination among actors and collaborative practices were brought up as part of the solution. From the 2000s, there has been a stronger emphasis on governance aspects within the sector, trying to capture the increasingly fragmented nature of transport policy.

Hence, from the 1980s the traditional “command-and-control” function of the state (=government) has increasingly given way to a more facilitating function (=governance). This means that the state assumes the role of facilitating agent by letting nongovernmental actors into the process while taking a more removed, coordinating, and steering role (Joss, 2015, p. 63). In short, governance scholars are concerned with the “relational organization” that the government has taken in transport policy, where multiple interdependent actors take part in and influence policy making; they are also concerned with how these relations are designed, and how constituting governance mechanisms, such as trust, contracts, and power structures, among others, are used to influence decision-making processes.

Governance has increasingly gained a position in both academic and wider policy discourses. However, it is worth mentioning that the emphasis on governance among transport scholars is still rather new. Saetren (2005) extensive literature review on policy studies in political science and public administration journals shows that there are limited studies of the transport sector. Transport studies did not even account for 1% of the reviewed studies (Saetren, 2005, p. 571). An analysis by Marsden and Reardon (2017) of 100 papers from the two leading transport policy journals shows that questions of governance are largely ignored in the transport literature (p. 238); only four papers considered the “ends or aims” of policy, that is, the goals, objectives, or settings (p. 244). See a more in-depth discussion in Hansson (2019).

Understanding Transport Governance—Key Features

Central to the governance perspective is the emphasis on the decision-making processes, analyzing how policies are shaped, implemented, and reshaped. Policy making consists of many decisions that, when combined, constitute a policy process. In transport policy, the decisions vary; they may include the choice of technology, the characteristics of the services to be provided, or the sources from which infrastructure and services are to be financed, among others (Gwilliam, 2008). Decisions are taken throughout the policy process (see Figure 1 in Hansson, 2019). This includes decisions that set the policy on the agenda or formulate a policy problem. Decisions also relate to policy adaptation, such as shaping policies from the national level to a local setting. The policy process also consists of decisions about how to implement policies and evaluate their effects, as well as who should be involved in this process. The policy process can be transferred through decisions on multiple governance levels; for example, a policy shaped at the EU level might also be implemented on a local government level within a municipality. In shaping these decisions, various activities take place and are driven by different actor constellations. These can be politicians and officials at various levels of government, as well as other actors, such as interest groups, consultants, private firms, and transport users. Hence, transport policy

Table 1 Two main governance modes, constructed from transport policy literature.

	<i>Policy-situated governance</i>	<i>Tender-based governance</i>
Goal	Innovation, sustainable mobility, and urban development	Service provision
Roles and responsibilities	More loosely defined, includes formal and informal structures	Well-defined, formalized structures (contract hierarchy)
Collaborative relationship	Mutual understanding	Command – execute
Coordinating mechanism	Trust	Contract
Coordinator	Coalition	Public or semi-public actors

Source: Author's own design.

consists of the interrelationships of actors who combine their resources in policy making. The governance perspective tries to identify the role that institutions, structures, and techniques have in managing these processes, as well as the relationships between the multiple actors involved (Joss, 2015, p. 63).

The policy literature differentiates between horizontal, vertical, and multi-level governance. The combination of vertical and horizontal governance is what often is referred to as multi-level governance (Peters, 2001). Vertical governance mark relations between higher and lower levels of government. Horizontal governance represents the relationships between public organizations and nonpublic organisations that are involved in a decision-making process. It is assumed that the public sector is more than a multilevel system in which the lower levels simply implement national policies; instead, lower-level governments, for example, counties and municipalities, hold their own interests as well as their own authority and legitimacy (Hansson & Torsteinsen, 2019). International organizations might also influence national or local policies. For example, countries of the European Union need to negotiate coherently, since these countries must first develop common negotiating positions with their country and then take those positions into discussions with other member nations (Hansson & Nerhagen, 2019; Peters, 2001). Hence, there are intertwined relations and responsibilities among the levels of government and public and private actors; these combined relations constitute multi-level governance networks, which all are part of effecting decisions in one or several phases of the policy process.

The transport policy literature covering aspects of governance is found in multiple transport settings. On one hand, there are studies analyzing policy making on a local level, for example in relation to sustainable transport systems or mobility plans (Tønnesen et al., 2019). There are scholars analyzing the effects of EU policies and how these are shaped and re-shaped in relation to national policies (Hansson, 2019). There are also studies on governance and public transport planning and service provision (Hansson, 2011). There is also a discussion on partnerships that can be related to governance (Sørensen & Longva, 2011). These different thematic threads bring a heterogenic and sometimes conflicting perspective on what governance might be. This paper proposes two main types of governance modes, which have been drawn from the various literature on governance in transport policy: (1) policy-situated governance and (2) tender-based governance (see Table 1). With a base in these two modes, the central features of governance in transport policy will be discussed.

Policy-situated governance is found in several settings, it can be where actor constellations are active in either influencing the content of the policies (in the agenda-setting phase or the policy-formulation phase) or in urban regions where the policies are to be implemented. Policy-situated governance consists of a variety of actors, such as political parties, public officials, enterprises, consultants, and sometimes end users such as citizens. The name “policy-situated governance” relates to a specific policy issue that is to be dealt with, for example, urban development, sustainable mobility, mobility planning, etc.

Tender-based governance is found in the end phases of the policy cycle when policies are transferred into practical implementation. Tender-based governance is sprung from the privatization policies that have a view of governing services from a distance and is highly evident in countries that use competitive tendering in the service provision of public transport.

A central feature of governance is understanding how actors are organized in terms of the responsibilities and resources they have at hand. These roles and responsibilities can be more or less formalized (Paulsson & Isaksson, 2019). Policy-situated governance features more loosely defined roles and responsibilities within the network; for example, a smart mobility process can hold experimental processes in which a broad range of actors are included (Bakker et al., 2014; Docherty et al., 2018). The decision-making processes within these networks hold both formal and informal structures (Hrelja et al., 2017). Tender-based governance holds a more formal and clear division of roles between the actors, based on a contract hierarchy. The contract defines the roles and responsibilities of the private actor(s), while the roles of the public actors are defined by the hierarchy that exists within the public organization's authorities (Hansson & Longva, 2014). The “strategy-tactics-operation framework” (STO), introduced to the transport sector by van de Velde (1999), has also been an influential perspective in this setting. It is a variation of control hierarchy and provides a model for comparing the organizational structures and the development of contractual agreements in different countries. It has allowed a range of issues to be framed and has provided an understanding of various stakeholders' functions, taking either a strategic, tactical, or operational role (van de Velde, 1999; Wong & Hensher, 2018).

Within the transport governance literature, there is a normative view that collaboration between organizations is a critical factor in reaching successful policy implementation. Behind this view is the belief that collaboration might generate new, integrated knowledge and capacities that are not available within individual organizations (Joss, 2015). This rationale is especially evident in policy studies on smart mobility or innovations in urban transport systems, in which shared understanding through the exchange of

knowledge, information, and expertise is seen as central to successful policy implementation (Meyer, et al., 2005). However, studies have also brought forward a more critical perspective on knowledge sharing. Governance constellations hold path dependency traces and new, innovative aspects and knowledge that are difficult to incorporate into existing constellations; this results in a reproduction of old views instead of the shaping of new ones (Hansson, 2019; Hansson & Nerhagen, 2019). In tender-based governance settings, there is also the belief that combining actor resources makes successful policy implementation more likely; however, this is based on a division of competence and labor, in which one part is contracted to execute a service. Within this structure, knowledge exchange/sharing can take place; however, this is defined within the premises of the contract.

In governance structures, actors bring different goals and agendas to the process, making it important to not only clarify roles and responsibilities but also power dimensions and asymmetric relations. There is a power dependency within governance structures, meaning that different actors employ power resources, such as political, economic, and informative resources (Docherty et al., 2018; Hansson, 2011). These resources combined are important for the overall network and are used to negotiate power throughout the process (Hansson, 2011). In tender-based governance, power relations might be easier to control because a contract can include some of these aspects. In policy-situated governance settings, power dimensions are complex and difficult to identify and deal with. Trust is brought forward as a central mechanism in overcoming power dimensions in policy-situated governance settings. Longva and Osland (2007) make the distinction between thin and thick trust. Tender-based relationships can be referred to as “thin trust,” while “thick trust” can be found in complex, intertwined networks that go beyond a competitive tendering regime (Longva & Osland, 2007). However, Stanley and Hensher (2008) stress the importance of also achieving a “trusting partnership” in contract-based settings (negotiated contracts). Various papers have discussed how to create and maintain a trusting partnership. In this setting, it is central that the authority and the provider devote considerable effort to developing the foundations for a trusting relationship during the precontractual phase (Stanley & Hensher, 2008).

There is often an actor constellation that takes on the role of mediating between the various interests and priorities in a governance process; this actor constellation is sometimes called “meta-governor” (Hansson, 2013). The mediating function is helpful in bringing forward actors and gaining a mutual understanding of the issue at hand (Tønnesen et al., 2019). By assuming this coordinating role, the meta-governor may also wield considerable influence over other actors (Hansson, 2013). In tender-based governance settings, the coordinating role is mainly given to public actors, for example, the transport authority (Hansson, 2013); in policy-situated governance, this coordinating role can consist of a more heterogenous constellation (Hrelja et al., 2017). A study by Wikström, et al. (2016) on the electrification of commercial fleets in urban areas also addresses the relevance of a single actor, “policy entrepreneur,” who takes on a coordinating role and pushes policies forward. The concept of “system builders” has also been used for explaining a similar role (Palm & Fallde, 2016).

Future Challenges

Over time, there has been a search for innovative mechanisms for making government work better and serve society better (Peters, 2001, p. 2). This chapter has addressed the governance turn in transport policy and described central features that explain the composition of governance structures. The governance turn in transport policy has been driven by both privatization reforms as well as reforms concerned with how to deal with climate change and promote a more sustainable transport system. These reforms have created a need for enhanced coordination among transport actors and new governance structures. Governance structures affect the transport policy process; they can be found in the early phases of the policy process, such as in the agenda-setting phase, or in the end parts of the policy process, such as when implementing policies.

It should not be given that a governance approach to transport policy delivers more effective or legitimate policy solutions. Practical applications of governance processes in transport policy have pointed out several challenges, underlining what can be a considerable divergence between the normative ideals of governance and actual practice.

Governance, from a policy theory perspective, is still limited within transport studies, even though we have seen an increasing number of scholars taking interest in this field during the last 10 years. As a final point, two aspects that need more research will be put forward.

More collaborative forms of governance require special efforts to ensure transparency and accountability. Governance processes are less transparent when private firms are involved in the early phases of decision making. Compared to the more traditional view of politics, in which politicians are ultimately responsible for decisions while other actors might be involved in the planning or implementation of the decisions, the roles in governance are more diffused (Hansson & Longva, 2014). The aspect of transparent decision making and accountability (“Who was involved in the decision making, and who can be held accountable for actions and the decisions taken?”) is only vaguely addressed in the transport literature, compared to other government sectors. This paper has addressed the notion of contracts and trust, but there is still much work to be done in relation to transparent decision-making and accountability.

Studies have also shown that governance mechanisms designed to reconcile policy goals may in fact intensify tensions between them (Joss, 2015). There is an intense discussion on policy deviance and how to achieve policy integration (Hull, 2008). In this context, it has been argued that governance networks might contribute to policy discrepancy. There are decoupling tendencies, created by governance structures, which mean that some policy elements are separated from the original policy process in order to deal with unwanted conflicts in the network; this results in fragmented policy proposals and difficulties in getting an overview of the proposed effects (Hansson, 2019). Studies have also addressed the problem of parallel policy making (Fallde, 2011). Parallel policy

making reflects “a fundamental lack of integration of sustainable mobility in policies and plans of strategic importance, which hinders effective policy integration” (Isaksson et al., 2017).

The paper has brought forward two governance mode structures constructed from findings in the literature on transport governance and policy. In urban and regional settings, policy-situated governance and tender-based governance might coexist and respond to different policy problems. Combining governance modes with multiple policy-making processes might be central to understanding why policies are not successfully implemented and how conflicting policy goals can be addressed. Studies are welcomed that further develop these governance modes (and perhaps identify more), both in terms of how they work separately as well as how they might conflict with or enhance parallel policy making.

References

- Bakker, S., Maat, K., van Wee, B., 2014. Stakeholders interests, expectations, and strategies regarding the development and implementation of electric vehicles: The case of the Netherlands. *Trans. Res. Part A: Policy Prac.* 66, 52–64, doi:10.1016/j.tra.2014.04.018.
- Beesley, M.E., Glaister, S., 1985. Deregulating the bus industry in Britain—(C) a response. *Trans. Rev.* 5 (2), 133–142, doi:10.1080/01441648508716590.
- Docherty, I., Marsden, G., Anable, J., 2018. The governance of smart mobility. *Trans. Res. Part A: Policy Prac.* 115, 114–125, doi:10.1016/j.tra.2017.09.012.
- Fallde, M., 2011. Miljö i tanken?: Policyprocesser vid övergången till alternativa drivmedel i kollektivtrafiken i Linköping och Helsingborg 1976–2005. (No. 534) [Doctoral thesis, Linköping University Electronic Press]. DiVA Database. Available from: <http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-70095>.
- Gwilliam, K.M., 2008. A review of issues in transit economics. *Res. Trans. Econ.* 23 (1), 4–22, doi:10.1016/j.retrec.2008.10.002.
- Gwilliam, K.M., Nash, C.A., Mackie, P.J., 1985. Deregulating the bus industry in Britain—(B) the case against. *Trans. Rev.* 5 (2), 105–112, doi:10.1080/01441648508716589.
- Hansson, L., 2011. The tactics behind public transport procurements: An integrated actor approach. *Eur. Trans. Res. Rev.* 3 (4), 197–209, doi:10.1007/s12544-011-0057-2.
- Hansson, L., 2013. Hybrid steering cultures in the governance of public transport: A successful way to meet demands? *Res. Trans. Econ.* 39 (1), 175–184, doi:10.1016/j.retrec.2012.06.011.
- Hansson, L., 2019. Public administrators' roles in the policy adaptation of transport directives: How knowledge is created and reproduced. *Trans. Policy*, doi:10.1016/j.tranpol.2019.10.008.
- Hansson, L., Longva, F., 2014. Contracting accountability in network governance structures. *Qual. Res. Account. Manag.* 11 (2), 92–110, doi:10.1108/GRAM-04-2014-0032.
- Hansson, L., Nerhagen, L., 2019. Regulatory measurements in policy coordinated practices: The case of promoting renewable energy and cleaner transport in Sweden. *Sustainability* 11 (6), 1687, doi:10.3390/su11061687.
- Hansson, L., Torsteinsen, H., 2019. Providing 'hard' local government services in a multi-level, multi-actor system. *Scand. J. Public Admin.* 23 (2), 3–11.
- Hensher, D.A., Wallis, I.P., 2005. Competitive tendering as a contracting mechanism for subsidising transport: The bus experience. *J. Trans. Econ. Policy* 39 (3), 295–322, <https://www.jstor.org/stable/20053970>.
- Hrelja, R., Monios, J., Rye, T., Isaksson, K., Scholten, C., 2017. The interplay of formal and informal institutions between local and regional authorities when creating well-functioning public transport systems. *Int. J. Sustain. Trans.* 11 (8), 611–622, doi:10.1080/15568318.2017.1292374.
- Hull, A., 2008. Policy integration: What will it take to achieve more sustainable transport solutions in cities? *Trans. Policy* 15 (2), 94–103, doi:10.1016/j.tranpol.2007.10.004.
- Isaksson, K., Antonson, H., Eriksson, L., 2017. Layering and parallel policy making – Complementary concepts for understanding implementation challenges related to sustainable mobility. *Trans. Policy* 53, 50–57, doi:10.1016/j.tranpol.2016.08.014.
- Joss, S., 2015. Sustainable cities: Governing for urban innovation. Palgrave Macmillan, New York, NY.
- Longva, F., & Osland, O., 2007. “Organizing Trust”. On the institutional underpinning and erosion of trust in different organizational forms in public transport. Paper presented at the 9th International Conference on Competition and Ownership in Land Passenger Transport (Thredbo 9), Lisbon Technical University.
- Marsden, G., Reardon, L., 2017. Questions of governance: Rethinking the study of transportation policy. *Transp. Res. Part A: Policy Prac.* 101, 238–251, doi:10.1016/j.tra.2017.05.008.
- Meyer, M.D., Campbell, S., Leach, D., Coogan, M., 2005. Collaboration: The key to success in transportation. *Trans. Res. Rec.* 1924 (1), 153–162, doi:10.1177/0361198105192400120.
- Palm, J., Fallde, M., 2016. What characterizes a system builder? The role of local energy companies in energy system transformation. *Sustainability* 8 (3), 256, doi:10.3390/su8030256.
- Paulsson, A., Isaksson, K., 2019. Networked authority and regionalised governance: Public transport, a hierarchy of documents and the anti-hierarchy of authorship. *Environ. Plan. C: Polit. Space* 37 (6), 985–1004, doi:10.1177/2399654418818553.
- Peters, B.G., 2001. *The Future of Governing*. University Press of Kansas, Lawrence, KS.
- Sætre, H., 2005. Facts and myths about research on public policy implementation: Out-of-fashion, allegedly dead, but still very much alive and relevant. *Policy Stud. J.* 33 (4), 559–582, doi:10.1111/j.1541-0072.2005.00133.x.
- Stanley, J., Hensher, D.A., 2008. Delivering trusting partnerships for route bus services: A Melbourne case study. *Trans. Res. Part A: Policy Prac.* 42 (10), 1295–1301, doi:10.1016/j.tra.2008.05.006.
- Sørensen, C.H., Longva, F., 2011. Increased coordination in public transport—Which mechanisms are available? *Trans. Policy* 18 (1), 117–125, doi:10.1016/j.tranpol.2010.07.001.
- Tønnesen, A., Krogstad, J.R., Christiansen, P., Isaksson, K., 2019. National goals and tools to fulfil them: A study of opportunities and pitfalls in Norwegian metagovernance of urban mobility. *Trans. Policy* 81, 35–44, doi:10.1016/j.tranpol.2019.05.018.
- van de Velde, D.M., 1999. Organisational forms and entrepreneurship in public transport: Classifying organisational forms. *Trans. Policy* 6 (3), 147–157, doi:10.1016/S0967-070X(99)00016-5.
- Wikström, M., Eriksson, L., Hansson, L., 2016. Introducing plug-in electric vehicles in public authorities. *Res. Trans. Bus. Manag.* 18, 29–37, doi:10.1016/j.rtbm.2016.01.009.
- Wong, Y.Z., Hensher, D.A., 2018. The Thredbo story: A journey of competition and ownership in land passenger transport. *Res. Trans. Econ.* 69, 9–22, doi:10.1016/j.retrec.2018.04.003.

Transferability of Urban Policy Measures

Paul Martin Timms, Institute for Transport Studies University of Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	77
Good Practice Guidelines	77
Global Level	77
EU Level	78
A Snapshot of the State-of-the-Art in Research on Policy Transfer in 2011	78
Research on Urban Transport Policy Transfer Since 2011	80
Suggestions for Further Research	81
References	81

Introduction

This chapter is concerned with the processes by which transport policy measures are transferred to specific urban locations from elsewhere. Following this introduction, Section 2 provides overviews (firstly on a global level and then on an EU level) of guidelines describing generic *good practice* concerning policy measures, which provide an important mechanism for achieving the first step in the transfer process, which is to start the circulation of information. Such guidelines, which are typically easily available for free on the Internet, have an intended audience of decision-makers and informed citizens/stakeholders. They are generally visually attractive, avoid using overly technical/academic language, and usually provide particular case studies illustrating the good practice principles being advocated. Inevitably, though, such *broadcasting* of good practice can pay little attention to the specific needs and contexts of cities that might implement the measures and so cannot reflect upon the full transfer process, except in a relatively simplistic fashion. Furthermore, in order to maximize their readership, they inevitably describe policy measures in a simplified idealized way, leaving out (or at least playing down) some of the “messier” complex details about their implementation, thus leading to an *information gap*. The job of filling this gap has been taken on by academic researchers, whose concepts and research results provide the subject matter for the remainder of the chapter.

A key watershed in the development of academic research on policy transfer in transport studies was provided by the publication of a special issue in the journal *Transport Policy* on “Transferability of urban transport policy” (Ison et al., 2011), which provided a comprehensive snapshot of different approaches to researching the subject at the time. These papers are discussed in Section 3, which also contains information about a parallel stream of research that saw the transfer process in terms of a *policy mobility* conceptual framework. Although not specifically concerned with transport policy, it was to play an important role, conceptually and methodologically, in influencing subsequent studies on transport policy transfer.

Section 4 provides an overview of academic research on urban transport policy transfer since 2011. Such research has taken profound directions since 2011: although the questions might be the same (or at least similar) the analysis goes deeper. Section 5 concludes the chapter with two suggestions for further research.

Good Practice Guidelines

Global Level

There are currently a number of international NGOs providing good practice guidelines on urban transport, including: the Institute for Transportation and Development Policy (ITDP)^a; the World Resources Institute – EMBARQ^b; and GIZ’s Sustainable Urban Transport Project (SUTP)^c. The roles of these NGOs are described by Sosa López and Montero (2018) as:

“politically complex: sometimes they collaborate with government officials, and in other occasions, they become their most forceful opponent. They alternately identify themselves as civil society groups that represent citizens, and other times they rely on their connections with global philanthropy and international development banks to present themselves as transport experts.” (p138)

While providing information on a full range of urban transport measures, these NGOs are particularly associated with the promotion of Bus Rapid Transit (BRT) (Sengers and Raven, 2015; Paget-Seekins, 2015). In particular, ITDP produces an extensive

^a <https://www.itdp.org/>

^b <https://www.wri.org/tags/embarq>

^c <https://www.sutp.org/en/>

online BRT Planning Guide^d and the BRT Standard (ITDP, 2016). The latter provides a detailed point-based scoring system which enables BRT planners to “certify” their BRT corridor as “basic BRT, bronze, silver, or gold,” thus placing the BRT corridor “within the hierarchy of international best practices.” While such an approach is no doubt useful in providing much technical information that needs to be taken into account when planning a BRT, the hierarchical scoring system risks being over-prescriptive, particularly as it is unable to take into account local contextual factors. From a transferability point of view, it is thus problematic.

EU Level

The European Union (EU) is of particular interest with respect to transport policy transferability, due to the large number of cross-national applied transport research projects that have been funded by the European Commission (EC) over the past 30 years (Rommerts, 2012, TRIP, 2014 and 2016). Inevitably, questions are raised in such projects about cross-national transferability. A recent product of EU applied research of particular relevance to this chapter is the second edition of the “Guidelines for Developing and Implementing a Sustainable Urban Mobility Plan [SUMP]” (Rupprecht Consult, 2019), which includes a step entitled “Measure Selection”. Although the guidelines are prescriptive in terms of process, they are less prescriptive, in comparison with the BRT Standard, with respect to the actual measures that might be adopted in any particular location. These guidelines, along with many other associated guidelines about particular aspects of the SUMP process (such as how they might be applied in different sizes of city) are available from the European Local Transport Information Service (ELTIS) website^e, which also has a large store of one-page descriptions of case studies (mainly from the EU but also from other parts of the world).

May (2015) describes the background for the development for the first edition of the SUMP Guidelines, listing some of the EU research projects that directly or indirectly made contributions. These include an early contribution by the PROSPECTS project, which produced the Decision Makers’ Guidebook (DMG) (May, 2005). The DMG was intentionally designed to be reader-friendly, using humorous cartoons to illustrate various aspects of the policy-making process, including the transfer of policy measures (shown in Box 1). As can be seen, the DMG emphasised that decisions about appropriate policy measures are dependent upon local context: they cannot simply be “imposed from above.”

The recommendation of the DMG to record experience (see Box 1) has been facilitated by the existence of the large number of EU research projects mentioned earlier. However, the resulting quantity of available information directly leads to practical questions as to how any decision-maker might make use of it in the policy transfer process. Macário and Marques (2008) developed a “10 Step” scheme, as illustrated in Figure 1, based upon their experience in EU research projects. A simplified “light” version of this scheme, for use in urban locations with less resources for carrying out a full transferability analysis, is described by Timms (2014). Whether “heavy” or “light,” all such ex-ante transferability analyses require an assessment of potential measures (Step 8 in Figure 1), leading to key context-specific questions such as:

- what is the role of public/stakeholder participation?;
- are the measures contested by different groups?;
- what objectives are the measures intended to achieve? (and from whose perspective(s)?);
- who is supplying the information from other locations? (and what are their motivations?);
- is the information about other locations “reliable”? (and from whose perspective(s)?);
- is the experience of some locations more influential than that from other places (and why?);
- which technical tools (if any) might be used in analyzing transferability?

All these questions lead to the need to understand policy transfer as more than just “broadcasting” good practice guidelines, and provide (some of the) questions to be tackled by academic researchers, as discussed in the next two sections.

A Snapshot of the State-of-the-Art in Research on Policy Transfer in 2011

As mentioned in the introduction to this chapter, a key moment in the development of academic research on policy transfer in transport studies was provided by the publication of a special issue in the journal *Transport Policy* on “Transferability of urban transport policy” (Ison et al., 2011). Of the six papers in the special issue, Marsden and Stead (2011) provided a review of the state-of-the-art on concepts and evidence, while the other five papers described case studies covering different aspects of the transfer process (Marsden et al., 2011; Timms, 2011; Bray et al., 2011; Rye et al., 2011; Attard and Enoch, 2011). In order to provide a consistent approach, the authors of all six papers were asked to use a common framework made up of questions derived from Dolowitz and Marsh (2000): Why transfer?; Who is involved?; What is transferred?; From where?; What is the degree of transfer?; What are the constraints?; How is transfer demonstrated?; and Does it succeed? Two issues of particular relevance to the current chapter can be highlighted here. First, Timms (2011), in the context of EU transport policy transfer, emphasized the potential differences between “top down” perspectives on transfer (as might be exemplified in the type of guidelines reviewed in Section 2) and the “bottom-up” perspectives of specific city transport planners (identified through semi-structured interviews). One conclusion to this research was that city planners receive information from a wide variety of sources, including a variety of professional and personal networks, and

^d <https://brtguide.itdp.org/branch/master/guide/>

^e <https://www.eltis.org/>

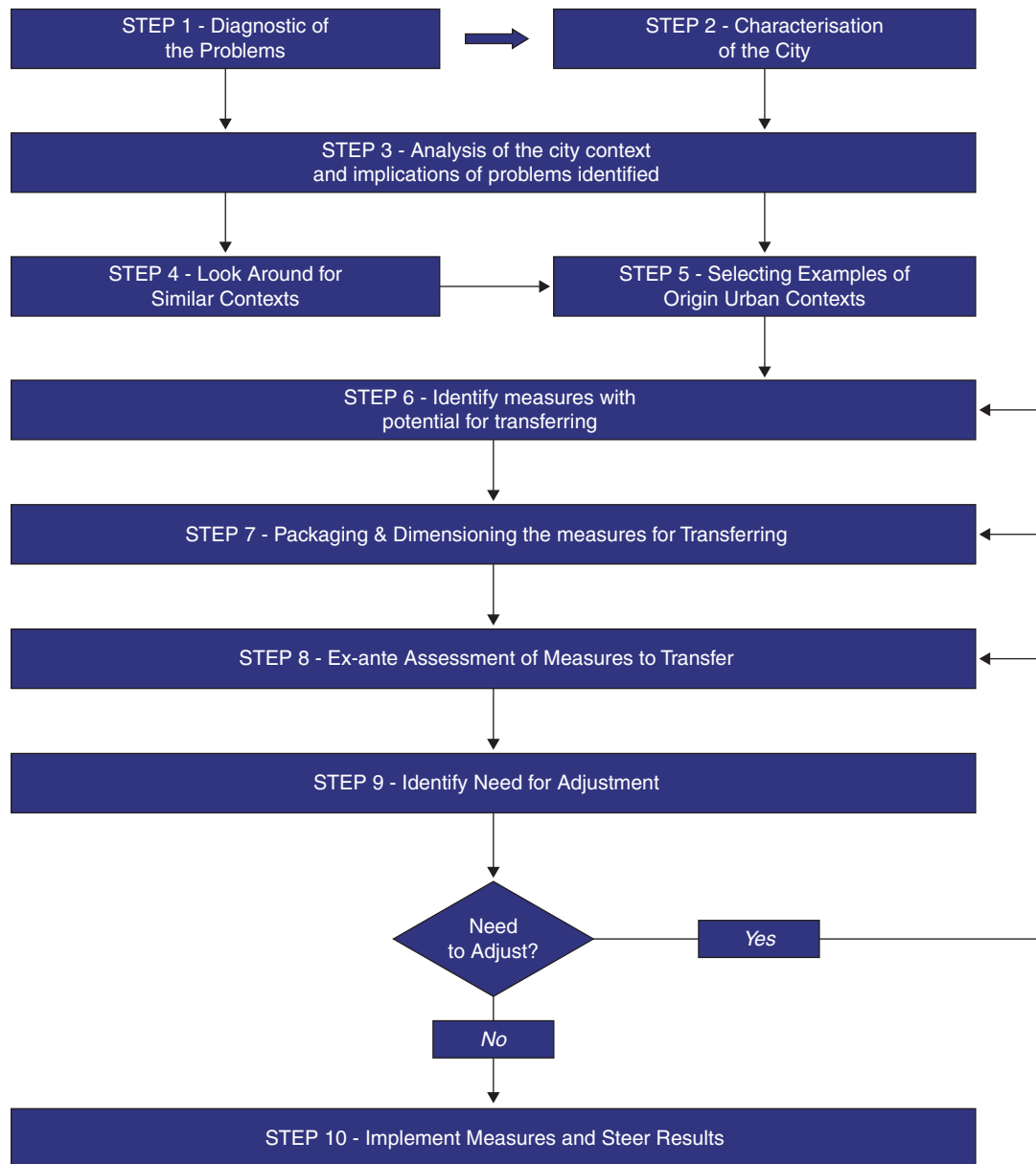


Figure 1 10-step transferability process. Source: Macário and Marques (2008).

such information interacted with that provided by “top down” guidelines in complex ways. Second, the paper by Attard and Enoch (2011) presented a “rich” description of the ex-post transfer of a specific policy (road pricing) to one particular city (Valletta, Malta). Although this was the only paper in the special issue that took this approach, it was an approach that would be adopted more frequently in later research (as will be seen in Section 4).

At the same time as the publication of the special issue in *Transport Policy*, research was being undertaken independently to enhance and develop various high level concepts about the policy transfer process, including: *urban policy mobilities and global circuits of knowledge* (McCann, 2011); *transfer-diffusion to mobility-mutation* (Peck, 2011); *institutional isomorphism* (Peck, 2011); *policy tourism* (González, 2011); *worlding practices* (Roy, 2011); and *a range of interests in the policy transfer “business,” such as consultants, advocates, evaluators, gurus, and critics* (Peck and Theodore, 2010). These concepts are all described within a context of “fast” policy transfer consistent with globalized neoliberalism, and make an important contribution to understanding the types of complexity that were identified in the questions posed at the end of Section 2. Many of the transport case studies to be reported in Section 4 make use of these, or similar, concepts, citing the authors mentioned here.

Table 1 Case studies in urban transport policy transfer since 2011.

<i>What is transferred?</i>	<i>From where?</i>	<i>To where?</i>	<i>Authors</i>
Bus Rapid Transit (BRT)	Bogotá	South African cities	Wood (2014a, 2014b, 2014c)
Weekly street closure (Ciclovía) and BRT	Bogotá	Guadalajara	Montero (2017a and 2017b)
Weekly street closure (Ciclovía) and BRT	Bogotá	Worldwide	Montero (2018)
Monorail, Light Rail Transit (LRT), Tram, BRT	Kuala Lumpur, Mumbai, Las Vegas, Moscow, Seattle, Sydney, Jakarta, Putrajaya, Malacca, Rio de Janeiro, Kaohsiung, Suzhou and Casablanca	Penang	Connolly (2019)
Transit-oriented Development (TOD)	Worldwide	Cities in the Netherlands	Thomas and Bertolini (2015)
TOD and BRT	Worldwide	Jinan and Chengdu	Thomas and Deakin (2017)
Public bicycle programs	Paris, Lyon, Montreal and Washington DC	European and North American cities	Parkes et al. (2013)
Public bicycle programs	Huangzhou	Mainly cities in China, but also in the Philippines, France and Hong Kong.	Ma (2017)
Cycling measures	Berlin	Manchester	Sheldrick et al. (2017)
Cycling measures	Amsterdam and Copenhagen	Portland	Ibsen and Olesen (2018)
Bike-sharing, subway, BRT, congestion charging	Chengdu, Shenzhen, Tokyo, Seoul, Singapore	Beijing, Shanghai	Chun et al. (2019)
Sustainable land-use and transport planning concepts	Cities in the Netherlands	Mainly cities in Europe, but also in India, Japan and the USA	Pojani and Stead (2014)
Public transport systems and integration, road pricing, demand management, land-use policies (e.g. TOD), urban realm improvements, cleaner vehicles	Worldwide	Lyon, Nancy, Edinburgh, Leeds, Bremen, Stockholm, Copenhagen, Seattle, Dallas, San Francisco, Vancouver	Marsden et al. (2012)
Public space policy	Worldwide	Hanoi	Söderström and Geertman (2013)
Removal of car-based transport infrastructure	San Francisco, New York, Seoul, Toronto and Seattle	Vancouver	Farmer and Perl (2018)
Urban freight transport measures	Belo Horizonte, São Paulo, Vancouver, New York, Barcelona, Utrecht	Cariacica	Timms (2014)

Research on Urban Transport Policy Transfer Since 2011

Since 2011, there has been a large increase in multisectoral academic research on urban policy transfer, frequently using the concepts used by the authors mentioned at the end of Section 3. Some of this research includes transport as one of a number of sectors being considered, awarding transport a more or less important role. Furthermore, there is a large amount of recent research on urban transport policy that makes implicit reference to policy transfer. Reviewing either (or both) of these topics comprehensively (i.e., multisectoral policy transfer and urban transport policy) would go beyond the scope of the present chapter. Therefore, the current section is limited to research since 2011 that concentrates exclusively (or at least mainly) upon the urban transport sector, makes the transfer of transport policy measures a central aspect of the research, and follows on (methodologically) from the research reported in Section 3. Two significant high level conceptual contributions to such research are: Vigar (2017), who explains how transferred knowledge can be incorporated in *communicative transport planning*; and Monios (2017), who distinguishes transport policy transfer from *policy churn*, defined as “changing a policy without establishing a clear link between the reasons for failure of the existing policy and how these will be overcome by the new policy.” In general, though, most of the research on urban transport policy transfer since 2011 has been case-study based: Table 1 gives information about the transport measures and origins/destinations of transfer involved.

A striking aspect of Table 1 is that the range of policies has been relatively limited, with emphasis upon bus rapid transit (BRT), transit-oriented development (TOD), and cycling measures. In general, the research provides many rich insights into the motivations of the particular actors involved in policy transfer. For example, Montero (2017) quotes a study tour participant from

Guadalajara to Bogotá who had heard Enrique Peñalosa (ex-mayor of Bogotá and worldwide “policy entrepreneur” for BRT) give a talk in Guadalajara:

“We thought it would be a good idea to go to Bogotá to see if it was true all that [Peñalosa] told us about the city: how in a city which had 50% of Guadalajara’s per capita income, 40 years of guerrilla, and with rampant drug trafficking ... how could he do that? We didn’t really believe him, we thought he was probably lying ... We needed to see it ourselves to believe it” (p340).

The overall impression from these case studies is that, as might be expected, *good (or best) practice* is far more nuanced than is implied in the guidelines described in Section 2. Furthermore, the case studies provide insights into “policy transfer failures” in two distinct senses: first what might be positively learnt in a particular city about a policy that “failed” somewhere else, for example, a failed plan for an expressway removal in Toronto (Farmer and Perl, 2018), and failed monorail and LRT schemes in Kuala Lumpur, Mumbai, Las Vegas, Moscow, Seattle, Sydney, Jakarta, Putrajaya and Malacca (Connolly, 2019); and second when the transfer process itself was deemed to be a failure, such as the transfer of TOD and BRT to Chengdu (Thomas and Deakin, 2017).

Suggestions for Further Research

This chapter has identified an evolution in thinking about the transfer of urban transport policy measures, taking into account both applied research (as used in the development of good practice guidelines) and academic research (which seeks to understand the complexities involved in the policy transfer process). Two suggestions are given here for further research:

- Rekhviashvili and Sgibnev (2018a) have argued that, in general, more attention should be paid in transport studies to urban transport workers, especially those with precarious contracts of employment, and in Rekhviashvili and Sgibnev (2018b) they discuss the position(s) of Uber drivers and marshrutka^f drivers. Furthermore, as can be seen in Table 1, there is in general a lack of research on the policy transfer of measures concerning informal public transport. It is thus suggested that specific attention should be made on the transfer of regulatory and other measures concerning informal public transport, including measures directly affecting informal transport workers, covering modes such as motorbike taxis (Ehebrecht et al., 2018, Diaz Olvera et al., 2020), auto-rickshaws (Harding et al., 2016) and other forms of paratransit (Cervero and Golub, 2007, Mateo-Babiano, 2016, Phun and Yai, 2016, Jennings and Behrens, 2017).
- Much emphasis has been put on the need for public participation in urban transport planning, both in SUMP guidelines (Rupprecht Consult, 2019, Lindenau and Böhler-Baedeker, 2016) and by transport studies researchers (McAndrews and Marcus, 2015, Sagaris, 2018, Kębłowski et al., 2019). More research needs to be carried out as to how the views of “ordinary citizens” can be incorporated in the public participation aspects of policy transfer processes. Research could pay particular attention to situations where such citizens have expert knowledge on other locations due to having spent long or short periods of time in other countries, or through having international social media connections (McArthur, 2019).

References

- Attard, M., Enoch, M., 2011. Policy transfer and the introduction of road pricing in Valletta, Malta. *Transp. Policy* 18, 544–553.
- Bray, D.J., Taylor, M.A.P., Scrafton, D., 2011. Transport policy in Australia—Evolution, learning and policy transfer. *Transp. Policy* 18, 522–532.
- Cervero, R., Golub, A., 2007. Informal transport: A global perspective. *Transp. Policy* 14, 445–457.
- Chun, J., Moody, J., Zhao, J., 2019. Transportation policymaking in Beijing and Shanghai: Contributors, obstacles, and process. *Case Stud. Transp. Policy* 7, 718–731, 2019.
- Connolly, C., 2019. Worlding cities through transportation infrastructure. *Environ. Plan. A* 51 (3), 617–635.
- Dolowitz, D., Marsh, D., 2000. Learning from abroad: the role of policy transfer in contemporary policy making. *Governance* 13 (1), 5–24.
- Ehebrecht, D., Heinrichs, D., Lenz, B., 2018. Motorcycle-taxis in Sub-Saharan Africa: Current knowledge, implications for the debate on “informal” transport and research needs. *J. Transp. Geog.* 69, 242–256.
- Farmer, D., Perl, A., 2018. The role of policy learning in urban mobility. *Urban Res. Pract.* 13 (1), 77–96.
- González, S., 2011. Bilbao and Barcelona ‘in motion’: how urban regeneration ‘models’ travel and mutate in the global flows of policy tourism. *Urban Stud.* 48, 1397–1418.
- Harding, S.E., Badami, M.G., Reynolds, C.C.O., Kandlikar, M., 2016. Auto-rickshaws in Indian cities: Public perceptions and operational realities. *Transp. Policy* 52, 143–152.
- Ibsen, M.E., Olesen, K., 2018. Bicycle urbanism as a competitive advantage in the neoliberal age: The case of bicycle promotion in Portland. *Int. Plan. Stud.* 23 (2), 210–224.
- Ison, S., Marsden, G., May, A.D., 2011. Transferability of urban transport policy. *Transp. Policy* 18, 489–491.
- ITDP, 2016. The BRT Standard. <https://www.itdp.org/2016/06/21/the-brt-standard/>.
- Jennings, G., Behrens, R., 2017. The case for investing in paratransit strategies for regulation and reform. Volvo Research and Educational Foundations (VREF). <http://www.vref.se/>.
- Kębłowski, W., Van Criekingen, M., Bassens, D., 2019. Moving past the sustainable perspectives on transport: An attempt to mobilise critical urban transport studies with the right to the city. *Transp. Policy* 81, 24–34.
- Kumar, M., Singh, S., Ghate, A.T., Pal, S., Wilson, S.A., 2016. Informal public transport modes in India: A case study of five city regions. *IATSS Res.* 39, 102–109.
- Lindenau, M., Böhler-Baedeker, S., 2016. CHALLENGE Participation Manual: Actively engaging citizens and stakeholders in the development of Sustainable Urban Mobility Plans. <http://www.sump-challenges.eu/content/participation>.
- McAndrews, C., Marcus, J., 2015. The politics of collective public participation in transportation decision-making. *Transp. Res. Part A* 78, 537–550.
- McArthur, J., 2019. The production and politics of urban knowledge: Contesting transport in Auckland, New Zealand. *Urban Policy Res.* 37 (1), 45–61.
- McCann, E., 2011. Urban policy mobilities and global circuits of knowledge: Toward a research agenda. *Ann. Asso. Am. Geogr.* 101 (1), 107–130.
- Ma, L., 2017. Site visits, policy learning, and the diffusion of policy innovation: Evidence from public bicycle programs in China. *J. Chin. Pol. Sci.* 22 581–599.
- Macário, R., Marques, C.F., 2008. Transferability of sustainable urban mobility measures. *Res. Transp. Econ.* 22 (1), 146–156.

^f Marshrutkas are public transport minibuses that are widespread in post-Soviet cities

- Marsden, G., Stead, D., 2011. Policy transfer and learning in the field of transport: A review of concepts and evidence. *Transp. Policy* 18, 492–500.
- Marsden, G., Frick, K.T., May, A.D., Deakin, E., 2011. How do cities approach policy innovation and policy learning? A study of 30 policies in Northern Europe and North America. *Transp. Policy* 18, 501–512.
- Marsden, G., Frick, K.T., May, A.D., Deakin, E., 2012. Bounded rationality in policy learning amongst cities: Lessons from the transport sector. *Environ. Plan. A* 44, 905–920.
- Mateo-Babiano, I., 2016. Indigeneity of transport in developing cities. *Int. Plan. Stud.* 21 (2), 132–147.
- May, A.D., 2005. *Developing Sustainable Urban Land use and Transport Strategies: A Decision-makers' Guidebook*. 2nd edition. Available from: <http://www.konsult.leeds.ac.uk/dmg/>.
- May, A.D., 2015. Encouraging good practice in the development of sustainable urban mobility plans. *Case Stud. Transp. Policy* 3, 3–11.
- Monios, J., 2017. Policy transfer or policy churn? Institutional isomorphism and neoliberal convergence in the transport sector. *Environ. Plan. A* 49 (2), 351–371.
- Montero, S., 2017a. Study tours and inter-city policy learning: Mobilizing Bogotá's transportation policies in Guadalajara. *Environ. Plan. A* 49 (2), 332–350.
- Montero, S., 2017b. Persuasive practitioners and the art of simplification: Mobilizing the 'Bogotá Model' through storytelling. *Novos Estudos CEBRAP* 36 (1), 59–75.
- Montero, S., 2018. Leveraging Bogotá: sustainable development, global philanthropy and the rise of urban solutionism. *Urban Stud.* 1–19, <https://doi.org/10.1177/0042098018798555>.
- Díaz Olvera, L., Plat, D., Pochet, P., 2020. Looking for the obvious: Motorcycle taxi services in Sub-Saharan African cities. *J. Transp. Geogr.*, in press.
- Paget-Seekins, L., 2015. Bus rapid transit as a neoliberal contradiction. *J. Transp. Geogr.* 48, 115–120.
- Parkes, S.D., Marsden, G., Shaheen, S.A., Cohen, A.P., 2013. Understanding the diffusion of public bikesharing systems: Evidence from Europe and North America. *J. Transp. Geogr.* 31, 94–103.
- Peck, J., 2011. Geographies of policy: From transfer-diffusion to mobility-mutation. *Prog. Human Geogr.* 35 (6), 773–797.
- Peck, J., Theodore, N., 2010. Mobilizing policy: Models, methods, and mutations. *Geoforum* 41 (2), 169–174.
- Phun, V.K., Yai, T., 2016. State of the art of paratransit literatures in Asian developing countries. *Asian Transp. Stud.* 4 (1), 57–77.
- Pojani, D., Stead, D., 2014. Going Dutch? The export of sustainable land-use and transport planning concepts from the Netherlands. *Urban Stud.* 52 (9), 1558–1576.
- Rekhviashvili, L., Sgibnev, W., 2018a. Placing transport workers on the agenda: the conflicting logics of governing mobility on Bishkek's Marshrutkas. *Antipode* 50 (5), 1376–1395.
- Rekhviashvili, L., Sgibnev, W., 2018b. Uber, Marshrutkas and socially (dis-)embedded mobilities. *J. Transp. Hist.* 39 (1), 72–91.
- Rommerts, M., 2012. The role of EU-supported projects in policy transfer in urban transport. Doctoral thesis. <https://discovery.ucl.ac.uk/id/eprint/1348582/>.
- Roy, A., 2011. Urbanisms, worlding practices, and the theory of planning. *Plan. Theory* 10 (1), 6–15.
- Rupprecht Consult (editor), 2019. *Guidelines for developing and implementing a sustainable urban mobility plan*, Second Edition. https://www.eltis.org/sites/default/files/sump-guidelines-2019_mediumres.pdf.
- Rye, T., Welsch, J., Plevnik, A., de Tomassi, R., 2011. First steps towards cross-national transfer in integrating mobility management and land use planning in the EU and Switzerland. *Transp. Policy* 18, 533–543.
- Sagaris, L., 2018. Citizen participation for sustainable transport: Lessons for change from Santiago and Temuco, Chile. *Res. Transp. Econ.* 69, 402–410.
- Sengers, F., Raven, R., 2015. Toward a spatial perspective on niche development: The case of bus rapid transit. *Environ. Innov. Societ. Transit.* 17, 166–182.
- Sheldrick, A., Evans, J., Schliwa, G., 2017. Policy learning and sustainable urban transitions: Mobilising Berlin's cycling renaissance. *Urban Stud.* 54 (12), 2739–2762.
- Söderström, O., Geertman, S., 2013. Loose threads: The translocal making of public space policy in Hanoi. *Singapore J. Trop. Geogr.* 34, 244–260.
- Sosa López, O., Montero, S., 2018. Expert-citizens: Producing and contesting mobility policy in Mexican cities. *J. Trans. Geogr.* 67, 137–144.
- Thomas, A., Deakin, E., 2017. Managing partnerships for sustainable development: the Berkeley-China sustainable transportation program. *Case Stud. Trans. Policy* 5, 45–54.
- Thomas, R., Bertolini, L., 2015. Policy transfer among planners in transit oriented development. *Town Plan. Rev.* 86 (5), 537–560.
- Timms, P.M., 2011. Urban transport policy transfer: 'top down' and 'bottom up' perspectives. *Transp. Policy* 18 (3), 513–521.
- Timms, P.M., 2014. Transferability of urban freight transport measures: a case study of Cariacica (Brazil). *Res. Transp. Bus. Manag.* 11, 63–74.
- TRIP, 2014. Thematic Research Summary: Land use and transport planning. <https://trimis.ec.europa.eu/content/land-use-and-transport-planning>.
- TRIP, 2016. Research Theme Analysis Report: Urban Mobility. <https://trimis.ec.europa.eu/content/urban-mobility>.
- Vigar, G., 2017. The four knowledges of transport planning: Enacting a more communicative, trans-disciplinary policy and decision-making. *Transp. Policy* 58, 39–45.
- Wood, A., 2014a. Moving policy: global and local characters circulating bus rapid transit through South African cities. *Urban Geogr.* 35 (8), 1238–1254.
- Wood, A., 2014b. The politics of policy circulation: Unpacking the relationship between South African and South American cities in the adoption of bus rapid transit. *Antipode* 47 (4), 1062–1079.
- Wood, A., 2014c. Learning through policy tourism: circulating bus rapid transit from South America to South Africa. *Environ. Plan. A* 46, 2654–2669.

Accessibility Tools for Transport Policy and Planning

Benjamin Büttner, Technical University of Munich, München, Germany

© 2021 Elsevier Ltd. All rights reserved.

Background and Development Over Time	83
Implementation Gap	83
Accessibility Concept	83
Accessibility Measures	84
Accessibility Tools	84
Application of Accessibility in Planning Practice and Policy Making	85
Critical Reflection for New Horizons	85
References	86

Background and Development Over Time

The concept of accessibility in regards to integrated land-use and transport planning has been a popular research topic for quite a while now (Hansen, 1959). The value of accessibility indicators as planning tools has also been well established, not only as a theoretical means of examining cities, but also as a tool for evaluating transportation policies (Koenig, 1980). While these tools have changed and improved over time, there has been discussion surrounding the practicality and value of different types of accessibility indicators (Geurs and Ritsema van Eck, 2001). Whether or not this wide array of accessibility tools is actually useful or helpful to practitioners is a prominent discussion topic (Silva and Larsson, 2018). The disconnect between the intended purposes of these tools and the actual implementation by practitioners creates an “implementation gap” that needs to be addressed (te Brömmelstroet et al., 2016).

The development of accessibility tools has not been static. Instead, tools have evolved and changed over time in order to take advantage of new technology and meet new demands. For years, tools focused on mathematical accuracy and precision as new technology allowed for tools to measure accessibility in ever increasing detail. However, more detail does not necessarily mean more useful. Finding a balance between these two characteristics is a challenge that still needs to be addressed by academics and practitioners.

The actual value of the abundance of accessibility tools is debatable. On one hand, a push for precision allows for the urban environment to be examined in more detail and with more accuracy, but on the other hand, this increase in complexity and detail is not necessarily useful to practitioners and can create a divide between the intended purposes of the tools and the actual needs of the practitioners. The “implementation gap” that is present in accessibility tools will be discussed further in the second section of this chapter.

Despite the proliferation of accessibility tools—even those that might not be useful to practitioners—new accessibility tools will likely need to be created in order to study the rapidly changing transportation landscape brought on by privatization in transportation markets, the development of new technology, and increased concern for the environmental impacts of transportation. The uncertainty of the future raises many questions about accessibility that need to be researched. Details regarding the future of accessibility tools and accessibility research will be addressed in the fourth and final section of this chapter.

Implementation Gap

The implementation gap in accessibility tools takes shape in three different forms: gaps in the accessibility concept, gaps in the accessibility measures, and gaps in the accessibility tools. More details about these implementation gaps will be discussed in the following sections.

Accessibility Concept

Despite the fact that accessibility as a concept has been known and used by planning practitioners for years, many modern planning policies still lack the requirement to take accessibility into account in urban and transport planning. In a survey of policies in 16 countries, it was found that only three of the countries surveyed had some sort of requirement to take accessibility into account in urban and transport planning (Hull et al., 2012b). The United Kingdom, Norway, and Germany had some sort of requirement to consider accessibility, while in other countries (Sweden, Belgium, Spain, Greece, and Denmark) it was only recommended that planners use an accessibility instrument. Finland, Australia, Poland, Italy, Portugal, Slovenia, and the Netherlands did not have a requirement to perform an accessibility assessment.

This survey highlights a significant issue with accessibility instruments. Despite the abundance of available tools that practitioners could use to measure accessibility in any number of ways for a variety of circumstances, many practitioners are not even required to consider accessibility. The reason for this is not always clear, but the concept of accessibility may not be well understood and the accessibility tools may be seen as complicated black boxes (te Brömmelstroet, 2014).

There is a conceptual mismatch when it comes to accessibility and what it exactly means to take accessibility into account when planning in land-use and transport. Accessibility can take many forms; for whom and for what purpose accessibility should be provided can vary tremendously and a much more nuanced approach is needed to take into account all of the different accessibility possibilities (Silva and Larsson, 2018). Practitioners and municipalities often only consider accessibility as a single goal without contemplating the many different shades and variations that accessibility could be realized. This way of thinking highlights a general misunderstanding of accessibility as a concept (Silva and Larsson, 2018).

Accessibility Measures

As discussed in the previous section, what exactly is included in the concept of accessibility can vary significantly. There are many different variables that could be measured in order to take accessibility into account. For example, time accessibility to jobs or public transport, but also cost accessibility to specific goods or services (Büttner, 2016). That being said, there is often a gap in the accessibility measures that can be used and those that are actually used or should be used. Access to jobs is a measurement that is often used in order to quantify accessibility, which is important, but not the be-all and end-all of accessibility. However, this might be the deciding factor for determining whether or not an infrastructure project receives funding or gets built. While job accessibility is important, using this single measurement in order to make decisions is not a very comprehensive approach and does not take into account the all of the everyday needs of actual people (Silva and Larsson, 2018).

In general, accessibility measures often examine accessibility from one of four main perspectives: infrastructure-based, location-based, person-based, or utility-based (Geurs and van Wee, 2004). Each of these perspectives has a different focus and level of suitability for different needs. However, it is important to note that while these different perspectives are all for measures of accessibility, they are not interchangeable. Geurs (2018) argues that time and time again we can see that conclusions of accessibility are greatly dependent on the type of measure that is used. This highlights the significance of the gap between accessibility measures that are available and those, which are used. While it may be possible to use one measure to assess accessibility, it might not be the most appropriate measure.

In the case of the Netherlands, Geurs (2018) breaks down Dutch national and regional transport planning and identifies five main limitations involving accessibility measures. These five main limitations highlight common instances where accessibility measures might be used by practitioners, but possibly inappropriately.

The first limitation is that planners might understand the significance of measuring accessibility for transportation planning, but other areas that would also benefit from using accessibility measures, such as public health and education, might be ignored. This “sectoral policy approach” misses opportunities to include accessibility measures in areas where they are needed.

The second limitation is that the interactions of land-use and transport are often ignored. While value can be gained by adding accessibility measures to a land-use analysis, land-use and transport are often considered separately.

The third limitation is that policies often focus exclusively on travel time and ignore other transport measures, such as cost or comfort. Individual components are also often ignored. The lack of individual differentiation means that it is not always clear what different levels of accessibility are available to different population groups.

The fourth limitation is that measures to address small-scale urban mobility, such as walking or cycling, and multimodality are often left out of accessibility indicators. While car and public transport are significant, focusing on them alone will not tell the whole story.

The fifth limitation is that the effects on equity are often left out of accessibility indicators. Addressing this limitation, as well as the other four limitations, is necessary if practitioners and academics are going to close the gap between the accessibility measures that get used and the measures that should be used.

Accessibility Tools

One of the most significant implementation gaps for accessibility tools is the gap between the tools that get developed and those that actually get used. Among the plethora of accessibility tools that are constantly being hatched up by academics, a relatively small percentage actually gets used by practitioners. While the cause of this phenomenon is up for debate, there are a number of factors that contribute to the lack of acceptance of these tools.

One possible explanation for the implementation gap of accessibility tools is the disconnect between planning practitioners and the tool developers. The two groups may have different goals and may be unaware of the needs of the other (Silva and Larsson, 2018). The developers of accessibility tools need to walk a thin line between creating measures that accurately represent the intricacies of the urban environment while also being approachable and understandable by all of the stakeholders involved (Hull et al., 2012a). Unfortunately, many accessibility tools often end up too far on one side of that line.

There are a number of barriers to using accessibility instruments that either prevent practitioners from using them, or prevent them from being used effectively. These barriers can be grouped into two main groups, technical/resource barriers and political barriers (te Brömmelstroet et al., 2014). Practitioners are often unsure if their organizations have sufficient familiarity with

accessibility instruments and if their organizations have sufficient resources, such as computational power and funding, in order to effectively use the accessibility instruments (Büttner et al., 2018).

Application of Accessibility in Planning Practice and Policy Making

As discussed in the previous section, there are gaps in the field of accessibility tools. However, it is important to note the accessibility tools that are being applied to planning practice and policy making. Thoroughly examining these can help shine light on which attributes of these tools are assets to the planning process and which attributes are a hindrance to the same process.

Unfortunately, using accessibility tools in planning practice does not always guarantee success and can produce mixed results. An example of this is the ABC location policy in the Netherlands. Enacted in 1990, the Dutch ABC location policy categorizes land into three main types (A, B, or C), then recommends where different types of businesses should be located based on accessibility to public transport. In general, A locations are located close to public transport and have very strict parking regulations, B locations are well-connected to public transport, but outside of city centers, and C locations have no parking limitations and typically have no parking restrictions. The idea is to integrate land-use and transport accessibility into a single policy.

Despite the intention to use an accessibility policy to coordinate land-use and transport the Dutch ABC location policy has been criticized for a few shortcomings. While the intended purpose of the plan is to combat traffic congestion by encouraging companies to locate in appropriate locations in relation to public transport, the actual effects on car-use have been negligible (Dieperink and Driessen, 2000). One possible reason for the limited effects is the poor alignment between the accessibility profiles of businesses and what businesses actually wanted (Dieperink and Driessen, 2000). This means that businesses that the government thought should be located in central areas close to transit were often more interested in being located near motorways and having ample parking.

Another possible reason for the limited success of the ABC location policy is that demand for office space in the 1990s was higher than expected (Schwanen et al., 2004). This meant that even if a business with an accessibility profile that required better access to transit really wanted to locate in an area categorized as type A, the business might not be able to locate there because of office space demands.

Germany has its own planning guidelines that are used to bring accessibility into the planning process. RIN (Richtlinien für integrierte Netzgestaltung) is a guideline that has been in use since 2008, and which specifies how far away people should be from different types of urban centers. These guidelines are noteworthy because they formally integrate public transport accessibility into the spatial planning process.

In Switzerland, the Agglomeration Policy is a national framework for integrating land-use and transport planning in urban regions. One of the key features of the policy is that it coordinates transport planning across multiple jurisdictions as well as across national borders. This is significant because the policy addresses the fact that people do not confine their lives to a single jurisdiction.

In addition to policy measures, easily-used accessibility tools are also available online. The WebCAT tool from Transport for London is an example of this. The interactive web tool visualizes public transport accessibility in London, not only for a base year, but also for several scenarios in the future. While this tool provides less in-depth analysis and detail than other accessibility tools, its primary asset is that it is approachable and user-friendly. This tool can easily be used by practitioners with different skill levels, as well as the general public.

The accessibility tools and policies described above are examples of how accessibility has been integrated into the planning process. Attributes of successful accessibility policies include formal integration of public transport accessibility into spatial planning process, as seen in Germany, cross-jurisdiction and cross-border accessibility, as seen in Switzerland, and approachability/ease-of-use, as seen with WebCAT in London. The shortcomings of the ABC location policy in the Netherlands also highlighted the importance of properly aligning the needs and desires of the different parties and stakeholders involved (Wulffhorst et al., 2017).

Critical Reflection for New Horizons

While the future of accessibility tools is not entirely clear, there are some noteworthy trends and concepts that are worth examining. New planning priorities and ideas may alter the form that future accessibility tools will take.

Climate change and its detrimental effects have moved to the forefront of public discourse in recent years. While this has taken shape in the form of calls to action—and criticism for inaction—there could and should be some overflow into the field of accessibility tools. Municipalities and governments may start to consider new accessibility measures, such as CO₂ emissions, that have previously been left out of accessibility research. Examples of such metrics can already be seen in some recent research (Kinigadner et al., 2019). Kinigadner et al. (2019) look at the effect firm location has on commuting-related CO₂ emissions and demonstrates a method for assessing this. The analysis revealed significant variation in commuting-related CO₂ emissions between firms located in central areas and firms located in peripheral areas. It is possible that newer emission-based measures—such as the one used in this analysis—will become more prevalent in planning practice as municipalities and governments focus more on meeting environmental goals.

Accessibility tools have changed and adapted over the years, but the transport landscape has been relatively stable. However, new technology and transport services have altered the urban-mobility environment and will likely continue changing it in the near future. Ride-sharing, ride-hailing, and car-sharing services have brought private business interests into the transport sector in a way

that was previously unseen. New technologies, such as electric scooters and free-floating bike-sharing systems, have added new mobility choices that were previously unavailable. Future technology, such as autonomous vehicles, will likely further disrupt the transport ecosystem and introduce new challenges. Accessibility research will most likely need to adapt in order to create new tools that can adequately assess the effects of new services and technology on accessibility.

Future accessibility tools might move toward being web-based and take advantage of open-source data. Examples of these types of tools already exist, but may increase in number in the near future. GOAT (Geo Open Accessibility Tool)^a is an example of a web-based accessibility tool that can be used to visualize local accessibility as well as build scenarios on a neighborhood scale while focusing on active modes. The tool is easily approachable, but provides the opportunity to increase the complexity of an analysis to fit the user's specific needs. Tools like GOAT bridge the gap between academics and practitioners by being user-friendly, but also flexible and detailed.

Accessibility planning faces many challenges in the future. Not only will new tools and better coordination with practitioners be needed in order to address the implementation gap, but new tools will also likely be needed in order to evaluate the changing urban mobility environment. With this in mind, an open platform was developed under www.accessibilityplanning.eu. The Meta Accessibility platform is collecting and sorting accessibility tools worldwide in respect to its purpose, functionality and target group. By this, suitable tools can be aligned to specific issues of different user groups (ranging from decision-makers, planning practitioners, and academics to citizens). Taking into account their respective needs the implementation gap of accessibility tools can eventually be overcome.

References

- Büttner, B., 2016. Sharp increases in mobility costs: A trigger for sustainable mobility. In: Wulforst, G., Klug, S. (Eds.), *Sustainable Mobility in Metropolitan Regions – Insights from interdisciplinary research for practice application*, Springer, Wiesbaden, doi:10.1007/978-3-658-14428-9_8.
- Büttner, B., Kinigadner, J., Ji, C., Wright, B., Wulforst, G., 2018. 2018. The TUM accessibility atlas: visualizing spatial and socioeconomic disparities in accessibility to support regional land-use and transport planning. *Netw. Spat. Econ.* 14 (2), 385–414, doi:10.1007/s11067-017-9378-6.
- Dieperink, C., Driessen, P., 2000. The ABC of Dutch location policy: Lessons in logic. *Eur. Spat. Res. Policy* 7, 5–19.
- Geurs, K.T., 2018. Transport Planning With Accessibility Indices in the Netherlands (International Transport Forum Discussion Papers No. 2018/09). International Transport Forum. Available from: <https://doi.org/10.1787/c62be65d-en>.
- Geurs, K.T., Ritsema van Eck, J.R., 2001. Accessibility Measures: Review and Applications, No. RIVM Report 408505 006. Urban Research Centre, Utrecht University.
- Geurs, K.T., van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: Review and research directions. *J. Trans. Geogr.* 12, 127–140, doi:10.1016/j.jtrangeo.2003.10.005.
- Hansen, W.G., 1959. How accessibility shapes land use. *J. Am. Inst. Plan.* 25, 73–76, doi:10.1080/01944365908978307.
- Hull, A., Bertolini, Luca, Silva, C., 2012a. Chapter 5. CONCLUSIONS, in: *Accessibility Instruments for Planning Practice*. COST Office, S.I., pp. 241–251.
- Hull, A., Papa, E., Silva, C., Joutsiniemi, A., 2012b. Chapter 4. ACCESSIBILITY INSTRUMENTS SURVEY, in: *Accessibility Instruments for Planning Practice*. COST Office, S.I., pp. 207–237.
- Kinigadner, J., Büttner, B., Wulforst, G., 2019. Beer versus bits: CO2-based accessibility analysis of firms' location choices and implications for low carbon workplace development. *Appl. Mobil.* 4, 200–218, doi:10.1080/23800127.2019.1572053.
- Koenig, J.G., 1980. Indicators of urban accessibility: Theory and application. *Transportation* 9, 145–172, doi:10.1007/BF00167128.
- Schwanen, T., Dijst, M., Dieleman, F.M., 2004. Policies for urban form and their impact on travel: The Netherlands experience. *Urban Stud.* 41, 579–603, doi:10.1080/0042098042000178690.
- Silva, C., Larsson, A., 2018. Challenges for Accessibility Planning and Research in the Context of Sustainable Mobility (International Transport Forum Discussion Papers No. 2018/07). Available from: <https://doi.org/10.1787/8e37f587-en>.
- te Brömmelstroet, M., 2014. Use of accessibility instruments Assessing Usability of Accessibility Instruments, COST Office, Amsterdam, pp. 1–5.
- te Brömmelstroet, M., Curtis, C., Larsson, A., Milakis, D., 2016. Strengths and weaknesses of accessibility instruments in planning practice: Technological rules based on experiential workshops. *Eur. Plan. Stud.* 24, 1175–1196, doi:10.1080/09654313.2015.1135231.
- te Brömmelstroet, M., Curtis, C., Larsson, A., Milakis, D., 2014. Conclusions and discussion. COST Action TU1002 – Assessing Usability of Accessibility Instruments, COST, Brussels, pp. 173–183.
- Wulforst, G., Büttner, B., Ji, C., 2017. The TUM Accessibility Atlas as a tool for supporting policies of sustainable mobility in metropolitan regions. *Trans. Res. Part A: Policy Prac.* 104, 121–136, doi:10.1016/j.tra.2017.04.012.

^a Accessible under www.open-accessibility.org

Demand Responsive Transport

Marcus Enoch, School of Architecture, Building and Civil Engineering, Loughborough University, Leicestershire, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	87
Defining and Characterizing DRT Systems	88
Characteristics of DRT Systems	88
Delivery Attributes	88
Organizational Aspects	89
Financial Elements	90
The Development of DRT	91
Phase 1: 1920–1969	91
Phase 2: 1970–2019	91
Phase 3: 2020 Onwards	92
Future Implications	92
References	93
Further Reading	93

Introduction

"[Ever] since the French mathematician and philosopher Blaise Pascal began operating what is thought to be the world's first public transport service in 1662, public transport has invariably meant some type of vehicle (or groups of vehicles) shuttling back and forth, terminal to terminal. While there have been variations, e.g., services with formal stops and schedules versus stop anywhere and no schedules, the basic public transport offer has not changed much."

Agarwal et al., (2019, pp. 77)

Unfortunately, while the need for mass transport systems able to move lots of people in as efficient a way as possible continues to become ever more vital to our society, the ability of the traditional methods of achieving that (particularly buses) is becoming less and less effective at doing so. Thus, in Great Britain for instance, the modal share of the bus (as measured as a percentage of billion passenger kilometers) fell from 42% of all trips in 1952, to only 5% in 2017—that is, from 92 to 38 billion passenger kilometers. By contrast car use increased from 58 billion passenger kilometers or 27% of all trips to 670 billion passenger kilometers (83%) over the same period (Department for Transport, 2018). In understanding the reasons for this, perhaps the classic explanation of what has happened is the so-called vicious cycle of decline (Mohring, 1972). Under this model, as society got richer then more people bought cars; and this increased traffic congestion which slowed down buses and led to less people taking the bus; and made the bus still less attractive because either services were cut or fares were raised to cover the increased costs of operating the buses; which meant more people bought cars and increased traffic congestion further; and so less people used buses etc. Not only that, but as car ownership increased, then new housing, retail and employment sites began to be located in areas not well served by public transport which in turn reinforced the steady decline in public transport use still further (Rodrigue, 2017). More recent analysis reported that bus use is suffering due to the availability of viable alternatives to the bus, changing public attitudes, and a series of wider social and economic trends (Bray and Bellamy, 2019).

One long-proposed solution for addressing these problems facing conventional public transport services, has been to introduce a form of public transport that is better suited to operating in these conditions of lower demand, that is, demand responsive transport (DRT). As a concept, DRT systems are appealing to operators in areas and/or at times of low demand, because they theoretically provide public transport users with a service that is higher quality than that offered by a bus, but for a fare that is cheaper than a taxi, and at a cost that is lower than for a conventional big bus. Yet despite the potential scope and appropriateness of DRT for modern transport needs, in practice they remain a niche concern. For instance, a survey of British local authorities revealed there to be a relatively small number of DRT schemes in operation (369 from a response rate of 47% of councils, crudely suggesting a total of around 800 DRT services across the country) compared with roughly 22,000 bus services—that is, about 4% of services (Davison et al., 2014; Stagecoach, 2015). That said, a 2010 survey of 194 transit agencies across the United States and Canada (a response rate of 45%) found that 39% operated some form of DRT for some services (Potts et al., 2010). Moreover, this also showed that DRT schemes were being introduced by a growing number of agencies each decade: thus eight introduced DRT schemes in the 1980s, 18 in the 1990s, 35 in the 1990s and 40 in the decade up to 2010.

Such a status has previously been ascribed to three main barriers: technological, institutional and economic (Cervero, 1997; Enoch et al., 2004; Mulley et al., 2012). In particular, technological challenges tended to relate to optimizing the booking,

scheduling, and routing functions, while institutional barriers included navigating the diverse range of licensing regimes for operators, drivers, vehicles and routes or service areas, which in turn had major implications on insurance, subsidy, tax, VAT, safety and several other operational questions. Meanwhile economic issues focused on the business model underpinning DRT. In sum, small vehicles generally are unable to generate sufficient revenue from the relatively low numbers of passengers often paying relatively low fares to cover the high costs of provision (particularly the driver costs).

The aim of this chapter therefore, is to look at exactly what DRT is, at its evolution over time, and at how it might develop in the future.

Defining and Characterizing DRT Systems

The scope of what are DRT systems and what are not is actually rather blurred, but for the purpose of completeness in this chapter the term will refer to “an intermediate form of public transport, somewhere between a regular service route that uses small low floor buses and variably routed highly personalized transport services offered by taxis” (Brake et al., 2004, p. 324). Similarly, Mageean and Nelson (2003) defined DRT as providing “on-demand [transport] for passengers using fleets of vehicles scheduled to pick up and drop off people in accordance with their needs” (p. 255). They add that it is seen by some as a tool that could fill the gap between a fixed route bus and a taxi in order to meet the needs of certain members of the population. Other related terms include paratransit (of which DRT could be considered as a subset) and flexible transport systems, which is a broadly similar concept. Fundamentally, DRT services may not have fixed routes or timetables but are available to unrelated members of the general public and are usually operated by vehicles that are smaller than a conventional bus.

Characteristics of DRT Systems

There are a wide range of types of DRT. Critical service design features are focused on how the service is delivered, organized, and financed.

Delivery Attributes

The key delivery attributes include how the service is routed, timetabled, accessed by the user, and on the vehicle size chosen. Spatially, DRT services can either follow fixed routes; be semi-flexibly routed and deviate from a fixed route on request (either via the booking system or when already on the vehicle); else be fully flexibly routed and operate within a defined zone. Related to this, DRT services can operate between a single trip origin and a single destination; between a single origin/destination and multiple destinations/origins; or between multiple origins and destinations.

Temporally, DRT can either operate to a fixed schedule (through a number of fixed timing points); operate on-demand, that is, when requested to do so by passengers; else be unscheduled, whereby operators provide services in areas and/or at times when they can expect to be busy—that is, they respond to “historical” patterns of demand.

These alternatives are illustrated in Fig. 1.

Accessing DRT services can be done in various ways. At its simplest, passengers turn up to a terminus or stop and physically request to be allowed to board at the time of departure. Alternatively, passengers can telephone or book through the Internet or via an app.

Lastly, vehicle sizes used for DRT typically range from options including 4–5 seat shared taxi-type services, to 7–8 seat taxis, to 15–20 seat minibuses, to 25–40 seat midi-buses. Decisions as to which sized vehicle to employ are clearly dependent on expected peak loadings, but also on a range of institutional factors around licensing and tax regimes for example (see later for more on this).

Planning the more complex service patterns in an effective and efficient way, particularly for systems with larger numbers of vehicles, has only really become viable with the development of advanced computer software packages (such as provided by Trapeze) since the mid-1990s to process booking requests and make appropriate vehicle and driver scheduling allocations. They can also require passengers to book (sometimes several hours) in advance, and/or require travelers to be flexible in their departure and arrival times. Typically there is a trade-off between service quality for the passenger, and the degree of operational efficiency that is possible. One other routing choice is between whether a service is made door-to-door or merely checkpoint-to-checkpoint, with the former option perhaps being better for the passenger being picked up, but worse in terms of slower journey times for the passengers already on the vehicle and the overall operational performance.

In all these planning and operational regards, technology has been rapidly advancing. Most recently, “Mobility-as-a-Service” (MaaS) solutions allow transport providers to bundle a range of service attributes together into packages, which are then offered to users who can then select the option that best meets their needs via digital transport service platforms that enable users to access, pay for, and get real-time information on, a range of public and private transport options including DRT services (Enoch, 2018). MaaS platforms thus create efficient and dynamic marketplaces for transport due to transparent information flows and minimal transaction costs, and hence could revolutionize how transport generally and DRT specifically is provided and consumed (see later).

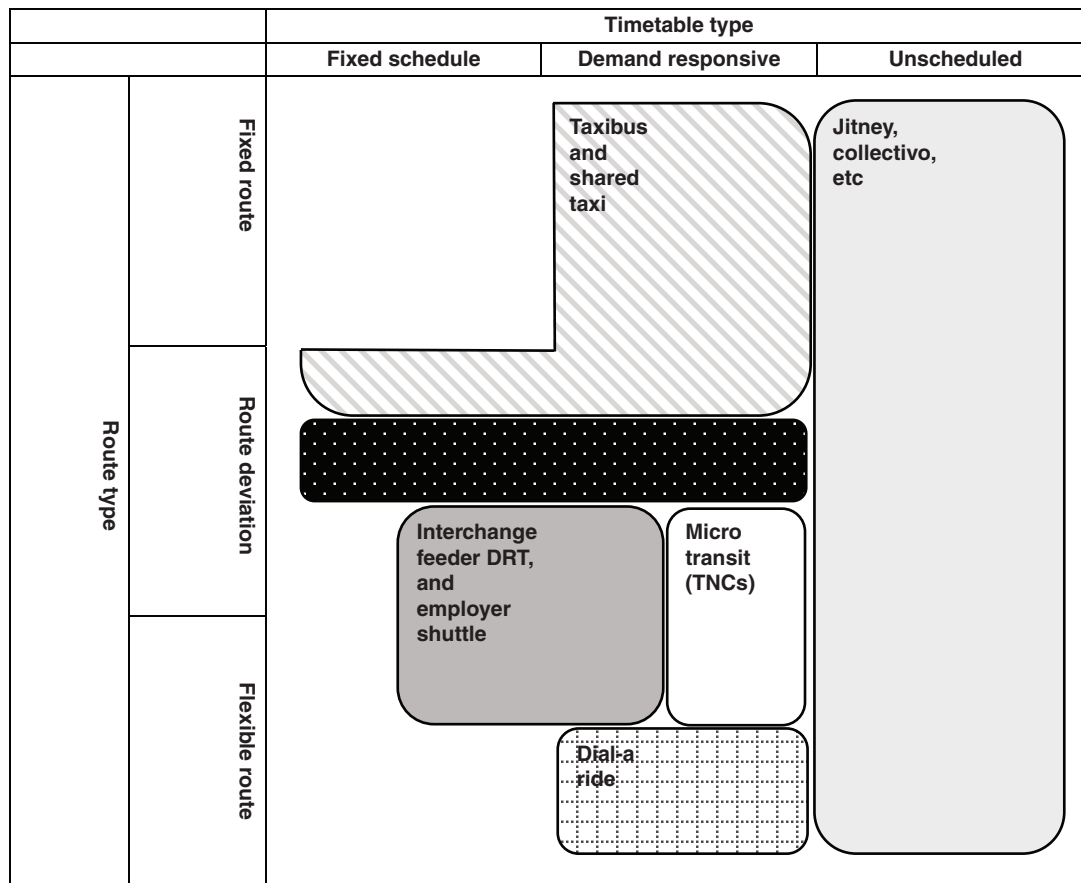


Figure 1 Example DRT scheme types by routing and timetabling options. Source: Adapted from *Cervero (1997)* and *Enoch et al. (2004)*.

Organizational Aspects

The organizational aspects include questions around why and how DRT systems are introduced and operated, which in turn can impact on the types of operators and regulatory frameworks involved. There have been two main motivations for introducing DRT systems—either for commercial reasons or to meet public policy goals. Thus, the majority of jitney operations in cities across the United States from the early 1900s, and a whole range of similar systems that still operate throughout South America, Africa, Asia and Eastern Europe (e.g., Dolmus in Turkey and Mashrukta in Kenya) are operated for profit (*Cervero, 1997*). Typically in these cases, the need for transport is not being met by private vehicles nor by conventional public transport, and hence the DRT operators step in to serve those people. By contrast the public policy driven schemes have generally been targeted at enabling citizens living in areas where demand is insufficient to justify operating conventional public transport services to access facilities (e.g., Regiotaxi in the Netherlands, InterConnect and Cango in the UK) (*Enoch et al., 2004*).

In identifying current and future markets for DRT, *Davison et al. (2012)* devised the following framework in terms of market and product position and potential (**Table 1**).

From this UK 2012 analysis, it can be seen that DRT is currently strongly associated with rural areas and mobility impaired and older people, but that this was predicted to change to a more suburban and working age-type landscape. Interestingly, evidence as to where the most recent DRT service models are now being applied is seemingly following this path, but in addition is focused on large urban population centers.

In terms of ownership, the commercially operated schemes are generally operated by private operators—often owner-drivers or cooperatives, though not always. Meanwhile the public policy driven schemes are run by a much broader range of publicly, privately and socially-owned providers from the very small to multinational organizations.

Finally there are a whole range of institutional complications that can interfere with the development of DRT and indeed other forms of new mobility systems that relate to the diverse range of licensing regimes for operators, drivers, vehicles and routes or service areas, which in turn have major implications on insurance, subsidy, tax, VAT, safety and several other operational questions (*Enoch et al., 2004*). Meanwhile DRT operations are also often excluded from a whole range of institutional and financial support due to policies and financing mechanisms that define public transport according to the existing large vehicle/corridor system model. One approach in simplifying this could be to regulate the day-to-day aspects of each transport operator (driver licensing, subsidy allocation, etc.) almost as they are now, but then to classify operators by their level of specialism (e.g., occasional, regular, specialist),

Table 1 Demand responsive transport market and product position and potential

	<i>Existing/strong products Market penetration</i>	<i>New/less certain products Product development</i>
Existing/strong markets	Rural, general public Without access to car and conventional public transport Mobility impaired Non-emergency patient transport Shopping, suburban and rural, general public Educational establishment for students with special educational needs Market development	Urban orbital, general public Airport access for passengers and employees (Market penetration in USA and Europe) Workplaces, outside urban core, employees Hospitals and other destinations, specialist needs Diversification
New/less certain markets	Educational establishments, rural, students Trip attractors, rural areas, tourists Integrated DRT supply for the general public	Suburban, general public (Product development in USA) Entertainment venues in urban centers Sport venues, ticket holders Meeting and conference venues, employees Services and good to rural areas

Source: Davison et al. (2012).

whereby the more specialist the operator, the tighter the regulations but the greater the operational benefits and opportunities (Enoch and Potter, 2016). This would mean that DRT modes for example, would no longer be forced into operating pre-conceived service patterns (constrained, for example, by limitations on number of seats, timetable schedules, route/area restrictions); and should allow for occasional providers to “safely” join the transport market as the opportunity arises.

Financial Elements

Establishing and maintaining the financial sustainability of a DRT system is perhaps the most challenging issue facing scheme promoters. Indeed, lack of financial viability is one of the commonest reasons for DRT schemes failing. This is partly a result of operating in areas of low demand. While for public policy schemes (which are often aimed at reducing levels of social exclusion) this is due to the inability to price the service at a rate that does not get close to reflecting the costs of providing it. In practice, this means that the traditional “big bus” commercial funding model, whereby one large vehicle on a major route packed with passengers for one journey in the morning and one in the evening can essentially cross-subsidize journeys for the rest of the day and across the network is not an option. The two fundamental approaches available to operators to secure a financially sustainable future are to raise revenue or to cut costs. Raising revenues can be done either through increasing monies from fares, through attracting more users and/or by increasing fares; or else by drawing in more budget from other sources. Thus, for the commercially driven “jitney-style” schemes on high-demand corridors that are otherwise poorly served by public transport, operators aim to fill up their vehicles with as many fare-paying riders as possible, often not even setting off until they are full. Next, airport shuttles in the United States charge a fare somewhere between the public transport and airport taxi rates, so reflecting the intermediate quality of the journey experience. By contrast, most public policy motivated scheme promoters are loath to charge more than a standard bus fare (generally with good reason) in spite of the far higher cost of providing those services, and hence their services require significant levels of subsidy. Interestingly, one slight exception is that of St Joseph Transit in Kansas City, Kansas, which charges a \$0.50 supplement for passenger wishing to deviate from the route (Potts et al., 2010). Consequently, marketing services can be crucially important ways of increasing DRT patronage, which Potts et al. (2010) recommend doing through community groups, events, targeted mailings, internet sites and posters/leaflets/brochures. Other mechanisms for covering costs, are to generate ancillary revenue streams, rely on local authority subsidies, or attract new monies from alternative external sources. Thus, for many years Lincolnshire County Council has offered its call centre services to neighboring DRT-operating local authorities and has successfully bid for additional funding from various national and European pots. And the PickMeUp DRT scheme launched in 2018 in Oxford, UK is partly financially supported by employers.

Cutting costs has been done by rationalizing service levels and by making efficiency savings. For instance, Wiltshire County Council, UK consolidated its DRT provision following a strategy review in 2006, while Flintshire Council, UK ended up replacing the demand responsive component of its Deeside Shuttle service in 2015 to reduce passenger subsidy levels. Further efficiency savings have also been sought through different transport providing agencies in Somerset, UK sharing access to their vehicles (e.g., community transport, social services, education, public transport operators and voluntary groups) through a brokerage-based Integrated Passenger Transport Unit approach. Enoch et al. (2007) added that it is also necessary for agencies to secure political support based on a strong strategic vision of how DRT is integral to delivering social inclusion, environmental and economic goals—and so make the services less vulnerable to cuts than otherwise. Finally on cost cutting, the development of autonomous vehicles

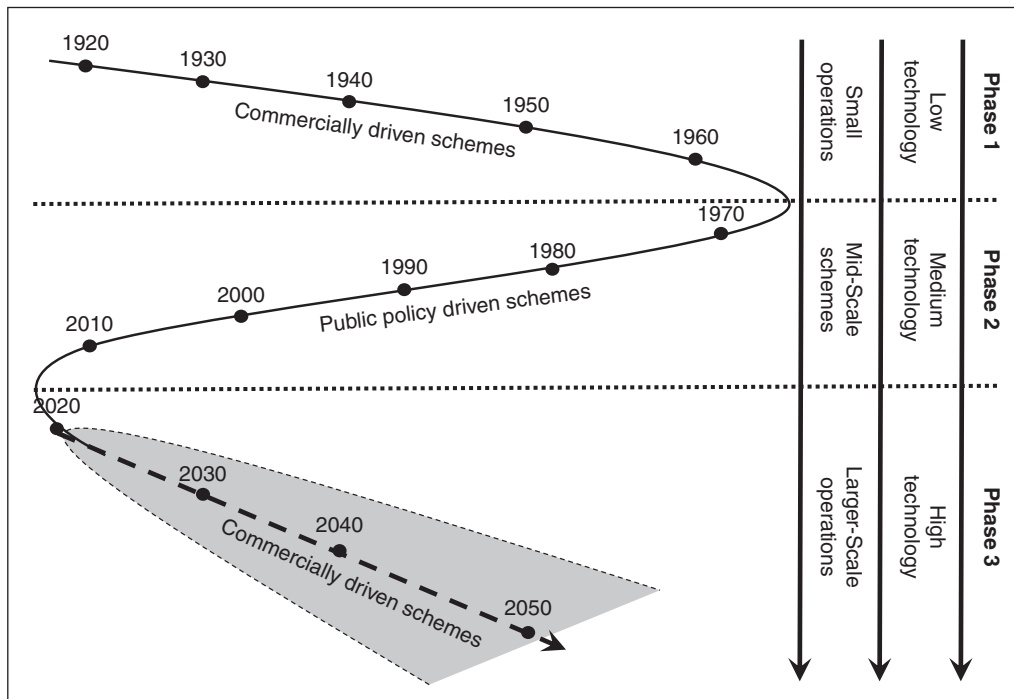


Figure 2 The evolution of DRT systems.

could have significant impacts in the medium term, with a number of tests currently being carried out with driverless buses around the world.

The Development of DRT

Pulling these strands together, looking back at the evolution of DRT over the past century in many countries around the world, we can see two distinct DRT operational phases in place, while we now seem about to enter a third (see Fig. 2).

Phase 1: 1920–1969

Under the first phase, “informal” DRT services such as jitneys have operated across the world since the early 20th century. These were/are relatively simple and low-technology minibus or shared taxi-based systems that operate for a profit and typically serve low-price market niches on busy corridors that are underserved by conventional public transport services. Passengers generally turn up at a terminus or hail from the roadside but can book ahead by phone in some cases. Services usually operate on demand as opposed to by a fixed timetable, along corridors with drivers sometimes deviating on request, and fares are broadly comparable to bus fares—sometimes slightly less, or slightly more, depending on whether they are perceived as being of lower or higher quality. Examples of such schemes still operate in countries as diverse as Tunisia, Russia, Mauritius, Kenya, the Philippines, Cuba, Northern Ireland, and the United States.

Phase 2: 1970–2019

The second “dial-a-ride” operational model began to emerge in the 1970s and 1980s, when public policy driven service types began to be trialed, for instance in cities such as Milton Keynes, UK and Merrill, Wisconsin (Flusberg, 1976). This “medium technology” model operated on the basis of passengers booking a seat on a minibus for a trip to be made within a specific time window usually at least an hour (often 24 h or more) in advance, so that the operator could then manually allocate the drivers and vehicles to the most efficient service pattern within a particular operational area (Kirby et al., 1974; Nutley, 1990). Unfortunately however, these operations were nearly always heavily subsidized and often failed for all but the most specialist trips due to the labor-intensiveness of the system, high labor costs and inefficiencies such as many service vehicles travelling without or with few passengers. Another reason for low performance was that the systems were simply not attractive to consumers (Khattack and Yim, 2004).

Subsequently, advances in software, computers, digital maps, expert systems, remote communications, in-vehicle computers and GPS from the late 1990s led to DRT steadily becoming more technologically viable in the late 1990s (Duffell, 2003). In Britain and

some countries elsewhere, these developments were stimulated for a period by the availability of government funding aimed at overcoming social exclusion and resulted in a sudden explosion in the take up of DRT, primarily by local authorities, but also by community transport (CT) operations, commercial public transport operators and even employers (Enoch et al., 2004). In recent years, however, with the removal of the government subsidy, many of the local authorities administering these schemes have faced difficult choices over how to proceed.

Phase 3: 2020 Onwards

Going forward, there are already indications that we are entering a third operational phase known as “micro-transit,” this time based on so-called transportation network companies (TNCs). In this context, sizeable multi-national companies such as Uber and Lyft are now trialing a range of innovative mobility services, including some like Uber POOL and Lyft Line, as well as smaller start-ups such as Go-Opti—which use on-line brokerage portals known as Digital Service Platforms to match taxi and minibus drivers with (unrelated) people wanting to make similar journeys, to encourage them to make their trips together for a lower fare than if they made them separately. This high-technology approach is commercially driven and appears to have the potential to be profitable at some point in the future in some areas at least. This is by virtue of “big data” and “machine learning” methods being used to vary prices for users and payments to drivers according to the relevant levels of demand and supply in real-time and across the network. Unfortunately though, they have not yet made money, and indeed several other micro-transit schemes such as Charriot, Slide, Kutsoplus and Bridj have subsequently stopped operating.

TNC operators are also in the process of working with local authorities in the United States and Canada to provide such DRT services as an alternative to buses—whereby the councils subsidize trips made by the Uber or Lyft to help them meet public policy goals—though in practice one imagines only a few of these are likely to be shared between unrelated passengers.

Future Implications

More broadly, Enoch (2015) and Enoch and Potter (2016) suggest that such a shift toward these microtransit DRT-type modes is likely to accelerate due to a range of factors that are pushing away from the conventional modes of car, bus, taxi and indeed traditional DRT. These include:

- More elderly people who will no longer be able to drive but who need access to places buses serve poorly; more younger people excluded from car ownership by high insurance costs and competing demands on their incomes;
- A number of deep-rooted economic and social factors that are leading to travel demand becoming increasingly dispersed in time, space and across functions.
- A growing culture of “collaborative consumption”;
- Increasing pressures on the global economy and the impact of the austerity agenda in many countries on revenue budgets. Expensive public transport infrastructure projects will be difficult to fund and bus subsidies hard to justify compared to commercial paratransit;
- The political desires to deregulate policy sectors and promote “choice” as a means of improving service quality;
- The increasingly blurred boundaries within the intermediate transport mode supplier sector and the increasing range of “new mobility solutions”;
- The increasing desire to better integrate transport options to create a more user-friendly transport system, through spatial, temporal, ticketing, information and seat brokerage MaaS mechanisms;
- The widespread adoption of big data technologies such as the Internet, smartphone, and GPS tracking technology; and
- The likely adoption of autonomous vehicle and connected autonomous vehicle technologies.

Overall, as with public transport generally, traditionally DRT services were introduced and operated based on (poor quality) ad hoc historic, averaged and aggregated data; operators were legally required to provide the services they are registered to whether used or not; and DRT users had very little direct influence on the type/level of service they received. In future, and with MaaS, transport users and providers will instantaneously communicate their needs via Digital Service Platforms (DSPs)—that is, two-sided markets or brokers—which, for a suitable charge, then make the most appropriate journey match (Djavadian and Chow, 2017). This means that:

- Users can now tell (multiple) operators of (multiple types of) transport service exactly the attributes they want from a transport service: time of arrival, origin and destination points, degree of flexibility, type/level of comfort required, and price they are prepared to pay;
- Providers, which now number many more and now offer many more different types of service than previously, can choose to respond, or not;
- Users, providers and regulators can readily monitor service availability and hence adjust their travel behavior, market strategy, or policy decisions accordingly.

In practice, providing that the regulatory agencies can overcome the various institutional challenges referred to earlier, such MaaS-type platforms could potentially mean that the supply of and demand for DRT and other transport services will be actively managed in near-real-time and in parallel—something which has not been possible before (Hensher, 2017). This should result in DRT services becoming more tailored to needs of users, while the supplier base will become more diverse and flexible. From this, it becomes possible to foresee instances where a commercial case for DRT begin to emerge, as the final barrier preventing the widespread adoption of DRT, that is, the economic viability, is overcome—particularly should driverless technology be adopted on a large scale.

References

- Agarwal, O.P., Kumar, A., Zimmerman, S., 2019. Public transit: from compulsion to choice. In: Agarwal, O.P., Kumar, A., Zimmerman, S. (Eds.), *Emerging paradigms in urban mobility*. Elsevier, Amsterdam, (Chapter 4), pp. 77–99.
- Brake, J., Nelson, J.D., Wright, S., 2004. Demand responsive transport: towards the emergence of a new market segment. *J. Transport Geogr.* 12 (2004), 323–337.
- Bray, J., Bellamy, S., 2019. What's driving bus patronage change? An analysis of the evidence base. Urban Transport Group. <http://www.urbantransportgroup.org> (last accessed 3 September 2019).
- Cervero, R., 1997. *Paratransit in America: redefining mass transportation*. Praeger, Westport.
- Davison, L.J., Enoch, M.P., Ryley, T.J., Quddus, M.A., Wang, C., 2012. Market niches for DRT. *Res. Transport. Bus. Manage.* 3, 50–61.
- Davison, L.J., Enoch, M.P., Ryley, T.J., Quddus, M.A., Wang, C., 2014. A survey of demand responsive transport in Great Britain. *Transport Policy* 31, 47–54.
- Department for Transport, 2018. *Transport statistics Great Britain*. Department for Transport, Crown Copyright, London.
- Djavadian, S., Chow, J.Y.J., 2017. An agent-based day-to-day adjustment process for modelling Mobility as a Service with a two-sided flexible transport market. *Transport. Res. B: Meth.* 104, 36–57.
- Duffell, J., 2003. Interview. Mobisoft, Witney, Oxfordshire (15 July).
- Enoch, M.P., 2015. How a rapid modal convergence into a universal automated taxi service could be the future for local passenger transport. *Technol. Anal. Strategy Manage.* 27 (8), 910–924.
- Enoch, M.P., 2018. Mobility as a service (MaaS) in the UK: change and its implications. *Future of Mobility Evidence Review*, Foresight, Government Office for Science, London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/766759/Mobilityasaservice.pdf (19 December).
- Enoch, M.P., Ison, S.G., Laws, R., 2007. Demand Responsive Transport in the UK: Routes for securing a financially sustainable future. In: *Transport, Mobility and Regional Development, Proceedings of the Regional Studies Association Winter Conference*, 23 November, London, 27–30.
- Enoch, M.P., Potter, S., 2016. Paratransit: the need for a regulatory revolution in the light of institutional inertia. In: Mulley, C., Nelson, J.D. (Eds.), *Paratransit: shaping the flexible transport future*. Emerald, Shipley, West Yorkshire, pp. 15–34.
- Enoch, M.P., Potter, S., Parkhurst, G., Smith, M.T., 2004. *Exploratory assessment of and innovations in demand responsive transport services - intermode*, Final Report. Department for Transport and Greater Manchester Passenger Transport Executive, London. <http://www.dft.gov.uk> (April).
- Flusberg, M., 1976. An innovative public transportation system for a small city: The Merrill, Wisconsin, Case Study. *Transport. Res. Record* 606, 54–59.
- Hensher, D.A., 2017. Future bus transport contracts under a mobility as a service (MaaS) regime in the digital age: are they likely to change? *Transport. Res. A Policy Pract.* 98, 86–96.
- Khattack, A.J., Yim, Y., 2004. Traveller response to innovative personalised demand responsive transit in the San Francisco Bay Area. *J. Urban Plan. Dev.* 42–55.
- Kirby, R.F., Bhatt, K.U., Kemp, M.A., McGillivray, R.G., Wohl, M., 1974. *Paratransit: neglected options for urban mobility*. The Urban Institute, Washington, DC.
- Mageean, J., Nelson, J.D., 2003. The evaluation of Demand Responsive Transport Services in Europe. *J. Trans. Geogr.* 11, 255–270.
- Mohring, H., 1972. Optimization and scale economies in urban bus operation. *Am. Econ. Rev.* 62 (4), 591–604.
- Mulley, C., Nelson, J.D., Teal, R., Wright, S., Daniels, R., 2012. Barriers to implementing flexible transport services: an international comparison of the experiences in Australia, Europe and USA. *Res. Trans. Bus. Manage.* 3, 3–11.
- Nutley, S.D., 1990. *Unconventional and community transport in the United Kingdom*. Gordon and Breach Science, London.
- Potts, J.F., Marshall, M.A., Crockett, E.C., Washington, J., 2010. *A guide for planning and operating flexible transportation services*, Report 140. Transit Cooperative Research Program, Transportation Research Board, Washington, DC.
- Rodrigue, J.P., 2017. *The geography of transport systems*. Routledge, New York.
- Stagecoach, 2015. *Bus industry frequently asked questions*. Stagecoach. Retrieved from <http://www.stagecoach.com/media/faqs/bus-industry-faqs.aspx> (December 15, 2015).

Further Reading

- Docherty, I., Marsden, G., Anable, J., 2018. The governance of smart mobility. *Transport. Res. A Policy Pract.* 115, 114–125.
- Lyons, G.D., 2001. Towards integrated traveller information. *Transport. Res. A* 34 (2001), 217–235, 2.

Transport and Climate Change

Robin Hickman, Christine Hannigan, Bartlett School of Planning, University College, London

© 2021 Elsevier Ltd. All rights reserved.

Introduction	94
Current Emissions From Transport	94
Sustainable Transport Amid a Growth Agenda	95
Visioning and Backcasting	96
Avoid-Shift-Improve	97
Technological Optimism	98
Regime and Structural Change	98
Conclusions: Improving Progress Toward Sustainable Mobility	99
Acknowledgments	99
See Also	99
References	99

Introduction

The climate change emergency is well-documented, even gaining widespread public attention as the ‘climate crisis’ is made more visible by activists such as Greta Thunberg (2019). We are aware that humanity faces an existential crisis in relation to climate change. Yet strategies to meaningfully reduce carbon dioxide (CO₂) emissions and limit environmental damage lack cohesion and political willpower. There are large problems in implementing appropriate projects, which significantly contribute to reducing CO₂ emissions, and this is particularly the case in the transport sector.

This chapter argues that current discourses and policy responses surrounding sustainable transport lack the boldness and urgency required to confront the climate emergency. Over reliance on technological solutions to decarbonize the transport network, amid premises of economic growth, fail to address the causal problems of increasing consumption and mobility and hence transport CO₂ emissions. Reducing transport-related CO₂ emissions demands radical cultural and behavioral changes at individual and societal levels—and this requires structural changes including political and cultural dimensions. The approaches used in transport planning can also be widened, to include a greater use of social science perspectives alongside the current strengths in quantitative, predictive analysis.

Current Emissions from Transport

During the last century, anthropogenic (human) activity has led to significant climate change, including surface temperature rises, over a relatively short time period (Intergovernmental Panel on Climate Change, 2014). Alongside there has been much intransigence and difficulty in changing policy approaches, project investments and ultimately in reducing CO₂ emissions.

Aggregate CO₂ emissions across all sectors, including industrial, residential and transport, are given in Fig. 1, using data from the Emission Database for Global Atmospheric Research (EDGAR). There are huge differences in CO₂ emissions by country: China is the largest aggregate emitter at over 10.5 GTCO₂ (7.6 tons CO₂ per capita), the US at 5.3 GTCO₂ (a higher 16.5 tons CO₂ per capita) – these two countries account for nearly 45% of global CO₂ emissions. There are other large emitters in India, Russia, Japan, Germany, Canada, Brazil, Mexico and the UK. Very large growth in emissions from 1990 to 2014 is seen in China, India and Brazil, with China increasing emissions by over 300% and India by 250% from 1990 to 2014 (Hickman et al., 2017). There are complexities in interpreting the data. For example, focusing on aggregate national emissions overlooks large individual and spatial differential within countries. In addition, allocating industrial emissions to producing countries rather than those consuming the products, and indeed ignoring historical emissions, means that the assumed accountability of industrialized countries is reduced.

Transport is a significant contributor to CO₂ emissions and as a sector is responsible for around 20% of all carbon emissions globally. Oil is the dominant fuel source for transport, with road transport accounting for over 80% of energy use—hence the use of fossil fuels makes transport a major contributor to climate change (Chapman, 2007).

Indeed, transport is the only sector increasing its CO₂ emissions over the past decades as any fuel efficiency gains are offset by increases in mobility (Hickman and Banister, 2014). Within OECD countries, transport accounts for 29% of all emissions. Countries differ quite markedly in percentage contribution of transport relative to all CO₂ emissions—the US at 33%, UK at 29%, India at 12% and China at 9% (World Bank, 2015). This reflects many factors, including the levels of mobility, the shape of the built environment, emissions from other sectors and power supplies.

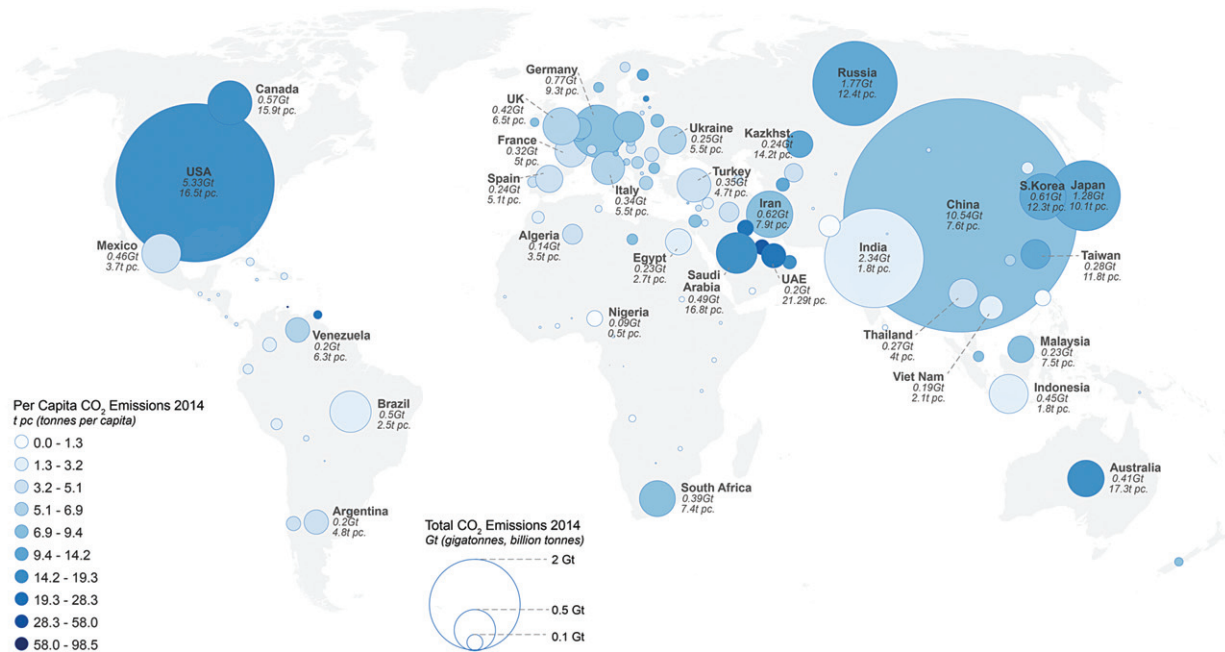


Figure 1 Global aggregate CO₂ emissions. Source: From Hickman et al. (2017), using EDGAR 2014 data.

At the aggregate level, the planet has already warmed 1.5°C above pre-industrial levels, causing “multiple observed changes in the climate system”. Even another half-degree increase will result in more extreme temperatures, floods, storms, precipitation, drought, which will in turn jeopardize water, food, and health security, disproportionately affecting the poor and vulnerable (Intergovernmental Panel on Climate and Change (IPCC), 2019). For example, the IPCC’s 2019 Special Report on Global Warming of 1.5°C warns:

“Such change would require the upscaling and acceleration of the implementation of far-reaching, multilevel, and cross-sectoral climate mitigation and addressing barriers. Such systematic change would need to be linked to complementary adaptation actions, including transformational adaptation . . . Current national pledges on mitigation and adaptation are not enough to stay below the Paris Agreement temperature limits and achieve its adaptation goals.”

It is still unclear how the transport sector can effectively respond to the climate change problems. There are few, if any, countries and cities with forward-looking strategies that demonstrate appropriate, quantified responses. The two most emitting modes (air and personal vehicles) are also the fastest-growing in use, and efficiency improvements in car technology are undermined by the increasing number and size of vehicles (Chapman, 2007). Indeed, significant streetspace reallocation, from the car to public transport, walking and cycling; and air travel demand management; are rarely discussed even in the ‘progressive’ transport planning cities. Consider, for example, London, a city with a well-known sustainable transport strategy (Transport for London, 2018), but which says little on progressive traffic space reallocation or longer distance inter-urban or international travel, with five major global airports in the city-region (Hickman et al., 2010).

Despite international bodies’ dire predictions and the escalation of climate activism in recent years, transport CO₂ emissions continue to move upwards. Achieving a global decrease of over two-thirds seems difficult to comprehend, and virtually impossible to achieve, given that the US, one of the planet’s most intense polluters, confirmed its withdrawal from the Paris Agreement by 2020. Even where cities and national governments introduce projects and initiatives to reduce CO₂ emissions, they are often offset by parallel increases in highway capacity. There is little analysis concerning how overall CO₂ emissions can be significantly reduced. Private industry, without financial incentives or legislative pressure, largely continues ‘business as usual’ operations, and travel behaviors remain unaffected.

Sustainable Transport Amid a Growth Agenda

The need to reduce CO₂ emissions—to consume, build and travel less—conflicts with economic growth agendas that promote the opposite of this. “Sustainable development” is typically defined in an abstract manner, without measurable indicators, providing room for interpretation and claims of sustainable development free of scrutiny (Castro, 2004; Swyngedouw, 2007).

The most commonly deployed definition, from the Brundtland Commission (World Commission on Environment and Development, 1987), is:

“Development that meets the needs of the present without compromising the ability of future generations to meet their own needs.”

However, as [Turcu \(2018\)](#) counters,

"It does not give an indication of what sustainability is precisely and what one may contemplate to measure. For example, what needs, what type of ability, which generations, and over what period of time?"

These issues remain unresolved in relation to sustainable mobility. The main application of sustainability in transport is through the conventional 'three pillars' concept that draws on environmental, social and economic dimensions. This assumes that equal weight is given to all pillars in the process and that a solution can be found that contributes well against each of the economic, social and environmental objectives. Yet, in practice, this rarely happens and the economic pillar is given much greater weight in the decision-making process ([Hickman, 2019](#)). All of this occurs within a growth paradigm, where transport investment is used to support urban development and economic growth. In the end, transport investment is sustaining growth rather than helping to achieve sustainable transport ([Hickman, 2017](#)).

The [OECD \(2000\)](#) attempt to give some definition to the process, defining environmentally sustainable transport as:

"Transport [that] does not endanger public health or ecosystems and meets needs for access consistent with (a) use of renewable resources below their rates of regenerations, and (b) use of non-renewable resources below the rates of development of renewable substitutes."

This aims to give greater weight and precision to the environmentally sustainable transport concept, using indicators that can be quantified. But, many countries and cities are working with only a fuzzy concept of sustainable mobility and this is interpreted in varied ways by different actors, becoming almost meaningless as a working concept.

Visioning and Backcasting

A visioning and backcasting approach can help to give a greater certainty to policy development and the direction of interventions. The concept of 'backcasting' was developed in the 1970s and 1980s to examine how different environmentally sustainable futures might be attained ([Lovins, 1977](#); [Robinson, 1982, 1990](#)). It was applied in the transport sector by authors such as the [OECD \(2000\)](#) to help understand what this might mean in terms of an environmentally sustainable transport system. The approach has since been developed and used in a range of transport studies, including in different contexts ([Åkerman and Höjer, 2006](#); [Geurs and Van Wee, 2000](#); [Hickman and Banister, 2007, 2014](#); [Hickman et al., 2010](#); [Soria-Lara and Banister, 2017](#), amongst others).

[Robinson \(2003\)](#) developed and pioneered the backcasting approach:

"Not to reveal what the future will likely be, but to indicate the relative feasibility and implications of different policy goals"

This is viewed in relation to varied and specific societal problems. Because of this, the various modeling approaches:

"Need to be able to show the implications of different user choices but not choose the most likely or optimal (e.g. least cost) solution. The resultant models act in part as accounting frameworks, which show the consequences of different behavioral choices, rather than predicting most likely outcomes" (ibid).

Backcasting is hence a method to analyze future options concerning long-term complex issues, involving many aspects of society as well as technological innovations and change ([Dreborg, 1996](#)). Rather than develop scenarios extending from the present-day, it considers pathways back from a future state ([Hickman and Banister, 2014](#)). The backcasting process is given in [Fig. 2](#).

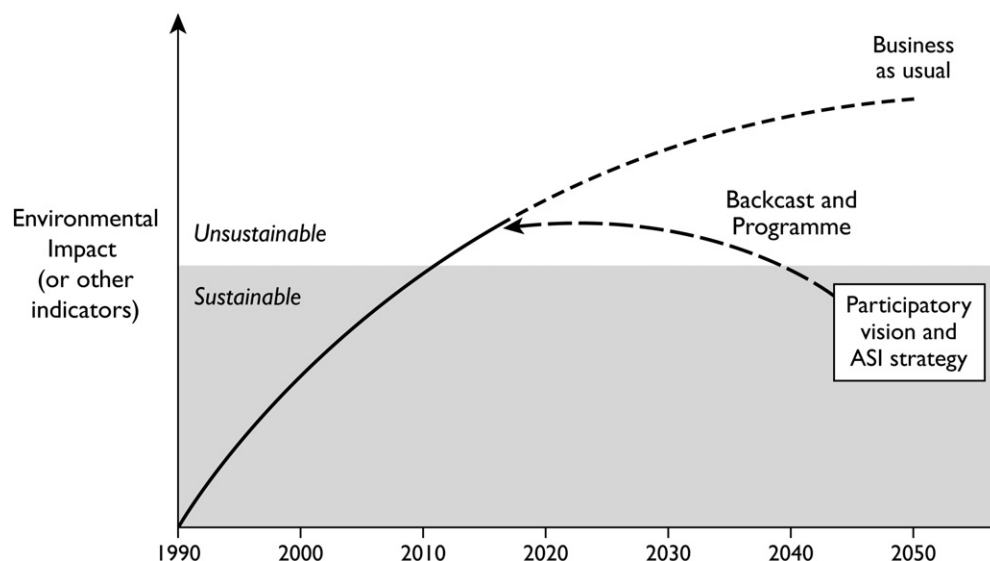


Figure 2 Visioning and backcasting. Source: [Hickman \(2020\)](#)

There are four key stages, all of which can be carried out in a participatory manner with local actors, stakeholders and even the public.

- Stage 1: Understanding of local problems and opportunities, BAU projections
- Stage 2: Development of vision for sustainable mobility, this may include key targets and an Avoid-Shift-Improve strategy (discussed below)
- Stage 3: 'Backcast' trajectory to give an implementation program for projects and initiatives
- Stage 4: Evaluation of progress and refinement of the vision, strategy and program

Although backcasting has been put forward as a useful process for implementing changes towards sustainable urban mobility (Hickman and Banister, 2014; Quist et al., 2011), its application still seems largely confined to academic research. There is little use in transport planning practice, for example in strategy development or appraising projects, which still seem entrenched in forecasting and 'predict and provide' analytical approaches.

The problem remains that there is too much motorized travel in many contexts—and forecasting-based approaches will struggle to break long-established trends. There needs to be a much greater effort to significantly increase the mode share of public transport, walking and cycling, across multiple cities and for travel between and beyond cities. There are progressive targets for achieving 80% or more of trips by noncar modes in selective cities, such as London (Transport for London, 2018)—but most urban areas do not demonstrate this awareness. Even for London, this will be difficult to achieve, particularly for travel in the outer suburbs. But, for wider cities, there are probably unsurmountable problems.

This is an exciting agenda for transport planning and transport planners—to provide transport systems that enable sustainable travel behaviors. However, it will be hugely difficult to retrofit existing dispersed cities with high quality public transport networks and sufficient densities to make public transport work at sufficient usage levels. Much of the public debate does not support the removal of traffic capacity and providing improved walking and cycling environments. This is why the visioning and backcasting approach can be very important: helping to discuss and agree future implementation pathways which may be very different to the business as usual trajectory in many contexts. For example, some of the North American dispersed cities have very poor public transport networks, and public transport, walking and cycling are used in less than 10% of trips. Hence there are different development pathways available to cities—but all must try to reduce private car usage due to the very large adverse impacts on society.

Avoid-Shift-Improve

The Avoid-Shift-Improve (ASI) framework (Dalkmann and Brannigan, 2007) offers a way forward for implementing sustainable mobility. It can be used to frame the strategy, which is constructed to implement the vision. Originating in Germany in the early 1990s, the ASI framework brought together the sustainable transport policies being developed at the time, and challenged the conventional "predict and provide" approaches, which often led to highway building. The ASI framework was embraced by International Non-Governmental Organisations (NGOs) as a wide-ranging approach for moving towards sustainable transport. It is similar in concept to the sustainable mobility paradigm (Banister, 2008), encouraging a range of interventions within a transport strategy (Fig. 3).

The measures within each theme can include (Hickman and Banister, 2019):

- Avoid: urban planning, affordable housing, development restraint in inaccessible locations, and reduced growth in international air;
- Shift: improved walking and cycling facilities and networks, multimodal public transport; use of urban highway re-prioritization, travel demand management, emissions-based road pricing, car parking supply restraint, Mobility as a Service, and urban logistics management;
- Improve: shared vehicle ownership, low emission vehicles and alternative fuels.

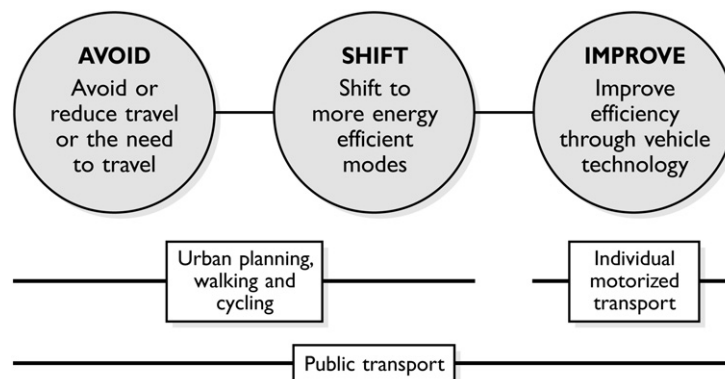


Figure 3 The ASI framework. Source: Based on Deutsche Gesellschaft für Internationale Zusammenarbeit GMBH (GIZ) (2019).

In some contexts there is progress being made with reduced levels of motorization and higher usage of public transport, walking and cycling. This has been described as the ‘peak car’ phenomenon (Goodwin and Van Dender, 2013; Millard-Ball and Schipper, 2010). A number of factors are attributed to this reduction in car use. These include the increasing cost of owning and maintaining a vehicle, including fuel; modifications to the built environment to de-prioritize cars over other modes of transport; urban densification; increased availability and lower (relative) prices of alternative long distance modes (rail, air); and a cultural change that devalues the status of the car. Many of these are impacts of the sustainable transport policies implemented over the last few decades.

Metz (2013) describes:

“A new era of travel . . . in which average per capita growth of daily travel has ceased”

This is attributed to a maturing of economies caused by a combination of access to more consumer choices, saturated markets and lessening demand for travel, combined with difficulties in traveling faster. It is difficult to travel greater distances if we end up spending too much time traveling. In addition, younger demographics are also eschewing becoming licenced drivers and car owner, and moving to urban centers where car ownership may not only be unnecessary but inconvenient. However, any peak car trend is limited to the Western industrialized countries, remains marginal, and is offset by very significant increases in motorization in the Global South countries. Sustainable urban mobility behaviors are yet to be conceived and implemented in many contexts.

Technological Optimism

A problem for many governments is that there is reliance on technological options to reduce transport CO₂ emissions, hoping that mobility levels and patterns can be maintained, but made cleaner (Hickman and Banister, 2014; Schwanen et al., 2011). As Lyons and Davidson (2016) remind us:

“Our fascination with technological advance, coupled with technological optimism, can cloud our ability to make sense of particular technological possibilities in terms of the extent to which they may become transformative and mainstream, and over what timescale”

New technologies are over-hyped, and particularly in transport. Often they are poorly delivered and any gains in emission reduction are offset by increased mobility. Similarly, Goulden et al. (2014) caution against technical solutions to solve the problem of high carbon mobility because of rebound effects, in which improvements in efficiency lead to increased demand and shape how efficiency is utilized. Rebounds (Sorrell, 2007) include improved fuel economy or increased comfort leading to longer distances traveled. The result is that CO₂ reductions are less than expected.

Regime and Structural Change

The Paris Agreement calls for regime changes, but most of the policies and plans used within transport planning are regime-compliant and not regime-changing.

“The reality on the ground in most countries . . . is that aspirations expressed in the Paris Agreement do not match-up with levels of action” (Ford et al., 2018).

“Business as usual” is likely to result in a 3°C increase in global surface temperatures, with uncertain catastrophic proportions. Regime compliancy is also embedded in the approaches and models used for transport and land use planning (Ford et al., 2018). Where drastic systemic change at institutional, societal, and individual levels is required, many governments instead prefer to implement voluntary incentives, or “nudges” that “create an illusion of choice, seeking to achieve collective goals without directly challenging the neoliberal paradigm” (Goulden et al., 2014).

Lyons and Davidson (2016) similarly question if planning and policy initiatives within the current global regime can lead to meaningful change. “A regime encapsulates the way of the world as we know it and within which the various actors engage.”

Transition theory perhaps offers a way forward in seeking to understand the different potential components of the transition to more sustainable behaviors, particularly when viewed through the multilevel perspective (Geels, 2012; Kemp et al., 2012; Schwanen et al., 2011). This includes niche, regime and landscape levels, all of which combine to influence travel behaviors. For example, the electrification of the vehicle fleet, or indeed other policy measures, can be viewed in terms of how the niche activity enters into the existing regime. The regime includes the key actors in the transport sector, such as the motor manufacturers, urban developers, bus and rail industry, government organizations, transport consultants, civil society and academia. All of these influence the transport system, shaping the infrastructure and network that is available. At the higher level are landscape factors, framing and influencing the regime and niche levels. These landscape factors may include climate change, socio-demographics and the political framework. This broader view of how change may occur helps us to understand that the use of infrastructure investment is also dependent on existing actor, network and power structures, the political and cultural context, and can change over time relative to changing knowledge and innovations. This is a more sophisticated understanding of potential change in travel behaviors and the factors that may be at play.

In recent years there has been progress made in drawing wider disciplines into transport research. Transport psychology has helped us to understand how individual attitudes may influence behaviors (Anable, 2005; Steg, 2005) alongside the transport infrastructure and built environment. This is complemented by an understanding of the wider sociology of travel, including social and cultural norms and political structures (Shove, 2014; Urry, 2006). Transport planning, in academic research at least, now draws on varied disciplines to help explain behaviors and the potential for change in terms of reducing transport CO₂ emissions

(Schwanen et al., 2011). This has been a great progression in helping to understand that enabling changes in travel and achieving reductions in transport CO₂ emissions are actually quite complex. This involves much more than building infrastructure, and potentially individual attitudes and societal structures also need to change. Building high quality public transport networks, improved cycling facilities and public space will help change individual attitudes and societal norms, but also the causality may be required and occur in reverse.

The current regime of motorization has arisen over the last century and has become entrenched in both its geographic reach and cultural symbolism (Lyons and Davidson, 2016). The car dominates the built environment and the fossil fuel industry is inordinately powerful in influencing transport policy to remain centered around the car. “Regime-compliant” policy pathways are often put forward to maintain the implicit reliance on motorization continuing uninterrupted. The conclusion is that the regime needs to change (Böhm et al., 2006) beyond the insignificant revisions usually made at the margins.

Conclusions: Improving Progress Toward Sustainable Mobility

The concentration of the global population into cities presents planners with the opportunity to influence travel behavior so that it can be much less carbon intensive. The projected urban growth can be the catalyst to reducing transport CO₂ emissions, as there will be high demand for new infrastructure and feasibly high proportions of the population will live close to public transport systems. But, implementing the ASI framework in multiple cities demands international cooperation, national and city leadership, and radical policy choices to be developed, agreed, and implemented.

Visioning and backcasting approaches offer a way forward for planning and policymaking, through which sustainable mobility futures can be discussed. Outdated “predict and provide” models should be replaced with scenario building and demand management. Forecasting-based approaches have become obsolete as they offer only marginal changes to the historical trajectories. Transport planning hence becomes a normative and participatory exercise and should be much less interested in predictive modeling and causality. Qualitative approaches, informed by the range of social science perspectives, can help strengthen the conventional use of quantitative analysis in transport planning. For example, we can ask what our desirable future cities and travel behaviors are—and seek to deliver these.

The ASI framework can be used to help define and apply a wide-ranging package of transport interventions within the strategy, including avoid, shift and improve measures. A greater radicalism is required in implementing public transport, walking and cycling projects, travel demand management measures and the reallocation of road space. There also needs to be a focus for transport planning beyond the infrastructure and built environment—the individual attitudes and societal structures need to be shaped to help enable sustainable travel behaviors. Without this, the radical and innovative projects cannot be implemented: the politicians and public will not accept them. There remain many current policy taboos (Gössling and Cohen, 2014), but these can be overcome as the level of public understanding of climate change and the potential solutions are discussed.

There is much discussion, in the transport industry and governmental organizations, concerning what we perceive as false promises of technical innovation to revolutionize the sector. The debate also needs to include reduced consumption and a more transparent understanding of sustainability itself. The major characteristic of backcasting is a focus on how desirable futures can be achieved. This can be the central approach to developing low CO₂ transport futures as it offers such a persuasive combination of vision development, definition of implementation pathway and delivery, all within a strengthened participatory and deliberative perspective. The climate crisis is now very evident—transport is the key sector that needs to improve its performance in reducing CO₂ emissions.

Acknowledgments

Thanks to Duncan Smith for the global aggregate CO₂ emissions map and Jamie Quinn for the remaining figures.

See Also

Land-use and Transport Planning; Evaluation Methods in Transport Policy and Planning; Transport and Air Quality Planning and Policy; Electric Mobility; The Politics of Mobility Policy; Energy and Transport Planning; Plateau Car

References

- Åkerman, J., Höjer, M., 2006. How much transport can the climate stand? Sweden on a sustainable path in 2050. *Energy Policy* 34, 1944–1957.
- Anable, J., 2005. ‘Complacent Car Addicts’ or ‘Aspiring Environmentalists’? Identifying travel behaviour segments using attitude theory. *Transport Policy* 12, 65–78.
- Banister, D., 2008. The sustainable mobility paradigm. *Transport Policy* 15, 73–80.
- Böhm, S., Jones, C., Land, C., Paterson, M., 2006. Part One: Conceptualizing automobility: introduction: impossibilities of automobility. *Sociol. Rev.* 54, 1–16.
- Castro, C., 2004. Sustainable development: mainstream and critical perspectives. *Organiz. Environ.* 17, 195–225.

- Chapman, L., 2007. Transport and climate change: a review. *J. Transport Geogr.* 15, 354–367.
- Dalkmann, H., Brannigan, C., 2007. Transport and Climate Change, Sourcebook Module 5e. Bonn: Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH.
- Deutsche Gesellschaft Für Internationale Zusammenarbeit GMBH (GIZ) 2019. TUMI Sustainable Urban Transport Graphic. Eschborn: GIZ.
- Dreborg, K.H., 1996. Essence of backcasting. *Futures* 28, 813–828.
- Ford, A., Dawson, R., Blythe, P., Barr, S., 2018. Land-use transport models for climate change mitigation and adaptation planning. *J. Transport Land Use* 11, 83–101.
- Geels, F.W., 2012. A socio-technical analysis of low-carbon transitions: introducing the multi-level perspective into transport studies. *J. Transport Geogr.* 24, 471–482.
- Geurs, K., Van Wee, B., 2000. Environmentally Sustainable Transport: Implementation and Impacts of Transport Scenarios for the Netherlands for 2030. National Institute for Public Health and the Environment.
- Goodwin, P., Van Dender, K., 2013. 'Peak Car' – Themes and Issues. *Transport Rev.* 33, 243–254.
- Gössling, S., Cohen, S., 2014. Why sustainable transport policies will fail: EU climate policy in the light of transport taboos. *J. Transport Geogr.* 39, 197–207.
- Goulden, M., Ryley, T., Dingwall, R., 2014. Beyond 'predict and provide': UK transport, the growth paradigm and climate change. *Transport Policy* 32, 139–147.
- Hickman, R., 2017. Sustainable travel or sustaining growth? In: Cowie, J., Ison, S. (Eds.), *The Routledge Handbook of Transport Economics*. Routledge, Abingdon.
- Hickman, R., 2019. The gentle tyranny of CBA in transport appraisal. In: Docherty, I., Shaw, J. (Eds.), *Transport Matters*. Policy Press, Bristol.
- Hickman, R., 2020. Transforming Urban Mobility: Introduction to Transport Planning for Sustainable Cities (online course). UCL, GIZ and Futurelearn, London.
- Hickman, R., Ashiru, O., Banister, D., 2010. Transport and climate change: simulating the options for carbon reduction in London. *Transport Policy* 17, 110–125.
- Hickman, R., Banister, D., 2007. Looking over the horizon: transport and reduced CO₂ emissions in the UK by 2030. *Transport Policy* 14, 377–387.
- Hickman, R., Banister, D., 2014. *Transportm Climate Change and the City*, Abingdon, Routledge.
- Hickman, R., Banister, D., 2019. Transport and the environment. In: Stanley, J., Hensher, D. (Eds.), *A Research Agenda for Transport Policy*. Edward Elgar, Cheltenham.
- Hickman, R., Smith, D., Moser, D., Schaufler, C., Vecia, G., 2017. Why the Automobile Has No Future: A Global Impact Analysis. Hamburg, Greenpeace Germany.
- Intergovernmental Panel on Climate Change, 2014. Climate Change 2014. Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva.
- Intergovernmental Panel on Climate Change (IPCC), 2019. Global Warming of 1.5 °C. An IPCC Special Report on the impacts of global warming of 1.5 °C above pre-industrial levels and related global greenhouse gas emission pathways. IPCC, Geneva.
- Kemp, R., Geels, F., Dudley, G., 2012. Sustainability transitions in the automobility regime and the need for a new perspective. In: Geels, F., Kemp, R., Dudley, G., Lyons, G. (Eds.), *Automobility in Transition? A Socio-Technical Analysis of Sustainable Transport*. Abingdon, Routledge.
- Lovins, A., 1977. *Soft Energy Paths: Toward a Durable Peace*. Ballinger Publishing, Cambridge, Mass.
- Lyons, G., Davidson, C., 2016. Guidance for transport planning and policymaking in the face of an uncertain future. *Transport. Res. Part A: Policy Practice* 88, 104–116.
- Metz, D., 2013. Peak car and beyond: The fourth era of travel. *Transport Rev.* 33, 255–270.
- Millard-Ball, A., Schipper, L., 2010. Are we reaching peak travel? Trends in passenger transport in eight industrialized countries. *Transport Rev.* 31, 357–378.
- Organisation For Economic Co-Operation Development, 2000. EST! Environmentally Sustainable Transport. Futures Strategies and Best Practice. Synthesis Report OECD, Paris.
- Quist, J., Thissen, W., Vergragt, P., 2011. The impact and spin off of participatory backcasting: from vision to niche. *Technol. Forecasting Social Change* 78, 883–897.
- Robinson, J., 1982. Energy backcasting: a proposed method of policy analysis. *Energy Policy* 10, 337–344.
- Robinson, J., 1990. Futures under glass: a recipe for people who hate to predict. *Futures* 22, 820–842.
- Robinson, J., 2003. Future subjunctive: backcasting as social learning. *Futures* 35, 839–856.
- Schwanen, T., Banister, D., Anable, J., 2011. Scientific research about climate change mitigation in transport: A critical review. *Transport. Res. Part A: Policy Practice* 45, 993–1006.
- Shove, E., 2014. Putting practice into policy: reconfiguring questions of consumption and climate change. *Contemporary Soc. Sci.* 9, 415–429.
- Soria-Lara, J., Banister, D., 2017. Participatory visioning in transport backcasting studies: methodological lessons from Andalusia (Spain). *J. Transport Geogr.* 58, 113–126.
- Sorrell, S., 2007. *The Rebound Effect*. Sussex Energy Group for UKERC, Oxford.
- Steg, L., 2005. Car use: lust and must. Instrumental, symbolic and affective motives for car use. *Transport. Res. Part A* 39, 147–162.
- Swyngedouw, E., 2007. Impossible "sustainability" and the post-political condition. In: Gibbs, D., Krueger, R. (Eds.), *The Sustainable Development Paradox*. Guilford Press, New York.
- Thunberg, G., 2019. *No One Is Too Small to Make a Difference*. Penguin, London.
- Transport For London, 2018. *Mayor's Transport Strategy*. GLA, TfL, London.
- Turcu, C., 2018. Sustainability indicators and certification schemes for the built environment. In: Bell, S., Morse, S. (Eds.), *Routledge Handbook of Sustainability Indicators*. Routledge, London.
- Urry, J., 2006. Inhabiting the car. *Sociol. Rev.* 54, 17–31.
- World Bank, 2015. *World Development Indicators, Country Data*. World Bank, Washington.
- World Commission on Environment and Development, 1987. *Our Common Future*. Oxford University Press, New York.

Taxicabs and Microtransit

David A. King, Arizona State University, Tempe, AZ, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	101
Taxi and Microtransit Regulation	101
History and Legality	101
Market Structure	102
Street Hail	102
Pre-Booked	102
Shared Vehicles	102
Microtransit	103
Regulatory Frameworks	103
Limits to Entry	103
Fares	103
Public Safety	103
Vehicle Requirements	104
Economics of Taxi Markets	104
Labor	104
Usage	104
Microtransit	105
Micromobility Markets	105
Mobility as a Service	105
Conclusions	105
References	105

Introduction

Taxicabs and microtransit are types of for-hire vehicle services that have received substantial attention by policymakers, planners and private investors since smart phones became ubiquitous, allowing for applications that connect riders and drivers with more ease and convenience than dispatch models traditionally used. The traditional taxicab market has been transformed by the development of ridehailing, creating expanded service and ridership for private trips. Microtransit maintains the on-demand aspects of taxicabs and ridehailing, but offers services with shared vehicles, from SUVs to small buses. There has been considerable interest by researchers and policymakers as to where expanded taxi and microtransit services fit within urban transportation systems, and whether they act as complements or substitutes for conventional transit.

Taxi and Microtransit Regulation

In most places, taxicabs are regulated by cities to manage fares, limit entry to market, ensure geographic reach, promote public safety and maintain other service characteristics. Microtransit is also locally regulated though with somewhat different requirements, depending on the city, state, or province. The for-hire vehicle industry has been subject to license requirements since its beginnings. The history of regulation includes cycles of strengthening and weakening regulations—the current ridehailing era is one of deregulation. Taxicabs are area of local transportation policy where there is substantial variation in regulatory approach across cities throughout the world. For both taxi services and microtransit, cities and transit agencies have explored some partnerships to improve service, though to date these have had mixed results for ridership costs.

History and Legality

Taxi services have been regulated since their invention. The cycle of regulation and deregulation is central to the history of these services. London and Paris were the first cities to regulate the taxi market, starting in the early 1600s. The cities required licenses for horse drawn hackneys and limited the total number of licenses issued. Such limits on the number of allowed vehicles have been a main feature of local regulations. While the original intent to ensure that operators met minimum quality and safety standards, limits to entry are problematic and lead to rent seeking and monopolistic behaviors.

There are many reasons why taxi services have been historically regulated, some of which can be assuaged through new technologies. First, unlimited entry leads to too many operators competing for too few fares, which depresses wages and is problematic for vehicle maintenance. Second, taxi services are credence goods, in that the riders do not know the correct value of the service. Regulating fares and service areas helps mitigate these concerns, and the smart phone apps of ridehailing and microtransit services can offer transparency in this regard. However, ridehailing firm Uber has also used their app to create “ghost cars” that give the impression of more service than exists, and drivers and passengers are shown different fares. This latter instance is a form of arbitrage, where Uber collects and keeps the difference between what drivers are willing to accept for a ride and what passengers are willing to pay. Third, for many reasons the operators of taxi services, and presumably microtransit, favor exclusionary regulation that keep profits flowing.

For hire services are regulated at local level in most cases. Cities, counties, and other regional governments design and enforce the regulatory environment. In the United States, states either allow cities to regulate taxicabs as part of their police powers, or in some instances states require that cities regulate the industry. There have been numerous court challenges questioning the legality of local taxi regulations but the abilities of cities to regulate the market has nearly always been upheld. The current era of new smart phone applications—discussed in more detail in a later section—represent a new line of legal inquiry that may yet prove the undoing of the historical regulatory structure, but early court cases from technology companies against the taxi industry and regulators present mixed success for the challenges.

Market Structure

Local control over taxi markets means that taxi markets are structured differently across cities. Microtransit is too new a service to have a clear sense of the regulatory structure. Hiring taxi services have historically been categorized as either street hail or dispatch models. In a few cities, such as New York, street hails are the only way to hire a Yellow Cab. In many other cities, street hails are illegal and taxicabs must be dispatched (or booked), yet other cities feature taxi stands, and often airports have their own systems for hiring cabs. There are also regulations about sharing taxis and where drivers are allowed to pick up passengers.

Street Hail

Street hails occur where taxi drivers cruise around in search of fares. Potential passengers flag—or hail—a car from the curb. Most cities prohibit street hails as a means of hiring a car as street hails increase the potential for discrimination based on geography or appearance. Street hails also require price regulation to eliminate fare negotiation at the point of hire. The largest justification for not allowing street hails is that taxi drivers will only cruise in the neighborhoods where they will be able to find an adequate number of fares. This means that taxis searching for street hails tend to cruise in a few parts of any city, usually in the Central Business District (CBD) or at the airports, meaning that much of the area that potentially can be served is not. More critically from a policy standpoint is that the concentration of taxi activity in a wealthy commercial area precludes taxi services in areas outside of the downtown.

Street hails are also potentially a problem with regard to discrimination based on what a passenger looks like. In New York City, the yellow cabs, which are limited to street hails by law, are required to take any fare for which they stop. Of course, this leaves open a great deal of interpretation. Stories of racial and ethnic discrimination are common where taxicabs simply would not stop for a hail. While the fares not picked up are hard to measure beyond anecdotes, more common are fares refused once a taxicab stops. Research on discrimination from ridehailing companies suggest that booking a ride through a smart phone app reduces discrimination of pick-ups ([Brown, 2019](#)).

Pre-booked

In contrast to hailing taxicabs many cities require pre-booking or dispatch models for service. Dispatch services work where the rider calls a central number, or in some cases the taxi driver directly, and a taxi is sent to pick them up. Pre-booked services also include cars that shuttle passengers to airports, limousines, and other commonly used services.

There are multiple justifications for pre-booked services. Cities that wish to prevent discrimination based on location or appearance seek to limit cab driver discretion in choosing fares. Absent dispatched services few taxis cabs would be found outside of central business districts or airports (or in some cases train stations). Beyond discrimination, pre-booked services prevent cruising for passengers, which is inefficient travel and wasteful for all involved outside of a few dense areas. Dispatched services are either operated as a first-call, first-served system where taxi cabs are sent as needed, or as a scheduled service.

Shared Vehicles

Most taxicabs are limited to transferring one set of passengers at a time. Perhaps a group of friends will share a ride and end up at multiple destinations, but multiple pick-ups are not allowed. There is a special group of taxi licenses that does allow for shared vehicles in some cities, and shared taxis are often encouraged in the wake of major disruption to conventional transit services. For instance, in the aftermath of Hurricane Sandy in October 2012, which knocked out parts of New York’s subway system for over a week, shared taxis and jitneys were indispensable for maintaining mobility within the city.

Many cities regulate how multiple passengers are treated by the fare. Some cities allow for a separate charge for each additional passenger while others require a single fare for each trip. For passengers, being able to split the fare is a way to reduce the cost of taxi rides and may induce additional taxi trip making. Splitting taxi fares is common when taxicabs are used for social occasions more than for commuting or airport transfers.

Microtransit

Microtransit is a separate but similar category from shared vehicles, and part of a broader category of Mobility on Demand (MOD). Microtransit is defined by the use of smart phone apps through which riders request rides from vehicles that range in size from large SUVs to small buses (Shaheen and Cohen, 2019). The services are multipassenger, and typically dynamically routed based on passenger demand, though in some instances there have been microtransit programs with designated stops. Designated stops increase travel time reliability and simplify routing, but decrease convenience to riders.

Regulatory Frameworks

Limits to Entry

License requirements that limit the total number of for hire vehicles allowed are very common in cities throughout the world. Ostensibly, regulators seek to determine the optimal number of taxicabs for perceived and real demand, then set the limits accordingly. Yet, in practice, it is extremely difficult to know how many taxicabs are actually required for any given demand, and by maintaining an artificial ceiling in the number of vehicles allowed actual demand is not revealed. Rather rider demand with regard to the constraints of the number of available taxis and the costs of the fares are revealed.

The iconic New York City yellow cabs, for instance, are limited to about 13,500 medallions. The Haas act of 1937 initially placed a cap on the number of allowed franchises in order to mitigate the negative effects of the “taxi wars.” Too many cars searching for fares reduced the earning potential of drivers where fares did not cover costs and the competition for passengers actually reduced service quality and made passengers and operators worse off. Under the medallion system, each licensed yellow cab is required to have a physical shield, about the size of a grapefruit, stamped on the hood of the vehicle. Medallions in New York City grant the right to the holder to cruise for hails anywhere in the city as well as protection from price competition (fares are set by the Taxi and Limousine Commission of New York City). This experience is typical for cities around the world, and there have long been ebbs and flows about the proper number of medallions. In the United States, however, there has been no correlation between the number of taxi licenses and population size (King et al., 2012).

Fares

Taxi fares are regulated to ensure adequate earnings to cover wages and operating costs as well as to protect passengers from fraudulent behavior. Taxi services are a well-used example of credence goods. Credence goods are economic goods with asymmetric information and potential for fraud. Auto mechanics and physicians are also common examples. Credence goods rely on the seller to be credible as the buyer has little way of knowing the true value of the good or service being providing. While everybody knows that they should receive a second opinion when a doctor provides a diagnosis, soliciting competing bids for taxi travel is impractical for all parties. Regulated fares avoid the problems with credence good by eliminated the need for negotiation. This not only serves the customers well, it speeds up the process of hiring a cab since the amount to be paid is not haggled over on each trip.

Part of fare policy is the fare media used for payments. While taxicabs now nearly universally accept credit cards, cash is allowed. For ridehailing and microtransit, payment is typically through the app, which requires access to a credit card. Access to credit cards is problematic for households without bank accounts, who generally pay for taxis and transit with cash (King and Saldarriaga, 2017). Unbanked households may be low income and unbanked by choice, or perhaps undocumented immigrants who are unable to open an account. As new technologies foster payments through credit cards, access to such cards becomes an issue of equity.

There are also issues of integration between taxi, ridehailing and microtransit apps with public transit and mapping apps. Mapping services such as Google Maps seek to provide complete information for travel choices when queried for directions. Transit agencies also want to act as a clearinghouse for all transport options. Some service providers seek to standardize transactional data specifications so that on-demand trip requests can be traded among service providers similar to how airline booking works (Larsen et al., 2019). Allowing for trip trading could foster many firms servicing trip requests and improve overall service, but at the expense of individual firms keeping their clients within their own system.

Public Safety

Public safety is the key consideration for regulators. Many transportation scholars and economists argue that taxicab regulations should be limited to those that ensure public safety, but even with safety there is substantial disagreement as to what is required. The bare minimum regulation requires that all taxicabs and drivers have adequate insurance and licenses. With the apps that new firms use to match drivers and passengers, concerns about personal safety have been minimized through the argument that drivers and passengers are known to each other and the companies, which is a deterrent to harassment or crime. There are reported stories that

dispute this, and while overall taxi and microtransit services are safe, there are real concerns about behavior. This is especially true for shared rides, where the drivers may be required to keep the peace.

Another aspect of safety is shift lengths. With conventional taxis, shifts are fairly well defined. Where the driver is working their own schedule through an app, it is difficult to assess how long they have been on the road. Most driving occupations have limits on consecutive hours a driver can work. With gig work, where drivers may turn the apps on and off per their own preferences, these requirements are difficult to enforce. Maintaining healthy drivers who operate safely is an area of interest for regulators.

Vehicle Requirements

New York City is famous for yellow cabs, and London is equally famous for black cars. These are iconic aspects of the cities by design. Cities recognize the value of distinctive colors and shapes of valuable taxi services. New York's yellow cabs have inspired cities across the globe to adopt similar shades of yellow to mark their official taxicabs. By signaling through vehicle standardization, passengers are assured of a safe trip for the stated fare.

Vehicle age is also regulated. In New York City, taxicabs must be replaced after three years. Prior to Chicago changing its regulations, many old New York taxis became new Chicago taxis. (The market for used New York City taxis also partially explains why other cities have many yellow taxis.) While in service most cities require regular inspections of the vehicles to ensure safety.

Less common than vehicle age, color, or propulsion are strict requirements for the specific make of taxicab. Most cities have guidelines for what cars can be used, places such as London and New York go further in requiring a particular model of car. Uber and other ridehailing firms have some standards about what types of cars can be used by drivers on their platforms, though these vary by location.

Economics of Taxi Markets

Labor

Labor concerns are prevalent in taxicab regulation. Unregulated taxi markets are subject to intense competition that drives down costs and profits. Unfortunately for taxi and ridehailing drivers, their wages are the only part of the cost of doing business that are not fixed. Cars, insurance and gas are all exogenously priced and there is little ability to reduce these costs. Wages, however, can get squeezed to the point where taxi driving is no longer sustainable. While this is exactly how the market economy works, the challenge for taxicabs is the process of reaching equilibrium. Passengers who depend on taxis and cities that rely on taxis to complement their transit systems cannot afford to go through periods of over and under supply of taxis. This is one reason why cities attempt to limit the number of licenses.

Currently, the employment model of ridehailing companies (e.g., Uber) is under legal challenges in cities throughout the world. Many cities, states, and provinces have sued to force the classification of drivers as employees rather than independent contractors. The outcome of these legal challenges is unknown at the time of this writing, so details will be omitted. However, classification of drivers as employees would substantially increase the costs of business for these companies, and would affect the overall market for these services.

For microtransit, labor may be from direct employees or the driver is paid as a percentage of fares. Major ridehailing firms have developed pooled ride services where drivers are paid a share of each fare. Other private transit providers, such as Bridj or Chariot, have developed services where drivers are paid as employees. None of these attempts at private microtransit have proved widely successful, and the private transit operators have either gone out of business or only license their software at this point. In some places, transit agencies have been experimenting with microtransit as part of their overall services. The Kansas City Area Transportation Authority has been experimenting with microtransit since 2016, and currently offers a mix of on-demand services that include taxis, ridehailing and microtransit.

Usage

The advent of ridehailing services such as Uber and Lyft have expanded for-hire vehicle travel beyond traditional taxi markets. The 2017 National Household Travel Survey (NHTS) was the first large travel survey to ask about these new services. Since ridehailing firms are private, getting reliable data about usage has been challenging. From the introduction of Uber in 2009 to 2017, the for-hire vehicle market share doubled in the United States. For-hire vehicles might account for only 0.5% of all trips, but the percentage of all Americans who use ridehailing in any given month is nearly 10%. The growth trend has been greater in mid-sized and large cities, and among younger individuals and wealthier households (Conway et al., 2018).

There is substantial interest in using taxis, ridehailing and microtransit as a first mile/last mile service to bring potential riders to transit stations. While there have been many highly publicized pilot projects testing this idea, the evidence that these services are used for first mile access is weak (King et al., 2020). At issue is these services are a premium service, with high marginal costs through fares. As transit riders are sensitive to fare costs, additional costs of a taxi or microtransit fare inhibits transit use. In addition, first mile access represents a transfer, and how well such transfers are timed will affect use. Without subsidy to the user, there is little reason to think that these services will be widely used for first mile access to transit.

Microtransit

A lot of the interest in microtransit stems from historical observations about the role of jitneys in urban transport. Jitneys have long operated in the space between sanctioned public transportation and the private automobile. Jitneys take many forms: vans, cars, shared-taxis, minibuses, buses, and even motorcycles; and operated in numerous ways: semi-fixed routes, fixed routes with variable stops, and complete passenger control over pick-ups and drop-offs. In many cities, such as New York and Miami, modern day jitneys (known as Dollar Vans in New York) serve largely immigrant communities. Yet the jump from informal transit to formal systems is difficult and often fails (King and Goldwyn, 2014).

Micromobility Markets

Micromobility is a transit service that is ripe with promise but not yet with a clear market. As most entrants into microtransit have been private firms with the hopes of finding a path to profitability, most microtransit initiatives have proven more challenging than expected. Early private leaders in generating microtransit interest have gone out of business. When the Kansas City Area Transportation Authority piloted a partnership with Bridj in 2016, they served less than 1500 people in the first year. The challenge for microtransit is that it is very difficult to scale the business to something that can be profitable in many markets. The early success of companies like Chariot was that they kept their operations small, limited to a few routes. This limited market for microtransit certainly exists, but there is not yet evidence that microtransit can expand the transit market to the degree that ridehailing has expanded the taxi market.

Mobility as a Service

One aspect of microtransit that has received substantial attention is the idea of Mobility-as-a-Service (MaaS). In this form, microtransit is one part of the suite of services that a MaaS operator might offer. The appeal of MaaS is that by combining multiple potential modes of travel, riders are offered a suite of services that collectively offer a substitute to the utility of cars. There have been limited interventions of MaaS, mostly in Europe, and while popular with small groups of people these apps and services have not yet gained widespread use (Zipper, 2020).

Ultimately, the challenges facing taxis, ridehailing and microtransit come down to finding a business model that is profitable. Microtransit and ridehailing are yet to find a profitable market-based approach to new services. Partnerships with cities and transit agencies remain a promising way forward to improve services and reduce public costs, but this is a substantially different approach than what has been tried to date.

Conclusions

Taxicabs are an important urban transport mode that has received new attention with the advent of smart phone apps. Also due to smart phones and new location technologies, microtransit has garnered interest as modern jitneys. While taxis and ridehailing have expanded the user base over the past decade, experiments with microtransit have been inconclusive for their overall impact.

As a locally regulated transportation system that requires very little public outlay to develop, taxicabs and microtransit represent an important opportunity to improve transportation choices for all people. As researchers work with the wealth of new digital data from these services, patterns of usage and demand are being revealed that help policymakers engage with urban transport planning and policy with better knowledge and improved understanding of how on-demand mobility options help people navigate the city.

Regulators must deal with economic concerns of rent seeking and asymmetric information, plus complicated issues surrounding labor. As private operators, taxi drivers typically earn less than unionized transit workers and have limited ability to raise fares. Yet in many markets, taxicabs are a useful stepping stone for immigrants and low skill workers. Regulators must balance the needs of the drivers and operators with the broader public.

Lastly, taxicabs and microtransit are unlikely to succeed without public assistance, either through direct subsidies or regulations. Regulations can limit competition and ensure profitability. Direct subsidies can help travelers who need new options for daily travel live without a car, or can help transport vulnerable people with high quality service in the event that they need it. Fixed-route transit is critical for cities, but not the appropriate response to many transportation problems. Taxis and microtransit can potentially help fill these gaps, but are unlikely to do so if they are developed solely as private services that are limited in scale and scope.

References

- Brown, Anne E., 2019. Prevalence and mechanisms of discrimination: evidence from the ride-hail and taxi industries. *J. Plan. Educ. Res.*, 0739456X19871687..
- Conway, Matthew, Deborah, Salon, David, King, 2018. Trends in taxi use and the advent of Ridehailing 1995–2017 evidence from the US National Household Travel Survey. *Urban Sci.* 2 (3.), doi:10.3390/urbansci2030079.
- King, D.A., Matthew Wigginton, Conway, Deborah, Salon, 2020. Do for-hire vehicles provide first mile/last mile access to transit? *Transport Findings*, doi:10.32866/001c.12872.
- King, David A., Goldwyn, Eric, 2014. Why do regulated jitney services often fail? Evidence from the New York City group ride vehicle project. *Transport Policy* 35, 186–192.

- King, David A., Jonathan R. Peters, Matthew W. Daus 2012. "Taxicabs for Improved Urban Mobility: Are We Missing an Opportunity?".
- King David, A., Saldarriaga, Juan Francisco, 2017. Access to taxicabs for unbanked households: an exploratory analysis in New York City. *J. Public Transport.* 20 (1), 1.
- Larsen, Niels, Roger Teal, David King, Candace Brakewood, 2019. "Development of a Transactional Data Standard for Demand Responsive Transportation." TCRP Research Report 210. Washington, D.C. <http://worldcat.org/issn/25723782>.
- Shaheen, Susan, Cohen, Adam, 2019. Shared ride services in North America: definitions, impacts, and the future of pooling. *Transport Rev.* 39 (4), 427–442, doi:10.1080/01441647.2018.1497728.
- Zipper, David, 2020. The problem with 'Mobility as a Service.'. *CiyLab* 5,. <https://www.bloomberg.com/news/articles/2020-08-05/the-struggle-to-make-mobility-as-a-service-make-money>.

Land-use and Transport Planning

Luis A. Guzman, Universidad de los Andes, Department of Civil and Environmental Engineering, Bogotá, Colombia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	107
Basic Concepts	108
Integrating the Land-Use and Transport Systems	108
Future Land-Use and Transport Challenges	109
Implications for Urban Planning Research and Practice	110
Conclusion	111
Acknowledgment	111
Biography	111
See Also	111
References	112
Further Reading	112

Introduction

Mobility is a necessity and enables people to take advantage of the city amenities, thereby fulfilling a variety of needs (Bertolini, 2012). However, mobility can also have negative effects, such as congestion, traffic injuries, urban sprawl, air pollution and climate change. Traffic congestion usually appears when many people want to go to the same destination, at the same time and generally using less sustainable transport modes, such as cars. Those effects, known as transport externalities, have a negative influence on the quality of urban life. When talking about quality of life in cities, the most common dark spot is traffic congestion and the amount of time spent in traffic. Consequently, in urban transport, it is not the travel itself that is important to people but rather what travel allows them to do. Human activities respond to a variety of needs and wishes: residing, working, studying, enjoyment, social interaction, and also mobility. The spatial location of these activities will define travel patterns, generating/ attracting more or less trips, long or short, of people and goods. The number of trips is related to the household characteristics and land-use configurations, and its location within the urban area. The activities need different mobility levels and its relationship with transport is reflected in its spatial organization. In turn, the spatial interaction depends on the intensity of movements between the locations where activities take place.

Economic activities are produced, managed or administrated by companies, which are at different locations. In the same way, residential activities are produced by households. The spatial organization allows companies and households to make use of the differences between spatial locations to satisfy a variety of performance needs. For example, a company might need a lot of land and labor, so it would look for locations where the land is cheaper. Another company might concentrate high-quality facilities and qualified employees. Regarding households, decisions with respect to location of residence are a function of a wide range of attributes, valued in different ways according to household characteristics, where usually a trade-off between housing and transport costs is present.

However, all cities are different and within each city there are differences. There are rich and poor cities, with high or low accessibility levels, with different needs, where there are producers and consumers, jobs and households. The fragmentation of production and consumption, and the specific location characteristics of activities, generate large flows of people, goods and information.

Although each city has different mobility needs and uses different ways of satisfying them, in general all cities have something in common: activity concentration areas that produce/attract the most trips, are the most important factors that make up the organization of urban space. Then, the development of cities is conditioned by land-use and transport systems, from public transport to cars; the transport system has contributed to the creation of land-use patterns. Moreover, at the same time, the activity types condition the travel behavior and the transport system use.

Therefore, a balance of activity location, vocation and intensity is necessary for maintaining the correct functioning of the urban environment. To create or improve a livable urban environment, the requirements of human activities have to be balanced against each other. These requirements and decisions are directly linked to mobility. One of those effects is the modal share. This will depend on the distance, cost, travel purpose, and also, the travelers/employers characteristics. Generally, travel costs and travel purposes depend on land-use patterns. This relationship, where trip and location decisions co-determine each other, reveals the need for coordinated land-use and transport planning. There is a growing interest in the connection between land-use and transport all around the world. This is an essential component in transport policy and planning, aimed at creating more sustainable development across urban areas in the world.

Basic Concepts

Conventional transport planning gave way to the integration of land-use and transport planning, based on the principle that mobility needs should be met considering the land-use configuration, in order to reduce travel distances and offer better options for use of sustainable transport modes. Land-use and transport planning is a complex task since it covers various aspects related to changes in travel demand and land-use patterns at different scales (time and space).

Land-use and transport planning is an interdisciplinary field that works under a framework that includes political, socioeconomic, cultural and technological changes, where many concepts are used. Land-use encloses several subsystems such as physical infrastructure, workplaces, residences, land as well as the outcome of a complex urban real-estate market. Therefore, land-use planning serves the process of balancing competing demands on scarce urban land. On the other hand, transport, in addition to facilities, includes road capacity, public transport services, frequency and reliability, travel time and costs, and the resulting flows of people and goods. Transport planning is a process to design and improve the provision, operation and management of facilities and services for the transport modes to achieve sustainable mobility.

The theoretical impacts of land-uses on travel patterns basically depend on three principal dimensions, known as 3Ds: density (population, jobs), diversity (land-use mix), and urban/neighborhood/street design (Cervero and Kockelman, 1997). Urban areas with high densities, shops and services, well mixed land-uses, and nonmotorized-friendly designs encourage the use of sustainable transport modes. On the other hand, the impact of transport on land-use is measured in the accessibility of a zone: high accessibility levels increases the attractiveness of a location, which in turn, encourages urban development. Therefore, the land-use developments basically depend on accessibility, which is given by the transport system. Focusing on passenger transport, accessibility can be defined as the extent to which land-use and transport system enable people to reach activities (destinations) by a combination of transport modes (Geurs and van Wee, 2004). This approach shows there is a relationship between the land-use and transport systems.

Beyond land-use and transport planning as isolated disciplines, this integration has a large potential to influence mode choice and reduce travel demand on motorized transport modes, and is a key element to achieve urban sustainability. Within an institutional framework, laws, and regulations, land-use and transport planning is a field of research and practice that supports and informs decision-making and policy evaluation, activity location and accessibility improvements, which can (should) change over time.

Integrating the Land-Use and Transport Systems

The integration of planning is important because urban land-use and transport are directly related as was mentioned before. In addition, the impacts of transport on land-use happen at different time scales, and are measured through changes in the accessibility of a location. When land-use changes (function, intensity), there are also changes in human activities and in passenger and good flows, which also change accessibility levels. Accessibility is shaped by transport system infrastructure (supply, capacity and connectivity), which is not uniform across the territory.

For instance, the construction of transport infrastructure (public transport station, intermodal transport hub, road interchange) will generate an improvement in the accessibility, modifying travel patterns. This will also favor the concentration of residential or economic activities, which will generate new travel demand, and probably a change in land prices. This change in the urban structure depends on how much accessibility is improved, the increase in the supply of services and goods given by the transport improvement, the urban regulation, and of course, the local economic conditions. These processes will favor the location of new activities and a reorganization of the urban spatial structure.

Accordingly, land-use and transport planning are the main components of a dynamic feedback system and must be a coordinated process. This concept can be summarized as follows (Wegener and Fuerst, 2004):

- The land-use diversity (residential, services, commerce, industrial) over the urban area determines the locations of human activities such as residing, studying, leisure, or working.
- The distribution of the different human activities in space requires spatial interactions between zones in the transport system to overcome the distance between trip origins and destinations.
- The distribution of infrastructure in the transport system creates opportunities for spatial interactions and can be measured in terms of accessibility.
- The distribution of accessibility in space codetermines location decisions. This will result in changes of the land-use system. The cycle continues.

This feedback cycle is explained by the solid arrows in Fig. 1. As shown, the transport system has a direct effect on accessibility, which is not uniform. Then, accessibility slowly influences land-use patterns: the construction and development of workplaces and housing increases gradually and depends on other variables such as the economic and regulatory context. In the same way, the effects of human activities on transport infrastructure is quite slow, since transport networks are the most permanent elements of the physical structure of cities. Finally, the location of human activities in the land-use system gives rise to a demand for spatial interaction in the form of people or goods transport. These interactions can adjust in a matter of days or hours due to congestion or travel demand changes.

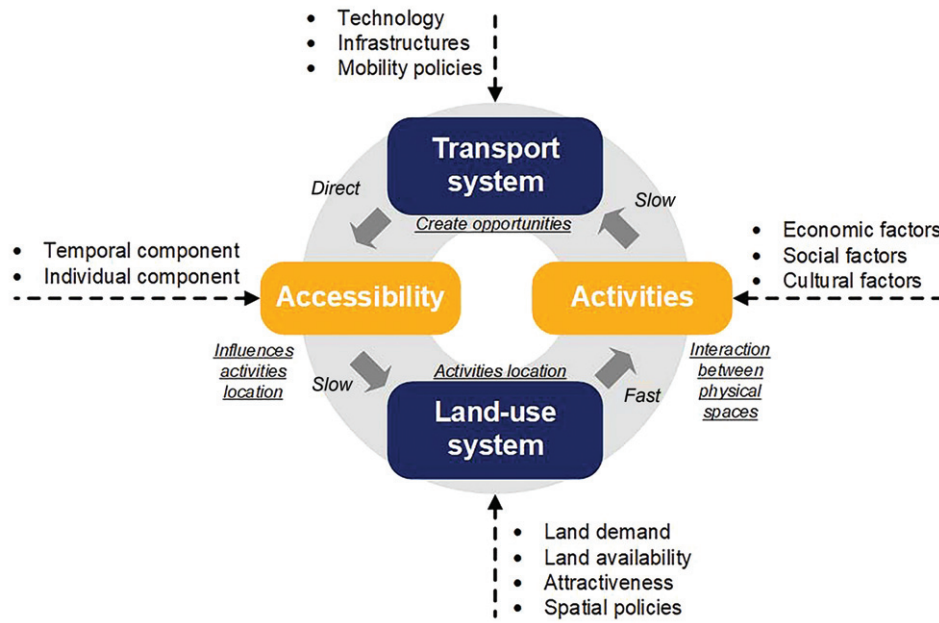


Figure 1 Land-use and transport feedback cycle. Source: adapted from Wegener and Fuerst (2004) and Bertolini (2012).

This reciprocal relationship is co-determined by external factors (dashed arrows in Fig. 1). For instance, the development of transport systems is not only determined by the travel demand but also by regulations, funding, technological innovation, and mobility policies. Land-use developments depend also on the city regulations, availability of land, real-estate market, environmental conditions and land-use policies. In addition, preferences for households and companies play an important role in the urban development process. Finally, besides land-use and transport systems, temporal component of accessibility is a key element for people since it includes the availability of opportunities at different times of the day, and the time available for travelers to participate in certain activities. The individual component of accessibility, reflects the needs (sociodemographic characteristics), abilities (physical condition, transport supply), opportunities (socioeconomic characteristics) (Geurs and van Wee, 2004), and capabilities of people (Cao and Hickman, 2019).

Although this definition is simple and easy to understand, the mechanisms through which the feedback cycle interacts between systems has been difficult to isolate and measure empirically, mainly because the complex interactions among several forces beyond the observed structure of the land-use and transport systems (Acheampong and Silva, 2015). Despite these important nuances, the land-use and transport feedback cycle provides a useful framework for planning and evaluates principles aiming at supporting sustainable transport modes at different levels. Sustainable land-use and transport planning require a deep knowledge of the urban and regional area, and a clear idea of the policy objectives.

Future Land-Use and Transport Challenges

Although several studies have concluded that urban structural variables (e.g., 3 Ds) have statistically significant influence on travel behavior, the results of other studies of different cities are sometimes inconsistent (Zhang et al., 2019). Despite this, there is still a consensus about the influence of activity locations, densities and the land-use mix on transport demand and travel patterns. These discrepancies may be due to diversity of urban environments and cultures, and should be studied in detail. The challenge is also to reconcile the short-term effects and long-term ones (Kasraian et al., 2016).

Land-use and transport planning is essential for well-informed decision making processes in urban environments. However, to help local authorities build appropriate plans that consider this dynamic feedback, and therefore improve policies' effects on city livability and reach sustainability goals, land-use and transport planning must try to bridge the gap between research and practice. First, city plans reflect a concern with accessibility in one way or another, but few explicitly articulate the relationship between land-use and transport systems over time and its side effects. Second, the proliferation of activity-based travel demand models, many times using big data sources, sometimes present difficulties in their integration with land-use models at disaggregated level. Third, the research and practice are different according to the region. The urban challenges of the cities in the Global North are considerably different from those in the Global South. Here is where most of the world's population reside, and cities in the Global South are characterized by rapid growth in population, motorization and urban footprint, without proper planning and control instruments. Integrated planning in these cities is crucial, but also the data collection and the development of evaluation tools. Therefore, operational land-use transport interaction (LUTI) models that recognize data limitations should be implemented, particularly in

Global South cities, mitigating existing knowledge gaps to improve urban planning in practice. Moreover, due to the high cost of urban infrastructure, it is necessary to incorporate concepts, methodologies and instruments of land value capture and pricing measures for private transport.

In addition, the current capabilities of the current models need to integrate complementary analysis, such as energy consumption, environment, health, or equity. With respect to integrating forecasting of possible impact of urban policies in the future, it is necessary to incorporate uncertainty explicitly into the land use and transport planning process. Cities change, and when the decision making involves long term decisions, the level of uncertainty, can be particularly high. Further research is needed to understand uncertainty over time and across different contexts and to develop and apply new methodologies to handle these challenges. Urban policies may also affect other levels, such as spatial and economic spillovers. Despite the great advances in recent years, big challenges still remain for land-use and transport modeling and their application (Moeckel et al., 2018).

In respect to the analysis of planning results, many tools/process can favor solutions that produce the greatest aggregate social benefit. This can hide large inequalities within local territories. A great mobility system is not sufficient to achieve sustainable development, including the reduction of inequality gaps. Sustainable long-term urban planning is essential, balancing accessibility and location with even distributions.

Finally, it is necessary to overcome spatial barriers for social, education and economic activities through an inclusive transport system. An initial step must be education: if students are taught to think on land-use and transport issues as an inseparable bundle with a dynamic relationship, mobility and land-use development issues will be one and the same.

Implications for Urban Planning Research and Practice

The landscape imposes certain conditions on urban development and therefore on the activities organization. As seen, the transport system influences the land uses and in turn, the land uses influence travel patterns. The combination of activities and accessibility constitute the local attractiveness of an area, which in the long-term determines the land use pattern and ultimately its urban form. This two-way interaction can be quantified from the perspective of location and mobility. The spatial location of the transport system is a central component of the definition of the spatial structure of a city. This system will allow activities to be carried out more or less easily (through improved accessibility). The scales of analysis, whether global, regional or local cannot be disregarded.

Reflections about the limitations of traditional approaches to urban planning, as isolated disciplines, are essential in the context of global development agendas that increasingly recognize the social implications for land-use and transport planning and its interaction with other areas of planning. Here, it is quite important to understand the importance and the changing role of urban transport in the modern world as a policy that can enable or hinder access to opportunities, services and goods that help build human, financial and social capital. That is, the integration of land use and transport planning is a key factor to improve the urban quality of life.

Consequently, land use and transport planning should aim to reduce the rate of growth of individual and motorized trips, supporting and encouraging use of public transport, walking and cycling, and balancing land use distribution. In order to achieve sustainable urban development, land use and transport planning goals have to be in line with the global challenges such as those related to poverty, inequality, climate change, environmental and social justice. Then, the following principles for successful land use and transport policies can be summarized in (Wegener and Fuerst, 2004):

1. Basic criteria of sustainable mobility and land use are only successful if the use of the private vehicle (car/motorcycle) is less attractive; that is, more slower or expensive (in the short term). Assuming that driving must be cheap is incorrect.
2. Land use policies to encourage the 3Ds without ancillary measures that make private vehicle use less attractive, will have a limited effect as people will continue to make longer trips to maximize their opportunities within their travel time and travel costs budgets. However, these policies are important in the long term as they provide the future conditions for a less private transport dependent urban lifestyle.
3. Pricing measures aiming to make private vehicles less attractive are very effective in the short term to reduce distances and travel times, as well as their modal share. However, effectiveness depends on a spatial structure that is not very dispersed.
4. Fears that land use and transport measures oriented to limiting private vehicles use in urban centers (including pedestrianization) are detrimental to the economic viability of these areas are baseless.
5. Policies aiming to improve public transport and encourage its use, in general do not lead to a significant reduction in private vehicle use, unless they are accompanied by other complementary measures.

Considering accessibility as a structural axis in the discussion about urban policies can enable integral land use and transport planning and evaluation of urban systems. This is possible as integrated planning forces planners, practitioners and decision makers to consider the interaction between its different components and highlights conditions of vulnerability and inequality within the territory. Furthermore, an integral vision of this relationship in its different scales and temporal dynamism can account for possible side effects of urban policies. For instance, if current practices in land use and transport planning are greatly improved, incentives for urban expansion may not be generated in the long term, and would lead to a potential balance in urban development. Therefore, if the ultimate goal of integrated land use and transport planning is constructing more sustainable cities, each strategy and instrument must contribute to one or more of the key policies, as shown in **Table 1**. Here, a high-level summary is provided, suggesting impacts

Table 1 A strategy contribution matrix

Key strategy element	Reducing the need for travel	Reducing car use	Improving public transport	Improving road network performance
Land-use	4	2	2	1
Infrastructure	–	2	5	3
Pricing	1	5	3	1
Attitudes	2	4	3	3
Management	1	2	4	5
Information	3	2	–	–

Minor contribution = 1; Major contribution = 5.

Source: May et al. (2005)

in which the different types of strategies can contribute to those in different categories. For instance, land use measures contribute most to reducing the overall need to travel in the long term, but pricing measures are the most effective way of reducing the level of private transport use in the short term. Other strategies can help to improve public transport and road network performance. The most important message is that there is no unique solution to urban problems: an efficient land use and transport planning practice must involve measures from many disciplines and sectors.

Integrated land use and transport policies have been promoted as a more realistic and effective approach to solving urban transport problems, rather than using isolated measures. With the integration of those two systems it will be easy to achieve greater performance from the overall urban policy. To make this a reality, the integration of different policy instruments involving land use planning measures, infrastructure provision, information, management and pricing must be taken into account. It is also necessary to integrate other policy sectors, such as environment and human health. The drawback is to understand how different packages of measures interact with each other: if they are a simple sum of effects, or there are synergies of some kind; and then identify the most successful combinations (May et al., 2006). This mix of effects has an impact on the spatial organization of cities over time.

Conclusion

The growing urbanization across the world has demonstrated that the urban environment and its land use and transport planning policies should be regarded as a crucial research topic more than ever before. When comparing land use and transport measures, latter measures are more direct and effective in achieving sustainable urban transport goals in the short term. Nevertheless, a support for long-term land use measures is essential for the development of more inclusive, safe, livable and less car dependent cities. In summary, integrated land use and transport planning, integrated with other relevant disciplines, are essential to achieve sustainable urban development.

Acknowledgment

Acknowledge any kind of support and/or financial assistance here when applicable.

Biography

Luis A. Guzman is a professor at the School of Engineering at Universidad de los Andes (Bogotá, Colombia). He holds an M.Sc from the Universidad de los Andes, and a Ph.D. cum laude in Systems of Civil Engineering, Urban and Transport Planning at the Universidad Politécnica de Madrid. His research interests include urban mobility, transport and land-use interaction and social, economic and spatial analysis of inequalities related to urban transport and policy evaluation in Latin America. He is consultant and adviser in different urban transport projects in Colombia. Consultant to the World Bank, adviser the Urban Development Institute of Bogotá on issues of value capture instruments for financing transport infrastructure. Adviser of the Bogotá Urban Planning Office in the reform of the Land-Use Planning Master Plan, in transport planning. Author of several articles published in international journals related to the evaluation of transport policies, poverty, equity and urban structure.

See Also

Transport Planning in the Global South; Transitions and Disruptive Technologies in Transport Planning; Urban Regeneration and Transportation Planning; Social and Distributional Impact Assessment in Transport Policy; Mobility Planning for Healthy Cities; Transport Planning and Management and its Implications in Chinese Cities; Mobility Planning and Policies for Older People;

Transport Policy and Governance; Accessibility Tools for Transport Policy and Planning; Transport and Climate Change; Externalities and External Costs in Transport Planning; Public Engagement in Transport Planning

References

- Acheampong, R.A., Silva, E., 2015. Land use–transport interaction modeling: a review of the literature and future research directions. *J. Transport Land Use* 8 (3), 11–38, doi:10.5198/jtlu.2015.806.
- Bertolini, L., 2012. Integrating mobility and urban development agendas: a manifesto. *disP - The Planning Review* 48 (1), 16–26, doi:10.1080/02513625.2012.702956.
- Cao, M., Hickman, R., 2019. Understanding travel and differential capabilities and functionings in Beijing. *Transport Policy* 83, 46–56, doi:10.1016/j.tranpol.2019.08.006.
- Cervero, R., Kockelman, K., 1997. Travel demand and the 3Ds: density, diversity, and design. *Transport. Res. Part D: Transport Environ.* 2 (3), 199–219, doi:10.1016/S1361-9209(97)00009-6.
- Geurs, K.T., van Wee, B., 2004. Accessibility evaluation of land-use and transport strategies: review and research directions. *J. Transport Geogr.* 12 (2), 127–140, doi:10.1016/j.jtrangeo.2003.10.005.
- Kasraian, D., et al., 2016. Long-term impacts of transport infrastructure networks on land-use change: an international review of empirical studies. *Transport Rev.* 36 (6), 772–792, doi:10.1080/01441647.2016.1168887.
- May, A.D. et al., 2005. *Decision Makers' Guidebook*. Leeds. Available from: https://www.fvv.tuwien.ac.at/fileadmin/_migrated/content_uploads/DMG_English_Version_2005_02.pdf.
- May, A.D., Kelly, C., Shepherd, S., 2006. The principles of integration in urban transport strategies. *Transport Policy* 13 (4), 319–327, doi:10.1016/j.tranpol.2005.12.005.
- Moeckel, R., et al., 2018. Trends in integrated land use/transport modeling: An evaluation of the state of the art. *Journal of Transport and Land Use* 11 (1), doi:10.5198/jtlu.2018.1205.
- Wegener, M., Fuerst, F., 2004. Land-use transport interaction: state of the art. *SSRN Electronic J.* Dortmund 119, doi:10.2139/ssrn.1434678.
- Zhang, Q., et al., 2019. Household trip generation and the built environment: does more density mean more trips? *Transport. Res. Rec.* 1–11, doi:10.1177/0361198119841854.

Further Reading

- Guzman, L.A., Peña, J., Carrasco, J.A., 2020. Assessing the role of the built environment and sociodemographic characteristics on walking travel distances in Bogotá. *J. Transp. Geogr.* 88, 102844, doi:10.1016/j.jtrangeo.2020.102844.
- Litman, T., 2019. *Evaluating transportation land use impacts. Considering the Impacts, Benefits and Costs of Different Land Use Development Patterns*. Victoria, Canada. <https://www.vtpi.org/>.
- Bertaud, A., 2018. *Order without Design: How markets shape cities*. In *Order without Design: How markets shape cities*. The MIT Press, Cambridge, MA 432.
- Leibowicz, B.D., 2017. Effects of urban land-use regulations on greenhouse gas emissions. *Cities* 70, 135–152, doi:10.1016/j.cities.2017.07.016.
- Kasraian, D., Maat, K., Stead, D., van Wee, B., 2016. Long-term impacts of transport infrastructure networks on land-use change: an international review of empirical studies. *Transport Rev.* 36 (6), 772–792, doi:10.1080/01441647.2016.1168887.
- van Wee, B., Bohte, W., Molin, E., Arentze, T., Liao, F., 2014. Policies for synchronization in the transport–land-use system. *Transport Policy* 31, 1–9, doi:10.1016/j.tranpol.2013.10.003.
- Cervero, R.B., 2013. Linking urban transport and land use in developing countries. *J. Transport Land Use* 6 (1), 7, doi:10.5198/jtlu.v6i1.425.
- Suzuki, H., Cervero, R., Iuchi, K., 2013. *Transforming cities with transit: transit and land-use integration for sustainable urban development*. Washington, DC: The World Bank. Available at: <https://openknowledge.worldbank.org/handle/10986/12233>.
- Heath, G.W., Brownson, R.C., Kruger, J., Miles, R., Powell, K.E., Ramsey, L.T., 2006. The effectiveness of urban design and land use and transport policies and practices to increase physical activity: a systematic review. *J. Phys. Activity Health* 3 (1), S55–S76.
- Ewing, R., Cervero, R., 2010. Travel and the built environment. a meta-analysis. *J. Am. Plan. Assoc.* 76 (3), 265–294, doi:10.1080/01944361003766766.
- van Wee, B., 2002. Land use and transport: research and policy challenges. *J. Transport Geogr.* 10 (4), 259–271, doi:10.1016/S0966-6923(02)00041-8.
- Geurs, K.T., van Eck, J.R., 2001. *Accessibility measures: review and applications. Evaluation of accessibility impacts of land-use transportation scenarios, and related social and economic impact*. Netherlands Environmental Assessment Agency. Utrecht.

Parking

Stephen Ison, Lucy Budd, Leicester Castle Business School, De Montfort University, Leicester, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	113
Types of Parking Space	113
Parking Management	114
Parking at the Workplace	115
Parking Charge at the Workplace	115
Workplace Parking Levy	115
Retail Parking	116
Residential Parking	116
Conclusions	116
References	117

Introduction

Car parking is a complex and problem rich area of study which encompasses issues relating to types of parking, availability, price, quality, different motivations for use (shoppers, leisure users, commuters, employers, and residents), different users (male/female, able bodied, or disabled) and the authorities who are tasked with managing this scarce resource as efficiently and effectively as possible with due consideration to land use planning, congestion, and the environmental impacts of private vehicle use (Ison and Mulley, 2014). When undertaking a journey by car, a parking space is required at the destination (and also possibly en-route in the case of a motorway service station or other intervening opportunity). In deciding where to park, cost, availability, and personal utility are all important in choosing not only where to park and the time to devote to the total journey, but also whether an alternative mode of travel is preferable to driving or indeed whether the journey is undertaken at all. This means that when developing a coherent parking policy there is a need to consider work, retail, leisure, education and residential parking, patterns of mobility, car ownership levels, the changing demographics, and mobility needs of the population, car ownership levels and the availability and frequency of public transport provision.

This chapter seeks to detail the types of car parking which are available as well as car parking management and workplace, retail, and residential parking considerations.

Types of Parking Space

Local authorities, private companies and, increasingly, residents (home-owners) provide and manage car parking spaces. Table 1 provides details of the types of parking which can be found in the majority of towns and cities worldwide.

Local authority on-street parking is parking on the roadside. Local authorities have direct control over it and as such can manage its supply or demand via the use of regulation and/or the price mechanism. Payment, where demanded, is usually made via a pay and display parking meter or (as is the case in many European and US cities) by phone.

Local authority off street surface or multistorey car parks refer to car parks that are not on the public road but that members of the public can utilize. Spaces are charged for either by means of pay and display machines or payment before exiting via a barrier system. There is normally a maximum period of stay and a sliding scale of pricing based on duration and the time of day or period of the week. Off-street parking of this type is readily available in towns and cities throughout Europe and North America (Rye and Koglin, 2014).

Privately owned public off street parking is available for public use and is charged for in the same way as local authority off-street parking is administered. It is increasingly common today to see payment through smart phones. This has the benefit of being *cashless* and has the dual advantage of aiding the motorist in that they do not have to carry cash in order to make payment (which often requires the correct amount) and the operator is spared the task of emptying the meters. It does, however, rely on good cellular network coverage and car drivers having a charged mobile telephone.

Private nonresidential parking is essentially off-street parking owned and controlled by a specific company or an office development. Spaces are often provided free for those associated with the organization and act as a perk of employment although charged spaces are increasingly common. This form of provision has been important both in terms of recruitment and retention of staff (Marsden, 2014). Supply of this type of parking space has its origins in planning regulations with developers being required to supply adequate provision for the size of development. This has, however, been seriously curtailed, in the UK at least, via restricting planning consent which seeks to restrict car parking spaces and encourage public transport use. The workplace parking levy (WPL)

Table 1 Types of car parking spaces

<i>Ownership</i>	<i>On or Off-street</i>	<i>User</i>	<i>Charging mechanism</i>
Local authority	On street [roadside] Off street	General Public General Public [surface or multistorey car parks]	Charged and free Mostly charged but can be free
Privately owned	Off street	Public	Charged
	Off street	Private nonresidential parking	Free
	Off street	Residents	Free (although some residents are now renting out parking spaces on their residential driveways)

Source: Adapted from Ison (2014).

has been advocated as a means of dealing with the issue of private nonresidential parking spaces. This is covered in more detail in a later section.

Residential parking essentially refers to spaces which are associated with privately owned houses and flats. While historically it was only residents who used such spaces there has been a growing trend in recent years for these spaces to be made available for use by nonresidents, particularly in locations near major generators of traffic such as hospitals, universities, and large sports stadia. This phenomenon will be dealt with in more detail below.

Parking Management

Bates and Leibling (2012) estimate that, on an average a car spends approximately 80% of its time parked at its owner's address, between 3% and 4% being driven and 16% parked at a point of destination be that work, retail, leisure, or education related. The 3%-4% could be spent in a congested urban environment with the social costs that entails, thus requiring some form of *demand management* while the 16% will most probably be in an urban area taking up valuable space that could be used more productively. This supports the view that the notion of "free parking" does not exist (Shoup, 2005).

Parking management, namely policies aimed at a more efficient use of scarce parking spaces, is required as part of a package of transport demand management measures to avoid the anarchic situation evidenced in many mega cities world-wide with hyper-congestion, environmental degradation, safety and health issues. As stated by the UK House of Commons Transport Committee (2013 p.5) "Effective parking strategies help to reconcile the competing demands of different road users. Parking restrictions are used to manage congestion and ensure that there is clear and fair access to public roads. The enforcement of parking restrictions should help to ensure that the needs of residents, shops, and businesses are met. Local authorities have primary responsibility for setting parking policy and enforcement strategies on local roads." Parking management can:

- encompass policies that lead to a more efficient use of parking resource;
- be an effective measure to reduce the level of traffic and its associated emissions;
- encourage motorists to utilize modes other than driving alone;
- form part of a "package" of transport measures; and
- be used as a demand management measure.

While parking management has been used extensively by local authorities in tackling congestion and localized air pollution it is often perceived as a second best measure when compared to congestion pricing which charges directly for the road space being used. In saying this, it has been a much easier approach to implement given its broader public appeal than say road pricing although there is often policy conflict between it being a demand management measure and its revenue generating capability and its impact on the local economy. Parking management can often be seen as reactive addressing an issue when it becomes evident such as commuters parking in residential streets.

Rye and Koglin (2014) detail the reasons for enacting parking management policies, namely:

- In certain localities there is inadequate or underutilized parking supply;
- There is competing demand for parking spaces from commuters, residents, and leisure users;
- This competing demand is likely to result in *parking search* (or cruising) as vehicles drive around searching for a parking space;
- In certain localities, most notably older- city centers containing terraced housing, there is a lack of supply, requiring the enactment of some form of permit system as a means of allocating demand;
- The issue of monitoring and enforcement of parking controls;
- Badly parked vehicles can result in safety issues and can impede access by emergency services vehicles.

The management of parking can take a number of forms depending on which type of parking is being considered, whether it is on-street or off-street. On-street parking management is in many cases the easiest to control either via the regulatory process by the use of double yellow lines (which, in the UK, denote the prohibition of on-street parking) or the pricing mechanism by the installation of parking meters which oblige drivers to purchase a ticket from a road side machine. This involves payments by cash,

card or mobile phone with drivers required to input their number plates to prevent loss of revenue when tickets are transferred between vehicles.

The introduction of double yellow lines and equivalent road markings has the dual effect of freeing up road capacity and also restricting the number of termination points at a destination. There is however, a downside to this approach. Limiting the number of spaces can lead to an increase in search time which clearly has an opportunity cost but also has the potential to increase the level of congestion and air pollution in the area where regulations are in place. In addition, there is the likelihood of encouraging through traffic since the supply of termination points has been curtailed and the road network freed up.

Technological solutions can also be used to enhance the efficient use of parking spaces so as to meet demand and also reduce parking search. One such solution is SF Park used in San Francisco in which sensors are embedded in on-street parking spaces in particular locations so that real-time availability of each space can be monitored and those seeking a space made aware of the fact that availability and dynamic pricing charges have been enacted as a means of reconciling demand with supply.

The use of parking meters can be configured so that the number of long stay spaces is reduced and that of short stay increased. This will have the effect of reducing the number of commuter journeys while increasing availability for shoppers and leisure users and also encouraging the turnover of spaces.

Another form of restraint involves what is known as *Controlled Parking Zones* (CPZ) where preference is afforded to local residents who purchase an annual permit allowing them to park in a particular zone. This is usually in areas where demand outstrips supply, most notably where there is a high proportion of terraced housing and in locations subject to excessive on-street commuter parking. In almost all cases, dedicated spaces are specifically demarcated and reserved for the use of disabled motorists (irrespective of whether they are the driver or a passenger). In the EU, these spaces are reserved for use by vehicles displaying a Blue Badge which has been authorized by the local authority due to specific health or mobility needs.

Parking at the Workplace

In terms of managing parking, one of the main issues is parking at the workplace and whether this can be controlled by the organization themselves or by the local authority.

Parking Charge at the Workplace

Employers may seek to implement a site-based parking management policy-based on the use of permits, which can either be issued free or charged for. Seeking to implement a site-based policy may be based on a number of factors namely:

- The site location is constrained, possibly in an urban area as is the case with many hospitals and universities both being significant generators of traffic but with limited availability of land for parking.
- Congestion maybe an issue both in and around the constrained site and on the radial routes used to access the site.
- Parking management is included as part of the companies travel plan.

As with any measure of this kind employee acceptance is a key issue. Questions raised when such a measure is being considered/implemented include the availability of alternative modes of transport, not least in respect to public transport and its frequency and reliability in peak commuter periods, and if a parking charge is in operation is it perceived as being fair and equitable. In terms of fairness and equity a number of organizations have linked their employee parking charge to their vehicle emissions and salary grade, in that the higher their vehicle emissions and their salary grade then the higher the parking charge they pay. As for occasional staff users a book of vouchers based on some proportion of the cost of a full permit might be issued. Visitors could be subject to a pay-and-display system with, say, the first 30 min free, 2 h and beyond charged with an increasing fee.

Workplace Parking Levy

One of the issues faced in relation to parking management is the issue of private nonresidential parking spaces, that is, office spaces provided by organizations for their employees which the authorities have limited ability to control directly. As a result, there is ongoing interest in the WPL which is a scheme designed to charge employers a levy for the number of parking spaces they have available on their site. The WPL was first proposed in the UK in 1998 as part of the Government's White Paper on the Future of Transport: A New Deal for Transport Better for Everyone. In the UK, the Government subsequently gave power to local authorities to implement a WPL with the proviso that the revenue was hypothecated for transport improvements. To date only Nottingham has implemented such a charge in the UK but other cities most notably Bristol, Edinburgh, Glasgow, Leicester and Reading plus the London Boroughs of Camden and Hounslow are, at the time of writing, showing interest. The WPL scheme is based on the notion of charging employers for each parking space they provide for their employees above a certain number. A dedicated chapter is available on the topic of the WPL, focusing on the Nottingham scheme, so the following offers only a short bullet point introduction.

- The Nottingham WPL was fully implemented in April 2012 and is currently the only WPL scheme in existence in the UK;
- The objectives for the Nottingham scheme are to:
 - address the issue of congestion through encouraging employers to reduce their number of parking spaces [and thus their WPL bill] and as such limit the number of parking destination points for motorists;

- use the hypothecated revenue raised from the WPL to secure “high quality sustainable public transport” that offers realistic alternatives to the private car;
- assist business in developing “smarter travel choices” as part of dedicated travel plans;
- have no significant impact on business investment decisions in the city;
- create no significant displaced parking issues;
- The main driver for opting for the WPL option in Nottingham was to enable a quicker income stream generation than would have been possible if a road pricing scheme had been introduced. A stream required so as to part fund the new tram line infrastructure development;
- Those employers providing 10 or more employee parking spaces are levied per space annually.

Retail Parking

Historically large out of town shopping developments have been a feature of retail activity particularly in the UK and US. Such locations have tended to offer free parking on a large scale. It is clear that with the advent of on-line Internet shopping companies such as Amazon are having a significant impact on the demand for this type of retail experience. As such land used for this purpose is likely to change being made available for redevelopment such as housing use.

In addition, the traditional high street is being significantly affected by on-line shopping and the change of purchasing habits. Thus parking associated with retail shopping in town and city centers and the pricing of those spaces is increasingly being called into question as ways are being sought to halt the decline in high street footfall. A related issue with respect to parking provision in urban areas is that of multistorey car parks. In many instances this comprises ageing infrastructure in need of replacement or retrofit. In addition, the users are increasingly being identified as an ageing population with a range of disabilities.

Residential Parking

There has been a growing world-wide trend in recent years for private accommodation, namely houses and flats to hire out their off-street parking spaces for use by nonresidents in certain locations. This unregulated residential car parking provision has emerged in recent years and forms an online virtual parking marketplace. This has involved residents seeking to address the needs of those motorists wanting to park, the cost of that parking and the lack of available car parking spaces around major generators of private motor vehicle traffic be that in the vicinity of airports, hospitals, or major sporting events. On-line parking spaces are accessed via websites which seek to match individuals who want to rent out private spaces with drivers looking for somewhere to park their vehicles. The individual parking providers upload a profile detailing the location, type, and importantly the price of their parking space or spaces and motorists search for a suitable space by location/keyword.

The use of residential parking spaces offers the user a different product type which is normally cheaper, thus appealing, and also affords a personal service which could involve the user/s being dropped off, for example, at an airport, valeting the user's car whilst away or even looking after the customers dogs. As regards the provider then renting out what is usually an unused parking space is a relatively easy means of supplementing their income or in the case of those that are unemployed then their only source of income.

Clearly there are implications not least in that it could represent a challenge, in terms of competition to traditional providers, which could be a local authority or airport, it could lead to traffic congestion and safety issues as a result of inconsiderate parking and there are also security issues raised by leaving a vehicle in the driveway of what is essentially a strangers property (see [Budd et al., 2013](#)).

Conclusions

Parking is a complex and at times controversial topic not least since a change in one area has implications to other aspects of the system. It has an economic, political and technological dimension, is multifaceted and pervades all aspects of transport demand management. It is an area subject to incremental change such as the introduction of a CPZ which is more likely to gain public and political support than large step changes. Parking management, comprising the efficient utilization of an important resource, is a significant element of transport policy and extensively used by local authority planners and increasingly individual organizations as a means of addressing the issue of demand and supply and the related congestion and environmental implications. Parking controls and pricing decisions can be used widely as a means of managing urban transport and improving network traffic flows. As such, over time there has been an increase in on-street parking controls in the form of time restrictions or varied pricing tariffs in addition, to new initiatives such as the introduction of the WPL as a strategy aimed at managing parking.

There is a range of emerging issues associated with parking and parking management not least the:

- increased use of and role for technology most notably parking apps and schemes such as *SF Park*;
- ways in which the Internet has been used to stimulate the online virtual parking marketplace allowing for a new area of parking supply with the associated implications;

- changing nature of parking demand given the advent of on-line shopping and the associated decline of the high-street;
- sophisticated use of parking management at the organizational level;
- desire to impact on private nonresidential parking spaces via the introduction of the WPL and the associated hypothecation of the revenue raised.

One thing is certain and that is the field of parking is dynamic and will continue to change. For example, the advent of the autonomous vehicle is likely to negate the need for large areas of expensive land in urban area being devoted to car parking. As such, urban land use will most probably see significant reconfiguration.

References

- Bates, J., Leibling, D., 2012. Spaced Out Perspectives on Parking Policy. RAC Foundation, London.
- Brooke, S., Ison, S.G., Quddus, M.A., 2014. On-street parking search: review and future research direction. *Transport. Res. Rec.* 65–75, TRB, Washington D.C, USA, No 2469.
- Budd, L., Ison, S.G., Budd, T., 2013. An empirical examination of the growing phenomenon of off-site residential car parking provision: the situation at UK airports. *Transport. Res. A: Policy Practice* 54, 26–34.
- Ison, S.G., Budd, L., 2016. Parking chapter 9. In: Bliemer, C.J., Mulley, C., Moutou, C. (Eds.), *Handbook on Transport and Urban Planning in the Developed World*. Edward Elgar Publishing, Cheltenham, UK, pp. 145–161.
- House of Commons Transport Committee, 2013. Local authority parking enforcement, Seventh Report of Session 2013–2014 (Vol. I, HC 118), The House of Commons, London: The Stationery Office.
- Ison, S.G., 2014. Parking Management Policy: It's potential in improving urban flows Bruxelles. European Automobile Manufacturers Association (ACEA), Belgium.
- Ison, S.G., Mulley, C. (Eds.), 2014. *Parking: Issues and Policies*. Emerald Publishing Limited, Bingley.
- Marsden, G., 2014. Parking policy. In: Ison, S.G., Mulley, C. (Eds.), *Parking: Issues and Policies*. Emerald Publishing Limited, Bingley.
- Rye, T., Koglin, T., 2014. Parking management. In: Ison, S.G., Mulley, C. (Eds.), *Parking: Issues and Policies*. Emerald Publishing Limited, Bingley.
- Shoup, D., 2005. *The High Cost of Parking*. American Planning Association, Chicago, IL.

Transport Planning in the Global South

Daniel Oviedo, Mariajose Nieto-Combariza, Development Planning Unit, University College London, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Evolution of Passenger Transport Planning: In Search of Efficiency	118
Adoption and Development of the UTP in the Global South	118
Challenges and Opportunities for Transport Planning in the Global South	120
Rapid Urbanization and Transition to Automobile Use	120
Paradigm Shifts and the Resurgence of Public Transport, Cycling, and Walking	121
Institutional Weakness Versus Growing Scope for New Policies	121
Equity, Segregation, and Socioeconomic Vulnerability	122
Informality and Innovation	122
Looking Forward: Gaps in Knowledge and Priorities for Research in a Rapidly Changing Environment	123
References	123
Further Reading	123

Evolution of Passenger Transport Planning: In Search of Efficiency

The Urban Transport Planning (UTP) process is a formalized planning methodology designed specifically for assessing current and expected transport conditions in a city and developing and prioritizing strategies for their improvement in the short and long term. The core components of the process are transport system analysis and forecasting, land-use and urban development assessment and forecasting, and goals, policy, and plans formulation based on technical criteria (Dimitriou, 2013).

With the evolution of information technologies and an increasing ease to collect reliable data at larger scales for planning purposes, the UTP process became highly influenced by advances in demand modeling. In particular, the development of the four-step transport demand model, used for planning and assessment of transport projects up to this date, became a central driver of the evolution of the UTP Process. This model seeks to simulate and forecast travel conditions and estimate the effects of different interventions on transport demand. The model is concerned with the stages of decision making of the individual traveler. This is one of the aspects that has received the most attention from the academic and practitioner community since the development of the UTP process, to the extent that it is often mistaken with the UTP itself (Dimitriou, 2013).

The evolution of this model has been guided more by the refinement of the tools rather than a thorough conceptual debate (Bates, 2007). The coherence and compatibility between the unified framework for transport modeling and planning behind the four-step model and the economic theory of utility maximization, and supply-demand balance, as well as the rapid evolution of computing capabilities has allowed mainstream techniques to grow in scale as new and more complex problems can be addressed (Bates, 2000). Modern modeling techniques for four-step transport demand analyses depend largely on big data sets which allow for high statistical representativeness at different levels of aggregation (Ortuzar and Willumsen, 2011).

In comparison with the components of the four-step model, other techniques and submodels for urban development analysis and forecasting included in the UTP process have experienced a slower development despite early studies dating from the 1950s and 1960s (Dimitriou, 2013). This is the case of tools to analyze urban development and interactions of land-use and transport in the context of the UTP process that seek to provide inputs and feedback to traffic submodels. The complexity of these interactions and the still active debates in relation to the conceptual and practical implications of transport and land use models have inspired an important number of works and tool developments in this area. However, lack of consensus related to the best approach to transport and land-use modeling have “made theories and software almost inseparable” (Ortuzar and Willumsen, 2011, p. 493). Land-use based analysis and forecasting are almost as data intensive as the previously described four-step model of the UTP process. However, as with previous submodels, these models and analyses require assumptions on public and private sector policies, decision-making processes, and human behavior for their effectiveness (Dimitriou and Gakenheimer, 2011). The UTP has also adopted formal methodologies for comparing the intended and unintended effects of transport interventions over the years, such as Cost-benefit analysis (CBA).

Adoption and Development of the UTP in the Global South

Approaches to modern UTP initially developed in the United States of America and Europe were later adopted in countries of the global south during the 1970s and 1980s. This adoption was often facilitated by international development agencies. Such technical and technological transfer of transport planning approaches to developing countries have led to disenchantment with city-wide, long-range efforts at UTP. While most theories applied today in the global south for transport planning come directly or indirectly

from western models, the pillars, limitations and requirements of mobility, and accessibility in the urban south are substantially different from the context in which these theories and knowledge originated.

Although, geographical, economic, and even social elements play a significant role in current transport planning conducted under the UTP process, the overarching rule of technocratic approaches has governed not only the past evolution of transport planning in the industrialized world, but also the way it was transferred to, and further developed in the global south. Between the 1960s and 1980s the UTP process suffered significant changes in industrialized countries because of revised policy goals related to environmental and social sustainability, while the developing world saw countless applications of the process in its initial form. The formalization and widespread distribution of methodologies and frameworks for UTP in the developed world and the role of development organizations on urban development expedited international dialogue and adoption of the UTP in many cities of developing countries.

During the 1960s, pressing increases in demand for private motorized transport in developing cities led local governments to seek advice from international consultants. This led to methods from the early definitions of the UTP to be applied in developing countries with minor adaptation to the context of developing cities by consultants under limitations of time and resources imposed by their contracts. In the years that followed this initial technical transfer of approaches to UTP, various North American, European, and Japanese consulting firms started offering their expertise to developing countries at a larger scale, profiting from refinement in modeling techniques and increasing computing capabilities. These studies often led to transport master plans and guidelines for development later adopted by local practitioners (Cervero, 2013). In addition, influence from foreign and local construction companies profiting from infrastructure-centered plans contributed to the consolidation of ambitious transport plans aiming at reforming urban structures through large infrastructure investments with the aims of modernization, increasing economic output, and employment generation (Vasconcellos, 2014).

Financial aid and capacity building initiatives are inextricably influenced by policy priorities and dominant knowledge in the international agenda (Dimitriou, 2013). In this regard, increasing importance of the role of transport for development in the agenda of international aid agencies from the developed world in the 1970s sought to respond to the transport challenges of cities in the developing world. Increase in urban transport lending by the World Bank grew from US\$32 to US\$113 million per year for technical assistance and from 0 to US\$13.1 million for training between 1972 and mid-1980s. International cooperation missions focused primarily on preparing local practitioners and civil servants in transport to design and implement policies under the UTP. In addition, many studies funded with international development aid produced planning guidelines based on the UTP process for large and medium-sized cities in the developing world, which were later incorporated to requirements for qualifying for external and internal urban transport financing (Dimitriou and Gakenheimer, 2011).

International development agencies are not the only agents involved in the transfer of policies and planning practices to the Global south. Urban policy has circulated at increasing speed due to several factors leading to “fast policy”. Better transport and communication have enabled rapid visits to other cities and new forms of evaluation allow rapid engagements with other policies. Intermediation from institutions such as think tanks and consulting companies has grown considerably since the 1970s, which profit primarily from policy models they produce and circulate. Recent decades have witnessed a “triumph of politics over policy”, which have influenced policy turnover times aligning them with political cycles. Considering that automobile owners have traditionally been a powerful lobby in the developing world, this aided development of infrastructure-incentive and traffic-centered policies for a good part of the 1980s and 1990s.

Prescriptions from structural adjustment policies fostered investments in additional network capacity in the 1980s in search for improvements in economic performance (Dimitriou and Gakenheimer, 2011). These policies not only added to a rapid change in the structure of transport networks in developing cities, but also to the dissemination and adoption of transport modeling and design tools.

Although many limitations identified in early models for demand forecasting have been resolved throughout the years, higher model complexity has increased requirements for technical and financial resources to access such tools. This becomes a daunting challenge for the adoption of travel demand models in contexts where resources are limited, as well as a problem of coherence between evolving regulations seeking to incorporate better modeling in transport planning regulations and the capacity of local governments to finance, and make use of such information (Cervero, 2013; Dimitriou and Gakenheimer, 2011).

These priorities in conditions for development aid changed in the 1980s and 1990s, incorporating new concerns related to increasing efficiency of public transport. A policy paper published by the World Bank in 1986 for the transport sector, which included priorities for UTP in developing countries, stressed the need for efficient transport management, involvement of private sector, reduction in subsidies, and incentives to competition under a deregulation framework. This echoed the approaches to planning and policy development in many cities that had started to develop their own solutions to urban transport issues using local ingenuity and increased funding from both local and international sources (i.e., Curitiba, Bogotá, Mexico, Lagos, Bangkok).

Developments in the late 1990s and 2000s were more concerned with the issues of sustainability and incorporated some long-standing needs for capacity building, institutional strengthening and focus on public transport, and empowerment of nonmotorized alternatives. While significant changes in urban transport development were taking place in some cities of the global south, these did not influence significantly the framework for UTP and policy evaluation. Discussions on environmental sustainability following the oil crisis of the 1990s strengthened the need for sustainable transport development and focus on efficient mobility systems centered on public transport investments. These interventions aimed at using data from technical appraisals to forecast travel demand and design policy instruments that allowed efficient traffic management that reduced the environmental footprint of urban transport development. This shift in the planning logic marked the inclusion of more rigorous criteria of environmental appraisals in transport policies throughout the world and defined new conditions for local and international financing for infrastructure projects.

Challenges and Opportunities for Transport Planning in the Global South

Tools for transport planning have experienced continuous refinement in the last three decades, although modeling has been a central issue even in modern transport studies, which becomes a point of debate. The principles of rational choice and utility maximization that underpin most transport demand models can be considered as too narrow for reflecting the multiplicity of needs, preferences and drivers of transport users with different social identities, beliefs, and principles. The framework of the UTP has suffered relatively minor changes in its general structure despite incorporation of new concepts and methodologies in its different stages in the industrialized world. This is partly explained by shifts in policy priorities and emerging concerns in relation to inclusiveness and sustainability of transport, which have influenced from basic concepts governing transport modeling to tools for strengthening decision making approaches. However, such changes in policy objectives and priorities, as well as the widespread adoption of better-adapted forms of planning in global south cities are still challenged by rapid urbanization and population growth, the increasing dependency on car use, high degrees of poverty, social inequality, and weak institutions. These challenges can also often present opportunities of innovation and localized developments, which need to be acknowledged if the positive evolution of transport planning in global south cities is to continue.

Rapid Urbanization and Transition to Automobile Use

There is a strong correlation between urbanization, per capita income and motorization (Vasconcellos, 2014), which links economic development and its associated increases in purchasing power to the rise of individual mobility and motorization rates. Car ownership has grown rapidly in emerging economies such as China, India, Mexico, and Russia (Pojani and Stead, 2017). Increased use of private automobiles and motorcycles contributes to a substantial increase in traffic congestion and travel times. Various case studies also point to a rapid increase in traffic fatalities of air quality concerns. Such is the case of cities in Brazil, Nigeria, Vietnam, and Indonesia, while large Indian cities are some of the most polluted in the world.

While it is expected that increases in per capita GDP are accompanied by an increase in individual mobility and motorization rates, abundant evidence from industrialized countries shows that motorization rates reach peaks in ownership of private vehicles that remain constant despite further increases in purchasing power. Rates of around 600 vehicles per 1000 inhabitants are among the highest values recorded in these contexts.

Car and motorcycle ownership rates can be largely explained by income and population distribution. With the rise of the middle classes and the transition from low to middle income levels there is an expected increase in both automobiles and motorcycles in cities of the global south. Nonetheless, motorization rates in the global south are still low compared to industrialized countries. The number of cars per 1000 inhabitants varies from 55 in Ecuador, to 185 vehicles in Mexico. Countries in the lowest range of motorization rates such as Peru and Bolivia are between 51 and 56 cars per 1000 inhabitants. Asian cities are continuing to grow very rapidly. In many cities the number of vehicles is doubling every 5–7 years.

Evolution of the personal vehicles industry has led to increasingly affordable cars and motorcycles. The latter has a much lower income threshold for purchasing and daily operation, which has led to a rapid growth in motorcycle ownership. In Uruguay, the number of motorcycles per 1000 inhabitants reaches 141 vehicles, while in Brazil and Colombia this figure is 81 and 68, respectively (Hidalgo and Huizenga, 2013). These figures show a tendency to increase as a result of economic growth and a reduction of the price of vehicles being distributed. The number of motorcycles in Brazil increased by 38% annually between 2000 and 2010 while in Colombia and Mexico growth rates has been between 14.7% and 16.4% per year, respectively. On average, cars have increased by about 6% annually in most parts of the global south.

This increase in rates of motorization has consequences on environmental sustainability as a result of externalities such as congestion and air and noise pollution, which stands in contrast with the comparatively low share of demand served by private vehicles. In cities like La Paz, Rio de Janeiro, and Montevideo, less than 20% of daily trips are made by private vehicles. In other cities, such as Bogotá, Belo Horizonte, Porto Alegre, Santiago, Caracas, and Guadalajara, this percentage is still below 30%. These figures contrast with increasing percentages of journeys made by public transport and nonmotorized modes. 45% of people in Maputo, Mozambique, walk as their main mean of transport. In Rio de Janeiro, while in Belo Horizonte, Santiago, León, Guadalajara, and Curitiba, nonmotorized travel represents between 32% and 42% of daily trips. In most cities this is complemented by at least 40% of trips being completed by public transport (Hidalgo and Huizenga, 2013).

Uteng and Lucas (2018) argue that one of the encompassing features of transport planning in developing countries is the central role that automobility infrastructure keeps having, despite the predominance of use by their dwellers of nonmotorized modes. Sagaris and Arora (2017) show that despite more than 70% of trips in Delhi and Santiago are made by public transport, walking, and cycling, key transport policies such as time/space restrictions on rickshaws and a major highway project (la Costanera Norte) keep imposing a system of automobility. Urban street use in developing cities often has a conflicting nature, with a complex pattern of coexistence between pedestrians, vehicles, vendors, and even animals, which also makes interventions more difficult (Dimitriou and Gakenheimer, 2011; Vasconcellos, 2014). In Hanoi, street vending was banned from 63 streets giving privilege to traffic flow over vendors in the urban core.

Pedestrians and users of nonmotorized vehicles account for most of the traffic related deaths in countries with lower uses of motorized vehicles. While in Delhi 43% of road users killed were pedestrians whilst in the Netherlands, Norway, Australia, and USA pedestrians account for less than 18% of the deaths with motorized four wheelers account for at least more than half the casualties. The growth in motorization is accompanied by inequalities in exposures to risks of traffic accidents, air, and noise pollution, as well

as the increased decrease in quality of life, which tends to affect more severely socially vulnerable groups and nonusers of motorized transport. The World Health Organisation recognizes transport-related effects such as exposure to pollution and road accidents as some of the leading global causes of death. These tend to have a much higher effect in global south cities, where most of the population is in a vulnerable position to deal with the aftermath of traffic-related injuries and sickness.

Pojani and Stead (2015) argue that medium-sized cities in the developing world can offer greater potential for more sustainable transformations than megacities. They generally have a smaller ecological footprint, and in principle, their size allows for flexibility in terms of urban expansion, adoption of “green” travel modes, and environmental protection. However, there is the limitation of smaller budgets and technical resources to implement transport programs and projects and probably higher levels of financial, and climatic vulnerability. Also, compared to larger cities in developing countries, generally medium-sized cities are less dense engendering less efficient public transport systems, lower shares of public transport use, and consequently more transport related energy consumption per citizen.

Paradigm Shifts and the Resurgence of Public Transport, Cycling, and Walking

In recent years, investment in public transport infrastructure of high capacity has increased in cities across the global south. An example of such a trend is emerging Bus Rapid Transit (BRT) systems. According to data from BRTData (2020), (104) cities in Africa, Asia, and Latin America have developed BRT systems that add to up to 3590 km in 2020. Something similar has occurred with cycling infrastructure throughout the developing world. By 2011, 327 cities in Latin America had exclusive lanes for bicycles (85% of them in Brazil). Availability of infrastructure by type of transport mode shows an increasing effort by national and local governments to increase supply of public transport and to provide better conditions for nonmotorized travel, although the latter is much more limited.

High densities and mixed uses are necessary for public transport and nonmotorized modes to be financially feasible and practical. To have compact urban forms land-use controls are necessary and consequently these measures have important results in terms of mobility patterns. Measures to manipulate a city’s shape, density, decentralization, and building layout between others can have a substantial effect in city issues such as congestion and pollution (Jenks et al., 2000). In the global south, these characteristics variate highly, with most Asian cities is being highly dense and sub-Saharan Africa cities showing large urban footprints with very low density (Pojani and Stead, 2015).

In China, there has been a renaissance of cycling—in 2005 10 million e-bikes were sold, multiplying by 250 the number sold in 1998. Weinert et al. (2007) attribute such increase to several convergent elements: bike-sharing programs with electric technology that allows for longer distances to be covered, safety regulations, lower e-bike prices, growing incomes, and regulations against gasoline vehicles. In other cities like Bogotá and Sao Paulo, cycle networks have been rapidly expanded, in China motorcycles lanes have been transformed into bicycle lanes (Dimitriou and Gakenheimer, 2011) and Daejeon and Rio de Janeiro have bike sharing systems with fleets of over 200 bikes (Midgley, 2011). This is supported by programs seeking to foster active lifestyles though temporary closures of roads during weekends for walking, cycling, and do physical activity such as Vida Chile, Bogotá’s Sundays Ciclovía, and the more recent cycle days in Kigali, Rwanda.

Institutional Weakness Versus Growing Scope for New Policies

A common trait in global south cities is the absence of analysis to develop a planning process based on specific contexts due primarily to a lack of standardized systems of data collection (Uteng and Lucas, 2018). Pojani and Stead (2017) identify lack of institutional capacity and lack of cross-agency collaboration as main challenges emerging cities face to plan for more sustainable and equitable transport systems. Most developing cities share the need of shaking off accumulated inefficiency in transport practices, institutional weakness, and consolidated power of privately managed public transport with very weak regulation.

Mainstream response to such challenges has entailed structural institutional and operational changes often materialized on high investments in infrastructure. However, weak institutional and physical integration between different sectors of urban and transport planning and operation have led to limited progress in the modernization of urban transport. Financial difficulties in the management, operation, and development of urban transport reflect more structural issues such as poor management, poor institutional coordination, and the lack of comprehensive transport, and land use planning integration. This is compounded by lack of technical training for civil servants and technical staff at local planning offices and a frequent lack of incentives for high-quality professionals to join the public sector.

Lack of effective enforcement of traffic laws and regulations, coupled with weak enforcement institutional capacity and corruption pose another significant challenge for planning, regulation, and decision-making. These are often compounded by complex power relations associated with context-specific cultural and social values. Demand management, road safety, and space management require the resources to enforce the law or otherwise they do not work.

Some of these challenges can be aided by technological developments seeking to strengthen capacity for management and regulation of transport supply, and demand. The most common are traffic signaling, and surveillance, tracking systems, electronic ticketing services and toll collection, bus management systems, and information services (Pojani and Stead, 2017). Implementation of technological aids to improve such conditions depends on the availability of financial resources and requires a change on the predominantly indifferent attitude by local professionals and lack of user trust.

Awareness-raising campaigns as a strategy to reduce environmental impact of mobility patterns in cities are more likely to succeed in global south urban spaces when they are low-cost activities such as car-free days or free vehicle inspection days. Support of charismatic political leaders is important and often even crucial for their reception (Pardo, 2001). Public attitude toward road pricing as a mechanism to reduce car use has been shown to become more positive after the programs have been implemented.

An advantage of command and control measures to manage car use in developing countries is that rationings and banning are interpreted to affect everyone equally, making them politically easier to implement. In Bangkok, a program to restrict all newly registered cars to use exclusively in nonpeak hours was implemented with positive results, while in Guangzhou only locally registered motorcycles can drive in the city. Other cities like Sao Paulo, Bogotá, and Mexico City have imposed temporary restriction based on license plate numbers to reduce congestion and pollution.

Equity, Segregation, and Socioeconomic Vulnerability

Developing cities tend to be spatially and socially segregated with high levels of spatial inequalities. High-income gated communities well-connected to the transport systems stand in stark contrast with low-income informal settlements bypassed in the processes of infrastructure and services provision. Poor neighborhoods also have very poor local supply of employment and other opportunities, which reinforces the negative effects of urban geography in their accessibility and inclusion. In most global south cities, poverty and peripherality are highly correlated, while most informal and poor neighborhoods share a generalized lack of infrastructure and transport supply and investment.

The intersection between poverty, segregated urban forms, and low supply of affordable transport, leads the poor in global south cities to spend a disproportionate share of their income in transport. Carruthers et al. (2005) found that in 2002 in Buenos Aires, the average public transport expenditure to commute to work as a share of income was 31.6% for the bottom income quintile. Transportation expenditures as a share of income were 6.4% and 11.7% for the bottom income quintile in Montevideo, Uruguay and Córdoba, Argentina, respectively. Carruthers et al. (2005) estimate that transportation expenditures as a share of total household expenditures, averaged over the entire population, were 15% in Yaoundé, Cameroon. In African and South-East Asian cities, poor populations walk because they cannot afford public transport fares, trading off affordability for very long walking trips highly exposed to traffic accidents and pollution.

Recent innovative transport strategies have changed some of these narratives in specific cities. Medellín's cable car system has improved accessibility of low-income hillside neighborhoods while increasing access to jobs and influencing social cohesion in the city (Dávila, 2013). Ropeways, air gondolas, or cabled propelled transit systems are fully segregated systems with much limited capacity than other technologies and a considerably lower capital investment. The aerial operation has made this technology a convenient alternative to overcome challenging geographies such as hilly terrains, rivers, harbors or dense, and historic neighborhoods.

Informality and Innovation

An unclear division between formal and informal modes is a common characteristic of governance and operation of transport in global south cities (Uteng and Lucas, 2018). The formal–informal dichotomy in practice is more a continuum of state regulation and unwritten but generalized rules and practices that emerge from established and changing social norms and traditions. Regulatory and service gaps create spaces for private, predominantly unregulated services, that grow without state intervention and sometimes dominate transport markets as in various cities in South Africa.

Informal transport services provide on-demand mobility for people that depend on transit to be mobile while representing an income source for low-skilled workers and the only alternative for areas where there is no formal transit supply (Cervero and Golub, 2007). They involve externalities such as increased congestion, pollution, accidents, and sometimes violence among owners and operators. In Nigeria, informal transport services have grown rapidly since the collapse of formal public transport system that could not cope with growing operation cost encompassed by fare regulation preventing fare increases.

Informal transport solutions have a dominant and pervasive presence in developing countries, but it remains difficult to access information about the lives of those involved in providing informal transport using formal approaches. New approaches to informal transport services are necessary to include these services that are increasingly marginalized in transport planning when cities in the process of policies of modernization in the global south. Similarly, ICT for collaborative paratransit mapping can be a method to close the gap in official planning of a key mobility system and to create a more inclusive transportation planning dialogue.

Across the Global South, the dominant emerging trend amongst paratransit services is perhaps the mushrooming of potentially disruptive digital platforms related to their operation and use. Widespread availability of mobile phones in Asian and Latin American cities has enabled new forms of mobility and changed the organization and operation of informal transport services. Over 150 mobility companies in Africa provide services from journey planning to ride-hailing and cashless fare collection (e.g., Gona in Lagos). In Latin American cities, a fertile ecosystem for start-ups has given rise to either homegrown initiatives, or adaptations of platforms from elsewhere to local conditions. Local and global entrepreneurs have introduced new forms of unregulated services such as “microtransit” and ride-hailing. The operation of these market entrants has been hampered by a lack of public sector regulation and violent acts on drivers by incumbent operators.

Looking Forward: Gaps in Knowledge and Priorities for Research in a Rapidly Changing Environment

The challenge for integration of mainstream approaches to UTP and theories concerning the social dimensions of transport interventions in the global south requires more emphasis on the international and interdisciplinary dialogues on urban development. Global south cities need to overcome the fragmentation in the planning process and develop a framework that can encompass the different dimensions of transport and urban development in all its stages. Social considerations have been limited in practice to the evaluation and monitoring spheres of the planning process, with only some project-specific initiatives to address poverty in some developing cities. Similarly, the influence of shifts in policy priorities and guidelines had a more perceivable effect on the detailed implementation of the UTP process rather than its overall structure.

A recent change of international interest towards poverty alleviation in development has presented an opportunity for transport strategies to play a more active role as part of programs aiming to reduce poverty in urban environments. This has involved the inclusion of new criteria for policy design and evaluation and a debate in developing contexts in relation to the objectives of transport master plans and infrastructure interventions that has not yet been fully explored in the international literature. Moreover, research on approaches to address the specific needs of disadvantaged populations that inform policies in their favor is another gap in current research and practice of transport planning in the global south. The diversity of societal environments, traditions and conventions that underpin gender relations in transport is another relevant area of research as the gendered dimension of transport planning has only been addressed in a handful of works in urban transport research in global south cities. There is a need for additional debates around ways of incorporating a stronger social dimension in theory and practice around urban transport in cities of developing countries.

Other priorities for research in these contexts include the examination of the links between transport and land-use and the added complexities associated with informality in housing, labor and transport in developing cities. Moreover, the contribution of different formal and informal innovations and the role of ICT in urban transport provision, mobility, and accessibility. This includes not only research on the effect of ride-hailing, micromobility, and technological transitions for sustainable transport, but also the many homegrown innovations that entrepreneurs and governments across Sub-Saharan Africa, South-East Asia, and Latin America are producing daily. Finally, the links between transport and health in global south cities, where a large share of the population is suffering from many transport-related causes of death and where the characteristics of society and the built environment may increase vulnerability to health-related risks and concerns are key research priorities.

Such research priorities take place in increasingly changing environments. Recent events in relation to the COVID-19 pandemic have forced researchers and practitioners to rethink the role of transport in society and to adapt its planning and operation. Considering the pace of demographic and economic growth in cities in the global south and their still large gaps in access to basic public services, such challenges are only likely to grow in scale and relevance. A predominantly younger population and expected economic growth annual rates of 4%–5% on average even in sub-Saharan Africa pose challenges that escape the priorities of research and planning in industrialized societies. Issues of urban governance and rapidly changing political environments are likely to pose challenges for new research and planning direction, which need to adapt to transitions to accelerate economic and social development. Despite poverty, inequality, and slow economic structural transformation and violence, global south cities will be at the center of development policy and practice in the following years and transport will continue claiming a key role in achieving more sustainable and inclusive cities, calling for larger focus in these contexts and posing an exciting challenge for researchers and practitioners of transport planning globally.

References

- Bates, J., 2007. History of demand modeling. In: Hensher, D. A., Button, K.J. *Handbook of Transport Modelling*. Elsevier Science, Oxford, pp. 11–33.
- Carruthers, R., Dick, M., Saurkar, A., 2005. Affordability of Public Transport in Developing Countries. Retrieved from <https://openknowledge.worldbank.org/handle/10986/17408>.
- Cervero, R., Golub, A., 2007. Informal transport: a global perspective. *Transp Policy* 14 (6), 445–457, doi:10.1016/j.tranpol.2007.04.011.
- Cervero, R.R., 2013. Linking urban transport and land use in developing countries. *J Transp Land Use* 6, 7–24, doi:10.5198/jtlu.v1.425.
- Dávila, J., 2013. Urban mobility and poverty: Lessons from Medellín and Soacha, Colombia (J. Dávila, Ed.).
- Dimitriou, H., 2013. *Transport Planning for Third World Cities* (Routledge Revivals).
- Dimitriou, H., Gakenheimer, R., 2011. *Urban Transport in the Developing World: A Handbook of Policy and Practice*.
- Gwilliam, K., 2003. Urban transport in developing countries. *Transp Rev* 23 (2), 197–216, doi:10.1080/01441640309893.
- Hensher, D.A., Button, K.J., 2000. *Handbook of Transport Modelling*. Pergamon. Elsevier, Amsterdam (NL).
- Pojani, D., Stead, D., (Eds.), 2017. *The Urban Transport Crisis in Emerging Economies*. <https://doi.org/10.1007/978-3-319-43851-1>.
- Sagaris, L., Arora, A., 2017. Cycling for social justice in democratizing contexts: Rethinking “sustainable” mobilities. In: *Urban Mobilities in the Global South*. Routledge, pp.19–40.
- Uteng, T.P., Lucas, K., 2018. The trajectories of urban mobilities in the Global South. In *Urban Mobilities in the Global South* (pp. 1–18). <https://doi.org/10.4324/9781315265094-1>.
- Vasconcellos, E., 2014. *Urban Transport Environment and Equity: The case for developing countries*.

Further Reading

- Anas, A., Lindsey, R., 2011. Symposium: transportation and the environment reducing urban road transportation externalities: road pricing in theory and in practice. *Rev. Environ. Econ. Policy* 5 (1), 66–68, doi:10.1093/reep/req019.
- Bocarejo, J.P., Portilla, I.J., Velásquez, J.M., Cruz, M.N., Peña, A., Oviedo, D.R., 2014. An innovative transit system and its impact on low income users: the case of the Metrocable in Medellín. *J. Transp. Geogr.* 39, 49–61, doi:10.1016/j.jtrangeo.2014.06.018.

- BRTData, 2020. Global BRTData. Retrieved April 20, 2020, from <https://brtdata.org/>.
- Carrigan, A., King, R., Velasquez, J.M., Raifman, M., Duduta, N., 2013. Social, environmental, and economic impacts of bus rapid transit: case studies from around the world. *Embarq* 151.
- Carruthers, R., Dick, M., Saurkar, A., 2005. Affordability of Public Transport in Developing Countries. Retrieved from <https://openknowledge.worldbank.org/handle/10986/17408>.
- Cervero, R., Golub, A., 2007. Informal transport: A global perspective. *Transp Policy* 14 (6), 445–457, doi:10.1016/j.tranpol.2007.04.011.
- Cervero, R.R., 2013. Linking urban transport and land use in developing countries. *J. Transp. Land Use* 6, 7–24, doi:10.5198/jtlu.v1.425.
- Dávila, J., 2013. Urban mobility and poverty: Lessons from Medellín and Soacha, Colombia (J. Dávila, Ed.). Retrieved from <http://discovery.ucl.ac.uk/1366633/>.
- Dimitriou, H., 2013. Transport Planning for Third World Cities (Routledge Revivals). Retrieved from <https://books.google.co.uk/books?hl=en&lr=&id=JkEareHgeOkC&oi=fnd&pg=PP1&dq=urban+transport+planning+dimitriou+developing+world&ots=XgX7akQcAt&sig=IDs6Xlmnf8J3M4Vb2G38pl1QMhC>.
- Dimitriou, H., Gakenheimer, R., 2011. Urban transport in the developing world: A handbook of policy and practice. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=09kM2SiGTwsC&oi=fnd&pg=PR1&dq=xavier+Godard,+2011+transportation+urban&ots=CwBhTBCQig&sig=u_xnvfhN04-071q0i9tFli-5BQ.
- Ehebrecht, D., Heinrichs, D., Lenz, B., 2018. Motorcycle-taxis in sub-Saharan Africa: Current knowledge, implications for the debate on “informal” transport and research needs. *J. Transp. Geogr.* 69, 242–256, doi:10.1016/j.jtrangeo.2018.05.006.
- Evans, J., O'Brien, J., Ch Ng, B., 2018. Towards a geography of informal transport: mobility, infrastructure and urban sustainability from the back of a motorbike. *Trans. Inst. Br. Geogr.* 43 (4), 674–688, doi:10.1111/tran.12239.
- Falavigna, C., Hernandez, D., 2016. Assessing inequalities on public transport affordability in two latin American cities: Montevideo (Uruguay) and Córdoba (Argentina). *Transp. Policy* 45, 145–155, doi:10.1016/j.tranpol.2015.09.011.
- Gwilliam, K., 2003. Urban transport in developing countries. *Transp. Rev.* 23 (2), 197–216, doi:10.1080/01441640309893.
- Harber, J., 2018. One hundred years of movement control: Labour (im)mobility and the South African political economy. In: Priya Uteng, T., Lucas, K. (Eds.), *Urban Mobilities in the Global South*. Routledge, Oxon, UK (Chapter 8).
- Harpham, T., Boateng, K.K.A., 1997. Urban governance in relation to the operation of urban services in developing countries. *Habitat Int.* 21 (1), 65–77, doi:10.1016/S0197-3975(96)00046-X.
- Henser, D.A., Button, K.J., 2000. *Handbook of Transport Modeling*. Pergamon/Elsevier, Amsterdam (NL).
- Hidalgo, D., Huizenga, C., 2013. Implementation of sustainable urban transport in Latin America. *Res. Transp. Econ.* 40 (1), 66–77, doi:10.1016/j.retrec.2012.06.034.
- Jenks, M.J., Burgess, M.J.R., Acioy, C., Allen, A., Barter, P.A., Brand, P., 2000. *Compact Cities: Sustainable Urban Forms for Developing Countries*. Taylor & Francis.
- Klopp, J., Cavoli, C., 2017. The Paratransit Puzzle: Minibus Mapping and Transportation Planning in Maputo and Nairobi. Retrieved from
- Lehmann, P., Brenck, M., Gebhardt, O., Schaller, S., Süßbauer, E., 2014. Barriers and opportunities for urban adaptation planning: analytical framework and evidence from cities in Latin America and Germany. *Mitig. Adapt. Strateg. Glob. Chang* 20 (1), 75–97, doi:10.1007/s11027-013-9480-0.
- Levy, C., 2013. Travel choice reframed: “deep distribution” and gender in urban transport. *Environment and Urbanization*. Retrieved from <http://eau.sagepub.com/content/early/2013/03/05/0956247813477810.abstract>.
- Midgley, P., 2011. Bicycle-sharing schemes: enhancing sustainable mobility in urban areas, United Nations. Department of Economic and Social Affairs 8, 1–12.
- Millard-Ball, A., Schipper, L., 2011. Are we reaching peak travel? Trends in passenger transport in eight industrialized countries. *Transp. Rev.* 31 (3), 357–378, doi:10.1080/01441647.2010.518291.
- Mohan, D., 2002. Road safety in less-motorized environments: future concerns. In *International Journal of Epidemiology*.
- Ortuzar, J. de D., Willumsen, L.G., 2011. Modelling Transport. In *Modelling Transport*. <https://doi.org/10.1002/9781119993308>.
- Oviedo, D., Levy, C., Dávila, J.D., 2017. Constructing wellbeing, deconstructing urban (im)mobilities in Abuja, Nigeria. In *Urban Mobilities in the Global South*, pp. 173–194. <https://doi.org/10.4324/9781315265094-10>.
- Peck, J., Theodore, N., 2001. Exporting workfare/importing welfare-to-work: exploring the politics of Third Way policy transfer. *Polit. Geogr.* 20 (4), 427–460.
- Peck, J., Theodore, N., 2010. Mobilizing policy: models, methods, and mutations. *Geoforum* 41 (2), 169–174.
- Pojani, D., Stead, D., 2015. Sustainable urban transport in the developing world: Beyond megacities. *Sustainability* 7 (6), 7784–7805, doi:10.3390/su7067784.
- Pojani, D., Stead, D. (Eds.), 2017. *The Urban Transport Crisis in Emerging Economies*, doi:10.1007/978-3-319-43851-1.
- Rynning, M.K., Priya Uteng, T., Lucas, K., 2018. Epilogue: Creating planning knowledge through dialogues between research and practice. In: Priya Uteng, T., Lucas, K. (Eds.), *Urban Mobilities in the Global South (Epilogue)*. Routledge, Oxon, UK.
- Schipper, L., Marie-Lilliu, C., Gorham, R., 2000. Flexing the Link between Transport and Greenhouse Gas Emissions - A Path for the World Bank.
- Shah, A., 2006. Corruption and decentralized public governance. World Bank Policy Research Working Paper. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=877331.
- Shoup, D., 2017. *The high cost of free parking: Updated edition*. Routledge.
- Stead, D., 2008. Effectiveness and acceptability of urban transport policies in Europe. *Int. J. Sustainable Transp.* 2 (1), 3–18, doi:10.1080/15568310701516614.
- Susilo, Y.O., Joewono, T.B., 2017. Indonesia. https://doi.org/10.1007/978-3-319-43851-1_6.
- Thibert, J., Osorio, G.A., 2014. Urban segregation and metropolitics in Latin America: The case of Bogotá Colombia. *Int. J. Urban Reg. Res.* 38 (4), 1319–1343, doi:10.1111/1468-2427.12021.
- UNCRD-IDB, 2011. Sustainable transport survey FTS Latin America. Retrieved from <http://www.uncrdlac.org/fts/>.
- Vasconcellos, E., 2014. Urban Transport Environment and Equity: The case for developing countries. Retrieved from <https://books.google.co.uk/books?hl=en&lr=&id=02B9AwAAQBAJ&oi=fnd&pg=PP1&dq=Urban+Transport,+Environment+and+Equity:+the+case+for+developing+countries+by+Eduardo+A+Vasconcellos&ots=qMUBomAfS-&sig=t8uSScMxwNA66avWjoLpcZWtCtgw>.
- Wegener, M., 2013. The future of mobility in cities: challenges for urban modeling. *Transp Policy* 29, 275–282, doi:10.1016/j.tranpol.2012.07.004.

Planning for Children's Independent Mobility

E. Owen D. Waygood*, Raktim Mitra†, *Transport Engineering, Department of Civil, Geotechnical, and Mining Engineering Polytechnique Montréal, Montreal, QC, Canada; †School of Urban and Regional Planning, Ryerson University, Toronto, ON, Canada

© 2021 Elsevier Ltd. All rights reserved.

Glossary	125
What is Children's Independent Mobility?	125
Why is it Necessary to Consider Children's Independent Mobility (CIM) in Planning?	125
What Influences Children's Independent Mobility (CIM)?	126
What is Urban Planning's role in CIM	127
Current and Future Considerations	128
A Need for Change in Planning Perspectives	129
Biographies	130
References	130
Further Reading	130

Glossary

Children's Independent Mobility (CIM) the freedom of a child to make trips or explore their neighborhoods on their own or with friends, without the accompaniment of an adult.

What is Children's Independent Mobility?

Children's independent mobility (commonly referred to as CIM in literature and practice) can be understood in a number of ways. Broadly, CIM is defined as the freedom of a child to make trips or explore their neighborhoods on their own or with friends, without the accompaniment of an adult. In more open terms, it relates to a situation where a child is *allowed* to make a trip or explore without an adult (called a "license"), whether they conduct such trips/activities or not. Typically, CIM is measured and studied in terms of making or having a license to walk/cycle/scoot without adult supervision to and from various destinations such as school, a friend's home, or parks. A child's independence can also be limited to certain conditions such as crossing at controlled intersections, not cycling on roads without segregated bicycle lanes, or not traveling at night; thus traffic-safety condition limits. Finally, with regard to transportation modes, particular focus is placed on a child's use of active travel modes (e.g., walking, scooting, cycling) and public or shared transport (e.g., bus, taxi, rail).

Why is it Necessary to Consider Children's Independent Mobility (CIM) in Planning?

CIM has a wide range of positive impacts for children (Waygood et al., 2017). These impacts can be organized with respect to their domain of well-being (Fig. 1). For children, CIM may affect their physical, social, psychological, cognitive, and economic well-being.

Physical well-being relates to benefits or harms that relate to the physical state of the child such as physical activity gained through active travel or exposure to outdoor air. Social well-being impacts relate to the social state of the child and include impacts such as the likelihood of a child meeting with friends or the child's community connections. Psychological well-being relates to the child's psychological state and relates to impacts such as travel satisfaction, life satisfaction, and happiness. Cognitive well-being is a sub-category of psychological well-being and generally relates to learning (formally or informally). Its impacts include increasing knowledge of one's living environment, being alert/awake and ready to learn. Finally, economic well-being relates to the economic state of the child and includes impacts such as the cost of travel or parental time use (as transport theory generally aims to reduce travel time of adults so that they can do productive work). Many of the benefits gained relate to active travel (e.g., physical activity, social interactions, satisfaction, being alert, low financial cost), though benefits also relate to public transportation options as these allow a greater range and thus increase the possible destinations a child could reach independently.

Children's independent mobility also relates to many current transport issues such as climate change emissions and energy use, road safety, transport costs, and congestion. An independently mobile child would use active travel modes (such as walking or cycling) and public transport for reaching various locations, and as a result, emissions are generally lower. Active travel modes do not create significant danger to others, and fatalities when traveling by public transport are extremely low. A child who is walking, bicycling, or scooting is a more vulnerable road user compared to those in motor vehicles, however, motor vehicle travel is

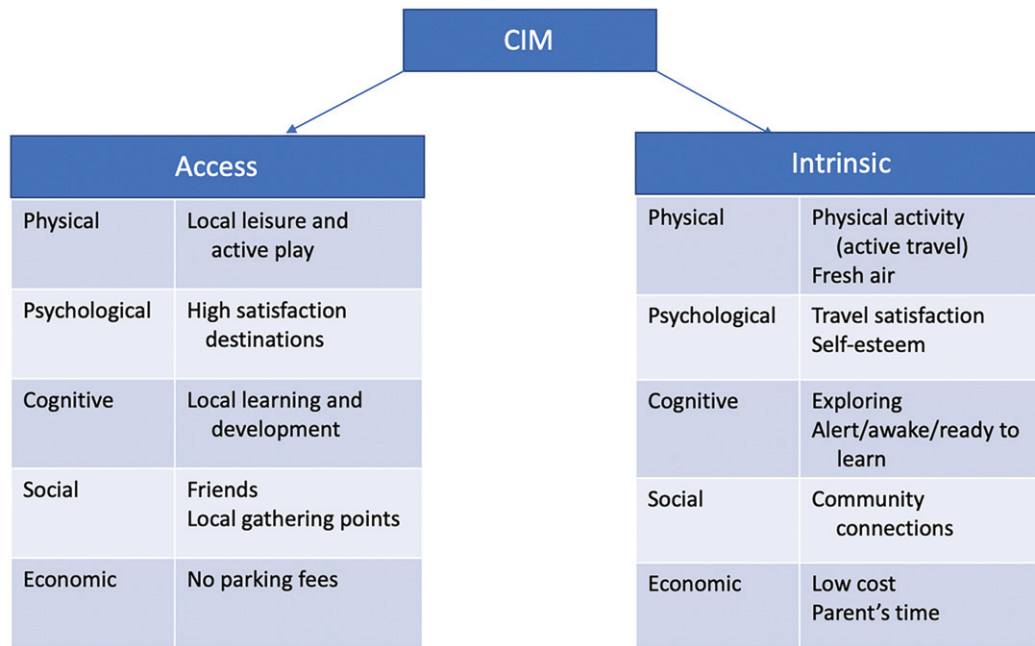


Figure 1 Conceptualization of how CIM relates to child well-being with examples.

intrinsically dangerous due to its speed for both the child in the vehicle and children outside the vehicle. In terms of transport costs, travel by active and public modes are less expensive than travel by car though people may not perceive it as such. Congestion is a problem of space, and public transport is the most efficient use of space followed by active modes. Finally, children who travel independently acquire the skills and familiarity with such modes through socialization and practice, and are thus more capable and likely to use the modes as adults.

Within planning, children can be considered as a marginalized group, as their inclusion and needs are not overtly taken into account. A lack of consideration to CIM has likely contributed to the modern problems of high chauffeuring burden and schools being located near heavy traffic. In contrast, planning for CIM creates travel contexts where most road users are able to travel safely and independently, including adults who do not use private motor vehicles (whether by choice or not). It also ties into the United Nation's Convention on the Rights of the Child ([United-Nations, 1989](#)), which states at various points that the child's survival, development, and well-being must be ensured. As CIM is positively linked to many such measures, its inclusion in planning is essential to protect the rights of children and develop cities for all.

What Influences Children's Independent Mobility (CIM)?

The influences on CIM can be best understood by examining a household's transportation decision-making process and the social and environmental contexts within which these decisions are made. A child's mobility outcome is the result of decisions around three key components, namely: the destination, the accompaniment (i.e., whether a child is traveling alone or with someone else, which could be other children or adults), and the transportation option or mode of travel. Adult household members and children negotiate these outcomes based on: (1) the importance of an activity (obligatory versus discretionary), (2) a child's perceived ability to independently travel and explore, and sometimes (3) the caregivers' availability and ability to facilitate a trip. Each household will make these decisions within different contexts, and these contexts will influence CIM outcomes. After several repeated occurrences, a transportation outcome may become habitual, unless any change in the social-ecological environment warrants a reassessment ([Mitra and Manaugh, 2019](#)).

A limited but emerging literature has examined various aspects of the social, physical, and policy environments that may influence a child's transportation outcome and consequently, CIM. These potential influences are summarized in [Fig. 2](#), using a social-ecological model framework. A child's physical and cognitive maturity as well as self-efficacy (i.e., their belief in themselves to conduct the trip without an adult) are important indicators of their license to independently travel and explore. A household's social and economic conditions and available transportation options will also influence travel behavior outcomes. Though in some countries girls' mobility are found to be more restricted, the relationship between gender and CIM are not conclusive. Age is often used as a proxy for a child's physical, cognitive and social development and experience, and it is generally expected that CIM increase with age. However, at what age CIM is allowed is found to vary by culture and context. Children in Japan and Finland are found to

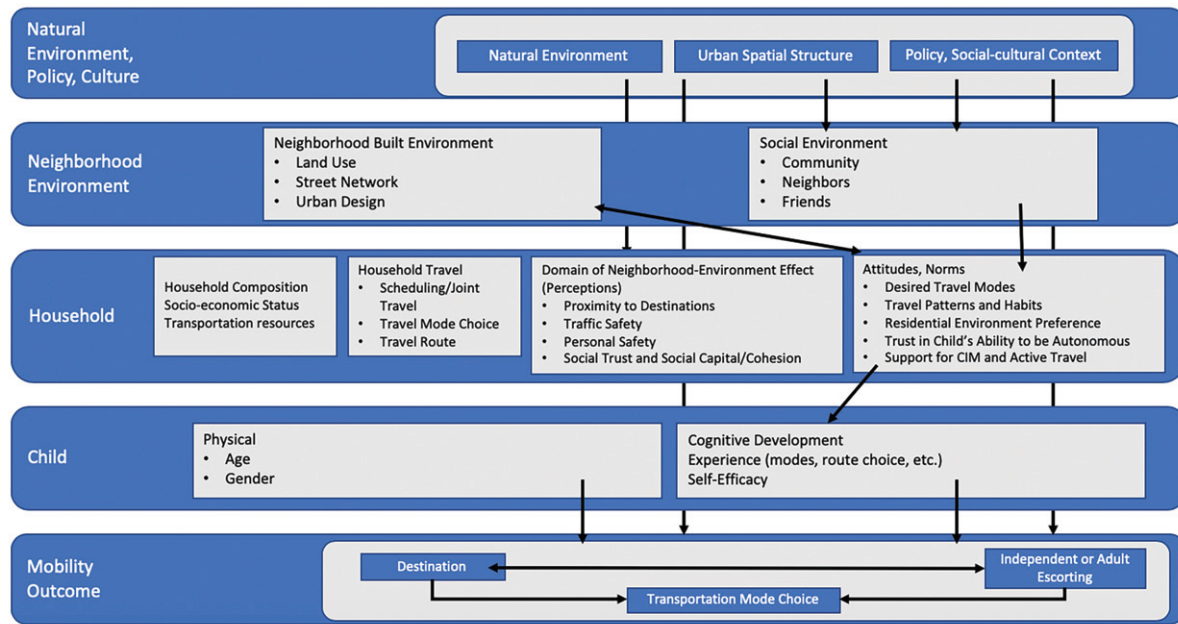


Figure 2 A social-ecological model of children's independent mobility and transportation outcomes. Source: Modified from Mitra and Manaugh (2019).

have considerable CIM at age 7, whereas in other countries such as the United Kingdom and the United States, CIM is not common until children are 10 or even older.

The built- and social-environment of a neighborhood may also influence a child's mobility outcomes by forming social expectations, travel-related attitudes, and perceptions of "walkable/bikeable" distance and safety. More broadly, the natural environment (e.g., topography, climatic conditions) and policy context (e.g., school selection and bussing policy) may also influence CIM, be it enabling or restricting CIM. Finally, all these potential influences may vary depending on the context within which a household is making their transportation decisions.

What is Urban Planning's role in CIM

Emerging evidence emphasizes the role of urban planning policy and practice on a child's mobility, and in particular, their independent mobility. In nearly all studies on CIM, distance between locations was identified as a key barrier to CIM, as most children would make these independent or unsupervised trips using active modes of travel, including walking and bicycling. As such, accessibility to both residential (e.g., a friend's house) and non-residential (e.g., school, playground) destinations is a key element of the "planned" built environment, that may enable or restrict CIM in urban, suburban and rural communities. Communities with a higher population density can support a greater diversity of local opportunities such as schools, playgrounds, and sport/leisure and cultural activities within a short distance. However, higher density development, sometimes called vertical neighborhoods, generally reduces residential space, availability of public spaces where children can meet and play, and also may result in crowded streets that are considered unsafe for a child. Partly as a result, families with children across North America and Europe have moved away from urban downtown locations which are actually accessible, diverse and typically enable CIM, to more suburban neighborhoods with larger residential space but limited opportunities to walk/cycle for a child. This reduction in CIM and reduced density of such developments can lead to lower levels of social and physical play, and higher traffic speeds along with vehicle dependence keep traffic danger high. Recognizing the apparent "absence" of children from downtown neighborhoods, some cities are investing in making their urban neighborhoods child-friendly. While more common in Europe (including Stockholm and Helsinki), North American cities such as Vancouver and Toronto have also developed guidelines to make families with children more "welcome" in vertical neighborhoods. Ensuring a balance between spaciousness, proximity, land use diversity, and traffic safety remains one of the key challenges in such efforts.

There are several critical challenges to address when planning communities for children. Parents obviously play a crucial role in CIM. Two key factors that emerge in studies on CIM are fear of traffic (traffic danger) and fear of strangers (personal security). Traffic danger appears to be the larger of the two concerns, and both are often found to increase with distance from destinations. Contemporary parenting culture emphasizes constant supervision of children and this approach propels some of these concerns. For example, some studies have shown that walking to school has declined over time, despite no changes in average travel distance to

school. In other words, the perception of what is considered a walkable and safe distance might have changed over the past decades across North America and Europe.

Another role that parents play pertains to attitudes and skills relating to CIM modes, including walking and bicycling. If a parent expects that a child will be able to travel independently, they will generally facilitate this by practicing with the child and helping them develop the necessary skills; also known as socializing. They may also choose a residential location that will be conducive to such travel (known as the residential self-selection). What is encouraging is that parental and children's attitudes can potentially be altered by systematic programming focusing on skills and experiences. Examples of such programming include walking school buses, bicycle training at schools, and doing class excursions by public transport. National walking school bus programs have existed in Japan for decades and have now become popular in other countries in the global north, though most implemented programs in western countries are ad-hoc and not coordinated across schools or school boards. Examples for safe bicycle training include Austria's *Verkehrsgarten* where children can train and develop their bicycle skills in a traffic-free zone. In the Netherlands, where the largest mode share for trips to elementary school is by bicycle, theory on general traffic rules use is given at school through programs such as Streetwise, and bicycle exams occur on the street.

With regard to the planning of the built environment, street design can facilitate or limit vehicular traffic, and as such will increase or limit traffic danger. For example, wider streets facilitate higher vehicular flow and speed and are thought to improve the safety of vehicle occupants, but they decrease the safety of other road network users. Higher speed traffic (above 30 km/h) increases the likelihood of a fatal collision when a child is involved in such an event. Restraining vehicle speeds through street design can increase CIM by not only reducing collision severity, but also by reducing the road crossing distance. Traffic engineering at controlled intersections, such as signal length (i.e., time allocated to pedestrians to cross a street during a green light) is also important. Previous standards for crossing times were roughly 1.2 m/s, based on physically capable adults, but a recommended speed to include most residents including children is 0.8 m/s (NACTO, 2013). With regard to street layouts, curvilinear roads, typical of many suburban developments, were originally designed to restrict vehicle movement by narrow lane width, reducing sight lines, and circuitous roads. However, traffic-engineering standards resulted in large vehicle lanes and long sight lines, which do not restrict vehicle movement. Circuitous roads limit through traffic, but without more direct active travel paths, result in longer walking/cycling distances between destinations even when the direct path distance may be short. Grid-patterns are easy to understand and thus navigate, but their scale will affect CIM.

In recognition of the importance of street safety, urban planners, traffic engineers and community organizations have adopted several approaches that may improve walking/bicycling to common destinations such as elementary schools, and by doing that, enable CIM. The most common approach is limiting vehicle speeds in school zones (often to 30 km/h). Programs such as Safe Routes to School (SR2S) in the US and School Travel Planning (STP) in Canada aim to improve the likelihood of active travel through education and infrastructure changes (National Center for Safe Routes to School, 2010). School travel corridors are another means of promoting active travel and CIM by limiting speeds on such routes and making infrastructure changes such as those mentioned earlier. Limiting through traffic to only local residents during school commute hours can reduce traffic danger in a more zonal approach (as opposed to just certain routes). This is applied in Japan where generally no through traffic is allowed within 500 m of elementary schools during school commute times.

The other components of built-environment planning—the density and diversity of land use—are also important for CIM. Population density by itself can increase the likelihood of desirable destinations whether that be playmates or a music teacher living within a reasonable CIM distance. However, without a mix in land uses, the diversity of possible destinations, and thus activities, is limited. CIM is found to be higher in areas with a greater diversity of land uses as they facilitate independent trips to schools, parks, libraries, recreation, shops, health services, and so on, which would all be located within close proximity of a child's home. Higher population density may also stimulate CIM as local play destinations are more likely to have other children present with whom the child can play, and thus provide the critical mass for group play.

Current and Future Considerations

Children's independent mobility, and children's mobility in general, has attracted much attention from public health in recent years. Transportation has long been known to be dangerous due to collisions, but more and more evidence demonstrates a wide range of negative impacts from sedentary lifestyles, emissions and noise. Collisions and sedentary lifestyles are perhaps more commonly discussed as collisions can result in fatalities and debilitating injuries, while sedentary lifestyles have been shown to relate to various physical health problems relating to body weight and cardio-vascular health. Exposure to vehicular emissions are increasingly linked to issues such as children's asthma, cancer, and even attention deficit disorder. Noise and vibrations are linked to sleep problems and cognitive ability (Waygood et al., 2017).

Over the past 2 decades, the Swedish approach to road safety, named "Vision Zero" has increasingly been adopted worldwide (Kristianssen et al., 2018; Ludvigsson et al., 2017). This approach differs from previous paradigms where many collisions were blamed on user error. Vision Zero takes the approach that the system must work to design out fatalities and serious injuries resulting from human error. As such, many residential streets are being redesigned/regulated to accommodate all users including children by restricting vehicle speed, and where speeds are above 30 km/h, by having protected infrastructure. Protected infrastructure for vulnerable users must not be seen as a cost of active travel, but a cost of high mobility as it is the source of the danger and must pay for its externalities.

As the new mobility options such as shared vehicles and autonomous vehicles continue to emerge, the role they will play in children's travel and their independent mobility is yet to be seen. For shared vehicles such as bicycles, the sizes are often too large for children under 12 and (bi-)laws may restrict their use, especially in the case of electric bicycles. Electric scooters also generally have similar restrictions in terms of age, and it is not uncommon for both electric vehicles to require helmets, which limits spontaneous/unplanned use. The use of such modes by children is not well studied, but anecdotal evidence suggests that teenagers do use them, and they do provide additional mobility options that could decrease reliance on high-cost personal motor vehicles.

With respect to autonomous vehicles, very limited research has been conducted on their potential impact on children. Proponents suggest that such vehicles would improve mobility for those without access to private motor vehicles, including children. However, in the case of children, parents would need to be confident with the safety of such systems and would likely need to authorize their dependents to use them. If safety standards are felt to be sufficient and costs were reasonable, it is not unforeseeable that families would use them, as it would facilitate children's travel to diverse destinations at greater distance without the time burden. If, like other common-use technologies, unlimited use service packages are created, this may lead to even further reductions in active travel, though it may be deemed independent. How society would react to such a vehicle killing a child is yet to be seen as well. People may accept human error, but it is not clear whether they would accept programming or equipment error on the part of autonomous vehicles. The overall impact is very difficult to anticipate, but it may result in a Wall-E type mobility future where sedentary lifestyles are completely facilitated.

On a more human side, the inclusion of children in planning is increasing. Although not common, guidance and examples are available, and outcomes support the argument that children's perspective differs from adults' and their inclusion in the planning process results in more inclusive design, making a more equitable system. Transportation planning is perhaps an area where the inclusion of children's needs is most limited and their voices, either directly or through advocates, is rarely taken into account. The basis of traditional transportation planning is value of time, which relates to a person's workforce contribution, which does not exist (for the most part) for children. The approach that planning for adults, and thus parents, includes planning for children raises questions of inter-generational equity and has perhaps led to the current situation where more and more children are dependent on adults for their travel as their needs are not properly considered.

A Need for Change in Planning Perspectives

Approaches to transportation planning are shifting from facilitating high-speed, high-flow traffic to enabling travel by all people (i.e., "basic mobility") and limiting negative health and economic externalities relating to human mobility. Public health considerations will increasingly help support the arguments for better planning for children's independent mobility. However, CIM is a multifaceted issue where addressing only one part of the problem will not likely result in significant changes. Improved infrastructure may make it possible for children to travel by active or public transportation options, but if their destinations are not accessible by those modes (due to distance or routes), only small gains may be achieved. As well, the advantages and disadvantages of separated versus (low speed) shared road spaces are not clear. If social norms related to children's travel do not shift away from an expectation of constant adult accompaniment for all daily trips, then improving safety and destination diversity may only produce small shifts in travel behavior. As such, urban planners and traffic engineers must look to build (either new or retrofit) infrastructure that accommodates safe active travel for all ages and skill levels, diversify destination options through mixed land use and supportive population densities, while also actively involving parents so that their habits and social norms shift over time. As such, improving children's independent mobility requires not only creative thinking in traffic and urban design, but also the adoption of inter-disciplinary approaches involving insights from the fields of sociology, psychology, and marketing. Law enforcement also plays a role, as it can help limit dangerous road user behavior. Ideally, street design would guide people to safe behavior, but this is not always possible. However, law enforcement may only occur at certain times if it is reliant on police presence. The role of automated enforcement, and its acceptance, is not well studied in the context of children's travel.

Although much has been learned about children's independent mobility or CIM in short, more remains to be explored. An exhaustive list is not possible here, but several avenues of investigation are proposed. For example, with regard to the benefits of CIM, while an emerging literature has focused on the relationship between CIM and physical health, the other domains of well-being are understudied. Our understanding of transportation infrastructure designs that are child-friendly is in its infancy, and the potential CIM gains as a result of such infrastructure is largely unknown. Finally, the development of planning tools to better facilitate the inclusion of children's independent mobility in our communities is badly needed. To address these issues, some have recommended including children in the transport planning process, however, in the current practice that has only been marginally addressed.

Biographies



E. Owen D. Waygood graduated in 2009 with a PhD in Civil Engineering from Kyoto University (Kyoto, Japan). After a position as a research associate and then research fellow at the Centre for Transport and Society at the University of the West of England (Bristol, United Kingdom), he held a position of Assistant and then Associate Professor of Transport Planning at Laval University (Quebec, Canada). In 2018, he was recruited by Polytechnique Montréal as an Associate Professor of Transport Engineering. He has published research on children's transport, physical activity, and social connections, sustainable transport, and transport behavior change. He has been a co-guest editor for special issues on transport and child wellbeing, and transport and wellbeing published with the journal *Travel Behaviour and Society*. He was the lead editor for the book, *Transport and Children's Wellbeing*, published in 2019. He conducts research on sustainable transport modes and how to increase their use.



Raktim Mitra is an Associate Professor of Urban and Regional Planning at Ryerson University (Toronto, Canada), where his research focuses on the intersection between the neighborhood built environment and transport behavior, and the impact on health outcomes such as physical activity and the quality of life. Mitra is the co-director of TransForm research laboratory of transport and land use planning at Ryerson University, co-chair of a paper review subcommittee (Bicycle Transportation) at the Transportation Research Board. His previous contributions include co-editing a special issue on the topic of transport and land use in childhood for the *Journal of Transport and Land Use*, and he is also the co-editor of the recently published book *Transport and Children's Wellbeing*.

References

- Kristianssen, A.-C., Andersson, R., Belin, M.-Å., Nilsen, P., 2018. Swedish vision zero policies for safety—a comparative policy content analysis. *Safety Sci.* 103, 260–269.
- Ludvigsson, J.F., Stiris, T., Del Torso, S., Mercier, J.-C., Valiulis, A., Hadjipanayis, A., 2017. European Academy of Paediatrics Statement: vision zero for child deaths in traffic accidents. *Eur. J. Pediatr.* 176 (2), 291–292.
- Mitra, R., Manaugh, K., 2019. A Social-Ecological Conceptualization of Children's Mobility. In: Waygood, E.O.D., Friman, M., Olsson, L.E., Mitra, R. (Eds.), *Transportation and Children's Well-being*. Elsevier, Amsterdam, pp. 81–100.
- NACTO, 2013. *Urban Street Design Guide*, National Association of City Transportation Officials. Island Press.
- National Center for Safe Routes to School, 2010. *Safe Routes to School: Case Studies From Around the Country*. National Center for Safe Routes to School, Chapel Hill, United States.
- United-Nations, 1989. *Convention on the Rights of the Child*, New York.
- Waygood, E.O.D., Friman, M., Olsson, L.E., Taniguchi, A., 2017. Transport and child well-being: an integrative review. *Travel Behav. Soc.* 9, 32–49.

Further Reading

- Christensen, P., O'Brien, M., 2003. *Children in the City: Home Neighbourhood and Community*. Routledge.
- Freeman, C., Tranter, P.J., 2011. *Children and Their Urban Environment: Changing Worlds*. Routledge.
- Hillman, M.(Ed.), 1993. *Children Transport and Quality of Life*, Policy Studies Institute, London, pp. 97.
- Hillman, M., Adams, J., Whitelegg, J., 1990. *One False Move . . . A Study of Children's Independent Mobility*. Policy Studies Institute, London.
- Gleeson, B., Sipe, N., 2006. *Creating Child Friendly Cities: Reinstating Kids in the City*. Routledge.
- Larouche, R.(Ed.), 2018. *Children's Active Transportation*, Amsterdam, Elsevier.
- O'Brien, C., Tranter, P.J., 2006. Planning for and with children and youth: insights from children about happiness, well-being and walking, Walk21-VII, "The Next Steps", The Seventh International Conference on Walking and Liveable Communities, Melbourne, Australia.
- Shaw, B., Bicket, M., Elliott, B., 2015. *Children's Independent Mobility: An International Comparison and Recommendations for Action*. Policy Studies Institute, London.
- Waygood, E.O.D., Friman, M., Olsson, L.E., Mitra, R., 2019. *Transport and Children's Well-being*. Elsevier.

The Politics of Mobility Policy

Geoff Vigar, School of Architecture Planning and Landscape, Newcastle University, Newcastle, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	131
Facilitating Infrastructure Supply Through Predict and Provide	131
Sustainable Mobility and Global Exemplars	132
Governance for Sustainable Mobility Policy	133
Conclusions	133
References	134
Further Reading	134

Introduction

It has been slowly revealed in planning and transport disciplines that far from acting in a singular public interest, professionals tend to deliver policies in line with their own experience. The modernist, universalizing principles that dominated thinking and policy in such disciplines in the second half of the 20th century in actuality proved to be highly damaging to many citizens' health and quality of life and resulted in poor outcomes for nature and the biosphere. The built reality of modernist transport planning also favored individualized transportation options. The subsequent dominance of cars in urban and rural landscapes was closely linked with ideas of personal freedom, which in planning practice meant a freedom to drive one's own car. The policy aim, not always explicitly stated, was to keep motor vehicle traffic moving at as high speed as was practicable.

The consumption of car use embodied a set of economic, emotional, cultural and esthetic responses described in Urry's idea of a "system of automobility" (Urry 2004). The rise to dominance of such a system is now revealed as not the politically neutral, "technical," response it might have been thought of at the time. Rather, automobility is a sustained political project underpinned by lobbying and vested interests and through the often unconscious bias inherent in practices enacted by professionals. Such practices frequently favor certain societal groups such as the able-bodied, men, professional classes, and majority ethnic populations. They also prioritize the human species over others and wider ecological conditions. In short transport policy in most nations and cities is socially and environmentally unjust. Flows of transport finance disproportionately benefit the already hypermobile: big infrastructure projects unduly favor car owners or those able to afford expensive rail fares and multiple flights a year. Lower income residents are typically disproportionately reliant on buses and on walking where public spending is inadequate and the quality of service is often poor (Lucas, 2004). Such residents are also disproportionately more likely to live in environments polluted by the highly mobile.

Transport disadvantage is not just related to poverty and low income. Disadvantage has a sociological dimension closely related to age, gender, and ethnicity. In terms of age, the young and the old tend to be disproportionately affected by the system of automobility. As countries around the world find themselves with increasingly ageing populations there are varying opinions on the policies that promote human flourishing among older people. While car use is often central to many older citizens, public transport, and good walking infrastructure is of equal if not greater importance, especially to those with fewer financial resources. Similarly children have long been seen as victims of policies designed to encourage the free flow of car traffic. A decline in child health, a rise in obesity, and asthma, alongside the psychological damage caused by the limiting of childhood freedoms known to previous generations, have all been highlighted in research dating back to the early 1970s (Hillman et al., 1990). As such, universal design principles that better acknowledge these systemic biases need to drive mainstream policy debate.

Finally, while there has been considerable advancement in ecological legislation around the globe in past decades, much contemporary policy is still anthropocentric. Policy and investment strategies are rarely strongly shaped by the highly damaging impacts to nonhuman species and local and global ecological conditions that arise from human hypermobility, air and car travel in particular.

Facilitating Infrastructure Supply Through Predict and Provide

Strategic policy directions, and the methods, education systems and practices that support such directions, tend to cohere into dominant paradigms or assemblages of practices, habits, theories, behaviors and artifacts. Such systems of meaning then rarely change dramatically: dominant paradigms emerge and become hard to displace. The 20th century saw the emergence and embedding of "predict and provide" as an overarching paradigm for transport planning, underpinned by neo-classical, "orthodox" thinking derived principally from the disciplines of economics and engineering (Vigar, 2002). Given the pressure to accommodate rises in car use in most nations, either among elites or mass populations, such a paradigm resulted in the provision of more road

space for private vehicles, and limited attention to public transport, cycling and walking. A neglect of alternatives to the private car was evident even while such modes were popular and thus this form of planning became a self-fulfilling prophecy as road space grew and other modes were made less attractive as they became starved of investment.

Historically, the politics of such predict and provide thinking was rarely considered by the planners, politicians and engineers implementing it. Catering for forecast increases in car ownership and use was simply good government and the political debate was typically at what speed and cost could road space be delivered to meet the projected growth. Similarly decline in public transport and cycle usage was predicted and systems were shrunk and not catered for, creating a self-fulfilling policy prophecy.

Slowly the politics of such choices was often revealed, at first through citizens mobilizing to protect urban spaces and neighborhoods from urban motorway construction in major cities like New York and London. Then, in many nations, the impact of interurban motorway building received sustained attention from those protecting the place qualities and ecology of rural areas. In part to bypass accusations of NIMBYism, such protests challenges the forecasting assumptions on which predict and provide was based. Latterly issues of climate change and the impacts on nonhuman species of facilitating private mobility became significant features in transport research and advocacy, if not always influential considerations in transport policy and decision-making.

Criticism of predict and provide methodologies often centered on the narrowness of techniques derived from economics and engineering disciplines. However, when it comes to the politics surrounding large infrastructure projects, as Bent Flyvbjerg's research has consistently shown, even narrow means-end rationality centered on economic cost-benefit tends to be ignored or willfully manipulated (Flyvbjerg, 2003). Here politics takes on its rawest form with decisions made with little regard to the outcomes of evaluation methods. To some extent this is a fair reflection as large-scale projects will have unforeseen and unknowable consequences given the time it takes to implement them and the many exogenous shocks to economic and other systems during that period. But the reality is that such initiatives are often the pet schemes promised to electorates and/ or those that are seen to carry large amounts of political capital, rather than projects that will improve the city for the citizens most in need or for ecological conditions.

At the other end of the spectrum much policy and decision-making is often carried on in a "path dependent" fashion with little political oversight. Here conservative transport policy communities tend to deliver largely what they delivered before and budgets tend to be tweaked rather than fundamentally reviewed. This has been highly problematic given the "new" policy goals of climate change, air quality and public health which demand very different policies to those that came before.

Thus, the "predict and provide" approach has proved remarkably resilient despite such challenges, and beyond a few alternatives it lives on in many nations and places well into the 21st century. However, much research suggests otherwise, its intuitive ideas of catering for growth and dealing with congestion by increasing infrastructure supply still resonates with many politicians and citizens in particular.

Sustainable Mobility and Global Exemplars

While the research evidence is clear that urban car use must be managed in some ways, it has proved hard to challenge the intuitive arguments underpinning predict and provide. There are also powerful vested interests perpetuating the system of automobility, whether capitalist - the spending on car advertising globally amounted to some US\$50bn^a in 2018 for example - or "hypermobile" affluent citizens who lobby to protect their lifestyles built around private vehicle use. We should also note the considerable lobbying effort worldwide directed at autonomous vehicle development which could go some ways to softening some of automobility's externalities, but in all likelihood is not going to tackle many of the social and environmental impacts of vehicle usage and could make some conditions worse (Pangbourne et al., 2020).

The most coherent paradigm mobilized to displace "predict and provide" is David Banister's sustainable mobility paradigm (SMP) which foregrounds environmental concerns but also acknowledges the deep inequities present in actually existing transport systems throughout most of the world (Banister, 2005). That said, some commentators emphasize the SMP's dependence on existing, off the shelf, solutions such as pedestrianization and argue that while these do have benefits, they rarely tackle the deep inequities present in mobility in the wider city (Kębłowski et al., 2018). Experiments in free public transport is one example of an alternative that might provoke more radical change and tackle wider injustices.

While a few such experiments are taking root across the globe, many towns and cities are struggling to implement planning and transport strategies that deliver more sustainable mobility practices. The embeddedness of the car in everyday life has made taking back road space for walking, cycling, and public transport prioritization difficult beyond central area pedestrianisations. Urry's "System of Automobility" is deeply embedded in many nations and aspired to, consciously or not, in others, supported by a huge capitalist infrastructure of advertising and government lobbying and in many nations hardwired into the thought worlds and practices of many professionals.

Where more sustainable mobility policies are arising, they are built not just on the long-standing debates highlighted above, but also on increasing evidence about the positive impacts of place making approaches from exemplar cities that have prioritized high-quality public spaces across the city not just downtown. These spaces can then be connected by high quality, safe walking, and cycle networks and affordable, efficient public transport. Urban development schemes in such exemplar cities, Bogota and Copenhagen for example, are often implemented with an emphasis on design quality, where a strong and confident city government has been a

^a <https://www.zenithmedia.com/wp-content/uploads/2019/03/Automotive-adspend-forecasts-2019-executive-summary.pdf>

constant feature, and sometimes with the consistent support of ruling governance regimes irrespective of the electoral cycle (e.g., Emanuel, 2019). In such places attempts to shape travel behavior go beyond central areas to retrofit existing neighborhoods to prioritize active travel and encourage people to see the street as public space and not a canal for car traffic. Critically, these packages of measures have demonstrated considerable economic benefits with many such cities being among the most successful on their respective continents.

Governance for Sustainable Mobility Policy

While economic issues are no doubt important, transport policy is often seen purely through the lens of its role in facilitating economic growth. Slowly the significance of transport policy in meeting climate and ecological targets as well as shaping the health of citizens, has begun to emerge. There is no doubt that research is playing an important role here, from increasing evidence of the impact of for example automobile traffic on human lung function in cities through exhaust emissions and tyre and braking systems, to transport's role in the climate emergency. Governance systems have generally been slow to adapt to such evidence and often favor light green, technocratic, approaches to deal with them, such as the promotion of electric vehicles.

But change can be rapid. The presence of a powerful city mayor can be one way that cities, such as London and New York, have been able to reasonably rapidly implement more sustainable policies, often where new mayors have been elected partly on the basis of changing the transport status quo. The key in reshaping policy direction in both these cities was to act quickly despite the opposition, to demonstrate that the city did not become grid locked when demand management measures and restrictions on car use in particular were implemented. Such measures have been most successful when complemented by affordable supply improvements, to show that the city worked better for most people than it did before (Sorenson et al., 2014).

An alternative to powerful mayoral action lies in more participative methods whereby citizens, through referenda or citizen juries or other deliberative methods, opt for such policies. There is contrasting evidence of the willingness of citizens to embrace elements of sustainable mobility where they see much of this agenda as a restriction on their current mobility practices centered on the use of private cars. But in neighborhoods where car traffic in particular might be dominant, citizens are more likely to opt for redesigned streets and public spaces than when imposed by "expert" planners and engineers. And this can be a "way-in" to lobbying for such policies at more strategic levels.

Such processes of co-design can help citizens imagine transitions to more sustainable mobility and more liveable urban futures. To realize such imaginings, creativity is needed at the center of any approach—a vision of what might be—as many of the global sustainable transport exemplars have emphasized. Design processes become research processes allowing for uncertainty in ways that traditional transport methods falsely seek to eliminate. Storytelling, making and enacting are vital to such processes using physical artifacts, sketches, 3D models, and virtual objects, all as provocations within the process as well as visions of the future. The temporary urbanism movement has shown how demonstrating a new vision through the use of traffic cones, straw bales and paint before implementing an expensive scheme is a powerful way of getting buy-in from businesses and citizens. Such open-ended and iterative approaches allow citizens to talk back to designs, informing and refining the proposals, standing in contrast to traditional linear processes of transport modeling, land-use planning, and for that matter traditional consultation. Temporary interventions are increasingly powerful tools to demonstrate that change is possible and that initial reservations among a majority of citizens may be misplaced.

The materiality of many such interventions can help dismantle established power relations, requiring engineers and planners to let go of some of their traditional forms of expertise. One role of engineers and planners in such strongly participatory process is to provide their knowledge of models and their limitations, of antecedent and precedent in terms of what might be possible, using the global exemplars we all know about and the research evidence about the links between transport behavior and economic growth, environmental impact, and social justice. But this is just one package of knowledge in a widely distributed urban intelligence. Such processes are not easy, but if cities are to "Copenhagenize," putting sustainability and place making at the center of their city visions, relying solely on traditional transport and urban planning approaches is unlikely to work (Colville-Andersen, 2018). Transitioning to a radically different mobility system requires demonstrations of what might be possible, both at the city scale and in neighborhoods, in the context of local circumstances and priorities, and with citizen representatives actively engaged in some form or another.

These approaches might be incremental, as in Copenhagen where car parks in the city center were restricted and closed slowly over time, while the city authorities went about building bike lanes and improving the pedestrian environment. Or it might be done at speed such as in Bogota in the late 1990s and early 2000s, driven by a mayor with radical ideas. Fundamentally both cities succeeded because there was a strong, well-articulated vision and desire for a better city. The willingness to experiment, to do things differently, was also important, as was political support for the project regardless of electoral or economic cycle. While there is a lot of talk of smart cities and the power of new technology, we should not forget that decades of top-down modernist planning prioritizing the car have left us in desperate need of ordinary things: benches, crossings, well-kept pavements, car free areas and cycle paths all greatly helping connect, strengthen and enliven communities. The global exemplars emphasize these needs strongly.

Conclusions

Buckminster Fuller asserted that change did not come from fighting the existing rationality but by building alternatives that render the existing obsolete. Movements such as Extinction Rebellion and lots of urban social movements are trying to do this and we see

some of this evident in the transport planning approaches of exemplar cities where demonstrating alternatives has been a key feature in facilitating more strategic change.

Some critical voices suggest that measures such as bike lanes and pedestrianization are the form of environmental gentrification, forcing the priorities of hipsters on less vocal majorities. But such experimentation can be grounded in broadly-based participatory democratic forms—as in the case of streets renaissance in New York and in Bogota—underpinned by local leadership capable of providing the framework within which such experimentation can emerge.

Many will also point to the rise of urban neoliberalism and the dominance of narrow forms of market rationality and of vested interests standing in the way of more equitable and just transport policy and spending. Considerable money is disbursed, and many are employed, in the transport realm often in a very path dependent way with little preconception of how budgets could be deployed differently. Indeed, considerable resources are wasted each year on defending policies and plans that have not been developed adequately with the participation of interested citizens and which lead to a great deal of protest. But global exemplar cities such as New York show how all sides of the political spectrum can get behind reclaiming streets from private motor vehicles and as Jane Jacobs noted in that City over 60 years ago, the urban street can be the most fundamental part of city life for citizens and we might do well to adopt this as the overarching principle for future urban transport planning.

References

- Banister, D., 2005. *Unsustainable Transport: City Transport in the 21st Century*. Routledge, London.
- Emanuel, M., 2019. Making a bicycle city: infrastructure and cycling in Copenhagen since 1880. *Urban Hist.* 46 (3), 493–517.
- Flyvbjerg, B., 2003. *Megaprojects and Risk*. Cambridge, CUP.
- Hillman, M., Adams, J., Whitelegg, J., 1990. *One False Move*. PSI, London.
- Kębłowski, W., Van Criekingen, M., Bassens, D., 2018. Moving past the sustainable perspectives on transport: An attempt to mobilise critical urban transport studies with the right to the city. *Trans. Policy* 81, 24–34.
- Lucas, K., 2004. *Running on Empty: Transport, social exclusion and environmental justice*. Policy Press, Bristol, UK.
- Pangbourne, K., Mladenovic, M., Stead, D., Milakis, D., 2020. Questioning mobility as a service: Unanticipated implications for society and governance. *Trans. Res. Part A* 131, 35–49.
- Sørensen, C.H., Isaksson, K., Macmillen, J., Åkerman, J., Kressler, F., 2014. Strategies to manage barriers in policy formation and implementation of road pricing packages. *Trans. Res. Part A* 60, 40–52.
- Urry, J., 2004. The 'system' of automobility. *Theo. Cult. Soc.* 21 (4–5), 25–39.
- Vigar, G., 2002. *The Politics of Mobility*. Spon, London.

Further Reading

- Banister, D., 2005. *Unsustainable Transport: City Transport in the 21st Century*. Routledge, London.
- Colville-Andersen, M., 2018. *Copenhagenize. The definitive guide to global bicycle urbanism*. Island Press, Washington, DC.
- Docherty, I., Shaw, J., 2019. *Transport Matters*. Bristol University Press, Bristol.
- Jacobs, J., 1961. *The death and life of great American cities*. New York, Random House.
- McArthur, J., 2018. The Production and Politics of Urban Knowledge: Contesting Transport in Auckland, New Zealand, *Urban Policy and Research*. doi:10.1080/08111146.2018.1476229.
- Parsons, R., Vigar, G., 2018. 'Resistance was futile!' Cycling's discourses of resistance to UK automobile modernism 1950-1970. *Plan. Perspec.* 33 (2), 163–183.
- Schiller, P.L., Kenworthy, J.R., 2018. *An Introduction to Sustainable Transportation*. Routledge, New York.
- Vigar, G., 2017. The four knowledges of transport planning: enacting a more communicative, trans-disciplinary policy and decision-making. *Trans. Policy* 58, 39–45.
- Vigar, G., Varna, G., 2019. *Connecting Places*. In: Docherty, I., Shaw, J. (Eds.), *Transport Matters*, BUP, Bristol.

Technology Enabled Data for Sustainable Transport Policy

Susan M. Grant-Muller*, Mahmoud Abdelrazek[†], Hannah Budnitz^{‡,§}, Caitlin D. Cottrill[¶], Fiona Crawford^{||}, Charisma F. Choudhury*, Teddy Cunningham**, Gillian Harrison*, Frances C. Hodgson*, Jinhyun Hong^{††}, Adam Martin*, Oliver O'Brien^{‡‡}, Claire Papaix^{§§}, Panagiotis Tsoleridis*, *University of Leeds, Leeds, West Yorkshire, United Kingdom; [†]Department of Civil, Environmental and Geomatic Engineering, University College London, London, United Kingdom; [‡]University of Birmingham, Edgbaston, Birmingham, United Kingdom; [§]Transport Studies Unit, University of Oxford, Oxford, United Kingdom; [¶]University of Aberdeen, Aberdeen, Scotland, United Kingdom; ^{||}University of the West of England, Bristol, United Kingdom; ^{**}University of Warwick, Warwick, United Kingdom; ^{††}Department of Urban Studies, University of Glasgow, Glasgow, Scotland, United Kingdom; ^{‡‡}Department of Geography, University College London, London, United Kingdom; ^{§§}Systems Management Strategy Department, University of Greenwich, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Glossary	135
Introduction	135
New Offerings for Policy Makers	136
Transport Network Modeling	136
Emerging Modeling and Analysis Methods	136
Cross-Sectoral Modeling	137
Positively Disrupting Travel Choice and Behavior	137
Data Challenges for Policy Makers	137
Missing Data and Bias in NEDF	137
Complementarity between NEDF and Traditional Mobility Data and Modeling	137
Skills and Engagement	138
The Role of Data Ecosystems	138
Collection and Access	138
Sharing, Use, and Privacy	139
Business Models	139
Summary of Opportunities and Challenges of Technology Enabled Data for Sustainable Transport Policy	139
Acknowledgment	140
See Also	140
References	140
Further Reading	141

Glossary

GDPR General Data Protection Regulations
NEDF New and Emerging Data Forms
CDR Call Data Records
GPS Global Positioning System

Introduction

Technology enabled data (as a subset of the broader big data landscape), have been highlighted as a strategic priority in policy agendas and national/international funding (EC, 2019; Lalawat, 2018; NITRD, 2016; NSFC, 2018). The UK Department for Transport has identified data and digital technologies “as a key enabler for innovation and a foundation for enabling the future of mobility”, whilst “improved transport data will allow the creation of innovative business models which have the potential to change the way people and goods are moved” (DfT, 2019). More broadly, the “rigorous analysis of high-quality data can enable citizens to hold governments to account, drive improvements in public services by informing choice, and feed innovation and growth across sectors by providing a robust evidence base for policymaking and practice” (ESRC, 2016). This article focuses on the potential of technology enabled data (hereafter referred to as “New and Emerging Data Forms” (NEDF)) considering the policy needs of the transport sector and sustainable mobility specifically.

NEDF include, but are not limited to (ESRC, 2016; OECD, 2013):

- *Internet data*, derived from social media and other web-enabled interactions (including data gathered by connected people and devices, e.g., mobile devices, sensors, wearable technology, Internet of Things),

- *Tracking data*, monitoring the movement of people and objects (including GPS/geolocation data, traffic and other transport sensor data, CCTV),
- *Satellite and aerial imagery* (e.g., Google Earth, Landsat, infrared, radar mapping)

Data need for sustainable transport policy may differ in one important respect from many other sectors. The efficient and sustainable operation of the transport sector and its ability to meet the mobility needs of society relies on understanding the unconstrained demand, dynamic interdependencies, physical interactions, and variety of *externalities* generated by multiple individual mobility choices. In summary, as individuals choose to travel and enter the transport system by whatever mode, their presence in the system impacts on others, on the functioning of the system and on the environment.

From an alternative perspective, many of the data need for mobility are clearly similar to those of other sectors (health, education, energy, and more), including the needs for privacy, data security, data quality, timeliness, and governance structures. The transport sector equally has a need for improved understanding of the value proposition (business model) for a range of actors in the sector, but chiefly, the value for members of the public who may generate the data.

A set of additional requirements for mobility data has been apparent despite a wealth of established data forms in the sector (large scale census, household surveys, embedded fixed-base sensors and more). The emergence of NEDF as a result of recent technology advances and utilized solely or in conjunction with more established data forms raises further questions about their contribution to the data need for sustainable mobility and policy makers.

New Offerings for Policy Makers

Transport Network Modeling

The modeling of individual movements, the state of the transport network and characteristics such as mode share have played a long-term and key role in the development of tactical and strategic transport policy. Highly refined models have been developed using traditional data sources, however with known issues due to the input data available. NEDFs have the potential to replace or enhance traditional transport data in creating high quality origin-destination matrices for transport modeling (Cuauhtemoc et al., 2017). Traditional transport data can be expensive to collect and may only offer a crosssection of travel patterns. In comparison, new forms of data are often collected automatically, digitally, at a high level of temporal granularity, and may be more comprehensive at a lower cost. More variable and customized outputs and more realistic representations of local transport supply and demand offer new opportunities to policy-makers as they appraise projects.

Data generated by mobile phone networks (call data records (CDRs) and network connections) have received particular attention, though other forms of technology enabled data, for example, from smartcards, social media, and GPS within mobile apps or vehicles, also enable model development (Harrison et al., 2020). Examples include detailed and comprehensive spatial data on cycling and walking patterns from Strava Metro data, tracking of passengers through stations through WiFi/Bluetooth Connections and testing short-term forecasting methodologies.

Other opportunities for established transport network modeling include:

- With repurposed data, the sample size may be considerably larger than could be the case if primary data were generated. It may also be possible to link the data of multiple travelers (e.g., friends traveling in groups).
- Many new data are dynamic and often available in real time. This makes it easier to update models, calibrate the models in real-time (or close to real-time), and provide timely analyses to inform current policy decisions.
- The continuous generation of data enables panel data to be constructed (i.e., long series of observations of the same individual). This can support analyses of within-individual changes and reduce the risk of bias inherent in cross-sectional analyses.
- Passively generated data may be more reliable than traditional sources such as surveys, which may exhibit response (or nonresponse) bias, arising from fatigue for instance.
- The combination (or linkage) of related available datasets to new forms of mobility data may help to overcome the lack of specific attributes needed for effective modeling and decision-making.

Emerging Modeling and Analysis Methods

New modeling and analysis methods are being adopted by the transport sector to harness new data. Research using NEDF, sometimes in combination with, or in comparison to, other data sources, has expanded from building and validating transport models to wider “smart cities” applications. These assess the intensity of activities over time and location, including during irregular events. Simulation models can test the impacts of policy options on outputs other than network efficiency (e.g., environmental impact), such as agent-based modeling or system dynamics modeling, and choice modeling to assess preferences of individuals. An approach gaining popularity for the analysis of large-scale databases in the transport sector is machine learning, part of the more general field of artificial intelligence. This comprises techniques able to uncover patterns inside a dataset (training set) which then use that knowledge to predict, or to recognize patterns, inside new datasets. Several statistical techniques and algorithms developed in computer science have been implemented for combining “redundant” sources of information, such as Bayesian inference, artificial neural networks, fuzzy logic and Kalman filters. Examples of such type of data sources are the traditional inductive loop detectors and the emerging sensor data, measuring the traffic flows in a road network.

Cross-Sectoral Modeling

There is an increasing need amongst policy makers at all levels to develop cross-sectoral policies, which consider and satisfy objectives from multiple spheres, including public health, environment, land use, and social equity. Transport traverses all these sectors so technology enabled data for sustainable transport policy has significant influence on the appraisal and implementation of cross-sectoral aims. Existing practices from areas such as healthcare and the finance industry may provide some guidance for the treatment of sensitive data, particularly as location identifiers are now considered “personal data” under General Data Protection Regulation (GDPR) guidance. From the vantage of healthcare data protection and public trust are key concerns, whereas research and emerging practices in financial technology may also provide valuable insights into the regulation and use of data. Early collaborative engagement of specialists/policy makers in these areas, each with their own objectives can lead to a true cross-sector model rather than an “adapted” single sector model. Furthermore, there is also the potential to detect previously unconsidered mutual benefits or trade-offs using the new forms of data.

Positively Disrupting Travel Choice and Behavior

There is a substantial body of work dedicated towards using NEDF to analyze and better understand the route and mode choice of travelers, which can lead to more effective policy decisions. This includes the use of GPS data to better understand route choice, CCTV and Wi-Fi data to map crowds and Bluetooth data to classify types of road user. Research has also focused on how to use the information to subsequently influence planning and policy decisions. An example of a potential application is to use knowledge of anticipated congestion at stations (from passenger tracking) to encourage users to travel via another route or at another time ([Transport for London, 2017](#)).

Although these NEDF have been instrumental in encouraging positive travel choice change through better understanding of travel patterns, there is a need for more research into whether (and if so, how) these data forms can be used to directly and actively influence travel choice and behavior. Furthermore, there is currently a paucity of research into whether these NEDFs are causing travel behavior change, as opposed to merely facilitating it through allowing researchers and practitioners to have access to more varied datasets.

Data Challenges for Policy Makers

Missing Data and Bias in NEDF

Most NEDF are not primarily collected as mobility data and therefore they often suffer from biases and missing data, including people who are missing from the data entirely, and missing or inaccurate information on individual trips (or other variables, such as gender). Social media data, for example, has value for the transport sector ([Grant-Muller et al., 2015](#)) but may have been generated by a small number of “superusers” ([CEDR, 2017](#)). Whilst nearly 90% of the population of Europe currently have access to a mobile phone, the majority of which are smartphones ([GSMA, 2018](#)), not everyone has a mobile phone or uses it at the same rate. It is worth noting that smartphone location data refers to the movement of the device, which may be independent of the movement of the phone owner or user. In most cases, however, research on NEDFs acknowledges bias in the data but does not account for it. This is problematic since sampling bias and a lack of demographic information are key barriers to the wider use of NEDFs by practitioners and policy makers ([Lee et al., 2016](#)).

Many traditional data collection activities such as household travel surveys are designed to ensure (or at least aim for) area coverage and general population representation. While representation has been increasingly difficult to achieve ([Bonnell et al., 2015](#)), many new data sources (such as Strava Metro data, travel times and speeds from Uber Movement, or event data sourced through Twitter) may be more specifically limited to certain segments of the population. Consequently there is a risk of a biased or incomplete data picture if used to infer beyond their original context. This is especially problematic in policy making if it is to be holistic and equitable. The representativeness of the population covered by NEDFs varies across data types and may depend on app use (e.g., Instagram, Twitter) or device usage (e.g., mobile phone, desktop) which varies across age groups and socioeconomic groups ([Ofcom, 2019](#)). Providing wearable GPS devices to all participants can help to reduce the risk of bias associated with smartphone ownership ([Rasmussen et al., 2015](#)), and app usage, although the devices can be challenging to use for some older people ([Schmidt et al., 2019](#)).

Other issues relate to:

- Missing trip characteristics such as trip purpose or mode.
- With repurposed data the variables that are generated in the dataset cannot be determined by the end users.
- In combining new data sources to enhance modeling, careful attention must be paid to the ways in which data are linked and any underlying biases or errors that may be propagated.

Many areas, therefore, require further research, but the most pressing matter is likely to be the lack of sociodemographic information. Without this information it is not possible to identify which social categories and population segments are missing from the data or even attempt to address biases.

Complementarity between NEDF and Traditional Mobility Data and Modeling

In the short and medium term, the use of NEDF to augment traditional data sources by reducing known biases should be explored (Misra et al., 2014) and an effective way to utilize NEDFs in the future is likely to involve combining multiple forms of data. This approach creates added value from the differing attributes each source may have. For example, supplementing or replacing bespoke instruments for Household Travel Surveys with data collected through other means shows great promise for the transport sector. In such cases, the complementary use of different sources has the potential to add important information for under-represented groups (Birkin, 2018) and missing variables. Conversely, survey data can directly capture important behavioral information of a small sample, while emerging data sources can provide continuous mobility observations (Kusakabe and Asakura, 2014).

Even where it is not possible to link data on the same individuals in different datasets, the new forms of data may provide important small-area geospatial information to enhance existing individual-level datasets. Data fusion techniques can be applied with different independent sources of information to provide a better estimate of the parameters included in the system analyzed (Zhao et al., 2015). CDRs are one type of new mobility data form from a wider group of large-scale mobility data. Social network data provides a further opportunity to enrich mobility data. Shared similarities in mobility patterns, objectives, and mode choices serve as an input to optimal transport design and sustainable transport policy.

The integration of new data sources with main stream modeling does not necessarily imply that methodological barriers exist. An example where integration has taken place is the fusion of aggregate census data and disaggregate mobility data from sensors using an agent-based modeling framework, specifically an integral part known as population synthesis (Beckman et al., 1996). The result of that process is to obtain individuals with added semantic information important for policy-oriented transport modeling but previously only included at aggregate level in the census. This is coupled with high-resolution spatio-temporal mobility information from the raw sensing data. For examples see Wu et al. (2014) and Ali et al. (2016).

Skills and Engagement

Despite the wide policy implications of NEDF research and an increased willingness for data owners to consider how NEDF could be used for policy, a lack of sufficient academic engagement with policy-makers can be a barrier to their application (Marsden and Reardon, 2017). Most of the skills needed to optimize the use of NEDF are currently focused in academia, whilst stakeholders also need to identify other resources required to realize the data potential. Thus, a key gap is directing sufficient energy to responding to policy needs and disseminating the policy implications of research on NEDF to inform decision-making.

Emerging datasets in the form of unstructured data, such as textual (e.g., from social media) or video (from CCTV cameras in the road network) require different analysis tools. Text analytics is gaining in popularity as social media usage increases. Specifically, Natural Language Processing, a text analytics set of algorithms, provides the ability to summaries important latent semantic information in a specific text. Video analytics, on the other hand, can be carried out in an automatic way with the use of image processing techniques, such as convolutional neural networks.

In summary, the opportunities provided by new data have resulted in the development and adoption of new methods into transport modeling and policy-making. Most of the new skills required originate in fields outside transport, including statistics, computer science, economics and epidemiology, but these experts lack the necessary theoretical or policy background. The transition period for both groups can be challenging, but of paramount importance for the successful cooperation between them.

Other implications include:

- Data processing and model development costs may be higher due, for example, to data cleaning issues or lack of skills/resources.
- A challenge in leveraging the advantage of dynamism in the data lies in the ability to process and conduct computations in real time.
- The opportunity to further evolve transport models may give rise to methodological or skills-based challenges.

The Role of Data Ecosystems

When considering the use of new forms of data for sustainable transport policy, it is critical to address issues arising from the ecosystem of data collection, access, sharing, use, and archiving. Practices associated with these considerations will vary depending on a number of factors, including the status of the collecting agency (public, private, or third-sector), the type of data being collected, the sensitivity of the relevant data attributes (for example, are the data considered “personal” under the EUs GDPR), and their intended uses and users. As data resources continue to evolve, it will be necessary to ensure that the practices being followed in their use are compatible with legal and regulatory requirements; acceptable from a societal standpoint; adequately and accurately recorded; and future-proofed for ongoing value.

Collection and Access

Many traditional data collection activities are conducted in-house or via contractual agreements, with the contracting organization having a degree of agency over both the attributes and format of data collected. In the emerging mobility data ecosystem, however, it is not uncommon for data to be obtained or purchased from third-party vendors or open data sources. In such instances, issues of data standardization, formatting, and the practices followed when collecting (including ethical considerations) may be of concern.

Obtaining consistent data over time may be particularly difficult. Ensuring that appropriate metadata is available will be necessary for incorporation of such data resources into current practices, to allow attributes to be accurately represented in the data ecosystem. An associated concern is data ownership, as commercial considerations may preclude the sharing of “raw” data in favor of data that have been aggregated or otherwise modified to both reduce the potential disclosure of sensitive data and maintain the value of such data in commercial applications.

Sharing, Use, and Privacy

While many new forms of mobility data have been posited for incorporation into transport modeling and planning processes, they may also be used for more immediate decision making on the part of travelers or transport service organizations. Twitter data, for example, has been explored for use in the identification of traffic events and disruptions (Gu et al., 2016) as well as information provision to the traveling public (Cottrill et al., 2017). Services and apps such as Waze, Google Maps, and Moovit, on the other hand, make use of crowd-sourced data (or volunteered geographic information – VGI) to provide real-time information on public transport routes and traffic conditions to travelers and decision makers (Attard et al., 2016).

Emerging privacy issues can affect the potential value of new forms of mobility data. For example, many new datasets (e.g., Strava, Twitter, etc.) provide location information that could be highly disclosive and contravene the GDPR. Research using CDR data has already demonstrated the uniqueness of mobility traces (De Montjoye et al., 2013). To safeguard their commercial interests as well as their users’ privacy, many third-party vendors have adopted new privacy controls and changed the characteristics of their data products. The changes seen in the granularity of Strava data products for example could happen to other new forms of mobility data in the near future. Constructive conversation and clear understanding about the value of these new data between researchers (planners) and third-party vendors are necessary to ensure the quality of data and its value. Research should also continue into methods to increase privacy whilst minimizing data loss, for example Gao et al. (2019).

Business Models

The largely innovation-driven emergence of micro-mobility modes has disrupted established governance practices and highlighted issues around the commodification of data. City stakeholders, commercial entities and the individual all have a claim to “rights” concerning the location data generated. The balancing of the “public good” alongside those rights and the commercial imperatives has become contentious. New data for mobility stimulate the development of new business models. New business models mean that there are opportunities to generate new value; however, the development of a new business model requires explicit choices about the value proposition and choices about the importance of relative values for groups of different users. A summary of the opportunities and challenges for NEDF in contributing to sustainable policy at broad level is provided in Table 1.

Table 1 Opportunities and challenges for NEDF in contributing to sustainable policy

<i>Sustainable transport policy objective</i>	<i>Opportunities</i>	<i>Challenges</i>
Economy	• New types of business models related to data ownership/generation may arise • Real-time calibration and analysis supports efficiency in smart networks, including improved travel times. • Opportunity to develop more cost-efficient cross-sectoral models • More cost-effective to obtain larger and longer-term datasets than traditional methods, especially if secondary data.	• Increased model development and policy appraisal costs as new skills and resources required • Maintenance costs of monitoring smart networks • Availability and costs of commodified data
Environment	• More finely grained understanding of travel behaviors can lead to the design of more efficient networks and better land-use • Combining with environmental sensor information can identify and monitor areas of high pollution • Real-time calibration and analysis can contribute to more reactive and resilient (future-proof) smart networks	• Energy and emissions associated with data generation, processing and storage
Social	• More reliable long-term data of larger groups of individuals • Data value propositions allow individuals ownership of their own data • Potential to develop policies that can influence individual travel choices • Potential to target specific segments / sub-groups	• Representativeness of the population and sample bias may accentuate hitherto neglected segments of the population • Lack of socio-demographic information • Raises issues of data governance, especially concerned with individual privacy

Summary of Opportunities and Challenges of Technology Enabled Data for Sustainable Transport Policy

In conclusion, whilst the data needs for sustainable transport policy are complex and differ from those of many other sectors, NEDF enable profound new insights into human choices around mobility. Whilst these new data forms bring new challenges, especially in terms of sharing and privacy, the value proposition for the various stakeholders involved is increasingly recognized. Given the cross-sectoral and re-purposed nature of many NEDF, the most significant challenge may be a closer working of policy decision-making bodies to capitalize on the potential that NEDF offer.

Acknowledgment

This note is coauthored by contributors and participants at a workshop sponsored by the Alan Turing Institute, the UK's national centre for Data Analytics and AI on 26/6/19, led by Prof. Susan Grant-Muller, Fellow of the Alan Turing Institute. In forming the research questions and framing the key points above, we have cross-referenced [OECD \(2013\)](#).

See Also

Mobility as a Service; Demand Responsive Transport; Car Sharing; Connected and Autonomous Vehicles: Priorities for Policy and Planning; Modeling and Simulation for Transport Planning; ITS for Transport Planning and Policies; Sensors and Data Driven Approaches in Transport

References

- Ali, A., Kim, J., Lee, S., 2016. Travel behaviour analysis using smart card data. *KSCE J. Civil Eng.* 20 (4), 1532–1539.
- Attard, M., Haklay, M., Capineri, C., 2016. The potential of Volunteered Geographic Information (VGI) in future transport systems. *Urban Plan.* 1 (4), 6–19.
- Beckman, R.J., Baggerly, K.A., McKay, M.D., 1996. Creating synthetic baseline populations. *Transport. Res. Part A* 30 (6), 415–429.
- Birkin, M., 2018. Big data: big data for social science research (Ubiquity symposium). *Ubiquity* 2018, 1–7.
- Bonnel, P., Bayart, C., Smith, B., 2015. Workshop synthesis: comparing and combining survey modes. *Transport. Res. Proc.* 11, 108–117.
- CEDR, 2017. Traffic Management Methods, Contractor Report 2017-04. Link: https://www.cedr.eu/download/Publications/2017/CEDR-Contractor-Report-2017-04_Call-2013-Traffic-Management-End-of-Programme-Report.pdf [accessed 29 Jul. 2019].
- Cottrill, C., Gault, P., Yeboah, G., Nelson, J.D., Anable, J., Budd, T., 2017. Tweeting Transit: An examination of social media strategies for transport information management during a large event. *Transport. Res. Part C: Emerging Technol.* 77, 421–432.
- Cuauhtemoc, A., Erath, A., Fourie, P.J., 2017. Transport modelling in the age of big data. *Int. J. Urban Sci.* 21, 19–42.
- De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the Crowd: the privacy bounds of human mobility. *Scientific Rep.* 3 (1376), 1–5.
- DfT, 2019. Areas of Research Interest 2019. Research and Evidence. In: Department For Transport, H.G., Uk, (ed.). Available online from: <https://www.gov.uk/government/publications/dft-areas-of-research-interest> (accessed 03/09/19).
- EC, 2019. Big Data [Online]. Available: <https://ec.europa.eu/digital-singlemarket/en/policies/big-data> [Accessed 2nd January 2019].
- ESRC, 2016. <https://esrc.ukri.org/files/funding/funding-opportunities/new-and-emerging-forms-of-data-policy-demonstrator-projects-call-spec/> [Accessed 2nd January 2019].
- Gao, J., Sun, L., Cai, M., 2019. Quantifying privacy vulnerability of individual mobility traces: A case study of license plate recognition data. *Transport. Res. Part C: Emerging Technol.* 104, 78–94.
- Grant-Muller, S.M., Gal-Tzur, A., Minkov, E., Kuflik, T., Nocera, S., Shoor, I., 2015. In: *Transport policy: Social media and user-generated content in a changing information paradigm*. In: Social media for government services. Springer, Cham, pp. 325–366.
- GSMA, 2018. The Mobile Economy: Europe 2018. London, UK. Available from: <https://www.gsma.com/r/mobileeconomy/europe/>.
- Gu, Y., Qian, Z.S., Chen, F., 2016. From Twitter to detector: real-time traffic incident detection using social media data. *Transport. Res. Part C: Emerging Technol.* 67, 321–342.
- Harrison, G., Grant-Muller, S.M., Hodgson, F.C., 2020. New and emerging data forms in transportation planning and policy: Opportunities and challenges for “Track and Trace” data. *Transport. Res. Part C: Emerging Technol.* 117, 102672.
- Kusakabe, T., Asakura, Y., 2014. Behavioural data mining of transit smart card data: a data fusion approach. *Transport. Res. Part C: Emerging Technol.* 46, 179–191.
- Lalawat, P., 2018. How Indian Government is using Big Data analytics to improve economy and public policy [Online]. <https://www.analyticsinsight.net/howindian-government-is-using-big-data-analytics-to-improve-economy-and-publicpolicy/> [Accessed 23rd January 2019].
- Lee, R.J., Sener, I.N., Mullins III, J.A., 2016. An evaluation of emerging data collection technologies for travel demand modelling: from research to practice. *Transport. Lett.* 8 (4), 181–193.
- Marsden, G., Reardon, L., 2017. Questions of governance: rethinking the study of transportation policy. *Transport. Res. Part A: Policy Pract.* 101, 238–251.
- Misra, A., Gooze, A., Watkins, K., Asad, M., Le Dantec, C.A., 2014. Crowdsourcing and its application to transportation data collection and management. *Transport. Res. Rec.* 2414 (1), 1–8.
- NITRD, 2016. The Federal Big Data Research and Development Strategic Plan. In: EXECUTIVE OFFICE OF THE PRESIDENT, N.S.A.T.C., (ed.). The Networking and Information Technology Research and Development Program. Online: <https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf> [Accessed 7th January 2019].
- NSFC, 2018. National Science Fund Guide to Programs 2018. Online: http://www.nsf.gov.cn/english/site_1/pdf/NationalNaturalScienceFundGuidetoPrograms2018.pdf [Accessed 2nd January 2019].
- OECD, 2013. New data for Understanding the Human Condition, [available on-line] <http://www.oecd.org/sti/inno/new-data-for-understanding-the-human-condition.pdf> [Accessed 15 Aug. 2019].
- Ofcom, 2019. Adults;1; Media use and attitudes report 2019. https://www.ofcom.org.uk/_data/assets/pdf_file/0021/149124/adults-media-use-and-attitudes-report.pdf [Accessed 29 Jul. 2019].
- Rasmussen, T.K., Ingvarsson, J.B., Halldórsdóttir, K., Nielsen, O.A., 2015. Improved methods to deduct trip legs and mode from travel surveys using wearable GPS devices: a case study from the Greater Copenhagen area. *Comput. Environ. Urban Syst.* 54, 301–313.
- Schmidt, T., Kerr, J., Kestens, Y., Schipperijn, J., 2019. Challenges in using wearable GPS devices in low-income older adults: Can map-based interviews help with assessments of mobility? *Transl. Behav. Med.* 9 (1), 99–109.
- Transport for London, 2017. Review of the TfL WiFi pilot. [online] London. Available at: <http://content.tfl.gov.uk/review-tfl-wifi-pilot.pdf> [Accessed 18 Jul. 2019].

- Wu, L., Zhi, Y., Sui, Z., Liu, Y., 2014. Intra-urban human mobility and activity transition: evidence from social media check-in data. *PLoS ONE* 9 (5), e97010.
- Zhao, Y., MacKinnon, D.J., Gallup, S.P., 2015. Big Data and Deep Learning for Understanding DoD Data. *CrossTalk*. <http://static1.1.sqspcdn.com/static/1/702523/26357333/1435733758490/201507-Zhao.pdf?token=jqwUi1OU3pCs4L5K9NTgjFY5VBQ%3D> [Accessed 18 Jul. 2020].

Further Reading

- Bwambale, A., Choudhury, C.F., Hess, S., 2019. Modelling trip generation using mobile phone data: a latent demographics approach. *J. Transport Geogr.* 76, 276–286.
- Crawford, F., Watling, D., Connors, R., 2018. Identifying road user classes based on repeated trip behaviour using Bluetooth data. *Transport. Res. Part A: Policy Pract.* 113, 55–74.
- Grant-Muller, S., Hodgson, F., Malleson, N., Snowball, R. (2017, June). Enhancing Energy, Health and Security Policy by Extracting, Enriching and Interfacing Next Generation Data in the Transport Domain (A Study on the Use of Big Data in Cross-Sectoral Policy Development). In *2017 IEEE International Congress on Big Data (BigData Congress)*, pp. 515–518. IEEE.
- Lee, R.J., Sener, I.N., Mullins III, J.A., 2016. An evaluation of emerging data collection technologies for travel demand modelling: from research to practice. *Transport. Lett.* 8 (4), 181–193.
- Nitsche, P., Widhalm, P., Breuss, S., Brändle, N., Maurer, P., 2014. Supporting large-scale travel surveys with smartphones – A practical approach. *Transport. Res. Part C: Emerging Technol.* 43, 212–221.
- Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: travel demand estimation using big data resources. *Transport. Res. Part C: Emerging Technol.* 58, 162–177.
- Wang, Z., He, S., Leung, Y., 2018. Applying mobile phone data to travel behaviour research: a literature review. *Travel Behav. Soc.* 11, 141–155.

Car Sharing

Cyriac George, Tanu Priya Uteng, Department of Mobility and Organisation, Institute of Transport Economics (TOI), Oslo, Norway

© 2021 Elsevier Ltd. All rights reserved.

Introduction	142
What is Car Sharing?	142
Types of Car Sharing	143
Types of CS by Business Model	143
Types of CS by Operational Model	143
Impacts of Car Sharing	144
Implications for Policy and Planning	144
Further Research	145
References	145
Further Reading	146

Introduction

Car sharing (CS) is an emerging mobility practice that emphasizes access to a vehicle over ownership. CS platforms allow users to rent vehicles on a self-service basis for short periods of time. The first viable CS platforms were established in Switzerland and Germany in the 1980s, after which similar attempts were made in dozens of urban contexts in North America and Europe. The earliest platforms were member-owned cooperatives, which were followed up by commercial [CS enterprises](#) starting in the mid-1990s and peer-to-peer (P2P) platforms in the 2010s.

The most common ways to differentiate between the various types of CS are based on business and operational models. CS is also part of the broader sharing economy and is one among many new forms of mobility that challenge the dominance of privately owned vehicles in contemporary transport behavior and planning. Although CS does not rely on radically new technologies, its novel use of existing technology has potentially broad implications with respect to mobility, urban development and the environment.

What is Car Sharing?

CS refers to the mobility practice whereby members of a platform have rental access to cars on a short-term self-service basis through a formal service provider or intermediary. A more in-depth understanding of this definition can be gained by considering the distinctions between (1) formal and informal car sharing, and (2) car sharing and other types of mobility that emphasize vehicle access over ownership.

First, people have been sharing cars for as long as cars have been around. Such informal types of CS can be as simple as friends or family members borrowing a car from one another. Scholarly analysis of CS focuses almost exclusively on the formal variety. “The distinctive criterion is whether a central service structure exists that coordinates the activities of multiple users of a car and whether any legal form of association exists that own the cars. In private car sharing, it is usual for one person to hold legal ownership rights to the car, and access to the car is organized in an informal way” (Truffer, 2003, p. 154).

Second, when asking what CS is, it is worthwhile to consider what CS is not. Most importantly, CS is often confused with other forms of shared mobility that emphasize vehicle access over ownership, namely ride sharing and ride sourcing. The clearest distinction is that CS offers access to a vehicle that is operated by the user whereas ride sharing and ride sourcing offer access, as a passenger, to a vehicle that is operated by someone else. The Shared Use Mobility Center provides the following definitions (Shared Use Mobility Center Sumc, 2015[SUMC], 2016, p. 5-8):

- Car sharing: “a service that provides members with access to an automobile for short-term—usually hourly—use”.
- Ride sourcing: “platforms to connect passengers with drivers who use personal, noncommercial, vehicles”.
- Ride sharing: “adding additional passengers to a preexisting trip...Unlike ride sourcing, ride sharing drivers are not ‘for-hire’”.

Some definitions of CS provide general descriptions of the practice but leave room for ambiguity with respect to these and other related types of mobility. For example, according to Shaheen et al. (1998, p. 35) “The principle of carsharing is simple: individuals gain the benefits of private cars without the costs and responsibilities of ownership.” While this description is true, it is more a normative commentary on CS rather than a definition. It, furthermore, includes other types of car use like vehicle leasing.

Definitions that are more specific are often a list of criteria; Millard-Ball (2005, p. 2-1) provided the following:

- An organized group of participants
- One or more shared vehicles

- A decentralized network of parking locations (“pods”) stationed close to homes, workplaces, and/or transit stations
- Usage booked in advance
- Rentals for short time periods (increments of one hour or less)
- Self-accessing vehicles

Given that the authors were writing about CS in North America during the 1990s and early 2000s, it should come as no surprise today that their criteria do not take into consideration newer forms of CS like P2P CS that do not conform to the parking station/pod model. The authors follow their list up by citing the State of Washington’s definition of CS as being the most concise and comprehensive:

“A membership program intended to offer an alternative to car ownership under which persons or entities that become members are permitted to use vehicles from a fleet on an hourly basis.” (Washington State Legislature, 2005)

While the Washington definition is a relatively good one, it leaves room to include traditional rental car services that often require time-consuming paperwork and verification procedures during vehicle pick-up. A key feature of modern formal CS is that it is self-service, meaning that a user can access a vehicle without interacting with another human being. Similarly, Frenken (2013, p. 9) defines CS as “a system that allows people to rent locally available cars at any time and for any duration”. While this is generally true, it leaves out important elements of CS like membership and self-service access.

An additional source of confusion is that in some markets, the UK in particular, CS is known as car clubs, whereas the term CS often refers to ride sharing, carpooling and ride sourcing (Millard-Ball, 2005, p 2-1; Shaheen et al., 2015).

It is likely that the definition of CS will have to be updated to keep up with advances in technological innovation. For example, user operation is one of the distinguishing characteristics of CS; if autonomous vehicle technology develops further and removes the need for a human operator, the boundaries between CS, ride-sourcing and ride-sharing begin to blur and will potentially disappear.

Types of Car Sharing

There are several types of CS. To understand this variety, one can differentiate platforms based on business model or operational model.

Types of CS by Business Model

Broadly speaking, business model can refer to the organization’s mission, or what Millard-Ball (2005) refer to as organizational form, of which there are three main types: for-profit, nonprofit, and cooperative. In practice, cooperatives are a subset of nonprofit as just about all cooperative platforms are nonprofit, but not all nonprofit platforms are cooperatives.

The largest CS platforms in the world are for-profit corporate ventures. Zipcar, for example, is a subsidiary of the Avis Budget group, a traditional rental car company. Similarly, Share Now is a joint venture of the German automobile manufacturers Daimler and BMW. Whereas profit-driven corporations have greater access to venture capital and incentive to expand, nonprofit platforms have greater access to support from governments or foundations (*ibid.*). Furthermore, nonprofit platforms tend to focus on limited geographic contexts, usually one city, which is more likely to engender familiarity and trust from local stakeholders (Brook, 2004).

CS business models can also refer to the relationship between the customer and the service provider. Here as well, there are three main types: business-to-consumer (B2C), business-to-business (B2B) and peer-to-peer (P2P). There could theoretically be a fourth consumer-to-business (C2B) category in which individuals could rent out their private vehicles to businesses and institutions, but this model is yet to take off. Although CS platforms across all three business models have embraced information and communication technology (ICT) to develop their service offerings, the P2P model relies more heavily on such technology to carry out transactions—without smartphones, P2P CS would be virtually impossible. These categories are not mutually exclusive—platforms often have a hybrid business model that offers both B2B and B2C services.

Types of CS by Operational Model

The other main typology for differentiating CS platforms is based on operational model, which refers to how the vehicles are used. Martin and Shaheen (2016) identify four types—roundtrip, one-way, peer-to-peer (P2P), and fractional ownership. It is worth noting at the outset that P2P platforms are unique in that they represent a distinct operational model as well as business model.

Round-trip, also known as traditional or station-based CS, requires that vehicles be picked up and dropped off at the same location. It is the most common and oldest type of CS platform and spans across all business models. Nonprofit and cooperative platforms are almost exclusively round-trip in nature. A CS station is effectively a parking space, or set thereof, that must be made available exclusively for a shared vehicle. Such stations must be easily accessible to users—they are typically located in densely populated mixed-use residential urban neighborhoods (Shaheen and Cohen, 2013, p. 14).

Users of one-way CS can pick up and drop off vehicles in different locations, usually within a limited geographic area. They are almost always but not exclusively “free-floating” platforms for which the pick-up and drop-off locations are any available legal parking space; some one-way platforms operate on a station to station basis (Shaheen et al., 2015, p. 525). Some of the earliest

attempts at formal CS in the 1970s like Procotip in Montpelier, and Witkar in Amsterdam were one-way platforms. These were short lived because of technical and organizational challenges. Current one-way platforms rely on relatively recent technologies like smartphones and GPS, that were not available in prior decades.

ICT innovations have also played a central role in the development of P2P CS which involves private automobile owners renting out their vehicles to other users. The CS platform operators provides a digital interface that facilitates transactions, for which they receive a share of rental costs. In a traditional P2P CS schemes, both the provider and the user are private individuals. [Stocker and Shaheen \(2017, p. 10\)](#) also site hybrid and marketplace CS platforms. With Hybrid P2P, traditional CS organizations supplement their own fleet by renting privately owned vehicles on a short-term basis from individuals or other third parties. Marketplace P2P functions similarly to traditional P2P but less standardization with respect to price, pick-up and drop-off procedures, and dispute resolution.

Other kinds of CS that are less common include the following ([Martin and Shaheen, 2016](#); [Shaheen and Cohen, 2018](#)):

- Fractional ownership, also known as fractional leasing, refers to multiple users subscribing to a fleet of vehicles owned by a third party.
- Institutional fleets are shared cars used by staff, students, and members of businesses, government bodies, universities, and similar organizations.
- Park and ride CS aims to solve the “first and last mile” problem associated with public transit outside of urban cores.
- Guest or visitor CS platforms targets users that are “away from home” or in-transit, often at sites such as holiday resorts and airports.
- Research pilots, which are not intended to be commercially viable

Impacts of Car Sharing

The first scholarly analyses of CS were pilot projects in the 1980s, such as Mobility Enterprise at Purdue University and STAR in San Francisco, that were testing the feasibility and potential impact of such services. As the number of CS services in Europe and North American increased over the next decade, especially economically viable ones, broader analyses of commercial offerings, public response, and overall impact became possible. Whereas the initial studies focused on individual service providers and case cities, as the CS industry matured, subsequent studies began to take into consideration industry-wide effects and lifecycle impacts of car sharing.

The impacts of CS are often measured by a multitude of environmental, economic and social indicators. Environmental impact, especially as it relates to GHG emissions and climate change, are the most studied and written about. The most common indicators used for determining the environmental impacts of CS are:

- Vehicle holding at the household level
- Vehicle miles traveled/Vehicle km traveled (VMT/VKT)
- GHG emissions
- Modal splits/relationship between CS and other modes of mobility

By the early 2000s, the CS market in North America and Europe underwent a period of maturity. There was a marked increase in the number of national and international commercial enterprises, as opposed to local cooperatives. The most notable of these were Zipcar, Car2go, DriveNow, and Sunfleet, all of which were either spearheaded by or eventually acquired by an incumbent automobile manufacturer or traditional car rental company. During this time, scholarly analysis of CS began to expand beyond individual programs and cities. [Martin et al. \(2010\)](#) carried out one of the first studies of CS that used survey data from a broad set of user participants ($N = 6281$) in the United States and Canada. The study found that car sharing removed between 9 and 13 vehicles from the road for each shared vehicle deployed—a total reduction of 90,000–130,000 vehicles.

Concurrently, a similarly broad study of CS in Europe, co-financed by the European Union and entitled “Momo car sharing”, was carried out using survey data from 84 CS service providers that operated a fleet of over 3500 vehicles, mostly in Germany. The study found that CS generally reduced the CO₂ emissions footprint of vehicles by 15–20% ([Loose, 2010](#)).

As the popularity of CS increased, and newer forms of CS (i.e., free floating and P2P platforms) emerged, larger studies supported the claim that even these new types of CS reduce carbon emissions. According to [Firnknorn and Müller \(2011\)](#) who studied the free-floating service car2go in Germany, 13.5% of a total 17,000 members were expected to reduce their car ownership, which corresponds to a net reduction of almost 2000 vehicles. The same study indicates that CS users emitted 53%–60% less CO₂ and that even when assuming the highest level of fuel consumption, car2go users could increase their VMT by 167% and maintain an overall reduction in CO₂ emissions.

Implications for Policy and Planning

From a policy perspective, how a practice such as car sharing can undergo formation and reformation in a manner that it can sustain itself needs to be understood and taken forward in policies and programs.

Car sharing has a vital spatial dimension. It requires not just the existence of users, but a geographic concentration of them as well. Spatial proximity to transit stops, high frequency of public transport services, walking/cycling infrastructure and to a great extent, restricted parking are essential material elements of car sharing. In municipal periphery, with lower concentrations of shared cars and users, car sharing becomes much less attractive. This suggests that in addition to requiring a sufficient number of users, car sharing is the type of innovation that is viable only when its users are located close to one another.

As with other forms of sharing, car sharing is predicated on reputation or trust in other users and the institution that governs the collective (Botsman, 2012). By allowing users to operate independently of one another, automobility has developed without the need for such trust. Car sharing necessarily makes the user more dependent on other users, not only temporally and spatially, but also with respect to how well they treat the car. This trust is often carried forward at the institutional level as well. Currently, the cooperative model strikes the perfect balance between alternative business model and institutional trustworthiness. If car sharing and other shared goods and services are to gain in popularity, there must be ample institutional trust or reputation that is capable of off-setting any suspicion or lack of trust users may have among one another.

Furthermore, for creating effective policies and solutions, it is essential to analyze the behavior of car sharing users across multiple life stages. This set of knowledge, on user variations, will offer the necessary tools with which the practice of car sharing can be upscaled to a mainstream or dominant mode of transport. The relationship between user and promoter circumstances and conditions is not static, but coevolutionary. Promoters need to respond to how users actually use their products and services.

Additionally, government needs to engage in understanding the different retention and recruitment mechanisms at play and support the most effective mechanisms for car sharing. For example, rather than or in addition to using the language of cost savings, convenience and environmental sustainability to market car sharing, there is an opportunity frame car sharing as an integral part of an urban lifestyle which builds on neighborhoods created on the principle of accessibility by walking, bicycling and public transport.

Further Research

The current landscape of urban mobility is in a state of flux. Shared mobility is fast emerging as one of the most sustainable options to streamline daily mobilities. Car sharing, along with other shared modalities like bikesharing, E-scooters etc. need to be integrated in ongoing discussions on multimodality and Mobility-as-a-service (MaaS). Additionally, there is a need for broader long-term and multicontext studies to understand the temporal, geographic and demographic particularities of car sharing.

The impact of car sharing has been primarily assessed through changes in VKT and vehicle holdings. This is necessarily so because of the ease of plotting these standardized impacts to feed standardized assessments. But such research remain incomplete in multiple ways. First, they do not engage with quantifying noncarbon emissions like nitrous-oxide, carbon monoxide that are not associated with VKT as compared with carbon dioxide. The nexus between increased battery-electric shared cars and health implications needs further attention, as the levels of suspended particulate matter (SPM) which are a leading cause of respiratory harm among urban residents remain largely understudied in the context of car sharing (Uherek et al., 2010). Third, we lack knowledge on ways in which autonomous car sharing will compete with public transport in the future. The kind of business models needed to bring together these two modules together remains largely unexplored. Fourth, studies analyzing the impacts on landuse and urban form as a repercussion of increased shared mobility are needed. Finally, though the demographic effects have been partially analyzed, the cultural, social, and health implications (e.g., another pandemic like Covid-19 which brought the world to a standstill in 2020) of shared automobility in different contexts needs further examination. Furthermore, the interlocking of CS and the constructs of suburbanization, shopping malls, elderly homes, assisted living, business parks, special economic zones (SEZs), knowledge clusters, labor mobility, segregation, gentrification, conceptualizations of masculinity, trust, status, freedom, epidemic outbreaks and the general evolution of daily lives needs be further studied and discussed from different methodological, and disciplinary angles.

References

- Brook, D., 2004. Carsharing—Start Up Issues And New Operational Models. Transportation Research Board 83rd Annual Meeting, Washington, DC, Citeseer.
- Enterprise Carshare, retrieved 14.10.2018: <https://www.enterprise-carshare.com/us/en/locations.html>.
- Firnknorn, J., Müller, M., 2011. What will be the environmental effects of new free-floating car-sharing systems? The case of car2go in Ulm. *Ecol. Econ.* 70 (8), 1519–1528.
- Frenken, K., 2013. Towards a prospective transition framework. A co-evolutionary model of socio-technical transitions and an application to car sharing in The Netherlands. *International Workshop on the Sharing Economy*. Utrecht.
- Loose, W., 2010. The state of European car-sharing. Project Momo Final Report D, 2.
- Martin, E., Shaheen, S.A., Lidicker, J., 2010. Impact of carsharing on household vehicle holdings: results from North American shared-use vehicle survey. *Transport. Res. Rec.* 2143 (1), 150–158.
- Martin, E., Shaheen, S., 2016. Impacts of Car2Go on vehicle ownership, modal shift, vehicle miles traveled, and greenhouse gas emissions: an analysis of five North American Cities. *Transportation Sustainability Research Center, UC Berkeley*.
- Millard-Ball, A., 2005. Car-sharing: Where and How it Succeeds. *Transportation Research Board*.
- Shaheen, S., Cohen, A.P., 2013. Carsharing and personal vehicle services: worldwide market development and emerging trends. *Int. J. Sustain. Transport.* 7 (1), 5–34.
- Shaheen, S., Chan, N., Bansal, A., Cohen, A., 2015. Definitions, Industry Developments, and Early Understanding. University of California Berkeley Transportation Sustainability Research Center, Berkeley, CA. http://innovativemobility.org/wp-content/uploads/2015/11/SharedMobility_WhitePaper_FINAL.pdf.
- Shaheen, S.A., Chan, N.D., Micheaux, H., 2015. One-way carsharing's evolution and operator perspectives from the Americas. *Transportation* 42 (3), 519–536.
- Stocker, A., Shaheen, S., 2017. Shared automated vehicles: Review of business models. *International Transport Forum*.
- Shaheen, S., Cohen, A., 2018. Shared ride services in North America: definitions, impacts, and the future of pooling. *Transport Rev.* 1–16.

- Sumc, S.-U.M.C., 2015. Shared-Use Mobility-Reference Guide. Los Angeles, CA, Chicago.
- Truffer, B., 2003. User-led Innovation Processes: The development of Professional Car Sharing by Environmentally Concerned Citizens. *Innovation. Eur. J. Soc. Sci. Res.* 16 (2), 139–154.
- Uherek, E., Halenka, T., Borken-Kleefeld, J., Balkanski, Y., Bernsten, T., Borrego, C., Gauss, M., Hoor, P., Juda-Rezler, K., Lelieveld, J., Melas, D., Rypdal, K., Schmid, S., 2010. Transport impacts on atmosphere and climate: Land transport. *Atmosph. Environ.* 44 (37), 4772–4816.
- Washington State Legislature, Revised Code of Washington 87.70.010 Definitions, retrieved 23.09.2018: <https://app.leg.wa.gov/rcw/default.aspx?cite=82.70.010>.

Further Reading

- Burghard, U., Dütschke, E., 2019. Who wants shared mobility? Lessons from early adopters and mainstream drivers on electric carsharing in Germany. *Transport. Res. Part D: Transport Environ.* 71, 96–109.
- Lempert, R., Zhao, J., Dowlatabadi, H., 2019. Convenience, savings, or lifestyle? Distinct motivations and travel patterns of one-way and two-way carsharing members in Vancouver, Canada. *Transport. Res. Part D: Transport Environ.* 71, 141–152.
- Priya Uteng, T., Julsrud, T.E., George, C., 2019. The role of life events and context in type of car share uptake: Comparing users of peer-to-peer and cooperative programs in Oslo, Norway. *Transport. Res. Part D: Transport Environ.* 71, 186–206.
- Clean fleets, 2015. Increasing efficiency of administration's fleet management – Car-sharing in Bremen Clean Fleets Case Study. Retrieved on 03.09.2018: http://www.clean-fleets.eu/fileadmin/files/documents/Publications/case_studies/Clean_Fleets_case_study_-_Bremen_Car-Sharing_integration.pdf.
- Hleb, M., Perković, M., 2015. Report on the introduction of electric municipal car-sharing scheme in Koprivnica. Implementation Status Report, CIVITAS.
- Kent, J.L., Dowling, R., 2013. Puncturing automobility? Carsharing practices. *J. Transport Geogr.* 32, 86–92.
- Puschmann, T., Alt, R., 2016. Sharing economy. *Bus. Inform. Syst. Eng.* 58 (1), 93–99.
- Shaheen, S., Stocker, A., 2015. Carsharing for Business, Zipcar Case Study & Impact Analysis (Information Brief).

Toward a More Holistic Understanding of Mega Transport Project (MTP) Success

John Ward, University College London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

What is a Mega Transport Project?	147
What are the Typical Performance Issues During MTP Delivery?	148
The Strategic Success Paradoxes	149
MTPs as Interdependent and Open “System of Systems”	149
Path-Dependent Appraisal Methods and Narrow Consideration of Impacts?	149
Evaluation of Sustainable Development Goals (SDG) Remains Under-Developed	150
Participation Practices Need “Opening Up” to Identify and Mediate Conflicts Over Impacts	150
Conclusions – Breaking the Paradoxes?	151
References	152

What is a Mega Transport Project?

Mega transport projects (MTPs) are described as large-scale physical assets, which may be linear (such as roads, railways, bridges, tunnels, waterways) or nodes (such as airports, ports, stations and their associated infrastructures) or a combination of these. MTPs can be standalone projects, or part of a program of projects, defined by a local, regional, or national strategy for example. In terms of “large scale” the classical literature defines MTPs according to their cost and size, although more recently we see a move toward using complexity and impacts as key elements in defining MTPs. It follows that most texts define MTPs according to one or more of four interrelated subcategories:

- **Cost:** definitions commonly specify a lower threshold for capital expenditure. For example, academic studies have adopted a minimum construction cost for MTPs of US\$1 billion (see Flyvbjerg et al., 2003; Hauswirth et al., 2004; Merrow, 1998; OMEGA, 2011; Priemus et al., 2008). Governments and commercial practice, in contrast, may adopt lower thresholds, such as in the UK, where large scale “National Infrastructure Projects” are those over £50 million pounds (IPA, 2016) or in Germany where Large Scale projects are classed as above €100 million Euro (Rothengatter, 2019). Classifications based on cost alone represent a proxy for a broader set of factors, for example the sheer size of budgets required for MTPs tends to require approval at the highest level, making them an inherently political decision. However, a threshold minimum value could be argued as rather arbitrary, as there are examples of nationally significant infrastructure projects which fall below these thresholds in countries with lower GDP and more modest transport budgets.
- **Complexity:** MTPs are often described as complex, given they are seldom single physical assets, but heterogenous combinations of hard infrastructure (physical and mechanical assets) and soft infrastructure (communication systems and the various institutions involved in their planning, delivery, and operation) which are packaged into a project to achieve a given set of objectives. Complexity here can refer to the state arising due to the interaction and feedback between these various subprojects and their stakeholders (internal complexity). However, the interdependencies between the project and its multidimensional contexts can also be wide-ranging and complex, giving rise to manifold uncertainties and risk throughout the project lifecycle, including inadequate, unreliable or misleading information; and conflicts between decision making, policy, and planning (Flyvbjerg, 2007).
- **Size:** MTPs are usually physically large, extending over long distances or large surface areas. For example, Phase 1 of the High Speed 2 railway currently under construction in the UK will run on 140 miles of dedicated track, and will require the acquisition of 70 km² of land, equivalent in surface areas to a medium size town (NAO, 2018). It follows that they are resource intensive undertakings, both in materials they consume, and the labor needed for their construction. They also have long planning and implementation horizons, often measured in decades and with target life expectancies of between 50 and 100 years. These unusually large spatial and temporal dimensions are important as they can result in significant nonhomogenous impacts distributed across geographies and economies, both intended and unintended, and with the subsequent production of project “winners” and “losers” (OMEGA Centre, 2011).
- **Impacts:** MTPs are undertaken by their promoters in the belief they will have transformational outcomes or impacts, via concentration–distribution type models (Sturup and Low, 2019). For example, MTPs are frequently perceived as critical to the “success” of major urban, metropolitan, regional, and/or national development because of their potential to affect significant socio-economic and territorial change (OMEGA Centre, 2011, 2012). It is argued that it is these impacts which set MTPs aside from conventional projects. MTPs shouldn’t be considered merely as larger and more expensive versions of traditional transport infrastructure investments. Rather, they should be regarded as a totally “different breed” of project (Capka, 2004) because of their diverse outcomes and impacts that frequently go well beyond the physical assets that are being delivered. However, the ability of MTPs to bring about positive socio-economic impacts can be contentious, for example there is limited evidence to support the

development of high-speed lines as a means for rebalancing economies or “leveling up” (Tomaney and Marques, 2013). MTPs can also bring about unintended consequences, such as increasing social inequity (Lee, 2018), uncertainties surrounding CO₂ emissions and ecological impacts (Cornet et al., 2018) and encourage procurement practices against the public interest (OMEGA Centre, 2011). Essentially, MTP can yield significant effects across the social, economic, technical, political, legal, and environmental domains. In this regard an important distinction must be made between project outputs, outcomes and impacts (Turner and Zolin, 2012) where outputs are the manifestation of the physical asset at construction closeout (e.g., completion of new road/rail bridge with three river crossings), outcomes are the direct effect of the output, say during the first few months/years of operation (e.g., increased passenger flows between A + B), and impacts are the longer term effects of the project output and outcome (e.g., reconfiguration of urban space and socio economic profile). It is also important to stress outputs/outcomes and impacts fall across a range of stakeholders who they affect in different ways at different times (Fahri et al., 2015).

MTPs have in recent decades rapidly grown in number, size, complexity, and cost. The increasing demand for MTPs is variously cited as driven by: the need to renew existing major assets as they come to the end of their life (Söderlund et al., 2017); high level economic objectives, such as distribution and leveling up (Tomaney and Marques, 2013; Institute for Government, 2019); forces of global competition and the “Great War of Independence on Space” (Castells, 1996); the necessity for counter cyclical investment (UN, 2009); technological showcasing (Frick, 2008); sustainable development (Cornet et al., 2018); and even national prestige (SMEC, 2001). It follows that MTPs currently feature as significant aspects in many of the development agendas worldwide and they are highly politicized, indeed, of all the factors impacting success and failure, studies cited the role of politics as a key driver (OMEGA Centre, 2012).

However, MTPs are also much quoted as “paradoxical” in nature (Arena and Molloy, 2010; Flyvbjerg et al., 2003; Jennings, 2012; Samset and Volden, 2016). In particular because, despite their seemingly ever-increasing popularity, they are consistently deemed “unsuccessful” because of their failure to meet their original expectations in terms of delivery outcomes, typically in terms of outturn cost, on-time delivery and specification. Arguments based upon these metrics have tended to dominate the international narrative about project success (see Morris and Hough, 1987; Flyvbjerg et al., 2003; Samset, 2012). Whilst more recently there have been attempts to also emphasize positive outcomes to balance the somewhat negative perspective of project performance (Rothengatter, 2019) these are in the minority.

What are the Typical Performance Issues During MTP Delivery?

Inspired by the MTP paradoxes, significant research has been undertaken to both identify explanations of MTP performance and to suggested solutions. MTP development, common with most infrastructure types, comprises a series of phases or steps, starting with project conception, through planning, appraisal, funding, delivery, operation and decommissioning, or renewal. Research to date has tended to focus on the earlier phases of this lifecycle, with some of the most well-known and influential, for example, studying the links between selected aspects of the planning and delivery phases, or “project front end”. Of these studies, Sanderson (2012) identifies three explanation types for megaproject performance as follows:

- **Type 1: Strategic rent seeking behavior:** Project promoters pose vested interests in seeking to misrepresent nonviable projects by inflating benefits and reducing costs. This rent-seeking behavior goes unchecked because the incentives to produce deceptive and over-optimistic estimates of project viability are much stronger than the disincentives for doing so, leading to situations akin to the “survival of the un-fittest” (Flyvbjerg, 2009). Given the typically lengthy timescales involved with MTP development and realization, “there is a lack of proper accountability for project promoters, typically politicians, because they are often not in office when the actual viability of a project can be assessed” (Sanderson, 2012: 436). Essentially this explanation assumes project promoters are capable of accurate estimates of viability, but these are not in their interests to produce as they expose the inefficiencies of their invested project. Resolutions of these issues include the legal requirement for risk analysis, the de-politization of the early decision-making phases, and policies aimed at increasing accountability (Davidson and Huot, 1989; Wachs, 1989; Flyvbjerg et al., 2002, 2003, 2005; Flyvbjerg, 2009).
- **Type 2: misaligned and under-developed governance:** the core argument here is that megaprojects are by their very nature turbulent, and underperformance of many megaprojects is best explained by the presence of insufficient governance arrangements that cannot effectively respond to the turbulence caused by the myriad risks inevitably associated with these endeavors. Differently from type 1, here promoters lack the capacity and capability to perform. Researchers from this perspective suggest a need for a better grasp of project turbulence through the design of adequate project governance frameworks (Miller and Lessard, 2000; De Meyer et al., 2002; Miller and Hobbs, 2009; Morris, 2009; Winch, 2009).
- **Type 3: diverse project cultures and rationales:** here “megaprojects are typically characterized by multiple and diverse discourses, cultures and rationalities rather than by a singular, shared rationality as is assumed by more orthodox, technicist perspectives” (Sanderson, 2012: 437). This means that different actors understand the individual elements of project and there interaction and interdependencies with others in very different and competing ways. Here “performance problems are an almost inevitable result of the normal day-to-day practice of managers trying to cope with an organizational environment that is complex, ambiguous, and often highly conflictual” (Clegg et al., 2002; Pitsis et al., 2003; van Marrewijk et al., 2008). Solutions include “generic management processes associated with building trust, sense-making, organization learning, and building an appropriate organizational culture” (Atkinson et al., 2006: 688).

Lehtonen (2019) terms these perspectives as those of the “mainstream rationalists” who seek to solve the presumed chronic problems of MTPs through calls for greater accountability, rigorous and independent ex-ante evaluation, and more effective ex-post monitoring and enforcement designed to minimize strategic, self-interested behavior. However, this research is far from conclusive, with Pitsis et al. (2018) amongst others identifying a broad range of further work to be done within this sphere.

The Strategic Success Paradoxes

There is potentially also another set of paradoxes, this time concerning strategic project success (OMEGA, 2012). There is a dearth of broader understandings and definitions of success beyond solely the narrow “iron triangle” metrics within which to frame and understand project performance against long-term impacts (OMEGA, 2012; Rothengatter, 2019). In particular, there are fundamental questions about what kinds of root problems MTPs are trying to solve, and if embarking on large scale, nonreversible, supply side physical infrastructure projects is always the best way forward in a world of fast behavioral change and rapid onset of disruptive technologies? There are questions regarding the ultimate outcome of the projects – how far do these go in solving some of the major problems facing society? (Fahri et al., 2015; Lehtonen, 2019; OMEGA, 2011; Söderlund et al., 2017; Samset and Voldon, 2016).

The very low rates of ex-post project evaluation mean projects continue to be developed despite a lack of rigorously validated understandings of their wide-ranging impacts, and tangible contributions to society. Whilst one of the challenges to establishing the success of a project can be attributed to “conditional causality” (Li, 2008) which can obscure the relationship between project outputs and outcomes, there is a well-established body of literature presenting methods of evaluation (Vanday, 2012) which can be adapted for application to MTPs. However, there remain few examples in the public domain of post project evaluations of MTP that go beyond time/cost/specification and economic assessments of project performance. The lack of evaluation has prevented practice from establishing widely accepted frameworks and procedures for broader evaluations, or for normalizing the routine collection of information required to fuel such evaluations. These oversights have limited our ability to more fully understand long-term impacts.

Lehtonen defines the few researchers responding to these issues as the “alternative approach” literature. These studies “emphasize the need for learning and reflexivity, as well as the complexity and dynamism of megaprojects” (Lehtonen, 2019:149). Here there is the more complex and intractable set of issues regarding project success against long-term project outcomes and impacts. Classic examples of this are Sydney Cross Harbor Tunnel (Turner and Zolin, 2012) or Highspeed 2, an HSR current under construction in the UK. These issues are typically broad, long ranging, complex, and interdependent.

Whereas the literature concerning the operational phases and their relationship with MTP success is both copious and tightly defined within the realm of project management and associated disciplines, the research concerning the success of MTP projects in terms of their outcomes and impacts is more sparse and spread across many disciplines. Below we bring together some the various strands of research as they relate to the blocking and inducement mechanisms to understanding more about long-term MTP project impacts and success.

MTPs as Interdependent and Open “System of Systems”

Some authors present a case for the need of more “open-systems” approaches to decision-making than current planning paradigms typically permit. There is a strand of literature which conceives MTP projects as “exploratory”, almost “organic”, allowing for unexpected outcomes to become recognized and accepted as part of an “emergent order”, replacing in some instances more planned orders (see OMEGA Centre, 2011). The two-way relationship between MTP decision-making and the external policy environment not only has these projects responding to economic and political changes over time, but also has them frequently acting as catalysts for major transformational change either by design or default. A important text underpinning such broader definitions of project success is “Planning as Wicked Problems”, from public policy research, which established that planning in general, and hence MTP decision making and development in particular, are beset by interdependent challenges of complexity, ambiguity, conflict, attenuation and knowledge, and institutional silos, amongst others (Rittel and Webber, 1973), and the metrics of success are “better” or “worse” rather than deterministic.

This contrasts with the observed tendency for MTPs to be planned and executed somewhat within their modal silos (Rosenberg et al., 2014). This can also be true of the guiding policy frameworks and institutions, which identify their need. For example, the UK’s National Policy Statements (NPS) or the projects identified by the National Infrastructure Plan together identify infrastructure investment going forward, but as such neither provides a complete joined up, systemic and strategic definition of infrastructure need. Indeed, many of the sector-specific NPS are badly out of date. The literature on the management of portfolios of projects encourages the consideration of project interdependencies beyond departmental silos (Hall et al., 2016; Young and Hall, 2015), and thus offers a step toward reconciling this situation. However this strand of research is currently sparse, leading to the open-systems and “wicked” nature of MTPs being commonly overlooked. This means their strategic influence, interdependency relationships with other infrastructure projects and associate developments, plus long term and operational phases can be somewhat neglected in decision making, limiting the ability to plan and control such impacts over the longer term.

Path-Dependent Appraisal Methods and Narrow Consideration of Impacts?

A common theme in the literature focuses on the inadequacies (and inappropriate applications) of *ex-ante* project appraisal methodologies negatively contributing to excessively narrow frames of project appraisal (see in particular, Dimitriou, 2009;

Macharis and Nijkamp, 2013; OMEGA, 2012; Schutte, 2010). Reinforcing this critique, there is also a growing body of literature expressing concern regarding the excessive importance given during project appraisal to overtly economic tools such as Cost Benefit Analysis (CBA) (see Alexander, 2006; Brown et al., 2001; Metz, 2008; Nss, 2006; as well as the exclusion of many project stakeholders from the planning and appraisal process (see Colomb, 2010; Haezendonck, 2007; Macharis and Bernardini, 2015; Macharis et al., 2009).

Authors of this body of literature have emphasized the need to ensure more holistic and transparent assessment of project proposals by employing methodologies such as Multi-Criteria Analysis (MCA) more extensively and increasing the participatory character of the appraisal exercise (see for example, Ward et al., 2016). This is seen as not only potentially enhancing sustainable development attributes in the evaluation of such mega projects but also better linking their policy and planning aims with their performance, when compared with measurement by more traditional metrics of project appraisal that focus on traditional narrow “iron triangle” delivery metrics discussed above. Whilst forms of MCA have been routinely applied to the planning of MTPs, such as via the Web TAG methodology (DfT, 2019), or even the airport commissions study on Airport Capacity to inform policy (Airports Commission, 2015) there is little evidence of such appraisal frameworks over-riding the outcomes of CBA.

Furthermore, whilst MTPs are traditionally funded by the public sector, the ongoing constraints on public budgets mean Public Private Partnerships (PPPs) are growing in importance globally as the favored procurement route. Some see the practicalities of PPPs as placing them at odds with aspirations for more inclusive and open project appraisal with adequate consideration of the public interest (see Shaoul et al., 2006; Siemiatycki, 2009).

Evaluation of Sustainable Development Goals (SDG) Remains Under-Developed

When we consider the intended outcomes and impacts of MTPs, traditionally, the underlying rationale of many MTPs, especially from the perspective of the State, has been the delivery of economic growth on the basis of the economic benefits they are predicted to generate or support (often measured through travel time savings). But it is well known that there are potentially significant environmental costs associated with these developments across all scales from local to global (Sturup and Low, 2019). Whilst MTPs remain developed and operated via fossil fuel there will be significant greenhouse gas emissions (Cornet et al., 2018). Gellert and Lynch (2003) highlight MTPs often cause displacement (either biophysical or social) both directly, such as habitat loss or eviction, and indirectly such as aquifer damage causing knock-on effects, or loss of access to resources. The creation of waste products from MTPs can also create issues of environmental injustice (Agyeman et al., 2002). MTPs furthermore can dramatically change local environments, causing severance and damage social cohesion (Lee, 2018). They can exacerbate equity issues, and they are frequently opposed by people living in their shadow. Sturup and Low (2019) highlight that undesirable social and environmental lifestyle habits can be locked in by the long life of MTPs which then become barriers to sustainable development (SD).

Of late, the emphasis on economic growth has been challenged by a broader agenda of multiple development aims in support of declared United Nations (UN) visions of sustainable development. These visions potentially re-define the order of development priorities that MTPs should contribute to, and even question the fundamental viability of MTPs. The Brundtland Report's (UN, 1987) definition of sustainable development implies an important shift from sustainability as a primarily ecological concept to a framework that also emphasizes the economic and social dimensions of development. The report underlines the belief that sustainable development is not just a matter of balance between its sometimes conflicting social, environmental, economic, and institutional dimensions, but also reflective of a need to respect the ecological limits of development as far as they are known (Rees and Wackernagel, 1994). When trying to apply concepts of sustainable development to MTPs it becomes clear that establishing measurable criteria across all domains of what is still a contested concept raises numerous challenges not least on how to combine a set of disaggregated criteria to represent “sustainability”. There can also be considerable time lag between identification and practical application of such concepts and criteria: an evaluation of the performance of 30 contemporary MTPs against SDG principles found very low adherence to SDG, mainly due to the projects having been conceived or constructed before SDG were integrated into decision making (Ward and Skayannis, 2019). Furthermore, whilst Environmental Impact Assessment (EIA) has become a mandatory aspect of many MTP developments, it's coverage of sustainability is only partial (Bruhn-Tysk and Eklund, 2002) and Social Impact Assessment, which offers further coverage of sustainability, is generally under applied (Vancley, 2017). Other examples of tools to introduce elements of sustainable development applied to MTP include DfT's Web TAG or the Asian Development Bank's STAR framework, developed by the bank to monitor progress in its investments toward supporting accessible, safe, environmentally friendly, and affordable transport (Véron-Okamoto and Sakamoto, 2014). However, whilst these tools are applied ex-ante, they are seldom deployed ex-post as tools evaluate project success against SD derived goals, and to arrive at a consistent outcome. Even when applied ex-ante, there are questions regarding the consistency of decision making outcome and the priority of information obtained via current SD appraisal tools, as illustrated by the recent successful legal challenge to the Heathrow Second Runway due to the failure to recognize commitments to climate change targets in consenting documents (Court of Appeal, 2020) suggesting key climate change policy was side-lined from core decision making and consenting processes.

Participation Practices Need “Opening Up” to Identify and Mediate Conflicts Over Impacts

The stakeholder theory literature often categorizes stakeholders by into two types, such as primary and secondary (Clarkson, 1995) where primary stakeholders are those with a direct involvement in, or impact on, the outcome of the project

(determined via power, legitimacy and urgency criteria) such as suppliers, government departments, etc., and secondary are those with less immediate impacts on projects such as community groups, lobbyists, environmental groups, etc., Whilst the literature highlights the importance of effective and early engagement with key stakeholders is seen as critical, and this has been mirrored in legislation in many countries (McAdam et al., 2010), the main stream rationalist literature has tended to focus on primary stakeholders. Whilst the consideration of secondary stakeholders and the impacts that fall upon them have been relatively under-researched “little is known about how to align project objectives with those of the secondary stakeholders” (Maddaloni and Davis, 2018:544). This shortfall in consideration for secondary stakeholders is therefore significant when considering the success of an MTP in terms of impacts.

The driver for stakeholder consultation is usually a need to avoid conflict, stemming from the unsuccessful Decide, Act, Defend (DAD) strategy of defending technocratically determined projects via the public arena (Beierle and Cayford, 2001). However, recent modes of participation have their own drawbacks. “Unless the nature of the consultation is so thorough and proactive as to obviate the need for conflict in the first place, we see consultation as affording opposition groups another opportunity to protest” (McAdam et al., 2010 p407). Critically, thorough and proactive consultation can result in the “opening up” of the infrastructure planning and design process to a plurality of new stakeholders to identify a broader set of potential risks and opportunities, and to reduce the premature discarding of options, perhaps of a more sustainable nature (Steinemann, 2001; Stirling, 2008) due to agency agendas, path dependent behaviors (Hellström et al., 2013) and solution templating.

Importantly, early consideration of these differing viewpoints allows the project to be scoped and specified appropriately from the outset (Rosenberg et al., 2014). Reducing the use of “straw men” or project alternatives predesigned to be rejected by the process. Problem definitions may furthermore be less than rigorously identified, leading to proposals treating symptoms rather than causes and structural outcomes. Issues also arise concerning the lack of meaningful public involvement in the decision-making process, with this involvement usually coming too late to influence the development of alternatives (Adelle and Weiland, 2012; Ward and Skayannis, 2019).

It should be appreciated that it also ensures an adequate redundancy of information is provided which is inevitably necessary when involving multiple different stakeholder groups in dynamic complex decision-making (Rosenberg et al., 2014). While the identification of the relevant actors and the analysis of their mutual relationships and agendas can be very challenging and lengthen the decision-making process, this advocated “opening up” to a plurality of voices and lines of argument makes for more effective consultation and is more likely to capture key issues that need to be addressed. This is also important given the traditional Planning regimes are adversarial in nature (Rydin et al., 2018).

Conclusions – Breaking the Paradoxes?

This article has sought to achieve three goals: the first is a common definition of MTPs, the second to review a selection of literature regarding causes and solutions for the MTP output or delivery performance paradoxes and the third to discuss a further set of MTP paradoxes, where MTP performance against project outcomes and impacts remains systematically unevaluated despite enormous investments which have been made, and continue to be made in such projects.

With regards to a common definition of MTPs, this article has discussed four fundamental criteria which have been repeatedly used in the literature to define them. What is clear is that definitions of MTPs remain generic and general, suggesting these have yet to merge into a mature body of knowledge.

The research driven performance paradox seeks to understand why MTPs consistently “fail” when measured against delivery performance “Iron Triangle” metrics (budget, schedule, and specification). Whilst the iron triangle of project management was conceived as a tool to represent trade-offs between these three features, they have become uncoupled, independent metrics to measure success. The literature finds numerous explanations for delivery failure against these metrics and offers various solutions which have been categorized here under three schools of thought. However, despite these efforts there is little evidence of the various prescriptions converting into empirically observed improvement in performance. Evaluation of project management intervention appears distinctly lacking from within the field. This suggests research is required to not only diagnose the various problems with MTP development, but to shed light on why the various prescribed solutions are not translating into greater success in delivery practice.

In terms of further strategic paradoxes this article suggests that promoters and governments continue to invest in these projects without firm evidence that their long term outcomes and impacts are likely to be as expected. In fact, many rationales driving investment transpire as highly contestable. In these terms, the challenge to MTP decision-making becomes more demanding: the first challenge of how to overcome the irreducible complexity and uncertainty which characterizes the decision-making for such large-scale infrastructures remains, but it is enveloped by questions around how to effectively shape them, so that they achieve their expected impacts for society and then to be able to capture their performance against these expected impacts. What is clear is ex-post evaluation frameworks remain narrow, under-developed, and chronically under applied so lesson learning does not happen to the degree that many government guidance documents would require or expect, and this issue is magnified with respect to the difficulties in incorporating dimensions of sustainable development and stakeholder participation into both appraisal and evaluation. Authors suggest a holistic appraisal methodology would do much to address many important issues surrounding project impacts that are all too often sidelined by a “business as usual” mentality, however, as this article highlights there is a significant lack of validated, empirically derived data with which to populate such frameworks.

References

- Adelle, C., Weiland, S., 2012. Policy Assessment: The state of the Art. *Impact Assess. Proj. Apprais.* 30, 25–33.
- Alexander, E.R., 2006. In: *Evolution and Status: Where is Planning Evaluation Today and How did it get Here?* in *Evaluation in Planning - Evolution and Prospects* edited by Alexander, E.R. Ashgate Press, Aldershot, pp. 3–16.
- Airports Commission, 2015. Airports Commission Final Report, available at: <https://www.gov.uk/government/publications/airports-commission-final-report>.
- J., Agyeman, Robert, D., Bullard Bob, Evans, 2002. Exploring the Nexus: Bringing Together Sustainability, Environmental Justice and Equity. *Space. Polity.* 6 (1), 77–90, doi:10.1080/13562570220137907.
- Arena, L., Molloy, E., 2010. The Governance Paradox in Megaprojects.
- Atkinson, R., Crawford, L., Ward, S., 2006. Fundamental uncertainties in projects and the scope of project management. *Int. J. Constr. Proj. Manag.* 24 (8), 687–698.
- Beierle, T.C., Cayford, J., 2001. Evaluating Dispute Resolution as an Approach to Public Participation. *Resources for the Future*, Washington, DC.
- Brown, M., Milner, S., Bulman, E., 2001. 'Assessing Transport Investment Projects: A Policy Assessment Model' in *Transport Policy and Research: What Future?* In: Giorgi, L., Pohoryles, R.J. (Eds.), Vermont, Ashgate, Burlington, 44–89.
- De Meyer, A., Loch, C.H., Pich, M.T., 2002. Managing project uncertainty: from variation to chaos. *Sloan Manag. Rev.* 60–67, Winter.
- Sara Bruhn-Tysk, Mats Eklund, 2002. Environmental impact assessment—a tool for sustainable development?: A case study of biofuelled energy plants in Sweden. *Environ. Impact Assess. Rev.* 22 (2), 129–144, doi:10.1016/S0195-9255(01)00104-4, ISSN 0195-9255.
- Capka, R.J., 2004. Megaprojects – They Are a Different Breed. *J. Public. Roads.* 68 (1), 2–9.
- Castells, M., 1996. *The Rise of the Network Society, The Information Age: Economy, Society and Culture Vol. I.* Oxford, UK, Cambridge, Massachusetts.
- Clarkson, M.B.E., 1995. A stakeholder framework for analysing and evaluating corporate social performance. *Acad. Manage. Rev.* 20 (1), 92–117.
- Clegg, S.R., et al., 2002. Governmentality matters: designing an alliance culture of inter-organizational collaboration for managing projects. *Organ. Stud.* 23 (3), 317–337, doi:10.1177/0170840602233001.
- Colomb, C., 2010. The Perspective of the Social Planner: Incorporating Principles of Sustainable Development within the Design and Delivery of Major Projects, *An international study with particular reference to Mega Urban Transport Projects for the Institution of Civil Engineers and the Actuarial Profession*, OMEGA Working Paper 6, OMEGA Centre, University College London, London.
- Cornet, Y., Dudley, G., Banister, D., 2018. High Speed Rail: Implications for carbon emissions and biodiversity. *Case Stud. Transp. Policy* 6 (3), 376–390, doi:10.1016/j.cstp.2017.08.007.
- Court of Appeal, 2020. *R (Friends of the Earth) v Secretary of State for Transport and Others [2020] EWCA Civ 214.*
- Davidson, F.P., Huot, J.-C., 1989. Management trends for major projects. *Proj. Apprais.* 4 (3), 133–142.
- Dimitriou, H.T., 2009. Globalization, Mega Transport Projects and Private Finance, Paper presented at 4th VREF International Conference on Future of Urban Transport, Volvo Research and Educational Foundations (VREF), Gothenburg, Sweden, April 19–21.
- DfT, 2019. Transport analysis guidance. Available at: <https://www.gov.uk/guidance/transport-analysis-guidance-tag>.
- Fahri, Johan, Biesenthal, Christopher, Pollack, Julien, Sankaran, Shankar, 2015. Understanding Megaproject Success beyond the Project Close-Out Stage (2015). *Constr. Econ. Build.* 15 (3), 48–58. Available at SSRN: <https://ssrn.com/abstract=2889833>.
- Flyvbjerg, B., 2007. Policy and planning for large-infrastructure projects: problems, causes, cures. *Environ. Plann. B: Plann. Des.* 34 (4), 578–597, doi:10.1068/b32111.
- Flyvbjerg, B., 2009. Optimism and misrepresentation in early project development. In: Williams, T.M., Samset, K., Sunnevåg, K.J. (Eds.), *Making Essential Choices with Scant Information: Front-End Decision Making in Major Projects*. Palgrave Macmillan, Basingstoke, pp. 147–168.
- Flyvbjerg, B., 2009. Survival of the unfittest: why the worst infrastructure gets built—and what we can do about it. *Oxford Rev. Econ. Policy.* 25 (3), 344–367. Accessed October 29, 2020. <http://www.jstor.org/stable/23607068>.
- Flyvbjerg, B., Skramis Holm, M.K., Buhl, S.L., 2002. Underestimating costs in public works projects: error or lie? *J. Am. Plann. Assoc.* 68 (3), 279–295.
- Flyvbjerg, B., Bruzelius, N., Rothengatter, W., 2003. *Megaprojects and Risk: An Anatomy of Ambition*. Cambridge University Press, Cambridge.
- Flyvbjerg, B., et al., 2005. How (in) accurate are demand forecasts in public works projects? The case of transportation. *J. Am. Plann. Assoc.* 71 (2), 131–146.
- Frick, K.T., 2008. *The Cost of the Technological Sublime: Daring Ingenuity and the new San Francisco-Oakland Bay Bridge*. University of California Transportation Center, UC Berkeley. Retrieved from <https://escholarship.org/uc/item/2d00f48t>.
- Gellert, P.K., Lynch, B.D., 2003. Mega-projects as displacements. *Int. Soc. Sci. J.* 55, 15–25, doi:10.1111/1468-2451.5501002.
- Haezendonck, E., 2007. Transport Project Evaluation in a Complex European and Institutional Environment in *Transport Project Evaluation: Extending the Social Cost–Benefit Approach* edited by Haezendonck, E., Edward Elgar, Cheltenham, pp. 1–8.
- Hall, J.W., Tran, M., Hickford, A.J., Nicholls, R.J. (Eds.), 2016. *The Future of National Infrastructure: A System-of-Systems Approach*.
- Hauswirth, D.B., Hoffman, D.M., Kaine, H.F., Ozobu, I.L., Thomas, C.L., Wong, P.W. 2004. Collaborative Leadership – Stories in Transportation Mega Projects: A “lessons learned” approach to collaborative leadership in mega project management, University of Maryland, University College and US Department of Transportation, Washington, D.C.
- Hellström, M., Ruuska, I., Wikström, K., Jäfs, D., 2013. Project governance and path creation in the early stages of Finnish nuclear power projects. *Int. J. Constr. Proj. Manag.* 31 (5), 712–723. ISSN 0263-7863, <https://doi.org/10.1016/j.ijproman.2013.01.005>.
- Infrastructure Projects Authority, 2016. *National Infrastructure Delivery Plan 2016–2021*, ISBN 978-1-910835-77-7.
- Institute of Government, 2019. Big vs. small infrastructure projects: does size matter? <https://www.instituteforgovernment.org.uk/printpdf/5272>.
- Jennings, W., 2012. Executive politics, risk and the mega-project paradox. In: Lodge, M., Wegrich, K. (Eds.), *Executive Politics in Times of Crisis*. pp. 239–263. https://doi.org/10.1057/9781137010261_13
- Di Maddaloni, F., Davis, K., 2018. Project manager's perception of the local communities' stakeholder in megaprojects. An empirical investigation in the UK. *Int. J. Proj. Manag.* 36 (3), 542–565, doi:10.1016/j.ijproman.2017.11.003.
- McAdam, D., Boudet, H., Davis, J., Orr, R., Scott, W., Levitt, R., 2010. “Site fights”: explaining opposition to pipeline projects in the developing world. *Sociol. Forum.* 25 (3), 401–427. Retrieved September 17, 2020, from <http://www.jstor.org/stable/40783509>.
- Shaoul, J., Stafford, A., Stapleton, P., 2006. Highway robbery? A financial analysis of design, build, finance and operate (DBFO) in UK roads. *Transp. Rev.* 26 (3), 257–274.
- Siemiątycki, M., 2009. Delivering transportation infrastructure through public-private partnerships: planning concerns. *J. Am. Plann. Assoc.* 76 (1), 43–58, doi:10.1080/01944360903329295.
- Stirling, A., 2008. ‘Opening up or Closing down: Participation, pluralism and diversity in the social appraisal of technology’. *J. Sci. Technol. Hum. Values* 33, 262–294.
- Lee, J., 2018. Spatial Ethics as an evaluation tool for the long-term impacts of mega urban projects: An application of Spatial Ethics Multi-criteria Assessment to Canning Town Regeneration Project, London. *Int. J. Sustain. Dev. Plan.* 13 (4), 541–555.
- Miller, R., Hobbs, B., 2009. The complexity of decision-making in large projects with multiple partners: be prepared to change. In: Williams, T.M., Samset, K., Sunnevåg, K.J. (Eds.), *Making Essential Choices with Scant Information: Front-End Decision Making in Major Projects*. Palgrave Macmillan, Basingstoke, pp. 375–389.
- Morrow, E.W., 1988. *Understanding the Outcomes of Megaprojects: A qualitative analysis of very large civilian projects*, RAND Corporation, Santa Monica, California.
- OMEGA Centre, 2011. *OMEGA Project 2: Lessons for Decision-makers: A Comparative Analysis of Selected Large-scale Transport Infrastructure Projects in Europe, USA and Asia-Pacific*, OMEGA Centre, University College London, London.
- OMEGA Centre, 2012. *OMEGA Project 2: Lessons for Decision-makers: A Comparative Analysis of Selected Large-scale Transport Infrastructure Projects in Europe, USA and Asia-Pacific - Executive Summary Report*, OMEGA Centre, University College London, London.

- Markku Lehtonen, Ecological Economics and Opening up of Megaproject Appraisal: Lessons from Megaproject Scholarship and Topics for a Research Programme, Ecological Economics, Volume 159, 2019, Pages 148–156, SSN 0921-8009, <https://doi.org/10.1016/j.ecolecon.2019.01.018>.
- Li, J., 2008. Three Dimensions Structure Model Based on Project Evaluation. In WICOM'08. 4th International Conference, Dalian, 12–14 October 2008. Dalian: IEEE. doi: <http://dx.doi.org/10.1109/wicom.2008.1840>.
- Macharis, C., Bernardini, A., 2015. Reviewing the use of multi-criteria decision analysis for the evaluation of transport projects: time for a multi-actor approach. *Transp. Policy*. 37, 177–186.
- Macharis, C., De Witte, A., Ampe, J., 2009. The multi-actor multi-criteria analysis methodology (MAMCA) for the Evaluation of Transport Projects', *Theory and Practice*. J. Adv. Transp. 43 (2), 183–202.
- Macharis, C., Nijkamp, P., 2013. Multi-Actor and Multi-Criteria Analysis in Evaluating Mega-Projects'. In: Priemus, H., van Wee, B. (Eds.), *International Handbook on Mega-Projects*. Edward Elgar, Cheltenham.
- Miller, L., Lessard, D.R., 2000. *Strategic Management of Large Engineering Projects: Shaping Institutions, Risks and Governance*, The MIT-Press, Cambridge, MA.
- Metz, D., 2008. The myth of travel time saving. *Transp. Rev.* 28 (3), 321–336.
- Morris, P.W.G., Hough, G.H., 1987. *The Anatomy of Major Projects: A Study of the Reality of Project Management*. John Wiley & Sons, Winchester.
- Morris, P.W.G., 2009. Implementing strategy through project management: the importance of managing the project front-end. In: Williams, T.M., Samset, K., Sunnevåg, K.J. (Eds.), *Making Essential Choices with Scant Information: Front-End Decision Making in Major Projects*. Palgrave Macmillan, Basingstoke, pp. 39–67.
- Nass, P., 2006. Cost-benefit analyses of transportation investments: neither critical nor realistic. *J. Crit. Realism* 5 (1), 32–60.
- NAO, 2018. Investigation into land and property acquisition for Phase One (London – West Midlands) of the High Speed 2 programme, London.
- Pitsis, T.S., Clegg, S.R., Marosszeky, M., Rura-Polley, T., 2003. Constructing the Olympic dream: a future perfect strategy of project management. *Organ. Sci.* 14 (5), 574–590.
- Pitsis, A., Clegg, S., Freeder, D., Sankaran, S., Burdon, S., 2018. Megaprojects redefined – complexity vs. cost and social imperatives". *Int. J. Manag. Proj. Bus.* 11 (1), 7–34, doi:10.1108/IJMPB-07-2017-0080.
- Priemus, H.B., Flyvbjerg, B., van Wee, B., 2008. *Decision-Making on Mega-Projects: Cost-Benefit Analysis, Planning and Innovation*. Edward Elgar Publishing, Cheltenham.
- Rees, W.E., Wackernagel, M., 1994. Ecological Footprints and appropriated carrying capacity: Measuring the natural capital requirements of the human economy. In: Jansson, A.M., Hammer, M., Folke, C., Costanza, R. (Eds.), *Investing in Natural Capital: The Ecological Economics Approach to Sustainability*, Island Press, Washington D.C.
- Rittel, H.W.J., Webber, M.M., 1973. Dilemmas in a general theory of planning. *Policy Sci* 4, 155–169, doi:10.1007/BF01405730.
- Rosenberg, G., Carhart, N., Edkins, A.J., Ward, J., 2014. Development of a Proposed Interdependency Planning and Management Framework. International Centre for Infrastructure Futures, London, UK.
- Rothengatter, W., 2019. Megaprojects in transportation networks. *Transp. Policy* 75, A1–A15, doi:10.1016/j.tranpol.2018.08.002.
- Rydin, Y., Natarajan, L., Lee, M., Lock, S., 2018. Black-boxing the evidence: planning regulation and major renewable energy infrastructure projects in England and Wales. *Plan. Theory. Pract.* 19 (2), 218–234, doi:10.1080/14649357.2018.1456080.
- Samset, K., 2012. *Beforehand and Long Thereafter – A Look-Back on the Concepts to Some Historical Projects*. Ex Ante Academic Publisher, Norwegian University of Science and Technology, Trondheim, Norway.
- Samset, K., Volden, G.H., 2016. Front-end definition of projects: Ten paradoxes and some reflections regarding project management and project governance. *Int. J. Constr. Proj. Manag.* 34 (2), 297–313.
- Sanderson, J., 2012. Risk, uncertainty and governance in megaprojects: a critical discussion of alternative explanations. *Int. J. Proj. Manag.* 30 (4), 432–443, doi:10.1016/j.ijproman.2011.11.002.
- Schutte, I.C., 2010. *The Appraisal of Transport Infrastructure Projects in Municipal Spheres of Government in South Africa with Reference to City of Tshwane*, Ph.D. Thesis, University of South Africa, Pretoria.
- SMEC, 2001. *The Management of Megaprojects in International Development*, Snowy Mountain Engineering Company (SMEC). Holdings Ltd, Cooma New South Wales.
- Sturup, S., Low, N., 2019. Sustainable development and mega infrastructure: an overview of the issues. *J. Mega Infrastru. Sustain. Dev.* 1 (1), 8–26, doi:10.1080/24724718.2019.1591744.
- Söderlund, J., Sankaran, S., Biesenthal, C., 2017. The past and Present of Megaprojects'. *Proj. Manag. J.* 48 (6), 5–16, doi:10.1177/875697281704800602.
- Steinmann, A., 2001. Improving alternatives for environmental impact assessment. *Environ. Impact Assess. Rev.* 21 (1), 3–21.
- Tomaney, J., Marques, P., 2013. Evidence, policy, and the politics of regional development: the case of high-speed rail in the United Kingdom. *Environ. Plann. C: Government and Policy* 31 (3), 414–427, doi:10.1068/c11249r.
- Turner, R., Zolin, R., 2012. Forecasting success on large projects: developing reliable scales to predict multiple perspectives by multiple stakeholders over multiple time frames'. *Proj. Manage. J.* 43 (5), 87–99, doi:10.1002/pm.21289.
- UN, 2009. The role of public investment in social and economic development, New York [available at https://unctad.org/system/files/official-document/webdiae20091_en.pdf].
- Vancly, F., 2004. The triple bottom line and impact assessment: how do TBL, EIA, SIA, SEA and EMS relate to each other? *J. Environ. Assess. Pol. Manag.* 6 (3), 265–288. Retrieved September 17, 2020, from <http://www.jstor.org/stable/enviassepolimana.6.3.265>.
- Vancly, F., 2012. Guidance for the design of qualitative case study evaluation, available at: http://ec.europa.eu/regional_policy/sources/docgener/evaluation/doc/performance/Vancly.pdf.
- Vancly, F., 2017. The potential contribution of social impact assessment to megaproject developments. In: Lehtonen, M., Joly, P.-B., Aparicio, L. (Eds.), *Socioeconomic Evaluation of Megaprojects: Dealing with Uncertainties*. Routledge, pp. 181–198.
- Véron-Okamoto, A., Sakamoto, K., 2014. Toward a Sustainability Appraisal Framework for Transport, ADB Sustainable Development Working Paper Series, Manila, Philippines.
- Wachs, M., 1989. When planners lie with numbers. *J. Am. Plann. Assoc.* 55 (4), 476–479.
- Ward, E.J., Dimitriou, H.T., Dean, M., 2016. Theory and background of multi-criteria analysis: Toward a policy-led approach to mega transport infrastructure project Appraisal. *Res. Transp. Econ. Special Edition* 58, 21–45.
- Ward, E.J., Skayannis, P., 2019. Mega transport projects and sustainable development: lessons from a multi case study evaluation of international practice. *J. Mega Infrastruct. Sustain. Dev.* 1 (1), 27–53, doi:10.1080/24724718.2019.1623646.
- Winch, G.M., 2009. *Managing Construction Projects: An Information Processing Approach*, 2nd edition. Wiley-Blackwell, Oxford.
- van Marrewijk, A., Clegg, S.R., Pitsis, T.S., Veenswijk, M., 2008. Managing public-private megaprojects: paradoxes, complexity and project design. *Int. J. Proj. Manag.* 26, 591–600.
- Young, K., Hall, J.W., 2015. Introducing system interdependency into infrastructure appraisal: from projects to portfolios to pathways. *Infrastruct. Complex.* 2 (2), doi:10.1186/s40551-015-0005-8.

Equity Considerations in Transport Planning

Karel Martens, Technion - Israel Institute of Technology, Israel Institute of Technology, Haifa, Israel

© 2021 Elsevier Ltd. All rights reserved.

Introduction	154
Equity	154
Equity as an Impact of Transport Planning	155
Equity as a Goal of Transport Planning	156
Transport Planning as a Means to Promote Socio-Economic Equity	157
Conclusion	159
Biography	159
See Also	159
References	159
Further Reading	160

Introduction

Transport planning refers to the processes of decision-making that shape the future of transport systems at local, regional, national, and international scales. A decision-making process is inevitably a process of choice. The outcome of transport planning will thus always favor certain concerns over others, certain investments over others, certain transport modes over others, certain areas over others, and so on – and, therefore, certain people over others. Transport planning is thus inescapably a normative activity that raises, or should raise, equity concerns.

This obvious understanding has not always been explicitly acknowledged. In the early days of the field, the late 1950s and the 1960s, transport planning was typically seen as a largely technical issue, an issue to be left to transport engineers who were to design the best possible transport systems based on the latest knowledge and technologies. This vision was obviously not shared by all and the practice of transport planning, with its focus on catering for rapidly growing car ownership and use, was criticized by some from its very inception (e.g., Mumford, 1963 [1953]; Jacobs, 1961). Indeed, during the 1970s many streams of academic research developed that were highly critical of the emerging “modern” transport landscapes, of the power relations embedded in them, and of their uneven impacts on people (notably, the bodies of literature on gender and transport, spatial mismatch, space-time geography, and transport-related social exclusion). Despite its critical stance, much of this work did neither confront transport planning head-on nor provide clear directions for an improved practice. As a result, transport planning and equity have remained two worlds apart for much of the 20th century.

That has radically changed since the early 2000s. Building on the highly critical studies linking transport and social exclusion (Lucas, 2012), the literature on transport and equity has burgeoned and philosophies of social justice have started to invade transport writing (Verlinghieri and Schwanen, 2020). Studies have started to address transport planning more head-on from an equity perspective. Yet, there is still no general agreement on how to relate to “equity considerations” within the context of transport planning. Analyzing a range of arguments, three ways to look at the issue can be discerned. But before turning to these, it is essential to briefly reflect on the term equity.

Equity

The term equity directs the attention to what differently positioned people and population groups receive, experience, enjoy or suffer—and to the question what different people ought to receive, experience, or enjoy. Typically, three different dimensions of equity are distinguished: distribution (the benefits and burdens or “goods” and “bads” that people receive), recognition (the way people are being addressed and related to, as individuals and as members of a group), and representation (the ability of persons and groups to co-determine the decisions that shape their lives). While each of these are important for transport planning, in what follows I will focus primarily on the first dimension: the distribution of goods and bads. While this narrows down the understanding of equity, it is the topic of distribution that has received most attention in the transport literature and it is also this topic that requires most domain-specific knowledge. At the same time, such a narrow perspective on equity has been criticized and the scope of equity will therefore be expanded again towards the end of this chapter.

Equity understood as distribution encompasses a wide range of topics, as transport planning distributes a host of goods and bads, albeit often indirectly and often without aiming for a particular pattern of distribution across the population. A host of partly overlapping goods can be distinguished, including tangible goods like transport infrastructure (kilometers of roads, proximity to a train station) and more abstract goods like travel speeds or accessibility. Likewise, a broad range of “bads” are produced by transport

infrastructures and their use, including noise pollution, local air pollution, climate impacts, and safety risks. Over the past two decades, more and more studies have been published analyzing the distribution of these and other goods and bads related to transport (Lucas and Martens, 2019).

Many of these empirical studies do not tackle the issue of equity head-on. Often, studies equate disparities with inequalities and inequalities with inequities. These terms, however, have a distinct meaning. The terms differences, disparities, and inequalities are in its essence descriptive terms. The terms inequities and injustices, in turn, imply a normative judgment of these differences. Equity requires explicit normative judgment. It requires researchers and transport professionals alike to leave their comfort zone of technical knowledge and expertise to explicitly address questions of right and wrong. Studies mapping disparities thus do not directly address equity, even though they may be framed in the language of equity and even though such studies are likely to raise questions of equity.

The fact that equity, justice, and fairness are normative concepts does not imply that they are subjective or entirely personal concepts, that what is just or fair is in the eye of the beholder and is beyond rational debate. Rather the opposite is true (see Taylor 1985). It is no coincidence that the terms “justice” and “justification” are linguistically related – and are so in many languages. The relationship between the two terms suggests that what is just can be justified, while what is unjust cannot be justified. This does not imply that injustices are never justified in everyday life (unfortunately, quite the contrary is true), but it does suggest that any justification for an injustice is likely to be rejected after due consideration. This link between justice and justification underscores that equity, justice and fairness are not subjective but intersubjective notions.

Intersubjective agreement of what justice is or at least what it is not is strikingly high in a number of key policy domains (Jeekel and Martens, 2017). For instance, in most European and many non-European countries, there is broad support for health care systems based on a fundamental notion of solidarity, with solidarity implying that people ought to have equal access to a reasonable minimum of health care services irrespective of their ability to pay for these services. A comparable broad agreement exists about the free provision of basic education to children, irrespective of ability, income, gender, or race. The fact that there is broad agreement in these domains confirms that justice is not merely a matter of personal taste, but that people can come to a (rough) agreement about what constitutes justice in a particular domain of concern.

The field of transport planning stands out from other main domains of government intervention, not only because there is no agreement (yet) on what constitutes justice in this domain, but also because a debate on what might constitute justice in transport is virtually absent. In what follows, I turn to three possible perspectives on equity in transport planning. The description of these perspectives is intended to inform the debate about justice in transport which is so direly needed.

Equity as an Impact of Transport Planning

In the first perspective on equity and transport planning, the traditional approach to transport planning is largely accepted and equity is seen as an impact of planning decisions (e.g., Schweitzer and Valenzuela, 2004; and many others). The traditional approach views transport planning as “the field of government intervention that aims to ensure effective and efficient movement” (Shiftan et al., 2007). This understanding has gone hand in hand with a focus on the performance of the transport system, as a well-functioning transport system was and often still is seen as a precondition for ensuring “effective and efficient movement”. A well-functioning transport system, in turn, is typically equated with a congestion-free system. While this principle applies to both the road and public transport networks (and nowadays is even applied to bicycle network planning, for instance in the Netherlands), in practice it has implied a focus on solving congestion on road links.

This focus on congestion has shaped transport planning from its very start. It went hand in hand with the measurement of travel time losses and forecasting of time savings that may be generated by transport investments. Over the years, transport planning has slowly broadened its horizon to account more systematically for other impacts of transport investments, most notably environmental impacts and traffic safety implications. These concerns have affected transport planning to some extent, sometimes resulting in more environmentally-sensitive road planning, sometimes leading to a (limited) shift away from roads towards more sustainable modes of transport. Likewise, concerns over safety have led to better road design and – in some instances – to more priority for walking and cycling. More recently, there have been calls to also address the social consequences of transport planning. In most countries, however, congestion reduction has remained an overarching *raison-d’être* for transport planning.

At first glance, the increasing interest in equity adds an additional “impact” to this growing list of concerns. After all, equity can be seen as yet another concern to take into account when considering alternative transport futures. But at a closer look, equity actually adds an additional *dimension* to the analyses that need to be conducted within the context of transport planning. This is so because equity requires an understanding of the distribution of each benefit and cost over different population groups. When equity is taken seriously, it is no longer sufficient to analyze, for instance, the increase (or decrease) in air pollution due to a transport intervention. It is no longer sufficient to determine whether a project generates net benefits. Rather, it becomes necessary to analyze in detail *which* population groups benefit and which lose out. Moreover, in order to determine whether the expected distribution of impacts is fair, a normative judgment is essential, which in turn requires an explicit equity yardstick. Equity, even if understood as an impact, thus adds a layer of complexity to the traditional approach to transport planning.

This distributional concern is slowly entering policy making, although it is certainly not part and parcel of the standard practice of transport planning. Some major advances can be observed in the UK and the USA. In the UK, the national government introduced a requirement for a Social and Distributional Impact (SDI) appraisal of all new road projects over £10 million. First introduced in

2011, this requirement has since been updated and refined (Department for Transport, 2014). The purpose of SDI appraisal is to ensure that vulnerable population groups are not adversely affected by new projects (Martens and Lucas, 2018).

The SDI guidelines are to some extent similar to the requirements in the USA, where some major advances have been made in the analysis of equity impacts. The vast (economic) disparities between (ethnic) groups have triggered a high sensitivity for distributional issues in the USA, not in the least because of the influential environmental justice movement. This has resulted in a series of federal directives and guidelines requiring transport planning authorities to conduct equity analyses. These guidelines have indeed helped to shape practice, with an increasing number of metropolitan planning organizations and other authorities starting to conduct distributive analyses, first primarily regarding the burdens of transport and more recently also regarding the benefits (Karner and Niemeier, 2013). Given the lack of clear guidance on how to analyze and normatively assess equity impacts, it is no surprise that these analyses vary in multiple dimensions, including the benefits and burdens addressed, the population groups distinguished, the indicators used, and the (implicit or explicit) standards of justice employed. In many cases, transport planning bodies do not adopt any explicit equity principle to assess the results of the distributive analyses (Martens and Golub, 2018). In other cases, they adopt the principle of proportionality following legal requirements that public agencies identify and avoid “disproportionately high and adverse” effects on minorities and weaker communities (Schweitzer and Stephenson, 2007). The principle of proportionality requires that population groups receive a share of benefits or burdens that does not deviate too much from their share in the total population, leaving substantial flexibility to public authorities to determine whether observed distributive patterns are at odds with requirements of equity or not. Perhaps even more important is the lack of clarity in policy documents regarding the link between conducted distributive analyses and proposed interventions and policies. In many cases, the mapping of equity impacts seems to be conducted as an add-on that does little to change the primary concern in transport planning with congestion mitigation. The general impression is that as long as the impacts of proposed transport plans and projects are not blatantly disproportionate, distributive impacts have few implications for policy priorities (Lowe, 2014).

While the practices in the UK and the USA can be criticized on numerous grounds, it is also clear that other countries could learn from the experiences in analyzing a range of equity impacts related to transport plans and programs. At the same time, these experiences give little direction on how to weigh equity impacts vis-à-vis the many other impacts generated by transport interventions. Like negative externalities, attention for equity may ultimately do little to change the practice and priorities of traditional transport planning, as long as it is merely seen as an impact of transport planning.

Equity as a Goal of Transport Planning

The second perspective on transport planning and equity seeks to put transport planning on a par with other major domains of government intervention, such as health care, education and housing. Like transport, each of these domains is concerned with the delivery, broadly conceived, of a particular good to citizens. In contrast to transport, equity is not seen in these domains as a mere impact but rather as a key goal of government policies, with regulations, taxes and subsidies designed to deliver equity in each of the domains. For instance, many countries have enshrined a right to housing in their constitution and have set up extensive policy frameworks encompassing housing supply and demand measures with the purpose of providing adequate housing for all citizens. While in none of these domains there is complete agreement about what fairness exactly requires from government, there tends to be broad support for the core justice principles, among decision-makers, politicians and citizens, across a very large share of the political spectrum.

The second perspective on transport planning and equity takes the policy domains of health care, education, and housing as its model. This both broadens and narrows concerns over equity. It broadens these concerns, because equity now becomes an *objective* of transport planning rather than an impact. It narrows down the concerns, because this second perspective draws all attention to one key outcome of transport policies: accessibility. This focus does not imply that the (distribution of) the externalities related to transport are irrelevant. Rather, it is argued that these externalities, such as air and noise pollution, can be properly seen as impacts, while providing accessibility is the very purpose of transport systems—and thus should be treated differently from these impacts in the transport planning process.

The understanding that the provision of accessibility is the very purpose of transport planning goes back to at least the early 1970s, but has gained in strength since the 1990s, as symbolized by Cervero’s classic paper calling for a paradigm shift from (auto) mobility to accessibility (Cervero, 1997). The call for a focus on accessibility is based on the fundamental insight that travel is first and foremost a means to an end, a means to access destinations dispersed over space. Hence, many authors have argued that transport planning should not focus on the measurement and enhancement of mobility, but should rather analyze and enhance accessibility (Levine et al., 2019).

From the perspective of justice and fairness, accessibility does not merely represent the “service” people receive from the transport system. Rather, it captures the much more fundamental notions of choice and freedom. Accessibility, understood as the range of activities and destinations people can reach within a reasonable budget of time, costs, and effort, represents the choices available to persons. The higher a person’s accessibility level, the more destinations are available, the larger her freedom to choose from multiple options, and hence the broader the possibilities for the person to execute her agency and give direction to her life. Accessibility can thus be seen as a, perhaps rather technical, measure of freedom, with freedom understood in a positive rather than a negative way, that is, defined as “freedom to” rather than “freedom from” (Berlin, 1958). It is this understanding that makes the distribution of accessibility of particular importance from the perspective of justice.

The equity perspective adds a fundamental layer to transport planning's focus on accessibility. From an equity perspective, it is not sufficient to assess transport interventions based on the extent to which they enhance accessibility "in general" or "on average". Rather, it becomes key to determine how accessibility is distributed over different population groups and how proposed interventions are expected to change this pattern of distribution. Moreover, in order to assess the fairness of these (changes in) accessibility patterns, a normative assessment of accessibility patterns is required, based on an explicit equity yardstick or justice principle.

This raises the core question of this second perspective on transport planning and equity: What is a fair distribution of accessibility? A range of authors have grappled with this question in recent years (see [Verlinghieri and Schwanen, 2020](#)) and have explored the relevance and implications of a range of social justice philosophies for transport. The depth of the academic debate has been limited so far, as few authors have taken up the challenge of developing a systematic and robust defense of a particular principle. Moreover, a societal debate, so essential as part of any call for justice, is still largely absent. Partly because these debates remain underdeveloped, no broad agreement on a proper justice principle for the domain of transport has emerged yet ([Martens and Lucas, 2018](#)). Different authors have proposed different principles, including sufficientarian and more egalitarian principles. The former type of principles set a minimum standard in terms of accessibility ([Martens, 2017](#)). More egalitarian approaches might call for the adoption of a range constraint on accessibility, implying a maximum difference between persons with the highest and lowest level of accessibility (e.g., [Golub and Martens, 2014](#)).

No agreement about the proper principle for the distribution of accessibility is in sight yet. But more important than the exact justice principle to be adopted, is the acceptance that the distribution of accessibility *should* be subject to constraints of justice. This is so, because the acceptance of this proposition—that is, that the pattern of accessibility is a concern of justice—has profound implications for the practice of transport planning. It implies that transport planning can no longer be defined as "the field of government intervention that aims to ensure effective and efficient movement". It is this traditional understanding of the purpose of transport planning, with its focus on the functioning of the transport system rather than on the service people receive from the system, that has been responsible for the existing disparities in accessibility, especially between people with and without access to a car, with far-reaching implications for affected people (see [Lucas, 2012](#)). When it is accepted that the distribution of accessibility is a concern of justice, transport planning will have to relate systematically to the existing patterns of accessibility and the impacts of proposed interventions on these patterns. This would lead to a fundamental reformulation of the purpose of transport planning. In general terms, without specifying a principle of justice, the field of transport planning would then be defined as "the field of governance that seeks to *guarantee a fair level of accessibility* for all through the regulation, operation, maintenance, and improvement of the transport system" (see [Martens, 2019](#)). Such a formulation could be easily adapted whatever principle of justice would be agreed upon after due societal debate.

While current transport planning practice is still far removed from this ideal, some countries have made some advancements in this direction. In Flanders, the Law on Mobility Policy, adopted in 2001 and again in 2009, formalizes five objectives among which the goal "to provide everyone with the opportunity to be mobile (. . .) with the aim of full participation of everyone in society", with the latter part of this goal implicitly underscoring the importance of accessibility. While this goal has failed to shape the practice of transport planning, it has resulted in highly detailed guidelines requiring improvements in public transport across the Flanders region. Likewise, the EU-endorsed guidelines for Sustainable Urban Mobility Plans formulated six objectives, the first of which read "ensure all citizens are offered transport options that enable access to key destinations and services" ([European Commission-EC, 2013](#); the objective no longer features in the most recent guidelines). Even more ambitious in aim and scope has been the UK's attempt at accessibility planning ([Department for Transport, 2004](#)). The purpose of accessibility planning was to systematically determine the extent to which population groups at risk of social exclusion could access key destinations for employment, healthcare, education, food shopping, and leisure—and to subsequently propose interventions in the (public) transport system as well as in service delivery to address any gaps. While accessibility planning was disconnected from national-level transport investments and policies and thus failed to address a major cause of disparities in accessibility ([Lucas, 2012](#); [Martens, 2017](#)), it was a bold attempt to seriously address accessibility issues at the local level.

To conclude, the second perspective on transport planning and equity implies profound changes to the practice of transport planning. It would not only require transport planners to analyze the functioning of the transport system in an entirely different way, but also adopt a very different problem definition. It would redirect the focus away from congestion on the network, towards population groups experiencing an unfair level of accessibility. Furthermore, it would demand the adoption of different analytical tools, indicators, and appraisal techniques, ultimately resulting in infrastructure investments, service provision and transport policies quite different from what is currently observed in much of the world.

Transport Planning as a Means to Promote Socio-Economic Equity

In the third perspective on transport planning and equity, concerns over social justice take center stage. This perspective is based on the observation that transport planning as practiced over the past seven decades has been part and parcel of the powerful processes of domination and oppression. These processes have systematically disadvantaged particular groups in society, including low-income households, ethnic minorities, migrants, and women, to the advantage of privileged groups ([Bullard et al., 2004](#)). From this perspective, transport planning is seen as one of the culprits in generating inequalities, poverty and disadvantage in society, rather than as a liberating force enhancing people's freedom of movement. Transport infrastructures, it is argued, are not merely passive artifacts, but shape "(im)mobile subjects" and have produced "unfair privilege" and "disadvantage" as each other's mirror images

(Sheller, 2015). Such a role is deemed unacceptable from a social justice perspective. Transport planning should therefore radically change and become a force of empowerment rather than a tool of oppression (Cook and Butz, 2019), that is, a vehicle to advance justice well beyond the transport domain itself. Hence, I will refer to it as the “radical approach to transport planning”.

It may be clear that merely mapping equity impacts, as suggested in the first perspective presented in this chapter, is not sufficient from the standpoint of this radical approach. Such mapping exercises often fail to shape policy priorities and even if they do so, they tend to maintain the status quo rather than advance radical change. Indeed, the tendency to adopt the principle of proportionality, while intuitively reasonable, does just that: maintain the status quo (Martens and Golub, 2018). This is so, because the principle of proportionality is typically applied to the benefits and burdens generated by an *intervention*, while ignoring existing disparities between population groups. The mapping of equity impacts is thus more likely to be part of the problem than a step towards gradually and systematically eliminating injustices in society.

In a similar vein, the radical approach is also wary about the call to provide a “fair” level of accessibility, as proposed in the second perspective (Verlinghieri and Schwanen, 2020). The argument here is that whenever fairness allows for deviations from equality, such deviations are likely to work to the detriment of disadvantaged groups. From the radical perspective, a principle like sufficient accessibility for all members of society is all but a fair principle, as it is likely to imply that currently disadvantaged groups will have to accept a “sufficient” level of accessibility, while advantaged groups may be expected to enjoy much higher levels of accessibility, either obtained through “free” market transactions or through their influence on inevitably political decision-making processes (Vanoutrive and Cooper, 2019). The same would hold for other types of benefits and burdens generated by transport investments. For instance, if transport planning merely upholds legal norms regarding air pollution, it is virtually inevitable that weaker population groups, in terms of income or ethnicity, are the ones ending up living in the most polluted neighborhoods, certainly in countries with a weakly developed social housing sector.

The radical approach to transport planning also substantially expands the scope of equity. It is not sufficient to address the distribution of goods and bads generated by (the use of) transport infrastructures (Cook and Butz, 2019). Equity in transport planning requires an explicit and open debate about what exactly is at stake—going beyond distribution to concerns over inclusion and representation of “oppressed and disenfranchised groups”, who should be placed “front and center” (Sheller, 2018, p. 28-29). While the other two perspectives tend to be supportive of more inclusionary decision-making processes, they also treat concerns over distribution separately from concerns over representation and recognition. In the radical perspective, no such disconnect is possible.

This call for justice to encompass distribution, recognition and representation is fed by a critical assessment of the presumed progressive role of governments. Rather than assuming that governments act in the common interest and that reason and evidence can convince them to adopt fairness principles in transport, the radical perspective underscores the limited impact of knowledge and argument in the face of entrenched interests. Thus, the approach underscores that “[e]very inch (. . .) will be politically contested” (Sheller, 2018, p. 166), as there are powerful forces seeking to maintain the status quo. Where the two approaches discussed above perhaps implicitly assume that professionals and decision-makers will slowly change practice for the better, gradually embracing concerns over equity and even adopting principles of justice, proponents of this third perspective are far more wary about the willingness to change among key actors. Hence, they call for more radical action. This outlook resonates, for instance, in the work of Karner et al. (Karner et al. 2020) who call for a shift from a state-centered “transport equity” approach towards a society-centered “transport justice” approach, with the latter embracing models of social change that seek to transform state policies from the outside or even create new systems of bottom-up governance detached from those currently sanctioned by the state.

While representation and recognition are key in the radical perspective, questions of distribution remain important. The radical approach insists that, historically, transport planning has contributed to socio-economic disparities in society precisely through the way it has distributed a range of goods and bads. The approach therefore calls for the correction of these historic wrongs, in line with the notion of restorative justice guiding the environmental justice movement in the USA and beyond (Sheller, 2018). The notion of restorative justice requires transport planning to correct past wrongs and thus to become a force of equalization rather than oppression. Such a corrective approach has been adopted in countless environmental restoration processes in the USA (Cole and Foster, 2000), but has seen little following in the domain of transport. Transport planning working in the vain of restorative justice would go beyond merely mapping equity impacts or guaranteeing everyone their “fair share”. Rather, the aim would be to eradicate the existing disparities between population groups, disparities that often are repeated along lines of deprivation such as ethnicity, immigration status, and gender. This applies as much to the vast disparities in accessibility levels, as to the lasting injustices generated by large-scale infrastructures that have ripped apart entire neighborhoods and communities and have burdened them with dangerous pollution levels. Transport planning should correct these wrongs, dedicating substantial budgets to investments and policies that serve disadvantaged groups, so that transport becomes a resource of empowerment rather than a dimension of deprivation.

While radical in nature, some of elements of this perspective have been visible in USA planning practices. An early regulation from the US Department of Transport, dating back to 1970, providing guidelines on how to interpret the Civil Rights Act of 1964 in the domain of transport, explicitly called on transport agencies to take affirmative action to remove the effects of past discrimination (Martens and Golub, 2018). While subsequent guidelines and regulations were less radical in character, the notion of equality and restorative justice is still apparent in transport planning practice in the USA today. For instance, some metropolitan planning organizations explicitly refer to the call for restorative justice in the equity analyses conducted for their regional transport plans. Particularly explicit is the metropolitan planning organization for the Chicago area. In the official “Comprehensive Regional Plan 2040”, in a section discussing the relationship between residential segregation, “equitable access to opportunity” and economic opportunity, it is explicitly stated that “in designing [transport] policies and making investments, it is important to take actions that

do not perpetuate these [employment and income] inequities, and to correct them if possible” (Chicago Metropolitan Agency for Planning [CMAP] 2013b, 48). While such calls are promising steps in the right direction, the impact of equity analyses on decision-making still seems limited at best (see [Lowe, 2014](#)).

Conclusion

The past decade has seen a surge of academic work on transport and equity, as well as an increasing concern in practice over the uneven impacts of transport investments and policies. Much of the academic work is empirical in nature, mapping the uneven geographies in a range of dimensions, from accessibility to air pollution and from traffic safety to transport-related well-being ([Lucas and Martens, 2019](#)), without explicitly addressing the implications for transport planning. In parallel, an increasing number of authors have started to explore the nexus between equity and transport planning, resulting in different perspectives as highlighted in this chapter.

The time seems ripe for change. A range of cities have radically changed their policies, away from a priority for cars, in an effort to enhance urban livability, promote sustainability and create more inclusive transport landscapes. Climate change is forcing a fundamental rethinking of current transport systems and policies, with its heavy reliance on fossil fuels, at all spatial scales. Key institutions, like the EU and The World Bank, have started to address equity issues ([European Commission-EC, 2013](#); [The World Bank, 2005](#)). These developments all suggest that equity will become an even more dominant issue in transport in the future. Whether this will lead to a radical change in transport planning and policy remains to be seen. Transport still stands out from other policy domains, such as housing, education, and health care, with their firm roots in principles of justice. There is still little recognition that the domain of transport deserves or even requires a comparable underpinning. Entrenched perspectives and interests often still dominate. The transport profession is still dominated by an engineering perspective with its ambition to deliver a transport system that ensures “effective and efficient movement”, with little attention for how a supposedly well-functioning transport system serves diverse population groups. Many actors still see transport investments as a way to promote economic development, even if there is little proof of economic impacts of additional investments in regions with an already advanced transport system. And powerful interests related to the car and road building industries obviously have little interest in supporting a radical overhaul of current transport planning practices. The fact that a (car-owning) majority of the population is benefiting from the unjust transport landscapes created over the past decades is certainly also not helpful for achieving change.

We thus conclude that while a “normalization” of the transport domain is urgently called for, such a normalization is not likely to happen without struggle and effort. Academic scholarship can contribute its small share to this struggle by highlighting the inherently political nature of transport planning and by proposing alternative models to the status quo. Hopefully, this article has made its own small contribution in this direction.

Biography

Karel Martens is Associate Professor and Chair of the Graduate Program in Urban and Regional Planning at the Faculty of Architecture and Town Planning, Technion - Israel Institute of Technology (Haifa, Israel), where he also heads the Fair Transport Lab. He holds a master degree in Spatial Planning (1991) and a PhD in Policy Sciences (2000), both from Radboud University, the Netherlands. He has authored numerous publications on the nexus between transport and equity. His book *Transport Justice: Designing Fair Transportation Systems* has been described by colleagues as “groundbreaking”, a “landmark” and “a revolution.”

See Also

Mobility Planning/Policies for Older People; Land Use and Transport Planning; Planning for Bus Priority; Demand Responsive Transport; Community Transport; Public Transport Subsidy and Regulation

References

- Berlin, I., 1958. Two concepts of liberty. *Four Essays on Liberty* 118–172.
- Bullard, R.D., Johnson, G.S., Torres, A.O. (Eds.), 2004. *Highway Robbery: Transportation Racism and New Routes to Equity*. South End Press, Cambridge.
- Cervero, R., 1997. Paradigm shift: from automobility to accessibility planning. *Urban Futures* (22), 9–20.
- In Cook, N., Butz, D. (Eds.), 2019. *Mobilities, Mobility Justice and Social Justice*. Routledge, Oxon/New York.
- Department for Transport, 2004. *Guidance on Accessibility Planning in Local Transport Plans*. UK Department for Transport, London.
- Department for Transport, 2014. WebTAG UNIT A4.2 Distributional Impact Appraisal. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/487601/TAG_unit_a4.2_distributional_impact_appraisal_jan_14.pdf.
- European Commission-EC, 2013. *Guidelines: Developing and Implementing a Sustainable Urban Mobility Plan*. Brussels, Directorate-General for Mobility and Transport.
- Golub, A., Martens, K., 2014. Using principles of justice to assess the modal equity of regional transportation plans. *J. Transport Geogr.* 41, 10–20.
- Jacobs, J., 1961. *The Death and Life of Great American Cities*. Random House, New York/Toronto.

- Jeekel, H., Martens, K., 2017. Equity in transport: learning from the policy domains of housing, health care and education. *Eur. Transp. Res. Rev.* 9 (53).
- Karner, A., London, J., Rowangould, D., Manaugh, K., 2020. "From Transportation Equity to Transportation Justice: Within, Through, and Beyond the State." *J. Plan. Lit* 35, 440–459.
- Karner, A., Niemeier, D., 2013. Civil rights guidance and equity analysis methods for regional transportation plans: a critical review of literature and practice. *J. Transp. Geogr.* 33, 126–134.
- Levine, J., Grengs, J., Merlin, L.A., 2019. *From Mobility to Accessibility: Transforming Urban and Transportation Planning*. Cornell University Press, Ithaca and London.
- Lowe, K., 2014. Bypassing Equity? Transit Investment and Regional Transportation Planning. *J. Plan. Educ. Res.* 34 (1), 30–44.
- Lucas, K., 2012. In: Geurs, K.T., Krizek, K.J., Reggiani, A. (Eds.), *A critical assessment of accessibility planning for social inclusion. Accessibility Analysis and Transport Planning: Challenges for Europe and North America*, Edward Elgar, Cheltenham, pp. 228.
- Lucas, K., 2012. Transport and social exclusion: where are we now? *Transp. Policy* 20, 105–113.
- Lucas, K., Martens, K. (Eds.), 2019. *Measuring Transport Equity*. Elsevier.
- Martens, K., 2019. A people-centred approach to accessibility. *Paris International Transport Forum Discussion Papers*. OECD Publishing, 27.
- Martens, K., Golub, A., 2018. A fair distribution of accessibility: interpreting civil rights regulations for regional transportation plans. *J. Plan. Educ. Res.*, Available online 28 September 2018.
- Mumford, L., 1963. *The Highway and the City*. Harcourt, Brace & World, New York.
- Schweitzer, L., Stephenson, M., 2007. Right answers wrong questions: environmental justice as urban research. *Urban. Stud.* 44 (2), 319–337.
- Schweitzer, L., Valenzuela, A., 2004. Environmental injustice and transportation: the claims and the evidence. *J. Plan. Lit.* 18 (4), 383–398.
- Sheller, M., 2015. Racialized mobility transitions in Philadelphia: connecting urban sustainability and transport justice. *City Society* 27 (1), 70–91.
- Shitka, Y., Button, K.J., Nijkamp, P. (Eds.), 2007. *Transportation Planning*. Edward Elgar, Cheltenham.
- Taylor, C., 1985. *Philosophical Papers Volume 1 and 2*. Cambridge University Press, Cambridge.
- The World Bank, 2005. *Equity and Development: World Development Report 2006*. World Bank/Oxford University Press, Washington.
- Vanoutrive, T., Cooper, E., 2019. How just is transportation justice theory? The issues of paternalism and production. *Transp. Res. Part A: Policy Pract.* 122, 112–119.

Further Reading

- Martens, K., 2017. *Transport Justice: Designing Fair Transportation Systems*. Routledge, New York/London.
- Martens, K., Lucas, K., 2018. Perspectives on transport and social justice. *Handbook on Global Social Justice*. G. Craig, Edward Elgar, Cheltenham/Northampton, pp. 351–370.
- Sheller, M., 2018. *Mobility Justice: The Politics of Movement in an Age of Extremes*, Verso.
- Verlinghieri, E., Schwanen, T., 2020. Transport and mobility justice: evolving discussions. *J. Transp. Geogr.* 87, 102798.

Planning for Rail Transport

Simon P. Blainey, Transportation Research Group, Faculty of Engineering and Physical Sciences, University of Southampton, Boldrewood Campus, Southampton, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	161
Purpose	161
Route Planning	161
Capacity	162
Rolling Stock	163
Stations and Terminals	164
Customer Experience	164
Resilience	164
Revenue and Other Benefits	164
Conclusions	165
Biography	165
See Also	165
References	165
Further Reading	165

Introduction

While planning for rail transport shares some common features with planning for other transport modes, rail systems have a number of distinct requirements. This chapter provides a high level overview of the range of issues which need to be considered when planning for rail transport, with a particular focus on those issues which are specific to rail or where the requirements for rail transport differ markedly from those for other modes. Planning for rail transport could involve planning an entirely new rail network, extensions or improvements to an existing network, or changes to services and operations on an existing network, and the scope of this planning will clearly affect the range of factors which must be considered. However, while it would not be necessary when planning minor service changes to review all the factors considered when planning for a new network, the service planning might well still be constrained by the decisions taken when the network was first constructed. It is therefore useful to have an understanding of why these decisions were taken.

It is common when discussing rail transport to differentiate between the requirements of passenger and freight traffic, even though on many rail systems around the world both types of traffic operate side by side on the same tracks. In fact, passenger and freight traffic have many requirements in common, and this chapter attempts to recognize this where possible by considering different traffic types in parallel. However, where these requirements differ the chapter focuses on planning for passenger traffic, due to the limited space available here.

Purpose

The most fundamental questions that have to be answered when planning for rail transport relate to the intended purpose of the infrastructure or service that is being planned. For example, what market is the route intended to serve (or create)? Will the route be used primarily by passenger or freight traffic? Is the route intended to reduce journey times or create additional capacity? Are the new services targeted primarily at short distance commuters or long distance leisure travelers? The answers to these and similar questions will significantly influence the design characteristics of the infrastructure or service that is developed, and a lack of certainty or clarity in this regard can make it extremely difficult to gain support for a scheme. Once the purpose for a scheme has been established, it is then possible to move on to plan the characteristics of that scheme.

Route Planning

It is very rare for complete railway networks to be planned as a coherent whole, and even when such planning does take place (most commonly for urban metro networks) the timescales and costs involved mean that a phased approach to construction is usually adopted. Changing external conditions will often mean that the later phases are then altered or even cancelled. It is perhaps equally rare for rail infrastructure to be planned in an integrated way with investment in other transport infrastructure, especially roads, so that different transport modes complement rather than compete with one another. This is unfortunate, as this lack of integration can

have a negative impact on the business case for transport investments. In particular, if there is a desire to generate mode shift from road to rail (e.g., for environmental reasons), it is best to avoid making parallel investments in both rail and road infrastructure on the same corridor.

It is more common for railway routes to be planned on an individual basis, and these routes will normally be intended to connect two or more key points, and potentially a number of key intermediate points between them. Where railways are built primarily for freight transport these points will usually be a source of raw materials or goods (such as a mine, factory or port), and the market or hub for these goods (such as a factory, distribution center, or outlet for onward transport). In contrast, where railways are built primarily to serve passenger traffic or for a mixture of passenger and freight traffic these key points will usually be settlements, and it is more likely that routes serving these traffic types will also aim to connect intermediate sources of traffic.

In order to maximize potential traffic on the route, the initial aim when planning a route is often to minimize the journey time and/or the transport costs between the start and end points. However, there will almost inevitably be trade-offs between directness and (1) the number of intermediate locations which can be served, and (2) the engineering constraints which have to be overcome and the consequent impact on construction costs. Most rail infrastructure managers in the 21st century set maximum acceptable levels of gradient and curvature (e.g., [RSSB, 2011](#)), which then restrict the potential route options. These are likely to be influenced by the planned operational speed, as in general steeper gradients and sharper curvature will reduce construction costs (as fewer earthworks will be required) but also reduce maximum permitted speeds. The choice of power source for the trains will also influence the maximum acceptable gradient, and may indeed determine whether or not particular routes are viable. For example the Pennsylvania Railroad's extension to Manhattan and Long Island in the early 20th century involved steep gradients on the approaches to tunneled river crossings which were only operationally feasible following the development of electrically-powered locomotives ([Nock, 1979](#), p.34).

As railways are not usually built in undeveloped and uninhabited landscapes, these route options will almost certainly also be constrained by the existing human and natural environment. The topology and geology of the area crossed by a new railway clearly have the potential to affect the feasibility of construction and to increase costs, for example by requiring additional reinforcements to earthworks in order to compensate for unfavorable ground conditions. Increasing concern over the potential for infrastructure to cause environmental damage means that in the 21st century railway planners will also need to minimize the impact of the route on sensitive or protected areas, such as key wildlife habitats and ancient woodlands. There may also be a need to prevent visual "damage" to the landscape.

Railway planning is likely to be particularly constrained in areas where there is already a high density of human activity, that is, in towns and cities. It is usually desirable to minimize the requirement for demolition of existing buildings, as the need to compensate and relocate building owners and occupants will substantially increase scheme costs. The construction of new surface transport corridors through built-up areas also brings the risk of community severance, by inhibiting access between areas on either side of the new route. For these reasons the most cost-effective solution for building railways into urban areas is usually to construct them either overground or (more commonly) underground. The costs of such construction are though still substantial, meaning that the unit costs of railway construction in urban areas will be significantly higher than in rural areas.

As well as the impacts caused by the direct "footprint" of the railway route, there will also be broader impacts on the surrounding environment which need to be considered at the planning stage. For example, train operations may generate substantial levels of noise and vibration, which can have adverse impacts on the local population and also in some circumstances on elements of the built environment. Depending on the power source used for the trains, operations may also contribute to local air pollution, for example near large stations or depots when diesel engines are used ([Thornes et al., 2017](#)). Planners may therefore need to design mitigation measures for these impacts. In practice it is unlikely that a single "optimal" route connecting two points will exist which best accommodates all the constraints discussed above, and several route options will therefore usually be developed for appraisal.

Capacity

Planning an appropriate level of capacity is crucial in order to ensure a railway can accommodate the traffic which it is intended to carry while also keeping costs to a minimum. The capacity of a railway can, in simple terms, be defined as the volume of traffic which that railway is capable of transporting within a given period of time. This overall capacity is then a function of three components: (1) the volume of traffic which can be carried on individual trains; (2) the number of trains which can operate over the railway within a given time period; and (3) the volume of traffic which can be transferred on/off the trains at stations and terminals within that time period. All three of these components will need to be considered when planning for new routes or services, but this section will focus on component 2 as the other components are considered in sections Rolling Stock and Stations and Terminals.

The number of trains that can be operated on a route or network is itself dependent on several other factors. The first of these is the timetable, which sets out the timings and routings for all trains operating on the network, and therefore also implicitly describes the interactions between those services. Timetable planning is both a determinant of and constrained by the capacity of railway networks. When planning for additional services on an existing route, timetabling (and therefore the amount of spare capacity available) will be constrained by the existing infrastructure unless funding is available for enhancements. In contrast, when planning for a new route or network it is advisable to develop the timetable for the route alongside the planning of the infrastructure to ensure that sufficient infrastructure capacity is provided. Two key aspects of the rail infrastructure act to constrain the timetabling of trains, namely the track layout and the signaling design.

The number of trains that can operate on a route will be determined by both the number of tracks provided and the design of “nodes” where tracks interact (Armstrong & Preston, 2017). For example, railways with low traffic levels may only have a single track to accommodate traffic in both directions, with occasional loops provided to allow trains operating in opposite directions to safely pass. In contrast, railways with high traffic levels may have multiple tracks for trains operating in each direction to allow faster and slower trains to be separated. In such circumstances the frequency of crossings allowing trains to switch between the “fast” and “slow” lines will further influence the number of trains which can be accommodated on the route. Similarly, on railways with low traffic levels simple flat junctions are usually provided, where trains operating on to a diverging route cross at the same level the tracks carrying trains in the opposite direction. However, where traffic levels are higher then investment in “grade-separated” junctions may be justified, where the diverging tracks cross other lines via an over- or under-bridge.

Conventionally, most railway routes accommodating anything more than a low density of traffic have been controlled by a “fixed block” signaling system. Such systems divide each track on a route into a series of sections, with only a single train permitted to occupy each section at any given time, and lineside signals used to control access to these sections. The minimum length of these “block sections” is dictated by the minimum distance required to bring a train to a halt using the braking systems available. The length of the block sections will therefore determine the number of trains that can be accommodated on a route, and thus its capacity. Capacity can though be increased by providing “multiple aspect” signals which provide train drivers with advance warning of a danger signal ahead, allowing them to gradually reduce their speed. In recent decades cab signaling systems have been developed which can permit a further increase in line capacity by making use of “moving block” sections. By continuously monitoring the position and speed of all trains on a route, these systems mean that the minimum safe distance to the train in front can be maintained at all times, rather than only at the points where trains pass fixed lineside signals.

The characteristics of the timetable developed for a route will also have an impact on the ability of that route to accommodate additional services. Capacity will be maximized if the services on a route are entirely homogenous (i.e., they all operate at the same speeds and have the same stopping patterns). In contrast, capacity will be reduced if a route is required to accommodate a mix of service types, such as fast intercity services alongside heavy freight services and high capacity local commuter services serving closely-spaced stations. This is because trains operating at different speeds and calling at different sets of intermediate locations will rapidly catch up with one another, and then be delayed until the track layout permits a faster train to overtake a slower one. This problem will be exacerbated if trains have slow acceleration and braking characteristics, which can be a particular issue with heavy freight trains. The capacity issues associated with heterogeneous traffic types can be mitigated to some extent by “flighting” trains in a timetable, where trains with similar speeds and calling patterns are planned to follow one another along a route. However, this may lead to a sub-optimal timetable from the passenger’s point of view, where they are offered several trains to a given destination in quick succession followed by a long gap until the next service.

Rolling Stock

At the time of writing the usual choice of power source for train operations is between electric and diesel power. Electric power (supplied direct to the train via either overhead wires or third/fourth rails) has a number of advantages in terms of operational performance, operating cost and environmental impacts, but also has substantial infrastructure installation costs, which mean that it may be prohibitively expensive for long and/or lightly used routes. “Bi-mode” trains which have both diesel generators and electric current collection equipment are currently in vogue in some contexts (notably the UK), but although they reduce the use of diesel power on electrified sections of route the additional weight associated with diesel engines and fuel tanks means that their energy consumption and performance are inferior to fully electric trains. While some alternative power supply options such as hydrogen fuel cells and batteries are starting to emerge, development has been relatively slow and they are yet to be applied at a fleet-wide scale.

The design characteristics of the rolling stock (trains) required for the route will clearly be dependent on the type(s) of service to be operated. For freight traffic a range of different (and highly specialized) wagon designs are available to carry different types of traffic. For passenger traffic, features like the number of carriages, density and type of seating, and the number and position of doors should be determined based on the characteristic of the route and timetable they will be used for. For example, high frequency urban metro services with frequent stops will usually be operated using relatively long trains with several wide doors per carriage and high density seating with wide gangways to maximize capacity and minimize boarding/alighting times. In contrast, trains for long distance intercity services will tend to have only one or two doors per carriage (usually at the ends), as with fewer stops reducing station dwell times will be less of a priority. However, they will normally be provided with lower density and more comfortable seating (perhaps with tables) and space for luggage storage. They are also more likely to be provided with additional facilities such as toilets, food provision, power sockets, and wifi, although the latter facility is increasingly also an expectation of passengers on metro and suburban operations. Clear information provision is a necessity on any passenger service, and this should be considered carefully when planning rolling stock, taking into account all user groups (e.g., those with hearing impairments will not be able to access verbal announcements about service disruption). Ensuring that trains are accessible to all user groups (including, but not confined to, those with restricted mobility) is also a key consideration, and is now a legal requirement in some contexts. Finally, it is important to remember that rail services are almost always in competition with other modes, and therefore that rolling stock design should aim to provide a positive customer experience (see section Customer Experience).

Stations and Terminals

Planning for railway stations and freight terminals has a number of similarities with planning equivalent facilities for other modes, particularly other public passenger transport modes. Integration with other transport modes is usually of prime importance, because unlike private road transport for virtually all passenger traffic and for much freight traffic rail cannot provide a “door to door” service. Railway stations provide a point of entry to (or exit from) the transport system, an interface between different transport modes, and a facility to store traffic transferring between these different modes (whether passengers or freight). Station planning should therefore be integrated with broader transport planning for the area around the station, in order to ensure that access and egress trips are effectively facilitated.

In general, the aim of station design should be to make the passenger’s transfer through the station as seamless, comfortable and efficient as possible, with similar principles applied to freight terminals. However, the way in which this can best be achieved will clearly vary with the differing characteristics of individual locations and traffic flows. Facilities provided at stations can be divided into those which are required for the operation of the train services (such as tracks and power supplies) and those required to enable passengers to access these services (such as waiting areas, lifts, and information screens). While in theory the only absolutely essential facility at a passenger station is a means of boarding/alighting from a train, in the form of a platform and/or steps on the vehicle, in practice at most stations additional facilities will be provided to enhance either the customer experience or the profitability of the station. The range of facilities provided will usually increase in proportion with the level of traffic at that station, with the smallest stations having little more than a basic shelter and timetable, while the largest stations have a range of retail, catering and entertainment outlets which can make them trip destinations in their own right.

Whatever the size of the station, active and informed planning of the facilities provided will almost certainly deliver a better outcome for the users (and therefore generate more business for the operator). In particular, information provision and wayfinding are key at stations of any size, both with regard to the rail service offered and to facilities for onward travel and key destinations in the surrounding area. At larger stations this may include the provision of dedicated customer-facing staff, while at smaller stations it will rely solely on signage (whether fixed or interactive).

Customer Experience

Planning for both stations and rolling stock links into the more general issue of planning to provide a high level of customer experience for railway users. In recent years, there has been an increased focus on this issue in some countries as railway operators have realized that it can play a key role in attracting discretionary travelers (those who are either making non-essential journeys or who have an alternative transport mode readily available) to travel by rail. Providing a good customer experience is particularly crucial when either the standard of service provided by other modes improves or when rail is unable to gain a competitive advantage over other modes based on price or journey time alone.

Customer experience can be influenced by a wide range of issues, with key factors to consider including information provision, service reliability, response to service disruptions, the quality of the station and on-board environment, and the provision of ancillary services. In general operators should aim to make their passengers (and freight customers) feel like valued customers rather than an encumbrance to the smooth operation of the rail service.

Resilience

Planning for resilience is closely linked to planning for a good customer experience, as a resilient railway will usually also be one which is more attractive to customers. The issue of resilience has become increasingly prominent in planning across a range of infrastructure sectors, with increased awareness of both the vulnerability of infrastructure systems to a range of disruptors and the interdependencies between different infrastructure systems. In order to effectively plan for resilience in railway operations it is necessary to obtain a comprehensive understanding of the vulnerabilities and likely failure points within the railway network, and to identify interdependencies with other systems which could lead to cascading or cross-system failures (Pant et al., 2016). Alongside this, operators will also need to carefully consider the maintenance requirements of their infrastructure and rolling stock, and plan this in a way which minimizes disruption to operations. Together this will then facilitate the production of comprehensive contingency plans for a range of potential disruptions, and the identification of trigger points for the implementation of these plans.

Revenue and Other Benefits

The main direct source of revenue for most passenger railways comes from fares, and it is therefore necessary to plan an effective ticketing system and structure for the route or service. Determining what price to charge passengers and what level of price differentiation to apply is not always straightforward, and can have a significant impact on usage patterns. For example, operators may wish to apply time or service quality-based differentiation to ticket prices, charging passengers more to travel at busy times or obtain a higher level of on-board service. Time-based differentiation can also act as a form of demand management in situations

where travel patterns are highly peaked, reducing the need to provide additional peak capacity. Increasingly rail operators are adopting more sophisticated yield-management systems with the aim of maximizing train occupancy, although by reducing flexibility this can potentially have a negative impact on the customer experience. Operators may also offer discounts to specific groups, whether this is to regular travelers in the form of season tickets, or particular demographic groups (such as students or the elderly) in the form of discount cards. As well as planning a suitable fare structure, operators will need to plan for revenue protection in order to ensure that “fare dodging” is minimized. Tickets are not the only source of revenue available to rail operators, with a number of potential additional revenue streams such as car parking and catering outlets. The potential for revenue generation does though need to be set against the risk that this could reduce customer satisfaction, for example, if charges are levied for facilities which are perceived to be “essential” such as toilets or wifi access.

Alongside direct revenue generation, railways can often deliver a range of other societal and economic benefits (Whitelegg, 1987), such as reductions in congestion and pollution and increased levels of accessibility to employment and key services. This means there can often be a strong argument for subsidizing rail operations which do not meet their direct financial costs, but it is important that any claimed benefits for a particular rail route or service are backed up by robust evidence and methodologies in order that the limited money available for subsidizing transport services is allocated to the most appropriate recipients.

Conclusions

The previous sections have provided an overview of the range of factors which need to be considered when planning a railway route or service. It is clear from this discussion that such planning is potentially highly complex, and in order for it to be undertaken effectively planners must consider the interdependencies between railways and other transport, infrastructure and land use systems. Nonetheless, if planned well railways have the potential to play a very significant role in addressing the transport challenges facing 21st century societies. They remain unrivalled in their capability to efficiently transport large numbers of people to and from urban centers and along interurban corridors, and are perhaps easier to decarbonize than any other mechanized transport mode. It therefore seems clear that planning for rail transport will continue to be an important task for the foreseeable future.

Biography

Simon Blainey is an Associate Professor in Transportation at the University of Southampton, with 13 years of experience in transport-related research. He has particular expertise in railway demand and operational modeling, and in strategic transport modeling and planning. He is programme lead for the University of Southampton’s MSc in Transportation Planning and Engineering, and leads the module on “Railway Engineering and Operations.”

See Also

Demand for Passenger Transport; Congestion, Allocation and Competition on the Railway Tracks; Regulation and Competition in Railways; Rail Cost Functions; Rail Freight; Introduction to Rail Transport; The Geography of Rail Transport; Railway Management; Service Network Design for Railroads Moving Freight; Subway Systems (Planning) LPT; Regulation of Rail Infrastructure and Services; Light Rail; High Speed Rail

References

- Armstrong, J., Preston, J.M., 2017. Capacity utilisation and performance at railway stations. *J. Rail Trans. Plan. Manag.* 7 (3), 187–205.
- Nock, O.S., 1979. *Railways of the USA*. Adam and Charles Black, London.
- Pant, R., Hall, J.W., Blainey, S.P., 2016. Vulnerability assessment framework for interdependent critical infrastructures: Case-study for Great Britain’s rail network. *Eur. J. Trans. Infrastruct. Res.* 16 (1), 174–194.
- RSSB, 2011. *Railway Group Standard GC/RT5021 Track System Requirements*, Issue Five.
- Thornes, J.E., Hickman, A., Baker, C., Cai, X., Saborit, J.M.D., 2017. Air quality in enclosed railway stations. *Proc. Inst. Civil Eng. Trans.* 170 (2), 99–107.
- Whitelegg, J., 1987. Rural railways and disinvestment in rural areas. *Reg. Stud.* 21 (1), 55–63.

Further Reading

- Blainey, S.P., Preston, J.M., 2013. A GIS-based appraisal framework for new local railway stations and services. *Trans. Policy* 25, 41–51.
- Blainey, S.P., Preston, J.M. (Eds.), 2021. *Sustainable Railway Engineering and Operations*. Emerald, Bingley.
- Glover, J., 2013. *Principles of Railway Operation*. Ian Allan, Hersham.
- Hall, C., 2016. *Modern Signalling Handbook*, 5th Ed. Ian Allan Publishing, Hersham.

- Hansen, I.A., Pachl, J (Eds.), 2008. Railway Timetable and Traffic. Eurail Press, Hamburg.
- Harris, N., Godward, E.W. (Eds.), 1991. Planning Passenger Railways. A&N Harris, Glossop.
- Institution of Railway Operators (2014) Operators Handbook (2nd edition), Stafford.
- Kichenside, G., Williams, I, 2016. Two Centuries of Railway Signalling, 2nd Edition. Oxford Publishing Company, Shepperton.
- Profillidis, V.A, 2014. Railway Management and Engineering, 3rd Ed. Ashgate, Farnham.

Connected and Autonomous Vehicles: Priorities for Policy and Planning

Dr. Alexandros Nikitas, Huddersfield Business School, University of Huddersfield, Huddersfield, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	167
Projections About CAVs Benefits and Side Effects and Why it is Important to Plan	167
The Key Themes for CAVs Policy: Understanding Where We Are and Avoiding the Bumps Ahead	168
Conclusions and Key Policy and Industry Recommendations	171
Biography	171
References	171
Further Reading	172

Introduction

Connected and Autonomous Vehicles (CAVs) will revolutionize automobility as we know it by passing on the vehicle control and driving responsibilities from the hands of human beings to autopilots with immense Artificial Intelligence (AI), Machine Learning (ML) and wireless connectivity capabilities. More specifically, on the one hand CAVs will be connected, meaning that they will be able to communicate information wirelessly with other vehicles (vehicle-to-vehicle or V2V communications) and with other sources including built environment infrastructure (vehicle-to-infrastructure or V2I communications), control centers supervising the transport system operations and external sources such as weather forecasting operatives (vehicle-to-everything or V2X communications). On the other hand, CAVs will be autonomous, meaning that they will perform driving tasks without human input. CAVs will scan their surroundings through a combination of sensors, cameras, radars and a Light Imaging Detection and Ranging (LIDAR) system and through AI and advanced control systems they will interpret the collected sensory information to detect obstacles and choose the most fitting navigation path.

CAVs are also expected to be the key apparatus for shifting to a transformative automobility narrative crafted by synergies between automation, connectivity, which define the identity of CAVs anyway, but also between electrification and shared use, which are of critical importance if future mobility is to promote sustainability. The synergistic effects between vehicle automation, connectivity, sharing and electrification can multiply significantly the benefits of CAVs as also noted by [Milakis et al. \(2017\)](#). CAVs are projected therefore, to be the immediate successor of the human-driven, conventionally fuelled, privately owned, unconnected vehicle that was the epicenter of urban development and defined the ways cities have been designed and built for many decades ([Nikitas et al., 2019](#); [Thomopoulos and Nikitas, 2019](#)).

Henceforward, the paper introduces, in the next section some of the key opportunities and challenges that may be directly associated with a full-scale launch of CAVs; these are listed as CAVs projected benefits and side effects. The paper also introduces 10 critical areas that policy-makers, planners, and future mobility providers need to look carefully when developing policy instruments, designing planning initiatives, and putting together strategic visions. Finally, a conclusion section describes the pathway to the driverless transition and the role of CAVs in a diverse future urban mobility era.

Projections About CAVs Benefits and Side Effects and Why it is Important to Plan

CAVs have gained traction across the globe with automotive industries, technology providers, ride-hailing companies, and policy stakeholders setting out plans and making multi-billion investments to enter dynamically and from a pole position, the arena of automated, connected and digitized transport. This is because CAVs are projected to be at the epicenter of a new and unprecedented travel eco-system with transformative powers for cities, societies, and economies ([Nikitas et al., 2020](#)).

CAVs may offer, at least in theory, an overwhelming potential for positive change in terms of reducing traffic accident, injury and death rates, road traffic congestion levels and travel delays, CO₂ and greenhouse gas emissions, noise nuisance, energy consumption, mobility-related social exclusion for those currently unable to drive, inefficient routing, parking requirements and enabling improved in-vehicle productivity and relaxed driving experience ([Fagnant and Kockelman, 2015](#); [Nikitas et al., 2017](#); [Thomopoulos and Givoni, 2015](#)).

Nonetheless, at the same time CAVs are considered a disruptive technology that will impose an entirely different mobility paradigm with many diverse challenges, some of which cannot be even predicted for now. If not implemented in a systematic, well-calculated, and cautious way, CAVs could create new barriers relating to vulnerability to hacking, software and hardware flaws, loss of privacy and personal space, increased traffic through greater car dependence or just running empty vehicles, behavioral adaption and user resistance problems, integration problems with other components of the new travel eco-system and loss of many current driving-related jobs ([Nikitas et al., 2020](#)).

The stakeholders responsible for the transition to CAVs should be aware of this wide range of opportunities and challenges, each of which may (or may not) be associated with an eventual large-scale launch of this technology. Effective planning and a pro-active approach are therefore necessary and will enable a smoother and more resource-efficient shift.

The Key Themes for CAVs Policy: Understanding Where We Are and Avoiding the Bumps Ahead

There are many issues that remain uncertain about how this CAV-oriented revolution will unfold. At the same time there are some givens that could inform the direction of the strategic vision and planning behind CAVs implementation; areas that can be highlighted as vital ones and need more research and development contributions. Academia, industry, business, and governing authorities in every level need to intensify their efforts in creating strategies that address these thematic areas.

These critical areas of priority, categorized according to what is known today, refer among others to: *technology maturity; legislation and responsibility; crisis and employment ethics; infrastructure, built environment and land use; integration with the non-connected modes; traffic safety and accident prevention; cyber security and privacy of personal space and data; business models; traffic congestion and travel behavior; and finally acceptability, trust and customer readiness.*

Technology: CAV-related technology has developed a lot. Millions of miles have been already performed by AVs in segregated, in most cases, environments where interaction with other vehicles is closely monitored and significantly controlled. But these early driverless cars are still basic prototypes of what is forecasted to be a fully operational CAV; they can only detect and avoid objects on typical (and usually pre-selected) transport scenarios but are not yet able to anticipate and fully respond to complex traffic behavior that is atypical or unprecedented and usually comes from non-motorized travelers. Teaching the car how to respond to what can be described as ‘the long tail of unlikely events’ such as a plastic bag blowing across the motorway, a couch sitting in the middle of the road (Waldrop, 2015) or unorthodox pedestrian or cyclist behavior is vital, and is an achievement that is yet to be attained in terms of technology (Nikitas et al., 2019). Also, the AV prototypes close to production today, are still seriously lagging in terms of their ability to communicate in real-time and thus connect with other vehicles and the road infrastructure. Technology needs to shape CAVs per se but also re-shape the whole travel eco-system that will host them.

Legislation: Despite being the cornerstone of every socio-technical innovation establishment, legislation has not yet reached the same level of maturity as technology and thus it is still a key barrier to CAVs implementation. In order for CAVs to deliver on their promise without becoming a disruptive technology that for a short-term, at least, could jeopardise the urban landscape equilibrium there is a need for them to operate based on a legal framework that is completely different from the one dealing with the conventional human-driven vehicle of today. Road traffic regulations, liability allocation, responsibility patterns, enforcement strategies, data privacy laws, road standards and safety assurance regimes are not yet fully (or at all) developed and there are no consistent universal standards. Policy-makers need to work systematically in establishing laws and regulations that will support the legal and safe implementation of CAVs.

Crisis and employment ethics: At the same time CAVs are underpinned by a moral dimension, unprecedented in the design of socio-technical innovation, creating questions that unless convincingly addressed they can bring any progress to a stalemate. What should be the computer’s decision-making criteria for choosing from a set of crash trajectory alternatives? What is more important in an unavoidable crash scenario, the self-interest or the public good? CAVs might need to choose between for instance running over pedestrians or sacrificing their passengers to save the pedestrians (Bonnefon et al., 2016) and there is not yet a clear consensus of what is the “right” and “ethical” option. Establishing what constitutes ethical behavior in emergency situations is a hard philosophical problem (de Sio, 2017) that policy-makers primarily, and mobility providers secondarily, need to solve. A second ethical issue of transformative proportions that decision-makers need to address refers to CAVs’ potential to create an unprecedented labour market disruption. How should we respond to the possible loss of millions of driving-related jobs? How can we replace all these jobs on the short- and long-term and how could we re-skill the people affected to be re-employable? Should we even proceed with a full-scale launch of CAVs if we cannot guarantee that the job market will rebound? Those responsible for the transition need to implement CAVs without creating a new colossal layer of employment-related social exclusion.

Infrastructure and land use: Refiguring out and rebranding infrastructure while planning built environments so that they include CAVs must be prioritized by researchers, policy-makers and smart mobility investors. There is a need for building a more tailored road environment that caters for CAVs’ special requirements in terms of connectivity, communication, coordination and digitalization. This smarter road transport environment will need to include sensors, data capture capabilities, state-of-the-art CCTV and traffic signal systems, the ability to be responsive to changes in the environment, electrical charging provision, adaptable street lighting and more importantly the ability to “talk” to different vehicles and road users (Nikitas et al., 2019). Land use policies and plans should support and be supported by CAV services. Transit Oriented Development (TOD) employing CAVs as first and last-mile solutions and neighborhood feeders to more mainstream mass transit systems may be an uneasy union because of their traditionally contradictory roles but future planning efforts need to blend them successfully and propose them as a balanced solution to land use development (Knowles et al., 2020).

Integration: Creating CAVs with the capacity to detect and understand the movement of other forms of non-connected and non-motorized transport modes and respect their travel needs is essential not only for safety and accident prevention but also in terms of achieving efficiencies and traffic optimization. The travel eco-system of the future should not operate under the presence of a CAV-induced dichotomy with some parts of the system understanding and co-functioning with CAVs and some not; there will be the need for a unified, homogeneous and integrated future mobility system. Human and machines speak different languages for now and this

has to change. Human-machine interactions need to be unproblematic, smooth, and effective. Until technology and policy-making bridge this gap between driverless cars and, for example, active travelers, CAVs will not be ready to be safely integrated in the mobility paradigm and maximize their potential in efficiency optimization.

Traffic safety: As long as human drivers are safer than driverless vehicles, it will be unethical to sell them (Sparrow and Howard, 2017). CAVs will need to surpass the human decision-making process when it comes to accident prevention and there is a pathway to do so since human error, the principle factor generating traffic accidents today (Thomas et al., 2013), can be eliminated or minimized with machines driving. By doing so, CAVs will not only help create significantly less traffic crashes, injuries, and deaths but erase some of the key side-effects of accidents, including production losses and massive costs on our healthcare systems. However, if the transport system is a mixed traffic system including large numbers of semi-autonomous and non-autonomous vehicles, low levels of vehicle connectivity and interaction with non-motorized travelers, the possible benefits of CAVs may be in jeopardy. It is still unclear whether CAVs may actually be more prone to accidents in a scenario where there is so much heterogeneity in the system, and where there is limited integration, synchronization, and interoperability.

Cyber security and privacy: One significant factor that may control to a disproportionate degree the acceptability, approval, and uptake narratives is how CAVs could effectively respond to their vulnerability to cyber threats and data sharing exploitation that may have negative effects in both security and privacy. The increased level of the computational functionality and connectivity that CAVs will be able to provide will increase their exposure to prospective vulnerabilities (Parkinson et al., 2017) and opportunities for data mismanagement. CAVs capacity to store and transmit travel, transaction and lifestyle data, make them potentially attractive targets for hackers. Such information can be easily sold for financial gains, or these systems could be used to inflict physical harm by extremists or used for illegal purposes by drug traffickers (Taeihagh and Lim, 2019). Unauthorized hacking and cyber-terrorism on the one hand, and private data sharing and loss of personal space (especially if CAVs end up being shared) on the other, are all policy and planning challenges that need to be addressed effectively for the travelers of the future.

Business models: The transport industry cannot move to the future under the conventional monolithic industrial structures, business models, and operational practices that have defined over 100 years of automotive history. The traditional auto-manufacturers will need to recraft their business approaches, having to compete and form synergies with a new breed of high-tech competitors including mega-players such as Google, Tesla, and Uber. Business models and their respective policy paths need to be flexible and resilient to uncertainties as recommended by Lyons (2016), while their viability should be assessed in cost structure competitiveness terms (Bösch et al., 2018). One of the paradigm-shifting concepts that will revolutionize the current modus operandi is the shared automated vehicle (SAV) which will convert the notion of travel from one that is largely carried out by privately held personal vehicles to fleet services by driverless, demand-responsive vehicles, shared (or for hire) across a mix of users (Fagnant and Kockelman, 2018). Litman (2017) argues that private vehicles might be completely displaced by SAVs that according to Narayanan et al. (2020) can be classified into car-sharing, ride-sharing, and mixed systems. SAVs are projected to be the key to the new holistic transport provision regime known as Mobility-as-a-Service (MaaS) that will replace privately owned transport with personalized mobility packages that give access to multiple travel modes, smart ticketing, and real-time information on a need-only basis via powerful digital platforms. A private ownership-based automation paradigm may be problematic because of the high cost of the vehicles and the need for a higher level of coordination that isolated ownership cannot offer (Nikitas et al., 2019). Private vehicle ownership might still find ways to survive but for the first time there will be a serious possibility that it will be severely overshadowed by car-sharing and ride-sharing services.

Traffic congestion and travel behavior: CAVs, in an ideal scenario, could produce traffic congestion savings due to their adaptive cruise control measures and traffic monitoring systems by minimizing accelerations and braking in freeway traffic and by optimizing road space use (Fagnant and Kockelman, 2015). This could increase fuel economy and congested traffic speeds significantly depending on V2V communication and how traffic-smoothing algorithms are implemented (Atiyeh, 2012). At the same time though, unoccupied vehicle miles traveled could become a real concern that would add an extra layer of road traffic congestion that does not exist today. Although, vehicle ownership may be reduced or even vanished, by providing an unprecedented ease of access, CAVs are projected to enable more people to travel more regularly and for longer distances and could even tempt individuals currently using public transport services to replace some of these trips with SAV services. Unoccupied vehicle trips and unsustainable travel behavior habits may create a congestion-generating puzzle that can only be solved if CAVs are implemented following strict sustainability rules. Investment in automated public transport, travel demand management and traffic calming measures could become even more necessary in the era of CAVs, to balance out these possible planning and policy side-effects.

Customer readiness, acceptability, and trust: Policy-makers and planners need to pro-actively recognize that customers may not be ready yet, despite the entire media buzz, for the monumental cultural transition that a driverless revolution implies. While in hypothetical scenarios, a priori acceptability of CAVs could be likely for many drivers today (Payre et al., 2014) the universal acceptance of CAVs, when these arrive to replace conventional cars, is not definite (Nikitas and Nikitas, 2015; Nikitas et al., 2017). The public may need to overcome socio-psychological barriers that stand in the way of widespread adoption (Shariff et al., 2017) including according to Schoettle and Sivak (2014) and Kyriakidis et al. (2015) concerns about security and cyber security threats, driving performance, unoccupied vehicles increasing road traffic, lacking the joy of driving and regulation absence. Users might need to be persuaded that a shift is safe, sustainable and beneficial to them. Trust is also a process that needs time and transport innovations generally require a period of familiarization (Nikitas et al., 2018). With acceptability being the “maker or breaker” of a smooth shift to CAVs, the key stakeholders need to, on the one hand, invest in marketing campaigns that make customers aware of the benefits of autonomous driving with an emphasis on word-of-mouth and consumer-to-consumer marketing, and on the other hand develop education, training, consultation, and engagement exercises that help people acknowledge what CAVs have to offer.

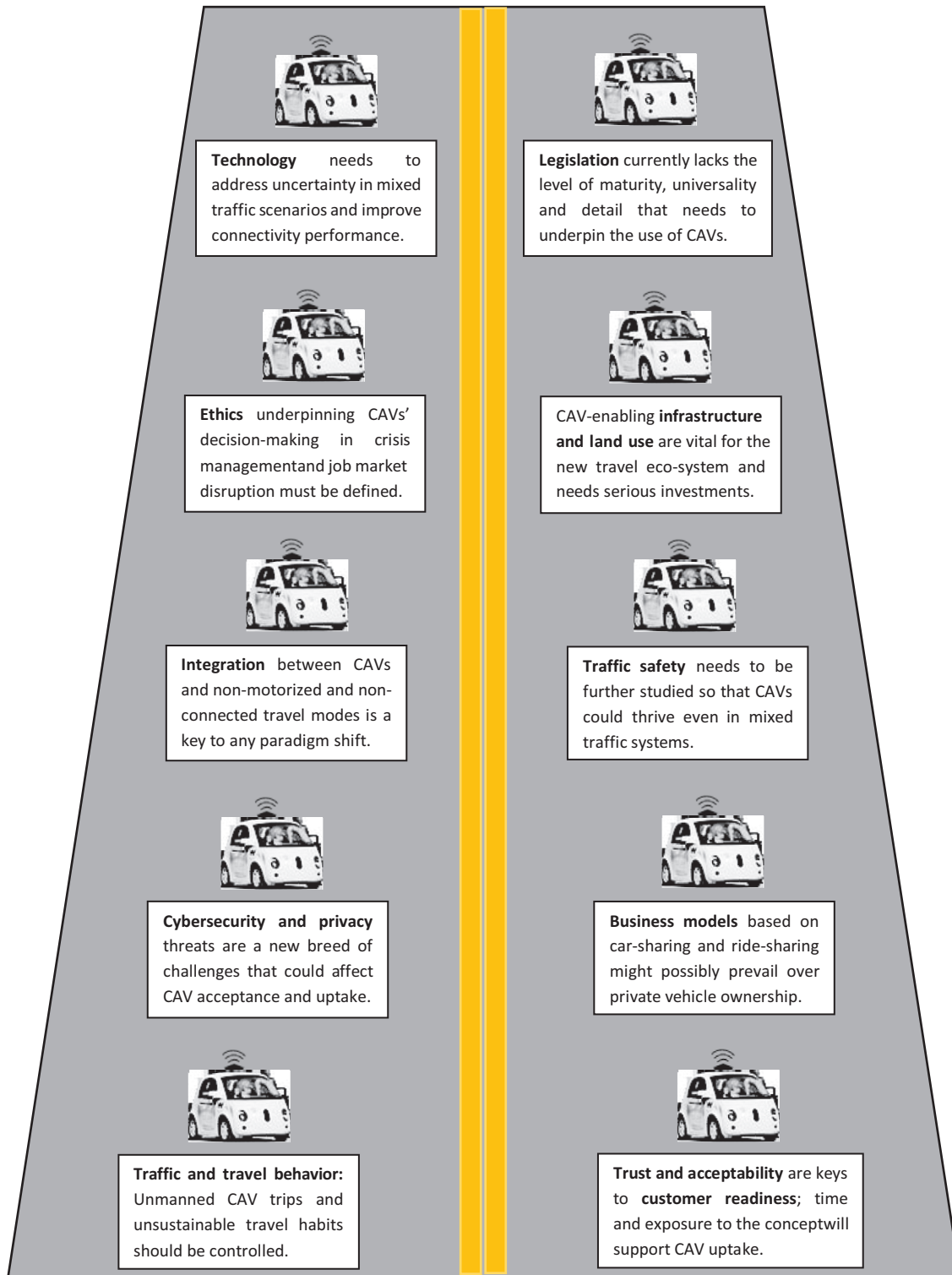


Figure 1 Priority areas for the planning and policy of connected and autonomous vehicles.

Fig. 1 presents an illustration summarizing the key priority areas for the development of CAVs when considering the needs for transport policy and planning. It ultimately provides a concise roadmap about the magnitude and diversity of the steps and challenges underpinning a shift to a CAV-centric mobility paradigm.

Conclusions and Key Policy and Industry Recommendations

Transitioning to CAVs is a lengthy and complicated procedure that requires from policy-makers, planners and mobility providers a strategic approach based on patience, perseverance, and flexibility. Investments on the technology of CAVs alone are not enough; technology on its own, much like Sochor and Nikitas (2016) described it, can only be “one of the several tools in the toolbox of mobility.” Legislation, infrastructure, education, research, and development investments are also critical and should be consistent, balanced, and significant. Implementing CAVs cannot and should not be rushed. This transition will not be about getting there first, but about getting there safely and securely. A full-scale launch must occur only when planning efforts reach a level of maturity that will answer most of today’s unanswered questions. CAVs should incrementally grow on a tactical and systematic basis. They should start from small-scale controlled schemes studied in living labs to allow research and development stakeholders to understand them better, and only then expand in a controlled and calculated way. Piloting and trailing exercises designed to raise awareness and promote engagement with customers and the general public, are critical for any such transition. Some trials may go wrong before the right formula for successfully operating CAVs is found; thus the key stakeholders of future mobility should be ready to respond to failures, accidents, and the backlash that these might generate. It should be clear that CAVs are not a panacea that will fix all that is wrong with transport today; CAVs cannot and will not solve all the traffic problems on their own but they can be a significant component to a more holistic solution (Nikitas et al., 2019). To maximize their potential for positive change and help “building” the smart cities of tomorrow, CAVs might need to potentially work under the principles of MaaS and be complemented by other powerful mobility initiatives like TOD, automated public transport, travel demand management tools, and traffic calming measures.

Biography



Dr. Alexandros Nikitas is a Senior Lecturer in Transport for Huddersfield Business School at the University of Huddersfield, United Kingdom. He is the Deputy Director of the School’s Behavioral Research Center and a co-investigator of the University’s European Structural and Investment Funds project LCR LEP Supply Chain. He is a member of the Editorial Boards of the Journals *Research in Transportation Business & Management* and *Case Studies on Transport Policy* and has produced more than 70 papers and conference outputs in his career thus far. He is conducting research regarding the societal importance of autonomous, connected, shared and sustainable transport. Earlier in his career Dr. Nikitas was a Senior Researcher in Urban Futures and Transportation for Chalmers University of Technology, Sweden where he worked on FP7 and H2020 initiatives. He has been an invited visiting scholar for Tongji University and Chang’an University in China and EC’s Joint Research Centre, Ispra, Italy. He was also a Local Councilor for his hometown Drama, Greece between 2011 and 2014.

References

- Atiyeh, C., 2012. Predicting traffic patterns, one Honda at a time. *MSN Auto*, June, 25, pp. 106–136.
- Bonnefon, J.F., Shariff, A., Rahwan, I., 2016. The social dilemma of autonomous vehicles. *Science* 352, 1573–1576.
- Bösch, P.M., Becker, F., Becker, H., Axhausen, K.W., 2018. Cost-based analysis of autonomous mobility services. *Transp. Policy* 64, 76–91.
- de Sio, F.S., 2017. Killing by autonomous vehicles and the legal doctrine of necessity. *Ethic. Theory Moral Pract.* 20 (2), 411–429.
- Fagnant, D.J., Kockelman, K., 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transp. Res. Policy Pract.* 77, 167–181.
- Fagnant, D.J., Kockelman, K.M., 2018. Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. *Transportation* 45 (1), 143–158.
- Knowles, R.D., Ferbrache, F., Nikitas, A., 2020. Transport’s historical, contemporary and future role in shaping urban development: re-evaluating transit oriented development. *Cities* 99, 102607.
- Kyriakidis, M., Happee, R., de Winter, J.C., 2015. Public opinion on automated driving: results of an international questionnaire among 5000 respondents. *Transp. Res. Traffic Psychol. Behav.* 32, 127–140.
- Litman, T., 2017. *Autonomous Vehicle Implementation Predictions*. Victoria Transport Policy Institute, Victoria, Canada.
- Lyons, G., 2016. Transport analysis in an uncertain world. *Transp. Rev.* 36 (5), 553–557.
- Milakis, D., Van Arem, B., Van Wee, B., 2017. Policy and society related implications of automated driving: a review of literature and directions for future research. *J. Intell. Transp. Syst.* 21 (4), 324–348.
- Narayanan, S., Chaniotakis, E., Antoniou, C., 2020. Shared autonomous vehicle services: a comprehensive review. *Transp. Res. Emerg. Technol.* 111, 255–293.
- Nikitas, A., Avineri, E., Parkhurst, G., 2018. Understanding the public acceptability of road pricing and the roles of older age, social norms, pro-social values and trust for urban policy-making: the case of Bristol. *Cities* 79, 78–91.
- Nikitas, A., Kougiyas, I., Alyavina, E., Njoya Tchouamou, E., 2017. How can autonomous and connected vehicles, electromobility, BRT, hyperloop, shared use mobility and mobility-as-a-service shape transport futures for the context of smart cities? *Urban Sci.* 1 (4), 36.
- Nikitas, A., Michalakopoulou, K., Njoya, E.T., Karampatzakis, D., 2020. Artificial intelligence, transport and the smart city: definitions and dimensions of a new mobility era. *Sustainability* 12 (7), 2789.

- Nikitas A., Nikitas G., 2015. Social dilemmas in vehicle automation. The 2015 Royal Geographical Society (with IBG) Annual International Conference, Special Transport Geography Research Group session in The Spaces of Road Transport Automation, September, Exeter, United Kingdom.
- Nikitas, A., Njoya, E.T., Dani, S., 2019. Examining the myths of connected and autonomous vehicles: analysing the pathway to a driverless mobility paradigm. *Int. J. Automotive Technol. Manage.* 19 (1/2), 10–30.
- Parkinson, S., Ward, P., Wilson, K., Miller, J., 2017. Cyber threats facing autonomous and connected vehicles: future challenges. *IEEE Transact. Intell. Transp. Syst.* 18 (11), 2898–2915.
- Payre, W., Cestac, J., Delhomme, P., 2014. Intention to use a fully automated car: attitudes and a priori acceptability. *Transp. Res. Traffic Psychol. Behav.* 27 (B), 252–263.
- Schoettle, B., Sivak, M., 2014. A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia. University of Michigan, Ann Arbor, Transportation Research Institute.
- Shariff, A., Bonnefon, J.F., Rahwan, I., 2017. Psychological roadblocks to the adoption of self-driving vehicles. *Nat. Hum. Behav.* 1 (10), 694–696.
- Sochor, J., Nikitas, A., 2016. Vulnerable users' perceptions of transport technologies. *Proc. Institut. Civil Eng. Urban Des. Plan.* 169 (3), 154–162.
- Sparrow, R., Howard, M., 2017. When human beings are like drunk robots: driverless vehicles, ethics, and the future of transport. *Transp. Res. Emerg. Technol.* 80, 206–215.
- Taeihagh, A., Lim, H.S.M., 2019. Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transp. Rev.* 39 (1), 103–128.
- Thomas, P., Morris, A., Talbot, R., Fagerlind, H., 2013. Identifying the causes of road crashes in Europe. *Ann. Adv. Automotive Med.* 57, 13.
- Thomopoulos, N., Givoni, M., 2015. The autonomous car—a blessing or a curse for the future of low carbon mobility? An exploration of likely vs. desirable outcomes. *Eur. J. Fut. Res.* 3 (1), 14.
- Thomopoulos, N., Nikitas, A., 2019. Smart urban mobility futures: editorial for special issue. *Int. J. Automotive Technol. Manage.* 19 (1-2), 1–9.
- Waldrop, M.M., 2015. No drivers required. *Nature* 518 (7537), 20.

Further Reading

- Bagloee, S.A., Tavana, M., Asadi, M., Oliver, T., 2016. Autonomous vehicles: challenges, opportunities, and future implications for transportation policies. *J. Modern Transp.* 24 (4), 284–303.
- Fraedrich, E., Heinrichs, D., Bahamonde-Birke, F.J., Cyganski, R., 2019. Autonomous driving, the built environment and policy implications. *Transp. Res. Policy Pract.* 122, 162–172.
- Legacy, C., Ashmore, D., Scheurer, J., Stone, J., Curtis, C., 2019. Planning the driverless city. *Transp. Rev.* 39 (1), 84–102.
- Schwarting, W., Alonso-Mora, J., Rus, D., 2018. Planning and decision-making for autonomous vehicles. *Ann. Rev. Contr. Robot. Autonom. Syst.* 1, 187–210.
- Zmud, J., Goodin, G., Moran, M., Kalra, N., Thorn, E., 2017. Advancing automated and connected vehicles: policy and planning strategies for state and local transportation agencies. National Cooperative Highway Research Program (NCHRP) Research Report 845. doi: 10.17226/24872.

Gendered Mobility

Sheila Mitra-Sarkar, Institute of Public Urban Affairs, San Diego State University and Principal, FutureTrans, San Diego, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	173
Different Facets of Women's Work	173
Gendered Work	174
Gendered Work and Mobility: Complex Social Constructs	174
Women "Opting-Out" of Careers or "Pushed-out" due to Policies	174
Gender Pay Gaps, Family Penalty, and Commuting Costs	175
Flexible Schedules and Telecommuting	175
Telecommuting for Climate Action and Family Friendly Policies	175
Transportation Policies that support Women in Poverty	175
Role of Transit for Female-headed Minority Households	176
Needs of Women in Public Transit	176
Womens's Fear of Victimization	176
Private Vehicle Ownership and Poverty	178
Unpacking Mental Health among Women	178
Gender and Well-Being Studies	179
Importance of Activity Models in Improving Gendered Mobility	179
Discussion	180
Conclusion	181
References	181
Further Reading	183

Introduction

In 1994, Rosenbloom and Burns published their article *Why working women drive alone: Implications of Travel Reduction Programs*. The authors asserted that since the 1970s, there was a significant increase in the role of the car coupled with the declining use of transit and car-pooling. The growing concerns of pollution, and traffic congestion led to the passing of The Intermodal Surface Transportation Efficiency Act (ISTEA) and the Clean Air Act Amendments of 1990. Both acts mandated that employers and governmental agencies must incorporate Travel Demand Management (TDM) programs aim at reducing car trips to work, so they focus on raising fees for parking cars, changes in work schedules, and alternative modes to get to work. Through their study, Rosenbloom and Burns (1994) showed that TDM programs did not alter travel patterns because cars are a necessity and not a luxury for most working women with families. The researchers who had advocated travel reduction programs did not fully understand that punitive measures to reduce vehicle trips would disrupt the work-life balance of working women. In this article, I take the conversation and deep dive into gendered work and how that shapes what I call "gendered mobility".

Different Facets of Women's Work

If we truly want to improve mobility for the women, we must understand and recognize visible and invisible work that they routinely complete. Throughout the day, women make intricate decisions and accompanied trips based on their role as employees and moms. In *The Time Bind*, Hochschild (1997) conducted series of interviews of couples and one of them was Mario and Deb Alcala. Mario was a regular shift worker with 20 additional overtime hours. Deb had 5 h of overtime along with rotating 7-day shift. The couple had three children. Hochschild describes Deb's schedule as a railroad switchman constantly switching her schedule and travel patterns to meet the demands of childcare and work (Hochschild, 1997).

Hochschild's (1997) narrative plays out in every household in the U.S. and in other countries. Women usually are in charge of managing their family schedules. In most cases, the unpaid work hours and the accompanied trips are not fully reported at regional and local levels. OECD (2017) has been collecting data on paid and unpaid work minutes per day by gender for selected countries since 1960s. In 2014, women between the ages 15–64 years worked less hours but had much higher unpaid hours than the men (15–64 years) and for some countries, Greece, Japan, Korea, Turkey, and Portugal there is marked difference in unpaid work each day (OECD, 2017). The effect of gendered work affects the time of the trips, complexity of the trips, trip mode heavily influenced by the socio-economic status of the women. If we want to focus on improving mobility for the women, we have to acknowledge the multiple roles taken by women both as workers and caregivers.

Gendered Work

In the 1960s and early 1990s, progressive middle-class white women entered the labor market (Becker, 2010; Tossi, 2002). By 2000, 74.8% of men and 60% of women in the United States were in the workforce (Blau and Kahn, 2007). It did not take long, for family scholars to observe that women's trips had two demanding sets of responsibilities, one was home-related and the other was work-related (Ridgeway and Correll, 2004). Although the role of women in society changed, they were still expected to take full responsibilities of the socio-emotional needs of the family members (Levinger, 1964; Rothschild, 1983). At the same time, feminist scholars observed that the rise in women's income did not increase men's domestic work. They argued that household labor continued to be mostly assigned to woman. This categorization of household work as "wifely" and "husbandly" roles reproduced dominant and subordinate status based on gender (Berk, 1985; West and Zimmermann, 1987).

Until the 1980s, women's household work trips were not recognized and remained invisible because it was unpaid. Arlene Daniel's (1987) gave "dignity" to women's work (inside and outside home) by reminding researchers that women's unpaid work and trips are complex and should not remain invisible. Other family scholars too have insisted that women's contribution in family, social, and community life that was unpaid must be included along with their paid work (Shelton and John, 1996). Around the same time, sociologists started talking about the emotional support (emotional sustenance, socioemotional role, therapeutic role) provided by women at home, work, and as volunteers (Berk, 1985; Hochschild, 1989; West and Zimmerman, 1987). Emotional support work makes others feel cared for and loved and enhances others' well-being (Erickson, 1996). These activities require time, effort, and skill and is known as "emotion work" (Hochschild, 1979). Erickson (2005) explained that showing appreciation, being a good listener, empathizing with another person's feelings (without reciprocation) for the long haul represents emotion work of the highest order (Erickson, 2005). There is no data collected on the number of minutes of "emotion work" by gender.

Women have self-imposed restrictions on the distances they are willing to travel for work or the number of hours they can work because they shoulder bulk of the routine housework childcare, and chauffeuring services (Ren and Kwan, 2009). Unfortunately, the self-imposed job constraints and/or reduced working hours per week affect their take home pay and career prospects (Golden, 2015).

Women's role in "concerted cultivation" of children is considered unpaid emotional labor (Lois, 2010). In her 2003 book, *Unequal Childhoods*, Lareau explained "concerted cultivation" style is used by the middle-class families to raise their children. The middle-class kids are driven to soccer practice and band recitals, and many other extra-curricular activities throughout the year (McKenna, February 16, 2012). This form of emotional labor requires many unpaid trips. Other literature dealt with the barriers to mobility for those specially-abled mothers and the consideration for "intensive mothering where paid care-takers or nannies take on the hours and trips associated with children (Christopher, 2012; Frederick, 2018). There is indeed a gap in the literature that quantifies the value of unpaid labor and the number of trips in transportation research.

Gendered Work and Mobility: Complex Social Constructs

Right around the time when sociologists, family scholars were writing about the invisible work by women (Berk, 1985; Hochschild, 1979; West and Zimmerman, 1987), feminist geographers were insisting that daily travel patterns from home to work should not be gender blind (Radcliffe, 2006). Transportation scholars started noticing that the women had shorter work trip lengths on average but with complex nonwork components (Hanson and Pratt, 1995; Wachs, 1998). Gordon et al. (1989) analyzed the Nationwide Personal Transportation Study (NPTS) surveys of 1977 and 1983-83 and found these large increases in nonwork trips were taking place mostly during peak hours and that women encountered higher levels of day-time fixed nonwork trip constraints than men. Using 1990 NPTS data, Strathman et al. (1994) found that women completed errands on their way to and from work 42% of the time, while men completed errands 30.6% of the time. The trip chains were higher for women in all categories (simple work, complex work, simple nonwork, and complex nonwork) and they varied with family and life-cycle status (Srinivasan, 2000; Strathman et al., 1994).

In households with children, women create complex trip chains substantially more than women in households without children, or than men. Many of the trip chains are to fulfill the growing need of concerted cultivation, but travel behavior models have not generated models looking into the increased trips due to concerted cultivation. Single mothers, especially of small children, are far more likely to create trip chains than either single fathers or mothers in households with two adults (McGuckin and Murakami, 1999). Nonwork travel is among the more challenging travel activities to study. Most importantly because they vary by socio-economic characteristics and gender. It is nonsystematic, and in many cases, fixed. They constitute the majority of travel activities, including shopping, personal business, and accessing healthcare, schooling, and others (Horner and O'Kelly, 2007).

Women "Opting-Out" of Careers or "Pushed-out" due to Policies

Among married-couple families with children, 63.0% (22.3 million) households had both parents employed (U.S. Bureau of Labor Statistics, 2020). However, for 37% of the families, the father worked and the mother worked part time or did not work. In these households, the women took up most of the household and childcare work.

Workplaces are far from family-friendly (Moore, 2018). In her study of 54 "opt-out" mothers, Stone (2007) found that addition to work, the unavailability, inability, and/or unwillingness of husbands' to take up family responsibilities also played an important

role in quitting. VerBrueggen and Wang (2019) showed that in the US 46% of the mothers whose husbands earned \$250,000 or higher a year were not in the labor force. Also that 35% of mothers whose husbands made less than \$25,000 a year were stay-at-home moms because of unaffordable childcare. While those married to husbands with an income between \$50,000 and \$75,000 are the least likely to stay at home; only 25% of them were out of the labor force.

Gender Pay Gaps, Family Penalty, and Commuting Costs

Traditional gender roles and women's greater responsibility for unpaid work has continued to negatively affect women's labor market outcomes. Blau and Kahn (2017) used Panel Study of Income Dynamics (PSID) microdata over the 1980–2010 period to assess the wage gap by gender. They reported that the female/male earnings ratio was around 60%. Blau and Kahn (2017) found that women's relative wages began to rise sharply in the 1980s, but became slower in the later years. By 2014, women full-time workers earned about 79% of what men did on an annual basis and about 83% on a weekly basis (Blau and Kahn, 2017). Economists mention that certain gender-specific factors explain the lower earnings of women. Some of these factors that cause wage differentials such as shorter hours, when children are young, changes the balance between wages and commuting costs as the commuting cost is not fully recovered.

Flexible Schedules and Telecommuting

While women have forged ahead in labor participation, the gendered role has caused distinct pressure to seek professions that enable them to fulfill the work associated with homes (Chung and van der Horst, 2017). Women select jobs that allow them to have a flexible schedule to support nonwork trips (Williams and Bornstein, 2006). The concept of flexibility or schedule control focuses on control over *when* work is done instead of *how* it is done (Kelly and Moen, 2007). This concept if used well, could assist women to keep their jobs and not worry about how to create work-life balance and reduce the need to travel.

Telecommuting workers are more likely to be managerial and professional workers with higher salaries and higher educational attainment. Schedule control in Germany showed that men benefited more from flex-control through overtime and income because when mothers work from home, they are more likely than fathers to do so during standard work hours (Glass and Noonan, 2016).

The Federal Highway Administration (FHWA) has been collecting data on the number of people working from home. Based on this data for the US, it seems that the telecommuting jobs are geared more towards the managers, and those who are low-income (Table 1).

Telecommuting for Climate Action and Family Friendly Policies

In the US, The Clean Air Act amendments of 1990 established telecommuting programs in cities with severe air quality conditions to investigate whether a reduction in emissions could be achieved. The 2010 Telework Enhancement Act renewed provisions for telecommuting policies in US federal agencies (Lari, 2012). While telecommuting is still often presented as a sustainable travel policy, its use remains limited and its effects on reducing overall travel time and peak period travel unclear. The researchers and employers need to shift the focus of telecommuting from an environment-friendly policy to a family friendly, gender friendly policy (Glass and Noonan, 2016).

Transportation Policies that support Women in Poverty

In 2016, women made up nearly 67% of the 24 million workers in low-wage jobs in the US (defined as jobs that typically pay \$11.50 per hour or less), though they make up slightly less than half (47%) of the workforce. Low-wage workers are a racially diverse group, and disproportionately females. Ross and Bateman (2019) reported that 52% are white, 25% are Latino or Hispanic, 15% are Black,

Table 1 Longitudinal trend in home-based Working by Income, Occupation and Home Location

Poverty level	1995	2001	2009	2017
Household Income				
<\$50,000	59	28	17	30
≥\$50,000 and <\$75,000	21	20	16	15
≥\$75,000 and <\$100,000	11	21	22	12
≥\$100,000	10	31	46	42
Occupation				
Sales/service		23	20	29
Clerical/administrative support		7	6	7
Manufacturing/construction		8	6	13
Professional/managerial/technical		62	67	51
		0	1	0

Source: Own Working from Federal Highway Administration (2017). Highway Statistics 2017 - Policy Federal Highway Administration. [online] Dot.gov. Data from: Table 1.

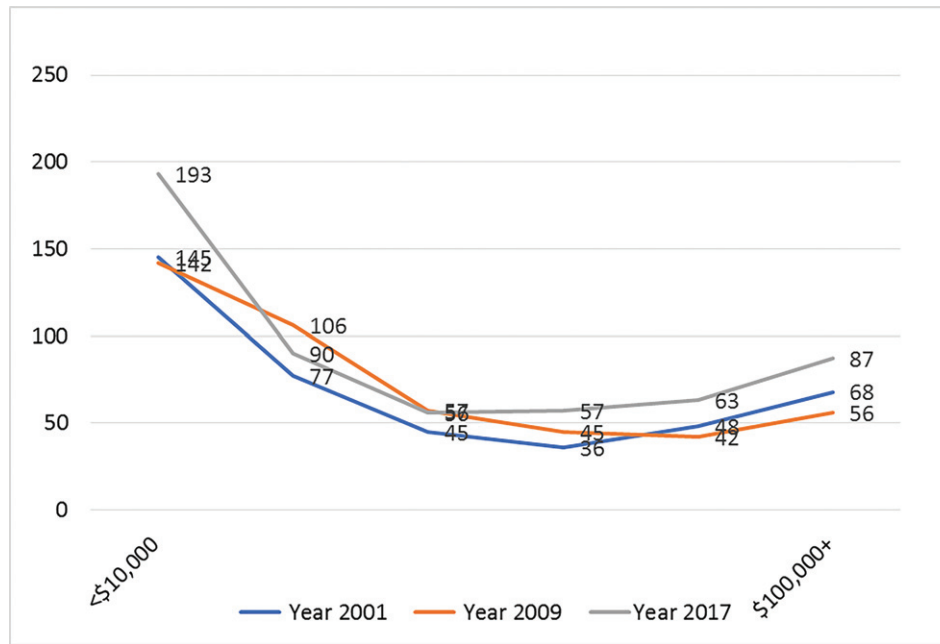


Figure 1 Number of Annual Transit Trips (in millions) by income range. Source: Own Working from: [Federal Highway Administration \(2019\)](#). *Transit Trend Analysis 2017*. [online] National Household Travel Survey. Data from: Table 2.

and 5% are Asian American. Those below the poverty level are disproportionately Blacks (21.7%) and Hispanics (18.3%), and proportionately more women are below the poverty level ([Ross and Bateman, 2019](#)). These women are in no position to influence policy changes or advocate for improved transportation system ([Ferragina, 2013](#); [McLanahan et al., 1994](#); [Townsend, 1987](#)).

Role of Transit for Female-headed Minority Households

A disproportionate number of female-headed households used transit because it provides mobility for those who are carless ([Blumenberg and Pierce, 2012](#); [Rosenbloom, 2006](#)). African-American women and Latina with low average wages rely more on public transit when compared to white women in the US ([McLafferty and Preston, 1991](#); [Metro, 2019](#)).

Although the overall number of transit trips have dropped in 2017 for all income groups, the households with lower income made the most transit trips ([Fig. 1](#)). Immigrants in general, commute by public transit at twice the rate of native-born commuters in metro areas ([Fig. 2](#)). However, since access to jobs is closely related to how reliable and efficient the public transport systems are, immigrants may start off by being transit-dependent but as years pass, they progressively opt for cars over transit as seen in [Fig. 3](#) ([Blumenberg and Evans, 2010](#)).

Needs of Women in Public Transit

Women with better access to transit have lower rates of unemployment. Efficient public transportation systems enable women to access jobs and economic opportunities ([Blumenberg and Pierce, 2012](#)). And yet, transit agencies fail to recognize that due to gendered division of labor, women have different travel patterns than men. In an initiative led by Metro's Women and Girls Governing Council, Los Angeles County Metropolitan Transportation Authority (LA Metro), became the first transit agency in the U. S. that acknowledged that there was a gender data gap and published, *Understanding How Women Travel (UHWI)*. After collecting qualitative and quantitative gender data through a survey that was completed by 2600 residents; qualitative data was collected through pop-up engagements at key rail stations; and focus groups with women riders who were diverse, such as, immigrants, have disabilities, or are experiencing homelessness ([Metro, 2019](#)).

Women's Fear of Victimization

Transit users spend a considerable amount time in public places, bus stops, transit stations, transit vehicles. Women's safety is most compromised in such place and in public transport. Mass transit moves large amounts of people through a city and shapes the crime pattern along a limited number of paths. The women who responded to Metro's survey identified safety concerns as the top barrier to use transit. Over 80% of women transit users felt unsafe walking to transit stations, waiting for transit and riding transit in the

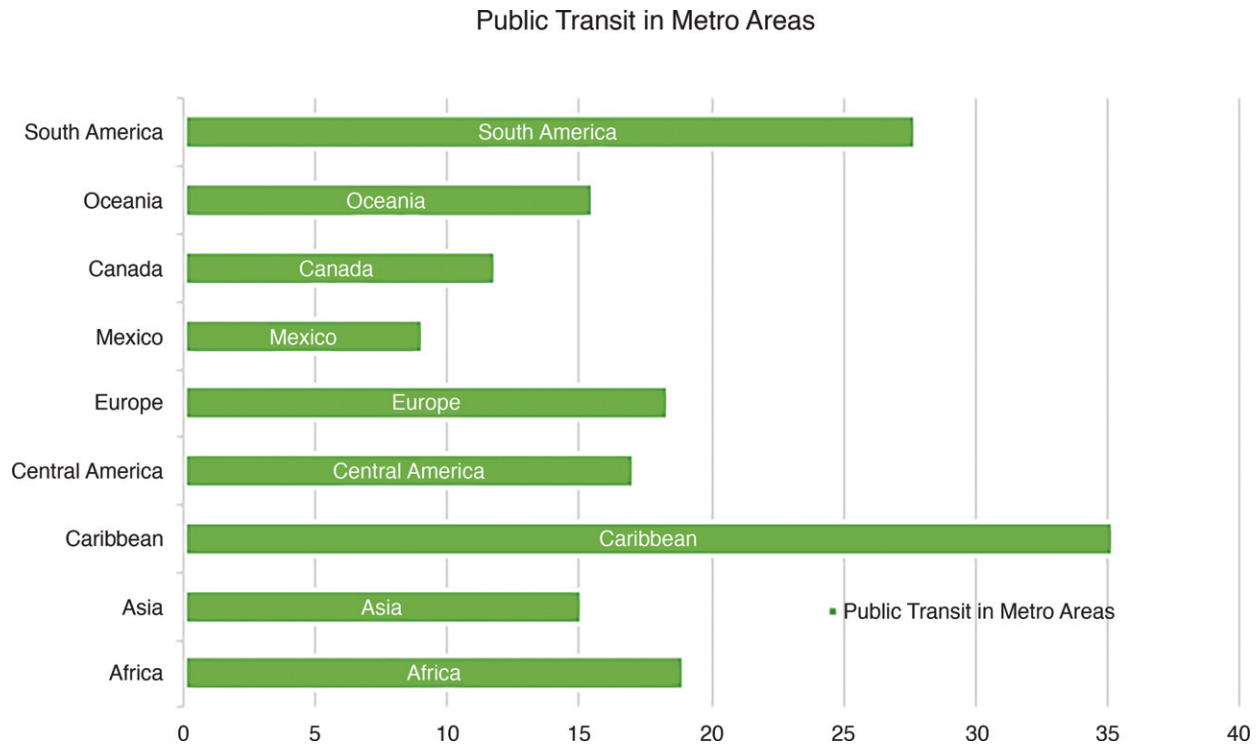


Figure 2 Transit usage (Percent) by place of birth among immigrant workers. Source: Own Working from Chatman, D.G. and Klein, N. (2009). *Immigrants and Travel Demand in the United States. Public Works Management & Policy*, 13(4), pp.312–327. Data from: Table 6.

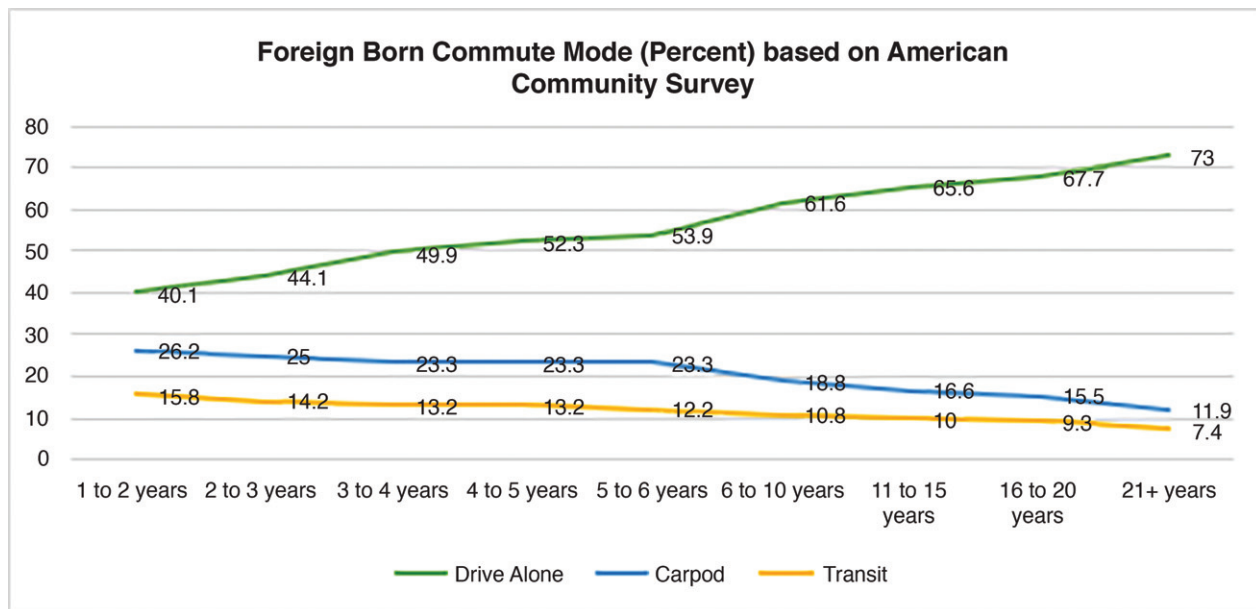


Figure 3 Foreign Born Commute Mode (Percent) based on American Community Survey. Source: Own Workings from Chatman and Klein (2009) Table 3 *Commute Mode by Nativity and Years in the US* p. 315.

evening Men felt unsafe too, but the responses were lower, although waiting at the transit stops was found to be least unsafe (70%) (Metro, 2019).

It is difficult to objectify women's fear when they are in public places at night. Desolation, darkness and lack of activity or natural surveillance in a transportation setting contribute to anxiety and the fear that no one will be there to help if a crime occurs (Loukaitou-Sideris, 2006). Fear of victimization in public settings is influenced by age, race, class, cultural and educational

Table 2 Responses (Percent) of Women by age and transit usage (300-700 respondents who responded to these question)

	Young	Inactive	Casual
Depending on the place			
Station and surroundings	44	28	39
When the transport is running	36	32	45
Depending on this circumstance			
In poorly lit	28	18	25
When people show bad manners	55	43	58
Depending on this circumstance			
When there are not many passengers	49	37	54
Weekdays	21	10	14
Time of the Day			
8.30 p.m.- 10.30 p.m.	26	19	27
After 10.30 p.m.	23	12	20

Source: Own Working from: D'Arbois de Jubainville, H. and Vanier, C. (2017). Women's avoidance behaviors in public transport in the Ile-de-France region. *Crime Prevention and Community Safety*, 19(3-4), pp.183-198. Data from Table 3.

background, sexual orientation, prior victimization experiences, and disability status (Gardner et al., 2017). Fear modifies choices women make as pedestrians and passengers (time of journey, where they wait for a train, where on a train they sit). Whether being rooted in real or only perceived danger, fear has significant effect on women's travel patterns. Women adopt a range of behavioral mechanisms when in public to choosing specific routes, modes, and transit environments over others to completely avoiding particular transit environments, bus stops and/or walking and biking (Condon et al., 2007; Loukaitou-Sideris, 2006). How does this fear relate to social exclusion and trip avoidance?

The micro-spaces of transportation nodes influence the perception of safety in systems. Micro-spaces refer to façade, the density and height of the buildings, the number and types of streets and entrances, transit stops but also urban design features, such as seating, storefronts and other physical environments (Ceccato and Moreira, 2020; Loukaitou-Sideris, 2006). The very nature of transit stops facilitate robberies. Transit stops are public spaces which allow easy entry and exit, and standing around is not deemed suspicious and the potential targets are most likely walking (Block and Block, 2000). Some researchers contend that subway stations are also crime generators (Block and Block, 2000; Piza and Kennedy, 2003). While analyzing the French national victimization survey (2010-13), D'Arbois de Jubainville and Vanier, 2017 found that young women feared traveling late in the evening felt most unsafe in station surroundings, presence of rude passengers, and when there are fewer people in the stations (Table 2). Surveys on security conducted in the United State, United Kingdom, and in India since the mid-1980s found that waiting at the bus stop was a major source of concern and fear among women (Levine et al., 1986; Loukaitou-Sideris, 1999; Mitra-Sarkar and Partheeban, 2011).

Private Vehicle Ownership and Poverty

Public transit systems in most metropolitan areas are slow, inconvenient, and lack sufficient coverage to rival the automobile (Pendall et al., 2014). Whilst many researchers have argued that private vehicle ownership is associated with higher incidence of employment (Blumenberg and Pierce, 2016; Smart and Klein, 2018), the costs of owning and maintaining a car may be greater than the income gains associated with car ownership. Low-income buyers are least likely to get the best prices while buying cars.

Unpacking Mental Health among Women

The US economy defines an "ideal worker" as a male full-time worker who puts in the hours no matter what and is willing to relocate if necessary (Williams, 2004). Naturally, a female is needed to assist this "ideal worker" to shine (Cerrato and Cifre, 2018; Woods and Eagly, 2015). Unfortunately, such expectations causes work-family conflict and stress. The notion that women stand up to stress better than men appear to mean that women can perform the lion's share of domestic work and childcare and not feel the stress (Bianchi, 2011; Coltrane and Messino, 2000).

Almeida and Kessler (1998) developed a personal interview to assess the occurrence of daily stressors called the Daily Inventory of Stressful Events (DISE). Almeida et al. (2020) used 2010 DISE on the U.S. National sample of adults aged 25-74 ($N = 1031$), to determine the prevalence as well correlates of daily stressors. Women had more frequent days in which they reported one stressful event, $t(1030) = 5.1$, $P < .01$ and reported higher levels of physical symptoms, $t(1030) = 3.9$, $P < .01$.

The magnitude and types of stressors that women face is different. Torquati et al. (2019) included seven longitudinal studies, with 28,431 unique participants to find out if shift work was associated with an increased overall risk of adverse mental health outcomes. They found that female shift workers were more likely to experience depressive symptoms than female non-shift workers (odds

ratio = 1.73; 95% CI = 1.39, 2.14). In 2019, the World Health Organization (WHO) recognized burnout as a syndrome linked to chronic stress at work. Burnout is commonly defined as Malach's triad of emotional exhaustion, depersonalization or cynicism, and feelings of diminished personal accomplishment in the context of the work environment (Malach-Pines, 2005).

The consequences of work-home conflict and their impact on depression have been reported as early as 6 months into the first postgraduate year and are more common among women physicians than among men (Guille et al., 2017). Women physicians employed full time spend 8.5 additional hours per week on child care and other domestic activities, including care for elderly parents (Jolly et al., 2014).

Improved technology has made working from home more attractive but the line between family and work have diffused producing work-family conflict (WFC) which is considered as a psychosocial risks (Cerrato and Cifre, 2018). WFC negatively affects both health and general life such as work performance and work satisfaction and also decreases family satisfaction (Cerrato and Cifre, 2018).

Gender and Well-Being Studies

Well-being refers to optimal psychological experience and functioning (Stevenson and Wolfers, 2009). A significant body of research has used Diener and Emmon's (1984) Subjective Well-Being (SWB) and Ryff's (1989) research on Psychological Well-Being (Huppert, 2009; Ryan and Deci, 2006). Subjective Well-Being scales have also been used to identify whether the commute to work has an impact on the SWB of workers (Choi et al., 2013; Ettema et al., 2010, 2012; Novaco and Gonzalez, 2009). These studies have shown that commuting has not increased happiness for men and women, but it adversely affects the women's well-being even more so. Women make decisions about commuting under a different set of constraints than men, and commuting takes away time from competing demand of trip-chaining and day-time fixity constraints (drop-off and pick-up). Employers fail to recognize how the commute affects staff differently according to their life situation (Roberts et al., 2011).

Stevenson and Wolfer examined men's and women's subjective well-being in the United States over 35 years using data from the General Social Survey (GSS). This survey is a nationally representative sample of about 4500 respondents since 1972. The authors found happiness greater for men relative to women. This finding of a decline in women's well-being relative to that of men, raises questions about whether modern social constructs (work and life) have made women worse off.

Social exclusion is often misunderstood, poorly defined and poorly measured although a key determinant of psychological well-being of single parents, elderly in poverty, unemployed, and immigrants (Santana, 2002). Social exclusion refers to a disadvantage, cause by consequences such as the absence of transport to enable personal mobility and affects individuals and groups who are denied access to the goods, services, activities, and resources (Social Exclusion Unit, 2003; Sen, 2000). Research has shown that increasing trip making and improving a person's social capital and sense of community is likely to reduce risks of social exclusion and increase wellbeing (Lucas, 2012).

Importance of Activity Models in Improving Gendered Mobility

Activity-based models and the study of entire trip chains became the focus of model development in the 1990s (Misra and Bhat, 2000). The activity-based travel demand analysis viewed travel as a derived demand to complete household activities (Axhausen and Gärling, 1992). Activity-based travel research has received much attention from travel demand modelers because they best capture complex travel episodes, such as multipurpose and multistep travel (Bhat and Koppelman, 2003; McNally and Rindt, 2008). The five important features include:

1. Travel derived from activity participation;
2. Activity-based approach that focuses on sequences of patterns of activities;
3. Individual's activity both planned and executed in the household (family) context;
4. Activities spread throughout a 24-hour period in a continuous manner, rather than discrete events;
5. Travel and location choices could be constrained (work, picking up, school) or unconstrained (shopping and leisure) (Chu et al., 2012)

Activity-based travel modelers have broadened the data collection adding new dimensions to include nonworkers. Misra and Bhat (2000) analyzed weekday activity travel data of 1190 nonworkers with 2327 individuals with at least one out of home activity for the city of San Francisco. The mean age was 57.5 (SD = 17.7), mostly Caucasian and slightly higher number of females. The study was important because it was one of the first to analyze the activity patterns of the nonworking households. Their exploratory analysis corroborated some of the earlier findings on activity sequencing while adding a better understanding of the activity patterns of older couples.

The activity-based modeling has the potential to be robust indicators of socio-economic health of the workers and nonworkers. Indeed social exclusion and material deprivation should be important components of any activity-based models that looks at gender differences. Transportation researchers have begun to include nonwork trips and trip chaining in their analysis and have found that women continue to do complex non-work trips, unfortunately, many of these trips still remain invisible as the data often remains uncaptured and there have not been enough interest to look at intra-gender differences in activity based modeling (Susilo et al., 2012; Uteng and Cresswell, 2012).

Discussion

Feminist transportation planners are becoming increasingly aware that “gendered mobility” affects women’s economic mobility, career choices and psychological well-being. Women make up almost half of the workforce, and a record 40% are breadwinners for their families. According to the 2018 US Census, workers in female-householder families (mostly African-Americans and Hispanics) worked full-time, year-round at a greater rate. And, women continue to be over-represented in occupational roles that allow flexibility, such as customer service, healthcare, and counseling occupations, where they are faced with emotionally charged encounters where they have to remain reserved (Glomb and Tews, 2004; Hochschild, 1989). And, yet, transportation agencies and policy makers have not made it a priority. In 2014, the Planning and Women Division of the American Planning Association (APA) and the Women’s Planning Forum of Cornell University, surveyed 300 local governments to ask about whether plans account for the needs of women. The overwhelming majority (98%) responded that they do not (Khanna, 2002). Gender issues have been least accommodated in transit planning although women are the predominant users (Loukaitou-Sideris et al., 2002; Metro, 2019; McLafferty and Preston, 1991).

There is a desperate need to spend time and understand what is truly expected of each woman to accomplish her work every day, both as a worker and as a mother in many cases. There are still a considerable number of questions as to the issues of why women “opt-out” of careers. Why are employers unwilling to give flexible working to women? Why women are cautious about choosing flex schedule? Why female doctors who graduate at the same rate as the men then opt for fewer hours? The OECD (2014) study showed that in every country, men are able to fit in valuable extra minutes of leisure each day while women spend more time doing unpaid housework.

Emotional labor demands have effects on career mobility, occupational segregation, poor job security and satisfaction, wage gap stress and well-being. As mentioned earlier, low wages make women responsible for more housework, childcare and emotional work at home and men can “buy” out time by getting higher pay. Women continue to make decisions about commuting under different set of constraints than men and it becomes a self-fulfilling cycle of seeking part-time or low paying jobs (Brines, 1994; Erickson, 2005). Commuting takes away time from competing demands, and women opt for shorter commutes and lower bargaining power (Chatterjee, 2019).

In 2001, the gap between the number of miles driven by men and women were considerably higher for all age groups (Fig. 4A), this dropped by 2017 but men still drive more miles for all age groups (Fig. 4B). There is a need of a better understanding of these numbers and find out whether these numbers translate into fewer opportunities in career, social exclusion, material deprivation and adverse mental health outcomes for women.

Women have to negotiate cultural constructs of family responsibilities and more recently, neurosexism (Fine, 2008) where working mothers struggle against the natural wiring of female brains to be the caregivers. The tug of war between career circuits and maternal circuits increase stress and anxiety and diminishes brainpower for mothering (Brizendine, 2007; as explained by Fine, 2008).

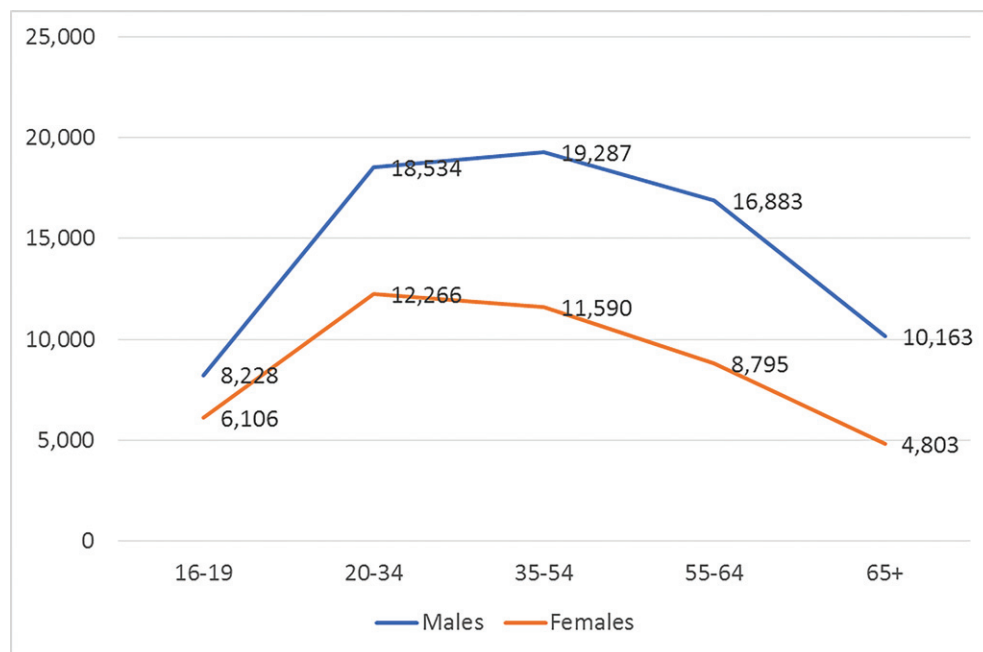


Figure 4 (A) Average Annual Miles per Licensed Driver by Gender 2001 in the US. (B) Average Annual Miles per Licensed Driver by Gender 2017 in the US. Source: Own Working Federal Highway Administration Highway Statistics 2011.

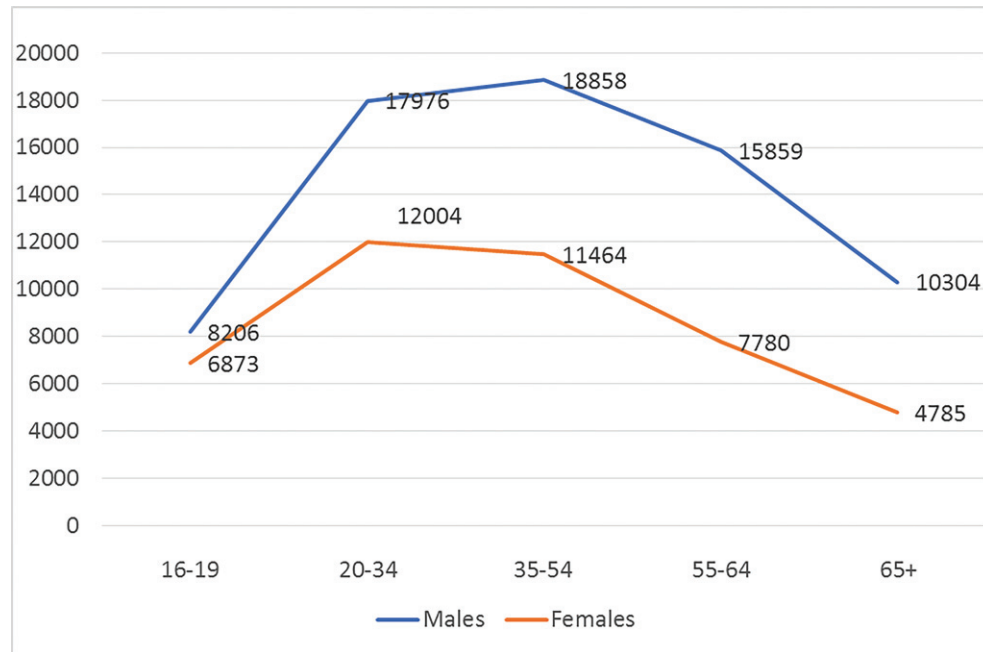


Figure 4 (Continued)

Conclusion

There is a lot to be done to first address the data gap and to instill in transportation researchers that gender is crosscutting. There is also the need for researchers to start questioning whether they should design research to change policies and infrastructure that perpetuate gender bias and stereotypes.

References

- Almeida, D.M., Charles, S.T., Mogle, J., Drewelies, J., et al., 2020. Charting adult development through (historically changing) daily stress processes. *Am. Psychol.* 75 (4), 511–524.
- Almeida, D.M., Kessler, R.C., 1998. Everyday stressors and gender differences in daily distress. *J. Personal. Social Psychol.* 75 (3), 670–680.
- Axhausen, K.W., Gärling, T., 1992. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport Rev.* 12 (4), 323–341.
- Becker, D., 2010. Women's work and the societal discourse of stress. *Feminism Psychol.* 20 (1), 36–52.
- Berk, S.F., 1985. *The Gender Factory: The Apportionment of Work in American Households*. [online] [www.springer.com](https://www.springer.com/gp/book/9781461294610). Springer US. Available at: <https://www.springer.com/gp/book/9781461294610> [Accessed 15 Sep. 2020].
- Bhat, C.R., Koppelman, F.S., 2003. Activity-based modeling of travel demand. In: Hall, R.W. (Ed.), *Handbook of Transportation Science*. International Series in Operations Research & Management Science, vol 56. Springer, Boston, MA, doi:10.1007/0-306-48058-1_3.
- Bianchi, S.M., 2011. Family change and time allocation in American families. *Ann Am. Acad. Polit. Soc. Sci.* 638 (1), 21–44.
- Blau, F.D., Kahn, L.M., 2007. The gender pay gap: have women gone as far as they can? *Acad. Manage. Perspect.* [online] 21 (1), 7–23. Available at: <http://www.jstor.org/stable/4166284> [Accessed 25 Oct. 2020].
- Blau, F.D., Kahn, L.M., 2017. The gender wage gap: extent, trends, and explanations. *J. Econ. Lit.* 55 (3), 789–865.
- Block, R., Block, R.B., 2000. The Bronx and Chicago - Street robbery and the environs of rapid transit stations. In: Goldsmith, V., McGuire, P.G., Mollenkopf, J.H., Ross, T.A. (Eds.), *Analysing Crime Patterns: Frontiers and Practice*. Sage, Thousand Oaks, pp. 137–257.
- Blumenberg, E., Evans, A., 2010. Planning for demographic diversity: the case of immigrants and public transit. *J. Public Transport.* 13 (2), 23–45.
- Blumenberg, E., Pierce, G., 2012. Automobile ownership and travel by the poor. *Transport. Res. Rec.: J. Transport. Res. Board* [online] 2320 (1), 28–36. Available at: <https://journals.sagepub.com/doi/10.3141/2320-04> [Accessed 23 Nov. 2019].
- Blumenberg, E., Pierce, G., 2016. The drive to work. *J. Plan. Educ. Res.* 37 (1), 66–82.
- Brines, J., 1994. Economic dependency, gender, and the division of labor at home. *Am. J. Sociol.* 100 (3), 652–688. Retrieved January 13, 2021, Available at: <http://www.jstor.org/stable/2782401>.
- Brizendine, L., 2007. *The Female Brain*. Bantam Press, London.
- Bureau of Labor Statistics, 2020. U.S. Department of Labor, Employment Characteristics of Family—2019 <https://www.bls.gov/news.release/pdf/famee.pdf> Accessed on April 30, 2020.
- Ceccato, V., Moreira, G., 2020. The dynamics of thefts and robberies in São Paulo's Metro, Brazil. *Eur. J. Crim. Policy Res.*.
- Cerrato, J., Cifre, E., 2018. Gender inequality in household chores and work-family conflict. *Front. Psychol.* 9.
- Chatman, D.G., Klein, N., 2009. Immigrants and travel demand in the United States. *Public Works Manage. Policy* 13 (4), 312–327.
- Chatterjee, K., Chng, K., Clark, S., Davis, B., et al., 2019. Commuting and wellbeing: a critical overview of the literature with implications for policy and future research. *Transp. Rev.* 40 (1), 5–34.
- Choi, J., Coughlin, J.F., D'Ambrosio, L., 2013. Travel time and subjective well-being. *Transport. Res. Rec.: J. Transport. Res. Board* [online] 2357 (1), 100–108. Available at: <https://trid.trb.org/view/1241986> [Accessed 22 Jul. 2020].
- Christopher, K., 2012. *Transport. Res. Rec.: J. Transport. Res. Board* [online] 26 (1), 73–96.
- Chu, Z., Cheng, L., Chen, H., 2012. A review of activity-based travel demand modeling. *CICT*.

- Chung, H., van der Horst, M., 2017. Women's employment patterns after childbirth and the perceived access to and use of flexitime and teleworking. *Human Relat.* 71 (1), 47–72.
- Coltrane, S., Messineo, M., 2000. The perpetuation of subtle prejudice: race and gender imagery in 1990s television advertising. *Sex Roles* 42 (5/6), 363–389. Available at: <https://link.springer.com/article/10.1023/A:1007046204478> [Accessed 17 May 2020].
- Condon, S., Lieber, M., Maillochon, F., 2007. Feeling unsafe in public places: understanding women's fears. *Revue française de sociologie* 48 (5), 101.
- DeVault, M.L., 1991. Family discourse and everyday practice: gender and class at the dinner table. *Syracuse Scholar* (1979–1991) 11 (1), Article 2.
- Diener, E., Emmons, R.A., 1984. The independence of positive and negative affect. *J. Personal. Social Psychol.* 47 (5), 1105–1117.
- D'Arbois de Jubainville, H., Vanier, C., 2017. Women's avoidance behaviours in public transport in the Ile-de-France region. *Crime Prevent. Commun. Safety* 19 (3–4), 183–198.
- Erickson, B.H., 1996. Culture, class, and connections. *Am. J. Sociol.* 102 (1), 217–251.
- Erickson, R.J., 2005. Why emotion work matters: sex, gender, and the division of household labor. *J. Marr. Family* 67 (2), 337–351.
- Ettema, D., Friman, M., Gärling, T., Olsson, L.E., Fujii, S., 2012. How in-vehicle activities affect work commuters' satisfaction with public transport. *J. Transport Geogr.* 24, 215–222.
- Federal Highway Administration, 2014. 2009 Poverty Status of Selected Groups. [online] Available at: <https://nhts.ornl.gov/briefs/PovertyBrief.pdf> [Accessed 8 Aug. 2020].
- Federal Highway Administration, 2019. Highway Statistics 2017 - Policy Federal Highway Administration. [online] Dot.gov. Available at: <https://www.fhwa.dot.gov/policyinformation/statistics/2017/> [Accessed 15 Nov. 2019].
- Ferragina, E., 2013. The socio-economic determinants of social capital and the mediating effect of history: Making Democracy Work revisited. *Int. J. Comparative Sociol.* 54 (1), 48–73.
- Fine, C., 2008. Will working mothers' brains explode? The popular new genre of neurosexism. *Neuroethics* 1, 69–72. Available at: <https://doi.org/10.1007/s12152-007-9004-2>
- Frederick, A., 2018. Disabling fields, enabling capital: mothers with disabilities and the concerted cultivation habitus. *Disability Stud. Quart.* [online] 38 (4). Available at: <https://dsq-sds.org/article/view/6162/5141> [Accessed 16 Jul. 2020].
- Gardner, N., Cui, J., Coiacetto, E., 2017. Harassment on public transport and its impacts on women's travel behaviour. *Austr. Planner* 54 (1), 8–15.
- Glass, J.L., Noonan, M.C., 2016. Telecommuting and earnings trajectories among American women and men 1989–2008. *Social Forces* 95 (1), 217–250.
- Globb, T.M., Tews, M.J., 2004. Emotional labor: a conceptualization and scale development. *J. Vocational Behav.* 64 (1), 1–23.
- Golden, L., 2015. Irregular Work Scheduling and Its Consequences Policy Brief 394. [online] Economic Policy Institute. Available at: <https://www.epi.org/publication/irregular-work-scheduling-and-its-consequences/>.
- Gordon, P., Kumar, A., Richardson, H.W., 1989. Gender differences in metropolitan travel behaviour. *Reg. Stud.* 23 (6), 499–510.
- Guille, C., Frank, E., Zhao, Z., Kalmbach, D.A., et al., 2017. Work-family conflict and the sex difference in depression among training physicians. *JAMA Inter. Med.* 177 (12), 1766. Available at: <https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2659558> [Accessed 5 Feb. 2020].
- Hanson, S., Pratt, G., 1995. *Gender, Work, and Space*. Google Books. Routledge.
- Hochschild, A.R., 1989. *The Second Shift: Working Parents and the Revolution at Home* by Abe Books. [Accessed 25 Jul. 2020].
- Hochschild, A., 1979. Emotion work, feeling rules, and social structure. *Am. J. Sociol.* 85 (3), 551–575. Available at: <http://www.jstor.org/stable/2778583>.
- Hochschild, A.R., 1997. *The Time Bind: When Work becomes Home Becomes Work*. Metropolitan Books, New York [Accessed 25 Jul. 2020].
- Horner, M.W., O'Kelly, M.E., 2007. Is non-work travel excessive? *J. Transport Geogr.* 15 (6), 411–416.
- Huppert, F.A., 2009. Psychological well-being: evidence regarding its causes and consequences. *Appl. Psychol. Health and Well-Being* 1 (2), 137–164.
- The Intermodal Surface Transportation Efficiency Act (ISTEA), 1991. Clean Air Act Amendments <https://www.congress.gov/bills/102nd-congress/house-bill/2950>.
- Jolly, S., Griffith, K.A., DeCastro, R., Stewart, et al., 2014. Gender differences in time spent on parenting and domestic responsibilities by high-achieving young physician-researchers. *Ann. Inter. Med.* [online] 160 (5), 344–353. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4131769/> [Accessed 24 Sep. 2019].
- Kelly, E.L., Moen, P., 2007. Rethinking the ClockWork of work: why schedule control may pay off at work and at home. *Adv. Devel. Hum. Resour.* [online] 9 (4), 487–506. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4295783/> [Accessed 1 Jan. 2020].
- Khanna, M., 2002. Mind the Gender Gap [online] American Planning Association. Available at: <https://www.planning.org/planning/2020/feb/mind-the-gender-gap/>.
- Lari, A., 2012. Telework/workforce flexibility to reduce congestion and environmental degradation? *Procedia - Soc. Behav. Sci.* 48, 712–721.
- Levine, N., Wachs, M., Shirazi, E., 1986. Crime at bus stops: a study of environmental factors. *J. Architect. Plan. Res.* [online] 3 (4), 339–361. Available at: <http://www.jstor.org/stable/4302882> [Accessed 25 Jul. 2020].
- Levinger, G., 1964. Note on need complementarity in marriage. *Psychol. Bull.* 61 (2), 153–157.
- Lois, J., 2010. The temporal emotion work of motherhood: homeschoolers' strategies for managing time shortage. *Gender Soc.* [online] 24 (4), 421–446. Available at: <http://www.jstor.org/stable/25741191> [Accessed 24 Apr. 2020].
- Loukaitou-Sideris, A., 1999. Hot spots of bus stop crime. *J. Am. Plan. Assoc.* 65 (4), 395–411.
- Loukaitou-Sideris, A., 2006. Is it safe to walk? 1 Neighborhood safety and security considerations and their effects on walking. *J. Plan. Lit.* 20 (3), 219–232.
- Loukaitou-Sideris, A., Liggett, R., Iseki, H., 2002. The geography of transit crime. *J. Plan. Educ. Res.* 22 (2), 135–151.
- Lucas, K., 2012. Transport and social exclusion: where are we now? *Transport Policy* 20, 105–113.
- Malach-Pines, A., 2005. The burnout measure short version. *Int. J. Stress Manage.* 12 (1), 78–88.
- McGuckin, N., Murakami, E., 1999. Examining trip-chaining behavior: comparison of travel by men and women. *Transport. Res. Rec.: J. Transport. Res. Board* 1693 (1), 79–85.
- McKenna, L., 2012. Explaining Annette Lareau, or, Why Parenting Style Ensures Inequality. [online] The Atlantic. Available at: <https://www.theatlantic.com/health/archive/2012/02/explaining-annette-lareau-or-why-parenting-style-ensures-inequality/253156/> [Accessed 24 Jul. 2020].
- McLafferty, S., Preston, V., 1991. Gender, race and commuting among service sector workers. *Professional Geographer* 43 (1), 1–15.
- McLanahan, S., Garfinkel, I., Casper, L., 1994. The Gender Poverty Gap: What Can We Learn From Other Countries? [online] RePEc - Econpapers. Available at: <https://EconPapers.repec.org/RePEc:il:iswps:112> [Accessed 24 Jul. 2020].
- McNally, M., Rindt, C., 2008. The Activity-Based Approach. [online] Semantic Scholar. Available at: <https://www.semanticscholar.org/paper/The-Activity-Based-Approach-McNally-Rindt/3bc7dd0bc9163e105717cfd4e9f9c2cca5bab14> [Accessed 24 Aug. 2020].
- Metro, 2019. Understanding how Women Travel. [online] Los Angeles Metro. Available at: http://libraryarchives.metro.net/DB_Attachments/2019-0294/UnderstandingHowWomenTravel_FullReport_FINAL.pdf [Accessed 4 Jun. 2020].
- Misra, R., Bhat, C., 2000. Activity-travel patterns of nonworkers in the San Francisco bay area: exploratory analysis. *Transport. Res. Rec.: J. Transport. Res. Board* 1718 (1), 43–51.
- Mitra-Sarkar, S., Partheeban, P., 2011. Abandon All Hope, Ye Who Enter Here: Understanding the Problem of "Eve Teasing" in Chennai, India. [online] trid.trb.org. Available at: <https://trid.trb.org/view/1101828> [Accessed 6 Jun. 2020].
- Moore, T.S., 2018. Why don't employees use family-friendly work practices? *Asia Pacific J. Human Resour.* 58 (1), 3–23.
- Novaco, R.W., Gonzalez, O.I., 2009. In: *Commuting and well-being. Technology and Psychological Well-being*. Cambridge University Press, pp. 174–205.
- OECD, 2017. Employment: Time spent in paid and unpaid work, by sex. [online] OECD.org. Available at: <https://stats.oecd.org/index.aspx?queryid=54757> [Accessed 25 May 2020].
- OECD Development Center, 2014. *Gender, Institutions and Development 2014*. OECD International Development Statistics.
- Pendall, R., Hayes, C., George, A., Mcdade, Z., Dawkins, C., Jeon, J., Knaap, E., Blumenberg, E., Pierce, G., Smart, M., 2014. Driving to Opportunity: Understanding the Links among Transportation Access, Residential Outcomes, and Economic Opportunity for Housing Voucher Recipients. [online] Available at: <https://www.urban.org/sites/default/files/publication/22461/413078-Driving-to-Opportunity-Understanding-the-Links-among-Transportation-Access-Residential-Outcomes-and-Economic-Opportunity-for-Housing-Voucher-Recipients.PDF> [Accessed 24 Oct. 2020].
- Piza, E., Kennedy, 2003. Transit stops, Robbery, and Routine Activities: Examining Street Robbery in the Newark, NJ Subway Environment. [online] Available at: <http://proceedings.esri.com/library/userconf/proc04/docs/pap1303.pdf> [Accessed 24 Oct. 2020].
- Radcliffe, S.A., 2006. Development and geography: gendered subjects in development processes and interventions. *Progr. Hum. Geogr.* 30 (4), 524–532.
- Ren, F., Kwan, M.-P., 2009. The impact of geographic context on e-shopping behavior. *Environ. Plan. B: Plan. Design* 36 (2), 262–278.

- Ridgeway, C.L., Correll, S.J., 2004. Unpacking the gender system. *Gender Society* 18 (4), 510–531.
- Roberts, J., Hodgson, R., Dolan, P., 2011. "It's driving her mad": Gender differences in the effects of commuting on psychological health. *J. Health Econ.* [online] 30 (5), 1064–1076. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0167629611000853> [Accessed 8 Jun. 2020].
- Rosenbloom, S., 2006. Understanding Women's and Men's Travel Patterns: The Research Challenge. [online] *trid.trb.org*. Available at: <https://trid.trb.org/view/783304> [Accessed 24 Jun. 2020].
- Rosenbloom, S., Burns, E., 1994. Why working women drive alone: implications for travel reduction programs. *Transport. Res. Rec.* 1459, 35–41, [Accessed 25 Jul. 2020].
- Ross, B., Bateman, N., 2019. Meet the Low Wage Workforce [online] *Brookings Institute*. Available at: https://www.brookings.edu/wp-content/uploads/2019/11/201911_Brookings-Metro_low-wage-workforce_Ross-Bateman.pdf [Accessed 15 Jun. 2020].
- Rothschild, J., 1983. *Machina ex dea* feminist perspectives on technology. Kronberg-Taunus Pergamon Press, New York Oxford Toronto Sydney Paris Frankfurt [Main I.E.].
- Ryan, R.M., Deci, E.L., 2006. Self-regulation and the problem of human autonomy: does psychology need choice, self-determination, and will? *J. Personal.* 74 (6), 1557–1586. Available at: <http://doi.wiley.com/10.1111/j.1467-6494.2006.00420.x> [Accessed 28 Feb. 2019].
- Santana, P., 2002. Poverty, social exclusion and health in Portugal. *Soc. Sci. Med.* 55(1), 33–45.
- Sen, A., 2000. Social Exclusion: Concept, Application, and Scrutiny. Available at: <https://www.adb.org/sites/default/files/publication/29778/social-exclusion.pdf> [Accessed 15 Jun. 2019].
- Shelton, B.A., John, D., 1996. The division of household labor. *Annu. Rev. Sociol.* [online] 22 (1), 299–322. Available at: <http://www.jstor.org/stable/2083433> [Accessed 24 Aug. 2020].
- Smart, M.J., Klein, N.J., 2018. Disentangling the role of cars and transit in employment and labor earnings. *Transportation* 47 (3), 1275–1309.
- Social Exclusion Unit, 2003. Making the Connections: Final Report on Transport and Social Exclusion. [online] Available at: https://www.ilo.org/wcmsp5/groups/public/—ed_emp/—emp_policy/—invest/documents/publication/wcms_asist_8210.pdf [Accessed 24 Jul. 2020].
- Srinivasan, S., 2000. Linking Land Use and Transportation: Measuring the Impact of Neighborhood-scale Spatial Patterns on Travel Behavior. [online] Available at: http://web.mit.edu/uis/theses/surmeeta_phd/diss1.pdf [Accessed 24 Jun. 2020].
- Stevenson, B., Wolfers, J., 2009. The Paradox of Declining Female Happiness. [online] *www.nber.org*. Available at: <https://www.nber.org/papers/w14969> [Accessed 24 Mar. 2020].
- Strathman, J.G., Dueker, K.J., Davis, J.S., 1994. Effects of household structure and selected travel characteristics on trip chaining. *Transportation* 21 (1), 23–45.
- Susilo, Y.O., Williams, K., Lindsay, M., Dair, C., 2012. The influence of individuals' environmental attitudes and urban design features on their travel patterns in sustainable neighborhoods in the UK. *Transp. Res. D Transp. Environ.* 109 (17), 190–200. Available at: https://www.sciencedirect.com/science/article/pii/S1361920911001544?casa_token=7u2FPdMFPz8AAAAA:3CFXAPqr8aRiOwe3ah7Z2M5LvupDeFn8Vg0DD1-gM_1TL-CL2e1bEUWkPczHMVdC2jQf7EQ-oQ [Accessed 19 Nov. 2020].
- Torquati, L., Mielke, G.I., Brown, W.J., Burton, N.W., Kolbe-Alexander, T.L., 2019. Shift work and poor mental health: a meta-analysis of longitudinal studies. *Am. J. Public Health* 109 (11), e13–e20.
- Tossi, M., 2002. A Century of Change: the U.S. Labor Force, 1950–2050, *Monthly Labor Review* [online] pp.15–28. Available at: <https://www.bls.gov/opub/mlr/2002/05/art2full.pdf> [Accessed 17 Apr. 2020].
- Townsend, P., 1987. Deprivation. *J. Social Policy* [online] 16 (2), 125–146. Available at: <https://www.cambridge.org/core/journals/journal-of-social-policy/article/deprivation/071B5D2C0917B508551AC72D941D6054> [Accessed 4 Apr. 2020].
- Uteng, M.T.P., Cresswell, P.T., 2012. *Gendered Mobilities*. Google Books. Ashgate Publishing, Ltd.
- Wachs, M., 1998. *The Automobile and The Automobile and Gender: Gender: An Historical Perspective An Historical Perspective*. [online] Washington, DC United States: Federal Highway Administration. Available at: <http://www.fhwa.dot.gov/ohim/womens/chap6.pdf>.
- West, C., Zimmerman, D.H., 1987. Doing gender. *Gender Soc.* 1 (2), 125–151. Available at: <http://www.jstor.org/stable/189945>.
- Williams, J.C., 2004. The Maternal Wall [online] *Harvard Business Review*. Available at: <https://hbr.org/2004/10/the-maternal-wall>.
- Williams, J.C., Bornstein, J., 2006. "Opt Out" or Pushed Out?: How the Press Covers Work/Family Conflict The Untold Story of Why Women Leave the Workforce. [online] Available at: <https://pdfs.semanticscholar.org/3449/76804a825eadb1817ed14d0b9bc9c10008e4.pdf>.
- Woods, W., Eagly, A.H., 2015. Two traditions of research on gender identity. *Sex Roles* 73 (11–12), 461–473.

Further Reading

- Blumenberg, E., Smart, M., 2010. Getting by with a little help from my friends . . . and family: immigrants and carpooling. *Transportation* 37 (3), 429–446.
- Blumenberg, E., 2016. Why low-income women in the US still need automobiles. *Town Plan. Rev.* 87 (5), 525–545.
- Federal Transit Administration, 2015. The National Transit Database (NTD). (Washington, DC). Available online at <https://www.transit.dot.gov/ntd> (Accessed February 1, 2018).
- Redelmeier, D.A., May, S.C., Thiruchelvam, D., Barrett, J.F., 2014. Pregnancy and the risk of a traffic crash. *Can. Med. Assoc. J.* 186 (10), 742–750.

Modeling and Simulation for Transport Planning

Michela Le Pira*, **Giuseppe Inturri†**, **Matteo Ignaccolo***, *Department of Civil Engineering and Architecture, University of Catania, Catania, Italy; †Department of Electric, Electronic and Computer Engineering, University of Catania, Catania, Italy

© 2021 Elsevier Ltd. All rights reserved.

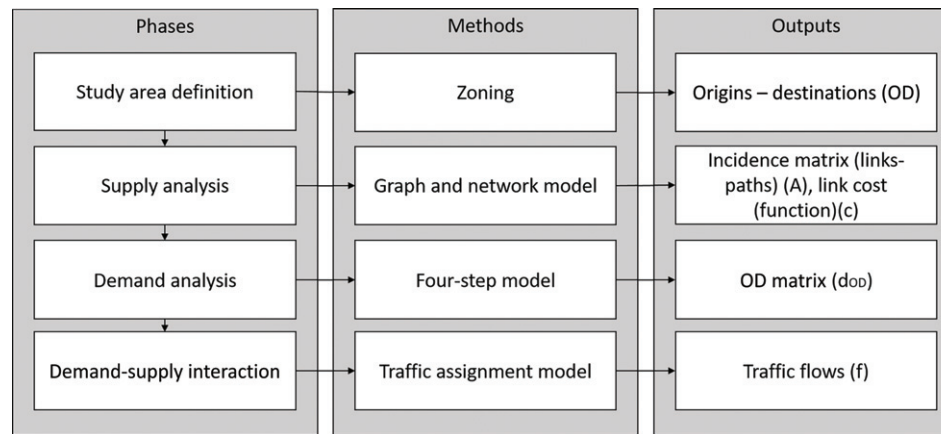
Introduction to Transport Planning	184
Models and Simulation for Transport Planning	185
Limits and New Trends of Modeling and Simulation for Transport Planning	186
Land Use Transport Interaction Models	187
Activity-Based Travel Demand Models	187
Big Data for Multimodal Transport Modeling	187
Freight Transport Models	188
Participatory Decision-Support Models	188
Conclusion	188
Acknowledgment	188
Biography	189
References	189
Further Reading	190

Introduction to Transport Planning

Transport systems are inherently complex because they affect the social, economic, and environmental dimensions of a territorial community, with several impacts and feedbacks not easy to be foreseen (see, e.g., the Braess paradox, [Steinberg and Zangwill, 1983](#)). Further complexity is added by the procedures needed for the construction and operation of transport systems, and mostly for the many actors involved with often conflicting interests ([Le Pira, 2018](#)). Public Administrations, at the different territorial levels (national, regional, and local) are responsible of decisions regarding the transport system. This is usually done through a set of documents that constitute a Transport Plan. It has the main objective to address the management of a sequence of decisions and actions about infrastructures, operation, and regulations of the overall transport system and their impacts on the community and the environment. Transport planning is a decision-making process based on rationality, aimed at defining and implementing transport system operations ([Ortúzar and Willumsen, 2011](#)). It is meant to effectively achieve aims and objectives as a result of a technical and political process, through a set of decisions that will inevitably favor some interests and expectations at the expense of others. It is also easy to understand why transport planning is not just about writing a plan with long-term decisions, but it is a rather dynamic process, for example, a sequence of elaborations (plans or individual projects) aimed at different decisions made in different moments.

Mathematical models play a crucial role, through their capability to simulate the transport system impacts and performances under different scenarios and, in the end, to assist the decision-maker during the planning process. Models must be considered just as tools to support the planning process, but they cannot substitute it ([Ortúzar and Willumsen, 2011](#)). The traditional planning process can be subdivided in macro-activities that can be categorized in technical and decision activities on one side, and analytical and modeling parallel phases (functional to the previous ones) on the other. Starting from the definition of the objectives and constraints, the analysis of the transport demand and supply and the traffic counts, in the present situation, are used to set up and calibrate a simulation model of the transport system. The model is then used to simulate users' behavior and vehicle flows on the different components of the transport network and predict the impacts of alternative plans and projects, under different scenarios. The outcome of the models forms the basis for decision-making. By comparing the results obtained from the model and from technical assessments (evaluation phase), a final choice is made, and the process must be followed up by an adequate monitoring of the plan.

Transport planning is losing its simple function of a rational choice among technical alternatives and is more and more being included within the policy process. In this respect, even if a transport plan is meant to increase the net welfare of a community, the benefits will never be equally distributed among its different actors and groups. Transport planning is therefore mostly about managing the public decision-making process. Besides, advances in new information and communication technologies (ICT) are rapidly changing the way transport systems and mobility are conceived, especially at the urban level. The concept of "smart city" becomes fundamental in a modern society: the "smart city of the future," that is, "a city in which ICT is merged with traditional infrastructures, coordinated and integrated using new digital technologies" ([Batty et al., 2012](#)), should be characterized by services which allow to optimize trips and to facilitate everyday activities (see e.g., the debate around "Mobility as a Service"), thus efficient transport systems are fundamental. This should be coupled with the goal of "sustainable mobility", which has been hindered by car dependence and the increased decentralization of cities, and which requires actions to reduce the need to travel and trip lengths, to



$$f = A F = A P (C) d = A P (A^T c) d$$

Figure 1 Framework for mobility analysis.

encourage modal shift and greater efficiency in the transport system (Banister, 2008). In this respect, the emergence of Sustainable Urban Mobility Plans (SUMP) has created a new direction for planning in Europe.

All these issues pose several challenges to policy-making and planning and point to the need for models capable of aiding decision-making. In this respect, research is abundant. However, the gap between model sophistication and user-friendliness is still too wide, especially considering the importance of involving non-expert citizens and stakeholders in increasingly participatory transport planning processes.

This chapter aims to provide an overview of the main approaches to modeling and simulation for transport planning and underline future research and practice-ready directions.

Models and Simulation for Transport Planning

Models and simulation tools have been developed to forecast future traffic conditions on transport networks, compare scenarios of alternative transport plans or projects, and estimate the impact of traffic or travel demand management strategies. Transport systems have been traditionally studied as the result of interaction between two main components, that is, the demand (users' trip flows – passengers or freight – per time unit) and the supply (infrastructure networks and related costs). The fundamentals of transport modeling were developed in the United States during the 1950s and later on imported into the United Kingdom in the early 1960s, following an evolutionary approach rather than revolutionary, as clearly explained in the “Handbook of Transport Modelling” by Hensher and Button (2008)^a.

Demand is usually represented via an OD (origin-destination) matrix, reproducing all the trips made in a given study area, appropriately divided into a discrete number of (internal/external) traffic zones. Supply is modeled as a mathematical representation of a network composed of nodes and weighted links, according to their fixed or variable (flow-dependent) costs, leading to respectively uncongested or congested networks.

Behavioral models are typically used to model transport demand, while graph theory is used to reproduce supply. The four-step model (FSM) is widely used to study transport demand and its interaction with supply, especially at the urban level. FSM is based on four sequential sub models: (1) trip generation, (2) trip distribution, (3) mode choice, and (4) route choice. Each one reproduces a traveler different decision, that is to perform a trip from a given origin, to which destination, by which transport mode, and route, all to choose from a set of alternatives available to him/her. Since the first two usually reproduce long term decisions linked to land use patterns (e.g., home-to-work trips), they are modeled via statistical descriptive models. On the contrary, mode and route choices are highly dependent on individual behaviors that can vary according to the state of the system. This is why behavioral models are used to reproduce these two choices. The economic analysis of discrete individual choices makes use of random utility maximization models (Block and Marschak, 1959). Discrete choice models (DCM) are basically econometric models aimed at analysing the behavior of a decision-maker when choosing among different (discrete) alternatives, assuming a rational decision-maker that tends to maximize his/her own utility. Varying model specifications leads to different models. The Multinomial Logit Model is widely used in transport planning to estimate travel demand, given a set of transport alternatives (i.e., mode choice or route choice) defined by certain attributes (Ben-Akiva and Lerman, 1985; Cascetta, 2009).

^aThe reader is encouraged to refer to this Handbook and to Ortúzar and Willumsen (2011) for a comprehensive overview of fundamental concepts of transport modelling that, for space reason and diversity of scope, cannot be covered by the present contribution.

Once the demand is known, the analysis of its interaction with the supply allows to study the performance of the transport system and the potential impacts of new scenarios. A simple scheme reproducing the different phases, methods and outputs to study the transport system of a given area is reported in **Figure 1**. Starting from a network and an OD demand, which allows to build an incidence matrix (A) accounting for the links pertaining to the alternative paths, flows in network links (f) are determined by path flows (F), which depend on the results of the route choice model, where the probability (P) of choosing a route is affected by its cost C .

Given a known travel demand distribution and transport network, traffic assignment models are used to estimate the flows on the different elements of the network and thus calculate performance indicators (times, costs, impacts on environment, etc.) consistent with the planning objectives. Most of the modeling efforts have been devoted to such tools. Traditional assignment models are homogeneous, considering only one type of user (e.g., road vehicles) and one type of network (e.g., road network). They assume that all travelers take the minimum cost route among the one available per each origin/destination pair. At same time, the cost of each route changes as a consequence of the increasing flow, until a so called user-equilibrium is reached ([Wardrop, 1952](#)), when no one can reduce his/her cost by unilaterally choosing another route. Many mathematical formulations and algorithms have been proposed and used to solve the user equilibrium problem ([Sheffi, 1985](#); [Florian and Hearn, 2003](#); [Gentile and Noekel, 2009](#)).

The vast majority of strategic transport models still uses traditional static traffic assignment (STA) procedures with travel time functions in which traffic flow can exceed capacity, delays are predicted in the wrong locations, and intersections are not properly handled ([Bliemer et al., 2013](#)). Dealing with time-varying flows both on the supply side (network costs) and demand side (OD trips' rate) is the challenge researchers face when dealing with Dynamic Traffic Assignment (DTA) in network modeling. The aim is to determine how traffic flow patterns (on links and paths) evolve over time and space in the network ([Peeta and Ziliaskoupolos, 2001](#)). The main difference with STA is that DTA models require path delay operators (to describe flow propagation in the network and queuing phenomena at intersections), rather than the monotonically increasing-with-flow path-delay function used with static assignment. This is to take into account that traveling along a path requires a given time, during which traffic conditions vary in such a way that is not predictable before departure ([Hensher and Button, 2008](#)).

Many software packages are today available to support analyses based on traffic simulations^b. They are traditionally divided into microscopic, mesoscopic and macroscopic traffic simulation, according to the level of description of vehicles and traffic flows. While attention has been traditionally paid to developing macroscopic models able to reproduce the transport system of large areas (cities, regions, and states), in the last years microscopic models have emerged as fundamental tools to increase the realism and the level of details of simulations. They are based on the description of the cinematic of each vehicle within a traffic flow, where acceleration, deceleration, speed and lane changing depend on the speed and position of the other vehicles. They also allow modeling of the behavior of individual heterogeneous components of the transport system, for example, pedestrian flows ([Helbing and Molnar, 1995](#)). More recently, macro and micro approaches have been integrated, for example, in large-scale microsimulation models like MATSIM ([Balmer et al., 2008](#)).

Limits and New Trends of Modeling and Simulation for Transport Planning

Though FSM has been used in professional practice by transport planners in the last decades, still a set of limits has to be outlined:

- *Independence assumptions.* While models are theoretically sound when considered separately, the independence assumption of the four submodels is an oversimplification required more for an easy tractability than for logical consistency.
- *The “predict and provide” paradigm.* The spatial resolution, based on the traffic analysis zone (TAZ) approach, tends to facilitate the ease of movement of motorized vehicles. Obtaining a high level of service for a road network translates into increasing the average speed of vehicles and lowering traffic intensity; this automatically determines the request for new infrastructures to increase road capacity, which consequently translates into further dependence on the use of the car. In this respect, the FSM responds to the so called “predict and provide” approach.
- *Transport supply management scenarios.* Asking for more capacity fulfils the presumptions of the dominant car culture, accepted by policy-makers and the public at large. Cost-effective alternatives are rarely proposed, and a transport supply management approach is preferred to satisfy trips generated by growth and land use; vice versa, transport demand management measures deployed to induce a change in travel behavior are less frequently considered ([Lewis, 1998](#)).
- *Multimodal network models.* A car-oriented culture and planning can lead to an inaccurate modeling of multimodal networks. Models usually account for separate OD matrices for private and public transport trips, and are hardly able to reproduce their interaction in unique simulation framework. Besides, even if research is abundant on sophisticated modeling approached to transit assignment ([De Cea and Fernández, 2007](#)), simulations used in the common practice are usually simplified and based on scheduled timetables, and are not able to readjust demand assignment taking into account real-time network conditions.
- *Trip-based travel demand modeling.* Given that different trip purposes are modeled separately, the scheduling and spatio-temporal inter-relationships between all trips and activities encompassing the individual's activity-travel pattern are not considered ([Bhat and Koppelman, 1999](#)).
-

^bThe reader can refer to the book of [Barceló \(2010\)](#) for a detailed description of most of the cited traffic simulation softwares.

Land use transport interaction. Finally, the FSM fails to consider the complex two-way dynamic link between land-use choices and the transport system (Kii et al., 2016)

It is thus important to fill the gap between modeling efforts and planning. In this respect, the role of transport modeling for urban planning has been changing from focusing on transport as a single issue to a more holistic view of mobility in relation to land use organization, environment and society as a whole, and from a relatively simple institutional context to a complex one with multiple participating stakeholders with often conflicting goals (Bertolini et al., 2008). Besides, traditional transport planning has been more focused on the passenger side, neglecting the important part played by freight transport and the potential of freight models, especially at the urban level (Tavasszy and de Jong, 2014; Gatta et al., 2017).

In the following, some of these new trends that have the potential to overcome these limitations will be briefly described.

Land Use Transport Interaction Models

Over the past 60 years, a considerable amount of research has led to operational Land Use and Transport Interaction (LUTI) models as decision support systems for assessing the impacts of land-use decisions on transport and vice versa, as well as evaluating large-scale transport investments. LUTI models integrate a Land-Use sub model and a Transport System sub model. The main aim is to integrate and predict households' residential and job location choice, the associated daily activity-travel patterns as well as transport mode and route choice (Acheampong and Silva, 2015). From the early LUTI aggregate spatial interaction-based models, drawing on the gravity analogy (Lowry, 1964), to aggregate utility-based and micro-simulation LUTI models (Kii et al., 2016), the need to capture the complex individual behavior dynamics and the relevant high resolution of spatial analysis has led to the adoption of agent-based modeling (ABM) for their capability to capture emergent phenomena (Le Pira et al., 2015). While the use of behavioral rules is similar to other disaggregate simulation techniques, ABM approaches allow agents (e.g., household members, travelers) to learn, modify, and improve their interactions with their environment.

While a FSM aims at high level of service for traffic flows and is car-oriented, LUTI models aim at increasing spatial and individual accessibility, that is the ease to reach a destination (active accessibility), and the ease for a destination to be reached (passive accessibility) (Geurs et al., 2010).

Relying on LUTI models able to predict accessibility as a proxy indicator of sustainability allows to change the transport planning paradigm from the "predict and provide" of FSMs to the "predict and prevent" approach. Planning moves from a simple reaction to provide greater infrastructural capacity to the expected growth in travel demand, to the prevention of unwanted results through proactive and prescriptive policies on the management of transport demand and land use. The natural consequence is that more and more researchers and practitioners now focus on a more balanced view of mobility and accessibility, completely encompassing the sustainable mobility paradigm (Banister, 2008).

Activity-Based Travel Demand Models

The activity-based approach recognizes travel as a derived demand from the need to pursue activities distributed in space and time. The key areas of investigation of this approach, therefore, include the demand for activity participation, the spatio-temporal constraints within which activity-travel behavior occurs, the complex interpersonal dynamics resulting from the interaction among household members and social networks, and activity scheduling and trip-chaining behavior in time and space (Bhat and Koppelman, 1999).

While trip-based models deal with aggregated units of analysis (households within a traffic zone), activity-based models are disaggregate, using a utility-econometric approach or microsimulation, for example, SimMobility (Alho et al., 2017). Research on LUTI models follows two lines: one is more flexible and complex, and involves intensive modeling at the local scale; the other involves simplified modeling at regional, national and global scales. The co-evolution of these two approaches is expected to contribute to sustainability science (Kii et al., 2016). There are a number of constraints imposed by microsimulation models. In addition to the large input data requirements, such models are slow to execute and require several running times, outputs between runs are also subject to significant stochastic variation and uncertainty (Acheampong and Silva, 2015).

Big Data for Multimodal Transport Modeling

With respect to the last point, significant opportunities come today by the widespread deployment of ICT, sensors and mobile devices (e.g., GPS, smartphones). The chance to record and collect data about traveling behaviors of individuals, position, route choice and speed of vehicles has opened a new field of research and professional practice falling under the topic "Big data and transport". In this respect, the known methods for updating an historic OD matrix by traffic detector counts and route choice behavior assumptions can benefit today of large and ubiquitous availability of Floating Car Data (Vogt et al., 2019). The explosion of "Big Data" availability together with the development of artificial intelligence techniques (e.g., machine learning) on more and more powerful computers is even questioning the need of transport modeling for traffic predictions. Real-time travel information deriving from the demand side (e.g., passenger location and desired destination), from the supply side (vehicle position and scheduled routing), coupled with GIS-based web platforms and fast effective dispatcher algorithms, have given birth to the possibility of dynamic transport planning, leading to the Mobility as a Service (MaaS) paradigm. It is a new mobility model that enables the

integration of the currently fragmented transport modes (collective, individual, public, private, and shared) and mobility services a traveler needs to conduct a trip (planning, booking, access to real time information, payment and ticketing). König et al. (2016) defined it as a multimodal and sustainable mobility service addressing customers' transport needs by integrating planning and payment on a one-stop-shop principle. Public transport should act as the backbone of this new paradigm, thus a correct planning based on multimodal network models is fundamental.

Freight Transport Models

Freight transport is an important aspect in transport planning. In this respect, increasing population, falling trade barriers, and declining transport costs fuelled by increasing consumption levels and the customisation of products and services have led to a continuous increase of freight flows (Tavasszy and de Jong, 2014). Only in recent years it became a public policy problem, mainly due to its impact on the environment and the community as a whole. Freight models have been developed in parallel with passenger models, even if at a slower pace. However, their inclusion in current transport planning practice is still lacking, due to the quite recent awareness of freight negative impacts on sustainability. Freight models can be categorized according to the levels of decisions, that is strategic, tactical and operational, which relate to three different markets, that is commodity, inventory logistics services, and transport logistics services, with direct and mutual dependence between decisions at the different layers (Tavasszy and de Jong, 2014). However, it is difficult to combine all the levels in a comprehensive model^c. Some recent attempts go in this direction, for example the MASS-GT model (de Bok and Tavasszy, 2018) or SimMobility Freight (Alho et al., 2017).

Understanding actor behavior and the consequences of their actions is fundamental in freight models. Especially at the urban level, urban freight transport poses several challenges to transport policy-making, being characterized by both public and private interests and trade-off between freight delivery efficiency and city sustainability (Gatta et al., 2017). Based on this premise, freight modeling tends to be more comprehensive and empirically sound, while also taking into account the behavioral components of the different actors and their heterogeneous preferences (Marcucci et al., 2017).

Participatory Decision-Support Models

The gap between modeling effort and planning practice is still a problem to overcome, especially if one considers the involvement of non-experts in the planning process. Besides, while there are many methods for stakeholder involvement, a deeper analysis of the underlying social dynamics is still missing and a modeling approach can be used to help the participatory process (Le Pira, 2018). In this respect, even if public participation is considered important for the success of a decision-making process, in general it is not well-structured and there is a lack of group decision-support methods and models that can guide stakeholders' involvement towards well-thought-out decisions. Le Pira et al. (2017) provide an overview of participatory decision-support methods and models, according to the desired level of engagement, from stakeholder identification and analysis to stakeholder consultation and participation in the decision-making process. These include social network analysis, multi-criteria decision-making methods, Participatory GIS, DCM, and ABM. In particular, a combination of DCM and ABM has proved to be useful to forecast stakeholders' reaction to policy change, but also to predict the outcome of an interaction process aimed at consensus building (Marcucci et al., 2017; Le Pira et al., 2017). Still, these models are still very theoretical and need to be tested in practice to verify their usability and added value to decision-making.

Conclusion

This chapter addressed the topic of modeling and simulation for transport planning, presenting the state of the art and some future research directions. After an introduction to transport planning challenges and the rationale of using models and simulations as decision-support tools, the main modeling approaches to reproduce and study transport systems have been presented. Given the limitations of traditional approaches, and the gap between research and practice, recent trends and modeling efforts have been briefly described to underline future research and practice-ready directions. To sum up, when dealing with modeling and simulation for transport planning, some fundamental concepts should be taken into consideration: (1) land use and transport interaction, (2) activity-based approach, (3) big data for transport modeling, (4) freight transport models and (5) participatory decision-support models. There are still avenues for new research, but it should be necessarily supported by planning implementation to bridge the gap between theory and practice.

Acknowledgment

This paper has been partially supported by the project of M. Le Pira "AIM Linea di Attività 3 – Mobilità sostenibile: Trasporti" (unique project code CUP E66C18001380007) under the programme "PON Ricerca e Innovazione 2014-2020 – Fondo Sociale Europeo, Azione 1.2 "Attrazione e mobilità internazionale dei ricercatori"

^c For a comprehensive review of modeling approaches to freight transport the reader can refer to Tavasszy and de Jong (2014).

Biography



Michela Le Pira is currently a Research fellow/lecturer specialized in Transport Planning at the University of Catania (Italy). She holds a PhD in transport engineering. As a PhD student, she focused on public participation in transport planning supported by agent-based modeling and multi-criteria decision methods. As a Postdoc researcher at University of Roma Tre, she worked on participatory decision-support methods for sustainable urban freight transport policy-making. As a research fellow at the University of Catania and guest researcher at TU Delft (Netherlands), she is currently investigating new mobility services for both passengers and freight.



Giuseppe Inturri is Associate Professor in Transport at the Department of Electric, Electronic and Computer Engineering at the University of Catania. He is also Rector's Delegate for the promotion of sustainable mobility and Rector's Delegate for the Italian University Network for Sustainable Development. He has research interests in land use and transport planning, sustainable mobility, multi-criteria analysis, public participation in transport decisions, agent based simulation and transport modeling. He is author of more than 80 publications in international journals, book chapters and conference proceedings.



Matteo Ignaccolo is Full Professor of Transport Planning (scientific sector ICAR/05) at the University of Catania. He teaches several Transport courses for Transport, Civil and Environmental Engineers and he is a member of the committee of the Doctorate in "Evaluation and mitigation of urban and land risks". He is author of over 100 publications on topics like: urban transport systems, sustainable mobility, air transport systems, technical/economic feasibility of transport interventions, and agent-based modeling applied to transport planning. He is currently the National President of AIIT (Italian Association for Traffic and Transport Engineering) and the Editor-in-Chief of the scientific journal "European Transport / Trasporti Europei."

References

- Acheampong, R.A., Silva, E., 2015. Land use–transport interaction modeling: A review of the literature and future research directions. *J. Trans. Land Use* 8 (3).
- Alho, A., Bhavathrathan, B.K., Stinson, M., Gopalakrishnan, R., Le, D.T., Ben-Akiva, M., 2017. A multi-scale agent-based modelling framework for urban freight distribution. *Transp. Res. Proc.* 27, 188–196.
- Balmer, M., K. Meister, and K. Nagel. 2008. Agent-based simulation of travel demand: Structure and computational performance of MATSim-T. ETH, Eidgenössische Technische Hochschule Zürich, IVT Institut für Verkehrsplanung und Transportsysteme.
- Banister, D., 2008. The sustainable mobility paradigm. *Transp. Policy* 15, 73–80.
- Barceló, J., 2010. Fundamentals of traffic simulation, 45. Springer, New York, pp. 439.
- Batty, M., Axhausen, K., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzonis, G., Portugali, Y., 2012. Smart cities of the future. *Eur. Phys. J. Special Topics* 214, 481–518.
- Ben-Akiva, M.E., Lerman, S.R., 1985. Discrete choice analysis: Theory and application to travel demand, 9. MIT press, Cambridge, Massachusetts; London, England.
- Bertolini, L., le Clercq, F., Straatemeier, T., 2008. Urban transportation planning in transition (editorial). *Trans. Policy* 15 (2), 69–72.
- Bhat, C.R., Koppelman, F.S., 1999. Activity-based modeling of travel demand. In: Hall, R. (Ed.), *Handbook of Transportation Science*, 23. Springer, US, pp. 35–61.

- Bliemer, M., Raadsen, M., Romph, E.D., & Smits, E.S., 2013. Requirements for traffic assignment models for strategic transport planning: A critical assessment. Paper presented at: Proceedings of the 36th Australasian Transport Research Forum 2013, ATRF, Brisbane, Australia, 2-4 October, 2013.
- Block, H.D., & Marschak, J., 1959. Random orderings and stochastic theories of response (No. 66). Cowles Foundation for Research in Economics, Yale University.
- Cascetta, E., 2009. *Transportation System Analysis. Models and Applications*, 2nd ed. Springer, New York.
- de Bok, M., Tavasszy, L., 2018. An empirical agent-based simulation system for urban goods transport (MASS-GT). *Proc. Comp. Sci.* 130, 126–133.
- De Cea, J., Fernández, E., 2007. Frequency-based transit-assignment models'. *Handbook Trans. Model.* 1, 575–589.
- Florian, M., Hearn, D., 2003. Network equilibrium and pricing. *Handbook of Transportation Science*, Springer, Boston, MA, pp. 373–411.
- Gatta, V., Marcucci, E., Le Pira, M., 2017. Smart urban freight planning process: integrating desk, living lab and modelling approaches in decision-making. *Eur. Transp. Res. Rev.* 9, 32.
- Gentile, G., & Noekel, K., 2009. Linear User Cost Equilibrium: the new algorithm for traffic assignment in VISUM. In *Proceedings of European Transport Conference 2009*.
- Geurs, K., Zondag, B., De Jong, G., de Bok, M., 2010. Accessibility appraisal of land-use/transport policy strategies: More than just adding up travel-time savings. *Transp. Res. Part D Trans. Environ.* 15 (7), 382–393.
- Helbing, D., Molnar, P., 1995. Social force model for pedestrian dynamics. *Phys. Rev. E* 51 (5), 4282.
- Hensher, D.A., Button, K.J., 2008. *Handbook of transport modelling* (No. 1). Elsevier, Amsterdam.
- Kii, M., Nakanishi, H., Nakamura, K., Doi, K., 2016. Transportation and spatial development: An overview and a future direction. *Transp. Policy* 49, 148–158.
- König, D., Eckhardt, J., Apaaja, A., Sochor, J. & Karlsson, M., 2016. Deliverable 3: Business and operator models for MaaS. MAASiFIE project funded by CEDR. Available at: http://www.vtt.fi/sites/maasifie/PublishingImages/results/cedr_mobility_MAASiFIE_deliverable_3_revised_final.pdf.
- Le Pira, M., 2018. Transport planning with stakeholders: an agent-based modelling approach. *Int. J. Trans. Econ.* 45, 15–32.
- Le Pira, M., Inturri, G., Ignaccolo, M., Pluchino, A., Rapisarda, A., 2015. Simulating opinion dynamics on stakeholders' networks through agent-based modeling for collective transport decisions. *Proc. Comp. Sci.* 52, 884–889.
- Le Pira, M., Marcucci, E., Gatta, V., Ignaccolo, M., Inturri, G., Pluchino, A., 2017. Towards a decision-support procedure to foster stakeholder involvement and acceptability of urban freight transport policies. *Eur. Transp. Res. Rev.* 9, 54.
- Lewis, S.L., 1998. Land use and transportation: Envisioning regional sustainability. *Transp. Policy* 5 (3), 147–161.
- Lowry, I.S., 1964. *A Model of Metropolis*. Santa Monica, CA: RAND Corporation.
- Marcucci, E., Le Pira, M., Gatta, V., Ignaccolo, M., Inturri, G., Pluchino, A., 2017. Simulating participatory urban freight transport policy-making: Accounting for heterogeneous stakeholders' preferences and interaction effects. *Transp. Res. Part E* 103, 69–86.
- Ortúzar, J.D., Willumsen, L., 2011. *Modelling Transport*, 4th ed. Wiley, Chichester, West Sussex, United Kingdom.
- Peeta, S., Ziliaskopoulos, A.K., 2001. Foundations of dynamic traffic assignment: The past, the present and the future. *Netw. Spat. Econ.* 1 (3–4), 233–265.
- Steinberg, R., Zangwill, W.I., 1983. The prevalence of Braess' paradox. *Transp. Sci.* 17 (3), 301–318.
- Sheffi, Y., 1985. *Urban Transportation Networks*. Prentice-Hall, Englewood, NJ.
- Tavasszy, L., de Jong, G., 2014. *Modelling freight transport*. Elsevier, London, Waltham, USA.
- Vogt, S., Fourati, W., Schendzielorz, T., Friedrich, B., 2019. Estimation of origin-destination matrices by fusing detector data and floating car data. *Transp. Res. Proc.* 37, 473–480.
- Wardrop, J.G., 1952. Some theoretical aspects of road traffic research. *Proceedings of the Institute of Civil Engineers*, Part II, pp. 325–378.

Further Reading

- Axhausen, K.W., Gärling, T., 1992. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transp. Rev.* 12 (4), 323–341.
- Daganzo, C.F., Sheffi, Y., 1977. On stochastic models of traffic assignment. *Transp. Sci.* 11 (3), 253–274.
- Frederix, R., Viti, F., Tampère, C.M., 2013. Dynamic origin–destination estimation in congested networks: Theoretical findings and implications in practice. *Transportmet. A: Transp. Sci.* 9 (6), 494–513.
- Gentile, G., 2016. Solving a dynamic user equilibrium model based on splitting rates with gradient projection algorithms. *Transp. Res. Part B* 92, 120–147.
- Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transp. Res. Part B* 42 (9), 759–770.
- Mahmassani, H., 2005. *Transportation and Traffic Theory: Flow, Dynamics and Human Interaction* (Proceedings of the 16th International Symposium on Transportation and Traffic Theory). Emerald Publishing, Untied Kingdom.
- Train, K.E., 2009. *Discrete choice methods with simulation*. Cambridge university press, New York, USA.

Externalities and External Costs in Transport Planning

Silvio Nocera, Università IUAV di Venezia – Department of Architecture and Arts, Venice, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	191
Main Negative Externalities in Transport	191
Positive Externalities in Transport	193
Economic Evaluation of Externalities	193
Conclusions	195
References	195
Further Reading	197

Introduction

In their book, [Sinha and Labi \(2007\)](#) divide transport costs by the source of cost incurrence into: (1) *agency costs* (the ones incurred by service providers), (2) *user costs* (monetary and nonmonetary costs incurred by the consumers of the transport services, that is passengers, shippers, and truckers), and (3) *community costs*, that represent the costs incurred by the community as a whole, including entities not directly involved with the use of the facility. Any side effect of travel that generates a cost for one of these three categories may be defined as a secondary cost, or **externality**. It is important to underline that such costs are generally not borne by transport users and hence not taken into account when they make a transport decision.

[Black \(2010\)](#) states that the unsustainability margins of transport are unavoidable, and defines a sustainable transport system as “one that provides transport and mobility with renewable fuels while minimizing emissions detrimental to the local and global environment, and preventing needless fatalities, injuries, and congestion.” This definition is not exactly avant-garde, as the discussion about externalities started really with [Pigou \(1920\)](#), who developed the works by [Sidgwick \(1887\)](#) and [Marshall \(1890\)](#). It specifies, however, that externalities are inevitable within transport, and also that enlightened transport planning can only aim at their containment. Starting from the theory of [Coase \(1960\)](#), [Arrow \(1969\)](#) first explained how externalities could be internalized by the creation of additional markets. Concrete investigations have been also made in various terms of policy ([Attard and Ison, 2015](#); [Button, 1990](#); [de Palma and Lindsey, 2011](#); [Eliasson and Jonsson, 2011](#); [Eliasson, 2009](#); [Small, 1992](#)). Literature on externalities also focused on advanced modeling ([Arnott et al., 1993](#); [Hensher, 2018](#); [Hess and Scarpa, 2010](#)), modal use and concurrency ([Cavallaro et al., 2018a](#); [Mulley et al., 2017](#)), and climate change ([Cavallaro et al., 2018b](#); [Hensher, 2008](#); [Nocera and Cavallaro, 2012](#); [Nocera et al., 2018a, 2018b](#)).

As a last preliminary concept, the meaning and the possible use of the terms “externalities” and “external costs” should be cleared. External costs may be seen as the monetary valuation of externalities, as they represent the costs that are caused from mobility without being included in the price of mobility itself. For example, if someone decides to use his/her car, he/she pays for the car itself, the gasoline, the insurance, but generally neither for the air pollution, nor for the noise caused. Such costs of mobility must be borne by others, and are defined “external” for this reason.

Both external costs and externalities share some important properties, and for this reason they are often used as synonyms in literature ([Bigazzi and Figliozzi, 2013](#); [Bigazzi and Figliozzi, 2013](#); [Lindsey et al., 2019](#); [Lindsey et al., 2019](#)). First of all, they both refer to side effects of transport. Furthermore, and most importantly: as specified above already, they represent real costs not included in the market price of transport and not borne by the users. For this reason, the implementation of a sort of compensation—generally under the form of a tax or a subsidy—should be required for economic efficiency to be achieved. Since such compensation is often not included, externalities prevent the accomplishment of the perfect concurrency. This pushes away the system from economic efficiency, opening up the possibility for second-best solutions, which have thoroughly been discussed in mathematical terms by [Verhoef \(2000\)](#) and [Proost \(2011\)](#), and particularly in association with externalities by a large number of authors (among many others, [Button, 2006](#); [Diamond, 1973](#); [Tikoudis et al., 2018](#)).

Main Negative Externalities in Transport

The main negative externalities of transport concern the traffic accidents, air pollution, climate change, noise and congestion, and are described in [Table 1](#).

There is no universal agreement in the literature on what lies aside of the contents of [Table 1](#). According to [Sinha and Labi \(2007\)](#), [DeLucchi and McCubbin \(2011\)](#), [Friedrich and Quinet \(2011\)](#), other externalities may include visibility, agriculture, materials, forestry and soil, water, energy, and landscape.

Congestion is the impedance that vehicles impose on each other, due to speed–flow relationships, in conditions where the use of the transport system approaches its capacity ([European Conference of Ministers of Transport ECMT, 2000](#)). As the flow of vehicles

Table 1 Most commonly considered external costs in transport planning

	Source	Public abatement	Nature of external cost	Policy instrument used
Congestion	Too many users of the same facilities increase in-vehicle cost and schedule delay cost	Increase of capacity, use of ITS	Mainly time and schedule delay costs	Congestion pricing, gasoline taxes, regulation
Traffic accidents	More intensive or more mixed use influences the average probability and severity of accidents	Adaptation of road equipment, emergency services, etc.	Mainly health, loss of life, material damage	Traffic regulations, pricing to reduce demand
Air pollution	Exhaust of combustion engines (car, trucks, bus, airplane, power stations)		Mainly health, loss of life	Standards on car emissions and quality of fuels
Noise	Vehicle, airplanes, trains	Noise walls, more silent road surfaces, tire design	Discomfort, health	Standards on car, banning use of certain equipment at night Tradable permits for night flights
Climate change	Fossil fuel use by vehicles, diesel-fuel-powered trains, airplanes		Long-term disruption of climate (sea level, cooling costs, water supply etc.)	Standards on fuel consumption (minimum efficiency) CO ₂ taxes Tradable permits

Source: Modified from Proost (2011).

on a given road increases, from a certain threshold value each additional vehicle is not only operating at an ever higher private cost, but it also causes an increase in the cost to other vehicles already in circulation (Danielis, 2001). In this sense, congestion is an externality internal to the examined transport modality and not *primarily* affecting the community. However, there is an interdependence between congestion and other externalities. For example, it decreases the energy efficiency of engines and, consequently, it increases air pollution. At the same time, the severity of the accidents is reduced, but their quantity increases, thus posing (still open) issues about the relationship between congestion and accident risks (Ricardo-AEA, 2014). The Roadway Congestion Index (Schrank and Thomas, 2009) can be a valid tool to calculate it for specific areas.

External accident costs do not refer to the entire costs derived from a generic crash, but include only the part that is not covered by risk-oriented insurance premiums. Thus, the calculation of the external component is a function of the level of accidents, but also of the insurance system: a rough distinction includes fatal, injury and property-damage-only crashes (Sinha and Labi, 2007). According to Ricardo-AEA (2014), categories related to accident include “medical costs, production losses, material damages and administrative costs, and the risk value as a proxy to estimate pain, grief and suffering caused by traffic accidents in monetary values.” These can be calculated as the sum of the expected cost due to an accident for the persons exposed to risk, the accident cost for the rest of the society (output loss, material costs, police and medical costs) and the expected cost for the relatives and friends of the persons exposed to risk. When referring to external costs, the last category is mostly considered. The evaluation of accident costs can be based on contingent valuations, which assess the willingness to pay for safety by transport users.

Air pollution refers to the presence of toxic chemicals or compounds in the air, at levels that can be health threatening. Such compounds may be found either in a gaseous form or in a solid form (as particulate matter suspended in the air). *Criteria Air Pollutants* represent the main air pollutants. Black (2011) includes in this category carbon monoxide (CO), nitrogen oxide (NO_x), particulate matter of size 10 µm or less (PM₁₀), sulfur dioxide (SO₂), lead (Pb), and Ozone (O₃). Sinha and Labi (2007) present a slightly different list, inserting also particulate matter of size 2.5 µm or less (PM_{2.5}), volatile organic compounds (VOCs) and ammonia (NH₃). The quantification of their emission in the atmosphere is function of several factors, which include the characteristics of the travel (e.g., the speed), the facility (e.g., the type of infrastructure and its status of maintenance), the driver and his/her behaviors, the vehicle (e.g., size or maintenance) and the environment (including temperature and humidity). Typically, effects of air pollution can be evaluated by adopting the dose–response method: this method focuses on the quantification of the impact that the emissions have on human health, environment and economic activities, calculated for the population potentially affected by the phenomenon.

Noise is an externality harmful to human health in at least two different ways. Indeed, it may produce psychological consequences (resulting in annoyance, nervousness and behavioral disorders), or physiological ones, resulting in long-term impacts such as hypertension and myocardial infarction. Impacts can vary significantly, being functions of several factors such as duration, frequency and intensity of the noise. Specific indicators have been introduced, in order to make this comparable. Day-evening-night level (L_{den}) is a main descriptor of noise level based on energy equivalent noise level (L_{eq}) over a whole day with a penalty of 10 dB (A) for night time noise (23.00–7.00) and an additional penalty of 5 dB(A) for evening noise (European Environmental Agency, 2001). Noise can be evaluated by adopting the hedonic pricing method, but due to its numerous and well-known limitations (Arguea and Hsiao, 1993), this may be integrated with other evaluations, such as the avoidance cost or the contingent valuation method.

Table 2 Methods to assign an economic value to transport externalities

Externality	Methods to assign an economic value
Congestion	Travel time delay, Contingent valuation method
Accident	Contingent valuation method, Dose/response assessment, Hedonic prices
Air pollution	Dose/response assessment, Avoidance cost, Contingent valuation method, Hedonic prices
Noise	Contingent valuation method, Hedonic Pricing Method, Avoidance cost, Dose/response assessment
Climate change	Damage costs, Avoidance costs

Sources: Danielis, 2001; Transportation Research Board, 2001; Ricardo-AEA, 2014.

Climate change can be defined as a change in the pattern of weather, and related changes in oceans, land surfaces and ice sheets, occurring over time scales of decades or longer (Australian Academy of Science, 2019). According to recent estimations (EC, 2019), it constitutes about 15% of transport externalities at the EU level, but its share is growing. Greenhouse gases (GHGs) are responsible for the greenhouse effect, a phenomenon through which the atmosphere absorbs the infrared radiation emitted by the planet's surface as a result of exposure to solar radiation, thus increasing the mean temperature of the planet. GHGs include carbon dioxide (CO₂), methane (CH₄), nitrous oxides (N₂O), ozone (O₃), and fluorinated gases (F-gases). Their internalization in transport costs is object of much debate (Nocera and Cavallaro, 2014). Prices determined by carbon trading or by carbon tax cannot be considered representative: high volatility and price fluctuations make the former unsuitable for long-term forecasts (Ellerman et al., 2008), whereas the latter depends on political choices. Alternative techniques such as the *damage costs* and the *avoidance costs* methods are preferred (Clarkson and Deyes, 2002). *Damage costs* assesses the future physical impacts of climate change and link them with their effects on the economy and society. *Avoidance costs* quantifies the money required to avoid an increase of GHG levels, to reduce such emissions or to remove GHG from the atmosphere.

Each of the externality previously presented can be internalized in transport costs by the assignment of a fair unitary economic value—a point that will be more specifically covered in Section 4. **Table 2** summarizes the methods for this specific aim.

Positive Externalities in Transport

The question of whether transport may yield also important external benefits should have a clear positive answer, even if this concept can be amplified by some parties when it comes to a comparison with external costs: frequent air travelers may for instance claim that the external benefits of air transport largely exceed its external costs, thus casting doubt on the necessity to restrict airplane use.

This point will not be discussed into detail. It should be, however, beyond dispute that contemporary societies are largely influenced by transport, inducing added value and employment in several fields. Besides the negative externalities mentioned in the previous section, it should hence be added that the different positive aspects are indeed benefits for society (Verhoef, 1994).

As a general concept, positive externalities of transport networks may include the ability to provide emergency services, increases in land value due to improved accessibility, and general agglomeration benefits. The last point is explained brilliantly in de Palma et al. (2019). The creation of value for some public areas efficiently served by public transport can also be interpreted as a positive social consequence of the improvement of accessibility. It is not exactly undisputed, but a positive correlation between real-estate property values and transport is generally acknowledged in literature: Efthymiou and Antoniou (2013) show that proximity to transport infrastructure has a direct impact on house purchase prices and rents, which is either positive or negative depending on the type of the transport system. Landis et al. (1995), Benjamin and Sirmins (1996), and Welch et al. (2016) also highlight a potential value of efficient transport systems, while negative effects on real-estate value have been reported by Chen et al. (1998) and Gadziński and Radzinski (2016).

Other positive externalities include the so-called “Mohring effect” (Mohring, 1972): it is the observation that urban public transport exhibits considerable economies of scale if users' waiting time is included in the cost function. Because waiting time forms part of the costs of transport, the Mohring effect implies increasing returns to scale for scheduled urban transport services and brings with it the implication that without subsidization, frequencies will be lower than socially optimal.

The quantification of those positive impacts is not straightforward, as the valuation techniques employed may often yield biased estimates (see section 4 for details). Legaspi et al. (2015) developed a practical framework for estimating the wider economic benefits generated from transport investments, while Ignaccolo et al. (2016) examined the potential impact of land use and transport policies on external costs through the evaluation of a simple ideal indicator. They report that high-density urban areas, close destinations and accessible with a broad spectrum of opportunities for low-impact transport (public transport, cycling, walking), reduce environmental impacts and transport externalities.

Economic Evaluation of Externalities

There is a keen interest in determining the overall costs and benefits of transport, hence including also externalities in the context. While it is taken for granted that transport has considerable repercussions of environmental and social nature, opinions strongly

diverge when it comes to the conceptual definition of external benefit and cost, to their identification in the case of transport, and especially to their monetary quantification.

A first difficulty derives from the fact that there is a univocal unit of measurement only for some impacts (land consumption, number of people involved in accidents, noise, vibrations, and so on). Other impacts may have no aggregate unit of measurement (air, water, and soil pollution), or strongly rely upon subjective perceptions (visual impacts).

Furthermore, there are time differences between the manifestation of damage and its perception. Damages can be immediate, short-term (noise, visual impacts, bad smells, dust), or medium–long term (respiratory diseases and genetic alterations linked to air and water pollution). The farther the effect is, the greater the difficulty of estimating its economic value.

The spatial dimension of the environmental impacts of transport is also very diversified. A significant part of the impacts is concentrated in the surroundings of the area in which the transport activity takes place and is located precisely (noise, visual impacts, vibrations, bad smells, and the most pollutants), while other impacts cover more ground (smog, acid rain, tropospheric ozone) or are even global (ozone layer reduction, greenhouse effects). The uncertainties connected to the complex and sometimes imprecise temporal and spatial definitions of the impacts also add to the scientific uncertainties on the definition of the relationship between exposure to pollution and effects on health, on human psychology and on ecological balance.

Having taken note of the above, literature shows that the main methods for estimating the environmental value are the following ones:

1. **Dose–response method** requires information on the effect that a change in a particular chemical or pollutant has on the level of the consumer's utility (Lu et al., 2019; Pearce and Turner, 1990);
2. **Contingent valuation** elicits information on willingness to pay or willingness to accept compensation for an increase or decrease in some usually nonmarketed goods or services (Bravo-Moncayo et al., 2017; Pethig, 1994);
3. **The Hedonic price method** is based on the underlying assumption that the price of a property is related to the stream of benefits derived from it. The method relies on the hypothesis that the prices which individuals pay for commodities reflect both environmental and nonenvironmental characteristics. The implicit prices are sometimes referred to as hedonic prices, which relate to the environmental attributes of the property (Braden and Kolstad, 1991; Seo et al., 2014);
4. **The Travel-Cost method** is a widely used surrogate market approach that relies on information on time and travel costs to derive a demand curve for a recreational site. This curve is in turn used to estimate the consumers' surplus or value of the site to all users (Hanauer and Reid, 2017; Zhang et al., 2015);
5. **The Avoidance Cost approach** (also known as mitigation or control cost) quantifies the money required to avoid a certain environmental damage (Hanley and Splash, 1993; Harper et al., 2016);
6. **The Damage Cost method** assesses the future physical impacts of a certain external effect and links them with their effects on the economy and society (Maibach et al., 2008; Massiani, 2015).

It is not possible to prove that a method is in all cases better than all the others, but some criteria (theoretical reliability, repeatability of the results, comparability with other estimates) can be established to evaluate its constancy. The choice of the method will then mainly depend on the type of externality to be evaluated, on the available data, on the level of information of the user, on the financial resources and on time availability.

Given the above and in the light of the context expressed earlier in Section 2, the so-called intersectoral externalities (i.e., those regarding only the producers or the users of a certain transport service) are mainly evaluated through market prices (even if some not univocally), often using the damage cost method: gains in excess of fuel and lubricants due to congestion to be paid by car drivers belong to this category, as well as nonrefunded acts of vandalism on transit vehicles, and losses in time for passengers. This latter point is a very complicated issue: over the past five decades numerous studies have been carried out, providing substantial insights on its impact, generally with the aim of ensuring consistency and good practice in project appraisal (Abrantes and Wardman, 2011; Carrion and Levinson, 2012; Peer et al., 2014).

Environmental externalities (i.e., those borne from the community) can be divided into two further categories. In the first one, it is possible to obtain with relative ease the duration and the costs of the interventions necessary to eliminate the negative effects. The economic estimation of these externalities is strongly linked to market prices and can hence be obtained almost univocally through avoidance costs (even if sometimes, damage costs can be also used). For example, the traffic noise on a railway line belongs to this category, if it is possible to eliminate it with opportune mitigation works, that is noise barriers. Avoidance cost method estimates potential expenditures on damage avoidance or protection (Phaneuf and Requate, 2016).

A different situation occurs where the economic estimation cannot be obtained univocally, but needs models or speculations that incorporate a huge degree of uncertainty and/or may properly leave substantial room to the beliefs of the analysts. The estimations of traffic incidents to vehicles made by insurance companies belongs to this category, as well as the estimations of future climate damages. In this case, the estimation is made upon market prices, but incorporates large subjective beliefs as well, so that a fair estimation of the externality cannot be obtained univocally. The use of indicators is recommended in this case, even if the damage cost method can be used in some borderline situations.

A last category refers to the externalities not of pure consumption but referred to the ecology, the landscape and generally to the aesthetic-cultural components. These are defined as “cultural externalities” and are generally only captured by that component of the community, which relies on its cultural and skill profile. Based on all these considerations, externalities may be clustered under the scheme shown in Table 3.

Table 3 A basic taxonomy for the externalities in transport

	<i>Actors involved</i>	<i>Description</i>	<i>Methods mostly used for the internalization</i>	<i>Examples</i>
Intersectoral externalities	Producers or the users of a certain transport service	Their value refers generally directly to market transactions	Damage costs	Congestion, noise, accidents
Environmental externalities	Community	It is possible to estimate in monetary terms the costs necessary to eliminate the effects caused by the externalities The estimation cannot be obtained straightforwardly but needs models or speculations that incorporate a huge degree of uncertainty and/or may properly leave substantial room to the beliefs of the analysts	Avoidance costs. Sometimes, damage costs Indicators. Damage costs	Noise Accidents, Climate change
Cultural externalities	Community	They cannot be quantified in monetary terms, but only through ordinal scales		

It should be finally noted that most of the taxonomies referred to externalities may be tenuous, as the same impact can concern more than one category previously defined. For instance, the impact caused by a vehicle into action on an urban congested road in a historical center causes at the same time: (1) increased consumption of fuel and lubricants, whose external costs may be easily calculated through market transactions; (2) some externalities (i.e., noise) for which the external costs may be calculated by planning and computing the monetary value of the works required for impact mitigation—even though this is not always possible; (3) emission of primary pollutants that can impact on the health of the community, and whose impact can be calculated only through parameters and models; (4) carbon and methane pollution, that impact on climate change, and are an externality in the literal sense of the word; (5) some other externalities, such as visual intrusion and, once again, emission of primary pollutants that can alter the landscape and cause damage to historical buildings.

Conclusions

This chapter has discussed the concepts of externality and external costs in transport planning. On the negative side of the coin, both of them contradict the polluter-pays principle: individuals who do not benefit from an activity have to bear (part of) its costs. This is not only unfair; but also it leads to market distortions and inefficient market outcomes. In spite of these acknowledged negative consequences, externalities are widespread, especially those resulting from transport onto the environment, and tend to overshadow the positive consequences of mobility.

Methods to quantify externalities may be rather complex and do not always produce satisfactory results, due to the huge degree of uncertainty that generally affects the process. This statement is generally valid, but the case of GHG emissions is emblematic: here, unitary values of a ton of CO_{2eq} present a range up to six orders of magnitude (Nocera and Tonin, 2014). Far from being a limitation for a correct appraisal of such externalities, this point should be rather a stimulus to reduce the uncertainties in the definition of fair unitary values.

Finally, effective policies will have a decisive role to play when it comes to ensuring that scarce public funds are channeled towards clean and future-proof transport solutions. Particularly, the implementation of more efficient and alternative transport forms will be central to reducing the negative consequences of transport, in parallel with demand measures to foster a modal shift toward active, shared and public mobility. A more environmental-friendly organization of the entire mobility system will be needed to reach the goals of sustainability and climate change. This, rather than a more difficult objective of reducing mobility volumes, might be a more realistic option for society.

References

- Abrantes, P.A.L., Wardman, M.R., 2011. Meta-analysis of UK values of travel time: an update. *Transport. Res. Part A: Policy Pract.* 45 (1), 1–17.
- Arguea, N.M., Hsiao, C., 1993. Econometric issues of estimating hedonic price functions: with an application to the U.S. market for automobiles. *J. Econometr.* 56 (1-2), 243–267.
- Arnott, R., de Palma, A., Lindsey, R., 1993. A structural model of peak-period congestion: a traffic bottleneck with elastic demand. *Am. Econ. Rev.* 83 (1), 161–179.
- Arrow, K., 1969. The organization of economic activity: Issues pertinent to the choice of market versus nonmarket allocation. In *The analysis and evaluation of public expenditures: The PPB system*, ed. Joint Economic Committee. Government Printing Office, Washington, DC.

- Attard, M., Ison, S., 2015. The effects of road user charges in the context of weak parking policies: The case of Malta. *Case Stud. Transport Policy* 3 (1), 37–43.
- Australian Academy of Science, 2019. What is climate change? Online at: <https://www.science.org.au/learning/general-audience/science-climate-change/1-what-is-climate-change>.
- Benjamin, J.D., Sirmans, G.S., 1996. Mass transportation, apartment rent and property values. *J. Real Estate Res.* 12 (1), 1–8.
- Bigazzi, A.Y., Figliozzi, M.A., 2013. Marginal costs of freeway traffic congestion with on-road pollution exposure externality. *Transport. Res. Part A: Policy Pract.* 57, 12–24.
- Black, W.R., 2010. *Sustainable Transportation: Problems and Solutions*. The Guilford Press, New York, NY, USA.
- Braden, J.B., Kolstad, C.D. (Eds.), 1991. *Measuring the Demand for Environmental Quality*. Elsevier, Amsterdam, The Netherlands.
- Bravo-Moncayo, L., Naranjo, J.L., Pavón García, I., Mosquera, R., 2017. Neural based contingent valuation of road traffic noise. *Transport. Res. Part D: Transport Environ.* 50, 26–39.
- Button, K.J., 1990. Environmental externalities and transport policy. *Oxford Rev. Econ. Policy* 6 (2), 61–75.
- Button, K.J., 2006. The political economy of parking charges in “first” and “second-best” worlds. *Transport Policy* 13 (6), 470–478.
- Carrión, C., Levinson, D., 2012. Value of travel time reliability: a review of current evidence. *Transport. Res. Part A: Policy Pract.* 46 (4), 720–741.
- Cavallaro, F., Danielis, R., Nocera, S., Rotaris, L., 2018a. Should BEVs be subsidized or taxed? A European perspective based on the economic value of CO₂ emissions. *Transport. Res. Part D: Transport Environ.* 64, 70–89.
- Cavallaro, F., Giaretta, F., Nocera, S., 2018b. The potential of road pricing schemes to reduce carbon emissions. *Transport Policy* 67, 85–92.
- Chen, H., Rufo, A., Dueker, K.J., 1998. Measuring the impact of light rail systems on single-family home values: a hedonic approach with geographic information system application. *Transp. Res. Rec.* 1617 (1), 38–43.
- Clarkson, R., Deyes, K., 2002. *Estimating the Social Cost of Carbon Emissions*. Department of Environment, Food and Rural Affairs, London, UK.
- Coase, R.H., 1960. The problem of social cost. *J. Law Econ.* 3, 1–44.
- Danielis, R., 2001. La teoria economica e la stima dei costi esterni dei trasporti. Online: <http://www2.units.it/danielis/wp/wp079.pdf>.
- DeLucchi, M., McCubbin, D., 2011. External costs of transport in the United States. In: DePalma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *A Handbook of Transport Economics*. Edward Elgar Publishing Ltd, Northampton, MA, USA.
- de Palma, A., Lindsey, R., 2011. Traffic congestion pricing methodologies and technologies. *Transport. Res. Part C: Emerging Technol.* 19 (6), 1377–1399.
- de Palma, A., Papageorgiou, Y.Y., Thisse, J.-F., Ushchev, P., 2019. About the origin of cities. *J. Urban Econ.* 111, 1–13.
- Diamond, P., 1973. Consumption of externalities and imperfect corrective pricing. *Bell J. Econ. Manage. Sci.* 4, 526–538.
- Efthymiou, D., Antoniou, C., 2013. How do transport infrastructure and policies affect house prices and rents? Evidence from Athens, Greece. *Transport. Res. Part A: Policy Practice* 52, 1–22.
- Eliasson, J., 2009. A cost-benefit analysis of the Stockholm congestion charging system. *Transport. Res. Part A: Policy Practice* 43 (4), 468–480.
- Eliasson, J., Jonsson, L., 2011. The unexpected “yes”: explanatory factors behind the positive attitudes to congestion charges in Stockholm. *Transport Policy* 18 (4), 636–647.
- Ellerman, A.D., Webster, M.D., Parsons, J., Jacoby, H.D., McGuinness, M., 2008. *Cap and Trade: Contributions to the Design of a U.S. Greenhouse Gas Program*. MIT Center for Energy and Environmental Policy Research.
- European Conference of Ministers of Transport, 2000. *Traffic Congestion in Europe*, Round Table 110. Paris, France.
- European Environmental Agency, 2001. Glossary. Online at: <https://www.eea.europa.eu/help/glossary/eea-glossary/Iden>.
- Friedrich, R., Quinet, E., 2011. External costs of transport in Europe. In: DePalma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *A Handbook of Transport Economics*. Edward Elgar Publishing Ltd, Northampton, MA, USA.
- Gadziński, J., Radziński, A., 2016. The first rapid tram line in Poland: how has it affected travel behaviours, housing choices and satisfaction, and apartment prices? *J. Transport Geogr.* 54, 451–463.
- Hanauer, M.M., Reid, J., 2017. Valuing urban open space using the travel-cost method and the implications of measurement error. *J. Environ. Manage.* 198 (2), 50–65.
- Hanley, N., Splash, C.L., 1993. *Cost of Benefit Analysis and the Environment*. Edward Elgar Publishing Ltd, Cheltenham, UK.
- Harper, C.D., Hendrickson, C.T., Samaras, C., 2016. Cost and benefit estimates of partially-automated vehicle collision avoidance technologies. *Accid. Anal. Prevent.* 95-A, 104–115.
- Hensher, D.A., 2008. Climate change enhanced greenhouse gas emissions and passenger transport – What can we do to make a difference? *Transport. Res. Part D: Transport Environ.* 13 (2), 95–111.
- Hensher, D.A., 2018. Tackling road congestion – what might it look like in the future under a collaborative and connected mobility model? *Transport Policy* 66, A1–A8.
- Hess, S., Scarpa, R., 2010. Specification and interpretation issues in behavioural models used for environmental assessment. *Transport. Res. Part D: Transport Environ.* 15-7, 367–369.
- Ignaccolo, M., Inturri, G., Le Pira, M., Capri, S., Mancuso, V., 2016. Evaluating the role of land use and transport policies in reducing the transport energy dependence of a city. *Res. Transport. Econ.* 55, 60–66.
- Landis, J., Guhathakurta, S., Huang, W., Zhang, M., Fukui, B., 1995. Rail transit investments, real estate values, and land use change: a comparative analysis of five California rail transit systems. *Research Monograph*. Available from: <http://escholarship.org/uc/item/4hh7f652>.
- Legaspi, J., Hensher, D., Wang, B., 2015. Estimating the wider economic benefits of transport investments: the case of the Sydney North West Rail Link project. *Case Stud Transport Policy* 3-2, 182–195.
- Lindsey, R., de Palma, A., Silva, H.E., 2019. Equilibrium in a dynamic model of congestion with large and small users. *Transport. Res. Part B: Methodol.* 124, 82–107.
- Lu, Y., Sun, G., Gou, Z., Liu, Y., Zhang, X., 2019. A dose–response effect between built environment characteristics and transport walking for youths. *J. Transport Health* 14, Article 100616.
- Maibach M., Schreyer C., Sutter D., van Essen H.P., Boon B.H., Smokers R., Schroten A., Doll C., Pawlowska B., Bak, M., 2008. *Handbook on estimation of external cost in the transport sector*. IMPACT, Delft, The Netherlands.
- Marshall, A., 1890. *Principles of Economics*. Macmillan, London.
- Massiani, J., 2015. Cost-benefit analysis of policies for the development of electric vehicles in Germany: methods and results. *Transport Policy* 38, 19–26.
- Mohring, H., 1972. Optimization and scale economies in urban bus transportation. *Am. Econ. Rev.* 62, 591–604.
- Mulley, C., Hensher, D.A., Cosgrove, D., 2017. Is rail cleaner and greener than bus? *Transport. Res. Part D: Transport Environ.* 51, 14–28.
- Nocera, S., Cavallaro, F., 2012. Economic evaluation of future carbon impacts on the Italian Highways. *Procedia - Social Behav. Sci.* 54, 1360–1369.
- Nocera, S., Cavallaro, F., 2014. The ancillary role of CO₂ reduction in urban transport plans. *Transport. Res. Proc.* 3, 760–769.
- Nocera, S., Tonin, S., 2014. A joint probability density function for reducing the uncertainty of marginal social cost of carbon evaluation in transport planning. *Adv. Intel. Syst. Comput.* 262, 113–126.
- Nocera, S., Irranca Galati, O., Cavallaro, F., 2018a. On the uncertainty in the economic evaluation of carbon emissions from transport. *J. Transport Econ. Policy* 52-1, 68–94.
- Nocera, S., Ruiz-Alarcón Quintero, C., Cavallaro, F., 2018b. Assessing carbon emissions from road transport through traffic flow estimators. *Transport. Res. Part C: Emerging Technol.* 95, 125–148.
- Pearce, D.W., Turner, R.K., 1990. *Economics of Natural Resources and the Environment*. Johns Hopkins University Press, Baltimore, MD, USA.
- Peer, S., Knockaert, J., Koster, P., Verhoeft, E.T., 2014. Over-reporting vs. overreacting: Commuters’ perceptions of travel times. *Transport. Res. Part A: Policy Pract.* 69, 476–494.
- Pethig, R. (Ed.), 1994. *Valuing the Environment: Methodological and Measurement Issues*. Springer.
- Phaneuf, D.J., Requate, T., 2016. In: *A Course in Environmental Economics - Theory, Policy, and Practice*. Cambridge University Press, Cambridge, UK.
- Pigou, A.C., 1920. *The Economics of Welfare*. Macmillan, London.
- Proost, S., 2011. Theory of external costs. In: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), *A Handbook of Transport Economics*. Edward Elgar Publishing Ltd, Northampton, MA, USA.
- Ricardo-AEA, 2014. *Update of the Handbook on External Costs of Transport*. Final Report. UK: RICARDO-AEA.
- Schrank, D.L., Thomas, T.J., 2009. *Urban Mobility Report*, College Station. Texas Transportation Institute.

- Seo, K., Golub, A., Kuby, M., 2014. Combined impacts of highways and light rail transit on residential property values: a spatial hedonic price model for Phoenix, Arizona. *J. Transport Geogr.* 41, 53–62.
- Sidgwick, H., 1887. *Principles of Political Economy*, 2nd ed. Macmillan, London.
- Sinha, K.C., Labi, S., 2007. *Transportation Decision Making: Principles of Project Evaluation and Programming*. John Wiley and Sons, New York, NY, USA.
- Small, K.A., 1992. Using the revenues from congestion pricing. *Transportation* 19, 359–381.
- Tikoudis, I., Verhoef, E.T., van Ommeren, J.N., 2018. Second-best urban tolls in a monocentric city with housing market regulations. 117A, 342–359.
- Transportation Research Board, 2001. Economic Implications of Congestion. Online at: http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_rpt_463-a.pdf.
- Verhoef, E.T., 1994. External effects and social costs of road transport. *Transport. Res. Part A* 28A (4), 273–287, 199.
- Verhoef, E.T., 2000. The implementation of marginal external cost pricing in road transport Long run vs short run and first-best vs second-best. *Papers Reg. Sci.* 79 (3), 307–332.
- Welch, T.F., Gehrke, S.R., Wang, F., 2016. Long-term impact of network access to bike facilities and public transit stations on housing sales prices in Portland, Oregon. *J. Transport Geogr.* 54, 264–272.
- Zhang, F., Wang, X.H., Nunes, P.A.L.D., Ma, C., 2015. The recreational value of gold coast beaches Australia: an application of the travel cost method. *Ecosyst. Services* 11, 106–114.

Further Reading

- D. Banister *Transport Policy and the Environment* 1998 Spon London, UK.
- K. Button *Transport Economics* 3rd edition 2010 Edward Elgar Cheltenham, USA.
- J. Cowie *The Economics of Transport—A Theoretical and Applied Perspective* 2010 Routledge London, UK.
- A. de Palma R. Lindsey E. Quinet R. Vickerman *A Handbook of Transport Economics* 2011 Edward Elgar Publishing Ltd Northampton, MA, USA.
- H. Folmer E.C. van Ierland *Valuation Methods and Policy Making in Environmental Economics* 1989 Elsevier, Amsterdam The Netherlands.
- Z. Li *Transportation Asset Management—Methodology and Applications* 2019 CRC Press Boca Raton FL, USA.
- Maibach, M., Schreyer, C., Sutter, D., Van Essen, H.P., Boon, B.H., Smokers, R., Schroten, A., Doll, C., Pawlowska, B., Bak, M., 2007. Handbook on estimation of external cost in the transport sector (No. 07.4288.52). CE Delft, The Netherlands..
- C. Nash B. Matthews *Measuring the Marginal Social Cost of Transport. Research in Transportation Economics, Volume 14* 2005 Elsevier Oxford, UK.
- S. Nocera *Feasibility Decisions in Transportation Engineering – Strategies for Transport Evaluation* 2010 McGraw-Hill Milan, Italy.
- J.-P. Rodrigue *The Geography of Transport Systems* 3rd edition 2013 Routledge London, UK.
- K.C. Sinha S. Labi *Transportation Decision Making: Principles of Project Evaluation and Programming* 2007 John Wiley and Sons New York, NY, USA.
- K.A. Small E. Verhoef *The Economics of Urban Transportation*. 2nd ed 2007 Routledge London, UK.

Planning for Public Transport with Automated Vehicles

Gonalo Homem de Almeida Rodriguez Correia, TU Delft, Faculty of Civil Engineering and Geosciences (CiTG), Transport & Planning Department Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Glossary	198
Introduction	198
Applications of Vehicle Automation in Public Transport	199
What Do Cities Want From Vehicle Automation?	200
Potential Impacts on Urban Mobility	201
Travelers' Behavior	201
Transport Modeling Studies	201
Planning for AVs as PT	202
Conclusions	202
Acknowledgments	203
See Also	203
References	203
Further Reading	204

Glossary

AV Automated vehicle
ODD Operational design domain
PT Public transport
SAV Shared automated vehicle
TNC Transport network company (Uber-like systems)

Introduction

The objective of this article is to give the reader an overview of possible applications of automated vehicles (AVs) technology within the scope of public transport (PT) systems planning and operations. This shall include shared automated vehicles (SAVs) as well. Given the initial state of development of these services, it is naturally not possible to present consolidated knowledge about their application, let alone on the impacts of these systems since they are mostly being tested in pilot studies as of the past few years. Nevertheless, it is already possible to disclose some of their applications and point out several advantages and disadvantages given the modeling studies and pilot testing. The article is not meant to be a literature review on the topic, but it does resort to relevant literature when this supports some of the planning challenges and opportunities of using AVs as PT avoiding merely speculating around the topic. In this article, we will be considering Level 4 and Level 5 automation levels as reference according to the SAE definition of automation (SAE International, 2016) meaning that vehicles can drive themselves without any need for a human driver either in all operational design domains (ODDs) (Level 5) or a specific ODD (Level 4). This does not preclude the fact that currently in most of the pilots, there is the requirement for the presence of a steward on board. The article is focused on road transport, leaving aside driverless trains or metro systems (Grogan, 2012) or drones, which are more associated with logistic systems (Kunze, 2016). One does not intend to estimate the time at which these systems will be fully matured for regular operation in our cities, this is still subject to controversy and high uncertainty (Nieuwenhuijsen et al., 2018).

Research and applications in the field of automated driving have been advancing quite fast, especially in the past 5 years (Ainsalu et al., 2018; de Almeida Correia et al., 2018), which has much to do with the moment that Google started spreading the word about their Google car project. The so-called Firefly car, built just for studying AVs, was unveiled in 2014. However, before that, there have been decades of technological development that led to the possibility of transforming dreams into reality. There are many studies related to the technology of the vehicles focused on obstacle detection and vehicle motion (Katrakazas et al., 2015), but also on studying traffic flows performance with AVs, specifically at analyzing the capacity gains with automation and connectivity (cooperation between vehicles) (van Arem and Tsao, 1997; Schakel et al., 2010). Nevertheless, this article is not intended to look at the technology challenges of vehicle automation but rather to look at their potential application and impacts in PT systems and how to plan for these.

In traditional PT systems, namely, bus systems, there is an interest in automation with a view to increasing flexibility in routing and pathfinding as well as to save personnel costs (Winter et al., 2018), specifically drivers, which account for up to 70% of PT operational costs (Currie, 2018). With the lower costs, in a future stage where the vehicles themselves are not too expensive, it would

be possible to cover more areas and eventually increasing equity in the accessibility of different regions. Moreover, without the drivers, it is much easier to adapt the service to the existing demand and not be hindered by drivers' working conditions, namely, shifts (Alessandrini et al., 2014). There are nowadays many pilots being run with vehicles that are called pods, with only a few seats, which are supplied by specialized companies aiming at testing the concept and also at making a profit out of the wish of cities and regions to join the movement. For an overview of many of these pilots, the reader may consult (Ainsalu et al., 2018).

It is also possible to observe great interest in SAV systems, which come fostered by the great growth of transport network companies (TNCs) such as Uber and Lyft. These are demand responsive and tackle some of the limitations of traditional PT (Boesch et al., 2016; Fagnant and Kockelman, 2018; Martinez et al., 2014; Scheltes and Correia, 2017). At the same time automation comes with the promise of facilitating several of the daily operations of carsharing, namely, relocating vehicles between different areas of a city, since current carsharing companies have considerable costs moving those vehicles around with drivers (Kek et al., 2009). This will represent a merging between TNCs and carsharing systems, since there will be no way of differentiating them.

In the following sections of this article, the different PT AV systems are defined along with their envisioned applications based on state of the art research and on the state of the world. In several cases, there are real systems in operation, which can help shed some light on the potential impacts of using AVs for providing PT, particularly in urban regions. We focus on planning issues for this technology aiming at exploring the challenges and opportunities, especially for cities. Planning for AVs is a very recent challenge with most of the cities not having any strategy for this future; however, some cities have started a debate for planning for automation. Moreover, there are lessons learned, both from research and applications, that can help set some recommendations for future planning of PT systems with AVs and clarify concepts for the reader who is interested in this field.

The article is structured in the following way. In the next section, the main applications of vehicle automation in PT are presented, which also includes the connection to the private usage of AVs. This continues with a section that explores what cities mostly want as an application of AVs in their mobility systems. The next section focusses on presenting expected impacts from AV application as PT. This is followed by presenting some guidelines for answering some key questions of cities on how to plan for AVs. The chapter ends with the main conclusions that can be drawn from the article.

Applications of Vehicle Automation in Public Transport

In this article, we propose to distinguish between what can be called regular service, but also on-demand, bus services with AVs, and shared mobility systems with AVs (SAV systems). However, it should be noted that such distinction is becoming quite blurred as TNCs such as Uber (cars) and ViaVan (vans) are essentially supplying the same type of service to their users, which is basically on-demand highly flexible transport services that are door-to-door and take advantage of information technology available through smartphones.

Regardless of the possibility of such systems merging due to similarity of operations, studies, and pilots, as referred in the introduction, have been organized between looking at vans (small buses or pods) and cars as the main vehicle to provide PT with automation on the roads. Usually the former is associated with only one part of the trip such as first/last mile or a specific demand segment such as transport for disabled or elderly users, while the latter is typically associated with taxi systems that are intended to serve all mobility in a region in a door-to-door fashion either for single clients or in a ridesharing mode.

Fig. 1 shows a schematic diagram where these two applications are presented, leading to a series of questions that are very important for transport planners and are being addressed through research of different kinds: pilots, transport modeling studies,

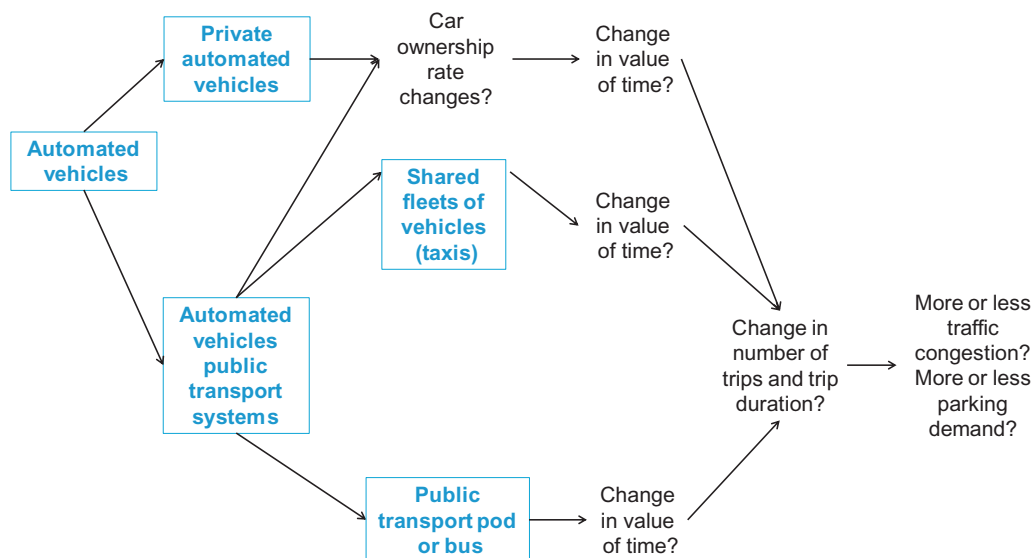


Figure 1 Two main usages of vehicle automation in public transport (PT) systems and relevant associated questions.

qualitative studies, and so on. The ultimate set of questions in this figure regards traffic congestion and parking demand, which are two of the most important variables to measure urban mobility performance.

Two main behavior questions are expressed in the figure: what happens to the value of travel time? And what happens to car ownership as a result of both these usages of AVs for PT? These will influence the number of trips done using the several modes of transport available as well as the trip duration in a certain area, which will then affect traffic congestion and parking demand. The real system is clearly more complex than what is expressed in [Fig. 1](#); however, this is simple enough to capture some of the major issues and opportunities of vehicle automation.

The value of travel time measures the trade-off between money and the experience or usefulness of traveling, and it is essentially a perceived indicator as it depends on the people, type of trip, and the mode of transport itself ([Mackie et al., 2001](#)). This is linked to the attractiveness of a mode when it compares to the other competing modes of transport. The general intuition is that a better experience inside AVs will lead to a lower value of travel time, therefore, a higher preference to spend time inside these vehicles compared to current ones. Nevertheless, if this added experience can be realized, it will most likely be on board of private cars, which today have to be driven by the traveler ([de Almeida Correia et al., 2019](#)). In the case of PT systems, the differences are not so obvious.

Private car ownership is linked to the experience in PT AVs, namely the value of travel time savings, but also on what will happen to the private cars themselves. It cannot be assumed that private cars will not evolve as much as PT AV systems. In fact, a great part of the innovation is being driven by carmakers not PT systems per se. These cars will obviously be used in the shared fleets of vehicles mentioned in [Fig. 1](#) as well, therefore benefits can be common to both private and public systems. What will be the most likely usage of these cars is dependent not only on travelers' behavior but also on the policies that will be implemented by public authorities. Currently, cities are already deciding on which vehicles should or should not be allowed to circulate in specific parts of their territories.

Like changes in car ownership, there is also much speculation around the travel times that people will be perceiving as acceptable when riding in an AV, which is again connected to value of travel time savings and also to the vehicle ownership itself. Once more this may make more sense to be discussed with reference to private AVs rather than PT systems where sharing a vehicle is in most cases part of the equation and being a passenger always a reality. It is for private cars that visions of moving living rooms are being put forward. For a SAV, or automated bus/pod, there is the possibility of a redesigned interior since the driver is no longer needed, but it is still difficult to understand how much that interior can be redesigned and customized to the needs of travelers.

The debate over the impacts of AVs in urban transport systems and the overall quality of the urban environments is at its peak. On one hand, AVs could contribute to taking passengers from traditional mass transport systems, such as trains and trams, thus contributing to higher energy spending. [Shaheen and Chan \(2016\)](#) analyzed the impacts of TNCs and concluded that many trips have been diverted from those modes to those new ways of traveling in cities. On the other hand, with AVs there is the possibility of routing these vehicles smartly trying to avoid traffic congestion ([Liang et al., 2018](#)) and parking away from the curb potentially releasing space for other usages. Nevertheless, for SAVs, the requirement for pick-up and drop-off locations is essential. In a recent modeling study by the ITF-OECD authors concluded that "striving to reconcile efficient and safe pick-up and drop-off activities with the productive occupation of space will be a central challenge going forward" ([ITF, 2018](#)). Despite attracting people away from traditional PT, SAVs can yield some benefits: with the use of information technology and the routing flexibility, there is the possibility of maximizing the occupancy of cars and usage of fleets of smaller vehicles 1-seat or 2-seats may be possible as well ([Wadud et al., 2016](#)).

What Do Cities Want From Vehicle Automation?

One of the first relevant pilots with automated shuttles has been the CityMobil and CityMobil2 projects funded by the European Union. The first project ran between 2006 and 2011, including several pilots and modeling studies through which it was concluded that "cybercars" would be better designed "as feeders for PT in areas of low population density" ([van Dijke and van Schijndel, 2012](#)). In CityMobil2 it was concluded that the greatest potential usage for PT AVs would be as "medium-size vehicle as on-demand transport services feeding conventional mass transits in the suburbs of large cities or on radial corridors as complementary mass transits with large busses and platoons of them and as main PT for small cities with personal vehicles" ([Alessandrini et al., 2014](#)).

The industry took the hint and nowadays companies like Navya (2014, France), EasyMile (2014, France), and Local Motors (2007, USA) are providing vehicles for pilot testing all over the world. The ParkShuttle system created by 2getthere (2005, Capelle aan den IJssel, The Netherlands) is the only mature system operational on a daily basis and even without a steward on board ([ATRA, 2020](#)), nevertheless, this is what can be traditionally classified as a personal rapid transit (PRT) system because the vehicles drive on a segregated lane guided by magnets on the road.

There could be several application contexts for on-demand automated bus services since all applications will benefit from savings in personnel costs as referred before, but the one that has been gaining more attention is the first/last-mile connection to higher capacity traditional PT systems such as rail. Recognizing that PT is not able to cover all areas of a city and country and that supplying lower density areas or covering off-peak periods can lead to high costs, automation is naturally being seen as an alternative to cover the connection to and from train or bus stations much like the CityMobil initial project concluded.

In a survey from 2017, Bloomberg asked 38 cities, which were identified as being actively preparing for AVs, what they expected from AV technology and the first and foremost anticipated usage of AVs is what they called the “last mile transit.” Naturally, with the great investment that has been done in the past decades and, in some cases, over the past century on expensive transport infrastructure such as trams and trains for moving masses of travelers, cities are not prioritizing the usage of private AVs in their streets or substituting those networks by AV-based systems. Even if private AVs may increase the efficiency of the usage of each family vehicle by being able to move around driverless to pick-up and drop-off family members (de Almeida Correia and van Arem, 2016). Research has been showing that the first/last-mile trip is many times the most undervalued stage of a multimodal transit trip. But this is a critical stage even for broader goals such as providing equity in job accessibility (Boarnet et al., 2017). Using flexible transport for these connections is not new, already with TNCs today, cities are trying to find ways of harnessing their flexibility and convenience to provide a first and last-mile solution by subsidizing these companies (Shaheen and Chan, 2016).

Potential Impacts on Urban Mobility

Travelers' Behavior

According to Fig. 1, the behavior component of the travelers is essential to ultimately explain what impacts will result in urban mobility systems and therefore the usefulness or nuisance of these systems. In that regard, the question of whether a change in the value of travel time will occur because of these systems when compared to existing ones is very important.

There are still scant studies in measuring the willingness of people to use AVs as PT and these tend to point for somewhat disappointing directions in terms of perceived advantages of such technologies in PT. Yap, de Almeida Correia and van Arem (2016) conducted a seminal stated preference survey in the Netherlands for assessing the AV for last mile of rail trips compared to the existing modes and concluded that the value of travel time increased rather than decreases, therefore, there was no perceived added comfort. There is still great enjoyment among the population in manually driving a vehicle (Kyriakidis et al., 2015); moreover, there may be an issue about safety in using this technology. In a survey from 2017, Deloitte found that quite a number of developed countries had a significant part of their populations feeling they could not trust automated driving. For example, in South Korea this share was 81% and Germany 72%. But it is striking that one year later the same survey had much more optimistic results with a decrease to 54 and 45% for those two countries, respectively (Deloitte, 2018). This goes to show that preferences and perceptions may be changing quite rapidly making it challenging for planners.

Paradoxically when asking people what is the benefit they mostly expect from automated driving people who are willing to use such vehicles tend to point toward safety as the most important one (Shabanpour et al., 2017). There seems to be a tendency for younger well-educated and tech-savvy males to be more willing to try these vehicles, although this is not conclusive (Fu et al., 2019). It is interesting though to notice how most of the time these behavioral studies, especially those that use stated preference techniques, have focused more on the usage of AVs as private transport rather than public vehicles, while for transport modeling studies the focus has been much more on SAVs.

Preferences of travelers will unsurprisingly be connected to what can be done inside an AV and how this compares to currently manually driven cars. One of the few studies that compares current PT vehicles to what the authors called automated driving transport service (ADTS), which is an automatically controlled door-to-door transport service, is the one (Ashkrof et al., 2019) in which it was concluded that “conventional cars and public transportation are perceived as being the least attractive alternatives concerning in-vehicle travel time on short- and long-distance commuting trips, respectively. Preference for ADTS lays between the car and public transportation, neither the best nor the worst alternative in all scenarios.” Showing that it is difficult to clearly mark the advantages of AVs in PT.

The elderly are often pointed as the great beneficiaries of vehicle automation because they would be able to use a car without having to drive one. This can pretty much be said about any group that does not have or does not want to have a driving license. Nevertheless, this is more related to private ownership of an AV rather than PT. Logically if it will be easier and cheaper to implement PT compared to before, this can also benefit people without a driver's license. The elderly (at least the elderly of today) “see benefits in increased mobility and therefore greater independence at a higher age as it allows disabled people to regain lost mobility” (Schmargendorf et al., 2018), and this cannot be ignored.

Transport Modeling Studies

Most transport modeling studies have been focusing on modeling SAV systems because these have a very different way of being operated compared to traditional PT systems. The dynamics of the movements of these vehicles on a network make the supply side of the transport system much more uncertain and highly dependent on the operational strategies that are taken to manage such systems. This had already been concluded regarding one-way carsharing systems whereby trips taken by some clients affect the supply that is going to be offered in the next minutes in a certain area of a city (Jorge and Correia, 2013). The added component of considering sharing the ride among different clients adds more complexity, hence the need to use agent-based disaggregated models where travelers and cars are individually represented (Martinez et al., 2015).

The most well-know studies about modeling the effects of SAVs in a city are probably those of Kara Kockelman (Fagnant and Kockelman, 2018) in Texas and the ITF-OECD model developed by Luis Martinez (Martinez and Viegas, 2017). Typically, these studies aim at estimating the number of required SAVs to satisfy an existent number of trips. Agent-based models explore different AV concepts and conclude that a fleet of shared AVs could reduce the number of vehicles needed to 10% of the current number in order

to deliver the same trips. [Fagnant and Kockelman \(2014\)](#) reach the same conclusion and indicated that it is possible to satisfy the same travel demand with each shared AV replacing around 11 conventional vehicles. However, research also shows that the number of vehicle kilometers traveled (VKT) is likely to increase due to unoccupied trips of SAVs. The increase of VKT could lead to congestion, which might negatively influence the environment because of pollution and the higher required amount of energy. Transport-mode choice is rarely integrated into such methods, which might hide the transfer of passengers from mass PT to car transport as [Shaheen and Chan \(2016\)](#) noted about TNCs.

For the planning of first/last-mile PT, there are several modeling studies that aim at understanding how AVs can play a role in providing access to mass PT. From a behavior point of view [Yap et al. \(2016\)](#) study, that has already been mentioned, showed that people may not find it very different from current options. At least as perceived back in 2016. Nevertheless, it is important to look at the operational side as well since this can determine the potential to use these vehicles as part of broader networks. [Liang et al. \(2016\)](#) studied the usage of shared electric AVs for accessing a railway station by optimally planning their operational area in a profit maximization perspective showing that these systems may be profitable for a certain catchment area. However, EVs are still a limitation due to their charging time. [Scheltes and Correia \(2017\)](#) used the same case study in the Netherlands for a simulation study of the same type of system showing that results on level of service offered to clients is very dependent on the operational strategies such as having or not proactive relocations of the vehicles or how to charge the vehicles if they are electric: one charging station next to the rail station or several ones scattered in the operation area. [Shen et al. \(2018\)](#) studied the integration of SAVs as first/last-mile access to a PT network of high capacity buses and concluded that the SAVs have the potential of improving level of service. [Winter et al. \(2018\)](#) studied the substitution of regular bus services by on-demand automated buses showing that the waiting costs, when added to a generalized cost function of the system performance, can go down considerably because of the expected reduction in waiting time. Nevertheless, most of these studies do not internalize traffic congestion and other more micro aspects such as interaction with other road users. In several cases it is difficult to see what the difference is between AVs and the existing taxi systems or demand-responsive transit systems constituted by human-driven buses besides the lower costs.

Planning for AVs as PT

Cities and regions are driven by goals associated with strategies designed to obtain quantifiable key performance indicators, which are not limited to mobility/accessibility indicators. City planning for AVs is not yet widespread. [Freemark et al. \(2019\)](#) in a recent survey paper found that few local governments in the United States have started to prepare for AVs and have mixed feelings about the effects of AVs. Nevertheless, in the same study, it was found that for those cities that have a plan the most frequent two goals for the implementation of AVs are “increase street safety” and “support transit.” “Improving sustainability” comes right after those two. Municipal authorities seem to find it important to regulate the usage of AVs, especially those who are more skeptical about the impacts of AVs. Nevertheless, the same survey finds that most city plans fail to present any concrete measure or policy for AVs, keeping the speech very general about what may happen with AVs and stating the need to see how this technology evolves. Some cities are suggesting focusing on the testing of technology, which seems to match what is happening in Europe as well where the effort is on piloting. There is a large number of officials who point for concerns with added VKT and urban sprawl associated with lower transit patronage and employment.

Cities will face challenges regarding governance mechanisms and regulations to accommodate a new reality with AVs that has potential impacts on so many subsystems managed by public authorities: movement of people and goods, PT management, traffic congestion, land use, safety, equity, and so on. But still, things are quite undefined from the transport system perspective. Coming back to [Fig. 1](#), it is quite hard still to foresee what usage of AVs will prevail let alone the impacts that can be expected. Specific challenges that cities are facing or will face in the future are summarized in [Table 1](#).

Conclusions

Planning for the usage of AVs as PT is still in its infancy. Despite many pilots where AVs are being tested, most of the work being done is on research aiming at estimating the impacts of different ways of using AVs as part of the transport system. Very few examples can be found of real PT systems integrated with other modes of road transport.

Cities are recognizing the need to plan for the mobility innovations that are being fostered by the TNCs and AV hype. This is reinforced by the fact that some are already facing troubles with TNCs, but regarding AVs, they are not yet developing regulations and specific policies other than promoting experimentation with this technology. This is fair in a sense, given the ill estimation of the impacts of AVs and the difficulty in estimating the probability of every scenario of application of these vehicles together with other system uncertainties in the future.

In this article, we have identified the different ways in which AVs may be used in the PT systems. We have distinguished private AVs, shared AVs, and pod-like systems and have put forward a simple conceptual model of several key behavioral questions associated with these options, namely, regarding the value of travel time and the frequency and length of the trips people will do in the future. It has been found that AVs used as PT may not bring a very different experience, but they do provide an opportunity for lowering the supply costs particularly as first/last mile of existing mass PT.

Table 1 Questions and tentative answers about planning challenges for automated vehicles (AVs) as PT

Challenges	Tentative answers
Regulate how private AVs can access public space? In a reality where private vehicles are being taken out of the cities what should be done with private AVs if they win the battle of the two markets? Amsterdam municipality intends to ban all petrol cars from the city by 2030, for example, (Boffey, 2019).	If private AVs become very attractive, this may lead to higher usage of private cars which ultimately, despite higher household trip satisfaction, may lead to more traffic congestion. Allowing them in certain high demand areas may exacerbate traffic congestion even because people may feel motivated to be comfortably inside one of such vehicles. Exclusion areas have to be defined if demand increases. Though one must not ignore the potential positive effects on traffic safety and efficiency.
Where and how should AVs be allowed to stop for pick-up and drop-off of their clients? Where to locate designated areas for this to happen? Should these movements be forbidden altogether?	Research is showing the challenge that may be posed by the stopping of these vehicles on the road. Some of the currently used parking spaces may have to be converted into stopping areas. Some parts of the road network may be forbidden for these operations. Current TNCs operations serve as a guide for the future.
How to implement AVs into PT systems? Should all possible usages for these vehicles be allowed or should priority be given for a specific usage?	It seems first/last-mile connections should be prioritized particularly in lower density areas of cities that struggle with the provision of good PT. These areas are also usually easier to convert when compared to more dense consolidated areas. Avoiding investing in gimmicks is a concern given the costs of these pilots and a shortage of investment capabilities of most cities.
Should investments be driven by the cities or by private companies? How should public authorities handle their interaction with these private stakeholders?	Current technology is being developed by private companies that want a take on future mobility systems. They are driven by profit. But many PT operators are already private in many cities around the world whereby the service is defined centrally, and concessions are given. The private initiative is not negative per se, but it may need regulation that is already applicable to TNCs, especially in reference to how SAVs will be offered. Competition and access to the market of automated TNCs will also require attention with clear rules on how to operate.
How to tune land use planning and the usage of AVs? Potential for sprawl or leverage for sustainable development?	Private AVs may eventually increase pressure toward farther regions from the city centers where traditional PT performs better. This will have to be controlled. But from a PT perspective, AVs can play an important role in fostering the adoption of transit-oriented development (TOD) policies in which the first/last-mile connections are essential to expanding catchment areas around rail systems, for example. The eventual conversion of parking spaces into quality public space can help move back people who can find themselves reconnected to city centers.

Acknowledgments

This research has been partially supported by the e-Hubs project, partially funded by the Interreg program. It has also been supported by the Susmo project funded by the Climate-KIC. The work is supported by the hEAT lab at TU Delft (research on Electric and Automated Transport), which the author co-directs.

See Also

Mobility as a Service; Mobility Planning and Policies for Older People; Demand Responsive Transport; Connected and Autonomous Vehicles: Priorities for Policy and Planning; Customer Satisfaction as a Measure of Service Quality in Public Transport Planning; Transport Demand Management; Public Transport Network Planning; Car Sharing and the Impact on New Car Registration; Planning for Bus Priority; Transitions and Disruptive Technologies in Transport Planning.

References

- Ainsalu, J., et al., 2018. State of the art of automated buses. *Sustainability (Switzerland)* 10 (9), 1–34, doi:10.3390/su10093118.
- Alessandrini, A., Cattivera, A., et al., 2014. CityMobil2: challenges and opportunities of fully automated mobility. In: Meyer, G., Beiker, S. (Eds.), *Road Vehicle Automation*. Springer International Publishing, Cham, pp. 169–184, doi:10.1007/978-3-319-05990-7_15.
- Alessandrini, A., Alfonsi, R., et al., 2014. Users' preferences towards automated road public transport: results from european surveys. *Transp. Res. Procedia* 3, 139–144, doi:10.1016/j.trpro.2014.10.099.
- Ashkrof, P., et al., 2019. Impact of automated vehicles on travel mode preference for different trip purposes and distances. *Transp. Res. Rec.* 2673 (5), 607–616, doi:10.1177/0361198119841032.
- ATRA, 2020. ATRA (Advanced Transit Association). Available from: <http://www.advancedtransit.org/advanced-transit/applications/rivium/> (accessed 06.10.2020).
- Boarnet, M.G., et al., 2017. First/last mile transit access as an equity planning issue. *Transp. Res. Part A Policy Pract.* 103, 296–310, doi:10.1016/j.tra.2017.06.011.

- Boesch, P.M., Ciari, F., Axhausen, K.W., 2016. Autonomous vehicle fleet sizes required to serve different levels of demand. *Transp. Res. Rec.* 2542, 111–119, doi:10.3141/2542-13.
- Boffey, D., 2019. Amsterdam to ban petrol and diesel cars and motorbikes by 2030. *The Guardian*, 3 May. Available from: <https://www.theguardian.com/world/2019/may/03/amsterdam-ban-petrol-diesel-cars-bikes-2030>.
- Currie, G., 2018. Lies, damned lies, AVs, shared mobility, and urban transit futures. *J. Public Transp.* 21 (1), 19–30, doi:10.5038/2375-0901.21.1.3.
- de Almeida Correia, G.H., van Arem, B., 2016. Solving the user optimum privately owned automated vehicles assignment problem (UO-POAVAP): a model to explore the impacts of self-driving vehicles on urban mobility. *Transp. Res. Part B Methodol.* 87, 64–88.
- de Almeida Correia, G.H., de, A., et al., 2018. Editorial of the TR Part C special issue on: "Operations with automated vehicles (AVs): Applications in freight and passenger transport". *Transp. Res. Part C Emerg. Technol.* 95, 493–496, doi:10.1016/j.trc.2018.07.014.
- de Almeida Correia, G.H., de, A., et al., 2019. On the impact of vehicle automation on the value of travel time while performing work and leisure activities in a car: Theoretical insights and results from a stated preference survey. *Transp. Res. Part A Policy Pract.* 119, 359–382, doi:10.1016/j.tra.2018.11.016.
- Deloitte, 2018. A reality check on advanced vehicle technologies. Evaluating the big bets being made on autonomous and electric vehicles. Available from: <https://www2.deloitte.com/us/en/insights/industry/automotive/advanced-vehicle-technologies-autonomous-electric-vehicles.html> (accessed 06.10.2020).
- Fagnant, D.J., Kockelman, K.M., 2014. The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. *Transp. Res. Part C Emerg. Technol.* 40, 1–13, doi:10.1016/j.trc.2013.12.001.
- Fagnant, D.J., Kockelman, K.M., 2018. Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. *Transportation* 45 (1), 143–158, doi:10.1007/s11116-016-9729-z.
- Freemark, Y., Hudson, A., Zhao, J., 2019. Are cities prepared for autonomous vehicles? Planning for technological change by U.S. local governments. *J. Am. Plann. Assoc.* 85 (2), 133–151, doi:10.1080/01944363.2019.1603760.
- Fu, M., Rothfeld, R., Antoniou, C., 2019. Exploring preferences for transportation modes in an urban air mobility environment: munich case study. *Transp. Res. Rec.* 2673 (10), 427–442, doi:10.1177/0361198119843858, 036119811984385.
- Grogan, A., 2012. Driverless trains: it's the automatic choice. *Eng. Technol.* 7 (5), 54–57, doi:10.1049/et.2012.0514.
- ITF, 2018. The shared-use city: managing the curb. Paris. Available from: <https://www.itf-oecd.org/shared-use-city-managing-curb-0>.
- Jorge, D., Correia, G., 2013. Carsharing systems demand estimation and defined operations: a literature review. *Eur. J. Transp. Infrastructure Res.* 13 (3), 201–220.
- Katrakazas, C., et al., 2015. Real-time motion planning methods for autonomous on-road driving: state-of-the-art and future research directions. *Transp. Res. Part C Emerg. Technol.* 60, 416–442, doi:10.1016/j.trc.2015.09.011.
- Kek, A.G.H.H., et al., 2009. A decision support system for vehicle relocation operations in carsharing systems. *Transp. Res. Part E Logist. Transp. Rev.* 45 (1), 149–158, doi:10.1016/j.tre.2008.02.008.
- Kunze, O., 2016. Replicators, ground drones and crowd logistics a vision of urban logistics in the year 2030. *Transp. Res. Procedia* 19, 286–299, doi:10.1016/j.trpro.2016.12.088.
- Kyriakidis, M., Happee, R., de Winter, J.C.F., 2015. Public opinion on automated driving: results of an international questionnaire among 5000 respondents. *Transp. Res. Part F Traffic Psychol. Behav.* 32, 127–140, doi:10.1016/j.trf.2015.04.014.
- Liang, X., Correia, G.H., de, A., van Arem, B., 2016. Optimizing the service area and trip selection of an electric automated taxi system used for the last mile of train trips. *Transp. Res. Part E Logist. Transp. Rev.* 93, 115–129, doi:10.1016/j.tre.2016.05.006.
- Liang, X., de Almeida Correia, G.H., van Arem, B., 2018. Applying a model for trip assignment and dynamic routing of automated taxis with congestion: system performance in the city of Delft, The Netherlands. *Transp. Res. Rec.* 2672 (8), 588–598, doi:10.1177/0361198118758048.
- Mackie, P.J., Jara-Diaz, S., Fowkes, A.S., 2001. The value of travel time savings in evaluation. *Transp. Res. Part E* 37, 91–106.
- Martinez, L.M., Viegas, J.M., 2017. Assessing the impacts of deploying a shared self-driving urban mobility system: an agent-based model applied to the city of Lisbon, Portugal. *Int. J. Transp. Sci. Technol.* 6, 13–27, doi:10.1016/j.ijtst.2017.05.005.
- Martinez, L.M., Correia, G.H.A., Viegas, J.M., 2014. An agent-based simulation model to assess the impacts of introducing a shared-taxi system: an application to Lisbon (Portugal). *J. Adv. Transp.* 49 (3), 475–495, doi:10.1002/atr.1283.
- Martinez, L.M., Correia, G.H.A., Viegas, J.M., 2015. An agent-based simulation model to assess the impacts of introducing a shared-taxi system: an application to Lisbon (Portugal). *J. Adv. Transp.* 49 (3), 475–495, doi:10.1002/atr.1283.
- Nieuwenhuijsen, J., et al., 2018. Towards a quantitative method to analyze the long-term innovation diffusion of automated vehicles technology using system dynamics. *Transp. Res. Part C Emerg. Technol.* 86, 300–327, doi:10.1016/j.trc.2017.11.016.
- SAE International, 2016. Summary of SAE International's levels of driving automation for on-road vehicle Global Ground Vehicle Standards (J3016).
- Schakel, W.J., van Arem, B., Netten, B.D., 2010. Effects of cooperative adaptive cruise control on traffic flow stability. In: 13th International IEEE Conference on Intelligent Transportation Systems. IEEE, 759–764. doi: 10.1109/ITSC.2010.5625133
- Schelles, A., de Almeida Correia, G.H., 2017. Exploring the use of automated vehicles as last mile connection of train trips through an agent-based simulation model: An application to Delft, Netherlands. *Int. J. Transp. Sci. Technol.* 6 (1), 28–41, doi:10.1016/j.ijtst.2017.05.004, Tongji University and Tongji University Press.
- Schmargendorf, M. et al., 2018. Autonomous driving and the elderly: perceived risks and benefits. *Mensch und Computer* 2017. doi: 10.18420/muc2018-ws11-0524.
- Shabanpour, R. et al., 2017. Consumer preferences of electric and automated vehicles. In: 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). IEEE, pp. 716–720. doi: 10.1109/MTITS.2017.8005606.
- Shaheen, S., Chan, N., 2016. Mobility and the sharing economy: potential to facilitate the first- and last-mile transit connections. *Built Environ.* 42 (4), 573–588, doi:10.2148/benv.42.4.573.
- Shen, Y., Zhang, H., Zhao, J., 2018. Integrating shared autonomous vehicle in public transportation system: a supply-side simulation of the first-mile service in Singapore. *Transp. Res. Part A Policy Pract.* 113, 125–136, doi:10.1016/j.tra.2018.04.004.
- van Arem, B., Tsao, H.-S.J., 1997. The development of automated vehicle guidance systems: commonalities and differences between the state of California and the Netherlands. *Advances in intelligent transportation system design*. Available from: <http://papers.sae.org/972672/> (accessed 04.12.2014).
- van Dijke, J.P., van Schijndel, M., 2012. CityMobil, advanced transport for the urban environment. *Transp. Res. Rec.* 2324 (1), 29–36, doi:10.3141/2324-04.
- Wadud, Z., MacKenzie, D., Leiby, P., 2016. Help or hindrance? The travel, energy and carbon impacts of highly automated vehicles. *Transp. Res. Part A Policy Pract.* 86, 1–18, doi:10.1016/j.tra.2015.12.001.
- Winter, K., et al., 2018. Performance analysis and fleet requirements of automated demand-responsive transport systems as an urban public transport service. *Int. J. Transp. Sci. Technol.* 7 (2), 151–167, doi:10.1016/J.IJTST.2018.04.004.
- Yap, M.D., Correia, G., van Arem, B., 2016. Preferences of travellers for using automated vehicles as last mile public transport of multimodal train trips. *Transp. Res. Part A Policy Pract.* 94, 1–16, doi:10.1016/j.tra.2016.09.003.

Further Reading

For an overview on all potential AV impacts the reader is recommended to consult:

- Milakis, D., van Arem, B., van Wee, B., 2017. Policy and society related implications of automated driving: a review of literature and directions for future research. *J. Intell. Transp. Syst.* 21 (4), 324–348, doi:10.1080/15472450.2017.1291351.

For an overview on modeling studies regarding automated vehicles and travel behavior the reader is recommended to consult:

Soteropoulos, A., Berger, M., Ciari, F., 2019. Impacts of automated vehicles on travel behaviour and land use: an international review of modelling studies. *Transp. Rev.* 39 (1), 29–49, doi:10.1080/01441647.2018.1523253.

For planning issues for AVs in transit systems:

Grush, B., Niles, J., 2018. *The End of Driving: Transportation Systems and Public Policy Planning for Autonomous Vehicles*. Elsevier, Amsterdam, doi:10.1016/C2017-0-03350-5.

Urban Congestion Charging in Transport Planning Practice

Ida Kristoffersson, Maria Börjesson, VTI Swedish National Road and Transport Research Institute, Stockholm, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Rationale for Introducing Congestion Charging	206
Urban Congestion Charging Systems Around the World	206
Effects of Congestion Charging	207
Short-Term Effects	207
Long-Term Effects	208
Designing a Congestion Charging System	208
Why are Congestion Charges so Rare?	209
Public Support	209
Political Support	211
Summary and Lessons Learned	211
References	212
Further Reading	213

Rationale for Introducing Congestion Charging

Congestion is increasing in many cities around the globe, reducing accessibility and increasing air pollution. One way of addressing the congestion problem is to increase the supply of road capacity by increasing road infrastructure investment or traffic management. Since space is scarce in growing cities, and since new road capacity is expensive, it is usually not possible to reduce congestion only by increasing the supply of road capacity. Hence, to combat congestion, measures impacting demand are usually needed.

If the demand for road space is larger than the supply, it is a scarce resource. As all scarce resources it has a cost: time or money cost. That is, if there is congestion, travelers will pay the cost by queuing. If demand is managed by a congestion charge, the cost of congestion will instead be paid by a charge which means that it is collected and can be re-used. Queuing time cannot however be collected and re-used. Hence, it is more efficient to use congestion charges than queues to allocate scarce road space. When the congestion charges are optimal, we have efficient use of scarce resources. Essentially the most valuable traffic is prioritized, which often has few other options. This is often freight transport, business trips and commuting trips. Of course, parking charges, annual vehicle tax, and fuel tax might also to some extent reduce the demand for car use, and thereby congestion, while generating revenue that can be re-used. However, they are less efficient in reducing congestion since they are not directly targeted toward the bottlenecks causing congestion.

Demand can be managed also by regulations. A possible regulation is to restricting vehicle traffic using a system of odd or even number plates. However, essentially the same problem remains as with queueing: the cost for the drivers not allowed to drive a given day can again not be collected and re-used as a monetary cost and is, therefore, less efficient than congestion charges. Moreover, it is not the most valuable traffic that it prioritized as in the case of congestion charging. Regulations using a system of odd or even number plates can also increase car ownership as an unwanted side effect. For this reason, transport economists have for decades advocated congestion charges to combat congestion (Vickrey, 1963; Small, 1983; Arnott et al., 1994).

However, even if congestion charging is an efficient way of allocating scarce road space, it will not solve the accessibility problem in the cities. Congestion charges will increase the generalized travel costs for most drivers. In other words, it does not solve the problem of high generalized travel costs or accessibility problems. Moreover, congestion charges will often imply that large sums are redistributed compared to net efficiency gains for society. For these reasons, it hands over power over the resources from drivers to decision-makers. The total gain and redistribution effect of congestion charges will be determined by how decision-makers choose to spend or circulate the money.

Urban Congestion Charging Systems Around the World

Already in the 1980s, Hong Kong carried out a trial (1983–85) with electronic congestion charging, but this was abandoned for integrity reasons (Hau, 1990). Since then, many other cities have had plans of introducing congestion charging, e.g. Edinburgh (McQuaid and Grieco, 2005; Rye et al., 2008), New York (Schaller, 2010), and Copenhagen (Rich and Nielsen, 2007). In many cases, these plans have gone quite far with detailed charging designs developed for the city in question, but in the end, the plans have failed in last minute, usually due to low public or political support.

Thus, for a long time, Singapore was the only example of an implemented urban congestion charging system. Singapore introduced a form of congestion charging called the Area Licensing Scheme (ALS) already in 1975 (Phang and Toh, 2004). This was a manual system of tolls around a restricted area. Traffic decreased in the restricted area, but congestion just shifted time and

place. Therefore, the system has been adjusted over the years—the biggest change being from a manual to an electronic road pricing (ERP) system in 1998. This also meant that congestion charging was made more dynamic with charges depending on time-of-day, vehicle size, and route taken.

The second large implementation of an urban congestion charging system occurred in London in 2003 due to major public concern about traffic congestion in the city (Leape, 2006). The London system differs from that in Singapore (and later on Stockholm and Gothenburg) in that it is not based around charging stations, rather it is a daily charge for driving within central London between 7 AM and 6.30 PM on workdays. An advantage of this kind of system is that also driving within the charging zone is charged, but the disadvantage is that it requires a very large number of video cameras which render the system very expensive.

The Stockholm (Eliasson et al., 2009) and Gothenburg (Börjesson and Kristoffersson, 2015) systems followed London and were introduced in 2006 and 2013, respectively. Both are congestion charging cordons around the inner cities. The Stockholm and Gothenburg charges are more similar to Singapore than London, given that they are time-of-day dependent and based on charging stations. A time-based charge called the Controlled Vehicular Access System was implemented in Valletta, Malta, in 2007, replacing a fixed annual charge to enter the city center (Attard and Ison, 2010). The introduction of congestion charging in Valletta benefitted from the successful experiences in London and Stockholm, but also from lessons of the failure to implement congestion charging in Edinburgh (Attard and Enoch, 2011). Furthermore, in Milan, a congestion charging system called Area C is in operation. Area C replaced the previous Ecopass scheme in 2012 and was approved as a permanent scheme in 2013 after an initial trial period (Percoco, 2013).

The Singapore, London, Stockholm, Valetta, Milan, and Gothenburg cases described above are the main implementations of *urban* congestion charging in operation today, which are developed with the objective (at least one of the objectives) to reduce congestion. Therefore, the remainder of this chapter will draw from these examples when describing the effects of urban congestion charging and lessons learnt over the years.

Congestion charging systems also exist in the small towns Durham, UK and Znojmo, Czech Republic, but the literature on these examples is scarce. A number of cities in Norway have toll rings that originally were developed to collect revenues in order to finance transport projects (Larsen and Ostmo, 2001). Since 2017, the Oslo charge is differentiated depending on peak and off peak, light, and heavy vehicles, as well as fuel type (where notably electric vehicles are fully exempt from the charge).

Effects of Congestion Charging

Short-Term Effects

Before congestion charging was introduced in e.g. London, Stockholm, and Gothenburg, many people were sceptical to whether this theoretically appealing policy would work in reality. Maybe road users would not care about a slightly higher cost or maybe they would take a very long time to adjust to the system. However, these apprehensions turned out to be incorrect. In fact, road users adapted very quickly to the new system. For example, in Stockholm, there was a decrease in car trips over the cordon already during the first month of the congestion charging trial of 27% (during charging hours) (Eliasson et al., 2009). The decrease in traffic flow diminished somewhat over time and stabilized at 22% during the trial. Also in London, the flow decrease was high with a reported decrease of 26% for chargeable vehicle movements when comparing spring 2003 to spring 2002 (Transport for London, 2003).

These are large decreases compared to other measures introduced to try to change mode choice of travelers going into city centers, for example, lowering the public transport fares, or improving bicycle lanes. A fairly large decrease in car travel due to cheaper public transport is shown in Cats et al., 2017, who state that completely fare-free public transport in Tallin, Estonia decreased car trips by about 10% in Tallin municipality, but it came along with a decrease of walking trips by 40%. Buehler et al., 2017 argue that it is a mix of policies to encourage walking, bicycling and public transport together with policies that discourage car use (in this case mainly parking charges and limitations on new road construction) that can explain the decrease in car use in central Berlin, Hamburg, Munich, Vienna, and Zurich in the last decades. However, a comparison of the two Scandinavian cities Stockholm and Copenhagen indicates that it can be difficult to reduce car use with more public transport or more cycling (Bastian and Börjesson, 2018). The two cities are similar in size. The share of cycling is 14% in the municipality of Copenhagen, but only 2% in the municipality of Stockholm (LSE Cities, 2013). However, Stockholm has a public transport share of 44% compared to Copenhagen's 34%. So, the share of car trips is virtually the same. The modal shares of Stockholm are very close to those of London. Hence, these numbers suggest that large increases in public transport often reduces cycling and vice versa.

The largest positive surprise regarding the short-term effects of congestion charges was the effect on congestion and car travel times. For example, in Stockholm, congestion did not only reduce close to the congestion charging stations—car travel delay times decreased far out in the network, even for car traffic that did not pass a charging station. The reason for this was network effects—that bottlenecks resolving affected car traffic upstream and car traffic going crosswise. In Stockholm, car travel delay times reduced mostly on arterial roads—by one third in the morning peak and one half in the afternoon peak (Eliasson et al., 2009). The reduction in delay times also leads to more predictable car travel times since the day-to-day uncertainty reduced considerably. The effects on congestion in London were similar to that of Stockholm, with a reported figure from 2003 of 32% reduction in delay per kilometer (Transport for London, 2003).

The case of Gothenburg shows however that smaller reductions in traffic flow, congestion and associated car travel times are to be expected if initial congestion levels are low. In Gothenburg, traffic flow over the cordon decreased by 12% during charging hours

when comparing 2013 with 2012 and travel times only reduced substantially on the innermost arterial roads (Börjesson and Kristoffersson, 2015).

Congestion reduction not only leads to shorter and more reliable travel times, it also leads to improved air quality in the local environment and thereby health benefits. It has been estimated that the Stockholm Trial reduced air-borne pollution by 10%–14% in the inner city (Eliasson, 2014a). It should, however, be noted that it is difficult to measure air quality effects, since it is heavily dependent on weather. Since urban congestion charging works by reducing traffic locally in congested cities, its climate effects are minor. The reduction in carbon dioxide is approximately proportional to the reduction in vehicle kilometers traveled, which in the case of Stockholm was estimated to 2%–3% in the county (City of Stockholm, 2006). Even though improved air-quality was (and still is) one of the main objectives of the congestion charging scheme in Milan, Percoco, 2013 shows that the charge did not manage to reduce pollution, presumably due to motorbikes not being charged.

One notable characteristic of congestion charging is that it is not easy to track what happens to the car trips that adjust to the charge. A substantial share of car trips that are priced off (primarily commute trips) switch to public transport, but far from all. But especially discretionary trips find other ways of adjusting (City of Stockholm, 2006). For instance, the trip can be cancelled, two trips can be combined into one or a trip can be moved to another time-of-day or another day. Since travel behavior varies over time, in the long and short run, it is often difficult for car travelers themselves to recall how and if they adjusted their behavior in response to the charge. Since congestion charging prioritizes the most valuable traffic, it is the trips facing the highest cost of adjusting (having no or few attractive alternatives) that stays on the road.

Furthermore, travel patterns are not as repetitive as transport modelers and perhaps stakeholders tend to think. Data from Stockholm has shown that occasional drivers, passing the cordon three days a week or less, constitute two-thirds of the vehicles passing the cordon (Eliasson, 2014a). Thus, only a small share of car traffic over the cordon is commuting to work on all weekdays. On the one hand, this makes it difficult to follow up and report adjustments to congestion charging. On the other hand, this feature of congestion charging makes it an appealing policy measure since it leaves the decision of which trips to adjust and in what way, to the travelers themselves.

Long-Term Effects

Long-term effects of urban congestion charging are not extensively discussed in the literature but some sources exist (Börjesson et al., 2012; De Lara et al., 2013; Percoco, 2014; Börjesson and Kristoffersson, 2018). However, it is well known that fuel price elasticities tend to increase over time since drivers tend to have more and better options to adapt in the longer time perspective (Goodwin et al., 2004). Increasing price elasticity over time could also be the case for congestion charging: for instance if drivers can find it easier to adapt residential location, workplace location, or family situation over time. One could also argue that the effect would wear off as people get used to the charges, but there is little evidence that this is the case. From current implementation evidence it seems as though there is an initial overreaction to the charges, but that already after a couple of months the effect has stabilized and does not diminish from there on, rather it increases slightly.

The longer a city has had congestion charging implemented, the more difficult it gets to disentangle the effect of congestion charging from other effects such as other travel demand management strategies, city planning and population growth. The baseline from the year before the introduction of congestion charging becomes more and more irrelevant as the years pass by. Some studies have however tried to assess the long-term effects of urban congestion charging a couple of years after introduction. For example, Börjesson et al., 2012 show that traffic reductions from the charges increased somewhat over time when external effects are controlled for.

Over the years, changes are made to the implemented charging systems. For example, in Stockholm, the system was extended to also include the motorway close to the city centers, and peak hour charges were raised both in Stockholm and Gothenburg. Börjesson and Kristoffersson (Börjesson and Kristoffersson, 2018) show that the price elasticities were substantially lower in the case of these system updates compared to the introduction, probably because the most price-sensitive traffic had already adapted when the charges were first introduced. Also, evidence from London (Evans, 2008) indicates small demand effects in response to adjustments in pricing levels.

In Singapore, the charges have been in place for a very long time and are used somewhat differently from other cities. The main objective of Singapore charges is to manage congestion (Goh, 2002). Charges are therefore adjusted twice a year during Summer and Christmas holidays, such that speeds are in the ranges of 20–30 km/h on arterial roads and 45–65 km/h on motorways. This more dynamic way of using the charges as a congestion management tool would be difficult for example in Sweden where a decision by the parliament is needed for every update of the charging scheme.

Designing a Congestion Charging System

The previous section showed that introducing congestion charging can help to reduce congestion in city centers, which is associated with travel time savings, improved local environment and health benefits. Traffic and other effects of congestion charging should however not be taken for granted. Designing a good congestion charging system is a difficult task and it is easy to end up with unintended effects. One likely pitfall is that the charging system results in second-best problems, moving congestion around in time and place. Any route choices available that can be used to avoid paying the charge, will be used. The case of Stockholm is an example

of a simpler situation resulting in few second-best problem, since the city is placed on islands, and it is easy to set charging stations on the bridges surrounding the city center. Another example of a few second-best problems is Valletta, which is situated on a peninsula with a limited number of access roads from the south-west only. A third example is the congestion charge of the New York City approved by the state legislature, designed as a toll ring around Manhattan south of 60th Street. A fourth example is the congestion tolls implemented at the San Francisco–Oakland Bay Bridge in 2010.

Gothenburg and many other cities where bottlenecks are not primarily bridges, on the other hand, posed more challenges to designers in terms of route choices and more charging stations were needed to surround the city center. Also in Gothenburg, the design had to be adjusted over time in order to avoid unintended barrier effects in the north part of the city that caused many citizens to complain about having to pay charges for e.g. driving to their nearest shop.

A sophisticated transport model is important to decide where to place charging stations and which charging levels should be used. The network assignment model will directly show if there are any undesirable route choice effects. Furthermore, reasonable price elasticities in the model will help to decide on the amounts to charge. The most difficult part is typically the dynamic effects of congestion charging. Since traditional transport models are developed to evaluate infrastructure investments, they are often not very good at evaluating time-of-day varying charges and network effects of reduced queuing upstream when bottlenecks are resolved. This is especially important to notice if a cost–benefit analysis (CBA) is carried out, since travel time savings is at the core of the CBA. Travel time savings are often difficult to calculate correctly using a static transport model. Evidence from Sweden show that this is, of course, more important if network effects are large as in Stockholm (Eliasson et al., 2013) compared to when they are small as in Gothenburg (West et al., 2016).

The design of a congestion charging system also includes deciding if there should be any exemptions to the charges. It is important that exemptions are carefully analyzed together with the proposed scheme since they often have large effects on the efficiency and equity of the system. Examples of exemptions include reduction of charge for residents in central London and full exemption from the charge for taxis and alternative fuel vehicles in Stockholm during the first years of operation. In Sweden, the trend has been toward fewer and fewer exemptions. To negotiate exemptions and start and end times of the charges with local stakeholders can be a way of building support for the charges, which the case of Valletta shows (Attard and Ison, 2010).

In theory (see the first section), in an efficient system each vehicle pays for the congestion cost they impose on other vehicles at the specific time and place. Thus, charges should vary not only by time-of-day but also by location and whether one is driving into or out of the city center. However, in practice, an important factor is that the design of the system needs to be easily communicated. In fact, one of the suggested reasons for the “no” in the referendum on congestion charging in Edinburgh was that the system design was not simple enough to be easily communicated to the public (Gaunt et al., 2007). The final design of charging amounts at each time, place and driving direction needs to consider not only efficiency but also equity. Indeed analyzing car use among low income citizens is a very important factor in equity consideration. A look at the greater picture is therefore necessary, including also the use of the system revenues.

Why are Congestion Charges so Rare?

Public Support

One of the main obstacles for introducing congestion charging, even though there is substantial evidence that it works, is often public resistance. Public resistance is not surprising since most travelers and voters will lose from the charges (even if traffic associated with high values of time will gain). The welfare increase for society and individual travelers will depend on how the revenue is recycled, which is often difficult to see for travelers and voters. Hence, if the public does not trust the government to spend the money wisely, this reduces the public support further. Moreover, introducing congestion charges implies that a new tax base is introduced. Once it is introduced, it is easier for decision-makers to increase the charges and extend the system. Again, the public will likely be against congestion charges if they do not trust the government not to increase the charges to become “too high.”

However, public support can be built for congestion charges. Interestingly, both Stockholm and Gothenburg, and several other cities, have reported that the public support increased shortly after the congestion charges were introduced. Other examples include London (Schade and Baum, 2007), Stockholm (Eliasson, 2014b; Eliasson and Jonsson, 2011), Trondheim, Bergen and Oslo (Tretvik, 2003), and United States (Zmud, 2008) quoted in (Anas and Lindsey, 2011). Several explanations for this phenomenon have been hypothesized. Based on an extensive two-wave survey of public attitudes before/after the introduction of congestion charges in Gothenburg, Börjesson et al., 2016 estimated models where respondents’ attitudes to congestion charges are explained by variables such as expected toll payments, value of time, socioeconomic factors, beliefs about effects, and attitudes to related issues such as environment, equity, taxation, government, and pricing policies in general. The authors concluded that status quo bias is the main contributing factor to the increased support in Gothenburg. Contrary to what is often assumed, “larger benefits than expected” does not play any role for the change in support in Gothenburg: the respondents had the same belief in the effect in the before and in the after wave of the survey.

However, the long-term trends in public support in the two Swedish cities indicate that trust in decision-makers might be important to build and keep support for congestion charges over time. Fig. 1 shows how public support for the charges developed in Stockholm. In 2004, over 45% of the citizens of the city of Stockholm stated that they would vote in favor of congestion charges in a referendum. The support fell, however, as the introduction approached and just before the introduction of the charges it had fallen below 40%. Once the charges were introduced, the support increased again and in the referendum in the autumn of 2006, 53% of the

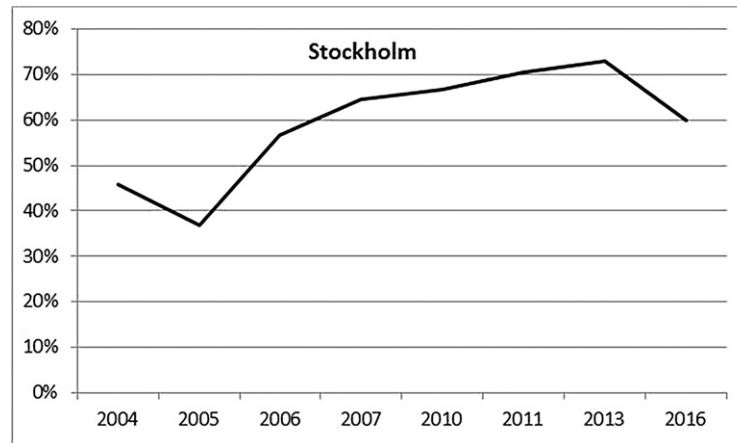


Figure 1 The share of respondents who stated that they would support the congestion charges in a referendum. The question is formulated as: “How would you vote in a referendum about the Stockholm congestion charges?”.

citizens of the city of Stockholm voted to keep the charges (excluding blank votes). Up until 2013, the support for the charges gradually increased to over 70%.

When the charges were increased and extended in 2016 the support for the charges dropped to 60%, which is almost down to the 2006 level. The decline in support could be caused by the extension of the system to an important bypass, the substantial charge increase, or the small effect it had on travel times (the effects are so small that the travelers did not notice them, except for some segments on the bypass). The decline in support could also be an effect of declining trust in the decision-makers: they communicated that the purpose of the congestion charge was to reduce congestion—whereas an important reason is to finance a big infrastructure package.

In Gothenburg, just over 30% supported congestion charges in 2011, and this support declined to 27% just before the charges were introduced (see Fig. 2). Just as in Stockholm, support increased after the introduction, but the referendum in September 2014 still resulted in 55% voting for abolishing the charges (the referendum was only consultative and was not implemented by the decision-makers because the revenue was necessary to fund a large infrastructure investment package (In 2007, Stockholm received a major transport road investment package, 50% funded by the revenues and 50% by the national government and the revenues were earmarked for a new bypass. It led the decision-makers in the Gothenburg region to seek a similar deal, resulting in a broad political coalition in the Gothenburg City Council to support the “West Swedish package”, partly funded by congestion charges (October 28, 2009). The largest investment in this package is the West Link (2.0 billion EUR), which is an 8-km-long rail link including a 6-km-long tunnel under central Gothenburg with BCR 0.45 (Mellin et al., 2011). The co-financing of the West Swedish package is the reason behind the increases in the charging levels in Gothenburg 2015 (it was not justified by congestion). In a poll in the autumn of 2014, just after the referendum, 51% stated that they supported the charges (It is unclear why the support in the referendum in autumn 2014 was lower than in the poll just after. One possibility is that 22% of the respondents in the poll that were undecided had

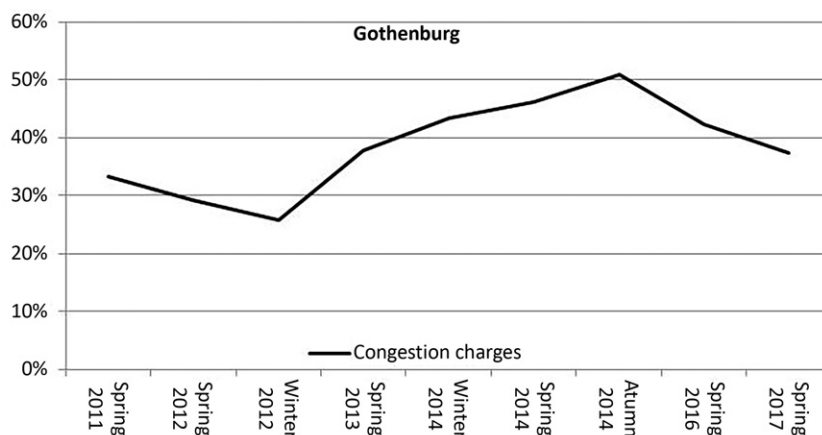


Figure 2 Share of respondents who state that they are positive or very positive to the congestion charges. The question is formulated as: “How positive or negative are you to congestion charges—part of the financing of the other parts of the package?”

a stronger tendency to vote against. Another possibility is that citizens that are more positive to the charges have a higher response rate in the poll.). After the increases in charging levels, support for the charges fell (as in Stockholm), and since then it has continued to decline.

Political Support

Many cities have failed to introduce congestion charges, for instance Edinburgh, Manchester, Helsinki, and Copenhagen. The lack of congestion charges is often attributed to low public support. However, the Swedish cases and for instance Edinburgh shows that political support might be just as important and that it is primarily determined by the power over the system and the revenues. Political support for the charging scheme was low in the case of Edinburgh and it was decided to hold a referendum about a complicated double-cordon charging scheme even though it was known already before the referendum that there existed a strong public opposition toward a double-cordon (Gaunt et al., 2007).

Before congestion charges were introduced in Stockholm in 2006, all the political parties (except the Green Party) were against them in Stockholm and Gothenburg. This has changed since the introduction of the Stockholm charges, to the extent that all established parties in Stockholm and Gothenburg are now in favor of the charges. The extension of the systems and the increased charging levels are supported by broad agreement in the city councils of both cities. This shift has been driven by the role that the revenues came to play in the negotiations for national grants for transport investments in the two cities. In Gothenburg, this is in spite of the low public support.

Summary and Lessons Learned

The currently operational charging systems in Singapore, London, Stockholm, Valletta, Milan, and Gothenburg show that congestion charges can be an efficient and effective policy measure for combating urban congestion. Moreover, ex-post cost-benefit analyses after the first-year show that congestion charges were socially beneficial in Stockholm as well as in Gothenburg and Milan (Eliasson, 2009; Rotaris et al., 2010; West and Börjesson, 2018). A key concern for policymakers in other cities/countries is obviously to what extent these effects are stable in the long term and whether they are transferable to their cities.

The finding that many of the effects experienced in the cities with operational charging systems are similar, with large differences in terms of city size, public transport use and car dependence, suggests that the positive traffic effects do not hinge upon some particular features of the city. This conclusion is also supported by a model-based study from Stockholm (Börjesson et al., 2014), using the transport model to forecast the effects of the Stockholm charges in scenarios where the road and public transport systems have been substantially modified. Hence, the traffic effects of congestion charges do in general seem to be transferable between cities with different characteristics. Having said that, the case of Gothenburg demonstrates that congestion charges are less efficient if initial congestion levels are low.

Based on the experiences from implemented urban congestion charging systems, the advice to policy-makers in cities planning to introduce congestion charging is to design the system with great care. Designing a system that effectively reduces congestion is more difficult than one could expect. There is a high risk of the second-best problem (i.e. moving congestion around in the network or causing new congestion). The design might, in general, be easier in cities like Stockholm and Valletta (as well as New York for instance) where bottlenecks are located on bridges or a few access roads, and where water is functioning as natural borders, and more difficult in cities like Gothenburg (and for instance Chicago) where the bottlenecks are more difficult to identify and isolate and where there are no natural borders preventing drivers to take undesirable route choices to avoid being charged. To design an efficient system that effectively reduces congestion, important policy advice is therefore to apply a carefully calibrated local transport model in the design process to test many designs and reject the bad ones. In the two Swedish cities, the best-practice transport model could predict the reduction in traffic volumes with high accuracy in the peak.

In several of the cities that have operational congestion charging systems, we have seen that public support increased shortly after the congestion charges were introduced. For this reason, having the referendum after a six-month trial proved to be crucial in Stockholm—had the referendum taken place just before the introduction of the system, a majority would have voted against introducing them (this has also happened for instance in Edinburgh). From a policy perspective, a trial is therefore favorable, and policymakers can study how public support develops before making the system permanent—but the problem is then to justify the investment of large resources in a system that might not survive after the trial. However, for cities like New York, where there are already tolls on the bridges that could be transformed to a congestion charging system, reducing the system investment cost substantially, a trial might be a good idea.

Even if congestion charges are very efficient, they should be used with caution. The net social benefit is often small in relation to the money that is redistributed. A particular problem with the large sums that are redistributed is that congestion charges often is a regressive tax instrument, for example in Gothenburg (West and Börjesson, 2018). This is because the level of car-dependence among low-income citizens is very important for the resulting equity effects. The policy lesson here is that in cities with high car dependence also among low-income groups, the spending of revenue to compensate low-income citizens is critical (for instance, the London system spent a considerable share of the revenues on additional buses in the local public transport service).

A current problem in some of the cities with operational charging systems is that the political focus and justification for congestion charges has increasingly shifted toward the financing of large (prestigious) investments with a low value for money,

that will not benefit low-income groups (or anyone in the coming 10–15 years). This might indeed be one reason why decision-makers in Gothenburg have not been able to build long-term stable public support for the charges there. Hence, in order to build long-term stable support and reduce problems of negative distribution effects, the money should be spent with great care. The Stockholm and Gothenburg cases indicate that the spending of the revenue on mega-investments does not necessarily help build support for congestion charges.

The decline in public support in Stockholm and Gothenburg after system revisions shows that public support cannot be taken for granted even after the introduction of congestion charges. The advice to policymakers is to keep the system as stable as possible over time. If the charges are gradually increased or extended, without large effects on the traffic, there is an increasing risk that drivers will find it more difficult to trust the government not to increase and extend the system only to increase revenue. To avoid such development, it is probably better if any increases in the charging levels are justified by increasing congestion levels, for example, than the need to collect revenue.



Ida Kristoffersson, Senior Research Leader in the field modelling and analysis of passenger transport at VTI Swedish National Road and Transport Research Institute. Her main research area is in development of travel demand models to meet the needs of evaluation of new policies and innovations given the increased attention to sustainability and digitalization of the transport sector. Research interests include transport demand modelling, transport pricing, cost-benefit analysis, and travel behavior.



Maria Börjesson, Professor in Economics at VTI Swedish National Road and Transport Research Institute and permanently affiliated professor at KTH Royal Institute of Technology. External affiliated to the choice modelling center, ITS Leeds. Research interests include transport cost-benefit analysis and appraisal, transport pricing, the impact of the transport system on employment and GDP, choice experiments and econometrics for valuation of non-market goods such as travel time, reliability, and security.

References

- Anas, A., Lindsey, R., 2011. Reducing urban road transportation externalities: road pricing in theory and in practice. *Rev. Environ. Econ. Policy* 19, doi:10.1093/reep/req019.
- Arnott, R., De Palma, A., Lindsey, R., 1994. The welfare effects of congestion tolls with heterogeneous commuters. *J. Transport Econ. Policy* 28, 139–161.
- Attard, M., Enoch, M., 2011. Policy transfer and the introduction of road pricing in Valletta, Malta. *Transport Policy* 18, 544–553.
- Attard, M., Ison, S.G., 2010. The implementation of road user charging and the lessons learnt: the case of Valletta, Malta. *J. Transport Geogr.* 18, 14–22.
- Bastian, A., Börjesson, M., 2018. The city as a driver of new mobility patterns, cycling and gender equality: travel behaviour trends in Stockholm 1985–2015. *Travel Behav. Soc.* 13, 71–87, <https://doi.org/10.1016/j.tbs.2018.06.003>.
- Börjesson, M., Kristoffersson, I., 2015. The Gothenburg congestion charge. Effects, design and politics. *Transport. Res. A: Policy Pract.* 75, 134–146.
- Börjesson, M., Kristoffersson, I., 2018. The Swedish congestion charges: ten years on. *Transport. Res. A* 107, 35–51.
- Börjesson, Eliasson, J., Beser-Hugosson, M., Brundell-Freij, K., 2012. The Stockholm congestion charges—5 years on. Effects, acceptability and lessons learnt. *Transport Policy* 20, 1–12.
- Börjesson, M., Brundell-Freij, K., Eliasson, J., 2014. Not invented here: transferability of congestion charges effects. *Transport Policy* 36, 263–271.
- Börjesson, M., Eliasson, J., Hamilton, C., 2016. Why experience changes attitudes to congestion pricing: the case of Gothenburg. *Transport. Res. A: Policy Pract.* 85, 1–16.
- Buehler, R., Pucher, J., Gerike, R., Götschi, T., 2017. Reducing car dependence in the heart of Europe: lessons from Germany, Austria, and Switzerland. *Transport Rev.* 37, 4–28.
- Cats, O., Susilo, Y.O., Reimal, T., 2017. The prospects of fare-free public transport: evidence from Tallinn. *Transportation* 44, 1083–1104.
- LSE Cities, 2013. Stockholm: Green Economy Leader Report. Produced by the Economics of Green Cities Programme at the London School of Economics and Political Science in partnership with the City of Stockholm.
- City of Stockholm, 2006. Facts and results from the Stockholm Trials.
- De Lara, M., De Palma, A., Kilani, M., Piperno, S., 2013. Congestion pricing and long term urban form: application to Paris region. *Regional Sci. Urban Econ.* 43, 282–295.
- Eliasson, J., 2009. A cost-benefit analysis of the Stockholm congestion charging system. *Transport. Res. A* 43, 468–480.
- Eliasson, J., 2014a. The Stockholm Congestion Charges: An Overview. Centre for Transport Studies, Stockholm 42, CTS Working Paper 7.
- Eliasson, J., 2014b. The role of attitude structures, direct experience and framing for successful congestion pricing. *Transport. Res. A* 67, 81–95.

- Eliasson, J., Jonsson, L., 2011. The unexpected "yes": explanatory factors behind the positive attitudes to congestion charges in Stockholm. *Transport Policy* 18, 636–647, doi:10.1016/j.tranpol.2011.03.006.
- Eliasson, J., Hultkrantz, L., Nerhagen, L., Rosqvist, L., 2009. The Stockholm congestion-charging trial 2006: overview of effects. *Transport. Res. A* 43, 240–250.
- Eliasson, J., Börjesson, M., van Amelsfort, D., Brundell-Freij, K., Engelson, L., 2013. Accuracy of congestion pricing forecasts. *Transport. Res. A: Policy Pract.* 52, 34–46.
- Evans, R., 2008. Demand elasticities for car trips to central London as revealed by the Central London Congestion Charge. *Transport For London Policy Analysis Division*.
- Gaunt, M., Rye, T., Allen, S., 2007. Public acceptability of road user charging: the case of Edinburgh and the 2005 referendum. *Transport Rev.* 27, 85–102.
- Goh, M., 2002. Congestion management and electronic road pricing in Singapore. *J. Transport Geogr.* 10, 29–38.
- Goodwin, P., Dargay, J., Hanly, M., 2004. Elasticities of road traffic and fuel consumption with respect to price and income: a review. *Transport Rev.* 24, 275–292.
- Hau, T.D., 1990. Electronic road pricing: developments in Hong Kong 1983–1989. *J. Transport Econ. Policy* 24, 203–214.
- Larsen, O., Ostmo, K., 2001. The experience of urban toll cordons in Norway: lessons for the future. *J. Transport Econ. Policy (JTEP)* 35, 457–471.
- Leape, J., 2006. The London congestion charge. *J. Econ. Perspect.* 20, 157–176.
- McQuaid, R., Grieco, M., 2005. Edinburgh and the politics of congestion charging: negotiating road user charging with affected publics. *Transport Policy* 12, 475–476.
- Mellin, A., Nilsson, J.-E., Pyddoke, R., 2011. Underlagsrapport till RiR 2011:28: Medfinansiering av statlig infrastruktur.
- Percoco, M., 2013. Is road pricing effective in abating pollution? Evidence from Milan. *Transport. Res. D: Transport Environ.* 25, 112–118.
- Percoco, M., 2014. The impact of road pricing on housing prices: preliminary evidence from Milan. *Transport. Res. A: Policy Pract.* 67, 188–194.
- Phang, S.-Y., Toh, R.S., 2004. Road congestion pricing in Singapore: 1975 to 2003. *Transport. J.* 16–25.
- Rich, J., Nielsen, O.A., 2007. A socio-economic assessment of proposed road user charging schemes in Copenhagen. *Transport Policy* 14, 330–345.
- Rotaris, L., Danielis, R., Marcucci, E., Massiani, J., 2010. The urban road pricing scheme to curb pollution in Milan, Italy: description, impacts and preliminary cost–benefit analysis assessment. *Transport. Res. A: Policy Pract.* 44, 359–375.
- Rye, T., Gaunt, M., Ison, S., 2008. Edinburgh's congestion charging plans: an analysis of reasons for non-implementation. *Transport. Plann. Technol.* 31, 641–661.
- Schade, J., Baum, M., 2007. Reactance or acceptance? Reactions towards the introduction of road pricing. *Transport. Res. A: Policy Pract.* 41, 41–48, doi:10.1016/j.tra.2006.05.008.
- Schaller, B., 2010. New York City's congestion pricing experience and implications for road pricing acceptance in the United States. *Transport Policy* 17, 266–273.
- Small, K., 1983. The incidence of congestion tolls on urban highways. *J. Urban Econ.* 13, 90–111.
- Transport for London, 2003. Congestion charging six months on.
- Tretvik, T., 2003. Urban road pricing in Norway: public acceptability and travel behaviour. Presented at the MC-ICAM Conference, Acceptability of Transport Pricing Strategies.
- Vickrey, W., 1963. Pricing in urban and suburban transport. *Am. Econ. Rev.* 53, 452–465.
- West, J., Börjesson, M., Engelson, L., 2016. Accuracy of the Gothenburg congestion charges forecast. *Transport. Res. A: Policy Pract.* 94, 266–277.
- West, J., Börjesson, M., 2018. The Gothenburg congestion charges: cost–benefit analysis and distribution effects. *Transportation* 1–30.
- Zmud, J., 2008. The public supports pricing if. A synthesis of public opinion studies on tolling and road pricing. *Tollways* 29–39.

Further Reading

- Bastian, A., Börjesson, M., 2005. The city as a driver of new mobility patterns, cycling and gender equality: travel behaviour trends in Stockholm 1985–2015 CTS working paper 2017:19.
- Ison, S., Rye, T., 2005. Implementing road user charging: the lessons learnt from Hong Kong, Cambridge and Central London. *Transport Rev.* 25, 451–465.
- Levinson, D., 2010. Equity effects of road pricing: a review. *Transport Rev.* 30, 33–57.
- Santos, G., 2005. Urban congestion charging: a comparison between London and Singapore. *Transport Rev.* 25, 511–534.
- Walker, J., 2018. Road Pricing: Technologies, Economics and Acceptability. IET

Energy and Transport Planning

Debbie Hopkins*, **Christian Brand†**, *Transport Studies Unit, School of Geography and the Environment, University of Oxford, Oxford, United Kingdom; †Transport Studies Unit and Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	214
Transport and Energy	214
Transport Planning and Reducing Transport Energy Demand	215
Beyond the “Minority World”	217
Conclusions	218
Acknowledgment	218
References	218

Introduction

Planning for future transport systems has always been—and probably always will be—a complex endeavor. It requires a certain degree of foresight into configurations of technologies, practices, norms and values which may emerge, and require different infrastructures, policies, regulations, and urban forms from those which are in place today. Effective planning and decision-making also requires understanding and appreciation of what does and does not work, and what the main uncertainties are. At the same time, existing physical, social and regulatory infrastructures, educational regimes and so on can lead to the replication of historical “best [planning] practice.” And unforeseen events—such as COVID-19—have the potential to radically shift what previously appeared to be immovable mobility practices (e.g., frequent air travel or car dependency).

The process of transport planning is not passive, but it actively participates in the construction of futures, which may be more or less equitable and sustainable. For example, through their everyday work and micro-decisions, planners and consultants assemble particular futures, which may lock-in or reject the status quo (Lissandrello et al., 2017; Pollock and Williams, 2010). It has, arguably, never been more important to recognize the role that transport planning plays in reinforcing, or rejecting high-carbon, energy-dependent transport practices. As transport has rightly been labeled the “roadblock to climate change mitigation” (Creutzig et al., 2015), questions of energy efficiency, energy inequality, the uptake and use of alternative fuels, and emissions reductions have come to the fore (Oswald et al., 2020). When combined with the inertia in the system infrastructure, vehicle fleet turnover and consumer behavior, the transport systems in developed economies are likely to look very similar in most respects in 2050 as they do today (Anable and Brand, 2019). This chapter focuses on road-based transport, as the major source of transport-related carbon emissions (Sims et al., 2014), and the largest absolute increase in energy consumption in the EU (European Environment Agency, 2019) whilst recognizing the relationships between different spaces: road, rail, water, and air.

Transport and Energy

Processes of movement require some form of energy. Roughly one-third of our energy goes into transportation. Active modes of transport (e.g., walking), which dominated social practice before the emergence and diffusion of the Internal Combustion Engine (ICE) and private motorized mobility, require bodily, “metabolic” energy. To supplement human bodily capabilities, a reliance on animals’ energy systems emerged, by way of horse riding, and horse and carts or coaches—affording a range of new possibilities to those able to use these modes (e.g., carrying heavier loads, traveling further, faster). Scholars from health and physical education have long been interested in the metabolic costs of “moving about” (e.g., Tucker, 1975), pointing to the in/efficiencies of different modes. Such logics of “efficiency” problematically pervade contemporary thinking on transport too, often contributing to a (re) enforcement of prioritizing “the most efficient” mode, which is often understood in terms of reducing bodily efforts, and increasing speed.

The widespread diffusion of ICEs shifted reliance from (animal/human) bodily metabolisms to fossil fuels. Fossil fuels grew in importance as the infrastructures (e.g., petrol stations) became established, know-how and competencies for the ICE grew (e.g., garage repair and maintenance), but perhaps more importantly, the powerful vested interests exerted dominance to prevent alternative fuel types and systems of mobility emerging. Fossil fuels were historically—and arguably remain—relatively abundant and accessible, and provide energy for mobility with high energy density, low weight, and small size of onboard energy storage necessary for mass transportation. Low costs and ease of transportation also contributed to the development of global chains of fossil fuel extraction to consumption. Just 12-30% of the energy from fuels put into conventional vehicles is used to move the vehicle (Department of Energy, ND). This figure depends on characteristics including, for instance, the drive cycle, driving style, and/ or

topography. The remaining 70–88% is lost to inefficiencies in the engine and driveline, and/or used to power in-vehicle accessories. Associated with these chains are well-documented exploitations and violence (Nixon, 2011; this is also the case for some non-fossil fuels, see for example: Branch and Martiniello, 2018). Moreover, and perhaps more often discussed, is the harm associated with the burning of fossil fuels, to both the environment and human health. The burning of fossil fuels is the principle source of carbon emissions (Höök and Tang, 2013). Between 30% and 50% of particulate matter (PM) and nitrogen oxides (NO_x)—the primary air pollutants affecting human health—in cities can be attributed to transport (European Commission, 2019), and diesel fuels in particular (WHO, ND), resulting in increased risk of cardiovascular and respiratory disease and cancer.

In Europe, road transport accounts for the largest proportion of energy consumption in the transport sector (73% in 2017), and in 2017 road transport energy consumption was 34% higher than in 1990 (European Environment Agency, 2019). The transport sector remains heavily reliant on fossil fuels. It is thought that globally, approximately 80% of transport fuels are derived from petroleum, and in the European Union in 2016, 95% of all fuels consumed were oil-derived (all fuels excluding biodiesel, biogas, biogasoline, electrical energy, natural gas, and solid biofuels) (European Environment Agency, 2019). For road transport, alternatives to fossil fuels include biofuels (e.g., bioethanol and biodiesel, either on their own or blended with fossil fuels), synthetic fuels (e.g., Fischer-Tropsch diesel), and electricity. Where the source of the electricity is renewable (e.g., solar, hydro, wind) there are significant benefits for an electro-mobility future, yet alone it is unlikely to achieve emission reductions within the timeframe required to prevent climate breakdown (Anable and Goodwin, 2019). This is because fuel type—and the pollutant content—is just one factor influencing transport energy consumption with other factors including technical efficiencies, modal choice, activity levels, lifestyles, socio-cultural factors, and population composition (Brand et al., 2019). Changing energy vectors and vehicle fleets also takes time; typically less than 10% of the car fleet is replaced each year, which implies a total transformation takes at least 12–15 years (Brand et al., 2020). This highlights the need to be thinking beyond energy supply, to pay closer attention to energy demand. In addition to this, questions have been asked of the sustainable resource use associated with electric battery production (see: Sovacool et al., 2020), with García-Olivares et al. (2018) showing that while a 100% renewable transport system may be feasible, it might not be optimal, as their modeling suggested it might not be compatible with broader sustainability objectives (e.g., decreased resources consumption).

The International Energy Agency (IEA) has noted that the transport sector is at a point of “critical transition,” whereby:

“existing measures to increase efficiency and reduce energy demand must be deepened and extended for compliance with the Sustainable Development Scenario (SDS). This process should be set in motion over the next decade, as any delay would require that stricter measures be taken beyond 2030, which could noticeably raise the cost of reaching climate targets. Combined efforts across all transport modes, accompanied by power sector decarbonisation, will play a crucial role for achieving SDS goals” (IEA, 2019: NP)

The Sustainable Development Scenario (SDS) is framed as an “integrated approach to energy and sustainable development,” outlining pathways to a major transformation of the global energy system, focused particularly on the three Sustainable Development Goals (SDGs) most closely related to energy: universal access to energy (SDG 7), health impacts of air pollution (part of SDG 3), and climate change (SDG 13). The SDS approach signals system-wide change, and points to the need for cooperation across different transport sectors, where, for instance, focusing on a single transport mode may disguise related trends occurring elsewhere in the transport system. An example of this can be seen in the work of Ottelin et al. (2014) who suggest that energy and emissions saved from reductions in car-based transport in urban areas might be replaced by air travel, potentially the result of a rebound effect (i.e., money saved from not owning a car is spent on air travel). Thus the “critical transition” of which the IEA speak acknowledges the need to go “further, faster” (CREDS, 2019), while also accounting for related sustainability objectives, and the connections between the transport system and other socio-economic sectors. Research on Just Transitions has gone some way to highlight the complex intersections of local fossil fuel extraction, economic systems, and decarbonization efforts (e.g., Healy and Barry, 2017), but multiple other connections exist with more or less visibility in academic and/or policy discourse.

Transport Planning and Reducing Transport Energy Demand

As noted earlier, there is a significant challenge to reducing the energy required for the current, and potential future transport demand. Planning and planners have the potential, through their daily practice; to lock in a private-car based transport future through prioritization of a built environment that perpetuates current practices, norms and values (e.g., logics of efficiency and speed, or use of cost-benefit analysis that favors travel time savings). The role of planners in low-carbon transitions has been discussed, with Cass et al. (2018: 166) noting how for planners’ “previous histories, policies, and other co-existing systems and networks affect the processes they hope to influence. Since planners’ actions are defined by these situations, following generic blueprints or transferable recipes for intervention is inherently problematic.” They point to the need for interventions, which pay attention to unique and situation-specific characteristics, which will be defined by historical and existing infrastructures, as well as “other interests and actors, both living and dead” (ibid.).

Thus, transport planning and transport planners have an important role to play in efforts to reduce transport-related emissions and energy demand. In characterizing what this role might be and what solutions will be optimal, research to date has shown that:

- technological efficiency will go some way to reducing emissions and energy consumption, but a wider suite of actions will be required to tackle the unsustainability of the current systems of transportation around the world (Anable and Goodwin, 2019)—thus waiting for technological solutions to mainstream (e.g., technological substitution such as EVs, CAVs) is misguided;
- Solutions need to include both passenger and freight/waste transport, and account for the multiple and complex relationships between transport and a range of social practices and economic systems;
- Solutions also need to account for the important interconnections between transport infrastructures (hard and soft) and practices;
- Close coordination between different sectors, and a range of actor groups is required for the development of a roadmap for decarbonizing and reducing the energy demand of the transport sector.

Scholars have pointed to a variety of ways that transport energy demand reductions, and decarbonization efforts can be accomplished, with important roles for transport and urban planning. This work has broadly agreed that four strategies need to occur to achieve sustainable mobility (e.g., Banister, 2008; Sims et al., 2014). These strategies call for a systemic, socio-technical approach to transport policy and planning. Within each strategy sits a suite of potential behavioral, policy/regulatory and technical responses, and as we go on to show, these solutions are highly connected, where for instance online shopping may reduce the need to travel, but then generates mobility for freight which could then be consolidated increasing the load factor of vehicles, but shifts consumption practices potentially increasing demand. The four strategies are identified as:

1. Reducing the need to travel (i.e., through travel substitution, combining trips, home working)
2. Reducing the distance of required travel (i.e., through land use planning measures, jobs and opportunities closer to home, etc.)
3. Shifting to low-carbon/ low-energy modes (i.e., through high quality infrastructures, promotions, personalized travel planning, etc.)
4. Increasing the efficiency of motorized mobility (i.e., through increased occupancy/load factors, vehicle design, alternative fuels, EVs, lower motorway speed limits, etc.)

People travel for a range of reasons; to get to work or education, for shopping, for leisure activities, to visit friends and relatives, or just for the sake of travel (and resultant emotions—happiness, excitement). Moreover, if we include both the movement of people and of things (i.e., freight, waste), then consideration of a broader suite of consumption practices come to the fore. This travel might be taken by a variety of transport modes; active (e.g., walking, cycling, e-biking, scooting), public (bus, train, coach), private (car, van), or a combination of these. Each mode results in different energy consumption and emissions based on the type of vehicle, occupancy rates, fuel types, distance traveled, and so on. Reducing the need to travel can—it is suggested—completely eliminate a portion of trips, and thus do away with the associated emissions and energy consumption (Brand et al., 2020). To do this requires mechanisms that fulfill the needs previously provided by corporeal mobility.

An obvious example of **travel substitution** is online shopping, where home delivery replaces the need for individual customers to travel to shops. From a passenger perspective, travel decreases, yet from the freight perspective it increases—pointing to the relational mobilities between passenger and freight. Consolidation of deliveries (ensuring a full delivery van) offers the greatest sustainability potential (Jaller and Pahwa, 2020)—and increases the “efficiency” of the transport, yet high rates of product returns and non-delivery (i.e., when the purchaser is not home to receive the parcel) contribute to an increased number of trips associated with purchases. Moreover, online shopping increases exposure to “consumerism and materialistic messages” (Guillen-Royo, 2019), with the potential to accelerate purchasing behaviors and shift social practices. For the freight sector, practices of dematerialization (e.g., e-readers replacing books and Spotify replacing CDs) have the potential to reduce the physical movement of material items, replacing it with digital transactions. Similarly, the emergence of 3D printing has the potential to reconfigure logistical networks. 3D printing could radically alter the construction sector, for instance, with the need for complex logistical patterns of mobility reduced, along with volumes of waste. Instead, materials would be “printed” on-site presenting an opportunity to rethink production, distribution, and consumption practices (Birtchnell and Urry, 2016), and potentially disrupting current transport systems and associated energy systems.

Digital technologies have a variety of potential further application, which may **reduce the need for mobility**. So-called “virtual mobilities” operate to digitize communication using information and communication technologies, including social media (e.g., Twitter, Facebook, TikTok), and communication apps (e.g., WhatsApp, WeChat, Telegram). It has been claimed that these technologies allow for communication, which may reduce the (perceived) need for physical co-presence. Yet research has shown that ICT is used for different types of communications than those, which take place face-to-face (Hopkins et al., 2019). Thus, assuming that technologies alone will facilitate substitution behaviors may be misguided. Instead, it is likely that conditions which allow for reduced travel need to be constructed—this might include flexible working policies, which enable some people to work from home thus reducing commuting mileage. It is important to note, however, that reduced daily mobility may be replaced by increased long-distance travel, resulting from rebound effects (Ottelin et al., 2014) or geographically stretched social networks (Urry, 2003). It is also important to acknowledge that digital technologies have often large energy demands and associated emissions (e.g., through data centers), and these need to be taken into account when digital platforms are used as substitutes to physical mobility.

In efforts to **reduce the distance of travel**, better linkages between transport planning and urban planning is required. This may contribute to more compact cities; a move which has the potential to both reduce the need for (motorized) travel, as well as the distance of travel—thus increasing the likelihood of shifting to low-carbon/low-energy transport modes. The per capita energy

consumption for transport in Houston, TX (urban sprawl) is a factor of about 10 higher than in a compact, multimodal city such as Copenhagen. It is important to highlight the difference here between calls for compact versus dense cities. A compact city is not necessarily dense, but offers relatively close proximity to the services that one might require in everyday life—shops, doctors, schools, and so on. Proximity is not the only measure of accessibility, but a sensitively designed compact environment may present some opportunities to reduce the need for or distance of travel for some members of the population. Moreover, cities are already responding to challenges including urban air pollution by way of developing schemes, which restrict—or ban—particular types of vehicles within particular portions of the city. These low-emission zones and congestion charges can provide a lever to promote alternatives to the private vehicle, particularly where investment in public transport provision and infrastructure for active modes occurs in tandem. Thus, urban planning, and the design of more compact environments may help to create the conditions in which increased public transport, walking and cycling can be adopted (Hickman et al., 2013), thereby decreasing energy consumption and harmful emissions.

Infrastructures—many of which are both carbon inducing (e.g., new roading projects) and hold embedded carbon—have the potential to lock-in high-carbon practices for future generations. Yet this does not need to be the case. There are examples where infrastructural lock-in has been overcome. For instance, where urban roading is re-purposed into urban parkways, where lanes of traffic are redesigned into separated cycleways. Creative use of existing infrastructure—even temporarily—can disrupt the dominance of high-carbon mobility in important ways. A prime example is the so-called Barcelona “superblocks” model (similar to Seattle’s “Home Zones”). Superblocks are “repurposed” neighborhoods of nine blocks where traffic is restricted to major roads around the outside, opening up entire groups of streets to residents, pedestrians and cyclists. It has been shown that Barcelona could save hundreds of lives and cut air pollution by a quarter if it fully implements its radical superblocks scheme to reduce traffic (Mueller et al., 2020). These examples are important because they show that change can—and does—happen. This is not intended to encourage the development of further roading infrastructure at the expense of investment in public and active modes, but just to show that the cycle can be reversed, infrastructural lock-in need not be viewed as a fixed barrier to decarbonizing transport or as justification for fatalism (“what can we do, we’re tied into car-based transport?”). There are examples of cities where decision makers and planners have achieved a transition from car dependence to pedestrian/cycling friendly urban form without radical changes to the built environment in the short term.

Current uncertainties of future transport systems are compounded by a variety of socio-technical innovations which are unsettling the existing regime, but the long-term implications of which are still broadly unknown. For instance, connected and autonomous vehicles (CAVs) are promised to respond to a variety of issues of the current transport system, including pollution and inequality; the growth in electric two-wheelers including e-bikes, e-cargo bikes, and e-scooters may extend active modes to a wider proportion of the population; mobility-as-a-service has the potential to reshape ownership models and increase sharing practices. Yet all of these (and other) innovations—which have the potential to reduce energy demand—are contentious and impacts are uncertain, with potential rebound effects (Sorrell et al., 2020), which may increase rather than decrease overall energy demand.

Beyond the “Minority World”

Without intending to repeat well-trodden discussions of the importance of acknowledging the heterogeneity of countries and cities around the world, it is also important to note the limitations to existing work on transport and energy, with much of the discourse and academic scholarship centered on a limited set of geographical regions for case studies and theorization.^a There are—of course—exceptions to this, with rising attention paid to China’s transport system and responses to issues associated with increased car ownership (Zhang et al., 2017; Yang et al., 2017). To here in this chapter, much (although not all) of the examples and case studies provided relate to the United Kingdom and northern Europe, where the authors are based. The narrative outside of Euro-America can work to homogenize transport systems and mobility practices, and understand these largely in relation to automobility (a private car dominant system of mobility). There are important reasons to ensure thinking on transport energy demand is not restricted to scholars and case studies of the Global North (also known as “Minority World” to reflect the relative current and future population sizes, relative to the “Majority World”). The dominance of Euro-American theorization and case studies across transport has been reported (e.g., Schwanen, 2018). A growing issue for transport planning is thinking through the implications of energy demand reductions in and for contexts beyond the existing cases.

Alternative issues arise in different geographies, for instance, Adom et al. (2018) point to the issue of constrained energy supply, demographic trends, economic growth, poor and unsafe infrastructures, and rapid urbanization in some African countries, as particular challenges for planning transport systems. The relationship between transport and economic growth is a complex one, which is characterized by a politics of developmentalism. Wang et al. (2019), for instance, show how urbanization, road infrastructure, economic growth, and industrialization are stimulating road transport energy consumption in Pakistan. In their analysis of MENA (Middle East and North Africa) countries, Saidi et al. (2018) found that energy consumption for transport works to stimulate economic activities and economic growth, but reflect on the need to account for the negative externalities of increased motorized mobility, pointing to the need for thoughtful infrastructure development, regulation and policy to prevent unrestrained increases in the physical infrastructures (with their own embedded energy and carbon) and associated automobility. In their

^a While Euro-America has dominated transport scholarship, other locales also have traction for specific reasons, for instance, South-East Asia for its application of ICT to leapfrog some dimensions of private car dependence, and South America for Bus Rapid Transit (BRT) systems.

investigation of West African countries, Adom et al. (2018: 919) found that rebound effects limited the energy conservation potential of pricing mechanisms, and concluded that “in the long-run, price disincentive tools have to be higher than a required price threshold in order to induce energy conservation behaviours in these economies.”

With all of this, there are important questions to ask of the multiple ways that colonial and postcolonial transport and urban planning have—and continue to—“shape the urban built environment, planning laws and institutional processes” (de Satgé and Watson, 2018: 31). There is, therefore, much work to do to offer more nuanced and sensitive assessments of transport energy demand and transport planning beyond Euro-America, and to understand this in connection to a broader suite of concerns. This may include seeking to integrate sustainable development scenarios to examine the complex intersections between energy and sustainable development as it plays out in a variety of contexts, cultures, urban forms and regulatory arrangements.

Conclusions

Mobility has long been viewed as integral and highly beneficial to social and economic life. Yet, as we write this (March 2020), novel Coronavirus COVID-19, still in its infancy, has already radically changed perspectives on the types of mobility that are perceived to be “necessary” and/or “essential.” The very concept of “necessary transport” has, for some, been upended. Suddenly, the previously immovable work trip is being conducted virtually. Some trips to visit family and friends have become “non-essential”. These events, which are hard to predict and plan for, offer the potential to reconfigure the importance placed on corporeal mobility, with early suggestions that COVID-19 could have substantial and significant implications for global aviation, but it is also likely that it will have an effect on more localized travel, including most forms of sustainable public transport and shared mobility. Working from home, where possible, is already one of the UK government’s recommendations while COVID-19 is in “containment” stage. Home delivery of supermarket goods and take away food are being recommended for those with symptoms—shifting the energy consumption from passenger to freight vehicles. Another potential implication could involve increased travel by private car, as people seek to avoid contact with others and thereby avoid shared (public) transport modes.

There are a host of issues associated with the dominant transport system, which is dependent on private mobility and fossil fuels. The technological shift to electrification is likely to be one part of the answer to the “net zero” clean energy challenge, but it is by no means the whole answer. To radically, systemically and quickly reduce the energy intensity and carbon dependency of the transport system will require major shifts in mindsets, ideas, norms, and practices as well as infrastructures, policies and regulations. Reducing energy demand therefore requires the support and coordination of multiple actor groups—including transport and urban planners. It requires out-of-the-box thinking, which challenges the status quo, and looks to alternative ways to configure transport systems which accounts for energy, but also pays attention to interconnected issues of access, equity, and affordability for diverse populations and geographies.

Acknowledgment

Both authors are partially funded by UK Research and Innovation, Grant agreement number EP/R035288/1 - Centre for Research on Energy Demand Solutions (CREDS).

References

- Adom, P.K., Barnor, C., Agradi, M.P., 2018. Road transport energy demand in West Africa: a test of the consumer-tolerable price hypothesis. *Int. J. Sust. Energy* 37 (10), 919–940.
- Anable, J., Brand, C., 2019. Energy, pollution and climate change. In: Docherty, I., Shaw, J. (Eds.), *Transport Matters*. Bristol Policy Press, p. pp. 452.
- Anable, J., Goodwin, P., 2019. Transport and mobility. In: Eyre, N., Killip, G. (Eds.), *Shifting the Focus: Energy Demand in a Net-Zero Carbon UK*. Centre for Research into Energy Demand Solutions, Oxford UK.
- Banister, D., 2008. The sustainable mobility paradigm. *Transp. Policy* 15, 73–80.
- Birtchnell, T., Urry, J., 2016. *A New Industrial Future? 3D Printing and the Reconfiguring of Production, Distribution and Consumption*. Routledge, Abingdon.
- Branch, A., Martiniello, G., 2018. Charcoal power: The political violence of non-fossil fuel in Uganda. *Geoforum* 97, 242–252.
- Brand, C., Anable, J., Morton, C., 2019. Lifestyle, efficiency and limits: Modelling transport energy and emissions using a socio-technical approach. *Energy Efficiency* 12, 187–207.
- Brand, C., Anable, J., Ketsopoulou, I., Watson, J., 2020. Road to zero or road to nowhere? Disrupting transport and energy in a zero carbon world. *Energy Policy* 139, 111334.
- Cass, N., Schwanen, T., Shove, E., 2018. Infrastructure, intersections and societal transformations. *Technol. Forecast. Soc. Change* 137, 160–167.
- Centre for Research on Energy Demand Solutions (CREDS), 2019. What we do. Available from: <https://www.creds.ac.uk/what-we-do/>.
- Creutzig, F., Jochem, P., Edelenbosch, O.Y., Mattauch, L., van Vuuren, D.P., McCollum, D., Minx, J., 2015. Transport: A roadblock to climate change mitigation? *Science* 350 (6263), 911–912.
- Department for Energy (ND) Where the Energy Goes: Gasoline Vehicles, Available at: <https://www.fueleconomy.gov/feg/atv.shtml>.
- de Satgé, R., Watson, V., 2018. African cities: planning ambitions and planning realities. In: de Satgé, R., Watson, V. (Eds.), *Urban Planning in the Global South*. Springer Nature, Switzerland.
- European Commission, 2019. Air quality: Traffic measures could effectively reduce NO₂ concentrations by 40% in Europe’s cities, EU Science Hub. 27 November 2019. Available from: <https://ec.europa.eu/jrc/en/news/air-quality-traffic-measures-could-effectively-reduce-no2-concentrations-40-europe-s-cities>.
- European Environment Agency, 2019. Final energy consumption in Europe by mode of transport. Available from: <https://www.eea.europa.eu/data-and-maps/indicators/transport-final-energy-consumption-by-mode/assessment-10>.
- García-Olivares, A., Solé, J., Osychenko, O., 2018. Transportation in a 100% renewable energy system. *Energy Conserv. Manage.* 158, 266–285.

- Guillen-Royo, M., 2019. Sustainable consumption and wellbeing: does on-line shopping matter? *J. Clean. Prod.* 229, s1112–s1124.
- Healy, N., Barry, J., 2017. Politicizing energy justice and energy system transitions: fossil fuel divestment and a 'just transition'. *Energy Policy* 108, 451–459.
- Hickman, R., Hall, P., Banister, D., 2013. Planning more for sustainable mobility. *J. Transp. Geograph.* 33, 210–219.
- Höök, M., Tang, X., 2013. Depletion of fossil fuels and anthropogenic climate change—a review. *Energy Policy* 52, 797–809.
- Hopkins, D., Bengoechea, E.G., Mandic, S., 2019. Adolescents and their aspirations for private car-based transport. *Transportation*. DOI: <https://doi.org/10.1007/s11116-019-10044-4>.
- International Energy Agency (IEA), 2019a. Tracking Transport. Tracking Report, May 2019. Available from: <https://www.iea.org/reports/tracking-transport-2019>.
- Jaller, M., Pahwa, A., 2020. Evaluating the environmental impacts of online shopping: A behavioural and transportation approach. *Transp. Res. Transp. Environ.* 80, 102223.
- Lissandrello, E., Hrelja, R., Tennøy, A., Richardson, T., 2017. Three performativities of innovation in public transport planning. *Int. Plan. Studies* 22 (2), 99–113.
- Mueller, N., Rojas-Rueda, D., Khreis, H., Cirach, M., Andres, D., Ballester, J., Bartoll, X., Daher, C., Deluca, A., Echave, C., Mila, C., Marquez, S., Palou, J., Perez, K., Tonne, C., Stevenson, M., Rueda, S., Nieuwen huijsen, M., 2020. Changing the urban design of cities for health: The superblock model. *Environ. Int.* 134, 105132.
- Nixon, R., 2011. *Slow Violence and the Environmentalism of the Poor*. Harvard University Press, Cambridge, MA, USA.
- Ottelin, J., Heinonen, J., Junnila, S., 2014. Greenhouse gas emissions from flying can offset the gain from reduced driving in dense urban areas.. *J. Transp. Geogr.* 41, 1–9.
- Oswald, Y., Owen, A., Steinberger, J.K., 2020. Large inequality in international and intranational energy footprints between income groups and across consumption categories. *Nat. Energy* 5, 231–239.
- Pollock, N., Williams, R., 2010. The business of expectations: How promissory organizations shape technology and innovation. *Soc. Studies Sci.* 40 (4), 525–548.
- Saidi, S., Shahbaz, M., Akhtar, P., 2018. The long-run relationships between transport energy consumption, transport infrastructure and economic growth in MENA countries.. *Transport. Res. A* 111, 78–95.
- Schwanen, T., 2018. Towards decolonised knowledge about transport. *Palgrave Commun.* 4 (79), 1–6.
- Sims, R., Schaeffer, R., Creutzig, F., Cruz-Núñez, X., D'Agosto, M., Dimitriu, D., Figueroa Meza, M.J., Fulton, L., Kobayashi, S., Lah, O., McKinnon, A., Newman, P., Ouyang, M., Schauer, J.J., Sperling, D., Tiwari, G., 2014. Transport. In: Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Farahani, E., Kadner, S., Seyboth, K., Adler, A., Baum, I., Brunner, S., Eickemeier, P., Kriemann, B., Savolainen, J., Schlömer, S., von Stechow, C., Zwickel, T., Minx, J.C. (Eds.), *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- Sorrell, S., Gatersleben, B., Druckman, A., 2020. The limits of energy sufficiency: A review of the evidence for rebound effects and negative spillovers from behavioural change. *Energy Res. Soc. Sci.* 64, 101439.
- Sovacool, B.K., Ali, S.H., Bazilian, M., Radley, B., Nemery, B., Okatz, J., Mulvaney, D., 2020. Sustainable minerals and metals for a low carbon future. *Science* 367 (6473), 30–33.
- Tucker, V.A., 1975. The energetic cost of moving about. *Am. Sci.* 63 (4), 413–419.
- UNA-UK, 2016. Climate 2020: Decarbonising Transport. 14 September 2016. Available from: <https://www.climate2020.org.uk/decarbonising-transport/>.
- Urry, J., 2003. Social networks, travel and talk. *Br. J. Sociol.* 54 (2), 155–175. doi:10.1080/0007131032000080186.
- Wang, Z., Ahmed, Z., Zhang, B., Wang, B., 2019. The nexus between urbanization, road infrastructure, and transport energy demand: empirical evidence from Pakistan. *Environ. Sci. Pollut. Res.* 26, 34884–34895.
- World Health Organization (WHO) (ND). Air pollution. Health, environment and Sustainable development. Available from: <https://www.who.int/sustainable-development/transport/health-risks/air-pollution/en/>.
- Yang, X., Jin, W., Jiang, H., Xie, Q., Shen, W., Han, W., 2017. Car ownership policies in China: Preference of residents and influence on the choice of electric cars. *Transp. Policy* 58, 62–71.
- Zhang, Y.-J., Liu, Z., Qin, C.-X., Tan, T.-D., 2017. The direct and indirect CO₂ rebound effect for private cars in China. *Energy Policy* 100, 149–161.

Customer Satisfaction as a Measure of Service Quality in Public Transport Planning

Laura Eboli, Gabriella Mazzulla, University of Calabria, Department of Civil Engineering, Rende, Calabria, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	220
The Concepts of Service Quality and Customer Satisfaction: A Historical Perspective	220
The Measurement of Service Quality in Public Transport	221
Analyzing Public Transport Service Quality in Terms of Passengers' Perceptions: Theoretical Approaches	221
New Challenges	222
Conclusion	223
References	223
Further Reading	223

Introduction

Transport systems have significant environmental, social, and economic externalities associated with them. These systems provide mobility and physical access to resources, goods, markets, and other amenities enhancing the quality of life. Therefore, several externalities positively affect people's lifestyles, in terms of economic growth, educational opportunities, land value, and so on. On the other hand, there are several negative externalities, such as air and noise pollution, energy consumption, traffic congestion, cost of living, and traffic crashes, with direct negative consequences on public health and welfare (World Bank, 1996; OECD, 1999). The increase of negative externalities has been expedited by the continuing modal shift in favour of the private car, which is the most detrimental form of motorized transport. Modal substitution would represent an important strategy of demand management for the achievement of a sustainable transport system, by providing, as an example, better modal alternatives like PT systems characterized by high-quality levels. Improving PT SQ is important in order to keep the current users and to attract new ones. With the aim to guarantee better PT services, all the actors involved in the PT business should assure well-organized PT systems and allow people's need for mobility to be satisfied under safe and comfortable conditions. PT agencies must provide methods for measuring and monitoring their performances.

There is a variety of aspects that characterizes a PT service. According to the Transit Capacity and Quality of Service Manual (TRB, 2003), SQ factors can be distinguished into two main broad categories: availability factors, and comfort and convenience factors. The availability factors include service coverage, scheduling, capacity, and information; on the other hand, the comfort and convenience factors include passenger loads, reliability, travel time, safety and security, cost, appearance, and comfort. An alternative classification of PT SQ factors can be found in the European Standard (EN 13816) by the European Committee for Standardization (CEN, 2002), which classifies the SQ factors into the main following categories: availability, accessibility, information, time, customer care, comfort, security, environmental impact. Aside from the various classifications, PT SQ depends on many different aspects, and just for this reason it is important to investigate these in order to identify which are the most important or the most critical for the users, with the final aim to satisfy customers and attract new ones.

For a long time the performance evaluation of PT has been carried out from the service managers' perspective. However, in the last few decades, SQ has become a major area of attention for practitioners, managers, and researchers, who have focused on the passengers' perspective.

This chapter wants to provide a picture of the state of the art concerning the SQ analysis in PT sector, with a particular attention to the perceptions of the passengers about the used services, in terms of satisfaction with the services. In the rest of the paper, we will clarify the relationship between SQ and customer satisfaction, and other fundamental concepts; then, we will give an overview of the methods adopted and proposed for measuring and analyzing SQ in PT, and finally we will provide new challenges of studying and analyzing SQ.

The Concepts of Service Quality and Customer Satisfaction: A Historical Perspective

SQ can refer to four points of view: the customer, the service provider, the service beneficiaries (customers and community), and the service partners' (operators, authorities, etc.). Starting from this assumption, some concepts linked to SQ can be better understood. The EN 13816 (CEN, 2002) defines four main concepts of SQ. More specifically, there is a level of quality, named as "quality sought," which is required by the customer, and a level of quality, named as "quality targeted," which the service provider aims to provide for the customers; this level is influenced by the level of quality sought. On the other hand, "quality delivered" is defined as the level of quality achieved on a day-to-day basis. Finally, there is the level of quality perceived by the customer, which is named as

“quality perceived.” In particular, the difference between “quality sought” and “quality perceived” may be taken as the degree of customer satisfaction. This definition has been largely supported in the literature; several researchers, in fact, view SQ as a means of achieving user satisfaction (Chen, 2008). Also according to the Transit Capacity and Quality of Service Manual (TRB, 2003), the concepts of SQ and customer satisfaction are interconnected. The manual defines SQ as the overall measured or perceived performance of PT service from the passenger’s point of view. Many researchers consider the customer’s point of view as the most relevant for evaluating PT performance.

Customer satisfaction in PT has been studied since the mid-1960s (TRB, 1999), and since the 1990s, the application of marketing techniques has provided transport researchers with a tool to study satisfaction with respect to travel (Fornell et al., 1996). The concept of satisfaction with travel has therefore been well established over time, and frequently discussed and used in the literature. The concept of SQ linked to the concept of customer satisfaction dates back to the 1990s, when Berry et al. (1990) pointed out that “customers are the sole judge of SQ.” Parasuraman et al. (1985) state that the perception of SQ is the result of a comparison of consumer expectations (i.e., quality sought) with actual service performance perceptions (i.e., quality perceived). Other authors, however, do not take expectations into consideration (Cronin and Taylor, 1992). While perceptions are usually measured by satisfaction ratings, there are varying interpretations of expectations. Teas (1993) stated that expectations could be interpreted as predictions of service, as an ideal standard, or as attribute importance. Smith (1995) contends that a number of researchers have directly substituted importance measures for expectations, affirming that measuring service attributes that are important to customers may be more meaningful to managers than measuring customer service expectations.

The relationship between SQ and satisfaction is not clear. In the literature, SQ usually accompanies satisfaction. This may be due to the similar nature of the two concepts, which both derive from the disconfirmation theory (Parasuraman et al., 1988). Some authors think that customer satisfaction causes perceived quality and others consider that SQ is a vehicle for satisfaction (e.g., Chen, 2008). More recently, since the beginning of the twenty-first century, a number of studies have focused on aiming to understand what drives satisfaction compared to loyalty, and it has become important to understand the differences between these two commonly used terms. In the PT context, loyalty is defined as a customer’s intention to use the service in the future based on previous experiences (TRB, 1999). Although it is possible to measure satisfaction without considering loyalty, the results of recent studies suggest that the reverse would be theoretically illogical as satisfaction tends to influence loyalty (van Lierop et al., 2018).

The Measurement of Service Quality in Public Transport

As largely discussed in the previous section, a widely adopted SQ measurement in PT is based on the customer point of view, even if SQ in PT is generally measured also from the service provider point of view by a range of disaggregate performance measures, which provide an “objective” evaluation of SQ. These measures can be derived from different sources of data; as an example, they can be calculated from information an agency would normally have in hand for other purposes (e.g., schedule data, system maps, complaint records, and so on). A particular form of manual data collection can be represented by the Passenger Environment Surveys (PES), which generally assess qualitative elements which are difficult to measure by any other way. In fact, PES use a “secret shopper” technique, according to which mystery riders travel through the PT system and rate a variety of trip attributes in order to provide a quantitative evaluation of factors that passengers would think of qualitatively (TRB, 2003).

On the other hand, the measurement of SQ from the customer point of view can be considered a subjective measure of SQ. It is generally derived from the well-known Customer Satisfaction Surveys (CSS); CSS are useful to help prioritize SQ improvement initiatives, measure the degree of success of past initiatives, and track changes in SQ over time (TRB, 2003). According to the CSS, users express their judgments about the services according to a predefined scale of measure. The kinds of scale of measurement most frequently used are the Likert, verbal and the numerical scale (Hill, 2003). User perceptions can be expressed also in terms of choice, which are collected from experiments based on Stated Preferences (SP) techniques (Hensher and Prioni, 2002). In this case, users are asked to make their choice among hypothetical services characterized by service aspects having certain levels of quality; by making the choice, the user gives indirectly a relevance to certain aspects rather than another aspects. An alternative kind of data is represented by the best-worst scaling, according to which users have to indicate which is the best factor and the worst one in terms of satisfaction and importance (Echaniz et al., 2019).

Analyzing Public Transport Service Quality in Terms of Passengers’ Perceptions: Theoretical Approaches

The analysis of SQ based on passengers’ perceptions has been extensively studied in the transport literature. There are many authors who worked on this issue and who continue to investigate and propose methods for analysing SQ in PT starting from data collected through surveys addressed to passengers. The literature is rich of studies analyzing different PT modes, from services offered by PT systems such as buses to railways services, or airport and airline services. Independently of the analyzed transport modes, the various studies differ in terms of approaches adopted for investigating SQ. The adoption or the development of the different methods has to be chosen as a function of the objective of the analysis. There are simple or more advanced methods and techniques depending on the level of detail that the analyst wants to achieve. The literature provides some techniques developed in the marketing field aimed to measure SQ in terms of customer satisfaction; the most famous being SERVQUAL (Parasuraman et al., 1985), which is an index function of the differences between expectation and perceptions of the users about a series of SQ factors. This method has been

largely applied in the literature, and several authors proposed also some modifications. This kind of method provides a measure of the level of quality of the service, based on the perceptions of the users about the single SQ aspects. There are similar methods such as the Customer Satisfaction Index (CSI) (Hill et al., 2003), which represents a measure of SQ on the basis of attributes' importance and satisfaction rates. Also for this method the literature provides several applications and modifications (e.g., the HSCI proposed by Eboli and Mazzulla, 2009). Next to these techniques providing an overall SQ measure, there are some practical techniques that provide only a measure of the level of quality of the single SQ factors. An example of these techniques is the Importance-Performance Analysis (IPA), which is a quadrant analysis that uses importance and performance as coordinates (Martilla and James, 1977). In general, the methods that adopt the importance concept can have two variants: the importance can be the importance directly stated by the passengers during the interview, or it can be calculated starting from the satisfaction rates, through correlations or regression weights of the satisfaction rates expressed for each SQ factor and the satisfaction expressed for the overall service.

Other interesting methods aiming to provide a level of SQ for each service factor is represented by procedures based on the combination of subjective and objective indicators calculated on the basis of users' perception about the service and measurements provided by the PT agency (e.g., Eboli and Mazzulla, 2011). The aim of the above mentioned methodologies is to obtain a measure of the overall SQ or of the quality of the single service factors, with the final goal to improve the level of SQ as a function of the criticalities registered through the application of the measurement method. On the other hand, there is a wide series of methodologies that allow the relevance of each service factor to be derived, with the final aim to understand which are the service aspects to which the service provider must point for improving the service and satisfying the users. There are simple techniques such as factor analysis, regression models, and more advanced methodologies such as Structural Equation Modeling (SEM), which allows the modeling of a phenomenon by considering both the unobserved "latent" constructs and the observed indicators that describe the phenomenon (de Oña et al., 2013). These models are generally applied by using satisfaction rates expressed according to different kinds of scale. An extension of these models can be found in the studies that introduce also the loyalty next to the more traditional concepts of SQ and customer satisfaction (e.g., de Oña et al., 2016a; Chen, 2016). Another method allowing the characteristics that most influence overall SQ to be identified is the classification and regression tree (CART) approach. The CART methodology can provide more useful and informative results than other approaches because of the "If-then" rules generated from the trees (de Oña et al., 2015).

There are other kinds of models that adopt data collected from experiments based on Stated Preferences (SP) techniques. These types of data are usefully analyzed through discrete choice models based on the Random Utility Theory (RUT), according to which users are rational decision-makers who make their choice by maximizing their utility. Although over the last few decades discrete choice models have been widely used for simulating the choice among different transport modes, within mode models have been proposed where the alternatives relate to a single transport mode, usually PT. Hensher and Prioni (2002) first proposed a methodology for measuring SQ in PT through the choice-based conjoint analysis. The work of Hensher have been followed by other authors who proposed more advanced models aimed to investigate on the relevance of the various SQ factors starting from the choices made by the users between the habitual service and more hypothetical services; by choosing his/her more attractive service the user indirectly expresses a judgement of importance about the service aspects. Examples of these models can be found in dell'Olio et al. (2011) and Eboli and Mazzulla (2008, 2010).

New Challenges

This section presents a picture of the most recent studies and the new trends in the analysis of SQ in PT. Some of these new methods represent advances of more traditional methods. In fact, some authors recently proposed more advanced SEM for better investigating the heterogeneity in passengers' satisfaction; as an example, Structural Equation Multiple Cause Multiple Indicator (SEM-MIMIC) models were proposed by Allen et al. (2020) and SEM Multi-Group Analysis (SEM-MGA) by Allen et al. (2019). Investigating the heterogeneity among users' is important in order to search for differences among users' perceptions, findings that can be useful for PT operators for identifying the most appropriate strategies for satisfying the various categories of users. Other authors pointed on an extension of the methods based on the use of SP data, and specifically for investigating not only on the opinions of PT current users, but also of the potential users, by introducing the concept of desired quality next to the concepts of expected and perceived quality (dell'Olio et al., 2011; Bellizzi et al., 2020). In this way, PT planners and operators could conveniently consider the opinions of potential users, and not only users, in order to satisfy the expectations of the users, but above all to accommodate the desired levels of quality of the potential users.

Another interesting challenge is represented by SQ analysis over space and time. Although SQ in PT has been widely investigated, only few studies considered the variability of passengers' opinions in terms of space and time, given that the major part of previous studies analysed passengers' satisfaction with PT services in a cross sectional manner only. However, it is common knowledge that passenger's perceptions about PT SQ attributes strongly depend on the context to which the passenger refers. For this reason, each attribute judged by users in terms of satisfaction or importance could show spatial and/or temporal variability. Recent examples of this kind of studies can be found in de Oña et al. (2016b), who proposed the use of the index numbers usually applied in the economic and industrial field for analyzing the variation of SQ over time, and Eboli et al. (2018) who investigated SQ at railway stations aiming to suggest a methodology for analyzing spatial variation of SQ attributes.

Conclusion

This chapter aimed to provide an exhaustive picture of the research and the methodologies proposed in the literature for assessing PT SQ. The work could represent a useful instrument for the researchers of the sector and practitioners interested in the measure of SQ based on passengers' perceptions. The chapter attempted to give also an idea of the future concerning the investigation on PT SQ.

References

- Allen, J., Eboli, L., Forciniti, C., Mazzulla, G., Ortuzar, J.D., 2019. The role of critical incidents and involvement in transit satisfaction and loyalty. *Trans. Policy* 75, 57–69, doi:10.1016/j.tranpol.2019.01.005.
- Allen, J., Eboli, L., Mazzulla, G., Ortuzar, J.D., 2020. Effect of critical incidents on public transport satisfaction and loyalty: an Ordinal Probit SEM-MIMIC approach. *Transportation* 47 (2), 827–863, doi:10.1007/s11116-018-9921-4.
- Bellizzi, M.G., dell'Olio, L., Eboli, L., Mazzulla, G., 2020. Heterogeneity in desired bus service quality from users' and potential users' perspective. *Trans. Res. Part A* 132, 365–377.
- Berry, L.L., Zeithaml, V.A., Parasuraman, A., 1990. Five imperatives for improving service quality. *Sloan Manag. Rev.* Summer 31 (4), 9–38.
- CEN/TC 320, 2002. Transportation—Logistics and services. European Standard EN 13816: Public passenger transport—Service quality definition, targeting and measurement. European Committee for Standardization, Brussels. Technical Report.
- Chen, C., 2008. Investigating structural relationships between service quality, perceived value, satisfaction, and behavioral intentions for air passengers: Evidence from Taiwan. *Trans. Res. Part A* 42 (4), 709–717.
- Chen, H.K., 2016. Structural interrelationships of group service quality, customer satisfaction, and behavioral intention for bus passengers. *Int. J. Sustain. Transp.* 10, 418–429.
- Cronin, J., Taylor, S., 1992. Measuring service quality: a reexamination and extension. *J. Mark.* 56, 55–68.
- de Oña, J., de Oña, R., Eboli, L., Forciniti, C., Mazzulla, G., 2016a. Transit passengers' behavioural intentions: the influence of service quality and customer satisfaction. *Transport. A* 12 (5), 385–412, doi:10.1080/23249935.2016.1146365.
- de Oña, J., de Oña, R., Eboli, L., Mazzulla, G., 2013. Perceived Service quality in bus transit service: A structural equation approach. *Trans. Policy* 29, 219–226, doi:10.1016/j.tranpol.2013.07.001.
- de Oña, J., de Oña, R., Eboli, L., Mazzulla, G., 2015. Heterogeneity in perceptions of service quality among groups of railway passengers. *Int. J. Sustain. Transp.* 9 (8), 612–626, doi:10.1080/15568318.2013.849318.
- de Oña, J., de Oña, R., Eboli, L., Mazzulla, G., 2016b. Index numbers for monitoring transit service quality. *Trans. Res. Part A* 84, 18–30, doi:10.1016/j.tra.2015.05.018.
- dell'Olio, L., Ibeas, A., Cecin, P., 2011. The quality of service desired by public transport users. *Trans. Policy* 18, 217–227.
- Eboli, L., Forciniti, C., Mazzulla, G., 2018. Spatial variation of the perceived transit service quality at rail stations. *Trans. Res. Part A* 114, 67–83, doi:10.1016/j.tra.2018.01.032.
- Eboli, L., Mazzulla, G., 2008. An SP experiment for measuring service quality in public transport. *Trans. Plan. Technol.* 31 (5), 509–523, doi:10.1080/03081060802364471.
- Eboli, L., Mazzulla, G., 2009. A new customer satisfaction index for evaluating transit service quality. *J. Public Transp.* 12 (3), 21–37.
- Eboli, L., Mazzulla, G., 2010. How to capture the passengers' point of view on a transit service through rating and choice options. *Trans. Rev.* 30 (4), 435–450, doi:10.1080/01441640903068441.
- Eboli, L., Mazzulla, G., 2011. A methodology for evaluating transit service quality based on subjective and objective measures from the passenger's point of view. *Trans. Policy* 18 (1), 172–181, doi:10.1016/j.tranpol.2010.07.007.
- Echaniz, E., Ho, C.Q., Rodríguez, A., dell'Olio, L., 2019. Comparing best-worst and ordered logit approaches for user satisfaction in transit services. *Trans. Res. Part A* 130, 752–769, doi:10.1016/j.tra.2019.10.012.
- Fornell, C., Johnson, M., Anderson, E., Cha, J., Bryant, B., 1996. The American customer satisfaction index: Nature, purpose, and findings. *J. Market.* 60 (4), 7–18.
- Hensher, D.A., Prioni, P., 2002. A service quality index for a area-wide contract performance assessment regime. *J. Trans. Econ. Policy* 36 (1), 93–113.
- Hill, N., 2003. How to Measure Customer Satisfaction. Gower Publishing, Hampshire.
- Martilla, J.A., James, J.C., 1977. Importance performance analysis. *J. Market.* 41 (1), 77–79.
- OECD (Organisation for Economic Co-operation and Development), 1999. Environment Policy Committee; Working Group on the State of the Environment. Indicators for the Integration of Environmental Concerns into Transport Policies; OECD Publications: Paris, France.
- Parasuraman, A., Zeithaml, V.A., Berry, L.L., 1985. A conceptual model of service quality and its implication for future research. *J. Market.* 49, 41–50.
- Parasuraman, A., Zeithaml, V.A., Berry, L.L., 1988. Servqual: A multiple item scale for measuring consumer perceptions of service quality. *J. Retail.* 64 (1), 12–40.
- Smith A.M. 1995. The consumer's evaluation of service quality: an examination of the SERVQUAL methodology. Doctoral dissertation, University of Manchester. British Thesis Service, D189377.
- Teas, R.K., 1993. Expectations, performance evaluation and consumer's perception of quality. *J. Mark.* 57 (4), 18–34.
- Transportation Research Board, 1999. A Handbook for Measuring Customer Satisfaction and Service Quality. TCRP Report 47. National Academy Press, Washington, D.C.
- Transportation Research Board, 2003. Transit Capacity and Quality of Service Manual. TCRP Report 100. National Academy Press, Washington, D.C.
- van Lierop, D., Badami, M.G., El-Geneidy, A.M., 2018. What influences satisfaction and loyalty in public transport? A review of the literature. *Trans. Rev.* 38 (1), 52–72, doi:10.1080/01441647.2017.1298683.
- World Bank, 1996. Sustainable Transport—Priorities for Policy Reform. World Bank, Washington, DC, USA.

Further Reading

- de Oña, J., de Oña, R., 2015. Quality of service in public transport based on customer satisfaction surveys: A review and assessment of methodological approaches. *Trans. Sci.* 49 (3), 605–622.
- dell'Olio, L., Ibeas, A., de Oña, J., de Oña, R., 2018. Public Transportation Quality Of Service. Factors, Models and Applications. Elsevier, Amsterdam.
- Friman, M., Fellelsson, M., 2009. Service supply and customer satisfaction in public transportation: The quality paradox. *J. Public Transp.* 12 (4), 57–69, doi:10.5038/2375-0901.12.4.4.
- Hansson, J., Pettersson, F., Svensson, H., Wretstrand, A., 2019. Preferences in regional public transport: a literature review. *Eur. Trans. Res. Rev.* 11, 38, doi:10.1186/s12544-019-0374-4.
- Hensher, D.A., 2007. Bus Transport: Economics, Policy and Planning. JAI Press, NY.
- Hill, N., 2000. Handbook of Customer Satisfaction and Loyalty Measurement. Gower Publishing, Hampshire.
- Redman, L., Friman, M., Garling, T., Hartig, T., 2013. Quality attributes of public transport that attract car users: A research review. *Trans. Policy* 25, 119–127.
- Transportation Research Board, 1998. A handbook: Integrating market research into transit management. TCRP Report 37.

- Transportation Research Board, 2002. A guidebook for developing a transit performance-measurement system transit cooperative research program (Vol. 88). Washington, DC: Transportation Research Board: United States Federal Transit Administration.
- Transportation Research Board, 2003. A guidebook for developing a transit performance-measurement system. TCRP Report 88; National Academy Press: Washington, D.C. 2003.
- Tyrinopoulos, Y., Antoniou, C., 2008. Public transit user satisfaction: Variability and policy implications. *Trans. Policy* 15, 260–272, doi:10.1016/j.tranpol.2008.06.002.
- van Lierop, D., El-Geneidy, A., 2016. Enjoying loyalty: The relationship between service quality, customer satisfaction, and behavioral intentions in public transit. *Res. Trans. Econ.* 59, 50–59.

Car Sharing and the Impact on New Car Registration

Mario Intini*, Marco Percoco†, *Department of Economics, Management and Business Law, University of Bari Aldo Moro, Bari, Italy;

†Department of Social and Political Sciences and GREEN, Università Bocconi, Milan, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	225
A Review of the Literature	225
Some Empirical Evidence From Italy	226
Conclusion	229
References	229

Introduction

The paradigm of sustainable mobility is currently very high on the agenda of policymakers because transport is widely believed to generate many externalities in terms of pollution, congestion, and accidents (Banister, 2008). Transport accounts for about 23% of energy-related carbon dioxide (CO₂) emissions and 15% of global greenhouse gas (GHG) emissions. According to the European Union, cars are responsible for approximately 12% of total EU CO₂ emissions (Bergantino et al., 2019). To tackle this issue, an increasing number of cities are adopting a wide variety of policies aimed at promoting the shift from private fuel-consuming cars to more sustainable transport means, such as public transport, bikes, and ecological vehicles.

This gradual shift has been enhanced by the introduction of tolls on roads (Percoco, 2013, 2014; 2016), incentives for green cars (Coat et al., 2009) and the expansion of public services (Shaheen and Cohen, 2013). More recently, sharing mobility has attracted the interest of policymakers because of its potential impact on car ownership and ultimately the number of kilometers travelled in urban areas. The potential of car sharing in reducing car registration is a topic of considerable attention in the automobile industry. Each shared car can be used by more than one person, and for this reason, the car sharing fleet is expected to substitute more private cars than the number of shared cars, thus reducing the total number of cars (Liao et al., 2020). The literature on the topic is very limited, although growing, especially from the empirical perspective. Nevertheless, in this article, we review existing studies and provide new evidence from Italian cities.

The European Commission's White Paper of 2011, namely "Roadmap to a Single European Transport Area—Towards a competitive and resource-efficient transport system (28/03/2011)" imposed a reduction of 60% of GHGs (compared to 1990 levels) in the transport industry by 2050 upon all European Union member states. The Italian government has attempted to intervene in private car utilization for 10 years by attempting to introduce and regulate alternative methods of mobility and pursuing a circulating fleet renewal. As a result, Italy saw the inception of the "Iniziativa Car Sharing" program (ICS), a voluntary agreement between some Italian municipalities that provided the first car-sharing services with a common organizational set-up. The free-floating approach immediately gained consumer confidence, and the number of registered users increased dramatically in a short period of time. For instance, the number of daily car-sharing rides in the city of Milan increased from 29 at launch to 5000 after four months and up to 10,000 after one year. In 2017, the registered users numbered 1.3 million people, 820,000 of whom were active users (Ciuffini et al., 2017). Milan has been the forerunner of the sustainable mobility paradigm, but several other major cities followed: Rome, Turin, Florence, Catania, and Bologna (Rotaris and Danielis, 2018). In this article, we review the literature on the impact of car-sharing and discuss some empirical evidence on new car sales in Italy.

A Review of the Literature

The impact of car-sharing on car ownership has been studied in a limited number of papers. Myers and Cairns (2009) conducted a survey of users of four UK car-sharing operators and found that the effect of sharing mobility on car ownership may take different forms: (1) the decision of users to sell their cars, resulting in a reduction in car ownership; (2) vehicles being scrapped, resulting in a decrease in the circulating fleet of cars, and (3) the decision to avoid a new car purchase, resulting in a reduction of the number of second cars owned by households. Myers and Cairns (2009) found an overall reduction in vehicles after the introduction of car-sharing because some users decided to sell their cars, not to buy a new car or to get their cars scrapped after they signed up with a sharing mobility service. Similar results were also obtained in Bremen (2005) for both private and corporate users by Maertins (2006) and Knie and Canzler (2005) in Germany and by Ueli et al. (2006) in Switzerland. The reduction in the number of vehicles is also supported by Martin et al. (2010) in various cities, Cervero and Tsai (2004) in San Francisco Bay Area, Price et al. (2006) in Washington DC, Zipcar (2005) in various cities, Lane (2005) in Philadelphia, Ryden and Morin (2005) in Belgium and Bremen city, Carplus (2014) in London and Ademe (2013) in Paris. A recent study of Liao et al. (2020), considering the Dutch population, show

that around 40% of car drivers are able to replace some of their private car trips by car sharing, and 20% indicated that they may forego a planned purchase or shed a current car if car sharing becomes available near to them.

A reduction in fuel consumption due to car-shared fleet efficiency as compared to the average of a circulating car fleet was detected by the *Mobility Genossenschaft* (2009) in Switzerland; by *Knie and Canzler* (2005) in Germany; by *Carplus* (2007) in the UK and, finally, by Information By E-mail Taxystop in Belgium. In their analysis, *Martin and Shaheen* (2011) transformed the fuel consumption savings into emissions reductions (measured in terms of avoided tons of GHG). The finding was an overall reduction of 34% of GHG emissions due to the efficiency of the new cars introduced by the car-sharing providers in North America. This methodology was also used by *Carplus* (2014), which updated his previous study and focused on the city of London, excluding other UK cities. The result was an estimated reduction of 410 kg of CO₂ per year in the UK's capital city. A decrease in the overall distance covered by private vehicles after the car-sharing was introduced (measured in vehicles per km) was found by *Martin and Shaheen* (2011) in North America; by *Cervero* (2003, *Cervero* (2003, 2004, 2008) and *Cervero and Tsai* (2004) in the Bay Area of San Francisco; by *Ryden and Morin* (2005) in Belgium and the German city of Bremen; by *Carplus* (2014) in London, by ICS (2009) in 10 Italian cities; by *Sioui et al.* (2013) in Montreal and by *Ademe* (2014) in Paris.

All these studies involved a station-based car-sharing service and showed that the users, at various magnitude, did reduce their habits in terms of distance covered after car-sharing was introduced in comparison to previous travel behaviors. For instance, the average daily reduction in terms of vehicles per km was 2% for the San Francisco Bay Area (*Cervero and Tsai*, 2004), and 45% in terms of vehicles per km in the German city of Bremen (*Ryden and Morin*, 2005).

In Montreal, *Sioui et al.* (2013) compared two analyses carried out simultaneously. The first studied a sample of 14,500 users of the *Communauto* car-sharing provider, while the second took into consideration the transportation routines of Montreal citizens, whose data were collected by the statistical offices of the city. The authors divided the sample according to the number of private cars owned by the households. Interestingly, people who owned one or more cars and were also car-sharing users drove much less than those who did not subscribe to a car-sharing service. In particular, in a household with one car, a car-sharing customer drove 28% less as compared to someone who was not a user. With two cars available in a household, the reduction reached 57%.

The aforementioned literature considered station-based car-sharing services, whereas comparatively little attention was devoted to free-floating services, which promise to be more flexible and pervasive. A notable exception is the work of *Martin and Shaheen* (2016), who studied a significant sample of *Car2go* users in five major North American cities, San Diego, Seattle, Vancouver, Calgary, and Washington DC, from 2014 to 2015, finding a reduction of up to 10% for newly registered cars. *Martin et al.* (2016) also found that the presence of car-sharing also reduced private car utilization (detected through the decrease in distance covered), but the magnitude of the reduction may not be precisely determinable. Specifically, GHG emissions were reduced by around 39,127 tons in a year (upper estimate) due to *Car2go* presence: -8137 tons in Calgary, -5371 tons in San Diego, -9080 tons in Seattle, -10,027 tons in Vancouver, and -6512 tons in Washington DC. The "modal shift" toward more environment-friendly forms of transportation implied that the car-sharing users in those cities did not change their habits regarding the use of bikes and trains. They traveled more by foot (+ 11%), and, finally, they reduced their frequency of use for both public means of transport (12% for the underground and 25% for public buses) and taxis (55%).

The negative effect of car-sharing on new car sales was detected by *Schmidt* (2020) in some German cities, where free-floating car-sharing had been present for longer.

Some Empirical Evidence From Italy

In this section, the empirical analyses used to determine the impacts of free-floating car-sharing's introduction are identified. Starting from a panel data layout, we estimated several versions of the following Eq. (1):

$$Y_{crt} = \alpha + \lambda_c + \lambda_r + \lambda_t + \rho D_{it} + X_{it}\delta + \varepsilon_{it} \quad (1)$$

where the dependent variable is (log) new car sales in city c in region r in year t . The unit of observation is at the city-region-year-model level in the first two panels, the month-year-model level in the third and fourth panels and the day-month-year-model level in the fifth and sixth panels. The dependent variable Y is the measure of "New Car Sales".

The fixed effects are treated as parameters to be estimated: c stands for city, r for region, d for day, m for months and y for years. The various λ terms indicate the fixed effects, and the vector X_{it} gathers the various control variables. D_{it} denotes the various treatments, and its parameter ρ is the object of the analysis, including, in some cases, its algebraic sign; ε indicates the error term.

The setup of the regression formula depends either on a temporal difference (times series analyses), which means that all the variables are observed before and after the introduction of car-sharing, or on a place-temporal difference (difference-in-difference analyses), which means that all the variables are observed before and after the introduction of car-sharing in various places.

The parameter ρ quantifies the size of the difference in explained variables before and after the launch of car-sharing between the treated variables and the control variables.

However, considering the possibility of autocorrelation between the independent variables—although mitigated by the presence of fixed effects—we investigated the statistical association between the dependent variables and the explanatory variables' coefficients. Finally, we added some controls, to which we assigned some theoretical importance, that may influence the explained variables. The data regarding the number of new car registrations were collected from the Italian Ministry of Infrastructure and

Table 1 List of cities in the sample

<i>Regions</i>	(1) <i>Piedmont</i>	(2) <i>Lombardy</i>	(3) <i>Tuscany</i>	(4) <i>Lazio</i>	(5) <i>Sicily</i>
Cities	Turin ^a Alessandria Asti Biella Cuneo Novara Verbania Vercelli	Milan ^a Bergamo Brescia Como Cremona Lecco Lodi Mantua Monza Pavia Sondrio Varese	Florence ^a Arezzo Grosseto Livorno Lucca Massa Pisa Pistoia Prato Siena	Rome ^a Frosinone Latina Rieti Viterbo	Catania ^a Agrigento Caltanissetta Enna Messina Palermo Ragusa Syracuse Trapani
N	8	12	10	5	9
Years	20–2017	20–2017	20–2017	20–2017	20–2017

^aCar-sharing cities; Control cities

Table 2 Descriptive statistics

<i>Variable</i>	<i>Observations</i>	<i>Mean</i>	<i>Std. Dev.</i>	<i>Min</i>	<i>Max</i>
New cars	352	8695.79	23165.17	239	215,917
Shared cars	352	56.8267	303.2727	0	3141
Population	352	216660.5	458277.3	21,536	287,3494
Males	352	102783.2	216737.7	10,018	136,2384
Young males	352	6686.869	14238.18	512	91,680
Income	352	22656.01	3527.111	14,095	33,831

Transport (MIT) for the years 2010–17. The dataset contains information relating to vehicles registered in the National Vehicle Archive, managed by the Italian Driver and Vehicle Licensing Agency.

Table 1 summarizes the cities taken into account for the two difference-in-differences, as well as those that were used as controls, whereas **Table 2** contains descriptive statistics for the main variables. It should be noted that as control cities, we have used all other provincial chief-towns in the region in order to control for region-specific common features (e.g., policies implemented by regional governments). The selection of the control group may appear to be arbitrary; however, in the robustness checks in **Table 4**, we increase the comparability across cities by means of a weighting procedure.

Table 3 presents the difference-in-difference results for the effect of free-floating car-sharing's introduction on the number of new cars registered. The coefficients in columns (1) and (2) are negative and significant, as are the coefficients in columns (3) and (4). These findings indicate that an increase in the number of shared cars causes a decrease in the number of new cars registered. In particular, the log-log model quantifies that if the number of car-sharing cars available in a city increases by 1%, the number of new cars registered decreases by 3.42% in the same year, with all else being equal. Additionally, the lead model (F.) quantifies that if the number of car-sharing cars increases by 1%, the number of new cars registered decreases by 2.79% in the following year, all else being equal. These results are also confirmed by the *DiD treatment*—the car-sharing dummy—which is negative and significant at the 5% level in both the models.

The forward notation (F.) stands for one-year-forward observations, as compared to the other set of variables. It is reasonable to believe that the decision regarding whether to buy a new car occurs not directly after but some time after car sharing has been adopted.

Table 4 presents the regression results after the implementation of entropy balancing through the Stata command “*ebalance*”, which generated dynamic weights (“*_webal*”). This function of multivariate reweighting was depicted by Hainmueller (2012) and serves to generate balanced samples that are useful in analyses characterized by binary treatments. The weights assigned to the control cities make them more similar and allow them to be compared with the treatment cluster.

The treatment coefficients depicted in the balanced regression remained qualitatively meaningful and statistically significant at the 1% level in the normal setup, as well as at the 5% level in the lead setup. From an economic point of view, the coefficients' magnitude is even enhanced, resulting in a stronger effect on the balanced framework.

Table 3 Difference-in-difference results

	(1) <i>Log new cars</i>	(2) <i>F. log new cars</i>	(3) <i>Log new cars</i>	(4) <i>F. log new cars</i>
Car-sharing dummy	−0.210** (0.0995)	−0.176** (0.0682)		
Log CS cars			−0.0342** (0.0148)	−0.0279*** (0.0103)
Log population	0.227 (3.869)	3.756 (3.550)	0.635 (3.989)	3.956 (3.619)
Log males	0.365 (3.259)	−2.179 (2.914)	0.0308 (3.345)	−2.347 (2.961)
Log young	−0.549 (0.365)	−0.332 (0.258)	−0.520 (0.373)	−0.315 (0.264)
Log income	0.443 (0.346)	0.00367 (0.161)	0.447 (0.351)	0.00651 (0.163)
Constant	1.428 (13.73)	−9.207 (11.84)	0.0586 (14.22)	−9.856 (12.14)
Observations	352	308	352	308
R-squared	0.989	0.992	0.989	0.992
Within R-sq.	0.104	0.077	0.112	0.082
City FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES

Robust standard errors in parentheses.

*** $p < 0.01$ ** $p < 0.05$ **Table 4** Difference-in-difference results after the implementation of entropy balancing:^{*}

	(1) <i>Log new cars</i>	(2) <i>F. log new cars</i>	(3) <i>Log new cars</i>	(4) <i>F. log new cars</i>
Car-sharing dummy	−0.229*** (0.0525)	−0.137** (0.0669)		
Log CS cars			−0.0405*** (0.00705)	−0.0237** (0.0105)
Log population	4.997 (16.49)	−0.846 (10.50)	7.521 (14.70)	0.142 (10.38)
Log males	−6.455 (15.81)	1.303 (9.532)	−8.704 (13.76)	0.414 (9.274)
Log young	−0.556 (0.561)	−0.717 (0.437)	−0.342 (0.462)	−0.616 (0.456)
Log income	4.874 (3.555)	2.849 (2.276)	5.012 (3.563)	2.922 (2.267)
Constant	−18.06 (71.67)	−16.45 (48.23)	−27.12 (73.66)	−20.26 (49.34)
Observations	352	308	352	308
R-squared	0.987	0.986	0.987	0.987
Within R-sq.	0.338	0.128	0.360	0.138
City FE	YES	YES	YES	YES
Year FE	YES	YES	YES	YES
Weights	YES	YES	YES	YES

Robust standard errors in parentheses.

*** $p < 0.01$ ** $p < 0.05$ * $p < 0.1$

Conclusion

The need to pursue improvements in quality of life for urban citizens is driving considerable changes in mobility patterns, the composition of the vehicle fleet and even modes of transport. Bike- and car-sharing services seem to play a pivotal role in transport policymaking. In particular, the introduction of the car-sharing service is considered efficient because the more intensive use of vehicles would lead to a negative economic and environmental impact.

In this chapter, a brief literature review was presented with cases from around the world and a study was discussed to assess the short-run impact of the introduction of car-sharing in Italian cities in terms of new car sales. We have found that a 10% increase in the number of shared cars caused a 2%–3% reduction in the number of new cars registered in the city, implying a certain effectiveness on the part of car-sharing in reducing the car ownership rate and, hence, in the long run, the size of the car fleet.

This article shows that the total number of vehicles in a city can be reduced through car sharing system. This result is important in the long term, since it creates a reduction in the amount of land and infrastructure needed for parking, open up more space for development reducing the spatial footprint of a city. Car sharing can eliminate parking concerns without requiring a total abandonment of own cars. Since free-floating car sharing involve access to public parking spaces, they are more dependent on the support of local authorities. To increase the incentives for car sharing, the government needs information regarding the scale of impact of car sharing on new car registration. This article is an empirical contribution in this branch of research.

Both the literature and the case study presented in this article point at positive effects of car sharing in terms of travel distance and car ownership. Although these results cohere with the expectation that sharing mobility decreases the car ownership rate, they must be taken with caution because they provide only short-run evidence. As the effect of car-sharing in the long run remains largely unknown, further research on this topic is needed to corroborate or reject the underlying rationales for policy intervention in the field of car sharing.

References

- Ademe, 2014. Enquête sur l'autopartage en trace directe, Rapport final.
- Banister, D., 2008. The sustainable mobility paradigm. *Transp. Policy* 15 (2), 73–80.
- Bergantino, A.S., Intini, M., Percoco, M., 2019. New car taxation and its unintended environmental consequences. GREEN Working Paper Series.
- Carplus: Monitoring Car Clubs, 2007. First Carplus Car Club Annual Members Survey Report. Leeds.
- Carplus: Monitoring Car Clubs, 2014. Carplus Car Club Annual Members Survey Report. London.
- Cervero, R., 2003. City carshare: first-year travel demand impacts. *Transp. Res. Rec.* 1839, 159–166.
- Cervero, R., 2004. Transit-oriented development in the United States: Experiences, challenges, and prospects, Report 102, Transit Cooperative Research Program, Washington, DC.
- Cervero, R., Tsai, Y., 2004. City carshare in San Francisco CA: second-year travel demand and car ownership impacts. *Transp. Res. Rec.* 1887, 117–127.
- Cervero, R., Arrington, G.B., 2008. Vehicle trip reduction impacts of transit oriented housing. *J. Public Trans* 11 (3), 1–17.
- Ciuffini, M., Orsini, R., Asperti, S., Gentili, V., Grossi, D., Milioni, D., Specchia, L., 2017. 2 Rapporto nazionale sulla Sharing Mobility. Fondazione per lo Sviluppo Sostenibile, Roma.
- Coad, A., De Haan, P., Woersdorfer, J.S., 2009. Consumer support for environmental policies: An application to purchases of green cars. *Ecological Economics* 68 (7), 2078–2086.
- Hainmueller, J., 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* 25–46.
- Iniziativa Car Sharing (ICS), 2009. Indici valoriali, rapporto di ricerca Ipr Marketing.
- Knie, A., Canzler, W., 2005. Die intermodalen Dienste der Bahn: Wirkungen und Potenziale neuer Verkehrsdienstleistungen. Gemeinsamer Schlussbericht von DB Rent und WZB.
- Verbundprojekt Inter-modi – Sicherung der Anschluss- und Zugangsmobilität durch neue Angebotsbausteine im Rahmen der Forschungsinitiative Schiene. Berlin.
- Lane, C., 2005. PhillyCarShare: First-year social and mobility impacts of car sharing in Philadelphia, Pennsylvania. *Transp. Res. Rec.* 1927, 158–166.
- Liao, F., Molin, E., Timmermans, H., van Wee, B., 2020. Carsharing: the impact of system characteristics on its potential to replace private car trips and reduce car ownership. *Transportation* 47 (2), 935–970.
- Maertins, C., 2006. Die Intermodalen Dienste der Bahn: Mehr Mobilität und weniger Verkehr? Wirkungen und Potenziale neuer Verkehrsdienstleistungen.
- Martin, E., Shaheen, S., Lidicker, J., 2010. Impact of carsharing on household vehicle holdings: results from a North American Shared-Use Vehicle Survey. *Transport. Res. Rec.* 2143, 150–158.
- Martin, E., Shaheen, S., 2011. Greenhouse gas emission impacts of carsharing in North America. *IEEE Trans. Intell. Transp. Syst* 12 (4), 1074–1086.
- Martin, E., Shaheen, S.A., 2016. Impacts of Car2go on vehicle ownership, modal shift, vehicle miles traveled, and greenhouse gas emissions: An analysis of five North American cities, mimeo.
- Mobility Genossenschaft, Geschäfts Nachhaltigkeitsbericht, 2008. Luzern, 2009.
- Myers, D., Cairns, S., 2009. Carplus annual survey of car clubs 2008/09. TRL Report.
- Percoco, M., 2013. Is road pricing effective in abating pollution? Evidence from Milan. *Transp. Res. D Transp. Environ.* 25, 112–118.
- Percoco, M., 2014. The effect of road pricing on traffic composition: evidence from a natural experiment in Milan, Italy. *Transp. Policy* 31, 55–60.
- Percoco, M., 2016. The impact of road pricing on accidents: a note on Milan. *Lett. Spat. Resour. Sci.* 9 (3), 343–352.
- Price, J., DeMaio, P., Hamilton, C., 2006. Arlington Carshare Program: 2006 Report. Arlington County Commuter Services, Division of Transportation, Department of Environmental Services, Arlington, VA.
- Rotaris, L., Danielis, R., 2018. The role for carsharing in medium to small-sized towns and in less-densely populated rural areas. *Transp. Res. Part A Policy Pract* 115, 49–62.
- Ryden C., Morin, E., 2005. Mobility services for urban sustainability: Environmental assessment. Moses Report WP6, Trivector Traffic AB, Stockholm, Sweden.
- Schmidt, P., 2020. The effect of car sharing on car sales. *Int. J. Ind. Organ.*, 102622.
- Sioui, L., Morency, C., Trépanier, M., 2013. How car sharing affects the travel behavior of households: A case study of Montréal, mimeo.
- Shaheen, S.A., Cohen, A.P., 2013. Carsharing and personal vehicle services: Worldwide market developments and emerging trends. *Int. J. Sustain. Transp* 7 (1), 5–34.
- Ueli, H., Matti, D., Schreyer, C., Maibach, M., 2006. Evaluation Car-Sharing. Schlussbericht. Swiss Bern, Federal Office for Energy (Eds.).
- Zipcar, 2005. Zipcar customer survey shows car-sharing leads to car shedding.

Evaluation Methods in Transport Policy and Planning

Niek Mouter, Delft University of Technology, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Evaluation Methods in Transport Policy and Planning	230
Cost–Benefit Analysis	230
Participatory Value Evaluation	231
Multicriteria Analysis	232
Environmental Impact Assessment	233
Social Impact Assessment	234
Conclusion	234
Biography	234
References	235

Evaluation Methods in Transport Policy and Planning

Transportation networks provide an array of benefits in the forms of goods delivery, access to services and personal mobility. However, transport can also result into several adverse effects such as damages to nature reserves, CO₂ emissions and noise pollution. Governments are in many ways involved in transport policy and planning. For instance, they determine regulations such as speed limits and they can decide to invest public budget in the extension and/or maintenance of transport infrastructure. In many cases, policy makers want to decide informed by the expected positive and negative impacts of a transport policy option. Various ex-ante evaluation methods are available which can be used to provide such information. This article surveys five evaluation methods that are used in transport policy and planning: Cost–Benefit Analysis (CBA), Participatory Value Evaluation (PVE), Multicriteria Analysis (MCA), Environmental Impact Assessment (EIA), and the Social Impact Assessment (SIA). The five methods will be introduced, and differences will be described.

Cost–Benefit Analysis

In virtually all western countries CBA is mandatory when national funding is asked for large transport projects (Mackie et al., 2014). A CBA measures the social desirability of a transport policy option in a systematic way based on economic theory. CBA quantifies the project's positive and negative effects and translates these quantified effects in monetary terms. Next, these monetary impacts are presented as present values using a discount rate which accounts for the notion that people prefer present impacts over future impacts (Mouter, 2018). Finally, present values are aggregated into a final indicator such as the net present value (NPV).

Welfare economics provides the theoretical underpinnings of CBA (Boadway and Bruce, 1984). One of the key concepts in welfare economics is the Pareto criterion which entails that the social welfare effect of a project is positive if it makes someone better off without making anyone worse off (Nyborg, 2012). A problem is that government policies will hardly ever be able to pass this criterion. For instance, when considering new transport infrastructure, taxpayers will pay the costs, and quite probably there will be negative external effects, for example noise or CO₂ emissions. A more practical concept is the Kaldor–Hicks efficiency criterion, which relaxes the Pareto conditions by adding the possibility of (potential) compensation. The Kaldor–Hicks efficiency criterion asserts that a policy (or other change) can be considered as welfare increasing if those who benefit can compensate those who suffer from it, creating a Pareto improvement after compensation. According to the Kaldor–Hicks efficiency criterion, the compensation does not actually have to take place: it is enough that it is theoretically possible. This implies that there is only a potential Pareto improvement, and not necessarily an actual Pareto improvement. Standard CBAs are generally based on the Kaldor–Hicks efficiency criterion. CBA assesses whether a transport project passes the Kaldor–Hicks efficiency test by expressing all the positive and negative effects of a policy in monetary terms and adding them up. If the sum is positive the project is considered to be welfare enhancing as those who benefit can theoretically compensate those who suffer.

Welfare economics provides strict procedures for the objects that have standing in the CBA analysis, for the impacts that are considered in the analysis and for the way different impacts are valued. In principle, welfare economics prescribes two principles when conducting a CBA being individualism and non-paternalism. Individualism implies that the preferences of individual citizens form the basis of a CBA (Sen, 1979) and nonpaternalism entails that individuals are conceived to be the best judge of their own welfare. In combination these postulations imply that the citizens and firms that are affected by the policy are the sole objects who have standing in a CBA study and their preferences are respected. In principle, preferences of experts, stakeholders and policy makers related to the impacts of the transport project do not play any role in the analysis as these actors are only consulted to provide methodologies for deriving the preferences of citizens. Welfare economics also provides CBA researchers with a clear frame of reference when selecting the impacts of transport policy options that should (not) be included in a CBA because only impacts that

affect the welfare of individuals should be included. For instance, citizens' preferences for the way that the benefits and burdens of a transport policy options are distributed across society are not part of the total net benefits in a CBA (Mouter et al., 2017a). Another consequence that is excluded from a CBA is public support for a transport policy option (Mouter, 2017).

Welfare economics also provides clear guidance for the weighting procedure that is used in CBA to evaluate the impacts of a transport policy option. CBA measures a project's societal value by transferring the project's societal effects in monetary terms using the notion of the amount of money individuals are willing to pay from their private income. For a positive effect, the WTP is the maximum amount which a person is prepared to pay for it. For negative effects, the willingness-to-pay is negative (then often called willingness-to-accept). For instance, the most dominant approach to derive the amount of money that individuals are willing to pay for reductions in travel time and accident risk relies on (hypothetical) route choice experiments in which travellers are asked to make a series of private choices between routes which differ in terms of travel time, accident risk and travel costs (e.g., Hensher et al., 2009). Impacts of transport policy options on landscape, nature and noise pollution are evaluated through analyzing the private decisions of individuals in the real estate market (e.g., Allen et al., 2015).

Participatory Value Evaluation

A core postulation of the CBA is that impacts of transport projects financed by the government are valued based on the amount of money individuals and parties are willing to pay from their private income. However, a range of scholars argues that the way in which individuals balance their own money against the attributes of government projects may substantially differ from the way that these individuals believe that their governments should trade-off public budget and impacts of public projects (e.g., Ackerman and Heinzerling, 2004; Sagoff, 1988). These academics believe that the amount of money individuals are willing to pay from their after tax income in (hypothetical) route choices provides crucial information for a toll company to determine the price that travelers should pay that want to use the road. In addition, these scholars accept that a real estate agency uses past choices from individuals in the real estate market to estimate a price for which houses in a new real estate project can be sold. However, the above-mentioned academics assert that the private WTP valuation paradigm fails to consider that private choices may not fully reflect citizens' preferences towards public policy.

The divergence between people's preferences in a private choice setting and a public choice setting has been recognized by various transport scholars. For instance, Mackie et al. (2001) argue that there is no good reason for the value that the individual is willing to pay to reduce travel time to be equal to the value that society at large attaches to the reassignment of time of that individual to other activities. Moreover, Jara-Díaz (2007) observes that there can be a stark difference between the values for use in public sector appraisal and the values which a commercial operator would wish to use in an analysis of the same project. The literature points out several reasons why it is problematic to infer the welfare of a government policy through individuals' private willingness to pay in the context of private consumer choices; (1) People may not be willing to contribute individually because the impact of their individual contribution is imperceptible, but people may be willing to contribute when the whole community is forced to contribute through a government intervention because the aggregate impact of this coordinated contribution can be substantial (Sen, 1995); (2) People value the same impact differently in a private sphere (grocery store) and a public sphere (ballot box) as moral considerations might be more salient in the latter context (e.g., Sagoff, 1988); (3) Valuing impacts of a government project through observing their consumer choices ignores that people may place a value on the way collective decisions are made.

Scholars developed so-called willingness to allocate public budget experiments (WTAPB) to alleviate the issues addressed above (e.g., Mouter et al., 2017b). In WTAPB experiments individuals make choices over alternative allocations of government budget across different government projects. The WTAPB approach thus aspires to infer welfare effects of (impacts of) government projects from individuals' preferences regarding the expenditure of public budget. The most important benefit of the WTAPB valuation approach is that individuals who believe that government budgets should be spent on different purposes than their own money can express these preferences (Mouter et al., 2020). A second advantage of the WTAPB valuation approach is that it bypasses the concern that WTP-based valuation is an inappropriate way to value impacts of government projects that are relatively difficult to translate into private income examples being landscape and nature. WTAPB does not require translation of government project impacts into private income. Instead, an impact of a government project is valued through the extent to which individuals are willing to sacrifice other impacts of government projects (Mouter et al., 2020). For instance, in a WTAPB experiment, individuals are asked to trade-off noise pollution against other impacts of governmental policy (e.g., reduction of mortality risk) which contrasts the WTP valuation approach in which individuals are asked to trade-off environmental impacts against private income. Moreover, the other two issues with the private WTP paradigm addressed in the previous paragraph ("people may be willing to contribute collectively, but may not be willing to contribute individually" and "people may place a value on the way collective decisions are made") are addressed in a WTAPB setting because impacts of transport projects are valued in a collective setting in which overall burdens and benefits of proposed transport projects are considered together in the context of a government decision (Mouter et al., 2019).

Although the WTAPB approach ameliorates various critiques of the private WTP valuation paradigm adopted in standard CBA, one clear downside of this valuation approach is that respondents are forced to spend the public budget as they have to make a choice between two or three alternative allocations of public budget. Therefore, preferences of individuals who believe that it is better to do nothing and/or reduce taxes instead of allocating public budget to one of the proposed projects are not respected. Participatory Value Evaluation (PVE) is a valuation method which has been developed to address this limitation of WTAPB. PVE establishes the desirability of government projects based on an experiment in which individuals select their preferred portfolio of government

projects given a constrained public budget (Mouter et al., 2021). The crucial difference between PVE and WTAPB is that participants in a PVE have the option to advise the government against allocating the budget to any (or some) of the projects that are considered in the PVE and shift the remaining budget to the next year (Mouter et al., 2020).

Recent literature on PVE distinguishes between two variants of PVE-experiments. First, the “fixed budget PVE format” which allows participants to allocate an earmarked amount of public budget and/or shift the public budget to the next year. To provide a stylized example of a “fixed-budget” PVE-experiment, assume that respondents can allocate a budget of 8 million euro to 16 projects all costing 1 million euro. In this context, respondents can allocate the budget to 8 or less projects. The second type of PVE experiment is called “the flexible budget PVE format”. In this format, participants are also allowed to change the public budget which will imply a change in taxes thereby affecting their after-tax income. In our stylized example participants would also be allowed to decrease the budget to, for instance 0 and this would imply an increase in private income through a reduction of taxes. On the other hand, respondents can also increase the budget which allows them to select more or even all the 16 projects. However, this would result in a tax increase and a reduction in their private income. Another difference between WTAPB experiments and PVE is that participants in a PVE allocate public budget to a portfolio of projects and, as a result, they can consider positive and negative synergies between projects and potential spatial equity concerns (Mouter et al., 2021). This is not possible in a WTAPB experiment where respondents select only one project from a limited set of projects. Hence, PVE allows individuals to express a broader set of preferences than WTAPB experiments (e.g., positive and negative synergy effects, preferences of individuals who believe that it is better to do nothing and/or reduce taxes instead of allocating public budget to one of the proposed projects).

The innovations of the PVE appraisal method are described in various papers (Dekker et al., 2019; Mouter et al., 2020, 2021). Mouter et al. (2021) explain that PVE can not only be considered as a valuation method which provides input for a CBA, but that the outcomes of a PVE can also be directly used for establishing the social welfare effects of government policy options. In that case, PVE can be considered as a full-fledged alternative to CBA. Dekker et al. (2019) present the technical details of such a welfare analysis. Mouter et al. (2021) also establish the extent to which CBA and PVE provide different policy recommendations and finally they aim to generate empirical insights into potential reasons why PVE and CBA produce different outcomes.

At present, only one large scale PVE has been conducted for the evaluation of transport projects (Mouter et al., 2021). In this PVE 2498 inhabitants from the Transport Authority Amsterdam were presented with 16 transport projects and related societal impacts. The total costs of the 16 projects was 386.5 million euros but with only 100 million euros to spend, it was not possible for the respondents to include all projects in their portfolio. A link to the demo version can be found via www.burger-begroting.nl and an English language version is available via <http://burgerbegroting.tbm.tudelft.nl/participatory-value-evaluation-transport-authority-amsterdam>. Mouter et al. (2021) show that in a PVE the societal value of transport projects is estimated using behavioral choice models which assume that participants aimed to select a portfolio of transport projects that in their view represents the portfolio which maximizes their “utility”. It is assumed that part of the value of an individual project is defined by the impacts that were explicitly presented to participants for each of the transport projects (e.g., reducing travel time, change in the number of traffic deaths) which is reflected in taste parameters (this is comparable to stated choice surveys which estimate the value of travel time savings and the value of a statistical life). Moreover, it is assumed that the (un)attractiveness of an individual project can also be defined by other considerations which is captured in project specific parameters. The authors find that projects which focus on improving traffic safety perform relatively good in the PVE, whereas car projects perform relatively good in the CBA analysis.

Multicriteria Analysis

The MCA, also known as multiple-criteria decision-making (MCDM) or multiple-criteria decision analysis (MCDA) typically encompass the following stages: (1) identification of criteria against which to test transport policy options; (2) weighting or scoring the different criteria to arrive at a ranking of options. A deep variety of MCA methods has been developed over the last years (Dean, 2018). A certain class of MCA approaches can be defined by the set of rules establishing the nature of options, objectives, criteria, scores and weights as well as the way in which objectives/criteria, scores and weights are used to assess, compare, screen in/out or rank options.

A distinction can be made between “sophisticated MCA methods” and “simple MCA methods”. Sophisticated methods use advanced mathematical principles and procedures to weigh criteria and rank options. One example of such a method is the Analytic Hierarchical Process (AHP) (Saaty, 1990). AHP begins by arranging the elements of the analysis in three main hierarchical levels. The overall goal of the decision-making problem at the top (e.g., satisfaction with a transport policy option); a set of decision criteria in the middle layers (e.g., factors that influence the satisfaction of transport policy options); and a group of competing options at the bottom (e.g., transport policy options). Next, the relative importance of each criterion with respect to the goal of the analysis is determined through a series of pairwise comparisons of criteria. The subjective judgments of experts or policy makers regarding the relevance of the different criteria are translated into a quantitative score and subsequently the most desirable option can be determined. Another example of a sophisticated MCA technique is the best-worst method (BWM) developed by Rezaei (2015). Similar to AHP, the BWM method starts with determining a set of decision-criteria. A key difference with AHP is that in the second stage decision-makers participating in a BWM are asked to select the best and the worst criteria after which they should determine the preference of the best criterion over all the other criteria and the preference of all the other criteria over the worst criterion. Based on these choices both the direction and the strength of the preferences of one criterion over the other are computed which provides the basis of selecting the best alternative.

Simple MCA methods use relatively rough procedures to score options or even abstain from scoring and ranking the options. One example is the UK Appraisal Summary Table which does not aim to provide a final ranking of the options. Simplified MCA techniques are very popular, mainly due to practicality reasons (Dean, 2018). MCA studies can be undertaken either in nonparticipatory (i.e., expert-led) or participatory manner. In nonparticipatory assessments, the analysis is carried out autonomously by one or more experts, according to a typical technocratic approach. A key argument in favor of this approach is that a group of scientists and trained experts is better suited to make complex decisions than the average citizen. By contrast, participatory techniques adopt a more collaborative decision-making style, with the direct involvement of the different interested and affected parties (i.e., problem stakeholders) in the analysis (Macharis and Bernardini, 2015). Although participatory MCA techniques have been championed by transport scholars in the literature it is not totally clear whether techniques have enjoyed real-world applications or constitute mere academic proposals (Dean, 2018).

CBA, PVE, and MCA are quite similar in several respects. The three appraisal methods all aim to provide policy makers with information to assess the desirability of a transport policy option and the methods all rely on transportation model results. The most important difference between CBA and PVE on the one hand and MCA on the other hand is that welfare economics provides the theoretical framework underlying CBA and PVE, whereas MCA methods are not built on this theoretical framework. CBA and PVE are both based on welfare economics which provides strict procedures for the objects which have standing in the CBA analysis, for the criteria/impacts that are considered in the analysis and for the way different impacts are valued. MCA analysts, on the other hand, have a larger degree of freedom when selecting criteria and developing procedures to determine the weights (e.g., Macharis and Bernardini, 2015). For instance, a MCA analyst can decide to include distributional aspects and public support for a transport policy option as separate criteria which affect the final outcome of a MCA. Both impacts cannot affect the final indicators of a CBA that is grounded in the Kaldor–Hicks efficiency criterion. PVE assumes that individuals define the attractiveness of a transport policy option based on standard impacts of a transport project such as reductions in travel time and accident risk (which are reflected in taste parameters) as well as other criteria that might not be on the radar of policy makers and experts (which are reflected in project specific parameters). Hence, PVE includes distributional impacts such as spatial equity (Mouter et al., 2021) and there is no fundamental difference between PVE and MCA when it comes to the inclusion of impacts in the evaluation.

Moreover, the weighting procedure that is used in CBA to evaluate the impacts of a transport policy option is very clear in the sense that the only criterion that defines the weight that should be assigned to an impact concerns the amount of money individuals are willing to pay from their private income. The same holds for a PVE in the sense that transport policy option are valued based on the willingness of individuals to allocate public budget to (the impacts of) the transport policy option. The weighting of impacts/criteria in a MCA can be partly based on translating impacts/criteria into monetary units, but the aggregation is also based on at least one other weighting method (e.g., scoring, ranking or weighting of a wide range of qualitative impact categories and criteria).

In sum, the three methods differ in terms of degrees of freedom for the analyst to select impacts and criteria that are relevant for the evaluation as well as weighting procedures. A CBA analyst has to follow strict procedures of welfare economics when selecting impacts and weighting procedures and as a result the degrees of freedom are limited; a PVE analyst has more liberty when selecting impacts although participants in the PVE (and not the analysts) eventually determine the relevance of a criterion for the evaluation. As stated above, a PVE analyst has no degrees of freedom when selecting a weighting procedure. Finally, a MCA analyst has large degrees of freedom in terms of selecting impacts and weighting procedures. The relatively large degree of freedom can be seen as a strength of MCA when benchmarked against CBA and PVE. For instance, Gühnemann et al. (2012) asserts that MCA seems to be better in measuring intangibles and soft impacts than CBA as these effects do not have to be converted into private income. For this reason, MCA has been used in many countries as complementary to CBA to capture impacts not properly accounted for by the latter method (e.g. Mackie et al., 2014).

Although the flexibility of MCA can be seen as a strength, this characteristic of the evaluation method has been criticized for the arbitrariness in the selection of the weights applicable to different criteria (Annema et al., 2015). Qualitative assessment and the imputation of value-laden weightings to different criteria may lead to subjective biasing. This arbitrariness in MCA has often been used by politicians to justify already preadopted decisions, which has undermined the rigor of the MCA appraisal method (Dobes and Bennett, 2010).

Environmental Impact Assessment

An Environmental Impact Assessment (EIA) is a comprehensive evaluation of the likely effects of a project that significantly affect the environment. It provides decision makers with an indication of the likely environmental consequences of their selected policies (Jay et al., 2007). Since the 1970s EIA has become increasingly more important in planning practice and has been introduced in national legislation worldwide (Cornero, 2010). Due to the legal requirements for EIA implementation, there are now strict guidelines for the EIA process (Cornero et al., 2010). The usual information contained in an EIA report is: (1) a description of project alternatives and a baseline description of the environment where the project is located; (2) a prediction of potential effects of the project on the environment; (3) measures envisaged to avoid, prevent or reduce the effects on the environment. The most important differences between EIA and the three methods that were previously discussed (CBA, PVE, and MCA) is that the EIA puts emphasis on a specific set of effects (environmental effects). A second difference is that an EIA also includes recommendations regarding the mitigation and management of negative environmental impacts. The purpose of the other methods that were discussed is not to provide such recommendations.

Social Impact Assessment

Social Impact Assessment (SIA) was introduced in the late 1970s in the United States due to the perception of EIA being an assessment method with a strong biophysical bias (Morgan, 2012). The primary purpose of a SIA is to bring about a more sustainable and equitable biophysical and human environment (Vanclay, 2003). A SIA includes the processes of analyzing, monitoring and managing the intended and unintended social consequences of planned interventions (e.g., Mottee and Howitt, 2018). Vanclay (2003) describes the core values of the SIA community and a set of principles to guide SIA practice. The role of SIA goes far beyond the ex-ante (in advance) prediction of adverse impacts and the determination of who wins and who loses (Vanclay, 2003). Esteves et al. (2012) argue that SIA practitioners also believe that there should be an emphasis on enhancing the lives of vulnerable and disadvantaged people, and in particular, that there should be a specific focus on improving the lives of the worst-off members of society. Hence, the normative postulations of SIA are more in line with Rawlsian ethics which differs from the utilitarian philosophy which undergirds CBA.

SIA is an anthropomorph methodology in the sense that analysts consider all issues that affect people. The main difference with CBA (and PVE) is that all social impacts are considered in a SIA and not only to the extent that humans are willing to pay (or willing to allocate public budget) for the impacts of the transport project. As a result, the degrees of freedom for the analyst to include impacts is relatively large likewise the MCA.

SIAs utilizes participatory processes to analyze the concerns of interested and affected parties. It heavily involves stakeholders in the assessment of social impacts, the analysis of alternatives, and monitoring of the planned intervention. The role of stakeholders in the analysis is much larger compared to the other evaluation methods discussed. MCA and EIA aim to involve stakeholders more and more, but stakeholder participation is an end of SIA in itself (Esteves et al., 2012). A typical SIA consist out of various steps examples being the identification of interested/affected people, selecting alternatives, assessing the social impacts, developing mitigation measures, developing strategies for dealing with residual or nonmitigatable impacts and proactively developing better outcomes (Vanclay, 2003). Hence, the focus of concern of SIA is a proactive stance to development and developing better outcomes, not just the identification or amelioration of negative or unintended outcomes.

Conclusion

This article surveyed five evaluation methods that are used in transport policy and planning: Cost-Benefit Analysis (CBA), Participatory Value Evaluation (PVE), Multi-Criteria Analysis (MCA), Environmental Impact Assessment (EIA) and the Social Impact Assessment (SIA). The main difference between the CBA and PVE is that PVE establishes the desirability of government projects based on people's advises regarding the allocation of the public budget toward (impacts of) government projects, whereas CBA establishes the desirability of government projects through analyzing people's trade-offs between their private income and impacts of government projects. The most important difference between CBA and PVE on the one hand and MCA on the other hand is that CBA and PVE are both based on welfare economics which provides strict procedures for the criteria/impacts that are considered in the analysis and for the way different impacts are weighted, whereas MCA analysts, have a larger degree of freedom when selecting criteria and developing procedures to determine the weights. The most important difference between EIA and SIA and the other three methods is that EIA and SIA put emphasis on a specific set of effects (environmental or improving the lives of the worst-off members of society). A second difference is that an EIA and SIA also include recommendations regarding the mitigation and management of environmental and/or social impacts.

Biography



Niek Mouter is an assistant professor in Infrastructure Project Appraisal at Delft University of Technology, the Netherlands, faculty Technology, Policy and Management. His research primarily revolves around two research lines: (1) The inclusion of equity and other ethically important considerations in the appraisal of infrastructure projects in general and Cost-Benefit Analysis in particular; (2) Improving the usability of Cost-Benefit Analysis for policy makers. He (co)developed a novel assessment approach called 'Participatory Value Evaluation'.

References

- Ackerman, F., Heinzelring, L., 2004. *Priceless: On Knowing the Price of Everything and the Value of Nothing*. The New Press, New York.
- Allen, M.T., Austin, G.W., Swaleheen, M., 2015. Measuring highway impacts on house prices using spatial regression. *J. Sustain. Real Estate* 7 (1), 83–98.
- Annema, J.A., Mouter, N., Rezaei, J., 2015. Cost-benefit analysis (CBA), or multi-criteria decision-making (MCDM) or both: politicians' perspective in transport policy appraisal. *Transport. Res. Proc.* 10, 788–797.
- Boadway, R.W., Bruce, N., 1984. *Welfare Economics*. B. Blackwell, New York.
- Cornero, A., 2010. Improving the implementation of Environmental Impact Assessment. Report. European Environmental Agency.
- Dean, M., 2018. Assessing the Applicability of Participatory Multi-Criteria Analysis Methodologies to the Appraisal of Mega Transport Infrastructure. Ph.D. Dissertation. The Bartlett School of Planning, University College London, UK.
- Dekker, T., Koster, P.R., Mouter, N., 2019. The economics of Participatory Value Evaluation. Tinbergen Institute Discussion papers 19-008/VIII.
- Dobes, L., Bennett, J., 2010. Multi-Criteria Analysis: Ignorance or Negligence?. Paper presented at the 'Australasian Transport Research Forum 2010', Canberra, Australia, 29 September – 1 October.
- Esteves, A., Maria, Franks, D., Vanclay, F., 2012. Social impact assessment: the state of the art. *Impact Assess. Project Apprais.* 30 (1), 34–42.
- Gühnemann, A., Laird, J.J., Pearman, A.D., 2012. Combining cost-benefit and multi-criteria analysis to prioritise a national road infrastructure programme. *Transport Policy* 23, 15–24.
- Hensher, D.A., Rose, J.M., Ortúzar, J. de D., Rizzi, L.I., 2009. Estimating the willingness to pay and value of risk reduction for car occupants in the road environment. *Transport. Res. Part A* 43 (7), 692–707.
- Jara-Díaz, S.R., 2007. *Transport Economic Theory*. Elsevier Science, Amsterdam.
- Jay, S., Jones, C., Slinn, P., Wood, C., 2007. Environmental impact assessment: retrospect and prospect. *Environ. Impact Assessment Rev.* 27 (4), 287–300.
- Macharis, C., Bernardini, A., 2015. Reviewing the use of multi-criteria decision analysis for the evaluation of transport projects: time for a multi-actor approach. *Transport Policy* 37, 177–186.
- Mackie, P.J., Jara-Díaz, S.R., Fowkes, A.S., 2001. The value of travel time savings in evaluation. *Transport. Res. Part E* 37, 91–106.
- Mackie, P.J., Worsley, T., Eliasson, J., 2014. Transport appraisal revisited. *Res. Transport. Econ.* 47, 3–18.
- Morgan, R.K., 2012. Environmental impact assessment: the state of the art. *Impact Assess. Project Apprais.* 30 (1), 5–14.
- Mottee, L.K., Howitt, R., 2018. Follow-up and social impact assessment (SIA) in urban transport-infrastructure projects: insights from the Parramatta rail link. *Australian Planner* 55 (1), 1–11.
- Mouter, N., 2017. Dutch politicians' use of Cost-Benefit Analysis. *Transportation* 44 (5), 1127–1145.
- Mouter, N., 2018. A critical assessment of discounting policies for transport cost-benefit analysis in five European practices. *Eur. J. Transport Infrastructure Res.* 18 (4), 389–412.
- Mouter, N., van Cranenburgh, S., van Wee, B., 2017a. An empirical assessment of Dutch citizens' preferences for spatial equality in the context of a national transport investment plan. *J. Transport Geogr.* 60, 217–230.
- Mouter, N., van Cranenburgh, S., van Wee, G.P., 2017b. Do individuals have different preferences as consumer and citizen? The trade-off between travel time and safety. *Transport. Res. Part A* 106, 333–349.
- Mouter, N., Koster, P.R., Dekker, T., 2020. An introduction to Participatory Value Evaluation. Working paper Tinbergen Institute 19-024/V.
- Mouter, N., Koster, P., Dekker, T., 2021. Contrasting the recommendations of participatory value evaluation and cost-benefit analysis in the context of urban mobility investments. *Transp. Res. Part A Policy Pract.* 144, 54–73.
- Mouter, N., Ojeda Cabral, M., Dekker, T., van Cranenburgh, S., 2019. The value of travel time, noise pollution, recreation and biodiversity: a social choice valuation perspective.. *Res. Transport. Econ* 76, 100733.
- Nyborg, K., 2012. *The Ethics and Politics of Environmental Cost-Benefit Analysis*. Routledge.
- Rezaei, J., 2015. Best-worst multi-criteria decision-making method. *Omega* 53, 49–57.
- Saaty, T.L., 1990. How to make a decision: the analytic hierarchy process. *Eur. J. Oper. Res.* 48, 9–26.
- Sagoff, M., 1988. *The Economy of the Earth: Philosophy, Law, and the Environment*. Cambridge University Press.
- Sen, A.K., 1979. Utilitarianism and welfarism. *J. Philos.* 76 (9), 463–489.
- Sen, A.K., 1995. Environmental evaluation and social choice: contingent valuation and the market analogy. *Jpn. Econ. Rev.* 46 (1), 23–37.
- Vanclay, F., 2003. International principles for social impact assessment. *Impact Assess. Project Apprais.* 21 (1), 5–12.

Transport and Air Quality Planning and Policy

Dr. Fabio Galatioto, Connected Places Catapult, Milton Keynes, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	236
The Role of Planning and Policies	236
Links Between Transport Planning, Policies and Air Quality	237
New Concepts, Methods and Tools to Provide New Insights	237
Exposome Paradigm	237
The PASTA Project	237
Modeling	238
The “New Normal”, Opportunities and Potential Impacts	238
Active Transport and Planning Role	238
Conclusions	239
Biography	239
See Also	239
References	239
Further Reading	240

Introduction

Transport systems are capable of producing significant benefits; however, they have produced over previous decades many negative externalities. For that, appropriate policies need to be devised to maximize its benefits and minimize its inconveniences. In urban areas, although many contributory sources to air pollution exist, road transport, in particular, continues to account for a significant proportion of emissions of all the main air pollutants (with the exception of SO_x) which has increased overtime. In 2017, in Europe the non-exhaust emissions of PM_{2.5} accounted for 46% of emissions from the road transport sector, compared with 18% in 2000, while for PM₁₀, the contribution increased from 32% in 2000 to 63% (EEA, 2017). In the UK, transport is responsible for approximately half of the NO_x emissions. In addition, these emissions tend to be emitted close to where people live and work, making the potential health impact even greater, by increasing the population exposure to higher levels. While emissions from road transport are mostly exhaust emissions arising from fuel combustion, including volatile organic compounds (VOCs), carbon monoxide (CO), and hydrocarbons (HCs) contributing to the formation of ground-level ozone and photochemical smog in the atmosphere, non-exhaust emissions contribute to both NMVOC (from fuel evaporation) and primary PM. An important consideration related to PM nonexhaust pollution in particular is that road transport contributes also from tyre- and brake-wear and road abrasion, which will still be generated even shifting 100% into ultra-low carbon vehicles such as full electric.

These two road-transport based particles sources make up 12% of the smallest and most dangerous type of PM, which can enter the bloodstream via the lungs. For Europe, the WHO estimate for the number of premature deaths attributed to air pollution is over 500,000 (WHO Europe, 2018), with 400,000 early deaths in the EU28. As a result, future strategies and policies should consider carefully the long terms benefits but more importantly the potential indirect or underestimated impacts. It is acknowledged that the allocation, design, and construction of transport infrastructure and services must be subject to careful planning, both by public and private agencies (Rodrigue, 2020).

The Role of Planning and Policies

First, a differentiation between policy and planning must be drawn, since the former usually relates to strategies and goals while the latter refers to concrete actions. Also, both must reflect the fundamental changes in society and contemporary issues and problems, hence policies and planning should constantly change.

- Transport policy deals with the development of a set of constructs and propositions that are established to achieve specific objectives relating to social, economic, and environmental conditions, and the functioning and performance of the transport system.
- Transport planning deals with the preparation and implementation of actions designed to address specific problems.

Over recent years, however, the strategic and wider impacts of planning choices have been difficult to capture, and overtime can cause additional elements of impact that year on year may result in air quality hotspots and local issues at pinch points.

There is no doubt that to plan for sustainable cities, we must consider the effect of transport policy on the environment in terms of air quality and its subsequent effect on health. As suggested by Engardt et al. (2011), there are various factors influencing air quality

within a planning horizon of 20–30 years: the global effect of climate change as well as local effects of urban form need to be considered. At the local level, the design of transportation systems and transport policies will influence the city's own contribution to air pollution. The travel choices made by inhabitants not only determine their contribution to urban emissions, but also influence their own health, which in turn determines the health sustainability of society as a whole. Empirical evidence from the last decade has shown that the characteristics of city and regional land use and transport planning are directly linked with air quality, individual exposure, uptake of pollutants and subsequent health outcomes. See for example [de Nazelle et al. \(2011\)](#), [Schindler and Caruso \(2014\)](#), and [Sider et al. \(2013\)](#).

It is important then, that local plans must commit to a compelling and clearly expressed place-based vision that has sustainable transport as well as health, climate change and environmental needs integrated from the start. Also, strategic and local plan producers must create collaborative partnerships with strategic stakeholders, transport service providers, and local communities going beyond statutory consultation ([CHIT, 2019](#)).

Links Between Transport Planning, Policies and Air Quality

Within cities can be observed considerable variation in the levels of environmental exposures to air pollution, noise, temperature variations and green spaces. Emerging evidence ([Nieuwenhuijsen, 2016](#)) suggests that urban and transport planning indicators such as road network, distance to major roads, and traffic density, household density, industry and natural and green space explain a large proportion of that variability. Personal behavior, including mobility, adds further variability to personal exposures, determines variability in green space and exposure, and can provide increased levels of physical activity.

Decision-makers need not only better data on the complexity of factors in environmental and development processes affecting human health, but also enhanced understanding of their linkages to be able to know at which level to target their actions. New research tools, methods and paradigms such as geographical information systems, smartphones, and other GPS devices, small sensors to measure environmental exposures and remote sensing are expected to provide crucial information for next generation of data driven plans and policies. Same data are expected to provide rich sources for the depth assessment of impacts and exposure, using new approaches such as the exposome paradigm ([Vrijheid, 2014](#)) together with citizens and cities observatories information.

While in cities there are often silos of urban planning, mobility and transport, parks and green space management, environmental management (public) health departments that tend to work largely independent and rarely together, multisectorial approaches are needed to tackle the environmental problems ([Gerike et al., 2016](#)).

For example, [Olowoporoku et al. \(2010\)](#) used a longitudinal study of the links between Local Air Quality Management and Local Transport Planning policy processes in England and observed that while substantial improvements in policy integration were happening within the selected case studies, it demonstrated that such improvements are often constrained by institutional complexities that create implementation gaps between national objectives and local decision-making outcomes.

New Concepts, Methods and Tools to Provide New Insights

In this section, the aim is to provide some inspiration looking at two alternative and innovative methods that recently have been proposed respectively to improve the Planning and Policies to achieve more sustainable and low polluted cities.

Exposome Paradigm

To get away from studying the “one exposure, one health outcome” associations, the new paradigm of “exposome” has been developed by [Vrijheid et al. \(2014\)](#). The paradigm envisages complex multilevel pathways and interactions with other environmental, socioeconomic, social, behavioral and life-style factors, and genetics. The exposome encompasses the totality of human environmental (i.e., nongenetic) exposures from conception onwards, complementing the genome ([Wild CP, 2005; 2012](#)). The dynamic nature of the exposome presents one of the most challenging features of its characterization. The increased use of new technologies including geographical information systems (GIS), sensors, remote sensing, combined with more traditional approaches has made possible to start assessing the exposome theory. First attempts have been made in large European projects such as HELIX (<http://www.projecthelix.eu>), EXPOSOMICS (<http://www.exposomicsproject.eu>), and HEALS (<http://www.heals.eu>).

The PASTA Project

The Physical Activity Through Sustainable Transport Approaches EU project, completed in October 2017, by using a mixed-method and multi-level approach that review interventions to improve outdoor air quality and public health, supported the development of policies aiming at increasing the physical activity of European citizens in the urban environment. Those resulting policies were consistently tested and applied in seven case study EU cities. A large-scale longitudinal survey involving 14,000 respondents was also carried out to assess the health impact of active mobility, with the final goal to inform the WHO's online Health Economic Assessment Tool for the health benefits of cycling and/or walking ([Gerike et al., 2016](#)).

Modeling

From a modeling point of view, a planning model should support the analysis of the effectiveness of policy decisions in terms of different aspects, such as environmental, economic, and health impact, taking into consideration users behavior response and its subsequent effect on the environmental and public health.

In the past, different studies have tried to link strategic transport planning models with air quality models, for example, [Affum et al. \(2003\)](#), [Boogaard et al. \(2012\)](#), [Hatzopoulou and Miller \(2010\)](#) or quantify the health impacts from transport related physical activity as well as changes to air pollution in [Woodcock et al. \(2013\)](#) applying an Integrated Transport and Health Impact Modelling Tool (ITHIM). More recently promoting smart, green and integrated transport has become the key priority in the strategic move towards sustainable development by the European Union and many other countries worldwide. This is reflected by the numerous studies on designing, modeling and promoting green transport systems (e.g., [Bao et al., 2017](#); [Kitthamkesorn et al., 2016](#); [Lin and Liao, 2016](#)); also, environmental constraints and efficiency as well as sustainability have been considered explicitly in many transport network design studies (e.g., [Li et al., 2014, 2016](#); [Maheshwari et al., 2016](#); [Yang et al., 2017](#)).

This however, has gradually changed the traditional and general approach taken by most transport planning and air quality models, where traffic information from the transport planning model, including traffic flow, speed and vehicle types on each roadway, are taken as input into air quality models. However, this traditional approach has shown in recent years several limitations and modeling usually represents scenarios based on assumptions and over time those assumptions can change drastically, requiring new model inputs and settings. Emerging approaches, such as the Agent-based modeling tools seems to go on the right direction and provide the right flexibility of the approach to be quickly adapted to changes.

Moreover, with recent pandemic events, also, more than ever, integration of planning, strategies and air quality modeling should aspire to solve and embed some of the existing identified gaps ([Wang and Connors, 2018](#)):

1. Wider access to health impact metrics;
2. The existing health impact should include both physical activity benefits and pollutant uptake;
3. Include localized effect on spatial pattern of air quality as a result of transport choices of residents;
4. More accurate exposure encountered by travelers on different modes at different locations throughout their trip; and
5. Exposure estimated based on a combination of users physical activity associated with different modes as well as the spatial pattern of air quality concentrations.

Those gaps have become particularly important since the development of the recent COVID-19 pandemic, which is likely to change in the long-term people behavior and the way we all work and live. This will have implications on the assumptions made on modeled scenarios for the next 5–10 years, which will definitely require a revisit and reassessment exercise. Similarly unexpected behavioral changes will emerge that may affect user choices, from the purchase of the house in more rural areas, to the impact on times and volumes of movements impacting the transport network.

The “New Normal”, Opportunities and Potential Impacts

As anticipated, recent events (COVID-19 pandemic), have definitely accelerated transitions that may have happened anyway or introduced completely new behaviors which will have a huge impact over next 5–10 years, such as massive increase of on-line shopping and remote working, the latter also causing unprecedented reduction of patronage in the public and shared transport, that although may be temporary may have a huge impact on the use of private motorized modes and hence increase of air quality problems, particularly in urban areas and occasionally in suburban areas and towns.

Another interesting emerging evidence is related to the impacts of the pandemic and lockdown on traffic volumes and economy. If we take as an example the UK, we have observed that although there was a visible large reduction of traffic volumes in April 2020, around –75% ([WAZE, 2020](#)), the economic impact within the same period has been –20.4% ([ONS, 2020](#)). Although 20% economic shrink was unprecedented, it is very interesting to observe the huge difference in the two reductions. This suggests that not all traffic on our roads was and is responsible or contributing to the economic growth and directly support the economy, but actually may contribute heavily to increase pollution and worst air quality without actually producing tangible economic benefits.

This is why it is important in the future to better understand and link transport movements to economy and identify those elements (e.g., commuting trips) that if removed (via remote working) will not impact the economy almost at all but produce enormous benefit especially in terms of urban pollution and overall transport GHG (Green House Gas) emissions. This capability should be embedded into future integrated models to allow a more holistic view and assessment of the benefits and costs both economic and environmental.

Active Transport and Planning Role

Before recent events, many studies and policies have suggested that more active travel such as walking and cycling would be beneficial to reduce short distance trips hence reduce local pollution and carbon emissions, but limited uptake in the large majority of cities and towns was observed. However, since March 2020 more and more frequently we have observed positive changes in terms of behavior of commuters and citizen. After the various lockdown and pandemic events, many people have started to consider and

move using active transport, making enormous progress in the shift towards this type of mode. In particular, while public transport is still subject to restrictions for social distancing, active transport choices in cities and towns need to be carefully planned and implications considered, in particular in terms of safety and security, generating virtuous circles. For example, a more pleasant roadside environment encourages walking and cycling. If more people choose to walk or cycle instead of drive, air pollution goes down and with it comes more demand for active travel infrastructure.

Planning then becomes a very important factor in the shift of citizens behaviors and choices, and the current situation offers an unprecedented opportunity to accelerate this process and make sure at planning and policy levels active transports are encouraged and prioritized, but that requires also a rebalance of the mobility infrastructure investment.

As indicated in the modeling section, in order to make sure models can truly represent the cost and benefit of the different choices including factors such as exposure, physical activities and long-term impacts of air pollution will help, especially in the long-term, to prioritize the best and more environmentally/sustainable form of transport systems and supporting infrastructures.

A final consideration is that among all modes of transport, active travel options are truly carbon free, while ultra low carbon or zero-carbon modes and options (electric vehicles, hydrogen, etc.) may have low carbon emissions at tail-pipe but they have embedded carbon that in most cases can offset the local benefits. For that many studies are suggesting an increasing importance of accounting for whole life carbon emissions to compare the greenhouse gas emissions among different options of low carbon vehicles and modes. Knobloch et al. (2020) for example have highlighted that based on a global average, and on current trajectories (half of all cars on the road could be electric vehicles by 2050), electric vehicles would achieve a 30% reduction in emissions per kilometer traveled, with this saving expected to increase as the adoption of renewable energy technologies grows. This, although is an important reduction, shows that to achieve net-zero carbon targets by 2040 or even 2050, policies and planning need to work together to make sure truly zero-carbon options such as active travel are facilitated and thought since the planning stage.

Conclusions

To conclude, this contribution has tried to highlight how transport and air quality planning and policy cannot anymore be separated and to truly achieve important and consistent results in the long-term, we all need to consider the different levels of complexity from large to local scale analysis and impact assessment. This ability in some ways does not exist currently and national authorities should focus on supporting initiatives that can develop and validate integrated approaches where planning, policies and impact assessment are equally important. It is acknowledged that this is not an easy task, but it is a key step that needs to be initiated to make sure truly sustainable and healthy cities will exist in a near future.

Biography



Dr. Fabio Galatioto is Senior Transport and Environmental specialist at Connected Places Catapult. Fabio has background in Transport Engineer and Modeling with almost 20 year experience in Transport and Environmental Monitoring and modeling, Air Quality, Sustainability & Impact Assessment and Traffic Safety. Since joining the Catapult in 2015 Fabio has been working to support SMEs and Academic partners and deliver innovative transport and environmental projects both at National and European Level.

See Also

Emerging Trends in Transport Demand Modeling in the Transition Toward Shared Mobility and Autonomy; Transport Planning and Management and its Implications in Chinese Cities; Transport Policy and Governance; Transport and Climate Change.

References

- Affum, J., Brown, A., Chan Y., 2003. Integrating air pollution modelling with scenario testing in road transport planning: The TRAEMS approach. *Sci. Total Environ.* 312 (1–3), 1–14.
- Bao, J., Xu, C., Liu, P., Wang, W., 2017. Exploring bikesharing travel patterns and trip purposes using smart card data and online point of interests. *Netw Spatial Econ.* 17 (4), 1231–1253.
- Boogaard, H., Janssen, N., Fischer, P., Kos, G., Weijers, E., Cassee, F., van der Zee, S., de Hartog, J., Meliefste, K., Wang, M., Brunekreef, B., Hoek, G., 2012. Impact of low emission zones and local traffic policies on ambient air pollution concentrations. *Sci. Total Environ.* 435, 132–140.
- CHIT, 2019. Better planning, better transport, better places. https://www.ciht.org.uk/media/10218/ciht-better-planning-a4_updated_linked_.pdf.

- De Nazelle, A., Nieuwenhuijsen, M., Anto, J., Brauer, M., Briggs, D., Braun-Fahrlander, C., Cavill, N., Cooper, A., Desqueyroux, H., Fruin, S., Hoek, G., Panis, L., Janssen, N., Jerrett, M., Joffe, M., Andersen, Z., van Kempen, E., Kingham, S., Kubesch, N., Leyden, K., Marshall, J., Matamala, J., Mellios, G., Mendez, M., Nassif, H., Ogilvie, D., Peiro, R., Perez, K., Rabl, A., Ragettli, M., Rodriguez, D., Rojas, D., Ruiz, P., Sallis, J., Terwoert, J., Toussaint, J.-F., Tuomisto, J., Zurbier, M., Lebre, E. 2011. Improving health through policies that promote active travel: a review of evidence to support integrated health impact assessment. *Environ. Int.* 37 (4), 766–777.
- Engardt, M., Johansson, C., Gidhagen, L., 2011. Web services for incorporation of air quality and climate change in long-term urban planning for Europe. *IFIP Advances in Information and Communication Technology* 359 AICT:558–565.
- EEA, 2017. National emissions reported to the Convention on Long-range Transboundary Air Pollution (LRTAP Convention). European Environment Agency.
- Gerike, R., de Nazelle, A., Nieuwenhuijsen, M., et al., 2016. Physical activity through sustainable transport approaches (PASTA): a study protocol for a multicentre project. *BMJ Open* 6 (1).
- Hatzopoulou, M., Miller, E. 2010. Linking an activity-based travel demand model with traffic emission and dispersion models: Transport's contribution to air pollution in Toronto. *Transp. Res. Part D: Transp. Environ.* 15 (6), 315–325.
- Kitthamkesorn, S., Chen, A., Xu, X., Ryu, S. 2016. Modeling mode and route similarities in network equilibrium problem with go-green modes. *Netw. Spatial Econ.* 16 (1), 33–60.
- Knobloch, F., Hanssen, S., Lam, A., et al., 2020. Net emission reductions from electric cars and heat pumps in 59 world regions over time. *Nat. Sustain.* 3, 437–447, doi:10.1038/s41893-020-0488-7.
- Lin, J.-J., Liao, R.-Y. 2016. Sustainability SI: Bikeway network design model for recreational bicycling in scenic areas. *Netw. Spatial Econ.* 16 (1), 9–31.
- Li, Z.-C., Wang, Y.-D., Lam, W.H.K., Sumalee, A., Choi, K. 2014. Design of sustainable cordon toll pricing schemes in a monocentric city. *Netw. Spatial Econ.* 14 (2), 133–158.
- Li, T., Wu, J., Sun, H., Gao, Z. 2016. Integrated co-evolution model of land use and traffic network design. *Netw. Spatial Econ.* 16 (2), 579–603.
- Maheshwari, P., Khaddar, R., Kachroo, P., Paz, A., 2016. Dynamic modeling of performance indices for planning of sustainable transportation systems. *Netw. Spatial Econ.* 16 (1), 371–393.
- Nieuwenhuijsen, M.J., 2016. Urban and transport planning, environmental exposures and health-new concepts, methods and tools to improve health in cities. *Environ. Health* 15, S38, doi:10.1186/s12940-016-0108-1.
- Olowoporoku, O., Hayes, E.T., Leksmono, N.S., Longhurst, J.W.S., Parkhurst, G.P., 2010. A longitudinal study of the links between Local Air Quality Management and Local Transport Planning policy processes in England. *J. Environ. Plan. Manag.* 53 (3), 385–403, doi:10.1080/09640561003613179.
- Rodrigue, J.-P., 2020. In: *The Geography of Transport Systems*. 5th edition. Routledge, New York, ISBN 978-0-367-36463-2, p. 456.
- Schindler, M., Caruso, G. 2014. Urban compactness and the trade-off between air pollution emission and exposure: lessons from a spatially explicit theoretical model. *Comput. Environ. Urban Syst.* 45, 13–23.
- Sider, T., Alam, A., Zukari, M., Dugum, H., Goldstein, N., Eluru, N., Hatzopoulou, M. 2013. Land-use and socio-economics as determinants of traffic emissions and individual exposure to air pollution. *J. Transp. Geogr.* 33, 230–239.
- Vrijheid, M., Slama, R., Robinson, O., Chatzi, L., Coen, M., van den Hazel, P., et al., 2014. The human early-life exposome (HELIX): project rationale and design. *Environ. Health Perspect.* 122, 535–544.
- Vrijheid, M. 2014. The exposome: a new paradigm to study the impact of environment on health. *Thorax* 69, 876–878.
- Wang, J.Y.T., Connors, R.D., 2018. Urban Growth, Transport Planning, Air Quality and Health: A Multi-Objective Spatial Analysis Framework for a Linear Monocentric City. *Netw. Spat Econ.* 18, 839–874, doi:10.1007/s11067-018-9398-x.
- Wild, C.P., 2005. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol. Biomark. Prev.* 14, 1847–2150.
- Wild, C.P., 2012. The exposome: from concept to utility. *Int. J. Epidemiol.* 41, 24–32.
- WHO Europe, 2018. Over half a million premature deaths annually in the European Region attributable to household and ambient air pollution. Available on-line: <http://www.euro.who.int/en/health-topics/environment-and-health/airquality/news/news/2018/5/over-half-a-million-premature-deaths-annually-in-the-european-region-attributable-to-household-and-ambient-air-pollution> [Accessed 2020].
- Woodcock, J., Givoni, M., Morgan, A. 2013. Health impact modelling of active travel visions for England and Wales using an integrated transport and health impact modelling tool (ITHIM). *PLoS ONE* 8 (1), e51462.
- Yang, X., Ban, X.J., Ma, R., 2017. Mixed equilibria with common constraints on transportation networks. *Netw. Spatial Econ.* 17 (2), 547–579.

Further Reading

- Banister, D., 2002. *Transport Planning*. Routledge, Abingdon/Oxford.
- Egan, N., 2019. Why Transport planning is vital to improve air Quality. <https://airqualitynews.com/2019/07/24/why-transport-planning-is-vital-to-improving-air-quality/#:~:text=Road%20transport%20is%20the%20largest,of%20combustion%20in%20the%20engine>.
- Woodcock, J., Edwards, P., Tonne, C., Armstrong, B.G., et al., 2009. Public health benefits of strategies to reduce greenhouse-gas emissions: urban land transport. *Lancet* 374 (9705), 1930–1943.

Cycling Policies

Esther Anaya-Boig, Centre for Environmental Policy, Imperial College London, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Historical Background	241
What are Cycling Policies About?	241
What are Cycling Policies For?	242
How are Cycling Policies Made?	243
What Works?	244
Summary and Conclusions	244
See Also	244
References	245
Further Reading	245

Historical Background

The history of cycling for transport is not geographically homogeneous, but there were some important events that happened in specific places in the US, the UK and other countries in Europe that influenced mobility at a global scale. Cycling policy probably started in the last quarter of the 19th century; when cycling spread all over the world and began to be used for mobility more than just for sports and leisure. Towards the end of that century, bicycles became affordable enough to reach the middle-classes and even played an important role in social movements such as the women's suffrage movement. At the same time and within very similar groups of users, cycling was critically contributing to the birth of the motoring industry transferring its knowledge in mechanical design and literally paving the way for the cars to arrive and spread all over the world (Reid, 2015). Around 1920s, motor cars arrived, and cycling became the poor people's transport. Thus started the era of automobility, the dominance of the motor car, not only for transport, but also as a social, economic, and cultural icon (Pooley, 2020). "Peak bicycle" in many cities happened between the 1930s and 1950s, variations depending on the impact of World War II (and other conflicts, such as the Spanish Civil War) in automobile and bicycle supplies. Some European cities' maps of the 1930s include cycling infrastructure, acknowledging that cycling was integrated in transport planning (Oldenziel et al., 2016). After the war, cars took over the streets very quickly, becoming especially prominent in wealthy cities and countries where the automobile industry flourished.

But in some places, cycling continued to be fully present as a mode of transport. These were places where economic resources were not sufficient for a notable uptake of cars (some big cities in the global south had a prominent use of bicycles until the 1990s) or where disruptive economic events and social movements created "perfect storms" against the hegemony of the car. In these places, cycling approaches were integrated back again in transport policies, but these were not easy nor rapid processes. These were the cases of The Netherlands and Denmark where cars took over from the 1950s until the early 1970s when automobility was disrupted by the energy crisis (1973) and questioned by civil society. The pressure of social movements was especially prominent in Amsterdam, in response to the high number of children killed in traffic crashes. Even in these cases, proper official cycle planning only arrived more than a decade later, in the 1980s. The remains of the 1930s cycling infrastructure were the starting point in Copenhagen and from there, cycling provision and practice increased until present times (Oldenziel et al., 2016).

Due to their specific historical contexts, "cycling paradises" in Denmark and The Netherlands might have a longer tradition of cycling policy than other places but it is not completely clear yet what should be done in other places to obtain similar cycling mobility levels.

What are Cycling Policies About?

Cycling policies are about cycling mobility. The term mobility is preferred to transport, with the aim of moving beyond the traditional narrow understandings of transport and towards a social approach (Banister, 2008; Urry, 2008). Very often the main dimension of cycling mobility that policies tend to be concerned with is infrastructure. These are the policies in relation to planning and designing cycle tracks and lanes, speed management and road safety. Perhaps correspondingly, the most cited literature in the field of cycling research focuses road safety, infrastructure and the role of the environment on behavior (Dill, 2019). This is despite of one of the most cited papers in cycling research already making the point one decade ago that increases in cycling require packages of different, complementary interventions—being infrastructure just one of them (Pucher et al., 2010).

It could be useful to think about integrated cycling policies, those in which the multiple dimensions of cycling mobility are taken into account. These dimensions would of course include cycling infrastructure as characterized above, but also (based on Anaya-Boig, 2021):

- Regulations: law and regulations set the conditions in which cycling can take place. They are both part of the context and instruments of policy that can be approved, modified and enforced. In spite of the relevance of regulations, the academic production on this dimension limits to a few studies about road safety and vehicle-related standards.

- Training and coaching: cycling mobility skills need to be learned; these are knowledge transfer policies. They are often part of the suggestions in cycling infrastructure and behavior studies but have been very scarcely studied in the literature.
- Governance: the interactions between different actors within the practice of policy making determine the outcomes of such policies. Some authors make a difference between participation and the more empowering and fair interaction of democratic planning.
- Communication: relates to the way cycling mobility is communicated in policy making, the discourse: the representation of cyclists, the language with its conscious and unconscious meanings and the information communicated by any other signs. The words used to designate people who cycle or who drive, or the images featured in a policy document, create meaning in relation to cycling mobility.
- Social and cultural movements: eminently disregarded in policy, cycling cultures are an essential source of information in order to understand cycling mobility and be able to project it into the future through planning and policies. Social and cultural movements are diverse, complex, and contextual. The history of cycling mobility in a specific place (as described in the introduction to this chapter) or the social movements created around the practice of cycling (the critical mass, for example) would be part of this dimension.
- Planning instruments: this dimension compiles all the documents and manifestations of cycle planning or strategies. Some institutions produce specific cycling instruments, others include cycling within their mobility planning, for example Sustainable Urban Mobility Plans (SUMPs). But it is also important how cycling mobility is integrated in other thematic planning instruments related to energy, the environment, health, economy, tourism, and others.

Incomplete approaches to cycling policy by the institutions involved in policy making might be influenced by a techno-centric approach (based on technological solutionism), in which engineering is the top discipline and the rest have not often been invited to join the debate. Interdisciplinary work is thus a suggested way forward (Walta, 2018).

What are Cycling Policies For?

Cycling policies aim at making cycling accessible. It is preferable to use “accessibility” rather than “promotion” as the goal of cycling policies. The term “promotion” has been profusely used in relation to cycling policy, but it comes from the marketing field, and it is meant to persuade people to buy a product. In promotion, it is not so important who the “buyers” are and whether the benefits (and burdens) of the product are reaching all of them and in which way. But research shows that cycling is not available for everyone who could benefit from it; women, low-income populations, lower educated people, migrants, refugees, ethnic groups are under-represented amongst cyclists. A marketed product might not be available for people with different needs and in a variety of contexts—they just cannot buy it. There are assumptions made by the promoters of the product (cycling) that create exclusion. For example, sportive, white, middle class men, might promote vehicular cycling as an approach to cycling infrastructure (i.e., sharing the road with motorized vehicles in all situations and for everyone). Some transport planners tend to plan for “people like them” (Ralph and Delbosc, 2017). Research has shown, though, that women, elderly and other groups in the population do not feel safe enough to practice cycling mobility in that way and prefer protected infrastructure (Aldred and Dales, 2017). The lack of social and cultural diversity in decision-making teams involved in cycling policies may be creating noninclusive outcomes and cycling inequalities.

Aiming at “accessibility” allows cycling policies to be universal but targeted at the same time—the use of the terms such as universal and targeted are borrowed from current debates in health policy (Carey and Crammond, 2017; Marmot, 2020). “Universal” refers to the aim of making cycling accessible for everyone. “Targeted” cycling policies mean that individuals with different personal and contextual factors might require a different combination and level of provision of cycling policies in order to have access to cycling (also referred to as “proportional”). A universal approach in itself would only be egalitarian (the same for everyone) but adding a targeted element would integrate proportionality, hence providing a more equitable outcome. An example of the tensions between universal and targeted approaches in cycling mobility are illustrated by migrant people and refugees arriving in a new location. The financial situations of these population groups are very precarious, and they need to be mobile, especially for job-seeking and caring purposes, with the lowest possible cost. Let us imagine they arrive in The Netherlands, the best possible environment to cycle, where cycling is actually the default mobility option; but they have not learnt how to cycle. Cycling would be of great help economically and socially for migrants and refugees but the fact that there is infrastructure does not really help that much. They need specific training programs, taking into account their specific needs, to provide them with access to cycling mobility (Mohammadi, 2019; van der Kloof et al., 2014).

The question about the purpose of cycling policies at the beginning of this section is thus very much linked to another one: “Who are cycling policies for?”, and both can be connected through accessibility. Not everyone should cycle but everyone should have access to cycling, even more so those who might be able to cover basic needs and have better opportunities by doing it. Making cycling accessible for all means that the effects of cycling policies need to be fairly (not equally) distributed. As a result, distributive justice requires a wider understanding of accessibility than used traditionally in transport planning (Pereira, 2018). Accessibility consists in reaching places and opportunities from a given location; cycling accessibility entails being able to do it cycling for the whole trip or combining it with other suitable and affordable means, for example, public transport. Access to cycling is not only provided by the proximity to infrastructure and other facilities, but also it requires to be affordable, to have the skills to do it, information, an appropriate vehicle; that is to say it depends on personal, social, cultural, and environmental characteristics. Ideally,

cycling policies would aim at providing access to cycling through a variety of dimensions (as the ones listed above), so that the population would then be able to select and combine those that enable them to cycle in their personal and contextual situations.

How are Cycling Policies Made?

Policy making often activates after realizing there has been a change in the situation and there is a need to control it, to make it safe and secure and to regulate its terms and conditions. This might have been part of the process that happened historically when a technological innovation such as the bicycle was introduced, as described in the first section. In relation to cycling mobility, these changes have intensified over the past couple of decades with public bike-sharing schemes, e-bikes, private dockless bike-share schemes and the use of e-scooters for mobility all over the world—all of which have been grouped under the new term “micromobility.” After this kind of changes, it takes time to respond with new policies. However, radical changes or disruptions can accentuate the need for policies and thus create greater opportunities for faster implementation (Marsden et al., 2020).

In the year 2020, a major disruption caused by the SARS-CoV-2 global pandemic has deeply affected mobility, causing the immobility of a considerable part of the population with lockdowns and quarantines; and the mobility of essential workers, most of them being exposed to high risk of infection in their job places. As population started to show with their practices, the safest way to move in this situation was perceived to be the individual transport. Population demands called for a response from policy makers. Some of the response and recovery policies consisted of increasing the provision of space for cycling. In some places, funding has been made available for the construction of cycle lanes. In other places, dedicated space for cycling has been implemented in a “tactical urbanism” fashion, that is, low cost, provisional, and pop-up cycle lanes. This situation brings to light some specific issues in relation to the production of cycling policy; one is about the distribution of space and the other is about governance. Additionally, there are the ongoing challenges of translation, that is bridging cycling research and policy making.

Space is one of the key resources that mobility policies need to manage. Cycling happens in a specific surface, that can be dedicated to cycles or shared with pedestrians or with other vehicles. This is usually street space in which cycles can be used or parked. Street space is contested in many ways, revealing that there are power relations about the use of space. Space for mobility is unequally distributed, dramatically skewed toward automobility. Cars occupy the public space for parking and traveling in a very inefficient way, and automobility is responsible for a considerable fraction of air pollution, traffic deaths and injuries and other undesirable impacts. Cycling mobility policies necessarily deal with the use of space; how much space is there for cycling, how is this space, where is it located, how is it used, how is it regulated. Thus, cycling policies inescapably need to challenge automobility, which has accumulated a disproportionate amount of street space over the last 70–90 years. The negotiation of space for cycling is at the core of cycling policies, with the outcomes of this struggle often being very disappointing. Conscious or unconsciously biased arguments in favor of keeping space for automobility can distract negotiators from what the real stakes are: accessibility. Although access to cycling is not only achieved through space, space is necessary to access cycling mobility. With the pandemic crisis, a heavier and urgent argument in favor of cycling mobility creates a window of opportunity to shift the balance and reclaim space for cycling. Automobiles are also perceived as pandemic-safe individual transport mobility options although their negative impacts persist in this situation and risks related to air pollution have even increased. The tension between both is maintained as the pandemic continues.

Governance seeks to disperse the state power amongst all kinds of public, voluntary and private organizations with which the government interacts (Marsden and Reardon, 2018). In the negotiations described above, power relations are preventing a more equitable outcome, which could be understood as a governance failure. Literature shows concerns over the social equity and the contribution to sustainability of cycling governance processes framed as participatory and with the goal of “promoting” “smart” cycling policies (Gironés and Vrščaj, 2018; Nikolaeva et al., 2019). Some authors propose a shift toward the democratization of cycle policy and planning in which the power is more balanced and communities can be empowered and feel more ownership of cycling policies (Sagaris and Arora, 2017). Teams of policy makers and planners should be multidisciplinary and inclusive, welcome and combine multiple expertise and points of view in order to break the above-mentioned tendency to plan for “people like them” (Ralph and Delbosc, 2017).

There are different institutions involved in cycling policy making; governments and consultancies are in the frontline. Advocacy groups, lobbies and platforms, and civil society should be involved through the governance process. Consultancies are usually the main source of technical knowledge, they are commissioned and supervised by governments. But in places where cycling levels are low, cycling mobility technical knowledge and even research, is often placed in advocacy groups, as described in the example of the Copenhagen first cycling plan, which was drafted by the cyclists’ federation in the 1970s and unofficially implemented reluctantly by the city council during the 1980s (Oldenziel et al., 2016). The process of formalizing, legitimizing, and institutionalizing cycling mobility knowledge is at different stages in different countries. In contrast, places with a longer tradition in cycle planning, may have formal training programs and approved technical manuals based on scientific evidence. Over the last decades, there has been a process of knowledge transfer from cycling advocates who have turned into technical consultants, city officials and researchers. However, there is still a gap between the different kinds of institutions that is preventing the circulation of knowledge. Policy makers cannot access research for diverse reasons (research is behind paywalls, lack of information literacy, lack of time to read academic papers, unfamiliarity with the formats and language, etc.). Evidence-based technical manuals for policy makers are still scarce but very useful for the translation of the scientific evidence into infrastructure design (see e.g., Parkin, 2018). On the other hand, researchers still seem to be assuming that producing the evidence is enough to convince policymakers. There is a lack of mutual exploration, scientists do not explore how policy making works so that they produce appropriate evidence and outcomes and policy

makers do not approach research because they do not perceive research outcomes as useful or accessible for them. Government and academia need to find ways to collaborate and establish a dialogue so that the activities of both institutions are more aligned, and for them to find ways to communicate more effectively.

What Works?

One of the most frequent questions in cycling research is why people cycle. Studies have found that both objective and perceived built and social environments are important. That is to say, infrastructure is important but the perception of, for example, comfort is important as well. The same happens with road safety and social support; objective measures of both show their relevance in cycling behavior but so does the perception of them—which does not necessarily correlate with the objective measure—. It is also known that cyclists feel safer in protected infrastructure and in roads with reduced speed limits. Protected infrastructure is even more important for certain groups to feel safe when cycling, like women and elderly people. In summary, factors like social background, attitude towards cycling, travel distance, infrastructure quality, and bicycle practice have been found to be important (Walta, 2018). But these findings are from studies undertaken in specific places and what they found for those places is not necessarily applicable in other places. Furthermore, historical and cultural contexts are usually not taken into consideration in the majority of these studies (Cox and Bunte, 2018). Hence, it should be questioned to what extent these findings can be generalized.

The reality is that there are big differences in the uptake of cycling mobility in different countries and even between cities within the same country. There is a tendency to assume that whatever seems to support people cycling in a specific place, could potentially make people cycle in other places. Cycling is increasing globally, and some experts argue that this is because noncycling countries have “adopted” cycling policies from cycling countries (Pucher and Buehler, 2017). Whether adoption is an adequate term in these cases has not been proved; cities do not necessarily acknowledge where they have transferred their cycle policies from, and mechanisms behind the transferability of cycling policies have not been made explicit nor sufficiently studied yet. Even if policy makers in a city claim that they are implementing “Dutch-style” cycling policies (e.g., infrastructure designs), they will never look and work exactly in the same way as they did in the original location where they were observed—if they were actually observed in any way. Other authors suggest to lower the intensity of the relation with the original policies by using inspiration rather than pretending to copy (Vigar, 2017). Transferability can support the design and implementation of cycling policies, but is it possible to know enough about a specific cycling policy in order to transfer it successfully? Multiple questions arise in parallel; why is that specific policy selected to be transferred? how is success defined? In transferability it is common to use the term “best practice”, but the evaluation of the original policies can be shallow and politically manipulated (Macmillen and Stead, 2014; Vigar, 2017). One further question related to transferability is about how important is finding that inspiration outside rather than locally. Some authors actually suggest that the way forward in cycling policy transferability is to realize local potential (Walta, 2018), which seems to support the importance of taking into account nonexpert knowledge (Attard and Enoch, 2011; Vigar, 2017) through more democratic governance structures (Sagaris and Arora, 2017). Transferability has been practiced across all transport policies, but it seems that cycling policy making has higher hopes in it, probably due to the lack of knowledge of policy makers and practitioners in places where cycling levels are low.

Summary and Conclusions

Cycling policies vary across cities and countries, as levels of cycling mobility do. The importance of historical and cultural contexts to understand the levels of cycling mobility is under-estimated. The object of cycling policies, cycling mobilities, is often reduced to the infrastructural dimension, whereas there are other dimensions that need to be integrated such as regulations, training and coaching, governance, communication, social and cultural movements and planning instruments. In order for cycling to be inclusive, the goal of cycling policies needs to shift from sheer promotion to making cycling accessible for everyone. The universality of cycling policies combined with a targeted approach for specific population groups allows more equitable outcomes. Mobility disruptions that reinforce the advantages of cycling mobility create windows of opportunity for policy makers to break through political barriers and reclaim the space that has been taken over by automobility. Democratized governance practices and inclusive teams of policy makers could produce more equitable cycling policies. Siloed institutions block knowledge transfer; more collaboration between policy-makers, researchers and the civil society is needed to better understand cycling mobility, balance power relations and co-create fairer policies. It is assumed that cycling policies from the “cycling paradises”—places that have high levels of cycling mobility—can be transferred to increase cycling levels elsewhere. However, the mechanism of cycling policy transfer has not been studied accurately enough, but rather relies on assumptions and is manipulated with political interest. There might be excessive expectations placed in cycling policy transfer that can be strengthened through re-valuing local potential, connecting all sources of knowledge and leveling power relations in the governance structures.

See Also

Cycling Economics; Bicycle and E-Scooter Safety; Bicycle Collision Avoidance Systems; Bicycle Infrastructure; Bicycles: the Safety of Shared Systems vs. Traditional Ownership; Bicycles for Urban Freight; Bicycle Sharing; Bicycle Sharing/Bikesharing; Road Modes; Cycling; Mobility as a Service (MAAS); Bicycle Sharing; Active Travel; Ride Sharing; Cyclists; Bike Sharing and Health; Electric Bikes and Health; Indirectly; The Politics of Mobility Policy; Gendered Mobility; Equity Considerations in Transport Planning

References

- Aldred, R., Dales, J., 2017. Diversifying and normalising cycling in London UK: an exploratory study on the influence of infrastructure. *J. Transp. Health* 4, 348–362, doi:10.1016/j.jth.2016.11.002.
- Anaya-Boig, E., 2021 Integrated Cycling Policy. A framework proposal for a research-based cycling policy innovation, In: Zuev, D., Psarikidou, K., Popan, C., (Eds.), *Cycling Societies: Innovations, Inequalities and Governance*. Routledge.
- Attard, M., Enoch, M.P., 2011. The role of policy transfer in the introduction of road pricing in Valletta, Malta. *Transport Policy* 8 (2), 544–553.
- Banister, D., 2008. The sustainable mobility paradigm. *New Dev. Urban Transp. Plan.* 15, 73–80, doi:10.1016/j.tranpol.2007.10.005.
- Carey, G., Crammond, B., 2017. A glossary of policy frameworks: the many forms of 'universalism' and policy 'targeting'. *J. Epidemiol. Community Health* 71, 303–307, doi:10.1136/jech-2014-204311.
- Cox, P., Bunte, H., 2018. Social practices and the importance of context. In: Graff, K., Bunte, H., Dziekan, K., Haubold, H., Neun, M. (Eds.), *Framing the Third Cycling Century German Environment Agency and European Cyclists' Federation, Dessau-Rosslau & Brussels*, pp. 122–131.
- Dill, J., 2019. Bicycle research 2019: More than helmets and head injuries. Jennifer Dill PhD. <https://jenniferdill.net/2019/08/14/bicycle-research-2019-more-than-helmets-and-head-injuries/> (accessed 8.7.20).
- Gironés, E.S., Vrščaj, D., 2018. Who benefits from smart mobility policies? The social construction of winners and losers in the connected bikes projects in the Netherlands. In: Marsden, G., Reardon, L. (Eds.), *Governance of the Smart Mobility Transition*. Emerald Publishing Limited, pp. 85–101, doi:10.1108/978-1-78754-317-120181006.
- Marsden, G., Anable, J., Chatterton, T., Docherty, I., Faulconbridge, J., Murray, L., Roby, H., Shires, J., 2020. Studying disruptive events: innovations in behaviour, opportunities for lower carbon transport policy? *Transport Policy* 94, 89–101, doi:10.1016/j.tranpol.2020.04.008.
- Macmillen, J., Stead, D., 2014. Learning heuristic or political rhetoric? Sustainable mobility and the functions of 'best practice'. *Transp. Policy* 35, 79–87, doi:10.1016/j.tranpol.2014.05.017.
- Marmot, M., 2020. Health equity in England: the Marmot review 10 years on. *Bmj* 368.
- Marsden, G., Reardon, L., 2018. Governance of the smart mobility transition. Emerald Publishing Limited, doi:10.1108/9781787543171.
- Mohammadi, S., 2019. Social inclusion of newly arrived female asylum seekers and refugees through a community sport initiative: the case of Bike Bridge. *Sport Soc.* 0, 1–25, doi:10.1080/17430437.2019.1565391.
- Nikolaeva, A., te Brömmelstroet, M., Raven, R., Ranson, J., 2019. Smart cycling futures: charting a new terrain and moving towards a research agenda. *J. Transp. Geogr.* 79, 102486, doi:10.1016/j.jtrangeo.2019.102486.
- Oldenziel, R., Emanuel, M., Bruhèze, A.A. de la, Veraart, F., 2016. Cycling cities: the European experience; hundred years of policy and practice. Foundation for the History of Technology, Eindhoven.
- Parkin, J., 2018. In: *Designing for Cycle Traffic: International Principles and Practice*. ICE Publishing, doi:10.1680/dtct.63495.
- Pereira, R.H.M., 2018. Distributive justice and transportation equity: inequality in accessibility in Rio de Janeiro. University of Oxford.
- Pooley, C.G., 2020. Mobility history of everyday. In: Kobayashi, A. (Ed.), *International Encyclopedia of Human Geography (Second Edition)*. Elsevier, Oxford, pp. 149–154, doi:10.1016/B978-0-08-102295-5.10293-8.
- Pucher, J., Buehler, R., 2017. Cycling towards a more sustainable transport future. *Transp. Rev.* 0, 1–6, doi:10.1080/01441647.2017.1340234.
- Pucher, J., Dill, J., Handy, S., 2010. Infrastructure, programs, and policies to increase bicycling: an international review. *Prev. Med.* 50, S106–S125, doi:10.1016/j.ypmed.2009.07.028.
- Ralph, K., Delbosc, A., 2017. I'm multimodal, aren't you? How ego-centric anchoring biases experts' perceptions of travel patterns. *Transp. Res. Part Policy Pract.* 100, 283–293, doi:10.1016/j.tra.2017.04.027.
- Reid, C., 2015. *Roads Were Not Built for Cars: How Cyclists Were the First to Push for Good Roads & Became the Pioneers of Motoring*, First Edition edition. ed Island Press, Washington, DC.
- Sagaris, L., Arora, A., 2017. Cycling for social justice in democratizing contexts: Rethinking "sustainable" mobilities. In: *Urban Mobilities in the Global South*. Routledge, pp. 19–40.
- Urry, J., 2008. Moving on the mobility turn. In: Canzler, W. (Ed.), *Tracing Mobilities: Towards a Cosmopolitan Perspective*. Routledge, London, p. 13.
- van der Kloof, A., Bastiaanssen, J., Martens, K., 2014. Bicycle lessons, activity participation and empowerment. *Case Stud. Transp. Policy, Social Exclusion in the Countries with Advanced Transport Systems* 2, 89–95, doi:10.1016/j.cstp.2014.06.006.
- Vigar, G., 2017. The four knowledges of transport planning: enacting a more communicative, trans-disciplinary policy and decision-making. *Transp. Policy* 58, 39–45, doi:10.1016/j.tranpol.2017.04.013.
- Walta, L., 2018. On the methodologies and transferability of bicycle research: a perspective from outside academia. *J. Transp. Land Use* 11., doi:10.5198/jtlu.2018.1458.

Further Reading

- Bonham, J., Johnson, M. (Eds.), 2015. *Cycling Futures*. University of Adelaide Press, Adelaide, Australia.
- Caimotto, M.C., 2020. Discourses of Cycling, Road Users and Sustainability: An Ecolinguistic Investigation. Springer International Publishing, Cham. <https://doi.org/10.1007/978-3-030-44026-8>.
- Cox, P. (Ed.), 2015. *Cycling Cultures*. University of Chester Press, Chester.
- Cox, P., Koglin, T. (Eds.), 2020. *The politics of cycling infrastructure: Spaces and (in)equality*. 1st ed. Policy Press, doi:10.2307/j.ctvvsqc63.
- Furness, Z.M., 2010. One less car: bicycling and the politics of automobility. In: *Sporting*, Temple University Press, Philadelphia.
- Horton, D., Rosen, P., Cox, P. (Eds.), 2007. *Cycling and Society*. 1st ed. Routledge, Aldershot, England, Burlington, VT.
- Gerike, R., Parkin, J., 2018. *Cycling futures: from research into practice*.
- Golub, A., Hoffmann, M.L., Lugo, A.E., Sandoval, G.F. (Eds.), 2016. *Bicycle Justice and Urban Transformation: Biking for all?* 1st ed. Routledge, London; New York.
- Longhurst, J., 2017. *Bike battles: a history of sharing the American road*.
- Oldenziel, R., Emanuel, M., Bruhèze, A.A. de la, Veraart, F., 2016. *Cycling cities: the European experience; hundred years of policy and practice*. Foundation for the History of Technology, Eindhoven.
- Parkin, J., 2012. *Cycling and Sustainability*. Emerald Group Publishing Limited, Bingley.
- Parkin, J., 2018. In: *Designing for Cycle Traffic: International principles and practice*. ICE Publishing, doi:10.1680/dtct.63495.
- Pucher, J.R., Buehler, R., 2012. *City cycling*, Edited by John Pucher and Ralph Buehler. ed. In: *Urban and Industrial Environments*, MIT Press, Cambridge, Mass.
- Sheller, M., 2018. *Mobility Justice: The Politics of Movement in an Age of Extremes*. Verso Books.
- Stehlin, J.G., 2019. *Cyclescapes of the Unequal City: Bicycle Infrastructure and Uneven Development*. Univ Of Minnesota Press.
- Vivanco, L.A., 2013. *Reconsidering the Bicycle: An Anthropological Perspective on a New*, 1st ed. Routledge, New York.

Community Severance

Paulo Anciaes*, **Jennifer S. Mindell†**, *Centre for Transport Studies, Civil Environmental and Geomatic Engineering, UCL (University College London), United Kingdom; †Health and Social Surveys Research Group, Research Department of Epidemiology and Public Health, UCL (University College London), United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

What is Community Severance?	246
Causes and Extent of the Problem	246
Direct Effects	248
Wider Effects	249
Behavior of Individuals and Organizations	249
Wider Effects on Individuals	250
Wider Effects on Communities	250
Equity Issues	250
Policies	251
Research	251
Conclusions	252
See Also	252
Relevant Websites	252
References	253

What is Community Severance?

Community severance occurs when transport infrastructure or vehicles using that infrastructure are a physical or psychological barrier to the movement of pedestrians and other modes of local transport, separating people from places and from other people. This may lead to changes in travel behavior and to wider negative effects on individuals, including restricted access to goods, services, resources, and opportunities; higher probability of social exclusion; fewer social contacts; and impaired health and well-being (Anciaes et al., 2015; Hérán, 2011; James et al., 2005; Mindell and Karlsen, 2012; Mindell et al., 2017). It can also lead to effects at the level of the community, including reduced vitality of local businesses, reduced social cohesion, and deterioration in the local environment (Appleyard and Lintell, 1972; Appleyard et al., 1981; Appleyard, 2020; Jacobs, 1961).

The concept of community severance is sometimes described as the “barrier effect” of transport. Although the two concepts are often used interchangeably, barrier effect tends to refer mainly to the loss of mobility and accessibility of pedestrians, while community severance also includes the wider negative effects on individuals and communities (Anciaes, 2015).

Causes and Extent of the Problem

The most extreme cases of community severance are those posed by linear transport infrastructures (such as motorways, dual-carriageway roads, and railways) where longitudinal movement by pedestrians is not allowed and transversal movement is limited to a few locations where grade-separated pedestrian crossing facilities are provided (e.g., footbridges or underpasses). In this type of infrastructure, severance is enforced by embankments, cuttings, or physical barriers (e.g., walls, fences, guard railings) that prevent pedestrians from crossing (Fig. 1). Rivers and canals can also cause severance, although this is usually attenuated by the amenity value they provide.

Severance can also be caused by the vehicles using the infrastructure, even in cases where crossing is physically possible. This is the case of roads with large amounts of motorized traffic, or traffic moving at fast speed (Fig. 2). Pedestrians find it difficult to cross these roads outside designated crossing facilities due to the lack of safe gaps in the traffic. This is aggravated in roads with many lanes; complex road layouts (e.g., roundabouts, large junctions, turning lanes); many large vehicles in the traffic (e.g., buses, heavy goods vehicles); vehicles parked on the curbside; and lack of dropped curbs or a median strip where pedestrians can stop while crossing. Public transport vehicles running on dedicated lanes (including tram and light rail lines, and bus rapid transit systems) may also cause severance.

Crossing facilities often aggravate, rather than reduce, severance. Underpasses are prone to flooding and tend to be perceived as unpleasant and unsafe, especially when they have insufficient lighting or are poorly designed and maintained (Fig. 3). To a smaller degree, these issues also apply to footbridges, which can also be problematic for people with fear of heights (Fig. 4). Using underpasses or footbridges also implies extra effort and detours to use stairs or ramps. In some cases, ramps (or lifts) are not



Figure 1 Severance caused by transport infrastructure (Dubai, United Arab Emirates).



Figure 2 Severance caused by motorized traffic (Shanghai, China).



Figure 3 Severance aggravated by crossing facilities (underpasses) (Chişinău, Moldova).



Figure 4 Severance aggravated by crossing facilities (footbridges) (Valparaíso, Chile).

provided, limiting the use of the facilities by older people and those with mobility impairments. At-grade crossing facilities, such as signalized crossings, also have problems, when waiting time is long or the time allowed for crossing is too short.

Large nonlinear transport infrastructure, such as car parks, railway stations and depots, ports, and airports, may also pose barriers to pedestrians. However, movement across railway stations is sometimes possible—stations can even be one of the few safe crossing points to cross the railway line.

In some cases, severance caused by transport infrastructure is compounded by the presence of other land uses with poor permeability for pedestrians, such as industrial areas, hospitals, and university campuses. This may lead to the isolation of some residential areas, which become enclosed by barriers to pedestrian movement on all sides.

Severance can originate from a variety of processes. For example, a new transport infrastructure may be aligned to cross an existing community. This is often due to geographic and technical constraints, but it can also be due to economic factors (e.g., lower value of the land crossed) or political factors (e.g., lack of political power of local communities, or political bias of decision makers) (Appleyard, 2020). A different process leading to severance is a progressive change to an existing infrastructure, such as the increase of road traffic volumes resulting from urban and economic growth. This is often accommodated by planners with changes to the road design, such as widening the carriageway and reducing the number of pedestrian crossing facilities.

Severance can also originate from the emergence of new residential areas around transport infrastructure crossing previously nonresidential land. This often happens due to urban growth, as previously “undesirable” areas crossed by transport infrastructure become one of the few possibilities for new land developments. The desirability of those areas, translated into land values, may increase because the presence of transport infrastructure increases accessibility by motorized modes. Again, political factors are relevant, as the conversion from nonresidential to residential land depends on land use regulations.

The problem of severance is spread around the world, especially in urban and suburban areas (see, e.g., Future of London, 2018). While some solutions are being implemented in some cities, reviewed later in this chapter, severance is intensifying in others. In large and fast-growing Asian cities, the increase in travel demand has often translated into massive multilevel roads, where the barrier effect is compounded by visual intrusion (Badami, 2009). In many developing countries, growing investment in roads has also failed to address the needs of pedestrians, even where walking is still the main mode of transport (Bradbury, 2014).

Direct Effects

The main direct effects of community severance are the loss of mobility (ability to move) and accessibility (ability to reach places) of pedestrians. In the presence of a barrier, pedestrians must either make a detour to reach crossing facilities (which implies extra effort and delays to walking trips) or cross in a place without facilities (with the associated risk of collision with vehicles). In both cases, the barrier can also be perceived by pedestrians as unpleasant due to exposure to noise, vibration, dust, and air pollution; feelings of intimidation caused by vehicles; and personal security issues of using an environment dominated by motorized vehicles and with few other pedestrians (Hine 1996; James et al., 2005; Mindell et al., 2017). Although most of the existing evidence deals with pedestrians, some of these effects also apply to cyclists. Detours and delays also affect the provision of services such as bus routes, post distribution, freight, waste collection, policing, and emergency services, as well as local trips by private car.

Regardless of the effect on movement across the barrier, there is also a psychological effect of being separated from the other side. This can be explained by the visual intrusion of the infrastructure and auxiliary structures (e.g., walls, fences, noise barriers, crossing facilities) or the noise and air pollution from traffic (Héran, 2011). This psychological effect may change the residents’ perceptions about their local area. Qualitative studies have shown, for example, that when barriers exist, residents will have a narrower “personal

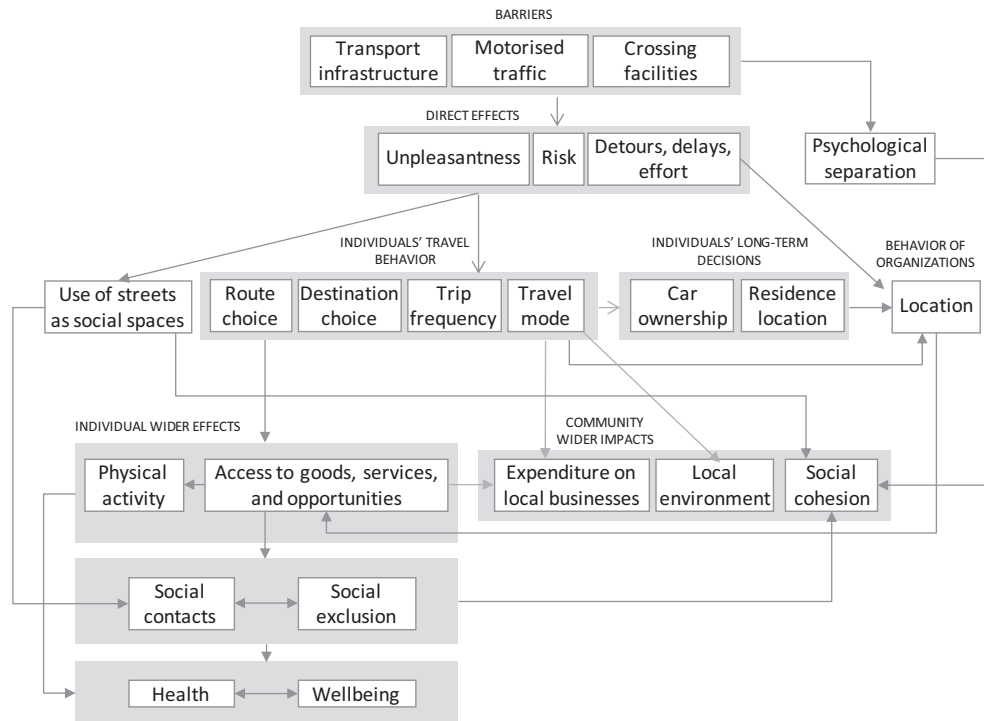


Figure 5 Chain of potential wider effects of community severance.

neighborhood," limited to their side of the barrier (Lassière, 1976). In addition, they will tend to have negative perceptions of the area next to the barrier on their side, and of the whole area on the other side.

However, the existence and extent of the direct effects of severance depend on how the barrier interacts with the local built environment. The separation of a residential area from another residential area or from a commercial or leisure area is more impactful than the separation from an industrial estate or undeveloped land. It is also likely that severance increases with proximity from the barrier, although not linearly, because people living next to the barrier may not dissociate the barrier effect from other effects, such as noise and air pollution. Feelings of severance may also extend far from the barrier, if the barrier disrupts the individuals' movement patterns. This is for example the case when difficult access to bus stops or a train station located on the other side of the barrier affects the frequency and destinations of non-local trips.

Feelings of severance also depend on perceptions about the barrier. The same traffic volumes and speeds are perceived differently by different people at different times of the day (Anciaes et al., 2017). Individuals who use the infrastructure for movement by car or public transport, or as destination for shopping, may not perceive the infrastructure as a barrier. Even underpasses and footbridges can become attractive when they have an innovative design or become shopping areas. Perceptions of severance may also decrease with the amount of time the individual has lived near the barrier. A new barrier disrupts existing mobility patterns and perceptions about neighborhoods but, with time, individuals may adapt to the barrier (Lassière, 1976). Finally, barriers can even be regarded positively if individuals wish to be separated from the communities on the other side.

Wider Effects

The direct effects of severance may lead to a complex chain of wider negative effects, mediated by changes in the behavior of individuals and organizations. These effects may be felt at the level of individuals or the whole community (i.e., residents, workers, and other people using the area). Fig. 5 synthesizes the chain of effects, which are described in detail the subsections below. It should be noted that although all the links in the figure are plausible, they are supported by varying levels of empirical evidence.

Behavior of Individuals and Organizations

The loss of mobility and accessibility due to the barrier is a constraint to the travel behavior of local residents. Individuals may change their routes to avoid the barrier, or change their destinations, going to places on their side of the barrier, rather than those on the other side (Anciaes et al., 2019; Hine 1996; Mindell et al., 2017). They may also reduce the number of trips or change travel mode, replacing walking and cycling with trips by motorized modes (car or bus), even when the destinations across the road are within

walking distance (Powers et al., 2019). The use of streets as social spaces is also likely to decrease (Appleyard and Lintell, 1972; Appleyard et al., 1981). In the longer term, severance may influence decisions such as whether to own a car and where to live.

Businesses and public organizations may also adapt to the presence of the barrier and to the changes in the individuals' travel behavior. Possible effects include more businesses catering to car users (in larger premises with car parks); and facilities being provided on both sides of the barrier, to cater to what have become separate markets. The routes of services such as post and waste removal may also change to reduce detours and delays. In the long term, the separation of communities may contribute to changes in school catchment areas and administrative areas.

Wider Effects on Individuals

The constraints that severance imposes on travel behavior reduce the individuals' ability to access goods and services, when the facilities providing them (e.g., shops and public facilities) are located on the other side of the barrier (James et al., 2005; Mindell et al., 2017). It also affects the ability to access opportunities such as employment, education, leisure, and other people. This may increase the probability of social exclusion and reduce the individuals' network of social contacts (Mackett and Thoreau, 2015).

Health and well-being can be affected through a variety of pathways. This can be a direct effect of the reduction in physical activity (via the suppression of walking trips), social exclusion, and reduction of social contacts; three pathways with solid evidence from public health research (see Mindell and Karlsen, 2012). Reduced accessibility also affects decisions to visit health-supportive destinations such as health centers, public parks, and shops offering healthy food. The stress and anxiety linked to collision risk, psychological effects of separation, and restrictions to access places and people may also affect mental health and subjective well-being.

Wider Effects on Communities

Local businesses are affected by the constraints posed by severance on the number of trips that local residents make and the travel mode they use. Evidence shows that although car users tend to spend more than pedestrians do on a single visit to shopping areas, they also make fewer trips, resulting in an overall negative effect on expenditure in local businesses (Jones et al., 2007). There is also a tendency for areas next to large transport barriers to have little economic activity due to low pedestrian flows (Jacobs, 1961). Severance is also linked to other negative economic outcomes, including income loss (due to constraints in accessing employment and education); depreciation of residential and commercial property prices; costs of extra trips by motorized modes; cost of detours to the routes of services; and congestion due to the increase in local traffic. In rural areas, there is also a loss of efficiency due to the separation of fields.

Social cohesion is also affected by the lower level of use of streets as social spaces; perceptions of separation and reduced social contacts within the community; and social exclusion of some individuals (Appleyard and Lintell, 1972; Appleyard et al., 1981). As noted, severance also influences perceptions of personal security and may even contribute to an increase in crime and vandalism, due to the reduction in the number of people using the streets and people's reduced sense of ownership and responsibility for their neighborhood.

Due to the replacement of walking trips with trips using motorized modes, severance also has environmental effects, both local (e.g., noise, air pollution) and global (e.g., use of nonrenewable resources, emissions of greenhouse gases). Outside urban areas, severance brought about by transport infrastructures and motorized traffic affects the movement of wildlife.

Equity Issues

Severance raises equity issues, linked to its effects on vulnerable groups. Due to collision risk and exposure to pollution on busy roads, adults will be unwilling to let children walk, cycle, or play outdoors unsupervised, affecting the children's independence and mental and physical development (Shaw et al., 2015). Due to their slower walking speeds, older people and those with mobility impairments are also vulnerable when crossing busy roads and may respond by suppressing walking trips (Musselwhite and Haddad, 2010). This restricts independent mobility and social contacts, two of the main components of healthy living for those groups. Older people, as well as women, also tend to be more concerned with personal security when using footbridges, underpasses, and roads dominated by vehicles and with few other pedestrians (James et al., 2005).

Some groups are also constrained in their actions to reduce severance due to the lack of alternative modes of transport, travel destinations, or walking routes. For example, low-income groups face more economic constraints to own and use a private car and may not be able to afford public transport. In most countries, women also tend to have lower levels of access to private vehicles and more demands on their time. In some cases, older people and some ethnic and religious groups are also dependent on local facilities and social contacts. Older people and those with impairments also find it more difficult to make detours to walking trips (Mindell et al., 2017).

Equity issues may also arise due to patterns in residence location of some groups. There is evidence that low-income groups tend to be disproportionately exposed to negative effects of transport such as noise and air pollution. This may be explained by the sorting of income groups to different levels of environmental quality, through housing markets: properties in cleaner and quieter areas have higher prices and rents, and may be unaffordable to low income groups. There is also some evidence of ethnic minorities being

disproportionately affected by noise and air pollution, which may be due to an income effect, segregation, or lack of political power of those groups. It is possible that these patterns also apply to severance, although empirical evidence has only recently started to emerge (Lara and Silva, 2019; Appleyard, 2020).

Policies

The most radical policy intervention to address severance is to remove the infrastructure causing the problem (Bocarejo et al., 2012). The most well-known example is the Cheonggye Expressway in Seoul, a multilevel motorway removed in 2005 to restore a stream and create a linear park. There is solid evidence of the success of this type of intervention, as shown by indicators such as increases in walking trips and in the use of public places, and increases in the value of properties in nearby areas (Kang and Cervero, 2009). However, when the removal of the infrastructure is accompanied by building a bypass in other areas, the problem is simply displaced. Even when bypasses are built in nonresidential areas, market processes may lead to the area becoming residential in the future, recreating severance. Rerouting the infrastructure to be aligned with a river or other natural feature also restricts access and reduces the amenity value of that feature.

Another option is to shift the infrastructure up or down. For example, the Central Artery in Boston was partly buried in a tunnel in 2003 and replaced with parks and a boulevard. This solution is expensive and not always technically feasible. A less expensive solution is sinking, without burying, the infrastructure, which allows for the replacement of grade-separated crossing facilities with at-grade facilities, although this still limits the number of available crossing points. Moving the infrastructure up, using flyovers, removes the barrier to pedestrian movement, but it does not remove, and can in fact increase, the psychological effect of separation, due to visual intrusion.

The barrier to pedestrian movement can also be reduced by adding new crossing facilities, which increase the number of available points where pedestrians can cross, reducing detours and increasing safety. Other options include modifying the type of existing facilities (e.g., replacing footbridges and underpasses with at-grade facilities), changing their location (to align with pedestrian desire lines), or changing their characteristics (e.g., improving the design and maintenance of footbridges and underpasses; and reducing crossing stages and allowing more time to cross in signalized crossings).

Another possibility is to make crossing easier at the points where no formal crossing facilities are provided. In the case of roads, this can be achieved by removing physical barriers; reducing the number of lanes assigned to motorized modes or the width per lane (reallocating the released space to pedestrians); or simply by adding dropped curbs and pedestrian refuges in the middle of the road. Adding a median strip extending along the road also allows pedestrians to walk along it and cross when a suitable gap in the traffic appears. "Courtesy crossings" have also become more common, using design features such as stripes, colored or textured surfaces, visual narrowing of the carriageway, and ramps, to encourage drivers to stop for pedestrians.

Severance can also be reduced by policies to change traffic, rather than the infrastructure. Road traffic volumes can be reduced by regulatory policies to restrict traffic (e.g., closure of roads to motorized traffic, road space rationing, one-way traffic schemes, high-occupancy lanes, removal of parking spaces) or by economic policies such as road pricing. Traffic speeds can be reduced by imposing lower speed limits, enforcing existing limits, or adding traffic calming design features. The problem of these interventions is that traffic volume and speed usually have an inverse relationship, so reducing one may increase the other.

Minor changes in the road environment can also reduce perceptions of severance in busy roads. Examples include improving pavements; removing obstructions; adding dropped curbs; and adding or improving lighting, greenery, street furniture, surveillance, and provision for the mobility-impaired. Even in the cases where crossing is not possible, improving the environment at the places where the local street network meets the road may also improve negative perceptions about the barrier.

Despite their potential benefits, it is not easy to make a political or business case for interventions to remove or reduce severance. This is partly because there are few established methods to objectively identify severance, measure its intensity, and assess the effect of interventions to reduce it (Anciaes et al., 2015). National-level official manuals for the assessment of transport projects often omit severance or mention it in a checklist of social or environmental impacts, without providing practical guidance. In some cases, simple qualitative scales are proposed, with a few loosely defined points. At the city level, only a few governments use quantitative indicators. For example, Transport for London uses a weighted average of number of lanes, traffic counts, and speeds. Methods for the monetization of severance are even rarer and when they exist, they capture only specific aspects of the problem (Anciaes et al., 2015). For example, the official guidance in Germany, Italy, and Australia currently suggests monetizing detours to walking trips by applying unit values of time. The tendency has been for less, rather than more, objectivity in the assessment of severance and in several countries (Sweden, Denmark, Switzerland) detailed methods to quantify and value severance have been dropped from official manuals over the years.

Research

There is a long, if discontinuous, history of research on community severance; the earliest studies dating from the 1950s. Appleyard and Lintell (1972) conducted the first high-impact study, producing a simple but powerful result that people living in roads with higher levels of motorized traffic in San Francisco had fewer friends and acquaintances and made fewer trips across the road, compared to those living in quieter roads. During the late 1970s and early 1980s, there were parallel efforts in several countries, with

governments commissioning comprehensive reports on severance. A study for the Department of Environment in the United Kingdom (Lassière 1976) documented the impact of busy roads on the travel behavior and neighborhood perceptions of local residents. In France, Loir and Icher (1983) looked at the interaction between different types of road infrastructure, urban dynamics, and people's perceptions in creating severance. In Sweden, Korner (1979) systematized the various effects of road barriers into different levels and in Norway, Lervåg (1984) reviewed methods to assess severance.

In the following 25 years, the only studies dealing specifically with severance were a few methodological reports for transport authorities in England (Clark et al., 1991) and New Zealand (Read and Cramphorn, 2001; Tate 1997). Research resumed only in the 2010s, with a new emphasis on interdisciplinarity. In the United Kingdom, the Street Mobility project at UCL (University College London) developed a suite of tools to identify and measure severance caused by busy roads (Mindell et al., 2017). In France, Frédéric Héran published the first book-length study on severance (Héran, 2011), proposing a broad approach to study the problem, integrating it with other transport and urban issues. Evidence on the direct and wider effects of severance has also been growing, with studies replicating Appleyard and Lintell's results in several countries (Hart and Parkhurst, 2011; Wiki et al., 2018). Evidence has also emerged on the impacts of severance on subjective well-being (Anciaes et al., 2019; Foley et al., 2017).

Several assessment methods have been proposed but have seldom been included in routine transport practice. Some of the methods rely on simple indicators of the size of the barrier, obtained by quantifying characteristics of road design (e.g., road width, number of lanes) and traffic (e.g., traffic volumes, proportion of large vehicles, traffic speed). More complex indicators measure the effects of the barrier, such as delays and detours. These indicators are sometimes fed into broader indicators of accessibility to local facilities or catchment areas for walking trips (Van Eldijk, 2019). Unit values are usually multiplied by the size of the affected population, defined as the population living within a certain radius from the barrier. A few authors have proposed indicators based on pedestrians' perceptions of safety (Tate, 1997) and observed road crossing behavior (Russell and Hine, 1996). Surveys and qualitative methods (e.g., participatory mapping) have also been used to capture the perceptions and priorities of local residents (Mindell et al., 2017).

There have been recent developments in methods to monetize the benefits of reducing severance. These methods apply standard approaches for the valuation of nonmarket goods, based on willingness to pay. Stated preference approaches (e.g., Anciaes and Jones, 2020) use surveys capturing people's choices among different scenarios, which allows for the calculation of unit values for the characteristics of road design, traffic, and crossing facilities. Revealed preference approaches (e.g., Kang and Cervero, 2009) estimate the value of severance as reflected in market prices, usually property prices. However, in most cases, revealed preference approaches have treated severance as a yes/no issue (the presence/absence of the road) or in terms of a single characteristic (usually traffic volumes). A problem common to both stated and revealed preference approaches is that they cannot capture the overall cost of severance, especially the costs of the wider effects on individuals and communities.

Despite these developments, research on severance still lags that of other negative effects of transport, such as noise and air pollution, and is still faced with unanswered questions and the lack of robust assessment methods. Research has focused almost exclusively on the effects of road traffic on pedestrians in roads without physical barriers. Little is known about more absolute types of barriers, such as those posed by motorways and by railways, where severance derives mainly from the infrastructure itself, and not from traffic. Research has also not moved beyond the study of "barrier" to that of "enclosure," in areas that are surrounded by barriers on all sides. There is also little knowledge on the effects of severance on cyclists and local motorized traffic. Another important gap, which is becoming a pressing policy issue, is how to address severance in fast-growing in cities in developing countries.

Conclusions

Community severance, or the barrier effect of transport infrastructure and traffic, reduces the mobility and accessibility of pedestrians and users of other modes of local transport, and can lead to changes to travel behavior and to wider negative effects on individuals and communities. The problem is more impactful for children, older people, women, and low-income groups. There are several possible solutions to remove or reduce community severance but few methods for assessing those solutions. Research on severance still lags that of other negative effects of transport, such as noise and air pollution, and available evidence does not yet cover all possible transport barriers and their effects.

See Also

Accessibility; Transportation Equity; Mobility; Pedestrian Crossing (Pelican, zebra, etc.); Pedestrians; Access to Transport; Wellbeing and Travel Satisfaction; Travel Behavior

Relevant Websites

<https://www.ucl.ac.uk/street-mobility/toolkit>
UCL Street Mobility Toolkit to assess community severance

References

- Anciaes, P.R., 2015. What Do We Mean By "Community Severance"? Street Mobility and Network Accessibility Working Paper Series, No.4. University College London, London. <http://discovery.ucl.ac.uk/1527807>.
- Anciaes, P.R., Jones, P., Mindell, J.S., 2015. Community severance: where is it found and at what cost? *Transport Rev.* 36 (3), 293–317.
- Anciaes, P.R., Jones, P., 2020. A comprehensive approach for the appraisal of the barrier effect of roads on pedestrians. *Transport. Res. Part A: Policy Pract.* 134, 227–250.
- Anciaes, P.R., Metcalfe, P.J., Heywood, C., 2017. Social impacts of road traffic: perceptions and priorities of local residents. *Impact Assess. Project Appraisal* 35, 172–183.
- Anciaes, P.R., Stockton, J., Ortegon, A., Scholes, S., 2019. Perceptions of road traffic conditions along with their reported impacts on walking are associated with wellbeing. *Travel Behav. Soc.* 15, 88–101.
- Appleyard, D., Lintell, M., 1972. The environmental quality of city streets: the residents' viewpoint. *J. Am. Inst. Plan.* 38 (2), 84–101.
- Appleyard, B., 2020. *Livable Streets 2.0*. Elsevier, Amsterdam.
- Appleyard, D., Gerson, M.S., Lintell, M., 1981. *Livable Streets*. University of California Press, Berkeley.
- Badami, M.G., 2009. Urban transport policy as if people and the environment mattered: pedestrian accessibility the first step. *Econ. Polit. Weekly* 44, 43–51.
- Bocarejo, J.P., Le Compte, M.C., Zhou, J., 2012. *The Life and Death of Urban Highways*. Institute for Transportation and Development Policy and EMBARQ, New York and Washington.
- Bradbury, A., 2014. Understanding Community Severance and its Impact on Women's Access and Mobility in African Countries. UK Department for International Development, London. <https://www.gov.uk/dfid-research-outputs/understanding-community-severance-and-its-impact-on-women-s-access-and-mobility-in-african-countries-literature-review>.
- Clark, J.M., Hutton, B.J., Burnett, N., Hathway, A., Harrison, A., 1991. The Appraisal of Community Severance. Transport and Road Research Laboratory, Crowthorne.
- Foley, L., Prins, R., Crawford, F., Humphreys, D., Mitchell, R., Sahlqvist, S., Thomson, H., Ogilvie, D., 2017. Effects of living near an urban motorway on the wellbeing of local residents in deprived areas: natural experimental study. *PLoS ONE* 12 (4), e0174882.
- Future of London, 2018. Overcoming London's Barriers. <https://www.futureoflondon.org.uk/knowledge/overcoming-barriers>.
- Héran, F., 2011. La Ville Morcelée - Effets de Coupure en Milieu Urbain. Economica, Paris.
- Hart, J., Parkhurst, G., 2011. Driven to excess - impacts of motor vehicle traffic on residential quality of life in residents of three streets in Bristol UK. *World Transport Policy Pract.* 17, 12–30.
- Hine, J., 1996. Pedestrian travel experiences - assessing the impact of traffic on behaviour and perceptions of safety using an in-depth interview technique. *J. Transport Geogr.* 4, 179–199.
- Jacobs, J., 1961. The curse of border vacuums., in *The Death and Life of Great American Cities*. Random House, New York., Chapter 14, pp. 336–352.
- James, E., Millington, A., Tomlinson, P., 2005. Understanding Community Severance, Part 1: Views of Practitioners and Communities. Report for UK Department for Transport. http://webarchive.nationalarchives.gov.uk/+/http://www.dft.gov.uk/adobe/pdf/163944/Understanding_Community_Sev1.pdf.
- Jones, P., Roberts, M., Morris, L., 2007. Rediscovering Mixed Streets – The Contribution of Local High Streets to Sustainable Communities. Policy Press, Bristol.
- Kang, C.D., Cervero, R., 2009. From elevated freeway to urban greenway: land value impacts of CGC project in Seoul, Korea. *Urban Studies* 46 (13), 2771–2794.
- Korner, J., 1979. Trafikanläggnings Barriäreffekter. Rapport - Trafikplanering, Chalmers Tekniska Högskola, Göteborg.
- Lara, D.V.R.L., Silva, A.N.R., 2019. Equity issues associated with transport barriers in a Brazilian medium-sized city. *J. Transport Health* 14, 100582.
- Lassière, A., 1976. The Environmental Evaluation of Transport Plans. UK Department of Environment, S.I.
- Lervåg, H., 1984. Vegen Som Barriere for Fotgjengere: Metodeskrivelser. NIBR Report 1984:13. NIBR, Oslo. <https://www.nb.no/nbsok/nb/78312bb838822a8d76cb8efff3ecabde?index=3#0>.
- Loir, C., Icher, J., 1983. Les Effets de Coupure de Voies Routières et Autoroutières en Milieu Urbain et Périurbain. CETUR-CETE de Bordeaux, Bordeaux. <http://temis.documentation.developpement-durable.gouv.fr/document.xsp?id=Temis-0002768>.
- Mackett, R.L., Thoreau, R., 2015. Transport, social exclusion and health. *J. Transport Health* 2 (4), 610–617.
- Mindell, J.S., Karlsen, S., 2012. Community severance and health: what do we actually know? *J. Urban Health* 89 (2), 232–246.
- Mindell, J.S., Anciaes, P.R., Dhanani, A., Stockton, J., Jones, P., Haklay, M., Groce, N., Scholes, S., Vaughan, L., 2017. Using triangulation to assess a suite of tools to measure community severance. *J. Transport Geogr.* 60, 119–129.
- Musselwhite, C., Haddad, H., 2010. Mobility, accessibility and quality of later life. *Qual. Ageing Older Adults* 11 (1), 25–37.
- Powers, E.F.J., Panter, J., Ogilvie, D., Foley, L., 2019. Local walking and cycling by residents living near urban motorways: cross-sectional analysis. *BMC Public Health* 19, 1434.
- Read, M.D., Cramphorn, B., 2001. Quantifying the Impact of Social Severance Caused by Roads. Research Report 201. Transfund New Zealand. <https://www.nzta.govt.nz/resources/research/reports/201/>.
- Russell, J., Hine, J., 1996. The impact of traffic on pedestrian behaviour - 1. Measuring the traffic barrier. *Traffic Eng. Control* 37, 16–18.
- Shaw, B., Bicket, M., Elliott, B., Fagan-Watson, B., Mocca, E., Hillman, M., 2015. Children's Independent Mobility: An International Comparison and Recommendations for Action. Policy Studies Institute, London. http://www.psi.org.uk/children_mobility.
- Tate, F.N., 1997. Social Severance. Research Report 80. Transfund New Zealand. <https://www.nzta.govt.nz/resources/research/reports/80/>.
- Van Eldijk, J., 2019. The wrong side of the tracks: quantifying barrier effects of transport infrastructure on local accessibility. *Transport. Res. Proc.* 42, 44–52.
- Wiki, J., Kingham, S., Banwell, K., 2018. Re-working Appleyard in a low density environment: an exploration of the impacts of motorised traffic volume on street livability in Christchurch New Zealand. *World Transport Policy Pract.* 24, 60–68.

Planning for Bus Priority

Claus H. Sørensen*, Fredrik Pettersson†, Joel Hansson†, *K2/VTI, Lund, Sweden; †Lund University, Lund, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	254
Why Bus Priority?	254
Types of Priority Measures	254
What are the Obstacles to Bus Priority?	256
Conflicts Over Space	256
Institutional Complexities	257
How to Manage Bus Priority Planning?	257
Acceptance	257
Collaboration	258
Funding Models	258
Guidelines and Action Plans	258
Concluding Remarks	258
Acknowledgment	259
Biography	259
See Also	260
References	260
Further Reading	260

Introduction

Why Bus Priority?

Well-functioning travel and transport in metropolitan regions is one of society's strategic challenges. Public transport is essential for more sustainable cities because it contributes to efficient use of road space, reduces emissions, and plays an important role in social sustainability. Most urban public transport in cities across the world is road based (Currie, 2016), and the vast majority of buses operate on streets in mixed traffic (Vuchic, 2007).

In mixed traffic, the service speed and reliability depend on traffic conditions. Because buses also need to stop to pick up and drop off passengers, the average speed of the bus services is inevitably lower than the average speed of cars on the same route. Consequently, buses in mixed traffic are in general not very competitive with car travel regarding speed and reliability (Vuchic, 2007). In order to improve the situation, there are several possible measures that prioritize on-road public transport services over other road users. The umbrella term for these measures is *transit priority measures* or, specifically for buses, *bus priority measures*.

The benefits of bus priority can be divided into primary and secondary benefits (Currie, 2016). Primary benefits include passenger travel time savings and increased reliability in the form of improved punctuality or regularity (TCRP, 2013). Secondary benefits stem from the travel time and reliability improvements, and these occur at increasing scales as the primary benefits of bus priority increase (Currie, 2016). One such secondary benefit is fleet and operating cost savings due to more efficient vehicle and personnel cycles. Alternatively, the increase in timetable efficiency can be used for increasing the frequency of service without increasing costs.

Frequency, travel time, and reliability are quality attributes that are of particular importance for customer satisfaction and for the loyalty of public transport passengers as well as for achieving a modal shift in favor of public transport (dell'Olio et al., 2018; Hansson et al., 2019; van Lierop et al., 2018). Thus, modal shift is another secondary benefit of bus priority. After a certain threshold, land use benefits may also be obtained because the capacity of buses well exceeds that of cars. For instance, by replacing a mixed traffic lane with a bus lane the total transport capacity can increase without increasing the road space (Litman, 2016). Finally, bus priority might also be associated with considerable road safety benefits (Goh et al., 2014).

As a result, approaches to justifying bus priority based only on travel time impacts are too simplistic. All primary and secondary benefits need to be considered, as should cost and benefit trade-offs for all road users (Currie, 2016).

Types of Priority Measures

Bus priority measures can be divided into two main categories, namely priority in space and priority in time (Reynolds et al., 2017). The priority in space category includes measures prioritizing buses in the road space, for example, bus lanes, bus gates, and measures related to bus stops. Since the introduction of the first bus lane in Chicago in the late 1930s, various priority in space measures have proliferated throughout the world (Dadashzadeh and Ergun, 2018). The priority in time category is focused on intersections and



Figure 1 Bus lanes can be an efficient use of road space. In this picture there are approximately twice as many passengers in the bus as there are people in the cars occupying the other two lanes of the road. Photo: Västtrafik/Stina Olsson.

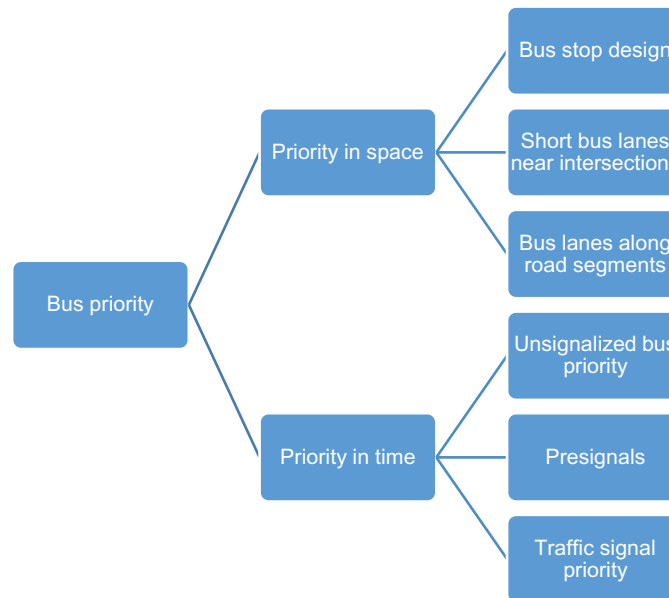


Figure 2 Categories of bus priority measures.

includes bus priority at traffic signals as well as priority for bus movements in unsignalized intersections. Sometimes other concepts are used, for instance, infrastructure strategies and traffic control strategies (TCRP, 2016), road design measures and traffic signal priority measures (Currie, 2016), or facility design-based measures and signal control-based measures (Diakaki et al., 2015). It should be noted that there is a slight difference in the two latter categorizations compared to the priority in space and time categories, and “traffic signal priority measures” and “signal control-based measures” exclude priority measures at unsignalized intersections. Early examples of priority in time measures relying on traffic signal control were implemented in European cities in the 1970s. In line with technological advances, more sophisticated variants of such measures have been developed and disseminated and are now commonplace in many cities around the world (Dinopoulou et al., 2013).

Based on the priority in time and priority in space categories, an example of how bus priority measures can be further arranged in subcategories is shown in Fig. 2. Each of these subcategories in turn contains a variety of priority measures (Currie, 2016; Dadashzadeh and Ergun, 2018; Diakaki et al., 2015; TCRP, 2016; Vuchic, 2007).

For *bus stop design*, the best options in terms of bus priority allow buses to stop directly in the travel lane, thereby also avoiding lane shifts while braking or accelerating. One such option is the bus bulb, a curb extension typically replacing a roadway that would otherwise be part of a parking lane (NACTO, 2016). Bus stop lengthening can also be seen as a priority measure, allowing the bus stop to serve more buses simultaneously and hence avoiding buses queuing to enter the stop.

In Fig. 2, bus lanes are separated in two categories—short bus lanes near intersections and bus lanes along (longer) road segments. *Short bus lanes near intersections* are designed either to avoid queues ahead of the intersections—through queue jump lanes, bypass lanes, or freeway access ramps—or to enable buses to pass through the intersection using exclusive infrastructure—for example, bus gates or right-turn lanes. To also improve speed and reliability between intersections, *bus lanes along road segments* are

important. A number of options regarding design and operations are available and are suitable for different situations as well as for determining the level of separation from other traffic. For instance, curbside bus lanes are usually slower than median bus lanes or fully segregated busways (Vuchic, 2007). The level of priority is also influenced by the separation method such as line markings, colored pavement, or physical barriers. Furthermore, there are operational aspects to take into account, most importantly if the bus lanes should be reserved for buses only or if other vehicle classes should be permitted in the bus lanes, including taxis, bicycles, high-occupancy vehicles, or trucks. The vehicle classes are ordered according to their typical impacts on bus services, and taxis generally have less impact and trucks generally have more impact than other vehicle classes (Vuchic, 2007).

In the priority in time category, measures are focused on intersections. In this category, *unsignalized bus priority* is a simple but effective type of priority measure, meaning that streets with bus services are prioritized and crossing streets have yield or stop signs. This type of priority measure also includes turn restrictions so that turning movements in general traffic that cause delays for buses are prohibited. The movement restrictions may be complemented with exemptions for buses so that buses are allowed to make movements that are prohibited for other vehicles. Turn restrictions and turn restriction exemptions are here categorized as unsignalized priority measures, but they are also possible to implement at signalized intersections.

Presignals are signals before an intersection, where the intersection itself may be signalized or unsignalized. Presignals are a support strategy for bus lanes that end before an intersection. The presignals allow buses to reach the intersection ahead of other traffic even though the bus lane is discontinued (TCRP, 2016).

In signalized intersections, there is a multitude of possible strategies for implementing *traffic signal priority* for buses. These strategies can be divided into passive and active priority measures. Passive signal priority (also called fixed-time or off-line signal priority) means that expected bus movements are considered when designing the signaling system, for example, by weighting the proportion of green time given to a phase that is used by buses. Real-time conditions of the bus operations are not considered in passive signal priority strategies. With active signal priority (also called real-time or traffic-responsive signal priority), buses are detected ahead of the intersection, triggering signal phase adjustments that favor the buses. Examples of such adjustments are green extension, early green, stage skipping, and stage reordering. Active signal priority can be conditional, taking into account the operational status of the bus and whether it is delayed, on time, or ahead of schedule. In contrast, unconditional strategies provide priority regardless of the status of the approaching bus. Furthermore, active signal priority can be reactive or proactive. Reactive strategies are typically applied at each intersection separately, whereas proactive strategies use model predictions to plan for arriving buses when they are one or more signals upstream (Diakaki et al., 2015).

Measures concerning bus operations, aimed at improved travel times and reliability, are also mentioned in connection with bus priority measures, including, for example, bus stop consolidation, route design, and fare payment changes (TCRP, 2016).

In both the priority in space and priority in time categories, the available types of measures can be designed in numerous ways resulting in various levels of bus priority. This is particularly valid for the broad concept of traffic signal priority, which has a large set of parameters and a wide variety of options. The different levels of bus priority in these options are possibly associated with different levels of implementation challenges; the higher the level of bus priority, the more challenging it is to implement. Bus lanes, for instance, are a narrower concept with fewer opportunities for compromise in terms of the level of priority. Typically, the greatest benefits of bus lanes are obtained where competition for space is the greatest and where they are the most difficult to implement.

What are the Obstacles to Bus Priority?

As both Reynolds et al. (2017) and Singerman Ray (2019) point out, there is surprisingly little research on the implementation of bus priority measures. Because bus priority is a central aspect in the planning and implementing of Bus Rapid Transit (BRT), the majority of literature that exists about the process of implementing priority measures is focused on such large scale and often programmatic types of investments. Additionally, the data used in these studies are usually collected in case studies situated in the Global South, for example in Latin America, Asia, and Africa (Lindau et al., 2014; Muñoz and Geschwinder, 2008; Munoz and Paget-Seekins, 2016; Nikitas and Karlsson, 2015; Poku-Boansi and Marsden, 2018).

Literature that describes conditions for the implementation of bus priority measures in other contexts, and where measures are not part of an upgrading to BRT standard, is more limited. Pettersson and Sørensen (2019), who studied the implementation of bus priority measures in “conventional” bus services in Scandinavian cities, concluded that there were both similarities and differences when comparing their results to the BRT literature. A key difference is that large-scale BRT projects often have strategic ambitions, for example relating strategic urban development objectives or ambitions to reform the public transport sector (Muñoz and Geschwinder, 2008; Munoz and Paget-Seekins, 2016; Poku-Boansi and Marsden, 2018). In such policy contexts bus priority measures are parts of high profile, often politically charged policy packages with far-reaching objectives extending beyond the limits of transport system effects. In contrast, Pettersson and Sørensen (2019) and Singerman Ray (2019) highlight the incremental and day-to-day character of implementing bus priority measures in conventional bus services. Here, the endeavors are initiated and carried out mostly by public administration officers.

Conflicts Over Space

Regardless of context and whether part of a BRT package or not, it is clear that the implementation of bus priority measures is difficult per se and is often fraught with conflict. A key conflict concerns the redistribution of space between various road users, which

becomes particularly evident concerning measures categorized as priority in space in the introduction above. Previous research on bus priority measures in a BRT context, for example, Nikitas and Karlsson (2015) and Munoz and Paget-Seekins (2016), often emphasize the conflict involved in reallocating road space from motorists to bus passengers. Removing on-street parking spaces and loading zones to create space for bus lanes is also controversial (Singerman Ray, 2019). Pettersson and Sørensen (2019) also found that the conflict over road space involves other road users than motorists, such as cyclists and pedestrians. Generally, in any city that aims to increase cycling and walking, conflicts between priority measures for buses and measures to increase cycling and walking are likely to occur. Apart from having to address the issue of road space allocation, the conflict between bus priority and nonmotorized transport modes is also about clashing objectives where ambitions to increase the operational speed of buses can be difficult to combine with ambitions to improve traffic safety (Pettersson and Sørensen, 2019).

A second, related conflict is that it can be difficult to speed up bus service while simultaneously facilitating urban development and protecting “urban qualities”. People living in areas served by buses may not consider faster buses as something positive, but rather as a threat to their quality of life. Concerns about increased noise and inadequate road safety along thoroughfares, as well as the elimination of stops, are issues that have engaged the public and impeded the implementation of priority measures. If large-scale infrastructure measures are needed to ensure high and consistent speed for buses, there is also a potential conflict where high value of exploitable land in the cities impedes the implementation of bus priority measures that require a lot of space (Pettersson and Sørensen, 2019).

Institutional Complexities

With the term institutions, we refer to the institutional framework for the formulation and implementation of priority measures. This includes the organizational structure for public transport and the funding structure for bus services. It also includes the division of labor between different public authorities and the formal administrative processes within which planning and implementation of priority measures take place. Furthermore, institutions also include the values and norms of the persons working in different organizations responsible for different aspects of implementing priority measures (March and Olsen, 1989; Rye et al., 2018).

In general, the planning and implementing of bus priority measures can be impeded by institutional complexities. Complexity can, for example, concern ambiguities about which organizations have the formal mandate, and hence the budgetary responsibility, for various measures (Muñoz and Geschwinder, 2008). This, in turn, can result in institutional inertia, that is that it takes (too) long to make the necessary decisions in all the agencies involved or that measures are not implemented at all (Pettersson and Sørensen, 2019; Wright and Hook, 2007).

Thus, an important dimension of institutional complexity in relation to bus priority is linked to funding structures, both for public transport in general and for priority measures per se. As mentioned in the introduction of the chapter, bus priority measures can reduce operating costs. Alternatively, the efficiency gains can be used to increase the frequency of service without increasing costs. However, the funding structure of public transport influences the possibility to recoup the benefits of investments in terms of improved operating economy and more passengers. For many types of priority measures, the costs for implementation are allocated to an organization acting as the infrastructure holder, while the benefits accrue to the organization operating the service. This means that unless there are direct links between the budgets of the infrastructure holder and the operator there will be no direct economic incentives to invest in bus priority (Pettersson and Sørensen, 2019).

The norms and values of officials working in different organizations are also a source of complexity because they can have different views of the problem, and hence divergent views about which types of solutions are justified. One example of how this can impede bus priority is if mobility is measured as flows of vehicles instead of flows of people (Lindau et al., 2014). In a context of congestion, emphasis is often on improving the level of service based on flows of vehicles rather than people, and this way of viewing the problem could hamper the implementation of both priority in space and priority in time measures. As mentioned in the discussion of conflicts about space, officials who work with public transport and those who work primarily with other means of transport, such as cycling, car travel, walking, or more general urban development, often have conflicting views (Pettersson and Sørensen, 2019).

How to Manage Bus Priority Planning?

It appears from the above that implementing bus priority measures is not an easy task, and it is entrenched with difficulties and obstacles. Here we will turn to four “tools” that appear relevant when aiming to introduce bus priority in cities.

Acceptance

Conflicts over space and institutional complexity imply that leadership becomes particularly important, and it is often seen as a key factor for resolving conflicts and gaining consensus (Lindau et al., 2014; Muñoz and Geschwinder, 2008; Munoz and Paget-Seekins, 2016; Singerman Ray, 2019; Wright and Hook, 2007). However, leadership can take many different forms. In the paradigmatic cases of introducing bus priority connected to BRT in the Latin American cities of Curitiba and Bogotá, it was the mayors with great visions of how to transform the city who were the political champions representing the necessary leadership. In studies of the implementation of bus priority measures in Scandinavia (Pettersson and Sørensen, 2019) and the US (Singerman

Ray, 2019), political champions play a less prominent role. When bus priority measures are implemented as small, incremental changes to the existing system, the role of champions within specific agencies seems more important. In such cases political leadership is still important, but in a more reactive manner, for example, by providing support for implementation once it is shown that priority measures are actually working (Singerman Ray, 2019). This highlights that politicians are dependent on the population they represent. Lack of public support for bus priority usually implies that the political leadership is not experienced. This makes it important to involve the public in the processes of planning, designing, and implementing bus priority measures (Pettersson and Sørensen, 2019).

Collaboration

Due to the institutional complexity in implementing bus priority measures, extensive collaboration among many stakeholders is needed for such measures to succeed. This collaboration can be within and among different municipal, state, and regional authorities as well as bus operating companies and other public transport authorities or private operators. Establishing some degree of consensus, coordination, and common understanding among a wide variety of actors is central, and the presentation of a united front of politicians and officials has also been suggested to be important to be able to “sell” the project to the public (Finn et al., 2011).

Through experiences on collaboration in public transport projects in Scandinavia, a number of lessons for successful collaboration have been recognized (Pettersson and Hrelja, 2018). Establishing a shared understanding of the purpose of the collaboration and building enduring relationships seem to be of utmost importance.

Funding Models

While institutional complexity prevails and constitutes an obstacle when planning and implementing bus priority measures, some institutional set-ups are more conducive than others to the establishment of bus priority measures, and funding models seem to be one institutional feature of significant importance.

The funding structures of public transport as well as bus priority measures differ across the world. Often, the local public authorities are responsible for and fund the local road network, and in many cases local or regional public authorities also subsidize the provision of public transport. In such cases, the funding scheme for public bus transport can be set up to incentivize the local authorities to invest in bus priority measures. This is the case, for example, if the local authority fully or partly pays the costs of the provided bus hours after ticket revenues are considered. In that case, implementation of bus priority measures making the bus service faster and more attractive will make it possible for a local authority to reduce bus hours while at the same time increasing ticket revenue from more passengers. Thus, local authorities can recoup the investments in bus priority measures and can establish positive business cases (Pettersson and Sørensen, 2019; Singerman Ray, 2019). In addition, external funding sources, such as state or regional-level support of investments in bus priority measures, can also constitute an important incentive for local authorities (Pettersson and Sørensen, 2019).

Guidelines and Action Plans

Guidelines for bus priority as well as local action plans for the implementation of bus priority measures can constitute another important tool for planning and implementing bus priority measures, and a number of guidelines and toolboxes for introducing bus priority measures already exist. Some of these have been developed in connection to BRT (Finn et al., 2011; The Swedish Knowledge Centre for Public Transport, 2016; Wright and Hook, 2007), while others have been developed in other national or local contexts (e.g., National Capital Region Transportation Planning Board, 2011; TRCP, 2013, 2016).

While guidelines mostly contribute with planning and technical recommendations, action plans can represent the political commitment to introduce or further develop bus priority initiatives. Many cities are working with such action plans, e.g. Greater Wellington in New Zealand (Greater Wellington, 2019), New York City in the US (New York City, 2019), and Stockholm in Sweden (ÅF Infrastructure, 2016). To take Stockholm as an example, the capital of Sweden has a mobility strategy including goals to increase the average speed on the trunk bus lines, and together with Region Stockholm the city has developed an action plan identifying measures that can be implemented in the short term (Pettersson and Sørensen, 2019; ÅF Infrastructure, 2016). Action plans can constitute an important arena for collaboration between different stakeholders involved in implementing bus priority measures.

Concluding Remarks

Although technical guidelines are important tools in the implementation of bus priority measures, such measures are first and foremost a political issue. Implementation of bus priority measures is fraught with conflicts between different societal goals and interests, and solutions always imply a redistribution of mobility options among different groups of the population. Some will gain and others will lose. The barriers for implementing bus priority measures are founded in conflicts between different road user groups, and political leadership is essential for managing the implementation barriers. Across countries and continents, processes of policy making and the execution of political leadership differ. There are, for example, huge differences between political leadership in Latin America and such leadership in Scandinavia, and thus direct transfer of bus priority policies from one context to another is

not advisable. Rather, each city must translate and adjust appropriate ideas for implementation of bus priority measures to the city's specific context.

Acknowledgment

The work on this article has been funded by and in collaboration with The Swedish Knowledge Centre for Public Transport (K2).

Biography



Claus Hedegaard Sørensen is a research leader at the Swedish Knowledge Centre for Public Transport (K2) and senior researcher at the Swedish National Road and Transport Research Institute (VTI). Claus is doing research on transport governance, and his research mostly focuses on environmental policy integration in transport, national transport planning, organization and collaboration within public transport, and the use and role of knowledge in transport policymaking. For the last couple of years he has researched and published mostly on governance of smart mobility. Thus, he has led and been involved in more research projects within that field.



Fredrik Pettersson is an associate senior lecturer at the Department of Transport and Roads, Lund University, and is involved in the Swedish Knowledge Centre for Public Transport (K2). His research interest is in the dynamics between different levels of decision-making and different organizations in the transition to a more sustainable transport system. In the last decade he has published research on national-level transport policy-making as well as planning and decision-making processes at the local and regional level.



Joel Hansson is a PhD student at Lund University and K2 – the Swedish knowledge centre for public transport. His PhD project aims at developing knowledge about effective and attractive regional public transport. Joel is the main author of the Swedish guidelines for Regional BRT. Before starting his PhD studies, Joel was a consultant, working with public transport strategies and analyses in Sweden.

See Also

Planning for Bus Rapid Transit (BRT); Light Rail; Policy and Governance

References

- Currie, G., 2016. Managing on-road public transport. In: Bliemer, M.C.J., Mulley, C., Moutou, C.J. (Eds.), *Handbook on Transport and Urban Planning in the Developed World*. Edward Elgar Publishing, Cheltenham, UK, pp. 471–497.
- Dadashzadeh, N., Ergun, M., 2018. Spatial bus priority schemes, implementation challenges and needs: an overview and directions for future studies. *Public Transport* 10, 545–570.
- dell'Olio, L., Ibeas, A., de Oña, J., de Oña, R., 2018. *Public Transport Quality of Service*. Elsevier, Amsterdam.
- Diakaki, C., Papageorgiou, M., Dinopoulou, V., Papamichail, I., Garyfalas, M., 2015. State-of-the-art and -practice review of public transport priority strategies. *IET Intelligent Transp. Syst.* 9 (4), 391–406.
- Dinopoulou, C., Diakaki, C., Papamichail, I., & Papageorgiou, M. (2013). Public transport priority strategies: progress and prospects. In *2nd international symposium and 24th national conference on operational research, Athens*.
- Finn, B., Heddebaut, O., Kerkhof, A., Rambaud, F., Sbert Lozano, O., Soulas, C. (2011) Buses with high level of service – Fundamental characteristics and recommendations for decision-making and research. Results from 35 European cities. Final report – COST action TU0603 (October 2007 – October 2011).
- Goh, K.C.K., Currie, G., Sarvi, M., Logan, D., 2014. Experimental microsimulation modeling of road safety impacts of bus priority. *Transport. Res. Rec.* 2402, 9–18.
- Greater Wellington, 2019. Bus Priority Action Plan. Draft. Retrieved from: <https://wellington.govt.nz/~media/services/parking-and-roads/bus-priority/files/wellington-bus-priority-action-plan-draft.pdf> (January 18, 2020).
- Hansson, J., Pettersson, F., Svensson, H., Wretstrand, A., 2019. Preferences in regional public transport: a literature review. *Eur. Transport Res. Rev.* 11, 38.
- Lindau, L.A., Hidalgo, D., de Almeida Lobo, A., 2014. Barriers to planning and implementing in Bus Rapid Transit systems. *Res. Transport. Econ.* 48, 9–15.
- Litman, T., 2016. When are Bus Lanes Warranted? Victoria Transport Policy Institute, Victoria, BC, Canada.
- March, J.G., Olsen, J.P. (1989) *Rediscovering institutions: the organizational basis of politics*, Free press.
- Muñoz, J.C., Geschwinder, A., 2008. Transantiago: a tale of two cities. *Res. Transport. Econ.* 22, 45–53.
- Munoz, J.C., Paget-Seekins, L. (Eds.), 2016. *Restructuring Public Transport through Bus Rapid Transit—An International and Interdisciplinary Perspective*. Policy Press, Bristol.
- (National Association of City Transportation Officials) NACTO., 2016. *Transit Street Design Guide*. Island Press, Washington, DC, USA.
- Nikitas, A., Karlsson, M., 2015. A Worldwide State-of-the-Art Analysis for Bus Rapid Transit: Looking for the Success Formula. *J. Public Transp* 18 (1).
- New York City, 2019. Better Buses Action Plan. Retrieved from: <https://www1.nyc.gov/html/brt/downloads/pdf/better-buses-action-plan-2019.pdf> (January 18, 2019).
- Pettersson, F., Hrelja, R., 2018. How to create functioning collaboration in theory and in practice—practical experiences of collaboration when planning public transport systems. *Int. J. Sustain. Transport.* 14 (1), 1–13.
- Pettersson, F., Sørensen, C.H., 2019. Why do cities invest in bus priority measures? Policy, polity, and politics in Stockholm and Copenhagen, *Transport Policy*, <https://doi.org/10.1016/j.tranpol.2019.10.013>.
- Poku-Boansi, M., Marsden, G., 2018. Bus rapid transit as a governance reform project. *J. Transport Geogr.* 70, 193–202.
- Reynolds, J., Currie, G., Rose, G., Cumming, A., 2017. Moving beyond techno-rationalism: new models of transit priority implementation. *Australasian Transport Research Forum 2017 Proceedings 27-29 November 2017, Auckland, New Zealand*.
- Rye, T., Monios, J., Hrelja, R., Isaksson, K., 2018. The relationship between formal and informal institutions for governance of public transport. *J. Transport Geogr.* 69, 196–206.
- Singerman Ray, R., 2019. The politics of prioritizing transit on city streets. *Transport. Res. Rec.* 2673 (3), 733–742.
- TCRP (Transit Cooperative Research Program), 2013. *Commonsense approaches for improving transit bus speeds: A synthesis of transit practice (TCRP synthesis 110)*. The National Academies Press, Washington, DC, USA.
- TCRP (Transit Cooperative Research Program), 2016. *A guidebook on transit-supportive roadway strategies (TCRP report 183)*. The National Academies Press, Washington, DC, USA.
- The Swedish Knowledge Centre for Public Transport, 2016. *Guidelines för attraktiv regional bustrafik – regional BRT*. Lund: K2. Retrieved from: http://www.k2centrum.se/sites/default/files/fields/field_bifogad_fil/guidelines_regional_brt_webb_20161201.pdf, December 3, 2019.
- van Lierop, D., Badami, M.G., El-Geneidy, A.M., 2018. What influences satisfaction and loyalty in public transport? A review of the literature. *Transport Rev.* 38 (1), 52–72.
- Vuchic, V.R., 2007. *Urban Transit Systems and Technology*. John Wiley & Sons, Hoboken, NJ, USA.
- Wright, L., Hook, W., 2007. *Bus Rapid Transit Planning Guide*. Downloadable from <https://www.itdp.org/wp-content/uploads/2014/07/Bus-Rapid-Transit-Guide-Complete-Guide.pdf>, July 3, 2018.
- ÅF-Infrastructure AB, 2016, <https://insynsverige.se/documentHandler.ashx?did=1863388> (January 18, 2020).

Further Reading

- Currie, G., 2016. Managing on-road public transport. In: Bliemer, M.C.J., Mulley, C., Moutou, C.J. (Eds.), *Handbook on Transport and Urban Planning in the Developed World*. Edward Elgar Publishing, Cheltenham, UK, pp. 471–497.
- National Capital Region Transportation Planning Board. Metropolitan Washington Council of Governments. 2011. Bus priority treatment guidelines. Washington, Retrieved from: https://nacto.org/wp-content/uploads/2015/04/bus_priority_treatment_guidelines_national_capital_region_trans_planning_board.pdf, December 3, 2019.
- Pettersson, F., Sørensen, C.H., 2019. Why do cities invest in bus priority measures? Policy, polity, and politics in Stockholm and Copenhagen. *Transport Policy*, Available online 26 October 2019.
- Reynolds, J., Currie, G., Rose, G., Cumming, A., 2017. Moving beyond techno-rationalism: new models of transit priority implementation, *Australasian Transport Research Forum 2017 Proceedings 27-29 November 2017, Auckland, New Zealand*.
- TCRP (Transit Cooperative Research Program), 2016. *A guidebook on transit-supportive roadway strategies (TCRP report 183)*. The National Academies Press, Washington, DC, USA.

High-Speed Rail and the City

Marie Delaplace, Gustave Eiffel University, Champs sur Marne, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	261
The Extension of HSR Throughout the World	261
HSR and the City: The Issue of the Wider Economic Impacts	262
HSR and Improving Cities' Accessibility	262
The Role of HSR with in Attracting Businesses and Tourists	263
HSR and Urban Projects	263
Conclusions	264
References	264

Introduction

High-speed rail (HSR) is characterized by high speed. But what is high speed? The International Union of Railways (UIC) defines HSR as trains running on services with a commercial speed of at least 250 km/h. However, according to European directive 96/48/EC, in some specific cases, speeds above 200 km/h can be considered high speed, if this speed is a significant improvement upon the previous rail service. Across the world, HSR speed has evolved. The commercial speed of the first Shinkansen did not exceed 218 km/h. The commercial speed of the first French *train à grande vitesse* (TGV) in 1981 was 260 km/h, while the commercial speed of the newest TGV in France can reach 350 km/h; the maximum commercial speed of the ETR 1000 train, which runs in Italy, is 360 km/h, while the Maglev (which uses a different technology to all others) between Pudong Airport and the Longyang Road in Shanghai has been running at 430 km/h since 2002.

HSR developed during the second half of the 20th century, first in Japan with the Shinkansen (or bullet train) between Tokyo and Osaka. This was introduced in time for the 1964 Tokyo Olympics. HSR subsequently expanded to European countries: in Italy with the Direttissima, which partially opened in 1978 between Rome and Florence; in France in 1981 with the TGV between Paris and Lyon; and in Germany in 1988, Spain in 1992, Belgium in 1997, the United Kingdom in 2003, and the Netherlands in 2009. Outside Europe, it was developed in China in 2003, South Korea in 2004 and Taiwan in 2007.

The year 2003 was a crucial date for HSR, as it marked its arrival in developing countries, with the first line from Qinhuangdao to Shenyang in China. A few years later, in 2009, HSR reached Turkey, with a first line between Ankara and Eskişehir. The last developing country to launch HSR was Morocco in 2018, with the line between Tangier and Kenitra (Delaplace, 2018).

The Extension of HSR Throughout the World

As of October 2019, there were 47,560 km of high-speed lines (HSLs) in the world, with a further 12,892 km under construction. More than 40,000 km of lines are planned worldwide for completion by 2050 (UIC, 2019). It must not be forgotten, though, that HSR is a heterogeneous category in terms of technology and network types.

In terms of rolling stock, high-speed trains are characterized by new, powerful engines, new and improved train stability systems, different train lengths, and sometimes new facilities for passengers. HSR technologies have had to be improved in order to maintain stability at high speed, enable trains to stop safely, and minimize noise and vibrations near the lines (Givoni, 2006). A high-speed train is not a radical innovation in itself, although technologies may differ from one country to another. This is typically because they have been developed at different times, within the context of specific technological trajectories. Some HSRs, such as the Shinkansen or the French TGV use the same technology as older trains—that is, wheels that run on rails; others, such as Maglev, which make use of magnetic levitation technology, are more innovative because the train has no direct physical contact with the rail, thus eliminating friction and enabling speeds of up to 600 km/h. In Italy, the Pendolino is characterized by a specific mechanism that allows it to tilt on bends. In some European countries, however, the gauge is different, causing interoperability issues. In certain countries several technologies coexist.

High-speed trains must also be analyzed in conjunction with the type of infrastructure on which they run, which also varies by country. Campos and de Rus (2009) defined four different types of systems: first, an exclusive exploitation model with specific and separate infrastructures for high-speed trains and conventional services respectively, as in Japan. The second kind of system is the French model, characterized by high-speed trains which can run on either specific new infrastructure or conventional lines, which are sometimes upgraded. The third type is the Spanish model, where HSLs are also used by conventional trains. The fourth and final model is characterized by high-speed and conventional services that run on both types of infrastructure, as is the case in Germany and Italy.

Moreover, the locations of stations differ depending on the countries and cities in question. In China where there is fast-growing and extensive urbanization, HSR has been a tool for creating new intermediate cities between major cities (Yin et al., 2015), with the station acting as the point of departure for the city's development. In France, as in Spain, HSR sometimes serves central stations and in other cases serves new peripheral stations located directly on the high-speed line. In France, however, some HSR services call at both types of station.

Lastly, the network structure of HSLs can also be quite different, and depends on the urban structure of the country concerned. In some cases, the network is centered on a very large city, sometimes the capital, as in the case of Paris (although some intersector HSLs do allow trains to bypass Paris) in France, Madrid in Spain, Seoul in South Korea, and Tokyo in Japan; in other cases, there are HSR hubs in many cities, as in Germany and China.

Whatever the technology and the network configuration, HSR is mainly used for passenger transport (albeit with different aims (commuting, business, tourism, etc.) and not for freight, although some countries did formerly use HSR for postal freight services, as in France up until 2014 (Strale, 2016). Since November 2018, HSR freight services have been running overnight in Italy between the Maddaloni–Marcianise rail freight terminal (Caserta province, north of Naples) and the Bologna Interport (See: <https://www.rail-freight.com/railfreight/2018/11/02/high-speed-freight-train-italy-hits-the-track-on-7-november/?gdpr=accept>). Other freight projects such as Euro Carex in Europe and a project initiated by the China Railway Rolling Stock Corporation (CCRC), China's leading train manufacturer (This project was part of China's Medium- and Long-Term Railway Plan (MLTRP)), have also been under discussion for a long time.

While HSR technology is not a radical technological innovation, the effects of this technology on local dynamism, also referred to as the wider economic impact of HSR, has been at the core of numerous debates.

HSR and the City: The Issue of the Wider Economic Impacts

In cities that benefit from it, HSR can be seen as an “improvement innovation” that is expected to increase accessibility to the city. This undoubtedly attracts businesses and/or tourists. In many cases, HSR is also a tool for urban development projects.

HSR and Improving Cities' Accessibility

A new high-speed infrastructure can improve the accessibility of cities as a result of higher speeds and the induced reduction of travel times. As such HSR in Europe, for example, has modified the map of European accessibility. It has brought peripheral regions closer to central areas, but has also increased the imbalances between the main cities and their hinterlands. Vickerman (2015) considers the benefits linked to improved accessibility and connectivity mainly concern large metropolitan areas. HSR strengthens the regional and national roles of cities served by HSR, speeding up the processes of concentration and metropolitanization (Urena et al., 2009). Monzon et al. (2013) show that extensions of HSR to peri-urban areas in Spain have increased spatial imbalances and contributed to a more polarized spatial development, as HSR extensions are only built taking into consideration network efficiency. Conversely, when HSR lines are implemented in order to serve regions homogeneously, it has a positive effect on spatial equity (Monzon et al., 2013).

HSL improves the daily accessibility. As a result, an increase in commuter traffic may occur owing to modal shift, with weekly travel replaced by daily travel, especially for high-income passengers (Klein, 2003). But HSR effects on accessibility are unequal. It is evident that the effects of HSR on accessibility are linked to the resulting speed improvements, which vary across countries and cities. These speed improvements depend on the technical characteristics of the HSTs and of HSR infrastructures (Campos and de Rus, 2009), and the physical and human geographical characteristics of the land through which the train runs. For example, commercial speed may be limited in urban environments, or in areas with tunnels or bridges. Moreover, some HSTs also run on conventional rail lines at lower speeds. Many a times they depend on the previous rail service speed (Garmendia et al., 2008). If the speed of the previous conventional rail service was especially low, the speed improvements linked to HSTs will appear more significant.

Accessibility improvements also depend on:

- The quality of the new service in terms of frequency (which is correlated to the size of the city, the number and type of destinations, whether it reduces or eliminates the need for transport connections, whether it enables return journeys within the same day, etc. (Delaplace, 2017); the type of HST service (international or interregional) (Willigers and Van Wee, 2011).
- The location of HSR stations. In their analysis of the case of Reims, in eastern France, Beckerich et al. (2019) show that HSR has improved the accessibility of both of Reims's stations, but in different ways depending on the type of station. At the peripheral station, access to the HSR network and the cities directly served by this network is better, while access to the capital (Paris) is better from the central station.
- The policies associated with HSR, especially concerning the quality of intermodal management at HSR stations. As Givoni and Banister (2012) point out, the train's journey time is not the only travel time to be taken into account. The amount of time it takes to get to and from stations can reduce the time savings generated by HSR. This is further influenced by supporting transport infrastructures, improving intermodality. Chen (2012) shows how accessibility of HSR stations is unequal among Chinese cities: conditions are much better in the larger cities where HSR stations are also typically intermodal transport hubs.

The Role of HSR with in Attracting Businesses and Tourists

HSR is expected to attract new high-level services in cities where it is implemented. In many instances, it can facilitate the development of business parks through the relocation of some activities (offices) in cities served by HSR, particularly in larger cities (Garmendia et al., 2008). In the case of an international HSR network, Rietveld et al. (2001) highlighted the arrival of international firms as a possible benefit. Concerning the attractiveness of firms around HSR stations, Willigers and Van Wee (2011) consider that HSR plays a role in the attractiveness of cities owing to improved accessibility at the intraregional level for firms already located in the same city or region before the arrival of HSR but not at the interregional level.

Areas around HSR stations also have the potential to attract businesses as a result of increased accessibility. This, however, is not always the case. In France, business parks around peripheral stations do not always attract enough business (Facchinetti-Mannone, 2013). In other cases, central HSR stations provide better intermodality, which allows for a greater concentration of activities in historic core districts. In Reims firms took longer to locate around the city's peripheral station than around its central station, with 70% being local companies (Becherich et al., 2019).

Improved accessibility and connectivity may result in increased productivity and competitiveness for companies located in cities connected to the HSR network. Nevertheless, according to Crozet (2015), the positive impacts of accessibility gains on productivity are conditional.

Improving the accessibility of a destination through HSR can induce the development of meetings, incentives, conferencing, and exhibitions (MICE) tourism, as well as urban tourism. These types of tourism appear to be the main benefactors of HSR, especially if the destination station is located in the heart of the town or city (Delaplace and Bazin-Benoit, 2017). For tourists, HSR replaces the fatigue and inconvenience of driving, congestion, and parking difficulties in city centers, as well as being a cheaper mode of travel. HSR therefore reinforces the touristic attractiveness of a destination and expands its market area. Furthermore, HSR is viewed as a major contributor to climate change efforts, reducing the impact of car and air travel on global emissions.

The results of studies conducted over the last 30 years show unequal benefits from HSR and the reinforcement in competition between tourist destinations (Wang et al., 2012) due to the varying degrees of improvement of their accessibility. An increase in tourist movements is mentioned in big cities in different cities.

Concerning China, a very specific case owing to the extensiveness of its network and the spatial dispersion of its tourist destinations, a certain heterogeneity exists according to location. There is a positive impact on tourism for less-developed western regions, while the impact is moderate for central regions and negligible for the east of the country (Chen and Haynes, 2019).

Another key consideration is the fact that rural tourism does not benefit from HSR, owing to the significant time losses and need for other modes of transport required to reach final destinations. HSR and tourism expansion do not always go hand in hand.

Lastly, a very new area of literature, focusing on the impact of HSR on destination choice, was investigated during the late 2010s (Delaplace et al., 2014). This literature contends that the effects also depend on the destination (the capital or intermediate cities, coastal destinations, theme parks, etc.) and sometimes on the type of tourists (domestic versus international) (Pagliara and Mauriello, 2020).

Furthermore, in many cases, the opening of an HSR service can also induce a decrease in length of stay.

HSR and Urban Projects

HSR can be sometimes described as an incremental innovation (Delaplace, 2017). It adds some new elements to the city. It is used in particular by cities as a tool for urban development and regeneration projects. Two examples include the areas around St Pancras International HSR station in London and Stratford International station, constructed as part of the London's 2012 Olympic Games. In both cases, railway-station districts are at the core of the contemporary urban fabric.

Central stations are renovated and transformed into urban hubs for various flows of transport and shopping activities. HSR generates urban renewal around old stations and urban projects around new stations and this helps to make the station a public space, and not just a rail or intermodal transport hub. By releasing land, HSR services can generate new opportunities in terms of business, residential or commercial real-estate developments (see for example stations in Lille, Le Mans, Reims, and Saint-Étienne in France). Local developers play an important role and their investment is perceived as a signal that leads other (national) developers to invest. These case studies also illustrate the role played by the city and local development agencies to attract developers.

Sometimes, HSR is used to create new centralities around peripheral stations. But these types of projects are not always successful (Beckerich et al., 2019; Facchinetti-Mannone, 2013).

The effects of HSR on urban change are also linked to policies that aim to improve the interconnection between different modes of transport in the city, and also with other cities. The station must be well connected to other transport modes and infrastructures, such as airports, motorways, and interregional rail transport, as well as local public transport (Rietveld et al., 2001), and at different levels (local, national, international, and inter-continental). But this is not always the case. Moreover, well-managed transfers to the city center are also required in the case of peripheral stations. The interconnection of these different networks requires cooperation between different local stakeholders, whether public or private (Delaplace, 2017).

Lastly, HSR can have a positive impact on the image of cities, in terms of modernity and prominence, which is often linked to urban projects around stations and the location of "firms with a status" (Rietveld et al., 2001). More broadly, HSR can raise the status of a whole region but also of the area around stations.

Conclusions

There is no evidence that HSR can always solve transport and regional development problems. One can conclude that the link between HSR and accessibility, economic activities and urban change is place-based: it depends on local considerations and policies. Geography and policies matter in this case!

With this in mind, what will the future hold for HSR? On the one hand, the future seems promising in Europe. The European Commission has expressed a wish for the European HSR network to be completed with interoperability issues resolved by 2050. The length of the HSR network as it stood in 2011 could be tripled by 2030. By this date, it is hoped that rail will be used for the majority of medium-distance passenger transport, as environmental concerns become ever more pressing. A modal shift from air travel to HSR would certainly benefit the environment, however this might be replaced by new induced traffic, contributing to the overall increase in emissions. Moreover, if one considers the total life cycle of an HSR line, the environmental benefits are even lower.

Nevertheless, the future seems to be characterized by HSR projects in numerous countries. In the United States a number of projects are underway, as well as in other parts of the world including Iran, India, Brazil, Malaysia, Egypt, Mexico, and Indonesia.

This paper is dedicated to Moshe Givoni, who passed away in 2019. He was an important researcher in the field of transportation in general and with regard to HSR in particular.



References

- Givoni, M., Banister, D., 2012. Speed: the less important element of the High-Speed Train. *J. Transport Geogr.* 22, 306–307.
- Beckerich, C., Benoit, S., Delaplace, M., 2019. Are the reasons for companies to locate around central versus peripheral high-speed rail stations different? The cases of Reims Central Station and Champagne-Ardenne Station. *Eur. Plan. Stud.* 27 (3), doi:10.1080/09654313.2019.1567111.
- Campos, J., de Rus, G., 2009. Some stylized facts about High-Speed Rail: a review of HSR experiences around the world. *Transport Policy* 16, 19–28.
- Chen, C.-L., 2012. Reshaping Chinese space-economy through high-speed trains: opportunities and challenges. *J. Transport Geogr. Elsevier* 22 (C), 312–316.
- Chen, Z., Haynes, K.E., 2019. High speed rail and China's New Economic Geography, Impact Assessment from the Regional Science Perspective New Horizons in Regional Science. Edward Elgar.
- Crozet Y., 2015. High speed rail and urban dynamics: wider or targeted economic effects? International Conference "High speed rail and the city", Urban Futures, Labex week 21–23.
- Delaplace, M., 2018. La grande vitesse ferroviaire dans les pays en développement et les possibles inégalités d'usage ; Une application au Maroc. *Les Cahiers Scientifiques du Transport* 73, 111–138.
- Delaplace, M., 2017. Assessing ex ante the wider effects of high-speed rail service in cities? The lessons drawn from a service innovation-based analysis. *Eur. Rev. Serv. Econ. Manage.* 3, 105–131, doi:10.15122/isbn.978-2-406-07120-4.p.0105.
- Delaplace, M., Bazin-Benoit, S., 2017. High-speed rail Services and tourism expansion: the need for cooperation. In: Albalade, D., Bel, G. (Eds.), *Evaluating High Speed Rail Routledge studies in transport analysis*. pp. 69–81.
- Delaplace M., Pagliara F., Perrin, J., Mermet, S., 2014. Can High Speed Rail foster the choice of destination for tourism purpose? *Procedia - Social and Behavioral Sciences*, EWGT2013-16th Meeting of the EURO Working Group on Transportation.
- Fachinetti-Mannone, V., 2013. Les nouvelles gares TGV périphériques : des instruments au service du développement économique des territoires ? *Géotransports. Transport et développement des territoires* n 1–2.
- Garmendia, M., Urena, J.-M. De, Ribalaygua, C., Leal, J., Coronado, J., 2008. Urban residential development in isolated small cities that are partially integrated in metropolitan areas by high speed train. *Eur. Urban Reg. Stud.* 15 (3), 249–264.
- Givoni, M., 2006. Development and impact of the modern high-speed train: a review. *Transport Rev.* 26 (5), 593–611.
- Klein, O., 2003. Le travail métropolitain: un outil géographique pour révéler l'usage sélectif de la grande vitesse. *L'espace géographique*, N°2/03.
- Monzon, A., Ortega, E., López, E., 2013. Efficiency and spatial equity impacts of high speed rail extensions in urban areas. *Cities* (30), 18–30.
- Pagliara, F., Mauriello, F., 2020. Modelling the impact of High Speed Rail on tourists with Geographically Weighted Poisson Regression. *Transport. Res. Part A* 132, 780–790.
- Rietveld, P., Bruinsma, F., Van Delft, H.T., Ubbels, B., 2001. Economic impacts of high speed trains. Experiences in Japan and France: Expectations in The Netherlands, *Serie Research Memoranda* (de : Faculte it der Economische Wetenschappen en Bedrijfskunde), No 20.
- Strale, M., 2016. High-speed rail for freight: potential developments and impacts on urban dynamics. *Open Transport. J.* 10 (Suppl–1 M6), 57–66, doi:10.2174/1874447801610010057.
- UIC, 2019. High Speed lines in the World, UIC High Speed Department, October, <https://uic.org/passenger/highspeed/>.
- Urena, J., Menerault, P., Garmendia, M., 2009. The high-speed rail challenge for big intermediate cities: a national, regional and local perspective. *Cities* 26 (5), 266–279.

- Vickerman, R., 2015. High-speed rail and regional development: the case of intermediate Stations. *J. Transport Geogr.* 42, 157–165.
- Vickerman, R., 2015. High-speed rail and regional development: the case of intermediate Stations. *J. Transport Geogr.* 42, 157–165.
- Wang, X., Huang, S., Zou, T., Yan, H., 2012. Effects of the high speed rail network on China's regional tourism development. *Tourism Manage. Perspect.* 1, 34–38.
- Willigers, J., Van Wee, B., 2011. High-speed rail and office location choices A stated choice experiment for the Netherlands. *J. Transport Geogr.* 19 (4), 745–754.
- Yin, M., Bertolini, L., Duan, J., 2015. The effects of the high-speed railway on urban development: International experience and potential implications for China. *Progr. Planning* 98, 1–52, doi:10.1016/j.progress.2013.11.001.

Public Engagement in Transport Planning

Miriam Ricci, Department of Geography and Environmental Management, University of the West of England, Bristol, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	266
What Does Engagement Mean? In What Forms can the Public be Engaged?	266
Who is/are 'the Public'?	267
What is the Rationale for and Advantages of Public Engagement?	267
What Aims and Objectives does Public Engagement in Transport Planning Seek to Achieve?	
How can These Objectives be Achieved in Practice?	267
What are the Challenges and Limitations of Public Engagement?	268
What Does Successful Public Engagement Look Like?	269
How has Public Engagement been used in Transport Planning to Date? What Does the Evidence Say?	269
Conclusions: What are the Prospects and Research Agenda for Public Engagement in Transport Planning?	270
References	271
Further Reading	271

Introduction

Public engagement is an umbrella term referring to the involvement of the public in policy and decision making. The past few decades have witnessed a trend toward increased public involvement in a wide range of national and local government domains, such as science and technology, health, the environment, and transport.

According to the definition offered by the Transport Planning Society (tps.org.uk), transport planning concerns the design, implementation and assessment of policies, plans and projects to improve and manage transport systems at a local, regional, national, and international level. As other key public policy domains, transport planning has wide-ranging implications on people's lives, the economy and the environment. Therefore, public engagement allows those who are affected by, or have a stake in, transport-related decisions to have an input into these decisions.

In order to discuss the topic of "public engagement in transport planning" in a meaningful way it is necessary to critically analyze the terminology used and the underlying concepts, by addressing the following questions in the context of transport policy and planning:

- What does engagement mean? In what forms can the public be engaged?
- Who is/are "the public"?
- What is the rationale for public engagement? What are the advantages of public engagement?
- What aims and objectives does public engagement seek to achieve? How can these objectives be achieved in practice?
- What are the challenges and limitations of public engagement?
- What does successful public engagement look like?
- How has public engagement been used in transport planning to date? What does the evidence say?
- What are the prospects and research agenda for public engagement in transport planning?

The rest of the article addresses each of the questions above.

What Does Engagement Mean? In What Forms can the Public be Engaged?

Other terms are often used as synonyms of public engagement, for example public or citizen participation, civic engagement, stakeholder engagement, collaborative or deliberative governance and so on. Although such varied terminology reflects the different policy and governance domains where public involvement is deployed, the ambivalence in conceptual definition can lead to ineffective application of public engagement.

Public engagement can be differentiated into three main categories according to the flow of information between the sponsors of the engagement (i.e., the decision-making organizations/institutions commissioning public engagement) and those who are being engaged (Rowe and Frewer, 2005). These categories are as follows: public communication, public consultation, and public participation.

Public communication involves the unidirectional flow of information (about a specific policy, plan, program, or project) from the sponsoring organization to the public. The engagement mechanisms (i.e., the activities or tools used to deliver engagement) under this category include communication campaigns using a variety of digital and nondigital media channels, open door events, public hearings or meetings where people can attend in person to view, hear and receive the information being communicated.

Public consultation refers to those public engagement mechanisms whereby the sponsor elicits information from the public, so the information flows unidirectionally from the public to the sponsoring body. This category includes face-to-face consultation and information-elicitation activities such as citizens' panel and focus group meetings, as well as consultation tools that do not require participants' physical presence, such as paper-based consultations, phone and online surveys and opinion polls.

Finally, public participation concerns those engagement initiatives where there is a bidirectional flow of information, or dialogue, between the sponsor and the public. Generally, participatory tools involve some mutual learning between the two parties and produce outputs that reflect (at varying degrees depending on the specific public engagement mechanism used) both parties' values and interests. Participation mechanisms include task forces, advisory boards, action planning workshops, citizens' juries, and consensus conferences.

Who is/are 'the Public'?

The public is also an ambiguous, heterogeneous concept. It normally refers to lay people or ordinary citizens but can also include individuals belonging to professional organizations in the public, private and third sector, as well as partisan or interest groups, who are normally referred to as "stakeholders". It can be argued that even ordinary citizens are stakeholders in that they have a stake in the issue under consideration. For example, in the case of transport policy and planning, citizens interact with the transport system in almost every aspect of their lives, for example as road users, and are affected by the positive and negative impacts of transport, for example the characteristics of the urban environment, traffic congestion, air and noise pollution, and global issues such as climate change.

In transport policy and planning, broad categories of public engagement recipients or participants might include transport service providers (infrastructure and operation); transport users, both individuals such as citizens using a range of passenger transport modes, and organizations such as businesses and retailers that rely on the transport system; transport professionals and consultants; trade unions; local communities; media; and financial institutions (Bickerstaff et al., 2002; Cascetta and Pagliara, 2013). These considerations around the notion of "the public" highlight the fact that there are in fact many publics, with multiple and often contrasting views and interests about a specific transport issue, strategy, policy, or intervention.

What is the Rationale for and Advantages of Public Engagement?

There are several normative and instrumental reasons underpinning the use of public engagement in contemporary democratic societies, which are relevant not just in the transport planning domain but also in other areas of public policy (Bobbio, 2019). From a normative point of view, public engagement can be considered a constitutive element of democratic governance, giving people the opportunity to understand what decisions are being made and for what purposes, so they can have their voices heard and have the possibility to influence such decisions. According to this view, public engagement, especially in its more participatory forms, can be a tool for citizens' empowerment. This can be considered part of a broader shift in governance style, from a conventional top-down, expert-dominated process to a more open, accessible and accountable decision-making process.

Involving the public can also be instrumentally advantageous. On the one hand, decisions made by considering perspectives and knowledge inputs from people bringing different types of expertise and interests can be more robust, lead to better outcomes, create a sense of ownership, increase mutual trust and reduce public opposition. On the other hand, the sole fact of including a diverse range of voices and inputs into decision-making can increase both the actual and perceived legitimacy and procedural fairness of the decision-making process, even when the outcomes may not satisfy the needs of all the involved participants.

What Aims and Objectives Does Public Engagement in Transport Planning Seek to Achieve? How can These Objectives be Achieved in Practice?

The specific objectives of public engagement in transport planning depend on the type and purpose of the decision-making process and the organization responsible for it. Decision making can be about developing transport strategies, local transport plans, and specific transport projects and interventions, ranging from major transport infrastructure projects to smaller local schemes.

A variety of organizations can commission and sponsor public engagement in transport policy and planning. These organizations and institutions are generally those with power over decision-making at local, regional, national, and international level. They may include transport authorities, local and regional authorities, national governments and their departments, agencies and ministries, and supra-national political or policy-setting entities such as the European Commission and other similar institutions.

Public engagement can be undertaken at different stages of the transport planning process, whether this concerns the development of strategies and plans, such as a cycling and walking strategy, or the implementation of an intervention "on the ground", such as a traffic calming scheme or a bus rapid transit line. These different stages are as follows (GUIDEMAPS Consortium, 2004):

1. Problem definition: this stage defines the problem/issue to be solved by the strategy/policy or the transport scheme/intervention. It also identifies who is affected by the problem, those who are responsible to address it and the key performance indicators to be used to monitor how the strategy/policy or schemes/interventions are addressing the problem.

2. Option generation: this stage identifies a set of options for the strategy/policy or features and characteristic of the scheme/intervention.
3. Option assessment: each option is assessed using a variety of techniques to establish its suitability to address the problem.
4. Formal decision-taking: in this phase a formal decision about the strategy/policy or scheme/intervention is taken, with sign off from the relevant responsible authority.
5. Implementation: this stage is about the implementation of the strategy/policy or the transport scheme/intervention, with the development of a management plan, communication and marketing plan, and evaluation plan.
6. Monitoring and evaluation: this stage takes place alongside the implementation phase and assesses the outputs, outcomes and impacts created by the strategy/policy or scheme/intervention.

Decision-making organizations such as municipalities and transport authorities may want to involve the public (i.e., ordinary citizens and stakeholder groups) in one or more of the stages listed above, depending on what their specific objectives are. For example, to develop a cycling and walking strategy, the city council might want to involve interest groups in specific phases of the decision-making process, such as problem definition, option generation and assessment. This in turn informs the various elements and features of the engagement process, which are part of the public engagement strategy.

Key questions for the public engagement strategy are as follows (GUIDEMAPS Consortium, 2004):

- Aims of public engagement: Why is the engagement process being undertaken? Is it appropriate in this context?
- Objectives and scope of public engagement: What elements of the strategy/scheme will be open to input from the public? How will public input be used to influence such elements? How much power will citizens and other stakeholders be given? How will public aspirations be managed?
- Identification of “the public” to be engaged: Who are the people who have a stake in the issue and therefore should be involved in the decision-making process? How can such people be identified? What techniques will be used to access them? Are there any groups who are “hard to reach”? What needs to be done to ensure inclusivity? If a venue is involved, is it suitable and accessible?
- Information and transparency: What information is provided to those people who are being engaged? Is the information adequate to ensure that all those involved can understand it and offer an informed view? How will the involved public be informed about how their input is being used at various stages of the planning process? How will participants be alerted about any unexpected changes in the process? How will feedback be provided to them?
- Engagement mechanisms: How will engagement be undertaken? What tools and techniques should be used? Are the selected engagement techniques best suited to the target groups? What is the required outcome of each engagement activity? What is the public being asked to contribute at each stage of the decision-making process?
- Timeframe: When should different activities take place? Are the start and end points of the engagement process clearly defined and agreed early in the process? Are engagement activities planned early enough for people to understand their added value and be fully engaged?
- Resourcing: Have resources required for the engagement strategy been considered as part of project budgeting? If an independent engagement consultant is employed, has a clear project brief been prepared including budget considerations?
- Evaluation: Is the public engagement process going to be evaluated? How? How will participants contribute to the assessment process? How will the organization use outcomes from the assessment to ensure better engagement practices in the future?

What are the Challenges and Limitations of Public Engagement?

Although public engagement is undeniably a constitutive part of transparent, democratic and inclusive governance styles, and in some cases a mandatory component of specific decision-making processes, it is not a panacea. Public engagement brings a whole host of challenges and limitations, related to the following issues (Flynn et al., 2011).

First, as outlined earlier in this article, the public is not a homogeneous entity which can be easily approached and consulted, but there are in fact many different publics or groups who have distinctive and often contrasting values, views and interests on a given issue. The challenge for the sponsor is about how to arrive at an authoritative decision in the collective interest, in the face of disagreement on what this collective interest might be. Important in this context is to recognize that technical expertise is not always homogenous either, as experts can and do disagree on which facts or evidence can be considered “certain”. Transport presents many “wicked problems” and because, by definition, any planning is about the future, uncertainties abound (Lyons and Davidson, 2016).

Second, different publics possess different kinds and levels of knowledge or expertise on a given issue. For example, transport professionals may bring technical expertise while residents and road users may bring experiential knowledge. However, these different publics may not be able to understand each other’s vocabularies and areas of expertise and may not have the same financial and time resources to be involved. The challenge for meaningful and genuine, as opposed to tokenistic, public engagement (Arnstein, 1969) is therefore about allowing for a truly fair dialogue and mutual learning between publics who may not have the same access to financial, social and cultural capital.

Third, even when public engagement is undertaken with the best intentions and the competent application of carefully selected methods, it may still not produce the outcomes that were originally envisaged. Unexpected events or circumstances might prevent reaching a decision that is accepted, or acceptable, by all the involved publics. Opposition and protest may still occur. Public engagement may not increase mutual trust, and might identify hidden and contested issues which had not, or could not, be

anticipated. Moreover, the bureaucratic and institutional cultures associated with traditional governance styles might be difficult to adapt to more open, participatory types of decision-making, leading to implicit biases in the process that might not be easy to recognize and remove (Irwin, 2007).

What Does Successful Public Engagement Look Like?

The issue of whether and to what extent public engagement is successful brings further questions, for example who defines success, according to what set of criteria? Drawing on an extensive review of the literature (Rowe et al., 2005) offer the following set of criteria as the basis to assess the success of public engagement processes:

- Representativeness: The engaged groups should comprise a broadly representative sample of the population with stakes in the issue under consideration.
- Independence: The participation process should be conducted in an independent, unbiased way.
- Early involvement: The public should be involved as early as possible in the process.
- Influence: The output of the engagement should have a genuine impact on policy/decision-making.
- Transparency: The process should be transparent so that the public can see what is going on and how decisions are being made.
- Resource accessibility: Public participants should have access to the appropriate resources to enable them to successfully fulfil their brief.
- Task definition: The nature and scope of the participation task should be clearly defined.
- Structured decision making: The participation exercise should use/provide appropriate mechanisms for structuring and displaying the decision-making process.
- Cost effectiveness: The engagement process should be cost effective and proportionate to the issue under consideration.

How has Public Engagement been used in Transport Planning to Date? What Does the Evidence Say?

Despite its growing importance and diffusion, the practice of public engagement is less developed in the transport planning domain compared to other areas of public policy, such as science and technology policy and health care (Nared, 2020). Nevertheless, some form of public engagement in transport decision making is generally required by law in most democracies across the globe.

It is perhaps not surprising that public engagement in transport planning appears to have received comparably less attention from the academic community than engagement in other policy and planning sectors. A review for the EU-funded project Challenge (Lindenau and Böhler-Baedeker, 2014) reveals that public engagement in transport planning varies widely across Europe, with several countries such as Germany, France, the UK, and Belgium having formal, mandatory consultation procedures for mid- and large-scale transport projects as well as for the development of transport strategies and Sustainable Urban Mobility Plans (SUMPs). A few countries deploy innovative engagement mechanisms such as participatory citizens' workshops and participatory transition management initiatives (e.g., in the Belgian city of Ghent).

A survey of 34 cities from across Europe and beyond, exploring their transport planning practices including the use of engagement mechanisms, found that most cities involved both stakeholders and citizens, but with varying degrees of engagement (Lindenau and Böhler-Baedeker, 2014). Stakeholders, such as public transport providers and sustainable transport organizations, were predominantly involved in the identification of transport and mobility problems, but only in few cases in other planning stages (e.g., specifying the vision and objectives, identifying possible solutions). Retailers and customers were seldom invited to participate. According to the survey's results, key barriers to successful public engagement include the lack of political will, limited financial and staff resources, lack of skills to undertake engagement, failure to establish a plan or strategy for participation, lack of interest from the public, an imbalance between different stakeholders' effective ability to provide input, lack of engagement culture especially in Eastern European post-communist countries.

Similar results emerged from a very recent survey-based study of the experience with participatory transport planning in the eight European metropolitan regions of Ljubljana, Oslo, Gothenburg, Helsinki, Budapest, Rome, Porto, and Barcelona (Nared, 2020). The findings show that public and stakeholder engagement varies according to the geographical scale involved. Participants' engagement is greater at the local level, where transport plans and interventions are perceived as more relevant to those involved. The participatory planning process takes longer than traditional top-down processes, but it can facilitate the implementation of the project/measure when the additional resources and time are justified. The survey found that the engagement process benefited from an inclusive selection of all the involved interest groups, the provision of an experienced engagement facilitator and, above all, transparency in demonstrating how the engagement process and its outcomes influenced the final plans, policies or interventions. Although significant differences in engagement practices and outcomes were found between the countries and regions involved, with Scandinavian countries displaying a more participatory planning culture, the use of participatory methods in transport planning is becoming increasingly important across Europe.

Most of the academic literature on the practice of public engagement in transport planning focuses on case studies from the UK. The emergence of public engagement in transport planning in the UK can be dated back to the late 90s, when a series of key policy documents issued by the Labour Government introduced the concept of engagement to put power closer to people, strengthen citizens' rights, make national and local government more open and accountable, and to ensure that public institutions were more

representative (Bickerstaff et al., 2002). Until then, transport had been a sector with very limited historical experience of public participation in the decision-making and policy process, which traditionally had been dominated by technical experts and politicians.

The 1998 White Paper on an integrated transport strategy announced the introduction of 5-year Local Transport Plans (LTPs), to be developed by highway authorities in England to set out investment proposals and targets in local transport matters such as traffic reduction, air pollution, public transport and active travel. Although the White Paper and subsequent official guidance documents required local authorities to undertake “extensive consultation” with the public and other stakeholders and that their opinions should “make a difference”, details of how this was to be achieved in practice were missing, as well as specific guidelines on who to involve and in what stages of the decision-making process. An evaluation of the early experience with public engagement in transport planning (Bickerstaff et al., 2002), based on a review of 58 LTP documents and a survey of the English local authorities responsible for their production, found that public engagement had been taken up almost universally across the sample, but with considerable variation in terms of the type and level of engagement. Information provision and consultation seemed to be the norm, but the study also found evidence of efforts to deploy more innovative forms of participatory engagement. The study concluded that, for most of the surveyed local authorities, the engagement process resembled the realms of tokenism on Arnstein’s ladder of citizen participation, due to such weaknesses as little engagement with different sectors of the public, few opportunities for effective deliberative interaction and mutual learning, and low policy transparency.

The Local Transport Act 2008 gave local authorities in England more autonomy and flexibility to produce LTPs in terms of their objectives, indicators, timescales and policy instruments. A content analysis of LTPs produced after 2008 revealed that some of the weaknesses identified above were still present and had a considerable impact on the inclusivity of these decision-making processes, especially for social groups more at risk of transport-related social exclusion, such as lone parents and ethnic minorities (Elvy, 2014).

Further insights on the practice of public participation in transport planning are provided by High Speed 2 (HS2), a major (and controversial) transport infrastructure project creating a high-speed rail link between London and the north of England (Cohen and Durant, 2019). Their critical analysis of the engagement process deployed in HS2 Phase I found that the approach was a classic example of “decide, announce, defend” (DAD), suggesting that such traditional and much criticized top-down approaches to decision making are still considered the default option.

Recent experiences in South American countries such as Brazil and Chile suggest that improvements in the practice of public participation are indeed taking place, with examples of local authorities deploying innovative digital technologies alongside more traditional face-to-face engagement techniques to encourage participation from a wider range of publics in the production of urban mobility plans or major transport infrastructure (Sagaris, 2018; Worker, 2014). However key challenges still persist, similarly to what has been found in other countries, notably the lack of institutional culture and practical arrangements that could facilitate more deliberative, collaborative participatory processes, and make them an established component of everyday transport planning.

Public engagement is an increasingly important and institutionalized component of transport decision-making also in North America and Australia. Studies looking at examples from these areas note that, whilst citizens’ participation is certainly a step forward towards more democratic decision-making processes, public engagement can add complexity rather than solve it.

A comparative analysis of citizen participatory mechanisms in Detroit, Michigan (a Citizens Advisory Committee involved in the development of a regional transit master plan) and Hamilton, Ontario (a Citizens’ Jury involved in the establishment of light rail transit) found that both these participatory mechanisms were considered important and useful by both the participants and the politicians who had established them. However, the conclusion was that neither mechanism had a significant impact on transit policies and outcomes, which were instead shaped by a range of political divisions at the local and regional level that no amount of public engagement could address (Sutcliffe and Cipkar, 2017).

Over the past few decades, transport policy in Australian cities has been embedded within metropolitan-wide strategic plans together with land use planning. The development of such plans has become an increasingly participatory and consensus-based process, designed around deliberation and inclusive public dialogue in the form of citizen juries, citizen panels and large-scale town hall meetings (Legacy et al., 2017). However, a critical analysis of three major inner-city road projects in Melbourne, Sydney and Perth revealed that the engagement process did not adequately and openly discuss the harder questions about infrastructure project priorities. Namely, how these projects should be financed and who will be served through the enhanced accessibility they are said to provide.

Conclusions: What are the Prospects and Research Agenda for Public Engagement in Transport Planning?

Public involvement in some form or another has become a feature of democratic governance, and transport policy and planning are no exception. However, despite the growth in engagement exercises within transport decision-making, in-depth evaluations of the extent to which these initiatives are successful, as well as of the underlying barriers and enabling factors, are very few. It seems that the research agenda proposed in 2002 by Bickerstaff and colleagues (Bickerstaff et al., 2002) is still very relevant today: what can public involvement methods achieve in transport-related decision-making processes? How far can public engagement deliver substantive outcomes in terms of policy, mutual learning and broader democratic benefits? What effects should public engagement have on transport policy and decisions? What can public engagement not achieve? How can public engagement mechanisms best be used to enhance transport planning? How can or should we evaluate and learn from the experience gained so far?

Regions and cities are currently faced with significant challenges such as climate change, air pollution, population growth and other key socio-economic trends. Transport represents both a cause and a solution to these issues, therefore more effective public engagement may be a way to ensure that policy decisions reflect shared visions of what the future should look like.

References

- Arnstein, S.R., 1969. A ladder of citizen participation. *J. Am. Inst. Planners*. Taylor & Francis Group 35 (4), 216–224, doi:10.1080/01944366908977225.
- Bickerstaff, K., Tolley, R., Walker, G., 2002. Transport planning and participation: the rhetoric and realities of public involvement. *J. Transport Geogr.* 10, 61–73, Available at: www.elsevier.com/locate/jtrangeo.
- Bobbio, L., 2019. Designing effective public participation. *Policy Soc.* Routledge 38 (1), 41–57, doi:10.1080/14494035.2018.1511193.
- Cascetta, E., Pagliara, F., 2013. Public engagement for planning and designing transportation systems. *Procedia - Social Behav. Sci.* Elsevier 87, 103–116, doi:10.1016/J.SBSPRO.2013.10.597.
- Cohen, T., Durrant, D., 2019. Public engagement and consultation: decide announce and defend? In: Docherty, I., Shaw, J. (Eds.), *Transport Matters*. Bristol University Policy Press, pp. 251–277.
- Elvy, J., 2014. Public participation in transport planning amongst the socially excluded: an analysis of 3rd generation local transport plans. *Case Stud. Transport Policy.* Elsevier 2 (2), 41–49, doi:10.1016/J.CSTP.2014.06.004.
- Flynn, R., Ricci, M., Bellaby, P., 2011. The mirage of citizen engagement in uncertain science: public attitudes towards hydrogen energy. *Int. J. Sci. Educ. Part B: Commun. Public Engage.* 1 (2), doi:10.1080/21548455.2011.585810.
- GUIDEMAPS Consortium, 2004. Successful transport decision-making. A project management and stakeholder engagement handbook.
- Irwin, A., 2007. Public dialogue and the scientific citizen. In: Flynn, R., Bellaby, P. (Eds.), *Risk and the Public Acceptance of New Technologies*. Palgrave Macmillan, UK, pp. 24–40.
- Legacy, C., Curtis, C., Scheurer, J., 2017. Planning transport infrastructure: examining the politics of transport planning in Melbourne, Sydney and Perth. *Urban Policy Res.* Routledge 35 (1), 44–60, doi:10.1080/08111146.2016.1272448.
- Lindenau, M., Böhler-Baedeker, S., 2014. Citizen and stakeholder involvement: a precondition for sustainable urban mobility. *Transport. Res. Procedia* 4, 347–360, doi:10.1016/j.trpro.2014.11.026.
- Lyons, G., Davidson, C., 2016. Guidance for transport planning and policymaking in the face of an uncertain future. *Transport. Res. Part A: Policy Practice.* Pergamon 88, 104–116, doi:10.1016/J.TRA.2016.03.012.
- Nared, J., 2020. Participatory transport planning: The experience of eight European metropolitan regions. In: *Urban Book Series*. Springer, pp. 13–29, doi:10.1007/978-3-030-28014-7_2.
- Rowe, G., et al., 2005. 'Difficulties in evaluating public engagement initiatives: reflections on an evaluation of the UK GM Nation? public debate about transgenic crops', *Public Understanding of Science*. Sage Publications Sage CA, Thousand Oaks, CA 14(4), pp. 331–352. 10.1177/0963662505056611
- Rowe, G., Frewer, L.J., 2005. A typology of public engagement mechanisms. *Sci. Technol. Human Values* 30 (2), 251–290, doi:10.1177/0162243904271724.
- Sagaris, L., 2018. Citizen participation for sustainable transport: Lessons for change from Santiago and Temuco, Chile. *Res. Transport. Econ.* Elsevier 69, 402–410, doi:10.1016/J.RETREC. 2018.05.001.
- Sutcliffe, J.B., Cipkar, S., 2017. Citizen participation in the public transportation policy process: a comparison of Detroit, Michigan, and Hamilton, Ontario. *Can. J. Urban Res.* 33–51, doi:10.2307/26290769, Institute of Urban Studies, University of Winnipeg.
- Worker, J., 2014. Why the public voice matters in urban mobility planning: Lessons from Brazil., *The City Fix*. Available at: <https://thecityfix.com/blog/public-participation-sustainable-urban-mobility-plan-brazil-cities-pac-jesse-worker/>.

Further Reading

- Grossardt, T., Bailey, K., 2018. *Transportation Planning and Public Participation*. Elsevier. doi: 10.1016/B978-0-12-812956-2.00011-6.
- Quick, K., 2014. 'Public Participation in Transportation Planning', in *Encyclopedia of Transportation: Social Science and Policy*, 1132–1137.

Long-Distance Travel

Giulio Mattioli*, **Muhammad Adeel†**, *Technische Universität Dortmund, Faculty of Spatial Planning, Department of Transport Planning, Dortmund, Germany; †University of Leeds, Institute for Transport Studies, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Scope	272
Definition	272
Measurement and Methods	273
Trends	274
Macro Drivers of Growth	274
Determinants of Long-Distance Travel Behavior	275
Policy Issues	275
Acknowledgment	276
See Also	276
References	276
Further Reading	276

Scope

This article focuses on long-distance passenger travel demand – the reader is referred to other articles in this Encyclopedia for a discussion of freight transportation and of infrastructure aspects in the long-distance segment. The focus of the discussion is mainly on private travel, given its prevalence in the long-distance travel segment, although we refer to business travel throughout.

Definition

While the concept of “long distance travel” (LDT) is commonplace in transportation research, a precise definition has remained elusive. Broadly speaking, LDT is defined based on one or more of the following criteria:

- spatial distance
- travel time
- frequency and regularity
- overnighing

Within transportation research, LDT is mostly defined based on spatial distance thresholds, although a range of values are used in the literature, with considerable variation across countries and between studies. The focus on spatial distance generally reflects an interest for transportation volumes and their negative consequences, which tend to be proportional to mileage. In countries in the Global North LDT accounts for a small share of trips, but for a significant share of total mileage (roughly between 30% and 50% depending on the country and method), and for an even higher share of emissions.

There is a parallel tradition of research, which defines LDT in terms of temporal duration. This reflects an interest for the (negative) consequences on individual well-being (e.g. health, social relationships) of spending an excessive amount of time traveling. Notably, the determinants and implications of “long-distance commuting” (whether on daily or weekly basis) have drawn much attention. Studies on “transport poverty” also highlight how excessive travel time can result in problems of time poverty—that is, reduced participation in activities that are essential for social inclusion—particularly in the Global South.

LDT is also often defined in terms of (low) frequency and (lack of) regularity, as when it is contrasted to “daily mobility” or “everyday travel”. Some definitions see LDT (or segments of it) as taking place outside a person’s “usual environment,” although this can be challenging to define precisely. The question of whether LDT is infrequent and irregular is a complex one. LDT is a ‘rare event’ among the population, as well as for most individuals, but at the same time there is a minority of frequent long-distance travelers for whom this is a regular occurrence (e.g., daily long-distance commuters, frequent flyers). These are responsible for disproportionately high shares of LDT mileage, which makes them worthy of investigation.

Finally, LDT is sometimes defined based on overnight stay in a non-home location, for two main reasons. First, overnighing-based definitions of LDT help reduce recall error in household travel surveys, as it is a memorable event for respondents. Second, overnight stays potentially generate tourism revenue, which explains why the official definition of tourists is “persons traveling outside their usual environment for all purposes of visit, and staying overnight but for not more than one consecutive year” (UNWTO–UNEP–WMO, 2008, p.121). Tourism is responsible for an important share of LDT, and this is drawing increasing attention from both transportation and tourism researchers.

There are thus multiple criteria to define LDT, and these are not always aligned. For some forms of LDT (e.g., occasional flight holiday to distant destination), all of the above criteria apply, while for others (e.g., daily long-distance commuting by car), only some of them do. This diversity of perspectives is not a limitation, as it reflects the multiplicity of practical interests that underlie research on LDT. Different definitions of LDT are appropriate, depending on whether the focus is on climate, quality of life or tourism. This does however require researchers to be explicit about the definitions they adopt, about the rationale for them, and to position themselves with respect to alternatives. It also raises issues for data collection, as discussed below.

Another way to deal with the definitional complexity of LDT is to see it as an umbrella concept covering multiple travel segments. [Frick and Grimm \(2014\)](#), for example, suggest a classification based on three criteria (trip purpose; overnighting; whether the trip is outside of one's usual environment), and consisting of eight segments:

- personal day trips
- long-distance everyday personal trips
- long-distance commuting
- long-distance everyday business trips
- business day trips
- holiday trips (5+ days)
- short holiday trips (2–4 days) and other personal overnight trips
- overnight business trips

Alternative segmentations have been proposed, some of which are based on distance. Empirical studies show that mode choice for long-distance journeys is sensitive to trip length, with cars dominating for “shorter” long-distance trips, and a greater share of rail and air travel for journeys over medium and very long distances.

Measurement and Methods

Even when LDT is defined in terms of spatial distance, its empirical assessment remains challenging. Accurate measurement of this travel segment is beset by methodological problems, some of which are inherently difficult to solve. Data collection on LDT tends to be fragmented across countries, disciplines, and surveys, which creates further challenges. The result is that our understanding of LDT lags behind that of everyday urban travel, a situation which contributes to its low salience on the research and policy agenda.

LDT is captured to some extent in conventional household travel surveys but, being a “rare event” for most individuals, it does not fit with the short-reporting period of most travel diaries. Hence a dedicated module on LDT is often included, with a longer reporting period (although this varies widely – from one week to 12 months). This, however, raises other methodological issues, mainly related to high response burden. In retrospective survey instruments, this tends to result in selective recall, soft refusal, and ultimately underreporting of long-distance trips ([Janzen et al., 2018](#)). Prospective instruments have their own issues in terms of respondent drop out. These risks must be traded-off against each other, as they all have implications for the accuracy of LDT estimates.

Other reasons for the underestimation of LDT are the undersampling of (inherently hard-to-reach) frequent long-distance travelers, and the multimodal nature of much LDT, as access and egress modes to transportation nodes like airports are not well-captured in travel surveys. Also, travel at the destination as part of overnight tours tends to be underestimated (or even not captured at all).

The difficulty of capturing LDT with conventional travel surveys has led some countries to develop dedicated LDT surveys. To these must be added national tourism demand surveys, and a range of tourism-related surveys conducted by public and private organizations. As these surveys differ in key methodological aspects such as operational definition of LDT, reporting period, inclusion of international travel, etc., the result is a fragmentation of the data landscape, which works against a comprehensive assessment of LDT levels and trends.

Methodological differences across countries add another layer of complexity. A significant share of LDT is across borders, particularly in world regions like Europe. Yet national surveys differ in key methodological details such as whether LDT is defined in terms of distance or overnighting, distance thresholds (ranging from 40 km to 100 miles), reporting period, and whether international travel is included. While there have been attempts to internationally harmonize LDT survey methods, to date they have had limited success. Internationally comparative studies of LDT are thus rare and beset by high levels of uncertainty.

Overall, data fragmentation and lack of harmonization make it necessary to integrate data from diverse sources in order to obtain overall estimates of LDT at a country (e.g., [Frick and Grimm, 2014](#)) or world region level ([van Goeurden et al., 2016](#)). This is suboptimal and can result in inconsistent measurement.

More broadly, the shortcomings of conventional household survey instruments in capturing LDT have led researchers to experiment with “passive data collection methods” based on mobile devices, although these have their own methodological shortcomings in terms of technical accuracy, data cleaning requirements, and sampling representativeness. The limited information available on individual attributes also constrains the opportunities for data analysis. Nonetheless, the results of studies based on mobile device data suggest that LDT demand is vastly underestimated by household surveys. [Janzen et al. \(2018\)](#) for instance analyzed mobile phone billing data of millions of French users, finding long-distance trip rates twice as high as those reported in the national travel survey.

Trends

While reliable figures on long-term trends in LDT are scarce, existing evidence points to a rapid growth since the mid-twentieth century. Schäfer et al. (2009) show how the large increase in global passenger travel distances between 1950 and 2005 was driven by a shift from low- to high-speed transportation. Their analysis shows that when a country reaches average mobility levels of circa 30 km per person per day, the modal share of the automobile starts to decline (on a pkm basis) in favor of “high-speed transportation modes” such as air travel and high-speed rail. This is associated with a shift from short- to long-distance travel, and a growing share of leisure trips. In high-mobility countries, there is also a shift from domestic to international LDT, as observed, for example, for Sweden (Frändberg and Vilhelmson, 2011). Globally, international tourist arrivals increased by more than 20 times between 1950 and 1995, far outstripping population growth (Gössling and Peeters, 2015).

Trends in the last three decades are better documented, and show the continuation, and possibly acceleration, of the trends just described. Global estimates (UNWTO and ITF, 2019) show that between 2005 and 2016 there has been a rapid increase in overnight tourism travel both domestically (+119%) and internationally (+65%), while same-day trips have doubled. van Goeuverden et al. (2016) find an increase in LDT trips (+8%) and volume (+27%) in Western Europe between 2001 and 2013. Since short-distance travel has stagnated over the same period, this results in an increase in LDT’s share of total pkm. Similar trends have been observed in several Global North countries, with faster growth in the private and leisure travel segments than in business travel. In countries like Sweden, growth has been particularly rapid for international travel with, for example, Larsson et al. (2018) finding a doubling of international flights between 1990 and 2014.

Overall, these historical trends suggest a continuous spatial extension of overall mobility (Frändberg and Vilhelmson, 2011). The “peak car” or “peak travel” phenomenon observed in some countries in the Global North could thus reflect a shift towards faster, long-distance travel modes (notably aviation) and international mobility (which is poorly captured in national travel surveys), rather than a turning point. Notably, there is suggestive evidence that peak car among young adults may be offset by increasing levels of long-distance, international, and air travel.

While it is often assumed that this rapid growth has gone along with a reduction of inequalities in participation to LDT, the empirical evidence is contrasted. Demoli and Subtil (2019) for example, show that the distribution of LDT among French residents has become more unequal between 1974 (when 10% of the population was responsible for 30% of LDT mileage) and 2008 (when this figure reached 60%).

Forecasts of future travel demand (Schäfer et al., 2009) show a continued increase, driven by the continuing shift towards high-speed transportation modes, whose share of total pkm is predicted to reach 45%–70% in “industrialized regions” and 15%–30% in “developing economies” by 2050 (assuming unconstrained infrastructure expansion and no travel demand management). Schäfer et al. (2009, p.51) argue that in such a scenario, “long-distance trips with a short stay at destination may become more normal,” as would associated “multilocal” and global lifestyles. Tourism forecasts (UNWTO and ITF, 2019) predict a 74% increase in overnight trips between 2016 and 2030, and a doubling of same-day visitors, with faster growth in the Asia and Pacific region, and an associated increase in air travel. At the time writing, it is not clear to what extent the COVID19 crisis of 2020 will change these forecasts.

Evidence on trends in work-related LDT is sparser and more ambiguous. Studies from the Global North generally show an increasing trend in long-distance commuting (e.g., Frändberg and Vilhelmson, 2011; Frick and Grimm, 2014), although this varies considerably between countries, over time, as well as depending on the (time- or space-based) definition. Viry et al. (2015) compare retrospective life-course data for three cohorts (born between the 1952 and 1981) in four EU countries and find no evidence of an increase in long-distance commuting and work-related “overnighting” over time.

Macro Drivers of Growth

The structural drivers of LDT growth are manifold and are investigated in various disciplines and strands of research. Schäfer et al. (2009) explain increasing travel demand since the 1950s simply as a result of economic growth, technological progress, and increasing affordability of faster, motorized travel modes. Yet the decreasing cost of travel, improvements in transportation and telecommunication technologies, and the expansion of related infrastructure tend to result in an expansion of activity spaces and social networks, which in turn encourage more LDT. This can be defined as the “institutionalisation” of LDT, that is, the self-reinforcing process whereby increasing levels of LDT become locked-in (Frändberg and Vilhelmson, 2010). With regard to tourism, further drivers of growth include increases in paid holiday time, the softening of border travel restrictions, the diffusion of information and communication technologies, as well as the globalization of technology, trade, and production (Scott and Gössling, 2015), which has direct implications for business travel. With regard to air travel, the rise of new business models such as low-cost carriers has played an important role.

Overall, greater ease of LDT has encouraged new lifestyles and practices, variously referred to as “transnationalism,” “transmigration,” and “residential multilocality,” which blur the distinction between migration, residential relocation and circular mobility. These may be more common among well-educated younger generations, and the youth travel segment is a key area of growth for the tourism industry. Transnational and multilocal practices are predicated on high levels of LDT, although their impacts on transportation have not received the attention that they deserve to date.

Looking forward, [Frick and Grimm \(2014\)](#) list a number of drivers that might contribute to future LDT growth including demographic trends (ageing, migration), as well as spatial (urbanization), economic (expansion of the service sector), and cultural (individualization) developments.

Determinants of Long-Distance Travel Behavior

Transportation research shows that multiple factors explain differences in short-distance travel behavior between individuals, including sociodemographic, economic, sociopsychological, and spatial attributes. The role of these factors in explaining LDT is comparatively less well understood, although existing evidence highlights both similarities and differences.

With regard to income, previous research has shown that LDT is more strongly associated with income than daily travel, and air travel is more strongly associated with income than car travel, even after controlling for intervening factors. Other socioeconomic characteristics consistently associated with higher levels of LDT in multivariate regression studies include (full-time) employment, being in education, higher levels of education, and male gender, while old age and living in households with children tend to reduce LDT.

Most studies on environmental attitudes and daily travel behavior find only a weak association between the two, and substantial evidence suggests that this “attitude-behavior gap” is even more pronounced for long-distance, holiday, and air travel. [Czepkiewicz et al. \(2019\)](#), on the other hand, find a strong association between cosmopolitanism and emissions from international travel, suggesting that LDT might well be influenced by other attitudinal dimensions.

While much research shows that in compact, inner city areas with mixed land use daily travel distances, (particularly by car) are lower, the relationship between urban form and LDT is in the opposite direction. [Czepkiewicz et al. \(2018\)](#) review the empirical literature on this point (mostly from Europe), concluding that levels of (non-business) LDT are higher in dense urban areas, even after controlling for common socio-economic predictors. This finding is mostly driven by international and air travel – when these are excluded, LDT levels are higher in low-density and peripheral areas. Five theoretical explanations for this relationship have been put forward ([Czepkiewicz et al., 2018](#)):

- a rebound effect, whereby the time and money saved on car use in everyday life are spent on LDT
- the “compensation hypothesis,” whereby for example, urban residents make up for lack of private green space with LDT to natural areas
- better access to LDT infrastructure
- lifestyles and sociopsychological characteristics conducive to more long distance travel
- greater dispersion of social networks

Overall, [Reichert et al. \(2016\)](#) argue that socioeconomic attributes affect emissions in daily travel and LDT in the same direction, while spatial attributes have opposite impacts on the two segments. A possible exception here is migration background, with some research showing that migrants have lower levels of car use, but higher levels of air travel and related emissions, partly because of more geographically dispersed social networks (see e.g., [Mattioli and Scheiner, 2019](#)). Yet to date most quantitative studies of LDT determinants do not consider migration background or social network dispersion.

The socioeconomic determinants of business LDT are broadly consistent with those of LDT as a whole, although further work- and workplace-related factors need to be considered such as sector of employment, business sector, position within organization, and the company’s locational and organizational strategies ([Aguilera and Proulhac, 2015](#)).

Multivariate studies on the determinants of long-distance commuting show similar results, with male gender, higher income, and higher education levels having a positive effect. Living in dense urban areas, and length of residence in the local area tend to reduce the chances of commuting over long distances, as does living in households with children (although this applies mostly to women).

Policy Issues

Historically, LDT has been a relatively overlooked topic in transportation studies and policymaking despite its disproportionate importance in terms of mileage, infrastructure use and associated impacts. It deserves greater consideration, not least due to its rapid growth and significant policy implications. We discuss two of them here: the contribution of LDT to climate change, and social inequalities in LDT demand.

The contribution of LDT to climate change is a topic that falls through the cracks between different sectors, disciplines and data sources. Accounts of sustainable transportation tend to have an urban bias, overlooking non-urban travel even though it accounts for 60% of global passenger transportation volume, and for two-thirds of projected growth to 2030 ([UNWTO and ITF, 2019](#)). Similarly, climate change has only recently gained prominence as a topic in tourism research, although estimates suggest that tourism travel-related emissions (of which half due to air travel) represent 22% of total (both passenger and freight) transportation emissions and 5% of overall man-made CO₂ emissions ([UNWTO and ITF, 2019](#)). These figures are lower-bound estimates, as 50% of travel-related tourism emissions are for aviation, where non-CO₂ climate-forcing impacts significantly magnify the global warming effect (see for example [Reichert et al., 2016](#)).

Historically, efficiency improvements in per passenger emissions have been more than offset by rapid growth in tourist numbers, a trend that is likely to continue. Since there is no easy “technological fix” to aviation emissions, some form of demand management policy will be necessary. This enhances the importance of understanding the dynamics and determinants of LDT demand. Emissions from personal travel could be reduced more effectively if the longest distance trips in relevant activity domains and geographies were the target of decarbonization or demand management. Conversely, reducing passenger transportation emissions will be extremely challenging unless LDT is addressed.

A second, less discussed policy issue is that participation in long-distance travel is very skewed toward privileged social groups. There are several good reasons to question inequalities in LDT, including: questions around the fairness of allocating public resources to the expansion of LDT infrastructure; the disproportionate contribution of frequent long-distance travelers to environmental externalities; the possibility that some forms of LDT may be instrumental to social inclusion, particularly as activity spaces and social networks have expanded.

Yet given the environmental impacts of LDT, the solution can hardly be a generalized increase in levels of travel activity – which points to trade-offs between the two policy issues discussed here. This is manifested for example, in the case of public investment in high-speed rail, which is sometimes aimed at encouraging a modal shift from air to rail, but can also be seen as inequitable, as it disproportionately benefits the high-income users of long-distance rail. As the importance of LDT becomes apparent, issues such as this will become more present in the transportation debate.

Acknowledgment

This research was funded by the German Research Foundation (DFG) (project ‘Long-distance society - Advancing knowledge of long-distance travel: uncovering its connections to mobility biography, migration, and daily travel’, SCHE 1692/10-1), and the UK Engineering and Physical Sciences Research Council (EPSRC) and Economic and Social Research Council (ESRC), Grant EP/R035288/1 (Centre for Research into Energy Demand Solutions - CREDS).

See Also

Demand for Passenger Transportation; Demand Management and Capacity Planning of Airports; Transportation Equity; Privatization and Deregulation of the Airline Industry; Demand for Air Travel and Income Elasticity; Freight Transport and Logistics; Travel Surveys; Infrastructure Transport Investments; Economic Growth and Regional Convergence; Transport Modes and Globalization; Introduction to Air Transport; Introduction to Rail Transport; Transport Modes and Sustainability; Air Transport

References

- Aguiléra, A., Proulhac, L., 2015. Socio-occupational and geographical determinants of the frequency of long-distance business travel in France. *J. Transp. Geogr.* 43, 28–35.
- Czepkiewicz, M., Árnadóttir, Á., Heinonen, J., 2019. Flights dominate travel emissions of young urbanites. *Sustainability* 11 (22), 6340.
- Czepkiewicz, M., Heinonen, J., Ottelin, J., 2018. Why do urbanites travel more than do others? A review of associations between urban form and long-distance leisure travel. *Environ. Res. Lett.* 13 (7), 073001.
- Demoli, Y., Subtil, J., 2019. Boarding classes. Mesurer la démocratisation du transport aérien en France (1974–2008). *Sociologie* 10 (2), 131–151.
- Frändberg, L., Vilhelmson, B., 2010. Structuring sustainable mobility: A critical issue for geography. *Geogr. Compass* 4 (2), 106–117.
- Frändberg, L., Vilhelmson, B., 2011. More or less travel: personal mobility trends in the Swedish population focusing gender and cohort. *J. Transp. Geogr.* 19 (6), 1235–1244.
- Frick, R. and Grimm, B., 2014. *Long-distance mobility. Current trends and future perspectives*. Institute for Mobility Research.
- Gössling, S., Peeters, P., 2015. Assessing tourism's global environmental impact 1900–2050. *J. Sustain. Tour.* 23 (5), 639–659.
- Janzen, M., Vanhoof, M., Smoreda, Z., Axhausen, K.W., 2018. Closer to the total? Long-distance travel of French mobile phone users. *Travel Behav. Soc.* 11, 31–42.
- Larsson, J., Kamb, A., Nässén, J., Åkerman, J., 2018. Measuring greenhouse gas emissions from international air travel of a country's residents methodological development and application for Sweden. *Environ. Impact Assess. Rev.* 72, 137–144.
- Mattioli, G. and Scheiner, J., 2019. The impact of migration background and social network dispersion on air and car travel in the UK. *51th Annual Universities' Transport Study Group*.
- Reichert, A., Holz-Rau, C., Scheiner, J., 2016. GHG emissions in daily travel and long-distance travel in Germany—Social and spatial correlates. *Transp. Res. Part D: Transp. Environ* 49, 25–43.
- Schäfer, A., Heywood, J.B., Jacoby, H.D., Waitz, I.A., 2009. *Transportation in a climate-constrained world*. MIT press, Cambridge, MA.
- Scott, D., Gössling, S., 2015. What could the next 40 years hold for global tourism? *Tour. Recreat. Res.* 40 (3), 269–285.
- UNWTO and ITF, 2019. Transport-related CO₂ emissions of the tourism sector – Modelling results. World Tourism Organisation, Madrid, Spain.
- UNWTO–UNEP–WMO, 2008. *Climate Change and Tourism: Responding to Global Challenges*. World Tourism Organisation. Madrid, Spain.
- van Goeverden, K., van Arem, B., van Nes, R., 2016. Volume and GHG emissions of long-distance travelling by Western Europeans. *Transp. Res. Part D: Transp. Environ.* 45, 28–47.
- Viry, G., Ravalet, E., Kaufmann, V., 2015. High mobility in Europe: An overview. In: Viry, G., Kaufmann, V. (Eds.), *High Mobility in Europe*, Palgrave Macmillan, London, pp. 29–58.

Further Reading

- Alcock, I., White, M.P., Taylor, T., Coldwell, D.F., Gribble, M.O., Evans, K.L., Corner, A., Vardoulakis, S., Fleming, L.E., 2017. Green'on the ground but not in the air: Pro-environmental attitudes are related to household behaviours but not discretionary air travel. *Glob. Environ. Change* 42, 136–147.

- Axhausen, K.W., Madre, J.-L., Polak, J.W., Toint, Ph. (Eds.), 2003. Capturing long-distance travel, Research Studies Press, Baldock.
- Bows-Larkin, A., Mander, S.L., Traut, M.B., Anderson, K.L., Wood, F.R., 2010. Aviation and climate change—the continuing challenge. *Encyc. Aerosp. Eng.* 1–11. [https://www.research.manchester.ac.uk/portal/en/publications/aviation-and-climate-change-the-continuing-challenge\(8f496302-f3e7-4092-act3-e296e9c3aaa4\)/export.html#export](https://www.research.manchester.ac.uk/portal/en/publications/aviation-and-climate-change-the-continuing-challenge(8f496302-f3e7-4092-act3-e296e9c3aaa4)/export.html#export).
- Dargay, J.M., Clark, S., 2012. The determinants of long distance travel in Great Britain. *Transp. Res. Part A: Policy Practice* 46 (3), 576–587.
- Dobruszkes, F., Ramos-Pérez, D., Decroly, J.M., 2019. Reasons for Flying. In: Graham, A., Dobruszkes, F. (Eds.), *Air Transport—A Tourism Perspective*. Elsevier, Amsterdam.
- Dowds, J., Harvey, C., LaMondia, J., Howerter, S., Ullman, H. and Aultman-Hall, L., 2018. Advancing Understanding of Long-Distance and Intercity Travel with Diverse Data Sources. National Center for Sustainable Transportation.
- Dubois, G., Peeters, P., Ceron, J.P., Gössling, S., 2011. The future tourism mobility of the world population: Emission growth versus climate policy. *Transp. Res. Part A: Policy Prac.* 45 (10), 1031–1042.
- EEA - European Environment Agency, 2014. Focusing on environmental pressures from long-distance transport. TERM 2014: transport indicators tracking progress towards environmental targets in Europe. Publications Office of the EU, Luxembourg.
- Frändberg, L., 2006. International mobility biographies: a means to capture the institutionalisation of long-distance travel? *Curr. Issues Tour.* 9 (4–5), 320–334.
- Gerike, R., Schulz, A., 2018. Workshop Synthesis: Surveys on long-distance travel and other rare events. *Transp. Res. Proc.* V 32 535–541.
- Holz-Rau, C., Scheiner, J., Sicks, K., 2014. Travel distances in daily travel and long-distance travel: What role is played by urban form? *Environ. Plan. A* 46 (2), 488–507.
- ICCT, 2019. The International Council on Clean Transportation, 2019. CO₂ emissions from commercial aviation, 2018. Working Paper 2019-16.
- Kuhnimhof, T., Collet, R., Armoogum, J., Madre, J.L., 2009. Generating internationally comparable figures on long-distance travel for Europe. *Transp. Res. Rec.* 2105 (1), 18–27.
- Mattioli, G., 2020. Towards a mobility biography approach to long-distance travel and 'mobility links'. In: Scheiner, J., Rau, H. (Eds.), *Mobility Across the Life Course A Dialogue between Qualitative and Quantitative Research Approaches*. Edgar, Cheltenham.
- Mitra, S.K., Saphores, J.D.M., 2019. Why do they live so far from work? Determinants of long-distance commuting in California. *J. Transp. Geogr.* 80, 102489.
- Recchi, E., Deutschmann, E. and Vespe, M., 2019. Estimating transnational human mobility on a global scale. Robert Schuman Centre for Advanced Studies Research Paper No. RSCAS, 30.

Urban Freight Policy

Laetitia Dablanç, University Gustave Eiffel, Marne-la-Vallée, France

© 2021 Elsevier Ltd. All rights reserved.

Introduction	278
Urban Freight Impacts Require New Policies	278
Access Rules, Low-Emission Zones, Tolls to Organize Urban Logistics	279
Negotiating With Freight Stakeholders	280
In Search of the Latest Innovations	280
Urban Logistics Planning: A New Field of Policy Intervention	281
Curbside Management: A New Sharing of Public Space	281
De-Polluting Commercial Vehicles in Cities	283
Data Collection, Data Analytics, Urban Freight Modeling, Policy Assessment: Huge Gaps in Policies	283
Conclusion	284
Acknowledgment	284
See Also	284
References	284
Further Reading	285

Introduction

Urban freight policy is the set of policies, mostly at local levels, that target the urban movement of goods in order to reduce its negative impacts (pollution, energy consumption, congestion, accidents) and increase its benefits (provision of urban jobs, promotion of the economic vitality of cities, city logistics innovation). With the development of e-commerce and fast deliveries, urban freight challenges have increased in cities worldwide and have become a topic for policy. The recent covid pandemic has raised awareness of urban freight from policy-makers. This is true for local governments but also at State and higher (European Union, US Federal level, etc.) levels of government. Guidelines are available for cities today, providing data and examples of best policy practices, but several areas of policy intervention are still missing, such as drastic reduction in CO₂ emissions from urban freight, or training and career development in distribution and warehousing. After an initial section on urban freight impacts, the main policy initiatives in urban freight are discussed: access rules and low emission zones; urban freight consultation initiatives; promotion of innovations; urban freight planning; curbside management; promotion of clean delivery vehicles; and data and modelling.

Urban Freight Impacts Require New Policies

Urban freight issues meet societal demands and therefore policies. One pollution peak after another in New Delhi (but also Paris or Milan) bring to the forefront the health risks posed by diesel, used by almost all delivery vehicles. Urban freight emits 52% of transportation-related NO_x in Paris (Coulombel et al., 2018). The stakes are also economic, since air pollution damages the image of major cities. Urban freight is also responsible for 20%–30% of transportation-related urban CO₂ emissions and contributes to climate change. This share is currently growing, and the OECD predicts that it may reach 50% by 2050. Protection against risks, from extreme weather events to terrorism, to pandemics, is a priority for local action, and among the targets that need protection are warehouses close to cities, especially those storing hazardous materials. The organization of the distribution of goods is at the forefront of the plans to prepare cities for exceptional events, such as recently demonstrated by the covid pandemic. On a more common occurrence, technological developments and new forms of consumption have made freight issues very visible for residents and policy-makers. When Amazon set up an urban warehouse in the center of Paris in the summer of 2016, the discussion began at the highest level: “This kind of facility is for retail, even if {Amazon} claim{s} to do logistics. What I am defending is not that they should be banned, but that they should fit into this retail qualification that corresponds to what they do and that they should be subject to the same rules” (Anne Hidalgo, Mayor of Paris). Self-employed couriers on bicycles or motorbikes on the streets of large cities worldwide pose new road safety problems and their situation challenges labor law (Dablanç et al., 2017). These on-demand “instant” delivery platforms also bring thousands of job opportunities to young people and migrants, re-introducing low-skilled jobs into urban centers.

Faced with these new challenges, urban freight transportation policies are taking shape. According to a recent count (Letnik et al., 2019), out of 129 medium and large European cities, 15 have adopted a “freight plan” and 84 include logistics measures in their urban mobility plans. Involvement in urban freight for the public sector can be manifold and at various levels. Municipalities are in charge of the urban road network, which is heavily impacted by commercial vehicle traffic. Cities are also interested in local economic development, and therefore have to make sure that the provision of goods and logistics services is adequate. They control land uses, including the land necessary for logistics activities (warehouses and terminals), and are in charge of local traffic and

parking regulations, including all regulations that relate to delivery vehicles. Finally, cities are concerned with environmental and social issues, which are strongly associated with urban freight activities. Metropolitan-wide agencies in charge of transportation infrastructure and metropolitan master plans are also involved in urban freight issues. Higher levels such as regional, national or federal, look at urban areas and promote policies related to air quality, climate change, job creation, and economic development. They set legislation such as air quality standards, organize data collection framework, and research programs that have a direct impact on urban freight local policies. This is also true at an international level, with an increasing number of organizations publishing reports and guidelines on urban freight (CAF, 2019). Major events such as the organization of the Olympic games provide opportunities to organize a change in urban freight operations citywide. After the 2012 Olympics for example, London developed a clear city logistics strategy (Browne et al., 2014). In many other cities, major infrastructure projects such as the implementation of a tramway or a subway system can also play that role.

Despite these many drivers, mandates and potential objectives, actual strategies of local public policies regarding freight and logistics are quite modest. Many cities still organize their urban freight policy around truck traffic and parking management. They view commercial traffic as something that should be banned or at least strictly regulated, and not as a service they should help organize. Local policies are often very parochial and can be conflicting. In the following sections, we address the main areas of policy intervention, mostly at the local level, and we identify achievements and remaining challenges.

Access Rules, Low-Emission Zones, Tolls to Organize Urban Logistics

The tool that local governments dealing with freight activities use primarily is truck access restriction. These restrictions are based on various criteria such as time windows, or the weight and size (length, surface) of the vehicles. The most famous truck ban is the London Lorry Ban, in place since 1975. Heavy goods vehicles over 18 tons of gross weight cannot circulate at night and weekends. All trucks in Seoul, Korea have been banned since 1979 from the central areas during working hours. Truck access rules based on weight or size policies have several drawbacks. They tend to promote small capacity vehicles (vans, light trucks), which increases total congestion and reduces the efficiency of urban freight. Regulating truck access requires enforcement and control, meaning a sufficient and well-trained staff. Truck access rules are also very local and can contradict each other from one municipality to another. In metropolitan Lyon, in France, there are more than 50 different regulations on truck weights, dimensions and delivery schedules (Heitz and Dablanc, 2019), forcing delivery drivers to choose which of the rules they will follow.

The most recent trends in access restrictions are environmental standards and road pricing. Both can be combined, as is the case for the London congestion charge. Urban tolls are not specific to commercial vehicles, while many environmental zones or “low emission zones” and “clean air zones” are. In cities worldwide, more and more rules target the age of vehicles or their level of atmospheric emissions. More than 200 major European cities have established low emission zones (<http://urbanaccessregulations.eu>). London is the largest (1500 km²) and most restrictive. It is well monitored, with hundreds of cameras recording vehicle plate registration numbers. It concerns heavy goods vehicles and large vans. An “ultra low emission zone” (ULEZ) applies to the city center, allowing only Euro 6 vehicles (built after 2014). By March 2021, only Euro 6 commercial vehicles will be allowed in all Greater London. Greater Manchester has also brought itself into line with London standards. In Madrid, Spain, the rule is based on a national vehicle classification. All Spanish cities should have a plan similar to Madrid’s from 2023, at the request of the central government. Several European cities—Amsterdam, Oslo—now plan for “zero-emission zones” (ZEZ).

New enforcement technologies have increased the efficiency of truck access policies. Automatic control systems such as automatic number plate recognition cameras, mobile enforcement, vehicle positioning and on-board equipment have been introduced in many cities in Asia and Europe. This technology comes at a cost and generates controversy mainly over privacy issues. Urban tolls for goods transport vehicles are, for their part, rare in cities. Singapore Electronic Road Pricing has operated since 1998, with specific fees for commercial vehicles. Oslo, Norway and Stockholm, Sweden, have tolls. Road pricing raises the question of price sensitivity, as operators only gradually change their transport methods when faced with a congestion charge. Since the introduction of the congestion charge in the inner city of London in 2003, the number of lorries in the area has remained almost stable, while at the same time the number of private cars has fallen by more than a third. Milan, which differentiates tariffs according to the age (and therefore pollution) of the vehicles, has a different experience. The average number of daily delivery vehicles in the restricted zone was 13,174 before the introduction of the system and 10,500 afterwards, with no apparent impact on the freight supply of the center of Milan (Rotaris et al., 2009). An interesting effect of pricing or access regulations based on the age of trucks is that they force small urban freight operators, often contractors to large companies, to change their operating practices and modernize.

The promotion of night deliveries (or rather early morning/late evening), and in parallel the reduction of noise caused by deliveries, are new targets of municipal policies. New York City provides subsidies to encourage businesses to deliver at off-hours. According to a survey made in New York City, businesses most likely to change to off peak deliveries are shippers doing their own deliveries (own account transportation), and receivers open during extended hours such as restaurants (Holguin-Veras, 2008). The PIEK program in the Netherlands covers all aspects of research and regulation to reduce noise from night deliveries: rolling stock, handling equipment and driving practices. The national government provides financial help for operators investing in silent delivery equipments, which are vehicles and handling equipment generating noise emissions below 65dB, for night deliveries. Tests have shown that companies delivering at night save 30% in delivery cost and 25% in diesel consumption (Fig. 1).



Figure 1 Encouraging silent delivery vehicles in Paris, France.

Negotiating With Freight Stakeholders

Cities are setting up consultation processes with delivery companies and their organizations. These groups provide collaborative opportunities between private companies and local administrations that otherwise are not willing to work together. “Freight Forums” or “urban logistics charters” have become part of local communities’ vocabulary. Examples are the Charter for Sustainable Urban Logistics in Paris signed by 80 carriers’ and shippers’ associations in 2013, and the Freight Quality Partnerships of the various boroughs of London, for which Transport for London, the Greater London transport agency, is responsible for overall coordination. The process of Living Labs, bodies for consultation, design and implementation that accompany an urban project, often has a variation for urban logistics (CITYLAB, 2019). The Multiactor multicriteria analysis (MAMCA) is proposed by Macharis et al. (2019) as a tool to better involve urban logistics stakeholders. It uses stakeholder objectives as evaluation criteria of the different solutions that are considered. More focused on research and experiments, the “Urban Freight Lab” of Seattle in the United States brings together the municipality, the University of Washington (which runs the laboratory) and the logistics, transportation and real estate industries. When these bodies have reached a certain longevity, they act as a forum for negotiating measures such as setting an urban pricing scheme or setting up a labelling system for carriers, like the Fleet Operator Recognition Scheme (FORS) in London, UK, which awards medals (bronze, silver, gold) to delivery companies on the basis of their environmental and social performance. Certification provides operators with a competitive advantage when bidding for contracts. It gives companies access to data, benchmark information, and training programs for their drivers.

Discussion forums between cities exist at a trans-national level, such as within POLIS, a network of European cities and regions dedicated to transport policies which leads numerous actions on freight transportation, or the Alliance for Logistics Innovation through Collaboration in Europe (ALICE), one of whose committees is dedicated to urban freight. ALICE’s roadmap for 2019 included “the search for new, optimized and sustainable urban logistics business models”. While such business models may not emerge from a European think tank, however well-intentioned it may be, but from the market’s response to actual regulatory or financial access conditions put in place in cities, platforms such as ALICE provide networks promoting the exchange of best practices.

In Search of the Latest Innovations

Many local governments are promoting innovation in urban logistics with efficiency and/or environmental mitigation as objectives. Electrically assisted cargo bikes are increasingly popular with cities. A world cargobike festival took place in Groningen in the Netherlands in June 2019 and another was held in Dublin, Ireland, a few days later, supported by the municipalities. Local administrations open bus and bike lanes to cargobikes, providing these new urban delivery vehicles with a considerable competitive advantage over vans and delivery trucks. In 2019, New York City authorized pedal assist cargo bikes from Amazon, DHL and UPS to operate in the city, using some of the municipal on-street unloading/loading zones at no fee. The rapid development of bike lanes following the covid pandemic further increases cargo-bikes’ advantage. On a more futuristic level, the city of Reykjavik in Iceland authorized Flytrex, a designer of unmanned automated vehicles (UAVs), and retail distributor Aha to organize a regular drone delivery line over one of the city’s bays. Swiss Post operates the drone delivery of pharmaceutical products in partnership with the municipalities of Zurich and Lugano. At the State and Federal levels in the US, authorizations are provided to delivery companies to test and operate drone delivery lines, such as UPS on campuses, and FedEx and Walgreens in association with Wing. Street robots such as Nuro, sidewalk robots such as Startship, are also developing.

In Paris, logistics innovation policy is promoted through the establishment of the “Rolling Lab”, a support program and a facility for urban logistics start-ups. But these innovation policies can be failing in several ways. Much of the innovation support is provided

through projects that focus mainly on high-cost demonstrators whose dissemination remains very limited. More fundamentally, cities are seeing the fast train of e-commerce innovations and deliveries pass by without a coordinated response, with each municipality identifying for example its own policy to manage data sharing, or to regulate digital delivery platforms, or to organize the implementation of automatic lockers in public spaces. Few coordinated public policy initiatives are engaged, in areas where market and business strategies are increasingly global.

It should also be noted that technologies for the management of urban logistics using public–private cooperation are still surprisingly limited. Delivery route optimization software do not take into account, for example, access rules or delivery schedules, nor lane closures related to road works. Similarly, municipalities do not have access to company data on urban freight. Taniguchi et al. (2018) propose ways to develop public–private partnerships through an integrated platform for innovative city logistics using big data from private companies and public agencies.

Urban Logistics Planning: A New Field of Policy Intervention

The logistics real estate market is dynamic in large cities and is seeing a diversification of demand in terms of warehouse sizes and formats. Cushman and Wakefield (2017) estimated the demand for e-commerce urban warehouses in Europe for year 2021, to be between 200,000 and 1 million square meters in cities such as Warsaw, Paris, or London. Logistics micro-hubs are also a sought-after commodity in cities. These developments are limited by competition on the urban land market from activities with higher rental yields (offices, housing). This situation favors logistics sprawl, that is the spatial decentralization of logistics buildings from dense areas to distant suburbs (Dablanc and Ross, 2012), which increases the total approach distances to deliver in the city. Decentralized logistics facilities tend to generate more truck-miles as delivery trucks need to cover much longer distances to reach urban destinations. This adds to regional congestion. A regional overview of logistics land uses is therefore needed. So far, few regional master plans set out guidelines for logistics land use and its links to mobility. As a positive example, the Chicago Metropolitan Agency for Planning sets provisions for better accommodating freight land uses in the communities and providing more efficient freight infrastructure (CMAP, 2018). Some issues are often left unresolved such as transit access to logistics developments. An Amazon's giant warehouse near Amiens in France opened in 2017 without anyone having thought of changing the bus traffic plan to site a bus stop near the warehouse.

A potential strategy is the development of logistics parks (sometimes called freight villages), partly through public investment. Some notable examples include *Güterverkehrszentren* (GVZ) at the outskirts of large German cities, or the Raritan Center in New Jersey. These logistics parks commonly include services for logistics companies located on the site, such as surveillance, catering, fueling and cleaning stations for trucks, overnight truck parking, and night accommodation for drivers. Former industrial areas (brownfields) can be used when developing freight villages. However, necessary remediation efforts (e.g., for the removal of polluted soil) should not be too expensive, as programs for logistics facilities development cannot tolerate high land costs.

At a local level, policies for city logistics now include the control of logistics land use through planning, zoning and regulatory provisions. Many cities impose the building of off-street delivery areas in new commercial or industrial developments. The Tokyo off-street parking ordinance compels all department stores, offices or warehouses to provide for loading/unloading facilities when they have a floor area of more than 2000 m². Cities should also be careful in avoiding inadequate regulations in their building and planning codes. Access to underground parking, for example, whether in public parks or in private buildings, can be very difficult for commercial vehicles, even small ones, because of height and size limits. The Chicago Downtown Freight Study of 2008 mentioned this as an important problem for truck drivers and the municipality consequently included recommendations for building provisions.

Requests for proposals have been made in Paris, in France, for converting urban gas stations and car parks into logistics facilities. Some municipalities welcome multistory warehouses. The logistics developer Prologis routinely builds several-story logistics terminals in downtown Tokyo. A three-story warehouse was opened in Seattle in 2017. The acceptance of such projects is increasingly conditioned to environmental criteria and to a careful attention to the integration of buildings into the urban environment. Some cities accept or even promote the development of multiactivity buildings, with logistics on street levels and other activities on the upper levels. One example is the Chapelle “logistics hotel” (a logistics hotel combines multiple levels and mixed uses) that was opened in Paris in 2018 (CITYLAB, 2018). Built on four levels, it combines an intermodal facility, a warehouse, public sports facilities, an urban farm, offices and a data center. The development of “urban consolidation centers” (UCCs) has often been identified as an interesting strategy to reduce urban freight impacts. These facilities provide a service of bundled and coordinated deliveries, often requiring public subsidies. Up to 200 such terminals existed in European cities in the 1990–2000s. However, due to operating costs, most of them closed down when municipalities could no longer subsidize them. Today, a few UCCs are operating, mostly in medium-size cities: Bristol in the U.K., Modena, Padua and other Italian cities, La Rochelle in France, Yokohama in Japan, and the city center of Gothenburg in Sweden (Figs. 2 and 3).

Curbside Management: A New Sharing of Public Space

Freight operators need good access to dedicated loading and unloading areas, be they public or private, on-street or off-street. The lack of sufficient delivery areas transfers delivery operations on traffic lanes (double-parking) or sidewalks, and leads to congestion



Figure 2 A micro-hub for FedEx by the Seine river in Paris, France.



Figure 3 Warehouse development in suburban areas, Atlanta, United States, 2018.

and traffic accidents. In busy urban areas and city centers, adequate on-street loading and unloading bays must be identified and their use must be better controlled. However, cities generally implement very local solutions for managing delivery vehicles on the road, without common operating procedures emerging. Municipalities have a rather case-by-case response to requests from carriers and retailers. The first national technical guideline on the subject in France was only published in 2013, much later than guidelines for bus stops or bike lanes for example. The City of Paris has organized its own procedures for the design of loading bays: a minimum of one every 100 m in the city streets is required, and a loading bay cannot be less than 10 m long to allow drivers to use a tailgate lift. In London and Paris, bus lanes are open to delivery trucks, but while in Paris trucks can enter them to deliver on the “Lincolns” (delivery areas set up half on the sidewalk and half on a protected bus lane), in London trucks can drive on the “bus and lorry lanes” but not load or unload from there. European municipalities differ considerably in the way they accept new delivery equipment in public spaces, such as automatic parcel lockers (widely refused in France, accepted in Germany) or parking permits for trailers serving as a temporary depot from which cargobikes or delivery droids supply neighborhoods (promoted in Belgium cities, not so much in other cities).

Time sharing is a good way to improve street parking capacity for trucks and vans. On Barcelona’s main boulevards, the two lateral lanes are dedicated to traffic during peak hours, deliveries during off peak hours, and residential parking during the night. The city is also renowned for its dedicated mobility motor squad consisting of 300 agents circulating with a motorbike and controlling all on-street parking activities including loading/unloading. This has prevented illegal long-term parking and made these zones always available to delivery truck drivers. A more ‘intelligent’ management of delivery areas is emerging. However, there are more projects than achievements. In Barcelona, delivery drivers have to register on a smartphone application (AreaDUM) with their vehicle registration number. Once they arrive at the delivery bay, the application identifies their location, which they must confirm. A 30-min window then opens to carry out the delivery operation. The urban planning concept of “complete streets” still needs to include freight and delivery activities. [Conway et al. \(2019\)](#) provides the first comprehensive strategy for a better integration of truck parking and loading infrastructure in street design and management. Diagrams and photos are provided for a more efficient design of street



Figure 4 An electric quadricycle for deliveries, Amsterdam, The Netherlands.

space accommodating multiple users, including commercial vehicles. In Seattle, the Urban Freight Lab looks at the “final 50 feet” of freight activities in cities, testing solutions such as new design for loading bays or common carrier delivery lockers (Urban Freight Lab, 2018).

De-Polluting Commercial Vehicles in Cities

Municipal authorities are seeking to improve the average condition of commercial vehicles’ urban fleets, which are much older than passenger cars and long distance freight vehicles. In the current state of the market and prices, a substantial increase in the fleet of clean vehicles, particularly electric delivery vehicles, cannot be expected without financial support for small operators, which is what many cities are doing, often in addition to national support. Oslo, in Norway, is the city with the widest range of financial and regulatory support for electric vehicles, but the market share of electric vehicles for commercial vehicles is only 2% of new vehicles, whereas it is now more than 50% for private cars: tax aid plays less of a role for commercial vehicles, which are already partially exempt. A detailed analysis taking into account the operating constraints of delivery companies, the continuous improvement of battery capacity and the maintenance of overall public subsidies for clean vehicles estimates the market share of electric delivery vehicles in French cities at only 13% in 2032 (Camilleri, 2018). Subsidies and charging stations are the two main challenges for policies targeting cleaner commercial vehicles in cities. The mayor of London introduced in 2020 grants of around €17,700 for each truck up to three trucks for operators in need of updating their fleets. A van scrappage scheme grant will be doubled. The question of installing charging stations, especially fast charging ones, on public roads is key but many municipalities are hesitant because of the cost of supportive policies. Avoiding costly investments, particularly municipal investments, could be justified in view of the promise made by car manufacturers of the introduction of vehicles with batteries allowing a driving range of more than 200 km and which have a higher load capacity than today’s vehicles. But in the meanwhile, conversion to electric vans is too slow, especially in tense markets such as New York, Los Angeles, Paris or London, in need of fast charging stations. According to the Rocky Mountain Institute, Shenzhen (China) is amongst the cities in the world with the highest number of electric delivery vans (70,000 by the end of 2019). This is due to two specific municipal policies: exemptions from freight vehicles’ time restriction; and strong municipal policy for the implementation of charging stations (Crow et al., 2019) (Fig. 4).

Data Collection, Data Analytics, Urban Freight Modeling, Policy Assessment: Huge Gaps in Policies

A successful urban freight policy includes good knowledge of logistics activities and the use of efficient tools of data collection, processing and modeling, and policy assessment. Urban freight indicators should be collected and examined overtime, in order for cities to understand major trends in commercial vehicle flows, consumer demands in online retail, innovations in logistics services, and impact of policies. However, data collection efforts in urban goods movements are broadly insufficient, whether for B2B data (business-to-business, freight trips made to or from businesses in cities) or for B2C or C2C data (Business to Consumer, Consumer to Consumer, resulting from online transactions). Specific urban freight surveys, of the type of household travel surveys made for passenger transportation, are expensive and municipalities or higher levels of governments are reluctant to pay for them. Examples of comprehensive efforts include: the Freturb urban freight model in France (Toilier et al., 2018), and production from the Center of Excellence for Sustainable Urban Freight Systems (Rensselaer Polytechnic Institute). New sources of data are explored, requiring local governments to partner with freight operators and other data owning business stakeholders such as e-retailers or

telecommunication operators. Anonymized big data from urban freight operations (private and public) are collected and processed by a neutral third party such as an academic center, then returned to municipalities and operators as user-friendly management/prediction tools. Projects to achieve such comprehensive tools are on their way but none is complete yet. In the shorter term, progress is expected on the use of policy assessment indicators by cities, such as the ones provided by Van Rooijen et al. (2008) or Holguin-Veras et al. (2018).

Conclusion

With the rapid changes in patterns of consumption, production and distribution, many of them accelerated by the recent covid pandemic, urban freight challenges have increased in cities worldwide and have become a topic for policy. Governments at various levels have a wide range of policy tools at their disposal to act on urban freight. Paradoxically however, they do not implement them to any great extent, and some areas such as the collection of data and use of urban freight models, training and career development in urban distribution and warehousing, or the reduction in urban freight carbon emissions, are not yet strong targets of policy. There is a lack of a global urban policy framework conducive to sustainable logistics. Basic targets are not yet achieved, such as the elimination of polluting commercial vehicles in cities, or enforcement of labor/market laws toward small transportation companies or digital workers operating illegally. The result is that the most innovative and sustainable freight operators find it difficult to impose themselves. There are many examples of progress made: low emission zones, improved curbside design for loading/unloading bays, the use of Intelligent Transport Systems (ITS) in many of the private, public and public-private city logistics activities, a better consideration of urban warehouses in the planning process, for example. But a more global approach to urban freight policy, on a metropolitan scale at least, with links to concepts such as circular economy, smart mobility, energy transition, has yet to be implemented in many cities around the world.

Acknowledgment

Metrofreight, a Center of Excellence funded by the Volvo Research and Educational Foundations, the French institute of science and technology for Transport, development and networks (IFSTTAR, University Gustave Eiffel), and the Horizon 2020 program from the European Commission have contributed to part of the research that was used and documented in this article.

See Also

Freight Transport and Logistics; Global Movement of Goods and Commodities

References

- Browne, M., Allen, J., Wainwright, I., Palmer, A., Williams, I., 2014. London 2012 Changing delivery patterns in response to the impact of the Games on traffic flows. *Int. J. Urban Sci.* 18 (2), 244–261.
- CAF, 2019. Estrategia CAF en Logística Urbana Sostenible y Segura, CAF Ediciones, 250p. Retrieved from: https://siip.produccion.gob.bo/noticias/files/BI_06012020da39d_5Logiscap.pdf (last accessed February 2, 2020).
- Camilleri, P., 2018. What future for electric light commercial vehicles? A prospective economic and operational analysis of electric vans for business users, with a focus on urban freight. PhD Thesis, University of Paris-East, Marne-la-Vallée, France.
- CITYLAB, 2018. Observatory of Strategic Developments Impacting Urban Logistics, Deliverable 2.1, Citylab Project, European Commission, 242 p.
- Chicago Metropolitan Agency for Planning (CMAP), 2018. 'Mobility', in ON TO 2050 Plan. Retrieved from: <https://www.cmap.illinois.gov/2050/mobility/freight> (last accessed February 2, 2020).
- Conway, A., et al., 2019. Complete Streets Considerations for Freight and Emergency Vehicle Operations. Albany, NY: New York State Energy Research Development Authority, United States. Retrieved from: <https://metrans.org/news/new-metrofreight-publication-a-guidebook-for-considering-freight-in-complete-street-design/> (last accessed February 2, 2020).
- Coulombel, N., Dablanc, L., Gardrat, M., Koning, M., 2018. The environmental social cost of urban road freight: evidence from the Paris region. *Transport. Res. Part D* 63, 514–532.
- Crow, A., Mullaney, D., Liu, Y., Wang, Z., 2019. A new EV horizon, insights from Shenzhen's path to global leadership in electric logistics vehicles. Rocky Mountain Institute report. Retrieved from: <https://rmi.org/insight/a-new-ev-horizon/> (last accessed February 13, 2020).
- Cushman and Wakefield, 2017. Urban logistics, the ultimate real estate challenge? 34 p.
- Dablanc, L., Morganti, E., Arvidsson, N., Woxenius, J., Browne, M., Saidi, N., 2017. The rise of on-demand 'instant deliveries' in European cities. *Supply Chain Forum* 18 (4), 203–217.
- Dablanc, L., Ross, C., 2012. Atlanta: a mega logistics center in the piedmont atlantic megaregion (PAM). *J. Transport Geogr.* 24, 432–442.
- Heitz, A., Dablanc, L., 2019. Freight in urban planning and local policies: results from a new survey in twenty French cities. City Logistics Conference, Dubrovnik, Croatia, June 13.
- Holguin-Veras, J., 2008. Necessary conditions for off-hour deliveries and the effectiveness of urban freight road pricing and alternative financial policies in competitive markets. *Transport. Res. Part A: Policy Practice* 42 (2), 392–413.
- Letnik, T., Marksel, M., Luppino, G., Bardi, A., 2019. Urban freight transport policies and measures implemented in strategic documents of European cities, a review. *Eur. Rev. Reg. Logistics* 2, 11–15.
- Macharis, C., Kin, B., Lebeau, P., 2019. Multi-actor multi-criteria analysis as a tool to involve urban logistics stakeholders. Chapter 13. In: Browne, M., Behrends, S., Woxenius, J., Giuliano, G., Holguin-Veras, J. (Eds.), *Urban logistics. Management, policy and innovation in a rapidly changing environment*. Kogan Page, London, pp. 274–292.
- Rotaris, L., Danielis, R., Marcucci, E., Massiani, J., 2009. The urban road pricing scheme to curb pollution in Milan: a preliminary assessment, working paper 122, Università de Trieste. Retrieved from: <http://www2.units.it/danielis/wp/wp122.pdf>. Last accessed on May 2, 2019.

- Taniguchi, E., Dupas, R., Deschamps, J.-C., Qureshi, A.G., 2018. Concepts of an integrated platform for innovative city logistics with urban consolidation centers and transshipment points City Logistics 3: Towards Sustainable Liveable Cities. In: Taniguchi, E., Thompson F.R.G., (Eds.), ISTE, Wiley, London, 129–146.
- Toilier, F., Gardrat, M., Routhier, J.L., Bonnafous, A., 2018. Freight transport modelling in urban areas: The French case of the FRETURB model. *Case Stud. Transport Policy* 6–4, 753–764.
- Urban Freight Lab, 2018. The Final 50 feet – Urban goods delivery system, research scan and data collection project. Final report, Seattle Department of Transportation, 176p. Retrieved from: https://depts.washington.edu/sctlctr/sites/default/files/SCTL_Final_50_full_report.pdf (last accessed February 3, 2020).
- Van Rooijen, T., Groothedde, B., Gerdessen, J.C., 2008. Quantifying the Effects of Community Level Regulation on City Logistics Innovations in City Logistics. In: Taniguchi, E., Thompson, R. (Eds.), Nova Science Publishers, Inc, 387–399.

Further Reading

- Browne, M., Behrends, S., Woxenius, J., Giuliano, G., Holguin-Veras, J., 2019. *Urban logistics. Management, Policy and Innovation in a Rapidly Changing Environment*. Kogan Page, London.
- Dablan, L., 2007. Goods transport in large European cities: difficult to organize, difficult to modernize. *Transport. Res. Part A Policy Practice* 41, 280–285.
- Giuliano, G., O'Brien, T., Dablan, L., Holliday, K., 2013. NCFRP Project 36(05) Synthesis of Freight Research in Urban Transportation Planning, Washington D.C.: National Cooperative Freight Research Program.
- Holguin-Veras, J., Amaya Leal, J., Sanchez-Diaz, I., Browne, M., Wojtowicz, J., 2018. State of the Art and Practice of Urban Freight Management Part I: Infrastructure, Vehicle-Related, and Traffic Operations. State of the art and practice of urban freight management Part II: Financial approaches, logistics, and demand management. *Transport. Res. Part A Policy Pract.*, doi:10.1016/j.tra.2018.10.037, DOI: 10.1016/j.tra.2018.10.036.
- Sanchez-Diaz, I., Browne, M. (Eds.), 2018. Topical Collection on Accommodating urban freight in city planning. *Eur. Transport Res. Rev.* 10, 55.

Regional Transport Planning

Chia-Lin Chen, Department of Geography and Planning, University of Liverpool, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	286
The Regional Concept of Transport Planning	286
Defining a Region	286
Origin of Regional Problems and Regional Policy	286
The Role of Transport in Regional Development	287
Key Themes of Implementation of Regional Transport Planning	287
Delimitation	287
Integrated Transport Planning	288
Planning Techniques and Goals	288
Institutional Reform and Governance	289
Summing Up	290
References	290
Further Reading	291

Introduction

Regional transport planning, which is undertaken at a subnational level and primarily influenced by changing institutional structure and political discourse, is one of the most elusive concepts of transport planning. A need for planning at a regional level was recognized in the 1920s in advanced western countries mainly on two inter-related grounds; namely, the interdependence of several municipalities and significant issues relating to rapid metropolitan growth beyond individual administrative boundaries, and the growing regional inequality between mega city-regions and other regions of the country. However, the role of transport in addressing these problems has been controversial. On the one hand, it is a challenge to define a suitable regional scope with an acceptable form of governance that can deal with common transport needs more effectively. On the other hand, an economic rationale tends to determine regional transport planning rather than equity concerns. This article aims to discuss the regional concept of transport planning and examine key themes underpinning the regional transport planning practice by drawing on the extensive experiences based in the USA, including delimitation, integrated transport planning, planning techniques and goals, and institutional reform and governance.

The Regional Concept of Transport Planning

Defining a Region

A “region” is an intermediate level between the basic local unit and the state (Dickinson, 1964). However, the definition of a region is slippery by nature. Keating (1998) provides three perspectives to defining a region; namely, administrative, economic, and cultural, arguing that these concepts could coincide and may conflict with one another. First, from an organizational viewpoint, a region is, ideally, a subnation division through formal government legislation. However, the extent to which regional institutions can exert their political power varies with nations. Second, from the economic perspective, a functional city region, generally beyond the fixed administrative boundary, reflects the dynamism and close interdependence among a network of cities based on their functional attributes, the travel to work pattern, and market linkages. Third, from a cultural angle, regions can be defined by history, language, and a sense of regional identity. The geographical scale of a cultural region may vary dramatically.

Origin of Regional Problems and Regional Policy

Hall (2002) distinguishes two scales of planning: the regional/local scale (city-region) and the national/regional scale to address regional problems. The first scale of regional planning reflects problem-oriented thinking. Patrick Geddes’ *Cities in Evolution* (1915) and Lewis Mumford’s *Culture of Cities* (1938) well capture a need for city-regional planning to tackle issues such as sanitation, transport and governance for an area of everyday living. The larger the size of urban expansion, the more urgent the demand for solving city regional problems. Similarly, Wallis (1994a) identifies three stages of social and economic restructuring in metropolitan areas in the USA, which underlie three waves of regionalism reform. The first wave corresponds with the growth of monocentric city-regions during the mass manufacturing and industrialization process. The second wave results in poly-centric metropolitan regions, where suburbs are adequate substitutes for central cities in production and distribution. The third wave attempts to address city-regional development in the globalised world, where a postindustrial society has a much higher proportion in services than manufacturing.

The second scale of regional planning—national/regional planning is different, but relevant to the first scale of city-regional planning, addressing the inequality of regional development. Issues relating to regional inequality hit hard in postindustrial western countries, while approaches to regionalization varied widely between different countries. France is especially renowned for its intervention in tackling excessive concentration of development around Paris (Dickinson, 1964). Before regional divisions emerged, an administrative tier named “*department*” was insufficient for dealing with regional issues. In the mid-1950s, programmed regions were assigned in the national plan and evolved in status by gaining legitimate responsibilities from 1982 onwards.

The Role of Transport in Regional Development

Transport plays essential roles in growth and development. Friedmann and Stuckey (1973) identified the role of transport in three aspects. First, transport can facilitate economic growth in response to demand. Second, transport drives development through a supply strategy that provides transport services ahead of demand, often attempting to solve problems encountered by lagging regions. In this regard, transport is decisive in reshaping spatial relationships; for example, the location of economic activities, a pattern of settlements, and social interactions. The co-evolution of transport and urban structure is exemplary. Third, apart from the connective functions, transport could be a divisive force to destroy a community through severance, which is evident in the environmentalist movement reacting to motorway construction.

Regional transport planning originates from a political view that in addressing emerging regional problems; for example, disparate development between regions. The under-developed regions which could not be left to experience market failure and thus needed to be tackled through regional policy and regional economic planning, in which transport plays a critical part. Oettle (1979) argues that public transport services should provide freedom of movement for underdeveloped regions.

However, the role of transport in addressing regional problems is often contested, in particular concerning rebalancing the effects of transport in lagging regions. Voigt (1979) argues that the instrumental character of transport in regional policy for addressing uneven regional development depends on the intended use of four spatially differentiated socio-economic structuring effects; namely, investment, income, capacity, and degree of freedom for a competitive strategy. Meanwhile, he also highlights the limitations of a transport policy as an instrument of regional policy without a coordinated package of measures that include both transport and nontransport policy areas.

National political economies in differing political settings and broader contexts play vital roles in accounting for varied intervention in transport systems to address regional problems. Bonnafous (1979) highlights that in France and most western European countries the notion of optimal allocation of resources was dominated by economic calculation and justified by profitability, meaning that infrastructure was only developed when an existing provision was demonstrably overloaded and holding back accumulation. However, in the 1970s, restructuring of space received unanimous priority for balancing regional development and the indirect effects of infrastructure. “The transport policy is then no longer assimilated to an optimum sectoral policy but is to be included in a much larger framework” (ibid., p.48). He stresses “transport can only be an auxiliary of a regional policy which employs all the other instruments of planning. None the less, good use of this auxiliary is no more evident” (ibid., p.56).

The UK government’s dominant view of value-for-money in transport investment results in funding resources concentrated in thriving growth places rather than lagging regions (Marshall, 2011). History has shown that transport planning at the regional level was promoted most strongly during 1997 and 2010, although the overall result might not have been as significant as expected. Since the abolishment of regional government in 2010, the England regions outside London struggled to exert proper regional transport planning, and the focus has been placed around city-regional transport planning efforts around large regional cities. Instead, regional transport planning has been implemented more effectively in Greater London, Scotland, and Northern Ireland where devolved power was secured at the regional level.

Blonk (1979) concluded that transport is a catalytic force, both an agent vital for industrial growth and an agent of decline where economic resources, conditions and human endeavor are insufficient to meet the competition of outside areas. Transport plans should support government policy; for example, land use, regional development, and urban policy. Timely coordination and regular reviews and updates of transport policy with regional development and land use planning are essential. Transport modes should be complementary to each other. Integrated transport models should reflect dynamic spatial patterns.

Key Themes of Implementation of Regional Transport Planning

Delimitation

Although the need for regionalization in the field of transport planning is widely acknowledged, defining a boundary for a regional concept of transport planning is contested. There are two available options. First, one can be mutually exclusive in dividing the nation into several regions; the second is to overlap with existing institutions. Which option to choose varies with different countries? Both options exist in the regionalization of France. However, in the USA, an exhaustive and permanent regionalization for transport planning purposes would not seem desirable (DeSalvo, 1973). Referring to Friedmann (1956) who suggests that in any developed economy basic spatial patterns can be expressed between (1) a central city and its surrounding region and (2) one city region and another, Friedmann and Stuckey (1973) argue that city and region can be accordingly treated “as a unit that is characterized by a tight pattern of functional interdependencies and articulated by a system of transport and communication facilities” (p.143). What should be considered when dividing workable transport regions? Should the boundary be identical for

different target groups; for example, freight and passengers and other transport modes? They summarize four approaches proposed by various scholars for delimitation; namely, (1) daily newspaper circulation and a Federal Reserve District boundary; (2) a standard metropolitan statistical area; that is, functional economic areas equivalent to the labor market of a city, based on commuter data; (3) a daily urban area (a radius of approximately 80 miles around urban centers; (4) an urban field, based on the metropolitan core areas and the inter-metropolitan periphery. Since the interaction of activities will determine the form of the region, they argue a region is primarily a taxonomical concept and should be delineated by specific activities on an ad-hoc basis. For instance, state-level intermodal coordination, an inter-state urban corridor, can avoid designating particular transport regions. Under such thinking, the fourth concept, which includes peripheries, literally tends to be ignored and become a remaining issue left by emphasizing metropolitan-centered regions, because those peripheral areas could not demonstrate their relationship with the metropolitan cores sufficiently (*ibid.*).

Integrated Transport Planning

As discussed earlier, the role of transport should be part of an overall regional policy, and the integrated planning process of transport and nontransport measures cannot be overstated. [Klaassen and Bourdrez \(1979\)](#) call for comprehensive transport planning, which should consider simultaneous horizontal and vertical integration. They argue that planning at the regional level should play a strategic role in the reconciliation of the top-down and bottom-up approaches to the entire planning process ([Klaassen and Paelinck, 1974](#)).

Horizontal integration denotes that transport planning should link with plans and development in various areas, including size, growth and composition of the population, education systems, labor market, economic activities, energy demand and supply, transport infrastructure, social infrastructure, living conditions and recreation, the environment and the ecological system, and land use possibilities ([Klaassen and Bourdrez, 1979](#)).

Vertical integration at the regional level implies that transport planning should consider transport plans at both higher and lower levels. Issues of vertical integration can be approached from two directions, bottom upwards and top downwards. Although local knowledge and needs are better grasped from the bottom upwards, difficulties may lie in lack of power at the regional level to coordinate the local conditions. Moreover, adding up the local needs does not necessarily lead to an optimal situation for all. From the top downwards, the issues include bureaucracy, inflexibility, insufficient local knowledge, and fewer practical solutions (*ibid.*).

[Klaassen and Bourdrez \(1979\)](#) argue that reform of local government responsibilities in the fields of traffic and transport might be a necessary precondition for establishing an effective and concerted planning process (*ibid.*, p.71). The aspects of reform include

1. societal integration
2. economic integration
3. administrative integration
4. integration over time
5. integration of modes of transport
6. integration of public and private passenger transport and of goods transport
7. integration of traffic plans and parking plans
8. spatial integration: coordination of inter-city traffic and urban traffic, long-distance rail transport, the primary road network, and the primary-secondary, tertiary, and urban road system

Planning Techniques and Goals

Development of new computer modeling and systems analysis techniques boosted regional transport studies in the late 1950s and 1960s, setting a substantial precedent for technical analysis at the regional level ([Goldman and Deakin, 2000:48](#); [Wallis, 1994a](#)). However, the capacity of transport modeling in reflecting dynamic regional development patterns is questionable. [Hall \(1967\)](#) argues that problems of a systematic methodology developed for metropolitan area transport studies lie in predicting land uses and the underlying social patterns, arguing that it is unlikely to generate satisfactory interpretation for the modeling work. Close communications among experts in different aspects of urban growth and change could improve the level of performance. Similarly, [Klaassen and Bourdrez \(1979\)](#) argue that conventional transport models assume traffic flows could develop exogenously, failing to consider reality.

Techniques used in the regional transport planning process have evolved in response to changing goals. In the early days, congestion issues derived from road building were addressed when efficiency was the primary concern. Cost-benefit analysis (CBA) is widely used to determine value-for-money through monetized calculation while failing to effectively take into account a multidimensional and broader range of goals in regional transport planning ([Aberle, 1979](#)). In the US, both the 20-year long-term plan and the 3-year shorter-term transport improvement program consider seven critical criteria, including economic vitality, safety and security, accessibility and mobility, quality of life (including environment and energy conservation), integration and connectivity, system management and operation, and preservation of existing systems ([McDowell, 1999, 10](#), cited in [Wolf and Fenwick, 2003](#)). A planning scheme justified on benefit-cost grounds could make the rich richer and the poor poorer, failing to address the inclusive and redistributive issues.

There is a general tendency to embrace new tools for assessing a broadened range of goals and strategies, including an increased emphasis on public involvement in regional transport planning. [Handy \(2008\)](#) develops a framework of performance measures,

arguing regional transport planning agencies have worked to revise planning processes and create new tools. She stresses the significant challenges metropolitan planning organizations face in response to a new transportation planning philosophy (*ibid.*).

Although there is a general acknowledgment of a broader range of goals, some equity issues are inherited and structural. Luna (2014) demonstrates that preexisting and persistent residential segregation results in procedural inequity for representation and participation in transport planning.

Institutional Reform and Governance

Institutional reform and governance contribute to effective metropolitan/regional transport planning with varied effects in different countries. There is a consensus that institutional planning is indispensable, including the organization of decision-making and distribution of power among different levels of government. Klaassen and Bourdrez (1979) emphasize that planners should consider government structure, the administrative system, the legal and fiscal framework, and the financial possibilities. Various regional structures, with or without legal status, have been created to address issues through avenues of actions (Dickinson, 1964). However, institutional reform does not necessarily mean a regional tier is required. DeSalvo (1973) argues that addressing regional transport problems should not mean they have to be dealt with by regional authorities.

In the context of the USA, supporters of metropolitan/regional transport planning agree that there are important mechanisms for ensuring local control over federal funding and that they deserve more broad authority to implement the plans they create. At the same time, critics have argued that they usurp legitimate functions of state governments and constitute an unnecessary layer of bureaucracy (Solof, 1998). Public choice theorists argue that “a public economy composed of multiple jurisdictions is likely to be more efficient and responsive than a public economy organized as a single, area-wide monopoly” (Bish and Ostrom, 1973, p.2). The idea of choice is constructed in a competitive environment where local governments are encouraged to facilitate, and individual households and businesses are free to make decisions about where they would like to be located. In this line, because regions are treated effectively as markets, and individual municipalities behave as firms, the boundaries of a region will vary according to the needs created for specific services. However, in this dominant perspective of competition, social welfare programs are not encouraged as taxes will increase costs. Consequently, the responsibility for the redistributive program might not be undertaken (Wallis, 1994a).

Institutional reform

The following paragraphs draw on the extensive experiences in the USA where regional transport planning practice has progressively evolved from the 1960s to illustrate how institutional reform and governance contribute to regional transport planning beyond an unnecessary layer of bureaucracy. Under the 1962 Federal-Aid Highway Act, metropolitan areas with a population of over 500,000 are required to establish a “continuing, cooperative, coordinated” planning process by representatives of local governments or forming metropolitan planning organizations (MPO) to develop regional transport plans.

MPO is defined as “a core area containing a large population nucleus, together with adjacent communities having a high degree of economic and social integration with that core” (the US Census Bureau cited in Solof, 1998). Some state laws create regional transport planning authorities for developing regional transport plans. For instance, the Puget Sound Regional Council is an MPO by the federal law as well as a regional transportation authority by the state law. MPO have been evolved with various forms. For instance, the planning responsibility of MPOs has been further strengthened over the years by the Intermodal Transportation Efficiency Act (1991) and the Transportation Efficiency Act for the 21st Century (1998) where both the 20-year long-term plan and the 3-year shorter-term transport improvement plan are required under the leadership of MPOs. The two plans are not just a wish list but are an identified list with revenue sources. In order to have a right balance in executing the federal mandates and seizing federal funding opportunities as well as being controlled by the state officials in matching funds, MPOs must “both coordinate local interests and achieve a workable agreement on the regional plan, and cooperate with the state highway department to have regional projects funded” (Goldman and Deakin, 2000:49).

The origin and evolution of MPO reform in regional transport planning reflects that the effectiveness of new institutional reform depends on both statutory capacity and internal capability. Before the creation of MPO, restructuring, for example, political unification is regarded as a means of securing comprehensive regional governance, such as annexation in various forms, which were achieved mainly through “legislation actions without resources to a local referendum” and subject to “strenuous objections of suburban municipalities” (Wallis, 1994a:161). The introduction of MPOs was characterized with procedural reforms designed to improve program coordination and comprehensive planning. In the first three decades following its introduction, MPOs gained modest responsibilities, and the basic framework remained intact: a relatively weak regional institution representing the interests of local agencies and often dominated by the State Department of Transport (DOT) (*ibid.*). In the 1990s, MPOs were empowered to be the leading agency in the planning process (including representatives of local governments, operators of major transport modes, and appropriate state officials) and to control some categories of federal planning funds. These new responsibilities would be carried out in partnership with state agencies and a variety of public and private interest groups to enable a consensus-based process for a vision of the region’s future and improving institutional capacity for sustained implementation was (Wallis, 1994b).

Institutional governance

In addition, to institutional reform of empowerment to the metropolitan planning organization, it is crucial to overcome the difficulties of institutional governance in coordinating two distinct and separate policy arenas, namely transport and land use planning in regional transport planning. Goldman and Deakin (2000) identify five types of partnership in order of increasing levels of interaction, shared responsibility, and role equality; namely, consultation, coordination, cooperation, consensus building, and

collaboration. Most MPO activities belong to the first three types of partnership, and it is evident that “a stronger MPO role has not been universally accepted and, in several regions, key actors, public and private, remain opposed to a strong MPO role” (p.70). However, they argue that lower-level partnership could still evolve to a higher-level collaboration, although this is not guaranteed. To devise useful models for progressive practice, they call for research into social learning aspects of partnership development and evolution of regional institutions.

Three factors appear to make metropolitan planning organizations more likely to engage in these coordination practices, namely environmental pressures and contemporary trends, a right balance of land use and transport-related representatives by state laws, and local capacity (e.g., actions by MPO staff and practical leaders) (Wolf and Fenwick, 2003).

Firstly, Federal programs in line with contemporary trends such as a Liveable Communities Initiative have encouraged linking transport and land use transport. However, land-use planning is not regarded as one of the seven factors in the federal mandate. Similarly, considerable progress in integrating transport and other policy areas has been made in various popular planning models, such as multimodal transport planning, station area development (transit-oriented development), in-fill development, and urban growth boundaries. One can imagine that the latest decarbonization plan will play a similar role in facilitating integration further.

Second, state laws that instruct a good balance of both land use and transport-related representatives on metropolitan planning organization boards are instrumental, but local practice varies. Under the new legislation for MPO responsibility, Goldman and Deakin (2000) find that geography-based representation dominates MPO boards, providing limited roles in which transport operators could become involved. For instance, despite these agencies being members of MPOs, they are often excluded from voting. Issues relating to lack of representation on MPO boards appear to be a significant obstacle to their ability to gain strong powers for regional governance. Seattle’s Metropolitan Municipal Council dissolved because its governance structure violated the constitutional principle of one person, one vote (Wallis, 1994a; cited in Goldman and Deakin, 2000).

Third, local capacity plays a role in the variation of coordinating land use and transport among MPOs. Goetz et al. (2002) argue that MPOs vary considerably in their ability to coordinate regional transport planning. Similarly, Wolf and Fenwick (2003) find there is a certain degree of coordination and a limited number of MPOs that lack coordination. Their analysis found three levels of coordination in practice. The first level involves some coordination, but these activities are not significant efforts because there is fear that the coordination would force local concessions for regional land use, and this would establish precedents for plans. The second level of coordination is bottom-up and heavily driven by local policy boards. The third level refers to those extensively and comprehensively coordinating transport and land use planning. Widely praised good models include Portland (Oregon), Minneapolis, St. Paul, and Seattle with the power given by state legislation to involve in coordination.

Wolf and Fenwick (2003) suggest four general strategies for coordinating transport and land use planning: namely, leveraging technical capacity (transport modeling, air quality conformity analyses, GIS services, and scenario analysis), initiating special projects (driven by fund programs and project advocates including MPO staff, policy leaders, and the community), developing ways to educate and involve the community (a strong and credible citizens’ advisory committee is commonly formed and charged with issues of congestion, growth and land use), and tying land-use issues to planning processes (land use considerations are included in either long-range plans or a specially focused corridor plan. In either case, the extent and significance of this land-use component cannot be readily determined because there are good intentions rather than actual plans for implementation) (ibid., 127).

Summing Up

This article examines the nature and development of regional transport planning, mainly drawing on the experience in the USA, which has established and progressively evolved its regional transport planning practice from the 1960s. Although different countries have different background and trajectory and political economy, this article discusses four critical themes including delimitation, integrated transport planning, planning techniques and goals, and institutional reform and governance, which are pivotal in implementing regional transport planning per se. Transport planning goals have been widened over time to embrace multiple-criteria objectives, including equity and inequality issue, thus a holistic and inclusive approach to integrating transport planning. Development should be instrumental in achieving effective regional transport planning and implementation in the dynamic and widening metropolitan and regional development process, including effective leadership, staff competence and credibility, development of a regional ethos, meaningful public involvement, and a cooperative relationship with the lower and higher levels of transport authorities, streamlined and efficient processes, integrated land-use planning, and accountability.

References

- Aberle, G., 1979. Transport and regional policy objectives in cost-benefit analysis. *Transport and Regional Development*. W. A. G. Blonk. Farnborough, Saxon House: 35-44.
- Bish, R.L., Ostrom, V., 1973. *Understanding Urban Government*. American Enterprise Institute, Washington, DC.
- Blonk, W.A. G. Ed., 1979. *Transport and Regional Development: An International Handbook*. Farnborough, Teakfield.
- Bonnafous, A., 1979. Underdeveloped Regions and Structural Aspects of Transport Infrastructure. *Transport and Regional Development*. W. A. G. Blonk. Farnborough, Saxon House.
- DeSalvo, J.S., Ed., 1973. *Perspectives on regional transportation planning*. Toronto, Lexington Books.
- Dickinson, R., 1964. *City and region: A geographical interpretation*. Routledge and Kegan Paul LTD, London.
- Friedmann, J., 1956. The concept of a planning region. *Land. Econ.* 32, 1-13.

- Friedmann, J., Stuckey, B., 1973. In: *The territorial basis of national transportation planning Perspectives on regional transportation planning*. J. S. DeSalvo. Lexington Books, Toronto, pp. 141–176.
- Goetz, A.R., et al., 2002. Metropolitan planning organisations: findings and recommendations for improving transport planning. *J. Federal*. 21 (1), 87–105.
- Goldman, T., Deakin, E., 2000. Regionalism through partnerships? Metropolitan planning since ISTEA. *Berkeley Plann. Rev.* 14, 46–75.
- Hall, P., 1967. New techniques in regional planning: experience of transport studies. *Reg. Stud.* 117–121.
- Hall, P., 2002. In: *Urban and Regional Planning*. Routledge, London.
- Handy, S., 2008. Regional transportation planning in the US: an examination of changes in technical aspects of the planning process in response to changing goals. *Transp. Policy* 15 (2), 113–126.
- Keating, M., 1998. *The New Regionalism in Western Europe- Territorial Restructuring and Political Change*. Edward Elgar, Cheltenham.
- Klaassen, L.H., Bourdrez, J.A., 1979. Integrated transport planning. In: *Transport and Regional Development*. W. A. G. Blonk. Saxon House, Farnborough, pp. 63–78.
- Klaassen, L.H., Paelinck, J.H.P., 1974. *Integration of Socio-economic and Physical Planning*. Rotterdam, NEI.
- Luna, M., 2014. Equity in Transportation Planning: An Analysis of the Boston Region Metropolitan Planning Organization. *Prof. Geogr.* 67 (2), 282–294.
- Marshall, T., 2011. Reforming the process for infrastructure planning in the UK/England 1990–2010. *Town Plann. Rev.* 82, 447–467.
- Oettle, K., 1979. In: *Obligations on transport for underdeveloped regions*. Transport and Regional Development. W. A. G. Blonk. Saxon House, Farnborough, pp. 17–34.
- Solof, M., 1998. *History of Metropolitan Planning Organisations*. Newark, North Jersey Transportation Planning Authority.
- Voigt, F., 1979. In: *Transport and regional policy: some general aspects*. Transport and Regional Development. W. A. G. Blonk. Saxon House, Farnborough, pp. 3–16.
- Wallis, A.D., 1994a. Inventing regionalism: the first two waves. *Natl. Civ. Rev.* 83, 159–175.
- Wallis, A.D., 1994b. Inventing regionalism: a two-phase approach. *Natl. Civ. Rev.* 83 (4), 447–468.
- Wolf, J.F., Fenwick, M., 2003. How Metropolitan Planning Organizations incorporate land-use issues in regional transportation planning. *State Local Government Rev.* 35 (2), 123–131.

Further Reading

- Sciara, G.C., 2017. Metropolitan transportation planning: Lessons from the past, institutions for the future. *J. Am. Plann. Assoc.* 83 (3), 262–276.
- Vickerman, R.W., 1991. In: *Infrastructure and Regional Development*. Pion, London.

Transport Project Financing

Romeo Danielis, Lucia Rotaris, Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche (DEAMS), Università degli Studi di Trieste, Italy

© 2021 Elsevier Ltd. All rights reserved.

Definition of Project Finance	292
Benefits and Risk	292
PF for Transport Projects	294
PF for Transport Projects in Theory	294
PF for Transport Projects in Practice	295
PF for Innovative Transport Projects	296
Conclusion	296
References	296

Definition of Project Finance

There are three possible financing approaches for a project: (1) corporate finance, based on the creditworthiness of the company, (2) object finance, based on the value of the asset, and (3) project finance (PF), used to finance capital-intensive long-term initiatives generating a predictable cash flow, such as the case for transport infrastructure. When the public sector (state government, local government, or transit agency) is the project initiator, PF can be referred to as public–private partnership (PPP or 3P). PF is used for self-contained projects, having a finite life, with high debt to equity ratio, and no or limited guarantees provided by the investors. The future cash flow is used for interest and debt repayment to investors and lenders (Daube et al., 2008).

Within the context of PF, a new company is specifically created for fundraising and the Project Company or Special Purpose Vehicle (SPV) carries out the project. A number of entities, each with a specific role (Fig. 1), are part of project. These are:

- the project company, characterized by an equity and debt structure enabling it to manage the cash flow generated by the project;
- the public sector, generally the project initiator, providing guarantees and acting as project sponsor;
- Lenders, providing debt and allowing the diversification of the project risks across different banks;
- Banks, acting as mandated lead arrangers, participants providing the funds, and advisors and consultants to the sponsors;
- the project sponsors. They are either institutional investors (private equity companies, infrastructure funds, insurance companies, pension funds, government agencies, investment banks, regional development banks) and/or companies (construction companies, suppliers, and operators) investing their capital in exchange for an equity stake in the project;
- Contractors, carrying out planning, design, and construction of the project;
- Input suppliers, providing materials and components of the infrastructure;
- Operators, responsible for maintaining, and operating the infrastructure after completion;
- Insurance companies, taking care of risk insurances;
- Consultants, providing due diligence and help selecting the bank consortium and defining the project structure (Bengtsson, 2019);
- The offtaker, that is, the party buying the product or service that the project produces.

PF is implemented via contractual agreements aimed at specifying the operating and revenue risks and at allocating them among all the parties involved in the project (Estache et al., 2008). The concession agreements assign the company in charge for developing the project the right to construct the infrastructure and to earn revenues by either charging an operator using the infrastructure (e.g., multimodal operators or freight forwarders using logistic platforms or intermodal centers), or by providing a service to the general public (e.g., toll roads). Several types of concession agreements are possible. For transport infrastructures, the most commonly used type is the build–operate–transfer, according to which the project company has only the right to operate the infrastructure and to earn revenues within a finite period, but it does not own the infrastructure which, in fact, belongs to the public sector. A similar and frequently used agreement is the build–transfer–operate. In this case, the ownership of the infrastructure is taken over by the public sector only after the completion of the project. Other types of agreements include leasing, joint ventures, contracting out, design–bid–build, design–build, design–build–operate, design–build–finance–operate, and pure operation and maintenance (Evenhuis and Vickerman, 2010; Kumari and Sharma, 2017).

Benefits and Risk

PF provides a large set of benefits both to investors and third parties. It involves, however, also risks. Along the lines suggested by Yescombe (2013), we divide PF benefits into two groups: the benefits to investors and the benefits to third parties. The former include:

- High leverage. It is one major reason for using project finance. Long-term investments do not offer a high return: high leverage improves the return for an investor.

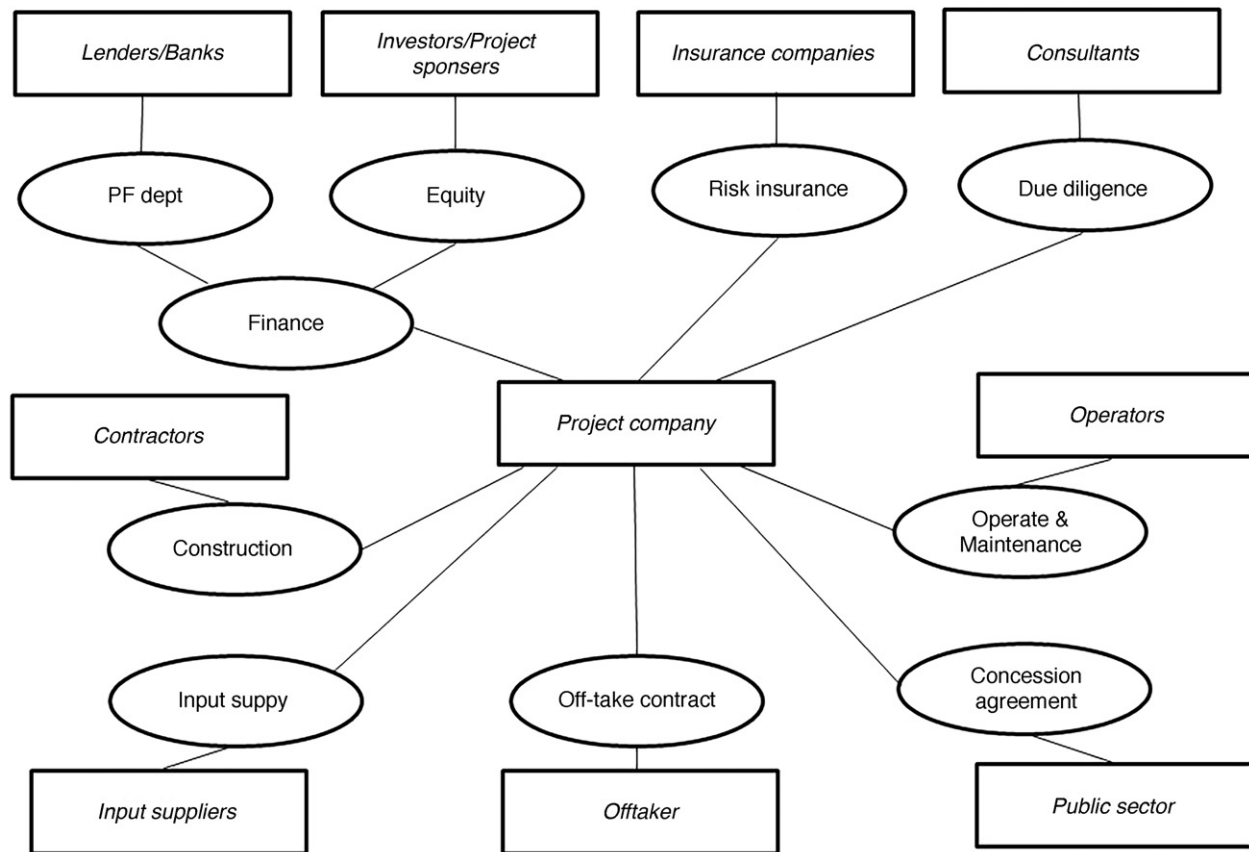


Figure 1 Stakeholders involved in PF. Source: Adapted from Yescombe (2013, p. 8).

- Tax benefits. If interest is tax deductible and dividends to shareholders are not, debt becomes even cheaper than equity, and hence encourages high leverage.
- Off-balance-sheet financing. If the investor has to raise the debt and then inject it into the project, this appears on the investor's balance sheet. On the contrary, a project finance structure may allow the investor to keep the debt off the consolidated balance sheet. This takes place more easily when the investor is a minority shareholder in the project.
- Risk limitation. An investor raising funds through PF does not normally guarantee the repayment of the debt: the risk is therefore limited to the amount of the equity investment.
- Risk spreading/joint ventures. When a project is too large for one investor, others may be brought in to share the risk in a joint-venture project company. This enables the risk to be spread among investors.
- Long-term finance. Project finance loans typically have a longer term than corporate finance.
- Enhanced credit. If the offtaker has a better credit standing than the equity investor, this may enable debt to be raised on better terms.

The benefits to third parties (offtaker or end users of the project) are the following:

- Lower product/service price: when a high level of debt is possible, the price of the product/service will be lower.
- Additional investment in public infrastructure: PF provides funding for additional investment in infrastructure that the public sector might otherwise not be able to undertake because of economic or financial constraints on the public-sector investment budget.
- Risk transfer: a PF contract transfers risks (e.g., project cost overruns) from the public to the private sector.
- Lower project costs: an important characteristic of PF is that the private sector might build and run such investments more cost-effectively than the public sector, even after allowing for the higher costs of project finance compared to public-sector finance. Such a property is dependent on:
 - the general tendency of the public sector to "over-engineer" or "gold-plate" projects;
 - greater private-sector expertise in controlling and managing the project construction and operation, given the private sector ability to offer incentives to managers;
 - the private sector being willing to assume the risk of construction and operation cost overruns;

- a “whole life” management approach rather than ad hoc arrangements for maintenance dependent on the availability of further public-sector funding;
- the absence of the “deal creep” phenomenon (i.e., increases in costs during the construction and operation phase).
- Third-party due diligence: benefits from independent due diligence and control of the project exercised by lenders, who want to ensure that all obligations are fulfilled and risks are adequately dealt with.
- Transparency: since a PF is self-contained, the true costs of the product or service can more easily be measured and monitored.
- Additional inward investments: when successful PF might open up new opportunities for infrastructure investments and can act as a showcase to promote further investments in the wider economy.
- Technology transfer: for developing countries, PF provides a way of producing market-based investments in infrastructures for which the local economy may have neither the resources nor the skills.

PF entails also several risks. [Yescombe \(2013\)](#) groups them into three main categories. First, there are commercial risks inherent to the project itself or to the market in which it operates. Second, the macro-economic risks (financial risks) related to external economic effects (e.g., inflation, interest rates, and currency exchange rates), and lastly, the political risks (country risks) related to the effects of government action (change of laws, contract disputes, expropriation) or political events such as war and civil disturbance, especially in case of cross-border financing or investments. Commercial risks comprise of the following:

- commercial viability risk: the overall financial and economic sustainability of the project;
- on time and on budget completion risks;
- environmental risks during construction or operation;
- operating risks related to achieving the planned costs and performance levels;
- revenue risks related to achieving the expected revenues;
- input supply risks related to acquiring materials or other inputs at the projected costs;
- contract mismatch risks related to the co-ordination of the project contracts;
- sponsor support risk: lack of support from the sponsors.

The core role of PF is to assess, manage, and properly insure against such risks. Risk analysis should be based on due-diligence, risk identification and quantification, and risk allocation through contracts including the lenders.

The risks differ by project phase. Construction projects typically have four distinct phases: planning, construction, operation, and decommissioning. A typical risk in the planning phase is the design error. In the construction phase the typical risk is cost overrun, while in the operating phase there is the risk that the demand for the asset is lower than expected (for further reading, see [Macário et al., 2015](#) for the Portuguese experience with PPP).

PF for Transport Projects

PF for Transport Projects in Theory

Several projects can be realized with PF including ([Weber et al., 2016](#)):

- transport assets (roads, railways, bridges, tunnels, ports, airports, inland waterways, tramways, subways, intermodal centers, and logistic platforms);
- public utilities (oil and gas networks, electricity transmission, allocation and distribution networks, water supply, treatment, disposal systems and waste disposal facilities, telephone and communication networks);
- natural resources projects (oil field or mining, renewable energy projects such as wind and solar parks);
- public buildings (schools, hospitals, administrative buildings, and social housing).

The projects financed via PF are construction projects in which an asset is either built (greenfield projects) or improved, upgraded, or expanded (brownfield projects) ([Rossi and Stepic, 2015](#)).

Infrastructure projects are particularly challenging since the project’s net present value (NPV) can be negative, making investments unattractive for the private sector. However, indirect benefits to communities may make them worthwhile ([Ehlers, 2014](#)). The cash flow time profile might also be unattractive, with high upfront investments and cash flows that cover costs occurring only in the last project phase ([Ehlers, 2014](#); [Sorge, 2004](#)). Finally, infrastructure projects are very complex since a large number of parties might be involved ([Ehlers, 2014](#); [Sorge, 2004](#)), and their success depends on the joint effort of all parties. Coordination problems, conflicts of interest, and free-riding can lead to project failure ([Sorge, 2004](#)). Infrastructure investments, however, can be financially attractive because often these projects offer long-term predictable and stable cash flows that allow the projects to have a high leverage ([Rossi and Stepic, 2015](#)). Additionally, some infrastructure projects protect against inflation, as payments, such as tolls, are connected to inflation security due to PPP arrangements ([Rossi and Stepic, 2015](#); [Weber et al., 2016](#)).

Transport projects are public goods whose consumption is nonexcludable and nonrival, preventing private sector from directly financing them. Moreover, they are strategic assets that critically determine the economic and social development of a country ([Aschauer, 1989](#); [Button, 1998](#); [Calderon and Servén, 2004](#)). They generate both positive (stimulating economic growth and increasing productivity) and negative (environmental impact, pollution, congestion, and visual impact) externalities, which are not accounted for by private investors when evaluating their financial returns. Finally, they give rise to natural monopolies or

oligopolies whose technical and operational standards are regulated by public authorities. For these reasons, governments have traditionally financed their construction, via either public-sector debts or using revenues collected from taxpayers. Moreover, they require large sums for their maintenance (Meunier and Quinet, 2010), but over time the availability of public funds in many countries has reduced.

PF for Transport Projects in Practice

This section provides a review of some country experiences with PF for transport projects. The focus is on the United States, Australia, and European countries.

In the United States, the federal surface transport programs are mostly funded through taxes on motor fuels that are deposited in the highway trust fund (Kirk and Mallett, 2013). The tax rates (in terms of cents per gallon), however, have not been increased since 1993. Prior to the 2007 recession, annual increases in driving and fuel use were sufficient to keep revenues rising steadily. This is no longer the case. Increases in fuel economy standards and the switch to electrification are likely to reduce motor fuel consumption and, hence, tax revenue in the years ahead. The US Congress has financed the federal surface transport program by supplementing fuel tax revenues with transfers from the US Treasury general fund, but it did not address concerns about the funding of surface transport programs over the longer term. Consequently, there has been a growing interest in improving transport infrastructure with private and non-grant funding sources, such as tolls, PPPs, and federal loan programs, but many projects may not be well-suited to alternative financing. Typically, the “public” in PPPs refers to a state government, local government, or transit agency. Since the United States has a comprehensive and relatively mature transport infrastructure network, about 85% of the total public capital investment in transport system in recent years is directed toward maintaining and repairing it, and not toward constructing new facilities or adding capacity to existing ones. Two decades of experience have shown that private investment is attracted to large, complex and expensive transport projects that add new capacity to the US system and can be supported by a revenue stream, usually through tolling. Thus, the overall market share for PPP projects in the overall US transport construction market has been—and likely will remain according to Kirk and Mallett (2013)—fairly small: less than 5% per year.

The value of PPPs contribution, however, should not be underestimated because, putting aside the possibility of significant increases in public funding, they will likely be a primary model for building new highway capacity in heavily congested urban areas in the decades ahead. It is estimated that since 1989, the private sector has been involved in financing 56 US highway projects, roughly worth \$46 billion. Reinhardt (2011) explored the most important objectives of state Department of Transport (DOT) in using PF for highway projects. He finds that the most prominent objectives are:

- developing projects that otherwise would be delayed;
- enabling the agency to expedite the award of the contract to avoid future cost escalation;
- enabling the agency to start project procurement despite funding shortfalls for the project;
- incentivizing project teams to accelerate the completion of the projects;
- enhancing agency’s ability to overcome cash flow constraints.

State DOTs pursue these objectives to develop the backlog of their delayed projects and use deferred payment mechanisms in anticipation of future funding. On the contrary, objectives such as obtaining financing services beyond in-house capabilities, transferring financing and interest rate risks to the private sector, and encouraging competition and innovation are ranked relatively low in the list of major objectives. The relative ranking of objectives, provided by the survey respondents in the study by Reinhardt (2011), shows that state DOTs typically think of PF more as an instrument to bridge their funding gaps and financing shortfalls and less as an innovative solution to gain life-cycle cost efficiencies, encourage competition, and transfer critical project risks to the private sector.

Australia, as many other countries, need to invest significantly in new infrastructure to meet the growing demand of an increasing population. PPPs are a suitable vehicle for some economic infrastructures and can bring design and construction innovation, incentivize management and operational efficiency, early project delivery and improved asset utilization. Regan et al. (2017) reviewed six PPP projects: the Eastlink Motorway and Peninsula Link Freeway in Melbourne, the Airport Link and Clem 7 Toll Roads in Brisbane, and the Lane Cove Tunnel and NorthConnex Toll Roads in Sydney. They found that the Clem 7, Eastlink, Peninsula Link and Lane Cove projects were delivered ahead of schedule. In the case of Clem 7, this is remarkable given the technical challenges of the complex tunneling task. Airport Link also required extensive tunneling, but ran into technical problems causing late delivery and over-budget. Forecasting errors were the main reasons why three of the toll roads projects were placed in administration and another sold at a very significant discount to cost price. The commercial failure of four of the six projects suggests that in the absence of a more collaborative approach to risk sharing, financiers are unlikely to support future toll road projects in Australia. The task for government is to strike a balance between risk allocation and value for money outcomes on a case-by-case basis to sustain the PPP model. Nonetheless, Regan et al. (2017) concluded that PPP proved that large and complex projects can be delivered on time, within budget and produce better standards of services than traditional alternative procurement methods.

During the last few decades, the European Union has been promoting the use of PPPs in order to accelerate the development of the TEN-T network for ensuring economic, social and territorial cohesion and increasing accessibility within the EU (Garrido et al., 2017). To this aim, several mechanisms have been placed at the disposal of the Member States including:

- grants from European Regional Development Fund (ERDF) and Cohesion Fund (CF);
- grants from the European Investment Bank (EIB);

- and a series of innovative instruments such as:
 - the Loan Guarantee Instrument for Trans-European Transport Network Projects (LGT), designed to cover traffic revenue shortfalls during the initial operating period (ramp-up);
 - the EU Project Bonds Initiative, aimed at enhancing the credit quality of project bonds issued by private companies;
 - the Marguerite Fund (MF), a pan-European infrastructure equity fund established to act as a catalyst in the development of infrastructures in the transport and energy sectors.

Garrido et al. (2017) analyzed the case study of Spain and concluded that despite the EU effort to encourage the use of EU funds to support PPP projects, the application of the ERDF, the Cohesion fund and the TEN-T budget line has been very limited. Regarding the innovative financial instruments, their implementation at the time of the writing of this chapter had been quite modest in Europe in general, but particularly in Spain. Instead, the EIB's lending activity to PPPs in Spain is currently the largest EU support source even though the share of EIB loans given to PPP projects compared to other loans has not increased over the last few years.

Villalba-Romero et al. (2015) analyzed quantitatively and qualitatively four PPP case studies concerning road construction: the M-6 in the UK, the R-2 in Spain, the A-23 in Portugal and the Attica in Greece. The sustainability performance is compared against the "conventional" project management. Sustainability is measured by three pillars: economic, social and environmental; while the project performance is measured in terms of quality, cost, and time. They conclude that two cases, that is, A23 Portugal and Attica Greece, can be considered successful, while M-6 UK is more a success than a failure and R-2 Spain can be considered neither a success nor a failure.

PF for Innovative Transport Projects

Beside traditional transport infrastructure projects, government are now struggling to finance the next generation of sustainable transport infrastructures addressing problems such as climate change, air pollution, automobile dependence, and diminishing supplies of hydrocarbons. Initiatives aimed at fostering public transit, carpooling, car sharing, cycling, and walking need adequate dedicated physical infrastructures such as large scale urban rapid transit lines, interurban railroads, dedicated carpooling and car sharing lines, side roads for cyclists and pedestrians, parking areas, and fast charging stations for electric shared vehicles. Technological innovation such as Intelligent Transport Systems (ITS), Cooperative ITS deployment, and Mobility As A Service systems, might represent cost-effective, high added value solutions in order to spread the use of vehicle sharing and of public integrated services (Carvalho et al., 2018). In the coming future automated vehicles used both for public and private transport will need special connected and automated transport infrastructures requiring large investments that public sector cannot possibly face alone. PF could play a relevant role allowing the development and spread of these disruptive innovations (Voegel et al., 2017). The empirical evidence, however, suggests that it might be problematic using PPPs to finance sustainable surface initiatives because of the need to guarantee ongoing capital and operating subsidies from governments to make these initiatives financially viable (Siemiatycki, 2012).

Conclusion

Many countries find it difficult to carry out successfully transport projects. Developing countries might not have public funds available, while developed countries experience public budget restrictions that limit the country's infrastructure development. PPP appears to be an interesting mechanism for empowering countries, both developing and developed, and bring together public and private sectors to enhance transport infrastructures. PPP is not, however, a panacea. In order for PPP to deliver satisfactory results, some necessary preconditions should be met. They include a sound project master and implementation plan, a thorough and realistic cost/benefit assessment, an appropriate risk allocation and risk sharing, a balanced adjustment between the public and the private interests, a transparent and predictable legal framework, and a stable political and economic environment. PPP entails, however, also risks, widely discussed in the literature from a theoretical and practical perspective. Common mistakes leading to unsatisfactory results are overestimation of the demand, misallocation of risk between public and the private partners, and lack of a thorough assessment among alternative projects. Pitfalls result also from an excessive number of agencies dissolving responsibility and introducing inefficiencies and bureaucracy in the PPP processes, frequent changes in the regulatory framework, and lack of information on the gap between the current and the expected costs. When not properly carried out PPP might lead to biased decisions, usually favoring new "gold plated" transport projects instead of improving the existing ones, and an insufficient risk transfer to the private partners, with the public bearing most of the financial burden.

References

- Aschauer, D., 1989. Is public expenditure productive? *J. Monet. Econ.* 23 (2), 177–200.
- Bengtsson, W.J., 2019. Unconscious Bias in Infrastructure Project Finance Decisions, Doctoral dissertation, Technische Universität. Available at: [https://www.researchgate.net/publication/338111111](#)
- Button, K., 1998. Infrastructure investment, endogenous growth and economic convergence. *Ann. Region. Sci.* 32 (1), 145–162.

- Calderon, C., Serven, L., 2004. The Effects of Infrastructure Development on Growth and Income Distribution, Policy Research Working Paper 3400. World Bank, Washington, DC.
- Carvalho, D., Vieira, J., Reis, V., Frencia, C., van Renssen, S., 2018. New ways of financing transport infrastructure projects in Europe, European Parliamentary Research Service. Available at: www.ep.europa.eu/stoa/.
- Daube, D., Vollrath, S., Alfen, H.W., 2008. A comparison of project finance and the forfeiting model as financing forms for PPP projects in Germany. *Int. J. Proj. Manage.* 26 (4), 376–387, doi:10.1016/j.ijproman.2007.07.001.
- Ehlers, T., 2014. Understanding the challenges for infrastructure finance. Bank for International Settlements Working Paper No. 454. Available at: <https://www.bis.org/publ/work454.pdf>.
- Estache, A., Juan, E., Trujillo, L., 2008. Public-private partnerships in transport. The World Bank. Sustainable Development Vice-Presidency. Available at <https://openknowledge.worldbank.org/bitstream/handle/10986/7602/wps4436.pdf?sequence=1>.
- Evenhuis, E., Vickerman, R., 2010. Transport pricing and public-private partnerships in theory: issues and suggestions. *Res. Transport. Transport Econ.* 30 (1), 6–14.
- Garrido, L., Gomez, J., de los Angeles Baeza, M., Vassallo, J.M., 2017. Is EU financial support enhancing the economic performance of PPP projects? An empirical analysis on the case of Spanish road infrastructure. *Transport policy* 56, 19–28.
- Kirk, R.S., Mallett, W., 2013. Funding and Financing Highways and Public Transportation Transport. Congressional Research Service, Washington, DC.
- Kumari, A., Sharma, A.K., 2017. Infrastructure financing and development: a bibliometric review. *Int. J. Crit. Infrastruct. Protect.* 16, 49–65.
- Macário, R., Ribeiro, J., Costa, J.D., 2015. Understanding pitfalls in the application of PPPs in transport infrastructure in Portugal. *Transport Policy* 41, 90–99.
- Meunier, D., Quinet, E., 2010. Tips and pitfalls in PPP design. *Res. Transport. Transport Econ.* 30 (1), 126–138.
- Regan, M., Smith, J., Love, P.E., 2017. Financing of public private partnerships: transactional evidence from Australian toll roads. *Case Stud. Transport Policy* 5 (2), 267–278.
- Reinhardt, W., 2011. The role of private investment in meeting US transportation infrastructure needs. *Public Works Financ.* 260, 1–31.
- Rossi, E., Stepic, R., 2015. Infrastructure Project Finance and Project Bonds in Europe. Macmillan, Palgrave.
- Siemiatycki, M., 2012. The theory and practice of infrastructure public-private partnerships revisited: the case of the transportation sector. Available at <http://www.ub.edu/graap/Final%20Papers%20PDF/Siemiatycki%20Mattii.pdf>.
- Sorge, M., 2004. The nature of credit risk in project finance. *BIS Q. Rev.* (December), 91–101.
- Voegel, T., Godziejewski, B., Macdonald, M., Grand-Perret, S., Merat, N., Rødseth, O., van Schijndel-de Nooij, M., 2017. Connected and Automated Transport, Studies and reports, EUROPEAN COMMISSION Directorate-General for Research and Innovation, Available at: https://ec.europa.eu/newsroom/horizon2020/document.cfm?doc_id=46276.
- Villalba-Romero, F., Liyanage, C., Roumboutsos, A., 2015. Sustainable PPPs: a comparative approach for road infrastructure. *Case Stud. Transport Policy* 3 (2), 243–250.
- Weber, B., Staub-Bisang, M., Alfen, H.W., 2016. Infrastructure as an Asset Class: Investment Strategy, Sustainability, Project Finance and PPP. John Wiley & Sons.
- Yescombe, E.R., 2013. Principles of Project Finance. Academic Press.

ITS for Transport Planning and Policies

Bruno Dalla Chiara, POLITECNICO DI TORINO (I), Engineering, Dept. DIATI – Transport Systems, Italy

© 2021 Elsevier Ltd. All rights reserved.

Definition of ITS	298
Reasons for ITS in Transport Planning	299
ITS Applications	300
ITS: the Main Components	302
Telecommunication Systems	302
Automatic Vehicle Location Systems (AVLS)	303
Automatic Identification Systems (AIS)	303
Data Collection for Traffic Flows or Passengers Through Sensors and Their Related Networks	303
Electronic Data Interchange (EDI)	303
Cartographic Databases and Geographic Information Systems (GIS)	303
Data Collection Through Automated Passenger Counting Systems and Traffic Sensors With the Related Networks	304
Automated Passenger Counting	304
Traffic Data Collection	304
Moving Sensors or Sensors in Motion	305
Static or Portable Sensors	305
Specific Weigh-In-Motion Sensors	306
Wireless Sensor Networks	306
Driver Assistance and Road Safety Technologies	306
The Technological Scenario: A Vision	307
Biography	308
References	308

Definition of ITS

ITS stands for Intelligent Transport System(s). Intelligent comes from the Latin *inter-lego*, which means “I link together” or “I link with or through”; another meaning of *inter-lego* is “I read through”, which signifies being able to read between the lines, in other words being intelligent.

A deep comprehension of “intelligent” implies a wide and long-lasting path for ITS, here intended as “inter-connected” transport systems, which are able to include both humans—that is, travelers, drivers or supervisors of a control room, who contribute with their personal intelligence—and goods. This interconnection also allows motionless communication to be achieved, even during traveling (i.e., remote working, remote-courses, tele-conferences, tele-seminars, and tele-diagnostics), as the person involved is devoted to communication and not necessarily in driving, when needed.

This interconnection, and the presence of humans within the connected sub-systems, should have at least one purpose: to improve the safety, security, quality or efficiency of transport systems for passengers and freight, by optimizing the use of natural resources—including energy sources for the traction and propulsion of vehicles—and respecting the environment.

To achieve such aims, certain procedures, systems (including algorithms) and devices (such as sensors) are required to allow the collection, communication, analysis and distribution of information and data to and from moving subjects, the transport infrastructure and information technology applications.

In the 2030–40s, some ITS devices could have simply been identified as “Transport systems” (TS), without the use of an adjective (intelligent) that nowadays is not taken for granted. In fact, the 2020s can have the goal of “ferrying” the present transportation towards ITS, in order to possibly one day consider ITS as being just ordinary TS (Fig. 1).

Since ITS includes information technologies and—in certain cases—the mobility of people, some of the related topics may also be defined as “Infomobility”; such a concept, which stands for “informed-mobility”, can involve road transport, or any other mode of transport, as well as their mutual interactions, but it remains a part of ITS.

It should be noted that transport systems operating on fixed guideways—such as railways, undergrounds, ropeways and automated people movers—are interconnected and monitored by humans on a regular basis, right from their conception, through control rooms (Fig. 2). Road transport was not conceived in the same way: interconnecting its components is a goal that needs to be achieved through transport planning and a policy.

Finally, many synthetic definitions of ITS have been proposed. A complete one, although not the only one, has been provided by an association of universities that deal with ITS: “Intelligent Transport Systems integrate telecommunications, electronics and information technologies with transport engineering in order to plan, design, operate, maintain and manage transport systems. This integration aims to improve safety, security, quality and efficiency of transport systems for passengers and freight, optimizing

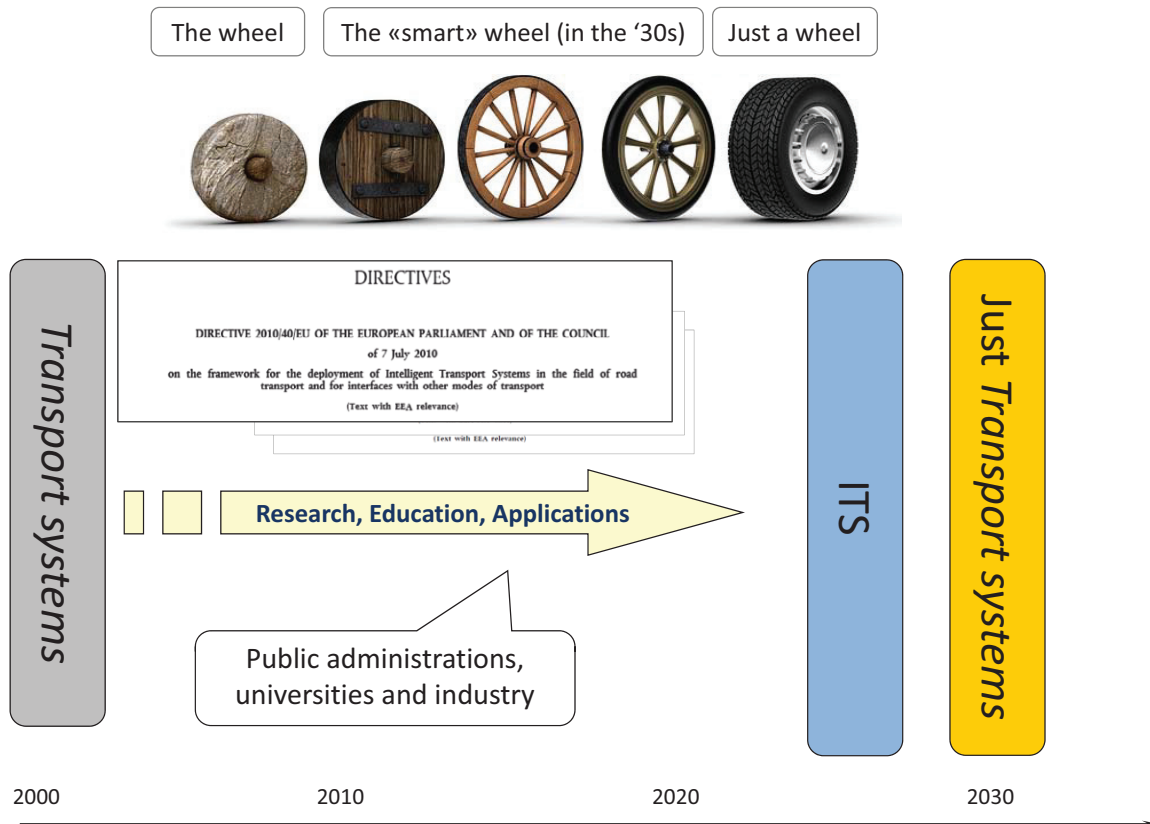


Figure 1 The transition of present transport systems toward renewed ones, passing through ITS, like the wheel in the past centuries



Figure 2 Control rooms for ITS: policy, planning, and operation.

the use of natural resources and respecting the environment. To achieve such aims, ITS require procedures, systems and devices to allow the collection, communication, analysis and distribution of information and data among moving subjects, the transport infrastructure and information technology applications". [ITS EDUNET, 2009].

Reasons for ITS in Transport Planning

A diffused, motorized, personal mobility—a long-term aim that might have inspired transport planning and policies at the beginning of the 20th century, when people and goods were mainly moved by horses and the first tramways—has generated also negative consequences over that century. These include environmental effects, unnecessary queues when traveling, traffic congestion and accidents, which are nowadays among the main causes of non-natural deaths for the under forties in many countries throughout the world. However, this pure motorized mobility is not necessarily informed, interconnected or intelligent.

Therefore, while one of the main goals of the long-term transport policy in the 20th century may have been diffused private motorization, the economic system of the 21st century has been marked by great changes, related to information technology applications, and is more environmentally oriented. These current trends, with reference to transport planning, are aimed at

improving the quality, safety and security as well as the energy efficiency of transport systems, many of which frequently already existed and were already in operation.

ITS is expected to facilitate this migration process towards systems with less queuing and optimised routes, through a better use of both road networks and energy, to prevent accidents, as well as the propagation of collisions, by informing the drivers, by fostering remote access to reservations (e.g., for parking areas equipped with plug-in electrified spots or for rail wagons devoted to road-rail transport of intermodal units) and payments (e.g., free-flow toll systems or free-flow access to parking garages), and by preventing injuries to people and damage to both the vehicles and the environment, in the broader sense. In this context, vehicles from necessity are becoming interconnected, in the pursuit of ITS.

ITS Applications

The main applications of ITS in *road transport* may be classified as follows (Dalla Chiara et al., 2017; Dalla Chiara et al., 2013).

1. **Monitoring traffic and road conditions**, such as vehicle flows, the related interruptions, the environmental situation and other information related to the road system, in order to:
 - a. obtain up-to-date information, including the occurrence of accidents or unexpected events, that can be transmitted to road and multimodal users, both in real-time along the infrastructure and possibly even before they initiate their journeys;
 - b. intervene in road system operations (opening/closing tollbooths according to the traffic, central traffic-light control systems, transit limitations, etc.);
 - c. plan possible interventions along the infrastructure, both under normal conditions and in the case of relevant events (a sports event, flooding, relevant damage of a road, etc.) in order to guarantee a given capacity, level of service and resiliency of a network or part of such a network;
 - d. schedule infrastructure maintenance works, by analysing the daily traffic volumes and trends, together with the environmental conditions.

Traffic data collection and automatic classification systems have the aim of detecting various traffic flow parameters, and this may involve just a simple counting, or establishing the position, spacing, headway, local and spatial speeds of vehicles, the traffic density or even the traffic volume. Different types of sensors or detectors are used to observe traffic parameters, and the use of one or more probe vehicles, equipped with sensors, moving within a traffic flow (a technique known as floating car data) is also possible (§ Data Collection through Automated Passenger Counting Systems and Traffic Sensors with the Related Networks and Fig. 3).

2. **Control and management of traffic-light systems**: traffic can be regulated through the use of stand-alone traffic lights, regulated at a local level, and by traffic control systems that are equipped with control rooms, at a wide area level, frequently on an urban scale.

The definition of a traffic-light control plan is normally determined on the basis of the observation of traffic flows (Fig. 4). These are estimated under the prevailing traffic conditions considering some typical daily trends, but they may also be influenced, in real time—through the abovementioned sensors—by a change in traffic patterns during the day or by the occurrence of emergency conditions that can affect a traffic flow, or even as a result of the introduction of priority policies (e.g., for public transport or emergency vehicles).

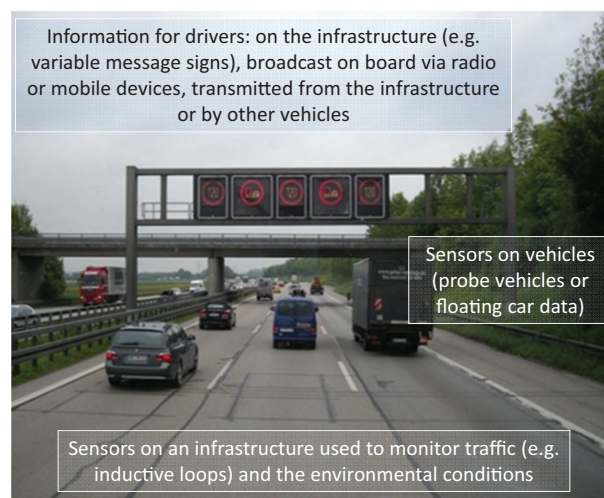


Figure 3 Monitoring traffic and road conditions for transport planning and traffic operation purposes.

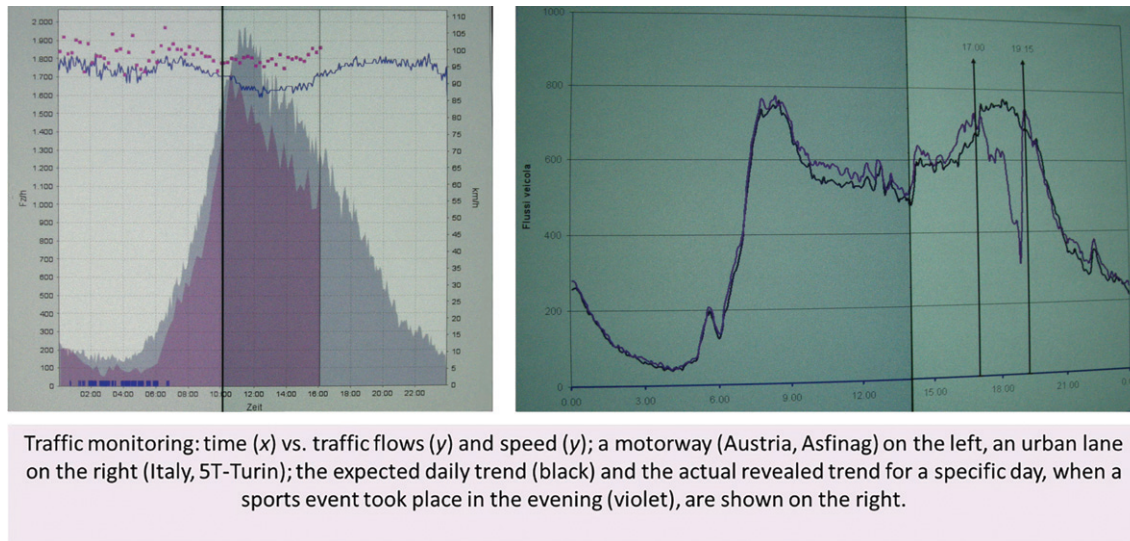


Figure 4 Traffic monitoring: typical daily trends and an actual traffic condition according to data collected by sensors.



Figure 5 Monitored parking areas for either plug-in hybrid or fully electric private and/or shared automobiles.

3. **Control and management of parking**, which is sometimes given the term “park pricing”, is related to the devices and systems adopted to monitor the occupancy and parking duration of vehicles in a given site, and to calculate the respective fees. The devices used to control on-street parking are generally sensors or trans-receivers that are located outside a vehicle (in a parking control system), inside a vehicle (either installed inside a car or on a mobile device) or application accessible through the web (on the cloud). The integration of a number of devices through telecommunication systems, with user information systems and software for data processing, is recognized as a (dynamic) parking guide system. The availability and booking of electric charging for plug-in vehicles should also be considered for recent cases of electro-mobility (Fig. 5).
4. **Control and management of the passage of vehicles (transits)**, through physical or virtual barriers, with or without vehicles stopping. In this case, the entry of a vehicle into an area that is only accessible through one or more protected entrances is allowed through the use of mechanical-electronic devices or by paying a toll. Thus, the control and management of transit points with vehicle stops requires the use of devices to actuate the barriers that prevent access by vehicles not equipped with the instruments necessary to open such barriers and to collect a toll and/or issue a payment ticket. In the case of transit points without vehicles having to stop, the equipment includes devices that recognize the passage of a vehicle, or a person, so that information can be exchanged between the vehicle and the toll collection, or authorizing structure, and then verified. This approach has been applied to transport systems operating on fixed guideways, such as trains and undergrounds, for many decades. As far as roads are concerned, when dealing with the control and management of transit, reference can also be made to *road-pricing*, which can be technologically achieved through the simultaneous use of accurate automatic location systems (§ ITS: the Main Components) and telecommunications. The control and management of transit points for freight transport through ITS is useful for intermodal terminals or port gates, as they can simplify and speed up operations as well as avoid possible human errors. In this case, systems that adopt Optical Character Recognition (OCR) or Radiofrequency transmission Identification (RFid) can be installed and, in some continents, they are also compulsory to facilitate intermodal (road-rail, road-maritime) operations.



Figure 6 Transmitting information on-board through broadcasting, infrastructure-to-vehicle or vehicle-to vehicle information.

5. **Enforcement systems**, which can be introduced to pursue the application of motor vehicle or traffic codes along a road network, often for safety reasons. This objective can be achieved by introducing ICT devices and systems to ensure infrastructures are used correctly, with the application of fines when necessary. Typical examples of these systems are those based on image recognition and sensors or devices to detect the passage of a vehicle (e.g., red light cameras at traffic lights or in limited urban areas) as well as speed violations.
6. **Traffic and traveler information systems**, which use data that are made available by the managers of the road infrastructure or traffic data providers. Information can be transmitted either by means of on-board systems or through ground systems. On-board information may imply either broadcasted information to all the vehicles moving in a given area (a region or an entire nation) or personalized information requested by an individual user. In the last century and at the beginning of the 21st century, road user information was primarily based on RF analogue transmission. However, more up-to-date systems exist, with the related services on the radio broadcast market, such as RDS-TMC (Radio Data System - Traffic Message Channel), DAB (Digital Audio Broadcasting) and DVB (Digital Video Broadcasting), all of which allow on-board information to be transmitted. Other alternative systems that can be used are short-range RF or infrared communication systems (DSRC, Dedicated Short-Range Communication) or even cellular networks (Fig. 6).
Data exchanged through ground devices can be transmitted to travelers through variable message signs or panels installed along a road/motorway network (Fig. 3) or along public transport routes or near interchange areas, and computer kiosks, as well as by means of infrastructure-to-vehicle communication (§ ITS: the Main Components).
7. **Driving assistance, on-board route guidance and navigation**, which are related to driving assistance systems pertaining to the instantaneous operation of the vehicle, information on the road network and any traffic limitations or immediate requests for assistance, the police and ambulance interventions.
Therefore, these pertain to a number of on-board sensors, which can help the drivers attain safer driving, to launch an emergency call, or use a navigation process that suggests how they should drive to a given destination at the minimum cost or in the shortest time.
8. **Control and management of passenger transport services**, a process which involves the monitoring of fleets of various categories of public transport vehicles, such as trams, buses, taxis and shared services (usually cars and bicycles, and nowadays also scooters and electric scooters). The operating conditions, functional conditions, actual occupancy (passengers on board; § Automated Passenger Counting) and position of the vehicles are in particular monitored.
9. **Integration of the above-mentioned applications**, which results in a wide-spectrum control of mobility and freight, that is, control that includes interoperability and compatibility with other systems at an urban, regional, national, and international level.

ITS: the Main Components

Considering the high number of vehicles on roads, which has been rising globally during the last century, many transport systems frequently struggle with congestion, in both space and time. Transportation by itself is no longer enough since society currently not only requires people to be moved and freight to be forwarded or handled, it also requires updated information. ITS can associate information to physical transport devices and can be decomposed into the following basic technological supports or classes of components [1, 2].

Telecommunication Systems

Telecommunication includes a vast array of technological solutions, networks and devices that allow information to be transmitted over various distances. Telecommunications lay the foundations for ITS, as all their applications involve the transfer of data and information, in real time when possible, between moving subjects and information technology applications (sensors, centralized control rooms).

These systems can include different mobile communication technologies:

- fixed networks, for public or private access;
- mobile networks for public or private access:
 - point-to-point and point-to-multipoint communication, with public-access mobile radio networks, such as 4G-5G, or with private mobile networks, such as PMR (private mobile radio), or be based on specific network services dedicated to road transport operators or applications, such as DSRC and co-operative driving, including vehicle-to-vehicle or infrastructure-to-vehicle communication;
 - broadcasting systems (such as DAB and DVB, § ITS Applications), to distribute information to moving vehicles and to travellers' mobile devices.

Automatic Vehicle Location Systems (AVLS)

AVLS is an ensemble of methods and technologies aimed at automatically determining the geographic location of a vehicle or of a personal device (i.e., a mobile terminal) and at making the information available or transmitting it to a requester. The main methods include:

1. the utilization of inertial instruments (such as accelerometers and gyroscopes) and odometers, which have been used on board for some decades;
2. automatic location systems based on tri-lateration or multi-lateration, such as GPS (USA), GLONASS (Russia) and GALILEO (European Union);
3. mobile cellular networks;
4. systems based on automatic identification devices distributed along a given pathway, which are being used more and more frequently, but not necessarily exclusively, for transport systems operating on a fixed guideway (trains, trams, undergrounds and the like).

These methods can be combined together to increase the continuity or the accuracy of each of them.

Automatic Identification Systems (AIS)

Automatic Identification Systems (AIS) are principally used to locate and identify certain attributes of objects passing through an open or closed system. The most important available AIS technologies are:

- radio frequency identification (RFid);
- bar-codes and bi-dimensional codes;
- magnetic-strip cards;
- smart cards;
- identification by means of video technology;
- biometric systems.

RFid and smart cards include, for example, ticketing systems for public transport. A possible application of these AIS is that of Near Field Communication (NFC), a bidirectional wireless connectivity (RF) that operates over a short range, which represents an evolution from contactless identification, or RFid, and other connectivity technologies. NFC allows bidirectional communication when two instruments are close, that is, within a few centimetres; a peer-to-peer connection between them is activated and both can send information.

Data Collection for Traffic Flows or Passengers Through Sensors and Their Related Networks

Traffic data concerning either vehicles or passengers can be measured using sensors of various types, each of which differs according to its application, performances and costs, that is, its purchase, installation and maintenance costs. These sensors and their related networks are analyzed later on (§ Traffic Data Collection).

Electronic Data Interchange (EDI)

EDI represents a simple concept: the use of direct links between computers and information technology applications, even between systems at different sites, to transmit data in a predefined way (e.g., EDIFACT, XML, . . .) that would otherwise usually be sent in printed form or in an informal, not easily manageable way.

Cartographic Databases and Geographic Information Systems (GIS)

A GIS allows data represented on a geographical base to be viewed, understood and interpreted in various ways that reveal relationships, patterns and trends in the form of maps, reports and charts.

When considering any ITS application, it is easy to verify whether it is constituted at least by a communication support (§ Telecommunication Systems) and possibly by one or more of the other components (§ Automatic Vehicle Location Systems (AVLS); Cartographic Databases and Geographic Information Systems (GIS)); therefore, the aforementioned groups of technologies, together with the already mentioned sensors, instruments and data collection systems, may constitute the basic bricks of any ITS: data collection systems are dealt with in detail in the following section.

Their applications can be expected, in various integrated forms, for the third decade of this century in order to attain:

- *smart roads*, which are expected to automatically recognize what is happening above them and then rapidly inform the travelers, through the use of staffed control rooms;
- *assisted driving* (§ Driver Assistance and Road Safety Technologies), as a probable intermediate passage towards automated vehicles.

Data Collection Through Automated Passenger Counting Systems and Traffic Sensors With the Related Networks

As has emerged from the previous section, data collection plays an important role in transport policies, transport planning, traffic engineering and in managing the capacity, in the field of traveler information, road management systems, parking management, as well as emergency management, where reliable traffic data are needed (Dalla Chiara et al., 2017; Dalla Chiara et al., 2013).

Automated Passenger Counting

Passenger counting has always been a key activity in the planning and development of public transport: many agencies measure part of their results on the number of people who use their service. The importance of an accurate analysis of passenger flows has been known for decades; the collection of data has traditionally been the duty of personnel appointed to report the load status on board. Manual collection is still the most widely used method in the world. However, this is unable to provide a continuous, fast and accurate control; hence the need for automated tools, which can be supported—if required—by traditional systems.

The options available nowadays include monitoring systems based on payment methods, such as fare collection machines and the use of smart cards. Technologies identified as APC (Automatic/ed Passenger Counting) are appropriate for this purpose and of great interest for the analysis of data. A great variety of passenger counting technologies are available on the market; no single solution can be considered as the best or the most economically convenient a priori. The modalities adopted to count people on tramways, buses and trains can therefore be summarized in the two following categories, namely:

1. Counting irrespective of a ticket:
 - a. monitoring of single individuals, usually by technologies on board the vehicle;
 - b. monitoring of the overall load on the means of transport, with technologies applied to the pneumatic suspensions or to the ground;
2. Counting related to the ticket, including the possibility of requiring checking-in and checking-out when using a public transport service.

In the former case, the counting can be based on either the detection (or attempted detection) of the single individuals or indirectly, usually by ascertaining the weight of the vehicle under the passenger compartment or on the infrastructure. Whatever method is selected, the objective is that of obtaining a reliable, accurate counting, or—at least—whose margin of error is contained within a few percentage units, compared to the actual number of people present on the public transport vehicle.

Counting irrespective of a ticket is carried out by means of appropriate sensors, which may detect the passage of people through gates, usually the bus, tramway and train doors. A possible solution is based upon the use of infrared sensors, which act as active switches; in the case of passive switches, pyroelectric sensors can be used. In some applications, only infrared ray switches are used, while double sensors (e.g., active and pyroelectric IR) are used to enhance the reliability of the counting.

Sensors applied to the pneumatic suspensions of a vehicle to reveal the weight, and consequently the number of passengers on the basis of an average weight, and video cameras with image recognition can also be successfully used for this purpose.

The counting of passengers by automatic means provides a huge amount of data, in terms of continuous flows of information, but also of diversified information which can be easily transmitted and correlated in order to acquire an overall view of the phenomenon. This context includes the capability of coupling AVLS with APCs in order to merge the data relevant to the load status, linked to the infeed and outfeed times or timetables at the stops, with the position data, for a continuous monitoring of a line.

Traffic Data Collection

Two primary detector technology categories can be distinguished (Klein, 2001):

1. with intrusive sensors, which need to be installed above or under the road surface;
2. with nonintrusive sensors, which implicitly offer the advantages of the traffic flow not being disrupted during their installation and their maintenance.

In addition to intrusive and nonintrusive detector technologies, other advanced methods, such as data from mobile phones, probe vehicles and remote sensing (by satellites or drones, when applicable and whenever conditions do not compromise their usage), could offer the possibility of detecting travel times and traffic volumes.

Moving Sensors or Sensors in Motion

Traffic and road conditions can be monitored by equipping one or more probe vehicles within a traffic flow, a technique that is known as floating car data. This permits the acquisition and real time analysis of space–time variables of motion, and thus of traffic, for comparisons—if necessary—with data from devices external to the vehicle (§ Static or Portable Sensors, together with sensors to measure environmental parameters and polluting emissions, etc.). Information on positions and speeds, acquired instantaneously by the floating car, can be determined by processing the signals coming from an automatic, satellite-based, location system and, in certain cases, by inertial instruments mounted on board such a vehicle.

Static or Portable Sensors

At present, the main instruments available for the automated collection of traffic data installed along a road network can be classified as follows (Klein, 2001).

- **Pneumatic instruments**, such as pneumatic road tubes. This was the first intrusive-type of traffic detector to be used, and it is still sometimes used for the short-term counting of traffic flows because of its simplicity and low cost. Pneumatic road tubes can also detect the volume and speed of vehicles, and classify them according to an axle count and spacing.
- **Inductive Loop Detectors (ILD)**, which are possibly the most commonly used sensors in traffic management applications; they operate by sensing disturbances to the electromagnetic field over a wire coil inserted underneath or inside a road pavement. A lead-in cable runs from a roadside pull box to the controller cabinet. Recent versions of ILD use higher frequencies to identify specific metal portions on the vehicle, which can then be used to classify the vehicles. Inductive loops are sensors that are commonly used for vehicle counting because of their high performance and low cost, compared to other technologies; they can be used to detect the volume, presence, occupancy, speed and class of vehicles using a single or a dual (speed-trap) loop configuration.
- **Magnetic sensors**, which are also known as magnetometers. They measure the disruption in the earth's magnetic field caused by the presence of a metal vehicle. Magnetic sensors are often used in place of loops on bridge decks, where ILDs cannot be installed, and in heavily reinforced road surfaces. Magnetic detectors can detect the volume, speed, presence and occupancy of vehicles, and they are not affected by inclement weather.
- **Piezoelectric instruments** are made up of a piezo-electric material that may convert a mechanical deformation into an electric signal. When a vehicle's axle passes over a piezoelectric device, it generates a voltage by causing electrical charges of opposite polarity to appear at the parallel faces of the piezoelectric crystalline material. The measured voltage is proportional to the force or weight of the vehicle. The adoption of this kind of detection system (generally called Weigh In Motion, WIM; § Specific Weigh-In-Motion Sensors) enables information to be acquired on the typology of the vehicle (load axles, overall weight) which can be added to more conventional information (traffic flows, occupation rates, headways or distance between vehicles, as well as to information on the speed and length of the vehicles themselves).
- **Microwave Radar Sensors**, which represent one the most important non-intrusive detector technologies. Two types of microwave detectors are available: Doppler Microwave Detectors and Frequency modulated Continuous Wave (FMCW) Detectors. The Doppler principle is used to calculate the speed of a vehicle from a continuous wave (CW) microwave radar that transmits electromagnetic energy at a constant frequency. Since only moving vehicles are detected by a CW Doppler radar, the presence of a vehicle cannot be measured by this waveform, and this kind of radar can only detect the volume, occupancy, classification and speed of vehicles. FMCW detectors, which are sometimes referred to as true-presence microwave detectors, transmit continuous frequency-modulated waves from the detection zone, and can detect both the speed and presence of vehicles.
- **Infrared sensors**, which can operate in either active or passive mode. In the former case, the detection zones are illuminated by means of infrared energy transmitted from laser diodes operating in the near infrared spectrum. A portion of the transmitted energy is reflected by any vehicle traveling through the zone. The reflected energy is focused, by means of an optical system, onto an infrared-sensitive element that is mounted at the focal plane of the optics. The infrared sensitive element converts the reflected energy into electrical signals that can be analyzed in real time. The presence of a moving or stationary vehicle is determined by measuring the round-trip propagation time of the transmitted infrared pulse. Active infrared detectors are affected by snow and rain, because short wavelengths cannot penetrate them. Vehicle speed is measured using two fixed beams, one pointed slightly ahead of the other. Passive sensors do not transmit energy: they instead detect the energy that is emitted or reflected from vehicles, road surfaces and other objects in the field of vision and from the atmosphere. Multi-zone passive infrared sensors can measure the speed and length of vehicles, as well as estimate the more conventional vehicle count and lane occupancy.
- **Ultrasound devices**, which are also known as ultrasonic sensors, generally transmit sound energy pressure waves at frequencies that are above the human audible range. Most ultrasonic sensors operate with pulse waveforms and they can provide vehicle count, presence and occupancy information. Pulse waveforms are used to measure distances from the road surface and vehicle surface, by detecting the portion of the transmitted energy that is reflected toward the sensor. The received ultrasonic energy is converted into electrical energy, which is analyzed by means of signal processing electronics. The speed of a vehicle can be measured by transmitting the pulse energy at two calibrated, closely-spaced incident angles. Constant-frequency ultrasonic sensors that measure speeds using the Doppler principle are more expensive than pulse models.

- **Passive acoustic systems**, which are able to measure the flow rate, occupancy and speed of vehicles by detecting acoustic energy in the form of audible sounds. The sounds are produced from a variety of sources from each vehicle, and from the interaction of the vehicle's tyres with the road. When a vehicle passes through the detection zone, a signal-processing algorithm recognizes an increase in sound energy and generates a vehicle presence signal. When the vehicle leaves the detection zone, the sound energy level drops below the detection threshold, and the vehicle presence signal is terminated.
- **Video technology**, which consists of Video Image Processors (VIP), detects vehicles and, in case of need, their number plates, by analyzing video images in order to determine the changes between successive frames and applying Optical Character Recognition (OCR) techniques; some versions also include the tracking of a vehicle in motion, sometimes finalized at road pricing. A VIP system usually consists of one or more cameras, a microprocessor-based computer to digitize and process the video imagery, a software to interpret the images or characters and convert them into traffic flow data. The image processing algorithms in the computer analyze variations in the color or gray levels in the groups of pixels contained within the video image frames. Heavy rain, strong wind and snow can reduce visibility, and the reflection of images from wet road surfaces also affects the performance of VIP systems.
- **RFid and radio wave identifiers**, including Bluetooth e WIFI scanners, are used to count traffic and to identify the origins-destinations of monitored devices, typically at a local scale. Their use is associated with—for example—the identification of road vehicles entering and leaving terminals, in order to quantify both the traffic and time requested for the transit time through the terminal.

Specific Weigh-In-Motion Sensors

Weigh-In-Motion (WIM) is a technological solution that allows the load exerted on a road by a moving vehicle to be measured in order to estimate the static load of the vehicle.

WIM can provide the real-time data of traffic flows, the total weights of vehicles, the weight of each axle, the type of axle and the classification of the vehicles themselves, and therefore results in numerous useful applications related to transport planning and policies, mainly pertaining to road pricing when pricing is also associated with the weight classes of the vehicles.

Their main applications are listed hereafter.

- **Traffic Monitoring and Management**: the management of traffic in real time, the classification of vehicles on the basis of their weights, and the protection of bridges from being overloaded.
- **Road surface design, monitoring and research**: estimation of the equivalent single axle load, the classification of roads (depending on the percentage of heavy-duty vehicles) and the study of the relationships between the axle load and damage to a road.
- **Toll Collection**: whereby an appropriate road user charge is imposed on vehicles on the basis of their axle loads.
- **Environmental studies**: to point out the relationships between emissions and the total weight of a vehicle.

A WIM station is usually composed of:

- a. proximity sensors, which are able to measure flows, headways or distances between vehicles, as well as the speed and lengths of different vehicles (§ Static or Portable Sensors). They indicate the presence or arrival of a vehicle, and they trigger the weight sensors to make them active in order to record at an appropriate moment;
- b. devices and detectors of various kinds, which are able to provide, through dedicated software, the weight of each axis, and the distances between the axles of each vehicle in order to obtain the different transited shapes (of the vehicles);
- c. a supply unit with a central unit, which is able to elaborate and manage information obtained from the sensors and from detectors, such as inductive loops, and to record them;
- d. a connection system of the different stations, by means of a modem, GSM or GPRS, or other communication supports;
- e. dedicated software programs for the analysis and statistical elaboration of the collected data, as well as software programs that are able to guarantee the quality of the measurements.

Wireless Sensor Networks

A Wireless Sensor Network (WSN) is a network made up of small sensor nodes (SNs) that communicate with each other using wireless communication. It combines distributed sensing, computation and wireless communication technologies. Conditions, such as temperature, noise, vibration, pressure, motion and pollutants, can be monitored at a large scale using a spatially distributed WSN. Because of its variety of functions and flexibility of deployment, numerous applications, including road traffic detection and more long-term transport policies, may be developed using WSNs. A WSN can constitute the base of a "smart road" (§ ITS: the Main Components).

Driver Assistance and Road Safety Technologies

As far as automobiles and, in general, road vehicles are concerned, their evolution during the second and third decades of the 21st century was closely related to ITS and in particular to their connectivity (vehicle-to-vehicle and vehicle-infrastructure communications) and assisted driving, which one day may be autonomous. Mere passive safety features (such as safety belts, headrests and air

bags), although still vital, have been associated with other technologies, and other safety factors have also been considered important since the beginning of this century, namely:

- **active safety**, which ensures that the driver has the greatest possible control of a vehicle and increases the likelihood of avoiding an accident;
- a **preventive system**, which prevents the occurrence of those conditions that could generate an accident. Such a system is relevant to foster concentration, comfort and any other aspect that can facilitate driving and inform a driver about the road conditions and potential hazards.

As far as active systems are concerned, this definition also includes road signs – whose preventive function is the very reason for their existence – with future communication systems or image detection on-board software that may be able to show the aforementioned road signs on the instrument panel of a vehicle. Moreover, the roads themselves, when well designed, developed and maintained, might also represent an element of active safety. The main safety devices or systems that have been quite common on newly designed cars since the second decade of this century include ABS (Anti Brake-Locking System), ESC (Electronic Stability Control), VDC (Vehicle Dynamic Control), BDC (Brake Dynamic Control), TCS (Traction Control System) and EBD (Electronic Brake Distribution). However, other active safety systems also exist, namely:

- anticollision systems;
- alert systems for the communication of hazards or obstacles;
- systems to detect a driver's behavior and conditions, or for the automatic correction of driving errors.

Modern road vehicles are generally equipped with such systems, that are usually recognized as Advanced Driver Assistance Systems (ADAS), as: Intelligent Speed Adaptation (ISA), Seat belt reminders, Alcohol Interlock Systems (AIS), In-vehicle event data recorders (EDR), Autonomous emergency braking systems, Anti-lock braking for motorcycles and Lane Keeping Warning Devices (LDW).

It is necessary to bear in mind that a great number of road accidents with casualties occur under poor visibility conditions, that is, at night or in the case of fog. Different types of sensors are, or could be, used to obtain information about what is occurring around and ahead of a vehicle. The most frequently used or usable ones in the automotive industry are ultrasound sensors, infrared sensors, radars, lidars and artificial vision.

Each sensor operates over a different range of electro-magnetic spectrum frequencies and supplies partial information on the space around the vehicle. A combination of these sensors might provide better results than the use of a single one. The first applications of these systems were for the automatic control of speed. Using either a microwave or laser radar, a vehicle manages to perceive the presence of another, slower vehicle in the same lane and therefore slows down to adapt to the speed of the preceding vehicle. This function is called Adaptive Cruise Control (ACC).

As far as the lateral control of a vehicle is concerned, sensor systems, based on cameras, may help a driver to keep to the same lane and prevent him/her from going off the road. Research is also being conducted on automatic steering control systems (lane keeping). Other examples of drive support functions are night vision (based on infra-red cameras), blind spot detection (the driver can be informed about the presence of an overtaking vehicle in the blind spot area on the side of the vehicle), sign recognition and support during parking maneuvers (parking sensors and, in more recent vehicles, steering-control). All these technologies allow a vehicle to be more integrated with the environment where it is driven and to support the interconnection of the driver, the vehicle itself and the infrastructure, thereafter pursuing ITS as defined above.

Inter-Vehicle Communication (IVC) systems play a significant role in enhancing the performance of traveler information systems. Inter-Vehicle Communication can be defined as the communication and cooperative exchange of data between vehicles via wireless technologies, within a range that can vary from a few to a few hundred meters; vehicle-to-vehicle (V2V) is another frequently used term when referring to direct communication between vehicles, without any external support.

Vehicle-to-Infrastructure (V2I), Infrastructure-to-Vehicle and Road-to-Vehicle Communication are all forms of cooperative wireless interaction between an infrastructure and a vehicle system, which are used to improve the safety and performance of both the infrastructure and the vehicles; V2I could also imply mobile Internet access, to provide support for the mobility management of vehicle users. V2I communication requires an infrastructure, through which the participating mobile units (vehicles) can transmit their data to the control center.

The Technological Scenario: A Vision

The first sets of technologies and applications of ITS for transport planning and policy were based on data collection and system monitoring, pertaining to the mobility of people, to traffic and to the transport of goods, with the related logistics. Such systems resort to already existing optical, electric, magnetic, radar, laser, acoustic, piezoelectric, pneumatic, pyroelectric and beam-based devices, and they also include Wireless Sensor Networks (WSN). The WSN paradigm consists of a large ensemble of heterogeneous, battery or self-powered sensors, interconnected to one another by means of wireless technology. The use of such interconnected wireless devices generates a widely and densely deployed infrastructure of sensors, which collect information about the utilization of the road surface by vehicles, their number and respective weights, as well as the environment conditions.

The merging or fusion of such a huge amount of data allows their points of departure and final destination (O/D) within the network to be known, in order to ascertain the flows, and information about the use of the network, the road surface conditions, the

rapid identification of accidents, traffic congestions, the level of service (LOS) of a road and the like to be rapidly identified. Such data can be compared or integrated with other data obtained from probe vehicles, through the so-called floating car applications. All these aggregated data can be used to attain a significant improvement in safety through the optimization of the management and control of a road network and through the tailoring of the road-maintenance schedules, and can even be associated with a Geographical Information System (GIS).

The communication of these data to an infrastructure manager can be either automatic or simply assisted, and they become instrumental in providing support for local applications as well as for more extensive purposes, such as flow monitoring and control through dynamic route guidance. All the collected and processed data can become useful at both the planning and operational level: for fleet management, for both passengers (i.e., public transport services, taxis, shared vehicles, multimodal platforms that is, the usage of different transport means through a common card, as the concept of mobility as a service, identified with the acronym MaaS) and for freight transport, as well as for the related logistic activities, such as the monitoring of containers, swap bodies and semi-trailers in intermodal transport contexts. On-board navigators, for private or personal use, may also resort to the information made available by the mobile Internet access scenario, and thus accidents, queues, dense fog and interruptions on carriageways can be avoided in real time, or the route taken to reach a certain place or event of interest may be changed.

Biography



Bruno Dalla Chiara is full professor at the Politecnico di Torino in the field of Transport Systems (since 2019) and was previously associate professor in the same university and academic field. He holds a Master of Science in Mechanical Engineering, with specialization in transport systems, engineering and economics (1993), and a Ph.D. in Transport Engineering, with specialization in ITS (1997). He holds or has held the chairs, at the Politecnico di Torino, in "Rail transport systems, urban transit and rope installations", "Transport systems and external logistics", "Mobility" and "Transport planning". He is the author of several publications, mainly concerning the engineering and design of transport systems, ITS, rail and intermodal transport.

References

- Dalla Chiara B., Defflorio F., Carboni A., 2017 "The basic technologies for ITS and their applications: the present and future of traffic and vehicle monitoring", Chapter 2. Volume "Intelligent Transport Systems (ITS): Past, Present and Future Directions", Editor: Gaetano Fusco, Series: Transportation Issues, Policies and R&D, 2017, ISBN: 978-1-53611-830-8, Nova Science Publishers, Hauppauge (NY), United States of America.
- Dalla Chiara, B., Bifulco, G.N., Fusco, G., Barabino, B., Corona, G., Rossi, R., Studer, L., 2013. "ITS nei trasporti stradali - Tecnologie, metodi ed applicazioni", a cura di B. Dalla Chiara, Ed. EGAF, pp. 1-477, Marzo 2013, ISBN 978-88-8482-477-6. ["ITS in road transport - Technologies, methods and applications", curated by B. Dalla Chiara, Ed. EGAF, pp. 1-477, March 2013, ISBN 978-88-8482-477-6].
- Klein, L.A., 2001. *Sensor Technologies and Data Requirements for ITS Applications*, Artech House, ISBN 978-1-58053-077-4.

Transitions and Disruptive Technologies in Transport Planning

Kate Pangbourne*, **Maria Attard†**, *Institute for Transport Studies, University of Leeds, Leeds, United Kingdom; †Department of Geography, Faculty of Arts, University of Malta, Msida, Malta

© 2021 Elsevier Ltd. All rights reserved.

Introduction	309
Concept of Transitions	309
Concept of Disruptive Technologies	310
Issues and Uncertainties Arising From Transitions and Disruptive Technologies	311
Conclusions	312
See Also	312
References	312
Further Reading	313

Introduction

The purpose of this chapter is to highlight how the continuous processes of socio-technical transition impact on the field of transport planning, and to reveal how at times, developments converge to make certain technological developments more “disruptive” to the status quo and to the ability to plan. The history of transport systems, at least over the last 300 years, is one that has been profoundly shaped by technological invention (Rodrigue et al., 2009) which is closely entwined with an ever-growing demand for transport as the developments increase human reach, both temporally and spatially and become more embedded in wider social practices (Rinkinen et al., 2020).

In transport, there is a lot of anticipation about the potential for new technologies and services to deliver solutions to some long-standing and rather intractable problems, such as congestion, air pollution, and carbon emissions. Concern in the health sector regarding the impacts of sedentarism arising from overdependence on cars also drive behavioral policy moves to increase active travel. It is increasingly important that cobenefits of change can be identified and promoted, and also that nontechnological solutions are deployed alongside.

Many of the current developments in transport involve the use of digital technologies in some form, and can be grouped under the descriptor “smart mobility.” However, this is a fuzzy concept, “perhaps more a label with currency than anything specific” (Marsden and Reardon, 2018, p. 2). Nevertheless, it is one which has achieved that currency precisely because there are a number of key shifts occurring, involving both electricity and digital communications technology (the “smart” part), which are already impacting on how we get around: electrification of vehicle fleets (which itself consists of several different technologies); growth in online retailing and working away from the workplace impacting on previous trip patterns; the convergence of mobile Internet, geospatial tools, and handheld portable devices underpinning the rise of the “app,” generating enormous amounts of new data and data traffic to and from consumers; the latter development is leading to a proliferation of new business models for shared and micro-mobility services such as station-based or free-floating bike hire, car sharing, ride-hailing, and parking management. Further away from widespread use are fully autonomous, or self-driving vehicles, though many existing vehicles increasingly have a number of automated features, such as hands-free parking, or lane control. The infrastructure itself is also increasingly instrumented, gathering and transmitting data that is used to monitor, manage, and enforce.

While traditionally these innovations are seen as meeting some kind of need or demand (to boost or unstick the economy for example), it is increasingly evident that technological innovation shapes that demand in ways that are usually unanticipated, even once the phenomenon is recognized (e.g., the concept of induced traffic demand). There has also been inertia in the deterministic forecasting systems that are in common use for transport planning and project appraisal, in which existing trends are extrapolated out into the future with little reference to either social or technological change. Before proceeding to explore the challenges (such as uncertainty) that these innovations or “disruptive technologies” pose to transport planning, the concept of transitions is explained.

Concept of Transitions

The concept of transition is one involving change, moving from one state to another state. As change is a continuous, uneven and inevitable process, this can pose some challenges for efforts to plan, steer, and control human activities. Furthermore, the nature of large-scale transition is often not clear until some way into the process—at what point during a process of change can we say that there has been a fundamental shift in the balance of power, the way things are done, in the state of an ecosystem or network or in terms of which technologies are dominant? Kemp and Loorbach (2003) suggest that transitions take one or two generations. Climate change is a particular example that illustrates how difficult it is to be certain about whether tipping points are close to being reached

or have already been passed, as any single incident of severe weather cannot with certainty be said to be “because of climate change,” but patterns of increasingly severe weather are certainly evident yet responses to the threat are highly variable (Eriksen et al., 2011). Transitions in ecological states are a driver of calls for human activities to change to be less damaging, to mitigate damage that is already occurring, or to adapt to changes that cannot be reversed. In this sense, transition is something that we seek as well as experience. As many of the changes that are seen as necessary are both societal and technological, including within the field of transport, the term socio-technical transition is a key concept, though there are related concepts, such as transformative change (e.g., O’Brien, 2016).

There are several strands of literature, utilizing a range of theories and conceptual frameworks to consider processes of transition and transformative change in the context of global challenges of climate change and sustainability, which collectively have influenced the emergence of the field of “transition studies” (Rauschmayer et al., 2015; but see also suggested further reading). What they tend to have in common is an emphasis on the connectedness of social and ecological processes, the pressing need for conscious and inclusive steering of change to avoid the worse consequences and the role of technology as just part of a more profound set of changes, not a solution in its own right. Some of the literature is theoretical in seeking explanations of what is (e.g., socio-technical transitions theory, and the associated multilevel perspective), whereas other contributions are more action oriented, such as transition management, which aims to operationalize transitions toward sustainability in governable ways, through a focus on the role of actors in processes (Jhagroe and Loorbach, 2015).

In terms of transport planning, the transition could be seen as *paradigmatic* (see Kuhn, 1962) with the new planning goals, tools, and processes linked to sustainable mobility being introduced in a rather piecemeal manner, whilst old planning goals, tools, and processes linked primarily to reducing congestion, still dominating the existing institutions and practices (Bertolini et al., 2008). The transition for transport planning remains therefore a complex challenge in many places around the world. A number of authors have provided insights into transitions thinking in transport planning, with some advocating a broader view to include case studies of nonwestern socio-technical transitions (see Hopkins et al., 2020). Some interesting examples are found in Ghosh and Schot (2019), Sengers and Raven (2014), and Sunio et al. (2020).

Concept of Disruptive Technologies

The concept of a technology being disruptive refers to the potential for introduced inventions to disturb how things have been done up until that point. Indeed, in relation to transitions, some degree of disruptiveness might be seen as the point, in that there is a perceived need to disturb the status quo. In the transport sector, this refers to the possibility that the new services and business models have a large impact on the practices of incumbents, often to the point of threatening their viability as going concerns. Indeed, the business models of the new actors are often explicitly posing a challenge to existing providers, as the promoters perceive an opportunity to profit from better meeting the needs of the customer or end-user in some way (Marsden and Reardon, 2018). Ride-hailing companies which match drivers to passengers are a commonly cited example of this, for the way they have used weak regulatory systems to aggressively compete against traditional taxi operators and licensed private hire companies, something which is explored in more detail below.

A new technology-enabled user-centric business model that integrates multiple transport modes is mobility as a service (MaaS). This is a platform-based concept that is at the forefront of the minds of decision-makers, though it is contested as to whether it is a truly new concept (Lyons et al. 2019). Whilst there are only a small number of extant examples of MaaS that are operating as more than just pilot schemes, there is an increasing amount of literature, gray and academic, exploring the potentials and implications of MaaS to provide integrated and efficient transport. Networks have formed to promote the concept (e.g., Europe’s MaaS Alliance or MaaS Scotland in the UK) and it is being encouraged by policy-makers. In Scotland, Transport Scotland have launched a Mobility as a Service Investment Fund to stimulate near-market pilot schemes and trials. In Finland, which is commonly held to be the origin of MaaS as a term, a legislative change was made in order to facilitate MaaS’ entry into the transport ecosystem, though this is not an approach that has yet been replicated elsewhere (Hensher et al., 2020). However, MaaS Alliance, which is a public–private partnership body, are clear that regulatory facilitation should happen, ultimately compelling transport service providers to digitize their practices and develop APIs to feed timetables, real time information and booking/payment facilities to end users via MaaS platforms. This makes the platform provider a new broker in the system, with the power to negotiate price with the service providers, and then develop bundles of services and payment options that are marketed to the consumer.

The disruptiveness of MaaS is the breaking of the direct relationship between transport service providers and their users. While this is perceived as reducing the cognitive effort of trip planning, there are likely to be some unanticipated negative impacts, such as for overall levels of demand for travel by different modes, as well as some socio-distributional effects with equity implications (Pangbourne et al., 2020). However, there are other individual developments, such as ride-hailing (Uber, Lyft), micro-mobility (bike and scooter sharing) and, to a certain extent the concept of shared autonomous vehicles, that have dominated much of the discussion about disruption in transport. The common denominator among this apparent diversity is the ubiquitous technologies, such as GPS and smartphones, that have facilitated their growth and development. Attard et al. (2016) provide a review of such technologies in the transport sector which have and will evidently continue to impact transport in the future.

The entry into the transport ecosystem of all of these innovations poses challenges for transport planning. First, the producers of these inventions are looking for markets and seeking to ensure that their products are accepted and facilitated. The alliances that form around certain developments, such as MaaS, are very active in pushing for acceptance, including advocating for legislative and

regulatory enablement. Second, it takes time for evidence to emerge as to the effects, good, or bad, of the technologies either as they are developed through industry and government supported trials and pilot schemes or once they are operating as start-ups funded wholly through the private sector. This hampers the ability to appraise and compare potential innovations and transport infrastructure projects. Third, the new forms of mobility create new affordances which impact on social practices in a mutually shaping relationship, which increases the challenge of steering for transport planning, as many of the shifts in practices are not directly within the sphere of transport. Finally, some technologies seem to have more fundamentally disruptive potential than others, and thus demand different levels of attention and generate differences of opinion about their worth from transport authorities and planners.

For example, [Stone et al. \(2018\)](#) examine the preparedness of Australian transport agencies in relation to planning for the deployment of autonomous vehicle (AV) technologies. On the whole, as the development of fully autonomous vehicles has been “five years” away for some time, despite public funding going into trials in many countries, it is very hard for transport planning to take the possibility of *profound* disruption into account in their usual planning horizons. Consequently, there is still a general tendency to take dichotomous positions based on utopian/dystopian visions and scenarios, with the position on car dependency often being a key determinant. [Stone et al. \(2018\)](#) suggest that some transport planners would argue for AV if they are designed to be integrated with existing public transport services, as well as with active travel (walking and cycling) because this would be the most efficient way of providing better accessibility for everyone while reducing overall traffic congestion and pollutants. However, if the business model used for AV is one of individual private use (owned or leased), then none of these benefits can be guaranteed, and congestion could worsen. [Stocker and Shaheen \(2017\)](#) reviewed work examining the likely impacts of different AV business models and found no clear picture, due to the small scale of existing trials, but identified a wide range of possible impacts. Given the almost total uncertainty over when the technology will be “street ready” and widely adopted, with the initial regulatory effort mostly focused on the safety and legal aspects of AV, there is little that transport planners can do to prepare strategically for the potential disruption, something that was revealed empirically by [Guerra \(2016\)](#).

Issues and Uncertainties Arising From Transitions and Disruptive Technologies

In the absence of empirical evidence of the effects of candidate innovations, we can look to past transitions for insight around the types of issues that can arise. [Geels \(2005\)](#) analyses the transition pathway in the shift to steam power on land and in international shipping, describing the disruption to the incumbent regime of horse and sail power. [Urry \(2004\)](#) analyses the transition to widespread car dependence and dominance through the conceptual lens of automobility. [Rinkinen et al. \(2020\)](#) describe many examples of interlocking technological and social developments that collectively shape and reshape the evolution of demand for energy and transport through their impact on social practices, including the timing of collective endeavors such as commuting or education and how that impacts on such diverse activities as transport, holidays, domestic tasks such as laundry, dishwashing, and cooking, and even when we watch television, with the concomitant impacts on the timing of peak energy or transport network usage.

The work of Geels et al. on the multilevel perspective of socio-technical transition highlights that there are many barriers to innovations moving from their niche (whether that is a laboratory or a demonstration project) to the regime level, where novelty has become more mainstream or resulted in a profound transition. A key barrier is the existence at regime level of “lock-in” mechanisms, which underpin the reproduction of existing ways of meeting transport needs, though some incremental change may occur ([Geels, 2012](#)). However, there are also multiple pathways: [Geels and Schot \(2007\)](#) describe transformation, reconfiguration, technological substitution and de-alignment/re-alignment (see also [Berkhout et al., 2004](#) or [Smith et al., 2005](#)).

Despite the abundance of technological and modal innovation, a transition to sustainable mobility has been slower and more difficult to achieve than in other sectors. With transport’s contribution to climate change at over 25% of the EU-28 greenhouse gas emissions in 2015, it is unlikely that innovative technologies alone will be enough to reach the decarbonization targets set by governments worldwide (see [Marsden et al., 2014](#)). Evidence from research shows that most of the so-called disruptive technologies are not achieving stated objectives of reducing energy consumption or modifying significantly travel behavior. Part of the problem lies in the potential uptake and impact of disruptive technologies in transport planning and transport policy in general, though the wider issue of nontransport social practices should not be underestimated, as highlighted in [Rinkinen et al. \(2020\)](#). For example, simultaneous sharing of smaller vehicles (various forms of lift-sharing) is commonly promoted as increasing sustainability through efficiency gains, but remains persistently unpopular though sequential sharing of vehicles from cars to bicycles to e-scooters enjoys a growing niche following ([Marsden et al., 2018](#)).

The rise of disruptive technologies have, over time, provided challenges to transport planning, and this is primarily influenced by the (cyber) libertarian approach adopted by many of these new technologies and the concerns by some over the underlying principles driving these new technologies and services ([Epstein, 2015](#)). The issues of regulation and the role of government have also been raised, in part, because of the lack of regulation, which has allowed for most of these disruptive technologies to start operating in the transport sector ([Attard et al., 2016](#)). The incumbent transport sector has generally been heavily regulated because of its focus on public infrastructure that support economies and social welfare, as well as for safety reasons. Part of the disruption is also featured in the new “role” of governments and politics to ensure sustainability in planning of future transport systems. So far, evidence of this is scarce, which suggests that further work needs to be undertaken to integrate disruptive technologies (even when they become mainstream such as, e.g., Uber) into the more traditional centralized transport planning systems. There also needs to be a wider public discussion about the role that transport performs in supporting livable places and longer term sustainability in its widest sense, that could inform more holistic appraisal ([Anciaes and Jones, 2020](#)). Concerns about the tendency for traditional

benefit-cost based appraisal to undermine the apparent viability of more sustainable transport investments have been raised consistently for at least a decade (CROSS REFERENCE CHAPTER 10645). More recently, the UK-based cycling charity Sustrans has characterized the issue with existing appraisal techniques as being one that could be resolved by more clearly basing it on a vision that put the least technological/most sustainable modes (i.e., walking, wheeling, and cycling) at the top of a hierarchy of modes, with shared modes and public transport coming next, before car-based modes (Sustrans, 2018).

Conclusions

In this chapter, we have described the concept of transitions, how this applies to the transport sector and challenges incumbent regimes through the potentially disruptive effects of many of the technological developments that continuously emerge in different forms of niche. As the niche innovations are of different types, fostered by different actors, have different rationales and unknown impacts, this creates a challenging environment for transport planning and governance, as many of the affordances of the innovations result in transitions in wider social practices, over which transport planning has no jurisdiction. There is little concrete evidence to understand how scaling-up of innovations affects both sustainability issues (such as social equity, environmental quality, and climate change mitigation) and the viability of existing transport services. There is also a lack of knowledge as to how to model and forecast future transport demand and envision desirable outcomes. This hampers the ability of transport planners to identify which innovations should be encouraged within a finite system as well as how they should be regulated and managed. We have focused only on innovations, which are currently very prominent either on the streets or in the discourse, and have not considered some of the other disruptive innovations that are being mooted or in research and development, such as flying cars, hyperloops or low earth orbit long-haul travel. Of those technologies that are still somewhat “sci-fi,” only autonomous vehicles seem close to realization, but nevertheless the prospect that one or more of these technologies might prove to be “the one” that solves all the problems can tend to distract policy-makers and planners from decisive action in the here and now, falling back instead on traditional solutions, which are often to provide increased road capacity, and often overlooking opportunities to make investments that encourage mode shift (itself a transition) to active travel and existing public transport options due to the limits of the currently available tools for appraising possible interventions.

See Also

Climate Change and Transport Policy; Car Sharing; Planning for Public Transport with Automated Vehicles; Ride Sharing; ITS for Transport Planning and Policy; Sensors and Data Driven Approached in Transport; Mobility as a Service; Electric Mobility; Technology Enabled Data for Sustainable Transport Policy; Connected and Autonomous Vehicles: Priorities for Policy and Planning; Evaluation in Transport; Transport Externalities in Policy and Planning

References

- Anciaes, P., Jones, P., 2020. Transport policy for livability – valuing the impacts on movement, place, and society. *Transport. Res. A: Policy Pract.* 132, 157–173.
- Attard, M., Haklay, M., Capineri, C., 2016. The potential of Volunteered Geographic Information (VGI) in future transport systems. *Urban Plann.* 1 (4), 6–19.
- Berkhout, F., Smith, A., Stirling, A., 2004. Socio-technological regimes and transition contexts. In: Elzen, B., Geels, F.W., Green, K. (Eds.), *System Innovation and the Transition to Sustainability: Theory, Evidence and Policy*. Edward Elgar, Cheltenham, pp. 48–75.
- Bertolini, L., le Clercq, F., Straatemeier, T., 2008. Urban transport planning in transition. *Transport Policy* 15, 69–72.
- Epstein, R., The libertarian: “Uber and innovation”, 2015, Hoover institution. Retrieved from <http://www.hoover.org/research/libertarian-uber-and-innovation>.
- Eriksen, S., Aldunce, P., Bahinipati, C.S., D'almeida Martins, R., Molefe, J.I., Nhemachena, C., O'brien, K., Olorunfemi, F., Park, J., Sygna, L., Ulsrud, K., 2011. When not every response to climate change is a good one: identifying principles for sustainable adaptation. *Climate Dev.* 3 (1), 7–20.
- Geels, F.W., 2005. The dynamics of transitions in socio-technical systems: a multi-level analysis of the transition pathway from horse-drawn carriages to automobiles (1860–1930). *Technol. Anal. Strategic Manage.* 17, 445–476.
- Geels, F.W., 2012. A socio-technical analysis of low carbon transitions: introducing the multi-level perspective into transport studies. *J. Transport Geogr.* 24, 471–482.
- Geels, F.W., Schot, J., 2007. Typology of sociotechnical transition pathways. *Res. Policy* 36 (3), 399–417.
- Ghosh, B., Schot, J., 2019. Towards a novel regime change framework: studying mobility transitions in public transport regimes in an Indian megacity. *Energy Res. Social Sci.* 51, 82–95.
- Guerra, E., 2016. Planning for cars that drive themselves. *J. Plann. Educ. Res.* 36, 210–224.
- Hensher, D., Mulley, C., Ho, C., Nelson, J., Smith, G. and Wong, J., Understanding MaaS: Past, Present and Future. 2020, Institute of Transport and Logistics Studies Working Paper. University of Sydney.
- Hopkins, D., Kester, J., Meelen, T., Schwanen, T., 2020. Not more but different: a comment on the transitions research agenda. *Environ. Innov. Soc. Transit.* 34, 4–6.
- Jhagroe, S., Loorbach, D., 2015. See no evil, hear no evil: the democratic potential of transition management. *Environ. Innov. Soc. Transit.* 15, 65–83.
- Kemp, R. and Loorbach, D., *Governance for sustainability through transition management*. Paper for EAEPE 2003 Conference, 2003, Maastricht.
- Kuhn, T.S., 1962. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, London.
- Lyons, G., Hammond, P., Mackay, K., 2019. The importance of user perspective in the evolution of MaaS. *Transport. Res. A: Policy Pract.* 121, 22–36, doi:10.1016/j.tra.2018.12.010.
- Marsden, G.R., Dales, J., Jones, P., Seagriff, E. and Spurling, N., All Change? The future of travel demand and the implications for policy and planning, 2018, Commission on Travel Demand, pp. 1–63.
- Marsden, G., Mullen, C., Bache, I., Bartle, I., Flinders, M., 2014. Carbon reduction and travel behaviour: discourses, disputes and contradictions in governance. *Transport Policy* 35, 71–78.
- Marsden, G., Reardon, L., 2018. Introduction. In: Marsden, Reardon, (Eds.), *Governance of the Smart Mobility Transition*. Emerald Points.

- O'Brien, K., *Climate change and social transformations: is it time for a quantum leap?* 2016, WIREs Climate Change.
- Pangbourne, K., Mladenović, M., Stead, D., Milakis, D., 2020. Questioning mobility as a service: unanticipated implications for society and governance. *Transport. Res. A: Policy Pract.* 131, 35–49.
- Rauschmayer, F., Bauler, T., Schöpke, N., 2015. Towards a thick understanding of sustainability transitions: linking transition management and social practice. *Ecol. Econ.* 109, 211–221.
- Rinkinen, J., Shove, E., Marsden, G., 2020. *Conceptualising Demand: A Distinctive Approach to Consumption and Practice*. Routledge.
- Rodrigue, J.-P., Comtois, C., Slack, B., 2009. *The Geography of Transport Systems*, 2nd edition. Routledge.
- Sengers, F., Raven, R., 2014. Metering motorbike mobility: informal transport in transition? *Technol. Anal. Strategic Manage.* 26 (4), 453–468.
- Smith, A., Stirling, A., Berkhout, F., 2005. The governance of sustainable socio-technical transitions. *Res. Policy* 34, 1491–1510.
- Stocker, A. and Shaheen, S., Shared automated vehicles: review of business models. International Transport Forum Discussion Paper No. 2017-09, 2017, Organisation for Economic Co-operation and Development (OECD), International Transport Forum, Paris.
- Stone, J., Ashmore, D., Scheurer, J., Legacy, C. and Curtis, C. Planning for disruptive transport technologies: how prepared are Australian transport agencies? In: Marsden and Reardon (Eds.), *Governance of the Smart Mobility Transition*, 2018, Emerald Points.
- Sunio, V., Argamosa, P., Caswang, J., Vinoya, C., 2020. The state in the governance of sustainable mobility transitions in the informal transport sector. *Res. Transport. Bus. Manage.* 100522.
- Sustrans, 2018. <https://www.sustrans.org.uk/our-blog/opinion/2018/october/transport-appraisal-a-pathway-to-poor-decision-making>. Accessed 15 October 2020.
- Urry, J., 2004. The 'system' of automobility. *Theory Culture Soc.* 21 (4–5), 25–39.

Further Reading

- Castán Broto, V., Glendenning, S., Dewberry, E., Walsh, C., Powell, M., 2014. What can we learn about transitions for sustainability from infrastructure shocks? *Technol. Forecast. Social Change* 84, 186–196.
- Cohen, M.J., 2012. The future of automobile society: a socio-technical transitions perspective. *Technol. Anal. Strategic Manage.* 24, 377–390.
- Geels, F., Kemp, R., Dudley, G., Lyons, G. (Eds.), 2012. *Automobility in Transition? A Socio-Technical Analysis of Sustainable Transport*. Routledge, New York.
- Geels, F.W., Schot, J., 2007. Typology of sociotechnical transition pathways'. *Res. Policy* 36, 399–417.
- Kemp, R., Rip, A., Schot, J., 2001. Constructing transition paths through the management of niches'. In: Garud, R., Karnoe, P. (Eds.), *Path Dependence and Creation*. Manwah, NJ, Lawrence Erlbaum Associates, pp. 269–302.
- Kivimaa, P., Virkamäki, V., 2014. Policy mixes, policy interplay and low carbon transitions: the case of passenger transport in Finland. *Environ. Policy Governance* 24, 28–41.
- Hopkins, D., Schwanen, T., 2018. Governing the race to automation. In: Marsden, Reardon, (Eds.), *Governance of the Smart Mobility Transition*. Emerald Points.
- Lyons, G., Davidson, C., 2016. Guidance for transport planning and policymaking in the face of an uncertain future. *Transport. Res. A: Policy Pract.* 88, 104–116.
- Mäkinen, K., Kivimaa, P., Helminen, V., 2015. Path creation for urban mobility transitions. *Manage. Environ. Q.* 26, 485–504.

Community Transport: Filling the Gaps for Those in Need of Mobility

Ian Shergold, University of the West of England, Bristol, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	314
Delivery Models and Services Offered	314
Funding of Services	315
Evidence of Benefits	316
Challenges Facing Community Transport	316
Funding	317
Meeting Demand	317
Operational Issues	317
Looking to the Future	317
Drivers of Growing Demand	317
New Transport Solutions (Technology and Operating Models)	318
Responding to the Climate Emergency	318
Conclusions	318
References	318
Further Reading	319

Introduction

The UK Community Transport Association (CTA), a representative body for providers in the UK, defines community transport in the following way:

“Community transport is about providing flexible and accessible community-led solutions in response to unmet local transport needs, and often represents the only means of transport for many vulnerable and isolated people, often older people or people with disabilities”
(CTA UK, 2020)

Other countries also have services for these groups, but may use different terminology, for example, paratransit in the USA or special transport services in Japan.

Community transport can be described as “formal” flexible transport, in that services are often regulated in respect of vehicles, drivers and operations. This as opposed to “informal” flexible transport, small-vehicle (often unregulated) passenger services with flexible routes, schedule and fares that are a common element of transport systems in the developing world (Wright et al., 2014). Informal flexible services serve a more traditional public transit customer (but confusingly can also be termed paratransit).

Delivery Models and Services Offered

Globally, a range of approaches, or delivery models are used to help those with mobility needs. In the UK community, transport is normally provided by voluntary sector organizations for the benefit of their local community. These organizations typically operate on a not-for-profit basis, with services regulated under UK legislation (the 1985 Transport Act, section 19 and 22), which describes permissible operating regimes. Perhaps the key differentiator between public transport and community services in the UK being the use of volunteer drivers in the latter.

This network of services has grown organically over the years, with little centralized planning or organization, meaning no standard level of provision across the country (this also seen in other countries like Australia). This has led services to develop where there is need, and where there are people looking to serve that need. As a consequence, there are approximately three thousand operators providing services under section 19 and 22 rules in Great Britain (DfT, 2018), with over 15 million passenger trips delivered in England in 2013/14 (House of Commons Library, 2015). In some areas this has meant overlapping services, or different services provided by different organizations. Thus, one settlement might be served by a community minibus to access healthcare as well as a community bus serving trips to shopping and a separate scheme providing mobility for the young unemployed. Rural providers of community transport have additional challenges, needing to deliver their services to a more diffuse population, and over longer distances as facilities and services might be located further from where people live.

Australia also commonly adopts a not-for-profit model based on community need, which provides targeted and flexible services. It has a significant network of community services, which have developed differently in the various Australian states, with no single overriding model or definition. The USA takes a different approach, here services are commissioned by local public transportation companies to supplement fixed-route public transport for the less-able in order to meet equality legislation. These services may be

delivered by private transportation companies, community groups or not-for-profit organizations. The availability (and use) of paratransit services in the US was further boosted by the Americans with Disabilities Act 1990 (ADA), which expressly mandated such an alternative being available. Other community transport-style services using elements of these differing approaches can be seen in various European countries, and in nations such as Canada, Japan, and New Zealand.

In the UK, community transport operators often emerge from local community groups or voluntary sector organizations with concerns for specific groups in society. Operations range in scale from a group of concerned citizens in a rural setting offering a car-based community scheme through to large organizations such as Hackney Community Transport (see <http://www.hctgroup.org/>) in the UK, who provide services in London and several other cities (as well as operating commercial bus services to generate income). Other community services might be operated by local branches of national charities, or by community hubs, which offer a range of welfare and wellbeing services including transport. In other instances, a common journey destination (i.e., a Doctors surgery) might also initiate a service.

Operators of services in the UK (and elsewhere) utilize many different types of vehicles, and deliver solutions in both rural and urban areas. They support journeys for purposes ranging from health and education to employment and social activities. In general, services are demand responsive in respect of routing (although not always in respect of timing), taking people from door to door. Increasingly though this is being supplemented by fixed route and scheduled services, such as community buses (again driven by volunteers). The latter often emerge as replacements to conventional public bus routes that are no longer commercially viable—a particular issue in rural areas.

A key feature of the UK approach is that the organization running the service is acting for social purposes, and not for profit. Conversely, in the USA private transportation companies are often used to deliver services, which can mean high(er) costs per passenger, particularly for demand responsive door-to-door operations. Typically, US paratransit has utilized accessible minibuses (to facilitate wheelchair access), although increasingly providers are now looking to reduce costs by using taxi or rideshare options as an alternative.

Important goals for operators are to help people to stay independent, participate in their communities and be able to access vital services (such as health). As an example, Ealing Community Transport (ECT), a larger UK provider based in London have the following objectives for their services (ECT, 2015):

- To provide accessible transport services for people with disabilities who find it difficult or impossible to use conventional passenger transport.
- To address social deprivation with transport for individuals and groups who may be characterized as socially deprived, for example for people with low income without cars who would otherwise be excluded from the skills development or jobs market.
- To counter geographical isolation through services for individuals and groups who are poorly served by conventional passenger transport networks. For example in rural areas, remote parts of urban estates or areas without transport in the evening or weekends.
- To assist community cohesion through transport for community and voluntary groups enabling them to provide services and to respond to the needs of their community.

The needs identified in these objectives emerge across a wide range of groups within society, who will each have individual requirements to access a wide range of services and destinations. Those countries with limited levels of funding for community-style services may at times have to focus on more specific groups of people or particular journey purposes—that is, access to healthcare in some UK services. This has meant the frail and disabled in Australia for example (Mulley and Nelson, 2012), whilst US paratransit has concentrated on those with problems accessing public transport, who are unable to navigate or access the public bus system, or have a temporary need for mobility following injury or a limited duration disability. Services in Japan also focus on the elderly, and the less-able, but importantly, in rural areas they may also provide transport for the public, to make up for shortfalls in scheduled public transport (Wright et al., 2014).

Funding of Services

UK community transport relies on volunteers to perform driving and sometimes administration roles, but looks to central and particularly local government, or other organizations for financial resources to purchase vehicles and to operate services. The latter has included Parish councils (neighborhood-based administrations with limited powers and resources), and doctors' surgeries who see benefits for their patients. Local communities may fundraise to support services, and commercial organizations may sponsor or fund new vehicles as part of their corporate social responsibility (CSR) activities. Passengers will also typically pay a small distance-based fee for the journeys they take and an annual membership or registration fee in order to access the service. In some instances, passengers will be able to use concessionary passes that offer free travel on public transport to cover their travel costs.

State or government funding is commonplace in services across the world, but may take different routes to service operators. Prior to 2018, the Australian government provided direct funding to operators for community transport services. This has now changed to a system whereby government monies fund individuals' who source and pay for transport from a personal budget (Mulley et al., 2018). Australian operators are not permitted to charge fares, but may accept donations from passengers (Mulley and Nelson, 2012). Services in Japan are reliant on local authority subsidy, and as noted above may subsidize routes by carrying other passengers in some instances. US paratransit services are funded by local public transportation agencies, and may charge fares (up to twice the equivalent public transit fare is allowable).

Larger UK operators may also operate commercial services to cross-subsidize community operations. This could be minibus services for education, or health and social care services for a local authority accessed through commercial tender and bid processes, or by operating traditional local bus services. In all instances, these operations are run with paid staff, as if they were a commercial company, with strict demarcation between the two parts of the organization. Operators may also offer their vehicles for private hire outside of operating hours, to again subsidize community services. The scale of larger operators does provide economies in respect of costs of administrative staff, depots and technology, but does risk breaking the link between the service and the community it serves.

Countries that make extensive use of volunteers (like the UK), can minimize the costs of running a service, but this then creates a problem of recruiting enough volunteers. Many volunteers are themselves older people, and more likely themselves to be constrained by health issues and their own ability to access and use transport. This seems to be less of an issue in some other countries, with Australian services seemingly able to draw from a wider age range (Denmark and Stevens, 2016). The use of commercial operators in the US helps to avoid this problem there.

Evidence of Benefits

Studies of community transport in the UK claim a range of benefits, which are likely to be replicated in services in other countries. For example, the access to mobility it provides is seen to be of significant importance in tackling isolation and promoting social inclusion (Canning et al., 2015). Interactions with the driver can become “*very significant and something to look forward to*” for those whose opportunities for social interaction are limited (Martikke and Jeffs, 2009). Several studies have looked to quantify these benefits monetarily in respect of loneliness and social isolation, with one proposing that community transport could save the UK National Health Service (NHS) between £0.4bn and £1.1bn a year in related costs (Deloitte and Charity, 2015).

Community transport can also play an important role in tackling poor accessibility by providing services when mainstream transport is not viable (Canning et al., 2015). Even if there are alternative public transport services available, reaching a stop may be problematic. It may be too far from a person’s home, there may not be a footway, and there may be no street lighting or benches to take a rest. Taxis could provide door-to-door services, but many see them as expensive, and the driver’s responsibility only comes to the kerbside. Community transport volunteers will often address what might be termed the last 10 yards/m, and collect and return people from and to their homes.

Public transport and particularly taxis may not be suitable for the disabled, unable to carry a wheelchair for example, whereas vehicles used by community transport are usually accessible, and drivers will have received training in supporting such passengers. This can facilitate independence, and counters transport-related social exclusion for (older) people with limited mobility, who may not be able to leave their house otherwise, and for those with disabilities who lack accessible transport. Some users would not be able to live independently without community transport and would need to relocate, or move into a care setting (DHC & TAS, 2011). Community transport also provides people in rural areas with access to key services (Canning et al., 2015), helping to maintain rural communities. There are economic benefits, allowing people to go shopping and access employment, and community transport can help build social capital in communities (ECT, 2015).

Where community transport services are delivered by voluntary organizations they may also provide further social benefits. Drivers and the organization they work for are operating the service for the benefit of the people and groups they are carrying, and therefore are likely to have a more positive attitude to their passengers. Through regular contact, they will also develop relationships with passengers that will directly address issues such as isolation and loneliness, in effect providing a form of befriending service for passengers.

Several health benefits are seen from the use of community transport. These include earlier detection and treatment of health conditions and thus significant savings for the UK health service (Age Scotland, 2013), as well as facilitating people to get out more, and be more active (Canning et al., 2015). Perhaps most importantly, the use of community transport for health-related journeys is linked to a reduction in the number of missed medical appointments (termed “Did Not Attend” or DNAs), which can be a significant cost to health services (Countryside Agency, 2014).

Because benefits are seen across a range of policy areas, comprehensive measurement can be difficult, and as a result several studies have looked to the ‘social return on investment’, (SROI) methodology (Cabinet Office, 2009), and this is becoming a more commonplace approach to quantify benefits of community services. Others have considered the costs of replacing community transport with commercial transport, suggesting the resultant costs would be an order of magnitude greater than current local authority spending in the areas studied (DHC and TAS, 2011). This in part because community transport delivers more than just mobility.

Challenges Facing Community Transport

Whilst the evidence above suggests significant benefits come from community transport-style services, there are also important challenges facing provision worldwide. Foremost amongst these are growing demand and the need for sufficient funding, problems exacerbated by a decade of austerity measures in many countries.

Funding

Community transport in the UK is primarily funded via the state, through local authorities at a town, city or county level. Reductions in central government funding to these bodies over recent years has led to cuts in local services such as community transport. In some locations, this has meant a rationalization of operators and their services. In recent years, small amounts of capital funding (to purchase vehicles) have come from central government, but crosssubsidy from commercial or commissioned services remains a key source of revenue funding. This can put community transport services into direct competition with commercial providers, who may perceive an unfair advantage in respect of costs for operators using volunteers in parts of their business. This has led to lengthy legal disputes in recent years in the UK and remains a contentious issue.

Funding approaches are also changing, with some countries exploring the use of a person centered funding (PCF) model. This gives “clients” a budget to purchase mobility along with other services they require. This could mean that current block budgets to providers become fragmented across different types of service (e.g., some people may choose to use taxis or rideshare services if they are comfortable with doing so, and perceive a comfort or time benefit). This might mean changes in business models for community operators and potentially puts clients at risk of not being able to access mobility (CTAUK, 2018). The PCF model has been implemented recently in Australia, but at present other countries (such as the UK) have not yet taken such an approach.

Meeting Demand

Current services may struggle to meet demand, which generally exceeds supply, so that some clients miss out and others may get limited access to services. Women currently outlive men, but older women are less likely to drive (often through lack of confidence) meaning growing numbers of older women are looking to community services for their mobility solutions. Ageing populations also mean more older people with chronic health conditions that may preclude other modes of travel, and necessitate additional health-related trips. More generally, medical appointments can necessitate long journeys (particularly in rural areas), and be time consuming, meaning some operators may need to ration such trips. In Australia a shift to older and frailer clients (aged 80-90) has implications for the nature of services that have to be provided (Battellino, 2009), perhaps making them less efficient as well.

There are also supply-side constraints facing operators, who must maximize the use of their vehicles by aggregating as much demand as possible. Thus, it may be that there are conflicting demands for transport to healthcare and shopping at the same time, potentially disappointing one or other set of users. Commonly minibuses are used as transport, but they may only be carrying a small number of passengers in a 16-seat vehicle, an inefficient use of resources. Prioritizing service delivery then becomes important, raising questions as to what purpose is more important (healthcare or shopping for example). This can create issues in respect of when will a journey actually be confirmed for a user? What if greater demand emerges later, will already confirmed routes then be re-arranged for example? There is also the issue of whether there is sufficient funding to pay for routing and booking systems for this growing demand.

Operational Issues

Community transport in all countries also has to deal with a range of operational issues. For example, there are inconsistencies in service quality and service provision across providers in the UK and elsewhere. In recognition of such issues, the UK government has provided some limited funding to try to improve standards and service levels. In Australia, the establishment of accreditation and operation standards were seen as a response to similar issues (Battellino, 2009), whilst greater co-ordination of vehicle use to avoid duplication and underutilization of transport resources has been proposed in the US (TCRP, 2004). Many of these issues are potentially addressed with new technological solutions (e.g., booking and routing systems), but these can be expensive, and beyond the reach of many smaller operators.

Looking to the Future

Several developing and emerging trends will affect the need for community transport services in the future, including population ageing and health trends, new technology and sector and government responses to the climate emergency.

Drivers of Growing Demand

A key user group for community transport are the elderly, and their numbers are growing globally (UN, 2017). The overall number and proportion of older people in populations will both grow in coming decades. With increasing life expectancy, this also means that there are greater numbers of those in their eighties and nineties, a group more likely to have physical and cognitive issues that impinge on their mobility. Notwithstanding growing car use by this age cohort, more people will be unable to use a car because of health constraints. These might be health issues related to the normal processes of aging (declining eyesight, personal mobility and cognitive decline) alongside the health outcomes of more widespread longer life (e.g., dementia). All of these factors suggest greater future demand for “public” transport, and specialist services for older people. This at the same time as countries like the UK have seen a decrease in the availability and supply of public transport over the last few decades as public funding has diminished, and commercial viability of services has become increasingly difficult.

New Transport Solutions (Technology and Operating Models)

One of the key changes in transport in recent years has been the arrival of new technology, in particular the Internet and smartphone. This has revolutionized the way in which many people access information about travel, and purchase their tickets. Those who will be older people in coming decades are experiencing and adopting these new technologies, and will likely expect them as services in their later life. These developments have also enabled the growth of rideshare (through Transport Network Companies - TNCs), offering almost instantaneous mobility, at seemingly affordable prices. Alongside near real-time delivery from other Internet-based companies, developments such as these are creating a new expectation in service levels.

Paratransit services in the USA have developed sophisticated technological solutions in response to the demands of providing services that meet the legislative guidelines. Similar solutions for scheduling, routing and booking are now in use in the US and Europe. The costs though can prove prohibitive for smaller organizations providing community transport. The use of digital technology is seen as a means to meet challenging demands in transport, helping with efficiency, lowering costs and optimizing operations. Greater use of technology is not without risk though, and there are concerns that reliance on technology risks losing human contact with passengers.

The concept of mobility as a service (MaaS), where multiple modes of transport might be combined by travelers and commoditized as a product may also be relevant. There is potentially a role within MaaS for community transport to become another element in wider public transport networks and services. This might offer opportunities to commercialize surplus space on community transport-e.g. selling spare seats to other travelers who would not qualify as users, or to buy “community miles” on public transport services. It would facilitate development of services nationally (or greater homogenization). MaaS could bring economies of scale and more effective booking systems. It might allow for greater use of vehicles, reduce empty seats, and make the best use of volunteer time. One risk might be the loss of the “local” nature and personal touch of services. The MaaS approach might provide a way for community operators to adjust to person centered funding, offering a wider package of mobility choices, and allowing clients to choose from them to meet their needs (Mulley et al., 2018)

Responding to the Climate Emergency

Globally there needs to be radical change in transport, including a move to shared mobility to help reduce the impact of the transport sector (around a quarter of all UK emissions are transport related - DBEIS, 2020). This may have positive benefits for community transport, as it already provides a model for how local transport can respond to mobility needs. It is possible to see how a more extensive network of shared, local demand responsive transport services would offer a more efficient transport solution in response to the need to reduce emissions significantly over the coming years.

Conclusions

Community transport services are highly valued by users, but operators can struggle to meet demand. This situation is likely to persist, and potentially worsen as trends in demographic change and the transport landscape increase future demand. Some new models of operation may be helpful, and there may be lessons to learn from other new entrants to the transport landscape in recent years-in particular their use of technology. There are also risks here though with concerns about digital literacy of older people, and digital divides between the younger and older old. Looking forward 30-40 years, it may be that the much-heralded arrival of driverless vehicles will offer a solution for those who need community transport. Until then there are patently benefits to be gained from adequately funding the existing community model to meet the growing demand for services.

References

- Age Scotland, 2013. Driving Change: The Case for Investing in Community Transport. [Online] <https://www.ageuk.org.uk/> [Accessed 2/3/2020].
- Battellino, H., 2009. Transport for the transport disadvantaged: A review of service delivery models in New South Wales. *Transport Policy* 16 (3), 123–129.
- Cabinet Office: Office of the Third Sector, 2009. SROI guide. A guide to Social Return on Investment. Cabinet Office, London.
- Canning, S., Thomas, R., Wright, D.S., 2015. Research into the social and economic benefits of community transport in Scotland. Transport Scotland, Edinburgh, Scotland.
- Countryside Agency, 2004. The Benefits of Providing Transport to Health-Care in Rural Areas, Final Report to the Countryside Agency. CAG Consultants and the TAS Partnership Ltd.
- Community Transport Association UK (CTAUK), 2018. Door to Door Call for Evidence CTA Response [Online] <https://ctauk.org/> [Accessed 2/3/2020].
- Community Transport Association UK (CTAUK), 2020. What is community transport? [Online] <https://ctauk.org/> [Accessed 2/3/2020].
- Deloitte and ECT Charity, 2015. Tackling Loneliness and Isolation through Community Transport. Ealing Community Transport (ECT) Charity. Greenford, Middlesex, UK.
- Denmark, D., Stevens, N., 2016. Community transport in Australia. In: Mulley, C., Nelson, J.D. (Eds.), *Paratransit: Shaping the Flexible Transport Future*, Emerald, pp. 263–287.
- Department for Business, Energy & Industrial Strategy (DBEIS), 2020. 2018 UK greenhouse gas emissions: final figures - statistical summary. [Online] <https://data.gov.uk/> [Accessed 2/3/2020].
- DiT, 2018. The use of section 19 and section 22 permits in providing road passenger transport in Great Britain: aligning domestic legislation with EU Regulation 1071/2009. Impact assessment (IA) Department for Transport. London.
- DHC and TAS, 2011. Value of Community Transport Economic Analysis. DHC in conjunction with TAS. Edinburgh, Scotland.
- ECT Charity, 2015. Why Community Transport Matters. Ealing Community Transport (ECT) Charity. Greenford, Middlesex, UK.
- House of Commons Library, 2015. Community Transport. Research Briefings. House of Commons Library, London. UK. [Online] <https://researchbriefings.parliament.uk/> [Accessed 2/3/2020].

- Martikie, S., Jeffs, M., 2009 Going the Extra Mile: Community Transport Services and their Impact on the Health of their Users, Transport Resource Unit, Greater Manchester Centre for Voluntary Organisation. In Canning, S., Thomas, R., & Wright, D.S.; 2015. Research into the social and economic benefits of community transport in Scotland. Transport Scotland, Edinburgh, Scotland.
- Mulley, C., Nelson, J.D., 2012. Recent developments in community transport provision: comparative experience from Britain and Australia. *Procedia - Social Behav. Sci.* 48, 1815–1825.
- Mulley, C., Nelson, J.D., Wright, S., 2018. Community transport meets mobility as a service: on the road to a new a flexible future. *Res. Transport. Econ.* 69, 583–591.
- Transit Cooperative Research Program (TCRP), 2004. Strategies to increase coordination of transport services for the transport disadvantaged. TCRP Report 105, Transport Research Board, Washington, DC.
- UN Department of Economic and Social Affairs (Population Division). 2017. *World Population Ageing*. United Nations. New York.
- Wright, S., Emele, D., Fukumoto, M., Velaga, N.R., Nelson, J.D., 2014. The design, management and operation of flexible transport systems: comparison of experience between UK, Japan and India. *Res. Transport. Econ.* 48., 330–338.

Further Reading

- DCMS (Department of Culture, Media and Sport), 2017. *UK Digital Strategy 2017*, GOV.UK, London.
- DfT. 2011. *Community Transport. Guidance for Local Authorities*.
- Nelson, J.D., Wright, S., Thomas, R., Canning, S., 2017. The social and economic benefits of community transport in Scotland. *Case Stud. Transport Policy* 5 (2), 286–298.

Fundamental Emerging Concepts and Trends for Environmental Friendly Urban Goods Distribution Systems

Sandra Melo, CEiiA – Center of Engineering and Development, Porto, Portugal

© 2021 Elsevier Ltd. All rights reserved.

Glossary	320
Introduction	320
Urban Goods Distribution and Transport Planning	321
Sustainable Mobility of Goods	322
Emerging Urban Goods Distribution Trends	322
Conclusions	323
Biography	323
References	323

Glossary

Autonomous delivery vehicles are electrically powered motorized vehicles that can deliver items or packages to customers without the intervention of a delivery person.

Crowdshipping involves persons who are already traveling carrying parcels, using their spare carrying capacity in vehicles to carry parcels for other people.

Logistics is the function which ensures the right materials, right quantity, right quality, right place, right time, right method, right customer satisfaction and right cost.

Parcel lockers are locker banks that allow receivers to pick up goods, as well as carriers to exchange goods between modes.

Supply chain is the mechanism which guarantees that products or services go from an origin point to a destination, passing through various stages, including transformations, storage, and transportation.

Urban goods distribution is the process which ensures that the resources needed in urban areas are positioned in the right places, at the right times, in the quantity required and at a certain cost, whilst not neglecting the interests of the main stakeholders involved.

Introduction

The movement of goods generates conflicts with other functionalities within the urban structure. In recent years, this conflicting interaction has intensified its impacts on urban environment, thanks to the increase of flows to be delivered to the city center through growing diversity and complexity of delivery channels and processes. The intensity of said impacts lead to pressure placed by society on freight and logistics-related activities to perform the operations based on a “negotiated balance” with environmental protection, economic growth and social equity. Such pressure has sought to influence the industry perspective, where the goals are related to cost imperatives and service objectives such as the maximization of delivery frequency, avoidance of split consignments, maximization of vehicle load, reduction of the mileage traveled and reduction of the overtime payments to drivers. The integration of environmental protection and social equity issues into freight and logistics-related activities has been publicly scrutinized by society, as far as the minimization of disturbance, congestion and pollution that are frequently associated with freight movements in urban areas are concerned. The importance of this scrutiny is such that it has contributed to the definition of determinant drivers for the development of research on emergent logistics processes and services.

In such a context, it is noticeable that research carried out on the topic of logistics and transport planning has strongly reflected concern with the impact on sustainable urban development. Studies embracing an environmental perspective have focused on freight solutions that might contribute to an improvement of the urban environment quality: that is lower environmental impacts, less noise, and less congestion. Research projects that have adopted an economic perspective have tried to find solutions to optimizing the transport of goods into the city, while minimizing the number of trips, maximizing companies’ profits and avoiding congestion periods. Similarly, research work focusing on societal equity has been dedicated to maximizing the efficiency of all systems, mitigating the negative impacts associated with urban goods distribution activity and thus, promoting mobility and sustainability within the system. When seen together, these different perspectives and respective goals reflect the aim of designing environmentally friendly and efficient distribution systems to supply the urban centers, while endeavoring to integrate the main (and sometimes conflicting) perceptions of society and companies.

The design of systems that are simultaneously efficient and environmentally friendly is anything but straightforward. It requires balancing a complex system of variables and stakeholders in a struggle for the use of scarce territory with a significant demand in terms of correlated functionalities. With the emergence of Industry 4.0 and its side effects, the current unresponsive prospects toward those systems can be significantly reversed. Further comprehensive reviews on the Industry 4.0 and overviews of its indissociable and challenging concepts of Internet of things, cyber-physical systems, Big Data, cloud computing, mobile-based systems, social media-based systems can be found in [Maslarić et al. \(2016\)](#); [Hofmann and Rüsch \(2017\)](#); [Trappey et al. \(2017\)](#); [Barreto et al. \(2017\)](#). For the purpose of the scope of this article, the Industry 4.0 paradigm creates value from information that is extracted and refined from data, which enables mass customization, the possibility of free flow of data within processes and services and opportunities for incorporating logistics and transport planning solutions within smart city technological initiatives.

In this article, fundamental concepts arising from the most recent societal trends and industry changes will be looked at. This article, makes explicit the interpretation and implications of the visions assumed by the concepts. The concepts are by no means unambiguous; there is not one uniform approach and no single general application of the concepts and it is doubtful whether such a thing would–could or should–ever exist. The definition of the concepts is highly dependent on the specific context and can serve different users with different priorities and concerns. These concepts constitute determinant guidelines for keeping abreast of the most likely developments in the short term.

Accordingly, this article explores the opportunities for the introduction of some emerging concepts in freight transport and urban goods distribution-related activities. This research is exploratory, and the main contributions inform and provide insights for developing policies, plans, or initiatives to create efficient and environmentally friendly distribution systems.

The following sections present the interpretation the concepts assume and correlate them with the emerging trends in urban freight transport and goods distribution.

Urban Goods Distribution and Transport Planning

The concept of “urban goods distribution” is closely correlated with the concepts of “supply chain” and “logistics”, both in terms of content and formal context.

Supply chain encompasses efforts involved in producing and delivering a final product or service, from the supplier’s supplier to the customer’s customer. There are few companies that can produce end products for end-customers from raw materials on their own, without the assistance of other organizations. The company that produces the raw material is often not the same company that sells the end products to the end customer. In order to provide end products to the end customers, a network of actors is involved in activities (such as purchasing, transforming, and distribution) to produce products and/or services. The series of companies (actors) and respective processes that interact in this producing and delivering process through the flow of products, the flow of information and the flow of money, is named supply chain.

With the growth of the Internet of Things, on the basis of predictive management tools, one can envisage in the short term an increasingly important role for the information flows as the main supportive driver of supply chain processes and services. The value of the information flow will be reflected in terms of performance, scalability, robustness, and flexibility in freight and logistics processes and services.

Supply chain is the mechanism which guarantees that products or services go from an origin point to a destination, passing through various stages, which may include transformation, storage, and transportation.

Logistics is a transversal function that exists in all structures, sometimes in more obvious ways; it works as the integrated system that brings the necessary processes, activities and logistic resources to materialize the supply chain into relation with each other. This means that logistics is what ensures that products pass all the stages in an effective way and arrive in accordance with the required conditions. The Council of Logistics Management (2020) defines the concept as “the process of the supply chain of planning, implementing and controlling the efficient, cost-effective flow and storage of raw materials, in-process inventory, finished goods and all the related information from the origin point until the consumption point, considering the service requirements of customers”. It is a less visible function that ensures the good functioning of the supply chain. The article defines logistics as the function that contributes to the creation of time, place and even form through the process management that allow companies to get the right goods to the right place at the right time, in the right conditions and at the right cost.

Logistics is the function which ensures the right materials, right quantity, right quality, right place, right time, right method, right customer satisfaction, and right cost.

When logistics activities are carried out in urban areas, they reveal unique characteristics that make them different from general logistics activities. For this reason, freight and logistics-related activities in urban areas are usually labeled specifically as “city logistics”, “last mile,” or “urban goods distribution”. Everything that is consumed in urban areas or the materials used in industrial processes brought by the urban goods distribution system. The urban goods distribution system differs from long haul and other freight transportation segments ([Jaller and Pahwa, 2020](#)). Urban goods distribution refers to the last step taken to move and store a product from the supplier stage to a customer stage in the supply chain; accordingly, it deals with much more complex network environments. It differs from general logistics to the extent that urban goods distribution is an activity that optimizes the logistics and transport activities by private companies while taking traffic congestion, energy consumption and environmental damage within the framework of a market economy into consideration. Most definitions of the concept imply that the distribution activity is closely related to and even dependent on, transports and highlights the need to balance private companies’ aims and societal concerns. It

concentrates mainly on goods transport but, as supported by [Muñuzuri et al. \(2005\)](#), the concept can also include service vehicles (inspections, installations, technical service, and emergencies) and other commercial uses (sales representatives, company cars).

The electrification, automation, and shared mobility transformations constitute promising alternatives for improving the efficiency of last mile operations, particularly in terms of reduction of costs, time, and carbon emissions. Consequently, the number of initiatives that incorporate automation and electrification in the last mile in recent years has significantly grown, some of them with a considerable level of market readiness, while others still require more research. The advent of these technologies is imminent, but there are still barriers that do not allow for their wide diffusion, despite their potential benefits.

Urban goods distribution moves goods to the final customer, using a transport system and giving rise to social, environmental, and economic/financial impacts. It is defined as the process which ensures that the resources needed in urban areas are positioned in the right places, at the right times, in the quantity and quality required and at a certain cost, whilst not neglecting the interests of the main stakeholders involved.

Sustainable Mobility of Goods

People use a transport system to satisfy their needs and mobility is a derived need which surfaces when basic human needs and related activities cannot be satisfied in one place. A transport system affords the individual the possibility of traveling, and thus from a transport planning point of view, mobility measures how well the transport system fulfils its purpose, providing the individual with the possibility to participate in spatially based activities. Thus, mobility takes place to fulfill specific plans and projects, through a physical and/or virtual movement, and is influenced by several elements, such as lifestyles, spatial characteristics, and accessibility. All ingredients are generally given only if there is adequate means of moving people, goods, and ideas. Thus, greater mobility means increased quality of life for the individual (greater freedom to choose activities and greater amount of time to dedicate to them). In this sense, mobility plays a key role in development and depends essentially on the means enabling one to move. This definition, common to most publications on mobility, is generally applied to passengers. Indeed, little has been written, and is known, about mobility for freight transport compared to passenger transport; thus, much fewer studies focus on sustainable freight transport mobility. This lack is due to the restricted availability of data concerning freight transport but is somewhat attenuated by the fact that mobility concepts and indicators used for freight transport can be relatively similar to the ones applied for passenger transport.

Adapting the definitions presented above to the particular topic of logistics and urban goods distribution, “mobility” is the ease of movement, dependent on an (efficient) transport system and a diversity of (sustainable) options to get to a final destination, where the consumption needs are met at moderate costs to transporters and to society and as close as possible to the planned times (reliability).

Emerging Urban Goods Distribution Trends

With the emerging requirements of Industry 4.0, of which the Internet of Things is just one of its offshoots, it is expected that fleet management operations will go through substantial changes in the short term ([Ahani et al., 2016](#)). **Autonomous delivery vehicles** are electrically powered motorized vehicles that can deliver items or packages to customers without the intervention of a delivery person. Autonomous trucks, autonomous robots, or even drones have been tested for last mile delivery of goods in urban areas. Currently, their market is growing rapidly and is very dynamic ([Figliozi and Jennings, 2020](#)). It is expected that due to the emergence of flexible manufacturing systems, the rising demand for customized autonomous delivery vehicles and the adoption of industrial automation by SMEs, the landscape for autonomous delivery vehicles will evolve fast.

Current systems are well known and widely implemented in manufacturing, medicine, and logistics. In these systems, the fleet is organized in a centralized way. Tasks like motion planning and allocation of tasks are done by a central entity. However, driven by future requirements such as flexibility, robustness, and scalability, the current trend is toward the distribution of the total intelligence of a system to its components: each device gets a part of the total intelligence, so as to be able to operate independently, in an endeavor to achieve the same global goals ([De Ryck et al., 2020](#); [Killeen et al., 2019](#)).

The benefits of autonomous delivery vehicles can include cheaper delivery costs for businesses, faster service to customers, energy conservation and sustainability, safety for delivery personnel, and accuracy in delivering the right package to the right customer. Despite the fact that there are numerous studies on the mechanical, electrical, and computational design of robots ([Figliozi and Jennings, 2020](#)), academic publications studying the impacts and implications of autonomous delivery vehicles for last mile logistics are few and far between. Examples of such research work is [Vleeshouwer et al. \(2017\)](#), who studied a small robot delivery service, showing that costs can be reduced significantly, but that the robot capacity is low and that robots are not economically feasible. [Jennings and Figliozi \(2019\)](#) also analyzed the regulation and characteristics of sidewalk and road autonomous delivery vehicles in the USA, with the results showing that footpath autonomous delivery vehicles can provide substantial cost, time, and vehicle travel savings.

From a literature review and reports on the use of these solutions in experiments and initiatives, one can agree that autonomous delivery vehicles can be feasible if companies scale up or cooperate to increase robot utilization. When compared with a conventional delivery van with a human driver, autonomous delivery robots are also more economical when delivery routes are relatively short, and their limited mileage range is not a restriction. Given the fast pace of technological development, it is expected that autonomous delivery vehicles will experience significant improvement in terms of their development and implementation in the short and medium term.

While companies are adapting their fleet management activities to the emergence of autonomous delivery vehicles, cities are providing solutions such as parcel lockers. **Parcel lockers** are locker banks that allow receivers to pick up goods, as well as carriers to exchange goods between modes. Parcel lockers are located near residences, offices, public transport terminals, and shopping centers. Some cities, such as Lisbon (Portugal), also locate them in underground parking areas with surveillance cameras that ensure easy loading and unloading operations as well as security conditions. Trucks and vans can be used to carry parcels to locker banks, and these can also then be picked up by couriers for delivery to the final customers within the respective area. These infrastructures can play a significant role in minimizing home delivery trips, as they provide an effective and reliable option for online shoppers.

Parcel lockers are expected to increase the efficiency of urban goods distribution activities and, simultaneously, to contribute to achieving a more environmentally friendly process. Whether operated by single carriers or multiple ones, reductions in financial costs (vehicle operations), emissions and congestion are expected to be achieved in the short term.

Lastly, to introduce a nonphysical trend, the **crowdshipping** phenomenon is also worthy of mention. Crowdshipping involves persons who are already traveling carrying parcels, using their spare carrying capacity in vehicles to carry parcels for other people. Crowdshipping has a significant potential for reducing the number of freight vehicles operating in urban areas and reducing the operating costs for carriers.

This mechanism, which still has security, and liability-related issues that need solving, can be supported by the use of allocation methods or apps that match travelers' routes with delivery requests. There is a strong complementary potential when this system is used together with parcel lockers to achieve the design of more effective and environmentally friendly systems.

Conclusions

As with other initiatives and solutions that rely on combining sharing urban infrastructures with the intrinsic profile and needs of personal and commercial drivers, as the one provided in Melo et al. (2019), the emergent concepts mentioned in this article can contribute to achieving a negotiated balance between industry and research. These concepts by no means represent the only way of achieving efficient and environmental friendly urban goods distribution systems. The selection of concepts was based on the provision of a general overview of the subject matter plus the presentation of emerging trends and their particular potential dissemination in the short to medium term, and the respective implications. Further research and context must be explored, for instance, taking unusual events like pandemics, security rules, and the provision of unmanned health care into consideration. Parcel lockers and autonomous delivery vehicles, as well as voluntary crowdshipping, are initiatives that can play a relevant role when those events take place, but additional trends should also be explored in detail.

Biography

Sandra Melo (Ph. D.) currently works at CEiiA (Center of Engineering and Development, Portugal). She has a consolidated experience in the assessment of mobility solutions, with a major focus on investigating the performance of last mile "best practices", taking both public and private stakeholders' objectives into consideration. She has coordinated and developed R&D projects on the evaluation of the impact of new modes, services and mobility systems, integrating KPIs for traffic, energy, environment, operational, and economic targets. She also has an extensive experience as a consultant on transport policy and planning, a technical coordinator of national and international projects, both in academia and the industry, and a trainer and lecturer on sustainable freight transport, transport policy evaluation, and transport behavior. She is the author of scientific papers on the topic of the sustainable mobility of passengers and freight and has been awarded four times competitive nominal prizes in recognition of her scientific contribution to academia. She is also co-chair of Cluster 3: Logistics and Freight of the NECTAR scientific network and chair of the Portuguese Transport Studies Group (GET).

References

- Ahani, P., Arantes, A., Melo, S., 2016. A portfolio approach for optimal fleet replacement toward sustainable urban freight transportation. *Transp. Res. D Transp. Environ.* 48, 357–368.
- Barreto, L., Amaral, A., Pereira, T., 2017. Industry 4.0 implications in logistics: an overview. *Procedia Manuf.* 13, 1245–1252.
- De Ryck, M., Versteyhe, M., Debrouwere, F., 2020. Automated guided vehicle systems, state-of-the-art control algorithms and techniques. *J. Manuf. Syst.* 54, 152–173.
- Figliozzi, M., Jennings, D., 2020. Autonomous delivery robots and their potential impacts on urban freight energy consumption and emissions. *Transp. Res. Procedia* 46, 21–28.
- Hofmann, E., Rüsch, M., 2017. Industry 4.0 and the current status as well as future prospects on logistics. *Comput. Ind.* 89, 23–34.
- Jaller, M., Pahwa, A., 2020. Evaluating the environmental impacts of online shopping: a behavioral and transportation approach. *Transp. Res. D Transp. Environ.* 80, 102223.
- Jennings, D., Figliozzi, M., 2019. Study of sidewalk autonomous delivery robots and their potential impacts on freight efficiency and travel. *Transp. Res. Rec.* 2673 (6), 317–326.
- Maslarić, M., Nikolić, S., Mirčetić, D., 2016. Logistics response to the industry 4.0: the physical Internet. *Open Eng.* 1.
- Melo, S., Macedo, J., Baptista, P., 2019. Capacity-sharing in logistics solutions: a new pathway towards sustainability. *Transp. Policy* 73, 143–151.
- Muñuzuri, J., Larrañeta, J., Onieva, L., Cortés, P., 2005. Solutions applicable by local administrations for urban logistics improvement. *Cities* 22 (1), 15–28.
- Killeen, P., Ding, B., Kiringa, I., Yeap, T., 2019. IoT-based predictive maintenance for fleet management. *Procedia Comput. Sci.* 151, 607–613.
- Trappey, A.J., Trappey, C.V., Govindarajan, U.H., Chuang, A.C., Sun, J.J., 2017. A review of essential standards and patent landscapes for the Internet of Things: A key enabler for Industry 4.0. *Adv. Eng. Inform.* 33, 208–229.
- Vleeshouwer, T., Duin, R.V., Verbraeck, A., 2017. Implementatie van autonome bezorgrobots voor een kleinschalige thuisbezorgdienst. *Vervoerslogistieke werkdagen 2017 Vervoerslogistieke werkdagen*.

Plateau Car

David Metz, Centre for Transport Studies, University College London, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	324
Evidence for Plateau Car	324
Demand Saturation	324
Demographic Change	327
'Peak Car' in Big Cities	327
Impact of Technology	328
Conclusion	329
Biography	329
See Also	329
References	329
Further Reading	330

Introduction

There is good evidence that the average distance traveled per person by car in many developed economies ceased to grow towards the end of the last century. This phenomenon has been termed "Peak Car", by analogy with "peak oil", which refers to the expected peaking and decline in output of this finite resource (Goodwin and Van Dender, 2013). However, for car use, the evidence points to a cessation of growth as the prime effect, with possible long-term decline not yet generally apparent. Accordingly, the term "Plateau Car" has been proposed to designate the phenomenon (Metz, 2013a).

Evidence for Plateau Car

National travel surveys commissioned by governments are an important source of data on travel behavior. Time series data for three key parameters from the English National Travel Survey are shown in Fig. 1 (NTS, 2018; Table 0101). The average distance traveled by all modes (other than international travel by air) increased from 4500 miles per person per year in the early 1970s to reach 7000 miles around 2000, falling thereafter and stabilizing at about 6500 miles in recent years. Three quarters of this distance is currently traveled by car, driver and passenger together; car use per person has declined somewhat since 2002 (Fig. 2).

The US National Household Travel Survey found growth in average daily travel by all modes increasing from 25.1 person miles in 1983 to reach 40.2 miles in 2001, growth then ceasing (39.0 miles in 2017) (NHTS, 2017; Table 14). Average distance per person traveled by private vehicle was 30.85 miles in 1990 and 30.45 miles in 2017 (NHTS, 2017; Table 12).

Data for distance traveled by car per capita in 11 developed economies are shown in Fig. 3, estimated from official statistical sources for total distance traveled by car and for total population. Car use grew until around the turn of the century, when it stabilized or fell in all but one case (Poland). Millard-Ball and Schipper (2011) found evidence for a leveling out or saturation of total passenger travel since the early years of the 21st century for eight developed countries.

Demand Saturation

Cessation of growth of demand for a product or service is a common feature of all markets. A new product that offers benefits to users is taken up, initially by early adopters with the more cautious following later. High levels of ownership, as for instance found with many household appliances, characterize a mature market in which growth has substantially ceased and demand is said to have saturated. It is plausible that the cessation of growth in surface travel and car use reflects demand saturation for daily travel for the purpose of gaining access to a wide choice of destinations offering services and opportunities.

The growth of average distance traveled in the last century shown in Fig. 1 is a consequence of an increase in speed, given the unchanged average travel time over the period, and is largely attributable to growth of car ownership. Conventional economic appraisal of transport investments that result in faster travel assumes that the main user benefit is the saving of travel time [10011, 10097]. However, the observed invariance of travel time implies that the main benefit is better access to desired destinations and the associated increase in opportunities and choices [10045]. Access is subject to diminishing returns as choice and opportunities increase. However, access also increases with the square of the speed of travel, being proportional to the area of a circle, the radius of

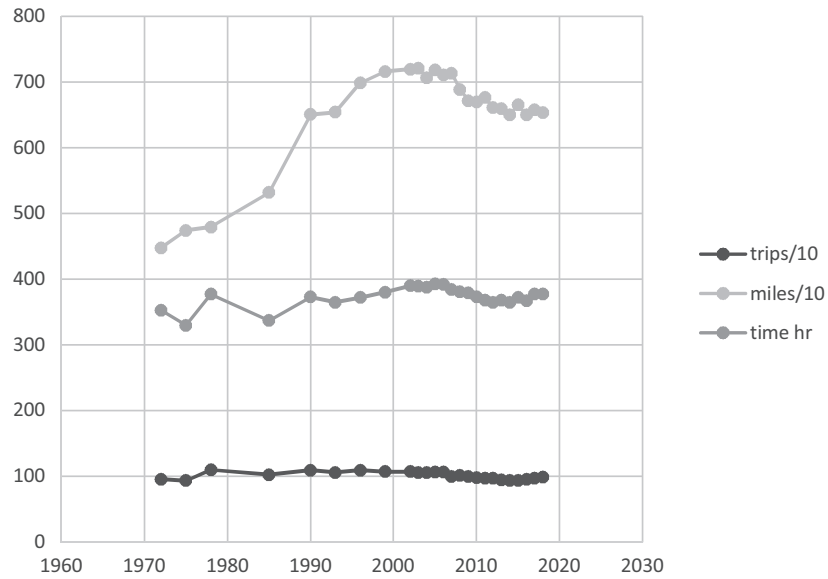


Figure 1 National Travel Survey. Source: *NTS, (2018) Table 0101*.

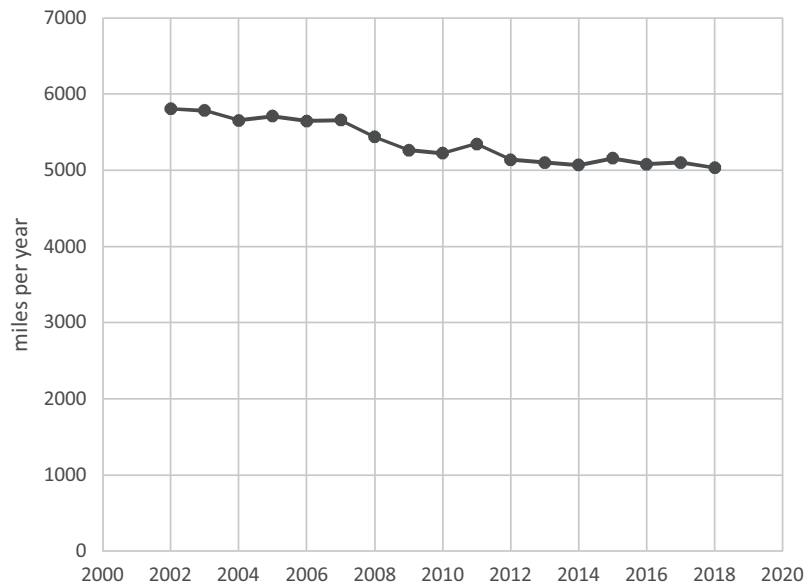


Figure 2 Distance traveled by car per person (driver and passenger). Source: *NTS, (2018) Table 0303*.

which is proportional to speed. This combination of diminishing returns while increasing with the square of speed yields a saturation function.

There is some evidence in support of the relevance of saturation in explaining cessation of growth of per capita car use. **Fig. 4** shows the proportion of households in England owning one or more cars, which increased steadily until the end of the last century when growth ceased, suggesting both saturation of demand for car ownership and the main cause of the cessation of average distance traveled by all modes of travel [10452].

UK data on access to key services by journey time indicates high proportions of potential users having access within reasonable travel times. For example, for access to family doctors (termed General Practitioners, GPs), 71% of users are within 15 min travel time by public transport/walking and 96% within 30 min, while 87% are within 15 min by bicycle and 98% by car (*JTS, 2017*; Table 0201). Similar high levels of access are found for other services, including employment, schools, food stores, and town centers.

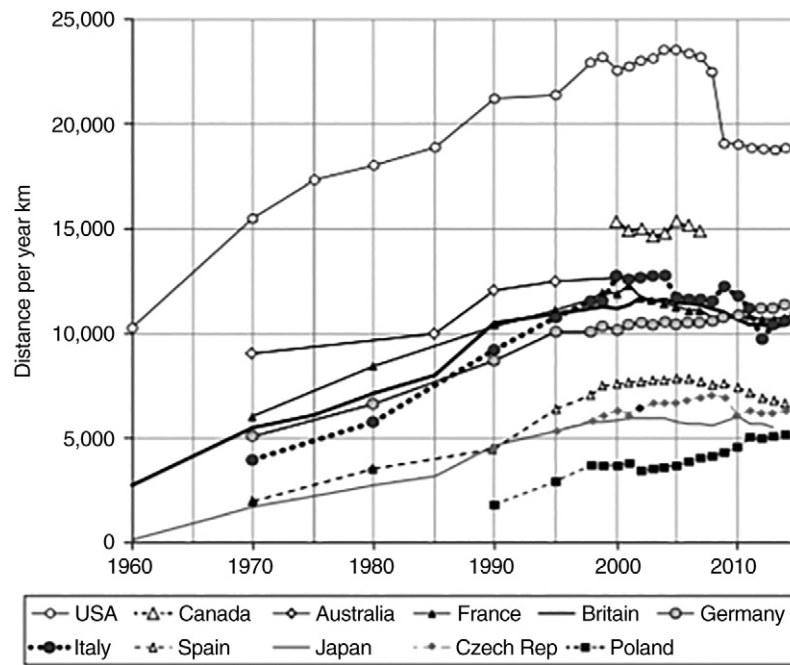


Figure 3 Distance traveled by car per person. Source: Dr Kit Mitchell, from UN ECE Transport Statistics for total distance traveled by car and official population statistics for total population.

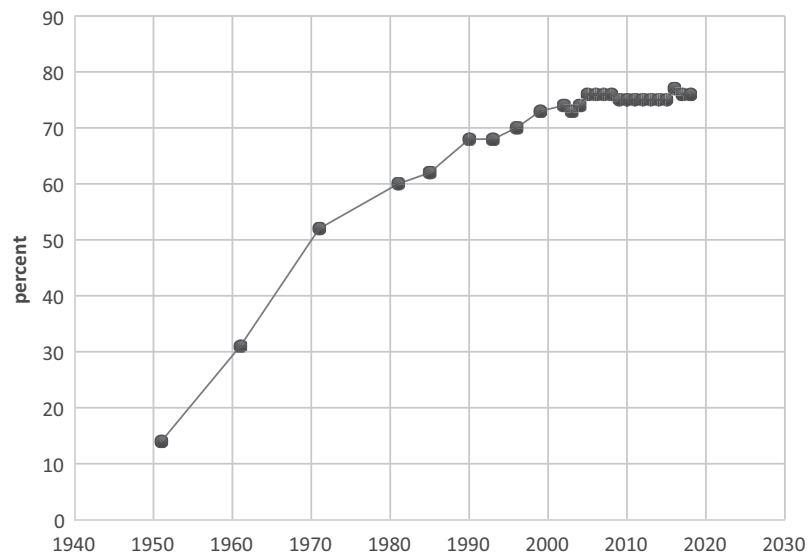


Figure 4 Proportion of UK households owning one or more cars. Source: NTS, (2018) Table 0205.

Journey time statistics can be used to infer levels of choice of key services. For instance, the populations of a majority of English localities have access on average to five or more GPs within a 30-min journey by public transport/walking, and almost all localities have such choice within 15 min by car (Metz, 2013b). A study of access to supermarkets in Britain found that 80% of the urban population had access to three or more large stores within 15 min by car, and 60% to four or more (Metz, 2010). Such high levels of access and choice are consistent with the proposition that demand saturation is contributing to the cessation of growth of per capita car use. In the case of supermarkets, this has come about through (1) the growth of car ownership and (2) the supermarket chains taking advantage of road construction that made land accessible for new large stores on the edge of urban areas, both trends now largely played out. Although comparable data for other developed economies are not available, it seems likely that similar saturation phenomena are occurring.

Demographic Change

There are a number of demographic changes that are contributing to the cessation of growth of per capita car use (Metz, 2016). Movement of populations from country to city is a long-term global trend, reinforced in recent years in developed economies by the shift from manufacturing to business services and the “knowledge economy” that prefers to locate in city centers. Contributing to urbanization is the process that economists term “agglomeration”, which offers particular advantage to the business services sector. Concentrations of firms in one geographic area benefit from learning, sharing, and matching. Firms acquire new knowledge by exchanging ideas and information, both formally and informally; they share inputs via common supply chains and infrastructure; and they benefit by matching jobs to workers from a deep pool of labor with relevant skills. Agglomeration leads to urban development and population growth at higher densities, despite the high land prices, rents and other costs. This growth has increased congestion on the urban road network and so made car use less attractive, but at the same time has improved the economic viability of public transport (Metz, 2019).

The growth of the economy in cities, as well as the expansion of city center universities, has attracted young people to move to vibrant cities to study, work, and live. This has contributed to a significant change in travel behavior amongst young people in developed economies. A recent comprehensive review of the research evidence and survey data noted a trend since the mid-1990s for successive cohorts of young people (ages 17–29) to own and use cars less than their predecessors (Chatterjee et al., 2018) [10429]. This contrasts with the baby boomers, born from 1946 to 1964, who led a rapid, prolonged and persistent growth in car ownership and use.

Factors contributing to this trend away from car use include the cost of car ownership (not least, high insurance charges for younger drivers), problems of parking in cities and on campuses, and the viability of alternatives such as bicycles, public transport, shared car use, and smart-phone apps to summon a taxi. However, according to Chatterjee et al. (2018), the main causes lie largely beyond the transport system and include increased participation in higher education, for which the car is not part of the lifestyle; the use of digital communications and social media; and more generally a delayed transition to what was traditionally seen as adulthood—commitment to a career, getting married, home ownership, starting a family, with stable employment a strong determinant of being a car driver.

An important question is how this shift away from car use by young people will affect the way they travel as they get older. Stokes (2013) has shown that those who start to drive later drive less when they do start; for instance, for those now in their thirties in Britain, if they learnt to drive when age 17, now drive 10,000 miles a year on average, while if they learnt at age 30, they are likely to drive around 6500 miles a year. This reduced mileage seems likely to reflect greater experience of alternative modes to the car gained before learning to drive, as well as living in places where such alternatives are viable, in particular for journeys to work. Chatterjee and colleagues concluded that while there are many uncertainties about the travel behavior of future cohorts of young people, as well as about how this may change as they get older, it is nevertheless hard to envisage realistic scenarios in which all these uncertainties combine to re-establish earlier levels of car use.

‘Peak Car’ in Big Cities

The discussion thus far has focused on per capita travel demand. Population growth is, of course, an important determinant of total travel demand. The impact of urban population growth on car use is well illustrated by the experience of London, for which exceptionally extensive travel statistics are available (TfL, 2019). After a period of population exodus, when inhabitants left poor-quality, overcrowded housing for a better life beyond the city boundaries, decline reversed around 1990 as the built environment improved and agglomeration economics increased productivity, generating higher incomes. The population of London increased quite rapidly, from 6.8 million in 1991 to 8.9 million in 2018.

Despite this population growth, road vehicle traffic in London as a whole, about 80% of which is car traffic, has not increased over the past 20 years—indeed has fallen by about 10% (Metz, 2015; TfL, 2019, Table 9). This is due largely to the limited capacity of the road network. Plans were proposed in the 1960s to build more roads to accommodate the expected growth in car use. An initial section of elevated motor road was constructed westwards from central London, but this was seen as damaging to the urban fabric and plans for similar new roads around central London were largely abandoned. Thus London has essentially retained its historic road network, which has constrained the growth of traffic. Indeed, there has been a reduction in the capacity of the road network for cars as the result of reallocation of road space to bus and cycle lanes and pedestrian uses.

Given that London’s population has been increasing but car traffic has fallen, it follows that the share of journeys by car must have declined, as is indeed observed. In 1993, when the relevant data series starts, 50% of all trips in London were by car (driver and passenger), but this declined to 35% in 2018 (TfL, 2019; Figure 2.3). Over the same period, the share of walking trips held steady at 25%, while cycling grew from a low base to reach about 3% of all journeys. The decline in car trips was compensated by the growth in use of public transport, which currently account for 37% of all journeys.

Fig. 5 shows an estimate of the share of journeys by car in London in the century 1950–2050, based on historic data, Transport for London’s recent data mentioned in the previous paragraph, and future extrapolation that takes account of projected population growth and assumes no addition to road capacity but continued investment in urban rail. From 1950 car ownership and use grew while at the same time the population was falling. Car use peaked at 50% of all journeys at around 1990, the time when the decline in population reversed. As the population subsequently grew, the historic road network acted as a capacity constraint, which meant that the share of trips by car had to fall, a trend projected to continue out to 2050, when some 27% of trips would be by car, on the stated

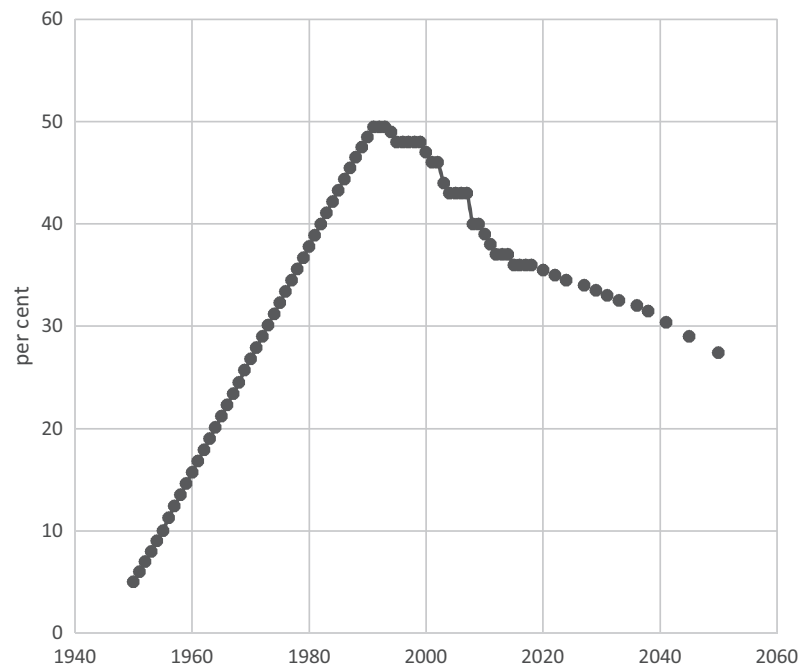


Figure 5 Share of journeys by car in London, 1950–2050. Source: author's estimates and *Transport for London* (Metz, 2015).

assumptions. The Mayor of London has more ambitious plans, aiming to reduce car use to only 20% by 2041, driven by concerns about health impacts (Mayor, 2018).

What is shown in Fig. 5 is the clearest illustration of the concept of “Peak Car”, a peak followed by decline in the *share* of trips by car. This has been termed “Peak Car in the Big City”, to distinguish it from peak or plateau car, referring to the distance traveled per person as discussed above (Metz, 2015).

While London provides the best data sources, there is evidence for a similar phenomenon in other large cities. A substantial study identified a peaking of car use in terms of car trips per day in Paris, Berlin, Vienna and Copenhagen, as well as London; on the other hand, data from US cities indicates a continuing very high level of car use (Jones, 2018). There is supporting evidence of a decline in car traffic in a number of UK cities, particularly in their centers, including Manchester and Birmingham (Metz, 2013a), as well as in the major Australian cities (Newman and Kenworthy, 2011).

A broad distinction may be made between cities with historic central areas where the street pattern limits car use and the population density make public transport economically viable; and more recent cities built at low density with the car in mind as the main means of mobility. This distinction is broadly between European and North American cities, although with central parts of the some of the latter having high density, and suburbs of most generally being of low density. The available evidence suggests a general phenomenon whereby successful cities with well-established centers attract people to work, study, visit and live; the total population and population density both increase; the authorities recognize that the road network cannot be expanded to accommodate more car use without damaging the urban environment; hence decisions are made to invest in public transport, particularly rail, which provides speedy and reliable travel compared with cars, buses and taxis on congested roads, as well as to improve facilities for walking and cycling. A common characteristic is the existence of a historic city center that people find attractive for work and leisure, where declining car use increases the attraction. In the absence of an appealing downtown district, population growth may lead to continued low-density development, with no mode shift away from car use.

In the 20th century, increasing prosperity was associated with increasing car ownership and use in developed economies. In the 21st century, increasing prosperity is associated with *decreasing* car use in big cities having attractive centers, growing populations and that can afford to invest in an extensive network of high-quality public transport as an alternative to the car on congested roads. A question for low income economies, where car ownership is still relatively low, is whether the peaking of car use in large, densely populated cities use can be avoided, transitioning instead to the modest mode share envisaged in the developed economy cities discussed above [10624].

Impact of Technology

Innovative technologies have been central to the historic development of transport. Until the coming of the railway, travel generally was at walking pace, with few able to take advantage of horse-drawn vehicles, which were not much faster on poor roads. Time has

always been a constraint on the distance that could be traveled. In the 19th century, the railway harnessed the energy of coal to achieve higher speeds and mass mobility that transformed economic and social geography. In the 20th century, the motor car using oil, the other fossil fuel, allowed speedy door-to-door travel, with further gains in access to desired destinations, opportunities and choices, within the time constraint of about an hour a day average travel time. The cessation of growth of average distance traveled by car in part reflects the inability to travel both faster and more safely on uncongested roads, and the difficulty of mitigating road traffic congestion.

There are new technologies, both deployed and entering the market, that will change how cars are propelled and used. Electric propulsion in place of the internal combustion engine will eliminate tailpipe emissions of carbon dioxide and noxious pollutants, but will not change the basic features of the car. Automated vehicles will reduce, or even eliminate, the role of the driver, but are unlikely to permit faster travel on existing road systems. Digital navigation devices are used widely, offering optimal routing through congested networks and estimates of journey time that mitigate the uncertainty associated with road traffic congestion. Digital platforms provide improved ways of locating vehicles for hire and sharing vehicles. All these innovations improve journey quality without having much impact on the speed of travel and are therefore unlike past transport innovations that offered step-changes in velocity (Metz, 2019).

Innovative technologies may affect car use by offering alternatives to travel from home. Online shopping is a likely contributory factor to the downward trend in shopping trips in England—15% decline in numbers of trips and 18% in distance, between 2002 and 2018 (NTS, 2018). Journeys to work in England have also been declining, from 7.1 per worker per week around 1990 to 5.7 in 2014, attributed to people working fewer days a week, more able to work from home, and more self-employment (DfT, 2016). The coronavirus pandemic of 2020 accentuated these trends, as well as demonstrating the scope for interactions made possible by personal travel to be replaced by interactions mediated virtually by digital technologies. On the other hand, digital technologies make possible a wider range of personal contacts, which may prompt a post-pandemic revival of travel to reinforce relationships through face-to-face engagement. The net long-term impact on travel behavior is unclear.

Conclusion

As the developed economies moved from the 20th to the 21st centuries, car use per capita generally ceased to increase, bringing to an end nearly two centuries of growth of personal travel that began with the development of the railways in the first part of the 19th century. The evidence suggests that the main contributory factors to this cessation of growth were saturation of demand for daily travel, reflecting high levels of access achieved; population growth in successful cities where space for car use is limited; and technological limitations that prevent faster travel, given constraints on travel time. Despite the deployment of new transport technologies, car use per capita seems unlikely to change much in the future.

Biography

David Metz is honorary professor in the Centre for Transport Studies, University College London, where his research focuses on how demographic, behavioral and technological factors influence travel demand. After an initial career in biomedical research, he took senior posts in a number of UK government departments, as both policy and scientific advisor, including 5 years as Chief Scientist at the Department of Transport. His recent research has been summarized in a book entitled 'Driving Change: Travel in the Twenty-First Century' (Agenda Publishing, 2019).

See Also

Passive Prevention Systems in Automobile Safety; Estimation of Value of Time; The Rule-of-a-Half and Interpreting the Consumer Surplus as Accessibility; Car Ownership and Car Use: A Psychological Perspective; Next Generation Travel: Young Adults' Travel Patterns; Transport Planning in the Global South

References

- Chatterjee, K., Goodwin, P., Schwanen, T., et al., 2018. Young People's Travel – What's Changed and Why? Review and Analysis. Report to Department for Transport. University of the West of England, Bristol.
- DfT, 2016. Commuting trends in England, 1988–2015. Department for Transport, London.
- Goodwin, P., Van Dender, K., 2013. 'Peak Car' – Themes and Issues. *Transport Rev.* 33 (3), 243–254. This editorial introduces articles in a special issue devoted to Peak Car.
- Jones, P., 2018. Urban Mobility: Preparing for the Future, Learning from the Past. CREATE Project Summary and Recommendations. <http://www.create-mobility.eu>.
- JTS, 2017. Journey Time Statistics. Department for Transport, London.
- Mayor, 2018. Mayor's Transport Strategy. Greater London Authority, London.
- Metz, D., 2010. Saturation of demand for daily travel. *Transport Rev.* 30 (5), 659–674.
- Metz, D., 2013a. Peak car and beyond: the fourth era of travel. *Transport Rev.* 33 (3), 255–270.

- Metz, D., 2013b. Mobility, access and choice: a new source of evidence. *J. Transport Land Use* 6 (2), 1–4.
- Metz, D., 2015. Peak car in the big city: reducing London's greenhouse gas emissions. *Case Stud. Transport Policy* 3 (4), 367–371.
- Metz, D., 2016. Changing demographics. In: Bliemer, M., Mulley, C., Moutou, C. (Eds.), *Handbook on Transport Urban Planning in the Developed World*, Edward Elgar, Cheltenham, UK, pp. 69–81.
- Metz, D., 2019. *Driving Change: Travel in the Twenty-First Century*. Agenda Publishing, Newcastle upon Tyne.
- Millard-Ball, A., Schipper, L., 2011. Are we reaching Peak Travel? Trends in passenger numbers in eight industrialised countries. *Transport Rev.* 31 (3), 357–378.
- Newman, P., Kenworthy, J., 2011. 'Peak Car Use': understanding the demise of automobile dependence. *World Transport Policy Practice* 17 (2), 31–39.
- NHTS, 2017. *Summary of Travel Trends: 2017 National Household Travel Survey*. Washington DC: US Department of Transportation, Federal Highway Administration.
- NTS, 2018. *National Travel Survey 2018*, Department for Transport, London.
- Stokes, G., 2013. The prospects for future levels of car access and use. *Transport Rev.* 33 (3), 360–375.
- TfL, 2019. *Travel in London: Report 12*. Transport for London, London.

Further Reading

- Metz, D., 2016. Changing demographics. In: Bliemer, M., Mulley, C., Moutou, C. (Eds.), *Handbook on Transport and Urban Planning in the Developed World*, Edward Elgar, Cheltenham, UK, pp. 69–81.
- Metz, D., 2019. *Driving Change: Travel in the Twenty-First Century*. Agenda Publishing, Newcastle upon Tyne.
- Newman, P., Kenworthy, J., 2015. *The End of Automobile Dependence*. Island Press, Washington DC.
- 'Peak Car' – Themes and Issues. *Transport Reviews*, 33(3), 243–254. A special issue devoted to Peak Car.

Emerging Trends in Transport Demand Modeling in the Transition Toward Shared Mobility and Autonomy

Patrizia Franco, Connected Places Catapult, Milton Keynes, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Glossary	331
Introduction	332
Literature Review	332
The Era of MaaS	332
Changing Travel Behaviors	333
Modeling to Assess the Impact of MaaS and NMS	333
Stakeholders Perspective	334
Local Authorities' Perspective	334
Transport Operators' Perspective	335
Data Required to Model MaaS and NMS	335
Demand Modeling—Key Changes	336
Conclusions	337
Acknowledgment	338
Biography	338
References	338

Glossary

NMS New Mobility Services linked to the sharing economy and the development of new ways to reach customers (ride-hailing, ride-sharing, car share, bike share, carpooling, micromobility, including e-scooters) using new modes of transport.

MaaS Integration of various forms of transport services into a seamless integrated mobility service accessible on demand and via a single portal, with user-centricity and customers at its heart.

TAG Transport Appraisal Guidelines developed by the UK Department for Transport. TAG provides information on the role of transport modeling and appraisal, and how the transport appraisal process supports the development of investment decisions to support a business case.

DRT Demand Responsive Transport flexible mode of transportation that adapts to the demands of its user groups, often accessible via a digital platform or mobile app.

ABM Agent-based modeling uses a synthetic population of agents to understand complexities in human behavior. Agents are entities with their own behavior, preferences and activities to fulfill. When it comes to model mobility, agents can be either the user of the service (static agent since their plan consist of a sequence of trips or activities) or vehicles (dynamic agents with no prior predefined activities). Activity based Modeling is a specific ABM where agents have activity plans to fulfill.

Synthetic Population Designed to resemble a real-world population with respect to sociodemographic characteristics and spatial information. In agent-based modeling for the transport sector, each agent in the population is associated also with daily activity plans.

Agents Social elements/persons in a model that can interact with each other and an environment, and pass information between each other. Agents might represent animals, individuals, households, organizations, or even entire countries.

Properties Agents have properties. A property might be a memory, a state, a characteristic, or attribute, such as age, gender, income, speed, or health. Properties are discrete, and can be binary (yes/no), or numerical (e.g., on a 1–100 scale).

Environment This is the virtual world in which agents act and interact. It can be 2D, or 3D. A neutral medium with no effect on agents whatsoever, or a prime determinant of their ability to act. It can be abstract and imagined, or a replication of real-world buildings, cities, or vast geographies (road networks extending to few blocks up to cities or regions).

Rules These govern what happens when agents interact (or come into contact) with each other, or their environment. They may also govern how learning and adaptation occur within an environment. These rules may be pre-programmed, or automatically inferred in the case of some smart ABM platforms.

Introduction

According to the United Nations, the population residing in cities is projected to reach 60% by 2030. Changes in travel behavior and car ownership are pivotal to enable a switch to low-carbon transport solutions. Within the transport sector, road transport is considered the biggest emitter (more than 70% of all GHG emissions in 2014). The delivery of seamless, personalized, shared and integrated mobility has the potential to enable the Mobility as a Service (MaaS) vision, however, there is a lack of understanding around the drivers allowing people to share and no best practice is available to assess the performance of the services and their impact/interaction with the public transport.

Autonomy, flexibility, and digitalization are changing the way we travel, however, business models established around on-demand, ride-sharing services have not proven to be successful in the long term. The process of trial and error established to attract demand randomly produced a negative effect on congestion, a decline in patronage on the public transport, with whom mobility services seek a spontaneous integration. The process to launch and maintain these services is missing some fundamental understanding on travel behaviors and customers' mobility needs, not currently captured in the market research carried out ahead of the launch. Perceived bias and habits should be considered in a comprehensive analysis that takes in consideration current travel patterns.

Local and transport authorities currently are unable to assess the impact that New Mobility Services (NMS), Demand Responsive Transport and Mobility as a Service (MaaS) might have on travel behaviors and any unintended consequences that might derive from the introduction of these on-demand mobility services.

Transport models currently in use to assess the demand for travel are unable to consider the full complexity of travelers' needs and the contribution shared mobility makes in a public transport system, both in urban and rural environments.

The increased complexity of travel demand models to fit modern lifestyle, and the demand for flexibility led to travel patterns being more difficult to predict, as well as a reduced availability of private transport for certain population groups with a subsequent increase in social exclusion. New methodologies are emerging to take into account the end-to-end users' journey, multimodality and the ability to represent shared mobility and autonomous flexible services. Moreover, the role of big data and new data sources, such as location-based data (i.e., mobile network data and global positioning systems data), integrated with data fusion processes and machine learning, have increased the spatial and temporal granularity of next generation demand models.

This article presents an evidence base of current market needs, current industry capabilities and defines the problem space in the implementation of NMS and the adoption of MaaS schemes.

Literature Review

In the transport modeling sector there are currently no robust tools and techniques, covered by guidance, to allow a commissioner of transport services, such as a local authority, to assess the investment case for introducing NMS or MaaS. New business models to support mobility services within the transport sector are flourishing around Europe and the rest of the world, however, identifying the most suitable locations for their introduction, and potential demand for the services, is more often a process of trial and error, as none of the key modeling software packages have the capability to fully predict demand for these services and assess the impacts of their introduction on the more complex and less flexible Public Transport ecosystem. New services are, therefore, often viewed as negatively disruptive, especially if not designed to complement existing public transport services or promote more sustainable active modes (walking and cycling), as well as being deployed in a potentially inefficient way.

NMS have great potential to provide greater flexibility for travelers and deliver better connectivity between modes, which have historically been designed and operated within respective silos. If they are designed around the needs of the customer and integrated appropriately with existing transport offerings, these services can provide new cost-effective, on-demand solutions and be used to manage demand on congested networks by incentivising alternatives, helping to maximize the capacity and resilience of existing infrastructure.

Due to the emerging nature of these NMS, transport operators and local authorities also have no corresponding guidance for the process new service providers must go through to demonstrate the likely impacts on network operations in order to receive licenses to operate. Current standards and guidelines for scheme assessment through modeling (such as those presented in the UK Department for Transport's TAG) are based around assessing and appraising investments in infrastructure and provide a limited opportunity for industry to evolve to consider investments in NMS as viable alternatives.

The Era of MaaS

Transport modeling has been traditionally facing significant challenges to realistically represent mobility. Many uncertainties, mainly but not only related to human behavior, have hindered the development of robust models, able to holistically capture the diverse dynamics of transport networks. In an emerging transportation future, where NMS and MaaS concepts will be widespread, users are likely to perceive traveling substantially differently. Although the term MaaS has been extensively used over the last few years, no formal definition has yet been agreed. This could be partly explained by the fact that MaaS is a relatively new concept which has not yet fully developed. Most definitions suggested by various entities revolve around principles such as integration, user-centricity and personalization. [Mulley \(2017\)](#) focuses on the disutility nature of traveling and states that "The concept of MaaS recognizes the way in which travel is typically a disutility for the traveler who is motivated only by consuming at the destination." Transport Systems Catapult definition highlights the range of available choices to the traveler, defining MaaS as a "new concept that

offers consumers access to a range of vehicle types and journey experiences" (Datson, 2016). Mars and Arup (Falconer et al., 2018) highlight the dual nature of MaaS distinguishing two inherent elements, namely the "physical transport provision" and the "medium for accessing the service". MaaS Alliance, a public-private partnership aiming to create a common foundation approach towards MaaS, defines it as "the integration of various forms of transport services into a single mobility service accessible on demand". Atkins et al. (2015) propose a broader but more comprehensive definition: MaaS is "The provision of transport as a flexible, personalized on-demand service that integrates all types of mobility opportunities and presents them to the user in a completely integrated manner to enable them to get from A to B as easily as possible".

Changing Travel Behaviors

According to recent studies, travel behavior is gradually getting more difficult to predict (Department for Transport, 2018; Ferreira et al., 2007; Holmberg et al., 2016). Various reasons have contributed to this, including the modern lifestyle, the increased demand for autonomy and flexibility, the diverse travel patterns, as well as, the reduced availability of private transport for certain population groups.

Another factor affecting travel behavior stems from the digitalization of the transport system, which allow travelers to make better decisions regarding their journeys. Due to the availability of accurate and timely information, travelers now can optimize their journeys and adjust their travel patterns accordingly, often leading to unexpected travel behavior.

In addition, the imminent technological advances (e.g., autonomous and connected vehicles and unmanned aerial vehicles) are very likely to radically change the rules of mobility. For instance, the flexibility and ease of use of autonomous vehicles could unlock new segments of travel demand which are currently suppressed, such as the elderly population and those who are mobility impaired (Harper et al., 2016). This increased demand could put current transport networks at, or over, their available capacity and therefore leave them unable to cope.

In light of the potential for travel behavior to change significantly, a thorough assessment of the effect that the widespread introduction of NMS would have on the transport system, requires advanced modeling techniques which are not readily available (Ben-Akiva et al., 2007; Jittrapirom et al., 2017). Moreover, Intelligent Transport Systems (ITS) have become an essential component of transport networks and many studies have examined its impact (Dia, 2002; Chorus et al., 2006), and its ongoing role which is generally neglected in most of the currently available transport models.

Traditional transport models (e.g., four-step modeling) face difficulties in realistically representing the complex dynamics of the emerging transport environment (Jain, 2016; Mladenovic and Trifunovic, 2014; Rodier et al., 2003). The methodologies applied in these models were developed under a different transport environment, but many of the underlying assumptions may not hold true in the era of NMS.

A four-stage model approach is unable to represent the shared element: the traditional division between highway assignment and public transport modeling does not allow the model to represent the interaction between private on-demand shared services and fixed scheduled public transport service (Franco et al., 2020a). The second main barrier in using strategic modeling tools at macroscopic level is represented by the aggregated trip-based approach adopted. A deep knowledge on how users travel over a day and what is their travel patterns is required. This is necessary in order to provide a valid alternative to private owned cars and implementing the integration of mobility services envisaged by MaaS. For example in Franco et al. (2018), the analysis of highly granular disaggregated Mobile Network Data (MND) demonstrated that a higher number of trips in an area is not providing a necessary condition that indicates an area with potential customers for a mobility service. Hence, the only viable solutions to model NMS would be the adoption of activity-based modeling approach which can derive either simple or complex travel patterns for users.

In many cases though, certain simplifications can enable NMS to be simulated by conventional methodologies (e.g., park-and-ride, regular taxis, fixed route transit). As a result, existing transport demand modeling software can be reconfigured in order to address NMS in a conventional way, but these approaches cannot always model NMS in the level of detail that is required for the operators and the Local Authorities seeking for a balanced integration between NMS and public transport services.

Modeling to Assess the Impact of MaaS and NMS

The first step toward the development of robust demand modeling tools and methodologies requires an in-depth understanding of NMS nature, as well as the challenges hindering their introduction, as shown in Fig. 1. Transport models can contribute to assess different implementation strategies for MaaS with their ability to provide a virtual testbed for scenario testing and the possibility to run sensitivity analysis. This can help to identify and frame uncertainties before embarking into NMS trials. Moreover, transport models can help to get a better understanding of potential problems caused by the introduction of NMS and their subsequent effects on multiple levels (e.g. land use, jobs, growth, congestion, public transport).

Transport modeling could also be used to identify scenarios where reduction of patronage from public transport sector is likely and assess mitigation strategies. It could be used to quantify accessibility improvements for areas covered by NMS and facilitate the estimation of induced demand due to such improvements. From a commercial point of view, proper use of specialized modeling tools could minimize the risk of failure for NMS by assessing their potential prior to their actual roll-out.

However, it is important to define the problem to solve and identify the customer for the model, choosing the appropriate tool and datasets according to their needs and the stakeholders involved in the creation of the model. When mobility services and

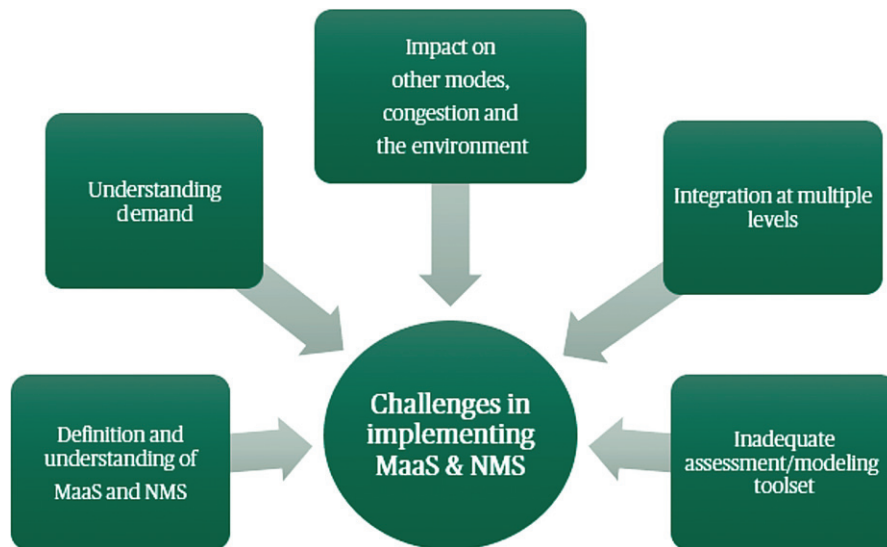


Figure 1 Challenges in implementing New Mobility Services and Mobility as a Service.

mobility on demand are introduced key questions that a model should be able to address look quite different depending on each stakeholders' perspective.

During the "Business Case for New Mobility Services (NMS): Demand Modeling Tools" project (CPC, 2019), a series of stakeholder engagement activities were undertaken with the aim to define the problem space, evaluate the market needs and the existing modeling landscape capabilities.

Stakeholders Perspective

Each stakeholder has different views on what MaaS is and what should deliver. Transport modeling and specifically demand modeling is identified as a way to de-risk the introduction of NMS and turn them in a sustainable viable proposition to link up traditionally less flexible public transport services.

Personal preference, accessibility, and attitude toward shared mobility play a big role in understanding demand. Estimating the effect NMS have on travel choices both on a short-term basis (e.g. planning and operational purposes) as well as on a long-term basis (e.g. development purposes) can be a challenge. Moreover, NMS should be inclusive, increase access to transport and minimize transport poverty, defined as lack of access to transport services. NMS are not expected to be used for long journeys but only to cover those first/last mile journeys during commuting to work.

Integration at multiple levels with the more complex public transport ecosystem should be introduced to allow users to access MaaS through a single application which provides access to mobility, with a single payment channel. However, there are no adequate tool to predict demand for NMS, hence hindering the implementation of sustainable MaaS schemes around the world.

Local Authorities' Perspective

When mobility services and mobility on demand are introduced, key questions that a model should be able to address are different depending on each stakeholders' perspective. From a transport/local authority's point of view the most interesting questions are focusing on the wider social, economic, and environmental benefits (or disbenefits) that the introduction of NMS could entail as shown in Fig. 2.

Local authorities see NMS as a way to increase the integration of mobility services, to enhance the actual provision of public transport services, and sometimes substitute it in areas where the level of service is too low to justify a continuous PT service running. They would like to introduce innovations to optimize public transport services, however they do not know how to implement the appropriate combination of mobility services to implement MaaS.

The ability to provide a virtual testbed for scenario testing and the possibility to run sensitivity analysis can help to identify and frame uncertainties before embarking into NMS trials. Transport models can help to get a better understanding of potential problems caused by the introduction of NMS and their subsequent effects on multiple levels (e.g., land use, jobs, growth, congestion, public transport).

Transport modeling could also be used to identify scenarios where monopolization of the existing transport market could occur and assess mitigation strategies. It could be used to quantify accessibility improvements for areas covered by NMS and facilitate the estimation of induced demand due to such improvements.

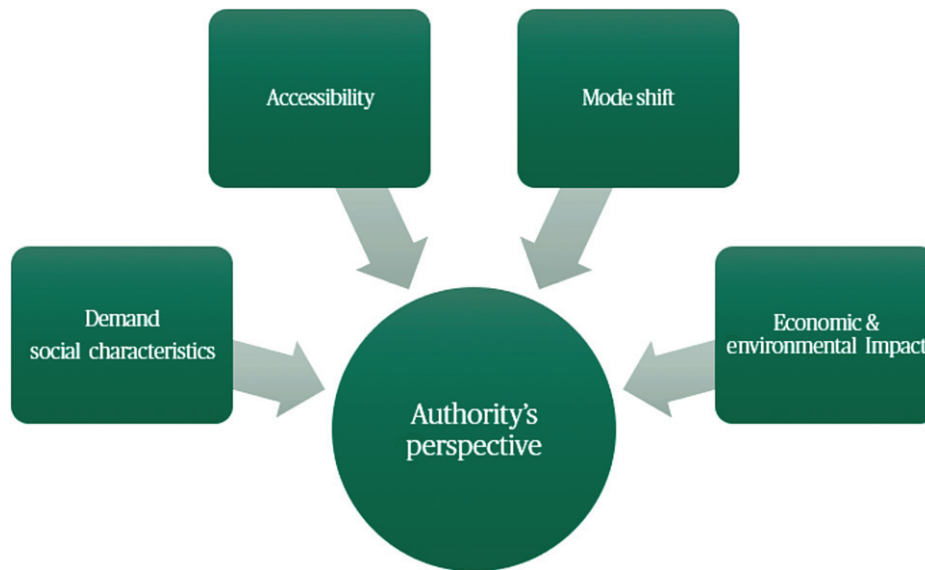


Figure 2 Key areas of concern from a Transport or Local authority's point of view regarding NMS and MaaS.

Transport Operators' Perspective

Transport operators are mostly interested in transport models which will enable them to prepare for a successful and smooth introduction of mobility services, minimizing the risk of failure for NMS by assessing their potential prior to their actual roll-out. Some key areas of concern from a transport operator's perspective regarding the introduction of NMS and MaaS are related to revenue, costs and operations of the service.

These key areas could be informed by the application of specialized transport modeling tools. Such tools can enable for instance, the design of an appropriate pricing structure as well as the estimation of the required infrastructure investments (e.g. required fleet size and personnel). Moreover, sensitivity analysis and simulations can be used to assess the cost of multiple implementation strategies without the need of expensive trials. Finally, a relevant transport model can provide insight regarding the identification of appropriate target groups for the service and the locations where to run it.

Another area of great importance to transport operators is related to the day-to-day operations and how these could be optimized in order to provide a better service for the customer while minimizing costs, such as optimum routing and efficient allocation of the service to the identification of the most appropriate area of coverage. A special mention was made regarding the importance of modeling tools (especially simulation-based ones) to help with the mitigation of disruption and unexpected events. Operators and transport providers are keen on identifying ways to deal with disruptive events while at the same time maintaining an acceptable level of service.

Data Required to Model MaaS and NMS

Transport models do not differ from other models in terms of data requirements since they also depend on a wide range of information to make accurate predictions. Although significant improvements have taken place over the last decades regarding data availability, many general concerns around the usability of data are hindering the use of data-driven approaches to transport modeling. These data requirements are grouped and presented in Fig. 3.

A better use and enhancement of already available sources (e.g., census, travel surveys, traffic counts) which have reliably assisted transport modelers for many years should be promoted, and therefore efforts should be concentrated towards their improvement (e.g., acceleration of their update, greater granularity). For instance, census data usually include most of the required information for most transport models, but their 10-year update cycle can quickly render them obsolete.

The recent technological advances, especially in the field of telecommunications and personal connectivity (e.g., smartphones, Internet of Things, Location-based data such as Mobile Phone data and GPS) has enabled the acquisition of very fine-grained data in vast quantities. The issue with using them for transport modeling purposes usually derives from the lack of standards to anonymise data. Data providers and transport modelers should cooperate in creating methodologies that will enable the use of data at a highly disaggregated level without any compromise on the protection of personal information. As an example, Mobile Network Providers have at their disposal detailed information regarding the mobility patterns of their customers. These data can be used to inform a transport model (Franco et al., 2020), but no widely accepted methodology as to how this information can be securely anonymised and aggregated without losing information at disaggregate level has been presented so far.

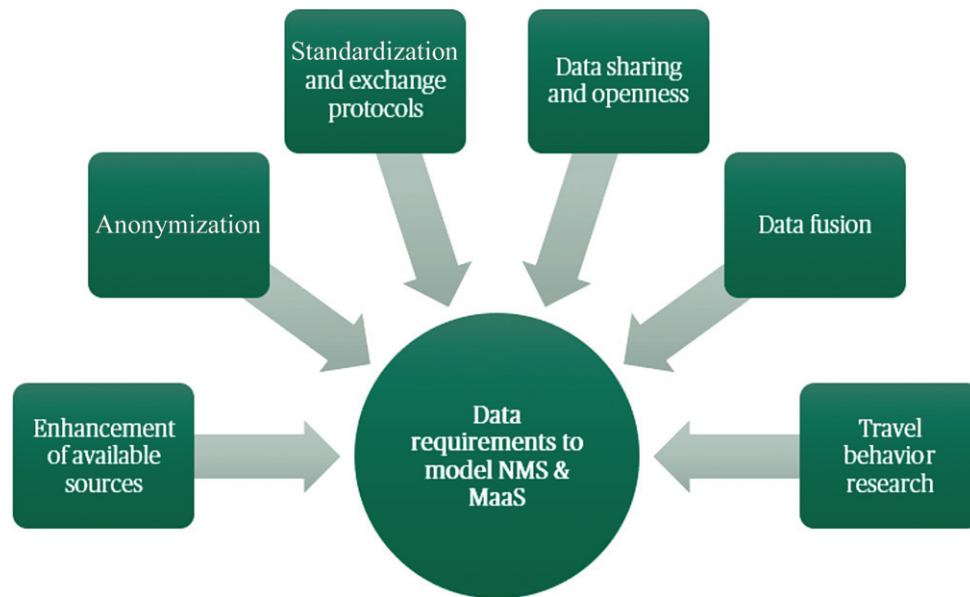


Figure 3 Identified data requirements.

Generally, a data-driven approach to transport planning is perceived as more secure and fit to model new trends and technologies, however, since mobile network data is used and collected for other industries, data mining and fusion is fundamental to provide a good data input for transport models. Especially for modeling the impact of MaaS, an activity-based approach would be preferable to represent the end-to-end user journey because provides the possibility to represent real-world and complex travel patterns when built using highly disaggregated data.

Demand Modeling—Key Changes

Modeling demand for mobility services is substantially different from the traditional approach due to the introduction of the “user centricity” element. The provision of a fully inclusive mobility service to satisfy all users’ needs, relies on the ability to model the end-to-end users’ journeys and the interdependence existing between one trip and another. This requires a better understanding of travel behavior and their attitude to share, so the likely uptake of new trends in the mobility sector can be derived.

Modeling tools have been experimenting more on Mobility on Demand and Demand Responsive Transport (DRT) with a specific focus on fleet operations and management without considering demand modeling as an integral part of the business model.

There are few elements that need to be considered when it comes to modeling MaaS:

- The ability to represent NMS’ integration with existing public transport systems;
- The ability to consider door-to-door, multimodal journeys and the correlation between each trip;
- The shared element of on-demand mobility services;
- The ability to provide an overview of the demand for travel over 24 h to serve current users but also capture latent demand in different time of the day and for a large-scale area.

Based on these characteristics, Transport Agent-based Models (ABM) are becoming a viable alternative solution to the traditional trip-based approach. A specific characteristic of this type of model is that they require a higher granularity data and are computationally intensive. All features that can be addressed with the data rich environment created in smart cities and cloud computing.

Connected Places Catapult has already explored this solution in the two Innovate UK funded projects “Mobility on Demand Laboratory Environments” (MODLE) (Franco et al., 2018, 2020a) and Merge Greenwich (Segui-Gasco et al., 2019), where ABMs are using a data-driven approach to test flexible Demand Responsive Transit, operated either with traditional or automated vehicles. ABMs can be developed either using a trip-based approach or an activity-based approach. In the latter case, the activity-based approach allows to model the complex interaction between flexible private mobility services and the more complex public transport ecosystem. In the MODLE simulation, demand and travel behaviors were informed by anonymised and aggregated Mobile Phone Network Data (MND). MND reveal real travel patterns from users using trip-chains data to derive an activity-based approach to model complex travel patterns for the daily activity plans of the synthetic population. Results showed that the traditional datasets based on the use of disconnected trips produce a heavy congested network which does not provide useful insights in the behaviors of users (Franco et al., 2020).

The requirements identified for creating demand models to represent and appraise NMS are reported in Table 1.

Table 1 Requirements for demand models to model mobility (Franco et al., 2020c).

Requirements to create a large-scale demonstrator for modelling and appraising New Mobility Services'	
Model	• Agent Based Model with Activity based approach: • End-to-end users' journeys • Multimodality • Contribution of shared Mobility towards increase in patronage in Public Transport • Microlevel approach to include new technologies, such as CAVs, for shared mobility services running with autonomous and conventional vehicles using dynamic agents • Appraise impact on the network and the environment
Data landscape	• Location based data (Mobile Network Data, GPS, satellite data, ...) as data input to generate daily travel patterns from users • Use of highly disaggregated data to make use of big data source coming from sensors and new technologies • Use of existing Data Centres, with consistent and long records of events for calibration and validation purpose of the model • Appraise the Impact of New Mobility Service on the environment using a highly granular environmental data
Other factors to consider for shared mobility	• Attitudes of people towards sharing to understand the uptake of New Mobility Services and how the Mobility as a Service offering will be affected by this propensity • Changes in the value of time for ride-sharing mobility services and how this will affect the costs in the model • Understanding the willingness to pay when users are presented with a range of conventional and new transport options • The use of Artificial Intelligence to accelerate the procedure to automate the travel pattern recognition • Integration with legacy models for strategic modelling and optimization of the mobility service operations

Based on the requirements mentioned in [Table 1](#), a new methodology was created during the 2019 "Demand Modeling and Assessment through a Network Demonstrator" (DeMAND) project, which allows to generate a large-scale agent-based model, which represents a regional digital twin ([Franco et al., 2020b](#)).

The DeMAND agent-based model has demonstrated the ability to model a large-scale area, using only readily available network information and latest generation of Mobile Network Data providing a tool to inform the transport authorities in the introduction of shared mobility. The use of agent-based modeling and the behavioral models enhance the public transport services through a bottom-up approach that holistically consider all transport choices available to users and their preference for travel, but also the different values of time associated with each mode of transport (i.e. active travel, public transport, shared mobility, private car). The methodology can also be extended in newly developed areas, where little information is available and in rural areas, which are typically considered of low demand for travel.

Conclusions

To implement MaaS and NMS, a full understanding of travel behaviors and transport choices is pivotal to assess the impacts on congestion, the environment and health of population. The public transport sector was adversely affected by the introduction of NMS and as such transport models could help in identifying impact, gaps, lack of service provision and integration of systems at various level. However, the potential of this methodology for demand modeling can stretch further given that the agent-based model with activity-based approach can offer the opportunity to test various policy interventions. For example changing the demand for travel, such as the restrictions introduced during the Covid-19 pandemic, either during the lock down phase or during the reopening of the activities, the possibility to target specific segments of the population or restricting certain activities.

Traditional traffic modeling software platforms are not designed to allow for the introduction of new modes of transport or to measure the impact on people's activities following the introduction new policy interventions, especially at large-scale, as they focus on aggregated trip-based, siloed mode-based approaches, whereas the focus in this new domain needs to consider people-centric,

multi-modal, end-to-end journeys, as well as grouping individuals according to their socio-demographic characteristics and even personal preferences during traveling.

Transport is expected to face multiple challenges and therefore tools able to provide insight regarding its future are essential. The tools currently available to transport operators and authorities, are not adequate to answer the complex questions related to mobility and will most likely require significant enhancements or complete re-engineering. Even though the need for development of more sophisticated transport models has already been recognized (Department for Transport, 2018; European Commission, 2016), there seems to be a long way until transport models will be capable to accurately represent travel behavior in the era of NMS and MaaS.

The Connected Places Catapult has started to develop tools and methodologies capable of providing answers. The data preparation approach and open source modeling tools trialed during the DeMAND project offer promising advanced capabilities in this field over commercial tools, however agent-based models do not yet offer a stable, de-risked platform from which the modeling industry can take over. Agent-based Models are subject to calibration and validation issues due to lack of guidance and standardization. Further investment in this research and convening the knowledge of industry and academia will help to stimulate the development path of modeling tools and techniques as well as leading the ability to create demonstrator models capable of modeling the impact of NMS at scale.

Acknowledgment

The author would like to thank the UK Department for Transport for funding the “Business Case for New Mobility Service: Demand Modelling tools” to advance the knowledge on large-scale demand agent-based modelling for urban and rural mobility solutions; and Nila Sari from the Department for Transport for the constructive feed back, Dr. Djibril Kaba and Dr. Haris Ballis, and other colleagues at the Connected Places Catapult for their support during the development of the project.

Biography



Patrizia Franco is a Principal Technologist in Demand Modelling at Connected Places Catapult (created in 2019 and formerly known as Transport Systems Catapult and Future Cities Catapult) interested in research and innovation in new and emerging mobility services.

Former researcher in Transport Modelling and Data Analysis at Newcastle University (UK), she has more than 15 years' experience in transport planning and policy, strategic transport models, public transport systems and is specialised in Agent-based and Activity-based Modelling to model MaaS and Demand Responsive Transit to achieve greater sustainability and resilience in smart cities.

Patrizia is currently technical lead for collaborative projects working with the UK Department for Transport on urban and rural mobility issues, such as “Assessing Sustainable Transport Solutions (AsSeTS) for Rural mobility” and DeMAND (Demand Modelling and Assessment through a Network Demonstrator) to improve the knowledge base around data-driven demand modelling tools and de-risk the uptake of NMS in Urban and Rural Mobility.

References

- Atkins, Flood, A., Christopher, M., 2015. Journeys of the Future: Introducing Mobility as a Service. Available at: <http://www.atkinsglobal.com/en-gb/uk-and-europe/about-us/reports/journeys-of-the-future>.
- Ben-Akiva, M., et al., 2007. Towards disaggregate dynamic travel forecasting models. *Tsinghua Sci. Technol.* 12 (2), 115–130, doi:10.1016/S1007-0214(07)70019-6.
- Connected Places Catapult, 2019. Business Case for New Mobility Services: Demand Modelling tools - Executive Summary project report https://s3.eu-west-1.amazonaws.com/media.cp.catapult/wp-content/uploads/2020/10/27175058/Business-case-for-New-Mobility-Services_Demand-Modelling-tools_Executive-Summary.pdf.
- Chorus, C.G., Molin, E.J.E., Van Wee, B., 2006. Use and effects of advanced traveller information services (ATIS): a review of the literature. *Transp. Rev. Routledge* 127–149, doi:10.1080/01441640500333677.
- Datson, J., 2016. 'Mobility As a Service: Exploring the Opportunity for Mobility As a Service in the UK', (July), 1-52.
- Department for Transport, 2018. Appraisal and Modelling Strategy: Informing Future Investment Decisions, June 2018.
- Dia, H., 2002. An agent-based approach to modelling driver route choice behaviour under the influence of real-time information. *Transp. Res.: Part C. Pergamon* 10 (5–6), 331, doi:10.1016/S0968-090X(02)00025-6.
- European Commission, 2016. Important notice on the second Horizon 2020 Work Programme. Available from: https://ec.europa.eu/research/participants/data/ref/h2020/wp/2016_2017/main/h2020-wp1617-transport_en.pdf (Accessed: 18 September 2018).
- Falconer, R., Zhou, T., Felder, M., 2018. 'Mobility-as-a-Service The value proposition for the public and our urban systems'.
- Ferreira, L., Charles, P., Tether, C., 2007. Evaluating flexible transport solutions'. *Transport. Plan. Technol.* 30 (2–3), 249–269, doi:10.1080/03081060701395501.
- Franco, P., Johnston, R., McCormick, 2020a. Demand Responsive transport: generation of activity patterns from mobile phone network data to support the operation of flexible mobility services. - Special Issue on Developments in Mobility as a Service (MaaS) and Intelligent Mobility. *Transp. Res. Part A: Policy Pract.* Vol. 131, January 2020, 244–266.
- Franco, P., Kaba, D., Close, S., Galatioto, F., Jundi, S., Johnston, R., Sari, N., 2020b. A new methodology to appraise demand for shared mobility services using mobile network data and activity-based modelling. *European transport Conference* (9–11 September 2020).
- Franco, P., Kaba, D. & Sari, N. (2020c) Framework for a large-scale demonstrator to model demand for New Mobility Services. *Transport Research Arena, Helsinki, Finland. Proceedings of 8th Transport Research Arena TRA 2020, April 27–30, 2020, Helsinki, Finland.*

- Franco, P., Johnston, R., McCormick, E., 2018. Role of Intelligent Transport Systems applications in the uptake of mobility on demand services, United Nation "Transport and Communications Bulletin for Asia and the Pacific, 2018, No. 88 - Intelligent Transport Systems", https://www.unescap.org/publications?field_publication_series%3A9126.
- Harper, C.D., et al. 2016. Estimating potential increases in travel with autonomous vehicles for the non-driving, elderly and people with travel-restrictive medical conditions, *Transp. Res. Part C: Emerging Technol.* Pergamon, 72, pp. 1-9.
- Holmberg, P.-E., et al., 2016. Mobility as a Service: Describing the Framework. Available at: https://www.viktoria.se/sites/default/files/pub/www.viktoria.se/upload/publications/final_report_maas_framework_v_1_0.pdf (Accessed: 21 September 2018).
- Jittrapirom, P., et al., 2017. Mobility as a service: a critical review of definitions, assessments of schemes, and key challenges. *Urban Plan.* 2 (2), 13., doi:10.17645/up.v2i2.931.
- Mladenovic, M.N., Trifunovic, A., 2014. The shortcomings of the conventional four step travel demand forecasting process. *J. Road Traffic Eng.* 60 (1), 5–12.
- Mulley, C., 2017. Mobility as a services (MaaS)—does it have critical mass? *Transp. Rev.* 247–251, doi:10.1080/01441647.2017.1280932.
- Rodier, C., et al., 2003. Car-sharing and car-free housing: predicted travel, emission, and economic benefits a case study of the sacramento, california region', University of California, Davis. Institute of Transportation Studies. Research report, 5059(August), p. 19p. Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.199.3791&rep=rep1&type=pdf>.
- Segui-Gasco, P., Ballis, H., Parisi, V., Kelsall, D., North, R., Busquets, D., 2019. Simulating a rich ride-share mobility service using agent-based models. *Transportation*, doi:10.1007/s11116-019-10012-y.
- TAG Unit M2.1 Variable Demand Modelling <https://www.gov.uk/government/publications/tag-unit-m2-variable-demand-modelling>.2000.

Policy and Planning for Walkability

Carlos Cañas Sanz^{*,†}, Maria Attard^{*,†}, ^{*}University of Malta, Msida, Malta; [†]Institute for Climate Change and Sustainable Development, University of Malta, Malta

© 2021 Elsevier Ltd. All rights reserved.

The Surge of Interest in Walkability	340
Walkability Policy Objectives	342
Policies to Ensure Pedestrians' Rights	342
Policies to Assure Pedestrian Accessibility	342
Policies to Provide Pedestrian Safety and Security	342
Policies to Promote Walking as a Health-Enhancing Activity	343
Policies to Encourage a Modal Shift	343
Walkability Policy Instruments	344
Engineering: Urban and Transport Planning for Walkability	345
Education and Encouragement: Raising Public Awareness on Walkability and Pursuing Behavioral Change	346
Enforcement: Regulations and Fiscal Measurements to Improve Walkability	346
Evaluation of Walkability Policy and Planning	346
Measuring Walking Activity and Experiences	346
Measuring Walkability	347
Measure the Impact of Walkability Interventions in Walking Activity and Experience	347
Conclusion	347
References	347
Further Reading	348

The Surge of Interest in Walkability

There is a global trend to advocate for walking and more pedestrian-friendly places (European Commission, 2013; UN-Habitat, 2014; WHO, 2018). This widespread policy agenda can be attributed to decades of multidisciplinary research and planning exercises showing the positive impacts of more walkable places and people walking (Berg et al., 2017; Mackenbach et al., 2014). To illustrate these findings, the report "Cities alive: Towards a Walking World" (ARUP, 2016) lists 50 benefits categorized in social, economic, environmental, and political sectors (Fig. 1).

Pedestrians travel on foot to reach specific destinations, engage in cultural and socio-economic activities, walk and use public spaces for physical or restorative exercises, with the possibility of combining different purposes in the same walk. The positive impacts, together with the ubiquity of walking and the disparity of their purpose, have drawn the attention of a wide variety of policy makers who increasingly see walkability as a key factor to address current challenges in the areas of public health, urban and transport planning, economic and social development, environmental sustainability, and cultural enrichment. This surge of interest came after a steady decline in walking activity in western societies in the 1970s, mainly attributed to suburban sprawl and the creation of car-oriented urban landscapes. Such developments along with an increasing sedentary lifestyle and diet changes, were directly related to new crisis, such as the obesity pandemic, traffic-related high mortality, lack of social cohesion, and environmental deterioration (Marshall et al., 2009). In the search for solutions, researchers, practitioners, and advocates focused on the alleged benefits of walkable places and active walking communities.

The development of walkability research was heavily influenced by the particularities of different disciplines that initially addressed this topic. Transport professionals initially focused on large-scale determinants, such as the urban fabric, land use and transport networks, to study pedestrian accessibility and connectivity to destinations and services (Cervero and Kockelman, 1997). Meanwhile, urban planners focused on the streetscape and pedestrian activity in public spaces (Gehl, 1987), where the street is redefined as an arena for social interaction and not just a means of access (Appleyard, 1981). Health scientists, on the other hand started to address the obesity pandemic at the end of the 20th century with new studies concluding that moderate physical activity constitutes a key determinant on health (Hardman and Stensel, 2009) and subsequently identifying walking as the most accessible and affordable source of health-enhancing physical activity (Owen et al., 2004). The subsequent adoption of transport methods and planning principles by health scientists prompted a new multidisciplinary approach (Frank et al., 2005).

The surge of empirical studies proving health benefits made walkability a political priority. This encouraged different disciplines to use this new measurable concept to investigate further implications of pedestrian activities and pedestrian-friendly environments. Economists focused on the associations between walkability and property value, store sales, and cost savings related to public health, transport and energy consumption (Litman, 2018). Sociologists addressed the relationships between walkability and social cohesion, interaction, public participation, and social justice (Anciaes, 2011). Furthermore, geographers and behavioral scientists

Positive impacts of walkable places and pedestrian communities

SOCIAL BENEFITS		ENVIRONMENTAL BENEFITS	
Health and wellbeing		Virtuous cycles	
Promoting active lifestyles Addressing the obesity crisis Reduction of chronic disease Improving mental health		Decreasing dependency on non-renewable resources Optimising land use Addressing air pollution Reducing ambient noise Improving urban microclimate Increasing permeable surface for water drainage	
Safety		Liveability	
Improving traffic safety Increasing passive surveillance Reducing crimes		Beautification of streets landscape and public spaces Implementing sittability and recreational facilities	
Placemaking		Transport efficiency	
Promoting vibrant urban experiences Enhancing sense of place Encouraging art and cultural initiatives		Reclaiming underused space from vehicles Encouraging a model shift from motor vehicle travel Promoting flexible commuting schemes Increasing permeability in the urban fabric Bridging barriers	
Social cohesion and equality		ECONOMIC BENEFITS	
Broadening competitiveness Fostering social interaction Strengthening community identity Developing intergenerational integration Encouraging inclusiveness		Local economy	
POLITICAL BENEFITS		Boosting prosperity Supporting local business Enhancing creative thinking and productivity	
Leadership		City attractiveness	
Fostering competitiveness Building public consensus		Enhancing city branding and identity Promoting tourism Encouraging inward investments Attracting creative class	
Urban governance		Urban regeneration	
Promoting citizen empowerment Encouraging participation of multiple stakeholders Enhancing civic responsibility		Increasing land and property values Activating street facades	
Sustainable development		Cost savings	
Promoting sustainable behaviors Addressing city resilience Planning opportunities Allowing flexible and micro-solutions Promoting cultural heritage		Shrinking congestion costs Construction and maintenance cost savings Reducing healthcare costs	

Figure 1 Benefits of walkability (ARUP, 2016).

attempted to define an overall common ground for walkability research where all the relevant elements are integrated and interconnected in proposed frameworks (Andrews et al., 2012). Whilst initially policy making addressed walkability by focusing on isolated and specific issues, such exercise to tackle obesity or minimizing pedestrian injuries in case of collisions, it gradually combined policy objectives and measures to address more multidisciplinary issues, such as sustainable urban mobility and liveability. The COVID-19 pandemic has further accentuated the importance of walkability as a policy to increase walking and reduce infection risks in crowded public transport.

Walkability Policy Objectives

Policies aiming at enabling and promoting pedestrian-friendly places and pro-active pedestrian communities are broad in scale and range. This section presents international policies developed to support and enable pedestrian mobility and other activities. Although the majority of these policies have no legally binding force for individual states, they constitute a guide for policy making and legal reform at national and local scale.

Policies to Ensure Pedestrians' Rights

The 1988 European Charter of Pedestrians' Rights constitutes the first attempt to establish a framework to integrate a pedestrian-friendly approach into policies and practices. This document elaborates upon a list of pedestrian rights and concepts. First, the need for healthy and safe pedestrian environments. Second, the need for pedestrian accessibility to destinations based on walking distance and pedestrian accessibility at street level, including physically and mentally impaired pedestrians with limited mobility. Third, the need to preserve public space for socio-cultural interactions. Fourth, the need to balance or reverse the car-oriented infrastructure, giving back public space to pedestrians and limiting the negative impacts of excessive car use. And finally, the need for an integrated multimodal transport system, with significant importance to public transport, to aid pedestrians when walking becomes unfeasible.

In a more recent context, the 2006 International Charter for Walking (Walk21) was recognized by the World Health Organization (WHO) as a comprehensive framework to help authorities refocus existing policies and create a culture where people choose to walk.

Policies to Assure Pedestrian Accessibility

While mobility is an integral part of life, accessibility is necessary to participate in society and can be seen as a prerequisite for obtaining income and wealth. The 2017 European Pillars of Social Rights includes the right to equal opportunities and access to the labor market, education and to goods and services available to the public. Based on this concept, transport systems should offer accessibility to destinations regardless of the means of transport. In this case, pedestrian accessibility to destinations is based on proximity and availability of goods and services within walking distance, which can be assessed at urban scale.

A second perspective for pedestrian accessibility is based on the need for universal design of the built environment to be usable by all people, regardless of their physical or mental capabilities. This type of accessibility is usually assessed at street level and can be guided by the principles of the 2006 United Nations Convention on the Rights of Person with Disabilities. Although such provisions are promoted across financial instruments such as the European Structural Funds, there is currently no obligation at EU level to meet accessibility criteria for new or renovated infrastructures. The third perspective on pedestrian accessibility is based on the availability of public space dedicated to pedestrians to engage in cultural and socio-economic activities, which can be assessed through the study of urban land use distribution. Based on principles of equity and social justice, the main concern in this respect is the large amount of public space that is often dedicated to the movement and parking of cars at the expense of pedestrians (see for example [European Commission, 2004](#)).

Policies to Provide Pedestrian Safety and Security

The main concern on pedestrian safety is related to collisions with motor vehicles. At least 51,300 pedestrians were killed on EU roads over the period 2010–18. While the number of pedestrian deaths has decreased by 2.6% on average each year in the EU over the period 2010–18, this trend is inferior to the annual reduction of motorized road user deaths over the same period (3.1%) (ETSC, 2020).

Since 2003, legislation has been introduced at EU level to reduce pedestrian injury risks from traffic. These measures are limited to crash-friendly car fronts, advanced braking systems or blind-spot mirrors, which are technical requirements where the EU has legal competence over Member States. However, the EU has no legal binding force on other important actions, such as speed management, adequate infrastructure for nonmotorized transport and separation of dangerous mixed traffic. Since these issues are mainly related to urban management, most of the actions have to be in accordance with the Commission's 2009 Action Plan on Urban Mobility, which can only be incorporated into the legal framework by national and local authorities (see also next section).

One concept that is dictating the current policy on transport safety with relevant implications to pedestrians is the Vision Zero approach, conceived by the Swedish Government and Swedish Industry in 1994. It has since influenced policies, materializing in the EU road safety policy framework for 2021–2030 "Next steps toward Vision Zero". It specifically recognizes the responsibility for road

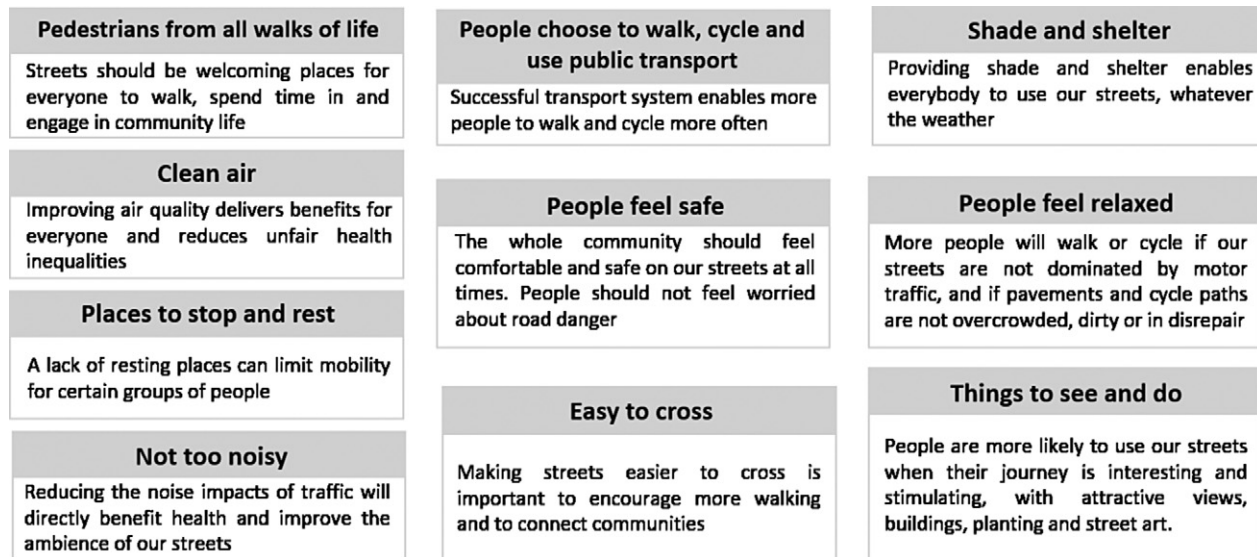


Figure 2 The 10 healthy streets indicators (Transport for London, 2017).

traffic safety as being shared between system designers and road users. One of its key principles is a 30 km/h speed limit based on the physical tolerance for a pedestrian to be hit by a well-designed car.

Although the main risk for pedestrian safety remains primarily linked to conflicts with traffic, there are growing concerns related to various forms of crime and terrorism across a number of cities worldwide further limiting pedestrian safety and security.

Policies to Promote Walking as a Health-Enhancing Activity

There is extensive literature on the physical and mental health benefits of walking as a moderate exercise and recreational activity, as well as the positive impacts of pedestrian-friendly environments to public wellbeing (Kelly et al., 2017). The WHO recommends adults to do at least 150 min of moderate-intensity aerobic physical activity throughout the week, performed in at least 10 min duration, which can be easily achieved by walking. Thus commuting on foot is increasingly integrated as an important contributor to transport-related health strategies (WHO, 2018).

At city level, London started to develop pedestrian strategies in the late 90s and published the first Walking Plan in 2004 addressing the environmental, socio-economic and health benefits of a more walkable city. Further on, the 2014 Transport Action Plan constituted a noteworthy example of transport and public health policy. This plan takes a “whole street approach” based on 10 Healthy Street Indicators (Fig. 2) to make streets more inviting for walking and cycling (Transport for London, 2017). The 2018 Walking Action Plan further elaborates on this Healthy Streets approach and envisions a future walkable city with specific and ambitious targets. It proposes four main actions: building and managing streets for people, planning and designing for walking, integrating walking with public transport, and leading a culture of change. It also consolidates the political willingness to achieve them with the appointment of the first ever walking and cycling commissioner.

Policies to Encourage a Modal Shift

The introduction of the mass-produced car in the 20th century represented a revolution in mobility and a milestone in the democratization of movement. The strive for speed and individual freedom based on car ownership led to a dominant car-oriented development where cities were planned to cater for the needs of private cars to the detriment of any other means of transport, and colonized most of the public space previously reserved for pedestrians. However, car-dependent cities started to experience a deterioration in the wellbeing of urban population as a result of traffic congestion, pollution, and road accidents.

The 2010 European Transport Policy White Paper aimed at reducing private car use while promoting more sustainable mobility at national scale. At the same time and considering the rapid urbanization across Europe, transport policies turned their attention to urban areas with the publication of the 2007 Green Paper: Towards a new culture for urban mobility. The need for rethinking urban mobility led the EU to develop a series of proposals and formulate collective policies to be implemented locally with the overall aim of promoting an urban mobility that allows for economic development, while preserving the quality of life of inhabitants and the protection of the environment.

One of the main proposals was to optimize a multimodal transport system, organizing comodality between public transport and individual modes of transport, and encouraging active travel. At this stage, European urban mobility was closely linked to environmental protection, healthy environment, land use planning, housing, and social aspects of accessibility and mobility. Thanks to this urban scale approach, walking has since been included as a relevant means of transport in European policies. The

2011 White Paper indeed considers walking as a relevant contributor to clean urban transport and commuting. The concept of Sustainable Urban Mobility Plans (SUMP) provides for a paradigm shift by focusing on people rather than directly on transport, with the aim of creating the much needed shift to greener alternatives to car ownership and use (European Commission, 2013).

This transport planning shift was reinforced further after the publication of the United Nations Sustainable Development Goals, with one particularly addressing sustainable cities and communities. All these current policy strategies set an encouraging scenario for the inclusion and development of walkability in the coming years.

Walkability Policy Instruments

The walkable environment, understood as the combination of the built, natural and social environment, is considered one of the main determinants influencing walking behavior. However, socio-economic status and personal abilities, as well as different purposes of walking, may require different types of walkable environments. Furthermore, it is recognized that the amount and frequency of walking is also influenced by personal mental states and the way in which individuals perceive the environment, as well as the individual's attitudes, norms and other behavioral controls.

In view of the disparity of walking determinants, numerous walkability socio-ecological models interlink all the conceptual factors that may explain walking behavior (Moudon and Lee, 2003). The need for organizing multiple factors and study their relative importance led Alfonzo (2005) to develop the hierarchy of "walking needs", proposing that in order for a person to choose to walk, the main determinants are feasibility, accessibility, safety, comfort and pleasantness. This approach serves as a framework to study how environmental characteristics are perceived to contribute toward experiences that affect people's decision to walk. Similarly, other researchers identified different pedestrian needs, such as usefulness, enjoyment, equity or sense of belonging (Mehta, 2008). All these pedestrian needs can be grouped in three categories. First, feasibility refers to the actual possibility for the walk to be achieved and it is strongly linked to personal capabilities and circumstances, such as physical and mental condition, economic resources and time availability. A second group includes accessibility, connectivity, linkage with other modes and usefulness, which refer to the proximity and variety of certain destinations within walking distance. Finally, there is a range of rather subjective concepts linked to pedestrian safety, comfort, pleasantness and sense of belonging that affect the walking experience. Figure 3 presents a conceptual framework on walking determinants.

In order to enable and promote walking and other pedestrian activities, its interdisciplinary and multiscale nature requires the collaboration of a wide range of experts, agencies and institutions to develop comprehensive programs and policies to combine certain measures and interventions. Although initial pedestrian programs were primarily focused on improving the walkable

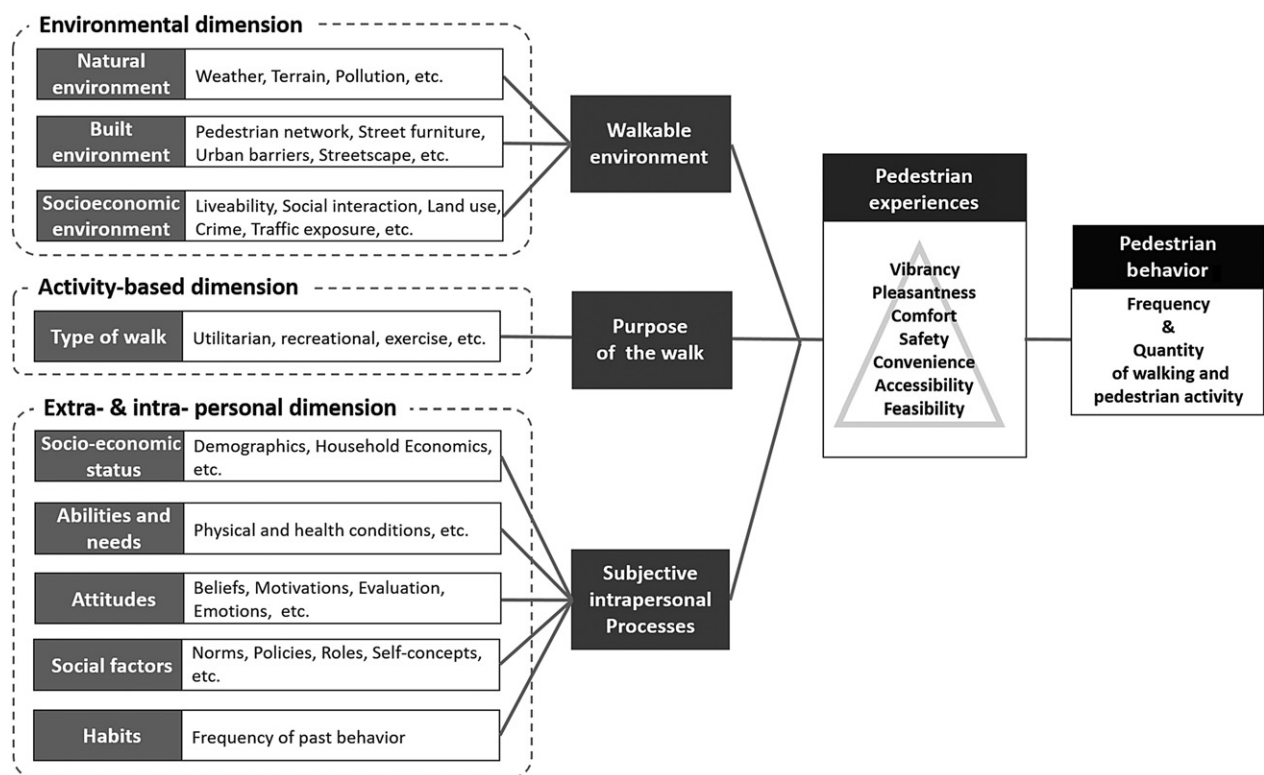


Figure 3 Determinants of pedestrian experience and behavior.

Table 1 The '6-Es' for walkability policy

Engineering	Conduct environmental and urban planning to make physical improvements in land use, urban fabric, streetscape, and pedestrian infrastructure toward a more pedestrian-friendly environment. Implement transport planning to provide a multimodal transport system that support pedestrian mobility with affordable and efficient public transport.
Education	Organize awareness campaigns and educational programs on the positive impacts of pedestrian-friendly places and proactive pedestrian communities. Campaigns on the negative impacts of car-dependency, physical inactivity and social isolation.
Encouragement	Implement behavioral change methods to promote walking over private car use, including incentives to pedestrians and disincentives for private car use. Organize events and activities to promote walking. Promote partnerships between governmental institutions with civil, environmental and cultural organizations to organize pedestrian related activities, such as walking to school, car-free events in public open spaces, and competitions amongst walkable cities.
Enforcement	Enforce the laws to prevent unsafe and illegal behavior by drivers and pedestrians to obey traffic laws and respect shared public space or the one allocated to each other.
Equity	Conduct urban and transport planning that takes into consideration all people, regardless of their physical or mental capabilities. Promote walkable places that give children and the elderly independence of movement in urban areas.
Evaluation	Assess the impact and success of the approach. Evaluate if the program or initiative results in an increase of the frequency, quantity and quality of walking and other pedestrian activities. Identify unintended consequences or opportunities to improve the effectiveness of the approach.

environment, as the walkability concept developed, more comprehensive interventions started to appear. The '4-E' approach (Sisiopiku and Cruzado, 2002) combines engineering, education, enforcement, and encouragement concepts. This approach has since been widely accepted by government bodies and advocacy groups in addressing safety concerns and planning for transport systems in general and pedestrian mobility in particular. The program was then upgraded to the "6-E" including the concepts of equity and evaluation (Buttazzoni et al., 2018). Table 1 shows how this approach can be applied to walkability policy.

Engineering: Urban and Transport Planning for Walkability

Walkability policies tend to define an appropriate environment and attempt to make the necessary interventions to provide a more walkable and pedestrian-friendly place. However, what initially may seem like a straight forward question—what makes a more walkable or pedestrian-friendly place?—has been actually answered in many different ways (Forsyth, 2015), leading to the current heterogeneous theoretical and operational groundwork of walkability research and policy.

Walkability is often defined as the extent to which the built environment supports and encourages walking and other pedestrian-related activities. Some authors identify specific elements that arguably make a place more walkable, such as continuously linked pavements, pedestrian crossings, sitting areas, street lighting and signage, as well as aesthetic factors and visual interest along pedestrian routes (Ewing and Handy, 2009). Others include a list of environmental characteristics that a place should provide, such as safety, comfort and convenience. In this regard, the "3Ds" approach defines walkability in terms of "density, diversity and design" (Cervero and Kockelman, 1997), later updated to the "5Ds" by including "distance to transit and destination accessibility," while the "7Cs" is based on the concepts of "connected, convivial, conspicuous, comfortable, convenient, coexistence and commitment" (Cambra and Moura, 2019).

The disparity in the term walkability can be further increased as the same characteristics are often considered differently. For instance, the recurrent term safety is often inconsistently defined among studies, as some refer to the risk of pedestrian-motor vehicle collisions, whereas others include hazards related to falls while walking and exposure to air and noise pollution. In addition, safety is sometimes considered from the perspective of crime in public space. Consequently, the actual definition of safety in a walkability policy strongly influences the specific determinants to consider in further interventions. For this reason, when the definition of the walkable environment includes intangible concepts, such as convivial or conspicuous, policymakers need to further elaborate on them to provide relevant and useful information for policy implication.

Despite the heterogeneity of desirable walkable environments, two main interventions on the build environment can be identified. The first one, based on the "3D" approach of urban density, diversity and design, requires the integration of land use planning and urban design to ensure walking proximity to certain destinations. The aim of these type of interventions is to plan a nonsegregated, dense and diverse urban land use distribution that offers a wide range of locations, activities and services to pedestrians, where the proximity to such places can be enhanced through continuous, well-connected and traversable pedestrian networks that include public transport to complement walking.

The second type of approach is based on physical interventions at streetscape level to improve pedestrian experience in terms of safety, comfort and pleasantness. This is often done by ensuring adequate, continuous and unobstructed pavements, safe pedestrian crossing and the availability of certain street furniture, such as sitting areas, street lights, and pedestrian signage, including accessible design for pedestrians with physical and mental impairments. Furthermore, these approaches often aim to minimize pedestrian exposure of traffic and to enhance positive characteristic and elements of the streetscape, such as enclosure, human scale and transparency, as well as the presence of urban green areas and public open spaces (GDCl, 2016).

Education and Encouragement: Raising Public Awareness on Walkability and Pursuing Behavioral Change

Apart from major investments to improve the walkable environment, there are a series of approaches known as “soft measures” or “smarter choices” to increase walking activity through behavioral change. They are intended to influence the perception of the general public to present walking as a feasible and desirable means of transport, especially for short trips or as a key opportunity within multimodal transport. This promotion of walking is often linked to the health, socio-economic and environmental benefits of commuting on foot when compared to other means of transport. They are not based on any legal sanction or reward and therefore individuals, business and public entities can opt to whether adopt them or not.

In the EU context, the European Mobility Week gives all Member States the opportunity to explore the role of urban transport and walkability concepts to tackle urban challenges. Since 2002, the EU provides campaign resources with an extensive repository of successful behavioral actions towards sustainable mobility (more info on www.mobilityweek.eu). Similarly, since 2011 the European Union grants the Access City Awards that recognizes a city’s willingness, ability and efforts to become more accessible for people with disabilities.

Enforcement: Regulations and Fiscal Measurements to Improve Walkability

Pedestrian-friendly places often require specific regulations to avoid conflicts between different means of transport and the urban land use allocated to each other, as well as to prevent unsafe and illegal behavior by drivers and pedestrians. Specific regulatory instruments can control the accessibility to certain vehicles to influence traffic conditions in parts of the urban areas with high pedestrian traffic. Another key measure is speed limitation and the distribution of urban land use for specific means of transport, with particular attention to on-street parking. In a more comprehensive approach, land use regulations require the assessment of new developments to comply with local travel plans.

Sustainable mobility plans often include fiscal measures to discourage private car use while incentivising active travel. These include economic incentives, compensatory schemes and social assistance through workplace, school and residential travel plans, as well as public transport subsidies to complement pedestrian mobility. Deterrents to private car users include pricing for parking and road use, as well as fuel and other car related taxation.

Evaluation of Walkability Policy and Planning

In order to design more walkable places and assess the impact of the plans, it is necessary to measure walkability and the actual walking activity, before and after an intervention. This information provides crucial input for further planning interventions and investment decisions.

Measuring Walking Activity and Experiences

Assessment methods to measure walking activity can be grouped in three categories, although they are often combined. First, direct observations are used to count the number of pedestrian passing through a street or entering a certain location. Second, indirect objective methods measure the frequency, intensity and duration of walking activity. These have been facilitated by smartphone technologies that allow for greater geographic and temporal detail. Third, direct subjective methods include survey questionnaires, interviews and walking diaries, which can be used to better understand walking behavior.

While walkability concepts were entirely assimilated among transport and urban planners at the beginning of the 21st century, there are still three major challenges to implement walkability interventions. First, walking is often ignored or not adequately measured in transport surveys, as a result walking is heavily underestimated and misrepresented. Second, current data collection methods differ widely, making comparability difficult. And third, collecting walking data is often time consuming and resource intensive, which presents particular challenges for accurate measurement. The International Walking Data Standard (Sauter et al., 2016) proposes a detailed set of requirements for walking data that are consistent and comparable, identifying five key performance indicators (see Table 2).

As the concept of walkability moved beyond the limited idea of walking as means of transport or moderate physical exercise, more walkability approaches linked to liveability studies focused on recreational walks as restorative exercise, cultural and socio-

Table 2 Key performance indicators to measure walking activity (Sauter et al., 2016).

1	Share of people who have made at least one walking stage on the survey day
2	Average number of daily walking trips per person
3	Average daily time walked per person
4	Average daily distance walked per person
5	Mode share of walking based on: stages, main mode, time, and distance

economic activities. Plans for recreational or leisure walking activities that quantify the use of public space by pedestrians are increasingly recognizing the importance of assessing not only walking activity but also pedestrian experiences.

Measuring Walkability

In order to measure the walkability of a place, it is necessary to first establish the theoretical and operational definitions of walkability, which are based on the objectives and specific characteristics of each initiative. These definitions guide the identification of the components and characteristics that determine the walkability of a place, which in turn can be observed and measured.

For large-scale components, such as terrain, urban fabric, land use, and pedestrian network, the most common methods include the interpretation and analysis of secondary data, like existing cartography, aerial images, and remote sensing. When the scale of analysis includes components and characteristics at a block or street level, the most conventional methods are pedestrian checklists, walkability audits, surveys, questionnaires, and interviews.

After all the necessary data have been collected, the final step is to process, interrelate and interpret it to assess the walkability of the area under study. There is a wide range of walkability assessment approaches, which could be broadly categorized in three main groups. First, some assessments exclusively focus on the urban fabric morphology. Second, walking accessibility assessments focus on the availability and spatial distribution of destinations within walking distance, including their characteristics and the ease of reaching them. And third, other assessments include and interrelate elements and characteristics of the walkable environment that influence walking activity and experience, to then calculate an overall walkability variable. They are commonly known as walkability composite indices (Hall and Ram, 2018), with a distinction between macro- and micro-scale assessments, the later known as walkability path indices.

Measure the Impact of Walkability Interventions in Walking Activity and Experience

Walkability interventions require an assessment of their effectiveness in influencing walking behavior. However, there is a current gap in understanding how the magnitude of a change in walkability relates to a change in pedestrian volumes and walking experience (Cambra and Moura, 2019). Some of the evidence remains methodologically weak due to the previously mentioned challenges related to lack of data on walking and the disparity of walkability assessments. Thus, it has become challenging to compare planning strategies, as well as generalize and validate results to support further policies. Aldred (2019) praises the use of quasi-experimental techniques, improvements in routine monitoring of smaller schemes, and the use of new big data source, while pointing out that more qualitative research could help develop a more sophisticated understanding of behavior change. As a good practice example, Cambra and Moura (2019) conducted a before-after walkability assessment using a validated methodology, opting for longitudinal analysis of the pedestrian volumes in the intervention sites and control areas, and quasi-longitudinal surveys on the walking experience in the area.

Conclusion

More people walking in pedestrian-friendly environments is increasingly being seen as a key policy objective to improve public health, promote more efficient and sustainable urban and transport planning, enhance community liveability and boost local economy. As a result, current regional and national policy strategies are setting an encouraging scenario for the development and inclusion of walkability concepts at local level to help in the improvement of sustainable walkable urban areas and prosperous pedestrian communities.

Walkability has become a ubiquitous and multidisciplinary policy topic, as walking and other pedestrian activities take place on a daily basis, but people decide whether to walk or not and how much they walk depending on personal attitudes and capabilities, environmental determinants, social norms and other behavioral controls. Consequently, walkability policies require a comprehensive approach to enable and encourage people to choose to walk by addressing specific physical improvements of the walkable environment, behavioral changes through education and promotion, and the establishment of fiscal, administrative and social measures to encourage walking while discouraging the excessive use of private car use, as well as sedentary and other unhealthy or dangerous lifestyles.

References

- Appleyard, D., 1981. *Livable Streets*. University of California Press, Berkeley.
- Aldred, R., 2019. Built environment interventions to increase active travel: a critical review and discussion. *Curr. Environ. Health Rep.* 6 (4), 309–315.
- Alfonzo, M.A., 2005. To walk or not to walk? The hierarchy of walking needs. *Environ. Behav.* 37 (6), 808–836.
- Anciaes, P.R., 2011. Urban transport, pedestrian mobility and social justice. A GIS analysis of the case of Lisbon Metropolitan Area. Doctoral Dissertation. University of London.
- Andrews, G.J., Hall, E., Evans, B., Colls, R., 2012. Moving beyond walkability: On the potential of health geography. *Social Sci. Med.* 75, 1925–1932.
- ARUP, 2016. *Cities alive. Towards a walking world*. Retrieved from <https://www.arup.com/perspectives/publications/research/section/cities-alive-towards-a-walking-world> (accessed 9 April 2020).

- Berg, P., Sharmeen, F., Weijs-Perree, M., 2017. On the subjective quality of social interactions: Influence of neighborhood walkability, social cohesion and mobility choices. *Transport. Res. Part A* 106, 309–319.
- Buttazzoni, A.N., Van Kesteren, E.S., Shah, T.I., Gilliland, J.A., 2018. Active school travel intervention methodologies in North America: a systematic review. *Am. J. Prevent. Med.* 55 (1), 115–124.
- Cambra, P., Moura, F., 2019. How does walkability change relate to walking behaviour change? Effects of a street improvement in pedestrian volumes and walking experience. *J. Transport Health* 16, 11–25.
- Cervero, R., Kockelman, K., 1997. Travel demand and the 3Ds: density, diversity, and design. *Transport. Res. Part D: Transport Environ.* 2 (3), 199–219.
- ETSC (European Transport Safety Council), 2020. How safe is walking and cycling in Europe? PIN Flash Report 38. Retrieved from https://etsc.eu/wp-content/uploads/PIN-Flash-38_FINAL.pdf (accessed 9 April 2020).
- European Commission. (2013). A concept for sustainable urban mobility plans. Together towards competitive and resource-efficient urban mobility. Retrieved from https://ec.europa.eu/transport/sites/transport/files/themes/urban/doc/ump/com%282013%29913_en.pdf (accessed 9 April 2020).
- European Commission, 2004. Reclaiming city streets for people – chaos or quality of life? Retrieved from https://ec.europa.eu/environment/pubs/pdf/streets_people.pdf (Accessed 9 April 2020).
- Ewing, R., Handy, S., 2009. Measuring the unmeasurable: urban design qualities related to walkability. *J. Urban Stud.* 14 (1), 65–84.
- Forsyth, A., 2015. What is a walkable place? The walkability debate in urban design. *Urban Design Int.* 20 (4), 274–292.
- Frank, L.D., Schmid, T.L., Sallis, J.F., Chapman, J.E., Saelens, B.E., 2005. Linking objectively measured physical activity with objectively measured urban form: findings from SMARTRAQ. *Am. J. Prevent. Med.* 28 (2), 117–125.
- Gehl, J., 1987. *Life between Buildings: Using Public Space*. Van Nostrand Reinhold, New York.
- GDCI (Global Designing Cities Initiative), 2016. *Global Street Design Guide*. National Association of City Transportation Officials. Island Press, Global Designing Cities Initiative.
- Hall, C.M., Ram, Y., 2018. Walk score and its potential contribution to the study of active transport and walkability: A critical and systematic review. *Transportation Research Part D*.
- Hardman, A.E., Stensel, D.L., 2009. Physical activity and health. In: *The Evidence Explained*, 2nd ed. Routledge, New York.
- Kelly, P., Murphy, M., Mutrie, N., 2017. The health benefits of walking. In: Mulley, C., et al. (Eds.), *Walking. Connecting sustainable transport with health* Transport and Sustainability 9, Emerald, UK, pp. 61–80.
- Litman, T., 2018. Economic value of walkability. Victoria Transport Policy Institute. Retrieved from <http://www.vtpi.org/walkability.pdf> (accessed 9 April 2020).
- Mackenbach, J.D., Rutter, H., Compornolle, S., Glonti, K., Oppert, J.M., Charreire, H., De Bourdeaudhuij, I., Brug, J., Nijpels G., Lakerveld J., 2014. Obesogenic environments: a systematic review of the association between the physical environment and adult weight status, the SPOTLIGHT project. *BMC Public Health*, 14, 233.
- Marshall, J.D., Brauer, M., Frank, L.D., 2009. Healthy neighbourhoods: walkability and air pollution. *Environ. Health Perspect.* 117, 1752–1759.
- Mehta, V., 2008. Walkable streets: pedestrian behaviour, perceptions and attitudes. *J. Urbanism* 1, 215–245.
- Moudon, A., Lee, C., 2003. Walking and bicycling: an evaluation of environmental audit instruments. *Am. J. Health Promotion* 18 (1), 21–37.
- Owen, N., Humpel, N., Leslie, E., Bauman, A., Sallis, J.F., 2004. Understanding environmental influences on walking: review and research agenda. *Am. J. Prevent. Med.* 27 (1), 67–76.
- Sauter, D., Pharoah, T., Tight, M., Martinson, R., Wedderburn, M., 2016. International walking data standard. Treatment of walking travel surveys. Internationally standardized monitoring methods of walking and public space. *Walk 21*. Retrieved from <http://www.measuring-walking.org/> (accessed 9 April 2020).
- Sisiopiku, V.P., Cruzado, I., 2002. Safe ways to school. *Adv. Architecture Series* 14, 863–871.
- Transport for London, 2017. Guide to the healthy street indicators. Retrieved from <https://healthystreetscom.files.wordpress.com/2017/11/guide-to-the-healthy-streets-indicators.pdf> (accessed 24 April 2020).
- UN-Habitat, 2014. A New Strategy of Sustainable Neighbourhood Planning: Five Principles. Nairobi: UN-Habitat Urban Planning and Design Branch. Retrieved from https://unhabitat.org/wp-content/uploads/2014/05/5-Principles_web.pdf (accessed 9 April 2020).
- WHO, 2018. Healthier and happier cities for all. A transformative approach for safe, inclusive, sustainable and resilient societies. Retrieved from http://www.euro.who.int/__data/assets/pdf_file/0003/361434/consensus-eng.pdf?ua=1 (accessed 9 April 2020).

Further Reading

- Abley, S., 2005. Walkability scoping paper. Land Transport New Zealand. Retrieved from <http://levelofservice.com/walkability-research.pdf> (accessed 9 April 2020).
- Mulley, C., Glebel, K., Ding, D. (Eds.), 2017. *Walking. Connecting sustainable transport with health*. Transport and Sustainability Series 9. Emerald, UK.
- Gehl Institute, 2017. How to use the Public Life Tools. Retrieved from https://gehlpeople.com/wp-content/uploads/2020/03/PL_Complete_Guide.pdf.

Public Transport Subsidy and Regulation

Jonathan Cowie, Edinburgh Napier University, School of Engineering and the Built Environment, Edinburgh, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	349
Rationale for State Intervention in Public Transport Markets	349
Participative Requirement/Transport Poverty	349
Use of Land Space	350
Reduced Environmental Impact of Mobility	350
Safety	350
The Mohring Effect	350
Forms of Revenue Subsidy Payment	351
Regulation of Public Transport	351
Qualitative Regulation	351
Quantitative Regulation	352
Problems with Regulation	353
Limits Free Enterprise	353
Always a Second Best Solution	353
Asymmetry of Information	353
Regulatory Capture	354
Concluding Remarks	354
Biography	354
See Also	354
References	354
Further Reading	355

Introduction

Public transport in most parts of the world is both heavily subsidized and closely regulated, hence this chapter examines some of the key factors behind these two issues. In many senses, the need for subsidy and regulation lies at the very heart of the planners vs. the market dichotomy in terms of the provision of all transport services, not just those relating to public transport. In particular when operated along purely free market principles, this results in a multitude of issues, primarily due to the fact that such markets suffer from a high degree of market failures. While the term “market failure” can be very clearly defined, in this context we simply describe it as occurring where the market mechanism does not come up with the “right” answer. This is specifically in terms of allocating the appropriate user to the appropriate mode. It therefore requires intervention by government bodies (mainly in the form of transport authorities) to coordinate and plan such activities in order to correct or reduce the effect of such failures. This intervention can come in the form of direct provision (the government does it itself, such as is the case with education), financial (subsidy) or direct command (regulation), or some permutation of all three. This chapter therefore will concentrate on the last two mentioned and outline the main reasons for such market failures and hence the mechanisms used to correct for these.

Rationale for State Intervention in Public Transport Markets

Before measures of either subsidy or regulation are implemented in public transport markets, there needs to be a clear rationale for such interventions. As stated, this is due to the existence of market failure, and primarily that the market undervalues public transport. This is for a number of reasons, all of which can be related to the key roles played by public transport in modern society, which are listed below.

Participative Requirement/Transport Poverty

Not everyone in the country either owns or has access to (as a driver or passenger) private transport i.e. the car. The UK 2018 National Travel Survey (DfT, 2018) for example showed that 24% of households had no access to a car. More concerning, car availability has been found to be strongly related to income, with earlier travel surveys showing 47% of those on the lowest quintile lived in a household with no car, compared to a mere 8% in the highest quintile. Furthermore, 50% of single parent families did not have any access to private transport. Even these figures however tend to underestimate the scale of the problem, as they mask those individuals who may live in a household that does own a car, but one which is in constant use by another household member. Access therefore is still a major issue.

Public transport supports Barr's idea of transport as a "participative requirement" (Barr, 2020), that is, it is required in order to fully participate within society. Education is a similar need, where a lack of education (at a basic level) will not kill you but may make it difficult to "get on" in the world. Contrast this with an "absolute requirement," which relates to food, shelter and health services, which represent essentials for continued wellbeing. With a participative requirement however, the basic assumption is that some kind of minimum provision is required. In other words, not everyone needs an Oxbridge education and a Rolls Royce but rather access to some minimum level of provision, which in the transport context is normally fulfilled by public transport.

This also links into the idea of "transport poverty," which as Lucas (2018) highlights, arises from one of four/five factors; personal/general availability, economic barriers, time poverty and the available options are considered to be unsafe. In terms of mobility, public transport may be viewed as a participative requirement that can even alleviate transport poverty. Subsidizing such services can also be viewed as a progressive measure. Bueno Cadena et al. (2016) found that the effect of a highly subsidized public transport fare policy in Madrid resulted in greater uptake from those residing in poorer, rather than more prosperous areas of the city. On social equity grounds therefore, this provides a strong rationale for it to be subsidized.

Use of Land Space

At certain points in the day (and in some severe cases at all times during the day), there is an over provision of private vehicles on particular sections of land, that is, roads. Roads have a limited capacity, which in many instances is exceeded. This in turn creates congestion. In some sense, congestion purely on its own is not a problem as this is a private decision. In other words, if an individual chooses to sit in a traffic jam every day on the way to and from work, then that could be viewed as a choice and one they are free to make. In this context however, the problem is that congestion costs money in terms of loss of working time and is therefore economically detrimental. Estimates from Smith et al. (2016) suggest that traffic congestion costs the UK economy £8.4bn every year, with the DfT (2014) estimated a congestion cost of £1.9bn on the strategic road network alone. In simple terms, the efficient working of the economy is dependent upon an efficient transport system, that is, one that actually moves people and freight. Moving people in larger sized groups takes up less land space, either on a bus, LRT, metro, or a train, and hence reduces congestion. Based on a study in Melbourne for example, Nguyen-Phuoc et al. (2018) found that the withdrawal of all forms of public transport would increase the number of severely congested roads to almost 150%, resulting in gridlock across large areas of the city. Given the context, these may be considered to be very much low side estimates of what the general case would look like.

Reduced Environmental Impact of Mobility

Partly related to congestion, this concerns the more general issue of the adverse impact of travel on the natural environment and hence how this can be reduced. Creutzig and He (2009) for example highlight that transport was responsible for 23% of world energy related greenhouse gas emissions, with 74% of that total as a result of road vehicle usage. While it is difficult beyond the naive view to highlight public transport's role in this debate, Hensher (2008) in running a number of policy scenarios designed to reduce transport's overall contribution of greenhouse gas emissions, highlighted publicly funded investment in public transport as a key area. Without such investment, any proposed policy induced change would simply result in a short-term behavioral change and a gradual drift back to the status quo. Public transport therefore offers the possibility of reducing individual carbon footprints and provide more sustainable transport modes on a mass scale.

Safety

Private cars represent a hazard to personal safety, particularly those that travel closely together. Although slightly dated now, figures from Evans, (2003) show that in terms of accidents per kilometer traveled, the train is twice as safe as the bus and the bus twice as safe again as the car. To a certain extent however, this aspect has to be treated with a deal of caution, as there may be a trade-off between the speed of private transport and the safety of public transport—slower vehicles cause less injuries and fatalities. Nevertheless, it is easier to manage the overall safety of a given transport system if individual movements are combined on a larger scale, as this allows a tighter degree of regulation and coordination of movement. In the case of mainline rail infrastructure, for example, every movement is controlled by signal, hence significantly reducing the risk of collisions.

The Mohring Effect

More academic arguments for the payment of subsidy are based around the idea of "welfare benefits," with Xu et al. (2018) highlighting how this would transmit into usage through a theoretical stance based upon the Mohring effect (Mohring, 1972). This suggests that increasing the frequency of public transport services will reduce generalized cost for users in the form of decreased waiting times, and hence lead to an increase in the demand. In some senses, this is akin to the idea of the virtuous circle. This is evidenced by Jara-Diaz and Gschwender (2009), who found that making users bear the whole cost of public transport provision, i.e. unsubsidized, results in a reduction in frequency regarding its optimal social level; in simple terms, less is provided than is socially desirable. Hence subsidy is required to ensure the socially optimum level of provision. To that argument it can be added that in more contemporary times potential users of public transport services are unwilling to pay the full cost of the provision of such services to the standards they would expect. The only way this gap can be bridged, and hence any form of Mohring effect realized, is through the payment of subsidy. In some respects this is akin to White's argument (White, 2010), that higher levels of usage will result in higher subsidies being paid, not lower, the recent experience in London being a classic example (DfT, 2019).

It should be noted however, that Xu et al. (2018) also highlight stances that suggests the Mohring effect is far more restricted. Notably, they cite Hensher (1998), who suggested that lowering service fares through subsidy only had a marginal impact on public

transport modal share and also [Savage \(2004\)](#), who put forward the commonly held view that the payment of subsidy leads to inefficiencies and general management slack, hence largely offsetting any potential gains.

Forms of Revenue Subsidy Payment

The most basic division of forms of subsidy would split these into two, capital and revenue, the former where subsidy is targeted at capital investments and the latter operations. This section will only consider the second of these, which are primarily aimed at reducing the operating deficit between revenue and costs. In many senses, the form of revenue subsidy which is used reflects the institutional framework under which public transport is provided. In contemporary times how this subsidy is paid has become far more formalized, again reflecting changes in the institutional framework. Specifically subsidy payment arrangements have been put on a far more statutory basis between state and provider, whether the latter be a public or private sector organization. Previously, [ESCAP \(2003\)](#) highlighted that in the case of state railways, these were often legally required to meet any demand (both passenger and freight) at fixed prices, with no binding provisions for any compensatory payments, that is, subsidy. In terms of the European Union, Regulation 1191/69¹ was an early piece of legislation that introduced the idea of public service obligations (PSO) in recognition of the requirement placed on an operator to provide an (unprofitable) public service. For such services, the operator would then be entitled to compensation from the specifier of the PSO, normally a transport authority.

In terms of the available options, revenue subsidy can be allocated in one of three ways:

- direct award (deficit financing)
- through a competitive tender process (absolute and relative contracts)
- negotiated contract

As the name suggests, direct award is where the subsidy is paid directly to a single provider and has normally been in the form of deficit financing. [Matulova and Fitzova \(2018\)](#) for example give a good overview of how this actually “operates” with reference to the Czech urban public transport sector, where most (still publicly owned) operators receive a form of loss compensation based upon a PSO. This includes an element of profit in order to retain the required technical conditions, although in practice profit levels for most operators have been set at zero. The operator therefore negotiates directly with the transport authority in relation to services to be provided, the fares to be charged and the investments required to support such operations. The financial deficit this produces is then paid as a form of subsidy.

In competitive tendering, the service specification is drawn up by the transport authority and then put out to tender. Most will specify the frequency and fare to be charged but may still leave some freedom with regard to vehicle standards and the general development of the service over the period of the contract. In other cases, tenders can be very specific as regards all service elements, such as in London. Due to the tightness of such tenders, the lowest bidder is most commonly awarded the contract. Competitive tendering has been in place in several parts of the world for many years, hence difficult if not impossible to give an overview of the main issues in a single chapter. Interested readers are therefore directed to the long running Thredbo series of conferences (see, e.g., [Alexandersson et al., 2018](#)), where many of these issues have been discussed and developed.

Generally speaking, tendering processes have very high transaction costs, hence can be expensive. A further issue is that it can result in an inconsistency in the provision of services, with operators unwilling to heavily invest (both financially and in terms of corporate identity) during the period of the contract. This is why [Hensher and Wallis \(2005\)](#) see a role for negotiated contracts, particularly where the incumbent is recognized as an efficient provider, a greater focus needs to be placed on innovation to grow patronage and incentives need to be provided to encourage the investment to support such activities. Such arrangements would also negate the high transaction costs and reduce business uncertainty associated with contract work, thereby unlocking potential investment opportunities.

Having considered the main issues behind the subsidization of public transport, we now examine the issue of regulation.

Regulation of Public Transport

Regulation, in the general sense, comes in two main forms, either qualitative or quantitative, the latter often described as economic regulation, with the main divisions in the public transport context shown in [Fig. 1](#).

Qualitative Regulation

Qualitative regulation is primarily focused on licensing, and hence in the context of public transport, ensuring that all vehicles are “safe” and fit for purpose, drivers are competent to a set minimum level to operate such vehicles and those responsible for operation are professionally competent to ensure that operations are carried out in a manner that meet minimum safety standards. Without such regulation, we would have market failure, due to the issue of a lack of “perfect information.” Hence to give a very simple example, a bus user does not need to be a highly qualified automotive engineer in order to assess if the bus they are about to board is “safe,” as the regulatory framework performs that function on their behalf. The same is true of the abilities of the driver. Hence, users can use such services in the knowledge that all aspects of operation conform to some minimum standards that are deemed to be “safe.” In the absence of any quantitative regulation, then any operator meeting these minimum standards is licensed to operate

¹ All provisions of Regulation 1191/69 have subsequently been superseded by EU Regulation 1370/2007.

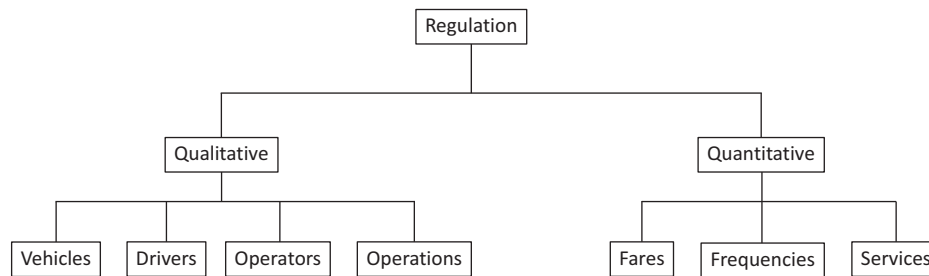


Figure 1 Main forms of regulation of public transport services.

public transport services. Note however that in some instances, transport authorities may specify criteria that are above the legal minimum, either to ensure the use of higher quality vehicles, or for some other (wider public interest) purpose; Transport for London for example require all buses to be painted red irrespective of the brand of individual operator of the service, as red buses are synonymous with London. In most countries, there will be a specific government body that ensures all such standards are upheld, and in the case of the UK, this comes under the remit of the Traffic Commissioners. Any breach of rules by operators may result in financial penalties or ultimately the suspension or complete withdrawal of the licence.

In many if not most developing countries, there is what is generally referred to as the “informal sector,” for which a very simple definition would be that public transport services are provided by small private operators without any form of regulation or other official endorsement (Chavis and Daganzo, 2013). While such a framework does provide some advantages, the main drawbacks have been identified as low vehicle and passenger convenience standards, poor safety records and significant contributions to environmental and congestion externalities (Venter, 2013). In many respects, all of these problems underline the need for qualitative regulation in the provision of public transport services.

Quantitative Regulation

Quantitative regulation on the other hand relates to placing conditions on the operator with regard to economic behaviour in the market, hence such controls can be applied with regard to the fare (or maximum fare) to be charged, the minimum frequency to be operated, the specific routes to be run and those “allowed” to operate those routes. In all such cases, this will be necessary because in the absence of such restrictions, operators would be in a position where the profit motive would outweigh the wider public interest. This generally occurs where there is a monopoly provider, that is, another example of market failure. Hence, the role of quantitative regulation is to ensure that public transport services are provided and consistent with the public interest, and not purely on the basis of (short term) profit.

Ideally, such measures should be overseen by an independent industry regulator, which has led some to observe that the “rise” of the regulatory state should provide for a depoliticization and depublicization of control of the regulated industry (Majone, 1997). As a consequence, the industry, in this context public transport, can be run to meet the wider needs of the public interest without being influenced by political considerations.

In terms of regulatory measures, we will only consider those relating to the fare to be charged, with the main instruments used outlined below.

Set the Fare

This is the simplest, and perhaps the most commonly used, form of fare regulation in the provision of public transport services, with numerous examples from across the world. The setting of the fare is illustrated in Fig. 2.

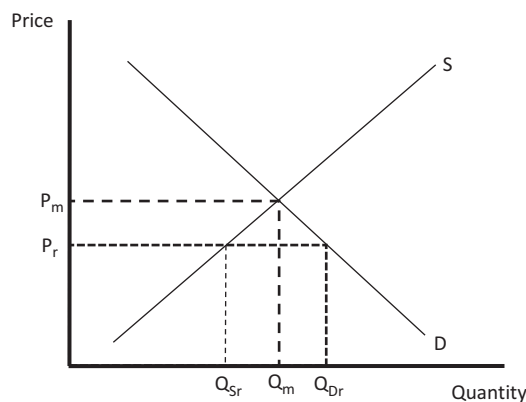


Figure 2 Fare regulation, set the fare.

Fig. 2 shows a generic market for public transport services, hence as the fare decreases, more will wish to travel by this service, shown by the demand curve (D), and as the price rises, more operators will be willing to provide services, shown by the supply curve (S). Where the two meet is known as market equilibrium, hence in this example this would give a market price of P_m , with the level consumed at Q_m . For one or more of the reasons previously outlined, this may be viewed as “too low” a level from a social welfare perspective. In order to encourage usage therefore, the regulator specifies a lower price, in this example P_r . Note that in order to be effective, that is, achieve wider public interest goals, any regulated fare needs to be set below the market fare.

One reason for fully illustrating what may appear to be a very simple concept is that without any further action, at the regulated fare demand would increase to Q_{Dr} but operators would only be willing to provide, that is, find it financially viable, quantity Q_{Sr} . This would result in massive overcrowding and completely unmet demand, which in a normal market situation would be addressed through increasing the price. With a regulated cap however, this cannot occur. Other inventory measures therefore need to be taken. In this case, this would normally be in the form of paying operators a subsidy in order to enable an increase in the level of supply to occur and to meet the demand at the regulated fare. Hence, subsidy and regulation are very much linked as part of an overall plan of public intervention.

Set Changes in the Fare

Rather than specify the actual fare to be charged, regulation can be in the form of specifying the amount the fare can change from one year to another. This is normally done in relation to the level of inflation, originally put forward by Littlechild (1983) in the form of RPI-X%, hence the retail price index (i.e., inflation) minus a factor specified by the regulator. In terms of public transport, this has been commonly applied in the case of passenger rail (Mirrlees-Black, 2014), a framework that has been in place over the full 23-year period of British private sector operation. One issue with such an approach, and particularly prevalent in the UK rail example, is that it assumes price differentials do not need to change over time, and hence by implication, were set at the “correct” level when the regulation was introduced.

Specify the Rate of Return (Profit) to be Gained

Rather than stipulate the fare or the change in the fare, the regulated price can be based on a rate of return. Hence over the regulatory period, all costs and investment needs are assessed, a profit rate is added, and from demand forecasts a fare level can be specified. This has been extensively used in the regulation of the British rail infrastructure provider (Network Rail), hence all operating costs and investment needs are appraised over a five year “control period,” and track access charges set accordingly to deliver the specified rate of return.

Problems with Regulation

Regulatory reform, and arguably regulation in general, has been described as often seen as a road paved by good intentions, but leading to “policy hell” (Lodge, 2002). Why this is the case is because the whole concept of regulation is to attempt to correct for some market failure, hence the planners intervene in the market to produce the “right” outcome. Through such endeavors however, this can introduce a wide variety of other issues. These were best summarized by Cowie (2010) as the only problem with regulation is the fact that it is needed at all. All of the following factors are extensions of this basic idea.

Limits Free Enterprise

von Mises (1949) put forward the theorem of consumer sovereignty, which to summarize, states that those firms that best serve the consumer are the ones that prosper the most. In many ways, this is the whole premise of the free market. Any regulatory framework restricts what the “market” can achieve. Hence as examples from the public transport context, restrictions on market entry may prevent the development of innovative new routes, while fare regulation may preclude the introduction of any user focused ticketing options. Innovation therefore is stagnated, and hence rather than consumer sovereignty, what results is regulatory authorities’ perspectives of what is in the best interests of the consumer, which can be quite different. The basic argument therefore is that regulation tends to lead to conservatism in the provision of public transport services.

Always a Second Best Solution

Regulation will always be a second-best solution, the “best” solution being viewed as the market. Given the full set of assumptions of perfect competition, the market will ensure that prices are as low as possible in order to ensure consumers maximize benefit and firms make a fair return (i.e., normal profits) which leaves sufficient funds for investment. The market also ensures efficiency in production, as inefficient firms will be driven out of business. All of these factors are reduced in a regulated market, as market mechanisms are subsumed by the regulatory framework, which will never be as efficient as the former. It can in some extreme instances offer protection for inefficient operators.

Asymmetry of Information

Cowie (2010) highlights that in order to regulate an industry effectively, the regulator requires a high level of data in order to plan and control operations. That data however is generated by those they are attempting to regulate, hence the operator knows more about their own business than the regulator does. Combined with cumbersome regulatory procedures (i.e. bureaucracy), this can lead to ineffective measures being used, and even where these may better reflect the reality of the situation, such measures may be easily avoided, with the consequence of regulatory frameworks becoming ineffective.

Regulatory Capture

Stigler (1971) put forward the idea of regulatory capture, which is a generic term to describe the situation where the regulator acts in the interests of the industry rather than in pursuit of wider public interest goals. This can be for a multitude of reasons. “Capture” need not reflect capture by the industry, but can also refer to the situation where the status and importance of the regulator rises with the importance of the industry, hence measures instigated by the regulator are not as “tough” on the industry as would be consistent with the public interest. Again therefore, the regulatory framework becomes ineffective.

Concluding Remarks

This chapter has examined the issues of subsidy and regulation in the provision of public transport services, in which the main principles behind such actions have been outlined. Primarily, both measures lie at the very heart of the market v the planners dichotomy, with government intervention in whatever form targeted at correcting for failures associated with provision strictly along market principles. Such issues however are closely associated with the very basic idea of what should be provided by the state and what should be left to individual responsibility, and to what extent such actions are supported by public funds.

Public transport is the only viable mass mover, hence a key element in reducing the detrimental impacts of mass mobility on both modern society and the natural environment more generally. It cannot do so however along purely market-based principles, as this results in high elements of market failure, particularly an under valuation of such services. Intervention, in the form of subsidies and regulatory measures, are required to produce a social welfare maximizing outcome. In other words, one that ensures the allocation of transport services results in the use of modes, particularly public transport, that are far more consistent with the idea of the wider public interest rather than one based solely on private benefits.

Biography

Jonathan Cowie is a lecturer in transport economics at Edinburgh Napier University, where he teaches masters level modules on public transport, transport policy and transport economics and appraisal. His research interests are best summarized as the supply side economics of transport markets, and he has published numerous articles on bus and rail economics. Jonathan is author of the book *The Economics of Transport*, and a longtime member of the Scottish Economic Society and the Chartered Institute of Highways and Transportation.

See Also

Planning for Public Transport with Automated Vehicles; Customer Satisfaction as a Measure of Service Quality in Public Transport Planning; Public Transport Network Planning

References

- Alexandersson, G., Hensher, D., Steel, R., 2018. Competition and ownership in land passenger transport (selected papers from the Thredbo 15 conference). *Res. Transport. Econ.* 69, 1–620.
- Barr, N., 2020. *The Economics of the Welfare State*. OUP, Oxford.
- Bueno Cadena, P., Manuel Vassallo, J., Herraiz, I., Loro, M., 2016. Social and distributional effects of public transport fares and subsidy policies. *Transport. Res. Rec.* 2544, 47–54, doi:10.3141/2544-06.
- Chavis, C., Daganzo, C., 2013. Analyzing the structure of informal transit: the evening commute problem. *Res. Transport. Econ.* 39, 277–284.
- Creutzig, F., He, D., 2009. Climate change mitigation and co-benefits of feasible transport demand policies in Beijing. *Transport. Res. D: Transport Environ.* 14, 120–131.
- Cowie, J., 2010. *The Economics of Transport; A Theoretical and Applied Perspective*. London, Routledge.
- DfT, 2014. National Policy Statement for National Networks. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/387222/npsnn-print.pdf, last accessed 29th August 2020.
- DfT, 2018. National Travel Survey: England 2018. Department for Transport, London.
- DfT, 2019. Transport Statistics Great Britain, 2019. Department for Transport, London.
- ESCAP, 2003. The Restructuring of Railways. Working Paper No. ST/ESCAP/2313. Available at: <https://www.unescap.org/sites/default/files/RailwayRestructuring.pdf>, last accessed 8th February 2020.
- Evans, A., 2003. Estimating transport fatality risk from past accident data. *Accid. Anal. Prev.* 35, 459–472.
- Hensher, D., 1998. Establishing a fare elasticity regime for urban passenger transport. *J. Transport Econ. Policy* 32 (2), 221–246.
- Hensher, D., 2008. Climate change, enhanced greenhouse gas emissions and passenger transport – what can we do to make a difference? *Transport. Res. D* 13 (2), 95–111.
- Hensher, D., Wallis, I., 2005. Competitive tendering as a contracting mechanism for subsidising transport. *J. Transport Econ. Policy* 39 (3), 295–321.
- Jara-Diaz, S., Gschwendner, A., 2009. The effect of financial constraints on the optimal design of public transport services. *Transportation* 36, 65–75.
- Littlechild, S., 1983. Regulation of British Telecommunications' Profitability. Report to the Secretary of State, Department of Industry, London.
- Lodge, M., 2002. The wrong type of regulation? Regulatory failure and the railways in Britain and Germany. *J. Public Policy* 22 (3), 271–297.
- Lucas, K., 2018. Transport poverty and inequalities. *Eur. Transport Res. Rev.* 10, 16–18.
- Majone, G., 1997. From the positive to the regulatory state. *J. Public Policy* 17 (2), 139–167.
- Matulova, M., Fitzova, H., 2018. Transformation of urban public transport financing and its effect on operators' efficiency: evidence from the Czech Republic. *Central Eur. J. Oper. Res.* 26, 967–983.

- Mirrlees-Black, 2014. Redirections on RPI-X regulation in OECD countries. *Utilities Policy* 31, 197–202.
- Mohring, H., 1972. Optimization and scale economies in urban bus transportation. *Am. Econ. Rev.* 62 (4), 591–604.
- Nguyen-Phuoc, D., Currie, G., De Gruyter, C., Young, W., 2018. Exploring the impact of public transport strikes on travel behavior and traffic congestion. *Int. J. Sustain. Transport* 12 (8), 613–623. doi:10.1080/15568318.2017.1419322.
- Savage, I., 2004. Management objectives and the causes of mass transit deficits. *Transport. Res. A: Policy Pract.* 38 (3), 181–199.
- Stigler, G., 1971. The theory of economic regulation. *Bell J. Econ. Manage. Sci.* 2 (1), 3–18.
- Smith, A.C., Holland, M., Korkeala, O., Warmington, J., Forster, D., Apsimon, H., Oxley, T., Dickens, R., Smith, S., 2016. Health and environmental co-benefits and conflicts of actions to meet UK carbon targets. *Climate Policy* 16 (3), 253–283.
- Venter, C., 2013. The lurch towards formalisation: lessons from the implementation of BRT in Johannesburg South Africa. *Res. Transport. Econ.* 39, 114–120.
- von Mises, L., 1949. *Human Action: A Treatise on Economics*. Yale University Press, New York.
- White, P., 2010. The conflict between competition policy and the wider role of the local bus industry in Britain. *Res. Transport. Econ.* 29 (1), 152–158.
- Xu, P., Wang, W., Wei, C., 2018. Economic and environmental effects of public transport subsidy policies: a spatial CGE model of Beijing. *Math. Probl. Eng.* 12, doi:10.1155/2018/3843281, Article ID: 3843281.

Further Reading

- Cowie, J., 2010. *The Economics of Transport*. Routledge, London (Chapters 10 & 11).
- Evans, A., 1985. Equalising grants for public transport subsidy. *J. Transport Econ. Policy* 19 (2), 105–138.
- Hensher, D., 2018. Public service contracts: the economics of reform with special reference to the bus sector. In: Cowie, J., Ison, S. (Eds.), *The Routledge Handbook of Transport Economics*. Routledge, London.
- Stigler, G., 1971. The theory of economic regulation. *Bell J. Econ. Manage. Sci.* 2 (1), 3–18.
- Vuchic, V., 2007. *Urban Transit: Operations, Planning and Economics*. John Wiley & Sons, New Jersey (Chapters 8 & 9).
- White, P., 2017. *Public Transport (The Natural and Built Environment Series)*. Routledge, London (Chapter 1).

Urban Regeneration and Transportation Planning

Thomas Vanoutrive, University of Antwerp, Antwerp, Belgium

© 2021 Elsevier Ltd. All rights reserved.

Urban Regeneration Strategies	356
Sustainable Mobility and Regeneration	356
Transit-Oriented Development	358
Experiences Around the World	358
Challenges and Criticisms	359
References	360
Further Reading	360

Urban Regeneration Strategies

For decades, both researchers and practitioners have acknowledged that transportation and urban development are inextricably interwoven. Despite this awareness, a recurrent theme is that transportation and urban planning are still separate worlds. To overcome this lack of integration, many attempts have been made to understand the causes of this separation, and concepts have been developed to close the gap. This situation is clearly visible in the wave of urban regeneration since the late 1980s.

Urban regeneration is usually associated with a more positive attitude toward cities, following a period when cities were predominantly perceived as problems. As the geographer PJ Taylor once stated: “‘Urban problems’ dominated urban research in the second half of the 20th century to create what might almost be termed a ‘hate literature’ on cities.” (Taylor, 2004, p. 3). Following decades of car-based suburbanization and massive job losses in inner-city areas, cities had developed into concentrations of unemployment, crime, poverty, and decay. While the prime focus has been on Western cities, urban areas elsewhere did not obtain a better rating.

In the decades preceding urban regeneration in the strict sense, renewal and redevelopment programs in the 1960s and 1970s invested in deprived neighborhoods. The emphasis was on the improvement of the conditions of the urban poor and on physical redevelopment. In contrast, since the 1980s, an entrepreneurial policy discourse has emerged with competitiveness and social cohesion as leading principles (Furbey, 1999). Urban regeneration is thus a historically specific concept, and is associated with the increased attractiveness of cities, both in discourse and practice. This has also been labeled “urban renaissance”, and one of the characteristics is that policymaking at the urban level has gained importance viz-a-viz the national level.

In particular since the early 1990s, a number of cities have become iconic examples of urban regeneration in which the bad image of urban decay was replaced by a view of cities as hotspots of creativity, cosmopolitanism and prosperity. The organization of the 1992 Olympics in Barcelona was accompanied by a regeneration program with, among other things, major investments in public space (Fig. 1). Five years later, in 1997, a Guggenheim Museum was opened in Bilbao, Spain. The building was designed by architect Frank Gehry and appeared in magazines all over the world. These are two leading examples of culture-led urban regeneration, and both large and small cities around the world have tried to copy the Barcelona model or the Bilbao (or Guggenheim) effect (González, 2011). In general, this type of urban regeneration reflects the transition from an industrial to a services-based economy, with a discursive emphasis on creative industries such as art, design, advertising, and film. Not only are these activities considered important economic sectors, they are also seen as magnets for investments by other sectors. Culture-led urban regeneration strategies have been accompanied by investments in transportation infrastructures and public spaces. Nevertheless, mega-events and architecture attracted more attention, and also the British regeneration policy in the 1990s paid only limited attention to transportation infrastructure (Dabinett et al., 1999).

Sustainable Mobility and Regeneration

While transportation has not always been the main focus in urban regeneration policy discourse, there have been many interactions between these two worlds. Particularly in France and some neighboring countries, the development of High-Speed Rail networks since the 1980s was linked to regeneration investments in and near stations. In line with culture-led urban regeneration strategies, several railway stations were designed by leading architects. But urban regeneration and transportation have been integrated in a more profound way than the merely architectural.

The urban regeneration discourse is a contemporary of that on sustainable mobility. Unsurprisingly, the two have important commonalities, and interacted significantly. A common theme is how pedestrian zones, bicycle infrastructure, and inner-city public transportation contribute to make city centers more attractive for residents, tourists and employees, in particular in the service industry. While in many regions, tram and railway networks had been reduced in the decades following World War II, particularly



Figure 1 With the organization of the 1992 Olympics, Barcelona became a best practice for urban regeneration strategies around the world (the Montjuïc Communications Tower designed by the famous architect Santiago Calatrava).



Figure 2 Light rail and trams have become a key ingredient in sustainable urban development programs (the Vauban district in Freiburg).

the 1990s and following decades witnessed investments in trams, light rail, tram-trains and similar systems. Often-cited examples include Strasbourg and Nantes in France, and Freiburg and Karlsruhe in Germany (Fig. 2). These examples show that sustainable modes of transportation can be (re)introduced.

Notwithstanding the possibility of reintroducing transportation modes and changes in policy, planners promote a decade-long, sustained strategy to integrate land use and transportation planning (Hickman et al., 2013). The prime example in the literature of

such a sustained strategy is the continuity between the 1947 Finger Plan in Copenhagen and more recent policies that locate urban development alongside transit lines (Knowles, 2012). The ‘fingers’ are the corridors or development axes which have their origin in the city center.

Especially in the United States, the smart growth and new urbanism movements in design have been successful advocates of urban regeneration. Transportation is a key dimension of these approaches which promote compact, walkable, mixed land use neighborhoods as an alternative to urban sprawl, in line with the 3Ds of travel demand: Density, Diversity, and Design. The first D, Density, enables efficient investments in transit as well as easy access to many activities at walking distance, as does the second D, Diversity or mixed land use. The third D, Design, is important to make the environment suitable for walking (Cervero and Kockelman, 1997). According to the website NewUrbanism.org the aim is “Giving people many choices for living an urban lifestyle in sustainable, convenient and enjoyable places, while providing the solutions to peak oil and climate change” and this is followed by the equation “Renewable Energy + Electric Transportation + Walkable Urbanism” (“New Urbanism,” n.d.). This illustrates that transportation is a key component of new urbanism and smart growth.

Transit-Oriented Development

Since the rise of urban regeneration strategies, proposals to integrate transportation and land use have matured and have inspired practitioners and researchers all over the world. Transit-Oriented Development (TOD) is presumably the most influential model in this respect. TOD encompasses that urbanization is concentrated around transit stations. The areas in the vicinity of stations are walkable neighborhoods with high densities in particular within the half mile radius of a station. TOD has its origins in North America, but has traveled from there to research institutes and planning offices in Europe, Australia, Asia, and elsewhere (Ibraeva et al., 2020).

Within the TOD approach, not all station areas need to be developed in the same way. The aim is to balance the quality of transit services and the activities in the area around the station. This idea is translated into the node-place model (Bertolini, 1999). The node value indicates how accessible a location is in transportation terms and, as a consequence, emphasizes the connectivity with other locations in terms of frequency and travel time. The place value is a measure which indicates the diversity and intensity of the activities performed in the area surrounding the node. Fig. 3 shows the different possible (im)balances between node and place value.

Experiences Around the World

Policy strategies such as TOD, culture-led urban regeneration, and compact city development travel around the world and are part of a global planning discourse. Attempts to link mega-events like the Olympic Games or soccer World Cup to urban regeneration could be found in locations as diverse as Cape Town, Rio de Janeiro, Beijing, and London. As mentioned earlier, TOD is particularly influential in the US where the concept was developed before it spread all over the world. However, every city is unique and strategies materialize differently in different locations.

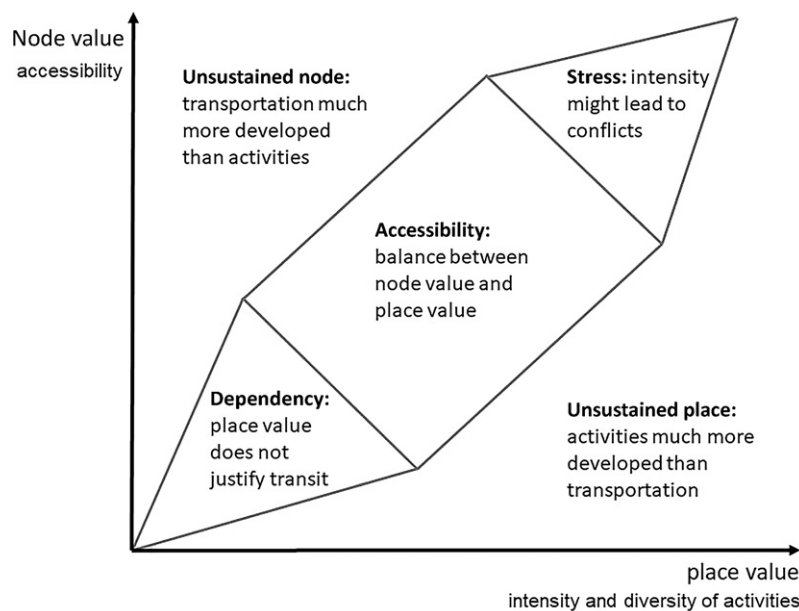


Figure 3 The node-place model. Source: Based on: Bertolini (1999), p. 202.



Figure 4 In the design of Houten, the Netherlands, bicycles have priority over other vehicles.

The expansion of High-Speed Rail and the strategy to develop stations as engines of growth is mainly observed in Western Europe, and later on, in China. Note that the 1964 Tokyo Olympics and the opening of the Tokaido Shinkansen high-speed rail connection between Tokyo and Osaka the same year predate urban regeneration in the strict sense. In many European cities considerable investments have been made in trams, tram-trains and light rail systems, but Europe has by no means a monopoly on such systems. Many examples of light rail predate urban renaissance but have been refurbished and extended since the wave of urban regeneration.

In Latin America, Bus Rapid Transit (BRT) is the most well-known vehicle for urban development. Although separate bus lanes and corridors require substantial investments, BRT is cheaper when compared to rail-based transit. Dedicated, segregated lanes, right-of-way infrastructure, and modern vehicles contribute to comfortable, fast, and frequent transportation. The rise of BRT systems runs in parallel with a more positive attitude toward cities in Latin America, which challenges the commonly made association between urban areas and crime. Examples of cities with a BRT system are Curitiba (Brazil), Bogotá and Cartagena (Colombia), Quito (Ecuador), Lima (Peru), Guadalajara (Mexico), Buenos Aires (Argentina), and Santiago (Chile). And also outside Latin America, BRT was implemented in places like Johannesburg (South Africa) and Guangzhou (China). While competitiveness generally is a leading concept in urban regeneration and transportation planning, Latin American BRT policies often emphasize equity and the role of transit in delivering accessibility to vulnerable population groups (Mejía-Dugand et al., 2013).

In the context of urban regeneration, the cornerstone of transportation planning is transit. However, station areas are dependent on the surrounding area. That is the reason why the creation of walkable neighborhoods is an inherent part of TOD and similar strategies. Apart from increases in sidewalk width and the provision of safe crossing facilities, cities have established car-free streets. This is part of what has been called the pedestrianization movement which involves the transformation of city streets into places where people can walk safely and comfortably. Organizations like Walk21 promote these measures via conferences and other activities (Walk21, 2020). Besides walking, the bicycle is promoted as a sustainable and healthy mode of transportation, but not in all locations. Topography, climate but also cultural reasons may explain why biking is less universally promoted than walking. Common bicycle-promoting measures include the provision of bike lanes, storage facilities, especially near stations, and bike share systems, which can be seen as an indicator for urban regeneration. Copenhagen and several cities in the Netherlands act as sources of inspiration for bicycle-oriented policies (Fig. 4).

Investments in transit and the promotion of walking and cycling are one side of the transportation planning coin, the other side is the discouragement of car use. This corresponds to the idea that policies need both sticks and carrots to transform urban transportation. The implementation of road pricing in the form of urban cordon charges is a “stick” to discourage car use which is intensely discussed (Eliasson, 2014). Examples of urban road pricing can be found in Singapore, Stockholm, and London. While the focus of pricing is on congestion mitigation, the aim is also to make cities more healthy and livable, which is also the aim of Low Emission Zones where the most polluting vehicles are not allowed to enter. Related measures include parking restrictions, traffic calming and banning motorized vehicles in some streets (Nieuwenhuijsen and Khreis, 2016).

Challenges and Criticisms

Emphasizing the need to integrate land use planning and transportation policy has been a constant in the literature. The associated challenge is to foster cooperation between stakeholders, in particular planning authorities, public transportation operators, but also developers. The financial dimension is of prime importance. The development of transportation infrastructure requires high initial

investments, and the same holds for land acquisition and large scale real estate developments. These investments may be offset by higher property prices near public transportation stops. Various value capture mechanisms have been proposed to ensure that benefits not only flow to developers, but also to those who invest in the infrastructures.

The effect of TOD-like as well as other types of urban regeneration strategies on real estate prices is a popular topic of study, but most attention goes to the effects on travel behavior. Besides transit use, the level of walking is a key indicator. Given the diversity of settings, densities, and contexts, it is not possible to provide universally applicable effect sizes for car use, transit patronage or the share of walking. Looking at this issue from a different angle, some actors do not classify projects as TOD when the share of sustainable travel is below, for example, 50% (Ibraeva et al., 2020). Regarding the effects, residential self-selection complicates the estimation since people with a positive attitude towards transit and walking have a preferences for living in walkable environments with good transit access (Cervero, 2007).

What TOD, BRT, compact city, sustainable mobility and culture-led and other urban regeneration strategies have in common is that such concepts have become part of an international planning discourse in which best practices from cities around the world are used to promote a better integration of transportation and land use planning. However, concerns have been raised regarding the transferability of policies originating in different contexts (Thomas et al., 2018). International policy discourses highlight in the first place factors which are easy to measure or visualize. Clear examples of easily transferable elements include population and job density, transportation technologies, and the half-mile radius around transit stations. The implementation of universally applicable models may result in cartographic reductionism and a static view of place which ignore the complexity inherent to particular places (Qviström et al., 2019). Furthermore, a restricted focus on station areas can turn the focus away from wider social effects.

A common concern with urban regeneration is gentrification. Investments in public space, housing, and transportation infrastructure generally increase the attractiveness of neighborhoods. As a consequence, rising property prices often displace the original residents, and the characteristics of the refurbished neighborhoods may no longer serve the needs of the original population. This is, for example, the case when investments in office buildings and connections to business districts are made in a largely working-class district. Urban regeneration strategies illustrate how well-intended policies to do something about urban decay, necessarily promote particular groups and ways of life (Enright, 2013). The wave of urban regeneration and TOD strategies discussed here clearly reflects the growing importance of the service sector in urban areas.

References

- Bertolini, L., 1999. Spatial development patterns and public transport: the application of an analytical model in the Netherlands. *Plan. Practice Res.* 14, 199–210, doi:10.1080/02697459915724.
- Cervero, R., 2007. Transit-oriented development's ridership bonus: a product of self-selection and public policies. *Environ. Plan A* 39, 2068–2085, doi:10.1068/a38377.
- Cervero, R., Kockelman, K., 1997. Travel demand and the 3Ds: density, diversity, and design. *Transport. Res. Part D: Transport Environ.* 2, 199–219, doi:10.1016/S1361-9209(97)00009-6.
- Dabinett, G., Gore, T., Haywood, R., Lawless, P., 1999. Transport investment and regeneration. Sheffield: 1992–1997. *Transport Policy* 6, 123–134, doi:10.1016/S0967-070X(99)00013-X.
- Eliasson, J., 2014. The Stockholm congestion pricing syndrome: How congestion charges went from unthinkable to uncontroversial. Centre for Transport Studies CTS Working Paper 2014:1, Stockholm.
- Enright, T.E., 2013. Mass transportation in the Neoliberal city: the mobilizing myths of the Grand Paris Express. *Environ. Plan A* 45, 797–813, doi:10.1068/a459.
- Furbey, R., 1999. Urban 'regeneration': reflections on a metaphor. *Critical Social Policy* 19, 419–445, doi:10.1177/026101839901900401.
- González, S., 2011. Bilbao and Barcelona 'in Motion'. How Urban Regeneration 'Models' travel and mutate in the global flows of policy tourism. *Urban Studies* 48, 1397–1418, doi:10.1177/0042098010374510.
- Hickman, R., Hall, P., Banister, D., 2013. Planning more for sustainable mobility. *J. Transport Geogr.* 33, 210–219, doi:10.1016/j.jtrangeo.2013.07.004.
- Ibraeva, A., Correia, G.H., de, A., Silva, C., Antunes, A.P., 2020. Transit-oriented development: a review of research achievements and challenges. *Transport. Res. Part A: Policy Practice* 132, 110–130, doi:10.1016/j.tra.2019.10.018.
- Knowles, R.D., 2012. Transit oriented development in Copenhagen, Denmark: from the Finger Plan to Ørestad. *J. Transport Geogr.* 22, 251–261, doi:10.1016/j.jtrangeo.2012.01.009.
- Mejía-Dugand, S., Hjelm, O., Baas, L., Ríos, R.A., 2013. Lessons from the spread of Bus Rapid Transit in Latin America. *J. Clean. Prod.* 50, 82–90, doi:10.1016/j.jclepro.2012.11.028.
- New Urbanism [WWW Document], n.d. URL <http://newurbanism.org/> (accessed 2.17.20).
- Nieuwenhuijsen, M.J., Khreis, H., 2016. Car free cities: pathway to healthy urban living. *Environ. Int.* 94, 251–262, doi:10.1016/j.envint.2016.05.032.
- Qviström, M., Luka, N., De Block, G., 2019. Beyond circular thinking: geographies of transit-oriented development. *Int. J. Urban Reg. Res.* 43, 786–793, doi:10.1111/1468-2427.12798.
- Taylor, P.J., 2004. *World City Network: A Global Urban Analysis*. Routledge, London; New York.
- Thomas, R., Pojani, D., Lenferink, S., Bertolini, L., Stead, D., van der Krabben, E., 2018. Is transit-oriented development (TOD) an internationally transferable policy concept? *Regional Stud.* 52, 1201–1213, doi:10.1080/00343404.2018.1428740.
- Walk21, 2020. Walk21 Leading the Walking Movement [WWW Document]. URL <https://www.walk21.com/> (accessed 2.28.20).

Further Reading

- Banister, D., 2005. *Unsustainable Transport: City Transport in the New Century*, 1st ed. Routledge, London; New York Transport, development and sustainability.
- Bertolini, L., 2017. *Planning the Mobile Metropolis: Transport for People, Places and the Planet*, Planning, Environment, Cities. Palgrave Macmillan, London.
- Calthorpe, P., 1993. *The Next American Metropolis: Ecology, Community, and the American Dream*. Princeton Architectural Press, New York.
- Cervero, R., Guerra, E., Al, S., 2017. *Beyond Mobility: Planning Cities for People and Places*. Island Press, Washington, DC.
- Couch, C., Fraser, C., Percy, S. (Eds.), 2003. *Urban Regeneration in Europe*. Blackwell Science Ltd, Oxford, UK. <https://doi.org/10.1002/9780470690604>.
- Hickman, R., Banister, D., 2014. *Transport, Climate Change and the City*, Routledge Advances in Climate Change Research. Routledge, Taylor & Francis Group, London; New York.
- OECD, 2018. *Rethinking Urban Sprawl: Moving Towards Sustainable Cities*. OECD. <https://doi.org/10.1787/9789264189881-en>.
- Ross, B., 2014. *Dead End: Suburban Sprawl and the Rebirth of American Urbanism*. Oxford University Press, Oxford, New York.

Social and Distributional Impact Assessment in Transport Policy

Laura Walker*, Angela Curl†, *Social and Equalities Impact Assessment Consultant (AECOM), Birmingham, United Kingdom; †University of Otago, Christchurch, New Zealand

© 2021 Elsevier Ltd. All rights reserved.

Introduction	361
Social and Distributional Impact Assessments	362
Case Study: Highways England Approach to Assessing Equality Impacts of New Major Highway Schemes	364
Practical Challenges and Guidance	364
Limitations of Existing Assessment Frameworks	365
Evidence, Evaluation, and Monitoring	365
Spatial and Temporal Coverage for Assessment	365
Engagement and Expertise	366
Conclusions	366
References	366
Further Reading	367

Introduction

The way in which we move around has profound impacts on society, the environment and the economy. Social impacts relate to any impacts that the transport system has on people. These impacts can be positive, for example providing access to essential services, social connections, providing physical activity and enhancing wellbeing. At the same time there are a broad range of negative impacts associated with mobility. Around 1.35 million people die each year as a result of road traffic injuries ([World Health Organization, 2020](#)) and many more are injured or traumatized by the impacts. Transport related emissions are a major contributor to climate change and local air quality, which both impact human health and wellbeing. Car dependence is associated with sedentary lifestyles and inactivity, contributing to rising levels of obesity and poor physical and mental wellbeing. Transport also constitutes a substantial proportion of household budgets which can be a cause of financial distress.

A broad range of transport related social impacts have been considered in academic literature and in practice. [Jones and Lucas \(2012\)](#) suggest that these fall into five broad categories:

- Accessibility (potential)
- Movement and activity (realized)
- Health related outcomes (road casualties and injuries; air quality; noise; physical activity; intrinsic value; mental health)
- Finance related (affordability)
- Community related (social interactions, personal safety & fear of crime, forced relocation).

Differences in the way different groups in society are affected by transport policies or interventions are known as distributional impacts. Considering the distributional impacts is important when considering transport equity or fairness.

In general, better off populations benefit most from current transport systems, while also inflicting most harm on others ([Brand and Preston, 2010](#)). The level of accessibility provided by the transport system and land use varies spatially and according to socio-demographic characteristics, such as income, disability, gender, age and ethnicity. Those that face difficulties accessing goods and services because of poor transport can be at risk of transport related social exclusion, with adverse consequences such as unemployment, missed appointments, poor uptake of education, difficulties in accessing healthy food and social isolation ([Lucas and Jones, 2012](#)). Negative outcomes, such as respiratory diseases, chronic health conditions and road trauma associated with the transport system are felt differently by different social groups, both within and between countries. For example, 93% of the world's road traffic fatalities occur in low or middle income countries, despite only owning 60% of vehicles in the world ([World Health Organization, 2020](#)). Poor air quality and transport crashes are more prevalent in more deprived areas ([Schweitzer and Zhou, 2010](#)). Breathing the air in the 1m above surface height exposes people to higher concentrations of air pollution from traffic meaning that children, especially those in pushchairs are particularly impacted by traffic pollution ([Sharma and Kumar, 2018](#)). Globally, indigenous populations have benefited least from the mobility provided by high carbon transport systems and are least responsible for associated emissions yet are often most vulnerable to climate change ([Markkanen and Anger-Kraavi, 2019](#)).

Transport policies and interventions have a fundamental role to play in ensuring that the transport system achieves positive outcomes for all members of society. It is important that decision makers consider how a proposed transport policy might impact people (social impacts) and how these impacts will be distributed across society (distributional impacts).

For some nations, the recognised rights and entitlements of people under international and national human rights or equalities legislation has resulted in distributional impact assessment approaches focused on those with protected status. For example, in the UK, national and local transport authorities must adhere to the Human Rights Act 1998, which incorporates articles of European

Convention of Human Rights (EHRC). The Public Sector Equality Duty (PSED) set out under section 149 of the UK Equality Act 2010 (UK Government, 2020) requires public bodies to pay due regard to impacts of policies, plans and projects with regard to groups of the population with 'protected characteristics'. These protected characteristics are age, sex, disability, religion, race, marriage/civil partnership, sexual orientation, transgender and pregnancy and maternity. In many countries the rights of indigenous peoples are recognised through the United Nations Declaration on Human Rights (United Nations, 2007). Such nations have a commitment to incorporating equity and rights of indigenous peoples in all decision making processes and that should, but often does not, extend to transport decision-making.

Social impacts have not been considered in transport appraisal to the same extent as economic or environmental impacts (Anciaes and Jones, 2020). A failure to consider social impacts has contributed to the negative impacts we see arising from the transport system, and unequal distribution of these impacts. Social and distributional impact assessments in transport planning and policy are therefore important from a social justice perspective, helping to redress existing social inequities and ensuring transport investment is fair and achieving positive social outcomes, rather than inflicting harm.

This article discusses some key principles for social assessment and an outline of social and distributional assessments for transport intervention. A Highways England case study describes how a bespoke tool is being used to capture the equality impacts of new major highways schemes. The article ends with an overview of some of the key challenges to assessing social and distributional impacts.

Social and Distributional Impact Assessments

A variety of approaches to measuring social impacts exist through national transport appraisal guidelines in many countries. However, they often do not adhere to all of the principles (Fig. 1) and are often considered an "add on" to existing cost-benefit analysis frameworks, and approached as a technical exercise. It can often be difficult to ascertain how much of a bearing social impact assessments have on decision making, relative to monetised impacts included in a cost-benefit analysis (Hickman, 2019).

Fig. 2 provides an outline process for social impact assessment within transport policy. Existing social problems and issues should be identified (Stage 1) prior to formation of relevant transport policy (Stage 2). However, in practice, social impact assessment is more often integrated into the policy process at Stage 3 (once a project or policy has been proposed and at the option appraisal stage) rather than informing new transport projects or policies.

Stage 3 focuses on assessing social impacts of the transport policy or intervention. The types of social impacts assessed varies across nations and institutions. For example, the English transport appraisal guidance, WebTAG assesses the following social impacts: physical activity, security, severance, journey quality, option values and non-use values, accessibility impacts, and affordability impacts (DfT, 2019). Australia has on guidance on qualitatively assessing impacts of transport projects on accessibility and affordability (Anciaes and Jones, 2020). Waka Kotahi - New Zealand Transport Agency has a Social Impact Guide, which identifies a number of social impacts: air quality, noise, vibration, water quality, (changes to) transport modes, social connectedness, community severance, changes to facilities, changes to local movement patterns, safety, economy or public health (Waka Kotahi NZ Transport Agency, 2016).

Principles of Social Impact Assessment

1. To ensure more sustainable and equitable development. Impact assessment should promote community development, capacity, and social capital.
2. SIA should take a proactive approach to achieving better outcomes, rather than simply seeking to mitigate negative or unintended outcomes.
3. SIA should inform the design of the policy in an adaptive fashion
4. Social, economic, and environmental impacts are connected in a complex system. Impact pathways need to be articulated and second and higher order, wider impacts, considered.
5. Evaluation is a key component so that future analyses can learn from the results of past activities. The approach is therefore reflexive, evaluative and continually developing.
6. SIA can be prospective and retrospective. For example, the social impacts of unplanned events could be analysed retrospectively.
7. Participatory processes and local knowledge should be used to analyse concerns of those affected and use stakeholder knowledge in the assessment of impacts, appraisal of alternatives and monitoring and evaluation processes

Figure 1 Principles of Social Impact Assessment based on Vanclay (2003)

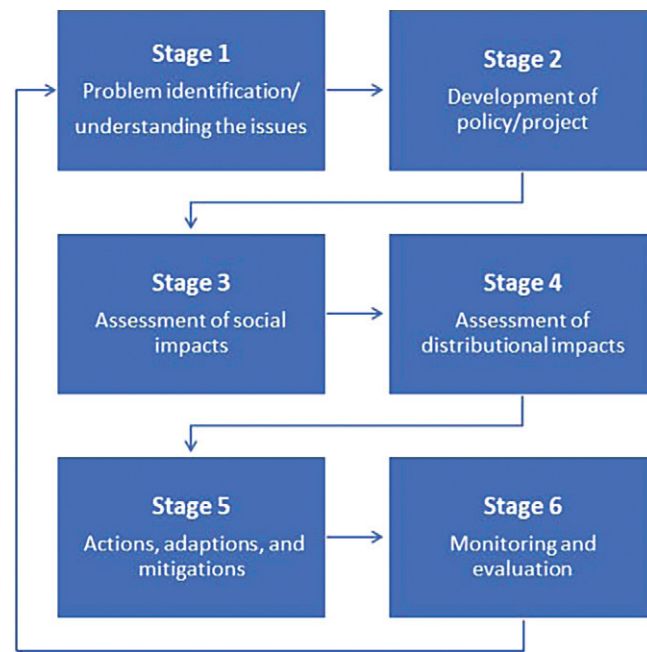


Figure 2 Typical process for Social and Distributional Impact Assessment

Stage 4 is the assessment of distributional impacts and understanding how transport policies and interventions impact different groups, including through participatory decision making processes which are a key component of social impact assessment (Vanclay, 2003). Distributional analyses consider how impacts (benefits and burdens) of transport systems are distributed among the population and whether there are inequalities among different groups of people or places. Assessments in practice and research have tended to focus on deprivation or low incomes, but also include: indigenous groups, gender, age, ethnicity, geographic location, household structure (e.g., children / single parents), household tenure, deprivation, car ownership, disability, faith, economic activity, educational qualifications, being in receipt of state benefits and indices of multiple deprivation (Lucas and Jones, 2012).

The English transport appraisal guidance (WebTAG) includes a distributional impact assessment of: user benefits; noise; air quality; accidents; severance; security; accessibility and personal affordability. Each of the social impacts noted above are rated on a 7 point scale for equity (DfT, 2015). French guidance suggests an equity index and mapping analyses and Sweden has guidelines for measuring gender equality in local transport (Anciaes and Jones, 2020).

Equality Impact Assessments (EqIAs) are frequently used as a tool by public bodies in the UK to screen, assess and monitor equality impacts, with some public bodies choosing to integrate an assessment of human rights implications into the same tool. An EqIA can provide an assessment of how transport interventions could potentially result in disproportionate or differential effects on the above protected characteristic groups. A group may be disproportionately affected by an identified impact as a consequence of a higher than average proportion of that group residing in an area. For example, severance of community in an area with a higher than average older population. A group also would be disproportionately affected should evidence show that their protected characteristic result in them being more vulnerable to the impact. For example, research shows that children are more susceptible to changes in noise level and air quality than other age groups. In other words, the uneven social outcomes as well as uneven distribution of transport related resources are important.

Scotland's Transport Appraisal Guidance (STAG) specifically advises on using the outputs of an EqIA to support the appraisal of accessibility and social inclusion impacts of transport project or policy (Transport Scotland, 2014). Furthermore, the Scottish Government are also undertaking Children's Rights and Wellbeing impact assessments on transport laws, policies and measures in line with the Children and Young People (Scotland) Act 2014. In addition to the protected characteristics groups included in the Equality Act 2010, the Fairer Scotland Duty also places a legal responsibility on particular public bodies in Scotland to pay due regard to actively consider how they can reduce inequalities of outcome, caused by socio-economic disadvantage, when making strategic decisions. Island Communities impact assessments are currently being developed to ensure that the interests of island communities are being considered as included under the Island (Scotland) Act 2018.

Stage 5 includes the development of actions for removing or minimising the negative impacts identified in assessment as well as enhancements for increasing beneficial outcomes. The actions could be wide ranging and include policy changes, partnership working with transport and non-transport service providers, environmental and construction processes and financial compensation.

Finally, **Stage 6** should involve the creation of a program to monitor impacts and the implementation of action plans, including assessment of social and distributional impacts. Evaluations should identify any further issues and inform future policy development.

Case Study: Highways England Approach to Assessing Equality Impacts of New Major Highway Schemes

Highways England (HE) is the governing organisation for England's strategic road network comprising of motorway and major highways or 'A' roads. It is responsible for the maintenance, operation and improvement of the national highways network totaling around 4300 miles (6920 km). Its aim is to ensure the road network is free flowing, safe, accessible and integrated whilst supporting the economy and resulting in long-term sustainable benefit to the environment (Highways England, 2015). It is also responsible for developing new major road schemes and seeking relevant planning consent from the UK Government for the implementation of these schemes.

HE produces EqIAs across a range of its policies and practices including the development of new major schemes as evidence of paying due regard to the Public Sector Equality Duty (PSED), despite EqIA not being a compulsory part of the transport appraisal process. As well as identifying impacts of proposed schemes on the groups with protected characteristics included within the Equality Act 2010 it also considers vulnerable and non-motorised users of the highway network, including pedestrians, cyclists, motorcyclists and equestrians. HE has developed the Equality, Diversity and Inclusion tool (EDIT) specifically to support the EqIA process for major highways schemes. The EDIT is a two-stage process designed to be used for assessing equality impacts of new major highways schemes. Stage one comprises an initial screening of potential equality impacts to determine whether a full EqIA is required. The screening process first identifies the following within an area of interest (typically one of the ten HE operating regions (UK Government, 2020)).

- the largest **numbers** of all people;
- the largest **numbers and proportions** of people from particular protected characteristic groups;
- destinations such as schools, hospitals, religious buildings and care homes; and
- concentrations of all of the above - people, equality groups and destinations.

The screening process also requires the user to determine whether a proposed scheme is likely to result in changes to the existing transport network, such as:

- pedestrian or community severance,
- access to public or community facilities or employment opportunities,
- Public transport usage or access,
- the streetscape and the pedestrian environment,
- Crossings on the highway network
- Physical accessibility; and
- Highway user experience.

Construction impacts such as temporary changes to the highway or footways, diversions and changes to routes, noise, dust and light environmental impacts, temporary construction employment and changes in access to facilities and services are also considered at this stage. Following the input of this information into the tool, an EDIT 'score' is provided and an initial EqIA screening is prepared. The tool provides a recommendation to proceed to Stage 2 of EDIT and a full EqIA as appropriate is based on the score and screening outcome.

Stage 2 of the EDIT tool captures further scheme design details including the main beneficiaries of the scheme, type of location (i. e., urban/rural) and potential impacts on nonmotorised users. Additional evidence obtained through other impact assessments (e. g., environmental impact assessments, WebTAG social and distributional impact assessments, health impact assessments and economic appraisal reports) is used to provide further information on potential adverse and beneficial effects.

The impacts of the construction phase of the scheme are also taken into consideration with regard to the environmental, health and accessibility impacts on the population. Social value outcomes associated with forecast job creation and skills legacy also forms part of the information input into the tool.

The EqIA draws upon the information gathered for EDIT to identify disproportionate and differential effects of identified potential impacts across protected characteristic groups. It also assesses feedback from public consultation and engagement with key groups as well as providing a two way input into the communication plan for the project by ensuring that consultation and engagement activities are fully inclusive.

Following the completion of the EqIA, action plans are developed and may include proposed design or policy changes and appropriate mitigation measures as well as programmes to monitor impacts over time. The EqIA and EDIT are considered to be 'live' documents recording and monitoring identified impacts and outcomes of are mitigation measured as a project progresses through different stages of development.

Practical Challenges and Guidance

The challenges to undertaking social and distributional impact assessments of transport policy and interventions are wide ranging. For the purpose of this article they are summarised under the following themes:

- Limitations of existing transport appraisal frameworks;
- Evidence, evaluation and monitoring;

- Spatial and temporal coverage of assessment; and
- Engagement and expertise.

Limitations of Existing Assessment Frameworks

As outlined in this article, there are a number of approaches to incorporating social and distributional impacts assessment in transport policy and project appraisal. However, these tend to be subsidiary, rather than driving processes. It is therefore unclear to what extent such assessments influence decisions or improve outcomes (Hickman, 2019). Relatedly, social impacts tend to be framed as negative side-effects of proposed transport policies or interventions, which need to be mitigated. Transport policy, however can help achieve social, sustainability and equity goals, if social impacts are considered from the outset.

Existing assessment frameworks have been critiqued for not keeping up with shifting policy paradigms and modern transport planning, which requires greater attention be paid to developing healthy, equitable and sustainable transport systems (Anciaes and Jones, 2020). Hickman (2019) suggested a mismatch between the social factors included in UK transport appraisal (accidents, physical activity and journey quality) and Sustainable Development Indicators (social capital, social mobility, life expectancy and housing provision).

Cost-benefit analyses are necessarily partial (Hickman, 2019) and while important social impacts may be noted they are often disregarded as unmeasurable (Anciaes and Jones, 2020) and excluded from appraisals (Searle and Legacy, 2019). Omitting social impacts can undermine the investment decisions by failing to account for the full range of impacts of a proposal (Searle and Legacy, 2019). The prioritisation of monetisable impacts, in particular journey time savings, can lead to investment decisions which prioritise road building and reduce the potential of transport investment to contribute to broader social policy and sustainability goals (Hickman, 2019). Considering how major projects perform in terms of meeting budget and time targets rather than improving communities can lead to negative outcomes such as community segregation (Mottee et al., 2020). Reliance on technical assessments and “objective” calculations means that cost benefit analyses are not seen as suitable for addressing issues related to social exclusion (Thomopoulos et al., 2009).

Cost benefit analyses are based on utilitarian principles of justice which seek to maximise aggregate welfare and do not consider the gains or losses of different groups (Thomopoulos et al., 2009). Searle and Legacy (2019) compare equity and social impacts to species extinction in that they are impossible to monetise in transport planning appraisal. Greater attention needs to be paid to aligning proposed policies or interventions with wider strategic objectives, even where impacts cannot be quantified (Hickman, 2019).

Evidence, Evaluation, and Monitoring

Social Impact Assessment is reflexive and evaluative (Vanclay, 2003) which means there is a need for good evaluation evidence to inform future practice. The difficulty in quantifying and monetising some social and equity impacts demands a stronger evidence base on the actual impacts of transport policies and interventions, to support analyses. Social impacts are part of a complex system, and understanding and articulating impact pathways, feedback loops and wider impacts, and changes in places and populations, beyond what can be accounted for in direct causal models is an important, yet challenging, part of the process (Anciaes and Jones, 2020; Vanclay, 2003). The long-term impacts and complex pathways to outcomes makes it harder to attribute social impacts to policies and interventions.

In practice, there has been limited evaluation and monitoring of transport interventions with regard to social impacts. Many evaluation frameworks are based on quantitative data collection to measure key performance indicators that can be directly attributed to interventions.

The literature focuses on three main types of transport related inequalities: transport related resources (car ownership/proximity to infrastructure); observed daily travel behavior (trip frequency/distances/travel time) and transport accessibility levels (Pereira et al., 2017). Social and equity impact assessments dominated by spatial analyses tend to assume that everyone in a particular place is impacted in the same way. However, health outcomes and inequities arising as a result of transport inequities feature less (Hosking et al., 2019). There is a need for a better understanding of differential outcomes. It may be that transport policies and interventions do not result in transport related inequities but do lead to inequitable outcomes in terms of health outcomes because people are impacted differently. There has been an increase in the use of epidemiological methods to understand the impacts of place based, built environment and infrastructure interventions on active travel. However, these have rarely focused on subgroup or equity analyses (Hosking et al., 2019).

Spatial and Temporal Coverage for Assessment

Assessments are often limited to the population in a given geographical area but impacts can be much wider, especially over longer timescales. Similarly, consideration needs to be made as to whether social impacts would be limited to the resident population of a given study area or those working in or visiting an area. Often, a detailed demographic breakdown is only available for the resident population. Census area boundaries may also provide an arbitrary boundary with limited relevance to the potential geographical impacts of a transport intervention.

Furthermore, most traditional demographic datasets provide only a static snapshot of the resident population, making it harder to assess the wider spatio-temporal travel patterns and needs of the population. The use of census and other secondary datasets can

also result in the risk of using obsolete data resulting in the omission of minority or underrepresented groups, for example where new communities have recently moved into an area.

Limited research has considered the uneven mobility offered by the public and active transport system at different times of day and the differentiated impacts based on gender, income and race (Smeds et al., 2020). Gaps in transport provision at night are particularly likely to impact low-income shift workers for example (Chandra et al., 2017). In addition to differential service availability at night time low income groups and women have been identified in the literature as facing fear-based exclusion at night time (Smeds et al., 2020).

Engagement and Expertise

Quantitative data alone does not provide a nuanced understanding of social issues and challenges affecting a community. Introducing new transport policies, plans and projects to an area without community engagement could lead to further exacerbation of transport issues and inequality. Engagement with local communities can be used to fill data gaps and gain a deeper understanding of existing issues. Standard consultation approaches are often perceived to take place once a decision has already been made, to inform rather than engage with the community. However, existing consultation processes often fail to include underrepresented groups, many of which are most vulnerable to negative social impacts leading to further potential social exclusion. There is also a need to recognise and value different types of evidence, including community experiences and understandings, and expert evaluations. Understanding social impacts is not a purely technical exercise but requires meaningful community engagement and social science expertise (Hickman, 2019). Failure to engage with social scientists results in overly technical, apparently objective assessments of social impacts, seen as a tick box exercise, which is not well suited to understanding complex social processes.

Conclusions

Assessing the social and distributional impacts of proposed transport policies and interventions is crucial to promoting healthy sustainable and equitable transport systems. Traditional transport planning appraisal systems have not adapted well to changing transport planning paradigms which require a focus on the ways in which transport systems contribute to quality of life, not just the efficiency of the system. Transport systems can contribute positively to social, equity and environmental objectives meaning that consideration of social impacts needs to take place at all stages of the policy process, and not simply be seen as a mitigation exercise and the option appraisal stage, as often is the case.

Approaches to assessing social and distributional impacts are often moulded to fit existing technical approaches. However, the political nature of decision making in transport needs more explicit consideration (Mattioli et al., 2020). A technical assessment may draw attention to differential impacts, but whether or not such distribution is fair is a political and ethical decision. A focus on equity should also take into account whether the systems and decision-making processes are fair. It may be that inequalities in transport can be considered fair, and in some cases necessary to achieve equitable outcomes, provided everyone has good levels of accessibility to essential services. Jones and Lucas (2012) recommend use of minimum standards of transport provision, equity and fairness. Ensuring that transport policies and interventions meet agreed minimum standards and maximum harms could help address some of the challenges of detailed social and distributional impact assessments outlined in this article. More fundamentally, it is important to remove systematic marginalisation and discrimination, and unfair differences in outcomes or exposures.

Case study information produced with kind permission from Highways England

References

- Anciaes, P., Jones, P., 2020. Transport policy for liveability – Valuing the impacts on movement, place, and society. *Transport. Res. Part A: Policy Pract.* 132, 157–173.
- Brand, C., Preston, J.M., 2010. '60-20 emission'—The unequal distribution of greenhouse gas emissions from personal, non-business travel in the UK. *Transport Policy* 17, 9–19.
- Chandra, S., Jimenez, J., Radhakrishnan, R., 2017. Accessibility evaluations for nighttime walking and bicycling for low-income shift workers. *J. Transport Geogr.* 64, 97–108.
- DfT, 2015. TAG unit A4-2 distributional impact appraisal. London.
- DfT, 2019. In: TAG unit A4-1 social impact appraisal. London.
- Hickman, R., 2019. The gentle tyranny of cost-benefit analysis in transport appraisal. *Transport Matters: Why transport matters and how we can make it better* 131.
- Hosking, J., Macmillan, A., Jones, R., Ameratunga, S., Woodward, A., 2019. Searching for health equity: validation of a search filter for ethnic and socioeconomic inequalities in transport. *Syst. Rev.* 8, 94.
- Jones, P., Lucas, K., 2012. The social consequences of transport decision-making: clarifying concepts, synthesising knowledge and assessing implications. *J. Transport Geogr.* 21, 4–16.
- Lucas, K., Jones, P., 2012. Social impacts and equity issues in transport: an introduction. *J. Transport Geogr.* 21, 1–3.
- Markkanen, S., Anger-Kraavi, A., 2019. Social impacts of climate change mitigation policies and their implications for inequality. *Climate Policy* 19, 827–844.
- Mattioli, G., Roberts, C., Steinberger, J.K., Brown, A., 2020. The political economy of car dependence: a systems of provision approach. *Energy Res. Social Sci.* 66, 101486.
- Mottee, L.K., Arts, J., Vancly, F., Miller, F., Howitt, R., 2020. Reflecting on How Social Impacts are Considered in Transport Infrastructure Project Planning: Looking beyond the Claimed Success of Sydney's South West Rail Link. *Urban Policy Res.* 1–14.
- Pereira, R.H.M., Schwanen, T., Banister, D., 2017. Distributive justice and equity in transportation. *Transport Rev.* 37, 170–191.
- Schweitzer, L., Zhou, J., 2010. Neighborhood air quality, respiratory health, and vulnerable populations in compact and sprawled regions. *J. Am. Plan. Assoc.* 76, 363–371.
- Searle, G., Legacy, C., 2019. Australian mega transport business cases: missing costs and benefits. *Urban Policy Res.* 37, 458–473.
- Sharma, A., Kumar, P., 2018. A review of factors surrounding the air pollution exposure to in-pram babies and mitigation strategies. *Environ. Int.* 120, 262–278.

- Smeds, E., Robin, E., McArthur, J., 2020. Night-time mobilities and (in)justice in London: Constructing mobile subjects and the politics of difference in policy-making. *J. Transport Geogr.* 82,, 102569.
- Thomopoulos, N., Grant-Muller, S., Tight, M.R., 2009. Incorporating equity considerations in transport infrastructure evaluation: current practice and a proposed methodology. *Eval. Program Plan.* 32, 351–359.
- Transport Scotland, 2014. STAG Technical Database: Accessibility and Social Inclusion.
- UK Government, 2020. Roads Managed by Highways England.
- United Nations, 2007. United Nations Declaration on the Rights of Indigenous Peoples (DRIP) UN GA Resolution 61/295.
- Vancly, F., 2003. International principles for social impact assessment. *Impact Assess. Project Apprais.* 21, 5–12.
- Waka Kotahi NZ Transport Agency, 2016. Social Impact Guide. People Place and Environment. Wellington.
- World Health Organization, 2020. Road Traffic Injuries. Fact Sheets.

Further Reading

- AEA Group, 2011. Knowledge Review of the Social and Distributional Impacts of Department for Transport Climate Change Policy Options.
- Lucas, K., Pangbourne, K., 2014. Assessing the equity of carbon mitigation policies for transport in Scotland. *Case Stud. Transport Policy* 2, 70–80.
- Social Exclusion Unit, 2003. Making the Connections: Final Report on Transport and Social Exclusion.

Mobility Planning for Healthy Cities

Ersilia Verlinghieri, University of Oxford, Oxford; University of Westminster, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	368
How Current Transport Systems Affect Public Health	368
How Current Transport Systems Affect Human Wellbeing	369
Planning Transport: a Policy Approach to a Complex Problem	369
A Matter of Politics and Power	370
Planning in the Public Domain	371
Conclusion	372
References	372
Further Reading	373

Introduction

Building healthier cities has been a core objective of state planning since Victorian times. With time, and with the entrance of a variety of actors and interests in shaping planning, this holistic principle has been over-ridden by other focuses. The attention has shifted from planning as a state-directed process to implement a clear societal objective (health, justice), to the means and processes to potentially achieve those, with increasing attention given to ideas of growth or innovation (Barton et al., 2015; Ferreira et al., 2020).

The COVID-19 pandemic has demonstrated how this can be quickly reverted, with public health becoming an incredibly powerful reason for implementing radical and rapid changes in our urban environments. This experience, combined with increasing awareness in the public domain of the damaging health effects of past planning choices, can support the implementation of key shifts also in the way transport systems are planned. Transport plays in fact a key role both in the reproduction of damaging health impacts and in the development of new urban environments in which health is at the centre.

This article explores how transport planning can contribute to improving human health and wellbeing. It will explore the negative health outcomes linked to oil-based and motorized transport systems to then discuss and consider the way transport and mobility can be linked to societal wellbeing. Subsequently, potential policy approaches to plan mobility for healthy cities, considering both their advantages and limitations, will be considered. In this, an analysis will be developed to consider the political aspects of transport and mobility planning, highlighting the importance of an approach that considers issues of power and economy and that gives voice to different planning actors, including citizens.

How Current Transport Systems Affect Public Health

There is nowadays substantive evidence of the negative impacts that transport can have on public health. These negative impacts are mainly linked to the widespread use of motorized and oil-based transport modes, especially the private car (and, increasingly, the airplane). Car dependency produces negative health impacts both directly, affecting individual health, and, indirectly, determining changes in behaviors and the environment (Khreis et al., 2016).

In the first place, transport systems designed around the private car, with their high consumption of public space and high speeds, are substantially responsible for an enormous number of deaths worldwide, which has been globally accepted in a “rather unquestioning manner” (Wells and Beynon, 2011, p. 2494). The WHO (2018) reports that approximately 1.35 million people die each year worldwide as a result of road traffic crashes, with road traffic being the leading cause of death for people aged 5–29. These crashes affect disproportionately noncar users, showing that fundamental justice issues permeate the way pedestrians and cyclists suffer the most from someone else’s use of a motor vehicle (Culver, 2018). Research has shown how this effect is sharpened in areas of higher deprivation or in poorer countries where traffic incidents’ victims are often young adults from less wealthy groups (Culver, 2018; Wells and Beynon, 2011).

Secondly, the high number of road vehicles generates substantial changes to the environment, which impact human health. Air pollution is recognized to be main cause of vast negative health effects such as cardio-vascular, cerebral-vascular and respiratory mortality and morbidity, decreased lung function in children, diabetes and, recently, coronavirus fatalities. Motorized transport constitutes a key cause of urban air pollution: nitrogen oxides, carbon monoxide, and volatile organic compounds are released in great quantities both via exhaust emissions, whilst particulate matter is released as a consequence of wear and tear phenomena (e.g., tyres and brakes) (Khreis et al., 2016). Air pollution, similarly to what happens for traffic incidents, is also disproportionately felt by minorities and vulnerable groups across cities worldwide. For example, research has shown that in the USA Black and Hispanic Americans are disproportionately affected by air pollution mostly produced by white Americans consumption (Tessum et al., 2019). Similarly, in the UK, low-income households have the highest levels of exposure (Barnes et al., 2019).

Together with augmenting air pollution, motorized transport is responsible for other changes to the environment that is detrimental to human health. High noise levels, local temperature raises, reduction of green space, and biodiversity loss have vast and lasting effects on physical and mental health, negatively impacting human life worldwide (Khreis et al., 2016). At the same time, motorized transport is the second largest and still growing contributor to global greenhouse gas emissions and especially carbon dioxide (CO₂), the backbone of climate change, and therefore is directly responsible for warmer weather and extreme weather events (Chapman, 2007). Those are reported to have substantial health impacts for all and especially most vulnerable and less mobile individuals. Climate change is likely to increase thermal-related mortality, water/food/vector-borne diseases, skin cancers, malnutrition, respiratory diseases linked to augmented air pollution, as well as direct increase in injury, and deaths as effect of extreme climate events (Nichols et al., 2009). Moreover, extreme climate event, as well as long-term environmental changes such as droughts, will generate important mental health effects, damage livelihoods of already vulnerable population with resulting lack of access to land or clean water, forced mass migrations, and broadening of societal inequalities (UNDP, 2007).

Finally, a transport system relying primarily on motorized transport promotes a sedentary lifestyle which is also a main causes of cardiovascular diseases, cerebrovascular disease, cancer (colon, breast, and lung), type 2 diabetes, dementia, anxiety, depression, obesity (Khreis et al., 2016). A less car-centered transport could instead promote the adoption of active forms of transport that can answer to the damaging health impacts of sedentary lifestyle. Research is increasingly showing how the benefits of active travel are at least one order of magnitude greater than the damages suffered from air pollution, so even cycling in much polluted cities will provide important health benefits! (Tainio et al., 2016).

How Current Transport Systems Affect Human Wellbeing

Together with directly impacting individual health, motorized transport affects other aspects of human life that are core to achieving wellbeing for all. These become especially evident when we adopt a definition of wellbeing which links to the Aristotelian tradition of wellbeing as “eudamonia”, that is the realization of one’s true potential and autonomy.

Transport has a direct responsibility in allowing people to access jobs, education, health, social networks, and the concept of accessibility has become key in a wide range of studies in transport geography (Litman, 2007). At the same time, the way we are able to travel or move and the way we experience it—trapped in a crammed bus in a very hot day, or smoothly driving our car on a quiet road—shapes our mood (Avila-Palencia et al., 2018), and also the meaning we give to places, our identity, and social relations (Doody, 2020). Transport affects also wellbeing at a more-than-individual level; there are strong interlinks between each individuals’ mobility practices and the fact that movements and the possibility for movements are collectively determined, both as a mother drives her kids to school or as a bus stops the flow of car traffic, or a ban on informal transport hinders access to livelihood for a working parent. A malfunctioning transport system, which might force some groups to rely on very low quality and unsafe transport means, or even not have access to any, can generate key societal inequalities and crucially impact the wellbeing of wide groups of the population. Malfunctioning transport systems are linked to phenomena of community severance, loss of social network, and most importantly, social exclusion and poverty.

Transport and mobility scholars have been examining these links and have highlighted how, in a system designed around the private car, uneven access to it, often or coupled with very limited availability of alternatives (both because of affordability and reach of services) generates phenomena of “transport poverty”, which includes issues related to lack of mobility, accessibility, and affordability of transport (Lucas et al., 2016). Transport poverty, which results in forced immobility or high part of one’s income spent to allow mobility, heavily impacts people’s ability to live a fulfilling and meaningful life. Research has shown also that such effects are particularly felt by those groups who are already disproportionately impacted by transport-related air pollution, even though they emit the least (Jephcote et al., 2016). The effects of motorized transport are in fact strongly unequally distributed with relation to income, gender, race, age, sexual orientation, and their intersection, and, as highlighted in relation to climate change effects, are strongly impacting poorer populations in nonwestern countries.

Planning Transport: a Policy Approach to a Complex Problem

Popular substantial pathways to tackle these impacts are suggested in terms of both *technological improvements* (adoption of greener technologies for motorized transports which reduce air pollution and adoption of smart technologies to increase traffic safety), *modal shifts* with promotion of active travel to solve both air pollution issues and damages linked to sedentary lifestyles, *land use changes and other urban policy measures* which would reduce number of trips and cars, need for travel, and would allow more green space or walkable environments (see more in Nieuwenhuijsen, 2020).

In support, further actions are now being coordinated across academia and the policy realm, often promoted by the WHO, with the aim to, for example, develop strategies to promote transport safety reducing car accidents and enforcing safety standards (World Health Organization, 2017b), fully assess health impacts of different transport interventions (World Health Organization, 2017a), and increase awareness of the links between transport-related air pollution and climate change effects. All these efforts recognize that addressing the complicated links between transport and health is a complex matter which requires coordination at different levels (Nieuwenhuijsen, 2020).

Firstly, transport planning is a complex subject which intersects with several domains of research and knowledge. Although traditionally addressed as part of an engineering problem, and lately within frameworks of classical economics, it is increasingly evident that *interdisciplinary efforts* are required to understand what transport and mobilities are about, their links to wellbeing, and which directions of change are possible. Increasingly, methodologies and theories from epidemiology, health science, social science, anthropology, geography, but also artificial intelligence, psychology, have been able to create a more adequate understanding of transport systems, in their being an intricate “ensembles of infrastructures, technologies, practices, regulation, meanings, values, affects enabling, and shaping the (non)movements through physical space of people, goods, and waste” (Schwanen, 2019, np). Amongst other results of these efforts, it is worth highlighting the development of sophisticated forms of Health Impact Assessment (often including participatory aspects) that are able to provide a strong narrative for policy makers evaluating alternative options for future transport scenarios (Nieuwenhuijsen et al., 2017).

Secondly, given that transport crosses several aspects of planning and policy, *integration at different levels of policy making* is a crucial requirement to be able to implement meaningful changes. In this regard, cities are increasingly becoming protagonists in coordinating efforts at different levels of local governance, for example by adopting Sustainable Urban Mobility Plans, or coordinating at a regional, national and international scale with new emerging networks and exchange exercises such as the World Health Organization’s (WHO) Healthy Cities. They have also opened substantially to the intervention of other actors, from universities and NGOs, to partnerships with private stakeholders and consultants to design different solutions. They have supported the emergence of Urban Living Labs as experiments of different forms of mobility (Voytenko et al., 2016), or adopted new forms of governance embracing the idea of transition management (Loorbach, 2010).

Thirdly, there is a growing awareness of the importance of *integrating health objective with sustainability objective*. Efforts towards improving the uptake of active travel or reducing overall the need for travel are crucial in this sense. Similarly, technologies which increase engine efficiency, or which allow citizens to approach “Mobility as a Service” can be of key importance. Amongst others, the WHO THE-PEP (WHO Transport, Health and Environment Pan European Programme) include important guidelines for working to integrate transport planning with health and environment.

These strategies, often combined, are allowing several (Western!) cities to start adopting healthier transport solutions, reducing carbon emission and fuelling the adoption of active travels. Key recent examples include Barcelona, with the innovative idea of superblocks, localized pedestrian areas at the level of a street block (Rueda, 2018), Paris, which has adopted the idea of a 15 min cities, boosting the adoption of walking and cycling, and Milan, equally promoting an idea of a polycentric and hierarchical transport system, especially in response to COVID-19 (Pisano, 2020).

However, there are still broad questions linked to the possibility of these strategies to last and to be implemented in other geographies. Most importantly, attention should be paid also to the shortcomings of certain planning and policy choices, especially if we adopt wellbeing and not a restricted idea of health as “being free from harm” as a definition of the final aim of planning.

A Matter of Politics and Power

The aforementioned pathways to implement change come with a series of shortcomings.

First, a pure focus on *infrastructural and technological change* is unable to provide the required change and transport planning too often seems to remain trapped in a fascination with innovation. As analyzed by Reigner and Brenac (2019), for example, the current emphasis on automation as a panacea to traffic incidents is showing to be distracting planners from other most suitable options, such as investing in public transportation or active mobility. A focus on automated driving is pushing forward a narrow conception of health as simply “being free from harm” and therefore missing the option of reforming transport systems via also improving urban environments, social life (e.g., by in promoting cycling or walking, or land use changes) and therefore wellbeing. Similarly, limitations are also linked to an obsession with *mode change* that does not account for the problematic nature of considering modal choices as a purely individual and rational decision (Ferreira et al., 2020), and promotes instead paternalistic solutions aimed at “educating citizens”. As stressed also by Reigner and Brenac (2019), the “individualization” of health, both in terms of who is responsible for it—the single rational individual—and at which level of intervention health issues ought to be solved—by primarily modifying individual behaviors and choices—is effectively missing to account for the interconnected nature of mobilities and wellbeing. For example, road accidents are often considered being consequence of individual misbehavior, whilst instead research is increasingly showing that they are most likely linked to the way the transport system itself is designed.

Second, a *holistic and interdisciplinary* approach to transport planning is often hindered by issues of translation between different disciplinary silos, which have their own specific understanding of which knowledge and language is more authoritative. This is evident in the tendency to prioritize quantitative methods and accounts over more qualitative methodologies, although the latter have demonstrated to contribute substantially to widen the understanding of the relationship between transport and wellbeing. Although there is a call for breaking these silos, obstacles remain given the issues with translation between departments and people embedded in different disciplinary traditions. Moreover, there are key issues in translating academic knowledge into a language and format accessible to policy makers. These obstacles further hinder the creation of a holistic way of transforming transport systems and are particularly challenging to address given also the unequal distribution of centers of knowledge production around transport which prioritizes knowledge coming from Western, English speaking countries.

Third, there are issues linked to the “*one size fits all*” idea which often is embedded in knowledge-exchange exercises, such as the international frameworks mentioned earlier, or promoted by consultants who are increasingly popular in transport planning.

Implementing effectively “best practices” ideas imported from other places and contexts reveals to be most of the times highly problematic. For example, authors have shown how even extremely flexible frameworks such as the idea of Bus Rapid Transit, promoted strongly by the World Bank to solve transport issues in developing countries, can create great issues when moved from one context to the other. A project that does not acknowledge the intricate local political economy of other transport and mobility systems in place can fail both in terms of meeting the needs of local populations or in being correctly implemented (Rizzo, 2015). Similar issues can also emerge when trying to pedestrianize a local area without accounting for the time that a specific change in land use might take or broader political conflicts around local authority (as happened e.g., also in the superblock case of Barcelona, see Scudellari et al., 2019).

Fourth, even when success stories are presented, there is a key duty in trying to understand *by whom* these stories are told and *who* is effectively benefitting from the success. In the complex world of contemporary cities, policies and transformations would always be contested and controversial and, sadly, most of the time benefits who is already better off. Even designing and implementing interventions able to truly respond to both the need to reduce carbon emissions and, at the same time, care for everyone’s health and wellbeing can pose several challenges especially in terms of equity. For example, amongst several issues that planners ought to consider is the phenomena called “green gentrification”: the increasing costs of living (and moving to and from an area) and subsequent displacement of most vulnerable residents as consequence of greening policies and urban renewal, often including pedestrianizations, cycle lanes and new urban parks (Gould and Lewis, 2016). Equity analysis or, even more in depth, looking at whose capabilities and needs are enabled or hindered by new urban or transport policies, is therefore a duty to ensure that healthy cities are as such for all.

Overall, it is clear that far from being a technical exercise, transport planning is a highly political matter, with research increasingly demonstrating the importance of analyzing the role of different actors and powers in shaping the way choices are made and implemented. Above all, it is evident the predominance and persistence of the automobile in today’s transport systems is far from being merely a result of policy choices. As shown by Mattioli (2020) in a key paper recently published, the key role given to the automobile in western (and increasingly nonwestern) mobility systems is the result of a number of interests that different powerful actors have been able to impose on the local, regional, national, and global agenda. Changing this trend cannot be achieved just by coordinated policy efforts.

Planning in the Public Domain

In 1987, Friedmann published a key text in which the idea of what planning is was explored in detail. Although much time has passed, the book is still insightful. It proposes to consider planning as a process of translating knowledge into action in the public domain, which is not only enacted by policy makers, but performed as an activity in which different actors, actions, and epistemologies converge. In this way Friedman includes urban social movements and other grassroots actors in the definition of planning, recognizing that “oppositional movements are essential to a healthy society, [and] point to the possibility of a fuller humanity” (298).

As other authors have highlighted (see e.g., Schwanen and Nixon, 2020), opening up planning to more participatory forms of decision making and urban transformation is particularly important, especially when what we are looking for is the redefinition of the paradigm of transport planning away from an idea of speed and efficiency (or profit) and toward health and wellbeing. It is also crucial when we want to re-politicize the idea of planning and understand urban change as composed as a myriad of interests and needs who are often conflictual. This opening, however, is different from simply and uncritically opening up the “public domain” to private actors and stakeholders who might push for shaping urban development in favor of specific economic interests (as often is the case with the increasing popular use of Public Private Partnerships which often results in privatization of public assets (Mirafteb, 2004)).

Opening up planning to participatory forms of decision making is a process through which actors such as local residents, citizen groups and especially vulnerable or hard to reach groups can have a voice in shaping voicing their needs and linking them to the direction of urban development. This can be done by complementing usual planning methods with more in depth connections with the lives and stories of who would be impacted by these changes, for example, via using ethnography, diaries, open meetings, and narrations. These data and opening can help tailoring responses to what is most appropriate to the content, in terms of what is culturally feasible, integrated with the local landscape and economies, and responding to the needs of the most vulnerable in the first place (see Fig. 1).

At the same time, such opening can be achieved by linking with the critical work of social movements that have historically reported, analyzed and helped tackling several issues in terms of urban equity and transport justice (Sheller, 2018). Amongst several other examples, it is of great inspiration the work of the Untokening, a USA “multiracial collective that centers they lived experiences of marginalized communities to address mobility justice and equity” (<http://www.untokening.org/summary>). By using critical race theory, the collective is helping planners rethinking the politics of public transport, walking and cycling, and building mobility justice. More simply, it is cycling activists that have played a crucial role in pushing the bike revolution in the Netherlands (Feddes, 2019) and that are now sustaining daily the implementation of Low Traffic Neighbors and walking and cycling environment in the UK, Europe, and elsewhere, often supporting in a unique manner the adoption of active travel in more disadvantaged areas (Schwanen and Nixon, 2020).



Figure 1 Regeneration plan for Kampala. An exemplification of overlapping preexisting solutions to different contexts without accounting for the culture, environments, needs of the local population https://issuu.com/gcoffeng/docs/igg021_smart_moving_kampala_for_pri.

Conclusion

In this article, the links between transport planning, public health, and human wellbeing have been explored, highlighting the current efforts that policy makers, planners, and international institutions are promoting. The article has highlighted the potentials of an interdisciplinary, coordinated approach that contextualizes health in the broader picture of transport and climate change. At the same time, the limitations that these approaches might have when not accounting for the contested and political nature of transport planning are exposed. The proposal is made to open up further to a process of politicizing the debate including marginalized voices and grassroots actors in the planning sphere.

What is suggested is a very difficult journey that might need a shift in the way resources for planning are considered and distributed, combining the reliance on fairly low-cost technologies and infrastructures (such as cycle ways and pedestrianizations) with more resources invested for public engagement and interdisciplinary conversations. Most importantly, intellectual resources should be invested in shifting the paradigm of planning toward opening up to new voices and discussion, destabilizing the predominance of the idea of speed and efficiency (and profit) as drivers of urban development and promoting planning as a complex, interdisciplinary, adaptive and multiscale and dynamic search for localized, and nonnormative solutions to local needs.

References

- Avila-Palencia, I., Int Panis, L., Dons, E., Gaupp-Berghausen, M., Raser, E., Götschi, T., Gerike, R., Brand, C., De Nazelle, A., Orjuela, J.P., Anaya-Boig, E., Stigell, E., Kahlmeier, S., Iacorossi, F., Nieuwenhuijsen, M.J., 2018. The effects of transport mode use on self-perceived health, mental health, and social contact measures: a cross-sectional and longitudinal study. *Environ. Int.* 120, 199–206.
- Barnes, J.H., Chatterton, T.J., Longhurst, J.W.S., 2019. Emissions vs. exposure: Increasing injustice from road traffic-related air pollution in the United Kingdom. *Transp. Res. D Transp. Environ.* 73, 56–66.
- Barton, H., Thompson, S., Burgess, S., Grant, M., 2015. *The Routledge Handbook of Planning for Health and Well-Being*. Routledge, Abingdon, Oxon, New York, NY.
- Chapman, L., 2007. Transport and climate change: a review. *J. Transp. Geogr.* 15, 354–367, doi:10.1016/j.jtrangeo.2006.11.008.
- Culver, G., 2018. Death and the Car: On (Auto) Mobility, Violence, and Injustice. *ACME* 17, 144–170.
- Doody, B.J., 2020. Becoming 'a Londoner': Migrants' experiences and habits of everyday (im)mobilities over the life course. *J. Transp. Geogr.* 82, 102572.
- Feddes, F., 2019. *Bike City Amsterdam: How Amsterdam became the cycling capital of the world*. Bas Lubberhuizen.
- Ferreira, A., von Schönfeld, K.C., Tan, W., Papa, E., 2020. Maladaptive planning and the pro-innovation bias: considering the case of automated vehicles. *Urban Sci.* 4, 41.
- Friedmann, J., 1987. *Planning in the Public Domain: From Knowledge to Action*. Princeton University Press, Princeton.
- Gould, K.A., Lewis, T.L., 2016. *Green Gentrification: Urban Sustainability and the Struggle for Environmental Justice*. Routledge.
- Jephcote, C., Chen, H., Ropkins, K., 2016. Implementation of the Polluter-Pays Principle (PPP) in local transport policy. *J. Transp. Geogr.* 55, 58–71, doi:10.1016/j.jtrangeo.2016.06.017.
- Khreis, H., Warsow, K.M., Verlinghieri, E., Guzman, A., Pellecuer, L., Ferreira, A., Jones, I., Heinen, E., Rojas-Rueda, D., Mueller, N., Schepers, P., Lucas, K., Nieuwenhuijsen, M., 2016. The health impacts of traffic-related exposures in urban areas: understanding real effects, underlying driving forces and co-producing future directions. *J. Transp. Health.* 3, 249–267, doi:10.1016/j.jth.2016.07.002.
- Litman, T., 2007. *Evaluating Accessibility for Transportation Planning*. Victoria Transport Policy Institute.
- Loorbach, D., 2010. Transition management for sustainable development: a prescriptive. *Complexity-Based Governance Framework*. *Governance* 23, 161–183, doi:10.1111/j.1468-0491.2009.01471.x.
- Lucas, K., Mattioli, G., Verlinghieri, E., Guzman, A., 2016. Transport poverty and its adverse social consequences. *Proc. Inst. Civil Eng. – Transp.* 169, 353–365, doi:10.1680/jtran.15.00073.
- Mattioli, G., Roberts, C., Steinberger, J.K., Brown, A., 2020. The political economy of car dependence: A systems of provision approach. *Energy Res. Soc. Sci.* 66, 101486.
- Mirafteb, F., 2004. Public-private partnerships: The trojan horse of neoliberal development? *J. Plan. Edu. Res.* 24, 89–101, doi:10.1177/0739456X04267173.
- Nichols, A., Maynard, V., Goodman, B., Richardson, J., 2009. Health climate change and sustainability: A systematic review and thematic analysis of the literature. *Environ. Health. Insights* 3, doi:10.4137/EHI.S3003, EHI.S3003.

- Nieuwenhuijsen, M.J., 2020. Urban and transport planning pathways to carbon neutral, liveable and healthy cities: a review of the current evidence. *Environ. Int.* 140, 105661, doi:10.1016/j.envint.2020.105661.
- Nieuwenhuijsen, M.J., Khreis, H., Verlinghieri, E., Mueller, N., Rojas-Rueda, D., 2017. Participatory quantitative health impact assessment of urban and transport planning in cities: a review and research needs. *Environ. Int.* 103, 61–72, doi:10.1016/j.envint.2017.03.022.
- Pisano, C., 2020. Strategies for post-covid cities: An insight to Paris En commun and Milano 2020. *Sustainability* 12, 5883, doi:10.3390/su12155883.
- Reigner, H., Brenac, T., 2019. Safe, sustainable . . . but depoliticized and uneven – A critical view of urban transport policies in France. *Transp. Res. Part A: Policy. Pract.* 121, 218–234, doi:10.1016/j.tra.2019.01.023.
- Rizzo, M., 2015. The political economy of an urban megaproject: The Bus Rapid Transit project in Tanzania. *Afr. Aff. (Lond)* 114, 249–270, doi:10.1093/afraf/adu084.
- Rueda, S., 2018. Superblocks for the design of new cities and renovation of existing ones Barcelona's Case. In: Nieuwenhuijsen, M., Khreis, H. (Eds.), *Integrating Human Health into Urban and Transport Planning*, Springer International Publishing.
- Scudellari, J., Staricco, L., Vitale Brovarone, E., 2019. Implementing the Supermanzana approach in Barcelona. Critical issues at local and urban level. *Journal of Urban Design* 1–22. <https://doi.org/10.1080/13574809.2019.1625706>
- Sheller, M., 2018. *Mobility Justice*. Verso London, Brooklyn, NY.
- Tessum, C.W., Apte, J.S., Goodkind, A.L., Muller, N.Z., Mullins, K.A., Paoletta, D.A., Polasky, S., Springer, N.P., Thakrar, S.K., Marshall, J.D., Hill, J.D., 2019. Inequity in consumption of goods and services adds to racial-ethnic disparities in air pollution exposure. *Proc. Natl. Acad. Sci.* 116 (13), 6001–6006.
- Tainio, M., de Nazelle, A.J., Götschi, T., Kahlmeier, S., Rojas-Rueda, D., Nieuwenhuijsen, M.J., de Sá, T.H., Kelly, P., Woodcock, J., 2016. Can air pollution negate the health benefits of cycling and walking? *Prevent. Med.* 87, 233–236, doi:10.1016/j.ypmed.2016.02.002.
- UNDP, 2007. *Making Globalization Work for All*. United Nations Development Programme.
- Uteng, T.P., Lucas, K., 2017. *Urban Mobilities in the Global South*. Routledge.
- Voytenko, Y., McCormick, K., Evans, J., Schliwa, G., 2016. Urban living labs for sustainability and low carbon cities in Europe: towards a research agenda. *J. Clean. Prod. Adv. Sustain. Sol.* 123, 45–54, doi:10.1016/j.jclepro.2015.08.053.
- Wells, P., Beynon, M.J., 2011. Corruption, automobility cultures and road traffic deaths: The perfect storm in rapidly motorizing countries? *Environ Plan A* 43, 2492–2503, doi:10.1068/a4498.
- World Health Organization, 2017a. Health economic assessment tool (HEAT) for walking and for cycling. [WWW Document]. URL <https://www.euro.who.int/en/health-topics/environment-and-health/Transport-and-health/publications/2017/health-economic-assessment-tool-heat-for-walking-and-for-cycling.-methods-and-user-guide-on-physical-activity,-air-pollution,-injuries-and-carbon-impact-assessments-2017>(accessed 10.2.20).
- World Health Organization, 2017b. *Save LIVES: a Road Safety Technical Package*. World Health Organization, Geneva.

Further Reading

- Cole, H.V.S., Anguelovski, I., Baró, F., García-Lamarca, M., Kotsila, P., Pulgar, C.P., del Shokry, G., Triguero-Mas, M., 2020. The COVID-19 pandemic: power and privilege, gentrification, and urban environmental justice in the global north. *Cities Health* 0, 1–5, doi:10.1080/23748834.2020.1785176.
- Enright, T., 2019. Transit justice as spatial justice: learning from activists. *Mobilities* 14, 665–680, doi:10.1080/17450101.2019.1607156.
- Khreis, H., May, A.D., Nieuwenhuijsen, M.J., 2017. Health impacts of urban transport policy measures: a guidance note for practice. *J. Transp. Health.* 6, 209–227, doi:10.1016/j.jth.2017.06.003.
- Schwanen, T., 2019. Transport geography, climate change and space: opportunity for new thinking. *J. Transport Geogr.* 81, 102530, doi:10.1016/j.jtrangeo.2019.102530.
- Schwanen, T., Nixon, D.V., 2020. Understanding the relationships between wellbeing and mobility in the unequal city: The case of community initiatives promoting cycling and walking in São Paulo and London. In: Keith, de Souza Santos, (Eds.), *Urban Transformations and Public Health in the Emergent City*. Manchester University Press, pp. 79–101.

Transport Demand Management

Begoña Guirao, Department of Transport Engineering, Regional and Urban Studies. Universidad Politécnica de Madrid, Madrid, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	374
The Origin and Evolution of Transport Demand Management (TDM)	374
A Disruptive Tool for Transport Management: Mobility as a Service (MaaS)	376
Threats of the MaaS Model and Need of New TDM Strategies	377
Planning TDM Strategies for a Transition Time	378
Conclusion	378
References	378

Introduction

Transport Demand Management (TDM) has traditionally comprised all mobility strategies with the aim to reduce private car use and promote public transport in big cities. The effectiveness of TDM strategies concerned the improvement of environmental and sustainable levels of the entire transport system. During the last decades, political pressure to improve air quality indicators in cities has increased and added more to the value of TDM strategies. The United Nations emphasized this issue in 2016, when more than 91% of urban dwellers were breathing unsafe air, with the result of 4.2 million deaths reported to be related to ambient air pollution (United Nations, 2018). Additionally, more than half of the global urban population is exposed to air pollution levels at least 2.5 times higher than the safety standards, with transport pollution being a major contributor to these figures (United Nations, 2018).

The resolution adopted by the United Nations General Assembly on 25 September 2015 defined the 2030 Agenda for Sustainable Development and one of the goals, the 11th Sustainable Development Goal was directly related to making cities and human settlements inclusive, safe, resilient, and sustainable (United Nations, 2015). The role of TDM strategies to achieve the 11th goal commands greater importance, and demands effective and quick solutions. Reduction of air pollution and consumption of fuel are not only the main objectives of new TDM strategies but also the reduction of noise and vibration, as well as accidents or community severance. Fortunately, new ICTs, the wide spread of smartphone usage, the innovation in vehicle manufacturing (electric and autonomous) and the emergence of the sharing economy offer a wide variety of opportunities to design new TDM strategies which up to now were unthinkable.

This chapter's contribution lies not only in describing the origin and evolution of Transport Demand Management but also in the proposal of a new TDM approach based on the technological change of the urban and interurban transport environment.

The Origin and Evolution of Transport Demand Management (TDM)

The origin of Transport Demand Management (TDM), also called "Mobility Management," took place at the end of last century (Nijkamp et al., 1998; Banister, 1999), when the absolute growth in travel, especially by car and plane, started to create a serious concern in terms of pollution (emission of harmful gases). At the urban level, reducing private car travel became the main objective of the first TDM strategies. Specifically, the target was the reduction in vehicle kilometers from trips made by environmentally damaging modes of transport, like the private car. At European level, transport policy programs reflected this concern on the increase of private car traffic (CEC, 1992, 1993). The electric car market had not yet taken off and smart technologies were not so powerful to connect transport users to operators, and other users. The context in which Transport Demand Management originated should not limit the concept of TDM to the strategies of solely reducing private car travel. TDM should be defined as the **design of a group of strategies devoted to improve the efficiency of the transport system exploring demand side measures**. The key issue is what we understand by "efficiency of the transport system" at any time in history.

Marshall and Banister (2000) were the first to understand the complexity of exploring only demand side measures when designing TDM strategies to reduce private car travel in cities. In order to reduce private car travel, an alternative of transport should be provided to private car users (like mode substitution or modal shift to public transport). Even decreasing the distance traveled by private car requires the adoption of other complex policy measures linked to the labor market (such as, the implementation of teleworking) and the change of location of activity centers in cities. Direct **push measures** (i.e., tax increase on fossil fuel) should be combined with **pull measures** (i.e., improvement of public transport or ticket price reduction) to achieve larger behavioral responses compared to the individual push measures. Marshall and Banister (2000) proposed the design of "strategic packages" to ensure that combining push and pull measures can achieve a tangible reduction in travel.

Some years later, Litman (2003) collected an "Online TDM Encyclopedia" showcasing 48 TDM strategies implemented all over the world, including push measures (measures to directly reduce private car use) and pull measures (to indirectly reduce private car

Table 1 Short- term TDM strategies included in the online TDM encyclopedia

<i>Incentives to shift mode</i>	<i>Improved transport options</i>
Bicycle and pedestrian encouragement	Address security concerns
Congestion pricing	Alternative work schedules
Distance-based pricing	Bicycle improvements
Commuter financial incentives	Bike/transit integration
Fuel tax increases	Carsharing
High occupant vehicle (HOV) priority	Guaranteed ride home
Pay-as-you-drive insurance	Park and ride
Parking pricing	Pedestrian improvements
Road pricing	Ridesharing
Vehicle use restrictions	Shuttle services
	Taxi service improvements
	Telework
	Traffic calming
	Transit improvements universal design

Source: Adapted from [Litman \(2003\)](#).

use). Following the approach of [Marshall and Banister \(2000\)](#), [Litman \(2003\)](#) emphasized the integrated vision of TDM planning, providing information on how various specific strategies can be combined for maximum effectiveness. Indeed, one of the major contributions of Litman was the identification of 48 different strategies already implemented, a very complete inventory of TDM actions, that provides the profile of the TDM type implemented at the end of the 20th century. [Table 1](#) and [Table 2](#) show the inventory of actions identified by Litman. [Table 1](#) groups short-term actions directly linked to the most common push and pull measures related to the reduction of private car travel, and [Table 2](#) shows more general long-term measures linked to land use management and policy reforms.

Each TDM strategy was defined by a combination of possible actions and registered with a short description and a list of indicators: how it was implemented, travel impacts, costs and benefits, equity impacts, application, relationship with other TDM strategies, stakeholders, barriers to implementation, best practices, case studies, references, and resources for more information.

The inclusion of equity impacts to the description of the TDM measures ([Litman, 2003](#)) reflects the start of a growing concern about the social fairness of transport policy measures ([Martens, 2016](#)). TDM can push users with lower incomes with less accessibility from their homes to key destinations (e.g., labor destinations) to reduce driving and shift to public transport modes. Fair transport accessibility is a key issue to achieve social equity in transport and a minimum standard of accessibility to the main destinations is required to respect individuals' rights and protect disadvantaged groups. According to [Lucas et al. \(2015\)](#), methodologies to assess new transport policies should include theories of *egalitarianism* and *sufficientarianism* in combination with an accessibility-based analysis using the Lorenz curve and Gini index. Accessibility studies need more research and a more complete understanding than traditional approaches to incorporate the concepts of equity and justice in transport. This indeed was the development trajectory for the evolution of TDM and its assessment.

Each TDM listed by [Litman \(2003\)](#) was also rated on a seven-point scale, including criteria related to their travel impacts and their ability to achieve some objectives. The range of possible objectives defined by [Litman \(2003\)](#) shows, in some way, the range of criteria of TDM efficiency used until 2003. TDM was defined, at the beginning of this chapter, as a group of strategies devoted to

Table 2 Long- term TDM strategies included in the online TDM encyclopedia

<i>Land use management</i>	<i>Policy and planning reforms</i>	<i>Implementation of support programs</i>
Car-free districts	Car-free planning	Access management
Clustered land use	Comprehensive transportation market reform	Campus transportation management
Location efficient development	Institutional reforms	Data collection and surveys
New urbanism	Least cost planning	Commute trip reduction
Parking management	Regulatory reform	Intelligent transportation systems
Smart growth		Freight transportation management
Transit oriented development (TOD)		School trip management
Street reclaiming		Special event management
		TDM marketing
		TDM programs
		Tourist transport management

Source: Adapted from [Litman \(2003\)](#).

improve the efficiency of the transport system exploring demand side measures. The list of efficiency criteria includes not only congestion reduction or transport choice, but also road and parking savings, consumer savings, road safety, environmental protection, efficient land use and community livability. Environmental protection has also gained importance among the evaluation criteria over the years. Some authors have underlined that effectiveness of transport policies concerns directly the improvement of the environmental performance of the transport system (Vieira et al., 2007). The recent UN Sustainable Development Goals (United Nations, 2015) further support this approach.

In 2016, Ison and Rye (2016) proposed an alternative classification of TDM measures aimed at influencing people travel behavior. They divided TDM actions into economic measures, land use measures and administrative measures, but introducing new typologies of actions as those based on “information for travelers” (travel information before a trip is undertaken or carsharing) or “substitution of communications for travel” (such as tele-working and e-shopping). These new types of TDM reflect the new trends of mobility which started to emerge in 2016.

The assessment of TDM strategies, which have already been implemented is a complex issue. Accurate information on users' behavior changes is not easy to obtain and while information exists, it is scattered across different administrations and public access to this data is sometimes limited. In recent years new ICTs and internet-based information resources have been developed and information in support of TDM planning and evaluation has improved. Prior to the recent technological changes, three different types of studies (Eriksson et al., 2010) have been carried out to assess users' behavioral responses to TDM:

- Field experiments to measure cause-effect relationships (e.g., user surveys, traffic flow measured by road gauging stations)
- Estimation of transport elasticities, based mainly on stated preference studies
- Analysis of hypothetical TDM measures

Literature on field experiments shows that changes in the cost of transport (fuel cost, public transport tickets) definitely influences travel behavior (Heath and Gifford, 2002). Research on transport elasticities tries to design indicators to measure the extent to which travel demand is sensitive to changes on public transport services or private car pricing (push and pull TDM measures), looking for a ratio of the proportional change of behavior to the proportional change in services, fares or prices. According to some literature a 10% increase in fuel prices only generates between 1% and 3% car use reduction (Dargay, 2007), while improvements in public transport, such as a 10% service increase, provide a higher percentage of public transport ridership (5%) (Evans, 2004). The debate between the efficiency of push measures against pull measures has been very intense, but a general consensus on the need of mixed solutions as a package within the TDM strategy has been consistent over the years (Banister, 2008; Vieira et al., 2007). Although this scientific consensus is a good goal for the evolution of TDM, the emergence of the concept of “MaaS” will open a new scenario for debate in this area.

A Disruptive Tool for Transport Management: Mobility as a Service (MaaS)

During the last decade mobility has experienced several disruptive changes which greatly complicate the analysis of transport demand management measures. These include:

- Widespread use of ICTs among the population (particularly smartphones)
- Social acceptance and confidence in the new consumption forms brought forward by of the “sharing economy”
- Decisive progress in the development and implementation of autonomous vehicles
- Important advances in the development and implementation of electric and hybrid vehicles

Some of these disruptive changes are interdependent. For example, the acceptance of the sharing economy has been promoted by the mass use of smartphones among the population and the design of easy-to-use applications (apps) which need no specialized digital skills. Additionally, the implementation of autonomous vehicles assumes a continuous digital communication between users, road infrastructure and vehicle operators guaranteed by the mass use of ICTs.

In this context, the concept of MaaS has emerged turning the old TDM vision on its head. New technologies including autonomous vehicles and transport systems are being closely linked to a service concept, in a scenario where mobility is no longer consumed as a “good” (based on private vehicle ownership), but rather as availability to access the transport system. The majority of MaaS forms offer door-to-door multi-modal integrated mobility services, being an alternative to private vehicle ownership. MaaS is provided via digital platforms connecting users and service operators, so users have previously to be subscribed to smartphone applications designed by operators. There is no scientific consensus on the requirements of a service to be considered a MaaS form and how to compare different MaaS services. According to the literature (Pangbourne et al., 2020), this term was firstly used in a doctoral thesis read in Finland in 2014 and, some years later, a good inventory of MaaS definitions and concepts was also produced by Sochor et al. (2018), who analyzed the more frequent key words used to define MaaS: integration, multimodality, user-centric, on-demand, multimodality, flexible, platform, door-to-door, one-stop shop and seamless transport systems. One of the most completed definitions was provided by Arthur D. Little (2018), who considers MaaS an integrated, flexible and user-oriented mobility service developed through a digital platform (app), combining public and private transport providers. This platform should not only help to plan the trip but also to pay a true mobility package (a multimodal trip) covering the needs of each customer segment. Obviously the report by Arthur D. Little (2018) defined an advanced type of MaaS service, but Sochor et al. (2018) were

conscious that currently implemented MaaS does not achieve this advanced level of integration, and went on to identify four level of possible “integration” in order to classify the topology of each MaaS:

- Level 0. No integration (separate services)
- Level 1. Integration only of information (multimodal travel planner, price info)
- Level 2. Integration of booking and payment (single trip, find, book, and pay)
- Level 3. Integration of service offer (subscription, contacts)
- Level 4. Integration of societal goals (policy, social incentives)

The progressive implementation of MaaS has generated a debate within the scientific community on the threats that this new mobility approach can represent for transport equity and sustainability and which TDM strategies should be the more suitable within the context of this new development. As level 4 is rarely achieved, [Sochor et al. \(2018\)](#) raised their concern with respect to societal goals and highlighting the role of public administrations in particular. More recently, a number of authors have underlined the adverse effects that MaaS can generate if not organized properly ([Wong et al., 2020](#), [Pangbourne et al., 2020](#)). Based on flexible travel, modal integration and shared mobility, these new platforms will generate induced trips, potentially increasing congestion, and compromising road capacity. If modal efficiency is not achieved, sustainability problems will come back and, in the longer term, MaaS would also modify the urban structure, impacting the land use features in each territory. Indeed [Wong et al. \(2020\)](#) have been pioneers in considering the need for a type of MaaS based on a government-contracted model, with road pricing being incorporated as part of the package price, depending on the mode efficiency, time of day or land use characteristics.

Threats of the MaaS Model and Need of New TDM Strategies

In terms of the design of a new TDM strategy, the existence of MaaS systems should be incorporated as an additional variable in the urban scenario, not as a solution itself. The potential of introducing and increasing new road trips is one of the many threats associated with the implementation of MaaS platforms due to their increased levels of accessibility. Scientific evidence of this fact has been recently revealed in different studies that analyze the new forms of mobility that integrate MaaS and now are currently in operation (like Free Floating Car Sharing). Surveys on future travel behavior have also revealed that autonomous vehicles will induce further new road trips and in the long term, change the shape and structure of our cities.

Autonomous vehicles could increase the efficiency of TDM actions, in terms of traffic safety and management, but their only existence does not guarantee better TDM strategies. The main targets of the autonomous vehicle are improving traffic flows and increasing road safety. Autonomous vehicles will allow for narrowing of lane widths in streets, reducing distances between vehicles (headways) and removing traffic signaling at intersections; and in the long term, reducing the levels of investment in road infrastructure ([Talebpour and Mahmassani, 2016](#)). Extra-capacity of the road network is the first promise of the autonomous vehicle innovation and can be viewed as a panacea for their on-demand mobility potential. But due to the time savings, an induced demand is also expected not only from existing drivers (the time saved in travel will be invested on traveling further or developing more trips) but also from non-drivers. In this induced additional flows, operational trips (zero-occupancy vehicles to relocate vehicles or recharge electric batteries) should also be considered. [Harb et al. \(2018\)](#), showed some of the impacts that autonomous vehicles will have on people’s travel behavior with a 76% increase in km traveled, longer trips and more displacements in the evenings than previously. Considering all trips in the new system, 20% would carry no passengers.

The percentage of private ownership of autonomous vehicles is a key issue for the sustainability of the whole transport system. Even thinking on the hypothesis of a universal transport system based on “automated taxis” (robo-taxis or taxibots), all operated by a mobility provider ([Enoch, 2015](#)), TDM strategies have to be carefully designed to avoid urban sprawl and road congestion. Automated taxis would replace private cars (due to externalities), conventional taxi (due to the lower cost of automated taxis), and buses (due to their failure to satisfy point to point demand). Urban sprawl would be promoted by this new transport system and the threat of road congestion may also limit the number of trips per family/person and the mobility provider will have to restrict the individual freedom to travel. As an alternative, TESLA has proposed a scenario where autonomous vehicles are privately held but hired out, when not used, for ridesharing. This scenario improves some of Enoch’s scenario freedom limitations but the temporal efficiency will depend on the proportion of private vehicles over the total and the complex demand management of the whole system (spatial and temporal peak hours). All these first studies reveal the need of a comprehensive digital interconnectedness of all involved actors, not only all the modal operators but also public administrations and regulators. Indeed, new TDM strategies should be promoted on the basis of this new structure.

More evidence generated by new types of MaaS systems can be found in the new generation of carsharing: the so called FFCS (Free-floating carsharing). FFCS can be considered a type of MaaS with the topology of Level 0. There are platforms in cities that offer integration of travel information (Level 1), including all carsharing operators and it is possible to plan a multimodal trip, obtaining an estimation of the total price, but in the majority of them the booking and payment has been carried out through each FFCS company web. As level 4 is rarely achieved, these systems can also represent a threat for the transport system sustainability, and new TDM strategies should incorporate this approach. If new FFCS users are transferred from public transport, are displaced from active modes and increase the number of vehicle-km traveled, then the positive FFCS effects on the transport system should be questioned. Moreover, FFCS cars, even electric ones, occupy a space in the traffic flows, influencing the level of service of the road urban traffic

flows where public transport operates. It would be necessary to study if, in the majority of the FFCS displacements, trips are also shared or not (vehicles are occupied only by a sole driver with no passengers). Finally, FFCS cannot be conceived without the existence of an efficient public transport network that would guarantee a return or next trip should a car be unavailable at the time required and at the location of the user. There is indeed lack of evidence in this respect. Similar concerns arise with the use of electric scooters and moto-sharing, such as the displacement of active travel modes, road safety issues and the growing conflict over pavement space in many cities.

In order to conclude, new TDM strategies should incorporate certain measures depending of the level of MaaS implemented in the transport system (0–4), the stage of penetration of autonomous and electric cars (percentage over the total car population) and variables conditioning free-floating carsharing like for example, the on-street parking availability.

Planning TDM Strategies for a Transition Time

Despite the progress in the electric and autonomous vehicles manufacturing and the current implementation of collaborative and connected mobility systems, there are three premises that TDM strategies will have to work with:

- 1) The co-existence of more than a decade of conventional transport systems with more advanced ones.
- 2) The impossibility of advanced transport systems, like many MaaS operations, to independently self-manage taking into account only their economic interests.
- 3) The need for an administration and regulatory framework, mainly in cities, able to manage transport demand by ensuring environmental sustainability and social equity.
- 4) The need, through new technologies, to encourage ride-sharing in addition to shared vehicles.

Until now, advanced transport systems (like carsharing systems) have been implemented with little regulation in cities. The market share of the FFCS system, for example, is very small compared to other modes of transport, but the increased use of autonomous electric vehicles or robo-taxis may change completely this scenario.

TDM future strategies should be based on transition phases, depending on the urban and socio-economic structure of each city and its previous public transport systems. Public local administrations should lead the design of a “level 4” MaaS to manage and coordinate the mobility infrastructure. Additionally, future TDM strategies should support pull measures over push measures, giving priority to individual incentives change behavior. Today real-time and on-demand information is easily available and information of individual movements can be tracked through GPS systems and smartphones. Therefore, the incentives policy should be based on sustainable individual behavior (sharing trips, using public transport, taking an electric taxi, etc.), being complemented with push measures such as road and parking pricing and fuel taxes. The lack of accurate information on users' behavior changes had been until now the weakest part in the assessment of implemented TDM measures, but ICTs have changed this scenario. Effectiveness of new proposals of TDM strategies is supposed to be tested easily now and this fact should encourage also the design of innovative and holistic TDM strategies. In the end new technologies should be at the service of social equity, economic progress and environmental concerns.

Conclusion

This chapter was drafted in March and April 2020, in full state of lockdown caused by the coronavirus pandemic. Mobility has declined dramatically as a result of this confinement in households, and the labor market has had to experience tele-working in large scales across many cities. In some cases with very positive results, for example the reduce levels of pollution in major cities worldwide. This historic event will affect mobility behavior not only in the coming months but also in the next decade, potentially reducing commuting trips for a larger variety of jobs than ever before. New ICTs and the first experience with the first MaaS models should inform transport planners to envision and design a transport system operation that is fully coordinated by an administration (public or mixed public–private) capable of ensuring environmental sustainability and social equity.

References

- Arthur D. Little, 2018. The Future of Mobility 3.0. Reinventing mobility in the era of disruption and creativity. Retrieved from: https://www.adlittle.com/sites/default/files/viewpoints/adl_uitp_future_of_mobility_3.0_1.pdf.
- Banister, D., 2008. The sustainable mobility paradigm. *Transp. Policy* 15, 73–80.
- Banister, D., 1999. Some thoughts on a walk in the woods built environment. *Travel Reduct. Policy Prac.* 25, 162–167.
- CEC, 1992. The Future Development of the Common Transport Policy: A Global Approach to the Construction of a Community Framework for Sustainable Mobility. Bulletin of the European Communities. Supplement 3/93. Commission of the European Communities, Brussels.
- CEC, 1993. Towards Sustainability. A European Community Programme of Policy and Action in Relation to the Environment and Sustainable Development (Fifth Environmental Action Programme), pp 1–93. Commission of the European Communities, Brussels. Retrieved from: <https://ec.europa.eu/environment/archives/action-programme/env-act5/pdf/5eap.pdf>.
- Dargay, J., 2007. The effect of prices and income on car travel in the UK. *Transp. Res. Part A* 41, 949–960.
- Enoch, M.P., 2015. How a rapid modal convergence into a universal automated taxi service could be the future for local passenger transport. *Technol. Anal. Strategic. Manag.* 27, 910–924.

- Eriksson, L., Nordlund, A.M., Garvill, J., 2010. Expected car use reduction in response to structural travel demand management measures. *Transp. Res. Part F* 13, 329–342.
- Evans, J.E., 2004. Traveler response to transportation system changes. Chap. 9: Transit scheduling and frequency. TCRP report 95. Transportation Research Board, Washington, DC.
- Harb, M., Xiao, Y., Circella, G., Mokhtarian, P.L., Walker, J.L., 2018. Projecting travelers into a world of self-driving vehicles: Estimating travel behavior implications via a naturalistic experiment. *Transportation* 45, 1671–1685.
- Heath, Y., Gifford, R., 2002. Extending the theory of planned behavior: Predicting the use of public transportation. *J. Appl. Soc. Psychol.* 32, 2154–2189.
- Ison, S., Rye, T., 2016. The Implementation and effectiveness of Transport Demand Management Measures. *An International Perspective*. Routledge: Taylor and Francis Group
- Litman, T., 2003. The Online TDM Encyclopedia: mobility management information gateway. *Transp. Policy* 10, 245–249.
- Lucas, K., van Wee, B., Maat, K., 2015. A method to evaluate equitable accessibility: Combining ethical theories and accessibility-based approaches. *Transportation* 43, 473–490.
- Marshall, S., Banister, D., 2000. Travel reduction strategies: intentions and outcomes. *Transp. Res. Part A* 34, 321–338.
- Martens, K., 2016. *Transport justice: Designing fair transportation systems*. Routledge, London.
- Nijkamp, P., Rienstra, S.A., Vleugel, J.M., 1998. Design and Assessment of Long Term Sustainable Transport System Scenarios. In: Button, K., Nijkamp, P., Priemus, H. (Eds.), *Transport Networks in Europe: Concepts, Analysis and Policies*. Edward Elgar, Cheltenham, pp. 307–332.
- Pangbourne, K., Mladenović, M., Stead, D., Milakis, D., 2020. Questioning mobility as a service: Unanticipated implications for society and governance. *Transp. Res. Part A* 131, 35–49.
- Sochor, J., Arby, H., Karlsson, I., Sarasini, S., 2018. A topological approach to Mobility as a Service: A proposed tool for understanding requirements and effects, and for aiding the integration of societal goals. *Res. Transp. Bus. Manag.* 27, 3–14.
- Talebpoor, A., Mahmassani, H.S., 2016. Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transp. Res. Part C* 71, 143–163.
- United Nations, 2015. Resolution adopted by the General Assembly on 25 September 2015. Transforming our world: the 2030 Agenda for Sustainable Development. Retrieved from: https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E.
- United Nations, 2018. The Sustainable Development Goals Report 2018. Retrieved from: <https://unstats.un.org/sdgs/files/report/2018/TheSustainableDevelopmentGoalsReport2018-EN.pdf>.
- Vieira, J., Moura, F., Viegas, J.M., 2007. Transport policy and environmental impacts: The importance of multi-instrumentality in policy integration. *Transp. Policy* 14, 421–432.
- Wong, Y.Z., Hensher, D.A., Mulley, C., 2020. Mobility as a service (MaaS): Charting a future context. *Transp. Res. Part A* 131, 5–19.

Planning and the Global Movement of Goods and Commodities

Christopher Clott, Chris Petrocelli, State University of New York Maritime College, New York City, NY, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	380
Risk Factors in the New World Related to the Global Movement of Goods	380
General Risk Factors to Consider	381
Transportation Planning Characteristics	381
Method for a Business Startup: Definition, Configuration and the Impacts on Transportation Planning	382
Tools to Assist in International Transportation Planning for the Movement of Goods	382
Project Management	382
Supply Chain and Transport Process Modeling	384
Outcomes—Activities Don't Matter, So Don't Measure Them!	386
Objective Factors	386
Subjective Factors	386
Conclusions—How do You Measure?	386
References	387

Introduction

Planning the global movement of goods and commodities has never been a simple procedure. Cost, quality, and delivery have always been key needs in developing transportation plans. In this new period where we are confronted by uncontrollable events such as the COVID-19 virus, we must also think about resiliency as a critical global supply need. A worldwide pandemic combined with increased nationalism and more regional sourcing requirements have upended traditional thinking. When COVID-19 first hit the industry, the initial widespread response was to safeguard employees by following the guidelines set by the World Health Organization (WHO) and the governments of the countries they operate in, such as providing hand sanitizers and working from home (*Business and Financial Times*, April 2020). Given the various uncertain trajectories of these events it is important to understand how to manage through these risk factors. This chapter will focus on the overall process of setting up executable transport planning models based on transport policies and plans for the global movement of goods that are relevant to the details of the business at hand. Given the challenges we face today, there needs to be a vision and road map using some key principles to be successful.

The need and use of transportation services for the movement of goods must take into consideration all expectations of the party paying for these services. Some requirements for transportation services must satisfy a primary need such as speed, cost, quality, and reliability. At times requirements may shift at different points in the supply chain as well as the external business factors that may influence transport requirements.

This section will provide the basic instruction in order to apply a method for supporting transportation requirements in a business startup. This includes an overview of some typical requirements that would need to be modeled under given assumptions. This also will include a series of steps with the outcome being a successfully planned and implemented transportation system for the business at hand. The chapter will also provide some insights on various control charts and modeling concepts to allow for the user community and all stakeholders involved to have a good understanding of the factors being considered when planning transportation for goods movement.

Risk Factors in the New World Related to the Global Movement of Goods

With all that has been taking place in 2020 related to the impact of the COVID-19 virus, there are several risk factors that need to be considered as you look across supply chains as they relate to transportation planning for the future. Change is inevitable and the old global supply chain playbook that was in place being used by many US and multinational companies is now being rewritten to cover new factors including how to deal with various risk factors as it relates to pandemic conditions. We have just experienced and are continuing to experience the impact of the COVID-19 virus across the supplier base, transport services, and the customer demand over the past months. The past practices and process controls that used to manage transportation and planning aspects of the business will have to change to accommodate for a well-managed transportation system.

During the initial phase of the COVID-19 virus, the supply chain and transportation issues first began with sourcing goods from the People's Republic of China and then expanding to Hong Kong immediately after the Chinese New Year. During this period, firms experienced a reduction in the supply of air transport from this key manufacturing and logistical region. This had a large impact on

the amount of goods that could be moved by air. This was caused by the drastic reduction of passenger aircrafts from the region with only limited freighter aircraft flying. Keeping up with the day-to-day challenge to be sure that all demand requirements are met drove the cost of air transport up by 300%–400% from Hong Kong and China airports to the United States during the first months of 2020 ([FreightWaves Newstex Finance and Accounting Blogs, 2020-05-21](#))

One of the key factors contributing to the disruption of the transport systems was the human factor. So much of the activities require interchanges/exchanges of goods and documents which go from one person to another. Increased data analytics and the digitalization of information flow will accelerate changes to transportation planning but there is still a need at this time for skilled individuals to manage their interactions. Additional factors that came into play were airport operations scheduling personnel limitations and limited hours and days of operation. This then spread onward to the United States and Europe as the effects of the virus impacted operations both on the ground as well as local delivery requirements and operations for local trucking services to have personnel and operations to accommodate normal demand.

General Risk Factors to Consider

- Environmental risk: Where across the supply chain as a whole is it vulnerable to external forces? While the type and timing of extreme external events may not be able to be forecasted, their impact needs to be assessed.
- Supply risk: How vulnerable is the business to disruptions in supply? Risk may be higher due to global sourcing, reliance on key suppliers, poor supply management, and so on. It is critical that given recent developments in tariff changes, shift in manufacturing capabilities (i.e., sourcing product Vietnam rather than China) and COVID-19 type disruptions that adequate alternatives exist when needed.
- Demand risk: How volatile is demand? Does the “bullwhip” effect cause demand amplification? Are there parallel interactions where the demand for another product affects the demand for ours? This has become a serious issue during COVID-19 for the US retail market. Once goods began to move into the North American region in late February, the US retail came to a standstill due to the initial days of COVID-19. By early March, most US retailers had closed their stores. This caused a bottleneck in getting goods into the US distribution centers due to stoppage of local air, sea, and trucking operations.
- Process risk: How resilient are our processes? Do we understand the sources of variability in those processes, for example, manufacturing? Where are the bottlenecks? How much additional capacity is available if required? Companies that have achieved a level of automation in the management of transportation transactions were better positioned to handle the process challenges of COVID-19. During this period while employees were forced to work from home companies that still relied on manual, paper systems for transportation planning and transaction processing had real challenges replicating the environment. Companies that had already migrated to some form of transportation planning and execution systems were able to minimize disruptions.
- Control risk: How likely are disturbances and distortions to be caused by our own internal control systems? For example, order quantities, batch sizes and safety stock policies can distort real demand. Our own decision rules and policies can cause “chaos”-type effects. During the early stages of the COVID-19 virus the information was not clear as to what the conditions were day to day in China and its effect on Hong Kong for air and sea transport. It was critical to rely on multiple sources (Carriers, Forwarders, other shippers, and contacts) to determine what the current situation was on the ground. It was a daily process to collect the communications from various sources and provide a summary to the management team to explain the conditions and impacts on transportation planning and execution ([Coyle, 2011](#)).

Transportation Planning Characteristics

Several characteristics need to be taken into consideration in order to properly plan the transportation required. These characteristics are outlined below.

- Time sensitivity is an important factor to take into consideration while looking at the transportation requirement. The same product can potentially move in various time methods to minimize cost if scheduling allows for slower methods or modes of transportation. The same cargo can be moved from a factory to a destination via air or sea depending on the time built in the overall supply chain for transport days.
- Distance: How far will the goods need to travel from origin to destination? Will transport be completed in hours, a day, or several days or weeks? This needs to be taken into consideration when performing inventory planning of the goods being shipped.
- Domestic or international: Does the transport require the movement of goods over an international border? If the goods are crossing an international border the proper documentation and markings on the goods need to be included in the process. Documents such as a commercial invoice, packing list, and certificate of origin are common documents required. Factors such as pandemic related country restrictions may become a factor in this decision.
- Commodity: What is the product or goods being transported? The type of commodity will have big influence on the types of service available which will be reflected in the service days and cost of the transportation service.
- Volume/frequency: The volume or amount of product to be transported as well as the frequency of how often transport services are required are two key factors that will drive the pricing structure for transportation. More frequent higher volume shippers will

enjoy benefits of lower pricing due to commitment on volume and regularity of shipping. Carriers and folders will always honor these types of customers in that they can plan their base business around such steady volumes

- Macro-market conditions: Several macro-economic market factors must be taken into consideration at any time during planning of transportation requirements for the movement of goods to facilitate a business supply chain. These factors tend to have different types of life cycles and influences on the overall cost and service quality of a transportation system or method selected. The globalization of industries has in some ways created macro-economic forces on product supply and transportation. There continues to be a downward pressure on price of service forcing competitors to handle business at extremely low margins to compete. Customers requiring the use of transportation services have been able to take more control given the increased digital information and communications available today as tools to aid the procurement activities (Wang et al., 2018).

Method for a Business Startup: Definition, Configuration and the Impacts on Transportation Planning

Process mapping is not a new technique for analyzing operational efficiency. Its effectiveness rests in the ability to pictorially portray how seemingly disparate processes are connected; to illustrate the essential information needed to drive the work; and to ultimately illustrate how process flow relates to organizational roles and responsibilities (Martin, 2005). Below is an outline of the project deliverables needed to get through a successful startup to a steady state of operations:

1. Determine the scope of transportation services required
 - a. Provide objective, professional guidance, and support in the area of supply chain design and implementation for Phase 1 operations of the company direct to consumer business.
 - b. Apply all knowledge, analytical skills, and industry contacts to facilitate result-oriented outcomes.
2. Defining Phase 1 operational objectives/deliverables
 - a. Document business requirements and assumptions for Phase 1 operations.
 - b. Configure supply chain model for Phase 1 operations as defined.
 - c. Design and implement the required operational application systems for Phase 1 operations.
 - d. Sourcing, order management, inventory management, and other management requirements.
3. Supply chain physical distribution modeling
 - a. Profile and documentation of all business requirements and modeling assumptions.
 - b. Sourcing strategy—buying terms, trade compliance, transport from supply sources.
 - c. Packaging strategy—bulk to consumable unit repackaging, kitting services.
 - d. Stocking model—strategy, volumes.
 - e. Demand management strategy—order cycle requirements for Phase 1 customers.
 - f. Physical distribution modeling.
 - g. Define/document each segment of the supply chain (all stocks and flows).
 - h. Determination Phase 1 assumptions and model for physical distribution to support business startup.
4. Supply chain operations systems determination
 - a. Profile and document business process assumptions to support Phase 1 business operations.
 - b. Define the scope of Phase 1 system processes required to support.
 - c. Procurement, international/domestic transport, trade compliance, repacking, warehouse/inventory management, CRM, order processing, shipment tracking.
 - d. Define decision making criteria to select an overall IT/systems strategy and select each business application required for Phase 1 operations.

Tools to Assist in International Transportation Planning for the Movement of Goods

Consider for a moment how the conventional organization processes orders goods. Typically, there will be a sequence of activities beginning with an order entry. The point of entry may be within the sales or commercial function but then it goes to credit control, from where it may pass to production planning or, in a make-to-stock environment, to the warehouse. Once the order has been manufactured or assembled, it will become the responsibility of distribution and transport planning (Martin, 2005).

Project Management

Fig. 1 is an example of a project plan format (Gantt chart) used to define tasks and manage each task toward project completion. This tool helps to define each project task, ownership, and timing. This format also assists in developing the overall project plan and timeline to get to the implementation of Phase 1 operations.

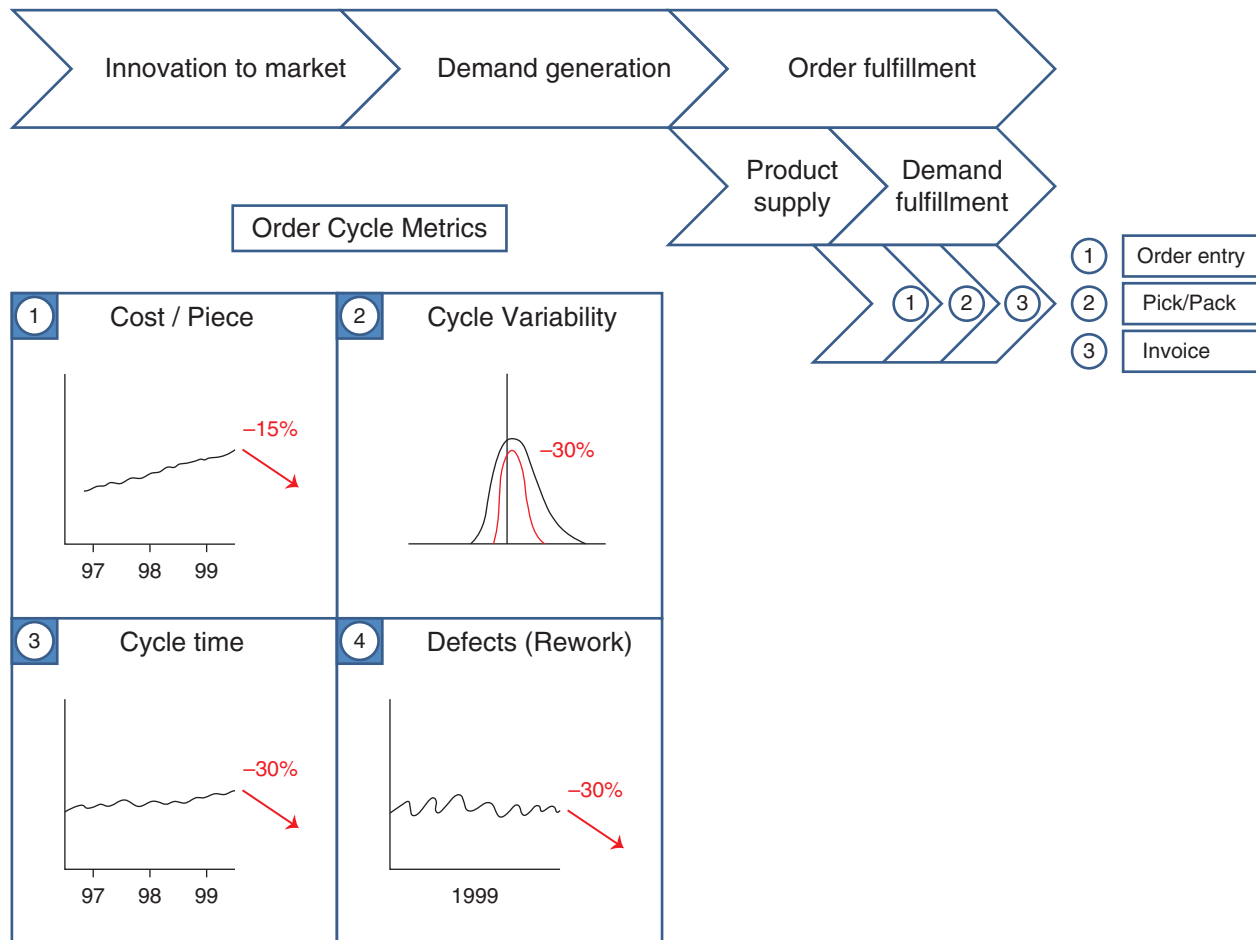


Figure 2 Order cycle metrics. *Source: Author created.*

Supply Chain and Transport Process Modeling

Fig. 2 illustrates the overall business process and each major subprocess. The first step is to organize the business process into a series of major processes and some other processes. Each major process represents a cash to cash segment of the business. Each subprocess represents a process step or segment of the overall business process. Each subprocess step is defined by determining the process start and process end. It is important to also understand the base unit of measure that is being used. This could be per unit, per piece, per carton or another unit of measure as defined by the company.

Cycle time is critical to measure in both absolute time from start to finish as well as variability around the mean of the cycle time. It is also critical to measure the defect rate in the process to be sure that time is not wasted on rework items such as reshipping re invoicing, and so on.

Fig. 3 is an example of a specific business process step within an overall supply chain. This example is the subprocess of moving goods from source to market region warehouse. This is one subprocess step within the product procurement cycle. Each process step uses a similar template to capture key characteristics for each supply chain process step for the company business.

Other modeling tools to consider when planning transportation are shown below:

- Procurement and sourcing practices—single sourcing SKUs will be scrutinized to be sure the supply chain is secure and consider alternative sources if needed.
- Supply network design—near sourcing will come back into play to shorten the process segments required to get goods from sources to markets.
- Supplier monitoring—24/7 supplier monitoring will become the standard to detect early defect conditions.
- Demand drives transport needs.
- Macro factors—detecting changes for influences such as the COVID-19 virus, political conditions, or other macro conditions that may disrupt the transport services.

When considering the best configuration for a new business startup there are several factors that must be incorporated into the transportation planning. The business configuration will drive the types or levels of transportation required for the complete supply

Business Process Step

Shipping goods from Factories to Market Region DCs - Once goods are produced they are exported to market region distribution centers

Profile

- Service Providers: Air/Sea transport providers
- Annual Spend:

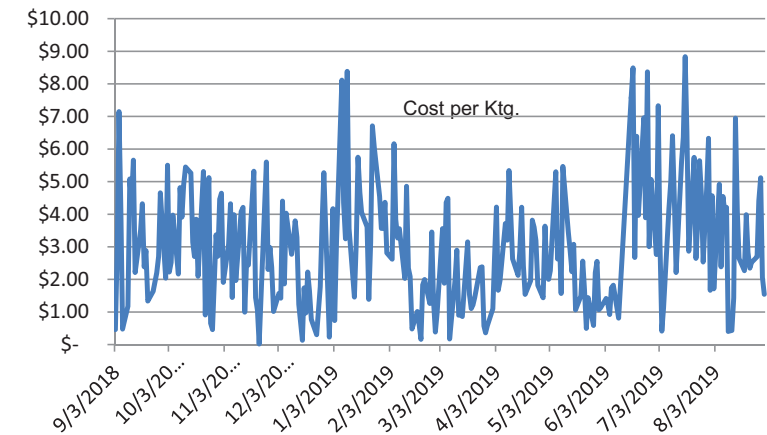
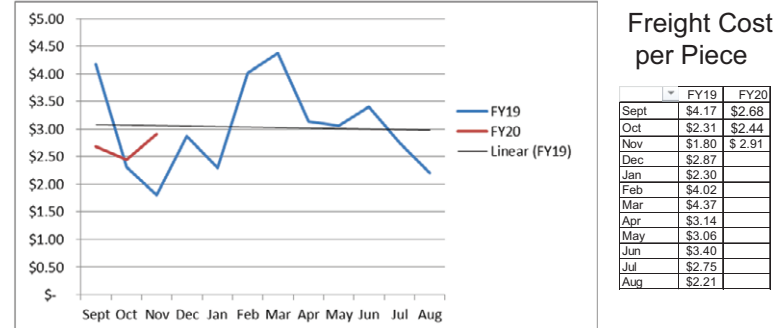
Q1 Accomplishments

- Evaluated possible Supply Chain Visibility Tools and created a process flowchart
- Supported the FOB and Lead Times enhancement development
- Completed the initial shipments for Samsung using the Shanghai bonded warehouse
- Completed rate review reducing air transport cost

Q2 Plans

- Next steps regarding Supply Chain visibility
- Complete pricing review for upcoming season
- Rework Management – review invoice error log to spot and act on root causes
- Convert delivery rate basis from pieces to weight (kgs)

Key performance indicators



chain (Martin, 2005). Transportation planning and execution required for the complete business process will include the need to move materials and/or finished goods from place to place dependent upon specifications including:

- Subvendor to vendor transport—usually commodities. Often the transport requirements for moving goods from a subvendor to a vendor are for raw materials considered commodities in raw form. Examples would be rolls of steel, rolls of fabric, large bulk volumes of oil, and raw minerals. During the early days of the COVID-19 pandemic, textile mills in Italy had shut down limiting the availability of fabric for garment manufacturers.
- Vendor to distribution center—tends to be the finished goods for the business. This transportation requirement can range from international air or sea to domestic truckload, less-than-truckload (LTL) or parcel (small package) depending on the type of commodities. Some business relationship may require direct shipping from the vendor or factory to the customer.
- Distribution to customer shipping requirements continue to become more and more stringent with many customer specific requirements. Customers today are now sourcing not only locally within country but also within a geographic region as well as on a global level. This requires customer specific delivery from not only distribution centers but potentially other points in the supply chain to be taken into consideration when planning transportation for customer requirements in general and will include customer to distribution (returns).

To satisfy the business requirement for each segment of transportation it is important to be sure all key characteristics that drive the service quality, speed, and cost are considered. Such transport requirements will need to be considerate of factors such as unusual size or weight per shipment based on the type of commodity being transported from sub vendor to vendor also the consideration of the timing and quality of service is a factor as well (Arroyo and Holmen, 2017; Reichhart and Holweg, 2007).

Outcomes—Activities Don't Matter, So Don't Measure Them!

Proper metrics are required to immediately detect potential issues. These metrics are derived from the key drivers that impact the success of the transport system in place. The drivers tend to be related to cost, time, and quality of service. Firms can become focused on activities and effort that are put into a project or initiative to solve a problem rather than the outcomes. Managing and measuring of transport activities is an art unto itself and firms that lose sight of the reason why proper metrics for important activities are critical can have detrimental effects (Herden, 2020).

So what is important to measure? One method in the transportation planning process is to break down the key metrics through benchmarking to find the largest drivers of the overall process goal (Reh, 2019). If specific numeric values, we are looking to measure are cost related, then we need to look at the key drivers that would most highly influence the cost such as distance shipped or the weight of the product being shipped.

Objective Factors

When we think about objective factors that are driving transportation cost, the Pareto method to define these factors is used. By isolating each variable in the transportation equation, one can determine what the largest key drivers are. Key drivers to solve are cost per unit shipped or quality of one aspect of service (tendering, invoicing). Key drivers for transportation include commodity type, weight, and distance.

Subjective Factors

Subjective factors of the performance of a transport system or process can be the experience of the participants that are impacted by the activities. Participants can have different focus depending on the influence. Focus can be on only cost, time, capacity, and quality of services or any combination. A subjective view can be the baseline or starting point—all be it a perception or view. Having a subjective point of view to progress to a more objective point of view on transportation factors that are critical to the performance of the transport process. Sometimes samples can be used but may not represent the overall population of the activity.

Subjective inputs could be anecdotal information suggesting that “costs are going up” or “shipments are more frequent.” These are usually suggested by individuals involved directly or indirectly in transport moves.

Conclusions—How do You Measure?

With the rapidly changing global environment we are facing new challenges and it becomes extremely important to evaluate the transportation planning systems and execution systems to similar companies in your industry and compare to other industry sectors. This becomes critical with such rapidly changing requirements due to factors such as influences of COVID-19 pandemic where material handling and human interaction around products and equipment within warehouse internal operations now must be rethought to include this additional risk factor.

The case for benchmarking suggests that a particular process in your firm can be strengthened. Some organizations benchmark as a means to improve discrete areas of their business and monitor competitors' shifting strategies and approaches. Regardless of the

motivation, cultivating an external view of your industry and competitors is a valuable part of effective management practices in a world that is constantly changing. The first step to evaluate practices must be to determine any gaps or opportunities to improve incorporating the changing conditions. Something to consider when benchmarking current activities is the level of comfort with the current system and the likelihood that changes can be adopted in a positive manner.

It is always helpful to determine what is “best in class” for like companies in order to compare a firms’ practices to a “best in class” level. This can be done by internal or external resources. Some corporations set process standards for all divisions to achieve based on best in class benchmarking. Determining the gaps between the current system and the best in class can create a roadmap to identify the gaps to close between the current system utilized and a best in class system.

When looking at measuring the performance of a transportation system it is important to understand the key drivers that need to be measured. There are usually a few primary and secondary metrics that will drive the majority of the behavior you are modeling. These metrics are distance, weight, type of commodity, as well as supply/demand characteristics of the macro-market for the transportation route. Measuring key performance indicators should be the activities or aspect of the transport that are driving the critical behaviors. These factors can be identified both subjectively as well as objectively.

The impact of the COVID-19 virus influences all activities. The new risk factors that need to be considered will require a rewrite of the playbook. Regional supply chains will become much more attractive over global supply chains in order to limit risk factors in today’s environment.

References

- Arroyo, P.E., Holmen, E., 2017. Integration in loosely coupled garment supply chains. *J. Global Oper. Strategic Sourcing*: Bingley 11 (3), 357–383, doi:10.1108/JGOSS-10-2017-0042.
- Bolstorff, P., Robert, R., 2007. AMACOM-American Management Association, ISBN-13: 978-0-8144-0926-8.
- Business and Financial Times, April 28, 2020. <https://www.msn.com/en-za/news/other/covid-19-and-the-logistics-transportation-industry/ar-BB13kX6k>.
- Coyle, J.J., 2011. *Transportation: A Supply Chain Perspective*. Thomsen-South Western.
- FreightWaves, Newstex Finance & Accounting Blogs, 2020-05-21.
- Herden, T., 2020. Explaining the competitive advantage generated from analytics with the knowledge-based view: the example of logistics and supply chain management. *Business Res. (Göttingen)* 13 (1), 163–214, doi:10.1007/s40685-019-00104-x.
- Lehmacher, W., Sharma, S., 2020. Smart Supply Chain Decisions in the New Normal, Supply and Demand Chain Executive, June.
- Christopher, M., 2011. *Logistics and Supply Chain Management: Creating Value-adding Networks*, 4th edition. Prentice Hall/Financial Times, New York. ISBN-13: 978-0-273-68176-2.
- Christopher, M., 2005. *Logistics and Supply Chain Management: Creating Value-adding Networks*, 3rd edition. Prentice Hall/Financial Times, New York. ISBN-13: 978-0-273-68176-2.
- Reichhart, A., Holweg, M., 2007. Creating the customer-responsive supply chain: a reconciliation of concepts. *Int. J. Oper. Prod. Manage. (Bradford)* 27 (11), 1144–1172, doi:10.1108/01443570710830575.
- Reh, J., 2019. The Importance of Benchmarking in Improving Business Operations, July 26. <https://www.thebalancecareers.com/overview-and-examples-of-benchmarking-in-business-2275114>.
- Wang, H., Nozick, L., Xu, N., Gearhart, J., 2018. Modeling ocean, rail, and truck transportation flows to support policy analysis. *Maritime Econ. Logist. (London)* 20 (3), 327–357, doi:10.1057/s41278-018-0108-x.

Public Transport Network Planning

Corinne Mulley, John D. Nelson, The University of Sydney, Sydney, NSW, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	388
Objectives and Purpose of Network Planning	388
The Role of Economics and its Impact on Network Planning	389
Objectives for Public Transport	389
Principles of Planning	389
Areas of Low Density and Demand	392
Issues in Planning	392
Conclusions	392
Biography	393
References	393
Further Reading	394

Introduction

This article provides a framework for understanding the principles and good practice in public transport network planning. It begins by considering the objectives and purpose of network planning before examining the role of economics in underpinning network planning activities and the higher level objectives of public transport provision. The entry then turns to the principles of network planning, including a discussion of timeframes and the necessity of understanding how network planning outcomes fit into the demand for travel by the end user. This is followed by discussing the adjustments necessary to provide good network planning for areas of low demand. The penultimate section examines issues and areas where more attention is needed in the future. The entry ends with a short contribution on the role of public transport in the future, in the post-COVID-19 world.

Objectives and Purpose of Network Planning

The nature and scope of network planning tasks are partly defined by the institutional set up of the area under consideration and partly shaped by the mode being planned. However, in defining the scope it is imperative that the area under consideration be carefully defined and that the aspirations for the level of service be clearly outlined. The planning needs to understand how existing resources contribute to the different goals of network planning and recognize that planning in the short, medium and long term have different issues.

The definition of network planning tasks need to be sensitive to the goals and objectives which should include allowing public transport operators to provide services that customers are willing to pay for, should support social policy as well as providing efficient transport and contribute to an overarching goal of transport sustainability. At all stages of network planning, it is important to define the indicators of success that relate the aspirations of the objectives to the outcome of the network planning. Monitoring change is necessary and the definition of key indicators at the planning stage is important if post implementation evaluation is to be targeted.

Public transport networks influence land use structure and vice versa. In network planning, these need to be positively considered so that land use and public transport strategies become mutually reinforcing. This is seen in cities where new public transport infrastructure is being established when, for example, a new metro is associated with the building up of densities around proposed stations or where new stations are associated with deliberate transit-oriented design (TOD) developments to exploit the greater accessibility of the new transport link. It is also seen where new land use developments, for example, the building of a new housing development, where new public transport links need to be provided if the development is not to become car-centric. In particular a strong public transport—land use strategy will connect the component parts of the area under consideration so that public transport can shape urban growth. This is because the market will recognize how public transport is useful in providing necessary connections. One way in which a public transport network can be extended is through the use of complementary modes such as micromobility solutions such as (e)bicycles and scooters and carpooling for the “first and last mile” (1LM) of a journey.

The institutional framework which determines how public transport is organized and which body is responsible for its delivery does make a difference to the objectives and purpose of network planning. The institutional framework is really a continuum going from one extreme of market competition to another extreme of coordinated planning. In the market competition end of the continuum there is little role for network planning with operators determining the price and quantity mix of public transport provision. As a result, users will be faced with large differences in service levels as operators will provide more services on routes with high patronage. Experience of market competition in bus services (in the UK and in New Zealand) led to excess capacity being

provided on “good” routes, increased innovation for services which added to the operators’ “bottom line”, more efficient production, but a network that was not stable and for which it was difficult to provide integrated ticketing and comprehensive passenger information. At the other extreme of coordinated planning, there is a significant role for network planning and the objective of public transport being a viable alternative to the car can be met through the provision of an integrated network with operators being able to exploit economies of scale. Whilst coordinated planning allows the development of long-term planning often based on the cross subsidy between routes it also provides the opportunity for the development of monopoly outcomes and weak market incentives. In reality, institutional frameworks tend to be located somewhere along this continuum with network planning being more relevant to the coordinated planning—sometimes called the public service—approach.

The institutional framework also has other impacts on the success of network planning. The coordination of the bodies responsible for public transport regulation and delivery is an important constraint on success. Regional organization of public transport is often cited as an important success factor but it is in fact a necessary but not sufficient factor (it is possible to have regional organization but fragmented coordination which gives rise to less favorable public transport planning outcomes). Network planning should at a minimum be concerned with the spatial area of the labor market if it is to avoid cross boundary issues for users who are commuting. An institutional framework will be more successful if it works to provide supportive and complementary policies to reinforce the overarching transport policy—often called “push and pull” policies—such as expensive or limited car parking in an urban center. Finally, stable funding is a pre-requisite for good network planning as users are not sympathetic to frequently changing networks.

In summary therefore, networks that perform well are characterized by sufficient stability so that the public transport system can influence urban development and allow users to create and maintain sustainable transport patterns, a robust and simple structure which provides public transport options for all citizens.

The Role of Economics and its Impact on Network Planning

Economics is fundamentally about the allocation of resources. For individuals making trips, they are broadly trading time and money when making their decision as to how to travel. For example, if a journey by public transport was to take ten times longer than by car it is likely to become a car journey. Car journeys may be shed in favor of a public transport trip if there is no parking at the destination or if parking is expensive. Governments have a higher level role in ensuring mobility for their community is undertaken in ways that meet the higher objectives of policy for example, in relation to congestion or environmental pollution. Governments also have to make the connections that are unlikely to be made by individuals, in particular the links between land use and transport planning and the links between public transport and health.

Individuals are typically only aware of the trips they make. They are not usually aware of the network that makes up the public transport system—the network that provides accessibility to destinations that are part of daily activities. In contrast, governments should have a commitment to the network as a whole and focus on enhancing the “network effect” when undertaking network expansion or improvements so that new projects and enhanced corridors add more to the network than if taken in isolation.

Objectives for Public Transport

Public transport policy objectives are multidimensional. They encompass economic aspects, such as the reduction in congestion through mode shift away from car travel, environmental sustainability as the transport sector is a significant contributor to greenhouse gas emissions and social/quality of life aspects including the positive health impacts from public transport use (more activity, walking to access public transport, and egress to final destination) and equity, social justice, and engagement impacts on social inclusion.

People choose to use public transport when it is high frequency, when the journey times compete well with a car journey and when they cannot park at their destination; that is when the experience can be made as seamless as possible. Frequency is important because network planning is all about understanding the coverage versus frequency trade off. If budgets are fixed, then more frequency means less coverage and vice versa. This links of course to the objectives for public transport. If an objective is patronage growth or environmental impact reduction, then the network planning must deliver higher frequency since this is most likely to get travelers out of their cars and onto public transport (Currie and Wallis, 2008). If an objective is to have greater equity or social inclusion, then the network planning must deliver coverage. To meet both objectives means that network planning needs to find the optimum balance between frequency and coverage, ensuring that individuals losing coverage are provided with accessibility in a way alternative to conventional public transport services.

Of course, the trade off between coverage and frequency flows through to performance measurement which only makes sense relative to the goals and objectives of the public transport policy.

Principles of Planning

In the short-term, network planning focuses on buses. In multimodal systems, in the short-term any rail based system is fixed: the timetable can be changed, the frequency and stopping patterns can be changed but not the route. So for urban areas in particular, bus

provides the flexibility to meet changing or new “needs” of accessibility in the short-term. Hence, the principles of this section relate to network planning for the bus network.

There are two approaches to network planning—the tailor-made approach and the ready-made approach (Nielsen et al., 2005). The tailor-made approach fits the needs of different users and fits the needs of different times of day whereas the ready-made approach provides a simple and stable network of lines throughout the day but varies the frequency and vehicle size according to the needs of different users and times of day (see Fig. 1).

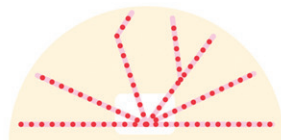
The Tailor-made approach



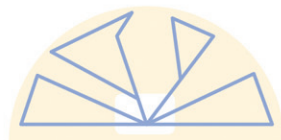
A dense, normal frequency network most of the day



Reduced and low frequency network on holidays



Express lines in peak periods

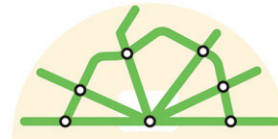


Evening and night lines



Service lines for the elderly and disabled

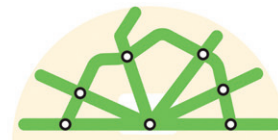
The Ready-made approach



A basic high frequency network most of the day



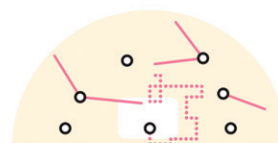
Same network, reduced frequencies on holidays



Same network, higher frequencies in peak periods



Same network, reduced frequencies evenings and nights



Local lines and demand-responsive services for all users



One stable, easy-to-use network for all at all times

Figure 1 Two different approaches to network design (Nielsen et al., 2005, p. 35).

The problem with taking the tailor-made approach is that services and networks can be very complicated to supply and an inefficient use of the total resource for public transport. Moreover, for the user, it can become complicated to understand with different networks from the dense, normal frequency network that exists for most of the day, a reduced network at weekends, a different but also reduced network for evening and night buses and so on. This has led to the development of the ready-made approach which develops the network in two stages. The first is the development of the trunk routes—those lines within the network which will have high frequency because of heavy demand and which will attract priority measures on the road network. This stage requires the concentration of available resources. The second stage is to provide more dispersed and flexible (demand responsive) services so as to serve all citizens. The major task of network planning is to find, for the available budget, the right balance between these—the frequency and coverage trade off discussed above. The second stage of development is discussed in the next section and is concerned with areas of low density and demand.

In the first stage of development, the ready-made approach argues for the concentration of resources on corridors, using timetable coordination to increase the frequency as density increases. Concentration of resources means removing the “wiggly” routes so that, for example, two low frequency lines which operate in close proximity should be amalgamated to provide one more direct route with enhanced frequency and no change in resources expended. The design principle is “one section one line” so as to provide a simple network, moving away from the single seat approach whereby many destinations are served by lines with no transfer. The “one section one line” approach has operational advantages since an operational disturbance, for example a traffic accident, only affects that section and is not transmitted to the rest of the network. The network planning skill in the “one section one line” approach is to link the sections so as to have the minimum number of routes. The best way of doing this is to link sections which have similar demands on different sides of the urban area center so that services do not have to terminate in the urban center where space is at a premium and where termination would simply add to congestion. Simplicity also saves resource with wiggly routes with many stops costing up to 250% more than a straight line with centrally placed stops (Nielsen, 2005, p. 133).

This first stage of development of the network requires a good frequency on these trunk lines. There are three main classes of frequency (Balcombe, 2004). The “forget the timetable” frequency is one whereby users do not need to consult the timetable and they arrive at stops randomly; in London, UK, this has been empirically estimated to be between 6 and 12 departures/h. The “fixed minute” timetable is reserved for lower demand lines and in Norway has been empirically determined as an hourly service. The final category is the frequency for flexible on-demand services where demand cannot support an hourly service. Each of these different frequency types have operational necessities: the “forget the timetable” frequency must have services which are spaced at equal intervals and the “fixed minute” service must keep to its timetable. The approach to flexible transport services is discussed in more detail below.

Seeking to provide an optimum frequency for trunk lines is another trade off for the network planner. In the peak, a “forget the timetable” frequency is essential. Increasing the frequency much beyond this can give rise to diminishing returns with smaller headways potentially leading to bus on bus congestion. Typically, if demand is such that a headway of 2 min or better is needed, then it is a better alternative to open up a new line, linking centers which are not currently served by direct routes. This not only releases pressure on the existing lines as well as providing alternative travel destinations but also has the impact of growing the network through exploiting the network effect. An alternative way to release pressure is to plan to introduce a high capacity route in the form of Bus Rapid Transit (BRT) infrastructure although this is not usually feasible in the short-term; a more recent compromise is “BRT lite” which is less likely to include fully segregated busways.

Whilst difficult to do effectively, network planning should seek to exploit the timetable in a way that as lines converge toward the city center, the frequency of service increases. The skill here is to organize the services from the outer suburbs which are necessarily less frequent, to build to a service pattern of equal intervals for users as the services from suburbs converge toward the city center.

In the ready-made approach, frequency is the key variable used to adjust for demand. Operators will supply more frequency in the peaks, reduce the frequency in the peak shoulders, and reduce again for evenings and holidays. This way a simple and stable network is maintained that meets the needs of users.

It is worth noting that if frequency is good, walking further to a stop will not have a big impact on the combined walk and wait time. Indeed, there is evidence that users will walk further to a more frequent stop (Mulley et al., 2018). It is also important to have good frequency across the network—both radially and across the city—this speeds up journeys and makes interchange between lines short.

The “elephant in the room” of the ready-made approach to network planning is its reliance on interchange. With a simple, stable network using the “one section one line” principle, it must be accepted that interchange will be needed between lines operated by the same mode and, in multimodal systems, between lines operated by different modes. Interchange, of course, allows the best of the mode to be exploited and transfer to another mode when it is better at providing the desired service. Accepting interchange means that services do not need to concentrate on providing single seat journeys and allows greater coverage. The planning skill is to maximize the number of direct journeys by knowing what passengers want and having high enough frequency that interchange is not a barrier to travel.

Interchange can be used to release resources. Bearing in mind that super high frequency can lead to bus on bus congestion, a tailor made network can be redesigned so that a previous design of single seat services from a number of suburbs, all converging to a common route at the city center is approached, becomes one where interchange is made at the start of the common route. This can release resources to improve the frequency of lines to the suburbs and provide a frequency that does not give rise to congestion as the city center is approached. Even with having to interchange, total travel does not need to increase as shown by Nielsen (2005, p.110).

Recognizing that transfers and interchange are necessary in a ready-made approach means that network planning should endeavor to minimize the cost of interchange. This can be done by ensuring timetable coordination, ensuring that route information is presented accessibly, removing fare penalties and creating short and easily understood interchanges. The empirical evidence is that higher levels of transfer are associated with higher public transport modal shares—showing the benefit of the network effect (Terzis and Last, 2000).

Areas of Low Density and Demand

In the second stage of development, the network planning strategy needs to take an alternative approach. Where there is low density and/or low demand, a strategy of a high frequency, stable network will not work: users would need to walk considerable distances to stops and providing high frequency would be an inefficient use of resource.

Strategies such as the “hub and spoke” approach of airlines can be used successfully when arrival and departure times to key interchanges are planned. More personalized, flexible transport services can be planned, typically using smaller vehicles, as an important part of the overall network.

Flexible transport services (FTS) can be used to feed into a planned network of fixed route services. FTS are usually operated with smaller vehicles and are provided on-demand. It has been argued that the family of FTS can be extended to include taxis and private hire, on demand car services such as Uber and Lyft and car pooling (Nelson and Wright, 2016). Flexible feeder services can help to secure the viability of the fixed route service as well as extending accessibility beyond the walkable catchment of the fixed route service. Typically, subsidy per passenger for this type of service is high even though the cost of operating smaller vehicles is lower than that of the “big bus” on the conventional fixed route. CallConnect in Lincolnshire, England is an example of a successful service. Established in 2001 and continuing to the present, CallConnect is used to feed a quality network of fixed route local bus services designed to improve transport links to destinations throughout the county. The 6 FTS feeder services offer guaranteed connections (with the availability of a taxi as a last resort) and the service offering also includes 14 separate area-based FTS services. Importantly, the flexible feeder services help secure the viability of the fixed route services and have resulted in the level of unmet need in the communities served being reduced by 90%. CallConnect (described in Wang et al., 2015) replaced a less effective fixed route service when it was introduced and the subsidy required by CallConnect is approximately the same as was required for the fixed route services it replaced.

Issues in Planning

The nature of public transport is changing as the boundaries between traditional fixed route services and private shared modes increasingly blurs, particularly as the new mobility modes and the influence of the sharing economy (manifested *inter alia* as bike-, car- and ride-sharing) become more prevalent. Passengers continue to show an increased preference for App-enabled personalization of travel opportunities (Nelson and Wright, 2019). The emerging Mobility as a Service (MaaS) context is helping to shape the future landscape for public transport and making the case for public transport to be situated at the core of MaaS offerings (Hensher et al., 2020).

Providing a network of lines is only part of the network planning task. However, some tasks are only possible in the longer term (more than 5 years), such as urban and regional structural planning and investment planning for new infrastructure. Other tasks, such as budget planning, fares policy and operational aspects are needed in the short-term to facilitate the delivery of the network plan. Changing the network plan takes time (between 2 and 5 years) and should be undertaken along with an analysis of the previous network from monitoring data and with information on then current mode choice behavior.

Conclusions

Good network planning should take account of the urban structure in locating demand, existing infrastructure and networks. The creation of a ready made network providing a simple and stable structure will build on the mutually supporting planning of land use and transport. In all cases, it must be remembered that the journey on the vehicle is only a small part of the overall journey so far as the traveler is concerned.

It is important to recognize that whilst interchange is seen as a penalty, a network plan predicated on interchange can provide access to more destinations with fast journey times. As a result, interchange based network plans can provide a higher use of public transport overall.

Whilst outside the scope of this entry, it is worth noting that different topographies can cause constraints: urban areas bisected by rivers, terrain subject to many changes in gradient. Also the institutional arrangements of the area—its history, its politics—can also act as a constraint on what can be done.

At the time of writing the world has experienced a sudden collapse in trip making behavior on account of the COVID-19 global pandemic. Following the introduction of stay at home directives global public transport usage has collapsed; Australian public transport usage, at April 2020, is 80% lower than 12 months earlier and in the UK, public transport usage is more than 90% lower in

some areas. With this comes a consequent fall in revenue for public transport with risk shared amongst different parties depending on the institutional framework but ultimately most of the risk will fall on governments.

It is clear that public transport patronage is not going to quickly rebound to pre-COVID-19 levels; this is primarily because of perceived anxiety over the potential health hazard of being in close proximity to others. Social distancing on public transport is difficult for a system designed to be mass transit. Advice on physical distancing varies and ranges from 1 to 2 m, depending on local and national contexts. But even with more relaxed distancing (1 m distance + gap between people in each seat row) a capacity of around 30% for buses and 50% for trains may be the best outcome that can be achieved. If public transport is to retain as many former passengers as possible, the traveling peaks need to be spread although it will be challenging to accommodate staggered start times be it for education, retail or workplaces. Earlier work on spreading the peak may be relevant here with, for example, a study of peak spreading in the University sector can be found in [Daniels and Mulley \(2012\)](#). Other measures that may be introduced include a move to rear door boarding, temperature checks, banning of cash payments, marshalling of people queuing, use of masks (where recommended) and provision of hand sanitisers. It is also likely that globally there will be greater working from home where this is possible.

There is compelling evidence that liveable cities need good public transport (achieved by good network planning) and the current imperative is to ensure that future public transport is safe, comfortable, and convenient. There may need to be a pause on larger mass transit infrastructure projects in the shorter-term in favor of a greater focus on enabling seamless intermodal journeys.

Biography



Corinne Mulley was the inaugural Chair of Public Transport at the Institute of Transport and Logistics Studies at the University of Sydney. She is now Professor Emerita at the University of Sydney and an Honorary Professor at Aberdeen University. Corinne is a transport economist and is active in transport research at the interface of transport policy and economics, in particular on issues relating to public transport. She has provided both practical and strategic advice on transport evaluation, including economic impact analysis, benchmarking, rural transport issues, and public transport management. Corinne's research is motivated by a need to provide evidence for policy initiatives and she has been involved in such research at local, state/regional, national/federal and European levels.



John Nelson is Chair in Public Transport at the Institute of Transport and Logistics Studies (ITLS), University of Sydney which he joined in 2019 from the University of Aberdeen where he was Sixth Century Chair in Transport Studies and Director of the Centre for Transport Research. Before moving to Aberdeen in 2007 he was Professor of Public Transport Systems at Newcastle University, UK. John is particularly interested in the application and evaluation of new technologies to improve transport systems (with a particular focus on public transport and shared transport solutions) as well as the policy frameworks and regulatory regimes necessary to achieve sustainable mobility.

References

- Currie, G., Wallis, I., 2008. Effective ways to grow urban bus markets – a synthesis of evidence. *J. Transport Geogr.* 16, 419–429.
- Daniels, R., Mulley, C., 2012. The paradox of public transport peak spreading: universities and travel demand management. *Int. J. Sustain. Transport.* 7 (2), 143–165.
- Hensher, D.A., Mulley, C., Nelson, J.D., Ho, C., Smith, G., Wong, Y., 2020. *Understanding Mobility as a Service (MaaS)*. Elsevier, Past, Present and Future.
- Mulley, C., Ho, C., Ho, L., Hensher, D., Rose, J., 2018. Will bus travellers walk further for a more frequent service? An international study using a stated preference approach. *Transport Policy* 69, 88–97.
- Nelson, J.D., Wright, S., 2016. Flexible transport management. In Bliemer, M., Mulley, C & Moutou, C. (Eds) *Handbook on Transport and Urban Planning in the Developed World*, Edward Elgar, pp. 452–470.
- Nelson, J.D., Wright, S., 2019. Themed volume of on "The Future of Public Transport". *Res. Transport. Business Manage.* 32, 1–2.
- Nielsen, G., Nelson, J., Mulley, C., Tegner, G., Lind, G., Lange, T., 2005. *Public transport—planning the networks*. HiTrans Best Practice Guide No. 2. ISBN: 82-990111-3-2.
- Terzis, G., Last, A., 2000. *Guide urban interchanges: a good practice guide*, final report. European 4th RTD Framework Programme.

Wang, C., Quddus, M., Enoch, M., Ryley, T., Davison, L., 2015. Exploring the propensity to travel by demand responsive transport in the rural area of Lincolnshire in England. *Case Stud. Transport Policy* 3 (2), 129–136.

Further Reading

Mees, P., 2000. *A Very Public Solution. Transport in the Dispersed City*. Melbourne University Press, Melbourne.

Mulley, C., Nelson, J.D., Ison, S.G. (Eds.). 2021. *Routledge Handbook of Public Transport*, Chapter 32. Routledge, Abingdon, Oxon.

Nielsen, G., Nelson, J., Mulley, C., Tegner, G., Lind, G. & Lange, T. 2005. *Public transport—planning the networks*. HiTrans Best Practice Guide No. 2. ISBN: 82-990111-3-2.

A Timely Perspective on Planning for Ageing Infrastructure

Anthony Perl, Urban Studies Program, Simon Fraser University, Vancouver, BC, Canada

© 2021 Elsevier Inc. All rights reserved.

The Need for a Timely Perspective on Planning Transport Infrastructure	395
Adapting the Half-Life Concept to Reassess Transport Infrastructure	396
Taking Stock of the Relationship Between Durability and Flexibility in Transport Infrastructure Development	396
Reconciling Durability with Flexibility: A Strategy for Developing Resilient Transport Infrastructure	397
Conclusion	398
Acknowledgment	398
References	398
Further Reading	399

The Need for a Timely Perspective on Planning Transport Infrastructure

The prevailing paradigm in transport planning is often said to be characterized by an imperative to “predict and provide” (Banister, 2008). This objective quickly narrows to “provide” – building new infrastructure to meet projected demand for mobility in order to prevent future congestion. Meanwhile, the “Sustainable Transport” planning paradigm has highlighted the need to “green” the technology supporting mobility and adjust travel behavior to reduce the ecological impacts of mobility (Schiller and Kenworthy, 2018), but has not adopted any major changes in the design and planning of transport infrastructure. Alongside calls to invest less in road construction, sustainability advocates usually call for investing and building more public transport infrastructure, particularly railway infrastructure (Banister et al., 2011). But while adjusting the mix of infrastructure supply is seen to advance sustainability, transport planners have yet to consider building new infrastructure differently. This is an overly narrow approach for what needs to be done when demand for new mobility grows, the main trigger of initiatives to supply more transport infrastructure.

The current mobility landscape is very dynamic, and at times even volatile. Rapidly increasing use and promotion of non-motorized transport especially in European cities juxtaposed against the major increase in motorization levels in developing economies are two examples of disruptive trends, although in opposing directions along the sustainability spectrum. Jones (2016) for example describes the evolution of cities in North-west Europe from predominantly being car based to being more people based and this change is explained in greater detail for some European cities by Buehler et al. (2017). On the other hand, Ma et al. (2012) forecast that the growth in demand for motorized transport in China is expected to continue expanding at a fast pace.

Recent technological developments and the incorporation of information and communication technologies (ITC) into transport operations have been driving changes in mobility, and more broadly in society. The rise of shared mobility transport options delivered by companies like Uber as well as the spread of the car-sharing phenomena could transform the ways that transport infrastructure will be used in the future. Furthermore, the adoption path and pace of autonomous vehicles, which are widely recognized as a disruptive technology will likely reshape our future mobility demand and consequently our transport system (Guerra, 2016). From ride hailing to drones to autonomous motor vehicles, these new mobility technologies are still in full flux, which adds to the uncertainty currently facing transport planners. This predicament is often referred to as “deep uncertainty” (Lyons and Davidson, 2016) and succinctly explained as “we know only what we do not know” (Wall et al., 2015).

In scholarship about “planning for the future,” a new strategy for managing change under conditions of deep uncertainty has emerged according to which “a planner should create a strategic vision of the future, commit to short-term actions, and establish a framework to guide future actions. A plan that embodies these ideas allows for its dynamic adaptation over time to meet changing circumstances.” (Haasnoot et al., 2013, p. 485). This new planning approach, referred to as Dynamic Adaptive Planning (DAP) is well designed to deal with transport infrastructure, given its inflexible attributes and long lifespan. Yet this approach to developing dynamic and adaptive attributes usually refers to the plan for delivering infrastructure and not its functional design. Wall et al. (2015), for example, describe a DAP for the San Francisco–Oakland Bay bridge replacement in which the infrastructure function remained fixed. Extending the DAP process to infrastructure redesign could have opened possibilities to make the structure serve “dynamic and adaptive” mobility functions in the future.

The rigid characteristics designed into most transport infrastructure constrain both efforts to transition to low-carbon mobility (Givoni and Banister, 2013) and to advance the “sustainable mobility paradigm” (Banister, 2008). Instead of automatically reinforcing existing technology and modal functions, new modes of thinking about an infrastructure’s future can offer the opportunity to adapt original designs and thus facilitate sustainability transitions. Innovation can be triggered by a formal assessment threshold discussed in the next section.

Table 1 The Four R Framework for transport infrastructure adaptation

Strategy	Description
Renew	Refurbish existing infrastructure to continue its current transport function. May include an upgrade to increase mobility.
Redesign	Enable infrastructure to support different forms of mobility through accommodating new transport technology and/or modes.
Repurpose	Decommissioning infrastructure from its original mobility function and using the structure and its surrounding space for activities other than transportation.
Remove	Dismantling the infrastructure and removing it from the previously occupied space.

Source: Adapted from *Givoni and Perl (2017)*

Adapting the Half-Life Concept to Reassess Transport Infrastructure

Ernest Rutherford received the Nobel prize in Chemistry for discovering that the decay of radioactive isotopes can be measured through a half-life ([Rutherford, 1900](#)). This is the point when the isotope has emitted half of its radioactivity. Such a half-life measure can also reshape our understanding of transport infrastructure by highlighting the change in its physical condition and underscoring the need to reassess its future. Like radioactive decay, the deterioration of transport infrastructure is unavoidable, and the half-life offers a focus that can trigger planning for what to do about infrastructure degradation.

Once built, any transport infrastructure begins to degrade and loses some of its utility, either because it cannot support as much mobility as it did upon inauguration, or because demand declines for the mobility using that infrastructure. Action should be taken before these trends reach their logical conclusion, in order to extend the useful life of that infrastructure. Four distinct strategies to meet the future are possible, which together have been labeled the 4R Framework for adapting transport infrastructure ([Givoni and Perl, 2017](#)). These approaches are outlined in [Table 1](#). The options include: *Renewing* the infrastructure to extend its physical life while preserving its original purpose and design; *Redesigning* the infrastructure to broaden its function to supporting a different mix of mobilities; *Repurposing* the infrastructure to transform its function toward supporting a non-transport use; or *Removing* the infrastructure completely, and thus clearing the space it occupied for other activities and new structures, or even returning the space to a natural environment.

Keeping infrastructure in a state of good repair to prolong its lifespan is a well-established transportation management practice. The most common strategy for maintaining infrastructure's future value is thus to *Renew* its existing mobility capability. But as future mobility demands evolve, and new technology diffuses, going beyond refurbishment to introduce different forms of capacity for mobilities can arise through infrastructure *Redesign*. Sometimes, this strategy accommodates multiple modes on the same infrastructure. At other times, a new mode is substituted for the original one, changing the role which that infrastructure was intended to play.

Deciding to redistribute the space that had been occupied by infrastructure to another purpose besides transport can create a new sense of place, often enabling the surrounding space to be redeveloped in ways that would have been incompatible with most mobility infrastructure. This approach is embraced by the strategy to *Repurpose* infrastructure for non-mobility uses. And an even more dramatic shift can occur when transport infrastructure is completely dismantled and extracted from the space that it once occupied, following a decision to *Remove* it.

Each alternative scenario should at least be considered as transport infrastructure approaches its half life. Then, the most plausible and desirable future can be identified, and the adaptive path toward some degree of change in that infrastructure's form and function can be planned out. Realizing greater levels of adaptation can be facilitated by decoupling the base of transport infrastructure from the superstructure over which mobility operates on it. This attenuation enables an optimization of durability and flexibility, attributes which are both needed to enhance infrastructure's future value and which are discussed in the next two sections.

Taking Stock of the Relationship Between Durability and Flexibility in Transport Infrastructure Development

Ancient societies have revealed little concern about the costs of building monumental infrastructure that would last forever, such as the Egyptian pyramids. Roman highways were also built to last, and small stretches of this infrastructure are still used ([Berechman, 2003](#)). But once economists introduced the practice of discounting infrastructure costs and benefits ([Parker, 1968](#)) it became common for infrastructure to be built for a fixed lifespan—after which benefits were not to be counted on. This time frame is typically set at 30–60 years after construction.

Even if its lifespan is not eternal, infrastructure's indivisibility and irreversibility do require a significant up front financial commitment. These attributes justify building *durable* infrastructure to last many decades. This same logic can also justify oversupplying infrastructure, through designing a larger than needed capacity for example, in order to meet future demand and thus avoid rising construction costs ([Zembri and Libourel, 2017](#)).

Over time, new transport technologies supplement earlier ones in meeting mobility needs, ([Grübler, 2004](#)). However, a new transport technology does not completely displace its predecessors, at least not immediately. For example, not only are railways still

operating in the “automobile age,” in recent decades they have experienced a resurgence that some have termed the second railway age (Banister and Hall, 1993). Gröbler’s (2004) analysis reveals how new transport technology contributes to meeting mobility needs through an incremental addition to established capacity that usually builds up over time.

Since each mode’s infrastructure is designed differently, investing in the upkeep of all this accumulated infrastructure often proves to be wasteful. This became evident with the decline in demand for rail travel during the mid-20th century. The development of the “car society” in the United Kingdom revealed the costs of the Victorian exuberance for building railway infrastructure to last forever. Designing durable infrastructure thus means creating structures that are difficult to adapt in the face of change.

What seemed unimaginable to Victorian railway builders, turned out to be inevitable – growth trends do not go on indefinitely. Recent evidence suggests that a similar dynamic of change could be in store for the automobile – new technologies (e.g., internet), new behaviors (e.g., ride sharing), and demographic changes appear to be shifting demand (Polzin and Chu, 2014; Stapleton et al., 2017).

Durable infrastructure also works to lock in spatial and social effects across a society. Motorways that enabled Americans to travel further and faster in the 20th century have embedded spatial and social outcomes that continue to reinforce automobile dependence during the 21st century (Squires, 2008). Such lock-in effects are compounded by the reluctance to write off fixed assets like motorways and railways even when their productivity is much reduced and external costs are growing. But periodically, concentrated change in mobility does occur.

Gilbert and Perl (2010) have defined a transport revolution as a substantial change in passenger or freight movement that occurs in less than 25 years. Such tipping points are catalyzed by disruptive transport technologies or radical reorganization of business models (e.g., deregulation) and they drastically change infrastructure’s required attributes in a short-time period. Thus, in times of rapid change, the benefit of flexibility in infrastructure design increases, and durable infrastructure’s inflexibility decreases its value by constraining its adaptation. Building flexibility into infrastructure through planning and design holds the key to avoiding both the sunk cost trap and the lock-in effects that impose future burdens on society. Thus, infrastructure that retains its value after a transport revolution can be judged to be both durable and flexible. Creating the capacity for such an optimal balance of durability and flexibility requires reconciling these two fundamental infrastructure design attributes.

Reconciling Durability with Flexibility: A Strategy for Developing Resilient Transport Infrastructure

While the challenges of developing infrastructure that can retain its value over time might appear daunting, a planning strategy that can adapt to alternative futures that emerge after the infrastructure’s half-life arrives shows considerable promise. With constraints on economic productivity and profit growing over time, investing in renewal and expansion of transport infrastructure becomes harder. This dilemma has been recognized both in transport and broader economic studies. Gifford (1994) claimed that rigid infrastructure is overvalued in urban transportation planning while Dixit and Pindyck (2000) discussed the burden created by overinvestment in a firm’s production capacity. The goal of reconciling durability with flexibility in transport infrastructure is thus both apparent and necessary. Looked at strategically, investing in transport infrastructure would make more sense if that infrastructure could be designed to be flexible and adaptable to changing needs.

An effective approach to adaptive infrastructure planning and design that combined flexibility with durability can be found in the 19th century French practice of separating the responsibility for developing transport infrastructure and superstructure. Walton (1998) details how the French state, which possessed a highly capable engineering bureaucracy, the *Corps des Ponts et Chaussées*, took on responsibility for designing rail rights of way and building their subgrades, bridges, and tunnels. This infrastructure was then transferred to private management on long-term leases, which enabled private carriers to finance and build the tracks, signals, and stations required for railway operations. *Superstructure* became the legal term used to differentiate these private fixed assets from the publicly planned and owned right of way that was identified as *infrastructure*.

By rediscovering this physical and organizational distinction in the 21st century, transport infrastructure can once again obtain the advantages from durability while the superstructure connected with it gains the flexibility to support more than one transport technology, serve multiple modes, or even enable a mix of transport and nontransport functions. Ongoing adaptation of durable infrastructure and flexible superstructure can be facilitated by enabling separate organizations to design and plan these components before and after the half-life point in time arrives. Such a strategy can accommodate adaptation to changing technology, demographics, and societal needs, as the demand for mobility evolves.

The value of Victorian rail infrastructure been enhanced using such an adaptive redesign strategy. Some railway bridges no longer used by trains have been converted to serve road transport, by reconfiguring their superstructure. And when rail travel demand started growing again in the 21st century, some decommissioned rail rights of way, which had been retained were reactivated with a new rail superstructure. Similarly, the ‘Rails to Trails’ program has adapted derelict US rail infrastructure by introducing superstructure for walking and cycling. The resulting trails preserve the transport right of way that could be further adapted to support future mobility needs (see for example Beiler et al., 2015).

Superstructures can become flexible, even without physical redesign, by organizing different uses at different times. Examples include repurposing urban motorways on Sundays for walking, cycling, and non-motorized transport, which started in South American cities (known as *Ciclovías*), and similarly in Paris for the summer holidays along the Seine (Bertolini, 2017, p. 93).

Assessing the trade-offs between durability and flexibility in each of the 4R framework’s strategic scenarios can benefit from accounting for the option value and opportunity cost of alternative designs (Litman, 2009, pp. 1-13; OECD, 2006, p. 21). Building

durable infrastructure increases the opportunity cost of removing it later, but can also augment the option value of enabling an adapted superstructure in future, even if such a design might increase initial costs. The opportunity cost of not enabling a flexible superstructure to permit future Repurposing or Redesign ought to be considered when designing new infrastructure. In a similar way, the option for removing infrastructure, because it might not be adaptable to future mobility, needs to be kept in mind.

Optimizing designs for option value and opportunity cost led to some of the infrastructure built for the 2012 Summer Olympics in London to be dismantled after the games. This strategy saved both construction and maintenance costs, while retaining the option for future reuse of both the venue and the materials that had been combined for a particular purpose at a given point in time (Brown and Cresciani, 2017).

Conclusion

Transportation asset management practices most typically embrace preserving infrastructure in its existing functionality, by assembling the resources needed for ongoing maintenance (e.g., Mohammadi et al., 2019). Although such preservation remains an important option, pursuing it as a singular goal is too limited to maintain the value of infrastructure over time, especially when the costs of a high-carbon mobility lock-in are taken into account (Givoni and Banister, 2013). Although it can be challenging to reimagine multiple alternatives at an infrastructure's half-life, planning, and designing adaptable (i.e., flexible) infrastructure will pay dividends through increasing and prolonging the value of the combined infrastructure and superstructure.

Effective and resilient infrastructure adaptation can be facilitated by relying on three key principles. First, recognizing that because an infrastructure inevitably deteriorates and loses utility over time, planners should identify a "half-life" point where options should be assessed for adapting to the future. This approach aligns with Haasnoot et al.'s (2013, p. 486) idea of an "adaptation tipping point" Second, once the half-life point has been recognized, four alternative design scenarios need to be considered. These are: *Renew*, *Redesign*, *Repurpose*, or *Remove*. These alternatives should already have been identified in the infrastructure's original design principles and even partly built into its physical form to facilitate implementing strategic transitions the future. Third, a distinction that is both organizational and technical must be made between the infrastructure (the foundation), which should be durable, and the superstructure it supports, which should be flexible.

Embracing both the temporal strategy for anticipatory renewal at or near the infrastructure's half life and the substantive analysis of four scenarios for improving utility in the future use provides the greatest opportunity to extend and enhance the value of transport infrastructure as it ages over time.

Acknowledgment

This understanding of how to adapt infrastructure to enhance its value over time originated in a collaboration with the late Dr. Moshe Givoni, Associate Professor of Geography at Tel Aviv University. Dr. Givoni was an accomplished and widely respected transport scholar who advanced understanding about mobility in those who encountered him or his work. The author would like to dedicate this encyclopedia contribution to Dr. Givoni's memory.

References

- Banister, D., 2008. The sustainable mobility paradigm. *Transp. Policy* 15 (2), 73–80.
- Banister, D., Anderton, K., Bonilla, D., Givoni, M., Schwanen, T., 2011. Transportation and the environment. *Ann. Rev. Environ. Resour.* 36, 247–270.
- Banister, D., Hall, P., 1993. The second railway age. *Built Environ.* 19 (3/4), 157–162.
- Beiler, M.O., Burkhardt, K., Nicholson, M., 2015. Evaluating the impact of Rail-Trails: A methodology for assessing travel demand and economic impacts. *Int. J. Sustain. Transp.* 9 (7), 509–519.
- Berechman, J., 2003. Transportation-economic aspects of Roman highway development: the case of Via Appia. *Transp. Res. Part A* 37, 453–478.
- Bertolini, L., 2017. *Planning the mobile Metropolis – transport for people, places and the planet*. Palgrave, London.
- Brown, L., Cresciani, M., 2017. Adaptable design in Olympic construction. *Int. J. Build. Path. Adap.* 35 (4.), 397–416.
- Buehler, R., Pucher, J., Gerike, R., Götschi, T., 2017. Reducing car dependence in the heart of Europe: lessons from Germany, Austria, and Switzerland. *Transp. Rev.* 37 (1), 4–28.
- Dixit, A.K., Pindyck, R.S., 2000. Expandability, reversibility, and optimal capacity choice. In: Brennan, M.J., Trigeorgis, L. (Eds.), *Project Flexibility, Agency and Competition: New Development in the Theory and Application of Real Option*, University Press, Oxford.
- Gifford, J.L., 1994. Adaptability and flexibility in urban transportation policy and planning. *Technol. Forecast. Soc. Change* 45, 111–117.
- Gilbert, R., Perl, A., 2010. *Transport Revolutions: Moving people and freight without oil*, Second ed. Routledge, New Society Publishers, Gabriola Island, British Columbia.
- In Givoni, M., Banister, D. (Eds.), 2013. *Moving Towards Low Carbon Mobility*. Edward-Elgar, Cheltenham, OECD location is Paris, France.
- Givoni, M., Perl, A., 2017. Rethinking transport infrastructure planning to extend its value over time. *J. Plan. Edu. Res.* 40 (1), 82–91.
- Grübler, A., 2004. *Technology and Global Change*. Cambridge University Press, Cambridge.
- Guerra, E., 2016. Planning for cars that drive themselves - metropolitan planning organizations, regional transportation plans, and autonomous vehicles. *J. Plan. Edu. Res.* 36 (2), 210–224.
- Haasnoot, M., Kwakkel, J.H., Walker, W.E., 2013. Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. *Glob. Environ. Change* 23, 485–498.
- Jones, P., 2016. The evolution of urban transport policy from car-based to people-based cities: is this development path universally applicable? In: *Proceedings of the 14th World Conference on Transport Research*. Elsevier: Shanghai, China.

- Litman, T., 2009. Transportation cost and benefit analysis: Techniques estimates and implications, Second ed. Victoria Transportation Policy Institute, Victoria.
- Ma, L., Liang, J., Gao, D., Sun, J., Li, Z., 2012. The future demand of transportation in China: 2030 Scenario based on a hybrid model. *Proc. Soc. Behav. Sci.* 54 (4), 428–437.
- Mohammadi, A., Amador-Jimenez, L., Nasiri, F., 2019. Review of asset management for metro systems: Challenges and opportunities. *Transp. Rev.* 39 (3), 309–326.
- Lyons, G., Davidson, C., 2016. Guidance for transport planning and policymaking in the face of an uncertain future. *Transp. Res. Part A* 88, 104–116.
- OECD, Organisation For Economic Co-Operation Development, 2006. Cost-Benefit Analysis and the Environment: Recent Developments. OECD Publishing.
- Parker, R.H., 1968. Discounted cash flow in historical perspective. *J. Account. Res.* 6 (1), 58–71.
- Polzin, S., Chu, X., 2014. Peak vehicle miles traveled and postpeak consequences? *Transp. Res. Rec.* 2453, 22–29.
- Rutherford, E., 1900. A radioactive substance emitted from Thorium compounds. *Philos. Mag.* 49 (5), 1–14.
- Schiller, P., Kenworthy, J., 2018. An introduction to sustainable transportation: Policy, planning and implementation. Routledge, Abingdon, Oxon.
- Stapleton, L., Sorrell, S., Schwanen, T., 2017. Peak car and increasing rebound: A closer look at car travel trends in Great Britain. *Transp. Res. Part D* 53, 217–233.
- Squires, R., 2008. The interstate sprawl system. *Society* 45 (3), 277–282.
- Wall, A.T., Warren, W.E., Marchau, V.A.J., Bertolini, L., 2015. Dynamic adaptive approach to transportation-infrastructure planning for climate change: San-Francisco-Bay-Area case study. *J. Infrastruct. Sys.* 21 (4), 05015004.
- Walton, J.R., 1998. Changing patterns of trade and interaction since 1500. In: Butlin, R.A., Dodgshon, R. (Eds.), *An Historical Geography of Europe*, Oxford University Press, Oxford, pp. 313–334.
- Zembri, P., Libourel, E., 2017. Towards oversized high-speed rail systems? Some lessons from French and Spanish cases. *Transp. Res. Proc.* 25, 368–385.

Further Reading

- Moffat, D., 2004. The art of modern transit station design. *Places* 16 (3), 75–77.

The Role of Media in Transport Planning and the Transport Policy Process

Özgül Ardiç*, J.A. Annema†, *Turkish State Railways (TCDD), Ankara, Turkey; †TU Delft, Jaffalaan 5, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	400
The Relationship Between the Media and the Public Policy Process	400
Main Concerns of the Literature on the Media and Transport Policy	401
Reciprocal Relationship between the Media and the Transport Policy Process	401
Characteristics of the Media Coverage	404
Media Influence on the Transport Policy Process	405
Conclusion	406
References	407
Further Reading	407

Introduction

The role of the media in the policy process for different transport policy measures is often questioned by scholars from various perspectives. For instance, some scholars explored the whole policy process over time and argued that the media presentation of transport policies led to some specific policy decisions and events (e.g., Norley, 2011). In fact, some scholars already assumed that the media presentation of transport policies could change the direction of a policy process and, therefore, examined how specific transport policy measures were presented by the media (e.g. Ryley and Gjersoe, 2006). But so far no comprehensive picture has been presented regarding the role of the media in transport policy planning and the transport policy process by scrutinizing the relationship between the media and the transport policy process.

This article aims to explore the role of media in transport planning and the transport policy process, based on a literature review. The Scopus database was used to select papers on the topic of this chapter. First, papers which included the word “media” in titles, abstracts or keywords, and the phrase “transport policy” in any parts, were selected. This search resulted in around 1500 papers being found. After that, the authors scanned the abstracts (and sometimes the main text) of these papers to determine the final set of papers. Papers which did not present any research questions or data analysis related to the relationship between the media and transport policy process and only mentioned this topic with reference to findings of other papers were excluded. As a result, our sample ended up having 17 papers in total.

The article is organized as follows. Section 2 presents a brief overview of the relationship between the media and the public policy process. Section 3 elaborates on the selected literature on the relationship between the media and the transport policy process, based on this overview. Finally, Section 4 presents conclusions to the chapter and ideas for future directions in this area of research.

The Relationship Between the Media and the Public Policy Process

The media is a major communication platform linking policy actors and the public in the public policy process and providing information exchanges among them. However, the relationship between the media and the public policy process is very complex, dynamic and interactive and involves a continuous interplay among numerous policy actors (e.g., politicians, interest groups), public and the media. With the development of digital technologies during the last decade, the relationship between media and other actors has become even more complex. Individuals and all policy actors can create their own news content by sharing their messages with others via various social media channels (e.g., Twitter). They do not have to rely on only traditional media and can search for numerous news sources on the Internet. But in the meantime, the traditional media outlets have been launching their new content on the Internet and incorporating social media platforms to their websites. So, in this way, the traditional media still maintains its dominant voice (Weimann and Brosius, 2015). To properly comprehend such a multifaceted relationship between the media and the public policy process, we need to probe into its two major characteristics: “reciprocity” and “conditionality” (Van Aelst, 2014; Voltmer and Koch-Baumgarten, 2010).

Reciprocity of the relationship between the media and the public policy process means that the media can influence the public policy process, while the reverse is true as well. That is to say, on the one hand, politicians may change the direction of a policy process or their policy communication according to the media coverage of their policies, as they assume that the public is highly influenced by the media. On the other hand, politicians, as well as other policy actors, all try to have their messages and policy activities in the media in order to communicate with each other and the public (Koch-Baumgarten and Voltmer, 2010; Van Aelst, 2014; Walgrave and Van Aelst, 2006). This process is “so intertwined that it is hardly possible to tell them apart” (Van Aelst and Vliegthart, 2013, p. 392). However, while some scholars observed that the influence of the media on the public policy

process is more dominant, others found that the influence of the public policy process on the media is stronger despite the fact that “the influence works both ways” (Van Aelst and Vliegenthart, 2014, p. 394). Against this background, in Section 3 we analyze the literature on the reciprocal relationship between the media and the transport policy process. We are especially interested to know which direction the influence has—is it from the transport policy process to the media or from the media to transport policy processes?

Both directions of influence—from the public policy process to the media and from the media to the public policy process—are conditional on various factors. Despite the fact that both influences do not run independently, for the ease of understanding we firstly explain how the influence of the public policy process on the media coverage is conditional, and then the other way around.

The influence of the public policy process on the media, namely the extent to which a public policy process is covered by the media, is conditional on the characteristics of the policy content (e.g., technical issues), policy events (e.g., routine politics or extraordinary policy events) and specific constellations of policy actors in the policy field concerned (e.g., a few conflicting prominent actors or numerous actors). Public policies where the characteristics of the policy content, events and actor constellations do not comply with news factors are not considered newsworthy and are ignored by the media (McQuail, 2010; Mencher, 2006). News factors are “criteria of news selection on which all journalists tend to agree” (Tresch, 2009, p. 70). For example, they include the prominence of the actors involved, the relevance of an issue, negativity, high level of conflict, geographical proximity, or the unexpectedness of an event (Mencher, 2006; Tuggle et al., 2004). The public policy process are filtered by the media according to these news factors and, thus, are not always reported, or are not fully covered. However, the media coverage of public policies may vary across media organizations as they may use slightly different news factors depending on their political slant, news tradition, editorial policies, reader profile or financial and organizational resources (Ardıç et al., 2015). For instance, the coverage may differ between popular and quality newspapers, as popular newspapers may select news based on news factors such as “negativity” or “personification” while quality ones may emphasize news factors like “controversy”, “relevance”, or “eliteness” more frequently because of differences in their commercial goals, reader profiles and editorial policies (Boukes and Vliegenthart, 2017). In the next section, based on these observations, we review the literature to explain how the transport policy process is covered in different types of media outlets.

The other direction of influence flows from the media to the public policy process. The media may have an impact on the public policy process in two ways: either rhetorically (impact on policy debate) or substantially (impact on actual policy decisions). In general, studies report strong media impact on a policy debate but rather weaker impact on actual policy measures (Van Aelst, 2014), so we discuss which types of influence are mentioned more frequently in the literature review results in the next section.

Whether or not the media has an influence on either a policy debate or actual policy decisions, the influence is contingent on the characteristics of the media input and the stages of the public policy process. Three characteristics of media input seem to determine the extent to which the media has an impact on the public policy process: the type of issue, the tone of the news and the type of media outlet (Van Aelst, 2014; Walgrave and Van Aelst, 2006). Regarding the type of the issue covered, it is argued that the media has more impact on a specific public policy when policy issues “without media would simply be not observable” (Walgrave and Van Aelst, 2006, p. 93). For example, the media coverage of foreign policy issues is expected to have a stronger impact on policy processes than that of domestic policy issues, as the media is usually the only information source of foreign policy issues for public and policy actors. Regarding the tone of the news, it is said that negative news leads to a stronger impact on public policies than positive news, as negative news urges political actors to give faster political reactions. Furthermore, not all types of media outlets have the same impact on the public policy process. It is speculated that newspapers have a stronger impact than TV news and “reliable and respected news outlets have more impact than marginal and dubious news sources” (Walgrave and Van Aelst, 2006, 93). Finally, Van Aelst and Vliegenthart (2013) show that newspapers with large audiences have stronger impacts.

Besides media input, media influence is considered to be conditional on the stages of the public policy process. The stages of the public policy process are the different phases of the process, such as “problem recognition, policy formulation, decision making, implementation, and evaluation” (Jann and Wegrich, 2007, p. 43). While some studies assume that media influence is observed during the initial phase of the public policy process when the policy problem is recognized, others show that the influence of the media is very likely to occur at other stages of the public policy process as well (Van Aelst, 2014). In the next section, we will explain in more detail the extent to which the media influence changes according to the different stages of a transport policy process and according to the characteristics of the media input: issue type, tone or type of media outlet.

Main Concerns of the Literature on the Media and Transport Policy

Table 1 lists all papers selected and summarizes their main findings. The columns of the table represent the following information about the papers, respectively: authors, transport policy issue and the country (and city) of investigation, and the main findings of the papers.

Reciprocal Relationship between the Media and the Transport Policy Process

Analyzing the road pricing policy process and its media coverage in the Netherlands, Ardıç et al. (2015) show that the relationship between the media and transport policy process is indeed reciprocal as suggested by Van Aelst (2014) and Van Aelst and Vliegenthart (2013). According to their findings, the influence from the road pricing policy process to the media seems to be more evident. That is

Table 1 Papers and main findings

<i>Papers</i>	<i>Transport policy issue</i>	<i>Country</i>	<i>Reciprocal relationship between the media and the transport policy process</i>	<i>Characteristics of the media coverage</i>	<i>Media influence on the transport policy process</i>
Connor and Wesolowski (2004)	Traffic accidents	The US	Not specified	Tone and issues: Restraint use is underreported; crashes involving drivers under the influence of alcohol and teen drivers are overrepresented; dramatic presentation with a victim/villain storyline Objectivity assessment: Not accurate (compared to official records) Media outlets: Minor differences among media outlets	Not specified
Beullens et al. (2008)	Traffic accidents	Belgium	Not specified	Tone and issues: Crashes are reported in a personalized and emotionalized manner; crashes involving young people are overrepresented; contributing factors to crashes are hardly mentioned	Not specified
Daniels et al. (2010)	Traffic accidents (motorcycle crashes)	Belgium	Not specified	Tone and issues: More serious injuries, crashes involving cars and happening on Saturdays and during the day are more frequently reported Objectivity assessment: Biased (compared to official records)	Not specified
De Ceunynck et al. (2015)	Traffic accidents	Belgium	Not specified	Tone and issues: More severe crashes, crashes involving young and female fatalities, bus or heavy goods vehicles, crashes happened on motorways, in Antwerp (compared to other provinces) and at the weekends are more frequently reported Objectivity assessment: Biased (compared to official records) Media outlets: No significant differences among media outlets	Not specified
Macmillan et al. (2016)	Traffic accidents (cycling safety and crashes)	The UK (London, Birmingham, Bristol, Cambridge)	Not specified	Tone and issues: Fatalities among females and younger individuals were more frequently reported; the media coverage of cyclist fatalities has increased tenfold, despite stability in the total number of cyclist fatalities in reality Media outlets: No significant differences among media outlets (only differences between newspapers in London and other cities)	Not specified
Vigar et al. (2011)	Transport policy package (inc. road pricing)	The UK (Manchester)	Not specified	Tone and issues: Mostly negative; charging measure is much more frequently reported than other measures Objectivity assessment: Not objective Media outlets: Considerable variation among media outlets; local media outlets are even-handed; the right wing press stressed the tax element more	Not specified

(continued)

Table 1 Papers and main findings (cont.)

<i>Papers</i>	<i>Transport policy issue</i>	<i>Country</i>	<i>Reciprocal relationship between the media and the transport policy process</i>	<i>Characteristics of the media coverage</i>	<i>Media influence on the transport policy process</i>
Ryley and Gjersoe (2006)	Road pricing policy	The UK (Edinburgh)	Not specified	Tone and issues: More negative than positive, becoming increasingly negative over time; scheme design, impact on business and congestion, fairness and equity were more frequently reported; impact on business and fairness were mostly reported negatively Objectivity assessment: Not objective Media outlets: One newspaper was more negative than the other; both newspapers changed tone over time	Not specified
Hamilton (2011)	Road pricing policy	Sweden (Stockholm)	Not specified	Tone and issues: Negative; operation cost of the system is frequently covered	Not specified
Ardıç et al. (2013)	Road pricing policy (two road pricing proposals)	The Netherlands	Not specified	Objectivity assessment: Not objective (compared to the average of media coverage) Media outlets: Two popular newspapers were negative and mixed, three quality newspapers were positive; privacy, impact on business sector and housing market were more frequently reported by popular ones, impact on the environment, road safety and car fleet, and labor market by quality ones	Not specified
Ardıç et al. (2015)	Road pricing policy (two road pricing proposals)	The Netherlands	The statements or actions of policy actors received media coverage, and the media coverage then stimulated the policy debate in each policy event	Tone and issues: Mainly positive; financial impact on households, impact on congestion, technical system and privacy issues were the most frequently reported; financial impact on households and privacy issues are the most negatively reported; the media coverage of two proposals have different characteristics.	The negative coverage by one newspaper influenced the policy debate; the negative coverage by one newspaper might have some impact on the removal of one proposal from the agenda; the influence of the media on the policy debate was rather indirect, meaning that policy actors mostly reacted to the messages from other policy actors in the media, rather than to the media coverage itself. The influence during policy stages: policy formulation and decision making
Nygrén et al. (2012)	Carbon-based car tax reform	Finland	Not specified	Tone and issues: Mainly positive; short-term impacts rather than long-term impacts are reported Objectivity assessment: Not objective	Not specified
Judge (2002)	Investment in motorway	Poland	Not specified	Tone and issues: Financial problems of the motorway program are more frequently reported than the environment impact and regional development issues	Not specified

(continued)

Table 1 Papers and main findings (*cont.*)

<i>Papers</i>	<i>Transport policy issue</i>	<i>Country</i>	<i>Reciprocal relationship between the media and the transport policy process</i>	<i>Characteristics of the media coverage</i>	<i>Media influence on the transport policy process</i>
Aldred et al. (2019)	Investments in cycling	The UK	Not specified	Tone and issues: Negative Objectivity assessment: Inaccurate (personal evaluation of interviewee)	Negative media coverage has hardly any impact on actual policy decisions (as stated by stakeholders) The influence during policy stages: problem recognition and policy formulation
Norley (2011)	Investments in public transportation	Australia (Sydney)	Not specified	Tone and issues: Positive, the media negatively covered delays and cancellations of public transport investments by the government	The positive media coverage and campaign of a newspaper have possibly some impact on actual policy decisions The influence during policy stages: problem recognition, policy formulation and decision making
Daniels et al. (2011)	Investments in public transportation	Australia (Sydney)	Not specified	Not specified	The positive campaign of a newspaper has possibly some impact on actual policy decisions The influence during policy stages: problem recognition, policy formulation and decision making
Wong and Fryxell (2004)	Fleet operations	China (Hong Kong)	Not specified	Not specified	Smaller firms are more strongly influenced by the media coverage (than bigger ones) with regard to adopting environmentally friendly fleet operations. The influence during policy stages: across all stages
Harding et al. (2016)	Auto-rickshaws	India	Not specified	Tone and issues: Largely negative, auto-rickshaws are covered as polluting and unsafe, a cause of congestion and the drivers as greedy	Negative media coverage (among other things) led to penalties and stringent regulations for auto-rickshaw drivers. The influence during policy stages: policy implementation and evaluation

to say, firstly the road pricing policy debate received some media coverage by attracting the media attention and then in reverse this media coverage influenced the policy debate. Furthermore, they analyzed that the media's impact on the policy debate was mainly indirect, meaning that policy actors reacted to statements of other actors in the media rather than to the media itself. But, this impact slightly varies according to the type of media outlet and tone of coverage.

Characteristics of the Media Coverage

Most studies in [Table 1](#) present research on characteristics of a media coverage and reflect on the extent to which the specific transport policy process has led to these characteristics in the media coverage. The studies examined four transport policy issues: traffic accidents, pricing policies (road pricing and car taxation), transport infrastructure investments (e.g., in road, public transportation, or cycling), and auto-rickshaw policy. Reviewing the findings of these studies in detail, we observe that the media coverage of these policy issues displays some similar characteristics, and also, that certain characteristics of a transport policy process seem to attract media attention more frequently.

Studies analyzing the media coverage of traffic accidents reported that traffic accidents involving young people ([Beullens et al., 2008](#); [Connor and Wesolowski, 2004](#); [De Ceunynck et al., 2015](#); [Macmillan et al., 2016](#)) or females ([De Ceunynck et al., 2015](#); [Macmillan et al., 2016](#)) and more severe crashes or serious injuries ([Daniels et al., 2010](#); [De Ceunynck et al., 2015](#)) were more frequently reported and were reported in a personalized and emotionalized manner ([Beullens et al., 2008](#); [Connor and Wesolowski, 2004](#)). [Connor and Wesolowski \(2004\)](#), [Daniels et al. \(2010\)](#) and [De Ceunynck et al. \(2015\)](#) compared these characteristics of the media coverage with those in official accident records and concluded that the media coverage was not objective,

in that some issues or events were over or underrepresented in the media. The more frequent media coverage of these characteristics seems to be related to their higher newsworthiness: scholars stated that crashes involving women fatalities were more frequently covered in the media because the low frequency of such crashes in reality made them more unexpected, and thus more newsworthy (De Ceunynck et al., 2015). Crashes involving teenagers were more newsworthy because they were more dramatic, compared to those involving adults (Connor and Wesolowski, 2004). An increase in the popularity of cycling in London made cycling accidents a more relevant issue and thus more newsworthy and raised the media's interest to cover cyclist fatalities, despite the fact that, in reality, no increase in cyclist fatalities could be observed. This caused a divergence in the media coverage of newspapers in London and other cities (Macmillan et al., 2016). But other than this, no difference in traffic accident coverage among media outlets was reported by other studies (see Table 1).

From the studies which examined the media coverage of road pricing policies and car taxation, two were conducted for British road pricing proposals, by Ryley and Gjersoe (2006) and Vigar et al. (2011), and one was for a Swedish proposal, by Hamilton (2011). They found that the road pricing policies were reported negatively by the media. On the other hand, the other two road pricing studies, one conducted for Dutch road pricing proposals by Ardiç et al. (2015) and one for a Finnish carbon-based car tax proposal by Nygrén et al. (2012), found that the media coverage for the proposals was positive. But all studies concluded that the media was not objective in its reporting of these proposals. As to the issues covered by the media, scheme design, the impact on business and congestion, and the financial impact on various households groups (equity/fairness) were more frequently reported during the Edinburgh road pricing policy process in the UK. Of these, the impact on business and various households groups had mostly negative coverage, as stated by Ryley and Gjersoe (2006). Likewise, in the Netherlands, Ardiç et al. (2015) reported that the financial impacts on various households groups, the impact on congestion, the technical system and privacy were the most frequently reported issues, and of these, the impact on various households groups and privacy were mostly negative.

Ardiç et al. (2015) explained that the characteristics of media coverage were the reflection of the Dutch road pricing policy process in as far as the specific policy debate and events and policy content complied with news factors. For instance, one pricing proposal during the 12 years of analysis of the Dutch policy debate on road pricing had more coverage than the other one because this one was more relevant, thus more newsworthy than the other one because of its comprehensive system design which was expected to influence the financial situation and mobility behavior of the whole population. Or, the privacy issue was almost always negatively covered by the media during the 12 years of the analysis period, except during one policy event in which a technical report was published. The findings of the report were positive about the privacy aspects of the proposal and most of the prominent actors reacted positively to these findings. An extraordinary event (the publication of this report) and the reactions of prominent individuals (political leaders) made these positive events and reactions newsworthy for the media and thus led to the positive media coverage of this issue.

Furthermore, Ardiç et al. (2013), Ryley and Gjersoe (2006), and Vigar et al. (2011) indicated that the characteristics of road pricing media coverage (the tone and issues reported) varied across different media outlets. Ardiç et al. (2013) observed the differentiation between popular and quality newspapers: two popular newspapers were usually negative and mixed, three quality newspapers were usually positive over the 12 years when the policy process was analyzed. Yet, why such differences occur between quality and popular newspapers (e.g., editorial policies, reader profile or commercial concerns), or what new factors are relatively more important to each newspaper, is not known. Ardiç et al. (2013), however, mentioned a possible editorial influence on the tone of one newspaper's coverage.

The remaining papers, which address the extent to which the transport policy process is covered by the media, are about four different policy issues from four different countries. Aldred et al. (2019) reported that investments in cycling in the UK had negative and inaccurate coverage, according to the opinions of stakeholders. Examining media coverage of a Polish motorway investment program, Judge (2002) found that the media reported the financial problems of motorway programs more frequently than the environmental impact and regional development issues. It seems that "the urgency of domestic financing issues, ... the dominance of external funders" on the Polish public agenda made the financing aspect of the program more relevant, and therefore more newsworthy for the media. Norley (2011), on the other hand, indicated that the media coverage of investment in public transport was positive in Australia. In addition to that, a newspaper even started a campaign to promote public transport on the public agenda (Norley, 2011). Therefore, editorial policies seemed to play a role in the positive coverage of public transport in Australia. Finally, Harding et al. (2016) analyzed the media coverage of auto-rickshaws in India and found that media coverage was largely negative and auto-rickshaws were covered as polluting and unsafe, a cause of congestion and the drivers as greedy. Such media coverage was very likely to be a reflection of "considerable criticism from public, ... and policy makers" about auto-rickshaws, as stated by Harding et al. (2016, p.143). Such opposition from the public and policy makers should have increased the newsworthiness of these negative aspects of auto-rickshaws and, thus, amounted to media coverage of these aspects.

Media Influence on the Transport Policy Process

Not many of the papers in Table 1 investigated the influence of the media on the transport policy process. Only the study by Ardiç et al. (2015) analyzed the influence of media coverage on the transport policy debate, four other studies considered only the media's influence on actual policy decisions. Comparing media coverage of road pricing policies and parliamentary debates over time, Ardiç et al. (2015) found that the media coverage indeed influenced the parliamentary debate, but rather indirectly. That is to say, parliamentarians reacted, not to newspapers, but to the statements made by other parliamentarians and policy actors covered by the news. This indicates that, in this case, the media mainly acted as a communication platform. The findings of other studies were

related to the influence of media on actual policy decisions. Aldred et al. (2019) argued that negative media coverage had hardly any impact on actual policy decisions regarding investments in cycling in the UK. Daniels et al. (2011), Harding et al. (2016), Norley (2011), and Wong and Fryxell (2004), on the other hand, claimed that the media possibly had some impact on actual policy decisions. The analyses of Wong and Fryxell (2004) and Aldred et al. (2019) are based on the views of stakeholders, but the empirical evidence in the studies of Daniels et al. (2011), Harding et al. (2016), and Norley (2011) is not very clear.

According to Ardiç et al. (2015) and Norley (2011), some newspapers are relatively more influential or play a more active role (like other policy actors) in the policy process. When analyzing references made to one popular newspaper during parliamentary meetings, Ardiç et al. (2015) pointed out that the negative coverage of this newspaper was likely to be one of the factors which played a role in the policy decision to abandon one of the road pricing proposals. In addition, the analysis of Ardiç et al. (2015) proved that this newspaper had a stronger impact on the policy debate than other newspapers. This stronger impact could be because of its wider readership or its negative coverage about the road pricing proposal, as Van Aelst and Vliegenthart (2013), Walgrave and Van Aelst (2006) stated. Similarly, Norley (2011) stated that the editorial office of one newspaper possibly played a role in transport policy planning by organizing a campaign to promote public transportation in Australia.

Finally, the studies mentioned in the previous paragraphs reported that media influence was observed across almost all stages of the transport policy process, with the policy recognition and policy formulation stages being more profound. They do not support the findings of studies which argue that media influence is limited to the policy recognition stages of a policy process (see Van Aelst, 2014).

Conclusion

This article aimed to explore the role of the media in transport planning and the transport policy process. To achieve this aim, we scrutinized the relationship between the media and the transport policy process, based on findings in existing literature and we used theoretical perspectives developed in the field of public policy and the media. Our analysis indicated that this relationship is reciprocal. That is to say that the influence of the media on the transport policy process, and its reverse, the influence of the transport policy process on the media, both exist. The current state of research in the field of transport policy presents ample evidence that the characteristics of the specific transport policy process (related to policy content, events, and actors) are reflected on the media coverage to some extent, whereas limited evidence is found regarding the influence of the media on the transport policy process, and the circumstances under which such influence is observed.

The results of our review support the research findings of public policy presented in Section 2 in a way that certain characteristics of a transport policy process attract media attention more frequently. Unexpected, extraordinary or dramatic events, speeches by prominent actors (both negative and positive), and relevant issues concerning or influencing many people's lives during the transport policy process are likely to be considered newsworthy by the media and receive media coverage. Most studies reviewed made an inference to the extent to which the media was objective in its reporting of the transport policy process and interestingly, they all concluded that the media did not report the policy process objectively. But, their findings in this regard (except one about traffic accidents) do not have a strong empirical basis as they did not use an independent measure of reality (media input, namely transport policy process) to compare to the media coverage in their measurement of media objectivity. "Reality" is not clearly defined in the studies of Nygrén et al. (2012), Ryley and Gjersoe (2006), and Vigar et al. (2011). Ardiç et al. (2013) consider the average of five newspapers as "reality" and compare each newspaper's coverage to this so-called reality to measure their objectivity. However, we note that it is not easy to find any decent measurement for "reality" for public policy debate and process (as it is for traffic accidents) to compare this to media coverage in order to evaluate the objectivity of media reporting (see detailed discussion about measurement of objectivity (McQuail, 1999)). No difference is found among newspapers in terms of the media coverage of traffic accidents, but with respect to road pricing policies, studies reported considerable differences among newspapers. It is likely that, because of the politicized nature of road pricing policy processes during the policy formulation period, as stated by Ardiç et al. (2013), Ryley and Gjersoe (2006), and Vigar et al. (2011), political slants, reader profile and the editorial policies of newspapers become more evident in the way of reporting and so they diverged their coverage accordingly.

With respect to the influence of the media on the transport policy process, Ardiç et al. (2015) provided empirical evidence regarding the influence of the media on the road pricing policy debate. Their findings imply that the media mainly acted as a kind of communication platform during the policy process. Namely, parliamentarians reacted, not to newspapers, but to the statements of other parliamentarians and policy actors covered by the news. The rest of the studies discuss whether the media influenced actual policy decisions. But, they do not seem to present very strong empirical evidence regarding the influence of media on actual policy decisions. Nevertheless, it should be admitted that, methodologically, it is very difficult to distil the influence of the media on actual policy decisions because public policy making is very complex and numerous factors play a role in actual policy decisions. Regarding the contingency of media influence on the type of media outlet, it is interesting that some newspapers seem to be relatively more influential, or play a more active role (like other policy actors) in the policy process. One newspaper in the Dutch road pricing policy process and the one in policy process regarding public transport in Australia influenced policy debate (possibly actual policy decisions) played an active role. Finally, media influence was observed across almost all stages of the transport policy process, but, in the studies that were reviewed, it is more often observed during the policy recognition and policy formulation stages.

To sum up, this article has shown that the media and the transport policy process influence each other reciprocally, but this influence is conditional on various factors. Most strikingly, research on this topic seems to be in its infancy. More research and more

specific case studies are needed to better understand the influence of the media on the transport policy process. Furthermore, all studies in this review involve only one country. Crosscountry differences in terms of political and media systems might give more insight into the relationship between the media and the public policy process. Thus, it might be useful to examine the role of media and the transport policy process by comparing cases from different countries in one study.

References

- Aldred, R., Watson, T., Lovelace, R., et al., 2019. Barriers to investing in cycling: stakeholder views from England. *Transport. Res. Part A: Policy Practice* 128, 149–159.
- Ardıç, Ö., Annema, J.A., van Wee, B., 2013. Has the Dutch news media acted as a policy actor in the road pricing policy debate? *Transport. Res. Part A: Policy Practice* 57, 47–63.
- Ardıç, Ö., Annema, J.A., van Wee, B., 2015. The reciprocal relationship between policy debate and media coverage: the case of road pricing policy in the Netherlands. *Transport. Res. Part A: Policy Practice* 78, 384–399.
- Beullens, K., Roe, K., Van den Bulck, J., 2008. Television news' coverage of motor-vehicle crashes. *J. Safety Res.* 39 (5), 547–553.
- Boukes, M., Vliegthart, R., 2017. A general pattern in the construction of economic newsworthiness? Analyzing news factors in popular, quality, regional, and financial newspapers. *Journalism*, 1464884917725989.
- Connor, S.M., Wesolowski, K., 2004. Newspaper framing of fatal motor vehicle crashes in four Midwestern cities in the United States 1999–2000. *Injury Prevent.* 10 (3), 149–153.
- Daniels, R., Gordon, C., Mulley, C., 2011. Mass media and mass transit: a newspaper's campaign on public transport planning in Sydney.
- Daniels, S., Brijis, T., Keunen, D., 2010. Official reporting and newspaper coverage of road crashes: a case study. *Safety Sci.* 48 (10), 1469–1476.
- De Ceunynck, T., De Smedt, J., Daniels, S., et al., 2015. Crashing the gates—selection criteria for television news reporting of traffic crashes. *Accid. Anal. Prevent.* 80, 142–152.
- Hamilton, C.J., 2011. Revisiting the cost of the Stockholm congestion charging system. *Transport Policy* 18 (6), 836–847.
- Harding, S.E., Badami, M.G., Reynolds, C.C., et al., 2016. Auto-rickshaws in Indian cities: public perceptions and operational realities. *Transport Policy* 52, 143–152.
- Jann, W., Wegrich, K., 2007. Theories of the policy cycle. In: Fischer, F., Gerald, M., Mara, S. (Eds.), *Handbook of Public Policy Analysis: Theory, Politics, and Methods*. CRC/Taylor & Francis, Boca Raton.
- Judge, E.J., 2002. Environmental and economic development issues in the Polish motorway programme: a review and an analysis of the public debate. *Eur. Environ.* 12 (2), 77–89.
- Koch-Baumgarten, S., Voltmer, K., 2010. Conclusion: the interplay of mass communication and political decision making – policy matters! In: Koch-Baumgarten, S., Voltmer, K. (Eds.), *Public Policy and Mass Media: The Interplay of Mass Communication and Political Decision Making*. Taylor & Francis, Hoboken.
- Macmillan, A., Roberts, A., Woodcock, J., et al., 2016. Trends in local newspaper reporting of London cyclist fatalities 1992–2012: the role of the media in shaping the systems dynamics of cycling. *Accid. Anal. Prevent.* 86, 137–145.
- McQuail, D., 1999. *Media performance: Mass communication and the public interest*. SAGE, London.
- McQuail, D., 2010. *McQuail's Mass Communication Theory*. SAGE, Los Angeles.
- Mencher, M., 2006. *Melvin Mencher's News Reporting and Writing*. McGraw-Hill, Boston.
- Norley, K., 2011. Urban rail infrastructure—the path from comprehensive transport plans to the recent experience.
- Nygrén, N.A., Lyytimäki, J., Tapio, P., 2012. A small step toward environmentally sustainable transport? The media debate over the Finnish carbon dioxide-based car tax reform. *Transport Policy* 24, 159–167.
- Ryley, T., Gjersoe, N., 2006. Newspaper response to the Edinburgh congestion charging proposals. *Transport Policy* 13 (1), 66–73.
- Tresch, A., 2009. Politicians in the Media: Determinants of Legislators' Presence and Prominence in Swiss Newspapers. *Int. J. Press/Politics* 14 (1), 67–90.
- Tuggle, C.A., Carr, F., Huffman, S., 2004. *Broadcast News Handbook: Writing, Reporting and Producing in a Converging Media World*. McGraw-Hill, Boston.
- Van Aelst, P., 2014. 12 Media, political agendas and public policy. In: Reinemann, C. (Ed.), *Political Communication*. p. 231.
- Van Aelst, P., Vliegthart, R., 2013. Studying the tango. *Journal. Stud.* 15 (4), 392–410.
- Vigar, G., Shaw, A., Swann, R., 2011. Selling sustainable mobility: the reporting of the Manchester Transport Innovation Fund bid in UK media. *Transport Policy* 18 (2), 468–479.
- Voltmer, K., Koch-Baumgarten, S., 2010. Introduction: mass media and public policy—is there a link? *Public Policy and the Mass Media* Routledge 19–32.
- Walgrave, S., Van Aelst, P., 2006. The contingency of the mass media's political agenda setting power: toward a preliminary theory. *J. Commun.* 56 (1), 88–109.
- Weimann, G., Brosius, H.-B., 2015. A new agenda for agenda-setting research in the digital era. *Political Commun. Online World*. Routledge 26–44.
- Wong, L.T., Fryxell, G.E., 2004. Stakeholder influences on environmental management practices: a study of fleet operations in Hong Kong (SAR) China. *Transport. J.* 22–35.

Further Reading

- Voltmer, K., Koch-Baumgarten, S., 2010. Introduction: mass media and public policy—is there a link? *Public Policy and the Mass Media*. Routledge 19–32.
- Walgrave, S., Van Aelst, P., 2006. The contingency of the mass media's political agenda setting power: toward a preliminary theory. *J. Commun.* 56 (1), 88–109.
- Van Aelst, P., 2014. 12 Media, political agendas and public policy. In: Reinemann, C. (Ed.), *Political Communication*. p. 231.

Travel Plans

Stephen Potter*, **Marcus Enoch†**, *School of Engineering and Innovation, Open University, Milton Keynes, Buckinghamshire, England; †School of Architecture, Building and Civil Engineering, Loughborough University, Leicestershire, England

© 2021 Elsevier Ltd. All rights reserved.

Defining and Characterizing the Travel Plan	408
Emergence	409
Motivations for Promoting or Adopting Travel Plans	409
Stagnation	410
Barriers to Travel Plan Take Up	410
Transformation	411
References	412

Defining and Characterizing the Travel Plan

Travel Plans have had a variety of names (Green Commuter Plan, Green Transport Plan, Sustainable Travel Plan, Mobility Management among others) and have been defined in a number of ways. These include:

- “A travel plan sets out to combat over-dependency on cars by boosting all the possible alternatives to single occupancy car use It cannot only benefit the environment but can produce financial benefits and productivity improvements.” (NBTN/ Department for Transport – *The Essential Guide to Travel Planning* 2008).
- “Travel planning is an effective business management tool which can be used to generate cost savings, lending companies a competitive advantage, and which has additional benefits for the environment and the health of employees.” (Transport for London, - *Business on the move*, 2005).
- “A package of measures to meet the needs of individual site(s) aimed at promoting greener, cleaner travel choices and reducing reliance on the car. It involves the development of a set of mechanisms, initiatives and targets that together can enable an organisation to reduce the impact of travel and transport on the environment whilst also bringing a number of benefits to the organisation and to staff.” (Energy Efficiency Best Practice Programme, 2001)

The differences in definition often reflect the perspective from which Travel Plans are considered. A city or transport planning authority will emphasize the transport system and environmental benefits that a Travel Plan might deliver; a business organization would take a business efficiency perspective and a funder of an organization (e.g., an education authority funding schools) might take a more regulatory/compliance approach.

In synthesizing meaning from a range of definitions, [Enoch \(2012\)](#) noted that the common themes emerging from these definitions are:

- Travel Plans are not an instrument themselves, but a delivery mechanism or strategy for other mostly transport-focused measures.
- Travel Plans are delivered by an additional “agent” that is not a part of the “traditional” transport policy institutional structure. This agent may be an employer, a school, a shopping center or any organization with responsibility for a site that generates traffic impacts.
- Travel Plans are initiated in two ways: either by the organization concerned or by a government body.
- Travel Plans seek to deliver transport and related benefits to the community as well as some direct organizational benefits to the participating “delivery agents.”
- Travel Plans are to some extent “site-specific,” that is tailored to the specific contextual circumstances.
- Travel Plans to some extent deliver a package or a strategy of a variety of transport instruments.

[Enoch \(2012\)](#) uses a definition that stems from these points, in that Travel Plans are:

“a mechanism for delivering a package of transport measures targeted at a specific site by an agent with a strong relationship with the local transport users to deliver transport and wider goals to the organisation and society as a whole”

(Enoch 2012, p.58)

In practice Travel Plans have been applied to a number of sectors, including schools ([Baslington, 2008](#)), residences ([De Gruyter et al., 2015](#)), visitor attractions ([Guiver and Stanford, 2014](#)) and hospitals ([TfL, 2006](#); [NHS 2006](#)), as well as retail, events and transport interchanges, though they are mostly associated with employers ([Roby, 2010](#)), and with the commute trip in particular. This chapter examines the evolution of the Travel Plan in its three stages so far, namely: emergence, stagnation, and transformation.

Emergence

For many years some employers have helped staff travel to work. Factories in remote locations might provide works buses for staff and often special transport arrangements are made for late-night shift workers. The Travel Plan concept—that employers and organizations should take responsibility to manage how staff, customers, and visitors travel to their site in response to wider public policy reasons—is a more recent development. The first initiative of this sort can be traced back to the 1940s, when as part of the war effort to conserve fuel, Boeing in Seattle introduced a car lift sharing scheme and a promotional poster of the time declared “When you ride ALONE, you ride with Hitler!”. After the war ended, the need to conserve fuel ceased and so too did car share schemes. Another external jolt led to their return, when the 1973 Yom Kippur War and the overthrow of the Shah of Iran in 1979 led large companies such as Amoco in Houston, 3M in Minneapolis-St Paul and Aerospace Corporation in Los Angeles to introduce measures to help their staff get to work (Shreffler, 1986).

The first environmental requirements for employers to manage the travel of their staff came as part of Californian air-quality legislation in the 1970s. The Dutch adopted Travel Plan type measures in the 1980s and local authorities in the United Kingdom (notably Nottingham City Council) began promoting “green commuter plans” in the early 1990s. The promotion of Travel Plans became UK Government policy from 1997, with workplace and school Travel Plans featuring in the 1998 and 2004 transport policy White Papers (DETR, 1998 and DfT, 2004). Travel Plans have also been adopted in other countries throughout Europe (EPOMM, 2013), as well as in Australia, New Zealand, Canada, and Japan (Enoch, 2012).

Motivations for Promoting or Adopting Travel Plans

A distinction from conventional transport policy is that Travel Plans are not primarily implemented by public bodies through planning, regulations, or infrastructure investments (e.g., building roads, cycleways, or metro systems). Instead, the implementation and management of a Travel Plan is by the organization whose site or operations generates traffic impacts. The varying emphases in the Travel Plan definitions above represent this shift in how public policy engages with such organizations. A crucial aspect for Travel Plan success is how such engagement is achieved and what motivates organizations to adopt and maintain a Travel Plan.

The involvement of organizations whose operations generate traffic impacts is both a strength and a weakness of the Travel Plan approach. The main weakness is that the vast majority of employers and other institutions do not see solving transport problems as their responsibility. To date, rather than adopting an integrated management approach, employers have tended to treat transport matters as either separate self-contained issues or a matter that is largely outside their control and not their responsibility. These issues have included:

- Road congestion affecting delivery reliability and costs (as well as staff punctuality)
- Congestion of on-site parking
- Transport-related planning conditions required for site development
- Changes to the tax treatment of transport benefits in the remuneration package (company cars, mileage allowances, etc.) and other transport-related human resource issues
- Transport and company environmental policies (including environmental requirements of export markets and other supply chain pressures)
- Transport effects upon brand image and public relations
- Transport impacts upon company quality initiatives.
- Meeting Corporate Social Responsibility goals, climbing “green league tables,” corporate image
- Recruitment and retention

As in all good management practice, it is crucial not to treat these seemingly disparate issues in isolation: this leads to ad hoc “fire fighting” that is costly and ineffective. The key to a cost-effective and successful approach is to recognize that organizations are dealing with a series of challenges and opportunities stemming from changes in the transport environment. A strategic, integrated business approach is needed, and this is where Travel Plans come into their own.

Nevertheless, companies generally consider Travel Plans only when some other pressing reason forces them to do so (Rye, 2002). In some cases, a Travel Plan can arise from a crisis of on-site parking congestion, or a perception that a company’s transport problems are harming its business image. For example, it could be embarrassing to a university that teaches environmental management if it could not show that it practices what it preaches by having an effective Travel Plan. However, in the majority of cases the most pressing reason arises when a company wishes to move to a new site or expand its existing site, and the local authority forces it, through a condition of planning consent, to manage the ways in which employees or customers travel. This approach is commonly used in the United Kingdom, Australia, and the United States, and is often tied to the principle that new developments generate an additional burden on surrounding infrastructure and services that needs to be mitigated (De Gruyter et al., 2018).

Outside of Travel Plans being required for new developments, in some places they were introduced as a measure for all employment sites in a city. This approach was first trialed in 1985 in several districts of California to combat air pollution, with the concept transferred to the Netherlands in 1989 following a study tour of Dutch policy makers looking to solutions for worsening traffic congestion (Hoogma, 2006). For some public sector organizations, a Travel Plan might be required directly by Government or as a condition of finance.

When consistently applied, Travel Plans can cut car use by worthwhile amounts. The best employer travel plans in the United Kingdom secured a reduction in car use of between 10% and 20%, while some feel that up to 30% is a possibility (Cairns et al., 2004). In the United States, where mandatory Travel Plans have been in use, in several cases a 30% cut in car use has been achieved.

Stagnation

Despite the potential promised by the concept, Travel Plans have never really established themselves as a mainstream part of transport policy. Despite the many apparent benefits afforded, in most locations Travel Plans have remained a relatively niche concern, with the overall impact on the transport network being correspondingly marginal. Thus, Rye (2002) concludes that, although there is clear evidence that Travel Plans have an effect at the site level, their potential for a system-wide effect was unrealized. He estimated that the contribution of Travel Plans to congestion reduction at the UK level to be less than 1% at the time of publication (2002), and anecdotal evidence suggests that nothing significant has occurred since that might have changed this to any degree.

Barriers to Travel Plan Take Up

Within organizations, Travel Plans were easily marginalized and discarded if they conflicted with other priorities. This particularly applied to where Travel Plans had been a regulatory imposition (planning or sector requirement), which tended to lead to a “minimum cost for compliance” approach. If the specification was simply to produce a Travel Plan, then the document was logged and no further actions taken to implement or maintain it. Even if the specification required the Travel Plan to be implemented with certain measures or to achieve a car use reduction target, it might fall into disuse when the regulatory pressure eased. Around the 2008 financial crisis and subsequent recession, many Travel Plans were simply discarded.

An in-depth study of 25 Travel Plans (Roby 2008), explored this issue, noting that all were originally driven by external factors, such as a planning condition or government directive. Eventually, about a third started to identify internal benefits of the Travel Plan, particularly reducing car parking costs, but also a positive impact on recruitment and retention. The majority, however, remained motivated by external requirements and had not developed an internal rationale for their existence.

Even if internal benefits could be identified, these were often too small to justify much resource being devoted to a Travel Plan or the benefits could not be financially captured. A key structural point is that the main benefits often fall outside the department that funds the Travel Plan. It is quite usual for a Travel Plan to be the responsibility of a Facilities or Estates department. This is because of their institutional responsibilities and they link to a local authority or the site management side of their funding body – the organizations that impose a Travel Plan requirement. The cost of running a Travel Plan can be offset by reductions in other expenditure within a Facilities department’s budget; hence it is not surprising that reducing car parking costs was the main internal benefit identified in Roby (2008). However, cross-departmental costs and benefits are more difficult to manage. Staff recruitment and retention is the responsibility of personnel, and they might be reluctant to transfer monies to facilities even if the Travel Plan does provide them with real benefits.

Issues that produce impacts across a business need to move up the management hierarchy, but, as noted by Enoch (2012) Travel Planning largely became “ghettoised” - a function undertaken by junior staff in the lower echelons of a Facilities department. In a very few cases, Travel Plans (or components of Travel Plans) came to be recognized of sufficient importance that they were supported by the managerial board who allocated them strategic resources. This was so for a number of university Travel Plans. One example is the University of Hertfordshire, which has developed a bus network serving its campuses and the surrounding areas. Cranfield University and the University of Falmouth are among others that also run their own bus services. The reason these are important for these universities is that they link into an issue of strategic importance—student recruitment and retention. The bus services provide a much-valued benefit to fee-paying students who increasingly expect good accommodation, social facilities and access. This Travel Plan measure has become of importance to student recruitment. The bus routes are public services and have developed to become profitable, so are doubly valued as now being a profit center as well as contributing to student recruitment and retention.

A second key factor in the stagnation and decline of Travel Plans has been that they were part of a demand management strategy for transport policy that came to be abandoned. Furthermore, Travel Plans were not just a demand management initiative, but they were something that involved a different means of implementation. Traditional transport policy is a civil-engineering focused area, built around state infrastructure projects funded by general taxation. A shift to demand management transport measures has largely retained this regime approach; instead of infrastructure projects for roads, the same approach and skills are used for infrastructure projects for public transport and cycle ways etc. But Travel Planning involves a different approach and skill set. It is about sustaining initiatives, not one-off civil engineering project actions; it is about building partnerships, consulting, educating and empowering users, implemented by a dispersed range of stakeholders—the role of the state (particularly local authorities) is significantly different. It is more like a management process than “traditional” transport engineering.

Enoch (2012) detailed the institutional barriers for Travel Planning in more detail, noting that this existed within Government and within organizations and was a product of how Travel Planning itself had emerged. Broadly, he concluded that Travel Plans had become vulnerable because they do not map onto the structures of the organizations in which they take place, nor the organizational structures and operating regimes of Government. The skills and culture clash between travel planners and the contexts in which they operate resulted in their stagnation and decline.

Furthermore, following the 2008 recession the ethos of transport policy shifted away from the demand management approach in which Travel Plans were rooted. In the UK context, after an initial flirtation in 2010 with austerity across the board, in 2016 the focus of the Conservative Government for transport shifted toward investing in major infrastructure projects to generate economic growth, albeit with a significant emphasis on public transport projects, such as London Crossrail and the High Speed 2 rail project. Major road expansions also took place including the use of “Smart” motorways to enlarge existing motorway capacity, an approach that did not require any engagement at a detailed level with employers or other organizations. To date, this emphasis on infrastructure expansion has not been viewed as incompatible with the growing climate change agenda as that was being addressed by measures that were compatible with the existing policy implementation approach. This was through traditional fiscal and regulatory measures to promote a fuel shift to electric vehicles combined with the decarbonisation of electricity generation. Transport policy reverted to what it was able to do—civil engineering projects and a policy regression to an emphasis on infrastructure expansion. In such a policy climate, Travel Plans became an irrelevance.

Transformation

Yet policy contexts rarely stand still for long, and circumstances have continued to evolve regarding Travel Plan services, although often not labeled or presented under the term “Travel Plan.” There has been a dramatic change in the ecosystem of organizations providing “plug and play” Travel Plan services for employers and other organizations, many of which have come from the digital economy. These range from global players such as Uber to much smaller enterprises like Vectare, which provides niche consultancy software and managed transport solutions to schools seeking to deliver safer and more efficient transport options to pupils. Another example is the sustainable transport consultancy Go Travel Solutions.

In addition, a whole suite of data generating and analytics companies such as BlockDox, Vivacity and Elephant Wi-Fi have moved to operate in the transport space, making it much more straightforward for organizations to fully understand and proactively influence the travel patterns of their employees, students, patients, and visitors at an ever diminishing cost and effort. Ten years ago, such data gathering and analysis was ad hoc, expensive, and tied up staff resources. Now travel information is found in the standard portfolio of data gathering and analytics service companies.

Another (related) change has been the rise of corporate “wellness” programs, as supported by Government policy announcements such as the work, health, and disability green paper: *Improving Lives* (DWP and DHSC, 2017), whereby companies support and encourage employees to look after their health by, for example, stopping smoking.

A further strand stimulating the return of Travel Plan type actions are two attitudinal shifts. First has been a steady increase in the “sharing economy.” This is typified by the highly successful Airbnb accommodation service, but there are transport examples where people share their car journeys with others in return for a charge and sharing material possessions such as cars or bikes through a Digital Service Platform (see www.earthshare.org). Second, there have been signs that the public is becoming much more supportive of policies aimed at addressing air quality, congestion, and climate change that require actions traditional transport policies cannot deliver. Hence, many cities around the world have recently introduced regulations to limit access to polluting vehicles into urban centers such as the Crit’Air Low Emission Zone scheme in France, and License Plate Restriction policies throughout Latin America and China (Ramos et al., 2017), which suggests that there may be also scope for Travel Plans focused on addressing these agendas.

Subsequent to these however, three key events during the first few months of 2020 have once again fundamentally changed many of the core assumptions underpinning transport policy. First, there was the ruling, on 27th February in the UK Court of Appeal in London that the Heathrow third runway could not go ahead because it is not consistent with the Paris Agreement on Climate Change (Carrington, 2020). Second, and related to this was the release of the Decarbonising transport: Setting the Agenda White Paper by the Department for Transport on 26th March (DfT, 2020), which committed the UK Government to prioritizing active travel and public transport. Interestingly, while it does not refer to Travel Plans explicitly, it does acknowledge the importance of “place-based solutions” as one of six strategic priorities which could potentially be viewed at a more neighborhood scale than the local authority level envisaged currently (pp. 7).

Third, and perhaps most significant are the ramifications of the global Corona virus pandemic set to significantly impact on how people travel in future (Enoch and Warren, 2020), and also on the financial ability of the Government to continue to spend on infrastructure improvements once the crisis is over. In addition, it is clear that one key element in ensuring the effectiveness of Government policy actions, around social distancing measures in particular, is to have employers “on side.” Indeed, one oft-cited reason for people traveling in spite of Government instructions not to during late March 2020 was that they were required to do so by their employers to keep their job (McCrum et al., 2020). This realization, that engagement with the factors behind the generator of traffic is needed and not a response to them, may well have been one reason for a guidance for employers and businesses on the corona virus note being issued by Government in early April 2020, which pointed out the need to encourage working from home policies as its first piece of advice—essentially a Travel Plan measure (DEBIS and PHE, 2020).

Taken together, these technological, attitudinal, and policy drivers present a situation in which the concept of the Travel Plan could well re-emerge, but likely in a much less structured and visible incarnation than when they were first adopted. Hence, many (perhaps most) organizations may well find themselves implementing one or more of the “plug-and-play” tools provided by a growing emergent digital cottage industry comprising suppliers of business-facing (travel-impacting) solutions without ever realizing that they are engaged in putting a Travel Plan in place. Related to this shift toward “shadow” Travel Plans among some organizations has been a parallel downplaying of the traditional Travel Plan concept within Government, as highlighted in the Decarbonising Transport report mentioned earlier. Yet in the final analysis, this “invisibility” may actually prove to be a blessing, as

long as the whole suite of supporting information, fiscal, and regulatory policies across Government around managing car use and promoting active travel and public transport use are put in place to encourage organizations to make the “right choices” during the course of their day-to-day operations.

References

- Baslington, H., 2008. School travel plans: overcoming barriers to implementation. *Transp. Rev.* 28 (2), 239–258.
- Cairns, S., Sloman, L., Newson, C., Anable, J., Kirkbride, A. and Goodwin, P., 2004. *Smarter Choices – Changing the Way We Travel*, UCL, Robert Gordon University and Eco-Logica. Final report to Department for Transport, London.
- Carrington, D., 2020. Heathrow third runway ruled illegal over climate change, *The Guardian*, 27 February. Available from: <https://www.theguardian.com/environment/2020/feb/27/heathrow-third-runway-ruled-illegal-over-climate-change>. Last accessed 6 April 2020.
- De Gruyter, C., Rose, G., Currie G., 2015. Enhancing the impact of travel plans for new residential developments: Insights from implementation theory. *Transp. Policy* 40, 24–35.
- De Gruyter, C., Rose, G., Currie, G., Rye, T., van de Graaff, E., 2018. Travel plans for new developments: A global review. *Transp. Rev.* 38 (2), 142–161.
- Department of the Environment, Transport and the Regions (DETR), 1998. *New Deal for Transport: Better for Everyone*, (White Paper), London, The Stationery Office.
- Department for Business, Energy and Industrial Strategy (DEBIS) and Public Health England (2020) Guidance for employers and businesses on coronavirus (COVID-19), UK Government, 6 April. Available from: <https://www.gov.uk/government/publications/guidance-to-employers-and-businesses-about-covid-19/guidance-for-employers-and-businesses-on-coronavirus-covid-19>. Last accessed 6 April 2020.
- Department for Transport, 2020. Decarbonising Transport: Setting the Challenge, Crown Copyright, London, 26 March. Available from: <https://www.gov.uk/government/publications/creating-the-transport-decarbonisation-plan>. Last accessed 6 April 2020.
- Department of Work and Pensions and Department of Health and Social Care, 2017. Work, health and disability green paper: Improving Lives, DWP and DHSC, 30 November 2017. Available from: <https://www.gov.uk/government/publications/guidance-to-employers-and-businesses-about-covid-19/guidance-for-employers-and-businesses-on-coronavirus-covid-19>. Last accessed 6 April 2020.
- Department for Transport (DfT), 2004. White Paper: The Future of Transport: a network for 2030, London, Department for Transport, Cmd 6234.
- Energy Efficiency Best Practice Programme, 2001. *A Travel Plan Resource Pack for Employers*, Energy Efficiency Best Practice Programme, London, The Stationery Office.
- Enoch, M.P., 2012. Sustainable transport mobility management and travel plans. Ashgate, Farnham.
- Enoch, M., Warren, J., 2020. Coronavirus is a once in a lifetime chance to reshape how we travel. *The Conversation*, 2nd April. <https://theconversation.com/coronavirus-is-a-once-in-a-lifetime-chance-to-reshape-how-we-travel-134764>.
- European Platform on Mobility Management, 2013. Mobility management: The smart way to sustainable mobility in European countries, regions and cities, EPOMM, Brussels. Available from: http://epomm.eu/docs/file/epomm_book_2013_web.pdf.
- Guiver, J., Stanford, D., 2014. Why destination visitor travel planning falls between the cracks. *J. Dest. Market. Manage.* 3 (3), 140–151.
- Hoogma, R., 2006. Company/commuter mobility management in the Netherlands, European Conference on Local Energy Action, Brussels, 7–8 February.
- McCrum D., Cundy A, Evans J and Gray A, (2020.) The workers with no choice but to keep on with the job, *Financial Times*, 4 April. Available from: <https://www.ft.com/content/48e4a311-6a7c-4f80-b541-7893a5f97ea4>. Last accessed 6 April 2020.
- NHS Estates, 2006. *Sustainable Development: Transport*. Department of Health and Social Care.
- National Business Travel Network and Department for Transport, 2008. *The Essential Guide to Travel Planning*, Department for Transport, London.
- Ramos, R., Cantillo, V., Arellana, J., Sarmiento, I., 2017. From restricting the use of cars by license plate numbers to congestion charging: Analysis of Medellin, Columbia. *Transp. Policy* 60, 119–130.
- Roby, H., 2008. Viewpoint: ‘Policymakers may see travel plans as a ‘green’ tool, but do employers see them the same way?’. *Local Transp. Today* 498, 18.
- Roby, H., 2010. Workplace travel plans: past, present and future. *J. Transp. Geogr.* 18 (1), 23–30.
- Rye, T., 2002. Travel plans: Do they work? *Transp. Policy* 9 (4), 287–298.
- Shreffler, E.N., 1986. Transport management organisations: An emerging public/private partnership. *Transp. Plan. Technol.* 10 (4), 257–266.
- Transport for London, 2005. Travel planning is the smart choice for London businesses, Business on the Move Conference, Guildhall, London, 28 April.
- Transport for London, 2006. *Developing and implementing travel plans: A good practice guide for the NHS in London*. Available from: <https://www.eltis.org/sites/default/files/tool/nhs-travel-plan-guide-part-1.pdf>.

Home Deliveries and Their Impact on Planning and Policy

Rosário Macário, Av Rovisco Pais 1, Lisboa, Portugal

© 2021 Elsevier Ltd. All rights reserved.

Introduction: The Concept	413
Factors Influencing Home Deliveries	413
Political Factors	413
Economic Factors	414
Social Factors	414
Technological Factors	415
Environmental Factors	415
Legal and Regulatory Factors	415
Typology of Home Deliveries	416
Home Deliveries Within the Broader Urban Logistics Framework	416
References	417
Further Reading	417

Introduction: The Concept

Sustainable Urban Mobility is today a significant concern from almost all cities in the developed and developing world. For a long time it was exclusively focused on the movement of people, but in the last years a growing consensus was established that urban mobility must encompass the sustainable movement of freight, without which living in urban areas is impossible. It is thus essential to understand how freight distribution integrates within the transport system and look at Sustainable Urban Logistic Plans (SULP) as a mean to manage urban logistics.

At the EU level, the President of the European Commission Ursula von der Leyen announced the European Green Deal (European Commission, 2019), within the priorities for 2019–24. This represents a growth strategy aiming to transform EU into a prosperous and equitable society, with efficient utilization of resources and competitive, that aims to reach 2050 with zero emissions and with economic growth decouples from resource utilization. A set of measures and objectives have been set, namely, the transition to sustainable and intelligent mobility, leading the transition to a clean and circular economy.

Since the 80s, several research projects were developed addressing many aspects of the organization of freight distribution in the urban environment. It is worth highlighting from the outset the SULP planning cycle developed by the NOVELOG project (Fig. 1) that identifies the six steps of the planning cycle for urban logistics, divided into two main phases, the preparation and the goal setting.

Home deliveries are a concept covering the movements encompassing the distribution of freight for household consumption. The citizens often initiate home deliveries process ordering their needed goods (e.g., from supermarkets or other retail shops). Other cases, the supplier can begin the process. Home deliveries are one of the most critical segments of urban freight distribution, currently growing in direct relation with the development of the digital world and Internet shopping, and one of the trends with high impact on freight traffic in urban areas. Eurostat reports that “60% of people in the EU aged 16–74 shopped on-line during the year prior to the 2019 survey, compared with 56% in the 2018 survey. Compared with 2009, the share of on-line shoppers had almost doubled from 32%” (European Commission, 2020).

Home deliveries result from consumers buying physical goods (not services) from on-line stores on the Internet (i.e. business to consumer or B2C). It is often confused with e-commerce, which is a broader concept entailing any commercial transaction between organizations and persons. Besides B2C, some websites (e.g., eBay) promote the trade of second-hand goods or homemade products, between consumers (consumer to consumer), also included in home deliveries.

Factors Influencing Home Deliveries

It is crucial to understand which factors influence home deliveries. For this analysis, we use the PESTEL methodology, which acronym represents the domains of analysis, being: Political; Economic; Social; Technological; Environmental; Legal, and regulatory.

Political Factors

Transport and logistics is a sector with high contribution to greenhouse gas emissions as it uses mostly fossil fuels (ref year 2020). The EU Commissioner Velez (European Commission, 2018) puts forward the priority area for the EU, being:

- Promoting the adoption of clean vehicles and alternative fuels.
- Increase the modal share of the most sustainable modes in the transport chain.



Figure 1 SULP planning cycle. Source: NOVELOG.

There is a strong political interest in shifting freight mobility to more sustainable solutions, and this is a driver for the organization of home deliveries within broader sustainable plans. The SULP – Sustainable Urban Logistic Plans have a recent history. The first EU project dedicated to the development of SULP dates from 2012 to 2015 (ENCLOSE), and more information is available in TRIMIS portal.

Economic Factors

Freight transport, of any type, is closely related to the economic activity of the area. In periods of economic growth, the population increases its available income and, consequently, the consumption of goods and services increases and, in turn, freight transport demand increases. The opposite is also true when economic contraction occurs. The World Economic Forum reports that from 2014 to 2019 e-commerce tripled (WEF, 2020), and no figures are available for Home Deliveries as defined above. However, the graphic presented in Fig. 2 presents the evolution of e-commerce items considering two dimensions, for the years 2017 and 2019: vertical axe reflects the quantity of items researched on-line; and, the horizontal axe reflects the percentage of on-line shopping. The increase of online shopping significantly fostered by digitalization leads to an increase of freight distribution not only in the direct delivery but also in returns due to any misfitting reason.

Main delivery points are: households, convenience stores, workplaces, and lockers. However, the real impact of urban logistics dynamic is yet to be accurately calculated, although we know that the type of vehicle used in the distribution is critical for the level of emissions produced. Besides, the more home deliveries increase more situations of potential conflict and accidents accrue with pedestrians and other vehicles, as well as a significant contribution to congestion.

Social Factors

European Union has a high urbanization rate. According to Eurostat I in 2014, 380 million inhabitants lived in urban areas. There is a clear increasing trend, and the projection for 2050 is to have between 450 and 550 million inhabitants (JRC, 2019).

Moreover, there is an intensive ageing process in our societies. Consumption increases with increased urbanization and consumption patterns change with ageing population. In both cases, home deliveries will increase.

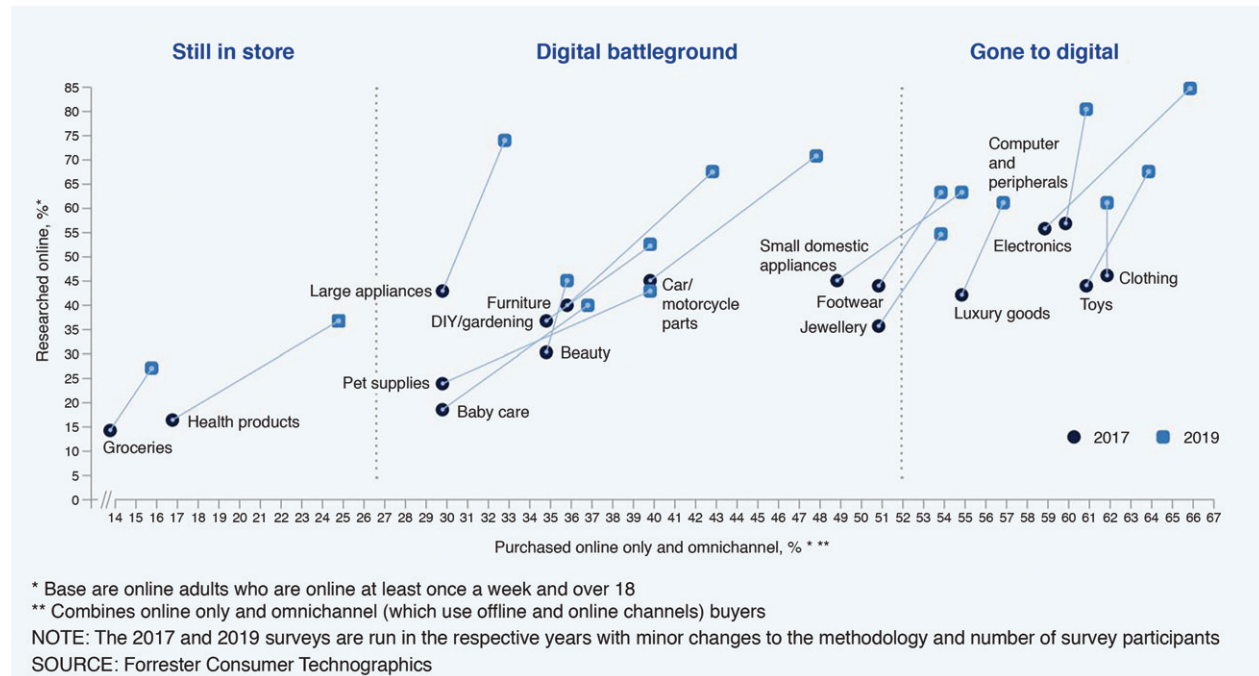


Figure 2 Increase of e-commerce from 2017 to 2019. Public source: WEF.

The speed of living is also changing, and this is made evident in consumption patterns of the new generations. The industry tries to match the needs of society and respective changes. During key informant interviews the following examples were reported as being the reality of 2019: Zara (from Inditex) changes clothes portfolio every 4 weeks. All products are ephemeral; Amazon implemented deliveries in 24–48 h, and delivers in 1 hour in several cities; while, Farfetch delivers clothes anywhere in the world in 24–48 h.

Technological Factors

Technological issues have been a driver for the development of new services for home deliveries. The domains where more intense impact is detected are: information systems; vehicle technology; infrastructure technology; operational management; and, monitoring.

Technology enables the sophistication of home delivery services, in particular in the communication with clients building a trusting relationship with the service.

Environmental Factors

The environmental domain is critical. Transport is responsible for 27% of greenhouse emissions and equivalent consumption of nonrenewable energy (EEA, 2019). Urban logistics is responsible for about 23% of carbon emissions in UE. No specific information exists for home delivery, but the trend is to maintain pressure over the whole urban logistic segments for the reduction of negative externalities.

Internalization of externalities implies an increase of direct operational costs for the transport operator. Besides, the global complexity of the urban system and existing restrictions (e.g., traffic, accesses, parking) provides large room for improvement in efficiency, although incentives are required to succeed.

Legal and Regulatory Factors

Intervention of public entities in urban logistics in general and, in particular, home deliveries, is recent. Regulation is still in an infant phase, but both internal (supported by the agents) and external costs (supported by the society) are drivers for change. It is recognized that important regulatory gaps exist, made evident namely through the growing pollutant emissions in urban area or the contribution to congestion with additional vehicle kilometers done by the deliveries whenever a lack of infrastructure to loading and unloading exists. From the operators perspective there is also a reduced efficiency of services and logistic operations whenever vehicles are in operation with a reduced load factor, or with overlapping routes, or forced to long periods of loading and unloading.

Despite being a sector primarily dominated by private entrepreneurs there are still many measures that can be implemented by the public authorities that can contribute to improve the overall performance of the sector. These measures should consider both the internal costs (beared by the operators) and the external costs (beared by the society).

Typology of Home Deliveries

Business models for home deliveries are blooming since 2010. Many innovative models have been developed. Papoutsis and Nathanail (2018) provide an extensive list of alternative measure for home deliveries, with identification of the research project where more detailed information can be found regarding cases where each measure was implemented.

We highlight here some of the most relevant measures, although we insist that reading the reference provides added value to the reader.

The most relevant measures for home deliveries are:

<i>Measure</i>	<i>Type of measure</i>	<i>Location of implementation</i>	<i>Research project for more details</i>
Pooling of inner city freight delivery	Cooperation	Aachen	C-LIEGE
Electric delivery vehicles for final deliveries	Vehicle Technology	Paris	TURBLOG
New delivery service exclusively using “cargo cycles” or electrically powered tricycles	Infrastructure and vehicle technology	Paris	TURBLOG
Pack Stations 24/7 (e-commerce) with lockers	New business model	Szczecin	C-LIEGE
Private logistic services, Dabbawala	New business model	Mumbai	TURBLOG
Time windows for deliveries	Regulatory	Bristol	START
Lockers	Technology and regulatory	Paris	Technology and regulatory

Home Deliveries Within the Broader Urban Logistics Framework

It is rare to find examples of integrated approaches to regulation and strategic planning in urban logistics the most common situation is to implement isolated measures with specific objectives. An integrated approach would imply tackling all the aspects that are upstream and downstream the act of freight distribution (see Fig. 1) and understand the effects that those aspects have in the performance of urban logistics as a whole.

The study of urban freight tasks is complex and heterogeneous, as it is strongly related with many other aspects of the urban system, such as: urban passenger system, land use, regional development, employment, etc. When considering urban freight planning it is necessary to devote some effort towards understanding its integration within urban mobility planning. For example, considering metropolitan areas urban freight must be considered in the master plan and a roadmap should be developed with guidelines for the development of business and commerce, in particular for home deliveries which have often a long supply chain that must be integrated without frictions between the different transport modes used (Reis and Macário, 2019). In a small town, access and parking regulations can be useful but it is insufficient, so innovative measures have to be considered to reduce the flow of vehicles associated with home deliveries. Some of the good practices that have provided good results are the establishment of pick-up points, or intelligent lockers, etc., Should not be forgotten that each city has its one life cycle and the respective pace of development which represents an indication of the city capacity to change.

In what concerns urban freight we can find three stages of development (Macário, 2013) for every city: the first stage, the restriction stage, characterized by urban areas that show no urban logistic policies and where regulation is only restrictive; the second stage, the new practices stage, when policy documents start being published and opening innovation avenues; the third stage, policy and practice stage, when cities start organizing their SULP. The following table identifies macro measures where home deliveries are included (Macário, 2013).

Table 1 Measures for different stages of development with relevancy for home deliveries

<i>Urban stage of development</i>	<i>Type of measure</i>	<i>Examples</i>
Restriction stage	Access restriction measures	Access restrictions according to vehicle characteristics (weight and volume)
Restriction stage	Legislative and organizational measures	Cooperative logistic systems, encouraging night deliveries, public private partnerships, intermediate delivery depots
New practice stage	Territorial management	Creation of loading and unloading areas, of load transfers, mini logistics platforms, and drop-off points
New practice stage	Technological measures	GPS, track and tracing systems, route planning software, intelligent transport systems, adoption of nonpolluting vehicles and vehicles adapted to urban characteristics (size and propulsion)
Policy and practice	Infrastructural measures	Implementation of urban distribution centers and peripheral storage facilities
Policy and practice stage	Policy oriented	Identification of logistic profiles, development of Master plans, recognizing the need for different solutions for different market segments

Rationalization of the freight distribution process (from the economic, spatial, and temporal perspectives) is required. This means reducing the flow of goods yet keeping the adequate level of distribution to satisfy consumer's needs.

References

- Macário, R., 2013. Modelling for public policies inducement of urban freight business development. In: Ben-Akiva, M., Meersman, H., van de Voorde, E. (Eds.), *Freight Transport Modelling*.
- McKinnon, A., 2018. *Decarbonizing Logistics: Distributing Goods in a Low Carbon World*, 1st ed. Kogan Page.
- Papoutsis K., 2020. Retail logistics cost and policy impact: what is the total cost to secure innovation for a greener supply chain, University of Antwerp.
- Reis, V., Macário, R., 2019. *Intermodal Freight Transportation*. Elsevier Publisher.
- WEF, 2020, The Future of the last mile ecosystem (http://www3.weforum.org/docs/WEF_Future_of_the_last_mile_ecosystem.pdf).

Further Reading

- European Commission, 2018. Study on urban logistics - "The integrated perspective", https://ec.europa.eu/transport/themes/urban/studies_da.
- European Commission, 2018. https://ec.europa.eu/commission/commissioners/2019-2024/valean/announcements/speech-commissioner-valean-tran-committee-european-parliament_en.
- European Commission, 2019. https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_pt.
- European Commission, JRC, 2019. The future of cities, <https://urban.jrc.ec.europa.eu/thefutureofcities/urbanisation#the-chapter>.
- European Commission, 2020. <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/DDN-20200420-2>.
- EEA-European Environmental Agency, 2019. <https://www.eea.europa.eu/data-and-maps/indicators/transport-emissions-of-greenhouse-gases/transport-emissions-of-greenhouse-gases-12>.
- NOVELOG, 2019. New cooperative business models and guidance for sustainable city logistics, novelog.eu.
- TRIMIS, 2018. <https://trimis.ec.europa.eu/project/energy-efficiency-city-logistics-services-small-and-mid-sized-european-historic-towns>.

Planning for Safe and Secure Transport Infrastructure

Per Erik Gårder, University of Maine, Orono, ME, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	418
Modes and Infrastructure	419
Risks When Traveling by Different Modes: Rail and Air	419
Risks When Traveling by Automobile	420
Risks for Unprotected Road Users	420
Pedestrian Safety	421
Conclusions and Discussion	423
Biography	424
See Also	424
References	424

Introduction

What should be the transport policy with respect to safety and security? If we start by looking at what constitutes reasonable policies, it makes sense to strive toward safe and secure systems for all modes of transportation. Few, if any, politicians or other decision makers would advocate for an unsafe or nonsecure system. Therefore, the more difficult question is to define what the term “safe and secure” means, and how we can plan for and build infrastructure that gives us safe and secure systems.

Safety and security policies may be set on local, regional, national, and international levels. An example of an international-level “safety” convention is the Rights of the Child, adopted in 1959 by the UN General Assembly, which defines children’s rights to protection. The act certainly does not have a focus on traffic safety, but one of the four core principles of the convention is the child’s right to life, survival, and development. Another of the core principles is “devotion to the best interests of the child,” and a third is “respect for the views of the child.” The convention is supported worldwide with barely any exceptions. The United States has signed the convention, but it has not been ratified by the US Congress, and the United States is thereby the only United Nations member state that is not a party to it. There are also conventions and standards that are international but not covering all or most of the world, such as European Union agreements.

As the title indicates, this article has a focus on transportation infrastructure. For a system to be safe and secure, it is obvious that one cannot see infrastructure in an isolated way. Motor-vehicle crashes, bicycle and pedestrian falls, and to a lesser extent rail, ferry, and airplane crashes, are typically caused by human error. But the likelihood of human error can certainly be influenced by infrastructure design. It is often stated that 90% of automobile crashes are caused by human error, but 90% of the crashes can be avoided through engineering measures. Though that latter part of the statement includes changes both to vehicles and to roadways, so infrastructure cannot do it all alone. The emphasis of this article is on personal safety and to some extent security. Long-term security with respect to preserving our current transportation infrastructure also could have a focus on climate change and global warming. With oceans rising, much of our road system as well as rail and airports may be flooded. These aspects are not covered in this article.

If a person wants to travel from A to B, their desire is typically to be able to do so as quickly as possible—or in a reasonable amount of time—with high comfort, at a low cost. They may also desire that it is an interesting experience, that it gives them exercise (if walking or riding a bicycle), and that they are not exposed to high levels of pollution. People typically assume that they will get to their destination safely, except that some people are fearful of walking or riding a bicycle through an “unsafe” neighborhood. Some people also have a fear of flying and think there is a high probability that their airplane will crash. Such fears can definitely be justified, especially if they fly by helicopter or a single-engine small propeller plane, but commercial jets are the safest mode of transport we have. Still, even commercial airplanes do crash, and when two Boeing 737 Max airplanes recently crashed—Lion Air in October 2018 and Ethiopian Air in March 2019—not only did people fear that the specific model of plane was not safe, but authorities around the world also acted and within days of the second crash grounded all planes of that type. As travelers, we demand that airplanes are “safe,” but we accept much higher risks when traveling by other modes, especially those where we are in control as operators.

We are exposed to multiple risks to our health when traveling. These include accidents or involuntary crashes, voluntary crashes with the intent to commit suicide, or homicide, caused by operators—for example, drivers or pilots—and crashes caused by terrorists. Other events during travel that can hurt us physically or psychologically include assaults of different seriousness—from verbal abuse to rape and murder—and the risk of catching a disease, for example, a virus that leads to an illness such as COVID-19. When this article is written, in the early fall of 2020, COVID-19 has gone from being widespread only in a small part of China to entering nearby countries and in March 2020 becoming a pandemic spread more-or-less everywhere on earth. It has moved from

continent to continent primarily by airplanes but also by cruise ships, and then locally by people traveling from town to town. For travelers themselves, walking or bicycling, as well as traveling in a private automobile, are considered safe with respect to catching the virus as long as the person is physically distanced by 2 m or more from others, whereas traveling in confined spaces in airplanes and other public modes of transportation is considered a risk factor for transmission. The fact that the virus is spread while traveling is not primarily an infrastructure-related issue and will therefore not be covered here in detail. However, terminal and station designs can have a great impact on such risks. In general, outdoor environments are considered safer than indoor rooms, especially if the latter do not have adequate ventilation. Air conditioners circulating but not filtering the air adequately may be the biggest “infrastructure” risk factor. The width of sidewalks and multiuse paths can certainly also influence a person’s ability to physically distance themselves from other people. So is the number of users of a system.

There are many other health issues relating to traffic. Exposure to air pollution can lead to cancer and many other illnesses. The World Health Organization (WHO) estimated in March 2014 “that in 2012 around 7 million people died—one in eight of total global deaths—as a result of air pollution exposure” (WHO, 2014). Traffic is just one of many sources of air pollution but a significant part. A 2013 EPA study attributed an estimated 29,000 premature deaths in the United States alone to ozone and PM_{2.5} emitted by road traffic (Fann et al., 2013). Moving roads away from where people live can reduce the death rate. However, going to electric vehicles would have a much greater effect, and building an electric distribution network is an infrastructure-related decision.

In addition, findings by the WHO show that chronic exposure to noise in cities can make people sick and can even kill people. Traffic noise alone may account for 3% of deaths from heart attacks and strokes in Europe. Given that 7 million people around the world die each year from heart disease, the global death toll from traffic noise could be around 210,000 (MetroUK, 2008). Installing noise barriers along highways and locating homes and airports apart are examples of infrastructure investments making a difference.

Modes and Infrastructure

This article covers infrastructure and not vehicle or operator characteristics. Safety and security aspects of infrastructure include:

- Walking: Design of multiuse paths, footpaths, sidewalks, and roadways
- Bicycling: Design of multiuse paths, bike paths, bike tracks, bike lanes, and mixed roadways
- Travel by private motor vehicle, including two-wheelers and trucks: Roadways and rest areas
- Travel by bus: Design of bus stops, bus terminals, and roadways
- Travel by train: Design of train stations and railways, including grade crossings
- Travel by airplane: Design of airports
- Travel by ferry and cruise ships: Design of ports and terminals

The short overview here cannot be exhaustive but will touch on some of these issues.

Risks When Traveling by Different Modes: Rail and Air

It is obvious that people accept some type of risk when they undertake certain activities, such as mountain climbing or downhill skiing. People traveling by car are also subject to a clear risk of getting injured, and since they typically undertake trips “voluntarily,” it can be argued that they implicitly accept the risks they are subjected to.

It also can be argued that people accept higher risks when they are in control of operating a vehicle than if they are the passenger of one. If we injure ourselves when driving a car, and it was primarily our own fault that the crash happened, we may not feel that we should blame the road administration even if what caused the injuries were collision with trees within the right-of-way area, and those trees could have been removed. On the other hand, if someone is killed when traveling by bus or airplane, many would blame the driver as well as the company operating the system, or the manufacturer of the vehicle/airplane, irrespective of if the driver/pilot were partially at fault or not. Why is that? Why should travelers not be equally safe when operating a mode themselves as when being passengers?

Let us look at the risk of getting killed when using some of our most common transportation systems. The risk of serious injuries should also be covered, but for brevity, only fatalities are discussed here partly because the definition of death is more uniform, and data quality is typically higher.

In EU-28, a total of 219 rail passengers were killed in the 5-year period, 2012 to 2016 (European Union Agency for Railway, 2018), giving an average death toll of 44 per year. That includes passengers killed in collisions and derailments as well as falling on board or when alighting or boarding trains and so on, excluding only confirmed suicides. Overall, with 44 fatalities per year, EU-28 saw around 0.086 fatalities per million inhabitants. For the average person, the annual risk of a fatality—the mortality rate—would be around 1 in 12 million. Is that safe? Another way of looking at it is that the average European in 2018 traveled 925 km/year by train (calculated as 472 billion km/510 million people), giving an average fatality rate to go by train in Europe of 0.09 fatalities per billion passenger-kilometer, or one fatality per 10.8 billion kilometers. Is that “safe”?

If we look at scheduled commercial air traffic in the United States, to get another estimate of actual fatality risk, and include all crashes and other accidents since January 2002, we have had 131 people killed on US airlines. Only one person was killed in a “large” US airliner between February 2009 and the writing of this article. Another three people were killed in the United States traveling by a

foreign airline (Asiana Airlines) when it crashed in 2013 for a total of 134 fatalities in the United States in the last 18+ years. So, from 2002 to the beginning of 2020, 18 years, we saw on average 7.4 fatalities/year, or around 0.025 fatalities per million inhabitants in the United States. And, if using 2018 US passenger travel mileage, 730.7 billion passenger-miles by plane, which is 1176 billion passenger-kilometer, as exposure, we get an estimate of passenger flight distance. Travel volumes have increased over the 18 years, and the average travel mileage may be around 1000 billion km/year. That means we saw around 0.007 fatalities per billion km, or 1 fatality/140 billion km. Is that safe? Well, it is about 13 times safer than being on a European train. And, if we look at the latest 10 years only, with four fatalities on commercial airplanes in the United States, we would be at around 0.0004 fatalities per billion person-kilometer. The risk of a fatality in the last 10 years would then be around one per 3000 billion km. Maybe we could agree that this is “safe”?

How many additional deaths related to European rail or US air traffic occur if we include “passengers” killed at terminals and stations? These fatalities could be caused by accidents or terrorist attacks or natural disasters such as hurricanes and earthquakes. The answer is that very few people have died in such events in recent years on these continents. However, if we include transport to and from airports and terminals, as motor-vehicle passengers or pedestrians, there are numerous fatalities every year.

Risks When Traveling by Automobile

Most travel do not involve US airlines or European trains, which were discussed earlier. If we look at automobile travel, do motor vehicles keep their occupants “safe”? One meaning of “safe” would be having a fatality rate below or similar to European trains, 0.09 fatalities per billion km? Or better, below the risk of flying in recent years in the United States, with 0.0004 fatalities per billion km. Let’s look at the US roadway system, using 2017 data. The United States that year had an official VMT of 3,212,670 million vehicle-miles according to the Federal Highway Administration (FHWA). With an average of 1.3 occupants per vehicle, that translates to 6,719,940 million people-kilometer. And, if we accept only 0.0004 fatalities per billion km of travel—to be as safe as air travel—we would accept only three automobile-related fatalities in the United States in 2017. Alternatively, if we accept our highway motorists to be no less safe than people in European trains, we would accept up to 0.09 fatalities per billion passenger-kilometer or 605 people killed. In reality, the United States saw over 37,000 fatalities in 2017 and 23,551 of those were occupants of motor vehicles according to the US National Highway Traffic Safety Administration (NHTSA). That is over 8000 times higher than what we could have as an objective required safety level, the safety level reached by commercial airlines in the United States or around 40 times higher than that of trains in Europe. Safety-wise, European roads do a bit better than US roads but are still far riskier than the potential desired levels as outlined earlier. The EU has seen right around 12,000 motor-vehicle occupant fatalities per year in later years (2013–16).

How can we make travel by automobile safer? In most countries, a great number of crashes are of rear-end type. These can be reduced by having low-speed variation between different users and along roads. Abrupt and drastic changes of speed is the worst, such as congestion on high-speed roads and change of speed from “high” to zero as is the case when a signal turns amber and then red. Signalization should therefore probably not be used where speeds are above 50 km/h. Well-built roundabouts give predictable speeds, since everybody slows down to 30 km/h or less even when there are no conflicting vehicles. In addition, high-speed highways should have variable speed limits, strictly enforced, that are reduced when traffic volumes are high. However, rear-end crashes cause few serious injuries and deaths. Rather, lane departure crashes make up a majority of fatalities. These can be divided into run-off-road and head-on collisions. Both these types can be mitigated with rumble strips, see for example [Gärder and Davies \(2006\)](#). However, rumble strips will far from eliminate these crashes. Cable barriers or concrete New-Jersey-style barriers are much more effective in eliminating head-on collisions and keeping vehicles away from unsafe roadside areas and should be used not only on multi-lane highways.

Risks for Unprotected Road Users

Unprotected road users have higher risks than automobile occupants. That is true for nonmotorist as well as for powered two-wheelers. If we look at motorcyclists, NHTSA reported that 5,172 motorcyclists and 23,708 passenger-vehicle occupants were killed in 2017 in the United States in spite of motorcycle crashes accounting for only 0.6% of vehicle kilometers traveled on America’s roads ([FHWA, 2019](#)). The risk per kilometer is much higher for motorcyclists than for automobile occupants. Still, passenger-car travel can be seen as the greater threat to public health since more people are killed in cars than on motorcycles.

If we look at the safety of nonmotorized, active modes of transportation, there is great variation between countries and between age groups as well as gender. Findings by [Buehler and Pucher \(2017\)](#) show fatality rates per kilometer traveled in Germany and the United States. We find that the fatality rate for senior (above age 65) pedestrians is roughly twice as high as for the population as a whole in both the United States (21.5 vs. 9.7 per 100 million kilometers walked) and Germany (3.8 vs. 1.9). Similarly, the fatality rate for senior cyclists is much higher than average in both the United States (7.6 vs. 4.7 per 100 million km cycled) and Germany (4.2 vs. 1.3). Children have lower fatality rates per kilometer walked than the population as a whole in both countries: 2.9 versus 9.7 (United States) and 0.9 versus 1.9 (Germany). Children also have slightly lower fatality rates per 100 million kilometers cycled: 4.1 versus 4.7 (United States) and 0.9 versus 1.3 (Germany). So, if we were to accept only 0.0004 fatalities per billion kilometers traveled—the approximate crash risk when flying in the United States—the risk for walking and bicycling is between 20,000 times too high (for children bicycling in Germany and 500,000 times too high (for older pedestrians walking in the United States). Obviously, the risk varies from location to location, and not all elderly people face the same risks, and risks may be even higher in other countries than

the United States. But most of us would and should agree that people often are not safe when riding a bicycle or walking. Since these modes are the most dangerous of main modes, it makes sense to focus more on these. And, if looking at mortality numbers, walking in particular is a safety issue, so the rest of this article will focus on pedestrian safety and infrastructure needs.

Pedestrian Safety

How can we make our transportation system safer for pedestrians? Obviously, pedestrian safety cannot always be maximized since the safety of bicyclists and motorists may at times conflict with the safety of pedestrians. Efficiency of traffic movements and environmental concerns also need attention. Moreover, construction and maintenance costs are important factors. Having said that, using designs, which are safe for pedestrians, typically also leads to high safety for other road users. Safety for everybody combined with efficient mobility, that is, short delays for motorists accomplished by the concept “quick but slow” often go hand in hand, such as when utilizing roundabouts along arterial streets rather than signalized intersections. One of the first studies to quantify the “quick but slow” concept was performed along South Golden Road in Golden, Colorado in the United States (Isebrands et al., 2008). This study showed a reduction in travel times by 10 s over an 800-m segment (quicker), while the 85-percentile speeds were reduced from 76 to 53 km/h. Safety was improved. Prior to construction of the roundabouts, there were 10 injury crashes per year, and in the four years after the roundabouts were built, only 1 injury crash was reported (Hartman as cited in Isebrands et al., 2008). Studies of pedestrian safety in Europe have shown that installation of single-lane roundabouts reduce pedestrian crashes by about 75% to 80% compared to signalization (Brude and Larsson, 2000; Schoon and van Minnen, 1994).

There are many health benefits connected to active transportation. Walking is probably the form of exercise that has the fewest negative side effects in the form of injuries—as long as injuries caused by collisions with motor vehicles do not happen. Empirical research by the US National Cancer Institute and Harvard University, following 661,000 adults during a 14-year period, found that just over an hour a day of moderate exercise, such as brisk walking, is optimal for longevity (Reynolds, 2015). But if pedestrians are exposed to excessive levels of air pollution from high volumes of traffic, these benefits decrease (Amorim et al., 2013).

There are obviously people with health problems who cannot walk, at least not far. However, if we include riding a wheelchair under the general umbrella of “walking,” most of us can and do walk, at least sometimes. So why do we not walk more? Probably because health benefits are not immediate. We do not want to sacrifice minutes today for a gain later in life. We feel that we do not have time to spend 30 min on walking when we can get there in 5 min by car.

It is important that children have safe and secure opportunities to walk to school. Walking may become a lifelong habit if it is started in childhood. Parents often worry more about a child being kidnapped or assaulted than about traffic, but as is pointed out by the British author Warwick Cairns, author of *How to Live Dangerously*, “. . . an average child would have to stand outside alone for 750,000 hours unsupervised to be kidnapped by a stranger. At that point, the child would be 85 years old” (USA Today, 2015). So, while we inflate some risks, we tend to neglect other risks, such as obesity and diabetes, which may occur if children are driven to their destinations.

People of all ages need to have “inviting” environments to be encouraged to walk, for example, to shops and restaurants. A building's entrance should be at the street with safe access directly from a sidewalk. Parking should be below, behind or on the side of buildings—not in front. It is also important for pedestrians to see others walking in an area. Mixed development—with restaurants, shops, offices, and apartments—means that there will be pedestrian activity during daytime as well as evening hours. Making the facilities attractive can by itself promote safety and security. The risk of assault, rape, or other violent crimes are reduced with pedestrian volume. And if drivers see many pedestrians near a street, they are more likely to slow down and yield to pedestrians. But safe infrastructure is still needed. Education and enforcement is seldom a substitute for safe physical layout but can be important complements.

Pedestrians should not only be allowed to walk “everywhere” but should be encouraged to do so. Providing sidewalks is a part of this. One type of roadways where pedestrians are not allowed is the motorway; however, other major roads should have sidewalks along them whenever there are substantial numbers of pedestrians, unless there are nearby parallel footpaths. The question becomes what constitutes a “substantial” number? Is one pedestrian an hour enough to motivate sidewalks? Can economic arguments allow us to keep operating an arterial-road-lacking sidewalk even if there are hundreds of pedestrians walking along it every day? If so, should the speed along it be restricted to 30 km/h even in rural areas? The current standards in Sweden are that rural roads need separate pedestrian and bicycle facilities when the combined pedestrian and bicycle volume is 100 per day (during summer months). If the pedestrian and bicycle flow is below 100 per day and annual average daily vehicle traffic (AADT) is above 2000, paved shoulders are acceptable. Mixed traffic is acceptable if AADT is below 2000 (Trafikverket, 2012). Finnish guidelines are similar but require separation of rural arterials already at 50 pedestrians or bicyclists a day if AADT is above 2000 (Gårder, 2018). Are those standards good enough? Maybe not if we are aiming for Vision Zero. We probably cannot get to zero risk but could we make walking as safe as flying by airplanes? As earlier pointed out, we would need to reduce risks by 99.99% for that to be the case.

A sidewalk is more attractive if it is set back from the roadway by a grass or vegetation strip (Gårder, 2018). Making the strip at least 2 m wide and having stem trees and/or utility poles separating vehicles from pedestrians will increase perceived safety, and a greater distance will reduce water and mud splashing from car tires. But a strip also increases actual safety if the sidewalk is removed from the clear zone. A *Policy on Geometric Design of Highways and Streets* (AASHTO, 2018), the so-called “Green Book,” defines a clear zone as the “unobstructed, traversable area provided beyond the edge of the travelled way for recovery of errant vehicles.” This guideline allows sidewalks to be within the clear zone, but it is clear that cars may traverse into it with some probability. For local roads and streets, the Green Book specifies, “a clear zone of 2 to 3 m or more from the edge of the travelled way, appropriately graded

with relatively flat slopes and rounded cross-sectional design, is desirable.” For higher-speed rural collector roads and arterials, the Green Book refers to the AASHTO *Roadside Design Guide* (AASHTO, 2011). That publication recommends that clear zones have a width of 9–10 m for flat, level terrain adjacent to a straight section of a 100 km/h highway with an average daily traffic of 6000 vehicles. The guide further states that for horizontal curves the clear zone should be increased by up to 50%. It would be prudent to not have sidewalks within this zone unless a barrier separates it from the road.

However, being on a sidewalk, no matter how close to a road, is not the most dangerous location for a pedestrian. Many crashes happen when pedestrians walk along roads lacking sidewalks and others when pedestrians cross roads. Several studies have shown that the lowest level of safety for pedestrians is found when they cross streets (Elvik et al., 2009). In order to reach a destination on foot, people need to cross streets and roads. The goal should be to provide “absolute” safety for pedestrians when crossing streets. Grade-separated crossings are not only expensive but may also not be used. If absolute safety cannot be provided, what is an acceptable safety level? That was discussed earlier, but Vision Zero goals—no fatalities and no serious injuries—should obviously remain the long-term goal.

In some jurisdictions, pedestrians have legal priority wherever they cross a roadway adjacent to an intersection. That is also the case in between intersections in some living-street environments (Woonerven, Home Zones, and so on) and in shared-space areas. However, in most locations, pedestrians have formal priority only in marked crosswalks and even there they may not be given priority. It can be noted that there are many gap acceptance studies where researchers have looked at pedestrians accepting and rejecting gaps between cars at marked crosswalks, but a few, if any, studies where someone looks at drivers accepting and rejecting gaps between pedestrians in crosswalks. Most of the time it is obviously up to pedestrians to find gaps even when they have the (formal) right-of-way. The absurdity of how pedestrians are treated needs to be highlighted. Pedestrians in marked crosswalks should not have to wait for a gap. Drivers crossing the crosswalk should look for gaps between pedestrians. Until pedestrians are given priority, crosswalks may cause more harm than good, since they often instill a false sense of safety, as highlighted more than 40 years ago by Herms (1972). Rather, we should provide crossing points at locations where the roadway is not wider than one (narrow) lane between raised refuge islands, at places where cars are traveling at “low” speed, probably never above 30 km/h. Encouraging pedestrians to cross where top speeds are above 50 km/h should never be done unless traffic volumes are extremely low and sight distances are good. If such conditions prevail, marking crosswalks is not needed. Marking crosswalks can lead to more crashes, especially on multilane streets (Zegeer, 2005; Zegeer et al., 2002). That does not mean that we as planners or engineers do not have an obligation to help pedestrians safely cross urban streets as well as rural roads. And that should apply to pedestrians with visual impairments as well as mental or physical handicaps. Geometric features that help pedestrian safety include more or less anything leading to lower speeds and shortened crossing distances (Gärder, 2004; 2018).

One reason speed is important for safety is that pedestrians accept shorter vehicle time gaps when speed increases. Pedestrians, especially the elderly, tend to go for constant distance gaps (Yannis et al., 2013). Lower walking speeds of the elderly and people with disabilities is also an issue. For drivers, higher speed is correlated to an increased likelihood and severity of a crash, but for pedestrians lower crossing speeds lead to greater risk (Shinar, 2008). Low vehicle speeds can be achieved simultaneously with narrow crossing distances if the roadway width is so narrow that it by itself slows down traffic. This is difficult to achieve if we want wide trucks and busses to be able to pass through. Very effective is a 2.1-m curb-to-curb distance, as has been done at some locations in the United Kingdom (see Fig. 1). If curb-to-curb width is much wider than that, for example, 2.8 m, many motorists will drive at speeds well over 30 km/h.



Figure 1 Example of a narrow travel lane slowing down drivers.

Another way to reduce speed is to make the geometry induce horizontal or vertical g-forces on the vehicle so that it becomes uncomfortable—or impossible—to drive fast. Small intersections or mid-block traffic circles can do this, and they can reduce median speeds to around 25 km/h and 85-percentile speeds to around 30 km/h. The Watt's hump as well as speed cushions can also achieve speeds in the 30 km/h range, whereas most speed tables with 8–10 cm height have 85-percentile speeds in the 45 to 50 km/h range (Johansson et al., 2003).

A relatively new invention is the “Actibump,” which is a “trap-door” bump activated only when a vehicle is going a set number of km/h above the posted speed limit. Actibump is a commercially developed product by the Swedish company Edema and that or similar devices can be used at crosswalks where low speeds need to be ensured. An alternative to this product is a system where drivers get a red light when they speed. Los Angeles, California, has experimented with such a system since 2014 on North Figueroa Street, at Amabel Street. There is a Your-Speed activated sign which triggers a downstream traffic signal to turn red if an approaching motorist exceeds the speed limit by more than 8 km/h (5 mph), forcing the motorist to come to a complete stop (S. Skehan, personal communication, March 18, 2015). A system like this can be combined with a red-light-running camera if motorists start ignoring the signal. An advantage with a system like this compared to direct camera enforcement of speeds is that it is politically easier to implement.

A Swedish study from the 1980s found that scramble—that is, pedestrians having a simultaneous walk phase on all crosswalks while all cars are stopped—is ineffective in promoting pedestrian safety unless red-walking frequencies are very low, and that red-walking frequencies become high when wait times are long, which is the case when scramble is used (Gårder, 1989). The study also found that left-turning vehicles are involved in many pedestrian crashes so that there is no easy way of using signalization as a way to improve pedestrian safety (Gårder, 1989). Other studies confirm that left-turning vehicles pose more of a threat to pedestrians than right-turning ones (in countries where people drive on the right). Reasons include higher speeds and opposing traffic-blocking view of the crosswalk (Lord, 1996; Lord et al., 1998).

Analysis of crashes in general shows that few multivehicle crashes happen without any of the parties taking evasive maneuvers prior to the impact. A study by Walz et al. (1983) found that only 18% of the fatal crashes had collision speeds equal to the traveling speed of the striking vehicle. In 62% the collision speed was reduced by at least 20% before striking a pedestrian. And lower speeds lead to higher probability of having started braking. They also reported that a reduction of the speed limit from 60 to 50 km/h was accompanied by a 25% reduction in pedestrian fatalities. However, the classic study on speed safety relationships is by Teichgräber (1983) who reported a probability of death for a pedestrian hit by a car as 10% at 20 km/h, 20% at 30 km/h, 30% at 40 km/h, 60% at 50 km/h, and 95% at 70 km/h. A Swedish study, based on the largest in-depth pedestrian accident study ever undertaken (at least at that time), shows that for adult pedestrians hit by the front of passenger cars, the fatality risk at 50 km/h is more than twice as high as the risk at 40 km/h and more than five times higher than the risk at 30 km/h (Rosén and Sander, 2009). And studies of real traffic show that pedestrians face 50 times higher risk when crossing a four-lane street posted at 55 km/h than when crossing a low-speed, two-lane street where drivers are traveling at speeds around 30 km/h (Gårder, 2004).

In conclusion, pedestrians are usually not killed or even seriously injured when hit by a car moving at a speed of 30 km/h or less at impact. Ensuring low speeds and short crossing distances should make most pedestrians safe. However, pedestrians who are intoxicated, or distracted, may walk out in front of a moving car and be seriously injured at moderate speeds. As previously pointed out, it should not be up to pedestrians to look for gaps between cars, but drivers should look for gaps between pedestrians. ITS functions can be structured to aid pedestrians with this (Leden et al., 2014). Auto-brake and eventually autonomous vehicles will lead to higher safety, but we need to design our infrastructure to be safe before that happens.

Conclusions and Discussion

People should be encouraged to use the safest available modes. That is typically trains and airplanes. Providing infrastructure for these modes should be a policy priority around the world. Automobile traffic is today far from safe in all countries, and we need to improve this, partially by redesigning our infrastructure. However, we also need to encourage walking as a mode of transportation since walking promotes better health and produces no air pollution. It must be “safe” to walk. Arterial and collector roads must have sidewalks. The sidewalks should be separated from the roadway by a curb and/or wide separation strip in higher-speed areas. Safe crossing points need to be provided. Pedestrians' interests rather than automobile traffic should be the policy priority.

The numbers referred to in the introductory sections relate to unintentional crashes, so-called accidents. If we add intentional crashes, suicides and homicides, the numbers do not change drastically. The same is the case if we add terrorism and piracy. Carjackings are fairly common in some countries, including South Africa and the United States, but a few of them lead to murder or death in any other way. Of course, death is not the only outcome that should be analyzed, and worldwide there are annually thousands of carjackings leading to traumatic experiences. Road piracy where people are robbed is also a safety issue. Piracy on our oceans where ships are captured sometimes leads to people being kidnapped and held for years, and in some years scores of people are killed.

If we look at transportation related illnesses that injure or kill travelers, we should look at excess injuries and deaths, rather than total injuries and deaths that occur while people are traveling. People spend, on average, about an hour a day traveling, mostly by motor vehicles. Obviously, people can have medical issues such as heart attacks and strokes during these times. However, the incidence rates of some illnesses may be higher when traveling than at other times. For example, sitting still in airplanes for long periods may lead to higher incidents of blood clots than “normal” activities. Road rage, or stress, while driving or bicycling can lead

to increased blood pressure and pulse rates, which in turn lead to stroke and other illnesses. However, infrastructure design should seldom be blamed for this.

Biography



Per Erik Gårder is a professor of Civil Engineering at the University of Maine, US, since 1992. He has a PhD from Lund University in Sweden, and he was, from 1983 to 1992, a faculty member of the Royal Institute of Technology (KTH) in Stockholm, Sweden. His research interest focuses on forecasting, designing, and evaluating facilities with emphasis on traffic safety.

See Also

The Concept of “Acceptable Risk” Applied to Road Safety Risk Level; Aggressive Driving and Road Rage; Bicycle and E-Scooter Safety; Automatic Vehicles and Connected Vehicles; Aviation Safety—Commercial Airlines; Carjacking; Motorcycle and PTW Safety; Railroad Safety; Recreational Boating Safety; Safety Culture; Speed Limits on Urban Streets; Suicides

References

- AASHTO, 2011. Roadside Design Guide, fourth ed. American Association of State Highway and Transportation Officials.
- AASHTO, 2018. A Policy on Geometric Design of Highways and Streets, seventh ed. American Association of State Highway and Transportation Officials, Washington, DC (referred to as Green Book).
- Amorim, J.H., Joana Valente, J., Cascão, P., Rodrigues, V., Pimentel, C., Miranda, A.I., Borrego, C., 2013. Pedestrian exposure to air pollution in cities: modeling the effect of roadside trees. *Adv. Meteorol.* 2013, 7. Article ID 964904. Available from: <https://www.hindawi.com/journals/amete/2013/964904/>.
- Brude, U., Larsson, J., 2000. What roundabout design provides the highest possible safety? *Nordic Road Transp. Res.* 12, 17–21.
- Buehler, R., Pucher, J., 2017. Trends in walking and cycling safety: recent evidence from high-income countries with a focus on the United States and Germany. *Am. J. Public Health* 107 (2), 281–287.
- Elvik, R., Høy, A., Vaa, T., Sørensen, M., 2009. The Handbook of Road Safety Measures, second ed. Emerald Group, United Kingdom.
- European Union Agency for Railway, 2018. Report on Railway Safety and Interoperability in the EU 2018. Available from: https://www.era.europa.eu/sites/default/files/library/docs/safety_interoperability_progress_reports/railway_safety_and_interoperability_in_eu_2018_en.pdf (accessed 12.02.2020).
- Fann, N., Fulcher, C.M., Baker, K., 2013. The recent and future health burden of air pollution apportioned across U.S. sectors. *Environ. Sci. Technol.* 47 (8), 3580–3589.
- FHWA, 2019. US Department of Transportation Federal Highway Administration Website: Motorcycle Safety in Research. [Updated: Monday, December 2, 2019]. Available from: <https://cms7.fhwa.dot.gov/research/research-programs/safety/motorcycle> (accessed 26.01.2020).
- Gårder, P., 1989. Pedestrian safety at traffic signals: a study carried out with the help of a traffic conflicts technique. *Accid. Anal. Prev.* 21 (5), 435–444.
- Gårder, P., 2004. The impact of speed and other variables on pedestrian safety in Maine. *Accid. Anal. Prev.* 36 (4), 501–596.
- Gårder, P., 2018. Providing for pedestrians. In: Dominique Lord, S. Simon Washington, S (Eds.), *Safe Mobility: Challenges, Methodology and Solutions* (Transport and Sustainability, vol. 11). Emerald Publishing Limited, Melbourne, Victoria, Australia, pp. 207–228.
- Gårder, P., Davies, M., 2006. Safety effect of continuous shoulder rumble strips on rural interstates in Maine. *Transportation Research Record No 1953*, pp 156–162. National Research Council, Washington, DC.
- Hermes, B., 1972. Pedestrian Crosswalk Study: Accidents in Painted and Unpainted Crosswalks. *Transportation Research Record No. 406*, Transportation Research Board, National Research Council. Washington, DC.
- Isebrands, H., Hallmark, S., Fitzsimmons, E., Stroda, J., 2008. Toolbox to Evaluate the Impacts of Roundabouts on a Corridor or Roadway Network. Center for Transportation Research and Education, Iowa State University, Ames, Iowa, USA.
- Johansson, C., Gårder, P., Leden, L., 2003. Towards Vision Zero at zebra crossings—a case study in Malmö Sweden on traffic safety and mobility for children and elderly. *Transp. Res. Rec.* 1828, 67–74.
- Leden, L., Gårder, P., Schirokoff, A., Monderde-i-Bort, H., Johansson, C., Basbas, S., 2014. A sustainable city environment through child safety and mobility—a challenge based on ITS? *Accid. Anal. Prev.* 62, 406–414.
- Lord, D., 1996. Analysis of pedestrian conflicts with left-turning traffic. *Transp. Res. Rec.* 1538, 61–67.
- Lord D., Smiley, A., Haroun, A., 1998. Pedestrian accidents with left-turning traffic at signalized intersections: characteristics, human factors and unconsidered issues. Paper presented at the 77th Annual TRB Meeting, TRB, National Research Council, Washington, DC.

- MetroUK, 2008. Noise kills thousands every year. Available from: <https://metro.co.uk/2007/08/22/noise-kills-thousands-every-year-42454/?ito=cbshare> (accessed 29.05.2020).
- Reynolds, G., 2015. Exercise in just the right dose. New York Times. April 21, 2015, p. D6.
- Rosén, E., Sander, U., 2009. Pedestrian fatality risk as a function of car impact speed. *Accid. Anal. Prev.* 41, 536–542.
- Schoon, C., van Minnen, J., 1994. The safety of roundabouts in the Netherlands. *Traffic Eng. Control* 142–148.
- Shinar, D., 2008. *Traffic Safety and Human Behavior*. Emerald Group Publishing Limited, Bingley, United Kingdom.
- Teichgräber, W., 1983. Die Bedeutung der Geschwindigkeit für die Verkehrssicherheit. *Z f Verkehrssicherheit* 2, 53–63.
- Trafikverket, 2012. Krav för vågars och gators utformning. Publikation 179.
- USA Today, April 27, 2015. Let 'free range' kids roam to their homes. USA Today 7A.
- Walz, F.H., Hoefliger, M., Fehlmann, W., 1983. Speed limit reduction from 60 to 50 km/h and pedestrian injuries. In: *Twenty-Seventh Stapp Car Crash Conference Proceedings with International Research Council on Biokinetics and Impacts (IRCOBI)*, pp. 311–318. Society of Automotive Engineers, Warrendale, PA.
- WHO, 2014. 7 million premature deaths annually linked to air pollution. [Published March 25, 2014]. Available from: <https://www.who.int/mediacentre/news/releases/2014/air-pollution/en/> (accessed on 29.05.2020).
- Yannis, G., Papadimitriou, E., Theofilatos, A., 2013. Pedestrian gap acceptance for mid-block street crossing. *Transp. Plann. Technol.* 36, 450–462.
- Zegeer, C.V., Stewart, J.R., Huang, H.H., Lagerwey, P.A., 2002. Safety effects of marked vs. unmarked crosswalks at uncontrolled locations. Executive Summary and Recommended Guidelines, FHWA-RD-01-075. US Department of Transportation, Federal Highway Administration, McLean, VA.
- Zegeer, C., 2005. Safety effects of marked versus unmarked crosswalks at uncontrolled locations. Final Report and Recommended Guidelines, FHWA Publication Number: HRT-04-100, Turner-Fairbank Highway Research Center, McLean, Virginia, USA.

Sensors and Data Driven Approaches in Transport

Mohammad Sadrani, Constantinos Antoniou, Department of Civil, Geo and Environmental Engineering, Technical University of Munich, Munich, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	426
Travel Time Estimation	427
Origin–Destination Matrices Estimation	427
Safety Analysis	428
Driver Behavior Monitoring	428
Transportation Mode Inference	429
Summary	430
Acknowledgments	430
References	430

Introduction

Sensors are an essential part of any Intelligent Transportation System (ITS), which can play a vital role in collecting real-time information for transportation systems. The applications of advanced sensor technologies to surveillance, control, and management of transit networks have become increasingly widespread in practical terms. Infrastructure-based sensor technologies, counting sensors, image sensors, inductive loop, microwave radar, video camera, and more recently Bluetooth, Wi-Fi signal and smartphone-based sensors are the most commonly-used technologies for long- and short-term traffic detection (Antoniou et al., 2011; Zhao et al., 2019). Both traffic operation agencies and highway users can reap the benefits of having access to comprehensive and precise knowledge of the real-time traffic information. For example, from the viewpoint of traffic management agencies, designing network-level travel time surveillance can bring real-time information on both temporal and spatial traffic state patterns, leading to a more precise travel behavior analysis. In addition, one of the simplest ways for traffic operation agencies to detect bottleneck locations and the spatiotemporal propagations of the traffic queues is to use raw traffic data (e.g., vehicle count, time–mean speed, or occupancy) collected from fixed traffic sensors. Accordingly, in the past few years, a large amount of literature has been devoted to the problem of optimally locating sensors on a traffic network in order to measure traffic flows accurately (e.g., Bianco et al., 2001; Hu and Liou, 2014).

In general, traffic sensor technologies can be categorized into three different groups, including intrusive, nonintrusive, and off-roadway technologies (Bottero et al., 2013). Intrusive sensors are fixed in or on the pavement, whereas nonintrusive traffic sensors can be installed overhead or on the side of the pavement with minimum disruption to normal traffic operations. Off-roadway technologies refer to those that do not need any specific equipment to be mounted under the pavement or on the roadside, for example, this group involves probe vehicle technologies with GPS – WAAS (EGNOS, in Europe) and mobile phones, which can be also combined with automatic vehicle identification (AVI) systems, and remote sensing technologies using images from satellite. The data collected using sensor systems can then be used to monitor traffic congestion levels and incidents on freeways, predict travel times, estimate OD flows, and support decision making for transport agencies (Sun et al., 2016).

With the rapid development of new technology, there are several tools to obtain detailed information through on-vehicle equipment, such as Wi-Fi and Bluetooth-equipped devices, which can provide high-resolution data for the entire trip of vehicles equipped with such sensors. Indeed, the sensors can capture the public parts of the Bluetooth or Wi-Fi signals within their coverage radius, which can make it possible to use several different matching algorithms to detect and track a device, when it is connected to a sensor. For example, if a vehicle, which has been equipped with a Bluetooth device, travels along a freeway corridor, it can be logged and time stamped at time t_1 by a sensor at location 1. After crossing a certain distance, it can be again detected and time stamped at time t_2 by means of a sensor at downstream location 2. Wi-Fi signal sensors can also discover the signals produced from Wi-Fi modules installed into different mobile devices. The Wi-Fi modules are defined based on IEEE 802.11 standard (IEEE 802 Group, 2010). The basic unit for data exchanged between devices is known as frame. Several different types of frames are defined in the standard protocol, namely Beacon, Acknowledgment (ACK), Data, and Probe. Each access point periodically sends Beacon frames to show its availability. If a mobile device is connected to an access point, information is transferred using Data or ACK frames; otherwise it sends Probe frames to look for existing access points. The information that can be detected using Wi-Fi signal sensors are: Media Access Control (MAC) addresses of the mobile device and the access point, frame type, time stamp, and signal strength correlated directly with the distance between Wi-Fi sensor and mobile device. Accordingly, by detecting the MAC address at multiple locations over time, a person will be tracked unless he/she turns off Wi-Fi or switches to airplane mode (Ji et al., 2017).

In summary, embedding sensors into a complex transport network can considerably increase the socio-economic benefits by providing sufficiently detailed traffic state information to users and operators. Moreover, a broad range of data mining and prediction methods can be employed for analyzing raw traffic data collected using sensor technologies.

The present article provides a review on the applications of traffic sensors, ranging from traditional ones to mobile sensors in smartphones, to surveillance of transportation networks, including travel time estimation (Travel Time Estimation Section), origin–destination matrices estimation (Origin–Destination Matrices Estimation Section), safety analysis (Safety Analysis Section), driver behavior monitoring (Driver Behavior Monitoring Section), and transportation mode inference (Transportation Mode Inference Section).

Travel Time Estimation

One of the main factors in the evaluation of a transport network is the congestion level measured in terms of delays and travel times. Real-time analysis of travel time is of great importance for traffic management and for providing information to commuters. Hence, the problem of locating traffic sensors on a transit network for travel time estimation is becoming a common trend in transportation research. Travel time estimation is practically performed from data collected on vehicle flows, speeds, queues, densities, and so on. The required data can be collected either manually (questionnaires, surveys, census, or even standing at street corners and counting vehicles), or through installing sensors (hardware and software) (Gentili and Mirchandani, 2018). The manual methods can lead to issues such as high respondent burden, small sampling rates, high data collection costs, short survey duration, and low data quality.

Chen et al. (2004) investigated the AVI reader location problem for both travel time and OD prediction applications. Three different location section criteria were taken into account when deciding locations for AVI readers: the minimization of the number of AVI readers, the maximization of the coverage of OD pairs, and the maximization of the number of AVI readings. To satisfy all three objectives, the problem was formulated as a multiobjective integer-optimization problem. Sherali et al. (2006) also developed a discrete optimization model for optimally locating AVI readers to estimate roadway travel times. A solution method based on the Reformulation–Linearization Technique combined with semi-definite programming concepts was proposed to solve the problem. Ban et al. (2009) proposed an optimization framework and a polynomial solution method for finding the optimal locations of point detectors employed to estimate freeway travel times. The objective of the model was to minimize the deviation of the actual and estimated travel times on link segments along a freeway route. Antoniou et al. (2013) developed a methodology for the identification and short-term prediction of traffic state, taking advantage of the ever-increasing accessibility of traffic data collected from emerging sensor technologies. For reliable travel time estimation, Park and Haghani (2015) developed a deterministic, a stochastic, and a dynamic formulation to find the optimal location of Bluetooth sensors, taking reorganization of sensors in a freeway corridor into account as well. Indeed, this method allows for the inclusion of traffic uncertainties into the problem, thus investigating scenarios which are more likely to occur.

In the vast majority of the previous studies, the active traffic sensors are totally installed at nodes of a traffic network and not at links; hence, the adjacent links of those nodes should be equipped with sensors, leading to a high cost. Zhu et al. (2018) developed a novel data-driven link-based network sensor location approach with the aim of maximizing travel time information gain. The underlying assumption is that the active sensors can be installed at links rather than at nodes. Moreover, the influence of route differentiation, as well as the uncertainty of travel time data are taken into account to robustly locate active sensors.

Mobile sensors in smartphones enable collection of high-resolution vehicle performance data in a cost-effective way. In recent years, there have been a growing number of publications focusing on the use of cellular phones with GPS data as the most common source of real time information to estimate travel times. For example, in the work of Amirian et al. (2016), the authors used several machine learning methods to improve the estimation of travel time for pedestrian mode. The dataset contains several different features, such as age, length of journey, change of elevation, day of week, time of day, weather condition, and gender. For instance, the length of journey for current day attribute was recorded using digital pedometer sensors on smartphones. The results showed that a random forest method can lead to a better performance compared to other methods tested, including gradient boosting, least squares, least-angle regression (LARS), least absolute shrinkage and selection operator (LASSO), K nearest neighbors, and Adaboost methods.

Origin–Destination Matrices Estimation

So far, a large number of studies have been carried out on the estimation and prediction of OD flow using data collected by traffic sensors in a cost-effective manner rather than the estimation of network origin–destination matrices based on traffic flow observations (Hadavi and Shafahi, 2016). The quality of OD demand estimation can be directly affected by the amount and quality of traffic data collected using different types of traffic sensors. Hence, finding the optimal number and installation locations of traffic surveillance sensors in view of the existing budget constraints of transit agencies, which is known as the network sensor location problem (NSLP), is of great importance for the estimation of a set of unknown network OD matrices at a high level of accuracy (Bianco et al., 2001).

The NSLP can be classified into two sub problems (Gentili and Mirchandani, 2012): the sensor location flow-observability problem (e.g., traffic flow inference), and the sensor location flow-estimation problem (e.g., OD estimation problem, travel time estimation). Three methods have been employed to solve the NSLP in the literature (Hu and Liou, 2014):

1. The mathematical optimization method with the aim of maximizing flow covers or minimizing specific requirements under a set of constraints, for example, budget constraints.
2. The algebraic method, in which one-step or iterative actions including matrix algebra based on specific incidence matrices, for example, the link–path incidence or the link–node incidence matrices are used to find a set of deployed links.

3. Graph theory, in which the spatial correlations between nodes and links within a network structure are disintegrated to find a set of critical links.

The space-state formulations based on Kalman Filtering are known as an efficient method for the estimation of time-dependent OD flows (Antoniou et al., 2007). Accordingly, Barceló Bugeda et al. (2012) developed a novel recursive linear Kalman-Filter approach for the estimation of time-dependent OD matrices from the measurements of traffic variables, taking the advantages of travel times and traffic counts collected by tracking Bluetooth equipped vehicles and traditional detection technologies. Hu et al. (2001) also developed a Kalman Filtering approach for the estimation of dynamic OD matrices, in which time-varying parameters as state variables were incorporated into the model formulation. Indeed, the Kalman-Filter method presented is able to cope with the nonlinear relationship existing between the state variables and the observations.

Antoniou et al. (2004) developed a methodology for the incorporation of automated vehicle identification (AVI) data into O-D estimation. Indeed, O-D estimation and prediction was carried out for both cases: the base case (no AVI information) and in the presence of AVI information. The results showed that the inclusion of additional data sources (e.g., information obtained from AVI systems) can improve the accuracy of the O-D estimation. Zhou and Mahmassani (2006) developed a novel OD-demand-estimation method in order to extract OD-demand distribution information from AVI counts without the inclusion of market-penetration rates and identification rates of AVI tags in the demand-estimation problem. A nonlinear ordinary least-squares estimation model was developed to amalgamate AVI counts with other available data sources into a multi-objective optimization framework. Ji et al. (2017) proposed a hierarchical Bayesian model to predict trip-level OD flow matrices and a period-level OD flow matrix using sampled OD flow data collected by the combination of farebox and Wi-Fi signal sensors data on a public transportation system. The model is also evaluated empirically on a real bus corridor, demonstrating the possibility of merging farebox and Wi-Fi signal sensors data to improve data quality and enlarge dataset that enable transit operators to attain sufficiently detailed transit route-level passenger travel patterns, and to derive latent aspects of passengers' travel behavior as well.

While a considerable amount of literature has been published on the NSLP, there is a notable paucity of well-controlled studies that seek to identify an optimal sensor configuration minimizing performance monitoring errors while taking the probability of sensor failures into account. To fill this gap, Danczyk et al. (2016) developed a probabilistic optimization model for the sensor allocation problem along freeway corridors, which minimizes the traffic measurement errors while taking all the plausible failure scenarios into consideration.

Safety Analysis

Traffic network surveillance using sensor technologies is also of great importance in real-time safety analysis, which can be applied to enhance incident detection capabilities and improve traffic control systems by exploring the relationship between real-time traffic features and crash occurrence (Yuan et al., 2018).

Petraki et al. (2020) used high-resolution driving behavior data collected from sensors of smartphones of 303 different drivers to investigate driver behavior at both road segment and junction levels. The dataset was also accompanied by traffic features (traffic flow and speed) collected via 26 loops installed over specific measuring positions on the two urban expressways under study. For the data analysis, two log-linear regression models were calibrated for road segments, as well as two linear regression models were developed for junctions. The results demonstrated that the number of harsh events can raise in road segments as the mean traffic flow per lane increases in the respective areas. Moreover, the increase of the average occupancy in junctions can lead to a rise in harsh accelerations, as well as when the average speed raises, more harsh deceleration events happen. The driving patterns along a route can be extracted from microscopic traffic measures (speed and acceleration profiles) obtained through the mobile sensor data. In the work of Paleti et al. (2017), mobile sensor data from smartphones was employed to derive latent driving patterns and to explore their relationship with traffic crash incidences. The findings showed that the microscopic traffic measures can significantly improve the data fit of crash frequency models.

The advent of several novel sources of information with an unprecedented volume, termed "big data", and state-of-the-art machine learning methods can support transportation researchers in finding appropriate solutions for the highway traffic safety system. Huang et al. (2020) have attempted to investigate the possibility of utilizing deep learning methods, which can automatically extract high-level features from input data, to detect crash occurrence and predict crash risk. The authors developed a deep learning algorithm on traffic data (volume, speed, and sensor occupancy) collected from roadside radar sensors for crash detection and risk estimation. The results indicated that a deep learning architecture can lead to a better crash detection performance and similar crash prediction performance compared to shallow models. Formosa et al. (2020) also proposed a deep learning methodology to leverage large imbalanced data for predicting real-time traffic conflicts. A large amount of disaggregated traffic data and on-board sensors data from an instrumented vehicle were collected over a section of the UK M1 motorway to construct the model. The authors took several different influencing factors into account in detecting traffic conflicts, such as speed variance, traffic density, speed, and weather conditions.

Driver Behavior Monitoring

With the rapid development of information and communication technologies (ICT), several different emerging data sources, such as GPS data, smartcard records data, roadside sensor data, and smartphone sensor-based data, have appeared and been applied to

supplement or substitute for traditional survey data to extend human travel behavior research. Of these, however, smartphone sensor-based data are the most widely applied and promising type, due to their growing penetration in the population, and their built-in location and motion sensors (Wesolowski et al., 2014). For example, smartphone-based vehicle positioning capability has resulted in collecting a massive data volume on the real-time location and dynamics of users, thereby providing a tremendous opportunity to monitor the operation of transit networks. Smartphone-based vehicle positioning methods have emerged since 2008 with the advent of GPS antennas and navigation maps in smartphones (Kanarachos et al., 2018).

In addition to GPS sensor, most of the modern smartphones are equipped with multiple motion sensors, such as accelerometer and magnetometer sensors, which can provide an unprecedented opportunity for the monitoring of the motion status of mobile phone users, thus improving the performance of transportation network surveillance. Indeed, these motion sensors enable transportation researchers to broaden their investigations in the area of travel behavior research by providing them with a considerable data volume collected on the real-time location and dynamics of users (Wang et al., 2018). The accelerometer sensors are used for estimating the acceleration and motion status of users in different directions, such as going up or down, standing, walking, cycling, and driving (Lane et al., 2010). The magnetometer sensors are used for detecting the orientation of a mobile phone device relative to the Earth's magnetic north. Overall, the recording of driving behavior in different conditions, either from smartphone sensors or from on-board device sensors has contributed to the collection of high-resolution data and objective measurements.

GPS positioning accuracy can be affected by a wide variety of factors, including the number of visible satellites, weather, and surrounding conditions such as buildings and trees (urban canyon effects) (Groves, 2011). Christopoulos et al. (2018) showed that the accuracy of the smartphone GPS position signal is on average 3 m, nevertheless, it can reduce significantly while crossing bridges and arriving at the urban area. Much of the available literature on smartphone-based vehicle positioning has focused on maintaining positioning accuracy when the GPS signal is weak (Chen et al., 2015). For instance, a possible way of enhancing the level of GPS position accuracy is to share the GPS information between smartphone of road travelers and other infrastructure objects of recognized GPS location. Nonetheless, the above-mentioned idea has been merely investigated in a simulation framework or using Dedicated Short-Range Communications (DSRC) vehicle-to-vehicle (V2V) communications (Bento et al., 2017).

The growing penetration of advanced smartphones equipped with a wide variety of built-in sensors (e.g., GPS, accelerometer, and gyroscope) and network interfaces allow for a more efficient development of connected applications used to monitor and recognize driving behavior in real time. Johnson and Trivedi (2011) developed a state-of-the-art system in which a Dynamic Time Warping (DTW) algorithm and smartphone based sensor-fusion (accelerometer, magnetometer, gyroscope, GPS, video) are used to detect driving events and driving style. The authors assessed the performance of multiple sensor fusion sets to detect lateral and longitudinal movements. Wahlström et al. (2015) developed a novel framework for detecting dangerous vehicle cornering events based on an updated test statistic estimated by an unscented Kalman filter applied to the global navigation satellite system (GNSS) measurements. They also evaluated the possibility of using inertial measurement unit (IMU) measurements in order to enhance the detection accuracy. Li et al. (2016) proposed a Safe Drive system which can determine the driving conditions with leveraging the built-in smartphone sensors. The authors proposed a precise driving condition classification algorithm-based on data collected from GPS and accelerometer sensors on smartphones. The findings showed that the algorithm can reach 87% classification accuracy in complex circumstances.

Castignani et al. (2017) developed an adaptive driving maneuver detection approach based on a Multivariate Normal distribution (MVN) model to detect risky driving maneuvers using smartphone sensors and GPS data. The results showed that the model can adapt to various road circumstances and vehicles. Saiprasert et al. (2017) proposed three various algorithms to detect and classify driving events using data collected from motion sensors of smartphones. The algorithms can also classify whether these driving events are aggressive. The experimental results indicated that a pattern-matching algorithm can outperform a rule-based algorithm for driving event detection in lateral and longitudinal axes. Recently, Johnson and Rajamani (2020) proposed a state-of-the-art algorithm to determine the position of a smartphone on a moving vehicle using the accelerometer, gyroscope and GPS sensors available on a phone. This detection can be used to recognize whether the smartphone is being used by a driver or a passenger on the vehicle. Hence, it can be useful for automatically restricting texting features when the phone is on the driver's seat.

Transportation Mode Inference

As discussed earlier, smartphones and their sensors are spreading more rapidly than expected as devices that can be used for monitoring driver behavior with leveraging their GPS trajectories data. In recent years, smartphone-based transportation mode inference has attracted considerable attention, both scholarly and popular. Kanarachos et al. (2018) carried out an extensive review of various sensor fusion approaches that are used for smartphone-based travel modes classification (see Table 1). The comparison has been carried out based on a Cybernetics paradigm, in which the signals, decision-making technique, sensor fusion level, and performance are taken into account.

Up to now, a large number of machine learning methods have been tested for the travel modes classification. As a comprehensive investigation, Xiao et al. (2015) demonstrated that Bayesian Networks can outperform Support Vector Machine (SVM), as well as SVM-based approaches are superior to Neural Networks. For example, they showed that Bayesian Networks can, respectively, lead to an improvement of 2.5% and 7% on the training and test sets compared to SVMs.

Table 1 Comparison of various sensor fusion approaches for smartphone-based classification of transportation mode

Reference	Signals			Method	Fusion	Accuracy
	Acceleration	Speed	GPS position			
Byon et al. (2009)	✓		✓	Neural Networks	Smartphone based fusion	60%–98%
Biljecki et al. (2013)		✓		Fuzzy expert	Smartphone based fusion	92%
Byon and Liang (2014)	✓		✓	Neural Networks	Smartphone based fusion	74%–83%
Xiao et al. (2015)	✓		✓	Bayesian Network	Smartphone based fusion	93%–95%
Assemi et al. (2016)	✓		✓	Multinomial Logistic Regression Model	Smartphone based fusion	95%
Martin et al. (2017)	✓		✓	Movelets, k-nearest neighbors, feature extraction	Crowd sensing-based fusion	89%
Martin et al. (2017)	✓		✓	Movelets, random forests, feature extraction	Crowd sensing-based fusion	97%

Source: Adapted from Kanarachos et al. (2018).

Summary

Sensors, which are a key part of any ITS, play a crucial role in collecting real-time travel data used to monitor traffic congestion levels, incidents on freeways, predict travel times, estimate OD flows, and support decision making for transport agencies. So far, a broad spectrum of sensor technologies has been evolved and used in the area of transportation systems, ranging from infrastructure-based sensor technologies to built-in smartphone sensors. The present chapter provided a wide-ranging review on the applications of various sensor technologies in travel time estimation (Travel Time Estimation Section), origin-destination matrices estimation (Origin-destination Matrices Estimation Section), safety analysis (Safety Analysis Section), driver behavior monitoring (Driver Behavior Monitoring Section), and transportation mode inference (Transportation Mode Inference Section). Thus far, researchers have mostly developed travel demand models based on raw traffic data collected through traditional sensor capabilities, which are expensive and can lead to issues such as short survey duration, and small sampling rates. On the other hand, emerging sensor technologies, such as smartphone-based sensors, can support transportation researchers to leverage state-of-the-art machine learning methods with a tremendous amount of mobility data for discovering daily travel behavior. Hence, this article has particularly focused on the review of the previous studies in which modern sensor technologies, such as Bluetooth, Wi-Fi signal, and smartphone-based sensors, are used for the monitoring of the dynamics of users.

Acknowledgments

This research has been supported through the TUM International Graduate School of Science and Engineering-IGSSE (MO3 project).

References

- Amirian, P., Basiri, A., Morley, J., 2016. Predictive analytics for enhancing travel time estimation in navigation apps of Apple, Google, and Microsoft, in: Proceedings of the 9th ACM SIGSPATIAL International Workshop on Computational Transportation Science. pp. 31–36.
- Antoniou, C., Balakrishna, R., Koutsopoulos, H.N., 2011. A synthesis of emerging data collection technologies and their impact on traffic management applications. *Eur. Transp. Res. Rev.* 3, 139–148.
- Antoniou, C., Ben-Akiva, M., Koutsopoulos, H.N., 2007. Nonlinear Kalman filtering algorithms for on-line calibration of dynamic traffic assignment models. *IEEE Trans. Intell. Transp. Syst.* 8, 661–670.
- Antoniou, C., Ben-Akiva, M., Koutsopoulos, H.N., 2004. Incorporating automated vehicle identification data into origin-destination estimation. *Transp. Res. Rec.* 1882, 37–44.
- Antoniou, C., Koutsopoulos, H.N., Yannis, G., 2013. Dynamic data-driven local traffic state estimation and prediction. *Transp. Res. Part C Emerg. Technol.* 34, 89–107.
- Assemi, B., Safi, H., Mesbah, M., Ferreira, L., 2016. Developing and validating a statistical model for travel mode identification on smartphones. *IEEE Trans. Intell. Transp. Syst.* 17, 1920–1931.
- Ban, X.J., Herring, R., Margulici, J.D., Bayen, A.M., 2009. Optimal sensor placement for freeway travel time estimation. In: *Transportation and Traffic Theory 2009: Golden Jubilee*. Springer, pp. 697–721.
- Barceló Buggedá, J., Montero Mercadé, L., Bullejos, M., Serch, O., Carmona Bautista, C., 2012. A kalman filter approach for the estimation of time dependent OD matrices exploiting bluetooth traffic data collection, in: *TRB 91st Annual Meeting Compendium of Papers DVD*. pp. 1–16.
- Bento, L.C., Bonnifait, P., Nunes, U.J., 2017. Cooperative GNSS positioning aided by road-features measurements. *Transp. Res. Part C Emerg. Technol.* 79, 42–57.
- Bianco, L., Confessore, G., Reverberi, P., 2001. A network based model for traffic sensor location with implications on O/D matrix estimates. *Transp. Sci.* 35, 50–60.
- Biljecki, F., Ledoux, H., Van Oosterom, P., 2013. Transportation mode-based segmentation and classification of movement trajectories. *Int. J. Geogr. Inf. Sci.* 27, 385–407.
- Bottero, M., Dalla Chiara, B., Defflorio, F.P., 2013. Wireless sensor networks for traffic monitoring in a logistic centre. *Transp. Res. Part C Emerg. Technol.* 26, 99–124.
- Byon, Y.-J., Abdulhai, B., Shalaby, A., 2009. Real-time transportation mode detection via tracking global positioning system mobile devices. *J. Intell. Transp. Syst.* 13, 161–170.

- Byon, Y.-J., Liang, S., 2014. Real-time transportation mode detection using smartphones and artificial neural networks: performance comparisons between smartphones and conventional global positioning system sensors. *J. Intell. Transp. Syst.* 18, 264–272.
- Castignani, G., Dermann, T., Frank, R., Engel, T., 2017. Smartphone-based adaptive driving maneuver detection: a large-scale evaluation study. *IEEE Trans. Intell. Transp. Syst.* 18, 2330–2339.
- Chen, A., Chootinan, P., Pravinongvuth, S., 2004. Multiobjective model for locating automatic vehicle identification readers. *Transp. Res. Rec.* 1886, 49–58.
- Chen, D., Cho, K.-T., Han, S., Jin, Z., Shin, K.G., 2015. Invisible sensing of vehicle steering with smartphones, in: *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. pp. 1–13.
- Christopoulos, S.-R.G., Kanarachos, S., Chronos, A., 2018. Learning driver braking behavior using smartphones, neural networks and the sliding correlation coefficient: road anomaly case study. *IEEE Trans. Intell. Transp. Syst.* 20, 65–74.
- Danczyk, A., Di, X., Liu, H.X., 2016. A probabilistic optimization model for allocating freeway sensors. *Transp. Res. Part C Emerg. Technol.* 67, 378–398.
- Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., Yuan, J., 2020. Predicting real-time traffic conflicts using deep learning. *Accid. Anal. Prev.* 136, 105429.
- Gentili, M., Mirchandani, P.B., 2018. Review of optimal sensor location models for travel time estimation. *Transp. Res. Part C Emerg. Technol.* 90, 74–96.
- Gentili, M., Mirchandani, P.B., 2012. Locating sensors on traffic networks: Models, challenges and research opportunities. *Transp. Res. part C Emerg. Technol.* 24, 227–255.
- Group, I. 802. 1. W., 2010. IEEE standard for information technology–Telecommunications and information exchange between systems–Local and metropolitan area networks–Specific requirements–Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 6: Wireless Access in Vehicular Environments. *IEEE Std* 802.
- Groves, P.D., 2011. Shadow matching: A new GNSS positioning technique for urban canyons. *J. Navig.* 64, 417–430.
- Hadavi, M., Shafahi, Y., 2016. Vehicle identification sensor models for origin–destination estimation. *Transp. Res. Part B Methodol.* 89, 82–106.
- Hu, S.-R., Liou, H.-T., 2014. A generalized sensor location model for the estimation of network origin–destination matrices. *Transp. Res. Part C Emerg. Technol.* 40, 93–110.
- Hu, S.-R., Madanat, S.M., Krogmeier, J.V., Peeta, S., 2001. Estimation of dynamic assignment matrices and OD demands using adaptive Kalman filtering. *J. Intell. Transp. Syst.* 6, 281–300.
- Huang, T., Wang, S., Sharma, A., 2020. Highway crash detection and risk estimation using deep learning. *Accid. Anal. Prev.* 135, 105392.
- Ji, Y., Zhao, J., Zhang, Z., Du, Y., 2017. Estimating bus loads and OD flows using location-stamped farebox and Wi-Fi signal data. *J. Adv. Trans* 2017.
- Johnson, D.A., Trivedi, M.M., 2011. Driving style recognition using a smartphone as a sensor platform, in: *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 1609–1615.
- Johnson, G., Rajamani, R., 2020. Smartphone localization inside a moving car for prevention of distracted driving. *Veh. Syst. Dyn.* 58, 290–306.
- Kanarachos, S., Christopoulos, S.-R.G., Chronos, A., 2018. Smartphones as an integrated platform for monitoring driver behaviour: the role of sensor fusion and connectivity. *Transp. Res. part C Emerg. Technol.* 95, 867–882.
- Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T., 2010. A survey of mobile phone sensing. *IEEE Commun. Mag.* 48, 140–150.
- Li, Y., Zhou, G., Li, Y., Shen, D., 2016. Determining driver phone use leveraging smartphone sensors. *Multimed. Tools Appl.* 75, 16959–16981.
- Martin, B.D., Addona, V., Wolfson, J., Adomavicius, G., Fan, Y., 2017. Methods for real-time prediction of the mode of travel using smartphone-based GPS and accelerometer data. *Sensors* 17, 2058.
- Paleti, R., Sahin, O., Cetin, M., 2017. Modeling the impact of latent driving patterns on traffic safety using mobile sensor data. *Accid. Anal. Prev.* 107, 92–101.
- Park, H., Haghani, A., 2015. Optimal number and location of Bluetooth sensors considering stochastic travel time prediction. *Transp. Res. Part C Emerg. Technol.* 55, 203–216.
- Petraki, V., Ziakopoulos, A., Yannis, G., 2020. Combined impact of road and traffic characteristic on driver behavior using smartphone sensor data. *Accid. Anal. Prev.* 144, 105657.
- Saiprasert, C., Pholprasit, T., Thajchayapong, S., 2017. Detection of driving events using sensory data on smartphone. *Int. J. Intell. Transp. Syst. Res.* 15, 17–28.
- Sherali, H.D., Desai, J., Rakha, H., 2006. A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times. *Transp. Res. Part B Methodol.* 40, 857–871.
- Sun, Z., Jin, W.-L., Ng, M., 2016. Network sensor health problem. *Transp. Res. Part C Emerg. Technol.* 68, 300–310.
- Wahlström, J., Skog, I., Händel, P., 2015. Detection of dangerous cornering in GNSS-data-driven insurance telematics. *IEEE Trans. Intell. Transp. Syst.* 16, 3073–3083.
- Wang, Z., He, S.Y., Leung, Y., 2018. Applying mobile phone data to travel behaviour research: a literature review. *Travel Behav. Soc.* 11, 141–155.
- Wesolowski, A., Stresman, G., Eagle, N., Stevenson, J., Owaga, C., Marube, E., Bousema, T., Drakeley, C., Cox, J., Buckee, C.O., 2014. Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones. *Sci. Rep.* 4, 5678.
- Xiao, G., Juan, Z., Zhang, C., 2015. Travel mode detection based on GPS track data and Bayesian networks. *Comput. Environ. Urban Syst.* 54, 14–22.
- Yuan, J., Abdel-Aty, M., Wang, L., Lee, J., Yu, R., Wang, X., 2018. Utilizing bluetooth and adaptive signal control data for real-time safety analysis on urban arterials. *Transp. Res. part C Emerg. Technol.* 97, 114–127.
- Zhao, J., Xu, H., Liu, H., Wu, J., Zheng, Y., Wu, D., 2019. Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors. *Transp. Res. part C Emerg. Technol.* 100, 68–87.
- Zhou, X., Mahmassani, H.S., 2006. Dynamic origin-destination demand estimation using automatic vehicle identification data. *IEEE Trans. Intell. Transp. Syst.* 7, 105–114.
- Zhu, N., Fu, C., Ma, S., 2018. Data-driven distributionally robust optimization approach for reliable travel-time-information-gain-oriented traffic sensor location model. *Transp. Res. part B Methodol.* 113, 91–120.

Planning Active Travel and School Transport

Fahimeh Khalaj, Dorina Pojani, Sara Alidoust, The University of Queensland, Australia

© 2021 Elsevier Ltd. All rights reserved.

Children's and Youth's Active Travel as a Planning Concern	432
Active School Travel in Western Europe and the Anglosphere	432
School Travel in Other Settings	433
The Road Ahead: Planning for Active School Transport	433
References	434
Further Reading	435

Children's and Youth's Active Travel as a Planning Concern

Active and independent travel among children and youth became an area of planning inquiry in the 1970s. That era saw a shift from “outdoor” and “walking” children to “indoor” and “chauffeured” children. In response, two seminal works were published in the UK: Ward’s “The Child in the City” (1978) and Hillman’s et al. “One False Move” (1990). These authors advocated for transport safety and security measures on behalf of children and youth. Their concerns stemmed from the post-war suburbanization movement, which led to increasing distances and intense car use in middle-class peripheries. Meanwhile, urban cores became depleted of people and resources and consequently, safety and security therein declined. Attitudes changed too: the outside world came to be seen as a hostile, dangerous place where children were likely to be harmed by cars or adults. In combination, these trends limited both urban and suburban children’s free mobility and access to the city.

In the last decades, with growing concerns over obesity, climate, and liveability, studies on the travel of children and youth travel have multiplied, especially in Western Europe and North America (see Hillman, 1999; Mackett, 2013; McDonald, 2005; O’Brien et al., 2000). These studies have revealed that, compared to previous generations, the travel patterns of children here have changed enormously. The key findings from studies set in Western Europe, the Anglosphere, and elsewhere are unpacked below.

Active School Travel in Western Europe and the Anglosphere

Many western-based studies, most of them cross-sectional, have focused on the school commute of children and youth. Studies are so numerous that, by now several systematic reviews are available (Aranda-Balboa et al., 2020; Costa et al., 2020; D’Haese et al., 2015; Davison et al., 2008; Faulkner et al., 2009; Ikeda et al., 2018a, 2018b; Larouche et al., 2014b; Lubans et al., 2011; Pont et al., 2009; Rothman et al., 2018; Sirard and Slater, 2008; Wong et al., 2011).

The reviews show that parents have adapted to modern urban living by confining children to the home or by escorting and/or chauffeuring them in cars. Car travel has become a form of mobile childcare, while allowing children out unchaperoned is now regarded as a marker of neglectful parenting. Children themselves often mirror their parents’ habits and views in regard to travel (Zwerts et al., 2010). There is a strong positive association between active school travel and other physical activities (and cardiovascular fitness) but findings are less unanimous on the assumed link between active school travel and lower body mass index (BMI) (Faulkner et al., 2009; Sirard and Slater, 2008).

Based on these reviews, the rates of active travel to school are in the following ranges: 20%–78% in continental Europe; 47%–68% in the UK; 5%–49% in the US; 19%–33% in Australia, and 15%–61% in New Zealand. Some reviews separate statistics for walking and cycling, whereas others lump them together under the banner of “active travel.” Where the data are reported separately, staggering differences in walking and cycling rates are revealed. For example, in the UK a solid 40%–69% of school-children walk to school whereas cycling rates are abysmal by comparison, ranging between 0.2% and 9%. In the US, between 4% and 63% of children walk to school depending on the setting but only 0.8%–14% cycle.

The main factors consistently found to associate with active school travel in Western European settings and the Anglosphere are (not necessarily in order of importance): short home-school distances; lower socio-economic background; urban character of residential neighborhood; densely populated school neighborhood; land-use mix along route; conducive caretakers’ schedules; supportive parental attitudes; lower parental safety and/or security concerns; small school size; male gender of child; availability of pedestrian infrastructure and absence of busy road crossings or freeways en route; fewer or no cars in the household; and school zoning policies (e.g., forcing children to attend a school within their catchment area). Less common correlates include: older age of child; minority ethnic background; social support; peer influence; positive relationships with neighbors; street design (intersection density, pedestrian route directness, length and density of street segments, street aesthetics, street topography); lower parental education level; parental employment and marital status; and children’s positive attitudes toward active school travel.

The foregoing findings are subject to the caveat that, comparing studies on the school commute is rather difficult due to the heterogeneity of sample sizes, family composition, school systems, spatial units, cultural traits, and environmental features, all of which have been examined as potential predictors of active school travel (Mitra and Buliung, 2012). Even the definition of “child” and “youth” varies among studies—in terms of age range, with 5 or 6 being the lower limit and 18 being the upper limit. Some studies separate the “to” and “from” school trips because in western settings the return trip is complicated by afterschool activities and caregiver schedules (given that many children are chauffeured)

School Travel in Other Settings

Studies in settings other than Western Europe and the Anglosphere are much less extensive. Case studies are only available for: Albania (Pojani and Boussauw, 2014); Brazil (Becker et al., 2017; Costa et al., 2012; Dias et al., 2019; Melo et al., 2013); mainland China (Yang et al., 2017, 2019; Zhang et al., 2017); Colombia (Arango et al., 2011); Cyprus (Loucaides et al., 2010); India (Singh and Vasudevan, 2018); Iran (Mehdizadeh et al., 2017; Samimi and Ermagun, 2013; Shokoohi et al., 2012a, 2012b); Japan (Mori et al., 2012; Waygood and Taniguchi, 2020); Ghana (Siiba, 2020); Hong Kong (Barnett et al., 2019; Leung and Loo, 2020; Yang et al., 2020); Mexico (Jáuregui et al., 2015, 2016); the Philippines (Tudor-Locke et al., 2003), Taiwan (Lin and Chang, 2010), and Vietnam (Trang et al., 2012). Two systematic reviews are available, one for Africa (Larouche et al., 2014a) and one for Brazil (Ferrari et al., 2018). Hence generalizations are clearly difficult.

Based on these studies and reviews, the rates of active travel to school range from 20%–98% in Asia to 19%–79% in South-eastern Europe to 41%–71% in Latin America to 20%–100% in Africa. Again, few reports provide separate statistics for walking and cycling. Where separate figures are provided, cycling rates are shown to be much lower than walking rates. For example, a study based in China reports a 57% walking rate and a 12% cycling rate. Similarly, a study conducted in Albania reports a 67% walking rate and a 1% cycling rate.

With regard to determinants, the most consistent findings are not too different from those unveiled by the western-based studies discussed above, and include: short home–school distances; lower socio-economic background; protective attitude toward girls; urban character of school neighborhood; supportive built environment features (e.g., availability of sidewalks and parks); street connectivity and safety; parental safety perceptions; children’s age; parental travel patterns (level of auto-dependency); fewer or no cars in the household; less maternal education; and nonworking mothers.

Less commonly mentioned factors include: high tree-shade density along sidewalks; male gender of child; maternal driving licence status; parental physical activity habits; children’s attitude toward active travel; peer support; street aesthetics and topography; street size; higher population density along school route; mixed land-use pattern along school route; milder weather conditions; higher number of adults in the household; and availability/accessibility of public transport stops along school routes. Some factors emerge only in particular contexts. For example, in studies based in Africa, factors such as fear of rape, drawing, or attacks by wild animals deterred children, especially girls, from engaging in active travel to school.

The Road Ahead: Planning for Active School Transport

Children face many barriers to active school travel, particularly in wealthier and more car-oriented settings. Parental concerns around active school travel are mostly justified: road injuries are a reality and the quality of public outdoor spaces, especially in the poorer sections of big cities, has declined. Air, noise, and visual pollution caused by moving and parked cars, as well as poor urban design, with spaces unfriendly to walking, are widespread outside (historic) city centers. These concerns notwithstanding, automobility starting at an early age poses major risks from an environmental and public health perspective, as children’s current travel behavior may affect their future travel behavior as adults. Also, reduced physical activity and lack of exposure to the outdoor environment may affect children’s physical and mental health.

Hence, active school transport should be considered as a priority concern in public policy. Socio-economic and demographic variables that correlate with active travel (lower socio-economic background of child; male gender of child; older age of child; conducive caretakers’ schedules; and fewer or no cars in the household) may be difficult to modify. However, environmental and psychological variables (short home–school distances; urban character of residential and school neighborhood; supportive built environment features such as availability of sidewalks, parks, direct routes); lower parental safety and/or security concerns; and supportive parental attitudes and less auto-dependency in the family) can and should be tackled to enable and encourage more children and youth to walk and cycle to school. Following is a list of the most popular programs in this regard, which have been adopted around the world. Most have been reviewed by Duggan et al. (2018).

Walking School Bus. In a Walking School Bus, participating parents or other adults take turns acting as “drivers”: they set a walking route along which they pick up children and deliver them safely to school (Collins and Kearns, 2010; Kingham and Ussher 2005, 2007; Pérez-Martín et al., 2018). The program is thought to have originated in Australia in 1992 but there is evidence to suggest that similar practices had been in place in Japan since at least the early 1960s. In the new millennium, the Walking School Bus has expanded to Canada, Denmark, New Zealand, the United Kingdom, and the United States. Despite its benefits, this program tends to suffer from a shortage of adult volunteers sharing the workload. Therefore, it is concentrated in wealthier neighborhoods (Collins and Kearns, 2010; Kingham and Ussher, 2005, 2007).

Safe Routes to School. This program was first initiated in the United States in 2005 with the aim of boosting traffic safety along the school routes and in school environs (Duggan et al., 2018; McDonald and Aalborg, 2009). Strategies have included footpath improvements, traffic calming features, speed reduction policies, and provision of safe bicycle parking facilities at schools. Public awareness and traffic education campaigns targeting both parents and pupils are also part of the program. Despite these efforts and investments, the program has been only moderately successful in altering children's travel-to-school behavior and shifting children from cars to bicycles or walking (McDonald and Aalborg, 2009).

Active & Safe Routes to School. This is the Canadian version of the "Safe Routes to School" program. It was introduced in three Toronto schools in the mid-1990s, and later spread to more than 1000 schools in all Canadian provinces (Duggan et al., 2018). In 2006, the program was supplemented by "School Travel Planning", which expanded to more than 120 schools across Canada by 2012.

Active Travel Program. This program commenced in Northern Ireland in 2013. It largely targets primary schools, although secondary schools are also encouraged to participate. In participating schools, an "Active Travel Officer" is devoted to organizing a range of curricular lessons as well as extracurricular activities to teach children the benefits and risks of active commuting. The program has been successful in increasing the rate of active school commutes by 15% over a 3-year period. It has also led to a decrease in traffic congestion in the vicinity of participating schools (Duggan et al., 2018).

TravelWise School. This program commenced in New Zealand in 2002, and by 2014, more than 20,000 schoolchildren were involved within the Auckland region alone. The basic aim was to encourage school communities (including teachers, administrators, children, and parents) to take the lead in identifying local road safety problems and devise a complete plan for reducing traffic accidents and congestion (Morton, 2005). Thanks to TravelWise School, schoolchildren's car crashes rates have been successfully reduced and 12,000 car trips have been eliminated. Integrating "safe cycling and walking" lessons in the school curricula and securing the commitment of local police have been key ingredients to the program's success (Duggan et al., 2018).

School Travel Plans & ModeShift. Since 2003 the United Kingdom has been mandating all schools to develop "School Travel Plans." Funding to local authorities has been provided for this purpose, which has allowed schools to dedicate a special officer to the task of developing and promoting active travel plans. Overall, School Travel Plans have succeeded in reducing car use for the school commute, especially where combined with ModeShift. "ModeShift" is a measure to monitor school travel plans and their outcomes, which was initiated in 2007. As part of "ModeShift," an online "Sustainable Travel Accreditation and Recognition Scheme" was introduced, in which school travel officers could report progress, achievements, and failures (Duggan et al., 2018).

Your Move. This program was first established in 1995 in Perth, Australia, as "TravelSmart." The program ran for over twenty years, before the name was changed to "Your Move" in 2015 to align with other local initiatives. By 2013, 130 schools in Western Australia had been included in the program. Your Move operates on multiple fronts: it provides information, tips, and lesson plans via a website; it organizes campaigns and educational programs; it offers incentives and rewards to individual participants; and it provides infrastructure improvement grants to local councils. Your Move has enabled travel behaviour change among schoolchildren and in participating schools, car trips have dropped and active travel has increased (Duggan et al., 2018).

In 2020, the Covid-19 pandemic has had implications for active school travel. There is concern and some anecdotal evidence that active travel habits formed earlier may have been reversed or weakened. Unfortunately, due to fears around virus transmission in open spaces, children may be traveling by car more than before.

References

- Aranda-Balboa, M., Huertas-Delgado, F., Herrador-Colmenero, M., Cardon, G., Chillón, P., 2020. Parental barriers to active transport to school: a systematic review. *Int. J. Public Health* 65, 87–98.
- Arango, C.M., Parra, D.C., Eyler, A., Sarmiento, O., Mantilla, S.C., Gomez, L.F., Lobelo, F., 2011. Walking or bicycling to school and weight status among adolescents from Montería, Colombia. *J. Phys. Activity Health* 8, S171–S177.
- Barnett, A., Sit, C.H., Mellecker, R.R., Cerin, E., 2019. Associations of socio-demographic, perceived environmental, social and psychological factors with active travel in Hong Kong adolescents: the iHealt (H) cross-sectional study. *J. Transport Health* 12, 336–348.
- Becker, L., Fermiro, R., Lima, A., Rech, C., Añez, C., Reis, R., 2017. Perceived barriers for active commuting to school among adolescents from Curitiba, Brazil. *Revista Brasileira de Atividade Física & Saúde* 22, 24–34.
- Collins, D., Kearns, R.A., 2010. Walking school buses in the Auckland region: a longitudinal assessment. *Transport Policy* 17, 1–8.
- Costa, F.F., Silva, K.S., Schmoelz, C.P., Campos, V.C., De Assis, M.A.A., 2012. Longitudinal and cross-sectional changes in active commuting to school among Brazilian schoolchildren. *Prevent. Med.* 55, 212–214.
- Costa, J., Adamakis, M., O'Brien, W., Martins, J., 2020. A scoping review of children and adolescents' active travel in Ireland. *Int. J. Environ. Res. Public Health* 17, 2016.
- D'Haese, S., Vanwolleghem, G., Hinckson, E., De Bourdeaudhuij, I., Deforche, B., Van Dyck, D., Cardon, G., 2015. Cross-continental comparison of the association between the physical environment and active transportation in children: a systematic review. *Int. J. Behav. Nutr. Phys. Activity* 12, 145.
- Davison, K.K., Werder, J.L., Lawson, C.T., 2008. Peer reviewed: Children's active commuting to school: current knowledge and future directions. *Prevent. Chronic Disease* 5.
- Dias, A.F., Gaya, A.R., Pizarro, A.N., Brand, C., Mendes, T.M., Mota, J., Santos, M.P., Gaya, A.C.A., 2019. Perceived and objective measures of neighborhood environment: association with active commuting to school by socioeconomic status in Brazilian adolescents. *J. Transport Health* 14, 100612.
- Duggan, M., Fetherston, H., Harrison, B., Lindberg, R., Parisella, A., Shilton, T., Greenland, R., Hickman, D., 2018. Active school travel: pathways to a healthy future. Australian Health Policy Collaboration, Melbourne, Victoria.
- Faulkner, G.E., Bulling, R.N., Flora, P.K., Fusco, C., 2009. Active school transport, physical activity levels and body weight of children and youth: a systematic review. *Prevent. Med.* 48, 3–8.
- Ferrari, G.L.D.M., Victo, E.R.D., Ferrari, T.K., Solé, D., 2018. Active transportation to school for children and adolescents from Brazil: a systematic review. *Revista Brasileira de Cineantropometria & Desempenho Humano* 20, 406–414.
- Hillman, M., Adams, J., Whitelegg, J., 1990. *One False Move: A Study of Children's Independent Mobility*. London, Policy Studies Institute.

- Hillman, M., 1999. "The Impact of Transport Policy on Children's Development." Paper presented at the Canterbury Safe Routes to School Project Seminar, Canterbury Christ Church University College, 29 May.
- Ikedo, E., Hinckson, E., Witten, K., Smith, M., 2018a. Associations of children's active school travel with perceptions of the physical environment and characteristics of the social environment: a systematic review. *Health Place* 54, 118–131.
- Ikedo, E., Stewart, T., Garrett, N., Egli, V., Mandic, S., Hosking, J., Witten, K., Hawley, G., Tautolo, E.S., Rodda, J., 2018b. Built environment associates of active school travel in New Zealand children and youth: a systematic meta-analysis using individual participant data. *J. Transport Health* 9, 117–131.
- Jáuregui, A., Medina, C., Salvo, D., Barquera, S., Rivera-Dommarco, J.A., 2015. Active commuting to school in Mexican adolescents: evidence from the Mexican National Nutrition and Health Survey. *J. Phys. Activity Health* 12, 1088–1095.
- Jáuregui, A., Soltero, E., Santos-Luna, R., Hernández-Barrera, L., Barquera, S., Jáuregui, E., Lévesque, L., López-Taylor, J., Ortiz-Hernández, L., Lee, R., 2016. A multisite study of environmental correlates of active commuting to school in Mexican children. *J. Phys. Activity Health* 13, 325–332.
- Kingham, S., Ussher, S., 2005. Ticket to a sustainable future: an evaluation of the long-term durability of the Walking School Bus programme in Christchurch, New Zealand. *Transport Policy* 12, 314–323.
- Kingham, S., Ussher, S., 2007. An assessment of the benefits of the walking school bus in Christchurch, New Zealand. *Transport. Res. Part A: Policy Pract.* 41, 502–510.
- Larouche, R., Oyeyemi, A.L., Prista, A., Onyewera, V., Akinroye, K.K., Tremblay, M.S., 2014a. A systematic review of active transportation research in Africa and the psychometric properties of measurement tools for children and youth. *Int. J. Behav. Nutr. Phys. Activity* 11, 129.
- Larouche, R., Saunders, T.J., Faulkner, G.E.J., Colley, R., Tremblay, M., 2014b. Associations between active school transport and physical activity, body composition, and cardiovascular fitness: a systematic review of 68 studies. *J. Phys. Activity Health* 11, 206–227.
- Leung, K.Y., Loo, B.P., 2020. Determinants of children's active travel to school: a case study in Hong Kong. *Travel Behav. Soc.* 21, 79–89.
- Lin, J.-J., Chang, H.-T., 2010. Built environment effects on children's school travel in Taipei: independence and travel mode. *Urban Stud.* 47, 867–889.
- Loucaides, C.A., Jago, R., Theophanous, M., 2010. Prevalence and correlates of active traveling to school among adolescents in Cyprus. *Central Eu. J. Public Health* 18.
- Lubans, D.R., Boreham, C.A., Kelly, P., Foster, C.E., 2011. The relationship between active travel to school and health-related fitness in children and adolescents: a systematic review. *Int. J. Behav. Nutr. Phys. Activity* 8, 1–12.
- Mackett, R.L., 2013. Children's travel behaviour and its health implications. *Transport Policy* 26, 66–72.
- McDonald, N.C., 2005. Children's travel: patterns and influences. Doctor of Philosophy University of California at Berkeley.
- McDonald, N.C., Aalborg, A.E., 2009. Why parents drive children to school: implications for safe routes to school programs. *J. Am. Plan. Assoc.* 75, 331–342.
- Mehdizadeh, M., Mamdoohi, A., Nordfjaern, T., 2017. Walking time to school, children's active school travel and their related factors. *J. Transport Health* 6, 313–326.
- Melo, E.N., Barros, M., Reis, R.S., Hino, A.A.F., Santos, C.M., Farias Junior, J.C., 2013. Is the environment near school associated with active commuting to school among preschoolers? *Revista Brasileira de Cineantropometria & Desempenho Humano* 15, 393–404.
- Mitra, R., Buliung, R.N., 2012. Built environment correlates of active school transportation: neighborhood and the modifiable areal unit problem. *J. Transport Geogr.* 20, 51–61.
- Mori, N., Armada, F., Willcox, D.C., 2012. Walking to school in Japan and childhood obesity prevention: new lessons from an old policy. *Am. J. Public Health* 102, 2068–2073.
- Morton, T., 2015. Travelwise to school: delivering school travel plans in the New Zealand environment. *Australasian Transport Research Forum (ATRF)*, 28TH, 2005, Sydney, New South Wales, Australia.
- O'Brien, M., Jones, D., Sloan, D., Rustin, M., 2000. Children's independent spatial mobility in the urban public realm. *Childhood* 7, 257–277.
- Pérez-Martín, P., Pedrós, G., Martínez-Jiménez, P., Varo-Martínez, M., 2018. Evaluation of a walking school bus service as an intervention for a modal shift at a primary school in Spain. *Transport Policy* 64, 1–9.
- Pojani, D., Boussauw, K., 2014. Keep the children walking: active school travel in Tirana, Albania. *J. Transport Geogr.* 38, 55–65.
- Pont, K., Ziviani, J., Wadley, D., Bennett, S., Abbott, R., 2009. Environmental correlates of children's active transportation: a systematic literature review. *Health Place* 15, 849–862.
- Rothman, L., Macpherson, A.K., Ross, T., Buliung, R.N., 2018. The decline in active school transportation (AST): A systematic review of the factors related to AST and changes in school transport over time in North America. *Prevent. Med.* 111, 314–322.
- Trang, N.H., Hong, T.K., Dibley, M.J., 2012. Active commuting to school among adolescents in Ho Chi Minh City, Vietnam: change and predictors in a longitudinal study 2004 to 2009. *Am. J. Prevent. Med.* 42, 120–128.
- Tudor-Locke, C., Ainsworth, B.E., Adair, L.S., Popkin, B.M., 2003. Objective physical activity of Filipino youth stratified for commuting mode to school. *Med. Sci. Sports Exercise* 35, 465–471.
- Samimi, A., Ermagun, A., 2013. Students' tendency to walk to school: case study of Tehran. *J. Urban Plan. Develop.* 139, 144–152.
- Shokoohi, R., Hanif, N.R., Dali, M., 2012a. Influence of the socio-economic factors on children's school travel. *Procedia-social Behav. Sci.* 50, 135–147.
- Shokoohi, R., Hanif, N.R., Dali, M.M., 2012b. Children walking to and from school in Tehran: associations with neighbourhood safety, parental concerns and children's perceptions. *Procedia-Social Behav. Sci.* 38, 315–323.
- Siiba, A., 2020. Active travel to school: Understanding the Ghanaian context of the underlying driving factors and the implications for transport planning. *J. Transport Health* 18, 100869.
- Singh, N., Vasudevan, V., 2018. Understanding school trip mode choice—The case of Kanpur (India). *J. Transport Geogr.* 66, 283–290.
- Sirard, J.R., Slater, M.E., 2008. Walking and bicycling to school: a review. *Am. J. Lifestyle Med.* 2, 372–396.
- Ward, C., 1978. *The Child in the City*, London Architectural Press.
- Waygood, E.O.D., Taniguchi, A., 2020. Japan: Maintaining high levels of walking. In: Waygood, E.O.D., Friman, M., Olsson, L.E., Mitra, R. (Eds.), *Transportation and Children's*, Elsevier, Amsterdam, Netherlands.
- Wong, B.Y.-M., Faulkner, G., Buliung, R., 2011. GIS measured environmental correlates of active school transport: a systematic review of 14 studies. *Int. J. Behav. Nutr. Phys. Activity* 8, 1–22.
- Yang, Y., Hong, X., Gurney, J.G., Wang, Y., 2017. Active travel to and from school among school-age children during 1997–2011 and associated factors in China. *J. Phys. Activity Health* 14, 684–691.
- Yang, Y., Lu, Y., Yang, L., Gou, Z., Zhang, X., 2020. Urban greenery, active school transport, and body weight among Hong Kong children. *Travel Behav. Soc.* 20, 104–113.
- Yang, Y., Xue, H., Liu, S., Wang, Y., 2019. Is the decline of active travel to school unavoidable by-products of economic growth and urbanization in developing countries? *Sustain. Cities Soc.* 47, 101446.
- Zhang, R., Yao, E., Liu, Z., 2017. School travel mode choice in Beijing, China. *J. Transport Geogr.* 62, 98–110.
- Zwerts, E., Allaert, G., Janssens, D., Wets, G., Witlox, F., 2010. How children view their travel behaviour: a case study from Flanders (Belgium). *J. Transport Geogr.* 18, 702–710.

Further Reading

- Bartlett, S., Hart, R., Satterthwaite, D., De La Barra, X., Missair, A., 2016. *Cities for Children: Children's Rights, Poverty and Urban Management*. Routledge.
- Brendan, G., Neil, S., 2006. *Creating Child Friendly Cities: New Perspectives and Prospects*. Taylor and Francis, London.
- Christensen, P., 2003. *Children in the City: Home Neighbourhood and Community*. Routledge, London.
- Christensen, P., Hadfield-Hill, S., Horton, J., Kraftl, P., 2017. *Children Living in Sustainable Built Environments: New Urbanisms, New Citizens*. Routledge, London.
- Driskell, D., 2017. *Creating Better Cities with Children and Youth: A Manual for Participation*. Routledge, London.

- Lange, A., 2018. *The Design of Childhood: How the Material World Shapes Independent Kids*. Bloomsbury Publishing, USA.
- Larouche, R., 2018. *Children's Active Transportation*. Elsevier, Amsterdam.
- Sarah, L.H., Gill, V., 2000. *Children's Geographies: Playing, Living, Learning*. Taylor and Francis, London.
- Waygood, E.O.D., Friman, M., Olsson, L.E., Mitra, R., 2019. *Transport and Children's Wellbeing*. Elsevier, San Diego.

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

Page left intentionally blank

INTERNATIONAL ENCYCLOPEDIA OF **TRANSPORTATION**

EDITOR-IN-CHIEF

Roger Vickerman

*School of Economics, University of Kent, Canterbury, United Kingdom
and*

Transport Strategy Centre, Imperial College, London, United Kingdom

VOLUME 7

**Transport Psychology
Transport Sustainability and Health**

SECTION EDITORS

Prof. Carlo G. Prato

*School of Civil Engineering, The University of Queensland
Brisbane, Australia*

Roger Vickerman

*School of Economics, University of Kent, Canterbury, United Kingdom
and*

Transport Strategy Centre, Imperial College, London, United Kingdom



Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Copyright © 2021 Elsevier Ltd. All rights reserved

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-08-102671-7

For information on all publications visit our
website at <http://store.elsevier.com>

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

Publisher: Oliver Walter
Acquisitions Editor: Oliver Walter
Content Project Manager: Natalie Lovell
Associate Content Project Manager: Manisha K and Ramalakshmi Boobalan
Designer: Matthew Limbert

EDITORIAL BOARD

Editor in Chief

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Section Editors

Maria Börjesson

Professor of Economics, VTI Swedish National Road and Transport Research Institute; Affiliated Professor at Linköping University, Sweden

Per Gårder

Department of Civil and Environmental Engineering, University of Maine, Orono, ME, United States

Prof. Kevin P.B. Cullinane

School of Business, Economics and Law, University of Gothenburg, Gothenburg, Sweden

Prof. Edward C.S. Chung

Department of Electrical Engineering, Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

Chandra R. Bhat

Center for Transportation Research (CTR), The University of Texas at Austin, TX, United States

Edoardo Marcucci

Department of Political Sciences, University of Roma Tre, Rome, Italy; Department of Logistics, Molde University College, Molde, Norway

Prof. Maria Attard

Department of Geography, Faculty of Arts, University of Malta, Msida, Malta

Prof. Carlo G. Prato

School of Civil Engineering, The University of Queensland, Brisbane, Australia

Roger Vickerman

School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

Page left intentionally blank

INTRODUCTION



Roger Vickerman

In an increasingly globalized world, despite reductions in costs and time, transportation has become even more important as a facilitator of economic and human interaction; this is reflected in technical advances in transportation systems, increasing interest in how transportation interacts with society and the need to provide novel approaches to understand its impacts. This has become particularly acute with the impact that Covid-19 has had on transportation across the world, at local, national, and international levels. This Encyclopedia brings a cross-cutting and integrated approach to all aspects of transportation from work in many disciplinary fields, engineering, operations research, economics, geography, and sociology to understand the changes taking place.

Transportation is both influenced by, and an influencer of, changes in the economy and society. Increasing speeds have reduced journey times and made the world a smaller place as globalization has affected both where people live and work and from where they source their goods and materials. Increasing volumes of traffic, often on old and life-expired infrastructure, lead to congestion and delays. Constraints on public budgets have led to increasing pressure on the private sector to fund improvements requiring new and innovative financial solutions. While there are clear differences in the nature of the pressures felt in the developed and less developed economies, there is an increasing recognition that in all societies, there is an accessibility problem such that certain groups become disadvantaged by the lack of access to reliable and cost-effective transport. While the problems are clearly multidimensional, research on transportation is often constrained by single disciplinary approaches and this carries over into the practice of transport planning and policy. The Encyclopedia cuts across these artificial boundaries by taking an approach that emphasizes the interaction between the different aspects of research and aims to offer new solutions to understand these problems. Each of the nine sections is based around a familiar dimension of work on transportation, but brings together the views of experts from different disciplinary perspectives. Each section is edited by an expert in the relevant field who has sought chapters from a range of authors representing different disciplines, different parts of the world, and different social perspectives. In this way, the work is not just a reflection of the state of the art that serves as a starting point for researchers and practitioners, but also a pointer toward new approaches, new ways of thinking, and novel solutions to problems.

The nine sections are structured around the following themes: Transport Modes; Freight Transport and Logistics; Transport Safety and Security; Transport Economics; Traffic Management; Transport Modeling and Data Management; Transport Policy and Planning; Transport Psychology; Sustainability and Health Issues in Transportation. Some of the chapters provide a technical introduction to a topic while others provide a bridge between topics or a more future-oriented view of new research areas or challenges. While there is guidance to cross-referencing between chapters, readers are encouraged to explore the tables of contents of all the sections to get a full understanding of the issues. Much of the Encyclopedia was completed before the Covid-19 pandemic and clearly this will have changed the situation in many areas covered by this work; the advantage of this type of reference work is that relevant updates will be possible in future editions.

The Encyclopedia has only been possible because of the cooperation of a large number of people. Robert Noland, Georgina Santos, Xiaowen Fu, and Dick Ettema served as an Editorial Advisory Board identifying possible editors of sections and advising on the overall structure. The Section Editors, Edoardo Marcucci, Kevin Cullinane, Per Garder, Maria Börjesson, Edward Chung, Chandra Bhat, Maria Attard, and Carlo Prato,

carried out the work of identifying potential authors of individual chapters, commissioning these, encouraging authors and reviewing drafts. More than 600 authors and co-authors of chapters are, however, are ultimately responsible for the success of this venture. Thanks are also due to the key people at Elsevier, particularly the Publishers, Andre Wolff and Oliver Walter, and the Project Managers, Sophie Harrison and Natalie Bentahar; they have shown exemplary patience over more than 3 years in bringing this work to fruition.

LIST OF CONTRIBUTORS TO VOLUME 7

Adelaide Cassia Nardocci

School of Public Health - USP, 715 Dr. Arnaldo Ave, SÃ£o Paulo, Brazil

Alireza Ermagun

Department of Civil and Environmental Engineering, Mississippi State University, Mississippi State, MS, United States

Amanda N. Stephens

Monash University Accident Research Centre, Monash University, Clayton, Australia

An-Magritt Kummeneje

Department of Civil Engineering, Kharazmi University, Faculty of Engineering, Tehran, Iran

Andrew Hill

School of Psychology, The University of Queensland, QLD, Australia

Andry Rakotonirainy

Centre for Accident Research and Road Safety, Kelvin Grove, QLD, Australia

Ann Williamson

Transport and Road Safety (TARS) Research, University of New South Wales, UNSW, Sydney, NSW, Australia

Aslak Fyhri

Institute of Transport Economics, Oslo, Norway

Barry Watson

Queensland University of Technology (QUT), Kelvin Grove, QLD, Australia

Birgitta Gatersleben

University of Surrey, School of Psychology, Guildford, United Kingdom

Bruno Dalla Chiara

POLITECNICO DI TORINO (I), Engineering, Dept. DIATI & Transport Systems, Italy

Bryan E. Porter

Old Dominion University, Norfolk, VA, United States

Bryan Matthews

University of Leeds, Leeds, United Kingdom

Carlo G. Prato

The University of Queensland, Brisbane, QLD, Australia

Carlos C. Silva

Center for Computer Graphics, Campus de Azur m, Guimar es, Portugal

Christopher R Cherry

Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, United States

Christopher Stokes

Centre for Automotive Safety Research, University of Adelaide, SA, Australia

Cl udia Aparecida Soares Machado

Department of Transportation Engineering - USP, 83 trav.2 Prof. Almeida Prado Ave, SÃ£o Paulo, Brazil

Clodoveu Augusto Davis

Department of Computer Science - UFMG, 2 Reitor - P res Albuquerque St, Belo Horizonte, MG 31270-901, Brazil

Cristiano Martins Monteiro

Department of Computer Science - UFMG, 2 Reitor P res Albuquerque St, Belo Horizonte, Brazil

D. Dodou

Delft University of Technology, Delft, The Netherlands

Daniele Contini

Institute of Atmospheric Sciences and Climate, ISAC-CNR Lecce, Italy

David Rodwell

Queensland University of Technology (QUT) Centre for Accident Research and Road Safety - Queensland (CARRS-Q), Kelvin Grove, QLD, Australia

David Rojas-Rueda

Department of Environmental and Radiological Health Sciences, Colorado State University, Fort Collins, CO, United States

Devajyoti Deka

Alan M. Voorhees Transportation Center, Edward J. Bloustein School of Planning and Public Policy, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States

Elisabete F. Freitas

University of Minho, Campus de Azurém, Guimarães, Portugal

Emanuel A. Sousa

Center for Computer Graphics, Campus de Azurém, Guimarães, Portugal

Emilian Szczepański

Warsaw University of Technology, Faculty of Production Engineering, Narbutta, Warsaw, Poland

Erik Jenelius

Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Ersilia Verlinghieri

Transport Studies Unit, School of Geography and the Environment, University of Oxford, Oxford, United Kingdom

Eva Merico

Institute of Atmospheric Sciences and Climate, ISAC-CNR Lecce, Italy

Feng Guo

Virginia Tech Transportation Institute, Blacksburg, VA, United States

Fernando Tobal Berssaneti

Institute of Energy and Environment - USP, 1289 Prof. Luciano Gualberto Ave, São Paulo, Brazil

Filip Fors Connolly

Department of Sociology, Umeå University, Umeå, Sweden

Gregoire Larue

Centre for Accident Research and Road Safety, Kelvin Grove, QLD, Australia

Haneen Khreis

Center for Advancing Research in Transportation Emissions, Energy, and Health (CARTEEH), Texas A&M Transportation Institute (TTI), College Station, TX, United States

Hanne Beate Sundfør

Institute of Transport Economics, Oslo, Norway

Hodaya Levy

Department of Cognitive Sciences, Hebrew University of Jerusalem, Jerusalem, Israel

Ilona Jacyna-Golda

Warsaw University of Technology, Faculty of Production Engineering, Narbutta 85, Warsaw, Poland

Ioni Lewis

Queensland University of Technology (QUT), Kelvin Grove, QLD, Australia

Isabel Kreibitz

Chemnitz University of Technology, Chemnitz, Germany

J.C.F. de Winter

Delft University of Technology, Delft, The Netherlands

Jake Olivier

Transport and Road Safety (TARS) Research, University of New South Wales, UNSW, Sydney, NSW, Australia

Jan Eusebio

Transport and Road Safety (TARS) Research, University of New South Wales, UNSW, Sydney, NSW, Australia

Jeffrey P. Michael

Center for Injury Research and Policy, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

Jeremy Woolley

Centre for Automotive Safety Research, University of Adelaide, SA, Australia

Joanne L. Harbluk

Human Factors and Crash Avoidance, Multi-Modal and Road Safety Programs, Ottawa, Ontario, Canada

John D. Bullough

Lighting Research Center, Rensselaer Polytechnic Institute, Troy, NY, United States

John Pearson

Council of Deputy Ministers Responsible for Transportation and Highway Safety, Ottawa, Ontario, Canada

Johnathon P. Ehsani

Center for Injury Research and Policy, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

Jonas Bårgman

Vehicle Safety Division at the Department of Mechanics and Maritime Science, Chalmers University of Technology, Gothenburg, Sweden

José Alberto Quintanilha

Institute of Energy and Environment – USP, 1289 Prof. Luciano Gualberto Ave, São Paulo, Brazil

Josef F. Krems

Chemnitz University of Technology, Chemnitz, Germany

Judith Charlton

Monash University Accident Research Centre, Clayton, VIC, Australia

Kanok Boriboonsomsin
University of California at Riverside, Riverside, CA, United States

Karel A. Brookhuis
University of Groningen, Faculty of Behavioural & Social Sciences, Department of Psychology, Groningen, The Netherlands

Katherine M. White
Queensland University of Technology (QUT), Kelvin Grove, QLD, Australia

Konrad Lewczuk
Air Force Institute of Technology, Ksiecia Bolesława 6, Warsaw, Poland

Kristie L. Johnson
Alexandria, VA, United States

Kristie Young
Monash University Accident Research Centre, Clayton, VIC, Australia

Lars-Göran Mattsson
Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

Lena Levin
Swedish National Road and Transport Research Institute, Sweden

Lewis M. Fulton
University of California, Davis, CA, United States

Lisa Dorn
Cranfield University, Cranfield, Bedfordshire, United Kingdom

Marco Diana
Politecnico di Torino - DIATI, Torino, Italy

Marianna Jacyna
Warsaw University of Technology, Faculty of Transport, Koszykowa, Warsaw, Poland

Mario Mongiardini
Centre for Automotive Safety Research, University of Adelaide, SA, Australia

Mariusz Izdebski
Warsaw University of Technology, Faculty of Transport, Koszykowa, Warsaw, Poland

Mark J. Nieuwenhuijsen
ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain; Universitat Pompeu Fabra (UPF), Barcelona, Spain; CIBER Epidemiologia y Salud Publica (CIBERESP), Madrid, Spain

Mark J.M. Sullman
Department of Social Sciences, University of Nicosia, Cyprus

Mark S. Horswill
School of Psychology, The University of Queensland, QLD, Australia

Michał, Kłodawski
Warsaw University of Technology, Faculty of Transport, Koszykowa, Warsaw, Poland

Mike Lenné
Monash University Accident Research Centre, Clayton, VIC, Australia

Milad Mehdizadeh
Iran University of Science and Technology (IUST), School of Civil Engineering, Department of Transportation, Narmak, Tehran, Iran

Mohammed Elhenawy
Centre for Accident Research and Road Safety, Kelvin Grove, QLD, Australia

Mohsen F. Zavareh
Iran University of Science and Technology (IUST), School of Civil Engineering, Department of Transportation, Narmak, Tehran, Iran

Narelle Haworth
Queensland University of Technology (QUT), Centre for Accident Research and Road Safety-Queensland, Brisbane, QLD, Australia

Orit Taubman - Ben-Ari
Bar Ilan University, Ramat Gan, Israel

Oscar Oviedo-Trespalacios
Centre for Accident Research and Road Safety – Queensland (CARRS-Q), Queensland University of Technology (QUT), Kelvin Grove, QLD, Australia

Paul Boase
Multi-Modal and Road Safety Programs, Ottawa, Ontario, Canada

Paulo Rui Anciaes
Centre for Transport Studies, Civil Environmental and Geomatic Engineering UCL (University College London), London, United Kingdom

Paweł, Gołda
Air Force Institute of Technology, Ksiecia Bolesława 6, Warsaw, Poland

Payam Dadvand
Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain

Piotr Goł, Ebiowski
Warsaw University of Technology, Faculty of Transport, Koszykowa, Warsaw, Poland

Raphael Grzebieta
Transport and Road Safety (TARS) Research, University of New South Wales, UNSW, Sydney, NSW, Australia

Richard Tay
MIT University, Melbourne, VIC, Australia

Roderick J. Lawrence
Geneva School of Social Sciences (G3S), University of Geneva Geneva, Switzerland

Roger L. Mackett
Centre for Transport Studies, University College London, London, Great Britain

Roger Vickerman
School of Economics, University of Kent, Canterbury, United Kingdom

Rosa Surinach
United Nations Human Settlements Programme –(UN-Habitat), Barcelona, Spain

Salvador Rueda Palenzuela
President of the Urban Ecology and Territorial Foundation, Barcelona, Spain

Sandra Rosenbloom
The University of Texas at Austin, Austin, TX, United States

Scott D. Jackson
Department of Environmental Conservation, University of Massachusetts, Amherst, MA, United States

Sherrie-Anne Kaye
Queensland University of Technology (QUT) –Centre for Accident Research and Road Safety – Queensland (CARRS-Q), Kelvin Grove, QLD, Australia

Sigal Kaplan
The Hebrew University of Jerusalem, Jerusalem, Israel

Sjaanie Koppel
Monash University Accident Research Centre, Clayton, VIC, Australia

Soheil Sohrabi
Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX, United States

Sonali Nandavar
Queensland University of Technology (QUT), Kelvin Grove, QLD, Australia

Sonja Forward
Swedish Road and Transport Research Institute, Sweden

Sonja Haustein
Technical University of Denmark, Department of Technology, Management and Economics, Lyngby, Denmark

Türker Özkan
Department of Psychology, Middle East Technical University (METU), Ankara, Turkey

Teresa Senserrick
Queensland University of Technology (QUT) –Centre for Accident Research and Road Safety - Queensland (CARRS-Q), Kelvin Grove, QLD, Australia

Thomas D. Berry
Christopher Newport University, Newport News, VA, United States

Timo Lajunen
Department of Psychology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Tommy Gørling
Centre for Finance and Department of Economics, School of Business, Economics, and Law, University of Gothenburg, Göteborg, Sweden

Torbjørn Rundmo
Norwegian University of Science and Technology (NTNU), Department of Psychology, Trondheim, Norway

Tova Rosenbloom
Research Institute of Human Factors in Road Safety, Management Department, Bar Ilan University, Tel Aviv, Israel

Trond Nordfjrn
Norwegian University of Science and Technology (NTNU), Department of Psychology, Trondheim, Norway

Yi Wen
Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, United States

CONTENTS OF ALL VOLUMES

<i>Editorial Board</i>	<i>v</i>
<i>Introduction</i>	<i>vii</i>
<i>Contributors to Volume 7</i>	<i>ix</i>

VOLUME 1

Introduction to Transportation Economics	1
Market Failures and Public Decision Making in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	2
Demand for Freight Transport <i>José Holguín-Veras and Diana G. Ramírez-Ríos</i>	7
Cost Functions for Road Transport <i>José Manuel Vassallo</i>	13
Future of Urban Freight <i>Russell G. Thompson</i>	20
Operation Costs for Public Transport <i>Marco Batarce</i>	26
Natural Monopoly in Transport <i>André de Palma and Julien Monardo</i>	30
Freight Costs: Air and Sea <i>Yulai Wan and Dong Yang</i>	36
Transport Production and Cost Structure <i>Ricardo Giesen and Darío Farren</i>	46
The Concept of External Cost: Marginal versus Total Cost and Internalization <i>Sofia F. Franco</i>	56
Value of Time <i>John J. Bates</i>	67
Valuation of Carbon Emissions <i>Svante Mandell</i>	72
Valuation of Travel Time Variability Using Scheduling Models <i>Katrine Hjorth</i>	76

Value of Crowding <i>Daniel Hörcher</i>	84
What Drives Transport and Mobility Trends? The Chicken-and-Egg Problem <i>Nathalie Picard</i>	89
Pricing Principles in the Transport Sector <i>Bruno De Borger and Stef Proost</i>	95
Long-Run Versus Short-Run Valuations <i>Stefanie Peer</i>	102
Value of Noise <i>Jon P. Nelson</i>	106
The Value of Life and Health <i>Henrik Andersson</i>	114
The Value of Security, Access Time, Waiting Time, and Transfers in Public Transport <i>Raquel Espino, Juan de Dios Ortúzar, and Luis I. Rizzi</i>	122
Demand for Passenger Transportation <i>Kenneth A Small and Robin Lindsey</i>	127
Real-World Experiences of Congestion Pricing <i>Charles Raux</i>	134
Distributional Effects of Congestion Charges and Fuel Taxes <i>Jonas Eliasson</i>	139
The Bottleneck Model <i>Dereje Abegaz and Yili Tang</i>	146
Dynamic Congestion Pricing and User Heterogeneity <i>Kathrin Goldmann and Gernot Sieg</i>	150
Economics of Parking <i>Daniel Albalade and Albert Gragera</i>	159
Loss Aversion and Size and Sign Effects in Value of Time Studies <i>Andrew Daly</i>	165
Intertemporal Variation of Valuations <i>James Fox</i>	170
The Rebound Effect for Car Transport <i>Bruno De Borger, Ismir Mulalic and Jan Rouwendal</i>	174
Elasticities for Travel Demand: Recent Evidence <i>Fay Dunkerley, Charlene Rohr and Mark Wardman</i>	179
Parking Price Elasticities <i>Stephan Lehner</i>	185
The Pareto Criterion and the Kaldor Hicks Criterion <i>Harald Minken</i>	190
Social Discount Rates <i>Lars Hultkrantz</i>	195
Cross-Elasticities between Modes <i>Mark Wardman and Jeremy Toner</i>	201
First-Best Congestion Pricing <i>Moez Kilani</i>	209

Ethical Aspects-Can We Value Life, Health, and Environment in Money Terms? <i>Jose M. Grisolia and Ken Willis</i>	216
Car tolls, Transit Subsidies for Commuting, and Distortions on the Labor Market <i>Ioannis Tikoudis¹ and Kurt Van Dender¹</i>	221
Demand Management and Capacity Planning of Airports <i>Stephen Ison and Lucy Budd</i>	227
Dealing With Negative Externalities: Low Emission Zones Versus Congestion Tolls <i>Valeria Bernardo, Xavier Fageda, and Ricardo Flores-Fillol</i>	231
The Rule-of-a-Half and Interpreting the Consumer Surplus as Accessibility <i>Mogens Fosgerau and Ninette Pilegaard</i>	237
Producer Surplus <i>Chau Man Fung</i>	242
The Robustness of Cost-Benefit Analyses <i>Morten Welde and James Odeck</i>	249
The GDP Effects of Transport Investments: The Macroeconomic Approach <i>James Laird and Daniel Johnson</i>	256
The Mohring Effect <i>Hugo E. Silva</i>	263
Public Transport Fare and Subsidy Optimization <i>Qianwen Guo and Zhongfei Li</i>	267
Transportation Equity <i>Rafael H. M. Pereira, and Alex Karner</i>	271
Impact of Transport Cost-Benefit Analysis on Public Decision-Making <i>Niek Mouter</i>	278
Causal Inference for Ex Post Evaluation of Transport Interventions <i>Daniel J. Graham</i>	283
Transport Cost and Location of Firms <i>Adelheid Holl</i>	291
Commuting, the Labor Market, and Wages <i>Jan Rouwendal, and Ismir Mulalic</i>	297
How to Buy Transport Infrastructure <i>Johan Nyström</i>	302
Procurement of Public Transport: Contractual Regimes <i>Andrew Smith, and Chris Nash</i>	308
The Mono-Centric City Model and Commuting Cost <i>Zhi-Chun Li, and Ya-Juan Chen</i>	315
Value of Time in Freight Transport <i>Gerard de Jong</i>	321
The Economics and Planning of Urban Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	326
Incentives in Public Transport Contracts <i>Andreas Vigren</i>	332
Transportation Improvements and Property Prices <i>Alex Anas</i>	337

Transport Infrastructure Effects on Economic Output: The Microeconomic Approach <i>Patricia C. Melo</i>	347
Wider Economic Impacts of Transport Investments <i>Anthony J. Venables</i>	355
Employment Effects of Transport Infrastructure <i>Anna Matas, and Javier Asensio</i>	360
Regulatory Reforms and Competition in Public Transport <i>Didier van de Velde</i>	365
How to Finance Transport Infrastructure? <i>Tiziana D'Alfonso, and Giuseppe Catalano</i>	371
Congestion, Allocation and Competition on the Railway Tracks <i>John Armstrong, and John Preston</i>	378
Public Private Partnership <i>David Meunier, and Emile Quinet</i>	385
Airline Economics <i>Anming Zhang, Yahua Zhang, and Zhibin Huang</i>	392
Privatization and Deregulation of the Airline Industry <i>Achim I. Czerny, and Hao Lang</i>	397
Price Discrimination and Yield Management in the Airline Industry <i>Guoquan Zhang, Colin C.H. Law, Yahua Zhang, and Hangjun Yang</i>	404
Regulation and Competition in Railways <i>Chris Nash, and Andrew Smith</i>	409
Cycling Economics <i>Bert van Wee</i>	414
The Economic Rationale for High-Speed Rail <i>Ginés de Rus</i>	419
Rail Cost Functions <i>Kristofer Odolinski, and Phill Wheat</i>	425
Transport and International Trade <i>Siri Pettersen Strandenes</i>	431
Maritime Economics: Organizational Structures <i>Kevin P.B. Cullinane</i>	436
Port Planning and Investment <i>Jasmine Siu Lee Lam</i>	443
Estimating the Capital Stock of Transport Infrastructure <i>Heike Link</i>	449
The Economics of Reducing Carbon Emissions From Air and Road Transport <i>Olga Ivanova</i>	457
Regulation and Financing of Toll Roads <i>Marco Ponti</i>	464
Are Megaprojects too Transformational for Cost-Benefit Analysis? <i>Tom Worsley</i>	470
Economics of Transportation Safety <i>Ian Savage</i>	476

Cost Overruns of Transportation Infrastructure Projects <i>James Odeck, and Morten Welde</i>	483
The Downs-Thomson Paradox <i>Joel P. Franklin</i>	490
Policy Instruments for Plug-In Electric Vehicles: An Overview and Discussion <i>Jake Whitehead, Patrick Plötz, Patrick Jochem, Frances Sprei, and Elisabeth Dütschke</i>	496
Vertical and Horizontal Separation in the European Railway Sector and Its Effects on Productivity <i>Pedro Cantos-Sánchez</i>	503
How will Autonomous Vehicles Impact Car Ownership and Travel Behavior <i>Patrick M. Bösch, Felix Becker, Henrik Becker, and Kay W. Axhausen</i>	508
Policy Instruments to Reduce Carbon Emissions from Road Transport Computable General Equilibrium Analysis in Transportation Economics <i>Johannes Bröcker</i>	520
Estimation of Value of Time <i>Stefan Flügel, and Askill H. Halse</i>	527
The Taxation of Car Use in the Future <i>Griet De Ceuster, and Inge Mayeres</i>	534
Cost-Benefit Analysis and Other Assessment Techniques: Contrasts and Synergies <i>Paolo Beria</i>	540
Demand for Air Travel and Income Elasticity <i>Jing Lu, Yucan Meng, Changmin Jiang, and Cheng Lv</i>	547
Generalized Cost for Transport <i>Jeppé Rich</i>	555
The Impact of Electric Vehicles on Energy Systems <i>Patrick E.P. Jochem, Jake Whitehead, and Elisabeth Dütschke</i>	560
Uber versus Taxis <i>Georgina Santos</i>	566
Contract Efficiency in Public Transport Services <i>Philippe Gagnepain, and Marc Ivaldi</i>	572
Company Cars <i>Stefan Gössling</i>	580
Transportation Network Companies (TNCs) and the Future of Public Transportation <i>Susan Shaheen, and Adam Cohen</i>	584
Public Transport in Low Density Areas <i>Jani-Pekka Jokinen, Leif Sörensen, and Jan Schlüter</i>	589
Market Failures in Transport: Direct and Indirect Public Intervention <i>Federico Boffa, and Alberto Iozzi</i>	596
The Braess Paradox <i>Anna Nagurney, and Ladimer S. Nagurney</i>	601
VOLUME 2	
Introduction to Transportation Safety and Security <i>Per Gårder</i>	1

The Concept of “Acceptable Risk” Applied to Road Safety Risk Level <i>Claes Tingvall</i>	2
Crash Not Accident <i>Robert A. Scopatz</i>	6
Age and Gender as Factors in Road Safety <i>Marion Sinclair</i>	11
Aggressive Driving and Road Rage <i>James E.W. Roseborough, Christine M. Wickens, and David L. Wiesenthal</i>	17
Aircraft Maintenance and Inspection <i>Alan Hobbs</i>	25
Airport Security <i>Richard W. Bloom</i>	34
Transport Safety and Security: Alcohol <i>James C. Fell</i>	40
Animal Crashes <i>Michal Bil</i>	53
Attenuators <i>Simonetta Boria</i>	63
ATV, Snowmobile, and Terrain Vehicle Safety <i>David P. Gilkey, and William Brazile</i>	77
Automobile Safety Inspection <i>Subasish Das</i>	85
Aviation Safety: Commercial Airlines <i>Clarence C. Rodrigues</i>	90
Aviation Safety, Freight, and Dangerous Goods Transport by Air <i>Glenn S. Baxter and Graham Wild</i>	98
Bicycle Collision Avoidance Systems: Can Cyclist Safety be Improved with Intelligent Transport Systems? <i>Lars Leden</i>	108
Bicycle Infrastructure <i>Rock E. Miller</i>	115
Bicycles, E-Bikes and Micromobility, A Traffic Safety Overview <i>Höskuldur Kröyer</i>	125
Bicycles: The Safety of Shared Systems Versus Traditional Ownership <i>Mercedes Castro-Nuño and José I. Castillo-Manzano</i>	139
Bridge Safety <i>Per Erik Garder</i>	144
Carsharing Safety and Insurance <i>Elliot Martin and Susan Shaheen</i>	150
Carjacking <i>Terance D. Mieth, Christopher Forepaugh, Tanya Dudinskaya</i>	157
Collision Avoidance Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	164
Collision Avoidance Systems, Automobiles <i>Erick J. Rodríguez-Seda</i>	173

Connected Automated Vehicles: Technologies, Developments, and Trends <i>Azra Habibovic and Lei Chen</i>	180
Construction Zones <i>Jalil Kianfar</i>	189
Costs of Accidents <i>Ulf Persson</i>	196
Critical Issues for Large Truck Safety <i>Matthew C. Camden, Jeffrey S. Hickman, Richard J. Hanowski, and Martin Walker</i>	200
Demerit Points and Similar Sanction Programs <i>Matúš ťucha and Kristýna Josrová</i>	210
Driver State and Mental Workload <i>Dick de Waard and Nicole van Nes</i>	216
Drugs, Illicit, and Prescription <i>Rune Elvik</i>	221
Education, Training, and Licensing <i>Matúš ťucha and Kristýna Josrová</i>	228
Elderly Driver Safety Issues <i>Mark J King</i>	233
Emergency Response Systems <i>Frances L. Edwards</i>	240
Emergency Vehicles and Traffic Safety <i>Shamsunnahar Yasmin, Sabreena Anowar, and Richard Tay</i>	247
Encouragement: Awards and Incentives <i>Fred Wegman</i>	255
Enforcement and Fines <i>Matúš ťucha and Ralf Risser</i>	263
Epidemiology of Road Traffic Crashes <i>Sherrie-Anne Kaye, Judy Fleiter, and Md Mazharul Haque</i>	269
Evacuation Planning and Transportation Resilience <i>Karl Kim</i>	276
Exposure: A Critical Factor in Risk Analysis <i>Frank Gross, PhD, PE</i>	282
Passenger Ferry Vessels and Cruise Ships: Safety and Security <i>Wayne K. Talley</i>	290
Fuel Economy Standards: Impacts on Safety <i>Kenneth T. Gillingham and Stephanie M. Weber</i>	296
Hazardous Materials Transport <i>Dr. Arjan Vincent van der Vlies</i>	304
Head-on Crashes <i>John N. Ivan</i>	311
Helicopters in Emergency Medical Response <i>Stephen J.M. Sollid, M.D., PhD</i>	316
Horizontal and Vertical Geometry <i>Victoria Gitelman</i>	322

Human Factors in Transportation <i>Alison Smiley, Christina (Missy) Rudin-Brown</i>	331
In-Depth Crash Analysis and Accident Investigation <i>Yong Peng, Helai Huang, and Xinghua Wang</i>	346
Incident Detection Systems, Airplanes <i>Ivan Ostroumov and Nataliia Kuzmenko</i>	351
Inequality and Traffic Safety <i>Miles Tight</i>	358
Lighting <i>John D. Bullough</i>	361
Macroscopic Safety Analysis <i>Mohamed Abdel-Aty and Jaeyoung Lee</i>	367
Motor Vehicle Crash Reportability <i>John J. McDonough</i>	380
Nominal Safety <i>Per Erik Garder</i>	386
Parking Lots <i>Maxim A. Dulebenets</i>	392
Passenger Van Safety <i>Saksith Chalermpong and Apiwat Ratanawaraha</i>	401
Passive Prevention Systems in Automobile Safety <i>B. Serpil Acar</i>	406
Pedestrian Safety, Children <i>Mette Møller</i>	415
Pedestrian Safety, General <i>Muhammad Z. Shah, Mehdi Moeinaddini, and Mahdi Aghaabbasi</i>	420
Pedestrian Safety, Older People <i>Carlo Lui</i>	429
Visually Impaired Pedestrian Safety <i>Robert S. Wall Emerson</i>	435
Photo/Video Traffic Enforcement <i>Charles M. Farmer</i>	439
Powered Two- and Three-Wheeler Safety <i>Fangrong Chang, Helai Huang, and Md. Mazharul Haque</i>	443
Railroad Safety <i>Xiang Liu and Zhipeng Zhang</i>	451
Railroad Safety: Grade Crossings and Trespassing <i>Rahim F. Benekohal and Jacob Mathew</i>	466
Recreational Boating Safety: Usage, Risk Factors, and the Prevention of Injury and Death <i>Amy E. Peden, Stacey Willcox-Pidgeon, and Kyra Hamilton</i>	477
Refuge Islands <i>Christer Hydén</i>	487
Risk Perception and Risk Behavior in the Context of Transportation <i>Martina Raue and Eva Lerner</i>	494

Road Diets	500
<i>Robert B. Noland</i>	
Road Safety Audits	508
<i>Xiao Qin</i>	
Roadside Safety Barriers	513
<i>Dean C. Alberson</i>	
Road Safety Management in Selected Countries	519
<i>Paul Boase and Brian Jonah</i>	
Roadway Pavement Conditions and Various Summer and Winter Maintenance Strategies	529
<i>Kamal Hossain, PhD P. Eng</i>	
Safety of Roundabouts	539
<i>Khaled Shaaban</i>	
Rumble Strips, Continuous Shoulder, and Centerline	549
<i>AnnaAnund and AnnaVadeby</i>	
Safety Culture	554
<i>Tor-Olav Nvestad</i>	
Safety Data Quality Management	560
<i>Robert A. Scopatz</i>	
School Bus Safety	564
<i>Yousif A. Abulhassan</i>	
School Campus Traffic Circulation	568
<i>Dimitrios Nalmpantis</i>	
Sexual Violence in Public Transportation	576
<i>Vania Ceccato</i>	
Shared Space	584
<i>Allison B. Duncan</i>	
Side Area Safety and Side Slopes	593
<i>José María Pardillo-Mayora</i>	
Simulators	602
<i>Nichole L. Morris, Curtis M. Craig, Jacob D. Achtemeier, and Peter A. Easterlund</i>	
Sleep-Related Issues and Fatigue	611
<i>Prof Marion Sinclair and Estelle Swart</i>	
Speed Governors and Limiters	617
<i>Christer Hydén</i>	
Speed Limits on Rural Highways	622
<i>Peter Tarmo Savolainen and Timothy Jordan Gates</i>	
Speed Limits on Urban Streets	632
<i>Anna Bray Sharpin, Claudia Adriazola-Steil, Ben Welle, and Natalia Lleras</i>	
Speed-Reducing Measures	641
<i>Vinod Vasudevan</i>	
Striping, Signs, and Other Forms of Information	648
<i>Kasem Choocharukul and Kerkritt Sriroongvikrai</i>	
Suicides	656
<i>Brendan Ryan</i>	

Surrogate Measures of Safety <i>Nicolas Saunier and Aliaksei Laureshyn</i>	662
Targeting Transit: the Terrorist Threat and the Challenges to Security <i>Brian Michael Jenkins</i>	668
The Swedish Vision Zero: A Policy Innovation <i>Matts-Åke Belin</i>	675
Tire Safety <i>Saied Taheri</i>	681
Town Gates: Section on Transport Safety and Security <i>Charles Tijus</i>	685
Traffic Flow Volume and Safety <i>Athanasios Theofilatos and Apostolos Ziakopoulos</i>	692
Traffic Safety and Security of Taxis and Ride-Hailing Vehicles <i>Zhe Wang, Helai Huang, and Ye Li</i>	699
Traffic Signals and Safety <i>Andrew P. Tarko</i>	706
Tunnels, Safety and Security Issues-Risk Assessment for Road Tunnels: State-of-the-Art Practices and Challenges <i>Konstantinos Kirytopoulos, Panagiotis Ntzeremes, and Konstantinos Kazaras</i>	713
Understanding, Managing, and Learning from Disruption <i>Karl Kim</i>	719
Use/Analysis of Crash Data and Underreporting of Crashes <i>Mohammadali Shirazi and Dominique Lord</i>	726
Utility Poles <i>Lai Zheng and Tarek Sayed</i>	731
Value of Life and Injuries <i>David A. Hensher</i>	737
Effects of Weather <i>Maria Pregnolato, Amirhassan Kermanshah, and Wisinee Wisetjindawat</i>	742
Wrong-Way Driving on Motorways <i>Huaguo Zhou and Md Atiquzzaman</i>	751

VOLUME 3

Freight Transport and Logistics <i>Sharon Cullinane and Kevin Cullinane</i>	1
Expanding the Perspective of Logistics and Supply Chain Management <i>David J. Closs</i>	8
Logistics and Supply Chain Management Performance Measures <i>David B. Grant and Sarah Shaw</i>	16
Economic Regulation/Deregulation and Nationalization/Privatization in Freight Transportation <i>Wayne K. Talley</i>	24
Freight Transport Policy <i>Luca Zamparini and Aura Reggiani</i>	29

Planning and Financing Logistics Spaces <i>Nicolas Raimbault</i>	35
Supply Chain Risk Management: Creating the Resilient Supply Chain <i>Richard Wilding</i>	41
Transportation Safety and Security <i>Maria G. Burns</i>	47
Resilience in Freight Transport Networks <i>Zhuohua Qu, Chengpeng Wan, and Zaili Yang</i>	53
Environmental Sustainability in Freight Transportation <i>Lisa M. Ellram</i>	58
Sustainable Logistics, CSR in Logistics, and Sustainable Supply Chain Management <i>Maria Björklund and Maja Piecyk-Ouellet</i>	64
Information Sharing and Business Analytics in Global Supply Chains <i>Prof Usha Ramanathan and Prof Ramakrishnan Ramanathan</i>	71
Logistics Information Systems <i>Petri Helo and Javad Rouzafzoon</i>	76
Factors Affecting the Selection of Logistics Service Providers <i>Aicha AGUEZZOUL</i>	85
Logistics Service Performance <i>Kee-hung Lai, Jinan Shao, and Yongyi Shou</i>	89
The World Bank's Logistics Performance Index <i>Christina K. Wiederer cwiederer, Jean-François Arvis, Lauri M. Ojala, and Tuomas M. M. Kiiski</i>	94
Outsourcing Logistics Functions <i>Evi Hartmann, Hendrik Birkel, and Matthias Kopyto</i>	102
Freight Transport and Logistics in JIT Systems <i>James H. Bookbinder and M. Ali Aelkū</i>	107
Supply Chain Finance <i>Erik Hofmann</i>	113
Packaging Logistics <i>Jesús García-Arca, Alicia Trinidad González-Portela Garrido, and J. Carlos Prado-Prado</i>	119
The Bullwhip Effect <i>Jan C. Fransoo and Maximiliano Udenio</i>	130
Blockchain Applications in Logistics <i>Yingli Wang</i>	136
Logistics in Asia <i>Shong-lee Ivan Su</i>	143
Logistics in the Developing World <i>Charles Kunaka</i>	150
Freight Network Modeling <i>Lóránt Tavasszy and Yousef Maknoon</i>	157
National Freight Transport Models <i>Gerard de Jong</i>	162
Freight Flows in Cities <i>Genevieve Giuliano</i>	168

Urban Logistics and Freight Transport <i>Michael Browne, José Holguín-Veras, and Julian Allen</i>	178
Omni-Channel Logistics <i>Tom Van Woensel</i>	184
Humanitarian Logistics <i>Gyöngyi Kovács and Diego Vega</i>	190
Optimization of Humanitarian Logistics <i>M. Teresa Ortuño, Jose M. Ferrer, Inmaculada Flores, and Gregorio Tirado</i>	195
Event Logistics <i>Rev. Ruth Dowson and Dan Lomax</i>	201
Reverse Logistics <i>Dale S. Rogers and Ronald S. Lembke</i>	208
E-Tailing and Reverse Logistics <i>Sharon Cullinane</i>	219
Green Routing of Freight Vehicles <i>Tolga Bektaş</i>	224
Freight Mode Choice <i>Hyun Chan Kim and Alan Nicholson</i>	231
The Value of Time in Freight Transport <i>María Feo-Valero, Amaya Vega, and Bárbara Vázquez-Paja</i>	236
Behavioral Research in Freight Transport <i>Edoardo Marcucci, Valerio Gatta, and Michela Le Pira</i>	242
Container (Liner) Shipping <i>Theo Notteboom</i>	247
Bulk Shipping Markets: An Overview of Market Structure and Dynamics <i>Manolis G. Kavussanos and Stella A. Moysiadou</i>	257
Ferries and Short Sea Shipping <i>Lourdes Trujillo and Alba Martínez-López</i>	280
Shipping and the Environment <i>Karin Andersson, Selma Brynolf, Lena Granhag, and J. Fredrik Lindgren</i>	286
Energy Efficiency of Ships <i>Harilaos N. Psaraftis</i>	294
Seaports <i>Mary R. Brooks and Geraldine Knatz</i>	299
Port Hinterlands <i>Francesco Parola, Giovanni Satta, and Francesco Vitellaro</i>	305
Seaports as Clusters of Economic Activities <i>Peter W. de Langen</i>	310
Port Efficiency and Effectiveness <i>Lourdes Trujillo, María Manuela González, Casiano Manrique-De-Lara-Peñate, and Ivone Pérez</i>	316
Container Port Automation <i>Michael G.H. Bell</i>	323
Optimizing Crane Operations in Ports Scheduling of Liner Container Shipping Services	335

<i>Yuquan Du and Qiang Meng</i>	
Dry Ports	344
<i>Gordon Wilmsmeier and Jason Monios</i>	
Arctic Shipping	349
<i>Yufeng Lin, David G. Babb, and Adolf K.Y. Ng</i>	
Airfreight and Economic Development	355
<i>Kenneth Button</i>	
Air Freight Logistics	361
<i>Keith Debbage and Neil Debbage</i>	
Air Freight Marketing	369
<i>Lucy Budd and Stephen Ison</i>	
Drones in Freight Transport	374
<i>Oliver Kunze</i>	
Duty of Care in the Selection of Motor Carriers	382
<i>Thomas M. Corsi</i>	
Carrier Selection for Less-Than-Truckload (LTL) Shipments	388
<i>Dinçer Konur, Gonca Yildirim, and Bahriye Cesaret</i>	
Decarbonizing Road Freight Transport	395
<i>Heikki Liimatainen</i>	
The Rebound Effect in Road Freight Transport	402
<i>Tooraj Jamasb and Manuel Llorca</i>	
Autonomous Goods Transport	407
<i>Heike Flømig</i>	
Rail Freight	413
<i>Dr Allan Woodburn</i>	
Rail Freight Vehicles	423
<i>Maksym Spiryagin, Qing Wu, Peter Wolfs, Colin Cole, Valentyn Spiryagin, and Tim McSweeney</i>	
Eurasia Rail Freight: Enablers and Inhibitors of Future Growth	436
<i>Hendrik Rodemann and Simon Templar</i>	
Intermodal and Synchromodal Freight Transport	456
<i>Tomas Ambra, Koen Mommens, and Cathy Macharis</i>	
Pipelines	463
<i>Matthew E. Oliver</i>	
3D Printers and Transport	471
<i>Wouter P.C. Boon and Bert van Wee</i>	
The Physical Internet and Logistics	479
<i>Eric Ballot and Shenle PAN</i>	
Bicycles for Urban Freight	488
<i>Barbara Lenz and Johannes Gruber</i>	
China's Belt and Road Initiative	495
<i>Paul Tae-Woo Lee</i>	

VOLUME 4

Introduction to Traffic Management <i>Edward C.S. Chung</i>	1
Urban Motorway Management <i>John Gaffney and Hendrik Zurlinden</i>	2
Ramp Metering Application <i>John Gaffney and Hendrik Zurlinden</i>	10
City Wide Coordinated Ramp Meters <i>John Gaffney and Hendrik Zurlinden</i>	21
Variable Speed Limits for Traffic Efficiency Improvement <i>José Ramón D. Frejo and Bart De Schutter</i>	33
Hard Shoulder Running <i>Justin Geistefeldt</i>	41
High-Occupancy Vehicle (HOV) and High-Occupancy Toll (HOT) Lanes <i>Roxana J. Javid, Jiani Xie, Lijiao Wang, Wenruifan Yang, Ramina Jahanbakhsh Javid, and Mahmoud Salari</i>	45
Reversible Lanes: Guidelines, Operation and Control, Research Directions <i>Gowri Asaithambi, Venkatesan Kanagaraj, and Madhuri Kashyap</i>	52
Electronic Toll Collection <i>Azusa Toriumi</i>	60
Road Pricing-Theory and Applications <i>Kian Keong Chin</i>	68
Road Pricing 1: The Theory of Congestion Pricing <i>Timothy D. Hau</i>	74
Road Pricing 2: Short- and Long-Run Equilibrium of Road Transportation <i>Timothy D. Hau</i>	83
Road Pricing 3: The Implications for Pricing Public Transportation <i>Timothy D. Hau</i>	90
Road Pricing 4: Case Study-The Implementation of Electronic Road Pricing in Hong Kong <i>Timothy D.</i>	103
Advanced Travelers Information Systems (ATIS) <i>Chintan Advani and Ashish Bhaskar</i>	106
Travel Time Reliability <i>Sharmili Banik, Anil Kumar, and Lelitha Vanajakshi</i>	109
Traffic Incident Management <i>Ruimin Li</i>	122
Traffic Incident Detection <i>Shuyan Chen and Yingjiu Pan</i>	128
Bottleneck <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	134
Flow Breakdown <i>Kentaro Wada, Toru Seo, and Yasuhiro Shiomi</i>	143
Recurrent Congestion <i>Takahiro Tsubota</i>	154

Nonrecurrent Congestion <i>Takahiro Tsubota</i>	158
Freeway to Arterial Interfaces <i>Abolfazl Karimpour and Yao-Jan Wu</i>	162
Arterial Road Management <i>Jiaqi Ma, Yi Guo and Adekunle Adebisi</i>	169
Signalized Intersections <i>Chaitrali Shirke</i>	178
Protected Phase <i>Jiarong Yao, Yumin Cao, and Keshuang Tang</i>	185
Permitted Phase <i>Yumin Cao, Jiarong Yao, and Keshuang Tang</i>	196
Hook Turns: Implementation, Benefits, and Limitations <i>Sara Moridpour and Amir Falamarzi</i>	203
Traffic Signal Coordination <i>Rahim F. Benekohal</i>	213
Emergency Vehicle Priority (Preemption): Concept and Advancements <i>Chaitrali Shirke</i>	221
Signalized Roundabouts <i>Yetis Sazi Murat and Rui-jun Guo</i>	227
Turbo Roundabouts: Design, Capacity and Comparison With Alternative Types of Roundabouts <i>Marco Guerrieri and Raffaele Mauro</i>	238
Capacity of an Intersection <i>Ashish Verma and Milan Mathew Thomas</i>	247
Delay <i>Shinji Tanaka</i>	258
Queue Length <i>Shinji Tanaka</i>	263
Local Area Traffic Management <i>Michael A.P. Taylor</i>	268
On-Street Parking <i>Jun Chen and Guang Yang</i>	278
Off-Street Parking <i>Jun Chen and Guang Yang</i>	285
Parking Information Systems <i>Behrang Assemi and Douglas Baker</i>	289
Performance-Based Parking Management <i>Douglas Baker and Behrang Assemi</i>	299
Transit Priority <i>Wanjing Ma, Qiheng Lin, and Ling Wang</i>	307
Adaptive Bus Control <i>Monica Menendez</i>	315
Transit Fare Collection <i>Mahmoud Mesbah and Kamal Khanali</i>	325

Transit Information Systems <i>Ankit Kumar Yadav and Nagendra R Velaga</i>	331
Tram Lane Configurations and Driving Rules <i>Farhana Naznin</i>	338
Railway Crossing <i>Chunliang Wu and Inhi Kim</i>	341
Pedestrian Crossing (Crosswalk) <i>Miho Iryo-Asano, Wael K.M. Alhajyaseen, and Koji Suzuki</i>	346
Bike-Sharing System: Uncovering the “1Success Factors” <i>S.K. Jason Chang and Amanda Fernandes Ferreira</i>	355
Airspace Systems Technologies-Overview and Opportunities <i>Banavar Sridhar and Gano B. Chatterji</i>	363
Air Traffic Flow and Capacity Management <i>Li Weigang and Cristiano P Garcia</i>	380
Port Management <i>Maria G. Burns</i>	390
Port Performance Measurement from a Multistakeholder Perspective <i>Min-Ho Ha, Zaili Yang, and Young-Joon Seo</i>	396
Transport Modeling and Data Management <i>Chandra Bhat</i>	407
Computational Methods and Data Analytics <i>Bilal Farooq and David López</i>	408
Activity-Based Models <i>Renato Guadamuz and Rajesh Paleti</i>	414
Advanced Traveler Information Systems <i>Stephen D. Boyles</i>	418
Demand-Responsive Transit, Evaluation Studies <i>Sebastián Raveau</i>	423
Location Choice Models <i>Adam Wilkinson Davis</i>	428
Full Feedback and Equilibrium Modeling in Urban Travel Forecasting <i>Yu (Marco) Nie, Jun Xie, and David Boyce</i>	432
The Use and Value of Geographic Information Systems in Transportation Modeling <i>Ming Zhang</i>	440
Latent Demand and Induced Travel <i>Charisma F. Choudhury</i>	448
ICT, Virtual and In-Person Activity Participation, and Travel Choice Analysis <i>Jacek Pawlak and Giovanni Circella</i>	452
Public Transit Ridership Forecasting Models <i>Ipsita Banerjee, Deepa L, and Abdul Rawoof Pinjari</i>	459
Spatial Mismatch, Job Access, and Reverse Commuting <i>Gian-Claudia Sciarra</i>	468

Microsimulation and Agent-Based Models in Transportation <i>Milos Balac</i>	473
Choice Models in Transportation <i>Naveen Chandra Iraganaboina and Naveen Eluru</i>	477
Multi-Criteria Decision Analysis <i>Zhanmin Zhang and Srijith Balakrishnan</i>	485
The National Household Travel Survey Data Series (NPTS/NHTS) <i>Nancy McGuckin</i>	493
Route Choice and Network Modeling <i>Emma Frejinger and Maëlle Zimmermann</i>	496
Departure Time Choice Modeling <i>Khandker Nurul Habib</i>	504
Traveler Responses to Congestion <i>David T. Ory and Gayathri Shivaraman</i>	509
Origin-Destination Demand Estimation Models <i>William H.K. Lam, Hu Shao, Shuhan Cao, and Hai Yang</i>	515
Parking Demand Models <i>S.C. Wong, Zhi-Chun Li, and William H.K. Lam</i>	519
Pavement Management Systems <i>Senthilmurugan Thyagarajan</i>	524
Residential Location Choice Models <i>Shlomo Bekhor and Sigal Kaplan</i>	531
Transport Demand Management <i>Feiyang Zhang and Becky P.Y. Loo</i>	537
Traffic Flow Analysis <i>H. Michael Zhang and Jia Li</i>	544
Autonomous Vehicles and Transportation Modeling <i>Annesha Enam, Felipe de Souza, Omer Verbas, Monique Stinson, and Joshua Auld</i>	557
Ride-Hailing and Travel Demand Implications <i>Felipe F. Dias</i>	564
Transportation Modeling and Planning Software <i>Joel Freedman</i>	569
Transportation Statistics and Databases <i>Taha Hossein Rashidi</i>	574
Travel Surveys <i>Stacey G. Bricka</i>	587
Travel Demand Forecasting: Where Are We and What Are the Emerging Issues <i>Thomas F. Rossi</i>	590
Travel Model Calibration and Validation <i>Ram M. Pendyala</i>	596
Trip Chaining Analysis <i>Cynthia Chen and Yusak Susilo</i>	606
Vehicle Ownership Models <i>Dr. Sören Groth and Prof. Dr. Dirk Wittowsky</i>	612

Bicycle Sharing/Bikesharing <i>Catherine Morency and Jean-Simon Bourdeau</i>	617
Carsharing <i>Shiva Habibi and Frances Sprei</i>	623
Urban Recreational Travel <i>Long Cheng and Frank Witlox</i>	629
Place Perception and Travel Behavior <i>Kathleen Deutsch-Burgner and Konstadinos G. Goulias</i>	635

VOLUME 5

Transport Modes <i>Edoardo Marcucci</i>	1
Infrastructure Transport Investments, Economic Growth and Regional Convergence <i>Xavier Fageda and Cecilia Olivieri</i>	2
Transport Modes and an Aging Society <i>Charles B.A. Musselwhite and Theresa Scott</i>	6
Sustainable Mobility Paths <i>Erling Holden and Geoffrey Gilpin</i>	13
Vehicles that Drive Themselves: What to Expect with Autonomous Vehicles <i>Michele D. Simoni and Kara M. Kockelman</i>	19
Transport Modes and Tourism <i>Ila Maltese and Luca Zamparini</i>	26
Transport Modes and Accessibility <i>Bert van Wee</i>	32
Transport Modes and Globalization <i>Jean-Paul Rodrigue</i>	38
Transport Modes and Cities <i>Erick Guerra and Gilles Duranton</i>	45
Transport Modes and Remote Areas	51
Modeling Mode Choice in Freight Transport <i>Lóránt Tavasszy and Gerard de Jongb</i>	57
Travel Mode Choice as Reasoned Action <i>Sebastian Bamberg, Icek Ajzen, and Peter Schmidt</i>	63
Energy Consumption of Transport Modes <i>Zissis Samaras and Ilias Vouitsis</i>	71
Transport Modes and People With Limited Mobility <i>Roger L Mackett</i>	85
Transport Modes and Commuters <i>Colin G. Pooley</i>	92
Shopping and Transport Modes <i>Antonio Comi</i>	98
Transport Modes and Health <i>Jennifer S. Mindell and Sandra Mandic</i>	106

Multimodality in Transportation <i>Sören Groth and Tobias Kuhnimhof</i>	118
Transport Modes and Disasters <i>Brian Wolshon</i>	127
Big Data for Public Transport Planning <i>Jan-Dirk Schmöcker</i>	134
Active Transport: Heterogeneous Street Users Serving Movement and Place Functions <i>Regine Gerike, Stefan Hubrich, Caroline Koszowski, Bettina Schröter, and Rico Wittwer</i>	140
Electric Vehicles <i>Christine Eisenmann, Daniel Görges, and Thomas Franke</i>	147
Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service <i>Susan Shaheen, PhD and Adam Cohen</i>	155
Adoption of new travel information platforms <i>Sigal Kaplan</i>	160
ICT and Transport Modes <i>Galit Cohen-Blankshtain</i>	165
Mode Choice and Life Events <i>Joachim Scheiner</i>	171
Introduction to Air Transport <i>Milan Janić</i>	178
The History of Air Transportation <i>Richard P. Hallion</i>	192
The Geography of Air Transport <i>Lucy Budd and Stephen Ison</i>	198
The Future of Air Transport <i>Rico Merkert and James Bushell</i>	203
Air Transport and Its Territorial Implications <i>Lanfranco Senn</i>	208
Next Generation Travel: Young Adults' Travel Patterns <i>Tobias Kuhnimhof and Scott Le Vine</i>	215
Airport Network Planning and Its Integration with the HSR System <i>Francesca Pagliara, Juan Carlos Martín, and Concepción Román</i>	222
Airport Management <i>Peter Forsyth and Hans-Martin Niemeier</i>	229
Airport Regulation <i>Achim I. Czerny</i>	234
Air Route Planning and Development <i>Renan Peres de Oliveira and Gui Lohmann</i>	241
Airline Management <i>Sveinn Vidar Gudmundsson</i>	249
Air Cargo <i>Volodymyr Bilotkach</i>	258

Airline Regulation <i>Andrew R. Goetz</i>	263
Air Vehicles Classification <i>Vincenzo Torre</i>	269
Aircraft Manufacturing <i>Antonio Sollo</i>	290
A Geography of Road Transport in Cities <i>Cristian Domarchi and Juan de Dios Ortúzar</i>	300
The Future of Road Transport <i>Preston L. Schiller</i>	306
Road Transportation and Territorial Scale <i>Ana M. Condeço-Melhorado</i>	315
Road Modes: Walking <i>Kevin Manaugh, PhD Associate Professor</i>	320
Bus Public Transport Planning and Operations <i>Ehab Diab and Ahmed El-Geneidy</i>	326
Street Design for Active Travel <i>Bruce Appleyard</i>	333
Transit Planning and Management <i>Zakhary Mallett and Marlon G Boarnet</i>	349
Road Infrastructure: Planning, Impact and Management <i>Jos Arts, Wim Leendertse, and Taede Tillema</i>	360
Road Transport Planning at the Urban Scale <i>David A. King and Kevin J. Krizek</i>	373
Road Traffic Regulation: Road Pricing and Environmental Quality <i>Marco Percoco</i>	378
Car Ownership and Car Use: A Psychological Perspective <i>J.L. Veldstra, A.B. Ünal, E.M. Steg</i>	384
Road Transport: E-Scooters <i>Gysele Lima Ricci and Klaus Bogenberger</i>	391
Railway Station and Network Planning <i>Ingo Arne Hansen</i>	399
Introduction to Rail Transport <i>Chris Nash and Tony Fowkes</i>	406
The History of Rail Transport <i>Carlo Ciccarelli, Andrea Giuntini, and Peter Groote</i>	413
The Geography of Rail Transport <i>Frédéric Dobruszkes and Amparo Moyano</i>	427
Rail Transport and Territorial Scale <i>Prof. Andrés Monzón and Dr. Elena López</i>	437
Railway Management <i>Vassilios A. Profillidis</i>	444
Railway Terminal Regulation <i>Nacima Baron</i>	454

Service Network Design for Freight Railroads <i>Teodor Gabriel Crainic</i>	464
Subway Systems <i>Guillaume Monchambert, Daniel Hörcher, Alejandro Tirachini, and Nicolas Coulombel</i>	471
Railway Company Management <i>Vilius Nikitinas, Skaistė Miliauskaitė</i>	479
Regulation of Rail Infrastructure and Services <i>Javier Campos</i>	485
Rail Vehicle Classification <i>Christos Pyrgidis and Alexandros Dolianitis</i>	490
Introduction to Maritime Shipping <i>Christa Sys and Thierry Vanelslander</i>	508
The Geography of Maritime Transport <i>César Ducruet and Justin Berli</i>	517
The Future of Maritime Transport <i>Harilaos N. Psaraftis</i>	535
Maritime Transport and Territorial Scale <i>Brian Slack</i>	540
Containerization and the Port Industry <i>Hercules Haralambides</i>	545
Port Management <i>Lourdes Trujillo, Daniel Castillo Hidalgo, and Manuel Herrera</i>	557
Economic and Environmental Regulation in the Port Sector <i>Beatriz Tovar and Alan Wall</i>	563
Maritime Route Planning <i>Johan Woxenius</i>	570
Inland Waterway Transport and Inland Ports: An Overview of Synchromodal Concepts, Drivers, and Success Cases in the IWW Sector <i>Behzad Behdani, Bart Wiegmans, and Yun Fan</i>	577
Maritime Company Passenger Management/Liner Industry <i>Claudio Ferrari and Alessio Tei</i>	587
Cruise Industry <i>Athanasios A. Pallis and Aimilia A. Papachristou</i>	593
International Maritime Regulation: Closing the Gaps Between Successful Achievements and Persistent Insufficiencies <i>Laurent Fedi</i>	600
Ship Classification <i>Gareth C. Burton and Mimosa T. Miller</i>	607
The Shipbuilding Industry and its Interactions With Shipping <i>Paul William Stott</i>	617
Methods for Designing Public Transport Networks <i>Zain Ul Abedin and Avishai (Avi) Ceder</i>	625
Space Transportation <i>Mark Hempsell</i>	638

Pipelines	646
<i>Franco Cotana and Mattia Manni</i>	
Women and Transport Modes	656
<i>Priya Uteng, PhD and Yusak Susilo, PhD</i>	
Transport Modes and Big Data	665
<i>Hannah D Budnitz, Emmanouil Tranos, and Lee Chapman</i>	
Transport Modes and Inequalities	671
<i>Caroline Mullen</i>	
Railway Traffic Management	678
<i>Francesco Corman</i>	
Indoor Transportation	684
<i>Lutfi Al-Sharif</i>	
Bicycle as a Transportation Mode	697
<i>Raktim Mitra and Paul M. Hess</i>	
Urban Air Mobility: Opportunities and Obstacles	702
<i>Adam Cohen and Susan Shaheen, PhD</i>	
Transport Modes and Sustainability	710
<i>Long Cheng, Jonas De Vos, and Frank Witlox</i>	

VOLUME 6

Introduction to Transport Policy and Planning	1
<i>Maria Attard</i>	
Workplace Parking Levy	2
<i>Stephen Ison and Lucy Budd</i>	
Air Transport	7
<i>Lucy Budd and Stephen Ison</i>	
Mobility as a Service	12
<i>Milo N. Mladenović</i>	
Bicycle Sharing	19
<i>Cyrille Médard de Chardon</i>	
Light Rail	31
<i>Fiona Ferbrache</i>	
Planning Tourism Travel	39
<i>Luca Zamparini</i>	
Transport Planning and Management and its Implications in Chinese Cities	44
<i>Mengqiu Cao</i>	
Road Safety	51
<i>George Yannis and Eleonora Papadimitriou</i>	
Mobility Planning and Policies for Older People	59
<i>Charles B A Musselwhite</i>	
Electric Mobility	64
<i>Graham Parkhurst</i>	
Transport Policy and Governance	73
<i>Lisa Hansson</i>	

Transferability of Urban Policy Measures <i>Paul Martin Timms</i>	77
Accessibility Tools for Transport Policy and Planning <i>Benjamin Büttner</i>	83
Demand Responsive Transport <i>Marcus Enoch</i>	87
Transport and Climate Change <i>Robin Hickman and Christine Hannigan</i>	94
Taxicabs and Microtransit <i>David A. King</i>	101
Land-Use and Transport Planning <i>Luis A. Guzman</i>	107
Parking <i>Stephen Ison and Lucy Budd</i>	113
Transport Planning in the Global South <i>Daniel Oviedo and Mariajose Nieto-Combariza</i>	118
Planning for Children's Independent Mobility <i>E. Owen D. Waygood and Raktim Mitra</i>	125
The Politics of Mobility Policy <i>Geoff Vigar</i>	131
Technology Enabled Data for Sustainable Transport Policy <i>Susan M. Grant-Muller, Mahmoud Abdelrazek, Hannah Budnitz, Caitlin D. Cottrill, Fiona Crawford, Charisma F. Choudhury, Teddy Cunningham, Gillian Harrison, Frances C. Hodgson, Jinhyun Hong, Adam Martin, Oliver O'Brien, Claire Papaix, and Panagiotis Tsoleridis</i>	135
Car Sharing <i>Cyriac George and Tanu Priya Uteng</i>	142
Toward a More Holistic Understanding of Mega Transport Project (MTP) Success <i>John Ward</i>	147
Equity Considerations in Transport Planning <i>Karel Martens</i>	154
Planning for Rail Transport <i>Simon P. Blainey</i>	161
Connected and Autonomous Vehicles: Priorities for Policy and Planning <i>Dr. Alexandros Nikitas</i>	167
Gendered Mobility <i>Sheila Mitra-Sarkar</i>	173
Modeling and Simulation for Transport Planning <i>Michela Le Pira, Giuseppe Inturri, and Matteo Ignaccolo</i>	184
Externalities and External Costs in Transport Planning <i>Silvio Nocera</i>	191
Planning for Public Transport with Automated Vehicles <i>Gonçalo Homem de Almeida Rodriguez Correia</i>	198
Urban Congestion Charging in Transport Planning Practice <i>Ida Kristoffersson and Maria Börjesson</i>	206

Energy and Transport Planning <i>Debbie Hopkins and Christian Brand</i>	214
Customer Satisfaction as a Measure of Service Quality in Public Transport Planning <i>Laura Eboli and Gabriella Mazzulla</i>	220
Car Sharing and the Impact on New Car Registration <i>Mario Intini and Marco Percoco</i>	225
Evaluation Methods in Transport Policy and Planning <i>Niek Mouter</i>	230
Transport and Air Quality Planning and Policy <i>Dr Fabio Galatioto</i>	236
Cycling Policies <i>Esther Anaya-Boig</i>	241
Community Severance <i>Paulo Anciaes and Jennifer S. Mindell</i>	246
Planning for Bus Priority <i>Claus H. Sørensen, Fredrik Pettersson, and Joel Hansson</i>	254
High-Speed Rail and the City <i>Marie Delaplace</i>	261
Public Engagement in Transport Planning <i>Miriam Ricci</i>	266
Long-Distance Travel <i>Giulio Mattioli and Muhammad Adeel</i>	272
Urban Freight Policy <i>Laetitia Dablanc</i>	278
Regional Transport Planning <i>Chia-Lin Chen</i>	286
Transport Project Financing <i>Romeo Danielis and Lucia Rotaris</i>	292
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	298
Transitions and Disruptive Technologies in Transport Planning <i>Kate Pangbourne and Maria Attard</i>	309
Community Transport: Filling the Gaps for Those in Need of Mobility <i>Ian Shergold</i>	314
Fundamental Emerging Concepts and Trends for Environmental Friendly Urban Goods Distribution Systems <i>Sandra Melo</i>	320
Plateau Car <i>David Metz</i>	324
Emerging Trends in Transport Demand Modeling in the Transition Toward Shared Mobility and Autonomy <i>Patrizia Franco</i>	331
Policy and Planning for Walkability <i>Carlos Cañas Sanz and Maria Attard</i>	340

Public Transport Subsidy and Regulation <i>Jonathan Cowie</i>	349
Urban Regeneration and Transportation Planning <i>Thomas Vanoutrive</i>	356
Social and Distributional Impact Assessment in Transport Policy <i>Laura Walker and Angela Curl</i>	361
Mobility Planning for Healthy Cities <i>Ersilia Verlinghieri</i>	368
Transport Demand Management <i>Begoña Guirao</i>	374
Planning and the Global Movement of Goods and Commodities <i>Christopher Clott and Chris Petrocelli</i>	380
Public Transport Network Planning <i>Corinne Mulley and John D. Nelson</i>	388
A Timely Perspective on Planning for Ageing Infrastructure <i>Anthony Perl</i>	395
The role of media in transport planning and the transport policy process <i>Özgül Ardiç and J.A. Annema</i>	400
Travel Plans <i>Stephen Potter and Marcus Enoch</i>	408
Home Deliveries and their Impact on Planning and Policy <i>Rosário Macário</i>	413
Planning for Safe and Secure Transport Infrastructure <i>Per Erik Gårder</i>	418
Sensors and Data Driven Approaches in Transport <i>Mohammad Sadrani and Constantinos Antoniou</i>	426
Planning Active Travel and School Transport <i>Fahimeh Khalaj, Dorina Pojani, and Sara Alidoust</i>	432
VOLUME 7	
Introduction to Transport Psychology <i>Carlo Prato</i>	1
From Self-reports to Auto-Tech-Detect (ATD)-based Self-reports in Traffic Research <i>Türker Özkan and Timo Lajunen</i>	2
Observational Field Studies in Traffic Psychology <i>Tova Rosenbloom and Hodaya Levy</i>	8
Driving Simulators <i>Karel A. Brookhuis</i>	14
Naturalistic Driving Studies: An Overview and International Perspective <i>Johnathon P. Ehsani, Joanne L. Harbluk, Jonas Bärghman, Ann Williamson, Jeffrey P. Michael, Raphael Grzebieta, Jake Olivier, Jan Eusebio, Judith Charlton, Sjaanie Koppel, Kristie Young, Mike Lenné, Narelle Haworth, Andry Rakotonirainy, Mohammed Elhenawy, Gregoire Larue, Teresa Senserrick, Jeremy Woolley, Mario Mongiardini, Christopher Stokes, Paul Boase, John Pearson, and Feng Guo</i>	20

A Detailed Approach to Qualitative Research Methods <i>Sonja Forward and Lena Levin</i>	39
Behavioral Change <i>Sonja Haustein</i>	46
Habitual Behavior <i>Carlo G. Prato</i>	54
Driving Behavior and Skills <i>Timo Lajunen and Türker Özkan</i>	59
The Multidimensional Driving Style Inventory <i>Orit Taubman - Ben-Ari</i>	65
Risk Perception in Transport: A Review of the State of the Art <i>Trond Nordfjrn, An-Magritt Kummeneje, Mohsen F. Zavareh, Milad Mehdizadeh, and Torbjørn Rundmo</i>	74
Social-Symbolic and Affective Aspects of Car Ownership and Use <i>Birgitta Gatersleben</i>	81
Pitfalls of Statistical Methods in Traffic Psychology <i>J.C.F. de Winter and D. Dodou</i>	87
Data Analysis: Structural Equation Models <i>Marco Diana</i>	96
Data Analysis: Integrated Choice and Latent Variable Models <i>Carlo G Prato</i>	102
Explaining Data Analysis Using Qualitative Methods <i>Lena Levin and Sonja Forward</i>	107
ITS for Transport Planning and Policies <i>Bruno Dalla Chiara</i>	113
Driver Aggression and Anger <i>Mark J.M. Sullman and Amanda N. Stephens</i>	121
Speeding: A "Tragedy of the Commons" Behavior <i>Bryan E. Porter, Thomas D. Berry, and Kristie L. Johnson</i>	130
Cycling as a Mode Choice: Motivational Psychology <i>Sigal Kaplan</i>	136
Motorcyclists <i>Narelle Haworth</i>	144
Drivers' Hazard Perception Skill <i>Mark S. Horswill and Andrew Hill</i>	151
Driver Education and Training for New Drivers: Moving beyond Current 'Wisdom' to New Directions <i>Teresa Senserrick, Oscar Oviedo-Trespalacios, David Rodwell, and Sherrie-Anne Kaye</i>	158
Road Safety Advertising: What We Currently Know and Where to From Here <i>Ioni Lewis, Barry Watson, Katherine M. White, and Sonali Nandavar</i>	165
Traffic Law Enforcement Theories and Models <i>Richard Tay</i>	171
Satisfaction with Travel and the Relationship to Well-Being <i>Tommy Gørling and Filip Fors Connolly</i>	177
	182

Electromobility: History, Definitions and an Overview of Psychological Research on a Sustainable Mobility System <i>Josef F. Krems and Isabel KreiÃŸig</i>	
Sharing: Attitudes and Perceptions <i>Yi Wen and Christopher R Cherry</i>	187
Women's Travel Patterns, Attitudes, and Constraints Around the World <i>Sandra Rosenbloom</i>	193
Driver Stress and Driving Performance <i>Lisa Dorn</i>	203
Introduction to Sustainability and Health in Transportation <i>Roger Vickerman</i>	225
Sustainable Development Goals and Health <i>Rosa Surinach</i>	226
Human Ecology <i>Roderick J. Lawrence</i>	234
Car- Free Cities <i>Haneen Khreis and Mark J. Nieuwenhuijsen</i>	240
Superblocks Base of a New Model of Mobility and Public Space. Barcelona as an Example <i>Salvador Rueda Palenzuela</i>	249
Resilience of Transport Systems <i>ErikJenelius and Lars-GöranMattsson</i>	258
Impact of Shipping to Atmospheric Pollutants: State-of-the-Art and Perspectives <i>Daniele Contini and Eva Merico</i>	268
Noise Pollution From Transport <i>Marianna Jacyna, Emilian Szczepański, Konrad Lewczuk, Mariusz Izdebski, Ilona Jacyna-Gotda, Michał, Kłodawski, Paweł, Gołda, Piotr Goł, and Ebiowski</i>	277
Visual Impacts From Transport <i>Paulo Rui Anciaes</i>	285
Light Pollution <i>John D. Bullough</i>	292
Wildlife Crossings and Barriers <i>Scott D. Jackson</i>	297
Environmental Justice, Transport Justice, and Mobility Justice <i>Devajyoti Deku</i>	305
Transport Noise and Health <i>Elisabete F. Freitas, Emanuel A. Sousa, and Carlos C. Silva</i>	311
Climate Change and Health, Related to Transport <i>Ersilia Verlinghieri</i>	320
Urban Greenspace, Transportation, and Health <i>Payam Dadvand and Mark J. Nieuwenhuijsen</i>	327
Transport Access and Health <i>Alireza Ermagun</i>	335
Social Exclusion and Health, Related to Transport <i>Roger L. Mackett</i>	341

Burden of Disease Assessment <i>David Rojas-Rueda</i>	347
Achieving a Near-Zero CO <i>Lewis M. Fulton</i>	353
Disabled Travelers <i>Bryan Matthews</i>	359
Health Impacts of Connected and Autonomous Vehicles <i>Soheil Sohrabi</i>	364
Electric Vehicles and Health <i>Kanok Boriboonsomsin</i>	372
Shared Mobility Opportunities and their Computational Challenges for Improving Health-Related Quality of Life <i>Cristiano Martins Monteiro, Cláudia Aparecida Soares Machado, Adelaide Cassia Nardocci, Fernando Tobal Berssaneti, José Alberto Quintanilha, and Clodoveu Augusto Davis</i>	376
Bike Sharing and Health <i>David Rojas-Rueda and Mark J. Nieuwenhuijsen</i>	384
E-Bikes and Health <i>Aslak Fyhri and Hanne Beate Sundfør</i>	393
<i>Index</i>	399

Introduction to Transport Psychology

Carlo G. Prato, School of Civil Engineering, The University of Queensland, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Transport systems are made of infrastructure (e.g., roads, bridges, stations, ports) and vehicles (e.g., bicycles, cars, trains, vessels), which most likely explains why economics and engineering have played an essential role in their study since the discipline known as transport engineering was born. However, transport systems are really made of people who use the vehicles on the infrastructure to become pedestrians, cyclists, car drivers, truck drivers, bus passengers, etc. This explains why transport psychology has been flourishing to answer questions about the people who populate transport systems from a behavioral, a cognitive and a social perspective. This also explains why transport psychology has been working toward helping provide answers from the transport demand perspective, in particular in relation to the role of attitudes and perceptions as latent concepts, and the safety perspective, in particular in relation to the role of hazard perception and driving behavior. The wealth of knowledge in transport psychology has certainly been flourishing also because of a wealth of methods that have become increasingly accessible in terms of psychological theories, methodological advancements and data collection techniques.

The articles in this section present the foundations of the discipline from theoretical, methodological and application perspective. All the articles provide a unique contribution to discovering the most compelling theories, the most appropriate data collection methods, the most innovative data analysis techniques, and the most popular issues up for debate. Extensive space has been dedicated to the collection of data, from the design of qualitative methods to the use of self-reports in a variety of contexts, culminating then in the use of the latest technology in driving simulator and naturalistic driving studies. Alongside data collection discussions, issues with their analysis have then been disentangled, starting from proper guidelines for qualitative analysis and statistical methods in traffic psychology, and arriving to models that integrate advanced econometrics with latent variable models that capture attitudes, perceptions, social influence, behavioral control, and all the elements of the most popular psychological theories. The focus of the section shifts then to key theories related to the demand side of behavioral studies, focusing for example not only on the affective and social aspects of choices, and the risk side of behavioral studies, concentrating on risky behavior such as speeding and hazards as well as their trigger such as distraction and aggression. Last, the focus of the section moves towards key issues such as vulnerable road users like cyclists and motorcyclists, educational and campaign programs for promoting wisdom (especially in young drivers), law enforcement theories for increasing safety, and perspectives on future mobility and the sharing economy.

Several of the chapters have cross-cutting themes with articles in other sections, from the closely related Transport Safety and Security to Transport Modelling and Data Management, from Sustainability and Health Issues in Transportation to Transport Policy and Planning. Readers are invited to explore these sections as well to have an even broader picture of the issues and fully appreciate the multidisciplinary nature of transportation studies.

From Self-Reports to Auto-Tech-Detect (ATD)-Based Self-Reports in Traffic Research

Türker Özkan*, Timo Lajunen†, *Department of Psychology, Middle East Technical University (METU), Ankara, Turkey; †Department of Psychology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

© 2021 Elsevier Ltd. All rights reserved.

The Content and Situation of Self-Reports in Traffic Research	2
The Usage of Self-Reports in Traffic Research	2
Self-Reports in Driver Skills and Behavior	2
Some Paradoxes of Self-reports in Driver Behavior	3
Self-Reports in Self-Assessments	3
Self-Reports in Accidents, Near Accidents, and Mileage	3
Accidents	3
Near Misses	3
Exposure	4
Memory and Social Desirability	4
Impression Management and Self-Deception in Self-Reports	4
Coping with Socially Desirable Responding in Self-Reports of Driving	5
Future Directions and/or Corrections: Auto-Tech-Detect (ATD)-Based Self-Reports in Traffic	5
Conclusions	6
References	6
Further Reading	7

The Content and Situation of Self-Reports in Traffic Research

In recent decades, the popularity of self-reports in traffic research including the number of publications on self-reported traffic accidents (Kamaluddin et al., 2018) has increased drastically. Self-reports include a great variety of different methods including questionnaires and inventories, interviews, focus groups (e.g., Kua et al., 2007), and driving diaries (e.g., Joshi et al., 2001). As Lajunen and Özkan (2011) stated that common features in all these diverse self-report measures are that participants are aware that they are participating a study, they are asked to actively reply to more or less structured questions, and that their responses are taken “face valid,” that is, answers are scored and analyzed based on the responses and not, for example, according to response time or another behavioral or physiological measurement. In self-reports, the content of the responses in this way is assumed to reflect a respondent’s reality.

The Usage of Self-Reports in Traffic Research

Literature search shows that self-report methodology has been used for a great variety of research including attitudes, opinions, beliefs, emotion, cognitive processes, and behaviors—basically any aspect of driving. While acknowledging that self-reports can be the most appropriate tool or even the only tool for many research purposes (e.g., for measuring attitudes, beliefs, and opinions), self-reports have space to improve especially the self-reports of past accidents, near-misses, mileage, and road user/driver skills and behavior (Lajunen and Özkan, 2011).

Self-Reports in Driver Skills and Behavior

Although in literature self-reports have been used for measuring both driving skills (performance) and driving style (driver behavior), we can claim that self-report methodology fits better to studies of driver behavior than performance because of several reasons (Lajunen and Özkan, 2011). First, driver behavior refers to driving style, i.e. the everyday way of driving which driver prefers and drivers are mostly aware about. Drivers’ awareness of their driving skills can be assumed to be much lower, because the basic motor and perceptual processes are automatic and do not need attention (Lajunen and Özkan, 2011).

Second, driver behavior consists of behaviors, which are often included in Highway Code or unwritten norms. Because of this normative character of most of the driver behaviors, drivers actually know what the ideal normative behavior would be and can compare their behavior to those norms (e.g., try to drive just slightly above the speed limit). Opposite to driver behavior, it is usually difficult to determine the “skill norms” of driving and drivers are mostly unaware of their skill level. Because of these two issues, high awareness level, and the existence of an absolute norm, self-reports fit better to studies of driver behavior than driver performance (Lajunen and Özkan, 2011).

Some Paradoxes of Self-Reports in Driver Behavior

The paradox in self-reports of attention or memory mistakes is that most of the errors go unnoticed (Lajunen and Özkan, 2011). As Bjørnskau and Sagberg, (2005) accurately pointed out: "Unconscious errors may be hard to remember precisely because they are unconscious." We can assume that drivers (e.g., elderly drivers, drivers with slight dementia) prone to cognitive failures and forgetting while driving have also difficulties in remembering their mistakes when answering self-reports like Driver Behavior Questionnaire (DBQ). In addition, asking "violating drivers" to honestly report their violations is the second paradox in the DBQ and self-reports in general: we assume that people who like to violate social norms in traffic would respect the honesty required in self-reports. It is more likely that drivers who do not respect general rules and norms in traffic would also not be exactly honest when answering self-report surveys (Lajunen and Özkan, 2011).

Self-Reports in Self-Assessments

Drivers' views of their skills can be studied with self-reports but direct measurement of real driving skills with a self-report is almost impossible (Lajunen and Özkan, 2011). Driver Skill Inventory (DSI) (Lajunen and Summala, 1995) provides another example of measurement of driver skills with self-reports. However, DSI does not actually even pretend to measure skills but rather a driver's skill and safety orientation. Drivers are asked to evaluate themselves as drivers by indicating their strong and weak sides in driving. Hence, the external criterion (compared to "other drivers") or absolute criterion (frequency of errors) was replaced by internal comparison. Direct measurement of driver skills or errors is very difficult and possible unreliable with self-report instruments. However, self-reports can be used reliably to measure a driver's view or opinion about his or her skills. While remembering individual mistakes is difficult, drivers are usually having a general idea of themselves as drivers, which can be measured by self-report instruments (Lajunen and Özkan, 2011).

Self-Reports in Accidents, Near Accidents, and Mileage

While self-reports are not adequate measures of unconscious processes or norm violations which drivers would not like to report honestly, self-reports work well when recording driver attitudes, self-evaluated skills, and beliefs. Interestingly, most of the self-reported drivers' behaviors are evaluated in terms of how risky they are, that is, the potential in causing a serious accident. Hence, accident risk seems to be the ultimate criterion for driver behaviors and, therefore, either objective or self-reported number of past accidents is used as a criterion for safe driving (Lajunen & Özkan, 2011).

Accidents

In most studies concerning behavioral correlates of individual differences in road-traffic accident risk, a driver's accident history (number and/or the severity of accidents) has been used as a criterion for safety. This accident data is collected either from drivers' self-reports or from the official statistics. However, both are subject to systematic and random error and are, therefore, somewhat biased (Elander et al., 1993). The advantage of asking drivers to report accidents is that minor crashes can also be recorded. In addition, driver's self-reports are usually more detailed than official reports, since drivers can be asked very specific questions. However, comparisons between self-reported and state-recorded numbers of accidents have shown some under-reporting of accidents in self-reports (Lajunen and Özkan, 2011).

Self-reports of accidents are easily biased by intentional or unintentional misrepresentation (Elander et al., 1993). The latter source of bias can be caused either by a different definition of reportable accidents between drivers or by simple forgetting. Chapman and Underwood (2000) concluded that "there is reason to suspect that even severe accidents may be routinely forgotten by normal drivers over periods as short as a year." They also suggested that minor accidents are forgotten at much higher rates. While self-reports of accidents seem to be more or less problematic and the degree of accuracy seems to depend on various factors (e.g., time period, age of the participants, way of conducting the survey), official accident records from the police, hospitals, or insurance companies seem to have many shortcomings too. Forgetting, various definitions of accidents or deliberate under-reporting typical to self-report do not distort the official accident records, but official records have some other limitations. First, the police or insurance companies' accident records do not include minor damages, and secondly, some driver groups, like older drivers, can be over-represented in these records for reasons not related to their risk of accident (Elander et al., 1993).

In addition to the age and gender bias as well as not including minor accidents, official accident records are not always available because of data protection acts and because the period of accidents recorded might be limited. Also, the type and role of the parties (e.g., culpability) in the accident are rarely available. Studies about self-reports of accidents and state records show that both recording methods have considerable shortcomings and strengths. As a conclusion, it seems that the best self-report method for recording accidents is a retrospective self-report of maximum of 5 years. Shorter the reporting period is, smaller the underreporting bias due to forgetting should be. If possible, self-reports should be complemented with state accident records (Lajunen and Özkan, 2011).

Near Misses

Since the forgetting rate gets higher when the time period for recall gets longer and when minor accidents are asked, we could suppose that self-reports of serious accidents, say, in the previous year would be the best estimate. However, accidents and especially serious accidents are (luckily) infrequent. This means that samples in accident studies should be enormous, which in turn, is

impossible for most of the studies. One approach to this problem is to record near-accidents, which are extremely frequent events (Chapman and Underwood, 2000) and may lead to a crash in less optimal conditions. By definition, near-accidents can be assumed as a good candidate for a general measure of safety.

Chapman and Underwood's (2000) study shows that while near-accidents are very frequent and, thus, easier to use in statistical analyses than actual accidents, the high forgetting rate makes them an unreliable measure of risk. Near-accidents can best be recorded by using driving diary technique, in which drivers keep a log about their near accidents for a couple of weeks. Since near-accidents are recorded as soon as possible after they have happened, the forgetting rate stays low while the number of near-accidents is much higher.

Exposure

Mileage is one of the main factors measured in studies about driver behavior. Mileage as total life-time mileage or annual mileage and type of exposure are extremely important variables, because mileage (and exposure) largely determines a driver's likelihood of being involved in an accident. Moreover, particular driver characteristics may lead only to particular forms of driving behavior which may, thereby, affect liability only for accidents of certain types (Elander et al., 1993).

The effects of mileage on driving style and accidents depend largely on the type of exposure. Type of roads usually driven (motorways versus country roads, traffic density), time of year (winter versus summer) and day, time of the day spent on the road (daylight versus night), and usual purpose of driving (work-related versus free-time) all affect the likelihood of accident involvement and driving style. Special effort should therefore be directed to measure these factors related to exposure in studies on driver style and accident involvement, since exposure may be related to psychological variables under investigation. Ignoring the role of driving experience and exposure can increase error variance and reduce the true associations between psychological variables and accident frequency (Elander et al., 1993).

In driver behavior studies, mileage is traditionally measured with either driving diaries or questionnaires, which both are based on self-reports. As self-reports, self-reports of mileage share in some degree the same problems and bias sources as self-reports of accidents. The accuracy of self-reports has been questioned in several studies, and in-vehicle recording devices have been suggested as an alternative to self-reports (Blanchard et al., 2009). While in-car recordings certainly provide a more unbiased and accurate measure of exposure than self-reports, their usefulness is still very limited to certain types of studies. For example, large-scale population studies do not allow direct recordings of driving frequency and amount. Moreover, many studies require anonymous participation and, thus, direct measurements cannot be considered.

While mileage self-reports share some of the problems with self-reports of accidents, for example, drivers forget to report some trips, we can assume that self-reports of mileage are less biased than self-reports of accidents because of several reasons. An accident is a single event but the estimation of mileage is continuous, which makes it naturally easier to evaluate. While accidents can be embarrassing or even traumatic events which one would like to forget, mileage is a rather neutral issue. Moreover, there are more techniques for reporting exposure than accident involvement: drivers can be asked to report their lifetime mileage, mileage in a certain period of time (year, months, week, day), frequency of driving in days or number of trips in addition to the type of exposure (e.g., in-city traffic, highways, night driving). Drivers can also be related to their life-time mileage to other more objective measures like ownership and use frequency of their car (Lajunen and Özkan, 2011).

Memory and Social Desirability

According to Chapman's and Underwood's (2000) study about forgetting near-accidents, about 80% of incidents were no longer reported after a delay of up to two weeks (Chapman and Underwood, 2000). Surprisingly, experimental studies about the effect of socially desirable responding (self-deception or lying) show that the social desirability bias is not a considerable problem in the Driver Behavior Questionnaire responses (Lajunen and Summala, 2003; Sullman and Taylor, 2010). Still, future users of the self-report measures should bear in mind that forgetting and socially desirable responding are potential sources of bias in driver behavior (Lajunen and Özkan, 2011). These biases might interact with sample characteristics, which means that forgetting and socially desirable responding is more pronounced among certain driver groups. The best way to reduce the bias related to forgetting and socially desirable responding is to use the DBQ always as an anonymous research instrument and not for evaluating individual drivers (Lajunen and Özkan, 2011).

Impression Management and Self-Deception in Self-Reports

A large number of studies involving different measures of socially desirable responding have shown that it consists of two distinct factors which can be called "impression management" and "self-deception," or in other words "other-deception" and "self-deception" (e.g., Paulhus, 1984). In this distinction, impression management refers to the deliberate tendency to give favorable self-descriptions to others and comes, therefore, close to lying and falsification (e.g., Paulhus, 1984). The deliberate nature of impression management is also manifested in the finding that impression management increases in public settings and seems to be a situation-dependent phenomenon (Paulhus, 1984). It has been recommended that impression management should be controlled when it is conceptually independent of the trait being assessed but still contributes to the self-reported scores of that trait (Paulhus, 1984; Paulhus and Reid, 1991).

The self-deception factor can be characterized as a positively biased but subjectively honest self-description. In contrast to impression management, self-deception is an unintentional tendency and not influenced by the anonymity versus public context

manipulation (Paulhus, 1984). Unlike impression management, self-deception has been reported to be intrinsically linked to positive personality constructs such as psychological adjustment like high self-esteem (Paulhus and Reid, 1991). These findings suggest that self-deception could be used for the purposes of gaining pleasure (ego enhancement) as well as avoiding pain (denial) and, therefore, provides aid for coping with negative life events and threatening information (Paulhus, 1984; Paulhus and Reid, 1991). Self-deception appears to be more of a personality construct than simply a distorting factor. Paulhus (Paulhus, 1984; Paulhus and Reid, 1991) suggests that self-deception should not be controlled if it is an intrinsic aspect of the personality construct concerned. It should be noted, however, that the bias caused by self-deception may not be constant over situations. Some situations can be more threatening than others and, therefore, elicit a stronger need for self-deception.

Coping with Socially Desirable Responding in Self-Reports of Driving

One possibility is to give up using self-reports of driving and accidents and to rely only on observed behavioral data and official accident records as some researchers seem to suggest (AfWahlberg et al., 2009; AfWahlberg, 2010). However, the objective measures have serious methodological limitations too as studies using an instrumented vehicle, simulator or laboratory tests show. Official accident records suffer their own sources of bias too (Lajunen and Özkan, 2011). The other possibility is to let them the use of self-reports of driving to go unchallenged and accept the small social desirability bias as innate characteristics of self-reports. Self-report research methodology offers various ways of coping with socially desirable responding. First, emphasis on anonymity and confidentiality in instructions and procedures (e.g., sealed envelopes, large group data collection) all reduce the effect of socially desirable responding. Second, scales for socially desirable responding like DSDS can be added to studies and their effect statistically controlled. Scales for controlling impression management, self-deception, careless answering style etc. can be easily be designed and embedded to such instruments as the Driver Behavior Questionnaire.

Future Directions and/or Corrections: Auto-Tech-Detect (ATD)-based Self-Reports in Traffic

It is well-known that cars already consist of different types of in-vehicle technologies that can be used for collecting driver's information while driving. Besides, Global Positioning System (GPS) devices including smartphones provide measures of driving exposure or travel patterns. It is possible that self-report bias in mileage estimate can be best solved by not using self-reports but rather relying on, for example, GPS based in-car recordings. Since this is not possible in many survey studies, the best remedy for reducing self-report bias is to use as many estimates of mileage as possible. For example, total annual or monthly mileage, frequency of driving (number of trips per time unit) or proportion of long trips can be asked and the final estimates can be based on several self-report indicators (Lajunen and Özkan, 2011). Such a combination of ATD based and pure self-reports allow researchers and practitioners to more easily compare and/or combine self-reported and ATD-based travel-related behavior. Furthermore, autonomous vehicles are on the way of travelling on roads at different levels of autonomy collecting, monitoring, controlling data, and behavior.

We can claim that, thus, diverse self-report measures tend to change its focus (i.e., data collection method, sample size, age of respondents, user type of respondents, recall period, follow-up precision and data quality, and ways of data collection—recruitment vs anonymity) while ATD based self-reports supply new advantages such as comparability with other data, reduction of desirability bias, precision at response rate, road user characteristics and outcome measures (e.g., Kamaluddin et al., 2018). As a result, participants are implicitly and/or technologically aware that they are not only participating in a study but also a participative complementary element of measures of driving. They are implicitly and/or technologically “asked” to actively reply to more or less structured questions or rather conditions, and that their responses are taken “technologically pin valid,” that is, answers are scored and analyzed based not only on the responses but also, for example, according to response time or other behavioral or physiological measurement. In ATD based self-reports, the content of the responses in this way is not assumed but counted as to reflect a respondent's reality.

It can also be claimed that some paradoxes based on social desirability, attention capacity or memory losses, and others of self-reports in driver skills and behavior may be solved by the support of ATD based self-driving capacities. Besides, direct measurement of driver skills or errors can be possible with reliable ATD based self-report instruments and general or holistic view of driver skills can be calculated by developing new indexes including both ATD and pure self-evaluations on attitudes, beliefs norms, and capacities.

In other words, ATD based data will strengthen “what” parts of behavior and skills and even accidents because we will have more precise measures and data sets. Ironically, “why” parts of behavior and skills and even accidents may go more center of traffic studies. Thus, “pure” self-report measures may be even more critical into some extent. Norms, time, and context including traffic culture/climate (e.g., Özkan and Lajunen, 2015) seem to be embedded or anchored to behavior and skills or even accidents to some extent.

Özkan et al. (2019), for instance, measured self-reported seatbelt use, observed use among those who self-reported was compared with observed seatbelt use in the two city populations. They found that an increase in population-based observed seatbelt use rate among drivers both in Afyon and Ankara was also reflected in the increase in self-reported use rate, and the observed rate among those who self-reported. Over reporting in self-reported use rate decreased over time, however, the statistically significant difference with population-based observed rate existed over time. In other words, the self-reported seat belt use among individuals might be anchored to the general usage rates in the population.

Özkan et al. (2006) interpreted changes in DBQ factor structure by the classification of alpha, beta and gamma (Golembiewski et al., 1976). ATD-based data and/or responses and “pure” self-reports may need to be evaluated separately or together in this direction in this new phase of traffic research. Golembiewski et al. (1976) defined alpha change as a change, which “involves a variation in the level of some existential state, given a constantly calibrated measuring instrument related to a constant conceptual domain.” Alpha change represents an unbiased measure of variation between Time 1 and Time 2 in the observed phenomenon (e.g., less competitiveness between T1 and T2). According to Golembiewski et al. (1976), beta change is characterized by “a variation in the level of some existential state, complicated by the fact that some intervals of the continuum associated with a constant conceptual domain have been recalibrated.” It refers to an observed variation in some state where the apparent change is due to an instrument that has been recalibrated by the participant between assessments (e.g., adaptation to risk). By Golembiewski et al. (1976) definition, gamma change “involves a redefinition or re-conceptualization of some domain, a major change in the perspective or frame of reference within which phenomena are perceived and classified, in what is taken to be relevant in some slice of reality.” Golembiewski et al. (1976) showed that gamma change might be indicated through a comparison of factor structures over time. In addition, both alpha and beta change should be taken into account when comparing factor structures. It is possible that some drastic events (e.g., involvement in a traffic accident) may change a driver’s view (“pure” self-reports) of driving or driving itself (ATD-based responses) and traffic safety in general. Interdependency between some variables, Characteristics, road user types, etc. should also be examined as possible reasons and types of change in reporting or responding in traffic.

Of course, these possibilities can open a new phase about reliability, content, construct, and criterion validity or predictive validity issues of self-report measures in traffic research. Sure, this may require researchers and practitioners to have more comprehensive models and/or theories of a road user/human or ATD based data behavior. The problem may not be the lack of reliable and valid data rather rich in data or the lack of powerful and real models of humans on roads. As stated by Bailey and Wundersitz (2019), carefully defining terms (e.g., crashes); assuring confidentiality of responses; identifying individuals who erroneously believe inappropriate driving behaviors such as speeding are really quite safe; considering the use of pre-determined response categories (e.g., 1–2 times a week); and considering having a control group not exposed to a campaign or other intervention can be critical for reducing noise in data collection and modeling phases. Finally, data protection laws and/or regulations may also draw the lines of a new phase of the combination of “pure” self-reports and ATD-based ones in traffic research.

Conclusions

When evaluating the role and evolution of self-reports in traffic research, we still fully agree with Reason et al. (1990) that the tool (i.e., like DBQ) is a powerful means to measure behaviors that are “too private to be detected by direct observation” but that at the same time “DBQ responses are several stages removed from the actuality of what goes on behind the wheel.” The same applies to self-reports in general: they are able to reveal information which is not available with any other measurement methods. The gap mentioned by reason between the reality and the picture given by self-reports may not be possible to erase but at least it can be considerably shortened with adequate use of self-report methodology and the advance support of ATD-based data and responses.

References

- Af Wahlberg, A.E., 2010. Social desirability effects in driver behavior inventories. *J. Saf. Res.* 41 (2), 99–106.
- Af Wahlberg, A.E., Dorn, L., Kline, T., 2009. The effect of social desirability on self reported and recorded road traffic accidents. *Transport. Res. Part F: Traffic Psychol. Behav.* 13 (2), 106–114.
- Bailey, T.J., Wundersitz, L.N., 2019. The relationship between self-reported and actual driving-related behaviours: a literature review. *Case Rep. Ser.*. <https://www.researchgate.net/publication/334364350>.
- Björnskaug, T., Sagberg, F., 2005. What do novice drivers learn during the first months of driving? Improved handling skills or improved road user interaction?. In: Underwood, G. (Ed.), *Traffic & Transportation Psychology. Theory and Application*. Elsevier, Oxford, UK, pp. 129–140.
- Blanchard, R.A., Myers, A.M., Porter, M.M., 2009. Correspondence between self-reported and objective measures of driving exposure and patterns in older drivers. *Accid. Anal. Prev.* 42 (2), 523–529.
- Chapman, P., Underwood, G., 2000. Forgetting near-accidents: the roles of severity culpability and experience in the poor recall of dangerous driving situations. *Appl. Cogn. Psychol.* 14, 31–44.
- Elander, J., West, R., French, D., 1993. Behavioral correlates of individual differences in road-traffic crash risk: an examination of methods and findings. *Psychol. Bull.* 113, 279–294.
- Golembiewski, R.T., Billingsley, K., Yeager, S., 1976. Measuring change and persistence in human affairs: types of change generated by OD designs. *J. Appl. Behav. Sci.* 133–157.
- Joshi, M.S., Senior, V., Smith, G.P., 2001. A diary study of the risk perceptions of road users. *Health Risk Soc.* 3 (3), 261–279.
- Kamaluddin, N.A., Andersen, C.S., Larsen, M.K., Meltøfte, K.R., Varhelyi, A., 2018. Self-reporting traffic crashes – a systematic literature review. *Eur. Transport Res. Rev.* 10, 26.
- Kua, A., Korner-Bitensky, N., Desrosiers, J., 2007. Older individuals’ perceptions regarding driving: focus group findings. *Phys. Occup. Ther. Geriatr.* 25 (4), 21–40.
- Lajunen, T., Özkan, T., 2011. Self-report instrument and methods. In: Porter, B. (Ed.), *Handbook of Traffic Psychology*. Elsevier Ltd., London (Chapter 4).
- Lajunen, T., Summala, H., 1995. Driving experience, personality, and skill and safety-motive dimensions in drivers’ self-assessments. *Personal. Indiv. Differ.* 19, 307–318.
- Lajunen, T., Summala, H., 2003. Can we trust self-reports of driving? Effects of impression management on driver behaviour questionnaire responses. *Transport. Res. Part F: Traffic Psychol. Behav.* 6, 97–107.
- Özkan, T., Lajunen, T., 2015. A general traffic (safety) culture system (G-TraSaCu-S). *TraSaCu Project, European Commission, RISE Programme*, <https://www.researchgate.net/publication/319241963>
- Özkan, T., Lajunen, T., Summala, H., 2006. Driver behaviour questionnaire: a follow-up study. *Accid. Anal. Prev.* 38, 386–395.
- Özkan, T., Gupta, S., Hoe, C., Lajunen, T., Hyder, A., 2019. The stability of self-reported and observed seatbelt use: an anchor case study from Turkey. *Technical Report* <https://www.researchgate.net/publication/331895871>.

- Paulhus, D.L., 1984. Two-component models of socially desirable responding. *J. Person. Soc. Psychol.* 46 (3), 598–609.
- Paulhus, D.L., Reid, D.B., 1991. Enhancement and Denial in socially desirable responding. *J. Personal. Soc. Psychol.* 60 (2), 307–317.
- Reason, J., Manstead, A., Stephen, S., Baxter, J., Campbell, K., 1990. Errors and violations on the roads: a real distinction? *Ergonomics* 33, 1315–1332.
- Sullman, M.J.M., Taylor, J.E., 2010. Social desirability and self-reported driving behaviours: should we be worried? *Transport. Res. Part F: Traffic Psychol. Behav.* 13, 215–221.

Further Reading

- Evans, L., 2004. *Traffic Safety*. Science Serving Society.
- Porter, B.E., 2011. *Handbook of Traffic Psychology*. Academic Press.
- Reason, J.T., 1990. *Human Error*. Cambridge University Press, New York.
- Shinar, D., 2017. *Traffic Safety and Human Behavior*. Emerald Group Publishing.

Observational Field Studies in Traffic Psychology

Tova Rosenbloom*, **Hodaya Levy†**, *Research Institute of Human Factors in Road Safety, Management Department, Bar Ilan University, Tel Aviv, Israel; †Department of Cognitive Sciences, Hebrew University of Jerusalem, Jerusalem, Israel

© 2021 Elsevier Ltd. All rights reserved.

Introduction	8
Studies Based on Existing Databases	8
Controlled Observations	9
Uncontrolled Observations Field Experiments (Field Studies)	9
Electronic Observations	11
Integration of Methods in Observational Field Studies	11
Conclusions	12
References	12

Introduction

Research in traffic psychology relates to investigation of various human behaviors of road users such as drivers, passengers, pedestrians, bicyclists, motor bicyclists and others. This chapter is aimed to review studies which are based on elements of observational methods and will examine how each method is closer than nonobservational methods to real-life behavior of road users. Overall, research in traffic psychology can be sorted by proximity (or distance) to authentic behavior. The farthest method is using questionnaires. This method is a popular way for measuring attitudes in regards with unsafe behavior such as traffic violations that imply on the people's behavior. Indeed, theories like PBT (Ajzen, 2011) try to measure the variables that mediate between attitudes and real behavior (e.g., perceived norms, perceived control, and behavioral intention). As to behavior measured by questionnaires, using self-report questionnaires (e.g., DBQ—Driving Behavior Questionnaire, Reason et al., 1990) is a well-established method to examine how people behave on the roads.

Studies based on simulations of on-road situations (computerized or by simulators) try to get closer to real behavior by studying road users' behavior in simulated situations. Mostly, the validity as well as the reliability of these methods are high. However, it is still likely that people would react differently in real-life situations. Other studies measure people's reactions in labs (experimental studies). Lab studies can be well-designed and control all the “noisy” variables. Such studies enable the researchers to measure precisely the behavior by objective parameters rather than by estimation. The problem with these studies is that they do not reflect the reality and ignore situations that may intervene between the lab and the real-world situations.

Naturalistic observations are focused on people's behavior, including drivers, pedestrians, and vulnerable road users acting in traffic environment. The studies based on this type of research document the behavior that happens here and now. If well-designed and conducted “by the book”, this genre of research is better than others due to its directness. It does not examine what people are thinking about their behavior, what is their attitude regarding it or what they report about themselves. It simply examines real behavior. In addition, naturalistic observations happen in the real settings rather than in labs or in simulated situations, so we can realize how people really behave in everyday life situations (e.g., Riaz, 2016)

Naturalistic observations have weaknesses as well. Generalizability is one of the problems, because we cannot be sure that the people behave the same in various locations. This may be solved by quantity (observing huge numbers of road users) and variety (various locations, various time of the day, etc.). Data collection might be also problematic, due to subjective estimation of the observers. For that reason, researchers should put an emphasis on good training for their observers. Another solution for this issue can be using electronic devices in order to collect data more objectively and accurately.

Hence, field studies (expanded later) try to combine the advantages of both lab experiments with observations. For example, Zivotofsky et al. (2012) compared the estimated road crossing time as pedestrians of young and elderly individuals. It included precrossing estimation and measuring the actual time it took them to cross the street. Following the actual street crossing pedestrians were asked to estimate the time it took them to cross the road.

Studies Based on Existing Databases

An optional way to relate to data of road users is analyzing big data bases. This includes demographic details, data of traffic violations (e.g., Rosenbloom and Eldror, 2013), drivers' performance on-road (Flannagan et al., 2019), using smartphones while driving in trucks (Hickman and Hanowski, 2012) etc. SHRP 2 Naturalistic Driving Study (Wu and Xu, 2017) is a good example for a large database of naturalistic driving that documented 300 travels and analyzed different aspects of safe driving. This study was based on video clips of the road as well of the drivers' faces. It examined the influence of some external factors on the drivers' behavior and

concluded that the vehicle type, the traffic flow, pedestrians, and drivers' age were the most crucial for explanation of drivers' behavior. In another study, a different group of road users was studied: Hussein et al. (2015) analyzed pedestrians as well as drivers' behaviors in signalized intersections in order to predict future intentions of pedestrians in relation to other road users' action.

Rosenbloom and Eldror (2013) compared police records of 1549 impounded drivers with 1354 controls regarding their matching violations performed prior to the application of the impoundment regulation, comparing accident and traffic-violations involvement in the subsequent year. This has been made in order to examine the effectiveness of vehicle impoundment system as a punishment.

The above-mentioned studies were based on analyzing existing databases. As a matter of fact, these studies were based on data of real behavior, but they still do not investigate behavior directly. The next section deals with direct observations of road users' behavior.

Controlled Observations

Usually, studies in traffic psychology deals with topics such as yielding of drivers to pedestrians, dilemma zone or amber light dilemma, safe road crossing of pedestrians, drivers' risky behavior, and road users' distraction. Observations are made with or without controlled interventions. The following studies have been conducted with interventions. Michael et al. (2000) examined the effect of signs shown to drivers on tailgating while driving at urban roads in normal traffic hours. The observations that have been taken along 10 weeks were recorded by video cameras. At certain places, a sign of "Please don't tailgate" was held by an experimenter and at other places another sign of "help prevent crashes please don't tailgate" was held. The different signs yielded different reactions of drivers and all reactions were recorded. Tailgating was measured by seconds of getting to the first vehicle.

The following study has been based on a rather complex design: Bertulis and Dulaski (2014) conducted a sophisticated design of their study. The purpose of this study was to measure how driver speed affects the rates of yielding for pedestrians. The study was designed to be "staged" with the exact same pedestrian for each driver and have the exact same amount of time to react and decide whether they would yield or not. The observer had a hand-held speed radar gun and a notebook to record the number of yields during attempted crossing events by the person. The study was conducted on the basis of the selective method, meaning the pedestrian (an experimenter) steps into the street as a driver is approaching, but with enough time for the driver to notice and brake. The observer first used the radar gun to measure the speed of at least a dozen vehicles and obtain the 85th percentile speed. Then, only the free-flowing vehicles going within two mph of the 85th percentile speed were recorded. All the locations had the same features regarding the geometry, lane configuration, parking of vehicles and surrounding land use.

Nonetheless, controlled observations have been made not only to study interaction between different road users but also to see how people manage with dual task on the roads. In order to test how smartphone tasks are associated with inattention blindness of pedestrians while crossing the street. Chen et al. (2016) designed a study that combined both field experiment together with interviews. All in all, 2556 pedestrians crossed the street and underwent the interview and were sorted to distracted and nondistracted pedestrians. This was determined by cameras' records. In addition to demographic details that have been collected through the interviews, inattention blindness and deafness were tested by placing a clown walking by them playing the national anthem. The two groups of pedestrians were asked if they noticed something unusual in their surroundings.

In line with the previous study, another field experiment was designed by Rosenbloom (2006) for the purpose of studying the effect of talking with cell phone while driving on speeding and safe gap keeping. The study was based on observations made by a passenger sitting by the driver through spontaneous driving. The manipulation made by the researchers was initiating two phone calls during the travel to which the drivers responded and talked along the travel. The observers sampled two different periods to register: 10 min during conversation and 10 min of nonconversation and compared the speeding and gap keeping behaviors of the drivers. After the travel, the drivers were asked if talking on phone while driving is disturbing.

The study combined spontaneous manipulated observations on pedestrians as well as personality and demographic questionnaires (Rosenbloom, 2006). Students were observed by crossing a signalized intersection (in green or red light). After reaching the other side of the intersection they were asked to participate in a survey that included Sensation Seeking Scale V (SSS; Zuckerman, 1994)

and demographic details. The study was carried out in four conditions: with or without the presence of a policeman and hurrying or not hurrying. The associations of the presence of the policeman, the time, the level of SS with the crossing (in green or in red) were examined.

Uncontrolled Observations Field Experiments (Field Studies)

Unlike controlled observations, uncontrolled observations do not involve any manipulations or interventions of the researchers. Observation means (human observers or video cameras, etc.) are placed in the location of observation to monitor the behaviors of the road users. The advantage of unobtrusive observations over other data collection methods is that people behave more spontaneously and in an authentic way when they are not aware of being observed. This may increase the validity of the study. This chapter makes a distinction between "human" observation and electronic devices which collect data such as video cameras, eye tracking systems and physiological monitoring equipment. For instance, in order to study factors that affect signaling rate on the road, researches placed human observers in junctions to record of 5600 vehicles. Observers used a chart to tick behaviors such as: the

direction of turn, whether the driver signaled or not and how the signaling was done (Faw, 2013). In a similar way, Rosenbloom et al. (2009a, 2009b) conducted a study to examine how often drivers commit driving violation based on the type of location they drive (i.e., city, town, or a village). The researchers examined five different type of violations: seat belt, safety seat for a child, using a phone without a speaker, give-way sign, and stopping in an undesignated area. Trained observers recorded, using a grid- observation, drivers in their natural environment to see which location are characterized with the overall highest rate of driving violations.

One interesting way in observational field studies is to accompany drivers and track their behavior without their knowledge. The present study aimed to do that and to examine the effects of road familiarity on driving performance. Observers joined female drivers for a trip drive as their friends. Drivers were not aware of their participation in the study. The observations were performed on a sample of different drivers who were driving on similar routes in terms of road quality, traffic volume and other road conditions. The more and the less familiar road sections had similar infrastructural and environmental characteristics. In order to mask the observation, the observers told the drivers that they had to complete some paperwork. Each participant drove in two conditions. A familiar condition, operationally defined as the drivers' residence towns and the nearby areas. Less familiar condition, operationally defined as other towns. At the end of the drive, the observer talked to the driver about the amount of tickets she received, average driving distance per month and license seniority (Rosenbloom et al., 2007).

Other studies used observations on pedestrians. A study that aimed to trace crossing behavior of pedestrians compared two groups: individual pedestrians and groups of pedestrians. Observers examined whether or not pedestrians crossed at a red-light and to what extent it happened in each of the research groups. Observers recorded variables like date, estimated age, gender, traffic volume, and durations of the green light using an observational grid (Rosenbloom, 2009).

Some observational studies examine specific populations among pedestrians. For instance, military officers are expected to show more compliance with the rules. Hence, Rosenbloom (2011) checked how values such virtue, honor, patriotism and subordination, which are often expected from officers, are exhibited when they crossroads compared to soldiers and civilians. Six parameters defined a safe crossing and the officers or soldiers were recognized by their uniforms. Close to an army base, and in the center of a city, observers registered the behavior of pedestrians in similar conditions like location, time and traffic volume.

Another example of specific populations is examining gender differences on driving and walking. Four crossroads for pedestrians were chosen. Observers watched all behaviors of both genders, using an observational grid which describes the whole processes of crossing. This study was part of a bigger research aimed to model pedestrians' behaviors. The experimenters stood on the opposite sidewalk, so they could easily check where a pedestrian was looking. The experimenter was equipped with a digital camera with a voice recorder and orally noted the pedestrian behaviors according to the observation grid. Experimenters learnt all behaviors by heart in order to record behaviors faster (Tom and Granié, 2011).

Pedestrians are sometimes involved in accidents because of their own technological or social distractions. Observers recorded pedestrians talking on the phone or to another person or holding hands during crossing the road. The researchers aimed to study to what extent these behaviors affect the duration of the crossing, violations of signs and looking at both sides of the road. Observers had a timer to check the duration of crossing and used a special table, prepared in advance, to record the behaviors (Thompson et al., 2012).

Often, observational studies look at various types of populations at the same time to compare their behaviors. This happens when researchers want to see how one behavior affects the safety of the people on the road. One of these studies compared the behaviors of pedestrians crossing the road with or without using the phone. Observers had an observational grid to record all information. They had both details about demographics and type of behavior, and they had a control group to compare how long it took them to cross the road (time matched control) and their demographics (demographic matched control; Hatfield and Murphy, 2007).

Another research studied pedestrians from a different perspective. The study looked at the connection between violation of pedestrians and motor vehicles in traffic light intersections and the rate of collisions between pedestrians and motorized vehicles. The method (GIS method) included recognizing dangerous junctions and in terms of accidents' rate and recording of the violations (Cinnamon et al., 2011).

The following study focused on the effect of religion on pedestrian behavior. The aim was to investigate the differences between religious and nonreligious pedestrians in urban intersections in two cities: one secular and one ultra-orthodox. Observers looked at behaviors such as running a red-light, crossing where there is no crosswalk, walking along the road, failing to check for traffic prior to crossing, and taking a child's hand when crossing. Three observers observed different groups: children, adults and total number of pedestrians crossing (Rosenbloom et al., 2004). Furthermore, the authors of the previous study were interested to reveal if the phenomenon of ultra-orthodox pedestrians appear solely in ultra-orthodox places. The following study used on observation grid to look at ultra- Orthodox and secular pedestrians in neighboring ultra-orthodox and secular cities. The crosswalks observed were similar to each other and the aim was to track the tendency to cross in a red light. Four trained observers watched the followings: observing children and young pedestrians, observing adult pedestrians, observing elderly pedestrians and last two observers were responsible for tracking the total number of pedestrians crossing the road and the total number of vehicles passing by, respectively (Rosenbloom et al., 2008a,b).

Not only adult pedestrians were in the focus of researchers who made observations. The following study depicted behaviors of children crossing a crosswalk. Two hundred sixty-nine children were observed between the ages of 7–11 near to an elementary school. Some children crossed with the help of an adult and some did not. Unsafe behaviors such as not stopping at the curb, not looking before crossing, attempting to cross when a car is nearing and running across the road were noted. Observation took place on hours that children usually go to school at three similar intersection. The study question was whether and how an accompany of an adult change the road crossing behavior of the children (Rosenbloom et al., 2008a,b).

Electronic Observations

Unlike human observation that is based on trained observers that record people's behavior, electronic observations rely on cameras that are placed either on the roads or in the drivers' car or other electronic systems that can monitor the road users' actions and/or other parameters on-road. One of the first and the most famous study of Naturalistic driving was conducted in the United States with equipped 100 cars were observed (Klauer et al., 2006) for a year. The researchers were interested in the information inside and outside the cars before an accident (or near to accident) occurred.

A study by Tivesten and Dozza (2015) observed drivers to examine, which factors lead to engagement of drivers in visual manual phone activities. The study aimed to answer questions such as: which contexts affect the tendency to be involved in phone activities, which contexts affect the timing of use on the phone and whether self-regulation increase the safety margins of drivers. The researchers collected information from 100 Volvo cars, which was part of a bigger project of EuroFOT using naturalistic driving data, and accompanied them for a year. The cars were driven in real traffic by the main drivers and other household members. The drivers all lived in the same area. No advanced driver assistance system was activated during the first 3–4 months of driving, which is the data used for the study. Researchers used cameras, sensors and other devices to collect the data. The data collection covered travel in its entirety, that is, from the moment the engine started until it was switched off. Though study has not met the full criteria of naturalistic observation (since the drivers were aware that they were observed all the time), drivers were still in their natural environment of driving. Other studies were aimed to look at main risk factors happen right before an accident. Specifically looking at the moments before it happened. Dingus et al. (2016) analyzed collision situations using of video recorder measurements of 3542 drivers that were recruited to SHRP 2 NDS system. The study categorized three age groups and specifically examined drivers under the age of 25 and above the age of 65 which are the ages in which most accidents occur.

In some cases, observations are taken in order to reveal what are the psychological circumstances that yield anger of road users. Precht et al. (2017) examined the correlation exists between anger and accidents and tried to look for the causes- whether it is because aggressive driving or because cognitive load. The study used SHRP2 records and analyzed 10 min videos when drivers were caught angry and the driving violations or mistakes they did through this time. They created behavioral factors to look at and watched which of these behaviors happened in a way that they could conclude for the reason behind the violations or mistakes in driving.

Distraction of drivers is often studied through experimental procedure but also by naturalistic observations. Wu and Xu (2018) conducted a study that looked at the connection between road familiarity and doing secondary task which are a cause of distraction. The reasoning for using naturalistic driving data is because it is hard to collect data about distractive activities like eating, putting on make-up and so on during driving simulations on a lab. The researches define a familiar road as one that drivers pass through at least once a week, and not familiar as one that driver pass through once a year.

Integration of Methods in Observational Field Studies

In order to gain the advantages of naturalistic observations on the one hand and of self-report techniques on the other, some studies combine naturalistic observations with other procedures such as simulations, questionnaires, etc. A combined method of both observations and questionnaires was used to study pedestrians' behavior when crossing the road and their perception of the walking environment before and after installing a marked crosswalk. Researchers examined the crossing behavior 14 months before the crosswalk was installed and after 6 weeks after its installation. Observation was video-based, and questionnaires were conducted consecutive to the crossing, so the sample of the observation and the questionnaires would be similar (Havard and Willis, 2012).

Rosenbloom and Pereg (2012) adopted this trend and examined how long pedestrians wait before crossing the first, the second and the third road divided by an island. The authors hypothesized that though in a narrow island there would be a negative correlation between the first and the second crossing and the waiting time, in a wide island the correlation would be positive since pedestrians perceive the wide two-stage crossing as isolated units. The research integrated two methods: an observational experiment and a questionnaire. Two trained experimenters managed the research. One observer recorded with a stopwatch the waiting time of the pedestrians from the first to the third crossing, checking how long it took the pedestrians to cross each unit. Then, the experimenter noted the information in the observational grid. The second experimenter stood at the end of the third crossings with a questionnaire to ask pedestrians for their demographic details. Only pedestrians who agreed to participate in the research by answering the questionnaire were included in the sample.

A unique study was conducted by Pantangi et al. (2020). This study examined the influence of High Visibility Enforcement Programs (HVEs) on diverse types of aggressive driving behavior, that is, speed, failure to maintain distance and dangerous lane change. Researchers asked whether using such a program reduces aggressive behavior on the road. The study used SHRP2 and NDS. The uniqueness of this study is the development of graded indices to compare the intensity and time period of speed and failure to maintain distance when different levels of aggressive behavior should be perceived. First, a campaign was conducted to inform the residents of the selected areas about the existence of the enforcement plan. Second, road cameras were hacked to track behavior and police officers carried out enforcement.

Signal detection theory is known as a paradigm that is appropriate for studying in labs. Rosenbloom and Wolf (2002b) translated this paradigm to terms of naturalistic behavior. In this study, the researches aimed to ask how the signal detection paradigm can serve to examine drivers' decisions in life traffic dilemmas. Observers joined a driver, with their awareness and permission, as silent passengers. The observers joined the drivers for their driving every day for three to four weeks. All drivers filled sensation-seeking and

demographic questionnaire. Observers were recruited based on their personal interview and filled an approach to driving questionnaires to minimize any of their biases. Only observers that were found to be innocent and well-balanced decision-makers on the road were recruited. The same researchers tried to test drivers for sensation seeking by computerized simulation and responding to it. The study was constructed of three experiments. In each experiment participants were exposed to road dilemma (either by a simulation, or in reality or acting it by themselves). The decision they had to make was always binary: to do or not to do (Rosenbloom and Wolf, 2002a).

Traffic Conflict Technique is an alternative method to analyze statistics of accidents. A conflict situation is when two road-users arrive each other at the same time and space, so if their movement stays unchanged, an accident is irresistible. This study aimed to examine those conflict situations. Data were collected using a video recorder placed on the way and an observational grid at the hand of trained observers. Observation procedures included locating, recording and evaluating the speed and distances, sketching the situation and describing the causes of the conflict. Observers also had audio recorders to report on the traffic situation and the causes of the conflicts. Observers recorded basic data on each conflict, such as date, time of day, light conditions, weather, mixed road users, etc. They also recorded all the circumstances that may contribute to understanding the possible reasons for the occurrence of the conflicts. To facilitate measurement, they mapped the distances between places using power poles, billboards, etc. (Uzundu et al., 2019).

Theories of social psychology are sometimes assimilated into empirical study. In this case, the theory of social norms was translated into a design of study that examined how children, accompanied by adults to cross streets, are influenced by a campaign based on the principles of this theory in an Ultra-Orthodox Jewish community in Israel. In each crosswalk, two trained observers stood on the side of the road and collected data: the first observer documented pedestrians' behaviors while the second was responsible for tracking the total number of vehicles passing. When recording the crossing behaviors of pairs of a young child and an accompanying adult/older child, the observers recorded the behavior of the pair as guided by the accompanying adult or older child. In the second study of the experiment, five trained interviewers who were acquainted with the culture administered questionnaires to the children with permission of their parents (Rosenbloom et al., 2009a, 2009b).

Conclusions

This chapter introduced a large scale of studies based on naturalistic observations in various versions—either as uncontrolled observations, through controlled observations taken by human observers or by electronic systems. Finally, combined methods were introduced that show how one method can complement the disadvantage of another method. In the future, it is likely that new gadgets will be recruited for data collecting in real-life observations, so we will gain many insights regarding human behavior on road. This will surely increase safety on roads and prevent fatalities.

References

- Ajzen, I., 2011. The theory of planned behaviour: reactions and reflections. *Psychol. Health* [online] 26 (9), 1113–1127.
- Bertulis, T., Dulaski, D.M., 2014. Driver approach speed and its impact on driver yielding to pedestrian behavior at unsignalized crosswalks. *Transport. Res. Rec.* 2464 (1), 46–51.
- Chen, P.-L., Saleh, W., Pai, C.-W., 2016. Texting and walking: a controlled field study of crossing behaviours and inattentive blindness in Taiwan. *Behav. Inform. Technol.* 36 (4), 435–445.
- Cinnamon, J., Schuurman, N., Hameed, S.M., 2011. Pedestrian injury and human behaviour: observing road-rule violations at high-incident intersections. *PLoS ONE* 6 (6), e21063.
- Dingus, T.A., Guo, F., Lee, S., Antin, J.F., Perez, M., Buchanan-King, M., Hankey, J., 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proc. Natl. Acad. Sci. U.S.A.* 113 (10), 2636–2641.
- Faw, H.W., 2013. To signal or not to signal: that should not be the question. *Accid. Anal. Prev.* [online] 59, 374–381.
- Flannagan, C., Bärman, J., Bálint, A., 2019. Replacement of distractions with other distractions: a propensity-based approach to estimating realistic crash odds ratios for driver engagement in secondary tasks. *Transport. Res. F* 63, 186–192.
- Hatfield, J., Murphy, S., 2007. The effects of mobile phone use on pedestrian crossing behaviour at signalised and unsignalised intersections. *Accid. Anal. Prev.* [online] 39 (1), 197–205.
- Havard, C., Willis, A., 2012. Effects of installing a marked crosswalk on road crossing behaviour and perceptions of the environment. *Transport. Res. F* 15 (3), 249–260.
- Hickman, J.S., Hanowski, R.J., 2012. An assessment of commercial motor vehicle driver distraction using naturalistic driving data. *Traffic Injury Prev.* 13 (6), 612–619.
- Hussein, M., Sayed, T., Reyad, P., Kim, L., 2015. Automated pedestrian safety analysis at a signalized intersection in New York city: automated data extraction for safety diagnosis and behavioral study. *Transport. Res. Rec.* 2519 (1), 17–27.
- Klauer, S., United States National Highway Traffic Safety Administration (2006). *The Impact of Driver Inattention on Near-crash/Crash Risk: An Analysis using the 100-car Naturalistic Driving Study Data*. U.S. Department of Transportation, National Highway Traffic Safety Administration, Washington, D.C.
- Michael, P.G., Leeming, F.C., Dwyer, W.O., 2000. Headway on urban streets: observational data and an intervention to decrease tailgating. *Transport. Res. F* 3 (2), 55–64.
- Pantangi, S.S., Fountas, G., Anastasopoulos, P.C., Pierowicz, J., Majka, K., Blatt, A., 2020. Do high visibility enforcement programs affect aggressive driving behavior? An empirical analysis using Naturalistic Driving Study data. *Accid. Anal. Prev.* 138, 105361.
- Precht, L., Keinath, A., Krems, J.F., 2017. Effects of driving anger on driver behavior – results from naturalistic driving data. *Transport. Res. F* 45, 75–92.
- Reason, J., Manstead, A., Stradling, S., Baxter, J., Campbell, K., 1990. Errors and violations on the roads: a real distinction? *Ergonomics* 33 (10–11), 1315–1332.
- Riaz, M.S., 2016. *Naturalistic Behaviour Observation of Road Users: A Scoping Review*. [online] Research Gate. Available at: https://www.researchgate.net/publication/309536489_Naturalistic_Behaviour_Observation_of_Road_Users_A_Scoping_Review.
- Rosenbloom, T., 2006. Driving performance while using cell phones: an observational study. *J. Saf. Res.* 37 (2), 207–212.
- Rosenbloom, T., 2009. Crossing at a red light: behaviour of individuals and groups. *Transport. Res. F* 12 (5), 389–394.
- Rosenbloom, T., 2011. Traffic light compliance by civilians, soldiers and military officers. *Accid. Anal. Prev.* 43 (6), 2010–2014.
- Rosenbloom, T., Wolf, Y., 2002a. Sensation seeking and detection of risky road signals: a developmental perspective. *Accid. Anal. Prev.* 34 (5), 569–580.
- Rosenbloom, T., Wolf, Y., 2002b. Signal detection in conditions of everyday life traffic dilemmas. *Accid. Anal. Prev.* 34 (6), 763–772.

- Rosenbloom, T., Pereg, A., 2012. A within-subject design of comparison of waiting time of pedestrians before crossing three successive road crossings. *Transport. Res. F* 15 (6), 625–634.
- Rosenbloom, T., Eldror, E., 2013. Vehicle impoundment regulations as a means for reducing traffic-violations and road accidents in Israel. *Accid. Anal. Prev.* 50, 423–429.
- Rosenbloom, T., Nemrodov, D., Barkan, H., 2004. For heaven's sake follow the rules: pedestrians' behavior in an ultra-orthodox and a non-orthodox city. *Transport. Res. F* 7 (6), 395–404.
- Rosenbloom, T., Perlman, A., Shahar, A., 2007. Women drivers' behavior in well-known versus less familiar locations. *J. Saf. Res.* 38 (3), 283–288.
- Rosenbloom, T., Ben-Eliyahu, A., Nemrodov, D., 2008a. Children's crossing behavior with an accompanying adult. *Saf. Sci.* 46 (8), 1248–1254.
- Rosenbloom, T., Shahar, A., Perlman, A., 2008b. Compliance of ultra-orthodox and secular pedestrians with traffic lights in ultra-orthodox and secular locations. *Accid. Anal. Prev.* 40 (6), 1919–1924.
- Rosenbloom, T., Ben-Eliyahu, A., Nemrodov, D., Biegel, A., Perlman, A., 2009a. Committing driving violations: an observational study comparing city, town and village. *J. Saf. Res.* 40 (3), 215–219.
- Rosenbloom, T., Sapir-Lavid, Y., Hadari-Carmi, O., 2009b. Social norms of accompanied young children and observed crossing behaviors. *J. Saf. Res.* 40 (1), 33–39.
- Tivesten, E., Dozza, M., 2015. Driving context influences drivers' decision to engage in visual-manual phone tasks: Evidence from a naturalistic driving study. *J. Saf. Res.* 53, 87–96.
- Tom, A., Granié, M.-A., 2011. Gender differences in pedestrian rule compliance and visual search at signalized and unsignalized crossroads. *Accid. Anal. Prev.* 43 (5), 1794–1801.
- Thompson, L.L., Rivara, F.P., Ayyagari, R.C., Ebel, B.E., 2012. Impact of social and technological distraction on pedestrian crossing behaviour: an observational study. *Inj. Prev.* 19 (4), 232–237.
- Uzundu, C., Jamson, S., Lai, F., 2019. Investigating unsafe behaviours in traffic conflict situations: An observational study in Nigeria. *J. Traffic Transport. Eng. (Engl. Ed.)* 6 (5), 482–492.
- Wu, J., Xu, H., 2017. Driver behavior analysis for right-turn drivers at signalized intersections using SHRP 2 naturalistic driving study data. *J. Saf. Res.* 63, 177–185.
- Wu, J., Xu, H., 2018. The influence of road familiarity on distracted driving activities and driving operation using naturalistic driving study data. *Transport. Res. F* 52, 75–85.
- Zivotofsky, A.Z., Eldror, E., Mandel, R., Rosenbloom, T., 2012. Misjudging their own steps. *Human factors. J. Hum. Factors Ergon. Soc.* 54 (4), 600–607.
- Zuckerman, M., 1994. *Behavioral Expressions and Biosocial Bases of Sensation Seeking.* Cambridge University Press.

Driving Simulators

Karel A. Brookhuis, University of Groningen, Faculty of Behavioural & Social Sciences, Department of Psychology, Groningen, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	14
History and Motivation (Wikipedia was Consulted for the Facts, Figures, and Names)	14
A Small Side Step; The Flight Simulator	14
Simulation of Motor Vehicles	15
Applications	16
Simulator Validity	17
Simulator Fidelity	18
Simulator Sickness	18
Conclusion	19
References	19

Introduction

Learning to drive common motor vehicles or learning to fly small airplanes is, until the very moment of writing, for a large part still a matter of sitting in the chair behind the steering-wheel or the steering-shaft, and start learning by doing it under supervision, in most countries. However, traffic in the air, on the water and on the road is increasingly dense and complex, leading to more victims and costs, while the quickly decreasing acceptance of the underlying accident numbers is approaching a bending within the traffic population. Travelers, and in their footprint politicians, are looking for and demanding drastic measures to enhance traffic safety. The requirements of steering the vehicles safely and efficiently through the air-, water-, and road-ways are increasingly demanding, calling for more adequate, stricter training, and assessment of pilots, helmsmen, and drivers. In a number of countries, this call has led to changes in licensing which in turn stimulated the development of specific simulators. Therefore, simulators are slowly becoming indispensable in most transport areas these days, for obvious reasons, of which safety, cost reduction, and flexibility are the most pregnant ones. Pilot and driver training is much more efficient in simulators, enabling trainers to confront their pupils with a variety of both common and critical situations and events, and everything in between at will.

History and Motivation (Wikipedia was Consulted for the Facts, Figures, and Names)

In 1959, the TX-2 computer was developed at MIT, enabling the integration of man-machine interfaces. From that moment on, a lot happened around computers and graphics. In 1966, the University of Utah recruited David Evans who founded a department around a computer science program, and computer graphics began to flourish under his guidance. His new department became the worldwide famous research center for computer graphics, during more than 10 years, attracting many talented computer scientists. With the general introduction of specialized graphics computers in the 1980s of the last century, the development to take care of the representation and manipulation of (traffic) environment images in 3D started. Based on metal-oxide-semiconductor (MOS) within VLSI (very-large-scale integration) technology, 16-bit central processing unit (CPU) microprocessors and the first graphics processing unit (GPU) chips revolutionized computer graphics, enabling real high-resolution graphics. Then, in the nineties 3D modeling evolved on a mass scale, stimulated by the enhancing quality of computer graphics interface (CGI) generally. This upheaval heralded home computers, enabling to take on the rendering tasks that had thus far been limited to “heavy” graphics computers and workstations such as the famous silicon graphics series, once top of the bill but costing a small fortune. A host of 3D modelers became available for powerful Microsoft Windows and Apple Macintosh machines, enabling university departments, research institutes, and companies to fully exploit the GPU toward the prominence in graphics it still enjoys to date. Driving simulators par excellence profited from this development.

A Small Side Step; The Flight Simulator

Bob Carson developed the *Flight Simulator* in 1977, in fact a series of amateur flight simulator programs for MS-DOS and Mac OS, later followed by versions on Microsoft Windows Operating Systems. It turned out to be the longest-running, best-known, and most comprehensive home flight simulator programs on the market. The flight simulator made you an “airplane pilot” flying all over the world, bringing most of the world’s main airports within reach. No competition for real airplane simulators, of course, but a sign of the time. And, training airplane pilots, being military or civilian, in fact did the same but under strict circumstances, confronting the pilots “in spe” with the earlier mentioned necessary variety of both common and critical situations and events.



Figure 1 Mock-up of a truck, offered to the Dutch ANWB for truck driver training.

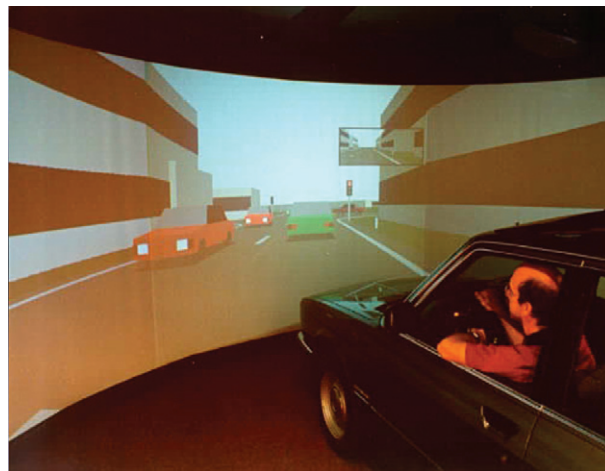


Figure 2 Driving simulator mockup and projection (Anno 1996).

Simulation of Motor Vehicles

The first generation of driving simulators, or at least sort of, was primitive, for instance, just a mock-up for basic training to change gears. Before and around the Second World War, trucks were equipped with a rather complex gear changing machinery, the double clutch, which could be trained in a mock-up ([Fig. 1](#)) to avoid damage of the gear box in real trucks.

In the 1970s, several more realistic simulators were developed around the world, some of which were based on television cameras hanging over a running belt mimicking a road, pictures displayed on televisions in front of a steering wheel, and car seat with pedals. The steering wheel would move the camera in lateral directions while gas and brake pedals would control the speed of the running belt. In this way, data could be collected from lateral and longitudinal control of the driver that participated in an experiment, see [Blaauw \(1979\)](#) for an example. This setup would enable studies concerning continuous driving under different conditions, such as research into the effects of sleepiness with time on task. [Blaauw \(1982\)](#) was one of the first to compare driving in a simulator and on the road, in an instrumented vehicle. He came to the conclusion that this type of simulators was well able to produce good absolute and relative validity for longitudinal vehicle control. Lateral vehicle control offered only good relative validity (see also section "Validity").

Driving simulators came to a first level of maturity with the introduction of graphics computers in the eighties. From that moment on, interaction between the driver and the traffic environment could be studied in quickly enhancing graphic surroundings. In a first, simple the traffic environment could be displayed on CRTs at eye height for the "driver," who was sitting on a chair with a steering wheel and pedals in front. Or, more sophisticated, could be projected on screens in front of a mockup to give it a real look and feel ([Fig. 2](#)).

Enhancement of the feeling of driving a real motor vehicle on different types of road came with adding a moving base, enabling 14 degrees tilt all sides ([Fig. 3](#)).



Figure 3 Medium advanced, moving based simulator (Anno 2014).

Applications

The present driving simulators provide an opportunity to evaluate the driving capacity of people that potentially deviate negatively from “normal” driving. For many years now, formal driving capability assessments for a driver license comprised both clinical and on-road driving components, whilst the build-up creates risk issues for the driving evaluator, the public, and the driver. In particular, young drivers, older adults, sick drivers, handicapped drivers, and drivers under influence of psycho-active substances should be able to detect their shortcomings and learn how to compensate or avoid. Simulator evaluations can identify those shortcomings, potentially adding improvement tools to clinical tools, increasing driving capacity (Gibbons et al. 2014). The present on-road evaluation of driving skills cannot match highly standardized driving simulators, that are relatively cheap, cost-effective and time efficient. Gibbons et al. (2014) studied differences of simulator composition with one or three screens and between older and middle-aged drivers. They found a lot of congruence between the simulator conditions, and concluded that a one-screen simulator sufficed for their specific purpose.

Another important research strand for simulator research is the influence of psychoactive substances on driving ability. As early as 1982, O’Hanlon et al. (1982) demonstrated that 10 mg of one of the most popular anxiolytics of the benzodiazepine family, diazepam, affected the driving capacity of professional drivers dangerously. Many people use prescription drugs on a daily basis, which in most cases implies driving motor vehicles. Whereas a host of medicinal drugs is put on the market the past 40 years, testing all of them to disastrous effects on driving capacity has not been done, partly because it is illegal to do that on the road in most countries. Simulators would be the solution. The Center for Addiction and Mental Health (CAMH) in Toronto and McGill University used the Virage Simulation VS500M driving simulator to study and measure the effects of cannabis on driving. Among young recreational cannabis users, a 100-mg dose of cannabis by inhalation had no effect on simple driving-related tasks, but there was significant impairment on complex tasks, especially when these were novel. These effects, along with lower self-perceived driving ability and safety, lasted up to 5 h after use. Veldstra et al. (2012) investigated the effects of 20-mg oral cannabis (Dronabinol) in a driving simulator, and found modest effects on driving performance. But the impairing effects that were observed were negative for traffic safety since they were comparable to driving under the influence of 0.5‰ BAC (legal limit in many countries) or more. Drivers did not seem to be able or prepared to compensate for possible impairing effects. Using simulators enables the exploration of the boundaries of driving safety under influence.

Many vehicle manufacturers operate driving simulators, for example, BMW, Ford, Renault, and FIAT. Many universities also operate simulators, for research. In addition to studying driver training issues, driving simulators allow the manufacturers to study driver behaviour under conditions in which it would be illegal and/or unethical to place drivers. For instance, studies of driver distraction by dashboard gadgets would be dangerous and unethical to do on the road. Brookhuis et al. (2008) investigated the effects of a congestion warning assistant that would warn the drivers in advance of a looming congestion (Fig. 4). The assistant should enable them to either adapt their speed in advance of even taking an alternative route, a facility that is more or less standard in advanced navigation systems these days.

With the increasing use of various in-vehicle information systems (IVIS) among which satellite navigation systems, smartphones, music and e-mail systems, simulators are playing an important role in assessing the safety and utility of such devices. Lately, much of these developments boil down into motor vehicle automation research. Automated driving technology has a great potential to enhance road transportation and improve traffic safety. Automated vehicles are expected to considerably reduce the number of accidents caused by human errors, increase fast and fluent traffic flow, increase comfort by letting the driver engage in alternative tasks and ensure mobility for all people, including older and impaired individuals (Kyriakidis et al., 2017). However, human beings tend to be poor supervisors of automation. With respect to Automated vehicles in particular, up to Level 4 automation, human drivers will be a key component, because they should operate the vehicle in conditions not supported by the automation, and will be expected (Level 4), or even required (Levels 2 and 3), to resume manual control when needed. A number of simulator studies have already demonstrated that the human driver cannot be expected to perceive and compensate for automation failure. Merat et al. (2014) conducted a driving simulator study to investigate drivers’ ability to resume control from a highly automated vehicle when



Figure 4 The congestion warning display as positioned in the simulator.

automation was switched off and manual control had to be resumed. They showed that whether automation transition to manual is based on intervals or on demand, it takes drivers some 35–40 s to resume lateral control of the vehicle in a stable manner. The results of this study indicate that if drivers are out of the loop, due to control of the vehicle in a limited self-driving situation (Level 3 automation), their ability to regain control of the vehicle is better if they are expecting automation to be switched off, that is, staying in the loop. Another problem of Automated vehicles is the perception and reaction of their “fellow” traffic participants in the immediate environment. Forecasting what the Automated vehicle will do might lead to irregular behavior of pedestrians and cyclists. A lot of research is still needed to determine when and where full automation could be allowed and installed, for which simulators are indispensable.

Simulator Validity

Blaauw (1982) was one of the first who compared on-road and simulator driver behaviour. While such comparative studies are not easy to carry out (see also Carsten and Jamson, 2011), they remain the most appropriate means to evaluate and improve driving simulator validity. Most of them relied on behavioral measures to capture the differences between on-road and simulated driving (e.g., Blana and Golias, 2002). More recently, few studies assessed the level of mental workload when driving in a real vehicle or in a simulator and showed contradictory results (e.g., Veldstra, 2014). Blaauw (1982) proposed to assess driving simulator validity based on physical and behavioral criteria. Physical validity corresponds to the extent to which software and hardware components of the simulator reproduce the physical reality of driving. Behavioral validity has been defined as the correspondence between driver behavior in a simulator and on the road. Blaauw (1982) also made a distinction between relative and absolute behavioral validity. Absolute validity is established with the sameness of numerical values in the behavioral measures in the two different environments. Relative validity is concluded when the differences due to the experimental manipulation are of the same order and direction and have a similar or identical magnitude in real and simulated driving conditions (for a discussion, see Veldstra et al., 2015).

Veldstra et al. (2015) and Veldstra (2014) compared simulator and on-road studies into effects of diverse alcohol levels and dronabinol (10 and 20 mg) on driving parameters, among which lateral control over the vehicle. They concluded that the driving simulator proved to be equally sensitive for demonstrating drug-induced effects as on the road driving especially for the road tracking task (Fig. 5). However, since inter individual variability was higher in the simulator than on the road more caution should be taken when testing for equivalence of treatment effects to effects found for alcohol. Nevertheless, the effects of alcohol on lateral control in the simulator indicate a similar, exponential progress on Delta SDLP (Difference between the three BAC levels and placebo), as the exponential effect of comparable alcohol levels on accident risk (Borkenstein, 1964).

Veldstra (2014) investigated the validity of the driving simulator for speed research: comparing speeding in the simulator to speeding in private cars. Fifteen drivers whose driving behaviour was followed in their private cars via so-called black boxes including Global Positioning System (GPS)-technology also drove in the driving simulator. The main parameter was the percentage of speed violations conducted. Road types with speed limits of 30, 50, 80, 100 or 120 km/h were analyzed for speed violations too. The same road types were reiterated in the driving simulator. The results indicated that speed violations committed in the simulator and in the private car depended on road type. Speed violation patterns were similar in relative terms for both settings, but only in the higher speed ranges. Furthermore, the drivers’ perceived physical correspondence between driving in a private car and simulator was related

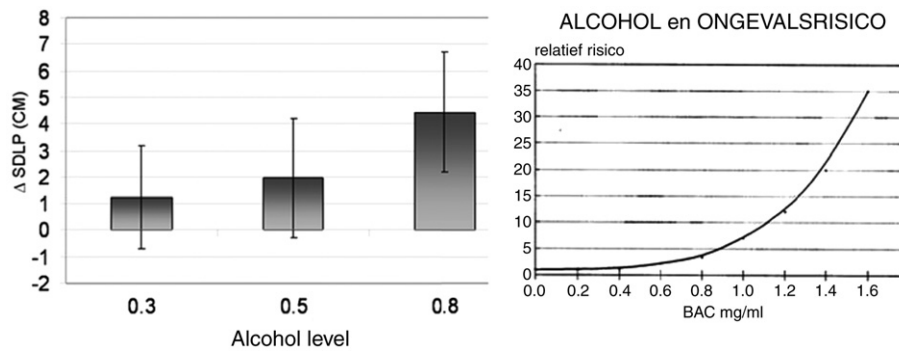


Figure 5 Delta SDLP (+95% CI), that is, difference to placebo, for three alcohol level conditions (A), the “Borkenstein curve,” relating BAC and relative accident risk (B).

to the extent that speed patterns in both environments were similar. This suggests that the driving simulator is a relatively valid instrument to assess speed violations, predominantly in the higher speed ranges.

Most traffic accidents can be attributed to driver impairment, for example, inattention, workload, fatigue, etc. It is now technically feasible, and tested the limits in simulators, to monitor and diagnose driver behaviour with respect to impairment with the aid of a limited number of (in-vehicle) sensors. A valid framework for the evaluation of driver impairment, and a method to assess absolute and relative criteria was proposed (see [Brookhuis et al., 2003](#)). Criteria for when impairment is below a certain threshold, leading to accidents, have been established in the simulator. Accidents and mental workload (either high or low) were related to each other, in conjunction with the origins of mental workload such as information overload, boredom, fatigue, or factors such as alcohol and drugs. The traffic environment and traffic itself will only gain in complexity, at least in the near future, with the rapid growth in numbers of automobiles and electronic applications. The study of the consequences of this forecast will naturally concentrate in simulators.

Simulator Fidelity

Some studies have paid attention to hardware components of the simulator to improve physical fidelity, but it is relatively well established that behavioral validity is a more important criterion than physical fidelity ([Veldstra et al., 2015](#)). Nevertheless, in driving simulators, the control over, and measurement of driver state is relatively easily derived (although skills are required and it is time consuming). In simulators of certain minimum fidelity, conditions may be created that cannot be matched in the real world, with respect to flexibility, safety and measurement accuracy. The laboratory-equivalence of serious driving simulator environments allows full control with respect to environmental conditions, scenarios and stimuli, and enables physiological measurement of state-related parameters such as heart rate and brain activity. In 2007, The National Advanced Driving Simulator developed by the National Highway Traffic Safety Administration (NHTSA) to conduct human factors research on driver behavior, was the most advanced, high fidelity simulator in the world. The simulator consists of a dome with a vehicle cab inside. The vehicle is attached to a motorized turntable that allows the dome to rotate and simulate different driving conditions. 64 feet of longitudinal and lateral travel and 330 degrees of rotation are used to give motion cues to the driver inside. Different makes and models of car cabs can be utilized. However, soon other institutes in different countries built similar, comparable machines.

The Latest development, according to their developers TNO in the Netherlands, is the Desdemona motion platform, considered unique in the world. It combines a centrifuge design for G-loading with 5 additional degrees-of-freedom. Making it the ideal simulator for motion critical training scenario's. The advanced motion simulation has been investigated for driving under armor in difficult off-road conditions, among others. The motion solutions include one-to-one terrain simulation, and make use of high-fidelity vehicle models.

Simulator Sickness

While driving simulators are a valuable tool for assessing multiple dimensions of driving performance under relatively safe conditions, researchers, and practitioners must be prepared for participants that suffer from simulator sickness. [Brooks et al. \(2010\)](#) describe multiple theories of motion sickness and discuss methods for assessing and reacting to simulator sickness symptoms. Individuals can possibly be identified who are unable to complete a driving simulator study due to simulator sickness with greater than 90% accuracy. Older participants are found to have a greater likelihood of simulator sickness than younger participants. Possible explanations for increased symptoms may be investigated in a simulator, as well as implications for research ethics and simulator sickness prevention.

Simulator sickness is a form of motion sickness (for an overview see, [Johnson, 2005](#)). Motion sickness (MS), more specifically seasickness, has been known and documented since at least the ancient Greeks. Simulator sickness (SS) has been known and

documented since the first helicopter flight simulator was procured in the mid-1950s. The most common signs and symptoms of MS are pallor, cold sweating, nausea, and vomiting. Vomiting is rare in SS. The most common signs and symptoms of SS involve visual symptoms, disorientation, and nausea. Only individuals with intact and normally functioning vestibular systems are susceptible to MS and SS. Several measurement techniques exist. The most practical is the Motion Sickness Assessment Questionnaire (MSAQ, Gianaros et al., 2001). The incidence of SS can range from very low to exceedingly high, depending upon the simulator, tasks, conditions, and population. In most individuals, the symptoms of SS subside in less than 1 h. Residual aftereffects lasting longer than 12 h are relatively rare. Adaptation is the single most effective solution to the problem of MS and SS. Most individuals adapt within a few sessions, some individuals require considerable exposure to adapt, and 3%–5% of individuals never adapt. Susceptibility to SS varies depending upon many factors, including age, flight experience, and prior history of MS. Performance of cognitive, perceptual, and psychomotor tasks is largely unaffected by MS and SS. Motivation to perform, however, is strongly affected.

Conclusion

Simulators are indispensable for training purposes and may be very convenient if not necessary for research, specifically when safety and flexibility are at stake. The time may come that governments demand authorized reports on testing new drugs before marketing and sales to ambulant patients who would regularly drive motor vehicles. Or authorities want to have established that driver support facilities do not dangerously distract drivers. To go after these limits, driving simulators may be given a central role in the near future.

References

- Blaauw, G.J., 1979. Simulatie en wegontwerp (Simulation and road design). Report IZF 1979-C21, Institute for Perception TNO, Soesterberg, The Netherlands, 1979.
- Blaauw, G.J., 1982. Driving experience and task demands in simulator and instrumented car: a validation study. *Hum. Factors* 24, 473–486, doi:10.1177/001872088202400408.
- Blana, E., Golias, J., 2002. Differences between vehicle lateral displacement on the road and in a fixed-base simulator. *Hum. Factors* 44, 303–313, doi:10.1518/0018720024497899.
- Borkenstein, R.F., Crowther, R.F., Shumate, R.P., Zeil, W.W., Zylman, R., 1964. The Role of the Drinking Driver in Traffic Accidents. Department of Police Administration, Indiana University, Bloomington.
- Brookhuis, K.A., De Waard, D., Fairclough, S.H., 2003. Criteria for driver impairment. *Ergonomics* 46, 433–445.
- Brookhuis, K.A., Van Driel, C.J.G., Hof, T., Van Arem, B., Hoedemaeker, M., 2008. Driving with a congestion assistant; mental workload and acceptance. *Appl. Ergon.* 40, 1019–1025.
- Brooks, J.O., Goodenough, R.R., Crisler, M.C., Klein, N.D., Alley, R.L., Koon, B.L., Logan, W.C., Ogle, J.H., Tyrrell, R.A., Wills, R.F., 2010. Simulator sickness during driving simulation studies. *Accid. Anal. Prev.* 42 (3), 788–796, doi:10.1016/j.aap.2009.04.013.
- Carsten, O., Jamson, A.H., 2011. Driving simulators as research tools in traffic psychology. In: Porter, B.E. (Ed.), *Handbook of Traffic Psychology*. Academic Press, London, pp. 87–96.
- Gianaros, P.J., Muth, E.R., Mordkoff, J.T., Levine, M.E., Stern, R.M., 2001. A questionnaire for the assessment of the multiple dimensions of motion sickness. *Aviation Space Environ. Med.* 72, 115–119.
- Gibbons, C., Mullen, N., Weaver, B., Reguly, P., Bédard, M., 2014. One- and three-screen driving simulator approaches to evaluate driving capacity: evidence of congruence and participants' endorsement. *Am. J. Occup. Ther.* 68, 344–352, doi:10.5014/ajot.2014.010322.
- Johnson, D., 2005. Introduction to and Review of Simulator Sickness Research (PDF). Research Report 1832. U.S. Army Research Institute for the Behavioral and Social Sciences.
- Kyriakidis, M., de Winter, J.C.F., Stanton, N., Bellet, T., van Arem, B., Brookhuis, K., Martens, M.H., Bengler, K., Andersson, J., Merat, N., Reed, N., Flament, M., Hagenzieker, M., Happee, R., 2017. A human factors perspective on automated driving. *Theor. Issues Ergon. Sci.* 20 (3), 223–249, doi:10.1080/1463922X.2017.1293187.
- Merat, N., Jamson, A.H., Lai, F.C.H., Daly, M., Carsten, O.M.J., 2014. Transition to manual: driver behaviour when resuming control from a highly automated vehicle. *Transport. Res. F* 27, 274–282.
- Veldstra, J.L., 2014. When the party is over. Investigating the effects of alcohol, THC and MDMA on simulator driving performance. PhD thesis, University of Groningen, Groningen, The Netherlands.
- Veldstra, J.L., Bosker, W.M., De Waard, D., Ramaekers, J.G., Brookhuis, K.A., 2015. Comparing treatment effects of oral THC on simulated and on-the-road driving performance: testing the validity of driving simulator drug research. *Psychopharmacology (Berl)* 232 (16), 2911–2919, doi:10.1007/s00213-015-3927-9.
- Veldstra, J.L., Karel, A., Brookhuis, K.A., Dick de Waard, D., Molmans, B.H.W., Verstraete, A., Skop, G., Jantos, R., 2012. Effects of alcohol (BAC 0.5‰) and ecstasy (MDMA 100mg) on (simulated) driving performance and traffic safety. *Psychopharmacology* 222 (3), 377–390.

Naturalistic Driving Studies: An Overview and International Perspective

Johnathon P. Ehsani*, Joanne L. Harbluk[†], Jonas Bärgrman[‡], Co-author for European Union Case Study: Ann Williamson[§], Jeffrey P. Michael*, Co-authors for Australian Case Study: Raphael Grzebieta[§], Jake Olivier[§], Jan Eusebio[§], Judith Charlton[¶], Sjaanie Koppel[¶], Kristie Young[¶], Mike Lenné[¶], Narelle Haworth[¶], Andry Rakotonirainy[¶], Mohammed Elhenawy[¶], Gregoire Larue[¶], Teresa Senserrick[¶], Jeremy Woolley^{**}, Mario Mongiardini^{**}, Christopher Stokes^{**}, Co-authors for Canadian Case Study: Paul Boase^{††}

John Pearson^{‡‡}, Co-author for Chinese Case Study: Feng Guo^{§§}, *Center for Injury Research and Policy, Department of Health Policy and Management, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States; [†]Human Factors and Crash Avoidance, Multi-Modal and Road Safety Programs, Ottawa, Ontario, Canada; [‡]Vehicle Safety Division at the Department of Mechanics and Maritime Science, Chalmers University of Technology, Gothenburg, Sweden; [§]Transport and Road Safety (TARS) Research, University of New South Wales, UNSW, Sydney, NSW, Australia; [¶]Monash University Accident Research Centre, Clayton, VIC, Australia; ^{||}Centre for Accident Research and Road Safety, Kelvin Grove, QLD, Australia; ^{**}Centre for Automotive Safety Research, University of Adelaide, SA, Australia; ^{††}Multi-Modal and Road Safety Programs, Ottawa, Ontario, Canada; ^{‡‡}Council of Deputy Ministers Responsible for Transportation and Highway Safety, Ottawa, Ontario, Canada; ^{§§}Virginia Tech Transportation Institute, Blacksburg, VA, United States

© 2021 Elsevier Ltd. All rights reserved.

Naturalistic Driving Studies: An Introduction	21
What is Naturalistic Driving?	21
Strengths of NDS	21
Limits and Considerations of NDS	23
Ethical Issues Related to NDS	23
Case Studies	23
Australia	24
Main Research Question	24
Sample Size and Inclusion Criteria	24
Instrumentation Installed	24
Days/Months/Years of Observation per Participant	24
High-Level Description of Analyses	24
Driver Distraction	24
Safety at Intersections	24
Main Findings	25
Lessons Learned	26
Funding Sources and Related Issues	26
Other Issues	26
Canada	26
Main Research Question	26
Sample Size and Inclusion Criteria	27
Passenger Vehicle Participants	27
Truck Participants	27
Instrumentation Installed	27
Days/Months/Years of Observation per Participant	27
High-Level Description of Analyses	27
Main Findings	28
Lessons Learned	29
Funding Sources and Related Issues	29
Other Issues	29
China	30
Main Research Question	30
Sample Size and Inclusion Criteria	30
Instrumentation Installed	30
Days/Months/Years of Observation per Participant	30
Main Findings	30
Lessons Learned	30
Funding Sources and Related Issues	30
Other Issues	30
European Union	31
Main Research Question	31
Sample Size and Inclusion Criteria	31
Instrumentation Installed	31

Days/Months/Years of Observation per Participant	32
High-Level Description of Analyses	32
Main Findings	32
Lessons Learned	32
Funding Sources and Related Issues	33
Other Issues	33
United States	34
Main Research Questions	34
Sample Size and Inclusion Criteria	34
Instrumentation Installed and Categories of Data Collection	34
High-Level Description of Analyses	35
Main Findings	35
Lessons Learned	36
Funding Sources and Related Issues	36
Other Issues/Additional Information	36
Synthesis of Lessons from across the Case Studies	36
What is the Future of NDS?	36
Acknowledgment	37
References	37
Further Reading	38

Naturalistic Driving Studies: An Introduction

What is Naturalistic Driving?

Road safety researchers and practitioners need good data to understand factors related to crash risk and many other aspects of driver performance and behavior to improve road safety. These data can come from a range of sources, such as police crash records, roadside observation, experiments on test tracks or driving simulators, and surveys. Each of these can provide meaningful data, depending on the nature of the research question being addressed. A common limitation is their inability to provide an objective measure of behavior across an extended period of time or capture high-risk events as they occur, to shed light on what happens prior to the event, during the event and after the event.

Naturalistic driving studies (NDS) address this gap by enabling prospective, continuous data collection of a range of behaviors on the road in real-world traffic environments. NDS typically involve instrumentation of the vehicle a participant drives routinely in order to observe and document their normal driving behaviors. To date, a number of large scale national and international NDS have been conducted (some of which are described as case studies in this chapter), dozens of smaller studies have been conducted, and many more are planned or underway.

Therefore, it is somewhat surprising that there is no accepted definition of what constitutes NDS or what a study must include to be classified as “naturalistic”. Rather than requiring any specific data collection technologies or methods that NDS should use, researchers have proposed principles or frameworks that NDS should follow. For example, Van Schagen and colleagues suggest that NDS should “provide insight into driver behavior during everyday trips by observing in detail the driver, the vehicle and the surroundings through unobtrusive data gathering equipment and without experimental control” (van Schagen et al., 2011). Lotan and colleagues offer a framework for NDS that spans a range of basic to high-resolution data collection technologies according to the scale of the study (Lotan et al., 2017). Specific technologies are recommended depending on the balance between resolution and scale (Fig. 1).

Typically, an NDS would involve a data collection system that includes a computer that receives and stores continuous data from accelerometers, a global positioning system (GPS), and cameras. Video cameras may capture a driver’s face, the dashboard, areas reachable by the driver’s hands, and the forward and rear roadway. Data collection systems could be custom built, assembled using off-the-shelf components, or use the capabilities of a smartphone, depending on the purpose of the study. While the majority of NDS have been conducted using cars and trucks, NDS have also examined behaviors involving other transportation modes such as motorcycles, bicycles, scooters, and walking.

In order to address the lack of an inclusive definition of a NDS, we offer the following definition: *An NDS is the prospective collection of continuous high-resolution behavioral data in a cohort of road users during uncontrolled driving on public roads without experimental manipulation for a defined period of time, usually for a minimum of a few weeks but could much longer.*

Strengths of NDS

Continuous data collection from multiple sources (e.g., video, GPS, accelerometers, etc.) means that NDS can generate extremely large amounts of data in a relatively short time. Given the richness of the data they supply, naturalistic studies have emerged as an

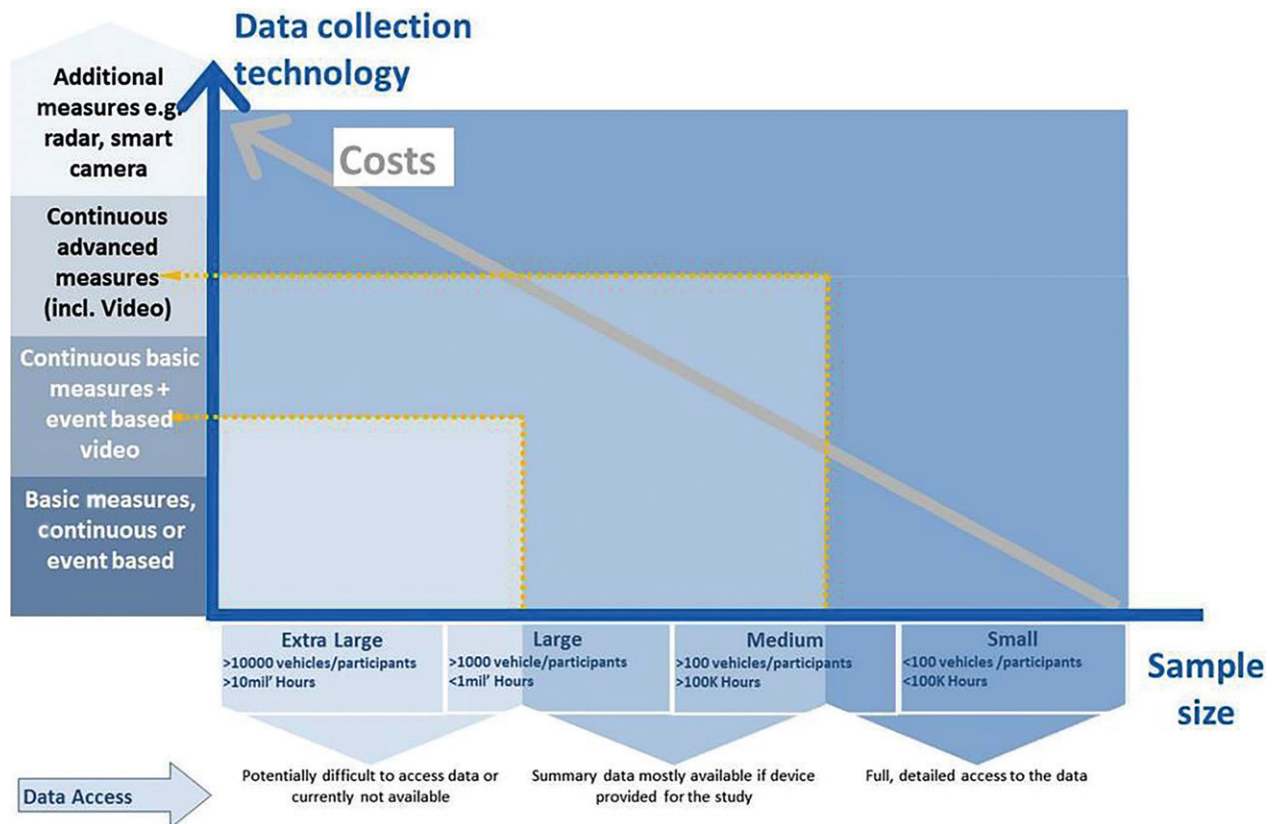


Figure 1 Framework for Naturalistic Studies. Source: from [Lotan et al. \(2017\)](#).

important tool in the study of road safety. NDS offer a range of behavioral, vehicle and environmental data that can be analyzed repeatedly from a variety of perspectives. NDS are not reliant on the memory of a driver or an observer, or on a specific set of measures or questions. Findings from NDS have informed important research policy debates in the areas of distracted driving ([Klauer et al., 2014](#)), advanced driver assistance systems ([Fitch and Hanowski, 2012](#)), impaired driving ([Dingus et al., 2016](#)), and vulnerable road users ([Ehsani et al., 2020](#); [Li et al., 2017](#)) to name a few.

NDS fill an important role in understanding both existing and emerging behavioral risks, supplementing empirical evidence from crash reports and experimental studies such as simulator research. NDS provide insight into the role of driver behavior in the contexts and conditions in which they drive. They can provide new insights on long-standing problems such as driver fatigue and can contribute to our understanding of contemporary issues such as texting while driving. In the case of distracted driving, police crash reports were used to identify the number and type of crashes that appeared to be related to distraction. Simulator studies were then used to identify how specific distractions could affect certain driving scenarios. Naturalistic studies then informed researchers and policymakers of how distractions translated into actual crash risk. In addition, NDS can provide information about when a distraction was present but no crash occurred and to identify factors that might be protective. Importantly, NDS data tell us not only what is wrong, but provide the necessary evidence to develop mitigations to improve safety.

In order to extract meaningful information from large amounts and various sources of data, novel analytical and statistical methods are being developed to analyze NDS ([Bärgman, 2016](#); [Guo, 2019](#)). While a review of the full range of methods is beyond the scope of this article, one specific approach which is unique to the analysis of NDS video data will be discussed in detail here.

In order to analyze NDS video data, researchers have adapted a widely used method from epidemiology called the case-control study ([Breslow, 1996](#)). In this design, driver behavior in a random sample of routine, uneventful driving segments becomes the *control* sample. The *cases* are made up of driver behavior a few seconds before a crash ([Guo and Hankey, 2009](#)). By comparing driver actions during the cases and controls, the association between certain behaviors and crash risk can be established. For example, several NDS have found that dialing a cell phone while driving is associated with a significantly increased risk of a crash ([Dingus et al., 2016](#); [Klauer et al., 2014](#); [Simons-Morton et al., 2014](#)). In order to reach this conclusion, researchers viewed the control videos and observed how frequently drivers were dialing a hand held cell phone. They also observed how often drivers dialed a cell phone during cases and found that it occurred more frequently just before a crash. Activities that require drivers to take their eyes off of the forward roadway result in the greatest risks.

An important point to keep in mind is that due to their nonexperimental nature, NDS are typically not suitable for conclusions about cause and effect to be definitively drawn ([Carsten et al., 2013](#)). Rather, NDS provide the ability to draw associations which can be observed in the data.

Limits and Considerations of NDS

A key limitation of NDS is the resource intensity of the studies. Instrumentation of vehicles is expensive and complicated. This tends to limit the sample sizes of studies and raises questions about generalizability of the findings. In addition to the resources needed for data collection, an equal or greater allocation of resources is needed to clean, process and analyze the data. This process can be prolonged by issues in data quality or an extended period of “on-boarding” for researchers as they learn how to use and interpret data from multiple, complex datasets.

A further limitation of NDS is that most studies rely on volunteer participants, who are likely to be safer, less risky drivers who are willing to contribute their data for safety research. Combined with the fact that serious crashes are rare events in an individual driver's lifetime, this means crashes occurring in NDS tend to be relatively minor. Some have argued that NDS findings have limited generalizability for safety interventions or policy development because they include crashes that are not representative of serious events that need to be prevented (Knippling, 2015). However, this discounts the value of NDS in advancing an understanding of driving behaviors other than serious crashes. There is currently no other way to observe what drivers usually do when they are driving normally and how often and when they engage in risky behaviors or how they behave in near misses. Furthermore, safety intervention and policy development is rarely reliant on a single source of data. NDS represent one of a constellation of methods and research approaches that can form the basis of effective policy and intervention development.

A common question about NDS is whether the presence of data collection equipment in the vehicle might change participant driving behavior. It seems plausible that the presence of in-vehicle data collection equipment could feel invasive for study participants and may influence their driving behavior, at least initially. The single study on this topic is inconclusive (Ehsani et al., 2017). Qualitative research on the topic found that while some participants quickly forget about the instrumentation, others never become accustomed to its presence, even after an extended period. However, there is no objective evidence that awareness of the instrumentation is associated with crash risk or other risky driving behaviors.

A more subtle, but equally important challenge to the validity of NDS comes from the case-control study design that forms the basis of many key NDS findings related to crash risk. When researchers are reviewing videos of crash events (cases) and routine driving (controls), they are unlikely to be blinded to the outcome of the event. This may result in unequal searching for explanatory behaviors, with a less intense search in the control videos. This “detection bias” is a well-documented issue in clinical case control studies (Haynes et al., 2005; Kopec and Esdaile, 1990) but has not been widely discussed in NDS literature. This limitation of NDS is particularly pronounced when the behavior in question could be subjectively interpreted, for example driving too fast for conditions. Reporting the inter-rater reliability of all human-coded driving behaviors in NDS studies could be an important step in documenting this limitation.

Ethical Issues Related to NDS

NDS studies present distinctive ethical issues related to participant privacy. For example, video data from multiple camera views may capture the behavior of the driver, passengers other road users. Origin and destination information obtained from GPS data could include the place of residence and place of work which may make it possible to identify an individual. In the event of a serious crash, data collected from cameras and accelerometers could provide evidence for who was at fault. For these reasons, rigorous ethics reviews are recommended prior to undertaking NDS studies. Extensive ethics training of all personnel who work with the resulting data is also recommended.

To the extent possible, additional steps to safeguard participants should be taken. For example, most studies analyzing the GPS data will remove the first few kilometres/miles around the origin and destination, reducing the privacy risks. Confidentiality agreements could also be put in place to secure the privacy of the study participants. In the United States, the National Institutes of Health issues a Certificate of Confidentiality which allows data collection with the provision that if these data were ever subpoenaed, they would be considered protected. These Certificates of Confidentiality can be obtained for any study in the United States (National Institutes of Health, 2020). Specific requirements and processes differ from country to country.

Case Studies

In order to provide practical insight for those considering using NDS methods, investigators of several large-scale NDS in Australia, Canada, China, the European Union, and the United States were asked to provide a description of the studies they had conducted according to the list of questions below.

1. Main research question
2. Sample size and inclusion criteria
3. Instrumentation installed
4. Duration of observation per participant
5. High level description of analyses
6. Main findings
7. Lessons learned
8. Funding sources and related issues
9. Other issues

Australia

Main Research Question

The Australian Naturalistic Driving study (ANDS) was launched in 2015 and is Australia's first large scale naturalistic driving study. The objective of the ANDS project was to collect driving data for experienced drivers over 4 months of normal driving. The project was not designed to collect data on crashes, but the data collected will provide insights into how crashes were avoided. The strength of the ANDS dataset is that it contains a longitudinal sample of ordinary driving for experienced Australian drivers and can show us what individual drivers usually do.

Sample Size and Inclusion Criteria

The ANDS sample included 347 vehicles/drivers and 62 secondary drivers. The inclusion criteria were designed to capture normal driving by reasonably to very experienced drivers:

- 21–70 years old, with the aim for an even split between male/female and across 20–35, 36–50, 51–70 year age groups;
- drive a personal vehicle for 4 months and no intention to sell the vehicle during the study period;
- hold an unrestricted driver's licence in the States of New South Wales (NSW) or Victoria (VIC);
- be, or have permission from, the registered owner of the vehicle to participate in the study;
- reside in Sydney, regional NSW, Melbourne, or regional VIC with aim for half in each state and a 70:30 split for metro and regional areas;
- vehicle is sedan, coupe, hatchback, station-wagon or four-wheel-drive/SUV; and,
- drive at least 10 trips per week.

Instrumentation Installed

The Data Acquisition System (DAS) developed by Virginia Tech Transport Institute (VTI) for the SHRP2 project was installed in each vehicle to record driver, vehicle and other road user behaviors in normal and safety-critical situations. Each DAS unit incorporates multiple sensors (e.g., video cameras, a still camera, GPS, radar, accelerometers). A Mobileye system that includes a "smart" forward facing camera and proprietary algorithms was also installed to provide additional automated triggers for potential collisions.

Days/Months/Years of Observation per Participant

Four months of instrumented driving for each participant. Each participant used their own vehicle which ensured they were familiar with the vehicle and were encouraged to drive when, where and as often as they usually do.

High-Level Description of Analyses

Processes were developed that would allow researchers to download and prepare the data in an efficient manner (over 60TB of data resides in a server in the USA) and to house the data on Australian servers for quicker access. Scripts were developed to summaries, visualize, and reduce the large quantity of trips to a more manageable format. This allowed researchers to pinpoint trips of interest depending on the research question and to test the feasibility of possible research questions that the data may be able to answer. These scripts have also been used to identify errors in the data and new scripts have been written to correct or mitigate these deficiencies where possible.

Driver Distraction

Distracted driving is one of the key research areas that has been addressed using ANDS data. Analyses have examined engagement in secondary tasks during everyday driving and the role that various driver characteristics and driving context variables play in influencing the initiation of secondary tasks. These analyses have utilized a small subset of ANDS data (186 randomly selected trips, equating to 2623 min of driving). These trips were manually coded using the four video camera views provided by the DAS. The categorical variables coded for each secondary task event included: secondary task type, passenger presence, driving context, self-regulatory behavior (task interruptions) and any incidents that occurred while the driver was engaged in the secondary task.

The detection and labeling of distraction events as described above were done manually and required watching ANDS videos which is expensive and time-consuming. The use of state-of-the-art machine learning techniques could automate the distraction coding time. Reducing/annotating a random sample of the videos and using it to train a machine learning model to carry out the data reduction for the rest of the dataset is possible. We used pre-trained deep neural networks and random forest to detect driver distraction in the ANDS video data (Elhenawy et al., 2019) (see Fig. 2).

Safety at Intersections

The ANDS data have also been used to analyze how infrastructure designs can affect road user behavior, and in particular speed. Speed data collected during the ANDS were analyzed to identify potential correlations between drivers' free-flow travel speed

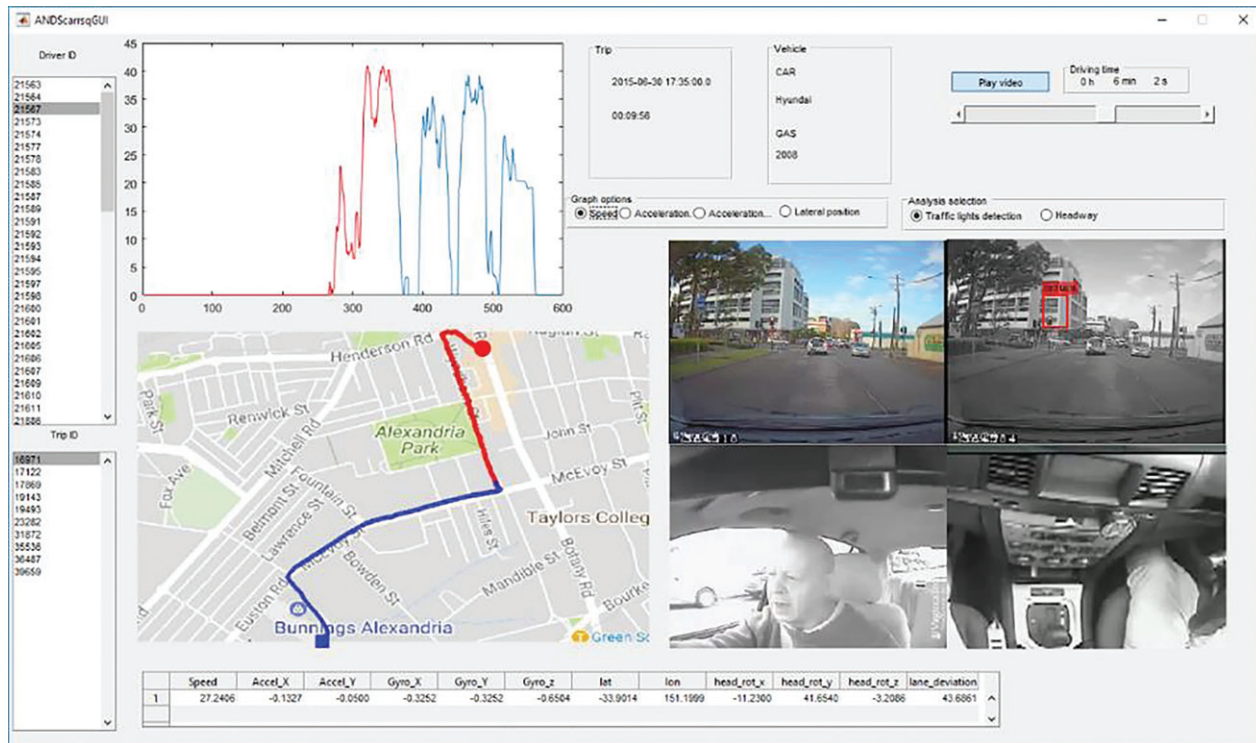


Figure 2 Visualization of ANDS data (traffic light state detection—top right quadrant).

through common types of intersections in Australia and the intersection design characteristics as well as to evaluate the associated level of fatal and serious injury crash risk incurred by drivers. The analysis was based on a sample of 93 intersections in Sydney.

Further analysis is looking at the following:

- Speed choice including compliance with speed limit value and speed choice in the context of factors like lane width and traffic calming devices, and situations in which speeding increases crash risk, and by how much.
- Driver headway including changes of time and distance from the vehicle ahead in different driving contexts.
- Interactions with vulnerable road users including investigation of crash and critical incident causation factors; gap and speed choice when passing cyclists; and anticipating pedestrian behaviors at crossings (lane and speed choice; when turning right and left).
- Fatigue including investigation of level of fatigue as a function of distance traveled and time of travel; degree of fatigue versus type of driving (urban versus rural) and analysis of the effects of fatigue on driving performance
- Interactions with intelligent transport systems (ITS) including driver responses to collision warning and crash avoidance technologies.

Main Findings

We have been able to identify and summaries all 192,000 trips recorded during the study, allowing researchers to understand driving habits, including mean trip durations and mean distances. We have also devised methods, including machine learning, to tackle and understand such a large dataset (over 60TB), such as pinpointing particular events of interest, trips and finding and correcting anomalies within the trip data.

Key findings to emerge from the distracted driving analyses are:

- Secondary task engagement is highly prevalent among Australian drivers, with drivers spending approximately 45% of their total driving time engaged in nondriving tasks.
- Older Australian drivers engage in secondary tasks just as frequently as their younger counterparts, but these tasks tend to be shorter and less risky than task engaged in by younger drivers.
- Drivers make strategic decisions regarding when to engage in secondary tasks, such as waiting until the vehicle is stationary; however, they do not appear to consider some contextual factors that may impact risk, such as weather and light conditions.
- Once engaged, drivers interrupt (or temporarily disengage from) only a small percentage (13.5%) of secondary tasks, with 87% of these tasks interrupted to re-engage in the driving task.

Key findings from the use of machine learning to detect and label distraction events.

- The proposed algorithm, when applied to the dashboard camera, achieved 0.609, 0.218, and 0.325 for true positive rate (TPR), the false positive rate (FPR) and the precision respectively. Moreover, for the face camera, it achieved 0.748, 0.344, and 0.651 for the TPR, the FPR and the precision, respectively. These promising results suggest the use of machine learning has potential to automate data reduction tasks from videos from naturalistic studies (Elhenawy et al., 2019).

Key findings from the analysis of travel speeds through common types of intersections in Australia (Mongiardini et al., 2020) include:

- The type of intersection (i.e., roundabout, priority, signalized) is the main design parameter to be found to have a statistically significant influence on driver speed behavior.
- Roundabouts are characterized by average and 85th-percentile travel speeds considerably lower than priority and signalized intersections and they are the only type of common intersection designs that appear to be capable of guaranteeing a low level (<10%) of fatal or serious injury risk.

Lessons Learned

The main lessons from this project are that projects of this nature are expensive and require considerable and continuing financial support. Specific statistical and computing skills are required to manage and query the large and complex datasets of heterogeneous measurements, including geospatial databases. Nevertheless, NDS provide a valuable opportunity for in-depth analysis of how normal drivers interact with their vehicle and the road system which can be used into the future.

Funding Sources and Related Issues

The ANDS is led by TARS Research Centre at UNSW Sydney and involves road safety research centers from three other universities: Monash, Adelaide, and Queensland University of Technology. This consortium obtained government funding totaling around \$3.5 million and \$1 million in-kind, in partnership with road authorities and industry partners.

Other Issues

The dataset is intended to be a resource to answer a multitude of road safety questions; however, there are some caveats on the usage of the data. All partners of the consortium can apply to a governance committee to conduct analyses using the ANDS dataset. Other parties can, in collaboration with members of the consortium, also conduct analyses once permission is obtained from this committee. All researchers must sign and comply with a Data Confidentiality Agreement.

Canada

Main Research Question

The overarching goal for the Canadian Naturalistic Driving Study (CNDS) was to produce a Canadian database for research by a broad range of users with varied capabilities and interests including transportation researchers, academic researchers, government researchers both federal and provincial/territorial, industry and others. The CNDS was conducted using the SHRP2 NDS approach to instrumentation and procedures with the result that the Canadian data can function as an additional data collection site and can be combined or compared with the six US sites on many dimensions. While there are commonalities across all seven sites, there are also a number of important differences with respect to issues such as distracted driving laws.

Some examples of specific research questions were:

- What are the contributing factors to crashes and near-crashes?
 - What are the differences between different roadway characteristics? Urban roads and rural highways?
 - What is the impact of winter conditions?
- What is the prevalence of risky behaviors? Under what conditions do they occur and which drivers are involved?
 - Speeding
 - Drowsiness
 - Impairment
- What can we learn about driver distraction and inattention?
 - What is the incidence of secondary task engagement?
 - Which age groups of drivers are most/least likely to engage?
 - What are the devices and tasks?

- When do they engage?
- How does the behavior of the Canadian drivers compare with that of the US drivers, given the differences in distraction-related legislation?

The ultimate goal is to propose countermeasures based on the findings.

Sample Size and Inclusion Criteria

CNDS data were collected on passenger vehicles and trucks.

Passenger Vehicle Participants

There were 149 passenger vehicle participants made up of teen drivers (aged 18–19), young adult drivers (aged 20–29), middle-aged drivers (aged 30–64), and senior drivers (aged 65 and older). Drivers' own vehicles were instrumented. A total of 150 vehicles were instrumented since a few drivers were replaced. Drivers participated for 24 months, 18 months, or 12 months and data were collected from June 2013 to October 2015 (Klauer et al., 2018).

Truck Participants

There were 25 commercial truck drivers who completed the study. Of the 25 trucks that were instrumented, 5 were driven on long routes (7–10 days on and 3–4 days off) and 2 were double-trailers. Questionnaires and assessments were those used in other commercial vehicle NDS and included driver abstracts and driver logs. Truck data were collected for approximately 12 months. Drivers were recruited through local trucking firms (Sarkar et al., 2018).

The CNDS was based in the city of Saskatoon located in the Canadian province of Saskatchewan. This location had a major advantage for recruitment as the participants were recruited through the province of Saskatchewan Government Insurance (SGI) program. They sent letters to target populations of interest based on desired characteristics (age, gender, postal code, commuter into Saskatoon). Some additional participants were recruited through fliers and media announcements. They were paid \$450 CDN per year for their participation.

Instrumentation Installed

The CNDS used the same DAS that was used in the SHRP2 project with the notable exception of a right over-the-shoulder camera which provided a better view of drivers' hands and their interactions with technology. Truck instrumentation was similar to that in passenger vehicles. The DAS system collected four full-time video streams of forward view, driver view, over the driver shoulder/center stack and right rear view as well as periodic still cabin images that were blurred to preserve passenger anonymity (Fig. 3)

A wide range of vehicle data were collected including accelerometer data, GPS, forward radar, turn signals and vehicle network data, including accelerator, brake pedal activation, ABS, steering wheel angle, seat belt information, airbag deployment, etc. An audio incident recording system enabled the driver to provide a brief account of an event that was useful when reviewing video events later.

At the beginning of the study, drivers completed an extensive battery of assessments and questionnaires addressing functional abilities (perception, cognition, psychomotor, physical), psychological testing (risk taking, sensation seeking), health including medical conditions and medications, sleep factors, safe driving knowledge and driver history.

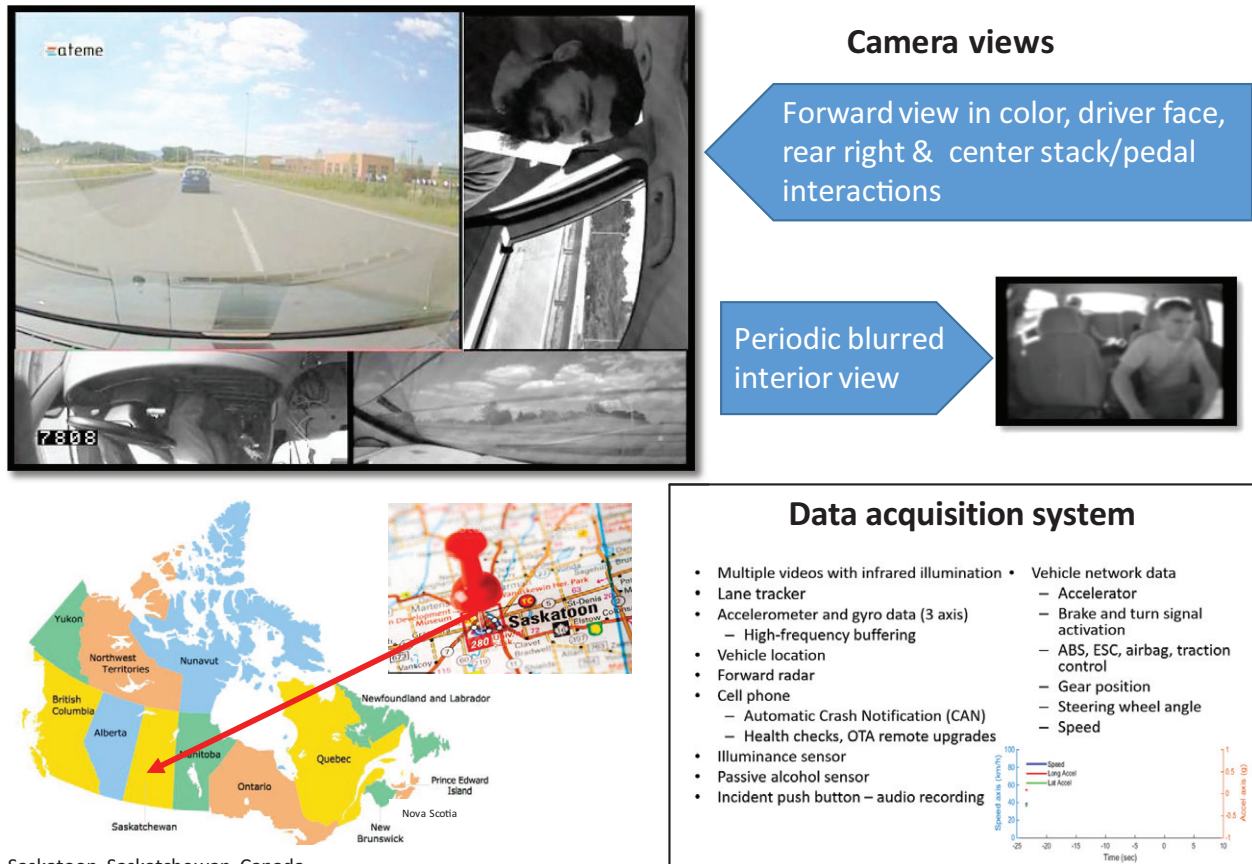
Days/Months/Years of Observation per Participant

Data were collected for the full duration that drivers drove their vehicles. Passenger vehicle drivers drove their vehicles for 12, 18, or 24 months. Truck drivers drove the fleet trucks for approximately 12 months.

High-Level Description of Analyses

Details of data treatment and reduction procedures:

- Postprocessing of data involved multiple steps involving quality control, filtering over data spikes, detection of spurious data.
- Summaries of driver demographics and driving history; physical and psychological test results.
- Summary measures for trip descriptions: trip length, duration, start time, stop time, minimum and maximum for speed and acceleration.
- Vehicle information such as vehicle age, type and condition, and available technologies.
- For video data: Manual Video coding and annotation; Distraction coding.
- Running event triggers to detect crashes, near crashes, type and severity.
- Produced baseline event records necessary for analyses.
- Several data analysis tools developed to enhance data exploration on the website.
- In addition, several Webinars have been provided for CNDS and SHRP2 data users. They have covered many topics including website use and data analysis methods.



The CNDS INSIGHT website provides a full description of the assessment and testing batteries, vehicle data captured and is available at <https://insight.canada-nds.net/home>. The website provides project background documentation, aggregated and summary graphs and access to the Data Dictionaries to assist users in planning their data analysis approaches. The data and associated documentation for the Canadian Natural Truck Driving Study (CNTDS) are available at <https://insight.canadatruck-nds.net>.

Main Findings

Summary of data for passenger vehicles:

- 83 crashes and 301 near-crashes
- 1,904,813 vehicle kilometers traveled
- 53,718 vehicle hours traveled

Summary of data for trucks:

- 34 crashes, 60 near crashes
- 1,304,000 vehicle kilometers traveled
- 27,500 vehicle hours traveled

Some differences between US and Canadian drivers:

Legislation Differences between Canada and the USA (Han *et al.*, Under Review):

- Secondary task involvement was a contributing behavior to more crashes for US drivers (63.1%) than Canadian drivers (44.7%). However, the Canadian findings were higher than would be expected considering the more stringent Canadian distraction laws.

Distraction & Cell Phone Use (Klauer *et al.*, Under Review):

- Overall, hand-held cell phone use is much lower for Canadian drivers (1.5%) than for US Drivers (5.9%).
- Overall young adult Canadian drivers have much less cell use engagement compared with teen drivers in Canada; whereas in the US, the teen and young adult drivers have similar patterns of cell phone use.
- Canadian teen drivers are the only driver age group that is similar to US drivers in that their cell phone use is similar to US teen drivers.

Crash Types & Contributing Factors (Harbluk et al., 2018):

- The Canadian crashes were most often rear end striking (34%), followed by intersection (22%) and run off road (8%) crashes.
- As might be expected, slippery road conditions contributed to more Canadian than US crashes (23.1% vs US 1.8%).

Lessons Learned

Early on, it became apparent that part of the NDS work would involve educating people on this research approach, where it fits in with other road safety research approaches, and what the many benefits would be not only in the immediate but also longer term. The NDS approach made it possible to ask new types of questions about road safety and to greatly improve the understanding of driver behavior.

Prior NDS work conducted in the 100 Car Study, SHRP2 and the European NDS projects provided a strong basis for the CNDS proposal. Aligning the Canadian project with the SHRP2 and adopting its design, equipment, implementation and data collection and reduction methods was efficient. The benefits continue as the CNDS database is used by researchers since the SHRP2 data provide opportunities for combining and contrasting data in analyses to address research questions.

NDS datasets are large, complex and must be kept secure to protect the integrity of data and the privacy of the participants. There can be considerable costs associated with this. Researchers who use the data must consider in their planning the associated costs and procedures for data access and processing if required.

In addition:

- Cell chip differences were found between those used in the US and those in Canada. The chips in the DAS needed to be replaced and reintegrated.
- Shipping across the border was subject to considerable paperwork & delays.
- Although driving was to take place only in Canada, GPS data clearly indicated that one truck driver did not follow this requirement and drove into the USA.

Funding Sources and Related Issues

The CNDS was conducted under auspices of Council of Deputy Ministers Responsible for Transportation & Highway Safety (<https://comt.ca>) with funding support from the Canadian Federal Government (Transport Canada), the provinces and territories, and the Canadian Council of Motor Transport Administrators (CCMTA).

The success of the project was due to the work and cooperation among:

- Council of Deputy Ministers Responsible for Transportation and Highway Safety
- Transportation Association of Canada (TAC)
- Canadian Council of Motor Transport Administrators (CCMTA)
- Saskatchewan Government Insurance (SGI)
- Saskatchewan Highways
- University of Saskatchewan Psychology Department
- SaskTel
- Transport Canada
- Second Strategic Highway Research Program (SHRP2)
- Transportation Research Board
- Virginia Tech Transportation Institute (VTI)
- Federal Motor Carrier Safety Administration (FMCSA)
- And the many drivers who participated

Other Issues

There are general challenges for NDS projects since they are very large, complex, run over long durations etc. There are specific concerns associated with data collection, use and analysis:

- There are challenges associated with the possibility and ease of data sharing. While there is considerable value in sharing data and conducting analyses across projects, in reality this can be quite difficult for various reasons due to differences in variable definitions, methods of data reduction, etc. The challenges become greater when sharing between countries.
- A general issue when using large data sets such as those from NDS, is for users to have sufficient and appropriate understanding of the data (and their limitations) and knowledge of appropriate analysis methods and how to apply them to the data.
- Working with human data requires researchers to respect and adhere to the policies that are necessary for privacy and security requirements. These requirements are essential to keep our commitment to those who participated in the study, but they may be a new requirement for those not used to working with human data. These requirements are likely to differ from country to country and these requirements are an evolving area within countries themselves.

China

Main Research Question

The overall objective of the study was to evaluate driver behavior, traffic conditions, and traffic safety in Shanghai, China.

Sample Size and Inclusion Criteria

The Shanghai Naturalistic Driving study (SHNDS) was a collaborative research study conducted by the Virginia Tech Transportation Institute (VTTI), the General Motors Company, and Tongji University. The SHNDS was the first large-scale NDS conducted in China using the same data collection system employed in the Second Strategic Highway Research Program (SHRP 2) NDS. The study included five experimental vehicles that collected data over a period of 3 years.

The study included 60 drivers from Shanghai, China. Each SHNDS participant drove one of the five experimental vehicles. To ensure that the data represented regular Shanghai driver behavior, the study excluded professional drivers or taxi drivers and avoided participants from the automobile industry. Other inclusion criteria including owning a vehicle, driving every day, as well as with a reasonable number of years of driving experience.

Instrumentation Installed

The SHNDS instrumented five research vehicles, including two Buick Lacrosses, two Chevy Cruzes, and one Cadillac DTS. The study used the same DAS developed by the VTTI for the SHRP 2 NDS project. The multichannel video configuration was similar to the SHRP 2 NDS, with front, back, driver's hands, and face views. The backup camera was updated to a higher resolution at later stage for better identifying following vehicle manoeuvres. The DAS continuously collected key driving information, including GPS, radar, three-dimensional acceleration, yaw rate, vehicle brake information, etc., from ignition-on to ignition-off.

In addition to driving data from the DAS, the data collection system also recorded information from a commercially available off-the-shelf Mobileye active warning system (AWS). The Shanghai NDS used an A-B design approach with 1 month of the AWS muted (i.e., not showing visual or audio alerts to drivers, but still functioning and recording in the background) followed by 1 month of the AWS active (i.e., showing active visual and audio alert signals to drivers).

Days/Months/Years of Observation per Participant

Each participant drove an experimental vehicle for 2 months. As noted, the AWS was off during the first month and was turned on during the second month. No specific instructions were provided and drivers drove the vehicles as they normally would. Note: due to holidays and other disruptions, the data collection period did vary by participant.

High level description of analyses:

The SHNDS data collected over 5000 h of continuous driving data from 60 drivers. The study identified 14 crashes and 115 near-crashes. Several subsequent projects were conducted, including both proprietary research and research in the public domain.

Main Findings

There are unique characteristics of driver behavior and traffic conditions. For instance, mopeds account for a considerable share of road traffic in Shanghai and a disproportionately large share of severe crashes. Glaser et al. (2017) analyzed the characteristics of moped-car conflicts and identified patterns of on-road vehicle-moped conflict configurations in Shanghai, suggesting that the conflicts could be attributed to the relatively complex traffic environment and risky behavior of moped riders. The study also indicated lower prevalence of hard steering in Shanghai compared to the United States. The Shanghai NDS also been used to evaluate the effects on mobile phone use on drivers' control behavior and lane-change behavior.

Lessons Learned

This multipartner international NDS brings challenges for logistics management and technical support. The unique traffic characteristics and driver behavior associated with the study may require tailored DAS configuration. A thorough evaluation of these characteristics can guide the data collection effort to maximize the research outputs.

Funding Sources and Related Issues

The SHNDS was jointly funded by the General Motors Company, Tongji University, and the National Surface Transportation Safety Center for Excellence at VTTI.

Other Issues

The SHNDS is based on a convenience sample of young-to-middle aged, experienced drivers in a Chinese megacity. The data might not represent the entire driver population and traffic conditions found in China.

European Union

Main Research Question

Stating one main research question in a typical naturalistic driving study is difficult, as the study often aims to provide data for a wide range of research areas. In UDrive, analyses targeted the areas of everyday and risky driving, distraction and inattention, vulnerable road users, and eco-driving. There was one subproject in UDrive per research area, with different researchers (although with some overlap) in each—a typical arrangement for projects funded by the European Commission, as EC calls are typically quite specific about the areas to be addressed. Furthermore, data from cars, trucks and powered two-wheelers (PTWs) were collected in UDrive; each vehicle modality had specific research questions.

Some examples of research questions in UDrive were:

Risky driving

- What environmental factors influence engagement in risky driving behaviors?
- How does traffic culture or country influence engagement in risky behaviors?

Driver distraction and inattention

- Which secondary tasks do truck and car drivers engage in? When, how often and for how long are they performed?
- Do driving task complexity and secondary task complexity influence the decision to engage in secondary tasks?
- Do drivers adapt their safety margins when performing secondary tasks?

Vulnerable road users

- Which factors influence whether car drivers perform a shoulder check before a right turn (UK: left turn) in an urban intersection?
- Which factors influence the lateral distance maintained when a car starts to overtake and passes a cyclist?
- What are the time headways between cars, trucks and powered two-wheelers?

Eco-driving

- When do drivers brake, and is it necessary to brake in each instance?
- Do drivers shift gears in order to avoid high engine speeds and high fuel consumption?

One general research question for the analyses of the car data in UDrive (applicable to many of the specific research questions) was: What are the differences across the five countries where data collection was performed, with respect to (some performance indicator)?

Sample Size and Inclusion Criteria

In UDrive, 287 drivers/riders drove a total of 192 cars, trucks and powered two-wheelers. Data was collected from approximately 230,000 trips with a total duration of more than 50,000 h (2.4 million kilometers). There was one main driver per vehicle, while the remaining participants were secondary drivers who drove the vehicles less.

A total of 120 cars, with 192 main and secondary drivers, were instrumented. The cars were driven in France (30 cars/45 participants), Germany (20/30), the Netherlands (10/33), Poland (30/31), and the United Kingdom (30/53). The participants in France, Germany, Poland, and the UK drove their own cars with the UDrive DASs installed for a duration of one to two years. The participants in the Netherlands drove cars leased by a UDrive partner and provided free of charge. This leasing option was created to increase the number of participants and get data from the Netherlands. All the cars were Renaults (Clio or Megane).

Thirty-two Volvo distribution trucks (48 main and secondary drivers) drove in the Netherlands. The trucks belonged to five different fleet owners and drove on both urban and rural roads (including highways). Finally, 40 Piaggio scooters (47 main and secondary drivers) were driven primarily in a small city in Spain.

Because the project was limited to using only two Renault car models, Volvo trucks and Piaggio scooters, the inclusion criteria were somewhat different across data collection sites (e.g., adherence to criteria were relaxed as there were substantial problems finding drivers of the particular Renault models in Poland and Germany).

Instrumentation Installed

Because there was no suitable off-the-shelf DAS, one was developed as part of UDrive. The installed system had different configurations across the three vehicle modalities, but the same base unit. In its most comprehensive configuration, it collected up to eight video streams simultaneously. The quality, resolution, and sample rate were different for the different cameras—optimized for the specific purpose of the individual video views. For cars and trucks, the DAS collected data from the vehicle's controller area network (CAN) and a MobileEye "smart camera". All vehicles were instrumented with complementary accelerometer, angular rate sensors and GPS.

For cars and trucks, approximately 270 measures were collected, while for the PTWs, there were only around 30 measures, as no CAN-bus or MobileEye data were collected. In addition to the collected data, approximately 50 map attributes were added to the data in postprocessing, and throughout the UDrive analysis an additional 400 measures were created (across all annotations, researchers and analyses). All data were encrypted at the source (in the DAS) and in all data transactions.

Days/Months/Years of Observation per Participant

Most car drivers drove their own instrumented vehicle for 1–2 years. However, drivers in the Netherlands used leased vehicles for approximately 6 months. Most secondary drivers drove substantially less than the primary drivers.

The trucks were mostly shared by the fleet drivers, who typically drove the vehicles for between 1 and 2 years. A subset of the scooter riders had the Piaggios for 1 to 2 years, while the others were used by two sets of drivers (two sequential “waves” of data collection), who rode them for less time.

High-Level Description of Analyses

A substantial part of the analysis in UDrive was dependent on reducing video data to numerical or categorical data through manual annotation. This process varied from annotating entire trips with respect to a range of predefined secondary tasks performed by the drivers to noting down event data, such as weather and road conditions and drivers’ glance behaviors before and during a right turn (left in the UK). In addition to the manual processing, algorithms were developed to automatically extract specific events, such as events in which a car is coming up to a bicyclist on a rural road (using data from the MobilEye system). However, the extraction algorithms were never perfect, so manual confirmation of each event was needed.

Typically, all analysis started with plotting the data, but after that, the researchers in UDrive used a wide range of analysis techniques, ranging from simple descriptive statistics to linear mixed-effect models and machine learning. Details of the analyses can be found in the four UDrive deliverables (Carsten et al., 2017; Dotzauer et al., 2017; Heijne et al., 2017; Jansen et al., 2017).

Note that a tool for managing the data, handling the manual annotations (data-reduction process) and calculating derived measures, events etc., was developed in UDrive. All researchers working with the data used the tool, be it for annotations or writing algorithms in scripts (all in Matlab). The process was such that analysts developed algorithms (e.g., derived measures or filters to extract events) on a small set of the data/trips. Then, approximately weekly, the changes were “pushed”; all scripts were applied to all trips in the database. Statistical analysis and, for example, modeling were then performed on the results (e.g., calculated performance indicators), but independently of the tool processing the data (e.g., using software such as R, Python, Matlab, SAS, or SPSS).

At the end of UDrive we asked all researchers involved in the analysis to estimate what proportion of their UDrive analysis-time was spent on (1) data preparation (excluding manual annotations), (2) interpretation and documentation, and (3) statistics and modeling. The results showed that approximately 75% (SD 10%) of the time was spent on data preparation, while the remaining 25% was split evenly between the other two categories. This is worth keeping in mind when budgeting for naturalistic data projects.

Main Findings

The following describes a selection of the key results:

- There were substantial differences by country in the time drivers used their mobile phones while driving, ranging from almost never (close to 0% of the time in Germany) to approximately 10% (in Poland). Furthermore, the German drivers only performed secondary tasks for 2% of the driving time, while the Polish drivers performed secondary tasks almost 20% of the time.
- Many car drivers overtake bicyclists riding on the roadway with smaller lateral clearance than is recommended, or even legal.
- Car drivers from the Netherlands check their blind spot prior to making a right turn six times more often (27%) than car drivers from France, Poland, or the UK.
- Unexpectedly, car drivers do not keep a shorter time-headway to a powered two-wheeler than to another car.

Lessons Learned

For UDrive, an important limitation was the use of Renault cars exclusively. Including a wider range of vehicle brands would have substantially improved the generalizability of the results. The main reason other brands were not used was lack of access to the CAN-bus data, so having more car manufacturers involved to provide access would have been good. Previous European NDS, such as the EuroFOT project (a Naturalistic Field Operational Test of active safety systems funded by the European Commission that ended in 2012) (Benmimoun et al., 2013), did have more car manufacturers involved in the project, and got more access to CAN-data. However, an issue in EuroFOT was the very limited sharing of data between test sites, making aggregation difficult and reducing statistical power. That is, EuroFOT collected data from more car brands, allowing for more detailed CAN-bus data collection, but the data from each brand was typically collected and analyzed by project partners with which the car manufacturers already had an established collaboration (i.e., “trusted” partners). This compartmentalization prevented in-depth comparisons across countries of, for example, behaviors. However, since the focus of EuroFOT was on system evaluation, this limitation was less of a problem than it would have been for UDrive.

Furthermore, the relatively low number of vehicles and participants per country in UDrive lowered the statistical power of the analyses greatly. Also, even if the number of kilometers driven was high, it was not high enough to include many near-crashes, let alone crashes. If I were to design the study today, I would split it into two: one with everyday driving data with multiple cameras, such as in UDrive; the other study would install smaller, much less expensive event data recorders with video in a much larger set of vehicles (Lotan et al., 2017). The latter study would increase the number of participants for which data is collected (increasing the

statistical power for at least part of the analysis), and capture at least a fair amount of near-crashes. Both studies together would provide a more complete picture of traffic safety than either alone.

Another aspect of data collection is collected data versus analyzable data. In UDrive, most of the collected car data could be (and was) used for analysis. However, for the truck and PTWs, there were significant issues, so an unexpectedly large amount of the data could not be analyzed. In fact, more than half of the collected truck data was unusable (not analyzable). The most common reason was that a positive driver confirmation could not be made: there was no image of the driver, so it was impossible to confirm that the person driving had consented to being part of the study. (Another reason was that even if a positive identification could be made, some camera view(s) were missing). Because the trucks were used by multiple drivers (drivers shared trucks at the company they worked), and not all drivers were part of the study, a “turn-off”-button was added; when it was pressed, nothing was recorded for that trip. However, instead of using the button, drivers damaged the cameras or covered them. We actually had a good on-line monitoring tool for detecting these issues, but because the trucks were not easily reachable (and the actual data collection team was far from “on their toes” in this regard), it could take weeks before the issue was fixed, and then the cameras were damaged again.

For future data collections using fleets, an important pre-requisite would be to make sure that all drivers are on-board with the data collection, and that the data collection team has processes in place for fast repairs. Note that the EuroFOT project also had issues with truck equipment being damaged, reducing the amount of analyzable data.

For the PTWs, the issue was different. Before enrolling in the study, all PTW riders stated how much they expected to ride. Many of the participants substantially overestimated the time they spent riding. Although some drivers had to return the PTW so other riders could be recruited, this issue resulted in a much smaller final dataset than expected for the PTW part of the study. It is recommended that the recruitment strategy and—even more importantly—the participant replacement process be improved to enable quick returns and participant replacements.

In all studies I have been involved in, the DASs have been a bottleneck. The development or integration of the system has taken more time than expected, and piloting the data acquisition has often been cut short due to time constraints. However, there have typically been good reasons for not choosing off-the-shelf solutions. Consequently, I believe that there is a need to allow substantially more time for piloting. Funding agencies must appreciate that need and allow projects to have longer durations.

The tool for annotation and data processing in UDrive was set up in a way that I found problematic. Admittedly, it had the benefit of giving everyone access to all measures, events and annotations, and it had good traceability in terms of demonstrating exactly how each derived measure was calculated from other measures (and/or events), even for chains of derived measures. However, having worked with an alternative (in the EuroFOT project) where data was processed as needed, and scripts, rather than processed data, were shared, I strongly believe in a more decentralized data processing solution. That is, analysts can choose any tool of their choice to interact with the data, and, possibly, when needed, share scripts for specific algorithms and event extractions.

Based on my experience working with NDS data from both Europe and the US over the last 14 years, I strongly recommend making sure that there is some forum or database where quality issues are documented so that they can be easily found and referenced and acknowledge (and promoted) by the organizations handling the data. The documentation must be available to all current and potential users of the data. Typically, smaller experiments and data collections in the domain of traffic safety are performed by very few users, who “know their data”.

However, as NDS data are often used by many researchers in many different organizations over several years, there is a danger that these other users could mistakenly assume that all the data are high-quality. Typical issues include synchronization and sensor bias (see, e.g., Appendix A. Supplementary data in Bärghman et al., 2017), as well as the correctness of the data tracking surrounding road users (e.g., from radar or smart camera data). All three issues can be found in UDrive data—as well as in, for example, SHRP2 data. However, the danger does not lie only in the (possibly poor) quality of the data. If others have created derived measures or extracted events, it is easy to think that they represent a certain thing when they actually represent something different (or may even be incorrect). Without extensive documentation being available, or, even better, talking to the individuals who made the calculations, there is a substantial risk that new users will make incorrect assumptions about the data.

Funding Sources and Related Issues

UDrive won a European Commission grant proposal for €8.0 million in funding for person months. UDrive partners (primarily industry) contributed in kind with an additional €2.7 million. This funding was split among the 19 partners. Additional funding was available for DAS installation, and data acquisition, maintenance and accessibility within the project. However, a main issue (at least so far) for projects funded by the European Commission is that there is no European-level funding mechanism to manage data and keep it available for researchers after the end of the project. As a result, individual organizations have to seek national funding or finance the management and access themselves; as a result, the data are less available to other organizations, valuable knowledge about the details of the dataset can be lost, and—in the worst case—the organizations have to delete costly parts of the dataset or shut down infrastructure.

Other Issues

In a unique approach to data access in the UDrive project, one of the partners (SAFER traffic safety research center) managed all the data (from all data collection sites) and set up a remote access platform. Authorized IP addresses had access to the data and analysis tools, while all processing and calculations were run at SAFER. That is, all involved partners used virtual desktops to do everything

from calculating performance indicators to watching and annotating videos of the drivers (without actually downloading any data). A process for extracting any results (figures, tables etc.) was in place—with the extraction handled centrally.

This remote access platform was the result of substantial discussion and effort that were put into developing the Data Protection Concept (DPC); see [Gellerman et al., 2016](#) for a more detailed overview) (Gellerman, Svanberg and Kotiranta, 2016). The DPC considered data protection, and the associated risks, across the entire data management chain. It also outlined a process that would allow access to research on the UDrive data on equal terms (equal trust in all) for all partners. To gain access, UDrive partners needed to provide a document describing how they handle digital and physical access to the data (through the virtual desktops). This is different from, for example, how US SHRP2 data are accessed: sensitive data (e.g., videos of the participants) can only be accessed by on-site visits to Virginia Tech Transportation Research Institute and a specific site of the Federal Highway Administration (FHWA).

The DPC can be used as a starting point for future (at least European) naturalistic driving studies' data access protocols and processes. However, the General Data Protection Regulation (GDPR) has since become a reality, changing some of the fundamentals of conducting NDS in Europe. For example, deidentification through automated face removal is likely a prerequisite for conducting new large NDS across at least parts of Europe—particularly for the external views (pedestrians, bicyclists etc.). In UDrive we experimented with de-identification, but did not find a good enough tool. When required (depending on the laws of individual countries) we blurred part of the screen instead, but that degraded the usefulness of the data. However, automated driver identification (to verify that the person is among those who provided written consent) would have substantially reduced the efforts needed in pre-processing the data. In UDrive, humans watched still frames from every single trip to make this verification. (Data from other drivers were purged from the database before it was accessible for analysis.) Car license plate de-identification (removal or blurring) is also needed in several European countries. In addition, GDPR increase the requirements for informing each individual driver about each new study done on the data, even if they have signed an informed consent form agreeing to such future research. Given that naturalistic driving data has a very wide range of uses, and is intended to be used for many years, this is a complication that needs to be handled when working with naturalistic driving data.

United States

The primary goal of the SHRP2 Safety research program was to address the role of driver behavior and performance with the goal of improving road safety. To make progress, it is necessary to understand drivers, their vehicles and the driving environment, and most importantly, the interactions of these factors and their impacts on safety. The approach taken was to conduct the SHRP2 naturalistic driving study that was unprecedented in size and scope and will be used to support safety research for many years.

Main Research Questions

The main goal of this study was to collect and compile detailed naturalistic driving data across a large and geographically diverse sample. These data are valuable to transportation safety researchers allowing them to explore how people drive under everyday conditions and also to document the activities or behaviors that precede crashes or near misses, enabling determination of crash risk.

The SHRP2 dataset is useful for research in a broad range of areas including:

- driver behavior and characteristics (e.g., distraction, aggression, speeding, teen drivers, older drivers medical conditions . . .)
- development of driver training programs, education and policy
- improved road design and infrastructure (e.g., road curvature and intersections, rural roads, lighting)
- vehicle systems and development of advanced vehicle safety features
- countermeasure development
- and much more

Sample Size and Inclusion Criteria

More than 3500 drivers participated, ranging in age from 16 to 98. Data collection ran for a 3-year period with most drivers participating for 1 to 2 years. Participants needed to be at least 16 years old, licensed and drive at least 3 days a week. Drivers' own vehicles were instrumented and they received compensation of \$500 USD a year. The sample includes the primary drivers of the vehicles as well as some secondary drivers. The study was conducted at sites in six US states, Florida, Indiana, New York, North Carolina, Pennsylvania, and Washington providing data on a wide range of driving environments.

Instrumentation Installed and Categories of Data Collection

There were three categories of SHRP2 data collection, data on the driver characteristics using a battery of assessments and tests, vehicle sensor data and full time video. Each vehicle was equipped with an onboard DAS which collected continuous data.

Driver Descriptions & Assessments

Demographics & Driving Knowledge	Visual, Perceptual & Cognitive Abilities
Age, Gender, Education etc.	Vision & Cognitive Tests
Driving History Questionnaire	Conner's Continuous Performance Test
Driving Knowledge Survey	Clock Drawing Assessment
Driver's Participation Experience	Physical Strength Tests
Driver Exit Interview	
Psychological Tests	Health Conditions
Barkley's ADHD Screening test	Medical Conditions & Medication
Risk perception Questionnaire	Sleep Habits Questionnaire
Risk Taking Survey	Medical Conditions & Medication at Exit
Sensation Seeking Scale Survey	
Driver Behavior Questionnaire	

Vehicle Data Acquisition:

Multiple Videos	Illuminance sensor
Machine vision eyes forward monitor	Passive alcohol sensor
Machine vision lane tracker	Incident push button
Accelerometer Data (3 axis)	Video
Rate Sensors (3 axis)	Audio (only on incident push button)
GPS	Turn signals
Latitude, Longitude, elevation, time, velocity	Vehicle network data
Forward radar	Accelerator
X and Y position	Brake pedal activation
Xdot and Ydot velocities	ABS
Cell Phone	Gear position
DAS system health checks	Steering wheel angle
Remote upgrades	Speed
	Seat Belt Information
	Airbag deployment
	And many more . . .

Video Data:

The DAS system collected four full-time video streams of forward view, driver view, over the driver shoulder/center stack and right rear view as well as periodic still cabin images that were blurred to preserve passenger anonymity. Full time video enables detailed analyses of driver glance behavior, interactions in the vehicle as well as information about the surrounding driving environment.

High-Level Description of Analyses

After data collection, postprocessing of data involved multiple steps including quality control, filtering over data spikes and detection of spurious data.

The SHRP2 dataset is expected to be useful for researchers for decades. It contains:

- More than 35 million miles of continuous data from over 5.5 million trip files
- More than 1900 light-vehicle crashes, over 6900 near-crashes
- Baseline driving event records necessary for analyses
- More than 1 million hours of video
- Summary data provided on drivers, vehicles, trips & events
- Video coding and annotation, coding for distraction
- Time series information
- Several data analysis tools developed to enhance data exploration
- Data dictionaries
- Information on reduced datasets
- And more . . .

Main Findings

The SHRP2 data have been the data source for many published research projects and continue to be widely used. These projects have provided insights into a wide variety areas including driver behavior, crash risk, the driving environment, infrastructure, weather and technologies. A list with brief descriptions of projects is provided at: <https://docs.google.com/document/d/1GQxjCvz0ZQbCtxekbamLK0uKsEUINmUaSPWJS7t2fBk/edit?usp=sharing>

Lessons Learned

There are several reports published since the completion of SHRP2 data collection that provide useful overviews and discussions. Antin and colleagues provides the methods used to design this large-scale study, recruit participants and presents highlights of the collected data. Their discussion also covers the management of the study and its strengths and weakness (Antin et al., 2019).

A technical report from the project provided several considerations for future projects (Dingus et al., 2014). Human participant protection and ethical review procedures took much longer to complete than anticipated. These tasks were complicated by differing requirements at the multiple study locations. Additional delays arose from procuring and repairing DAS components and from design modifications to the DAS units.

Recruitment procedures required re-evaluation and modification during the participant recruitment process, especially those for specialty populations like young or old drivers. Careful planning but also adaptability and creativity were required to successfully manage a study with such a large number of participants. In NDS studies, data collection and reduction occur over long periods of time. Analyses that are conducted on partial data prior to complete data collection may lead to confusion or misleading results. This can also be disruptive to data management processes.

Funding Sources and Related Issues

The second Strategic Highway Research Program (SHRP2) is a collaborative partnership of the Federal Highway Administration (FHWA), the American Association of State Highway and Transportation Officials (AASHTO), and the Transportation Research Board (TRB).

Other Issues/Additional Information

Information about the SHRP2 data and data access is available at the SHRP2 InSight Data Access Website at <https://insight.shrp2nds.us> (Transportation Research Board of the National Academy of Science, no date). It provides a complete overview of the data set, including updates on new events, new trip summary measures, new enhancement projects, and new documentation as well as data sets that have been developed and are available for use. Links to webinars covering many topics including website use and data analysis methods have been provided for SHRP2 data users.

A citation repository for data sets derived from SHRP 2 NDS is available on the Dataverse at <http://dataverse.vtti.vt.edu/dataverse/shrp2nds>.

The Roadway Information Database (RID) is a companion database to the SHRP2 NDS that provides roadway elements and conditions. These two databases are linked, providing an unprecedented resource to study driver behavior within the context of actual roadway characteristics and driving conditions. More information on the RID is available at <https://ctre.iastate.edu/roadway-information-database-rid/>.

Now that data collection phase has completed, current projects and information can be found on websites supported by implementation partners. The Federal Highway Administration (FHWA; <https://www.fhwa.dot.gov/goshrp2>) and the American Association of State Highway and Transportation Officials (AASHTO; <http://shrp2.transportation.org/Pages/Safety.aspx>) are responsible for implementing the resulting products known as SHRP2 Solutions.

Synthesis of Lessons from across the Case Studies

These case studies represent a cross section of the largest and most sophisticated NDS that have been conducted to date. The issues raised by the authors may not reflect the views of all the investigators and project consortium members involved, but they provide an insight into the logistical and analytical challenges involved in conducting NDS. Some common points across case studies include:

- NDS demand substantial investment, both in terms of data collection and analysis
- Funding agencies may be unfamiliar with this type of research and require education
- Be prepared for problems in the field related to data collection
- If possible, allow for extensive pilot testing of data collection methods
- Allow flexibility in study execution and inclusion criteria once data collection is underway
- Once data collection is complete, data coding and analysis need to be rigorous and consistent
- Recognize there will be a time-lag between the time that data collection is completed and when findings are available

When considering whether to conduct an NDS, researchers may also find it useful to contact those who have been involved in previous studies to address specific questions about their approach

What is the Future of NDS?

NDS are increasingly used to understand human interaction with advanced driver assistance features. NDS methods are also being used to observe how highly automated vehicles interact with other road users and infrastructure. One example is work being conducted by the Queensland University of Technology (QUT) in Australia. They are using NDS during an on-road pilot of

cooperative intelligent transport systems (C-ITS), which will involve a Field Operational Test (FOT) of 6 vehicle-to-infrastructure (V2I) use cases with up to 500 privately-owned vehicles participating (Lewis and Rakotonirainy, 2020).

Canadian (CNDS) data are being used for research supporting advanced driver assistance systems and highly automated vehicle applications (Perez et al., 2016). For example, the Canadian and SHRP2 NDS data have been used to develop naturalistic driving data baselines for highly automated commercial motor vehicles in highway applications (Krum et al., Under Review). SHRP2 NDS data are also being used as experimental stimuli in laboratory based studies (Ehsani et al., 2020).

NDS data collection equipment continues to become increasingly miniaturized, portable or nomadic, and less expensive (Grimberg et al., 2020). These lower-cost data collection methods, such as smartphone applications or Bluetooth-paired devices, can automatically begin data collection when a participant drives a vehicle, can be programmed to record event-triggered video before and after an event, and continuously collect acceleration, braking and GPS data. These data collection approaches allow for larger sample sizes that are not constrained by those who live in close proximity to where vehicle instrumentation can be installed and maintained. They may also allow the participation of drivers who do not own a vehicle or may not receive permission to fit instrumentation into a vehicle (e.g., teenage drivers). Furthermore, it could extend study participation to those who are traditionally under-represented in NDS research such as minorities and those from lower-income households.

Hybrid studies combining data collection methods are also being conducted. In these studies, a small number of participants' vehicles are equipped with high-resolution NDS data collection equipment, while a larger sample use a simplified continuous data collection method, such as the data collection capabilities of smartphone. The findings of these two groups are combined to generate a more generalizable set of findings.

In summary, NDS provide an excellent addition to established research methods in transportation research, offering an unparalleled insight into road user behavior and providing much needed knowledge to improve road safety.

Acknowledgment

No funding sources to report.

References

- Antin, J.F., et al., 2019. 'Second strategic highway research program naturalistic driving study methods'. *Safety Sci.* 119, 2–10, doi:10.1016/j.ssci.2019.01.016.
- Bärgman, J., 2016. 'Methods for analysis of naturalistic driving data in driver behavior research'. Chalmers University. Available from: <http://publications.lib.chalmers.se/records/fulltext/244575/244575.pdf>.
- Benmimoun, M. et al., 2013. 'euroFOT: Field Operational Test and Impact Assessment of Advanced Driver Assistance Systems: Final Results', in *Proceedings of the FISITA 2012 World Automotive Congress*. Springer (Lecture Notes in Electrical Engineering), Berlin, Heidelberg, pp. 537–547. doi:10.1007/978-3-642-33805-2_43.
- Breslow, N.E., 1996. Statistics in Epidemiology: The Case-Control Study. *J. Am. Stat. Assoc.* 91 (433), 14–28, doi:10.1080/01621459.1996.10476660.
- Carsten, O. et al., 2017. 'UDRIVE deliverable 43.1', Driver Distraction and Inattention, of the EU FP7 Project UDRIVE (www.udrive.eu).
- Carsten, O., Kircher, K., Jamson, S., 2013. 'Vehicle-based studies of driving in the real world: The hard truth?'. *Accid. Anal. Prevent.* 58, 162–174, doi:10.1016/j.aap.2013.06.006.
- Dingus, T.A., et al., 2014. Naturalistic Driving Study: Technical Coordination and Quality Control. National Academies of Sciences, Engineering, and Medicine, Washington, DC. Available from: <https://www.nap.edu/download/22362> (Accessed: 26 July 2020).
- Dingus, T.A., et al., 2016. 'Driver crash risk factors and prevalence evaluation using naturalistic driving data'. *Proc. Natl. Acad. Sci.* 113, 2636–2641, doi:10.1073/pnas.1513271113.
- Dotzauer, M., et al., (2017) UDRIVE deliverable 42.1 Risk factors, crash causation and everyday driving of the EU FP7 Project UDRIVE. UDRIVE Consortium. Available from: http://doi.org/10.26323/UDRIVE_D42.1 (Accessed: 31 July 2020).
- Ehsani, J.P., et al., 2017. 'Teen drivers' awareness of vehicle instrumentation in naturalistic research'. *J. Safety Res.*, doi:10.1016/j.jsr.2017.10.003.
- Ehsani, J.P., Seymour, K.E., et al., 2020. Developing and testing a hazard prediction task for novice drivers: a novel application of naturalistic driving videos. *J. Safety Res.* 73, 303–309, doi:10.1016/j.jsr.2020.03.010.
- Ehsani, J.P., Gershon, P., et al., 2020. Learner driver experience and teenagers' crash risk during the first year of independent driving. *JAMA Pediatr.*, doi:10.1001/jamapediatrics.2020.0208.
- Ehennawy, M., et al., 2019. Detecting Driver Distraction in the ANDS Data using Pre-trained Models and Transfer Learning, in *8th International Symposium on Naturalistic Driving Research*.
- Fitch, G.M., Hanowski, R.J., 2012. 'Using Naturalistic Driving Research to Design, Test and Evaluate Driver Assistance Systems', in Eskandarian, A. (ed.) *Handbook of Intelligent Vehicles*. Springer, London, pp. 559–580. doi: 10.1007/978-0-85729-085-4_21.
- Gellerman, H., Svanberg, E., Kotiranta, R., 2016. 'UDRIVE Data Protection Concept', in *Proceedings of the ITS World Congress*.
- Grimberg, E., Botzer, A., Musicant, O., 2020. Smartphones vs. in-vehicle data acquisition systems as tools for naturalistic driving studies: a comparative review. *Safety Sci.* 131, 104917, doi:10.1016/j.ssci.2020.104917.
- Guo, F., 2019. Statistical methods for naturalistic driving studies. *Annu. Rev. Stat. Appl.* 6 (1), 309–328, doi:10.1146/annurev-statistics-030718-105153.
- Guo, F., Hankey, J.M., 2009. Modeling 100-car safety events: a case-based approach for analyzing naturalistic driving data: final report. Report. Virginia Tech. Virginia Tech Transportation Institute. Available from: <https://vtechworks.lib.vt.edu/handle/10919/7406> (Accessed: 21 July 2020).
- Han, S., et al., Under Review. Evaluation of Cell Phone Use Legislation While Driving: An International Comparison of Canada and the United States.
- Harbluk, J., et al., 2018. Crash Types & Contributing Factors: Canadian & US Naturalistic Driving Studies, in: *Seventh International Symposium on Naturalistic Driving Research*, Blacksburg, Virginia.
- Haynes, R.B., et al., 2005. *Clinical Epidemiology: How to Do Clinical Practice Research*.
- Heijne, V., Ligterink, N., Stelwagen, U., 2017. Potential of eco-driving. UDRIVE Deliverable D45. 1. EU FP7 Project UDRIVE Consortium.
- Jansen, R., et al., 2017. UDRIVE deliverable 44.1 Interactions with vulnerable road users, of the EU FP7 Project UDRIVE. UDRIVE Consortium. Available from: http://doi.org/10.26323/UDRIVE_D44.1 (Accessed: 31 July 2020).
- Klauer, C., Pearson, J., Hankey, J., 2018. An Overview of the Canada Naturalistic Driving and Canada Truck Naturalistic Driving Studies, in: *Transportation Association of Canada Annual Meeting*, Saskatoon, Saskatchewan, Canada.

- Klauer, S., et al., Under Review. *Canada Distraction Analysis*. Report by VTTI for Transport Canada.
- Klauer, S.G., et al., 2014. Distracted driving and risk of road crashes among novice and experienced drivers. *N. Engl. J. Med.* 370, 54–59, doi:10.1056/NEJMsa1204142.
- Knipling, R.R., 2015. 'Naturalistic Driving Events: No Harm, No Foul, No Validity', in *Proceedings of the 8th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design: driving assessment 2015*. Driving Assessment Conference, Salt Lake City, Utah, USA: University of Iowa, pp. 197–203. doi: 10.17077/drivingassessment.1571.
- Kopeck, J.A., Esdaile, J.M., 1990. Bias in case-control studies. A review. *J. Epidemiol. Community Health* 44 (3), 179–186, doi:10.1136/jech.44.3.179.
- Krum, A.J., et al., Under Review. *Naturalistic Driving Data Baseline for Highly Automated Commercial Motor Vehicle On-Highway Application*. Federal Motor Carrier Safety Administration, USDOT, Washington D.C.
- Lewis, I., Rakotonirainy, A., 2020. 'Ipswich Connected Vehicle Pilot'. Personal Communication.
- Li, G., et al., 2017. Longitudinal Research on Aging Drivers (LongROAD): study design and methods. *Injury Epidemiol.* 4 (1), 22, doi:10.1186/s40621-017-0121-z.
- Lotan, T. et al. (2017) The potential of naturalistic driving studies with simple data acquisition systems (DAS) for monitoring driver behaviour. report. Loughborough University. Available at: https://repository.lboro.ac.uk/articles/report/The_potential_of_naturalistic_driving_studies_with_simple_data_acquisition_systems_DAS_for_monitoring_driver_behaviour/9353960 (Accessed: 6 July 2020).
- Mongiardini, M., et al., 2020. 'Evaluate travel speeds and associated risk of casualty crashes through intersections in Australia using naturalistic driving data.', in: *Australasian Road Safety Conference*, Melbourne.
- National Institutes of Health, 2020. NOT-OD-20-075: Notice of Transition to New System for Issuing Certificates of Confidentiality for Non-NIH Funded Research. Available from: <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-075.html> (Accessed: 22 July 2020).
- Perez, M., et al., 2016. Transportation Safety Meets Big Data: The SHRP 2 Naturalistic Driving Database, 計測と制御, 55(5), pp. 415–421. doi: 10.11499/sicej.55.415.
- Sarkar, A., et al., 2018. Overview and Preliminary Analysis of Canadian Truck Naturalistic Driving Study, in: *Seventh International Symposium on Naturalistic Driving Research*, Blacksburg, Virginia, p. 15.
- van Schagen, I., et al., 2011. Towards a large scale European Naturalistic Driving study: final report of PROLOGUE: deliverable D4.2. report. Loughborough University. Available from: https://repository.lboro.ac.uk/articles/report/Towards_a_large_scale_European_Naturalistic_Driving_study_final_report_of_PROLOGUE_deliverable_D4_2/9353405 (Accessed: 10 July 2020).
- Simons-Morton, B.G., et al., 2014. Keep your eyes on the road: young driver crash risk increases according to duration of distraction. *J. Adolescent Health* 54, S61–S67, doi:10.1016/j.jadohealth.2013.11.021.
- Transportation Research Board of the National Academy of Science (no date) The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset: SHRP2 NDS Data Access. Available from: <https://insight.shrp2nds.us/> (Accessed: 23 September 2020).

Further Reading

For further details on how to design and analyse an NDS see:

- Klauer, S.G., Ehsani, J.P., Simons-Morton, B., 2016. Using naturalistic driving methods to study novice drivers. In *Handbook of Teen and Novice Drivers: Research, Practice, Policy, and Directions*. CRC Press, pp. 409–419.
- Further details about selected NDS:
- Jonathan F. Antin, Suzie Lee, Miguel A. Perez, Thomas A. Dingus, Jonathan M. Hankey, Ann Brach, 2019. Second strategic highway research program naturalistic driving study methods. *Safety Science*, 119, 2–10.
- N. van Nes, J. Bärghman, M. Christoph, I. van Schagen, 2019. The potential of naturalistic driving for in-depth understanding of driver behavior: UDRIVE results and beyond. *Safety Science*, 119, 11–20. <https://doi.org/10.1016/j.ssci.2018.12.029>.
- Bärghman, J., et al., 2017. The UDrive dataset and key analysis results. SAFER/Chalmers. Available from: <http://www.UDrive.eu>.

A Detailed Approach to Qualitative Research Methods

Sonja Forward, Lena Levin, Swedish Road and Transport Research Institute, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	39
Interviews	40
Preparation	40
Interview Protocols	40
Sample Size	41
Recruitment	42
The Interview	42
The Opening of the Interview	43
During the Interview	43
End the Interview	44
Focus Groups	44
Summary and Conclusion	44
References	45
Further Reading	45

Introduction

The interest in qualitative methods and research has steadily increased since the late 1960s. According to [Bryman \(1997\)](#) it is even older than that, since the Chicago school from the 1920s and 1930s can be seen as an early example of qualitative research. [McLeod \(2019\)](#) pointed out that the use of qualitative methods was a reaction to scientific studies in psychology, such as behaviorism. Behaviorism being more interested in human behavior and actions rather than how humans' constructs meanings and interpret different events.

In research a distinction is made between qualitative and quantitative research because they are based on different methods, which means that they relate differently to the topic of the study.

The goal of quantitative research is to collect as much data as possible, which is then manipulated through various statistical analyses. Data, in this case, can easily be converted into numbers. The goal is to explain the cause and effect (causality), have a sufficiently broad and representative material so that the results can be generalized using a well-defined approach, which can be replicated by other researchers. Qualitative research, on the other hand, is more interested in the research participant's own interpretation of the situation. The ambition is to present the human part of a story ([Jacob and Furgerson, 2012](#)). It recognizes that the social world is complex and dynamic. To achieve this, the researcher has to be actively involved with the participant and be prepared to accept that "multiple realities" exists ([Banister et al., 1994](#)). The goal is not to generalize the findings, instead it is regarded as situational and conditional ([Arksey and Knight, 1999](#)).

Qualitative research takes a phenomenological approach. This means that every understanding of social reality must be based on people's own experience of this reality. The phenomenologist tries to perceive things from the individual's point of view. Although the results can, in some cases, be converted into numbers, this is not the main purpose.

However, quantitative and qualitative research can also benefit from each other. For instance, qualitative research can form the basis for a larger quantitative study. The results can contribute to the design of questionnaires that are then distributed to a larger population. Quantitative studies can contribute with background knowledge prior to the design of qualitative studies and it can help in developing theories.

Compared with quantitative research qualitative studies are more labor intensive, which means that the number of participants in the study, is usually relatively small making it difficult to generalize the results. On a contextual level, qualitative studies can also result in knowledge that is transferable to similar situations, environments, and participants with similar background knowledge (e.g., persons who are novice drivers and their first lessons at a driving school). [Table 1](#) summarizes the differences between quantitative and qualitative research.

Despite the differences between quantitative and qualitative research all science is in some form or another based on sensory experience. The assumption is sometimes that quantitative research methods are objective, versus qualitative ones are subjective. From a positivistic stance, getting access to subjective accounts cannot be done in an objective way ([Kesseba et al., 2018](#)). However, regardless of the method used the researcher(s) need to decide on the topic, how to measure it, analyze it and then how to interpret the results.

All events in life which we observe would be completely meaningless if the brain did not also help to interpret what is happening. If we observe a man running across a street and if we see a car at the same time, we assume that the man is running to avoid an

Table 1 General differences between quantitative and qualitative research

<i>Dimensions</i>	<i>Quantitative</i>	<i>Qualitative</i>
1. Aim	Preparation	Explore the participants interpretation of the situation
2. The relationship between researcher and participants	Distanced	Close
3. The researcher's approach	Outsider	Insider
4. The relationship between theory and research	Confirmation and testing	Develops gradually
5. Strategy	Structured	Unstructured and semistructured
6. The results	Establish general laws, and generalizations (nomothetic)	More specific and focus on the individual (idiographic)
7. The image of social reality	Static and somewhat external in relation to the actor	Process-oriented and something that is socially constructed by the actors
8. Information	Hard, reliable	Rich, in-depth

Source: Developed from *Bryman (1997)*.

accident. We have thus linked two different phenomena and assumed that these in some way affect each other. Our own and others' previous experiences and the situation help us in this. The person is on the street, which leads to a different interpretation than if he was on the pavement. We have thus created a meaning with what is happening.

It could therefore be argued, that it is impossible for us to observe something entirely objectively because without interpretation, our data is meaningless. Interpretation is influenced by cultural, political and personal interests, or as *Banister et al. (1994)* stressed:

"Research is always carried out from a particular standpoint, and the pretence to neutrality in many quantitative studies in psychology is disingenuous. It is always worth considering, then the "position of the research", both with reference to the definition of the problem to be studied and with regard to the way the researcher interacts with the material to produce a particular type of sense" (p 13).

As described in *Table 1*, the quantitative approach is described as nomothetic which can be used to develop theories. Although a theory governs quantitative research, to a greater extent than qualitative research, *Collins and Stockton (2018)* would argue that theories influence almost all aspects of research and that this also include qualitative studies. The problem arises when the theory is regarded as facts which cannot be disputed.

Although qualitative research has many things in common it is important to point out that there are different forms of qualitative research. They include; diary accounts, in-depth interviews, documents, focus groups, case study research, ethnography, participant observations, and life history.

In this article, the focus will be on in-depth interviews although we will touch on some other methods.

Interviews

In an interview the information obtained is via communication, either face-to face or via telephone, or other media. It can help in discovering new territory, beliefs, and meanings. Sometimes it can explore matter(s) which are not possible using quantitative means since they are too complex (*Banister et al., 1994*). However, as *Arksey and Knight (1999)* pointed out interviewing is not only one research method but several. The only thing they have in common is the conversation between an interviewer who asks the question and an interviewee who responds. In the following section, both the design and execution of interviews will be discussed.

Preparation

Before conducting an interview, a great deal of preparation is needed. To start with it is important to understand the topic being studied. Although this could lead to preconceptions, which would be regarded as the enemy of qualitative research. But despite this it could be argued that the benefits from understanding the topic are much greater than the costs (*Arksey and Knight, 1999*). It can help in connecting with the person being interviewed but also enable the interviewer to ask appropriate follow-up questions (i.e., probes). A good understanding of the research literature enables the researcher to develop questions which are relevant and if possible, differ from previous research (*Jacob and Furgerson, 2012*).

Interview Protocols

An interview "protocol" can be anything from structured to unstructured. If the interview is structured, it means that the questions are closed and the answers usually short. Anything that deviates from this is considered disruptive.

Example:

Driving in 60 km/h in an urban area makes my driving better adjusted to other drivers

Strongly disagree 1 2 3 4 5 6 7 *Strongly agree*

The example shows that a structured interview is similar to a questionnaire used in quantitative research. The difference is that the interview is carried out by an interviewer rather than by post or email. For this reason, it is possible to replicate the study and determine if the results have an internal consistency over time (i.e., reliability). In order to test its reliability, the study must follow the same order every time it is performed which is only possible if the format is structured. Because there are so many similarities with quantitative methods, one may ask if such a study can really be called a qualitative study? The problem with this approach is that it is guided too much by already established theories and as a consequence important aspect will be missed. Structured answers can also give a distorted picture of reality because many areas cannot be quantified on a scale from e.g. 1–7.

Unlike an interview based on a structured interview guide, the semistructured one is different. The guide includes fewer questions which allows the interviewer to tailor his/her questions to the person interviewed, rather than following a strict code of practice. One important advantage with using a semistructured approach is that the questions can be adopted and adjusted to the person being interviewed. Another important advantage is that the interviewer might discover something he/she had not anticipated.

In a study aimed to explore driver's attitudes and behavior the methodology chosen was semistructured interviews (Forward, 2006). The following questions, which were inspired by the Theory of Planned Behavior (Ajzen, 1991), tried to elicit the participant's views about the behavior outlined in a scenario describing a driving violation:

- What would the consequences be if you commit this offence?
- Can you think about any reason why you might be driving in the way indicated?
- Can you think about why you would refrain from it?
- Do you believe that people important to you would approve or disapprove of this kind of driving?
- How easy is it to avoid carrying out this behavior?
- Is there anything, or anybody, which could make you to lose control over the situation?
- Have you ever been in a similar situation before?
- Will you drive like this in the near future?

At the end of the interview additional questions were asked about their: age, driving skills, driving experience, if their mood affected their driving, their definitions of a "safe driver", accident involvement, and/or other offences.

The results from this study generated some issues which could not have been achieved using a quantitative study. For instance, the immediate response to the question about consequences of the act varied and depended on the perception of the situation. In a serious situation the response was purely affective and immediate whereas in a low risk situation it was more considered.

Unstructured interviews use open questions. The absence of a template means that the interviewer decides the direction. In some cases, the purpose of the study may be very vague. In other cases, the interviewer becomes aware of the purpose much later. During the course of study, one may encounter something unexpected, which means that the question may change direction during the course of an interview. Hence, the formulation of the theory take place at the same time as the information is collected.

How to construct the questions?

- The language must be easy to understand and suitable for the group to be studied
- Avoid questions which might provoke
- Avoid leading questions:
- Avoid assumptions
- Avoid too personal questions
- If the questions are sensitive, ask them in past tense rather than in the present
- Embed the question. For instance, the question of drink driving can be answered more correctly if it is embedded with other questions about drinking
- Finally test the questions in a pilot study

Sample Size

As previously pointed out qualitative studies are more time consuming than quantitative studies and as a consequence the sample has to be smaller. Since the purpose of the study is not to generalize the results, but to explore meanings, the sampling technique can be fairly flexible. For instance, the researcher might interview one person, who in turn nominates others ("snowball sampling").

A number of studies have discussed the appropriate number of people taking part in the study. Mason (2010) analyzed 560 different studies that had used qualitative methods ranging from structured to unstructured. The results showed that the most common sample size was between 20 and 30 interviews (followed by 40, 10–25). Since both semistructured and unstructured interviews are rather time consuming, sometimes more than one hour, conducting 20–30 interviews would be a considerable task. Especially if we also consider that the interviews will be transcribed (see Data Analysis article). A conclusion is that the sample needs to be large enough to be able to elaborate and uncover important issues and many different perspectives but at the same time not too large, so it becomes repetitive.

Since qualitative research are more concerned with meaning rather than generalization and replication [Mason \(2010\)](#) would argue that: “There is a point of diminishing return to qualitative sample—as the study goes on more data does not necessarily led to more information” (p 1). With other words the research has come to a point of saturation where additional interviews will not add anything new ([Kvale and Brinkmann, 2009](#); [Silverman, 2004](#)).

One problem with this is that saturation is many times used as a reason for not including more interviews although very rarely this is proved ([Mason, 2010](#)) “the cut off between adding to emerging findings and not adding, might be considered inevitably arbitrary” (p 16). Despite this, the author argued that saturation should be a guiding principle for qualitative data.

[Charmaz \(2006\)](#) pointed out that the sample size also depends on the aim of the study. A small study might reach saturation quicker than a larger one. For instance, if the study wants to discuss assessment of novice drivers then fewer interviews might be needed than if the aim was to study the experience of driver training, from start to finish. She also pointed out that the experience of the interviewer is important. An experienced interviewer might be able to collect rich and meaningful data with only a fraction of the interviews needed for a less experienced researcher. Hence the skill of the interviewer is important which we will return to later.

Recruitment

The process of recruitment and the information provided to the person being interviewed is very critical. [Potter and Hepburn \(2005\)](#) stressed at least two questions; the first one, who is being recruited and second, what information is provided to the interviewee.

During the process of recruitment, the participant needs to be well informed. What should and should not be included in such information depends many times on standards recommended by the Ethical committee in each country. However, the following information will help the person to decide if they want to take part or not:

- Who you are and who you work for?
- The aim of the study
- Why the project is being carried out?
- Who benefits from the results?
- Why you want this person to be involved?
- What the participation in the study means
- What kind of questions will be asked?
- What will happen to the data collected?
- Are the person's answers confidential?
- How will the data be stored and who will get access to it?
- Who they can contact if they have any questions?

If they agree to participate the date, time, and location needs to be specified. In this case, the interviewer needs to be flexible which sometimes means meeting after office hours.

The Interview

Interviews are conducted by people from various backgrounds and some might be very blasé believing that it is very easy and common sense ([Arksey and Knight, 1999](#)). This is clearly not the case in research since the interviewer need a range of skills in order to conduct a meaningful interview. “Asking questions and getting answers is much harder task than it may seem at first” ([Fontana and Frey, 2000](#), p. 645).

The qualitative interview can be conducted in different ways:

- One interviewer and one interviewee
- More than one interviewer—one silent and one holding the conversation
- More than one participant—for example, husband and wife
- Group of people who know each other
- Focus groups—where the participants do not know each other, have been recruited via for example, advertisement
- Telephone interview or other media
- One by one

Remember that if only one person is being interviewed then two people on, so to speak, the opposite side of the table can be intimidating.

Data can be registered using:

- Notes
- Some form of recording device
- Both notes and recorder

If notes are being used by the same person carrying out the interview, then only brief notes should be taken. The important thing is to maintain eye contact and really listen to what is being said.

The Opening of the Interview

Both the opening and the ending of the interview need to be done in a professional way using a script (Jacob and Furgerson, 2012). When opening the interview, the interviewer needs to emphasize, yet again, how important it is that they take part. Give an approximate time of the interview but at the same time add that they can stop it whenever they want. Stress that their name will not be included in any report. Finally ask if they accept that the interview is being recorded. The introduction should also be used as form of “ice breaker” with the aim to make the interviewee relaxed. Sometimes, it might be good to start with some easy to answer questions about their background etc.

During the Interview

Depending on the structure of the interview the protocol is used in different ways. If it is structured, it should be followed in an orderly sequence. If on the other hand, it is semistructured the order depends very much on the person being interviewed. The important thing is to make the interview, as much as possible, as an “ordinary” conversation.

Arksey and Knight (1999) described an interview as a “jam session” and the interviewer as a “jazz musician”. The interviewer needs to improvise and adjust the question to the situation. This also applies to how the questions are being asked, what words to use and tone. Sometimes it might be necessary to use a language which the interviewee is more familiar with. Using jargon and too many difficult words can make the interviewee feel ill at ease and sometimes even feel inferior to the interviewer. The balance of power is important, and it is worth remembering that the interviewer has more power than the interviewee and therefore efforts have to be made to reduce this imbalance. Although, this power dynamic has been challenged since it would appear to depend on who the interviewer and the interview are (Riley et al., 2003)

As previously discussed, the experience of the interviewer is important and as Charmaz (2006) pointed out the difference between experienced and less experienced interviewers when it came to the interview itself. This is if we by experience mean that they understand the topic discussed in more detail than the less experienced one. Background information would then help in asking subsequent questions and not only take everything at face values. Although understanding the topic is important, it is also crucial that the interviewer understand how this, together with his/her own experiences, interests, etc, can influence the interviewees response and later how the information will be analyzed (see Data Analysis article). To be self-aware helps in preventing taking control over the interview and ending up with answers which fit preconceived ideas.

To overcome this some researchers, suggest that some form of “reflexivity” is helpful. It means that the interviewer examines both themselves and the relationship with the person interviewed. This would be done to ensure that the interviewer during the conversation stays open-minded, curious, and avoid premature interpretations (Alvesson, 2002; Alvesson and Sköldberg, 2000).

However, this should obviously not lead to excess selfreflection at the expense of attending to the person being interviewed (Riley et al., 2003). The aim is to understand what the interviewer themselves brings to the situation.

Another problem, closely related to the above, is social desirability or response bias. This can happen if the interviewee becomes aware of the interviewer’s standpoint which will influence the response. For instance, if the questions is about speeding in traffic the interviewee may adapt and develop his/her answers in a certain direction that he/she expects the interviewee wants to hear or to legitimize his/her own actions. For example, “No, I never drive too fast or” “I only drive too fast on motorways and roads where there are separate lanes for cars and unprotected road users”. Hence, the person tries to want to conceal the fact that they have committed an offence, or at least rationalize it. Social desirability can also happen even if the interviewer is nonjudgmental if the topic discussed is sensitive and not accepted in wider society.

This was also noted in a study on drinking and driving (Forward et al., 2007). The interviewees were very vague about the number of times they had been drinking and driving. Sometimes they firmly stated that it only had happened once. In order to find out if this was true the interviewer gently returned to this question at a later stage. The second time this was discussed it became clear that it was far from the first time and the reason for doing the same has also changed. To achieve this more honest account of their “stories” was only possible after building some trust with the interviewees. Obviously, drinking and driving is a serious offence so a negative reaction from the interviewer would have affected the interview in a negative way.

Hence, the art of interviewing involves being emphatic but at the same time detached. In connection with this the issue of selfdisclosure has been debated. The recommendations are sometimes to share own experiences with the person being interviewed as a way to connect (Arksey and Knight, 1999). In contrast, others have argued that it can be counter-productive since it can be intimidating in Arksey and Knight (1999). It can also distract the interviewer since they are reminded of past events which can shift the focus from the interviewee to the interviewer. It is therefore argued that report can be established by being a good listener, nonjudgmental and with a genuine interest in understanding the person being interviewed.

During an interview it is not unusual that the interviewee deviates from the subject and starts to talk about something else. This needs to be handled with care and the task it to gently bring them back to the subject without embarrassing them. The question should also be if it is a deviation or not, perhaps it is just another way to answer the question.

To summarize:

- Concentrate on the conversation
- Adjust the order of the questions after the conversation not the script

- Add appropriate probes
- Show that you are interested
- Make the interviewer feel at ease
- Be positive about the person you are interviewing. Do not show any negative emotions (impatience, irritation, etc.)
- Accept what they have to say without condemnation.
- Be aware of body language
- Have a good understanding of the subject and a great deal of selfawareness

End the Interview

Always ask the interviewee if they want to add anything over and above what has been discussed. After that tell them that, you are pleased with their contribution and that it was valuable. Let them know what will happen later and if they will be able to see a draft of the report. If that is not the case tell them that they will be able to receive the final report. Sometimes it might be necessary with a shorter second interview if question arises during transcription.

Focus Groups

Most of what is said about individual semistructured interviews also apply to focus groups which is a form of data collection in conversational format with more than one interviewee. The characteristic of focus groups methodology is that it is often stated that groups, rather than individuals within groups, are the main unit of analysis (Kreuger, 1994; Morgan, 1988). Ideally, focus groups give some insights to the “public” discourse on certain issues, which might be different from “private” views expressed in individual interviews. Furthermore, it is stated that the interaction within groups generates a particular type of data and one advantage of focus groups is the dynamic interaction. “Focus groups can be viewed as performances in which the participants jointly produce accounts about proposed topics in a socially organized situation” (Smithson, 2000, p.105). Focus groups have been described as particularly useful in early stages of research as a means for eliciting issues which participants think is relevant or in development processes where for example, search for new knowledge and contradiction can be studied (cf. Smithson, 2000; Wibeck, 2014).

Limitations of focus groups include the tendency for certain types of socially acceptable opinion to emerge, and for certain types of participants to dominate the research process. The problem of a dominant voice can be dealt with by making the focus groups homogenous for example, in terms of experience, age, education, and sex (Smithson, 2000, p. 107). The ideal size is stated between five and eight/ten participants. With larger groups it can be difficult to moderate the discussion, and it might be risk for subgroups talking simultaneous with each other. The moderator should encourage different group members to speak, for example, if there is one quiet participant. But it need not be a problem if some of the participants are more silent than others; silence is a well-known feature of human interaction (Poland and Pederson, 1998). The focus group method is not merely, as is sometimes argued a quick way to pick up relevant themes around a topic, but a social event that includes performances by all concerned. Moderating focus groups is like other forms of interviewing a skill that has to be learnt by education and training.

Finally, with regard to interviews and focus groups, evaluations need to be made on a regular basis, for nothing is ever prefect and we can always improve.

Summary and Conclusion

The use of qualitative methods in transport research has become increasingly popular. The reason for this is that new territory and events can be discovered and explored, which is not always possible using quantitative methods. The focus is on participants own meaning of events rather than collecting data, which can be quantified. Indeed, the social world is complex and dynamic and cannot always be translated into numbers. In this chapter the focus is on interviews which are one important method, amongst others. To conduct interviews careful planning is necessary. It starts with preparations which include a sound understanding of the topic in question. A protocol is then being constructed, which can be anything from structured to unstructured, all depends on the central aim of the study. However, in this article the use of structured interviews is not recommended if the researcher wants to explore something with as open a mind as possible. The sample size is another important topic to consider and in qualitative studies there is no absolute requirement, instead the notion of “saturation” is often used, which means that additional interviews are unlikely to increase the understanding of the topic. The interview itself starts with an opening, which clearly describes the aim of the study and additional practical information. The actual interview follows, which is not something to be blasé about, believing that it is very easy and common sense. Instead it can be described as a “jam session” in which the interviewer needs to improvise and adjust the questions to the situation. The ending of the interview is as important as the introduction, perhaps the interviewee wants to add additional information and know what will happen next. The chapter ends with a short description of focus groups, which is not dissimilar to interviews but with more than one person. One important element which the interviewer, or the moderator, needs to master is the ability to let all the people in the group feel included and also have a voice. Finally, if a qualitative study is conducted with the respect it deserves it can generate valuable information, which can be used by itself, but it can also form the basis for quantitative studies using a larger sample.

References

- Alvesson, M., Skoldberg, K., 2000. *Reflexive Methodology: New Vistas for Qualitative Research*. Sage, London.
- Alvesson, M., 2002. *Postmodernism and Social Research*. Open University, Philadelphia, PA.
- Arksey, H., Knight, P., 1999. *Interviewing for Social Scientists*. Sage Publication, London.
- Banister, P., Burman, E., Parker, I., Taylor, M., Tindall, C., 1994. *Qualitative Methods in Psychology – A Research Guide*. Open University Press, Buckingham.
- Bryman, A., 1997. *Kvantitet Och Kvalitet I Samhällsvetenskaplig Forskning*. Studentlitteratur, Lund.
- Charmaz, K., 2006. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analyses*. Sage, Thousand Oaks, CA.
- Collins, C.S., Stockton, C.M., 2018. The central role of theory in qualitative research. *Int. J. Qual. Methods* 71., doi:10.1177/1609406918797475.
- Fontana, A., Frey, J.H., 2000. The interview: From structured questions to negotiated text. In: Denzin, N.K., Lincoln, Y.S. (Eds.), *Handbook of Qualitative Research*. 2nd ed. SAGE Publications, Thousand Oaks, CA, pp. 645–672.
- Forward, S.E., 2006. The intention to commit driving violations - A qualitative study. *Transport. Res. Part F* 9, 412–426.
- Forward, S.E., Linderholm, I., Forsberg, I., 2007. Alkohol i trafiken: djupintervjuer med personer som fälltts för rattfylleri. (Drinking and driving. In-depth interviews with convicted drivers). VTI report 553, Swedish National Road and Transport Research Institute, Linköping; Sweden.
- Jacob, S.A., Furgerson, S.P., 2012. Writing interview protocols and conducting interviews: Tips for students new to the field of qualitative research. *The Qualitative Report*, 17(T&L Art. 6), 1–10. Retrieved from <http://www.nova.edu/ssss/QR/QR17/jacob.pdf>.
- Kesseba, K., Awolowo, I., Clark, M., 2018. Qualitative research methodology: a neo-empiricist perspective. In: BAM2018 Proceedings. British Academy of Management (BAM). Copyright and re-use policy. Retrieved from <http://shura.shu.ac.uk/information.html> Sheffield Hallam University Research Archive <http://shura.shu.ac.uk>.
- Krueger, R.A., 1994. *Focus Groups: A Practical Guide For Applied Research*, Second Edition. Sage, Newbury Park.
- Kvale, S., Brinkmann, S., 2009. *Interviews Learning the Craft of Qualitative Research Interviewing*. Sage, London.
- Mason, M., 2010. Sample Size and Saturation in PhD Studies Using Qualitative Interviews. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, [S.I.], v. 11, n. 3, Aug. 2010. ISSN 1438-5627 Retrieved from <https://www.qualitative-research.net/index.php/fqs/article/view/1428>.
- McLeod, S.A., 2019, July 30. Qualitative vs. quantitative research. *Simply Psychology*. Retrieved from <https://www.simplypsychology.org/qualitative-quantitative.html>.
- Morgan, D.L., 1988. *Focus Groups as Qualitative Research*. Sage, CA, Newbury Park.
- Poland, B., Pederson, A., 1998. Reading between the lines: interpreting silences in qualitative research. *Qual. Inq.* 4, 293–312.
- Potter, J., Hepburn, A., 2005. Qualitative interviews in psychology: problems and possibilities. *Qualit. Res. Psychol.* 2, 281–307, doi:10.1191/1478088705qp045oa.
- Riley, S., Schouten, W., Cahill, S., 2003. Exploring the dynamics of subjectivity and power between researcher and researched. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, [S.I.] V4 (2), 1438–5627. Retrieved from <https://www.qualitative-research.net/index.php/fqs/article/view/713/1544>.
- Silverman, D., 2004. *Qualitative Research. Theory, Method and Practice*, 3rd edition. Sage, London.
- Smithson, J., 2000. Using and analysing focus groups: limitations and possibilities. *Int. J. Social Res. Methodol.* 3, 103–119, doi:10.1080/136455700405172.
- Wibeck, V., 2014. Enhancing learning, communication and public engagement about climate change – some lessons from recent literature. *Environ. Educ. Res.* 20 (3), 387–411, doi:10.1080/13504622.2013.812720.

Further Reading

- Silverman, D., 2016. *Qualitative Research*, 4th edition. Sage, London.
- Wilkinson, S., 2004. Focus group research. In: Silverman, D. (Eds.), *Qualitative Research. Theory Method and Practice*. Sage, London, pp. 177–199.

Behavioral Change

Sonja Haustein, Technical University of Denmark, Department of Technology, Management and Economics, Lyngby, Denmark

© 2021 Elsevier Ltd. All rights reserved.

Introduction	46
Theories of Behavior and Behavioral Change	46
Theories that Explain Transport Behavior as Deliberate Choices	46
Theories of Behavior Change	48
Habits	49
Dual-Process and Dual-System Models	50
Behavior Change as a Consequence of Life Course Events	50
Segmentation Approaches as a Basis for Behavior Change	51
Conclusions	52
Biography	52
See Also	52
References	53
Further Reading	53

Introduction

Human behavior in traffic, such as risky, distracted, or unskilled driving, is the main cause of road traffic crashes, which are among the leading causes of death worldwide. In addition, motorized transport contributes to noise and air pollution and related health problems and is responsible for a significant share of greenhouse gas emissions. While we observe clear declines of emissions in other sectors, emissions from transport are still higher today than in 1990 (EC, 2020).

In the near future, technological solutions alone can neither sufficiently reduce emissions nor road traffic crashes. Actually, technological achievements like higher fuel efficiency or faster travel time often result in rebound effects towards an equally or even less sustainable behavior. Increasing energy efficiency of cars has, for example, been offset with higher shares of larger, more heavy cars, and owning an "eco-car" may increase driving or consumption in other areas as owners feel they have "done their bit" (Sorrell et al., 2020). Automated driving, which is suggested to reduce safety risks, pollution, and congestion, may bring new challenges when people move farther out of cities as they can better utilize travel time (Nielsen and Haustein, 2018). It may not only increase travel demand but also lead to new types of crashes when people engage in distracting activities as a consequence of (too high) trust in the autopilot (Körber et al., 2018). Thus, the highly needed transition to sustainable transport cannot rely on technological solutions alone but requires behavioral change. To achieve this change, we need to consider what we already know about human behavior and how it can be altered.

Theories of Behavior and Behavioral Change

In transport psychology, behavioral change is understood as the adaption of transport behavior as a consequence of changes in psychological factors and processes, which are assumed to determine behavior. Such changes can be triggered by psychological interventions, by changes in the life-course, or in the social and physical environment (e.g., traffic infrastructure, road safety legislation, social norms, and policies).

This section presents the most relevant theories of behavior and behavioral change, as knowledge about the psychological determinants and processes of behavior change delivers important starting points for the design of interventions. Indeed, it has been shown that interventions that are based on a theoretical framework are more effective in changing behavior than non-theory-based interventions (e.g., Webb et al., 2010).

Theories that Explain Transport Behavior as Deliberate Choices

The most often applied theoretical framework in transport research to explain transport behavior is Ajzen's Theory of Planned Behavior (TPB). TPB is an extension of the Theory of Reasoned Action (TRA), which was introduced by Fishbein and Ajzen. Both theories assume that people make reasoned choices aiming at maximizing personal benefits. According to TRA and TPB, a specific behavior (e.g., driving within the speed limit) results from the intention to engage in this behavior. The stronger the intention, the more likely it is that the intention results in the respective behavior. Intention is influenced by attitude toward the behavior, which is the evaluation of the consequences of the behavior (good/bad), and subjective norm, which is the perception of social approval or

support of the behavior. Subjective norm was first only measured by injunctive norm or related beliefs (e.g., “My friends don’t mind if I drive faster than allowed”) but has later been supplemented by descriptive norm, which is the behavior that is observed in others (e.g., “My friends often drive faster than allowed”). Descriptive norm has been found to be more relevant than injunctive norm for both safety-relevant behavior (e.g., Møller and Haustein, 2014) and mode choice (e.g., Eriksson and Forward, 2011) and is thus a relevant addition to TPB.

While TRA only considers behaviors under complete volitional control, TPB can additionally explain behaviors that are under incomplete volitional control. This extension is possible because of the inclusion of an additional factor, called perceived behavioral control (PBC). PBC describes how easy or difficult a person perceives the conduction of a target behavior (e.g., cycle to work). PBC may be completely determined by actual behavioral control (e.g., if a person cannot take the bike because of snow on the cycling path) but in most cases PBC will differ from actual behavioral control as different people perceive the same situations or environments as more or less supportive for the conduction of a behavior. PBC is assumed to be a direct predictor of both intention and behavior. Thus, according to TPB, intention and PBC jointly determine behavior. When applied for explaining mode choice, PBC mostly relates to the perception of the transport infrastructure. Haustein and Hunecke added the construct of perceived mobility necessities (PMN) to better account for perceived requirements and constraints resulting from the personal living situation, like combined demands from family and work, which require a high level of mobility. PMN as well as actual constraints of the living situation have been found to hamper the use of public transportation and to encourage car use, while the effect on cycling depends on the predominant mobility culture (Thorhauge et al., 2020).

TPB has successfully been applied to explain a wide range of transport behaviors, ranging from the use of different transport modes to safety-relevant behaviors like speeding or drunk driving.

What makes the TPB attractive is that it is a parsimonious model while still being open for the inclusion of additional factors. Indeed, a variety of determinants have been added to complement TPB, most importantly habits and personal norm—constructs that will be introduced later in this chapter. In addition, as the standard attitude measures are often reduced to a positive versus negative evaluation of the behavior or to functional aspects like time and money, attitude has been exchanged or complemented by symbolic and affective motives. In case of mode choice, this includes to what extent the use of a specific transport mode is related to fun and passion (excitement), status and prestige, or freedom and autonomy. Some people perceive freedom when driving a car others when cycling; some are proud about their large SUV others about their electric car or bike; in particular symbolic meanings of transport modes do not only differ individually but also based on social and cultural context (e.g., Ashmore et al., 2020) and are related to self-identity (e.g., as a cyclist), a factor that has also been added to TPB.

Personal norm (PN), is the central variable of the Norm Activation Model (NAM) and is defined as the perceived moral obligation to engage in a behavior. NAM was originally developed by Schwartz and Howard to explain altruistic behavior and has later been expanded to pro-environmental behavior, like the use of active transport modes. NAM views pro-environmental behavior as the consequence of the activation of PN. Similar to altruistic behavior (e.g., helping an unknown person in an emergency situation), many pro-environmental actions are associated with behavioral costs, which may prohibit the activation of PN. Whether PN is activated or not, depends on several conditions. First, people must be aware that a specific behavior (e.g., air travel) has negative environmental consequences (climate change), referred to as *awareness of need or problem awareness*. Second, people need to be aware that their behavior contributes to the problem and thus feel responsible for the consequences of their behavior, instead of ascribing responsibility to the industry or the government (= *ascription of responsibility*). Moreover, people will be more likely to perform the behavior when they think that it will help to solve the problem (*outcome efficacy*). Outcome efficacy depends on the expectation that others will also act pro-environmentally—if not, this, similar as denial of consequences and denial of responsibility, can be used to justify one’s own unsustainable behavior. Finally, NAM states that people must perceive the ability to act according to their personal norm (e.g. be able to use other modes or to omit the trip), which is called self-efficacy and can be compared to TPB’s PBC. NAM is less formalized than TPB and has more complex assumptions, leading to a higher variability on how it has been applied and tested.

While TRA and TPB focus on optimizing one’s own benefits, and NAM focusses on acting in accordance with moral norms, Lindenberg and Steg’s Goal-Framing Theory (GFT) integrates NAM’s normative goal and TPB’s gain goal and additionally considers hedonic goals to influence behavior. All three goals are assumed to frame the way people process information and how they act. What goals are predominant depends both on people’s underlying values and situational factors.

An even more parsimonious model to explain behavior is proposed by the Campbell Paradigm, which has been revived by Kaiser in the domain of environmental psychology. In this model, sustainable travel behavior (e.g., cycling to work) is the result of two compensatory effects: an individual’s environmental attitude and the behavioral costs of the behavior. In case of cycling, costs can result from constraints of the transport environment (e.g., unsafe cycling infrastructure, long distance to work), from not aligning with social norms (e.g., in low-cycling countries), or bad weather conditions, for example. As cycling in Denmark or the Netherlands is less socially “costly” (it may even be more stigmatized not to cycle) and is also facilitated by a dense and safe cycling infrastructure, it does not require high environmental attitudes to become a cycling commuter. By contrast, abstaining from air travel has been identified as one of the most difficult environmental behaviors, which can explain the high discrepancy between attitude and behavior, when measured in a conventional way (separating attitude and behavior). With the General Ecological Behavior scale, Kaiser developed an instrument to assess environmental attitude based on the difficulty of self-reported pro-environmental behavior, presuming that attitude is apparent in the behaviors that a person engages in.

From an intervention perspective, the Campbell Paradigm’s assumptions imply that both lowering the behavioral costs of sustainable travel behavior (e.g., by providing a more supportive cycling environment, better public transport, or by increasing the

Table 1 Intervention strategies targeting selected psychological factors

<i>Variables</i>	<i>Intervention strategies</i>
Attitude	Increasing positive beliefs about the desired behavior, decreasing negative beliefs, for example, through trying out a new behavior (e.g., cycling to work, testing an electric vehicle/e-bike) Providing information about the benefits of the behavior (e.g., w.r.t. health, safety, environment, . . .)
Subjective Norm	Social models Communicating descriptive and injunctive norms Correcting misperceptions of social norms Changing legislation (e.g., ban of diesel cars; seat belt law)
PBC	Providing knowledge and skills (e.g., education, training, courses, trials)
Awareness of need	Information, education
Awareness of consequences	Personalized information about environmental/safety effects of own behavior, individual feedback
Ascription of responsibility	Hypocrisy paradigm (inducing cognitive dissonance)
Personal norm	Commitment

Source: Adapted from Klöckner (2015).

costs of alternative modes) and promoting pro-environmental attitudes are both effective strategies to achieve more sustainable travel behavior (Taube et al., 2018).

Table 1 summarizes the main psychological factors of the presented models and shows the related interventions. A change of attitudes and the underlying beliefs can, for example, be achieved by providing people with the opportunity to test the new behavior, for example through access to an electric vehicle (Jensen et al., 2013), or by encouraging people to cycle to work for a test period (Gatersleben and Appleton, 2007). Whether this results in higher PBC as well, depends on actual barriers of the behavior (e.g., driving range of electric vehicles, availability of charging stations, cycling infrastructure), which need to be addressed if one wants the efforts to try a new behavior to result in an overall positive re-evaluation of that behavior.

In the context of aberrant driving behavior, PBC relates to the ease with which people can avoid the violating behavior and has been identified as the most relevant TPB predictor of engaging in drunk driving (Parker et al., 1992). Thus, here it would be relevant to support people in better planning their activities beforehand, so they do not get into a situation, they in the end find difficult to control. Another strategy is to teach practices, which support their actual and perceived feasibility to control their behavior. Yet, it is also relevant to communicate the individual responsibility for risky behavior and the possibility to exert control over it.

In many areas of road safety, an actual increase of skills (e.g., hazard-perception skills) is more relevant than the perception of (improved) skills. With regard to driving skills, an overestimation of own competences and the belief to be able to control all kind of traffic situations are actually potential safety risks, and thus here the aim is to align objective and subjective skills rather than to increase PBC.

To increase subjective norm, the communication of what others, peers, or role models do can be an effective strategy. However, the right communication is crucial, otherwise the effects can be contrary to what is intended (Keizer & Schultz, 2019). Also the change of legislation can result in a change of subjective norm, as has for example been shown for the introduction of seat belt laws that are generally accepted. This requires, however, that people are made aware of the legislation.

To support personal norm, a private or public commitment to engage in a specific behavior (or abstain from another) can be used and has for example shown a supportive effect in an intervention to break car use habits (Matthies et al., 2006). Another behavior change tool addressing moral motivation is the inducement of hypocrisy as introduced by Aronson and colleagues. Here, cognitive dissonance between people's attitudes and their actual behavior is created by first having people to advocate the importance of a specific behavior (e.g., that it is relevant to avoid flying to protect the environment) and afterwards asking them about their own failure in the respective behavior. The wish to reduce cognitive dissonance is expected to cause increased compliance with one's own attitudes and norms. The hypocrisy paradigm has successfully been applied with regard to several pro-environmental behaviors, however, as it requires face-to-face interaction, its use in practice is quite limited.

Theories of Behavior Change

While the action models presented above aim to explain single behavioral decisions, another group of theories describe the *process* of behavior change. The most prominent model is the Transtheoretical Model (TTM), which has been introduced by Prochaska and DiClemente in relation to health-related behavior. Since then, the TTM has been applied in various behavioral contexts, including transport behavior. TTM considers behavioral change a process of five stages: In the *Precontemplation* stage, people have no intention to change their behavior, either as they do not consider it as problematic or have given up. In the *Contemplation* stage people become aware of the need to change their behavior, while they in the *Preparation* stage actually form the intention and plan to take action, that

Table 2 Intervention strategies related to stages of change

<i>Stage of change</i>	<i>Intervention strategies</i>
Precontemplation	Activating social norms Activating personal norms: Enhancing problem awareness; increasing acceptance of personal responsibility Enhancing goal setting and goal commitment
Contemplation	Providing information about the pros and cons of different behavioral alternatives Helping selecting the best alternative Enhancing PBC
Preparation/Action	Supporting behavioral planning (if-then planning) Providing information on how to overcome potential barriers
Maintenance	Providing feedback about successful goal achievement Helping to cope with the temptation to relapse into the old behavior Providing social support (also relevant in other stages)

Source: Adapted from Bamberg (2013) and Bamberg and Schulte (2018).

is they get ready for change. In the *Action* stage people actually change their behavior, while the *Maintenance* stage focusses on stabilizing the new behavior and preventing relapse. Stage progression is not necessarily a linear process and people may be stuck at specific stages or even fall back to earlier stages as a consequence of external or internal barriers. A critique of the TTM is that psychological factors and processes related to stage progression remain rather vague.

To address this weakness, Bamberg developed a new stage model that goes back to the Model of Action Phases (MAP) by Heckhausen and Gollwitzer. The MAP stresses the goal-directed nature of behavior change, and suggests to break the process of change down to four stages, each characterized by one of the following tasks: deliberating, planning, acting, and evaluating. In the Stage model of Self-regulated Behavioral Change (SSBC), Bamberg specifies through which psychological factors and processes people approach from one to the next stage of behavior change by reverting to assumptions of TPB and NAM.

Support for the SSBC comes from a longitudinal study that followed people interested in the purchase of electric vehicles, in which 85% of transitions between the stages occurred along the paths suggested by the stage model (Klößner, 2014). The model also provides concrete suggestions for psychological interventions to trigger stage progression. This has for example been tested in a phone-based social-marketing campaign aiming to reduce car use (Bamberg, 2013). In the study, people received interventions (personal phone call and specific information material) that were adapted for their specific stage of change, resulting in a significant reduction of car use and increased use of public transportation.

The model has also inspired a Danish cycling campaign (Nielsen and Haustein, 2019). Here, more general information about cycling and its health benefits were provided to raise awareness among people who did not intend to increase cycling yet. For people who had the goal to increase cycling, a customized smart-phone app (a “cycling coach”) helped to set a realistic goal, and goal-achievement was supported by feedback, self-monitoring, and social support. Campaign activities of a municipality were positively related to the number of cycling trips, while controlling for possible confounding factors.

Table 2 provides an overview of intervention strategies for each of the four stages of behavior change. A study that reviewed the effectiveness of different behavioral change techniques to promote walking and cycling found that the inclusion of self-monitoring and intention-formation techniques is most promising (Bird et al., 2013). These measures are most likely even more effective, when applied in the matching stage of behavior change.

Habits

While the previous models referred to behavior (or behavior change) as a consequence of a deliberate intention or decision process, behavior differs in the degree to which it is preceded by a deliberate intention. Some behaviors are performed in the same way as always, rather automatic, without thinking about it—here we speak of a habit, a concept that was first introduced by Triandis.

The more frequently a behavior is performed under the same stable circumstances and leads to the desired outcome, the more likely it will be performed again and turns into a habit that is activated automatically by specific situational cues. This process requires less cognitive resources than deliberate behavioral decisions. However, if people have first established a habit, it is more difficult to change the behavior again.

Habits have, for example, been identified as important predictors of car use and have successfully been integrated into TPB. As strong habits lead to less interest in information about alternative behaviors, established intervention techniques often fail. Therefore, substantial changes of the situational conditions under which the behavior takes place have been suggested as a precondition to change habits. This could, for example, be the provision of a free public transport pass, which should ideally be combined with an intervention targeting psychological factors or processes (also referred to as “Downstream-Plus-Context-Change Interventions,” Verplanken and Wood, 2006) to avoid that behavior goes back to normal when the incentive is removed.

Dual-Process and Dual-System Models

Attempts to integrate conscious, deliberate processes that influence behavior (as explained by models like TPB), and automatic or impulsive processes (including habits), are provided by Dual-process or Dual-system models. These approaches suggest that two separate processes or systems determine our behavioral choices. The first one (also referred to as system 1, implicit or impulsive system) works unconscious, fast and automated and may result in irrational outcomes; the other one works conscious and slower, deliberate and more rational. Depending on factors, such as self-control, cognitive load, or mood, the one or other system is supposed to be dominant.

Implicit evaluative associations, also referred to as implicit attitudes, are part of the impulsive system and are suggested to influence explicit attitudes and behaviors. In contrast to explicit attitudes, which can be self-reported, implicit attitudes are not immediately consciously accessible and are according to Greenwald and Banaji's definition "traces of past experience." Different computer-based measures exist to measure implicit attitudes, such as the Implicit Association Test (IAT), or the Go/No-go Association Task (GNAT). In contrast to self-reported attitudes, these measures are not subject to self-report bias, as participants are not aware of what is measured. While the IAT is widely used in different fields of psychology, its validity and reliability has also been questioned.

A measure to change implicit attitudes is the Approach-Avoidance Task (AAT), in which participants respond to pictures by pulling or pushing a joystick. The specific instructions result in that the majority of pictures with the behavior/subject that should be avoided are pushed away (without knowing the purpose of the task), while the majority of pictures with the desired behavior/subject is pulled toward oneself, resulting in a change in implicit attitudes. The measure has been applied successfully to change implicit attitudes and subsequent behavior like alcohol consumption. More recently, the AAT has also been applied to change implicit drunk driving attitudes (Martinussen and Sømnhovd, 2019).

With regard to dual-system models, different ways on how the impulsive and reflective system influence behavior have been proposed. It has been suggested that spontaneous behavior is better predicted by the impulsive system, and deliberate choices are better explained by the reflective system (*dissociation pattern*). However, there is also empirical support for both systems contributing independently to the explanation of the same behavior (*additive pattern*). Finally, it has been suggested that both system interact to influence behavior (*interactive pattern*). A longitudinal study focusing on the question how implicit and explicit attitudes influence intention and behavior of physical activity (Muschalik et al., 2018) could not find a direct effect of implicit attitudes on behavior but a moderating effect, which support the assumed interactive pattern. With regard to future interventions, this would imply that it is effective to target both implicit and explicit attitudes to achieve behavior change.

Behavior Change as a Consequence of Life Course Events

While psychological models mainly focus on behavior change as a consequence of changes of intra-psychic processes, the interdisciplinary Mobility Biographies Approach deals with behavior change as a consequence of key events in the life course. These events make it likely that current mobility behavior is reconsidered and may be subject to change and existing travel habits are disrupted. Müggenburg et al. (2015) provide a theoretical framework for the Mobility Biographies Approach, in which three types of key events are distinguished:

1. Life events, separated into private events (e.g., child birth, children leaving parental home, road traffic crashes) and events related to the professional career (e.g., entering labor market, starting university; changing the workplace, retirement)
2. Adaptations in long-term mobility decisions, which are residential location and long-term modal decisions (e.g., car ownership, licensing, seasonal ticket)
3. Exogenous interventions, which may be targeted (e.g., provision of information, free public transport tickets, shared bikes) or untargeted (e.g., closure of a bridge, new infrastructure)

According to the framework, these events influence each other, are influenced by long-term processes related to ageing or socialization, period and cohort effects, and determine everyday mobility decisions. In case of long-term and everyday mobility decisions, reciprocal influences are assumed.

Among the events that have been examined most often is residential relocation. Scheiner and Holz-Rau, for example, examined its effect on mode choice and found that moving to suburban locations was followed by an increase in car use, while the opposite was the case for moving into the city. A similar study by de Vos et al. (2018) found that travel attitudes often influence the choice of the new location and that both travel attitudes and travel mode choice change after relocation, aligning over time. This stresses the reciprocal relationship between attitude and behavior that is often ignored in empirical tests of psychological models, although the alignment of attitudes and behavior as postulated in Festinger's Theory of Cognitive Dissonance is generally not questioned or even part of theoretical assumptions.

The Mobility Biographies Approach has, for example, been used as a framework to interpret mobility trajectories of car sharing users (Jain et al., 2020). Here, it has been shown that the reconsideration of car ownership is often triggered by key life events or residential relocation, while car sharing mostly facilitated changes in mobility behavior rather than causing them. Life events, such as retirement, children leaving home or child birth, have thus been suggested as optimal occasions for promoting car sharing either as a means to abolish a car or to avoid buying an additional car.

While the Mobility Biographies Approach mainly focusses on mobility and mode choice, also traffic safety, in particular driving behavior, can be influenced by life events, for example the involvement in a road traffic collision. Crash involvement can lead to posttraumatic stress (PTS), which has been found related to hostile-aggressive driving or anxious driving behavior—both with negative consequences for road safety. It can even lead to driving cessation, which is a key event in itself that can have negative consequences for health, well-being, activity involvement and social inclusion. Apart from PTS, more recently, attention has also been placed to positive psychological consequences of traumatic experiences, referred to as post-traumatic growth (PTG). However, the concept of PTG, the relation to PTS and the joint impact on driver behavior change is still subject to investigation.

Whether and to what extent events in the life course change our behavior is related to a variety of demographic, psychological, and geographic factors and is thus difficult to generalize. But as important life events interrupt our daily routines and behavioral patterns, intervention research understands them as opportunities to change undesired automated behaviors in a positive direction.

Segmentation Approaches as a Basis for Behavior Change

While interventions can be informed by a theoretical framework, as has been shown in the previous sections, another approach to design interventions are user segmentations. This approach is motivated by the aim to create more tailored interventions, which are assumed to be more effective than “one-fits-all” approaches.

A classic example of a user segmentation is the differentiation between captive and choice users of a transport mode. However, in the last two decades much more nuanced user segmentations have been suggested. One of the first ones was developed by Mette Jensen and is based on sociological analysis. She distinguished between passionate car drivers, daily life car drivers, leisure time car drivers, cyclists/public transport users of heart, cyclists/public transport users of convenience and cyclists/public transport users of necessity. Subsequent segmentation approaches with focus on mobility behavior, as for example suggested by Anable or Hunecke, were mostly informed by psychological theory, in particular TPB and NAM, and created based on standardized items and cluster analysis. Ideally, the resulting cluster profiles provide valuable information on how to design and promote interventions to achieve shifts from car to pro-environmental transport modes. Several other areas of application exists, for example changing air travel (Davison et al., 2014) or departure time choice (Haustein et al., 2018),

all of them with roots in behavioral theory. Instead of these post hoc approaches, where the resulting segments are not known beforehand, segments can also be defined a priori, for example, when dividing people into stages of change based on Bamberg's SSBC, which can be done by related stage-diagnostic questions.

In a traffic safety context, segmentations more often take personality traits and/or driving style as a basis to define subgroups of road users that require different types of interventions as traits have been found related to safety-relevant behavior, such as speeding or road anger expression. While certain personality traits are also related to environmental attitudes and behaviors, these are (so far) seldom used in the context of segmentations related to mobility behavior.

Besides approaches that have their roots in psychological and sociological theory, geographic segmentations compare users based on the city, district or neighborhood they live in, certain characteristics of these (e.g., density, land use mix) or their combination. As a specific form of a geographic segmentation (influenced by sociology and psychology), one can consider so-called *mobility cultures*. What mobility culture a region belongs to, does not only depend on spatial factors and the transport environment but also on predominant travel patterns, attitudes and norms as well as mobility-related policies and discourses. While mobility cultures are traditionally described qualitatively, *urban mobility cultures* (Klinger et al., 2013) and *European mobility cultures* (Haustein and Nielsen, 2016) have been identified and described based on quantitative data.

Demographic segmentations divide people into age groups or household types or more specific life-stages or life-styles, which combine several demographic factors, which are linked to travel behavior. In a traffic safety context, the subgroup of young male drivers is particular relevant due to a higher engagement in risky driving, and a particular focus is also placed on the growing group of older drivers. Also travel or traffic behavior itself is used to group people—in a mobility context groups of interest could be car commuters with a short trip to work (thus a potential to change mode choice), in traffic safety it could be people engaging in aberrant driving or road anger expression—in these cases the aim would be to identify target groups for specific behavior change interventions.

What kind of segmentation approach is relevant to choose, highly depends on the purpose of the segmentation, and all approaches have specific pros and cons. Demographic and geographic segmentations allow for an easy identification of members, however, knowledge about the segments' specific needs is restricted. Psychographic segmentations provide better starting points for information and communication strategies based on the attitudinal profiles of the segments, but this requires a higher effort with regard to data collection and analysis.

Although based on different kind of factors, different segmentations approaches may result in the identification of similar groups as has been shown in a comparison of different segmentation approaches of older road users. It can also make sense to mix different kind of factors to combine the advantages of several approaches. Table 3 provides an overview of common approaches in transport research and practice and their main purposes.

Table 3 Segmentation approaches in transport

<i>Approach</i>	<i>Grouping variables (examples)</i>	<i>Purpose (examples)</i>
Psychographic segmentations	Personality traits, attitudes, values, norms, perceived barriers, risk perception, sensation seeking, change intention	Social marketing, basis for psychological interventions
Geographic segmentations	Countries, cities, districts, population density, neighborhood characteristics, traffic infrastructure	Monitoring, basis for policy and long-term infrastructural interventions
Demographic segmentations	Age, gender, occupation, household form	Travel demand modelling, destination choice, residential choice
Behavioral segmentations	Mode choice, trip frequency, driving behavior, traffic violations	Monitoring, identification of target groups for training, courses, incentives, travel demand management

Source: Adapted from *Haustein and Hunecke (2013)*.

Conclusions

Current psychological models—or at least their empirical operationalization in transport research—often assume a directed causal relation from attitude to behavior. However, there is increasing evidence of a reciprocal relationship. This is also supported by the Theory of Cognitive Dissonance, which assumes that people tend to align their attitudes and behaviors to avoid mental discomfort. From an intervention perspective, this implies that it is both relevant to try to change people's attitudes and to provide opportunities to test or practice new behaviors. The latter may lead to a re-evaluation of attitudes and perceived barriers—if the social and physical environment supports that the new behavior results in a positive experience. Thus, a combination of efforts to improve the social and physical environment with tailored psychological interventions seems most promising to achieve a change of behavior that results in increased health, well-being, and environmental protection.

Biography



Sonja Haustein is a Senior Researcher and Psychologist at the Technical University of Denmark. Her research aims at understanding transport related choices and thereby providing the basis for interventions that facilitate green, safe, and inclusive transport. Key research areas include behavioral change, new transport modes and services, active travel, segmentation approaches, and psychological interventions.

See Also

The Rebound Effect for Car Transport; The Theory of Planned Behavior; Modeling Passenger Transport Mode Choice; Habitual Behavior; Social, Personality and Affective Aspects of Driving; Mode Choice and Life Events; Young Drivers; Older Drivers; Driver Aggression & Anger; Speeding Behavior; Electromobility: Attitudes and Perceptions; Carsharing; Data Analysis: Classification; Transport Modes and Health; Driver Distraction; Sharing: Attitudes and Perceptions; Future Mobility: Attitudes and Perceptions; Driving Behavior & Skills

References

- Ashmore, D.P., Thoreau, R., Kwami, C., Christie, N., Tyler, N.A., 2020. Using thematic analysis to explore symbolism in transport choice across national cultures.. *Transport* 47 (2), 607–640.
- Bamberg, S., 2013. Applying the stage model of self-regulated behavioral change in a car use reduction intervention. *J. Environ. Psychol.* 33, 68–75.
- Bamberg, S., Schulte, M., 2018. Processes of change. In: *Environmental Psychology: An Introduction*. pp. 307–318.
- Bird, E.L., Baker, G., Mutrie, N., Ogilvie, D., Sahlqvist, S., Powell, J., 2013. Behavior change techniques used to promote walking and cycling: a systematic review. *Health Psychol.* 32 (8), 829.
- Davison, L., Littleford, C., Ryley, T., 2014. Air travel attitudes and behaviours: the development of environment-based segments. *J. Air Transport Manage.* 36, 13–22.

- De Vos, J., Ettema, D., Witlox, F., 2018. Changing travel behaviour and attitudes following a residential relocation. *J. Transport Geogr.* 73, 131–147.
- EC (European Commission), 2020. Transport emissions. A European strategy for low-emission mobility. Retrieved from: https://ec.europa.eu/clima/policies/transport_en (21-6-2020).
- Eriksson, L., Forward, S.E., 2011. Is the intention to travel in a pro-environmental manner and the intention to use the car determined by different factors? *Transport. Res. D: Transport Environ.* 16 (5), 372–376.
- Gatersleben, B., Appleton, K.M., 2007. Contemplating cycling to work: attitudes and perceptions in different stages of change. *Transport. Res. A: Policy Pract.* 41 (4), 302–312.
- Haustein, S., Hunecke, M., 2013. Identifying target groups for environmentally sustainable transport: assessment of different segmentation approaches. *Curr. Opin. Environ. Sustain.* 5 (2), 197–204.
- Haustein, S., Nielsen, T.A.S., 2016. European mobility cultures: a survey-based cluster analysis across 28 European countries. *J. Transport Geogr.* 54, 173–180.
- Haustein, S., Thorhauge, M., Cherchi, E., 2018. Commuters' attitudes and norms related to travel time and punctuality: a psychographic segmentation to reduce congestion. *Travel Behav. Soc.* 12, 41–50.
- Jain, T., Johnson, M., Rose, G., 2020. Exploring the process of travel behaviour change and mobility trajectories associated with car share adoption. *Travel Behav. Soc.* 18, 117–131.
- Jensen, A.F., Cherchi, E., Mabit, S.L., 2013. On the stability of preferences and attitudes before and after experiencing an electric vehicle. *Transport. Res. D: Transport Environ.* 25, 24–32.
- Keizer, K., Schultz, P.W., 2019. Social norms and pro-environmental behaviour. In: Steg, L., de Groot, J.I. (Eds.), *Environmental Psychology: An Introduction*, Second ed.. Wiley, Hoboken, pp. 179–188.
- Klinger, T., Kenworthy, J.R., Lanzendorf, M., 2013. Dimensions of urban mobility cultures—a comparison of German cities. *J. Transport Geogr.* 31, 18–29.
- Klößner, C.A., 2014. The dynamics of purchasing an electric vehicle—a prospective longitudinal study of the decision-making process. *Transport. Res. F: Traffic Psychol. Behav.* 24, 103–116.
- Klößner, C.A., 2015. *The Psychology of Pro-Environmental Communication: Beyond Standard Information Strategies*. Palgrave Macmillan, Basingstoke.
- Körber, M., Baseler, E., Bengler, K., 2018. Introduction matters: manipulating trust in automation and reliance in automated driving. *Appl. Ergon.* 66, 18–31.
- Martinussen, L.M., Sørhovd, M.J., 2019. Avoiding drunk driving with a push. *Eur. Transport Res. Rev.* 11 (1), 16.
- Matthies, E., Klößner, C.A., Preißner, C.L., 2006. Applying a modified moral decision making model to change habitual car use: how can commitment be effective? *Appl. Psychol.* 55 (1), 91–106.
- Møller, M., Haustein, S., 2014. Peer influence on speeding behaviour among male drivers aged 18 and 28. *Accid. Anal. Prev.* 64, 92–99.
- Müggenburg, H., Busch-Geertsema, A., Lanzendorf, M., 2015. Mobility biographies: a review of achievements and challenges of the mobility biographies approach and a framework for further research. *J. Transport Geogr.* 46, 151–163.
- Muschalik, C., Elfeddali, I., Candel, M.J., de Vries, H., 2018. A longitudinal study on how implicit attitudes and explicit cognitions synergistically influence physical activity intention and behavior. *BMC Psychol.* 6 (1), 18.
- Nielsen, T.A.S., Haustein, S., 2018. On sceptics and enthusiasts: what are the expectations towards self-driving cars? *Transport Policy* 66, 49–55.
- Nielsen, T.A.S., Haustein, S., 2019. Behavioural effects of a health-related cycling campaign in Denmark: evidence from the national travel survey and an online survey accompanying the campaign. *J. Transport Health* 12, 152–163.
- Parker, D., Manstead, A.S., Stradling, S.G., Reason, J.T., Baxter, J.S., 1992. Intention to commit driving violations: an application of the theory of planned behavior. *J. Appl. Psychol.* 77 (1), 94.
- Sorrell, S., Gatersleben, B., Druckman, A., 2020. The limits of energy sufficiency: a review of the evidence for rebound effects and negative spillovers from behavioural change. *Energy Res. Soc. Sci.* 64, 101439.
- Taube, O., Kibbe, A., Vetter, M., Adler, M., Kaiser, F.G., 2018. Applying the Campbell Paradigm to sustainable travel behavior: compensatory effects of environmental attitude and the transportation environment. *Transport. Res. F* 56, 392–407.
- Thorhauge, M., Kassahun, H., Cherchi, E., Haustein, S., 2020. Mobility needs, activity patterns and activity flexibility: how subjective and objective constraints influence mode choice. *Transport. Res. A* 139, 255–272.
- Verplanken, B., Wood, W., 2006. Interventions to break and create consumer habits. *J. Public Policy Market.* 25 (1), 90–103.
- Webb, T.L., Joseph, J., Yardley, L., Michie, S., 2010. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. *J. Med. Internet Res.* 12 (1), 1–27.

Further Reading

- Bamberg, S., 2013. Changing environmentally harmful behaviors: a stage model of self-regulated behavioral change. *J. Environ. Psychol.* 34, 151–159.
- Deutsch, R., Strack, F., 2006. Duality models in social psychology: from dual processes to interacting systems. *Psychol. Inquiry* 17 (3), 166–172.
- Gärling, T., Ettema, D., Friman, M. (Eds.), 2014. *Handbook of Sustainable Travel*. Springer, New York, NY, USA.
- Hunecke, M., Haustein, S., Böhrer, S., Grischkat, S., 2010. Attitude-based target groups to reduce the ecological impact of daily mobility behavior. *Environ. Behav.* 42 (1), 3–43.
- Kaiser, F.G., Byrka, K., Hartig, T., 2010. Reviving Campbell's paradigm for attitude research. *Personal. Soc. Psychol. Rev.* 14, 351–367.
- Klößner, C.A. (Ed.), 2015. *The Psychology of Pro-Environmental Communication: Beyond Standard. Information Strategies*. Palgrave Macmillan, Basingstoke.
- Kroesen, M., Handy, S., Chorus, C., 2017. Do attitudes cause behavior or vice versa? An alternative conceptualization of the attitude-behavior relationship in travel behavior modeling. *Transp. Res. Part A Policy Pract.* 101 (1), 190–202.
- Steg, L., de Groot, J.I. (Eds.), 2019. *Second ed. Environmental Psychology: An Introduction*. Wiley, Hoboken, NJ.
- Scheiner, J., Rau, H. (Eds.), 2015. *Mobility and Travel Behaviour Across the Life Course*. Edward Elgar.

Habitual Behavior

Carlo G. Prato, The University of Queensland, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	54
Models of Habit	55
Models Based on the Theory of Planned Behavior	55
Models Based on the Norm Activation Model	55
Models Based on the Value Belief Norm Theory	56
Models Based on Other Theories	56
Habit Measurement	56
Future Prospects	57
References	57

Introduction

When observing travel behavior, it is quite common to note that travel patterns repeat themselves day to day, week to week, and at times even month to month if not year to year (Gärling and Axhausen, 2003). A very common interpretation of this phenomenon is that travel behavior is habitual, but is it really habit or rather the repetition of behavior?

From a social psychology perspective, habit is a concept that is often included in models that explain human behavior (Triandis, 1977). A certain behavior will have an occurrence probability that depends on intention and habit under physiological excitement and objective conditions. Intention and habit exist in a balance that implies that an increase in habit leads to a decrease in intention, and vice versa. Intention is generally defined as the probability that is consciously assigned to a certain behavior: positive attitudes toward the behavior, social norms by relevant others, and situational control are determinants of intention strength (Ajzen, 1991). Habit is generally defined as behavior that has become automatic in reaction to certain cues and keys: it is not intentional and that is the reason why habit makes people do what they have always done even when their intentions are to do something else and behave differently (Neal et al., 2006). Habit is strengthened with an increase in the frequency that these cues and keys are given and an intensification in the learnt tendency to repeat past reactions (Wood and Neal, 2007). As habitual behavior accounts for a large proportion of ordinary activities, it is no surprise that it plays a significant role in one of them: travel.

From a travel behavior perspective, habits are considered to be more complex than repeated behavior (Forward, 2014; Lanzini and Khan, 2017; Ouellette and Wood, 1998; Verplanken and Aarts, 1999). A caveat in travel behavior research is that repeated behavior is not necessarily habitual behavior: replicating a behavior might in fact manifest from the intention (e.g., to drive to work) being formed repeatedly. Accordingly, the key difference concerns the amount and/or depth of information that is searched and processed prior to performing the behavior. Habit has been hypothesized theoretically (Gärling and Garvill, 1993) and proved empirically to be the choice to perform a behavior without deliberation (Gärling and Garvill, 1993; Verplanken et al., 1997). Information processing is in fact the key in distinguishing habit from repeated behavior, as the former implies that the choice is script-based with less information being searched and processes, whereas the latter implies that the choice is rational and goal-directed with more information being processed and the same outcome being obtained. However, a relation exists between habit and repeated behavior, as the goal orientation motivates people to replicate their behavior and then repetition supports habit formation. Habit becomes then important in travel behavior, as repeating one type of behavior implies that the choice of behavior is not dependent only on reason (Aarts et al., 1997) and, when the habit occurs once, it reduces the probability of the choice of a different behavior (Verplanken et al., 1997).

Accordingly, consensus exists that habit allows predicting the future travel behavior of people. However, is the frequency of past behavior the best predictor of future behavior? Even though the model of interpersonal behavior measures habit strength as the number of repetitions of a certain behavior (Triandis, 1977), a difference exists between habit and repeated behavior, and hence a difference exists between modeling habit or inertia. While habit reveals itself as inertia, the opposite is not necessarily true as a behavior can be inertial (i.e., repeated many times) but not habitual. However, the literature on travel behavior seems to use the two terms interchangeably, also because it is very complex to disentangle the different components of inertia given that an inertial behavior can become habitual over time. A wealth of applications has been observed, especially within a microeconomic approach modeling choices for travel modes (e.g., Bamberg et al., 2003; Chatterjee, 2011; Cherchi and Manca, 2011; Gardner, 2009; Gärling and Axhausen, 2003; Kaplan et al., 2017), vehicles (Bauer, 2018; Jansson et al., 2009), parking (van der Waerden et al., 2015), routes (Bogers et al., 2005; Kaplan and Prato, 2012), residential locations (Ralph and Brown, 2017), and departure times (Thorhauge et al., 2020). The objective of these applications is often to influence travel choices from unsustainable, and generally inertial if not habitual, solutions to sustainable, and generally more rational and deliberate, ones. While it appears obvious that a change of routine could be driven by an increased awareness of other alternatives (Fujii et al., 2001), the observation of travel behavior shows

that behavior change is quite difficult to observe. However, at least some level of behavior change cannot be avoided if carbon emissions from transport are to be reduced significantly (Anable, 2005; Banister, 2008, 2011), and this is the reason why it is fundamental to understand how to model and measure habit within the current models of travel behavior. Understanding habit is in fact fundamental if the challenge is to break unsustainable and carbon-intensive habits (see, e.g., Gärling and Axhausen, 2003; Matthies et al., 2006; Verplanken and Wood, 2006).

Given these premises, this article looks at the behavioral theories where habit is prevalently incorporated, the measurement methods that capture habit and its strength, and the directions that research into habitual behavior could pursue.

Models of Habit

When studying travel behavior, three major frameworks may be considered (Lanzini and Khan, 2017), namely the ones based on the Theory of Planned Behavior (TPB, Ajzen, 1991), the ones based on the feeling of moral obligation (e.g., Nordlund and Garvill, 2002), and the ones based on habit (e.g., Aarts and Dijksterhuis, 2000). Habit is therefore one of three fundamental research frameworks, and this explains why it has been considered for long time as integral to the process of explaining travel behavior (e.g., Banister, 1978; Verplanken et al., 1994). While some studies confused on the concept of habit (e.g., Friedrichsmeier et al., 2013), most studies concentrated on habit as an extension of the existing theoretical models that deal in particular with travel mode choice and its changes as the focal point of studying sustainable transport choices. The theoretical models used for explaining travel the explanation of travel mode choice mostly include the TPB (Ajzen, 1991), the Norm Activation Model (Schwartz, 1977), and the Value Belief Norm Theory (Stern et al., 1999). This section looks at the applications of these theories for capturing habit in behavioral models.

Models Based on the Theory of Planned Behavior

The most commonly used theory for explaining travel behavior, and in particular travel mode choice, is the TPB in either the original or a modified version. Generally, studies hypothesize the relation between habit and intention, and then verify empirically whether this relation exists and what its strength is. The most prevalent approach complements the TPB with a measurement of habit (e.g., Bamberg and Schmidt, 2003; Verplanken et al., 1998), while an alternative approach extends the theory by adding the measurement of habit (e.g., Donald et al., 2014; Forward, 2014).

Considering the three main behavioral predictors (i.e., attitudes, subjective norms, perceived behavioral control) that form an intention that precedes behavior, it is possible to measure the strength of habits with respect to the three behavioral predictors and reveal whether they are strong, thus making the relation between intention and behavior weak (Verplanken et al., 1998). Alternatively, it is possible to map the influence of past behavior on future behavior by inducing change to context conditions and hence measuring habit as a mediator of this influence (Bamberg et al., 2003), or simply add habit as behavioral predictor and make the prediction of intention and behavior more accurate (Bamberg and Schmidt, 2003).

Extending the model with moral norms, descriptive norms, and environmental commitment allows explaining travel behavior via intention and habit, while showing that unsustainable behavior is often determined by both intention and habit, while sustainable behavior is affected only by intention (Donald et al., 2014). Extending the model with the Trans theoretical Model of Behavior Change further elaborates the concept by having six phases that not only define, but also describe the formation of habit (Forward, 2014): (1) precontemplation, (2) contemplation, (3) preparation, (4) action, (5) maintenance, and (6) termination. This model extends the TPB that focuses only on preparation and action by working on the awareness of alternative behavior (in the contemplation phase) and the continuation of the changed behavior (in the maintenance phase) that are essential to break through strong habits. An essential element of this model is the recognition that, although positive consequences are the results of the behavior change, negative consequences need to be mapped as well if the behavior is to be maintained and become a new habit.

Models Based on the Norm Activation Model

Another commonly used theory for explaining human behavior is the Norm Activation Model (Schwartz, 1977). Generally, studies map travel behavior and in particular mode choice and its change (e.g., Hunecke et al., 2001; Nordlund and Garvill, 2002).

Considering the processes of activation of moral norms and their transformation into action, and having these processes explain altruistic and environmental behavior, three predictors of behavior are considered: awareness of consequences, assignment of responsibility, and personal norms. The extension of the model to habit assumes that the personal norm of sustainable travel behavior would be a strong predictor of travel choices only without the effect of a contradictory habit (Klöckner et al., 2003; Klöckner and Matthies, 2004). A strong habit stops in fact the norm activation process and mediates the effect of personal norms on travel behavior, while starting directly from the context.

An evolution of the model considers whether people do not behave in accordance with norms, triggering therefore protection mechanisms such as the refusing responsibility for their actions or reframing the context in which they act (Matthies et al., 2006). A further evolution of the model considers the effect of socialization, in relation to both peers and household members, on the choice of travel mode through habit (Haustein et al., 2009; Klöckner and Matthies, 2012). Summarizing, extending the Norm Activation Model to the effect of habit can help explain why people are not able to behave according to their moral motivation.

Models Based on the Value Belief Norm Theory

A third commonly used theory to represent behavior is the Value Belief Norm Theory (Stern et al., 1999). The theory is generally the combination of theoretical concepts of pro-environmental behavior that has been used to explain travel behavior and in particular travel mode choices (e.g., Eriksson et al., 2006; Nordlund and Garvill, 2002; Nordlund and Westin, 2013). The theoretical concepts that are combined in the Value Belief Norm Theory are the Theory of Values (Schwartz, 1992), the Norm Activation Theory (Schwartz, 1977) and the New Environmental Paradigm (Stern et al., 1995).

According to this theory, specific travel behavior is originated by a causal chain of variables that include values, beliefs, and personal norms. Personal norms for sustainable travel behavior are activated by beliefs that external conditions threaten individual values, and accordingly people might act to limit this threat. An empirical evaluation of the theory analyses attempts to break the habitual use of nonsustainable travel modes via extensive discussions with an expert with a focus on potential behavioral change and “implementation plans” (Eriksson et al., 2008). The discussions change the perspective of the behavior into becoming more intentional and hence the habit of nonsustainable behavior into becoming weaker, while the relation between this behavior and the personal norm becomes stronger. The hypothesis is that habit can be broken via interventions focused on influencing pro-environmental values and norms for which there is a belief to act in compliance with them to achieve sustainable behavior.

Models Based on Other Theories

Apart from habits, all these theories include attitudes as predictors of behavior. The relation between habit and attitudes can be expected to be similar to the relation between habit and intention, as intention can be viewed as formed by attitudes, subjective norms, and behavioral control (Ajzen, 1991), or by past behavior, attitudes, and subjective norms (Ouellette and Wood, 1998). In the travel behavior context, the assumption is that habits and attitudes interact in the prediction of travel mode choices in the future. If habit is strong, then the relation between attitude and behavior is weak, and vice versa: then, increasing strength of habit implies a predictiveness loss in attitudes (Aarts et al., 1997; Verplanken et al., 1994, 1997).

Another approach is the mapping of motivations and habit as determinants of travel behavior (Gardner, 2009). Habit forms from repeated behavior that result from the strong intention to perform that behavior. If habit is strong, the behavior is not formed any more by motivation, in line with the inversely proportional relation theorized by Triandis (1977). If habit is not considered, then the role of decision processes in travel behavior choices may be overestimated: accordingly, measurement of habit needs to be included when measuring decision-making processes. Considering habit, attitudes, and motivation helps models to become more applicable in practice.

Habit Measurement

From the theoretical approaches presented in the previous section, it is evident that one of the key aspects for the inclusion of habit in the different theories is its measurement. How is habit operationalized by researchers in the context of the theoretical foundations?

The frequency of behavior is an indicator of strength of habit (e.g., Bamberg, 2000; Triandis, 1977), but also past behavior is at times defined as habit (Ouellette and Wood, 1998). Then, the strength of habit was also measured, with strong habits defined as the ones emerging often and in stable contexts, and weak habits defined as the ones reflecting unusual effectiveness or changeable circumstances (Wood et al., 2002, 2005). However, these definitions do not disentangle properly repeated behavior from habit.

Accordingly, an alternative approach was proposed on the basis of the concept of habit as a script. The response-frequency method (Verplanken et al., 1994) presents situations or scripts that occur naturally and ask participants to choose quickly a travel mode. An increased frequency of a certain choice is assumed to indicate a more script-based or habitual behavior under the assumption that a developed habit is general to different situations and is triggered by having to travel from one place to another. As a criticism of the method was the necessary extensive piloting and demanding administration, an alternative measure was created. The self-reported habit index (Verplanken and Orbell, 2003) records answers to 12 items that comment on a certain behavior on a disagree/agree scale. The assumption is that the responses capture the strength of habit by focusing on the history of repeated behavior, the difficulty of behavioral control, the absence of conscious behavior, the effectiveness of the behavior and an identity level.

More recently, habit strength was measured as the product of context stability and past repeated behavior (Wood et al., 2005). One advantage would appear to be that this measure covers both the preconditions of habit formation in a single index, although one disadvantage (common to other measures) is that this measure can be confounded with other effects (e.g., attitudes) with unknown stability over time. Accordingly, context stability seems to be an essential requisite for the prediction of future behavior and the product is a measure of strength.

Most of the theories described previously are investigated within a microeconomic approach and several studies (as mentioned in the introduction) are based on discrete choice analysis of travel behavior in different contexts. The measurement of habit in these studies is generally based on past repeated behavior that is either obtained from travel diaries or investigated with ad hoc surveys. If habit is the repeated performance of behavior, a legitimate question concerns how people arrive to specific behavioral sequences in terms of combinations of purpose, time of day, departure time, travel mode, and route. Possibly, past behavior is used again because it is the easiest and less risky solution to the need to move from one place to another. Accordingly, even though habitual choice is incorporated into transport modeling and planning, the understanding of the line between intention and habit is often blurred.

Models from revealed preference data tend to avoid direct indicators of habit. Potential indicators would be vehicle kilometers travelled and public transport trips in a unit of time for travel mode choice, measures of the activity space of each person for destination choice, and most frequent departure time for departure time choice. Models from stated preference data include more often indicators of repeated behavior. Examples of indicators are from questions about car ownership, car availability, and use of travel modes in the last week or month, and ownership of season ticket for public transport. When models are to be used for forecasting purposes, simulation at the disaggregate level does not imply problems in the short-term where the repeated behavior over time could be employed to measure habit, but could give problems in the long-term where the assumption of context stability would be difficult to justify.

Even though it appears different to measure habit because of the fact that repeated behavior is the most used indicator, even though it is conceptually something else (as extensively discussed in the previous sections), it is essential to better understand habitual choice to answer the question: what does it take to make people choose sustainable travel behavior? The success of travel demand management measures requires the ability of people to incorporate these measures into their habits.

Future Prospects

The obvious consideration about the literature on habitual travel behavior is its highly diversified character (Schwanen et al., 2012): various theoretical frameworks and measurement methods exist, although arguably a lot of research is more pragmatic in that attempts to evaluate whether interventions to change travel behavior produce the desired effect. A positive development from the methodological perspective (e.g., structural equation models, hybrid choice models) has allowed the theories to be used for empirical modeling studies that incorporate habit indicators and hence explain behavior (and habitual behavior) in more depth.

The key learning when reflecting on sustainable transport choices is that breaking habits to allow for behavioral change requires automaticity to be replaced by reasoned action. The equilibrium between habit and intention is then crucial, as the disruption of automatic responses to keys and cues is possible only when the generation of implementation intentions facilitates the emergence of new habits (Bamberg, 2000; Eriksson et al., 2008). Moreover, the recognition of the value-action gaps between intention and actual behavior is also essential, as it is true that understanding the relation between attitudes and habit is fundamental for modeling behavior, but it is also true that the gap is a major challenge for attitude theories. Incorporating habit may mitigate the value-action gaps (Schwanen et al., 2012), but using the frequency of past behavior to measure habit is somewhat tautological and cannot be confirmed empirically, especially in cross-sectional studies.

What can then be done when thinking about behavior change? Two premises must be considered (Schwanen et al., 2012): (1) the realization that habits are even more central to everyday behavior than many "habit psychologists" suggest; (2) the notion that habit itself has to be reconsidered. While we think that human beings respond automatically to certain cues, we need to think that habit is a tendency or a force. Schwanen et al. (2012) suggest that the second premise emerges from traditions in thinking about habit: the first comes from Descartes and Kant, for whom habit is a routine process that threatens reflective thought; the second comes from Aristotle, for whom moral virtue results from habit that is the mechanization of activity. The first perspective is the one that dominates the travel behavior literature presented above, while the second perspective is about stability and change. Accordingly, the literature should investigate the latter by (1) shifting the focus from repeated behavior to habit formation and renewal in order to understand habit breaking, (2) shifting the focus from the mind to the body and the environment in order to consider them to the same extent, and (3) shifting the focus on one instrument (e.g., information provision) while transcending the behavior/technology dualism. Relevantly, this effort should involve not only users of transport systems but extend to the industry, retailers, advertising companies, media, nontransport agencies, and departments, namely all the factors that contribute to the collective sense-making of different mobility forms (Schwanen et al., 2012).

A different approach is then needed because behavior change towards sustainable transport is needed to start the decarbonization of our communities from transport choices. Research might need more creativity to understand how to move towards carbon-free travel behavior by breaking habits that lead a vast amount of the population in cities all around the world to choose to travel by a private car with an internal combustion engine that is propelled by oil. Perhaps different theories and methods might need to be tested empirically with all the aforementioned stakeholders participating: auto-ethnographies, video-diaries, qualitative studies focused on discussions between researchers, and participants (also focusing on diaries and ethnographies), role-play studies engaged with perhaps performance arts (Schwanen et al., 2012). This visionary proposition will need to be accepted by an academy that often resists less traditional methods, with a positivist approach that should provide insights that will help understand the mechanisms and strategies through which travel habits can be transformed into new carbon-free ones.

References

- Aarts, H., Dijksterhuis, A., 2000. The automatic activation of goal-directed behaviour: The case of travel habit. *J. Environ. Psychol.* 20, 75–82.
- Aarts, H., Verplanken, B., van Knippenberg, A., 1997. Habit and information use in travel mode choice. *Acta Psychol.* 96, 1–14.
- Anable, J., 2005. 'Complacent car addicts' or 'aspiring environmentalists'? Identifying travel behaviour segments using attitude theory. *Transp. Policy* 12 (1), 65–78.
- Ajzen, I., 1991. The theory of planned behaviour. *Organ. Behav. Hum. Decis. Processes* 50, 179–211.
- Bamberg, S., 2000. The promotion of new behavior by forming an implementation intention: Results of a field experiment in the domain of travel mode choice. *J. Appl. Soc. Psychol.* 30, 1903–1922.

- Bamberg, S., Ajzen, I., Schmidt, P., 2003. Choice of travel mode in the theory of planned behavior: the roles of past behavior, habit, and reasoned action. *Basic. Appl. Soc. Psychol.* 25, 175–187.
- Bamberg, S., Rölle, D., Weber, C., 2003. Does habitual car use not lead to more resistance to change of travel mode. *Transportation* 30, 97–108.
- Bamberg, S., Schmidt, P., 2003. Incentives, morality, or habit? Predicting students' car use for university routes with the models of Ajzen Schwartz and Triandis. *Environ. Behav.* 35 (2), 264–285.
- Banister, D., 1978. The influence of habit formation on modal choice: a heuristic model. *Transportation* 7, 5–18.
- Banister, D., 2008. The sustainable mobility paradigm. *Transp. Policy* 15 (1), 73–80.
- Banister, D., 2011. Cities, mobility and climate change. *J. Transp. Geogr.* 19 (6), 1538–1546.
- Bauer, G., 2018. The impact of battery electric vehicles on vehicle purchase and driving behavior in Norway. *Transport. Res. Part D: Transp. Environ.* 58, 239–258.
- Bogers, E., Viti, F., Hoogendoorn, S., 2005. Joint modeling of advanced travel information service, habit, and learning impacts on route choice by laboratory simulator experiments. *Transp. Res. Rec.* 1926, 189–197.
- Chatterjee, K., 2011. Modelling the dynamics of bus use in a changing travel environment using panel data. *Transportation* 38, 487–509.
- Cherchi, E., Manca, F., 2011. Accounting for inertia in modal choices: some new evidence using a RP/SP dataset. *Transportation* 38, 679–695.
- Donald, I.J., Cooper, S.R., Conchie, S.M., 2014. An extended theory of planned behaviour model of the psychological factors affecting commuters' transport mode use. *J. Environ. Psychol.* 40 (12), 39–48.
- Eriksson, L., Garvill, J., Nordlund, A.M., 2006. Acceptability of travel demand management measures: The importance of problem awareness, personal norm, freedom, and fairness. *J. Environ. Psychol.* 26 (1), 15–26.
- Eriksson, L., Garvill, J., Nordlund, A.M., 2008. Interrupting habitual car use: the importance of car habit strength and moral activation for personal car use reduction. *Transport. Res. F: Traffic Psychol.* 11 (1), 10–23.
- Forward, S.E., 2014. Exploring people's willingness to bike using a combination of the theory of planned behavioural and the transtheoretical model. *Eur. Rev. Appl. Psychol.* 64 (3), 151–159.
- Fujii, S., Gärling, T., Kitamura, R., 2001. Changes in drivers' perceptions and use of public transport during a freeway closure: effects of temporary structural change on cooperation in a real-life social dilemma. *Environ. Behav.* 33, 796–808.
- Friedrichsmeier, S.E., Matthies, E., Klöckner, C.A., 2013. Explaining stability in travel mode choice: An empirical comparison of two concepts of habit. *Transport. Res. Part F: Traffic Psychol. Behav.* 16 (1), 1–13.
- Gardner, B., 2009. Modeling motivation and habit in stable travel mode contexts. *Transport. Res. Part F: Traffic Psychol. Behav.* 12 (1), 68–76.
- Gärling, T., Axhausen, K.W., 2003. Introduction: habitual travel choice. *Transportation* 30 (1), 1–11.
- Gärling, T., Garvill, J., 1993. Psychological explanations of participation in everyday activities. In: Gärling, T., Golledge, R.G. (Eds.), *Behaviour and Environment: Psychological and Geographical Approaches*. Elsevier/North-Holland, Amsterdam, The Netherlands, pp. 270–297.
- Haustein, S., Klöckner, C.A., Blöbaum, A., 2009. Car use of young adults: The role of travel socialization. *Transport. Res. Part F Traffic Psychol. Behav.* 12 (2), 168–178.
- Hunecke, M., Blöbaum, A., Matthies, E., Höger, R., 2001. Responsibility and environment: ecological norm orientation and external factors in the domain of travel mode choice behavior. *Environ. Behav.* 33, 830–852.
- Jansson, J., Marel, A., Nordlund, A., 2009. Elucidating green consumers: a cluster analytic approach on proenvironmental purchase and curtailment behaviors. *J. Euromarket.* 18, 245–267.
- Kaplan, S., Monteiro, M.M., Anderson, M.K., Nielsen, O.A., Dos Santos, E.M., 2017. The role of information systems in nonroutine transit use of university students: Evidence from Brazil and Denmark. *Transport. Res. Part A Policy. Pract.* 95, 34–48.
- Kaplan, S., Prato, C.G., 2012. Closing the gap between behavior and models in route choice: The role of spatiotemporal constraints and latent traits in choice set formation. *Transport. Res. Part F: Traffic Psychol. Behav.* 15 (1), 9–24.
- Klöckner, C.A., Matthies, E., 2004. How habits interfere with norm directed behavior—A normative decision-making model for travel mode choice. *J. Environ. Psychol.* 24, 319–327.
- Klöckner, C.A., Matthies, E., 2012. Two pieces of the same puzzle? Script based car choice habits between the influence of socialization and past behavior. *J. Appl. Soc. Psychol.* 42, 793–821.
- Klöckner, C.A., Matthies, E., Hunecke, M., 2003. Problems of operationalizing habits and integrating habits in normative decision-making models. *J. Appl. Soc. Psychol.* 33, 396–417.
- Lanzini, P., Khan, S.A., 2017. Shedding light on the psychological and behavioral determinants of travel mode choice: A meta-analysis. *Transport. Res. Part F: Traffic Psychol. Behav.* 48, 13–27.
- Matthies, E.A., Klöckner, C.A., Preißner, C.L., 2006. Applying a modified moral decision making model to change habitual car use: how can commitment be effective? *Appl. Psychol.* 55 (1), 91–106.
- Neal, D.T., Wood, E., Quinn, J.M., 2006. Habits: A repeat performance. *Curr. Dir. Psychol. Sci.* 15, 198–202.
- Nordlund, A.M., Garvill, J., 2002. Value structures behind proenvironmental behavior. *Environ. Behav.* 34 (6), 740–756.
- Nordlund, A., Westin, K., 2013. Influence of values, beliefs, and age on intention to travel by a new railway line under construction in northern Sweden. *Transport. Res. Part A Policy Pract.* 48 (2), 86–95.
- Ouellette, J.A., Wood, W., 1998. Habit and intention in everyday life: the multiple processes by which past behavior predicts future behavior. *Psychol. Bull.* 124, 54–74.
- Ralph, K.M., Brown, A.E., 2017. The role of habit and residential location in travel behavior change programs, a field experiment. *Transportation* 46, 719–734.
- Schwanen, T., Banister, D., Anable, J., 2012. Rethinking habits and their role in behaviour change: the case of low-carbon mobility. *J. Trans. Geogr.* 24, 522–532.
- Schwartz, S.H., 1977. Normative influences on altruism. In: Berkowitz, L. (Ed.), *Advances in Experimental Social Psychology*. Academic Press, San Diego, pp. 221–279.
- Schwartz, S.H., 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Adv. Exp. Soc. Psychol.* 25, 1–65.
- Stern, P.C., Dietz, T., Abel, T.D., Guagnano, G.A., Kalof, L., 1999. A value-belief-norm theory of support for social movements: The case of environmentalism. *Hum. Ecol. Rev.* 6 (2), 81–97.
- Stern, P.C., Dietz, T., Guagnano, G.A., 1995. The new environmental paradigm in social psychological perspective. *Environ. Behav.* 27, 723–745.
- Thorhauge, M., Swait, J., Cherchi, E., 2020. The habit-driven life: Accounting for inertia in departure time choices for commuting trips. *Transport. Res. Part A Policy Pract.* 133, 272–289.
- Triandis, H.C., 1977. *Interpersonal Behaviour*. Brooks/Cole, Monterey, CA.
- van der Waerden, P., Timmermans, H., da Silva, A.N.R., 2015. The influence of personal and trip characteristics on habitual parking behavior. *Case Stud. Transport Policy* 3, 33–36.
- Verplanken, B., Aarts, H., 1999. Habit, attitude, and planned behaviour: is habit an empty construct or an interesting case of goal-directed automaticity? *Eur. Rev. Soc. Psychol.* 10, 101–134.
- Verplanken, B., Aarts, H., van Knippenberg, A., 1997. Habit, information acquisition, and the process of making travel mode choices. *Eur. J. Soc. Psychol.* 27, 539–560.
- Verplanken, B., Aarts, H., van Knippenberg, A., Moonen, A., 1998. Habit versus planned behavior: A field experiment. *British J. Soc. Psychol.* 37, 111–128.
- Verplanken, B., Aarts, H., van Knippenberg, A., van Knippenberg, C., 1994. Attitude versus general habit: antecedents of travel mode choice. *J. Appl. Soc. Psychol.* 24, 285–300.
- Verplanken, B., Orbell, S., 2003. Reflections on past behaviour: a self-report index of habit strength. *J. Appl. Soc. Psychol.* 33, 1313–1330.
- Verplanken, B., Wood, W., 2006. Interventions to break and create consumer habits. *J. Public Policy Market.* 25 (1), 90–103.
- Wood, W., Neal, D., 2007. A new look at habits and the habit-goal interface. *Psychol. Rev.* 114, 843–863.
- Wood, W., Quinn, J.M., Kashy, D.A., 2002. Habits in everyday life: thought, emotion, and action. *J. Personal. Soc. Psychol.* 83, 1281–1297.
- Wood, W., Tam, L., Guerrero Witt, M., 2005. Changing circumstances, disrupting habits. *J. Personal. Soc. Psychol.* 88, 918–933.

Driving Behavior and Skills

Timo Lajunen*, **Türker Özkan***, *Department of Psychology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway;
†Department of Psychology, Middle East Technical University (METU), Ankara, Turkey

© 2021 Elsevier Ltd. All rights reserved.

Driver Behavior and Performance	59
Driver Behavior Questionnaire	59
Structure	59
Reliability and Validity	60
Use and Considerations	61
Driving Skill Inventory	61
Structure	61
Reliability and Validity	62
Use and Considerations	62
Conclusions	63
References	63
Further Reading	64

Driver Behavior and Performance

Car driving can be divided into two separate but interrelated components: driver behavior and driver performance. Driver behavior refers to a driver's personal driving style, that is, a way how a driver chooses to drive. Driving style or behavior includes such behaviors as preferred speed in general and in a situation given, preferred following distance, overtaking frequency, scanning of the mirrors and environment for potential hazards, and the degree of compliance with traffic rules. The driving style's central issue is the driver's level of risk-taking: driving style can be described in terms of the crash risk. Since driving behavior or style consists of driving habits, it becomes established over a period of years and can be modified similarly to any habitual behavior. Driving style does not necessarily improve with driving experience but is rather related to a driver's attitudes and personality (Elander et al., 1993). For example, drivers with hostile attitudes to other drivers and aggressive personality are more likely to have a risky driving style than drivers with positive attitudes toward others and traffic safety and nonaggressive character.

Driving skills, or in other words, a driver's level of performance and driving abilities, include information-processing and motor skills. As any skills, driving skills can be expected to improve with practice and training, that is, with driving experience (Elander et al., 1993). The importance of practice can be seen in the steep decline of crash risk early months of independent driving after getting a full license. In addition to practice, driving skills are influenced by the driver's general cognitive abilities. The role of the driver's general information-processing and motor ability is emphasized when some of those skills have declined either temporarily (e.g., driving while intoxicated) or permanently (e.g., Alzheimer's disease).

Fig. 1 displays how driving skills (i.e., driver performance) and driving style (i.e., driver behavior (Summala, 1988)) interact and finally determine a driver's crash risk. In this model, a crash is an outcome of both an error and a violation. Risky driving style characterized by violations against safe driving practices reduces the safety margin, that is, time or space available for corrective actions (Summala, 1988), while shortcomings in driving skills including both perceptual and vehicle handling skills increase the likelihood of an error. A driver's perceptual-motor skills are influenced by driving experience and general cognitive abilities. On the other hand, the driving style, including the tendency to violate formal and informal traffic rules and to deviate from safe practices, reflects a person's personality characteristics (e.g., sensation seeking, aggressiveness), attitudes (e.g., hostility), and lifestyle (e.g., alcohol use). In this way, risky personality characteristics may increase the likelihood of risky driving and traffic violations, which can lead to acceptance of narrow safety margins and, thus, less time for corrective actions in the case of an error. For example, a person scoring high in sensation and thrill-seeking is more likely to exceed the speed limit habitually.

Driver Behavior Questionnaire

Structure

Driver Behavior Questionnaire (DBQ) (Reason et al., 1990) is one of the most widely used self-report instrument for measuring aberrant driving. Document search by using the search term "Driver Behavior Questionnaire" (in title, abstract, or keywords) yielded 244 hits in Scopus and 1960 hits in Google Scholar (27.07.2020). The DBQ was initially developed by Reason and colleagues to investigate "the possible distinctions between aberrant behaviors" (Reason et al., 1990). The original fifty items were selected to cover five aberrant driving classes, that is, slips, lapses, mistakes, unintended violations, and deliberate violations. Later, Parker et al.

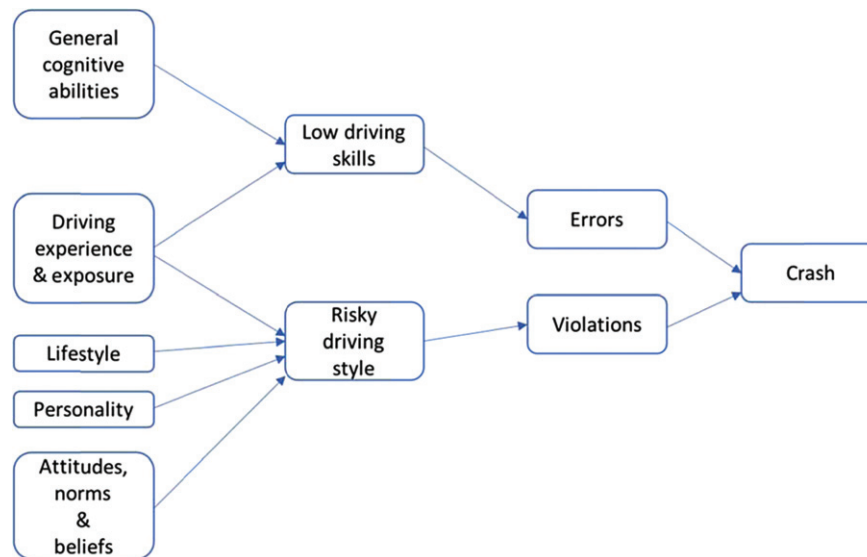


Figure 1 Two pathways to crash: low driving skills and risky driving style.

surveyed over 1600 British drivers and identified a threefold typology of aberrant driving behaviors including (1) “lapses” as “absent-minded behaviors with consequences mainly for the perpetrator, posing no threat to other road users,” (2) “errors” as “typically misjudgments and failures of observation that may be hazardous to others,” and (3) violations as “deliberate contraventions of safe driving practice” (Parker et al., 1995). Thus, an important distinctive feature between errors and violations is that errors are unintentional and violations intentional, but both are potentially risky. This distinction matches the “two pathways to crash”—driver skills and behavior—displayed in Fig. 1. The DBQ measures the failures of driver skills (error) and driving style (violations) separately. The violations were later divided into highway code violations and interpersonally aggressive violations (Lawton et al., 1997).

The last addition to DBQ is the Positive Driver Behaviors Scale (PDBS) (Özkan and Lajunen, 2005). Özkan and Lajunen (2005) aimed at extending the DBQ toward an omnibus scale, including not only aberrant driver behaviors. The positive driver behaviors aim at “taking care of smooth traffic flow or paying attention to other road users” or “politeness” (Özkan and Lajunen, 2005). The Positive Driver Behaviors Scale contains 14 items, which measure politeness and consideration for other road users.

Both the DBQ and the Positive Driver Behaviors Scale are frequency scales with the following instruction:

“For each item, you are asked to indicate HOW OFTEN, if at all, this kind of thing has happened to you. Base your judgments on what you remember of your driving over, say, the past year. Please indicate your judgments by ticking ONE of the boxes in the grid next to each item. The response alternatives are: Never, Hardly ever, Occasionally, Quite often, Frequently, Nearly all the time.”

Reliability and Validity

The DBQ has been translated into many languages, and it has been used in a large number of countries. Since there are different versions of the DBQ available from the Swedish (DBQ-SWE) 104-item scale (Åberg and Rimmö, 1998) to 9-item mini-DBQ (Martinussen et al., 2013), it is difficult to give an overview about the reliability and validity of the scale. It can be said, in general, that the DBQ scales “violations,” “errors,” and “lapses” show acceptable reliability (reliability coefficient > 0.70) in the majority of studies. The Positive Driver Behaviors Scale has been used only in four studies, and the reliability coefficients have ranged from 0.84 to 0.92 (Ersan et al., 2020; Öz et al., 2014; Özkan and Lajunen, 2005; Shen et al., 2018).

The structural (construct) validity of the DBQ has been investigated in many studies because the fit of the original supposed factor structure in a new driver population is an essential step (after translation) when adapting an existing self-report instrument to a new country, culture or language area. The DBQ factor structure has been investigated mostly by using exploratory factor analysis (EFA) but much less frequently with confirmatory factor analysis (CFA) in which the fit of the assumed (original) DBQ factor structure is investigated by calculating its fit to the newly obtained empirical data. In general, we can conclude that the original three-factor structure of the DBQ based on lapses, errors, and violations can be globally replicated with some minor deviations. As with the reliability analyses, it is difficult to draw simple conclusions about the DBQ construct validity, because of the great variety of forms of the DBQ used in different studies. It is also somewhat common that the researchers add their items to the DBQ depending on their interests, which naturally may lead to small deviations from the original factor structure.

Based on their meta-analysis about the DBQ as a predictor of crashes, De Winter and Dodou (2010) concluded that among the various versions of the DBQ, the two-factor solution (errors and violations) is the most replicable of the DBQ structures (De Winter and Dodou, 2010). Mattsson et al. (Mattsson, 2012, 2014; Mattsson et al., 2015) analyzed the DBQ factor structure by using

Exploratory Structural Equation Model (ESEM). While [Mattsson \(2012\)](#) seemed to find support for the four-factor structure of slips, lapses, aggressive violations, and highway code violations, he called for an update of the theoretical structure and the instrument itself based on recent theories of attention and human error ([Mattsson, 2012](#)). While the study by [Mattsson \(2012\)](#) has been criticized for a too-small sample size ([De Winter, 2013](#)), [Martinussen et al.](#) investigated both first- and second-order factor structures of the DBQ by using a sample of 4335 drivers ([Martinussen et al., 2013](#)). The results showed a great fit for the three-factor (violations, errors, lapses) version of the DBQ. It can be concluded that especially the two-factor (errors and violations) and three-factor (errors, violations, lapses) solutions seem to show high construct validity and reliability.

The criterion validity or predictive validity of the DBQ is difficult to assess. Since the DBQ measures aberrant (i.e., relatively infrequent) behavior, it is challenging to design experimental studies for establishing the correlation to real driver behavior. Correlation to driving in a simulator or a short test drive might not capture aberrant behavior. Correlation to other similar driver behavior instruments is also a questionable method for establishing concurrent validity because those instruments used as criteria can be even less reliable and valid than the DBQ. One way to approach the possible predictive validity of the DBQ is the correlation to risky driving outcomes, that is, a driver's involvement in crashes or penalties. In their meta-analysis of studies containing more than 45,000 respondents, [De Winter and Dodou \(2010\)](#) concluded that the DBQ errors and violations are about equally strong predictors of self-reported crashes. The same authors have found recently similar relations between the DBQ violations score correlates with actual (i.e., non-self-reported) crashes in the level of individual drivers ([De Winter et al., 2018](#)) and at the level of countries ([De Winter and Dodou, 2016](#)). The cumulating research evidence indicates that the DBQ error and violations scores are related to drivers' crash risk, which can be accepted as some degree of concurrent validity. However, it should be noted that the DBQ is a self-report of risky behavior, not crash involvement. The only real proof of validity would be a demonstrated relationship to those driving behaviors measured in actual driving, which the DBQ aims at measuring.

Use and Considerations

Researchers interested in measuring aberrant (risky) driver behavior with the DBQ should bear in mind that the DBQ is a self-report instrument in which the respondents have to report how often they have committed certain risky and sometimes illegal behaviors. As with any self-reports and especially with self-reports of unwanted behaviors, several sources of bias might influence the answers. The DBQ "lapses," for example, measure memory failures (forget where you left your car in a car park), "errors" are cognitive failures (nearly hit a cyclist who has come up on your inside), and "violations" include deliberate rule or norm violations (sound your horn to indicate your annoyance to another road user). So, the DBQ asks forgetful drivers to remember the occasions when they forgot something and the reckless drivers to report their transgressions honestly.

According to Chapman's and Underwood's (2000) study about forgetting near-accidents, about 80% of incidents were no longer reported after a delay of up to 2 weeks ([Chapman and Underwood, 2000](#)). In the DBQ, most of the behaviors are near-misses or even less than that, so we can expect that forgetting biases the DBQ answer. In terms of the DBQ violations, we could expect that socially desirable responding biases the answers to socially non-acceptable items such as those measuring aggression or driving under the influence of alcohol. Surprisingly, experimental studies about the effect of socially desirable responding (self-deception or lying) show that the social desirability bias is not a considerable problem in the DBQ responses ([Lajunen and Summala, 2003](#); [Sullman and Taylor, 2010](#)). Still, future users of the DBQ should bear in mind that forgetting and socially desirable responding are potential sources of bias in the DBQ. These biases might interact with sample characteristics, which means that forgetting and socially desirable responding is more pronounced among certain driver groups.

The best way to reduce the bias related to forgetting and socially desirable responding is to use the DBQ always as an anonymous research instrument and not for evaluating individual drivers. The instruction of the original DBQ reads: "No one is perfect. Even the best drivers make errors or commit violations at some time . . ." ([Reason et al., 1990](#)). When using the DBQ in research, it is essential to emphasize that the behaviors listed in the DBQ are not for detecting bad drivers but for studying real driving.

Driving Skill Inventory

Structure

Like the DBQ, the Driver Skill Inventory (DSI) ([Lajunen and Summala, 1995](#)) utilizes the difference between driving skills and driving style ([Fig. 1](#)). The DSI contains two self-report scales: perceptual-motor skills and safety skills. The DSI asks the respondent to evaluate his or her abilities—either perceptual-motor or safety—in terms of strength to her or his other skills. The instruction states that

"There are many differences between drivers, especially in different components of driving. We all have strong and weak components. Please indicate your strong and weak components by ticking ONE of the boxes in the grid next to each item."

The response scale includes response alternatives "Definitely weak," "Weak," "Neither weak nor strong," "Strong," and "Definitely strong." Hence, the two types of skills—perceptual-motor vehicle handling skills and safety skills—are measured in relation to each other ([Lajunen et al., 1998](#)). In some studies, a comparison to "the average driver" is used ([Martinussen et al., 2014](#)). The choice of instruction depends on the aim of the use and drastically changes the functioning of the instrument. Internal criterion ("own strong and weak characteristics") indicates whether the driver sees himself/herself mostly as a skillful driver with high vehicle

handling skills or as a safe driver, who avoids risks. If an “average driver” is used as a reference point, the DSI measures a respondent’s view of other drivers as much as himself/herself as a driver.

When the internal reference point is used, a driver’s skill or safety orientation may be calculated by dividing the safety skill score by the perceptual-motor (vehicle handling) skill score. A score above one means that a person is safety-oriented, and the score below one means that the person sees his or her perceptual-motor skills stronger than his or her safety skills. While the safety skill and perceptual-motor skills scores can be used as separate correlates in analyses, the ratio of these skills might be more useful in some situations.

A Scopus search (“Driver Skill Inventory” in all fields) returned (4 August 2020) 24 hits while Google Scholar search found 311 documents. While the DSI does not have such popularity as the DBQ, the DSI has been used in many countries and translated into several languages.

Reliability and Validity

Unlike with the DBQ with many alternative scales with different factor structures, the DSI studies have almost solely relied on the original two-factor structure containing a perceptual-skill and safety skill scale (Lajunen and Summala, 1995). The factor structure of these factors seems to be rather stable across countries and different driver groups. While most of the factor analytical studies of the DSI have utilized exploratory factor analysis (EFA), Ostapczuk et al. (2017) investigated the validity of two orthogonal DSI factors by using confirmatory factor analysis (CFA) among less experienced ($N = 130$) and experienced German drivers ($N = 1199$) and found the two-factor structure having satisfactory construct validity (Ostapczuk et al., 2017). Özkan et al. (2006) used a different approach and studied the fit of the theoretical structure by using EFA with target rotations and similarity indexes in samples collected in the UK, the Netherlands, Finland, Greece, Iran, and Turkey. Interestingly, the two-factor structures of DSI proved to be highly congruent in “safe” Northern and Western European countries while the safety skills factor found in Greek, Iranian, and Turkish sample was relatively incongruent in spite of high factor similarity found in perceptual-motor skills (Özkan et al., 2006). Recent Chinese studies applying EFA have found (Wu et al., 2020; Zhang et al., 2018, 2019), a three-factor structure the most appropriate. On the other hand, some other Chinese studies have found support to the original two factors (Shi et al., 2018; Xu et al., 2018). One reason for these small deviations might be in the sizes and types of samples: student samples, non-professional drivers and professional drivers might conceptualize safe and skillful driving differently. Small sample sizes, such as in Özkan et al. (2006), are likely to produce less stable results than large representative samples. In sum, the DSI two-factor structure based on perceptual-motor skills and safety skills seems to show high construct validity in different countries.

All studies applying the DSI have reported some type of reliability coefficient for the two DSI factors. In general, the scales have shown satisfactory internal consistency (reliability coefficients being mostly above 0.80 and often over 0.90). In the three-factor solutions, the reliability coefficients of some scales have been less impressive (around 0.70) but acceptable (Wu et al., 2018; Zhang et al., 2019). Since the internal consistency evaluated with the alpha reliability coefficient depends directly on the number of items, the scales in the three-factor solution might show lower internal consistency than the two-factor solution. It should be noted that basically all studies applying the DSI have found the two factors internally consistent.

Validation of the DSI is more complicated than the DBQ, because the DSI measures a driver’s view of his or her perceptual-motor skills or safety skills either by using an internal reference or by comparing to the “average driver.” It is essential to understand that the DSI is not an objective measure of driver skills, but rather an instrument for investigating a driver’s view of his or her skills. This view can be seen dominated by either “vehicle handling” or “safety concern” and may not reflect a driver’s real objectively measured driving skill level. Therefore, any kind of other skill measures (e.g., performance in a test track or rule knowledge) would not serve as a proper criterion for validation. Strictly speaking, it is impossible to say anything about the actual concurrent or predictive validity of the DSI because of the subjective nature of the instrument.

One way to get some idea of DSI’s possible predictive validity is to investigate if the DSI scales are related to crashes and tickets. A general finding is that the safety skills correlate negatively with crashes (Özkan and Lajunen, 2006; Özkan et al., 2006; Warner et al., 2013) and especially penalties or demerit points (Özkan et al., 2006; Wu et al., 2018; Xu et al., 2018). The relationship between high perceptual-motor skill score and crashes is much less reported (Özkan and Lajunen, 2006; Özkan et al., 2006) while the positive relationship between perceptual-motor skills and penalties seem to be more established (Özkan et al., 2006).

The above-described studies are based on self-reports and, therefore, the crash and penalty reports may be biased by socially desirable responding or forgetting. Eensoo et al. (2010) obtained the traffic violations data from the police database and the data on traffic collisions from the national traffic insurance database (Eensoo et al., 2010) and, thus, can be considered free of the problems related to self-reporting. They found out that among other variables, the overestimated driving skills in DSI were the strongest predictor of speed limit exceeding (Eensoo et al., 2010).

Relationships between DSI scales, outcome variables (crashes, penalties), and risky driving are often presented as validation of the DSI. Instead of being a proof of validity, these relationships found in several studies show that a driver’s view of himself or herself either as safety-oriented and rule-following driver or as a technically skillful driver is reflected in risky behavior and, finally, in crashes and penalties.

Use and Considerations

The DSI is a self-report instrument, so we might expect that the responses suffer from the same sources of bias as the other self-report measures. However, it should be noted that the DSI requires the respondent to evaluate the strong and weak components in their

driving. Due to this instruction, the DSI can be expected to be less vulnerable to socially desirable responding. A driver who wants to give an embellished view of himself or herself would very likely exaggerate both perceptual-motor and safety skills, resulting in a low score in safety orientation. Unlike the DBQ, the DSI does not include items, which are inherently harmful or socially undesirable. Therefore, the respondents should not have any reason to embellish their answers or lie. Ostapczuk et al. (2017) studied the relationship between the DSI and socially desirable responding (SDR) measured by the Driver Social Desirability Scale (DSDS) (Lajunen et al., 1997) and concluded that “The DSI seemed to be only marginally contaminated by SDR” (Ostapczuk et al., 2017).

Conclusions

Both DBQ and DSI are widely used scales, which have shown satisfactory reliability and construct validity. The effect of socially desirable responding, that is, impression management or self-deception, seem to be marginal. Both instruments have been demonstrated to correlate with risky driving and crashes.

References

- Åberg, L., Rimmö, P.A., 1998. Dimensions of aberrant driver behaviour. *Ergonomics* 41, 39–56.
- Chapman, P., Underwood, G., 2000. Forgetting Near-Accidents: The Roles of Severity, Culpability and Experience in the Poor Recall of Dangerous Driving Situations.. *Appl. Cognit. Psychol.* 14, 31–44.
- De Winter, J.C.F., 2013. Small sample sizes, overextraction, and unrealistic expectations: A commentary on M. Mattsson. *Accid. Anal. Prevent.* 50, 776–777.
- De Winter, J.C.F., Dodou, D., 2010. The driver behaviour questionnaire as a predictor of accidents: a meta-analysis. *J. Safety Res.* 41, 463–470.
- De Winter, J.C.F., Dodou, D., 2016. National correlates of self-reported traffic violations across 41 countries. *Personal. Individual Diff.* 98, 145–152.
- De Winter, J.C.F., Dreger, F.A., Huang, W., Miller, A., Soccolich, S., Ghanipour Machiani, S., Engström, J., 2018. The relationship between the Driver Behavior Questionnaire, Sensation Seeking Scale, and recorded crashes: A brief comment on Martinussen et al and new data from SHRP2. *Accid. Anal. Prevent.* 118, 54–56.
- Eensoo, D., Paaver, M., Harro, J., 2010. Factors associated with speeding penalties in novice drivers C3 - Annals of Advances in Automotive Medicine - 54th Annual Scientific Conference. 287–294.
- Elander, J., West, R., French, D., 1993. Behavioral correlates of individual differences in road-traffic crash risk: an examination of methods and findings. *Psychol. Bull.* 113, 279–294.
- Ersan, Ö., Üzümcüoğlu, Y., Azik, D., Findik, G., Kaçan, B., Solmaz, G., Özkan, T., Lajunen, T., Öz, B., Pashkevich, A., Pashkevich, M., Danelli-Mylona, V., Georgogianni, D., Berisha Krasniqi, E., Krasniqi, M., Makris, E., Shubenkova, K., Xheladini, G., 2020. Cross-cultural differences in driver aggression, aberrant, and positive driver behaviors. *Transport. Res. Part F: Traffic Psychol. Behav.* 71, 88–97.
- Lajunen, T., Corry, A., Summala, H., Hartley, L., 1997. Impression management and self-deception in traffic behaviour inventories. *Personal. Individual Diff.* 22, 341–353.
- Lajunen, T., Corry, A., Summala, H., Hartley, L., 1998. Cross-cultural differences in drivers' self-assessments of their perceptual-motor and safety skills: Australians and finns. *Personal. Individual Diff.* 24, 539–550.
- Lajunen, T., Summala, H., 1995. Driving experience, personality, and skill and safety-motive dimensions in drivers' self-assessments. *Personal. Individual Diff.* 19, 307–318.
- Lajunen, T., Summala, H., 2003. Can we trust self-reports of driving? Effects of impression management on driver behaviour questionnaire responses. *Transport. Res. Part F: Traffic Psychol. Behav.* 6, 97–107.
- Lawton, R., Parker, D., Manstead, A.S.R., Stradling, S.G., 1997. The role of affect in predicting social behaviors: The case of road traffic violations. *Journal of Applied Social Psychology* 27, 1258–1276.
- Martinussen, L.M., Lajunen, T., Møller, M., Özkan, T., 2013. Short and user-friendly: the development and validation of the Mini-DBQ. *Accid. Anal. Prevent.* 50, 1259–1265.
- Martinussen, L.M., Møller, M., Prato, C.G., 2014. Assessing the relationship between the Driver Behavior Questionnaire and the Driver Skill Inventory: Revealing sub-groups of drivers. *Transport. Res. Part F: Traffic Psychol. Behav.* 26, 82–91.
- Mattsson, M., 2012. Investigating the factorial invariance of the 28-item DBQ across genders and age groups: An Exploratory Structural Equation Modeling Study. *Accid. Anal. Prevent.* 48, 379–396.
- Mattsson, M., 2014. On testing factorial invariance: A reply to JCF de Winter. *Accid. Anal. Prevent.* 63, 89–93.
- Mattsson, M., Fearghal, O., Lajunen, T., Gormley, M., Summala, H., 2015. Measurement invariance of the Driver Behavior Questionnaire across samples of young drivers from Finland and Ireland. *Accid. Anal. Prevent.* 78, 185–200.
- Ostapczuk, M., Joseph, R., Pufal, J., Musch, J., 2017. Validation of the German version of the Driver Skill Inventory (DSI) and the Driver Social Desirability Scales (DSDS). *Transport. Res. Part F: Traffic Psychol. Behav.* 45, 169–182.
- Öz, B., Özkan, T., Lajunen, T., 2014. Trip-focused organizational safety climate: investigating the relationships with errors, violations and positive driver behaviours in professional driving. *Transport. Res. Part F: Traffic Psychol. Behav.* 26, 361–369.
- Özkan, T., Lajunen, T., 2005. A new addition to DBQ: Positive driver behaviours scale. *Transport. Res. Part F: Traffic Psychol. Behav.* 8, 355–368.
- Özkan, T., Lajunen, T., 2006. What causes the differences in driving between young men and women? The effects of gender roles and sex on young drivers' driving behaviour and self-assessment of skills. *Transport. Res. Part F: Traffic Psychol. Behav.* 9, 269–277.
- Özkan, T., Lajunen, T., Chliaoutakis, J.E., Parker, D., Summala, H., 2006. Cross-cultural differences in driving skills: A comparison of six countries. *Accid. Anal. Prevent.* 38, 1011–1018.
- Parker, D., Reason, J.T., Manstead, A.S.R., Stradling, S.G., 1995. Driving errors, driving violations and accident involvement. *Ergonomics* 38, 1036–1048.
- Reason, J., Manstead, A., Stephen, S., Baxter, J., Campbell, K., 1990. Errors and violations on the roads: A real distinction? *Ergonomics* 33, 1315–1332.
- Shen, B., Qu, W., Ge, Y., Sun, Z., Zhang, K., 2018. The relationship between personalities and self-report positive driving behavior in a Chinese sample. *PLoS ONE* 13.
- Shi, J., Xiao, Y., Tao, L., Atchley, P., 2018. Factors causing aberrant driving behaviors: A model of problem drivers in China. *J. Transport. Safety Security* 10, 288–302.
- Sullman, M.J.M., Taylor, J.E., 2010. Social desirability and self-reported driving behaviours: Should we be worried? *Transport. Res. Part F: Traffic Psychol. Behav.* 13, 215–221.
- Summala, H., 1988. Risk control is not risk adjustment: The zero-risk theory of driver behaviour and its implications. *Ergonomics* 31, 491–506.
- Warner, H.W., Özkan, T., Lajunen, T., Tzamaloukas, G.S., 2013. Cross-cultural comparison of driving skills among students in four different countries. *Safety Sci.* 57, 69–74.
- Wu, C., Chu, W., Zhang, H., Özkan, T., 2018. Interactions between driving skills on aggressive driving: Study among Chinese drivers. *Transport. Res. Rec.* 2672, 10–20.
- Wu, L., Li, H., Ding, H., Zhang, L., 2020. Who has better driving style: Let data tell us C3. *Advances in Intelligent Systems and Computing* 1037, 467–485.
- Xu, J., Liu, J., Sun, J., Zhang, K., Qu, W., Ge, F.Y., 2018. The relationship between driving skill and driving behavior: Psychometric adaptation of the Driver Skill Inventory in China. *Accid. Anal. Prevent.* 120, 92–100.

- Zhang, W., Zhang, X., Feng, Z., Liu, J., Zhou, M., Wang, K., 2018. The fitness-to-drive of shift-work taxi drivers with obstructive sleep apnea: An investigation of self-reported driver behavior and skill. *Transport. Res. Part F: Traffic Psychol. Behav.* 59, 545–554.
- Zhang, Z., Ma.F.T., Ji.F.N., Hu, Z, Zhu, W., 2019. An assessment of the relationship between driving skills and driving behaviors among Chinese bus drivers. *Adv. Mech. Eng.* 11 (1), 1–10.

Further Reading

- Evans, L., 2004. *Traffic Safety*. Science Serving Society.
- Porter, B.E., 2011. *Handbook of Traffic Psychology*. Academic Press.
- Shinar, D., 2017. *Traffic Safety and Human Behavior*. Emerald Group Publishing.

The Multidimensional Driving Style Inventory

Orit Taubman – Ben-Ari, Bar Ilan University, Ramat Gan, Israel

© 2021 Elsevier Ltd. All rights reserved.

Driving Styles as a Multidimensional Concept	65
The Multidimensional Driving Style Inventory: Theoretical Principles and Conceptualization	65
Sociodemographic and Driving-related Variables	66
MDSI and Personality	68
MDSI and Familial and Social Contexts	69
Appendix 1: Examples of Studies Using the MDSI	71
References	72

Driving Styles as a Multidimensional Concept

Continuous efforts to understand human road behavior and prevent risky driving and traffic violations have led to the concept of driving styles. Driving style is defined as the way the driver chooses to drive or habitually drives, including decisions on speed, headway, and level of attentiveness and assertiveness behind the wheel (Elander et al., 1993). This is not the same as driving skill, which reflects performance or the ability to maintain control of the vehicle and respond adaptively to various complex and demanding traffic contexts (Elander et al., 1993). Whereas driving skill is a behavioral component and is expected to improve with practice or training, driving style is a psychological characteristic influenced by the individual's values and needs in general, as well as specific attitudes and beliefs regarding driving. In other words, driving style is a subjective feature, while driving skill is considered more objective and can therefore be observed and assessed externally and/or technically.

Over the years, the concept of driving style has evolved from a unidimensional perception of risky driving as a reflection of drivers' recklessness and dangerous behavior in general, to a broader multidimensional view of driving behaviors as deriving from a variety of motivational factors. Several attempts have been made to assess driving styles, and particularly their negative and potentially harmful aspects, such as stress, errors, violations, and driving anxiety, as well as aggressive, risky, or dangerous driving. The tools developed for this purpose include: Driving Behavior Inventory (DBI, Gulian et al., 1988, 1989); Driving Style Questionnaire (DSQ, French et al., 1993); The Attitudes to Driving Violations (ADVS, West and Hall, 1997); Driver Behavior Questionnaire (DBQ, Reason et al., 1990); Drivers Behavior Questionnaire (Furnham and Saipe, 1993), Driving Vengeance Questionnaire (DVQ, Wiesenthal et al., 2000); and Dula Dangerous Driving Index (DDDI; Dula and Ballard, 2003). All these instruments share the premise that driving styles reflect individual differences and a habitual way of driving, and thus are fairly stable (Sagberg et al., 2015).

In contrast, the *Multidimensional Driving Style Inventory* (MDSI), first published by Taubman – Ben-Ari, Mikulincer, and Gillath in 2004, is based on the assumption that driving styles may be more comprehensive, including positive as well as negative characteristics of the driver, and may stem from both conscious and unconscious factors that come together to determine how a person drives. The MDSI was designed to conceptualize an individual's habitual driving style as a driving-specific behavior that can explain the extent of involvement in car crashes and traffic violations, both directly and in combination with more general sociodemographic and personality characteristics (Taubman – Ben-Ari et al., 2004). Since its publication, the MDSI has been used in studies around the world, in addition to Israel, where the instrument was originally developed (see Table (Appendix 1) for examples of published international uses of the inventory).

The Multidimensional Driving Style Inventory: Theoretical Principles and Conceptualization

Three principles underlie construction of the MDSI. One, that the complex and multidimensional nature of the phenomenon of driving is best reflected through driving styles. Two, that existing definitions and scales of driving styles need to be integrated into a single, multidimensional conceptualization. Three, that whereas most earlier research related to driving behaviors associated with crashes, it is important to broaden the scope to include characteristics of driving in general in order to understand the whole range of variables that predict how people drive, including safe, careful, and considerate behaviors.

This approach led to the design of a scale assessing four differentiated driving styles: (1) reckless and careless; (2) anxious; (3) angry and hostile; and (4) patient and careful. The *reckless and careless style* refers to deliberate violations of safe driving norms and the seeking of sensations and thrills while driving. It is the most salient reflection of risky driving behavior, characterized by driving at high speeds, passing in no-passing lanes, and driving while intoxicated. The *anxious style* relates to feelings of vigilance, nervousness, and tension, as well as ineffective engagement in relaxing activities during driving. This style characterizes people who feel unsure of themselves as drivers, and reflects their sense of helplessness and insecurity. The *angry and hostile style* refers to expressions of

irritation, rage, and hostility on the road, such as cursing other drivers, honking the horn, or flashing headlights. It indicates aggression accompanied by a motivation to beat out other vehicles in competitive traffic situations. The *patient and careful style* relates to well-adjusted driving behaviors, such as planning ahead, paying full attention to the road, displaying patience, courtesy, and calm behind the wheel, and obeying the traffic rules. It characterizes drivers who are considerate and sympathetic to other drivers' needs, and who exhibit understanding of complex traffic situations.

Initial studies revealed eight coherent, valid, and reliable factors that could be grouped into these four global styles (Taubman – Ben-Ari et al., 2004): risky and high velocity driving (reckless and careless style); anxious, dissociative, and distress reduction driving (anxious style); angry and hostile driving; and careful and patient driving. Most subsequent studies have therefore used the four-factor version of the inventory. However, reasonable to high reliabilities and good validity have been found for both the four global styles and the eight factors. In addition, some adaptations of the MDSI contain six factors.

The MDSI has been translated into numerous languages (among them, Italian, Spanish, Romanian, and Mandarin Chinese), and used in various countries (e.g., Australia, England, Italy, Argentina, Spain, the United States, the Netherlands, Switzerland). It has been culturally adapted for Argentina (Poó and Ledesma, 2013; Trógolo et al., 2020), Spain (Padilla et al., 2020), Romania (Holman and Havârneanu, 2015), and China, where two different versions have been developed (Sun and Chang, 2019; Wang et al., 2018). It has also been validated in the Netherlands and the Flemish-speaking region of Belgium (Hooft van Huysduynen et al., 2015), Malaysia (Karjanto et al., 2017), Bulgaria (Totkova and Racheva, 2019), and Italy (Freuli et al., 2020), as well as among professional drivers in Colombia (Useche et al., 2019b). Overall, the results of these studies attest to the universal applicability and robustness of the inventory.

Sociodemographic and Driving-related Variables

Most of the sociodemographic characteristics examined over the years have not yielded significant associations with the four driving styles. This suggests that driving styles are not affected by background variables such as education, health, economic status, or family status, but rather represent relatively stable and universal traits. However, findings do show quite systematically that gender is a strong predictor of driving styles, with men engaging in the reckless and careless and the angry and hostile styles more than women, and women adopting the careful and patient style more than men. Age appears to play a role as well, so that the three maladaptive driving styles tend to lessen over the years, whereas the careful and patient style becomes more dominant as people get older (Taubman – Ben-Ari and Skvirsky, 2016).

In regard to driving-related variables, all the indicators of risky driving (e.g., involvement in car crashes, number of traffic violations, assessment of risky driving as dangerous), whether assessed by self-report techniques or by the use of a simulator or a car equipped with an in-vehicle data recorder (IVDR), have been found to be coherently and systematically related to the MDSI driving styles in the expected directions. These findings not only indicate that maladjusted driving is associated with a higher frequency of risky behaviors on the road, but also provide validation of the MDSI as an efficient measurement tool. This is valuable information, enabling the inventory to be used to identify baseline driving styles before road safety interventions, as well as for post-intervention assessment. It may also be employed to enhance drivers' awareness of their own and others' driving styles, either by clinicians in individual or family sessions or by safety officers as part of programs aimed at effecting organizational culture change. The MDSI can therefore inform the efforts of those striving to enhance traffic safety at all stages of their work: planning, dissemination, and assessment.

Following are the background variables associated with each of the driving styles.

Reckless and careless style: Men are systematically found to engage in the reckless and careless driving style more than women (e.g., Holland et al., 2010; Navon and Taubman – Ben-Ari, 2019; Padilla et al., 2020; Poó and Ledesma, 2013; Pugnetti and Elmer, 2020; Sun and Chang, 2019; Taubman – Ben-Ari, 2018; Taubman – Ben-Ari and Yehiel, 2012). Similar results have been yielded by lower age (e.g., Farah et al., 2007; Gwyther and Holland, 2012; Hooft van Huysduynen et al., 2018; Taubman – Ben-Ari et al., 2004; Taubman – Ben-Ari and Yehiel, 2012). In addition, drivers who have been involved in any kind of car crash consistently score higher on this style than those who have not (Liang and Lin, 2018; Poó et al., 2013; Sun and Chang, 2019; Taubman – Ben-Ari et al., 2004; Taubman – Ben-Ari and Skvirsky, 2016), as do drivers with a history of traffic violations or who report both ordinary and aggressive violations in questionnaires such as the DBQ (e.g., Holman and Popusoi, 2018; Liang and Lin, 2018; Padilla et al., 2020; Sun and Chang, 2019; Wang et al., 2018). Along the same lines, reoffenders endorse the reckless and careless style more than non-offenders (Faílde-Garrido et al., 2016; Padilla et al., 2020), and the rate of occupational driving crashes in the previous two years among professional drivers was significantly and positively linked to this driving style (Useche et al., 2019b).

The same picture emerges from a set of self-report relevance assessments. Thus, both self-reported **risky driving habits** and the **willingness of drivers to take risks while driving** were positively associated with the reckless and careless style, while an inverse association was found between this style and the **assessment of reckless driving as risky** (Taubman – Ben-Ari and Skvirsky, 2016).

Exposure, or the average amount of time the individual drives on a regular basis, has also been related to this driving style. Higher exposure has been positively associated with high speed driving, suggesting that the more time a person spends behind the wheel, the greater their tendency to drive recklessly (Holland et al., 2010). This might be related to the finding that young drivers who have their own car report higher reckless and careless driving than those who do not own a car (Taubman – Ben-Ari and Skvirsky, 2016). However, as with age, driving experience, that is the length of time the individual has had a driving license, appears to moderate risky driving, correlating negatively with the reckless and careless style (Gwyther and Holland, 2012; Taubman – Ben-Ari, 2018).

Furthermore, the reckless and careless style has been found to correlate with **performance measures** collected in a driving simulator (e.g., Bellem et al., 2018; Farah et al., 2007). Higher scores on this style were positively associated with speed, number of completed passing maneuvers, critical gaps (Farah et al., 2007), speeding, braking, steering, lateral positioning, and maintaining distance from the vehicle ahead (Hooft van Huysduynen et al., 2018). Similarly, this style corresponded with real driving data collected by an IVDR, correlating positively with the rate of risky events recorded (Taubman – Ben-Ari, Eherenfreund-Hager et al., 2016). Finally, lower hazard perception in various domains, such as social, financial, and health, has been associated with higher reckless and careless driving (Padilla et al., 2020).

Anxious style: Women tend to endorse the anxious style more than men (e.g., Holland et al., 2010; Poó and Ledesma, 2013; Poó et al., 2013; Sun and Chang, 2019; Taubman – Ben-Ari, 2006, 2014a). In addition, a negative association has been found between age and the anxious style (e.g., Navon and Taubman – Ben-Ari, 2019; Taubman – Ben-Ari et al., 2004), and its dissociative component in particular (Trógolo et al., 2014).

Drivers who had committed **traffic violations** also scored higher on this style than those who had not or those who reported fewer violations, (e.g., Padilla et al., 2020; Sun and Chang, 2019; Wang et al., 2018). In the same vein, DBQ ordinary violations (Holman and Popusoi, 2018), aggressive violations, and errors and lapses (Wang et al., 2018) were all positively associated with the anxious driving style, as was dissociative driving, which correlates substantively and positively with the DBQ subscales Lapses and Errors (Padilla et al., 2020). Similarly, both younger and older drivers who had been involved in a **car crash** resulting in fatalities reported higher anxious driving than those without this history (Taubman – Ben-Ari and Skvirsky, 2016). On the other hand, higher **exposure** (e.g., Gwyther and Holland, 2012; Holland et al., 2010) and lower **experience** (e.g., Holman and Havârneanu, 2015; Karjanto et al., 2017; Sun and Chang, 2019) have been associated with the anxious driving style, indicating that the more a person drives and the more experience they have behind the wheel, the less anxious their driving.

In regard to **performance measures**, higher critical passing gaps in a driving simulator were found among drivers scoring higher on the anxious style (Farah et al., 2007). This style also correlated negatively with the rate of risky events recorded by an IVDR (Taubman – Ben-Ari, Eherenfreund-Hager et al., 2016). However, self-reported **risky driving habits** have been positively associated with the anxious driving style (Taubman – Ben-Ari and Skvirsky, 2016).

Angry and hostile style: Men report driving in this manner more than women (e.g., Poó et al., 2013; Padilla et al., 2020; Pugnetti and Elmer, 2020; Taubman – Ben-Ari, 2006, 2018), and lower age is consistently associated with higher endorsement of the angry and hostile driving style (e.g., Farah et al., 2007; Gwyther and Holland, 2012; Navon and Taubman – Ben-Ari, 2019; Sun and Chang, 2019; Taubman – Ben-Ari, 2014a). Furthermore, drivers who had committed **traffic violations** and been involved in **car crashes** scored significantly higher on this style (e.g., Liang and Lin, 2018; Padilla et al., 2020; Poó et al., 2013; Sun and Chang, 2019; Taubman – Ben-Ari and Skvirsky, 2016), including professional drivers (Useche et al., 2019b, 2020). Similarly, higher scores on both DBQ ordinary and aggressive violations were associated with the angry and hostile style (Holman and Popusoi, 2018; Wang et al., 2018), and traffic reoffenders scored higher on this style than nonoffenders (Padilla et al., 2018, 2020).

In addition, a set of self-report measures relating to driving revealed a negative association between this style and the **assessment of reckless driving as risky**, and positive associations with the reported frequency of **risky driving habits** and **willingness to take risks while driving** (Taubman – Ben-Ari and Skvirsky, 2016). Driving **experience** also correlated negatively with the angry and hostile style (Gwyther and Holland, 2012). As to **performance measures**, participants with lower critical passing gaps in a driving simulator scored higher on this driving style (Farah et al., 2009), and the rate of risky events assessed by an IVDR was positively associated with the angry and hostile style (Taubman – Ben-Ari, Eherenfreund-Hager et al., 2016).

Patient and careful style: Women tend to endorse the patient and careful style more than men (e.g., Holland et al., 2010; Poó and Ledesma, 2013; Pugnetti and Elmer, 2020; Taubman – Ben-Ari, 2018; Taubman – Ben-Ari and Yehiel, 2012), and it has also been related to older age (e.g., Hooft van Huysduynen et al., 2018; Padilla et al., 2020; Poó et al., 2013; Sun and Chang, 2019; Taubman – Ben-Ari and Yehiel, 2012). In addition, this style has been associated with lower previous involvement in **car crashes** (Poó et al., 2013; Sun and Chang, 2019; Taubman – Ben-Ari et al., 2004; Taubman – Ben-Ari and Skvirsky, 2016), and fewer self-reported driving violations (Fáilde-Garrido et al., 2016; Holman and Popusoi, 2018; Wang et al., 2018), with non-offenders scoring higher than others (Padilla et al., 2018, 2020). Moreover, the patient and careful style correlated positively with the DBQ positive driving behaviors and negatively with the dangerous driving behaviors (Padilla et al., 2020; Wang et al., 2018). Higher **assessment of reckless driving as risky** and lower **risky driving habits** and **willingness to take risks behind the wheel** have also been associated with the patient and careful style (Taubman – Ben-Ari and Skvirsky, 2016). In line with these findings, in an on-road experiment examining when drivers choose to deactivate full-range adaptive cruise control (ACC), Varotto et al. (2018) found that patient and careful drivers resumed manual control or regulated the target speed when the sense of risk was higher in low-risk situations and lower in high-risk situations. Thus, risk perceptions were related to this driving style not only when measured by self-report instruments, but also in simulated driving.

Higher **exposure** has been negatively associated with the patient and careful style (Holland et al., 2010), indicating that the more a person drives, the less patient their driving. However, driving **experience** has yielded a positive association with this style (Gwyther and Holland, 2012; Holman and Havârneanu, 2015; Karjanto et al., 2017; Sun and Chang, 2019), suggesting that over the years a person's driving tends to become more patient and careful (Taubman – Ben-Ari and Skvirsky, 2016). In regard to **performance measures**, higher critical passing gaps in a simulator were found among drivers scoring higher on the patient and careful style (Farah et al., 2007), and the rate of risky events recorded by an IVDR correlated negatively with this style (Taubman – Ben-Ari, Eherenfreund-Hager et al., 2016).

MDSI and Personality

The question of the associations between the MDSI factors and personality traits, cognitive skills, and emotional characteristics has also aroused the interest of researchers. Some of the variables investigated are directly connected to risky behavior or aggressive tendencies, both validating the relevant driving styles and deepening our understanding of their fundamental premises. Others are more distal, enabling us to extend what can be observed and understood from external behavior to the underlying motivations and deeper levels of the self that lead to the choice of a particular driving style. The associations found between personal variables and each of the driving styles are presented below.

Reckless and careless style: This mode of driving reflects a maladaptive approach by which individuals consciously choose to take risks to achieve enjoyment and arousal. It is usually accompanied by positive affect stemming from the pleasure and gratification generated by the thrills and excitement of the driving experience. Thus, this style, which endangers both the driver and other road users, is spurred by a self-focused motivation (Taubman – Ben-Ari, 2008) reflected in the desire to satisfy one's ever-growing need for sensations. Self-esteem, sense of coherence, and differentiation of self, which represent adaptive and healthy personality traits (Antonovsky, 1979; Bowen, 1978; Rosenberg, 1979), have all been inversely associated with the reckless and careless driving style (Miller and Taubman – Ben-Ari, 2010; Taubman – Ben-Ari, 2014a; Taubman – Ben-Ari et al., 2004). Significant inverse correlations have also been found between this style and the Big Five dimensions (John and Srivastava, 1999) of agreeableness and conscientiousness (Taubman – Ben-Ari and Yehiel, 2012; Wang et al., 2018). In contrast, sensation seeking, which is frequently associated with risky behaviors, has been positively related to the endorsement of this style among both younger and older drivers (Holman and Havârneanu, 2015; Karjanto et al., 2017; Sun and Chang, 2019; Poó and Ledesma, 2013; Taubman – Ben-Ari et al., 2004; Taubman – Ben-Ari and Skvirsky, 2016), and especially among young men (Miller and Taubman – Ben-Ari, 2010; Poó and Ledesma, 2013). Similarly, positive correlations have been found between the reckless and careless driving style and the aggression–hostility personality trait (Poó and Ledesma, 2013), impulsivity (Totkova 2020), and sensitivity to reward, all of which are indications of intense sensation seeking (Padilla et al., 2020). Other traits positively associated with the reckless and careless style relate to difficulty regulating emotional states, such as impulse control difficulties (Navon and Taubman – Ben-Ari, 2019; Trógolo et al., 2014), nonacceptance of emotional responses, and difficulties engaging in goal-directed behavior (Trógolo et al., 2014). Moreover, emotion regulation was found to play an important role in explaining why young drivers tend to engage in this driving style, with the younger the driver, the more they reported emotion regulation difficulties, and the more they reported emotion regulation difficulties, the more they tended to drive recklessly (Navon-Eyal and Taubman – Ben-Ari, 2020). In addition, lower trait forgivingness, which reflects a difficulty regulating negative emotions, contributed to a higher reported tendency to drive recklessly (Navon and Taubman – Ben-Ari, 2019).

Finally, among young drivers, general conformity with authority has been inversely related to the reckless and careless style (Taubman – Ben-Ari and Katz-Ben-Ami, 2012). In other words, drivers endorsing this style tend to focus on fulfilling their desire for adventure and the immediate gratification of their needs, coupled by a lack of respect for authority and the law.

Anxious style: This is another maladaptive driving style that endangers the driver and other road users, but for completely different reasons. Paradoxically, it is motivated by the need to protect the self along with the fear of others' behavior. Anxious drivers are characterized by hesitation and the inability to make effective decisions. They lack confidence in their driving skills, and view other drivers as an additional source of stress (Taubman – Ben-Ari, 2008). The personality characteristics found to correlate positively with the anxious driving style include: trait anxiety and neuroticism, which are basic signs of psychological imbalance (Miller and Taubman – Ben-Ari, 2010; Poó and Ledesma, 2013; Taubman – Ben-Ari et al., 2004; Taubman – Ben-Ari and Yehiel, 2012); external locus of control (Holland et al., 2010), which refers to the belief that the outcomes of events are controlled by external forces, such as fate, luck, or powerful others (Rotter, 1966); higher alcohol dependence (Padilla et al., 2020); sensitivity to punishment, which is a reflection of a greater sensitivity to the possible dangers of the road (Padilla et al., 2020); and impulsive sensation seeking, which has been associated in particular with the dissociative and distress reduction components of this driving style (Holman and Havârneanu, 2015; Poó and Ledesma, 2013). In contrast, personality characteristics considered constructive resources have been shown to correlate inversely with the anxious style. These include self-esteem (associated with the dissociative component; Taubman – Ben-Ari et al., 2004), extraversion (Taubman – Ben-Ari et al., 2004), the Big Five dimension of conscientiousness (Taubman – Ben-Ari and Yehiel, 2012; Wang et al., 2018), and desire for control (Holman and Havârneanu, 2015; Karjanto et al., 2017). Among young drivers, sense of coherence (Taubman – Ben-Ari, 2014a) and perceived self-efficacy as a driver (Miller and Taubman – Ben-Ari, 2010) have also yielded inverse associations with anxious driving.

More impulse control difficulties and limited access to emotion regulation strategies, two dimensions of emotional dysregulation (Gratz and Roemer, 2004), have also been associated with higher anxious driving (Navon and Taubman – Ben-Ari, 2019), as have lack of emotional awareness, lack of emotional clarity, nonacceptance of emotional response, difficulties engaging in goal-directed behavior, and limited access to emotion regulation strategies. In addition, emotion regulation fully mediated between age and the anxious driving style, so that the younger the driver, the more they reported emotion regulation difficulties, which in turn was related to higher endorsement of the anxious style (Navon-Eyal and Taubman – Ben-Ari, 2020). Finally, lower trait forgivingness contributed to a higher reported tendency to drive in the anxious style (Navon and Taubman – Ben-Ari, 2019). Anxious drivers are therefore impelled by their fears and lower ego resources.

Angry and hostile style: This style, which is also maladaptive, relates to aggressive driving and the desire to compete with other road users, who are considered less competent. It is expressed in violent behavior on the road accompanied by negative emotions such as anger, antagonism, rage, nervousness, and irritability. Although the driving behavior represented by this style may resemble

that of the reckless and careless style, the difference lies in the underlying motivation. Here the driver is not looking for excitement, but is fulfilling a need to win out over others and vent frustrations (Taubman – Ben-Ari, 2008). Thus, it is not a reflection of happiness and pleasure, but of anger and impatience. Angry and hostile driving has yielded positive correlations with driving anger dimensions (as measured by three factors: traffic obstructions, illegal behavior, and hostile gestures; Padilla et al., 2020), sensation seeking (Holman and Havârneanu, 2015; Miller and Taubman – Ben-Ari, 2010; Poó and Ledesma, 2013), the aggression–hostility personality trait (Poó and Ledesma, 2013), higher difficulties in impulse control (Navon and Taubman – Ben-Ari, 2019; Trógolo et al., 2014), higher desire for control (Taubman – Ben-Ari et al., 2004), and the Alcohol Dependence Syndrome (Padilla et al., 2020). Inverse correlations have been found with the Big Five factors of agreeableness and conscientiousness (Taubman – Ben-Ari and Yehiel, 2012), forgivingness (Navon and Taubman – Ben-Ari, 2019), sense of coherence, and perception of meaning in life (Taubman – Ben-Ari, 2014a). Moreover, difficulties regulating emotions fully mediated between age and this style, so that the younger the driver, the more they reported emotion regulation difficulties, which in turn was associated with higher endorsement of the angry and hostile style (Navon-Eyal and Taubman – Ben-Ari, 2020).

Thus, drivers adopting the angry and hostile style are both aggressive and easily irritated, and have less faith in the benevolence of the world. They are impatient of other drivers or complex traffic situations, and tend to react both impulsively and vehemently.

Careful and patient style: This is an adaptive and positive driving style that represents commitment to safety, higher values of caring and empathy for others and the environment, and respect for the law. Careful drivers are motivated by conscientiousness, consideration, and awareness of the consequences of their driving for other road users. Emotions are rarely involved in this driving style. Instead, it is regulated by systematic planning ahead, self-control and restraint (Taubman – Ben-Ari, 2008), and low emotional conflicts or difficulties (Navon-Eyal and Taubman – Ben-Ari, 2020). The careful style has been positively related to self-esteem, desire for control (Holman and Havârneanu, 2015; Karjanto et al., 2017; Taubman – Ben-Ari et al., 2004), differentiation of self (especially among young female drivers; Miller and Taubman – Ben-Ari, 2010), the Big five factors of agreeableness, conscientiousness, and openness (Taubman – Ben-Ari and Yehiel, 2012; Wang et al., 2018), sense of coherence (Taubman – Ben-Ari, 2014a), and higher forgivingness (Navon and Taubman – Ben-Ari, 2019). It has also been associated with greater perception of meaning in life (Taubman – Ben-Ari, 2014a), and a higher ethical stance of idealism (Holman and Popusoi, 2018), reflecting the principle that a morally correct action will always result in a positive outcome for others, and therefore leading to a reluctance to break rules (Forsyth, 1992). Among young drivers, it has also been associated with higher perceived self-efficacy as a driver (Miller and Taubman – Ben-Ari, 2010).

In contrast, careful driving has been shown to correlate inversely with sensation seeking (Holman and Havârneanu, 2015; Poó and Ledesma, 2013; Taubman – Ben-Ari et al., 2004), impulsivity (Totkova, 2020), trait anxiety and neuroticism (Taubman – Ben-Ari et al., 2004), and the emotion regulation dimensions of difficulties with impulse control and emotional awareness (Navon and Taubman – Ben-Ari, 2019; Trógolo et al., 2014). In other words, drivers displaying this style drive safely due not only to their tendency to obey traffic laws, but also to their idealism, values of prudent driving, care for others, and the tendency to be positive and considerate.

MDSI and Familial and Social Contexts

The third set of variables that has been examined in relation to the MDSI are environmental spaces in which driving takes place, including the family as a whole, individual family members (parents, siblings, grandparents), peers, and friends. This issue is relevant mainly to young drivers, as it relates to the context in which their driving style is shaped during their early experience behind the wheel.

Most investigations of family members have focused on the influence of the parents, and have produced considerable evidence of the intergenerational transmission of driving styles (Gil et al., 2016; Miller and Taubman – Ben-Ari, 2010; Taubman – Ben-Ari et al., 2005). This suggests that parents represent a model for their offspring, probably from early childhood, when children observe their parents' driving behavior long before they themselves have a license (Taubman – Ben-Ari, 2014b).

In addition to modeling, the general atmosphere in the family has also been related to the children's driving styles. Hence, studies have shown that the higher the family's level of cohesion and adaptability (two aspects of Olson's FACES; Olson, 1991), the lower the young driver's tendency to endorse any of the three maladaptive styles and the higher their preference for the careful and patient style (Gil et al., 2016; Taubman – Ben-Ari and Katz-Ben-Ami, 2013).

A third component of the parent-child relationship that has been examined extensively in this context is the multidimensional concept of Family Climate for Road Safety (FCRS; Taubman – Ben-Ari and Katz-Ben-Ami, 2013), which relates to seven dimensions of the family's influence on teens' driving habits: Modeling, Feedback, Communication, Monitoring, Commitment to Safety (represented negatively in the Family Climate for Road Safety Scale as Noncommitment to Safety), Messages, and Limits. Studies reveal that all positive aspects of FCRS are inversely associated with the reckless and careless style among young drivers, whereas higher family's lack of commitment to safety is positively related to this style (Skvirsky et al., 2017; Taubman – Ben-Ari and Katz-Ben-Ami, 2012, 2013). Furthermore, noncommitment to safety and messages were found to mediate the relationship between the family's perceived cohesion and father's reckless and anxious driving styles on the one hand, and novice male drivers' driving styles on the other. More specifically, higher endorsement of the reckless and careless style by fathers and family relations characterized by lower cohesion were related to young drivers' greater perception of the family as uncommitted to safety. This, in turn, was associated with higher endorsement of both the reckless and careless and the angry and hostile styles, as well as a lower preference for the patient and careful style, on the part of the young driver (Gil et al., 2016).

Reckless and careless style: Father–son dyads produced the strongest associations in this style, indicating the likelihood that the sons of fathers who model reckless driving will adopt the same style (Taubman – Ben-Ari, 2016). Examination of FCRS among parents of young drivers showed that their own reckless and careless driving style correlated inversely with the positive aspects of family climate, and positively with noncommitment to safety (Taubman – Ben-Ari, 2015). Moreover, negative correlations were found between parents' positive FCRS dimensions and the reckless and careless style reported by their children, while a positive association emerged between parents' noncommitment to road safety and their children's reckless driving style (Taubman – Ben-Ari, 2016).

In addition, a study of young drivers who also reported on their older siblings' driving style (Taubman – Ben-Ari, 2018) found that the higher the young driver perceived their sibling to exhibit the reckless and careless style, the higher their own endorsement of this manner of driving, with both adopting it to a similar degree. Furthermore, the higher the siblings' conflictual relationship, the more the young driver tended to adopt the reckless and careless driving style.

In a study investigating the issue of grandparents (Taubman – Ben-Ari, Findler et al., 2016), higher endorsement of the reckless driving style was significantly related to higher grandparents' reports of tension when driving their grandchildren, as well as to less positive attitudes to the use of child restraints.

Peer pressure is also linked to the reckless and careless style of young drivers, so that greater attentiveness to peers was associated with higher maladaptive driving and lower safety (Taubman – Ben-Ari and Katz-Ben-Ami, 2012). Moreover, lower peer commitment to safe driving was related to this driving style (Skvirsky et al., 2017).

Anxious style: Intergenerational transmission has been found for this manner of driving as well (Gil et al., 2016; Miller and Taubman – Ben-Ari, 2010; Taubman – Ben-Ari et al., 2005). In addition, the higher the family's level of cohesion and adaptability (Olson, 1991), the lower the young driver's tendency to drive anxiously (Gil et al., 2016; Taubman – Ben-Ari and Katz-Ben-Ami, 2013). The anxious driving style was also found to correlate positively with the FCRS dimensions of monitoring and limits (Taubman – Ben-Ari and Katz-Ben-Ami, 2012, 2013), indicating that firm discipline on the part of parents does not guarantee the adaptive behavior of their children, but rather leads to greater anxiety and nervousness on the road. Links have also been found between the anxious style and a lower tendency in the family to communicate and set boundaries for young drivers (Skvirsky et al., 2017). Moreover, among parents of young drivers, the positive FCRS dimensions (save for monitoring) correlated inversely both with their own anxious style (Taubman – Ben-Ari, 2015), and with their offspring's anxious driving (Taubman – Ben-Ari, 2016). Finally, a positive association was found between parents' noncommitment to road safety and their children's adoption of the anxious driving style (Taubman – Ben-Ari, 2016).

Young drivers' self-reported anxious style was similarly associated with their perception that their sibling drove in this manner. Here the young driver showed an even higher level of anxiety behind the wheel than their older sibling. Relations between the two played a role in this driving style as well: the warmer and closer the relationship, the lower the young drivers' anxious driving, while the more conflictual the relationship, the greater their tendency to drive anxiously (Taubman – Ben-Ari, 2018).

Furthermore, here too, anxious driving style was significantly related to higher grandparents' reports of tension when driving their grandchildren, and to less positive attitudes to the use of child restraints (Taubman – Ben-Ari, Findler et al., 2016).

Higher peer pressure (Skvirsky et al., 2017; Taubman – Ben-Ari and Katz-Ben-Ami, 2012) and perceived social costs of driving with friends (Skvirsky et al., 2017) were also positively associated with the anxious driving style among young drivers.

Angry and hostile style: Evidence of intergenerational transmission of the angry and hostile driving style (Gil et al., 2016; Miller and Taubman – Ben-Ari, 2010; Taubman – Ben-Ari et al., 2005), as well as the influence of the general atmosphere in the family, has also been found. Hence, studies have shown that the higher the family's level of cohesion and adaptability, and the higher the positive aspects of FCRS, the lower the young driver's tendency to endorse this style (Gil et al., 2016; Taubman – Ben-Ari and Katz-Ben-Ami, 2013). The angry and hostile style has also been related to a lower level of communication and feedback in the family and a lack of its commitment to safe driving (Skvirsky et al., 2017; Taubman – Ben-Ari, 2016). For this style as well, positive correlations emerged between siblings (Taubman – Ben-Ari, 2018), and a more conflictual relationship between them was associated with higher angry driving among young drivers.

The angry and hostile driving style was also related to higher grandparents' reports of tension when driving their grandchildren, and less positive attitudes to child restraints (Taubman – Ben-Ari, Findler et al., 2016). In addition, among young drivers, this style was positively associated with peer pressure (Taubman – Ben-Ari and Katz-Ben-Ami, 2012) and with a lack of commitment to safe driving by peers (Skvirsky et al., 2017).

Careful and patient style: Intergenerational transmission of the careful driving style has been observed as well (Gil et al., 2016; Miller and Taubman – Ben-Ari, 2010; Taubman – Ben-Ari et al., 2005). This is an important indication of the positive effect parents can have on their children's driving habits. Similarly, a positive atmosphere in the family, that is, a high level of cohesion and adaptability, has been related to a higher preference for the careful and patient style among young drivers (Gil et al., 2016; Taubman – Ben-Ari and Katz-Ben-Ami, 2013). Furthermore, the positive aspects of FCRS have been associated with the careful driving style, along with a negative correlation with noncommitment to safety (Taubman – Ben-Ari and Katz-Ben-Ami, 2012, 2013).

As with the other driving styles, the careful and patient style is shared among siblings (Taubman – Ben-Ari, 2018), and was found to correlate with a warm and close relationship between them. Thus, a positive relationship between siblings can be considered a protective element, similar to that of parents who drive in this style.

In addition, endorsement of the careful driving style among grandparents was associated with grandparents' positive attitude to child restraints (Taubman – Ben-Ari, Findler et al., 2016). This style was correlated among young drivers with lower peer pressure (Skvirsky et al., 2017; Taubman – Ben-Ari and Katz-Ben-Ami, 2012). Another interesting finding is reported by a study using the

MDSI in New Zealand: with a passenger in the car, drivers were less reckless, less angry, and more patient, all of which are features conducive to safer driving (Charlton and Starkey, 2020).

Appendix 1: Examples of Studies Using the MDSI

<i>Authors and year of publication</i>	<i>Country</i>	<i>Language</i>	<i>Version, Number of items and driving styles</i>
Taubman – Ben-Ari et al. (2004)	Israel	Hebrew	Original version; 44 items; eight driving styles: dissociative, anxious, risky, angry, high velocity, distress reduction, patient, careful
Taubman – Ben-Ari et al. (2005).	Israel	Hebrew	Original version; 44 items; four driving styles: reckless and careless, anxious, angry and hostile, patient and careful
Holland et al. (2010)	UK	English	Original version; 44 items; eight driving styles: dissociative, anxious, risky, angry, high velocity, distress reduction, patient, careful
Rossi et al. (2012)	Italy	Italian	Original version; 44 items; four driving styles: reckless and careless, anxious, angry and hostile, patient and careful
Freuli et al. (2020)	Italy: validation study	Italian	Original version; 40 items; eight driving styles: dissociative, anxious, risky, angry, high velocity, distress reduction, patient, careful
Poó et al. (2013)	Argentina: validation of MDSI-S	Spanish	Version translated into the Spanish spoken in Argentina and adapted to the Argentinian sociocultural driving context (MDSI-S); 40 items, of which 17 are Argentinian additions; six driving styles: risky, dissociative, angry, careful, anxious, distress reduction
Faílde-Garrido et al. (2016)	Spain	Spanish	MDSI-S; 40 items; six driving styles: risky, dissociative, angry, careful, anxious, distress reduction
Useche et al. (2019a)	Colombia	Spanish	MDSI-S; 40 items; six driving styles: risky, dissociative, angry, careful, anxious, distress reduction
Useche et al. (2019b)	Colombia: validation of the MDSI for professional drivers	Spanish	Original version adapted and validated for professional drivers; 33 items; four driving styles: reckless and careless, anxious, angry and hostile, patient and careful
Padilla et al. (2020)	Spain: validation of MDSI-Spain	Spanish	Version adapted to the Spanish spoken in Spain (MDSI-Spain); 34 items, 23 of which were taken from the Argentinian version; six driving styles: reckless, anxious, careful, angry, dissociative, distress reduction
Hooft van Huysduynen et al. (2015)	Netherlands and Belgium: validation study	English and Dutch	Adapted version of original inventory; 24 items; five driving styles: careful (including items from the risky style), angry, anxious, dissociative, distress reduction
Holman and Havârneanu (2015)	Romania: validation of MDSI-RO	Romanian	Version translated into Romanian and adapted for Romanian socio-cultural driving context (MDSI-RO); 41 items (including 5 Romanian additions tapping the Romanian-specific norms of violation of rules as an irrational driving style), and 36 items corresponding to items in previous versions found to be culturally equivalent in the Romanian context; seven driving styles: anxious, patient and careful, risky, angry, distress reduction, dissociative, irrational
Karjanto et al. (2017)	Malaysia		Adapted version used in the Netherlands; 24 items; four driving styles: careful, risky, anxious-dissociative, angry
Skvirsky et al. (2017)	Australia	English	Original version; 44 items; four driving styles: reckless and careless, anxious, angry and hostile, patient and careful
Bellem et al. (2018)	Germany	German	Original version; 44 items; eight driving styles: dissociative, anxious, risky, angry, high velocity, distress reduction, patient, careful
Wang et al. (2018).	China: validation of MDSI-C	Mandarin Chinese	Brief version translated into Mandarin Chinese (MDSI-C), developed from original MDSI, MDSI-S, and MDSI-RO; 12 items; four driving styles: risky, angry-high velocity, careful, anxious
Sun and Chang (2019)	China: validation of MDSI-C	Mandarin Chinese	Version translated into Mandarin Chinese and adapted for China; 32 items, of which 30 are from the original inventory, and 2 items are added to the risky driving style; six driving styles: risky, anxious, angry, distress reduction, careful, dissociative
Totkova and Racheva (2019)	Bulgaria: validation of MDSI-BG	Bulgarian	Version translated into Bulgarian and adapted for the sociocultural driving context of Bulgaria (MDSI-BG); 57 items, of which 44 items are from the original inventory, 12 from the Argentinian version, and 7 from the Romanian version; eight driving styles: risky, anxious, high velocity, dissociative, patient and careful, angry, distress reduction, irrational

References

- Antonovsky, A., 1979. Health, Stress, and Coping. Jossey Bass.
- Bellem, H., Thiel, B., Schrauf, M., Krems, J.F., 2018. Comfort in automated driving: An analysis of preferences for different automated driving styles and their dependence on personality traits. *Transport. Res. Part F: Traffic Psychol. Behav.* 55, 90–100, doi:10.1016/j.trf.2018.02.036.
- Bowen, M., 1978. Family Therapy in Clinical Practice. Jason Aronson.
- Charlton, S.G., Starkey, N.J., 2020. Co-driving: passenger actions and distractions. *Accid. Anal. Prevent.* 144, 105624, doi:10.1016/j.aap.2020.105624.
- Dula, C.S., Ballard, M.E., 2003. Development and evaluation of a measure of dangerous, aggressive, negative emotional, and risky driving. *J. Appl. Soc. Psychol.* 33 (2), 263–282, doi:10.1111/j.1559-1816.2003.tb01896.x.
- Elander, J., West, R., French, D., 1993. Behavioral correlates of individual differences in road-traffic crash risk: An examination of methods and findings. *Psychol. Bull.* 113 (2), 279–294, doi:10.1037/0033-2909.113.2.279.
- Faílde-Garrido, J.M., García-Rodríguez, M.A., Rodríguez-Castro, Y., González-Fernández, A., Lameiras Fernández, M., Carrera Fernández, M.V., 2016. Psychosocial determinants of road traffic offences in a sample of Spanish male prison inmates. *Transport. Res. Part F: Traffic Psychol. Behav.* 37, 97–106, doi:10.1016/j.trf.2015.12.004.
- Farah, H., Bekhor, S., Polus, A., Toledo, T., 2009. A passing gap acceptance model for two-lane rural highways. *Transportmetrica* 5 (3), 159–172, doi:10.1080/18128600902721899.
- Farah, H., Polus, A., Bekhor, S., Toledo, T., 2007. Study of passing gap acceptance behavior using a driving simulator. *Adv. Transport. Stud.* 23–31.
- Forsyth, D.R., 1992. Judging the morality of business practices: the influence of personal moral philosophies. *J. Bus. Ethics* 11 (5–6), 461–470, doi:10.1007/BF00870557.
- French, D.J., West, R.J., Elander, J., Wilding, J.M., 1993. Decision-making style, driving style, and self-reported involvement in road traffic accidents. *Ergonomics* 36 (6), 627–644, doi:10.1080/00140139308967925.
- Freuli, F., De Cet, G., Gastaldi, M., Orsini, F., Tagliabue, M., Rossi, R., Vidotto, G., 2020. Cross-cultural perspective of driving style in young adults: Psychometric evaluation through the analysis of the Multidimensional Driving Style Inventory. *Transport. Res. Part F: Traffic Psychol. Behav.* 73, 425–432, doi:10.1016/j.trf.2020.07.010.
- Furnham, A., Saïpe, J., 1993. Personality correlates of convicted drivers. *Personal. Individual Diff.* 14 (2), 329–336, doi:10.1016/0191-8869(93)90131-L.
- Gil, S., Taubman – Ben-Ari, O., Toledo, T., 2016. A multidimensional intergenerational model of young males' driving styles. *Accid. Anal. Prevent.* 97, 141–145, doi:10.1016/j.aap.2016.09.004.
- Gratz, K.L., Roemer, L., 2004. Multidimensional assessment of emotion regulation and dysregulation: development, factor structure, and initial validation of the difficulties in emotion regulation scale. *J. Psychopathol. Behav. Assess.* 26 (1), 41–54, doi:10.1007/s10862-008-9102-4.
- Gulian, E., Glendon, I., Matthews, G., Davies, R., Debney, L., 1988. Exploration of driver stress using self-reported data. In: Rothengatter, T., de Bruin, R. (Eds.), *Road User Behaviour: Theory and Research*. Van Gorcum & Co B.V., Assen, The Netherlands, pp. 342–347.
- Gulian, E., Matthews, G., Glendon, A.I., Davies, D.R., Debney, L.M., 1989. Dimensions of driver stress. *Ergonomics* 32 (6), 585–602, doi:10.1080/00140138908966134.
- Gwyther, H., Holland, C., 2012. The effect of age, gender and attitudes on self-regulation in driving. *Accid. Anal. Prevent.* 45, 19–28, doi:10.1016/j.aap.2011.11.022.
- Holland, C., Geraghty, J., Shah, K., 2010. Differential moderating effect of locus of control on effect of driving experience in young male and female drivers. *Personal. Individual Diff.* 48 (7), 821–826, doi:10.1016/j.paid.2010.02.003.
- Holman, A.C., Havârneanu, C.E., 2015. The Romanian version of the multidimensional driving style inventory: Psychometric properties and cultural specificities. *Transport. Res. Part F: Traffic Psychol. Behav.* 35, 45–59, doi:10.1016/j.trf.2015.10.001.
- Holman, A.C., Popusoi, S.A., 2018. Ethical predispositions to violate or obey traffic rules and the mediating role of driving styles. *J. Psychol. Interdiscipl. Appl.* 152 (5), 257–275, doi:10.1080/00223980.2018.1447433.
- Hooft van Huysduynen, H., Terken, J., Eggen, B., 2018. The relation between self-reported driving style and driving behaviour. A simulator study. *Transport. Res. Part F: Traffic Psychol. Behav.* 56, 245–255, doi:10.1016/j.trf.2018.04.017.
- Hooft van Huysduynen, H., Terken, J., Martens, J.B., Eggen, B., 2015. Measuring driving styles: A validation of the multidimensional driving style inventory. *AutomotiveUI '15: Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, September 2015, pp. 257–264. <https://doi.org/10.1145/2799250.2799266>.
- John, O.P., Srivastava, S., 1999. The Big Five trait taxonomy: history, measurement and theoretical perspectives. In: Pervin, L., John, O.P. (Eds.), *Handbook of Personality: Theory and Research*. 2nd ed. Guilford Press, pp. 102–138.
- Karjanto, J., Md Yusof, N., Terken, J., Hassan, M.Z., Delbressine, F., Hooft van Huysduynen, H., Rauterberg, M., 2017. The identification of Malaysian driving styles using the multidimensional driving style inventory. *MATEC Web of Conferences* 90, 01004, doi:10.1051/mateconf/20179001004.
- Liang, B., Lin, Y., 2018. Using physiological and behavioral measurements in a picture-based road hazard perception experiment to classify risky and safe drivers. *Transport. Res. Part F: Traffic Psychol. Behav.* 58, 93–105, doi:10.1016/j.trf.2018.05.024.
- Miller, G., Taubman – Ben-Ari, O., 2010. Driving styles among young novice drivers – The contribution of parental driving styles and personal characteristics. *Accid. Anal. Prevent.* 42 (2), 558–570, doi:10.1016/j.aap.2009.09.024.
- Navon, M., Taubman – Ben-Ari, O., 2019. Driven by emotions: The association between emotion regulation, forgiveness, and driving styles. *Transport. Res. Part F: Traffic Psychol. Behav.* 65, 1–9, doi:10.1016/j.trf.2019.07.005.
- Navon- Eyal, M., Taubman – Ben-Ari, O., 2020. Can emotion regulation explain the association between age and driving styles? *Transportation Research Part F: Traffic Psychology and Behaviour* 74, 439–445, doi:10.1016/j.trf.2020.09.008.
- Olson, D.H., 1991. Three dimensional (3-D) circumplex model and revised scoring of FACES. *Family Process* 30 (1), 74–79, doi:10.1111/j.1545-5300.1991.00074.x.
- Padilla, J.L., Castro, C., Doncel, P., Taubman – Ben-Ari, O., 2020. Adaptation of the multidimensional driving styles inventory for Spanish drivers: convergent and predictive validity evidence for detecting safe and unsafe driving styles. *Accid. Anal. Prevent.* 136, 105413, doi:10.1016/j.aap.2019.105413.
- Padilla, J.L., Doncel, P., Gugliotta, A., Castro, C., 2018. Which drivers are at risk? Factors that determine the profile of the reoffender driver. *Accid. Anal. Prevent.* 119, 237–247, doi:10.1016/j.aap.2018.07.021.
- Poó, F.M., Ledesma, R.D., 2013. A study on the relationship between personality and driving styles. *Traffic Injury Prevent.* 14 (4), 346–352, doi:10.1080/15389588.2012.717729.
- Poó, F.M., Taubman – Ben-Ari, O., Ledesma, R.D., Díaz-Lázaro, C.M., 2013. Reliability and validity of a Spanish-language version of the multidimensional driving style inventory. *Transport. Res. Part F* 17, 75–87, doi:10.1016/j.trf.2012.10.003.
- Pugnetti, C., Elmer, S., 2020. Self-assessment of driving style and the willingness to share personal information. *J. Risk Finan. Manage.* 13 (3), 53, doi:10.3390/jrfm13030053.
- Reason, J., Manstead, A., Stradling, S., Baxter, J., Campbell, K., 1990. Errors and violations on the roads: A real distinction? *Ergonomics* 33 (10–11), 1315–1332, doi:10.1080/00140139008925335.
- Rosenberg, M., 1979. *Conceiving the self*. Basic Books.
- Rossi, R., Gastaldi, M., Gecchele, G., Meneguzzi, C., 2012. Comparative analysis of random utility models and fuzzy logic models for representing gap-acceptance behavior using data from driving simulator experiments. *Procedia - Social and Behavioral Sciences* 54, 834–844, doi:10.1016/j.sbspro.2012.09.799.
- Rotter, J.B., 1966. Generalised expectancies for internal versus external locus of control of reinforcement. *Psychol. Monographs: Gen. Appl.* 80 (1), 1–28, doi:10.1037/h0092976.
- Sagberg, F., Selpi, Piccinini, G.F.B., Engström, J., 2015. A review of research on driving styles and road safety. *Human Factors* 57 (7), 1248–1275, doi:10.1177/0018720815591313.
- Skvirsky, V., Taubman – Ben-Ari, O., Greenbury, T.J., Prato, C.G., 2017. Contributors to young drivers' driving styles – a comparison between Israel and Queensland. *Accid. Anal. Prevent.* 109, 47–54, doi:10.1016/j.aap.2017.08.031.
- Sun, L., Chang, R., 2019. Reliability and validity of the Multidimensional Driving Style Inventory in Chinese drivers. *Traffic Injury Prevent.* 20 (2), 152–157, doi:10.1080/15389588.2018.1542140.

- Taubman – Ben-Ari, O., 2006. Couple similarity for driving style. *Transport. Res. Part F: Traffic Psychol. Behav.* 9 (3), 185–193, doi:10.1016/j.trf.2005.11.001.
- Taubman – Ben-Ari, O., 2008. The Israeli driver's profile. A report prepared for the Israeli Ministry of Transportation – The National Authority for Road Safety.
- Taubman – Ben-Ari, O., 2014a. How are meaning in life and family aspects associated with teen driving behaviors? *Transport. Res. Part F: Traffic Psychol. Behav.* 24, 92–102, doi:10.1016/j.trf.2014.04.008.
- Taubman – Ben-Ari, O., 2014b. The parental factor in adolescent reckless driving: The road ahead. *Accid. Anal. Prevent.* 69, 1–4, doi:10.1016/j.aap.2014.02.011.
- Taubman – Ben-Ari, O., 2015. Parents' perceptions of the family climate for road safety. *Accid. Anal. Prevent.* 74, 157–161, doi:10.1016/j.aap.2014.10.023.
- Taubman – Ben-Ari, O., 2016. Parents' perceptions of the Family Climate for Road Safety: Associations with parents' self-efficacy and attitudes toward accompanied driving, and teens' driving styles. *Transport. Res. Part F: Traffic Psychol. Behav.* 40, 14–22, doi:10.1016/j.trf.2016.04.006.
- Taubman – Ben-Ari, O., 2018. What potential role do siblings play in young drivers' driving styles? *Transport. Res. Part F: Traffic Psychol. Behav.* 58, 19–24, doi:10.1016/j.trf.2018.05.028.
- Taubman – Ben-Ari, O., Eherenfreund-Hager, A., Prato, C.G., Eherenfreund-Hager et al., 2016. The value of self-report measures as indicators of driving behaviors among young drivers. *Transport. Res. Part F: Traffic Psychol. Behav.* 39, 33–42, doi:10.1016/j.trf.2016.03.005.
- Taubman – Ben-Ari, O., Findler, L., Noy, A., Porat-Zyman, G., 2016. When grandparents drive their grandchildren. *Transport. Res. Part F: Traffic Psychol. Behav.* 39, 54–64, doi:10.1016/j.trf.2016.03.004.
- Taubman – Ben-Ari, O., Katz-Ben-Ami, L., 2012. The contribution of family climate for road safety and social environment to the reported driving behavior of young drivers. *Accid. Anal. Prevent.* 47, 1–10, doi:10.1016/j.aap.2012.01.003.
- Taubman – Ben-Ari, O., Katz-Ben-Ami, L., 2013. Family climate for road safety: a new concept and measure. *Accid. Anal. Prevent.* 54, 1–14, doi:10.1016/j.aap.2013.02.001.
- Taubman – Ben-Ari, O., Mikulincer, M., Gillath, O., 2004. The multidimensional driving style inventory – scale construct and validation. *Accid. Anal. Prevent.* 36 (3), 323–332, doi:10.1016/S0001-4575(03)00010-1.
- Taubman – Ben-Ari, O., Mikulincer, M., Gillath, O., 2005. From parents to children—similarity in parents and offspring driving styles. *Transport. Res. Part F: Traffic Psychol. Behav.* 8 (1), 19–29, doi:10.1016/j.trf.2004.11.001.
- Taubman – Ben-Ari, O., Skvirsky, V., 2016. The multidimensional driving style inventory a decade later: Review of the literature and re-evaluation of the scale. *Accid. Anal. Prevent.* 93, 179–188, doi:10.1016/j.aap.2016.04.038.
- Taubman – Ben-Ari, O., Yehiel, D., 2012. Driving styles and their associations with personality and motivation. *Accid. Anal. Prevent.* 45, 416–422, doi:10.1016/j.aap.2011.08.007.
- Totkova, Z., 2020. Interconnection between driving style, traffic locus of control, and impulsivity in Bulgarian drivers. *Behav. Sci.* 10 (2), 58., doi:10.3390/bs10020058.
- Totkova, Z., Racheva, R., 2019. The Bulgarian version of the Multidimensional Driving Style Inventory: Psychometric properties. *Behav. Sci.* 9 (12), 145., doi:10.3390/bs9120145.
- Trógolo, M.A., Melchior, F., Medrano, L.A., 2014. The role of difficulties in emotion regulation on driving behavior. *J. Behav. Health Social Issues* 6 (1), 107–117, doi:10.5460/jbhsi.v6.1.47607.
- Trógolo, M.A., Tosi, J.D., Poó, F.M., Ledesma, R.D., Medrano, L.A., Domínguez-Lara, S., 2020. Factor structure and measurement invariance of the multidimensional driving style inventory across gender and age: An ESEM approach. *Transport. Res. Part F: Traffic Psychol. Behav.* 71, 23–30, doi:10.1016/j.trf.2020.04.001.
- Useche, S.A., Cendales, B., Alonso, F., Montoro, L., Pastor, J.C., 2019a. Trait driving anger and driving styles among Colombian professional drivers. *Heliyon* 5 (8), e02259, doi:10.1016/j.heliyon.2019.e02259.
- Useche, S.A., Cendales, B., Alonso, F., Orozco-Fontalvo, M., 2020. A matter of style? Testing the moderating effect of driving styles on the relationship between job strain and work-related crashes of professional drivers. *Transport. Res. Part F* 72, 307–317, doi:10.1016/j.trf.2020.05.015.
- Useche, S.A., Cendales, B., Alonso, F., Pastor, J.C., Montoro, L., 2019b. Validation of the Multidimensional Driving Style Inventory (MDSI) in professional drivers: How does it work in transportation workers? *Transport. Res. Part F: Traffic Psychol. Behav.* 67, 155–163, doi:10.1016/j.trf.2019.10.012.
- Varotto, S.F., Farah, H., Toledo, T., van Arem, B., Hoogendoorn, S.P., 2018. Modelling decisions of control transitions and target speed regulations in full-range Adaptive Cruise Control based on Risk Allostasis Theory. *Transport. Res. Part B: Methodol.* 117, Part A, 318–341, doi:10.1016/j.trb.2018.09.007.
- Wang, Y., Qu, W., Ge, Y., Sun, X., Zhang, K., 2018. Effect of personality traits on driving style: Psychometric adaption of the multidimensional driving style inventory in a Chinese sample. *PLoS ONE* 13 (9), e0202126, doi:10.1371/journal.pone.0202126.
- West, R., Hall, J., 1997. The role of personality and attitudes in traffic accident risk. *Appl. Psychol. Int. Rev.* 46 (3), 253–264, doi:10.1111/j.1464-0597.1997.tb01229.x.
- Wiesenthal, D.L., Hennessy, D., Gibson, P.M., 2000. The Driving Vengeance Questionnaire (DVQ): The development of a scale to measure deviant drivers' attitudes. *Violence Victims* 15 (2), 115–136, doi:10.1891/0886-6708.15.2.115.

Risk Perception in Transport: A Review of the State of the Art

Trond Nordfjærn*, **An-Magritt Kummeneje***, **Mohsen F. Zavareh†**, **Milad Mehdizadeh‡**, **Torbjørn Rundmo***, *Norwegian University of Science and Technology (NTNU), Department of Psychology, Trondheim, Norway; †Department of Civil Engineering, Kharazmi University, Faculty of Engineering, Tehran, Iran; ‡Iran University of Science and Technology (IUST), School of Civil Engineering, Department of Transportation, Narmak, Tehran, Iran

© 2021 Elsevier Ltd. All rights reserved.

Introduction	74
What is Transport Risk Perception?	74
Definition and Measurement of Transport Risk Perception	74
Associations Between Risk Perception and Driving Behavior	75
Relations Between Risk Perception and Demand for Risk Mitigation in Transport	76
Risk Perception and Sustainable Transport Mode Use in Urban Areas	77
General Mode Use in Urban Areas	77
Active Mode Use in Urban Areas	77
Risk Perception and Mode Use on School Trips	78
Summary and Conclusions	79
Acknowledgment	79
References	79
Further Reading	80

Introduction

The aim of this chapter is to give an overview of the state of the art concerning the role of transport risk perception for driving behavior, demand for risk mitigation, and use of sustainable urban transport modes. The review summarizes theory and peer-reviewed empirical findings regarding risk perception and transport. Transport is delimited to surface modes typically available in urban areas, such as car, tram, metro, walking, and cycling. Risk perception is also important in sectors such as aviation and shipping, but this is beyond the scope of the current review.

As shown in **Fig. 1**, the review is divided into four sections. The first section discusses the risk perception concept with a focus on operational definition and theoretical foundations. The second section focuses on risk perception in relation to an important variable specifically in the road traffic system, namely driving behavior. The third section broadens the scope to the entire surface transport system, and focuses on risk perception in relation to demand for transport risk mitigation exerted by the public upon the transport authorities. The last section maintains this scope, but with a particular focus on transport mode use in urban areas. This section discusses risk perception in relation to general and active sustainable mode use in urban areas and, more specifically, mode choices on school trips. The latter concerns risk perception among parents when they choose transport modes for their children on school travels. This focus is viable as children represent the next generation of mode users and transport habits are usually developed in young age (see, e.g., [Verplanken et al., 1997](#); [Verplanken and Orbell, 2003](#)).

What is Transport Risk Perception?

In research, risk perception has been conceived as an endogenous as well as exogenous variable. The first-mentioned type of research concerns which factors that may influence on the “level” of perceived risk, and the last-mentioned focuses on whether perceived risk is associated with lay peoples’ demand for risk mitigation, choice of travel mode, and risk-taking behavior in traffic. In addition, there has been attention on the concept and how it should be measured. The current review will mainly focus on risk perception as exogenous variable as the concept primarily is interesting because it predicts transport behavior.

Definition and Measurement of Transport Risk Perception

The psychometric approach previously dominated risk perception research. It focused on the unique and subjective qualities of perceived risk ([Slovic, 1992, 2000](#)). [Fischhoff et al. \(1978\)](#) showed that risk sources had a pattern of qualities and that these qualities were related to perceived risk. Nine general properties of the risk source were found to be important; volunteerism of risk, immediacy of effect, knowledge about the risk by the person who was exposed to the potentially-hazardous risk source, knowledge about the risk in science, control over the risk, newness, whether the risk was chronic/catastrophic, was common/dread, and how severe the consequences were judged to be. The degree to which these factors were related to potentially hazardous activities or technologies

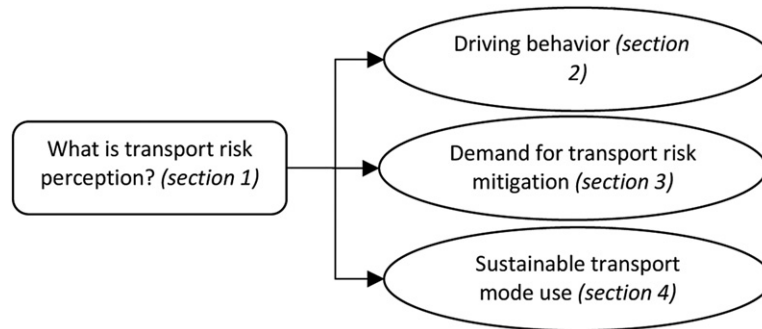


Figure 1 Chapter outline.

was believed to determine how the risk source was perceived (Fischhoff et al., 2000). Principal component analysis (PCA) of the nine properties across a large number of risk sources showed that dread and severity of consequences were the most important dimensions. The last-mentioned dimension concerned the evaluation of the probability that a mishap or illness would be fatal. Accordingly, the last-mentioned property should be considered as a measure of the perceived probability of a fatal incident occurring, that is, judgment of severity of consequences.

Risk perception is often considered to be a pure cognitive construct. It can be argued that dread reflects emotion (Sjøberg, 1999). Emotions, such as worry and concern, should therefore be seen as an effect of perceived risk rather than a part of the construct. Precognitive processes, for example, anticipated worry, may in accordance with the risk-as-feeling approach (Loewenstein et al., 2001) influence perception of risk. “Risk-as-feelings” may be distinguished from “risk-as-analysis” (Slovic et al., 2001). Risk-as-analysis is the process where cognitive reasoning, reflection and logic are used to evaluate a risk source. This process is identical to perceived risk. The risk-as-feelings framework (Loewenstein et al., 2001) refers to intuitive and instinctive reactions to a risk source. Consequently, emotions such as worry and concern are relevant for risk perception, however, primarily because they can influence and be influenced by cognitive evaluations, that is, of perceived risk. Perception of risk as a construct applies only to cognitively mediated processes. In accordance with the studies mentioned above, risk perception is most frequently understood to be the subjective assessment of the probability of an accident or any other event with negative consequences and the anticipated judgement of severity of consequences if such an event should take place (see, e.g., Bubek et al., 2012; Champ and Brenkert-Smith, 2016; Sjøberg, 1998, 2004). Thereby, we define transport risk perception by subjective evaluations of probability and potential severity of consequences of accidents or security incidents in transport.

The focus on the measurement of subjective assessment of probability and judgement of severity of consequences is in accordance with the two main factors in Quantitative Risk Assessments (QRAs). In addition, exposure to the risk source may be taken into QRA estimates (Rausand and Utne, 2009). QRAs are based on a reflective model of risk. In such a model, the parameters included into the measurement need not by necessity belong to an underlying construct to be part of the estimate of the “objective” risk level. Thus, QRAs may consist of parameters that are not necessarily related to one another. Contrary to what is the case in QRAs, risk perception is in the psychometric paradigm believed to be a formative construct. This makes PCA and confirmatory factor analysis (CFA) relevant for examining the dimensional structure of perceived risk. Rundmo and Nordfjrn (2017) carried out structural equation modeling (SEM) aimed at testing the dimensional structure of perceived risk in transport. A formative theoretical model showed a bad fit to the data. Entering worry and concern into such a formative model in addition to probability assessment and judgment of consequences, did not improve the fit of the model.

There is no empirical evidence in transport safety research showing that perceived risk is a formative construct. A theoretical model hypothesizing assessment of probability and judgement of severity of consequences to be independent factors was shown to be satisfactory fitted to the data. Subjective probability assessment and judgments of severity of consequences may be independent factors not belonging to any overall psychological concept forming the judgement, that is, where the judgement “reflects” the overall construct “risk perception.” Rundmo and Nordfjrn (2017) showed, however, that probability and consequence evaluations were significantly associated in lay people’s subjective judgements of transport risk. When probability assessment was added as an exogenous variable, it explained a moderate, however significant, percentage of the variance in judgement of severity of consequences. This was the case in judgement of public as well as private travel modes. Consequently, it may be proposed that the subjective probability assessment may be formative for judgements of severity of consequences. The higher the probability was assessed to be, the larger was also the consequences believed to be if an event with negative consequences should take place, and vice versa.

Associations Between Risk Perception and Driving Behavior

One of the main safety challenges in transport is present in the road traffic sector. A substantial amount of accidents in this sector is caused by risky driving behavior, such as speeding and rule violations. Aligned with the definition outlined in the previous section, traffic risk perception is the subjective interpretation of probability and potential severity of consequences regarding road traffic

accidents. Traffic risk perception is mainly of interest because it may be an antecedent of driving behavior (Machin and Sankey, 2008), such that high perceived risk could be associated with a lower probability of conducting a risk activity and low perceived risk could be associated with higher risk-taking propensity. This assumption aligns with theories that incorporate risk perception to predict risk or health behaviors, such as the Health Belief Model (Janz and Becker, 1984) and Protection Motivation Theory (Rogers, 1983). The state of the art reflects discrepant findings regarding the association between risk perception and driving behavior. For instance, some studies have shown that high traffic risk perception may be associated with less speeding and other rule violations (e.g., Machin and Sankey, 2008; Steinbakk et al., 2019; Ulleberg and Rundmo, 2003) as well as cell phone use while driving (Oviedo-Trespalacios et al., 2017). Common for these studies is that they incorporated affective components of risk into the operationalization of traffic risk perception, or did not accommodate both probability and consequence assessments into the measurement instrument.

There are several studies that call into question the association between risk perception and driving behavior. Rundmo and Iversen (2004) found that perceptions of traffic risk related weakly to self-reported driving behavior. These findings have been replicated in a series of cross-cultural studies conducted in several countries spanning across Europe, Asia and Sub-Saharan Africa, where risk perception either was found to have a modest (e.g., Simsekoglu et al., 2013) or weak association with driving behavior (e.g., Lund and Rundmo, 2009; Nordfjrn et al., 2011). Although Ulleberg and Rundmo (2003) detected a significant correlation between high risk perception and less risk-taking driving behavior, the role of risk perception vanished when attitudes and personality traits were accommodated with traffic risk perception in a structural equation model. Further, Ivers et al. (2009) found no substantial associations between driving behavior and risk perception. The association between risk perception and driving behavior has also been indirectly questioned in cross-country studies. Nordfjrn and Rundmo (2009) found that Norwegians (where driving behavior tend to be safe) reported substantially lower traffic risk perception than individuals from Ghana (where driving behavior tend to be rather unsafe). Given that high risk perception is an antecedent of safe driving behavior, one may have expected the opposite tendency to occur. Similar tendencies were also noted in an earlier study (Hayakawa et al., 2000) where Japanese respondents reported higher traffic accident probabilities than individuals from the United States, despite that road traffic fatality rates are higher in the latter country. A potential explanation is that people habituate and desensitize to the specific hazards and risks present in their own specific traffic environments. This assumption has received support in recent studies using the hazard perception paradigm (Bazilinskyy et al., 2020; Ventsislavova et al., 2019).

Research also indicates that traffic risk perception is associated with demographic characteristics. Empirical evidence for these relations was found in a study where drivers from various age groups rated the accident risk of 100 slide-projected traffic situations (Tränkle et al., 1990; see also Sivak et al., 1989). The results suggested that young males reported less risk than older male drivers, whereas the age effect was not present among females. Research also supports that males tend to underestimate their personal risk when driving and overestimate their driving competence compared with other demographic groups (Glendon et al., 1996). The gender effect in risk perception also seems to be present in other transport domains, such as bicycling (Prati et al., 2019). Deery (1999) showed that young drivers tend to underestimate accident risk across different hazard situations, and that they have higher risk tolerance than more experienced drivers. The research points to the fact that risk perception is subject to demographic variation, and may be an explanatory factor for demographic differences in traffic safety.

Relations Between Risk Perception and Demand for Risk Mitigation in Transport

The previous section discussed how traffic risk perception relates to risky behavior specifically within the road traffic sector. The current section expands the focus to the general surface transport system, and to transport risk mitigating demands exerted upon policymakers and experts by the general public. The association between risk perception and demand for risk mitigation is mainly interesting because it may affect mode use. If people have high risk perception and demand for risk mitigation regarding a certain sector of the transport system (e.g., public transport), and the demand is not addressed by appropriate countermeasures, people may change to other modes (e.g., car) that they perceive to be safer (Phun et al., 2018). Demand for risk mitigation may be more relevant for those who mainly use the public transport sector, as Nordfjrn and Rundmo (2018) reported that such demands were more strongly related to intentions of using public transport than demands for risk mitigation in the road traffic sector.

One of the core challenges regarding demand for transport risk mitigation is that the general public and experts tend to diverge in their perceptions of risk. Studies have reported that experts tend to focus on the probability component of risk perception, whereas the general public stresses the consequence component (e.g., Rundmo and Moen, 2006). The same study suggested that this could be due to that the consequence component is more strongly associated with affect, such as worry, whereas the probability component is more strongly related to cognition and a more analytical approach to risk. Overall, Rundmo and Moen (2006) demonstrated that probability assessments were of minor importance in demand for risk mitigation and that the severity of consequences primarily was significant because they were associated with anticipatory worry, which was a significant predictor variable of demand for risk mitigation. One could expect that experts tend to rely more on cognitive evaluations than the general public which may rely more strongly on affective heuristics to a risk source. In accordance with this assumption, Sjøberg (1999) reported that the expected severity of consequences of a risk source was associated with demand for risk mitigation, whereas probability estimates and such demands were weakly associated in a general public. Contradictory to previous research, Rundmo and Nordfjrn (2013) concluded that risk perception did not predict demand for risk mitigation. Risk perception was, however, found to predict worry which in turn was a significant predictor of demand for risk mitigation (see also Baron et al., 2000). Further, the researchers reported that worry has

an indirect effect on demand for risk mitigation through priority of safety. Similar relations have been demonstrated more recently as [Rundmo and Nordfjrn \(2017\)](#) showed that probability assessments of accidents were not related to demand for risk mitigation in transport, while judgement of severity of consequences directly as well as indirectly was associated with anticipatory worry. In addition to judgement of severity of consequences, worry was a significant predictor of demand for risk mitigation. It seems like that a high demand for risk mitigation is likely when the risk is perceived as high, and the tolerance is low. The higher a risk is perceived and the more worry it induces, the less risk tolerance is expected.

The risk as feelings framework (see section "What is Transport Risk Perception") has been utilized to investigate demand for risk mitigation in transport. [Moen \(2008\)](#) found that both the consequence and probability components related to perceived risk, while the consequence component was a somewhat stronger predictor of feelings than probability estimates. Of note, feelings had a stronger relation to priority of safety than risk perception. Priority of safety was in turn a strong predictor of demand for risk mitigation. This research suggests that risk perception is not all that important for demand for risk mitigation, at least not directly, but risk perception may still be relevant as this construct has mediated effects on demand for risk mitigation by its relation to affect. Several previous studies tended to take for granted that the larger the risk is, the more people would like it to be mitigated. [Rottenstreich and Hsee \(2001\)](#) refuted this assumption, and found that when an outcome has emotional implications, people tended to be insensitive to variations in probability. As such the relative importance of probability depends on the outcome, and the role of probability assessment may be negligible when the outcomes are vivid compared with pallid outcomes.

Risk Perception and Sustainable Transport Mode Use in Urban Areas

General Mode Use in Urban Areas

In addition to being a predictor of driving behavior and safety cognitions such as demand for risk mitigation, risk perception may also be relevant for general mode use in the urban public. People in urban areas usually have several surface transport modes available, such as metro, tram, cars, bicycling and walking. Risk perception may influence the decision-making process between these alternatives. In addition to evaluations of probability and severity of consequences of transport accidents and injuries, transport risk perception includes subjective evaluations of probability and severity of consequences related to security matters. Security is especially viable in public transport and covers the risk of being subject to, for example, physical assaults, theft, and terrorism ([Nordfjrn and Rundmo, 2018](#); [Rundmo and Nordfjrn, 2019](#)). Security is not equally important to traffic safety by use of private motorized travel modes, such as car, motorbike, moped and scooter. Meanwhile, both safety and security are relevant for behavior among cyclists and pedestrians ([Kummeneje and Rundmo, 2019, 2020](#); [Kummeneje et al., 2019](#)).

[Rundmo et al. \(2011\)](#) investigated a number of risk judgements in relation to urban mode use. Opposing previous studies which showed that the consequence component is more important for risk decisions, the findings showed that the probability component was more important for mode use than the consequence component. Those who used public transport more often perceived the probability of an accident to be less compared to the two groups of respondents who used private modes more often than public modes. Accident probabilities were also rated to be overall higher for private modes compared to public modes. [Roche-Cerasi et al. \(2013\)](#) reported that respondents who mainly used public transport in urban travel assessed higher probability of experiencing criminality, such as violence, in public transport than those who mainly used private modes, such as car. Those who used private modes were generally more worried about accidents, whereas those who used public transport were more focused on security issues such as criminality and terrorism. These findings align with other studies (e.g., [Backer-Grondahl et al., 2009](#)) which showed that people worry more about accidents than unpleasant incidents on private modes, while they worry more about unpleasant incidents than accidents on public modes. However, challenging the findings of previous studies the results reflected rather negligible associations between mode use and risk perception.

Some studies have also revealed indications of avoidance behavior due to risk perception. For instance, [Nordfjrn et al. \(2014\)](#) found that high risk perception of terrorism and major accidents as well as risk perception related to theft and violence predicted more car use among Norwegian commuters. These findings are opposed by [Elias and Shiftan \(2012\)](#) who concluded that risk perception had a weak relation with the intention to use public transport and that people focus on other priorities, such as advantages of using car, rather than health risks when choosing to use a car. [Nordfjrn and Rundmo \(2018\)](#) found that individuals who had experienced a negative security incident on public transport in recent years had elevated risk perception and worry across public and private transport modes. The findings further reflected that high risk perception in public transport predicted a lower intention of using public transport modes, whereas high risk perception in private motorized transport predicted a stronger intention to use public modes. For transport risks where consequences are perceived to be low, for example, for cyclists and for being a pedestrian during daytime and in spring and summer, probability assessments have been found to have a direct association with how often people cycled and walked during winter and summer ([Kummeneje and Rundmo, 2019, 2020](#)).

Active Mode Use in Urban Areas

Active mode use, such as walking and cycling, is given high priority in European transport policy. The use of active modes is seen as a pro-environmental as well as a health promoting behavior. Several previous studies have examined risk perception and worry in relation to active travel. For instance, [Moen and Rundmo \(2006\)](#) and [Oltedal and Rundmo \(2007\)](#) found that the respondents were less worried about being in an accident when cycling and walking compared with using other private travel modes. The respondents

also perceived the probability of being in an accident as the lowest as a pedestrian and cyclist compared with other private transport modes (car, motorcycle, scooter). Another interesting finding was that the respondents reported that the consequences of being in an accident were greater when walking than when cycling, but still lower than when using motorized transport modes.

According to [Chaurand and Delhomme \(2013\)](#), the study of risk perception related to cycling has received relatively little attention. The same is true for pedestrians, and few studies to date have focused on cyclists' and pedestrians' perceived risk and worry. Much of the research in this field has been conducted to aid engineers and city planners in designing and improving the road infrastructure. In most of these studies the respondents have been asked to rate their overall risk perception of a route described by video-clips, simulations, pictures, or surveys. In these studies, the perceptions of risk and feelings related to the road infrastructure or traffic environment have been in focus (e.g., [Lawson et al., 2013](#); [Llorca et al., 2017](#); [Manton et al., 2016](#); [Moller and Hels, 2008](#); [Rankavat and Tiwari, 2016](#)). The findings are difficult to compare across these studies due to inconsistent operational definitions and measures of risk perception.

In recent studies, risk perception of active modes has been operationalized by measuring both the respondents' assessment of probability and judgment of the severity of consequences of different types of hazards when cycling or walking ([Kummeneje and Rundmo, 2018, 2019, 2020](#); [Kummeneje et al., 2019](#)). [Kummeneje and Rundmo \(2018\)](#) and [Kummeneje et al. \(2019\)](#) found seasonal differences in cyclists perceived risk and worry about being in an accident. The cyclists perceived the risk to be higher and were more worried when cycling in winter conditions compared with summer conditions. Risk perception and worry were further found to be important predictors of both the decision to cycle and the frequency of cycling during wintertime. In a study of pedestrians' risk perception and worry, [Kummeneje and Rundmo \(2019\)](#) found a strong association between risk perception and worry among pedestrians. Further, they found that worry was moderately associated with walking frequency during night-time, and how often individuals walked alone outdoors during night-time. These associations were stronger for people without access to a private car. No associations were found between worry and walking frequency during the daytime. In a different study, [Kummeneje and Rundmo \(2020\)](#) found that risk perception and worry not only were important for travel behavior, but also for risk-taking behavior and conflicts with other road users when cycling.

Risk Perception and Mode Use on School Trips

Risk perception could also play a role when people make transport mode choices on behalf of others, such as for school children. Among school-aged children (6–17 years), it is evident that pupils younger than 12 years are less likely to accurately perceive traffic risks ([Van Goeverden and De Boer, 2013](#)). Due to this fact, parents are more likely to make decisions regarding their children's travel mode in trips to/from school ([D' Haese et al., 2011](#); [Mehdizadeh et al., 2017a](#)). In addition to contextual variables, such as distance from home to school and parental socioeconomic characteristics, perceived risk about available transport options may have a pivotal role in mode choice decisions in school trips. For instance, due to a higher perceived risk regarding the walking/cycling route to school, parents may prefer not to let their children walk/cycle to school, even when the school is located close to their residential location. Consequently, it can be concluded that different aspects of risk perception concerning various travel modes can impact on mobility styles and transport mode choices. Higher perceived risk regarding active and public transport modes may lead to more use of motorized transport. Such reasoning is in line with psychological theories, such as the Protection Motivation Theory ([Rogers and Prentice-Dunn, 1997](#)) and the Health Belief Model ([Rosenstock, 1974](#)).

Although there is much research on mode choice behavior in school trips, less attention has been devoted to the investigation of risk perception in school trips. [Mehdizadeh et al. \(2017b\)](#) analyzed the relative roles of parental risk perceptions and worry toward different transport options on pupils' mode use in the Middle East context. Both major cognitive components of risk perception, that is, the probability and consequence estimates of accident involvement, were significantly associated with using active modes instead of motorized modes. It was also concluded that worry could impact on parental mode choice decisions. The authors showed that parents who had high risk perception and more worry about walking, were more likely to choose motorized modes for their children in school trips. In a similar way, after controlling for the effects of child's gender and distance from home to school, [Fallah Zavareh et al. \(2020\)](#) found that parental worry was significantly higher in the group of non-walking pupils than the group of walking pupils. In another study [Mehdizadeh et al. \(2018\)](#) showed that parents worried the most about letting pupils walk alone to school compared to other modes of transport.

Although very few studies focused on the role of risk perception on school trips, several studies of parental worry about children as pedestrians have been conducted ([Mammen et al., 2012](#); [Peterson et al., 1990](#); [Salmon et al., 2007](#)). [Mammen et al. \(2012\)](#) investigated differences in parental worry about children's school travel. They found that parents who escorted their children to school worried more than other parents about the possibility of strangers and bullies approaching their child and about the traffic volume around the school. [Salmon et al. \(2007\)](#) reported an association between the child's use of active transport to school and parents' concern that their child may be injured in a road accident. [Peterson et al. \(1990\)](#) investigated parents' feelings of worry about different types of injury, including their children being injured by a motor vehicle when walking. Overall, the results showed that parents reported low feelings of worry about injuries.

Parental risk perception can also influence willingness to accept interventions designed to change travel behavior on school trips. In one of the few studies that focused on this, [Fallah Zavareh et al. \(2020\)](#) reported that more parental worry was associated with a weaker intention to let non-walking pupils walk in the presence of a proposed IT solution which informed parents about children's entrance to and exit from the school. Parental worry had no significant association for the group of walking pupils. Predictors of parental risk perception in relation to children's commuting to school have been subject of a few previous studies. [Nevelsteen et al.](#)

(2012) investigated the relationship between perceived traffic safety of parents and found that parental safety perception can be explained by the age and gender of the child as well as traffic infrastructure near home or destinations frequently visited. Lam (2001) reported that the age of the child, gender of the parent, employment of the parent, living environment, and previous injury experience were significantly associated with parental risk perception of children's road safety.

In addition to socio-demographic and situational predictors, risk perception may be influenced by individual competence (Slovic, 2007). For instance, Rudner (2012) found that the interaction of parental conceptions regarding the environment and their evaluations of the children's competencies (risk control, skill in dealing with the risk and their knowledge about potential hazards) influence how parents perceive risk, which in turn may affect parental decisions for children's independent mobility. Studying a group of non-walking pupils, Fallah Zavareh et al. (2020) also showed that positive attitudes toward safety responsibilities in transport, and trust in experts' knowledge were significantly associated with parental worry about children's risk of being involved in a road crash when walking to school.

Summary and Conclusions

The state of the art suggests that risk perception is relevant for driving behavior, demand for risk mitigation and urban transport mode use. However, several studies did not operationalize risk perception in line with risk theory, but incorporated affective components, such as worry, into the risk perception construct. It might be valid to include operationalizations of affect as well as other risk judgements (e.g., risk tolerance, priorities of safety, risk aversion etc.) into the prediction models when examining the role of risk perception for transport behavior. Future studies could benefit by a more restricted and clear-cut operationalization of traffic risk perception as perceived probability and consequences of accidents. Affect and other related risk judgements could be incorporated as separate constructs predicting the outcome variables. This would facilitate comparison of findings across studies as well as risk perception research which is more aligned with psychological risk theory.

Acknowledgment

The review did not receive any specific support and/or funding.

References

- Backer-Grondahl, A., Fyhri, A., Ulleberg, P., Amundsen, A.H., 2009. Accidents and unpleasant incidents: worry in transport and prediction of travel behavior. *Risk Anal.* 29, 1217–1226.
- Baron, J., Hershey, J.C., Kunreuther, H., 2000. Determinants of priority for risk reduction: the role of worry. *Risk Anal.* 20, 413–428.
- Bazilinskyy, P., Eisma, Y.B., Dodou, D., de Winter, J.C.F., in press. Risk perception: a study using dashcam videos and participants from different world regions. *Traffic. Inj. Prev.* 21, 347–353.
- Bubek, P., Botzen, W.J., Aerts, J.C., 2012. A review of risk perceptions and other factors that influence flood mitigation behavior. *Risk Anal.* 32, 1495–1495.
- Champ, P.A., Brenkert-Smith, H., 2016. Is seeing believing? Perceptions of wildfire risk over time. *Risk Anal.* 36, 816–830.
- Chaurand, N., Delhomme, P., 2013. Cyclists and drivers in road interactions: a comparison of perceived crash risk. *Accid. Anal. Prev.* 50, 1176–1184.
- D'Haese, S., De Meester, F., De Bourdeaudhuij, I., Deforche, B., Cardon, G., 2011. Criterion distances and environmental correlates of active commuting to school in children. *Int. J. Behav. Nutr. Phys. Act.* 8, 88–97.
- Deery, A.H., 1999. Hazards and risk perception among young novice drivers. *J. Saf. Res.* 30, 225–236.
- Elias, W., Shifftan, Y., 2012. The influence of individual's risk perception and attitudes on travel behavior. *Transp. Res. A* 46, 1241–1251.
- Fallah Zavareh, M.F., Mehdizadeh, M., Nordfjærn, T., 2020. "If I know when you will arrive I will let you walk to school": the role of information technology. *J. Saf. Res.* 267–277.
- Fischhoff, B., Slovic, P., Lichtenstein, S., 1978. How safe is safe enough? A psychometric study of attitudes towards technological risks. *Policy Study J.* 9, 127–152.
- Fischhoff, B., Slovic, P., Lichtenstein, S., Combs, B., 2000. How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. In: Slovic, P. (Ed.), *The Perception of Risk*. Earthscan, London, pp. 80–104.
- Glendon, A.I., Dorn, L., Davis, D.R., Matthews, G., Taylor, R.G., 1996. Age and gender differences in perceived accident likelihood and driver competencies. *Risk Anal.* 16, 755–762.
- Hayakawa, H., Fischbeck, P.S., Fischhoff, B., 2000. Traffic accident statistics and risk perceptions in Japan and the United States. *Accid. Anal. Prev.* 32, 827–835.
- Ivers, R., Senserrick, T., Boufous, S., Stevenson, M., Chen, H.-Y., Woodward, M., Norton, R., 2009. Novice drivers' risky driving behavior, risk perception, and crash risk: findings from the DRIVE study. *Am. J. Public Health* 99, 1638–1644.
- Janz, N.K., Becker, M.H., 1984. The Health Belief Model: a decade later. *Health. Educ. Q.* 11, 1–47.
- Kummeneje, A.-M., Rundmo, T., 2018. Subjective assessment of risk among urban work travel cyclists. In: Haugen, S., Barros, A., Gulijk, C., van Kongsvik, T., Vinnem, J.E. (Eds.), *Safety and Reliability: Safe Societies in a Changing World*. Taylor & Francis, London, pp. 459–466.
- Kummeneje, A.M., Rundmo, T., 2019. Risk perception, worry and pedestrian behaviour in the Norwegian population. *Accid. Anal. Prev.* 13, 1–9.
- Kummeneje, A.M., Rundmo, T., 2020. Attitudes, risk perception and risk-taking behaviour among regular cyclists in Norway. *Transp. Res. F: Traffic Psychol. Behav.* 69, 135–150.
- Kummeneje, A.M., Ryeng, E.O., Rundmo, T., 2019. Seasonal variation in risk perception and travel behaviour among cyclists in a Norwegian urban area. *Accid. Anal. Prev.* 124, 40–49.
- Lam, L.T., 2001. Parental risk perceptions of childhood pedestrian road safety. *J. Saf. Res.* 465–478.
- Lawson, A.R., Pakrashi, V., Ghosh, S., Szeto, W.Y., 2013. Perception of safety of cyclists in Dublin city. *Accid. Anal. Prev.* 50, 499–511.
- Llorca, C., Angel-Domenech, A., Agustín-Gómez, F., García, A., 2017. Motor vehicles overtaking cyclists on two-lane rural roads: analysis on speed and lateral clearance. *Saf. Sci.* 92, 302–310.
- Loewenstein, G.F., Weber, E.U., Hsee, C.K., Welch, N., 2001. Risk as feelings. *Psychol. Bull.* 127, 267–286.
- Lund, I.O., Rundmo, T., 2009. Cross-cultural comparisons of traffic safety, risk perception, attitudes and behaviour. *Saf. Sci.* 47, 547–553.
- Manton, R., Rau, H., Fahy, F., Sheahan, J., Clifford, E., 2016. Using mental mapping to unpack perceived cycling risk. *Accid. Anal. Prev.* 88, 138–149.
- Machin, M.A., Sankey, K.S., 2008. Relationships between young drivers' personality characteristics, risk perceptions, and driving behavior. *Accid. Anal. Prev.* 40, 541.

- Mammen, G., Faulkner, G., Buliung, R., Lay, J., 2012. Understanding the drive to escort: a cross-sectional analysis examining parental attitudes towards children's school travel and independent mobility. *BMC Public Health* 12, 862.
- Mehdizadeh, M., Mamdoohi, A., Nordfjærn, T., 2017a. Walking time to school, children's active school travel and their related factors. *J. Transp. Health* 6, 313–326.
- Mehdizadeh, M., Nordfjærn, T., Mamdoohi, A.R., Mohaymany, A.S., 2017b. The role of parental risk judgements, transport safety attitudes, transport priorities and accident experiences on pupils' walking to school. *Accid. Anal. Prev.* 102, 60–71.
- Mehdizadeh, M., Nordfjærn, T., Mamdoohi, A., 2018. The role of socio-economic, built environment and psychological factors in parental mode choice for their children in an Iranian setting. *Transportation* 45, 523–543.
- Moen, B.-E., 2008. Risk perception, priority of safety, and demand for risk mitigation in transport. Doctoral thesis, Norwegian University of Science and Technology, Department of Psychology, Trondheim.
- Moen, B.-E., Rundmo, T., 2006. Perception of transport risk in the Norwegian public. *Risk Manag. Int. J.* 8, 43–60.
- Moller, M., Hels, T., 2008. Cyclists' perception of risk in roundabouts. *Accid. Anal. Prev.* 40, 1055–1062.
- Nevelsteen, K., Steenberghen, T., Van Rompaey, A., Uyttersprot, L., 2012. Controlling factors of the parental safety perception on children's travel mode choice. *Accid. Anal. Prev.* 45, 39–49.
- Nordfjærn, T., Rundmo, T., 2009. Perceptions of traffic risk in an industrialised and a developing country. *Transp. Res. F: Traffic Psychol. Behav.* 12, 91–98.
- Nordfjærn, T., Jørgensen, S., Rundmo, T., 2011. A cross-cultural comparison of road traffic risk perceptions, attitudes towards traffic safety and driver behaviour. *J. Risk Res.* 14, 657–684.
- Nordfjærn, T., Simsekoglu, Ö., Lind, H.B., Jørgensen, S.H., Rundmo, T., 2014. Transport priorities, risk perception and worry associated with mode use and preferences among Norwegian commuters. *Accid. Anal. Prev.* 72, 391–400.
- Nordfjærn, T., Rundmo, T., 2018. Transport risk evaluations associated with past exposure to adverse security events in public transport. *Transp. Res. F: Traffic Psychol. Behav.* 53, 14–23.
- Olteidal, S., Rundmo, T., 2007. Using cluster analysis to test the cultural theory of risk perception. *Transp. Res. F: Traffic Psychol. Behav.* 10, 254–262.
- Oviedo-Trespalacios, O., King, M., Haque, M., Washington, S., 2017. Risk factors of mobile phone use while driving in Queensland: Prevalence, attitudes, crash risk perception, and task-management strategies. *PLoS One* 12, e0183361.
- Peterson, L., Farmer, J., Kashani, J.H., 1990. Parental injury prevention endeavors: a function of health beliefs? *Health Psychol.* 9, 177–191.
- Phun, V.K., Kato, H., Yai, T., 2018. Traffic risk perception and behavioral intentions of paratransit users in Phnom Penh. *Transp. Res. F: Traffic Psychol. Behav.* 55, 175–187.
- Prati, G., Fraboni, F., De Angelis, M., Pietrantonio, L., Johnson, D., Shires, J., 2019. Gender differences in cycling patterns and attitudes towards cycling in a sample of European regular cyclists. *J. Transp. Geogr.* 78, 1–7.
- Rankavat, S., Tiwari, G., 2016. Pedestrians risk perception of traffic crash and built environment features – Delhi, India. *Saf. Sci.* 87, 1–7.
- Rausand, M., Utne, I.B., 2009. *Risikooanalyse: Teori og Metoder*. Akademia, Trondheim.
- Roche-Cerasi, I., Rundmo, T., Sigurdson, J.F., Moe, D., 2013. Transport mode preferences, risk perception and worry in a Norwegian urban population. *Accid. Anal. Prev.* 50, 698–770.
- Rogers, R.W., 1983. Cognitive and physiological processes in fear appeals and attitude change: a revised theory of protection motivation. In: Cacioppo, J., Petty, R. (Eds.), *Social Psychophysiology*. Guilford Press, New York.
- Rogers, R.W., Prentice-Dunn, S., 1997. Protection motivation theory. In: Gochman, D.S. (Ed.), *Handbook of Health Behavior Research*. 1. Personal and Social Determinants. Plenum Press, New York, pp. 113–132.
- Rosenstock, I.M., 1974. Historical origins of the health belief model. *Health Educ. Behav.* 2, 328–335.
- Rottenstreich, Y., Hsee, C.K., 2001. Money, kisses, and electric shocks: on the affective psychology of risk. *Psychol. Sci.* 12, 185–190.
- Rudner, J., 2012. Public knowing of risk and children's independent mobility. *Progr. Plann.* 78, 1–53.
- Rundmo, T., Iversen, H., 2004. Risk perception and driving behaviour among adolescents in two Norwegian counties before and after a traffic safety campaign. *Saf. Sci.* 42, 1–21.
- Rundmo, T., Moen, B.-E., 2006. Risk perception and demand for risk mitigation in transport: a comparison of lay people, politicians and experts. *J. Risk Res.* 9, 623–640.
- Rundmo, T., Nordfjærn, T., Iversen, H.H., Olteidal, S., Jørgensen, S.H., 2011. The role of risk perception and other risk-related judgements in transportation mode use. *Saf. Sci.* 49, 226–235.
- Rundmo, T., Nordfjærn, T., 2013. Predictors of demand for risk mitigation in transport. *Transp. Res. F: Traffic Psychol. Behav.* 20, 183–192.
- Rundmo, T., Nordfjærn, T., 2017. Does risk perception really exist? *Saf. Sci.* 93, 230–240.
- Rundmo, T., Nordfjærn, T., 2019. Judgement of transport security, risk sensitivity and travel mode use in urban areas. *Sustainability* 11, 1908.
- Salmon, J., Salmon, L., Crawford, D.A., Hume, C., Timperio, A., 2007. Associations among individual, social, and environmental barriers and children's walking or cycling to school. *Am. J. Health Promot.* 22, 107–113.
- Simsekoglu, Ö., Nordfjærn, T., Fallah Zavareh, M., Hezaveh, A.M., Mamdoohi, A.R., Rundmo, T., 2013. Risk perceptions, fatalism and driver behaviors in Turkey and Iran. *Saf. Sci.* 59, 187–192.
- Sivak, M., Soler, J., Tränkle, U., Spagnhol, J.M., 1989. Cross-cultural differences in driver risk-perception. *Accid. Anal. Prev.* 21, 355–362.
- Sjøberg, L., 1998. Worry and risk perception. *Risk Anal.* 18, 85–93.
- Sjøberg, L., 1999. Consequences of perceived risk: demand for risk mitigation. *J. Risk Res.* 2, 129–149.
- Sjøberg, L., Moen, B.E., Rundmo, T., 2004. Explaining Risk Perception: An Evaluation of the Psychometric Paradigm in Risk Perception Research. Rotunde Publication, Trondheim.
- Slovic, P., 1992. Perception of risk: reflections of the psychometric paradigm. In: Krinsky, S., Golding, D. (Eds.), *Social Theories of Risk*. Praeger, Westport, CT, pp. 117–152.
- Slovic, P., 2000. Perception of risk. In: Slovic, P. (Ed.), *Perception of Risk*. Earthscan, London, pp. 220–231.
- Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G., 2001. The affect heuristics. In: Gilovic, P., Griffin, D., Kahneman, D. (Eds.), *Intuitive Judgements: Heuristics and Biases*. Cambridge University Press, Cambridge, England.
- Slovic, P., 2007. *The Perception of Risk*, 6th ed. Earthscan Publications, Trowbridge, New York.
- Steinbakk, R.T., Ulleberg, P., Sagberg, F., Fostervold, K.I., 2019. Speed preferences in work zones: the combined effect of visible roadwork activity, personality traits, attitudes, risk perception and driving style. *Transp. Res. F: Traffic Psychol. Behav.* 62, 390–405.
- Tränkle, U., Gelau, C., Metker, T., 1990. Risk perception and age-specific accidents of young drivers. *Accid. Anal. Prev.* 22, 119–125.
- Ulleberg, P., Rundmo, T., 2003. Personality, attitudes and risk perception as predictors of risky driving behaviour among young drivers. *Saf. Sci.* 41, 427–443.
- Van Goeverden, C.D., De Boer, E., 2013. School travel behaviour in the Netherlands and Flanders. *Trans. Policy* 26, 73–84.
- Ventsislavova, P., Crundall, D., Baguley, T., Castro, C., Gugliotta, A., Garcia-Fernandez, P., Zhang, W., Ba, Y., Li, Q., 2019. A comparison of hazard perception and hazard prediction tests across China, Spain and the UK. *Accid. Anal. Prev.* 122, 268–286.
- Verplanken, B., Orbell, S., 2003. Reflections on past behavior: a self-report index of habit strength 1. *J. Appl. Soc. Psychol.* 33, 1313–1330.
- Verplanken, B., Aarts, H., Van Knippenberg, A., 1997. Habit, information acquisition, and the process of making travel mode choices. *Eur. J. Social Psychol.* 27, 539–560.

Further Reading

- Henne, H.M., Tandon, P.S., Frank, L.D., Saelens, B.E., 2014. Parental factors in children's active transport to school. *Public Health* 128, 643–646.
- Muromachi, Y., 2017. Experiences of past school travel modes by university students and their intention of future car purchase. *Transp. Res. A* 104, 209–220.

Social-Symbolic and Affective Aspects of Car Ownership and Use

Birgitta Gatersleben, University of Surrey, School of Psychology, Guildford, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Glossary	81
Introduction	81
Symbolic Aspects of the Car and Identity	81
Brand Image and the Self; Personality and Self-Congruency Theory	82
Cars as Symbols of Success, Status, and Wealth	82
Autonomy, Territoriality, and Attachment	83
Affect	83
Explaining Car Use and Willingness to Change	84
Electric Vehicles	84
Autonomous Cars	84
Conclusion	85
References	85

Glossary

Symbolic aspects of cars the extent to which a car is seen as or used as a symbol of a person's unique characteristics or group membership (including broad social categories such as socio-economic status).

Brand-image congruency the extent to which (perceived) characteristics of a car brand reflect the personality characteristics of the car owner.

Costly-signals of status the extent to which expensive cars are seen and used as symbols (or costly signals) of status and wealth.

Personal and social identity the ways in which people describe themselves and others reflecting their unique characteristics and group membership.

Affective aspects of cars any emotional aspects associated with buying, owning and driving a car such as stress, pleasure, thrill, relaxation as well as social emotions such as pride.

Car pride the self-conscious emotion derived from the appraisal of owning and using cars as a positive self-representation.

Introduction

People buy and drive cars for many reasons. Instrumental reasons, such as flexibility, cost, and speed are important, but other factors, such as status, prestige, pride, and driving pleasure also play a significant role in understanding car ownership and use. The idea that cars have symbolic and affective value that is (related to but) distinct from their instrumental value is well understood by the car industry and has been discussed in academic and non-academic literature for several years (Birdwell, 1968; Diekstra and Kroon, 2003; Evans, 1959; Flink, 1975; Gossling, 2017; Marsh and Collett 1986; Stokes and Hallett, 1992).

The aim of this chapter is to provide an overview of this literature. It will start with defining concepts, outlining theory and discussing research evidence. The chapter will look ahead to review how these factors help explain car use, and responses to travel demand management strategies. It will end with an exploration of the role of these factors for future developments in car technology, looking at electric vehicles and autonomous driving.

Symbolic Aspects of the Car and Identity

Symbolic aspects are associated with the extent to which a car is seen as or used as a symbol of a person's unique characteristics or group membership (including broad social categories such as socio-economic status). The symbolic value of cars is strongly related to social identity theory. Within psychological literature it is well known that people use symbols to relate to each other and the world at large (Dittmar, 2007). Material symbols help people make sense of their social world. Highly visible possessions such as cars enable people to make important quick social judgments about others (their values, intentions, beliefs, status, etc.), especially when more detailed information is not available (e.g., in brief interactions with new people). They also enable people (more or less intentionally) to manage the impressions they make on others. Visual material objects such as the car are extremely useful in these processes of impression formation, and impression management (Schlenker, 1982).

Three interrelated types of symbolic function of cars can be distinguished (Dittmar, 2007). First, cars can be categorical symbols, or signs of a social identity that people can use to express their social standing, wealth, and status and/or to signify group membership. Second, cars can signify special relationships with specific individuals. Finally, cars can have self-expressive function. They can symbolize the owner's personal identity, representing their values, personalities, attitudes, and unique qualities.

The first and latter of these topics have been studied extensively, as will be shown below. The second has received less attention. However, material objects such as gifts or heirlooms have been shown to be prominent symbols of special relationships (Dittmar, 2007). Cars can also fulfil such symbolic function as is demonstrated by this quote from one of the participants of Fraine et al. (2007):

My dad gave it [her car] to me for my 18th birthday and even though I should get rid of it because it's causing me a lot of mechanical problems, I don't want to get rid of it because of that sentimental value [young female] Fraine et al., 2007, p 208.

Many people will have special memories of others associated with their car ownership and use. The car plays an important function in developing and maintaining social relationships with important others—to visit friends, join clubs or ferry the children around (Murtagh et al., 2012).

The symbolic value of cars has been identified through different research methods. Qualitative studies have found that people talk about cars as symbols of prestige (as a symbol of wealth), having an exciting life, masculinity, and being a valued member of society (Hiscock et al., 2002). Experimental studies have demonstrated that people respond differently to others when they are seen with a high or low status car (Dun and Searle, 2010). In self-report surveys, people rate the symbolic value of cars higher than that of other transport modes (Steg, 2005).

Brand Image and the Self; Personality and Self-Congruency Theory

Cars are ideal visual material symbols. They are cultural symbols that reflect socially shared stereotypes of groups of people. One can design an easy (and perhaps fun) matching game with images of cars (e.g., a pink VW Beetle, a blue Peugeot, and a black Mercedes) and characteristics of people (gender, income, job type, personality, etc.). Car advertisers label cars with personality characteristics. Websites publish quizzes to tell people “what your car says about you”. For instance, “Owning a cherry red sports car means you take financial risks”; “Driving a Ford means you're friendly but direct”; “BMW drivers consider themselves knowledgeable”; “SUV drivers are pedantic about safety”; and “Driving a Toyota means you play it safe” (www.hotcars.com).

The personality of cars and people have been studied for quite a few years. In 1959, Evans studied the perceived personalities characteristics of two common cars (Ford and Chevrolet) and their owners to examine whether people drive cars that match their personality. Birdwell (1968) found significant congruence among the perceived personality characteristics of car owners and their cars. This was particularly so for higher priced cars. Many more studies since then have demonstrated that this self-image congruency can explain car brand preferences (Kressmann et al., 2006), car purchase intentions (Ericksen, 1996), and satisfaction with a purchased car (Jamal and Al-Marri, 2007). Such effects have been found for people's perceptions of their actual as well as their ideal self (who they would like to be; Lönnqvist et al., 2019). Self-image congruency theory (Sirgy, 1982, 1986) suggests that consumer choices are influenced by the extent to which consumers perceive a product's image to be consistent with their self-image. Clearly people are drawn towards cars that reflect how they (wish to) see themselves.

Cars as Symbols of Success, Status, and Wealth

Veblen coined the term conspicuous consumption in 1899 to describe how discretionary income was used by the middle classes to display their status and prestige. Spending a lot of money on things you do not need is the ultimate symbol (or costly signal) of status and wealth. Cars are expensive, making them excellent “costly signals” of status and wealth.

The status value of cars has been shown to influence social perceptions. Dunn and Searle (2010) demonstrated that the same male model was perceived as more attractive by women when presented with a ‘high status’ (Silver Bentley Continental GT) rather than a ‘neutral status’ car (Red Ford Fiesta ST). Men were not influenced by this status manipulation. Attractiveness ratings were also studied by Sundie et al. (2011) who demonstrated that a display of status specifically enhances short-term sexual attraction. Men presented with a high status (Porsche) rather than a low status car (Honda) were rated as more desirable for short-term sexual relationships. They were also rated more open to such short-term sexual relationships (Sundie et al., 2011). The status value of the car also affects how men rate each other. Hennighausen et al. (2016) showed that men with luxury cars were rated more attractive, intelligent and ambitious by other men, but they were also more likely seen as a rival, competitor, and mate poacher. These studies suggest that the flaunting a high-status car may communicate to others a very specific type of attractiveness; as someone to have fun with, but not to establish long-term relationships or friendships with.

Car status is related to anger and aggression on the road (Galovski and Blanchard, 2004). Doob and Gross (1968) demonstrated that when a road is blocked by a high-status car 50% of frustrated drivers result into horn honking whereas 84% of frustrated drivers honked when the road was blocked by a low status car. These findings were repeated in more recent work which showed that when a high-status car loiters in front of a traffic light it results in less aggressive responses (honking) from others in comparison to a low status car (McGarva and Steiner, 2000). McGarva and Steiner (2000) also found that drivers of high-status cars are more likely to respond aggressively. When car status of aggressor and frustrater are examined together status similarity appears to result into less aggression.

High-status cars not only change the way people see others they also affect the way we see ourselves and can support a positive sense of self. [Sivanathan and Pettit \(2010\)](#) demonstrated that when people were asked to imagine owning a high-status car (rather than low status) they coped better with negative feedback. Moreover, people with low self-esteem were willing to pay more for a high-status car and low self-esteem helped explain why lower income participants were willing to spend more money on a high-status car. Based on four studies they suggest that participants use status consumption to help restore self-integrity, boost self-esteem, and “restore psychological pain” ([Sivanathan and Pettit, 2010](#)).

Of course there are individual differences. Some people attach more importance to owning the right type of car than others. Male drivers tend to value symbolic and affective outcomes more than female drivers ([Steg et al., 2001](#)). Conspicuous consumption has been shown to be higher among lower income groups who spent a greater proportion of their income on buying luxury status products ([Landis and Gladstone, 2017](#); [Sivanathan and Pettit, 2010](#)). Moreover, people with a stronger materialistic value orientation (who believe that acquiring wealth and material possessions will improve their wellbeing and status; [Richins, 2004](#)) are more likely to want to drive an expensive car for a day, believe this would make them feel powerful and that others would be impressed ([Gatersleben, 2011](#)). Flaunting a high-status car has also been shown to be linked personality. People (particularly men) who score higher on narcissism and disagreeableness are more likely to drive high-status cars ([Lönqvist et al., 2019](#)).

Autonomy, Territoriality, and Attachment

Over time people can develop strong bonds with their car and car ownership and use can become an important part of people’s sense of self ([Goodwin, 1997](#); [Mann and Abraham, 2006](#); [Steg, 2005](#)). Belk (1991) refers to Sartre (1943) who suggested that there are three ways in which possessions may become part of the self: through mastery and control of the object, by creating an object, and through getting to know the object. Following this logic the car can become part of people’s sense of self when they learn to drive and master the object, through ownership and personalization, and through use and experience.

[Fraine et al. \(2007\)](#) examined car attachment through the lens of territoriality theory and demonstrated that for some people the car can become an important primary territory, which is reflected in the way they refer to their car as important to their sense of self, as well as their use of territorial marking (personalizing the car) and behavioral responses to territorial invasion. Higher car attachment was associated with stronger negative reactions to territorial invasion: tailgating. Others have also shown that those who identify more strongly with their car are more likely to find it annoying if someone were to touch their car, write in dust on their car, change settings on the dashboard or park too close ([Watts, 2008](#)).

Possessions are psychologically important to people when they provide control and a sense of mastery ([Dittmar, 2007](#); [Furby, 1978](#)). Owning and driving a car can provide a sense of control over people’s social and physical world that goes well beyond what a human being can achieve without the car. This is something that no other mode of transport can provide in the same way ([Diekstra and Kroon, 2003](#); [Flink, 1975](#); [Goodwin, 1995](#); [Marsh and Collett, 1986](#)). Autonomy and related aspects such as a sense of control, independence and freedom are often rated by people as the key advantage of cars ([Hiscock et al., 2002](#); [Lupton, 2002](#); [Mann and Abraham, 2006](#); [Steg, 2003](#); [Stradling et al., 1999](#)). The convenience provided by cars gives people a sense of control over their lives ([Hiscock et al., 2002](#)), whereas the ‘unreliability [of public transport] does the opposite’ ([Hiscock et al., 2002](#)).

Autonomy related to a car is twofold ([Jensen, 1999](#)): (1) freedom from others (privacy) and (2) freedom from time schedules ([Hiscock et al., 2002](#); [Lupton, 2002](#); [Mann and Abraham, 2006](#)). The first is linked to Fraine’s perspective of the car as a primary territory. Autonomy has a strong instrumental basis; for instance, being able to close the car door to control privacy or having the option to choose travel time and route (e.g., [Hiscock et al., 2002](#); [Lupton, 2002](#); [Mann and Abraham, 2006](#); [Steg, 2003](#); [Stradling et al., 1999](#)). The sense of control and mastery that car ownership and use brings plays an important role in explaining car attachment (Belk, 1991).

Affect

Affective aspects refer to the any emotional aspects associated with buying, owning and driving a car such as pleasure, thrill, relaxation as well as social emotions such as pride. Affective aspects can refer to both negative and positive emotional experiences such as feeling stressed, excited, pleasant, or bored. It is well known that driving (to work) can be stressful ([Wener and Evans, 2011](#)). However, car commuter journeys are more likely to evoke feelings of relaxation, freedom, and control ([Gatersleben and Uzzell, 2007](#)), especially when compared to public transport (e.g. [Jensen, 1999](#)). For many people the car is a place to relax and unwind after a day of work (e.g., [Steg et al., 2001](#); [Gatersleben and Uzzell, 2007](#); [Mann and Abraham, 2006](#)). As one of Lupton’s (2002) respondents noted: ‘it’s a release sometimes . . . get onto the road . . . and chill out’ (p. 280).

Positive affective experiences of car use are associated with the two autonomy functions of the car: freedom from others and from timetables, as is nicely demonstrated by these two quotes from [Fraine et al. \(2007\)](#):

“I feel relaxed when I sit in my car. Sometimes, if I’m not going for a drive, I’ll just go sit in it and put on the radio. And in a way I’m in my element a little bit yeah. No distractions” [young male] [Fraine et al., 2007](#), p 208. Yeah for me it’s a link to the outside world. It’s freedom. I can remember there’s one time where I didn’t have access to it and you really feel trapped [adult male] [Fraine et al., 2007](#) p 208

Positive affective experiences of driving have been linked to symbolic aspects. [Zhao and Zhao \(2015\)](#) coined the term car pride defined as the self-conscious emotion derived from the appraisal of owning and using cars as a positive self-representation. Car pride captures symbolic aspects such as the degree to which people believe owning and driving a car provides prestige and distinction (classy car, distinguish you from ... give you prestige), and affective aspects related to car attachment (I'm fond of my car), driving pleasure (exciting, pleasure) and autonomy (control). The link between symbolic and affective aspects is also made by others. [Bergstad et al. \(2011\)](#), for instance, found an empirical distinction between affective-symbolic aspects (kick out of driving, prestige, sense of power, pleasure, etc) and instrumental-independence aspects (flexibility and control go where you want), but not between affective and symbolic aspects.

Explaining Car Use and Willingness to Change

The perceived instrumental, symbolic and affective benefits (and costs) of the car predict car ownership and use ([Steg et al., 2001, 2005](#)). Moreover, the car is rated more positively than other modes on almost all aspects ([Steg, 2003](#)). Many other studies have confirmed these findings demonstrating that car ownership and use for a range of different purposes (leisure, work) is associated with these perceived benefits ([Anable and Gatersleben, 2005; Bergstad et al., 2011; Bălău, 2019; Ingeborgrud and Ryghaug, 2019; Van Tan and Fujii, 2008](#)).

Psychological attachment to the car has also been linked to levels of car use ([Fraine et al., 2007](#)) and people who are emotionally attached to their car tend to drive more compared to those who are less emotionally attached ([Nilsson and Küller, 2000](#)). Car attachment is also linked to greater resistance to any proposed changes to car ownership and use ([Goodwin, 1995; Sheller, 2004](#)). The more car drivers derive a sense of personal identity from driving, the less willing they are to reduce their car use ([Stradling et al., 1999](#)). People who feel more psychologically attached to their car express greater reluctance to accept a range of travel demand management strategies ([Nilsson and Küller, 2000](#)) and a reduced willingness to accept car sharing schemes ([Paundra et al., 2017](#)). The extent to which travel demand measures are perceived to threaten people's identities (as a good parent, as a car user) has been shown to predict resistance to accept these measures and change behavior ([Murtagh et al., 2012](#)).

Electric Vehicles

Although psychological attachment to cars may be associated with lower willingness to reduce car use or accept measures that curtail car use, this may not necessarily be the same for people's willingness to adopt new car technologies ([Griskevicius et al., 2010; Ingeborgrud and Ryghaug, 2019; Schuitema et al., 2013](#)). The symbolic value of these cars has been shown to be positively related to willingness to adopt these new technologies.

Experimental studies found that participants were more likely to choose a hybrid car rather than a standard luxury car when they were more concerned about their status (in an imagined job interview: [Griskevicius et al., 2010](#)). However, these effects were only found when the hybrid car was more expensive. They were also stronger when choices were made in public than in private (a shop or on-line). This work is in line with the ideas of costly signaling and conspicuous consumption suggesting that more sustainable products may be more attractive when they can be used as costly signals of status.

The relatively importance of symbolic aspects for people's willingness to adopt electric vehicles was also shown by [Schuitema et al. \(2013\)](#) who found that both the perceived hedonic (excitement, pleasure) and symbolic (pride, embarrassment) value of electric vehicles (plug in and hybrid) predicted willingness to buy electric cars. [White and Sintov \(2017\)](#) found that intentions to buy and rent electric cars were higher when they were rated higher as symbols of environmentalism and social innovation. As predicted by self-congruency theory people with a strong pro-environmental self-identity were more likely to have positive perceptions of electric vehicle attributes.

Autonomous Cars

A major technological innovation in the car industry is the autonomous car. People's willingness to adopt such cars is strongly associated with safety concerns and trust in the system. [Fraedrich and Lenz \(2016\)](#) asked German drivers to evaluate a range of autonomous driving scenarios (full autonomy, parking, on demand vehicles) and found acceptance very low. Open ended comments on the scenarios suggested that perceived instrumental costs and benefits of autonomous cars were mixed, affective evaluations were overwhelmingly negative and very few comments referred to symbolic aspects. However, symbolic aspects did appear important in a study among 159 UK car owners ([Bird, 2011](#)). Respondents rated a fictitious normal car and a (semi or fully) autonomous version of that car on a range of values, following [Schuitema et al. \(2003\)](#). The autonomous cars were rated significantly lower than the "normal" car on conservatism (safety, security) and hedonism (pleasure, excitement) but significantly higher on status (pride, power, wealth), see [Fig. 1](#).

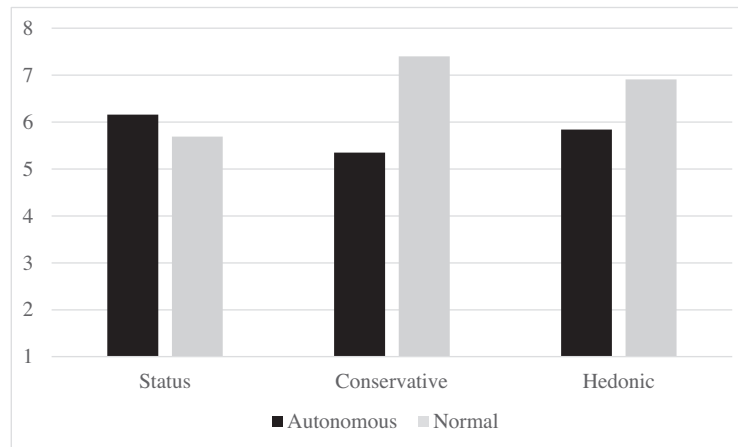


Figure 1 The perceived status, conservative and hedonic value of “normal” and autonomous cars.

Conclusion

Cars are bought and used for many reasons. This chapter demonstrated that the extent to which cars (are seen to) reflect a person’s unique qualities and group membership (status, personality) affects purchase behavior, driver behavior as well as the social interactions. It also showed that positive symbolic and affective experiences influence car use, driver behavior, as well as people’s willingness to change their behavior, accept travel demand measures and adopt new car technologies.

Cars are clearly more than instrumental objects that get us from A to B, although the instrumental aspects of the car are intricately linked to the symbolic and affective aspects. Changes in the material aspects of cars and the social contexts are likely to change those symbolic and affective aspects. For instance, a high-status car will lose its symbolic status value when it is owned by many, technological advances will change perceptions of what is new or exciting and changes in costs will alter its value as a costly signal.

Overall though symbolic and affective aspects of car ownership are important, in understanding why and how people buy and use cars, how they relate to each other and in helping us to understand, predict and plan future demand use management and technological innovations.

References

- Anable, J., Gatersleben, B., 2005. All work and no play? The positive utility of travel for work compared to leisure journeys. *Transport. Res. Part A: Policy Practice* 39, 163–181.
- Ashmore, D.P., Christie, N., Tyler, N.A., 2017. Symbolic transport choice across national cultures: theoretical considerations for research design. *Transport. Plan. Technol.* 40 (8), 875–900.
- Băbuț, M., 2019. Symbolic and affective motives, constraints and self-efficacy among Romanian car buyers. *J. Market. Consumer Behav. Emerging Markets* 8 (1), 14–29.
- Belk, R.W., 1988. Possessions and the extended self. *J. Consumer Res.* 15, 139–168.
- Bergstad, C.J., Gamble, A., Hagman, O., Polk, M., Gärling, T., Olsson, L.E., 2011. Affective–symbolic and instrumental–independence psychological motives mediating effects of socio-demographic variables on daily car use. *J. Transport Geogr.* 19 (1), 33–38.
- Bird, M., 2011. Understanding people’s perceptions of autonomous cars. Unpublished dissertation, University of Surrey.
- Birdwell, A.I., 1968. A study of influence of image congruence on consumer choice. *J. Business* 41, 76–88.
- Carr, H.L., Vignoles, V.L., 2011. Keeping up with the Joneses: Status projection as symbolic self-completion. *Eur. J. Social Psychol.* 41 (4), 518–527.
- Diekstra, R.F.W., Kroon, M., 2003. Cars and behaviour. In: Tolley, R. (Ed.), *Sustainable Transport. Planning for Walking and Cycling in Urban Environments*. Woodland Publishing, Cambridge, pp. 252–265.
- Dittmar, H., 2007. In: To have is to be? Psychological functions of material possessions. In: *Consumer culture identity and well-being*. Psychology Press, pp. 43–66.
- Doob, A.N., Gross, A.E., 1968. Status of frustrator as an inhibitor of horn-honking responses. *J. Social Psychol.* 76 (2), 213–218.
- Dunn, M.J., Searle, R., 2010. Effect of manipulated prestige-car ownership on both sex attractiveness ratings. *Br. J. Psychol.* 101 (1), 69–80.
- Eyben, F., Wöllmer, M., Poitschke, T., Schuller, B., Blaschke, C., Färber, B., Nguyen-Thien, N., 2010. Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. *Adv. Human–Computer Interaction* 2010.
- Ellaway, A., Macintyre, S., Hiscock, R., Kearns, A., 2003. In the driving seat: psychosocial benefits from private motor vehicle transport compared to public transport. *Transport. Res. Part F* 6, 217–231.
- Ericksen, M.K., 1996. Using self-congruence and ideal congruence to predict purchase intention: a European Perspective. *J. Euro –Marketing* 6 (1), 41–56.
- Evans, C.F.B., 1959. Psychological and objective factors in the prediction of brand choice: Ford versus Chevrolet. *J. Business* XXII, 340–369.
- Evans, G., Wener, R., Phillips, D., 2002. The morning rush hour: predictability and commuter stress. *Environ. Behav.* 34 (4), 521–530.
- Flink, J.J., 1975. *The Car Culture*. MIT Press, Massachusetts.
- Fraine, G., Smith, S.G., Zinkiewicz, L., Chapman, R., Sheehan, M., 2007. At home on the road? Can drivers’ relationships with their cars be associated with territoriality? *J. Environ. Psychol.* 27 (3), 204–214.
- Fraedrich, E., Lenz, B., 2016. Taking a drive, hitching a ride: autonomous driving and car usage. In: *Autonomous Driving*. Springer, Berlin, Heidelberg, pp. 665–685.
- Galovski, T.E., Blanchard, E.B., 2004. Road rage: a domain for psychological intervention? *Aggression and Violent Behavior* 9 (2), 105–127.
- Gatersleben, B., 2011. The car as a material possession: Exploring the link between materialism and car ownership and use. In: Lucas, K., Blumenberg, E., Weinberger, R. (Eds.), *Auto Motives. Understanding Car Use Behaviours*, Emerald, pp. 137–150.

- Gatersleben, B., 2007. Affective and symbolic aspects of car use: a review. In: Gärling, T., Steg, L. (Eds.), *Threats to the quality of urban life from car traffic: Problems, causes, and solutions*, Elsevier, Amsterdam, pp. 219–234.
- Gatersleben, B., Uzzell, D., 2007. The journey to work: exploring commuter mood among driver, cyclists, walkers and users of public transport. *Environ. Behav.* 39, 416–431.
- Griskevicius, V., Tybur, J., Van den Bergh, B., 2010. Going green to be seen: status, reputation, and conspicuous conservation. *J. Personal. Social Psychol.* 98, 392–404.
- Goodwin, P.B. (Ed.), 1995, *Car Dependence*. London: RAC Foundation for Motoring and the Environment.
- Goodwin, P.B., 1997. Mobility and Car Dependence. In: Rothengatter, J.A., Carbonell Vaya, E. (Eds.), *Traffic and Transport Psychology. Theory and Application*. Pergamon, Oxford, pp. 449–464.
- Gossling, S., 2017. *The Psychology of the Car: Automobile Admiration, Attachment, and Addiction*. Elsevier.
- Hiscock, R., Macintyre, S., Kearns, A., Ellaway, A., 2002. Means of transport and ontological security: do cars provide psycho-social benefits to their users? *Transport. Res. Part D* 7, 119–135.
- Hennighausen, C., Hudders, L., Lange, B.P., Fink, H., 2016. What if the rival drives a Porsche? Luxury car spending as a costly signal in male intrasexual competition. *Evolutionary Psychol.* 14 (4), 1474704916678217.
- Hotcars. <https://www.hotcars.com/what-a-drivers-choice-of-car-says-about-their-personality/>.
- Ingeborgrud, L., Ryghaug, M., 2019. The role of practical, cognitive and symbolic factors in the successful implementation of battery electric vehicles in Norway. *Transport. Res. Part A: Policy Practice* 130, 507–516.
- Jamal, A., Al-Marri, M., 2007. Exploring the effect of self-image congruence and brand preference on satisfaction: the role of expertise. *J. Market. Manage.* 23 (7–8), 613–629.
- Jensen, M., 1999. Passion and heart in transport – a sociological analysis on transport behaviour. *Transport Policy* 6, 19–33.
- Joo, Y.K., Lee, J.E.R., 2014. Can “The Voices in the Car” Persuade Drivers to Go Green?: Effects of Benefit Appeals from In-Vehicle Voice Agents and the Role of Drivers’ Affective States on Eco-Driving. *Cyberpsychol. Behav. Social Network.* 17 (4), 255–261.
- Kressmann, F., Sirgy, M.J., Herrmann, A., Huber, F., Huber, S., Lee, D.J., 2006. Direct and indirect effects of self-image congruence on brand loyalty. *J. Business Res.* 59 (9), 955–964.
- Landis, B., Gladstone, J.J., 2017. Personality, income, and compensatory consumption: Low-income extraverts spend more on status. *Psychol. Sci.* 28 (10), 1518–1520.
- Lois, D., López-Sáez, M., 2009. The relationship between instrumental, symbolic and affective factors as predictors of car use: A structural equation modeling approach. *Transport. Res. Part A: Policy Practice* 43 (9–10), 790–799.
- Lönnqvist, J.E., Ilmarinen, V.J., Leikas, S., 2019. Not only assholes drive Mercedes. Besides disagreeable men, also conscientious people drive high-status cars. *Int. J. Psychol.*
- Lupton, D., 2002. Road Rage: Drivers’ Understandings and Experiences. *J. Sociol.* 38 (3), 275–290.
- Macintyre, S., Ellaway, A., Der, G., Ford, G., Hunt, K., 1998. Do housing tenure and car access predict health because they are simply markers of income and self esteem? A Scottish Study. *J. Epidemiol. Community Health* 52 (10), 657–663.
- Mann, E., 2006. *Cognitive and Affective Antecedents of Commuter Transport Mode Choices*. Unpublished doctoral dissertation, University of Sussex, UK.
- Mann, E., Abraham, C., 2006. The Role of Affect in UK Commuters’ Travel Mode Choices: An Interpretative Phenomenological Analysis. *Br. J. Psychol.* 97 (2), 155–176.
- Marsh, P., Collett, P., 1986. *Driving Passion: The Psychology of the Car*. Jonathan Cape, London.
- McGarva, A.R., Steiner, M., 2000. Provoked driver aggression and status: A field study. *Transport. Res. part F: Traffic Psychol. Behav.* 3 (3), 167–179.
- Moody, J., Zhao, J., 2019. Car pride and its bidirectional relations with car ownership: Case studies in New York City and Houston. *Transport. Res. Part A: Policy Practice* 124, 334–353.
- Murtagh, N., Gatersleben, B., Uzzell, D., 2012. Multiple identities and travel mode choice for regular journeys. *Transport. Res. Part F: Traffic Psychol. Behav.* 15 (5), 514–524.
- Murtagh, N., Gatersleben, B., Uzzell, D., 2012. Self-identity threat and resistance to change: Evidence from regular travel behaviour. *J. Environ. Psychol.* 32 (4), 318–326.
- Nilsson, M., Küller, R., 2000. Travel behaviour and environmental concern. *Transp. Res. Transp. Environ* 5 (3), 211–234.
- Pandra, J., Rook, L., van Dalen, J., Ketter, W., 2017. Preferences for car sharing services: effects of instrumental attributes and psychological ownership. *J. Environ. Psychol.* 53, 121–130.
- Rezvani, Z., Jansson, J., Bodin, J., 2015. Advances in consumer electric vehicle adoption research: A review and research agenda. *Transport. Res. Part D: Transport Environ.* 34, 122–136.
- Sachs, W., 1984. *Die Liebe zum Automobil*. In: Ein Rückblick in die Geschichte unserer Wünsche, Reikbeck bei Hamburg., Rohwolf.
- Schlenker, B.R., 1982. Translating actions into attitudes: An identity-analytic approach to the explanation of social conduct. In: Berkowitz, L. (Ed.), *Advances in Experimental Social Psychology*, vol. 15. Academic Press, New York.
- Schuitema, G., Anable, J., Skippon, S., Kinnear, N., 2013. The role of instrumental, hedonic and symbolic attributes in the intention to adopt electric vehicles. *Transport. Res. Part A: Policy Pract.* 48, 39–49.
- Sheller, M., 2004. Automotive Emotions. *Feeling the Car. Theory Culture and Society* 21 (4/5), 221–242.
- Sivanathan, N., Pettit, N.C., 2010. Protecting the self through consumption: Status goods as affirmational commodities. *J. Exp. Social Psychol.* 46 (3), 564–570.
- Sirgy, M.J., 1982. Self-concept in consumer behavior: A critical review. *J. Consum. Res.* 9 (3), 287–300.
- Sirgy, M.J., 1986. *Self-Congruity: Toward a Theory of Personality and Cybernetics*. Praeger Publishers/Greenwood Publishing Group.
- Steg, L., Vlek, C., Slotegraaf, G., 2001. Instrumental-reasoned and Symbolic-affective Motives for Using a Motor Car. *Transport. Res. Part F* 4, 151–169.
- Steg, L., 2005. Car use: lust and must. Instrumental, symbolic and affective motives for car use. *Transport. Res. Part A* 39, 147–162.
- Steg, L., 2004. Car use lust and must. In: Rothengatter, T., Huguenin, R.D. (Eds.), *Traffic and Transport Psychology*. Elsevier, Oxford, pp. 443–452.
- Steg, L., 2003. Can public transport compete with the private car? *IATSS Res.* 27 (2), 27–36.
- Stradling, S.G., Meadows, M.L., Beatty, S., 1999. Factors affecting car use choices. Report to the Department of the Environment, Transport and the Regions. Transport Research Institute, Napier University, Edinburgh.
- Stokes, G., Hallett, S., 1992. The role of advertising and the car. *Transport. Rev.* 12 (2), 171–183.
- Sovacool, B.K., Axsen, J., 2018. Functional, symbolic and societal frames for automobility: implications for sustainability transitions. *Transport. Res. Part A: Policy Pract.* 118, 730–746.
- Sundie, J.M., Kenrick, D.T., Griskevicius, V., Tybur, J.M., Vohs, K.D., Beal, D.J., 2011. Peacocks, Porsches, and Thorstein Veblen: Conspicuous consumption as a sexual signaling system. *J. Personal. Social Psychol.* 100 (4), 664.
- Stets, J.E., Burke, P.J., 2000. Identity theory and social identity theory. *Social Psychol. Quart.* 63, 224–237.
- Van Tan, H., Fujii, S. (2008). Psychological Dimensions of Attitudes Toward Car and Public Transport: Symbolic-Affective, Instrumental, and Social Orderliness (No. 08-2901)
- Veblen, T., 1899. *The Theory of the Leisure Class*. Macmillan, New York, NY.
- Watts, L., 2008. An in-depth exploration into the nature of human-car attachment and its implications for pro-environmental travel behaviour. University of Surrey, Unpublisheddissertation.
- Wener, R.E., Evans, G.W., 2011. Comparing stress of car and train commuters. *Transp. Res. Traffic Psychol. Behav.* 14 (2), 111–116.
- White, L.V., Sintov, N.D., 2017. You are what you drive: environmentalist and social innovator symbolism drives electric vehicle adoption intentions. *Transport. Res. Part A: Policy Pract.* 99, 94–113.
- Zhao, Z., Zhao, J., 2015. Car pride: Psychological structure and behavioral implications. In 94th Annual Meeting of the Transportation Research Board, Washington, DC.

Pitfalls of Statistical Methods in Traffic Psychology

J.C.F. de Winter, D. Dodou, Delft University of Technology, Delft, The Netherlands

© 2021 Elsevier Ltd. All rights reserved.

Introduction	87
The <i>P</i> -Value is Not the Same as the Probability That the Null Hypothesis is True	87
Everything is Correlated	89
Letting the Statistical Test Depend on a Statistical Test	91
Violation of Independence	92
Conclusion	94
Supplementary Materials	95
References	95

Introduction

In traffic psychology, researchers typically conduct an experiment, the data of which are subsequently analyzed. For example, a researcher may conduct a driving simulator experiment with different time-budget conditions to evaluate how drivers reclaim control from an SAE Level 3 automated vehicle. For each driver and time-budget condition, scores on measures are obtained, such as the maximum absolute steering wheel angle, self-reported criticality of the situation, and take-over time. In other cases, the traffic psychologist may investigate individual differences, for example, by analyzing accident records or data obtained from Internet-based questionnaires.

The available experimental data are invariably subjected to statistical tests. We analyzed the papers of 129 experiments on automation-to-manual take-overs (list of papers available in a meta-analysis by [Zhang et al., 2019](#)) and found that *P*-values were used in as much as 95% of the 129 studies. ANOVAs (65% of studies), usually combined with post hoc (Bonferroni-corrected) *t*-tests, as well as paired/unpaired *t*-test presented separately (46% of studies), were among the most common statistical procedures. Nonparametric test variants (e.g., Wilcoxon test) were common as well. In areas of traffic psychology that are concerned with individual differences, on the other hand, correlation coefficients and data reduction techniques, such as exploratory factor analysis, are typically used ([Fig. 1](#)).

When we were asked to contribute an article about statistics in traffic psychology, we realized that there are already many articles and books that detail how statistical tests should be conducted, sometimes accompanied by an extensive step-by-step plan tailored to a specific software package (e.g., [Field, 2013](#)). It does not appear scientifically valuable to repeat such information. What appears valuable, however, is to report on common pitfalls and misconceptions regarding statistics in the field of traffic psychology.

This article was written based on the first author's subjective experiences as a researcher and peer reviewer in the field. The pitfalls were selected based on estimates of risk, where risk is a function of how often these pitfalls occur and how severe their consequences are. There are, of course, many other pitfalls of statistical methods, but the list below is regarded to be of direct relevance to the traffic psychology researcher.

The *P*-Value is Not the Same as the Probability That the Null Hypothesis is True

In a large proportion of papers in traffic psychology, statistical tests are performed. These tests yield *P*-values, based on which statements are made about hypotheses. Formally, the *P*-value indicates how extreme one's observations are under the assumption of a null hypothesis. A critical thing that tends to be overlooked in null-hypothesis significance testing is that *P*-values do not allow for making direct claims about whether the null hypothesis is true or false.

A thought experiment can prove the point. Suppose one investigates psychic abilities, such as whether participants can predict the outcome of a random number generator (see [Bem, 2011](#)). This hypothesis is certainly intriguing and spectacular if true. However, it is impossible to be correct, given the current knowledge of physics, and the fact that, as far as we know, no person in the world can systematically outplay the roulette table at casinos. Any significant *P*-value ($P < 0.05$) must therefore be a false positive (see [Wagenmakers et al., 2011](#) for a similar argumentation).

The above message is illustrated in [Fig. 2](#). Let us assume that, in a given subfield of traffic psychology, as much as 95% of the hypotheses are false. That is, there is only a 5% probability that the effect one hopes to find exists in reality. In such a case, even if assuming a large sample size and corresponding statistical power of 80%, then still only 46% of the statistically significant effects reflect a true effect. In other words, it is more likely that a significant effect ($P < 0.05$) is a false positive than a true positive.

This type of Bayesian reasoning corresponds to the argumentation of John Ioannidis in his well-known work "Why most published research findings are false" ([Ioannidis, 2005](#)). [Fig. 2](#) provides an example where statistical power is reasonably high

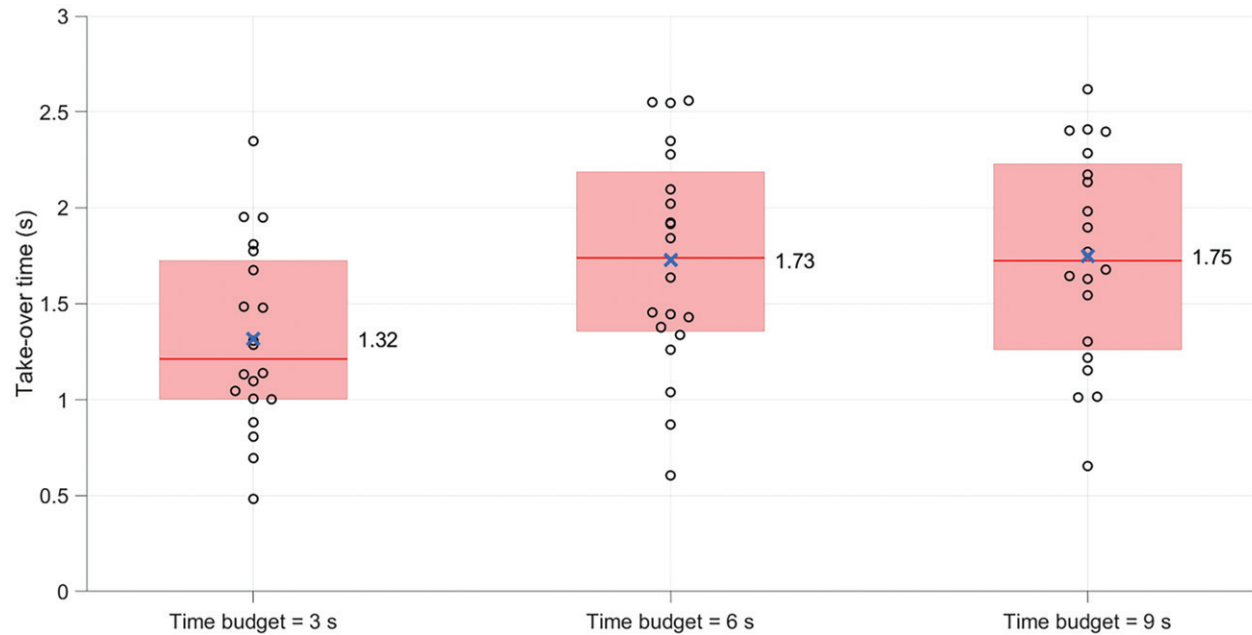


Figure 1 Example (simulated) data for an experiment on automation-to-manual take-overs, with three time-budget conditions. A sample size of 20 per condition was assumed. The blue cross and the number next to each bar indicate the group mean.

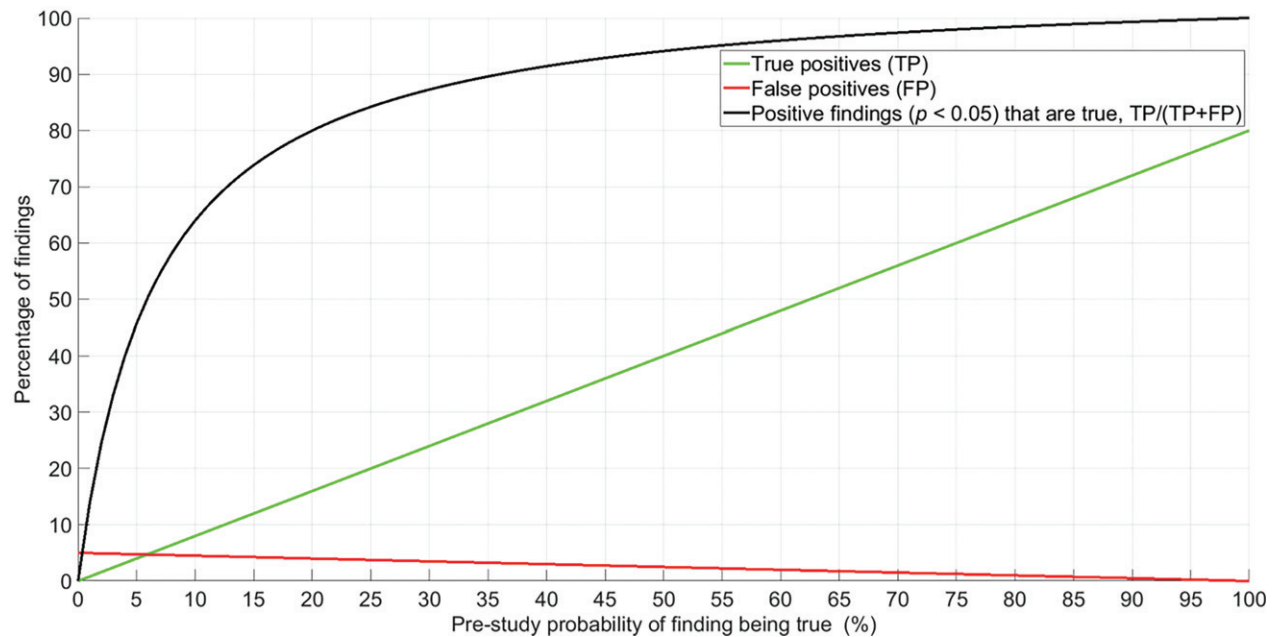


Figure 2 Illustration of the effect of the pre-study probability that a research finding is true for statistical power $(1 - \beta) = 80\%$. For a significance level (α) of 5% and a pre-study probability of 5%, only 46% of the detected statistically significant effects are true positives.

(80%). Using the same figure, it can be understood why low statistical power (e.g., small expected effects or small sample sizes) reduces the probability that a research finding is true. For example, if power were 60% instead of 80%, then there would be fewer true positives, and, as a result, only 39% of the positive (i.e., statistically significant) results would be true positives. In other words, low statistical power does reduce not only the probability of detecting an effect but also the likelihood that an observed significant effect is a true effect (Button et al., 2013). Note that the example in Fig. 2 does not even consider the phenomenon of bias, where researchers may unconsciously or consciously distort research findings because researchers generally like to find significant effects (Ware and Munafò, 2015). Bias is discussed in the third section of this article.

In traffic psychology, effects are often well established and replicable, because experimental conditions are clearly operationalized (e.g., human-machine interface conditions, traffic conditions in a driving simulator), and the driver's inputs (e.g., steering

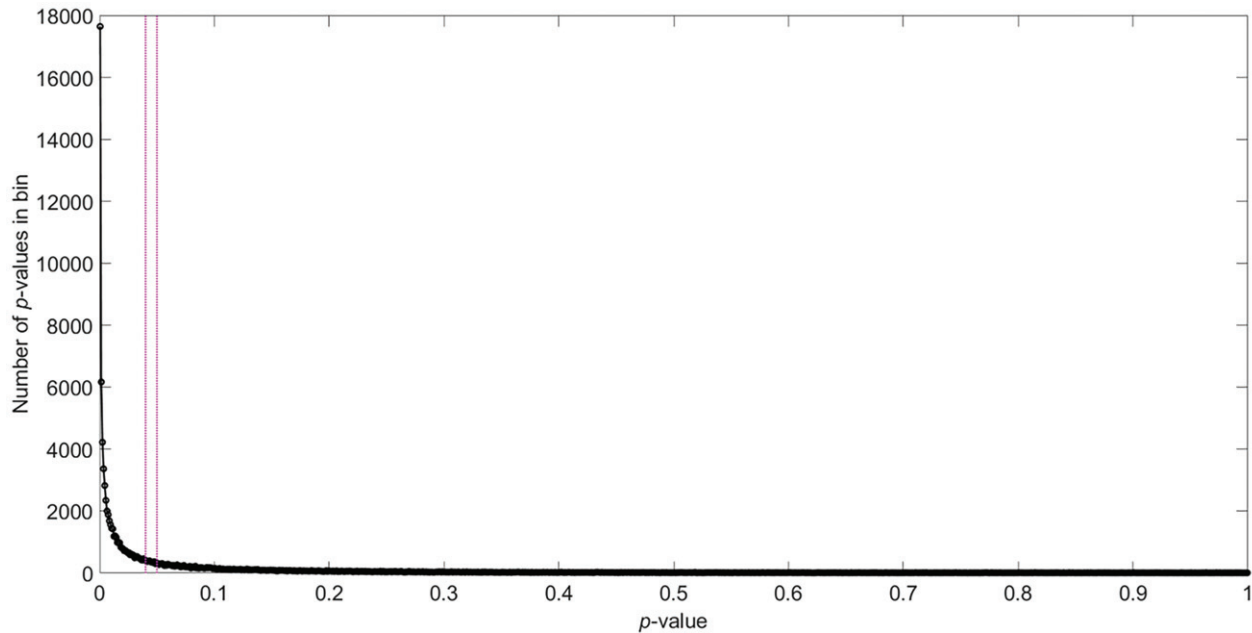


Figure 3 Distribution of P -values (so-called P -curve) when submitting two samples ($n = 20$ each), with Cohen's $d = 0.8$, to an independent-samples t -test. The vertical dashed lines are drawn at $P = 0.04$ and $P = 0.05$, respectively. The bins in this figure are 0.001 wide.

wheel angle, throttle and brake inputs) are easily measurable and causally related to the task conditions. For example, the effect of time budget and take-over request modality on take-over times are known to be strong and reproduced many times (see Zhang et al., 2019, for a meta-analysis). However, in subfields of traffic psychology, researchers may sometimes make rather spectacular, surprising, or intriguing claims. Examples may be the effects of emotional priming, mindfulness, or pleasant smells in the car on driving behavior. While the existence of such effects cannot be ruled out, the underlying mechanisms of how these would affect driving behavior are less clear, and hence the credibility of the findings is low. When these types of findings are obtained via small samples and supported by a barely significant effect (e.g., $P = 0.04$), scepticism is appropriate. Conversely, if an effect is supported by theory or the P -value is small (e.g., $P < 0.001$), then the finding is more likely to be replicable.

Incidentally, it should be noted that P -values below 0.05 are more prevalent in the literature than they should be. It is well established that significant findings ($P < 0.05$) are more likely to appear in papers than nonsignificant findings ($P > 0.05$), a phenomenon known as publication bias. Research also shows that the use of P -values is becoming increasingly common in the abstracts of papers (Chavalarias et al., 2016; De Winter and Dodou, 2015), and there is some evidence for “convenient” downwards rounding of P -values to be able to claim “ $P \leq 0.05$ ” (Hartgerink et al., 2016).

The expected rarity of P -values in the $P = 0.04$ – 0.05 region can be illustrated using a computer simulation. Suppose one wishes to compare the means of two samples. The first sample is drawn from a population with a mean of 0 and a standard deviation of 1; the second sample is drawn from a population with a mean of 0.8 and also a standard deviation of 1. In other words, the true effect size is large (Cohen's $d = 0.8$). If one draws random samples of $n = 20$ each and performs 100,000 independent-samples t -tests, then most P -values will turn out to be very small. In this specific case, only 3.6% of the P -values are between 0.04 and 0.05 (see Fig. 3 for the results of the simulation).

In summary, a P -value slightly below 0.05 does not imply that the alternative hypothesis is true; one should consider the plausibility of the hypothesis (Ioannidis, 2005). Furthermore, if there truly is a strong effect in the population, P -values just below 0.05 are rare, and small P -values (e.g., $P < 0.001$) are more likely.

Possible solutions to the above-mentioned problems are the use of larger sample sizes (Button et al., 2013) and the use of lower alpha values (e.g., 0.005 instead of 0.05; Benjamin et al., 2018; but see Lakens et al., 2018, for counterarguments). Some researchers in traffic psychology have used Bayesian statistics to counteract the problem of making wrong inferences based on null hypothesis significance testing (see Körber et al., 2016 for a study on Bayesian statistics in automation-to-manual take-overs). Although Bayesian statistics are popular in several areas of transportation research, its use in traffic psychology and other human-related fields is still limited (see Chavalarias et al., 2016 for a survey of biomedical papers).

Everything is Correlated

In traffic psychology, researchers sometimes have access to extremely large databases. For example, it has been found in a study of over 10,000 drivers that self-reported violations are predictive of self-reported accidents (Wells et al., 2008) and that demographic variables are predictive of the willingness to use a driverless shuttle (Nordhoff et al., 2018). Similarly, accident statistics of hundreds

of thousands of drivers have revealed interesting relationships between crash involvement, traffic tickets, and individual characteristics (Factor, 2014). For such large sample sizes, however, even extremely small effects are statistically significantly different from 0 (Nunnally, 1960; Standing et al., 1991), an inconvenience which Meehl (1990) referred to as the “crud factor” (p. 108).

The bottom line is that a statistically significant effect does not imply that this effect is important. For example, the correlation between self-reported violations and self-reported accidents is about 0.15 (De Winter et al., 2015), which may be interesting for testing theories about personal predispositions but probably not very useful for identifying accident-prone drivers. Colleague Evans (2009) once explained that, in principle, wearing a second t-shirt contributes to a reduction of road traffic deaths, as this extra layer of clothing has a protective effect in absorbing energy during a crash. However, as he pointed out, this, of course, does not mean that researchers should advise drivers to put on an extra t-shirt before entering their cars.

The above concerns about the meaninglessness of null hypothesis significance testing do not only apply to small effects; they also apply to strong effects, in particular in the context of human performance and behavior. A common phenomenon in the analysis of human performance, including driving, is that performance measures are almost always correlated. For example, a driver who is skilled in lane-keeping will probably also be skilled in car following, and a positive statistically significant correlation coefficient between performance measures (e.g., standard deviation of lateral position and standard deviation of headway) is virtually guaranteed. Horn and McArdle (2007) noted: “practically every variable that in any sense indicates the good things in life—health, money, education, and so forth, and athletic and artistic abilities, as well as cognitive abilities—correlate positively with every other such thing that everything is significantly correlated” (p. 219).

In questionnaire research as well, strong positive correlations between variables are the rule rather than the exception, due to method effects. This problem, which has been vigorously highlighted by a few authors in traffic psychology before (in particular Af Wählberg, 2010, 2017), seems to have fallen on deaf ears in a large portion of researchers.

The problem of methods effects in questionnaire research can be illustrated using a small computer simulation (see also De Winter et al., 2019). First, let us assume two constructs (e.g., driving errors and driving violations), which are uncorrelated in the population. Let us also assume that these constructs can be measured using a five-point Likert scale (e.g., 1 = never, 2 = hardly ever, 3 = occasionally, 4 = quite often, 5 = frequently). The distribution of the correlation coefficients, assuming a sample size of 300 respondents, is shown in Fig. 4 (this distribution is based on 100,000 repetitions) and averages at $r = 0.00$, as expected. Next, one-third of the respondents (100 of 300) are simulated to have a different notion of quantity and the meaning of words such as “occasionally”. Let us assume that these respondents report on average 2 points lower than a bias-free sample (i.e., they report “never” instead of “hardly ever” and “occasionally”, “hardly ever” instead of “quite often”, and “occasionally” instead of “frequently”). Additionally, one-third of the respondents report on average 2 points higher than a bias-free sample (i.e., more toward “frequently”). The distribution of the overall sample, yielding an average correlation between the two variables of 0.55, is provided in Fig. 4. What this simulation shows is that method effects alone, in this case, response style (“nay-saying”, “yea-saying”), can cause strong correlations (see also De Winter and Dodou, 2017).

The problem illustrated here seems to be strikingly common in traffic psychology. One group of unnamed authors, for example, evaluated the acceptance of automated vehicles using the Unified Theory of Acceptance and Use of Technology (UTAUT). They

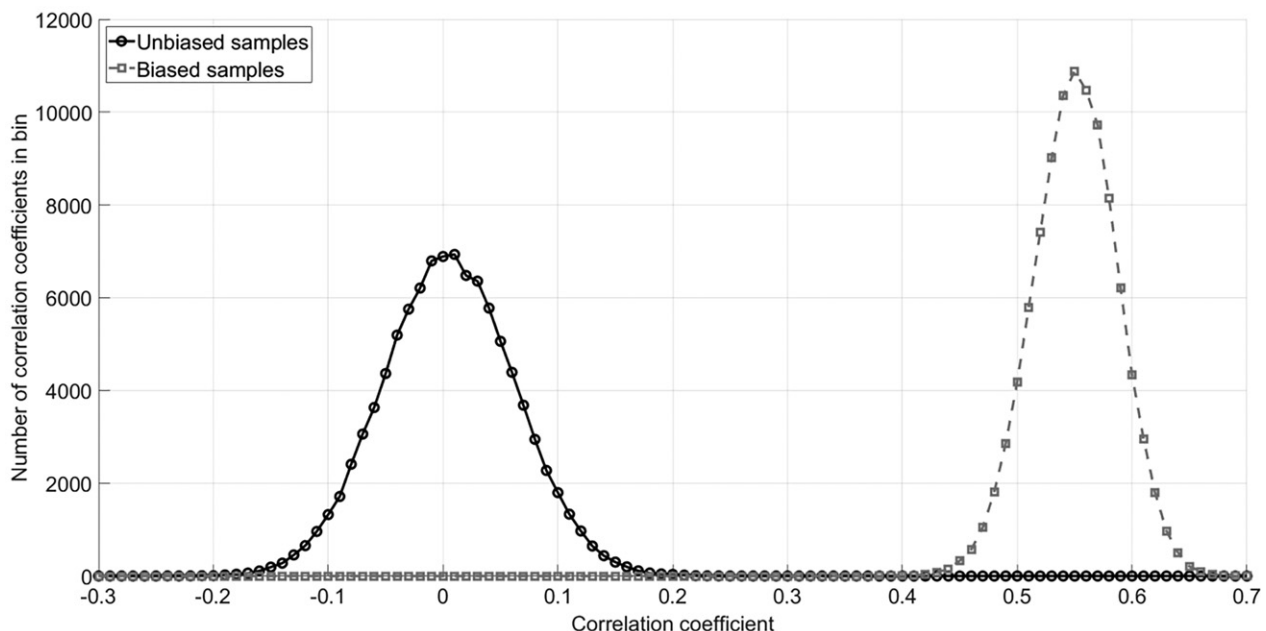


Figure 4 Distribution of Pearson product-moment correlation coefficients between two variables for a sample size of 300 when the correlation in the population is 0 (unbiased samples) and when the population correlation is 0 but response style contaminates the results (biased samples). The bins in this figure are 0.01 wide.

found positive correlations ($r = 0.5$ to 0.6 , all P -values < 0.001) between behavioral intentions to use an automated vehicle, on the one hand, and effort expectancy, social influence, facilitating conditions, and hedonic motivation, on the other. These findings, although correct from the viewpoint of null hypothesis significance testing, do not imply that one can predict real-life behavioral intentions that strongly. The positive correlations probably reflect a method effect or a form of social desirability or self-esteem which translates into response style.

The bottom line is that effect sizes should be reported and interpreted. Researchers should not focus solely on P -values: even trivially small effects can yield highly significant P -values, and strong effects are, of course, statistically significantly different from 0, but may not reflect the true state of the world. Possible remedies to common method biases are to measure predictor and criterion variables at different points in time, the use of forced-choice survey items (Bartram, 2007; Brown and Maydeu-Olivares, 2011; Nederhof, 1985), and the use of a criterion variable that is recorded rather than self-reported (Af Wählberg, 2017).

Letting the Statistical Test Depend on a Statistical Test

In traffic psychology, it often happens that researchers pretest their data. For example, a researcher may test for normality before deciding whether to conduct a parametric t -test or ANOVA or a nonparametric Wilcoxon rank-sum test, Kruskal–Wallis test, or Friedman test. A problem with pretesting is that the “numbers don’t remember where they came from” (Lord, 1953, p. 21), and that such tests may not have sufficient power for detecting violations of the assumptions of the follow-up tests (Rasch et al., 2011; Zimmerman, 2011).

Suppose a researcher is faced with two samples of data which he would like to statistically compare. The researcher decides to conduct a t -test or a Wilcoxon rank-sum test based on the result of a Shapiro–Wilk test of normality. The issue here is that, if the sample size is higher, the statistical power of the Shapiro–Wilk test will tend to be higher as well, and so the outcome of the Shapiro–Wilk test depends on sample size as well as the idiosyncratic nature of the dataset. It is much more efficient to decide beforehand which statistical test to use. For example, it is already well known that Driver Behaviour Questionnaire (DBQ) item data are heavily tailed (e.g., Mattsson, 2012). It may therefore be better to use tests which are robust to tailed distributions (e.g., Wilcoxon rank-sum tests, signed-rank tests, and Spearman’s rank-order correlations). No pretest is needed if prior research already reveals which type of test is most appropriate.

A computer simulation could prove the point. Suppose one wishes to test whether the central tendency of two distributions is different. The first distribution has a mean of 0 and a standard deviation of 1, and the second distribution has a mean of 1 and a standard deviation of 1 (i.e., Cohen’s $d = 1$). Furthermore, the sample size is 20 per group, and the first group features an outlier. More specifically, the value for the 20th participant in the first group equals 4 (i.e., the mean + 4 standard deviations).

Table 1 shows the results of a simulation study of a two-stage process, with 100,000 repetitions. That is, 100,000 times, 20 samples were drawn from the first distribution, and 20 samples were drawn from the second distribution, and the two samples were compared using an independent-samples t -test and a Wilcoxon rank-sum test, as well as a conditional test. In the conditional test, an independent-samples t -test was conducted when a Shapiro–Wilk test used on the first sample yielded $P \geq 0.05$, and a Wilcoxon rank-sum test was conducted when the same Shapiro–Wilk test yielded $P < 0.05$.

Table 1 reveals that the statistical power is higher for the Wilcoxon rank-sum test than for the t -test, which can be explained by the outlier. After all, it is well established that the t -test is not particularly robust to outliers. The conditional approach improves statistical power substantially compared to the independent-samples t -test but is still less powerful than the Wilcoxon rank-sum test.

Using an insurance policy analogy (Anscombe, 1960), it can be wise to choose a nonparametric test such as a Wilcoxon rank-sum test or a Welch’s t -test (Lakens, 2015a) and not use a pretesting approach. Such a test yields a slight loss of power compared to the regular t -test if the normality assumption is met. However, the insurance premium is only small for the high protection it offers against outliers or unequal variances.

More damaging outcomes can occur when the researcher chooses the statistical test based on statistical significance. For example, a researcher may decide to remove outliers and subsequently perform the statistical test again, or may first perform a regular t -test, and, if the result is not significant, perform a Wilcoxon-rank-sum test. Dropping or merging of experimental conditions, optional stopping, or adding covariates/moderator variables (e.g., age, gender) are similar examples. These types of so-called questionable research practices, which can dramatically increase false-positive rates, have been previously identified in psychological research (Bakker and Wicherts, 2014; Nelson et al., 2018), and may well be exacerbated in traffic psychology, an area that has been progressing in parallel to mainstream psychology and engineering.

Table 1 Results of a simulation study for detecting a significant difference between two groups in the presence of an outlier in one of the two groups

	Statistical power (% of 100,000 repetitions yielding $P < 0.05$)
Independent-samples t -test	56.7%
Wilcoxon rank-sum test	72.3%
Conditional test	66.9%

Note. The Shapiro–Wilk test yielded a significant effect in only 60.5% of the 100,000 repetitions.

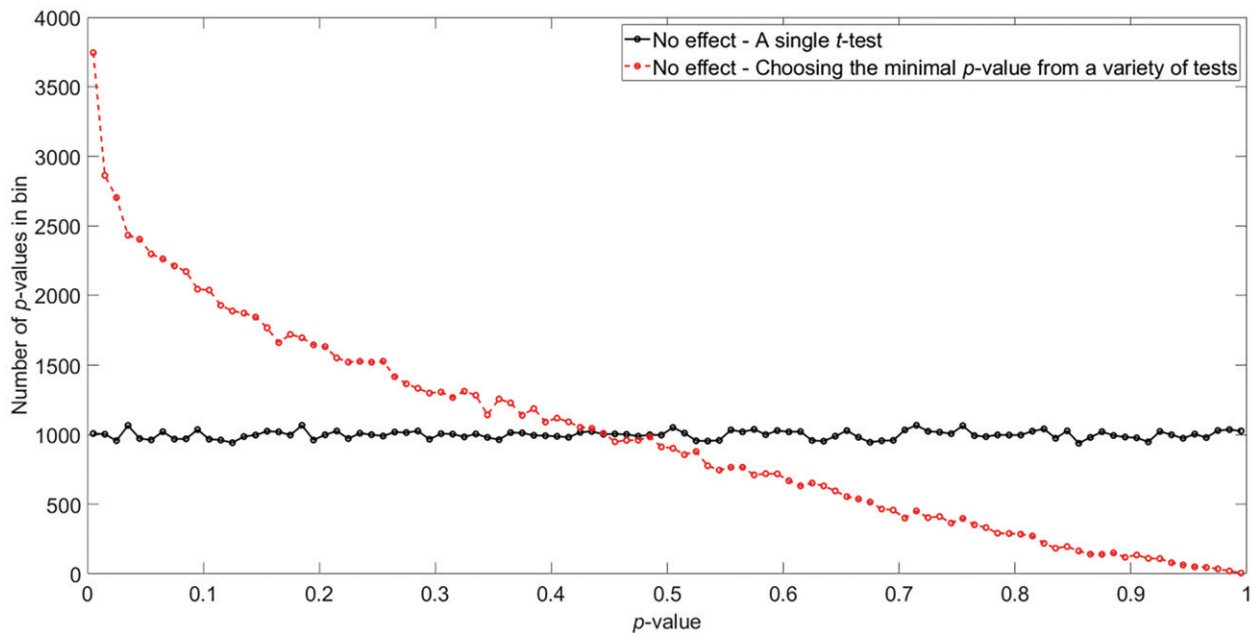


Figure 5 *P*-value distribution (*P*-curve) in case there is no effect in the population. By “trying out” various statistical test and reporting the “best” effect, the false positive rate may increase substantially (from 5.0% to 14.1% in this simulation). The bins in this figure are 0.01 wide.

Fig. 5 illustrates the result of independent-samples *t*-tests in black. In this case, two samples with equal means and standard deviations were compared, $n = 20$ per group. The results of this simulation, performed with 100,000 repetitions, show that the distribution of the *P*-value is uniform (see also Lakens, 2015b).

The red line in Fig. 5 shows the results of statistical tests for the same data, but here we simulated researchers who engaged in questionable research practices. More specifically, the simulated researchers all performed an independent-samples *t*-test, a paired-samples *t*-test, an independent-samples *t*-test after automated outlier removal, and an independent samples *t*-test for subgroups. The latter could be done, for example, if the samples consist of 10 males and 10 females, and the researchers test for a difference in means for males and females separately. If the researchers use the minimum *P*-value among all five options, the *P*-values become lower than they should be. In this case, the false-positive rate has become 14.1%, that is, 14.1% of the 100,000 cases yielded $P < 0.05$.

In traffic psychology, a popular method is to perform a stepwise regression analysis. A stepwise analysis works by including statistically significant predictors only and removing variables when they are not significant, and suffers from similar problems as the problems as illustrated in Fig. 5. Further critique of stepwise regression analysis is provided by Antonakis and Dietz (2011) and Thompson (1989), amongst others.

In summary, the statistical test should not depend on another statistical test: Two-stage approaches can yield suboptimal statistical power and, more worryingly, peeking at the data and adjusting the statistical test accordingly increases the prevalence of false positives. Methods such as preregistration of one’s hypotheses (Nosek et al., 2018) can help prevent the issues mentioned in this section.

Violation of Independence

It often happens that the traffic psychologist computes correlation coefficients to explore relationships between variables collected in the experiment. For example, the researcher may be interested in the relationship between take-over time and the maximum absolute steering wheel angle. Fig. 6 shows a scatter plot of take-over time and maximum absolute steering wheel angle, as could have been obtained from a simulator-based experiment. In this case, the association between the two variables is moderate and not significantly different from 0 ($r = 0.27$, $P = 0.247$).

Typically, experiments contain multiple repetitions of the same scenario. For example, there may have been five identical take-over trials per time budget condition. In traffic psychology, the first author has seen it happening on numerous occasions that researchers inadvertently treat those repeated measures as independent. For the sake of argument, let us assume that participants are consistent with respect to themselves. This is a reasonable assumption in behavioral research, where in the long run, participants exhibit stable behavioral patterns (excluding learning, mood swings, etc.). Thus, it is assumed that the 20 participants output practically the same take-over time and the same maximum absolute steering wheel angle for each of the identical five take-over trials.

The corresponding scatter plot is drawn in Fig. 7; it contains the same data as in Fig. 6, except that there are now 100 data points because the repeated measures for the same persons are plotted separately, whereas the markers in Fig. 6 were assumed to reflect the

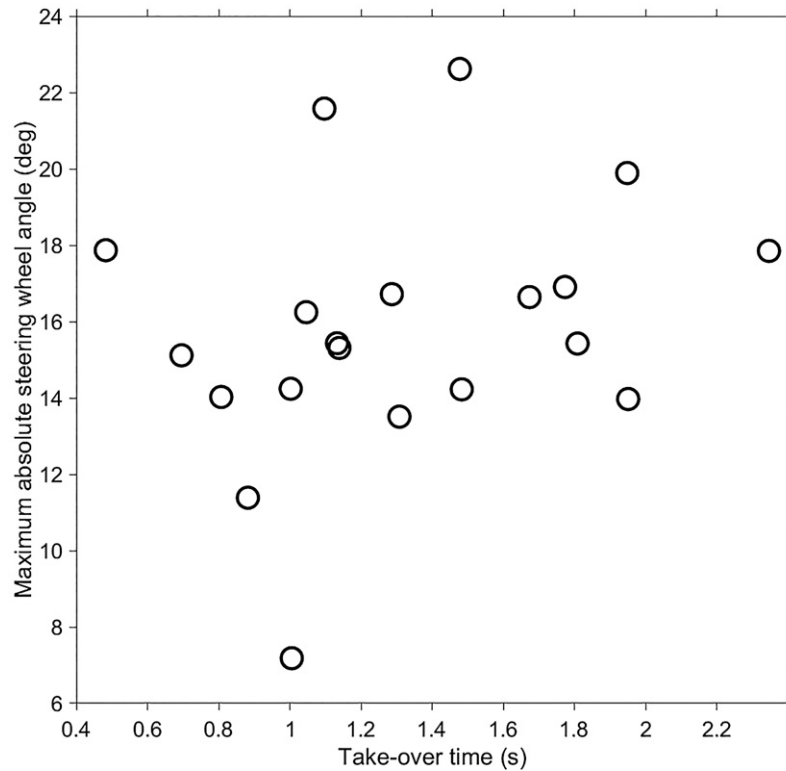


Figure 6 Example (simulated) data for an experiment on automation take-overs. The horizontal axis shows the take-over time (same data as the data in Fig. 1 for the 3 s time budget condition), and the vertical axis shows the maximum absolute steering wheel angle.

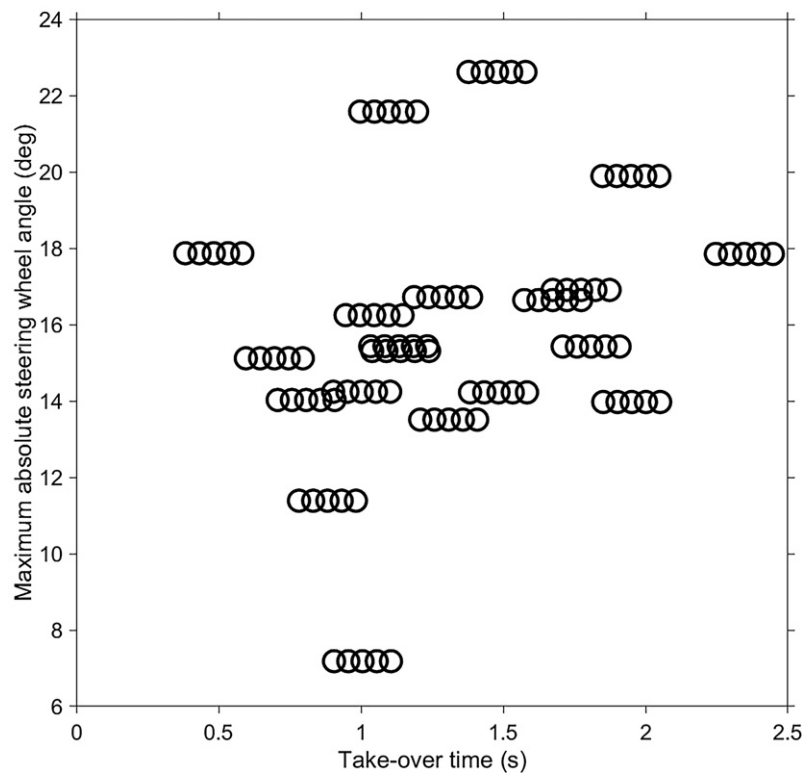


Figure 7 Example (simulated) data for an experiment on automation take-overs. The horizontal axis shows the take-over time, and the vertical axis shows the maximum absolute steering wheel angle. The data are the same as in Fig. 6, except that each value is now repeated five times. A small variation in take-over time was added (within the range of -0.1 s to 0.1 s) to assure that the markers are not overlapping.

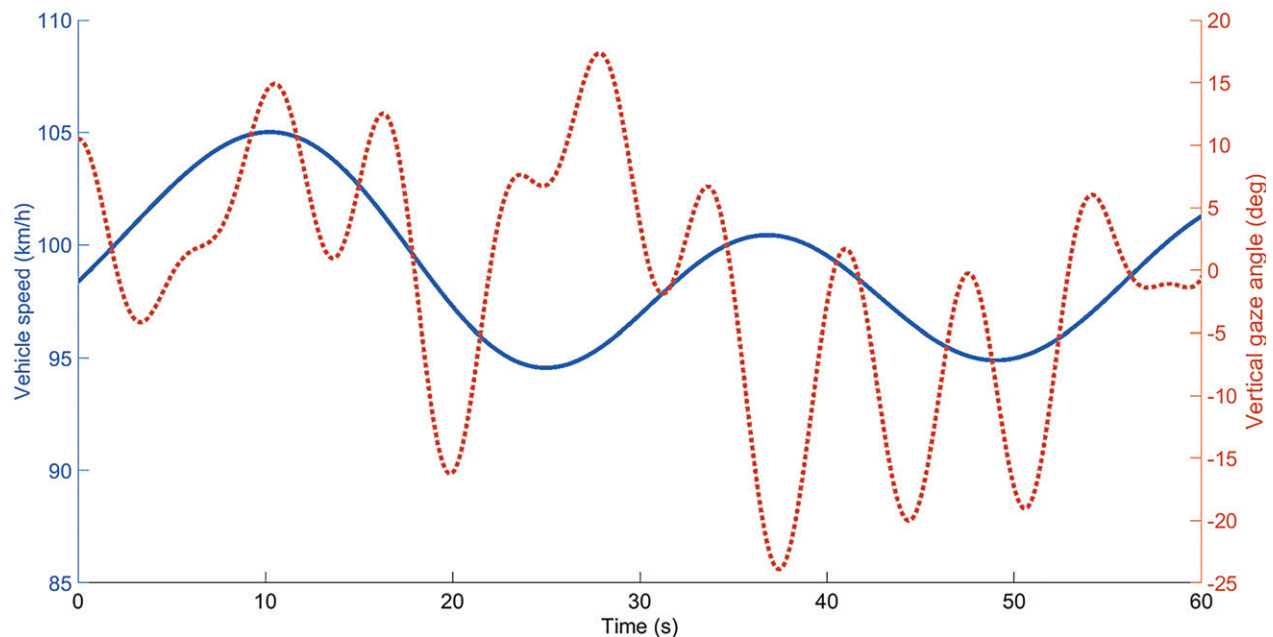


Figure 8 Example (simulated) data for vehicle speed and the driver's vertical eye movements. These data were generated using a multi-sine approach and a random number generator.

average of the five trials. The correlation coefficient, as depicted in Fig. 7, is the same as before ($r = 0.27$), but the P -value is clearly lower ($P = 0.007$). The reason for the reduced P -value is that repeated samples are incorrectly treated as independent. In other words, the researcher has artificially multiplied the sample size by a factor 5.

Violation of independence can come in different guises. It occurs not only in the computation of P -values for correlation coefficients but also in other types of statistical tests (e.g., ANOVAs) and other data types, such as time series. Assume that a researcher computes the relationship between two signals. For example, the traffic psychologist may be interested in whether drivers are more likely to look at the road instead of the dashboard when the vehicle speed is higher. Accordingly, the researcher computes the correlation coefficient between vehicle speed in km/h and the vertical gaze angle in degrees, as measured with an eye-tracker in the car. In the example shown in Fig. 8, which involves simulated data, a correlation coefficient of 0.22 is observed. It often happens that the researcher claims that such a correlation is statistically significant from 0, by using all sampling instances as inputs. For example, the data shown in Fig. 8 have been recorded at 50 Hz, so 1 min of data involves 3000 samples, and the corresponding P -value is 8.37×10^{-34} . However, this P -value is clearly misleading and incorrect, as the adjacent sample points are correlated. In fact, the two signals shown in Fig. 8 were generated using a random number generator and are independent.

In summary, researchers in traffic psychology should ensure that the participants are the unit of analysis, not the repeated measurements of the same participants. It is also possible to treat multiple independent trials as the unit of analysis, an approach which tends to be taken in psychophysics (Smith and Little, 2018).

Conclusion

This article highlighted a number of common pitfalls in the use of statistics in traffic psychology. The take-home message of this work is that a single statistically significant P -value cannot be used to disprove one's null hypothesis, especially when the P -value is barely significant (e.g., $P = 0.04$) or when it is a surprising discovery. Furthermore, the article pointed out that effect sizes should always be considered and that decision-making should not be based on P -values only: sometimes even extremely small and practically insignificant effects are statistically significantly different from zero, or strong effects may exist which are not particularly interesting because "everything is correlated". We also showed that two-stage statistical procedures are inefficient and that data peeking increases the prevalence of false positives. Fourth, and finally, we highlighted the issue of violation of independence, which we believe is a common problem in the field.

The current article is far from complete, and there are a variety of other pitfalls that can be acknowledged. Other common mistakes are (1) the usage of the Kaiser criterion (e.g., counting the number of eigenvalues greater than 1 for deciding on the number of factors to retain in factor analysis), (2) confusing standardized and nonstandardized coefficients in a regression analysis, and (3) multicollinearity in the regression model (e.g., inputting strongly correlated variables, such as driver age and driver experience in years).

We hope that the present article can be of use to the applied researcher. Recognition of the presented pitfalls may improve the robustness of published findings in the area of traffic psychology.

Supplementary Materials

A MATLAB script that reproduces the analyses, figures, and tables is available at <https://doi.org/10.4121/13141289>

References

- Af Wählberg, A., 2017. Driver Behaviour and Accident Research Methodology: Unresolved Problems. CRC Press, Taylor and Francis Group, Boca Raton, FL.
- Af Wählberg, A.E., 2010. Social desirability effects in driver behavior inventories. *J. Safety Res.* 41, 99–106, doi:10.1016/j.jsr. 2010.02.005.
- Anscombe, F.J., 1960. Rejection of outliers. *Technometrics* 2, 123–146, doi:10.2307/1266540.
- Antonakis, J., Dietz, J., 2011. Looking for validity or testing it? The perils of stepwise regression, extreme-scores analysis, heteroscedasticity, and measurement error. *Pers. Individ. Differ.* 50, 409–415, doi:10.1016/j.paid.2010.09.014.
- Bakker, M., Wicherts, J.M., 2014. Outlier removal and the relation with reporting errors and quality of psychological research. *PLOS ONE* 9, e103360, doi:10.1371/journal.pone.0103360.
- Bartram, D., 2007. Increasing validity with forced-choice criterion measurement formats. *Int. J. Sel. Assess.* 15, 263–272, doi:10.1111/j.1468-2389.2007.00386.x.
- Bem, D.J., 2011. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *J. Pers. Soc. Psychol.* 100, 407–425, doi:10.1037/a0021524.
- Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Cesarini, D., 2018. Redefine statistical significance. *Nat. Hum. Behav.* 2, 6–10, doi:10.1038/s41562-017-0189-z.
- Brown, A., Maydeu-Olivares, A., 2011. Item response modeling of forced-choice questionnaires. *Educ. Psychol. Meas.* 71, 460–502, doi:10.1177/0013164410375112.
- Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376, doi:10.1038/nrn3475.
- Chavalarias, D., Wallach, J.D., Li, A.H.T., Ioannidis, J.P., 2016. Evolution of reporting P values in the biomedical literature 1990–2015. *JAMA* 315, 1141–1148, doi:10.1001/jama.2016.1952.
- De Winter, J.C.F., Dodou, D., 2015. A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ* 3, e733, 10.7717/peerj.733.
- De Winter, J.C.F., Dodou, D., 2017. In: Human subject research for engineers. Cham: SpringerBriefs in Applied Sciences and Technology. Springer.
- De Winter, J.C.F., Dodou, D., Stanton, N.A., 2015. A quarter of a century of the DBQ: some supplementary notes on its validity with regard to accidents. *Ergonomics* 58, 1745–1769, doi:10.1080/00140139.2015.1030460.
- De Winter, J.C.F., Kováčová, N., Hagenzieker, M., 2019. Cycling Skill Inventory: assessment of motor-tactical skills and safety motives. *Traffic Inj. Prev.* 20, 3–9, doi:10.1080/15389588.2019.1639158.
- Evans, L., 2009. Reason, replication, and other scientific basics need increased emphasis to better advance safety knowledge. In: SWOV and Friends Symposium, The Hague, the Netherlands.
- Factor, R., 2014. The effect of traffic tickets on road traffic crashes. *Accid Anal. Prev.* 64, 86–91, doi:10.1016/j.aap.2013.11.010.
- Field, A., 2013. Discovering Statistics using IBM SPSS Statistics, 4th ed. Sage Publications.
- Hartgerink, C.H., Van Aert, R.C., Nuijten, M.B., Wicherts, J.M., Van Assen, M.A., 2016. Distributions of p-values smaller than .05 in psychology: what is going on? *PeerJ* 4, e1935, doi:10.7717/peerj.1935.
- Horn, J.L., McArdle, J.J., 2007. Understanding human intelligence since Spearman. In: Cudeck, R., MacCallum, R.C. (Eds.), *Factor Analysis at 100. Historical Developments and Future Directions*. Lawrence Erlbaum Associates, pp. 205–247.
- Ioannidis, J.P., 2005. Why most published research findings are false. *PLOS Med.* 2, e124, doi:10.1371/journal.pmed.0020124.
- Körber, M., Radlmayr, J., Bengler, K., 2016. Bayesian highest density intervals of take-over times for highly automated driving in different traffic densities. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60, 2009–2013, doi:10.1177/1541931213601457.
- Lakens, D., 2015a. Always use Welch's t-test instead of Student's t-test [blog]. Retrieved from <http://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html>.
- Lakens, D., 2015b. Comment: What p-hacking really looks like: A comment on Masicampo and Lalande (2012). *Q. J. Exp. Psycho.* 68, 829–832, 10.1080%2F17470218.2014.982664.
- Lakens, D., Adolfs, F.G., Albers, C.J., Anvari, F., Apps, M.A., Argamon, S.E., Buchanan, E.M., 2018. Justify your alpha. *Nat. Hum. Behav.* 2, 168–171, doi:10.1038/s41562-018-0311-x.
- Lord, F.M., 1953. On the statistical treatment of football numbers. *Am. Psychol.* 8, 750–751, doi:10.1037/h0063675.
- Mattsson, M., 2012. Investigating the factorial invariance of the 28-item DBQ across genders and age groups: an Exploratory Structural Equation Modeling Study. *Accid. Anal. Prev.* 48, 379–396, doi:10.1016/j.aap.2012.02.009.
- Meehl, P.E., 1990. Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant it. *Psychol. Inq.* 1, 108–141, doi:10.1207/s15327965pli0102_1.
- Nederhof, A.J., 1985. Methods of coping with social desirability bias: a review. *Eur. J. Soc. Psychol.* 15, 263–280, doi:10.1002/ejsp.2420150303.
- Nelson, L.D., Simmons, J., Simonsohn, U., 2018. Psychology's renaissance. *Annu. Rev. Psychol.* 69, 511–534, doi:10.1146/annurev-psych-122216-011836.
- Nordhoff, S., De Winter, J., Madigan, R., Merat, N., Van Arem, B., Happee, R., 2018. User acceptance of automated shuttles in Berlin-Schöneberg: A questionnaire study. *Transp. Res. Part F: Traffic Psychol. Behav.* 58, 843–854, doi:10.1016/j.trf.2018.06.024.
- Nosek, B.A., Ebersole, C.R., DeHaven, A.C., Mellor, D.T., 2018. The preregistration revolution. *Proc. Natl. Acad. Sci.* 115, 2600–2606, doi:10.1073/pnas.1708274114.
- Nunnally, J., 1960. The place of statistics in psychology. *Educ. Psychol. Meas.* 20, 64–650, doi:10.1177/001316446002000401.
- Rasch, D., Kubinger, K.D., Moder, K., 2011. The two-sample t test: pre-testing its assumptions does not pay off. *Stat. Pap.* 52, 219–231, doi:10.1007/s00362-009-0224-x.
- Smith, P.L., Little, D.R., 2018. Small is beautiful: In defense of the small-N design. *Psychon. Bull. Rev.* 25, 2083–2101, doi:10.3758/s13423-018-1451-8.
- Standing, L., Sproule, R., Khouzam, N., 1991. Empirical statistics: IV Illustrating Meehl's sixth law of soft psychology: everything correlates with everything. *Psycho. Rep.* 69, 123–126, doi:10.2466/pr0.1991.69.1.123.
- Thompson, B., 1989. Why won't stepwise methods die? *Meas. Eval. Couns. Dev.* 21, 146–148, doi:10.1080/07481756.1989.12022899.
- Wagenmakers, E.J., Wetzel, R., Borsboom, D., Van Der Maas, H.L., 2011. Why psychologists must change the way they analyse their data: the case of psi: comment on Bem (2011). *J. Pers. Soc. Psychol.* 100, 426–432, doi:10.1037/a0022790.
- Ware, J.J., Munafò, M.R., 2015. Significance chasing in research practice: causes, consequences and possible solutions. *Addiction* 110, 4–8, doi:10.1111/add.12673.
- Wells, P., Tong, S., Sexton, B., Grayson, G., Jones, E., 2008. Cohort II: A study of learner and new drivers. Volume 1 - Main report (Report No. 81). Department for Transport, London.
- Zhang, B., De Winter, J., Varotto, S., Happee, R., Martens, M., 2019. Determinants of take-over time from automated driving: A meta-analysis of 129 studies. *Transp. Res. Part F: Traffic Psychol. Behav.* 64, 285–307, doi:10.1016/j.trf.2019.04.020.
- Zimmerman, D.W., 2011. A simple and effective decision rule for choosing a significance test to protect against non-normality. *Br. J. Math. Stat. Psychol.* 64, 388–409, doi:10.1348/000711010X524739.

Data Analysis: Structural Equation Models

Marco Diana, Politecnico di Torino – DIATI, Torino, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction and Illustrative Examples	96
SEM Ambits of Use	97
SEM Use in Travel Behavior Research	98
Future SEM Prospects	99
See Also	99
References	100
Further Reading	101

Introduction and Illustrative Examples

Structural equation models (SEMs) include a broad range of data analysis techniques that are commonly used in social sciences to unravel complex patterns of (inter)relationships among different variables. Two main kinds of techniques can be seen as specific instances of a SEM:

- path analysis, that is in turn a generalization of multiple regression analysis, in which several different dependencies among different variables are jointly considered;
- confirmatory factor analysis, that can be applied to identify the underlying latent constructs of a larger set of manifest variables, that is, variables that can be directly measured through some experimental activity.

SEM in their most general form can therefore contain both (1) a structural model, that is describing the set of dependence relationships that the researcher is willing to study among the considered variables, and (2) a measurement model, aimed at capturing the relationships between a set of manifest variables, or indicators, with their corresponding latent variables, or unobserved constructs.

Fig. 1 shows an example of SEM in travel behavior research, where only a structural model is needed since all variables are manifest (Kroesen, 2017). No measurement model is then specified in this instance. The goal is to simultaneously model the levels of use of five different transport modes, in terms of total distance traveled in a day. Eight exogenous socioeconomic variables are considered, along with an additional set of three mediating variables that are affected by the characteristics of the subject on one hand, and that are in turn influencing the final outcome on the other. This multilevel model structure is one of the attracting features of SEM, since it allows for a more realistic representation of the reality, which seldom reflects a neat distinction between endogenous and exogenous variables. In this example, the total effects of a given socioeconomic characteristic on modal usages are in fact obtained by compounding the direct effects (8×5 relationships) plus the indirect effects through the mediating variables, thus offering a clearer insight of the phenomenon under consideration. Considering a different context, it has also to be noted that a multistage structure is an asset when dealing with panel data or dynamic models.

The conceptual model of **Fig. 1** shows that all possible structural relationships represented by arrows were estimated, but it is also possible to specify a more parsimonious model, where only a subset of these are considered. Within a SEM, it is also possible to estimate correlations among endogenous variables rather than structural relationships, or nonrecursive models where directions of causation are not unique and the existence of “causal loops” is tested, provided that the model itself is identified. Expanding the above example, a researcher might postulate that the levels of use of cars or public transport might in turn have an influence on the ownership of bikes and e-bikes.

A complementary example of SEM is provided in **Fig. 2**, this time considering a measurement model that was proposed to quantitatively define the concept of primary, or intrinsic utility of traveling (Diana, 2008). Following standard conventions, latent constructs are represented by ovals and manifest indicators by rectangles. No exogenous variables or dependence patterns among variables are considered here: arrows are linking the different latent variables with their respective indicators. Given the absence of structural relationships, this model can be seen as a confirmatory factor analysis that is empirically testing a theory, aimed at showing how the primary trip utility can be indirectly observed through a set of indicators. These latter are presented in the lower part of the figure and consist in a set of 20 items $y_1, \dots, y_{20}$ measured in a survey through semantic scales, therefore dealing with ordinal rather than metric variables.

Such manifest variables are not directly related to the primary utility construct, but they rather load on six different latent dimensions or aspects that can help in better understanding the phenomenon under consideration. In turn, the latter latent dimensions load on the second-order factor that is capturing the primary utility. Again, the possibility of defining multiple stages not only in structural but also in measurement models is one key feature of the method. The figure is showing with Greek letters the model parameters that were estimated, namely the first- and second-order factor loadings (respectively λ and γ), the variance of the first- and second-order measurement errors (respectively ϵ and ζ). Additionally, a measurement error

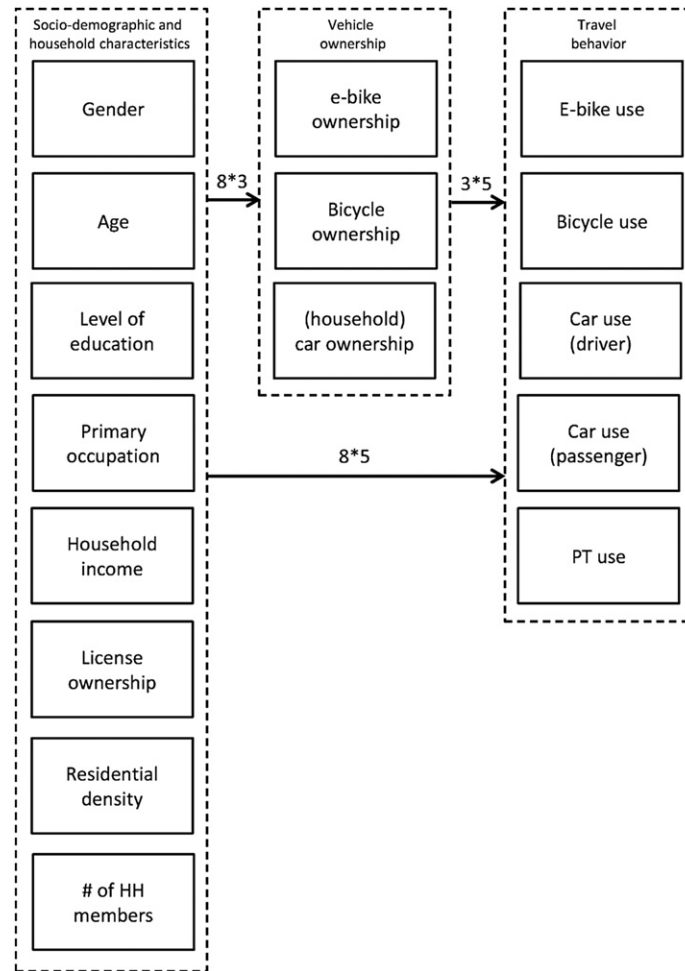


Figure 1 An example of SEM involving a path diagram. Source: Kroesen, 2017.

covariance between two manifest items was also estimated, but of course additional ones could have been included as well based on theory guidance.

As already mentioned, more general instances than the two above examples are commonly studied, where some of the variables in the structural model (either endogenous or exogenous) are latent and both structural and measurement models are therefore jointly estimated.

The SEM concept has been developed from the 1970s and it is nowadays one of the most widely used methods in social sciences. The present contribution will not deal with technical issues such as SEM statistical assumptions, model specification, estimation, fit assessment and validation, since they are widely and thoroughly covered by existing manuals, at different levels of technical discussion and for all kinds of readership. The “Further readings” section presents an essential selection of these sources, including texts with different technical levels. Besides, many different SEM softwares have been developed in recent years and they generally reached a maturity stage that allow the analyst to enjoy a wide range of functionalities in a relatively user-friendly environment. Examples include programs such as *EQS*, *LISREL*, and *Mplus*, but also dedicated applications or modules of statistical analysis softwares such as *CALIS* for the SAS environment, *Amos* for SPSS, *SEPATH* for STATISTICA and the *lavaan* and *sem* packages for the open source software R.

In the following, the intention is rather to offer some guidance on when developing a SEM might be good option in the transport research domain, in view of the current state of the art and available alternatives. The related discussion is developed in the following subhead. To better illustrate this point, some examples of recent SEM applications dealing with different topics within the travel behavior research sector will then be presented. The chapter will then end with some speculation on the potential of this method to serve the transportation research community in the (near) future.

SEM Ambits of Use

The two previous examples have shown how the flexibility of SEM can be exploited in a wide range of situations. Yet such flexibility is also inherently bringing some limitations in its best ambits of use. In particular, SEM has no underpinning theory, unlike e.g. discrete

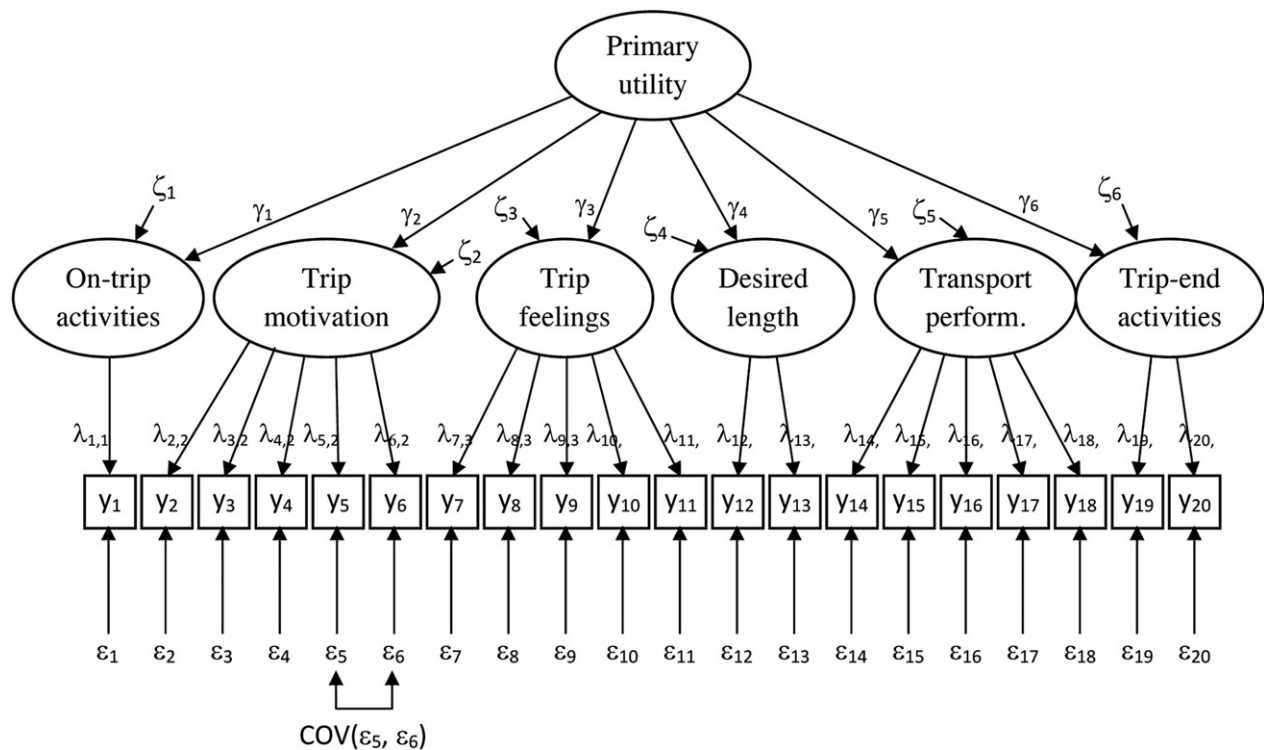


Figure 2 An example of confirmatory factor analysis through SEM. Source: [Diana, 2008](#).

choice models, which are based on the consumer theory. The researcher is left with the responsibility of adopting or developing a sufficiently robust theoretical background before specifying a SEM, thus making SEM a confirmatory rather than an exploratory research technique. Conversely, the lack of such framework should not lead to considering the flexibility of the tool simply as a convenient way to build a model that fits at best the available dataset. This is perhaps the single most common misuse of SEM that is encountered in the literature.

Traditionally, SEM has been seen as a technique that can only handle continuous endogenous variables. However, estimation methods are also available whenever the model outcomes are rather ordinal or binary, albeit quite larger sample size are required in such cases. On the other hand, SEM cannot deal with dependent variables having more than two categories, that can be rather analyzed through multivariate logistic regression or its counterpart in the econometric domain, the multivariate logit model. Concerning the latter, the relatively recent rise of the Integrated Choice and Latent Variables (ICLV) models that can jointly estimate structural relationships and factor loadings of latent variables gives the analyst a valuable alternative to SEM whenever (1) the outcome variable is categorical and (2) the theoretical background of the consumer theory is appropriate for the problem under consideration.

In common with other techniques making use of latent variables, SEM has been criticized for being a descriptive method that is of little use for forecasting purposes. Indeed, SEM can give additional insights that can be later used to better specify forecasting models, while avoiding or at least properly considering common specification errors or related issues such as multicollinearity, ecological fallacy or the confusion between correlation and causation, just to name a few.

SEM Use in Travel Behavior Research

SEM started being consistently used by the transport research community only at the end of the 20th century. The interested reader is referred to [Golob \(2003\)](#) for a review of these early applications and a discussion on the most promising fields where the technique can be applied, which is largely still valid nowadays. In the following, we will build on that work to identify some more recent lines of research in the transportation sector that have been developed by employing SEM techniques, without any claim of being exhaustive given the nature of the present contribution.

Due to the possibility of modeling complex relationships across different variables, one of the earlier SEM applications was related to activity-based models. Many studies resort on the method to clarify the dynamics of time allocation among various activities and the related travel durations and modal choices, considering a range of personal characteristics and inter-personal interactions ([Cheng et al., 2019](#); [Fyhri and Hjorthol, 2009](#); [Sharmeen et al., 2014](#); [Scheiner, 2020](#); [Susilo and Liu, 2016](#); [Wang and Law, 2007](#)). A related field is the study of the intertwined relationship between built environment and travel choices, where SEM can

help in sorting out self selection and simultaneity biases (Mokhtarian and Cao, 2008; Wang and Lin, 2019), along with the mediating role of additional variables such as car ownership (Van Acker and Witlox, 2010).

The link between psychological constructs (such as attitudes, perceptions, satisfaction, and wellbeing) and travel-related intentions and behaviors is another typical field where SEM has been successfully employed (Bamberg et al., 2007; De Oña et al., 2016; Ingvardson and Nielsen, 2019; Lai and Chen, 2011; Lois and López-Sáez, 2009; Kaplan et al., 2019; Kroesen, 2014; Scheiner and Holz-Rau, 2007). SEM seems particularly promising when the focus is on innovative transport services and specific social groups, where choice alternatives might not be well known and different behavioral theories might be appropriate (Bennett et al., 2019; Kroesen, 2017; Mohamed et al., 2016; Neoh et al., 2018; Wang et al., 2019; Will and Schuller, 2016).

Travel safety research can be seen under a similar viewpoint, whenever the goal is to relate risk and safety perceptions or personality traits with behaviors and distraction, mainly in relation with car- and two-wheeled vehicles usage or walking. Indeed, this research domain has possibly witnessed the largest proportion of SEM studies in more recent years (e.g., Chataway et al., 2014; Chen and Donmez, 2016; Dinh et al., 2020; Jovanović et al., 2011; Kummeneje and Rundmo, 2019; Mallia et al., 2015; Susilo et al., 2015; Tazul Islam et al., 2017; Useche et al., 2018; Wong et al., 2018; Zhao et al., 2019). Additional applications include the effects of road and driver characteristics, environmental factors and urban forms on traffic safety (Kamel and Sayed 2019; Lee et al., 2008; Najaf et al., 2018).

Nowadays, travel demand management strategies (such as travel behavior change programs or travel nudging) are often implemented along with more concrete interventions on the transport system offer to maximize the effectiveness of the interventions. Studying how such a mixture of hard and soft (or push and pull) measures interact usually requires putting forward a behavioral theory, often developed by environmental psychologists, that can be empirically tested through SEM (Bamberg et al., 2011).

Future SEM Prospects

The above reviewed state of the art of SEM in the transport sector can give us some hints on how such methods can contribute in tackling future research endeavors in this area.

To date, the common mindset of transport modelers and practitioners is understandably to explain observed mobility behaviors on the basis of different factors, including sociodemographics, psychological constructs and characteristics of the transport system. In turn, decision-making processes are often based on an economic evaluation of different alternatives, such as in cost-benefit analysis, that is largely eased by a consistent theoretical framework that is directly giving key inputs such as value of travel times or elasticity values. Yet travel choices might not be purely endogenous, since they can in turn influence other elements in the life of individuals. Failing to acknowledge such loops might lead to a poor understanding of mobility phenomena and undermine the effectiveness of advice given to policy makers to tackle the complex problems that affect contemporary transport systems.

The above caveat seems particularly critical when dealing with specific issues that are, however, gaining importance in the study of transport systems. One good example is the impact of disruptive technologies, such as the role of ubiquitous travel information in shaping mobility behaviors, driver reactions to improved safety devices or the potential diffusion of the connected, autonomous, shared and electric (CASE) paradigm in road transport (Diana, 2010). Related technology applications call for a behavioral reaction from the traveler side, who might range from an aprioristic refusal to an enthusiastic adoption. Such reactions go much beyond a rational decision-making context, with the added complexity of a high degree of dynamism of the whole process (early adopters versus latecomers) that can negatively affect the reliability of model-based forecasts, especially if based on cross-sectional experiments. In such situations, taking not for granted the usual attitude – behavior direction of causation with the mediating role of the utility construct could lead to better insights.

Moving from research topics to research methods, SEM can be seen as a tool that may work well together with some agnostic, or perhaps even uninformed, research approaches, such as the analysis of big data, that are breaking through the “comfort zone” made of well-established practices and consolidated paradigms of the transportation research community and that are capturing a lot of attention of both decision makers and the general public. Those alternative approaches are often carried out by scholars, such as physicists and computer scientists, that simply “let the data speak”, by applying data mining methods and algorithms that they master in the transportation research field, like they are used to do in any other domain. There is indeed a good potential to strengthen the connections between the “big data” and “small data” research communities dealing with transport issues (Chen et al., 2016), and SEM could be a flexible enough tool to help achieving this objective. Data can indeed speak, but SEM can be very useful in actually understanding what data are telling us.

See Also

Are Travel Choices Derived from the Rationality Assumption?; Human Factors in Transportation; Data Analysis: Reduction; Data Analysis: Latent Variable and Latent Class Models

References

- Bamberg, S., Hunecke, M., Blöbaum, A., 2007. Social context, personal norms and the use of public transportation: two field studies. *J. Environ. Psychol.* 27 (3), 190–203.
- Bamberg, S., Fujii, S., Friman, M., Gärling, T., 2011. Behaviour theory and soft transport policy measures. *Transport Policy* 18 (1), 228–235.
- Bennett, R., Vijaygopal, R., Kottasz, R., 2019. Attitudes towards autonomous vehicles among people with physical disabilities. *Transport. Res. Part A: Policy Pract.* 127, 1–17.
- Chataway, E.S., Kaplan, S., Nielsen, T.A.S., Prato, C.G., 2014. Safety perceptions and reported behavior related to cycling in mixed traffic: a comparison between Brisbane and Copenhagen. *Transport. Res. Part F: Traffic Psychol. Behav.* 23, 32–43.
- Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transport. Res. Part C: Emerging Technol.* 68, 285–299.
- Chen, H.Y.W., Donmez, B., 2016. What drives technology-based distractions? A structural equation model on social-psychological factors of technology-based driver distraction engagement. *Acc. Anal. Prevent.* 91, 166–174.
- Cheng, L., Chen, X., Yang, S., Wu, J., Yang, M., 2019. Structural equation models to analyze activity participation, trip generation, and mode choice of low-income commuters. *Transport. Lett.* 11 (6), 341–349.
- De Oña, J., de Oña, R., Eboli, L., Forciniti, C., Mazzulla, G., 2016. Transit passengers' behavioural intentions: the influence of service quality and customer satisfaction. *Transportmetrica A: Transport Sci.* 12 (5), 385–412.
- Diana, M., 2008. Making the "primary utility of travel" concept operational: a measurement model for the assessment of the intrinsic utility of reported trips. *Transport. Res. Part A: Policy Pract.* 42 (3), 455–474.
- Diana, M., 2010. From mode choice to modal diversion: a new behavioural paradigm and an application to the study of the demand for innovative transport services. *Technol. Forecasting Social Change* 77 (3), 429–441.
- Dinh, D.D., Vu, N.H., McIlroy, R.C., Plant, K.A., Stanton, N.A., 2020. Examining the roles of multidimensional fatalism on traffic safety attitudes and pedestrian behaviour. *Safety Sci.* 124, 104587.
- Fyhri, A., Hjorthol, R., 2009. Children's independent mobility to school, friends and leisure activities. *J. Transport Geogr.* 17 (5), 377–384.
- Golob, T.F., 2003. Structural equation modeling for travel behavior research. *Transport. Res. Part B: Methodol.* 37 (1), 1–25.
- Ingvardson, J.B., Nielsen, O.A., 2019. The relationship between norms, satisfaction and public transport use: A comparison across six European cities using structural equation modelling. *Transport. Res. Part A: Policy Pract.* 126, 37–57.
- Lai, W.T., Chen, C.F., 2011. Behavioral intentions of public transit passengers—The roles of service quality, perceived value, satisfaction and involvement. *Transport Policy* 18 (2), 318–325.
- Jovanović, D., Lipovac, K., Stanojević, P., Stanojević, D., 2011. The effects of personality traits on driving-related anger and aggressive behaviour in traffic among Serbian drivers. *Transport. Res. Part F: Traffic Psychol. Behav.* 14 (1), 43–53.
- Kamel, M.B., Sayed, T., 2019. Accounting for mediation in cyclist-vehicle crash models: a Bayesian mediation analysis approach. *Accident Anal. Prevent.* 131, 122–130.
- Kaplan, S., Wrzesinska, D.K., Prato, C.G., 2019. Psychosocial benefits and positive mood related to habitual bicycle use. *Transport. Res. Part F: Traffic Psychol. Behav.* 64, 342–352.
- Kroesen, M., 2014. Assessing mediators in the relationship between commute time and subjective well-being: structural equation analysis. *Transport. Res. Rec.* 2452 (1), 114–123.
- Kroesen, M., 2017. To what extent do e-bikes substitute travel by other modes? Evidence from the Netherlands. *Transport. Res. Part D: Transport Environ.* 53, 377–387.
- Kummeneje, A.M., Rundmo, T., 2019. Risk perception, worry, and pedestrian behaviour in the Norwegian population. *Accident Anal. Prevent.* 133, 105294.
- Lee, J.Y., Chung, J.H., Son, B., 2008. Analysis of traffic accident size for Korean highway using structural equation models. *Accident Anal. Prevent.* 40 (6), 1955–1963.
- Lois, D., López-Sáez, M., 2009. The relationship between instrumental, symbolic and affective factors as predictors of car use: A structural equation modeling approach. *Transport. Res. Part A: Policy Pract.* 43 (9–10), 790–799.
- Mallia, L., Lazaras, L., Violani, C., Lucidi, F., 2015. Crash risk and aberrant driving behaviors among bus drivers: the role of personality and attitudes towards traffic safety. *Accident Anal. Prevent.* 79, 145–151.
- Mohamed, M., Higgins, C., Ferguson, M., Kanaroglou, P., 2016. Identifying and characterizing potential electric vehicle adopters in Canada: a two-stage modelling approach. *Transport Policy* 52, 100–112.
- Mokhtarian, P.L., Cao, X., 2008. Examining the impacts of residential self-selection on travel behavior: a focus on methodologies. *Transport. Res. Part B: Methodol.* 42 (3), 204–228.
- Najaf, P., Thill, J.C., Zhang, W., Fields, M.G., 2018. City-level urban form and traffic safety: a structural equation modeling analysis of direct and indirect effects. *J. Transport Geogr.* 69, 257–270.
- Neoh, J.G., Chipulu, M., Marshall, A., Tewkesbury, A., 2018. How commuters' motivations to drive relate to propensity to carpool: evidence from the United Kingdom and the United States. *Transport. Res. Part A: Policy Pract.* 110, 128–148.
- Scheiner, J., 2020. Changes in travel mode use over the life course with partner interactions in couple households. *Transport. Res. Part A: Policy Pract.* 132, 791–807.
- Scheiner, J., Holz-Rau, C., 2007. Travel mode choice: affected by objective or subjective determinants? *Transportation* 34 (4), 487–511.
- Sharmeen, F., Arentze, T., Timmermans, H., 2014. An analysis of the dynamics of activity and travel needs in response to social network evolution and life-cycle events: a structural equation model. *Transport. Res. Part A: Policy Pract.* 59, 159–171.
- Susilo, Y.O., Joewono, T.B., Vandebona, U., 2015. Reasons underlying behaviour of motorcyclists disregarding traffic regulations in urban areas of Indonesia. *Accident Anal. Prevent.* 75, 272–284.
- Susilo, Y.O., Liu, C., 2016. The influence of parents' travel patterns, perceptions and residential self-selectivity to their children travel mode shares. *Transportation* 43 (2), 357–378.
- Tazul Islam, M., Thue, L., Grekul, J., 2017. Understanding traffic safety culture: implications for increasing traffic safety. *Transport. Res. Rec.* 2635, 79–89.
- Useche, S.A., Montoro, L., Alonso, F., Tortosa, F.M., 2018. Does gender really matter? A structural equation model to explain risky and positive cycling behaviors. *Accident Anal. Prevent.* 118, 86–95.
- Van Acker, V., Witlox, F., 2010. Car ownership as a mediating variable in car travel behaviour research using a structural equation modelling approach to identify its dual relationship. *J. Transport Geogr.* 18 (1), 65–74.
- Wang, D., Law, F.Y.T., 2007. Impacts of Information and Communication Technologies (ICT) on time use and travel behavior: a structural equations analysis. *Transportation* 34 (4), 513–527.
- Wang, D., Lin, T., 2019. Built environment, travel behavior, and residential self-selection: A study based on panel data from Beijing, China. *Transportation* 46 (1), 51–74.
- Wang, Y., Gu, J., Wang, S., Wang, J., 2019. Understanding consumers' willingness to use ride-sharing services: the roles of perceived value and perceived risk. *Transport. Res. Part C: Emerging Technol.* 105, 504–519.
- Will, C., Schuller, A., 2016. Understanding user acceptance factors of electric vehicle smart charging. *Transport. Res. Part C: Emerging Technol.* 71, 198–214.
- Wong, I.Y., Smith, S.S., Sullivan, K.A., 2018. Validating an older adult driving behaviour model with structural equation modelling and confirmatory factor analysis. *Transport. Res. Part F: Traffic Psychol. Behav.* 59, 495–504.
- Zhao, X., Xu, W., Ma, J., Li, H., Chen, Y., 2019. An analysis of the relationship between driver characteristics and driving safety using structural equation models. *Transport. Res. Part F: Traffic Psychol. Behav.* 62, 529–545.

Further Reading

- Bollen, K.A., 1989. Structural equations with latent variables. Wiley, ISBN 0-471-01171-1.
- Kline, R.B., 2016. Principles and practice of structural equation modelling, Fourth edition. The Guilford Press ISBN 9781462523344
- Tarka, P., 2018. An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality Quantity* 52 (1), 313–354.
- Ullman, J.B., Bentler, P.M., 2012. Structural equation modelling. In: Weiner, I.B. (Ed.), *Handbook of Psychology*, Second edition, Volume 2. Wiley, pp. 661–690, doi:10.1002/9781118133880.hop202023.

Data Analysis: Integrated Choice and Latent Variable Models

Carlo G. Prato, The University of Queensland, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	102
Model Framework	103
Model Applications	104
Conclusions and Future Prospects	104
See Also	105
References	105
Further Reading	106

Introduction

The study of human behavior has always attracted the interest of economists, engineers, psychologists, planners, and marketing experts who have attempted to explain the past behavior via a behavioral analysis approach and predict future behavior via a predictive modeling approach. Discrete choice models are the foundation of the study of human behavior from both a behavioral perspective, where the analysis focuses on revealing features and anomalies of choice behavior, and a predictive perspective, where the analysis concentrates on emphasizing the regularities of choice behavior (Ben-Akiva et al., 2002).

Travel behavioral modelers have traditionally used discrete choice models for the purpose of explaining the past behavior and predicting the future behavior of travelers while relying on basic concepts from the neo-classical economic theory. Specifically, choice models are based on the concept of utility, a latent construct that captures the value that a decision-maker associates with the available alternatives, and a maximization principle, an abstraction of the decision-maker as a rational actor who chooses the alternative with the highest utility in the attempt to maximize their personal wellbeing. Utility is typically mapped as a function of observed characteristics of the decision-maker (e.g., gender, age, income) and the alternatives (e.g., travel time, public transport fare, tolls), and assumptions about the stochastic component of the function allows for the calculation of the probability of choosing an alternative among all the ones that are available to the decision-maker. However, it is plausible to think that choice behavior is not only determined from observed characteristics of the decision-maker and the alternatives that are easy to measure, but also by information processing, perceptions, attitudes, beliefs, context, external constraints as individual preferences are formed because of psychological and sociological reasons (McFadden, 1986). As recognized by psychologists, who have a greater interest in the behavioral analysis when compared to economists and engineers who have a greater interest in the predictive analysis, latent constructs such as attitudes, perceptions, norms, affects, and beliefs most likely influence the choice behavior of decision-makers (Bamberg and Schmidt, 2001; Gärling et al., 2003). Accordingly, models are needed to integrate these unobservable constructs within the mapping of the utility if we were to capture the determinants of choice behavior.

Three approaches were initially proposed in the literature to recognize the importance of unobserved characteristics of the decision-maker (e.g., attitudes, perceptions, norms) in the explanation of their choice behavior (Bhat and Dubey, 2014). The first approach used parametric or nonparametric random distributions to account for the intrinsic preference for alternatives and the sensitivity to attributes of the alternatives to vary across decision-makers (e.g., Bhat, 1998; Revelt and Train, 1996). However, the model estimation was econometrically inconsistent because the attitudes and perception might have been correlated with the observed variables that were used in the utility functions to explain the choice behavior. The second approach utilized indicators of attitudes as explanatory variables in the mapping of the utility (e.g., Bhat and Koppelman, 1993; Koppelman and Hauser, 1978). However, the model estimation was econometrically inconsistent because the indicators might have not been representing directly the underlying attitudes that are related to the choice and consequently have been generating a measurement error of the latent variables. The third approach applied factor analysis to reduce the number of indicators into a manageable number of latent variables that are "error-free" explanatory variables. However, the model estimation was econometrically inconsistent because the latent variables might have been specific to individual alternatives or might have interacted with variables that vary across alternatives.

More recently, the Integrated Choice and Latent Variable (ICLV) model (see Bolduc et al., 2005; Walker and Ben-Akiva, 2002) was proposed as an econometrically consistent answer to the need of incorporating attitudes, perceptions, norms, affects and beliefs into discrete choice models. The objective of the model was to gain a broader and deeper understanding of the choice behavior of decision-makers by combining observed variables with psychometric measures associated with individual attitudes, perceptions, etc. Advancements in behavioral and social sciences made the task easier thanks to the progress in the development of survey instruments (e.g., batteries of Likert scale items for rating the level of agreement or disagreement on a symmetric scale), which allowed to measure a large variety of latent variables that could relate to choice behavior, as well as the elaboration of decision-making theories, which provided a theoretical support to the study of choice behavior (see, e.g., Ajzen, 1991; Homer and

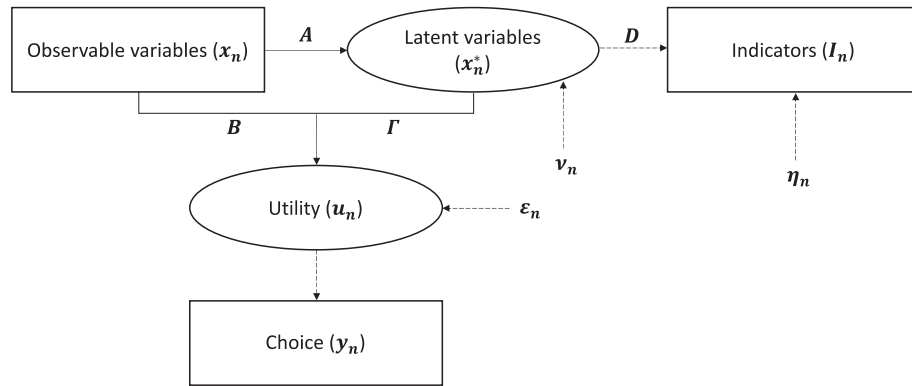


Figure 1 The ICLV model framework. Source: Adapted from Vij and Walker (2016).

Kahle, 1988; Triandis, 1977). The testing of both theories and scales relies often on variations of Structural Equations Modeling (SEM), which is suitable for capturing causal relations between a variety of observed and latent variables, but is less seemly for explaining choice behavior as it lacks the depth in that regard of the utility maximization framework at the foundation of discrete choice modeling. The ICLV model closes the methodological gap by combining the latent variable modeling used by behavioral and social scientists with the discrete choice modeling utilized by econometricians, hence removing all the restrictions related to the assumptions required for SEM estimation.

Given the purpose of gaining a broader and deeper understanding of choice behavior, the ICLV model is very important from the behavioral perspective as it provides the opportunity for recognizing that the role of latent variables in explaining travel behavior may be just as relevant as the one of observed variables that have been traditionally used in utility functions for a long time. However, questions exist about the effectiveness of the ICLV model from the predictive perspective, as studies exist that support the assumption that it enhances not only the explanatory power but also the predictive ability of models (Bolduc and Daziano, 2010; Temme et al., 2008), but also studies exist that express concerns about whether the framework is beneficial to predict behavior (Chorus and Kroesen, 2014; Vij and Walker, 2016). As basic questions remain unanswered about whether an ICLV model fits the data better than a model without latent variables, or whether estimates from an ICLV model help answer policy questions that are already answered to by a model without latent variables, the remainder of this chapter presents the framework in its conceptual and mathematical formulation, discusses the range of application, and considers the future of ICLV models.

Model Framework

The ICLV model framework is presented in Fig. 1. The model for the latent constructs of attitudes, perceptions, norms, affects, and/or beliefs has two components: (1) measurement equations capture the relations between indicator variables and latent constructs, and (2) structural equations express the relations between observed exogenous variables and latent constructs as well as between the latent constructs themselves. The latent variables are assumed to be measured by multiple indicators that can be observed on a continuous, ordinal, or nominal scale, and in the latter cases the measurement equations map the continuous latent constructs to the ordinal or nominal scale of the indicator variables (e.g., on Likert-type scales). The model for the choice behavior of the decision-maker has also two components: (1) measurement equations indicate which alternative is chosen among the available ones that the decision-maker is aware of, and (2) structural equations map the utility of the alternatives as functions of both observed and latent variables. The choice is assumed to be consistent with the paradigm (e.g., utility maximization, regret minimization) that the decision-maker follows to choose the preferred alternative (e.g., the one with associated the highest utility, the one with associated the minimum regret).

Mathematically, four sets of equations represent the described components of the ICLV model (Vij and Walker, 2016):

$$I_n = Dx_n + \Gamma x_n^* + \eta_n \quad (1)$$

$$x_n^* = Ax_n + \nu_n \quad (2)$$

$$y_{ni} = \{1 \text{ if } u_{ni} \geq u_{nj} \text{ for } j \in \{j \dots J\}, 0 \text{ otherwise}\} \quad (3)$$

where x_n is the $(K \times 1)$ vector of observed variables, x_n^* is the $(M \times 1)$ vector of latent variables, I_n is the $(R \times 1)$ vector of indicators to measure the latent variables, u_n is the $(J \times 1)$ vector of utilities of each of the alternatives as perceived by decision-maker n , D is the $(J \times M)$ matrix of parameters and η_n is the $(R \times 1)$ vector of error terms of the measurement equations (1) of the latent variable model, A is the $(M \times K)$ matrix of parameters and ν_n is the $(M \times 1)$ vector of error terms of the structural equations (2) of the latent variable model, y_{ni} is the choice indicator equal to 1 if n chooses alternative i in the measurement equations (3) of the choice model,

and $\mathbf{B}$ and $\mathbf{\Gamma}$ are the $(J \times K)$ and $(J \times M)$ matrices of parameters and ε_n is the $(J \times 1)$ vector of error terms of the utility functions (4) of the choice model.

While the framework is conceptually easy to comprehend and its mathematical representation is econometrically consistent, the use of ICLV model is not straightforward for three reasons (Bhat and Dubey, 2014): (1) restrictions in the specifications used in applications; (2) difficulties in the estimation of the models; (3) computational requirements of the models. Initial applications of the ICLV model assumed that there was no correlation between the error terms of the structural equations of the latent variable model (i.e., the elements of vector ν_n) as well as between the error terms of the utilities of the alternatives in the choice model (i.e., the elements of vector ε_n). However, these assumptions were restrictive as there might have been unobserved factors that impact the formation of multiple attitudes for the same decision-maker or capture correlation for different alternatives (Temme et al., 2008). Also, most applications of the ICLV model used simulated maximum likelihood for model estimation, similarly to mixed logit models. However, the demands of integrating over a mixture of probabilities routinely causes convergence problems and increases computational times. Accordingly, more recently, a Bayesian Markov Chain Monte Carlo simulation approach (Daziano and Bolduc, 2013a) and a maximum approximate composite marginal likelihood (MACML) inference approach (Bhat and Dubey, 2014) were proposed to simplify the estimation of ICLV models and reduce the dimensionality of the integral.

Relevantly, significant advancements in optimization techniques and conspicuous improvements in computational power have contributed to the increase in the number of applications of the ICLV model. Even more relevantly, the availability of estimation software such as Biogeme (Bierlaire, 2020), Mplus (Muthén and Muthén, 2005) and more recently Apollo (Hess and Palma, 2019) have created a level of accessibility that is crucial for the progress in the science at the foundation of ICLV models, whether it is from the methodological or the application perspective. De facto, the leaps and bounds in understanding human behavior in the travel demand context would have not been possible without these substantial contributions.

Model Applications

Application of the ICLV model to travel behavior and travel demand analysis has focused mostly on the role of attitudes and perceptions in choice behavior, and more marginally on other latent constructs such as norms, beliefs and socialization. The present list does not intend to be exhaustive, but to be indicative of the directions that research prevalently move toward. Most of the interest in the role of attitudes and perceptions is likely driven by the interest in understanding better determinants of choice of sustainable transport solutions. Several studies have investigated the effect of environmental concerns (e.g., Daziano and Bolduc, 2013b; Hess et al., 2013), attitudes toward vehicle features (e.g., design, spaciousness, technology, etc.) attitudes toward vehicle leasing (e.g., Glerum et al., 2013; Kim et al., 2017) and heterogeneity in the population for preferences toward electromobility (Bahamonde-Birke et al., 2017; Kim et al., 2014) on the decision-making process of consumers wanting to buy alternative fuel vehicles. Some studies focused on the propensity toward car sharing being related to environmental concerns and attitudes toward sharing from a general perspective (Efthymiou and Antoniou, 2016; Kim et al., 2017), while other studies have looked at attitudes toward the environment to explain the choice of active travel (Kamargianni and Polydoropoulou, 2013; Maldonado-Hinarejos et al., 2014) or more in general the choice of travel mode (Paulssen et al., 2014; Vij et al., 2013). The versatility of the ICLV model framework has led also to its application to alternative contexts such as security on rail (Daly et al., 2012), car route choices (Prato et al., 2012), freight decision-making (Bergantino et al., 2013), and bicycle route choices (Bhat et al., 2015).

Other studies have investigated latent variables beyond attitudes and perceptions. Personality traits, perceptual factors and travel attitudes were revealed to be latent variables that affect travel mode choice not only directly, but also indirectly via a mediation effect of some latent variables on others. The effect of affective factors and habit on travel mode choice was investigated according to the Theory of Reasoned Action (Tudela et al., 2011), while inertia was measured as a latent habitual effect in travel mode choice (Cherchi et al., 2014). The effect of perceived mobility necessities was considered within the context of departure time choice (Thorhaug et al., 2019), and a hierarchical model was proposed where “values” at the lower level of the psychological construct affected attitudes at the higher level of the same construct, which in turn had an effect on travel mode choices (Paulssen et al., 2014). A similar hierarchical model was used to capture the effect of social interaction on the formation of psychological factors and the decision-making process (Kamargianni and Polydoropoulou, 2013). Even with the significant advancements from the methodological and computational perspectives, and the notable increase in the number of studies in the last decade, the possibilities for the application of the ICLV model to the understanding of travel behavior and the analysis of travel demand are still substantial.

It should be noted that important research directions for the application of ICLV models are the investigation of whether the models introduce benefits when compared with a more traditional choice model without latent variables, especially when considering their effectiveness in drawing policy conclusions (Vij and Walker, 2016), as well as the analysis of the efficiency of estimation methods that would make the model even more accessible than it already is.

Conclusions and Future Prospects

As mentioned in the introduction, while it is obvious that the ICLV model provides substantial contribution to the understanding of human behavior from the behavioral analysis perspective, recent research has posed the question of the actual benefit of using ICLV models from the predictive analysis perspective, and in particular the effectiveness of using these models to draw policies (see, e.g.,

Chorus and Kroesen, 2014; Vij and Walker, 2016). Recent research has also showed that a reduced form model without latent variables fits the choice data at least comparably with respect to the original ICLV model from which it originated, and has concluded that it is important to recognize this truth not to diminish the benefits that are related to the use of the model (Vij and Walker, 2016).

It is however important not to discard the significant advantages of the application of ICLV models to understand human behavior for travel demand analysis, and also not to ignore some necessary considerations for their applications moving forward (Vij and Walker, 2016). Firstly, the comprehension of the behavior benefits substantially from the estimation of ICLV models because it allows for testing complex behavioral theories and imparting meaning to underlying sources of heterogeneity thanks to the identification of relations between variables. Choice models without latent variables would not be able to identify otherwise those relations on the basis of observed characteristics of decision-maker and alternatives. Secondly, the estimation of ICLV models can help correct for omitted variable bias and measurement error when relaxing the assumption of causality that would make the estimates from the ICLV model and its reduced form asymptotically equal. The omitted variable bias and in particular the measurement errors are related to the estimation of models in the early days when psychological and social factors were considered in choice models. Thirdly, the estimation of ICLV models can help the analyst to improve the ability to predict outcomes from the same models because of the reduction of the variance of the model estimates that increases the precision of policy outputs such as marginal rates of substitution. Large errors are in fact associated with the calculation of the latter, especially when computing the ratios between parameters that are distributed across the population. Fourthly, the estimation of ICLV models opens the possibility for the analyst to measure the impact of latent variables on behavior via measures such as demand elasticities and willingness to pay, predict behavioral change as a function of changes in the latent variables over time, and support the design of policies for behavioral change in ways that would simply not possible with choice models without latent variables.

Overall, there are possible guidelines to the process of evaluating that studies bring to the table tangible gains when estimating an ICLV model rather than its reduced form (Vij and Walker, 2016). Relevantly, future research should consider that the ICLV model is a powerful tool in the context of an increasing number of tools available to modelers. As long as analysts are conscientious in terms of the data that they collect, thoughtful in terms of the design of the model framework, and creative in terms of clarifying the value of the ICLV model to practitioners and policy-makers, it can be expected that the model will continue to flourish and eventually provide a substantial contribution also from the predictive perspective.

See Also

Are Travel Choices Derived from the Rationality Assumption?; Human Factors in Transportation; Data Analysis: Reduction; Data Analysis: Structural Equation Models

References

- Ajzen, I., 1991. The theory of planned behavior. *Organ. Behav. Hum. Decis.* 50 (2), 179–211.
- Bahamonde-Birke, F.J., Kunert, U., Link, H., de Dios Ortúzar, J., 2017. About attitudes and perceptions: finding the proper way to consider latent variables in discrete choice models. *Transportation* 44 (3), 475–493.
- Bamberg, S., Schmidt, P., 2001. Theory-driven evaluation of an intervention to reduce the private car-use. *J. Appl. Soc. Psychol.* 31 (6), 1300–1329.
- Ben-Akiva, M., McFadden, D., Train, K., Walker, J., Bhat, C., Bierlaire, M., Bolduc, D., Boersch-Supan, A., Brownstone, D., Bunch, D.S., Daly, A., 2002. Hybrid choice models: progress and challenges. *Market. Lett.* 13 (3), 163–175.
- Bergantino, A.S., Bierlaire, M., Catalano, M., Migliore, M., Amoroso, S., 2013. Taste heterogeneity and latent preferences in the choice behaviour of freight transport operators. *Transport Policy* 30, 77–91.
- Bhat, C.R., 1998. Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling. *Transport. Res. A: Policy Pract.* 32 (7), 495–507.
- Bhat, C.R., Dubey, S.K., 2014. A new estimation approach to integrate latent psychological constructs in choice modeling. *Transport. Res. B: Methodol.* 67, 68–85.
- Bhat, C.R., Dubey, S.K., Nagel, K., 2015. Introducing non-normality of latent psychological constructs in choice modeling with an application to bicyclist route choice. *Transport. Res. B: Methodol.* 78, 341–363.
- Bhat, C.R., Koppelman, F.S., 1993. An endogenous switching simultaneous equation system of employment, income, and car ownership. *Transport. Res. A: Policy Pract.* 27 (6), 447–459.
- Bierlaire, M., 2020. A short introduction to PandasBiogeme. Technical report TRANSP-OR 200605. Transport and Mobility Laboratory, ENAC, EPFL.
- Bolduc, D., Ben-Akiva, M., Walker, J., Michaud, A., 2005. Hybrid choice models with logit kernel: applicability to large scale models. In: Lee-Gosselin, M., Doherty, S. (Eds.), *Integrated Land-Use and Transportation Models: Behavioural Foundations*. Elsevier, Oxford, pp. 275–302.
- Bolduc, D., Daziano, R.A., 2010. On estimation of hybrid choice models. In: *Choice Modelling: The State-of-the-Art and the State-of-Practice: Proceedings from the Inaugural International Choice Modelling Conference*. Emerald Group Publishing Limited, pp. 259–287.
- Cherchi, E., Meloni, I., de Dios Ortúzar, J., 2014. The latent effect of inertia in the modal choice. In: *Proceedings of the 13th International Conference on Travel Behavior Research*, pp. 517–534.
- Chorus, C.G., Kroesen, M., 2014. On the (im-)possibility of deriving transport policy implications from hybrid choice models. *Transport Policy* 36, 217–222.
- Daly, A., Hess, S., Patruni, B., Pologlou, D., Rohr, C., 2012. Using ordered attitudinal indicators in a latent variable choice model: a study of the impact of security on rail travel behaviour. *Transportation* 39 (2), 267–297.
- Daziano, R.A., Bolduc, D., 2013a. Covariance, identification, and finite-sample performance of the MSL and Bayes estimators of a logit model with latent attributes. *Transportation* 40 (3), 647–670.
- Daziano, R.A., Bolduc, D., 2013b. Incorporating pro-environmental preferences towards green automobile technologies through a Bayesian hybrid choice model. *Transportmetrica A: Transport Sci.* 9 (1), 74–106.
- Efthymiou, D., Antoniou, C., 2016. Modeling the propensity to join carsharing using hybrid choice models and mixed survey data. *Transp. Policy* 51, 143–149.
- Gärling, T., Fujii, S., Gärling, A., Jakobsson, C., 2003. Moderating effects of social value orientation on determinants of pro-environmental behavior intention. *J. Environ. Psychol.* 23, 1–9.
- Glerum, A., Stankovikj, L., Thémans, M., Bierlaire, M., 2013. Forecasting the demand for electric vehicle: accounting for attitudes and perceptions. *Transport. Sci.* 48 (4), 483–499.

- Hess, S., Palma, D., 2019. Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application. *J. Choice Model.* 32, 100170.
- Hess, S., Shires, J., Jopson, A., 2013. Accommodating underlying pro-environmental attitudes in a rail travel context: application of a latent variable latent class specification. *Transport. Res. D: Transport Environ.* 25, 42–48.
- Homer, P.M., Kahle, L.R., 1988. A structural equation test of the value-attitude-behavior hierarchy. *J. Personal. Soc. Psychol.* 54 (4), 638–646.
- Kamargianni, M., Polydoropoulou, A., 2013. Hybrid choice model to investigate effects of teenagers' attitudes toward walking and cycling on mode choice behavior. *Transport. Res. Rec.* 2382, 151–161.
- Kim, J., Rasouli, S., Timmermans, H., 2014. Expanding scope of hybrid choice models allowing for mixture of social influences and latent attitudes: application to intended purchase of electric cars. *Transport. Res. A: Policy Pract.* 69, 71–85.
- Kim, J., Rasouli, S., Timmermans, H.J., 2017. The effects of activity-travel context and individual attitudes on car-sharing decisions under travel time uncertainty: a hybrid choice modeling approach. *Transport. Res. D: Transport Environ.* 56, 189–202.
- Koppelman, F.S., Hauser, J.R., 1978. Destination choice behavior for non-grocery-shopping trips. *Transport. Res. Rec.* 673, 157–165.
- Maldonado-Hinarejos, R., Sivakumar, A., Polak, J.W., 2014. Exploring the role of individual attitudes and perceptions in predicting the demand for cycling: a hybrid choice modelling approach. *Transportation* 41 (6), 1287–1304.
- McFadden, D.L., 1986. The choice theory approach to marketing research. *Market. Sci.* 5 (4), 275–297.
- Muthén, L.K., Muthén, B.O., 2005. *Mplus: Statistical Analysis with Latent Variables: User's Guide*. Muthén & Muthén, Los Angeles, CA.
- Paulssen, M., Temme, D., Vij, A., Walker, J.L., 2014. Values, attitudes and travel behavior: a hierarchical latent variable mixed logit model of travel mode choice. *Transportation* 41 (4), 873–888.
- Prato, C.G., Bekhor, S., Pronello, C., 2012. Latent variables and route choice behavior. *Transportation* 39 (2), 299–319.
- Revelt, D., Train, K., 1996. Incentives for appliance efficiency in a competitive energy environment. *ACEEE Summer Stud. Proc.* 3, 123–129.
- Temme, D., Paulssen, M., Dannewald, T., 2008. Incorporating latent variables into discrete choice models – a simultaneous estimation approach using SEM software. *Business Res.* 1, 220–237.
- Triandis, H.C., 1977. *Interpersonal Behaviour*. Brooks/Cole, Monterey, CA.
- Thorhauge, M., Cherchi, E., Walker, J.L., Rich, J., 2019. The role of intention as mediator between latent effects and behavior: application of a hybrid choice model to study departure time choices. *Transportation* 46 (4), 1421–1445.
- Tudela A., Habib, K.M.N., Carrasco, J.A., Osman, A.O., 2011. Incorporating the explicit role of psychological factors on mode choice: A hybrid mode choice model by using data from an innovative psychometric survey. Presented at 2nd International Choice Modelling Conference, Leeds, United Kingdom, July 2011.
- Vij, A., Walker, J.L., 2016. How, when and why integrated choice and latent variable models are latently useful. *Transport. Res. B: Methodol.* 90, 192–217.
- Vij, A., Carrel, A., Walker, J.L., 2013. Incorporating the influence of latent modal preferences on travel mode choice behavior. *Transport. Res. A: Policy Pract.* 54, 164–178.
- Walker, J., Ben-Akiva, M., 2002. Generalized random utility model. *Math. Soc. Sci.* 43, 303–343.

Further Reading

- Kahneman, D., 2002. Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice. Nobel Prize Lecture, Stockholm, Sweden.
- McFadden, D., 2000. Disaggregate behavioural travel Demand's RUM Side. A 30-year retrospective. Nobel Prize Lecture, Stockholm, Sweden. In: Hensher, D.A. (Ed.), *Travel Behaviour Research: The Leading Edge*. Pergamon Press, Oxford.
- Walker, J., 2001. *Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables*. Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Explaining Data Analysis Using Qualitative Methods

Lena Levin, Sonja Forward, Swedish National Road and Transport Research Institute, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	107
Theoretical Perspectives	107
Transcription	108
Categorization	108
Validity and Reliability	109
Publishing	110
Summary and Conclusions	111
References	111

Introduction

Qualitative research data can be collected through a wide range of methods, such as interviews, focus groups, and participant and non-participant observations. The choice of method/methods are usually not decided until the research perspective and questions are formulated. Sometimes a combination of qualitative and quantitative methods might be used. For example, to discover commuting patterns among people in a city area and their experiences of the same, a quantitative method, like travel survey, might be used together with in-depth interviews and observations. How data is being analyzed varies between disciplines such as anthropology, case studies, ethnography, history, language studies, literature studies, psychology, sociology, social psychology, and a range of hybrid and interdisciplinary research traditions. Hence, how the data is being analyzed depends not only on the theoretical perspective but also on the study contexts and the research question.

This article connects to the previous article on qualitative methods (Chapter XX) and focuses on how qualitative data can be analyzed and then presented. However, it is important to bear in mind that many different methods can be used when analyzing data, some of them this article will only touch upon. For instance, qualitative text analysis, ethnographic fieldwork, news media analysis, image analysis, and analysis of political discourses and social media (for an extensive review see [Silverman, 2004; 2016](#)). Despite this, the general agreement is that analysis is an ongoing, iterative process that begins in the early stages of data collection and continues throughout the study ([Bradley et al., 2007](#)).

Theoretical Perspectives

Quantitative and qualitative studies are different in so far as the first are built on numbers, amounts or quantities; while the latter are built on stories, accounts, or thick descriptions. Inherent in this dichotomy is the view that quantitative studies generally adopt a deductive process (i.e., begins with a theory which is tested), while qualitative studies generally adopt an inductive process (i.e., generate a new theory from data). Although, this distinction is true in general, it does not fully, and accurately, describe the processes adopted by quantitative and qualitative researchers in practice ([Hyde, 2000](#)).

According to literature on qualitative methods (e.g., [Bitektine, 2007; Hyde, 2000](#)), both inductive reasoning and deductive reasoning may support the knowledge production in qualitative research. Inductive reasoning starts with observations of specific instances, seeking to increase and develop the theory through knowledge about the phenomena under investigation. Deductive reasoning starts with a hypothesis and an established theory and seeks to determine if the theory is supported or rejected. Thus, the analysis process of qualitative data can be done either on the basis of the theory guiding the research questions or on the basis of the salient issues that arises in the text itself or on the basis of both approaches. In any case the approach taken will influence the findings and how these are reported ([McGowan et al., 2020](#)).

Types of theories which can be relevant include:

- **Individual theories**—for example, theories of individual development, cognitive behavior, personality, learning, individual perception, and interpersonal interactions
- **Group theories**—for example, theories of family functioning, ageing, gender relations, informal groups, work teams, supervisory and employment relations, and interpersonal networks
- **Organizational theories**—for example, theories of bureaucracies, organizational structure and functions, hierarchy in institutions, and interorganizational relationship
- **Societal theories**—for example, theories on urban development, international behavior, political and cultural institutions, technological development, transport systems, and market functions ([Yin, 2009](#), p. 31)

Some theoretical approaches cut across these theory types, for instance: decision making theory can involve individuals, organizations, and social groups. Another example of cross cutting is the evaluation of public supported programs, such as a national program for traffic safety. Public programs typically include both theories on society and theories on groups or individual behavior.

Transcription

Qualitative researchers who conduct interviews, focus groups, and conversation analysis are often working with transcriptions of audio and/or video recordings. However, there are many variations of how to construct and deal with transcriptions.

By transcription is meant to transfer utterances (by one or more speakers) to written notation, which is performed for the purpose of studying the structure, form, and/or content of the speech or conversation.

In a review of the literature, including 46 journal articles and 30 book chapters or sections from books, [Davidson \(2009\)](#) divides the transcription of qualitative research data into three broad areas: (1) how transcription is defined and understood, (2) how transcription is conducted, and (3) how transcription is reported in research studies.

A general view is that transcription has to be selective, interpretive, and representational. In linguistic anthropology, transcription has been viewed as “cultural practice” or “cultural activity” and transcripts as artifacts that possess “temporal-historical dimensions” ([Duranti, 2007](#), p. 302). Further, transcription is understood to reflect theory and to shape it ([Du Bois, 1991](#)) as researchers “reflexively document and affirm theoretical positions” ([Mischler, 1991](#), p. 271, in [Davidson, 2009](#), p.37) during the process of transcription and analysis.

Researchers make choices, and these are inherently related to the theoretical perspective and how researchers locate themselves and others in the research process. Because it is impossible to record all features of talk and interaction from recordings, all transcripts have to be selective in one way or another. The detail and the scope of the transcript can therefore vary. As [McLellan et al. \(2003\)](#) argue: “at some point, a researcher must also settle on what is transcribed /.../ despite all best intentions, the textual data will never fully encompass all that takes place during an interview” (p. 65).

It may, for example, be limited to certain verbal aspects or include a number of para- or extralinguistic aspects during and between the spoken accounts. According to [Kvale \(2007\)](#), transcripts are artificial constructions from an oral to written mode of communication, and a researcher must make choices in regards to what extent a textual document should also include nonlinguistic observations (e.g., facial expression, body language, and length of pauses) or be transcribed verbatim and identify specific speech patterns, vernacular expressions, intonations, or emotions. [Poland and Pederson \(1998\)](#), for instance, stated that what is not said is just as important as what is said, and thus to grasp the meaning researchers need to include contextual information regarding silence or pauses in conversation (cf. [Goffman, 1981](#)). However, how detailed the transcript should be depends on the aim of the study and the theoretical and methodological framework. One rule of thumb is to make an exact reproduction of what is being communicated and not try to edit the text too early ([McLellan et al., 2003](#)). For instance, correct grammar and mispronunciations.

The practice of transcription might also be divided at several levels. There are both plain and more complicated systems. Some research areas provide detailed information following central notation systems (cf. [Jefferson, 1990/2004](#); [Mondada, 2007](#)) and the emerging field of interaction analysis in mobile settings ([Haddington et al., 2013](#)). Computer technology nowadays makes it possible to both make good recordings and to view recordings as well as developing transcripts on the computer screen using software programs developed specifically for transcription.

Furthermore, language implications might be considered by researchers. Transcriptions that encompass translation from one language to another are especially complex and challenging. It might require the use of interpreters, for example, and transcribers other than the researcher if the researcher is not a native speaker of the language used ([Duranti, 1997](#)). Many researchers use hired professional transcribers, although transcribers often need instruction in relation to the specific purposes of the study and the transcription requirements of it.

According to [Davidson \(2009\)](#) the researchers need to clarify aspects of the transcription for themselves but also to others involved in the research and paid transcribers. Furthermore, the transcription process, which includes selectivity, needs to be articulated in the written reports of the research.

Categorization

The analysis often starts during the transcription procedure when the researcher is carefully listening to audio recordings of interviews or watching video recordings of the interaction between informants. What is regarded as data varies. Some researchers consider research recordings to be “data” and use transcription as summoned representations, where others argue that transcripts are the data.

Qualitative researchers often organize their data in categories, based on concepts, and some form of coding system. Typically, at the first stage of the categorization process the transcripts are analyzed line by line, and pertinent excerpts are given provisional conceptual descriptions (codes). The next stage involves the search for relationships between conceptual labels and to develop categories. At the final stage, categories are integrated and refined.

The following steps represent a systematic approach that allows for open discovery of emergent concepts with a focus on generating taxonomy, themes, or theory ([Bradley et al., 2007](#)):

- Reading for overall understanding.
- Coding, provides the analyst with a formal system to organize the data but also to document additional links within and between concepts described in the data. Codes can be described as tags or labels, which are assigned to whole documents or segments of

documents. According to Bradley et al. (2007) some experts argue that a single researcher conducting all the coding is both sufficient and preferred. However, analysis conducted by one person may not be possible to be repeated by others who have different backgrounds. In this case it would therefore be important to clearly present the researcher's biases and philosophical approaches. In contrast, others recommend that the coding process involve a team of researchers, preferably with differing backgrounds.

- Developing the code structure. How this should be developed depends on if the coding should be inductive or more deductive. See examples here.

Grounded theory encourages the development of codes as purely inductive (Strauss and Corbin, 1998). As more data are reviewed, the specifications of codes are developed and refined to fit the data. To ascertain whether a code is appropriately assigned, the analyst compares text segments to segments that have been previously assigned the same code and decides whether they reflect the same concept. Through this process, the code structure evolves inductively, reflecting "the ground," that is, the experiences of participants.

In contrast, if a deductive approach is used to develop a code structure, the initial step is to use a theory. Although this process started earlier when the interview guide was developed. For instance, if the study is based on the Theory of Planned Behavior, the guide would include questions about: attitudes, norms, and perceived behavioral control (Ajzen, 1991). However, it could be argued that such approach would be rather limited, leaving little room for new discoveries (McGowan et al., 2020). Instead, both an inductive and deductive approach is sometimes used. In this case the structure and the coding are based on the theory, but at the same time nonrelated factors are included.

In order to develop a code structure, five code types have been identified, which can be used to generate taxonomy, themes, and theory: (1) conceptual codes and subcodes identifying key concept domains and essential dimensions of these concepts; (2) relationships codes—links between other concepts; (3) participant perspective codes, which identify, for example, if the participant appears positive, negative, or indifferent about a particular experience or part of an experience; (4) participant characteristic codes; and (5) setting codes.

Interview analysis can be focused on one or more of these code perspectives, depending on the research objectives. In interaction analysis (e.g., focus groups and conversation) the analysts do not talk so much about codes but on building categories from data collections of recurrent and deviant cases—often demonstrated by excerpts from the transcriptions.

Next sections present two examples of how to approach data analysis. The first uses a coding method to extract the interview persons' experiences of a disruptive occasion and the second uses multimodal sequence analysis to identify activities in a situated learning process and the participants' interaction in the specific mobile setting of car driving.

A model of analysis, developed by Graneheim and Lundman (2003) and also practiced by Nyberg et al. (inpress), is based on a matrix organizing the utterances in three levels. The transcribed interviews are read through several times to obtain a sense of the whole and then the transcriptions are extracted into "meaning units," condensed meaning units and codes (see Table 1). In this case the codes represent interviewees experiences of losing their driving license due to visual field injury.

The various codes were in this example compared based on differences and similarities and sorted into tentative categories. Finally, the tentative categories were discussed by a research team and then revised by the main author (Nyberg).

The next example comes from a study called "Driver Training in Practice" (DTriP). The aforesaid analysis was inspired by conversation analysis (CA) (see e.g., Garfinkel, 1967; Goffman, 1981; Sacks, 1992) and so called "tacit knowledge," which is often taken for granted by an experienced driver. Learning to drive is a process of interaction where practical knowledge is communicated from the driver instructor and acted on by the novice trainee driver (cf. Zemel and Koschmann, 2014). Both talk, eye gaze and body movements are important for the practice of learning driving skills. Hence, more than 100 h of driving lessons was recorded in order to monitor what resources the participants use during driving practice (Broth et al., 2019; Levin et al., 2017). The analysis started by thorough listening and watching the video recordings and doing multimodal transcriptions (cf. Mondada, 2007). Recurrent activities, which were of interest due to the research questions were chosen, assigned a unique name, and stored in a data base. These findings constituted collections for further analysis and subcategorization. During the analysis the interaction was analyzed sequence by sequence, and the transcriptions were constructed in increased detail so as to collect as much information as possible; also print screen (photos) from the video recordings or sketches of situations were constructed (see e.g., Broth et al., 2019; Fig. 1).

Validity and Reliability

In quantitative research both validity and reliability are important criterion for describing its quality. Validity is about making sure that the measuring instrument is adapted to what you intend to study and that it reflects real events. Reliability, traditionally, means that a study conducted at a later stage, using the same methods, should arrive at the same results as the previous on (i.e., consistent

Table 1 Excerpt from Nyberg et al. (work in progress)

Meaning unit	Condensed meaning unit	Code
"I was deeply shocked because this is a big tragedy in my life"	Was deeply shocked	Shock
	Great tragedy in life	Tragedy



Figure 1 Example of video data from a research project using small go-pro cameras recording the interaction of driver training (DTriP, The Swedish Research Council #2012-5376). Selected part of excerpt 1 (ts2_13_k2_28m11s_28m21s) from Broth et al. (2019). *INST*, Instructor; *TRDR*, trainee driver.

03 **INST** släpp ga+sen_z
 let go of the gas
 inst ->+gaze rvm->
 04 **&(0.4)&(0.1)+la:stbilen vill köra om dej_z**
 the lorry wants to overtake you
 inst ->+gaze ahead
 tdvr &gz lsvm&gaze ahead

over time). In qualitative research this is not always possible since it is more concerned with exploring “meaning,” in certain contexts, than replication and generalization to a large population (Halliday, 2016; Mason, 2010). Instead, criterions such as credibility, transferability, and trustworthiness are used, and validity and reliability are not regarded as two separate entities (Golafshani, 2003).

Other methods can therefore be used to assess the reliability in qualitative research and provide transparency of the study’s structure. This would mean that the researcher has to justify every move they make (Halliday, 2016). “The general way of approaching the reliability problem is to make as many steps as operational as possible and to conduct research as if someone were always looking over your shoulder” (Yin, 2009, p. 38).

The question is also if the study has been analyzed in a detached way or if the results partly reflects the researchers preconceived ideas. The pre-understandings and conditions need to be clear (Silverman, 2004). According to Thomas (2003) the trustworthiness of findings can be assessed by a range of techniques, such as: (1) independent replication of the research, (2) comparison with findings from previous research, (3) triangulation within a project, (4) feedback from participants in the research (so-called “member checks”), and (5) feedback from users of the research findings.

The accuracy of the interpretation of data from an interview study can also be checked using so-called “member checks,” which means that participants in the study are able to review, comment, and/or correct the material: (1) review a transcript of their own interview, (2) a copy of preliminary findings, and (3) a draft copy of a research manuscript. Member checks can be useful for obtaining participant approval when using quotations and where anonymity cannot be guaranteed (Thomas, 2017).

Publishing

Qualitative studies usually result in a large amount of data, which means that even at this stage some form of selection is needed. Thus, the writing process should be considered a stage in the analysis process.

Publishing in high impact scientific journals versus popular science magazines or stakeholder reports are quite different (see Table 2) but also show some similarities. In all cases the message needs to be clear and the presentations be adapted to the readers’

Table 2 Audiences and their expectations

Audience	Expectations
Academic colleagues	Theoretical, factual, or methodological insights; scientific dialogue
Policy makers	Practical information and knowledge development relevant to policy issues; problem solving
Practitioners	A theoretical framework for understanding clients/customers better; factual information; practical suggestions for better procedures; reform of existing practices
The general public	New facts; ideas for reform of current practices or policies; reflecting their everyday life and confirmations that others share their own experience of particular problems; improvements; guidelines for how to manage better or get better service from practitioners or institutions

Source: Adapted from Silverman, 2001. *Interpreting Qualitative Data. Methods for Analysing Talk, Text and Interaction*. Sage, London, p. 267.

knowledge and interest. Clark and Thompson (2016, p. 2) point out that the writer should nail the key message: strong, clear, and concise about what the research found and why this is important. With too many messages, the main points of the paper are lost. If there are no clear messages, the readers are wondering “what’s the point?”

Most qualitative, inductive studies report between three and eight main categories and often pick out exceptions to illustrate these findings in their publications. Inexperienced researchers often try to report 10 or more major categories. This indicates that they have probably not finished the process of combining the smaller categories into larger categories or that they have not made the crucial decisions about which categories are the most important (Thomas, 2003).

Quotes from transcriptions are used as verifications of analysis results and can be considered as part of the validation procedure in qualitative research. Quotes typically illustrate the results and verify the story when writing and publishing. Quotes and excerpts for publications should be chosen to exemplify recurrent and/or specific findings or deviant cases, which are discussed in the text. Note that the academic journals often need concise messages about the research findings and convinces why it is important—a large number of quotes without scientific discourse does not serve these purposes.

Caulley (2008) argues that qualitative researchers should be more generous in using concrete details from fieldnotes or notes from research diaries. It is important for the researcher to be a careful observer of details. However, it is also crucial to handle data carefully with respect to interviewees and participants’ integrity who may not have given their consent to become identifiable in either scientific reports or popular science media.

Note that researchers usually select journals only after most, or all, of the manuscripts have been written, which is a mistake. The advice from editors is to view the work more in terms of how it fits with specific journals and use this as a help when writing the paper. Editors usually have a vision of who are their readers and intended readers and the kinds of articles they want in their journal. “Try to get a better sense of this vision by taking opportunities to communicate with editors at conferences, on social media, and via e-mail” (Clark and Thompson, 2016, p. 2–3).

Summary and Conclusions

In transportation research, various analysis methods are practiced to understand complexity of human behavior but also their needs and motivations. This article discusses analysis of qualitative data, using examples from audio recorded interviews and video recordings of sequential interaction. Qualitative analysis is built on stories, accounts, or “thick” descriptions. The approach can be described as inductive (i.e., starts with an observation and ends with a theory) although a deductive approach can also be used (i.e., starts with a theory, which is tested). How the data is being analyzed depends on the theoretical perspective but also on the study contexts and the research question.

The analysis in qualitative research often starts during the transcription procedure when the researcher is listening to audio recordings of interviews or watching video recordings of the interaction between informants. Invariably this results in a large amount of information and all transcripts have to be selective in one way or another. The next step is to organize the data in categories, based on concepts, and in some coding systems, which help to organize data in collections or clusters. One conclusion is that coding systems can be either theory or data driven. The article also discusses the assessment of validity and reliability in qualitative research, which is different from quantitative studies. In qualitative research they are not regarded as two separate entities and criterions such as: credibility, transferability, and trustworthiness are used instead.

The selective process finishes when writing the results, and they should be considered a stage in the analysis process. Quotes from transcriptions are used as verifications of analysis and can be considered as part of the validation procedure in qualitative research. They typically illustrate the results and verify the story when writing and publishing quotes and excerpts, which should be chosen to exemplify recurrent and/or specific findings or deviant cases, which are discussed in the text.

References

- Ajzen, I., 1991. The theory of planned behaviour. *Organ. Behav. Hum. Decis. Process.* 50, 179–211.
- Bitektine, A., 2007. Prospective case study design: qualitative method for deductive theory testing. *Organ. Res. Methods* 11, 160–180.
- Bradley, E.H., Curry, L.A., Devers, K.J., 2007. Qualitative data analysis for health services research: developing taxonomy, themes and theory. *Health Serv. Res.* 42, 1758–1772, doi: 10.1111/j.1475-6773.2006.00684.x.
- Broth, M., Cromdal, J., Levin, L., 2019. Telling the other’s side. Formulating others’ mental states in driver training. *Lang. Commun.* 65, 7–21.
- Caulley, D.N., 2008. Making qualitative research reports less boring. The techniques of writing creative nonfiction. *Qual. Inq.* 14, 424–449.
- Clark, A., Thompson, D.R., 2016. Five tips for writing qualitative research in high-impact journals. *Int. J. Qual. Methods* 17, 1–3.
- Davidson, C., 2009. Transcription: imperatives for qualitative research. *Int. J. Qual. Methods* 8, 35–52.
- Du Bois, J.W., 1991. Transcription design principles for spoken discourse research. *Pragmatics* 1, 71–106.
- Duranti, A., 1997. *Linguistic Anthropology*. Cambridge University Press, Cambridge, UK.
- Duranti, A., 2007. Transcripts, like shadows on a wall. *Mind Cult. Act.* 13, 301–310.
- Garfinkel, H., 1967. *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, NJ.
- Goffman, E., 1981. *Forms of Talk*. Blackwell, Oxford.
- Golafshani, N., 2003. Understanding reliability and validity in qualitative research. *Qual. Report* 8 (4), 597–606. Retrieved from <http://nsuworks.nova.edu/tqr/vol8/iss4/6>.
- Graneheim, U.H., Lundman, B., 2003. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Educ. Today* 24, 105–112.
- Haddington, P., Mondada, L., Nevile, M. (Eds.), 2013. *Interaction and Mobility*. De Gruyter, Berlin and Boston.

- Halliday, A., 2016. *Doing & Writing Qualitative Research*, third ed. Sage, London.
- Hyde, K.F., 2000. Recognising deductive processes in qualitative research. *Qual. Mark. Res.* 3, 82–89.
- Jefferson, G., 1990/2004. Glossary of transcript symbols with an introduction. In: Lerner, G.H. (Ed.), *Conversation Analysis: Studies from the First Generation*. John Benjamins, Amsterdam, pp. 13–31.
- Kvale, S., 2007. *Doing Interviews*. Sage Publications, Thousand Oaks, Calif.
- Levin, L., Cromdal, J., Broth, M., Gazin, A.-D., Haddington, P., McIlvenny, P., Melander, H., Rauniomaa, M., 2017. Unpacking corrections in mobile instruction: error-occasioned learning opportunities in driving, cycling and aviation training. *Linguist. Educ.* 38, 11–23.
- Mason, M., 2010. Sample size and saturation in PhD studies using qualitative interviews [63 paragraphs]. *Forum Qual. Sozialforschung* 11 (3). Art. 8. <http://nbn-resolving.de/urn:nbn:de:0114-fqs100387>.
- McGowan, L.J., Powell, R., French, D.P., 2020. How can use of the theoretical domains framework be optimized in qualitative research? *Br J. Health Psychol.* 25 (3), 677–694.
- McLellan, E., MacQueen, K.M., Neidig, J.L., 2003. Beyond the qualitative interview: data preparation and transcription. *Field Methods* 15, 63–84.
- Mischler, E., 1991. Representing discourse: the rhetoric of transcription. *J. Narrative Life History* 1, 255–280.
- Mondada, L., 2007. Commentary: transcript variations and the indexicality of transcribing practices. *Discourse Stud.* 9, 809–821.
- Nyberg, J. et al., forthcoming. Distrust in authorities. A Swedish example from interviews with individuals who have had their driving license withdrawn due to visual field loss.
- Poland, B., Pederson, A., 1998. Reading between the lines: interpreting silences in qualitative research. *Qual. Inq.* 4, 293–312.
- Sacks, H., 1992. *Lectures in Conversation*. Blackwell, Oxford.
- Silverman, D., 2001. *Interpreting Qualitative Data. Methods for Analysing Talk. Text and Interaction*. Sage, London.
- Silverman, D., 2004. *Qualitative Research. Theory, Method and Practice*, third ed. Sage, London.
- Silverman, D., 2016. *Qualitative Research*, fourth ed. Sage, London.
- Strauss, A., Corbin, J., 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, Thousand Oaks, CA.
- Thomas, D.R., 2003. A general inductive approach for qualitative data analysis. *Am. J. Eval.* 27, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.462.5445&rep=rep1&type=pdf>.
- Thomas, D.R., 2017. Feedback from research participants: are member checks useful in qualitative research? *Qual. Res. Psychol.* 14, 23–41, <http://dx.doi.org/10.1080/14780887.2016.1219435>.
- Yin, R., 2009. *Case Study Research. Design and Methods*. Sage, London.
- Zemel, A., Koschmann, T., 2014. Put your fingers right in here: Learnability and instructed experience. *Discourse Stud.* 6 (2), 163–183.

Driver Distraction

Oscar Oviedo-Trespalacios*, **Michael A. Regan†**, *Centre for Accident Research and Road Safety – Queensland (CARRS-Q), Queensland University of Technology (QUT), Kelvin Grove, QLD, Australia; †Research Centre for Integrated Transport Innovation (rCITI), University of New South Wales, Sydney, NSW, Australia

© 2021 Elsevier Ltd. All rights reserved.

Driver Distraction and Inattention	113
Types of Distractions and Triggered Responses	113
Impact on Driving Performance	114
Factors Moderating Driver Distraction	114
Evidence Implicating Distraction as a Traffic Safety Problem	115
Prevention and Management of Distracted Driving	116
Countermeasures for Distracted Driving	116
Looking to the Future	116
Concluding Comments	118
Acknowledgment	118
See Also	118
References	119
Further Reading	120

Driver Distraction and Inattention

There is converging evidence, reviewed in this chapter, that driver distraction is a significant road safety problem. Driver distraction is one mechanism of driver inattention. [Regan et al. \(2011\)](#) conceptualized this relationship by developing a taxonomy of driver attention using previous classifications of attentional failures identified as having contributed to crashes in in-depth crash studies (e.g., [Hoel et al., 2010](#); [Treat 1980](#); [Van Elslande and Fouquet 2007](#); [Wallén Warner et al., 2008](#)). [Regan et al. \(2011\)](#) defined driver inattention as “insufficient or no attention to activities critical for safe driving” (p. 1776) and proposed that driver inattention is induced by five attentional mechanisms (identified in [Fig. 1](#)), one of which they labeled “Driver Diverted Attention”; which is synonymous with driver distraction (see [Fig. 1](#)). [Regan et al. \(2011\)](#) defined Driver Diverted Attention (i.e., driver distraction) as “the diversion of attention away from activities critical for safe driving toward a competing activity, which may result in insufficient or no attention to activities critical for safe driving” (p. 1776). This definition was grounded on an earlier definition formulated by [Lee et al. \(2009, p. 34\)](#) that was subsequently endorsed by an international group of experts convened by the International Organisation for Standardization (ISO): “Driver distraction is the diversion of attention away from activities critical for safe driving toward a competing activity”.

Types of Distractions and Triggered Responses

Many different sources of distraction have been identified in crash and naturalistic driving studies, and these have been categorized as follows ([Regan and Hallett, 2011](#)): objects (e.g., mobile phone, advertising billboard); events (e.g., crash scene, lightning); passengers (e.g., child, adult); other road users (e.g., cyclists, pedestrians, other vehicles); animals (e.g., dog); and internal stimuli (e.g., that trigger day dreams and thoughts). A source of distraction has certain “modal properties” (e.g., visual, auditory; [Hallett et al., 2011](#)) which, along with other properties, may trigger a diversion of attention away from activities critical for safe driving.

Different “types” of distraction can be classified according to the sensory modality through the diversion of attention is triggered. [Regan and Hallett \(2011\)](#) conceptualized six types of distraction, distinguished by the diversion of attention towards things that:

- we see (“visual distraction”)
- we hear (“auditory distraction”)
- we smell (“olfactory distraction”)
- we taste (“gustatory distraction”; e.g., the taste of a rotten piece of fruit), and
- we feel (tactile distraction; e.g., the feel of a large spider crawling on one’s leg).

The actions that a driver may perform while engaged in a competing activity are potentially enormous (e.g., looking, reaching, touching, etc.). However, the behavioral effects triggered by these driver actions are finite in number. [Hallett et al. \(2011\)](#) have referred to these behavioral effects as “triggered responses”, and have characterized them (for distracted drivers) as follows:

- Eyes off the Road – driver takes eyes off activities critical for safe driving
- Mind off the road – driver takes mind off activities critical for safe driving

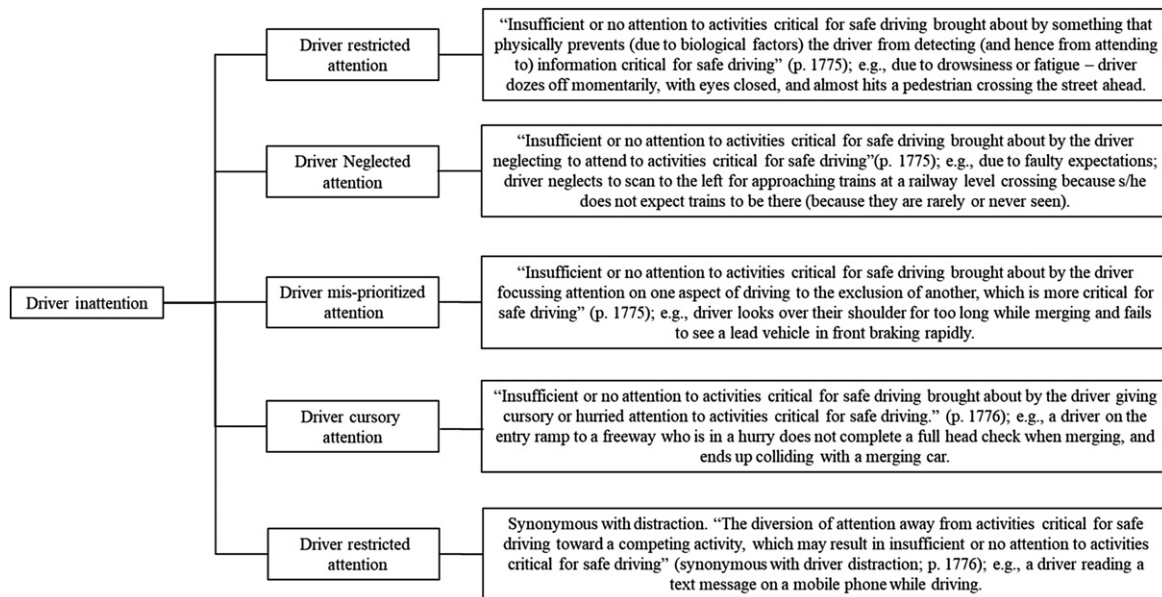


Figure 1 Mechanisms of Driver Inattention Source: [Regan et al. \(2011\)](#).

- Ears off the road – driver takes ears off activities critical for safe driving (as a result of having one's mind off the road), and
- Hands or feet off (vehicle) controls – driver takes hands and/or feet off activities critical for safe driving.

Different competing tasks induce distinctive triggered responses, often in combination. For example, looking at an advertising sign will take a driver's "Eyes off the Road" and their "Mind off the road" when thinking about the message(s) it conveys.

Impact on Driving Performance

The triggered responses induced by the different types of distraction have been found to interfere with, and degrade, driving performance in various ways (see [Bayly et al., 2008](#); [Bruyas, 2017](#); [Horberry and Edquist, 2009](#); [Oviedo-Trespalcacios et al., 2016](#)).

Competing activities that mainly take drivers' eyes off the road have been found to affect some specific aspects of driving performance:

- the selection of information (e.g., failing to detect relevant roadway information; spatially concentrated gaze on the forward road centre when the driver's eyes are returned to the forward roadway);
- information processing (e.g., longer reaction times to braking lead vehicles and roadway warnings; "change blindness" that inhibits the detection of changes in the road scene); and
- vehicle control (e.g., degraded lane keeping performance; reduced speed; increased following distance).

Competing activities that mainly take a driver's *mind* off the road are also known to affect specific aspects of driving performance:

- selection of information: reduced checking of speedometer and rear-view mirrors, and less attention to hazards in the periphery due to spatially concentrated gaze on the forward road centre;
- information processing: inattention blindness, missing red lights, and the inability to remember some things that have been seen during a drive;
- vehicle control: more hard braking, more navigation errors, changes in lane keeping performance, changes in following distances and speed, and fewer lane changes; and
- more traffic rule violations: speeding, red light running, crossing solid lines.

Factors Moderating Driver Distraction

The same competing activity (e.g., typing a text message) can have different effects on activities critical for safe driving depending on a range of moderating factors. These include ([Oviedo-Trespalcacios et al., 2019a](#); [Young et al., 2009](#)):

- Driver characteristics: such as age, gender, driving experience, driver state (e.g., drowsy, drunk, angry), familiarity with and amount of practise with the competing task, and personality (e.g., the propensity to take risks) ([Huth and Brusque, 2014](#); [Oviedo-Trespalcacios et al., 2020a](#); [Young et al., 2009](#)).

- Driving task demand: characteristics of the driving task itself, such as traffic conditions, weather conditions, road conditions/design, the number and type of vehicle occupants, the ergonomic quality of vehicle cockpit design, and vehicle speed (Li et al., 2020; Onate-Vega et al., 2020; Oviedo-Trespalcacios et al., 2017, 2020a; Young et al., 2009).
- Secondary task demand: competing (secondary) tasks may degrade driver performance depending on (1) how similar the task is to driving sub-tasks (e.g., whether it requires visual and/or manual control actions similar to those required for driving), (2) its complexity, (3) whether or not it can be ignored, (4) how predictable it is, (5) how easily it can be adjusted, (6) how easy it is for the task to be interrupted and resumed, and (7) how long it takes to perform the task (Oviedo-Trespalcacios et al., 2020a; Regan et al., 2011; Young et al., 2009).
- Self-regulation: the ability of a driver to regulate behavior in the face of, or in anticipation of, a competing activity in order to compensate for its potentially adverse effects (Oviedo-Trespalcacios et al., 2019a; Young et al., 2009). This can occur at different levels of driving control: strategic (e.g., turning on mobile phone applications that prevent mobile phone use while driving); tactical (e.g., limiting or ceasing phone use in complex traffic conditions); and operational (e.g., compensating for increased demands by increasing headway distance or reducing speed) (Chen et al., 2020; Li et al., 2019; Oviedo-Trespalcacios, 2019a; Saifuzzaman et al., 2015).

Evidence Implicating Distraction as a Traffic Safety Problem

Both crash archives and crash risk studies have associated distraction with increased risk of road crashes.

Crash studies using official archives are frequently found in the literature. A review of studies using police archives from New Zealand and the United States found that 10%–12% of police-reported crashes were linked to driver distraction (Gordon, 2009). A more recent investigation in the United States using the Fatal Accident Reporting System (FARS) database reported that 7.7% of all fatal crashes involved distraction (Qin et al., 2019). In Australia, Beanland et al. (2013) found that 13.9% of casualty crashes were associated with in-vehicle sources of distraction, such as mobile phones or animals/insects in the vehicle. In Norway, Sundfør et al. (2019) found that 2%–4% of all fatal crashes involved mobile phone use while driving, whilst interaction with other sources of in-vehicle distraction, such as GPS, laptop or tablet computers, video cameras, backing cameras, and passengers, contributed to 8% of all fatal crashes. Eby and Kostyniuk (2003) determined that single-vehicle-run-off-the-road crashes and rear-end crashes are the most common types of crash associated with driver distraction.

Crash studies using archival data lack behavioral detail around the circumstances of driver distraction: many crashes considered to involve distraction in the United States, for example, are listed as “distraction/inattention details unknown” in official crash data (NHTSA 2016). To overcome this limitation, crash risk studies have been undertaken that produce this behavioral detail and estimate changes in crash risk associated with distracted driving.

Crash risk studies provide information about the increased driving risk posed by driver involvement in a distraction-related activity over and above that of the risk posed by “normal” driving. Naturalistic driving studies (NDS), where driver and vehicle behavior are recorded using video, accelerometers and other unobtrusive instrumentation, for long periods of time, have provided crash risk estimates for a wide range of distracting activities. The *Strategic Highway Research Program (SHRP 2)* naturalistic driving study in the United States is the largest and most comprehensive such study in the world, having involved nearly 3500 drivers for three years. From the SHRP 2 study, Dingus et al. (2016) reported that observable distractions, overall, increased the odds of having a crash by a factor of 2.0.

Another comprehensive naturalistic driving study in the United States was conducted by Fitch et al., (2013) to investigate changes in crash risk due to handheld and hands-free phone use among 204 drivers during a period of 4 weeks. Fitch et al. (2013) found handheld mobile phone use, overall, to increase the odds of having a crash by a factor of 1.39, while Dingus et al. (2016) reported SHRP 2 data confirming that interaction with a handheld mobile phone, overall, increased the odds of a crash by a factor of 3.6.

Table 1 summarizes the results from these key studies and identifies increases in risk associated with driver engagement in a range of other distracting activities. Noteworthy are increases in risk associated with talking on a handheld phone, ranging from 0.79 (Fitch, 2013) to 2.2 (Dingus et al., 2016). Recently, Young (2017) recalculated the odds ratio of handheld mobile phone conversations using the SHRP 2 data after controlling for selection and confounding bias, which have been noted by others as issues in some of these studies (see Wijayaratna et al., 2019). This resulted in an odds ratio of between 0.94 and 0.92. Neither value is significantly different from 1, implying that there is no increase in risk associated with talking on a handheld mobile phone. These values are similar to findings reported in previous naturalistic studies (Fitch et al., 2013).

Distraction can also derive from driver exposure to sources external to the vehicle. Generally, these have not received the same level of attention in the published research as in-vehicle distractions. Around 30% of distraction-related crashes derive from driver engagement with sources outside the vehicle. These include animals, architecture, advertising signage, construction zones/equipment, crash scenes, incidents (e.g., road rage), insects, landmarks, road signs, road users, scenery, other vehicles and weather (e.g., lightning) (Gordon, 2009).

In a Systematic Review conducted by Oviedo-Trespalcacios et al. (2019b), there was a trend in the literature of reporting an increased crash risk associated with driver interaction with digital billboards. It is known from the work of Dingus et al. (2016) that an extended eye glance duration to an external object increases the odds of having a crash by a factor of 7.1 (OR). Roadside advertising signs are designed deliberately to attract and maintain driver attention to information that is irrelevant to driving.

Table 1 Summary of crash risk findings from naturalistic driving studies

<i>Study</i>	<i>Observed distraction</i>	<i>Odds ratio (95% CI)</i>
Fitch et al. (2013)	Mobile phone visual-manual task (i.e., text messaging/browsing, locate/answer, dial, push to begin/end use, and end handheld phone use)	1.73 (1.1–2.7) ^a
	Mobile phone handheld talk	0.79 (0.43–1.44)
	Mobile phone portable hands-free talk	0.73 (0.36–1.47)
	Mobile phone integrated hands-free talk	0.73 (0.36–1.47)
Dingus et al. (2016)	Observable distraction (overall)	2.0 (1.8–2.4) ^a
	Total mobile phone handheld	3.6 (2.9–4.5) ^a
	Mobile phone handheld talk	2.2 (1.6–3.1) ^a
	Mobile phone browse	2.7 (1.5–5.1) ^a
	Mobile phone handheld dial (a number)	12.2 (5.6–26.4) ^a
	Mobile phone handheld text	6.1 (4.5–8.2) ^a
	Mobile phone handheld browse	2.7 (1.5–5.1) ^a
	In-vehicle information systems	4.6 (2.9–7.4) ^a
	Extended eye glance to an external object	7.1 (4.8–10.4) ^a
	Reading and writing (including with tablets)	9.9 (3.6–26.9) ^a
	Personal hygiene activities	1.4 (0.8–2.5) ^a
	Eating	1.8 (1.1–2.9) ^a
	Reaching for objects inside the vehicle (excluding mobile phones)	9.1 (6.5–12.6) ^a
Young (2017)	Mobile phone handheld talk	0.94 (0.30–2.3) or 0.92 (0.45–1.7)

^aSignificantly increase driving risk

Prevention and Management of Distracted Driving

Traditionally, the responsibility for managing distraction has been on the driver. This view implies that when engaging in distraction drivers are solely to blame for any adverse outcomes. Generally, this “victim blaming” approach has, to date, been the status quo of distraction prevention. However, safety experts concur that this approach is unsuitable for dealing with distracted driving (or any other risky behavior; Tingvall and Haworth 1999; Tingvall et al., 2009; Young and Salmon, 2015). Drivers tend to be, and will continue to be, distracted due to a number of factors that are often difficult to control. For example, recent research has shown that problematic mobile phone use has characteristics that resemble addiction (Oviedo-Trespalcacios et al., 2019c). The connection between problematic phone use and distracted driving suggests that the involvement of wider mental health stakeholders may be needed to prevent mobile phone use while driving.

Given that current prevention strategies mainly targeting the driver (i.e., their motivation and beliefs about the risks of engaging with secondary tasks) have shown limited effectiveness, there has been a move recently towards considering the role of all stakeholders in society with responsibilities for the management of driver distraction. The premise is that distracted driving is a product of wider systems factors that are imposed on drivers (Parnell et al., 2016). Nowadays, some groups of stakeholders, directly linked with distracted driving, have not assumed their responsibilities (Oviedo-Trespalcacios et al., 2019c; Parnell et al., 2017; Young and Salmon 2015). An example is the often-complacent role of the creators of mobile phone applications used by drivers, who have the technology available to prevent it (Galitz, 2017). There is no doubt that effective prevention of distraction lies in the active participation of a wider group of stakeholders in society that hold responsibility over distraction. Guaranteeing the active participation and cooperation of major stakeholders such as government, industry, drivers, technology developers etc. in the prevention of distraction is a key task for road safety practitioners and policymakers.

Countermeasures for Distracted Driving

Many countermeasures have been developed in an attempt to prevent and mitigate distracted driving. Unfortunately, very few have been evaluated to establish their effectiveness. Table 2 includes countermeasures supported by empirical research, with a focus on those which have reduced the occurrence or impact of distracted driving. The countermeasures are organized following the hierarchy of controls for the management of risks. The most effective are those that focus on “Elimination” - the elimination of the risk - and the least effective are those that focus on the provision of Personal Protective Equipment (PPE) to protect the body from injury.

Looking to the Future

Distraction is likely to continue to be a road safety problem, even as vehicles become increasingly automated. For partially automated vehicles, there is currently no evidence that these reduce distraction. On the contrary, partial automation has been

Table 2 Countermeasures for distracted driving

<i>Types of countermeasures</i>	<i>Definition</i>	<i>Countermeasures</i>
Elimination	Involves complete physical removal of the hazard and risk it creates. This might be difficult to implement in certain circumstances such as distracted driving from mobile phone use, as drivers have reported difficulties separating from their phones (George et al., 2018). Nonetheless, elimination should always be considered and implemented before the other methods.	Fully Automated Vehicles (FAV) SAE Level 4-5: Vehicles that do not require a human operator to control or monitor the vehicle. Therefore, it is anticipated that as the driver will not have responsibility to control or monitor the vehicle distraction will not play a significant role. Projected benefits of these vehicles might only be observed in 25-30 years (Clark et al., 2016; Dia 2015).
Substitution	Involves substituting the risks with lesser hazardous practice or material. It must be implemented in the very early phases of development.	Workload managers: Driver-support systems that present driving and non-driving related information in such a way that road-users are not distracted. The technology could reduce distraction from driver-assistance systems (driving-related information; Teh et al., 2018), but the effectiveness of managing non-driving related information is unknown.
Engineering Controls	Involves controlling identified hazards through the modification or addition of physical safety features to equipment or machinery.	Road design: PIARC (2016) recommends using road engineering treatments such as concrete or steel side barriers, wire rope side and median barriers, speed humps, rough shoulders, variable speed signs, etc. to mitigate driver distraction. SAE Level 3 Partial Driving Automation Vehicle (also known as autopilot): The vehicle can perform key control tasks, but drivers are supposed to maintain steering and take over control of the vehicle at all times. Research suggests that drivers are likely to engage in distraction while driving partially automated vehicles (Lin et al., 2018), potentially resulting in road crashes. Advanced Driver-Assistance Systems (ADAS; available in SAE Level 1-2 vehicles): ADAS support drivers with some automation features (e.g., cruise control, lane centering, etc.) and can also provide warnings for upcoming collisions (e.g., forward collision warning, etc.), thus preventing and mitigating some distraction-related crashes. Evidence suggests that ADAS could, to the contrary, increase distraction through poorly designed warnings/alarms (Biondi et al., 2014; Thompson et al., 2018), and in-vehicle information systems (IVIS) can elicit means for drivers to engage in non-driving tasks (Oviedo-Trespalcacios et al., 2019d). Blocking technology: Includes mobile phone applications and hardware which seek to block mobile phone use while driving. This technology is effective in reducing mobile phone use while driving (Oviedo-Trespalcacios et al., 2020b; Albert and Lotan, 2019; Ponte et al., 2016), but acceptance is low among drivers (Oviedo-Trespalcacios et al., 2019e, 2020c; Ponte et al., 2016; Reagan and Cicchino, 2020). Feedback systems: Delivers information to drivers on their performance. Research has found feedback systems that consider parental norms and real-time feedback are associated with a smaller duration of off-road glances (Merrikhpour and Donmez, 2017). Legislation: Legislation banning the use of mobile phone devices whilst driving. Very few such bans have been evaluated, and there are mixed reports about the effectiveness of bans across jurisdictions. Police enforcement: Police operations to increase compliance of legislation banning different forms of mobile phone use while driving. Research has shown that police officers are unable to correctly enforce legislation in many circumstances due to lack of resources, poor visibility inside vehicle, etc. (Nevin et al., 2017; Rudisill et al., 2019). Driving for work procedures/ policies: Implementation of Workplace Health and Safety (WH&S) organizational
Administrative Controls	Involves changing the way individuals work through limiting exposure to a hazard. This method is not always reliable or effective as it is prone to variability of human performance.	

(continued)

Table 2 Countermeasures for distracted driving (*cont.*)

<i>Types of countermeasures</i>	<i>Definition</i>	<i>Countermeasures</i>
Personal Protective Equipment (PPE)	Involves protecting the body from injury (e.g., gloves, safety glasses, earplugs). PPE is deemed to be the least effective method as it does not eliminate the hazard. Regarding distracted driving, PPE corresponds to post-crash safety measures such as seatbelts, airbags and other passive safety elements to minimize the impact of a crash.	procedures/policies to reduce distracted driving and enhance safety of drivers. Research has found truck and bus drivers working for organizations that enforced texting bans have lower texting and driving prevalence in comparison to companies without bans (Hickman et al., 2010). Educational programs: These programs include strategies on how to reduce mobile phone use while driving. Some examples include “<i>It Can Wait</i>”, “<i>Just Drive—Take Action Against Distraction</i>”, and “<i>Steering Teens Safe</i>”. The programs were found to be effective to a certain extent. Training programs: The “<i>Forward Concentration and Attention Learning (FOCAL)</i>” research-led training program consists of three stages: pre-test, training, and post-test. Drivers who received the FOCAL training engaged in fewer in-vehicle glances longer than 2 s by roughly 25 percentage points when compared to the placebo group (Unverricht et al., 2019). Seatbelts, Vehicle Design, and Emergency Care: Prior research has found these measures to reduce the severity of crashes (Bhattacharyya and Layton, 1979).

shown to create distraction due to decreased engagement with the driving task (Cunningham and Regan 2018b; Kanaan et al., 2020; Regan et al., 2020). A study in China with Tesla drivers, for example, found that drivers often engage in distracting activities while using the autopilot system (Lin et al., 2018). Similar findings have been reported in the U.S., where drivers of vehicles with ADAS, such as adaptive cruise control, report engaging more in mobile phone use and texting (Dunn et al., 2019). Matthews et al., (2019) showed that autopilot systems elicit subjective symptoms of fatigue and loss of alertness that last even after the autopilot system has been deactivated. Likewise, as vehicles become more automated, the human-machine interfaces for the technologies that automate driving could themselves become sources of distraction. Evidence already exists showing that automation actions and alerts that are unexpected, because of a lack of training, lack of situational awareness, or some other mechanism, may create “automation surprises” (Hollnagel and Woods, 2005); and, in doing so, distract drivers. An understanding of, and development of countermeasures for, driver distraction at all levels of automation is a future concern for traffic safety.

Concluding Comments

In this chapter we have introduced the reader to the field of driver distraction: its definition and mechanisms, its impact on traffic safety, prevention approaches, and countermeasures. The focus of the chapter has been on driver distraction, although we acknowledge that there are other road users vulnerable to the effects of distraction, including riders and pedestrians. Greater coordination between stakeholders with responsibility for preventing and managing driver distraction, and a greater emphasis on the evaluation of distraction prevention and mitigation countermeasures, will be important overarching strategies in managing distraction into the future. As appropriately concluded by the European Commission (2018), countermeasures to distraction should be derived from the best available and reliable scientific knowledge of the underlying mechanisms of distraction, the prevalence of different distractive activities and the risks associated with these activities.

Acknowledgment

Dr Oscar Oviedo-Trespalacios’ contribution to this chapter was partially funded by an Australian Research Council Discovery Early Career Researcher Award [DE200101079].

See Also

Naturalistic Studies; Driving Behavior & Skills; Young Drivers; Driver Education and Training

References

- Albert, G., Lotan, T., 2019. Exploring the impact of "soft blocking" on smartphone usage of young drivers. *Accid. Anal. Prevent.* 125, 56–62.
- Bayly, M., Young, K.L., Regan, M.A., 2008. Sources of distraction inside the vehicle and their effects on driving performance. In: *Driver distraction: Theory, effects and mitigation*. CRC Press, p. 191.
- Beanland, V., Fitzharris, M., Young, K.L., Lenné, M.G., 2013. Driver inattention and driver distraction in serious casualty crashes: Data from the Australian National Crash In-depth Study. *Accid. Anal. Prevent.* 54, 99–107.
- Bhattacharyya, M.N., Layton, A.P., 1979. Effectiveness of seat belt legislation on the Queensland road toll—an Australian case study in intervention analysis. *J. Am. Stat. Assoc.* 74 (367), 596–603.
- Biondi, F., Rossi, R., Gastaldi, M., Mulatti, C., 2014. Beeping ADAS: Reflexive effect on drivers' behavior. *Transport. Res. Part F: Traffic Psychol. Behav.* 25, 27–33.
- Bruyas, M.P., 2017. Impact of mobile phone use on driving performance: review of experimental literature. In: Regan, M.A., Lee, J.D., Victor, T., W., (Eds.), *Driver Distraction and Inattention*. CRC Press.
- Chen, Y., Fu, R., Xu, Q., Yuan, W., 2020. Mobile Phone Use in a Car-Following Situation: Impact on Time Headway and Effectiveness of Driver's Rear-End Risk Compensation Behavior via a Driving Simulator Study. *Int. J. Environ. Res. Public Health* 17 (4), 1328.
- Clark, B., Parkhurst, G., Ricci, M., 2016. Understanding the socioeconomic adoption scenarios for autonomous vehicles: A literature review. Project report. University of the West of England, Bristol.
- Cunningham, M., Regan, M., 2018. Automated vehicles may encourage a new breed of distracted drivers. *The Conversation*.
- Dia, H., 2015. Driverless cars will change the way we think of car ownership. Retrieved from <https://theconversation.com/driverless-carswill-change-the-way-we-thinkof-car-ownership-50125>.
- Dingus, T.A., Guo, F., Lee, S., Antin, J.F., Perez, M., Buchanan-King, M., Hankey, J., 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proc. Natl. Acad. Sci.* 113 (10), 2636–2641.
- Dunn, N., Dingus, T., Socolich, S., 2019. Understanding the Impact of Technology: Do Advanced Driver Assistance and Semi-Automated Vehicle Systems Lead to Improper Driving Behavior?
- AAA Foundation for Traffic Safety. Retrieved from https://aaafoundation.org/wp-content/uploads/2019/12/19-0460_AAAFTS_VTTI-ADAS-Driver-Behavior-Report_Final-Report.pdf.
- Eby, D., Kostyniuk, L.P., 2003. Driver distraction and crashes: An assessment of crash databases and review of the literature.
- European Commission, 2018. Driver distraction 2018. European Commission, Brussels, Belgium.
- Fitch, G.M., Socolich, S.A., Guo, F., McClafferty, J., Fang, Y., Olson, R.L., Perez, M.A., Hanowski, R.J., Hankey, J.M., Dingus, T.A., 2013. The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk (No. DOT HS 811 757).
- Galitz, S., 2017. Killer cell phones and complacent companies: how apple fails to cure distracted driving fatalities. *U. Miami L. Rev.* 72, 880.
- George, A.M., Brown, P.M., Scholz, B., Scott-Parker, B., Rickwood, D., 2018. I need to skip a song because it sucks": Exploring mobile phone use while driving among young adults. *Transport. Res. part F: Traffic Psychol. Behav.* 58, 382–391.
- Gordon, C.P., 2009. Crash studies of driver distraction. In: Regan, M.A., Lee, J.D., Victor, T.W. (Eds.), *Driver Distraction and Inattention*. CRC Press.
- Kanaan, D., Donmez, B., Kelley-Baker, T., Popkin, S., Lehrer, A., Fisher, D.L., 2020. 9 Driver Fitness in the Resumption of Control. *Handbook of Human Factors for Automated, Connected, and Intelligent Vehicles*. CRC Press.
- Lee, J.D., Young, K.L., Regan, M.A., 2009. Defining Driver Distraction. In: Regan, M.A., Lee, J.D., Young, K. (Eds.), *Driver Distraction: Theory, Effects, and Mitigation*. CRC Press.
- Li, X., Oviedo-Trespalacios, O., Rakotonirainy, A., Yan, X., 2019. Collision risk management of cognitively distracted drivers in a car-following situation. *Transport. Res. Part F: Traffic Psychol. Behav.* 60, 288–298.
- Li, X., Oviedo-Trespalacios, O., Rakotonirainy, A., 2020. Drivers' gap acceptance behaviours at intersections: A driving simulator study to understand the impact of mobile phone visual-manual interactions. *Accid. Anal. Prevent.* 138, 105–486.
- Lin, R., Ma, L., Zhang, W., 2018. An interview study exploring Tesla drivers' behavioural adaptation. *Appl. Ergon.* 72, 37–47.
- Hallett, C., Regan, M.A. and Bruyas, M.P., 2011, September. Development and validation of a driver distraction impact assessment test. In *The electronic proceedings of the 2nd international conference on driver distraction and inattention*. pp. 5-7.
- Hickman, J.S., Hanowski, R.J., Bocanegra, J., 2010. Distraction in commercial trucks and buses: Assessing prevalence and risk in conjunction with crashes and near-crashes.
- Hoel, J., Jaffard, M., Van Elslande, P., 2010, April. Attentional competition between tasks and its implications. In *European Conference on Human Centred Design for Intelligent Transport Systems*, 2nd, 2010, Berlin, Germany.
- Hollnagel, E., Woods, D.D., 2005. 2005. Joint Cognitive Systems: Foundations of Cognitive Systems Engineering. CRC press.
- Horberry, T., Edquist, J., 2009. Distractions outside the vehicle. In: Regan, M.A., Lee, J.D., Young, K. (Eds.), *Driver distraction: Theory, effects, and Mitigation*, pp. 215.
- Huth, V., Brusque, C., 2014. Drivers' adaptation to mobile phone use: interaction strategies, consequences on driving behaviour and potential impact on road safety Driver Adaptation to Information and Assistance Systems: chapter 9. In: *Driver Adaptation to Information and Assistance Systems*.
- National Highway Traffic Safety Administration. (NHTSA), 2016. Visual-manual NHTSA driver distraction guidelines for portable and aftermarket devices. National Highway Traffic Safety Administration (NHTSA), Department of Transportation (DOT), Washington, DC.
- Nevin, P.E., Blunar, L., Kirk, A.P., Freedheim, A., Kaufman, R., Hitchcock, L., Maeser, J.D., Ebel, B.E., 2017. I wasn't texting; I was just reading an email . . .": a qualitative study of distracted driving enforcement in Washington State. *Injury Prevention* 23 (3), 165–170.
- Matthews, G., Neubauer, C., Saxby, D.J., Wohleber, R.W., Lin, J., 2019. Dangerous intersections? A review of studies of fatigue and distraction in the automated vehicle. *Accid. Anal. Prevent.* 126, 85–94.
- Merrikhpour, M., Donmez, B., 2017. Designing feedback to mitigate teen distracted driving: A social norms approach. *Accid. Anal. Prevent.* 104, 185–194.
- Onate-Vega, D., Oviedo-Trespalacios, O., King, M.J., 2020. How drivers adapt their behaviour to changes in task complexity: the role of secondary task demands and road environment factors. *Transportation research part F: traffic psychology and behaviour* 71, 145–156.
- Oviedo-Trespalacios, O., Haque, M.M., King, M., Washington, S., 2016. Understanding the impacts of mobile phone distraction on driving performance: A systematic review. *Transport Res. part C: Emerging Technol.* 72, 360–380.
- Oviedo-Trespalacios, O., Haque, M.M., King, M., Washington, S., 2017. Effects of road infrastructure and traffic complexity in speed adaptation behaviour of distracted drivers. *Accid. Anal. Prevent.* 101, 67–77.
- Oviedo-Trespalacios, O., Haque, M.M., King, M., Washington, S., 2019a. 2019a. Mate! I'm running 10 min late": An investigation into the self-regulation of mobile phone tasks while driving. *Accid. Anal. Prevent.* 122, 134–142.
- Oviedo-Trespalacios, O., Truelove, V., Watson, B., Hinton, J.A., 2019b. The impact of road advertising signs on driver behaviour and implications for road safety: A critical systematic review. *Transport. Res. Part A: Policy Pract.* 122, 85–98.
- Oviedo-Trespalacios, O., Nandavar, S., Newton, J.D.A., Demant, D., Phillips, J.G., 2019c. Problematic use of mobile phones in Australia . . . is it getting worse? *Front. Psychiatry* 10, 105.
- Oviedo-Trespalacios, O., Nandavar, S., Haworth, N., 2019d. How do perceptions of risk and other psychological factors influence the use of in-vehicle information systems (IVIS)? *Transport. Res. part F: Traffic Psychol. Behav.* 67, 113–122.
- Oviedo-Trespalacios, O., Williamson, A., King, M., 2019e. User preferences and design recommendations for voluntary smartphone applications to prevent distracted driving. *Transport. Res. part F: Traffic Psychol. Behav.* 64, 47–57.

- Oviedo-Trespalacios, O., Afghari, A.P., Haque, M.M., 2020a. A hierarchical Bayesian multivariate ordered model of distracted drivers' decision to initiate risk-compensating behaviour. *Anal. Methods Accid. Res.* 26, 100121.
- Oviedo-Trespalacios, O., Truelove, V., King, M., 2020b. 2020b. It is frustrating to not have control even though I know it's not legal!": A mixed-methods investigation on applications to prevent mobile phone use while driving. *Accid. Anal. Prevent.* 137, 105412.
- Oviedo-Trespalacios, O., Briant, O., Kaye, S.A., King, M., 2020c. Assessing driver acceptance of technology that reduces mobile phone use while driving: The case of mobile phone applications. *Accid. Anal. Prevent.* 135, 105348.
- PIARC, 2016. The Role of Road Engineering in Combatting Driver Distraction and Fatigue Road Safety Risks. Report 2016R24EN. World Road Association (PIARC), Paris, France.
- Parnell, K.J., Stanton, N.A., Plant, K.L., 2016. Exploring the mechanisms of distraction from in-vehicle technology: the development of the PARRC model. *Safety Sci.* 87, 25–37.
- Parnell, K.J., Stanton, N.A., Plant, K.L., 2017. What's the law got to do with it? Legislation regarding in-vehicle technology use and its impact on driver distraction. *Accid. Anal. Prevent.* 100, 1–14.
- Ponte, G., Baldock, M.R., Thompson, J.P., 2016. Examination of the effectiveness and acceptability of mobile phone blocking technology among drivers of corporate fleet vehicles (No. CASR140).
- Qin, L., Li, Z.R., Chen, Z., Bill, M.A., Noyce, D.A., 2019. 2019. Understanding driver distractions in fatal crashes: an exploratory empirical analysis. *J. Safety Res.* 69, 23–31.
- Reagan, I.J., Cicchino, J.B., 2020. Do Not Disturb While Driving—Use of cellphone blockers among adult drivers. *Safety Sci.* 128, 104753.
- Regan, M.A., Hallett, C., 2011. Driver distraction: definition, mechanisms, effects, and mitigation. In: Porter, B. (Ed.), *Handbook of Traffic Psychology*. Academic Press, pp. 275–286.
- Regan, M.A., Hallett, C., Gordon, C.P., 2011. Driver distraction and driver inattention: definition, relationship and taxonomy. *Accid. Anal. Prevent.* 43, 1771–1781.
- Regan, M.A., Prabhakaran, P., Wallace, P., Cunningham, M.L., Bennett, J.M., 2020. Education and training for drivers of assisted and automated vehicles (Austroads Report No. AP-R616-20).
- Rudisill, T.M., Baus, A.D., Jarrett, T., 2019. 2019. Challenges of enforcing cell phone use while driving laws among police: a qualitative study. *Injury Prevent.* 25 (6), 494–500.
- Saifuzzaman, M., Haque, M.M., Zheng, Z., Washington, S., 2015. Impact of mobile phone use on car-following behaviour of young drivers. *Accid. Anal. Prevent.* 82, 10–19.
- Sundfør, H.B., Sagberg, F., Høye, A., 2019. Inattention and distraction in fatal road crashes—results from in-depth crash investigations in Norway. *Accid. Anal. Prevent.* 125, 152–157.
- Teh, E., Jamson, S., Carsten, O., 2018. Design characteristics of a workload manager to aid drivers in safety-critical situations. *Cognit. Technol. Work* 20 (3), 401–412.
- Thompson, J.P., Mackenzie, J.R., Dutschke, J.K., Baldock, M.R., Raftery, S.J., Wall, J., 2018. A trial of retrofitted advisory collision avoidance technology in government fleet vehicles. *Accid. Anal. Prevent.* 115, 34–40.
- Tingvall, C., Haworth, N., 1999. Vision Zero—An ethical approach to safety and mobility, paper presented to the 6th International Conference Road Safety & Traffic Enforcement: Beyond 2000. In Department of Transportation Federal Highway Administration International Technology Exchange Program, April 2003.
- Tingvall, C., Eckstein, L., Hammer, M., 2009. Government and industry perspectives on driver distraction. In: Regan, M.A., Lee, J.D., Young, K. (Eds.), *Driver Distraction: Theory, Effects and Mitigation*. CRC Press, pp. 603–620.
- Treat, J.R., 1980. A study of precrash factors involved in traffic accidents. *HSRI Res. Rev.*
- Unverricht, J., Yamani, Y., Yahoodik, S., Chen, J., Horrey, W.J., 2019. November. Attention maintenance training: Are young drivers getting better or being more strategic?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 1991–1995). SAGE Publications, Sage CA: Los Angeles, CA.
- Van Elslande, P., Fouquet, K., 2007. Analyzing human functional failures' in road accidents. Final report. Deliverable D 5.
- Wallén Warner, H., Ljung, M., Sandin, J., Johansson, E., Björklund, G., 2008. Manual for DREAM 3.0, Driving Reliability and Error Analysis Method. Deliverable D5. 6 of the EU FP6 project SafetyNet, TREN-04-FP6TRSI2. 395465/506723. Chalmers University of Technology.
- Wijayarathna, K.P., Cunningham, M.L., Regan, M.A., Jian, S., Chand, S., Dixit, V.V., 2019. Mobile phone conversation distraction: Understanding differences in impact between simulator and naturalistic driving studies. *Accid. Anal. Prevent.* 129, 108–118.
- Young, R., 2017. Removing biases from crash odds ratio estimates of secondary tasks: A new analysis of the SHRP 2 naturalistic driving study data (No. 2017-01-1380). SAE Technical Paper.
- Young, K.L., Regan, M.A. and Lee, J.D (2009). Factors moderating the impact of distraction on driving performance and safety. In Regan, M.A., Lee, J.D. & Young, K. (Eds.), *Driver Distraction: Theory, Effects and Mitigation*. Florida, USA: CRC Press (Chapter 19)
- Young, K.L., Salmon, P.M., 2015. Sharing the responsibility for driver distraction across road transport systems: a systems approach to the management of distracted driving. *Accid. Anal. Prevent.* 74, 350–359.

Further Reading

- Fisher, D.L., Horrey, W.J., Lee, J.D., Regan, M.A. (Eds.), 2020. *Handbook of Human Factors for Automated, Connected, and Intelligent Vehicles*. CRC Press.
- Rupp, G., 2010. Performance metrics for assessing driver distraction (pp. 1–23). SAE.
- Parnell, K.J., Stanton, N.A., Plant, K.L., 2018. *Driver Distraction: A Sociotechnical Systems Approach*. CRC Press.
- Regan, M.A., Lee, J.D., Young, K. (Eds.), 2009. *Driver Distraction: Theory, Effects, and Mitigation*. CRC Press.
- Regan, M.A., Lee, J.D., Victor, T.W. (Eds.), 2013. *Driver Distraction and Inattention: Advances in Research and Countermeasures* (Vol. 1). Ashgate Publishing, Ltd..
- Oviedo-Trespalacios, O., 2018. 2018. Getting away with texting: Behavioural adaptation of drivers engaging in visual-manual tasks while driving. *Transp. Res. Policy Pract* 116, 112–121.

Driver Aggression and Anger

Mark J.M. Sullman*, Amanda N. Stephens[†], ^{*}Department of Social Sciences, University of Nicosia, Cyprus; [†]Monash University Accident Research Centre, Monash University, Clayton, Australia

© 2021 Elsevier Ltd. All rights reserved.

General Anger and Aggression	121
Aggressive Driving	121
Road Rage and Aggressive Driving	122
The Prevalence of Aggressive Driving	122
Is Aggressive Driving Increasing?	123
Aggressive Driving, Crash, and Injury Risk	123
Aggressive Drivers	124
The General Aggression Model for Aggressive Drivers	124
Driver and Situation Factors Leading to Anger and Aggression	125
Driver Anger and Appraisal Tendencies	126
Relationships Between Anger and Types of Aggression	126
Strategies for Reducing Anger and Aggressive Driving	126
Anger and Aggression in Autonomous Vehicles	127
Conclusions	127
References	127
Further Reading	129

General Anger and Aggression

The concepts of anger and aggression are two related but different constructs. The American Psychologists Association defines anger as “an emotion characterized by tension and hostility arising from frustration, real or imagined injury by another, or perceived injustice” (APA, 2020a). Aggression can be defined as “behavior aimed at harming others physically or psychologically” (APA 2020b). Thus, although strongly related, anger describes the emotion experienced, while aggression relates to the expression of anger in behavioral terms. The concept of driving anger originated in a clinical setting, where psychologists attempted to understand and treating problem anger in a driving context (Deffenbacher et al., 1986). These clinical findings spawned a program of research to explore whether the propensity to experience anger in the driving setting was a specific type of trait anger, which situations cause drivers to become angry, what drivers do when angry and the thoughts angry drivers have (Deffenbacher et al., 1994, 2002, 2003). Deffenbacher et al. (1994) defined driving anger as a situation specific form of trait anger, with research reporting low to moderate correlations with general trait anger (e.g., Deffenbacher et al., 1994). However, one of the most consistent findings has been that those higher in trait driving anger are more likely to become aggressive while driving (Berdoulat et al., 2013; Blankenship et al., 2013; Dahlen and White, 2006; Iversen and Rundmo, 2002; Lajunen et al., 1998; Lajunen and Parker, 2001; Li et al., 2014; Millar, 2007; Suhr and Nesbit, 2013).

Aggressive Driving

The term aggressive driving is used differently in the driving context to the definitions used by the APA for general aggression. For example, Lajunen and Parker (2001) distinguish between “aggressive driving” and “aggressive drivers” with the main difference being the ability to understand the intent of aggressive drivers; but not aggressive driving. The two most used definitions for aggressive driving reflect this distinction. The American Automobile Association (AAA, 2014) defines aggressive driving as: “Any unsafe driving behavior, performed deliberately and with ill intention or disregard for safety” (AAA Foundation for Traffic Safety, 2016); thus, requiring an understanding of motivations behind the behaviors. In contrast, the National Highway Traffic Safety Administration (NHTSA) define aggressive driving as “the operation of a motor vehicle that endangers or is likely to endanger persons or property” (Stuster, 2004). This definition removes the subjective element and is more suitable for enforcement purposes.

Behaviors that are considered aggressive, within these two definitions, include:

- Driving in excess of the speed limit (speeding);
- Following another driver too closely (tailgating);
- Dangerous overtaking;
- Weaving in and out of traffic;
- Preventing other drivers from merging or passing;

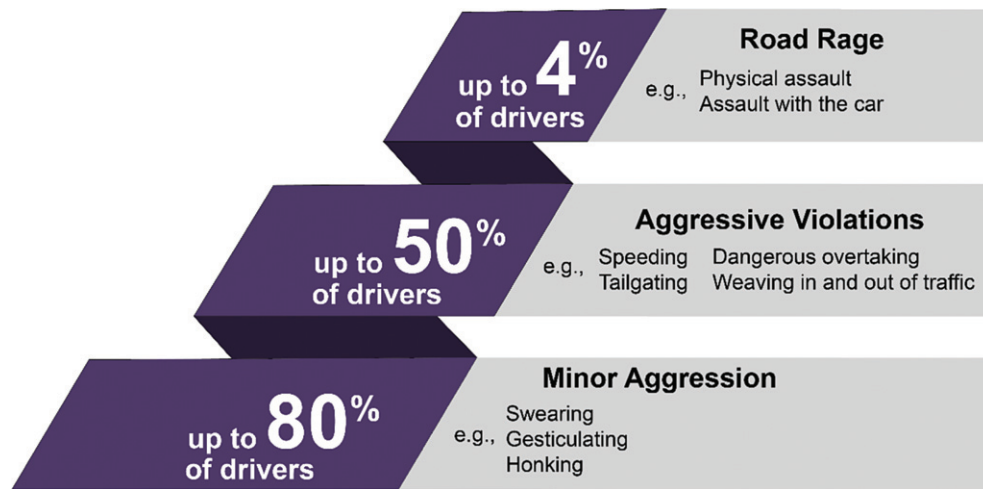


Figure 1 Types of aggressive driving and general prevalence. Source: NB Prevalence taken from AAA Foundation for Traffic Safety (2016), Stephens and Fitzharris (2019).

However, aggressive acts can range on a continuum from seemingly benign actions (such as mild gesticulating or swearing under one's breath) to extreme acts of aggression (see Fig. 1). Neither of these is considered within the above definitions, yet it is often these ends of the spectrum that drivers and researchers refer to as aggressive driving. Thus, a paradox in this field is that while most drivers can easily recount a time when another driver has been aggressive towards them, the definition of aggressive driving is not consistent across drivers, researchers or road safety agencies.

Road Rage and Aggressive Driving

A further inconsistency in aggressive driving research is the use of the term road rage. A number of researchers use road rage synonymously with aggressive driving, referring to minor acts of horn honking, or swearing as road rage. However, this oversimplifies a complex problem and can lead to a disregard of aggression as a genuine road safety concern (Elliott, 1999).

The NHTSA defines road rage separately as a criminal offense, while aggressive driving is a traffic violation (Stuster, 2004). This definition recognizes the extreme and violent nature of road rage, which could be deemed assault. As can be seen in Fig. 1, we propose that aggressive behaviors should be considered across two broad categories and road rage as an extreme separate construct. Thus, we suggest aggressive driving may include:

- minor aggressive behaviors, which are not enforceable and may constitute culturally common behaviors, such as swearing or gesticulating, horn honking, or flashing lights;
- enforceable aggressive driving behaviors that, as per the NHTSA definition, includes the use of a vehicle that endangers or is likely to endanger other road users and, as per the AAA definition, may be the results of ill intention or a disregard for safety.

These two categories may both be considered within the aggressive driving framework, particularly for research purposes. In addition, we propose that, as per the NHTSA definition, road rage can be defined as:

- Extreme acts of violence punishable as a criminal offense, which includes physical assault.

Road rage is, therefore, not considered to be aggressive driving and will not be included in the following literature review.

The Prevalence of Aggressive Driving

The inconsistent definitions for aggressive driving make it hard to understand its prevalence. In addition, as certain proxy behaviors (i.e., speeding) are considered to be road safety issues in their own right, the prevalence of aggressive driving may be understated. Nonetheless, research consistently shows that most drivers will experience aggression from other drivers at some point, and that the majority of drivers will also act aggressively themselves (Europe, 2003).

The largest global survey on aggressive driving was conducted in 2003 and included over 13,000 drivers from 23 countries (Europe, 2003). These included countries in Europe, the UK, the USA, South America, Russia, and Australasia. The results showed that, globally, it was common to receive aggression from others, with between 26% and 66% of respondents in each country reporting being the victim of aggression in the previous year. The types of aggression surveyed included minor aggressive acts, such as verbal abuse, gesticulating and flashing lights—which were the most frequently reported—to enforceable aggression, such as being pursued by another driver, which was considerably less frequent. Physical attacks were the least common with 1%–10% of those

who had reported receiving aggression, reporting it in this form. Interestingly, the types of aggression most frequently received differed across countries, suggesting a cultural element to aggressive driving and/or the perception of others' behavior as aggressive.

Perpetrating aggressive driving behavior is as common, if not more common as receiving it. In the global study, between 36% and 69% of drivers had been aggressive at least once in the previous year (Europe, 2003). More recent national studies have shown that most drivers will be aggressive at least once in a 12-month period. Data from the USA and Australia show that over 80% of drivers reported aggression (AAA Foundation for Traffic Safety, 2016; Stephens and Fitzharris, 2019), which included behaviors ranging from minor aggression to enforceable aggressive behaviors (see Fig. 1).

As would be expected, when considered across specific behaviors, the prevalence of aggressive acts decreases as the severity of the act increases. For example, Stephens and Fitzharris (2019) reported that 71% of Australian drivers sound their horns when angry; while more serious aggression such as tailgating were reported by 45% of drivers, and pursuing another driver when angry reported by 10% of drivers. Likewise, 51% of drivers in the USA reported tailgating, with 12% deliberately cutting off another vehicle and an alarming 4% exiting the vehicle to confront another driver (or to engage in road rage) (AAA Foundation for Traffic Safety, 2016). Interestingly, in the USA, sounding the horn was reported by 45% of the sample, indicating less drivers engaged in horn honking than tailgating or yelling at another driver (47%). This again highlights a culturally specific element of aggression and the repertoire of behaviors that are commonly used to express anger.

While concerning, these figures do not represent the frequency with which drivers engage in the behaviors every time they drive. Most drivers do not "always" act aggressively. For example, in a Canadian report of behaviors exhibited "often" while driving, 20% of drivers swore, while only 6% used their horn aggressively (Vanlaar et al., 2007). Thus, there is a notable decrease in prevalence when considering how often each driver engages in the behavior, rather than just whether they engage in the behavior or not.

Overall, the research shows that aggression remains common on the roads. However, in which aggression is displayed and the frequency of this will differ across driving cultures. Minor acts of aggression are more common than more extreme aggression, although a notable proportion of drivers still engage in dangerous aggression, just relatively infrequently.

Is Aggressive Driving Increasing?

While many drivers perceive that aggression is increasing (AAA Foundation for Traffic Safety, 2016; Vanlaar et al., 2008), there is limited evidence to suggest this is the case. As reported above, the general population prevalence of aggressive acts are similar for studies published in the 2000s (e.g., Europe, 2003) and more recent reports (e.g., AAA Foundation for Traffic Safety, 2016; Stephens and Fitzharris, 2019). Vanlaar et al. (2007) compared self-reported frequencies of aggression between different samples of drivers in Canada taken in 2001 and 2006 and showed no significant differences across the two time points. This was evident across minor aggressions, such as swearing, sounding the horn and gesticulating, as well as for enforceable aggression, such as speeding through red lights.

Møller and Haustein (2018) found slight increases in aggression, when comparing self-reports from independent samples of drivers in Denmark measured in 2005 and 2016. A higher percentage of drivers in the 2016 sample, compared to the 2005 sample, reported having yelled at another road user (19% cf 12%), gesticulating (13% cf 10%) and hitting another vehicle (2% cf <1%). However, there were no differences in the percentages of drivers reporting hitting other road users or threatening other road users. The main differences between the studies by Møller and Haustein (2018) and Vanlaar et al. (2007) was the frequency response rate, with Møller and Haustein (2018) comparing drivers who had engaged in each behavior over the previous 12 months against those who had not. In comparison, Vanlaar et al. (2007) compared the percentage of drivers who reported engaging in the behaviors often in the previous 12 months. Thus, there is no evidence to suggest that aggressive drivers increase the frequency of these behaviors over time, and only limited evidence to suggest that more drivers are adopting certain aggressive behaviors.

Aggressive Driving, Crash, and Injury Risk

Despite the inconsistency in defining aggression, and in understanding its prevalence, aggressive driving is consistently associated with crash risk. This finding is robust across different data collection methods.

Survey research has shown weak relationships between self-reported aggressive driving and self-reported involvement in crashes (Dahlen et al., 2012; Stephens and Sullman, 2015; Sullman et al., 2007, 2015, 2019). This is also evident for professional drivers (Sullman et al., 2002, 2013). A large population study of almost 13,000 Canadian drivers showed associations between the severity of aggression and crash involvement (Wickens et al., 2016). In their study, drivers reporting no aggression had lower odds of being crash involved, compared to those reporting minor aggression (OR = 0.65, 95% CI: 0.54–0.79), whereas those reporting both minor (e.g., shouting, swearing, or gesticulating) and serious driver aggression (e.g., threatening to hurt a driver or passenger in another vehicle) had around 80% higher odds of crash involvement, compared to those reporting minor aggression only (OR: 1.78; 95% CI: 1.04–3.03).¹

There has been criticism of self-report studies that aim to understand the relationships between behavior and crash outcomes.¹ The primary arguments are that self-reports may reflect (1) the frequency of driving, not the behavior and (2) are open to

¹ We refer the interested reader to the published commentary debate regarding the driver behaviour questionnaire and relationship with crashes (see DE WINTER, J. & DODOU, D. 2012. Response to second commentary on "The Driver Behaviour Questionnaire as a predictor of accidents: A meta-analysis". *Journal of safety research*

confirmatory bias leading to under-reporting, which is reflected in weak relationships between behavior and crashes. De Winter and Dodou (2012) provide persuasive evidence regarding the validity of the relationships that have emerged between self-reported driver behavior and crash outcomes, demonstrating that the relationships that emerge are weak because crash involvement is relatively rare. Specific to aggression, the relationships that have been found in survey research have been supported through methodologies using objective behavioral measurements, such as driving simulators (Deffenbacher et al., 2003; Stephens and Groeger, 2009), microtraffic simulations (Habtemichael and de Picado Santos, 2014), naturalistic driving studies (Dingus et al., 2016) and crash data analysis (Conner and Smith, 2014; Paleti et al., 2010).

Using objective crash data, aggressive driving has been shown to significantly increase the odds of crash involvement, with estimates ranging from a threefold increase (Habtemichael and de Picado Santos, 2014) to a 35-fold increase (Dingus et al., 2016). Furthermore, microscopic traffic simulation has shown drivers are three to six times more likely to crash when being aggressive (Habtemichael and de Picado Santos, 2014). Considered across specific behaviors, according to Habtemichael and de Picado Santos (2014), the odds of crashing are fourfold when speeding or weaving in and out of traffic and sixfold when tailgating.

Particularly compelling evidence for the relationship between aggression and crash involvement comes from Naturalistic Driving Studies (NDS), in particular the Second Strategic Highway Research Program (SHRP2) conducted in the USA (Antin, 2011). SHRP2 included vehicle sensor data and video data both from within and outside the vehicle, thereby offering a rich source of objective information regarding the types of aggressive events and the associated effects on driving performance (e.g., crashes). This highlights the fact that NDS are the most ecologically valid method to understand aggression within the driving context. Using the SHRP2 data, Dingus et al. (2016) found that drivers have 35 times the odds of being crash involved (95% CI = 17–70) when aggressive, than they do when not driving aggressively. These odds are comparable to when driving impaired by drugs or alcohol (36 times the odds [OR]; 95% CI = 17–76) and are considerably higher than overall distraction (OR: 2; 95% CI = 1.8–2.4) or distraction when dialing a hand-held mobile phone (OR: 12; 95% CI: 6–26). The confidence intervals indicate that, compared to distraction, aggression was a less frequent behavior, but with greater risks. However, considerably more research has been undertaken into driver distraction than driver aggression,² despite the increased risk presented by the latter.

Aggressive driving not only contributes to crash risk, but has also been shown as an exacerbating factor for injuries resulting from a crash (Conner and Smith, 2014; Paleti et al., 2010). Conner and Smith (2014) used hospital and crash records to estimate that 15% of people involved in aggressive driving crashes in Ohio (2005 and 2009) sustained injuries requiring hospitalization. Paleti et al. (2010) used data collected from 2005 to 2007 for the National Motor Vehicle Crash Causation Study in the USA to understand the relationship between aggressive driving and injury severity. They estimated that 31% of all nonincapacitating injuries and 14% all incapacitating or fatal injuries that occurred during the period could be attributed to crashes from aggressive driving. Therefore, there is consistent evidence to show aggressive driving contributes to crashes and to crashes involving injury.

Aggressive Drivers

Aggressive drivers tend to be younger and male (AAA Foundation for Traffic Safety, 2016; Conner and Smith, 2014; Deffenbacher et al., 2002, 2003; Stephens and Sullman, 2015; Sullman, 2015; Sullman et al., 2013, 2015, 2018; Paleti et al., 2010; Wickens et al., 2011). Indeed young males also tend to be over-represented in crash statistics (Li et al., 2003).

Unlike aggressive driving behaviors, research into aggressive drivers often focuses on the individual's intention or motivations for their behaviors. Aggression can be considered as a response to anger-provoking situations (Deffenbacher et al., 2002) or may form part of a more problematic driving style. The latter may include other illegal behavior patterns, such as drink driving (Paleti et al., 2010), speeding and mobile phone use (Stephens and Fitzharris, 2019; Wickens et al., 2011).

As highlighted in the prevalence statistics, aggressive drivers are unlikely to be aggressive every time they drive. Instead, aggression is a dynamic concept that relies on a unique combination of the person and the situation (Stephens and Groeger, 2009). This is best described using the general Aggression Model (GAM; Anderson and Bushman, 2002; see Fig. 2)

The General Aggression Model for Aggressive Drivers

As can be seen in the GAM (Anderson and Bushman, 2002), aggression relies on the emotional response a driver has to a situation (routes). This response will differ according to who the driver is and the circumstances of the situation (inputs). The resulting behavior (outcomes) will also depend upon the emotion-based appraisals made by the driver, who will evaluate what response is appropriate and the risk involved. An additional element of the GAM is that aggression provides a feedback loop, whereby successful outcomes, for example speeding past a slower driver, are likely to reinforce future aggressive behaviors.

Few studies have comprehensively examined aggressive drivers through the GAM model. However, a large amount of research has been conducted to understand isolated components of this model. This includes driver and situation factors leading to anger and aggression, how anger influences appraisals relating to the situation and appropriate responses, and the relationships between anger and aggression. Each of these are discussed below.

² *Scopus search on "aggressive driving" revealed 2,541 documents, compared to "driver distraction" with 5,915 (performed March 2020)

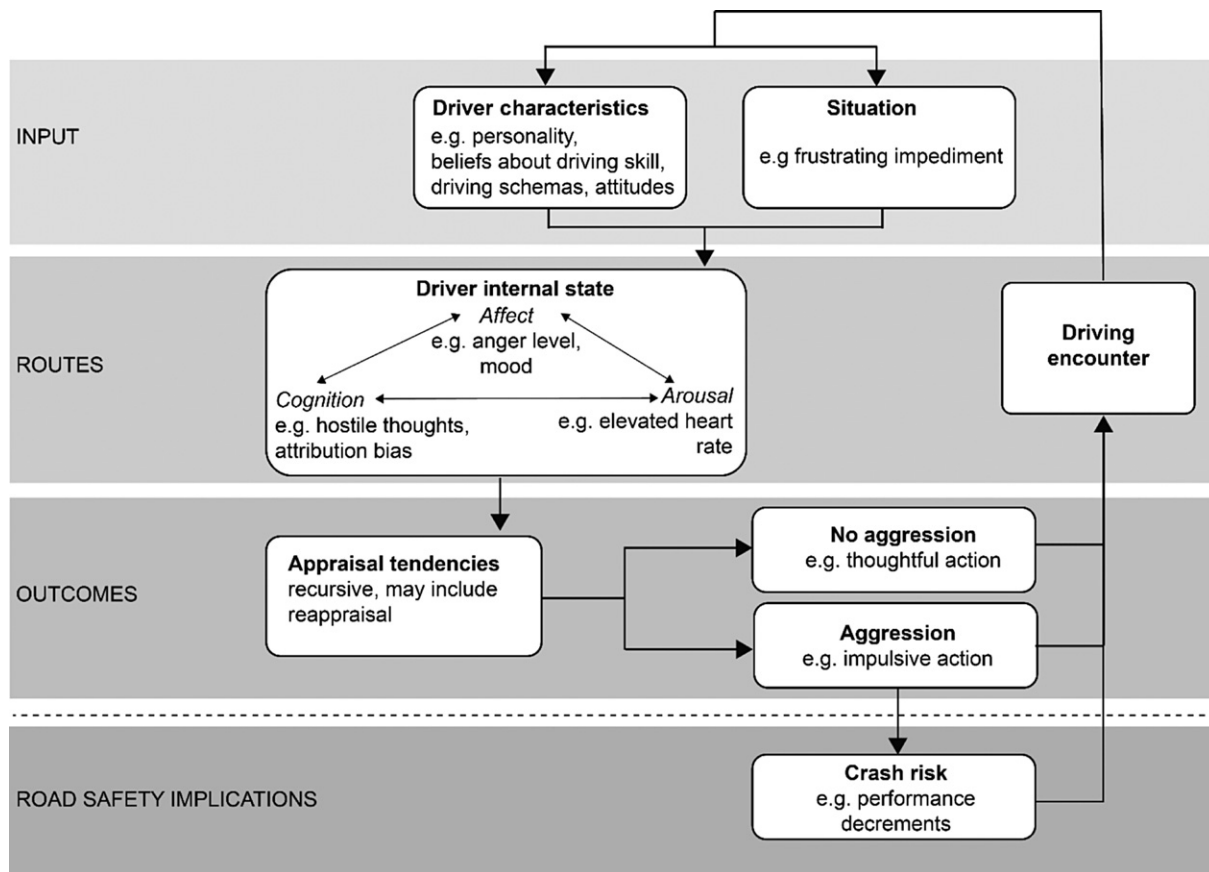


Figure 2 Representation of the General Aggression Model (Anderson and Bushman, 2002) for aggressive driving.

Driver and Situation Factors Leading to Anger and Aggression

Certain types of drivers are more prone to anger, and certain situations are more anger provoking than others (Deffenbacher et al., 1994). The Driving Anger Scale (DAS; Deffenbacher et al., 1994) is a widely used measure to understand a driver's tendency to become angry while driving (referred to as trait driving anger).³ The DAS measures the likelihood of becoming angry across six types of situations: slower lead drivers, traffic delays, hostile gestures from other drivers, police presence, discourtesy from others and illegal driving. The DAS has been used extensively to understand trait driving anger in a number of countries, including: New Zealand (Sullman, 2006), Europe (Björklund, 2008; Villieux and Delhomme, 2007), the UK (Lajunen et al., 1998), Asia (Li et al., 2014; Sullman et al., 2014), Turkey (Sullman et al., 2013), and Ukraine (Sullman et al., 2017). These studies have consistently shown that younger drivers tend to report more anger propensities than older drivers. This aligns with the age differences noted in both the prevalence of aggressive driving and overall crash statistics. Interestingly, when gender differences are found for anger tendencies, it is women who report higher anger than men (see Deffenbacher et al., 2016; Sullman, 2006). This misaligns with the fact that men tend to exhibit more aggressive behaviors than women and suggests other factors mediate the relationship between anger and aggression.

Research using the DAS has also shown that anger provoking situations vary across traffic cultures, and as such some researchers report different factor structures of the DAS (see Deffenbacher et al., 2016). For example, many researchers find police presence induces low levels of anger. However, situations that consistently provoke anger tend to involve progress impediments, perceived risk and hostility from other drivers (Deffenbacher et al., 2016). Stephens et al. (2019) recently developed the Measure for Angry Drivers to assess anger in modern day driving. They concluded that the three main types of anger-provocation situations continued to be progress impediments, danger from others and hostility. Thus, while the driving task and environment has changed considerably since the DAS was developed, the types of anger provoking situations have remained relatively constant. The only change is that some triggers relating to modern driving (i.e. other's being distracted at, traffic lights, by their mobile phone) are now included in the broader classifications, such as progress impediments.

In-line with the GAM, which suggests anger results from a combination of driver and situational factors, relationships between overall anger tendencies and general anger experienced while driving, tend to be moderate (Deffenbacher et al., 2003). Moreover,

³ We refer the interested reader to DEFFENBACHER, J., STEPHENS, A. & SULLMAN, M. J. 2016. Driving anger as a psychological construct: Twenty years of research using the Driving Anger Scale. *Transportation research part F: traffic psychology behaviour*, 42, 236-247. This provides a review of the research using the DAS

anger prone drivers do not report anger across every situation. Stephens and Groeger (2009) measured levels of anger while drivers drove in a simulator and found that anger propensities were unrelated to reported anger when driving in anger provoking situations. However, in relatively innocuous situations, drivers more prone to anger did report higher levels of anger. This again supports the theory that the relationship between trait anger and anger experienced is mediated by individual and situational factors.

Other driver characteristics can also influence the level of anger experienced and subsequent aggression. Personality traits related to both include: impulsiveness (Dahlen et al., 2012), sensation seeking (Stephens and Sullman, 2015), rumination (Suhr and Nesbit, 2013) and aggressive tendencies (Berdoulat et al., 2013). Lower levels of agreeableness (Dahlen et al., 2012), conscientiousness and emotional stability (Chraif et al., 2016) and higher levels of neuroticism have also been found to be associated with higher levels of anger and aggression (Iancu et al., 2016).

Driver Anger and Appraisal Tendencies

Not all angry drivers will display their anger aggressively. According to the GAM, this is due to the appraisal tendencies made while driving. The initial appraisals are instantaneous and rely on quick assessments of the situation in terms of its relevance and inconvenience to the driver, who is to blame and how legitimate their behavior was (Lerner and Keltner, 2001). The second stage of the appraisal process assesses perceived control and ability to overcome the situation through action. These processes often rely on pre-existing schemas or stereotypes (Allen and Anderson, 2017).

Driver appraisals are also influenced by mood. When angry, drivers may rely on more stereotypical evaluations (Stephens and Groeger, 2011) and make evaluations based on their current mood, rather than the circumstances. One example of this is that drivers in an angry mood are often more likely to have hostile thoughts about other drivers, or to attribute blame for situations that are beyond their control (Stephens and Groeger, 2014). This is exacerbated by the fact that many drivers tend to perceive other drivers as less skilled (Stephens et al., 2016b) and act aggressively to those of lower status (Stephens and Groeger, 2014). Thus, the appraisals made by drivers, influenced by driver state, determine whether there is an aggressive outcome and the form that this takes. These appraisals can be more significant in the determination of aggression, than one's anger tendencies (Stephens et al., 2016a).

Relationships Between Anger and Types of Aggression

A common tool to measure self-reported aggression is the Driving Anger Expression Inventory (DAX; Deffenbacher et al., 2002). The DAX has three types of aggressive responses to anger: verbal aggression (such as yelling at another driver); personal physical aggression (such as getting out of the vehicle to confront another driver); and using the vehicle to express anger (such as following too closely or speeding up). The DAX also includes items to measure adaptive or constructive responses to anger (such as telling yourself to ignore it). To date, there has been an overemphasis on the dangerous types of aggression, with relatively little research on predicting positive behaviors. However, research consistently shows that drivers most commonly try to deal with anger in a positive or constructive manner (Deffenbacher et al., 2002; Sullman, 2015; Sullman et al., 2013, 2015), with the aggressive forms of responses being less frequent.

Two recent meta-analyses reviewed the anger-aggression relationships (Bogdan et al., 2016; Zhang and Chan, 2016). Bogdan et al. (2016) compared 51 studies using self-reported anger tendencies with aggressive expressions of anger. Several of the included studies measured aggression using the DAX. They found that the strongest relationships were between trait anger and verbal expressions of anger, which they argued may be due to the ease of this form of communication. It may also be that other more aggressive forms of responses are often appraised as not being appropriate or viable for the situation, while verbal aggression is a versatile response type. Zhang and Chan (2016) examined the strength of relationships between anger and aberrant driving behaviors, including aggression. They also included 51 studies in their review and found that trait driving anger was related to aggressive, risky driving and driving errors, with the strongest relationship being between anger and aggressive driving. In both of these meta-analyses only weak to moderate relationships were noted. Furthermore, these consider only general tendencies for behavior, rather than more situation specific responses. Thus, while these studies could be more nuanced, they still provide a useful tool to understand the contribution of trait anger to different types of aggressive responses.

Strategies for Reducing Anger and Aggressive Driving

The GAM shows us that it is not the situation that leads to anger and aggression, but how the driver evaluates the situation. Thus, effective strategies for the reduction of anger or aggression lie in targeting the appraisal process. This can be achieved through cognitive strategies and training, key messaging and enforcement.

Cognitive strategies focus on changing the stereotypical response or negative schemas regarding other drivers and their behaviors⁴ (Deffenbacher, 2016). These may also include shifting the focus of drivers. Indeed, Stephens and Groeger (2009) found that

⁴ For a full description of cognitive strategies, we refer the interested reader to DEFFENBACHER, J. L. 2016. A review of interventions for the reduction of driving anger. *Transportation research part F: traffic psychology and behaviour*, 42, 411-421.

drivers who were asked to focus on the anger provocation in each situation had worse driving performance than those asked to rate the danger or difficulty. Mindfulness or resilience training may also offer promising results for the reduction of emotional reactivity. Increased mindfulness has been associated with less anger and aggression in drivers (Kazemeini et al., 2013; Stephens et al., 2018), while increased resilience has been associated with more adaptive ways of dealing with anger (Gras et al., 2016).

Enforcement also plays a key role in behavior change. The NHTSA found targeted enforcement led to a reduction in the proportion of aggression-as-the-primary-factor in crashes (Stuster, 2004). Evidence from other illegal behaviors, such as speeding and drink driving shows that drivers who engage in these behaviors do so because they do not perceive they will get caught and also underestimate the risk (Stephens et al., 2017a, 2017b). However, the risk of aggressive driving may be less clear, as a result of the inconsistency in defining it. This highlights the importance of clear definitions and proxy behaviors to operationalize it, as well as the need for this to be communicated to drivers.

Anger and Aggression in Autonomous Vehicles

While the role of anger in aggressive driving is now well established, it is less clear how emotions will influence driver operation in autonomous driving. This will be particularly relevant in the lower stages of automation, where drivers will still be required to take-over the driving task. There are some data to suggest that AV may provide a promising solution to the aggressive driving problem. Research has shown that drivers who acknowledge being angry and aggressive would readily accept in-vehicle systems designed to intervene and dissipate potential aggression in manual vehicles (Li et al., 2020). In particular, drivers report that auditory warnings that have a family member's voice would assist in reducing aggressive responses. However, studies of mood during AV driving have shown that a driver's mood can deteriorate during complex driving (Techer et al., 2019). The effects of this on take-over responses are not yet known. Finally, we do not know whether AV will be the solution for anger and aggression, or a tool that exacerbates this.

Conclusions

Aggression and driving anger are genuine road safety concerns, but there is currently no agreed definition of aggression. Road safety agencies use proxy behaviors that can be indicative of aggression to understand the problem, but these proxy behaviors vary across agencies and drivers may not know these behaviors are enforceable. In contrast, researchers often focus on aggressive drivers, trying to understand factors associated with their deliberate aggressive behaviors. Although both approaches are equally important, understanding intent allows the development of strategies to reduce aggression. These strategies are likely to be effective in changing how drivers appraise and respond to various driving situations. Nevertheless, enforcement can only be carried out on observed behaviors, irrespective of intent. Finally, we do not currently know the role that anger and aggression will play in transitioning or adopting AV; however, the evidence gathered to date suggests that AVs can be used to support drivers and to reduce anger and aggression.

References

- AAA Foundation For Traffic Safety, 2016. Prevalence of Self-Reported Aggressive Driving Behavior: United States, 2014.
- APA, 2020a. APA Dictionary of Psychology. Retrieved from <https://dictionary.apa.org/anger>.
- APA, 2020b. APA Dictionary of Psychology. Retrieved from <https://dictionary.apa.org/aggression>.
- Allen, J.J., Anderson, C.A., 2017. General aggression model. *The International Encyclopedia of Media Effects* 1–15.
- Anderson, C.A., Bushman, B.J., 2002. Human aggression. *Annu. Rev. Psychol.* 53.
- Antin, J.F., 2011. Design of the in-vehicle driving behavior and crash risk study: in support of the SHRP 2 naturalistic driving study, Transportation Research Board.
- Berdoulat, E., Vavassori, D., Sastre, M.T.M., 2013. Driving anger, emotional and instrumental aggressiveness, and impulsiveness in the prediction of aggressive and transgressive driving. *Accid. Anal. Prevent.* 50, 758–767.
- Björklund, G.M., 2008. Driver irritation and aggressive behaviour. *Accid. Anal. Prevent.* 40, 1069–1077.
- Bogdan, S.R., Măirean, C., Havârneanu, C.-E., 2016. A meta-analysis of the association between anger and aggressive driving. *Transport. Res. Part F: Traffic Psychol. Behav.* 42, 350–364.
- Chraïf, M., Anîşiei, M., Burtăverde, V., Mihăilescu, T., 2016. The link between personality, aggressive driving, and risky driving outcomes—testing a theoretical model. *J. Risk Res.* 19, 780–797.
- Conner, K.A., Smith, G.A., 2014. The impact of aggressive driving-related injuries in Ohio, 2004–2009. *J. Safety Res.* 51, 23–31.
- Dahlen, E.R., Edwards, B.D., Tubré, T., Zyphur, M.J., Warren, C.R., 2012. Taking a look behind the wheel: An investigation into the personality predictors of aggressive driving. *Accid. Anal. Prevent.* 45, 1–9.
- Dahlen, E.R., White, R.P., 2006. The Big Five factors, sensation seeking, and driving anger in the prediction of unsafe driving. *Personal. Individual Diff.* 41 (5), 903–915.
- de Winter, J., Dodou, D., 2012. Response to second commentary on "The Driver Behaviour Questionnaire as a predictor of accidents: a meta-analysis". *J. Safety Res.* 43, 94–98.
- Deffenbacher, J.L., Demm, P.M., Brandon, A.D., 1986. High general anger: Correlates and treatment. *Behav. Res. Ther.* 24 (4), 481–489.
- Deffenbacher, J., Stephens, A., Sullman, M.J., 2016. Driving anger as a psychological construct: twenty years of research using the Driving Anger Scale. *Transport. Res. Part F: Traffic Psychol. Behav.* 42, 236–247.
- Deffenbacher, J.L., 2016. A review of interventions for the reduction of driving anger. *Transport. Res. Part F: Traffic Psychol. Behav.* 42, 411–421.
- Deffenbacher, J.L., Deffenbacher, D.M., Lynch, R.S., Richards, T.L., 2003. Anger, aggression, and risky behavior: a comparison of high and low anger drivers. *Behav. Res. Therapy* 41, 701–718.

- Deffenbacher, J.L., Lynch, R.S., Oetting, E.R., Swaim, R.C., 2002. The driving anger expression inventory: a measure of how people express their anger on the road. *Behav. Res. Therapy* 40, 717–737.
- Deffenbacher, J.L., Oetting, E.R., Lynch, R.S., 1994. Development of a driving anger scale. *Psychol. Rep.* 74, 83–91.
- Deffenbacher, J.L., Petrilli, R.T., Lynch, R.S., Oetting, E.R., Swaim, R.C., 2003. The driver's angry thoughts questionnaire: a measure of angry cognitions when driving. *Cognit. Therapy Res.* 27 (4), 383–402.
- Dingus, T.A., Guo, F., Lee, S., Antin, J.F., Perez, M., Buchanan-King, M., Hankey, J., 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proc. Natl. Acad. Sci.* 113, 2636–2641.
- Elliott, B., 1999. Road rage: Media hype or serious road safety issue?.
- Europe, E.G., 2003. Aggressive behaviour behind the wheel.
- Gras, M.-E., Font-Mayolas, S., Patiño, J., Baltasar, A., Planes, M., Sullman, M.J.M., 2016. Resilience and the expression of driving anger. *Transport. Res. Part F: Traffic Psychol. Behav.* 42, 307–316.
- Habtemichael, F.G., de Picado Santos, L., 2014. Crash risk evaluation of aggressive driving on motorways: Microscopic traffic simulation approach. *Transport. Res. Part F: Traffic Psychol. Behav.* 23, 101–112.
- Iancu, A.E., Hoge, A., Olteanu, A.F., 2016. The association between personality and aggressive driving: A meta-analysis. *Romanian J. Psychol.* 18, 24–32.
- Iversen, H., Rundmo, T., 2002. Personality, risky driving and accident involvement among Norwegian drivers. *Personal. Individual Diff.* 33 (8), 1251–1263.
- Kazemeini, T., Ghanbari-E-Hashem-Abadi, B., Safarzadeh, A., 2013. Mindfulness based cognitive group therapy vs cognitive behavioral group therapy as a treatment for driving anger and aggression in Iranian taxi drivers. *Psychology* 4, 638.
- Lajunen, T., Parker, D., 2001. Are aggressive people aggressive drivers? A study of the relationship between self-reported general aggressiveness, driver anger and aggressive driving. *Accid. Anal. Prevent.* 33, 243–255.
- Lajunen, T., Parker, D., 1998. Dimensions of driver anger, aggressive and highway code violations and their mediation by safety orientation in UK drivers. *Transport. Res. Part F: Traffic Psychol. Behav.* 1, 107–121.
- Lerner, J.S., Keltner, D., 2001. Fear, anger, and risk. *J. Personal. Soc. Psychol.* 81, 146.
- Li, F., Yao, X., Jiang, L., Li, Y., 2014. Driving anger in China: psychometric properties of the Driving Anger Scale (DAS) and its relationship with aggressive driving. *Personal. Individual Diff.* 68, 130–135.
- Li, G., Braver, E.R., Chen, L.-H., 2003. Fragility versus excessive crash involvement as determinants of high death rates per vehicle-mile of travel among older drivers. *Accid. Anal. Prevent.* 35, 227–235.
- Li, S., Zhang, T., Liu, N., Zhang, W., Tao, D., Wang, Z., 2020. Drivers' attitudes, preference, and acceptance of in-vehicle anger intervention systems and their relationships to demographic and personality characteristics. *Int. J. Ind. Ergon.* 75, 102899.
- Millar, M., 2007. The influence of public self-consciousness and anger on aggressive driving. *Personal. Individual Diff.* 43 (8), 2116–2126.
- Møller, M., Hausteine, S., 2018. Road anger expression—changes over time and attributed reasons. *Accid. Anal. Prevent.* 119, 29–36.
- Paleti, R., Eluru, N., Bhat, C.R., 2010. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accid. Anal. Prevent.* 42, 1839–1854.
- Stephens, A., Bishop, C., Liu, S., Fitzharris, M., 2017. Alcohol consumption patterns and attitudes toward drink-drive behaviours and road safety enforcement strategies. *Accid. Anal. Prevent.* 98, 241–251.
- Stephens, A., Fitzharris, M., 2019. The frequency and nature of aggressive acts on Australian roads. *J. Australasian Coll. Road Safety* 30, 27–36.
- Stephens, A., Groeger, J.A., 2009. Situational specificity of trait influences on drivers' evaluations and driving behaviour. *Transport. Res. Part F: Traffic Psychol. Behav.* 12, 29–39.
- Stephens, A., Groeger, J.A., 2011. Anger-congruent behaviour transfers across driving situations. *Cognit. Emotion* 25, 1423–1438.
- Stephens, A., Groeger, J.A., 2014. Following slower drivers: Lead driver status moderates driver's anger and behavioural responses and exonerates culpability. *Transport. Res. Part F: Traffic Psychol. Behav.* 22, 140–149.
- Stephens, A., Hill, T., Sullman, M.J.M., 2016a. Driving anger in Ukraine: appraisals, not trait driving anger, predict anger intensity while driving. *Accid. Anal. Prevent.* 88, 20–28.
- Stephens, A., Koppel, S., Young, K., Chambers, R., Hassed, C., 2018. Associations between self-reported mindfulness, driving anger and aggressive driving. *Transport. Res. Part F: Traffic Psychol. Behav.* 56, 149–155.
- Stephens, A., Lennon, A., Bihler, C., Trawley, S., 2019. The measure for angry drivers (MAD). *Transport. Res. Part F: Traffic Psychol. Behav.* 64, 472–484.
- Stephens, A., Nieuwesteeg, M., Page-Smith, J., Fitzharris, M., 2017b. Self-reported speed compliance and attitudes towards speeding in a representative sample of drivers in Australia. *Accid. Anal. Prevent.* 103, 56–64.
- Stephens, A., Sullman, M.J.M., 2015. Trait predictors of aggression and crash-related behaviors across drivers from the United Kingdom and the Irish Republic. *Risk Anal.* 35, 1730–1745.
- Stephens, A.N., Trawley, S.L., Ohtsuka, K., 2016b. Venting anger in cyberspace: Self-entitlement versus self-preservation in# roadrage tweets. *Transport. Res. Part F: Traffic Psychol. Behav.* 42, 400–410.
- Stuster, J., 2004. Aggressive Driving Enforcement Evaluation of Two Programs. In , Administration, N.H.T.S., (ed.).
- Suhr, K.A., Nesbit, S.M., 2013. Dwelling on 'Road Rage': The effects of trait rumination on aggressive driving. *Transport. Res. Part F: Traffic Psychol. Behav.* 21, 207–218.
- Sullman, M.J.M., Hill, T., Stephens, A., 2018. Predicting intentions to text and call while driving using the theory of planned behaviour. *Transport. Res. Part F: Traffic Psychol. Behav.* 58, 405–413.
- Sullman, M.J.M., 2006. Anger amongst New Zealand drivers. *Transport. Res. Part F: Traffic Psychol. Behav.* 9, 173–184.
- Sullman, M.J.M., 2015. The expression of anger on the road. *Safety Sci.* 72, 153–159.
- Sullman, M.J.M., Gras, M.E., Cunill, M.O., Planes, M., Font-Mayolas, S., 2007. Driving anger in Spain. *Personal. Individual Diff.* 42, 701–713.
- Sullman, M.J.M., Meadows, M.L., Pajo, K.B., 2002. Aberrant driving behaviours amongst New Zealand truck drivers. *Transport. Res. Part F: Traffic Psychol. Behav.* 5, 217–232.
- Sullman, M.J.M., Stephens, A., Hill, T., 2017. Gender roles and the expression of driving anger among Ukrainian drivers. *Risk Anal.* 37, 52–64.
- Sullman, M.J.M., Taylor, J.E., Stephens, A.N., 2019. Multigroup invariance of the DAS across a random and an internet-sourced sample. *Accid. Anal. Prevent.* 131, 137–145.
- Sullman, M.J.M., Stephens, A., Yong, M., 2015. Anger, aggression and road rage behaviour in Malaysian drivers. *Transport. Res. Part F: Traffic Psychol. Behav.* 29, 70–82.
- Sullman, M.J.M., Stephens, A.N., Kuzu, D., 2013. The expression of anger amongst Turkish taxi drivers. *Accid. Anal. Prevent.* 56, 42–50.
- Sullman, M.J.M., Stephens, A.N., Yong, M., 2014. Driving anger in Malaysia. *Accid. Anal. Prevent.* 71, 1–9.
- Techer, F., Ojeda, L., Barat, D., Marteau, J.-Y., Rampillon, F., Feron, S., Dogan, E., 2019. Anger and highly automated driving in urban areas: The role of time pressure. *Transport. Res. Part F: Traffic Psychol. Behav.* 64, 353–360.
- Vanlaar, W., Simpson, H., Mayhew, D., Robertson, R., 2007. The Road Safety Monitor 2006: Aggressive Driving.
- Vanlaar, W., Simpson, H., Mayhew, D., Robertson, R., 2008. Aggressive driving: a survey of attitudes, opinions and behaviors. *J. Safety Res.* 39, 375–381.
- Villieux, A., Delhomme, P., 2007. Driving anger scale. French adaptation: Further evidence of reliability and validity. *Perceptual Motor Skills* 104, 947–957.
- Wickens, C.M., Mann, R.E., Ialomiteanu, A.R., Stoduto, G., 2016. Do driver anger and aggression contribute to the odds of a crash? A population-level analysis. *Transport. Res. Part F: Traffic Psychol. Behav.* 42, 389–399.
- Wickens, C.M., Mann, R.E., Stoduto, G., Ialomiteanu, A., Smart, R.G., 2011. Age group differences in self-reported aggressive driving perpetration and victimization. *Transport. Res. Part F: Traffic Psychol. Behav.* 14, 400–412.
- Zhang, T., Chan, A.H., 2016. The association between driving anger and driving outcomes: a meta-analysis of evidence from the past twenty years. *Accid. Anal. Prevent.* 90, 50–62.

Further Reading

- Banks, K.L., Nesbit, S.M., Murray, R.A., 2013. Driving anger and metacognition: the role of thought confidence on anger and aggressive driving intentions. *Aggressive Behav.* 39 (4), 323–334.
- Iancu, A.E., Hoge, A., Olteanu, A.F., 2016. The association between personality and aggressive driving: A meta-analysis. *Romanian J. Psychol.* 18, 24–32.
- Shinar, D., 1998. Aggressive driving: the contribution of the drivers and the situation. *Transport. Res. Part F: Traffic Psychol. Behav.* 1, 137–160.

Speeding: A “Tragedy of the Commons” Behavior

Bryan E. Porter^{*}, Thomas D. Berry[†], Kristie L. Johnson[‡], ^{*}Old Dominion University, Norfolk, VA, United States; [†]Christopher Newport University, Newport News, VA, United States; [‡]Alexandria, VA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	130
Review Framework	130
What We Know	131
Speeding Antecedents	131
Speeding Consequences	132
Speeding Countermeasures	133
Countermeasure Barriers	133
Traffic Safety Culture	133
Political Will	134
Next Steps for the Speeding Researcher	134
References	134
Suggested Reading	135

Introduction

Roads are a public good, most often paid for by public funds. Licensed drivers can drive anytime and go anywhere on them. We are free to do these things as individuals, taking advantage of the public space. Roads are a “commons” (Baden & Noonan, 1998). But commons’ spaces are often abused, for freedom to use them at will without perceived consequences for breaking rules lead to what Hardin (1968) termed the “tragedy of the commons” (p. 1244): the freedom of individuals abusing the commons, acting in their own self-interests without restraint, leads to ruin.

Porter and Berry (2004) argued abusing the roadway commons was a possible explanation for aggressive driving. Specifically, while we can use roads whenever and wherever we please, there are rules for that use. When those rules are broken in some cases, the abuse leads to risk for traffic crashes as well as for provoking aggression and anger in other road users. One rule-breaking behavior contributing to the commons’ tragedy is speeding. As Porter and Berry discussed, aggressive speeders use the road as their own property, ignoring the shared needs of others, in making weaving actions or placing others in harm’s way if they lose control of their vehicles. Such an individual’s behavior can benefit them at the expense of others, and when risk turns into crashes the commons has been ruined for all (particularly when lives are lost).

Our chapter continues this focus on speeding. We differentiate between “speeding” and “speed.” “Speeding” behavior involves driving at speeds that exceed a posted limit and are risky given conditions of the road, conditions of the vehicle, presence of other vehicles, pedestrians, or bicyclists, or other considerations. Speeding does not require an absolute numeric threshold, but rather a speed level that increases the likelihood of a crash. “Speed,” on the other hand, is just that: an absolute quantity that by itself does not imply or predict risk.

Speeding as a cause of fatal crashes is well-established. For example, in the United States (our home country), at least one driver was speeding in 26% of traffic fatalities (2018 data, latest available; National Center for Statistics and Analysis, 2020). In Virginia, our home state, speeding was involved in 42.2% of traffic fatalities (in 2019, latest data available; Virginia Highway Safety Office, 2020).

There are significant economic impacts, too. In the United States, the economic costs of speed-related crashes were \$52 billion in 2010, or an average of \$168 per person (Blincoe et al., 2015; also see CDC, 2014). Developing countermeasures to reduce speeding is worth the efforts required for public safety and economics.

Our chapter provides the latest, best practice information for understanding speeding’s prevalence as well as its countermeasures. Our goal is to produce concrete suggestions for how the behavior is shaped so that readers know how to build programs to address its reduction. However, for readers who wish to know more of the history and theoretical developments behind how the field developed these latest applications, we first provide a framework upon which the review is structured.

Review Framework

Speeding is a widely studied behavior from multiple disciplines. We simply cannot cover the full history in a brief chapter, nor a full coverage of all disciplines involved. We contain our review to published material mostly since 2011, roughly the past decade with few exceptions. Strategically, the current review adds to one we published previously, with readers encouraged to consider that earlier work for a longer-term understanding of speeding (Berry et al., 2011).

Table 1 Antecedents and consequences for speeding behavior

Category	Subcategory	Example	Reference
Antecedent	Structural—person	Age: sex/gender	Younger more likely than older Younger males and females in urban areas and younger males in rural areas more likely to speed. Men are 50% more likely than women to drive over speed limit Doroudgar et al. (2017), Richard et al. (2018), Richard et al. (2012)
	Structural—environment	Road design: roadway type	Roadway configuration and visual cues can alter speed (e.g., traffic calming) Travel speeds are lower in urban than suburban and rural areas on all road types Ewing (2017), Neuner et al. (2015), NHTSA (2018)
	Rules	Traffic norms perceptions/expectations	Speed limits impact speeding Safe speed influence by expectation of what would trigger a citation AutoInsurance Center (2020), Mannering (2009)
Consequences		Convictions	Citations reduce crashes 6% and violations 8%. Other research studies show no impact on receiving future citations Masten and Peck (2004)
		Crashes	At least one speeder in 26% of traffic fatalities in Virginia, speeding involved in 42.2% of traffic fatalities National Center for Statistics and Analysis (2020), Virginia Highway Safety Office (2019)
		Economic costs	Annual estimated cost of motor vehicle crashes in USA = 50 billion dollars Blincoe et al. (2015)

We organize material using the behavior-analytic model, a useful approach for understanding how behaviors impact public health injury risk (Sleet et al., 2006). Behavior analysis, or its large-scale partner, behavioral community psychology, focuses on antecedents leading to speeding and the consequences that follow which either increase or decrease the future likelihood of speeding. Further, our main interests are associated with observed speeding, not self-reported behaviors or reports of intention to speed.

With our goal to give the reader practical guidance, the review targets effective countermeasures that reduce speeding. We highlight countermeasures that also reduce crash and injury risk from speeding. To be even more consistent with modern trends in traffic psychology, we discuss the traffic culture context (Özkan and Lajunen, 2011). Scholars are acknowledging more commonly, and we believe finally so, that not all countermeasures are effective or generalize beyond their initial developmental setting and time. Culture impacts effectiveness.

Then, we conclude the chapter with our recommendations for next steps in speeding research. We lay a possible pathway for a future chapter to be written to update the one you are reading.

What We Know

The following are antecedents that predict and consequences that impact the likelihood of future speeding. These are the tools through which we can impact speeding. They are briefly summarized in text, but then presented in Table 1 for easy reference. In addition, they are organized by categories suggested by Mattaini (1996) to understand behavior within a cultural context.

Speeding Antecedents

Structural. Mattaini (1996) suggested structural antecedents could be categorized into those related to the *person* or *environment*. For the former, demographic variables are common antecedents. Consider roadways as congregate areas whereby people populate streets, roads, highways as drivers, passengers, and other road users with inherent demographic and motivational diversity. Over the years, age-related speed patterns have emerged in both attitudinal and behavioral measures. Young driver attitudes toward speeding tend to be more positive than older driver attitudes. This cognitive difference correlates consistently with age-related speeding behavior as observed in speed-related crashes.

Additionally, speed-related differences extend to gender differences among younger drivers. For instance, younger teenage men are more likely than teenage women to speed (38% vs. 25%, respectively; see Swedler et al., 2012). Younger men and women in urban areas are more likely to speed than those in rural areas (Doroudgar et al., 2017; Richard et al., 2018).

Gender, independent of age, also makes a difference. Men show consistently higher indices of risky behavior and the consequences associated with such risk taking. For each year from 1982 to 2018 male drivers were more likely than female drivers to be

involved in a speeding related fatality (FARS Data, posted 2019; [Insurance Institute for Highway Safety \[IIHS\], 2020](#)). Additionally, men as compared to women have further comorbidities including a greater likelihood to drive while intoxicated that involves the driver’s death. Also, men as compared to women report both using their cell phones and texting while driving in greater rates ([Forgays et al., 2014](#)). Because the activity of regulating speed is a primary task of a driver, factors playing a peripheral role but are none-the-less influential on driver attention, vigilance, decision making, and vehicular control should be considered in understanding the nature of a speeding driver.

For an environmental example, roadway type can impact speeding. Urban roadways have lower speeds on average than suburban and rural roadways (NHTSA, 2018).

Rules. Communities initiate traffic-safety laws to regulate the risk taking of drivers, and as such laws are a special type of countermeasure when setting speed limits and enforcement parameters. But laws as countermeasures derive their influence over behavior as socially derived and presented antecedents. Here, the antecedent law functions as an announcement to inform the public about a behavioral contingency. For instance, a speed limit sign posts a number indicating an upper end driver speed, but the sign is a symbol of a law indicating the potential consequence for violating the speed limit (a fine). When laws have been long established, they also can be a social norm (i.e., rule-governed behavior). When correlating different US state speeding laws with FARS’ speed related deaths, the type of law was seen to moderate a state’s population of driver speed ([AutoInsurance Center, 2020](#)).

Speeding Consequences

Crashes, injuries, fatalities, and economic decrements are important consequences of a speeding-related traffic crash. Such consequences were mentioned earlier in this chapter. Another type of consequence, traffic citations for speeding, is certainly one that can be manipulated as a countermeasure with effectiveness. We turn now to countermeasures—those actions communities can do to reduce speeding on their roadways. A summary of countermeasures is in [Table 2](#).

Table 2 Countermeasures to reduce speeding behavior

Category	Descriptor	Example	Reference
Education/ media	Media publicity Goal setting/social norming	Publicity campaign focused toward speed less effective than alcohol targeted Pilot study showed short-term improvement in compliance with speed limits	Phillips et al. (2011) , Newman et al. (2014)
Engineering	Road design Speed limiters Intelligent speed adaptation Autonomous vehicles	Roadway configuration and visual cues can alter speed Device restricts maximum speed. Study of speed limiters on 20 commercial vehicle fleets showed ~50% reduction in crash rate compared to truck without the limiters Vehicle adjusts to the speed limit or warns driver when over the speed limit. Small but lower proportion of drivers speed 8 or more mph over limit when using warning device. Speed alerting and speed controlling devices were effective at reducing speed of serious speeders in Netherlands, but one removed speeding returned Vehicle controls speed improving fuel economy and travel time, reducing crashes	Edquist et al. (2009) , Hanowski et al. (2012) , De Leonardis et al. (2014) , Van der Pas (2014) , Anderson et al. (2016)
Enforcement	Speed limits Variable speed limits reducing speed limits high-visibility Enforcement (HVE) automated enforcement	Higher speed limits associated with more fatalities Lower speed limits associated with fewer crashes, injuries, and fatalities. 3 mph reduction in average speed on a 30 mph road is expected to reduce injury crashes by 27% and fatal crashes by 49% Each 5 mph increase in state max speed limit, fatality rates increased 8% on interstates and freeways and 4% on other roads. Speed limit varies based on traffic and weather conditions. Preliminary study of Wyoming freeway VSL lead to reduction from 0.17 to 0.75 mph for every mph reduction in speed limit. Increasing number of localities are dropping speed limit citywide or on corridors to reduce crashes and severity of crashes. Reducing citywide speed to 20 mph in Bristol, UK, was associated with 63% reduction in fatal injuries. Enforcement and media used to highlight enforcement effort. Limited success as requires high levels of enforcement efforts Citation issued based on set number of mph over speed limit—consistent consequence. Reduction in the number of crashes. 20%–25% reduction in injury crashes. 8%–49% reduction for all crashes, 11%–4% for serious injury and fatal crashes. France installed 2756 cameras between 2003 and 2010 which is estimated to have prevented ~15,000 fatalities	AASHTO (2010) , Farmer et al. (2010) , Buddemeyer et al. (2010) , Bornioli et al. (2020) , Blomberg et al. (2012) , Decina et al. (2007) , Wilson et al. (2010) , WHO (2017)

Speeding Countermeasures

Education. Speeding education is used with all age groups. Predrivers receive the bulk of speeding related education via driver's education courses and driving tests. However, not every such education program has shown effectiveness in reducing speeding. For example, Phillips et al. (2011) found such programs were less effective for speeding than for alcohol and driving.

Another education-based countermeasure has used goal setting and social norming. Newman et al. (2014) evaluated this countermeasure's impact with a small group of industry drivers and demonstrated reductions in their speeding.

Engineering. Engineering countermeasures can be designed into the environment to reduce speeding. The most common methods include setting appropriate speeds, road design, and traffic controls. Setting appropriate speeds weighs safety and travel efficiency objectives (Forbes et al., 2012). Higher limits are associated with higher fatalities. In fact, a three-mph reduction in average speed on a 30 mph road can reduce injury and fatality crashes by 27% and 49%, respectively (ASHTO, 2010). Reducing speed limits has been linked to lower speeds on freeways (Buddemeyer et al., 2010) and within cities (Bornioli et al., 2020). And, when state maximum speed limits increase negative consequences follow. For each 5 mph increase, fatality rates increased 8% on interstates and freeways and 4% on other road types (Farmer et al., 2010). Reducing city limits to 20 mph in Bristol, UK produced a 63% reduction in fatal injuries (Bornioli et al., 2020).

Road design is considered with setting speed limits. Road design directly influences travel speeds with wider, straighter, more open roads spurring higher speeds than narrower, curved, and roadside-cluttered roads. Simple reconfigurations such as narrowing roadway markings, adding chicanes, or road diets reduce speed (Ewing, 2017; Neuner et al., 2015). Traffic controls can also aid in limiting speed. Road corridors or segments can be timed so that signal lights allow for maximum traffic flow when speed limits are obeyed (De Gier et al., 2011).

Other engineering countermeasures target the vehicle itself. Speed limiter devices can be installed to prevent going over a certain speed. These devices are typically found in school buses and commercial vehicles (Hanowski et al., 2012). Other devices can be installed that provide feedback when the driver goes over a certain speed. While such a device does not restrict speed, it provides a tactical, visual, or auditory alert (De Leonardis et al., 2014). Newer devices and apps allow parents to monitor their child's driving behavior, including speeding (Farmer et al., 2010). For lower volume conditions on road stretches without traffic controls, cruise control helps regulate the individual vehicle's speed.

More recently, technology is being developed and refined to implement intelligent transportation systems (ITS) where all vehicles and road users are linked together through infrastructure to help improve the safety and efficiency of roadways. Autonomous vehicles are the latest innovation being tested for such a system. While a fully automated vehicle fleet is decades in the future, automated vehicles could be programmed to follow speed limits and maintain following distances which could reduce speeding related crashes, fuel economy, and travel time all together (e.g., Anderson et al., 2016).

Enforcement. The most common countermeasure for speeding that applies consequences to the behavior in an attempt to prevent its future occurrence are the millions of citations issued by law enforcement officers each year resulting in fines and possibly court costs totaling hundreds of dollars (Mashhadi et al., 2017). Using citations attempts to deter speeding based on the fear of consequences. High visibility enforcement is an often-used technique combining enforcement and media outreach to promote the message that officers are out in force issuing citations. And there is evidence that these efforts can reduce speeding and its resulting crashes (Masten and Peck, 2004). But the success may require high levels of effort (Blomberg et al., 2012).

While law enforcement cannot issue citations to every motorist speeding past, speed cameras (automated enforcement) offer a consistent method to issue citations. Speed cameras allow a threshold to be set after which every motorist passing the location above the threshold is issued a citation. Public acceptance of speed cameras and traditional speed enforcement can be improved by ensuring that enforcement locations are data-driven, fair, and consistent with the goal of preventing crashes. And, in fact, these cameras are effective at reducing injury and fatal crashes (Decina et al., 2007; Høye, 2015; Wilson et al., 2010). In France between 2003 and 2010, speed cameras were linked to 15,000 prevented fatalities (WHO, 2017).

Countermeasure Barriers

The scholarly community has known the major antecedents and consequences for years. Countermeasures have been developed and tested as new technological innovations (e.g., photo enforcement) have been introduced, but the basic understanding has not changed. Rather, what is clear is that the largest issue with reducing speeding is to develop a countermeasure, and deploy it, in ways that breakthrough barriers created by other variables not regularly or fully incorporated in speeding studies: cultural differences and political will.

Traffic Safety Culture

Traffic safety culture, as a construct, is described by Özkan and Lajunen (2011) as bringing the person and environment interaction into traffic safety's sphere. Much like Lewin (1936) discussed behavior being a function of person in the environment, traffic safety culture considers how one's driving and walking behavior are dependent to some level on the environment in which those behaviors occur.

Özkan and Lajunen give many examples of how different levels of culture impact countermeasures' impacts on road safety (from the immediate environment of a person through to influences on citizens unique to their countries). Resources available to a state or

country influence road development, technological innovation, vehicle design, and economic support for countermeasures of varying costs. There are differences in governmental foci, educational opportunities, and urban versus rural spaces among different levels that also create cultural differences in public health. Such differences impact road safety outcomes.

While not a behavioral study, but rather a self-report one assessing drivers' violations and errors, Özkan et al. (2006) found that countries considered “safe” (Britain, The Netherlands, Finland) had drivers who reported more speeding on motorways than drivers from “dangerous” countries (Turkey, Greece, and Iran). Motorway speeding seemed to occur on better-constructed roads and was related to the amount of enforcement present. However, dangerous countries' drivers reported more aggressive driving acts and errors. The authors concluded less developed infrastructure, lower norms, and stress created the worse driving reports for the dangerous countries.

Acknowledging that cultural differences are important, countermeasures can be tailored to address influencers of a targeted area. Role models, particularly those who live in the same area of concern for the countermeasure, could demonstrate the behavior that is expected to increase safety. Such indigenous change agents (Roger, 1995) can be more effective motivating change than outside agencies or authority figures directing individuals about how to behave.

Political Will

In our experience over years of working with communities to change at-risk driver and pedestrian behavior, one barrier regularly presents itself. This barrier is particularly tall and thick when experts have suggested enforcement strategies for reducing risky driving. Enforcement, as reported here and in other settings, is among the most effective techniques for creating roadway behavior change. However, the will to pass legislation to allow enforcement, to consider automated enforcement in particular, and to listen to and absorb public scrutiny for restricting behaviors perceived by some as a freedom (e.g., freedom to drive faster than a limit on dry roads and no other driver nearby) is key to most successful countermeasures in roadway safety.

Policy makers who effectively garner public backing are successful in their ability to build community relationships and explain the need for legislative action. They too have a budget that can support countermeasures, such as automated enforcement (effective, but initially expensive to deploy). And these policy makers are successful communicating that the goal for their actions is to save lives, not fill the public coffers with citation payments. Professionals who wish to effectively build and launch a countermeasure to reduce speeding in a community, state, or country must have the support of these policy makers who have the political will to back and support the countermeasure. Those partners will also be critical in maintaining the gains of any countermeasure that is effective, for once professionals move on to other issues the policy makers remain to create legislation, allocate budgets, or both to ensure the countermeasure remains (e.g., automated enforcement) and gains sustained, or whether the risky behavior is likely to return.

Next Steps for the Speeding Researcher

The review provided here has two major conclusions. First, much of what we know about speeding and to reduce it to reduce injuries and fatalities has not changed from previous reviews. Appropriate road design, properly set speed limits, and strong enforcement programs that include automatic enforcement where feasible can reduce the harm caused by speeding behavior. Newer technologies, most notably autonomous vehicles, show promise for the future control of speeding. However, autonomous vehicles are not yet widespread to impact community-wide speeding outcomes.

Future researchers can make the biggest impact in the field by focusing on two topics. First, with the future technology in consideration, they can investigate what infrastructure needs will be required to properly use autonomous vehicles. While these vehicles will not solve all speeding concerns, nor do they remove fully the role of the human in driving risk, it seems clear the technology will only grow in importance. Researchers must not assume the speeding issue will disappear, and in fact while we have a mixed environment (human and machine drivers), the role of speeding may lead to worse outcomes if the speed variances on roadways increase.

Second, the importance of traffic culture and political will create tremendously important research opportunities. Countermeasures should be designed with both in mind; there may be regions where automated enforcement simply will not be accepted, leaving high-visibility enforcement the remaining best practice. There will be a greater need for researchers to test such adaptations and make no assumptions that everyone will be willing or able to use all countermeasures. What is not well known is how traffic culture moderates the effectiveness of countermeasures. Worse, there is even fewer research studies on the role of political will on countermeasure adoption. This one area alone would be well-worth a focus for a scholar's career.

In closing, we predict that the next review of speeding behavior and countermeasures will read much the same as this one and the earlier chapter we wrote (Berry et al., 2011) unless researchers take more time now on infrastructure, culture, and will. We look forward to the possibility that we will see such research take center stages in our journals soon.

References

- American Association of State Highway and Transportation Officials [AASHTO], 2010. Highway safety manual, first ed. Retrieved from: highwaysafetymanual.org
- Anderson, J.M., Kalra, N., Stanle, K.D., Sorenson, P., Samaras, C., Oluwatola, O.A., 2016. Autonomous vehicle technology: a guide for policymakers. RAND Corporation. Retrieved from: https://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR443-2/RAND_RR443-2.pdf
- AutoInsurance Center, 2020. The law vs. loss of life. Retrieved from: <https://www.autoinsurancecenter.com/the-law-vs-loss-of-life.htm>.

- Baden, J.A., Noonan, D.S., 1998. Preface: overcoming the tragedy.. In: Baden, J.A., Noonan, D.S. (Eds.), *Managing the commons*, second ed.. Indiana University Press, Bloomington and Indianapolis, Indiana .
- Berry, T.D., Johnson, K.L., Porter, B.E., 2011. Speed (ing). In: Porter, B.E. (Ed.), *Handbook of Traffic Psychology*. Elsevier Academic Press, San Diego, pp. 249–265.
- Blincoe, L., Miller, T.R., Zaloshnja, E., Lawrence, B.A., 2015. The economic and societal impact of motor vehicle crashes, 2010 (Revised) (No. DOT HS 812 013).
- Blomberg, R.D., Thomas III, F.D., Marziani, B.J., 2012. Demonstration and Evaluation of the Heed the Speed Pedestrian Safety Program (Report No. DOT HS 811 515). National Highway Traffic Safety Administration, Washington, DC. Retrieved from https://safety.fhwa.dot.gov/speedmgt/ref_mats/fhwsa1304/resources3/01%20-%20Demonstration%20and%20Evaluation%20of%20the%20Heed%20the%20Speed%20Pedestrian%20Safety%20Program.pdf.
- Bornioli, A., Bray, I., Pilkington, P., et al., 2020. Effects of city-wide 20 mph (30 km/hour) speed limits on road injuries in Bristol, UK. *Injury Prev.* 26, 85–88. Retrieved from: <https://injuryprevention.bmj.com/content/26/1/85>.
- Buddemeyer, J., Young, R.K., Dorsey-Spitz, B., 2010. Rural, variable speed limit system for southeast Wyoming.. *Transp. Res. Rec.* 9, 37–44.
- Center for Disease Control, 2014. Motor Vehicle Crash Deaths. Retrieved from: <https://www.cdc.gov/vitalsigns/motor-vehicle-safety/index.html> (accessed 01.02.2017).
- Decina, L.E., Thomas, L., Srinivasan, R., Staplin, L., 2007. Automated Enforcement: A Compendium of Worldwide Evaluations of Results (Report No. DOT HS 810 763). Retrieved from the NHTSA website: nhtsa.gov/DOT/NHTSA/Traffic%20Injury%20Control/Articles/Associated%20Files/HS810763.pdf.
- De Gier, J., Garoni, T.M., Rojas, O., 2011. Traffic flow on realistic road networks with adaptive traffic lights. *J. Stat. Mech. : Theory Exp.* 04, P04008.
- De Leonardis, D., Huey, R., Robinson, E., 2014. Investigation of the use and feasibility of speed warning systems (Report No. DOT HS 811 996). Retrieved from nhtsa.gov/staticfiles/nti/pdf/811996-InvestUseFeasSpeedWarnSys.pdf.
- Doroudgar, S., Chuang, H.M., Perry, P.J., Thomas, K., Bohnert, K., Canedo, J., 2017. Driving performance comparing older versus younger drivers.. *Traffic Inj. Prev.* 2, 41–46.
- Ewing, R., 2017. *US Traffic Calming Manual*. Routledge.
- Farmer, C.M., Kirley, B.B., McCart, A.T., 2010. Effects of in-vehicle monitoring on the driving behavior of teenagers. *J. Saf. Res.* 41 (1), 39–45.
- Forbes, G.J., Gardner, T., McGee, H., Srivivasan, R., 2012. Methods and practices for setting speed limits: an informational report (FHWA-SA-12-004). Federal Highway Administration. Retrieved from: https://safety.fhwa.dot.gov/speedmgt/ref_mats/fhwsa12004/fhwsa12004.pdf.
- Forgays, D.K., Hyman, I., Schreiber, J., 2014. Texting everywhere for everything: gender and age differences in cell phone etiquette and use. *Comput. Hum. Behav.* 31, 314–321.
- Hanowski, R.J., Bergoffen, G., Hickman, J.S., Guo, F., Murray, D., Bishop, R., Johnson, S., Camden, M., 2012. Research on the safety impacts of speed limiter device installations on commercial motor vehicles: Phase II (FMCSA-RRR-12-006). Retrieved from: <https://rosap.nhtl.bts.gov/view/dot/105>.
- Hardin, G., 1968. The tragedy of the commons. *Science* 162, 1243–1248.
- Høye, A., 2015. Safety effects of fixed speed cameras: an empirical Bayes evaluation. *Accid. Anal. Prev.* 82, 263–269.
- Insurance Institute for Highway Safety, 2020. Fatality facts 2018 – gender. Retrieved from: <https://www.iihs.org/topics/fatality-statistics/detail/gender>.
- Lewin, K., 1936. *Principles of Topological Psychology*. McGraw-Hill, New York.
- Masten, S.V., Peck, R.C., 2004. Problem driver remediation: A meta-analysis of the driver improvement literature.. *J. Safety Res.* 35, 403–425.
- National Center for Statistics and Analysis, 2020. Speeding: 2018 Data (Traffic Safety Facts. DOT HS 812 932). National Highway Traffic Safety Administration. Retrieved from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812932>.
- Newman, S., Lewis, I., Warmerdam, A., 2014. Modifying behaviour to reduce over-speeding in work-related drivers: an objective approach. *Accid. Anal. Prev.* 64, 23–29.
- Neuner, M., Chandler, B., Atkinson, J., Donoughe, K., Knapp, K., Welch, T., Crowe, R., 2015. Road Diet Case Studies (No. FHWA-SA-15-052). Federal Highway Administration, United States.
- NHTSA., 2018. National traffic speeds survey III: 2015 (Traffic tech: Technology transfer series. Report No. DOT HS 812 489). National Highway Traffic Safety Administration, Washington, DC.
- Ozkan, T., Lajunen, T., 2011. Person and environment: traffic culture. In: Porter, B.E. (Ed.), *Handbook of Traffic Psychology*. Elsevier Academic Press, San Diego, pp. 179–192.
- Özkan, T., Lajunen, T., Chliaoutakis, J.E., Parker, D., Summala, H., 2006. Cross-cultural differences in driving behaviours: a comparison of six countries. *Transport. Res. F: Traffic Psychol. Behav.* 9, 227–242.
- Phillips, R.O., Ulleberg, P., Vaa, T., 2011. Meta-analysis of the effect of road safety campaigns on accidents.. *Accid. Anal. Prev.* 43, 1204–1218.
- Porter, B.E., Berry, T.D., 2004. Abusing the roadway “commons”: understanding driving through an environmental preservation theory. In: Rothengatter, T., Huguenin, D. (Eds.), *Traffic and Transport Psychology: Theory and Practice*. Elsevier, Amsterdam, pp. 165–175.
- Mashhadi, M.M.R., Saha, P., Ksaibati, K., 2017. Impact of traffic enforcement on traffic safety. *Int. J. Police Sci. Manage.* 19 (4), 238–246.
- Mattaini, M.A., 1996. Public issues, human behavior, and cultural design. In: Mattaini, M.A., Thyer, B.A. (Eds.), *Finding Solutions to Social Problems: Behavioral Strategies for Change*. American Psychological Association, Washington, DC, pp. 13–40.
- Richard, C.M., Campbell, J.L., Lichty, M.G., Brown, J.L., Chrysler, S., Lee, J.D., Boyle, L., Reagle, G., 2012. Motivations for speeding, Volume I: Summary report. (Report No. DOT HS 811 658). National Highway Traffic Safety Administration, Washington, DC.
- Richard, C.M., Magee, K., Bacon-Abdelmoteleb, P., Brown, J.L., 2018. Countermeasures that Work: A Highway Safety Countermeasure Guide for State Highway Safety Offices, 9th ed. (Report No. DOT HS 812 478). National Highway Traffic Safety Administration, Washington, DC.
- Roger, E.M., 1995. *Diffusion of Innovations*, 4th ed. The Free Press, New York.
- Sleet, D.A., Trifiletti, L.B., Gielen, A.G., Simons-Morton, B., 2006. Individual-level behavior change models: Applications to injury problems. In: Gielen, A.C., Sleet, D.A., DiClemente, R.J. (Eds.), *Injury and Violence Prevention: Behavioral Science Theories, Methods, and Applications*. Jossey-Bass, San Francisco, CA, pp. 19–40.
- Swedler, D.I., Bowman, S.M., Baker, S.P., 2012. Gender and age differences among teen drivers in fatal crashes. *Ann. Adv. Autom. Med.* 56, 97–106.
- Virginia Highway Safety Office, 2020. 2019 Virginia Traffic Crash Facts. Richmond, Virginia.
- World Health Organization, 2017. Managing Speed. World Health Organization, Geneva. Retrieved from apps.who.int/iris/bitstream/10665/254760/1/WHO-NMH-NVI-17.7-eng.pdf.
- Wilson, C., Willis, C., Hendrikz, J.K., Le Brocq, R., Bellamy, N., 2010. Speed cameras for the prevention of road traffic injuries and deaths. *Cochrane Database System. Rev.* 11. Retrieved from speedcamerareport.co.uk/cochrane_sys_review_10.pdf.

Suggested Reading

- Porter, B.E. (Ed.), 2011. *Handbook of Traffic Psychology*. Academic Press, San Diego, CA.
- Barr Jr., G.C., Kane, K.E., Barraco, R.D., et al., 2015. Gender differences in perceptions and self-reported driving behaviors among teenagers. *J. Emerg. Med.* 48 (3), 366–370.
- Staton, C., Vissoci, J., Gong, E., Toomey, N., Wafula, R., Abdelgadir, J., Ratliff, C.D., 2016. Road traffic injury prevention initiatives: a systematic review and metasummary of effectiveness in low- and middle-income countries. *PLoS One* 11 (1), e0144971.

Cycling as a Mode Choice: Motivational Psychology

Sigal Kaplan, The Hebrew University of Jerusalem, Jerusalem, Israel

© 2021 Elsevier Ltd. All rights reserved.

Introduction	136
Behavioral Motivation	136
Expectations	137
The Theory of Planned Behavior	137
Need-Based Theories	139
Satisfaction	139
Extrinsic Satisfaction Sources	140
Intrinsic Satisfaction Sources	141
Further Research	141
References	142
Further Reading	143

Introduction

Cycling reduces mortality risk by 10%-28% and is associated with proven health benefits, including cardiovascular fitness, weight loss, improved cognitive ability, reduced stress, and longer life expectancy (Götschi et al., 2016). Even when considering accident risk and air pollution the health benefits of cycling considerably outweigh the negative externalities (de Hartog et al., 2010). Cycling also contributes to reducing air pollution and greenhouse emissions. Big data from Shanghai show that 1 million bike trips in the city center with an average length of 2.4 km result in CO₂ and NO_x emissions decreasing by 25,240 tons and 64 tons respectively (Zhang and Mi, 2018).

In light of the proven health and environmental benefits, as well as the ability to provide an affordable transport solution and reduce congestion, cities around the world are implementing bike-sharing and bicycle infrastructure solutions (Kaplan et al., 2018a, b). Currently, more than 700 cities operate bike-sharing programs, with the largest ranging between 20,000 and 90,000 bikes (Fishman et al., 2014). The public reaction to these systems is positive, with a 20% reduction in car use, up to three trips per bike daily, trips ranging between 15 and 20 min and up to 3-4 km traveled per trip (Fishman et al., 2014). While cycling in car-oriented countries is still evolving, evidence from Germany shows that cycling in a car-oriented country can rise to 10% nationwide and can exceed 30% in some cities, reaching the level of bicycle-friendly countries (Lanzendorf and Busch-Geertsema, 2014). Yet, policymakers face challenges associated with changes in behavioral norms and long-lasting car-oriented urban design. The common expression “if you build it, they will come” is used to justify investing in cycling infrastructure in places with low cycling rates and motivate the implementation of bike-sharing schemes, which is often coupled with limited cycling infrastructure (Félix et al., 2019). Policymakers facing such decisions need to promote changes in travel habits and norms as well as implement major changes in urban design and infrastructure. Hence, policymakers are concerned about the regional transferability of successfully promoting cycling habits. Many research and policy frameworks overlook two important aspects of cycling: motivations and trip satisfaction (Willis et al., 2013). The traditional perspective views regional differences in the success of cycling promotion as a result of spatial development patterns, transport and land-use policies, and cultural norms (Lanzendorf and Busch-Geertsema, 2014). However, this does not sufficiently support the transferability of cycling habits across regions and the ability to recreate the success of cycling-friendly countries. Three questions remain: if you provide cycling infrastructure, will they cycle? Will they also cycle under conditions of partial implementation? Can we effectively promote cycling in a gradual manner? The answers to these questions are embedded in motivational psychology. Unlike infrastructure design characteristics that depend heavily on local conditions, or a change in cultural norms that requires decades, motivational psychology can be generalized and is readily transferable across regions.

Behavioral Motivation

The ability to generalize behavioral motivation is fundamental to designing effective cycling policies and campaigns. While cycling can be discussed from a pure cost-benefit perspective (e.g., Wardman et al., 2007), it is better understood through discussing the broad sense of utility and travel satisfaction. Short-term trip-based mode choice is traditionally explained by functional monetary and time gains, which only partially explain cycling as a mode choice. Health benefits and hazards are also part of economic decision-making, although they are often difficult to calculate and perceive and are sensitive to temporal myopia bias. Rationalizing both excess car use and cycling has contributed to the development of a holistic and pluralistic utilitarian perspective that encapsulates general motivational theories and expands utility from “gain” and “loss” to “satisfaction” and “difficulty” (Gärling and Axhausen, 2003; De Vos et al., 2019).

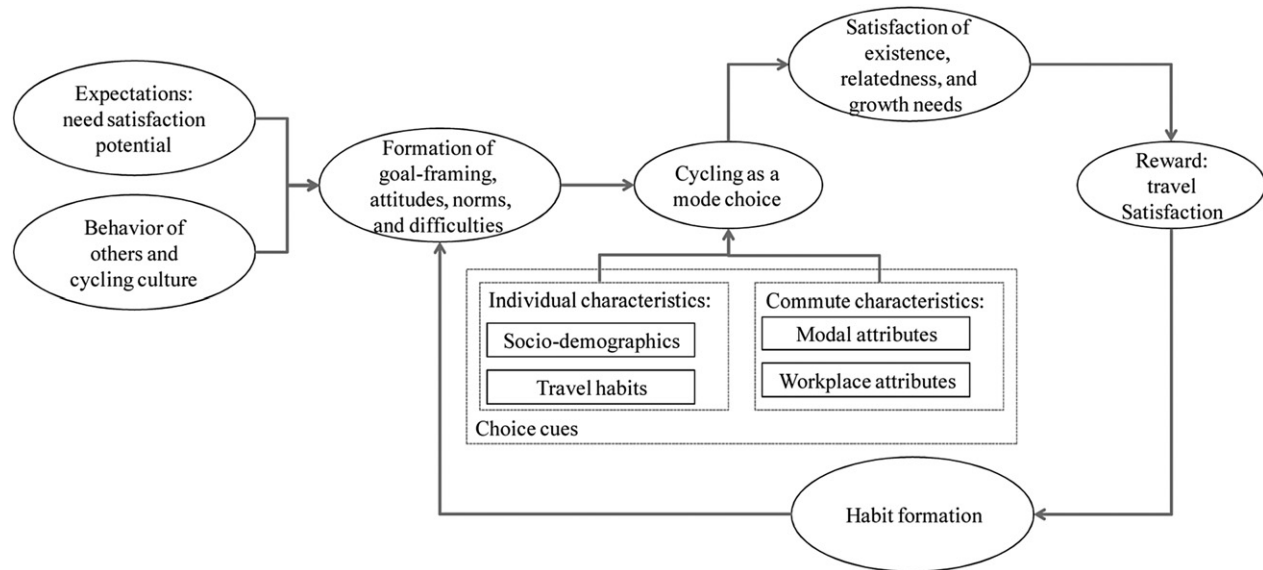


Figure 1 The habit formation loop.

The term “satisfaction” encompasses the opportunities in terms of time and monetary gains, health benefits, immediate rewards associated with hedonic enjoyment, and long-term rewards associated with self-concepts and general mood. Satisfaction can be viewed as the satisfaction of functional, social relatedness, and personal growth needs that form the basis of attitude formation (Alderfer, 1969) and the three-dimensional goal framing theory based on functional gain, normative, and hedonic values (Lindenberg and Steg, 2007). In contrast to the classical assumption that functional gains are the main motivators, higher-order needs and emotions can act in synergy or competition with functional values. Satisfaction also enables the broadening of the scope of utility to include the “new mobilities” paradigm (Sheller and Urry, 2006), which suggests that travel contributes to identity production by meeting higher-order needs of independence, self-actualization, self-esteem, and social relatedness. Satisfying social relatedness needs integrates conforming to social norms within the broader sense of satisfaction, thus adding the external motivation of subjective norms. The term “difficulties,” as the complementary pillar to “satisfaction,” refers to the perceived loss or an unsuccessful result, the perceived ability to perform an action, and the amount of volitional control (Ajzen, 1991). Cycling presents a broad range of potential difficulties, including functional constraints deriving from resource availability and accessibility, as well as physical difficulties and psychological barriers.

Cycling as a preferred choice and the transition to cycling as a habitual choice rely on positive perceptions of the travel experience to initiate the choice and reinforce repeated choice (Brette et al., 2014; Gärling and Axhausen, 2003). Similar to cost-benefit analysis, “satisfaction” and “difficulty,” which lead to the formation of goal framing, attitudes, and personal norms, can be assessed prior to the choice. Following the choice to cycle, derived satisfaction and difficulties are re-evaluated, leading to a habit formation loop upon a gratifying experience, as shown in Fig. 1. A possible form of the habit formation loop, where satisfaction leads to a proactive travel attitude, is documented by De Vos et al. (2019). The following sections describe at length each part of the loop.

Expectations

The Theory of Planned Behavior

The most prevalent guiding framework for internal cycling motivation is the theory of planned behavior (TPB; Ajzen, 1991), which has been applied to explain cycling intentions and behavior. The theory has been applied to examine the propensity of utilitarian and recreational cycling to be the main travel mode, and the acceptance and use of shared bikes and electric bikes. The TPB has been applied across population groups from adolescents to the elderly in both car-oriented and cycling-oriented countries around the world. It has been applied in a cross-sectional analysis to explain behavior at a specific point in time, and it has recently been applied to predict cycling behavior change (Bird et al., 2018; Nkurunziza et al., 2012). While the TPB has been found to be a rigorous approach, a unified scale or instrument has not been developed for its standardized application: application varies with respect to the choice of psychological constructs and measurement items. The TPB has been explored with a qualitative approach (Heinen and Handy, 2012; Janke and Handy, 2019) and validated using structural equation models (e.g., Cai et al., 2019; Fernández-Heredia et al., 2014; Kaplan et al., 2015) and inter-group comparisons between cyclists and non-cyclists (Félix et al., 2019). The TPB has been extended to include complementary dimensions and theories such as habit (de Bruijn et al., 2009), the socio-ecological model (Sigurdardottir et al., 2013), social identity theory (Lois et al., 2015), the technology acceptance model (Chen, 2016), the prototype willingness model (Frater et al., 2017), and the unified theory of acceptance and use of technology (Jahanshahi et al., 2020).

Cycling attitudes considered within the TPB are person-, behavior-, and product-centered attitudes. Human-centered attitudes derive from personal values. Although they are manifested through actions and choice, they are not directly associated with behavior. Examples are environmental attitudes and personal norms. Behavior-centered values focus on the attractiveness of cycling, encompassing perceived expectations regarding the value of cycling in terms of observed advantages and feelings. Behavior-based attitudes are a direct translation of Ajzen's (1991) original model conceptualization and are the most prevalent in cycling studies. Behavior-centered attitudes include normative, gain, and hedonic goal framing and can be cycling specific or comparative measures of cycling versus other modes. Normative goal framing focuses on moral obligation and conforming with personal norms. Gain goal framing focuses on the perception of results, and hedonic goal framing focuses on feelings associated with the behavior. Examples of a normative perspective are "I would pollute the environment less" (Lois et al., 2015) and "I feel morally obliged to travel by bike instead of using other transportation" (Han et al., 2017). Examples of a gain perspective are "cycling helps me save money compared to other transport modes," and "cycling helps me get exercise" (Lois et al., 2015; Sigurdardottir et al., 2013). Examples of a hedonic perspective are "I enjoy cycling" (Lois et al., 2015) and "cycling is interesting/pleasant/stimulating" (Frater et al., 2017). Product-centered attitudes are associated with cycling infrastructure, and with bike-sharing systems' quality, features, and usefulness. Product-based attitudes largely depend on the perceived product image, imagined prototype evaluation, and similarity to existing alternatives (Frater et al., 2017). In the case of cycling and the adoption of shared-bicycle systems, product-based evaluation depends on the subjective image or vision of the system, rather than its objective traits. Hence, product-based attitudes also provide normative, gain, and hedonic framing based on the envisioned system attributes and associated value. Examples of infrastructure-based attitudes are "availability of bikeways along the way" (Gao et al., 2018) and "the route has beautiful scenery" (Winters et al., 2011). Product-based normative goal framing includes ecological and altruistic values. An example is "the environmental performance of the bike-sharing company corresponds to my expectations" (Chen, 2016). Product-based hedonic perceptions correspond with innovative features or a positive experience using the system. Examples are "using the bike-sharing system is relaxing" (Chen, 2016), "I am interested in riding a high-tech electric bicycle," and "electric bicycles are fast and easy" (Kaplan et al., 2015).

Cycling subjective norms refer to the social circle of influence, including significant others (e.g., family, friends, and colleagues), and community social norms based on location, cycling culture, media activity, population strata, and government policies. Social norms include both visibility and word-of-mouth. Visibility refers to behavioral trends and the behavior of significant others, while word-of-mouth consists of conversations about cycling, expressed opinions and thoughts, and appreciation by others. Norms represent the social pressure to conform to expectations, social support and encouragement to do "the right thing," and the hope of becoming an opinion leader in one's social circle. Examples of word-of-mouth in the social circle representing social pressure and encouragement, respectively, are "people who influence my behavior think that I should use bike sharing" and "people who are important to me encourage me to use bike sharing" (Chen, 2016). An example of a community visibility norm is "people usually cycle in the city," while examples of community word-of-mouth are "people think that cycling for recreation is cool" and "my friends are concerned about environmental responsibility" (Kaplan et al., 2015). Examples of community policy norms are "my school encourages environmentally responsible behavior" and "driving in city centers will be banned" (Sigurdardottir et al., 2013).

Cycling difficulties include person-oriented difficulties, situational difficulties and constraints, and product or system-oriented ease-of-use (Fernández-Heredia et al., 2014; Kaplan et al., 2015). Person-based difficulties comprise bike availability constraints, concerns regarding the perceived ability to cycle, lifestyle constraints, and fear of cycling. Availability or accessibility constraints include lacking a bicycle, gear, facilities, and storage space. The perceived ability to cycle entails concerns regarding the required physical effort, physical fitness, and cycling skills or experience. Lifestyle constraints include dress code, trip chaining, traveling with children, the need to carry baggage, and the car as essential commuting or work tool (Félix et al., 2019). Situational difficulties include location-based concerns, unfavorable weather conditions, and parental consent for children and adolescents. Location-based concerns include hilly topography, long-distance, cycling or traffic congestion, quality and availability of facilities, and accident and crime rates (e.g., bike theft, vandalism, and harassment). A product-oriented difficulty is bicycle maintenance. System-oriented difficulties refer to infrastructure and operating the bike-sharing system features, such as using apps, finding the bikes, paying online, and returning the bike.

Perceived risk and fear of cycling in mixed traffic are among the most important barriers to cycling in terms of their prevalence and strength. Fear of cycling is particularly relevant in car-oriented or emerging cycling-oriented countries where cyclists share the road with mixed traffic. While fear of cycling has been incorporated into the TPB as a type of barrier, there is ample research on cycling fear as a standalone concept, its sources, and associated coping mechanisms. Fear of cycling can be measured by the perception of danger associated with cycling in mixed traffic. Recent studies show that fear is a multifaceted construct driven by explicit cyclist-motorist interactions and implicit cyclist perceptions of infrastructure layout and traffic (Chataway et al., 2014). Research on the topic includes cyclist-motorist road interactions and resulting emotions, and cyclists' coping strategies while cycling in mixed traffic. Among the first to identify cyclist-motorist tension were O'Connor and Brown (2010), who identified negative verbal and physical harassment of cyclists by drivers in Australia. These findings were corroborated by Kaplan and Prato (2016), whose results from Israel revealed several sources of driver aggression and three types of cyclist coping mechanisms, with the most prevalent being cycling avoidance. They showed that conflicts between cyclists and motorists involve a high perception of threat that leads to intense feelings of anxiety, fear, and stress, impeding cycling. Later, Möller and Haustein (2017) presented evidence that anger expression among drivers and cyclists also exists in cycling-friendly countries such as Denmark. Following Prati et al. (2017) findings that cyclists are viewed as a minority group, Kaplan et al. (2019b, 2019c) explored the ability of the social interaction theory to explain cycling fear and the willingness of drivers and cyclists to share the road in the Czech Republic and Israel. They found that the

willingness to share the road is associated with cycling cultural norms, the perception of drivers as distracted or aggressive, the perception of cyclists as vulnerable or assertive, and the perceived anger and anxiety resulting from cyclist-motorist road interactions (Kaplan et al., 2019c). The findings show that road infrastructure is needed to establish a safe cycling environment and promote cyclists' "right" to the road. Moreover, cyclist-motorist interactions and perceptions need to change through promoting mutual empathy in order to eradicate fear as a barrier to cycling.

Need-Based Theories

Ho et al. (2015) found that core values associated with cycling among recreational cyclists reflect the ability of cycling to create a sense of happiness and to fulfill basic and higher-order needs including belonging, self-esteem, and self-actualization. They postulated that cycling motivation conforms to the theory of human needs. Alderfer's (1969) ERG theory of needs (Kaplan et al., 2018a, 2018b), social identity (Lois et al., 2015), and Ryan and Deci. (2000) self-determination theory (SDT; Chi et al., 2020; Grayson et al., 2019) were proposed to study cycling motivation. As in the case of the TPB, a unified scale for cycling needs has not been developed and each study provides a different empirical interpretation. The ERG theory of needs postulates that people strive to satisfy existence, relatedness, and personal growth needs. In the context of cycling, existence needs are functional needs related to the ability to carry out daily activity patterns, save resources, gain independence, and maintain a healthy lifestyle. Relatedness needs in the context of cycling refer to belonging to a group or community, socializing, and gaining appreciation from significant others. Personal growth needs in the context of cycling refer to acting in accordance with personal values and promoting personal growth in terms of self-efficacy, competence, and autonomy. The theory of needs allows a non-hierarchical order of needs importance variation across individuals (Alderfer, 1969), meaning that relatedness and growth needs can take priority over functional needs. This is of the utmost importance for cycling as a mode choice, since it explains and rationalizes choices that deviate from the functional superiority embedded in the economic utilitarian approach.

Need-based theories differ from the TPB theory in three respects: (1) they measure attitudes toward the behavior as the ability to satisfy human needs; (2) they focus on social relatedness rather than conforming to social norms; and (3) they consider behavioral constraints that derive from self-efficacy, competence, and autonomy. While the TPB provides non-mediated behavioral motivation that accommodates internal motivation, external pressure, and perceived ability, need-based theories drive the goal framing underlying the TPB attitudinal constructs: the need to function and the need for social relatedness and personal growth. Much like a utility, needs can become behavioral motivators on the basis of expectations regarding the ability of a behavior to satisfy needs. Kaplan et al. (2018a) developed a need-based cycling motivational scale validated in Denmark as a cycling-oriented country and Poland as a car-oriented country (Kaplan et al., 2018b).

Existence needs motivate gain goal framing and thus trigger attitudes based on the ability to satisfy functional gains. Cycling related existence needs include saving time, money, parking search time and costs, getting exercise and fresh air, increasing the number of activities and the locus of opportunities, and relieving stress (Kaplan et al., 2018a). Existence needs relate to functional gains and can make a significant difference in the case of weakened populations and individuals with limited financial resources. They are also highly relevant in cycling-oriented cities, such as Copenhagen, where cycling superhighways, exclusive bicycle pathways and bridges, and bicycle preemption at traffic lights promote bicycle efficiency.

Relatedness needs include the need to belong and to create meaningful social relationships in order to gain acceptance, confirmation, support, understanding, and influence (Alderfer, 1969). Relatedness needs to motivate interacting socially, creating a social identity, and conforming to social norms in order to belong and gain appreciation. Relatedness needs can be satisfied by significant others, expanding social networks, and the community at large, with the strength of relatedness depending on the strength of the social ties and influence. Cycling relatedness needs motivate conforming to cultural cycling norms, taking part in cycling as a group activity, creating a cyclist social identity, and using cycling as a platform for social interaction (Chi et al., 2020; Grayson et al., 2019; Kaplan et al., 2018a, 2018b; Lois et al., 2015).

Growth needs entail a person utilizing their capacity to have an effect on themselves or their environment, enhancing capacity and autonomy, developing new capabilities, and finding the opportunities to realize their full potential (Alderfer, 1969). By definition, the ERG embodies the SDT constructs, which focus on three types of needs: competence, autonomy, and relatedness (Ryan and Deci, 2000). Growth needs trigger normative goal framing because they encompass the development of self-efficacy, autonomy, and self-identity by acting in agreement with normative values, increasing performance, and learning new skills. In interpreting growth needs, Lois et al. (2015), Grayson et al. (2019), and Chi et al. (2020) focused on growth as the effect on the self. Kaplan et al. (2018a) elicited a person's effect on themselves and their environment, self-identity formation, and acting in accordance with environmental values.

Satisfaction

Utilitarian and recreational cycling satisfaction has been explored for children (Westman et al., 2017), adults, and the elderly (Zander et al., 2013) in various countries. Cycling satisfaction is a desirable outcome contributing to feelings of enjoyment, long-term general well-being, and an increase in productivity (Luth et al., 2017; Zander et al., 2013). It also serves as a motivational factor through triggering the formation of cycling attitudes and nurturing the habit formation loop (De Vos et al., 2019; Ingvardson et al., 2020; Rowe et al., 2016). The importance of considering the dual role of satisfaction as both a motivator and a reward is demonstrated in de Kruijf et al. (2019) study, which examined expected versus actual e-bike travel satisfaction. Although they

found a statistically significant difference between expectations and actual satisfaction, the results show that on average, people can reasonably assess their long-term cycling satisfaction. Another important reason to consider cycling satisfaction is the interweave between leisure and work satisfaction (Luth et al., 2017). Both work and cycling satisfaction derive from balanced activity engagement and having a promotion focus; hence, encouraging cycling as a leisure activity can evoke satisfaction in other life domains.

Cycling is a common social practice during childhood encouraged in both bicycle-friendly and car-oriented countries. Yet, most research focuses on the travel satisfaction of adults, with little attention to children's cycling happiness, activation, and positive mood. Westman et al. (2017) reduced the knowledge gap by developing the Satisfaction with Travel for Children scale and applying it to assess the travel quality of children aged 10-15. Children who traveled by bicycle and school bus reported higher travel satisfaction than those who traveled by car. Children who cycled to school also had better mood, with higher levels of valence and activation. Overall, the study found that cycling can improve everyday travel for children.

Mood improvements as a result of cycling are expected due to improved blood circulation and the release of neurotransmitters that trigger the brain's reward system. It is interesting to investigate whether cycling has an effect on the mood of those with clinical mood conditions. While evidence of this is scarce, cycling has recently been identified as a means to improve the well-being and life satisfaction of the elderly, and to alleviate depression symptoms in clinical patients. Leyland et al. (2019) explored the influence of regular bike and e-bike riding on cognitive functioning and well-being. Well-being was measured using three questionnaires: the Psychological Well-Being scale, the Satisfaction in Life scale, and the Positive and Negative Affect scale. During the eight-week trial period, the participants cycled for an average of two hours a week. Regular bike users and e-bike cyclists showed a significant improvement in terms of life satisfaction, mood, and cognitive functioning. Zhao et al. (2017) longitudinal clinical trial results show that long-term bicycle riding helps to relieve depression symptoms among hemodialysis patients. While satisfaction was not measured directly, quality of life and clinical depressive symptoms improved.

While studies have found that cyclists are more satisfied with their commute than other mode users, few have explored the underlying reasons for cycling satisfaction (Willis et al., 2013). Satisfaction is pyramidal in structure, with various layers of psychological depth. Satisfaction in its most immediate form refers to mood change that can derive from both physiological reactions and psychological factors. Satisfaction can derive from external factors such as a pleasant environment and trip characteristics (Willis et al., 2013) and can be a result of the ability of cycling to relieve stress (Avila-Palencia et al., 2017), which is an underlying determinant of satisfaction (Elangovan, 2001). However, satisfaction has much deeper roots in the psychological structure. When looking at motivational goal framing, satisfaction can be described from the perspective of normative, gain, and hedonic goal attainment framing. Delving deeper into the psychological structure, satisfaction derives from meeting existence, relatedness, and growth needs, with the latter satisfied through noticeable improvements in self-concepts. The layers of satisfaction have been revealed in qualitative studies and validated using large samples and rigorous statistical models. Analyzing cycling through the lens of the theory of social practice, Spotswood et al. (2015) postulated that material, meaning, and competences motivate cycling behavior. Their interpretation of material as a means to reproduce daily life, meaning as normative beliefs, and competencies as skills acquirement coincides with the gain and normative goals of the ERG theory of human needs. Thus, the motivational role of satisfaction is also evident in the theory of social practice.

Extrinsic Satisfaction Sources

While there is much knowledge regarding the influence of the built environment on cycling frequency and prevalence, the relationship between the built environment and cycling satisfaction is less explored (Willis et al., 2013). Moreover, evidence from different locations yields some contradictory results, meaning that the results are not readily transferrable across geographical regions.

Infrastructure design and urban design are sources of cycling satisfaction in both cycling-friendly and car-oriented countries. Ho et al. (2015) qualitative results showed that cycling in a pleasant urban environment contributes to enjoyment. Their findings include statements such as "I enjoyed the scenic beauty along the cycling route," which show a direct linkage between the cycling environment and cyclists' sense of joy. Snizek et al. (2013) mapped cyclists' perceived positive and negative experiences along cycling routes in Copenhagen. Regression analysis results showed that cycling satisfaction depends on separated cycling paths, road hierarchy, commercial land use, and the share of green edges along the cycling path. de Kruijf et al. (2019) found that e-bike travel satisfaction in the Netherlands is positively related to green edges, while other characteristics of the built environment, such as aesthetics and street atmosphere, were insignificant. Sharma et al. (2019) performed a regression analysis to find the factors affecting satisfaction with the bicycle roads in Seoul, with a focus on engineering design. Their findings show the effect of engineering design and inducing cycling comfort on the quality of cyclists' experience. They found that satisfaction is related to good condition asphalt, wide cycle lanes, moderate slope, sufficient signage, separation from road traffic, and bicycle parking availability. Through analyzing a cycling infrastructure improvement case study, Skov-Petersen et al. (2017) showed that cyclists are sensitive to engineering improvements in terms of both satisfaction levels and behavioral change. User satisfaction with bike-sharing was also investigated, with the overall level of satisfaction found to be related to bicycle comfort, availability, customer response efficiency, and smartphone compatibility (Maioli et al., 2019).

Willis et al. (2013) explored the relationship between the built environment and the cycling satisfaction of several cyclist types: cycling enthusiasts, exercise-convenience motivated cyclists, active environmentalists, and year-round cyclists. While they did not find a significant difference between "satisfied" and "non-satisfied" cyclists, possibly due to the low variation in the urban

environment, they found that cycling distance is associated with a higher level of satisfaction among cycling enthusiasts. The authors suggested that intrinsic motivators related to economic saving, pleasure, and identity motivate satisfaction. This opened the door to understanding satisfaction through the lens of the normative, gain, and hedonic goal framing theory and through the lens of the theory of human needs.

Intrinsic Satisfaction Sources

Luth et al. (2017) explained cycling satisfaction as the result of being passionate about cycling and having a high promotion focus directed at cycling. Passion arises from the need to satisfy emotional needs, which are manifested through life goals. Looking through the lens of goal attainment (Lindenberg and Steg, 2007), studies show that satisfaction derived from goal attainment varies between recreational and utilitarian cyclists, and between car-oriented and cycling-oriented environments. Rowe et al. (2016) showed that for women recreational cyclists, enjoyment is derived from normative, gain, and hedonic goal attainment. Normative goal attainment seeks to pursue environmental, health, and fitness goals, along with travel independence, control, and autonomy, which give rise to feelings of empowerment. Gain goal framing arises from time and monetary savings, convenience versus other transport modes, increased activity participation, and increased spatial locus of opportunities. Notably, normative and gain goals are interlinked because gain goal attainment can help to attain normative goals. Hedonic feelings arise with sensation seeking, the stimulation of the visual and somatic sensory systems as a result of outdoor cycling in a pleasant environment, and the triggering of fond childhood memories. Rowe et al. (2016) showed that cyclists can articulate the goals attained, attain multiple goals simultaneously, and link goal attainment to enjoyment. Spotswood et al. (2015) suggested differences between the hedonic goal framing of recreational and utility cyclists, while normative and gain goal framing for the two groups are aligned. Spotswood et al. (2015) found that in a car-oriented environment, utilitarian cyclists tend to report less enjoyment and focus either on gain goals (i.e., monetary and time savings) or on normative goal framing (i.e., fitness improvement, sporty and valiant image) due to the stressful car-oriented cycling environment. They also found that while recreational cycling is associated with a sense of competence, utilitarian cycling is associated with bravery and commitment.

If one delves deeper into the psychological structure, cycling satisfaction is rooted in the fulfillment of existence, relatedness, and growth needs. Ingvardson et al. (2020) investigated the role of meeting human needs and its relationship to the travel satisfaction of car, public transport, and bicycle commuters in Denmark. The level of satisfaction of car users and bicycle riders was very similar, with 40% of the respondents indicating the highest level of satisfaction for both modes. The study shows that in a bicycle-oriented country, the bicycle is a competitive alternative to car driving. The findings reveal that while car driving satisfies functional needs and generates positive self-concepts, bicycle riding meets self-efficacy and togetherness, generates positive self-concepts, and triggers travel satisfaction. The analysis verified a statistically significant link between the fulfillment of needs and travel satisfaction. Kaplan et al. (2019a) explored the relationship between cycling self-concepts and mood improvements among cyclists in Australia. They measured happiness, enthusiasm, relaxation, and composure (calmness and focus) using a scale adapted from the profile mood states scale. Taking a domain-specific multidimensional approach, Kaplan et al. (2019a) developed a scale for measuring cycling self-concepts, encompassing four abilities: physical, social, emotional, and contextual. The cycling self-concept scale assessed physical, competence, social, and psychological self-concepts. Physical self-concepts include items that measure appearance and functioning, since the function is a more powerful and enduring self-concept than physical appearance. The physical self-concept scale refers to staying in shape, reaching fitness goals, feeling younger, developing strength, mobility and energy, and improving physical abilities. The competence scale addresses coping with challenges by overcoming topography, weather, and functional difficulties. The scale concerns self-actualization, self-esteem, optimism, and making the best of every situation. The social scale focuses on the perceived improvement in the ability to socialize by expanding the social network, having meaningful conversations, and bonding through a shared activity. The findings show both a direct relationship between cycling and mood improvement and an indirect relationship, with self-concepts as mediators between cycling and mood improvement.

Further Research

Cycling motivation research keeps expanding in new directions and there are important areas to be further explored. Three main research needs are presented.

Cycling socialization. Current studies focus mainly on adult behavioral change. However, current generations are responsible for the travel behavior of future generations. Baslington (2008) showed how travel socialization processes deliver travel behavior from parents to children. Sigurdardottir et al. (2013) showed that children take their parents as role models for cycling choice, influencing their view of themselves as future cyclists or car users. Yet, in-depth analysis of the socialization process has not been conducted. A possible research direction is using the symbolic interaction theory to explain cycling socialization processes. This theory postulates that people validate their self-concepts by regarding themselves from the perspective of others, as reflected through their own perceptions. Kaplan et al. (2019c) validated the theory for adult cyclists, but it has not been applied to explore cycling habit formation among children. Family climate is a powerful force shaping children's behavior through role models, feedback, communication, monitoring, commitment, and limits (Taubman-Ben-Ari and Katz-Ben-Ami, 2013). Currently, there is no research addressing the link between family climate and cycling among children.

Switching to cycling. Most studies focus on cycling as a standalone decision rather than a mode switching decision. The main competitor of cycling is the car, in both car-oriented and bicycle-friendly countries (Ingvardson et al., 2020). The switch from other transport modes to cycling is non-trivial due to difficulties related to individual characteristics, trip attributes, and the urban form. Currently, psychological factors are outside the choice context. A few studies document change or switching behavior (Bird et al., 2018; Kaplan et al., 2018a; Thigpen et al., 2019). Studies internalizing psychological motivation into modal choice using hybrid discrete choice models need to be elaborated so that choice psychology can also be used as an integral part of agent-based transport models. Moreover, according to the trans-theoretical model (Prochaska and DiClemente, 1983), behavioral change requires several stages: pre-contemplation, contemplation, preparation, action, maintenance, and relapse. In the context of cycling choice, these stages translate into awareness, cognitive assessment, motivation formation, preparation, action, maintenance, and switching back to other modes. Currently, only the cognitive assessment and motivation stages are addressed. While loyalty to bike-sharing systems has been investigated (Chen, 2016), the overall change process over time has not been documented. More information is needed regarding the overall change process, its time span, and the most important success factors. Moreover, the maintenance stage and possible relapse to other modes should be explored to expand cycling from a niche market to a common modal choice. Longitudinal studies tracking cycling behavior over time are needed.

Measuring satisfaction. Evaluating satisfaction from self-reports using satisfaction scales is a common practice. A few studies used self-reporting to track cycling satisfaction along cycling routes (Snizek et al., 2013). However, such studies are limited in terms of satisfaction point identification, self-report biases, and continuous satisfaction measurements. Shoval et al. (2018) showed how the emotional dimension of the city experience can be mapped using real-time tracking of emotional arousal (positive or negative) through physiological monitoring. Such monitoring could help in understanding the link between urban form and hedonic reactions to cycling, and how the intensity and length of the hedonic gratification influence habit formation. Satisfaction monitoring using physiological measures could also link cycling well-being with objective measurements of cycling health improvements. This research direction combines data science and psychology and pertains to the next generation of citizen-based science and data visualization.

References

- Ajzen, I., 1991. Theory of planned behaviour. *Organ. Behav. Hum. Decis. Processes* 50, 179–211.
- Alderfer, C.P., 1969. An empirical test of a new theory of human needs. *Organ. Behav. Hum. Perform.* 4, 142–175.
- Avila-Palencia, I., de Nazelle, A., Cole-Hunter, T., et al., 2017. The relationship between bicycle commuting and perceived stress: a cross-sectional study. *BMJ Open* 7, e013542.
- Baslington, H., 2008. Travel socialization: a social theory of travel. *Mode behavior. Int. J. Sustain. Transport* 2 (2), 91–114.
- Bird, E.L., Panter, J., Baker, G., Jones, T., Ogilvie, D., 2018. Predicting walking and cycling behaviour change using an extended theory of planned behaviour. *J. Transport Health* 10, 11–27.
- Brette, O., Buhler, T., Lazaric, N., Marechal, K., 2014. Reconsidering the nature and effects of habits in urban transportation behavior. *J. Inst. Econ.* 10, 399–426.
- de Bruijn, G.J., Kremers, S.P.J., Singh, A., van den Putte, B., van Mechelen, W., 2009. Adult active transportation adding habit strength to the theory of planned behavior. *Am. J. Prev. Med.* 36, 189–194.
- Cai, S., Long, X., Li, L., Liang, H., Wang, Q., Ding, X., 2019. Determinants of intention and behavior of low carbon commuting through bicycle-sharing in China. *J. Cleaner Prod.* 212, 602–e609.
- Chataway, E.S., Kaplan, S., Nielsen, T.A.S., Prato, C.G., 2014. Safety perceptions and reported behavior related to cycling in mixed traffic: a comparison between Brisbane and Copenhagen. *Transport. Res. Part F* 23, 32–43.
- Chen, S.Y., 2016. Using the sustainable modified TAM and TPB to analyze the effects of perceived green value on loyalty to a public bike system. *Transport. Res. Part A* 88, 58–72.
- Chi, M., George, J.F., Huang, R., Wang, P., 2020. Unraveling sustainable behaviors in the sharing economy: an empirical study of bicycle-sharing in China. *J. Cleaner Prod.* 260, 120962.
- Elangovan, A.R., 2001. Causal ordering of stress, satisfaction and commitment, and intention to quit: a structural equations analysis. *Leadership Organ. Dev. J.* 22, 159–165.
- Félix, R., Moura, F., Clifton, K.J., 2019. Maturing urban cycling: comparing barriers and motivators to bicycle of cyclists and non-cyclists in Lisbon, Portugal. *J. Transport Health* 15, 100628.
- Fernández-Heredia, A., Monzón, A., Jara-Díaz, S., 2014. Understanding cyclists' perceptions, keys for a successful bicycle promotion. *Transport. Res. Part A* 63, 1–11.
- Frater, J., Kuijer, R., Kingham, S., 2017. Why adolescents don't bicycle to school: does the prototype/willingness model augment the theory of planned behaviour to explain intentions? *Transport. Res. Part F* 46, 250–259.
- Fishman, E., Washington, S., Haworth, N., 2014. Bike share's impact on car use: evidence from the United States, Great Britain, and Australia. *Transport. Res. Part D* 31, 13–20.
- Gao, Y., Chen, X., Shan, X., Fu, Z., 2018. Active commuting among junior high school students in a Chinese medium-sized city: application of the theory of planned behavior. *Transport. Res. Part F* 56, 46–53.
- Gärling, T., Axhausen, K.W., 2003. Introduction: habitual travel choice. *Transportation* 30, 1–11.
- Götschi, T., Garrard, J., Giles-Corti, B., 2016. Cycling as a part of daily life: a review of health perspectives. *Transport Rev.* 36, 45–71.
- Grayson, A., Totzkay, D.S., Walling, B.M., Ingalls, J., Viken, G., Smith, S.W., Silk, K.J., 2019. Formative research identifying message strategies for a campus bicycle safety campaign using self-determination theory and the social norms approach. *Accid. Anal. Prev.* 133, 105295.
- Han, H., Meng, B., Kim, W., 2017. Emerging bicycle tourism and the theory of planned behavior. *J. Sustain. Tourism* 25 (2), 292–309.
- de Hartog, J.J., Boogaard, H., Nijland, H., Hoek, G., 2010. Do the health benefits of cycling outweigh the risks? *Environ. Health Perspect.* 118, 1109–1116.
- Heinen, E., Handy, S., 2012. Similarities in attitudes and norms and the effect on bicycle commuting: evidence from the bicycle cities Davis and Delft. *Int. J. Sustain. Transport* 6 (5), 257–281.
- Ho, C.I., Liao, T.Y., Huang, S.C., Chen, H.M., 2015. Beyond environmental concerns: using means-end chains to explore the personal psychological values and motivations of leisure/recreational cyclists. *J. Sustain. Tourism* 23 (2), 234–254.
- Ingvardson, J.B., Kaplan, S., de Abreu e Silva, J., di Ciommo, F., Shiftan, Y., Nielsen, O.A., 2020. Existence, relatedness and growth needs as mediators between mode choice and travel satisfaction: evidence from Denmark. *Transportation* 47, 337–358.
- Jahanshahi, D., Tabibi, Z., van Wee, B., 2020. Factors influencing the acceptance and use of a bicycle sharing system: applying an extended Unified Theory of Acceptance and Use of Technology (UTAUT) Case Studies on Transport Policy (in press).
- Janke, J., Handy, S., 2019. How life course events trigger changes in bicycling attitudes and behavior: insights into causality. *Travel Behav. Soc.* 16, 31–41.
- Kaplan, S., Manca, F., Nielsen, T.A.S., Prato, C.G., 2015. Intentions to use bike-sharing for holiday cycling: an application of the theory of planned behavior. *Tourism Manage.* 47, 34–46.

- Kaplan, S., Prato, C.G., 2016. "Them or Us": perceptions, cognitions, emotions, and overt behavior associated with cyclists and motorists sharing the road. *Int. J. Sustain. Transport*. 10 (3), 193–200.
- Kaplan, S., Wrzesinska, D.K., Prato, C.G., 2018a. The role of culture and needs in the cycling habits of female immigrants from a driving-oriented to a cycling-oriented country. *Transport. Res. Rec.* 2672 (3), 155–165.
- Kaplan, S., Wrzesinska, D.K., Prato, C.G., 2018b. The role of human needs in the intention to use conventional and electric bicycle sharing in a driving-oriented country. *Transport Policy* 71, 138–146.
- Kaplan, S., Wrzesinska, D.K., Prato, C.G., 2019a. Psychosocial benefits and positive mood related to habitual bicycle use. *Transport. Res. Part F: Traffic Psychol. Behav.* 64, 342–352.
- Kaplan, S., Mikolasek, I., Foltynova, H.B., Janstrup, K.H., Prato, C.G., 2019b. Attitudes, norms and difficulties underlying road sharing intentions as drivers and cyclists: evidence from the Czech Republic. *Int. J. Sustain. Transport*. 13 (5), 350–362.
- Kaplan, S., Luria, R., Prato, C.G., 2019c. The relation between cyclists' perceptions of drivers, self-concepts and their willingness to cycle in mixed traffic. *Transport. Res. Part F* 62, 45–57.
- de Kruijff, J., Ettema, D., Dijst, M., 2019. A longitudinal evaluation of satisfaction with e-cycling in daily commuting in the Netherlands. *Travel Behav. Soc.* 16, 192–200.
- Lanzendorf, M., Busch-Geertsema, Annika, 2014. The cycling boom in large German cities-empirical evidence for successful cycling campaigns. *Transport Policy* 36, 26–33.
- Leyland, L.A., Spencer, B., Beale, N., Jones, T., van Reekum, C.M., 2019. The effect of cycling on cognitive function and well-being in older adults. *PLoS ONE* 14, e0211779.
- Lindenberg, S., Steg, L., 2007. Normative, gain and hedonic goal frames guiding environmental behavior. *J. Social Issues* 63, 117–137.
- Lois, D., Moriano, J.A., Rondonella, G., 2015. Cycle commuting intention: a model based on theory of planned behaviour and social identity. *Transport. Res. Part F* 32, 101–113.
- Luth, M.T., Flinchbaugh, C.L., Ross, J., 2017. On the bike and in the cubicle: the role of passion and regulatory focus in cycling and work satisfaction. *Psychol. Sport Exerc.* 28, 37–45.
- Maioli, H.C., Correa de Carvalho, R., Dumke de Medeiros, D., 2019. SERVBIKE: riding customer satisfaction of bicycle sharing service. *Sustain. Cities Soc.* 50, 101680.
- Møller, M., Hausteine, S., 2017. Anger expression among Danish cyclists and drivers: a comparison based on mode specific anger expression inventories. *Accid. Anal. Prev.* 108, 354–360.
- Nkurunziza, A., Zuidgeest, M., Brussel, M., Van Maarseveen, M., 2012. Examining the potential for modal change: motivators and barriers for bicycle commuting in Dar-es-Salaam. *Transport Policy* 24, 249–259.
- O'Connor, J.P., Brown, T.D., 2010. Riding with the sharks: serious leisure cyclist's perceptions of sharing the road with motorists. *J. Sci. Med. Sport* 13, 53–58.
- Prati, G., Puchades, V.M., Pietrantonio, L., 2017. Cyclists as a minority group? *Transport. Res. Part F* 47, 34–41.
- Prochaska, J.O., DiClemente, C.C., 1983. Stages and processes of self-change of smoking: toward an integrative model of change. *J. Consult. Clin. Psychol.* 51 (3), 390–395, doi:10.1037/0022-006X.51.3.390.
- Rowe, K., Shilbury, D., Perkins, L., Hinckson, E., 2016. 417–430 Challenges for sport development: women's entry level cycling participation. *Sport Manage. Rev.* 19, 417–430.
- Ryan, R.M., Deci, E.L., 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychologist* 55, 68–78.
- Sigurdardottir, S.B., Kaplan, S., Møller, M., Teasdale, T.W., 2013. Understanding adolescents' intentions to commute by car or bicycle as adults. *Transport. Res. Part D* 24, 1–9.
- Sharma, B., Nam, H.K., Yan, W., Kim, H.Y., 2019. Barriers and enabling factors affecting satisfaction and safety perception with use of bicycle roads in Seoul South Korea. *Int. J. Environ. Res. Public Health* 16, 773.
- Sheller, M., Urry, J., 2006. The new mobilities paradigm. *Environ. Plann. A: Econ. Space* 38, 207–226.
- Shoval, N., Schvimer, Y., Tamir, M., 2018. Real-time measurement of tourists' objective and subjective emotions in time and space. *J. Travel Res.* 57, 3–16.
- Skov-Petersen, H., Jacobsen, J.B., Vedel, S.E., Nielsen, T.A.S., Rask, S., 2017. Effects of upgrading to cycle highways - an analysis of demand induction, use patterns and satisfaction before and after. *J. Transport Geogr.* 64, 203–210.
- Snizek, B., Nielsen, T.A.S., Skov-Petersen, H., 2013. Mapping bicyclists' experiences in Copenhagen. *J. Transport Geogr.* 30, 227–233.
- Spotswood, F., Chatterton, T., Tapp, A., Williams, D., 2015. Analysing cycling as a social practice: an empirical grounding for behaviour change. *Transport. Res. Part F* 29, 22–33.
- Taubman-Ben-Ari, O., Katz-Ben-Ami, L., 2013. Family climate for road safety: a new concept and measure. *Accid. Anal. Prev.* 54, 1–14.
- Thigpen, C., Fischer, J., Nelson, T., Therrien, S., Fuller, D., Gauvin, L., Winters, M., 2019. Who is ready to bicycle? Categorizing and mapping bicyclists with behavior change concepts. *Transport Policy* 82, 11–17.
- De Vos, Tim Schwanen, Veronique Van Acker, Frank Witlox, 2019. Do satisfying walking and cycling trips result in more future trips with active travel modes? An exploratory study. *Int. J. Sustain. Transport*. 13 (3), 180–196.
- Wardman, M., Tight, M., Page, M., 2007. Factors influencing the propensity to cycle to work. *Transport. Res. Part A* 41, 339–350.
- Westman, J., Olsson, L.E., Gärling, T., Friman, M., 2017. Children's travel to school: satisfaction, current mood, and cognitive performance. *Transportation* 44, 1365–1382.
- Willis, D.P., Manaugh, K., El-Geneidy, A., 2013. Uniquely satisfied: exploring cyclist satisfaction. *Transport. Res. Part F* 18, 136–147.
- Winters, M., Davidson, G., Kao, D., Teschke, K., 2011. Motivators and deterrents of bicycling: comparing influences on decisions to ride. *Transportation* 38, 153–168.
- Zander, A., Passmore, E., Mason, C., Rissel, C., 2013. Joy, exercise, enjoyment, getting out: a qualitative study of older people's experience of cycling in Sydney, Australia. *J. Environ. Public Health*, Article ID 547453.
- Zhang, Y., Mi, Z., 2018. Environmental benefits of bike sharing: a big data-based analysis. *Appl. Energy* 220, 296–301.
- Zhao, C., Ma, H., Yang, L., Xiao, Y., 2017. Long-term bicycle riding ameliorates the depression of the patients undergoing hemodialysis by affecting the levels of interleukin-6 and interleukin-18. *Neuropsych. Dis. Treatment* 13, 91–100.

Further Reading

- Moos, R.H., 1987. Person-environment congruence in work, school, and health care settings. *J. Vocat. Behav.* 31, 231–247.

Motorcyclists

Narelle Haworth, Queensland University of Technology (QUT), Centre for Accident Research and Road Safety-Queensland, Brisbane, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Definitions	144
Conflicting Perspectives	144
Diversity in Motorcycling	145
Purposes of Riding	145
Motorcycle Types	145
Riding Locations	146
Rider Demographics	146
Factors Contributing to Motorcycle Crashes	146
Failure by Drivers to See Motorcycles	146
Inexperience	146
Risky Attitudes and Behaviors	147
Measures to Improve Motorcycle Safety	147
Rider Training and Licensing	147
Enforcement of Road Rules	148
Improving Conspicuity	148
Helmets and Protective Clothing	148
Safer Motorcycles	148
The Future of Motorcycling	148
References	149
Further Reading	150

Definitions

The term “motorcyclist” includes both the rider who is controlling the motorcycle and any passengers who are seated behind (pillion passengers) or in a sidecar or other attachment. However, the definition and scope of the term motorcycle is more contested. The term “powered two-wheeler” (PTW) has been used in European publications to refer to mopeds, scooters, and motorcycles; and commonly includes similar three-wheeled vehicles. Many jurisdictions define mopeds in terms of engine capacity (usually lower than 50 cc) and top speed (often lower than 50 km/h). In general, mopeds and scooters have a “step-through” design and automatic transmission, while motorcycles must be straddled by the rider and have manual transmissions. There is no official definition of a scooter and they are recorded as mopeds or motorcycles in official databases, depending on their engine size and top speed.

The recent growth in sales of e-bikes and e-scooters has blurred the distinction with mopeds and scooters. Up to 60% of e-bikes sold in China (the world’s largest market) are faster and/or heavier than legally allowed (Wei et al., 2013), suggesting that they should be classified as electric mopeds or motorcycles. The “speed e-bikes” in Europe are sometimes regulated as mopeds. New classification systems have emerged in response to the proliferation of new technology. Under the Society of Automotive Engineers (SAE, 2019) new standard - SAE J3194 Taxonomy and classification of powered micromobility devices—a moped would likely be classified as a Mid-Weight Plus, Standard-width, Medium-speed seated scooter (WT4/WD1/SP3/C seated scooter).

Conflicting Perspectives

Motorcyclists and motorcycles polarize community views. PTWs have appeal for transport because of their low cost, high fuel efficiency, comparable speeds to cars, high manoeuvrability, less space needed for parking and ability to weave through traffic jams (Tiwari, 2015). Motorcycling for recreation is attractive as a lifestyle choice because of its image, the thrill of riding, the feeling of freedom, and to impress others (Watson et al., 2007). Yet many people in the general community, and in road safety, see motorcycling as a dangerous pursuit and motorcyclists as risk-takers. These factors are clearly of interest to traffic psychology.

Motorcycles are popular. Between 2013 and 2016, the number of two- and three-wheelers operating in the world increased by 10% and in 10 countries motorcycles make up more than 70% of the vehicle fleet (WHO, 2018). Motorcycles comprise a small minority of registered vehicles in many high income countries (e.g. 3% in the US in 2017, 4.5% in Australia in 2019) but their numbers increased by 57% in the US and 79% in Australia between 2001–05 and 2013–17 (ABS, 2005, 2007; BITRE, 2011, 2020; NCSA, 2019).

Table 1 Inter-relationships between motorcycle type, rider characteristics, purpose and location of riding and crash risk factors and characteristics

<i>Common characteristics</i>	<i>Recreation</i>	<i>Transport</i>	<i>Occupational</i>
Type of motorcycle	Larger or off-road	Often smaller, including mopeds and scooters	Often smaller, including mopeds and scooters
Location	Often rural	Largely urban	Largely urban but some farm use
Type of rider	Older and returning riders	Mixed	Lower income in LMICs
Risk factors	Road related factors (incl. curves), inexperience (incl. returning riders), speeding, drink riding	Driver failure to see, inexperience (esp. braking), filtering, speeding, lack of protective clothing	Nonuse of helmets in LMICs, esp. by passengers, and on farms, overloading, poor maintenance, speeding
Crash characteristics	Single vehicle incl. run-off-road and hit roadside objects, esp. on curves	Multivehicle esp. at intersections	Multivehicle esp. at intersections in urban areas, rollover on farms

Globally, 28% of those killed in road crashes are using motorized two- and three-wheelers. In the WHO South-East Asia and Western Pacific regions, these riders comprise 43% and 36% of all deaths, respectively (WHO, 2018). Despite their lower usage, they comprise 14% of all road crash fatalities in the US (2017; NCSA, 2019), 19% in Great Britain (2019; DfT, 2020c) and 18% in Australia (2019; BITRE, 2020). Along with bicyclists and pedestrians, motorcyclists are often classed as ‘vulnerable road users’: motorcyclists were nearly 27 times more likely to be killed than a car occupant per vehicle mile traveled in the US (NCSA, 2019) and 63 times more likely to be killed than a car occupant per billion passenger miles in Great Britain (2019, DfT, 2020c).

Road safety agencies and police are concerned about the increasing numbers of crashes and fatalities, but transport professionals and the motorcycle industry and enthusiasts point to the declining fatality rate. Indeed, the fatality rate per 10,000 registered motorcycles in the US dropped by 16% (from 6.9 to 5.8) from 2001–05 to 2013–17, while in Australia it dropped by 43% (from 5.5 to 3.1).

Diversity in Motorcycling

There are complex relationships among motorcycle type, rider characteristics, purpose of riding and distance ridden which influences the magnitude and nature of the resulting safety issues. Table 1 attempts to summarize these factors which are then discussed.

Purposes of Riding

Recreational riding is more common than commuter riding in North America, the United Kingdom, Northern Europe and Australasia but the opposite pattern is observed in Southern Europe, and in low- and middle-income countries (LMICs). In many LMICs, motorcycles are the only affordable means of motorized transport and high levels of congestion and inadequate public transport make motorcycles the most practical form of transport (OECD/ITF, 2015). Occupational use of motorcycles is common in some countries for postal deliveries, and in many LMICs motorcycle taxis have become widespread, where customers are transported as pillion passengers.

Motorcycle Types

Purpose of riding influences choice of type of motorcycle. Recreational riding is associated with larger motorcycles which can travel faster and provide more comfort. Smaller motorcycles and scooters are more common for commuter and occupational riding, given that they are less expensive to purchase and operate and more manoeuvrable in dense traffic.

Regulations can also affect choice of motorcycle type. Mopeds are much more common in countries where they are exempt from or are subject to less stringent licensing and registration requirements. In some European countries (e.g. the Netherlands, France), the minimum age for a moped licence is low, so moped use by young riders presents a significant road safety issue (Aare and Holst, 2003). Official estimates of moped numbers are not available in most global documents because they are not required to be registered in some countries.

Blackman and Haworth (2013) reported that the crash rates per registered vehicle were similar for mopeds, scooters and motorcycles, but the moped crash rate per vehicle-kilometer traveled was nearly four times that of motorcycles and larger scooters because the average distance traveled per year by mopeds was much lower. Amongst larger motorcycles, considerable research has focused on the extent to which engine capacity and power-to-weight ratio influence crash risk and severity, given that these characteristics can be regulated in licensing and registration systems. An analysis of Australian motorcycle crashes (Budd et al., 2018) found that crash risk increased with power-to-weight ratio but crash severity was more strongly influenced by engine size.

Super sport and sports types appear to be among the least safe motorcycle types (Budd et al., 2018; Teoh and Campbell, 2010) but comparisons of other types have produced varying results. Comparing the safety of different types of motorcycles is challenging because some types are ridden further per year than others and comparisons based on crashes per registered vehicle may mask higher crash risks per distance traveled. Another challenge in interpreting variations in observed safety among motorcycle types is that the characteristics of their riders differ, and these characteristics likely affect how safely they are ridden.

Riding Locations

Riding for commuting purposes occurs mostly on urban, relatively low-speed roads while recreational riders favor rural, high-speed, winding roads. These rural roads generally have lower in-built safety standards, poor maintenance and little police enforcement which together contribute to the higher crash risk of recreational riding (Haworth et al., 1997; Moskal et al., 2012).

Off-road motorcycle riding occurs for recreation ("dirt bike riding", or "trail bike riding") and agricultural purposes, although All-Terrain Vehicles (ATVs) have increasingly replaced motorcycles for this purpose. The amount of off-road riding is likely to be significant in countries because where off-road motorcycles comprise a large segment of motorcycles sold (e.g. US: 21% in 2019; Statista, 2020 and Australia: 38% in 2019; FCAI, 2020). In Australia, similar numbers of motorcyclists are hospitalised following on- and off-road incidents (AIHW, 2018).

Rider Demographics

Younger, older and returning riders

In many high-income countries, the growth in recreational riding has largely been among men who are middle aged and over. A US survey (MIC, 2019) reported that the median age of motorcycle owners in 2018 was 50, rising from 45 in 2012. Budd et al. (2018) reported that in Victoria, Australia, the number of motorcycle licences issued to older riders grew by almost 300% and the learner permits doubled over the period 2006–15. Older motorcyclists were more likely ride on open roads which at higher speeds makes them more vulnerable. The proportion of injury crashes involving riders aged 60 years and over doubled to 7% from 2005 to 2014 and these riders had poorer injury outcomes.

Older riders comprise continuing, new and returning riders, with the proportions depending on the definition used (Haworth et al., 2013). While findings vary, the crash risk of returning riders appears to be elevated in terms of distance ridden, but not per licence held. An elevated crash risk could result from a deterioration in motorcycle handling skills due to lack of practice, changes in motorcycle design and performance over time leading to unfamiliarity with the motorcycle, or rider attitudes and behaviors (Mulvihill and Symmons, 2010).

Despite the growth in numbers of older and returning riders and their contribution to crash numbers, crash rates remain highest for the youngest riders who, while numerically small, are inexperienced with both riding and the traffic system, and share the risk-taking characteristics of their peers.

Women who ride

Across high-income countries, a large majority of motorcyclists are male, although relatively more women ride scooters. The number of female motorcycle riders appears to be increasing in the US, with surveys showing an increase in the proportion of motorcycle owners who were women from 10% in 2009 to 19% in 2018 (MIC, 2018). In contrast, the representation of women among candidates passing the practical motorcycle tests in the UK has remained between 7.5 and 9.5% since 2009/10 (DfT, 2020a,b). Relatively little is known about the characteristics of women who ride, although a media release from a UK insurer (The Bike Insurer, 2018) stated that 8% of insurance quotes in 2017–18 were provided to women, whose average age was 37. The most common engine size was 125 cc, and average annual mileage was 3,000 miles (4800 km). Little information is available to compare the safety of male and female riders, although women appear to be more interested in safety-focused motorcycle training than men.

Factors Contributing to Motorcycle Crashes

The psychosocial factors contributing to motorcycle crashes include failure by drivers to see motorcycles, rider inexperience, and risky attitudes and behaviors. In addition, because they have only two wheels, motorcycles are much more sensitive to deficits in road design and maintenance and the unprotected nature of riders mean that collisions with roadside furniture and other objects contributes significantly to the severity of injuries in single-vehicle crashes. A more detailed summary of the factors contributing to motorcycle crashes is provided in OECD/ITF (2015).

Failure by Drivers to See Motorcycles

Many motorcycle crashes occur because of errors or violations by drivers of other motor vehicles. For example, almost half of the fatal motorcycle crashes in the US in 2017 involved collisions with other motor vehicles and in 42% of the two-vehicle crashes, the other vehicle was turning left across the path of the motorcycle which was going straight (NCSA, 2019). In many crashes of this type, the other driver claims to have not seen the motorcyclist. Drivers' failure to see motorcyclists results from poor sensory (or physical conspicuity) because the motorcycle is a small visual target and poor cognitive conspicuity because drivers do not expect to see motorcycles in countries where motorcycles make up a small proportion of the traffic stream.

Inexperience

Like novice drivers, inexperienced riders have higher crash risks, possibly related to their poorer braking skills or less developed hazard perception. Some inexperienced riders are unlicensed, inflating the representation of inexperienced riders in crashes. In the US in 2017, 29% of motorcycle riders in fatal crashes did not have valid licences (NCSA, 2019). In Australia, Budd et al. (2018) found that the odds of a more severe injury crash outcome were 25% higher if the rider was unlicensed, but it is unclear whether this reflects a true increase or unwillingness to report less severe crashes.

However, unlike drivers, many riders do not ride much and so remain inexperienced. For example, in the US in 2017, motorcycles traveled only 2312 miles per year on average. Motorcyclists who ride less have relatively higher crash risks per unit of distance traveled than those who ride more often (Harrison and Christie, 2005). Riding a new or borrowed is also associated with increased crash risk (Haworth et al., 1997).

Risky Attitudes and Behaviors

Psychosocial factors have largely been studied in relation to recreational riding, because most research has been undertaken in high-income countries where recreational riding is more common. However, the psychological underpinnings of observed behaviors may differ among countries. In some LMICs, nonwearing of helmets may reflect lack of laws and enforcement, and potentially tropical climates, rather than deliberate risk-taking. Similarly, nonwearing helmets may be a much stronger indication of risky attitudes in Australia (where nonhelmet use on public roads is rare) than in some of the US States where helmet use is not compulsory.

Motorcycling shares some psychosocial aspects with other forms of mobility, such as attribution and calibration errors and influences from peers.

The poor safety record of motorcycling may partly result from it being attractive to people with an increased propensity for risk taking (Horswill and Helman, 2003; Tunncliffe et al., 2012). In addition, some motorcycling research (e.g. Tunncliffe et al., 2012) suggests that the influence of “people I ride with” may be particularly important and that some risky behaviors may be motivated by a desire to keep up with the group.

Speeding is more common in motorcycle than car crashes, with 32% of motorcycle riders in fatal crashes being speeding versus 18% of drivers in the US in 2017 (NCSA, 2019). Motorcycles often travel faster than other vehicles in urban areas, which can result in crashes because they are traveling faster through intersections than other road users expect (OECD/ITF, 2015).

While alcohol is clearly incompatible with safe motorcycling, the extent of its involvement in motorcycle crashes differs widely across countries. In some LMICs where alcohol consumption rates are very high and motorcycles are the most common form of transport, a large proportion of crashes involve alcohol. In the US in 2017, 28% of motorcycle riders killed in motor vehicle crashes had an illegal level of alcohol (BAC 0.08 g/dL or higher) and another 7% had lower levels of alcohol (NCSA, 2019). Alcohol is generally more prevalent in motorcycle riders than car drivers in fatal crashes in the US but not in some other countries (e.g., Australia, France, Sweden are mentioned in OECD/ITF, 2015). In common with other vehicles, alcohol is more common in single- than multi-vehicle motorcycle crashes, and in fatal crashes than injury crashes.

Risky behaviors tend to co-occur which makes it difficult to assess the relative contribution of each factor to crash risk and severity (see review by Theofilatos and Yannis, 2015). For example, a host of studies in high-income countries have found that not wearing a helmet and drink riding co-occur, often also among males who are riding for recreation. Speeding and being unlicensed also tend to be part of this constellation of risky behaviors.

There is controversy regarding whether lane filtering is risky and its prevalence and legal status varies among jurisdictions (Clabaux et al., 2017). Many riders maintain that filtering reduces their risk of being hit from behind by a car, or that it prevents their motorcycle's engine overheating in slow-moving traffic. A series of French studies (Clabaux et al., 2017, 2019) demonstrated that filtering on urban roads increased crash risk four-fold, that the risk of hitting pedestrians in city centers was more than five times higher for filtering than for non-filtering motorcycles, and that filtering more than doubled crash risk on urban freeways. The increased risk may relate to filtering riders being less detectable or predictable by drivers, or other aspects of the risky behavior of these riders (Clabaux et al., 2017).

Riding without adequate protective clothing can increase crash severity and is an objectively risky behavior but it does not appear to correlate with other risky behaviors and attitudes (de Rome et al., 2011).

Measures to Improve Motorcycle Safety

Many of the general road safety measures (e.g. speed limits that align with infrastructure safety, reduction in impaired driving) also improve motorcycle safety, particularly to the extent that these measures contribute to drivers of larger vehicles being less likely to collide with motorcycles. Improvements to the road environment may be the most effective measures prevent or reduce the severity of crashes resulting from the factors outlined in the previous section. For example, dedicated turning lanes and installation of traffic signals (or changes to the phasing of existing signals) can reduce turn-in-front-of crashes that occur because turning drivers fail to see motorcycles. Improvements to skid resistance and widening of shoulders on curves can prevent run-off-road crashes on curves where motorcycles are deliberately or inadvertently speeding.

Rider Training and Licensing

Licensing requirements vary among jurisdictions and often depend on how the vehicle is classified. Internationally, the approach to moped licensing varies from no training or licence requirement to moped riders requiring a motorcycle licence. In some countries, the licensing age for mopeds is lower than for cars and motorcycles, and moped use is seen as an introduction to motorized independent mobility.

Vehicle-based restrictions are common in licensing systems, such as restricting the engine capacity and/or power-to-weight ratio of motorcycles that novices are allowed to ride, or requiring additional training and/or testing to allow access to more powerful motorcycles. While some of these requirements may have reduced motorcycle crashes by making motorcycling less attractive, their

effectiveness in reducing crash risk among those who ride has not been clearly demonstrated, possibly because confounding effects of age, type of motorcycle and distance ridden (Budd et al., 2018).

Attempts have been made to adapt graduated licensing schemes used for car licensing to motorcycle licensing but the differences in age and experience profiles between novice drivers and novice riders, the impracticality of supervision and the potential for slow accumulation of experience mean that simple adaptations may not be appropriate (Haworth and Rowden, 2010).

Evaluations of the effectiveness of training in improving rider safety have been inconclusive (e.g. Kardamanidis et al., 2010). Many of the same issues arise as in training as for car drivers—skills-based training potentially leading to over-confidence and greater likelihood to indulge in risky behaviors. There is a trend to encourage training to focus more on modifying rider attitudes and higher-order cognitive skills such as hazard perception but to motivate riders and trainers there is a need to include these components in licence testing. Post-licence training is undertaken by relatively few riders and often focuses on further development of vehicle control skills, which can result in increased risk-taking. Many authorities promote voluntary refresher training for returning riders but this is completed by relatively few returning riders.

Enforcement of Road Rules

Enforcement of road rules to deter illegal behaviors such as speeding is generally less effective for motorcyclists than other road users. Motorcycles can accelerate quickly, travel at high speeds and filter through traffic, making it difficult for police (in cars) to safely apprehend some high-risk riders. Difficulties in identifying motorcycle licence plates using camera-based (electronic) speed enforcement may be contributing to riders being less deterred from speeding in some jurisdictions. Additionally, there can be a mismatch between allocation of enforcement resources and motorcycle use. For example, most motorcycle riding occurs in daytime but most alcohol enforcement occurs at night to follow the drinking patterns of car drivers. Similarly, a lot of speeding by motorcycles occurs on rural roads which traditionally have lower levels of police enforcement because of their lower traffic volumes.

Improving Conspicuity

Requirements for hard-wiring motorcycle headlamps so that they are on even in daytime have reduced the number of multi-vehicle motorcycle crashes in a range of countries (see OECD/ITF, 2015). There is debate regarding whether the introduction of similar requirements for other vehicles has reduced the effectiveness of this measure for motorcycles. The European 2BESAFE project developed and tested a range of alternative headlamp configurations for motorcycles and identified a T-shaped arrangement was most effective.

There is some dispute regarding whether wearing brighter clothing improves the daytime detectability of motorcycles. Experimental studies have suggested that contrast with the background may be more important than brightness.

Attempts to increase drivers' expectancy of motorcyclists have shown only limited and short-term success (Hosking et al., 2007).

Helmets and Protective Clothing

Correct helmet use is associated with a 42% reduction in the risk of fatal injuries and a 69% reduction in the risk of head injuries to motorcyclists (Liu et al., 2008). For this reason, the focus of international efforts has been on increasing helmet use, particularly by enforcement of legislation that requires wearing of helmets that comply with performance standards (WHO, 2018).

Use of effective protective clothing can reduce the severity of non-fatal injuries to motorcyclists in crashes (de Rome et al., 2011). However, the protective performance of garments varies dramatically and riders may not be convinced of the protective value of what can be expensive items. To address this, various approaches have been developed to test protective clothing and/or provide information to consumers (de Rome, 2018).

Safer Motorcycles

Most of the effort in improving the safety of motorcycles has focused on active safety systems which prevent crashes, although development of passive systems such as motorcycle airbags has received some attention (see OECD/ITF, 2015). The effectiveness of antilock braking systems (ABS) has been identified in a range of studies (summarized by Savino et al., 2020). In a recent review, Savino et al. (2020) identified the following active safety systems for PTWs: antilock braking system, autonomous emergency braking, collision avoidance, intersection support, intelligent transportation systems, curve warning, human machine interface systems, stability control, traction control, and vision assistance. They noted that these systems are at different levels of development and there is a need for studies to examine the combined safety benefit of more than one technology.

One of the challenges faced continues to be rider unwillingness to adopt some new technologies, particularly those which reduce or limit rider control of the motorcycle (Beanland et al., 2013). Another factor limiting acceptance is affordability, particularly in LMICs and for lower-priced PTWs.

The Future of Motorcycling

Given increasing levels of urbanization and traffic congestion, the demand for motorcycles for transport is expected to increase. There has been a tendency to move from PTWs to cars in some LMICs as incomes increase, but congestion levels may result in owning both a car and a motorcycle and using them for different purposes (Eccarius and Liu, 2020). Bans on petrol motorcycles and tighter emission controls are expected to continue, leading to an expansion in the use of electric motorcycles (Eccarius and Liu, 2020). New vehicle categories may replace PTWs for short urban trips, with emerging evidence that standing electric scooters may be replacing mopeds in some environments. As part of the move towards a shared mobility, dockless motorcycle sharing schemes have begun in

Europe, largely using battery electric, fully automatic scooters. These schemes may play a role in future Mobility-as-a-Service (MaaS) systems (Eccarius and Liu, 2020). For recreation, 'time share' of motorcycles is growing in the US. While autonomous motorcycles are conceptually possible, rider demand may be low.

References

- Aare, M., Holst, H., 2003. Injuries from motorcycle and moped crashes in Sweden from 1987 to 1999. *Injury Control Safety Promotion* 10 (3), 131–138.
- Australian Bureau of Statistics (ABS), 2005. Survey of Motor Vehicle Use 9208.0 12 months ended 31 October 2004. Canberra, ACT.
- Australian Bureau of Statistics (ABS), 2007. Survey of Motor Vehicle Use 9208.0 12 months ended 31 October 2006. Canberra, ACT.
- Australian Institute of Health and Welfare (AIHW), 2018. Hospitalised injury due to land transport crashes. Cat. No. INJCAT 195. Injury research and statistics series no. 115. AIHW, Canberra. <https://www.aihw.gov.au/getmedia/2c2bb67a-4a4a-4b45-a030-aaa977c7292c/aihw-injcat-195.pdf.aspx?inline=true>.
- Beanland, V., Lenné, M.G., Fuessl, E., Oberlander, M., Joshi, S., Bellet, T., Banet, A., ReoBger, L., Leden, L., Spyropoulou, I., et al., 2013. Acceptability of rider assistive systems for powered two-wheelers. *Transport. Res. Part F Traffic Psychol. Behav.* 19, 63–76.
- Blackman, R., Haworth, N., 2013. Comparison of moped, scooter and motorcycle crash risk and crash severity. *Accid. Anal. Prevent.* 57, 1–9.
- Budd, L., Allen, T., Newstead, S., 2018. Current trends in motorcycle-related crash and injury risk in Australia by motorcycle type and attributes. Report 336. Monash University Accident Research Centre, Melbourne.
- Bureau of Infrastructure, Transport and Regional Economics (BITRE), 2011. Road deaths Australia. 2010 statistical summary. Canberra, ACT. <https://www.bitre.gov.au/statistics/safety>.
- Bureau of Infrastructure, Transport and Regional Economics (BITRE), 2020. Road trauma Australia 2019 statistical summary. <https://www.bitre.gov.au/statistics/safety>.
- Clabaux, N., Fournier, J.Y., Michel, J.E., 2017. Powered two-wheeler riders' risk of crashes associated with filtering on urban roads. *Traffic Inj. Prev.* 18 (2), 182–187.
- Clabaux, N., Fournier, J.Y., Michel, J.E., Perrin, C., 2019. Does filtering by powered two-wheelers present a risk for pedestrians in city centers? *J. Transport Health* 13, 224–233.
- de Rome, L., 2018. Stars or standards? A review of motorcycle protective clothing from the southern hemisphere. www.racfoundation.org.
- de Rome, L., Ivers, R., Fitzharris, M., Du, W., Haworth, N., Heritier, S., Richardson, D., 2011. Motorcycle protective clothing: protection from injury or just the weather? *Accid. Anal. Prevent.* 43, 1893–1900.
- Department for Transport (DfT), 2020a. Practical motorcycle test (Module 1) pass rates by gender, monthly, Great Britain: April 2009 to March 2020. <https://www.gov.uk/government/organisations/department-for-transport/series/driving-tests-and-instructors-statistics>.
- Department for Transport (DfT), 2020b. Practical motorcycle test (Module 2) pass rates by gender, monthly, Great Britain: April 2009 to March 2020. <https://www.gov.uk/government/organisations/department-for-transport/series/driving-tests-and-instructors-statistics>.
- Department for Transport (DfT), 2020c. Reported road casualties in Great Britain: provisional results 2019. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/904698/rccgb-provisional-results-2019.pdf.
- Eccarius, T., Liu, C.-C., 2020. Powered two-wheelers for sustainable mobility: a review of consumer adoption of electric motorcycles. *Int. J. Sustain. Transport* 14 (3), 215–231.
- Federal Chamber of Automotive Industries (FCAI), 2020. Australia's new vehicle market. <https://www.fc.ai.com.au/sales>.
- Harrison, W.A., Christie, R., 2005. Exposure survey of motorcyclists in New South Wales. *Accid. Anal. Prevent.* 37, 441–451.
- Haworth, N., Blackman, R.A., Fernandes, R., Ma, A., 2013. Identifying and characterising crashes of returning riders – A new approach. Proceedings of the 2013 Australasian Road Safety Research, Policing & Education Conference, Brisbane, 28–30 August. <https://eprints.qut.edu.au/65002/>.
- Haworth, N., Rowden, P., 2010. Challenges in improving the safety of learner motorcyclists. In Suggett, J., (Ed.) Proceedings of the 20th Canadian Multidisciplinary Road Safety Conference. Canadian Association of Road Safety Professionals, CD Rom, 1–16. <https://eprints.qut.edu.au/37849/>.
- Haworth, N., Smith, R., Brumen, I., Pronk, N., 1997. Case-control Study of Motorcycle Crashes (CR174). Federal Office of Road Safety, Canberra.
- Horswill, M.S., Helman, S., 2003. A behavioral comparison between motorcyclists and a matched group of non-motorcycling car drivers: factors influencing accident risk. *Accid. Anal. Prevent.* 35, 589–597.
- Hosking, S., Mulvihill, C., Edquist, J., Lenne, M., Stephan, K., Symmons, M., Murdoch, C., Archer, J., 2007. Evaluation of associative learning methods to train drivers to give way to motorcyclists. Report 278. Monash University Accident Research Centre, Melbourne.
- Kardamanidis, K., Martiniuk, A., Ivers, R.Q., et al., 2010. Motorcycle rider training for the prevention of road traffic crashes. *Cochrane Database Syst. Rev.* 10, CD005240.
- Liu, B.C., Ivers, R., Norton, R., Boufous, S., Blows, S., Lo, S.K., 2008. Helmets for preventing injury in motorcycle riders. *Cochrane Database Syst. Rev.* CD004333.
- Moskal, A., Martin, J.L., Laumon, B., 2012. Risk factors for injury accidents among moped and motorcycle riders. *Accid. Anal. Prevent.* 49, 5–11.
- Motorcycle Industry Council (MIC), 2018. Motorcycle ownership among women climbs to 19 percent. <https://www.mic.org/#/newsArticle>.
- Mulvihill, C., Symmons, M., 2010. A comparison of rider attitudes and behaviours between crash-involved and non crash-involved returned motorcyclists. Paper presented at the 2010 Australasian Road Safety Research, Policing and Education Conference, 31 August – 3 September 2010, Canberra. <http://casr.adelaide.edu.au/rsr/RSR;1;2010/MulvihillIC.pdf>.
- National Center for Statistics and Analysis (NCSA), 2019. Motorcycles: 2017 data (Updated, Traffic Safety Facts. Report No. DOT HS 812 785). Washington, DC: National Highway Traffic Safety Administration. <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812785>.
- OECD/ITF, 2015. Improving safety for motorcycle, scooter and moped riders. OECD Publishing, Paris.
- SAE, 2019. SAE J3194TM Taxonomy & Classification of Powered Micromobility Vehicles. <https://www.sae.org/binaries/content/assets/cm/content/topics/micromobility/sae-j3194-summary—2019-11.pdf>.
- Savino, G., Lot, R., Massaro, M., Rizzi, M., Symeonidis, I., Will, S., Brown, J., 2020. Active safety systems for powered two-wheelers: a systematic review. *Traffic Injury Prevention* 21, 78–86.
- Statista, 2020. Estimated U.S. motorcycle sales in 2019, by type <https://www.statista.com/statistics/252264/us-motorcycle-sales-in-units-by-type/>.
- Teoh, E.R., Campbell, M., 2010. Role of motorcycle type in fatal motorcycle crashes. Insurance Institute for Highway Safety, Arlington, VA.
- The Bike Insurer, 2018. New data reveals true story of UK's female motorcycle and scooter riders. <https://www.thebikeinsurer.co.uk/motorbike-news/industry/uk-female-motorcyclists-revealed/> (Accessed 4 August 2020).
- Theofilatos, A., Yannis, G., 2015. A review of powered-two-wheeler behaviour and safety. *Int. J. Injury Control Safety Promotion* 22, 284–307.
- Tiwari, G., 2015. Safety challenges of powered two-wheelers. *Int. J. Injury Control Safety Promotion* 22, 281–283.
- Tunnicliff, D.J., Watson, B.C., White, K.M., Hyde, M.K., Schonfeld, C.C., Wishart, D.E., 2012. Understanding the factors influencing safe and unsafe motorcycle rider intentions. *Accid. Anal. Prevent.* 49, 133–141.
- Watson, B., Tunnicliff, D., White, K., Schonfeld, C., Wishart, D., 2007. Psychological and social factors influencing motorcycle rider intentions and behaviour. Report RSRG 2007-04. Australian Transport Safety Bureau, Canberra. https://eprints.qut.edu.au/9103/1/road_rgr_200704.pdf.
- Wei, L., Xin, F., An, K., Ye, Y., 2013. Comparison study on travel characteristics between two kinds of electric bike. *Procedia – Social Behav. Sci.* 96, 1603–1610, doi:10.1016/j.sbspro.2013.08.182.
- WHO, 2018. Global Status Report on Road Safety 2018. World Health Organization, Geneva.

Further Reading

- Allen, T., Newstead, S., Lenné, M., McClure, R., Hillard, P., Symmons, M., Day, L., 2017. Contributing factors to motorcycle injury Crashes in Victoria. Australia Transportation Research Part F 2017 45, 157–168.
- Hamzah, A., Solah, M.S., Ariffin, A.H., 2013. Electric motorcycle risks on roads: An overview. In Proceedings of the Southeast Asia Safer Mobility Symposium 2013. Retrieved from http://www.academia.edu/download/39953561/SEA-SMS2013-Azhar-Electric_motorcycle_risks_on_roads_An_Overview.pdf
- Haworth, N., 2012. Powered two wheelers in a changing world – Challenges and opportunities. *Accid. Anal. Prevent.* 44, 12–18.
- Haworth, N., Debnath, A.K., 2013. How similar are two-unit bicycle and motorcycle crashes? *Accid. Anal. Prevent.* 58, 15–25.
- NHTSA, 2019. Motorcycle safety plan. https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13507-motorcycle_safety_plan_050919_v8-tag.pdf.
- WHO. Powered two- and three-wheeler safety. 2017. A road safety manual for decision-makers and practitioners. Geneva: World Health Organization.

Drivers' Hazard Perception Skill

Mark S. Horswill, Andrew Hill, School of Psychology, The University of Queensland, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

What is Drivers' Hazard Perception Skill and Why is it Considered Important?	151
How can Hazard Perception Skill be Measured?	151
How does Hazard Perception Change across the Life Span?	153
How can Hazard Perception Skill be Improved?	153
The Impact on Crash Risk of Introducing Hazard Perception Testing and Training	155
Acknowledgment	155
References	155
Further Reading	157

What is Drivers' Hazard Perception Skill and Why is it Considered Important?

Those drivers who are better able to predict dangerous situations on the road ahead tend to have fewer crashes than others. This skill has been referred to as *hazard perception*, and it has been a focus of research for over 50 years because of the potential opportunities it offers for both predicting and reducing drivers' crash risk (Horswill, 2016a). Studies have linked hazard perception skill to crash involvement, both retrospectively (Boufous et al., 2011; Cheng et al., 2011; Congdon, 1999; McKenna and Horswill, 1999; Quimby et al., 1986; Tüské et al., 2019) and prospectively (Horswill et al., 2015b; Wells et al., 2008). For instance, Horswill et al. (2015b) found that drivers who initially failed a test of hazard perception used for driver licensing had 25% more crashes during the year following the test than those who passed the first time (excluding parking or reversing incidents, or crashes when the driver's vehicle was stationary). Hazard perception has been found to be associated with crash involvement across the life span, including in samples of novice drivers (Horswill et al., 2015b), drivers aged over 65 years (Horswill et al., 2010a), and cross-age fleet driver samples (Darby et al., 2009).

Hazard perception skill has also been linked with other important crash predictors. For instance, distractions of various kinds have been shown to reduce hazard perception performance among both novice and experienced drivers (Borowsky et al., 2014, 2015, Horswill and McKenna, 1999; Savage et al., 2013). Sleepiness also significantly impacts hazard perception skill (Hamid et al., 2014) in both novice and experienced drivers (Smith et al., 2009), although it has a greater effect on novice drivers. With regards to speeding, drivers with better hazard perception have been found to drive slower (Grayson et al., 2003) and drivers trained in hazard perception tend to choose slower speeds than untrained drivers (McKenna et al., 2006). Alcohol consumption has also been found to impair hazard perception performance (Deery and Love, 1996; West et al., 1993).

This body of evidence has led to the suggestion that problems with hazard perception are implicated in causing a significant proportion of crashes (Horswill, 2016b) in the sense that, when a driver crashes due to driving too fast or being distracted, fatigued, or drunk, the crash may be attributable on some level to a failure of hazard perception. For instance, the crash may have occurred because the driver's judgement was impaired (e.g., by alcohol, distraction, or sleepiness) to the extent that they failed to notice a hazard sufficiently early. Similarly, drivers who fail to recognize a potential hazard on the road ahead might be more willing to drive faster than is appropriate in that situation.

How can Hazard Perception Skill be Measured?

Hazard perception is usually measured using a computer-based *hazard perception test* in which drivers view video footage (or computer-generated images) of traffic situations and make a timed response when they predict that a potentially dangerous situation is developing (Horswill, 2016b; see Fig. 1). Computer-based tests are generally favored over on-road measurement of hazard perception performance for a number of reasons, including the following (see Horswill, 2016b). First, high-risk events are too infrequent in real driving to allow drivers' responses to be meaningfully assessed during a single typical journey. Second, measuring when a driver anticipates a hazard is difficult in real driving because this might occur sometime before they have to act. Third, it is difficult to standardize assessment in real driving as no two drivers will encounter exactly the same set of traffic situations. Fourth, there are ethical issues involved in deliberately exposing drivers to hazardous situations for testing purposes.

To give one example of a computer-based hazard perception assessment, the test currently used for driver licensing in Queensland, Australia (Wetton et al., 2011) involves test-takers viewing a series of traffic clips shot from a driver's perspective. Test-takers use a computer mouse to click on any other road user (i.e., vehicle, cyclist, or pedestrian) that is likely to become involved in a traffic conflict with their vehicle (i.e., the car containing the camera). A *traffic conflict* is defined as a situation in which the camera car needs to take action by slowing down or changing course to prevent a crash. The time taken to respond to each traffic conflict is



Figure 1 Screenshot from a hazard perception test video, showing a view of the road shot from the driver's perspective (Hill et al., 2019).

averaged over all conflicts to provide an overall score. Evidence indicates that scores from this type of hazard perception test (including similar tests that employ a simple button-press response mode) reflect real driving behavior, even though taking the test does not involve real driving. For example, hazard perception test scores predict on-road driver performance ratings (Grayson et al., 2003; Mills et al., 1998; Ross et al., 2013; Wood et al., 2013), as well as crash involvement (Horswill et al., 2015b), and are able to distinguish high-risk driver groups (young novices) from lower-risk groups (mid-age experienced drivers) (Borowsky et al., 2009, 2010; Horswill et al., 2008; Manley et al., 2020; McKenna and Crick, 1991; Smith et al., 2009; Wallis and Horswill, 2007; Wetton et al., 2010).

It is worth noting that it is not a trivial exercise to create an effective and valid hazard perception test. For instance, an early prototype of the UK licensing hazard perception test failed to predict crash risk and failed to distinguish novice and experienced drivers (Grayson and Sexton, 2002), prompting a significant redesign. Researchers have argued that this is because only certain types of hazard scenes can be used for measuring this skill (Crundall et al., 2012; Grayson and Sexton, 2002; Sagberg and Bjørnskau, 2006; Wetton et al., 2011). The consensus seems to be that a scene that is appropriate for a hazard perception test needs to involve sufficiently subtle anticipatory cues to allow drivers who are good at hazard perception to display their superiority over those who are not as good. Tests containing hazards that, under high-vigilance test conditions, are equally obvious to all drivers (e.g., where test developers have primarily chosen hazards because they appear to represent high-frequency crash types) have no power to tell apart good and bad drivers, which is the fundamental objective of the test (Horswill, 2016b).

Another method for measuring hazard perception skill (known as a *hazard prediction test*) involves showing drivers clips of potentially dangerous traffic incidents, stopping each clip just before the incident, and then asking the test-taker to predict what will happen next (Jackson et al., 2009). The quality of drivers' predictions has been found to be associated with self-reported crash involvement (Horswill et al., 2020), as well as distinguishing high-risk (young novices) and lower-risk (older experienced drivers) groups (Crundall, 2016; Horswill et al., 2020; Lim et al., 2014; Ventsislavova and Crundall, 2018; Ventsislavova et al., 2019).

Eye tracking has also been used to measure hazard perception skill, by determining whether and when drivers look at locations in a traffic scene relevant to anticipating hazards. For example, Pradhan et al. (2009) tracked drivers' gaze during real driving and examined fixations on stationary areas of the road environment that a skilled driver ought to be monitoring (e.g., a side road where the view into that road is obscured). Similar techniques have been used with drivers viewing video or computer-generated images of traffic scenes (Taylor et al., 2013). One advantage of this approach is that drivers do not have to perform any task that is not part of normal driving (e.g., generating predictions out loud or pressing a button) to obtain a score. However, one criticism of this technique is that it assumes that drivers are always paying attention to what they are looking at, which is known not to be the case. Also, while novice/experienced differences have been found for eye-scanning patterns (Underwood et al., 2003), no direct association with crash involvement has yet been shown for scores derived from eye-tracking data.

Other researchers have created tests of hazard perception in which test-takers drive a simulated vehicle through a virtual reality traffic environment. One method of scoring hazard perception skill in this type of assessment has been to measure vehicle speed

during the approach to a hazard, where drivers with superior hazard perception would be expected to slow earlier (Crundall et al., 2012). An advantage of this approach is that, unlike other methods, it indicates whether drivers are capable of taking appropriate action once they have anticipated a hazard. Also, the very act of having to control a vehicle is likely to influence visual scanning behavior, such that drivers' gaze patterns may be more representative of real driving than in tests where the driver is instead watching a video. On the other hand, giving test-takers control over the vehicle may change the nature of the hazards from driver-to-driver, undermining the comparability of test scores across participants (e.g., if a driver is driving more slowly than others, then they may not need to anticipate hazards as early to maintain the same level of safety). A further issue is that the computer-generated imagery typically used in these simulators is considerably less realistic than the equivalent filmed scene. If drivers are aware that a situation is fake, then this may change their attitudes and behaviors toward any hazards.

As previously noted, on-road measures of hazard perception are generally considered more problematic than computer-based measures for most purposes. However, some researchers (Pradhan et al., 2009) have used eye-tracking to measure hazard perception skill during real driving, as discussed above. Others have set up staged hazards on closed-circuit tracks (Lee et al., 2008; Wood, 2002) and measured performance using indicators such as the time that drivers first fixated the hazard and whether they behaved appropriately on the approach to the hazard. This method has yielded novice/experienced driver differences (Lee et al., 2008). Driving instructor ratings of hazard perception during a single journey have also been used (Wood et al., 2013), though this technique has the additional complication that the rater is not blind to the driver's appearance, which could potentially influence their judgement (e.g., they may be biased toward giving young-looking people worse ratings because they are aware that younger drivers tend to have poorer hazard perception skill).

Other methods of measuring hazard perception have involved static traffic images (Pradhan et al., 2009; Scialfa et al., 2013; Tüské et al., 2019), schematic plans of traffic situations (Pollatsek et al., 2006), and retrospective judgements of traffic clips (Renge et al., 2005; Vlakveld, 2014).

How does Hazard Perception Change across the Life Span?

Hazard perception varies systematically across the life span, with younger drivers having a lower level of competence than mid-age drivers (Horswill, 2016a). This age-related difference is highly confounded with driving experience, which it presumably reflects, and is in turn reflected in the relative crash risk of these two groups. However, hazard perception performance peaks at around 55 years (Quimby and Watts, 1981). By the time drivers reach 75 years of age, the decline in their hazard perception skill compared with mid-age drivers becomes significant (Horswill et al., 2009).

These changes in hazard perception performance across the life span have been conceptualized in terms of drivers' mental representations of the traffic environment (Horswill, 2016a). Learners are thought to start out with a relatively impoverished understanding of the road environment, whereby they are more likely to simply react to dangerous events rather than proactively predicting them (Horswill and McKenna, 2004). As they gain experience, their mental representation of the traffic environment becomes gradually more sophisticated. In practical terms, this means that more experienced drivers have a better understanding of the subtleties of where and when potential hazards might arise. This allows them to make more accurate predictions as to what might happen next in any given situation, allowing them to anticipate where other road users might appear from, how they might behave, and where they might go (Horswill, 2016c). This ability to predict dangerous events earlier is thought to lead to safer behavior because it provides drivers with more time to deal with any dangerous situations that might arise, and also allows them to take action to reduce the risk of potential hazards before they transpire (e.g., by slowing down when approaching a location where another road user could potentially emerge without warning). This mechanism reflects the concept of hazard perception as situation awareness on the road (Horswill and McKenna, 2004), where drivers with poor hazard perception have poorer perception and comprehension of what is happening around them and are consequently less able to predict what might happen next.

As drivers enter the later stages of life, new problems arise. Due to declines in cognitive capacities, older drivers may not be able to process all of the information required to feed their ongoing dynamic understanding of the road environment as effectively as before. Also, older drivers may find it more challenging to perceive certain stimuli, due to deficits in their useful field of view (i.e., they are less likely to notice stimuli that appear in peripheral vision) and reduced contrast sensitivity (i.e., lower contrast objects become harder to discern). In addition, gradual age-related declines in reaction time, which begin in the mid-20s (Thompson et al., 2014), eventually progress beyond the point where driving experience can compensate (Horswill et al., 2008).

How can Hazard Perception Skill be Improved?

Given that hazard perception appears to be an important skill for crash risk, and some drivers are considerably more skilled than others, one potential strategy for reducing crashes is to improve the hazard perception skill of those who have not reached an optimal standard. However, a potential problem with this approach is that driver training in general has not been found to be a reliable method of decreasing crash risk (Beanland et al., 2013). In fact, in some cases, driver training has been shown to increase crash involvement (Mayhew et al., 1998). One explanation for these failures is the effect of training on driver confidence. It has been suggested that, while training may indeed increase drivers' skills, it simultaneously increases their confidence in those skills (Gregersen, 1996). Increased confidence has been linked to increased risk-taking (Horswill et al., 2004), as drivers believe their

improved competence enables them to cope better with dangerous situations. This in turn negates the safety benefit of their improved skills. In other words, it is not a foregone conclusion that improving hazard perception skill via training will automatically lead to reductions in crash risk.

Fortunately, the available evidence suggests that hazard perception training can improve drivers' hazard perception skill without increasing their confidence or their propensity to drive more dangerously. In terms of raising skill levels, hazard perception training has been found to improve hazard perception test scores not only for novice drivers (Wetton et al., 2013) whose mental representations of the traffic environment might be underdeveloped, but also for older drivers whose hazard perception may have declined due to factors not related to lack of experience, such as deteriorating cognition, vision, and reaction times (Horswill et al., 2010b, 2015a). In fact, even mid-age experienced drivers have been found to benefit from hazard perception training (Horswill et al., 2013b). In addition, hazard perception training has been found not to increase drivers' self-ratings of their overall driving skill or their hazard perception skill (Horswill et al., 2013b), suggesting that any skill improvements are unlikely to be neutralized by accompanying increases in risk-taking. There is even some evidence that hazard perception training can actually lead to *decreases* in risk-taking intentions (McKenna et al., 2006).

The underlying goal of most hazard perception training interventions (either implicitly or explicitly) is to give trainees greater insight into the expert driver's mindset and their more sophisticated mental model of the traffic environment (Horswill, 2016c). It has been suggested that this approach prevents increases in self-confidence because trainees generally receive repeated feedback indicating that they are noticing far less than would an expert. Increases in risk-taking may also be prevented because drivers with improved hazard perception become more aware of the dangers facing them. In addition, increased risk-taking usually makes the act of perceiving hazards more difficult (e.g., faster drivers have less time to notice hazards; tailgating drivers' view of hazards may be blocked by the vehicle in front), and this is likely to be more apparent to an individual who is actively searching for sources of risk.

A number of different training techniques have been found to improve hazard perception skill. Perhaps the most well-established of these is the use of commentary drives (Åbele et al., 2018, 2019; Cantwell et al., 2013; Crundall et al., 2010; Isler et al., 2009, 2011; McKenna et al., 2006; Young et al., 2017). This typically involves the trainee being asked to generate a running verbal commentary either during real driving or while watching a video shot from the perspective of a driver. The trainee is asked to verbalize what they are looking at, and the dangers they are anticipating. They are then given feedback on the quality of their commentary. For example, in a video-based training exercise, they might watch the same footage overdubbed with an expert driver's commentary to compare with their own attempt, before repeating the process with a new clip. This is designed to provide trainees with insight into the expert driver's mental representation of driving. Commentary drive training has been found to improve response times in computer-based hazard perception tests in a number of studies (Isler et al., 2009; McKenna et al., 2006; Wetton et al., 2013). It has also been found to lead to behavioral changes in a driving simulator, including reduced crash rates, earlier speed decreases when approaching hazards, and earlier braking (Crundall et al., 2010).

Another method of training hazard perception involves prompting drivers where to direct their attention while viewing videos, photos, or computer-generated images of traffic (Agrawal et al., 2018; Pradhan et al., 2009; Yamani et al., 2014). This approach is underpinned by the finding that novice drivers have different visual scanning patterns from experienced drivers (Crundall and Underwood, 1998). For instance, they are less likely to look at locations in the traffic environment that ought to be monitored in case hazards develop (Fisher et al., 2006). To give one example, the first version of the "Risk Awareness and Anticipation Training" (RAPT) system, developed in the US (Pollatsek et al., 2006), involved asking trainees to position markers on a bird's eye schematic view of a driving situation to indicate where hidden (from the perspective of a given vehicle) road users might be located. Trainees were also asked to indicate the locations that the driver of the target vehicle should be monitoring in anticipation of potential hidden road users emerging in a way that might cause a hazard. In a subsequent driving simulator journey, trained participants were found to monitor relevant locations to a greater extent than untrained participants. A later version of RAPT involved showing trainees a series of driver-perspective still images depicting a developing road scene, and asking them to indicate where they ought to be monitoring. Trainees' learnings were found to translate into more effective monitoring of key locations on the road ahead during real driving (Pradhan et al., 2009).

The hazard prediction test method described earlier (in which a clip of traffic cuts to black, and the driver is asked to predict what could happen next) has also been used for training. In the training exercise version (McKenna and Crick, 1997; Poulsen et al., 2010; Wetton et al., 2013), the driver is provided with feedback on their predictions (e.g., they are shown an expert driver's predictions and also see what actually happened after the cut-point). This training approach has been associated with improved hazard perception test scores, both in isolation (Wetton et al., 2013) and in combination with other strategies (Horswill et al., 2013b, Poulsen et al., 2010). One version of this technique, called "HazardTouch" (Shimazaki et al., 2017) involved trainees watching videos of real crashes. This resulted in trained drivers fixating earlier on the edges of blind spots while watching videos of driving and checking those locations more frequently than controls. Trainees also selected slower speeds before intersections during a real-world drive, as well as checking the intersections more frequently and for longer as they passed them.

A key factor for effective skill acquisition that is lacking in everyday driving is high quality and unambiguous performance feedback (Horswill et al., 2017). Given that the average driver only crashes once every ten years (Evans, 1991), it can be argued that, for the vast majority of their driving time, drivers are receiving feedback that their driving is good (in the sense that they are mostly avoiding crashes). This is consistent with the finding that most drivers rate their own driving as better than that of the typical driver (Svenson, 1981), where such ratings have consistently been found to have close-to-zero correlations with objective performance measures (Horswill et al., 2013a). That is, drivers have little insight into their own driving skill. In contrast, most other skilled tasks, such as sports or musicianship, involve rather more frequent performance feedback. It has been argued that effective improvement of

skills can only occur in the presence of meaningful feedback (Ericsson and Poole, 2016) and that this explains why drivers typically take decades to reach their peak in hazard perception. Horswill et al. (2017) created a training exercise designed to address this issue more directly than other hazard perception training methods. It involved drivers responding to clips from a mouse-click response-time hazard perception test, and then being given high quality feedback on their performance. The feedback involved viewing another version of each clip, which had been annotated with crosshairs showing the location and timing of their own response, that of an expert driver, and that of an average driver (Horswill et al., 2017). Trainees were also shown a graph comparing their performance with these other driver groups. This intervention was found to improve response times in a subsequent hazard perception test.

Repetition learning using hazard clips has also been found to be effective in improving hazard perception skill. Kahana-Levy et al. (2019) presented a series of traffic clips to trainees, repeating each clip containing a hazard three times, separated by other clips. They found that drivers fixated hazards earlier when subsequently shown new clips (see also Meir et al., 2014).

However, not all attempts at training hazard perception skill have been successful. One method found to be ineffective is classroom instruction (McKenna and Crick, 1997; Meir et al., 2014), which is of note given its ubiquity as a training method in general.

One limitation of the hazard perception training literature to date is that most interventions have typically only been evaluated using short-term follow ups, ranging from immediately after the training has been completed to a week. For training to reduce crash risk, effects need to last considerably longer than this. However, the limited longer-term research that has been conducted has yielded promising results. For example, Chapman et al. (2002) reported significant learning retention between 3 and 6 months following hazard perception training, while Horswill et al. (2015a) reported no significant decay in performance improvements after three months.

The Impact on Crash Risk of Introducing Hazard Perception Testing and Training

As described above, we have evidence that hazard perception test scores are linked with crash involvement, and we have evidence that hazard perception training improves hazard perception test scores and affects behaviors associated with crash involvement. However, it does not necessarily follow from this that introducing hazard perception testing into driver licensing procedures, or making hazard perception training available, will translate into crash reductions. Nevertheless, the admittedly limited evidence that does address these crash effects shows promise. First, in some European countries, "new paradigm" driver education programs have been introduced, which have a substantial focus on hazard perception (Washington et al., 2011). These programs have been associated with reductions in crash risk, albeit with some methodological caveats (Washington et al., 2011). Second, there is one randomized control trial in the literature examining hazard perception training effects on crash rates (Thomas et al., 2016). This involved a 10-min computer-based training intervention, based on the RAPT program described earlier (Pradhan et al., 2009). While the sample in this study was arguably under-powered, the intervention nonetheless produced a 25% reduction in crashes for male participants during the following year (albeit no reduction for females). Third, the introduction of a hazard perception test into the UK driver licensing system is estimated to have reduced non-low-speed public road crashes by 11.3% for drivers during the year after they have taken the test (Wells et al., 2008). This equates to the prevention of nearly 10,000 crashes per year. While more research to verify these outcomes would be desirable, they nonetheless suggest that measuring and manipulating drivers' hazard perception skill is a viable strategy for improving road safety.

Acknowledgment

The preparation of this article did not receive specific support.

References

- Øbele, L., Hausteijn, S., Martinussen, L.M., Møller, M., 2019. Improving drivers' hazard perception in pedestrian-related situations based on a short simulator-based intervention. *Transport. Res. F: Traffic Psychol. Behav.* 62, 1–10.
- Øbele, L., Hausteijn, S., Møller, M., Martinussen, L.M., 2018. Consistency between subjectively and objectively measured hazard perception skills among young male drivers. *Accid. Anal. Prev.* 118, 214–220.
- Agrawal, R., Knodler, M., Fisher, D.L., Samuel, S., 2018. Virtual reality headset training: can it be used to improve young drivers' latent hazard anticipation and mitigation skills. *Transport. Res. Rec.* 2672, 20–30.
- Beanland, V., Goode, N., Salmon, P.M., Lenne, M.G., 2013. Is there a case for driver training? A review of the efficacy of pre- and post-licence driver training. *Saf. Sci.* 51, 127–137.
- Borowsky, A., Horrey, W.J., Liang, Y., Garabet, A., Simmons, L., Fisher, D.L., 2015. The effects of momentary visual disruption on hazard anticipation and awareness in driving. *Traffic Injury Prev.* 16, 133–139.
- Borowsky, A., Horrey, W.J., Liang, Y., Simmons, L., Garabet, A., Fisher, D.L., 2014. Memory for a hazard is interrupted by performance of a secondary in-vehicle task. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58. SAGE Publications, Los Angeles, CA, pp. 12219–12223.
- Borowsky, A., Oron-Gilad, T., Parmet, Y., 2009. Age and skill differences in classifying hazardous traffic scenes. *Transport. Res. F: Traffic Psychol. Behav.* 12, 277–287.
- Borowsky, A., Shinar, D., Oron-Gilad, T., 2010. Age, skill, and hazard perception in driving. *Accid. Anal. Prev.* 42, 1240–1249.

- Boufous, S., Ivers, R., Senserrick, T., Stevenson, M., 2011. Attempts at the practical on-road driving test and the hazard perception test and the risk of traffic crashes in young drivers. *Traffic Injury Prev.* 12, 475–482.
- Cantwell, S., Isler, R., Starkey, N., 2013. The effects of road commentary training on novice drivers' visual search behaviour: a preliminary investigation. In: *Proceedings of the 2013 Australasian Road Safety Research, Policing & Education Conference*, 28th–30th August, Brisbane, Queensland.
- Chapman, P., Underwood, G., Roberts, K., 2002. Visual search patterns in trained and untrained drivers. *Transport. Res. F: Traffic Psychol. Behav.* 5, 157–167.
- Cheng, A.S.K., Ng, T.C.K., Lee, H.C., 2011. A comparison of the hazard perception ability of accident-involved and accident-free motorcycle riders. *Accid. Anal. Prev.* 43, 1464–1471.
- Congdon, P., 1999. VicRoads Hazard Perception Test, Can it Predict Accidents? Australian Council for Educational Research, Camberwell, Victoria, Australia.
- Crundall, D., 2016. Hazard prediction discriminates between novice and experienced drivers. *Accid. Anal. Prev.* 86, 47–58.
- Crundall, D., Andrews, B., Van Loon, E., Chapman, P., 2010. Commentary training improves responsiveness to hazards in a driving simulator. *Accid. Anal. Prev.* 42, 2117–2124.
- Crundall, D., Chapman, P., Trawley, S., Collins, L., Van Loon, E., Andrews, B., Underwood, G., 2012. Some hazards are more attractive than others: drivers of varying experience respond differently to different types of hazard. *Accid. Anal. Prev.* 45, 600–609.
- Crundall, D.E., Underwood, G., 1998. Effects of experience and processing demands on visual information acquisition in drivers. *Ergonomics* 41, 448–458.
- Darby, P., Murray, W., Raeside, R., 2009. Applying online fleet driver assessment to help identify, target and reduce occupational road safety risks. *Saf. Sci.* 47, 436–442.
- Deery, H.A., Love, A.W., 1996. The effect of a moderate dose of alcohol on the hazard perception profile of young drink-drivers. *Addiction* 91, 815–827.
- Ericsson, A., Poole, R., 2016. *Peak: Secrets from the New Science of Expertise*. Houghton Mifflin, Harcourt.
- Evans, L., 1991. *Traffic Safety and the Driver*. Van Nostrand Reinhold, New York.
- Fisher, D.L., Pollatsek, A.P., Pradhan, A., 2006. Can novice drivers be trained to scan for information that will reduce their likelihood of a crash? *Injury Prev.* 12, 25.
- Grayson, G.B., Maycock, G., Groeger, J.A., Hammond, S.M., Field, D.T., 2003. Risk, Hazard Perception and Perceived Control. Transport Research Laboratory.
- Grayson, G.B., Sexton, B., 2002. The Development of Hazard Perception Testing. Transport Research Laboratory.
- Gregersen, N.P., 1996. Young drivers' overestimation of their own skill – an experiment on the relation between training strategy and skill. *Accid. Anal. Prev.* 28, 243–250.
- Hamid, M., Samuel, S., Borowsky, A., Fisher, D.L., 2014. Evaluation of effect of total awake time on driving performance skills – hazard anticipation and hazard mitigation: a simulator study. *Transportation Research Board 93rd Annual Meeting*.
- Hill, A., Horswill, M.S., Whiting, J., Watson, M.O., 2019. Computer-based hazard perception test scores are associated with the frequency of heavy braking in everyday driving. *Accid. Anal. Prev.* 122, 207–214.
- Horswill, M.S., 2016a. Hazard perception in driving. *Curr. Direct. Psychol. Sci.* 25, 425–430.
- Horswill, M.S., 2016b. Hazard perception tests. In: Fisher, D.L., Caird, J.K., Horrey, W., Trick, L. (Eds.), *The Handbook of Teen and Novice Drivers*. CRC Press, Boca Raton, FL.
- Horswill, M.S., 2016c. Improving fitness to drive: the case for hazard perception training. *Austr. Psychologist* 51, 173–181.
- Horswill, M.S., Anstey, K.J., Hatherly, C.G., Wood, J., 2010a. The crash involvement of older drivers is associated with their hazard perception latencies. *J. Int. Neuropsychol. Soc.* 16, 939–944.
- Horswill, M.S., Falconer, E.K., Pachana, N.A., Wetton, M., Hill, A., 2015a. The longer-term effects of a brief hazard perception training intervention in older drivers. *Psychol. Aging* 30, 62–67.
- Horswill, M.S., Garth, M., Hill, A., Watson, M.O., 2017. The effect of performance feedback on drivers' hazard perception ability and self-ratings. *Accid. Anal. Prev.* 101, 135–142.
- Horswill, M.S., Hill, A., Jackson, T., 2020. Scores on a new hazard prediction test are associated with both driver experience and crash involvement. *Transport. Res. F: Traffic Psychol. Behav.* 71, 98–109.
- Horswill, M.S., Hill, A., Wetton, M., 2015b. Can a video-based hazard perception test used for driver licensing predict crash involvement? *Accid. Anal. Prev.* 82, 213–219.
- Horswill, M.S., Kemala, C.N., Wetton, M., Scialfa, C.T., Pachana, N.A., 2010b. Improving older drivers' hazard perception ability. *Psychol. Aging* 25, 464–469.
- Horswill, M.S., Marrington, S.A., McCullough, C.M., Wood, J., Pachana, N.A., McWilliam, J., Raikos, M.K., 2008. The hazard perception ability of older drivers. *J. Gerontol. B: Psychol. Sci. Social Sci.* 63, 212–218.
- Horswill, M.S., McKenna, F.P., 1999. The effect of interference on dynamic risk-taking judgments. *Br. J. Psychol.* 90, 189–199.
- Horswill, M.S., McKenna, F.P., 2004. Drivers' hazard perception ability: Situation awareness on the road. In: Banbury, S., Tremblay, S. (Eds.), *A Cognitive Approach to Situation Awareness: Theory and Application*. Aldershot, UK, Ashgate.
- Horswill, M.S., Pachana, N.A., Wood, J., Marrington, S.A., McWilliam, J., McCullough, C.M., 2009. A comparison of the hazard perception ability of matched groups of healthy drivers aged 35 to 55, 65 to 74, and 75 to 84 years. *J. Int. Neuropsychol. Soc.* 15, 799–802.
- Horswill, M.S., Sullivan, K., Lurie-Beck, J.K., Smith, S., 2013a. How realistic are older drivers' ratings of their driving ability? *Accid. Anal. Prev.* 50, 130–137.
- Horswill, M.S., Taylor, K., Newnam, S., Wetton, M., Hill, A., 2013b. Even highly experienced drivers benefit from a brief hazard perception training intervention. *Accid. Anal. Prev.* 52, 100–110.
- Horswill, M.S., Waylen, A.E., Tofield, M.I., 2004. Drivers' ratings of different components of their own driving skill: a greater illusion of superiority for skills that relate to accident involvement. *J. Appl. Social Psychol.* 34, 177–195.
- Isler, R.B., Starkey, N.J., Sheppard, P., 2011. Effects of higher-order driving skill training on young, inexperienced drivers' on-road driving performance. *Accid. Anal. Prev.* 43, 1818–1827.
- Isler, R.B., Starkey, N.J., Williamson, A.R., 2009. Video-based road commentary training improves hazard perception of young drivers in a dual task. *Accid. Anal. Prev.* 41, 445–452.
- Jackson, L., Chapman, P., Crundall, D., 2009. What happens next? Predicting other road users' behaviour as a function of driving experience and processing time. *Ergonomics* 52, 154–164.
- Kahana-Levy, N., Shavitzky-Golkin, S., Borowsky, A., Vakil, E., 2019. The effects of repetitive presentation of specific hazards on eye movements in hazard perception training, of experienced and young-inexperienced drivers. *Accid. Anal. Prev.* 122, 255–267.
- Lee, S.E., Klauer, S.G., Olsen, E.C.B., Simons-Morton, B.G., Dingus, T.A., Ramsey, D.J., Ouimet, M.C., 2008. Detection of road hazards by novice teen and experienced adult drivers. *Transport. Res. Rec.* 2078, 26–32.
- Lim, P.C., Sheppard, E., Crundall, D., 2014. A predictive hazard perception paradigm differentiates driving experience cross-culturally. *Transport. Res. F: Traffic Psychol. Behav.* 26, 210–217.
- Manley, H., Paisarnrisrisomuk, N., Hill, A., Horswill, M.S., 2020. The development and validation of a hazard perception test for Thai drivers. *Transport. Res. F: Traffic Psychol. Behav.* 71, 229–237.
- Mayhew, D.R., Simpson, H.M., Williams, A.F., Ferguson, S.A., 1998. Effectiveness and role of driver education and training in a graduated licensing system. *J. Public Health Policy* 51–67.
- McKenna, F.P., Crick, J.L., 1991. Hazard Perception in Drivers: A Methodology for Testing and Training. Transport and Road Research Laboratory.
- McKenna, F.P., Crick, J.L., 1997. Developments in Hazard Perception. Department of Transport.
- McKenna, F.P., Horswill, M.S., 1999. Hazard perception and its relevance for driver licensing. *J. Int. Assoc. Traffic Saf. Sci.* 23, 26–41.
- McKenna, F.P., Horswill, M.S., Alexander, J.L., 2006. Does anticipation training affect drivers' risk taking? *J. Exp. Psychol.: Appl.* 12, 1–10.
- Meir, A., Borowsky, A., Oron-Gilad, T., 2014. Formation and evaluation of act and Anticipate Hazard Perception Training (AAHT) intervention for young novice drivers. *Traffic Injury Prev.* 15, 172–180.
- Mills, K.L., Hall, R.D., McDonald, M., Rolls, G.W.P., 1998. The Effects of Hazard Perception Training on the Development of Novice Driver Skills. London, UK, Department for Transport.
- Pollatsek, A., Narayana, V., Pradhan, A., Fisher, D.L., 2006. Using eye movements to evaluate a PC-based risk awareness and perception training program on a driving simulator. *Hum. Factors* 48, 447.
- Poulsen, A.A., Horswill, M.S., Wetton, M.A., Hill, A., Lim, S.M., 2010. A brief office-based hazard perception intervention for drivers with ADHD symptoms. *Austr. N. Z. J. Psychiatry* 44, 528–534.

- Pradhan, A.K., Pollatsek, A., Knodler, M., Fisher, D.L., 2009. Can younger drivers be trained to scan for information that will reduce their risk in roadway traffic scenarios that are hard to identify as hazardous? *Ergonomics* 52, 657–673.
- Quimby, A.R., Maycock, G., Carter, I.D., Dixon, R., Wall, J.G., 1986. Perceptual Abilities of Accident Involved Drivers. Transport and Road Research Laboratory, Crowthorne, UK.
- Quimby, A.R., Watts, G.R., 1981. Human Factors and Driving Performance. Transport and Road Research Laboratory, Crowthorne, UK.
- Renge, K., Ishibashi, T., Oiri, M., Ota, H., Tsunenari, S., Mukai, M., 2005. Elderly drivers' hazard perception and driving performance. In: Underwood, G. (Ed.), *Traffic and Transport Psychology: Theory and Application*. Elsevier, London.
- Ross, R.W., Scialfa, C., Cordazzo, S., Bubric, K., 2013. Predicting older adults' on-road driving performance. In: 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design.
- Sagberg, F., Bjornskau, T., 2006. Hazard perception and driving experience among novice drivers. *Accid. Anal. Prev.* 38, 407–414.
- Savage, S.W., Potter, D.D., Tatler, B.W., 2013. Does preoccupation impair hazard perception? A simultaneous EEG and Eye Tracking study. *Transport. Res. F: Traffic Psychol. Behav.* 17, 52–62.
- Scialfa, C.T., Borkenhagen, D., Lyon, J., Deschênes, M., 2013. A comparison of static and dynamic hazard perception tests. *Accid. Anal. Prev.* 51, 268–273.
- Shimazaki, K., Ito, T., Fujii, A., Ishida, T., 2017. Improving drivers' eye fixation using accident scenes of the HazardTouch driver-training tool. *Transport. Res. F: Traffic Psychol. Behav.* 51, 81–87.
- Smith, S.S., Horswill, M.S., Chambers, B., Wetton, M., 2009. Hazard perception in novice and experienced drivers: the effects of sleepiness. *Accid. Anal. Prev.* 41, 729–733.
- Svenson, O., 1981. Are we all less risky and more skillful than our fellow drivers? *Acta Psychol.* 47, 143–148.
- Taylor, T., Pradhan, A.K., Divekar, G., Romoser, M., Muttart, J., Gomez, R., Pollatsek, A., Fisher, D.L., 2013. The view from the road: the contribution of on-road glance-monitoring technologies to understanding driver behavior. *Accid. Anal. Prev.* 58, 175–186.
- Thomas, F.D., Blomberg, R.D., Peck, R.C., Korbelak, K.T., 2016. Evaluation of the Safety Benefits of the Risk Awareness and Perception Training Program for Novice Teen Drivers. National Highway Traffic Safety Administration.
- Thompson, J.J., Blair, M.R., Henrey, A.J., 2014. Over the hill at 24: persistent age-related cognitive-motor decline in reaction times in an ecologically valid video game task begins in early adulthood. *PLoS One* 9, e94215.
- Toske, V., Šeibokaite, L., Endriulaitiene, A., Lehtonen, E., 2019. Hazard perception test development for Lithuanian drivers. *IATSS Res.* 43, 108–113.
- Underwood, G., Chapman, P., Brocklehurst, N., Underwood, J., Crundall, D., 2003. Visual attention while driving: sequences of eye fixations made by experienced and novice drivers. *Ergonomics* 46, 629–646.
- Ventsislavova, P., Crundall, D., 2018. The hazard prediction test: a comparison of free-response and multiple-choice formats. *Safety Sci.* 109, 246–255.
- Ventsislavova, P., Crundall, D., Baguley, T., Castro, C., Gugliotta, A., Garcia-Fernandez, P., Zhang, W., Ba, Y., Li, Q., 2019. A comparison of hazard perception and hazard prediction tests across China, Spain and the UK. *Accid. Anal. Prev.* 122, 268–286.
- Viakveld, W.P., 2014. A comparative study of two desktop hazard perception tasks suitable for mass testing in which scores are not based on response latencies. *Transport. Res. F: Traffic Psychol. Behav.* 22, 218–231.
- Wallis, T.S.A., Horswill, M.S., 2007. Using fuzzy signal detection theory to determine why experienced and trained drivers respond faster than novices in a hazard perception test. *Accid. Anal. Prev.* 39, 1177–1185.
- Washington, S., Cole, R.J., Herbel, S.B., 2011. European advanced driver training programs: reasons for optimism. *IATSS Res.* 34, 72–79.
- Wells, P., Tong, S., Sexton, B., Grayson, G., Jones, E., 2008. Cohort II: A Study of Learner and New Drivers. Department for Transport, London.
- West, R., Wilding, J., French, D., Kemp, R., Irving, A., 1993. Effect of low and moderate doses of alcohol on driving hazard perception latency and driving speed. *Addiction* 88, 527–532.
- Wetton, M.A., Hill, A., Horswill, M.S., 2011. The development and validation of a hazard perception test for use in driver licensing. *Accid. Anal. Prev.* 43, 1759–1770.
- Wetton, M.A., Hill, A., Horswill, M.S., 2013. Are what happens next exercises and self-generated commentaries useful additions to hazard perception training for novice drivers? *Accid. Anal. Prev.* 54, 57–66.
- Wetton, M.A., Horswill, M.S., Hatherly, C., Wood, J.M., Pachana, N.A., Anstey, K.J., 2010. The development and validation of two complementary measures of drivers' hazard perception ability. *Accid. Anal. Prev.* 42, 1232–1239.
- Wood, J.M., 2002. Age and visual impairment decrease driving performance as measured on a closed-road circuit. *Hum. Factors* 44, 482–494.
- Wood, J.M., Horswill, M.S., Lacherez, P.F., Anstey, K.J., 2013. Evaluation of screening tests for predicting older driver performance and safety assessed by an on-road test. *Accid. Anal. Prev.* 50, 1161–1168.
- Yamani, Y., Samuel, S., Fisher, D., 2014. Simulator evaluation of an integrated road safety training program. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Sage Publications, pp. 1904–1908.
- Young, A.H., Crundall, D., Chapman, P., 2017. Commentary driver training: effects of commentary exposure, practice and production on hazard perception and eye movements. *Accid. Anal. Prev.* 101, 1–10.

Further Reading

- Horswill, M.S., 2016. Hazard perception in driving. *Curr. Direct. Psychol. Sci.* 25, 425–430.
- Horswill, M.S., 2016. Hazard perception tests. In: Fisher, D.L., Caird, J.K., Horrey, W., Trick, L. (Eds.), *The Handbook of Teen and Novice Drivers*. CRC Press, Boca Raton, FL.
- Horswill, M.S., 2016. Improving fitness to drive: the case for hazard perception training. *Austr. Psychologist* 51, 173–181.
- Horswill, M.S., McKenna, F.P., 2004. Drivers' hazard perception ability: Situation awareness on the road. In: Banbury, S., Tremblay, S. (Eds.), *A Cognitive Approach to Situation Awareness: Theory and Application*. Ashgate, Aldershot, UK.

Driver Education and Training for New Drivers: Moving beyond Current ‘Wisdom’ to New Directions

Teresa Senserrick, Oscar Oviedo-Trespalacios, David Rodwell, Sherrie-Anne Kaye, Queensland University of Technology (QUT) Centre for Accident Research and Road Safety – Queensland (CARRS-Q), Kelvin Grove, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	158
Common Approaches to Novice Driver Education and Training	158
Foundations of Current Wisdom in Novice Driver Education and Training	159
Theoretical and Emerging Research into Novice Driver Education and Training Best Practices	159
Looking to the Future—Current and Emerging Driver Education and Training Needs	160
Concluding Comments	162
Acknowledgment	162
References	162
Further Reading	164

Introduction

Driver education and training are among the earliest and most universal behavioral countermeasures to road crashes. This particularly applies when first learning to drive a car-type vehicle and becoming an independently licensed driver, which is the focus of this article. Originally, such efforts centered on road rules and training of basic skills relating to vehicle handling and maneuvering in traffic in preparation for licensure. Over recent decades, however, there has been a considerable shift in expected outcomes, now also aiming for safer drivers post-licensure; with casualty crashes peaking early in this transition period ([Senserrick and Mitsopoulos-Rubens, 2013](#)).

In this context, current wisdom generally dictates that such efforts are ineffective. That is, the ways in which societies generally educate and train new drivers for licensure do not sufficiently protect them. Certainly, the latter is true. However, if we look closer at the literature, including more recent advances and emerging trends, it could be argued that true wisdom is still lacking. Recent high-quality evaluations of high-quality programs are scant. Approaches that we know are ineffective are popular and prevail, while more nuanced initiatives often require intensive efforts and funding for wider implementation and more compelling evaluation, and do not advance. In this article, we acknowledge this body of literature, but seek to highlight promising research and needs for new directions as the vehicles we drive and the ways in which we travel are progressively being transformed.

Before we elaborate, we acknowledge that much of the research specific to this review has been undertaken in countries in which this transition to independent driving predominantly occurs in the first years following the minimum licensing age. This includes Australia, where the authors are based, and as such this context contributes to the perspectives discussed in this article. Therefore, our focus commonly relates to young drivers, with this term implying a broad age range, in keeping with the United Nations definition of youth as ages 15–24 years. Roles for parents are also sometimes highlighted as relevant to this age group. We use the term driver education holistically, to refer to efforts to increase any driving-related knowledge, and driver training more narrowly, specific to driving skill attainment.

Common Approaches to Novice Driver Education and Training

Best practices in novice driver safety are structured around graduated licensing systems ([Mayhew et al., 2014](#)), that is, systems that progress the novice through various license stages of increasing risk as more driving experience or maturity are gained. Even outside of formal graduated systems, most jurisdictions set requirements for licensure that require or promote a learner period, during which road rules and other road safety knowledge must be gained and supervised-only driving is permitted in order to pass mandatory knowledge and practical driving tests. While education might be limited to driver handbooks, other official texts or individual web-based programs, a range of optional or mandatory government and private group-based initiatives have also evolved.

In some jurisdictions, learner driving can only be undertaken at specialist driving schools or under the direction of an accredited professional driving instructor (such as in China, Denmark, and Spain). In other jurisdictions, parents and other experienced drivers (sometimes with specific eligibility criteria and registration requirements) can also supervise a learner (such as in the United Kingdom and United States) or else learners choose between professional or lay options (such as in Austria and France). In the Australian context, both professional and lay supervision are permitted at learner drivers' discretion, and a mix of both is promoted.

One-on-one professional driving lessons taken on-road are considered to lay the groundwork or otherwise strengthen adherence to legal requirements (such as up-to-date road rule knowledge) and preparation for the practical driving test for independent licensure (such as, safe and preferred vehicle handling and maneuvering requirements); potentially also correcting any “bad habits” acquired from lay supervisors. Alternatively, lay supervision, commonly undertaken with parents, or other family and friends, is considered to broaden the experience learners gain. This is due to the greater variability in driving achieved, in contrast to the

typically one-hour professional lessons that commence at the learner's residence, therefore covering similar locations and often similar times of day and traffic conditions (see, e.g., [Groeger and Brady, 2004](#)).

While professional lessons clearly fall within the definition of training, lay supervision is typically characterized as "practice driving," which provides opportunity to develop a wide range of cognitive schema or mental representations of driving to facilitate rapid safer responses in critical situations. Falling somewhat between these are learner driver mentoring programs, in which volunteers provide supervision for otherwise disadvantaged youth who do not have access to lay supervisors, vehicles or funds for professional lessons in order to achieve the high level of learner driving (100–120 h) required by some Australian states (see, e.g., guidelines informed by previous research, [Soole et al., 2014](#)).

Group-based education initiatives for young drivers commonly are coordinated through schools and might also take place at school within regular school hours or at external venues at other times. Such programs address common young driver crash risks and might incorporate presentations from crash survivors, police, or emergency room surgeons, which aim to foster a deeper understanding of the risks inherent in driving or the potential consequences of risky driving to promote safer attitudes and behaviors. Peer-led or peer-to-peer education components also appears to be increasing, with some examples of improvements ([Buckley and Watson, 2014](#); [Cutello et al., 2020](#)).

Some programs also have been put in place to help prepare parents of young people for the tasks involved in supervising their child behind the wheel, although evaluations regarding the efficacy of these programs has been mixed ([Curry et al., 2015](#)). The more promising programs work to complement graduated driver licensing systems, such as a web-based program to increase the diversity of practice that is provided in parental supervised driving ([Mirman et al., 2014](#); [Toledo et al., 2012](#)), and a parent-teen agreement initiative (Checkpoints) that aims to increase parental limit setting on their children's driving when they first start to drive unsupervised ([Zakrajsek et al., 2013](#)).

Foundations of Current Wisdom in Novice Driver Education and Training

The United States pioneered an early Drivers' Ed model for learner drivers that is evident in these approaches and still integrated in some graduated licensing systems today. The approach required young drivers to undertake a combination of group-based classroom education focused on road safety knowledge and attitudes (minimum 30 h) plus in-vehicle training, either on-road, on a closed track or simulated (minimum 6 h). High-quality research evaluations of "state-of-the-art" applications in the United States and Australia in the 1970s and 1980s not only failed to identify a protective role, but there was some evidence participation resulted in earlier licensure, thereby increasing exposure to crashes and inflating risk (reviewed in [Senserrick, 2007](#)). Some programs even allowed time discounts on the learner tenure as an incentive, which continues to be demonstrated as counterproductive ([Begg and Brookland, 2015](#)). Further efforts to tailor such courses to the specific skill deficits of novices over the years, including in several European countries, have similarly been ineffective or counterproductive, particularly a practice broadly referred to as "skid training" to counter the identified over-representation of novices in a loss-of-control crashes (reviewed in [Senserrick and Mitsopoulos-Rubens, 2013](#)).

Such failures are variously attributed to the brevity or poor quality of the initiative or its evaluation, or a decay in skills from the time of learning to the time of need, especially school-based programs that target entire class levels and therefore often include many young people who do not yet have a permit or license to drive and might not do so for several months or even years later. A further critical factor in the counterproductive findings is the potential for such initiatives to increase novices' perceptions of their driving capabilities as much greater than actual—that is, a form of miscalibration rather than "overconfidence" as sometimes incorrectly labeled. This can lead the trained driver to take on driving situations that untrained drivers might avoid, such as driving in inclement weather or not sufficiently reducing their travel speed in such circumstances. Collectively, such work has led to an overarching conclusion that education and training are generally ineffective, basically because "it is not what you know, but what you do" that matters.

Theoretical and Emerging Research into Novice Driver Education and Training Best Practices

To some extent, this historical context continues to underlie much of the current "wisdom." However, there have been several more nuanced initiatives and exceptions to these global conclusions. In particular, a long literature focused specifically on "higher-order" hazard anticipation and perception skills has demonstrated not only that training can increase such skills, but that these skills transfer to real-world driving, can discriminate between novice, experienced and expert drivers, and increasingly also between crash- and non-crash-involved drivers ([Horswill et al., 2020](#); [Manley et al., 2020](#); [McDonald et al., 2015](#); [Thomas et al., 2016](#)). A recent study also suggests that hazard perception training can promote reduced engagement in secondary tasks ([Krishnan et al., 2019](#)), another prevalent crash risk factor, particularly among young drivers. Most jurisdictions in Australia now include hazard perception tests in their graduated licensing systems, which are currently being updated. This follows indications of the success of such tests in screening out drivers not ready to progress ([Boufous et al., 2011](#); [Horswill et al., 2015](#); [Wells et al., 2008](#)).

Attention to "high-order" skills, distinguished from basic vehicle-handling capacity, has increasingly expanded since emergence of a Finnish best-practice framework known as the Goals for Driver Education or GDE ([Hatakka et al., 2002](#)). The GDE presents as a hierarchical matrix with basic skills of "vehicle maneuvering" at the lowest level, successively increasing to maneuvering in traffic or

"mastery of traffic situations," to "goals and contexts for driving" that include decision making relative to specific driving situations (such as route planning, presence of peers and state of the driver), to "goals for life and skills for living" at the highest level, which moves beyond the specific driving context to broader cultural and social contexts (such as living in a location with a "car culture" or "cycling culture," and social and wider societal norms around drug use). Across the matrix for each of these levels are specific "knowledge and skills" and "risk-increasing factors," but also "self-evaluation and awareness skills." Best practice is considered to touch on all cells of the matrix, but not in a unidirectional manner, with lower levels considered to impact on higher levels and vice-versa. Reflecting on the historical context, traditional driver education has focused on the first "rows and columns" of the matrix, moved somewhat into the higher level of the goals and contexts for specific driving contexts, but rarely reached the wider societal context at the highest level, or included self-reflection components in relation to these or even lower levels.

More recent initiatives aligning with GDE self-evaluation and awareness skills include programs focused on "insight training," with the intent of giving novice drivers the ability to calibrate their perceptions of their driving skills with the reality of their driving skills and thereby foster safer attitudes towards driving risks (Horrey et al., 2015; Isler et al., 2011; Washington et al., 2011). Such programs teach drivers to self-assess and anticipate risk, focusing on higher-order skills as well as basic vehicle handling skills, such as demonstrating the physics of braking distance at different speeds and contrasting the long distances required to stop to the close following distances that drivers typically leave in traffic. At the highest broader societal level are "resilience-focused" programs, that aim more widely to empower youth to make informed decisions and adopt safer strategies not only to reduce their own exposure to risk, but also that of their friends and others, which may or may not include content specific to driving risks (Griffin et al., 2004; Senserrick et al., 2009). While there is reason for optimism regarding such approaches, more research is required to determine potential differential risks and benefits, particularly for young males (White et al., 2011).

Increasing attention to providing feedback to novices on their driving is a form of education that could be argued also to promote better self-reflection and self-assessment. This includes in-person feedback or feedback from the vehicle or on-line or via text messages from various programs and tests, as examples. Several studies have demonstrated that novice driving can be improved by providing feedback on their vehicle handling skills (e.g., Carney et al., 2010; Simons-Morton et al., 2013), speed management (e.g., Molloy et al., 2019), hazard perception (e.g., Horswill et al., 2017) and also headway in one small demonstration study (Arnold and VanHouten, 2020); but not impulse control in another body of work (Hatfield et al., 2018). There have been mixed results in also providing feedback to parents of young drivers about their child's driving, although one recent study suggests a key to the effectiveness of this approach is to provide parents with effective communication strategies (Peek-Asa et al., 2019).

Most recently, a GDE coding taxonomy has been developed to explore a range of higher-order spoken and non-verbal communications by professional instructors in learner driver lessons, using naturalistic driving methods (Watson-Brown et al., 2018). A pilot and larger main study in Australia found 15%–35% of such instruction was higher-order, with the main study finding a further 12% of chances for higher-order instruction were either not perceived or addressed with functional feedback only (Scott-Parker et al., 2014; Watson-Brown et al., under review). While seemingly low, somewhat comparable research in the United States on parental supervision of teen practice driving found between 6% to 15% could be identified as higher-order across three studies (Ehsani et al., 2015, 2017; Goodwin et al., 2014). Importantly, a survey of a corresponding cohort of young drivers in Australia found that the more professional higher-order instruction perceived the lower the engagement in both inattention and intentional risk-taking driving behaviors, which were mediated by a self-regulated safety orientation towards driving (Watson-Brown et al., 2019).

This is emerging research and needs further applications and evaluation, nonetheless, further exploration of the professional instruction in the Australian research found that the proportion of higher-order communication did not vary as might be expected according to the level of driving experience of the learner or instructor (Watson-Brown et al., under review). For example, there was no corresponding increase with increased learner driving hours, proportional use varied widely by instructor, but also on average was highest in lessons with learners who had less than 20 h of driving experience. This suggests, much like the established value in providing even experienced drivers with hazard perception training (e.g., Horswill et al., 2013), there might also be value in upskilling professional instructors, and possibly also parents, in how to extend their communications to learners drivers to be high-order, in order to deepen the learning and potentially better protect the driver once independently licensed.

Looking to the Future—Current and Emerging Driver Education and Training Needs

The nature of driving has been rapidly changing in recent years. Vehicles are being transformed as a consequence of the technological developments leading to automated driving. Although fully automated vehicles including driverless cars will not be commercially available in the short- or mid-term, the driving experience will be enhanced and supported by a wide range of technologies during the transition period. Advanced Driver Support Systems (ADAS), such as blind spot monitoring, lane departure warning, collision warning, and in-vehicle information systems, are being incorporated in increasing numbers in vehicles (Viereckl et al., 2016).

There is great expectation about the positive impact of these new systems on safe driving and road trauma (Cicchino, 2019; Spicer et al., 2018). However, ADAS bring new challenges for driver education and training. In particular, there are concerns about misperceptions or lack of awareness of ADAS functionalities among drivers, particularly those who learn to drive in vehicles with these technologies. This could result in negative behavioral adaptations among drivers such as taking advantage of or misusing the systems. For example, there are reports that when driving with adaptive cruise control drivers present delayed reactions times (Larsson et al., 2014; Shen and Neyens, 2017) and decreases in minimum time headway/time-to-collision (Bianchi-Piccinini

et al., 2015; Hoedemaeker and Brookhuis, 1998), as well as being more likely to engage in distractions (Dunn et al., 2019). In some extreme cases it has been reported that other road users reduce their protective behaviors in the presence of a vehicle known to have autonomous emergency braking (Kinosada et al., 2020). There is a need to understand the future education and training needs that ADAS and other technologies will bring for the transport system.

Education and training could have an important role in the safe up-take of in-vehicle technology. However, evidence suggests that they currently are insufficient and lack structure. A common finding from many studies around the world is that large groups of drivers do not even know if they possess a system. For example, in the Netherlands, it was found that the vast majority of drivers of vehicles with adaptive cruise control, lane departure warning, emergency brake and distance alert systems were largely unaware of owning ADAS (from 65% up to 83%; Harms et al., 2020). In another Dutch study, one-quarter of new car buyers had not received any information about ADAS fitted in the car (Boelhouwer et al., 2020). Another issue has been that even if drivers know that they possess a system, they do not understand the functioning or limitations of the ADAS. For example, a national cross-sectional study of drivers in Spain found that not knowing how to use the systems was one of the main reasons for not using ADAS, including fatigue detection systems and adaptive cruise control (Lijarcio et al., 2019). More concerning it seems that many drivers can be unaware of the limitations of the ADAS, such as the limitations of cameras, radars and other sensors that these technologies rely on, as demonstrated in United States research (McDonald et al., 2018). There is an important gap in upskilling drivers of all ages regarding the safe use of ADAS.

A major problem is that education and training about ADAS has been largely the responsibility of the driver. Generally, drivers are required to read the user manuals and complement this with other sources of information, such as via the internet. However, reports show that the majority of drivers use trial-and-error as opposed to user manuals or receiving instructions from the dealer (Harms et al., 2020). Drivers also report that user manuals can be difficult to understand, too long and contain insufficient information (Viktorová and Šucha, 2019). Inadequate training at the time of purchase is one of the most significant challenges for ADAS education and training. There is evidence that car sellers often do not have enough time to inform customers about ADAS, do not have the right training or education themselves, or show inconsistent quality (Abraham et al., 2018; Boelhouwer et al., 2020; McDonald et al., 2018). Other sources of education and training such as discussion with family or friends may provide insufficient information about ADAS risks (Oviedo-Trespalcacios et al., 2019).

This situation is extremely problematic because drivers may develop erroneous mental models of the ADAS that could result in misuse of the systems or encounter problems with ADAS during safety-critical situations. An example of this has been reported in the United States, with drivers of vehicles with lateral and longitudinal ADAS found to have greater odds of engaging in distracted driving behaviors (Dunn et al., 2019). Further, qualitative research has found that some drivers only realize the limitations of an ADAS while driving, such as only working within certain speed limits or issues operating in certain weather conditions (Viktorová and Šucha, 2019). Research from the United States reported that the current state of driver education on in-vehicle technologies is often misaligned with driver learning preference. Participants preferred receiving training at the point of sale or via a technology expert along with some other media, such as websites, vehicle manual, or in-vehicle computerized assistance (Abraham et al., 2018); yet as noted above, this does not appear to eventuate. Further developments are needed to guarantee effective educational and training activities on ADAS use.

Further, ADAS with visual interfaces, such as in-vehicle information systems (IVIS), are known to present distraction risks for many drivers (Oviedo-Trespalcacios et al., 2019). Given that IVIS are key elements of the human-machine interface, it is likely that their prevalence will increase (Statista, 2017). Traditional approaches to education of drivers have targeted prevention of distractions, whereas most recent vehicles require drivers to share and interact with an increased number of systems, such as IVIS and head-up displays. Therefore, divergent with traditional education and training approaches that focus on avoiding distractions, there is a need to upskill drivers to engage safely in distractions with these ADAS. An example of this is the U.S. FOrward Concentration and Attention Learning (FOCAL) educational program (Unverricht et al., 2019). FOCAL educational training develops drivers' capacity to self-regulate off road glances. Experiments after training confirm that the extent of distraction can be reduced. Drivers who received FOCAL training engaged in fewer in-vehicle glances longer than two seconds (generally accepted as the minimum safety threshold) compared to drivers who received traditional education on distraction-related risks and road rules. Future education and training initiatives will need to consider the need to upskill drivers in the correct interactions with ADAS that include safer and less safe distraction.

Another important concern in the field is that extensive experience with ADAS might result in skill atrophy when equivalent systems are not available, potentially increasing risk of missing detection of a pedestrian for example (Miller and Boyle, 2019). This particularly applies to young novice drivers, given common changing of vehicle use as a learner between a professional instructor's vehicle, their parents' vehicles and potential transition to their first own vehicle, which is more likely to be older and have fewer safety features (Eichelberger et al., 2015; Keall and Newstead, 2010). But also this applies to experienced drivers who engage in the rental/for-hire vehicle market, such as business travelers and tourists, who might switch to and from various vehicles as they travel to different locations. Such use might be short-term and likewise also demand for brief education and training initiatives. As yet however, evidence-based examples are lacking. There are still open questions about the role of education and training programs designed to allow drivers to maintain relevant skills transferring between vehicles with more or fewer ADAS and, continuing to the future, also higher and lower levels of automation.

On-road trials of automated vehicles (AVs) are currently underway in many countries. The Society of Automotive Engineers (SAE International, 2018) defines six levels of automation, ranging from no automation (Level 0) to full automation (Level 5). It has been predicted that AVs will reduce human error associated with crashes. However, more recent research has argued that AVs may increase

passive fatigue, that is, fatigue which is associated with low workload, such as monitoring vehicle systems rather than operating the vehicle (Matthews et al., 2019). Increases in such fatigue has potential to increase crash risk when a driver is requested to take back control of the vehicle. These new advances in vehicle technologies will change the role of a driver to one of observing the vehicle functions and intervening when requested (or when required), rather than actively operating the vehicle at all times. Therefore, driver education and training are likely to be vital in increasing users' understanding of this new human-automation interaction as well as adaptive changes needed when transitioning from higher to lower level vehicles and vice-versa.

The need for increased education and training can be demonstrated in media reports of crashes that have occurred in the United States involving self-driving vehicles. In 2018, a vehicle travelling in automated mode in Arizona hit a pedestrian, who later died in hospital. At the time of the crash the human operator of the vehicle was watching their phone and not monitoring the road environment. Similarly, a Tesla vehicle operating in self-driving mode in Florida in 2016 failed to distinguish a truck, resulting in a fatal collision. In both cases, the driver (or operator) should have been monitoring the road environment and have interacted by taking back control of the vehicle but failed to do so. These cases further highlight the need to educate drivers on vehicle capabilities, including such limitations associated with AV technologies.

To this end, some recent research in the AV literature has examined the effectiveness of different types of training programs, including, Virtual Reality systems, Head-Mounted display (i.e., a device which is worn on the head), visual (e.g., video) and written education resources, and experiencing automated functions in a driving simulator (e.g., Forster et al., 2020; Payre et al. 2017; Sportillo et al., 2018). While such research is in its infancy of assessing which type(s) of training programs are most effective, what has been found is that individuals who complete some type of training are more equipped to take back control of the vehicle compared to those who complete no (or limited) training. For example, in France, it was found training improved response times in a driving simulator study (Payre et al., 2017). Further, participants who received elaborated training (i.e., read text explaining automated driving, answered questions with the experimenter, and watched a short video on how to engage and take back control of the vehicle) reported greater trust in automated driving after the intervention, compared to participants who did not receive this prior training before the simulator drive. Once AVs are more commonly available and appropriate legislation has been introduced, there will be more opportunities for people to increase their skills and gain experience in using these AV functions on closed tracks and on open roads.

Concluding Comments

The ways in which we travel are rapidly evolving. Some research suggests that young people are delaying driver licensure for a variety of reasons (Thigpen and Handy, 2018). Travel modes have expanded to include more shared vehicles, as well as an array of electronic bicycles and "rideables" (such as e-scooters, e-unicycles), and demand for more livable cities is increasing. So too should be the demand for more innovative, forward-thinking driver education and training. Relying on traditional approaches to "tick the box" and historic "wisdom" that such approaches cannot be effective provides an excuse that prevents greater efforts to improve such practices. We argue that such complacency is not justifiable, particularly with the potential for counterproductive outcomes by relying on the status quo. There are important types of knowledge and skills that need to be taught, and examples that this can be done without miscalibration risk, and with true potential to play a protective role for novice drivers; one of the most at-risk driver groups. Therefore, it is imperative among both the academic community and industry that efforts continue, not to dismiss education and training, but to persevere to make optimal, agile improvements and demonstrate effectiveness through high-quality evaluation.

Acknowledgment

Dr Oscar Oviedo-Trespalacios is the recipient of an Australian Research Council Discovery Early Career Researcher Award [DE200101079] funded by the Australian Government.

References

- Abraham, H., Reimer, B., Mehler, B., 2018. Learning to use in-vehicle technologies: consumer preferences and effects on understanding. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 62, 1589–1593.
- Arnold, M.L., VanHouten, R., 2020. Increasing following headway in a driving simulator and transfer to real world driving. *J. Org. Behav. Manage.* 1–19.
- Begg, D., Brookland, R., 2015. Participation in driver education/training courses during graduated driver licensing, and the effect of a time-discount on subsequent traffic offenses: findings from the New Zealand Drivers Study. *J. Saf. Res.* 55, 13–20.
- Bianchi-Piccinini, G.F., Rodrigues, C.M., Leitão, M., Simões, A., 2015. Reaction to a critical situation during driving with adaptive cruise control for users and non-users of the system. *Saf. Sci.* 72, 116–126.
- Boelhrouwer, A., van den Beukel, A.P., van der Voort, M.C., Hottentot, C., de Wit, R.Q., Martens, M.H., 2020. How are car buyers and car sellers currently informed about ADAS? An investigation among drivers and car sellers in the Netherlands. *Transportation Research Interdisciplinary Perspectives*, p. 100103.
- Boufous, S., Ivers, R., Senserrick, T., Stevenson, M., 2011. Attempts at the practical on-road driving test and the hazard perception test and the risk of traffic crashes in young drivers. *Traffic Injury Prev.* 12, 475–482.
- Buckley, L., Watson, B., 2014. Best practice in peer-led curriculum content: informing an interactive program to improve passenger safety among high school seniors. In: *Proceedings of the 2014 Australasian Road Safety Research, Policing and Education Conference*. Retrieved from: <https://eprints.qut.edu.au/79320/>.

- Carney, C., McGehee, D.V., Lee, J.D., Reyes, M.L., Raby, M., 2010. Using an event-triggered video intervention system to expand the supervised learning of newly licensed adolescent drivers. *Am. Public Health Assoc.* 100, 1101–1106.
- Cicchino, J.B., 2019. Real-world effects of rear automatic braking and other backing assistance systems. *J. Saf. Res.* 68, 41–47.
- Curry, A.E., Peek-Asa, C., Hamann, C.J., Mirman, J.H., 2015. Effectiveness of parent-focused interventions to increase teen driver safety: a critical review. *J. Adolesc. Health* 57, S6–S14.
- Cutello, C.A., Hellier, E., Stander, J., Hanoch, Y., 2020. Evaluating the effectiveness of a young driver-education intervention: Learn2Live. *Transport. Res. F: Traffic Psychol. Behav.* 69, 375–384.
- Dunn, N., Dingus, T., Soccolich, S., 2019. Understanding the Impact of Technology: Do Advanced Driver Assistance and Semi-Automated Vehicle Systems Lead to Improper Driving Behavior? Retrieved from: https://aaafoundation.org/wp-content/uploads/2019/12/19-0460_AAAFTS_VTTI-ADAS-Driver-Behavior-Report_Final-Report.pdf.
- Ehsani, J.P., Klauer, S.G., Zhu, C., Gershon, P., Dingus, T.A., Simons-Morton, B.G., 2017. Naturalistic assessment of the learner license period. *Accid. Anal. Prev.* 106, 275–284.
- Eichelberger, A.H., Teoh, E.R., McCart, A.T., 2015. Vehicle choices for teenage drivers: a national survey of US parents. *J. Saf. Res.* 55, 1–5.
- Eshani, J., Simmons-Morton, B.G., Klauer, S.G., 2015. Developing and testing operational definitions for functional and higher order driving instruction. In: Paper presented at the 8th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Salt Lake City, Utah.
- Forster, Y., Hergeth, S., Naujoks, F., Krems, J.F., Keinath, A., 2020. What and how to tell beforehand: the effect of user education on understanding, interaction and satisfaction with driving automation. *Transport. Res. F: Traffic Psychol. Behav.* 68, 316–335.
- Goodwin, A.H., Foss, R.D., Margolis, L.H., Harrell, S., 2014. Parent comments and instruction during the first four months of supervised driving: an opportunity missed? *Accid. Anal. Prev.* 69, 15–22.
- Griffin, K.W., Botvin, G.J., Nichols, T.R., 2004. Long-term follow-up effects of a school-based drug abuse prevention program on adolescent risky driving. *Prev. Sci.* 5, 207–212.
- Groeger, J.A., Brady, S.J., 2004. Differential Effects of Formal and Informal Driver Training. Department for Transportation, Great Britain.
- Harms, I.M., Bingen, L., Steffens, J., 2020. Addressing the awareness gap: a combined survey and vehicle registration analysis to assess car owners' usage of ADAS in fleets. *Transport. Res. A: Policy Pract.* 134, 65–77.
- Hatakka, M., Keskinen, E., Gregersen, N.P., Glad, A., Hernetkoski, K., 2002. From control of the vehicle to personal self-control; broadening the perspectives to driver education. *Transport. Res. F: Traffic Psychol. Behav.* 5, 201–215.
- Hatfield, J., Williamson, A., Kehoe, E.J., Lemon, J., Arguel, A., Prabhakaran, P., Job, R.F.S., 2018. The effects of training impulse control on simulated driving. *Accid. Anal. Prev.* 119, 1–15.
- Hoedemaeker, M., Brookhuis, K.A., 1998. Behavioural adaptation to driving with an adaptive cruise control (ACC). *Transport. Res. F: Traffic Psychol. Behav.* 1, 95–106.
- Horrey, W.J., Lesch, M.F., Mitsopoulos-Rubens, E., Lee, J.D., 2015. Calibration of skill and judgment in driving: development of a conceptual framework and the implications for road safety. *Accid. Anal. Prev.* 76, 25–33.
- Horswill, M.S., Garth, M., Hill, A., Watson, M.O., 2017. The effect of performance feedback on drivers' hazard perception ability and self-ratings. *Accid. Anal. Prev.* 101, 135–142.
- Horswill, M.S., Hill, A., Wetton, M., 2015. Can a video-based hazard perception test used for driver licensing predict crash involvement? *Accid. Anal. Prev.* 82, 213–219.
- Horswill, M.S., Hill, A., Jackson, T., 2020. Scores on a new hazard prediction test are associated with both driver experience and crash involvement. *Transport. Res. F: Traffic Psychol. Behav.* 71, 98–109.
- Horswill, M.S., Taylor, K., Newnam, S., Wetton, M., Hill, A., 2013. Even highly experienced drivers benefit from a brief hazard perception training intervention. *Accid. Anal. Prev.* 52, 100–110.
- Isler, R.B., Starkey, N.J., Sheppard, P., 2011. Effects of higher-order driving skill training on young, inexperienced drivers' on-road driving performance. *Accid. Anal. Prev.* 43, 1818–1827.
- Larsson, A.F., Kircher, K., Hultgren, J.A., 2014. Learning from experience: familiarity with ACC and responding to a cut-in situation in automated driving. *Transport. Res. F: Traffic Psychol. Behav.* 27, 229–237.
- Lijarcio, I., Useche, S.A., Llamazares, J., Montoro, L., 2019. Availability, demand perceived constraints and misuse of ADAS technologies in Spain: findings from a national study. *IEEE Access* 7, 129862–129873.
- Keall, M.D., Newstead, S., 2010. Characteristics of vehicles driven by different driver demographics - how can safety vehicle choices be encouraged? Monash University Accident Research Centre Report #301. Monash University, Clayton, VIC. Retrieved from: <https://www.monash.edu/muarc/archive/our-publications/reports/muarc301>.
- Kinosada, Y., Kobayashi, T., Shinohara, K., 2020. Trusting other vehicles' automatic emergency braking decreases self-protective driving. *Hum. Factors* 1–16.
- Krishnan, A., Samuel, S., Yamani, Y., Romoser, M.R.E., Fisher, D.L., 2019. Effectiveness of a strategic hazard anticipation training intervention in high risk scenarios. *Transport. Res. F: Traffic Psychol. Behav.* 67, 43–56.
- Manley, H., Paisarnrisomsuk, N., Hill, A., Horswill, M.S., 2020. The development and validation of a hazard perception test for Thai drivers. *Transport. Res. F: Traffic Psychol. Behav.* 71, 229–237.
- Matthews, G., Neubauer, C., Saxby, D.J., Wohleber, R.W., Lin, J., 2019. Dangerous intersections? A review of studies of fatigue and distraction in the automated vehicle. *Accid. Anal. Prev.* 126, 85–94.
- Mayhew, D., Williams, A.F., Pashley, C., 2014. A new GDL framework: evidence base to integrate novice driver strategies. Traffic Injury Research Foundation, Ottawa, ON. Retrieved from: https://tirf.ca/wp-content/uploads/2017/01/ANewGDLFramework_EvidenceBaseToIntegrateNoviceDriverStrategies_6.pdf.
- McDonald, A., Carney, C., McGehee, D.V., 2018. Vehicle owners' experiences with and reactions to advanced driver assistance systems. American Automobile Association Foundation for Traffic Safety, Washington, DC. Retrieved from: https://aaafoundation.org/wp-content/uploads/2018/09/VehicleOwnersExperiencesWithADAS_TechnicalReport.pdf.
- McDonald, C.C., Goodwin, A.H., Pradhan, A.K., Romoser, M.R.E., Williams, A.F.A., 2015. Review of hazard anticipation training programs for young drivers. *J. Adolesc. Health* 57, S15eS23.
- Miller, E.E., Boyle, L.N., 2019. Behavioral adaptations to lane keeping systems: effects of exposure and withdrawal. *Hum. Factors* 61, 152–164.
- Mirman, J.H., Albert, W.D., Curry, A.E., Winston, F.K., Fisher Thiel, M.C., Durbin, D.R., 2014. TeenDrivingPlan effectiveness: the effect of quantity and diversity of supervised practice on teens' driving performance. *J. Adolesc. Health* 55, 620–626.
- Molloy, O., Molesworth, B.R.C., Williamson, A., 2019. Which cognitive training intervention can improve young drivers' speed management on the road? *Transport. Res. F: Traffic Psychol. Behav.* 60, 68–80.
- Oviedo-Trespalacios, O., Nandavar, S., Haworth, N., 2019. How do perceptions of risk and other psychological factors influence the use of in-vehicle information systems (IVIS)? *Transport. Res. F: Traffic Psychol. Behav.* 67, 113–122.
- Payre, W., Cestas, J., Dang, N.-T., Vienne, F., Delhomme, P., 2017. Impact of training and in-vehicle task performance on manual control recovery in automated car. *Transport. Res. F: Traffic Psychol. Behav.* 46, 216–227.
- Peek-Asa, C., Reyes, M.L., Hamann, C.J., Butcher, B.D., Cavanaugh, J.E., 2019. A randomized trial to test the impact of parent communication on improving in-vehicle feedback systems. *Accid. Anal. Prev.* 131, 63–69.
- SAE International, 2018. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles: J3016_201806. Retrieved from: https://www.sae.org/standards/content/j3016_201806/.
- Scott-Parker, B., Senserrick, T., Simons-Morton, B., Jones, C., 2014. Higher-order instruction by professional driving instructors: a naturalistic pilot study. Retrieved from: <https://acrs.org.au/article/higher-order-instruction-by-professional-driving-instructors-a-naturalistic-pilot-study/>.
- Senserrick, T.M., 2007. Recent developments in young driver education and training in Australia. *J. Saf. Res.* 38, 237–244.
- Senserrick, T., Ivers, R., Boufous, S., Chen, H.-Y., Stevenson, M., Norton, R., 2009. Young driver education programs that build resilience have potential to reduce road crashes. *Pediatrics* 124, 1287–1292.
- Senserrick, T., Mitsopoulos-Rubens, E., 2013. Behavioural adaptation and novice drivers. In: Rudin-Brown, M., Jamson, S. (Eds.), *Behavioural Adaptation and Road Safety: Theory, Evidence and Action*. Taylor & Francis Group, United States, pp. 245–263.

- Shen, S., Neyens, D.M., 2017. Assessing drivers' response during automated driver support system failures with non-driving tasks. *J. Saf. Res.* 61, 149–155.
- Simons-Morton, B.G., Bingham, C.R., Ouimet, M.C., Pradhan, A.K., Chen, R., Barretto, A., Shope, J.T., 2013. The effect on teenage risky driving of feedback from a safety monitoring system: a randomized controlled trial. *J. Adolesc. Health* 53, 21–26.
- Soole, D., Watson, B., Bates, L., 2014. Guidelines for developing and operating Learner Driver Mentor Programs. Retrieved from: <https://eprints.qut.edu.au/104256/>.
- Spicer, R., Vahabghaie, A., Bahouth, G., Drees, L., Martinez von Bülow, R., Baur, P., 2018. Field effectiveness evaluation of advanced driver assistance systems. *Traffic Injury Prev.* 19, S91–S95.
- Sportillo, D., Paljic, A., Ojeda, L., 2018. Get ready for automated driving using virtual reality. *Accid. Anal. Prev.* 118, 102–113.
- Statista, 2017. Shipments of the in-vehicle IVISs worldwide from 2015 to 2022 (in million units). Retrieved from: <https://www.statista.com/statistics/784966/in-car-infotainment-systems-shipments-worldwide/>.
- Thigpen, C., Handy, S., 2018. Driver's licensing delay: a retrospective case study of the impact of attitudes, parental and social influences, and intergenerational differences. *Transport. Res. A: Policy Pract.* 111, 24–40.
- Thomas, F., Rilea, S., Blomberg, R., Peck, R., Korbelak, K., 2016. Evaluation of the safety benefits of the Risk Awareness and Perception Training program for novice drivers (DOT HS 812 235). National Highway Traffic Safety Administration, Washington, DC. Retrieved from: https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/12913_-_evaluation_of_an_updated_version-031317_v3b_-_tagged.pdf.
- Toledo, T., Lotan, T., Taubnab - Ben-Ari, O., Grimberg, E., 2012. Evaluation of a program to enhance young drivers' safety in Israel. *Accid. Anal. Prev.* 45, 705–710.
- Unverricht, J., Yamani, Y., Yahoodik, S., Chen, J., Horrey, W.J., 2019. Attention maintenance training: are young drivers getting better or being more strategic? *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 63, 1991–1995.
- Viktorová, L., Sucha, M., 2019. Learning about advanced driver assistance systems—the case of ACC and FCW in a sample of Czech drivers. *Transport. Res. F: Traffic Psychol. Behav.* 65, 576–583.
- Viereckl, R., Koster, A., Hirsch, E., Ahlemann, D., 2016. Connected car report 2016: Opportunities, risk, and turmoil on the road to autonomous vehicles. Retrieved from: <https://www.strategyand.pwc.com/gx/en/insights/2016/connected-car-2016/connected-car-report-2016.pdf>.
- Washington, S., Cole, R.J., Herbel, S.B., 2011. European advanced driver training programs: reasons for optimism. *IATSS Res.* 34, 72–79.
- Watson-Brown, N., Scott-Parker, B., Senserrick, T., 2019. Association between higher-order driving instruction and risky driving behaviours: exploring the mediating effects of a self-regulated safety orientation. *Accid. Anal. Prev.* 131, 275–283.
- Watson-Brown, N., Scott-Parker, B., Senserrick, T., 2018. Development of a higher-order instruction coding taxonomy for observational data: Initial application to professional driving instruction. *Appl. Ergon.* 70, 88–97.
- Watson-Brown, N., Scott-Parker, B., Senserrick, T., 2020. Higher-order driving instruction and opportunities for improvement: Exploring differences across Learner driver experience. *J. Saf. Res. J. Safety Res.* <https://doi.org/10.1016/j.jsr.2020.08.002>.
- Wells, P., Tong, S., Sexton, B., Grayson, G., Jones, E., 2008. Cohort II: a study of learner and new drivers. Retrieved from: <https://www.roadsafetyobservatory.com/Evidence/Details/11556>.
- White, M., Cunningham, L., Titchener, K., 2011. Young drivers' optimism bias for accident risk and driving skill: accountability and insight experience manipulations. *Accid. Anal. Prev.* 43, 1309–1315.
- Zakrajsek, J., Shope, J.T., Greenspan, A.I., Wany, J., Bingham, C.R., Simons-Morton, B.G., 2013. Effectiveness of a brief parent-directed teen driver safety intervention (checkpoints) delivered by driver education instructors. *J. Adolesc. Health* 53, 27–33.

Further Reading

- Alonso Raposo, M., Ciuffo, B., Alves Dies, P., Ardente, F., Aurambout, J.-P., et al., 2019. The future of road transport—implications of automated, connected, low-carbon and shared mobility. EUR 29748 EN, Publications Office of the European Union, Luxembourg. Retrieved from: <https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/future-road-transport>.
- Bates, L., Filtiness, A., Watson, B., 2018. Driver education and licensing programs. In: Lord, D., Washington, S. (Eds.), *Safe Mobility: Challenges, Methodology and Solutions*. Emerald Group Publishing Limited, Bingley, UK, pp. 13–36.
- Regan, M., Prabhakaran, P., Wallace, P., Cunningham, M.L., Bennett, J.M., 2020. Education and training for drivers of assisted and automated vehicles. Austroads Research Report AP-R616-20. Austroads, Canberra Australia. Retrieved from: <https://austroads.com.au/publications/connected-and-automated-vehicles/ap-r616-20>.
- Watson-Brown, N., 2020. Operationalising theoretical frameworks for a best-practice model of higher-order driving instruction for learner drivers. PhD thesis. The University of Sunshine Coast, QLD. https://research.usc.edu.au/permalink/61USC_INST/1vg4fiv/alma99462008502621.

Road Safety Advertising: What We Currently Know and Where to From Here

Ioni Lewis, Barry Watson, Katherine M. White, Sonali Nandavar, Queensland University of Technology (QUT), Kelvin Grove, QLD, Australia

© 2021 Elsevier Ltd. All rights reserved.

Brief Background to Advertising as a Road Safety Intervention	165
What is the Role of Road Safety Advertising?	165
What are the Factors that Underpin Effective Road Safety Advertising?	165
Research Priorities Going Forward	168
Concluding Comments	169
References	169
Further Reading	170

Brief Background to Advertising as a Road Safety Intervention

As a road safety intervention, advertising campaigns represent a long-standing approach. In years gone by, more traditional broadcast media such as television, radio, and print have been the key dissemination channels of such campaigns. Of these broadcast media, perhaps not surprisingly given its potential reach, early meta-analytical evidence of road safety advertising campaigns supported the effectiveness of television relative to other message media (Elliott, 1993). Presently, the use of these traditional media types remains popular among road safety agencies; however, the communication landscape has changed quite dramatically over recent years due to the proliferation and increased use of technology that has led to the identification of new, digital and online communication approaches. Coupled with this changing communication landscape, is the need to be aware of different approaches to campaigns and campaign content and how best to ensure that the messaging reaches an intended audience and has the desired effect of changing attitudes and, ultimately, behavior.

As earlier articles of this book have alluded to, road trauma is a leading contributor to the global burden of disease and injury (Vos et al., 2020). With economies adversely impacted in post-COVID times, road safety authorities the world over will be especially focused on evidence-based interventions that can be delivered at relatively low cost. Thus, now more so than ever, it is critical to understand the role and effectiveness of road safety advertising as a countermeasure and to be aware of gaps in knowledge that future research and practice must address.

With this in mind, this article focuses on road safety advertising campaigns—what influences their effectiveness, what factors underpin successful campaigns, and what are the on-going research priorities. The article aims to identify what is known and what is yet to be known, thereby setting a research agenda for the future.

What is the Role of Road Safety Advertising?

Human factors have long been recognized as contributing to road crashes (Treat et al., 1979). Among the major contributors to such trauma are the “Fatal 5” (Queensland Police Service, 2020), which include speeding, drink driving, distraction, not wearing a seatbelt, and fatigue. The extent to which human decision-making underpins whether or not one engages in each of these behaviors, highlights the important role that advertising campaigns can play in raising awareness of the risks and strategies to prevent engaging in such risky behavior as well as, ultimately, changing attitudes and behaviors so that individuals opt to instead engage in safer, legal on-road behaviors. As Lewis et al. (2019a,b) outline in their review of the role of advertising from a Traffic Safety Culture perspective, road safety advertising can play a reinforcing or transforming role. Specifically, advertising may reinforce the purpose of traffic-laws and their associated enforcement and/or it may seek to “transform” community values and attitudes (see also Watson and Soole, 2013). For instance, with regard to the reinforcing role and in the case of drink driving, road safety advertising may seek to highlight the threat of police detection (e.g., random breath testing) associated with drink driving. In contrast, a transforming approach would focus on changing values and attitudes by highlighting the physical (e.g., death) and emotional (e.g., loss of loved one) harm inflicted upon the victims, their families, and the wider community from alcohol-related road crashes. The reinforcing and transforming approaches are not mutually exclusive and can co-occur within the same advertising campaign (Lewis et al., 2019a,b; Watson and Soole, 2013).

What are the Factors that Underpin Effective Road Safety Advertising?

The available evidence attests to an extensive and varied array of factors that influence the potential effectiveness of road safety advertising. Indeed, it is beyond the scope of this current article to be able to provide an exhaustive review of all such factors. In taking a step back though, before talking about factors that influence “effectiveness”, it also needs to be acknowledged that the

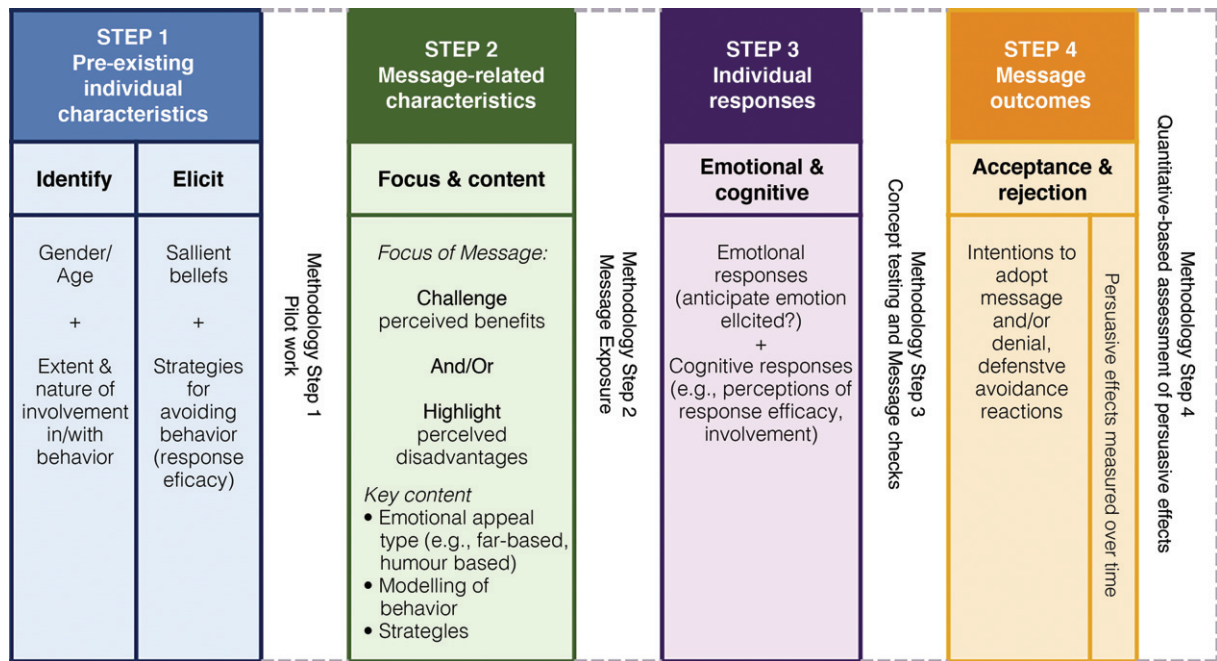


Figure 1 The step approach to message design and testing (SatMDT; Lewis et al., 2016a).

conceptualization of “effectiveness” itself is ambiguous in this context. Due to this conceptual ambiguity, any comparisons across studies can be akin to the adage of “comparing apples and oranges”. For instance, depending on campaign objectives, effectiveness may be assessed in terms of awareness raising, attitude change, or actual behavior change (Elliott, 1993). Appreciably, achieving evidence of having raised awareness of a campaign and/or issue may be more readily achievable (and measurable) than determining the extent to which a particular campaign may have achieved changes in a particular behavior, such as having drivers increase their compliance with the speed limit following their exposure to an anti-speeding campaign.

The challenges associated with defining message outcomes also extends to criticisms directed at evidence being based predominantly on effectiveness as defined in terms of the extent of acceptance achieved as opposed to the extent of message rejection. Theoretical and empirical evidence attests to the need to consider outcomes also in terms of the extent of rejection (e.g., Witte, 1992; see also Lewis et al., 2016a); however, relative to the focus on acceptance, fewer studies incorporate measures of rejection. This omission is problematic to the extent that evidence attests to such outcomes as not being mutually exclusive (Lewis et al., 2010; Tay and Watson, 2002; Witte, 1992) and, thus, to understand the *overall* or *net* effectiveness of a campaign or message, it is necessary to assess rates of both acceptance and rejection.

In the attempt to consolidate the extant evidence in the field in a way that could be used to develop message content and aid message evaluation, the Step approach to Message Design and Testing (SatMDT; Lewis et al., 2016a) was developed (see Fig. 1). The SatMDT is based on social psychological theories of decision making and persuasion including the Theory of Planned Behaviour (TPB; Ajzen, 1991), Social Learning Theory (Bandura, 1969), the Extended Parallel Process Model (EPPM; Witte, 1992), and the Elaboration Likelihood Model (ELM; Petty and Cacioppo, 1986). Given the framework has been outlined comprehensively elsewhere (see Lewis et al., 2016a, 2019a,b), a similar outline will not be presented herein. Instead, noting briefly that, as Fig. 1 of the SatMDT shows, the framework outlines the message design and evaluation process in a series of steps. At each step, the framework identifies a range of key considerations. For instance, the importance of considering message acceptance and rejection in any attempts to evaluate message effectiveness, as was discussed previously, is captured as key consideration within the Message Evaluation step, Step 4, of the framework. When developing message content, as addressed within Step 2 of the framework, factors including emotional appeal type and modeling of behavior are highlighted as evidence-based factors to consider. As will be discussed further shortly, the choice of an emotional appeal type represents a long-standing message factor associated with much contention with Janis and Feshback credited as having conducted the first study of fear and persuasion back in the early 1950s (see Janis and Feshback, 1953).

Of note, the utility of the framework in aiding in the design of road safety messages has been demonstrated across a number of studies and in relation to various road user behaviors and issues including, for example, speeding (in general community [Lewis et al., 2017]), in organizational fleets [Lewis and Newnam, 2011]), speeding in school zones (Glendon and Lewis, 2019), distraction (Gauld et al., 2014; Gauld et al., 2020), child pedestrian safety (Murray et al., 2016), child restraint use (Lewis et al., 2016b), tailgating/vehicle headway (Schramm et al., 2012), drink walking ([intoxicated pedestrians]; via comments provided by the first author for a campaign developed by the Queensland Department of Transport and Main Roads, <https://streetsmarts.initiatives.qld.gov.au/pedestrians/drink-walking>), as well as messaging encouraging use of cooperative and automated vehicles in the future when

they become publicly available (Lewis et al., 2020a). Moreover, its applicability has been demonstrated in relation to the development and/or evaluation of messaging via various media including, for example, television (Lewis et al., 2016b), highway Variable Message Signs ([VMS]; Schramm et al., 2012), portable road-side VMS (Glendon and Lewis, 2019), online/digital (Lewis et al., 2020b), radio (Lewis et al., 2017), and print advertisements (Lewis et al., 2020b). Without question, equally important to specifying from the outset what the intended objectives of a particular campaign are (i.e., raise awareness or change behavior) is to determine the intended media that will be used to disseminate the messaging.

As was noted earlier, and perhaps representing one of the longest standing controversies in the field of health persuasion has been the role of emotion and, in particular, whether or not to rely upon fear to achieve persuasion. Suffice to say, whether or not a proponent or opponent of using fear, evidence suggests the use of fear is to be done so cautiously and the relationship between fear and persuasion is complex. It is acknowledged that, as individuals we all fear different threats to different extents and, as such, it is more about ensuring the threatening stimulus is the most salient and likely to appeal to an intended target audience. This acknowledgment highlights it is important to appreciate that, even when designed to be a “fear appeal”, such an appeal may not actually elicit fear if the threat depicted is not considered a threat that one feels personally susceptible to or as severe. In other words, it is the cognitive processing and appraising of a threat and whether it is relevant and severe that determines whether or not one is likely to experience fear (and as is addressed within Witte’s 1992 Extended Parallel Process Model [EPPM]).

Furthermore, according to models such as the EPPM, it is also not just about the threat but also about the extent to which a message may incorporate strategies that an individual believes are useful in preventing or reducing the threat and they are strategies that they can see themselves implementing; defined in the literature as response efficacy and message self-efficacy, respectively (Witte, 1992). Indeed, meta-analytical evidence attests to the fact that perceived susceptibility and severity, as well as perceptions of efficacy (in terms of response and message self-efficacy) are all positive predictors of message acceptance (see Floyd et al., 2000); however, it is efficacy that is the most important predictor and, thus, crucial determinant likely to increase acceptance and minimize rejection (Lewis et al., 2010).

Despite evidence supporting the important role that efficacy plays in message effectiveness, recent research has suggested that response efficacy is not always incorporated within road safety advertising campaigns. For example, Diegelmann et al. (2020), in a study of mobile phone use while driving campaigns in the UK, reported that response efficacy was rarely incorporated within online campaigns. Similarly, Cismaru (2014) found that fewer than half of the “texting while driving” campaigns included in their analysis of campaigns in the UK, Canada, and USA, had incorporated response efficacy. Another study found only one out of seven road safety messages in Russia had incorporated response efficacy (Ngondo and Klyueva, 2019). Such findings signal the need for greater alignment between research and practice to the extent that response efficacy, a key factor recognized as enhancing acceptance and minimizing rejection, is not always incorporated within road safety advertising messages.

Although traditionally, strong shock-based approaches threatening potential death or injury in a road crash as a result of engaging in risky and/or illegal behavior represented a commonly implemented approach in Australian road safety advertising (Lewis et al., 2019a,b), recent campaigns have tended to not rely as heavily upon such strong physical threats or at least not to rely exclusively upon them as the only approach to presenting road safety messaging. This change is supported by evidence with a meta-analysis conducted by Carey et al. (2013) finding that exposure to such appeals was not a significant determinant of driving outcomes. Additionally, Lewis et al. (2008) found the persuasiveness of fear-evoking emotional appeals for anti-drink driving messages to be greater immediately after exposure but reduce over time, while Anakwah et al. (2015) found that fear-arousing messages did not alter the attitudes of risky drivers.

Research has also found that the likely effectiveness of threat appeals differs between males and females, with males less influenced by messages that incorporated threats of physical harm than females (Lewis et al., 2007a). Similarly, Tay and Ozanne (2002) found that females were more influenced by threat appeals, with a reduction in fatal crashes occurring among females, but not young males, following exposure to a threat-based campaign (Tay and Ozanne, 2002). A recent study by Gauld et al. (2020) examined the effectiveness of three negative emotion-based messages addressing driver distraction and, in particular, smartphone use while driving among young drivers. They found that females were significantly more likely to report the messages as being effective, while the males were significantly more likely to report rejection (i.e., avoidance or denial). In a study conducted by Rodrigues et al. (2015), the researchers explored gender differences with regards to drivers’ perceptions of the effectiveness of road safety advertising campaigns in Portugal. They found that females reported significantly higher intentions to enact the advice offered by the campaigns than males, further highlighting the impact of gender on road safety message effectiveness. To the extent that evidence also highlights that males and females differ in regards to their engagement in risky driving behaviors, such as speeding and mobile phone use while driving (Oviedo-Trespalacios and Scott-Parker, 2018; Oviedo-Trespalacios et al., 2019; Rhodes and Pivik, 2011), collectively, such evidence serves as an important reminder that, consistent with the first step of the SatMDT framework, knowing the intended target audience and what underpins their engagement in particular behaviors is the key first step in message development.

As an evidence-based approach to understanding more about the factors that underpin individuals’ engagement in particular behaviors, the Theory of Planned Behaviour ([TPB]; Ajzen, 1991) offers much insight into the cognitions underpinning individuals’ decisions. Evidence based on the model supports the fact that understanding more about individuals’ pre-existing beliefs about an issue or behavior aids in the prediction of their intentions to engage in risky or unsafe driving behaviors (Atombo et al., 2016; Rowe et al., 2016). Hence, it follows that the identification of these underlying beliefs influencing individuals’ decision-making would provide key foci for the development of targeted road safety advertising (Lewis et al., 2016a). Some studies have identified critical beliefs as potential targets for public education messages targeting young drivers’ engagement in speeding (Horvath et al., 2012) and

smartphone use while driving (Gauld et al., 2016). For instance, the desire to feel in control was a salient advantage of speeding, while infringements were a disadvantage (Lewis et al., 2013). Thus, a message to reduce speeding could focus on the monetary aspect of engaging in this behavior (i.e., waste of a day's pay), while another message could emphasize that drivers are more in control when they are not speeding, considering how they are now more attentive of their driving (Lewis et al., 2013). Given the importance of such underpinning beliefs, it is perhaps not surprising that the identification of beliefs in accordance with the TPB is a key aspect within Step 1 of the SatMDT framework.

Once this formative phase has been conducted and the insights used to inform message content, it is also as important to ensure that concept testing of the messaging occurs with members of the intended target audience. Such testing ensures that messages are functioning as intended with those in the target group. The importance of undertaking such testing is reflected in its being designated as a key step, specifically the third step, within the SatMDT framework.

While the effectiveness of fear-based approaches can vary across demographic groups, an important consideration is the likely effectiveness of alternative emotional appeals. A growing body of evidence supports the role that positive approaches such as humor and pride may play in the road safety advertising context (Lewis et al., 2008). That said, while supporting the role that positive emotions can play in the context of road safety advertising as an alternative to fear, the evidence also highlights that the effectiveness of humorous messaging may also vary across different demographics. For instance, Lewis et al. (2008) found, in a study on negative and positive anti-speeding messages, that while males tended to be more influenced by the positive messaging than the negative, fear-based messaging, for females, opposite findings emerged. Similarly, Plant et al., (2017) found that anti-speeding messages which evoked positive emotions were found to reduce young drivers' speeding behavior.

Similar to the use of fear, evidence has cautioned that the use of humor in road safety campaigns must be appropriate; for instance, humor should never be associated with serious, negative outcomes such as road crashes (Lewis et al., 2007b). One of the potential advantages of using positive emotion-based approaches is that they may more readily (relative to negative, fear-based approaches) enable the use of the modeling of positive or safe behaviors and the rewards or, at least the avoidance of negative outcomes, being shown as consequences of engaging in such positive, safe behaviors. In other words, by their nature, the use of positive behavioral modeling aligns well with the use of positive approaches which may seek to elicit more positive feelings in audience members as to having done the right thing and being able to feel positively about that. In contrast, a typical negative or fear-based approach couples depiction of an individual engaging in an unsafe, risky, and/or illegal behavior (e.g., drink driving) and accompanies it with the negative consequences that may result (e.g., being caught and fined by police, being injured in a road crash).

Research Priorities Going Forward

In this final section, we overview what we consider are some of the most pressing areas of need with respect to furthering understanding the role and effectiveness of advertising as a road safety intervention. We preface this section by acknowledging that our intent is not to offer an exhaustive list but rather to provide some suggestions regarding potential avenues to advance contemporary knowledge.

First, there is need to understand more about road safety advertising in the online context. Not surprisingly given that there are decades of campaigns that have been disseminated via traditional broadcast media, much of the available evidence has been on the use and effectiveness of messaging delivered via these more traditional platforms. However, traditional media is a *one-way communication* platform, whereas digital communication is *dynamic*, unpredictable, and one in which an individual can be a co-creator of content. As people curate their "self" online, factors relating to identity, impression management, and normative expectations are likely to influence their responses and likelihood to contribute to digital campaigns (e.g., whether to post, share, or like content). Although online analytics reveal the numbers of "likes" and shares that a campaign receives, the key question is whether an individual ultimately adopts its recommendations and, therefore, changes their attitudes and behaviors. Traditional models of persuasion offer no guidance on the anticipated correspondence between these analytics and behavior change. Thus, traditional frameworks can neither account for the intricacies and the complex interplay of factors that influence an individual's engagement with and the effectiveness of digital campaigns, nor can these frameworks determine the ability of such communications to change people's attitudes and behavior.

This need is only too apparent when considering that, in Australia, advertising spend on digital advertising is projected to reach over \$9 million in 2020 (Statista, 2020). With it fast approaching 30 years since a seminal meta-analysis on road safety advertising campaigns was published (Elliott, 1993), the field would benefit from a meta-analytic review that captured the effect sizes associated with various media including communication platforms available via the digital context. In 1993, Elliott concluded that it was television that was associated with the largest effect size relative to the other traditional media. This finding was not considered that surprising given television was the medium associated with the greatest potential reach. Nowadays, however, that is not necessarily the case with particular demographics, such as young adults, likely to rely more upon online and digital options for their communication and information-seeking needs (Murray and Lewis, 2011; Wundersitz et al., 2010).

Extending upon this, there is a need to understand more about the types of messaging approaches likely to be effective for online platforms. For instance, with regards to advertising campaigns in general (and not road safety advertising in particular), some research has explored the emotional appeal type of viral campaigns. Hsieh et al. (2012), for instance, found comedic or humorous online videos to play an important role in influencing positive attitudes towards a received video, as well as intentions to forward the video onto others. Specifically, these types of videos can elicit positive emotions in viewers and encourage them to share their

emotions with others (Hsieh et al., 2012). These findings are in line with another study conducted by Berger and Milkman (2012), who found that positive online content was more likely to be highly shared, even after controlling for the frequency of occurrence, and online content that evoked a de-activating emotion (e.g., sadness) was less likely to go viral.

Second, there remains a gap in knowledge when considering the role and effectiveness of different types of humor in road safety advertising (Lewis et al., 2018) and in relation to the persuasive effects of other types of emotion more broadly. There is also need to understand the relative effectiveness of different types of humor and other emotions more broadly in relation to different audiences.

Third, there is a need to continue efforts regarding the application of the most robust methodological approaches to assess the effects of road safety advertising. Much of the available evidence is based on self-report methods such as questionnaires (Borawska et al., 2020). Although such information provides insights into individuals' perceptions, it would also be useful to develop a deeper understanding of individuals' automatic and unconscious reactions towards different emotional appeals (Borawska et al., 2020). This understanding could be gained through various neuroscientific methods such as psychophysiological (e.g., heart rate, skin conductance; Ordoñana et al., 2009) and electrophysiological (e.g., EEG; Borawska et al., 2020) measures. Moreover, research has found objective measures (e.g., skin conductance response, GPS devices) could be used in conjunction with self-report measures to understand both the persuasive processing and outcomes of emotion-based anti-speeding advertisements as well as confirm the robustness of the respective methods (Kaye et al., 2016). Another important consideration when evaluating a campaign is to ensure comparisons are made between groups of individuals who have seen/heard a campaign message as well as those who have not. Those who do not see/hear the message can be considered as a baseline or control group. The responses from those who did see/hear the message can be compared with the control group and, thus, a more comprehensive understanding of the nature and extent of changes can be determined (Kaye et al., 2016; Robertson and Pashley, 2015).

Finally, and extending upon the preceding point, there is an ongoing need for road safety advertising campaigns to be evaluated and for these evaluations to be robust and theoretically underpinned (Lewis et al., 2016). Such evaluations ensure that the evidence-base continues to grow and that lessons are learned as to how and why a particular campaign was or was not successful. For example, there remains a need to explore the relative effectiveness of road safety advertising when it is used to support other initiatives versus being used in isolation.

Concluding Comments

This article provided a review of where knowledge currently stands regarding the role and effectiveness of advertising as a road safety intervention and concluded by highlighting some suggested directions for future research. While the nature of road safety advertising has evolved over the last two decades, it continues to be relied on extensively to bring about changes in road safety awareness, attitudes and behaviors and to support other initiatives. Hence, it is essential that research continues to identify whether the resources devoted to road safety advertising are being used in the most optimal way.

References

- Anakwah, N., Akotia, C.S., Osafo, J., Parimah, F., Sarfo, J.O., Aggrey, G.B., 2015. Risky driving attitudes in Ghana: is the use of fear-based messages operational? *Eur. Res.* 97 (8), 560–567.
- Ajzen, I., 1991. The theory of planned behavior. *Organ. Behav. Human Decision Process.* 50 (2), 179–211.
- Atombo, C., Wu, C., Zhong, M., Zhang, H., 2016. Investigating the motivational factors influencing drivers intentions to unsafe driving behaviours: speeding and overtaking violations. *Transport. Res. Part F: Traffic Psychol. Behav.* 43, 104–121.
- Bandura, A., 1969. *Principles of behavior modification*.
- Berger, J., Milkman, K.L., 2012. What makes online content viral? *J. Market. Res.* 49 (2), 192–205.
- Borawska, A., Oleksy, T., Maisson, D., 2020. Do negative emotions in social advertising really work? Confrontation of classic vs EEG reaction toward advertising that promotes safe driving. *PLoS ONE* 15 (5), e0233036.
- Carey, R.N., McDermott, D.T., Sarma, K.M., 2013. The impact of threat appeals on fear arousal and driver behavior: A meta-analysis of experimental research 1990–2011. *PLoS ONE* 8 (5), pe62821.
- Cismaru, M., 2014. Using the extended parallel process model to understand texting while driving and guide communication campaigns against it. *Social Market. Quart.* 20 (1), 66–82.
- Diegelmann, S., Ninaus, K., Terlutter, R., 2020. Distracted driving prevention: an analysis of recent UK campaigns. *J. Social Market.* 10 (2), 243–264.
- Elliott, B., 1993. Road safety mass media campaigns: A meta analysis (No. CR 118). Federal Office of Road Safety.
- Floyd, D.L., Prentice-Dunn, S., Rogers, R.W., 2000. A meta-analysis of research on protection motivation theory. *J. Appl. Soc. Psychol.* 30 (2), 407–429.
- Gauld, C.S., Lewis, I., White, K.M., 2014. Concealed texting while driving: What are young people's beliefs about this risky behaviour? *Safety Sci.* 65, 63–69.
- Gauld, C.S., Lewis, I.M., White, K.M., Watson, B., 2016. Young drivers' engagement with social interactive technology on their smartphone: critical beliefs to target in public education messages. *Accid. Anal. Prevent.* 96, 208–218.
- Gauld, C.S., Lewis, I.M., White, K.M., Watson, B.C., Rose, C.T., Fleiter, J.J., 2020. Gender differences in the effectiveness of public education messages aimed at smartphone use among young drivers. *Traffic Injury Prevent.* 21 (2), 127–132.
- Glendon, A.I., Lewis, I., 2019. Field testing anti-speeding messages. In 2019 Australasian Road Safety Conference, Adelaide, South Australia. <https://acrs.org.au/files/papers/arsc/2019/JACRS-D-19-00010-Glendon.pdf>.
- Horvath, C., Lewis, I., Watson, B., 2012. The beliefs which motivate young male and female drivers to speed: A comparison of low and high intenders. *Accid. Anal. Prevent.* 45, 334–341.
- Hsieh, J.K., Hsieh, Y.C., Tang, Y.C., 2012. Exploring the disseminating behaviors of eWOM marketing: persuasion in online video. *Electronic Commerce Res.* 12 (2), 201–224.
- Janis, I.L., Feshbach, S., 1953. Effects of fear-arousing communications. *J. Abnormal Social Psychol.* 48 (1), 78.
- Kaye, S.A., Lewis, I., Algie, J., White, M.J., 2016. Young drivers' responses to anti-speeding advertisements: comparison of self-report and objective measures of persuasive processing and outcomes. *Traffic Injury Prevent.* 17 (4), 352–358.

- Lewis, I., Watson, B., Tay, R., 2007a. Examining the effectiveness of physical threats in road safety advertising: The role of the third-person effect, gender, and age. *Transport. Res. Part F: Traffic Psychol. Behav.* 10 (1), 48–60.
- Lewis, I.M., Watson, B., White, K.M., Tay, R., 2007b. Promoting public health messages: Should we move beyond fear-evoking appeals in road safety? *Qualitative Health Res.* 17 (1), 61–74.
- Lewis, I., Watson, B., White, K.M., 2008. An examination of message-relevant affect in road safety messages: Should road safety advertisements aim to make us feel good or bad? *Transport. Res. Part F: Traffic Psychol. Behav.* 11 (6), 403–417.
- Lewis, I.M., Watson, B., White, K.M., 2010. Response efficacy: the key to minimizing rejection and maximizing acceptance of emotion-based anti-speeding messages. *Accid. Anal. Prevent.* 42 (2), 459–467.
- Lewis, I., Watson, B., White, K.M., Elliott, B., 2013. The beliefs which influence young males to speed and strategies to slow them down: informing the content of anti-speeding messages. *Psychol. Market.* 30 (9), 826–841.
- Lewis, I., Ho, B., Lennon, A., 2016b. Designing and evaluating a persuasive child restraint television commercial. *Traffic Injury Prevent* 17 (3), 271–277.
- Lewis, I., Watson, B., White, K.M., 2016a. The Step approach to Message Design and Testing (SatMDT): A conceptual framework to guide the development and evaluation of persuasive health messages. *Accid. Anal. Prevent.* 97, 309–314.
- Lewis, I., White, K.M., Ho, B., Elliott, B., Watson, B., 2017. Insights into targeting young male drivers with anti-speeding advertising: an application of the Step approach to Message Design and Testing (SatMDT). *Accid. Anal. Prevent.* 103, 129–142.
- Lewis, I., Watson, B., White, K.M., 2018, October. Exploring the effectiveness of different types of humour in road safety advertising campaigns. In *Australasian Road Safety Conference, 2018, Sydney, New South Wales, Australia*.
- Lewis, I., Forward, S., Elliott, B., Kaye, S.A., Fleiter, J.J., Watson, B., 2019a. Designing and evaluating road safety advertising campaigns. In: *Traffic Safety Culture*, Emerald Publishing Limited.
- Lewis, I., Elliott, B., Kaye, S.A., Fleiter, J.J., Watson, B., 2019b. In: *The Australian Experience with Road Safety Advertising Campaigns in Improving Traffic Safety Culture*. In *Traffic Safety Culture: Definition, Foundation and Application*. Emerald Publishing Limited, pp. 275–295.
- Lewis, I., Kaye, S.-A., Nandavar, S., Briant, O., Murray, C. 2020a. Beliefs influencing intended use of private and shared CAVs. Accepted for the 2020 International Conference of Transport and Traffic Psychology (ICTTP), Gothenburg, Sweden (conference did not occur in 2020 due to COVID).
- Lewis, I., Kerr, G., White, K.M., Nandavar, S., 2020b. Laughing or crying at distracted driving messages? Accepted for the 2020 International Conference of Transport and Traffic Psychology (ICTTP), Gothenburg, Sweden (conference did not occur in 2020 due to COVID).
- Murray, C., Lewis, I., 2011. Is there an app for that? Social media uses for road safety. In Healy, D., (Ed.) *Proceedings of the Australasian College of Road Safety National Conference 2011*. Australasian College of Road Safety, Australia, pp. 1–15.
- Murray, C., Lewis, I., Lennon, A., Vuong, K., Haworth, N., 2016. The development and evaluation of a child pedestrian safety campaign. In *Safety 2016 World Conference, 2016-09-18-2016-09-21*.
- Ngondo, P.S., Klyueva, A., 2019. Fear appeals in road safety advertising: an analysis of a controversial social marketing campaign in Russia. *Russian J. Commun.* 11 (2), 167–183.
- Ordoñana, J.R., Gonzalez-Javier, F., Espin-Lopez, L., Gomez-Amor, J., 2009. Self-report and psychophysiological responses to fear appeals. *Human Commun. Res.* 35 (2), 195–220.
- Oviedo-Trespalacios, O., Scott-Parker, B., 2018. The sex disparity in risky driving: a survey of Colombian young drivers. *Traffic Injury Prevent.* 19 (1), 9–17.
- Oviedo-Trespalacios, O., Nandavar, S., Newton, J.D.A., Demant, D., Phillips, J.G., 2019. Problematic use of mobile phones in Australia . . . is it getting worse? *Front. Psychiatry* 10, 105.
- Petty, R.E., Cacioppo, J.T., 1986. In: *The elaboration likelihood model of persuasion*. In *Communication and Persuasion*. Springer, New York, NY, pp. 1–24.
- Plant, B.R., Irwin, J.D., Chekaluk, E., 2017. The effects of anti-speeding advertisements on the simulated driving behaviour of young drivers. *Accid. Anal. Prevent.* 100, 65–74.
- Queensland Police Service, 2020, *The Fatal Five- staying safe on the roads*, Queensland Government, <https://www.police.qld.gov.au/initiatives/fatal-five-staying-safe-roads>.
- Rhodes, N., Pivik, K., 2011. Age and gender differences in risky driving: the roles of positive affect and risk perception. *Accid. Anal. Prevent.* 43 (3), 923–931.
- Robertson, R.D., Pashley, C.R., 2015. Road safety campaigns: What the research tells us. *Traffic Injury Research Foundation*. https://turf.ca/wp-content/uploads/2017/01/2015_RoadSafetyCampaigns_Report_2.pdf.
- Rowe, R., Andrews, E., Harris, P.R., Armitage, C.J., McKenna, F.P., Norman, P., 2016. Identifying beliefs underlying pre-drivers' intentions to take risks: An application of the Theory of Planned Behaviour. *Accid. Anal. Prevent.* 89, 49–56.
- Rodrigues, H.S., Manuel, O.F., Cardoso, P.R., 2015. The effect of drivers gender on the perception of Portuguese road safety communication campaigns. *Int. J. Bus. Manage.* 11 (4).
- Schramm, A., Rakotonirainy, A., Smith, S., Lewis, I., Soole, D., Watson, B., Troutbeck, R., 2012. Effects of speeding and headway related variable message signs on driver behaviour and attitudes. *Queensland Department of Transport and Main Roads, Australia*.
- Statista, 2020. Digital Advertising. Retrieved from <https://www.statista.com/outlook/216/107/digital-advertising/australia>.
- Tay, R.S., Ozanne, L., 2002. Who are we scaring with high fear road safety advertising campaigns? *Asia Pacific J. Transport* 4, 1–12.
- Tay, R., Watson, B., 2002. Changing drivers' intentions and behaviours using fear-based driver fatigue advertisements. *Health Market. Quart.* 19 (4), 55–68.
- Treat, J.R., Tumbas, N.S., McDonald, S.T., Shinar, D., Hume, R.D., Mayer, R.E., Stansifer, R.L., Castellan, N.J., 1979. Tri-level study of the causes of traffic accidents: final report. Executive summary. Indiana University, Bloomington, Institute for Research in Public Safety.
- Vos, T., Lim, S.S., Abbafati, C., Abbas, K.M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., Abdelalim, A., Abdollahi, M., 2020. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258), 1204–1222.
- Watson, B., Soole, D.W., 2013. Traffic safety culture in Australia: Contrasting community perceptions to drink driving & speeding. Presentation at the Transportation Research Board-Panel Session, 311.
- Witte, K., 1992. Putting the fear back into fear appeals: The extended parallel process model. *Commun. Monographs* 59 (4), 329–349.
- Wundersitz, L.N., Hutchinson, T.P., Woolley, J.E., 2010. Best practice in road safety mass media campaigns: a literature review. *Social Psychol.* 5, 119–186.

Further Reading

- Petty, R.E., Cacioppo, J.T., 1979. Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *J. Personality Social Psychol.* 37 (10), 1915.

Traffic Law Enforcement Theories and Models

Richard Tay, RMIT University, Melbourne, VIC, Australia

© 2021 Elsevier Ltd. All rights reserved.

Introduction	171
Deterrence Theory	171
Criminology Perspective	171
Social Learning Perspective	172
Economic Perspective	172
Deterrence Mechanism	172
Certainty of Punishment	172
Severity of Punishment	172
Immediacy of Punishment	172
Punishment and Punishment Avoidance	173
General and Specific Deterrence	173
Enforcement Models	173
Steady States of Violation Rates	173
Time and Distance Halo	173
Halo Effect, Learning and Violation Rate	174
Halo Effect and Enforcement Coverage	174
Learning Theory and Enforcement Scheduling	174
Economic Theory of Choice under Uncertainty	175
Deterrence Production Frontier	175
Conclusion	176
References	176

Introduction

Traffic collisions are a leading cause of deaths and serious injuries in many countries and cost societies an average of one-two percent of their gross national product (WHO, 2003). In an effort to reduce and prevention collisions, many countries have implemented a variety of engineering, education and enforcement strategies. The role of traffic law enforcement is to ensure that road users comply with the traffic laws by apprehending and punishing those who commit a violation. If traffic legislations are designed to improve safety, then reducing violations will result in fewer traffic collisions.

Traffic rules and regulations are designed to prevent risky road user behaviors and ensure a safe road environment. However, not all road users will obey all them and many road users may also unintentionally violate them. Regardless of intention, a traffic law violation may significantly increase the likelihood of a traffic collision, resulting in property-damage, injuries and deaths. Therefore, it is important that traffic laws be enforced adequately to reduce and prevent illegal and risky road user behaviors.

Although there are a multitude of laws covering a variety of areas, including vehicle standards and safe operations, most of the foci in the traffic safety literature have been on high risk driver behaviors such as speeding, drink-driving, not wearing seatbelt, driving while fatigued, and distracted driving. The need for and the effectiveness of the enforcement of these laws have been well research and documented in the literature.

The purpose of this book chapter is to provide a conceptual framework to inform the development and evaluation of traffic law enforcement, as well as to provide some possible explanation for some of the mixed results obtained in previous studies.

Deterrence Theory

The European "Escape" Project proposed a Compliance Model of traffic law enforcement that described the process: Legislation → enforcement → Objective Risk → Subjective Risk → Deterrence → Compliance (Mäkinen et al., 2003). Although the model also has a secondary path in which other support measures can affect the objective and subjective risks, deterrence is the main mechanism driving compliance in this model.

Criminology Perspective

In criminology and social psychology, the concept of deterrence is used to describe the prevention of criminal behavior through legal sanctions or punishment. Deterrence refers to compliance produced by the existence and administration of criminal law rather than

some other source (Meier and Johnson, 1977). Deterrence occurs when a particular sanction functions to prevent the occurrence of an act or is thought to decrease the probability of the act's occurrence. Traditionally, deterrence has been conceptualized as involving two distinct types of processes: specific deterrence and general deterrence. Specific deterrence operates when an individual who has committed a crime and has been sanctioned refrains from committing additional criminal acts for fear of another punishment. General deterrence occurs when a would-be offender refrains from committing a crime because another has been punished for offending.

Social Learning Perspective

In social learning theory, general deterrence is defined as the deterrent effect of indirect experience with punishment and punishment avoidance while specific deterrence is defined as the deterrent effect of direct experience with punishment and punishment avoidance (Stafford and Warr, 1993). This approach has several advantages over the traditional definition. First, it recognized the possibility that both general deterrence and specific deterrence could operate for any given person or in any population. Second, it treated punishment avoidance as analytically distinct from the experience of suffering a punishment. Third, it was compatible with contemporary learning theory, particularly the distinction between observational/vicarious learning and experiential learning.

Economic Perspective

Consistent with these views, the main focus of the deterrence theory in the field of economics is on increasing the individual's perceived expected cost of engaging in an illegal activity (Tay, 2005a, 2005b, 2005c, 2005d). An increase in the expected cost is posited to decrease the level of illegal activity consumed by the individual or decrease the likelihood of the individual to engage in the illegal activity. This increase in the perceived cost could be a result of an increase in any or all of the severity of the sanction, the probability of apprehension or the swiftness of punishment.

Deterrence Mechanism

There are three factors that influence the effectiveness of any deterrence policy based on the concept of punishment: certainty of punishment, severity of punishment, and immediacy of punishment. In the road safety arena, the severity of the punishment is determined mainly by the legislature whereas the swiftness of the punishment is determined mainly by the administrative or judiciary system. The main impetus of law enforcement therefore is in increasing the certainty of apprehension and punishment.

Certainty of Punishment

From an economic perspective, an increase in the perceived probability of apprehension will increase the expected cost of committing in violation and thus reduce the likelihood of the violation. The perceived probability of apprehension is a subjective probability that depends on the individual's information set. Two of the major determinants of this information set are the level of police presence on the roads and the hit (apprehension) rate. It can be argued that the former focuses more on general deterrence whereas the latter focuses more on specific deterrence. Thus, traffic law enforcement agencies need to consider optimal mix of covert versus overt operations, high traffic versus high violation areas, and manned versus automated enforcement (Tay, 2009, 2010).

Severity of Punishment

From an economic perspective, an increase in the severity of punishment will increase the expected cost of committing in violation and thus reduce the likelihood of the violation. Although the severity of punishment is determined mainly by the legislature, law enforcement officers often have some discretion, such as issuing a warning or a ticket. Ennis (1967) studied the effect of warning versus citation whereby each technique was tried on 100 speeding motorists and found that 33% of those issued a citation exceeded the speed again while 65% of those issued a warning exceeded the speed again 5 miles down the road. Thus, there is evidence that the severity of punishment has an effect through specific deterrence. However, Mann et al. (1991) found that increasing the severity of punishment, such as amount of fine, length of probation and length of jail, did not have any significant effect of collision reduction. Overall, the evidence suggests that the type of punishment is a better measure of severity than the absolute amount (within certain limits) of the same sanction.

Immediacy of Punishment

From an economic perspective, a delay in punishment, especially financial penalty, will decrease the expected cost of committing in violation and thus increase the likelihood of the violation. However, since the delay is small, this effect is unlikely to be significant. From a learning perspective, immediate feedback will provide a better association between violation and punishment. Moreover, a longer delay in punishment will also provide more opportunities to learn from punishment avoidance, especially when the certainty of punishment is low. Experimental studies on the effect of punishment on behavior generally indicate that immediate punishment

is far more effective than delayed punishment. Hence, all else held constant, manned enforcement which provides an immediate sanction should be more effective than automated enforcement which has a longer delay between the time of violation and the time when the driver received an infringement notice (Tay, 2009).

Punishment and Punishment Avoidance

One of the key assumptions underlying both the learning and economic perspectives of deterrence theory is that drivers learn from both punishment and punishment avoidance. Based on their own experience of punishment or punishment avoidance and the experience of others, drivers will revise their perceptions of the probability of being caught. This learning process is especially true for commuters and drivers who pass through a location regularly. In particular, many drivers will learn relatively quickly for fixed location enforcement, such as road safety cameras. This adaptive learning from punishment and maladaptive learning from punishment avoidance is critical consideration when designing the optimal enforcement strategy, in terms of spatial and temporal coverage, given a limited amount of resources (Tay and de Barros, 2009, 2011).

General and Specific Deterrence

Both general and specific deterrence are likely to be present in most types of enforcement activities. In the case of road safety cameras, there is no need for the actual presence of police. Instead, general deterrence is assured by the presence of camera installations and equipment, such as enforcement signage, the camera casings and poles. It should be stressed that the presence of camera casings can only provide a general deterrent effect and not specific deterrent effect. Tay (2005a, 2009) and Tay and de Barros (2011) argued that without some threshold level of punishment and specific deterrence, efforts aimed at general deterrence do not work because drivers will soon realize that few offenders ever get caught and punished. Therefore, the common belief that the presence of camera casings at camera-enforced locations is sufficient to deter drivers is only consistent with the mechanism of general deterrence but not specific deterrence and its overall effectiveness is unknown. Thus, traffic law enforcement agencies need to consider optimal mix of covert versus overt operations, high traffic versus high violation areas, and manned versus automated enforcement, to produce the maximum overall deterrence effect (Tay, 2009, 2010; Tay and de Barros, 2006, 2011).

Enforcement Models

Steady States of Violation Rates

We can assume that a significant proportion of the drivers are risk adverse or have a morale attachment to the law. These people will generally not violate the road rules and will form a limit or the upper threshold of the violation rate. However, there will be a small proportion of the drivers who are low to moderate risk risk-takers. They will speed, run a red light, not wear a seatbelt or drink-and-drive under some circumstances, such as driving under time pressure or under pressure from impatient drivers behind them, driving for a short distance, and when they think that they are able to get away with the offence without being caught. A significant portion of this group of errant drivers can be deterred from committing an offence if their perceived probability of apprehension is sufficiently high. Finally, there exist a very small proportion of high-risk drivers who will speed, run the red, not wear and seatbelt or drink and drive regardless of the perceived probability of apprehension. They include drivers who are very high-risk takers, drivers with very high value of time or high income relative to the monetary penalty. These drivers will form the lower threshold of the violation rate, until they are effectively remove from the driving, including driving unlicensed.

The above theoretical conceptualization leads to three simple hypotheses about violation rates. First, even if enforcement is absent permanently or for an extended period of time, the violation rate will not be 100% at most locations. There is an upper steady state limit. Second, even if the enforcement is present permanently or for an extended period of time, the violation rate will not be zero. There is a lower steady state limit. Third, when enforcement is intermittent, the violation rate at any location and time will be between these two thresholds and will depend on the durations of enforcement and the apprehension rate. The steady state thresholds or limits will depend on the type of behavior, risk profile of the driving population as well as the severity and swiftness of punishment.

Time and Distance Halo

Besides having a differential effect on different road user types, enforcement also has a differential effect by time and space. This is particularly true for transient behaviors such as speeding and red light running. Time halo can be defined as the length of time that the effects of enforcement on drivers' behavior continue after the enforcement operations have been ended. Distance halo is defined as the distance over which the effects of an enforcement operation is effective.

Although time and distance halos are commonly assumed to be a step function in practice, they should theoretically be considered as a decreasing function from the time when enforcement ceases or from the location of the enforcement. The rate of decrease will depend largely on the type of enforcement and the road users' risk profile. Red light camera, for example, may have a slower rate of decrease over time than mobile speed enforcement but has a faster rate of decrease over distance. Even for the same type of enforcement, such as mobile speed camera, some drivers will resume speeding much earlier (either by time or distance) than

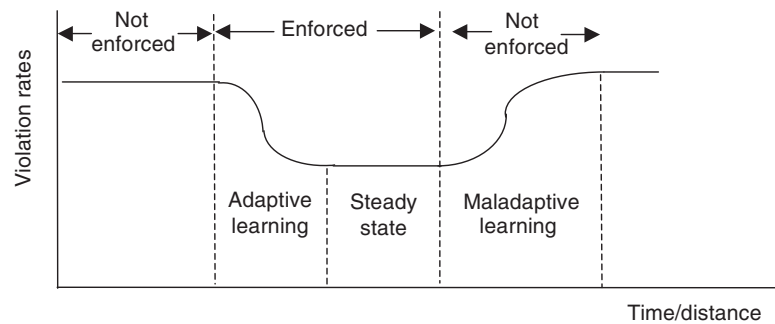


Figure 1 Adaptive learning and maladaptive learning.

others after they have passed the police vehicles. Assuming that the driving population is normally distributed, then the deterrence effect is expected to decrease exponentially with time and space.

Halo Effect, Learning and Violation Rate

As shown in **Fig. 1**, starting from an absence position, when enforcement is first present on at any location, the violation rates is expected to show an exponentially decreasing effect until it reaches a lower steady state threshold. We shall call this phase the adaptive learning phase where drivers learn over time that enforcement is present and decrease their likelihood of committing a violation. If the enforcement continues for a sufficient period of time, the violation rate will reach and remain at the lower steady state level. Conversely, when enforcement is absent or removed, drivers will learn eventually that there is no enforcement and increase their likelihood of committing a violation. This time period is associated with maladaptive learning and is typically known as the time and/or distance halo in traffic enforcement. [Tay and De Barros \(2009, 2011\)](#) found that the maladaptive learning period is much shorter than the adaptive learning period.

Halo Effect and Enforcement Coverage

It is a common practice for enforcement to be rotation around different locations due to time and budget constraints. Hence, it is important to understand the relationship between halo effects and violation rate over time and space to optimize the allocation of enforcement resources.

Once the time halo is ascertained, the maximum time between successive enforcement or the maximum length of nonenforced time to achieve effective coverage is known. The ratio of the required enforced time to nonenforced time will provide an estimate of the number of enforcement units required to effectively enforce a given number of locations or the number of locations that can be effectively enforced using a fixed number of enforcement units ([Tay and de Barros, 2009](#)).

It should be noted that distance has both an adaptive and maladaptive effect. This is especially true for highly visible enforcement. Some drivers will begin to change their behavior, such as slowing down, before they reach the enforcement point or location and will continue to drive slower for some distance after they passed the point of enforcement, especially for visible mobile enforcement units. These pre- and postdistance halos are critical parameters when considering the number of enforcement locations needed to effectively cover a given geographical area.

Learning Theory and Enforcement Scheduling

Theoretically, focusing the limited enforcement resources at a smaller number of locations will increase the likelihood of punishment for a smaller group of drivers, but decreases the likelihood for a larger group of drivers. On the other hand, spreading out the enforcement to a larger number of locations will increase the general deterrence but reduce the specific deterrence.

In practice, many road safety professionals have advocated that enforcement should be random and unpredictable, which is neither based on any behavioral theory nor any empirical evidence. Part of this misconception arises due to misconstruing enforcement as a negative reinforcement which is the removal of an undesirable thing from the person when he/she behaves in the desired manner. Negative reinforcement rarely occurs in law traffic enforcement. On the other hand, punishment is much more prevalent in traffic enforcement. Positive punishment which is the imposition of an undesirable thing to discourage bad behavior is quite common, including the imposition of demerit points, monetary fine or detention. Negative punishment which is the removal of a desirable thing to deter bad behavior, such as license revocation, also occurs in traffic enforcement although to a lesser degree. The empirical evidence shows that whereas negative reinforcement is more effective when administered intermittently (variable ratio), punishment is more effective when administered continuously. Hence, the practice of scheduling enforcement to be random and unpredictable should be taken with caution. The empirical evidence thus far suggested that focusing enforcement in smaller number of locations is more effective than randomly scheduling enforcement among a larger number of locations ([Tay and de Barros, 2011](#)).

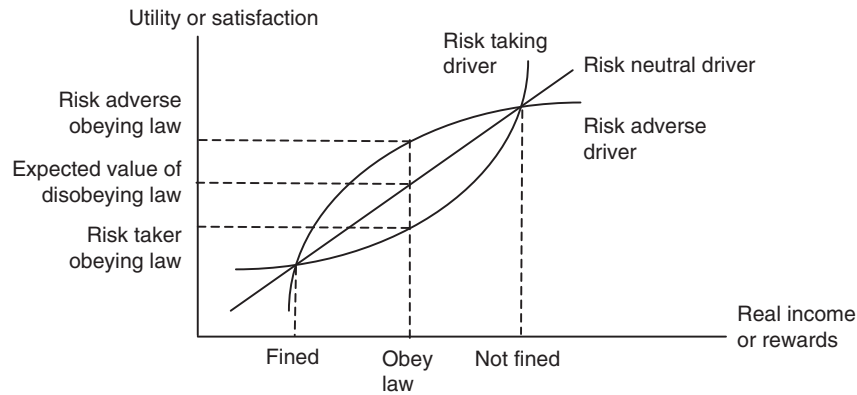


Figure 2 Risk propensity and violations under uncertainty.

Economic Theory of Choice under Uncertainty

In economics, a risk-taking consumer is assumed to always choose the higher risk alternative with a greater reward than and a less risky alternative with a smaller reward, as shown in Fig. 2. If traffic enforcement is uncertain, then a risk-taking consumer will always choose to disobey the traffic law because the expected utility or satisfaction from not obeying the law is higher than obeying the law. If traffic law enforcement efforts are targeted at high risk-taking drivers, then it should not be uncertain and unpredictable. However, if the enforcement efforts are targeted at the low-moderate risk drivers, then it should be unpredictable.

Deterrence Production Frontier

Since enforcement resources are limited, it is vital to optimize the mix between general and specific deterrence to maximize the safety or the overall deterrence effect. Many different strategies and experiments have been proposed, implemented and/or evaluated, with differing results. This is not surprising since the effectiveness of the strategy will depend on many operational factors, including the type of behavior, risk profile of the user population, enforcement method, enforcement time and location, and the overall level of enforcement or amount of resources used, which differ substantially across different studies.

Theoretically, different enforcement strategies will produce different combinations of general and specific deterrence. As an example, consider the deployment of road safety cameras, and for simplicity, consider two operational aspects of deployment: overttness and location. Safety cameras can be deployed overtly and visibility at high visibility or high traffic locations to maximize general deterrence or covertly at high violation locations to maximize specific deterrence. As shown in Fig. 3, given a fixed amount of enforcement resources, there are many different combinations of overt/covert and high visibility/high violation combinations that can be used. Some of the combinations are expected to maximize the amount of deterrence produced and will form the deterrence production frontier while other less effective strategies will be located inside the frontier. An overall safety or deterrence level higher than the frontier level can only be achieved by increasing the amount of enforcement resources.

To optimize the allocation of scarce law enforcement resources, only those combinations on or near the frontier should be considered. The best combination to use will depend on the profile of the road user population. More importantly, more research needs to be conducted to identify the operational factors influencing the effectiveness of different enforcement strategies as well as quantifying the relative effects of the different operational factors. Till date, very little research has been conducted to directly compared the different strategies and fewer have attempted to estimate the deterrence production function or safety production function for traffic law enforcement.

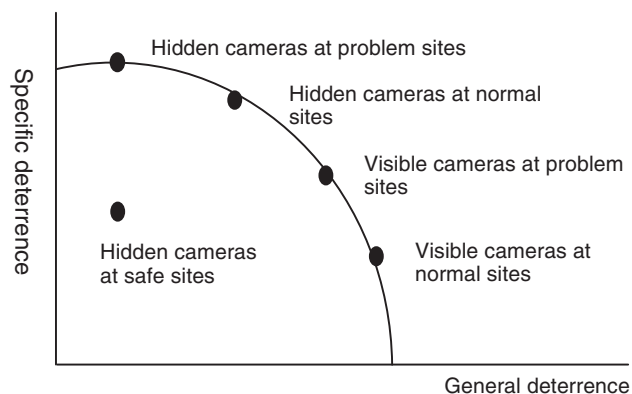


Figure 3 Deterrence production frontier.

Conclusion

Traffic law enforcement is one of the three pillars of road safety. The experience from around the world has shown that it is an effective method to reduce traffic violations and crashes. However, the design and implementation of the traffic strategies in different jurisdictions varies considerably. This article provides researchers and relevant authorities with a summary of some theories and models that can be used to design and evaluate traffic law enforcement programs.

References

- Ennis, P., 1967. Traffic citation are more effective. *Traffic Digest Rev.* 15, 11–16.
- Mäkinen, T., Zaidel, D., Andersson, D., et.al., 2003. Traffic enforcement in Europe: effects, measures, needs and future. The Escape Project, The European Commission.
- Mann, R., Vingilis, E., Gavin, D., Adlaf, E., Anglin, L., 1991. Sentence severity and the drinking driver: relationship with traffic safety outcomes. *Acc. Anal. Prevent.* 23 (6), 483–491.
- Meier, R., Johnson, W., 1977. Deterrence as social control: the legal and extralegal production of conformity. *Am. Sociol. Rev.* 42, 292–304.
- Stafford, M., Warr, M., 1993. A reconceptualization of general and specific deterrence. *J. Res. Crime Delinquency* 30 (2), 123–135.
- Tay, R., 2005a. General and specific deterrent effects of traffic enforcement: do we have to catch offenders to reduce crashes? *J. Transport Econ. Policy* 39 (2), 209–223.
- Tay, R., 2005b. The effectiveness of enforcement and publicity campaigns on serious crashes involving young male drivers: are drink driving and speeding similar? *Accid. Anal. Prevent.* 37 (5), 922–929.
- Tay, R., 2005c. Drink driving enforcement and publicity campaigns: are the policy recommendations sensitive to model specifications? *Accid. Anal. Prevent.* 37 (2), 259–266.
- Tay, R., 2005d. Deterrent effects of drink driving enforcement: some evidence from New Zealand. *Int. J. Transport Econ.* 32 (1), 103–109.
- Tay, R., de Barros, A., 2006. Optimal Strategy for the Deployment of Intersection Safety Camera, report submitted to Alberta Transportation and Transport Canada.
- Tay, R., 2009. The effectiveness of automated and manned traffic enforcement. *Int. J. Sustain. Transport* 3 (3), 178–186.
- Tay, R., de Barros, A., 2009. Minimizing red light violations: how many cameras do we need for a given number of locations? *J. Transport. Safety Sec.* 1 (1), 18–31.
- Tay, R., 2010. Speed cameras: improving safety or raising revenue? *J. Transport Econ. Policy* 44 (2), 247–257.
- Tay, R., de Barros, A., 2011. Should traffic enforcement be unpredictable? The case of red light cameras in Edmonton. *Accid. Anal. Prevent.* 43, 955–961.
- WHO, 2003. World Report on Road Traffic Injury Prevention, World Health Organization, Geneva.

Satisfaction With Travel and the Relationship to Well-Being

Tommy Gärling*, Filip Fors Connolly[†], *Centre for Finance and Department of Economics, School of Business, Economics, and Law, University of Gothenburg, Göteborg, Sweden; [†]Department of Sociology, Umeå University, Umeå, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	177
Satisfaction With Travel	177
Independent of Travel Mode	178
Private Motorized Travel Modes	178
Public Motorized Travel Modes	179
Nonmotorized Travel Modes	179
Combined Travel Modes	179
The Relationship to Well-Being	180
Travel-Related Emotional Well-Being	180
Spill-Over Effects on Emotional Well-Being	180
Travel-Related Life Satisfaction	180
Conclusions	180
References	181
Further Reading	181

Introduction

Travel is central to almost all human activities. Everyday travel includes journeys at short distances to school or work, shopping outlets, various other facilities, and visits to relatives or friends. Less frequent longer journeys are made to holiday destinations away from home. Longer journeys are also made for business. The benefits to individuals and society are huge; however, the benefits need to be balanced against costs of environmental damages, accidents, and unproductive time.

Purpose of travel, travel time, and experiences during travel have bearings on the two questions raised in the following: What makes everyday travel satisfying or dissatisfying, and how does satisfaction or dissatisfaction with everyday travel influence well-being? In order to answer the first question, research has focused on the central features of travel illustrated in Fig. 1. Addressing the second question, research has implemented knowledge and measurement methods developed in well-being research.

The purpose of destination-oriented travel is associated with need, obligation, or desire to engage in some activity at the destination, either work, sustenance, or leisure, corresponding to the common classification in transport policymaking and planning.

When populations of interconnected people in urban areas grew and spread spatially, faster modes were necessary to not increase total travel times. Motorized travel resulted in longer travel distances but only marginally longer travel times. Walking differs from motorized travel in being slower and hence is referred to as a slow travel mode. Cycling is another slow travel mode. Together walking and cycling are labeled active travel modes since they in general require more physical activity than passive motorized travel.

Motorized travel modes are private or public. In private transportation, vehicles are not shared or shared with relatives or friends, in public transportation vehicles are primarily shared with strangers. This results in different forms of social interaction. Another difference is the degree to which different modes restrict travel in time and space, with consequences for congestions, accident risks, and delays.

Satisfaction With Travel

A traditional approach is to assess satisfaction with everyday travel based on utility-maximization theory. According to this theory, choices people make that maximize utility result in satisfaction with the outcome of their choices. Satisfaction with travel is therefore derived from observed choices. However, since the outcome of a chosen alternative is not always experienced to be satisfying, the assumed correspondence between utility of choice outcomes and satisfaction has been challenged (DeVos et al., 2016). Several complementary methods of self-report ratings of satisfaction with travel have therefore been developed and used, but few have been subject to adequate psychometric assessments. An exception is the multi-item Satisfaction with Travel Scale (STS) validated and used in several studies.

The STS retrospectively measure satisfaction with any form of travel. Analogous to self-report measures of well-being (see next section), it combines a cognitive-evaluation dimension (e.g., ratings on a scale from high to low travel quality) with two dimensions that assess mood (a less transient feeling state) during travel. The theoretical basis for the mood dimensions is the affect circumplex defined by the two orthogonal principal dimensions pleasure versus displeasure and activation (arousal) versus deactivation. Items

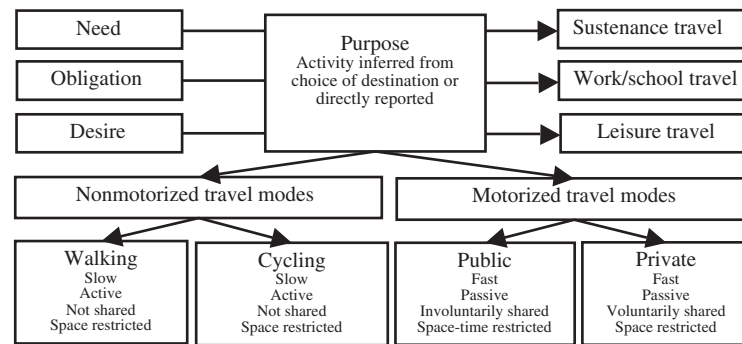


Figure 1 A classification of travel features. Source: Adapted from Gärling et al. (2019), with permission.

in the STS consist of validated rating scales that tap two dimensions oblique to the principal dimensions of the affect grid, one ranging from unpleasant deactivation to pleasant activation (e.g., ratings on a scale from tired to alert), and the other ranging from pleasant deactivation to unpleasant activation (e.g., ratings on a scale from calm to stressed). The three STS dimensions are distinct constructs positively correlated with each other and subsumed to a latent higher-order construct of overall satisfaction with travel. Measures of the dimensions have acceptable reliability, their relationships are invariant across primary travel modes, and they satisfactorily discriminate between the travel modes.

Although the object of measurements of satisfaction may be unspecified travel, it is more commonly satisfaction with a specified class of journeys or journey, or a specified travel mode for a specified class of journeys. The most frequent object is satisfaction with commutes to and from work by different travel modes (Chatterjee et al., 2020). Purpose, travel time, and experiences during travel are factors associated with travel mode that influence satisfaction with travel. In direct comparisons of travel modes satisfaction with the active commute modes of walking and cycling is higher than satisfaction with the passive commute mode car use, and satisfaction with car use is higher than satisfaction with the other passive commute mode public transport use. In the following, we first focus on aspects that are partly independent of travel mode, then aspects that are specific for the different travel modes.

Independent of Travel Mode

Some travel purposes have direct relationship to well-being, and such relationships carry over to satisfaction with travel for these purposes. An example is that satisfaction with travel to leisure destinations is higher than satisfaction with travel to work.

Longer travel times due to slow speed, long distances, interchanges, or delays decreases satisfaction with travel, and sometimes also too short travel times decrease satisfaction.

Satisfaction with travel is influenced by traveler characteristics. One is favorable or unfavorable pretravel attitude. Another example is that satisfaction increases with age.

Private Motorized Travel Modes

The spread of affluence made cars affordable to most people. As a result of owning a car, people could lead a life with which they were more satisfied, frequently preferring to live in low-density suburbs further away from their downtown workplace, occupying single-family homes with a garden. The downside is that car commuting to and from work increases accident risk, is stressful or boring, and time consuming.

The way the society today is spatially organized requires travel to different places for different purposes. Current travel demand thus goes beyond the home-to-work journey on weekdays, including as it does frequent multipurpose and multistop trips at different times during all week. For trips exceeding walking distance, the versatile car is then the travel mode that increases satisfaction with travel more than other modes. That the car is usually the fastest travel mode increases satisfaction further.

The status and identity associated with being a car owner is another source of satisfaction that has roots in the very early days of the car. Today the ubiquity of the car has led to it being perceived as an everyday tool, and anyone not owning a car is perceived as being less well-off.

The purpose for using a car is sometimes not activity participation; car use is thus not purely a demand derived from the purpose of travel. This is the case when people travel for the sake of traveling or travel to a destination via a more scenic route. Affective (e.g., enjoyment) and symbolic (e.g., showing others) factors are important additional motives for car use, in particular for men and younger people. Excess driving furthermore results from valuing driving itself. Yet, driving is also sometimes stressful or boring.

The various benefits of the car that have led to its frequent use may result in the development of a car-use habit. Important conditions for the development of a habitual behavior is that a person is able and motivated to repeat the behavior, and that the behavior is rewarded such that it is repeated a sufficient number of times in a stable context. These conditions frequently apply to car use. A car-use habit implies that information processing is changed such that the intention to travel to a specified destination automatically activates a car choice.

Table 1 Quality attributes in public transport

Category	Attribute	Definition
Perceived	Comfort	How comfortable the journey is regarding access to seating, noise levels, driver handling, and air conditioning
	Safety/Security	How safe from traffic accidents passengers feel during the journey as well as personal safety
	Convenience	How simple the service is to use and how much it adds to ease of mobility.
Physical	Aesthetics	Aesthetic appeal of vehicles, stations, and wait areas
	Reliability	How closely the actual service matches the route timetable
	Frequency	How frequently the service operates during different hours
	Speed	The time spent traveling from origin to destination
	Accessibility	The degree to which the service is available to as many as possible
	Price	The monetary cost of travel
	Ease of transfers/interchanges	How simple transport connections are including wait times
	Vehicle conditions	The physical and mechanical conditions of vehicles including frequency of breakdowns
	Information provision	How accurate the information is about routes and interchanges

Source: Adapted from Redman et al. (2013), with permission.

Public Motorized Travel Modes

Historically, public transport (the term “public transport” is predominantly used in Europe, Japan, and Australia, the interchangeable terms “mass transit” or “public transit” are used in North America and Southeast Asia (primarily in China, Hong Kong, and Singapore)) services (railways, tramways, and buses) originated in Europe in connection with the 19th century industrialization, urbanization, and separation of residences from workplaces. The industrialization both increased demand for fast mass transportation at a large scale of people and goods as well as providing the technical and economic means of meeting this demand.

Public transport is a more economical and environmentally friendly motorized travel mode than the car. However, at least in suburban and rural areas, it is less versatile and slower. These are main reasons why satisfaction with public transport is lower than satisfaction with car transport. In densely populated urban areas, public transport (in particular metro and commuter trains) is still attractive.

A number of attributes of public transport services that increase user satisfaction have been identified. A list of these attributes is shown in **Table 1**. Perceived attributes refer to users’ experiences directly measured by observing their responses. If, for instance, seats are replaced in vehicles to increase comfort, this may be assessed by on-board interviews. Physical attributes are measured without involvement of users. For instance, the effect of a revised timetable may be assessed by measures of punctuality, assuming it would reflect changes in user perceptions of reliability.

Satisfaction with travel does not solely depend on cognitive evaluations of features. Critical incidents generally defined as an event that stands out from what is normal by being perceived to be either negative or positive but the definition may include additional, more specific criteria depending on research field) during travel appear to influence satisfaction, partly due to their emotional impacts. In public transport critical incidents are primarily negative, for instance inappropriate treatment by employees, lack of punctuality, or crowding.

In many places around the world, great effort has been made to improve public transport services. This has paid off in increased patronage. Success factors are active collaboration between different stakeholders, high-quality bus systems, visual and distinct company brands, and integration of travel modes.

Nonmotorized Travel Modes

Walking and cycling declined when motorized travel increased. An upsurge has however been observed during recent years, possibly with the exception for children, low-income earners, and the elderly. Satisfaction with travel is also high for nonmotorized travel modes. Although traveler self-selection or unavailability of motorized alternatives are possible reasons, intrinsic attributes of the active travel modes are important. One factor standing out is that nonmotorized travel helps people engage in physical activity that they believe have positive health effects. Satisfaction is also influenced by other factors, including experienced excitement and variation in the environment, company of others, enjoyment of the activity (walking or cycling) itself, scenery, and weather.

Combined Travel Modes

Many journeys consist of several stages. Walking to and from the bus stop, train station, or car park before and after using the primary travel mode are common examples. Public transport journeys may also require changes from bus to train or from one bus or train line to another. How is satisfaction with the journey related to satisfaction with the stages of the journey? Is overall satisfaction influenced by satisfaction with each stage or do stages with high and low satisfaction have more influence? Does satisfaction with the end of the journey (reaching the destination) have a stronger influence? For commutes to and from work, integration is made by

averaging satisfaction with the stages, weighted by their duration. This may be different for journeys that are less routine. Duration-weighting is still consistent with that satisfaction is influenced by travel time.

The Relationship to Well-Being

Well-being has in the past frequently been equated with material living standard, indicated by, for instance, income or GDP per capita. In recent years the focus has, as recommended by the United Nations and OECD, changed to subjective experiences of well-being, referred to as life satisfaction and emotional well-being. In particular in wealthy countries, the roles of other than material conditions are increasingly becoming recognized. In addition to socioeconomic factors, social relations, health, and personality traits are important determinants. Furthermore, both life satisfaction and emotional well-being are strongly associated with the fulfillment of human needs. Satisfaction with everyday activities also play important roles. Travel enables people to engage in out-of-home activities that may increase their life satisfaction, at the same time as travel itself is an everyday activity that may influence emotional well-being (Ettema et al., 2010).

Travel-Related Emotional Well-Being

Emotional well-being is measured retrospectively by asking people to report the frequency (or duration) and intensity of positive and negative feelings recalled for a specified time interval. The measure captures the prevailing mood as well as transient feelings in response to external events. In the Day Reconstruction Method (DRM) people are asked to recall important activities during a preceding day and indicate for each activity how they felt when performing it. Still another method is to ask people in real time to report how they feel at the moment. If reported retrospectively, or momentarily at several points in time, a measure of emotional well-being is constructed as the difference between the frequency of positive and negative feelings.

Positive and negative critical incidents during travel influence emotional well-being measured retrospectively by the mood scales in the STS. Users of cars and active travel modes have more positive moods than users of public transport. Longer travel times make public transport users more bored and less relaxed. Activities performed during travel may counteract negative mood effects.

As assessed by the DRM, negative feelings are associated with commuting. This is consistent with that commuting causes stress, as revealed by performance measures. But everyday travel only seems to account for a few percent of the variance in aggregated measures of daily mood. Satisfaction with travel still has direct effects on weekly mood as well as indirect effects through feelings associated with frequent travel-dependent out-of-home activities during the week.

Spill-Over Effects on Emotional Well-Being

Long car commutes have stress after-effects which negatively affect activities at work and home. Yet, momentary measures of mood before and after morning commutes to work, by active and passive travel modes, indicate that mood is relatively stable. Neither do changes in mood after travel carry over to mood at the workplace.

Travel-Related Life Satisfaction

Life satisfaction is usually assessed by means of cognitive judgments, either a single judgment on a ladder ranging from 0 (worst possible life) to 10 (best possible life), or a multi-item measure referred to as the Satisfaction with Life Scale including questions such as "I am satisfied with my life" and "in most ways my life is close to my ideal".

Commuting has potentially negative effects because it limits time for physical and social leisure activities that are important for life satisfaction. Restricting travel may also cause social exclusion which has a strong negative association with life satisfaction.

Other travel than work commutes have direct effects on life satisfaction, either because travel is in itself a valued activity or because travel enables other valued activities (e.g., social visits).

Conclusions

Recent years have seen an accumulation of knowledge of what influences satisfaction with destination-oriented everyday travel and how it influences emotional well-being and life satisfaction. Perhaps the most important contribution to date is still the development and validation of feasible methods for measuring satisfaction and well-being, which otherwise would appear to be elusive constructs and therefore neglected in transport planning. However, this does not imply that methods that for decades has been the core in previous research and application should be replaced. Rather, new and old methods are possible to combine to the benefit of an increasing methodological sophistication and rigor.

Additional research is needed to further develop online methods for measuring satisfaction with travel. The aim should be to increase knowledge of how satisfaction is related to features of travel. This is needed for improving design and management of transportation systems such that satisfaction with travel services increases. Everyday travel (e.g., commuting) is not in general an activity that in itself promotes emotional well-being and life satisfaction. Increasing travel-related satisfaction is still essential since it would reduce the negative consequences of everyday travel for emotional well-being and life satisfaction.

References

- Chatterjee, K., Chng, S., Clark, B., Davis, A., De Vos, J., et al., 2020. Commuting and wellbeing: a critical overview of the literature with implications for policy and future research. *Transport Rev.* 40, 5–34.
- De Vos, J., Mokhtarian, P.L., Schwanen, T., Van Acker, V., Witlox, F., 2016. Travel mode choice and travel satisfaction: bridging the gap between decision utility and experienced utility. *Transportation* 43, 771–796.
- Ettema, D., Gärling, T., Olsson, L.E., Friman, M., 2010. Out-of-home activities, daily travel and subjective well-being. *Transport. Res. Part A* 44, 723–732.
- Gärling, T., Bamberg, S., Friman, M., 2019. The role of attitude in choice of travel, satisfaction with travel, and change to sustainable travel. In: Albarracín, D., Johnson, B.T. (Eds.), *Handbook of Attitudes: Applications*. Routledge, London, pp. 562–586.
- Redman, L., Friman, M., Gärling, T., Hartig, T., 2013. Quality attributes of public transport that attract car users: a research review. *Transport Policy* 25, 119–127.

Further Reading

- Ben-Akiva, M., Lehman, S.R., 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.
- Chng, S., Abraham, C., White, M.P., Hoffmann, C., Skippon, S., 2018. Psychological theories of car use: an integrative review and conceptual framework. *J. Environ. Psychol.* 55, 23–33.
- De Vos, J., Schwanen, T., Van Acker, V., Witlox, F., 2013. Travel and subjective well-being: a focus on findings, methods and future research needs. *Transport Rev.* 33, 421–442.
- Diener, E., Oishi, S., Tay, L. (Eds.), 2018. *Handbook of Well-Being*. DEF Publishers, Salt Lake City, UT, Published online (doi:nobascholar.com).
- Diener, E., Lucas, R.E., Schimmack, U., Helliwell, J.F., 2009. *Well-Being for Public Policy*. Oxford University Press, New York.
- Friman, M., Ettema, D., Olsson, L.E. (Eds.), *Quality of Life and Daily Travel*. Springer, New York.
- Friman, M., Gärling, T., 2017. Moving towards sustainable consumption: a psychological perspective on improvement of public transport. In: Jansson-Boyd, C.V., Zawisza, M. (Eds.), *Routledge International Handbook of Consumer Psychology*. Routledge, London, pp. 524–541.
- Gärling, T., Ettema, D., Friman, M. (Eds.), *Handbook of Sustainable Travel*. Springer Science, Dordrecht, The Netherlands.
- Novaco, R.W., Gonzales, O.I., 2008. Commuting and well-being. In: Amichai-Hamburger, Y. (Ed.), *Technology and Well-Being*. Cambridge University Press, New York, pp. 174–205.
- McFadden, D., 2001. Disaggregate behavioral travel demand's RUM side – A 30 years retrospective. In: Hensher, D.A. (Ed.), *Travel Behavior Research*. Elsevier, Amsterdam, pp. 17–63.
- Mokhtarian, P.L., 2019. Subjective well-being and travel: retrospect and prospect. *Transportation* 46, 493–513.
- Oliver, R.L., 2010. *Satisfaction: A Behavioral Perspective on the Consumer*. Sharpe, New York.

Electromobility: History, Definitions and an Overview of Psychological Research on a Sustainable Mobility System

Josef F. Krems, Isabel KreiBig, Chemnitz University of Technology, Chemnitz, Germany

© 2021 Elsevier Ltd. All rights reserved.

Introduction	182
History	182
What is it? Terminology	183
Psychological Issues	183
Range Stress	183
Eco-Driving	183
Noise	184
Acceptance	184
E-Bikes: Usage, Safety, and Mode Choice	184
Summary	185
See Also	185
Relevant Websites	185
References	185
Further Reading	186

Introduction

Electromobility is considered a key technology for sustainable mobility in the future. Currently in many countries the number of electric powered cars is increasing and infrastructure for charging is increasingly being developed. The adoption of electric powered cars essentially depends not only on technological progress like battery technology or carbon-light weight, but largely on psychological features such as acceptance that, aside from pricing, mostly depends on range and charging infrastructure.

Besides electric powered cars, electric bikes (e-bikes or pedelecs) are becoming a viable option for transport, particularly in urban areas. Psychological research is addressing questions of mode choice, route selection, and safety due to higher vehicle speeds and weights.

History

The history of electric powered vehicles began in 1881 when Gustave Trouvé presented the first e-car, the Trouvé Tricycle, a couple years ahead of Carl Benz's combustion engine powered car. The first German e-vehicle was the "Flockenwagen," introduced in 1888, which was based on a four-wheel carriage and is considered the first electric powered passenger car. Around 1900 many more electric powered cars existed than those with an integrated combustion engine (ICE). For example, it is estimated that more than 90% of cabs in New York were electric powered. Famous examples are the "Lohner-Porsche" (1900), "Columbia" (1904), and "Siemens Viktoria" (1905). The Lohner-Porsche was also the first example of a hybrid electric vehicle (HEV) that used four hub-mounted electric motors to directly control the wheels. The situation changed in the first half of the 20th century, particularly after World War I. The number of ICE cars steadily increased due to an extended range and easier handling, whereas the number of electric powered cars diminished. In the late 20th century, ICE vehicles were vastly more common. Only few concept cars were around, like the "BMW 1602" (1969) or the "VW Citystromer" (1982). In the first decade of the 21st century, the situation completely changed. Due to problems with air-pollution, noise, and climate change in general e-mobility awoke from hibernation. Regions like California sharply restricted the emission of CO₂ and forced manufacturers to develop alternative engines for vehicles. Nowadays, all major manufactures offer a broad variety of different EV types.

In 2010, the European Union had around 1500 newly registered e-cars. In 2019, this number increased to around 463,000, which is about 300-times more, but still significantly fewer than the number of newly registered cars with traditional ICE. World-wide, e-car registration in 2019 is estimated to be around 2,200,000 (up from around 200,000 in 2013), according to the German Association of the Automotive Industry (VDA). It is further estimated that the number of electric models available will surpass 200 in 2021 and that European EV production will reach about 4 million vehicles by 2025.

In 2010, about 300 charging stations were in use in Germany. Ten years later, around 30,000 stations are working. This is around 100 times more with a growth rate increasing exponentially.

What is it? Terminology

An electric car (e-car or e-vehicle, EV) or battery electric vehicle (BEV) is a vehicle propelled by one or more electric motors, using energy stored in rechargeable batteries.

A hybrid electric vehicle (HEV) is a vehicle containing an electric drive train combined with an ICE. A famous and bestselling example is the “Toyota Prius.”

A plug-in-hybrid electric vehicle (PHEV) contains a battery that can be recharged by being plugging into an external source of electric power or by an on-board engine or generator. A famous example is the “Toyota Prius PHV.”

Range-extended electric vehicles (REEV) or extended-range electric vehicles (EREV) are vehicles that include an auxiliary power unit (APU) known as range extender, but are mostly an ICE. The unit drives an electric generator which charges the battery to increase the vehicle’s range.

Recuperation or regenerative braking is a procedure to convert kinetic energy into electric power to be used immediately or later by storing the energy in the battery. It is an important mechanism to extend the range and is a main characteristic of EVs.

An electric bicycle (e-bike) is a bicycle with an integrated electric motor that supports the driver up to a certain speed. Regarding performance and driver support, there is a wide variety of types ranging from low-powered e-bikes, called pedelec (PEDal ELectric Cycle, usually up to 25 km/h with up to 0.25 kW power) to S-pedelecs (up to 45 km/h), which are classified differently from country to country (bicycle, moped, motor-scooter).

Psychological Issues

The impact of electromobility on traffic and transport with regard to acceptance and purchase decisions highly depends on psychological factors. In several user studies, most including predominately “early adopters” (specific sample—middle-aged men, high education, high income) or rather inexperienced users—the following topics were identified as most important:

Range Stress

Less range, compared to ICE cars, was considered an essential disadvantage of EVs more than 100 years ago and is still a main barrier towards BEV adoption (Li et al., 2017). In the last 10 years, the increased battery capacity has been a major research focus resulting in cars on the market currently that have a range of around 500 km. However, this is still less than a comparable conventional car. A second issue is charging infrastructure and duration. Compared to a conventional gas refill stop, charging batteries, even with fast-charging technology, still takes much longer. Also, the number of charging stations is still far below the number of regular gas-stations. However, large field studies (e.g., The MINI-E studies, Krems et al., 2011; Pearre et al., 2011) show that around 95% of all trips are shorter than 200 km and that users were satisfied with a range of around 160 km. Franke et al. (2012, 2017) distinguish between different range levels. “Cycle range” is the “objective” trip length given a certain state-of-charge (SOC) of the batteries. “Competent range” is the range possible if users operate the vehicle while maximizing energy saving. “Performant range” refers to the actual driver behavior, which is influenced by (lack of) skills, habits, and motivation. “Comfortable range” refers to the SOC percentage considered necessary for drivers to feel comfortable. Meanwhile in psychological research, the term “range anxiety” was shifted to “range stress,” which captures the driver’s psychological status in a much more appropriate way. In their model, Franke et al. (2012) assume that recognizing a low remaining range for a certain trip constitutes a conflict between available resources and eternal demands. This person–environment imbalance bears similarities with common stress definitions, particularly with the transaction theory of stress (Lazarus and Folkman, 1984). This model emphasizes that reducing stressors (e.g., increasing range) is only one coping strategy supplemented by individual appraisal processes and social support. Range is primarily a psychological barrier and must be considered together with mere increasing nominal battery capacity.

Eco-Driving

With regard to the limited range, energy saving strategies during driving (eco-driving strategies) are specifically relevant when driving BEVs compared to conventional vehicles. In terms of the range model (Franke et al., 2012), eco-driving strategies are considered appropriate to cope with range stress (i.e., in low on charge situations) and therefore potentially reduce negative feelings regarding limited range (Rauh et al., 2017). Eco-driving strategies comprise aspects of route choice (e.g., elevation profile), driving maneuvers (e.g., moderate acceleration) and driving-related actions such as the usage of auxiliary functions (e.g., air conditioning). Although most eco-driving strategies for BEVs are similar to those known from conventional vehicles, there are some EV-specifics, most relevant amongst them the optimal usage of regenerative braking (Helmbrecht et al., 2014; Jensen et al., 2020). Drivers’ growing experience has been studied to gain knowledge on BEV eco-driving strategies (Neumann et al., 2015) and their implementation in critical range situations (Günther et al., 2019). Additionally, experiencing a critical range situation combined with informing drivers about efficient eco-driving strategies has shown to improve even inexperienced BEV drivers’ eco-driving knowledge and acceptance (Günther et al., 2017). Similar to other pro-environmental behaviors, motivational aspects influence eco-driving (Neumann, 2016). Therefore, effective interventions such as persuasive elements in in-vehicle instrumentation (Günther et al., 2020) could enhance comfortable range and also strengthen the environmental benefit of EVs.

Noise

Compared to ICE vehicles, the BEV's engine noise emission is quite low resulting in "silent" driving at least during low speeds (<30 kph), where the tyre noise emission is negligible. This lack of noise is described twofold in the literature. On the one hand, BEV drivers and passengers clearly appreciated it, enhancing their driving pleasure (Bühler et al., 2014; Cocron et al., 2010; Daramy-Williams et al., 2019). Moreover, it offers a potential environmental benefit by reducing noise pollution, for instance in big cities (del Carmen Pardo-Ferreira et al., 2020). On the other hand the lack of noise emission and the corresponding later perception of BEVs at low speeds compared to conventional vehicles (Stelling-Kończak et al., 2015) has been discussed as a potential hazard, especially for vulnerable road users (VRUs; e.g., pedestrians or cyclists). Focusing on driving pleasure, user studies revealed that the lack of noise is considered as one of the main advantages of BEVs from a driver perspective (Bühler et al., 2014; Cocron et al., 2010), and, in turn, could even increase the chances for BEV acceptance (Bühler et al., 2014). Furthermore, experience effects show that initially positive ratings of silent driving even increased after BEV driving for a certain time period (Bühler et al., 2014; Carroll, 2010). Addressing safety concerns, several studies showed a relatively low perceived risk of traffic incidents reported by BEV drivers, which decreased after BEV driving for a longer period (Cocron and Krems, 2013). Additionally, experienced BEV drivers reported to be aware of potential hazards and therefore adopted more cautious driving styles (Cocron and Krems, 2013). There are also hints that inexperienced BEV drivers show such awareness and compensatory behavior (Cocron et al., 2014). Looking at research analyzing crash statistics, it is still unclear whether there is a higher incident rate for electric compared to conventional vehicles with VRU involvement, mainly due to missing exposure data and the few registered accidents (Stelling-Kończak et al., 2015; Pardo-Ferreira, 2020). Early in the discussion about the safety issue of low noise emission, additional vehicle sounds, so called Acoustic Vehicle Alerting Systems (AVAS), have been proposed as possible solution. Although drivers do not view this strategy as necessary or appreciated (Cocron and Krems, 2013; del Carmen Pardo-Ferreira et al., 2020), many countries (e.g., Japan, US, EU; e.g., European Commission, 2014) have introduced respective regulations demanding the implementation of additional sound to BEVs. Analyzing possible positive and negative AVAS consequences will certainly be a challenge for upcoming research on BEV acoustics.

Acceptance

Searching for strategies to enhance BEV adoption and reduce subjective barriers in this regard, BEV acceptance has been extensively investigated throughout the last two decades (Coffman et al., 2017; Li et al., 2017; Rezvani et al., 2015). According to Schade and Schlag (2003) acceptance is defined by "respondents' attitudes including their behavioral reactions after the introduction of a measure" (p. 47). Following this definition, BEV acceptance involves an attitudinal as well as a behavioral component, both of which have been addressed in research on this topic (compare also Li et al., 2017; Rezvani et al., 2015). A common theoretical background investigating BEV acceptance, including the willingness to buy such a vehicle, is provided by the theory of planned behavior (TPB; Ajzen, 1991), often extended by further factors such as experiencing BEVs or the assessment of BEV features (e.g., Li et al., 2017; Moons and De Pelsmacker, 2012; Rezvani et al., 2015; Schmalfuß et al., 2017;). In addition, a variety of psychological constructs and theories exist providing the basis for consumer BEV adoption research (see, e.g., Li et al., 2017; Rezvani et al., 2015 for further discussion). Examples describing and predicting BEV adoption and related intentions include the diffusion of innovation theory (Rogers, 2010; e.g., Peters and Dütschke, 2014), normative theories (e.g., value-belief-norm; Stern, 2000) but also theories associated with symbols and self-identity (e.g., Noppers et al., 2014; Schuitema et al., 2013). Reviewing the literature on BEV acceptance, Li et al. (2017) regard attitudes towards BEVs to be the most effective predictor of BEV acceptance. In addition, other psychological aspects like experiencing BEV features (e.g., low noise emission, acceleration, regenerative braking) and consumers' expectations as well as symbolic attributes have shown to be important factors influencing BEV acceptance (Li et al., 2017; Rezvani et al., 2015). Provision of direct BEV experience appeared to considerably impact the assessment of BEV features as well as attitudes (e.g., Li et al., 2017; Schmalfuß et al., 2017). This finding underlines the importance of the provision of hands-on experience for studying BEV acceptance and simultaneously narrows the external validity of studies with participants who had never previously experienced a BEV and its features, such as surveys with inexperienced participants (compare also Coffman et al., 2017; Li et al., 2017). The BEV market penetration is still far behind expectations and targets, possibly because of the limited driving range combined with potentially long charging durations as well as battery costs and the relatively high purchase price as the most important barriers (Coffman et al., 2017; Li et al., 2017; Liao et al., 2017). Additionally, factors like consumer characteristics (e.g., gender, age, education) as well as social norms or values have been seen to influence BEV acceptance (Coffman et al., 2017; Li et al., 2017).

E-Bikes: Usage, Safety, and Mode Choice

In recent years, e-bikes have become increasingly popular on our roads. In Germany, more than a third of sold bicycles are equipped with an electric motor. Across Europe, the number of sold e-bikes increased from 0.5 million in 2009 to 3.6 million in 2018 (Statista Research Department, 2020). Pedelects can help to reduce traffic congestion in cities and air-pollutions. Further, they are a way to obtain physical exercise for many people. Compared to conventional bicycles, pedelecs offer the opportunity to maintain speed with less effort regardless of hilly routes, individual physical fitness or trip length, therefore overcoming one of the main barriers of conventional bicycles (Fishman and Cherry, 2016). Due to their electric support, e-bikes are especially attractive for older age groups—in Germany, for example, half of all pedelec trips are made by persons 60 years and older (Nobis and Kuhnimhof, 2018).

This has mainly positive impacts, such as enhancing physical fitness and activity in older age, but also increases the risk of higher accident rates. Gehlert et al. (2018) found higher proportions of elderly involved in pedelec compared to bicycle accidents. The authors also observed more pedelec accidents involving inappropriate speeds, specifically for elderly cyclists. In general, e-bike accidents appeared to be more severe (more fatalities and severe injuries). Comparing conventional and e-bike usage, Schleinitz et al. (2017) found similar trip lengths and purposes for both vehicle types. Furthermore, e-bikes are driven only slightly faster than bicycles, with the exception of S-pedelecs, which are driven at considerably higher speeds. E-bikes offer a substantial environmental benefit—depending on country and study, many car or bus trips were replaced by riding an e-bike (Fishman and Cherry, 2016). Cargo e-bikes, enabling the user to carry large and heavy weights, are one solution discussed to reduce air pollution caused by private transport. In this regard, a Norwegian study showed decreased car use and increased cycling trips per day, as well as increased trip distances for families equipped with cargo e-bikes compared to a control group (Bjørnarå et al., 2019). Currently, some initiatives exist to establish sharing concepts for cargo e-bikes in cities to make this promising (but expensive) travel mode available to a large user group and thus reach their full potential.

Summary

In times of climate change, electric mobility is seen as one promising solution to reduce CO₂-emissions in the traffic sector given alternative energy usage as well as appropriate production and recycling conditions. The term electromobility comprises various vehicles propelled by electric drive — ranging from hybrid and fully electric cars to busses as well as e-bikes and cargo e-bikes — and includes infrastructure necessary for charging and managing the vehicles.

Initially, psychological research mainly concentrated on electric cars, focusing on BEV acceptance and users' interaction with specific features such as limited range, noise-reduced driving and regenerative braking. Research has shown a complex set of factors and interactions impacting the BEV adoption process ranging from BEV-related to individual and societal facets, which might be met by an equally comprehensive set of actions addressing many of the specific facets to push this process, if intended.

During the past few years, electric two-wheelers have become increasingly popular. Given their low weight and the potential to replace the car as a mode of transport, these vehicles even enhance the environmental benefit and additionally bear the potential of a more active lifestyle. Given their specifics (e.g., predominantly elderly cyclists, different speed profile, high acquisition costs—at least for cargo e-bikes), research, and policy are required to provide specific knowledge and implement appropriate measures to take the chance on this environmental friendly mode of transportation.

See Also

Future Mobility: Attitudes and Perceptions; Electric Mobility; Electric Vehicles; Sharing: Attitudes and Perceptions; Transport Modes and Sustainability; The Future of Road Transport; Shared Mobility: An Overview of Definitions, Current Practices, and Its Relationship to Mobility on Demand and Mobility as a Service; Sustainable Mobility Paths; Transport Modes and an Aging Society; Electric Vehicles and Health; E-Bikes and Health; Bicycle Sharing / Bikesharing; Transport Modes and Cities; Driving Behavior and Skills; Road Safety; Bicycle Sharing

Relevant Websites

<https://www.prognos.com/presse/news/detailansicht/1899/7eb78fa11ec616ec3d738ca9b11c2e/>
<https://www.transportenvironment.org/sites/te/files/publications/01%202020%20Draft%20TE%20Infrastructure%20Report%20Final.pdf>
<https://evadoption.com/category/battery-range/>
<https://www.bdew.de/energie/elektromobilitaet-dossier/energiewirtschaft-baut-ladeinfrastruktur-auf/>
https://en.wikipedia.org/wiki/Electric_vehicle
<https://www.statista.com/statistics/276036/unit-sales-e-bikes-europe/>

References

- Ajzen, I., 1991. The theory of planned behaviour. *Organ. Behav. Hum. Decis. Process.* 50 (2), 179–211.
- Bjørnarå, H.B., Berntsen, S., Te Velde, S., Fyhri, A., Deforche, B., Andersen, L.B., Bere, E., 2019. From cars to bikes—the effect of an intervention providing access to different bike types: a randomized controlled trial. *PLoS One* 14 (7), e0219304.
- Bühler, F., Cocron, P., Neumann, I., Franke, T., Krems, J.F., 2014. Is EV experience related to EV acceptance? Results from a German field study. *Transp. Res. Part F Traffic Psychol. Behav.* 25, 34–49.
- Carroll, S., 2010. *The Smart Move Trial: Description and Initial Results*. Cenex, London.
- Cocron, P., Bühler, F., Franke, T., Neumann, I., Krems, J.F., 2010. The silence of electric vehicles – blessing or curse? *Proc. 90th Ann. Meeting Transport. Res. Board, Washington, DC*.
- Cocron, P., Bachl, V., Fröh, L., Koch, I., Krems, J.F., 2014. Hazard detection in noise-related incidents – the role of driving experience with battery electric vehicles. *Accid. Anal. Prev.* 73, 380–391.

- Cocron, P., Krems, J.F., 2013. Driver perceptions of the safety implications of quiet electric vehicles. *Accid. Anal. Prev.* 58, 122–131.
- Coffman, M., Bernstein, P., Wee, S., 2017. Electric vehicles revisited: a review of factors that affect adoption. *Transp. Rev.* 37 (1), 79–93.
- Daramy-Williams, E., Anable, J., Grant-Muller, S., 2019. A systematic review of the evidence on plug-in electric vehicle user experience. *Transport. Res. Part D Transp. Environ.* 71, 22–36.
- del Carmen Pardo-Ferreira, M., Rubio-Romero, J.C., Galindo-Reyes, F.C., López-Arquillos, A., 2020. Work-related road safety: the impact of the low noise levels produced by electric vehicles according to experienced drivers. *Saf. Sci.* 121, 580–588.
- European Commission, 2014. Commission welcomes parliament vote on decreasing vehicle noise, IP/14/363, 02/04/2014. Retrieved from: http://europa.eu/rapid/press-release_IP-14-363_en.htm.
- Fishman, E., Cherry, C., 2016. E-bikes in the mainstream: reviewing a decade of research. *Transport Rev.* 36 (1), 72–91.
- Franke, T., Neumann, I., Bühler, F., Cocron, P., Krems, J.F., 2012. Experiencing range in an electric vehicle: understanding psychological barriers. *Appl. Psychol.* 61 (3), 368–391.
- Franke, T., Günther, M., Trantow, M., Krems, J.F., 2017. Does this range suit me? Range satisfaction of battery electric vehicle users. *Appl. Ergon.* 65, 191–199.
- Gehlert, T., Krölling, S., Schreiber, M., Schleinitz, K., 2018. Accident analysis and comparison of bicycles and pedelecs Framing the third cycling century. *Bridg. Gap Res. Pract.* 77–85.
- Günther, M., Kacperski, C., Krems, J.F., 2020. Can electric vehicle drivers be persuaded to eco-drive? A field study of feedback, gamification and financial rewards in Germany. *Energ. Res. Soc. Sci.* 63, 101407.
- Günther, M., Rauh, N., Krems, J.F., 2017. Conducting a study to investigate eco-driving strategies with battery electric vehicles—a multiple method approach. *Transp. Res. Procedia* 25, 2242–2256.
- Günther, M., Rauh, N., Krems, J.F., 2019. How driving experience and consumption related information influences eco-driving with battery electric vehicles—results from a field study. *Transp. Res. Part F Traffic Psychol. Behav.* 62, 435–450.
- Helmbrecht, M., Olaverri-Monreal, C., Bengler, K., Vilimek, R., Keinath, A., 2014. How electric vehicles affect driving behavioral patterns. *IEEE Intell. Transp. Syst. Mag.* 6 (3), 22–32.
- Jensen, A.F., Rasmussen, Th.K., Prato, C.G., 2020. A route choice model for capturing driver preferences when driving electric and conventional vehicles. *Sustainability* 12, 1149.
- Krems, J.F., Cocron, P., Franke, T., Dielmann, B., Keinath, A., Schwalm, M., 2011. Electromobility in megacities: will mobility patterns change? In: VDI-FVT. (Eds.), *Der Fahrer im 21. Jahrhundert*. VDI-Verlag, Düsseldorf, pp. 263–272.
- Lazarus, R.D., Folkman, S., 1984. *Stress, Appraisal, and Coping*. Springer, New York.
- Li, W., Long, R., Chen, H., Geng, J., 2017. A review of factors influencing consumer intentions to adopt battery electric vehicles. *Renew. Sust. Energ. Rev.* 78, 318–328.
- Liao, F., Molin, E., van Wee, B., 2017. Consumer preferences for electric vehicles: a literature review. *Transp. Rev.* 37 (3), 252–275.
- Möller, K., 2018. Frühe Elektromobilität auf der Straße: Kultur und Technik. In: Horstmann, Th., Döring (Hrsg.), P. (Eds.), *Zeiten der Elektromobilität*. VDE Verlag, Berlin, pp. 7–38.
- Moons, I., De Pelsmacker, P., 2012. Emotions as determinants of electric car usage intention. *J. Mark. Manage.* 28 (3–4), 195–237.
- Neumann, I., 2016. Electric vehicle energy efficiency – implications for the user interface, driving task and motivational aspects. Doctor Thesis, Chemnitz Technical University of Technology.
- Neumann, I., Franke, T., Cocron, P., Bühler, F., Krems, J.F., 2015. Eco-driving strategies in battery electric vehicle use—how do drivers adapt over time? *IET Intell. Transp. Syst.* 9 (7), 746–753.
- Nobis, C., Kuhnimhof, T., 2018. *Mobilität in Deutschland—MiD, Ergebnisbericht*.
- Noppers, E.H., Keizer, K., Bolderdijk, J.W., Steg, L., 2014. The adoption of sustainable innovations: driven by symbolic and environmental motives. *Global Environ. Change* 25, 52–62.
- Pearre, N.S., Kempton, W., Guensler, R.L., Elango, V.V., 2011. Electric vehicles: how much range is required for a day's driving? *Transp. Res. Part C Emerg. Technol.* 19 (6), 1171–1184.
- Peters, A., Dütschke, E., 2014. How do consumers perceive electric vehicles? A comparison of German consumer groups. *J. Environ. Policy Plan.* (September), 1–19.
- Rauh, N., Franke, T., Krems, J.F., 2017. First-time experience of critical range situations in BEV use and the positive effect of coping information. *Transp. Res. Part F: Traffic Psychol. Behav.* 44, 30–41.
- Rezvani, Z., Jansson, J., Bodin, J., 2015. Advances in consumer electric vehicle adoption research: a review and research agenda. *Transp. Res. Part D: Transp. Environ.* 34, 122–136.
- Rogers, E.M., 2010. *Diffusion of Innovations*. Simon and Schuster, NY.
- Schade, J., Schlag, B., 2003. Acceptability of urban transport pricing strategies. *Transp. Res. Part F: Traffic Psychol. Behav.* 6, 45–61.
- Schleinitz, K., Petzoldt, T., Franke-Bartholdt, L., Krems, J.F., Gehlert, T., 2017. The German naturalistic cycling study – comparing cycling speed of riders of different e-bikes and conventional bicycles. *Saf. Sci.* 92, 290–297.
- Schuitema, G., Anable, J., Skippon, S., Kinnear, N., 2013. The role of instrumental, hedonic and symbolic attributes in the intention to adopt electric vehicles. *Transp. Res. A: Policy Pract.* 48, 39–49.
- Schmalfuß, F., Mühl, K., Krems, J.F., 2017. Direct experience with battery electric vehicles (BEVs) matters when evaluating vehicle attributes, attitude and purchase intention. *Transp. Res. Part F: Traffic Psychol. Behav.* 46, 47–69.
- Stelling-Kończak, A., Hagenzieker, M., van Wee, B.V., 2015. Traffic sounds and cycling safety: the use of electronic devices by cyclists and the quietness of hybrid and electric cars. *Transp. Rev.* 35 (4), 422–444.
- Stern, P.C., 2000. Toward a coherent theory of environmentally significant behavior. *J. Soc. Issues* 56 (3), 407–424.

Further Reading

- Coffman, M., Bernstein, P., Wee, S., 2017. Electric vehicles revisited: a review of factors that affect adoption. *Transp. Rev.* 37 (1), 79–93.
- Daramy-Williams, E., Anable, J., Grant-Muller, S., 2019. A systematic review of the evidence on plug-in electric vehicle user experience. *Transp. Res. Part D Transp. Environ.* 71, 22–36.
- Franke, T., Neumann, I., Bühler, F., Cocron, P., Krems, J.F., 2012. Experiencing range in an electric vehicle: Understanding psychological barriers. *Appl. Psychol.* 61 (3), 368–391.
- Li, W., Long, R., Chen, H., Geng, J., 2017. A review of factors influencing consumer intentions to adopt battery electric vehicles. *Renew. Sust. Energ. Rev.* 78, 318–328.
- Rezvani, Z., Jansson, J., Bodin, J., 2015. Advances in consumer electric vehicle adoption research: a review and research agenda. *Transport. Res. Part D: Transport Environ.* 34, 122–136.

Sharing: Attitudes and Perceptions

Yi Wen, Christopher R Cherry, Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, TN, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	187
What is Shared Mobility?	187
Carsharing	188
Ridesourcing	189
Bikesharing	189
Scooter Sharing	190
A Disruption to Disruptive Transportation	191
References	191

Introduction

Shared mobility services such as ridesourcing, carsharing, and micromobility vehicle (e.g., bicycle and e-scooter) sharing are reshaping many aspects of our transportation system. In the last decade, large private sector investments met with rapid growth of technological innovation. Technology like location-based smartphone services and onboard vehicle telematics have allowed previous low-tech models of shared mobility to scale to billions of trips in a short period of time. Low tech models (e.g., conventional taxis, free-floating bikesharing) have been replaced or forced to adapt to the emergence of new competitors. Cities have been forced to revisit how they operate and prioritize the public right of way. The public has adapted to changing norms on sharing space, resources, and information. The rapid change in attitudes and perceptions of shared mobility options is becoming more understood as researchers have conducted both quantitative and qualitative studies focusing on the adoption of different sharing models. Some shared mobility services appear to be more popular than other, seemingly similar, technologies. Shared scooters, for example, have a much higher adoption rate compared to bikesharing. In their first year of launch, there were 38.5 million scooter trips in the United States in 2018, about the same number as bikesharing trips that had been growing annually for a decade ([National Association of City Transportation Officials, 2019](#)). The rate of adoption of shared scooters was faster than the rate of adoption of ridesourcing ([Bruce, 2018](#)). While this shift may be driven largely by technological advancements and innovative business models, how people perceive these shared mobility options plays an important role. This chapter presents a narrative on this dynamic evolution of perceptions and attitudes and summarizes what is known and unknown about how consumers perceive this evolving portfolio of transportation services. It also describes how different models and technologies have evolved and comments on what this may mean for transportation and society in the future. We will start this chapter by understanding what shared mobility is.

What is Shared Mobility?

Shared mobility, or shared-use mobility systems, are broadly defined as shared used of travel modes such as a vehicle or a bicycle that provide users with short-term access on an as-needed basis ([SAE International, 2018](#)). This definition includes many well-known shared modes such conventional taxis and public transportation, but also maturing models like bikesharing and carsharing and emerging models like e-scooter sharing. Collectively, transportation systems that fall under shared mobility are often grouped under mobility on demand (MOD) or mobility as a service (MaaS) categories. This contrasts with recent transportation systems that are either public (e.g., a public-sector-owned system providing transportation services through transit agencies) or private (e.g., private automobile ownership). MaaS frameworks break from those compartments, where the public sector, private sector, or individuals can provide market-driven transportation services through shared mobility platforms that are often enabled by smartphone applications or other technology to overcome barriers to previous shared mobility schemes.

Shared mobility terminology is evolving rapidly. The SAE International Standard J3163 ([SAE International, 2018](#)) aims to clarify terminology through standards development. From a consumer perspective, that document broadly describes shared mobility in three dimensions: (1) the travel modes, (2) the enabling mobility applications, and (3) the service and operational models. The travel modes refer to a variety of shared mobility options such as motor vehicles, bicycles, and e-scooters that are available to meet diverse mobility needs. The enabling mobility applications are smartphone applications that provide users with travel information (e.g., navigation, schedule) and/or assistance in using shared mobility services if needed. For example, users can check the availability of for-hire taxis in the area and reserve one through such applications. Last, differences in service and operational models capture the variations of the relationship between a user and the shared mobility service provider. For example, such relationship can be membership-based, for-hire, peer-to-peer (P2P), or come in other forms.

From an attitude and perception perspective, different models have different potential barriers. For example, some modes require interaction with a driver (who might be a professional driver or a contract driver, e.g., taxi or ridesourcing), some also require

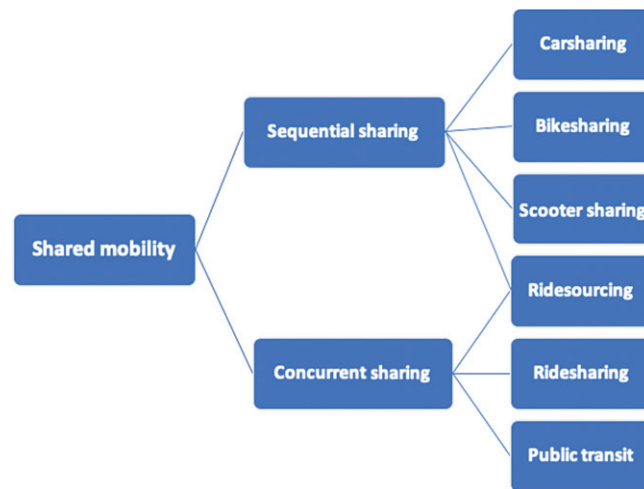


Figure 1 Classification of selected shared mobility travel modes. *Source: Adapted from Shared-Use Mobility Center and SAE International.*

interaction with passengers (e.g., pooled ridesourcing), and others require no interaction with others (e.g., bikesharing and carsharing). Shared mobility modes can be used concurrently by users sharing rides at the same time, or they can be used sequentially by users taking turns sharing vehicles (Shared-Use Mobility Center, 2020). Following this dichotomy in Fig. 1 we categorize the most important shared mobility modes. Ridesourcing bridges concurrent and sequential shared mobility models since ridesourcing can be pooled. Two of the concurrent modes, ridesharing (e.g., carpool) and conventional public transit are well documented in the literature. This chapter focuses on attitudes and perceptions of contemporary and emergent shared mobility modes; carsharing, ridesourcing, bikesharing, and scooter sharing.

Of these four emergent shared mobility modes, there are common characteristics that tie together their users. Shared mobility users tend to be younger, more often in carless or car-light households, and have higher incomes and education levels than the general population (Alonso-Almeida, 2019; Martin and Shaheen, 2011; Sikder, 2019; ter Schure et al., 2012). Some of that rider base may be drawn from the geographic distribution of services. Shared mode operators and regulators often make deliberate efforts to expand services to improve equitable access (Howland et al., 2017; Pan et al., 2020).

Carsharing

Carsharing is among the pioneers in the shared mobility industry. A carsharing program provides a group of users the access to a fleet of vehicles whose parking, insurance, fuel, and maintenance are usually covered by the service provider (SAE International, 2018). Depending on where users are instructed to return the shared vehicle, a carsharing system can be categorized as a one-way (also known as free-floating), round-trip, or a hybrid system. For example, a round-trip system requires users to return the vehicle to the same place where it is picked up and the user is responsible for the vehicle for the entire duration of their round-trip. A one-way system allows users to end their reservation wherever they end a (one-way) trip. One-way systems support more one-way trips like transit tours where carsharing supports the first- or last-mile. Variations of carsharing models also come from different types of ownership models of the shared vehicles. In cases where the shared vehicles are owned by some users and shared among other users within the system, known as a peer-to-peer (P2P) carsharing model. The P2P model exists in other shared mobility systems as well and we will revisit this concept in the following sections. It is of important to acknowledge this difference in shared vehicle ownership as it can affect how consumers perceive and use these services.

In North America, carsharing services first emerged in residential neighborhoods before diversifying into other market segments such as university campuses (Shaheen et al., 2009). Because carsharing services provide on-demand vehicle access at a lower price compared to owning a vehicle, it can be appealing to the carless population, who are likely the early adopters and the targeted consumers of such services. Carsharing users tend to reduce their overall car miles traveled, reduce transit use, while increasing their walk and bicycling trips (Cervero et al., 2007; Martin and Shaheen, 2011).

A number of factors are known to influence mode choice and the decision to use carsharing. Among them, high cost and long travel time to an accessible shared car are barriers to grow carsharing membership (Zhou and Kockelman, 2011). In P2P carsharing, where individuals provide short term rentals of personal cars, monetary benefits are the major positive contributor, while liability issues top as the primary deterrent for both vehicle renters and owners (Ballús-Armet et al., 2014).

Burkhardt and Millard-Ball (2006) found convenience and environmental reasons are consistently the main drivers of carsharing use. More recent studies confirm those findings and also identify value-seeking and lifestyle as another two important motives of using carsharing (Schaefer, 2013). Burkhardt and Dütschke (2019) used survey data from Germany and showed that perceived compatibility with daily life is the biggest motivator for both current and potential carsharing users. However, Hartl et al. (2018) suggested that though important, the role of sustainability can be secondary and have a marginally small influence on carsharing adoption.

With rapid development in vehicle technologies and carsharing business models, perceptions and attitudes of carsharing are likely to change in the near future with uncertain implications. Electrification and automation are two recognized trends in the automobile industry. An early study among electric vehicle (EV) carsharing program in Seoul, Korea found that the EV carsharing program is unlikely to reduce car ownership (Kim et al., 2015). Yoon et al. (2017) found that Beijing respondents in a stated preference survey did not value EVs over conventional fuels. However, Lavieri et al. (2017) suggested that early adopters of shared autonomous vehicles are likely to be younger population with a higher education level, who share same similarities with the early carsharing adopters. Besides carsharing, these changes may have meaningful impacts on another vehicle-based shared mobility mode—ridesourcing.

Ridesourcing

Ridesourcing, also sometimes referred to as ridehailing, refers to a set of services that match drivers and passengers through smartphone applications for pre-arranged or on-demand travel at an agreed price (SAE International, 2018). This is different than “ridesharing” as the latter term denotes pairing services of drivers and passengers based on origins and destinations, either formally or informally. For “ridesourcing,” this relationship relies on the driver providing a commercial service to the passenger, whereas for “ridesharing” the primary purpose is to share a journey. The most representative ridesourcing services are for-hire services such as Uber and Lyft while carpooling and vanpooling are the typical forms of ridesharing.

The commercial launch of Uber in 2009, now the Silicon Valley giant, marks the beginning of the ridesourcing industry. Given its short history, the popularity of ridesourcing is remarkable as more than one in every five American have used such services (Young and Farber, 2019). Yet, with the success of ridesourcing, their impacts on the transportation system inevitably become more apparent. Pan et al. (2020) analyzed taxi and ridesourcing trips in New York City and concluded that ridesourcing services are more equitable than the traditional taxi services, and the overall equality level of the taxi industry has been improved with the advent of ridesourcing services. However, the negative impacts are also concerning. There is growing evidence that ridesourcing may increase vehicle-kilometers traveled (VKT) and congestion (Erhardt et al., 2019; Henao and Marshall, 2019; Tirachini and Gomez-Lobo, 2020), which may diminish some of the motivation from environmentally conscious travelers to use the service. These findings draw into question the environmental sustainability or congestion merits of ridesourcing.

Existing findings regarding attitudes and perceptions of ridesourcing do not largely mirror the literature on carsharing despite some similarities. Technological familiarity is an important attribute of ridesourcing users (Alemi et al., 2019; Lavieri and Bhat, 2019). However, environmental motivations do not appear to be a focus of research in ridesourcing. Rather, many studies across the globe pointed out that concerns over trust, privacy, and personal safety are the major challenges for ridesourcing adoption (Alemi et al., 2019; Lavieri and Bhat, 2019; Ma et al., 2019).

Pooled (or shared) rides provide an interesting perspective to examine this attitudes and perceptions question. By pooling riders in a shared vehicle, this branch of shared mobility service can be promising in reducing congestion, VKT, and emissions induced by individual ridesourcing trips, and even by all other vehicles on the road. However, literature suggest that pooled ride users choose such service mainly because of reasons such as time and cost savings (Alonso-Almeida, 2019), and security of a seat (Tirachini et al., 2020) other than sustainability motives. Furthermore, Alonso-Almeida (2019) pointed out that car-centric individuals highly prefer individual rides over pooled rides. This is likely associated with the long-standing images of freedom and social status that owning and driving a car may represent in some cultures. In addition, Morris et al. (2020) noted that drivers providing on-demand pooled rides are less satisfied compared to individual rides for reasons such as unfair compensation with extra work in routing, pick-ups, and drop-offs. These findings may partly explain why ridesharing (e.g., carpooling, vanpooling) and similar ideas has struggled in many markets.

The widespread adoption of ridesourcing among the younger generation is partly attributed to the use of smartphones and mobile applications. Being technology-savvy has been seen as a driving factor both in carsharing and ridesourcing adoptions. Yet, these are not the only shared mobility options that are benefiting from the fast-developing technology. Bikesharing has its own story to tell.

Bikesharing

Bikesharing is the on-demand shared mobility service that provides access of bicycles to users at a variety of locations (SAE International, 2018). There are different versions of bikesharing systems depending on types of bicycles and the service models. For example, the shared bikes can be regular bicycles, electric bicycles (e-bikes), or fleets can include a combination of both. Likewise, the bikesharing system can be station-based, dockless, or a hybrid consisting of both options. Users can either choose to pay the fares based on the number of trips taken or sign up for a membership that typically provides unlimited trips to the system with a subscription fee.

Bikesharing systems have undergone three major generational updates since the prototype in Amsterdam in 1965 (DeMaio, 2009). The contemporary model is known as the fourth generation of bikesharing. This generation features the use of mobile technology, electronic payment, and location tracking. For example, users can find, reserve and unlock a shared bike through a smart phone application, pay through debit or credit cards, and return the bike through the same app. The fourth generation

bikesharing systems started as station-based systems. For a station-based system, users can only pick up and drop off the shared bikes from an existing list of docked stations within the system boundary though the drop-off station does not need to be the same station where the shared bike is picked up. Note that this is a different concept compared to round-trip idea in carsharing discussed previously. During this period, new ideas were experimented with bikesharing. For example, [Langford et al. \(2013\)](#) piloted a bikesharing system integrating e-bikes on the University of Tennessee campus.

Early bikesharing systems were free-floating dockless systems. In the late 2000s, the industry converged on docked systems to leverage technology at the station-level and improve operational efficiency. Still, some operators pushed dockless bikesharing systems that relied heavily on on-board telematics and smartphone integration (e.g., social bicycles in the United States and call-a-bike in Germany). Yet, the most important recent game changer in the bikesharing industry to date arrived in 2015. Two Chinese start-ups MoBike and Ofo started the launch of their dockless bikesharing models around the same time in selected Chinese markets. This new model received rapid popularity and was soon exported overseas at significant scale. In the United States, for example, number of dockless bikesharing trips went up from 1.4 million in 2017, to 9 million in 2018, reflecting a jump of 4% share of bikesharing trips to almost 20% ([National Association of City Transportation Officials, 2018, 2019](#)). As the name implies, the biggest difference between a dockless bikesharing system and a station-based system is that the dockless model has no docking stations for the shared bikes. Instead, users of the system can pick up and drop off a shared bike within the system boundary in the essentially same mechanism as a free-floating carsharing system. Because there is no station hardware associated with the operation, the cost of using the dockless systems are presumably cheaper to the operator and user, which may in part contribute to their early popularity.

Due to the distinct differences between station-based and dockless bikesharing systems, the user demographics and how respective user groups perceive and use bikesharing are likely different. [Fishman \(2016\)](#) examined and summarized the literature of bikesharing before the launch of dockless models and found that the major motivation to use bikesharing is convenience for station-based bikesharing users. In contrast to station-based systems, dockless bikesharing users revealed that while having a higher education level, a higher income level no longer appears to be as significant as for station-based bikesharing use ([Chen et al., 2020](#); [Kaviti et al., 2019](#)). As for the adoption challenges, previous studies identified safety of riding, mandated helmet use, and the sign-up process could be barriers for station-based bikesharing ([Buck et al., 2013](#); [Fishman, 2016](#)), while pro-walking attitudes discourage the use of dockless bikesharing ([Chen et al., 2020](#)).

The experience of dockless bikesharing systems in user penetration and adoption have impacts and implications beyond bikesharing. Notably, this may have suggested a viable business model for the shared mobility companies. In fact, as the dockless bikesharing market appeared to hit a point of market saturation, new companies gradually turned to shared scooters as the new fuel for ridership and revenues. This new wave of scooter sharing and investments is changing and even challenging the common understanding of transportation in both positive and negative senses.

Scooter Sharing

Scooter sharing is the on-demand shared mobility service that provides access of scooters to users at a variety of locations ([SAE International, 2018](#)). As of August 2020, most available shared scooters around the world are lightweight electric standing scooters ([SAE International, 2019](#)). In many ways, a scooter sharing system is similar to a dockless bikesharing system. However, unlike the modern bikesharing systems that started as station-based systems, scooter sharing launched and remains predominantly dockless. This, coupled with other features of this new mobility option, creates an array of opportunities and challenges. In summary, the two main concerns about shared dockless bikes and scooters are improper parking, and risky or illegal riding behaviors such as riding on the sidewalk.

Concerning improper parking, the results of observational studies do not seem to agree with the public perception. For example, a survey on scooter sharing in Bloomington, Indiana reported that more than 73% respondents prefer designated parking areas for scooters, and close to 60% respondents call for banning scooters from sidewalks, both suggesting that there is a perceived issue with shared scooters occupying public space. Yet, the literature in this area suggest something different. Three observational studies across seven cities found that scooters that are improperly parked only accounts for 0.8%–6% of observations ([Brown et al., 2020](#); [Fang et al., 2018](#); [James et al., 2019](#)). These studies imply that improper parking may not be as serious as how it is perceived.

As for risky and illegal riding behaviors such as sidewalk riding, it appears that the lack of adequate and infrastructure facilitates the tension between shared scooter riders and other road users. For example, even in Denver, one of the more bicycle friendly cities in North America, 34% of the respondents of a survey indicated that while walking, a scooter hit or almost hit them, and 19% of the respondents stated a similar experience when driving ([Denver Department of Public Works, 2019](#)). In the same survey, 28% of the respondents indicated they would use shared scooters more if there were more dedicated infrastructure for riding. [Portland Bureau of Transportation \(2019\)](#) conducted both an observational study and a survey regarding use of scooter sharing in the city. They noted that scooter sharing riders would use bike lanes on low-speed streets, and incrementally higher percentages of sidewalk riding is observed on streets with higher speed limits. Meanwhile, the Portland study revealed that when a neighborhood low-traffic and low-speed street network is available, no users ride on the sidewalk, even without dedicated bicycle infrastructure. And, only 8% riders use the sidewalk with a protected bike lane available. People who are walking, biking, or riding a scooter are among the least protected road users, and many of the them are inexperienced as bicyclists. It is understandable that they choose the infrastructure type that they feel most comfortable with. Without a proper riding infrastructure such as a protected bike lane, sidewalks that are

designed for pedestrian become over-designed for scooter riders while roadways that are designed to move vehicles are under-designed for scooters. Sidewalk riding decreases the comfort level both for pedestrian and scooter riders, both of whom value their own safety first. It is the lack of infrastructure that force and intensify the interactions among different groups of road users, and influence how they perceive this shared mobility service.

Regarding the shared scooter rider demographics, they are similar to other shared modes (young, educated, higher income). There is observational evidence that scooter sharing attracts more women and racial minorities than other shared modes (Portland Bureau of Transportation, 2019; San Francisco Municipal Transportation Agency, 2019; Sanders et al., 2020). Despite the observational evidence that scooter sharing attracts a more diverse group of users, many non-riders perceive scooter sharing negatively. Besides, the broader demographic and increased number of riders can expose the less skilled population to a higher risk level. Indeed, some evidence points to scooter riders being unfamiliar with rules of the road and driver and rider expectations, calling for a need for more intuitive and low stress transportation networks to support novice riders. Relevant guidelines and changes should be developed and enforced to meet these emerging needs for transportation. However, an unprecedented pandemic in 2019 may point us in a new direction for shared mobility.

A Disruption to Disruptive Transportation

Shared mobility has increasingly disrupted conventional transportation systems. While critics might point to substituted low-impact modes (e.g., transit, non-motorized modes) as evidence of the negative aspects of shared mobility, many of the benefits revolve around providing a diversity of services that can provide the appropriate technology for specific trips and reduce reliance on space- and energy-inefficient modes (e.g., single occupant cars) for all trips. Travelers can consume appropriately scaled transportation services that meet their immediate needs, which is positive for constrained urban environments. This disruption has been enabled by technology, but also cultural shifts that result in attitudes and behaviors that foster increasing adoption of shared modes for many demographics.

The COVID-19 pandemic has upended all aspects of society, including shared mobility systems. Transit system ridership immediately plunged by more than 80% in most markets, following slow rebounds as economies opened (Transit, 2020). Shared mobility services have also been dramatically impacted (StreetLight, 2020; Thigpen, 2020) and it is uncertain how attitudes toward infection risk, job status, and rider demographics will affect rider behavior. People are riding bicycles and walking for transportation more often, there is some evidence that bikesharing and shared scooter riders have different behaviors compared to transit riders (Campbell and Brakewood, 2017; The Light Electric Vehicle Education & Research Initiative, 2020). However, at the time of this writing, there is a tremendous level of uncertainty related to the role of shared mobility in future transportation systems. Resilient transportation systems require service diversity. Shared mobility systems have increased diversity of transportation services more in the last decade than any other period in recent history. Shared mobility is essential for the continued economic viability of cities. As entire economies evolve to increase resilience against shocks, shared mobility will continue to adapt to assure continued mobility and access to urban destinations.

References

- Alemi, F., Circella, G., Mokhtarian, P., Handy, S., 2019. What drives the use of ridehailing in California? Ordered probit models of the usage frequency of Uber and Lyft. *Transport. Res. C: Emerg. Technol.* 102, 233–248, doi:10.1016/j.trc.2018.12.016.
- Alonso-Almeida, M.d. M., 2019. Carsharing: another gender issue? Drivers of carsharing usage among women and relationship to perceived value. *Travel Behav. Soc.* 17, 36–45, doi:10.1016/j.tbs.2019.06.003.
- Ballús-Armet, I., Shaheen, S.A., Clonts, K., Weinzimmer, D., 2014. Peer-to-peer carsharing: exploring public perception and market characteristics in the San Francisco bay area, California. *Transport. Res. Rec.* 2416 (1), 27–36, doi:10.3141/2416-04.
- Brown, A., Klein, N.J., Thigpen, C., Williams, N., 2020. Impeding access: the frequency and characteristics of improper scooter, bike, and car parking. *Transport. Res. Interdiscipl. Perspect.* 4, 100099, doi:10.1016/j.trip.2020.100099.
- Bruce, O., 2018. The Bull Case for Micromobility. Retrieved from <https://micromobility.io/blog/2018/12/7/the-bull-case-for-micromobility>.
- Buck, D., Buehler, R., Happ, P., Rawls, B., Chung, P., Borecki, N., 2013. Are bikeshare users different from regular cyclists?: A first look at short-term users, annual members, and area cyclists in the Washington, D.C., Region. *Transportation Research Record* 2387 (1), 112–119, doi:10.3141/2387-13.
- Burghard, U., Dütschke, E., 2019. Who wants shared mobility? Lessons from early adopters and mainstream drivers on electric carsharing in Germany. *Transport. Res. D: Transport Environ.* 71, 96–109, doi:10.1016/j.trd.2018.11.011.
- Burkhardt, J.E., Millard-Ball, A., 2006. Who is attracted to carsharing? *Transport. Res. Rec.* 1986 (1), 98–105, doi:10.1177/0361198106198600113.
- Campbell, K.B., Brakewood, C., 2017. Sharing riders: how bikesharing impacts bus ridership in New York City. *Transport. Res. A: Policy Pract.* 100, 264–282.
- Cervero, R., Golub, A., Nee, B., 2007. City CarShare: longer-term travel demand and car ownership impacts. *Transport. Res. Rec.* 1992 (1), 70–80, doi:10.3141/1992-09.
- Chen, Z., van Lierop, D., Ettema, D., 2020. Exploring dockless bikeshare usage: a case study of Beijing, China. *Sustainability* 12 (3), 1238.
- DeMaio, P., 2009. Bike-sharing: history, impacts, models of provision, and future. *J. Public Transport.* 12 (4), 3.
- Denver Department of Public Works, 2019. Denver Dockless Mobility Program Pilot Interim Report. Retrieved from <https://www.denvergov.org/content/dam/denvergov/Portals/705/documents/permits/Denver-dockless-mobility-pilot-update-Feb2019.pdf>.
- Erhardt, G.D., Roy, S., Cooper, D., Sana, B., Chen, M., Castiglione, J., 2019. Do transportation network companies decrease or increase congestion? *Sci. Adv.* 5 (5), eaau2670, doi:10.1126/sciadv.aau2670.
- Fang, K., Agrawal, A.W., Steele, J., Hunter, J.J., Hooper, A.M., 2018. Where Do Riders Park Dockless, Shared Electric Scooters? Findings from San Jose, California.
- Fishman, E., 2016. Bikeshare: a review of recent literature. *Transport Rev.* 36 (1), 92–113, doi:10.1080/01441647.2015.1033036.

- Hartl, B., Sabitzer, T., Hofmann, E., Penz, E., 2018. "Sustainability is a nice bonus": the role of sustainability in carsharing from a consumer perspective. *J. Cleaner Prod.* 202, 88–100, doi:10.1016/j.jclepro.2018.08.138.
- Henao, A., Marshall, W.E., 2019. The impact of ride-hailing on vehicle miles traveled. *Transportation* 46 (6), 2173–2194.
- Howland, S., McNeil, N., Broach, J., Rankins, K., MacArthur, J., Dill, J., 2017. Current efforts to make bikeshare more equitable: survey of system owners and operators. *Transport. Res. Rec.* 2662 (1), 160–167, doi:10.3141/2662-18.
- James, O., Swiderski, J., Hicks, J., Teoman, D., Buehler, R., 2019. Pedestrians and e-scooters: an initial look at e-scooter parking and perceptions by riders and non-riders. *Sustainability* 11 (20), 5591.
- Kaviti, S., Venigalla, M.M., Lucas, K., 2019. Travel behavior and price preferences of bikesharing members and casual users: a capital bikeshare perspective. *Travel Behav. Soc.* 15, 133–145, doi:10.1016/j.tbs.2019.02.004.
- Kim, D., Ko, J., Park, Y., 2015. Factors affecting electric vehicle sharing program participants' attitudes about car ownership and program participation. *Transport. Res. D: Transport Environ.* 36, 96–106, doi:10.1016/j.trd.2015.02.009.
- Langford, B.C., Cherry, C., Yoon, T., Worley, S., Smith, D., 2013. North America's first e-bikeshare: a year of experience. *Transport. Res. Rec.* 2387 (1), 120–128, doi:10.3141/2387-14.
- Lavieri, P.S., Bhat, C.R., 2019. Investigating objective and subjective factors influencing the adoption, frequency, and characteristics of ride-hailing trips. *Transport. Res. C: Emerg. Technol.* 105, 100–125, doi:10.1016/j.trc.2019.05.037.
- Lavieri, P.S., Garikapati, V.M., Bhat, C.R., Pendyala, R.M., Astroza, S., Dias, F.F., 2017. Modeling individual preferences for ownership and sharing of autonomous vehicle technologies. *Transportat. Res. Rec.* 2665 (1), 1–10, doi:10.3141/2665-01.
- Ma, L., Zhang, X., Ding, X., Wang, G., 2019. Risk perception and intention to discontinue use of ride-hailing services in China: taking the example of DiDi Chuxing. *Transport. Res. F: Traffic Psychol. Behav.* 66, 459–470, doi:10.1016/j.trf.2019.09.021.
- Martin, E., Shaheen, S., 2011. The impact of carsharing on public transit and non-motorized travel: an exploration of North American carsharing survey data. *Energies* 4 (11), 2094–2114.
- Morris, E.A., Zhou, Y., Brown, A.E., Khan, S.M., Derochers, J.L., Campbell, H., Chowdhury, M., 2020. Are drivers cool with pool? Driver attitudes towards the shared TNC services UberPool and Lyft Shared. *Transport Policy* 94, 123–138, doi:10.1016/j.tranpol.2020.04.019.
- National Association of City Transportation Officials, 2018. Bike Share in the U.S.: 2017. Retrieved from <https://nacto.org/wp-content/uploads/2018/05/NACTO-Bike-Share-2017.pdf>.
- National Association of City Transportation Officials, 2019. Shared Micromobility in the U.S.: 2018. Retrieved from https://nacto.org/wp-content/uploads/2019/04/NACTO_Shared-Micromobility-in-2018_Web.pdf.
- Pan, R., Yang, H., Xie, K., Wen, Y., 2020. Exploring the equity of traditional and ride-hailing taxi services during peak hours. *Transport. Res. Rec.*, doi:10.1177/0361198120928338.
- Portland Bureau of Transportation, 2019. 2018 E-scooter Findings Report. Retrieved from <https://www.portlandoregon.gov/transportation/article/709719>.
- SAE International, 2018. J3163 Taxonomy and Definitions for Terms Related to Shared Mobility and Enabling Technologies. SAE International.
- SAE International, 2019. J3194 Taxonomy and Classification of Powered Micromobility Vehicles.
- San Francisco Municipal Transportation Agency, 2019. Powered Scooter Share Mid-Pilot Evaluation. Retrieved from https://www.sfmta.com/sites/default/files/reports-and-documents/2019/04/powered_scooter_share_mid-pilot_evaluation_final.pdf.
- Sanders, R.L., Branion-Calles, M., Nelson, T.A., 2020. To scoot or not to scoot: findings from a recent survey about the benefits and barriers of using E-scooters for riders and non-riders. *Transport. Res. A: Policy Pract.* 139, 217–227, doi:10.1016/j.tra.2020.07.009.
- Schaefers, T., 2013. Exploring carsharing usage motives: a hierarchical means-end chain analysis. *Transport. Res. A: Policy Pract.* 47, 69–77, doi:10.1016/j.tra.2012.10.024.
- Shaheen, S.A., Cohen, A.P., Chung, M.S., 2009. North American carsharing: 10-year retrospective. *Transport. Res. Rec.* 2110 (1), 35–44, doi:10.3141/2110-05.
- Shared-Use Mobility Center, 2020. What is Shared Mobility? Retrieved from <https://sharedusemobilitycenter.org/what-is-shared-mobility/>.
- Sikder, S., 2019. Who uses ride-hailing services in the United States? *Transport. Res. Rec.* 2673 (12), 40–54, doi:10.1177/0361198119859302.
- StreetLight, 2020. Understand the impact of COVID-19 on traffic, travel patterns, toll revenues and more. Retrieved from <https://www.streetlightdata.com/covid-transportation-metrics/>.
- ter Schure, J., Napolitan, F., Hutchinson, R., 2012. Cumulative impacts of carsharing and unbundled parking on vehicle ownership and mode choice. *Transport. Res. Rec.* 2319 (1), 96–104, doi:10.3141/2319-11.
- The Light Electric Vehicle Education & Research Initiative, 2020. COVID-19 Impacts on Transit, Bike Share, and Scooter Share Systems. Retrieved from <https://www.levresearch.com/coronavirus.html>.
- Thigpen, C., 2020. Rethinking Travel in the era of COVID-19 Survey Findings and Implications for Urban Transportation. Retrieved from.
- Tirachini, A., Chaniotakis, E., Abouelela, M., Antoniou, C., 2020. The sustainability of shared mobility: can a platform for shared rides reduce motorized traffic in cities? *Transport. Res. C: Emerg. Technol.* 117, 102707, doi:10.1016/j.trc.2020.102707.
- Tirachini, A., Gomez-Lobo, A., 2020. Does ride-hailing increase or decrease vehicle kilometers traveled (VKT)? A simulation approach for Santiago de Chile. *Int. J. Sustain. Transport.* 14 (3), 187–204, doi:10.1080/15568318.2018.1539146.
- Transit, 2020. How coronavirus is disrupting public transit. Retrieved from <https://transitapp.com/coronavirus>.
- Yoon, T., Cherry, C.R., Jones, L.R., 2017. One-way and round-trip carsharing: a stated preference experiment in Beijing. *Transport. Res. D: Transport Environ.* 53, 102–114, doi:10.1016/j.trd.2017.04.009.
- Young, M., Farber, S., 2019. The who, why, and when of Uber and other ride-hailing trips: an examination of a large sample household travel survey. *Transport. Res. A: Policy Pract.* 119, 383–392, doi:10.1016/j.tra.2018.11.018.
- Zhou, B., Kockelman, K.M., 2011. Opportunities for and impacts of carsharing: a survey of the Austin, Texas Market. *Int. J. Sustain. Transport.* 5 (3), 135–152, doi:10.1080/15568311003717181.

Women's Travel Patterns, Attitudes, and Constraints Around the World

Sandra Rosenbloom, The University of Texas at Austin, Austin, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	193
Starting Early; Gender Differences in Children's Travel	194
Women's Complicated Travel Patterns	194
Women's Mode Choice	195
Personal Security	196
Women and Acceptance of New Transportation Technology	197
Traffic Safety	198
The Industrial World	198
Developing Nations	199
Conclusions and Policy Implications	200
Biography	201
References	201
Further Reading	202

Introduction

We have known for decades that women have different transportation behavior, attitudes, and safety outcomes than men in both the developed and developing world. Analysts have continued to find major gender differences in travel and crash patterns even when controlling for important socio-economic factors that influence behavior such as age, household structure, life cycle, and employment status. These gender differences in travel behavior start early in people's lives and have endured long after travel behavior scholars predicted women and men's travel patterns would converge. They arise from a complicated web of remarkably persistent economic and educational factors, and, cultural, societal, and religious norms about women's role and rights to the city that substantially constrain their activities and choices even in the most industrial nations.

Indeed, some aspects of women's travel behavior have come to resemble men's as more women have obtained salaried employment outside the home as well as drivers' licenses and cars—and men have assumed more childcare and household duties. But converging trends in trip purpose and mode should not obscure the many gender differences which remain, differences which are far starker in the Global South, and act as serious constraints on women's ability to care for themselves and their families everywhere. These constraints are magnified by the failure of many societal institutions to respond to inherent physical differences in the safety and security needs of women and men—including vehicles that are not constructed to protect women in crashes or the constant sexual harassment that faces women in almost every mode of travel everywhere, but particularly the Global South. Nor should we ignore the fact that there are major but far less studied differences in travel patterns and needs between groups of women that often break along racial and ethnic lines. Finally, it is important to question why socio-demographic factors like income inequality and discrimination in labor and housing markets, factors associated with some differences between the sexes in travel behavior, so disproportionately impact women to begin with.

All transportation differences between women and men, and even between groups of women, are not indicative of deficiencies to be addressed or problems to be overcome of course. They can reflect personal attitudes and preferences. Yet these all these differences should be reflected in transportation policy and planning efforts to create transportation systems that respond appropriately to a variety of user needs and preferences, or "transportation markets." These differences assume even more importance when considering the need to support sustainable transportation choices when so many deficiencies in the transportation systems push women into increasing personal vehicle use. In addition many experts call for different approaches to transportation planning and modeling and project assessments by explicitly considering differential impacts on women and men.

The sections below address:

- gender travel differences that begin in childhood
- women's complicated travel patterns
- women's mode choices
- personal security
- women and new technology
- traffic safety
- conclusions and policy suggestions

Starting Early; Gender Differences in Children's Travel

More than four decades of research shows that parents often refuse to allow little girls to travel independently, or walk or cycle, without direct adult supervision, even as they allow comparable little boys to do so (McMillan et al. 2006; Yeung et al., 2018). Most studies show that girls are less likely to engage in active transport (walking and cycling) and more likely to be driven or accompanied to outside activities than comparable boys. Kepper et al. (2020) found that parents of female adolescents imposed more restrictions on their daughters' outdoor play in all types of neighborhoods. Colley and Buliung (2016) found that in Toronto girls were more likely to be driven to primary school than comparable boys and less likely to walk and that those patterns intensified over a 25-year period. Larouche et al. (2019) found that girls in New Brunswick (Canada) had lower odds of engaging in active transportation to school (cycling, walking) than comparable boys, especially in poor weather. Larsen et al. (2018) found that boys had higher odds of walking to school than girls in grades 5 and 6 in Toronto. A Welsh study found that girls were less likely to walk to secondary school than boys (Potoglou and Arslangulova, 2019). Some studies, however, did not find gender differences in some aspects of children's travel (e.g., Simons et al., 2018; Thigpen and Handy, 2018), but their methods and focus may explain these differential findings.

Parents' attitudes train girls to be more fearful of outdoor environments, discourage their independent travel, and accustom them to wait to be driven or accompanied as opposed to providing their own mobility (Rosenbloom and Plessis-Fraisard, 2011). It also puts a greater burden on parents, and usually the mother, to provide transportation for daughters. These patterns may well have life-long consequences as girls become parents; Aibar Solano et al. (2018) found, for example, that mothers were less likely to allow children to walk for a variety of purposes than were fathers in comparable households.

We have more limited information on children's travel in developing economies. A 2019 Iranian study found that girls who lived in a household with a car were less likely to walk to school for short trips than comparable boys (Hatamzadeh et al., 2017). The authors concluded that policymakers must address gender differences to increase rates of walking to school (in contrast to an earlier study in Tehran that did not find gender differences in active transportation to school Samimi and Ermagun, 2013).

Given the body of research that shows important gender differences among young children it is interesting to note that a significant share of studies of children's travel, particularly those focused on active school travel, do not separately analyze girls and boys' travel behavior nor their possibly differential response to various incentives designed to encourage active transportation to school (such as Safe Routes to School Programs).

Women's Complicated Travel Patterns

Women and men have had different travel patterns than one another for decades. Even in countries devoted to equality of the sexes, like the Nordic nations, comparable women and men have different attitudes and preferences about their family and childcare obligations which play out in their travel patterns. Early studies which found important gender differences in travel did not always control for salient variables; indeed some current studies still continue to compare all men to all women which obscures important socio-demographic factors like whether the women are employed outside the home and for how many hours per week, and the presence and age of children. Yet most recent studies are far more sophisticated, controlling for crucial variables—and they continue to show that women and men in industrial nations, even those in the paid labor force, have very different travel patterns because women are required to, or feel compelled to, spend more time on childcare and household responsibilities. The most common way salaried women respond is having generally shorter commutes than comparable men measured in distance and time, although those using slower modes sometimes incur longer commutes measured in time (Reuschke and Houston, 2020).

Women also have more complicated and variable trip-patterns than comparable men, in part because they accept a greater share of household and child- and eldercare responsibilities and alter their work trip commute in distance and time to accommodate those tasks. A 2019 study of travel-diary data from Australia, the UK, Spain, and Finland found that men and women often spend similar *total* time traveling, but that there were great disparities in trip purposes, mode choice, and how children's travel was arranged (Craig and van Tienoven, 2019). The researchers concluded that we continue to see a gendered distribution of labor and social parenting norms; with women the parent conducting child-serving travel no matter how egalitarian a society considered itself (i.e., Finland). A 2020 study found substantial differences in the travel patterns of comparable women and men but concluded that, rather than cultural roles or constraints, gendered labor market and employment structures created differences in commute distance by gender (Reuschke and Houston, 2020). The results are the same however.

Blumenberg et al. (2019) found that women in one car households in California often took that car—but rather than being a privilege it was an indication that the women had assumed the bulk of household and childcare duties. A recent German study found the same pattern; women in dual-earner families often took the sole car, accepting all childcare and household serving obligations, and had shorter commutes (Chidambaram and Scheiner, 2020). The German study also found that the gender imbalance was helped if and when the man took on additional household duties.

Han et al. (2020) found that dual earner households sometimes traded off activities requiring travel but that men tended to provide the routine trip of dropping kids off at school or childcare in the morning while women tended to take responsibility for the rest of the daily trips that arose. A Canadian study found that professional (employed) women spend more time than comparable men on home-serving tasks including shopping and taking care of children and other adults. These women also spent more time driving each day even if they worked the same number of hours per day as comparable men (Shirgaokar and Lanyi-Bennett, 2020). A Dutch study found that even when salaried women did not make a specific child- or household serving trip, they were responsible for

scheduling the trip for the other parent or adult in the household (Schwanen, 2007). These findings led both Craig and van Tienoven (2019) and Shirgaokar and Lanyi-Bennett (2020) to stress the need to understand gendered travel when adopting employment and transportation policies.

Gender gaps in travel are far greater in many countries in the Global South (Arman et al., 2018) generally linked to both less formal employment among women in turn linked to norms about what it is acceptable behavior for women (Cunha, 2006; Wafa et al., 2008). Women are forbidden or seriously constrained from using bicycles or electric scooters/three wheelers in many countries where the use of such a mode would substantially improve their mobility.

The most severe examples are in Muslim countries (Wafa et al., 2008). A Pakistani study found that women were substantially less likely to travel outside the home than comparable men, and were far more likely to walk when they did travel. The researchers concluded, as have those studying similar countries, that women's role was seen as a private one, with remaining in the home upholding family honor (Adeel et al., 2017). A later study by the same authors (Adeel and Yeh, 2018) found that 55% of women but only 6% of Pakistani men did not travel on any given day; women with *higher* levels of education and income who had children were the *least* mobile of all. The authors concluded this behavior was "triggered by a gender-based sociocultural environment which restricts female mobility due to family honor concerns."

Women's Mode Choice

The gap between women and men for some commonly used variables, such as having a driver's license, choosing to drive, and commute distance, has closed over time—but is still large enough to be significant. Colley and Buliung's (2016) study of 25 years of Toronto's "gendered transport" (1986–2011) finds that employed men drive more than comparable women and that the gender gap increases with the age of the traveler. The gender gap in driving has, however, decreased over the 25 years as more women entered the paid labor force, undertaken longer commutes, and head households. Women in 2011 were more likely to drive than in the past so other modes (public transit and walking) were not as intensely gendered in the Toronto region as they had been in 1986. Yet there was still a meaningful gap between women and men's modal choices and commute distances.

Research has consistently shown that women are more likely to use public transit and to walk for travel when they can than comparable men, although most studies show that as women become employed they are more likely to be auto drivers. Women have generally been far less likely to cycle than men in both industrial and developing nations (Sersli et al., 2020) although in countries with high cycling rates, like The Netherlands, there is a smaller gender gap. A 2016 study in the United Kingdom where cycling has been declining nationally for over 30 years, found that communities able to increase cycling overall saw no increase in the share of women cyclists and a decrease in the number of older cyclists (Aldred et al., 2020).

There is a large and still growing literature analyzing why women cycle substantially less than men particularly for transportation or commuting (as opposed to recreational cycling where women cycle slightly more often). In fact one scholar has argued that there is no further need for research on the topic; that we can and should formulate appropriate policies and programs to encourage women and men to cycle now based on the extant knowledge base. The literature shows that women tend to be worried about their safety and sometimes their security when cycling; they prefer grade separated cycling facilities rather than traveling with traffic (Aldred et al., 2017). A study in Brussels also found that women were substantially less likely to cycle than men due to concerns about traveling in lanes with motorized traffic and unpleasant auto fumes (de Geus et al., 2019). A US study of immigrants in San Francisco found that the lack of neighborhood safety was a significant deterrent to women cycling, an issue not of concern to most men (Barajas, 2020). There is some evidence that women also worry about their clothes, hair, and make-up when using cycling for their commute.

An important share of the literature on women and cycling focuses on the use of bike share programs, where bikes are available for rental at stations throughout a city generally by using a credit card. A Norwegian study found that a bike sharing program in Oslo was less accessible to, suited to, and used by women because they had significantly different travel patterns than the male cyclists for whom the system was defined (Bocker et al., 2020). The National Institute for Transportation and Communities (2019) found that disadvantaged women were suspicious of and concerned about bike share programs because some did not have credit cards. Others were worried that deposit charges would remain on their cards for some time reducing the amounts they could charge to those cards until the deposits were removed or they were afraid the bike might be damaged or lost after they returned it but the full costs would be charged to their cards.

The inability or unwillingness to cycle also reflects the societal roles that women assume or the constraints under which they operate (Prati, 2010). A 2019 Australian study found that women did not find cycling convenient for the myriad activities in which they had to engage (Bourke et al., 2019). Ravensbergen et al. (2020) argue that the planners must recognize the gender difference in cycling is not a biological or "natural" one but is simply a response to how society divides domestic and childcare responsibilities. The National Institute for Transportation and Communities (2019) found that disadvantaged women in three US cities reported multiple barriers to cycling, including the need to travel with multiple children and packages, the lack of space for a bicycle at their home or destination, and fear of assault or harassment when cycling. Women in many developing nations also fear assault or harassment and are kept from cycling for any purposes by cultural norms which view such behavior as inappropriate or even immoral (Iqbal et al., 2020). Ravensbergen et al. (2020) argue, however, that many of these issues are not intrinsic problems with cycling and could be addressed with better training and some changes to cycles (providing ways to carry packages or children for example).

Existing research has been limited by a failure to focus clearly on differences among groups of women (although there are exceptions). Many studies fail to control for *intersectionality*, racial and ethnic differences in travel patterns and attitudes that may persist even when controlling for income, employment status, and household characteristics. Women of color often face substantial discrimination in the labor market which substantially impacts when and where they work, which in turn impacts the transportation modes available to them or which they can afford. In the Global South women in general, but particularly women of color, are often unable to use regular buses, BRT, or rapid rail systems because they cannot afford more formal modes and/or those modes do not serve the far-flung location of their jobs, which often differ significantly from the jobs available to men. In industrial nations women of color are often forced to make long difficult public transit trips, traveling further for work than other women (or men) with the same income (Salon and Gulyani, 2010).

Personal Security

A significant barrier to women's travel is fear of crime, sexual harassment, and physical assault; women around the world make major changes in all aspects of travel in response to these fears (Adur and Jha, 2018; Chowdhury and van Wee, 2020; Gardner, et al., 2017). Brazilian researchers found that far fewer female Brazilian university students felt safe traveling to campus than male students; these women, often forced to walk lacking other alternatives, would often take longer circuitous routes where they felt they were less likely to be crime victims (Capasso da Silva and da Silva, 2000). More than 60% of female riders of the BRT system in Barranquilla, Columbia reported being victims of sexual harassment while using the system; overcrowded buses had the most negative impact on their sense of security but they were also concerned about traveling at night or alone and inadequate lighting (Orozco-Fontalvo et al., 2019). Kash (2019) found that 37% of women using public transit in Bogota and Soacha, Columbia reported substantial sexual harassment; they responded by adopting what she calls "defensive postures," avoiding travel at certain times, traveling in groups, seeking a defensible position in a transit vehicle, taking more expensive modes, and even abandoning a trip entirely.

Researchers in New Zealand found that many female public transit riders, particularly women of color, were fearful of being a victim and were extremely sensitive to transfer and wait times at public transport terminals because they felt especially vulnerable at such points. They carried open cell phones, constantly checked apps on vehicle arrival times, and wore headphones to avoid hearing verbal harassment (Chowdhury and van Wee, 2020).

The problems of sexual harassment in the Indian sub-continent are particularly well known and have traditionally been ignored by local authorities (Mitra-Sarkar and Partheeban, 2011; Malik et al., 2020). Gadekar (2016) found that women in the Indian sub-continent are conditioned from childhood to accept a negative image of themselves and to understand their limited rights to the city from the daily harassment commonly known as "eve-teasing." Women are forced to endure cat calling, lewd remarks, fondling, and rubbing by men as they pass on the street, wait at bus or train stops, or ride aboard vehicles (In 2012, in an incident that made international news and sparked domestic outrage, a group of males (including one boy) gang raped a female college student on a bus in Delhi; she was later flown to Singapore for medical assistance but died of her injuries. The men defending themselves by saying no decent women would have been traveling alone without a male member of her family (she was traveling with a male friend whom the group beat) (Phillips et al., 2015). The boy received three years confinement, one man committed suicide while in jail, and the remaining men were hanged in 2020 for the crime. There were mass protests by women around the country and the Indian government undertook a study of the conditions that led to the attack; they made some changes—additional lighting, enhanced patrols of public transit stops (although the incident occurred on a vehicle) and creating apps to report attacks—but critics charge that the government did little to address the basic cultural problems). Kuruvilla and Suhara (2014) found that their entire sample of 120 women university students in Kerala had been subjected to almost daily sexual harassment, entirely on public transit vehicles. A study in Labore (India) found that harassment at Bus Rapid Transit (BRT) stations and in buses was a major challenge to women travelers (Malik et al., 2020). A study in New Delhi found that women's perception of the likelihood of being sexually harassed on public transit was consistent with self-reported occurrences but they underestimated the likelihood of attacks on informal transport modes (like taxis and auto-rickshaws). They researchers also found that women viewed incidents of sexual harassment as much more serious than did men did (Madan and Nalla, 2016).

Women in India and Pakistan respond by travelling only in groups or with male family members or colleagues; they change the times at which they travel or their destinations, they avoid certain places (like public transit transfer points) and they carry their cell phones open in their hands at all times. Not surprisingly a number of women in developing nations use more expensive options when available (taxis) or report planning to buy automobiles to avoid the worst of these incidents (Malik et al., 2020; Natarajan, 2016) in spite of UN and international focus on promoting safe, affordable, accessible, and sustainable public transport. Hsu (2009) found in fact that women in both California and Taiwan who were fearful about harassment using public transit thought buying a car was the only reliable solution to the problem. Gekoski et al. (2015) noted that the same phenomenon among women who had the resources to purchase a car but concluded that women in countries with a wide income disparities often do not have access to cars and are likely transit captives.

Sexual harassment is largely ignored by transportation planners, public transit agencies, and police for a variety of reasons. Kash (2019) found that transit officials in Bogota often believed that victims were lying or mistaken, sexual harassment did not really hurt victims, victims were often at fault, it was "natural" for men to behave in these ways so their behavior could not be changed, and finally that transit officials could not be responsible for societal norms and values (the last not an uncommon response in multiple transit systems). The lack of appropriate official responses creates a vicious cycle in both the developed and developing world:

women do not know that they can report such actions or they realize the futility of doing so while transportation authorities use the lack of reports of harassment or assault as proof that there is no problem (Natarajan, 2016). The majority of these incidents are not viewed as crimes; even if they are, the perpetrators are generally gone before law enforcement can be summoned so police often do not bother to respond. Bystanders have no idea what to do or when. There are often jurisdictional issues as well even when harassment serious enough to be considered a crime occurs—do transit police or the police of a local jurisdiction respond when it occurred on a train or bus or at a transit stop?

Some transit operators in Bangladesh, Brazil, Dubai, India, Japan, Mexico, and Nepal provide women-only coaches on trains but this practice can divide families or colleagues traveling together. A British study noted that women-only coaches, although an effective way to reduce unwanted sexual behavior, were “. . . essentially a ‘short-term’ fix and reinforce a message that women must be contained and segregated in order to protect them” and as such they did not recommend this option for countries like Britain where they “. . . would be a retrograde step.” (Gekoski et al., 2015, Recommendations, unpaginated). Moreover, studies show that gangs of men often wait at the doors of such coaches to harass and attack women as they exit (although Graglia (2015) argues that women-only coaches in Mexico City were the basis of an important rights-based movement).

Loukaitou-Sideris (2009) surveyed a number of US public transit operators and found that very few had programs or policies that targeted the safety and security needs of women riders—even though most felt that women had distinct needs. In fact only a third believed that it was the responsibility of transit systems to address these issues because they were a larger societal problem (echoing the sentiments that Kash found among South American transit officials). Since that time a few large public transit operators in North America have developed policies or programs to make women (and men) who are victimized aware that these behaviors are either forbidden by the operator or illegal and to make reporting harassment easier. Some transit operators provide guidance to by-standers as well the victims themselves on what to do at the time of the attack and how to report; some have created special phone apps allowing victims and witnesses to report such incidents immediately. These efforts are often accompanied but advertising and education campaigns stressing that such behavior is unacceptable and can be reported/ But we have little information on the impact or effectiveness of such programs.

A 2015 study undertaken for the British Transport Police and the UK Department for Transport (Gekoski et al., 2015) found that transit operators in industrial countries had considered or (infrequently) adopted a variety of initiatives to address sexual harassment and violence. In addition to providing women-only coaches, these initiatives included better station and vehicle maintenance, emergency alarms or panic buttons or phones on vehicles or platforms, dedicated places at transit hubs to report harassment and assault, hotlines to report incidents, request stops on buses, real-time scheduling information and apps, and legislative changes providing harsher punishments for offenders. The study also found that many transit agencies improved surveillance in several ways: *formally* by dispatching high visibility staff and police as well as undercover operatives; *technologically* by deploying CCTV cameras on vehicles and at transport hubs, and *naturally* through better lighting and visibility throughout the network. They also developed community outreach programs and awareness campaigns to make it clear that this behavior was unacceptable and could and should be reported. The study found however, that there were numerous examples of national and international initiatives in this area, “Unfortunately there are few rigorous evaluations using before and after measures of crime/incidents or randomized control trials to provide evidence of whether such initiatives achieve their aims” (Gekoski, Executive Summary, unpaginated.)

Overall, sexual harassment and even physical attacks continue to be underreported, under-policed, and underestimated crimes and practices that make women’s travel difficult and even dangerous around the world. Few systems are undertaking major efforts to address these issues in either the Global South or the industrial world. The limited responses of so many transportation authorities, including their failure to even acknowledge let alone keep track of such incidents, impose serious burdens on women, the victims, without addressing the perpetrators at all.

Women and Acceptance of New Transportation Technology

Women have long been known to have different attitudes and preferences toward certain modes and travel choices; women are for example more concerned about the environment and making personal choices that support sustainability. It surprised some researchers then to find that women were suspicious of or viewed pooled transportation services, like Lyft and Uber, and automated vehicles, negatively. A study in the Czech Republic, for example, found that women were less likely to be positive toward connected and automated vehicles than comparable men (Havlickova et al., 2020). A Finnish study found that women held less positive feelings toward automated vehicles (Liljamo et al., 2018); Hohenberger et al. (2016) also found that German women were more likely to have negative emotions about automated cars and were unwilling to say they would use them. The gender differences were not explained by age or education effects. These authors suggested stressing ways to reduce women’s anxiety about such options and promoting the aspects that women did or might have pleasurable about automated vehicles.

All indications are that women’s responses to both current pooled options and potential shared (or individual) automated vehicles are a combination of security and safety concerns and lifestyle issues. Women are worried about cars that drive themselves and lack confidence in the power of technology to deal with all roadway problems. They are also anxious about boarding vehicles with strangers, particularly men, already aboard, especially if they have to share bench seating. Finally women’s attitudes are also conditioned by the childcare and domestic responsibilities they face. Women may prioritize time over money; they may simply not have the time for the route detours needed for other travelers. Moreover, they are often traveling with children or older relatives as well as carrying strollers and car seats and packages. And they may need to make multiple stops on any journey (Institute of

[Transportation Studies, 2019](#)). Women may find it hard to see how pooled resources now or in the future will fulfill many of their needs or address their personal safety and security concerns.

Traffic Safety

The Industrial World

Women drivers and pedestrians are generally far safer than comparable men in the industrial world; women have substantially fewer auto and pedestrian crashes than men although they are more likely to be seriously injured or die in crashes of comparable severity ([Rosenbloom and Herbel, 2009](#)). In 2017, women in the European Union accounted for less than a fourth of total traffic fatalities (as a pedestrian, driver, or passenger) and a third of the number of male traffic fatalities ([European Commission, Table 1, undated](#)). In 2017, women's traffic fatality rates were 24.0 per one million population in the EU28, a ratio that was unchanged since 2006 (EC, Table 2, undated). Women in fatal traffic crashes in the EU were substantially less likely to be driving a vehicle when killed in it. In fact in 2017, there were only three EU countries where driver fatalities accounted for 50% or more of all women's traffic deaths (the Netherlands, Austria, and Denmark) and more than a dozen EU countries where female driver fatalities constituted less than 20% of all women's traffic deaths. Women in the EU in fatal traffic crashes were largely dying as pedestrians or passengers in vehicles in spite of being involved in a lower absolute number of such crashes. Men in contrast were substantially more likely to die as the driver in a traffic fatality; there was no country in the EU 28 in which driver fatalities were less than half of annual male traffic fatalities. In the majority of EU countries in 2017 almost three-quarters (and sometimes more) of all male traffic fatalities were driver fatalities (EC, Figs. 8a and 8b, undated).

Traffic fatality rates say little about who is to blame but the EU did find that women drivers were less likely to be cited by the police for serious or fatal traffic crashes in five European countries between 2005 and 2008, although citation rates are not necessarily an accurate evaluation of crash responsibility (in part because one cause is often attributed to a crash due in reality to sequentially linked contributing factors). When cited as drivers by the police for auto crashes in which they were involved, women were less likely to be driving in the wrong direction or traveling at excess speeds but more likely to be stopping short or taking action prematurely than men (EC, Fig. 9, undated). These data, moreover, did not include information on the blood alcohol level of the drivers, although we know that alcohol is involved in large number of crashes and that men are more likely to have been drinking than women.

Good data by sex on driving or incurring a pedestrian crash while impaired are limited for the European Union. A 2012 report by the World Health Organization (WHO) found that 17% of all women's pedestrian crashes in Europe involved alcohol compared to 40% of men's pedestrian injury crashes. Among men 15–29, 46% of road injury crashes involved alcohol; the comparable figure for women that age was 18%. Among those 30–44 years of age, 25% of the road crashes of men that involved an injury were related to alcohol; the comparable figure for women was 25%. Alcohol was involved far more in injury related crashes in all modes in the Baltic countries and Central Europe for both sexes; in a few of those countries women's injury crash rates involving alcohol consumption equaled or exceeded men's ([WHO, 2012](#), p. 50).

US data show similar patterns, in spite of the fact that women drivers have outnumbered male drivers on US highways for a number of years. US women are safer drivers and die more often as passengers than drivers and as pedestrians more than passengers. In 2018 roughly 86% of both men and women 14 years of age and older were licensed drivers in the US—but that total included 2.5 million more female than male drivers (FHWA, Table DL20b, 2019). [US National Highway Safety Transportation Administration \(NHTSA\)](#) data show that in 2018 men constituted over three-quarters of drivers involved in a vehicle crash in which one or more occupants died. But women made up half of the occupants who actually died and almost two thirds of occupants injured in those vehicle crashes. Female drivers, however, were involved in a higher percent of the total number of crashes that involved injury (but not death) to one or more vehicle occupants than men but in fewer of the total number of crashes that involved only property damage. Yet in both cases women's crash involvement rate per 100,000 female drivers (not population) was lower, sometimes substantially, than men's (NHTSA, Table 57, undated).

Data from the [US Insurance Institute for Highway Safety](#) show that passenger death rates have fallen significantly in the 20-year period from 1998 to 2018, but faster for men than for women. The number of women who died as a passenger in a personal vehicle dropped almost 40% over that period but declined almost 46% for men. In 1998, women constituted 49.7% of all passenger fatalities but that rose slightly to 51.7% in 2018. In 1998, women were 29.8% of all pedestrian fatalities rising slightly to 30.3% in 2018. The more sobering statistic is that while the total number of driver and occupant deaths in traffic crashes declined substantially over the 20-year period, the number of pedestrian deaths increased 29% for women and 26% for men (IIHS, calculated from table, "Motor Vehicle Crash Deaths by type and gender, 1975–2018."). It's also important to recognize that pedestrian injuries and even deaths are officially underreported around the globe; researchers using hospital ER and admitting data find that the actual number of serious pedestrian crashes may be from three to four times higher than crashes reported by the police/official crash data—especially among the young and old. A British study that corrected for that problem found that women and older people were far more likely to be injured in a pedestrian fall than other travelers in ways not fully recognized by official data ([Aldred, 2018](#)).

The role of alcohol and other impairments also breaks along gender lines. US women drivers were also far less likely than men to be involved in fatal vehicle crashes (where anyone died) in which they were alcohol impaired. In 2018 women accounted for 26.4% of all alcohol-impaired drivers involved in a fatal crash and just under 20% of all those drivers with a blood alcohol level at or over 0.08 g/dl (the common dividing line for driving under the influence in most US states). This was a 1% increase in the number (not share) of female drivers with a blood level 0.08+ g/dl involved in a fatal crash over the 10-year period 2009–18 (NHTSA, Table 3,

2019)—in spite of the fact that the total number of drivers (of both sexes) involved in such crashes increased 12% and the number of women drivers on the road went up 8.4% over the same time period (computed from NTSA, Table DL20, 2010 and 2019). We would, of course, hope for alcohol related fatal and other crashes to decrease sharply due to better education and enforcement (as have total driver and occupant crashes), but women drivers continue to be a much smaller share of the impaired driving problem than male drivers.

The gender differences in crashes and crash outcomes are a result of a combination of factors. Women drivers have fewer fatal crashes because they drive less, do so more safely, and are less likely to be impaired (by illegal drugs or cannabis as well as alcohol) (Lloyd et al., 2020). Women are also more likely to follow the rules, wear their seat belts, and not speed (Rosenbloom and Herbel, 2009). A recent study of how drivers use GPS devices in their vehicles, for example, found that men were more likely to violate the speed limit and more frequently than women (Yared et al., 2020). At the same time, the data above show that women are more likely to die (or be seriously injured) in the crashes in which they are involved whether as a pedestrian, driver, or passenger.

Women's poorer vehicle crash outcomes are due to several factors. First, the protection systems in cars are designed largely for men and not for women (or for pregnant women). In fact, it is practically impossible for women, who are generally of shorter stature than the average man, to be sitting in the safest places in a car. Second, women are on average frailer and thus more likely to die in pedestrian crashes (World Health Organization, 2002). Vehicle concerns are not being well addressed because regulators have not required manufacturers to specifically do so; a recent EU study concluded that most safety assessments and vehicle tests use the average size male as the regulatory model, excluding average sized women (Linder and Svedberg, 2019). Thus we have not substantially increased our knowledge of how to make personal vehicles safer in a crash for women occupants (driver or passenger). We also lack good information on where and how a pregnant women should wear a seat belt although a recent study found that belted/restrained pregnant women suffered less severe injuries than unbelted women (Schellenberg, et al., 2020). We still do not know, however the impact of seat belts in a crash on the fetus; one study found that fetuses sometimes died in fairly minor crashes in which the mothers were not seriously injured.

There are also some indications that younger women may be closing the gender gap in crash rates (IIHS, undated). Yet some of these crashes may result from the societally imposed roles that many women assume; women who are driving with children are particularly prone to having their attention diverted from the driving task (Maasalo et al., 2016, 2017). Kelley-Baker and Romano (2014) find that the percent of women drinking and driving with children in the car has also steadily increased and at faster rates than among men (although the growth rates reflect the very low starting point for women). Yet most drinking drivers are still men and having a child in the car seems to induce both sexes to drive more safely overall (Maasalo et al., 2017).

Developing Nations

We lack good gender data on traffic crashes and outcomes in most developing nations—a significant problem in itself. The World Health Organization reported in 2002 that there was a steadily rising number of road traffic crashes in developing nations as a result of the substantial increase in motor vehicles and vehicle congestion. A 2020 study found that 93% of all global road traffic injuries take place in low- and middle-income countries; worldwide, traffic injuries are the eighth leading cause of deaths in these countries, costing an estimated 3%–5% of Gross Domestic Product, in addition to the major financial impact on families (Lomia et al., 2020). Georgia, a middle-income country in the Caucasus region of Eurasia, has one of the highest road traffic injury rates of any country in the European region—such crashes were the second leading cause of death among women 15–49 years old and the principal cause of death for women 15–35 in both 2006 and 2012 with no improvement over time.

Lomia et al. (2020) analyzed the traffic crashes of 843 Georgian women 15–49, 78 of whom (or 9.3%) had died in the crash; they did not conduct a comparative analysis with male crash victims. The researchers found that 71.8% of the fatal crashes occurred when women were the passenger in a personal vehicle; only 5.1% of women who died were driving the vehicle in which they were riding. Twenty percent of women who died were pedestrians; all other modes accounted for the remaining 3%. Most of the women who died did so at the scene of the crash (and not on the way to the hospital or at the hospital). Surprisingly the victims were equally divided across the income spectrum; 60% were married and 60% were aged 40–49. In fact, the researchers found that being employed, married, and having completed a secondary education was associated with a greater likelihood of dying in a traffic crash, a finding they could not fully explain. It is possible that such women just had greater total traffic exposure although the authors postulated that more highly educated and employed women not only had greater access to automobiles but were more likely to flaunt basic traffic laws because of their time-compressed schedules.

Men significantly outnumber women as road injury victims in developing nations however—by roughly double—largely but not entirely because women rarely use automobiles so men have higher exposure to the risk of vehicle related traffic injuries. Men in developing nations are also more likely to suffer traffic fatalities because of extremely risky behavior; a study in Karachi found that men were substantially more likely to jump on or off a moving bus amidst very heavy traffic. Pedestrian deaths are a leading cause of traffic deaths in developing nations, particularly among men, because of the large number of two- and three-wheeled motorized vehicles that mix with pedestrian traffic, the failure of all vehicle drivers to observe traffic laws designed to protect pedestrians, and poor or non-existent pedestrian facilities that force travelers into the roadway.

WHO (2002) noted that women are more likely significantly impacted by traffic crashes than men because they have limited access to medical facilities and medical insurance and lack money to pay for childcare or other domestic duties when they are too injured to perform these tasks. Moreover, when the male member of a household dies or is seriously injured in developing nations, the whole family may suffer financial devastation.

Conclusions and Policy Implications

Women and men's travel differs, differences that begin in early childhood and persist as learned behaviors. Most research shows that women and men's travel patterns still differ even as other factors (employment, commute distance, licensing and auto use, etc.) have narrowed the travel gap between men and women. Women are more likely to undertake commute trips than in the past since roughly two-thirds of all women in the industrial nations have salaried employment outside the home, more than 60% (and almost 90% in some countries) have drivers licenses, and many own or have primary access to a car. Women's travel patterns, however, still reflect an unequal distribution of household and childcare and eldercare obligations, all of which play out in their travel patterns. Women in most industrial nations make more trips, although they generally travel fewer miles, than comparable men; they have more complicated travel patterns because they often link multiple trips together far more than do men while providing transportation for children and older relatives traveling to places they themselves do not need to be.

Women in developing nations are often limited by cultural and religious norms to the kinds of out-of-home activities in which they can engage and the kinds of transportation options that are considered suitable for their use. Women, for example, are often sanctioned if they try to cycle or drive three-wheeled vehicles or motorized scooters, although those options are generally the only ones they could afford and would substantially increase their employment and educational opportunities and decrease their total travel time. In a number of predominantly but not exclusively Muslim, countries women are forbidden to work or even travel outside the home; in other countries when they do they are harassed as they walk or use public transit. In almost every country in the world far fewer women cycle than men, although this is a cheap and sustainable mode that could enhance their health and mobility.

It is inescapable that women are also far more likely to be victims of sexual harassment and violence as they travel in both the developed and developing world; public officials and transit agencies have done little to address these issues which pose substantial barriers to women's travel, creating anxiety and fear even among women traveling in industrial nations. Women are generally safer drivers than men at all ages but they are substantially more likely to die as passengers in a private vehicle than as a driver; moreover they are more likely to die in vehicle and pedestrian crashes of comparable severity because they are more fragile on average and public policies have not required automobile manufacturers to understand women's vehicle protection needs or to alter vehicle design to make those vehicles safer—even for pregnant women.

What we know now about women's travel patterns and needs also reveals the many things we do not know or on which we have not previously focused. Many researchers call for greater attention to intersectionality, understanding how gender interacts with race and ethnicity and class and immigration status to create important travel differences among various groups of women. Others note that members of the LGBTQ+ community often face similar safety and security issues but they or their needs are not included in these discussions. We also need to better understand the implications of all these gender differences for women's ability or willingness to use new technology, in an age with rapid changes in modes of service, communications, and dependence on the so-called gig or part time economy facilitated by online platforms. It is also important to question why socio-demographic factors like income inequality and discrimination in labor and housing markets, factors associated with some differences between the sexes in travel behavior, so disproportionately impact women to begin with.

The failure to recognize persistent differences in the travel patterns and needs of women and men, and among individual groups of women, play out in public policy decisions and project evaluations. Not all differences between women and men require remediation of course or raise equity concerns; they must be reflected in our planning efforts however in order to ensure that transportation plans and system decisions reflect and respond to the various transportation "markets" that exist in increasingly diverse communities. Substantial evidence suggests, for example, that women are not well served by massive rail projects or Bus Rapid Transit projects which many low- and middle-income nations are constructing (often with major funding from international development banks) because they do not well serve the spread-out employment and other destinations to which women travel and often make no provisions to ensure women's personal security. Worse governments often discontinue or underfund local buses, and try to stamp out informal transport services (taxis, jeepneys, etc.) on which many women in developing nations rely, when major guideway projects are built, locking women out of employment options and subsistence jobs in which they formally engaged. It is important to attempt to understand how different plans and projects will and impact women and men differentially and provide appropriate options for different users. [Adom-Asamoah et al. \(2020\)](#) for example, examined rural transport infrastructure projects in Ghana and found that they impacted women and men differently because their travel patterns were so different. Adom-Asamoah et al. concluded, "... the issue of gender must be re-negotiated and properly understood."

Many analysts feel we know enough about how to encourage more women to cycle either for transportation or recreation (or both) to enact proper policies and programs—yet most governments at all levels fail to do so. Few governments expend resources for pedestrian facilities and bus stops, even though the lack of such facilities or their poor state of repair pose serious safety problems to many women. Certainly these problems are worse in the Global South where women frequently walk as their only travel mode or to connect with public transit, often with children born and unborn and packages, etc. Yet the lack of adequate pedestrian and public transit infrastructure (safe and accessible bus stops, for example) plagues many industrial nations in whole sections of urban areas creating disproportionate problems for older women or those traveling with children, or wheeling baby carriages or strollers, or carrying packages, etc. and clearly inhibiting active travel and public transit use. Overall many travel behavior researchers, government officials, and policy analysts still fail to consider the persistent differences in how women and men evaluate, perceive, and use (or not) different modes, and their significant differences in travel patterns and purposes. Public policies as a result often fail to meet women's needs or make various travel modes more practical, safer, secure, and convenient for women and men travelers.

Biography

Dr Sandi Rosenbloom is a Professor of Community and Regional Planning at the University of Texas at Austin. She has long studied the travel patterns of women, older people, and those with disabilities. She helped organize the first international conference on women's travel issues in 1978 and has been a member of the Conference Steering Committee for the subsequent five international conferences (the sixth held at the National Academies of Science, Engineering, and Medicine Beckman Center in Irvine, California in September, 2019).

References

- Adeel, M., Yeh, A.G.O., Zhang, F., 2017. Gender inequality in mobility and mode choice in Pakistan. *Transportation* 44 (6), 1519–1534.
- Adeel, M., Yeh, A.G.O., 2018. Gendered immobility: influence of social roles and local context on mobility decisions in Pakistan. *Transport. Plann. Technol.* 41 (6), 660–678.
- Adom-Asamoah, G., Amoako, C., Adarkwa, K.K., 2020. Gender disparities in rural accessibility and mobility in Ghana. *Case Stud. Transport. Policy* 18 (1), 49–58.
- Adur, S.M., Jha, S., 2018. (Re)centering street harassment – an appraisal of safe cities global initiative in Delhi, India. *J. Gender Stud.* 217 (1), 114–1214.
- Aibar Solano, A., Mandic, S., Lanaspá, E.-G., Gallardo, L.O., Casterad, J.Z., 2018. Parental barriers to active commuting to school in children; does parental gender matter? *J. Transport Health* 9, 141–149.
- Aldred, R., 2018. Inequalities in self-report road injury risk in Britain; a new analysis of National Travel Survey Data focusing on pedestrian injuries. *J. Transport Health* 9 (1), 96–104.
- Aldred, R., Elliott, B., Woodcock, J., Goodman, A., 2017. Cycling provision separated from motor traffic; a systematic review exploring whether stated preferences vary by gender and age. *Transport Rev.* 317 (1), 29–55.
- Aldred, R., Woodcock, J., Goodman, A., 2020. Does more cycling mean more diversity in cycling? *Transport Rev.* 36 (1), 28–44.
- Arman, M.A., Khademi, N., de Lapparent, M., 2018. Women's mode and trip structure choices in daily activity-travel; a developing country perspective. *Transport. Technol* 41 (8), 845–877.
- Barajas, J.M., 2020. Supplemental infrastructure: how community networks and immigrant identity influence cycling. *Transportation* 47 (3), 1251–1274.
- Blumenberg, E., Schouten, A., Brown, A., 2019. Who's in the driver's seat? Gender and the division of car use in auto-deficit households. *Transport. Res. Board Abstr.* 2019.
- Bocker, L., Anderson, E., Utend, T.P., Throndsen, T., 2020. Bike sharing in conjunction to public transport; exploring spatiotemporal, age, and gender dimensions in Oslo, Norway. *Transportation, Part A* 138, 389–401.
- Bourke, M., Craike, M., Hilland, T.A., 2019. Moderating effect of gender on the associations of perceived attributes of the neighborhood environment and social norms on transport cycling behaviours. *J. Transport Health* 13, 63–71.
- Capasso da Silva, D., da Silva, A.N., 2000. Sustainable modes and violence; perceived safety and exposure to crime on trips to and from a Brazilian university campus. *J. Transport Health* 16.
- Chidambaram, B., Scheiner, J., 2020. Understanding relative commuting within dual-earner couples in Germany. *Transport. Res. A* 134 (1), 113–129.
- Cunha, C., 2006. Understanding the community impact; bicycling in sub-Saharan Africa. *Sustainable Transport* 18 (1), 24–25.
- Chowdhury, S., van Wee, B., 2020. Examining women's perception of safety during waiting times at public transportation terminals. *Transport Policy* 94 (1), 102–108.
- Colley, M., Buliung, R.N., 2016. Gender differences in school and work commuting mode through the life cycle: exploring trends in the Greater Toronto and Hamilton area 1986 to 2011. *Transport. Res. Rec.* 2598, 102–109.
- Craig, L., van Tienoven, P., 2019. Gender, mobility, and parental shares of daily travel with and for children; a cross-national time use comparison. *J. Transport Geogr.* 76, 93–102.
- Cunha, C., 2006. Understanding the community impact; bicycling in sub-Saharan Africa. *Sustainable Transport* 18 (1), 24–25.
- de Geus, B., Wuytens, N., Deliens, T., Kester, I., Macharis, C., Meeusen, R., 2019. Psychosocial and environmental correlates of cycling for transportation in Brussels. *Transport. Res. Part A* 123, 80–90.
- European Commission, 2017. Traffic Safety Basic Facts 2017. European Road Safety Observatory. Viewed on: ec.europa.eu/transport/road_safety/sites/roadsafety/files/pdf/statistics/dacota/bfs2017_gender.pdf.
- Gadekar, U., 2016. "Eve teasing" and its psychosocial influence among adolescent girls. *Int. J. Curr. Adv. Res.* 5 (6), 1028–1031.
- Gardner, N., Cui, J., Coiacetto, E., 2017. Harassment on public transport and its impact on women's travel behavior. *Austr. Planner* 54 (1).
- Gekoski, A., Gray, J.M., Hovath, M.A.H., Edwards, S., Emirali, A., Adler, J.R., 2015. What 'Works' in Reducing Sexual Harassment and Sexual Offenses on Public Transport Nationally and Internationally; A Rapid Assessment. British Transport Police and Department for Transport, London.
- Graglia, A.D., 2015. Finding mobility; women negotiating fear and violence in Mexico City's public transit system. *Gender Place Culture* 23 (5), 624–640.
- Han, B., Kim, J., Timmermans, H., 2020. Turn taking behavior in dual earner households with children; a focus on escorting routines. *Transportation* 47 (1), 203–222.
- Hatamzadeh, Y., Habbian, M., Khodai, A., 2017. Walking behavior across genders in school trips; case study of Rasht, Iran. *J. Transport Health* 5, 42–54.
- Havlickova, D., Gabrhel, V., Adamovska, E., Zamecnik, P., 2020. The role of gender and age in autonomous mobility; general attitude, awareness, and media preference in the context of the Czech Republic. *Transport. Sci.* 10 (2), 53–63.
- Hohenberger, C., Spörle, M., Welp, I.M., 2016. How and why do men and women differ in their willingness to use automated cars? The influence of emotions across different gender groups. *Transport. Res. Part A* 94 (3), 374–385.
- Hsu, H.-P., 2009. How does fear of sexual harassment on transit affect women's use of transit. *Proceedings of the 4th International Conference on Women's Travel Issues. Conference Proceedings, Vol. 46*, pp. 83–99.
- Institute of Transportation Studies, 2019. Is it OK to Get Into A Car with a Stranger? Risks and Benefits of Ride-pooling in Shared Automated Vehicles. University of California at Davis. Viewed on: <https://escholarship.org/uc/item/1cb6n6r9> 2019.
- Insurance Institute for Highway Safety, 2018. Fatality Facts 2018: Gender. Viewed on: <https://www.iihs.org/topics/fatality-statistics/details/gender> undated.
- Iqbal, S., Woodcock, A., Osmond, J., 2020. The effects of gender transport poverty in Karachi. *J. Transport Geogr.* 84.
- Kash, G., 2019. Transportation professionals' vision of transit assault: the problems of deproblematizing beliefs. *Transport. Res. Part A* 139 (2), 200–216.
- Kelley-Baker, T., Romano, E., 2014. Gender differences in risky driving in presence of younger passengers. In: *Women's Issues in Transportation, Proceedings, 5th International Conference on Women's Travel Issues*.
- Kepper, M.M., Staiano, A.E., Katzarzyk, P.T., Reis, R.S., Eyler, A.A., Griffith, D.M., Kendall, M.L., Banna, E.L., Denstel, K.D., Broyles, S.T., 2020. Using mixed methods to understand women's parenting practices related to their child's outdoor play and physical activities living in diverse neighborhood environments. *Health Place* 62.
- Kuruville, M., Suhara, F., 2014. Response patterns of girl students to eve-teasing; an empirical study in a university setting. *Int. J. Educ. Psychol. Res.* 3 (2).
- Larouche, R., Gunnell, K., Belanger, M., 2019. Seasonal variations and changes in school travel mode from childhood to late adolescence: a prospective study in New Brunswick, Canada. *J. Transport Health* 12, 371–378.
- Larsen, K., Larouche, R., Buliung, R.N., Faulkner, G.E.J., 2018. A matched pairs approach to assessing parental perceptions and preferences for mode of travel to school. *J. Transport Health* 11, 56–63.
- Liljamo, T., Liimatainen, H., Poilanen, M., 2020. Attitudes and concerns on automated vehicles. *Transport. Res. Part F* 59 (1), 24–44.
- Linder, A., Svedberg, W., 2019. Review of average sized male and female occupant models in European regulatory safety assessment tests and European laws; gaps and bridging suggestions. *Accid. Anal. Prev.* 177, 156–162.

- Lloyd, S.L., Lopez-Quintero, C., and Striley, C.W., 2020. Sex differences in driving under the influence of cannabis; the role of medical and recreational use. *Addict. Behav.* 110.
- Lomia, N., Berdzuli, N., Sharashidze, N., Sturua, L., Pestvenidze, E., Pedersen, A., 2020. Socio-demographic determinants of road traffic fatalities in women of reproductive age in the Republic of Georgia; Evidence from the National Age Mortality Study (2014). *Int. J. Women. Health.* 12, 527–537.
- Loukaitou-Sideris, A., 2009. How to Ease Women's Fear of Transportation Environments; Case Studies and Best Practices. Mineta Transportation Institute, San Jose State University. Viewed on: <https://transweb.sjsu.edu/sites/default/files/2611-women-transportation.pdf>.
- Maasalo, I., Lehoten, E., Pekkanen, J., Summala, H., 2016. Child passengers and driver culpability in fatal crashes by driver gender. *Traffic Inj. Prev.* 17 (5), 447–453.
- Maasalo, I., Lehoten, E., Summala, H., 2017. Young female drivers at risk while driving with a small child. *Accid. Anal. Prev.* 108, 321–331.
- Madan, M., Nalla, M.K., 2016. Sexual harassment in public spaces; examining gender differences in perceived seriousness and victimization. *Int. Criminal Justice Rev.* 1, 1–18.
- Malik, B.Z., Rehman, Z., Khan, A.H., Akram, W., 2020. Women's mobility via bus rapid transit: experiential patterns and challenges in Lahore. *J. Transport Health* 17.
- Manish, M., Nalla, M.K., 2016. Sexual harassment in public spaces; examining gender differences in perceived seriousness and victimization. *Int. Criminal Justice Rev.* 1 (18).
- McMillan, T., Day, K., Boarnet, M., Alfonso, M., Anderson, C., 2006. Johnny walks to school—Does Jane? Sex differences in children's active travel to school. *Children Youth and Environments* 16 (1), 75–89.
- Mitra-Sarkar, S., Partheeban, P., 2001. Abandon all hope, ye who enter here; understanding the problem of “eve teasing” in Chennai, India. In: *Transportation Research Board (Eds.), Proceedings of the 4th International Conference on Women's Travel Issues*, Vol. 2. No. 46, pp. 74–84.
- Natarajan, M., 2016. Rapid assessment of “eve teasing” (sexual harassment) of young women during the commute to college in India. *Crime Sci.* 5 (6).
- National Institute for Transportation and Communities, 2019. *Narratives of Marginalized Cyclists: Understanding Obstacles to Utilitarian Cycling Among Women and Minorities in Portland, OR*. <https://rosap.nhtl.bts.gov/view/dot/32299> 2017.
- Orozco-Fontalvo, M., Soto, J., Arevalo, A., Oviedo-Trespalacios, O., 2019. Women's perceived risk of sexual harassment in a Bus Rapid Transit (BRT) system: The case of Barranquilla, Columbia. *J. Transport. Health.*, doi:10.1016/j.jth.2019.100598.
- Phillips, M., Mostofian, F., Jetly, R., Pulhukudy, N., Madden, K., Bhandan, M., 2015. Media coverage of violence against women in India; a systematic study of a high profile rape case.. *BMC Women's Health* 15 (8), doi:10.1186/s12905-015-0161-x.
- Potoglou, D., Arslangulova, B., 2019. Factors influencing active travel to primary and secondary schools in Wales. *Transport. Plann. Technol.* 40 (1), 80–99.
- Prati, G., 2010. Gender equality and women's participation in transport cycling.. *J. Transport. Geogr.* 66 (4), 369–375.
- Ravensbergen, L., Builung, R., Laliberte, N., 2019. A critical review of the literature on gender and cycling. *Transport. Res. Board Abstr.*
- Ravensbergen, L., Builung, R., Sersli, S., 2020. Velomobilities of care in a low-cycling city. *Transport. Res. Part A* 134 (3), 336–347.
- Reuschke, D., Houston, D., 2020. Revisiting the gender gap in commuting through self-employment. *J. Transport Geogr.* 85.
- Rosenbloom, S., Herbel, S., 2009. The safety and mobility patterns of older women; do current patterns foretell the future? *Public Works Manage. Policy* 13 (4), 338–353.
- Rosenbloom, S., Plessis-Fraisard, M., 2011. Women's travel in developed and developing countries; two versions of the same story? In: *4th International Conference on Women's Travel Issues. Proceedings*, Vol. 1. National Academy Press, Washington, DC.
- Salon, D., Gulyani, S., 2010. Mobility poverty, and gender; travel choices of slum residents in Nairobi, Kenya. *Transport Rev.* 30 (5), 641–657.
- Samimi, A., Ermagun, A., 2013. Students' tendency to walk to school; case study of Tehran. *J. Urban Plann. Dev.* 139 (2), 144–152.
- Schellenberg, M., Ruiz, N.S., Cheng, V., Heindel, P., Roedel, E.Q., Clark, D.H., Inaba, K., Demetriades, D.J., 2020. *Surg. Res.* 254 (1), 96–101.
- Schwanen, T., 2007. Gender differences in chauffeuring children among dual earner families. *Profess. Geogra.* 59 (4), 447–462.
- Shirgaokar, M., Lanyi-Bennett, K., 2020. I'll have to drive there; how daily time constraints impact women's car use differently than men's. *Transportation* 47 (3), 1365–1392.
- Simons, A.K., Koekemoer, A., van Nierkerk, R., Govender, 2018. Parental supervision and discomfort with children wanting to walk to school in low-income communities in Cape Town, South Africa. *Traffic Inj. Preve.* 16 (2), 391–398.
- Thigpen, C.G., Handy, S.L., 2018. Effects of building a stock of bicycling experiences in youth. *Transport. Res. Rec.* 2672 (36), 12–23.
- U.S. National Highway Safety Administration, 2018. *Alcohol Impaired Driving. Traffic Safety Facts. 2018 Data. Report DOT HS 812 864*.
- Wafa, E., Newmark, L., Shifan, Y., 2008. Gender and travel behavior in two Arab communities in Israel. *Transport. Res. Rec.* 2038.
- World Health Organization, 2012. Regional Office for Europe. *Alcohol in the European Union; Consumption, Harm, and Policy Approaches*. Viewed on: apps.who.int/iris/bitstream/handle/10665/107301/e96457.pdf.
- World Health Organization, 2002. *Gender and Health. Gender and Road Traffic Injuries*. Viewed on: apps.who.int/iris/bitstream/handle/10665/6887/a85576.pdf?sequence=1.
- Yared, T.P., Patterson, E.S., Li, E.S.S., 2020. Are safety and performance affected by navigation system display size, environmental illumination, and gender when driving in both urban and rural areas? *Accid. Anal. Prev.* 142.
- Yeung, J., Wearing, S., Hills, A.P., 2008. Child transport practices and perceived barriers in active commuting to school.. *Transport. Res.* 42 (6), 895–900.

Further Reading

- Caragata, G., Wister, A., Mitchell, B., 2019. “You're a great driver/Your driving scares me” reactions of older drivers to family comments. *Transport. Res. Part F* 60, 485–498.
- Chakrabarti, S., Joh, K., 2019. The effect of parenthood on travel behavior: evidence from the California household travel survey. *Transport. Res. Part A* 120, 101–115.
- Kwen, G., 2019. Always on the defensive; the effect of transit sexual assault on travel behavior and experiences in Columbia and Bolivia. *J. Transport Health* 13 (2), 234–246.
- McGuckin, N., Contrino, H., Nakamoto, H., Sanchez, A., 2009. *Driving Miss Daisy: Older women as passengers*. In *Conference Proceedings 35: Research on Women's Issues in Transportation. Report of the 4th International Conference; Vol. 2. Technical Papers*. Transportation Research Board, Washington, DC.
- Sersli, S., Gislason, M., Scott, N., Winters, M., 2020. Riding alone and together; is mobility of care at odds with mothers' bicycling? *J. Transport Geogr.* 13.
- U.S. Federal Highway Administration, 2010. Office of Highway Policy Information. Table DL-20. Distribution of Licensed Drivers – 2009 by Sex and Percent in each Age Group and Relation to Population. Viewed on: fhwa.dot.gov/policyinformation/statistics/2009.
- U.S. Federal Highway Administration, 2019. Office of Highway Policy Information. Table DL-20. Distribution of Licensed Drivers – 2018 by Sex and Percent in each Age Group and Relation to Population. Viewed on: fhwa.dot.gov/policyinformation/statistics/2018.
- Williams, S., Qiu, W., Al-awwad, Z., Alfayez, A., 2019. Commuting for women in Saudi Arabia; metro to driving—options to support women's employment. *J. Transport Geogr.* 77 (1), 126–138.
- Institute for Transport and Development Policy, 2018. *Women and Children's Access to the City*. itdpdotorg.wpengine.com/wp-content/uploads/2018/08/women-and-children-access-to-the-City_ENG-VI_June-2018.pdf.
- International Transport Forum, OECD, 2018. *Women's Safety and Security: A Public Transport Priority*. [itf-oecd.org/sites/default/files/docs/womens-safety-security](https://www.itf-oecd.org/sites/default/files/docs/womens-safety-security), 43 pp.
- NYU Rudin Center, 2018. *The Pink Tax on Transportation; Women's Challenges in Mobility*. https://wagner.nyu.edu/files/faculty/publications/Pink%20Tax%20Survey%20Results_finaldraft4.pdf.
- OECD, 2018. *Understanding Urban Travel Behaviour by Gender for Efficient and Equitable Transport Policies*. www.itf-oecd.org/understanding-urban-travel-behaviour-gender-efficient-and-equitable-transport-policies.
- Resources for the Future, 2017. *The Effect of Vehicle Restrictions on Female Labor Supply*. www.rff.org/files/document/file/RFF%20pt-China%20Vehicle%20Lottery%20Effects%20on%20Women.pdf.

Driver Stress and Driving Performance

Lisa Dorn, Cranfield University, Cranfield, Bedfordshire, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	203
The Driver Behavior Inventory (DBI)	203
The Driver Stress Inventory	204
Stress Vulnerabilities and Performance Effects	204
Aggression (AGG)	211
Dislike of Driving (DIS)	220
Alertness (AL)/Hazard Monitoring (HM)	220
Thrill Seeking (TS)	220
Fatigue Proneness (FP)	221
Driving for Work	221
Conclusion	221
Acknowledgment	222
Biography	222
References	222
Further Reading	223

Introduction

Driver stress, as defined by the transactional model (Matthews, 2001b, 2002), is an outcome of the interaction of the person and the traffic situation and emphasizes that stress is an unfolding process in the dynamic interchange with other road users and situational encounters. According to the model, environmental stressors are appraised and may elicit maladaptive cognitions and stress responses that tax the person's ability to cope. These cognitive stress processes support two forms of outcome: subjective responses such as anxiety, anger and tiredness, and performance outcomes such as impairment of psychomotor control and changes in speed. The impact of these stressors is moderated by personality factors and further influence how external stimuli are appraised. The ongoing selection of coping strategies may be effective or ineffective for driver safety. For example, a driver who is vulnerable to frustration whilst driving and copes by becoming confrontational can be expected to exhibit a discernible pattern of conflict laden encounters leading to greater risk taking.

This chapter reviews key findings of the development and validity of these measures of driver stress and performance outcomes. The application of the instruments for assessing individual driver stress vulnerabilities for targeted delivery of fleet safety interventions are also discussed.

The Driver Behavior Inventory (DBI)

Work began in the 1980s at Aston University in the UK with a series of studies to design an instrument to measure dimensions of driver stress (Gulian et al., 1989, 1990) that hitherto had not been attempted in such a comprehensive way. Initial studies tested hypotheses based on the transactional theory of stress (Lazarus and Folkman, 1984). The Aston studies proposed that driver stress is a compound function of factors both intrinsic (traffic conditions) and extrinsic (personal life) to driving. Gulian et al. (1988) described the development of the driver behavior inventory (DBI) using a few items from Parry (1968) and Stokols et al. (1978) with the remainder emerging from interviews with drivers and their driving diaries. Analysis of the pilot questionnaire highlighted a cluster of traffic situations, as well as feelings about driving and other road users, which were associated with and predictive of driver stress. In a follow-up study using a modified DBI (Gulian et al., 1989), factor analysis of responses to 31 items using a fresh sample revealed a 5-factor solution. Scales derived from the factors showed reasonably high internal consistency and accounted for about 40% of the variance as well as a 'general driver stress scale' from a one-factor solution:

- Driving aggression (AGG)
- Dislike of driving (DIS)
- Heightened alertness when driving (AL)
- Irritation when overtaken (IO)
- Tension when overtaking (OT)
- General driver stress' factor (GEN) constituted about 18% of variance.

The DBI scales have been investigated in relation to demographic characteristics and driving history and the extent to which these stress vulnerabilities correlate with self-report measures of personality, mood, stress, and performance. Studies have fairly reliably shown that the DBI factors are not general personality factors, but relate to various behaviors and reactions while driving (Dorn and Matthews, 1992; Glendon et al., 1993). The DBI traits are only moderately correlated with standard personality measures such as neuroticism (Matthews et al., 1997) and have greater predictive validity for driving-related criteria than do standard personality measures (Dorn and Matthews, 1995).

The Driver Stress Inventory

The Aston team then undertook a series of validation and reliability studies and developed the transactional model further (Matthews et al., 1996, 1997, 1998; Matthews, 2002). From these studies, the DBI was refined and extended given that the overtaking factors was found not to be stable and loaded predominantly on to AGG. Assessing thrill-seeking and fatigue reactions to driving were also deemed important components of stress but were not included in the original DBI.

The newly formulated DSI measures five distinct dimensions of stress vulnerability (Matthews et al., 1997):

- Driving aggression (AGG): characterized by negative appraisals of other drivers that tend to generate feelings of anger and dangerous driving behaviors.
- Dislike of driving (DIS): associated with negative self-appraisal and generates negative mood states and worries which tend to interfere with task performance.
- Hazard monitoring (HM): reflects the active monitoring for hazards to preempt threat by the vigilant search for danger.
- Thrill-seeking (TS): defined by enjoyment of danger and increased risk taking.
- Fatigue proneness (FP): associated with errors and lapses and predictive of loss of task engagement reflected in a lack of willingness to engage in effortful, task-focused coping.

The DSI is therefore a more comprehensive measure of affective reactions to driving. The alphas coefficients for the scales were found to be higher than those for DBI scales (Matthews et al., 1997). However, the homogeneity of the DBI and DSI scales have repeatedly been found to be satisfactory (Hoare, 2001; Glendon et al., 1993; Gulian et al., 1989; Lajunen and Summala, 1995; Matthews et al., 1996). It can therefore be concluded that this measure of driver stress is psychometrically sound. However, Westerman and Haigney (2000) identified somewhat different factor solutions although the three largest factors were recovered: AGG, AL and DIS as found in later studies (Kontogiannis, 2006; Taubman-Ben-Ari, 2008). The DBI/DSI dimensions have mainly been tested on English and American (US/Canadian) car drivers but also seems to generalize across to different cultures (Finnish: Lajunen and Summala 1995; Japanese: Matthews et al., 1999a; Greek: Kontogiannis 2006; Chinese: Qu et al., 2016) and different professional driver groups (truck drivers: Bergomi et al., 2017; Hartley and Hassani, 1994; police drivers: Gandolfi and Dorn, 2003; bus drivers: Besharati and Kashani, 2017; Dorn et al., 2010).

Driver stress may directly affect the driver's choice of behavioral coping strategies so a separate measure of coping with driver stress called the Driving Coping Questionnaire (DCQ: Matthews et al., 1997) was also developed based on previous work (Matthews et al., 1993). The DCQ aimed to capture how people tend to cope with the stress of driving and includes the following dimensions:

- Confrontive coping (CONF): standing up to others to assert one's needs or relieving one's feelings through risk-taking e.g. tailgating a slow vehicle to encourage them to move out of the way.
- Task-focus (TF): making an effort to drive safely when demands are high e.g. slowing down and being more vigilant.
- Emotion-focus (EF): regulating feelings and thoughts about events, criticizing oneself for mistakes, e.g. thinking reassuring thoughts when upset, or reflecting on why another driver committed an unsafe action.
- Reappraisal (RE): looking on the bright side when driving is demanding e.g. viewing the drive as a learning experience.
- Avoidance (AV): physical or mental disengagement from events by suppressing negative feelings. For example, a driver might avoid a particularly congested route or seek to remain detached when driving is demanding.

A consistent relationship between the DBI and DCQ scales have been established (Hoare, 2001; Matthews et al., 1997) suggesting that coping mediates the associations between driver stress vulnerability and stress outcomes. For example, several studies show that CONF is positively associated with AGG and TS. EF is positively related to DIS with drivers high on DIS being more prone to negative emotional coping that diverts attention from the driving task toward self-blame and worry. On the other hand, TF is positively correlated with AL/HM suggesting that drivers who focus on the problem appear to direct their efforts toward the task itself by adopting rational strategies such as information seeking, taking precautions, and making plans. FP is linked to emotion-oriented driving and reducing discomfort.

Stress Vulnerabilities and Performance Effects

Tables 1 and 2 show that the validity of the DBI/DSI scales has been tested using three main methods showing converging evidence; self-report, driving simulator, and field studies with self-report methods being most commonly used. The self-report studies have generally found that the DBI/DSI scales correlate with existing validated instruments and linked with demographic characteristics

Table 1 Driver behavior inventory (DBI) review

Reference	Method	Scale	Participants	Results	Comments
Brewer (2000)	Self-report study	DBI	249 Australian bus drivers and non-bus drivers (94% male). Age ranged from 19 to 64	AGG was predicted by AL, OT, and IO	Correlations between the DBI and other questionnaires (e. g., Driving Anger Scale) not reported
Chen (2007)	Self-report study	DBI	194 Taiwanese driver participants; mean age = 41.6. At least one years' driving experience	High AGG drivers reported higher frequency of mobile phone use	AGG validated with risk taking behavior
Dorn and Matthews (1992)	Self-report study	DBI	Study 1: 60 British participants (55% male); mean age = 28.8. Study 2: 54 participants (51% male); mean age = 28.9	In study 1, neuroticism is associated with ineffective coping strategies and extraversion is associated with use of rational, problem solving strategies. In study 2, DIS was associated with high neuroticism and low affection	DBI traits only moderately correlated with standard personality measures and have greater predictive validity for driving-related criteria
Dorn and Matthews (1995)	Driving simulator study	DBI	Study 1: 73 British participants; mean age = 44.3. Study 2: 93 participants; mean age = 22.7 Samples combined = 166	One sample performed a passive' drive, and the second sample performed an "active" driving task. The DBI traits were more strongly related to mood than personality traits. In both studies, DIS correlated with greater post-task tension and depression. DIS was the strongest single predictor of post-drive mood. Prior to the drive, participants also rated accident risk, driving skill, and judgment, for themselves and for peers. DIS rated themselves worse than peers and DIS was the strongest single predictor of self-ratings	The DBI predicts emotional stress in simulated driving
Glendon et al. (1993)	Self-report study	DBI	61 British drivers (67% male); mean age = 41.4	Drivers completed the DBI on two occasions separated 5-month interval. Coefficients alpha, test-retest reliability, and correlational analyses between the five scales revealed reliable scales. GEN was found to be particularly robust. AL was less robust and minor changes improved their reliability	
Gulian et al. (1989)	Self-report study	DBI	112 British participants (27% male); mean age = 44.00; mean driving experience = 20.0 years	Factor analysis revealed a 5-factor solution with reasonably high internal consistency as well as a "general driving stress scale" from a one-factor solution. Cognitive Failure Questionnaire (CFQ) scores correlated with AGG, DIS, and GEN	The original DBI research paper
Gulian et al. (1990)	Self-report and diary study	DBI	152 British (59% male) mainly commuter participants; mean driving experience = 22.0 years	13 GEN items correlated with Daily Driving Stress (DDS) on at least three separate days	
Hartley and Hassani (1994)	Self-report study	DBI	755 Australian participants; 184 truck drivers with high violation involvement (HVI) (37%) and 151 with low violation involvement (LVI) (30%) 183 car drivers with HVI (37%) and 237 (47%) with LVI	PCA revealed a six-factor solution. HVI car drivers scored higher on AGG than LVI car drivers but truck drivers did not differ. For DIS, both groups of LVI drivers scored higher than HVI counterparts. Truck drivers reported that driving is more stressful than car drivers did and HVI truckers reported that driving more stressful than LVI truckers, but this relationship was reversed for car drivers	The dependent variables were composites of accidents. Confidence of control predicted accidents and convictions

(continued)

Table 1 Driver behavior inventory (DBI) review (*cont.*)

<i>Reference</i>	<i>Method</i>	<i>Scale</i>	<i>Participants</i>	<i>Results</i>	<i>Comments</i>
Hennessy and Wiesenthal, (1997)	Field study	DBI	27 Canadian student sample and 13 working participants (50% males); mean age = 28.85	Significant interaction found between the GEN and levels of traffic congestion with greater levels of stress reported in high congestion conditions. High GEN scorers showed even further elevation in state stress than those who were lower in GEN, under similar conditions	First field study to validate the DBI but participants studied during only one trip
Hennessy and Wiesenthal (1999)	Field study	DBI	60 Canadian commuters (50% males); mean age = 28.80	GEN was a predictor of state driver stress in both low congestion and high congestion road conditions. State driver stress in low congestion was predicted by time urgency and GEN in low congestion	Transactional nature of driver stress confirmed in field trials
Hennessy et al. (2000)	Field study	DBI	56 Canadian commuters (50% male); mean age = 26.50	GEN correlated with the State Driver Stress Questionnaire (SDSQ) and the Survey of Recent Life Experiences (SRLE). GEN contributed positively under low congestion and hassles by GEN interaction under high congestion. Hassles exposure moderately increased state driver stress for high GEN drivers, but reduced state driver stress for medium and low GEN drivers	Only one commute trip investigated
Hennessy et al. (2004)	Self-report study	DBI	192 Canadian commuters (36% male); mean age = 25.83	High GEN drivers reported greater mild driver aggression than low stress drivers. Male drivers reported greater frequency of past violent driving behaviors than female drivers and high GEN drivers reported greater violence than low stress drivers	The DSI scores are associated with more extreme driver violence
Hennessy and Wiesenthal (2001)	Field study	DBI	192 Canadian commuters (36% male); mean age = 25.83.	Gen positively correlated with Self Report Driver Aggression and negatively correlated with driver Age. GEN predicts mild driver aggression and driver violence. High GEN drivers were also found to report elevated incidents of past driving violence compared to low GEN drivers	Female sample was slightly younger and less experienced than the male sample
Hill and Boyle (2007)	Self-report study	DBI	914 American drivers (60% male); 80% commuted to work daily	18 driving scenarios (for freeway and non-freeway driving) were grouped using factor analysis to group driving conditions that have similar stress ratings using the DBI. Four factors revealed that Weather-related conditions and interactions with other drivers were skewed toward higher stress levels, while performing driving tasks and visibility-related conditions were skewed toward lower stress levels. Drivers found weather-related conditions and interactions with other drivers to be more stressful than contending with limited visibility	Findings show that stress depends on driver characteristics and the specific driving environment
Kontogiannis (2006)	Self-report study	DBI	714 Greek participants (66% male)	DSI and DCQ factor analyzed for a Greek sample supporting the main factor structure of AGG, DIS and AL.	Drivers high in AGG were more likely to be caught speeding and reported more accidents

(continued)

Table 1 Driver behavior inventory (DBI) review (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
Lajunen et al. (1997)	Self-report study	DBI	Sample 1: 201 Australian student participants (35% male); mean age = 30.5; mean driving experience = 11.7 years Sample 2: 203 Finnish student participants (32% male); mean age = 24.3; mean driving experience = 5.3 years	Males scored higher than females on and AGG was negatively correlated with age. DIS increased with age and decreased with driving experience. High AGG drivers were more likely to engage in CONF and DIS was associated with self-criticism For sample 1, DIS was negatively associated with Driver Social Desirability Scale (DSDS) Driver Self-Deception (DSD). AGG was negatively associated with DSDS-Driver Impression Management (DIM). AL was positively correlated with DSDS-DSD, DSDS. For sample 2, DIS was associated with DSDS-DIM, DSDS-DSD. AGG was negatively correlated with DSDS-DSD and DIM. AL positively correlated with all of the study variables. In the Finnish data, AL had a stronger correlation with DIM than DSD, but in the Australian data AL correlated with DSD, but not with DIM.	DBI factors related to biased driver perceptions
Lajunen et al. (1998)	Self-report study	DBI	Sample 1: 201 Australian student participants (35% male); mean age = 30.5; mean driving experience = 11.7 years. Sample 2: 203 Finnish student participants (32% male); mean age = 24.3; mean driving experience = 5.3 years	AGG correlated negatively with safety skills for both samples. AL correlated with safety skills and perceptual-motor skills in sample 1. DIS had a strong negative correlation with perceptual-motor skills for both samples and correlated negatively with safety skills for sample 1. Speed was negatively related to DIS, but positively related to AGG and AL. DIS, AGG and AL had a significant effect on evaluations of perceptual-motor skills. AGG predicts Safety Motives	Correlations between the DBI scales and driving and personality variables were similar in both countries
Lajunen and Summala (1995)	Self-report study	DBI	113 Finnish student participants, mean age = 23.9 years (28% male); mean driving experience = 5.2 years	Safety motives negatively correlated with AGG and skill was positively correlated with AGG and negatively correlated with the DIS. DIS predicts skill ratings and AGG predicts safety motive scores. AGG correlated with the Externality measured by the Montag Driving Internality and Driving Externality Scales (MDIE)	Neuroticism positively correlated with AGG and DIS confirming previous studies
Liu and Lee (2005)	Field study with instrumented vehicle	DBI	12 Taiwanese participants (50% male); mean age = 35.2; all participants had at least two years driving experience	High AGG and low AGG drivers drove across intersections with or without car-phone. High AGG drivers more likely to approach stop line at a constant speed and brake more intensely and accelerate after the traffic lights than drivers low on AGG. High AGG drivers had 23 instances of high-risk driving including amber-light running tailgating; weaving in and out of traffic; improper passing; passing on the shoulder; improper	A small sample size limited statistical power

(continued)

Table 1 Driver behavior inventory (DBI) review (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
Lucas and Heady (2002)	Self-report study	DBI	125 American participants (30% male); mean age = 36.0	lane changes; failure to yield and running stop signs and red lights Time urgency and commute satisfaction was measured via self-report. Commuters who had flexible scheduling reported lower levels of GEN than commuters without flextime	Only one commute trip used when asked to complete the questionnaires
Matthews (1996)	Driving simulator study	DBI	80 British participants (50% males) aged 18–30.	Three significant interactive effects on reaction time were found between DIS, stress condition and road curvature. Significant interactions were also found between DIS and task combination and the stress manipulation were moderated by individual differences in stress vulnerability, assessed by DIS. High DIS drivers under stress showed slower RT, reduced cuing and increased heading error. The data show some correspondence between these effects on objective performance and subjective response to stress. High DIS was associated with a maladaptive pattern of increased cognitive interference under stress	Supports the effort regulation hypothesis, that driver stress may lead to insufficient effort or active control of performance when task demands are low
Matthews and Desmond (1998)	Driving simulator study	DBI	256 British participants (50% male); mean age = 20.9	DIS and FP correlated positively with all fatigue scales, pre- and post-task. AGG was positively associated with pre- and post- boredom, and pre- and post-muscular fatigue. AL was negatively associated with pre- and post- boredom, pre-malaise, pre-muscular fatigue and positively associated with post-malaise. AGG positively correlated with psychoticism and DIS was positively associated with neuroticism. FP was negatively associated with extraversion and positively associated with neuroticism. FP was positively associated with pre- and post-disengagement, distress, and physical fatigue scores; AGG was positively associated with pre-disengagement, distress and physical fatigue scores, and post-disengagement; DIS was positively associated with pre- and post-distress, physical fatigue, and post-disengagement scores. AL was negatively associated with pre- and post-disengagement, post- distress and pre-physical fatigue scores. AL was a significant predictor of pre-disengagement and pre-physical fatigue; AGG and FP were significant predictors of pre-distress and of post- disengagement.	DIS and FP were the most consistent predictors of greater fatigue

(continued)

Table 1 Driver behavior inventory (DBI) review (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
Matthews et al. (1991)	Self-report study	DBI	Study 1: 159 British participants; mean age = 33.0. Study 2: 44 British participants; mean age = 43.0. Study 3: 49 British participants; mean age = 27.0. Study 4: 50 British commuters; mean age = 41.0.	Study 1: Age correlated negatively with AGG, IO, OT and GEN and positively correlated with AL. Neuroticism positively correlated with AGG, DIS and GEN. Psychoticism positively correlated with AGG and GEN and Extraversion correlated with AGG and OT. Speeding conviction associated with lower DIS, IO, OT and AL. Study 2: Age positively correlated with GEN. DIS, AL with frequency of hassles. AGG associated with indirect aggression, irritability verbal expression of aggression and total score. AL positively correlated with negativism and GEN positively associated with irritability and total score. Study 3: Sustaining attention negatively associated with DIS and GEN and positively associated with AL. Absent-minded memory failures correlated with GEN. Efficient processing of auditory stimuli negatively correlated with DIS. Efficient attention to others in social situations positively correlated with AL. Study 4: DIS negatively correlated with tense arousal and hedonic tone. GEN associated with energetic arousal and tense arousal.	GEN AGG and OT associated with minor accidents
Matthews et al. (1993)	Driving simulator study	DBI	61 participants	Three types of driving situation: (1) open-road tasks; (2) following tasks; and (3) overtaking tasks. AGG highly correlated with risk-taking only in overtaking studies. Drivers high on AL paid better attention to hazards	
Matthews et al. (1998)	Driving simulator study	DBI	Study 1: 75 British participants; mean age 44.20 (60% male) Study 2: 65 British participants; mean age 37.40 (50% male) Study 3: 65 British participants; mean age 36.90 (51% male) Study 4: 93 British participants; mean age 22.7 (50% male)	Positional variability associated with DIS in study 1. In study 2, mean speed positively associated with AL and Attention RT negatively associated with AL. Attention (distance) was positively associated with AL and AGG negatively associated with mean distance. In study 3 and 4, positional variability positively associated with DIS. In study 4, simulator errors positively correlated with DIS. AGG positively correlated with mean speed, overtakes, high-risk overtakes, and errors. DIS was negatively associated with ability, judgement, and skill and AL positively associated with ability, judgement, and skill. Age was significantly correlated with AGG and females scored higher on DIS than males. AGG associated with pre-drive tension in Sample 1, pre-drive depression in Sample 4, post-drive energy in Sample 4 and post-drive depression	Results broadly consistent with predictions according to the transactional model of driver stress

(continued)

Table 1 Driver behavior inventory (DBI) review (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
Matthews et al. (1996)	Driving simulator study	DBI	80 British participants (51% males); mean age = 20.9; mean driving experience = 3.4 years	in Sample 4. DIS was positively associated with pre-drive tension in Sample 1, post-drive tension in Sample 1 and post-drive depression in Samples 1 and 4. AL positively associated with energy in Sample 4. High-DIS drivers showed impaired control on straight roads but not curves. High-DIS drivers adapted to high levels of demand, but may be at risk of performance impairment when task requires relatively little active control.	Driving performance varies according to task parameters and individual differences
Matthews et al. (1999)	Self-report study	DBI	510 Japanese participants (50% male); mean age = 36.7	Males and younger drivers score higher on AGG OT and DIS and high scorers rated driving demands higher than low scorers. AGG was the strongest predictor of accident involvement, speeding conviction and other convictions.	Similar factor structure in Japan. AGG the strongest predictor of accident involvement
Maxwell et al. (2005)	Self-report study	DBI	245 British student participants (46.5% males); mean age = 32.44.	Triangular relationship reported between traffic violations, Propensity for Angry Driving Scale (PADS) and AGG. Drivers high in AGG are more likely to commit violations such as speeding and running red lights than their non-aggressive counterparts. Frustration experienced when slowed by other drivers	University students and staff participated
Shamoa-Nir and Koslowsky (2010)	Field study	DBI	226 drivers from Israel (23% male); mean age = 29.0; mean driving experience = 10.27 years	Participants observed driving their cars. Drivers high on DIS reacted emotionally when coping with stress.	Drivers observed as they entered a parking lot and may not be representative of everyday driving
Taubman-Ben-Ari (2008)	Self-report study	DBI	290 conscripts from Israel (53% male); mean age = 19.52 and at least 1 year of driving experience	Self-report measures assessed the perceived costs and benefits of driving, driver self-image, and self-efficacy as a driver. Factor analyses revealed four benefit factors and four costs and four self-image factors. Higher perception of benefits and costs were related to a higher AGG and Self-image of an impulsive driver related to higher AGG. Self-images of a cautious and a courteous driver related to lower AGG. High AL positively correlated with perceived benefit of sense of control, and negatively with perceived costs of distress and life endangerment. DIS inversely associated with perceived benefit of pleasure, and positively related to perceived costs of distress, damage to self-esteem, annoyance, and life endangerment.	DBI factors associated with costs and benefits of driving in an expected direction
Taubman-Ben-Ari et al. (2004)	Self-report study	DBI	328 Israeli participants (32% male); mean age = 31.78	A self-report scale constructed adapting DBI items and other scales to create four driving styles called the multidimensional driving style inventory (MDSI). Factor analysis revealed 8 main factors,	

(continued)

Table 1 Driver behavior inventory (DBI) review (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
				representing dissociative, anxious, risky, angry, high-velocity, distress reduction, patient and careful driving	
Van Rooy et al. (2006)	Self-report study	DBI	322 American Psychology undergraduates (26% male); mean age = 22.12	Construct validity of the DBI showed high correlations among AGG and the Driver Anger Scale (DAS) and Driver Vengeance Questionnaire (DVQ) providing evidence for convergent validity. DBI predicted traffic violations but this was not found for the DAS or DVQ	A confirmatory factor analyses failed to demonstrate discriminant validity between the scales
af Wählberg (2010)	Self-report study	DBI	7522 British participants in wave 1 (59% males; mean age = 21.7). 5633 participants in wave 2 = (59% males; mean age = 21.7)	Significant associations for offender drivers in waves 1 and 2 were found between AGG and self-reported collisions and $r = 0.50$, $P < 0.001$ and penalty points ($r = 0.50$, $P < 0.001$)	Significant associations between AGG and self-reported collisions for offender drivers
Wickens and Wiesenthal (2005)	Self-report and field study	DBI	42 Canadian participants (69% male)	Trait driver stress measured using GEN and participants completed the State Driver Stress Questionnaire (SDSQ) and the Job Stress Survey (JSS) to assess time urgency and perceptions of control during commute in a safe parking space. Time urgency negatively influenced cognitive appraisal of a stressful situation, leading to increased tension and possibly more aggressive driving behavior	High and low traffic congestion studied
Wickens et al. (2015)	Field study	DBI	170 undergraduate students (41% male); mean age = 23.80	GEN a significant predictor of vengeance measured by the Driving Vengeance Questionnaire (DVQ)	Vengeance scores were somewhat low across the sample

and driving history in an expected direction. Field studies have tested the predictive validity of the DBI versus experienced stress in actual driving situations (Hennessy and Wiesenthal, 1997, 1999; Hennessy et al., 2000), showing that the DBI scales are fairly strongly associated with how people react when driving in congested areas. The experimental studies using driving simulators have identified more precise effects of the DBI/DSI dimensions on driving and task performance in a controlled environment. These findings complement those from self-report and field studies. Collectively, the studies reviewed here reveal that different stress vulnerabilities elicit qualitatively different patterns of stress response and impact on driving performance. These findings will now be explored in more detail below.

Aggression (AGG)

This review shows that high AGG drivers tend to be motivated toward maintaining mastery over other drivers (Matthews, 2001b). High AGG drivers tend to appraise other drivers as hostile, and to cope confrontationally. This cognitive style is associated with both anger and dangerous driving. Anger has been investigated via correlational studies of AGG and simulator performance. Matthews et al. (1998a) conducted several experiments in which performance was assessed in different contexts: driving on an open-road with no other traffic, following a lead vehicle with overtaking prohibited, and driving in slow-moving traffic with opportunities to pass. These studies showed that AGG related to performance in the overtaking context only. In two studies, high AGG drivers drove faster, committed more errors and executed more high-risk overtakes with implications for the context-dependence of AGG effects on driving performance. The transactional model implies that the cognitions typical of aggression are likely to be elicited especially by the close proximity of other drivers, or by impedance due to factors such as congestion. Matthews et al. (1998b) showed that, when permitted to pass slow-moving cars, high AGG drivers drove faster, committed more errors and executed more high-risk passes. Although, AGG did not relate to speed in open-road driving suggesting that risky behaviors may be an outcome of a confrontive coping strategy. This hypothesis is confirmed with studies of passing behaviors (Emo et al., 2004, 2016) and further supports the

Table 2 A review of the driver stress inventory and the driver coping questionnaire

<i>Reference</i>	<i>Method</i>	<i>Scale</i>	<i>Participants</i>	<i>Results</i>	<i>Comments</i>
Bergomi et al. (2017)	Salivary-amylase and cortisol study	DSI	42 Italian bus drivers (88% male); mean age = 40.0; mean length of service = 10.5 years	Cortisol levels were positively correlated with FP at the middle of the work-shift and with AGG at the end of a day off. At the end of the work-shift, cortisol levels were negatively correlated with HM) and salivary amylase was positively correlated with TS	The only physiological study to validate the DSI scales
Besharati and Kashani, (2017)	Self-report study	BDBI	107 Persian bus drivers; mean age = 41.9; length of service mean = 15.1 years	Factor structure of the bus driver behavior inventory (BDBI) replicated. The odds of crash involvement decreases by 51%, for each one-unit improvement in HM. For each one unit increase in the RD score, the driver would be 3.2 times more likely to be involved in a crash and the odds of crash involvement would decrease by 51% and 45%, for each one-unit improvement in the FP and TS respectively	HM, RD, FP, and TS predict bus crash involvement
Danaf et al. (2015)	Driving simulator study	DSI	96 Lebanese student participants	State anger and AGG effects found at a specific time period by the occurrence of frustrating events	No interactions with other vehicles studied
Desmond and Matthews (2009)	Field and driving simulator study	DSI and DCQ	Study 1 = 58 Australian long-haul male truck drivers; mean age = 37.09; mean driving experience = 15.08. Study 2 = 104 (42% male); mean age = 32.1; mean driving experience = 12.6 years	DSI and DCQ predictive of state response to driving, and the association between FP and post-drive fatigue found study 1 and 2. Truck drivers apparently resilient to fatigue experienced increases in both fatigue and emotional and cognitive stress symptoms that suggests a risk of performance impairment. Shorter simulator drives in Study 2 did not elicit these reactions in all drivers. FP predicted a range of subjective state variables and also predicted change in fatigue, controlling for individual differences in pre-drive state. DIS and AGG also predicted post-drive fatigue and stress dimensions. FP was the most reliable DSI predictor of fatigue state change, whereas DIS related to cognitive interference, and AGG to anger	Results add to the validity of the DSI as a predictor of post-drive subjective states in both real and simulated driving
Desmond et al. (2001)	Field study	DSI	Study 1: 58 American male truck drivers; mean age = 37.09 Study 2: 104 drivers (41% male); mean age = 32.10	FP positively associated with pre-drive tense arousal, task-induced fatigue and anger. FP also associated with cognitive interference, and negatively associated with hedonic tone and positively associated with	High FP drivers may be at risk from fatigue during both prolonged and relatively short driving trips

(continued)

Table 2 A review of the driver stress inventory and the driver coping questionnaire (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
Dorn (2005)	Driving simulator study	DSI and DCQ	53 British male police drivers; mean age = 36.7; mean driving experience = 18.37 years	post-drive measures. High FP drivers in both studies showed elevated levels of state fatigue, tension, unpleasant mood and task-related cognitive interference. HM significantly associated with reduced speed when passing a hazard and a significant correlation between speed at passing the hazard and CONF. Proximity to the hazard was positively correlated with TF. Drivers who crossed the center line at potentially unsafe locations scored higher on DIS compared with police drivers who crossed the center line at safer locations. Drivers who crossed the center line at unsafe locations scored higher on TS	Results add to the validity of the DSI for a police driver group
Dorn and Gandolfi (2008)	Self-report study	DSI and DCQ	334 company car drivers (69% male); mean age = 41.26; mean driving experience = 18.94 years	The DSI and 40 items based on in-depth interviews with drivers at work and a literature review of the risks of driving for work were administered along with the DCQ and the DSDS (Driver Social Desirability Scale; Lajunen et al. 1997). The DSI plus additional items were subject to a PCA revealing 6 factors explaining 46.31% of the variance. Five factors were largely replicated from the DSI – AGG, TS, and HM but FP reversed to reveal a Fatigue Resistance factors (FR) and DIS reversed to form an Enjoyment of Driving factor (ENJ). An additional factor labelled Work Related Risk (WRR) describing risk taking when under time pressure. The DCQ PCA replicated the original 5 factors explaining 50.44% of the variance. WWR positively correlated with AGG, TS, and CONF and EF and negatively correlated with FR and TF	Drivers with two or more crashes reported higher levels of AGG compared with those with 1 or no accidents. Drivers with three or more crashes reported higher levels WRR compared with drivers with two crashes or less
Dorn and Garwood (2005)	Self-report study	DSI and DCQ	Study 1: 45 British bus drivers (98% male); mean age = 43.5. Study 2: 315 bus drivers took part: 297 males and 18 females, aged between 19 and 71 (mean age 46.3)	Factor analyses of the DSI and items generated from interview with bus drivers revealed 7 factors named the bus driver behavior inventory (BDBI) included FP, HM Relaxed Driving (RD), Patient Driving (PAT), Anxious Driving (ANX),	Lower HM and PAT was correlated with blameworthy incidents. EV, ANX and AV predicted blameworthy incidents

(*continued*)

Table 2 A review of the driver stress inventory and the driver coping questionnaire (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
Dorn et al. (2010)	Driving simulator study	DSI and DCQ	943 British bus drivers (87% male); Study 1: 315 drivers; mean age = 46.3; length of service = 8.9 years. Study 2: 557 drivers; mean age = 46.5; length of service = 8.9 years. Study 3: 71 trainee bus drivers; mean age = 36.0	TS, Incident Inevitability (II) and Evaluative Coping (EV). Blameworthy incident-involved bus drivers scored lower on HM and patient driving (reversed AGG). New factors were similar to the original DCQ factors. Blameworthy incident-involved drivers scored lower on EV and higher on ANX and AV Six factors from DSI and four factors from DCQ were confirmed in a PCA for Study 1. Study 2 test-retest correlations for all the BDRI components ranged from 0.68 to 0.86. HM negatively correlated with all crashes, all responsible crashes and sole responsible crashes. FP correlated with all crashes and EV negatively associated with all crashes and solely responsible crashes. For Study 3, trainees were tested on three simulator drives EF and AV associated with a deterioration of driving performance when under time pressure. AV was consistently and significantly associated with celebration across all drives	Higher HM predicts lower crash involvement. FP positively correlated with all crashes. EV negatively associated with all crashes and solely responsible crashes
Emo et al. (2016)	Driving simulator study	DSI and DCQ	112 American students (36% male); mean age = 20.90; mean driving experience = 5.02 years	AGG positively correlated with CONF. DIS was positively associated with EF and TF. HM positively associated with TF. TS was correlated with CONF and RE and negatively associated with TF. FP was positively correlated with EF. AGG correlated with the UWIST Mood Adjective Checklist (UMACL) anger scale and DSSQ factors of worry and distress. DIS correlated with worry and distress. HM was associated with worry and FP was positively associated with distress and negatively associated with Task Engagement. DIS was positively correlated with distance performance index and negatively associated with distance to oncoming and passes. HM negatively associated with passes and TS positively associated with passes and distance	Individual difference in driver stress responses and risk taking are confirmed

(continued)

Table 2 A review of the driver stress inventory and the driver coping questionnaire (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
Emo et al. (2004)	Driving simulator study	DSI and DCQ	112 American college students (36% male); mean age = 20.1	AGG and DIS correlated significantly with the Dundee State Stress Questionnaire (DSSQ) post-task distress and worry; FP was associated with high distress and low task engagement. HM correlated positively with worry and DIS correlated with all three "risk-taking" indices of performance: mean distance to lead vehicle at initiation of pass, mean distance to oncoming vehicle. HM was negatively associated with passes and TS was negatively associated with mean distance to lead vehicle at initiation of pass and passes. AGG was associated with post-task anger. CONF, EF, RE and AV was positively associated with the DSSQ post-task worry factor. EF was positively associated with distress and negatively with TF. EF was a significant predictor of post-task worry and CONF was correlated with the three "risk-taking" indices of driving performance and passes. CONF was associated with AGG and post-task anger	Dispositional coping factors predicted risk-taking behaviors, such as frequent passing and were not mediated by situational coping. This suggests that emotions during driving may shape habitual behavioral styles
Funke et al. (2007)	Driving simulator study	DSI	168 American undergraduate students (41% male); mean age = 20.37; mean driving experience = 4.95 years	Pre-task DIS was related to distress and worry and task engagement. Post-task DIS was associated with distress, worry and workload and with distress in the free driving condition; with distress in the lead following condition; and with distress, worry and workload in the automation condition, although it failed to predict performance	Adds to existing personality work suggesting that automation may have unexpected effects on worry in stress-vulnerable drivers
Gandolfi and Dorn (2005)	Self-report study	DSI and DCQ	333 British police drivers (82% male); mean age = 36.17; mean driving experience = 10.81 years	Administered DSI and interview-based items to police drivers seven factor structure established a new inventory called the Police Driver Risk Index (PDRI). The PDRI factors were HM, AGG, TS, Fatigue Resistance (FR), Risk inevitability (RI), Enjoyment of driving (ENJ) and Anxious driving (ANX). Three coping Factors replicated from the DCQ and 2 socially desirable responding factors (partially replicated from the DSDS)	Most DSI factors retained and additional factors emerged for a police driver sample

(*continued*)

Table 2 A review of the driver stress inventory and the driver coping questionnaire (*cont.*)

<i>Reference</i>	<i>Method</i>	<i>Scale</i>	<i>Participants</i>	<i>Results</i>	<i>Comments</i>
Garwood and Dorn (2003)	Self-report study	DSI and DCQ	543 British bus drivers (93% male); mean age = 45.0	AGG, FP and HM retained in factor analyses for bus driver sample and two additional factors were defined as Safe/Cautious Driving with TS items loading negatively and Enjoyment of Driving loadings in the opposite direction. Five factors were confirmed equivalent to the original DCQ	First study to develop a bus driver version of the DSI
Hennessy et al. (2016)	Self-report study	DSI	206 Canadian student sample (46% male); mean age = 22.57; mean driving experience = 6.23 years	Two versions of a simulator-based video of a vehicle crossing the center line into the path of oncoming victim vehicle were created by varying the time period elapsed prior to and following a near collision. For the "early" timing condition, the offending vehicle traveled without incident for 30s prior to the near collision. In the "late" timing condition the offending vehicle traveled for 4 mins without incident prior to the near collision. A Primacy Effect dominance in driver judgments for the early group was moderated by high HM drivers ratings. Primacy may be due to increased vigilance to early stimuli under a stressful condition, and high HM led to more negative ratings of "riskiness"	HM score exaggerated negative evaluations in the Early Timing group
Hoare (2001)	Self-report study	DSI	54 Australian bus drivers (98% male); modal age range = 40 - 59 years. 72% employed full-time with 9+ years driving experience (69%)	AGG, DIS, FP, TS, CONF and EF predictors of Need for Recovery, Job-Related Affective Well-Being and Physical Symptoms. AGG the single best predictor of Job-Related Affective Well-Being, whilst FP was the most reliable predictor of Need for Recovery	The DSI associated with work related factors
Kashani and Besharati (2019)	Self-report study	BDBI	177 Iranian male bus driver participants; mean age = 41.7; mean driving experience = 11.9 years	BDBI factors based on the DSI and DCQ (Dorn et al., 2010) were replicated. Older drivers were significantly more prone to fatigue and greater levels of Relaxed Driving compared with younger drivers.	HM predicted bus drivers' crash involvement
Machin (2001)	Self-report study	DCQ	108 Australian male bus drivers aged between 30 and 60 and most drivers had been working for at least 9 years	Evaluation of a fatigue management program using the DCQ found that higher scores on CONF were associated with greater job dissatisfaction and higher scores on TF were negatively associated with length of service. High EF positively associated with kilometers driven and number	More experienced employees reported similar outcomes to new employees

(continued)

Table 2 A review of the driver stress inventory and the driver coping questionnaire (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
				of traffic fines. High RE scores associated with number of hours spent driving and with lower job dissatisfaction. High AV scores also positively associated with a greater number of hours spent. Higher scores on the Job-related Affective Well-being Scale were strongly related to increased TF and RE. Lower Affective Well-being was strongly associated with greater reported use of CONF and EF	
Machin and Hoare, (2008)	Self-report study	DCQ	159 Australian bus driver participants (98% male).	Weekly Driving Hours positively correlated with CONF and AV. CONF and EF accounted for 5% of the variance in Need for Recovery. CONF was a predictor of Physical Symptoms. EF and RE were predictors of Positive Affect. CONF and EF were predictors of the variance in Negative Affect	Low response rate from an initial 500 coach drivers contacted
Matthews et al. (1996)	Self-report study	DSI and DCQ	602 participants Sample 1: British at work drivers (48% male) Sample 2: British students (48% male) Sample 3: American students (36% male)	Alpha coefficients for the scales ranged from 0.73 - 0.87 in British samples, and from 0.69 to 0.85 in the US sample. Highest positive correlations between AGG and TS and between DIS and FP. Age was positively correlated with HM and negatively correlated with TS and DIS. Males were higher than females on TS but females had higher scores on DIS and FP. Positive correlations found between AGG and CONF, between DIS and EF and between HM and TF. TS was also strongly correlated with CONF and both TS and AGG were related to reduced TF. AGG, DIS, FP and TS were related to Neuroticism. Openness positively correlated with TS and negatively correlated with DIS; extraversion negatively correlated with HM and DIS and positively correlated with TS	
Matthews et al. (1997)	Self-report study	DSI and DCQ	Sample 1: 339 British working drivers (48% male) Sample 2: 244 British student (51%) Sample 3: 219 American students (36% male)	A principal components analysis of the 83 DSI stress items was conducted and five factors extracted. DSI scores were compared (1) in accident involved and non-involved groups, and (2) in convicted and non-convicted groups. In the British sample, accident involvement related to higher	The study uses data from Matthews et al (1996). TS, HM and were associated with higher accidents and convictions

(continued)

Table 2 A review of the driver stress inventory and the driver coping questionnaire (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
Oz et al., (2010)	Self-report study	DSI	234 Turkish male drivers, professional and non-professional from four different driver groups (taxi drivers, $N = 69$; minibuses drivers, $N = 63$; heavy vehicle drivers, $N = 64$; and non-professional drivers, $N = 38$)	TS, AGG and lower HM. Convictions were associated with AGG, FP and a trend toward higher TS in the convicted group. In the US sample, both accidents and convictions related to AGG for accidents and convictions, and to TS for accidents. Differences for the relationship between AGG and TS between taxi and non-professional drivers, minibuses and heavy vehicle drivers, minibuses and non-professional drivers, and heavy vehicle and non-professional drivers. The correlations for AGG and FP were significant for the taxi and minibuses drivers and a significant difference was observed in the correlations between AGG and HM for the minibuses and truck drivers	Professional drivers report greater levels of driver stress than non-professional drivers suggesting different types of countermeasures
Qu et al. (2016).	Self-report study	DSI	244 Chinese participants aged between 18-66 (48% male)	Internal consistency coefficients ranged from 0.63 to 0.75. AGG, HM and FP correlated with all of the Driver Behavior Questionnaire (DBQ) subscales ranging from ± 0.17 to ± 0.46. DIS was positively associated with errors and lapses. TS was significantly associated with AGG and errors. AGG and FP were associated with accidents and minor accidents and HM was negatively associated with minor accidents. DIS and FP were negatively associated with mileage whilst TS was positively associated with mileage. Females reported greater levels of DIS and FP and lower levels of HM than male drivers did. Speeding and traffic sign violation offenders reported higher AGG and TS, while drivers with traffic sign violations reported decreased HM compared with non-offenders	AGG and FP scales predicted accidents and minor accidents. Higher HM associated with lower number of minor accidents
Rowden et al. (2011)	Self-report study	DSI	247 Australian employee participants; age ranged from 22 years to 69 years with the majority being male (77.7%)	DSI scales found to have acceptable reliability. AGG, DIS and FP positively associated with DBQ errors, lapses and violations. AGG FP and TS were positively correlated with Violations which was negatively associated with HM. Age was negatively related to all DSI scales except	Driver stress may be reciprocally related to stress in other domains, including work and domestic life

(continued)

Table 2 A review of the driver stress inventory and the driver coping questionnaire (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
				HM which was positively related. Gender differences were found for DIS, AGG, FP and HM in the expected direction. AGG and DIS were positively associated with the General Health Questionnaire (GHQ), Job Related Tension Scale (JRTS) and Daily Hassle; FP was positively associated with Daily Hassle and GHQ. Confirmatory factor analysis of the predictor variables revealed two factors related to DSI: negative affect and risk taking which influenced both lapses and errors, and the DSI risk-taking factor had the strongest influence on violations	
Susilowati and Yasukouchi (2012)	Self-report study	DSI and DCQ	35 Japanese participants: older group; mean age = 61.9; older age group; mean age = 69.5. 10 student participants in the young group (50% male); mean age = 23.3	Correlation for DSI and DCQ factors observed only in older drivers. FP was lower for younger drivers than older drivers. Older drivers also had a lower score for TS and AGG. Young drivers scored higher on EF whereas the older group tended to score moderately. For the older driver groups EF was correlated with RE and AV. For the older age group 2 AV was positively correlated with CONF, TF and EF	
Wishart et al. (2017)	Self-report study	DSI	892 Australian employee participants (59% male); mean age = 43	TS predicted all work-related risky driving behaviors taking into account age, gender and work driving exposure. High TS scorers were more likely to conduct unsafe driving behaviors including errors and violations and driving while distracted and fatigued. TS was the strongest predictor and accounted for the largest unique variance of all occupational risky driving behaviours	Variance explained by thrill and adventure seeking in driving error is relatively low (11.5%)
Wohleber and Matthews (2016)	Self-report study	DSI	127 American student participants (49% male); mean age = 20.16	Self-confidence in general driving ability and self-confidence in managing impairments correlated negatively with DIS and FP. Self-confidence in ability to compensate for impairments correlated positively with TS and negatively with HM. Ratings for the above average effect (AAE) were negatively associated with DIS. Additionally, impairment AAE was positively correlated with AGG. DIS	AAE was the significant predictor variable, and HM TS, and FP substantially improved prediction for DBQ errors

(*continued*)

Table 2 A review of the driver stress inventory and the driver coping questionnaire (*cont.*)

Reference	Method	Scale	Participants	Results	Comments
Zhang et al. (2009)	Driving simulator study	DSI	20 participants (50% male); Young group mean age = 20.7; mean driving experience = 5.4 years; Older group mean age = 72.2; mean driving experience = 55.1 years	was associated with reduced confidence on all self-estimation and overplacement High HM and FP associated with reduced speed in the presence of a hazard with velocity ratio correlating with these DSI factors. AGG and DIS did not show any significance with age. Under normal driving conditions, no effect of driver stress on driving behavior was seen	Presence of a hazard or driving under a critical condition has a significant influence on the driver's stress state

transactional model. It seems that dangerous behaviors are more likely to be elicited in interaction with other road users as hostile appraisals trigger impatience and a confrontive coping style, consistent with their higher crash frequency in real life (Matthews et al., 1999b).

Dislike of Driving (DIS)

DIS is associated with negative self-appraisal (Dorn and Matthews, 1995) and emotion-focused coping strategies such as self-blame that generate negative mood states and worries which tend to interfere with task performance. Higher levels of tension, anxiety, unhappiness and worry during both real and simulated drives have been reported (Dorn and Matthews, 1995; Matthews, 1996, 2001). Simulator studies of DIS show how the maladaptive cognitions associated with anxiety, such as emotion-focused coping, may produce performance impairment, especially in threatening environments. DIS correlates positively with variability of lateral position, indicating poor psychomotor control, with control errors and was negatively associated with total overtaking frequency (Matthews et al., 1998). Higher frequencies of driving errors have also been reported by high DIS drivers in self-report studies (Matthews et al., 1997). It appears that high DIS drivers are more behaviorally cautious so that there is little overall impact on safety although negative self-appraisals may divert attention from the driving task, resulting in impaired performance efficiency and behavioral distractibility.

Further studies investigated cognitive factors for effects of DIS on attentional performance, using dual-task methods (Matthews et al., 1996). Drivers performed an additional secondary task and task demands were independently manipulated by varying the curvature of the road (straight vs curved). The results showed the suppression of typical DIS effects under high task demands showing that when task demands are low, high DIS drivers may divert attention onto internal worries. DIS effects on performance were also accentuated following exposure to steering control disruption. Here, DIS was associated with poor control of lateral tracking and poor attention to secondary task stimuli presented on road-signs as they attempted to control a skidding vehicle (Matthews, 1996).

Alertness (AL)/Hazard Monitoring (HM)

The DBI AL factors and the DSI HM factor appear to be more narrow in scope from a stress perspective and relate to a specific style of task-focused coping by actively monitoring for hazards and detecting hazards faster (Dorn, 2005; Matthews et al., 1998; Zhang et al., 2009). AL/HM does not appear to be associated with negative mood or stress states compared with other dimensions (Matthews et al., 1997). AL is associated with safety skills and perceptual-motor skills (Lajunen et al., 1998) and negatively associated with pre- and post-disengagement (Matthews and Desmond, 1998). HM is negatively correlated with self-confidence ratings in ability to compensate for impairments (Wohleber and Matthews, 2016) and associated with reduction in speed when passing a hazard (Dorn, 2005; Zhang et al., 2009); negatively correlated with cortisol levels (Bergomi et al., 2017); decreasing odds of crash involvement (Besharati and Kashani, 2017); more negative evaluations of danger (Hennessy et al., 2016); a lower number of minor collisions (Qu et al., 2016) and bus crashes (Dorn et al., 2010; Kashani and Besharati, 2019).

Kontogiannis (2006) reported that AL increased with age, possibly due to increasing driving experience. However, Westerman and Haigney (2000) found that AL decreased with age perhaps due to an age-related decline in cognitive capacity. Further studies are required to investigate the role of strategy and capacity for different age groups.

Thrill Seeking (TS)

High TS is defined by enjoyment of danger and risk taking. Therefore, negative correlations with DIS have been reported (Matthews et al., 1996). TS has been found to be a significant predictor of violations in the DBQ (Matthews et al., 1997; Qu et al., 2016;

Westerman and Haigney, 2000) and is positively correlated with driving offences (Matthews et al., 1997; Rowden et al., 2011). TS is linked to CONF (Emo et al., 2016) and self-reported accident involvement (Besharati and Kashani, 2017; Matthews et al., 1997) and positively associated with number of passes, risky passes and reduced distance to other vehicles in simulator studies (Dorn, 2005; Emo et al., 2016). For drivers at work, TS also predicts all work-related risky driving behaviors taking into account age, gender and exposure. TS was the strongest predictor and accounted for the largest unique variance of all occupational risky driving behaviors (Wishart et al., 2017). High TS scorers reported unsafe driving behaviors including errors and violations.

Fatigue Proneness (FP)

Initial studies of FP showed that this dimension was a significant predictor of errors and lapses, (Matthews et al., 1997). In subsequent simulator studies, FP effects on performance have been studied primarily through experimental inductions of fatigue by having drivers perform a perceptually and cognitively demanding secondary task concurrent with driving. FP effects were found only when driving straight road sections, rather than curves (Matthews et al., 1996). The deficits shown by fatigued drivers during straight-road driving may reflect difficulties in mobilizing sufficient effort to maintain adequate performance in a monotonous situation, due to loss of motivation. Desmond et al. (2001) found that FP is most predictive of loss of task engagement following a fatiguing or long-duration drive. High FP drivers also showed elevated levels of state fatigue, tension, unpleasant mood and task-related cognitive interference. Matthews (2001) suggests that high FP drivers tend to be motivated toward avoiding discomfort when tired. These processes may be moderated by the driver's appraisal of the need for effort and their willingness to engage in task-focused coping depending on driving task variation demands.

The above studies show that the DBI/DSI measures and predicts individual differences in driver stress and performance. The research to date reveals that drivers are not passively experiencing "stress" due to external demands but are actively evaluating the events they encounter. The appraisal process appears to trigger personal concerns, such as their attitudes toward other drivers, and beliefs about their own driving competence and ability to cope. These kinds of behavior have implications for driver safety. Whilst crashes are difficult to predict, especially with self-report measures (af Wählberg and Dorn, 2015), it is encouraging that correlates of crash involvement with the DBI/DSI scales have been found. Tables 1 and 2 show that the DBI/DSI scales have been found to be correlated with road traffic accidents in several studies (Dorn and Gandolfi, 2008; Dorn et al., 2010; Kashani and Besharati, 2019; Kontogiannis, 2006; Matthews et al., 1991, 1997, 1999; af Wählberg, 2010). Minor crashes were predicted in two studies (Matthews et al., 1991; Qu et al., 2016) and partly replicated in Matthews et al (1996). Cognitive interference may explain the statistical link between stress and crash risk previously described (Matthews et al., 1998). Emotional stress impairs performance by adding to the driving task demands and may divert attentional resources, leading to performance decrements, particularly if the task is demanding.

Driving for Work

It has been well established that driving for work is stressful given that it may require the pressure of driving to schedules, often for long hours and with some form of job performance or delivery targets to achieve. DSI/DCQ Studies have shown that drivers at work are particularly prone to adverse stress and fatigue reactions as summarized in Tables 1 and 2 (Bergomi et al., 2017; Desmond and Matthews, 2009; Hoare, 2001; Matthews et al., 2001; Oz et al., 2010). Intercorrelations have also been found for the driver stress dimensions with general health, job related stress and daily hassles (Machin 2001; Rowden, et al., 2011; Wickens and Wiesenthal 2005; Wishart et al., 2017) suggesting a "spillover" effect. That is, driver stress may be reciprocally related to stress in other domains including work and domestic life and the scales have been found to predict various stress outcomes for professional drivers.

In view of these findings, several studies have called for the administration of the DSI/DCQ to assess individual differences in driver stress and coping (Dorn and Gandolfi, 2008; Dorn et al., 2010; Machin, 2001; Matthews et al., 1998; Rowden et al., 2011). Individual scores can then be used to deliver interventions to develop driver safety-enhancing appraisals and coping responses. In 2005, a fleet version of the DSI and DCQ scales including the DSDS (Lajunen et al., 1997) called the Fleet Driver Risk Index (FDRI) (Dorn and Gandolfi, 2008) was made commercially available to fleet-based organizations on a Cranfield University online platform called DriverMetrics. Since then variants for other professional groups have also been added for profiling professional drivers (police; Gandolfi and Dorn, 2005; bus; Dorn et al., 2010). Coaching interventions to regulate the emotional reactions to driving are used to address maladaptive cognitions, stress responses and coping to mitigate the risk of work-related crash rates. Research to evaluate the effectiveness of this approach using a control group is planned, although many case studies have shown significant reductions in crash rates since implementation. In professional driving settings, this multivariate assessment can also support the hiring of those drivers whose styles of cognition promote safety and thus are less likely to cause vehicular damage and injury and fatal crashes. Based on the research evaluated in this chapter, drivers with adaptive coping styles will probably have fewer personal injuries, cause less property damage, and therefore be a more productive employee.

Conclusion

The data reviewed here suggests an individual driver's propensity toward risk-taking is dependent on a variety of personality-based emotional factors, including enjoyment of risk, risk-taking associated with frustration, impatience and hostility to others and

ineffective coping strategies. About seventy key papers have studied the DBI/DSI/DBQ scales confirming its psychometric properties, stability over time, validity and predictive capacity for driver reactions, performance behaviors, and accidents. Research on its practical application as a driver risk assessment tool for stress management when driving for work is underway and is currently widely used across many countries within large fleet-based companies.

Acknowledgment

I am grateful to Gerry Matthews and Anders af Wahlberg for comments on an early version of this review and to Tetiana Hill for a literature search.

Biography

Dr Lisa Dorn is an Associate Professor of Driver Behavior and Director of the Driving Research Group at Cranfield University. Dr Dorn is also Founder and Director for DriverMetrics, a Cranfield University spin-out company to exploit her research in the design of educational interventions to improve driver safety. Dr Dorn has been a Principal Investigator for nearly three decades on a range of projects on driver behavior.

References

- Bergomi, M., et al., 2017. Work-related Stress and Role of Personality in a Sample of Italian Bus Drivers. 433–440.
- Besharati, M.M., Kashani, A.T., 2017. Factors contributing to intercity commercial bus drivers' crash involvement risk. *Arch. Environ. Occup. Health*.
- Brewer, A.M., 2000. Road rage: what, who, when, where and how? *Transport Rev.* 20, 49–64.
- Chen, L.Y., 2007. Driver personality characteristics related to self-reported accident involvement and mobile phone use while driving. *Saf. Sci.* 45 (8), 823–831.
- Danaf, M., Abou-Zeid, M., Kaysi, I., 2015. Modeling anger and aggressive driving behavior in a dynamic choice—latent variable model. *Accid. Anal. Prev.* 75, 105–118.
- Desmond, P.A., Matthews, G., 2009. Individual differences in stress and fatigue in two field studies of driving. *Transport. Res. Part F* 12 (4), 265–276.
- Desmond, P.A., Matthews, G., Bush, J., 2001. Individual differences in fatigue and stress states in two field studies of driving. In: *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*, pp. 1571–1575.
- Dorn, L., 2005. Professional driver training and driver stress: effects on simulated driving performance. In: Underwood, G. (Ed.), *Traffic and Transport Psychology*. Elsevier, Amsterdam.
- Dorn, L., Garwood, L., 2005. Development of a psychometric measure of bus driver behaviour. *Behavioural Research in Road Safety: 14th Seminar*. Department for Transport, London.
- Dorn, L., Gandolfi, J., 2008. Designing a psychometrically-based self-assessment to address fleet driver risk. In: Dorn, L. (Ed.), *Driver Behaviour and Training Vol. 3*, Routledge.
- Dorn, L., Matthews, G., 1992. Two further studies of personality correlates of driver stress. *Personal. Indiv. Differ.* 13, 949–951.
- Dorn, L., Matthews, G., 1995. Prediction of mood and risk appraisals from trait measures: two studies of simulated driving. *Eur. J. Personal.* 9, 25–42.
- Dorn, L., Stephen, L., Wahlberg, A., Gandolfi, E.J., 2010. Development and validation of a self-report measure of bus driver behaviour. *Ergonomics* 53 (12), 1420–1433.
- Emo, A.K., Matthews, G., Funke, G., 2016. The slow and the furious: anger, stress and risky passing in simulated traffic congestion. *Transport. Res. Part F: Traffic Psychol. Behav.* 42, 1–14.
- Emo, A.K., Matthews, G., Funke, G., Warm, J.S., 2004. Stress vulnerability, coping and risk-taking behaviours during simulated driving. In: *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1228–1232.
- Funke, G.J., Matthews, G., Warm, J.S., Emo, A., 2007. Vehicle automation: a remedy for driver stress? *Ergonomics* 50, 1302–1323.
- Gandolfi, J., Dorn, L., 2005. Development of the police driver risk index. In: Dorn, L. (Ed.), *Driver Behaviour and Training, Vol. II*. Aldershot, Ashgate, pp. 337–347.
- Garwood, L., Dorn, L., 2003. Stress vulnerability and choice of coping strategies in UK bus drivers. In: Dorn, L. (Ed.), *Driver Behaviour and Training*. Aldershot, Ashgate, pp. 55–64.
- Glendon, A.I., Dorn, L., Matthews, G., Gulian, E., Davies, D.R., Debney, L.M., 1993. Reliability of the driving behaviour inventory. *Ergonomics* 36, 719–726.
- Gulian, E., Glendon, A.I., Matthews, G., Davies, D.R., Debney, L.M., 1990. The stress of driving: A diary study. *Work Stress* 4, 7–16.
- Gulian, E., Matthews, G., Glendon, A.I., Davies, D.R., Debney, L.M., 1989. Dimensions of driver stress. *Ergonomics* 32, 585–602.
- Hartley, L.R., El Hassani, J., 1994. Stress, violations and accidents. *Appl. Ergon.* 25, 221–230.
- Hennessy, D.A., Jakubowski, R.D., Brittany, L., 2016. The impact of primacy/recency effects and hazard monitoring on attributions of other drivers. *Transport. Res. Part F: Traffic Psychol. Behav.* 39, 43–53.
- Hennessy, D.A., Wiesensthal, D.L., 1997. The relationship between traffic congestion, driver stress and direct versus indirect coping behaviours. *Ergonomics* 40, 348–361.
- Hennessy, D.A., Wiesensthal, D.L., 1999. Traffic congestion, driver stress, and driver aggression. *Aggressive Behav.* 25, 409–423.
- Hennessy, D.A., Wiesensthal, D.L., 2001. Gender, driver aggression, and driver violence: an applied evaluation. *Sex Roles* 44, 661–676.
- Hennessy, D.A., Wiesensthal, D.L., Kohn, P.M., 2000. The influence of traffic congestion, daily hassles, and trait stress susceptibility of state driver stress: an interactive perspective. *J. Appl. Biobehav. Res.* 5, 162–179.
- Hennessy, D.A., Wiesensthal, D.L., Wickens, C.M., Lustman, M., 2004. The impact of gender and stress on traffic aggression. Are we really that different? In: James, P., Morgan, (Eds.), *Focus on Aggression Research*, 157–174.
- Hill, J.D., Boyle, L.N., 2007. Driver stress as influenced by driving manoeuvres and roadway conditions. *Transport. Res. Part F: Traffic Psychol. Behav.* 10 (3), 177–186.
- Hoare, P.N., 2001. Stress traits and coping styles as predictors of stress symptoms in bus drivers. Unpublished undergraduate thesis, University of Southern Queensland.
- Kashani, A.T., Besharati, M.M., 2019. An investigation on the relationship between demographic variables, driving behaviour and crash involvement risk of bus drivers: a case study from Iran. *Int. J. Occup. Saf. Ergon.*, doi:10.1080/10803548.2019.1603012.
- Kontogiannis, T., 2006. Patterns of driver stress and coping strategies in a Greek sample and their relationship to aberrant behaviours and traffic accidents. *Accid. Anal. Prev.* 38 (5), 913–924.
- Lajunen, T., Corry, A., Summala, H., Hartley, L., 1997. Impression management and self-deception in traffic behaviour inventories. *Personal. Indiv. Differ.* 22 (3), 341–353.
- Lajunen, T., Corry, A., Summala, H., Hartley, L., 1998. Cross-cultural differences in drivers' self-assessments of their perceptual-motor and safety skills: Australians and Finns. *Personal. Indiv. Differ.* 24, 539–550.
- Lajunen, T., Summala, H., 1995. Driving experience, personality, and skill and safety-motive dimensions in drivers' self-assessments. *Personal. Indiv. Differ.* 19, 307–318.
- Liu, B.S., Lee, Y.H., 2005. Effects of car-phone use and aggressive disposition during critical driving manoeuvres. *Transport. Res. Part F* 8 (4–5), 369–382.
- Lucas, J.L., Heady, R.B., 2002. Flexitime commuters and their driver stress, feelings of urgency, and commute satisfaction. *J. Bus. Psychol.* 16, 565–571.

- Machin, M.A., 2001. Evaluating a non-prescriptive fatigue management strategy for express coach drivers: a report prepared for the Australian Transport Safety Bureau.
- Machin, M.A., Hoare, P.N., 2008. The role of adaptive and maladaptive driver coping styles in predicting fatigue, affective well-being and physical strain in bus drivers. *Anxiety Stress Coping* 21, (4).
- Matthews, G., 1999. Individual differences in driver stress and performance. In: *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 579–583.
- Matthews, G., 2001b. A transactional model of driver stress. In: Hancock, P., Desmond, P. (Eds.), *Stress, Workload and Fatigue*. Mahwah, Erlbaum, pp. 133–163.
- Matthews, G., 2002. Towards a transactional ergonomics for driver stress and fatigue. *Theor. Issues Ergon. Sci.* 3, 195–211.
- Matthews, G., Desmond, P.A., 1998. Personality and multiple dimensions of task-induced fatigue: a study of simulated driving. *Personal. Indiv. Differ.* 25, 443–458.
- Matthews, G., Desmond, P.A., Joyner, L., Carcary, B., Gilliland, K., 1996. Validation of the Driver Stress Inventory and the Driver Coping Questionnaire. Unpublished report. Also presented at the International Conference on Traffic and Transport Psychology, Valencia, Spain, 1996.
- Matthews, G., Desmond, P.A., Joyner, L., Carcary, B., Gilliland, K., 1997. A comprehensive questionnaire measure of driver stress and affect. In: Rothengatter, T., Vaya, E.C. (Eds.), *Traffic and Transport Psychology: Theory and Application*. Pergamon, Amsterdam, pp. 317–324.
- Matthews, G., Dorn, L., Glendon, A.I., 1991. Personality correlates of driver stress. *Personal. Indiv. Differ.* 12, 535–549.
- Matthews, G., Dorn, L., Hoyes, T.W., Glendon, A.I., Davies, D.R., Taylor, R.G., 1993. Driver stress and simulated driving: studies of risk taking and attention. In: Grayson, G.B. (Ed.), *Behavioural Research in Road Safety III*. Transport Research Laboratory, Crowthorne, pp. 1–10.
- Matthews, G., Dorn, L., Hoyes, T.W., Davies, D.R., Glendon, A.I., Taylor, R.G., 1998. Driver stress and performance on a driving simulator. *Hum. Factors* 40, 136–149.
- Matthews, G., Emo, A.K., Funke, G.J., 2005. The transactional model of driver stress and fatigue and its implications for driver training. In: Dorn, L. (Ed.), *Driver Behaviour and Training*, Vol. II. Ashgate, Aldershot, pp. 273–285.
- Matthews, G., Sparkes, T.J., Bygrave, H.M., 1996. Stress, attentional overload and simulated driving performance. *Hum. Perform.* 9, 77–101.
- Matthews, G., Tsuda, A., Xin, G., Ozeki, Y., 1999a. Individual differences in driver stress vulnerability in a Japanese sample. *Ergonomics* 42, 401–415.
- Matthews, G., Joyner, L., Newman, R., 1999b. Age and gender differences in stress responses during simulated driving. In: *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1007–1011.
- Maxwell, J.P., Grant, S., Lipkin, S., 2005. Further validation of the propensity for angry driving scale in British drivers. *Personal. Indiv. Differ.* 38, 213–224.
- Parry, M.H., 1968. *Aggression on the Road*. Tavistock Publications, London.
- Qu, W., Zhang, Q., Zhao, W., Zhang, K., 2016. Validation of the driver stress inventory in china; relationship with dangerous driving behaviours. *Accid. Anal. Prev.* 87, 50–58.
- Rowden, P., Matthews, G., Watson, B., et al., 2011. The relative impact of work-related stress, life stress and driving environment stress on driving outcomes. *Accid. Anal. Prev.* 43 (4), 1332–1340.
- Shamoa-Nir, L., Koslowsky, M., 2010. Aggression on the road as a function of stress, coping strategy and driver style. *Psychology* 1, 35–44.
- Stokols, D., Novaco, R.W., Stokols, J., Campbell, J., 1978. Traffic congestion, type A behavior, and stress. *J. Appl. Psychol.* 63, 467–480.
- Susilowati, I.H., Yasukouchi, A., 2012. Cognitive characteristics of older Japanese drivers. *J. Physiol. Anthropol.* 31, 2.
- Taubman-Ben-Ari, O., 2008. Motivational sources of driving and their associations with reckless driving cognitions and behaviour. *Eur. Rev. Appl. Psychol.* 58 (1), 51–64.
- Taubman-Ben-Ari, O., Mikulincer, M., Gillath, O., 2004. The multidimensional driving style inventory—scale construct and validation. *Accid. Anal. Prev.* 36 (3), 323–332.
- Van Rooy, D.L., Rotton, J., Burns, T.M., 2006. Convergent, discriminant, and predictive validity of aggressive driving inventories: they drive as they live. *Aggressive Behav.* 32 (2), 89–98.
- af Wahlberg, A.E., 2010. Re-education of young driving offenders; effects on self-reports of driver behaviour. *J. Saf. Res.* 41 (4), 331–338.
- af Wahlberg, A.E., Dorn, L., 2015. How reliable are self-report measures of mileage, violations and crashes? *Saf. Sci.* 76, 67–73.
- Westerman, S.J., Haigney, D., 2000. Individual differences in driver stress, error and violation. *Personal. Indiv. Differ.* 29, 981–998.
- Wickens, C.M., Wiesenhal, D.L., 2005. State driver stress as a function of occupational stress, traffic congestion, and trait stress susceptibility. *J. Appl. Biobehav. Res.* 10 (2), 83–97.
- Wickens, C.M., Wiesenhal, D.L., Roseborough, J.E.W., 2015. In situ methodology for studying state driver stress: a between subjects replication. *J. Biobehav. Res.* 20 (1).
- Wickens, C.M., Wiesenhal, D.L., Roseborough, J.E.W., 2015. Personality predictors of driver vengeance. *Violence Victims* 30 (1), 148–162.
- Wishart, D., Somoray, K., Rowland, B., 2017. Role of thrill and adventure seeking in risky work-related driving behaviours. *Personal. Indiv. Differ.* 104, 362–367.
- Wohleber, R.W., Matthews, G., 2016. Multiple facets of overconfidence: implications for driving safety. *Transport. Res. Part F: Traffic Psychol. Behav.* 43, 265–278.
- Zhang, Y., Jin, S., Garner, M., Mosaly, P., Kaber, D., 2009. The effects of aging and cognitive stress disposition on driver situation awareness and performance in hazardous situations. In: *Proceedings of the 17th Triennial Congress of the International Ergonomics Association*, Beijing, China.

Further Reading

- Matthews, G., Desmond, P.A., 2001. Stress and driving performance: implications for design and training. In: Hancock, P., Desmond, P. (Eds.), *Stress, Workload and Fatigue*. Mahwah, Erlbaum, pp. 211–231.
- Nesbit, S.M., Conger, J.C., Conger, A.J., 2007. A quantitative review of the relationship between anger and aggressive driving. *Aggression Violent Behav.* 12 (2), 156–176.
- Navaco, R.W., Gonzalez, O.I., 2009. *Commuting and well-being*. In: Yair, A.-H. (Ed.), *Technology and Well-being*. Cambridge University Press.
- Öz, B., Özkan, T., Lajunen, T., 2010. Professional and non-professional drivers' stress reactions and risky driving. *Transport. Res. Part F: Traffic Psychol. Behav.* 13 (1), 32–40.
- Wiesenhal, D.L., Hennessy, D.A., Totten, B., 2000. The influence of music on driver stress. *J. Appl. Soc. Psychol.* 30, 1709–1719.

Page left intentionally blank

Introduction to Sustainability and Health in Transportation

Roger Vickerman, School of Economics, University of Kent, Canterbury, United Kingdom; Transport Strategy Centre, Imperial College, London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Sustainability of transport systems has for some time been largely seen as a question of the impact of transport on the environment both at a global level through greenhouse gas emissions and at a more local level through local air pollution, noise, and visual intrusion. Whilst these issues remain important and at the forefront of much research and an increasingly dominant topic in policy debates, the overall relationship between transport and health has become increasingly important. That transport affects the health of people in a negative way is part of the evaluation of the environmental impacts. But the accessibility of transport to some sections of the population, defined by such factors as age, gender or disability, and hence their access to a variety of activities from employment to leisure to education and healthcare is now an increasing concern. The choice of mode of transport may impact on health; this can be negative through choosing to walk in an environment polluted by traffic or positive through choosing to cycle along a growing network of segregated cycleways. Within modes the growing popularity of e-bikes or e-scooters demonstrates a desire to use cars less, but with some conflicting views on their health benefits, including accident risks.

The articles in this section present a range of views across these topics. All demonstrate the vibrancy of research and the importance of the policy debate at local, national, and international levels and outline the future priorities in both research and policy. Topics covered not only include the various environmental impacts of transportation, both global and local, and their health implications; but also major policy and planning initiatives to improve ways of life such as car-free cities and green space. Social aspects such as access to facilities, social exclusion and the problems facing specific groups of the population are important topics. Rising issues covered include the role of autonomous and connected vehicles and the implications of increased use of electrically powered vehicles, as well as the potential of shared mobility. More philosophical topics such as the concept of human ecology and the rising concern for environmental and transport justice are included as well as the way transport forms a major element in the UN Sustainable Development Goals. Many of the chapters have crosscutting themes with chapters in other sections of the Encyclopedia, especially those dealing with Psychology, Transport Planning and Policy and Safety and Security and readers are invited to explore these to gain a fuller understanding of the issues.

Sustainable Development Goals and Health

Rosa Surinach, United Nations Human Settlements Programme (UN-Habitat), Barcelona, Spain

© 2021 Elsevier Ltd. All rights reserved.

Sustainable Development Goals	226
The Global Goals	226
SDGs Monitoring and Reporting	227
The Decade of Action (2020–30)	228
Health and Development	228
Shanghai Declaration on Promoting Health in 2030 Agenda	228
COVID-19 Pandemic's Impact in Global Development	229
Urban Environment and Health	230
The New Urban Agenda	230
Transportation and Health	230
Lack of Safety on the World's Roads	232
Sustainable Transport	232
Relevant Websites	233
References	233
Further Reading	233

Sustainable Development Goals

The Sustainable Development Goals (SDGs) are a universal call to action to end poverty, protect the planet and improve the lives and prospects of everyone. The Goals were adopted by all United Nations Member States in 2015, as part of the 2030 agenda for sustainable development (United Nations, 2015), which set out a 15-year plan to achieve the goals.

The SDGs are composed by 17 goals and 169 targets, which are integrated and indivisible, global in nature, and universally applicable, taking into account different national realities, capacities and levels of development, and respecting national policies and priorities. The Goals recognize the link between sustainable development and other relevant ongoing processes in the economic, social, and environmental fields.

The SDGs replaced the Millennium Development Goals (MDGs), which started a global effort in 2000 to tackle extreme poverty and hunger. For 15 years, the MDGs drove progress in several important areas: reducing income poverty, providing much needed access to water and sanitation, driving down child mortality, and drastically improving maternal health. Most significantly, the MDGs made huge strides in combating HIV/AIDS and other treatable diseases such as malaria and tuberculosis.

The MDGs were revolutionary in providing a common language to reach global agreement, and its achievements provide the international community with valuable lessons and experience to begin work on the new goals. However, the MDGs only applied to developing countries, while the SDGs apply universally to all UN Member States, and they are considerably more comprehensive and ambitious than the MDGs.

The SDGs were the result of over two years of intensive public consultation and engagement with civil society and other stakeholders around the works, which paid particular attention to the voices of the poorest and the most vulnerable.

The Global Goals

- Goal 1. End poverty in all its forms everywhere
- Goal 2. End hunger, achieve food security and improved nutrition and promote sustainable agriculture
- Goal 3. Ensure healthy lives and promote wellbeing for all at all ages
- Goal 4. Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all
- Goal 5. Achieve gender equality and empower all women and girls
- Goal 6. Ensure availability and sustainable management of water and sanitation for all
- Goal 7. Ensure access to affordable, reliable, sustainable and modern energy for all
- Goal 8. Promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all
- Goal 9. Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation
- Goal 10. Reduce inequality within and among countries
- Goal 11. Make cities and human settlements inclusive, safe, resilient, and sustainable
- Goal 12. Ensure sustainable consumption and production patterns
- Goal 13. Take urgent action to combat climate change and its impacts

SUSTAINABLE DEVELOPMENT GOALS



Figure 1 Icons of the SDGs. Source: From United Nations Sustainable Development site un.org/sustainabledevelopment.

- Goal 14. Conserve and sustainably use the oceans, seas and marine resources for sustainable development
- Goal 15. Protect, restore and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, and halt and reverse land degradation and halt biodiversity loss
- Goal 16. Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels
- Goal 17. Strengthen the means of implementation and revitalize the global partnership for sustainable development (**Fig. 1**).

SDGs Monitoring and Reporting

The Sustainable Development Goals are not legally binding. Nevertheless, countries are expected to take ownership and establish a national framework for achieving the 17 goals. Implementation and success rely on countries own sustainable development policies, plans, and programs.

Besides financial implications, the most significant challenge to the universal implementation of the SDGs include capacities for progress monitoring. Actions at national and subnational level to monitor progress require quality, accessible and timely data collection and regional follow-up and review.

The UN provide the Global Indicator Framework for Sustainable Development Goals ([United Nations, 2020a](#)) that allow to monitor and review the 17 goals and 169 targets using a set of global indicators. Governments are requested to develop their own national indicators to assist in monitoring progress.

The follow-up and review process is informed by an annual SDG Progress Report to be prepared by the Secretary-General. The report provides an overview of the world's implementation efforts to date, highlighting areas of progress and where more action needs to be taken.

The SDG Report 2020 states that an estimated 71 million people are expected to be pushed back into extreme poverty in 2020, the first rise in global poverty since 1998. In addition, the effects of COVID-19 pandemic are already visible. The more than one billion slum dwellers worldwide are acutely at risk from the pandemic, suffering from a lack of adequate housing, no running water at home, shared toilets, little, or no waste management systems, overcrowded public transport and limited access to formal health care facilities. ([United Nations, 2020b](#)).

The high-level political forum on sustainable development is the annual meetings playing a central role in reviewing progress toward the SDGs at the global level. The Forum adopts intergovernmentally negotiated political declarations.

In addition, understanding that 65% of the SDGs could not be achieved unless local governments are fully and equitably involved in implementation (UCLG, UN-Habitat, 2020), since 2016 local and regional governments monitor and report progress they have made in localizing SDGs and the other global agendas. In 2018, the city of New York was among the first to produce voluntary local reviews (VLRs), organic self-assessing documents, which aimed to complete the information and data on SDG implementation that national governments are already providing, yearly, at the high-level political forum.

The Decade of Action (2020–30)

In September 2019, the UN Secretary-General called on society to mobilize for a decade of action to accelerate SDGs implementation. The decade of action (2020–2030) is deployed on three levels: global action to secure greater leadership, more resources, and smarter solutions for the SDGs; local action embedding the needed transition in the policies, budgets, institutions and regulatory frameworks of local governments; and people action, including by youth, civil society, the private sector, the media, unions, academia, and other stakeholders.

Health and Development

Health is a cornerstone of development. Ensuring healthy lives and promoting well-being at all ages is essential to global sustainable development. In 2017, only around one third to half of the global population was covered by essential health services. If current trends continue, only 39%–63% of the global population will be covered by essential health services by 2030 (United Nations, 2020b).

Recognizing this interdependence, the SDGs provide an ambitious and comprehensive plan of action for ending the injustices that underpin poor health. Goal 3 is focused on good health and well-being. It takes into account widening economic and social inequalities, rapid urbanization, threats to the climate and the environment, the continuing burden of HIV and other infectious diseases, and emerging challenges such as noncommunicable diseases (NCDs). Also, universal health coverage will be integral to achieving SDG 3, ending poverty, and reducing inequalities.

But health is also part of many of the 17 SDGs' indicators and targets—and, at the same time, action on health-related aspects of these SDGs is essential to achieve many of the targets for SDG 3. Following a research led by the Barcelona Institute for Global Health (ISGLOBAL), at least 48 SDG targets are relevant to urban health, corresponding to 15 SDGs, mainly captured by SDG 3 and SDG 11 on inclusive, safe, resilient, and sustainable cities (Ramirez-Rubio et al., 2019) (Fig. 2).

In this sense, SDGs provide a platform for intersectoral work facilitating a “Health in All Policies” (HiAP) approach. HiAP is based on the recognition that our greatest health challenges—for example, noncommunicable diseases, health inequities and inequalities, climate change, and spiraling health care costs—are highly complex and often linked through the social determinants of health. The social determinants of health are the circumstances in which people are born, grow up, live, work and age, and the wider set of forces and systems affecting these circumstances: for example, economic and development policies, social norms, social policies, and political systems.

In this context, promoting healthy communities, and in particular health equity across different population groups, requires that we address the social determinants of health, such as public transportation, education access, access to healthy food, economic opportunities, and more. While many public policies work to achieve this, conflicts of interest may arise. Alternatively, unintended impacts of policies are not measured and addressed. This requires innovative solutions, and structures that build channels for dialogue and decision-making that work across traditional government policy siloes (WHO, 2014).

Shanghai Declaration on Promoting Health in 2030 Agenda

The outcome of the Ninth Global Conference on Health Promotion (Shanghai, 21–24 November 2016), jointly organized by the Government of China and WHO, was the concise Shanghai Declaration on Health Promotion (WHO, 2016). It was endorsed by the over 1260 participants of the Conference from 131 countries, including 81 ministers and 123 mayors.

The Shanghai Declaration recognizes health and well-being as essential to achieving sustainable development. The principal message is that health is a political issue and, therefore, political commitments are crucial. In this sense, the declaration pretends to provide guidance to UN Member States on how to reflect promoting health into national SDGs responses and how to accelerate action.

Health promotion is about enabling and empowering people, communities, and societies to take change of their own health and quality of life. The declaration highlights three pillars of health promotion: *Good governance*: strengthening governance and policies to make healthy choices accessible and affordable to all; and to create sustainable systems that make whole-of-society collaboration real. *Healthy cities*: creating greener cities that enable people to live, work, and play in harmony and good health. *Health literacy*: increasing knowledge and social skills to help people to make the healthiest choices and decision for their families and themselves.

One of the recommended actions on the healthy cities area is to design our cities to promote sustainable urban mobility, walking and physical activity through attractive and green neighborhoods, active transport infrastructure, strong road safety laws, and accessible play and leisure facilities.



Figure 2 SDGs related to urban health within a HiAP approach (Ramirez-Rubio et al., 2019).

COVID-19 Pandemic's Impact in Global Development

In 2020, the world faced a global health crisis unlike any other. Coronavirus (COVID-19) pandemic spread human suffering, destabilizing the global economy, and upending the lives of billions of people around the globe.

The pandemic shown that in rich and poor countries alike, a health emergency can push people into bankruptcy or poverty; hence what began as a health crisis quickly became a human and socio-economic crisis.

In March 2020, the United Nations Secretary-General, António Guterres, reported how the COVID-19 crisis was likely to have a profound and negative effect on sustainable development efforts. A prolonged global economic slowdown may adversely impact the implementation of the 2030 Agenda for Sustainable Development and the Paris Agreement on Climate Change (United Nations, 2020c).

The most vulnerable, including women, children, the elderly, and informal workers, were hit the hardest. The impact on the environment, on the other hand, was likely to be positive on the short term, as the drastic reduction in economic activity reduced CO₂ emissions and pollution in many areas. However, unless countries deliver on their commitment to sustainable development once the crisis is over and the global economy restarts, such improvements are destined to be short-lived.

The United Nations Department of Economic and Social Affairs (UNDESA) mapped how COVID-19 affected each of the SDGs, from disruption to food supplies (SDG 2) to increased levels of violence against women (SDG 5). This conceptual mapping, while simple, shows the value of the SDGs as a framework for understanding the intersecting flow-on effects of a health crisis (Fig. 3).

While the crisis is imperiling progress toward the SDGs, it also makes their achievement all the more urgent and necessary. It is essential that recent gains are protected as much as possible. A transformative recovery from COVID-19 should be pursued, one that addresses the crisis, reduces risks from future potential crises and relaunched the implementation efforts to deliver the 2030 Agenda and SDGs during the decade of action.

Urban Environment and Health

Cities have a huge responsibility and opportunity to promote healthier urban environments. Urbanization is one of the leading global trends of the 21st century that has a significant impact on health. Since 2007, over half of the world's population live in urban areas, a proportion that is expected to raise 66% by 2050 (UN-Habitat, 2018). As most future urban growth will take place in developing countries, the world today has a unique opportunity to guide urbanization in a way that protects and promote health.

Urban areas will be increasingly critical for achieving all SDGs and integrating the social, economic and environmental goals set forth in the 2030 Agenda. It is, therefore, no surprise that SDGs 3 and 11 are strongly interlinked. Both goals have explicitly targeted improving road safety and air quality. Target 4 for cities directly links to health by aiming to reduce mortality due to disasters. Additionally, all targets of Goal 11 that aim to improve the living and working conditions of people in cities will support the achievement of the health goal.

As an example, 7 million people die every year from exposure to fine particles in polluted air. Air pollution is a major cause of NCDs. Venerable populations are often at risk from a variety of sources, which may expose them to pollutants including vehicle emissions and industrial discharges. In an analysis of data from 3000 cities, 80% of them did not meet the recommended WHO air quality guidelines, a critical parameter being the presence of fine particulate matter, produced mainly by vehicle diesel-engines. The 60% of European cities, 20% of North American cities, and 40% of high-income Asian cities, fail to meet adequate air quality standards (Clos and Surinach, 2019).

Furthermore, the increasing effects of climate change are now very evident. Climate change affects many of the social and environmental determinants of health—clean air, safe drinking water, sufficient food, and secure shelter. According to WHO, between 2030 and 2050, climate change is expected to cause approximately 250,000 additional deaths per year, from malnutrition, malaria, diarrhea, and heat stress. Reducing emissions of greenhouse gases through better transport, food, and energy-use choices can result in improved health, particularly through reduced air pollution.

Effective planning policies may significantly reduce this problem. Urban planning decisions can create or exacerbate major health risks for populations, or they can foster healthier environments, lifestyles, and create healthy and resilient cities and societies.

On the other hand, urban design can help to reduce inequalities both in service provision and health status and lead to more inclusive and productive societies. Bad examples such as limited access, poor quality, or absence of open public spaces or whole districts built without taking into account the need for people to be able to access local amenities by walking drive an increase of NCDs pandemic. Noncommunicable diseases already account for nearly 70% of global deaths each year with rapid and unplanned urbanization being a major factor. Expanding networks for urban public transit, walking and cycling, particularly for vulnerable groups, and having access to more options for safe mobility can increase access to jobs and services, reduce risks of road injury and limit noise exposure (WHO, UNDP, 2016).

The New Urban Agenda

In 2016, the United Nations adopted the New Urban Agenda at the UN Conference on Housing and Sustainable Urban Development (Habitat III). The New Urban Agenda incorporates a new recognition of the correlation between good urbanization and development, and of the fundamental linkages between health and sustainable urban development. It further clarifies the importance of these linkages, and that health is not only about the provision of health care services, but also reflects on decades of experiences and advances in our understanding of how the shape and form of urban development influences the health of city residents (United Nations, 2016).

The New Urban Agenda recognizes that effective urban planning, infrastructure development, and governance can mitigate risks and promote the health and well-being of urban populations. There is, however, a need to further address how specific urban policies can contribute to reducing disease burdens.

Transportation and Health

Transport plays an essential role in our societies and economies. It provides access to jobs, education, services, amenities and leisure, while contributing to economic growth jobs and trade. At the same time, it has an impact on the environment and human health.



Figure 3 COVID-19 affecting all SDGs (United Nations, 2020c).

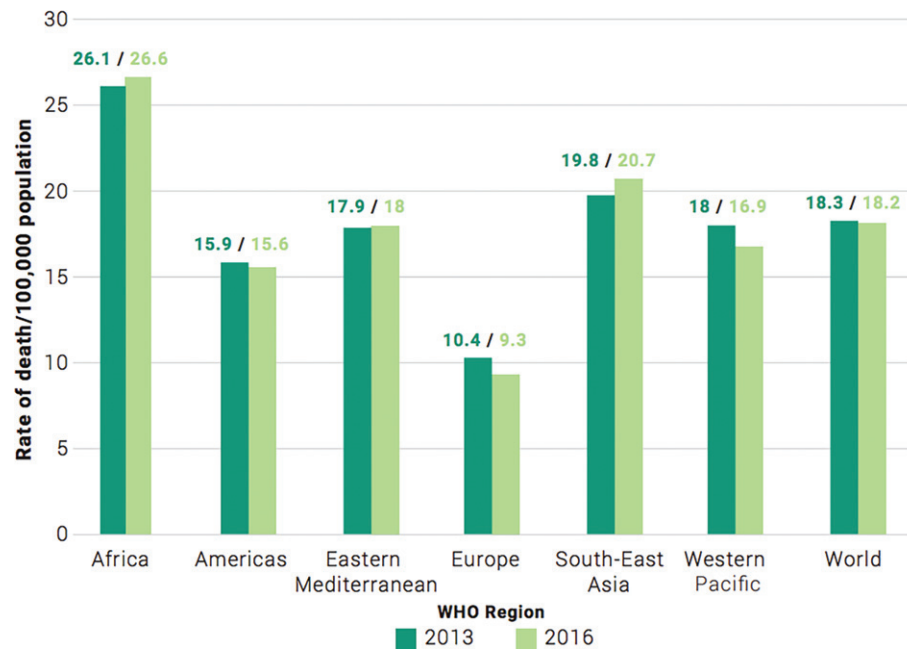


Figure 4 Rates of road traffic death per 100,000 (WHO, 2018).

Close to a quarter of global greenhouse gas emissions come from transport and these emissions are projected to grow substantially in the years to come, contributing to climate change. Other pollutants, most evidently in many urban centers, directly impact health; casualties and deaths from transport-related accidents are also on the rise (United Nations, 2020d).

Lack of Safety on the World's Roads

Approximately 1.35 million people die each year as a result of road traffic crashes and road traffic injuries are now the leading killer of children and young people aged 5–29 years.

According to the Global Status Report on Road Safety (WHO, 2018) the number of road traffic deaths continues to rise steadily. However, the rate of death relative to the size of the world's population remains constant. When considered in the context of the increasing global population and rapid motorization that has taken place over the same period, this suggests that existing road safety efforts may have mitigated the situation from getting worse.

Middle- and high-income countries reported progress. However, not a single low-income country demonstrated a reduction in overall deaths. In fact, the risk of a road traffic death remains three times higher in low-income countries than in high-income countries. The rates are highest in Africa (26.6 per 100,000 population) and lowest in Europe (9.3 per 100,000 population). In terms of regions, Americas, Europe and the Western Pacific are showing a decline in traffic deaths rates (Fig. 4).

The SDGs, in its target 3.6, call for a 50% reduction in the number of road traffic deaths by 2020. The current progress remains far from sufficient to achieve the target. In the Stockholm Declaration, ministers and representatives of international, regional and subregional governmental and nongovernmental organizations, and the private sector, recognize that SDG target 3.6 will not be met by 2020 and that significant progress can only be achieved through stronger national leadership, global cooperation, implementation of evidence-based strategies and engagement with all relevant actors including the private sector, as well as additional innovative approaches (Third Global Ministerial Conference on Road Safety: Achieving Global Goals 2030, 2020).

Sustainable Transport

The role of transport in sustainable development was first recognized at the 1992 United Nation's Earth Summit and reinforced in its outcome document—Agenda 21. The global attention to transport has continued in recent years. World leaders recognized unanimously at the 2012 United Nations Conference on Sustainable Development (Rio + 20) that transportation and mobility are central to sustainable development. Sustainable transportation can enhance economic growth and improve accessibility. Sustainable transport achieves better integration of the economy while respecting the environment, improving social equity, health, resilience of cities, urban–rural linkages, and productivity of rural areas.

In the 2030 Agenda for Sustainable Development, sustainable transport is mainstreamed across several SDGs and targets, especially those related to food security, health, energy, economic growth, infrastructure, and cities and human settlements.

Sustainable transport plays a fundamental role in overcoming the social exclusion of vulnerable groups. SDG target 11.2 explicitly calls on the international community to work toward sustainable transport for all people: "By 2030, provide access to

safe, affordable, accessible and sustainable transport systems for all, improving road safety, notably by expanding public transport, with special attention to the needs of those in vulnerable situations, women, children, persons with disabilities and older persons.”

It seeks to expand public transportation making it easy for vulnerable groups to use. Good mass transportation and high-density, compact urban design go hand in hand. Efficient planning will result in lower travel times and costs, especially if the right mix of walking and cycling facilities are combined with public transport and rapid transit systems. This must be done in such a way to reduce hazards for pedestrians and cyclists from motorized transport and, has been shown to greatly reduce the prevalence of NCDs. Encouraging less private motor vehicle use, impacts positively on air and noise pollution. Public transport can also help to promote equity in access to employment, educational and social amenities.

Evidence indicates that significant health gains can be made from restricting the use of diesel engines for both commercial and private vehicles in dense urban systems. The exhaust gases are carcinogenic, and emit a greater proportion of small particulate matter, which is associated with increased stroke, heart and respiratory disease, and early death (WHO, 2003).

In Mexico City, the reductions in air pollution resulted in a saving in 6100 days of lost work, together with reduced cases of bronchitis and deaths (WHO, UN-Habitat, 2016). Over the past 10 years, improved bus rapid transport systems serving around 1 million passengers a day, encouraged 10% of the population to leave their cars at home. Additionally, an innovative bike sharing program was developed which was integrated with the metro-rail and bus systems, offering one payment for all. Greater access to parks and green spaces was also made possible.

Providing access through sustainable transport will require advances in three key areas: policy development and implementation, financing, and technological innovation. If rigorous action is pursued in each of these areas, together they will drive change and help individuals, businesses and governments mobilize sustainable transport.

Relevant Websites

United Nations Sustainable Development: www.un.org/sustainabledevelopment

United Nations SDGs Indicators Framework: <https://unstats.un.org/sdgs>

United Nations Covid-10: www.un.org/coronavirus

New Urban Agenda: www.habitat3.org

World Health Organization Urban Health: <https://www.who.int/health-topics/urban-health>

References

- Clos, J., Surinach, R., 2019. Health Sustainable development goals and the new urban agenda. In: Nieuwenhuijsen, M., Khries, H. (Eds.), *Integrating Human Health into Urban and Transport Planning*. Springer, s.l., pp. 17–30.
- Ramirez-Rubio, O., et al., 2019. Urban health: an example of a “health in all policies” approach in the context of SDGs implementation. *Global Health* 15 (87), doi:10.1186/s12992-019-0529-z.
- Third Global Ministerial Conference on Road Safety: Achieving Global Goals 2030, 2020. Stockholm Declaration. Stockholm.
- UCLG, UN-Habitat, 2020. Guidelines for Voluntary Local Reviews. UN-Habitat, Nairobi.
- United Nations, 2015. *Transforming Our World: The 2030 Agenda for Sustainable Development*. United Nations, New York.
- United Nations, 2016. *The New Urban Agenda*. United Nations, Quito.
- UN-Habitat, 2018. *Tracking Progress Towards Inclusive, Safe, Resilient and Sustainable Cities and Human Settlements. SDG11 Synthesis Report – High Level Political Forum*. UN-Habitat, Nairobi.
- United Nations, 2020a. *Global Indicator Framework for the Sustainable Development Goals and Targets of the 2030 Agenda for Sustainable Development*. E/CN.3/2020/2. United Nations, New York.
- United Nations, 2020b. *The Sustainable Development Goals Report 2020*. United Nations, New York.
- United Nations, 2020c. *Shared Responsibility, Global Solidarity: Responding to the Socio-economic Impacts of COVID-19*. United Nations, New York.
- United Nations, 2020d. *Concept Note of the Second Global Sustainable Transport Conference Convened*. United Nations, New York.
- WHO, UNDP, 2016. *Noncommunicable diseases: what municipal authorities, local governments and ministries responsible for urban planning need to know*. World Health Organization.
- WHO, UN-Habitat, 2016. *Global report on urban health: equitable, healthier cities for sustainable development*. World Health Organization, UN-Habitat, Geneva.
- WHO, 2003. *Investing in Mental Health*. World Health Organization, Geneva.
- WHO, 2014. *Helsinki Statement on Health in All Policies. Contributing to Social and Economic Development: Sustainable Action Across Sectors to Improve Health and Health Equity*. World Health Organization, Helsinki.
- WHO, 2016. *Shanghai Declaration on Promoting Health in 2030 Agenda*. World Health Organization, Shanghai.
- WHO, 2018. *Global Status Report on Road Safety 2018*. World Health Organization, Geneva.

Further Reading

- Hook, W., Howe, J., 2005. *Transport and the Millennium Development Goals*.
- Grant, M., 2020. *Integrating Health in Urban and Territorial Planning: A Sourcebook*. UN-Habitat and World Health Organization.
- United Nations Secretary General's Advisory Group on Sustainable Transport, 2016. *Mobilizing Sustainable Transport for Development*. United Nations, New York.
- WHO, 2016. *Health as the Pulse of the New Urban Agenda*. World Health Organization, Quito.

Human Ecology

Roderick J. Lawrence, Geneva School of Social Sciences (G3S), University of Geneva Geneva, Switzerland

© 2021 Elsevier Ltd. All rights reserved.

Introduction	234
Theoretical Concepts and Framework	234
What is Human Ecology?	235
What is Health	237
Conclusion	238
Biography	238
References	238
Further Reading	239

Introduction

Mobility and transportation are fundamental constituents of human life. Both can be interpreted as interdisciplinary concepts that federate ideas and methods from several disciplines and professions. Historically, mobility and transportation are intentional processes that respond to fundamental human needs for secure and safe living conditions including protection from natural threats (e.g., avalanches, droughts, floods, and landslides), human threats (e.g., infectious diseases and warfare), as well as a sustained supply of food and water. The mobility of people goods and services not only responds to these essential needs that protect health and well-being, but also personal and group aspirations that express social status, group identity, and personal wealth, as well as specific kinds of lifestyles and ways of living.

Human ecologists have analyzed the transportation of goods, including commercial trade of agricultural produce in global food system (Dyball and Newell, 2015); and the mobility of humans including migration flows, residential mobility, and mass tourism (Fennell and Butler, 2003). An understanding of impacts of tourism on indigenous populations, land use and natural ecosystems, and local economic development, has been associated with environmental and human health impacts in tourist destinations.

There is a relationship between the viability and sustainability of human groups and the demands that they attribute to the natural resources of their habitat and the hinterlands (Lawrence, 2001). Given that natural and human-made ecosystems are finite, the sustainable uses of natural resources and the accumulation of waste products are only possible if human demands do not exceed flows of energy and matter into and out of ecosystems. During the 20th century there were important technological advances leading to more efficient uses of natural resources and lower levels of ejected waste. However, many have argued that, alone, technological innovation will not lead to sustainable development. This viewpoint is shared by several human ecologists, including the late Barry Commoner (1917–2012), and Stephen Boyden (1925–), as well as the authors of the Report of the World Commission on Environment and Development published in 1987.

There has been a large volume of empirical research on the carbon dependency of the transportation sector, and its large contribution to greenhouse gas emissions and ambient air pollution in cities Nieuwenhuijsen and Khreis (2019). These concerns have included impacts on population health and natural ecosystems in the framework of the Transport, Health, and Environment Pan-European Programme (PEP). From the 1990s, the PEP Secretariat led by the United Nations Economic Commission for Europe (UNECE) and the World Health Organization Regional Office for Europe (WHO-EURO) enlarged the public health framework beyond negative impacts of transportation, such as traffic accidents, to include concerns about relations between population health and lifestyles, notably the lack of physical activity of sedentary populations, especially children and young adults (van Schalkwyk and Mindell, 2018). These subjects can be studied independently from several disciplinary perspectives, as well as by interdisciplinary contributions including the natural and medical sciences, the humanities, and social sciences those that apply a human ecology perspective (Lawrence, 2001). Moreover, the PEP Secretariat has identified and promoted relations with the 2030 Agenda for Sustainable Development, by highlighting the areas where the PEP should intensify collaboration in implementing twelve sustainable development goals (SDGs) (World Health Organization, 2018).

The next section presents core principles of human ecology and distinguishes them from those of general and social ecology. Then principles of human ecology are used to formulate and apply a conceptual framework to analyze the impacts of mobility and transportation on individual, population, and planetary health.

Theoretical Concepts and Framework

Ernst Haeckel (1834–1919), a German zoologist, used the term “ecology” in 1866 to denote a science that studies multiple interrelations between organisms and their surroundings. Since the late 19th century, the term ecology has been interpreted in numerous ways including general, human, and social ecology (Lawrence, 2001).

The term “general ecology” was used by botanists and biologists during the last century to refer to the interrelations between animals, fungi, plants, and their immediate surroundings. The number of contributions about the science of ecology grew from the beginning of the 20th century. Animal and plant ecologists maintain that interactions between organisms and all the components of ecosystems follow principles that refer to their similarities and differences. A community of organisms develops from simple to more complex forms through a sequence of developmental stages known as succession. This term refers to the slow progression of changes in communities of animals and plants owing to changes in climatic, ecological, and geological conditions. This trend may become a climax state: Climax is a dynamic equilibrium state that is determined by the limiting factors of the climate, soil, or other ecological conditions (Pickett et al. 2001). Climax refers to the culmination of the evolution of animal and plant communities that correspond to the optimal development of the biomass with respect to specific ecological conditions. By using an analogy, some ecologists imply that human groups and communities are natural phenomena that develop by slow progression and succession processes. This interpretation means that psychological and social characteristics of human individuals and societies are analogous to biological factors; that competition between human beings is an innate biological process; and that climax is the outcome. The fundamental principles of human ecology challenge this analogy by accounting for the psychological, social, and cultural dimensions of human life including individual and collective behavior that can be adaptive, and capable of sustaining humans in adverse situations (Lawrence 2001).

What is Human Ecology?

Human ecology usually designates studies of the dynamic relationships between humans and the physical, biotic, cultural, and social characteristics of their environment and the biosphere. However, this is not the original meaning of this term which was first used by Ellen Swallow Richards (1842–1911) who founded ecofeminism. In her formulation of Euthenics, she proposed “human ecology” which she defined as a science for better living (Clarke, 1973). From an institutional perspective, human ecology developed in the context of a rapidly urbanizing city after the First World War in the Department of Sociology at the University of Chicago. It was championed by a coalition of researchers from several social science disciplines (including anthropology, demography, geography, psychology, and sociology). These researchers shared a concern about the effects of urban living on the daily life and well-being of the residents, especially minority groups of migrants and low-income households living in cities.

Today, human ecology usually refers to the study of the reciprocal relations between people, their habitat and the environment beyond their immediate surroundings. A conceptual model of human ecology formulated in Lawrence (2001) is reproduced in Fig. 1. This figure is not meant to be a descriptive model of people-environment relations that can provide a complete understanding of a complex and vast subject. Rather, it represents the systemic interrelations between sets of biotic, abiotic, and cultural factors that are combined in any human ecosystem. Hence, it does not concentrate only on specific components because it considers the whole system as the unit of study for people-environment relations. This integrated model can be applied to analyze different geographical areas from the scale of building sites to neighborhoods and cities. It is a synchronic representation of a human ecosystem that is open and linked to others; and it can account for processes that involve mobility and transportation. The model is meant to be applied at different times to explicitly address both short- and long-term perspectives. This temporal perspective can identify change to any of the specific components as well as the interrelations between them. For example, increasing mobility for leisure relies heavily on transportation by motor vehicles that have significant impacts on the consumption of fossil fuels, carbon dioxide emissions, ambient volumes of noise, and concentrations of ozone along major roads. These outcomes are known to have negative impacts on health.

Human ecology is explicitly interdisciplinary. The material and nonmaterial dimensions of human ecosystems, shown in Fig. 1, include genetic patrimony, especially the capacity of the human brain to interpret and transform land and other natural resources into a viable habitat; demographic characteristics such as the size and composition of human populations in mega-urban regions; the social organization of human groups in urban neighborhoods (including kinship relations and household structure); institutions including associations, rules and customs that regulate individual and collective behaviors; the local economy including all consumption and production processes; and, last but not least, the beliefs, knowledge, religion, and values of local populations (Lawrence, 2001).

The conceptual framework shown in Fig. 1 incorporates principles of coaction and coevolution, which should be applied in coordinated ways. These principles express mutual interactions between the ecological, biological, and cultural constituents inherent in human habitats. Consequently, if one of the constituents is changed, then others will also be modified too, because they are directly or indirectly related systemically. This conceptual framework enables an improved understanding of human habitats and contemporary mobility and transportation processes that could impact negatively on them. For example, buildings, villages and cities can be interpreted as metabolisms having systemic properties that are locality specific. The input and output of a system (e.g., a household or a community) can be interpreted as a metabolism, comprising a set of interrelated components, circumscribed by an open boundary with flows of material, and nonphysical entities across it (Dyball and Newell, 2015). The internal processes transform inputs into outputs, which are influenced by feedback loops and other regulatory mechanisms. The applications of this approach include material flow accounts and substance flow analysis. Since the 1990s, manmade ecological systems have been interpreted according to this kind of dynamic representation of material and energy flows which reply on mobility, energy, resources, and transportation. Another application is in the field of public health. Rutter et al. (2017) explain why a complex systems interpretation of health requires an understanding of the many variables that are interconnected.

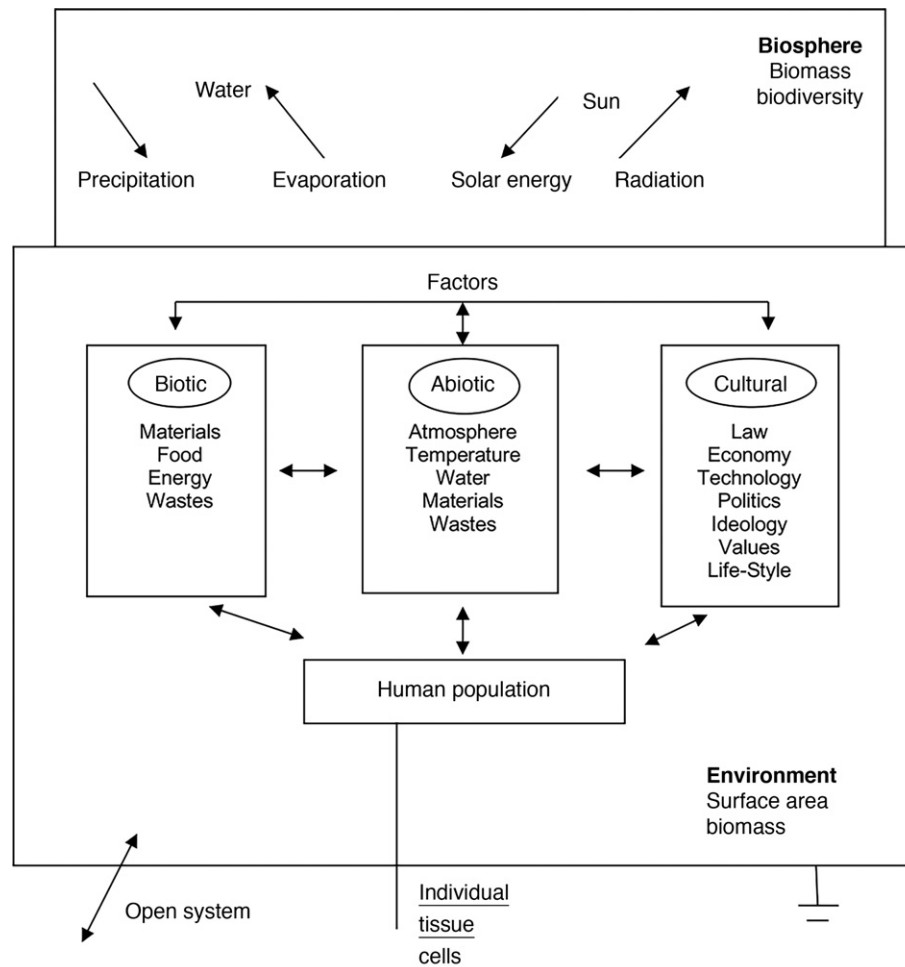


Figure 1 The holistic and systemic framework of a human ecology perspective showing the interrelations between biotic and abiotic factors, cultural, social, and individual human factors, and artefacts which are delimited by situations, habitats, or large ecosystem Source: [Lawrence, 2001](#)

A human ecology perspective encourages systemic thinking about the increasing incidence of obesity in many countries, extending simple approaches about changing characteristics of food production and consumption processes to include mobility, transport, and other characteristics of lifestyles that are influenced by advertising and marketing. A systemic approach admits that simple cause-effect interpretations that are common in epidemiological research are insufficient to understand the complexity of malnutrition and sedentary living in a rapidly urbanizing world. This means that intervention studies that consider modifying only one variable are unlikely to succeed.

An entry on Human Ecology is included in the UNESCO—Encyclopedia of Life Support Systems explained why the contribution of the Chicago School of Sociology used an inappropriate biological analogy to discuss the life, habitats, and reproduction of humans by comparing them with animals and plants ([Lawrence, 2001](#)). It also noted that the search for simple correlations between biological, economic, and geographical variables that excluded fundamental human values (including ethical values, worldviews, and political authority) could not explain the multiple meanings, geographical layout, and social organization of built environments, especially why different residential areas coexist in the same city.

Human ecology has been challenged owing to criticisms of contributions allied to the Chicago School, mainly between 1920 and the 1950s. These contributions, which could be termed the first generation of human ecology, were supplemented by a second generation, during the 1960s and 1970s. Gerald [Marten \(2001\)](#) explained there are fundamental differences between these two sets of contributions. The second generation includes the seminal contributions of Stephen Boyden, Barry Commoner, René Dubos, and Gerald Young, as well as international research programs, including the Man and Biosphere (MAB) Program coordinated by UNESCO for ecological studies of human habitats. This second generation of human ecology is distinctly different from the first owing to its concern with the “ecologic crisis” from local to global levels of the biosphere. Understanding this crisis requires an interdisciplinary, systemic framework like that applied using multiple research methods in specific cities including Hong Kong ([Boyden et al., 1981](#)). These contributions were the forerunners of recent socioecological systems research, including an increasing number of contributions known today as urban ecology and socioecological systems ([Ostrom, 2009](#)).

The term social ecology is questionable because of its ambiguous connotation. In essence, it does not refer explicitly to the unique capabilities of individual and collective human minds (especially the double articulation of written and spoken language not shared by primates); or human creativity to transform the bounty of nature into artifacts and built environments; or the capacity of the human mind to represent and simulate complex subjects or situations (including future cities). These intellectual capabilities of *homo sapiens* distinguish humans from all other living species (Boyden, 1992). Consequently, the term anthropologic, (rather than sociologic, also applicable to ants, animals and bees, for example) should be used when considering people-environment relations (Lawrence, 2001). This is the way human ecology is interpreted and represented in Fig. 1. It provides an alternative framework to biomedical explanations of individual and population health.

What is Health

Health is a dynamic condition resulting from the interrelations between the biological, chemical, economic, physical, and social environment of humans (Fig. 2). Health is place-based and should be considered as locality specific, not just population specific, and is influenced by changing living conditions. All the components of human habitats should be compatible with basic human needs and full functional activity, including biological reproduction over a long period. A key issue is how transport planning professionals position themselves critically in the context of increasing health and socioeconomic inequalities in cities (Litman, 2013).

Health is a fundamental human right, recognized in the Universal Declaration of Human Rights in 1948. Health is also an essential component of development, vital to a nation's economic growth and internal stability. Today, increasing health care costs, reduced expenditure on public medical services, and lack of universal health insurance, mean that health is increasingly influenced by wealth. Improved health plays a crucial role in reducing poverty (UN Habitat, 2010). In the field of health promotion, health is the capability of individuals and populations to achieve their potential and to respond positively to challenges of daily life. For example, human adaptability, a core concept of human ecology, is an intentional process to sustain health and well-being in situations of radical change, like the threat of individual and societal responses to epidemics, such as coronavirus in 2002-2003 and again in 2019-2020.

Health is an asset, and a resource for everyday life, rather than a standard, or outcome that ought to be achieved. This interpretation is appropriate for research about the multiple interrelations between health, human behavior, built and natural environments, because environmental and social conditions in specific localities do impact on human relations, they may induce stress, and they can have positive or negative impacts on health and well-being. This interpretation of health, founded on core principles of human ecology, implies that the capacity of the medical and healthcare sector to deal with health promotion and prevention is limited. Therefore, close collaboration with other sectors is necessary in order to improve health and well-being.

Health is both an individual and a collective condition that is influenced by decisions made by politicians and professionals in the field of the built environment, including housing, infrastructure, transportation, and land-use planning. These examples also confirm the need to “de-medicalize” common interpretations of sedentary lifestyles, obesity, and other societal disorders in order to account for the crucial role of the built environment, transportation, and active living in influencing health, well-being and life styles. Transportation policy and planning decisions can affect health and well-being in various ways. How people travel influences physical and mental health, including cancer, cardiovascular disease, diabetes, and accidental injury, notably four major causes of mortality (Litman, 2013). Resilience is an interdisciplinary concept used in the behavioral, medical, natural, and social sciences to explain how individuals and groups modify their behavior or components of human ecosystems in order to withstand the negative impacts of adverse conditions on their health and well-being.

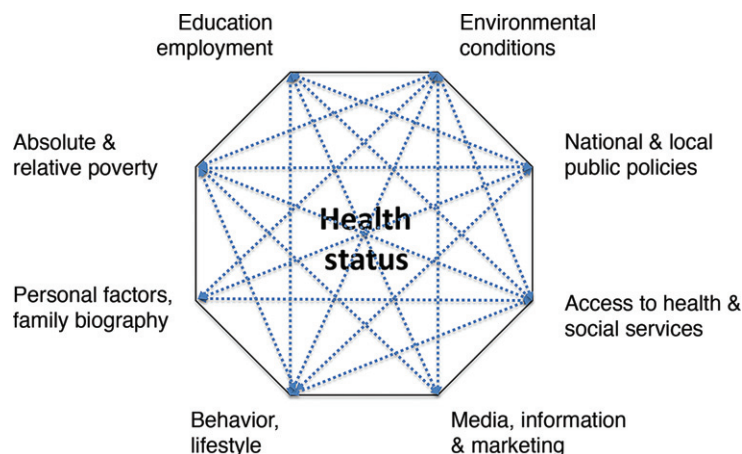


Figure 2 This diagram shows 8 sets of interrelated variables that influence health. Interdisciplinary research confirms that the availability and affordability of medical and health services explain only about 10% of those variables that influence health, whereas environmental conditions are estimated to represent about 25% of all variables. Source: © Roderick Lawrence

Ecological research and practice incorporate analyses of natural and human-made ecosystems that influence individual, group, and population health. Pioneers of ecological research on health that apply principles of human ecology include Stephen Boyden, Valerie Brown, Nancy Kreiger, and Anthony McMichael. The postulates of ecological public health emphasize that individual, group, socioeconomic, and environmental variables interact with others (e.g., public policy agendas) across different levels (Rayner and Lang, 2012). A holistic, systemic framework is a basic tenet shared by human ecology and ecological public health. This type of framework should be applied in interdisciplinary and transdisciplinary research. For example, Tretter and Löffler-Stastka (2019) apply core principles of human ecology to analyze relationships between individual and population health, household living, and ambient environmental conditions. Architecture, epidemiology, human ecology, geography, landscape planning, psychology, public health, sociology, statistics, and urban planning can converge and concert using innovative geographical information systems (GIS) and statistical methods to tackle the complexity of this subject. This innovative approach can improve current understanding and promote the application of new knowledge that can be used in policy definition and project implementation that explicitly accounts for health, mobility, and transportation.

Conclusion

The impacts of human societies on biophysical components of the biosphere and on humans themselves have been studied using principles of human ecology during the last 30 years. McMichael (2001), for example, analyzed adaptation processes used to maintain human health in response to negative influences from natural and manmade ecosystems. The term “evo-deviation” refers to a general biological principle that describes living conditions of human and other species that are different from those in their natural habitat. When these differences become large then unpredicted, perhaps irreversible, maladjustments may occur that impact negatively on health and well-being. Some physiological maladjustments during the last 10,000 years include diseases such as typhoid, cholera, smallpox, and more recently coronavirus and ebola.

Human societies have faced many demanding challenges in the context of rapid demographic, financial, political, and technological developments especially during the last two centuries. Today, societies are confronted by complex ecological, energy, and health challenges at local, regional, and global levels. If these challenges are to be confronted effectively, then societies need to address some fundamental issues, including their increasing mobility and greater reliance on transporting goods and services, because these have multiple, sometimes unforeseen environmental, financial, and social consequences, that impact negatively on health and well-being. Some of these driving forces (especially production and consumption processes that damage the ecological and biological components of ecosystems and human health) are challenging the equilibrium of symbiotic relations between human societies, their health and well-being, as well as living conditions provided by their habitat and the biosphere. This challenging situation is explicitly derived from the principle that anthropogenic factors have become major driving forces in the evolution of the biosphere and living conditions on Earth (Whitmee et al., 2015). Human ecology provides a comprehensive framework for understanding and responding to the public health challenges associated with mobility and transportation in a rapidly urbanizing world.

Biography



Roderick J. Lawrence, B. Arch (Hons), MLitt., DSc is Honorary Professor at the Geneva School of Social Sciences (G3S) at the University of Geneva since 2015. He is a member of the scientific advisory board of the Swiss Academies of Arts and Sciences “Network for transdisciplinary research” since 2009. He was Visiting Professor at the International Institute for Global Health at the United Nations University from 2014 to 2016. He was the founding Director of the Certificate of Advanced Studies in Sustainable Development at the University of Geneva from 2003 to 2015.

References

- Boyden, S., 1992. Bio-history: The interplay between human society and the biosphere. Past and Present. UNESCO, Paris.
- Boyden, S., Millar, S., Newcombe, K., O'Neill, B., 1981. The Ecology of a City and its People: The case of Hong Kong. Australian National University Press, Canberra.
- Clarke, R., 1973. Ellen Swallow: The Woman who Founded Ecology. Follet Publishing Company, Chicago IL.
- Dyball, R., Newell, B., 2015. Understanding Human Ecology: A Systems Approach to Sustainability. Earthscan, London UK.
- Fennell, D., Butler, R., 2003. A human ecological approach to tourism interactions. Int. J. Tour. Res. 5 (3), 197–210, doi:10.1002/jtr.428.
- Tretter, F., Löffler-Stastka, H., 2019. The human ecological perspective and biopsychosocial medicine. Int. J. Environ. Res. Public Health 16 (21), 4230, doi:10.3390/ijerph16214230.

- Lawrence, R., 2001. Human Ecology. In: Tolba, M.K. (Ed.), *Our Fragile World: Challenges and opportunities for sustainable development*, 1. EOLSS Publishers, Oxford, pp. 675–693.
- Litman, T., 2013. Transportation and Public Health. *Ann. Rev. Public Health* 34 (1), 217–233., doi:10.1146/annurev-publhealth-031912-114502.
- Marten, G., 2001. Human Ecology: Basic Concepts For Sustainable Development. Earthscan, London.
- McLaren, L., Hawe, P., 2005. Ecological perspectives in health research. *J. Epidemiol. Comm. Health* 59, 6–14, doi:10.1136/jech.2003.018044.
- McMichael, A., 2001. *Human Frontiers, Environments, and Disease: Past Patterns, Uncertain Futures*. Cambridge University Press, Cambridge UK.
- Nieuwenhuijsen, M., Khreis, H. (Eds.), 2019. *Integrating Human Health into Urban and Transport Planning*. E. Springer, Cham CH., pp. 89–112.
- Ostrom, E., 2009. A general framework for analysing sustainability of social-ecological systems. *Science* 325, 419–422, doi:10.1126/science.1172133.
- Pickett, A., Cadenasso, M., Grove, J., Nilson, C., Pouyat, R., Zipperer, W., Costanza, R., 2001. Urban ecological systems: Linking terrestrial, ecological, physical and socioeconomic components of metropolitan areas. *Ann. Rev. Ecol. Sys.* 32, 127–157.
- Rayner, G., Lang, T., 2012. *Ecological Public Health: Reshaping the Conditions of Good Health*. Routledge, Abington UK & New York.
- Rutter, H., Savona, N., Glanti, K., et al., 2017. The need for a complex systems model of evidence for public health. *Lancet* 390 (10112), 2602–2604, doi:10.1016/S0140-6736(17)31267-9.
- UN Habitat, 2010. *Hidden cities: Unmasking and overcoming inequalities in health in urban areas*. World Health Organization, Geneva.
- van Schalkwyk, M., Mindell, J., 2018. Current issues in the impacts of transport on health. *Brit. Med. Bullett.* 125 (1), 67–77, doi:10.1093/bmb/idx048.
- Whitmee, S., Haines, A., Beyrer, C. et al., 2015. *Safeguarding Human Health in the Anthropocene Epoch: Report of The Rockefeller Foundation–Lancet Commission on planetary health*. The Lancet (published online July 16) [http://dx.doi.org/10.1016/S0140-6736\(15\)60901-1](http://dx.doi.org/10.1016/S0140-6736(15)60901-1).
- World Health Organization Regional Office for Europe, 2018. *Making the (Transport, Health and Environment) Link*. World Health Organization, Copenhagen DK.

Further Reading

- V. BrownJ. HarrisJ. RussellTackling Wicked Problems Through the Transdisciplinary Imagination2010EarthscanLondon.
- B. CommonerThe Closing Circle: Nature, Man, and Technology1971Alfred A. KnopfNew York.
- N. KriegerEpidemiology and the web of causation: Has anyone seen the spider?Soc. Sci. Med.391994887903.
- Lawrence, R., 2019. Human ecology in the context of urbanization. In Nieuwenhuijsen, M., Khreis, H. (Eds.), *Integrating Human Health into Urban and Transport Planning*. Springer, Cham CH, pp. 89–112. https://doi.org/10.1007/978-3-319-74983-9_6.
- R. LawrenceJ. ForbatJ. ZuffereyRethinking conceptual frameworks and models of health and natural environmentsHealth2622019158179.doi:10.1177/1363459318785717.
- D. StokolsSocial Ecology in the Digital Age: Solving Complex Problems in a Globalizing World2018Academic PublishersLondon.
- F. WhiteL. StallonesJ. LastGlobal Public Health: Ecological Foundations2013Oxford University PressOxford.
- World Commission on Environment and Development (WCED), 1987. *Our common Future*. Report of the Commission on Environment and Development. United Nations, New York..

Car-Free Cities

Haneen Khreis^{*,†}, Mark J. Nieuwenhuijsen^{*,‡,§}, ^{*}Center for Advancing Research in Transportation Emissions, Energy, and Health (CARTEEH), Texas A&M Transportation Institute (TTI), Texas, United States; [†]ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain; [‡]Universitat Pompeu Fabra (UPF), Barcelona, Spain; [§]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

© 2021 Elsevier Ltd. All rights reserved.

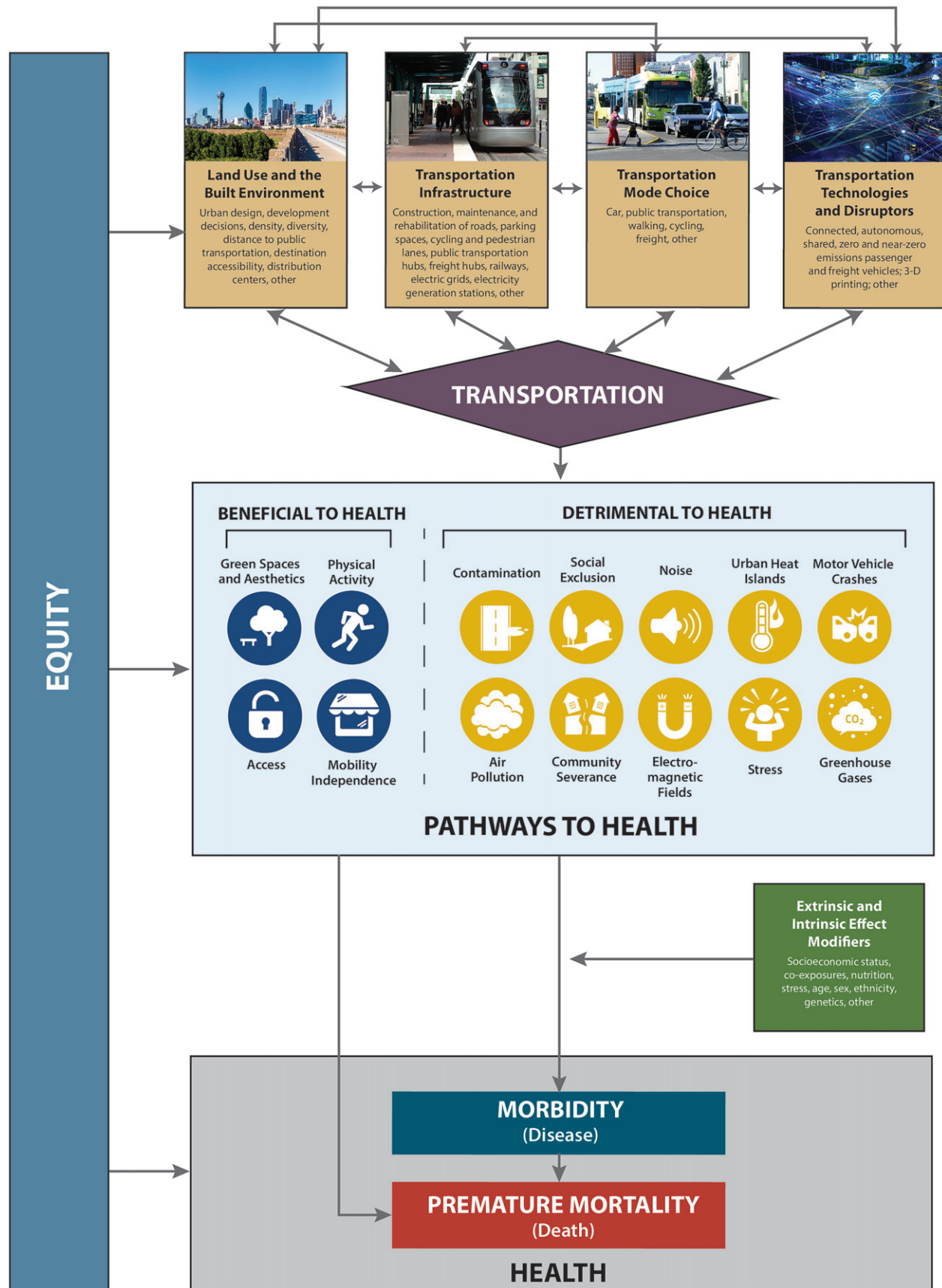
Introduction	240
Early Developments Toward Car-Free City Advocacy	242
Car-Free Initiatives From Around the World	242
Can We Retrofit Cities and Change Mode Share?	243
Key Reasons for Transitions	243
How to Get There?	244
What Holds us Back From Going Car-Free?	246
References	247

Introduction

Cities and their transportation systems have been profoundly changed and shaped by cars. At times it appears that cities cater more for cars rather than for citizens. The car is an important part of almost every transportation system that has brought increased access, mobility and convenience, and increased employment opportunities and economic growth. Many cities and their transportation systems have been developed to accommodate car travel as presented in large and continued investments in for example, the construction, maintenance, and rehabilitation of roadways, bridges and parking spaces. In turn, these decisions led to and reinforced lasting changes to land-use patterns such as urban sprawl (Frumkin, 2016). In many cities, car travel and associated infrastructure keep growing, as transportation planning and policy remain demand-driven. Planners tend to predict car travel and provide required infrastructure to facilitate car mobility, underlined by the belief that this mobility boosts the economy, although the role that further road construction plays in the economy is not so clear (Hensher and Button, 2001). This prevailing planning pattern creates a vicious circle that is hard to break as more car infrastructure leads to more car use, and results in less investment in other forms of mobility. In cities, cars are the most space-intensive form of transportation, both during use (Héran and Ravalet, 2008), and when not in use such as when parking (McCahill and Garrick, 2012). For example, cars consume as much as 70% of land in some American cities' downtowns (Crawford, 2000), and the average car spends about 95% of its life parked. The total number of vehicles in the world is estimated at 1.2 billion and projected to increase to 2.5 billion by 2050 (Voelcker, 2020). The investment in car travel and infrastructure has been enormous and tends to receive more financial support, including subsidies, than other modes of transportation.

While car travel brought along many benefits to some segments of the society, it also is responsible for significant adverse impacts on the environment and human health (Khreis et al., 2019; 2016). A recent study conceptualized the linkages between urban transportation and health and found that health impacts can occur through fourteen interrelated pathways; four of which can be beneficial to human health, and ten of which can adversely impact human health; resulting in premature mortality and increased morbidity rates across a wide range of health outcomes (Khreis et al., 2019). These linkages are illustrated in Fig. 1. In particular, relying on the private car for transportation can take up space in cities, reducing the amount of blue and green spaces, which have been both linked to many beneficial health effects. It can also reinforce and increase sedentary lifestyles, resulting in less physical activity which has been linked to many adverse health effects. As compared to active (walking and cycling) and public transportation, car travel results in more contamination, social exclusion, noise, urban heat islands, motor vehicle crashes, air pollution, stress, and greenhouse gases, which result in climate change (Khreis et al., 2019). Car travel can also result in community severance which refers to transportation infrastructure and/or motorized traffic (speed or volume of traffic) which divide places and people, interfering with the ability of individuals to access goods, services, and personal networks (Khreis et al., 2019). These linkages have been reviewed in details elsewhere (Khreis et al., 2019).

The public health burden which results from these linkages is significant, but modifiable. For example, each year, over 1.3 million deaths and 78 million injuries warranting medical care are due to motor vehicle crashes globally (Bhalla et al., 2014). Transportation-related noise and motor vehicle crashes were estimated to result in 1.7% and 1.9% of all-cause premature deaths in Houston, Texas, respectively, which is comparable to the death rate resulting from influenza, pneumonia or suicide in the United States (Sohrabi and Khreis, 2020). Also in Houston, air pollution in the form of fine particulate matter with a diameter of 2.5 micrometers or less (PM_{2.5}) was estimated to result in 7.3% of all-cause premature deaths in 2010, which is higher than the death rate associated with diabetes mellitus, Alzheimer's disease, or motor vehicle crashes in the United States (Sohrabi et al., 2020). In conservative global estimates, 184,000 deaths a year were attributable to traffic-related air pollution (Bhalla et al., 2014). Lelieveld and coworkers estimated that land transport-related air pollution is responsible for one-fifth of deaths from air pollution in the



© 2019 Texas A&M Transportation Institute

Fig. 1 Transportation and health conceptual model. Source Khreis et al. (2019).

United Kingdom, the United States, and Germany (Lelieveld et al., 2015), while recent estimates attributed a quarter of new childhood asthma cases per year in an English urban area to traffic-related air pollution (Khreis et al., 2018). Mueller and coworkers estimated that 20% of premature mortality annually is due to reduced physical activity, high exposure to air pollution, noise, and heat; and lack of access to green space in Barcelona, Spain, due to suboptimal urban and transportation planning (Mueller et al., 2017). Even though the mode share of car travel in Barcelona is around 20%, 60% of public space is taken up by car infrastructure such as roads and parking, which could be used in a more beneficial way for human health (BBC News, 2020).

Although it may be difficult to imagine a world without cars, many cities are beginning to shift their transportation solutions away from the private cars and toward active and public transportation, realizing that these latter modes are more environmentally friendly, healthier, and people oriented. The global climate change crisis has also been an important driver and a reason that is often cited by cities in this context. Some cities now refer to such solutions as car-free solutions, which for example include the implementation of car-free city centers, car-free designated areas, and car-free days.

Early Developments Toward Car-Free City Advocacy

A theoretical design for a car-free city of one million people was first proposed by J.H. Crawford in 1996 and further refined in his books, *Car Free Cities* (Crawford, 2000), and *Car Free Design Manual* (Crawford and Dimas, 2009). Later in 1997, the Lyon Protocol was published for the design and implementation of large car-free districts in existing cities (<http://www.carfree.com/lyon.html>). Created during a weeklong conference in Lyon, France, participants envisioned what would be needed to implement car-free districts. This included (1) identifying interested parties, (2) gathering necessary data, for example, on land-use, traffic flows, land ownership, demographics and air, noise, and water pollution, (3) developing preliminary concepts and proposals, (4) facilitating dialogue by building media attention, (5) engaging the political process and building a coalition to support implementing large car-free areas, (6) gradually implementing coherent car reduction measures, (7) engaging with, consulting, and listening to local communities, followed by (8) city planners developing final plans for full implementation. Although not specifically mentioning car-free cities, work by Newman and Kenworthy highlighted many of the problems with car dependence and the benefits of moving away from such development patterns (Newman and Kenworthy, 1999).

The concept of a car-free city is not a city free of motorized vehicles and this vision is agreed on by other authors (Loo, 2018). There are significant advantages related to motorized vehicles such as the movement of heavy goods, aiding immobile people, emergency medical services, and traveling long distances. The key point is effectively eliminating the need for and use of *private* motorized vehicles, especially in cities, which are already hotspots for adverse environmental exposures and health effects related to road transportation (Khreis et al., 2017). To reduce or even eliminate the need for private vehicles in cities, numerous steps are necessary. Most importantly, these steps involve providing and improving alternatives and in the long-term, adjusting city form and land-use to support active travel and (private) car-free living. However, to make such “hard” changes, numerous “soft” processes [see for example (Loo, 2018)] are required as the attitudes, beliefs, and knowledge of decision-makers and the general population have for decades been sold on the advantages of private cars without sufficiently considering its negative implications.

Car-Free Initiatives From Around the World

While there is no agreed-on definition or vision for what a car-free city should look like, examples from around the world give insight into cities’ visions, plans, and already implemented changes. Hamburg, Helsinki, and Oslo have recently announced their plans to become partly private car-free cities (Cathcart-Keays, 2020). Hamburg aims to develop a network of pedestrian and cycling paths to link together different parks and public spaces in the city, while enforcing a car ban on a number of urban roads (Cathcart-Keays, 2020). Helsinki aims to be car-free by 2025, creating a novel mobility-on-demand transportation system, where all shared (including bike and car share) and public transportation options can be utilized through a single payment option using smart-phones (ASLA, 2025), which may render private cars obsolete (Greenfield, 2014). The city’s vision is to make an array of transportation options, beyond the private car, flexible, cheap and well-coordinated, increasing their convenience and ease of use, so that they become competitive with private car ownership (Greenfield, 2014). Some new apartment developments as well, such as the Helsinki’s Kalasatama neighborhood, come with no parking (Peters, 2020). Similarly, Utrecht in the Netherlands is building a new high density, car-free neighborhood for 12,000 people (Boffey, 2020). Oslo made its city center car-free, completely banning some cars on some streets, removing 700 parking spots and replacing them with bike lanes and a different array of green spaces including plants and tiny parks (Peters, 2019). The city is also adding new trams and metro lines with more frequent departures and cheaper tickets (Peters, 2019). Similarly, York in the United Kingdom announced plans to ban most vehicles from its medieval center by 2023, adding an on-demand shuttle service and more bikes to help in the transition (Peters, 2020). Finally, Pontevedra is a small car-free city in Spain. Cars are banned from Pontevedra’s city center, creating a model for a pedestrian-friendly future. In the car-free zone, carbon dioxide (CO₂) emissions have been cut by 70%, nearly three quarters of car journeys were shifted to walking or cycling, and walkers are free to roam (Borgen, 2018).

Other cities have taken less drastic steps, in the same direction, which ban cars on specific days or only ban certain cars which for example do not meet certain emission standards (i.e., more polluting). In a plan known as Madrid Central, Madrid enacted a low-emission zone which bans all petrol vehicles registered before 2000 and all diesel vehicles registered before 2006, unless they meet specific exemptions or are used by residents of the area (Jones, 2018). Brussels, Dublin and London are also discussing banning

diesel vehicles in their city center, while every weekday Milan is offering public transportation tokens to riders who surrender their cars (Cathcart-Keays, 2020). Spearheaded by Mayor Anne Hidalgo, cars are banned from accessing the city center in Paris the first Sunday of every month, increasing access to cyclists, pedestrians and people who use rollerblades and scooters (Coffey, 2020). This ban excludes delivery vehicles, public transportation means and taxis who are still allowed to enter and leave the city center through designated access points and at a maximum speed limit of 20 kilometers per hour (Coffey, 2020). Hidalgo has also been championing the idea of a 15 minutes city, where all Parisians can walk or bike to works and daily errands within 15 minutes (Peters, 2020). Los Angeles adopted the South American-born idea “Ciclovía,” which is a car-free day that closes streets to traffic and sponsors an event to promote active transportation instead. During the program, known as “CicLAVia,” ultrafine particle (UFP) and PM_{2.5} concentrations were measured and found to be reduced by 21% and 49%, respectively, on the streets where cars were banned (Shu et al., 2016). The community wide reductions in PM_{2.5} were less at 12% (Shu et al., 2016). New developments in Utrecht in The Netherlands do not allow residents to own a car but provide neighborhood shared cars for farther trips (Peters, 2020), which is similar to a new development called Culdesac in Tempe, Arizona, which includes basic amenities onsite and is purposefully located near a light rail station that takes commuters to downtown Tempe (Peters, 2020). Car-free developments are becoming more common and can also be seen in England (Morris et al., 2009), and other cities across Europe, most notably including the German neighborhood: Freiburg (Melia, 2006). The district of Vauban in Freiburg, a neighborhood with more than 5000 residents on a former military base, represents that largest car-free development in Europe (Vauban Freiburg Wikipedia, 2020). The tram system in the city provides the backbone for transportation, where 70% of all public transportation trips are made by tram, and 30% are made by bus (Melia, 2006). In this neighborhood, cycling is the predominant mode of transportation where 75% of the working population cycles to work and only 16% of trips are made by car (Melia, 2006). Other cities including Bogota, Brussels, Chengdu, Copenhagen, Hyderabad, Ghent and Rome all have enacted or proposed different measures that aim at reducing car traffic including implementing car-free days, investing in cycling infrastructure and pedestrianization, restricting parking space and considerable increases in public transportation and green space provision (Cathcart-Keays, 2020; Peters, 2020).

Can We Retrofit Cities and Change Mode Share?

Compact cities like Spanish cities may be easier to refit to car-free cities than sprawled cities (Nieuwenhuijsen, 2020). The main challenges will be how to change existing infrastructure that is mainly used by cars to infrastructure for active and public transportation, and how to change people’s perceptions, attitudes, and behaviors and ultimately long-term lifestyles. Some cities have initiated strategies to create car-free spaces like the Superblocks in Barcelona (see below), and these could be transferable to other cities.

A large number of car trips are less than five kilometers (as high as 50%), and those could be replaced by more sustainable and healthier modes of transportation such as cycling (Nieuwenhuijsen, 2020). Cycling has many benefits as it increases physical activity and reduces adverse health outcomes such as premature mortality, it combines transportation with the gym, where on average cyclists weigh two kilograms less than car drivers, it does not cause air and noise pollution, it emits near zero CO₂, it uses much less space than the car, where one car can occupy the same parking space as ten bicycles, and cyclists tend to be happier than other transportation users, with better mental and physical well-being and less stress (Barcelona Institute for Global Health, 2020).

Given the relatively small transportation mode share of cars, Spanish cities (Table 1) are moving toward going car-free, but unfortunately the streets are still dominated by cars. This is partly because even though the mode share is relatively small, there are still hundreds of thousands of cars on the road, and because each car takes up a lot of space. Cities still spend too much money on keeping car drivers happy, and too little on other road and transportation mode users. Therefore, a more concentrated effort is needed to push for a reduction in car use and create car-free streets and neighborhoods in a move towards a car-free city. Also, any new urban development should be car-free and provide good alternatives. Supporting the local economy and creating 15-minute cities with dense and mixed land-use are essential (O’Sullivan, 2020).

Key Reasons for Transitions

The key reasons cited for car-free transitions by cities are to reduce greenhouse gases and mitigate the catastrophic effects of climate change, in addition to reducing air pollution (Nieuwenhuijsen and Khreis, 2016). Indeed, there is evidence that car-free transitions

Table 1 Mode share (%) in Spanish cities (2018)

	Active modes	Public transport	Private transport	Others
Barcelona	46.4	33.3	20.3	0.0
Madrid	40.0	35.0	20.0	5.0
Sevilla	32.3	20.1	43.4	4.2
Bilbao	54.5	22.3	23.2	0.0
Valencia	38.5	22.5	37.3	1.7

Source: Table prepared by Guillem Vich, Barcelona Institute for Global Health

can help achieve these goals. Measurements show that nitrogen dioxide (NO₂) levels, a pollutant which has been associated with many adverse health effects, dropped by 40% during car-free days in Paris (Willsher, 2015), and 20% in Leeds during a quasi-car-free day during the Tour de France (Nieuwenhuijsen and Khreis, 2016). Certain car-free developments cite a wider multitude of reasons including livability, and improved quality of life which is a municipal objective in Freiburg (Melia, 2006). As reviewed in detail elsewhere, the beneficial effects of such transitions would extend well beyond reducing greenhouse gases and air pollution to potentially increasing green space exposures, physical activity, access and social interactions, alongside reducing traffic noise, urban heat, stress and road travel injuries (Nieuwenhuijsen and Khreis, 2016), all associated with improved public health. For example, on Paris's car-free days, measurements show that noise dropped by 50% in the city center (Willsher, 2015). Another health impact assessment exercise modeled the potential environmental and health benefits of the Superblocks model in Barcelona, Spain. While the Superblocks are not exactly a car-free initiative, they do share various elements with the initiatives outlined above including restricting car use in residential blocks to residential traffic only with a maximum speed of 20 kilometers per hour and expanding the active and public transportation network, while increasing the provision of green and open space including plazas, parks, green corridors, green patches and general greening in and outside the Superblocks (Mueller et al., 2019). The health impact assessment estimated that 667 premature deaths could be prevented annually with the implementation of 503 Superblocks covering the city of Barcelona, where the greatest number of avoided deaths is due to reductions in traffic-related NO₂ (291 deaths), followed by noise (163 deaths), heat (117 deaths), green space development (60 deaths), and the increase in physical activity due to the shift from car and motorcycle trips to active and public transportation (36 deaths) (Mueller et al., 2019).

It is important to note that these benefits may be confined to the city centers or areas designated as car-free and that there are valid concerns that car traffic, and the associated adverse exposures, would be displaced to other areas, instead of being eliminated. Examples of this has been shown in the case of air pollution exposure on car-free days in the eastern Po Valley in Italy (Masiol et al., 2014). Research from that region (Masiol et al., 2014) and elsewhere (Burgen, 2018; Mueller et al., 2019) suggests that traffic restrictions should be imposed simultaneously on a larger scale to produce broader benefits to air quality and others, and avoid the exacerbation of environmental injustice, which is tightly linked to transportation planning and policy (Fuller and Brugge, 2020; Mueller et al., 2018). In addition, there are concerns that as the quality of life improves in certain areas where the car-free initiatives take place, gentrification will occur in these areas resulting in rising rents and pushing lower socioeconomic status populations out (known as forced migration) (Mueller et al., 2019). In order to mitigate such risks, traffic restrictions are best implemented city wide with a consistent and equitable distribution.

How to Get There?

While there is no clear consensus or definition of what a car-free city should look like, it seems like a good start for cities is to ban personal car travel within their boundaries, while excluding other essential vehicles like emergency vehicles and also potentially delivery vehicles which meet certain emission standards, keeping in mind that heavy-duty traffic is a key contributor to urban air pollution. The vision for a post-automobile city can go in two directions which are evident in cities' plans from around the world (Bieda, 2016). Cities can completely go car-free, which entails a city model completely devoid of cars. On the other hand, cities can substantially reduce car use, including having designated car-free areas and offering convenient, affordable and competitive alternatives that would ultimately render personal vehicles obsolete. Changes can be incremental, and some cities are already toying with the idea of becoming partly car-free with the implementation of small steps towards that goal including car-free days, low-emission zones, the development of car-free neighborhoods, introducing concepts like the 15-minute city and the Superblocks, as briefly overviewed above. These incremental changes may be a good approach to secure public acceptability as citizens start experiencing the beneficial and pleasant effects of car-free initiatives including improved air quality, reduced noise, and increased physical activity and opportunities for social interactions and cohesion.

To reduce or eliminate the need for personal car travel in cities, many steps need to be taken. Nieuwenhuijsen et al. (2019) suggested nine essential prerequisites underpinning the transition to car-free cities, as illustrated in Fig. 2.

1. Political vision and leadership
2. Mobility to accessibility paradigm shift
3. Alternative convenient and quality transportation means
4. Dedicated funding
5. Media strategy and public involvement and acceptability
6. Intensive data collection and analysis
7. Evaluation of current status, alternative scenarios, and post-evaluation of policies' impacts
8. Stakeholders' involvement and support
9. Detailed plan aligned with other high-level objectives and strategies.

Most importantly, cities need to provide convenient, affordable, and competitive transportation alternatives to the private car such as high-quality public transportation options that are frequent enough, easy to use and affordable, and safe segregated pedestrian and cycling paths, which are well-connected and aesthetically pleasing by for example adding green or blue spaces. Both active and public transportation networks need to be integrated as their integration can potentially allow for the full substitution of car trips by for example covering long distances by public transportation options and short distances such as the

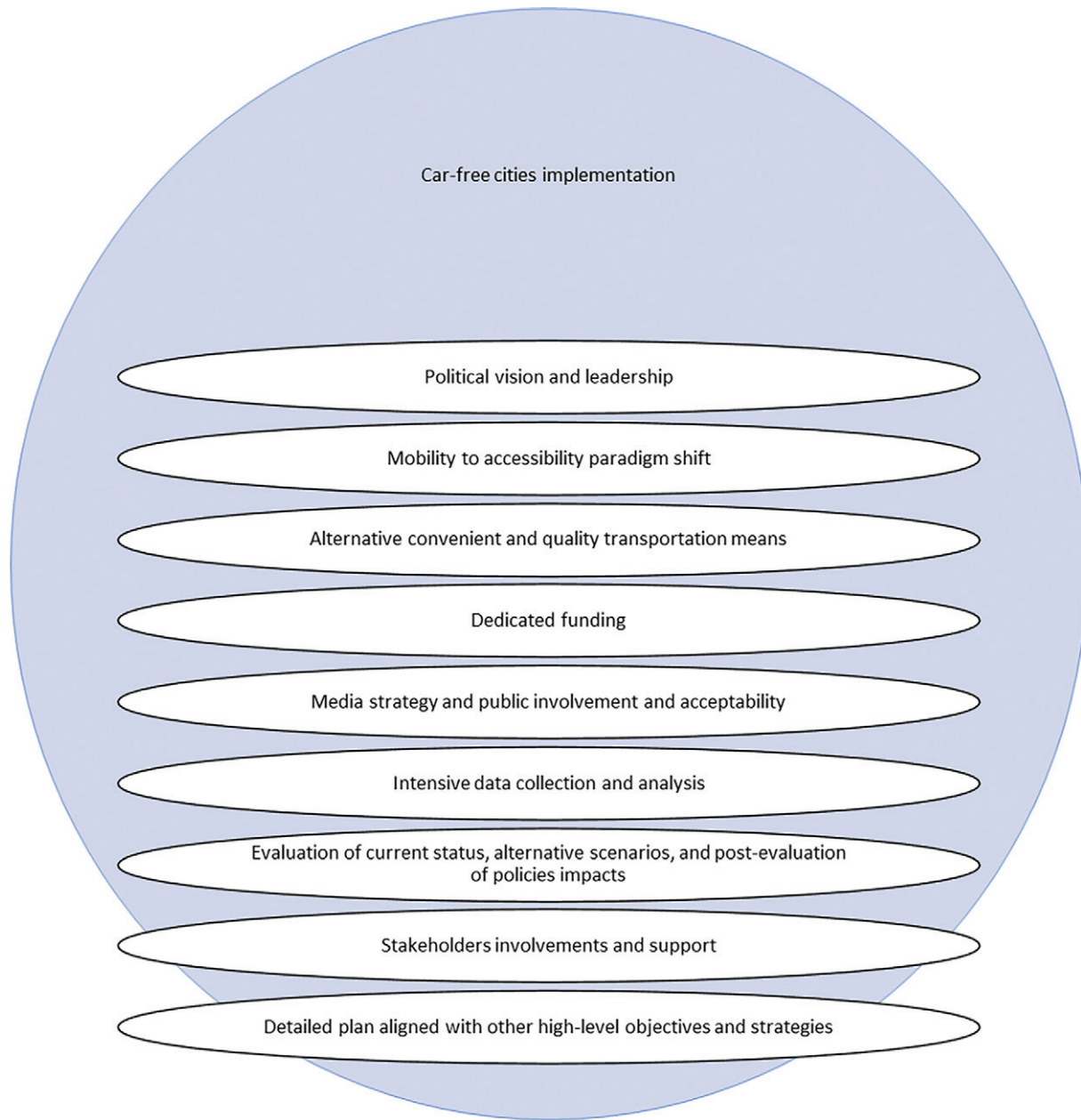


Fig. 2 Prerequisites for car-free cities' implementation. Source: Adapted from [Nieuwenhuijsen et al. \(2019\)](#).

last mile by active transportation options. To be fully functional, the coverage of transportation networks should not only focus on accessing jobs, but also other essential and non-essential destinations including medical care, education, healthy food, shopping, and social networks. The design of these networks should consider climatic factors in the cities where they are implemented. For example, in dry and hot cities, thermal discomfort may hinder the use of active and public transportation, while in humid climates, mechanical discomfort would be an important factor ([Böcker et al., 2016](#); [Shaaban et al., 2018](#); [Miao et al., 2019](#)). Planners and urban designers have options to improve thermal and mechanical discomfort in cities such as creating green and open spaces, including tree canopies, in the vicinity of cycling and pedestrian lanes, designing buildings with moderately nonuniform building heights to improve ventilation, using light colors on walking and cycling paths, and providing shaded paths and public spaces, and shaded or cooled public transportation hubs. Also, importantly, increasing building density and land-use diversity can make active and public transportation more convenient and viable options, and can help reduce exposure to adverse weather by making distances between origins and destinations shorter ([Negev et al., 2020](#)).

These changes are not a function of transportation planning alone but require integration with land-use planning. A paradigm shift from planning for mobility to planning for accessibility in transportation can set new directions in urban development and contribute to the emergence and sustainability of car-free cities. More research and data collection on the impacts of car-free

initiatives is also needed to build the evidence base and the case for such changes, which may be perceived as radical in certain contexts. Such research is currently missing and most of the documentation on city plans for going car-free comes from the gray literature and from Europe and North America, and too often does not include a pre-assessment and post-assessment, which could shed light on how successful these plans are in improving environmental and health conditions.

Citizen and business support for car-free plans is also essential and can be obtained through engaging these groups in the planning dialogue and demonstrating the benefits of such schemes as documented elsewhere or measured in pre- and post-assessments. Specifically, businesses are often opposed to the introduction of car-free zones or initiatives and represent a major barrier to the implementation and sustainability of such plans (Nieuwenhuijsen and Khreis, 2016). For example, car-free restrictions in Oslo have been opposed and hampered by trade associations and shopkeepers, on the basis that such plans may lead to poorer cities with less life and a reduced commercial activity in the city (Cathcart-Keays, 2020). This example from Oslo highlighted the importance of involving stakeholders in car-free transportation planning, including citizens, shopkeepers and trade associations, who in anecdotal evidence seemed happier and more accepting of these plans after their engagement and the revision of some elements to roll out changes in a more incremental manner (Cathcart-Keays, 2020). Similarly, recently in New Orleans, French Quarter residents rallied against proposals to make the historic neighborhood more pedestrian friendly, carrying signs which read as "Save our neighborhood—no pedestrian mall". The residents cited multiple reasons for their opposition including that the proposal will make it harder for them to unload groceries, hamper the work of carpenters who repair and restore historic buildings, and workers who tend bar, tote luggage and work kitchen stoves. Safety concerns were also voiced including that people who work in the French Quarter and get paid tips in cash prefer to get picked up outside work after their shifts, rather than walk with their cash, in an area known for higher rates of crime. Locals also suggested that the proposed plan caters more to tourists than local residents and workers (Reckdahl, 2020). Unfortunately, the car-free literature extent does not specifically covers areas and cities affected by tourism, or engage with issues of gentrification and safety in a comprehensive manner. Other actions which may facilitate the implementation and sustainability of car-free cities and initiatives include the alignment with other policy objectives such as reducing congestion or promoting green growth and the development of a timely media plan to explain the purpose and process of the proposed changes (Nieuwenhuijsen et al., 2019).

While the emergence of car-free cities and the more frequent adaptation of such plans by cities across the world is promising, there is currently no consensus on what these concepts mean or how best they can be implemented including how to overcome a wide range of barriers which are somewhat context specific. This is expected as implementing a car-free scheme in Barcelona will be extremely different from trying to implement one in Houston, as the city was built for cars. Current trends in moving cities' solutions away from the private car and to active and public transportation options signal a change in transportation planning and policy that unfortunately seems to be more restricted to Western Europe, where cities were initially built for walking and horse and carriage and tend to be compact. As European cities experiment with these concepts, it is imperative that research is conducted to measure the impacts of these schemes, not only on environmental exposures and health outcomes, but also on business activity, tourism, safety and economic outcomes. This will help build the case and evidence base for the implementation of car-free cities around the world.

What Holds us Back From Going Car-Free?

Decades of planning and investments in car infrastructure attracted cars to the cities and it will take decades to overturn this. Large car-oriented infrastructures continue to dominate with relatively small proportions of the budget allocated to and little work done for quality active and public transportation provision across most regions. There is an urgent need to rebalance and provide better and safer infrastructures and policy support for active and public transportation modes. For many, the car is still a status symbol and the fastest, easiest and most comfortable way to get around, and negative impacts such as air pollution, noise, heat island effects, and CO₂ emissions are ignored and not communicated well enough. Further, retail interest and car interest groups may be some of the biggest barriers, but their concerns are often unfounded.

A car-free city would provide a catalyst for better town planning by removing the need to facilitate car mobility and ensuring that urban areas are planned around people, functionality, and better built environments instead.

Furthermore, there are many other ongoing changes that may facilitate or hinder the process. Technological advancements such as electric cars and autonomous cars have been proposed as the solution to current issues such as air pollution, greenhouse gases, and road safety, but they have only a limited number of advantages and potentially many disadvantages. The use of electric cars may reduce tailpipe emissions (or at least displace them outside of the city), but it does not reduce non-tailpipe particulate matter emissions (Timmers and Achten, 2016) or the number of cars in the city, and they do not help to reduce other health issues such as deaths and injuries due to crashes, and the lack of daily physical activity. Autonomous vehicles may reduce the number of cars parked in the city, and to some extent the number of cars circulating in the city, but they also have the potential to increase car use in certain population groups, and may increase kilometers traveled as vehicles would need to travel between locations of disparate users, thereby increasing the exposures and health problems related to car use (Rojas-Rueda et al., 2020). If using autonomous vehicles is as easy as waiting for an elevator (an autonomous vertical vehicle) to travel up a few floors, many people will choose to wait rather than making the small effort to actively travel short distances. These technological solutions, which seem to have become a focal point of recent transportation policy, need more investigation and should be integrated within the detailed plans or proposals for future car-free cities. It is not that such technologies won't play a role in future transportation services, but they should not be considered a silver bullet to all current transportation problems.

There are yet many uncertainties in terms of acceptability and behavior change when introducing the car-free city and also the likely changes in terms of air pollution, noise, temperature, social cohesion, and physical activity. There is a research need to evaluate the health impacts of (planned) interventions in cities, including changes in perceptions and attitudes and health impact modeling of future scenarios, in addition to a more meaningful engagement with issues of gentrification, forced migration, the local economy and safety and crime. Further research and research synthesis are needed to build a strong evidence base, which is currently nonexistent. It is also vital to better understand and document how mobility patterns change with the introduction of car-free measures.

References

- ASLA, D., 2020. Helsinki Aims to Be Car-Free by 2025. Available from: <https://www.smartcitiesdive.com/ex/sustainablecitiescollective/helsinki-aims-be-car-free-2025/297026/>.
- Bhalla, K., et al., 2014. Transport for health: The global burden of disease from motorized road transport.
- Bhalla, K., et al., 2014. Transport for health: the global burden of disease from motorized road transport. In: Global Road Safety Facility, Institute for Health Metrics and Evaluation and World Bank, Washington DC.
- Bieda, K., 2016. Car-free cities: urban utopia or real perspective? In: Back to the Sense of the city, International Monograph Book, Centre de Política de Sòl i Valoracions.
- Böcker, L., Dijst, M., Faber, J., 2016. Weather, transport mode choices and emotional travel experiences. *Transp. Res. Part A: Policy Practice* 94, 360–373.
- Boffey, D., 2020. Forward-thinking Utrecht builds car-free district for 12,000 people. Available from: <https://www.theguardian.com/world/2020/mar/15/forward-thinking-utrecht-builds-car-free-district-for-12000-people>.
- Bracelona Institute for Global Health, 2020. I. 7 Ways that Bicycles Can Make Cities Healthier. Available from: https://www.isglobal.org/en/publication/-/asset_publisher/ijGAMKTWu9m4/content/7-maneras-en-que-las-bicicletas-pueden-hacer-las-ciudades-mas-saludables.
- Burgen, S., 2018. For me, this is paradise': life in the Spanish city that banned cars. Available from: <https://www.theguardian.com/cities/2018/sep/18/paradise-life-spanish-city-banned-cars-pontevedra>.
- Cathcart-Keays, A., 2020. Will we ever get a truly car-free city? Available from: <https://www.theguardian.com/cities/2015/dec/09/car-free-city-oslo-helsinki-copenhagen>.
- Cathcart-Keays, A., 2020. Oslo's car ban sounded simple enough. Then the backlash began. Available from: <https://www.theguardian.com/cities/2017/jun/13/oslo-ban-cars-backlash-parking>.
- Coffey, H., 2020. Paris to ban cars in city centre one sunday a month. Available from: <https://www.independent.co.uk/travel/news-and-advice/paris-car-free-sundays-city-centre-france-pedestrian-a8566991.html>.
- Crawford, J.H., 2000. Carfree cities, International Books, Alexander Numankade 17, 3572 KP Utrecht, Netherlands.
- Crawford, J.H., 2009. Carfree design manual. International Books.
- Frumkin, H., 2016. Urban sprawl and public health: Designing, planning, and building for healthy communities. Island Press, Washington, DC.
- Fuller, C.H. and Brugge, D., 2020. Environmental Justice in Traffic-Related Air Pollution, in: Khreis, H., et al., (Eds.) Elsevier.
- Greenfield, A., 2014. Helsinki's ambitious plan to make car ownership pointless in 10 years. Available from: <https://www.theguardian.com/cities/2014/jul/10/helsinki-shared-public-transport-plan-car-ownership-pointless>.
- Hensher, D.A., Button, K.J., 2001. Handbook of transport systems and traffic control. Elsevier Science.
- Héran, F., Ravalet, P.J., 2008. La consommation d'espace-temps des divers modes de déplacement en milieu urbain. Application au cas de l'île de France.
- Jones, S., 2018. It's the only way forward': Madrid bans polluting vehicles from city centre. Available from: <https://www.theguardian.com/cities/2018/nov/30/its-the-only-way-forward-madrid-bans-polluting-vehicles-from-city-centre>.
- Khreis, H., et al., 2016. The health impacts of traffic-related exposures in urban areas: Understanding real effects, underlying driving forces and co-producing future directions. *J. Transp. Health* 3 (3), 249–267.
- Khreis, H., et al., 2017. Commentary: How to create healthy environments in cities. *Epidemiology* 28 (1), 60–62.
- Khreis, H., de Hoogh, K., Nieuwenhuijsen, M., 2018. Full-chain health impact assessment of traffic-related air pollution and childhood asthma. *Environ. Int.* 114, 365–375.
- Khreis, H.G., Andrew, R.T., Zietsman J., Nieuwenhuijsen M.J., Mindell J.S., Winfree, G.D., et al., 2019. Transportation and Health: A Conceptual Model and Literature Review. T.C.f.A.R.i. T.E. College Station, Energy, and Health, May 2019. Available from: http://www.carteteh.org/wp-content/uploads/2019/04/14-Pathways-Project-Brief_Final-version_24April2019.pdf.
- Lelieveld, J., et al., 2015. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature* 525 (7569), 367–371.
- Loo, B.P., 2018. Realising car-free developments within compact cities. In: Proceedings of the Institution of Civil Engineers-Municipal Engineer, Thomas Telford Ltd.
- Masiol, M., et al., 2014. Thirteen years of air pollution hourly monitoring in a large city: Potential sources, trends, cycles and effects of car-free days. *Sci. Total Environ.* 494, 84–96.
- McCaill, C., Garrick, N., 2012. Automobile use and land consumption: Empirical evidence from 12 cities. *Urban Des. Int.* 17 (3), 221–227.
- Melia, S., 2006. On the road to sustainability: Transport and car-free living in Freiburg. Report for WHO Healthy Cities Collaborating Centre. Available from: www.carfree.org.uk/038.
- Miao, Q., Welch, E.W., Sriraj, P., 2019. Extreme weather, public transport ridership and moderating effect of bus stop shelters. *J. Transp. Geogr.* 74, 125–133.
- Morris, D., et al., 2009. Car-free development through UK community travel plans. *Proc. Inst. Civil Eng. Urban Des. Plan.* 162 (1), 19–27.
- Mueller, N., et al., 2017. Urban and transport planning related exposures and mortality: A health impact assessment for cities. *Environ. Health Perspect.* 125, 89–96.
- Mueller, N., et al., 2018. Socioeconomic inequalities in urban and transport planning related exposures and mortality: A health impact assessment study for Bradford, UK. *Environ. Int.* 121, 931–941.
- Mueller, N., et al., 2019. Changing the urban design of cities for health: the superblock model. *Environ. Int.* 105132.
- Negev, M., et al., 2020. Designing cities for health and resilience in a hot and dry climates. *British Medical Journal*.
- Newman, P., Kenworthy, J., 1999. Sustainability and cities: Overcoming automobile dependence. Island press.
- BBC News, B., 2020. Barcelona's car-free smart city experiment. Available from: <https://www.bbc.com/news/technology-50658537>.
- Nieuwenhuijsen, M., 2020. Enseñanzas del coronavirus: 8 medidas para hacer ciudades más habitables y saludables. Available from: <https://theconversation.com/ensenanzas-del-coronavirus-8-medidas-para-hacer-ciudades-mas-habitable-y-saludables-136807>.
- Nieuwenhuijsen, M.J., 2020b. Urban and transport planning pathways to carbon neutral, liveable and healthy cities: A review of the current evidence. *Environ. Int.* 105661.
- Nieuwenhuijsen, M.J., Khreis, H., 2016. Car free cities: Pathway to healthy urban living. *Environ. Int.* 94, 251–262.
- Nieuwenhuijsen, M., et al., 2019. Implementing car-free cities: rationale, requirements, barriers and facilitators. *Integrating Human Health into Urban and Transport Planning*, Springer, pp. 199–219.
- O'Sullivan, F., 2020. Paris Mayor: It's Time for a '15-Minute City'. Available from: <https://www.bloomberg.com/news/articles/2020-02-18/paris-mayor-pledges-a-greener-15-minute-city>.
- Peters, A., 2019. What happened when Oslo decided to make its downtown basically car-free? Available from: <https://www.fastcompany.com/90294948/what-happened-when-oslo-decided-to-make-its-downtown-basically-car-free>.
- Peters, A., 2020. Here are 11 more cities that have joined the car-free revolution. Available from: <https://www.fastcompany.com/90456075/here-are-11-more-neighborhoods-that-have-joined-the-car-free-revolution>.

- Rojas-Rueda, D., et al., 2020. Autonomous Vehicles and Public Health.. *Annu. Rev. Public Health* 41, 329–345.
- Shaaban, K., Muley, D., Elnashar, D., 2018. Evaluating the effect of seasonal variations on walking behaviour in a hot weather country using logistic regression. *Int. J. Urban Sci.* 22 (3), 382–391.
- Shu, S., et al., 2016. Air quality impacts of a CicLAvia event in Downtown Los Angeles CA. *Environ. Pollut.* 208, 170–176.
- Sohrabi, S., Khreis, H., 2020. Burden of disease from transportation noise and motor vehicle crashes: Analysis of data from Houston, Texas. *Environ. Int.* 136, 105520.
- Sohrabi, S., Zietsman, J., Khreis, H., 2020. Burden of disease assessment of ambient air pollution and premature mortality in urban areas: the role of socioeconomic status and transportation. *Int. J. Environ. Res. Public Health* 17 (4), 1166.
- Timmers, V.R., Achten, P.A., 2016. Non-exhaust PM emissions from electric vehicles. *Atmos. Environ.* 134, 10–17.
- Vauban Freiburg, Wikipedia, 2020.. Available from: https://en.wikipedia.org/wiki/Vauban,_Freiburg.
- V oelcker, J., 2014. 1.2 Billion Vehicles On World's Roads Now, 2 Billion By 2035: Report.Available from: https://www.greencarreports.com/news/1093560_1-2-billion-vehicles-on-worlds-roads-now-2-billion-by-2035-report.
- Willsher, K., 2015. Paris's first attempt at car-free day brings big drop in air and noise pollution.; Available from: <https://www.theguardian.com/world/2015/oct/03/pariss-first-attempt-at-car-free-day-brings-big-drop-in-air-and-noise-pollution>.

Superblocks Base of a New Model of Mobility and Public Space. Barcelona as an Example

Salvador Rueda Palenzuela, President of the Urban Ecology and Territorial Foundation, Barcelona, Spain

© 2021 Elsevier Ltd. All rights reserved.

Urban Functionality: Model of Mobility and Public Space	249
A New Urban Cell: The Superblock	249
Superblock, Base for a Functional and Urbanism Model: The Case of Barcelona	249
The Bus and Bicycle Network	251
The Pedestrian Network	252
Uses in Public Space and Citizen Rights: From Pedestrians to Citizens	252
The Green Network that Appears with the Implementation of Superblocks	253
The Impacts of the Current Mobility Model	254
Superblocks, a Tool That Helps to Mitigation and Adaptation to Climate Change (Climate Change Assessment)	256
References	257

Urban Functionality: Model of Mobility and Public Space

Mobility is one of the uses of public space, not the only one. Its purpose is, above all, the functionality of urban systems. Ensuring functionality and, at the same time, the rest of uses in the public space is not simple and is one of the objectives of the superblock concept.

Urban functionality, defined by the mobility and services patterns of each city, determines, largely, the quality and liveability of public space. There is a need to develop a model of mobility and a more sustainable public space, in order to guarantee a more accessible, comfortable, safe, and multifunctional public space where people can be citizens and exercise in the public space the rights to interchange, culture, leisure, expression, and demonstration, besides the right to move. With the current model of mobility, cities dedicate most of the public space to mobility and in these conditions the ultimate aspiration is to be a pedestrian, a mean of transportation. At least 75% of public space should be dedicated to citizen rights. Public space should acquire the maximum liveability but being, at the same time: comfortable (without noise, without air pollution, and with the greatest thermal comfort); attractive (with a high diversity of activities, with attractive activities and with the greatest biodiversity); and ergonomic (accessible, with liberated space to exercise all the rights and with a good relation between built heights and street widths).

A New Urban Cell: The Superblock

The superblock is a cell of nine blocks in the case of the example in Barcelona, defined by a network of basic roads that connects origins and destinations of the whole city (Fig. 1). When the cell is being reproduced along the urban system, the size accommodates the morphological and functional characteristics of the current city (it must be highlighted that the superblocks project is a urban recycling project), liberating the maximum surface of public space, which is nowadays occupied by traffic and, at the same time, guarantee the functionality and organization of the system. The new cell is defined by the perimeter basic roads, where through and connecting traffic circulate at a maximum speed of 50 km/h. Interior roads (intervias) of superblock constitute a local network with limited speed of 10 km/h, speeds which allow shared urban uses. A superblock cannot be crossed, which means that the movements in its interior only make sense if their origin or destination is in the intervias. Therefore, the neighborhood streets will be with little or no noise, or pollution, etc. and allow more than the 70% of space that is currently occupied by the through traffic for the movements by foot or bicycle.

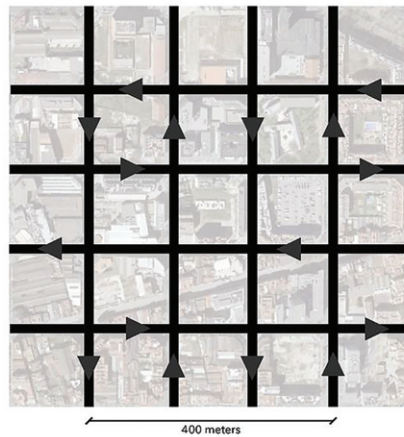
The reasons to choose the dimensions of 3×3 for the superblock are based on the characteristics of cars which, at a speed of 20 km/h (which is the average urban speed in Barcelona), will spend a similar amount of time to go around the superblock as a person walking around the block at about 4 km/h. With the presence of main crossings every 400 m, the traffic lights synchronization is much more efficient (with these distances one can think of prioritization of the traffic lights for public transport) and avoids disruption of the main flux because of turns (two out of three turns are avoided).

Superblock, Base for a Functional and Urbanism Model: The Case of Barcelona

Superblocks aim to be the basis of a functional model of any city, but they can also become the basis of a new urbanism model. The average number of people in a Superblock in Barcelona is higher than 6000 inhabitants. Every superblock emerges as a little city. In this section, we will focus on the superblock as the basis of a new functional model and the consequences for public space.

Road hierarchy in the new Superblock model

CURRENT SITUATION

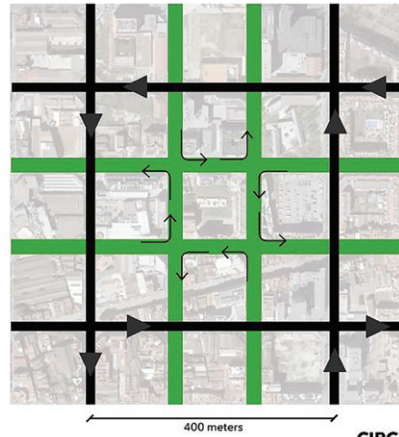


Basic network: 50 km/h



SOLE RIGHT IN STREET SPACE: MOBILITY
HIGHEST AIM: PEDESTRIAN.

SUPERBLOCK MODEL



Local network: 10 km/h



EXERCISE ALL THE RIGHTS THAT THE CITY OFFERS.
HIGHEST AIM: ACTIVE CITIZEN.

**CIRCULATING
VEHICLES DO
NOT PASS
THROUGH**

Figure 1 Networks scheme, current and future, based on superblocks. Source: BCNecologia

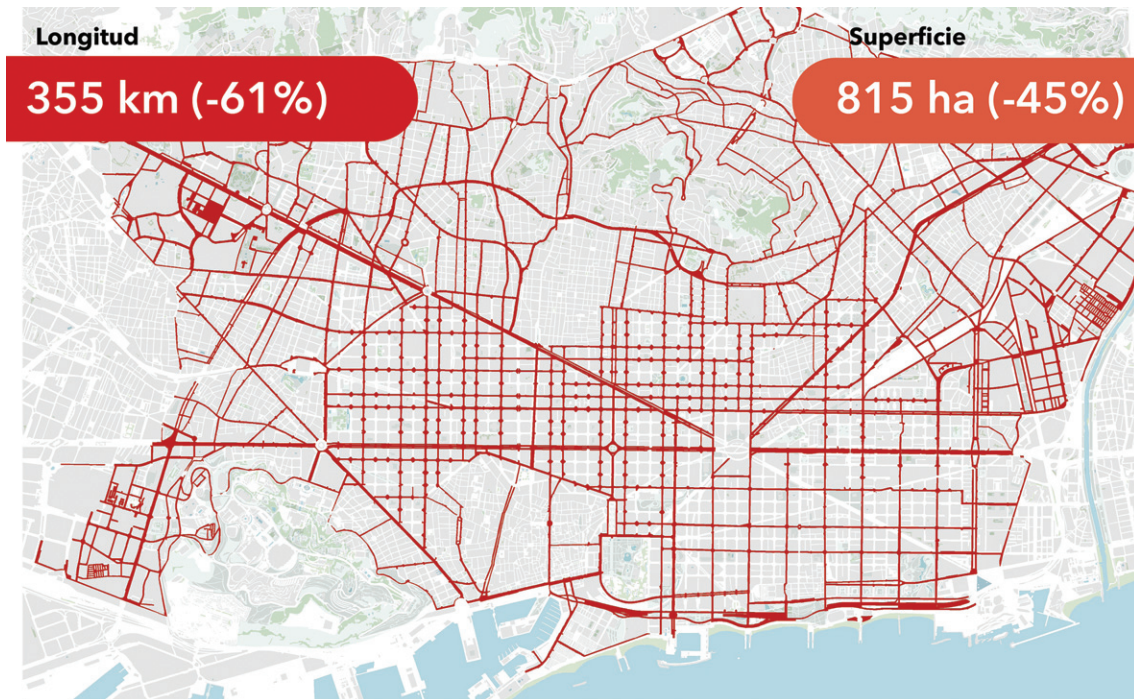


Figure 2 Map of superblocks in Barcelona. Public space (in red) dedicated to mobility. Source: BCNecologia

The defining roads of superblocks (in red) comprise a network of basic roads where the urban transport circulates: collective transport, private vehicle, emergencies, services, and, if the section allows it, bicycles. This network of basic roads, which results in the greatest orthogonality, allows access to the city at the highest speed admitted by law (50 km/h) (Fig. 2).

The basic network of superblocks reduces the length of the total of roads that nowadays is used by through traffic by 61% (see Fig. 2). This dramatic reduction does not lead to a proportional drop-off in the circulation of vehicles for the same level of service (the same speed of the vehicles circulating). In Barcelona, with a reduction of vehicles of about 13% we can achieve a level of service

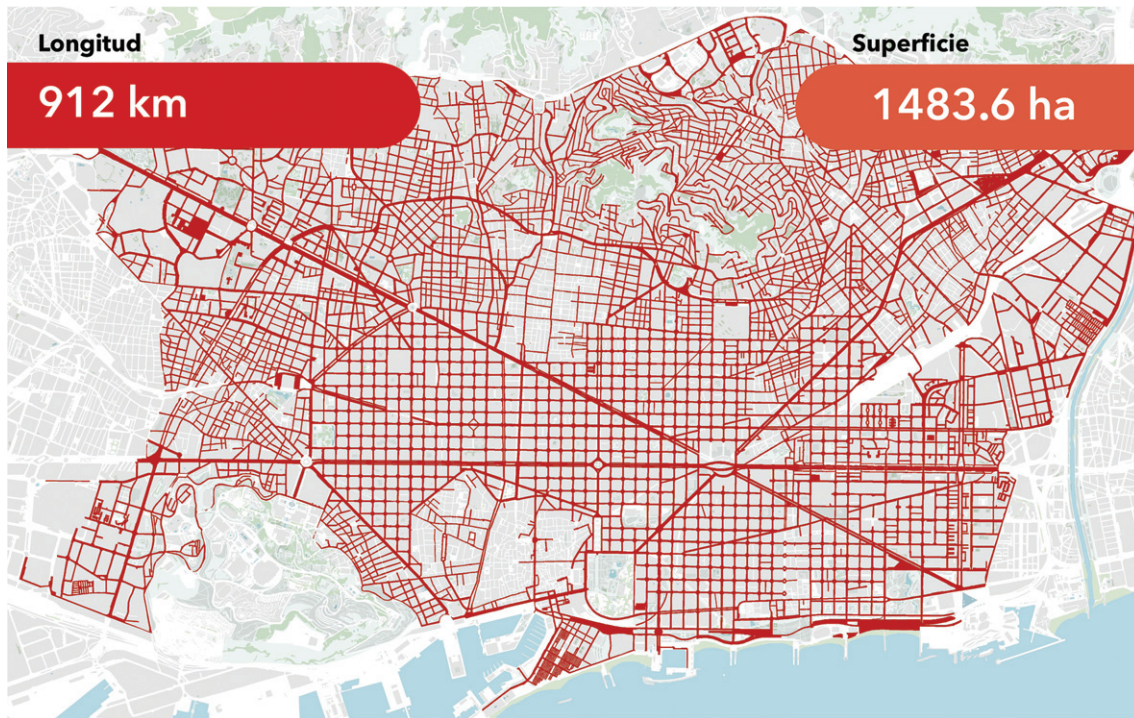


Figure 3 Public space of Barcelona dedicated to through traffic in the current situation. Source: BCNecologia

similar to current one. It maintains, thus, the functionality and organization of the system. **Fig. 3** shows that the space dedicated nowadays to through traffic is near 15 million of squared meters and the length of the roads dedicated to mobility nearly 912 km, or 85% of the total roads of Barcelona.

The Bus and Bicycle Network

Superblocks allow the integration of networks of through traffic (cars, buses, and bicycles) in its periphery, and priority of movements by foot and bicycle on the inside. Orthogonal networks are the most efficient in urban systems. With the new orthogonal network of buses, the topology of the new network, combined with the distance of the stops every 400 m, contributes to the increase of commercial speed of buses, more so than the traffic lights prioritization or the construction of bus lanes (**Fig. 4**). With the same number of buses that have a frequency of coming every 14–15 min nowadays, there will be a move to a frequency of buses coming around every 5 min in the future (it will be like a surface underground). The service will be the same in the center and in the periphery of the city. The design of the network allows an average waiting time at each stop of around 2 min, which for our mental clock does

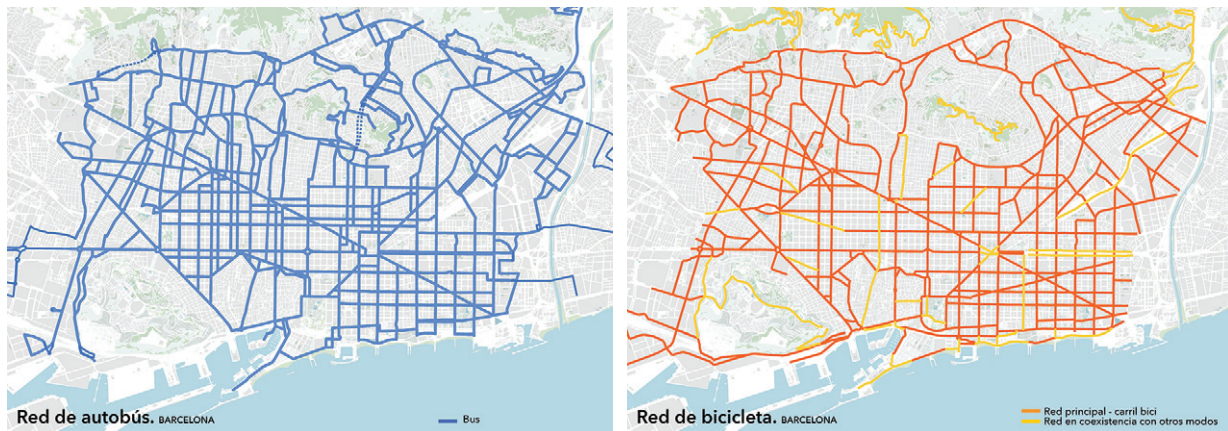


Figure 4 Bus network and bicycle network in Barcelona in a superblocks scenario. Source: BCNecologia

not feel like waiting. It is a network that connects any origin with any destination with only one transfer in 95% of cases. It will be an intelligible network like the underground and, the estimated number of transfers will be similar to the underground. In the example fabric, the perpendicularity of the lines of the network allows the presence of a unique stop for the horizontal and vertical lines in all the intersections (in the octagonal crossroads).

The bicycle network will also be adjusted to the superblocks structure (Fig. 4). Superblock's periphery hosts the transport network of bicycles, with exclusive lanes and shares the street section with buses and cars. The interior of the superblocks, at 10 km/h, allows bicycles in both directions, to cross the superblock. Their speed must be adjusted to the speed of pedestrians and the uses that are in place at that moment. If necessary, the rider needs to step off the bicycle. The conditions of intervias allow children go to school by bicycle or by foot without being accompanied by an adult.

The integration of electric engines for the transport is on the agenda of many cities. There is no doubt that the electric bicycle is the electric vehicle to promote. It does not pollute, it does not make noise, it practically does not consume energy (the consumed energy for a journey by electric bicycle, adding the used metabolic energy and the consumed electricity, is less than the energy metabolically consumed doing the same journey by foot); it allows an average person to overcome high slopes of up to 20%, it is healthy, and adjusts the effort to the context. During the summer it even refrigerates—allowing its use in the most severe season without sweating. The homologated motor works up to 25 km/h reducing the severity of accidents. The average distance of the classical bicycle is about 5 km, while electric bicycle increases it up to 10 km, which is the distance from one end to the other in the municipality of Barcelona.

The electric bicycle is, for a distance of 11 km and a speed 30% higher than the speed of a classic bicycle, the most competitive form of mobility.

The isoform distribution of the networks along the city gives a level of service which is equitable for the bus and bicycle networks, something previously only the car had.

The Pedestrian Network

At present, Barcelona has 230 ha of streets with specific space for pedestrians or with speeds limited to 20 km/h. If we add the surface used by the pedestrians in big avenues, the total is about 15.8% of the public space of roads. With the implementation of superblocks 6.22 million of square meters are liberated. This will make it the most important recycling proposal in the world, without demolishing a single building (Fig. 5).

Uses in Public Space and Citizen Rights: From Pedestrians to Citizens

Maybe the most radical aspect of the proposal is the reconversion of most of the urban space to multiple uses and rights. Radical because it goes to the roots of the meaning of public space. A city exists when, first, there is public space and, second, when they are reunited in a limited space by a determined number of legal entities that are complementary and “work” synergistically. We can find ourselves in an urbanization of undetached houses with space for the car to get to the garage. In this case, we talk about urbanized space, but not public space. In urbanization, it makes little sense to have a market, a cultural event, or even, find children playing with balls in the middle of the street.

A city starts to become a city when there is public space, since it is the “house of everybody”, the meeting place for interchange, leisure and staying, culture, expression, and democracy and, also, the movement. Public space makes us citizens and we are so when we have the possibility of occupying it for the exercise of all the rights mentioned above. Today, the limited possibility of exercising the citizen rights relegates us as pedestrians. Giving citizens public space that was lost due to the current model of mobility is key to

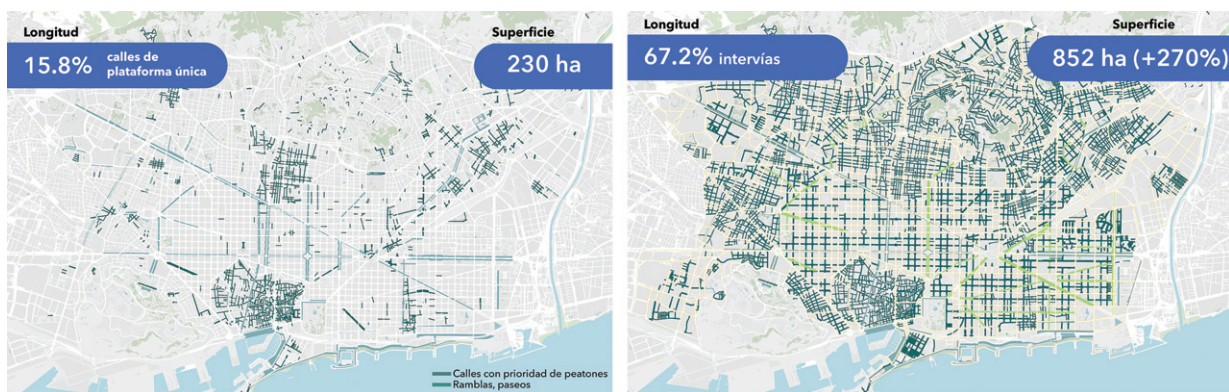


Figure 5 Map for pedestrians and other uses. Current scenario and with Superblocks. Source: BCNecologia

the new model of mobility and public space based on superblocks. Electric vehicles can reduce part of the noise (but not the noise that at some speeds comes from the friction between the tires and the bearing surface and not from the engine) and part of the air pollution (as almost half of the pollution caused by particles come from the “dust” lifted by the wheels, that comes from tires particles, brakes, bearing lubricant oils, etc., that, as it is known, contain heavy metals and components of high toxicity). What they cannot reduce is the space that they occupy, which in a compact city in general, and particularly in Barcelona, is the most scarce good.

Superblocks provides citizens with nature in almost 70% of the space of the city. Superblocks allow not only for the integration of transportation networks but also green networks. Furthermore, spaces that are not crossed by any network of mobility: cars, buses, and bicycles. Rights are obtained through speeds that are compatible with the use of the space by the most vulnerable people (for instance, the pass of blind people, children playing). If a Superblock allows a bus network, a car network, or a bicycle network with signalized lane, then it is no longer a Superblock because it is not compatible with all the rights.

The Green Network that Appears with the Implementation of Superblocks

Many urban planners consider that the permeability of the soil a good indicator of naturalization of an urban fabric. The presence of permeable soils rebalances the water cycle: it favors the infiltration of rain waters and retains water through the different vegetable surfaces. The vegetation protects the soil from excessive insulation and protects it from the compaction that provokes the direct impact of raindrops on the soil. By enabling the water to stay longer on surface, it increases the possibility that it infiltrates into the phreatic layers and reduces the risk of floods. It boosts the closing of the cycle of organic matter, by providing the urban soil with surfaces of compost generated by organic waste self-composting. Green spaces and land reservation for urban gardens constitute spaces for generating community among the inhabitants of a neighborhood or territorial unit. Surfaces with vegetable cover help to mitigate the emissions of CO₂, by fixing this gas through the photosynthetic process. Vegetable surfaces are, furthermore, potential captors of polluting particles and help to propitiate thermal comfort, minimizing the heat island effect. Besides, surfaces with trees proportionate acoustic and mechanic comfort, reducing the effect of noise and wind in the urban environment.

There are few areas with permeable soils in Barcelona (Fig. 6). The current green surface in the area of example is only 171.2 ha. The number of square meters per inhabitant is 2.7 m², very far below the 9 m²/inhab the WHO recommends. In a city like Barcelona, with such a lack of free spaces, Superblocks allow us to obtain better values of corrected compactness (balance between urban compression and decompression). The release of only the interiors of blocks in example, is clearly insufficient although necessary. In example, green occupied green spaces. Superblocks will allow us to restore part of the green space that is so needed.

With superblocks, green surfaces increase significantly to 403.7 ha of potential green, while maintaining city's functionality. It will increase from 2.7 m²/inhab to 6.3 m²/inhab for the whole area of Cerdà's Plan. In the Sant Martí area, it increases up to 7.6 m²/inhab. Street transformation and substituting cars with green space, allow us to obtain urban landscapes like the ones that are shown in Fig. 7. It shows the project presented by the City Council to the neighbors of the pilot superblock of Poblenou, for the section of Sancho d'Àvila between Llacuna and Roc Boronat streets (Fig. 8). As Oriol Bohigas said: “A street will have for Latins an infinity of values that a garden won't”. (Bohigas 1958)

A square has been and is the very place for public space. In the case of example of Barcelona, often there are sidewalks of only 5 m of width with little green space. (1.85 m²/inhabitants) With the Superblocks project, the number and surface areas of new squares that are in the junctions of the example fabric will multiply. In a Superblock type of 3 × 3 blocks, four new squares of about 1900 m² will be created (Fig. 8). 130 nodes will become complete squares of 1900 m², adding around 24.7 ha while there will be 20 new squares with a surface of 2/3rd of a complete surface, adding 3 more ha. In total there will be about 150 new squares with a surface of 27.7 ha.

Green from the interiors of block and green covers can also be added. Environmental benefits increase with the increase of urban green surface in height and surface. When you add the public space's green surface and green covers (with an estimated an

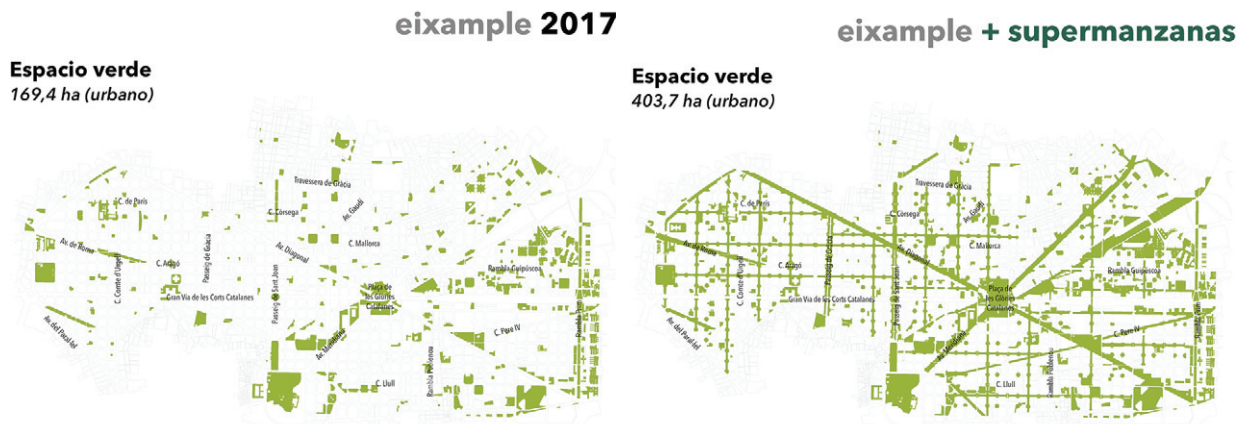


Figure 6 Green space in current situation and with Superblocks. Source: BCNecologia



Figure 7 Sections of a street with single platform in the interior of superblock and proposal of transformation of Sancho d'Àvila street between Llacuna and Roc Boronat streets. Sources: BCNecologia and Barcelona City Council

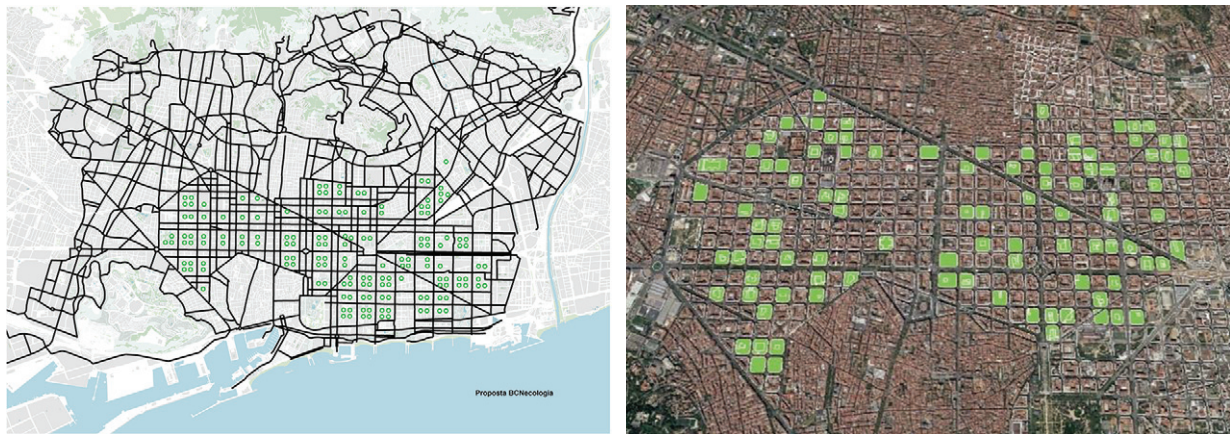


Figure 8 Junctions that become new squares in the superblocks model and block's interiors released and potentially releasable in central Example. Sources: BCNecologia and Barcelona City Council

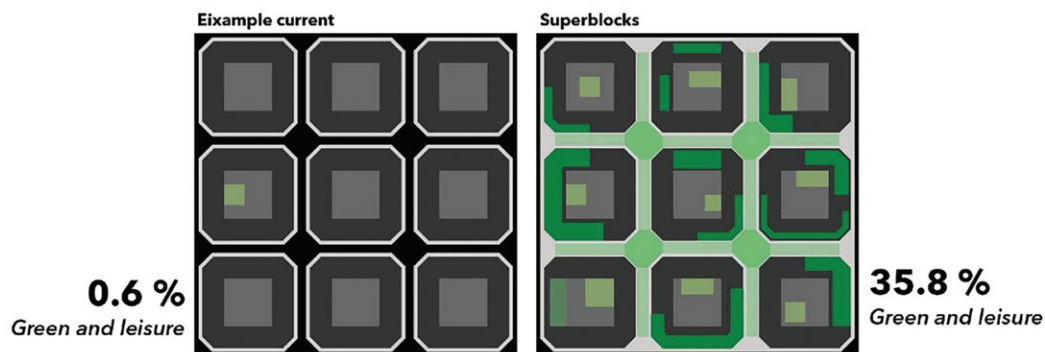


Figure 9 Percentage of current green surface and potential green spaces (in % of total surface) in an area equivalent to a superblock type where is included: public space green, blocks' interior patio's green and green covers. Source: BCNecologia

occupation of 30%), and the green surface in the interior of blocks (counting 1500 m² per block), green surface per inhabitant increases up to 9.6 m²/inhabitants (Fig. 9).

The Impacts of the Current Mobility Model

Mobility currently brings the biggest dysfunctions to the city. The use of public space is restricted to mobility and Barcelona dedicates more than 60% of public space and 85% of roads to traffic.

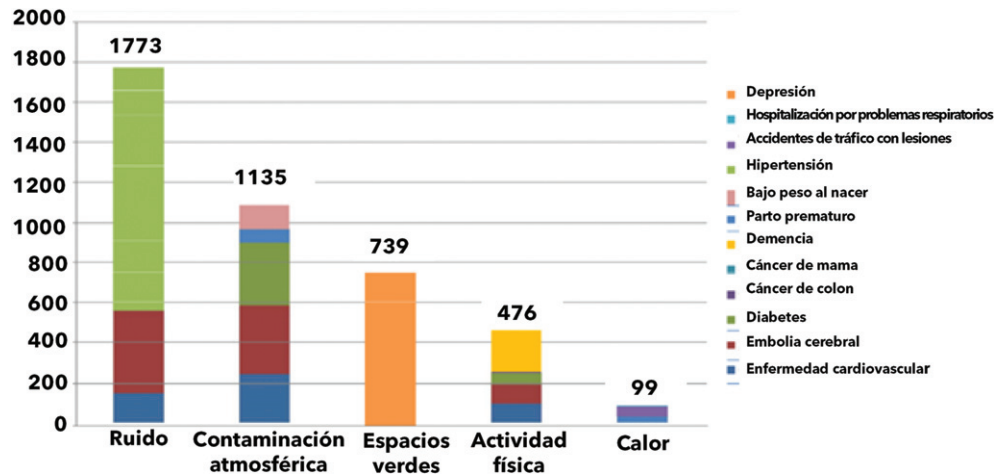


Figure 10 Morbidity in Barcelona by different reasons. Source: ISGlobal

Air pollution emitted by motorized traffic has an unacceptable impact on the health of population in the Metropolitan Area of Barcelona. In a study conducted by ISGlobal CREAL in an area of 56 municipalities, including Barcelona, it was estimated that air pollution causes: 3500 premature deaths per year; 1800 hospitalizations by cardiovascular reasons; 5100 cases of chronic bronchitis in adults; 31,100 cases of children bronchitis; 54,000 asthma attacks among children and adults. (Künzli and Pérez, 2007).

The impacts of air pollution on health are today the main problem to solve out of all the problems caused by the current model of mobility. The liveability index in most of the city is below minimal. The green surface in central example is 1.85 m² per inhabitant when the WHO recommends 9 m²/inhab. Example is also the district with the most traffic; almost 50% of the population is exposed to inadmissible levels of noise (diurnal values >65 dbA). The negative economic impact runs in the millions of Euros per year (according to the World Bank, for Spain it was 45,000 million of Euros/year in 2013, considering only the impact on the health). Black asphalt and the emissions of cars are responsible for the most important part of the urban heat island. This increase of more than 2° of average temperature (at summer nights can surpass the 5° of difference compared to periphery) is especially painful and in some cases mortal, for the most vulnerable people: elderly, children and sick people, particularly when the heat waves occur because of [climate change](#). Traffic accidents caused 30 deaths per year in Barcelona and more than 30 injuries per km/year in example. Visual intrusion and landscape deterioration (landscape as expression of the integration of different variables) convert Barcelona in a “pressure cooker” affecting 85% of the length of streets in the city.

The results of a study conducted by ISGlobal (Mueller et al., 2017) for Barcelona and its Metropolitan Area, show clearly the impact that some of the exposures have on morbidity of the citizens of Barcelona (Fig. 10).

The second (Mueller et al., 2019) calculates the number of premature deaths (and the causes) that could be avoided with the implantation of the superblocks in Barcelona (Fig. 11).

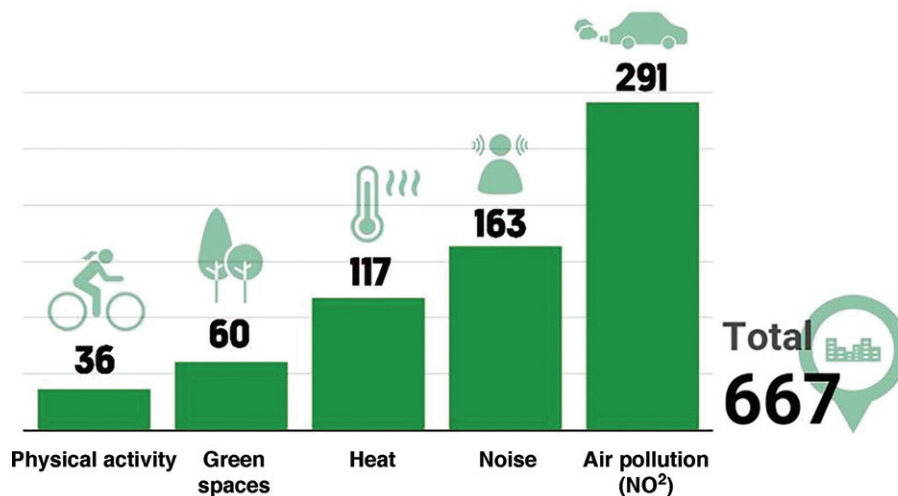


Figure 11 Annual premature deaths that the “superblocks” model could avoid in Barcelona. Source: ISGlobal

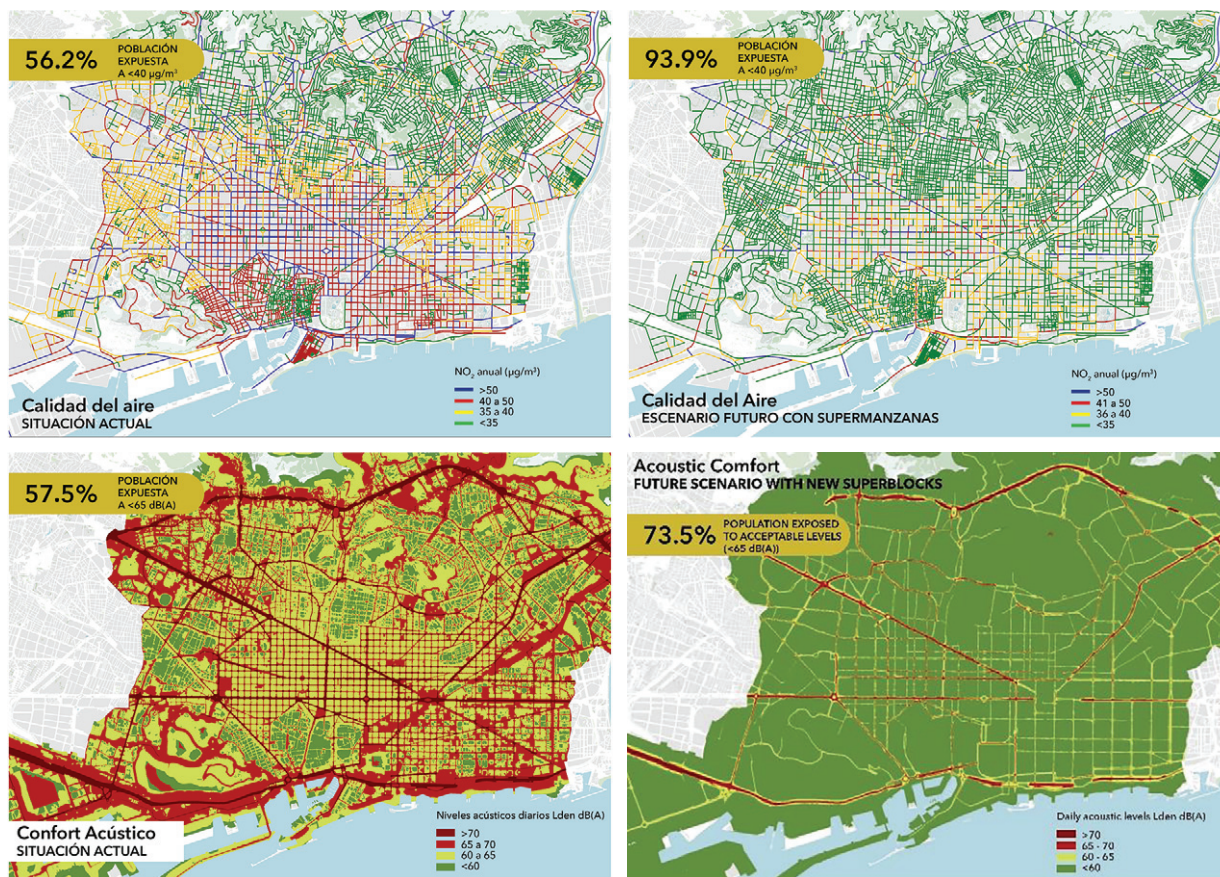


Figure 12 Air quality and urban noise in Barcelona. Exposed population (in %) to levels of NO_2 and decibels legally admissible. Current scenario and with superblocks. Source: BCNecologia

The result is a city that is not ready to address the great challenges of the beginning of century: sustainability in the information age. There is need for a new ecosystemic model with its corresponding system of proportions that includes, at the same time, the reduction of: polluting emissions, noise, energy and the increase of: green, staying spaces; legal entities diversity, also the one dense in knowledge. An urban model that extends along the city and takes into account the current modes of mobility. To address these serious problems, the Sustainable Urban Mobility Plan of Barcelona approved by the City Council (March 2015) proposes to extend the Superblocks throughout the city and reduce 21% of circulating vehicles. With this reduction it is estimated that the levels of air pollution in all the measurement stations will be below the guideline values. With a reduction of 21% of circulating vehicles, the level of service of traffic and the environmental conditions of roads in a superblocks scenario will be much better than now.

With this reduction of vehicles, the percentage of people exposed to levels below the guidelines for air pollution will be 94% (today it is 56%) and for noise 73.5% (nowadays it is 54%). The liveability index will increase significantly in all the neighborhoods of the city (Fig. 12).

Superblocks, a Tool That Helps to Mitigation and Adaptation to Climate Change (Climate Change Assessment)

Urban heat islands form due to the emissions of heat by part of the series of surfaces that configure physical parts of the city. The emission of heat to the atmosphere is a consequence of different factors, like for example: climate conditions of the place, urban morphology, distribution of materials of the different urban surfaces, (pavements, covers and façades) and also by the heat produced by the activity of the city (traffic, building's conditioning, etc.) (Martín-Vide, 2015).

Currently, cities are characterized by the predominance of impermeable surfaces. This aspect produces an important concentration of heat, especially in the asphalt pavements of urban roads and the buildings. One of the areas of discussion of climate change is precisely the mitigation of urban heat island because its impact has repercussions both for energy consumption and people's well-being, reducing the comfort levels during the day and above all during night. Heat absorbed during the day is freed during night, increasing notably the temperatures of summer nights. That is why it is important to implement measures that allow mitigating urban heat generation and reducing urban centers temperature.

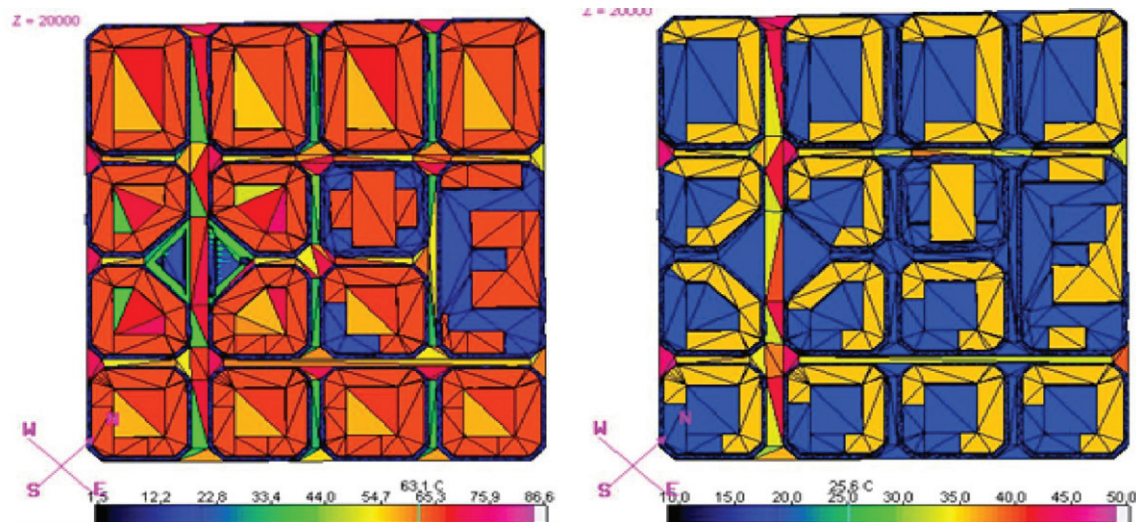


Figure 13 Thermal simulation comparing example in current situation and with superblocks. Source: BCNecologia

To reduce the current heat island effects, it is necessary to reduce the number of vehicles circulating and the consumption of fossil fuel energy, as well as to extend a green carpet that allows us to increase our resilience toward heat waves that will become more frequent with climate change. Superblocks will help in the mitigation and in the adaptation of the problem.

Through simulations of heat transfer for a summer day in Barcelona, one can demonstrate the repercussion of the current urban morphology and with Superblocks. The Superblock scenario will substitute the asphalted pavements in intervias with semipermeable pavements and the greatest cover of road trees. It will dedicate 30% of cover surfaces to green covers and release 1500 m² of interior patios of each of the blocks to green spaces. This has been analyzed in two areas of Superblocks in the left part of example (Fig. 13).

Results clearly indicate how the current situation of urban fabric in example is warmer than the Superblocks proposal. In terms of heat balance,¹ while the current fabric is 72.1 kW/m², the superblocks proposal obtains a balance of 46.3 kW/m², 35.9% less than the current situation. This energy translates into lower surface temperatures of pavements, buildings, and green spaces.

The heat balance by radiation refers to the sum of heat gained by radiation both in short and long wavelength. In incident solar radiation (short wave) a part is absorbed, and another proportion is reflected according to material's albedo.

Absorbed radiation accorded to the thermal capacity and emissivity of materials is equally emitted as long wave to the rest of materials. From the albedo characteristics and emissivity of materials in pavements, façades and covers, it is calculated the total of heat produced during the day in energy units of kW/m².

References

- Bohigas, O., 1958. En el centenario del Plan Cerdà. Cuadernos de arquitectura.
- Climate Change. The IPCC Scientific Assessment WMO/IPCC/UNEP. Cambridge University Press, Cambridge.
- Künzli, N., & Pérez, L. (2007). *Els beneficis per a la salut pública de la reducció de la contaminació atmosfèrica a l'àrea metropolitana de Barcelona*. CREAL.
- Martín-Vide, J., 2015. La isla de calor en el Área Metropolitana de Barcelona y la adaptación al cambio climático. Proyecto METROBS 2015. Ed. AMB.
- Mueller, N., Rojas-Rueda, D., Basagaña, X., Cirach, M., Cole-Hunter, T., Dadvand, P., Donaire-Gonzalez, D., Foraster, M., Gascon, M., Martinez, D., Tonne, C., Triguero-Mas, M., Valentín, A., Nieuwenhuijsen, M., 2017. Health impacts related to urban and transport planning: A burden of disease assessment. *Environ. Int.* 107, 243–257.
- Mueller, N., et al., 2019. Changing the urban design of cities for health: the superblock model. *Environ. Int.*
- Rueda, S., 1995. *Ecología Urbana: Barcelona i la seva Regió Metropolitana com a referents*. Ed. Beta Editorial.
- Rueda, S., et al., 2012. *Ecosystemic Urbanism Certification*. Ed. BCNecologia.
- Rueda, S., et al., 2014. *Ecological Urbanism: its application to the design of an eco-neighborhood in Figueres*. Ed. Agencia de Ecología Urbana de Barcelona.
- Rueda, S., et al., 2017. *Les superilles per al disseny de noves ciutats i la renovació de les existents. El cas de Barcelona. Papers-59: Nous reptes en la mobilitat quotidiana. Politiques publiques per a un model més equitatiu i sostenible*. Publicacions IERMB.

Resilience of Transport Systems

Erik Jenelius, Lars-Göran Mattsson, Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

© 2021 Elsevier Ltd. All rights reserved.

Introduction	258
Resilience and Related Concepts	258
Definition	259
Causes and Consequences of Disruptions	260
Tools to Study Resilience	262
Enhancing Resilience	264
Ongoing Trends	265
Concluding Remarks	265
Biographies	266
See Also	266
Relevant Websites	266
References	267
Further Reading	267

Introduction

Our daily lives as well as society at large have become increasingly dependent on critical infrastructure systems providing amenities, such as electric power supply, information and telecommunications (ICT), and not least, mobility and freight transport. Over time, the systems have become more complex and intertwined, with consequences of disruptions that are difficult to identify, understand, and mitigate. Because they are critical to society, their *resilience*, that is, ability to resist, absorb, adapt, and quickly recover from shocks, disruptions, and deliberate attacks, has become an important societal goal.

The transport system can conceptually be divided into three parts (**Fig. 1**):

1. technical and physical infrastructure including rails, roads, terminals, airports, ports, vehicles, signals, signs, etc.
2. direct and indirect users of the technology, that is, travelers, transporters, and society in general,
3. actors responsible for planning, operating and maintaining the infrastructure, and services under given regulations and budget restrictions.

Shocks to the transport system can initiate in any of the three parts and then propagate to the other parts. Infrastructure disruptions are caused by, for example, technical failures (bridge collapses, power outages, vehicle malfunctions, etc.), extreme operating conditions (floods, snow storms, etc.), or deliberate attacks. User disruptions can occur due to economic crises, conflicts, and health crises, such as the Covid-19 pandemic. Disruptions in the management may arise from, for example, political shifts or labor conflicts. Climate change will be an increasing threat primarily to the transport infrastructure. Ironically, the transport system has contributed to this threat through its hitherto almost total dependence on fossil fuels.

Societal shocks can be completely unexpected. While the risk of a pandemic was known, Covid-19 had cascading effects on the economy and the transport system to an extent never seen before. The surprisingly rapid global spread of Covid-19 is, in part, explained by the extreme level of mobility offered by today's transport system ([Musselwhite et al., 2020](#); [Peeri et al., 2020](#)). Most countries sought to suppress the spread by closing borders and restricting mobility locally and globally as well as locking down activities and businesses. This led to a drastic reduction in travel demand, followed by a corresponding reduction in supply. The effects were most pronounced for the aviation industry. Public transport, including rail, was also greatly affected. However, the transport infrastructure was left intact.

Resilience and Related Concepts

[Holling \(1973\)](#) introduced resilience as a term for an ecological system's ability to absorb changes and shocks and still persist. Another ecologist, [Pimm \(1984\)](#), defined resilience as the rapidity with which a system returns to equilibrium after a perturbation. Both these views are typically included in recent definitions of the resilience of transport systems. [Rose \(2007\)](#), for instance, expresses this as a distinction between *static resilience*, a system's capability to maintain its function, and *dynamic resilience*, the rate at which a system returns to equilibrium after a disturbance. [Bruneau et al. \(2003\)](#) understand a resilient system as one that has reduced the probabilities of shocks, that exposed to a shock experiences moderate consequences, and that recovers quickly. A large number of similar definitions have been suggested. We summarize them as follows (inspired in particular by [UNISDR, 2009](#)):

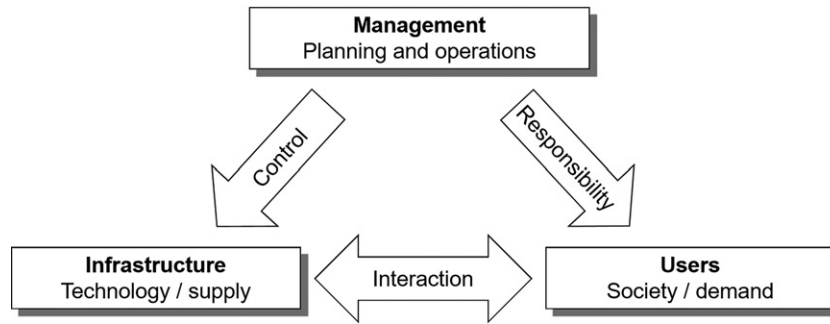


Figure 1 A trisecting model of the transport system.

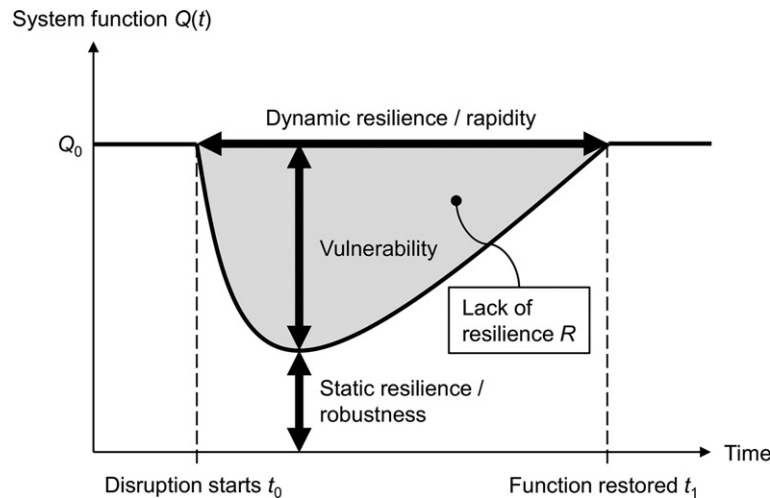


Figure 2 Stylized illustration of resilience and vulnerability of transport systems. Source: Adapted from Bruneau et al., (2003); McDaniel et al., (2008).

Definition

Resilience is the ability of a transport system to prepare for and to withstand, absorb and adapt to shocks, and to recover from the consequences in a timely and efficient manner.

Bruneau et al. (2003) specify four properties that characterize a resilient system: *robustness*, *redundancy*, *resourcefulness*, and *rapidity*. Robustness reflects the ability to withstand shocks without or with only limited degradation of service. Redundancy reflects the extent to which there are alternative units that can function as substitutes and hence absorb the consequences of degradation. Resourcefulness reflects the capacity to identify problems, prioritize actions, and mobilize necessary material and human resources to adapt and recover the system. Rapidity reflects the capacity to recover in a timely and efficient manner. The robustness and redundancy represent the static aspect of resilience, whereas resourcefulness and rapidity represent the dynamic aspect.

The relation between different concepts is illustrated in Fig. 2. The area between the straight horizontal line and the curved line below represents the total loss of function, that is, lack of resilience against a specific shock (Bruneau et al., 2003). (The lack of resilience R can be represented quantitatively as $R = \int_{t_0}^{t_1} [Q_0 - Q(t)] dt$, where Q_0 is the normal and $Q(t)$ the time-varying quality of service (system function) between the time t_0 , when the shock occurs, and the time t_1 , when the function is restored to its pre-shock level). This lack is obviously dependent on ex ante mitigation measures and ex-post adaptation measures that have been taken. The figure does not represent the extent to which the likelihood of shocks has been reduced but presents the situation conditional on that a shock has occurred. This may be termed *conditional resilience*.

There are many related concepts, in particular *risk* and *vulnerability*. Following Kaplan and Garrick (1981), *risk* can be defined as a set of triplets, each one consisting of a scenario, the probability, and the consequence in terms of damage resulting from the scenario. (Formally they define $\text{risk} = \{ \langle \text{scenario}_i, \text{probability}_i, \text{consequence}_i \rangle \mid i = 1, 2, \dots, I \}$). The scenario is a description of a chain of (unwanted, negative) events that can happen. The consequence corresponds to conditional lack of resilience. Measures taken by the system owner to strengthen the system become part of risk management.

Vulnerability of a transport system may be defined as a susceptibility to events that can result in considerable reductions in transport system serviceability (Berdica, 2002) or, with a shorter but similar formulation, as society's risk of transport system disruptions and degradation (Jenelius and Mattsson, 2015). Vulnerability studies aim at finding the weaknesses in a system, estimating the probabilities that they will lead to serious deterioration and the consequences thereof. The term vulnerability is

Table 1 Causes of disruptions in transport infrastructure.

Where Origin	Natural	Human-made unintentional	Human-made intentional
Inside	Technical failure	Accident, handling error, negligence	Malicious insider, labor market dispute
Outside	Natural disaster, earthquake, extreme weather, technical failure	Accident, handling error, negligence	Act of war, terrorist attack, criminal action, prank

Table 2 Examples of weather-related transport system disruptions

Weather condition	Transport mode	Possible infrastructure impacts	Possible effects on users
Heavy snowfall	Road	Drainage blocking, road and bridge destruction, road blocking	Slow driving speed, reduced accessibility, cancelled trips
	Rail	Track damage and blocking, electricity loss, freezing of switches	Stopped and cancelled trains, lower speeds, accidents
	Air	Blocked runways	Delays, cancelled flights
Heavy rain	Road	Drainage blocking, road and bridge destruction, road flooding	Slow driving speed, reduced accessibility, cancelled trips
	Rail	Track damage and blocking	Stopped and cancelled trains, lower speeds, accidents
Drought	Air	Blocked routes, closed airports	Delays, cancelled, and diverted flights
	Road	Forest fires, road damage and blocking	Reduced accessibility, cancelled trips
	Rail	Forest fires, track damage and blocking, electricity loss	Stopped and cancelled trains
	Waterway (inland)	Lower water levels	Slow speed, reduced carrying capacity, reduced accessibility
Heavy wind	Waterway	Ship and port damage, stopped or cancelled ferries	Delays, accidents
	Air	Reduced ground, take-off and landing speed, turbulence	Delays, accidents, discomfort
Sandstorm, volcanic ash	Air	Airspace closed, engine damage	Delays, cancelled, and diverted flights

Source: Adapted from Leviäkangas et al., 2011

often used even if there is no attempt to determine the probability of disruption. When a distinction is necessary, this may be termed conditional vulnerability. A vulnerability study can help the system manager identify the critical elements of a system and which preventive measures that would be most valuable. Vulnerability as opposed to resilience studies are thus more limited and are primarily focused on how the robustness of a system can be strengthened. In Fig. 2, conditional vulnerability corresponds to how much the system function will decrease when subject to a shock, or the normal level of function minus the robustness. (Formally, conditional vulnerability is $Q_0 - \min_{t \in [t_0, t_1]} Q(t)$).

Causes and Consequences of Disruptions

Disruptions in transport systems can take many forms (see Table 1). It is useful to distinguish between disruptions caused by “nature,” technical failures and unintentional errors on the one hand, and disruptions caused by malign individuals on the other hand. In the first case, there is a certain element of randomness as to when and where disruptions occur. Internal threats in the form of technical failures, accidents, and handling errors fall within the domain of system managers with more direct control compared to external causes. In some cases, a combination of both technical and human factors contributes to initiate and exacerbate an incident.

The transport system is also heavily exposed to external threats such as disasters, extreme weather, and technical failures in other infrastructure. The unexpectedness of many of these threats makes it difficult to protect the transport system. Table 2 lists a few examples of weather-related transport system disruptions. As climate change continues, extreme weather can be expected to become more common and its effects on transport more severe (Koetsee and Rietveld, 2009).

External threats also include deliberate attacks such as sabotage, terrorism, and acts of war. Tragically, almost all modes of transport have been subject to deadly terrorist attacks. An intelligent adversary can be expected to look for vulnerabilities in a system and where and when it is poorly protected. A malicious insider can be particularly well-informed about a system’s vulnerabilities. Since the transport system is critical for companies and people, it is also often exposed to labor market conflicts, not least in labor-intensive modes such as air, rail, and public transport. Table 3 lists some notable historical disruptive events and Figs. 3–6 show some of the sites where they occurred.

Different transport subsystems, or modes, have different resilience characteristics. The road transport system provides the majority of all inland passenger and ton kilometers transported. Road transport is characterized by relatively high levels of

Table 3 Examples of transport system disruptions.

Transport mode	Location, date	Initiating event	Transport system damage
Waterway (maritime)	Baltic Sea, 1994	Rough weather, damaged bow visor	Passenger ferry sank, 852 died from drowning and hypothermia
Rail	Madrid, 2004	Terrorist bomb	Train cars exploded, 193 died, some 2,000 injured
Air	Iceland, 2010	Volcanic ash cloud	European airspace closed for more than a week
All	Chile, 2010	Earthquake followed by tsunami	All transport infrastructure seriously damaged, officially 525 died, 25 missing
Public transport and other modes	New York, 2012	Hurricane Sandy	Flooding, subway and other traffic stopped for 5 days or more
Road	Genoa, 2018	Rain, technical failure under investigation	Collapsed bridge, 43 died and 16 injured

Source: Wikipedia retrieved May 15, 2020 Available from: https://en.wikipedia.org/wiki/MS_Estonia , https://en.wikipedia.org/wiki/2004_Madrid_train_bombings , https://en.wikipedia.org/wiki/Air_travel_disruption_after_the_2010_Eyjafjallaj%C3%B6kull_eruption , https://en.wikipedia.org/wiki/2010_Chile_earthquake , https://en.wikipedia.org/wiki/Effects_of_Hurricane_Sandy_in_New_York , https://en.wikipedia.org/wiki/Ponte_Morandi



Figure 3 Collapsed express highway in Santiago, Chile, under the 2010 earthquake. Source: Esteban Maldonado, CC BY-SA 2.0, Available from: <https://commons.wikimedia.org/w/index.php?curid=9670264>.



Figure 4 Remains of one of the trains after the terrorist bomb attack at Atocha Station, Madrid, Spain, 2004. Source: Ramón Peco, CC BY 2.0, Available from: <https://commons.wikimedia.org/w/index.php?curid=4677285>.

independence for the users and redundancy in routes. This implies that the users are well-equipped to handle disruptions through detours and rescheduling. Remote areas may be vulnerable to disruptions at a particular critical location, for example, a bridge, while urban areas can suffer from cascading congestion and queues. Even in dense networks, area-covering disruptions such as floods can have severe consequences (Jenelius and Mattsson, 2015).

Compared to roads, rail traffic is characterized by substantially more regulation and control. The system is dependent on other critical infrastructure, in particular the electric power grid. The dependence on labor, ICT systems as well as mechanical and electric systems is also high. Route redundancy is typically low, and overtaking stopped trains is difficult, if not impossible. These factors contribute to make the rail system relatively sensitive to disruptions.



Figure 5 The collapsed Morandi Bridge in Genoa, Italy, 2018. Source: Salvatore Fabrizio, CC BY 4.0, Available from: <https://commons.wikimedia.org/w/index.php?curid=71622568>.



Figure 6 Overlooking the Icelandic Eyjafjallajökull glacier and the ongoing volcano eruption from Hvolsvöllur, 2010. Source: Boaworm, CC BY 3.0, Available from: <https://commons.wikimedia.org/w/index.php?curid=10056072>

Air traffic is characterized by high safety and security concerns coupled with private market competition and strong dependencies on labor, fuel supply, ICT, and electric systems. The network consists of nodes (airports) with strict regulations and interdependencies between vehicles and journeys, and links (flights) with high degrees of freedom concerning flight path and speed. Disturbances that tend to originate at nodes can lead to knock-on delays and missed connections.

Public transport can offer efficient mobility in urban areas as well as base levels of accessibility in remote areas, and can be provided in the road, rail, waterway, and even air transport systems. Depending on the transport mode, public transport is thus subject to the vulnerabilities of each system discussed earlier. The system is further characterized by the schedule-based operations, and traffic disturbances may lead to long in-vehicle travel times, waiting times, crowding, and denied boarding.

Consequences of disruptions range from annoying delays, missed travel connections, interrupted trips, and delayed goods deliveries, to economic decline due to reduced transport opportunities and, not least, injuries and deaths due to accidents, disasters, or deliberate attacks. It is challenging to evaluate these very different dimensions in a one-dimensional consequence indicator. Cost-benefit analysis of transport policy measures may possibly serve as an example. The effects on accidents, the environment, accessibility, and wider economic benefits are then summarized in a monetary value according to a well-defined procedure. However, shocks have impacts in dimensions not typically included in traditional cost-benefit analysis. Taking Covid-19 as an example, although the consequences for the transport system were very significant, the major consequences were in terms of people's life, health, and the economy. Exceptional events may also affect citizens' perceived security and their confidence in the public institutions' ability to protect them. Such effects are difficult to value in monetary terms. Since different social strata and geographical areas are affected unevenly by a disruption, fairness aspects must also be taken into account.

Tools to Study Resilience

Data on past disruptions can provide valuable information to better prepare for future threats. This is especially relevant for natural and unintentional manmade disruptions. For intentional attacks, historical data has less relevance for where and when future attacks may occur, as an adversary likely wants to surprise a defender.

It is intuitive to conceptualize transport infrastructure as a network or graph, comprising a set of nodes (or vertices), and a set of links (or edges) in which each link connects two nodes. The appropriate network representation of the system depends on the scope

of the analysis. A common representation of the air transport system, for example, is to model airports as nodes and connections or specific flights as links. The basic node and link structure is known as the topology of the network. On top of the topology can be added other properties such as link weights (for example, lengths, travel times or costs) and capacity constraints. The most sophisticated network models involve detailed dynamic, sometimes stochastic, simulation of the interactions between vehicles and the stationary infrastructure (Bell and Iida, 1997; de Dios Ortúzar and Willumsen, 2011).

A network model can be used to analyze the resilience against disruptions represented as failures/removals of nodes or links. A simple measure of network connectivity is the relative size of the largest connected component of the network, that is, the largest number of nodes that can be reached from each other through at least one network route. Another measure is the average distance between all node pairs, where distance is defined as the length of the shortest path connecting each node pair. If the network is not connected, which may happen after link and node removals, the average distance becomes infinite. Latora and Marchiori (2001) introduce an efficiency indicator defined as the average across all node pairs of the reciprocals of the node pair distances. (In mathematical terms, the efficiency metric is defined as $E = \frac{1}{N(N-1)} \sum_{i \neq j \in V} \frac{1}{d_{ij}}$, where V is the node set, N is the number of nodes, and d_{ij} is the network distance between nodes i and j .) This indicator is well-defined also for nonconnected networks, and the change in efficiency can be used as a vulnerability metric. (Formally, the (conditional) vulnerability to a removal of node i can be defined as $E_0 - E_i$, where E_0 is the efficiency of the nominal network and E_i is the efficiency with node i removed).

There is an important connection between the resilience and vulnerability of a transport network and its topological structure, in particular the existence and location of central nodes and links (Reggiani, 2013; Zhang et al., 2015). A simple local measure of node centrality is the *degree*, that is, the number of links connecting directly to the node. In the air transport network example, the degree of a node is the number of routes or flights connecting an airport to other airports. High-degree nodes are sometimes known as hubs. The distribution of degrees among network nodes has attracted considerable interest. The network is said to be scale-free if the degree distribution follows a power law, at least asymptotically, that is, decreases as $k^{-\gamma}$ for large values of the degree k . The exponent γ is typically between 2 and 3. Scale-free networks are considered robust to random failures but vulnerable to successful attacks on high-degree nodes (Lin and Ban, 2013). A global measure of node centrality is *betweenness*, which is the number of shortest paths between all other node pairs that include the node. In the air transport network, high betweenness indicates that the airport is an important transfer location. Degree and betweenness are illustrated in Fig. 7.

The network-modeling paradigm can be extended to include the transport users, in particular travelers and goods. These can be modeled as travel demand between special source (origin) and sink (destination) nodes, which enters the network as link flows. More elaborate network models may incorporate congestion (flow-dependent link costs), demand routing mechanisms, dynamic and stochastic demand, etc. Including transport demand in the network model allows for more nuanced metrics of resilience against node and link removals. A useful measure is the *unsatisfied demand*, defined as the amount of flow, which is unable to reach its destinations. If each link is associated with a travel time (or cost), a meaningful resilience metric is the increase in total travel time (or cost) (Jenelius and Mattsson, 2015). Figure 8 illustrates the potential impacts of link removals.

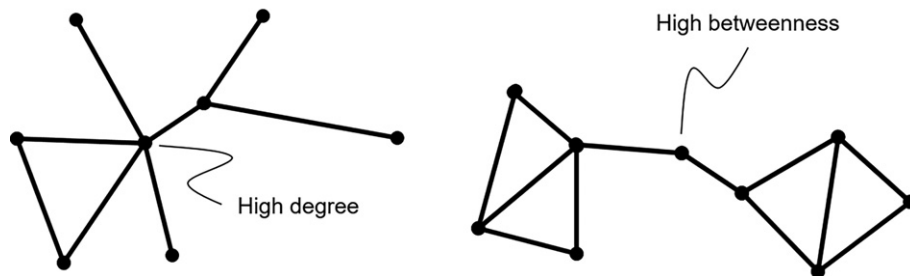


Figure 7 Nodes of high degree (left) and betweenness (right) centrality

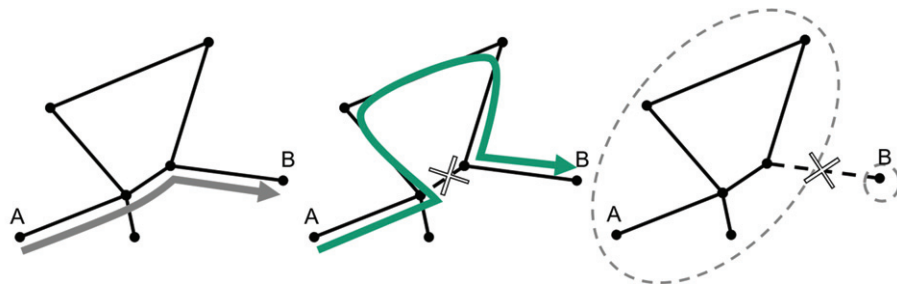


Figure 8 Left: Undamaged network, shortest path flow from source: A to sink B. Middle: Rerouting due to link removal, increased shortest path length. Right: Network fragmentation due to link removal, unsatisfied demand.

At the finest level of detail, agent-based models can represent each traveler and vehicle as an individual object with unique properties. Such frameworks can be used to capture the dynamic dimension of resilience, including the deterioration and eventual restoration of system function associated with a disruption (Malandri et al., 2018).

Different types of disruptive events require different modeling approaches. Disruptions caused by “nature” or unintentional human actions are often appropriately modeled as removals of randomly selected nodes or links. Random removals of multiple links or nodes can be studied as well but increases the combinatorial complexity.

Disruptions caused by intentional human interventions are likely targeted towards network components where they can achieve as much damage as possible. Data on past attacks will then say little about future attacks. It may be better to analyze who may be interested in and capable of performing such an attack. A commonly studied strategy is to attack the nodes with the highest degree, thereby disrupting as many local connections as possible. Another strategy is to target the global connectivity by attacking the nodes with the highest betweenness centrality or transport flow. Some studies evaluate resilience in terms of how network function deteriorates as increasing numbers of nodes are removed (for example, Zhang et al., 2016).

A characteristic feature of intentional attacks is that the strategy for targeting network components can adapt to measures taken by society to protect the system. Thus, protecting the most vulnerable components from attacks could divert attacks to other, less protected components. Here, *game theory* can provide insights into the risk for attacks across the transport. For example, consider a situation when an antagonist wants to disrupt freight transport by attacking a network link, while the transporter wants to route the freight to minimize the expected losses. In this case, optimal strategies for the transporter can be found by studying the *Nash equilibria* of the conflict, that is, strategy combinations under which no actor can single-handedly increase its payoff by changing strategy (Bell et al., 2008).

Enhancing Resilience

Since resilience is a multifaceted concept, a set of proactive and reactive measures is required to enhance it. To some extent, potential threats may be avoided. For actions such as investments in new railways, roads, and other physical infrastructure, national transport authorities have the capacity to make the decisions. To some extent, they can avoid sites that are exposed to natural threats like floods, landslides, and adverse weather. However, trying to protect the transport system from any conceivable threat or disruption would be prohibitively expensive. The most challenging example is climate change. If effective measures to stop or at least limit climate change are implemented, the expected increase in extreme weather can be reduced. To realize such a policy, decisions outside the transport system at an international political level are necessary.

For shocks like earthquakes, tsunamis, and deliberate attacks, it is impossible to predict with accuracy where and when they will occur. Other shocks are hardly possible to imagine in advance. This was the case with the transport impacts of Covid-19 through the measures taken to limit its spread. Therefore, reactive measures to withstand, absorb, and adapt to shocks are also necessary. The properties that characterize a resilient system, robustness, redundancy, resourcefulness, and rapidity, are helpful in structuring such considerations (Bruneau et al., 2003).

Robustness and redundancy are primarily associated with the static aspect of resilience. A robust transport system can withstand shocks with no or limited degradation of service. Roads, bridges, railways, drainage, and other critical technical elements can be built to withstand and absorb considerable external stress and planned to be regularly inspected, repaired and maintained. The most obvious example of increasing the resilience of a transport network through redundancy is to add links to increase the number of alternative routes. Different transport modes can partly substitute for each other. Reserve capacity for personnel and vehicles in, for example, public transport, can help coping with increased sick leave or sudden technical problems. The tools discussed in the previous section are valuable for identifying the elements in the transport networks that are most critical to strengthen and maintain, and the degree to which increased redundancy can absorb the consequences of degradations. Table 4 provides examples of policy instruments to enhance resilience against flooding in an Indian context.

Resourcefulness and rapidity are primarily associated with the dynamic aspect of resilience, that is, reducing the total consequences through timely mitigation and restoration. Resourcefulness reflects the ability to identify problems, prioritize actions, and mobilize necessary material and human resources to adapt and recover the system. The capacity to handle disruptions can be strengthened through a high level of readiness, including relevant equipment and personnel who are well-trained in handling the

Table 4 Examples of actions to enhance resilience against flooding

Phase	Mechanism	Action	Effect on transport
Proactive	Avoidance	Restricting development in low lying or vulnerable areas Slum relocation and rehabilitation	Fewer travelers exposed to flooding Fewer travelers exposed to flooding
	Robustness	Providing proper drainage facilities at vulnerable areas Replacement of impermeable road surface with permeable material in vulnerable areas	Decreased road inundation Decreased road inundation
	Redundancy	Construction of redundant infrastructure	More route options
Reactive	Adaptation	Rerouting people during flooding	Travelers less affected by flooding

Source: Based on Vajjarapu et al., 2020.

Table 5 Examples of recommended preventive actions to enhance resilience to extreme weather events*Preventive actions to enhance resilience to extreme weather events*

Consider whether there are potential “single points of failure” in the networks, which leave parts of the country at risk of having vital economic and social links severed. Identify a “critical network,” comprising those routes which are of national economic significance and should be maintained to a higher level.

All transport operators should have contingency plans to cope with extreme weather events.

Contingency plans should be regularly rehearsed and progressively extended.

All transport operators should ensure they have clearly agreed channels for receiving weather and flood forecasts.

All transport operators and authorities should implement a dedicated passenger and user communications plan for times of transport disruption.

In the face of an extreme weather event, transport operators should plan for the best practicable service which they can realistically deliver.

Transport operators should have checklists of resources which they will need as part of their recovery effort.

Rail infrastructure operator should liaise with its electricity suppliers to trace through the routes and sub-stations to ensure adequate system redundancy.

Contingency rail timetables should be prepared to match different weather events.

Rail infrastructure operator should have adequate resources available for clearing fallen trees.

Source: Adapted from Department of Transport, 2014

equipment and dealing with accidents and disasters. Adequate information systems to alarm rescue forces about disruptions and help them coordinate relief efforts are important. In addition, resources must be available to reduce the consequences for those directly or indirectly affected by a disruption. Travelers often have to wait for hours at airports, on stopped trains or in car queues. Such consequences can be partially mitigated by providing travelers with relevant information to facilitate their own decisions. To ensure that the transport system is well prepared to minimize the consequences of disruptions, a number of measures must be implemented. **Table 5** gives some examples from the United Kingdom of recommended preventive measures to improve the transport system’s resilience to extreme weather.

Ongoing Trends

Continued technological development can be expected to make infrastructure systems increasingly intertwined. Among the transport subsystems, the road system has been the least dependent on other infrastructure systems. The development of connected autonomous vehicles may change this through the need for wireless communication with other vehicles, infrastructure and global satellite navigation systems. Moreover, if a vehicle loses contact with the support systems, it must be able to stop safely. Mobility-as-a-service is a new concept that requires reliable wireless communication for ordering and operating the vehicles. Since the communication is vulnerable to intentional and unintentional disturbances, it is necessary to make the channels reliable and secure. Another important trend affecting particularly the road transport system is the electrification of vehicles, which increases the dependence on the electric power system. A resilient road transport system with self-driving electrified vehicles requires both interference-free wireless communication and reliable electricity supply.

Climate change has far-reaching consequences for transport. Extreme weather such as hurricanes, heavy rain, and drought are forecasted to become more common. This will lead to a higher frequency of disruptions in the transport networks due to, for example, storm-felled trees, landslides, floods, and forest fires. Further, policies aimed at reducing carbon emissions and reaching the Agenda 2030 sustainable development goals will by necessity create significant shocks to transport users through changes in vehicle technologies, regulations, taxes, fuel prices, ticket and shipping prices, etc.

The impact of the Covid-19 pandemic on mobility and demand for transport has been profound. To stop the spread of the virus, authorities locked down many activities and drastically restricted physical contacts and mobility. The responses to these actions illustrate the flexibility and adaptability that exists in society. The biggest adjustment made by individuals in both private and working life is arguably the change from physical to virtual contacts. Since this requires access and familiarity with, for example, video conferencing systems, it is reasonable to assume that many will continue to meet virtually also after Covid-19 once this effort has been made. Whether it will reduce travel in the long run remains to be seen. Covid-19’s stimulus of e-commerce is another effect that can be lasting. In the aviation sector, demand fell to almost zero and many companies have gone bankrupt or needed great financial support from governments to survive. When air traffic resumes to a fairly normal level, ticket prices and freight rates are likely to be higher than before the pandemic. Combined with the debate on the climate effects of aviation, this may lead to a significant, lasting reduction in air traffic. The increased familiarity with video conferences may further contribute to this. The impacts on transport resilience of these potential mobility effects are difficult to speculate about. However, what we have learned from Covid-19 is that human adaptability and ingenuity are very large when sudden changes occur. This is an unplanned and unexpected form of resilience.

Concluding Remarks

By resilience is meant a transport system’s ability to prepare for and to withstand, absorb and adapt to shocks, and to recover from the consequences in a timely and efficient manner. The significance of transport systems makes their resilience an important policy

issue. Transport also has a key role for other infrastructure systems in facilitating rescue and repair personnel to quickly get in place to handle the consequences after a disruption. The severity of shocks to the transport systems can vary widely, as well as the causes. Different transport systems exhibit different characteristics from a resilience perspective, depending, for example, on how the behavior of the users is regulated. In the road system, users can often make detours to avoid disrupted links. The rail system is substantially more regulated and rerouting is decided at the level of traffic control.

Research on resilience, risk and vulnerability of transport systems has exploded the last two decades. Tools to study resilience have been developed that can help identifying vulnerabilities in the systems. The tools are also helpful in evaluating measures to enhance resilience. Hopefully, the implementation of new research findings will help decision-makers to make the transport systems less vulnerable and more resilient.

Biographies



Erik Jenelius holds an MSc degree in engineering physics and a PhD degree in infrastructure from KTH Royal Institute of Technology, Stockholm, Sweden. He is an Associate Professor in Public Transport Systems and Head of the Division of Transport Planning at the same university. His research interests include data-driven traffic management, sustainable urban mobility, transport network and resilience analysis, and public transport systems planning and operations.



Lars-Göran Mattsson holds an MSc degree in engineering physics and a PhD degree in optimization from KTH Royal Institute of Technology, Stockholm, Sweden. He is a Professor Emeritus in Transport Systems Analysis at the same university. His main research interest is systems analysis for the public sector. In addition to transport research, this includes research in urban and regional planning, health care, and economics. A particular area of research is risk, vulnerability, and resilience of critical infrastructure.

See Also

Valuation of Travel Time Variability Using Scheduling Models; Resilience in Freight Transport Networks; Traffic Incident Management; Traffic Incident Detection; Advanced Travelers Information Systems (ATIS); Travel Time Reliability; Advanced Traveler Information Systems

Relevant Websites

www.instr.org

INSTR, International Symposium on Transport Network Reliability

www.trb.org

TRB, Transportation Research Board

www.gov.uk/government/organisations/department-for-transport

UK Department for Transport

www.undrr.org

UNDRR, United Nations Office for Disaster Risk Reduction

www.transportation.gov

US Department of Transportation

References

- Bell, M.G.H., Iida, Y., 1997. *Transportation Network Analysis*. Wiley, Chichester.
- Bell, M.G.H., Kanturska, U., Schmöcker, J.-D., Fonzone, A., 2008. Attacker-defender models and road network vulnerability. *Philos. Trans. Royal Soc. A* 366, 1893–1906.
- Berdica, K., 2002. An introduction to road vulnerability: What has been done, is done and should be done. *Transp. Policy* 9, 117–127.
- Bruneau, M., Chang, S.E., Eguchi, R.T., et al., 2003. A framework to quantitatively assess and enhance the seismic resilience of communities. *Earthquake Spectra* 19 (4), 733–752.
- Department for Transport, 2014. *Transport Resilience Review: A review of the resilience of the transport network to extreme weather events*. Available from: <https://www.gov.uk/government/publications/transport-resilience-review-recommendations>.
- de Dios Ortúzar, J., Willumsen, L.G., 2011. *Modelling Transport*, 4th Edition.. Wiley, Chichester.
- Holling, C.S., 1973. Resilience and stability of ecological systems. *Ann. Rev. Ecol. System.* 4, 1–23.
- Jenelius, E., Mattsson, L.G., 2015. Road network vulnerability analysis: Conceptualization, implementation and application. *Comp. Environ. Urban Sys.* 49, 136–147.
- Kaplan, S., Garrick, B.J., 1981. On the quantitative definition of risk. *Risk Anal.* 1 (1), 11–27.
- Koetsee, M.J., Rietveld, P., 2009. The impact of climate change and weather on transport: An overview of empirical findings. *Transp. Res. Part D* 14, 205–221.
- Latora, V., Marchiori, M., 2001. Efficient behavior of small-world networks. *Phys. Rev. Lett.* 87 (19), 198701.
- Leviäkangas, P., Tuominen, A., Molarius, R. et al., 2011. Extreme weather impacts on transport systems. *VTT Working Papers* 168, ISBN 978-951-38-7509-1.
- Lin, J., Ban, Y., 2013. Complex network topology of transportation systems. *Transp. Rev.* 33 (6), 658–685.
- Malandri, C., Fonzone, A., Cats, O., 2018. Recovery time and propagation effects of passenger transport disruptions. *Physica A* 505, 7–17.
- McDaniels, T., Chang, S., Cole, D., Mikawoz, J., Longstaff, H., 2008. Fostering resilience to extreme events within infrastructure systems: Characterizing decision contexts for mitigation and adaptation. *Global Environ. Change* 18, 310–318.
- Musselwhite, C., Avineri, E., Susilo, Y., 2020. The corona virus disease COVID-19 and implications for transport and health: Editorial. *J. Transp. Health* 16, 1–3.
- Peeri, N.C., Shrestha, N., Rahman, S., et al., 2020. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? *Int. J. Epidemiol.* 49 (3), 717–726.
- Pimm, S.L., 1984. The complexity and stability of ecosystems. *Nature* 307, 321–326.
- Reggiani, A., 2013. Network resilience for transport security: Some methodological considerations. *Transp. Policy* 28, 63–68.
- Rose, A., 2007. Economic resilience to natural and man-made disasters: multidisciplinary origins and contextual dimensions. *Environ. Hazards* 7 (4), 383–398.
- UNISDR, 2009. *UNISDR Terminology on Disaster Risk Reduction*. United Nations International Strategy for Disaster Risk Reduction, Geneva, Available from: https://reliefweb.int/sites/reliefweb.int/files/resources/Full_Report_2010.pdf.
- Vajjarapu, H., Verma, A., Allirani, H., 2020. Evaluating climate adaptation policies for urban transportation in India. *Int. J. Disast. Risk Reduct.* 47 (forthcoming). <https://doi.org/10.1016/j.ijdr.2020.101528>.
- Zhang, J., Hu, F., Wang, S., Dai, Y., Wang, Y., 2016. Structural vulnerability and intervention of high speed railway networks. *Physica A* 462, 743–751.
- Zhang, X., Miller-Hooks, E., Denny, K., 2015. Assessing the role of network topology in transportation network resilience. *J. Transp. Geog.* 46, 35–45.

Further Reading

- Bešinović, N., 2020. Resilience in railway transport systems: a literature review and research agenda. *Transp. Rev.* 40 (4), 457–478.
- Faturechi, R., Miller-Hooks, E., 2015. Measuring the performance of transportation infrastructure systems in disasters: A comprehensive review. *J. Infrastruct. Sys.* 21 (1), 04014025.
- Gu, Y., Fu, X., Liu, Z., Xu, X., Chen, A., 2020. Performance of transportation network under perturbations: Reliability, vulnerability, and resilience. *Transp. Res. Part E* 133, 101809.
- Mattsson, L.-G., Jenelius, E., 2015. Vulnerability and resilience of transport systems: A discussion of recent research. *Transp. Res. Part A* 81, 16–34.
- Murray, A.T., Matisziw, T.C., Grubisic, T.H., 2008. A methodological overview of network vulnerability analysis. *Growth Change* 39 (4), 573–592.
- Taylor, M., 2017. *Vulnerability Analysis for Transportation Networks*. Elsevier, New York.
- Twumasi-Boakyje, R., Sobanjo, J., 2019. Civil infrastructure resilience: state-of-the-art on transportation network systems. *Transp. A: Transp. Sci.* 15 (2), 455–484.
- Zhou, Y., Wang, J., Yang, H., 2019. Resilience of transportation systems: Concepts and comprehensive review. *IEEE Trans. Intel. Transp. Sys.* 20 (12), 4262–4276.

Impact of Shipping to Atmospheric Pollutants: State-of-the-Art and Perspectives

Daniele Contini, Eva Merico, Institute of Atmospheric Sciences and Climate, ISAC-CNR Lecce, Italy

© 2021 Elsevier Ltd. All rights reserved.

Introduction	268
Overview of Shipping Sector: Some Statistics	268
Global and Local Regulatory Framework	269
Assessment of Shipping Impact to Atmospheric Pollutants and Human Health	270
Future Perspectives and Conclusions	273
Acknowledgments	274
See Also	274
References	274
Further Reading	275

Introduction

Shipping is a growing transportation sector due to its reliability, affordability and global nature. Vessels getting bigger (i.e., tankers, bulk carriers, and container ships), faster and more efficient can carry billions of tons of goods along a few principal trade routes of the global trade, being shipping the most efficient mode of transportation. For example, in one year, a single large containership could carry over 200,000 container loads of cargo (<http://www.worldshipping.org/>). World seaborne trade volumes are estimated at about 80% of the global goods trade (UNCTAD, 2016). In Europe, currently almost 90% of the European Union (thereafter EU) import and export freight trade is seaborne (Karl et al., 2019), with a total of 2.1 million vessels calling in main European harbors in 2016 (EUROSTAT, 2019). Tourist shipping is also increasing in terms of passengers and size of cruise ships are continuously increasing (Contini et al., 2015), with a total of 397 million passengers in EU harbors in 2016, and a rise of 0.4% compared to the previous year (EUROSTAT, 2019).

Apart from economic growth and energy efficiency of ships, the environmental issues are needed to be investigated at different levels. It is important to discriminate between local impacts (near harbor areas) and open-sea traffic influence at continental and global scale. At local scale, harbor development could lead to a conflict of interest between stakeholders and local population (most of the harbors and terminals are located close to the urban areas), however, harbors are committed to being “green” while building prosperity for present and future generations in a mutual concept of “sustainability” and “economic and social scope.” In this perspective, because of the pressure on the environment and potential health effects, air quality was underlined as the main and fundamental goal within harbor management.

Shipping emissions predominantly consist in primary and secondary particulate matter (PM, mainly particles of small size), black carbon (BC), sulfur dioxide (SO₂), nitrogen oxides (NO_x), volatile organic compounds (VOCs), carbon oxide (CO), and carbon dioxide (CO₂). Trace elements like metals in PM (e.g., Vanadium - V - and Nickel - Ni - that are considered tracer of this emission) and PAHs (polycyclic aromatic hydrocarbons), in gas and particulate phases) are also important compounds released by ships. Further, taking into account that, for decades, heavy fuel oil (HFO) has been the cheapest, most efficient but also the most polluting fuel, ocean vessels powered by HFO have emitted greenhouse gases (GHGs). These pollutants are known to deteriorate air quality producing adverse health effects (Corbett et al., 2007; Oeder et al., 2015; Liu et al., 2016) and negative effects on climate (Fuglestedt et al., 2009; Eyring et al., 2010).

Thereby, atmospheric pollution due to ships raised in the last decades concerns in the scientific community and in policy-makers, and global efforts have been devoted to regulate and prevent risks from ship emissions, especially in harbor cities and coastal areas.

This work reviews the current knowledge of the impacts of shipping to atmospheric pollutants emissions, contributions to local air quality in coastal areas and harbor-cities, and the impacts to health indicators. The effects of legislation standards, relative to sulfur content in marine fuels and enforcement of Emission Control Areas (ECAs), on atmospheric pollution due to shipping is discussed. Furthermore, future perspectives and projections related to the increase of energy efficiency of ships and improvement on quality of fuels will be discussed.

Overview of Shipping Sector: Some Statistics

The massive increase in shipping has been driven by the growth in world trade depending on both institutional and technological factors. In 2018, Singapore, the Republic of Korea, Malaysia and the United States had the best worldwide connected hubs even if the top-20 bilateral connections were intra-regional, namely within Europe and within Eastern and South-Eastern Asia. Although financial crisis in 2009 and a relatively slow growth recorded since 2013, international seaborne trade volume reached 11 billion tons in 2018 (+2.7% compared to 2017), container traffic accounted for 793 million twenty-foot equivalent units (TEUs), that is larger than 4.7% with respect to the previous year (UNCTAD, 2019). As consequence of global trade, over recent years, tonnage has

increased considerably in all categories, especially for bulk carriers (about 71% of goods—coal, ores, grains, containers—were transported by dry cargoes in 2018), with an only decline for general cargo (i.e., oil tanker). Finally, in the same year the majority of harbor calls of ships, including ferries, roll-on roll-off, and passenger ships was from Norway.

Global GHG emissions, expressed on a CO₂-equivalent (CO₂e) basis, of total shipping registered by the International Maritime Organization (thereafter IMO) decreased in relative terms from 3.2% in 2007 to 2.5% in 2012. For 2012 this corresponds to 961 Mtons of CO₂e (816 Mtons are relative to international shipping), and about 97.6% is due to the sole CO₂ emissions (IMO, 2015). The multi-year (2007–12) average annual totals emissions are 20.9 Mtons and 11.3 Mtons for NO_x (as nitrogen dioxide - NO₂) and sulfur oxides, SO_x (as SO₂), from all shipping (IMO, 2015), representing about 15% and 13% of global NO_x and SO_x from anthropogenic sources reported in the IPCC Fifth Assessment Report (AR5). A successive study (Johansson et al., 2017) showed that global shipping emissions in 2015 were comparable to that of 2012 for NO_x but were lower (9.7 Mtons) for SO_x as a consequence of the use of low-sulfur content fuels in specific areas. Global emissions in 2015 (Johansson et al., 2017) were 1.5 Mtons (PM_{2.5}, i.e., particles with aerodynamic diameter less than 2.5 µm) and 1.35 Mtons (CO). The largest contributors at global level (more than 80% of the total) are container ships, cargo carriers, and tankers (i.e., emissions from commercial shipping). Broadly speaking, the largest emission of PM_{2.5} and SO_x, per unit area, occur in China Sea, in the sea areas of southern and south-eastern Asia, in the Red Sea, in the Mediterranean, in north Atlantic in proximity of EU coast, in the Gulf of Mexico and the Caribbean Sea, and along the west coast of United States. Shipping emissions over Europe represent 16%, 11%, and 5% of total NO_x, SO_x, and PM₁₀ (i.e., particles with aerodynamic diameter less than 10 µm) emissions even if there is a certain variability associated to the emission databases used (Russo et al., 2018). Given its importance in both health and radiative forcing, research efforts have also been devoted to investigate BC emissions from shipping that represent about 2% of global BC emissions (Lack and Corbett, 2012).

Global and Local Regulatory Framework

Environmental awareness on harbor system began to be of significant relevance as of the 1990s. The first adoption of institutional measures, by proposing a series of recommendations regarding environmental issues, was introduced in 1998 by the American Association of Ports Authorities (AAPA), which was based on an alliance between harbors of the United States, Canada, the Caribbean, and Latin America. Similarly, in Europe, the European Sea Ports Organization (ESPO) in 1994, published a first version of the Environmental Code of Practices.

As agency of United Nations, the International Maritime Organization, born in 1948 as Inter-Governmental Maritime Consultative Organization, or IMCO, and then changed to IMO in 1982, adopted the “International Convention for the Prevention of Pollution from Ships,” as modified by the Protocol of 1978 reaching to MARPOL 73/78. In 1997, a new Annex VI “Regulations for the Prevention of Air Pollution from Ships” was added entering into force on 19 May 2005. Successive amendments have been done for progressive improvement on the quality of fuels used for shipping to reduce NO_x and SO_x and PM emissions; furthermore, designated ECAs were foreseen. Regulation 14 sets the sulfur content of any fuel oil used on board ships, outside ECAs, not exceeding the following limits: 4.5% m/m prior to 1 January 2012; 3.5% m/m on and after 1 January 2012; and 0.5% m/m on and after 1 January 2020 (known as “IMO 2020”). A summary of regulations on the maximum sulfur content of marine fuels is shown in Fig. 1.

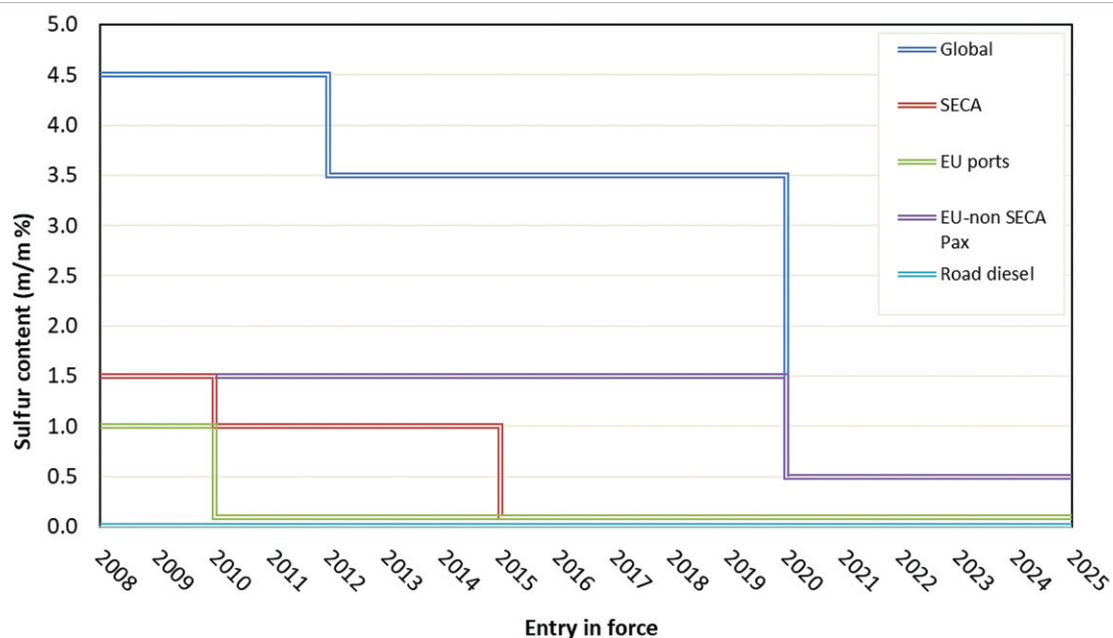


Figure 1 Sulfur content limits in marine fuels according to IMO and EU legislation.

Regarding NO_x, in October 2016, two new NO_x ECAs (the Baltic Sea and the North Sea) were introduced. As of 1 January 2021, within these areas all ships must respect defined mandatory engine standards or equivalent NO_x emission reduction technologies according to Tier III standards.

In addition to global IMO regulations, there are some local initiatives too. In Europe, after a long legislative effort, the latest limits for marine fuels are included in Directive 2016/802/EU, introduced by Directive 2012/33/EU amending Directive 1999/32/EC. As of 1 January 2015, EU zones designated as SO_x—ECAs (the Baltic Sea and the North Sea) have to ensure that ships use fuels with a sulfur content of no more than 0.1%. Furthermore, EU non-SECA passenger ships should use fuels with maximum sulfur content of 1.5% m/m and from 2010 onwards all ships at berth in EU harbors for more than 2 hours should use fuels with sulfur content of less than 0.1% by weight. China established (regardless of the MARPOL Convention) a sulfur limit of 0.5% m/m applied in some steps, designating three areas along the coastline—the Pearl River Delta, the Yangtze River Delta and the Bohai Rim waters coastline—defined “Domestic Emission Control Areas” (DECAs). First, as of 1 April 2016, the implementation of the regulation was effective only for berthing in specific harbors within DECAs, however, this limit was still higher compared to EU and US legislation (0.1% m/m). In 2018, extension to all harbors in the designated DECAs was applied until, starting January 1, 2019, all ships entering China’s coastal waters, during all operations, have been compliant with the limit of 0.5% m/m. Finally, several national initiatives have been proposed for future years (i.e., limits for inland waterways, extension of DECAs geographical boundaries, future IMO ECA implementation) in order to cope severe air pollution caused by shipping. In the United States, since 1 January 2014, the “Regulated California Waters” Phase II imposes to use 0.1% m/m sulfur content marine distillates within 24 nautical miles of the California coasts. In addition, CARB (California Air Resources Board) requires some ship categories (container, passenger, and refrigerated cargo) to connect shore power during berthing in California harbors.

To deal with the new IMO requirements ship owners have essentially three options: change to cleaner (with lower sulfur level) or alternative (i.e., liquefied natural gas, LNG) fuels, application of retrofitting technologies (i.e., scrubbers) and cold ironing. Each of these has strengths and weaknesses. Scrubbers require large initial investments due to installation on old engines of existing ships. According to a survey (EGCSA, 2018), exhaust gas cleaning systems are the best solution for the biggest ship companies (i.e., Maersk CEO), in fact 983 vessels (especially bulk carriers) were equipped with scrubbers as of 31 May 2018. Instead, although alternatives fuels could be a valid option, their availability and bunkering are still lacking, as well as potential incompatibility problems with current vessel’s engines. Finally, shore side electricity at berthing has been adopted in several harbors in Europe (i.e., Antwerp, Gothenburg), North America (Los Angeles, Vancouver, Seattle), in Alaska and California. This solution is recognized to reduce not only NO_x, SO_x, and PM emissions inside harbors near populated coastal cities, but also noise. However, barriers are represented mainly by energy and infrastructure costs, taxes, and technical issues due to different standards for the plug-in systems (voltage and frequency). In addition, the potential of shore side energy production through renewables should be further investigated.

Assessment of Shipping Impact to Atmospheric Pollutants and Human Health

The impact of shipping should be discussed considering, separately, emissions, contributions to concentrations (i.e., effects on air quality), and health effects. These are correlated but not linearly so that a reduction of emissions has different effects on air quality in specific sites depending on the distance from the emissions and the typical transport and dispersion conditions due to winds. Health effect are not related only to concentration of pollutants but to population-weighted concentrations (i.e., to population exposure) meaning that a degradation of air quality in a specific area has very different impact on public health depending on the density of population in that area.

Global pollutant emissions from ships are mainly due to commercial traffic (but this could not happen at local scale) and about 70% of these are released within 400 km from coastlines (Eyring et al., 2005). In harbor ship emissions represent only a small fraction of the global emissions associated with shipping (Dalsøren et al., 2009). However, they can have important environmental effect on coastal regions in Europe, Asia and USA, which often have harbors located near urban and industrial centers. They come from different sources: emissions during maneuvering, during hoteling because several ships at berth needs to power their services with auxiliary engines, harbor-related emissions during loading/unloading of ships including road traffic in the internal area of harbors. This last kind of emission is strictly related to harbor logistic and infrastructures. Hoteling phase is the largest contributor to local emissions in several harbors considering that this phase lasts generally much more than maneuvering. Recent studies in some Mediterranean harbors (Merico et al., 2017, 2019) showed that, at local scale, showed that in harbor emissions of PM and NO_x are typically comparable with those of road traffic (of a small/medium size town), emissions of SO₂ are generally significantly larger than those of road traffic because of the much more stringent standard of sulfur content in road vehicles fuels, emissions of CO are generally lower than those of road traffic. Globally, the implementation of the new 2020 standard will produce a 75% reduction in SO_x emission from ships and a reduction of about 49% of particulate matter including both primary and secondary sulfate (Sofiev et al., 2018). No significant reduction is foreseen for ship-related emissions of NO_x and CO₂. Recent studies in DECA area of Pearl River Delta (China) showed that the use of low sulfur fuels could actually increase VOCs emission factors of ships and this threaten to increase ozone (O₃) formation in harbor cities (Wu et al., 2020).

The influence of ship traffic is not easily identifiable in most areas where land-based sources (industrial settlements, railway, and urban agglomerates) co-exist and/or for complicated meteorological conditions. The quantification of the impact of harbors on local air quality has been performed with different approaches such as large (global, continental, and/or national) scale source-oriented modeling (Monteiro et al., 2018; Sofiev et al., 2018) or at local scale (Merico et al., 2019) using transport and dispersion codes based on estimates of shipping emissions. Other studies rely on receptor-oriented models based on specific measurement

campaigns in coastal or harbor areas and statistical methods or receptor models to estimate shipping contributions. Widely used receptor-oriented approaches are the use of high temporal resolution measurements correlated with wind conditions and ship traffic (Contini et al., 2011; Ledoux et al., 2018) or the use of chemical composition of PM and specifically the V concentration and the V/Ni ratio considered as tracer of ship emissions (Cesari et al., 2014; Viana et al., 2009). Each method has advantages and drawbacks and particular care should be placed in comparing results obtained with different approaches and having different uncertainty. Scientific studies based on receptor-oriented approaches for the evaluation of shipping contributions to PM_{2.5} or PM₁₀ in different coastal/harbor sites worldwide have been reviewed in this work and results are reported in Fig. 2 in terms of relative impacts compared to measured concentrations. In some areas more than one site have been analyzed or the same site was involved in more than one study. In these cases an interval of impacts is reported in Fig. 2.

Contributions to PM_{2.5} range between 3% and 9% in US sites, between 0.2% and 14% in Europe and are generally larger in East Asia especially near busy harbors in China (between 2.4% and 25%). In Europe there is a clear gradient with larger contributions in Mediterranean Sea compared to northern Europe. Another aspect is that the contributions to PM₁₀ are generally lower (in relative terms) compared to those to PM_{2.5}. This is because ship emissions are dominated by particles of small size. Limited studies compare the impact of shipping to particle mass concentrations (PM_{2.5} or PM₁₀) with impact to particle number concentration (PNC). These are not expected to be the same because mass concentrations are dominated by relatively large particles (diameters > 1 µm), instead, PNC is dominated by ultrafine particles (diameter < 0.3 µm) or nanoparticles (diameter < 0.1 µm) that have low contribution to aerosol mass. The results of some studies in the Mediterranean area in which impacts to PNC and PM_{2.5} are estimated using the same receptor-oriented approaches are reported in Fig. 4. These show that shipping emission contributions to PNC are expected to be 3–4 times larger than those to PM because combustion particles emitted by ships are dominated (in number) by ultrafine particles as observed also in other studies (Diesch et al., 2013; Gobbi et al., 2019).

The studies regarding impact of ships to gaseous pollutant concentrations done using receptor-oriented approaches are more limited. A review of current data for Europe is reported in Fig. 3. Results show that the relative contributions to SO₂ and NO_x are generally larger compared to the contributions to PM_{2.5} or PM₁₀. Contribution to CO is typically comparable with that to PM (Merico et al., 2019). At local scale, ship emissions bring to a depletion of O₃ because of titration due to NO releases (Merico et al., 2016; Ledoux et al., 2018). This is a local scale effect because at larger distances from the emissions, in diurnal hours, the excess of NO₂ due to ships could actually increase the production of O₃, especially during summer. Large-scale numerical modeling results indicated that shipping emissions enhance surface O₃ concentration by 5%–10% in the Mediterranean Sea (Aksoyoglu et al., 2016). Considering current thresholds of air quality standards in Europe and in several other areas worldwide, NO_x emissions from ships could lead to exceedances of legislation standard for NO₂ (or NO_x) concentrations representing a critical aspect that should be addressed in future mitigation policies.

There are evidences that the use of low sulfur fuels reduced the contribution to SO₂ concentrations and to secondary and primary PM in coastal areas (Tao et al., 2013; Contini et al., 2015; Liu et al., 2018). It has also been observed a reduction in V concentrations even if total contribution to metals in PM₁₀ was not significantly affected (Gregoris et al., 2016). Furthermore, shipping emissions could significantly contribute to PAHs in gaseous and particulate phases. In some Mediterranean harbors contributions from 50% to 70% (locally) was estimated and no significant reduction as a consequence of the use of low sulfur fuel was observed (Donato et al., 2014; Gregoris et al., 2016).

There are evidences, from epidemiological studies, of a negative impact of shipping to health. Estimates are done using population-weighted concentrations of ship-related pollutants and specific exposure-response (ER) functions. There are significant variability on the results obtained according to the spatial resolution of the analysis, to the population databases used and to the ER functions considered. However, all studies indicate relevant influence of shipping on mortality and morbidity. Globally, about 60,000 premature deaths in 2010 and about 90,000 premature deaths in 2012 due to cardiopulmonary diseases and lung cancer were attributed to exposure to particulate matter emitted from international shipping (Corbett et al., 2007; Winebrake et al., 2009), and 400,000 premature deaths per year due to shipping are predicted for 2020, under business-as-usual assumptions (Sofiev et al., 2018). In east Asia between 14,500 to 37,500 premature deaths were estimated as due to PM_{2.5} related to shipping (Liu et al., 2016). A study in the area of Sidney (Australia) showed for 2010/2011 that about 1.9% of the region-wide annual average population weighted-mean concentration of all natural and human-made PM_{2.5} was attributable to ship exhaust, and up to 9.4% at suburbs close to harbors. These figures lead to an estimate of 220 years of life were lost as a result of ship exhaust-related exposure (Broome et al., 2016). The use of low sulfur fuels decreases average ship-related concentrations (mainly for PM and SO₂), consequently, population exposure is lower and this reduces health effects of shipping emissions (Broome et al., 2016; Barregard et al., 2019). Globally, the projection of the new fuels standard enforced in 2020 will reduce premature mortality from 430,000 to about 250,000 and will also reduce morbidity by 54% (Sofiev et al., 2018). These figures represent about 2.6% global reduction in PM_{2.5} cardiovascular and cancer deaths and about 3.6% global reduction in childhood asthma.

Emissions from ships could alter radiative balance (influencing climate) through the simultaneous releases of pollutants having positive (CO₂, BC) and negative (SO₂) radiative forcing (RF). NO_x emission has both positive RF due to O₃ formation and negative RF due to enhancement in methane (CH₄) loss. The net balance on short term is a negative RF, around -0.4 W/m² in 2005 (Eyring et al., 2009), however, on long term, the accumulation of CO₂ due to its slow removal is expected to turn the cooling effect to a warming effect (Fuglestedt et al., 2009). The effect of using low-sulfur fuel is somewhat complicated because there is a decrease on secondary sulfate in ship-related aerosols that generally have a cooling effect. Projections indicate that this reduction of sulfur will change optical properties of released aerosol reducing cooling effect of ship aerosols by about 80% corresponding to a 3% increase in

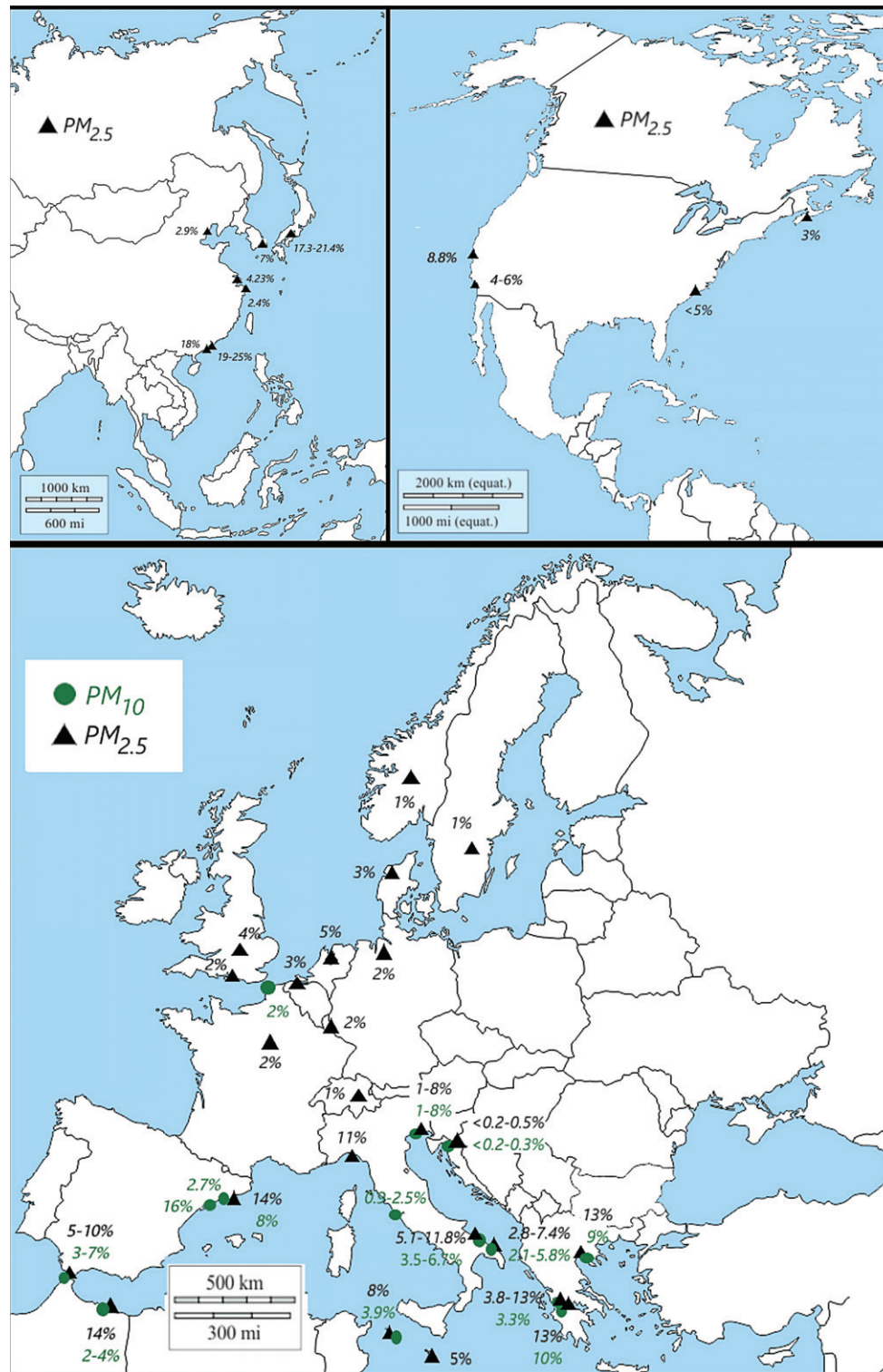


Figure 2 Relative (%) shipping contribution (excluded mixed industrial and shipping sources) to $PM_{2.5}$ (black triangle) and PM_{10} (green circle) worldwide. Source: Data is taken from the following references: Donateo et al. (2014), Lang et al. (2017), Nakatsubo et al. (2016), Saraga et al. (2019), Sarigiannis et al. (2017), Sorte et al. (2020), Viana et al. (2014).

current estimates of total anthropogenic radiative forcing (Sofiev et al., 2018). This suggests that future mitigation strategies and policies should target both aspects; health and climate, as two faces of the same medal. This could be achieved, for example, with actions targeting a joint reduction of air pollutants and GHGs.

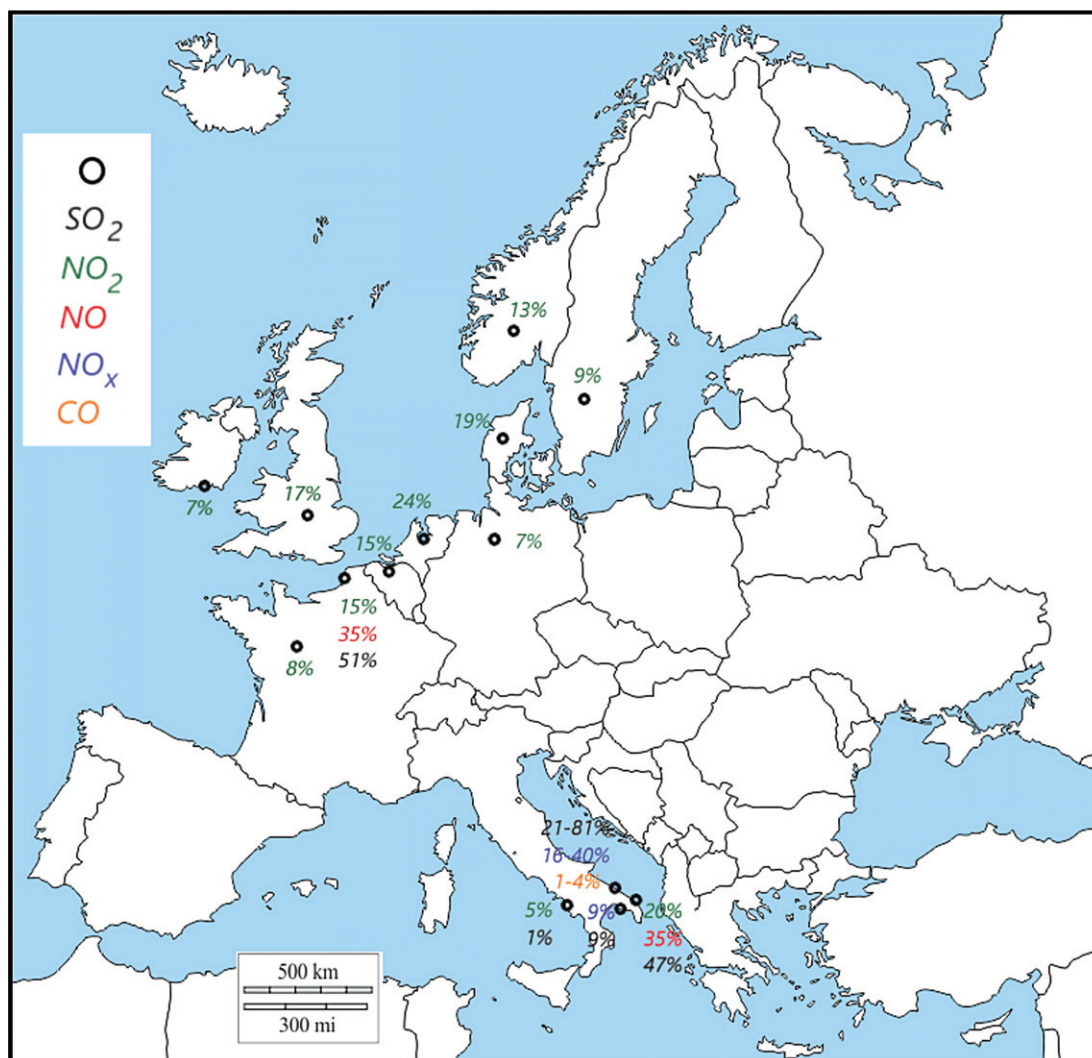


Figure 3 Relative (%) shipping contribution to gaseous pollutants across Europe. Source: Data is taken from the following references: Gariazzo et al. (2007), Ledoux et al. (2018), Merico et al. (2016; 2019), Murena et al. (2018), Viana et al. (2014).

Future Perspectives and Conclusions

A non-negligible share of global anthropogenic emissions of atmospheric pollutants (both in gas and particulate phases) and of GHGs is associated with shipping. At global level container ships, cargo carriers, and tankers are the largest contributors (more than 80% of total shipping emissions for most of the pollutants) but at local scale, in coastal areas and near harbor-cities, this generalization is not possible and cruise/passenger or ferries could have a prominent role.

Ship traffic contributions to atmospheric particulate matter (PM_{2.5}) worldwide ranges typically between 0.2% and 14% in Europe, with largest impacts in the Mediterranean area; between 3% and 9% in North America; and between 2.4% and 25% in Asia, especially in busy harbors in China. The general trend is that contribution of shipping to ultrafine (i.e., diameter < 0.3 μm) particle number concentration (PNC) is 3–4 times larger than the contribution to mass concentrations (either PM_{2.5} or PM₁₀). This would suggest that PNC could be a more suitable metric for investigating the role (and future trends) of shipping to atmospheric particulate matter. Contribution to gaseous pollutants like NO_x, SO₂, and CO are generally larger than those to particulate matter (PM_{2.5} or PM₁₀) and, considering current legislation thresholds in different continents, shipping could play a role in local exceedances of NO₂/NO_x thresholds.

The use of low-sulfur fuels as well as the establishment of ECAs (or DECAs) in specific areas brought to a reduction of SO_x, PM primary emissions, secondary sulfate in particles, and BC leading to a reduction of the impacts to measured concentrations especially near coastal and busy harbor areas. However, it had low or negligible effect on reduction of NO_x, total metals in PM (even if a certain reduction on specific metals like V was observed), and PAHs. Further, there are evidences that the use of low-sulfur fuels increases emission factors of VOCs with potential influence on secondary O₃ formation (that has negative effects on health). This suggests that future policies should take into account also non-criteria pollutants to maximize the positive effect on air quality. Epidemiological

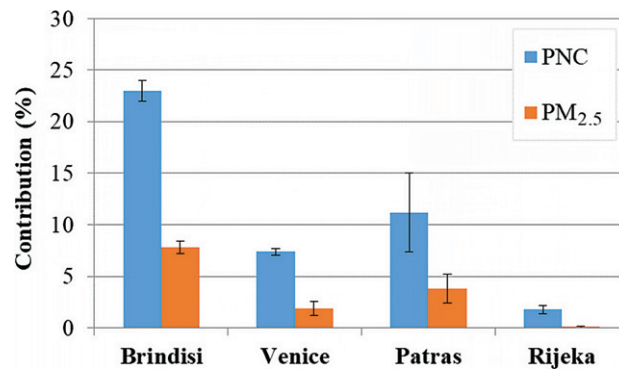


Figure 4 Relative (%) impacts to PM_{2.5} and PNC in four Mediterranean harbor cities.

studies showed that reduction of sulfur content of fuels brought a decrease of premature mortality and morbidity due to ship emissions globally and that ECAs/DECAs had similar beneficial effects at local scales in the interested areas.

The application of the new legislation (started in 2020) will bring a further reduction of shipping emissions. Projections indicate that this will lead to a reduction of 34% of ship-related premature mortality and 54% for morbidity. Despite this reduction there will still be about 250,000 deaths annually so that there will still be margins for further improvement. Specifically, further improvements on fuels are possible as well as establishment of additional ECAs, however, it has to be done with attention to the economic aspects in order not to affect the competitiveness of specific harbors or routes and with attention to climate. Low sulfur fuel reduces radiative cooling of ship-related aerosols and projection indicates a 3% increase of the total anthropogenic radiative forcing. Therefore, future stronger policies for shipping emissions should be focused on achieving both climate and health targets as could be done by jointly reducing atmospheric pollutants and GHGs. This becomes particularly important in climate hot spots like, for example, in exploiting new Arctic ship routes.

Future joint benefits for GHGs and air pollutants could come from the improvement of energy efficiency operational indicator (EEOI) of ships that will reduce directly fuel consumption. Further benefits, at local scale, could be achieved by improving sustainable and energy efficiency strategies and policies in harbors moving toward a “green” development of seaports. For example, harbor logistics could reduce emissions of harbor-related activities at loading/unloading of ships. This is a minor component of global emissions but could be relevant on local scale. Furthermore, cold ironing to partially/totally provide electricity to ships from the docks could significantly reduce emissions especially if electrical power is at least partially obtained from renewable sources.

Acknowledgments

The authors wish to acknowledge the contribution of the POSEIDON project “Pollution monitoring of ship emissions: an integrated approach for harbors of the Adriatic basin” (www.medmaritimeprojects.eu/section/poseidon, with the financial support of the European Territorial Cooperation, MED 2007-2013); of the CESAPO project “Contribution of Emission Sources on the Air quality of the Port-cities in Greece and Italy” (www.cesapo.upatras.gr, Interreg Greece-Italy 2007-13); and of the ECOMOBILITY project “ECOLOGical supporting for traffic Management in cOastal areas By using an IntelLIgent sYstem” (<https://www.italy-croatia.eu/web/ecomobility>, Interreg Italy-Croatia 2014-20).

See Also

Shipping and the Environment; Energy Efficiency of Ships; Arctic Shipping; Ro-Ro shipping

References

- Aksoyoglu, S., Prévôt, A.S.H., Baltensperger, U., 2016. Contribution of ship emissions to the concentration and deposition of air pollutants in Europe. *Atmos. Chem. Phys.* 156, 1895–1906.
- Barregard, L., Molnár, P., Jonson, J.E., Stockfelt, L., 2019. Impact on population health of Baltic shipping emissions. *Int. J. Environ. Res. Public Health* 16, 1954.
- Broome, R.A., Cope, M.E., Goldsworthy, B., Goldsworthy, L., Emmerson, K., et al., 2016. The mortality effect of ship-related fine particulate matter in the Sydney greater metropolitan region of NSW, Australia. *Environ. Int.* 87, 85–93.
- Cesari, D., Genga, A., Ielpo, P., Siciliano, M., Mascolo, G., et al., 2014. Source apportionment of PM_{2.5} in the harbour - industrial area of Brindisi (Italy): identification and estimation of the contribution of in-port ship emissions. *Sci. Total Environ.* 392–400, 497–498.
- Contini, D., Gambaro, A., Belosi, F., De Pieri, S., Cairns, W.R.L., et al., 2011. The direct influence of ship traffic on atmospheric PM 2.5, PM10 and PAH in Venice. *J. Environ. Manag.* 92, 2119–2129.
- Contini, D., Gambaro, A., Donato, A., Cescon, P., Cesari, D., et al., 2015. Inter annual trend of the primary contribution of ship emissions to PM 25 concentrations in Venice (Italy): efficiency of emissions mitigation strategies. *Atmos. Environ.* 102, 183–190.

- Corbett, J.J., Winebrake, J.J., Green, E.H., Kasibhatla, P., Eyring, V., et al., 2007. Mortality from ship emissions: a global assessment. *Environ. Sci. Technol.* 41 (24), 8512–8518.
- Dalsøren, S.B., Eide, M.S., Endresen, Ø., Mjelde, A., Gravir, G., et al., 2009. Update on emissions and environmental impacts from the international fleet of ships: the contribution from major ship types and ports. *Atmos. Chem. Phys.* 9, 2171–2194.
- Diesch, J.M., Drewnick, F., Klimach, T., Borrmann, S., 2013. Investigation of gaseous and particulate emissions from various marine vessel types measured on the banks of the Elbe in Northern Germany. *Atmos. Chem. Phys.* 13, 3603–3618.
- Donato, A., Gregoris, E., Gambaro, A., Merico, E., Giua, R., et al., 2014. Contribution of harbour activities and ship traffic to PM_{2.5}, particle number concentrations and PAHs in a port city of the Mediterranean Sea (Italy). *Environ. Sci. Pollut. Res.* 21, 9415–9429.
- EGCSA, 2018. <https://www.egcsa.com/>.
- EUROSTAT, 2019. <https://ec.europa.eu/eurostat>.
- Eyring, V., Isaksen, I.S., Bernsten, T., Collins, W.J., Corbett, J.J., et al., 2010. Transport impacts on atmosphere and climate: shipping. *Atmos. Environ.* 44, 4735–4771.
- Eyring, V., Isaksen, I.S.A., Bernsten, T., Collins, W.J., Corbett, J.J., et al., 2009. Assessment of transport impacts on climate and ozone: shipping. *Atmos. Environ.* 45, 4735–4771.
- Eyring, V., Kohler, H.W., Van Aardenne, J., Lauer, A., 2005. Emissions from international shipping: the last 50 years. *J. Geophys. Res. Atmos.* 110, D17305.
- Fuglestad, J., Berntsen, T., Eyring, V., Isaksen, I., Lee, D.S., et al., 2009. Shipping emissions: from cooling to warming of climates – and reducing impacts on health. *Environ. Sci. Technol.* 43 (24), 9057–9062.
- Gariazzo, C., Papaleo, V., Pelliccioni, A., Calori, G., Radice, P., et al., 2007. Application of a Lagrangian particle model to assess the impact of harbour, industrial and urban activities on air quality in the Taranto area Italy. *Atmos. Environ.* 41, 6432–6444.
- Gobbi, G.P., Di Liberto, L., Barnaba, F., 2019. Impact of port emissions on EU-regulated and non-regulated air quality indicators: The case of Civitavecchia (Italy). *Sci. Tot. Environ.* 719, 134984.
- Gregoris, E., Barbaro, E., Morabito, E., Toscano, G., Donato, A., et al., 2016. Impact of maritime traffic on polycyclic aromatic hydrocarbons, metals and particulate matter in Venice air. *Environ. Sci. Pollut. Res.* 23, 6951–6959.
- IMO, International Maritime Organization, 2015. Third IMO GHG Study 2014; London, UK, April 2015. Available from: <http://www.imo.org/>.
- Johansson, L., Jalkanen, J.-P., Kukkonen, J., 2017. Global assessment of shipping emissions in 2015 on a high spatial and temporal resolution. *Atmos. Environ.* 167, 403–415.
- Karl, M., Jonson, J.E., Uppstu, A., Aulinger, A., Prank, M., et al., 2019. Effects of ship emissions on air quality in the Baltic Sea region simulated with three different chemistry transport models. *Atmos. Chem. Phys.* 19, 7019–7053.
- Lack, D.A., Corbett, J.J., 2012. Black carbon from ships: a review of the effects of ship speed, fuel quality and exhaust gas scrubbing. *Atmos. Chem. Phys.* 12 (9), 3985–4000.
- Lang, J., Zhou, Y., Chen, D., Xing, X., Wei, L., et al., 2017. Investigating the contribution of shipping emissions to atmospheric PM_{2.5} using a combined source apportionment approach. *Environ. Poll.* 229, 557–566.
- Ledoux, F., Roche, C., Cazier, F., Beaugard, C., Courcot, D., 2018. Influence of ship emissions on NO_x, SO₂O₃ and PM concentrations in a North-Sea harbour in France. *J. Environ. Sci.* 71, 56–66.
- Liu, H., Fu, M., Jin, X., Shang, Y., Shindell, D., et al., 2016. Health and climate impacts of ocean-going vessels in East Asia. *Nat. Clim. Change* 6, 1037–1041.
- Liu, H., Jin, X., Wu, L., Wang, X., Fu, M., et al., 2018. The impact of marine shipping and its DECA control on air quality in the Pearl River Delta, China. *Sci. Total Environ.* 625, 1476–1485.
- Merico, E., Dinio, A., Contini, D., 2019. Development of an integrated modelling-measurement system for near-real-time estimates of harbour activity impact to atmospheric pollution in coastal cities. *Transp. Res. Part D* 73, 108–119.
- Merico, E., Donato, A., Gambaro, A., Cesari, D., Gregoris, E., et al., 2016. Influence of in-port ships emissions to gaseous atmospheric pollutants and to particulate matter of different sizes in a Mediterranean harbour in Italy. *Atmos. Environ.* 139, 1–10.
- Merico, E., Gambaro, A., Argiriou, A., Alebic-Juretic, A., Barbaro, E., et al., 2017. Atmospheric impact of ship traffic in four Adriatic-Ionian port-cities: comparison and harmonization of different approaches. *Transp. Res. Part D* 50, 431–445.
- Monteiro, A., Russo, M., Gama, C., Borrego, C., 2018. How important are maritime emissions for the air quality: at European and national scale. *Environ. Pollut.* 242, 565–575.
- Murena, F., Mocerino, L., Quaranta, F., Toscano, D., 2018. Impact on air quality of cruise ship emissions in Naples, Italy. *Atmos. Environ.* 187, 70–83.
- Nakatsubo, R., Oshita, Y., Aikawa, M., Takimoto, M., Kubo, T., Matsumura, C., Takaishi, Y., Hiraki, T., 2015. Influence of marine vessel emissions on the atmospheric PM_{2.5} in Japan's around the congested sea areas. *Sci. Total Environ.* 520, 134744.
- Oeder, S., Kanashova, T., Sippula, O., Sapcaru, S.C., Streibel, T., et al., 2015. Particulate matter from both heavy fuel oil and diesel fuel shipping emissions show strong biological effects on human lung cells at realistic and comparable in vitro exposure conditions. *PLoS One* 10 (6), e0126536.
- Russo, M.A., Leitao, J., Gama, C., Ferreira, J., Monteiro, A., 2018. Shipping emissions over Europe: a state-of-the-art and comparative analysis. *Atmos. Environ.* 177, 187–194.
- Saraga, D.E., Tolis, E.I., Maggos, T., Vasilakos, C., Bartzis, J.G., 2019. PM_{2.5} source apportionment for the port city of Thessaloniki, Greece. *Sci. Total Environ.* 650, 2337–2354.
- Sarigiannis, D.A., Handakas, E.J., Kerimenidou, M., Zarkadas, I., Gotti, A., et al., 2017. Monitoring of air pollution levels related to Charilaos Trikoupi Bridge. *Sci. Total Environ.* 609, 1451–1463.
- Sofiev, M., Winebrake, J.J., Johansson, L., Carr, E.W., Prank, M., et al., 2018. Cleaner fuels for ships provide public health benefits with climate tradeoffs. *Nat. Commun.* 9, 406.
- Sorte, S., Rodrigues, V., Borrego, C., Monteiro, A., 2020. Impact of harbour activities on local air quality: a review. *Environ. Pollut.* 257, 113542.
- Tao, L., Fairley, D., Kleeman, M.J., Harley, R.A., 2013. Effects of switching to lower sulphur marine fuel oil on air quality in the San Francisco Bay area. *Environ. Sci. Technol.* 47, 10171–10178.
- UNCTAD, 2016. Review of Maritime Transport, United Nation Publication.
- UNCTAD, 2019. Review of Maritime Transport, United Nation Publication.
- Viana, M., Amato, F., Alastuey, A., Querol, X., Saúl, G., et al., 2009. Chemical tracers of particulate emissions from commercial shipping. *Environ. Sci. Technol.* 43, 7472–7477.
- Viana, M., Hammingh, P., Colette, A., Querol, X., Degraeuwe, B., et al., 2014. Impact of maritime transport emissions on coastal air quality in Europe. *Atmos. Environ.* 90, 96–105.
- Winebrake, J.J., Corbett, J.J., Green, E.H., Lauer, A., Eyring, V., 2009. Mitigating the health impacts of pollution from oceangoing shipping: an assessment of low-sulfur fuel mandates. *Environ. Sci. Technol.* 43 (13), 4776–4782.
- Wu, Z., Zhang, Y., He, J., Chen, H., Huang, X., et al., 2020. Dramatic increase in reactive volatile organic compound (VOC) emissions from ships at berth after implementing the fuel switch policy in the Pearl River Delta Emission Control Area. *Atmos. Chem. Phys.* 20, 1887–1900.

Further Reading

- Abbasov, F., Earl, T., Jeanne, N., Hemmings, B., Gilliam, L., et al., 2019. One Corporation to Pollute Them All. Luxury cruise air emissions in Europe. In house analysis by Transport & Environment, June 2019. Available from: <https://www.transportenvironment.org/>.
- Ausmeel, S., Eriksson, A., Ahlberg, E., Kristensson, A., 2019. Methods for identifying aged ship plumes and estimating contribution to aerosol exposure downwind of shipping lanes. *Atmos. Meas. Tech.* 12, 4479–4493.
- Muntean, M., Van Dingenen, R., Monforti-Ferrario, F. et al., 2019. Identifying key priorities in support to the EU Macro-regional Strategies implementation – An ex-ante assessment for the Adriatic-Ionian and Alpine regions focusing on clean growth in transport and bioenergy. European Commission, Ipsra, JRC110395.
- Nakatsubo, R., Oshita, Y., Aikawa, M., Takimoto, M., Kubo, T., et al., 2020. Influence of marine vessel emissions on the atmospheric PM_{2.5} in Japan's around the congested sea areas. *Sci. Tot. Environ.* 702, 134744.

- POSEIDON, 2015. Inter-comparison of the impacts of maritime activities on air pollution in four port-cities of the Adriatic/Ionian Macroregion: trends, policy gaps and possible future strategies. project Final report. Available from: <http://www.medmaritimeprojects.eu/section/poseidon-redirect/outputs>.
- Rouil, L., Ratsivalaka, C., André, J.M., Allemand, N., 2019. ECAMED: a Technical Feasibility Study for the Implementation of an Emission Control Area (ECA) in the Mediterranean Sea. Synthesis report – January 11th, 2019. Available from: <https://www.ecologique-solidaire.gouv.fr/>.
- Schinas, O., Bani, J., 2012. The impact of a possible extension at EU level of SECAs to the entire European coastline. Note upon request of the European Parliament's Committee on Transport and Tourism. May 2012. Available from: <http://www.europarl.europa.eu/studies>.
- Viana, M., Fann, N., Tobias, A., Querol, X., Rojas-Rueda, D., et al., 2015. Environmental and health benefits from designating the Marmara Sea and the Turkish straits as an emission control area (ECA). *Environ. Sci. Technol.* 49 (6), 3304–3313.
- Wang, X., Shen, Y., Lin, Y., Pan, J., Zhang, Y., et al., 2019. Atmospheric pollution from ships and its impact on local air quality at a port site in Shanghai. *Atmos. Chem. Phys.* 19, 6315–6330.

Noise Pollution From Transport

Marianna Jacyna*, **Emilian Szczepański***, **Konrad Lewczuk***, **Mariusz Izdebski***, **Ilona Jacyna-Golda†**, **Michał Kłodawski***, **Paweł Gołda***, **Piotr Gołębiowski***, *Warsaw University of Technology, Faculty of Transport, Koszykowa 75, Warsaw, Poland; †Warsaw University of Technology, Faculty of Production Engineering, Narbutta 85, Warsaw, Poland; ‡Air Force Institute of Technology, Księcia Bolesława 6, Warsaw, Poland

© 2021 Elsevier Ltd. All rights reserved.

Noise Pollution Overview	277
Definition of Noise Pollution	277
Noise Indicators and Measures	277
Health Impacts From Noise Pollution	278
Road Traffic Noise	278
Sources of Traffic Noise	278
Road Traffic Noise Standards	279
Technical and Law Counteractions to Traffic Noise	279
Railway Noise	279
Sources of Railway Noise	279
Railway Noise Standards	280
Technical and Law Counteractions to Railway Noise	280
Aircraft Noise	280
Sources of Aircraft Noise	280
Aircraft Noise Standards	280
Technical and Law Counteractions to Aircraft Noise	281
Urban Noise Pollution	281
Sources of Urban Noise Pollution: Car Traffic, Railway Transport, Heavy Goods Vehicles	281
Noise Standards in Cities	282
Technical and Law Counteractions to Noise Pollution in Cities	282
Methods and Tools for Measurement and Modeling Noise Pollution	283
Noise Pollution Modeling	283
Road, Railway, and Aircraft Noise Calculation Methods	283
References	284
Further Reading	284

Noise Pollution Overview

Definition of Noise Pollution

The growth in travel demand over the last decades has led to a range of significant transport-related policy problems. Environmental externalities produced by transportation systems prevail. Noise pollution is one of the most pressing environmental problems associated with transportation and it has become an important element of the sustainable development. Noise is a growing concern among both the general public and policy-makers.

Noise pollution (sound pollution, environmental noise) is the propagation of anthropogenic noise with harmful impact on human health and well-being or animal life. It refers to the elevation of natural ambient noise levels due to sound-generating human activities, which may be wanted (music, sirens, sonars) or unwanted (i.e., transport-related noise).

Noise from transportation is the world's most prevalent form of environmental noise. The major source of noise pollution both inside and outside urban areas is road traffic. Noise from railways and aircrafts has a much lower impact in terms of overall population noise exposure, but both remain significant sources of localized noise pollution. European Environment Agency (EEA, 2019) has estimated that in Europe 78.2 million people are exposed to noise in excess of 55 dB L_{den} (55.1 million to noise in excess 50 dB L_{night}) due to road traffic, while the number exposed to the same level from railway is 10.3 million, from aircraft—3.0 million and that from industry—0.8 million.

Directive 2002/49/EC relating to the assessment and management of environmental noise (the Environmental Noise Directive—END) is the main EU instrument to identify noise pollution levels and to trigger the necessary action both at Member State and at EU level.

Noise Indicators and Measures

Environmental noise monitoring is the measurement of noise in an outdoor environment. The most common approach to sound intensity measurement is to use the decibel scale $I(\text{dB}) = 10\log_{10}(I/I_0)$. Decibels measure the ratio of a given intensity I to the threshold of hearing intensity I_0 , so that this threshold takes the value 0 decibels (0 dB).

Sound intensity is a sound power per unit area measured in Watts/m². Standard threshold of hearing intensity $I_0 = 10^{-12}$ Watts/m². The nominal dynamic range of human hearing is from the standard threshold of hearing I_0 to the threshold of pain $I_p = 10^{13} I_0 \approx 130$ dB.

Intensity of selected anthropogenic sounds: quiet library—30 dB, quiet room—40 dB, normal conversation—60 dB (intrusive), moderate traffic—75 dB, heavy traffic—85 dB (annoying), subway, motorcycle, truck traffic—90 to 100 dB (load), garbage truck, pneumatic drill—100 dB (very loud), emergency response siren, jet takeoff—120 dB, jackhammer—130 dB, jet engine—140 dB (painfully loud).

The audible spectrum of humans is from about 20 Hz to 20 kHz, but the upper limit in average adults is often closer to 15–17 kHz. All mechanical waves in this range can be associated with sound pollution.

Authoritative agencies set health protective noise thresholds, the excess of which may cause discomfort or lead to injuries and illnesses:

- US Environmental Protection Agency: day-night average sound level, indoors: $L_{dn} = 45$ dbA, dbA is an A-weighted decibels.
- World Health Organization:
 - equivalent continuous sound level, outdoors: $L_{eq} (16h) = 55$ dbA.
 - average nighttime noise level, outdoor: $L_{night} = 40$ dbA.

World Health Organization (WHO, 2018) strongly recommends reduction of average noise levels from road traffic below 53 dB L_{den} (45 dB L_{night}), from railway traffic below 54 dB L_{den} (44 dB L_{night}) and from aircraft below 45 dB L_{den} (40 dB L_{night}). It also recommends that policy-makers implement suitable measures to reduce noise exposure from transport in the population exposed to levels above the guideline values for average and night noise exposure.

Health Impacts From Noise Pollution

Exposure to noise can lead to auditory and non-auditory effects on health. Noise leads to hearing loss and tinnitus. Noise is a non-specific stressor having an adverse effect on human health, especially following long-term exposure. These effects are the result of psychological and physiological distress, as well as a disturbance of the organism's homeostasis and increasing allostatic load (WHO, 2018). Noise may negatively interfere with a child's learning and behavior, can damage physiological health, cause hypertension, high stress levels, sleep disturbances, and other harmful and disturbing effects (WHO, 2018). High noise levels can contribute to cardiovascular effects in humans and an increased incidence of coronary artery disease (Münzel et al., 2018). In animals, noise can increase the risk of death by altering predator or prey detection and avoidance and, interfere with reproduction and navigation (Kaselo and Tyson, 2004).

Road Traffic Noise

Sources of Traffic Noise

Noise from road transport vehicles has been of interest to researchers for many years. A great deal of work in this respect has been devoted to the noise to which a road user is exposed. Intensive work to reduce noise inside means of transport has continued to this day since the 1970s. Far less attention has been paid so far to reducing noise that affects the environment in which road transport vehicles operate (Guarinoni et al., 2012). Due to the number of road transport means, it is this mode of transport that generates the most noise. Both vehicle groups and individual vehicles are the sources of noise. Groups of road vehicles are most often found in urbanized areas and on very busy roads. The noise from such vehicles has the nature of continuous buzz or noise and its intensity may vary depending on the time of day and peak hours. Individual vehicles generate short-term noise that can occur anywhere.

Road noise is a combination of noise from the vehicle's drive system (engine noise) and noise caused by the interaction between the vehicle's tires and the road pavement (rolling noise). The noise level generated by a vehicle is mainly dependent on the speed at which the vehicle is running and the engine speed. As the speed increases, the engine noise gives way to rolling noise.

The vehicle drive system noise mainly consists of engine noise. Today it is usually an internal combustion engine. In addition to the combustion of the air-fuel mixture itself, noise is generated by the exhaust system, the air intake, fans, and auxiliary equipment.

Rolling noise is influenced by, among other things (Bernhard and Sandberg, 2005): noise from the impact of the vehicle's tyres on the road pavement, aerodynamic noise generated by the air pressed by the tyre in between the tread of the tyre and the road pavement, noise from vibrations of the tyre tread and unevenness of the road pavement, noise from slipping of the tyre when it comes into contact with the road pavement, and airborne noise generated by the air surrounding the vehicle.

This noise is further enhanced by a phenomenon known as the "horn effect." The geometry at the tyre/road point of contact forms the shape of a horn, which causes the high radiation of the noise emitted at this point. The tyre width, tread pattern and vehicle load also influence the rolling noise level. Of course, this type of noise will also be greatly influenced by the type of pavement on which the vehicle is running.

An additional source of noise from road transport is the horn. In many countries abuse of the horn can be punishable by police. Studies on noise from road vehicles do not take the horn sound into account.

Road Traffic Noise Standards

The main piece of EU legislation that defines the permissible levels of noise affecting humans and their environment is Directive 2002/49/EC (the Environmental Noise Directive—END). This Directive relates to the assessment and management of environmental noise, while indicating the measures necessary to reduce noise.

Another piece of legislation regulating noise from road transport vehicles is the European Directive (70/157/EEC) adopted in 1970. The Directive refers to the permissible noise level and exhaust system of motor vehicles. Since its adoption in 1970, it has been amended several times in order to adapt it to technical progress and changing vehicle structures. Currently, the content of this directive is included in directive No. 2007/34/EC of 2007. The directive introduces limits on the noise generated by motor vehicles depending on the vehicle category. The noise limit for passenger cars with a maximum of 9 seats is 74 dB. The noise limit for vehicles with a DMC of up to 3.5 and the engine power of not less than 150 kW intended for the carriage of cargo is 80 dB.

Another piece of legislation to limit noise emissions from road transport vehicles is directive (Directive 2001/43) of 2001. This directive applies to tyres of motor vehicles and their trailers and to their fitting. This directive divides motor vehicle tyres into classes. For each class of tyres and their tread widths, permissible noise limits are set, which vary from 70 to 79 dB.

Technical and Law Counteractions to Traffic Noise

Noise from means of transport can be reduced directly at its source or indirectly in the environment of its users (*At-source versus end-of-pipe measures*). The most commonly used methods of noise reduction in the user environment include the use of noise barriers, soundproof windows and proper urban planning (Bernhard and Sandberg, 2005).

Measures to control noise at its source include: legal regulations, regulations concerning the type of road pavement, regulations concerning the noise from the vehicle's drive system, regulations concerning the tyre noise level, the way the driver drives the vehicle, traffic management especially in urban areas (e.g., restrictions on traffic for heavy goods vehicles, reduction of traffic volume in selected areas, speed reduction, introduction of electric vehicles) combined with the transfer of passenger flows from private cars to public transport.

The research indicates that it is the most effective to eliminate noise already at its source (EC, 2004). So far, by far the most effective and cost-efficient method of noise reduction at source is the legislation that defines acceptable noise levels already at the stage of production of road transport vehicles, their equipment or tyres. Legislation towards setting noise limits for means of transport allows for noise reduction at a relatively low cost, as state-of-the-art engines and tyres already perform much better than current limits. Tightening the limits in this direction results in only minor additional costs for road vehicle manufacturers.

Railway Noise

Sources of Railway Noise

Railway is another mode of transport that causes environmental pollution through noise emissions. The basic sources of noise generated by rail transport include: noise from rolling stock, noise from the operation of equipment auxiliary to the operation of rail transport, aerodynamic noise, noise generated by signal equipment, and others (Murphy and King, 2014; Thompson, 2009).

Noise from rolling stock can be divided into four groups (EC, 2014):

- stationary noise,
- starting noise,
- the running noise, and
- noise inside the driver's cab.

Stationary noise is generated by railway vehicles when they are not in motion but are prepared for the run. Excessive noise from railway vehicles can be generated by a number of devices built on the vehicle—such as converters, motors, inverters, compressors, vehicle door closing and opening sounds, warning sounds before doors close or sounds generated by a system of toilets made in the so-called closed circuit. This noise is a nuisance both to those outside and inside the vehicle. Nevertheless, it is measured outside the vehicle.

Starting noise is associated with the railway vehicle or vehicles being put into motion. In order to do this, the traction vehicle must be powered at an appropriate level, which causes the noise emitted to be much higher than when running in uniform motion, for example. This noise is a nuisance both to those outside and inside the vehicle. Nevertheless, it is measured outside the vehicle.

Running (rolling) noise is the noise generated by the railway vehicle when running in uniform motion at 80 km/h and possibly 250 km/h. This noise is a nuisance both to those outside and inside the vehicle. Nevertheless, it is measured outside the vehicle.

The last type of noise from rolling stock relates to the interior of the driver's cab. A human being—an operator—a train driver drives a traction vehicle all the time. Therefore, the noise he is exposed to must meet strict standards due to the need for the highest level of safety and comfort. This noise is only onerous for people inside the vehicle and measurements are taken there.

The running of the train involves the need for authorized staff to operate a number of safety devices and to perform other accompanying rail transport operations. It is necessary, among others, to change the position of railway switches and derails, to

display a permitting signal at a semaphore, etc. Operation of these devices involves the emission of sound at appropriate levels. Therefore, there may be a risk of noise occurrence. Therefore, it is necessary to properly protect the devices against noise generation.

Aerodynamic noise is associated with a railway vehicle overcoming an obstacle resulting from an air barrier. When components of a railway vehicle try to “break through” the air medium, sound is generated. The higher the speed, the higher the intensity of the sound and the possibility of noise occurrence. When designing a vehicle, designers must act in such a way that the shape allows the least possible negative impact on noise generation.

The last group of devices that cause negative noise generation are signal devices mounted on the railway vehicle, which allow the driver to draw the attention of other traffic participants. Their use is associated with the need to maintain an adequate degree of safety. Currently three types of devices are used: whistle, high-pitched siren, and low-pitched siren. Each device should be audible from a suitable distance, but it must also comply with noise standards.

Railway Noise Standards

The limit values for stationary noise are expressed in terms of sound pressure. Under appropriate conditions, the equivalent continuous sound level for the vehicle concerned, the equivalent continuous sound level for the vehicle concerned taking into account the air compressor, and the corrected and time-constant sound level taking into account the impulse noise emitted by the exhaust shall be measured. The limit values depending on the type of railway vehicle are set in the range of 64–85 dB (EC, 2014).

Starting noise is also measured as a maximum corrected and time-constant sound level under appropriate conditions. Limit values are dependent on the type of vehicle and its power or maximum speed and take values between 80 and 87 dB (EC, 2014).

As already mentioned earlier, the running noise is measured at 80 km/h and/or 250 km/h under appropriate measurement conditions. Limit values express an equivalent continuous sound level and depend on the type of railway vehicle. For 80 km/h, the values are taken from 79 to 85 dB and for 250 km/h from 95 to 99 dB (EC, 2014).

The noise inside the driver’s cab shall be measured as an equivalent sound level close to the driver’s ear. Its limit values are dependent on the sound speed and emission values and they take values between 78 and 95 dB (EC, 2014).

Technical and Law Counteractions to Railway Noise

For the Member States of the European Union, the relevant document is Commission Regulation (EU) No 1304/2014 of 26 November 2014 on technical specifications for interoperability relating to the subsystem “Rolling stock — noise”, amending Decision 2008/232/EC and repealing Decision 2011/229/EU Text with EEA relevance (EC, 2014). This document describes the conditions for implementing interoperability (uninterrupted train running), including from the noise point of view. It is essential for the introduction of a single railway system throughout the European Union.

Aircraft Noise

Sources of Aircraft Noise

Air transport is one of the most developing branches of transport. Air transport allows to reach one’s destination very quickly and is one of the safest means of transport. Negative factors characterizing this branch of transport include its high emissivity of harmful compounds to the atmosphere and its noisiness. Sources of noise in air transport can be analyzed in several dimensions (Murphy and King, 2014), that is:

- noise generated by engines (propeller or jet engines), while the noise generated by jet engines results from the high speed of throwing gases into the atmosphere;
- aerodynamic noise, associated with airflow disturbances through elements such as wings, flaps or landing gear, can vary in different phases of flight.
- noise generated by equipment such as on-board generators for starting engines or pressure generation systems, and
- noise generated by ground operations such as aircraft taxiing, aircraft preparation, and maintenance passenger handling.

There are many factors that influence the noise level for particular sources (e.g., runway surroundings, weather conditions). Different aircrafts can generate different noise levels at different flight phases and at different airports.

Another source of noise is helicopter traffic. In this case, the engine generates noise through a “blade-slap.” This is the impact of the blade on the air turbulence generated by the preceding blade. In this case too, there is aerodynamic noise and generated by the helicopter’s additional equipment (Wilson, 2006).

Aircraft Noise Standards

International legal regulations on aircraft noise are implemented by specialized organizations that develop and implement international regulations governing the safety of international air navigation and support the development of air transport to ensure safe

and orderly development, such as the International Civil Aviation Organization (ICAO), Airport Council International (ACI), Directives of the European Union and European Parliament, US Federal Aviation Authority, Joint Aviation Authorities.

Air transport noise legislation addresses, inter alia, issues related to the implementation of noise-reducing procedures or appropriate air traffic management and planning both in the airspace and on the tarmac. In order to manage noise, ICAO has clarified a number of operational measures implemented at the time of take-off or landing, which are divided into the following areas:

- Operational measures reducing the level of emitted noise, for example, modern aircraft designs that minimize aerodynamic noise as well as the noise generated by aircraft propulsion system.
- Operational measures maximizing the distance between the aircraft and the ground.
- Operational measures targeting noise to less populated areas.

There are no specific provisions on the level of noise emitted by means of air transport, as well as actions and measures to protect against this type of noise. Individual solutions should be developed by individual authorities. However, as indicated in the WHO (2018) report, it is strongly recommended to reduce aircraft noise levels below 45 dB L_{den} for average noise exposure and below 40 dB L_{night} for night exposure.

Technical and Law Counteractions to Aircraft Noise

Measures to reduce aircraft noise can be considered under the following aspects: the development of modern engine propulsion systems, aircraft design and components, air traffic management methods, and methods to mitigate airport environmental noise.

Significant progress in aircraft design has been made since the 1950s of the 20th century, which lead to a reduction in noise level. It is known that today's aircrafts are 20 dB quieter than those designed in the beginning, which means that individual aircrafts are around 75% quieter (Guarinoni et al., 2012). Engine noise is one of the main factors influencing overall air traffic noise level. The reduction in noise results from a combination of changes in engine cycle parameters and design features. A key noise reduction measure is the renewal of an individual airline's fleet and the introduction of advanced aircraft technologies and solutions (Alonso et al., 2017) or the minimization of noise through appropriate aircraft design (Smith, 2004). An example is the creation of a silent SAX-40 aircraft with a Blended Wing Body (BWB) design (Graham et al., 2014).

Air traffic management plays an important role in the noise reduction process (Clarke, 2003). By using state-of-the-art navigation systems such as Area Navigation (RNAV) with the Global Positioning System (GPS) or Instrument Landing System (ILS) it is possible to control aircraft more precisely on approach to landing and take-off and thus to avoid inhabited or noise-sensitive areas.

The main measure to minimize aircraft noise around airports is the implementation of certain noise management measures. These measures include the development of models for the assignment of aircrafts with different engine types to specific runways at landing or take-off, the development of new arrival and departure procedures, the imposition of penalties on airlines for exceeding the noise limits. Minimizing the departure and arrival routes in the airspace of a terminal is another action carried out to reduce air noise.

The classification of noise abatement measures is shown in Table 1. In general, measures can be classified as legal actions as defined in directives, regulations, organizational actions to minimize noise through appropriate air transport traffic management, for example, route planning, and technical actions taking into account the application of the latest engineering and equipment solutions.

Urban Noise Pollution

Sources of Urban Noise Pollution: Car Traffic, Railway Transport, Heavy Goods Vehicles

Noise in urban areas is emitted by cars, trucks, motorcycles, buses, trams, and rail vehicles. It usually comes from traction systems (engine, braking system, exhaust system), wheel contact with road or rail and aerodynamic noise and vibration. The signals of rescue vehicles (fire brigade, ambulance, police), the operation of sanitary vehicles (e.g., garbage trucks), car alarms and horns are an

Table 1 Ways to reduce noise level in air transport

Legal actions	Organizational actions	Technical actions
• Flight reduction at night time • Financial penalties • Tax relief • Publication of best practice	• Assignment of appropriate aircraft types to runways • Organization of new arrival and departure procedures • The optimized route network in the terminal area • Noise maps	• Introduction of advanced air traffic management technologies such as RNAV • Fleet modernization • Noise barriers (barrier walls, embankments) • Quieter aircraft structures • New engine structures

important element of the city's noise environment. The noise is amplified and directed by tall buildings and compact developments along the streets.

The following traffic parameters are decisive factors for the noise level in the city: the traffic volume, the share of heavy vehicles and motorcycles, the speed of the vehicle flow, the traffic flow (stops and starts), the driving style and the use of horns (popular in Asian countries). The technical condition of vehicles (mainly related to age) has a major impact (Jacyna et al., 2017).

The noise level is also determined by the inclination of the carriageway section, the height of the location above the carriageway and the distance of the receiver from the carriageway, the texture and material of the pavement (concrete, asphalt, cobblestone, sound-absorbing surfaces), the terrain, the development of the surroundings (grass, shrubs, trees).

Noise Standards in Cities

The permissible noise level in a city may be limited by the authorities. Restrictions affect traffic organization and infrastructure design. During the daytime, restrictions to about 50–65 dB, WHO (2018) strong recommendation is 53 dB and at night to about 40–55 dB (WHO strong recommendation is 45 dB) and depending on the type of developments and function of the area. The limit values are often exceeded in the vicinity of most streets in the absence of protective devices, especially at night.

Technical and Law Counteractions to Noise Pollution in Cities

Reducing the noise level to the limit values is difficult, and therefore efforts are made to reduce the noise nuisance. A 3–5 dB reduction in the noise level has a noticeable effect.

Methods to reduce noise at source (emission zone):

1. Related to the vehicle and the driver:
 - a. Vehicle design (suspension, flow coefficient), engine design, type of tyres.
 - b. Technical condition and control of illegal modifications to exhaust systems.
 - c. Driving style and culture.
2. Design of infrastructure:
 - a. Location of the road and its surroundings—moving the road away from protected areas, elevation of the road (excavation, tunnel, partial covering, etc.),
 - b. Use of partially noise-absorbing construction materials.
 - c. Use of so-called “silent pavements” (3–5 dB noise reduction).
 - d. Reducing road inclination.
 - e. Maintaining road pavements in good condition.
 - f. Appropriate shaping of the excavation slope using greenery.
 - g. Use of non-contact tracks and prefabricated elements containing damping elements (track linings, mats under tracks—noise reduction 6–14 dB).
 - h. Use of grass.
3. Traffic organization:
 - a. Road hierarchy: layout of primary (traffic) and secondary (access and local) streets.
 - b. Active traffic control (increasing traffic flow), minimizing the number of stops.
 - c. Exclusion of traffic of selected groups of vehicles (temporary, area, competence).
 - d. Speed limiting (marking, speed cameras—reduction by 2 dB/10 km/h). Maintaining the speed of 30–50 km/h (with the predominant share of vehicles up to 3.5 t) results in a minimum noise level emission.
 - e. Using roundabouts (from 2 to 5 dB compared to a junction).
 - f. Promoting public transport, bicycles, and pedestrian traffic.

Reduction of noise at the receiver (nuisance zone):

- Objects located on the way of the sound wave—noise barriers (up to several dB), earth embankments (up to 25 dB), green belts (approx. 0.5–5 dB/1 m width of a dense hedge), non-residential buildings (e.g., garages, commercial buildings).
- Location, shape and acoustic insulation of buildings, noise barriers on (and in) the façade.
- Change of function of a building exposed to noise.

In EU cities with more than 100,000 inhabitants, at noise-prone areas, the necessity of making acoustic maps was introduced (Directive 2003/49/EC of the European Parliament and of the Council of 25 June 2002) relating to the assessment and management of environmental noise.

Methods and Tools for Measurement and Modeling Noise Pollution

Noise Pollution Modeling

An important aspect of assessing the impact of transport on the environment is noise measurement and prediction. Planning of infrastructure, transport systems, and their assessment requires the use of prediction models that take into account many factors related to the traffic, the characteristics of infrastructure and the environment. These models, extended by estimating the noise propagation, allow the preparation of noise maps for a selected area (Jacyna and Merkisz, 2014; Jacyna-Golda et al., 2017).

Noise prediction and propagation models are developed on the basis of actual traffic measurements. They aim at forecasting the noise intensity from the point of view of the receiver. Measurements are made under laboratory conditions (from a single noise source) as well as under real conditions for specific environmental conditions and vehicle structure. Such tests allow to develop a function taking into account various factors and indicating the noise level in a given place. The documentation of individual models contains information about the method of real measurements, that is, the measuring device used—a microphone, its location in space, the measurement procedure—continuous measurements, at time intervals, sampling, measurement conditions—traffic intensity, speed, pavement, weather conditions, ambient noise. On this basis, it is possible to repeat measurements, verify the model, and/or develop modifications to the model for specific traffic conditions.

The main group of models is statistical models based on the measurement sample and determining the noise level function for a given indicator. This group is most often used in the planning of transport systems by governmental entities. The second group of models is heuristic models, most often based on the use of artificial neural networks to forecast noise levels. The research indicates that models taking into account ANN are more accurate than statistical ones (Nedic et al., 2014), but they are most often used in academic applications.

As indicated by the Garg and Maji (2014), noise models may take into account various aspects and factors influencing noise levels, the most popular ones include:

- Source characteristics influencing emissions at the point of origin, including speed, traffic intensity, rolling noise, engine noise, aerodynamic noise, continuous or impulse emission, traffic dynamics characteristics (acceleration, deceleration), direction of traffic, characteristics of the means of transport—structure, type of tyres, load capacity.
- Characteristics of the infrastructure and the environment, including terrain, buildings and development density, noise barriers, vegetation, type of pavement, weather conditions, number of road lanes, occurrence of junctions, tunnels, and bridges.

Depending on the nature of the considerations and the type of transport (road, rail, air, industrial), micro, mezo, and macro scale models can be distinguished.

Road, Railway, and Aircraft Noise Calculation Methods

Road traffic is the most common element of research on noise in transport, to some extent due to the popularity and prevalence of this mode of transport. Many prediction models have been developed over the years. The following models are often used in road traffic, for example:

- FHWA TNM—a model developed in the United States by the Federal Highway Administration (FHWA), which is essentially a software tool for assessing road noise emissions. It includes five categories of vehicles (light vehicles up to 4500 kg, medium vehicles between 4500 kg and 12,000 kg, heavy vehicles more than 12,000 kg, buses more than nine passengers, and motorcycles). It is used to assess the noise emission from highways or road network based on adjustment to a sound reference level. The model contains an extensive database of reference emission values from different vehicle types and for different traffic conditions or pavements.
- CNOSSOS-EU—a model developed for the needs of the European Commission. This model includes five categories of vehicles (light—up to 3500 kg; medium—more than 3500 kg with two-axes, heavy vehicles—with three and more axes, moped and motorcycles; and open category). The calculation includes the estimation of rolling noise and engine noise, taking into account a wide range of parameters including but not limited to: traffic flow volume, speed, type of pavement, weather conditions or driver behavior. The model is also used to predict rail, aircraft, and industrial noise.

The above models are only selected examples used in practice and for research. Often these models are developed and used for specific countries (e.g., USA, UK, Germany, Switzerland, France, Japan, Scandinavian countries), the additional models/tools frequently used include the following: CoRTN model, NMPB-Routes, ASJ RTN-Model, RLS90, Harmonoise, Son road, Nord, CadnaA, TRANEX, SoundPlan. Information on the characteristics, comparisons, and application of the above models can be found in the papers (Garg and Maji, 2014; Khan et al., 2018; Murphy and King, 2014).

In rail transport, similar to road transport, models are being developed to assess noise emissions and spread. Here too, the already mentioned CNOSSOS-EU model applies. It takes into account; inter alia, the characteristics of the railway vehicle, the presence of barriers and devices to limit the spread of noise as well as the characteristics of the track. The model takes into account up to six types of noise sources in the assessment of noise emissions: traction, aerodynamic, rolling, squeal, impact, and bridge (Murphy and King, 2014). Other examples of models include: Reken-en Meetvoorschrift Railverkeerslawaii (RMR), Calculation of Railway Noise

(CRN), Mithra-Fer, Sound from Trains Environmental Analysis Method (STEAM), U.S. Department of Transport—Federal Transit Authority Noise Model.

In air transport, sophisticated prediction models are also used, which are an important element and indicator of noise reduction in a given area. Aircraft noise level can be measured using various measuring methods and models, for example, Integrated noise model (INM), SoundPLAN, Environmental Noise Model (ENM), FAA—integrated noise model, ECAC-CEAC Doc 29, Information on noise models can be found in the paper, for example, (Filippone, 2014; Murphy and King, 2014). In the era of Artificial Intelligence development technology, fuzzy models are the advanced models for assessing noise in the airport environment (Diop et al., 2013).

References

- Alonso, G., Benito, A., Boto, L., 2017. The efficiency of noise mitigation measures at European airports. *Transp. Res. Procedia* 25, 103–135.
- Bernhard, R.J., Sandberg, U., 2005. Tyre-pavement noise: where does it come from? In: *TRNews, Transportation Noise: Measures and countermeasures*. Transportation Research Board, Washington, D.C..
- Clarke, J.P., 2003. The role of advanced air traffic management in reducing the impact of aircraft noise and enabling aviation growth. *J. Air Transp. Manage.* 9 (3), 161–165.
- Diop, A., Alberto, C., Consenza, N., Marcellin, R., Faye, M.R., Mora-Camino, F., 2013. Assessing noise impact around airports: a fuzzy modeling approach. *Sch. J. Eng. Technol.* 1 (4), 226–231.
- EC, 2004. *EffNoise: Service contract relating to the effectiveness of noise mitigation measures: final Report, Volume I*. LÄRMKONTOR GmbH (contractor), Brussels.
- EC, 2014. Commission Regulation (EU) No 1304/2014 of 26 November 2014 on technical specifications for interoperability relating to the subsystem “Rolling stock — noise”.
- EEA, 2019. Exposure of Europe's population to environmental noise. European Environment Agency, Copenhagen.
- Filippone, A., 2014. Aircraft noise prediction. *Progr. Aerosp. Sci.* 68, 27–63.
- Garg, N., Maji, S., 2014. A critical review of principal traffic noise models: Strategies and implications. *Environ. Impact Assess. Rev.* 46, 68–81.
- Graham, W.R., Hall, C.A., Morales, M.V., 2014. The potential of future aircraft technology for noise and pollutant emissions reduction. *Transp. Policy* 34, 36–51.
- Guarironi, M., Ganzleben, C., Murphy, E., Jurkiewicz, K., 2012. Towards a comprehensive noise strategy. Policy Department Economic and Scientific Policy, European Parliament, Brussels.
- Jacyna, M., Merksiz, J., 2014. Proecological approach to modelling traffic organization in national transport system. *Arch. Transp.* 30 (2), 31–41.
- Jacyna, M., Wasiak, M., Lewczuk, K., Karon, G., 2017. Noise and environmental pollution from transport: decisive problems in developing ecologically efficient transport systems. *J. Vibroeng.* 19 (7), 5639–5655.
- Jacyna-Gotda, I., Gołębowski, P., Izdebski, M., Kłodawski, M., Jachimowski, R., Szczepański, E., 2017. The evaluation of the sustainable transport system development with the scenario analyses procedure. *J. Vibroeng.* 19 (7), 5627–5638.
- Kaselo, P.A., Tyson, K.O., 2004. Synthesis of Noise Effects on Wildlife Populations. Office of Research and Technology Services, Federal Highway Administration. Report No. DTFH61-03-H-00123 (September 2004). Available from: www.fhwa.dot.gov.
- Khan, J., Ketzel, M., Kakosimos, K., Sørensen, M., Jensen, S.S., 2018. Road traffic air and noise pollution exposure assessment—A review of tools and techniques. *Sci. Total Environ.* 634, 661–676.
- Münzel, T., Schmidt, F.P., Steven, S., Herzog, J., Daiber, A., Sørensen, M., 2018. Environmental noise and the cardiovascular system. *J. Am. Coll. Cardiol.* 71 (6), 688–697.
- Murphy, E., King, E., 2014. *Environmental Noise Pollution: Noise Mapping, Public Health, and Policy*. Elsevier, Newnes.
- Nedic, V., Despotovic, D., Cvetanovic, S., Despotovic, M., Babic, S., 2014. Comparison of classical statistical methods and artificial neural network in traffic noise prediction. *Environ. Impact Assess. Rev.* 49, 24–30.
- Smith, M., 2004. *Aircraft Noise*. Cambridge University Press, Cambridge.
- Thompson, D., 2009. *Railway Noise and Vibration: Mechanisms, Modelling and Means of Control*. Elsevier, Oxford.
- WHO, 2018. *Environmental noise guidelines for the European Region*. World Health Organization, Regional Office for Europe.
- Wilson, C.E., 2006. *Noise Control Revised Edition*. Krieger Publishing Company, Florida, Malbra.

Further Reading

- Daley, B., 2016. *Air Transport and the Environment*. Routledge, London.
- De Neufville, R., Odoni, A., 2003. *Airport Systems: Planning, Design and Management*. McGraw-Hill, New York.
- Licitra, G. (Ed.), 2012. *Noise Mapping in the EU: Models and Procedures*. CRC Press, Taylor & Francis Group, Boca Raton.
- Nicolopoulou-Stamati, P., Hens, L., Howard, C.V. (Eds.), 2006. *Environmental Health Impacts of Transport and Mobility*. Vol. 21, Springer Science & Business Media, Dordrecht, The Netherlands.
- Nilsson, M., Bengtsson, J., Klæboe, R. (Eds.), 2014. *Environmental Methods for Transport Noise Reduction*. CRC Press, Boca Raton.
- Rajakumara, H.N., Gowda, R.M., 2008. Road traffic noise prediction models: a review. *Int. J. Sust. Dev. Plan.* 3 (3), 257–271.

Visual Impacts From Transport

Paulo Rui Ancaies, Centre for Transport Studies, Civil Environmental and Geomatic Engineering UCL (University College London), London, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

What are Visual Impacts of Transport and Who is Affected?	285
Visual Intrusion—Impact on Users of Non-Motorized Modes and Non-Users	285
Transport Infrastructure	285
Motorized Vehicles	287
Consequences of Visual Impacts	287
The Subjectivity of Visual Impacts	288
How to Assess Visual Impacts	288
How to Reduce or Prevent Visual Impacts	289
The View From the Road—Impact on Users of Motorized Modes of Transport	290
See Also	291
References	291
Further Reading	291

What are Visual Impacts of Transport and Who is Affected?

The visual impacts of transport are the negative or positive effects of the physical appearance of transport infrastructure and motorized vehicles. These effects decrease the quality of trips and places and possibly affect travel behavior and use of public places; the well-being and residential satisfaction of the affected individuals; and the economic vitality and social cohesion of the affected neighborhoods.

Transport has two different types of visual impacts, depending on who is affected. The first type is felt by users of non-motorized modes (e.g., pedestrians, cyclists) and individuals who are not using the transport infrastructure causing the effect, but who live, work, shop, or visit the area where the infrastructure is visible. This impact is usually negative and has been described in the literature as “visual blight,” “visual intrusion,” or “visual pollution.” The second type of impact is felt by users of motorized vehicles (e.g., car and bus drivers and passengers). This impact is a mix of negative and positive aspects. The two types of impacts differ in terms of causes, consequences, solutions, and assessment methods. This article is therefore split into two parts, one for each type of impact, with the part on the impacts on users of non-motorized modes and non-users being lengthier because of the potentially extensive negative consequences of these impacts.

Visual Intrusion—Impact on Users of Non-Motorized Modes and Non-Users

Transport Infrastructure

Elevated roads and railways are the transport infrastructures most commonly recognized as causing negative visual impacts (**Fig. 1**). These infrastructures have a large imprint on the landscape and intrude on the visual field of pedestrians and people using public places next or underneath them, or looking at them from buildings. Elevated infrastructures can be disliked purely on aesthetical grounds, due to their size or appearance, or regarded as intimidating due to their size and position at higher level. The effect is aggravated in areas with multiple infrastructures at various heights and in areas with high-rise buildings, where some residents have elevated infrastructure near their windows.

Large transport infrastructures cause negative visual impacts even when not elevated. Motorways, multi-lane roads, and railways can be perceived negatively because of their size, which is incompatible with the scale of pedestrians, or because they are perceived as alien to the surrounding environment. Ports and airports may also be disliked because they are large single-use areas that include many buildings and other structures.

The negative visual impact often stems from specific elements of the infrastructure, rather than the whole infrastructure. Roundabouts, large road junctions, rail depots, areas with multiple rail service tracks, and large car parks, are particularly intrusive, as they occupy a large area that is off-limits to pedestrians (**Fig. 2**).

Often, the most visually disturbing elements are auxiliary structures. For example, overhead signs and billboards on motorways and roads intrude on pedestrians in streets nearby or below and on people at home. Again, size is the determining factor, as these signs are meant to be seen by people in vehicles moving fast. There is also extensive evidence that pedestrians dislike clutter caused by traffic signs and street furniture on urban streets, because of a visual disturbance effect above that caused by the obstructions that such elements cause on movement (**Fig. 3**).



Figure 1 Example of visual impact caused by elevated road infrastructure (Shanghai, China).



Figure 2 Example of visual impact caused by non-elevated road infrastructure (Brussels, Belgium).



Figure 3 Example of visual impact caused by clutter (Skopje, North Macedonia).

Lights from roads, ports, and airports also contribute to light pollution. Lights on tall poles along roads intrude on people's homes, affecting sleep patterns. Lights from transport infrastructure and vehicles, and their reflection, also contribute to urban sly glow, and lights on roads crossing natural areas disrupt wildlife. However, lighting also has positive aspects: it facilitates movement at night and increases the visibility of pedestrians, reducing fear of crime.

Visual intrusion may also arise because of elements that solve other problems. For example, noise barriers are often placed in areas where transport infrastructure cross residential areas, so they affect the view from peoples' homes. However, they may also reduce negative visual impacts if they are perceived as less obtrusive than the motorways or railways they cover—this depends on individual tastes and on the characteristics of the barrier, including the size, shape, materials, textures, and colors.

Another example is elements of the infrastructure that reduce the barrier effect for pedestrians. Pedestrians using footbridges are above large and busy transport infrastructures (roads or railways) and thus have a wider, unobstructed, sight of them. As an elevated element, footbridges also intrude on the visual field of pedestrians walking below and of people at home. Footbridges may also be perceived as unattractive because of their design, or lack of repair or maintenance. This is also the case of underpasses, which tend to be perceived as unattractive, unpleasant, and unsafe, because of insufficient lighting, poor design (lacking visually attractive features), and poor maintenance (leading to the appearance of visually unattractive features such as litter and graffiti).

While transport infrastructure has a mostly negative visual impact on pedestrians, some transport infrastructures can have positive impacts. This is for example the case of some iconic bridges (e.g., Sydney Harbour Bridge, Golden Gate Bridge in San Francisco) or even some visually attractive footbridges.

Motorized Vehicles

Pedestrians and people at home may also dislike the sight of motorized vehicles using the infrastructure, especially when the traffic has a large proportion of Heavy Goods Vehicles. People may dislike the design of vehicles or feel intimidated by their size, quantity, speed, or by particular situations (e.g., vehicles moving at fast speed in elevated roads). Sociologists have also noted that some pedestrians dislike the sight of private cars because of the perceived negative power relationship linked to the unequal distribution of roadscape.

These problems are exacerbated by glare from vehicle headlights on pedestrians at nighttime and reflections from the vehicles' metallic surfaces during daytime. This causes discomfort and reduces the amenity value of walking. It also causes distractions and reduces visibility, limiting pedestrians' ability to move safely along or across busy roads.

Motorized vehicles have negative visual impacts even when they are not moving. People often dislike the sight of parked cars as they are perceived to encroach on pedestrian space and more generally, on residential neighborhoods (Fig. 4). This effect is aggravated when cars are parked on pedestrian pavements.

Again, in a few cases, looking at motorized vehicles can be a positive experience. Trainspotting is a popular hobby in many countries. Some airports also have observations decks used by many passengers and non-passengers. Seeing ferryboats or canal boats can also be relaxing. Trams and funiculars are also a tourist attraction in many cities, such as Lisbon (Portugal) and Valparaíso (Chile).

Consequences of Visual Impacts

The negative impacts of transport on pedestrians and users of the affected area can lead to wider negative consequences, described below. While most of these have been mentioned in the literature, it is often only as a part of a list of hypothesized effects. There is still little empirical evidence on the magnitude of many of these effects, and on who is affected and under which circumstances—in contrast with the rich body of evidence produced about other negative local effects of transport, such as noise and air pollution.

Visual impacts reduce the perceived quality and people's enjoyment of the places (e.g., streets, squares, parks) that are next, under, or within sight of unattractive transport infrastructure. Often, the reaction is to avoid those places or, if that is not possible, to spend less time in them. The low visual quality of public places affects what Jan Gehl calls "optional" and "social" activities, that is, activities that are not "necessary" (e.g., waiting for bus) and only occur in places where people feel relaxed and comfortable (Gehl, 2010).



Figure 4 Example of visual impact caused by parked cars (Lisbon, Portugal).

The sight of visually intrusive transport infrastructure and traffic also reduces the perceived quality of walking and cycling trips as it limits the ability to perceive the surrounding environment and the enjoyment derived from walking and cycling. This restricts travel behavior, as people seek to avoid routes that are visually unappealing and choose other routes, which may be longer, have other problems (e.g., slopes), or have fewer trip attractors (e.g., shops). Visual impacts also create a psychological barrier effect that dissuades people from walking to the other side, separating neighborhoods and impeding people's ability to reach people and places on the other side.

Exposure to the visual impacts of transport can even affect people's image of cities and towns. The importance of visual perceptions has been recognized in the work of Kevin Lynch and Gordon Cullen in the 1960s (Lynch, 1960; Cullen, 1961), who called for planners to improve the visual quality of cities and towns. This has become imperative in the 21st Century, at a time when there is increased competition between cities to capture residents, business, and visitors based on high-quality public places.

For the individuals affected, the continued exposure to intrusive, unpleasant, or intimidating transport infrastructure and traffic, and the associated constraints on travel behavior and the use of public places, may have a negative impact on health and well-being. There is evidence, for example, that the view of one's home can aggravate or prevent the recovery from stress, depression, and even physical health problems.

The impacts on the quality of local places and the view from home also reduce residential and neighborhood satisfaction and can be a determinant of the choice of residential location. People may choose to live in different parts of the same neighborhood, or in a different neighborhood altogether, to avoid those impacts. However, empirical evidence has been non-conclusive, as the residential location choice is not explained by visual impacts alone, but by the cumulative effect of all negative aspects of living close to transport infrastructure, as well as the positive aspects (i.e., increased accessibility).

As people avoid the areas near or under transport infrastructure, these areas also become less attractive for commercial and leisure land uses and often become vacant—what Jane Jacobs called “border vacuums” (Jacobs, 1961). These can amount to a considerable proportion of a city's area—for example, New York City has an estimated 700 miles of elevated infrastructure. Some spaces may even become surrounded by large transport infrastructure on all sides, including above, and even when accessible to pedestrians, they become “no go” areas due to their unattractiveness—a situation described in fiction in J.G. Ballard's book “Concrete Island” (Ballard, 1974).

Unattractive places and neighborhoods also fare bad economically. There is extensive evidence that the quality of views influences the value of homes, and probably also of rents of shops and offices. It may also reflect on expenditure on businesses, via the effect on the number of trips, the time spent in places, and the propensity to spend money in those places. The price of hotel rooms and tourist attractions is also highly sensitive to the quality of views.

There is also a social cost. Areas next to walls or noise barriers or under large infrastructure tend to be neglected and attract vandalism and non-social behavior. The outcomes of some of this behavior (e.g., litter, unattractive graffiti) exacerbate the visual impact. All this affects people's perceptions of personal security, crime, attitudes towards the neighborhood, and ultimately community cohesion. At the same time, reduced walking and use of public places reduces social interaction, which shrinks local social networks and neighborhood social capital, and contributes to social exclusion of some people—again leading to reduced community cohesion.

The Subjectivity of Visual Impacts

Visual impacts are specific to a particular site at a particular time, affecting a particular group of people. As such, it is difficult to generalize research on these impacts, in contrast with other local impacts of transport, such as noise, air pollution, and even the barrier effect on pedestrians, which can be measured more or less objectively.

The problem depends on the spatial context—but there is not enough empirical evidence to know exactly how. For example, it is likely that the visual impacts of transport infrastructure are especially harmful when they are felt in natural areas, as the infrastructure will be perceived as an alien, unexpected, element. However, it could also be argued that visual impacts are more harmful in urban areas, in places that are used daily by people, on their way to work or walking around their neighborhood. Visual impacts may be attenuated in areas with attractive visual elements. But they can also be regarded as more impactful in their areas, as they detract from people's enjoyment of the attractive elements.

It also depends on the time context. Over time, individuals may adjust to the presence of visually disruptive transport infrastructure or motorized traffic. Local residents who experienced the change in the visual environment caused by a newly built infrastructure may also have different perceptions of the problem than those who have recently moved to the affected area.

The extent to which visual impacts are considered a problem also depends on personal perceptions. Visual preferences vary with culture. For example, in some cultures, minimalism is perceived positively, in others is perceived as boring. Studies on general visual preferences usually find differences according to age and gender, but there is not enough evidence to support his hypothesis in the case of transport infrastructure and motorized vehicles. It is likely, however, that the effect is stronger for people with mobility impairments, as it limits their ability to leave home or their neighborhoods.

How to Assess Visual Impacts

The most common approach to measure the visual impacts of transport infrastructure is to map viewsheds, that is, the area that is visible from a given point. The identification of intrusive infrastructure in viewsheds from a given point signals to a possible problem of negative visual impacts. Advances in GIS (Geographic Information System) data and methods, especially 3D models, have allowed for a more precise estimation of viewsheds. However, even with these methods, the approach is not usually taken over to the

next step, i.e. to measure the degree to which the infrastructure impacts on the people walking or spending time in the point where the viewshed is calculated.

Street audit tools (i.e., systematic assessments of the pedestrian environment) are sometimes used to capture visual aspects. However, these aspects are only a component in a larger assessment framework. Disaggregated results are not always presented. In addition, the aspects assessed usually refer to the whole visual character of the surrounding built environment, not the visual intrusion of transport infrastructure. The assessment also relies on the judgments of experts (usually only one), who may not be aware of how visual aspects affect the people who are exposed to them on a daily basis.

To address these limitations, surveys can be used to capture the perceptions and priorities of pedestrians or local residents with regards to the presence/absence of the transport infrastructure or particular characteristics of the infrastructure or motorized traffic. These surveys typically ask participants to rate or choose among different options, visualized in images, videos, and, increasingly, in immersive virtual reality environments.

In practice, visual impact assessments usually involve the production of various maps and a descriptive assessment synthesizing those maps. Assessments are often done in the case of new infrastructure but seldom in the case of changes that increase the visual impacts of existing infrastructure or interventions that reduce those impacts. There is also more guidance on how to assess the impacts on areas with natural or cultural interest, compared with impacts on residential areas.

The output of the assessment may also not be fully integrated into project appraisal. In fact, most official documents on transport appraisal mention visual impacts merely as an item on a list of social or environmental effects, not providing recommendations on measurement methods. As a result, visual impacts are either absent or included in a descriptive manner in cost-benefit analyses. As noted before, visual impacts have an economic value. However, it is not easy to pinpoint this value as a component of the value of the whole sensory experience of transport infrastructure and vehicles, which also includes exposure to dust, noise, vibration, and air pollution. Only a few academic studies have used stated preference surveys or hedonic analyses to capture trade-offs between property costs and visual aspects.

How to Reduce or Prevent Visual Impacts

There is increased political acceptance of interventions to remove obtrusive infrastructure, reflecting the shift of urban transport policy away from a car-centric paradigm and toward a livability paradigm (Anciaes and Jones, 2020). Elevated roads have been demolished in several cities. An emblematic project was the demolition of the Cheonggye Expressway in Seoul (South Korea) in 2005 to restore a stream and create a new park. The reduction of visual aspects is usually only one of the intended aims of these projects, the others being restoring pedestrian links and reducing noise and air pollution. Another solution is to screen the infrastructure with trees, rather than walls, which, as mentioned before, may aggravate the problem).

There is also growing interest in reconverting spaces under elevated roads and railways as shopping or leisure areas. Large-scale studies by the Design Trust for Public Spaces in New York (DTPS, 2015) and ARUP in several cities around the world (ARUP, 2018) have documented successful examples and provided guidelines for further interventions. A crucial component of these interventions is to make the spaces more visually attractive, with better lighting, greening, and public art. Unused elevated infrastructure can also be transformed into linear parks, for example the Promenade Plantée in Paris and the High Line in New York (Fig. 5).

Visual impacts can also be reduced by reducing traffic levels, especially of large vehicles. This can be achieved by restricting access of motorized vehicles to areas with high pedestrian traffic, at certain times.

There is also an increasing interest in the improvement of the public realm in roads, streets, and public places. Examples include reallocating road space to pedestrians and place activities (by restricting car movement or parking), the use of more visually appealing materials for street furniture and pedestrian pavements, and the removal of large and/or unnecessary street furniture.



Figure 5 Example of reconversion of elevated transport infrastructure (New York, United States).

Improving landscaping or adding trees also reduces the visual impact of roads. Issues such as light pollution tend to attract less attention, although there are regulations on the position and intensity of lights.

The View From the Road—Impact on Users of Motorized Modes of Transport

Visual aspects are also relevant for users of motorized modes (e.g., car, bus, and train drivers and passengers). Many of the aspects mentioned in the first part of this article also apply in this case, but there are important differences.

The view of the surrounding area contributes to trip quality and to the amenity value that drivers and passengers derive from the trip. Views of green or blue areas tend to be perceived positively, and views of industrial or derelict residential areas as negative. The research of Donald Appleyard and colleagues in motorways approaching large cities in the United States was one of the first to highlight the importance of the view from the road (Appleyard et al., 1964) and its implications for road design. A common result of this and other research is that, since cars move at a relatively fast speed, the visual experience of car users depends on the size of the visual field.

Car users are also affected by the view of the infrastructure they are using. In fact, people who use the same road frequently (for example, commuters) may not pay as much attention to the surrounding landscape (the way a visitor would do) as to details of the road and traffic, including walls and other elements at the road shoulder; the appearance of the road surface; signs; and the characteristics and motion of other vehicles.

Again, elevated roads and railways tend to have a negative visual impact, as they shrink the visual field of car users moving below. Complex junctions are also perceived as stressful to car drivers. The experience of driving in some junctions can even leave an imprint on the collective perception of wide regions around them. An example is the Gravelly Hill Interchange near Birmingham (England), popularly known since the 1960s as “Spaghetti Junction”—a nickname subsequently given to other similarly complex junctions around the world.

Despite being an auxiliary element, road landscaping is a crucial aspect of the visual experience of car users. Trees or green areas on the sides of the road or in median strips increase the amenity value of the trip and reduce speed and driver stress and fatigue. In contrast, clutter is perceived negatively and distracts and disorients drivers. Clutter is caused by the presence of an excessive number of elements (e.g., traffic signs, billboards) with different sizes, shapes, and colors, and placed at different positions, some at eye level, and others overhead.

Road lighting has both negative and positive effects on drivers. Overly bright road lighting or glare from other vehicles’ headlights at night distract the driver and pose safety risks, but also increase visibility of the road and other vehicles, addressing not only traffic safety but also personal security.

Car users may respond to visual impacts by changing their routes, avoiding the less attractive ones. They may even change their travel destinations to avoid unattractive places or the routes to access them. These behaviors are common for leisure trips, but for commuting or shopping trips are often unfeasible, because alternative routes may be longer or more congested. There is little empirical evidence on how car users trade-off the visual quality of different routes against trip length, duration, and cost.

The view that car users have from the road also has consequences on health and well-being. Regular exposure to unattractive views contributes to commuter stress, lower satisfaction with one’s residential and work areas, and lower subjective well-being. In contrast, exposure to nature-dominated drives may contribute to stress recovery and immunization.

Most of the methods for visual impact assessment mentioned in the first part of this article also apply to the case of car users, with some differences. For example, viewsheds represent what can be seen from the infrastructure. Official guidance on how to use the assessment methods, and on the range of aspects to consider, tends to be more detailed than in the case of impacts on non-users.



Figure 6 Example of road landscaping (Lima, Peru).

Guidelines mention both objective aspects (e.g., visibility, quality, maintenance of materials) and subjective aspects (e.g., harmony, character).

Visual aspects can become an important factor determining road alignment and design. This is especially the case of roads in areas with natural or cultural interest. There are programs in several countries that frame the planning and management of these roads—for example the National Scenic Byways program in the United States and the National Scenic Routes program in Norway.

The visual experience of car users can also be improved with small, but effective measures, even in busy roads in urban areas. Common interventions include adding, improving, or simply maintaining landscaped areas, and using more a visually appealing road pavement (Fig. 6). In some countries, fountains or public art have also been placed on roundabouts.

There are also regulations to reduce clutter, both in large infrastructure and in urban roads and streets. In the United States, the Highway Beautification Act of 1965 was one of the first laws aimed at improving the appearance of roads and the views from it, by regulating the number and characteristics of signs and billboards, as well as some unsightly land uses that can be seen from the road. In São Paulo (Brazil), the Lei Cidade Limpa (“Clean City Law”) outlawed the use of outdoor advertisements. Again showing the subjectivity of visual impacts, both of these examples have attracted controversy not only from the businesses that had placed the billboards but also from some road users who liked to see them.

See Also

Access to Transport and Health; Wellbeing and Travel Satisfaction; Planning for Safe and Secure Transport Infrastructure; Road Modes: Walking

References

- Anciaes, P., Jones, P., 2020. Transport policy for liveability – valuing the impacts on movement, place, and society. *Transp. Res. A: Policy Practice* 132, 157–173.
- Appleyard, D., Lynch, K., Myer, J., 1964. *The View From the Road*. MIT Press, Cambridge, United States.
- ARUP, 2018. *Under the Viaduct. Neglected Spaces No Longer*. ARUP, London.
- Ballard, J.G., 1974. *Concrete Island*. Jonathan Cape, London.
- Cullen, G., 1961. *The Concise Townscape*. Architectural Press, London.
- DTPS (Design Trust for Public Spaces), 2015. *Under the Elevated: Reclaiming Space, Connecting Communities*. DTPD and NYC Department of Transportation, New York.
- Gehl, J., 2010. *Cities for People*. Island Press, Washington.
- Jacobs, J., 1961. The curse of border vacuums. In: *The Death and Life of Great American Cities*. Random House, New York, Chapter 14, pp. 336–352.
- Lynch, K., 1960. *The Image of the City*. MIT Press, Cambridge, United States.

Further Reading

- Anciaes, P.R., Metcalfe, P.J., Heywood, C., 2017. Social impacts of road traffic: perceptions and priorities of local residents. *Impact Assess. Project Appraisal* 35, 172–183.
- Blumentrath, C., Tveit, M.S., 2014. Visual characteristics of roads: a literature review of people's perception of Norwegian design practice. *Transp. Res. A: Policy Practice* 59, 58–71.
- Bocarejo, J.P., Le Compte, M.C., Zhou, J., 2012. *The Life and Death of Urban Highways*. Institute for Transportation and Development Policy and EMBARQ, New York and Washington.
- Eliasson, J., Dillén, J.L., Widell, J., 2002. Measuring intrusion valuations through stated preferences and hedonic prices—a comparative study. *AET Papers Repository*. Available from: <https://aetransport.org/past-etc-papers/conference-papers-pre-2009>.
- Foley, L., Prins, R., Crawford, F., Humphreys, D., Mitchell, R., Sahlqvist, S., Thomson, H., Ogilvie, D., 2017. Effects of living near an urban motorway on the wellbeing of local residents in deprived areas: natural experimental study. *PLoS One* 12 (4), e0174882.
- Garré, S., Meeus, S., Gulincx, H. The dual role of roads in the visual landscape: a case-study in the area around Mechelen (Belgium). *Landscape and Urban Planning* 92, 125–135.
- Highways Agency (UK), 2010. *Landscape and Visual Effects Assessment*. Note 135/10.
- Jiang, L., Kang, J., Schroth, O., 2015. Prediction of the visual impact of motorway using GIS. *Environ. Impact Assess. Rev.* 55, 59–73.
- Kang, C.D., Cervero, R., 2009. From elevated freeway to urban greenway: land value impacts of CGC project in Seoul. Korea. *Urban Studies* 46, 2771–2794.
- Kaplan, R., 2001. The nature of the view from home: psychological benefits. *Environ. Behav.* 33, 507–542.
- Maffei, L., Masullo, M., Aletta, F., Gabriele, M., 2013. The influence of visual characteristics of barriers on railway noise perception. *Sci. Total Environ.* 445/446, 41–47.
- Parsons, R., Tassinary, L.G., Ulrich, R.S., Hebl, M.R., Grossman-Alexander, M., 1998. The view from the road: implications for stress recovery and immunization. *J. Environ. Psychol.* 18, 113–140.
- Taylor, N., 2005. The aesthetic experience of traffic in the modern city. *Urban Studies* 40, 1609–1625.
- Yamada, H., Shinohara, O., Amano, K., Okada, K., 1987. Visual vulnerability of streetscapes to elevated structures. *Environ. Behav.* 18, 733–775.

Light Pollution

John D. Bullough, Lighting Research Center, Rensselaer Polytechnic Institute, Troy, NY, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	292
Sky Glow	292
Quantity of Light	292
Spectral Distribution	293
Spatial Distribution	293
Recommendations for Minimizing Sky Glow	294
Light Trespass	294
Quantity of Light	294
Spectral Distribution	294
Spatial Distribution	294
Recommendations for Minimizing Light Trespass	295
Glare	295
Quantity of Light	295
Spectral Distribution	295
Spatial Distribution	295
Recommendations for Minimizing Glare	296
Outdoor Site-Lighting Performance System	296
See Also	296
References	296

Introduction

Although it is generally realized that the use of roadway lighting can contribute to reduced crashes and an improved sense of safety and security at night, there is also an increasing awareness among public authorities that outdoor lighting along roadways and other facilities can have negative impacts collectively referred to as light pollution (Riegel, 1973). Although this chapter focuses on those impacts as they affect human activity and observation, the potential effects of lighting on wildlife including insects and nocturnal animals are also beginning to be understood (Longcore and Rich, 2004).

This chapter is organized according to the three main light pollution impacts that have been studied:

- Sky glow
- Light trespass
- Glare

Within each of the sections of this chapter that address these aspects of light pollution, the relative impacts of the quantity, the spectral distribution, and the spatial distribution are discussed, followed by several simple recommendations for the design and specification of roadway lighting systems that can help to mitigate light pollution. Finally, a simple photometric system for quantifying light pollution impacts and comparing alternative lighting systems is described.

Sky Glow

Sky glow refers to the brightening of the nighttime sky caused by electric light sources present in the outdoor environment. Before the widespread application of roadway and other forms of outdoor lighting, the Milky Way galaxy was visible in the night sky throughout most, if not all, of the world, at least, on clear nights. This section discusses lighting related factors that influence sky glow.

Quantity of Light

Probably the most important factor related to sky glow is the amount of light (e.g., illuminance on the ground or luminance of the paved roadway surface) used in the specific lighting application. It has been understood for several decades (Waldram, 1972) that light reflected from the illuminated ground surfaces provides the majority of the contribution to increased sky luminance. This can

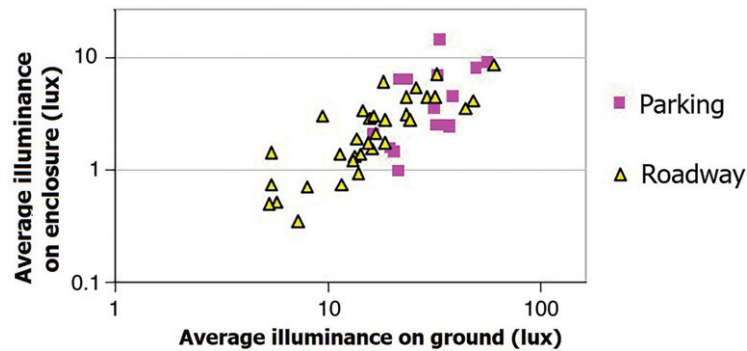


Figure 1 The average illuminance on an enclosure around a lighted road or parking lot is about one tenth of the average illuminance on the ground surface.

be critical because many recommended practices for roadway and outdoor illumination (CIE, 2010; IES, 2018) throughout the world are expressed in terms of the *minimum* recommended value for the average illuminance or luminance.

Only very recently have such guidelines begun to discuss the notion of not exceeding these minimum recommendations by a substantial amount. Indeed “ratcheting” light levels on roads in order to allay perceived concerns about security has resulted in a substantial number of overlighted roads and parking lots, with a direct impact on the amount of light reflected into the night sky. Figure 1 illustrates that the total amount of light coming from an outdoor lighting installation along a roadway or parking lot is proportional to the amount of light falling on the pavement. The average illuminance on an enclosure around a lighted road or parking lot (representing light that can contribute to sky glow) is about one tenth the average illuminance on the horizontal surface, which matches the average reflectance of asphalt pavement of about 10% (Brons et al., 2008).

Spectral Distribution

Recently there have been calls among the astronomical observation community to limit the short-wavelength (“blue”) output of outdoor lighting in order to help minimize sky glow impacts. The basis for these recommendations is Rayleigh scattering, which is proportional to the reciprocal of the fourth power of the wavelength of light. Rayleigh scattering occurs when the size of the scattering particles is similar to the wavelength of the light being scattered (for visible light, this is 400–700 nm). The majority of atmospheric light scatter that contributes to sky glow is caused by aerosols, which contain particles that are typically an order of magnitude larger than the wavelengths of light. Scattering from aerosols is a different type of scatter known as Mie scattering, which is much less sensitive to the wavelength of the scattering light. As a consequence, light from all visible wavelengths is scattered to a similar extent and the spectrum has a relatively small impact.

Bierman (2012) and Rea and Bierman (2015) calculated the impacts of changing the spectral distribution of outdoor lighting from a yellowish light source (high pressure sodium, HPS) with a correlated color temperature (CCT) of 2200 K to a cool-white light emitting diode (LED) light source with a CCT of 6500 K. The relative amount of scattered light depended in part on the aerosol content of the atmosphere (spectral differences were greater for dry, clear air relative to humid air) but were no larger than 10%–20%. Of interest, the perceived brightness from light with a higher CCT (Rea et al., 2011) could allow the light level from an LED source to be reduced relative to HPS, while still maintaining an equal or greater perception of brightness. Such a reduction would have a corresponding reduction in the amount of sky glow.

Near astronomical observatories, where the ability to detect faint celestial objects is critical, the use of low pressure sodium (LPS) sources for road lighting is often encouraged. LPS lamps emit nearly monochromatic light at a wavelength of 589 nm, appearing deep yellow in color. Filters are available that block this wavelength while allowing others to pass through, permitting astronomers a largely unimpeded view of the night sky.

Spatial Distribution

The spatial distribution can have obvious effects on sky glow; if a streetlight used for roadway lighting emits light in the upward direction toward the sky, most if not all of that light will contribute to sky glow. For this reason, roadway lighting recommendations (IES, 2011) have tended to promote the use of “fully shielded” fixtures that emit no upward light. Fortunately, even “cobrahead” style lights with a drop lens emit relatively little upward light even though a portion of the lens is visible from above the fixture. Typically, the amount of luminous flux from a roadway lighting fixture that goes upward is no more than about 0.5% of the total flux from the fixture. Relative to the amount of light reflected from illuminated surfaces, this direct contribution is small. The same cannot be said for “globe” style and some other decorative types of streetlighting fixtures, which emit a large amount of light in the upward direction (up to 50% for globe fixtures).

Recommendations for Minimizing Sky Glow

The following steps can help minimize sky glow from roadway lighting:

- Use light levels that are no higher than needed for the type of road being illuminated.
- Avoid installing fixtures that produce upward light.
- Near astronomical observatories, consider using LPS lighting systems.

Light Trespass

The phenomenon of light trespass occurs when light from a roadway lighting installation or other outdoor location is incident on nearby properties. This can be especially bothersome when light enters the windows of residences at night. This section addresses how different aspects of lighting can influence the perception of light trespass from roadway and other outdoor lighting.

Quantity of Light

Obviously, a lighting installation that casts much of its light onto adjacent properties rather than onto the location of interest (e.g., the road right-of-way) will potentially increase the likelihood of that installation producing light trespass. Some recommendations for preventing light trespass have stipulated upper limits for the illuminance that can fall onto the windows of nearby properties. These limits (CIE, 2003) can range from 1 to 25 lux depending upon the time of night (lower illuminances for postcurfew periods and higher for precurfew) and the type of location (lower for rural locations, and higher limits for urban locations). Adaptive roadway lighting that reduces light levels during hours when fewer vehicles and pedestrians are using the road, and when residents are more likely to be sleeping, can also be considered as a strategy for reducing light trespass.

Spectral Distribution

As described in the previous section of this chapter regarding the spectral distribution of light as it pertains to sky glow, the brightness appearance of a lighted space is influenced by the amount of short-wavelength (“blue”) light produced by the light source. Thus, it is possible that a roadway lighting fixture that produces light with a high CCT (cool-white appearance) would make an interior space look brighter than one with a lower CCT (warm-white appearance) even if the measured amount of light reaching a window from outside were the same. Nonetheless, this would be a relatively small impact compared to the quantity of light.

Spatial Distribution

In order to cast light onto the windows of adjacent residential properties, light from roadway fixtures needs to be emitted at relatively high angles from directly below the fixture. However, it should also be noted that the illuminance on a surface from a light source is inversely proportional to the square of the distance between the source and the surface. **Figure 2** illustrates the maximum vertical illuminance from an outdoor lighting installation for different distances from the property line. Right along the property line the maximum vertical illuminance is nearly 200 lux, higher than many light levels in residences during the daytime, and greater than any recommended maximum level for light trespass (CIE, 2003).

As the distance from the property line increases, however, the maximum vertical illuminance decreases sharply so that for a distance of 20 m from the property line the illuminance is only 1.5 lux. This illustrates the strong influence the location of streetlight poles as well as of windows can have on light trespass. In an urban area, building facades are often located very close to the street while in rural locations the building may be set far back from the street; each of these will have very different amounts of light potentially entering windows. Another

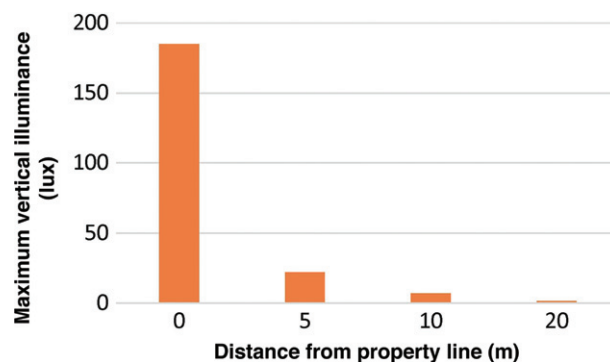


Figure 2 Maximum vertical illuminances from an outdoor lighting installation, as a function of the distance from the property line.

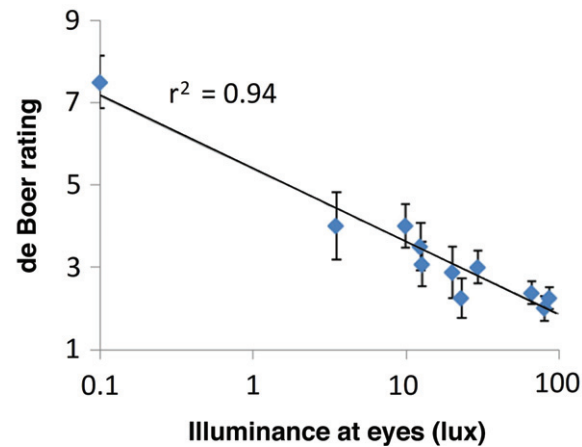


Figure 3 Ratings of discomfort glare on the *de Boer* (1967) scale (9 = just noticeable glare, 7 = satisfactory, 5 = just acceptable, 3 = disturbing, and 1 = unbearable) under nighttime conditions as a function of the illuminance at the eyes from a glare source.

mitigating factor for light trespass can be the installation of house-side shields on streetlighting fixtures to block light that could reach residential windows.

Recommendations for Minimizing Light Trespass

In order to minimize the potential for roadway lighting installations to contribute to light trespass, consider the following:

- Locate streetlight poles as far from adjacent properties as practical.
- Use adaptive lighting strategies to reduce light levels during late hours of the night.
- Use house-side shields to reduce the amount of light that can enter windows from streetlights.

Glare

Glare can be defined in two ways: one is the reduction in visibility from a bright light source caused by scattered light within the eyes, and another is the sensation of discomfort, annoyance or even pain created by a light that is much brighter than its surrounding environment. Roadway lighting can contribute to both types of glare because streetlights are by nature, bright points of light viewed during otherwise dark conditions. This section discusses how different aspects of lighting influence perceptions of glare.

Quantity of Light

At night, the intensity of a light source plays a major role in the degree to which that light source can create glare, as illustrated in [Figure 3](#). A secondary factor is the background light level; when a roadway is illuminated to a higher luminance, this actually reduces the impacts of glare because the contrast between the light source and the surrounding area is lower. This is why vehicle headlights are often judged as glaring at night, but when viewed under daytime conditions do not create any discomfort at all. The [IES \(2018\)](#) states that part of the reason for roadway lighting is in fact to reduce the effects of discomfort glare.

Spectral Distribution

Similar to the perception of brightness of a lighted scene, two bright light sources that have the same intensity but different spectral distributions will not necessarily be judged as equally uncomfortable. A source that produces more short-wavelength output (or has a higher CCT) will tend to elicit more discomfort than one that is dominated by longer visible wavelengths ([Bullough, 2009](#)).

Spatial Distribution

When roadway lighting began to be used widely on streets, a “cutoff” classification system was developed to define the maximum intensity at high angles from directly below a streetlighting fixture. Noncutoff streetlights had the highest maximum intensities; semi-cutoff streetlights intermediate values, and cutoff streetlights produced the lowest high-angle intensities, and as a consequence, cutoff luminaires tended to produce the least amount of discomfort glare. This system was used with success for many decades ([Bullough, 2002](#)) until it was replaced with a similar system ([IES, 2011](#)) that also addresses upward light from a street lighting fixture. Discomfort and disability glare are both also reduced when the distance between the glare source and the line of sight is increased. For this reason, streetlight fixtures should only be mounted on poles that are less than 5 m in height when the wattage of the lights is low.

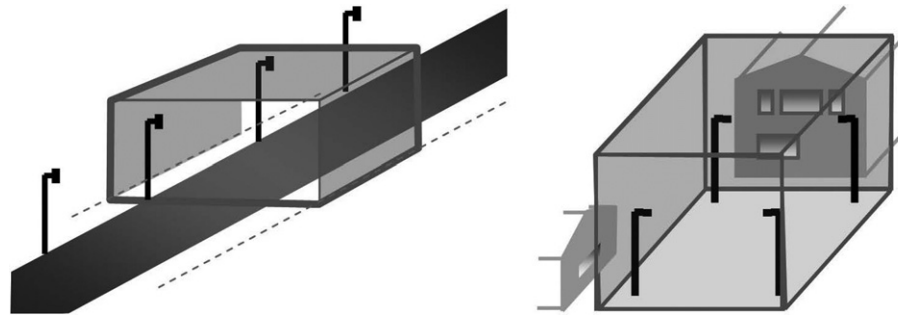


Figure 4 Virtual “box” enclosure used in the OSP system.

Recommendations for Minimizing Glare

In order to reduce the potential for roadway and other outdoor lighting installations to create glare, consider the following:

- Minimize high-angle intensity from the streetlight fixtures.
- Use lower CCTs when feasible as a secondary option.
- Mount bright streetlights higher than 5 m from the ground to maximize the angle between the streetlight and drivers’ or pedestrians’ line of sight.

Outdoor Site-Lighting Performance System

In order to compare lighting configurations in terms of light pollution impacts including the potential to produce sky glow, light trespass or glare, the outdoor site-lighting performance (OSP) system was developed (Brons et al., 2008). The basis for this system is a virtual ‘box’ enclosure around a lighted roadway or parking lot (Figure 4) that represents the boundary between a lighted location and the exterior parts of the environment that can be affected by streetlighting or other outdoor lighting.

Using the OSP system (Brons et al., 2008), the potential for sky glow is quantified by estimating the average illuminance on the upper surface and the tops of the side surfaces of the box leaving the lighted area. The potential for light trespass is quantified by estimating the maximum illuminance on the side surfaces leaving the lighted area. Glare is quantified using the illuminance from the lighting fixtures as well as the illuminance from the surrounding surfaces and the ambient environment (Bullough et al., 2008), at a height of 1.5 m along the side surfaces of the box. The OSP system is used in recommended practices for outdoor lighting at airports as a way to quantify and compare the potential to create light pollution (IES, 2015).

See Also

Lighting; Visual Impacts From Transport

References

- Bierman, A., 2012. Will switching to LED outdoor lighting increase sky glow? *Light. Res. Technol.* 44, 449–458.
- Brons, J.A., Bullough, J.D., Rea, M.S., 2008. Outdoor site-lighting performance: A comprehensive and quantitative framework for assessing light pollution. *Light. Res. Technol.* 40, 201–224.
- Bullough, J.D., 2002. Interpreting outdoor luminaire cutoff classification. *Light. Des. Appl.* 32, 44–46.
- Bullough, J.D., 2009. Spectral sensitivity for extrafoveal discomfort glare. *Mod. Opt.* 56, 1518–1522.
- Bullough, J.D., Brons, J.A., Rea, M.S., 2008. Predicting discomfort glare from outdoor lighting installations. *Light. Res. Technol.* 40, 225–242.
- Commission Internationale de l’Eclairage (CIE), 2003. Guide on the Limitation of the Effects of Obtrusive Light from Outdoor Lighting Installations, CIE 150. Commission Internationale de l’Eclairage, Vienna.
- Commission Internationale de l’Eclairage (CIE), 2010. Lighting of Roads for Motor and Pedestrian Traffic, CIE 115. Commission Internationale de l’Eclairage, Vienna.
- de Boer, J., 1967. Public Lighting. In: Philips Technical Library, Eindhoven, Netherlands.
- Illuminating Engineering Society, 2011. Luminaire Classification System for Outdoor Luminaires TM-15-11. Illuminating Engineering Society, New York.
- Illuminating Engineering Society, 2015. Outdoor Lighting for Airport Environments RP-37. Illuminating Engineering Society, New York.
- Illuminating Engineering Society, 2018. American National Standard Practice for Design and Maintenance of Roadway and Parking Facility Lighting RP-8-18. Illuminating Engineering Society, New York.
- Longcore, T., Rich, C., 2004. Ecological light pollution. *Front. Ecol. Environ.* 2, 191–198.
- Rea, M.S., Radetsky, L.C., Bullough, J.D., 2011. Toward a model of outdoor lighting scene brightness. *Light. Res. Technol.* 43, 7–30.
- Rea, M.S., Bierman, A., 2015. Spectral considerations for outdoor lighting: Consequences for sky glow. *Light. Res. Technol.* 47, 920–930.
- Riegel, K.W., 1973. Light pollution. *Science* 179, 1285–1291.
- Waldram, J.M., 1972. The calculation of sky haze luminance from street lighting. *Light. Res. Technol.* 4, 21–26.

Wildlife Crossings and Barriers

Scott D. Jackson, Department of Environmental Conservation, University of Massachusetts, Amherst, MA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	297
Impacts of Roads and Railroads on Fish and Wildlife	297
Effects on Habitat	297
Road Mortality	298
Barrier Effects of Transportation Infrastructure	298
Importance of Animal Movement	298
Roads, Highways, and Railroads as Barriers	299
Impacts of Increased Mortality and Reduced Wildlife Passage	300
Mitigating Transportation Impacts on Fish and Wildlife	301
Mitigation Techniques That Don't Involve Crossing Structures	301
Wildlife Crossing Structures	301
Road-Stream Crossings	302
Conclusion	303
Review Articles	303
Relevant Websites	303
References	304
Further Reading	304

Introduction

Roads, highways, and railroads are components of transportation infrastructure that are essential to functioning societies. Dense networks of roads generally occur where residential, commercial, and industrial development is most concentrated. Railroads and highways typically stretch out over long expanses of land, connecting people of different counties, parishes, states, provinces, and countries. Many of the environmental impacts of roads and railroad, and the vehicles that travel on them, are well known. Air, water, and noise pollution, the contamination of soil and vegetation, as well as automobile accidents are unavoidable consequences of maintaining elaborate networks of transportation infrastructure. While impacts to environmental quality, human health, and safety are well integrated into transportation policy, planning, and regulation, another environmental consequence of extensive transportation networks—the effects of roads, highways, and railroads on fish and wildlife populations—receives far less attention.

The effects of roads, highways, and railroads (transportation infrastructure) are most obvious when animal–vehicle collisions cause property damage, injury, or death, or by the presence of road-killed animals along roads and highways. Wildlife mortality caused by vehicle collisions is an important factor affecting the persistence of wildlife populations, but the impact of transportation infrastructure includes other elements that negatively affect wildlife populations. These include habitat loss and degradation, as well as the barrier effects of roads, highways, and railroads.

The ecological road-effect zone, the area of adjacent land that is ecologically affected by a road or highway, is variable in width depending on the size and characteristics of the roadway but averages around 600 m (Forman and Deblinger, 1999). There are approximately 64 million km of paved and unpaved roads on earth (CIA) and it is estimated that 83% of the continental United States is located within 1 km of a road or highway (Ritters and Wickham, 2003). The amount of new transportation infrastructure is growing rapidly in the developing world, and although the pace of new road construction has slowed in the developed world, existing roads/highways are being widened and traffic volumes are increasing. Collectively, the roads, highways, and railroads around the globe have had an enormous impact on fish, wildlife, and ecosystems. To conserve the Earth's biodiversity resources we must consider how we might reduce these impacts both for existing transportation infrastructure and for new roads, highways, and railroad yet to be constructed.

This chapter focuses on the impact of transportation infrastructure on fish and wildlife populations, and what we can do to avoid, minimize, or mitigate those impacts. Much of the research on this subject has examined the effects of roads and highways; thus, much of the chapter will focus on those impacts. However, the effects of railroads on wildlife populations are quite similar, and many of the concepts and examples described for roads and highways are just as relevant for railroads.

Impacts of Roads and Railroads on Fish and Wildlife

Effects on Habitat

Roads, highways, and railroads are poor habitat for most fish and wildlife. Wherever these elements of transportation infrastructure replace natural areas there is a direct loss of habitat. The amount of available habitat can be further reduced if wildlife avoid natural

areas associated with roads, highways, and railroads because of light, noise, edge effects, or the commotion caused by vehicular traffic. Wetland and aquatic habitats are degraded by changes in hydrology, water quality, water temperature, and loss of outside inputs such as leaf litter and woody debris. Degraded habitat may be able to sustain fewer individuals of a given species or may be rendered completely unsuitable as habitat.

Habitat fragmentation occurs when areas of contiguous habitat (e.g., forest) is dissected by transportation infrastructure, creating smaller patches of that habitat type, more edges (e.g., where forest abuts more open habitats such as grasslands or shrublands), and less core area (areas insulated from edges). Some wildlife species are edge sensitive, meaning that their survival or reproduction is reduced by edge predators, brood parasites (e.g., brown-headed cowbirds, *Molothrus ater*, in North America that deposit their eggs in other birds' nests), and changes in microclimate that are unfavorable for the species. As patches of habitat become smaller and more isolated, they support smaller populations of fish or wildlife. Some species are so vulnerable to edge effects that they are considered area-sensitive, meaning that they will not nest in habitat patches below a certain size.

Another aspect of roads and highways that can affect the amount and quality of habitat available to fish and wildlife is the secondary impacts of road building. Where roads go people and development usually follow. Road building facilitates development and extractive industries that can result in habitat loss and degradation that extends farther out from the road footprint and well into the future.

Any reduction in available habitat, due to habitat destruction or wildlife avoidance of areas near transportation infrastructure, reduces the fish and wildlife carrying capacity of the area. Smaller and more isolated populations are more vulnerable to the adverse effects of random (stochastic) events. These include environmental stochasticity, such as atypical weather conditions (unusually rainy or dry years, severe winters), or changes in the abundance of food, shelter, predators, competitors, parasites, and diseases. Very small populations are also vulnerable to demographic instability and genetic changes due to random chance. The smaller a population, the more susceptible it is to random or unusual events and the less likely that it will persist well into the future.

Road Mortality

An obvious impact of transportation infrastructure is the loss of individual animals to roadkill. Road mortality rates tend to be highest right after the roadway is opened to traffic and decreases over time as wildlife populations are diminished or eliminated. One study conducted in western France a year after a new motorway opened documented 2266 road-killed vertebrates representing 97 species over the course of 33 weeks (Lodé, 2000).

The impact of road mortality on wildlife populations may range from severe to negligible, even if the numbers killed on the road are high. It is not the absolute number of individuals killed that matters. It is whether the road mortality is compensatory or additive. For some species, human-caused mortality (roadkill or hunting) may have a minimal effect on populations if the individuals killed are likely to have died anyway, due to predation or winter mortality. For example, many species of frogs and toads produce large year classes of young that experience high rates of natural mortality. This is true for many prey species with high reproductive rates, such as rodents and rabbits. Furthermore, some species can compensate for elevated mortality rates through increased reproduction, becoming reproductive at an earlier age, or via immigration from surrounding areas.

At the other end of the spectrum are species for which any human-caused mortality is likely to be in addition to, rather than in place of, natural mortality. This typically occurs in species that have naturally low reproductive rates and relatively long life spans. Large carnivores, such as grizzly/brown bears, *Ursus arctos*, and wolverines, *Gulo gulo*, typically fall within this group. Road mortality has been identified as an important threat affecting endangered species/subspecies such as the Florida Panther, *Felis concolor coryi* (Maehr et al., 1991), Old World badger, *Meles meles* (van der Zee et al., 1992), and Iberian lynx, *Felis pardina* (Ferrerias et al., 1992). Turtles and tortoises are particularly vulnerable to population decline due to road mortality (Beaudry et al., 2008). These animals are generally characterized by low reproductive rates and high adult survival rates (in excess of 100 years for some species). Adult longevity makes up for low reproductive potential. Slow moving and often attracted to road shoulders for nesting, turtles and tortoises are vulnerable to vehicle strikes, and road mortality is usually additive, killing animals that are unlikely to die otherwise. Due to their low reproductive rates, they have very little capacity to compensate for the added adult mortality.

Transportation infrastructure can also increase wildlife mortality by serving as attractive nuisances. Grassy roadside verges can attract grazing animals, putting them in close proximity to vehicular traffic. Ungulates and North American porcupines, *Erethizon dorsatum*, in need of sodium will seek out road and highway shoulders for road salt. Spilt grain along roadsides or railroad tracks can attract and concentrate wildlife, creating opportunities for roadkill. Storm water management systems can attract amphibians to busy roadsides in order to breed or forage in artificial basins, which may not hold water long enough for deposited eggs to hatch and for larvae to develop and transform into juveniles.

Roads and highways also indirectly increase wildlife mortality by facilitating human access to areas that would be otherwise difficult to reach. Wildlife mortality due to hunting, poaching, or extractive industries, only adds to the direct toll taken by the roads, highways, and railroads themselves.

Barrier Effects of Transportation Infrastructure

Importance of Animal Movement

Movement is critically important to the survival of individual animals and the persistence of fish and wildlife populations. Even relatively sedentary species, such as freshwater mussels, have a dispersal life stage that facilitates genetic mixing and colonization of

vacant habitat. In North America, freshwater mussels produce parasitic microscopic larvae called glochidia that are released into water. These glochidia attach to gills or fins of particular species of fish (or aquatic salamanders) where they will complete their larval development and eventually drop off and, if they land in an area of suitable habitat, will continue their lives as juveniles, then adult mussels. The persistence of mussel populations is dependent on the presence of suitable fish hosts and an environment that allows those fish to move within their aquatic habitats (streams, rivers, lakes, and ponds).

Individual animals move within and between areas of suitable habitat for a variety of reasons. Some are routine daily movements to forage or seek shelter from predators or the elements. Animals move to access vital habitats such as breeding habitat, mineral licks (salt licks), and thermal refuges (especially important for cold-water fish). Some species migrate long distances between summer and winter habitat. Some are movements tied to reproduction, as individuals seek mates or suitable nesting, denning, or spawning habitat.

Many amphibian species undergo seasonal migrations from terrestrial overwintering habitat to aquatic breeding sites in the spring. Although many turtles are aquatic or semi-aquatic, all but a few species must lay their eggs on land. These nesting movements present a number of hazards (including roads) to adult female turtles. In order to mature and complete their life cycles, juvenile amphibians and hatchling turtles must complete these movements between aquatic and terrestrial habitats in reverse. Breeding movements may involve long-distance travel, such as when anadromous fish like salmon, river herring, and sturgeon migrate from the ocean up into freshwater streams and rivers to spawn. Eels are catadromous fish whose adults migrate from freshwater to the ocean and the juvenile life stages migrate back upstream to complete their development in freshwater habitats.

Another reason why movement is so important for fish and wildlife is that habitat conditions change. Some of that change is natural and expected such as when vegetative communities change due to natural succession, or when the arrangement of sediment and woody debris in streams shifts in response to seasonal high flows, causing changes in the distribution of riffles, pools, or runs. In other cases, change may be sudden and unexpected, such as fire, floods, droughts, or severe storms. The change in conditions may be short term or long term, but individual animals often need to relocate to areas with more favorable conditions.

Movement is also important at the population level. Populations are groups of individual animals that regularly interact and interbreed. As discussed above, movement is essential for reproduction and the routine interactions among individuals that make up populations. Constraints on movement often delineate one population from another, each population distinguished by demographic characteristics (birth, death, immigration, and emigration rates) that are independent from other populations. Animal movements help maintain the continuity of populations and barriers to movement can fragment populations, producing smaller populations with reduced long-term viability. Dispersal of individuals away from populations is important for regulating population density and maintaining abundance levels within the carrying capacity of their environment.

Despite their inherent vulnerability to environmental, demographic, and genetic stochasticity, small populations of fish and wildlife can persist for long periods of time when they are part of a network of local populations linked together via occasional movements of individuals between and among those populations. These networks of populations, otherwise known as regional populations or metapopulations, are very important for the long-term maintenance of local populations.

Within regional populations, the movement of individuals between local populations is rare enough that the local populations remain distinct from each other, but occurs often enough that it helps maintain the regional population over time. Dispersing individuals from neighboring populations may provide immigrants that supplement the local population and help it maintain population levels even when local reproduction is insufficient to do so on its own. Occasional movement of individuals between and among populations can provide an infusion of genes (gene flow) that can partially or entirely negate the genetic deterioration that often results from small population size. When local populations do fail, the movement of individuals from other nearby populations provides a source of colonizers that can reestablish populations in habitat where local extinctions have occurred. A balance between local extinctions and recolonizations can maintain networks of local populations far longer than isolated populations could persist in the absence of such population level exchanges.

Roads, Highways, and Railroads as Barriers

Mortality is a significant barrier associated with transportation infrastructure. Animals killed on the road fail in their attempts to cross roads, highways, or railroads. However, the barrier effects of transportation infrastructure go well beyond that. These additional factors generally fall into two categories: physical barriers and behavioral barriers.

Fish and other aquatic organisms are likely to encounter various barriers associated with road-stream crossings. Most, but not all, bridges provide good aquatic organism passage. Culverts, on the other hand, rarely provide full passage for all aquatic organisms that inhabit those streams. Because culverts usually do not span the entire width of the stream, they concentrate flows and result in velocities that are higher than would normally be encountered in the natural stream channel. Fish and other aquatic organisms that cannot swim faster than the velocity of the water are presented with a velocity barrier that can delay or completely obstruct upstream passage. Some species that can manage to swim faster than the current often cannot keep up the effort long enough to make it through the culvert (exhaustion barrier). Concentrated flows create more erosive power and often create scour pools at the outlet that leave the bottom of the culvert perched above the water surface in the receiving pool below. Some fish, such as some species of salmon and trout, are excellent jumpers and may be able to overcome sizeable outlet drops. However, many fish cannot jump at all or are weak jumpers, and these outlet drops represent significant obstacles (jump barriers) to passage. Undersized culverts will typically back water up at the inlet which can result in turbulence that can disorient fish (turbulence barrier) and the accumulation of substrate at the inlet resulting in inlet drops.

Fish often encounter multiple barriers when trying to pass through culverts. It is possible that a fish traveling upstream to spawn, or up into a small tributary to escape warm summer water temperatures, could encounter an outlet drop requiring a jump to get into the culvert, a velocity or exhaustion barrier, and end with a second jump barrier to exit the culvert over an inlet drop. For fish, or other aquatic organisms that can swim but cannot jump, they will be blocked at the outlet without even getting into the culvert.

For terrestrial wildlife, physical barriers include the surface of roads and highways, as well as the rails and ballast of railroads. Although railroad tracks hardly seem like insurmountable barriers to movement, turtles find it difficult to cross over the tracks and have sometimes been trapped between the rails. The ballast stones used along some rail lines generally contrast with the substrate of nearby areas. Animals, such as moles, that travel below ground would presumably find it difficult to cross railroad tracks atop ballast stone. Similarly, wide sections of paved road or highway would present a significant physical obstacle to moles and other burrowing animals.

There are other features of roads and highways that serve as barriers to wildlife passage. These include Jersey barriers, retaining walls, noise barriers, steep embankments, deep road cuts, and right-of-way fencing. On smaller roads, something as simple as traditional curbing, when combined with catch basins, can serve as traps for salamanders and small (e.g., hatchling) turtles. When encountering curbs, these animals seek a route around the barrier. As they walk along the curb and encounter a catch basin they often fall into and are trapped in the deep sumps of these basins.

Behavior barriers occur when characteristics of the road, highway or railroad, present conditions that are risky for wildlife to cross. Some wildlife avoid roadsides because of the noise, light and/or activity of vehicular traffic. Others are reluctant to expose themselves to predators by crossing areas that are essentially devoid of cover. Studies have documented that several species of small mammals are reluctant to cross even relatively small roads (Oxley et al., 1974; Mader, 1984; Swihart and Slade, 1984). Although some arboreal wildlife (species that spend most of their lives in tree canopies) will move across the ground and might be willing to cross over a road surface, other species generally stick to the trees and are very reluctant to travel on the ground. The wider the expanse of road the stronger the deterrent to passage. It is not hard to image that few animals will successfully cross a six or eight lane superhighway, most won't even try.

Impacts of Increased Mortality and Reduced Wildlife Passage

When considering the impacts of transportation infrastructure, it is important not to get too focused on impacts to individual animals, but to focus instead on the long-term impacts on fish and wildlife populations. The ability of populations to persist over time is a function of demographic rates (birth, death, immigration, emigration) which, in turn, are determined by demographic characteristics (abundance, spatial distribution, sex ratio, age structure, genetic structure) of those populations. To the extent that road mortality and the loss of connectivity affects these demographic rates and characteristics, they can affect population growth rates, and may push populations into steep or gradual declines. Here is a summary of the effects of roads, highways, and railroads on fish and wildlife populations.

- Roadkill leading to the loss of populations
- Reduced access to vital habitats or habitat features (breeding habitat, overwintering areas, thermal refuges, mineral licks, etc.)
- Reduced ability to relocate in response to changing habitat conditions
- Fragmentation of populations, creating smaller more isolated populations that are vulnerable to demographic and environmental stochasticity, catastrophes (fire, flooding, drought, etc.), and the loss of genetic variability due to genetic drift and inbreeding depression
- Disruption of processes (supplementation, gene flow, and recolonization) that maintain regional populations over time

Following are some examples of how roads, highways, and railroads negatively affect fish and wildlife populations.

Research on seven salmonid fish in western North America found that the percentage of areas with healthy populations was 58% in roadless watersheds but declined to 16% for watersheds with road densities exceeding 4 km/km². In a study of Agassiz's desert tortoise (*Gopherus agassizii*), Nafus et al., 2013 concluded that roads may decrease tortoise populations via cumulative mortality and reduced population growth rates due to the loss of reproductive adults. Steen and Gibbs, 2004 and Aresco, 2005 observed turtle populations that had sex ratios that were dramatically biased toward males, and attributed it to the effects of road mortality on nesting females.

Sawaya et al., 2019 documented demographic fragmentation of a wolverine population bisected by a major transportation corridor and concluded that sex-biased impacts on dispersal (females rarely crossed the highway) led to genetic isolation. Similar sex-biased impacts were documented for grizzly bears, *Ursus arctos horribilis* (Proctor et al., 2005). The authors of this study concluded that although movement of male bears across the highway was sufficient to mediate concerns about genetic connectivity, disruption of female dispersal would reduce the possibility for supplementation and re-colonization.

An early example of the impact of transportation infrastructure on population genetics was documented by Reh and Seitz, 1990. They found significant genetic differences in a population of common frog, *Rana temporaria*, that was surrounded by roads, a highway, and a railway, and presumably isolated from other populations of that species. Since then, many studies have documented the loss of genetic diversity due to road mortality and the loss of connectivity as well as genetic differentiation among populations separated by highways. Epps et al., 2005 looked at genetic diversity of 27 populations of desert bighorn sheep, *Ovis canadensis nelsoni*, and discovered a rapid reduction in genetic diversity due to isolation by highways, canals and developed areas. Studying wildlife affected by a freeway in southern California (USA), Riley et al., 2006 found genetic isolation in carnivore populations despite moderate levels of dispersal across the highway. A 2010 review of genetic effects of roads and highways (Holderegger and Di

Giulio, 2010) found that roads often decrease genetic diversity of affected wildlife populations due to reduced population size, and decrease functional connectivity as indicated by increased genetic differentiation of populations.

With information about demographic rates and characteristics, and how those rates and characteristics are affected by transportation infrastructure, ecologists can use population viability modeling to estimate persistence time for populations under various scenarios. Taylor and Goldingay, 2009 estimated the probability of extinction for the greater glider, *Petauroides Volans*, for a population in a remnant area of habitat in Australia with and without a land bridge to facilitate highway crossings. Their modeling suggested that if no dispersal across the road occurs, that subpopulation of this arboreal mammal has a very high probability of extinction within the next 100 years. Persistence improves dramatically when a moderate amount of dispersal is included in the model. Ramp and Ben-Ami, 2006 modeled population viability for the swamp wallaby, *Wallabia bicolor*, and found that a 20% decrease in female roadkill deaths had the potential to stem a projected population decline and improve long-term viability for this population.

Although some population extinctions occur soon after a motorway is opened to traffic, changes in demography and population genetics typically take many years to manifest and there is likely to be a significant lag time involved in transportation-related population loss. A study in Ontario, Canada, by Findlay and Houlahan, 1997 investigated species richness for birds, reptiles, amphibians, and plants in 30 wetlands relative to surrounding land use. They found a significant relationship between species diversity and the density of paved roads within 1.9 km of the wetlands. Further analysis of the data (Findlay and Bourdages, 2000) found that the relationship between road density and species diversity was stronger when they used the road network as it existed in 1944. This suggests that lower species diversity observed today may be the result of roads, highways, and railroads built many years ago, and that impacts of transportation infrastructure are likely to continue well into the future.

Mitigating Transportation Impacts on Fish and Wildlife

There are many techniques and approaches that have been used to mitigate the impacts of roads, highways, and railroads on fish and wildlife populations. Some are primarily focused on reducing animal-vehicle collisions; others seek to both reduce road mortality and enhance connectivity across motorways and railroads. There have been numerous studies that have evaluated these various approaches and provide information about their performance for various species. However, the settings and the species differ from case to case, and it is not easy to draw general conclusions about the efficacy of mitigation techniques.

Mitigation Techniques That Don't Involve Crossing Structures

Curbs: A simple way to reduce road mortality and improve crossing success for salamanders and turtles is to avoid traditional curbing in favor of gently sloping berms or country drainage (no curbs).

Fencing: Fencing is often used to prevent large mammals from accessing motorways. They are relatively ineffective for climbing animals or those that can easily burrow under the fences. Fencing may be successful at reducing animal-vehicle collisions but they can exacerbate the barrier effects of transportation infrastructure.

Warning signs: Wildlife warning signs are widely used because they are relatively inexpensive. Motorists, however, grow accustomed to the signs and any effects on driver behavior are typically short-term. Signs that are erected and taken down seasonally have a greater chance of being successful.

Animal detection systems: Used in combination with warning signs and flashing lights, animal detection systems alert motorists when large animals have been detected in (or in close proximity to) the road ahead. In the past, these systems have proven relatively ineffective because false positive detections can lead drivers to conclude that they cannot be trusted.

Crosswalks: Fencing or habitat manipulation can be used to concentrate wildlife into designated crossing zones, or crosswalks. These crossing areas are chosen for their good visibility and are accompanied by warning signs and, occasionally, animal detection systems. While in some cases, this approach has reduced the number of animal-vehicle collisions along a stretch of road, it does create collision hot spots where concentrations of animals may be on or around the roadway.

Reflectors, mirrors, and flagging models: These techniques are typically used to keep ungulates (deer, moose, and elk) off roads and highways. Reflectors and mirrors spray light from vehicle headlights into the surrounding landscape in hopes that wildlife will be alerted to the presence of vehicles and/or spooked into leaving the roadway. Flagging models use cutouts of deer rumps with erect tails to signal alarm to approaching deer. None of these techniques has been particularly effective at preventing animal-vehicle collisions.

Chemical repellents: A handful of chemical repellents have been tried in an effort to keep wildlife away from roads, with mixed success. Some studies report positive effects; others report no effect. The same repellent can successfully repel some species while actually attracting others. When successful, chemical repellents tend to yield only modest benefits.

Forested medians: Studies of arboreal mammals in Australia have documented the positive effects of forested medians. When the width of a highway is too great for mammals to glide across, successful crossings can occur using the forested median as a stopover point (van der Ree et al., 2010). Forested medians may also serve as habitat for small mammals (a genetic reservoir) that helps maintain genetic connectivity between populations that would otherwise be completely separated on opposite sides of a highway.

Wildlife Crossing Structures

Wildlife crossing structures, built specifically to facilitate wildlife passage across roads, highways, and railroads, come in many sizes and designs. Different wildlife species often have specific requirements and there are no designs that meet the needs of all species. Small and inexpensive structures can be used more frequently to make transportation infrastructure more permeable. Larger, more

expensive structures tend to accommodate more species, but with large spaces between structures that may not serve the needs of small animals and species with limited mobility.

Structures for arboreal wildlife: Canopy bridges and glider poles have proven successful at assisting arboreal species (including gliding mammals) across roads in Australia (Soanes et al., 2015) and may have at potential for promoting connectivity for tree dwelling and gliding animals in other parts of the world.

Amphibian tunnels: Underpass systems designed and located specifically to facilitate breeding migrations of frogs, toads and salamanders have an extensive history in Europe and their use has more recently spread to other countries. Various designs have been tried with some designs proving more effective than others. Open-top structures are preferred because they allow rain to enter the tunnels and keep the substrate wet during migrations.

Upland culverts: Culverts ranging in size from 0.5 to 5.5 m in diameter have been used specifically to facilitate wildlife movement under roads, highways, and railroads. Small culverts can be effective for species of wildlife like weasels, mink, stoats, and martens, that are comfortable entering confined spaces, but are unlikely to be used by most wildlife. The larger the culvert size, the more species likely to be accommodated.

Wildlife underpasses: Wildlife underpasses are relatively large structures designed to allow passage for large wildlife (e.g., ungulates and large carnivores). They typically can accommodate a wide range of species, although some species prefer to use smaller structures.

Wildlife overpasses (ecoducts): Wildlife overpasses, also known as green bridges or ecoducts, are bridges that allow wildlife to pass over roads and highways. These are very open structures that suit the needs of species, such as gray wolves, *Canis lupus*, and grizzly bears that are wary of confined spaces. Overpasses are most effective when they are wide (>30 m) and are designed with natural substrate and vegetation; some include cover objects and ponds. These are expensive, but highly effective structures that are used by a wide variety of wildlife.

Guide fencing: With rare exceptions, wildlife crossing structures are ineffective unless accompanied by guide fencing to funnel animals to the structure entrances. Fencing can be made of a variety of materials depending on the needs of wildlife for which the structures are intended. To guide large animals to structures, rows of boulders are sometimes used instead of fencing.

Stump rows: Rows of tree stumps can providing cover for small animals, such as mice, ground squirrels, lizards, and snakes, making it more likely that these species will use large structures like underpasses and overpasses. Rock piles or other cover objects might also work, depending on the target species.

There are many factors that can affect the success of wildlife crossings structures, including location, size, light, moisture, substrate, vegetation, cover objects, guide fencing, approaches, and human activity. For more detail about structure design and performance, consult some of the materials in the Further Reading section below.

Road-Stream Crossings

As long linear ecosystems, rivers and streams stretch out over large areas of the landscape. These river and stream networks are transportation systems for water, sediment, woody debris, and organisms. Roads, highways, and railroads are linear elements of human transportation infrastructure that traverse many of these same landscapes. It is inevitable that these two systems will frequently cross. Occasionally, waterways will disrupt the human transportation system, such as when a flood causes a culvert or bridge to fail. More often, it is roads, highways, and railroads, and more specifically the culverts (and sometimes bridges) associated with road-stream crossings, that disrupt river and stream networks.

The focus of traditional culvert design is almost exclusively on hydraulics and hydrology; how to get water efficiently across the road without disruption to the transportation system. Over time, it became clear that hydraulic design approaches often result in crossings that are impassable for some or all of the aquatic organisms that inhabit streams and rivers, as well as disrupting the transport of sediment and woody debris that is important for maintaining stream habitat.

The hydraulic approach can also be used to create passable stream crossings by designing culverts that are compatible with the swimming and leaping capabilities of particular target species (usually fish). When used in this way, crossings can be designed and constructed to allow passage for some species, but they rarely accommodate all the aquatic organisms that need to move throughout the stream network. Furthermore, we lack detailed information about the swimming and leaping abilities of most species of fish and aquatic wildlife; information needed to hydraulically design aquatic passage.

Stream simulation is an alternative approach for stream crossing design that seeks to create conditions (substrate, bed features, water depth, and velocity) within the crossing structure that mimic (i.e., fall within the natural range of variability) characteristics of the natural channel. Many bridges effectively meet the objectives of stream simulation. Stream simulation crossings are either embedded culverts or open-bottom structures (arches or three-sided boxes) that are as wide, or wider, than the width of the stream. To avoid insufficient water depths at low flow, channels are either maintained or constructed within the structures to achieve water depths and velocities at a range of flows that are comparable to those found in the natural channel.

Road-stream crossings also have the potential to provide passage for semi-aquatic and terrestrial wildlife. Many wildlife species will cross roads and highways under bridges, if those structures are large enough and provide relatively dry passage along one or both banks. Small animals can take advantage of culvert crossings on intermittent streams when they are dry. Some of the best opportunities for mitigating transportation infrastructure impacts on landscape connectivity are at road-stream crossings, where some kind of structure will already be planned or in place. The ideal situation for wildlife is when the roadway or railroad crosses the entire stream or river valley with a long bridge, also known as a viaduct.

Using stream simulation design for a crossing will generally make it passable for a number of semiaquatic and terrestrial wildlife, while also maintaining aquatic organism passage. To accommodate large wildlife species (e.g., elk, moose, grizzly bears) structures will generally need to be larger than necessary simply to achieve stream simulation. However, assuming the location is suitable, it might be more feasible to increase the size of a road-stream crossing than to construct a wildlife underpass or overpass. There are also modifications to culverts and bridges that can be used to expand the number of wildlife species capable of using road-stream crossings to get across transportation infrastructure.

Expanded bridges: Bridges can be expanded to provide a dry path with appropriate substrate along one or both banks.

Wildlife benches or shelves: Existing structures (bridges or culverts) that do not provide dry passage can be modified by the addition of benches (on the bottom of the structure) or shelves (attached to the sides of structures) to provide such passage. These can be relatively low-cost solutions that will facilitate movement for a variety of smaller animals.

Vole tubes: Voles are small mammals (rodents) that often travel through shallow tunnels in soil or dense vegetation. They are most comfortable in confined spaces and generally avoid open areas. Affixing tubes to wildlife shelves can significantly improve the use of these structures by voles (Foresman, 2003).

Conclusion

Historically, conservation efforts have focused on protecting large blocks of undeveloped land, rare and endangered species habitat, and providing recreational opportunities. As our natural areas have been increasingly fragmented by extensive road networks, residential, commercial, and industrial development and associated infrastructure, maintaining or enhancing connections among natural areas is now seen as essential for protecting the long-term viability of wildlife populations and the ecological integrity of ecosystems. To protect biodiversity for future generations, we must assess the connections among our natural areas and wildlife habitats, design strategies to protect existing connectivity, and mitigate barriers to species movement.

There are many tools and approaches available to mitigate the impacts of roads, highways, and railroads on fish, wildlife, and ecosystems. These tools and approaches can be deployed in combination and at a variety of scales to ensure an interconnected landscape that can support the long-term persistence of fish and wildlife populations.

At local scales, the focus may be on threatened and endangered species, and strategies to maintain landscape connectivity for small, less vagile species. This may involve multiple small crossing structures (e.g., amphibian and reptile tunnels) and a focus on improving passability at road-stream crossings. A regional approach is needed to maintain viable populations of wide-ranging species and to ensure that metapopulation dynamics necessary to maintain regional populations remain intact. Long-term maintenance of metapopulations for wide-ranging species, and those, like large carnivores, that typically occur at low densities, will require a large, landscape-scale approach. To be successful at the regional and large landscape scales will require the use of landscape analysis and modeling, multiple tools and approaches for reducing road mortality and increasing landscape connectivity, and the cooperation of multiple stakeholders and governmental jurisdictions. Regional and landscape-scale connectivity will only become more important over time, as climate change requires wildlife to shift their geographic distributions in response to changing habitat conditions.

Finally, we do not yet have the knowledge or technology to mitigate all the negative effects of transportation infrastructure on wildlife and ecosystems. Maintaining roadless areas, and closing roads (e.g., forestry roads) to expand roadless areas, are essential strategies for ensuring the long-term persistence of highly sensitive species, such as wolverines, wolves, and grizzly bears.

Review Articles

- Glista, D.J., DeVault, T.L., DeWoody, J.A., 2009. A review of mitigation measures for reducing wildlife mortality on roadways. *Landsc. Urban Plan.* 91, 1–7.
- Holderegger, R., Di Giulio, M., 2010. The genetic effects of roads: a review of empirical evidence. *Basic Appl. Ecol.* 11 (2010), 522–531.
- Taylor, B.D., Goldingay, R.L., 2010. Roads and wildlife: impacts, mitigation and implications for wildlife management in Australia. *Wildlife Res.* 37, 320–331.
- Trombulak, S.C., Frissell, C.A., inpress. Review of ecological effects of roads on terrestrial and aquatic communities. *Conserv. Biol.* 14 (1), 18–30.
- Van der Grift, E.A., van der Ree, R., Fahrig, L., Findlay, S., Houlahan, J., Jaeger, J.A.G., Klar, N., Madriñan, L.F., Olson, L., 2013. Evaluating the effectiveness of road mitigation measures. *Biodivers. Conserv* 22 (2), 425–448.

Relevant Websites

The Stream Continuity Portal: <https://streamcontinuity.org/>

International Conference on Ecology and Transportation: <https://icoet.net/>

Infra Eco Network Europe: <http://www.iene.info/>

Road Ecology Center: <https://roadecology.ucdavis.edu/frontpage>

Wildlife and Roads: A Resource to Help Mitigate Roads for Wildlife: <http://www.wildlifeandroads.org/index.cfm>

References

- Aresco, M.J., 2005. The effect of sex-specific terrestrial movements and roads on the sex ratio of freshwater turtles. *Biol. Conserv.* 123, 37–44.
- Beaudry, F., deMaynadier, P.G., Hunter Jr., M.L., 2008. Identifying road mortality threat at multiple spatial scales for semi-aquatic turtles. *Biol. Conserv.* 141, 2550–2563.
- CIA World Factbook: <https://www.cia.gov/library/publications/the-world-factbook/>.
- Epps, C.W., Palsbøll, P.J., Wehausen, J.D., Roderick, G.K., Ramey I.I., R.R., McCullough, D.R., 2005. Highways block gene flow and cause a rapid decline in genetic diversity of desert bighorn sheep. *Ecol. Lett.* 8 (10), 1029–1038.
- Ferreras, P., Aldama, J.J., Beltran, J.F., Delibes, M., 1992. Rates and causes of mortality in a fragmented population of Iberian lynx *Felis pardina* Temminck 1824. *Biol. Conserv.* 61, 197–202.
- Findlay, C.S., Bourdages, J., 2000. Response time of wetland biodiversity to road construction on adjacent lands. *Conserv. Biol.* 14 (1), 86–94.
- Findlay, C.S., Houlihan, J., 1997. Anthropogenic correlates of species richness in Southeastern Ontario wetlands. *Conserv. Biol.* 11, 1000–1009.
- Foresman, K.R. 2003. Small mammal use of modified culverts on the Lolo South Project of Western Montana—an update. Pages 342–343 in Proceedings of the international conference on ecology and transportation. C. L. Irwin, P. Garrett, and K. P. McDermott, editors. Center for Transportation and the Environment, North Carolina State University, Raleigh, NC.
- Forman, R.T.T., Deblinger, R.D., 1999. The ecological road-effect zone of a Massachusetts (USA) suburban highway. *Conserv. Biol.* 14, 36–46.
- Holderegger, R., Di Giulio, M., 2010. The genetic effects of roads: a review of empirical evidence. *Basic Appl. Ecol.* 11, 522–531.
- Lodé, T., 2000. Effect of a motorway on mortality and isolation of wildlife populations. *Ambio* 29 (3), 163–166.
- Mader, H.J., 1984. Animal habitat isolation by roads and agricultural fields. *Biol. Conserv.* 29, 81–96.
- Maehr, D.S., Land, E.D., Roelke, M.E., 1991. Mortality patterns of panthers in southwest Florida. *Proc. Annu. Conf. of Southeast. Assoc. Fish and Wildl. Agencies* 45, 201–207.
- Nafus, M.G., Tuberville, T.D., Buhlmann, K.A., Todd, B.D., 2013. Relative abundance and demographic structure of Agassiz's desert tortoise (*Gopherus agassizii*) along roads of varying size and traffic volume. *Biol. Conserv.* 12 (2013), 100–106.
- Oxley, D.J., Fenton, M.B., Carmody, G.R., 1974. The effects of roads on populations of small mammals. *J. Appl. Ecol.* 11, 51–59.
- Proctor, M.F., McLellan, B.N., Strobeck, C., Barclay, R.M.R., 2005. Genetic analysis reveals demographic fragmentation of grizzly bears yielding vulnerably small populations. *Proc. R. Soc. B* 272, 2409–2416.
- Ramp, D., Ben-Ami, D., 2006. The effect of road-based fatalities on the viability of a peri-urban swamp wallaby population. *J. Wildlife Manage.* 70 (6), 1615–1624.
- Reh, W.A., Seitz, A., 1990. The influence of land use on the genetic structure of populations of the common frog *Rana temporaria*. *Biol. Conserv.* 54, 239–249.
- Riley, S.P.D., Pollinger, J.P., Sauvajot, R.M., York, E.C., Bromley, C., Fuller, T.K., Wayne, R.K., 2006. A southern California freeway is a physical and social barrier to gene flow in carnivores. *Mol. Ecol.* 15, 1733–1741.
- Ritters, K.H., Wickham, J.D., 2003. How far to the nearest road? *Front. Ecol. Environ.* 1, 125–129.
- Sawaya, M.A., Clevenger, A.P., Schwartz, M.K., 2019. Demographic fragmentation of a protected wolverine population bisected by a major transportation corridor. *Biol. Conserv.* 236 (2019), 616–625.
- Soanes, K., Vesk, P.A., van der Ree, R., 2015. Monitoring the use of road-crossing structures by arboreal marsupials: insights gained from motion-triggered cameras and passive integrated transponder (PIT) tags. *Wildlife Research*, 2015 42, 241–256. URL <http://dx.doi.org/10.1071/WR14067>.
- Steen, D.A., Gibbs, J.P., 2004. Effects of roads on the structure of freshwater turtle populations. *Conserv. Biol.* 18 (4), 1143–1148.
- Swihart, R.K., Slade, N.A., 1984. Road crossing in *Sigmodon hispidus* and *Microtus ochrogaster*. *J. Mamm.* 65 (2), 357–360.
- Taylor, B.D., Goldingay, R.L., 2009. Can road-crossing structures improve population viability of an urban gliding mammal? *Ecol. Soc.* 14 (2), 13. <http://www.ecologyandsociety.org/vol14/iss2/art13/>.
- van der Ree, R., Cesarini, S., Sunnucks, P., Moore, J.L., Taylor, A., 2010. Large gaps in canopy reduce road crossing by a gliding mammal. *Ecol. Soc.* 15 (4), 35. [online] URL: <http://www.ecologyandsociety.org/vol15/iss4/art35/>.
- van der Zee, F.F., Wiertz, J., Ter Braak, C.J. F., van Apeldoorn, R.C., 1992. Landscape change as a possible cause of the badger *Meles meles* L. decline in The Netherlands. *Biol. Conserv.* 61, 17–22.

Further Reading

- Andrews, K.M., Nanjappa, P., Riley, S.P.D., 2015. *Roads and Ecological Infrastructure: Concepts and Applications for Small Animals*. Wildlife Management and Conservation. Published in association with The Wildlife Society by Johns Hopkins University Press, Baltimore (Maryland).
- Beckmann, J.P., Clevenger, A.P., Huijser, M.P., Hilty, J.A. (Eds.), 2010. *Safe Passages: Highways, Wildlife, And Habitat Connectivity*, Island Press, Washington, D.C., pp. 383.
- Forman, R.T.T., 2000. Estimate of the area affected ecologically by the road system in the United States. *Conserv. Biol.* 14 (1), 31–35.
- Forman, R.T.T., Sperling, D., Bissonette, J.A., Clevenger, A.P., Cutshall, C.D., Dale, V.H., Fahrig, L., France, R., Goldman, C.R., Heanue, K., Jones, J.A., Swanson, F.J., Turrentine, T., Winter, T.C., 2003. *Road Ecology: Science And Solutions*, Island Press, Covelo, CA.
- Forest Service Stream Simulation Working Group U.S., 2008. *Stream Simulation: An Ecological Approach To Providing Passage For Aquatic Organisms At Road-Stream Crossings*. San Dimas, US Forest Service Technology and Development Program.
- van der Ree, R., Smith, D., Grilo, C., 2015. *Handbook of Road Ecology*. Oxford, John Wiley & Sons.

Environmental Justice, Transport Justice, and Mobility Justice

Devajyoti Deka, Alan M. Voorhees Transportation Center, Edward J. Bloustein School of Planning and Public Policy, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	305
Before EJ, TJ, and MJ	305
Environmental Justice	306
Transport Justice (TJ)	307
Mobility Justice (MJ)	308
Conclusion	309
Biography	309
References	310
Further Reading	310

Introduction

Environmental justice (EJ), transport justice (TJ), and mobility justice (MJ) all evolved with the intent to benefit the transportation disadvantaged. There are vast differences between the concepts despite the tendency among some researchers to treat them as synonymous. EJ entered mainstream US transportation planning in the mid-1990s. On the other hand, the earliest writings on TJ and MJ can be traced back to the mid-2000s, but they have gained increased attention in more recent times, especially since 2016. It is not as if the transportation disadvantaged did not get attention from transportation agencies and researchers before the emergence of the concepts of EJ, TJ, and MJ. US researchers and agencies have expressed concerns about the transportation disadvantaged at least since the 1960s, when the Civil Rights Act and the first Urban Mass Transportation Act were passed and the initial writings on spatial mismatch appeared in response to the job access problem of minority populations. After a brief introduction to the research on the transportation disadvantaged from the 1960s to the 1980s, this article discusses the circumstances in which EJ, TJ, and MJ emerged, the similarities and differences between the concepts, and the manner in which each concept relates to transportation planning. Although justice in transportation is important around the world, the general context of this article is the US because that is where the earliest of the three concepts, EJ, evolved and TJ and MJ were founded, at least partially, as a critique of US regional transportation planning practices.

Before EJ, TJ, and MJ

One of the earliest studies to use the term “transportation disadvantaged” was [Altshuler \(1969\)](#), where only the poor and people with disabilities were considered deserving of receiving public transit subsidies because of their transportation disadvantages, namely, lack of affordability for the poor and bodily limitations for the people with disabilities. In a synthesis report by the Transportation Research Board ([TRB, 1976](#)), the term was used to describe people with disabilities, people in poverty, as well as older adults. These groups were identified as transportation disadvantaged because they could not travel with as much ease as other people. Recognizing that different people used the term to describe different population groups, including the poor, racial minorities, people with disabilities, children, older adults, and women, [McGuire \(1976\)](#) vehemently argued that people should not be considered transportation disadvantaged merely because of their race or ethnicity and suggested that the term be used to describe only people in extreme poverty or extreme disability. A number of studies in the 1970s made reference to the transportation disadvantaged ([Paaswell and Recker, 1974](#); Schnell, [Falcocchio and Cantilli, 1974](#)), but unlike [McGuire \(1976\)](#), they refrained from questioning who should and who should not be considered transportation disadvantaged.

[Altshuler et al. \(1979\)](#) contended that transportation disadvantage became a serious issue because of the increasing popularity of the automobile. That is because the greater use of the automobile was accompanied by a substantial reduction of public transit service, which decreased transportation options for people who could not afford or use an automobile. Thus, it was one of the earliest studies to point the finger at the automobile as the cause of transportation disadvantage—a tradition that has continued until the present day in the domains of TJ and MJ. Today, carless households are often considered as transportation disadvantaged, and for better or worse, public transit has been viewed by many as the savior of the transportation disadvantaged.

It is important to note the social and political environment of the period in which transportation disadvantage became a concern in the United States. The first Urban Mass Transportation Act was passed in 1964, which introduced federal subsidies for public transit ([Weiner, 2008](#)). Naturally, it prompted transportation professionals and researchers to ask who should be eligible to receive transit subsidies. Much of the early literature on the transportation disadvantaged, therefore, revolved around transit subsidies. The

Civil Rights Act was passed in 1964, which prohibited discrimination based on race, color, religion, sex, or national origin. This piece of legislation would have a profound effect on transportation agencies in the following decades. Another legislation to affect transportation policies in the United States was the National Environmental Policy Act of 1969, which continues to be a cornerstone for EJ today. The next federal law to have a profound effect on the transportation disadvantaged was the Americans with Disabilities Act of 1990, which prohibited discrimination on the basis of disability in the public as well as the private sector (Weiner, 2008). The law, as amended, required public transit agencies to implement measures to make stations and vehicles accessible to all people, including people with disabilities. The law also required transit agencies to provide paratransit service to people with disabilities who could not use fixed-route transit.

A concept from pre-EJ times that had a substantial impact on the thinking of researchers in the domain of the transportation disadvantaged is spatial mismatch. The concept is often tied to Kain's (1968) seminal study that focused on difficulties in job access for minority workers because of housing market discrimination. The concept evolved over the years through the works of Holzer (1991), Ihlanfeldt and Sjoquist (1998), and many others, bringing attention to the job access problem for low-skilled, low-income, and minority workers and, at the same time, adding the dimension of space to discussions on the transportation disadvantaged. Although the concept of spatial mismatch is not only about transportation, it helped to bring attention to the transportation problems of low-income and minority populations, especially in regard to their access to suburban job locations. In a sense, the spatial mismatch concept convinced many that the role of transportation planning is not merely to improve transportation, but also to improve economic well-being of low-income and minority populations.

Environmental Justice

Environmental justice advocacy, often referred to as a movement, began in the late 1970s and the term EJ most likely was used for the first time in the early 1980s in North Carolina in a fight against the dumping of environmental waste (Bullard and Johnson, 1997; Martinez-Alier et al., 2014). Since that time, the primary concern for EJ advocates, typically belonging to grassroots organizations, has been that racial minorities and low-income populations disproportionately suffer from pollution and other adverse effects of transportation. By the late 1990s, EJ advocacy had spread throughout the US as grassroots organizations with varied interests coalesced against transit agencies and highway authorities in many cities for unfair distribution of transportation benefits and burdens. Since that time, EJ advocacy has influenced diverse types of transportation decisions, including but not limited to tolling, congestion pricing, gasoline taxation, transit routing, transit fare, noise barriers, air pollution, and bikeshare programs.

EJ gained a tremendous boost in 1994, when President Clinton signed Presidential Executive Order No. 12898, titled "Federal Action to Address Environmental Justice in Minority Populations and Low-Income Populations." EJ entered mainstream transportation planning in 1997, when the US Department of Transportation (USDOT) issued its own order (DOT Order No. 5610.2) to make transportation policy and planning consistent with the Presidential Executive Order. The foundation for EJ is provided by Title VI of the Civil Rights Act and the National Environmental Policy Act (NEPA); the first providing the non-discrimination basis and the latter providing the implementation process. Presidential Executive Order No. 13166, titled "Improving Access to Services for Persons with Limited English Proficiency," signed by President Bush in 2000, added people with limited English proficiency (LEP) as another disadvantaged group requiring special attention in transportation and other public spheres.

The Federal Highway Administration (FHWA, 2015) and the Federal Transit Administration (FTA, 2012) continue to issue circulars to assist transportation agencies to comply with the USDOT's Title VI and EJ programs. The objective of EJ program is to ensure that the adverse effects of transportation policies, programs, and actions do not disproportionately affect minority populations and low-income populations and that the benefits and burdens of transportation are equitably distributed among all people. The Title VI program, consisting of both Title VI and EJ components, is concerned about non-discrimination on the basis of race, color, national origin, sex, age, disability, low-income, and limited English proficiency.

The USDOT (FHWA, 2015, p.11) defines minority and low-income population in the EJ context "as any readily identifiable group of minority and/or low-income persons who live in geographic proximity, and, if circumstances warrant, geographically dispersed/transient persons of those groups (such as migrant workers, homeless persons, or Native Americans) who will be similarly affected by a proposed FHWA/DOT program, policy, or activity." The terms "underserved population" and "traditionally underserved population" are often used to describe low-income and minority populations, but other transportation disadvantaged populations, such as children, older adults, and persons with disabilities are also considered underserved. Thus, underserved populations today are fairly similar to the transportation disadvantaged before EJ entered the transportation lexicon.

In the United States, metropolitan planning organizations (MPO) are entrusted with preparing regional transportation plans (FHWA/FTA, 2004). Regional transportation planning is a continuous process, beginning with the setting of regional goals and visions, followed by needs assessment, strategy identification, strategy evaluation, development of the plan document, development of the transportation improvement program, project development, and system operation/maintenance. EJ enters the planning process at many levels; while it is an essential part of public involvement and the NEPA process, many MPOs also undertake sophisticated data analysis to identify low-income and minority communities to ensure that the benefits and burdens of transportation programs and projects are equitably distributed. As MPOs typically work under resource constraints, the attention received by EJ varies from agency to agency. While some MPOs may make substantial efforts to go above and beyond what is required by laws and regulations, others can afford to make only the bare minimum effort to comply with the laws and regulations. EJ remains somewhat ad hoc in the regional planning process mainly because capabilities of MPOs vary substantially.

Thus, EJ is a concept that emerged in the late 1970s and early 1980s through grassroots advocacy that led to many successes, including new federal policies. After the EJ Executive Order was signed in 1994, EJ has become a formal part of transportation planning in the US. Unlike the concepts to be discussed in the following sections, EJ has a regulatory framework within which MPOs perform their activities. In contrast, TJ and MJ, at this time at least, are merely interesting and insightful concepts.

Transport Justice (TJ)

The term transport justice was popularized primarily by the book “Transport Justice: Designing Fair Transportation Systems” (Martens, 2017), but its normative approach can be traced back to earlier works of Martens and his peers (Martens, 2006; Martens, 2012; Martens et al., 2012). Although some authors (e.g., Beiler and Mohammed, 2016) have treated concepts of EJ and TJ as interchangeable, they are perhaps more dissimilar than similar when one enters the domain of justice in transportation.

The concept of TJ emerged as a critique of conventional transportation planning, especially in the United States, where regional transportation models are used to predict the impact of population and employment growth on network traffic and recommendations for improvement are made on the basis of the model results. The TJ proponents argue that this “predict-and-provide” practice, in the US and elsewhere, disproportionately benefits car users because more car use generates more congestion and more congestion leads to more highway investments benefiting only those who can afford to use cars. This practice benefits people who travel more by car and adversely affects people who cannot afford cars or travel less by cars. Another criticism of conventional transportation planning from the proponents of TJ focuses on the use of cost-benefit analysis (CBA) to make transportation investment decisions. They argue that the use of travelers’ income as an input in CBA unduly favors the wealthy.

In contrast to EJ, which evolved in the sphere of grassroots advocacy in the real world, the concept of TJ emerged as a justice-theoretic approach and to this day it has remained mostly theoretical. The proponents draw primarily from the writings of three social thinkers of the 20th century: Rawls (1971), Walzer (1983), and Dworkin (1981a, 1981b). On the basis of Rawls’s decision making under the veil of ignorance, the proponents argue that prudent people would agree to the distribution of social goods that maximize the welfare of the worst-off in transportation (i.e., the maximin principle). On the basis of Walzer, the proponents claim that transportation can be considered an independent sphere of social good. Using Dworkin’s insurance schemes against “option luck” and “brute luck,” the TJ proponents argue that society would prefer a distribution system that guarantees a sufficient level of social goods to everyone despite there being wide variations in welfare among people. Putting together all of the above, the TJ proponents suggest that there are two domains for the distribution of the transportation good: the domain of free exchange and the domain of justice. TJ is about the latter.

TJ is concerned about the distribution of only two social goods: potential mobility and accessibility. In contrast to revealed mobility, which reflects a person’s actual travel, potential mobility reflects the capacity to travel. In contrast to revealed accessibility, potential accessibility refers to a person’s capacity to participate in activities distributed over space rather than actual participation in activities. Thus, both concepts are based on capacity rather than the actual acts of travel and activity participation. According to the TJ proponents, accessibility is a more meaningful measure of the transportation good compared to potential mobility because mobility does not capture variations in the distribution of activities over space. Equating both potential mobility and accessibility to freedom of choice, TJ proponents hold that “ultimately accessibility best reflects the social meaning of the transport good in Western societies” (Martens, 2017, p. 53). In this regard, TJ conforms to the general acceptance among today’s planners and researchers that both land use and transportation are important rather than transportation alone.

Despite the mostly theoretical approach of the TJ concept, Martens (2017, p. 174) proposed a 10-step transportation planning model that should be of interest to both transportation planners and researchers. The sequential steps are:

1. Identify population groups by residential location, mode availability, and socioeconomic status
2. Assess their potential mobility and accessibility by multiple measures
3. Identify potential mobility and accessibility sufficiency thresholds through a democratic process
4. Identify population groups with insufficient potential mobility and accessibility
5. Assess the severity of accessibility insufficiencies for population groups by using an Accessibility Fairness Index
6. Rank population groups according to their contribution to overall accessibility deficiency
7. Identify the causes of the accessibility deficiencies for different population groups
8. Identify promising interventions for reducing accessibility deficiencies
9. Assess the effects and benefits of the proposed interventions by cost-effectiveness analysis
10. Implement the selected interventions and monitor impacts.

Two quantitative estimation methods involved in the process are important to note. In the second of the 10-step process, potential mobility is to be estimated by the potential mobility index (PMI), also termed as average aerial speed. Expressed conventionally, the PMI measure suggested by Martens (2017, p. 154–155) is given by

$$PMI_i = \frac{1}{N} \sum_{j=1}^n \frac{D_{ij}}{T_{ij}}$$

where PMI_i is the potential mobility index for zone i , N is the total number of zones, D_{ij} is the aerial distance between zone i and j ,

and T_{ij} is the travel time by transportation network between zone i and zone j . Obviously, if the numerator of the measure were network distance instead of aerial distance, PMI would be a measure of level of service used in conventional transportation planning and engineering and PMI would represent average network *speed* to destinations. However, using aerial distance instead of network distance has been recommended because aerial distance represents the maximum distance. Once PMI is estimated, according to the recommended method, multiple accessibility estimates are to be obtained based on the PMI estimates considering activities such as employment, shops and stores, healthcare establishments, etc.

The other measurement of note in the TJ approach is the Accessibility Fairness Index (AFI), which is meant to denote the severity of accessibility deficiency for a region. AFI has been defined by Martens (2017, p. 160) as follows,

$$AFI_r = \frac{1}{N} \sum_{j=1}^q n_i \left[\frac{z - y_i}{z} \right]^2$$

where r stands for region r , N is the total population in region r , z is the accessibility sufficiency threshold, q is the number of groups in region r experiencing accessibility levels below sufficiency threshold z , n_i is the population size of the i -th population group below the sufficiency threshold z , and y_i is the amount of accessibility shortfall experienced by population group i . A higher AFI score for a region means greater accessibility deficiency. The equation suggests, all else being equal, the larger the regional population size (N), the smaller will be the AFI (i.e., accessibility deficiency); the larger the size of each population group below the sufficiency threshold (n), the greater will be the AFI; the larger the amount of accessibility experienced by people below the accessibility threshold (y_i), the higher will be the AFI; the higher the accessibility threshold (z), the greater will be the AFI. Martens (2017) estimates AFI for the City Region of Amsterdam, consisting of 16 municipalities, by focusing on employment access and two travel modes: car and public transit.

Although not a quantitative measure similar to PMI and AFI, another methodological contribution from the TJ proponents could be of interest to transportation planners interested in the domain of justice. That method, presented in Martens et al. (2012), recommends considering accessibility gaps between those with the highest accessibility and those with the lowest accessibility in making transportation investment decisions. In this case, the TJ proponents deviate from the typical Rawlsian maximin principle and suggests that transportation investments should seek to minimize the gap between the highest and lowest accessibility levels instead of merely maximizing accessibility for those with the lowest accessibility.

On the whole, TJ has sought to add a justice-theoretic dimension to transportation that is missing in EJ or in transportation research prior to EJ. Unlike EJ, which evolved to address existing injustices, TJ uses a theoretical framework in an attempt to establish a long-lasting equilibrium that would be considered fair by all people. For that, the TJ proponents have relied heavily on Rawlsian liberal philosophy.

Despite both EJ and TJ seeking justice in transportation, there are several notable differences between the two. First, EJ evolved in the domain of grassroots advocacy, whereas TJ evolved in academia. EJ has evolved sufficiently to demand and obtain a regulatory framework in at least one country, the US, but TJ does not have such a framework in the US or any other country. Third, both benefits and burdens are important considerations for EJ, but only benefits, especially potential mobility and accessibility, are important for TJ. Fourth, TJ excludes sustainability issues such as air pollution from transportation and focuses entirely on transportation benefits. In contrast, EJ emerged primarily to address pollution and it today it addresses both benefits and burdens of transportation. Fifth, racial and ethnic considerations are highly important for EJ, but such considerations are beyond the scope of TJ. Sixth, while non-discrimination is the primary mechanism by which justice is sought in EJ, it has no role in TJ. Finally, TJ seeks to establish a stable justice equilibrium that would be acceptable to all, but EJ makes not such effort and instead focuses on transportation agencies being compliant to rules and regulations.

Mobility Justice (MJ)

Although both TJ and MJ emerged in the same timeframe to address issues in transportation justice, they are vastly different concepts when looked at through the lens of a transportation planner or researcher. If TJ encompasses only a minute sphere of transportation, MJ covers all of transportation, and beyond. Many other authors have contributed to the emerging body of literature on MJ, but sociologist Sheller's (2018a) "*Mobility Justice: The Politics of Movement in an Age of Extremes*" is undoubtedly one of the most influential pieces on the subject. Despite MJ's growing popularity in recent years, Sheller herself notes that the concept began to take shape with Sheller and Urry (2006). Sheller's other works, such as Sheller (2018b) and Sheller (2018c), as well the works of other authors (e.g., Everuss, 2019; Cook and Butz, 2019) provide additional insights. Considering the much wider scope of MJ, the discussion below is restricted to issues related to transportation only.

In contrast to typical transportation planning, where "mobility" is defined as the ability to travel or actual travel, in MJ studies it means all movements, including migration, as well as immobility. Proponents claim that MJ can help to build political alliances to effectively address both short- and long-term problems of mobility and immobility. They justify the need for the MJ paradigm as a means to address three contemporary crises faced by the world: (1) the climate crisis, (2) the urbanization crisis, and (3) the refugee crisis. Among the three, the last, obviously, has the least to do with the current practice of transportation planning because planners mostly focus on intra-urban or intra-metropolitan transportation of people and goods. "Mobility justice is not just about transportation within cities, though that is an important part of it, but also about smaller micro-mobilities at the bodily scale that are

inflected by racial and classed processes, gendered practices, and social shaping of disabilities and sexualities” (Sheller, 2018a, p. 2). The proponents argue that both mobility and immobility are important to MJ. They recognize that mobility is not always a desired outcome because sometimes people have to travel or migrate because of coercion or housing affordability issues. Rights are important to MJ and it stands for rights to mobility as much as it stands for rights to immobility.

MJ is based on insights from EJ, TJ, and social justice. Like TJ, MJ supports the establishment of a minimum threshold of accessibility through a deliberative process, but at the same time, it is also critical of TJ for its narrow focus on accessibility. Distributive justice—the central focus of TJ—is only one of a variety of justice considerations for MJ, for it also includes deliberative justice, procedural justice, epistemic justice, and restorative justice. Deliberative justice demands not only the right to deliberate, but also the recognition of the vulnerabilities of population groups. Procedural justice demands information and meaningful participation of all populations in the process of governance. Epistemic justice demands proactive knowledge production and the bridging of knowledge gaps. Finally, restorative justice demands reparations to those harmed by transportation and environmental pollution. Thus, MJ demands not only distributive justice in transportation, but also a meaningful and informative participatory process that results in actual distribution of compensation to those harmed by transportation projects and investments.

Despite their differences, TJ and MJ also have certain similarities. First, both TJ and MJ originated from a disenchantment with status quo transportation planning practices, especially its excessive emphasis on automobile travel. Second, both view roadway congestion as a serious problem caused by the automobile culture. Third, both disdain cost-benefit analysis as a technique for measuring transportation benefits. Fourth, both emphasize the importance of space. Finally, both seek to improve the welfare of the less privileged despite having different ideas about who the less privileged are.

However, the differences between TJ and MJ are starker than their similarities. First, TJ does not consider any specific population group other than those with high accessibility and low accessibility, whereas MJ makes reference to specific population groups in terms of demographic, economic, social, and cultural characteristics. Second, TJ does not seriously consider any doctrine of justice other than Rawlsian justice (e.g., utilitarianism, libertarianism, egalitarianism, etc.), whereas MJ provides an assessment of each doctrine. Third, TJ recommends quantitative methods for application in transportation planning, but MJ does not make such efforts. Fourth, TJ is barely concerned about the role of any stakeholders in the transportation planning process, whereas MJ emphasizes the important role of community organizations and recognizes the role of government. Fifth, TJ seeks a stable reflective equilibrium in the Rawlsian way, whereas MJ seeks both short-term and long-term solutions in a dynamic world. Finally, and perhaps most importantly, TJ avoids any discussion of the distribution of the adverse effects of transportation, whereas MJ places a far greater emphasis on the correction of the adverse effects of transportation than the distribution of transportation benefits.

Conclusion

This article summarized the concepts of EJ, TJ, and MJ as related to transportation planning. It demonstrated that there are many differences between the three concepts despite having some similarities. As of now, EJ has a regulatory framework in the US, whereas TJ and MJ are merely two conceptual approaches to address transportation problems of disadvantage populations. Although TJ and MJ may appear esoteric to transportation professionals who contribute to actual plans using technical professional skills, they should investigate how these new concepts can benefit the planning process and outcomes. Transportation researchers who have the understanding of both research and practice can play a meaningful role in the integration of TJ and MJ into transportation planning.

The transportation world is a dynamic one. While most of the concerns in current EJ, TJ, and MJ literature pertain to conventional cars and public transportation, the last decade has seen substantial improvements in newer vehicle technologies (e.g., automated vehicles and electric vehicles), app-based transportation (i.e., ridesourcing), and transit service delivery models (e.g., microtransit, subsidized ridesourcing, scheduling technology, etc.). At the present time, there is significant excitement among transportation researchers about the “Three Revolutions” in transportation, consisting of electric vehicles, automated vehicles, and ridesourcing (Sperling, 2018). There is speculation among researchers that the integration of these new technologies will enable the general population to forego car ownership and use as shared mobility services provided by private companies in automated vehicles will become ubiquitous and conventional public transit will remain available only in the densest urban corridors. The justice implications of such changes cannot be known at the present time. However, given the history of injustices related to the conventional automobile, there is reason to be vigilant about potential disparate (positive and negative) impacts of new automobile technologies and service delivery models on different populations. Transportation researchers and planners concerned about justice cannot afford to lose sight on the new developments in transport technologies because they have the potential to affect not only the future of cities and transportation, but also the livelihood of future generations.

Biography

Devajyoti Deka MA (Economics), MRP, MUP, PhD (Planning), is the assistant director for research at the Alan M. Voorhees Transportation Center of Rutgers, The State University of New Jersey, New Brunswick. His research at the center predominantly addresses the transportation disadvantaged, including people with disabilities, older adults, low-income and minority populations, and children. Between 2003 and 2008, he worked as the manager for systems planning and modeling at the North Jersey

Transportation Planning Authority, where he supervised the agency's environmental justice program and managed projects pertaining to regional capital investment strategy and transportation strategy evaluation.

References

- Altshuler, A., 1969. Transit subsidies: by whom, for whom? *J. Am. Inst. Plan.* 35 (2), 84–89, doi:10.1080/01944366908977578.
- Altshuler, A., Womack, J.P., Pucher, J.R., 1979. *The Urban Transportation System: Politics and Policy Innovation*. MIT Press, Cambridge, M.A.
- Beiler, M.O., Mohammed, M., 2016. Exploring transportation equity: Development and application of a transportation justice framework. *Transp. Res. Part D: Transp. Environ.* 47, 285–298, doi:10.1016/j.trd.2016.06.007.
- Bullard, R.D., Johnson, G.S., 1997. Just transportation. In: Bullard, R.D., Johnson, G.S. (Eds.), *Just Transportation: Dismantling Race and Class Barriers to Mobility*, New Society Publishers, Gabriola Island B. C., pp. 7–21.
- Cook, N., Butz, D. (Eds.), 2019. *Mobilities, Mobility Justice and Social Justice*. Routledge, London.
- Dworkin, R., 1981a. What is equality? Part 1: Equality of welfare. *Philos. Public Affairs* 10 (3), 185–246.
- Dworkin, R., 1981b. What is equality? Part 2: Equality of resources. *Philos. Public Affairs* 10 (4), 283–345.
- Everuss, L., 2019. Mobility Justice: a new means to examine and influence the politics of mobility. *Appl. Mobil.* 4 (1), 132–137, doi:10.1080/23800127.2019.1576489.
- Falcocchio, J.C., Cantilli, E.J., 1974. *Transportation and the disadvantaged: The Poor, the Young, the Elderly, the Handicapped*. Lexington Books, Lexington, M.A.
- FHWA (Federal Highway Administration), 2015. *Federal Highway Administration Environmental Justice Reference Guide*. US Department of Transportation, Washington, D.C. Available from: https://www.fhwa.dot.gov/environment/environmental_justice/publications/reference_guide_2015/. Accessed date: 13 June 2020.
- FHWA/FTA (Federal Highway Administration and Federal Transit Administration), 2004. *The Metropolitan Transportation Planning Process: Key Issues*. Report No. FHWA-EP-03-041 (5/04). Federal Highway Administration and Federal Transit Administration, US Department of Transportation, Washington, D.C.
- FTA (Federal Transit Administration), 2012. *Environmental Justice Policy Guidance for Federal Transit Administration Recipients*. FTA C 4703. US Department of Transportation, Washington, D.C. Available from: https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/FTA_EJ_Circular_7.14-12_FINAL.pdf. Accessed date: 13 June 2020.
- Holzer, H.J., 1991. The spatial mismatch hypothesis: What has the evidence shown? *Urban Stud.* 28 (1), 105–122, doi:10.1080/00420989120080071.
- Ihlanfeldt, K.R., Sjoquist, D.L., 1998. The spatial mismatch hypothesis: A review of recent studies and their implications for welfare reform. *Housing Policy Debate* 9 (4), 849–892., doi:10.1080/10511482.1998.9521321.
- Kain, J.F., 1968. Housing segregation, negro employment, and metropolitan decentralization. *Quart. J. Econ.* 82 (2), 175–197, doi:10.2307/1885893.
- Martens, K., 2006. Grounding transport planning on principles of social justice. *Berkley Plan. J.* 19, 1–17, doi:10.5070/BP319111486.
- Martens, K., 2012. Justice in transport as justice in accessibility: Applying Walzer's 'Spheres of Justice' to the transport sector. *Transportation* 39 (6), 1035–1053., doi:10.1007/s11116-012-9388-7.
- Martens, K., 2017. *Transport Justice: Designing Fair Transportation Systems*. Routledge, New York.
- Martens, K., Golub, A., Robinson, G., 2012. A justice-theoretic approach to the distribution of transportation benefits: Implications for transportation planning practice in the United States. *Transp. Res. Part A: Policy Prac.* 46 (4), 684–695, doi:10.1016/j.tra.2012.01.004.
- Martinez-Alier, J., Anguelovski, I., Bond, P., et al., 2014. Between activism and science: grassroots concepts for sustainability coined by Environmental Justice Organizations. *J. Polit. Ecol.* 21, 19–60, doi:10.2458/v21i1.21124.
- McGuire, C., 1976. Who are the Transportation Disadvantaged? Report No. DOT-BIP-WP-27-10-77. Metropolitan Transportation Commission, Berkeley; Department of Housing and Urban Development, Washington, D.C.; and US Department of Transportation, Washington, D.C.
- Paaswell, R.E., Recker, W.W., 1974. Location of the carless. *Transp. Res. Rec.* 516, 11–20.
- Rawls, J., 1971. *A Theory of Justice*. Harvard University Press, Cambridge, M.A.
- Sheller, M., 2018b. Racialized mobility transitions in Philadelphia: Connecting urban sustainability and transport justice. *City Soc.* 27 (1), 70–91, doi:10.1111/ciso.12049.
- Sheller, M., 2018a. *Mobility Justice: The Politics of Movement in an Age of Extremes*. Verso, London.
- Sheller, M., 2018c. Theorising mobility justice. *Tempo Soc.* 30 (2), 17–34, doi:10.11606/0103-2070.ts.2018.142763.
- Sheller, M., Urry, J., 2006. The new mobilities paradigm. *Environ. Plan. A* 38 (2), 207–226., doi:10.1068/a37268.
- Sperling, D. (Ed.), 2018. *Three Revolutions: Steering Automated, Shared, and Electric Vehicles to a Better Future*. Island Press, Washington, D.C.
- TRB (Transportation Research Board), 1976. *Transportation Requirements for the Handicapped, Elderly, and Economically Disadvantaged*. National Cooperative Highway Research Program: Synthesis of Highway Practice 39. Transportation Research Board, National Research Council, Washington, D.C.
- Walzer, M., 1983. *Spheres of Justice: A Defense of Pluralism and Equality*. Basic Books, New York, N.Y.
- Weiner, W., 2008. *Urban Transportation Planning in the United States*, 3rd ed. Springer Science and Business, New York, N.Y.

Further Reading

- Sen, A., 2009. *The Idea of Justice*. Harvard University Press, Cambridge, M.A.
- Rawls, J., 2001. *Justice as Fairness*. Harvard University Press, Cambridge, M.A.
- Holifield, R., Chakraborty, J., Walker, G. (Eds.), 2018. *The Routledge Handbook of Environmental Justice*. Routledge, New York, N.Y.

Transport Noise and Health

Elisabete F. Freitas*, Emanuel A. Sousa†, Carlos C. Silva†, *University of Minho, Campus de Azurém, Guimarães, Portugal; †Center for Computer Graphics, Campus de Azurém, Guimarães, Portugal

© 2021 Elsevier Ltd. All rights reserved.

Introduction	311
Relevant Concepts	311
Emission and Noise Sources	312
Continuous, Intermittent, and Impulsive Noise	312
Noise Annoyance	312
Noise Exposure–Response Relation (ERR) Curves	312
Environmental Noise Management	312
Response Mechanism to Noise Exposure	312
Short-Term Effects of Noise on Health	312
Short-Term Annoyance	312
Sleep Disturbance	313
Long-Term Effects of Noise on Health	315
Cardiovascular Disease	315
Effects on Cognitive Development	315
Mental Health	315
Other Long-Term Effects	316
Cancer	316
Diabetes	316
Obesity	316
Pregnancy	316
Future Research	316
References	317

Introduction

The 20th century brought about a profound transformation of the way humans travel. After centuries of animal-powered locomotion, the invention of the electric and combustion engines has fundamentally changed the way humans perceived distance and conceived commercial and labor relationships. A by-product of this transformation was the change in the way human settlements and activities sound. As a matter of fact, it can be argued that, in terms of acoustic landscape (also referred to as soundscape), the 20th century is characterized by the sound of the road, railway, maritime, and airborne motorized transportation (Schafer, 1997). A harmful subproduct of this transformation is that excessive noise originating from transport vehicles has become a pervasive health hazard, especially in urban settings.

Transportation noise has been listed as a major environmental risk factor in Europe, only supplanted by air pollution (Hänninen et al., 2014). The World Health Organization (WHO) acknowledges environmental noise (dominated by transportation) as an important public health problem (WHO, 2018) and estimates that in Europe it accounts for a disability-adjusted life-year (DALY) loss of 61,000 years for ischemic heart disease (IHD), 45,000 years for cognitive impairment in children, 903,000 years for sleep disturbance, 22,000 years for tinnitus and 654,000 years for annoyance (WHO Regional Office for Europe & JRC, 2011). Therefore, despite all the benefits that came along with increasing mobility, one cannot ignore the significant burden on the health of those who are more exposed to the sound of it.

This chapter explores state-of-the-art research on the health impacts of transportation noise. It begins by defining the relevant concepts and main determinants of the transportation noise problem, following with a presentation of short- and long-term effects of noise on health. General effects of transportation noise on public health are also addressed, concluding with indications of future research.

Relevant Concepts

To facilitate the comprehension of the following sections, a set of crucial concepts related to the transportation noise problem are defined next.

Emission and Noise Sources

According to ISO 3740:2019 (ISO, 2019), emission is defined as airborne sound radiated by a well-defined noise source (e.g., the machine under test) under specified operating and mounting conditions. The basic noise emission quantities are the sound power level of the source itself and the emission sound pressure levels at the vicinity of the source. Road, rail, and aircraft transportation account for most of the transportation noise and have been the focus of most studies on specific noise sources (Lercher, 2011). The range of noise levels generated by these sources is wide. It depends on the internal components of the vehicle—such as the engine, intake system, exhaust system, fans, and structural vibration—and the vehicle–medium interaction—namely, tire–road and wheel–rail contact, water movement, and air turbulence. For road and railway traffic, rolling noise is usually the dominating noise source under typical operational conditions. While for the first the sound is generated mainly by tires–pavement interaction, for the second, it is the combined roughness of wheel and rail. Noise from aircraft is primarily caused by the engines. During landing, where engines are often in idle, airframe noise can also become relevant (Sørensen et al., 2020).

Continuous, Intermittent, and Impulsive Noise

Transportation noise can be characterized by its duration and by the way its level fluctuates. A *continuous* or *steady* noise suffers negligibly small fluctuations of level (<5 dB) within a period of observation—for example, the sound of slow-moving traffic in a busy city center; an *intermittent* noise persists for more than 1 s and then abruptly drops to the level of the background noise, showing this pattern several times during the period of observation—for example, the sound of a train passing over the rail joints and squats; an *impulsive* noise consists in a series of bursts of sound energy, each having a duration of less than 1 s and representing a change of sound pressure level of 40 dB or more within 0.5 s (Rosati and Jamesdaniel, 2020)—for example, a road vehicle rolling over rumble strips.

Noise Annoyance

The most widespread and well documented subjective response to noise is “annoyance,” which is a multidimensional negative reaction that covers: (1) immediate behavioral noise effects aspects, such as “disturbance” and “interference in ongoing activities,” and (2) evaluative aspects such as “nuisance,” “unpleasantness,” and “getting on one’s nerves” (Guski et al., 1999). It may also include feelings of fear and mild anger related to a belief that one is being avoidably harmed (Stansfeld and Matheson, 2003).

Noise Exposure–Response Relation (ERR) Curves

It is the strength of association between exposure to environmental noise and long-term noise annoyance, as demonstrated by the percentage of highly annoyed residents (%HA) as a function of noise level (usually expressed in L_{den} , as discussed in a recent meta-analysis by (Guski et al., 2017)). The exposure–response relation can be different for different noise sources and combinations of noise sources. The current literature shows steeper increments in ERR curves for exposure to air transportation, followed by exposure to railway transportation, and with the lower rate of increment in %HA being found in exposure to road traffic (Guski et al., 2017).

Environmental Noise Management

Also designated by *environmental noise control*, according to Brown and Van Kamp (2015) includes technical interventions that embrace reduction of levels at the source (e.g., low noise road pavements and vehicle engines), positioning of outdoor barriers between source and receivers, and changes in the acoustic properties of building envelopes to reduce levels at receivers.

Response Mechanism to Noise Exposure

Noise exposure may be associated with stress responses, characterized by activation of the neuroendocrine system (i.e., hypothalamus–pituitary–adrenal (HPA) axis and sympathetic–adrenal–medulla axis), and a subsequent release of stress hormones (Cantuaria et al., 2018). For example, associations between exposure to transportation noise and stress-induced secretion of glucocorticoids, mostly cortisol in children and adults, were empirically demonstrated. In newborn infants, it is suspected that road traffic noise exposure may act as a potential environmental stressor in early postnatal life, which might have a possible physiological relevance later in life (Cantuaria et al., 2018).

There is a multitude of possible effects of transportation noise on health. They have mainly been grouped by the duration of symptomatology as short and long-term effects (Porter et al., 2000). Short-term effects are the result of annoyance expressed during one or few hours, while long-term effects or chronic annoyance designates a feeling that has been felt over months and years (Bartels et al., 2015). In the next sections, the distinction between short and long-term effects of traffic noise will be clarified by describing the most common health consequences of each.

Short-Term Effects of Noise on Health

Short-Term Annoyance

The annoyance response is seen by many researchers as a stress-reaction involving an environmental threat and individual physiological, emotional, cognitive, and behavioral responses, which can partly be remembered and integrated into a verbal long-term annoyance response (Guski et al., 2017).

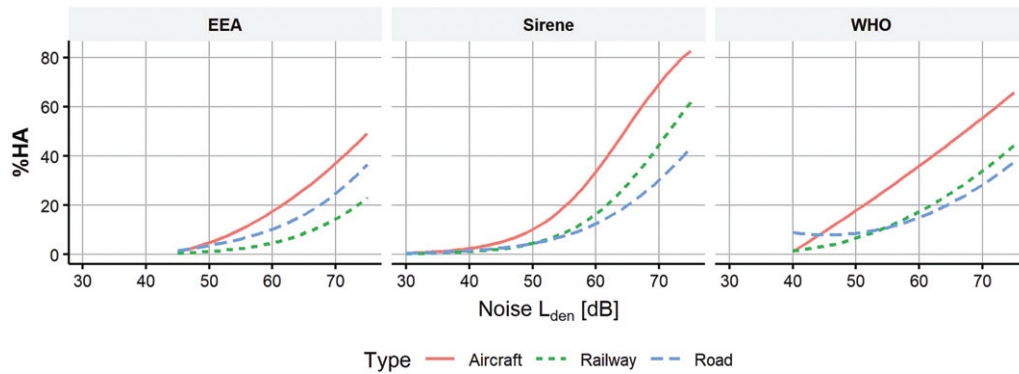


Figure 1 Examples of noise-response curves from the European Environment Agency (EEA) (Babisch et al., 2010), the Sirene project (Brink et al., 2019), and WHO (World Health Organization, 2018).

Different time dimensions of annoyance can be distinguished. Short-term annoyance refers to a response regarding noise exposure during a limited period, for instance, the previous night. Long-term annoyance describes a general feeling that has evolved over longer periods, of weeks, months, or even years. Long-term annoyance is considered to result from the accumulation of acute physiological responses to a noise event, and short-term to reactions in the day after a disturbed night's sleep, that is, acute responses resulting from awakenings in the previous night and possibly perceived fatigue and cognitive performance decrements. According to Porter et al. (2000), all levels of annoyance have the same causes and share the same characteristics despite their different orders of time. Also, there is some evidence that short-term annoyance is related to long-term annoyance.

The prediction of annoyance has been based on exposure-response curves. In general, for each traffic noise, these curves describe an increase in annoyance with rising exposure (Fig. 1). The European Environment Agency recommends those provided by the EU-position paper on dose-effect relations (Babisch et al., 2010) to predict the annoyance from road traffic, aircraft, and railways (separately for each noise source). Also, the WHO recommended in the 2018 guidelines (World Health Organization, 2018) the curves proposed in Guski et al. (2017). Recent studies indicate that people react more strongly to a given noise exposure level today than in the past, as it was shown in the Sirene project (Brink et al., 2019; Foraster et al., 2019).

Average noise levels are often selected as an acoustical measure to predict noise annoyance, but noise affected persons do not only react to a global noise exposure, characterized by the average noise level, but also to the characteristics of noise events, that is, to number, distribution, duration, level, and meaning of noise exposure. For example, Quehl et al. (2017) showed the importance of the number of overflights for the prediction of short-term annoyance due to nocturnal aircraft noise. Other approaches based on psychoacoustical indexes have also been explored to explain short-term annoyance. They have been used, for example, to characterize annoying auditory sensations evoked by railway pass-by noises in urban areas (Catherine et al., 2018) or to develop short-term annoyance models as Chun et al. (2018) did for combined aircraft and road traffic noises.

Sleep Disturbance

Sleep disturbance refers to internal/external interference with sleep onset or sleep continuity (Basner and McGuire, 2018). Undisturbed sleep is a prerequisite for high daytime performance, wellbeing, and health. Noise is a common cause of sleep disturbance because the sleeping brain retains its ability to process incoming acoustic stimuli (Jafari et al., 2018). The auditory system has a watchman function and constantly scans the environment for potential threats. Humans perceive, evaluate, and react to environmental sounds even while asleep. These reactions are part of an integral activation process of the organism that expresses itself, for example, as changes in sleep structure (sleep structure varies systematically over the course of the night, with deep sleep stage N3 (stage 3 or 4 according to the old classification) dominating the first half of the night and REM sleep and superficial sleep stages N1 and N2 (1 and 2) dominating the second half of the night) or increases in blood pressure and heart rate (Schmidt et al., 2013).

Noise has been shown to fragment sleep, reduce sleep continuity, and reduce total sleep time. The noise effects on sleep are well illustrated in Fig. 2. At short-term, noise-induced sleep disturbance include impaired mood, subjectively and objectively increased daytime sleepiness, and impaired cognitive performance (Basner and McGuire, 2018). Over long time-periods, health consequences such as cardiovascular and metabolic diseases may develop. Too much noise at night may also lead to chronic partial sleep loss, which plays a role in the current epidemics of obesity and diabetes (Sørensen et al., 2020).

There is evidence that physiological, unconscious reactions during sleep are different from psychological, conscious reactions during wakefulness, and cannot be predicted from annoyance surveys (Elmenhorst et al., 2019). During the night, transportation noise can often be described as intermittent (i.e., discrete noise events rather than a constant background noise level). When the organism differentiates a noise event from the background, physiologic reactions during sleep can be expected (Basner and McGuire, 2018).

Among those reactions is the awakening, which refers to events where a subject wakes up from sleep, regains consciousness, and recalls it in the next morning. There is evidence that the different modes of transportation affect awakenings in several manners. The

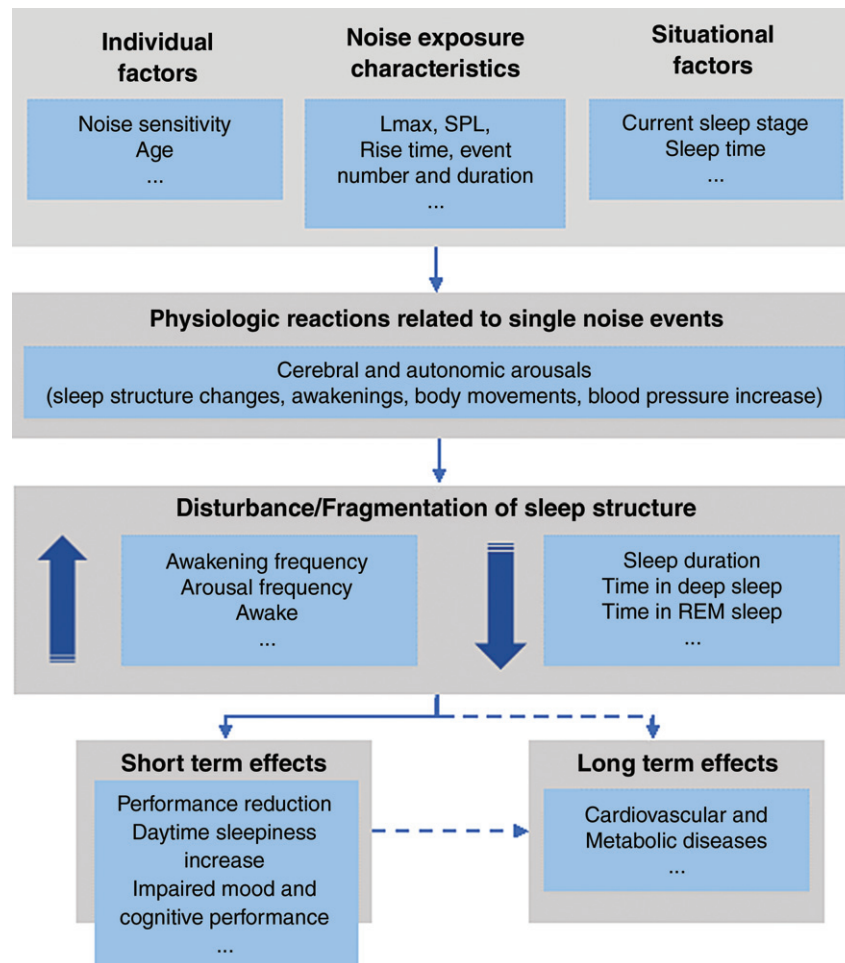


Figure 2 Effects of noise on sleep (Adapted from Basner and McGuire, 2018).

several traffic noise sources induce different awakening probabilities even at equal maximum A-weighted sound pressure level (SPL) and similar acoustical parameters, like rise time of the maximum A-weighted SPL of a noise event, number, and duration of noise events, as well as physiological parameters like current sleep stage, elapsed sleep time, and elapsed sleep time in the same sleep stage. At equal maximum A-weighted SPL the awakening probability due to the three traffic noise sources increased in the order aircraft < road < railway noise (Elmenhorst et al., 2019). This rank order of traffic modes was obtained in laboratory studies. Thus, it is different from expose-response curves obtained from self-reported studies. These studies are based on responses to questions, for example, the ones related to awakenings analyzed in (Basner and McGuire, 2018; Brink et al., 2019) or annoyance as discussed in (Lechner et al., 2019), where aircraft noise causes higher reactions, followed by railway, and then road traffic noise.

The number and acoustical properties of single noise events reflect better the actual degree of nocturnal sleep disturbance in a single night than equivalent noise levels (Basner and McGuire, 2018). As reported by Elmenhorst et al. (2019), the spectral characteristics of noise from the different traffic noise sources might explain the differences between reactions. Especially high frequencies that are filtered for aircraft noise via the atmosphere were found to explain differences in awakening probability between traffic noise sources. For railway noise, high-frequency components are more likely to induce event-related arousals and increases in heart rate than low-frequency events (Smith et al., 2019). Also, the fluctuations in freight train sounds, as well as its sharpness, were found to have an impact.

Cortical Awakenings caused by road, rail, and aircraft noise measured by polysomnography (simultaneous measurement of at least brain electrical potentials, eye movements, and muscle tone) showed a positive association between indoor maximum noise levels of single events and the probability of sleep stage transitions to wake or Stage 1 for all transportation modes (Basner and McGuire, 2018). Also, the noise levels at which the probability of an additional awakening may occur differs between transportation modes, and are between 33 and 38 dBA (Basner and McGuire, 2018).

A recent study confirmed effects from the nighttime transportation noise events on increased sleep electroencephalography arousal indices. In this study, sleep structure and continuity were not affected (Rudzik et al., 2020). Nevertheless, it was found that the amplitude of sleep spindles (which have a sleep-protective function and are relevant for memory consolidation) was consistently decreased during noise exposure compared with noise-free nights (Münzel et al., 2020).

Noise at the beginning of the night impairs the process of falling asleep but may be compensated later. In contrast, even short periods of noise toward the end of the sleeping period cause sleep disturbances. This finding is supported both by laboratory studies and observational studies on self-reported sleep quality in the morning (Münzel et al., 2020).

Long-Term Effects of Noise on Health

While short-term effects of transportation noise have been deeply studied and are well documented, there is a growing body of research concerning the effects of long-term exposure on physical and mental health. There is particularly strong evidence for a relation between noise and cardiovascular and metabolic disease, impairment of cognitive development (in children), and mental disease (Sørensen et al., 2020). More recent studies have also pointed to a relation between transportation noise and cancer (e.g., Hegewald et al., 2017) and metabolic diseases (e.g., Foraster et al., 2018).

Cardiovascular Disease

Several researchers have been studying the relation between transportation noise and cardiac disease, in both laboratory and real-life environments. Noise (particularly through sleep disruption and stress-related reactions) increases sympathetic and endocrine responses increasing on oxidative stress (Munzel et al., 2014) and inflammatory or immune responses (Münzel et al., 2017). This may lead to increases in blood viscosity, blood coagulation, and blood pressure (Lundberg, 1999), which are known risk factors for cardiac disease. Two cardiac conditions have been extensively studied concerning transportation noise: hypertension and ischemic heart disease (IHD). Regarding hypertension, several studies explored its relation with noise (Babisch, 2014; Babisch et al., 2009; Huang et al., 2015; Jarup et al., 2008; Schmidt et al., 2015; Van Kempen et al., 2006) with mixed results. Large scale studies like the ones undertaken in projects RANCH (Van Kempen et al., 2006) and HYENA (Jarup et al., 2008) have pointed to increased risk of hypertension resulting from long-term noise exposure to aircraft and road traffic. A subsequent meta-analysis found increases in the risk of hypertension from transportation noise, particularly aircraft noise (Babisch, 2014; Babisch and Van Kamp, 2009; Huang et al., 2015). A systematic review by WHO combining 37 studies found significant positive risk estimates for prevalent hypertension, for aircraft, railway and traffic noise (van Kempen et al., 2018). However, the WHO analysis was mostly comprised of cross-sectional studies and thus the quality of the evidence was ranked “very low.” More recent cohort studies (Dimakopoulou et al., 2016; Pyko et al., 2018) have also found association between noise and increased risk of hypertension.

IHD has also been consistently connected with long-term noise exposure. Meta-analysis has shown associations with road traffic (Babisch, 2008, 2014) and aircraft traffic (Vienneau et al., 2015). The systematic review by WHO evaluated 22 articles and found several associations between transport noise (aircraft, traffic, and rail) and IHD. Particularly, evidence for a relation with traffic noise was deemed as “high” quality, due to the existence of several cohort studies supporting the association.

Transportation noise may also increase the risk of other cardiac diseases such as stroke (Halonen et al., 2015; Héritier et al., 2019; Pyko et al., 2019; Seidler et al., 2018).

Effects on Cognitive Development

Transportation noise is hypothesized to have effects on cognitive development, so several research works focus on understanding the effects of noise in children at school age. Noise may affect children’s cognitive development through factors such as annoyance, distraction, reduction speech intelligibility, increased arousal, learned helplessness, frustration, and indirect effects caused by sleep disturbance (Basner et al., 2014; Shield & Dockrell, 2003).

Multiple studies support the association between noise exposure and impairments in reading and oral comprehension, memory, attention, and Standard assessment tests (SATs) (see Clark and Paunovic, 2018a for a review). The majority of the studies concerns with children exposure to aircraft or freeway traffic while in school (Sophie et al., 2014). Of particular interest is the study of Evans et al., who analyzed the consequences of the relocation of the Munich airport on children’s health and cognition (Evans et al., 1998; Hygge et al., 2002). They found that deficits in reading comprehension and long-term memory in children exposed to noise of the old airport disappeared two years after the relocation. Conversely, similar deficits developed in children exposed to noise originating from the new airport. The large-scale RANCH study comprising children from Spain, the Netherlands and the UK living near airports also found negative effects of aircraft noise exposure on reading comprehension and recognition memory, after adjusting for a range of socio-economic factors (Clark et al., 2006; Stansfeld et al., 2005).

A WHO systematic review (Clark and Paunovic, 2018a) found a significant association between exposure to aircraft noise and oral and reading comprehension as well as SATs tests, and performance on short- and long-term (Episodic) memory. The authors rated the evidence of both associations as medium quality. Evidence concerning the effects of traffic and rail noise was generally considered less consistent, which may be due to a limited number of studies, cross-sectional designs that do not allow to infer causality relations, and inconsistent performance assessment methods.

Mental Health

Transportation noise is regarded as environmental stress cause and continuous exposure is seen as a risk factor in the development of latent mental health issues (Berglund et al., 1999). A possible pathway for such effect is the increase of autonomic response and production of stress hormones, which may lead to the development of depression and anxiety disorders (Berglund et al., 1999).

Nevertheless, evidence for specific connections between noise and mental wealth is still scarce. A systematic review by WHO has pointed to a connection between exposure to road and railway noise and emotional conduct disorders, and road noise and hyperactivity in children (Clark and Paunovic, 2018b). The same report also evaluated evidence for connections between noise and measures of depression and anxiety, but they were generally considered low quality, since there are few studies, they are mostly cross-sectional, have small samples and vary substantially in methodological terms (Sørensen et al., 2020). Similar conclusions were drawn by (Sakhvidi et al., 2018) in another systematic review focused on children. As for adults, there is some recent evidence supporting a relation between noise and depression (Eze et al., 2020; Seidler et al., 2017). The cohort study by (Eze et al., 2020) shows that this relation may be mediated by annoyance, as exposure per se was only loosely related with depression. This is further supported by findings of the Gutenberg Health Study (Beutel et al., 2020) a longitudinal study with a large community sample exposed to multiple noise sources including aircraft, road, and railways. The study concluded that daytime baseline aircraft annoyance is a predictor of depression and anxiety.

Other Long-Term Effects

Cancer

A relation between transportation noise and cancer is hypothesized (Sørensen et al., 2020), possibly through long-term sleep disruption and induced stress, which are known to affect hormonal processes (Babisch et al., 2001). There is some evidence of a connection between transport noise and breast cancer (Hegewald et al., 2017; Sørensen et al., 2014), and colon cancer (Roswall et al., 2017), although research is still scarce.

Diabetes

There is some evidence of a correlation between long-term exposure to noise and the onset of diabetes. A recent meta-analysis by (Sakhvidi et al., 2018) has reviewed 15 studies on the relation between noise exposure (including several related with transportation noise) and diabetes mellitus. They found a 6% increase in the risk of diabetes mellitus for a 5 dB increase in noise exposure, mainly related to air and road traffic.

Obesity

There is evidence of a connection between noise exposure and obesity markers, such as body mass index (BMI) and waist circumference (Foraster et al., 2018; Pyko et al., 2015, 2017).

Pregnancy

There is some research relating transportation noise exposure with birth outcomes such as reduced weight, size of the newborn, and preterm birth (e.g., Smith et al., 2017; Wallas et al., 2019). A recent meta-analysis by (Dzhambov and Lercher, 2019) found only moderate-quality evidence for a relation between noise and lower birth weight, although they have identified trends related with other variables such as the reduced size for gestational age, highlighting the need for further studies.

Future Research

Exposure to transportation noise is today an important public health issue, especially for the urban population. While the issue has been the subject of intense research, there is still a considerable amount of unknowns, especially regarding the consequences of long-term exposure. Many of the studies conducted so far were done with small samples and/or relied on cross-sectional designs, making difficult the deduction of causality relations. Such seems to be one of the main conclusions of several systematic reviews commissioned by the WHO (Clark and Paunovic, 2018a, 2018b; van Kempen et al., 2018) as well as others conducted independently (Sakhvidi et al., 2018). Benefits could also be drawn from more unified approaches in the measurement and categorization of noise sources and the corresponding health outcomes, that would facilitate comparisons and meta-analysis.

Attention must also be paid to the role of confounders. Many recent studies now account for the role of air pollution exposure, which often comes hand to hand with noise exposure (Stansfeld et al., 2015; Vienneau et al., 2015). However, other factors need to be addressed, such as socio-economic factors, life-factors, gender, and pre-existing physical and mental conditions (Belojevic and Evans, 2012; van Kempen et al., 2012). Age-specific factors should also be accounted for. While children and elders are often pointed as more vulnerable populations, there is some evidence that they are less sensitive to noise annoyance than middle-age people (Van Gerven et al., 2009), which may indicate less risk of health effects. Possible explanations may be either a relation with maturation and decline of neurocognitive functions, respectively in children and elderly, or a higher average mental workload in middle age, which limits the ability to cope with noise (Van Gerven et al., 2009). Nevertheless, apart from annoyance, children may be prone to hindrances in cognitive development (see section "Effects on cognitive development"), while elders are generally more vulnerable to cardiovascular and metabolic diseases. The specificity of these age groups should therefore be studied.

Understanding the pathways through which noise affects physical and mental condition is also an important research goal. Chronic stress and sleep disturbance are often pointed as pathways to consequences in physical and mental health. The study by Eze et al. (2020) pointed to an annoyance pathway between noise and depression. Additional research is needed for understanding the hazard mechanisms and establish relations between short and long-term effects.

As transportation noise has severe consequences on health, the WHO proposed maximum limits to noise exposure. A provocative question can be asked of whether minimum limits should not also be proposed. New energy vehicles are generally less noisy than their combustion counterparts at lower speeds (like the ones practiced in urban centers). A reduction in urban noise is expected in the next years, even if only a modest adoption of these vehicles occurs (Campello-Vicente et al., 2017). However, if the change can bring positive health outcomes, there is the risk of an increase in the number of traffic crashes involving pedestrians. The vehicle industry has been trying to solve the issue by introducing artificial noises at low speeds. However, this solution can offset the benefits of a reduction in engine noise and may increase traffic noise annoyance. Understanding how low noise emission affects the vehicles' detection and localization by pedestrians, particularly by blind or visually impaired ones, is thus an important research goal if we intend to take full advantage of the noise reduction potential of electric motorization. So far, studies addressing the influence of vehicular noise on pedestrian crossing decision-making are rare, and their scope is limited. It seems that noise may not be the most important cue to ensure safe road crossing decisions by unimpaired adults (Pugliese et al., 2020; Soares et al., 2020), at least when the car is in sight. However, the same may not apply regarding the detection of nonvisible vehicles (Barton et al., 2012). In conclusion, more insight into the tradeoff between traffic noise and safety is relevant to mitigate noise effects on public health.

The new mobility paradigms allied to the alternative energy sources, with the consequent changes in noise patterns and characteristics of the vehicles, reinforce the need to keep studying and innovating on the approach to short-term effects of traffic noise on public health. Some researchers are already looking for acoustic and psychoacoustic alternative noise indicators to explain reactions to traffic noise.

Noise mitigation measures at the engineering level have been proposed to reduce the noise arriving at dwellings by acting on the source, propagation paths, and buildings (receptors). However, acting on the source is beneficial in many aspects as we all appreciate to walk on the street or open the window at home or work and find a pleasant, nonannoying soundscape. Carrying out more interdisciplinary research involving vehicles' and components' manufacturers, transportation and health professionals, and policy-makers will drive to cost-effective solutions to change to healthy and safe traffic noise levels and characteristics.

References

- Babisch, W., 2008. Road traffic noise and cardiovascular risk. *Noise Health* 10 (38), 27–33, doi:10.4103/1463-1741.39005.
- Babisch, W., 2014. Updated exposure-response relationship between road traffic noise and coronary heart diseases: A meta-analysis. *Noise Health* 16 (68), 1–9, doi:10.4103/1463-1741.127847.
- Babisch, W., Van Kamp, I., 2009. Exposure-response relationship of the association between aircraft noise and the risk of hypertension. *Noise Health* 11 (44), 161–168, doi:10.4103/1463-1741.53363.
- Babisch, W., Fromme, H., Beyer, A., Ising, H., 2001. Increased catecholamine levels in urine in subjects exposed to road traffic noise: the role of stress hormones in noise research. *Environ. Int.* 26 (7–8), 475–481, doi:10.1016/S0160-4120(01)00030-7.
- Babisch, W., Dutilleul, G., Paviotti, M., Backman, A., Gergely, B., McManus, B., Coelho, B., Hinton, J., Kephapoulos, S., van den Berg, M., Licita, G., Rasmussen, S., Blanes, N., Nugent, C., de Vos, P., Bloomfield, A., 2010. Good practice guide on noise exposure and potential health effects. In: EEA Technical report No. 11/2010.
- Bartels, S., Márki, F., Müller, U., 2015. The influence of acoustical and non-acoustical factors on short-term annoyance due to aircraft noise in the field – the COSMA study. *Sci. Total Environ.* 538, 834–843, doi:10.1016/j.scitotenv.2015.08.064.
- Barton, B.K., Ulrich, T.A., Lew, R., 2012. Auditory detection and localization of approaching vehicles. *Accid. Anal. Prev.* 49, 347–353, doi:10.1016/j.aap.2011.11.024.
- Basner, M., Babisch, W., Davis, A., Brink, M., Clark, C., Janssen, S., Stansfeld, S., 2014. Auditory and non-auditory effects of noise on health. *Lancet* 383 (9925), 1325–1332, doi:10.1016/S0140-6736(13)61613-X.
- Basner, M., McGuire, S., 2018. WHO environmental noise guidelines for the european region: a systematic review on environmental noise and effects on sleep. *Int. J. Environ. Res. Public Health* 15 (3), doi:10.3390/ijerph15030519.
- Beloevic, G., Evans, G.W., 2012. Traffic noise and blood pressure in low-socioeconomic status. African-American urban schoolchildren. *J. Acoust. Soc. Am.* 132 (3), 1403–1406, doi:10.1121/1.4739449.
- Berglund, B., Lindvall, T., Dietrich, S., Organization, W.H., 1999. Guidelines for Community Noise. <https://apps.who.int/iris/handle/10665/66217>.
- Beutel, M.E., Brähler, E., Ernst, M., Klein, E., Reiner, I., Wiltink, J., Michal, M., Wild, P.S., Schulz, A., Münzel, T., Hahad, O., König, J., Lackner, K.J., Pfeiffer, N., Tibubos, A.N., 2020. Noise annoyance predicts symptoms of depression, anxiety and sleep disturbance 5 years later. Findings from the Gutenberg Health Study. *Eur. J. Public Health* 30 (3), 516–521, doi:10.1093/eurpub/ckaa015.
- Brink, M., Schäffer, B., Vienneau, D., Foraster, M., Pieren, R., Eze, I.C., Cajochen, C., Probst-Hensch, N., Röösli, M., Wunderli, J.M., 2019. A survey on exposure-response relationships for road, rail, and aircraft noise annoyance: Differences between continuous and intermittent noise. *Environ. Int.* 125, 277–290, doi:10.1016/j.envint.2019.01.043.
- Brink, M., Schäffer, B., Vienneau, D., Pieren, R., Foraster, M., Eze, I.C., Rudzik, F., Thiesse, L., Cajochen, C., Probst-Hensch, N., Röösli, M., Wunderli, J.M., 2019. Self-reported sleep disturbance from road, rail and aircraft noise: Exposure-response relationships and effect modifiers in the SIRENE study. *Int. J. Environ. Res. Public Health* 16 (21), doi:10.3390/ijerph16214186.
- Brown, A.L., Van Kamp, I., 2015. A conceptual model of environmental noise interventions and human health effects. INTER-NOISE 2015-44th International Congress and Exposition on Noise Control Engineering.
- Campello-Vicente, H., Peral-Orts, R., Campillo-Davo, N., Velasco-Sanchez, E., 2017. The effect of electric vehicles on urban noise maps. *Appl. Acoust.* 116, 59–64, doi:10.1016/j.apacoust.2016.09.018.
- Cantuarua, M.L., Usemann, J., Proietti, E., Blanes-Vidal, V., Dick, B., Flück, C.E., Rüedi, S., Héritier, H., Wunderli, J.M., Latzin, P., Frey, U., Röösli, M., Vienneau, D., 2018. Glucocorticoid metabolites in newborns: a marker for traffic noise related stress? *Environ. Int.* 117, 319–326, doi:10.1016/j.envint.2018.05.002.
- Marquis-Favre, C., Pierre-Augustin, V., Jules, B., 2018. Short-Term Annoyance Due to Railway Noise in Urban Areas: Indices Accounting for Different Annoying Acoustical Features. In: *Euronoise*, Crete, Greece.
- Chun, C., Gwak, D.Y., Yoon, K., Lee, S., 2018. Short-term annoyance model of combined aircraft and road traffic noise based on partial loudness model. *J. Mech. Sci. Technol.* 32 (8), 3557–3562, doi:10.1007/s12206-018-0706-7.
- Clark, C., Martin, R., Van Kempen, E., Alfred, T., Head, J., Davies, H.W., Haines, M.M., Barrio, I.L., Matheson, M., Stansfeld, S.A., 2006. Exposure-effect relations between aircraft and road traffic noise exposure at school and reading comprehension: the RANCH project. *Am. J. Epidemiol.* 163 (1), 27–37, doi:10.1093/aje/kwj001.
- Clark, C., Paunovic, K., 2018a. WHO environmental noise guidelines for the european region: A systematic review on environmental noise and cognition. *Int. J. Environ. Res. Public Health* 15 (2), doi:10.3390/ijerph15020285.

- Clark, C., Paunovic, K., 2018b. Who environmental noise guidelines for the European region: A systematic review on environmental noise and quality of life, wellbeing and mental health. *Int. J. Environ. Res. Public Health* 15 (11), doi:10.3390/ijerph15112400.
- Dimakopoulou, K., Koutentakis, K., Papageorgiou, I., Kasdagli, M.-I., Sourti, P., Samoli, E., Houthuijs, D., Swart, W., Hansell, A., Katsouyanni, K., 2016. Is aircraft noise exposure associated with cardiovascular disease and hypertension? Results from a cohort study in Athens Greece. *ISEE Conf. Abstr.* 2016 (1), doi:10.1289/isee.2016.3651.
- Dzhambov, A.M., Lercher, P., 2019. Road traffic noise exposure and birth outcomes: an updated systematic review and meta-analysis. *Int. J. Environ. Res. Public Health* 16 (14), doi:10.3390/ijerph16142522.
- Elmenhorst, E.M., Griefahn, B., Rolny, V., Basner, M., 2019. Comparing the effects of road, railway, and aircraft noise on sleep: exposure–response relationships from pooled data of three laboratory studies. *Int. J. Environ. Res. Public Health* 16 (6), doi:10.3390/ijerph16061073.
- Evans, G.W., Bullinger, M., Hygge, S., 1998. Chronic noise exposure and physiological response: a prospective study of children living under environmental stress. *Psychol. Sci.* 9 (1), 75–77, doi:10.1111/1467-9280.00014.
- Eze, I.C., Foraster, M., Schaffner, E., Vienneau, D., Pieren, R., Imboden, M., Wunderli, J.M., Cajochen, C., Brink, M., Röösli, M., Probst-Hensch, N., 2020. Incidence of depression in relation to transportation noise exposure and noise annoyance in the SAPALDIA study. *Environ. Int.* 144, doi:10.1016/j.envint.2020.106014.
- Foraster, M., Eze, I.C., Vienneau, D., Schaffner, E., Jeong, A., Héritier, H., Rudzik, F., Thiesse, L., Pieren, R., Brink, M., Cajochen, C., Wunderli, J.M., Röösli, M., Probst-Hensch, N., 2018. Long-term exposure to transportation noise and its association with adiposity markers and development of obesity. *Environ. Int.* 121, 879–889, doi:10.1016/j.envint.2018.09.057.
- Guski, R., Felscher-Suhr, U., Schuemer, R., 1999. The concept of noise annoyance: how international experts see it. *J. Sound Vib.* 223 (4), 513–527, doi:10.1006/jsvi.1998.2173.
- Guski, R., Schreckenberg, D., Schuemer, R., 2017. WHO environmental noise guidelines for the European region: a systematic review on environmental noise and annoyance. *Int. J. Environ. Res. Public Health* 14 (12), doi:10.3390/ijerph14121539.
- Hälonen, J.I., Hansell, A.L., Gulliver, J., Morley, D., Blangiardo, M., Fecht, D., Toledano, M.B., Beevers, S.D., Anderson, H.R., Kelly, F.J., Tonne, C., 2015. Road traffic noise is associated with increased cardiovascular morbidity and mortality and all-cause mortality in London. *Eur. Heart J.* 36 (39), 2653–2661, doi:10.1093/eurheartj/ehv216.
- Hänninen, O., Knol, A.B., Jantunen, M., Lim, T.A., Conrad, A., Rappolder, M., Carrer, P., Fanetti, A.C., Kim, R., Buekers, J., Torfs, R., Iavarone, I., Classen, T., Hornberg, C., Mekel, O.C.L., 2014. Environmental burden of disease in Europe: assessing nine risk factors in six countries. *Environ. Health Perspect.* 122 (5), 439–446, doi:10.1289/ehp.1206154.
- Hegewald, J., Schubert, M., Wagner, M., Dräke, P., Prote, U., Swart, E., Mähler, U., Zeeb, H., Seidler, A., 2017. Breast cancer and exposure to aircraft, road, and railway-noise: a case–control study based on health insurance records. *Scand. J. Work Environ. Health*, doi:10.5271/sjweh.3665.
- Héritier, H., Vienneau, D., Foraster, M., Eze, I.C., Schaffner, E., De Hoogh, K., Thiesse, L., Rudzik, F., Habermacher, M., Köpfli, M., Pieren, R., Brink, M., Cajochen, C., Wunderli, J.M., Probst-Hensch, N., Röösli, M., 2019. A systematic analysis of mutual effects of transportation noise and air pollution exposure on myocardial infarction mortality: a nationwide cohort study in Switzerland. *Eur. Heart J.* 40 (7), 598–603, doi:10.1093/eurheartj/ehy650.
- Huang, D., Song, X., Cui, Q., Tian, J., Wang, Q., Yang, K., 2015. Is there an association between aircraft noise exposure and the incidence of hypertension? A meta-analysis of 16784 participants. *Noise Health* 17 (75), 93–97, doi:10.4103/1463-1741.153400.
- Hygge, S., Evans, G.W., Bullinger, M., 2002. A prospective study of some effects of aircraft noise on cognitive performance in schoolchildren. *Psychol. Sci.* 13 (5), 469–474, doi:10.1111/1467-9280.00483.
- ISO, 2019. ISO 3740:2019 Acoustics — Determination of sound power levels of noise sources — Guidelines for the use of basic standards. ISO, 2019.
- Jafari, Z., Kolb, B.E., Mohajerani, M.H., 2018. Chronic traffic noise stress accelerates brain impairment and cognitive decline in mice. *Exp. Neurol.* 308, 1–12, doi:10.1016/j.expneurol.2018.06.011.
- Jarup, L., Babisch, W., Houthuijs, D., Pershagen, G., Katsouyanni, K., Cadum, E., Dudley, M.L., Savigny, P., Seiffert, I., Swart, W., Breugelmans, O., Bluhm, G., Selander, J., Haralabidis, A., Dimakopoulou, K., Sourti, P., Velonakis, M., Vigna-Taglianti, F., Antonietti, M.C., . . . Zahos, Y., 2008. Hypertension and exposure to noise near airports: the HYENA study. *Environ. Health Perspect.* 116 (3), 329–333, doi:10.1289/ehp.10775.
- Lechner, C., Schnaiter, D., Bose-O'Reilly, S., 2019. Combined effects of aircraft, rail, and road traffic noise on total noise annoyance—a cross-sectional study in Innsbruck. *Int. J. Environ. Res. Public Health* 16 (18), doi:10.3390/IJERPH16183504.
- Lercher, P., 2011. Combined noise exposure at home. *Encyclop. Environ. Health* 764–777, doi:10.1016/B978-0-444-52272-6.00254-3.
- Lundberg, U., 1999. Coping with stress: neuroendocrine reactions and implications for health. *Noise Health* 1 (4), 67–74.
- Münzel, T., Daiber, A., Steven, S., Tran, L.P., Ullmann, E., Kossmann, S., Schmidt, F.P., Oelze, M., Xia, N., Li, H., Pinto, A., Wild, P., Pies, K., Schmidt, E.R., Rapp, S., Kröller-Schön, S., 2017. Effects of noise on vascular function, oxidative stress, and inflammation: mechanistic insight from studies in mice. *Eur. Heart J.* 38 (37), 2838–2849, doi:10.1093/eurheartj/ehx081.
- Munzel, T., Gori, T., Babisch, W., Basner, M., 2014. Cardiovascular effects of environmental noise exposure. *Eur. Heart J.* 35, 829–836.
- Münzel, T., Kröller-Schön, S., Oelze, M., Gori, T., Schmidt, F.P., Steven, S., Lahad, O., Röösli, M., Wunderli, J.-M., Daiber, A., Sørensen, M., 2020. Adverse cardiovascular effects of traffic noise with a focus on nighttime noise and the new WHO noise guidelines. *Annu. Rev. Public Health* 41 (1), 309–328, doi:10.1146/annurev-pubhealth-081519-062400.
- Porter, N.D., Kershaw, A.D., Ollerhead, J.B., 2000. Adverse Effects of Night-Time Aircraft Noise (Issue Report 9964). UK Civil Aviation Authority.
- Pugliese, B.J., Barton, B.K., Davis, S.J., Lopez, G., 2020. Assessing pedestrian safety across modalities via a simulated vehicle time-to-arrival task. *Accid. Anal. Prev.* 134, doi:10.1016/j.aap.2019.105344.
- Pyko, A., Andersson, N., Eriksson, C., De Faire, U., Lind, T., Mitkovskaya, N., Ögren, M., Östenson, C.G., Pedersen, N.L., Rizzuto, D., Wallas, A.K., Pershagen, G., 2019. Long-term transportation noise exposure and incidence of ischaemic heart disease and stroke: a cohort study. *Occup. Environ. Med.* 76 (4), 201–207, doi:10.1136/oemed-2018-105333.
- Pyko, A., Eriksson, C., Lind, T., Mitkovskaya, N., Wallas, A., Ögren, M., Östenson, C.G., Pershagen, G., 2017. Long-term exposure to transportation noise in relation to development of obesity—a cohort study. *Environ. Health Perspect.* 125 (11), 117005, doi:10.1289/EHP1910.
- Pyko, A., Eriksson, C., Oftedal, B., Hilding, A., Östenson, C.G., Krog, N.H., Julin, B., Aasvang, G.M., Pershagen, G., 2015. Exposure to traffic noise and markers of obesity. *Occup. Environ. Med.* 72 (8), 594–601, doi:10.1136/oemed-2014-102516.
- Pyko, A., Lind, T., Mitkovskaya, N., Ögren, M., Östenson, C.G., Wallas, A., Pershagen, G., Eriksson, C., 2018. Transportation noise and incidence of hypertension. *Int. J. Hyg. Environ. Health* 221 (8), 1133–1141, doi:10.1016/j.ijheh.2018.06.005.
- Quehl, J., Müller, U., Mendolia, F., 2017. Short-term annoyance from nocturnal aircraft noise exposure: results of the NORAH and STRAIN sleep studies. *Int. Arch. Occup. Environ. Health* 90 (8), 765–778, doi:10.1007/s00420-017-1238-7.
- Rosati, R., Jamesdaniel, S., 2020. Environmental exposures and hearing loss. *Int. J. Environ. Res. Public Health* 17 (13), 1–4, doi:10.3390/ijerph17134879.
- Roswall, N., Raaschou-Nielsen, O., Ketzel, M., Overvad, K., Halkjaer, J., Sørensen, M., 2017. Modeled traffic noise at the residence and colorectal cancer incidence: a cohort study. *Cancer Causes Control* 28 (7), 745–753, doi:10.1007/s10552-017-0904-0.
- Rudzik, F., Thiesse, L., Pieren, R., Héritier, H., Eze, I.C., Foraster, M., Vienneau, D., Brink, M., Wunderli, J.M., Probst-Hensch, N., Röösli, M., Fulda, S., Cajochen, C., 2020. Ultradian modulation of cortical arousals during sleep: effects of age and exposure to nighttime transportation noise. *Sleep* 43 (7), doi:10.1093/sleep/zsz324.
- Sakhvidi, F.Z., Sakhvidi, M.J.Z., Mehrparvar, A.H., Dzhambov, A.M., 2018. Environmental noise exposure and neurodevelopmental and mental health problems in children: a systematic review. *Curr. Environ. Health Rep.* 5 (3), 365–374, doi:10.1007/s40572-018-0208-x.
- Sakhvidi, M.J.Z., Sakhvidi, F.Z., Mehrparvar, A.H., Foraster, M., Davdand, P., 2018. Association between noise exposure and diabetes: a systematic review and meta-analysis. *Environ. Res.* 166, 647–657, doi:10.1016/j.envres.2018.05.011.
- Schäfer, M., 1997. Soundscape: Our Sonic Environment and the Tuning of the World. Knopf.
- Schmidt, F., Kolle, K., Kreuder, K., Schnorbus, B., Wild, P., Hechtner, M., Binder, H., Gori, T., Münzel, T., 2015. Nighttime aircraft noise impairs endothelial function and increases blood pressure in patients with or at high risk for coronary artery disease. *Clin. Res. Cardiol.* 104 (1), 23–30, doi:10.1007/s00392-014-0751-x.
- Schmidt, F.P., Basner, M., Kroger, G., Weck, S., Schnorbus, B., Muttray, A., Sariyar, M., Binder, H., Gori, T., Warnholtz, A., Münzel, T., 2013. Effect of nighttime aircraft noise exposure on endothelial function and stress hormone release in healthy adults. *Eur. Heart J.* 34 (45), 3508–3514, doi:10.1093/eurheartj/ehz269.

- Seidler, A., Hegewald, J., Seidler, A.L., Schubert, M., Wagner, M., Dröge, P., Haufe, E., Schmitt, J., Swart, E., Zeeb, H., 2017. Association between aircraft, road and railway traffic noise and depression in a large case-control study based on secondary data. *Environ. Res.* 152, 263–271, doi:10.1016/j.envres.2016.10.017.
- Seidler, A.L., Hegewald, J., Schubert, M., Weihofen, V.M., Wagner, M., Dröge, P., Swart, E., Zeeb, H., Seidler, A., 2018. The effect of aircraft, road, and railway traffic noise on stroke – results of a case-control study based on secondary data. *Noise Health* 20 (95), 152–161, doi:10.4103/nah.NAH_7_18.
- Shield, B.M., Dockrell, J.E., 2003. The effects of noise on children at school: a review. *Build. Acoust.* 10 (2), 97–116, doi:10.1260/135101003768965960.
- Smith, M.G., Ögren, M., Ageborg Morsing, J., Persson Waye, K., 2019. Effects of ground-borne noise from railway tunnels on sleep: a polysomnographic study. *Build. Environ.* 149, 288–296, doi:10.1016/j.buildenv.2018.12.009.
- Smith, R.B., Fecht, D., Gulliver, J., Beevers, S.D., Dajnak, D., Blangiardo, M., Ghosh, R.E., Hansell, A.L., Kelly, F.J., Ross Anderson, H., Toledano, M.B., 2017. Impact of London's road traffic air and noise pollution on birth weight: retrospective population based cohort study. *BMJ (Online)* 359, doi:10.1136/bmj.j5299.
- Soares, F., Silva, E., Pereira, F., Silva, C., Sousa, E., Freitas, E., 2020. The influence of noise emitted by vehicles on pedestrian crossing decision-making: a study in a virtual environment. *Appl. Sci. (Switzerland)* 10 (8), doi:10.3390/APP10082913.
- Sophie, P., Jean-Pierre, L., Hélène, H., Rémy, P., Marc, B., Jérôme, D., Joseph, L., Cyril, M., Frédéric, M., 2014. Association between ambient noise exposure and school performance of children living in an urban area: a cross-sectional population-based study. *J. Urban Health* 91 (2), 256–271.
- Sørensen, M., Ketzel, M., Overvad, K., Tjønneland, A., Raaschou-Nielsen, O., 2014. Exposure to road traffic and railway noise and postmenopausal breast cancer: a cohort study. *Int. J. Cancer* 134 (11), 2691–2698.
- Sørensen, Mette, Münzel, T., Brink, M., Roswall, N., Wunderli, J.M., Foraster, M., 2020. Transport, noise, and health. *Adv. Transport. Health* 105–131, doi:10.1016/b978-0-12-819136-1.00004-8.
- Stansfeld, S.A., Berglund, B., Clark, C., Lopez-Barrio, I., Fischer, P., Öhrström, E., Haines, M.M., Head, J., Hygge, S., van Kamp, I., Berry, B.F., 2005. Aircraft and road traffic noise and children's cognition and health: a cross-national study. *Lancet* 365 (9475), 1942–1949, doi:10.1016/S0140-6736(05)66660-3.
- Stansfeld, S., Matheson, M., 2003. Noise pollution: non-auditory effects on health. *Br. Med. Bull.* 68, 243–257, doi:10.1093/bmb/ldg033.
- Stansfeld, Stephen, Clark, C., 2015. Health effects of noise exposure in children. *Curr. Environ. Health Rep.* 2 (2), 171–178, doi:10.1007/s40572-015-0044-1.
- Van Gerven, P.W.M., Vos, H., Van Bortel, M.P.J., Janssen, S.A., Miedema, H.M.E., 2009. Annoyance from environmental noise across the lifespan. *J. Acoust. Soc. Am.* 126 (1), 187–194, doi:10.1121/1.3147510.
- Van Kempen, E., Van Kamp, I., Fischer, P., Davies, H., Houthuijs, D., Stellato, R., Clark, C., Stansfeld, S., 2006. Noise exposure and children's blood pressure and heart rate: the RANCH project. *Occup. Environ. Med.* 63 (9), 632–639, doi:10.1136/oem.2006.026831.
- van Kempen, Elise, Casas, M., Pershagen, G., Foraster, M., 2018. WHO environmental noise guidelines for the European region: a systematic review on environmental noise and cardiovascular and metabolic effects: a summary. *Int. J. Environ. Res. Public Health* 15 (2), doi:10.3390/ijerph15020379.
- van Kempen, Elise, Fischer, P., Janssen, N., Houthuijs, D., van Kamp, I., Stansfeld, S., Cassee, F., 2012. Neurobehavioral effects of exposure to traffic-related air pollution and transportation noise in primary schoolchildren. *Environ. Res.* 115, 18–25, doi:10.1016/j.envres.2012.03.002.
- Vienneau, D., Perez, L., Schindler, C., Lieb, C., Sommer, H., Probst-Hensch, N., Künzli, N., Röösli, M., 2015. Years of life lost and morbidity cases attributable to transportation noise and air pollution: a comparative health risk assessment for Switzerland in 2010. *Int. J. Hygiene Environ. Health* 218, 514–521, doi:10.1016/j.ijheh.2015.05.003.
- Wallas, A., Ekström, S., Bergström, A., Eriksson, C., Gruzdeva, O., Sjöström, M., Pyko, A., Ögren, M., Bottai, M., Pershagen, G., 2019. Traffic noise exposure in relation to adverse birth outcomes and body mass between birth and adolescence. *Environ. Res.* 169, 362–367, doi:10.1016/j.envres.2018.11.039.
- World Health Organization, 2018. Environmental noise guidelines for the European region.

Climate Change and Health, Related to Transport

Ersilia Verlinghieri, Transport Studies Unit, School of Geography and the Environment, University of Oxford, Oxford, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

The Strong Links Between Transport and Climate Change	320
High-Carbon Transport Systems	320
The Links Between Climate-Change and High-Carbon Transport Systems	321
The Health Implications of High-Carbon Transport Systems	321
The Health Implications of a Changing Climate	322
How We can Tackle Adverse Health Impacts and at the Same Time Reduce Climate Change by Rethinking our Transport Systems	323
Changes in the Physical Infrastructure of Transport Systems	323
Changes in Mobilities	323
Changes in Transport Planning and Governance	324
Acknowledgment	325
References	325

The Strong Links Between Transport and Climate Change

The reality of anthropogenic climate change is not just one of incremental deviation from typical temperature and climate conditions, with warmer summers for some, shorter winters for others and more common extreme weather events. Climate change means increasing likelihood of abrupt, unpredictable and drastic global changes that will suddenly disrupt our and other species' lives, livelihoods and wellbeing (Lenton et al., 2008). The magnitude of these damages is expected to be far greater what we have experienced during the recent COVID-19 pandemic, for the price in human lives, geographical extent, involvement of different life forms and irreversibility. It is not a surprise then that the rapidly changing climate is becoming a key concern for citizens across the world, and that transport policy and planning have been increasingly called into question for not adequately addressing what ought to be a key objective.

Evidence shows that reducing transport-related carbon emissions is a vital step toward the avoidance of catastrophic climate change and that, at the same time, transport is one of the most challenging sectors to decarbonise (IPCC, 2018). Despite multiple sustainable transport initiatives, emissions from the sector have stubbornly continued to grow. In 2017, transport worldwide accounted for 25% of carbon emissions and, between 2010 and 2017, for half of their global growth (IEA, 2018). Although the COVID-19 pandemic brought a temporary and unexpectedly substantial decrease in emissions, maintaining these positive effects, and effectively and permanently decarbonising the sector, ought to be a national and international priority. At the same time, there is evidence that this still constitutes a substantial challenge, as it requires a radical reconfiguration of lifestyles and cities, with a major reduction in car dependency, long-distance travel, carbon-intensive import-export markets, etc. (Schwanen, 2019).

In this chapter I will explore the way in which climate change is strongly linked to current high-carbon transport systems and human health in a negative, self-reinforcing cycle, the risks we face if no action is taken and the possibilities we still have to break this negative cycle.

High-Carbon Transport Systems

"High-carbon transport system" (HCTS) is a term that can help us describe the key features of most established transport systems^a in Western society, which are becoming increasingly popular in other global regions. HCTS include what the academic literature has described as the systems of "automobility" and "aeromobility," expanding paradigms of mobility practices, technological and infrastructural developments, land use patterns, and modes of production and consumption that lock-in car (and increasingly air-travel) dependency (Freund, 2014; Lassen, 2016; Paterson, 2007). This dominance of the private car and, more recently, long-distance hypermobility via flying, are the cardinal points of worldwide transport systems and, more broadly, oil-based economies (Mattioli et al., 2020). Automobility and aeromobility are planned on the assumption that favoring speed, efficiency, and economic growth equates with increasing social benefits (Schwanen, 2019). They constitute a salient feature of neoliberal urban development and are, evidently, a major contributor to global CO₂ emissions and air pollution (Freund, 2014; Paterson, 2007). They have key and widespread negative impacts on health and wellbeing across several geographies (Khreis et al., 2016). As depicted in the diagram below, HCTS directly generate health impacts and directly contribute to climate change, which, in turn, indirectly increases the negative health effects, as I will show in the next sections of this chapter (Figure 1).

^a With transport systems here I refer to what Schwanen (2019, np) calls "transport configurations – the ensembles of infrastructures, technologies, practices, regulation, meanings, values, affects enabling and shaping the (non)movements through physical space of people, goods and waste".

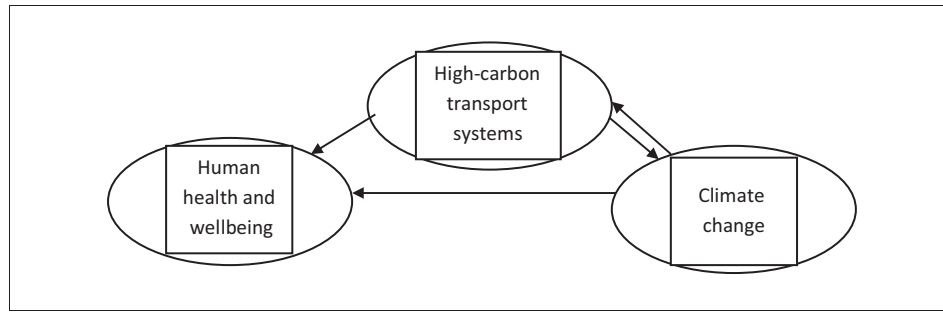


Figure 1 The relation between HCTS, health and climate change

The Links Between Climate-Change and High-Carbon Transport Systems

Interdisciplinary research efforts have now confirmed that the current drastic changes in climate are not just fluctuations of temperature as traditionally experienced in our planet's history, but substantial deviation from average temperature as a result of disproportionate release of carbon dioxide (CO₂) and other greenhouse gasses (such as black carbon and methane) in the atmosphere. CO₂ concentration are reaching a value higher than at any time in at least 800,000 years, a 40% increase with respect to historic trends (IPCC, 2018). These gasses are released as a result of most human activities involving combustion of fossil fuel and changes in land-use, including farming and deforestation (IPCC, 2019). To date, these carbon-intensive human activities have caused between 0.8 and 1.2 of global warming with the expectation of an increase to 1.5 by 2052, if there is no change (IPCC, 2018). The damaging effects of this anthropogenic global warming are becoming increasingly evident and are predicted to get much more severe. They include sea level rises, heat waves, heavy rainfalls and floods, forests dieback, biodiversity losses, increase risk of vector-borne diseases, etc. (Lenton et al., 2008).

The transport sector is the second largest and still growing contributor to global greenhouse gas emissions and especially CO₂, the backbone of climate change, therefore the name HCTS. Despite improvements in other sectors, at the beginning of 2020, transportation was still responsible for 24% of direct CO₂ emissions from fuel combustion, an immense responsibility in an era of climate crisis (FAOSTAT, 2020; IEA, 2020). This is the result of automobility and aeromobility being fundamentally oil-based: the high number of road vehicles used in Western countries and, increasingly, globally (GHO, 2020) accounts for almost three-quarters of the emissions from the transport sectors. Improvements related to electric vehicle adoption are rapidly outpaced by the higher emissions of increasingly popular SUVs, whilst aviation and shipping emissions continue to rise faster than any other sector (Chapman, 2007; IEA, 2020).

We can therefore clearly state that transport, as HCTS, is a key cause of anthropogenic climate change. This has been widely acknowledged, with climate change mitigation becoming a key topic in transport studies and increasingly transport policy (see for example (DfT, 2020). Although the magnitude of the work published in this direction is too large to be reviewed in this chapter [for a summary of works on climate change in transport geography see for example Schwanen (2019)] some key themes are emerging when considering mitigation pathways, as I will explore in the last section of this chapter.

At the same time, extreme weather events related to climate change have important impacts on transport systems, operations, and their resilience (Budnitz et al., 2020), and in recent years there has been an expansion in research considering potential climate change adaptation paths in the field, for example focusing on sea level rise impact on coastal transport infrastructure (Dawson et al., 2016), or changing travel behaviors as consequence of higher temperatures (Ferranti et al., 2017).

The Health Implications of High-Carbon Transport Systems

HCTS have notoriously negative impacts on human health and wellbeing. These can be grouped in three main areas: direct impact on individual health, secondary impacts on generalized health as result of changes in behaviors and the environment, and broader societal impacts.

First, as reported by the WHO (2020), approximately 1.35 million people die each year worldwide as a result of road traffic crashes. Road traffic is the leading cause of death for people aged 5–29 years, effectively generating an immense and silent genocide that has been globally accepted in a “rather unquestioning manner” (Wells and Beynon, 2011, p.2494). Although this effect is not necessarily linked to the type of combustion engine utilized, transport systems designed around the private car, with their high consumption of public space and high speeds, are substantially intertwined with an uneven system of “death and violence,” in which the ones that are most affected are non-car users. This effect is sharpened in poorer countries where traffic accident victims are often young adults from less wealthy groups (Culver, 2018; Wells and Beynon, 2011). As suggested by Culver (2018), fundamental justice issues permeate the way pedestrians and cyclists suffer the most from someone else's use of a motor vehicle. This injustice, as will be

evident in the next sections, applies not only to the distribution of damage related to traffic accidents, but also to the health impacts related to traffic emissions.

Second, the high number of road vehicles generate substantial changes to the environment which impact human health. HCTS do not only contribute to greenhouse gas emissions, but are also a key source of air pollution, being responsible for a high proportion of particulate matter circulating in the atmosphere as well as other pollutants such as NO, NO₂, carbon-monoxide, and other volatile organic compounds such as benzene and ethene. These are released both via exhaust emissions and as consequence of wear and tear phenomena (e.g., tyres and brakes), and are proven to generate vast negative health effects such as cardiovascular, cerebral-vascular and respiratory mortality and morbidity, decreased lung function in children, and diabetes (Landrigan et al., 2018; Khreis et al., 2016). Similar negative effects are also linked to the high level of traffic-related noise. Furthermore, as widely acknowledged, a vast burden of disease is linked to the sedentary life-style that car-dependency promotes. Lack of physical activity is linked to cardiovascular diseases, cerebrovascular disease, cancer (colon, breast, and lung), type 2 diabetes, dementia, anxiety, depression, obesity (Khreis et al., 2016). To these need to be added the unexpectedly substantial negative health impacts associated with the reduced availability of green space and biodiversity loss as a consequence of car-centred planning that, together with a culture of speed and efficiency linked to car-centered societies, has substantial damaging mental health and inflammatory outcomes (Nieuwenhuijsen, 2020).

Finally, key societal inequalities are embedded in HCTS which indirectly result in health issues. These are of great concern and their magnitude becomes clearer if we adopt a framework in which health is linked to wider discussions around human wellbeing. Although several definitions of what wellbeing entails are possible, it can be linked to the possibility for each individual to live a fulfilling life by, for example, being able to access and perform activities which she values (Sen, 2010). In this sense, transport is directly responsible for allowing people to access and perform most of these activities and therefore reach a suitable level of wellbeing.

Following this, or other complementary ideas around wellbeing, the transport and mobilities literature is increasingly highlighting the strong inequality outcomes of HCTS (Verlinghieri and Schwanen, 2020). The uneven access that different population groups have to a car in a world designed around it (Mattioli et al., 2018)^b, coupled with the very limited availability of alternatives (both because of affordability and reach of services) generate “transport poverty” (Lucas et al., 2016). This forced immobility or the high cost^c of mobility, heavily impacts people’s ability to access healthcare, education, appropriate livelihoods, social networks, and therefore their generalized wellbeing. Similarly, poorest groups are also the most affected by transport related air pollution, even though they emit the least (Jephcote et al., 2016). These impacts are unevenly distributed also with relation to income, gender, race, age, sexual orientation, and their intersection, and are particularly felt with new established HCTS in non-western countries (Uteng and Lucas, 2018).

The Health Implications of a Changing Climate

Warmer weather and extreme weather events are reported to have substantial health impacts for all and especially most vulnerable and less mobile individuals. Climate change is likely to increase thermal-related mortality, water/food/vector-borne diseases, skin cancers, malnutrition, respiratory diseases linked to augmented air pollution, as well as direct increase in injury and deaths as effect of extreme climate events (Nichols et al., 2009). Extreme weather will also make walking, that is the predominant mean of transport for the poor in most countries, unsustainable.

At the same time, extreme climate event as well as long-term environmental changes such as droughts, will generate important mental health effects, damage livelihoods of already vulnerable population with resulting lack of access to land or clean water, forced mass migrations and broadening of societal inequalities. As UNDP reported already in 2007, “increased exposure to drought, to more intense storms, to floods and environmental stress is holding back the efforts of the world’s poor to build a better life for themselves and their children” (UNDP, 2007; p.1).

Environmental and climate justice movements have been highlighting these dangers for decades and have developed a clear agenda that addresses the uneven distribution of climate change effects and the substantial changes needed in our society and economies to address those (Routledge et al., 2018). Their analysis points out the unjust outcomes of HCTS. While wealthier citizens in the West (or, as minorities in non-western countries) benefit from the availability and comfort of automobility and aeromobility, and of the related government and private sectors investments which strongly sustain their car-centered lifestyles (Mattioli et al., 2020) and their opportunity for seamless long-distance air-travels, most vulnerable groups across different geographies pay the price of this highly unsustainable systems. These people, that often rely solely on walking and, are effected both directly in their individual health, and more broadly and profoundly in terms of irreversible damages to their livelihoods, economies, social networks and cultures.

^b Or the high price of maintaining one for some groups forced to do so to reach their livelihoods, even though maintaining a vehicle costs an unacceptably high proportion of their disposable income.

^c Here with cost I intend to include also the non-monetary aspects that are linked to unequal transport systems, in terms of safety, discomfort, risk of harassment, etc. which are mostly paid by minorities.

How We can Tackle Adverse Health Impacts and at the Same Time Reduce Climate Change by Rethinking our Transport Systems

From the analysis presented, it is clear that there is a strong link between automobility, aeromobility, and climate change and that efforts to reduce the health impacts of both have to be directed toward shifting human mobilities away from the focus on the private car and, more generally, away from high-carbon transport systems.

Academics and policy makers have suggested several different pathways to reduce carbon emissions from transport. These act at different levels of transport systems that, for simplicity and following the definition from Timms et al. (2014), can be grouped as changes in *physical infrastructures*, *mobilities*, and *planning and governance*. As I show in the remainder of this chapter, if considered as a whole, these pathways might be able to direct us toward starting to challenge automobility and tackle its damaging impacts on health and the climate.

Changes in the Physical Infrastructure of Transport Systems

This level is traditionally considered central in tackling climate change. Technological change and modernization, as well as the construction of transport infrastructures, have been and still are at the core of transport research and policy, and attract substantial public and private investment. Vehicle and fuel efficiency, electrification, and automation are seen as the key premises for both less-carbon intensive and less polluting transport futures. In the United Kingdom, for example, they are the backbone of the £1.5 billion investment as part of the DfT's recent Road to Zero strategy (DfT, 2018), which includes end of sale of traditional diesel and petrol cars, substantive incentives for EVs purchase, development and charging infrastructures, wide adoption of low-emission fuels, and development and adoption of new low-emissions technologies.

Automation and smart technologies are also believed to facilitate a transition to less-carbon intensive transport systems as, with specific implementation conditions, they could provide new avenues for changing travel patterns and behaviors, with more potential for diminishing car ownership via on-demand mobility services, as well as the adoption of energy-saving driving practices and more efficient vehicles (Wadud et al., 2016).

However, there are some substantial challenges in tackling adverse health impacts of our transport systems and at the same time reducing climate change by solely improving transport technologies.

First, coordinating efforts toward carbon and air pollution reduction and human wellbeing, is a great challenge, and some of the solutions that might lower carbon emissions may have damaging social effects (Mattioli, 2016). For example, current EVs technologies do not reduce traffic accidents. Similarly, they are still very high in PM emissions (Timmers and Achten, 2016) and there is a lack of clarity on the precise full life-cycle emissions of both new vehicles and supporting infrastructure production. Automation and AI for shared mobilities come with huge privacy and data use issues that are still far to be solved. Moreover, some infrastructural developments have key distributional impacts, either because they encourage the phenomena of gentrification or more simply through their allocation of substantial incentives and resources to facilitate the movements of groups that are already among the most mobile (Schwanen, 2020). New technologies won't be affordable for great part of population and surely for the most vulnerable groups in the poorest countries. Directing urban development toward accommodating primarily technologies that are accessible only to high income groups risks exacerbating the transport poverty of vulnerable groups who already pay greatly in terms of health and wellbeing impacts.

Finally, there is a risk that the economic benefits of higher vehicle efficiency (or perceived safety as for automated cars) may further encourage car ownership, further locking-in, with substantial development of hard infrastructures for private mobility, car dependency, and preventing the development of alternative options (Henderson, 2020). Overall, there is a clear risk that the proposed technological advancement if considered on its own or as the core of any future transport strategy, will not be able to challenge the patterns of land use and travel behaviors, as well as the logic of efficiency, acceleration and growth that is fundamental to automobility and key in the worsening of the climate emergency (Schwanen, 2019).

Changes in Mobilities

In Timms et al. (2014), the definition of transport systems uses the term "mobilities" to refer to the movements of people and goods across space. Academically, the term is also used to indicate a research direction that, in the last decade, has highlighted the crucial role of movement in shaping not only transport systems, but the entire society (Sheller and Urry, 2006, 2016). The "mobilities tradition" complements traditional transport studies by looking beyond mere physical movements "from A to B," and focusing on the *meanings* and *practices* that individuals and society associate with these movements. By asking questions such as: "Why do we make journeys from one place to another?"; "What modes of transport do different people use?"; "How exactly do they use them?"; and "How do they shape what we are and the way we live?," it highlights the different dimensions that influence travel and movements and how they are differentially experienced by people or shape their lives.

The mobilities tradition can be of help when looking at changes to transport systems traditionally framed in terms of "modal shift" or "behavioral changes." These include changes in travel in terms of: (1) which vehicle is used -promoting the adoption, for example, of walking or cycling or public transport-, (2) frequency and length of journeys, -for example promoting trip-chaining or changes in land use-, (3) a combination of the two, for example, with the idea of Mobility as a Service (MaaS). Several strategies have been adopted to implement these changes, including information campaigns, investments in public

transport or walking and cycling infrastructures and trainings, taxation measures for vehicles, new ticketing options, changes in land-use (Nieuwenhuijsen, 2020).

These strategies are promising in their ability to be effective in reducing carbon emissions and, at the same time, promoting healthy outcomes. There is widespread evidence of the health benefits of walking and cycling, with the associated physical activity outweighing the risk of road accidents or increased air pollution exposure. Similarly, land use changes that, via greater density, diversification of use and better design, promote greater accessibility to key destinations by walking, cycling or public transport, result also in widespread health benefits. Among all, the example of the Barcelona's super blocks provide evidence of how well-designed land-use changes combined with strategies to recover public and green space and therefore reduce car use have enormous health benefits (Nieuwenhuijsen, 2020).

However, convincing people to give up their cars and shift to other ways of moving has proven to be a great challenge and the widespread and lasting adoption of low-carbon mobilities still has far to go. The mobilities literature provides an explanation for this, showing how specific travel practices are far from being the result of an individual rational and weighted choice. They are part of a broad socio-technical system in which available technologies and materials, cultural norms and values, personal experiences and habits interlock with specific travel practices (Stephenson et al., 2015). Ideas of freedom, privacy, comfort, and individuality associated with the private car are reinforced and replicated in car adverts, movies, societal discourses and personal experiences, making car dependence a much stronger habit that we might initially imagine (Paterson, 2007). At the same time, the adoption of new mobility practices, especially when including a new series of bodily movements, knowledge, skills, and resources, such as cycling or utilizing MaaS, might become an unsurmountable challenge for resource-scarce groups of the population whose survival is tied to the availability of the car. The gilets jaunes movement in France or the recent mobilizations in Chile have shown how much urban livelihoods are tight to transport pricing. Similarly, the same act of walking or biking might be perceived as socially unacceptable or extremely unsafe for certain groups, especially women or minorities (Law, 1999), so that certain investments once again do not respond to the needs of the most vulnerable.

Changes in Transport Planning and Governance

Beyond changing the type of vehicles we use or their fuels, or the way we use them, there is also the option of implementing wider and broader changes in the way transport systems are planned and governed. As a start, this means opening up transport planning and policy to dialogue with other perspectives and traditions; for example widening the connection with the health profession or by using instruments such as Health Impact Assessment tools, or giving substantial and continued space to participatory planning experiences (Nieuwenhuijsen, 2020).

However, this approach might still be not enough when the way mobilities can and are governed is fundamentally shaped by broader societal processes and political interests. Tackling car dependency cannot fail to acknowledge the profound economic interests that are behind its perpetuation. The automotive industry has been a major political player in locking in automobility worldwide and creating a self-reinforcing system. It has strongly supported the dismantling and underfunding of public transport; it has lobbied to pressure public investment in pro-car infrastructures and devolution of public space to make room for cars; it has joined real-estate developers to favor urban sprawl; it has actively supported the development of "car culture," with advertising and marketing, and intentionally shaping consumers' choices and practices (Mattioli, 2020). As Paterson (2007) reports, automobility has become fundamental to capital accumulation and a less carbon intense world ought to profoundly rethink this relationship. A similar discussion applies to the raising economic and lobbying power of the flight industry that, as we have witnessed during the COVID-19 outbreak, has a strong influence on governments' decisions. This means that there is a need to fundamentally politicizing the discussion around transport futures, being aware of the power relations and interests that shape decisions in transport and the possibilities to shift them.

A substantial and sustained shift away from HCTS and the health impacts and inequalities associated with them, together with the changes so far mentioned, requires a profound shift in the key objective of planning and policy, as well of our economies, away from being shaped by the interests of automobility and aeromobility. As increasingly suggested in several fields of geographical, sociological, and economic research, this can be done by decoupling the idea of speed and economic growth from the achievement of human wellbeing and separating the interests of capital from the pursuit of liveable cities (D'Alessandro et al., 2020; Ferreira and Schönfeld, 2020; Raworth, 2017). Frameworks that, for example, consider transportation and mobility as a common good can be highly promising in this direction (Nikolaeva et al., 2019).

At the same time, challenging HCTS requires considering the multi-scalar, cross-border, intersectional nature of the changes we are facing. As shown in this chapter, this can be achieved by going beyond considering transport only as a matter of individual movement, technologies and infrastructure, and embracing more of a mobilities perspective which allows us to account for the societal meanings, practices and political processes linked to those movements, infrastructures and technologies. This, as suggested by sociologist Mimi Sheller (2018) in her key book on mobility justice, can allow us to integrate the understanding of what ought to be changed in HCTS with the possibility of its implementation when strictly linked to broader socioeconomic elements as well as processes of climate change and justice, migrations, social inequalities, and urban crisis.

Finally, as Sheller (2018) also suggests and as mentioned before, it is perhaps by opening up to other voices such as environmental and social movements, citizens initiatives and grassroots experiences that true change could be brought about. The climate movement as well as transport activists have been working for decades in analyzing the limitations of current mobility systems as part of a broader context of systems of capitalism, extractivism and colonialism. With a logic of self-determination and profound

respect for the ecology, they have promoted an agenda that suggests “*leaving fossil fuels in the ground, reasserting peoples’ and community control over the production of food and renewable energy, massively reducing over-consumption, particularly in the Global North, respecting the rights of indigenous and forest people, and recognising the ecological and climate debt owed to the peoples in the Global South by the societies of the Global North necessitating the making of reparations*” (Routledge et al. 2018, 79). Their experience, in participatory dialogue with the voices of citizens and minority groups, can allow planners and researchers to develop solutions to the world-dominant paradigm of the car (and increasingly the aeroplane) that are tailored to the needs of specific populations and geographies, and at the same time recognize the global effects of local choices in transport and mobilities, both for the climate and the life on the planet.

Acknowledgment

I would like to greatly thank Dr. Hannah Budnizt and Simona Sulikova for their comments and suggestions that significantly improved the manuscript.

References

- Barnes, J.H., Chatterton, T.J., Longhurst, J.W.S., 2019. Emissions vs exposure: Increasing injustice from road traffic-related air pollution in the United Kingdom. *Transp. Res. Part D: Transp. Environ.* 73, 56–66, doi:10.1016/j.trd.2019.05.012.
- Budnizt, H., Tranos, E., Chapman, L., 2020. Responding to stormy weather: Choosing which journeys to make. *Travel Behav. Soc.* 18, 94–105, doi:10.1016/j.tbs.2019.10.008.
- Chapman, L., 2007. Transport and climate change: A review. *J. Transp. Geogr.* 15, 354–367, doi:10.1016/j.jtrangeo.2006.11.008.
- Culver, G., 2018. Death and the Car: On (Auto)Mobility, Violence, and Injustice 27.
- D'Alessandro, S., Cieplinski, A., Distefano, T., Dittmer, K., 2020. Feasible alternatives to green growth. *Nat. Sustain.* 3, 329–335, doi:10.1038/s41893-020-0484-y.
- Dawson, D., Shaw, J., Roland Gehrels, W., 2016. Sea-level rise impacts on transport infrastructure: The notorious case of the coastal railway line at Dawlish, England. *J. Transp. Geogr.* 51, 97–109, doi:10.1016/j.jtrangeo.2015.11.009.
- Department for Transport, 2020. Decarbonising Transport: Setting the Challenge. Department for Transport, London.
- Department for Transport, 2018. The Road to Zero : Next steps towards cleaner road transport and delivering our Industrial Strategy. Department for Transport, London.
- FAOSTAT [WWW Document], n.d. URL <http://www.fao.org/faostat/en/#data/EM> (accessed 7.2020).
- Ferranti, E., Chapman, L., Whyatt, D., 2017. A perfect storm? The collapse of Lancaster's critical infrastructure networks following intense rainfall on 4/5 December 2015. *Weather* 72, 3–7, doi:10.1002/wea.2907.
- Ferreira, A., Schönfeld, K.C. von, 2020. Interlacing planning and degrowth scholarship. *disP - Plan. Rev.* 56, 53–64, doi:10.1080/02513625.2020.1756633.
- Freund, P., 2014. The revolution will not be motorized: Moving toward nonmotorized spatiality. *Capital. Nat. Social.* 25, 7–18, doi:10.1080/10455752.2014.964504.
- GHO By category Registered vehicles - Data by country [WWW Document], n.d. WHO. URL <https://apps.who.int/gho/data/node.main.A995> (accessed 7.2020).
- Henderson, J., 2020. EVs Are Not the Answer: A Mobility Justice Critique of Electric Vehicle Transitions. *Annals of the American Association of Geographers* 1–18. <https://doi.org/10.1080/24694452.2020.1744422>.
- IEA, 2018. World Energy Outlook 2018. IEA.
- International Energy Agency, 2020. Global Energy Review 2020: The impacts of the Covid-19 crisis on global energy demand and CO2 emissions. OECD. <https://doi.org/10.1787/a60abbf2-en>.
- IPCC, 2019. Climate Change and Land. IPCC.
- IPCC, n.d. Global Warming of 1.5°C. IPCC.
- Jephcote, C., Chen, H., Ropkins, K., 2016. Implementation of the Polluter-Pays Principle (PPP) in local transport policy.. *J. Transp. Geogr.* 55, 58–71, doi:10.1016/j.jtrangeo.2016.06.017.
- Khreis, H., Warsow, K.M., Verlingheri, E., et al., 2016. The health impacts of traffic-related exposures in urban areas: Understanding real effects, underlying driving forces and co-producing future directions. *J. Transp. Health* 3, 249–267, doi:10.1016/j.jth.2016.07.002.
- Landrigan, P.J., Fuller, R., Acosta, N.J.R., et al., 2018. The Lancet Commission on pollution and health. *Lancet* 391, 462–512.
- Lassen, C., 2016. Aeromobility and Work: Environment and Planning A. Available from: <https://doi.org/10.1068/a37278>.
- Law, R., 1999. Beyond 'women and transport': towards new geographies of gender and daily mobility. *Prog. Human Geogr.* 23, 567–588, doi:10.1191/030913299666161864.
- Lenton, T.M., Held, H., Kriegler, E., Hall, J.W., Lucht, W., Rahmstorf, S., Schellnhuber, H.J., 2008. Tipping elements in the Earth's climate system. *Proc. Natl. Acad. Sci.* 105, 1786–1793, doi:10.1073/pnas.0705414105.
- Lucas, K., Mattioli, G., Verlingheri, E., Guzman, A., 2016. Transport poverty and its adverse social consequences. *Proc. Inst. Civil Eng. Transp.* 169, 353–365, doi:10.1680/jtran.15.00073.
- Mattioli, G., 2016. Transport needs in a climate-constrained world. A novel framework to reconcile social and environmental sustainability in transport. *Ener. Res. Soc. Sci.* 18, 118–128, doi:10.1016/j.erss.2016.03.025.
- Mattioli, G., Roberts, C., Steinberger, J.K., Brown, A., 2020. The political economy of car dependence: a systems of provision approach.. *Energy Res. Soc. Sci.* 66, 101486, doi:10.1016/j.erss.2020.101486.
- Mattioli, G., Nicolas, J.-P., Gertz, C., 2018. Household transport costs, economic stress and energy vulnerability. *Transp. Policy* 65, 1–4, doi:10.1016/j.tranpol.2017.11.002.
- Mattioli, G., Roberts, C., Steinberger, J.K., Brown, A., 2020. The political economy of car dependence: A systems of provision approach. *Ener. Res. Soc. Sci.* 66, 101486, doi:10.1016/j.erss.2020.101486.
- Nichols, A., Maynard, V., Goodman, B., Richardson, J., 2009. Health, climate change and sustainability: A systematic review and thematic analysis of the literature. *Environ. Health Insights* 3, EHI S3003, doi:10.4137/EHI.S3003.
- Nieuwenhuijsen, M.J., 2020. Urban and transport planning pathways to carbon neutral, liveable and healthy cities; A review of the current evidence. *Environ. Int.* 140, 105661, doi:10.1016/j.envint.2020.105661.
- Nikolaeva, A., Adey, P., Cresswell, T., Lee, J.Y., Nóvoa, A., Temenos, C., 2019. Commoning mobility: Towards a new politics of mobility transitions. *Trans. Inst. Br. Geogr.* 44, 346–360, doi:10.1111/tran.12287.
- Paterson, M., 2007. Automobile politics: ecology and cultural political economy. Cambridge University Press, Cambridge; New York.
- Raworth, K., 2017. Doughnut economics: Seven ways to think like a 21st-century economist. Chelsea Green Publishing, Chelsea.
- World Health Organization (WHO), 2020. Road traffic injuries. Available from: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>.
- Routledge, P., Cumbers, A., Derickson, K.D., 2018. States of just transition: Realising climate justice through and against the state. *Geoforum* 88, 78–86, doi:10.1016/j.geoforum.2017.11.015.

- Schwanen, T., 2020. Low-carbon mobility in London: A just transition? *One Earth* 2, 132–134, doi:10.1016/j.oneear.2020.01.013.
- Schwanen, T., 2019. Transport geography, climate change and space: opportunity for new thinking. *J. Transp. Geogr.* 81, 102530, doi:10.1016/j.jtrangeo.2019.102530.
- Sen, A., 2010. *The Idea of Justice*. Penguin UK, London.
- Sheller, M., 2018. *Mobility Justice*. Verso, London, Brooklyn, NY.
- Sheller, M., Urry, J., 2016. Mobilizing the new mobilities paradigm. *Appl. Mobil.* 1, 10–25, doi:10.1080/23800127.2016.1151216.
- Sheller, M., Urry, J., 2006. The new mobilities paradigm. *Environ. Plan. A* 38, 207–226, doi:10.1068/a37268.
- Stephenson, J., Hopkins, D., Doering, A., 2015. Conceptualizing transport transitions: Energy cultures as an organizing framework. *WIREs Ener. Environ.* 4, 354–364, doi:10.1002/wene.149.
- Timmers, V.R.J.H., Achten, P.A.J., 2016. Non-exhaust PM emissions from electric vehicles. *Atmos. Environ.* 134, 10–17, doi:10.1016/j.atmosenv.2016.03.017.
- Timms, P., Tight, M., Watling, D., 2014. Imagineering mobility: Constructing utopias for future urban transport. *Environ Plan A* 46, 78–93, doi:10.1068/a45669.
- UNDP, 2007. *Making Globalization Work for All*. United Nations Development Programme.
- Uteng, T.P., Lucas, K., 2017. *Urban Mobilities in the Global South*. Routledge, Abingdon, New York.
- Verlinghieri, E., Schwanen, T., 2020. Transport and mobility justice: Evolving discussions. *J. Transp. Geogr.* 87, 102798, doi:10.1016/j.jtrangeo.2020.102798.
- Wadud, Z., MacKenzie, D., Leiby, P., 2016. Help or hindrance? The travel, energy and carbon impacts of highly automated vehicles. *Transp. Res. Part A: Policy Prac.* 86, 1–18, doi:10.1016/j.tra.2015.12.001.
- Wells, P., Beynon, M.J., 2011. Corruption, automobility cultures and road traffic deaths: The perfect storm in rapidly motorizing countries? *Environ Plan A* 43, 2492–2503, doi:10.1068/a4498.
- WHO, 2020. Road traffic injuries, <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>.

Urban Greenspace, Transportation, and Health

Payam Dadvand, Mark J. Nieuwenhuijsen, Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain

© 2021 Elsevier Ltd. All rights reserved.

Overview	327
Urban Greenspaces	327
Potential Mechanisms	328
Stress Reduction/Cognitive Restoration	328
Mitigating Urban-Related Environmental Hazards	328
Enhancing Social Interaction and Cohesion	329
Increasing Physical Activity	329
Enriching Environmental Biodiversity	329
Health Benefits	329
Pregnancy Outcomes and Complications	329
Neurodevelopment in Children	330
Cognitive Function in Adults	330
Perceived General Health	330
Mental Health	330
Other Noncommunicable Diseases	330
Ageing	330
Mortality	330
Health Risks	330
Asthma and Allergy	331
Pesticide Exposure	331
Vector-Based and Zoonotic Infections	331
Accidental Injuries	331
Greenspace, Transportation, and Healthy Urban Living	331
Commuting	331
Green Replacement	331
Final Remarks	332
References	332

Overview

The rapid urbanization started in the last century is still ongoing. Currently, more than half of the global population resides in cities and this proportion is projected to increase to more than two-third by 2050 (UN Department of Economic and Social Affairs, 2015). Cities are the powerhouses of innovation and wealth creation where residents usually have better access to healthcare and other facilities (Bettencourt et al., 2007). However, living in urban areas is often associated with higher exposure to a number of environmental hazards such as air pollution, noise, and heat and limited access to natural environments. At the same time, urban lifestyle is mainly associated with less physical activity and higher psychological stress (Bettencourt et al., 2007). These environmental and lifestyle factors are proposed to contribute to the existing higher prevalence of a wide range of adverse health conditions such as psychological disorders and chronic noncommunicable diseases in urban areas (Cyril et al., 2013). Natural environments, including greenspace, have been related with improved physical and mental health and wellbeing. They are also recognized as a mitigation measure to buffer the aforementioned adverse health effects of urban areas. This chapter provides an overview of (1) what urban green spaces are; (2) the potential mechanisms through which exposure to greenspace could exert health effects; (3) the health effects associated with greenspace exposure; and (4) the potentials of greenspace and transportation to promote healthy urban living.

Urban Greenspaces

US Environmental Protection Agency (EPA) has defined green space as the “land that is partly or completely covered with grass, trees, shrubs, or other vegetation” which includes “parks, community gardens, and cemeteries” (US Environmental Protection Agency, 2017). The abundance and availability of greenspace in cities could be a function of several factors from which the urban planning and climate play key roles. A survey of 386 European cities (2009), for example, showed that while there was a general north-south decreasing gradient in the proportion of the greenspace coverage among these cities, still there were greener cities in south and less

green cities in north (Fuller and Gaston, 2009). The amount of green space available to people in cities also varies considerably from, for example, 1.9 m² per person in Buenos Aires, Argentina to 52.0 m² in Curitiba, Brazil. There are many types of green spaces in cities including parks, street green, and natural green, which can be captured by different maps or remote sensing methods (Gascon et al., 2016a). Hence, the estimation of the proportion of the city covered with greenspace or the available greenspace per capita depends on the definition of greenspace and the source of available data to calculate greenspace.

Potential Mechanisms

The mechanisms underlying health effects of the exposure to greenspace are not well-established yet but reducing stress, restoring cognitive function, enhancing social cohesion/interactions, reducing sedentary behavior/increasing physical activity, mitigating the exposure to air pollution, noise, and heat, and enriching micro- and macro-biodiversity and hence environmental microbial input have been suggested to be involved.

Stress Reduction/Cognitive Restoration

An extensive body of experimental and observational evidence has established the capability of the exposure to greenspace to reduce stress and restore cognition function. The *Stress Reduction Theory* proposes that greenspace, through properties such as curving sightlines, spatial openness, and the presence of water, could induce recovery from stress and help to diminish states of arousal and negative thoughts through psycho-physiological pathways (Ulrich, 1984). *Attention Restoration Theory* postulates that contact with nature with its innately delightful stimuli could modestly invoke indirect (i.e., effortless) attention and at the same time minimize the need for directed attention that together could restore the directed attention capability (Berman et al., 2008; Kaplan, 1995; Kaplan and Kaplan, 1989). These pathways have been suggested to play important roles in the health benefits of exposure to greenspace (Dadvand et al. 2016; de Vries et al. 2013).

Mitigating Urban-Related Environmental Hazards

The impact of greenspace on air pollution is complex and context-specific. On one hand, vegetations has been suggested to reduce air pollution by direct and indirect mechanisms (Givoni, 1991). The direct mechanism is through filtering air pollutants by vegetations, mainly based on dry deposition of pollutants (both particles and gases) through stomata uptake or non-stomata deposition on plant surfaces (Akbari, 2002; Givoni, 1991; Nowak et al., 2006; Paoletti et al., 2011). The indirect effect is proposed to be through cooling effect of plants that in turn reduces smog formation (Givoni, 1991). A study on the effects of greenspace surrounding home address on personal exposure to air pollution using personal air pollution monitors reported that higher surrounding greenspace was associated with reduced personal exposure to particulate air pollution but not nitric oxides (Dadvand et al., 2012). Another study also showed that higher greenspace within and surrounding schools is associated with lower indoor and outdoor levels of traffic-related air pollutants at school (Dadvand et al., 2015). The capability of vegetations to reduce air pollution is suggested to be type-specific with trees being the most efficient and grasses being the least efficient types (Givoni, 1991). Modelling studies of the amount of air pollution removed by tree canopies in continental US (Nowak et al., 2006) and Greater London (Tallis et al., 2011) have estimated that about 1%–2% of air pollution in these areas are removed by canopies. Experimental studies on the mitigation of the air pollution by roadside vegetation have reported inconsistent results with some reports that do not support such a mitigation (Baldauf et al. 2011; Hagler et al. 2012). Simulation studies of this mitigation effect have also indicated that roadside trees are able to generate a canyon effect with higher air pollution levels on the downwind and lower air pollution levels on the upwind side of the street (Baldauf et al., 2011; Buccolieri et al., 2009). Furthermore, biodegradation of vegetation residues could generate volatile organic compounds (VOCs), a family of air pollutants with potential health effects on humans. VOCs could also participate in complex photochemical reactions with other air pollutants such as nitric oxides and ozone and engage in the generation of biogenic secondary organic aerosols (Hoyle et al., 2011; Kesselmeier and Staudt, 1999). Although the interaction between greenspace and air pollution appears to be multifaceted and complex and the available evidence on such interaction is still limited and inconsistent, the available studies evaluating the mediator role of air pollution in the observed health benefits of the exposure to greenspace are suggestive for such a mediation. For example, a recent study of the association of the exposure to greenspace on cognitive development in children estimated that up to 60% of this association could be explain by the reduction of traffic-related air pollutants by greenspace (Dadvand et al., 2015).

The capability of greenspace to reduce temperature is well established. Shading, evapotranspiration (emission of water vapor into the atmosphere), and the micro-regulation of air movements and heat exchange are among the mechanisms through which vegetations could modulate the temperature (Bowler et al., 2010). A systematic review and meta-analysis of the available studies on such effect concluded that the temperature in urban parks is on average 1°C less than that of other nongreen areas in cities (Bowler et al., 2010). Given the existence of Heat Island Effect in urban areas, the capability of the greenspace to reduce temperature is of importance for promoting resilience in cities, especially by considering the occurring changes in the global climate.

The available evidence on mitigation of noise exposure by greenspace is still limited. However, these studies are suggestive for the buffering of the noise exposure and/or reduction of noise annoyance by the greenspace surrounding the homes and also by green facades (De Ridder et al., 2004; Dzhambov et al., 2018; Gidlöf-Gunnarsson and Öhrström, 2007). Schäffer et al. (2020) investigated

the effects of residential green on noise annoyance, accounting for different transportation noise sources as well as for the degree of urbanization (Schäffer et al., 2020). Increasing residential green was found to be associated with reduced annoyance due to road traffic and railway noise, but increased annoyance due to aircraft noise.

Enhancing Social Interaction and Cohesion

A cohesive society is defined as a society that “works towards the well-being of all its members, fights exclusion and marginalization, creates a sense of belonging, promotes trust, and offers its members the opportunity of upward mobility” (Organisation for Economic Cooperation and Development (OECD), 2011). Social interaction and cohesion have been related with improved perceived general health (Kawachi et al., 2008), lower morbidity, more longevity, and reduced inequality in health (Marmot et al., 2012). The body of evidence on the relation between the exposure to greenspace and social interaction and/or cohesion is still limited; however, it is generally supportive for such a relation (Dadvand et al., 2016, 2019; de Vries et al., 2013; Maas et al., 2009a; Sugiyama et al., 2008), with a few exceptions (Triguero-Mas et al. 2015). A small number of studies have also reported that social interaction and/or cohesion could mediate the association of the exposure to greenspace with perceived general health (Dadvand et al., 2016; de Vries et al., 2013; Maas et al., 2009a).

Increasing Physical Activity

Physical activity has been postulated to act as a mediator of the health of benefits of the greenspace exposure. However, the available evidence on the potential effect of greenspace exposure on physical activity is inconsistent with a considerable heterogeneity in the reported direction and strength of associations (Bancroft et al., 2015; Lachowycz and Jones, 2011; McGrath et al., 2015). A part of this heterogeneity could be because most of these studies have not accounted for the quality aspects of green spaces while these aspects are shown to affect the use of these spaces for physical activity (McCormack et al., 2010). The few studies evaluating the mediation of health benefits of greenspace exposure by physical activity are suggestive for a modest (Dadvand et al., 2016; de Vries et al., 2013) or no (de Keijzer et al., 2018; de Keijzer et al., 2019b) mediation role of physical activity in these benefits.

Enriching Environmental Biodiversity

Plants directly modulate the microbiome present in the rhizosphere (the below-ground microbial habitat provided by plant root system) and phyllosphere (the above-ground microbial habitats provided by plants) (Berendsen et al., 2012; Vorholt, 2012), and hence could indirectly modulate the environmental microbiome to which humans are exposed. Available reports suggest that bacterial diversity in humans’ faeces decreases with the level of urbanization, which is strongly associated with reduced environmental biodiversity (De Filippo et al., 2010; Yatsunenko et al., 2012). Human microbiome including gut microbiome has been shown to interact with the host organs, regulate systemic immune response and prevent chronic inflammation (Martinez, 2014). Therefore, the ability of greenspace to enrich immunoregulation-inducing microbial input from the environment (Rook, 2013) could be a potential mechanism underlying the relation between the greenspace exposure and human health. A study in adolescents, for example, reported that living near a forest during childhood could enhance the diversity of the skin microbiome, which in turn was associated with the reduced risk of allergic sensitization later in life (Hanski et al., 2012a).

Health Benefits

Exposure to greenspace has been related with a wide range of physical and mental health benefits. This exposure, for instance, has been related with improved perceived general health, better pregnancy outcomes and lower risk of pregnancy complications, enhanced neurodevelopment in children, improved mental health and cognitive function, lower risk of a number of chronic diseases (e.g., diabetes and cardiovascular conditions), decelerated ageing, and reduced premature mortality.

Pregnancy Outcomes and Complications

The most consistent evidence on the association of maternal greenspace exposure and pregnancy outcomes has been reported for the fetal growth. Higher greenspace surrounding maternal residential address during pregnancy has been associated with increased birth weight and reduced risk of low birth weight or small for gestational age (Akaraci et al., 2020; Zhan et al., 2020). The available evidence for this association for the length of pregnancy is inconsistent. While some studies are suggestive for an increased length of gestation (i.e., reduced risk of preterm birth) associated with higher greenspace surrounding maternal residential address (Grazuleviciene et al., 2015; Hystad et al., 2014; Laurent et al., 2013; Nichani et al., 2017), other studies have not supported this association (Agay-Shay et al., 2014; Dadvand et al., 2012a, 2012b). A limited number of studies have also evaluated the association between maternal greenspace exposure during pregnancy and the risk of pregnancy complications such as gestational diabetes and pregnancy-induced hypertensive disorders including preeclampsia. Although a recent systematic review and meta-analysis of the available literature (Zhan et al., 2020) did not find any statistically significant association between greenspace exposure and pregnancy complications, but these studies were generally supportive of a protective association.

Neurodevelopment in Children

The “biophilia hypothesis” proposes that humans have evolutionary bonds to nature (Kellert and Wilson, 1993; Wilson, 1984). Accordingly, contact with nature including greenspace is postulated to have a crucial role in the neurodevelopment of children (Kahn and Kellert, 2002; Kellert, 2005). Earlier experimental studies have shown that spending time and playing in greenspaces could, in short-term, reduce severity of symptoms and improve attention in children with attention deficit-hyperactivity disorder (ADHD) (Kuo and Taylor, 2004; Taylor et al., 2001; Taylor and Kuo, 2009; van den Berg and van den Berg, 2011). More recently, observational epidemiological studies have reported that higher residential surrounding greenspace and more time spent playing in greenspaces in long run could reduce risk of behavioral and emotional problems including ADHD (Amoly et al., 2014; Markevych et al., 2014b) and enhance cognitive development including attention and working memory (Dadvand et al., 2015, 2017; Wells, 2000).

Cognitive Function in Adults

In adults, the available evidence for the association of exposure to greenspace with cognitive function is still scarce and mainly deals with effects of visual access to greenspace on cognitive function (de Keijzer et al., 2016). These studies, all cross-sectional, are suggestive for improved cognitive functions including better direct attentional capacity and lower concentration problems associated with the visual access to greenspace (de Keijzer et al., 2016).

Perceived General Health

Exposure to greenspace has been consistently related with improved perceived general health in adults (Gascon et al., 2015) and self-satisfaction in children (Dadvand et al., 2019). Studies have shown that improved mental health and social cohesion and, to less extent, enhanced physical activity are among the main mechanisms underlying this association (Dadvand et al., 2016; de Vries et al., 2013).

Mental Health

The effect of exposure to greenspace on mental health is one of the most established health effects of this exposure. Higher exposure to greenspace has been associated with improved mood and lower risk of psychological distress and psychiatric conditions such as depression and anxiety and less likelihood of use of psychiatric medicine (Gascon et al., 2015; Mygind et al., 2019; Vanaken and Danckaerts, 2018).

Other Noncommunicable Diseases

The available evidence on the impacts of greenspaces on noncommunicable diseases other than asthma and allergy is still limited but is suggestive for a beneficial impact. More exposure to greenspace has been associated with lower risk of cardiovascular conditions, diabetes, and lower back pain in adults (Dalton et al., 2016; Maas et al., 2009b). In children, this exposure has been associated with lower blood pressure (Markevych et al., 2014a) and lower blood glucose level and less risk of hyperglycaemia (Dadvand et al., 2018).

Ageing

The available evidence on the potential effects of long-term exposure to greenspace on ageing is accumulating and is suggestive for the capability of this exposure to promote healthy ageing and improve mental and physical health and wellbeing of elderly (de Keijzer et al., 2020). Specifically, long-term exposure to greenspace has been associated with better cognitive function, physical capability, mental health and perceived wellbeing and lower risk of morbidities (de Keijzer et al., 2020). For example, a series of recent longitudinal studies conducted in the ageing Whitehall II cohort has shown that higher residential surrounding greenspace is associated with decelerated cognitive ageing (de Keijzer et al., 2018) and physical functioning decline (de Keijzer et al., 2019b) and a reduced risk of incident metabolic syndrome (de Keijzer et al., 2019a) in elderly.

Mortality

Two systematic reviews and meta-analyses of the available literature on the impact of exposure to greenspace on mortality have found that higher residential surrounding greenspace is associated with reduced all-cause premature mortality as well as cardiovascular mortality (Gascon et al., 2016b; Rojas-Rueda et al., 2019). More physical activity, lower exposure to air pollution, enhanced social cohesion, and improved mental health have been reported to mediate the association between exposure to greenspace and mortality (James et al., 2016).

Health Risks

In addition to the health benefits, green spaces could also potentially induce a number of adverse health outcomes including asthma and allergic conditions, increase exposure to pesticides, host reservoirs and vectors of infectious diseases, and enhance the risk of accidental injuries.

Asthma and Allergy

The available evidence on the impact of exposure to greenspace on asthma and allergic conditions in children is inconsistent. While some studies have related higher residential surrounding greenspace with increased risk of asthma and allergic conditions (Andrusaityte et al., 2016; DellaValle et al., 2012; Lovasi et al., 2013), others have not found such a relation or have even reported protective associations (Hanski et al., 2012b; Hind et al., 2017; Lovasi et al., 2008; Maas et al., 2009b; Pilat et al., 2012). The type of greenspace and the bioclimatic properties of the study region could explain, in part, such an inconsistency. One study, for example, has reported that while urban parks were related with higher risk of asthma and allergic attacks, natural greenspaces (e.g., forests) did not show such a relation (Dadvand et al., 2014). Another study conducted in seven birth cohorts in Australia, Canada, Germany, Netherlands, and Sweden, showed a notable between-centre heterogeneity in terms of the direction and strength of associations (Fuentes et al., 2016). A recent systematic review of the available evidence has suggested a potentially protective association of the greenspace exposure with asthma in children (Hartley et al., 2020).

Pesticide Exposure

Application of pesticides in green spaces, especially in agricultural lands, could expose individuals living nearby these spaces or those who use these spaces to these chemicals. This exposure, in turn, could lead to a range of health outcomes including cancers as well as adverse conditions in nervous, reproductive, endocrine, and immune systems (Blair et al., 2015).

Vector-Based and Zoonotic Infections

Green spaces could host reservoirs and vectors of infectious diseases, which could increase the risk of vector-borne diseases transferred by mosquitoes (e.g., Dengue fever or malaria), ticks (e.g., Lyme disease and tick-borne encephalitis), or sandflies (e.g., leishmaniasis) (WHO Regional Office for Europe, 2016). Exposure to animal faeces in green spaces could also induce zoonotic infections such as toxocariasis or toxoplasmosis (WHO Regional Office for Europe, 2016).

Accidental Injuries

Users of green spaces, especially children, could experience accidental injuries such as falls or drowning while they are in these spaces. However, at population level, the injuries that occur in green spaces account for a very tiny proportion of accidental injuries (WHO Regional Office for Europe, 2016) and they are to a large extent preventable through proper design of these spaces and putting in place preventive measures.

Greenspace, Transportation, and Healthy Urban Living

Commuting

Commuting routes with natural features could promote walking or cycling for commuting. Moreover, commuting through green environments could have mental health benefits as exposure to greenspace can reduce stress and improve mental health. In one of the few available studies, Zijlema et al. (2018) evaluated the association between natural, mainly green, environments and commuting, whether active or not, and the association between commuting (through green environments), whether active or not, and mental health in four European cities in Spain, the Netherlands, Lithuania and the United Kingdom (Zijlema et al., 2018). Data on commuting behavior (active commuting at least one day/week, daily green environment commuting) and mental health were collected through questionnaires. Adjusted multilevel analyses showed that daily green environment commuters were more often active commuters. Furthermore, there was no association between active commuting and mental health, but daily commuters through green environments had on average a 2.74 (95% confidence intervals (CIs): 1.66, 3.82) point higher mental health score than those not commuting through green environments. The association with mental health was stronger among active commuters (4.03, 95% CIs: 2.13, 5.94) compared to nonactive commuters (2.21; 95% CIs: 0.90, 3.51) when daily commuting through green environments. Briefly, the study showed that Daily commuting through green environments was associated with better mental health, especially for active commuters. But also that daily commuters through green environments were more likely to be active commuters. Active commuting itself was not associated with mental health. These findings suggest that cities should invest in green commuting routes for cycling and walking.

Green Replacement

Given the many benefits of the exposure greenspace, the health and wellbeing of residents in urban areas, where there is often a lack of greenspace, can be improved through increasing the amount, availability, and quality of green spaces (Nieuwenhuijsen et al., 2017; van den Bosch and Nieuwenhuijsen, 2017). Cities can be made healthier and more equitable for people, not by painting trees on walls, but by having a park near where people live, planting trees in the streets, and introducing urban gardens. Urban gardens may have additional benefits in terms of social cohesion, local food production and economy and, if done at a

sufficiently large scale, can contribute to the sustainability and self-sufficiency of cities. Most cities need more parks, which can also become a part of their identity and attraction. Also, green roofs and walls may transform the city, not only in terms of resilience but also in terms of visual attractiveness. Our current cities are too car dominated, and car infrastructure such as roads and parking take up much space that can be, otherwise, used for developing green spaces. Reducing space for cars and the number of cars could have the additional advantage that people have to switch to public and active transportation and thereby reducing the major urban-related environmental hazards (i.e., air pollution and noise) in cities and increasing physical activity in citizens (Nieuwenhuijsen and Khreis, 2016). New urban concepts like the Superblocks (Barcelona), car free cities (Hamburg) or the 15-min city (Paris) can create space for greenspace and a healthier transport environment that not only are more sustainable and liveable but also healthier (Nieuwenhuijsen, 2020). Although greening our cities is not the only solution to improve the health and wellbeing of residents, it can certainly make an important contribution, which is being increasingly recognized through new studies assessing the health impacts of urban and transport planning (Nieuwenhuijsen, 2020).

Final Remarks

Although greenspace can result in great health benefits and many cities are bringing in more greenspace, the magnitude of the link between transport, greenspace and health has not been well explored and needs further work. There are still questions such as for example what type of, how much and where do we exactly need more greenspace to get the greatest benefits. However, this should not stop us from greening of cities and creating greener environments for transport, particularly for active commuting, which is the most sustainable and healthy way of mobility.

References

- Agay-Shay, K., Peled, A., Crespo, A.V., Peretz, C., Amitai, Y., Linn, S., et al., 2014. Green spaces and adverse pregnancy outcomes. *Occup. Environ. Med.* 71, 562–569.
- Akaraci, S., Feng, X., Suesse, T., Jalaludin, B., Astell-Burt, T., 2020. A systematic review and meta-analysis of associations between green and blue spaces and birth outcomes. *Int. J. Environ. Res. Public Health* 17, 2949.
- Akbari, H., 2002. Shade trees reduce building energy use and CO₂ emissions from power plants. *Environ. Pollut.* 116, S119–S126.
- Amoly, E., Dadvand, P., Forns, J., López-Vicente, M., Basagaña, X., Julvez, J., et al., 2014. Green and blue spaces and behavioral development in barcelona schoolchildren: The breathe project. *Environ. Health Perspect.* 122, 1351–1358.
- Andrusaityte, S., Grazuleviciene, R., Kudzyte, J., Bernotiene, A., Dedele, A., Nieuwenhuijsen, M.J., 2016. Associations between neighbourhood greenness and asthma in preschool children in kaunas, lithuania: A case control study. *BMJ Open* 6, e010341.
- Baldauf, R., Jackson, L., Hagler, G., Vlad, I., McPherson, G., Nowak, D., et al., 2011. The role of vegetation in mitigating air quality impacts from traffic emissions. *EM Jan*: 1–3.
- Bancroft, C., Joshi, S., Rundle, A., Hutson, M., Chong, C., Weiss, C.C., et al., 2015. Association of proximity and density of parks and objectively measured physical activity in the united states: a systematic review. *Soc. Sci. Med.* 138, 22–30.
- Berendsen, R.L., Pieterse, C.M.J., Bakker, P.A.H.M., 2012. The rhizosphere microbiome and plant health. *Trends Plant Sci.* 17, 478–486.
- Berman, M.G., Jonides, S., Kaplan, S., 2008. The cognitive benefits of interacting with nature. *Psychol. Sci.* 19, 1207–1212.
- Bettencourt, L.M.A., Lobo, J., Helbing, D., Kühnert, C., West, G.B., 2007. Growth, innovation, scaling, and the pace of life in cities. *PNAS* 104, 7301–7306.
- Blair, A., Ritz, B., Wesseling, C., Beane Freeman, L., 2015. Pesticides and human health. *Occup. Environ. Med.* 72, 81–82.
- Bowler, D.E., Buyung-Ali, L., Knight, T.M., Pullin, A.S., 2010. Urban greening to cool towns and cities: A systematic review of the empirical evidence. *Landsc. Urban Plan.* 97, 147–155.
- Buccolieri, R., Gromke, C., Di Sabatino, S., Ruck, B., 2009. Aerodynamic effects of trees on pollutant concentration in street canyons. *Sci. Total Environ.* 407, 5247–5256.
- Cyril, S., Oldroyd, J.C., Renzaho, A., 2013. Urbanisation, urbanicity, and health: A systematic review of the reliability and validity of urbanicity scales. *BMC Public Health* 13, 513.
- Dadvand, P., de Nazelle, A., Figueras, F., Basagaña, X., Sue, J., Amoly, E., et al., 2012a. Green space, health inequality and pregnancy. *Environ. Int.* 40, 110–115.
- Dadvand, P., de Nazelle, A., Triguero-Mas, M., Schembari, A., Cirach, M., Amoly, E., et al., 2012b. Surrounding greenness and exposure to air pollution during pregnancy: An analysis of personal monitoring data. *Environ. Health Perspect.* 120, 1286–1290.
- Dadvand, P., Sunyer, J., Basagaña, X., Ballester, F., Lertxundi, A., Fernández-Somoano, A., et al., 2012b. Surrounding greenness and pregnancy outcomes in four spanish birth cohorts. *Environ. Health Perspect.* 120, 1481–1487.
- Dadvand, P., Villanueva, C.M., Font-Ribera, L., Martínez, D., Basagaña, X., Belmonte, J., et al., 2014. Risks and benefits of green spaces for children: A cross-sectional study of associations with sedentary behavior, obesity, asthma, and allergy. *Environ. Health Perspect.* 122, 1329–1335.
- Dadvand, P., Nieuwenhuijsen, M.J., Esnaola, M., Forns, J., Basagaña, X., Alvarez-Pedrerol, M., et al., 2015. Green spaces and cognitive development in primary schoolchildren. *Proc Natl Acad Sci U S A* 112, 7937–7942.
- Dadvand, P., Rivas, I., Basagaña, X., Alvarez-Pedrerol, M., Su, J., De Castro Pascual, M., et al., 2015. The association between greenness and traffic-related air pollution at schools. *Sci. Total Environ.* 523, 59–63.
- Dadvand, P., Bartoll, X., Basagaña, X., Dalmau-Bueno, A., Martínez, D., Ambros, A., et al., 2016. Green spaces and general health: roles of mental health status, social support, and physical activity. *Environ. Int.* 91, 161–167.
- Dadvand, P., Tischer, C., Estarlich, M., Llop, S., Dalmau-Bueno, A., López-Vicente, M., et al., 2017. Lifelong residential exposure to green space and attention: A population-based prospective study. *Environ. Health Perspect.* 125, 097016.
- Dadvand, P., Poursafa, P., Heshmat, R., Motlagh, M.E., Qorbani, M., Basagaña, X., et al., 2018. Use of green spaces and blood glucose in children; a population-based caspian-v study. *Environ. Pollut.* 243, 1134–1140.
- Dadvand, P., Hariri, S., Abbasi, B., Heshmat, R., Qorbani, M., Motlagh, M.E., et al., 2019. Use of green spaces, self-satisfaction and social contacts in adolescents: A population-based caspian-v study. *Environ. Res.* 168, 171–177.
- Dalton, A.M., Jones, A.P., Sharp, S.J., Cooper, A.J.M., Griffin, S., Wareham, N.J., 2016. Residential neighbourhood greenspace is associated with reduced risk of incident diabetes in older people: A prospective cohort study. *BMC Public Health* 16, 1171.
- De Filippo, C., Cavalieri, D., Di Paola, M., Ramazzotti, M., Poullet, J.B., Massart, S., et al., 2010. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc. Natl. Acad. Sci. U. S. A.* 107, 14691–14696.
- de Keijzer, C., Gascon, M., Nieuwenhuijsen, M.J., Dadvand, P., 2016. Long-term green space exposure and cognition across the life course: A systematic review. *Curr. Environ. Health Rep.* 3, 468–477.

- de Keijzer, C., Tonne, C., Basagaña, X., Valentín, A., Singh-Manoux, A., Alonso, J., et al., 2018. Residential surrounding greenness and cognitive decline: A 10-year follow-up of the whitehall ii cohort. *Environ. Health Perspect.* 126, 077003.
- de Keijzer, C., Basagaña, X., Tonne, C., Valentín, A., Alonso, J., Antó, J.M., et al., 2019a. Long-term exposure to greenspace and metabolic syndrome: A whitehall ii study. *Environ. Pollut.* 255, 113231.
- de Keijzer, C., Tonne, C., Sabia, S., Basagaña, X., Valentín, A., Singh-Manoux, A., et al., 2019b. Green and blue spaces and physical functioning in older adults: Longitudinal analyses of the whitehall ii study. *Environ. Int.* 122, 346–356.
- de Keijzer, C., Bauwelinck, M., Dadvand, P., 2020. Long-term exposure to residential greenspace and healthy ageing: A systematic review. *Curr. Environ. Health Rep.* 7, 65–88.
- De Ridder, K., Adamec, V., Bañuelos, A., Bruse, M., Bürger, M., Damsgaard, O., et al., 2004. An integrated methodology to assess the benefits of urban green space. *Sci. Total Environ.* 334–335, 489–497.
- de Vries, S., van Dillen, S.M.E., Groenewegen, P.P., Spreeuwenberg, P., 2013. Streetscape greenery and health: Stress, social cohesion and physical activity as mediators. *Soc. Sci. Med.* 94, 26–33.
- DellaValle, C.T., Triche, E.W., Leaderer, B.P., Bell, M.L., 2012. Effects of ambient pollen concentrations on frequency and severity of asthma symptoms among asthmatic children. *Epidemiology* 23, 55–63.
- Dzhambov, A.M., Markevych, I., Tilov, B., Arabadzhiev, Z., Stoyanov, D., Gatseva, P., et al., 2018. Lower noise annoyance associated with gis-derived greenspace: Pathways through perceived greenspace and residential noise. *Int. J. Environ. Res. Public Health* 15, 1533.
- Fuertes, E., Markevych, I., Bowatte, G., Gruzdeva, O., Gehring, U., Becker, A., et al., 2016. Residential greenness is differentially associated with childhood allergic rhinitis and aeroallergen sensitization in seven birth cohorts. *Allergy*, doi:10.1111/all.12915:n/a-n/a.
- Fuller, R.A., Gaston, K.J., 2009. The scaling of green space coverage in european cities. *Biol. Lett.* 5, 52–55.
- Gascon, M., Triguero-Mas, M., Martínez, D., Dadvand, P., Fors, J., Plasència, A., et al., 2015. Mental health benefits of long-term exposure to residential green and blue spaces: A systematic review. *Int. J. Environ. Res. Public Health* 12, 4354–4379.
- Gascon, M., Cirach, M., Martínez, D., Dadvand, P., Valentín, A., Plasència, A., et al., 2016a. Normalized difference vegetation index (ndvi) as a marker of surrounding greenness in epidemiological studies: The case of barcelona city. *Urban For Urban Green* 19, 88–94.
- Gascon, M., Triguero-Mas, M., Martínez, D., Dadvand, P., Rojas-Rueda, D., Plasència, A., et al., 2016b. Residential green spaces and mortality: A systematic review. *Environ. Int.* 86, 60–67.
- Gidlöf-Gunnarsson, A., Öhrström, E., 2007. Noise and well-being in urban residential environments: The potential role of perceived availability to nearby green areas. *Landsc. Urban Plan.* 83, 115–126.
- Givoni, B., 1991. Impact of planted areas on urban environmental quality: A review. *Atmos. Environ. Part B Urban Atmos.* 25, 289–299.
- Grazuleviciene, R., Danileviciute, A., Dedele, A., Vencloviene, J., Andrusaityte, S., Uzdaviciute, I., et al., 2015. Surrounding greenness, proximity to city parks and pregnancy outcomes in kaunas cohort study. *Int. J. Hyg. Environ. Health* 218, 358–365.
- Hagler, G.S.W., Lin, M.-Y., Khlystov, A., Baldauf, R.W., Isakov, V., Faircloth, J., et al., 2012. Field investigation of roadside vegetative and structural barrier impact on near-road ultrafine particle concentrations under a variety of wind conditions. *Sci. Total Environ.* 419, 7–15.
- Hanski, I., von Hertzen, L., Fyhrquist, N., Koskinen, K., Torppa, K., Laatikainen, T., et al., 2012a. Environmental biodiversity, human microbiota, and allergy are interrelated. *Proc. Natl. Acad. Sci. U. S. A.* 109, 8334–8339.
- Hanski, I., von Hertzen, L., Fyhrquist, N., Koskinen, K., Torppa, K., Laatikainen, T., et al., 2012b. Environmental biodiversity, human microbiota, and allergy are interrelated. *Proc. Natl. Acad. Sci. U. S. A.* 109, 8334–8339.
- Hartley, K., Ryan, P., Brokamp, C., Gillespie, G.L., 2020. Effect of greenness on asthma in children: A systematic review. *Public Health Nurs.* 37, 453–460.
- Hind, S., Mieke, K., Lillian, T., Michael, B., 2017. Asthma trajectories in a population-based birth cohort. Impacts of air pollution and greenness. *Am. J. Respir. Crit. Care Med.* 195, 607–613.
- Hoyle, C.R., Boy, M., Donahue, N.M., Fry, J.L., Glasius, M., Guenther, A., et al., 2011. A review of the anthropogenic influence on biogenic secondary organic aerosol. *Atmos. Chem. Phys.* 11, 321–343.
- Hystad, P., Davies, H.W., Frank, L., Van Loon, J., Gehring, U., Tamburic, L., et al., 2014. Residential greenness and birth outcomes: Evaluating the influence of spatially correlated built-environment factors. *Environ. Health Perspect.* 122, 1095–1102.
- James, P., Hart, J.E., Banay, R.F., Laden, F., 2016. Exposure to greenness and mortality in a nationwide prospective cohort study of women. *Environ. Health Perspect.* 124, 1344–1352.
- Kahn, P.H., Kellert, S.R., 2002. *Children and Nature: Psychological, Sociocultural, and Evolutionary Investigations*. Massachusetts Institute of Technology, Cambridge.
- Kaplan, R., Kaplan, S., 1989. *The Experience of Nature: A Psychological Perspective*. Cambridge University Press, New York.
- Kaplan, S., 1995. The restorative benefits of nature: Toward an integrative framework. *J. Environ. Psychol.* 15, 169–182.
- Kawachi, I., Subramanian, S.V., Kim, D., 2008. *Social Capital and Health*. Springer, New York.
- Kellert, S.R., Wilson, E.O., 1993. *The Biophilia Hypothesis*. Island Press, Washington, DC.
- Kellert, S.R., 2005. *Building for Life: Designing and Understanding the Human-Nature Connection*. Island Press, Washington.
- Kesselmeier, J., Staudt, M., 1999. Biogenic volatile organic compounds (voc): An overview on emission, physiology and ecology. *J. Atmos. Chem.* 33, 23–88.
- Kuo, F.E., Taylor, A.F., 2004. A potential natural treatment for attention-deficit/hyperactivity disorder: Evidence from a national study. *Am. J. Public Health* 94, 1580–1586.
- Lachowycz, K., Jones, A.P., 2011. Greenspace and obesity: A systematic review of the evidence. *Obes. Rev.* 12, e183–e189.
- Lauritzen, O., Wu, J., Li, L., Milesi, C., 2013. Green spaces and pregnancy outcomes in southern california. *Health Place* 24, 190–195.
- Lovasi, G.S., Quinn, J.W., Neckerman, K.M., Perzanowski, M.S., Rundle, A., 2008. Children living in areas with more street trees have lower prevalence of asthma. *J. Epidemiol. Community Health* 62, 647–649.
- Lovasi, G.S., O'Neil-Dunne, J.P.M., Lu, J.W.T., Sheehan, D., Perzanowski, M.S., MacFaden, S.W., et al., 2013. Urban tree canopy and asthma, wheeze, rhinitis, and allergic sensitization to tree pollen in a New York city birth cohort. *Environ. Health Perspect.* doi:10.1289/ehp.1205513
- Maas, J., Van Dillen, S.M.E., Verheij, R.A., Groenewegen, P.P., 2009a. Social contacts as a possible mechanism behind the relation between green space and health. *Health Place* 15, 586–595.
- Maas, J., Verheij, R.A., de Vries, S., Spreeuwenberg, P., Schellevis, F.G., Groenewegen, P.P., 2009b. Morbidity is related to a green living environment. *J. Epidemiol. Community Health* 63, 967–973.
- Markevych, I., Thiering, E., Fuertes, E., Sugiri, D., Berdel, D., Koletzko, S., et al., 2014a. A cross-sectional analysis of the effects of residential greenness on blood pressure in 10-year old children: Results from the giniplus and lisapplus studies. *BMC Public Health* 14, 477.
- Markevych, I., Tiesler, C.M.T., Fuertes, E., Romanos, M., Dadvand, P., Nieuwenhuijsen, M.J., et al., 2014b. Access to urban green spaces and behavioural problems in children: Results from the giniplus and lisapplus studies. *Environ. Int.* 71, 29–35.
- Marmot, M., Allen, J., Bell, R., Bloomer, E., Goldblatt, P., 2012. Who European review of social determinants of health and the health divide. *Lancet* 380, 1011–1029.
- Martinez, F.D., 2014. The human microbiome. Early life determinant of health outcomes. *Ann. Am. Thorac. Soc.* 11, S7–S12.
- McCormack, G.R., Rock, M., Toohey, A.M., Hignell, D., 2010. Characteristics of urban parks associated with park use and physical activity: A review of qualitative research. *Health Place* 16, 712–726.
- McGrath, L.J., Hopkins, W.G., Hinckson, E.A., 2015. Associations of objectively measured built-environment attributes with youth moderate-vigorous physical activity: A systematic review and meta-analysis. *Sports Med.* 45, 841–865.
- Mygind, L., Kjeldsted, E., Hartmeyer, R., Mygind, E., Stevenson, M.P., Quintana, D.S., et al., 2019. Effects of public green space on acute psychophysiological stress response: A systematic review and meta-analysis of the experimental and quasi-experimental evidence. *Environ. Behav.* 0013916519873376.

- Nichani, V., Dirks, K., Burns, B., Bird, A., Morton, S., Grant, C., 2017. Green space and pregnancy outcomes: Evidence from growing up in new zealand. *Health Place* 46, 21–28.
- Nieuwenhuijsen, M.J., Khreis, H., 2016. Car free cities: Pathway to healthy urban living. *Environ. Int.* 94, 251–262.
- Nieuwenhuijsen, M.J., Khreis, H., Triguero-Mas, M., Gascon, M., Davdand, P., 2017. Fifty shades of green: Pathway to healthy urban living. *Epidemiology* 28, 63–71.
- Nieuwenhuijsen, M.J., 2020. Urban and transport planning pathways to carbon neutral, liveable and healthy cities; a review of the current evidence. *Environ. Int.* 140, 105661.
- Nowak, D.J., Crane, D.E., Stevens, J.C., 2006. Air pollution removal by urban trees and shrubs in the united states. *Urban Forest. Urban Green.* 4, 115–123.
- Organisation for Economic Cooperation and Development (OECD), 2011. Perspectives on global development 2012: Social cohesion in a shifting world. OECD Publishing, Paris.
- Paoletti, E., Bardelli, T., Giovannini, G., Pecchioli, L., 2011. Air quality impact of an urban park over time. *Procedia Environ. Sci.* 4, 10–16.
- Pilat, M.A., McFarland, A., Snelgrove, A., Collins, K., Waliczek, T.M., Zajicek, J., 2012. The effect of tree cover and vegetation on incidence of childhood asthma in metropolitan statistical areas of texas. *HortTechnology* 22, 631–637.
- Rojas-Rueda, D., Nieuwenhuijsen, M.J., Gascon, M., Perez-Leon, D., Mudu, P., 2019. Green spaces and mortality: A systematic review and meta-analysis of cohort studies. *Lancet Planet Health* 3, e469–e477.
- Rook, G.A.W., 2013. Regulation of the immune system by biodiversity from the natural environment: An ecosystem service essential to health. *Proc. Natl. Acad. Sci. U. S. A.* 110, 18360–18367.
- Schäffer, B., Brink, M., Schlatter, F., Vienneau, D., Wunderli, J.M., 2020. Residential green is associated with reduced annoyance to road traffic and railway noise but increased annoyance to aircraft noise exposure. *Environ. Int.* 143, 105885.
- Sugiyama, T., Leslie, E., Giles-Corti, B., Owen, N., 2008. Associations of neighbourhood greenness with physical and mental health: Do walking, social coherence and local social interaction explain the relationships? *J. Epidemiol. Community Health* 62, e9.
- Tallis, M., Taylor, G., Sinnett, D., Freer-Smith, P., 2011. Estimating the removal of atmospheric particulate pollution by the urban tree canopy of london, under current and future environments. *Landsc. Urban Plan.* 103, 129–138.
- Taylor, A.F., Kuo, F.E., Sullivan, W.C., 2001. Coping with add the surprising connection to green play settings. *Environ. Behav.* 33, 54–77.
- Taylor, A.F., Kuo, F.E., 2009. Children with attention deficits concentrate better after walk in the park. *J. Atten Disord.* 12, 402–409.
- Triguero-Mas, M., Davdand, P., Cirach, M., Martínez, D., Medina, A., Mompert, A., et al., 2015. Natural outdoor environments and mental and physical health: Relationships and mechanisms. *Environ. Int.* 77, 35–41.
- Ulrich, R., 1984. View through a window may influence recovery. *Science* 224, 224–225.
- UN Department of Economic and Social Affairs, 2015. World urbanization prospects; the 2014 revision. United Nations, New York.
- US Environmental Protection Agency, 2017. Green streets and community open space. Available from: <https://www.epa.gov/G3/green-streets-and-community-open-space> [accessed July 30 2017].
- van den Berg, A.E., van den Berg, C.G., 2011. A comparison of children with adhd in a natural and built setting. *Child Care Health Dev.* 37, 430–439.
- van den Bosch, M., Nieuwenhuijsen, M., 2017. No time to lose - green the cities now. *Environ. Int.* 99, 343–350.
- Vanaken, G., Danckaerts, M., 2018. Impact of green space exposure on children's and adolescents' mental health: A systematic review. *Int. J. Environ. Res. Public Health* 15, 2668.
- Vorholt, J.A., 2012. Microbial life in the phyllosphere. *Nat. Rev. Micro* 10, 828–840.
- Wells, N.M., 2000. At home with nature effects of greenness on children's cognitive functioning. *Environ. Behav.* 32, 775–795.
- WHO regional Office for Europe, 2016. Urban green spaces and health - a review of evidence. Available from: <http://www.euro.who.int/en/health-topics/environment-and-health/urban-health/publications/2016/urban-green-spaces-and-health-a-review-of-evidence-2016> [accessed July 31 2017].
- Wilson, E.O., 1984. *Biophilia*. Harvard University Press, Cambridge.
- Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., et al., 2012. Human gut microbiome viewed across age and geography. *Nature* 486, 222–227.
- Zhan, Y., Liu, J., Lu, Z., Yue, H., Zhang, J., Jiang, Y., 2020. Influence of residential greenness on adverse pregnancy outcomes: A systematic review and dose-response meta-analysis. *Sci. Total Environ.* 718, 137420.
- Zijlema, W.L., Avila-Palencia, I., Triguero-Mas, M., Gidlow, C., Maas, J., Kruijs, H., et al., 2018. Active commuting through natural environments is associated with better mental health: Results from the phenotype project. *Environ. Int.* 121, 721–727.

Transport Access and Health

Alireza Ermagun, Department of Civil and Environmental Engineering, Mississippi State University, Mississippi State, MS, United States

© 2021 Elsevier Ltd. All rights reserved.

From Access to Healthiness	335
Access-Related Transport Barriers to Healthcare	336
Equity of Access to Health Care	338
Conclusion	339
Acknowledgment	339
References	339
Further Reading	340

From Access to Healthiness

From its source Latin *accedere*, the assimilated form of *ad* and *cedere*, access means “to approach.” It conveys “habit or power of getting into the presence of someone or something.” The contemporary meaning is different, but not too far from the origin. Access means freedom (Access, 2020). It offers neither what people will do nor what people want to do, rather what people could do. Access is a product of mobility and place, and immediately relates to the transport network and the relative location of human activities and housing. Mobility means the ability to move easily. It reflects network speed or travel time, and is complementary to access, rather than competitive. Mobility improvements can contribute to improving access. Yet mobility is just a fraction of the access concept. An illustrative example suffices: I draw it from my experience living in Chicago, the third-most-populous city in the United States, and the Starkville, a marquee American college town located in east-central Mississippi. Driving 30 min from downtown Starkville, a traveler can almost cover the entire city, but reaches only two Starbucks Coffeehouses. Conversely, while driving 30 min in downtown Chicago does not get you far, you can reach more than a dozen Starbucks Coffeehouses. This opens more opportunities including jobs, groceries, health care, education, and recreational facilities with a direct and indirect influence on health. Fig. 1 depicts how access to opportunities affects maintaining a healthy life style and well-being. Research declares access to hospitals (Love and Lindquist, 1995), primary care physicians (Schuurman et al., 2010), pharmacies (Cooper et al., 2009), grocery stores (Zenk et al., 2005), parks (Roemmich et al., 2006), trails (Wolf and Wohlfart, 2014), and playgrounds (Smoyer-Tomic et al., 2004) contribute to personal well-being.

Should not we deliver access to opportunities equitably? From the Civil War to the Civil rights movement, the signing of the Treaty of Paris 1783 to pulling out of the Paris Agreement in 2017, the United States has spent much of its existence grappling with the problem of equity. Throughout the 20th century, minority groups and low-income populations fought for equitable access to health care, jobs, education, and public transit. In 1955, African Americans in Montgomery, Alabama began The Bus Boycott, which ended up going before the Supreme Court and integrating transport across the United States. The Bus Boycott was the first ripple in a wave that led to the signing of the Civil Rights Act in 1964, under which, “All persons shall be entitled to the full and equal enjoyment of the goods, services, facilities, and privileges, advantages, and accommodations of any place of public accommodation. .. without discrimination or segregation on the ground of race, color, religion, or national origin.” In the 1980s, the fight for equity turned toward the environment, as studies commenced identifying poor and nonwhite residential communities, mainly in the South, as more likely to house hazardous waste facilities. However, another decade passed until this uneven dispersal of hazardous materials began to be addressed. In 1994, President Clinton signed Executive Order 12898 into law, ordering Federal agencies to prioritize environmental justice for poor and minority communities. The quest for justice escalated on April 15, 1997, when the U.S. Department of Transportation released DOT Order 5610.2 that reaffirmed its responsibility to bring justice to the transport industry.

Though history has made progress in the equity movement, there is still a sharp divide between income classes, minorities, the elderly, and gender in terms of equity of access to opportunities (Ermagun and Tilahun, 2020). For many this means the working poor, African Americans, Hispanics, the elderly, and people with disabilities are left in communities that have been segregated into the “haves”: affordable and efficient transport to work, health care, healthy food, education, and recreation, and the “have nots.” Transport authorities, indeed, have failed to equitably distribute transport access and have worked for the benefit of car owners. This biased approach begets transport-related social exclusion. This is a process by which people are limited to participate in social activities, in part due to either the lack of access to opportunities or mobility impairment (Preston and Rajé, 2007). Regardless of the opportunity, transport-related social exclusion leads to diminished levels of well-being particularly for underserved and underrepresented groups as illustrated in Fig. 1. This highlights the concern of transport access as a social equity issue.

Access-related equity is split up into (1) equity and social class and (2) equity and mobility need (Litman, 2002). The former aims to disclose the disparity of access in groups that differ in social, demographic, and economic characteristics. How access is distributed among low- and high-income groups targets this access-related equity category. The latter aims to show to what extent access meets the needs of the people with mobility impairments and special constraints. This category targets how access is provided for people

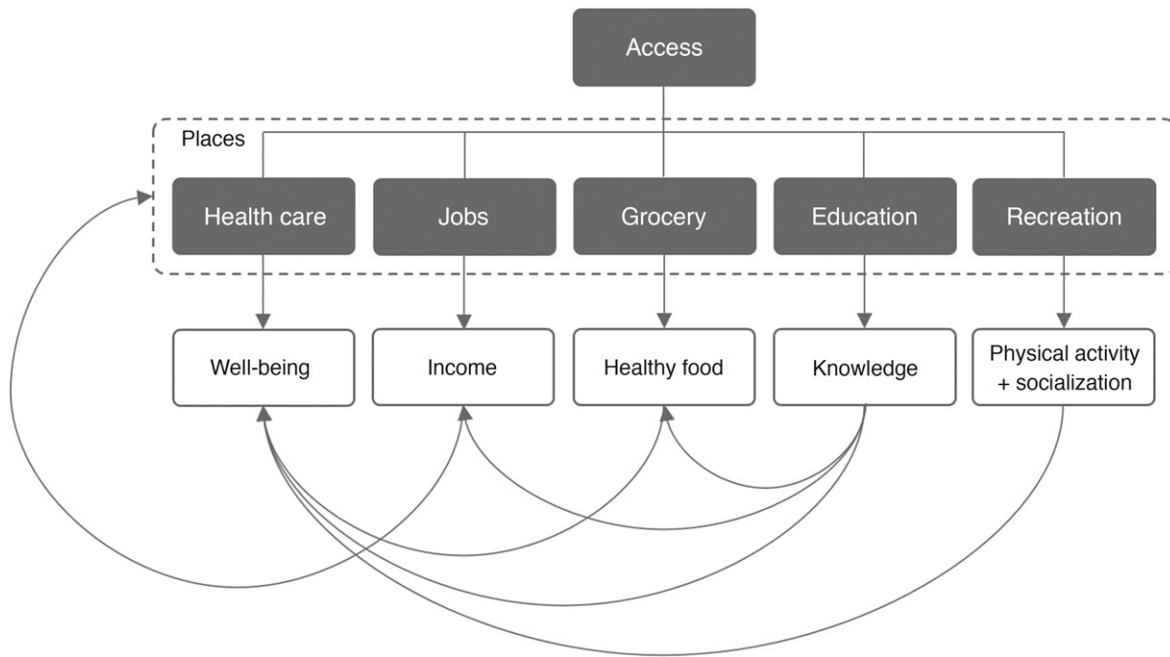


Figure 1 Direct and indirect relation between transport access to opportunities and health.

with disabilities. Access-related equity has been discussed at some length in the transport literature, leaning more toward employment opportunities (Borowski et al., 2018; Pyrialakou et al., 2016; Sanchez, 1999; Thakuriah and Metaxatos, 2000). Little is known about the equity of access to health-related services, despite the fact that transport access is often perceived as a major barrier to medication and health care access. A 2005 study estimates Americans who miss or delay nonemergency medical treatment every year due to transport barriers at 3.6 million (Wallace et al., 2005).

Access-Related Transport Barriers to Healthcare

It is conceived that inequity of access to health-related services is a consequence of several factors operating at a number of different levels as highlighted in Fig. 2. These factors range from socioeconomic and demographic characteristics of individuals to those operating at mobility and place levels. Inequity of access to health-related services is often reflected either as a lack of vehicle access or mobility impairment to reach health care facilities because of social status or a lacking access to health care facilities as a product of mobility and place.

Research unanimously indicates that a lack of vehicle access or mobility impairment is negatively correlated with the utilization of health care. Arcury et al. (2005) surveyed 1059 people from 12 counties in North Carolina. Of the respondents, consisting of 948 nonminority and 111 African Americans, individuals who had a driver's license were over two times more likely to visit a doctor regularly versus those who do not have a driver's license. Individuals who had friends and family they could rely on for transport

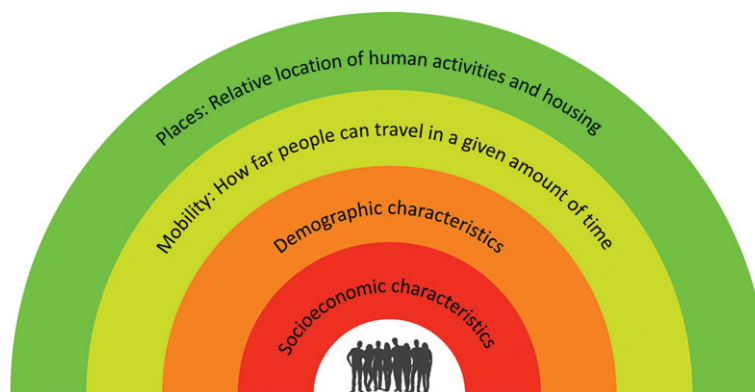


Figure 2 Determinants of inequity of access to health-related services.

Table 1 Summary of studies exploring equity of transport access to health care

Article	Area	Access measure	Travel threshold	Travel mode	Cohorts
Ermagun and Tilahun (2020)	US	Cumulative opportunities	30 min	PT	Minority, low-income, elderly, disabled, low-educated, carless
Pu et al. (2020)	Congo	Raster-based accessibility	2 h	W, M, PT	General population
Boisjoly et al. (2020)	Canada	2SFCA	45 min	PT	Vulnerable
Tao and Cheng (2019)	China	2SFCA	124 min	PT, PC	Elderly
Campbell et al. (2019)	Nairobi	Gravity	60 min	W, P, PC	Low and middle incomes
Kelobonye et al. (2019)	Australia	2SFCA	15 min	PT, PC	Residents location
Guimarães et al. (2019)	Brazil	Focus groups	–	W, PT	Low incomes
Mayaud et al. (2019)	Cascadia	Catchment area	30 min	PT	Low incomes, elderly
Madill et al. (2018)	Australia	Travel time	–	PT, PC	Residents location
Agbenyo et al. (2017)	Ghana	Travel distance	5 km	W, M, B, PT	General population
Higgs et al. (2017)	UK	2SFCA	15 min	PT, PC	Elderly
Ahern and Hine (2015)	Ireland	Focus groups	–	PT, PC	Elderly
Wang (2011)	Canada	2SFCA	11 min	PC	Immigrants
Harris et al. (2011)	South Africa	National Survey	–	W, PT, PC	Africans, poor, uninsured and rural
Paez et al. (2010)	Canada	Cumulative opportunities	–	PC	Elderly
Dai (2010)	US	2SFCA	15, 20, 25, 30 min	PC	African American segregation
Lovett et al. (2002)	UK	Travel time	0–40 min	PT, PC	Residents location

2SFCA, two-step floating catchment area; B, bike; M, motor; P, paratransit; PT, public transit; W, walking; PC, private car.

were more than 1.5 times as likely to see the doctor as those without a ride. Yang et al. (2006) conducted a cross-sectional study of low-income families, all of whom have State Children's Health Insurance Programs for their children, to determine what factors caused a caregiver to keep or miss an appointment. Of the caregivers who canceled their appointment, 50% said a transport issue, such as not having a ride, not having a public transit option, or traffic as the reason for canceled appointments.

Lack of access to health care facilities is often explored through the analysis of geographical location of people and their distance or travel time to health care facilities. The results are mixed. Most research shows rural patients have greater transport barriers to health care facilities than urban patients and highlights distance is a barrier. Few studies declare that there is no significant relationship between health care utilization and either geographical location of people or their distance to health care facilities. Blazer et al. (1995) collected a stratified random sample of 4162 residents of one urban and four rural counties of North Carolina, which was representative of more than 28,000 residents 65 years of age and older. They found that rural residents were more likely to use alternative medical facilities or to forgo treatment until it was absolutely necessary due to problems being transported to a medical facility. Participants cited the cost of traveling to the health care facility and, in many cases, having to depend on friends and family to transport them, was an influential factor. Probst et al. (2007) drawn data from 2001 US National Household Travel Survey to analyze effects of residence and race on burden of travel for care. They found rural residents, on average, traveled about 15 km (9 miles) further for care than urban residents, taking about six additional minutes to complete their trip. In a study of special needs families, Skinner and Slifkin (2007) showed children in rural areas were more likely than urban children to miss a treatment or appointment due to lack of transport to a care facility. Using a subsample of 102 family caregivers from 31 states, Kruzich et al. (2003) highlighted that since clinics emphasize parental participation in the therapies, facilities that are further away require the caregiver to take time away from their job which in turn increases their financial burden. In a survey of 781 diabetic adults, Littenberg et al. (2006) discussed that the farther the patient lives from their medical facility, the less likely they are to receive insulin treatment. This strongly affects diabetics who live in rural regions quality of life, as it becomes harder to control glucose levels. Strauss et al. (2006) had similar findings performing a cross-sectional analysis of 973 adults with diabetes in Vermont, New Hampshire, and northern New York. They pinpointed for every 35 km (22 miles) a patient would have to drive, their glucose measure increases by 0.25%.

Transport barriers disproportionately affect different demographic and socioeconomic cohorts. Minority and low-income populations, the elderly, disabled, low-educated, female, and people without a car face transport barriers more than their counterparts. Probst et al. (2007) showed African Americans were more likely than whites to have trips for care that required more than 30 min of travel, even though the distance traveled is much shorter. They posit this is because African Americans are more likely to rely on public transit as their main source for transport. Wallace et al. (2005) estimated the transport-disadvantaged population using data from (1) the Bureau of Transportation Statistics, (2) U.S. 2002 National Health Interview Survey, and (3) Medical Expenditure Panel Survey. They concluded that the transport-disadvantaged population are poorer, older, female, minority, and less educated. In addition, they showed children who lack access to medical care due to transport barriers are more concentrated in urban areas and people with chronic conditions and multiple diseases experience greater difficulty in accessing health care due to the lack of nonemergency medical transport. Conducting a telephone survey among 2097 rural community-dwelling elders in West Texas, Borders (2004) found that Hispanic elders are more likely to live out in rural areas and be more dependent on family for transport. This not only limits how often they can see their doctor but also the privilege of seeing the same doctor for continuing care.

Equity of Access to Health Care

The well-being of individuals, as alluded to in [Fig. 1](#), is associated with access. An underlying tone to the access to health care is the socioeconomic and demographic attributes amongst the disadvantaged and affluent groups in the society. Research has focused on how access to health care is distributed among different sociodemographic groups. [Table 1](#) lists the related research. It is not the intent of the summary of literature presented in [Table 1](#) to be all inclusive. Rather, its purpose is to provide a broad range of research time frames, locations, modes of transport, access thresholds, and research methodologies used to measure the equity of access to health care over the past 20 years. The geographical scope of studies specifies that the equity of access to health care is a need perceived around the globe spanning from Asia ([Tao and Cheng, 2019](#)), Africa ([Harris et al., 2011](#)), and Europe ([Higgs et al., 2017](#)) to America ([Mayaud et al., 2019](#)), and Australia ([Madill et al., 2018](#)). Research studies the equity of access to health care objectively by measuring access to health care ([Boisjoly et al., 2020](#); [Campbell et al., 2019](#); [Ermagun and Tilahun, 2020](#); [Kelobonye et al., 2019](#)) or subjectively by conducting focus groups or using national and local surveys ([Ahern and Hine, 2015](#); [Guimarães et al., 2019](#); [Harris et al., 2011](#)).

Measuring access to health care methods ranges from abstracted Euclidean or Manhattan distance to two-step floating catchment area and gravity measures. [Boisjoly et al. \(2020\)](#) quantified access to healthcare services by public transport in eight metropolitan regions of Canada using a two-step floating catchment area method. The study areas covered different population sizes from small to large regions, selected based on data availability with health care data from Canadian Institute for Health Information and the transport data from the General Transit Feed Specification. They showed individuals residing in large metropolitan areas were subjected to low levels of access to health care facilities due to spatial dispersion of the areas that affect transport. [Mayaud et al. \(2019\)](#) compared the equity of access to health care facilities by public transit in Portland, Seattle, and Vancouver, located in the Cascadia region. A catchment analysis was performed to analyze the socioeconomic makeup and determine access to health-related facilities. The findings indicated that Portland displays the highest inequity in health care facilities access, and Vancouver the least, owing to Vancouver's relatively compact geographic area and high population density. Over 75% of seniors are socially excluded from hospitals and over 50% from clinics in Portland, compared to 30% and 3%, respectively, in Vancouver. The inequality of access to health care facilities also exists among low-income neighborhoods in all three cities. This translates into proportionally higher transport costs for low-income area residents compared with high-income area residents. They argued access to health facilities in Portland are more centralized due to the setup of its public transport network. For Seattle and Vancouver, however, the public transport network is setup to distribute access more evenly. [Higgs et al. \(2017\)](#) argued that access to general practitioners (GP) is linked to the availability of transport amongst other problems such as socioeconomic factors, community isolation, and spatial distribution. Individuals who lack vehicle ownership, are poverty stricken, or are elderly rely on the use of public transport to access healthcare facilities. Using the South Wales case, they employed Enhanced Two-Step Floating Catchment Area techniques for measuring access GP surgeries by car and bus within a 15-min threshold time. Evidence showed there was a negative correlation between the percentage of persons aged over 65 years and distance to GP by either car or bus. For every percentage increase in persons aged over 65 years the distance to GP decreases by 0.09 km and 0.13 km by car and bus, respectively. There was also a negative association between percentage of persons aged over 65 years and the number of GPs located within a 15-min threshold traveling by car, but the association was not significant when traveling by bus. While mode of travel was an important consideration for determining access, the study suggested other factors provide more precise means to measure access—distance, density, and ease of access.

Measuring access to health care is often limited to the number of cumulative health care facilities reachable under an impedance function as the measure of access. This immediately ignores “trips not taken” due to barriers, which might be captured by subjective studies. Measures used in quantifying the equity of access to health care are not satisfactory tools to detect trips not taken and barriers impacting the actual behavior of individuals. In contrast, focus groups and designed surveys afford explanatory information about how transport, socioeconomic, and demographic barriers exclude vulnerable populations from reaching to health care facilities even if they are physically located in their catchment areas. [Guimarães et al. \(2019\)](#) developed a conceptual framework on health care access needs, which identified the land use interaction and interconnectivity between transport and health care as a meso-level satisfier. Transport elements were comprised of safety, comfort, speed, cost, information, public transit stops and routes, public transit service times, public transit frequency, and public transit punctuality. Health care elements were comprised of information, location, opening times, waiting times, adequacy, quality of care, referral system, coordination, and continuity. They used focus groups as a medium to collect information of 114 residents in 12 low-income neighborhoods in São Paulo, Brazil. The outcome of the focus group led to the frequency of the themes considered relevant to the participants from the standpoint of access to health care: proximity and remoteness, walking safety, public transport services, personal security, and quality of health care services. Respondents reported concerns of deficiencies and inadequacy of the transport facilities with issues not limited to lack of sidewalks, uneven walking surfaces, narrow pedestrian paths, and lack of traffic signals. These poor designs pose limitation to the safety of all individuals using the infrastructure—young, old, and disabled. Residents of low-income areas rely on public transit systems to get around for different purposes and are fully aware of the issues of mobility and access that arises. Lacking car ownership, people are left to grapple with the issues of availability, affordability, comfort, reliability, and overcrowding. Issues of crime and violence are dominant in low-income communities as revealed from the focus group. Residents are left concerned about their personal security especially for pedestrians who fear assault and public transit users who fear sexual assault as the crimes are dependent on the travel mode.

Conclusion

Access is the ease of reaching valued destinations by different modes of travel to accomplish the desired services and activities. As the ultimate goal in transport planning, access has a wide variety of factors that can affect it not limited to mobility, proximity, transport system connectivity, affordability, convenience, and social acceptability (Litman, 2008). The issues of transport are not just a mere dichotomy between health care and transport, or a trichotomy of places, socioeconomic, and demographic attributes. Barriers to health care are not a standalone issue. Individual level characteristics, mobility, and land use attributes affect access to health care. How individuals get access to health care is linked to how they get to these facilities. Most are not able to afford using private modes of travel and revert to public transit or walking. Both travel modes have challenges related to them, with more pockets in rural areas and areas with a higher percentage of minority, low-income families, the elderly, people with disability, and low-educated people. Individual level characteristics could affect how the marginalized live and survive. The challenges faced by the marginalized are greater compared to the affluent.

In the development of policy or legislation, vulnerable people were never completely ignored, but their needs were never taken seriously. They have been condemned to have a decent “level-of-access.” The initial stages of planning of cities and transport are critical to ensuring vulnerable people are accounted for to enable access. Policies should enable connected land use, transport, and the service system to allow for better access to much needed health and wellness services. Through studies, policies can be revised and recommendations can be made for improvements for planning transport to improve health care access for vulnerable people. It, however, must be noticed that providing access is a wicked planning problem and the contrast between the haves and the have nots often persists. First, there are mismatched incentives between transport networks and land regulations. There is no guarantee that the efficient choice of transport networks and land is harmonious. Second, transport networks are often regulated by the public sector, while the location of human activities is owned by the private sector. As transport networks are not built for free, the public sector might pursue investments that provide high access per dollar returns. If this happens, investments do not necessarily benefit people equitably. The first chapter of “A Political Economy of Access” notes this trade-off through the lens of equity and efficiency of access (Levinson and King, 2019).

It also must be noticed that the noisy confusion in the equity of access debates is in part due to understanding equity as the equal distribution of physical access to health care. However, what needs to be delivered equitably is not just the number of cumulative health care facilities reachable under an impedance function as the measure of access, but the means to lift barriers in the ability to reach health care facilities. This issue does not necessarily need to be tackled by changes in transport infrastructures. Supporting online consultations delivery initiatives or distributing and stocking healthy foods in local shops are efficient solution examples to lift transport barriers, rather than investing in transport infrastructure. The former initiative eliminates the need of many patients to travel to health care facilities and the latter initiative eliminates the barrier of distance to healthy foods for people with low mobility.

Overall, the intent here has not been to provide techniques to solve the actual problem of transport access to healthcare. Rather, the intent has been to emphasize that the problem exists, even in areas which involve some planning, and to distinguish facts from values. The intent has been to help city authorities, planners, and practitioners work with facts to effectively reach their values.

Acknowledgment

The author would like to thank Jacquelyn Erinne for her assistance in collecting information. All opinions are the responsibility of the author.

References

- Access, 2020. In Merriam-Webster.com. Retrieved from <https://www.merriam-webster.com/dictionary/access>.
- Agbenyo, F., Nunbogu, A.M., Dongzagla, A., 2017. Accessibility mapping of health facilities in rural Ghana. *J. Transport Health* 6, 73–83.
- Ahern, A., Hine, J., 2015. Accessibility of health services for aged people in rural Ireland. *Int. J. Sustain. Transport* 9 (5), 389–395.
- Arcury, T.A., Preisser, J.S., Gesler, W.M., Powers, J.M., 2005. Access to transportation and health care utilization in a rural region. *J. Rural Health* 21 (1), 31–38.
- Blazer, D.G., Landerman, L.R., Fillenbaum, G., Horner, R., 1995. Health services access and use among older adults in North Carolina: urban vs rural residents. *Am. J. Public Health* 85 (10), 1384–1390.
- Boisjoly, G., Deboosere, R., Wasfi, R., Orpana, H., Manaugh, K., Buliung, R., El-Geneidy, A., 2020. Measuring accessibility to hospitals by public transport: an assessment of eight Canadian metropolitan regions. *J. Transport Health* 18, 100916.
- Borders, T.F., 2004. Rural community-dwelling elders' reports of access to care: are there Hispanic versus non-Hispanic white disparities? *J. Rural Health* 20 (3), 210–220.
- Borowski, E., Ermagun, A., Levinson, D., 2018. Disparity of Access: Variations in Transit Service by Race, Ethnicity, Income, and Auto Availability.
- Campbell, K.B., Rising, J.A., Klopp, J.M., Mbilo, J.M., 2019. Accessibility across transport modes and residential developments in Nairobi. *J. Transport Geogr.* 74, 77–90.
- Cooper, H.L., Bossak, B.H., Tempalski, B., Friedman, S.R., Des Jarlais, D.C., 2009. Temporal trends in spatial access to pharmacies that sell over-the-counter syringes in New York City health districts: relationship to local racial/ethnic composition and need. *J. Urban Health* 86 (6), 929–945.
- Dai, D., 2010. Black residential segregation, disparities in spatial access to health care facilities, and late-stage breast cancer diagnosis in metropolitan Detroit. *Health Place* 16 (5), 1038–1052.
- Ermagun, A., Tilahun, N., 2020. Equity of transit accessibility across Chicago. *Transport. Res. Part D: Transport Environ.* 86, 102461.
- Guimarães, T., Lucas, K., Timms, P., 2019. Understanding how low-income communities gain access to healthcare services: a qualitative study in São Paulo, Brazil. *J. Transport Health* 15, 100658.

- Harris, B., Goudge, J., Ataguba, J.E., McIntyre, D., Nxumalo, N., Jikwana, S., Chersich, M., 2011. Inequities in access to health care in South Africa. *J. Public Health Policy* 32 (1), S102–S123.
- Higgs, G., Zahnow, R., Corcoran, J., Langford, M., Fry, R., 2017. Modelling spatial access to general practitioner surgeries: does public transport availability matter? *J. Transport Health* 6, 143–154.
- Kelobonye, K., McCarney, G., Xia, J.C., Swapan, M.S.H., Mao, F., Zhou, H., 2019. Relative accessibility analysis for key land uses: a spatial equity perspective. *J. Transport Geogr.* 75, 82–93.
- Kruzich, J.M., Jivanjee, P., Robinson, A., Friesen, B.J., 2003. Family caregivers' perceptions of barriers to and supports of participation in their children's out-of-home treatment. *Psychiatric Services* 54 (11), 1513–1518.
- Levinson, D.M., King, D.A., 2019. A Political Economy of Access: Infrastructure, Networks, Cities, and Institutions. Network Design Lab.
- Litman, T., 2002. Evaluating transportation equity. *World Transport Policy Pract.* 8 (2), 50–65.
- Litman, T., 2008. Evaluating Accessibility for Transportation Planning. Victoria Transport Policy Institute, Victoria, Canada.
- Littenberg, B., Strauss, K., MacLean, C.D., Troy, A.R., 2006. The use of insulin declines as patients live farther from their source of care: results of a survey of adults with type 2 diabetes. *BMC Public Health* 6 (1), 198.
- Love, D., Lindquist, P., 1995. The geographical accessibility of hospitals to the aged: a geographic information systems analysis within Illinois. *Health Services Res.* 29 (6), 629.
- Lovett, A., Haynes, R., Sünnerberg, G., Gale, S., 2002. Car travel time and accessibility by bus to general practitioner services: a study using patient registers and GIS. *Soc. Sci. Med.* 55 (1), 97–111.
- Madill, R., Badland, H., Mavoa, S., Giles-Corti, B., 2018. Comparing private and public transport access to diabetic health services across inner, middle, and outer suburbs of Melbourne, Australia. *BMC Health Services Res.* 18 (1), 286.
- Mayaud, J.R., Tran, M., Nuttall, R., 2019. An urban data framework for assessing equity in cities: comparing accessibility to healthcare facilities in Cascadia. *Comput. Environ. Urban Syst.* 78, 101401.
- Paez, A., Mercado, R.G., Farber, S., Morency, C., Roorda, M., 2010. Accessibility to health care facilities in Montreal Island: an application of relative accessibility indicators from the perspective of senior and non-senior residents. *Int. J. Health Geogr.* 9 (1), 52.
- Preston, J., Rajé, F., 2007. Accessibility, mobility and transport-related social exclusion. *J. Transport Geogr.* 15 (3), 151–160.
- Probst, J.C., Laditka, S.B., Wang, J.Y., Johnson, A.O., 2007. Effects of residence and race on burden of travel for care: cross sectional analysis of the 2001 US National Household Travel Survey. *BMC Health Services Res.* 7 (1), 40.
- Pu, Q., Yoo, E.H., Rothstein, D.H., Cairo, S., Malemo, L., 2020. Improving the spatial accessibility of healthcare in North Kivu, Democratic Republic of Congo. *Appl. Geogr.* 121, 102262.
- Pyrialakou, V.D., Gkritza, K., Fricker, J.D., 2016. Accessibility, mobility, and realized travel behavior: assessing transport disadvantage from a policy perspective. *J. Transport Geogr.* 51, 252–269.
- Roemmich, J.N., Epstein, L.H., Raja, S., Yin, L., Robinson, J., Winiewicz, D., 2006. Association of access to parks and recreational facilities with the physical activity of young children. *Prevent. Med.* 43 (6), 437–441.
- Sanchez, T.W., 1999. The connection between public transit and employment: the cases of Portland and Atlanta. *J. Am. Plan. Assoc.* 65 (3), 284–296.
- Schuurman, N., Berube, M., Crooks, V.A., 2010. Measuring potential spatial access to primary health care physicians using a modified gravity model. *The Canadian Geographer/Le Géographe Canadien* 54 (1), 29–45.
- Skinner, A.C., Slifkin, R.T., 2007. Rural/urban differences in barriers to and burden of care for children with special health care needs. *J. Rural Health* 23 (2), 150–157.
- Smoyer-Tomic, K.E., Hewko, J.N., Hodgson, M.J., 2004. Spatial accessibility and equity of playgrounds in Edmonton, Canada. *Canadian Geographer/Le Géographe canadien* 48 (3), 287–302.
- Strauss, K., MacLean, C., Troy, A., Littenberg, B., 2006. Driving distance as a barrier to glycemic control in diabetes. *J. Gen. Intern. Med.* 21 (4), 378.
- Tao, Z., Cheng, Y., 2019. Modelling the spatial accessibility of the elderly to healthcare services in Beijing, China. *Environ. Plan. B: Urban Anal. City Sci.* 46 (6), 1132–1147.
- Thakuriah, P., Metaxatos, P., 2000. Effect of residential location and access to transportation on employment opportunities. *Transport. Res. Rec.* 1726 (1), 24–32.
- Wallace, R., Hughes-Cromwick, P., Mull, H., Khasnabis, S., 2005. Access to health care and nonemergency medical transportation: two missing links. *Transport. Res. Rec.* 1924 (1), 76–84.
- Wang, L., 2011. Analysing spatial accessibility to health care: a case study of access by different immigrant groups to primary care physicians in Toronto. *Ann. GIS* 17 (4), 237–251.
- Wolf, I.D., Wohlfart, T., 2014. Walking, hiking and running in parks: a multidisciplinary assessment of health and well-being benefits. *Landsc. Urban Plan.* 130, 89–103.
- Yang, S., Zarr, R.L., Kass-Hout, T.A., Kourosh, A., Kelly, N.R., 2006. Transportation barriers to accessing health care for urban children. *J. Health Care Poor Underserved* 17 (4), 928–943.
- Zenk, S.N., Schulz, A.J., Israel, B.A., James, S.A., Bao, S., Wilson, M.L., 2005. Neighborhood racial composition, neighborhood poverty, and the spatial accessibility of supermarkets in metropolitan Detroit. *Am. J. Public Health* 95 (4), 660–667.

Further Reading

- Cooper, E., Gates, S., Grollman, C., Mayer, M., Davis, B., Bankiewicz, U., Khambhaita, P., 2019. Transport, health, and wellbeing: An evidence review for the Department for Transport. *Nat Cen Social Res.*
- Di Ciommo, F., Dupont-Kieffer, A., Lucas, K., Martens, K. (Eds.), 2019. Measuring Transport Equity. Elsevier Science Publishing Company Incorporated.
- Ermagun, A., Levinson, D., 2017. Transit makes you short: On health impact assessment of transportation and the built environment. *J. Transport Health* 4, 373–387.
- Gates, S., Gogescu, F., Grollman, C., Cooper, E., Khambhaita, P., 2019. Transport and inequality: an evidence review for the Department for Transport. *NatCen Social Res.*
- Lee, R.J., Sener, I.N., 2016. Transportation planning and quality of life: Where do they intersect? *Transport Policy* 48, 146–155.
- Levinson, D.M., King, D.A., 2019. A Political Economy of Access: Infrastructure, Networks, Cities, and Institutions. Network Design Lab.
- Neutens, T., 2015. Accessibility, equity and health care: review and research directions for transport geographers. *J. Transport Geogr.* 43, 14–27.
- Syed, S.T., Gerber, B.S., Sharp, L.K., 2013. Traveling towards disease: transportation barriers to health care access. *J. Commun. Health* 38 (5), 976–993.
- van Gaans, D., Dent, E., 2018. Issues of accessibility to health services by older Australians: a review. *Public Health Rev.* 39 (1), 20.
- Van Wee, B., Geurs, K., 2011. Discussing equity and social exclusion in accessibility evaluations. *Eur. J. Transport Infrastruc. Res.* 11 (4).
- Xiao, C., Goryakin, Y., Cecchini, M., 2019. Physical activity levels and new public transit: a systematic review and meta-analysis. *Am. J. Prevent. Med.* 56 (3), 464–473.

Social Exclusion and Health, Related to Transport

Roger L. Mackett, Centre for Transport Studies, University College London, London, Great Britain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	341
The Nature of Social Exclusion	341
The Relationship Between Social Exclusion and Health	342
Transport in the Relationship Between Social Exclusion and Health	342
Social Exclusion and the Ability to Travel	343
Barriers to Travel	343
The Role Transport Plays in the Relationship Between Social Exclusion and Health	343
Access	344
Travel as a Form of Physical Activity	344
The Externalities of Transport	344
Summing Up	344
Biography	345
References	345
Further Reading	346

Introduction

“Social exclusion” is a term first coined by Lenoir (1974) in France. It describes the conditions that some people find themselves in, living outside society and so not able to enjoy many of the opportunities that it offers in terms of material goods and support, when required, in return for contributing to it through labour and taxation. Society offers the opportunity to participate in activities to help maintain good health. It also provides facilities to aid recovery from illness, and support when ill. Those who are excluded from society may not be able to enjoy such benefits and so may have poorer health than other people. Part of the reason for this is that it is necessary to travel to receive some medical treatment, and to participate in some healthy activities. Those who are excluded from society may not have access to travel facilities for a variety of reasons. The purpose of this article is to examine evidence that underpins the relationship between social exclusion and health and the role that transport plays in this.

The Nature of Social Exclusion

There have been a number of reports and academic papers about social exclusion, and they use various definitions of social exclusion. O'Donnell et al. (2018) list nine different definitions of it and six of its converse, social inclusion. They point out that some focus on the problems associated with not being included in society while others mention the various aspects of life that people are excluded from, such as a range of employment and education opportunities. The definitions are at various levels of abstraction and consider various levels at which social exclusion can operate. The common point about all the definitions is that they reflect the multi-dimensional aspect of social exclusion. Kenyon et al. (2002) list the following dimensions:

- Economic
- Societal
- Social networks
- Organized political (ability to influence decision making at an organised level)
- Personal political (ability to make decisions over own life)
- Personal
- Living space
- Temporal
- Mobility

For each of these, they list the potentially exclusionary factors.

In this article, the definition used by the authors of the various sources of evidence will be accepted as being appropriate for the issues being addressed, but it should be borne in mind that there may be subtle differences that reflect the context of the work, and possibly, the disciplinary background of the authors. The key point is that people who are socially excluded do not have the ability, opportunities and resources to participate fully in the opportunities offered by society that the individuals wish to take advantage of (or would if they were aware of them).

The Relationship Between Social Exclusion and Health

Article 12 of the United Nations International Covenant on Economic, Social and Cultural Rights says “The States Parties to the present Covenant recognize the right of everyone to the enjoyment of the highest attainable standard of physical and mental health” (United Nations, 1966). While everyone may have the right to good health, it is quite likely that some people are unable to achieve it because they cannot access the facilities and opportunities to enable them to do so because they are excluded from society.

The World Health Organization (WHO) set up the Social Exclusion Knowledge Network (SEKN) that produced a report (Popay et al., 2008) that considered the nature of social exclusion and developed a conceptual framework for understanding social exclusion in the context of health inequalities. The report also examined evidence about some existing policies and actions being used to address social exclusion around the world. The report describes how social exclusion processes lead to health inequities as a result of a continuum of inclusion/exclusion characterized by inequalities in:

- Access to resources (means that can be used to meet human needs)
- Capabilities (the relative power people have to utilize the resources available to them)
- Rights

These operate across four dimensions: economic, political, social, and cultural. The report argues that social exclusion influences health directly, through its manifestations in the health system, and indirectly, by affecting economic and other social inequalities that influence health. These inequalities then contribute to social exclusion processes, creating a vicious circle.

The World Health Organization (2010a) argues that health systems have a key role in addressing the relationship between poverty, social exclusion and health through four mechanisms: financing, stewardship, service delivery and resource generation. Financing helps by ensuring universal availability of the services, stewardship involves engaging with other sectors that influence health, service delivery helps to ensure that the health needs of socially excluded people are targeted, and resource generation contributes to health equity by ensuring that education of future health and medical professionals includes understanding of the health needs of people who are socially excluded and how these can be addressed.

In 2008, the Tallinn Charter (World Health Organization, 2008) was endorsed by European ministers of health. It is a commitment by the 53 Member States of the WHO in Europe to ensure that health systems pay due attention to the needs of vulnerable people. Not only did it contain a commitment to providing good quality health services for all, including vulnerable people, in response to their needs, it also acknowledged the need to enable people to make healthy lifestyle changes. This implies that people who are socially excluded are not only entitled to access good quality services, but also, have the opportunity to have healthy lifestyles, both positive, such as through health enhancing physical activity, and by avoiding factors that can adversely affect health. The evidence below suggests that this is not always the case. The recognition of the commitment across Europe to adapt health systems to address the needs of people who are socially excluded is illustrated by the twenty-two cases studies described in a report by the World Health Organization (2010b).

Further evidence of the international recognition of the relationship between social exclusion and health is presented by van Bergen et al. (2019) who examined 22 case studies of the links in EU and OECD countries. They found an association between high levels of social exclusion (or low levels of social inclusion) and adverse mental and general health. For physical health, the evidence was unclear. They also looked at the evidence for people in groups at high risk of social exclusion (adults in HIV treatment, problematic drug users, and single mothers) and found an association between high social exclusion (or low social inclusion) and adverse mental health, but not for physical and general health.

The links between social exclusion and health in healthcare settings in have been examined in a scoping study by O'Donnell et al. (2018) who reviewed 22 instruments (questionnaires) used to measure social exclusion or inclusion. The focus of 19 of these was mental health. In all cases, the instruments included “social networks” and in 17 cases they also included “community.”

Some studies have focused on particular groups who may be at high risk of social exclusion. Watson et al. (2016) looked at health and social exclusion for homeless people in Ontario, Canada, by conducting interviews with 21 people who did not have a permanent residence that they could return to when they chose. Several themes emerged from the analysis in three domains: social environment, health behaviors, and health status. The three domains were interwoven. When participants did not feel supported by those around them in their “social environments,” it was difficult for them to feel motivated to practice “healthy behaviors.” The behaviors that helped participants cope with possible social exclusion, homelessness and health changes might lead to a decline in “health status.” Changes to participants’ behaviors and social environments could influence their health status, and vice versa.

Sacker et al. (2017) used the UK Household Longitudinal Study to examine social exclusion and health among older people who are another group who are at risk of social exclusion. They found that among older people in the United Kingdom, poor health is associated with greater social exclusion and, in turn, social exclusion is linked to health decline. They also found that use of a car, mobile phone and the internet are factors that protected older adults in poor health from social exclusion and suggest that designing age-friendly hardware and software might help support social inclusion in later life and have public health benefits.

Transport in the Relationship Between Social Exclusion and Health

The evidence discussed earlier has demonstrated the existence of a relationship between social exclusion and health. Most of the empirical evidence is cross-sectional and so cannot demonstrate the direction of causality. It can be hypothesized that the

relationship works both ways: people who are socially excluded have poorer access to healthcare facilities which lead to poorer health, for example, and some people who have poor health may not be able to work sufficiently to earn enough income to maintain a reasonable standard of living and so their standard of living declines and they finish up excluded from society.

An intermediary factor in the relationship may be transport because travel is often required to access healthcare and to maintain a healthy lifestyle, and those in poor health may have limited access to travel facilities to enable them to remain within society. These issues will be considered further.

Social Exclusion and the Ability to Travel

Some of the characteristics of people who are socially excluded may affect their access to travel:

- **Income:** People with low incomes travel less than richer people. This is illustrated by evidence for England in 2018, where people in the top income quintile travelled an average of 14,810 km a year, and those in the lowest quintile had an average of 6,534 km (Department for Transport, 2019a). This may be partly because poorer people have lower car ownership: in England 54% in the lowest income quintile owned one or more cars compared with the top highest quintile with 87%. They may also have more difficulty paying public transport fares. Horten and Reed (2010) showed that government spending in Britain on transport is strongly biased towards higher income groups because they are more likely to use road and rail than poorer people and they are the modes that tend to attract government expenditure rather than buses.
- **Disability:** This is illustrated by the fact that in England in 2018, the 9% of people aged 16+ with mobility difficulties made an average of 603 trips a year while those without a mobility difficulty made 1049 trips on average (Department for Transport, 2019a).
- **Age:** Older people travel less than younger people because they are more likely to have given up driving, have greater difficulty walking, and difficulty climbing stairs which may make using public transport more difficult and these problems tend to increase with age (Mackett, 2014b, 2017). Travel may be difficult for some young people who cannot afford to learn to drive or afford bus fares and so become excluded from the job market and leisure facilities (pteg, 2010). Younger children may live with a single parent who is employed and so unable to take them or unable to afford the cost of travel to after-school activities (pteg, 2010).
- **Gender:** More men than women hold driving licence: in England in 2018, 81% of males aged 17+ held a licence compared with 70% of females (Department for Transport, 2019a), but more women than men said that they felt unsafe using buses (Department for Transport, 2014).
- **Ethnicity:** In England, members of black and Asian ethnic groups make over 15% fewer trips and have lower car ownership than white people (Department for Transport, 2019a).

Barriers to Travel

Part of the reason for the differences discussed earlier is that barriers to travel exist and these affect some socially excluded people more than others. Some examples are shown as:

- **Cost:** As implied earlier, those with lower incomes travel less than wealthier people partly because of the cost of travel. One in four unemployed people say that their job search is inhibited by the cost of travel to interviews (Social Exclusion Unit, 2003), which means that the chances of obtaining employment are lower.
- **Availability of transport:** As discussed earlier, poorer people have less access to car than those who are wealthy and so are more dependent on public transport. It takes much longer to reach key services (centers of employment, shops, schools, and medical services) for those traveling by public transport and walking compared to those using a car, particularly in rural areas in England (Department for Transport, 2019b).
- **Psychological barriers:** Some people with mental health conditions are so anxious about using some modes of travel that they cannot use them (Mackett, 2019).
- **Physical barriers:** Sometimes small physical barriers such as characteristics of the footway and road crossings can prevent some less mobile older and disabled people from traveling (Titheridge et al., 2009).
- **Facilities:** Sometimes the lack of facilities discourages some people from traveling: for example, some older and disabled people need seats and bus shelters to enable them to make local journeys (Titheridge et al., 2009) and concern about the lack of toilet facilities can cause anxiety for some people (Mackett, 2019).
- **Information:** Nowadays, much travel information is available on mobile phones. While 95% of all households in the United Kingdom owned a mobile phone in 2018, only 85% of those in the lowest income decile did so, and only 67% of retired people living alone mainly dependent on state pensions did so (Office of National Statistics, 2019).

The Role Transport Plays in the Relationship Between Social Exclusion and Health

Transport plays three main roles in the relationship between social exclusion and health: through providing access to facilities that offer health care and opportunities to follow a healthy lifestyle, through active travel that offers health enhancing physical activity, and through externalities that may have adverse effects on health that affect excluded people disproportionately.

Access

Travel offers access to various facilities that provide health benefits: to hospitals and other healthcare services to help reduce the risk of ill-health and to provide treatment when necessary, to facilities such as gyms that help promote well-being, to green spaces that offer physical activity, and to shops that sell healthy food. The difficulties discussed earlier that some socially excluded people have traveling mean that they are not able to enjoy the same health benefits as other people.

Access to healthcare without private transport is particularly difficult. The [Social Exclusion Unit \(2003\)](#) found that one third of people who did not own a car found it difficult to get to hospital compared to 17% of those who own a car. In 2017, 98% of car users in the United Kingdom could reach a doctors' practice within 15 minutes travel and 31% could reach a hospital, whereas the equivalent figures for those traveling by other modes were 72% for going to the doctors' and 5% for visiting a hospital ([Department for Transport, 2019b](#)).

Eating a healthy, balanced diet is an important part of maintaining good health ([NHS, 2019](#)). There is evidence that low-income areas with poor access to healthy foods are associated with higher cardiovascular risk factors ([Morris et al., 2019](#)). [Corfe \(2018\)](#) has examined the barriers to eating healthily in the United Kingdom. The report suggests that 10.2 million individuals in Great Britain live in areas where individuals without a car or with disabilities that hinder mobility may find it difficult to access a wide range of healthy, affordable food products easily. There are places where it is difficult to buy healthy food, but there are many fast food outlets selling cheap, unhealthy food ([Bowyer et al., 2009](#)). In some parts of Britain, this is because large supermarkets have been located on the edge of urban areas ([Furey et al., 2001](#)).

[Braubach et al. \(2017\)](#) show that green spaces can contribute to health and well-being, and discuss the available evidence about the beneficial effects they offer such as improved mental health through stress alleviation, reduced risks of cardiovascular disease, obesity, diabetes and death, and improved pregnancy outcomes by supporting physical activity and reducing exposure to pollutants, noise and excessive heat. However, many green spaces require a car to access them, so people who do not have access to one, including many socially excluded people, may not enjoy these benefits.

Travel as a Form of Physical Activity

The [World Health Organization \(2018\)](#) says that physical activity has significant health benefits and so walking and cycling can contribute to being healthy. The National Travel Survey ([Department for Transport, 2019a](#)) shows that, while the average distanced walked each year decreases with income, probably because of higher car availability at higher income levels, the average distance cycled in a year increases with income. Those in the highest income quintile undertake more active travel (walking and cycling combined) than those in the lowest quintile. This implies that people with low incomes, who are more likely to be socially excluded, receive fewer health benefits from active travel than others.

The Externalities of Transport

As discussed earlier, transport provides benefits by enabling people to reach opportunities including those that provide health benefits. However, transport produces externalities, that is, effects on people other than those using it. These include vehicular emissions that cause atmospheric pollution, noise, and injuries and deaths from vehicle collisions.

The adverse impacts of cars are concentrated in places where cars are driven, which are not necessarily the places where the users live. Vehicle emissions have an adverse impact upon health ([Buckeridge et al., 2002](#); [Zhang and Batterman, 2013](#)). The highest emissions of pollutants from cars tend to be in deprived areas ([King and Steadman, 2000](#)).

[Halonen et al. \(2015\)](#) found that long-term exposure to road traffic noise in London was associated with small increased risks of mortality, particularly for stroke in the elderly even after correction for the effects of traffic pollution. Groups with lower socioeconomic position have been found to be subject to higher environmental noise exposure, in general ([Dreger et al., 2019](#)).

[Edwards et al. \(2006\)](#) have shown that there is a relationship between deprivation and road traffic injury risk in London. The strongest relationship found was for pedestrians, where the most deprived were over twice as likely to be injured as the least deprived.

The higher levels of traffic along roads in areas where socially excluded people tend to live can lead to community severance by limiting social interaction with friends and neighbors. This can cause lower levels of well-being ([Boniface et al., 2015](#)).

Summing Up

This article has demonstrated the link between social exclusion and health and ways in which transport can affect the relationship. Whilst there are many other links between social exclusion and health, for example, in the political, social and cultural realms, it is clear that transport plays a role. It can also play a part in addressing social exclusion. [Mackett and Thoreau \(2015\)](#) outline some examples of transport interventions that might help reduce social exclusion and improve health, for example, by helping people find employment and so increase their income. One example where a transport policy was introduced explicitly to help address social exclusion and maintain well-being was the introduction of free off-peak bus travel for older and disabled people in Great Britain in 2008, which appears to have been effective ([Mackett, 2014a](#)).

Improving health and reducing social exclusion are important policy objectives, and there are useful synergies between them, with transport being a useful factor to facilitate some of the links, implying that suitable investment in transport may help to achieve both objectives.

Biography

Roger Mackett BA, MSc, PhD, FCILT, FCIHT is Emeritus Professor of Transport Studies at University College London. He has researched into various aspects of transport policy including the barriers to access for older and disabled people, the use of the car for short trips and the effects of car use on children's lives. He is a member of the Disabled Persons' Transport Advisory Committee (DPTAC) and the Standing Committee on Accessible Transportation and Mobility of the US Transportation Research Board (TRB).

References

- Boniface, S., Scantlebury, R., Watkins, S.J., Mindell, J.S., 2015. Health implications of transport: evidence of effects of transport on social interactions. *J. Transp. Health* 2, 441–446. Available from: <https://www.sciencedirect.com/science/article/pii/S2214140515002224>.
- Bowyer, S., Caraher, M., Eilbert, K., Carr-Hill, R., 2009. Shopping for food: lessons from a London Borough. *Br. Food J.* 111, 452–474.
- Braubach, M., Egorov, A., Mudu, P., et al., 2017. Effects of urban green space on environmental health, equity and resilience. In: Kabisch, N., Korn, H., Stadler, J., Bonn, A. (Eds.), *Nature-Based Solutions to Climate Change Adaptation in Urban Areas*. Springer Open, Cham, Switzerland, pp. 187–205.
- Buckeridge, D.L., Glazier, R., Harvey, B.J., et al., 2002. Effect of motor vehicle emissions on respiratory health in an urban area. *Environ. Health Perspect.* 110, 293–300.
- Corfe, S., 2018. What are the barriers to eating healthily in the UK? Social Market Foundation. Available from: <http://www.smf.co.uk/wp-content/uploads/2018/10/What-are-the-barriers-to-eating-healthily-in-the-UK.pdf>.
- Department for Transport, 2014. Public attitudes towards buses. Available from: <https://www.gov.uk/government/statistical-data-sets/att01-attitudes-towards-buses>.
- Department for Transport, 2019a. National Travel Survey: 2018. Available from: <https://www.gov.uk/government/statistics/national-travel-survey-2018>.
- Department for Transport, 2019b. Journey Time Statistics, England: 2017. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/853574/journey-time-statistics-2017.pdf.
- Dreger, S., Schüle, S.A., Hilz, L.K., Bolte, G., 2019. Social inequalities in environmental noise exposure: a review of evidence in the WHO European region. *Int. J. Environ. Res. Public Health* 16, 1011, doi:10.3390/ijerph16061011.
- Edwards, P., Green, J., Roberts, I., Grundy, C., Lachowycz, K., 2006. Deprivation and Road Safety in London: A report to the London Road Safety Unit, London.
- Furey, S., Strugnell, C., McIlveen, H., 2001. An investigation of the potential existence of "food deserts" in rural and urban areas of Northern Ireland. *Agric. Hum. Values* 18, 447–457.
- Hälonen, J.I., Hansell, A.L., Gulliver, J., et al., 2015. Road traffic noise is associated with increased cardiovascular morbidity and mortality and all-cause mortality in London. *Eur. Heart J.* 36, 2653–2661.
- Horten, T., Reed, H., 2010. Where the money goes: How we benefit from public services, Trades Union Congress. Available from: <http://www.tuc.org.uk/sites/default/files/extras/wherethemoneygoes.pdf>.
- Kenyon, S., Lyons, G., Rafferty, J., 2002. Transport and social exclusion: investigating the possibility of promoting inclusion through virtual mobility. *J. Transp. Geogr.* 10, 207–219.
- King, K., Steadman, J., 2000. Analysis of air pollution and social deprivation, AEA Technology. Available from: <https://uk-air.defra.gov.uk/assets/documents/reports/cat09/aeat-r-env-Q241.pdf>.
- Lenoir, R., 1974. *Les Exclus: Un français sur dix*. Editions du Seuil, Paris.
- Morris, A.A., McAllister, P., Grant, A., et al., 2019. Relation of living in a "food desert" to recurrent hospitalizations in patients with heart failure. *Am. J. Cardiol.* 123, 291–296.
- NHS, 2019. Eat well. Available from: <https://www.nhs.uk/live-well/eat-well/>.
- O'Donnell, P., O'Donovan, D., Elmusharaf, K., 2018. Measuring social exclusion in healthcare settings: a scoping review. *Int. J. Equity Health* 17, 15. Available from: <https://doi.org/10.1186/s12939-018-0732-1>.
- Office of National Statistics, 2019. Percentage of households with durable goods by income group and household composition: Table A46. Available from: <https://www.ons.gov.uk/people-populationandcommunity/personalandhouseholdfinances/expenditure/datasets/percentageofhouseholdswithdurablegoodsbyincomegroupandhouseholdcompositionuktablea46>.
- Popay, J., Escorel, S., Hernández, M., et al., 2008. Understanding and Tackling Social Exclusion, Final Report to the WHO Commission on Social Determinants of Health. Available from: https://www.who.int/social_determinants/knowledge_networks/final_reports/sekn_final%20report_042008.pdf.
- Mackett, R.L., 2014a. Has the policy of concessionary bus travel for older people in Britain been successful? *Case Studies Transp. Policy* 2, 81–88. Available from: <http://dx.doi.org/10.1016/j.cstp.2014.05.001>.
- Mackett, R.L., 2014b. Overcoming the barriers to access for older people, Report produced for the Age Action Alliance. Available from: <http://ageactionalliance.org/wordpress/wp-content/uploads/2014/11/Overcoming-the-barriers-to-access-Nov-14.pdf>.
- Mackett, R.L., 2017. Older people's travel and its relationship to their health and wellbeing. In: Musselwhite, C.B.A., (Ed.) *Transport, Travel and Later Life*. Emerald, Bingley, Great Britain, pp. 15–36. Available from: <http://www.emeraldinsight.com/doi/full/10.1108/S2044-99412017000010001>.
- Mackett, R.L., 2019. Mental health and travel: Survey report. Report, Department of Civil, Environmental and Geomatic Engineering, University College London. Available from: <https://bit.ly/2lviXbs>.
- Mackett, R.L., Thoreau, R., 2015. Transport, social exclusion and health. *J. Transp. Health* 2, 610–617. Available from: <http://dx.doi.org/10.1016/j.jth.2015.07.006>.
- pteg, 2010. Transport and Social Inclusion: Have we made the connections in our cities? Report, Passenger Transport Executive Group. Available from: <http://www.pteg.net/resources/types/reports/transport-and-social-inclusion-have-we-made-connections-our-cities>.
- Sacker, A., Ross, A., MacLeod, C.A., Netuveli, G., Windle, G., 2017. Health and social exclusion in older age: evidence from Understanding Society, the UK household longitudinal study. *J. Epidemiol. Commun. Health* 71, 681–690. Available from: <https://jech.bmj.com/content/71/7/681>.
- Social Exclusion Unit, 2003. Making the connections: transport and social exclusion. Report, Social Exclusion Unit. The Stationery Office, London. Available from: https://www.ilo.org/empolicy/pubs/WCMS_ASIST_8210/lang-en/index.htm.
- Titheridge, H., Achuthan, K., Mackett, R., Solomon, J., 2009. Assessing the extent of transport social exclusion among the elderly. *J. Transp. Land Use* 2, 31–48. Available from: <https://www.jtlu.org/index.php/jtlu/article/view/44/56>.
- United Nations, 1966. International Covenant on Economic, Social and Cultural Rights. Available from: <https://www.ohchr.org/EN/ProfessionalInterest/Pages/CESCR.aspx>.
- van Bergen, A.P.L., Wolf, J.R.L.M., Badou, M., et al., 2019. The association between social exclusion or inclusion and health in EU and OECD countries: a systematic review. *Eur. J. Public Health* 29, 575–582. Available from: <https://doi.org/10.1093/eurpub/cky143>.
- Watson, J., Crawley, J., Kane, D., 2016. Social exclusion, health and hidden homelessness. *Public Health* 139, 96–102.
- World Health Organization, 2008. The Tallinn Charter: Health Systems for Health and Wealth. WHO Regional Office for Europe, Copenhagen. Available from: <http://www.euro.who.int/document/E91438.pdf>.
- World Health Organization, 2010a. Poverty, social exclusion and health systems in the WHO European Region. Available from: http://www.euro.who.int/__data/assets/pdf_file/0004/127525/e94499.pdf.
- World Health Organization, 2010b. Poverty and social exclusion in the WHO European Region: health systems respond. Available from: <https://www.inmp.it/eng/Publications/Books/Poverty-and-social-exclusion-in-the-WHO-European-Region-health-systems-respond>.
- World Health Organization, 2018. Physical activity, Factsheet. Available from: <https://www.who.int/news-room/fact-sheets/detail/physical-activity>.
- Zhang, K., Batterman, S., 2013. Air pollution and health risks due to vehicle traffic. *Sci. Total Environ.* 450–451, 307–316.

Further Reading

- Hills, J., Le Grand, J., Piachaud, D. (Eds.), 2002. *Understanding Social Exclusion*, Oxford University Press, Oxford.
- Khan, S., Combaz, E., McAslan FE., 2015. Social exclusion: topic guide. Governance and Social Development Resource Centre, University of Birmingham, Birmingham. Available from: <https://assets.publishing.service.gov.uk/media/57a0899a40f0b652dd0002f2/SocialExclusion.pdf>.
- Stanley, J.K., Hensher, D.A., Stanley, J.R., Vella-Brodrick, D., 2011. Mobility, social exclusion and well-being: Exploring the links. *Transp. Res. Policy Pract.* 45, 789–801.
- van Bergen, A.P.L., Wolf, J.R.L.M., Badou, M., et al., 2019. The association between social exclusion or inclusion and health in EU and OECD countries: a systematic review. *Eur. J. Public Health* 29, 575–582. Available from: <https://doi.org/10.1093/eurpub/cky143>.

Burden of Disease Assessment

David Rojas-Rueda, Department of Environmental and Radiological Health Sciences, Colorado State University, Fort Collins, CO, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	347
Burden of Disease Approach	347
Data Required for Estimate Burden of Disease	348
Other Health Indicators	349
Comparative Risk Assessment Approach	349
Examples of BoD and CRA	351
Conclusions	351
References	352

Introduction

The Burden of Disease (BoD) is a tool to integrate health metrics and evidence into public health and nonhealth sectors (such as transport). BoD helps to integrate morbidity and mortality in the same unit, estimating base on the life expectancy, and the natural history of the disease, those years that certain populations will live with a disability-related to specific health diagnosis and those years of life lost due to premature death. BoD is a tool that can help health professionals and stakeholders understand the impact of different diseases and causes of death in a particular population. BoD can also help to compare different risk factors and how different health and nonhealth interventions can impact public health. The BoD is a framework for integrating, analyzing, and communicate, information that is available on a population's health, along with some understanding of how that population's health is changing so that the information is more relevant for health policy and planning purposes (Lopez et al., 2006). Transport policies have been related to several health risk factors (e.g., air pollution traffic incidents, physical activity, etc.). BoD will help to measure the health impacts of transport policies, based on risk factors, providing a standard health metric to compare transport policies and interventions with a health perspective (Fig. 1).

Burden of Disease Approach

The BoD is a concept developed in the 1990s by the World Bank and the World Health Organization (WHO) to describe death and loss of health due to diseases, injuries, and risk factors for all regions of the world (Lopez et al., 2006). The burden of a particular disease is estimated by adding together the number of years of life a person loses as a consequence of dying early because of the disease based on their life expectancy [called Years of Life Lost (YLL)]; and the number of years of life a person lives with disability caused by the disease, based on the natural history of the disease and life expectancy [called Years of Life lived with Disability (YLD)]. Adding the YLL and YLD provides an estimate of disease burden, called the Disability-Adjusted Life Year (DALY). One DALY represents the loss of one year of life lived in full health and is the primary health unit used by the BoD approach. Besides the DALY, several other measures have been devised, including the Quality-Adjusted Life Year (QALY), the health-adjusted life expectancies (HALE), and the Healthy Life Year (HeaLY).

The DALY is the most common measure used by the BoD approach and combines in one measure the time lived with disability and the time lost due to premature mortality. The following formulas describe how DALYs can be estimated:

$$\text{DALY} = \text{YLL} + \text{YLD}$$

where, YLL = years of life lost due to premature mortality. YLD = years lived with disability

The YLL corresponds to the number of deaths multiplied by the standard life expectancy (from the geographical location) at the age at which death occurs. The basic formula for calculating the YLL for a given cause, age, or sex, is:

$$\text{YLL} = N \times L$$

where, N = number of deaths. L = standard life expectancy at the age of death (in years).

To estimate YLD, the number of disability or disease cases is multiplied by the average duration of the disease and a weight factor that reflects the severity of the disease on a scale from 0 (perfect health) to 1 (dead). The formula for YLD is:

$$\text{YLD} = I \times \text{DW} \times L$$

where, YLD = years lived with disability. I = number of incident cases. DW = disability weight. L = average duration of disability (years)

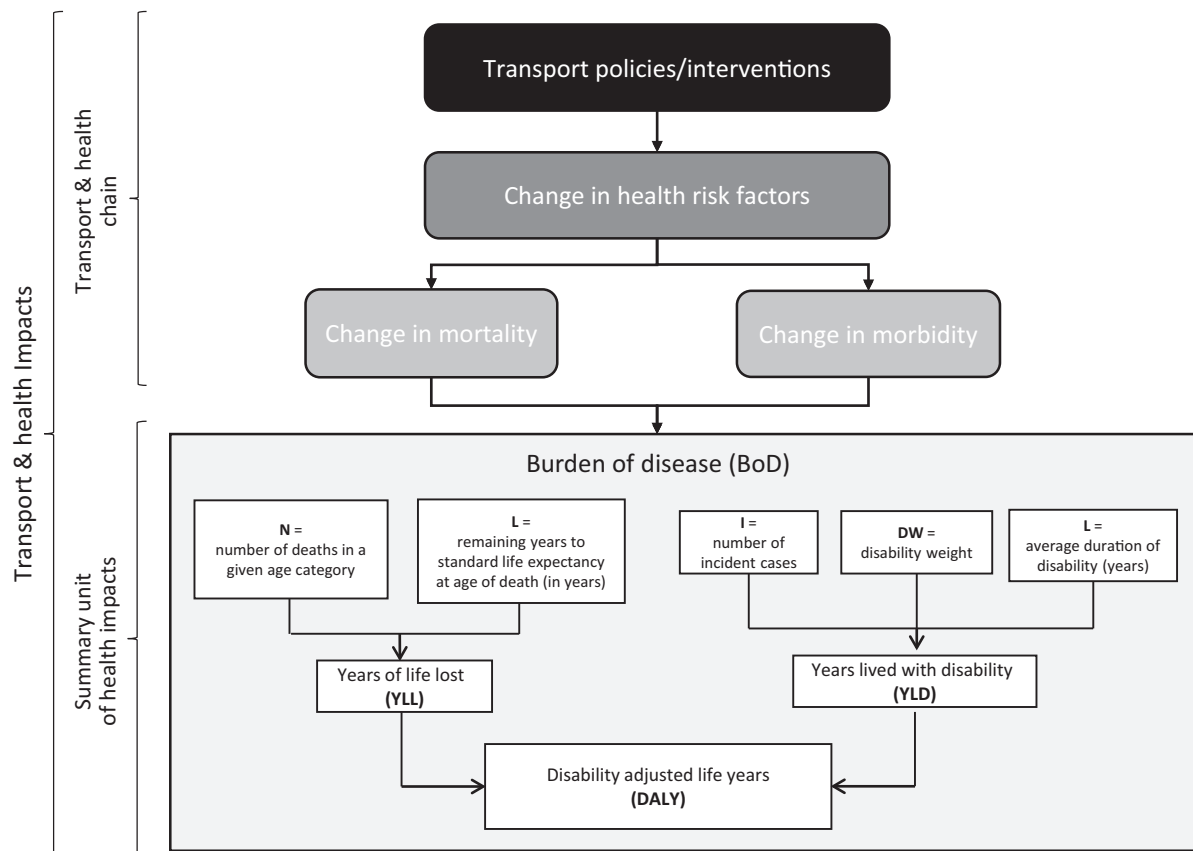


Figure 1 Burden of disease in the context of transport policies.

Data Required for Estimate Burden of Disease

Number of deaths: the total number of deaths are the reported and registered number of fatalities on the specific cause of death, in a specific population in each time. This value can be found in the death certificate, health, and population statistics of the region and country. The number of deaths can be reported in an aggregation of estimates of deaths by cause, age, and sex by geographical location. For the estimating of YLL, a specific number of deaths for the geographical location under study should be chosen.

Standard life expectancy: life expectancy refers to the number of years a person can expect to live. By definition, life expectancy is based on an estimate of the average age that members of a population group will be when they die. The life expectancy can be obtained from national and regional vital registrations, census, and surveys. Global life expectancy at birth in 2016 was 72.0 years (74.2 y for females and 69.8 y for males), ranging from 61.2 years in the African Region to 77.5 years in the European Region. Women live longer than men all around the world. The gap in life expectancy between the sexes was 4.4 years in 2016. There are two main options when choosing the life expectancy in the BoD approach, (1) can use the latest available life expectancy of the country or region where the BoD is estimated, and (2) also can be used the highest life expectancy reported around the world, for example, Japan has a life expectancy of 85.3 years in 2017. The main decision between choosing one option is based on the assessor judgment based on a representative approach versus an equity approach. The representative approach (option 1) will choose the life expectancy of the county or regions where a BoD is performed; an equity approach (option 2) will assume that all the humans have the same potential to survive as the maximum reported in the best-case scenario, in other words, if Japanese people have a life expectancy of 85.3 years any human been can have the same life expectancy under a proper healthy life style.

Number of incident cases: incidence is the number of newly diagnosed cases of a disease in a specific population and time. The BoD estimations should use similar time frames for the number of deaths and incident cases in the population to estimate DALYs. Incident cases can be found in the official records and health statistics of a region or county.

Disability weight: A disability weight is a weight factor that reflects the severity of the disease on a scale from 0 (perfect health) to 1 (equivalent to death). Disability weights, which represent the magnitude of health loss associated with specific health outcomes, are used to calculate years lived with disability (YLD) for these outcomes in a given population. For example, a disability weight for mild diarrhea could 0.074 compare to severe diarrhea 0.24. But more severe health outcomes could have more impact on disability such as treated spinal cord lesion at the neck with a disability weight of 0.58 or an untreated spinal cord lesion at the neck with a disability

weight of 0.0.73. The most recent list of disability weight is provided by the global burden of disease (GBD) study in 2013 (Salomon et al., 2015).

Average duration of disability: The average duration of disability is the disease duration until remission, cure, or death. This is also called the natural history of the disease, where the process of the disease that without management or treatment could result in disease remission or death, and with management or treatment would result in the disease cure or control. The disease duration is based on the disease-incidence approach and can be found on scientific publications for the specific diagnosis. Traditionally, the BoD has been estimated by an incidence-based method. However, since the first GBD study in 2010, the prevalence-based approach has become more widespread. The prevalence-based approach estimates all health-related losses over the course of a year, whereas the incidence-based approach captures the burden of disease in new diagnostic cases. Since the incidence-based YLD is calculated as the sum of future health loss related to disease incidence in the reference year, it will not reflect the prevalence of all YLDs, nor does it accurately capture the time between which a disease occurs and the age at which health loss begins.

Other Health Indicators

Quality-adjusted life years is a statistical health indicator that reflects health quantity and quality. It was developed in the late 1960s by economists and is mainly used for cost-effectiveness analyses of improvements in social welfare and clinical interventions. The QALY method can estimate the number of years lived and the quality of life during those years that can be attributed to an intervention. When combined with the cost of providing intervention, QALYs are used to develop cost-utility ratios required to generate a year of “perfect health,” a perception of life without pain or disease. The health-related quality of life (or utility values) in QALYs are not linked to specific diseases but rather are based on individuals’ opinions about their own health state (patient weights) or on the judgments of others (e.g., a representative sample of the population, study researchers, or health professionals) about a particular health state (community weights where perfect health value = 1 and death = 0). The community weights integrate biomedical and psychosocial aspects of the BoD and consider five quality of life dimensions: (1) mobility, (2) pain or discomfort, (3) self-care, (4) anxiety-depression, and (5) usual activities (Lajoie, 2015).

Healthy life year combines the healthy life lost due to death with that lost due to morbidity. The healthy life lost requires knowing the incidence rate, case fatality ratio, the age of disease onset, the age of death, and the life expectancy at these ages. The HeaLY includes three components for disability: case disability ratio (CDR, analogous to the case fatality ratio), the extent of disability, and duration of the disability. The CDR and duration of disability can be determined objectively, but the assessment of the extent of disability, which ranges from 0 (no disability) to 1 (equivalent to death), may have a substantial subjective element (Hyder et al., 2012). An essential difference between the HeaLY and DALY is that the starting point for the HeaLY is the onset of disease; the loss of healthy life is based on the natural history of the disease (as modified by interventions). The YLD component of the DALY is similar, but the YLL is based on mortality in the current year. In a steady-state, there is no difference, but when there is a changing incidence, the DALY approach can significantly underestimate the health impacts.

Health-adjusted life expectancy is a health measure that summarizes the expected number of years to be lived in what might be termed the equivalent of full health. HALE provides one of the best summary measures for the overall level of health. Health expectancy indices combine the mortality experience of a population with the disability experience. In some cases, the HALE is calculated using the prevalence of disability at each age in order to divide the years of life expected at each age according to a life table cohort into years with and without a disability. Mortality is captured by using a life table method, while the disability component is expressed by additions of the prevalence of various disabilities within the life table. This indicator allows an assessment of the proportion of life spent in disabled states. HALE does not relate to specific diseases but rather to the average extent of disability among that proportion in each age group that is disabled (Hyder et al., 2012).

These BoD units are especially useful for policy comparison that affects different causes of mortality and different disease diagnoses. The BoD gives the stakeholders a unit of comparison that summarizes mortality and morbidity, comparing different diagnosis, helping to understand more comprehensively the magnitude and direction of a specific intervention, plan, program, or policy. The main disadvantage of these measures (DALYs, YLL, and YLD) is the difficulty to be interpreted for nonhealth practitioners, who are, in most cases, the final audience of a public health intervention.

Comparative Risk Assessment Approach

The comparative risk assessment (CRA) approach compares the changes of an intervention, policy, program, or plan on health determinants and their impact on the BoD (mortality and /or morbidity) compared with a baseline scenario or business as usual (Fig. 2). The aim of the CRA approach is to provide quantitative estimates of the expected health impacts (e.g., changes in premature deaths, cases of the disease, DALYs, etc.) and the distribution thereof among the population.

The steps of a CRA are:

- *Hazard identification:* The intervention, policy, or program context and the relevant exposure (or a multitude of exposures) needs to be identified (e.g., type of pollutant), and the baseline exposure level distribution of the population under study needs to be quantified.

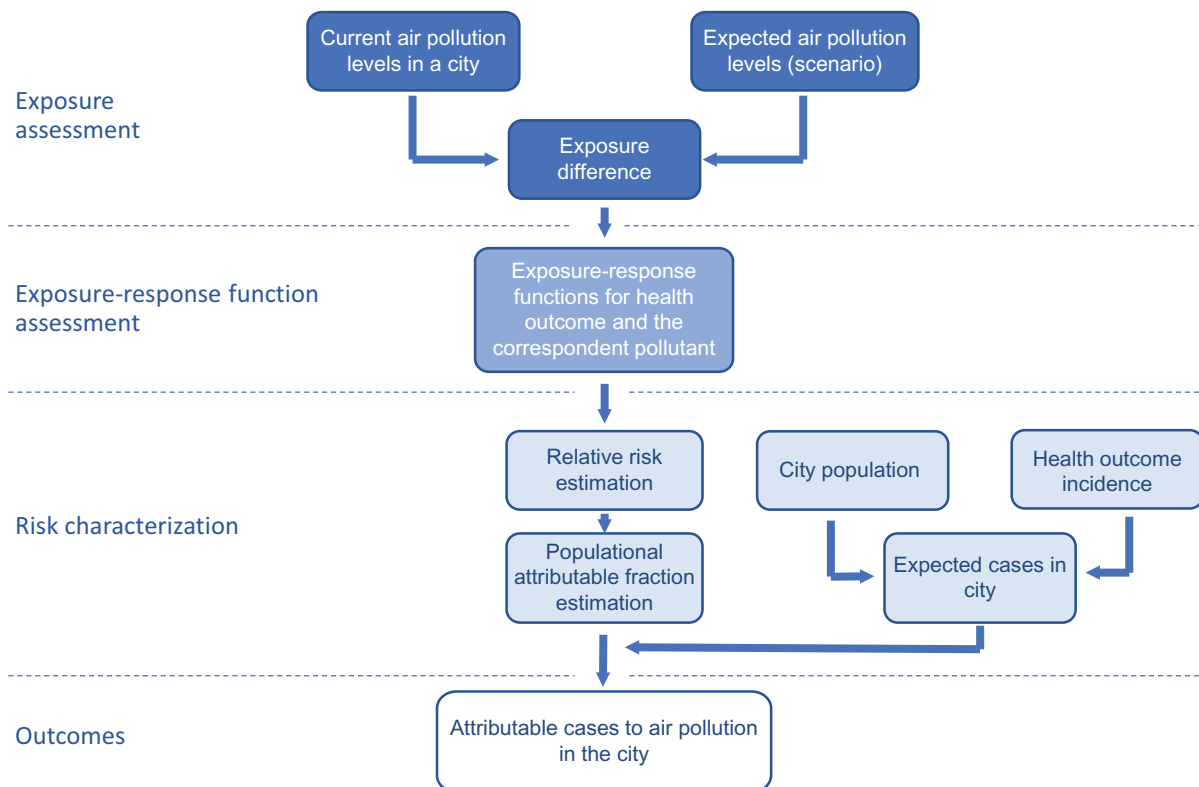


Figure 2 Comparative risk assessment framework applied to air pollution.

- *Exposure assessment:* The exposure level at baseline is compared with the aimed at the exposure level of the counterfactual scenario, and the resulting difference in exposure level is quantified.
- *Exposure-response assessment:* An exposure-response function (ERF) that quantifies the strength of association between the exposure and the health outcome needs to be available from the literature or available guidance. The ERF needs to be of the best available evidence. In an ideal case, the ERF was estimated particularly for the population under study, reflecting most accurately the level of risk the population is exposed to. However, in many cases, an ERF for the population under study is not available; therefore, the ERF should preferably be a pooled, generalized estimate of the overall effect coming from a meta-analysis (or large longitudinal study).
- *Risk characterization:* The risk estimate obtained from the ERF that most epidemiological studies report in terms of relative risk (RR), which is the ratio of incidence observed at two different exposure levels, quantifies the strength of association between the exposure and health outcome. The risk estimate is scaled to the difference in exposure level resulting from the comparison of the baseline exposure level with the counterfactual exposure level.

In addition to these steps, when a policy plan or program is assessed is essential to consider the following points:

- *Scenario(s) definition:* A counterfactual scenario (or a set of scenarios) is defined, which describes the possible exposure level that is aimed at with the proposed intervention, policy, or program. Various criteria may determine the choice of counterfactual exposure distribution. Ideally, counterfactuals should represent intervention, policy, or program actions that are realistic and can be implemented (e.g., motorized traffic reductions to reduce air pollution levels in cities) rather than counterfactuals that are not achievable (e.g., zero air pollution). [Murray and Lopez \(1999\)](#) looked into different counterfactual scenarios categories and identified exposure distributions corresponding to theoretical minimum risk (i.e., exposure distribution with lowest population risk), plausible minimum risk (i.e., imaginable exposure distribution), possible minimum risk (i.e., exposure distribution observed in some populations), and cost-effective minimum risk (i.e., considers costs of exposure reduction).
- *Health outcome definition:* The health outcome (or multiple health outcomes) of interest needs to be defined that was shown by previous epidemiological studies to be associated with the exposure. A causal relationship between the exposure and the health outcome of interest is a prerequisite.
- *Health data identification:* The total burden (TB) of the health outcome of interest needs to be available for the population under study. Incidence rates are preferred over prevalence proportions because only new cases expected over a given time (i.e., incident cases) are preventable under the assumption that exposure levels will change.

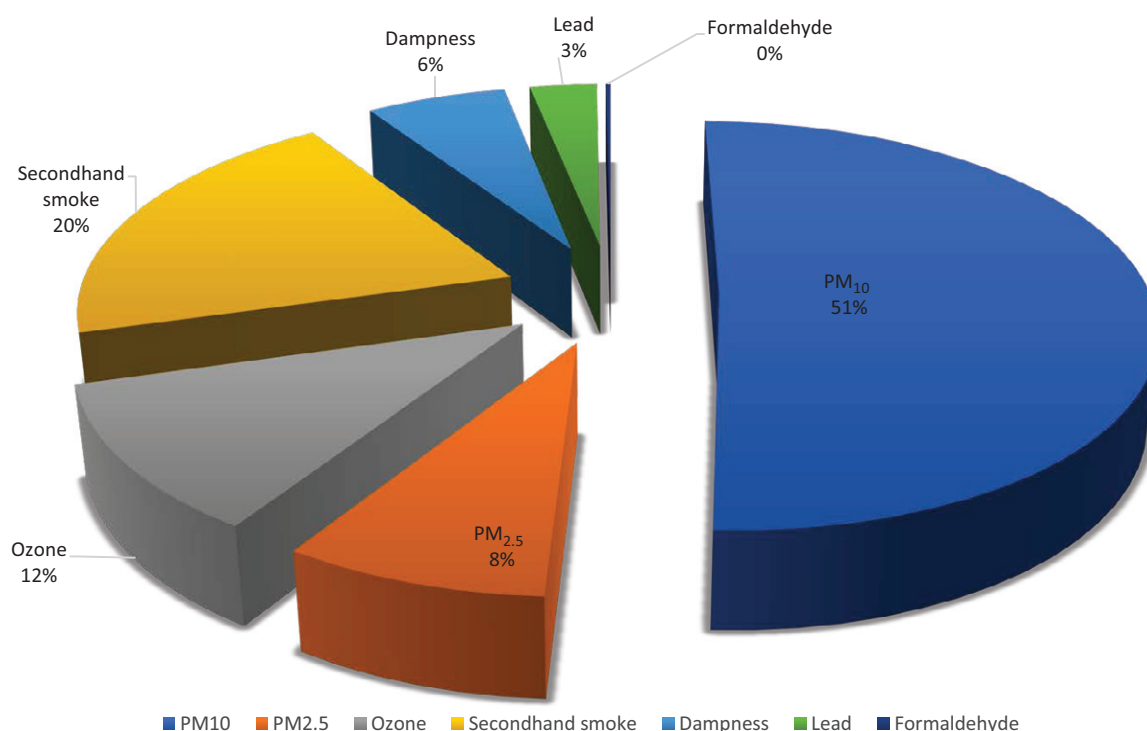


Figure 3 Environmental burden of childhood disease in Europe 28 in disability-adjusted life years (DALYs). Source: Modified from Rojas-Rueda et al. (2019).

Examples of BoD and CRA

The GBD project is the most extensive, and most know example of BoD. The GBD is published each year and include a global perspective on multiple causes of the and diagnoses, providing YLL and prevalence YLD (Vos et al., 2015). The GBD provides multiple resources to communicate their results and DALYs, YLL, YLD can be consulted by country or region in their website <https://vizhub.healthdata.org/gbd-compare/>. The GBD provides different results based on age, sex, year health outcome, and risk factor, offering a great resource to consult the BoD of a specific geographical location, point in time, and health outcome a comparison among years, locations, age, sex, and outcome. The WHO also provides a GBD, and their estimates are produced on average every 5 years (WHO, 2019). The GBD from the WHO also provides DALYs, YLL, and prevalence YLD, globally, by region, country. More specific BoD focus on the environment is also available; recent studies have estimated the environmental BoD (EBoD) in specific age groups and geographical locations, such as in Europe (Rojas-Rueda et al., 2019). This EBoD focuses on estimating the health burden (DALYs, YLL, and YLD), related to specific environmental risk factors (e.g., air pollution, lead, secondhand smoke) and specific health outcomes (e.g., asthma). In this case, the EBoD instead of comparing health outcomes, compares the health burden of different risk factors (Fig. 3), with the aim to highlight the risk factors that could require more urgent policy interventions to improve public health.

In comparison to BoD, the comparative risk assessment approach does not focus only on quantifying health outcomes and risk factors. The CRA adds the dimension of an intervention to these assessments, comparing different policy scenarios and their impact on health determinants (risk factors) and outcomes. For example, Mueller et al. (2017a, 2017b) attributed almost 660 premature deaths and almost 10,000 DALYs to breaching WHO (PM_{2.5}) air quality guidelines in Barcelona, Spain, recognizing that motorized traffic is the most important contributor to local air pollution levels. In another study, Khreis et al. (2019) looked into the air pollution childhood asthma burden in 18 European countries with over 64 million children and estimated that if minimum pollution levels recorded for NO₂, PM_{2.5}, and BC could be complied with, then over 135,000, 190,000, and 89,000 incident cases of asthma could be prevented, respectively. A recent study assessed the Barcelona Superblock Model, an innovative land use intervention with the aims to reclaim space for people, reduce motorized transport, promote active transport, increase urban greening, and mitigate effects of climate change, found that with full implementation of the model, current city-wide annual mean NO₂ concentrations of 47.2 µg/m³ could be reduced to 35.7 µg/m³, which in return would result in 291 preventable deaths and an increase of average life expectancy by 129 days (Mueller et al., 2019).

Conclusions

The BoD approach and the CRA are essential and valuable tools to measure multiple health outcomes, risk factors, and transport policy interventions. Public health practitioners and authorities guide their policies base on several health metrics, and currently,

DALY, YLL, and YLD are the most common health metric used to assess the population health and compare health and nonhealth sector policies. Examples like the GBD, the WHO BoD and a single project on EBoD or CRA should be considered as good examples of how transport policies should be evaluated and assessed on health.

References

- Hyder, A.A., Puwanachandra, P., Morrow, R.H., 2012. Measuring the health of populations: explaining composite indicators. *J. Public Health Res.* 1 (3), 222–228, doi:10.4081/jphr.2012.e35, Published 2012 Dec 28.
- Khreis, H., et al., 2019. Outdoor air pollution and the burden of childhood asthma across Europe'. *Eur. Respir. J.* 54 (4), 1802194, doi:10.1183/13993003.02194-2018.
- Lajoie, J., 2015. Understanding the measurement of global burden of disease – National Collaborating Centre for Infectious Diseases. Available from: <https://nccid.ca/publications/understanding-the-measurement-of-global-burden-of-disease/>. (Accessed: 26 May 2020).
- Lopez, A.D., et al., 2006. Global Burden of Disease and Risk Factors Editors AND PACIFIC THE CARIBBEAN. Available from: http://apps.who.int/iris/bitstream/10665/41864/1/0965546608_eng.pdf.
- Mueller, N., Rojas-Rueda, D., Basagaña, X., Cirach, M., Cole-Hunter, T., Dadvand, P., et al., 2017. Health impacts related to urban and transport planning: A burden of disease assessment. *Environ. Int.* 107, 243–257, doi:10.1016/j.envint.2017.07.020.
- Mueller, N., Rojas-Rueda, D., Basagaña, X., Cirach, M., Cole-Hunter, T., Dadvand, P., et al., 2017a. Urban and transport planning related exposures and mortality: A health impact assessment for cities. *Environ. Health Perspect.* 125 (1), 89–96, doi:10.1289/EHP220.
- Mueller, N., et al., 2020. Changing the urban design of cities for health: The superblock model. *Environ. Int.* 134, 105132, doi:10.1016/j.envint.2019.105132.
- Murray, C., Lopez, A., 1999. On the comparable quantification of health risks: lessons from the global burden of disease study. *Epidemiology* 10, 594–605.
- Rojas-Rueda, D., et al., 2019. Environmental burden of childhood disease in Europe. *Int. J. Environ. Res. Public Health* 16 (6), 1084, doi:10.3390/ijerph16061084.
- Salomon, J.A., et al., 2015. Disability weights for the Global Burden of Disease 2013 study. *The Lancet Global Health.* 3 (11), e712–e723, doi:10.1016/S2214-109X(15)00069-8.
- Vos, T., et al., 2015. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *The Lancet* 386 (9995), 743–800, doi:10.1016/S0140-6736(15)60692-4.
- WHO, 2019. Global Health Estimates 2016: Disease burden by Cause, Age, Sex, by Country and by Region, 2000–2016. Geneva, World Health Organization; 2018, https://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html.

Achieving a Near-Zero CO₂ Transportation System Worldwide by 2050

Lewis M. Fulton, University of California, Davis, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	353
What Vehicle/Fuel Technology Options are There for Achieving the Targets?	353
National/Regional Progress to Date and Targets	355
Scenarios for Achieving CO₂ Reduction Targets	356
How Can We Achieve the “Revolution”?	356
Concluding Remarks	357
References	357
Recommendations for Further Reading	358

Introduction

The world is experiencing rising CO₂ and other greenhouse-gas emissions, resulting in climate change. The Intergovernmental Panel on Climate Change (IPCC) has warned that continuing increases in GHGs could trigger up to 6°C of warming over the next 100 years, while cutting CO₂ dramatically can still limit temperature change to 2°C or less (IPCC, 2013, 2019).

The IPCC (2019) report shows that reaching global carbon neutrality (zero net emissions on the planet) by 2050–60 is consistent with a 1.5°C climate target. All sectors will need to reach close to zero to achieve this. This paper considers the challenge of reaching a near-zero CO₂ emissions target for transportation worldwide by 2050. It reviews recent evidence and considers the technology/fuels-oriented barriers and strategies for hitting this target, though does not delve deeply into policy. It also does not consider major changes in travel patterns that could contribute to CO₂ reductions. The focus is on the technologies and fuels available to achieve this, and their potential for hitting final and interim targets along the way.

Transportation is one of the fastest growing sectors, with among the fastest rates of increase in CO₂ emissions, globally, in recent years, reaching 8.3 Gt or 24% of total energy-related CO₂ emissions in 2018 (IEA, 2020). It is rising due to rising activity across all transportation modes (at least before the 2020 pandemic hit). This rising activity has been offset to some degree by improved efficiency of vehicles, but not nearly enough to offset this activity growth. Within transportation, passenger road travel (cars, two wheelers, and buses) emitted about 43% (3.6 Gt), while trucks emitted 29% (2.4 Gt). Aviation and shipping each emitted around 10% (a little less than 1 t) (IEA, 2019).

CO₂ emissions from these major transport modes are mostly related to combustion of fossil fuels (typically gasoline, diesel, fuel oil, and jet fuel). Use of these fuels must be cut dramatically and eventually eliminated to achieve a near-zero CO₂ future. CO₂ must also be cut to near zero on a life-cycle basis, including vehicle and fuel production. The energy carriers electricity and hydrogen, produced from renewable sources (or fossil with carbon capture, or nuclear) are the primary approaches to achieving very low carbon energy systems for transportation.

Whether the world's transportation sector can reach a near-zero CO₂ target by 2050 will depend on many factors, especially technology potential, cost, and policies to move markets rapidly toward using these. In transportation perhaps more than any other sector, personal behavior will play a major role; households around the world must be convinced to alter their travel patterns and, especially, to adopt low carbon technology vehicles and fuels. Businesses will as well, and they are more likely to make choices on a cost-minimizing basis than households, so they may be even more difficult to convince, if costs are high. On the fuels side, industry, and power sector players must be convinced (or required) to reduce the carbon intensity of the energy they provide, including electricity, hydrogen, biofuels, and “electrofuels,” (fuels derived from electricity) among others.

While this paper focuses on advanced, very low carbon vehicle technology and fuel options, it should be noted that two other strategies can provide important CO₂ reductions over the next 10–20 years: conventional vehicle fuel economy improvements and changes in travel patterns. In each case, perhaps up to a 50% improvement in energy intensity could be achieved with very strong actions (such as substituting smaller, highly efficient vehicles for larger, less efficient ones, and for travel, a large-scale substitution of mass modes such as trains and buses for individual modes such as cars), going much beyond 50% would likely be very challenging, without changing to very low carbon technologies and fuels.

What Vehicle/Fuel Technology Options are There for Achieving the Targets?

From a vehicle point of view, electric and fuel cell vehicles are the obvious technologies for using decarbonized electricity and hydrogen, and must be priorities for changing road and rail transportation systems. For shipping and especially air, it may be difficult to use these energy carriers, given the need for long distance travel and thus for very energy dense fuels. In such cases, advanced, low-

Table 1 Overview of zero-carbon fuel options and issues

	<i>Electric vehicles and electricity</i>	<i>Fuel cells and hydrogen</i>	<i>Biofuels</i>	<i>Synthetic hydrocarbons ("electrofuels")</i>
Key current trends	Global average EV sales were around 2.5% in 2019 and growing, with about 20 million in service worldwide	Hydrogen is not yet significantly used as a transportation fuel. Around 10–15,000 commercial fuel cell vehicles have been sold (mostly in California) using H ₂	Up to 10% blends of ethanol or bio-diesel/renewable diesel are used in some countries, but there virtually no advanced (cellulosic based) biofuels production	Currently no known synthetic hydrocarbon commercial sales for transportation fuel anywhere in the world
Key infrastructure issues	Vehicle home and public charging infrastructure developing; few fast charging systems in most areas but their need is uncertain	Relatively little hydrogen storage/refueling infrastructure; led by 40 stations in California	Can be compatible with current ICE vehicles and infrastructure. Limits on ethanol and biodiesel but drop-in gasoline/diesel replacement fuels can go to 100%	Can produce drop-in replacement fuels, 100% compatible with current ICE vehicles and infrastructure
Critical path issues	Battery cost reduction; charger installation to support rapid sales	H ₂ electrolysis cost reduction and infrastructure scaleup to help reduce costs	Advanced technology cost reduction and scaleup	Electrolysis, Fischer–Tropsch and other technology scale-up to help reduce costs
Timing issues	Advancing rapidly, potential for 20%–50% shares in many countries by 2030	Early deployment phase; potential for 5%–20% sales shares in most markets by 2030	Current ethanol and biodiesel fuels widely available; large scale production of advanced fuels unlikely by 2030	Large scale production unlikely by 2030
Benefits	Low energy, running cost, zero emissions	Operation in FCEVs, rapid fueling, and longer range than BEVs	Compatible with ICE vehicles, possible low WTW GHG emissions	Compatible with ICE vehicles, potential very low WTW GHG emissions
Costs	High vehicle costs but dropping. Driving cost tends to be low given high efficiency of battery electric vehicles	FCEV purchase costs up to 50% higher than gasoline vehicles. H ₂ driving cost up to 3× that for gasoline, but may drop significantly by 2030 if scale is increased	Conventional biofuels are within 20% of the cost of petroleum fuels; advanced biofuels up to 2× cost	Very high costs of production (\$10–30/gge) likely until high volumes, mature market achieved

Source: Compiled by author based on a range of technical reports including *Fulton et al. (2019)*, *Brynnolf et al. (2018)*, *IEA ETP (2020)*, and *Witcover and Williams (2020)*, among others.

carbon biofuels and “electro fuels” (liquid fuels derived from electricity via production of synthesis gases) are the most obvious pathways.

Table 1 summarizes many of the advantages, disadvantages, and costs of different fuel and energy carrier options. A brief introduction to each technology category (table columns) follows:

- **Vehicle electrification:** Battery electric and plug-in hybrid vehicles can help unlock the route to zero emissions driving, as electric grids are decarbonized. EVs typically run on grid electricity with carbon intensity determined by the local electricity generation mix, but the world’s grids must be mostly decarbonized by 2050 to achieve the carbon neutrality targets in any case.
- **Fuel cells and hydrogen.** Fuel cell vehicles, running on hydrogen, offer the possibility of fast refueling and longer driving range than battery electrics. Hydrogen derived from grid electricity via electrolysis will decarbonize as the grid does. In some cases, hydrogen production could be off grid, such as from wind or solar parks connected to hydrogen electrolysis plants.
- **Biofuels:** There are a wide range of bio-based fuels, including those derived from waste bio-streams, renewable natural gas, and purpose-grown crops. Electricity and hydrogen can be derived from biomass, though here we mainly focus on liquid biofuels that are compatible with internal combustion engine (ICE) vehicles. The net GHG effects of any bio pathway are highly variant, mainly determined by feedstock and fuel production emissions (both direct and indirect effects.) Advanced options tend to derive from cellulosic feedstocks with very low process CO₂ emissions, leading to net-neutral CO₂ emissions.
- **Synthetic hydrocarbons:** These are typically high quality “drop-in” gasoline, diesel, or jet airplane replacement fuels, that can be fully blended (up to 100%, eliminating fossil components) and used by today’s vehicles. They can be produced from direct air (or process CO₂) capture and thermochemical processing to produce methane or long-chain hydrocarbons. These fuels contain CO₂ that will be emitted upon combustion, but since the fuel is produced via carbon capture, they can achieve net zero CO₂ emissions.

A common theme is that all (except electricity) are expensive at this time, and will require large scales and learning to bring costs down to reasonable levels. Other barriers will also need to be overcome, such as provision of refueling infrastructure (especially for hydrogen). Biofuels and synthetic hydrocarbons have the advantage of being compatible with today’s ICE vehicles.

It is notable that most of the liquid fuels described are compatible in high blend shares with today’s gasoline and diesel vehicles, giving them an enormous advantage in terms of minimum required vehicle and refueling infrastructure changes and investment requirements. These include drop-in biofuels and synthetic gasoline, diesel, and jet fuel from electricity (electrofuels), likely using

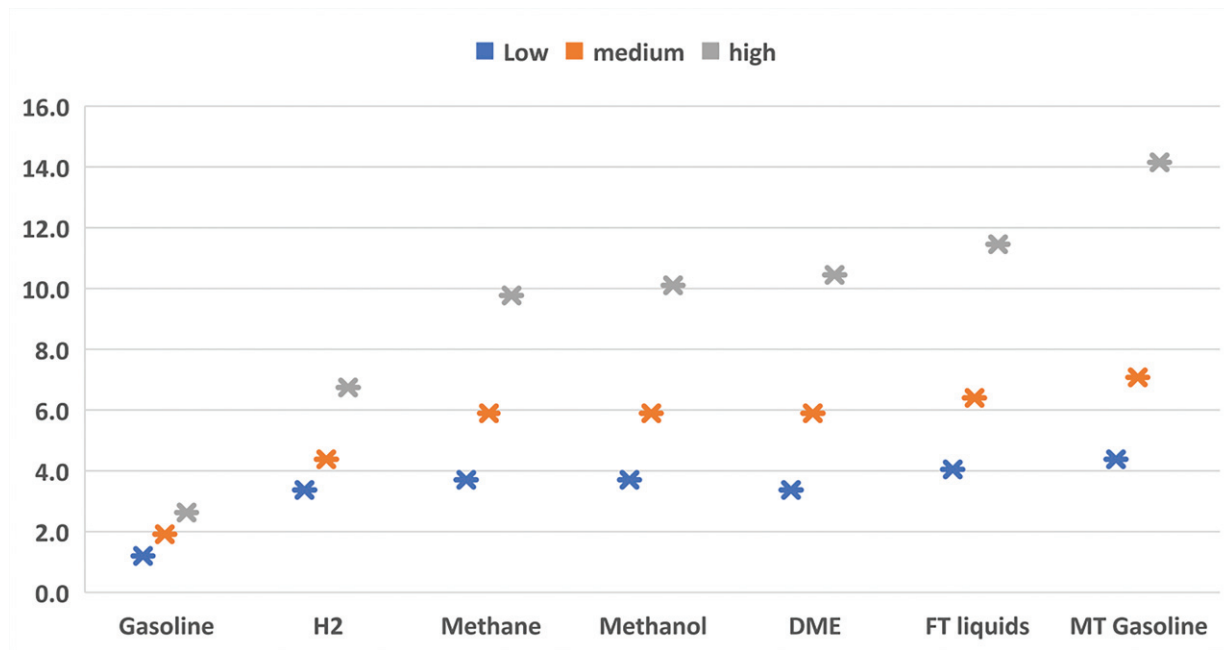


Figure 1 Recent and projected costs of various electrofuels, low/medium/and high cases (\$/gasoline-equivalent gallon). Source: Brynolf et al. (2018).

direct air capture. However, these fuels can be very expensive, and in the case of biofuels, the true life-cycle GHG emissions and some other sustainability factors are of great concern.

Scaling is also a major issue for these types of fuels, either in terms of sustainability (biofuels at high scale may be less sustainable than at low scale), or cost (both advanced cellulosic biofuels and most electrofuels may need very large-scale production to cut costs).

In fact, some of these fuels are already cost competitive, notably some conventional biofuels. For example, in the United States in 2019, the retail costs of traditional biodiesel and ethanol were competitive with diesel and gasoline respectively, though that reflects some differences in tax levels between them and petroleum gasoline and diesel (Clean Cities, 2020). Renewable diesel, made mainly from hydrotreated vegetable oils, and a drop-in fuel (up to 100% blends), is somewhat more expensive but still sells within 30% of the price of diesel fuel.

However, the production cost of advanced biofuels, such as ethanol or drop-in gasoline and diesel replacement fuels from cellulosic feedstocks, is much higher and very little is produced today. These costs may be up to twice the production costs of gasoline and diesel, though should get much close to those costs as plant sizes, capacities, and volumes rise, at least as suggested by modeling studies (Witcover and Williams, 2020).

But what about the other advanced fuel types with the potential for net-zero CO₂ operation? Fig. 1 summarizes the findings of a recent paper (Brynolf et al., 2018) containing a meta-analysis of 24 studies that provided detailed cost estimates for future, large-scale production of various types of electrofuels.

Clearly, there is potential for an eventual large scale electrofuels future where these fuels can be produced at something close to competitive with gasoline and diesel fuel, especially if those fossil fuel costs rise with rising oil price. But near term, low volume production costs are estimated to be very high, on the order of \$6–14 per gasoline-equivalent gallon. Such costs may be “non-starters” for most investors. Only governments could support a process of building capacity and achieving enough volume and experience to bring down these costs to eventually have a chance to be competitive without subsidies.

It is important to note that the table and figure do not account for vehicle efficiency using those fuels; in all liquid fuel cases, the efficiency of ICEs would be similar to those using gasoline or diesel; however for hydrogen, fuel cell vehicles could provide a 30%–50% efficiency benefit, making this option potential much cheaper from a fuel cost standpoint; however, this option also required a complete changeover to fuel cell vehicles, and investment in hydrogen system infrastructure (storage, fuel distribution, and refueling stations) to support this. With retail hydrogen at \$6/kg (about \$6/gallon gasoline-equiv), fuel cells would be able to operate at close to the driving cost of gasoline vehicles. And this type of cost may be the easiest to achieve of all the electrofuels. Thus, hydrogen and fuel cells are often considered to be among the most promising of these options.

National/Regional Progress to Date and Targets

So where do countries actually stand on some of these in terms of progress and plans? As shown in Table 2, as of late 2019 there has been little commercial scale progress in any area except electric vehicles. EV sales shares have been rising though the 2020 pandemic

Table 2 EV status and targets as of 2019 for key countries and regions

	<i>California</i>	<i>USA</i>	<i>EU</i>	<i>France</i>	<i>Norway</i>	<i>Japan</i>	<i>China</i>	<i>World</i>
2019 New light-duty electric vehicle (EV) sales share	8%	2%	3%	4%	56%	2%	5%	2.6%
Target EV sales shares or ICE ban year	15% by 2025, 100% by 2035	None	None	ICE ban by 2040	ICE ban by 2025	None	None	NA

Source: IEA GEVO (2020), various news reports of EV sales targets and ICE ban dates; CA governor's executive order, October 2020.

may slow the rates in some countries. Some countries (and the US state of California) have also set ambitious targets for future sales, or set a ban on ICE vehicle sales in the future.

The EV shares in these countries have grown fairly rapidly, though the question is whether this rate of increase will continue into the future. California notably recently set truck ZEV sales requirements to 2035, reaching 100% ZEVs (zero-emission vehicles) for cars and as high as 75% ZEVs in the sales mix for medium duty (delivery type) trucks. These can be battery electric, plug-in hybrid or fuel cell vehicles. Many other countries have set "soft" targets for such a phase out in the 2035–40 time frames, with Norway leading the way with a 2025 phaseout plan.

Scenarios for Achieving CO₂ Reduction Targets

In 2019, the IEA created a "Sustainable Development Scenario" that is believed would be consistent with a 1.7–1.8°C warming increase (WEO 2019). It projects to 2040, and the scenario would cut total energy-related GHGs between 2020 and 2040 by about 40%, on a path toward net-zero emissions globally by around 2060–65. Extracts from **Tables 2 and 3** of that WEO show a reduction in transportation CO₂ of about 35% between 2025 (its peak year) and 2040.

How does the IEA suggest achieving such reductions? Basically by introducing low carbon technology vehicles and fuels, and ramping up sales of these as fast as is reasonably possible. In particular, electric vehicles play a major role for cars and trucks, and low carbon fuels play a major role for shipping and aviation. IEA's scenario includes key milestones for 2030 and 2040. These are shown in **Table 3**, along with this author's estimates of where they would need to be by 2050.

The IEA scenario sets their 2030 and 2040 targets based on ambitious policy and technology progress, and even these do not put the world on a path to achieving near zero transportation by 2050. As per this author's additional targets for 2050, very rapid changes will be needed between 2040 and 2050 to hit targets. But if the world has made the types of progress envisioned by the IEA by 2040, this "big push" to 2050 should be possible in all transportation modes. Also, the figures in the table show global averages; in 2030 and 2040, leading countries will need to do even better than the IEA global averages shown.

Overall, the IEA is showing that even to achieve a global transportation system emissions reduction of 30% between 2017 and 2040, consistent with a broader 1.7–1.8°C target, a veritable revolution will be needed. And this revolution will need to happen faster in wealthier countries, more able to invest in new technologies and systems, with other countries following along and adopting technologies as they become cheaper and available in large volumes. The lead countries and regions appear likely to be Japan, China, the EU, and the US—or at least California.

How Can We Achieve the "Revolution"?

A look again at **Table 3** provides some sense of how fast the world must move in particular areas. For example, electric PLDVs reach 236 million by 2030, out of around 1.5 billion expected vehicles on the planet in that year. But to reach that stock level, sales must

Table 3 IEA scenario targets to 2040, compared to needed 2050 targets

	<i>2030 (IEA)</i>	<i>2040 (IEA)</i>	<i>2050 (This author)</i>
Light duty electric vehicle stocks	236 million (around 20% of world total)	933 million (around 60% of world total)	Near 100% of around 2 billion vehicles on world's roads
Average carbon intensity of world's truck fleets	33% reduction in average CO ₂ intensity vs. 2020	60% reduction	90%–100% reduction
Total shipping emissions	Similar to 2020	20% reduction	80% reduction
Total aviation emissions	Similar to 2020	10% reduction	80% reduction

Notes: 2030 and 2040 changes from IEA WEO (2019), 2050 targets set by this author given overall goals.

grow very rapidly and may need to account for at least 30% worldwide in that year, or around 30–40 million. This in turn suggests that electric vehicle sales in leading countries may need to be 40%–50% in 2030.

Similarly, freight truck carbon intensity drops by a third, meaning carbon intensity of new trucks in leading countries probably should drop by at least half. This can be achieved through a combination of conventional truck fuel economy improvement, and the introduction of electric and hydrogen trucking, with rapid growth in sales.

Shipping and aviation can be thought of in global terms (as in the table), and move more slowly than land transport, but leading countries still need to push harder, buying efficient new planes and promoting low carbon fuels; at least a 30% reduction in carbon intensity in these modes (taking into account travel growth) will be needed.

To achieve targets like this, in the next 10–15 years there must be a relatively strong reliance on electric vehicles for light duty; electric and hydrogen vehicles for trucks; and liquid biofuels for ships and aviation. The most advanced liquids are not yet available and this may slow progress for these vehicles, but the hope is that these vehicles can catch up post-2030. An additional challenge will be bringing down the high cost of such fuels, along with hydrogen. The most likely way to do this is to force the technologies and fuels into the market (regulations and pricing policies), and create the scales needed to achieve much lower costs.

Eventually, by 2050, there is a reasonable prospect that very large volumes of electricity, hydrogen, and advanced zero (net) carbon liquids will be available at reasonable prices, and allow a near-zero CO₂ transportation system to be realized.

Concluding Remarks

This review suggests that:

1. Reaching very low CO₂ emissions by the 2050–60 timeframe will require very rapid changes in the sales of vehicle technologies by 2030 and 2040. The changes will be unprecedented and will require very strong policy efforts around the world to achieve.
2. Some technologies, notably vehicle electrification and electricity carbon reduction, have a “head start” over most other technology/fuel options, and may be among the least expensive overall for consumers, given their efficiency and the declining cost of batteries.
3. Hydrogen can become a reasonably low-cost energy source when used with fuel cell vehicles, since these vehicles are very efficient, making this a front-runner option; however, it does require a complete change in vehicle technology and refueling infrastructure and an associated policy commitment.
4. Today’s biofuels (ethanol from grains and biodiesel or renewable diesel from plant oil seeds) provide a competitive near-term option, but with limited CO₂ reduction benefits and significant sustainability concerns. Longer-term options, mostly derived from cellulosic feedstocks, can help address these concerns and provide “drop-in” fuels, but are likely to be expensive.
5. There are other technology/fuel options that could provide zero-emissions transportation, notably, “electrofuels”—producing gaseous or liquid fuels from electricity with carbon capture. The liquid fuels could generally be compatible with today’s gasoline and diesel vehicles, so have a major advantage in avoiding the need for a vehicle technology transition.
6. However, the advanced biofuel and electrofuel options are very expensive in the near term and would require enormous investments (perhaps tens or hundreds of billions of dollars) to achieve large-scale volumes with a chance to reach long-term, more competitive production costs. Returns during a transition period could be low or negative, raising questions about who would be willing to make these investments. Strong government interventions will likely be needed.

Overall a range of technologies and fuels (energy carriers) exist to achieve very low carbon transportation systems by 2050–60, if strong commitments to this transition were made. But the costs of transitions may be high and strong policy pressure must be applied to ensure a rapid, smooth transition. The current pandemic (as of time of writing) has slowed travel growth, but also may be slowing the uptake of new technologies. It is highly uncertain what near-term or long-term effects it will ultimately have, but it must not stand in the way of a rapid transition to very low carbon technologies if the targets are to have any chance of being met.

References

- Brynolf, S., Taljegard, M., Grahn, M., Hansson, J., 2018. Electrofuels for the transport sector: a review of production costs. *Renewable Sustainable Energy Rev.* 81, 1887–1905. <https://research.chalmers.se/en/publication/500505>.
- Clean Cities Alternative Fuel Price Report, US Department of Energy, January 2020. https://afdc.energy.gov/files/u/publication/alternative_fuel_price_report_jan_2020.pdf.
- Fulton et al., 2019. Technology and Fuel Transition Scenarios to Low Greenhouse Gas Futures for Cars and Trucks in California, UC Davis, ITS-Davis Research Report-UCD-ITS-RR-19-35. <https://escholarship.org/uc/item/8wn8920p>.
- IEA ETP, 2020. Energy Technology Perspectives, International Energy Agency, OECD Paris. <https://www.iea.org/reports/energy-technology-perspectives-2020>.
- IEA GEVO, 2020. Global Electric Vehicle Outlook 2020, OECD Paris. <https://www.iea.org/reports/global-ev-outlook-2020>.
- IEA WEO, 2019. World Energy Outlook, International Energy Agency, OECD Paris, <https://www.iea.org/reports/world-energy-outlook-2019>.
- IEA, 2019. Transport sector CO₂ emissions by mode in the Sustainable Development Scenario, 2000–2030. <https://www.iea.org/data-and-statistics/charts/transport-sector-co2-emissions-by-mode-in-the-sustainable-development-scenario-2000-2030>.
- IEA, 2020. CO₂ Emissions from Fuel Combustion Overview, OECD Paris. <https://www.iea.org/reports/co2-emissions-from-fuel-combustion-overview>.
- IPCC AR5, 2013. Working Group 1, Climate Change 2013: The Physical Science Evidence, Summary for Policy Makers, Geneva. <https://www.ipcc.ch/report/ar5/wg1/summary-for-policy-makers/>.

IPCC, 2019. 1.5 Degree Special Report, summary of 1.5 degree emissions, concentrations and temperature pathways. https://www.ipcc.ch/site/assets/uploads/sites/2/2019/06/SR15_Full_Report_High_Res.pdf.

Witcover, J., Williams, R., 2020. Comparison of "advanced" biofuel cost estimates: trends during rollout of low carbon fuel policies. *Transport. Res. D: Transport Environ.* 79, 102211. <https://www.sciencedirect.com/science/article/pii/S1361920919312222?via%3Dihub>.

Recommendations for Further Reading

- Any of the IEA publications referenced in this report
- IPCC 1.5 degree report (referenced)
- [Fulton et al. \(2019\)](#) (referenced) for consideration of a transition scenario of this type for the US state of California.

Governor of California Executive Order N-79-20, September 2020. <https://www.gov.ca.gov/2020/09/23/governor-newsom-announces-california-will-phase-out-gasoline-powered-cars-drastically-reduce-demand-for-fossil-fuel-in-californias-fight-against-climate-change/>.

Disabled Travelers

Bryan Matthews, University of Leeds, Leeds, United Kingdom

© 2021 Elsevier Ltd. All rights reserved.

Introduction	359
Some Context	359
Disabled Traveler Statistics	360
Why are Disabled Travelers' Travel Patterns Different to Those of Nondisabled People?	360
Why do These Difficulties Exist and Persist?	361
Why Does This Matter?	361
What Can be Done?	362
Technology to Assist	362
Speculations About the Future	362
References	363
Further Reading	363

Introduction

Disability affects a substantial proportion of the world's population, and it is therefore important to understand issues affecting disabled travelers in order to ensure that the design and delivery of transport systems and services take these issues into account. The best estimates show that disabled people comprise approx. 15% of the world's population (WHO, 2019; United Nations, 2015). This proportion has been increasing, and current expectations are that these numbers will continue to rise over the coming decades. The upward trend in disability stems from the trend toward an ageing population, and the higher risk of disability amongst older age-groups, and the increasing prevalence of chronic health conditions, such as diabetes, cardiovascular disease, cancer, and mental health disorders. The proportion of the population affected is even more significant if we consider the extended concept of persons with reduced mobility (PRM). This includes not only disabled people, but also other people who face barriers to their mobility, including older people and people with temporary impairments (such as a broken limb). It is estimated that PRM comprise approximately a third of the population (isemoa.eu).

It is timely to reflect on the issues facing disabled travelers. First, it is now over two decades since countries began to enact legislation to promote greater inclusion of disabled people in society, and 15 years since the UN drew up the Convention on the Rights of Persons with Disabilities (UNCRPD). Secondly, we are currently in the midst of a technological shift affecting transport in potentially profound ways.

This article seeks to provide a concise review of the current state of knowledge in relation to disabled travelers. It will start with some context and then some statistics, before addressing the questions of why disabled people's travel patterns are different to those of nondisabled people, why this situation exists and why it matters. The article then turns to look at what can be done, to provide pointers to some examples of good practice and to consider the possible implications of technological and other future trends for disabled travelers. A set of further reading is then provided, for the interested reader to delve more deeply.

Some Context

How one defines "disability" is contested, but the social model of disability is arguably the more useful perspective for the purposes of public policy. The different ways of understanding and conceptualizing disability vary in relation to the extent to which it is viewed as a medical issue or a social issue; that is, whether the source of disability is viewed as emanating from the individual's medical condition or from society's failure to organize itself in a way that accommodates the diversity of the population (Oliver, 1983; and UPIAS, 1976). Viewing it from a social perspective has the effect of turning attention to how society can adapt and become more inclusive and accessible.

The social model of disability draws a distinction between a person's impairment and the disablement they experience. Impairments are defined as conditions or characteristics specific to individuals, such as deafness, blindness, autism, wheelchair-use etc. Disablement is defined as the oppression experienced by people with impairments. That is, impairment is something that belongs to the person, whereas disablement is something that is done to the person.

Having identified the distinction between disability and impairment, it is important to acknowledge the wide range of impairment types and, thereby, the diversity amongst disabled people, as this has a direct bearing on transport and travel. Impairment types are grouped differently by different organizations. One generic disaggregation is into physical, sensory, and cognitive impairments.

Table 1 Disability prevalence disaggregated by impairment for Great Britain (millions)

	2002/03	2003/04	2004/05	2005/06	2006/07	2007/08	2008/09	2009/10	2010/11	2011/12
Mobility	6.3	6.2	6.0	6.2	6.2	6.3	6.4	6.3	6.4	6.5
Difficulty with Lifting, carrying	5.8	5.9	6.0	6.1	6.0	6.0	6.1	6.0	6.1	6.3
Manual dexterity	2.4	2.6	2.5	2.5	2.6	2.6	2.7	2.6	2.7	2.8
Continence	1.2	1.3	1.2	1.6	1.5	1.5	1.5	1.5	1.7	1.8
Communication	1.3	1.7	1.9	2.0	1.9	2.0	2.0	2.1	2.0	2.2
Memory/concentration/learning	1.7	2.0	2.0	2.0	1.9	2.0	2.2	2.2	2.3	2.5
Recognizing when in danger	0.4	0.6	0.6	0.6	0.7	0.7	0.7	0.7	0.8	0.8
Physical coordination	N/A	N/A	2.2	2.2	2.4	2.4	2.4	2.4	2.6	2.7
Other	1.4	1.9	2.5	3.2	3.2	3.4	3.5	3.8	3.9	4.1
At least one impairment	10.4	10.1	10.1	10.8	10.4	10.6	10.9	11.0	11.2	11.6

Source: Gov.UK, (2014). *Official Statistics. Disability prevalence estimates 2002/03 to 2011/12 (Apr to Mar)* [online] Available from: <https://www.gov.uk/government/statistics/disability-prevalence-estimates-2002/03-to-2011/12-Apr-to-Mar> [Accessed 06/11/2020]

By way of example, **Table 1** shows the incidence of disability, disaggregated by impairment type for Great Britain. It can be seen that the total population of disabled people is some 11.6 m (approx. 20% of the total population), and yet the disaggregated numbers add up to more than this total. This is because many disabled people have more than one impairment.

To take just one comparative example as a means of illustrating the importance of understanding the role of impairment type, the experiences of a disabled traveler who is a wheelchair user will be different in a number of ways to the experiences of a disabled traveler who has a visual impairment. For instance, if a wheelchair user embarks on a journey and discovers that completing that journey necessitates the use of a flight of steps, they will either be unable to complete that journey or have to divert their journey in order to take a route that allows them to avoid those steps. In contrast, so long as the visually impaired person is aware of those steps, then their journey will not usually be disrupted. On the other hand, a visually impaired person waiting at a bus stop may find it extremely difficult to know when the specific bus service they need has arrived at the stop, whilst a wheelchair user would typically have no difficulty with this. Some aspects of being a disabled traveler will, however, be common across different impairment types, such as the potential for experiencing negative attitudes from other travelers or from transport staff.

Disabled Traveler Statistics

On average, disabled people travel substantially less than nondisabled people, and this is not simply a reflection of their preferences. Evidence shows that disabled people travel a third less than is the average for nondisabled people (NTS, 2014) in terms of distance traveled. Furthermore, Clery et al. (2017) look at proportions of the population who make fewer than 400 trips a year. They find that over 26% of disabled people who have difficulty with one or more transport mode make fewer than 400 trips per year, whilst amongst nondisabled people this figure is 13%.

Research also demonstrates differences in transport use by disabled people with different types of impairment. For instance, Clery et al. (2017) find that 64% of people with self-care difficulties, 59% of people with balance difficulties and 52% of people with sight difficulties never use public transport; by comparison, 42% of nondisabled people never use public transport.

In part, differences in travel patterns between disabled and nondisabled people can be explained by other differences between disabled and nondisabled people, such as employment rates (unemployment rates being much higher amongst disabled people than nondisabled people) and perhaps individual preferences. However, there is also evidence that disabled people would travel more if conditions were improved.

Why are Disabled Travelers' Travel Patterns Different to Those of Nondisabled People?

Disabled travelers face a range of difficulties when traveling that nondisabled travelers do not face, and this is probably the key reason why trip rates and travel patterns diverge. The difficulties faced by disabled people when making journeys are often referred to as "barriers", rendering parts of the transport system partially or entirely inaccessible. These barriers have been categorized into a number of groups:

1. Physical design (of the built environment and of transport infrastructure and vehicles);
2. Transport service availability/proximity;
3. Financial affordability;
4. Information and awareness; and
5. Training.

Thinking about physical barriers, a wheelchair user will find it difficult or impossible to use parts of the transport system that require the use of steps. Furthermore, if there is uncertainty about whether a particular journey made by a particular mode will or will not require the use of steps, this will serve to discourage a wheelchair user from making that journey by that particular mode. It is these sorts of considerations that suppress trip rates and constrained mode choice, and highlight the importance of the “trip chain” and of specific accessibility-related travel information.

The concept of the “trip chain” is important as, just as a chain is only as strong as its weakest link, a trip is only as accessible as its least accessible link. That is, any journey will be made up of a number of stages or links. These can include the walk from one’s front door to the nearest bus stop, accessing and waiting at that bus stop, purchasing a ticket and boarding the bus, locating a seat and so on. At any of these links, disabled people may experience difficulties due to one or other accessibility problem associated with that link. The concept of the trip chain highlights the potential for a significant accessibility problem at just one of these links to render the whole journey impossible.

In a transport system that comprises some parts that are accessible, some that are partially accessible and others that are entirely inaccessible, the role of specific accessibility-related information becomes vital. If a disabled traveler knows that each stage of the journey they wish to make is accessible for them, then this will give them greater confidence to make that journey. If, on the other hand, a disabled traveler knows that one or more stages of the journey they wish to make will be partially or entirely inaccessible for them, they will be able to plan accordingly—whether that be to seek assistance for those difficult stages of the journey, or to somehow make alternative arrangements.

Added to the groupings of barriers identified above, there is the specific issue of the safety of disabled travelers. It is observed that disabled travelers tend to self-regulate their behavior in order to avoid dangers (e.g., that might arise from their diminished reaction-times, increased physical frailty etc.) and mitigate accident risk. However, this tends to mean that they suppress their travel and mobility in some way, and so safety becomes closely linked with issues of access, inclusion, physical activity, wellbeing, and public health.

Why do These Difficulties Exist and Persist?

Thinking within the social model of disability, these difficulties stem from transport systems and services having been designed without regard to how disabled travelers would use them. The practice of “inclusive design” is a relatively recent one (Design Council, xxxx) and much of the transport system we have today dates back to an era when attitudes towards and awareness of disabled people were very different to those that exist today.

In more recent times, governments and international agencies have introduced legislation, regulations and guidelines in an effort to ensure that transport systems become more accessible for disabled travelers. The Americans with Disabilities Act (ADA), in 1990, was something of a landmark, creating a set of rights for disabled Americans and a set of obligations for service providers. Since then, many countries have implemented broadly similar legislation. At the international level, the United Nations has drawn up the Convention on the Rights of Persons with Disabilities (UNCRPD), which includes articles on accessibility and mobility, whilst the EU has introduced a comprehensive framework of passenger rights, explicitly encompassing disabled passengers.

In some instances, the legislation includes regulations and/or guidelines setting out how service providers should or must adapt services to ensure they are inclusive. For instance, in the UK, the Equality Act (2010) incorporates a set of Rail Vehicle Accessibility Regulations that set out the requirements of new rail vehicles. This approach of providing regulations has been praised for its clarity, but it has also been criticized for stifling innovation. In either case, there are calls for such regulations to be reviewed and updated periodically, to ensure that they continue to reflect the requirements of disabled travelers.

Whilst regulations can be monitored to ensure they are adhered to, rights must be enforced, and this can prove difficult. Some countries have introduced public enforcement bodies to assist with this, but even where these exist, enforcement is often down to private individuals bringing forward legal cases, which many are reluctant to do.

It is often argued that inclusive design costs relatively little if incorporated into project plans at the outset, and a variety of design guidance has emerged to enable inclusion to be designed in from the start. More recently, concern has focused on how disability access is taken into account as part of transport investment decision-making processes.

Why Does This Matter?

There are a number of ethical, legal, social, and economic arguments as to why travel for disabled people matters. From an ethical and legal perspective, it is argued that disabled people should have an equal right to travel and should not be subject to discrimination on the grounds of being disabled. These arguments have served to frame the legislation referred to above. Whilst they are quite commonly accepted, there remains some disagreement regarding how these rights fit alongside other sets of rights and about how to fund the cost implications arising from these rights. As a means of managing this disagreement, the concept of “reasonableness” has been introduced.

Social and economic arguments center on the diminished life opportunities disabled people face as a consequence of their travel being limited due to the constraining effects of inaccessible transport systems. That is, the suppression of disabled traveler’s mobility can lead to a range of social and economic problems, including heightened levels of loneliness, reduced levels of physical activity,

reduced levels of employment and reduced levels of economic, and cultural participation. Furthermore, these can then lead to consequent health problems, which also have both social and economic dimensions.

What Can be Done?

It is generally acknowledged that the key to tackling problems faced by disabled travelers is to address each of the key barriers to accessible travel, referred to above. This is likely to involve incorporating inclusive practices into the design of infrastructure and services, ranging from physical infrastructure, customer support, information provision and digital technologies. As a starting point, much will be achieved by complying with legislative regulations, as referred to above, and with industry guidelines (refs xx). Furthermore, there are a number of excellent compilations of good practice that have been compiled on behalf of the European Commission (MEDIATE, 2014; and xx, 2019).

Technology to Assist

The following broadly adopts the categorization set out in “New Transport Technology for Older People” (OECD, 2004), which remains a very relevant document. This makes a three-way division of technological applications, as to whether they are applied to a vehicle, a person or to the physical infrastructure.

“In-vehicle systems” for private cars have generally been developed by motor manufacturers based on commercial incentives to increase driver safety, confidence and comfort, with an awareness of the growing size of the older driver market. A range of systems exist including:

- Adaptive Speed Control;
- Night Vision Displays;
- Navigation Systems;
- Gap-Estimation Aids; and
- Workload Management Systems.

Personal technologies are being used to develop solutions for different groups of disabled travelers. A great deal of effort has focused on technologies to assist visually impaired people. To a lesser extent, solutions have also targeted wheelchair users and others with reduced mobility, and for people with hearing or other communications impairments. Key amongst these technologies is personal navigation aids—which can be disaggregated into two further subcategories:

1. Passenger guidance for visually impaired people seeks to make use of satellite navigation technologies, in conjunction with screen magnification and speech synthesis technologies, in order to provide route navigation and way finding information to people with a visual impairment. Passenger guidance based on GPS and mapping technologies, often referred to as Personal Positioning Systems (PPS), have been implemented via specialist and, more recently, more mainstream applications (apps) designed for internet-enabled mobile phones and smart phones, since the required technologies are often already integral to the mobile device. A number of such apps exist. Furthermore, the Royal National Institute of Blind People (RNIB) has been conducting ground-breaking work in the areas of Augmented Reality, Electronic recognition, and artificial vision. They see these as forming part of a “Blended technological” solution, whereby technologies serve to complement techniques already being used, such as the long white cane and the guide dog.
2. Tailored accessibility-related information and communication apps for wheelchair users and people with otherwise reduced mobility draw on digital maps and crowd-sourced data, often provided in real time.

New technologies have also been embedded into the physical infrastructure of the built environment. Triggered Information, via the use of Beacons, has been the subject of research and development, and the RNIB has recently launched its Update to their REACT system of talking beacons for visually impaired people. A further idea is that of “Electronic lead lines”, whereby RFID tags are fitted to the visually impaired person or to their long white cane, which then interact with tags in the environment. The aim is to enable the visually impaired person to be prompted and guided in unknown environments, but it could also serve to improve cognitive awareness of one’s surroundings. The Austrian “ways4all” project has undertaken ground-breaking work in this area, establishing several demonstration sites in Vienna. However, it is not possible to embed RFID tags everywhere, so the question returns to the issue of how to use what is already there to best effect.

Speculations About the Future

There are several scientifically exciting technological developments taking place, but a number of difficulties remain. Technology take-up, particularly amongst older age-groups and lower income groups seems to continue to be a problem. Likewise, technology abandonment has been shown to be commonplace, particularly where that technology has been designed especially for the elderly

and disabled market. This would seem to be a result of several factors including awareness of the technologies, cost to the individual, accessibility/user-friendliness of the technology itself, and a lack of availability of training specific to the technologies. Add to this the problem of battery life for the devices, and it would appear that new technologies might be best-seen as providing a complement—albeit a powerful one—to a travel environment that is itself inclusively designed.

Increasing levels of automation within the transport system is likely to have a number of advantages and disadvantages for disabled travelers. The issues encountered will differ by transport mode and by the specifics of the traveler's impairment.

On public transport, automation is likely to lead to a greater reliance on ticket machines and on online information and booking services, and reduced staffing in the system. These developments may make life easier for disabled travelers who have difficulties in accessing traditional ticket windows or with communicating via the spoken word. However, they may make life more difficult for those disabled travelers who require staff assistance with travel.

Increasing automation of private cars would be expected to make the driving task progressively less demanding. Ultimately, the driverless car (at level 5 of automation) might be expected to make private car-use universally accessible, even to those currently ineligible to drive such as blind and partially sighted people. However, ongoing consideration is needed to ensure that autonomous vehicles and that the infrastructure to support them is designed so as to be accessible. Furthermore, precaution is needed with regard to personal security issues and to what happens in the event of emergencies (such as the automation failing and the vehicle having to revert to manual drive).

Lastly, traveling in the wake of Covid-19 brings forward a range of new issues for disabled travelers. It is too early to say how travel will adapt in the coming years, but requirements for face-coverings and social distancing are two areas that pose difficulties. On the other hand, less crowded public transport and a greater potential to access opportunities remotely (without the need to travel), are likely to be of benefit to disabled travelers.

References

- Clery et al., 2017. Disabled People's Travel Behavior and Attitudes to Travel. Department for Transport.
 Isemoa.Eu, Accessible and energy-efficient mobility for all! [online] Available from: <http://isemoa.eu/> [Accessed 06/11/2020]
 National Travel Survey, 2014. Fact Sheet: Disability and Travel 2007-2014, NTS England: Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/533345/disability-and-travel-factsheet.pdf [Accessed 02/12/2020].
 OECD, 2004. New transport technology for older people; Summary and Conclusions of the Symposium on Human Factors of Transport Technology for Older Persons, held on 23-24 September 2003, Cambridge, Massachusetts. Organisation for Economic Co-operation and Development OECD, Paris.
 Oliver, M., 1983. Social Work with Disabled People. Macmillan, Basingstoke.
 UPIAS, 1976. Fundamental Principles of Disability. Union of the Physically Impaired Against Segregation, London.
 World Population Ageing, United Nations, 2015.

Further Reading

- Breemersch et al., 2018. Best practices guide on the carriage of persons with reduced mobility [online] European Commission, DG Mobility and Transport. Available from: <https://op.europa.eu/en/publication-detail/-/publication/67385059-df42-11e9-9c4e-01aa75ed71a1> [Accessed 06/11/2020].

Health Impacts of Connected and Autonomous Vehicles

Soheil Sohrabi, Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	364
CAVs Present Benefits and Risks to Public Health	364
First-order Impacts	364
Second-order Health Impacts	366
Quantifying the Impacts	367
Policy Recommendation	367
Conclusion	369
Acknowledgment	370
References	370
Further Reading	371

Introduction

The impacts of connected and automated vehicles (CAVs) on traffic flow, travel behavior, and infrastructure have been discussed in previous chapters (10110, 10110, 10636, and 10640). The focus of this chapter is on the health implications of CAVs, which has received relatively less research attention in recent years (Milakis et al., 2017). The discussion presented in this chapter links the previous chapters regarding the impacts of CAVs on transportation (10110, 10400, 10636, and 10640) and the discussion on the impacts of transportation on public health (10731, 10732, 10733, 10737, 10738, and 10742).

This chapter investigates three major questions surrounding CAVs and public health:

1. How do CAVs impact public health?
2. To what extent do CAVs impact public health?
3. What actions are needed before and after implementing CAVs?

To answer these questions, I collected and studied research on the health implications of automated vehicles (AVs), connected vehicles, and CAVs, with a special focus on recently published review studies (Crayton and Meier, 2017; Dean et al., 2019; Rojas-Rueda et al., 2020; Singleton et al., 2020; Sohrabi et al., 2020). The answers to these questions should benefit transportation engineers, urban planners, car manufacturers, the healthcare sector, and policymakers. It could help them understand the effects of CAV implementation on public health, allowing them to make more informed decisions regarding investments in CAVs and their deployment. It also is expected to urge stakeholders to intervene and mitigate negative health consequences and facilitate the positive impacts of CAV implementation.

CAVs Present Benefits and Risks to Public Health

There are two orders of impact when investigating CAVs health implications. First-order impacts refer to those resulting from changes in transportation—including the systems, infrastructure, land use, and activities behind it. Second-order impacts are consequences of the first-order impacts, as well as the policies and actions that govern the negative consequences of CAVs. In subsequent sections, the potential CAVs' first- and second-order health implications are discussed through 32 and 6 pathways, respectively (Fig. 1).

First-order Impacts

A study by Sohrabi et al. (2020) proposed a framework to identify the health impacts of AVs through the changes in transportation. To this end, first, the changes in transportation were investigated using seven points of impact, including transportation jobs; transportation equity; land use and the built environment; traffic flow; trip, mode and route choice; transportation infrastructure; and traffic safety. Then, 14 areas where transportation can affect public health were discussed—access, mobility independence, social exclusion, contamination, greenhouse gases, community severance, heat, noise, air pollution, stress, green spaces, physical inactivity, electromagnetic fields, and motor vehicle crashes. Ultimately, the AVs implementation was linked to public health through 32 pathways. Fig. 2 shows the proposed framework by Sohrabi et al. (2020). The pathways are labeled as having negative, positive, and uncertain impacts, given the direction of AVs' health implications through each pathway.

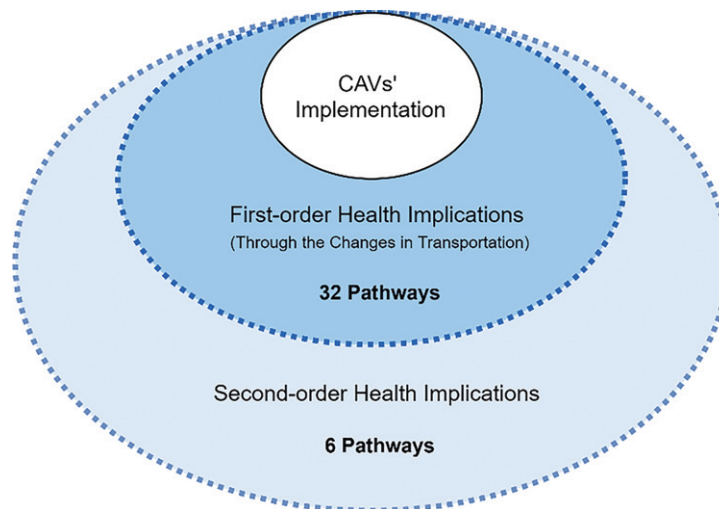


Figure 1 CAVs first- and second-order health implications.

Considering the communication capabilities of CAVs, along with its advantages and infrastructure requirements, the 32 pathways between CAVs and public health are identified. I have discussed the 32 pathways in the context of CAVs' seven points of impact on transportation:

1. *Transportation jobs*: The deployment of CAVs can eliminate driving jobs as well as vehicle service-related jobs (Groshen et al., 2019; Pettigrew et al., 2018). This loss of jobs can result in social exclusion, which has adverse health consequences (Holt-Lunstad et al., 2015).
2. *Transportation equity*: CAVs have the potential to provide mobility options for individuals with disabilities, the elderly, and nonlicensed drivers (Bennett et al., 2019a, 2019b). This would facilitate access to healthy food and healthcare, promoting public health by improving mobility independence and social inclusion (Holt-Lunstad et al., 2015).
3. *Land use and built environment*: Cheaper, more comfortable, and easier travel facilitated by CAVs would incentivize drivers to travel longer distances, ultimately resulting in urban sprawl (Milakis et al., 2017). As a result of urban sprawl, the total vehicle miles traveled (VMT) would increase (Childress et al., 2015), and accessibility would be reduced in urban areas (Milakis et al., 2017). On the other hand, there would be a reduction in parking needs (Zhang et al., 2015) and space would be freed in urban areas. The increases in VMT can be translated into increases in adverse public health impacts from traffic-related noise (WHO, 2018), heat (Cheng et al., 2014), greenhouse gases (GHG) (Haines et al., 2006), air pollution (Cesaroni et al., 2014; Kurt et al., 2016), and contamination (Jaishankar et al., 2014). Urban sprawl restricts accessibility and community interactions, resulting in health consequences from social exclusion (Holt-Lunstad et al., 2015). Potential changes in urban design can affect public health through accessibility to green spaces (Gascon et al., 2016) and active transportation (Kyu et al., 2016); however, the direction of these changes is unclear.
4. *Traffic flow*: Since CAVs can gather information by communicating with infrastructure and other vehicles, and sensing the environment, they drive more smoothly by optimizing the acceleration and deceleration (Milakis et al., 2017). Given that CAVs emit less harmful pollutants—including heat, GHGs, air pollutants, and contaminants—they have a low impact on human health. The implications CAVs have on health owing to noise are uncertain. On the one hand, traffic noise may be reduced by less acceleration and deceleration of CAVs. On the other hand, the expected increase in flow speed can increase overall noise emissions (WHO, 2018). While traffic-related stress is associated with adverse health impacts (Hackett and Steptoe, 2017; Kivimäki and Steptoe, 2018; Schnurr and Green, 2004), self-driving CAVs can eliminate the stress caused by driving and traffic congestion, promoting public health.
5. *Trip, mode, and route choice*: CAVs can improve the mobility of individuals with disabilities, those who are nonlicensed, and the elderly, resulting in increased travel demand (Milakis et al., 2017). In addition, more affordable travel options encourage a modal shift from public transit and active transportation to private cars (Milakis et al., 2017). As a result, an increase in VMT is expected after the implementation of CAVs. As discussed previously, the higher the VMT, the higher the levels of adverse health impacts owing to noise, heat, GHGs, air pollution, and contamination. Furthermore, the modal shift from active transportation to motorized vehicle transportation can reduce physical activity (Sohrabi et al., 2020) and negatively impact public health (Kyu et al., 2016). An increase in motorized traffic is associated with community severance (Mindell et al., 2017), which also has adverse health impacts (Cohen et al., 2014; Emond and Handy, 2012).
6. *Transportation infrastructure*: We expect a change in transportation infrastructure needs with the introduction of CAVs, depending on the potential changes in travel demand and transportation modal splits. Paved surfaces, such as roads and parking lots, are a source of heat in urban areas (Coseo and Larsen, 2014). Also, the growth in transportation networks contributes to community severance, as it occupies green spaces (Sohrabi et al., 2020). In addition to the negative health impact of transportation

1. Traffic accidents are a major source for organ donations; in 2019, 12% of all donations in the United States were from motor-vehicle crashes ([Organ Procurement and Transplantation Network, 2020](#)). Consequently, CAVs impacts on motor-vehicle crashes could result in changes in organ donation numbers ([Rojas-Rueda et al., 2020](#)). Given the uncertainties regarding CAVs impacts on traffic safety, the health implications of CAVs through changes in organ donation remain uncertain.

2. According to Singleton et al. (2020), CAVs can promote travel satisfaction, which (indirectly) affects life satisfaction (De Vos, 2019). Given that life satisfaction is associated with positive health impacts (Collins et al., 2009), CAVs have the potential to positively impact public health.
3. Shared CAVs can facilitate ridesharing, leading to substantial transportation and environmental benefits (Krueger et al., 2016). However, Rojas-Rueda et al. (2020) highlighted the potential negative health impacts of shared CAVs owing to the spreading of communicable diseases.
4. The partial and fully self-driving capabilities of CAVs allow passengers to use drugs and alcohol (Rojas-Rueda et al., 2020). At lower levels of automation, substance abuse can result in impaired driving and negative safety consequences.
5. The changes in transportation—particularly the need for new infrastructure—can result in changes in road constructions. Road construction creates particulate matter with diameters $<10\text{ }\mu\text{m}$, which has been associated with adverse health impacts (Giunta, 2020).
6. The work zone contributes to motor-vehicle crashes (Yang et al., 2015), and a reduction in road construction can result in improved traffic safety, promoting public health by decreasing the number of motor-vehicle crashes. As discussed before, the CAVs impacts on transportation infrastructure needs are not clear, as are the CAVs second-order health consequences through the fifth and sixth pathways.

Quantifying the Impacts

The literature on the health implications of CAVs is speculative, consisting of qualitative analyses of CAVs health impacts, rather than quantitative analyses (Sohrabi et al., 2020). Quantifying the health impacts of CAVs can help to assess the extent of their impact and address the uncertainties in their implications, as mentioned in previous sections. In addition, quantification is required to evaluate supporting policies and make more informed decisions on CAV implementations. Despite attempts to estimate CAVs health impacts (Pourrahmani et al., 2020), quantitative literature is still in the nascent stages, potentially because of the complexity of the problem. Quantifying CAVs health impacts involves interdisciplinary research, which requires integrating information, data, techniques, tools, and concepts from various fields of expertise to be able to tackle the complexity of the problem. In this section, I reviewed recent literature regarding the health-impact assessment (HIA) of transportation projects, discussed CAVs first-order health impacts and illustrated a roadmap and future requirements for quantification purposes.

HIA is used to measure and evaluate the potential health effects of policies, plans, or projects on the community, either quantitatively or qualitatively (Cole et al., 2019). There is an increasing amount of information on quantitative full-chain HIA of transportation projects and policies (Nieuwenhuijsen et al., 2017). Full-chain HIA integrates a set of tools and models to assess the health impacts of transportation projects from the source (Nieuwenhuijsen et al., 2017). Using a similar approach, the CAV implementation and recommended supporting policies can be evaluated. Fig. 3 shows the required workstream for full-chain quantitative HIA of CAVs (first-order) health impacts.

First, scenarios regarding the CAV implementation and supporting-policy need to be defined, considering stakeholder insights and needs, CAV technologies and their functionality, and CAV adoption rate. Next, changes in transportation need to be estimated for each scenario, based on the seven transportation points of impact. Travel demand modeling is the main workstream for ascertaining the extent of changes in transportation along with urban growth modeling and crash prediction modeling. CAVs roles in providing mobility options for individuals with disabilities and non-licensed drivers can be estimated after quantifying the changes in trips and mode choice. The reduction in transportation-related jobs is also dependent on changes in VMT and traffic safety (vehicle service-related jobs). Finally, the burden of disease estimation of the scenarios consists of three steps. First, the exposure of the population to environmental health determinants needs to be estimated. Among the fourteen transportation-related risk factors, GHG, air pollution, noise, and heat require further modeling. Second, the health risk estimations for each risk factor can be found using the exposure–response function from the epidemiology studies. Third, the estimated risks for each scenario should be compared with those of the base scenario (e.g., no-automation scenario) to evaluate the health implications of CAV implementation or supporting policies (Fig. 3).

Quantifying the second-order health impacts can be even more challenging, given higher levels of uncertainty stemming from CAVs mechanisms of impact and the complexities regarding potential health consequences.

Policy Recommendation

The potential negative health impacts of CAV implementation have been discussed in the previous section. Decision-makers, law-makers, the automobile industry, the health sector, transportation engineers, and urban planners need to intervene to mitigate these negative impacts. The required policies to govern the negative impacts of CAVs can be categorized into four groups: (1) those needed to govern the CAVs unintended implications, (2) those that regulate the emissions from CAVs themselves and exposure to them, (3) those to incentivize technologies and mobility advancements to support CAV implementation, and (4) those to promote research on CAVs implications. Table 1 summarizes the recommended policies.

Urban sprawl, the modal shift from transit and active transportation, and the induced demand after CAV implementation are major contributors to the increase in VMT and its negative health implications. Therefore, well-thought-out policies are required to

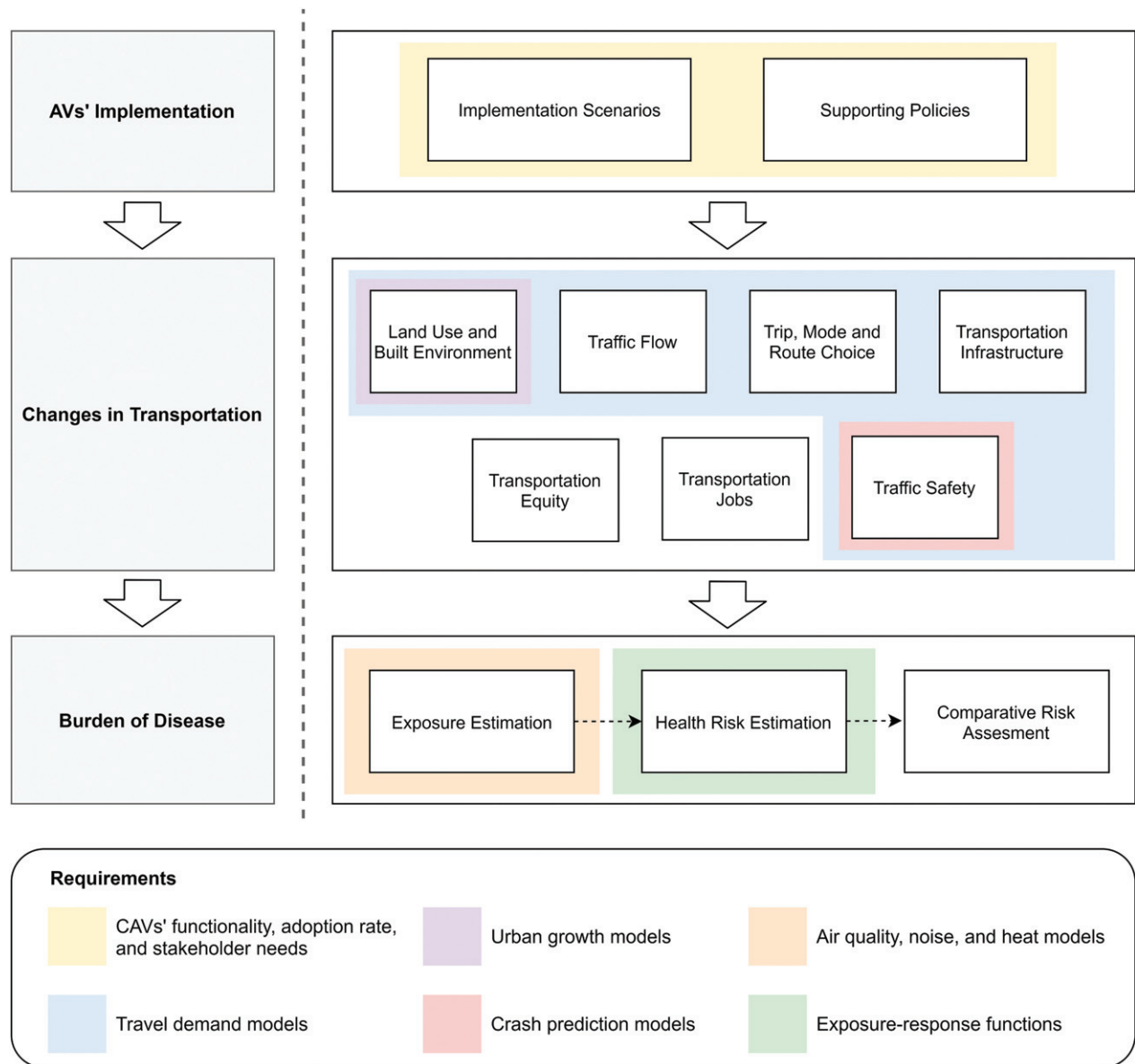


Figure 3 Suggested workstream for quantifying CAVs health implications.

control these issues. In addition, a gradual transition to CAVs mitigates the social and health impacts of job losses, as it enables workers displaced by automated driving technology to develop new skills and find new jobs ([Center for Global Policy Solutions, 2017](#)).

Regulating CAV-related emissions and exposure to them is another effective policy to mitigate CAVs health implications. In particular, policies are required to regulate CAVs noise emissions, exposure to EMFs, and substance abuse in CAVs. Promoting organ donation can also mitigate the potential shortage in organs donors in case of a reduction in motor vehicle crashes. Supporting policies to control the spread of communicable diseases can mitigate the negative health impacts of shared CAVs.

Using shared CAVs instead of privately owned CAVs can reduce the expected growth in VMT ([Fagnant and Kockelman, 2014](#); [Greenblatt and Shaheen, 2015](#); [Krueger et al., 2016](#)) and parking needs ([Zhang et al., 2015](#)). Therefore, the negative health implications of CAVs attributable to increases in VMT can be alleviated. Equipping CAVs with electric engines that produce less harmful vehicle emissions ([Buekers et al., 2014](#)) and contaminants, such as engine oil, can also reduce the negative impacts of CAVs on public health.

Finally, providing financial support for future research on the identification and quantification of CAVs health implications can promote the discussion regarding CAVs health impacts, lead to more informed decision-making regarding CAVs supporting policies, and result in more efficient implementations of this new technology. Specifically, future research is required to address the uncertainties regarding CAVs implications, with a focus on their health impacts.

Table 1 Recommended CAV supporting policies

Recommended policy		Pathway to health
Govern the negative impacts of CAVs on transportation and land use	Control urban sprawl	Access Social exclusion Contamination GHG Community severance Heat Noise Air pollution Green spaces Physical inactivity
	Control modal shift and induced demand	Contamination GHG Community severance Heat Noise Air pollution Physical inactivity
	Support job losses by a gradual transition to CAVs	Social exclusion
Regulate exposure to CAV-related emissions	Control noise emission	Noise
	Regulate substance abuse for CAVs	Motor vehicle crashes
	Promote organ donation	Organ donation
	Regulate exposure to EMFs	EMF
	Control communicable diseases spread in shared CAVs	Communicable disease
Incentivize the use of supporting technologies	Support the use and advancements in elective vehicles	Noise Air pollution GHG Heat Contamination
	Support shared CAVs	Contamination GHG Community severance Heat Noise Air pollution Physical inactivity
Research on CAVs implication	Support research to quantify the CAVs health implications and address the uncertainties in the impacts	Green spaces ^a Physical inactivity ^a Noise ^b Community severance ^c Heat ^c Green spaces ^c

^aFrom changes in land use and built environment.

^bFrom changes in traffic flow.

^cFrom changes in transportation infrastructure.

Conclusion

CAVs can have a significant effect on public health. Considering the first- and second-order health impacts of CAVs, 32 and 6 pathways between CAV implementation and public health were identified, respectively. CAVs can negatively impact public health through 19 pathways, and their impacts are uncertain through 10 pathways. This indicates the necessity for supporting policies to govern the CAVs negative health implications. Furthermore, the uncertainties regarding CAVs implications and policy

recommendations require further quantitative evaluation. This is interdisciplinary research that requires integrating information, data, techniques, tools, and concepts from various fields of expertise. Full-chain health impact assessments can help overcome the complexity of the problem and evaluate CAV implementation and supporting policies. Further research is required to address the limitations in the literature regarding CAVs health implications by (1) identifying CAVs impacts on the freight transport system, (2) investigating CAVs impacts in rural areas, (3) quantifying CAVs health impacts, and (4) evaluating the recommended policies to govern CAVs negative impacts.

Acknowledgment

The author would like to thank Dr. Dominique Lord and Dr. Haneen Khreis for their insightful feedback.

References

- Bennett, R., Vijaygopal, R., Kottasz, R., 2019a. Attitudes towards autonomous vehicles among people with physical disabilities. *Transport. Res. Part A: Policy Pract.* 127, 1–17.
- Bennett, R., Vijaygopal, R., Kottasz, R., 2019b. Willingness of people with mental health disabilities to travel in driverless vehicles. *J. Transport Health* 12, 1–12.
- Bila, C., Sivrikaya, F., Khan, M.A., Albayrak, S., 2017. Vehicles of the future: a survey of research on safety issues. *IEEE Trans. Intel. Transport. Syst.* 18, 1046–1065.
- Buekers, J., Van Holderbeke, M., Bierkens, J., Panis, L.I., 2014. Health and environmental benefits related to electric vehicle introduction in EU countries. *Transport. Res. Part D: Transport Environ.* 33, 26–38.
- Center for Global Policy Solutions, 2017. *Stick Shift: Autonomous Vehicles, Driving Jobs, and the Future of Work*. Center for Global Policy Solutions. Available from: <https://www.law.gwu.edu/sites/g/files/zaxdzs2351/f/downloads/Stick-Shift-Autonomous-Vehicles-Driving-Jobs-and-the-Future-of-Work.pdf> (June, 2020).
- Cesaroni, G., Forastiere, F., Stafoggia, M., Andersen, Z.J., Badaloni, C., Beelen, R., Caracciolo, B., De Faire, U., Erbel, R., Eriksen, K.T., 2014. Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE Project. *BMJ* 348, f7412.
- Cheng, J., Xu, Z., Zhu, R., Wang, X., Jin, L., Song, J., Su, H., 2014. Impact of diurnal temperature range on human health: a systematic review. *Int. J. Biometeorol.* 58, 2011–2024.
- Childress, S., Nichols, B., Charlton, B., Coe, S., 2015. Using an activity-based model to explore the potential impacts of automated vehicles. *J. Transport. Res. Board* 2493, 99–106.
- Cohen, J.M., Boniface, S., Watkins, S., 2014. Health implications of transport planning, development and operations. *J. Transport Health* 1, 63–72.
- Cole, B.L., Macleod, K.E., Spriggs, R., 2019. Health impact assessment of transportation projects and policies: living up to aims of advancing population health and health equity? *Annu. Rev. Public Health* 40, 305–318.
- Collins, A.L., Gleij, D.A., Goldman, N., 2009. The role of life satisfaction and depressive symptoms in all-cause mortality. *Psychol. Aging* 24, 696.
- Coseo, P., Larsen, L., 2014. How factors of land use/land cover, building configuration, and adjacent heat sources and sinks explain Urban Heat Islands in Chicago. *Landsc. Urban Plan.* 125, 117–129.
- Crayton, T.J., Meier, B.M., 2017. Autonomous vehicles: developing a public health research agenda to frame the future of transportation policy. *J. Transport Health* 6, 245–252.
- De Vos, J., 2019. Analysing the effect of trip satisfaction on satisfaction with the leisure activity at the destination of the trip, in relationship with life satisfaction. *Transportation* 46, 623–645.
- Dean, J., Wray, A.J., Braun, L., Casello, J.M., McCallum, L., Gower, S., 2019. Holding the keys to health? A scoping study of the population health impacts of automated vehicles. *BMC Public Health* 19, 1258.
- Emond, C.R., Handy, S.L., 2012. Factors associated with bicycling to high school: insights from Davis, CA. *J. Transport Geogr.* 20, 71–79.
- Fagnant, D.J., Kockelman, K.M., 2014. The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. *Transport. Res. Part C: Emerging Technol.* 40, 1–13.
- Gascon, M., Triguero-Mas, M., Martínez, D., Davdand, P., Rojas-Rueda, D., Plasencia, A., Nieuwenhuijsen, M.J., 2016. Residential green spaces and mortality: a systematic review. *Environ. Int.* 86, 60–67.
- Giunta, M., 2020. Assessment of the environmental impact of road construction: Modelling and prediction of fine particulate matter emissions. *Build. Environ.* 176, 106865.
- Goodall, N.J., 2014. Ethical decision making during automated vehicle crashes. *Transport. Res. Rec.* 2424, 58–65.
- Greenblatt, J.B., Shaheen, S., 2015. Automated vehicles, on-demand mobility, and environmental impacts. *Curr. Sustain./Renew. Energy Rep.* 2, 74–81.
- Groshen, E.L., Helper, S., Macduffie, J.P., Carson, C., 2019. Preparing US workers and employers for an autonomous vehicle future. Upjohn Institute for Employment Research 19-036.
- Hackett, R.A., Steptoe, A., 2017. Type 2 diabetes mellitus and psychological stress—a modifiable risk factor. *Nat. Rev. Endocrinol.* 13, 547.
- Haines, A., Kovats, R.S., Campbell-Lendrum, D., Corvalán, C., 2006. Climate change and human health: impacts, vulnerability and public health. *Public Health* 120, 585–596.
- Holt-Lunstad, J., Smith, T.B., Baker, M., Harris, T., Stephenson, D., 2015. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspect. Psychol. Sci.* 10, 227–237.
- Jaishankar, M., Tseten, T., Anbalagan, N., Mathew, B.B., Beeregowda, K.N., 2014. Toxicity, mechanism and health effects of some heavy metals. *Interdisciplinary Toxicol.* 7, 60–72.
- Judakova, Z., Janousek, L., 2019. Possible health impacts of advanced vehicles wireless technologies. *Transport. Res. Proc.* 40, 1404–1411.
- Kalra, N., 2017. *Challenges and approaches to realizing autonomous vehicle safety*, Santa Monica, CA, RAND Corporation.
- Kivimäki, M., Steptoe, A., 2018. Effects of stress on the development and progression of cardiovascular disease. *Nat. Rev. Cardiol.* 15, 215.
- Krueger, R., Rashidi, T.H., Rose, J.M., 2016. Preferences for shared autonomous vehicles. *Transport. Res. Part C: Emerging Technol.* 69, 343–355.
- Kurt, O.K., Zhang, J., Pinkerton, K.E., 2016. Pulmonary health effects of air pollution. *Curr. Opin. Pulmonary Med.* 22, 138.
- Kyu, H.H., Bachman, V.F., Alexander, L.T., Mumford, J.E., Afshin, A., Estep, K., Veerman, J.L., Delwiche, K., Iannarone, M.L., Moyer, M.L., 2016. Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013. *BMJ* 354, i3857.
- Milakis, D., Van Arem, B., Van Wee, B., 2017. Policy and society related implications of automated driving: a review of literature and directions for future research. *J. Intel. Transport. Syst.* 17, 1.
- Mindell, J.S., Ancaea, P.R., Dhanani, A., Stockton, J., Jones, P., Haklay, M., Groce, N., Scholes, S., Vaughan, L., 2017. Using triangulation to assess a suite of tools to measure community severance. *J. Transport Geogr.* 60, 119–129.
- National Highway Traffic Safety Administration, 2018. Critical reasons for crashes investigated in the national motor vehicle crash causation survey. NHTSA's National Center for Statistics and Analysis, U.S. Department of Transportation. Available from: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115> (January 2019).
- Nieuwenhuijsen, M.J., Khreis, H., Verlingieri, E., Mueller, N., Rojas-Rueda, D., 2017. Participatory quantitative health impact assessment of urban and transport planning in cities: a review and research needs. *Environ. Int.* 103, 61–72.
- Organ Procurement and Transplantation Network, 2020. *Deceased Donors Recovered in the U.S. by Circumstance of Death* [Online]. U.S. Department of Health & Human Services. Available from: <https://optn.transplant.hrsa.gov/data/view-data-reports/national-data/#> (August 2020) [Accessed].

- Pettigrew, S., Fritschi, L., Norman, R., 2018. The potential implications of autonomous vehicles in and around the workplace. *Int. J. Environ. Res. Public Health* 15, 1876.
- Pourrahmani, E., Jaller, M., Maizlish, N., Rodier, C., 2020. Health impact assessment of connected and autonomous vehicles in San Francisco, bay area. *Transport. Res. Rec.*, 0361198120942749.
- Rojas-Rueda, D., Nieuwenhuijsen, M.J., Khreis, H., Frumkin, H., 2020. Autonomous vehicles and public health. *Annu. Rev. Public Health* 41, 329–345.
- Schnurr, P.P., Green, B.L., 2004. *Trauma and Health: Physical Health Consequences of Exposure to Extreme Stress*, American Psychological Association.
- Singleton, P.A., De Vos, J., Heinen, E., Pudane, B., 2020. Potential health and well-being implications of autonomous vehicles. In: Dimitris, M., Thomopoulos, N., Van, W.E.E., B., (Eds.), *Policy implications of Autonomous Vehicles. Advances in Transport Policy and Planning*. Elsevier.
- Sohrabi, S., Khreis, H., Lord, D., 2020. Impacts of autonomous vehicles on public health: a conceptual framework and policy recommendations. *Sustain. Cities Soc.* 102457.
- Taeihagh, A., Lim, H.S.M., 2018. Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Rev.* 39, 103–128.
- WHO, 2018. *Environmental noise guidelines for the European Region* [Online]. Available from: <http://www.euro.who.int/__data/assets/pdf_file/0008/383921/noise-guidelines-eng.pdf?ua=1> (January 2019) [Accessed].
- Yang, H., Ozbay, K., Ozturk, O., Xie, K., 2015. Work zone safety analysis and modeling: a state-of-the-art review. *Traffic Injury Prevent.* 16, 387–396.
- Zhang, W., Guhathakurta, S., Fang, J., Zhang, G., 2015. Exploring the impact of shared autonomous vehicles on urban parking demand: an agent-based simulation approach. *Sustain. Cities Soc.* 19, 34–45.
- Zhang, Y., Lai, J., Ruan, G., Chen, C., Wang, D.W., 2016. Meta-analysis of extremely low frequency electromagnetic fields and cancer risk: a pooled analysis of epidemiologic studies. *Environ. Int.* 88, 36–43.

Further Reading

- Milakis, D., Van Arem, B., Van Wee, B., 2017. Policy and society related implications of automated driving: a review of literature and directions for future research. *J. Intel. Transport. Syst.* 17.
- Nieuwenhuijsen, M.J., Khreis, H., 2020. Transport and health; an introduction. *Adv. Transport. Health* 3–32, Elsevier.
- Sohrabi, S., Khreis, H., Lord, D., 2020. Impacts of autonomous vehicles on public health: a conceptual model and policy recommendations. *Sustain. Cities Soc.* 63, 102457.

Electric Vehicles and Health

Kanok Boriboonsomsin, University of California at Riverside, Riverside, CA, United States

© 2021 Elsevier Ltd. All rights reserved.

Introduction	372
Vehicle Emissions	372
Environmental and Health Impacts of Electric Vehicles	373
Discussion and Conclusions	374
References	375
Further Reading	375

Introduction

Electric vehicles (EVs) represent changes in both vehicle and fuel technologies from conventional internal combustion engine (ICE) vehicles. EVs are propelled partly or solely by one or more electric motors, using energy stored in rechargeable batteries. There are three major types of EVs: battery electric vehicles (BEVs), fuel cell electric vehicles (FCEVs), and plug-in hybrid electric vehicles (PHEVs). BEVs run only on electricity stored in batteries that can be recharged by the electrical grid. FCEVs also run only on electricity stored in batteries, but the electricity comes from hydrogen stored on the vehicle that is converted by fuel cell. Because BEVs and FCEVs do not have tailpipe emissions, they are sometime referred to as zero-emission vehicles. PHEVs are similar to BEVs, but in addition to batteries and electric motor(s), a PHEV also has a fuel tank and an ICE. PHEVs can run solely on electricity, during which there will be no tailpipe emission. But they can also be propelled by the ICE as in conventional gasoline or diesel vehicles when the battery is depleted. During that time, PHEVs will produce tailpipe emissions like how conventional vehicles do. Replacing the ICE with one or more electric motors results in a more energy-efficient vehicle powertrain. For instance, EVs convert over 77% of electrical energy from the grid to power the wheels whereas conventional gasoline vehicles only convert 12%–30% of energy from the fossil fuel ([U.S. Environmental Protection Agency, 2020](#)). Therefore, EVs provide significant energy efficiency and associated greenhouse gas (GHG) reduction benefits over conventional ICE vehicles.

In recent years, the impetus for EVs came from the climate target adopted in the 2015 Paris Agreement to limit the rise of the global temperature well below 2°C above the preindustrial level ([Rogelj et al., 2016](#)). To achieve this target, one of the several measures recommended by the International Energy Agency (IEA) is that at least 20% of all on-road vehicles be driven electrically by 2030. Several countries and jurisdictions have started to act on the IEA recommendation. For instance, California is leading the push in the United States to achieve zero-emission transportation. In response to state-mandated GHG emission targets, including 40% below 1990 level by 2030 and 80% below 1990 level by 2050, California has adopted a variety of incentives and regulations, such as providing EV purchase incentives for individuals and fleets, permitting single-occupant EVs to use carpool lanes, imposing EV sales mandates on vehicle manufacturers of both cars and trucks, and requiring public transit agencies to transition to 100% zero-emission bus fleets, among others.

Global cumulative passenger EV sales have grown exponentially, surpassing 4 million in fall 2018. China was driving this trend where it accounted for 37% of all passenger EV sales in the world since 2011, and around 99% of electric bus sales ([Bloomberg NEF, 2018](#)). With the rapid rise in sales, global passenger EV fleet is forecast to reach 500 millions by 2040, making up 31% of the world's passenger cars on the road ([Bloomberg NEF, 2020](#)). Early adopters of EVs, at least in Sweden and Germany, tend to be predominantly male, young, wealthy, and college-educated ([Langbroek et al. 2017](#); [Plötz et al., 2014](#)). In the United States, a survey of 1,094 people in San Francisco found that people with household income more than \$200,000 are 7% and 13% more likely to adopt BEV and PHEV, respectively, compared to people with household income less than \$75,000 ([Spurlock et al., 2019](#)). In the medium-duty sector, early adopters of EVs tend to be large fleets that utilize EVs in applications such as package delivery, fixed route transit, school bus, and utility ([Gladstein et al., 2020](#)). While heavy-duty EVs have not been fully adopted, there have been recent demonstrations of heavy-duty EVs in short haul applications such as drayage and regional distribution ([California Air Resources Board, 2020](#)).

Vehicle Emissions

Emissions associated with vehicles can be evaluated on a direct as well as a life-cycle basis. Direct emissions are emitted through the tailpipe, through evaporation from the fuel system and during the fueling process, and from brake and tire wear. Direct emissions include smog-forming pollutants (such as nitrogen oxides and volatile organic compounds), other pollutants harmful to human health (such as particulate matter and carbon monoxide), and GHGs, primarily carbon dioxide (CO₂). BEVs and FCEVs produce no direct emissions (except for from brake and tire wear), which specifically helps improve air quality in areas where they are used. On

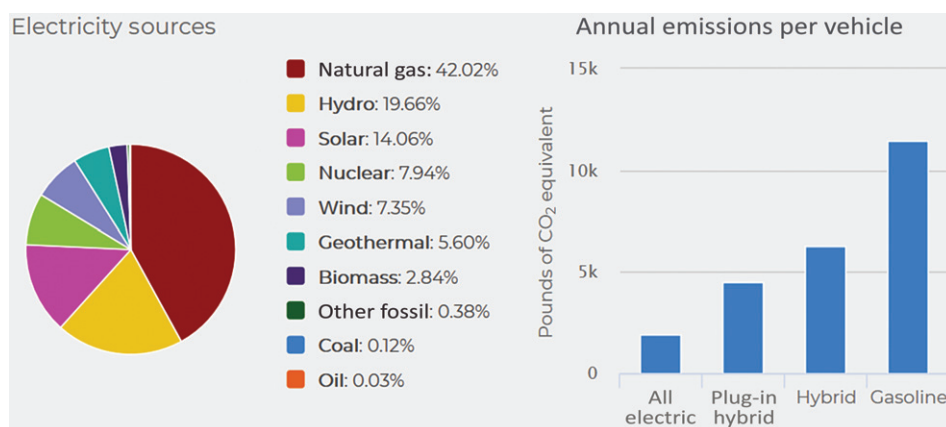


Figure 1 Average mix of electricity generation sources and annual life-cycle CO₂ emissions per vehicle in California, United States (U.S. Department of Energy, 2020b).

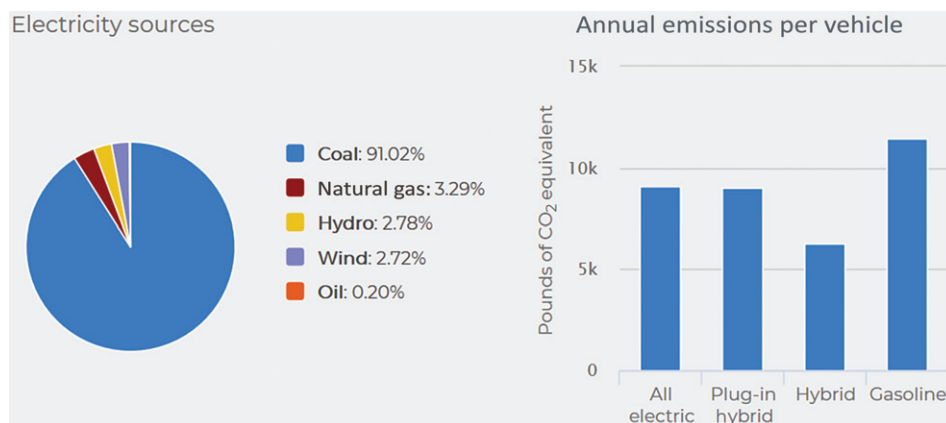


Figure 2 Average mix of electricity generation sources and annual life-cycle CO₂ emissions per vehicle in West Virginia, United States (U.S. Department of Energy, 2020b).

the other hand, PHEVs produce evaporative emissions from the fuel system as well as tailpipe emissions when operating on the ICE. However, because most PHEVs use smaller ICE and are more efficient than comparable conventional vehicles, they still produce lower tailpipe emissions even when operating on the ICE (U.S. Department of Energy, 2020a).

In addition to direct emissions produced during vehicle use, life-cycle emissions also include emissions associated with the production, processing, distribution, and recycling or disposal of vehicle and fuel. For example, for a conventional vehicle that uses gasoline as fuel, emissions are produced not only when gasoline is burned in the vehicle, but also when petroleum is extracted from the ground, refined into gasoline, and distributed to gas stations (U.S. Department of Energy, 2020a). Because of the broader scope, life-cycle emissions include more variety of harmful pollutants and GHGs beyond those included in direct emissions. For example, petroleum refining process also produces sulfur dioxide and chlorofluorocarbons (Liu et al., 2020).

All vehicles produce a large amount of life-cycle emissions. However, EVs typically produce lower life-cycle emissions than conventional vehicles because most emissions are lower for electricity generation than burning gasoline or diesel. The exact amount of these emissions depends on the mix of electricity generation sources, which varies by geographic location. Figs. 1 and 2 give examples of how the mix of electricity generation sources could significantly impact life-cycle emissions of EVs. According to Fig. 1, more than half of the electricity in California is generated from renewable sources such as wind and solar. Therefore, EVs in California have low life-cycle CO₂ emissions. On the other hand, Fig. 2 shows that almost 95% of the electricity in West Virginia is generated from fossil fuels (coal and natural gas), which makes EVs there have relatively higher life-cycle CO₂ emissions. By comparison, an EV in West Virginia produces four times as much life-cycle CO₂ emissions as the same EV in California. As of 2015, driving an average BEV would result in lower CO₂ emissions than driving a comparable gasoline vehicle in many parts of the U.S. (Nealer et al., 2015). That is now true anywhere in the U.S. (U.S. Department of Energy, 2020b).

Environmental and Health Impacts of Electric Vehicles

In terms of environmental and health benefits, BEVs and FCEVs produce no tailpipe emissions, and thus, do not expose travelers (drivers, passengers, transit riders, bicyclists, pedestrian, etc.) as well as people who live, work, or play near roadways to harmful

pollutant emissions as conventional ICE vehicles do. Exposure to these pollutant emissions has been associated with a variety of respiratory and cardiovascular diseases. Health damages from such pollutant exposure were estimated to range from 1.38 to 1.48 cents per vehicle miles traveled. Thus, replacing a conventional ICE vehicle with a BEV or a FCEV would result in approximately \$1,477 to \$1,893 (in 2015 U.S. dollars) of societal health benefits over the life of the vehicle or 120,000 miles (Malmgren, 2016). A modeling study of the Greater Toronto and Hamilton Area in Canada reveals that a shift to EVs would lead to cleaner air, save lives from respiratory and cardiovascular diseases, and provide billions in social benefits every year. The modeling results show that each BEV and FCEV that replaces a conventional gasoline vehicle would bring \$9,850 (in 2016 Canadian dollars) in social benefits from avoiding premature deaths, which is far higher than the \$5,000 purchase incentive per vehicle currently offered by the Canadian government (Environmental Defence and the Ontario Public Health Association, 2020). In addition, it is argued that the health benefits of EVs also justify the use of government incentives to build EV charging stations in support of the growing EV population (House and Wright, 2019).

BEVs and FCEVs producing no tailpipe emissions also has a side benefit in that there will be no chance of carbon monoxide poisoning, which has led to several deaths in the U.S. (Jean and De Puy Kamp, 2018). This can occur if leaving conventional ICE vehicles on in closed space (such as garage) for a long period, resulting in heightened level of carbon monoxide emission, which is odorless and colorless. Since BEVs and FCEVs require no gasoline or diesel, there will also be less exposure to evaporative emissions in garage and at gas station. While EVs still produce brake and tire wear emissions, they produce lower brake wear emissions than their conventional counterparts due to the regenerative braking technology that allows EVs to use brake pad less often. Other benefits of EVs include reduced noise pollution, faster acceleration and smoother ride, improved safety (as EVs tend to have lower center of gravity and so be less likely to roll over), lower maintenance costs, and reduced energy dependence (as electricity can be generated domestically).

On the other hand, emissions footprint of EVs also depends on how the electricity that powers them is generated. The increased demand for electricity to serve EVs may increase emissions at power plants (Nopmongkol et al., 2017). Electricity utilities often have peaker plant, which is supplemental power plant that operates only when demand for power is high. Peaker plants usually run on natural gas, and not on renewable energy sources. While natural gas power plants are cleaner than coal-fired power plants, they still emit a significant amount of nitrogen oxides and other emissions, which impacts the nearby communities. This means emissions are simply shifted from where EVs are driven to the power plants that generate electricity for these EVs, potentially resulting in environmental injustice. For example, while only 42% of the electrification in California comes from natural gas, half of the natural gas power plants in the state are concentrated in some of the most environmentally disadvantaged communities that are already overburdened with pollution. However, this will not be the case if the electricity that serve EVs is generated from renewable sources such as wind and solar.

In the life-cycle of a BEV that also includes its manufacturing and disposal, other important sources of emissions are the production of the battery and the vehicle itself (Tessum et al., 2014). Weis et al. (2016) compared BEVs and conventional gasoline vehicles, and found that BEVs have higher sulfur dioxide but lower carbon monoxide and volatile organic compound life-cycle emissions. Also, reduced direct emissions of nitrogen oxides from increased adoption of EVs may initially cause unexpected increase in secondary pollutants such as ozone and secondary particulate matter. Because of nonlinear production chemistry and titration of ozone, reduced nitrogen oxides could result in ozone enhancement in urban areas, further increasing the atmospheric oxidizing capacity and secondary particulate matter. This phenomenon has occurred in many cities around the world during the COVID-19 pandemic (Le et al., 2020). In addition, in the early days EVs used to cause unexpected safety issues because they were very quiet, making pedestrians and bicyclists unaware of their presence. Regulators in many countries have required or will be requiring EVs to generate warning sounds while traveling at low speed.

Discussion and Conclusions

BEVs and PHEVs make up the majority of EVs on the road today. While FCEVs are commercially available, they have not gained the same level of market acceptance and wide adoption as BEVs and PHEVs. This trend is expected to continued in the next few decades. The rapid growth in EV population has been contributed by a variety of EV-related incentives and regulations, which have spurred innovations (especially in battery technologies) that drive down the cost of EVs. It is anticipated that the price of a passenger EV will be in parity with a comparable ICE counterpart in the next few years. Modeling studies have shown significant environmental and health benefits of a large adoption EVs. As there are more and more EVs on the road, it will be interesting to observe these benefits of EVs in real world.

While passenger EVs have largely overcome market barriers over the last decade—they now have respectable driving range while also being more affordable—EV buses and trucks are currently going through that phase. Early adoptions and demonstrations of these heavier EVs have proven their technical and operational feasibility in many applications such as urban transit and short haul. However, it will still require further advancement in battery and charger technologies to make EV buses and trucks economically attractive, as compared to their ICE counterparts and other alternative fuel options. But once that becomes materialized and a large fraction of medium-duty and heavy-duty vehicles also turns into EVs, the environmental and health benefits to the society will amplify.

It is important to keep in mind that EVs need clean electricity from renewable sources in order to minimize their life-cycle emissions and ensure that the emissions and health impacts are not simply shifted from where EVs are used to communities near

power plants. And while EVs can help address many of the energy, climate, environmental, and public health issues, they will not solve traffic congestion problems. Therefore, a holistic approach that not only promotes a full transition to EVs and renewable energy sources but also manage whether, where, when, and how we travel should be taken in order to achieve a truly sustainable transportation future.

References

- Bloomberg NEF, 2018. Cumulative global EV sales hit 4 million. <https://about.bnef.com/blog/cumulative-global-ev-sales-hit-4-million/>.
- Bloomberg NEF, 2020. Electric vehicle outlook 2020: Executive summary. <https://bnef.turtl.co/story/evo-2020/?teaser=yes>.
- California Air Resources Board, 2020. Advanced technology demonstration projects. <https://ww2.arb.ca.gov/our-work/programs/low-carbon-transportation-investments-and-air-quality-improvement-program-0>.
- Environmental Defence and the Ontario Public Health Association, 2020. Clearing the Air: How electric vehicles and cleaner trucks can reduce pollution, improve health and save lives in the Greater Toronto and Hamilton Area, Stakeholder report, June.
- Gladstein, Neandross, & Associates, 2020. The state of sustainable fleet 2020. August, www.StateofSustainableFleets.com.
- House, M.L., Wright, D.J., 2019. Using the health benefits of electric vehicles to justify charging infrastructure incentives. *Int. J. Electric Hybrid Vehicles* 11 (2), 85–105.
- Jean, D., De Puy Kamp, M., 2018. Deadly convenience: Keyless car and their carbon monoxide toll. *The New York Times*, May 13.
- Langbroek, J.H., Franklin, J.P., Susilo, Y.O., 2017. Electric vehicle users and their travel patterns in Greater Stockholm. *Transport. Res. Part D: Transport Environ.* 52, 98–111.
- Le, T., Wang, Y., Lui, L., Yang, J., Yung, Y.L., Li, G., Seinfeld, J.H., 2020. Unexpected air pollution with marked emission reductions during the COVID-19 outbreak in China. *Science* 369 (6504), 702–706.
- Liu, Y., Lu, S., Yan, X., Gao, S., Cui, X., Cui, Z., 2020. Life cycle assessment of petroleum refining process: A case study in China. *J. Clean. Prod.* 256, 120422.
- Malmgren, I., 2016. Quantifying the societal benefits of electric vehicles. *World Electric Vehicle J.* 8, 996–1007.
- Nealer, R., Reichmuth, D., Anair, D., 2015. Cleaner cars from cradle to grave: How electric cars beat gasoline cars on lifetime global warming emissions. *Union of Concerned Scientist*, November.
- Nopmongkol, U., Grant, J., Knipping, E., Alexander, M., Schurhoff, R., Young, D., Jung, J., Shah, T., Yarwood, G., 2017. Air quality impacts of electrifying vehicles and equipment across the United States. *Environ. Sci. Technol.* 51 (5), 2830–2837.
- Plötz, P., Schneider, U., Globisch, J., Dütschke, E., 2014. Who will buy electric vehicles? Identifying early adopters in Germany. *Transport. Res. Part A: Policy Pract.* 67, 96–109.
- Rogelj, J., den Elzen, M., Höhne, N., Fransen, T., Fekete, H., Winkler, H., Schaeffer, R., Sha, F., Riahi, K., Meinshausen, M., 2016. Paris Agreement climate proposals need a boost to keep warming well below 2°C. *Nature* 534 (7609), 631.
- Spurlock, C.A., Sears, J., Wong-Parodi, G., Walker, V., Jin, L., Taylor, M., Duvall, A., Gopal, A., Todd, A., 2019. Describing the users: Understanding adoption of and interest in shared, electrified, and automated transportation in the San Francisco Bay Area. *Transport. Res. Part D: Transport Environ.*.
- Tessum, C.W., Hill, J.D., Marshall, J.D., 2014. Life cycle air quality impacts of conventional and alternative light-duty transportation in the United States. *Proc. Natl. Acad. Sci.* 111 (52), 18490–18495.
- U.S. Department of Energy, 2020a. Reducing pollution with electric vehicles. https://afdc.energy.gov/vehicles/electric_emissions.html.
- U.S. Department of Energy, 2020b. Emissions from hybrid and plug-in electric vehicles. https://afdc.energy.gov/vehicles/electric_emissions.html.
- U.S. Environmental Protection Agency, 2020. All-electric vehicles. <https://www.fueleconomy.gov/feg/evtech.shtml>.
- Weis, A., Jaramillo, P., Michalek, J., 2016. Consequential life cycle air emissions externalities for plug-in electric vehicles in the PJM interconnection. *Environ. Res. Lett.* 11 (2), 024009.

Further Reading

- Buekers, J., Holderbeke, M.V., Bierkens, J., Panis, L.I., 2014. Health and environmental benefits related to electric vehicle introduction in EU countries. *Transport. Res. Part D: Transport Environ.* 33, 26–38.
- International Energy Agency, 2019. Global EV outlook 2019. May.
- Tanvir, S., Hao, P., Boriboonsomsin, K., 2020. Emerging transportation technologies and implications for traffic-related emissions, air pollution exposure, and health. *Traffic-Related Air Pollution*, ISBN 978-0-12-818122-5, Elsevier, H. Khreis, M. J. Nieuwenhuijsen, J. Zietsman, and T. Ramani (Eds.), 511–530.

Shared Mobility Opportunities and Their Computational Challenges for Improving Health-Related Quality of Life

Cristiano Martins Monteiro^{*}, Cláudia Aparecida Soares Machado[†], Adelaide Cassia Nardocci[‡], Fernando Tobal Berssaneti[†], José Alberto Quintanilha[§], Clodoveu Augusto Davis Jr.^{*}, ^{*}Department of Computer Science – UFMG, 2 Reitor Píres Albuquerque St, Belo Horizonte, MG, Brazil; [†]Department of Transportation Engineering – USP, 83 trav.2 Prof. Almeida Prado Ave, São Paulo, SP, Brazil; [‡]School of Public Health – USP, 715 Dr. Arnaldo Ave, São Paulo, SP, Brazil; [§]Institute of Energy and Environment – USP, 1289 Prof. Luciano Gualberto Ave, São Paulo, SP, Brazil

© 2021 Elsevier Ltd. All rights reserved.

Introduction	376
Traffic-Related Pollution and Noise	377
Traffic-Related Air Pollution	377
Traffic-Related Noise	378
Active Transport	378
Electric Vehicles	378
Inclusive Transport	379
Shared Mobility on Emergency Evacuations	379
Mobility for People With Disabilities	380
Gender and Rural Barriers of Mobility	380
Mobility for Older People During COVID-19	380
Computational Challenges	381
Final Remarks	381
Acknowledgment	381
References	381

Introduction

The shared economy concept has expanded through different sectors and business models, bringing innovative solutions that can reshape people's and access to services (Cohen and Kietzmann, 2014; Cherry and Pidgeon, 2018). Amid shared economy models, the shared mobility modalities have grown in urban areas by providing more sustainable and affordable transportation services (Cohen and Kietzmann, 2014; Machado et al., 2018). In addition, shared mobility services bring economic opportunities by creating new space for profit and employment (Cherry and Pidgeon, 2018).

Although shared mobility has been extensively studied in recent years (Cohen and Kietzmann, 2014; Machado et al., 2018; McKenzie, 2020), the intersection among shared mobility and health subjects was not widely researched yet. This article discusses how shared mobility can help reduce respiratory and cardiovascular diseases in regions with dense traffic, and presents benefits and challenges of providing shared mobility services for vulnerable groups. The studied vulnerable groups were older people in rural areas far from healthcare services, people with physical or developmental disabilities, people living close to wildfires and needing to be evacuated, and low-income people in the status of transport-related social exclusion.

In order to illustrate the captivating multidisciplinary nature of this work, a word cloud was generated to visualize general subjects among the referenced works. Word cloud is a visual representation of words from a text, enhancing the most frequent words (Heimerl et al., 2014). In this case, the word cloud was built using all abstracts from the referenced works in the sections "Traffic-Related Pollution and Noise" and "Inclusive Transport." Referenced reports which do not have an abstract were represented by their first section inside the "Executive Summary" chapter. Referenced short papers with no abstract nor an equivalent section were represented by their first paragraph. Referenced abstracts or first section from "Executive Summary" chapter written in Portuguese or in British English were translated to American English. It was necessary to avoid replicas of the same word. Word replications varying only upper or lower case, and differing only in plural or singular were also eliminated. After these corrections, the word cloud was composed using 13,622 words, of which 2,798 words are unique. Fig. 1 shows the word cloud generated using the website "wordclouds.com" with default settings.

The words "air," "noise," "pollution," "exposure," "traffic" probably appear more often because they together summarize the causes of health issues caused by fossil-fueled vehicles, presented in the section "Traffic-Related Pollution and Noise." The words "access" and "disabilities" seem to represent the section "Inclusive Transport." Words related to the section "Traffic-Related Pollution and Noise" appeared more often than words related to the section "Inclusive Transport," probably because most references about exposure of traffic-related noise also evaluate the exposure of traffic-related air pollution, since both factors can occur together and be correlated. The topics about "Inclusive Transport" seems to be more extended through different subjects, therefore words are repeated less. Other smaller words on the word cloud evidence the particularities from the papers which compose the references of



Figure 1 Word cloud of referenced works.

this work. The following section presents the diseases caused by traffic-related air pollution and noise and presents environmentally friendly approaches to reduce those diseases. The subsequent sections, respectively, presents the transport-related social exclusion issues and discusses the computational challenges for the shared mobility opportunities.

Traffic-Related Pollution and Noise

This section is divided in four subsections: "Traffic-Related Air Pollution," "Traffic-Related Noise," "Active Transport," and "Electric Vehicles." The following subsection discusses the diseases caused by emissions of pollutants on air.

Traffic-Related Air Pollution

Amid the sustainable development goals addressed by the United Nations¹ there are the goals “Make cities inclusive, safe, resilient and sustainable” and “Take urgent action to combat climate change and its impacts.” Both goals aim to improve air quality in the city by diminishing emissions of carbon dioxide (CO₂) in order to reduce air pollution.

Air pollution is often related to adverse health outcomes in the population (Giles-Corti et al., 2016). These adverse outcomes include impaired lung development on prenatal child (Korten et al., 2017), childhood asthma (Gasana et al., 2012; (Khreis et al., 2019a, b; Alotaibi et al., 2019), childhood wheeze (Gasana et al., 2012), asthma exacerbation (Yang and Omaye, 2009), impaired lung function (Samet and Krewski, 2007), cardiovascular disease and mortality (Rajagopalan et al., 2018), and higher prevalence and risk of chronic kidney disease among the elderly population (Chen et al., 2018).

Furthermore, air pollutants from traffic and other combustion sources were associated with increased risk of infection by influenza (Croft et al., 2019). Recent works also associated air pollution with the vast diffusion of COVID-19 in the North Italy (Coccia, 2020; Conticini et al., 2020), and Wuhan, China (Yao et al., 2020). A prolonged exposure to air pollutants can induce a chronic respiratory system inflammation, even in young and healthy people, and weaken the subject's immune system. That may explain the higher prevalence and lethality of COVID-19, together with a higher average age of inhabitants in the North Italy region (Conticini et al., 2020).

The International Agency for Research on Cancer classified diesel engine exhaust as “Group 1 carcinogen,” known to be carcinogenic to humans². This classification was given after sufficient evidence that being exposed to diesel engine exhaust increases the risk of lung cancer (IARC, 2012). However, heavy-duty vehicles such as tractors, trucks, and buses usually are powered by diesel due to its relatively low-operational cost (Hannach et al., 2019). Furthermore, other fossil-fueled powered vehicles also emit air pollutants that are harmful to urban inhabitants and commuters (Mukaria et al., 2017; Anowar et al., 2017; CETESB, 2019; Nogueira et al., 2019).

Different urban environments have different amounts of pollutant particulate matter on air, making some places more harmful to the peoples' health than other places (Cheng et al., 2012; Targino et al., 2018; Nogueira et al., 2020). In Taipei, Taiwan, a study with measurements collected in July 2010 showed that the amount of fine particulate matter on air in the bus platform of the Taipei

¹ <https://www.un.org/sustainabledevelopment/sustainable-development-goals/>

¹ <https://www.cancer.org/cancer/cancer-causes/general-info/known-and-probable-human-carcinogens.html>

Bus Station was 2.0 times greater than the amount found on the waiting room, and 2.8 times greater than in the ambient monitoring station. Also, the station ticket inspectors were exposed from 1.6 to 1.8 times more fine particulate matter than in the Taipei outdoor atmosphere (Cheng et al., 2012).

In the city of Londrina, Brazil, a study with data collected from October through December 2014 showed that the concentration of black carbon inside buses was up to six times greater than on the outside of buses, for commuters walking or bicycling (Targino et al., 2018). Another study using measurements collected from May through December 2016 in the Metropolitan Area of São Paulo, Brazil, showed that commuters who wait for the bus in bus terminals on weekdays were exposed to 128% more coarse particulates than commuters who traveled outside the bus terminals. Besides, the concentration of black carbon on bus terminals was 3.3 times greater than in background sites (Nogueira et al., 2020).

Traffic-Related Noise

Traffic-related noise is known to be annoying and, therefore, it is also a decision factor while choosing where to live (Wardman and Bristow, 2004; Arsenio et al., 2006) since spatial characteristics such as traffic volume and length of nearby arterial roads can better explain the difference of noises in a city than temporal characteristics such as noise difference between workdays and weekends (Zuo et al., 2014). Furthermore, resting in a bedroom with a window facing a quieter traffic environment reduces the risk of sleeping problems (Bodin et al., 2015).

Besides the annoyance, traffic-related noise can bring psychological effects to people (Zuo et al., 2014). Those effects include sleep disturbance (Bodin et al., 2015), stress reaction (Dratva et al., 2010), hypertension (Fuks et al., 2017), cardiovascular events as consequence of physiological disorders (Recio et al., 2016), and diabetes (Clark et al., 2017). Also, traffic-related noise can increase the risk of myocardial infarction (Grazuleviciene et al., 2004; Babisch et al., 2005) and stroke (Münzel et al., 2018).

Children can also suffer the effects of traffic-related noise (Skrzypek et al., 2017; Foraster et al., 2019). A research based on 2715 children aged from 7 to 10 years, studying in 39 socioeconomically similar schools from Barcelona, Spain, showed that exposure to traffic-related noise outdoors was associated with smaller working memory and with greater inattentiveness, indicating that children exposed to traffic-related noise may have a slower cognitive development (Foraster et al., 2019). Another study with about 5136 children aged from 7 to 14 years old from Bytom, Poland, also found association between traffic-related noise and sleep disturbance for children exposed to the noise. The authors also found association between traffic-related noise and attention disorders (Skrzypek et al., 2017).

Air pollution and noise issues caused by fossil-fueled powered vehicles can be avoided during the city planning by setting homes, schools, facilities and other public areas far from dense traffic roads, and setting cycle lanes distanced from motor vehicle lanes (Giles-Corti et al., 2016). For the already built and working cities, common approaches consist in replacing fossil-fueled powered vehicles by active transport or electric vehicles. The following subsections present these approaches.

Active Transport

An approach would consist in shifting mobility from motor vehicles to active transport means, such as walking and bicycling (Nieuwenhuijsen and Khreis, 2016). Car-free and active mobility policies have grown through the recent years in order to promote more environmentally friendly and healthy cities. By doing so, it is possible to reduce road traffic fatalities (Otero et al., 2018), mitigate diseases with traffic-related air pollution or noise causes, and incentivize physical activities reducing the risk of diseases related to sedentarism (Nazelle et al., 2011; Nieuwenhuijsen and Khreis, 2016; Vich et al., 2019).

Practicing active mobility reduces the risks of cardiovascular diseases, diabetes, breast cancer (in women), colon cancer and dementia (Rojas-Rueda et al., 2013). Active mobility can also be based on the shared economy concept, creating services such as bikesharing. Bikesharing is an active shared mobility service which focuses on renting bikes on an as-needed basis (Machado et al., 2018; Cohen and Shaheen, 2018). Bikesharing can reduce congestion, bridge first-and-last-mile gaps in the transportation network and motivate multi-modal trips (Cohen and Shaheen, 2018).

Environmental factors can also motivate people to practice active mobility. A study with 127 participants sourcing 772 days of walking and location data in Barcelona, Spain, identified that while daily walking time was not statistically significant associated with participant's gender, employment status, or day type (workday or weekend), the participants' daily walking time was positively associated with existence of urban trees and large-scale open spaces such as beaches or large parks (Vich et al., 2019).

Regarding cyclists from places without car-free policies or lacking environment friendly spaces, their exposure to air pollution can be reduced by making healthier and preferred alternative routes. Although these routes may be longer, a reasonable trade-off between distance and cleaner air can be found by avoiding areas with concentration of pollutant particulate matters, for example (Anowar et al., 2017). Other approaches for improving health-related quality of life focus on replacing the fossil-fueled powered vehicles by other less pollutant ones. The following subsection presents the benefits of using electric vehicles.

Electric Vehicles

A direct benefit of replacing fossil-fueled powered vehicles by electric ones relies on reducing the transport-related air pollution. As an example, Nogueira et al. (2020) showed that renovating the diesel bus fleet of the Metropolitan Area of São Paulo by replacing the

buses with standard Euro II engines by buses with standard Euro V engines, and incorporating electric buses, would reduce CO₂ and nitrogen oxides (NO) emissions by up to 40%.

Incorporating electric vehicles on the cities is a promising way of reducing the carbon emissions on air (Ji et al., 2012; Nogueira et al., 2020). Even in regions where the electricity production consumed by the electric vehicles was generated from fossil-fuels, that production can happen far from urban centers, diminishing the inhabitants' health diseases caused by air pollution. Besides, there are electric vehicles planned for only one passenger, such as e-scooters. The electricity consumption for these vehicles is lower than the consumption of an electric car, avoiding the waste of produced electricity whether there is only one passenger per vehicle (Ji et al., 2012). Furthermore, electric vehicles can be offered as a mobility service following the concept of shared mobility. With shared mobility, the number of idle vehicles through the day can be reduced, not even needing to produce unnecessary vehicles. By doing so, it is possible to make mobility more sustainable (Machado et al., 2018).

An online survey disseminated in different Spanish cities between April and June 2018 identified the profile of e-scooter short-term rental clients. Analysis showed patterns of sociodemographic data, such as lower age and university degree, on more frequent clients (Aguilera-García et al., 2020). New lifestyle trends are also noticed in India, with more environmental concerns related to the rapid urbanization and increase in the middle-class population (Roy et al., 2018). Therefore, shared on-demand transport services together with sustainable awareness and low-cost necessities are shaping the emergent transportation modes in the cities (Machado et al., 2018; Roy et al., 2018; Aguilera-García et al. 2020). The following section presents the drawbacks caused by the lack of such services.

Inclusive Transport

According to the Social Exclusion Unit (2003) from the United Kingdom, transport problems can hamper people's access to services and opportunities, reinforcing their status of social exclusion. The lack of transportation means can also bring difficulties to build social networks and reduce family recreational day trips, being harmful to well-being and mental health. Few mobility options can also restrict people to walk to locally accessible grocery stores, being subject to higher grocery prices and having fewer healthy food options. Besides, the lack of transportation means can hamper opportunities of employment, education, and training that could increase income (Mackett and Thoreau, 2015).

The authors from Social Exclusion Unit (2003) defined transport accessibility as people being able to get to key services at a reasonable cost, in reasonable time and with reasonable ease. Also, accessible transport should be reliable, passengers should feel safe using it and people should be financially and physically able to access the transport. Low-accessibility transportation can cause problems for work, learning, healthcare, shopping, social, and sporting activities.

Transport-related social exclusion is often more properly explained by non-transport factors and issues of power, accessibility, and choice. For example, while people with money, time and good health to drive can choose to enjoy an attractive rural environment, inhabitants of these rural areas (lacking services such as healthcare and shopping) can be "forced" to own and maintain a vehicle in order to have access to essential services. Therefore, being "forced" to have a vehicle is a factor of experiencing transport-related social exclusion, differently from choosing to own a vehicle while having other reasonable options of mobility (Pooley, 2016).

Shared mobility modes represent a promising set of solutions for vulnerable groups that are not properly served by public transport (due to waiting time, trip duration or distance to get to a bus stop) neither can comfortably afford a private vehicle (Viegas and Martinez, 2017). The shared mobility concept can also be applied to evacuation situations (Li et al., 2018; Wong and Shaheen, 2019).

Shared Mobility on Emergency Evacuations

An online survey distributed to people impacted by the wildfires of October 2017 in Northern California, December 2017 in Southern California, and August 2018 in Carr evaluated the respondents' willingness to share transportation and shelter in future evacuations. Results showed that respondents are more likely to share transportation resources than housing resources due to the lesser level of commitment with the individual being helped. Thus, shared mobility through community aid can play a key role on saving lives during emergencies (Wong and Shaheen, 2019).

However, among those three wildfires, the percentage of respondents willing to share personal transportation while evacuating varied from 59% to 72%, and the percentage willing to share personal transportation before evacuation varied from 37% to 62%. Fewer respondents willing to offer transportation before evacuation is explained by the possible short time for preparation between the risk recognition of wildfire and the beginning of evacuation. Also, more than 30% of respondents from each evaluated wildfire stated that they would deviate at maximum ten minutes from evacuation route to pick-up a passenger (Wong and Shaheen, 2019). Even though ridesourcing services such as Uber and Lyft could play a major role during evacuations, respondents from vulnerable groups (older adults, individuals with disabilities and low-income individuals) expressed strong concern about ridesourcing availability and reliability with enough accommodation for disabled people or at a affordable cost (due to possible dynamic pricing) mainly in areas near the fire line (Li et al., 2018; Wong and Shaheen, 2019).

Older people from the wildfire areas said not to trust in the ridesourcing drivers and were highly concerned about their personal security, particularly while sharing rides. They also noted that drivers may need to evacuate their own families or be unwilling to help

in the evacuation process. People with disabilities were concerned about ridesourcing vehicles not being wheelchair accessible nor have enough space to store the wheelchair. Also, it is possible that drivers cancel trip requests after realizing the passenger's disability. Besides, ridesourcing drivers may not know how to assist people with disabilities, nor be trained to save people in a disaster. A proposed solution consisted in maintaining a neighborhood volunteer network similar to carpooling, creating partnerships with paratransit services, training drivers to identify and assist people with disabilities, focusing on drivers who live near but not in the areas which can be impacted by a wildfire, receiving support from government, and ensuring that mobility costs will not surge during the evacuation (Wong and Shaheen, 2019).

Mobility for People With Disabilities

Even without disasters and with a reasonable number of mobility options, people with disabilities can also be transport-related and socially excluded. Besides individuals with physical disabilities, people with developmental disabilities (including individuals diagnosed with autism spectrum disorder, cerebral palsy, intellectual disability, or traumatic brain injury) can also face barriers while using transportation means (Stock et al., 2011; Lubin and Deka, 2012; Wasfi et al., 2017). These barriers include lack of awareness of transport options (Wasfi et al., 2017; Stock et al., 2019), having to understand complex schedules of public transport or potential difficulty of transfers (Wasfi et al., 2017), need of advance scheduling for using demand-responsive paratransit services and difficulty on understanding onboard announcements (Lubin and Deka, 2012), high expenses in ridesourcing since people with developmental disabilities are less likely to have a driver's license or own a private vehicle, and issues of personal safety and security such as encountering with strangers, getting lost or passing by busy street crossings (Stock et al., 2011).

Proposed ways to mitigate those issues consists in training people with developmental disabilities to use specific technologies to simplify and guide them to get to their destinations. As an example of these technologies, there are navigation apps explaining with pictures and audios the next steps of using public transport, such as when to signal a stop and drop-off the bus (Stock et al., 2011; Stock et al., 2019). In the future, autonomous vehicles can enable less expensive ridesourcing services that could also reduce barriers of transport-related exclusion for low income individuals with disabilities (Rodier, 2020). Besides low income individuals with disabilities, autonomous vehicles can also reduce the gender gap in transport-related social exclusion (Singh, 2019).

Gender and Rural Barriers of Mobility

Transport-related social exclusion may have different impacts according with gender, since women are usually more concerned with their personal safety and security, and they are less likely to have a driver's license (Mackett and Thoreau, 2015; Singh, 2019). Regarding older people in rural areas, a study performed in the Republic of Ireland and Northern Ireland showed that although older men usually have a driver's license, their lifestyles make them more car dependent than old women without a driver's license and without a husband. Most of the men participating in that study wanted to be more independent in their trips, were not used to sharing transport space, and saw community transport as "feminised." The "feminised" stereotype is due to rural transport services being predominantly used by women to get to services more attractive to them, such as shopping and visiting clubs and social groups. The study also showed that since health services are usually centralized in order to provide better medical facilities and services, some rural areas have no transport network to access these services. In this case, patients are car-dependent to get to, at least, the first transfer to reach the medical facility. This problem affects both older males and females. Even when someone is able to drive (has a driver's license, a car, and is healthy to drive), his spouse or relative may not be able to drive and may need help, since rural areas commonly do not have a cheap taxi service available (Ahern and Hine, 2012).

Mobility for Older People During COVID-19

Regarding regions with public transport or shared mobility options, an important point about clients' health rely on cleaning the shared vehicles, mainly during epidemic scenarios such as from COVID-19. Since shared vehicles are expected to be reused by other clients (Machado et al., 2018) and viable COVID-19 virus were detected in plastic or stainless steel surfaces even after 72 hours of exposure (van Doremalen et al., 2020), the contamination by COVID-19 can spread through clients of shared mobility even if they used the same vehicle on different days.

Being contaminated by COVID-19 is dangerous mainly for older people. A study from Mainland China with data from 44,672 contaminated people as of February 11, 2020 showed that although 13,909 people with COVID-19 were at least 60 years old (equivalent to 31%), most of deaths (829 from 1023 deaths, equivalent to 81%) occurred to people at least 60 years old (Novel Coronavirus Pneumonia Emergency Response Epidemiology Team, 2020). About Italy, another study with data as of March 15, 2020 showed 22,512 confirmed cases and 1,625 deaths, of which 1,567 (equivalent to 96%) are from people who are at least 60 years old (Livingston and Bucher, 2020). Given the higher risks related to older people being contaminated by COVID-19, together with health complications of people with chronic diseases (such as respiratory system inflammation mentioned earlier in this work), it is expected that individuals in the risk group avoid using carpooling options or public transport for health reasons, and therefore can establish or reinforce their status as transport-related socially excluded people.

Computational Challenges

The availability, costs and company profits of different mobility services under diverse scenarios can be simulated using software (Mourad et al., 2019; Lage et al., 2019; Monteiro et al., 2019a). However, although the way such software is built can make them run faster and be more computationally efficient (Mourad et al., 2019; Monteiro et al., 2019b), some of those mobility problems become hugely complex as wider and more realistic scenarios are evaluated. In these cases, an impracticable time (in the order of years) would be required to find the best solution among all the possible ones (Garey and Johnson, 1979; Cormen et al., 2009).

Usually, extensive and realistic problems are dealt with using methods which do not guarantee that the solution achieved is the best among all possible. However, a computer running these methods can quickly find a good solution, and that found solution may often be reasonable for the problem at hand (Mourad et al., 2019). Nevertheless, as the problem becomes more complex, it becomes harder to find a good solution in a viable time (Ritzinger et al., 2016). This can become a challenge for companies of shared-mobility services, which may daily depend on such simulation or optimization software, since their prices are usually constrained by the prices of competitors. As an example, carsharing rentals should be lower than ridesourcing, otherwise the clients would prefer to simply call an Uber instead of having to walk to get to an available vehicle and also having to drive to their destination (Monteiro et al., 2019a).

Similar situations emerge while adopting the sustainable and inclusive services mentioned in the chapters “Traffic-Related Pollution and Noise” and “Inclusive Transport.” Planning the location of electric vehicle chargers is within the class of the hardest problems of Computer Science (Lam et al., 2014). This class is known as NP-Hard (Garey and Johnson, 1979; Cormen et al., 2009), which also contains the problem of optimizing the routes carried by a vehicle relocating shared bikes in order to bring them to areas of demand (Erdoğan et al., 2015). In addition, the NP-Hard class contains the problem of optimizing multiple workers to relocate (or consequently give maintenance and clean) the carsharing vehicles (Albinski, 2015), and it is also known that the carpooling problem proposed as a solution for community evacuation belongs to the NP-Hard class (Buchholz, 1997; Baugh et al., 1998). Regarding community evacuation, carpooling would have an additional hindrance, related to the probably short response time available to make decisions (Mourad et al., 2019) to save people in more isolated areas.

Other computational challenges can emerge in tasks of acquisition, processing and analysis of data. The lack of quality and enough observations in datasets about Vulnerable Road Users (VRU), particularly from developing countries, hamper the use and trust of standard spatial regression models in VRU analysis (Machado et al., 2015). Other problems happen while analyzing data from traffic accidents. Although public datasets of traffic accidents can be available about the same area, each of these datasets (official and unofficial) can present different traffic accident events, requiring that data integration techniques be applied to merge those data before analyzing them (Santos et al., 2017). In addition, low precision of georeferenced data about moving objects, such as vehicle GPS location history, can require preprocessing tasks in order to avoid analyzing abnormal locations recorded due to GPS errors (Monteiro et al., 2017).

Final Remarks

Besides the difficulty of finding good solutions and dealing with quality and integration of data, some services have additional challenges to be properly deployed by for-profit companies due to the involved costs. For example, organizing the staff to clean the vehicles between rentals in order to reduce the risks of contamination by diseases such as COVID-19 would require higher costs than only relocating vehicles according to demand. Also, offering accessible transport modes in rural areas with low demand requires more vehicle and operational costs than only serving the frequent trips. However, for these examples of contamination and rural areas, shared mobility modes can be more suitable than conventional public transport since buses and subway can easily spread contamination due to the people's agglomeration, and buses and subways are restricted to fixed routes and schedules. Furthermore, the additional cost for shared mobility could be overcome by community aid or support from the government.

Concluding, shared mobility acts as a promising source of trips, benefiting people in vulnerable groups, and enabling them to access essential services, such as healthcare. Shared mobility services can also reduce the number of vehicles in the cities, making them more sustainable, and healthy for all inhabitants. Fewer vehicles transiting, mainly regarding the fossil-fueled powered ones, can make the air cleaner and the environment less noisy, improving urban quality of life.

Acknowledgment

The authors acknowledge the support from CNPq and CAPES, Brazilian agencies in charge of fostering research and development.

References

- Aguilera-García, Á., Gomez, J., Sobrino, N., 2020. Exploring the adoption of moped scooter-sharing systems in Spanish urban areas. *Cities* 96, 1–13, 102424.
- Ahern, A., Hine, J., 2012. Rural transport—valuing the mobility of older people. *Res. Transp. Econ.* 34 (1), 27–34.
- Albinski, S.J., 2015. A branch-and-cut method for the vehicle relocation problem in the one-way car-sharing. Master's Thesis. KTH Royal Institute of Technology. Stockholm, Sweden. Retrieved from: <https://www.diva-portal.org/smash/get/diva2:812003/FULLTEXT01.pdf>.

- Alotaibi, R., Bechle, M., Marshall, J.D., Ramani, T., Zietsman, J., Nieuwenhuijsen, M.J., Khreis, H., 2019. Traffic related air pollution and the burden of childhood asthma in the contiguous United States in 2000 and 2010. *Environ. Int.* 127, 858–867.
- Anowar, S., Eluru, N., Hatzopoulou, M., 2017. Quantifying the value of a clean ride: How far would you bicycle to avoid exposure to traffic-related air pollution? *Transp. Res. Part A: Policy Prac.* 105, 66–78.
- Arsenio, E., Bristow, A.L., Wardman, M., 2006. Stated choice valuations of traffic related noise. *Transp. Res. Part D: Transp. Environ.* 11 (1), 15–31.
- Baugh Jr., J.W., Kakivaya, G.K.R., Stone, J.R., 1998. Intractability of the dial-a-ride problem and a multiobjective solution using simulated annealing. *Eng. Optimiz.* 30 (2), 91–123.
- Bodin, T., Björk, J., Ardö, J., Albin, M., 2015. Annoyance, sleep and concentration problems due to combined traffic noise and the benefit of quiet side. *Int. J. Environ. Res. Public Health* 12 (2), 1612–1628.
- Babisch, W., Beule, B., Schust, M., Kersten, N., Ising, H., 2005. Traffic noise and risk of myocardial infarction. *Epidemiology* 16 (1), 33–40.
- Buchholz, F., 1997. The Carpool Problem. Technical report, Institute of Computer Science, University of Stuttgart.
- CETESB, 2019. Qualidade do ar no estado de São Paulo 2018. São Paulo: CETESB, 2019. Retrieved from: <https://cetesb.sp.gov.br/ar/publicacoes-relatorios/>.
- Chen, S.Y., Chu, D.C., Lee, J.H., Yang, Y.R., Chan, C.C., 2018. Traffic-related air pollution associated with chronic kidney disease among elderly residents in Taipei City. *Environ. Pollut.* 234, 838–845.
- Cherry, C.E., Pidgeon, N.F., 2018. Is sharing the solution? Exploring public acceptability of the sharing economy. *J. Clean. Prod.* 195, 939–948.
- Clark, C., Sbihi, H., Tamburic, L., Brauer, M., Frank, L.D., Davies, H.W., 2017. Association of long-term exposure to transportation noise and traffic-related air pollution with the incidence of diabetes: A prospective cohort study. *Environ. Health Perspect.* 125 (8), 1–10, 087025.
- Coccia, M., 2020. Diffusion of COVID-19 Outbreaks: The Interaction between Air Pollution-to-Human and Human-to-Human Transmission Dynamics in Hinterland Regions with Cold Weather and Low Average Wind Speed (April 3, 2020). Working Paper CocciaLab n. 48/2020, CNR - National Research Council of Italy. doi: <http://dx.doi.org/10.2139/ssrn.3567841>. Retrieved from: <https://ssrn.com/abstract=3567841>.
- Cohen, B., Kietzmann, J., 2014. Ride on! Mobility business models for the sharing economy. *Org. Environ.* 27 (3), 279–296.
- Cohen, A. and Shaheen, S., 2018. Planning for shared mobility Chicago: American Planning Association. doi: <http://dx.doi.org/10.7922/G2NV9GDD>. Retrieved from: <https://escholarship.org/uc/item/0dk3h89p>.
- Conticini, E., Frediani, B., Caro, D., 2020. Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy? *Environ. Pollut.* 261, 1–3, 114465.
- Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2009. Introduction to Algorithms, 3rd. ed.. The MIT Press, England.
- Cheng, Y.H., Chang, H.P., Yan, J.W., 2012. Temporal variations in airborne particulate matter levels at an indoor bus terminal and exposure implications for terminal workers. *Aerosol Air Qual. Res.* 12 (1), 30–38.
- Croft, D.P., Zhang, W., Lin, S., Thurston, S.W., Hopke, P.K., van Wijngaarden, E., Squizzato, S., Masiol, M., Utell, M.J. and Rich, D.Q., 2019. Associations between Source-Specific Particulate Matter and Respiratory Infections in New York State Adults. *Environmental science & technology*.
- van Doremalen, N., Bushmaker, T., Morris, D.H., Holbrook, M.G., Gamble, A., Williamson, B.N., Tamin, A., Harcourt, J.L., Thornburg, N.J., Gerber, S.I., Lloyd-Smith, J.O., 2020. Aerosol and surface stability of SARS-CoV-2 as compared with SARS-CoV-1. *N. Engl. J. Med.* 382 (16), 1564–1567.
- Dratva, J., Zemp, E., Dietrich, D.F., Bridevaux, P.O., Rochat, T., Schindler, C., Gerbase, M.W., 2010. Impact of road traffic noise annoyance on health-related quality of life: Results from a population-based study. *Qual. Life Res.* 19 (1), 37–46.
- Erdogan, G., Battarra, M., Calvo, R.W., 2015. An exact algorithm for the static rebalancing problem arising in bicycle sharing systems. *Eur. J. Operat. Res.* 245 (3), 667–679.
- Foraster, M., Esnaola, M., Garca-Esteban, R., Lpez-Vicente, M., Rivas, I. and Sunyer, J., 2019, September. Exposure to road traffic noise and cognitive development in primary schoolchildren. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, 259 (5), pp. 4282–4285. Institute of Noise Control Engineering.
- Fuks, K.B., Weinmayr, G., Basagaña, X., Gruzdeva, O., Hampel, R., Oftedal, B., Sørensen, M., Wolf, K., Aamodt, G., Aasvang, G.M., Aguilera, I., 2017. Long-term exposure to ambient air pollution and traffic noise and incident hypertension in seven cohorts of the European study of cohorts for air pollution effects (ESCAPE). *Eur. Heart J.* 38 (13), 983–990.
- Garey, M.R., Johnson, D.S., 1979. Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., USA.
- Gasana, J., Dillikar, D., Mendy, A., Forno, E., Vieira, E.R., 2012. Motor vehicle air pollution and asthma in children: A meta-analysis. *Environ. Res.* 117, 36–45.
- Grzulewiciene, R., Lekaviciute, J., Mozgeris, G., Merkevičius, S., Deikus, J., 2004. Traffic noise emissions and myocardial infarction risk. *Polish J. Environ. Stud.* 13. (6).
- Giles-Corti, B., Vernez-Moudon, A., Reis, R., Turrell, G., Dannenberg, A.L., Badland, H., Foster, S., Lowe, M., Sallis, J.F., Stevenson, M., Owen, N., 2016. City planning and population health: A global challenge. *The Lancet* 388 (10062), 2912–2924.
- Hannach El, M.E., Ahmadi, P., Guzman, L., Pickup, S., Kjeang, E., 2019. Life cycle assessment of hydrogen and diesel dual-fuel class 8 heavy duty trucks. *Int. J. Hydro. Ener.* 44 (16), 8575–8584.
- Heimerl, F., Lohmann, S., Lange, S. and Ertl, T., 2014. Word cloud explorer: Text analytics based on word clouds. In 2014 47th Hawaii International Conference on System Sciences, IEEE, pp. 1833–1842.
- IARC, 2012. IARC: Diesel Exhaust Carcinogenic - IARC. Lyon, France: International Agency for Research in Cancer. Retrieved from: https://www.iarc.fr/wp-content/uploads/2018/07/pr213_E.pdf.
- Ji, S., Cherry, C.R., J. Bechle, M.J., Wu, Y., Marshall, J.D., 2012. Electric vehicles in China: Emissions and health impacts. *Environ. Sci. Technol.* 46 (4), 2018–2024.
- Khreis, H., Cirach, M., Mueller, N., de Hoogh, K., Hoek, G., Nieuwenhuijsen, M.J., Rojas-Rueda, D., 2019b. Outdoor air pollution and the burden of childhood asthma across Europe. *Eur. Resp. J.* 54 (4), 1–36.
- Khreis, H., Ramani, T., de Hoogh, K., Mueller, N., Rojas-Rueda, D., Zietsman, J., Nieuwenhuijsen, M.J., 2019a. Traffic-related air pollution and the local burden of childhood asthma in Bradford UK. *Int. J. Transp. Sci. Technol.* 8 (2), 116–128.
- Korten, I., Ramsey, K., Latzin, P., 2017. Air pollution during pregnancy and lung development in the child. *Paed. Resp. Rev.* 21, 38–46.
- Lage, M.O., Machado, C.A.S., Monteiro, C.M., Berssaneti, F.T., Quintanilha, J.A., 2019. Location suitable for the implementation of carsharing in the city of São Paulo. *Proc. Manufact.* 39, 1962–1967.
- Lam, A.Y., Leung, Y.W., Chu, X., 2014. Electric vehicle charging station placement: Formulation, complexity, and solutions. *IEEE Trans. Smart Grid* 5 (6), 2846–2856.
- Li, M., Xu, J., Liu, X., Sun, C., Duan, Z., 2018. Use of shared-mobility services to accomplish emergency evacuation in urban areas via reduction in intermediate trips—case study in Xi'an China. *Sustainability* 10 (12), 4862.
- Livingston, E., Bucher, K., 2020. Coronavirus disease 2019 (COVID-19) in Italy. *JAMA* 323 (14), 1335.
- Lubin, A., Deka, D., 2012. Role of public transportation as job access mode: Lessons from survey of people with disabilities in New Jersey. *Transp. Res. Rec.* 2277 (1), 90–97.
- Machado, C.A.S., Giannotti, M.A., Neto, F.C., Tripodi, A., Persia, L., Quintanilha, J.A., 2015. Characterization of black spot zones for vulnerable road users in São Paulo (Brazil) and Rome (Italy). *ISPRS Int. J. Geo-Info.* 4 (2), 858–882.
- Machado, C.A.S., Salles Hue de, N.P.M., Berssaneti, F.T., Quintanilha, J.A., 2018. An overview of shared mobility. *Sustainability* 10 (12), 4342.
- Mackett, R.L., Thoreau, R., 2015. Transport, social exclusion and health. *J. Transp. Health* 2 (4), 610–617.
- McKenzie, G., 2020. Urban mobility in the sharing economy: A spatiotemporal comparison of shared mobility services. *Comp. Environ. Urban Sys.* 79, 101418.
- Monteiro, C.M., Machado, C.A.S., de Oliveira Lage, M., Berssaneti, F.T., Davis Jr., C.A., Quintanilha, J.A., 2019a. Maximizing carsharing profits: An optimization model to support the carsharing planning. *Proc. Manufact.* 39, 1968–1976.
- Monteiro, C.M., Mateus, G.R., Davis Jr., C.A., 2019b. Computational performance of carsharing fleet-sizing optimization. In *GEOINFO* (2019), 111–122.
- Monteiro, C.M., da Silva, F.R. and Murta, C.D., 2017, July. Pré-processamento e Análise de Dados de Táxis. In *Anais do XLIV Seminário Integrado de Software e Hardware*. SBC.
- Mourad, A., Puchinger, J., Chu, C., 2019. A survey of models and algorithms for optimizing shared mobility. *Transp. Res. Part B: Method.* 123, 323–346.
- Mukaria, S.M., Wahome, R.G., Gatari, M., Thenya, T., Karatu, K., 2017. Particulate matter from motor vehicles in Nairobi road junctions Kenya. *J. Atmos. Pollut.* 5 (2), 62–68.

- Münzel, T., Schmidt, F.P., Steven, S., Herzog, J., Daiber, A., Sørensen, M., 2018. Environmental noise and the cardiovascular system. *J. Am. Coll. Cardiol.* 71 (6), 688–697.
- Nazelle, A., de Nieuwenhuijsen, M.J., Antó, J.M., Brauer, M., Briggs, D., Braun-Fahrlander, C., Cavill, N., Cooper, A.R., Desqueyroux, H., Fruin, S., Hoek, G., 2011. Improving health through policies that promote active travel: A review of evidence to support integrated health impact assessment. *Environ. Int.* 37 (4), 766–777.
- Nieuwenhuijsen, M.J., Khreis, H., 2016. Car free cities: Pathway to healthy urban living. *Environ. Int.* 94, 251–262.
- Nogueira, T., Dominutti, P.A., Vieira-Filho, M., Fornaro, A., Andrade, M.D.F., 2019. Evaluating atmospheric pollutants from urban buses under real-world conditions: Implications of the main public transport mode in São Paulo, Brazil. *Atmosphere* 10 (3), 108.
- Nogueira, T., Kumar, P., Nardocci, A., de Fatima Andrade, M., 2020. Public health implications of particulate matter inside bus terminals in Sao Paulo, Brazil. *Sci. Total Environ.* 711, 135064.
- Novel Coronavirus Pneumonia Emergency Response Epidemiology, Team., 2020. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020. *China CDC Weekly* 2 (8), 113–122.
- Otero, I., Nieuwenhuijsen, M.J., Rojas-Rueda, D., 2018. Health impacts of bike sharing systems in Europe. *Environ. Int.* 115, 387–394.
- Pooley, C., 2016. Mobility, transport and social inclusion: Lessons from history. *Soc. Inc.* 4 (3), 100–109.
- Rajagopalan, S., Al-Kindi, S.G., Brook, R.D., 2018. Air pollution and cardiovascular disease: JACC state-of-the-art review. *J. Am. Coll. Cardiol.* 72 (17), 2054–2070.
- Recio, A., Linares, C., Banegas, J.R., Díaz, J., 2016. Road traffic noise effects on cardiovascular, respiratory, and metabolic health: An integrative model of biological mechanisms. *Environ. Res.* 146, 359–370.
- Ritzinger, U., Puchinger, J., Hartl, R.F., 2016. A survey on dynamic and stochastic vehicle routing problems. *Int. J. Prod. Res.* 54 (1), 215–231.
- Rodier, C., 2020. The potential for autonomous vehicle technologies to address barriers to driving for individuals with autism. Technical Report: Mineta Transportation Institute; San José State University: San José, United States of America. doi: <https://doi.org/10.31979/mti.2020.1706>. Retrieved from: <https://transweb.sjsu.edu/sites/default/files/1706-Rodier-AV-Address-Barriers-Driving-Individuals-Autism.pdf>.
- Rojas-Rueda, D., Nazelle, A., de Teixidó, O., Nieuwenhuijsen, M.J., 2013. Health impact assessment of increasing public transport and cycling use in Barcelona: A morbidity and burden of disease approach. *Prev. Med.* 57 (5), 573–579.
- Roy, J., Chakravarty, D., Dasgupta, S., Chakraborty, D., Pal, S., Ghosh, D., 2018. Where is the hope? Blending modern urban lifestyle with cultural practices in India. *Curr. Opin. Environ. Sustain.* 31, 96–103.
- Samet, J., Krewski, D., 2007. Health effects associated with exposure to ambient air pollution. *J. Toxicol. Environ. Health Part A* 70 (3–4), 227–242.
- Santos, S.R., dos Davis Jr., C.A., Smarzo, R., 2017. Analyzing traffic accidents based on the integration of official and crowdsourced data. *J. Inform. Data Manag.* 8 (1), 67–82.
- Singh, Y.J., 2019. Is smart mobility also gender-smart? *J. Gender Stud.* 28 (1), 1–15.
- Social Exclusion Unit, S.E.U., 2003. Making the connections: final report on transport and social exclusion. Technical Report: Office of the Deputy Prime Minister: London: England. Retrieved from: https://www.ilo.org/wcmsp5/groups/public/-ed_emp/-emp_policy/-invest/documents/publication/wcms_asist_8210.pdf.
- Skrzypek, M., Kowalska, M., Czech, E.M., Niewiadomska, E., Zejda, J.E., 2017. Impact of road traffic noise on sleep disturbances and attention disorders amongst school children living in upper Silesian industrial zone Poland. *Int. J. Occup. Med. Environ. Health* 30 (3), 511.
- Stock, S.E., Davies, D.K., Herold, R.G., Wehmeyer, M.L., 2019. Technology to support transportation needs assessment, training, and pre-trip planning by people with intellectual disability. *Adv. Neurodev. Disord.* 3 (3), 319–324.
- Stock, S.E., Davies, D.K., Wehmeyer, M.L., Lachapelle, Y., 2011. Emerging new practices in technology to support independent community access for people with intellectual and cognitive disabilities. *NeuroRehabil.* 28 (3), 261–269.
- Targino, A.C., Rodrigues, M.V.C., Krecl, P., Cipoli, Y.A., Ribeiro, J.P.M., 2018. Commuter exposure to black carbon particles on diesel buses, on bicycles and on foot: A case study in a Brazilian city. *Environ. Sci. Pollut. Res.* 25 (2), 1132–1146.
- Vich, G., Marquet, O., Miralles-Guasch, C., 2019. Green streetscape and walking: Exploring active mobility patterns in dense and compact cities. *J. Transp. Health* 12, 50–59.
- Viegas, J. and Martinez, L., Transition to shared mobility: how large cities can deliver inclusive transport services; Corporate Partnership Board Report; International Transport Forum: Paris, France, 2017.
- Wardman, M., Bristow, A.L., 2004. Traffic related noise and air quality valuations: Evidence from stated preference residential choice models. *Transp. Res. Part D: Transp. Environ.* 9 (1), 1–27.
- Wasfi, R., Steinmetz-Wood, M., Levinson, D., 2017. Measuring the transportation needs of people with developmental disabilities: A means to social inclusion. *Disab. Health J.* 10 (2), 356–360.
- Wong, S. and Shaheen, S., 2019. Current State of the Sharing Economy and Evacuations: Lessons from California. Technical Report: Institute of Transportation Studies; University of California: Berkeley, United States of America. doi: <https://doi.org/10.7922/G2WW7FVK>. Retrieved from: <https://escholarship.org/content/qt16s8d37x/qt16s8d37x.pdf>.
- Yang, W., Omaye, S.T., 2009. Air pollutants, oxidative stress and human health. *Muta. Res./Genet. Toxicol. Environ. Mutagenesis* 674 (1–2), 45–54.
- Yao, Y., Pan, J., Liu, Z., Meng, X., Wang, W., Kan, H. and Wang, W., 2020. Ambient nitrogen dioxide pollution and spread ability of COVID-19 in Chinese cities. *medRxiv*. doi: <https://doi.org/10.1101/2020.03.31.20048595>. Retrieved from: <https://www.medrxiv.org/content/10.1101/2020.03.31.20048595v2.full.pdf>.
- Zuo, F., Li, Y., Johnson, S., Johnson, J., Varughese, S., Copes, R., Liu, F., Wu, H.J., Hou, R., Chen, H., 2014. Temporal and spatial variability of traffic-related noise in the City of Toronto Canada. *Sci. Total Environ.* 472, 1100–1107.

Bike Sharing and Health

David Rojas-Rueda^{*}, Mark J. Nieuwenhuijsen[†], ^{*}Department of Environmental and Radiological Health Sciences, Colorado State University, Fort Collins, CO, United States; [†]Barcelona Institute for Global Health, ISGlobal, Barcelona, Spain

© 2021 Elsevier Ltd. All rights reserved.

Introduction	384
Health Benefits of BSSs	386
E-Bikes and Bike Sharing	388
Helmets and Bike Sharing	390
BSS Users' Profiles	390
Conclusions	391
References	391
Further Reading	392

Introduction

Bike-sharing systems (BSSs) have been implemented in several cities around the world as policies to mitigate climate change, reduce traffic congestion, and promote physical activity and thereby health. Also, the cyclists do not produce air pollution or traffic noise and help toward the last mile trips. A BSS or bike-share scheme is a system in which bikes are made available for shared use to individuals on a very short-term basis. Often the time use is limited to allow multiple users per bike and rapid turnaround. BSS allow people to borrow a bike from one point and return it to a different point (**Fig. 1**). BSS has become very popular in cities across Europe, Asia and America, and in 2013 more than 500 BSS were implemented around the world (**Larsen, 2013**).

Interest in urban cycling is increasing and the number of bike share programs has grown rapidly over the last 5–10 years. Bike share programs have existed for almost 50 years, but the recent change in the technology used and interest in more active and liveable cities has a more widespread use of the programs. Wuhan and Hangzhou, China currently have the world's largest public bicycle share schemes, with 70,000 and 65,000 bikes, respectively (**China News, 2011**; **Meddin, 2011a**). In 2007, Paris launched Europe's largest scheme, with over 20,000 bicycles. New York City launched North America's largest bike share program, with 10,000 bicycles in 2013. **Shaheen et al. (2010)** summarize the benefits of bike sharing as flexible mobility, emission reductions, physical activity benefits, reduced congestion and fuel use, individual financial savings and support for multimodal transport connections. These factors have acted as a catalyst for the development of bike sharing globally.

The first bike share began in Europe in 1965 with fairly simple technology and the first large-scale bike-sharing program was launched in 1995, in Copenhagen as Bycyklen (City Bikes) with 1100 bikes (**Shaheen et al., 2010**). Currently the BSS in Paris called "Vélib", is the biggest in Europe with 23,600 bikes and 1800 stations; other BSS have also reached a considerable large size as London (12,000 bikes), Barcelona (6000), Lyon (4000) or Valencia, Seville, Milan or Brussels with more than 2000 bikes. In some countries like Spain, there has been a rapid increase in the number of BSS, almost doubling the number of systems implemented from 58 to 97 between 2008 and 2009 (**Oortwijn, 2015**). But China still has the largest system as mentioned before. Either way the presence of BSS can be found around the globe (**Fig. 2**).

Bike share programs can provide safe and convenient access to bicycles for short trips, such as running errands during lunch, and transit-work trips. Users do not need parking at their home and maintenance of the bikes is taken care of making it convenient for the users. Until recently, bike share programs worldwide had experienced limited success; in the last 5–10 years, innovations in technology have given rise to a new (third) generation of technology-driven bike share programs. These new bike share programs can dramatically increase the visibility of cycling and lower barriers to use by requiring only that the user have a desire to bike and a credit card or phone.

Bike share programs help increase cycling mode share, serve as a missing link in the public transit system, reduce a city's travel-related carbon footprint and provide additional "green" jobs related to system management and maintenance. They have had limited success though to get people out of the car and onto the bike; to a large extent new users are coming from the public transport system or previously walked. In the US, many cities are looking into bike share programs, but they have not yet been widely implemented. The US has a strong car dependency and limited infrastructure for cycling which makes it harder. Some recent systems failed and were abandoned. Poor design, underutilization and a lack of maintenance are among the potential pitfalls faced when building and implementing a bike share system.

For a bike program to be successful it is important that the correct technology and package of services involved be mated to the unique challenges that each program faces. Furthermore, there needs to be the political will and also the necessary infrastructure for safe cycling. It is therefore recommended that each city considering implementation of a bike share program have an independent assessment of the political backing, community needs, economics, technologies, logistical issues, service area, infrastructure and other challenges faced by a final system.

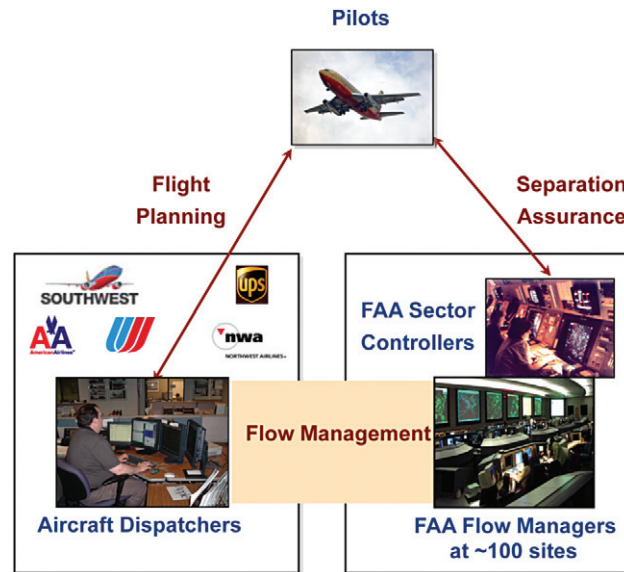


Figure 1 BSS steps.

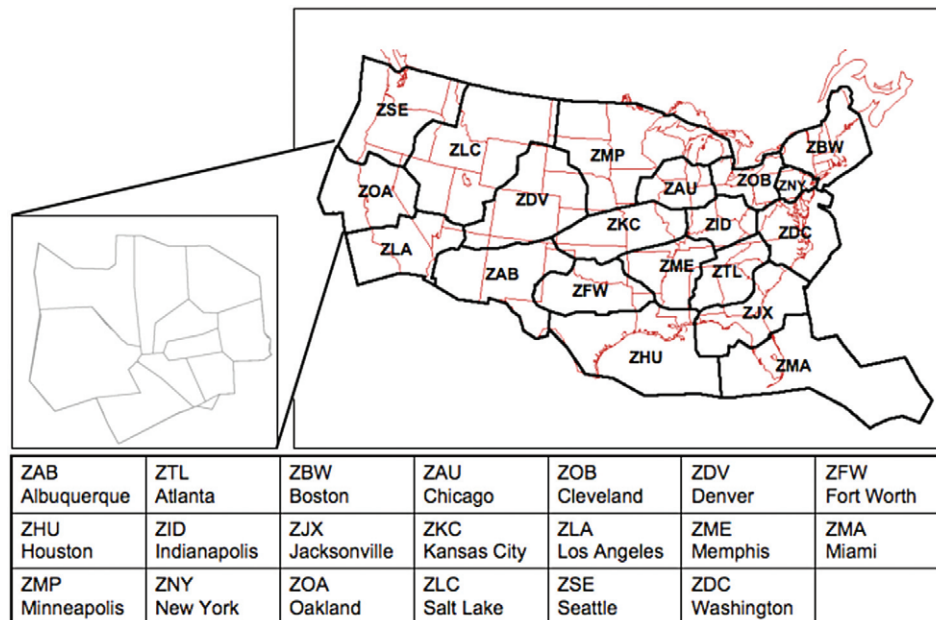


Figure 2 Pictures of the BSSs (David Rojas). (A) Marrakesh, Morocco, (B) Melbourne, Australia, (C) Fort Collins, United States (D) Mexico City, Mexico.

A key aspect of any bike share program is system and fleet maintenance and management. These activities can help to ensure the bike share system is in top operating order and sufficient bikes are available to accommodate all users. To ensure that bicycles are available at all stations, it is likely that bicycles will have to be redistributed from one station to another from time to time. Past performance of systems in Lyon and Paris indicates that many locations experience peak times of business when a rack will be either completely full or completely empty, making the rental or return of bikes impossible. Information about bicycle demand should be gathered through GPS units, radio frequency identification (RFID) tags and any other means used to track bicycle locations.

Innovative dockless mobile public-bikesharing programs have the potential to bring cycling back to places where cycling was common before but disappeared like China (Fig. 3). On the basis of GPS travel data from their bikes and a survey of more than 100,000 Mobike users, by April 2017—less than a year after the company's launch in Shanghai, China—Mobike Global estimated that of all journeys made by their users, the percentage of journeys by bicycle had increased from 5.5% to 11.6% (with the majority of journeys done to link with buses and trains) and the number of car journeys had halved (Ding et al., 2018).

However, mobile public-bike-sharing programs without docking stations are inherently subject to vandalism, misuse, and loss of bicycles. Cluttering of footpaths with bicycles and increasing traffic violations by cyclists have also been frequently reported. Despite



Figure 3 Dockless bikesharing system in Singapore (David Rojas).

the problems and challenges, mobile public-bike-sharing programs exemplify an alignment between private-sector interests and public health and sustainability, and offer unique opportunities to promote physical activity on a large scale. By the end of 2017, mobile public-bikesharing programs had spread to hundreds of cities across Asia, Europe, North America, and Australia, but unfortunately failed in many places due to competition, the abandoned bikes and lack of a sustainable business model.

BSSs may help normalize the image of cycling and reduce perceptions that cycling is “risky” or “only for sporty people.” Goodman et al. (2014) sought to compare the use of specialist cycling clothing between users of the London bicycle sharing system (LBSS) and cyclists using personal bicycles. They observed 3594 people on bicycles at 35 randomly selected locations across central and inner London. The 592 LBSS users were much less likely to wear helmets (16% vs. 64% among personal-bicycle cyclists), high-visibility clothes (11% vs. 35%) and sports clothes (2% vs. 25%). In total, 79% of LBSS users wore none of these types of specialist cycling clothing, as compared to only 30% of personal bicycle cyclists. This was true of male and female LBSS cyclists alike. They concluded that bicycle sharing systems may not only encourage cycling directly, by providing bicycles to rent, but also indirectly, by increasing the number and diversity of cycling “role models” visible.

Health Benefits of BSSs

It is well known that cycling is healthy due to physical activity. A number of studies have assessed the health implications of BSSs. As epidemiological studies are hard to conduct, they have mainly used health impact assessment methodology and have taken into account not only health benefits coming from physical activity but also risks from increased inhalation of air pollution and accidents.

The first to do so were Rojas-Rueda et al. (2011) in Barcelona (Fig. 4). They estimated the risks and benefits to health of travel by bicycle, using a bicycle sharing scheme, compared with travel by car in an urban environment in the new public bicycle sharing initiative, Bicing, in Barcelona, Spain (Fig. 5). At the time, the system had 181,982 subscribers. The primary outcome measure was all cause mortality for the three domains of physical activity, air pollution (exposure to particulate matter $<2.5 \mu\text{m}$ (PM_{2.5})), and road traffic incidents. The secondary outcome was change in levels of carbon dioxide emissions. Compared with car users the estimated increase annual change in mortality of the Barcelona residents using Bicing ($n = 181,982$) was 0.03 deaths from road traffic incidents and 0.13 deaths from air pollution. As a result of physical activity, were estimated 12.46 avoided deaths (benefit:risk ratio 77). Overall, taking to account risk and benefits, the total estimated health impact of the Bicing in Barcelona was 12.28 annual deaths avoided. As a result of journeys by Bicing, annual carbon dioxide emissions were reduced by an estimated 9,062,344 kg.

Woodcock et al. (2014) modeled the impacts of the bicycle sharing system in London on the health of its users (Fig. 5). They used total population operational registration and usage data for the London cycle hire scheme (collected April 2011–March 2012), surveys of cycle hire users (collected 2011), and London data on travel, physical activity, road traffic collisions, and particulate air pollution (PM_{2.5}) (collected 2005–12). Over the year examined the users made 7.4 million cycle hire trips (estimated 71% of cycling time by men). These trips would mostly otherwise have been made on foot (31%) or by public transport (47%). They found that the population benefits from the cycle hire scheme substantially outweighed harms (net change -72 DALYs (95% credible interval -110 to -43) among men using cycle hire per accounting year; -15 (-42 to -6) among women; note that negative DALYs represent a health benefit). When they modeled cycle hire injury rates as being equal to background rates for all cycling in central London, these benefits were smaller and there was no evidence of a benefit among women (change -49 DALYs (-88 to -17) among

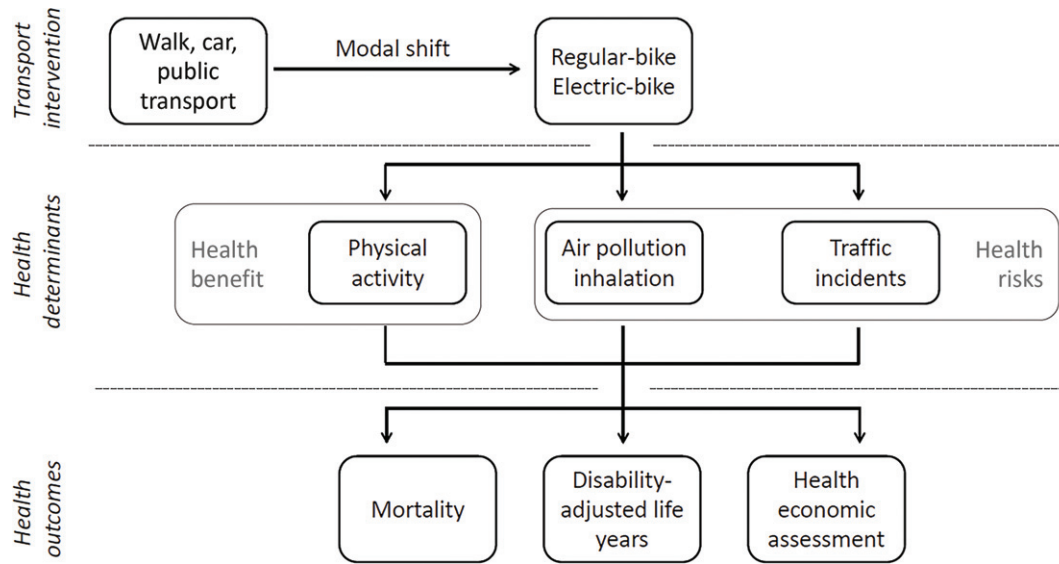


Figure 4 General framework used by health impact assessments on BSSs.



Figure 5 BSSs in Barcelona, Spain and London, UK. (David Rojas). (A) Barcelona, Spain. (B) London, UK.



Figure 6 Twelve European BSSs included in the health impacts assessment.

men; -1 DALY (-27 to 12) among women). This sex difference largely reflected higher road collision fatality rates for female cyclists. At older ages the modeled benefits of cycling were much larger than the harms. Using background injury rates in the youngest age group (15–29 years), the medium term benefits and harms were both comparatively small and potentially negative.

Otero et al. (2018) performed a multinational health impact assessment study to quantify the health risks and benefits of car trips substitution by bikes trips (regular-bikes and/or electric-bikes) from European BSS with >2000 bikes (Fig. 6). This one was the first health impact assessment including the impact of electric bicycles, taking into account the differences between physical activity, speed, and injury risk compared to regular-bikes. Four scenarios were created to estimate the annual expected number of deaths (increasing or reduced) due to physical activity, road traffic fatalities, and air pollution (Fig. 7). A quantitative model was built using data from transport and health surveys and environmental and traffic safety records. In addition, an economic assessment was included to translate mortality impacts in economic values using the value of statistical life. The study population was adult BSS users between 18 and 64 years old. Twelve BSSs (Barcelona, Brussels, Hamburg, Lille, Lyon, Madrid, Milan, Paris, Seville, Toulouse, Valencia, and Warsaw), were included in the analysis. The study found that between those 12 larger BSS in Europe the car substitution impact was very low (ranging from 4.5% to 12% of the BSS trips). In all scenarios and cities, the health benefits of physical activity outweighed the health risk of traffic fatalities and air pollution, with a ranged benefit:risk ratio of 9 (Milan) and 204 (Seville). It was estimated that 5.17 (95%CI: 3.11–7.01) annual deaths are avoided in the 12 BSSs, with the actual level of car trip substitution, corresponding to an annual saving of 18 million of Euros. It was also estimated that if all BSS trips replaced car trips, 73.25 deaths could be avoided each year (225 million Euros saving) in the 12 cities.

E-Bikes and Bike Sharing

E-bikes or electric bikes are also used as part of the bike sharing fleets in some cities (Fig. 8). Electric bikes can be classified in two different types, “standard assistance” e-bikes, and “high-assistance” e-bikes. Each type of e-bikes is related with different physical activity assistance and speed. The “standard assistance” e-bike is the most common type. In terms of physical activity, “standard assistance” e-bikes can be defined as an e-bike that requires 90% of the physical activity of a regular-bike, and “high assistance” e-bike was defined as an e-bike that requires 75% of the physical activity of a regular-bike (Otero et al., 2018). In terms of speed, “standard assistance” e-bikes were defined as an e-bike that increases in average 21% the speed of a regular-bike, and “high assistance” e-bike was defined as an e-bike that increases on average 33% the speed of a regular-bike (Otero et al., 2018). For traffic fatalities, e-bikes

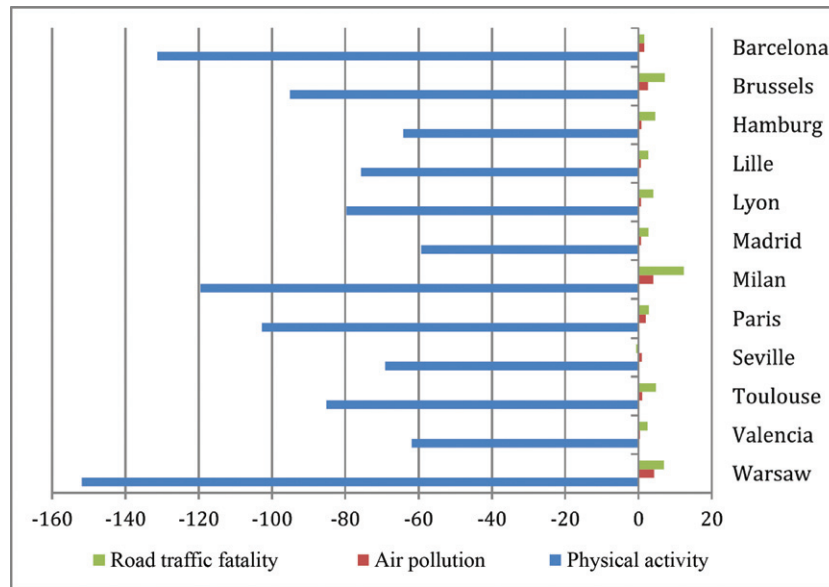


Figure 7 Deaths per 100,000 cyclists related to the health risks (road traffic fatality and air pollution inhalation) and benefits (physical activity).



Figure 8 E-bike public system in Copenhagen, Denmark and Madrid, Spain (David Rojas). (A) Copenhagen, (B) Madrid.

have been related to different levels of risk compared to regular bikes. Otero I. et al. reported an odds ratio of 1.92 (1.48–2.48) for traffic fatalities compared with a regular-bike, that means 92% more probabilities to suffer a traffic fatality in an e-bike vs. a regular bike. In their analysis, Otero I. et al. compared risk and benefits of e-bikes. This study estimated that although the electric bikes resulted in less health benefits, compared to the regular-bikes, electric bikes trips still providing more health benefits than traveling by car in those cities.

Helmets and Bike Sharing

The use of bicycle helmets to prevent or reduce serious head injuries is well established. However, it is unclear how to effectively promote helmet use, particularly in the context of bicycle-sharing programs. The bike sharing programs have the potential to increase physical activity and decrease air pollution, but anecdotal evidence suggests helmet use is lower among users of bicycle-sharing programs than cyclists on private bicycles. Kraemer et al. (2012) conducted a cross-sectional study to assess helmet use among users of a bicycle-sharing program in Washington, DC. Helmet use was significantly lower among cyclists on shared bicycles than private bicycles, highlighting a need for targeted helmet promotion activities.

Grenier et al. (2012) described bicycle helmet use among Montreal cyclists as a step toward injury prevention programming. Using a cross-sectional study design, cyclists were observed during 60-min periods at 22 locations on the island of Montreal. There were 1–3 observation periods per location. Observations took place between August 16 and October 31, 2011. A total of 4789 cyclists were observed. The helmet-wearing proportion of all cyclists observed was 46% (95% CI 44–47). Women had a higher helmet-wearing proportion than men (50%, 95% CI 47–52 vs. 44%, 95% CI 42–45, respectively). Youth had the highest helmet-wearing proportion (73%, 95% CI 64–81), while young adults had the lowest (34%, 95% CI 30–37). Visible minorities were observed wearing a helmet 29% (95% CI 25–34) of the time compared to Caucasians, 47% (95% CI 46–49). BIXI (bike sharing program) riders were observed wearing a helmet 12% (95% CI 10–15) of the time compared to riders with their own bike, 51% (95% CI 49–52). They concluded that although above the national average, bicycle helmet use in Montreal was still considerably low given that the majority of cyclists did not wear a helmet.

In a pilot study, Basch et al. (2014) estimated the prevalence of helmet use among a sample of Citi Bike program users in New York City. A total of 1054 cyclists were observed over 44 h and across the 22 busiest Citi Bike locations. Overall, 85.3% (95% CI 82.2, 88.4%) of the cyclists observed did not wear a helmet. Rates of helmet non-use were also consistent whether cyclists were entering or leaving the docking station, among cyclists using the Citi Bikes earlier versus later in the day, and among cyclists using the Citi Bikes on weekends versus weekdays. Improved understanding about factors that facilitate and hinder helmet use is needed to help reduce head injury risk among users of bicycle sharing programs.

Basch et al. (2015) studied the NYC's public bicycle share program, Citi Bike, the fastest growing program of its kind in the US, with nearly 100,000 members and more than 330 docking stations across Manhattan and Brooklyn. The purpose of this study was to assess helmet use behavior among Citi Bike riders at 25 of the busiest docking stations. The 25 Citi Bike Stations varied greatly in terms of usage: total number of cyclists ($N = 96$ –342), commute versus recreation (22.9%–79.5% commute time riders), weekday versus weekend (6.0%–49.0% weekend riders). Helmet use ranged between 2.9% and 29.2% across sites (median = 7.5%). A total of 4919 cyclists were observed, of whom 545 (11.1%) were wearing helmets. Incoming cyclists were more likely to wear helmets than outgoing cyclists (11.0% vs 5.9%, $P = .000$). NYC's bike share program endorses helmet use, but relies on education to encourage it. The data confirmed that, to date, this strategy has not been successful.

Mooney et al. (2019) counted cyclists over several hours at four locations in Seattle, WA. They categorized each rider according to whether he or she was wearing a helmet and to whether or not he or she was riding a bike share bike. Whereas 91% of riders of private bikes wore helmets, only 20% of bike share riders wore helmets. Moreover, in locations where a greater proportion of riders were on bikes share bikes, fewer riders of private bicycles wore helmets ($r = -0.96$, $P = 0.04$). The impact of bike sharing programs on helmet wearing norms among private bike riders warrants further exploration, but integration of bike helmet in the BSS could be an option to encourage bike helmet use among BSS users (Fig. 9).

BSS Users' Profiles

Ogilvie and Goodman (2012) sought to examine inequalities in uptake and usage of London's Barclays Cycle Hire (BCH) scheme. They obtained complete BCH registration data, and compared users with the general population. They examined usage levels by explanatory variables including gender, small-area income-deprivation and local cycling prevalence. 100,801 registered individuals made 2.5 million trips between July 2010 and March 2011. Compared with residents and workers in the central London area served by the scheme, registered individuals were more likely to be male and to live in areas of low deprivation and high cycling prevalence. Among those registered, females made 1.63 (95% CI 1.53, 1.74) fewer trips per month than males, and made under a fifth of all trips. Adjusting for the fact that deprived areas were less likely to be close to BCH docking stations, users in the most deprived areas made 0.85 (95% CI 0.63, 1.07) more trips per month than those in the least deprived areas. They concluded that females and residents in deprived areas were underrepresented among users of London's public bicycle sharing scheme. The scheme's planned expansion into more deprived areas has, however, the potential to create a more equitable uptake of cycling.

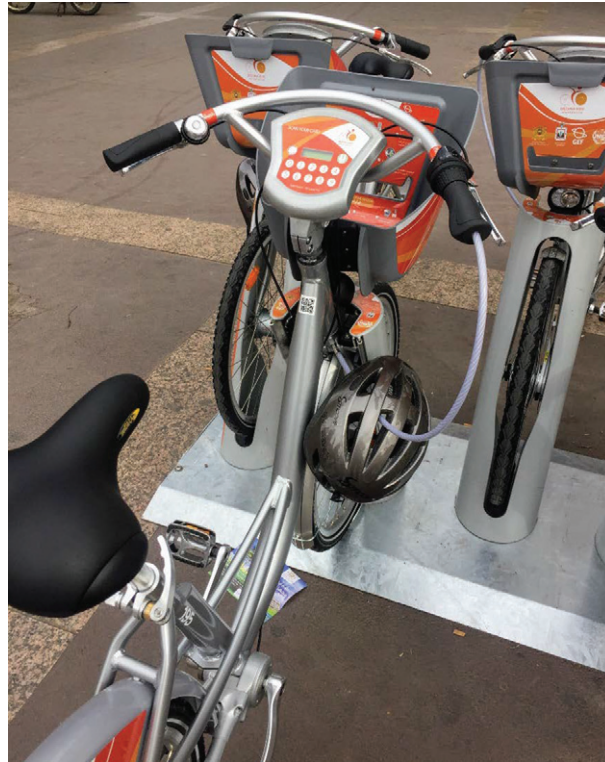


Figure 9 Helmet included in the Medina BSS in Marrakesh, Morocco (David Rojas).

Hosford et al. (2018) examined the socio-demographic and transportation characteristics of current, potential, and unlikely users of a public bicycle share program and identified specific motivators and deterrents to public bicycle share use. They used cross-sectional data from a 2017 Vancouver public bicycle share (Mobi by Shaw Go) member survey ($n = 1272$) and a 2017 population-based survey of Vancouver residents ($n = 792$). They categorized nonusers from the population survey as either potential or unlikely users based on their stated interest in using public bicycle share within the next year. Public bicycle share users in Vancouver tended to be male, employed, and have higher educations and incomes as compared to nonusers, and were more likely to use active modes of transportation. The vast majority of nonusers (74%) thought the public bicycle share program was a good idea for Vancouver. Of the nonusers, 23% were identified as potential users. Potential users tended to be younger, have lower incomes, and were more likely to use public transit for their main mode of transportation, as compared to current and unlikely users. The most common motivators among potential users related to health benefits, not owning a bicycle, and stations near their home or destination. The deterrents among unlikely users were a preference for riding their own bicycle, perceived inconvenience compared to other modes, bad weather, and traffic. Cost was a deterrent to one-fifth of unlikely users, notable given they tended to have lower incomes than current users.

Conclusions

BSSs have gained popularity over the recent years and can now be found in many cities around the World. Different technological improvements have been implemented (electric bikes, dockless, etc.) make it more attractive and easier for users. Although often introduced to reduce congestion and for climate mitigation reasons, they have only been able to absorb a small portion of car trips across cities. Despite the small impact on mode shift, they have been able to provide many health and health economic benefits, due to the increase of physical activity, and helping to normalize urban cycling through cities. BSSs are facing many challenges and opportunities, that should be tackle to improve their health benefits, such as provide more equally distributed bikes throughout the population, increase the use of helmets, and being able to attract more users from private motorized modes like cars or motorcycles. In general, BSSs should be seeing as a tool for public health promotion and prevention.

References

- Basch, C.H., Ethan, D., Rajan, S., Samayoa-Kozlowsky, S., Basch, C.E., 2014. Helmet use among users of the Citi Bike bicycle-sharing program: a pilot study in New York City. *J. Community Health* 39 (3), 503–507.
- Basch, C.H., Ethan, D., Zybert, P., Afzaal, S., Spillane, M., Basch, C.E., 2015. Public bike sharing in New York City: helmet use behavior patterns at 25 Citi Bike™ stations. *J. Community Health* 40 (3), 530–533.

- China News, 2011. Wuhan free rental bikes up to 70,000 intelligent rent but also the system starts. Retrieved August 9, 2012 from <http://www.chinanews.com/dl/2011/12-31/3575510.shtml>.
- Ding, D., Jia, Y., Gebel, K., 2018. Mobile bicycle sharing: the social trend that could change how we move. *Lancet Public Health* 3 (5), e215.
- Goodman, A., Green, J., Woodcock, J., 2014. The role of bicycle sharing systems in normalising the image of cycling: an observational study of London cyclists. *J. Transport Health* 1, 5–8.
- Hosford, K., Lear, S.A., Fuller, D., Teschke, K., Therrien, S., Winters, M., 2018. Who is in the near market for bicycle sharing? Identifying current, potential, and unlikely users of a public bicycle share program in Vancouver, Canada. *BMC Public Health* 18 (1), 1326.
- Kraemer, J.D., Roffenbender, J.S., Anderko, L., 2012. Helmet wearing among users of a public bicycle-sharing program in the district of Columbia and comparable riders on personal bicycles. *Am. J. Public Health* 102 (8), e23–e25.
- Larsen, J., 2013. Bike-Sharing Programs Hit the Streets in Over 500 Cities Worldwide. Earth Policy Inst. Washington, D.C.
- Meddin, R., 2011a. The Bike-sharing world: First days of summer 2011. Retrieved December 15, 2011, from <http://bike-sharing.blogspot.com/search?q=Brisbane>.
- Mooney, S.J., Lee, B., O'Connor, A.W., 2019. Free-floating bikeshare and helmet use in Seattle, WA. *J. Community Health* 44 (3), :577–579.
- Ogilvie, F., Goodman, A., 2012. Inequalities in usage of a public bicycle sharing scheme: socio-demographic predictors of uptake and usage of the London (UK) cycle hire scheme. *Prev. Med.* 55 (1), 40–45.
- Oortwijn, J., 2015. Bike-Sharing Systems to Grow to Multi-Billion Business. Available from: <http://www.bike-eu.com/sales-trends/nieuws/2015/10/bike-sharing-systems-togrow-to-multi-billion-business-10124878> [accessed 31 January 2017].
- Otero, I., Nieuwenhuijsen, M.J., Rojas-Rueda, D., 2018. Health impacts of bike sharing systems in Europe. *Environ. Int.* 115, 387–394.
- Rojas-Rueda, D., de Nazelle, A., Tainio, M., Nieuwenhuijsen, M.J., 2011. The health risks and benefits of cycling in urban environments compared with car use: health impact assessment study. *nBMJ* 343, d4521.
- Shaheen, S., Guzman, S., Zhang, H., 2010. Bikesharing in Europe, the Americas, and Asia. *Transport. Res. Rec.* 2143, 159–167, doi:10.3141/2143-20.
- Woodcock, J., Tainio, M., Cheshire, J., O'Brien, O., Goodman, A., 2014. Health effects of the London bicycle sharing system: health impact modelling study. *BMJ* 348, g425.

Further Reading

- Estevan, I., Queralt, A., Molina-García, J., 2018. Biking to school: the role of bicycle-sharing programs in adolescents. *J. Sch Health* 88 (12), 871–887.
- Loaiza-Monsalve, D., Riascos, A.P., 2019. Human mobility in bike-sharing systems: structure of local and non-local dynamics. *PLoS One* 14 (3), e0213106.
- Transport for London, 2010. Travel in London report 3. London: Transport for London, Retrieved from <http://www.tfl.gov.uk/assets/downloads/corporate/travel-in-london-report-3.pdf>.

E-Bikes and Health

Aslak Fyhri, Hanne Beate Sundfør, Institute of Transport Economics, Oslo, Norway

© 2021 Elsevier Ltd. All rights reserved.

Introduction	393
Definitions	393
E-Bikes	393
Physical Activity	393
Metabolic Equivalent of the Task	394
Active Transport and Health Benefits	394
Intensity of Physical Activity From Using E-Bikes	394
E-Bikes and Cardiorespiratory Fitness Improvement	394
Substitution Effects	395
Effects on Travel Behavior	395
Psychological Outcomes From Riding an E-Bike	396
Accident Risk From E-Bikes	396
E-Bikes and Health Summarized	396
References	397
Further Reading	398

Introduction

The use of E-bikes have seen a rapid growth during the last few decades. In Europe, the numbers of e-bikes increased from 588,000 in 2010 to 1,667,000 in 2016 (CONEBI, 2017).

Compared to conventional bicycles, e-bikes can maintain speed with less physical effort due to the electric-motor support, hence overcoming some of the common barriers to cycle commuting (De Geus and Hendriksen, 2015). However, given that e-bikes require less muscular effort, the question then is if the pedaling is sufficiently strenuous to give health benefits, and most importantly, what the public health effects are from them.

Definitions

E-Bikes

In general, e-bikes in a European, North American, and Australian context refers to bicycle-style e-bike design (i.e., the bicycle has functional pedals, but is assisted with an electric motor) as opposed to scooter-style e-bike design (with no pedals) (Fishman and Cherry, 2016). E-bikes following the regulations made by the European Union are formally known as electric pedal assisted cycle (EPAC), but are also known as pedelecs. The EU regulations states that the motor assistance is limited to 250 W, and caps assistance at 25 km/h (European Committee for Standardization, 2011). Pedelecs are in most countries legally classified as bicycles, as they require pedaling for electrical assistance to be provided (Fishman and Cherry, 2016). In the following we use the term *e-bike* to denote a bicycle of the pedelec type.

Physical Activity

Physical activity (PA) from cycling is defined as any bodily movement produced by skeletal muscles that requires energy expenditure (Warburton and Bredin, 2017). It has both an acute and long term effect, and reduces risk of several noncommunicable diseases (Warburton and Bredin, 2017). In general, too little physical activity can have a negative effect on health, and increase the risk of diseases such as heart attack, cancer and diabetes (Bauman, 2004; Warburton and Bredin, 2017). The intensity of an activity can be measured by means of oxygen uptake, metabolic equivalents, energy expenditure per minute, heart rate and power output (Åstrand and Rodahl, 2003). Physical activity is normally categorized as low, moderate, and vigorous intensity.

The health benefits of physical activity depends on the individual's baseline cardiorespiratory fitness (i.e., weight, maximal O_2 uptake, etc.) and the frequency, duration, and intensity of the activity performed (Gjerset, 1992). Maximal oxygen uptake ($VO_{2\text{-max}}$) is the amount of oxygen that the body can utilize in 1 min.

To gain health benefits, The World Health Organization (WHO) recommends that healthy adults engage in at least 150 min of moderate physical activity or 75 min of vigorous physical activity per week (World Health Organization, 2016). Behind these recommendations lies an understanding of the physiological mechanisms in which the higher the intensity, the shorter the duration

Table 1 Categories of MET levels and related activities

<i>MET</i>	<i>Physical activity</i>	<i>Examples</i>
Below 3	Light	Sitting, walking slowly
3–6	Moderate	Walking, housework
6 and above	Vigorous	Cycling, jogging, aerobics

MET, Metabolic equivalent of the task.

is needed for the same health gain. If an individual has a “low” baseline fitness, less duration, frequency and/or intensity is needed for enhanced health effect (Gjerset, 1992; Åstrand and Rodahl, 2003).

Metabolic Equivalent of the Task

In order to have an objective measure of physical activity we often refer to the Metabolic equivalent of the task (MET), which is an objective measure of rate of energy use relative to the person’s mass. One MET is defined as the energy used at rest (resting metabolism). **Table 1** shows categories of MET levels and related activities.

To be intense enough to promote and maintain health an exercise should be at least three METs, which is threshold for moderate intensity (Haskell et al., 2007).

Active Transport and Health Benefits

In many countries, there is an aim to increase cycling as a mode of transport. One reason for this is the potential for overall increased physical activity and subsequent population health benefits (Oja et al., 2011). Increasing physical activity levels through active transport can potentially give many individuals an adequate level of PA (Ainsworth et al., 2011; de Geus et al., 2007). Active commuting has been found to be associated with a lower risk of all-cause mortality and cancer incidence (Celis-Morales et al., 2017). The greatest gains in health outcomes from active travel are reported in the least active individuals (Oja et al., 2011).

Intensity of Physical Activity From Using E-Bikes

Several studies have compared power expenditure when cycling on an e-bike with a conventional bike (CB) (e.g., Berntsen et al., 2017; Gojanovic et al., 2011; Theurel et al., 2012), using different research designs and physiological measures. Typically, all of these studies find that people spend less energy and less time when riding an e-bike than a conventional bike for the same trip. As an example, a study where subjects performed trips at a self-selected pace ($N = 18$) found that VO_2 max during cycling with an e-bike (high assistance) was 55%, compared to 66% for an e-bike (standard assistance) and 73% for a conventional bike (Gojanovic et al., 2011).

The differences between e-bikes and CBs are most evident uphill (Berntsen et al., 2017; Gojanovic et al., 2011; Langford et al., 2017), which illustrates an interesting case of adaptation among users. When allowed to self-select pace and intensity, people may select a similar physiological intensity across activities regardless of the assistance (conventional bike vs e-bike), thereby resulting in similar physiological outcomes on flat and downhill segments. Since there is a lower threshold for how slowly the conventional bike can be ridden without losing balance, there is less opportunity to adjust effort levels to produce comparable intensity levels, when riding uphill.

A systematic review by Bourne et al. (2018) shows that even if e-bikes demand less effort than CBs, they do help to provide at least moderate intensity. Importantly, this finding is not dependent upon baseline activity levels: both active and inactive individuals get some exercise.

E-Bikes and Cardiorespiratory Fitness Improvement

Cycling on a regular basis (at least three times a week) with a conventional bike positively influences physical fitness (de Geus et al., 2009; Hendriksen et al., 2000; Moller et al., 2011; Oja et al., 2011). Given that e-bike users spend less energy than conventional bike users, the question then is, can e-bikes improve people’s cardiorespiratory fitness?

The above-mentioned studies indicate that there is potential. The percentage of peak $VO_{2\text{-max}}$ (i.e., the maximum amount of oxygen that the body can make use of in one minute) are found to range from 51 to 75 for e-cycling (Bourne et al., 2018). These values exceed the hypothesized minimum of threshold (45% of $VO_{2\text{-max}}$) required to improve cardiorespiratory fitness (Swain and Franklin, 2002).

Intervention studies have looked at e-cycling in a population over time (De Geus et al., 2013; Höchsmann et al., 2018; Lobben et al., 2018; Peterman et al., 2016). Peterman et al. (2016) conducted a study with 20 sedentary commuters, and found that after four weeks of e-bike commuting participants improved their glucose tolerance, VO_2 max (8% increment), and maximal power output. These results indicate that for inactive individuals, cardiorespiratory fitness could be improved.

However, since only a few of these studies included control groups, the number of participants tended to be quite low and study periods were limited, more research is needed to give firm conclusions about the long-term health impacts.

Substitution Effects

A key issue from a public health perspective is whether the health effect of increased cycling is cancelled out by a reduction in other physical activities—a *substitution effect* (Thomson et al., 2008).

In transport planning, cost benefit analysis (CBA) is a central tool. In a CBA all the measurable benefits of a given measure or investment are weighed against its cost. When planning investments for better cycling infrastructure, health benefits account for a large portion of the total benefits; as much as 55%–75% according to Slensminde (2004). However, one large question that is unresolved is the size of the substitution effect. In general, the empirical evidence for substitution is weak (Thomson et al., 2008).

Evidence seems to indicate that people have a total “time budget” from which increased time spent traveling would imply decreased time spent on other activities (Börjesson and Eliasson, 2012). If this is the case, people are likely to have a total (physical or psychological) “budget” for PA, and that increased cycling (with conventional bikes) would cause physical or mental fatigue affecting engagement in other activities negatively. However, this study was based on hypothetical self-reports, which arises some questions about the validity of the results. The finding is also somewhat counter to general knowledge from behavior and motivational impacts, that indicates that conducting one positive activity might entice people to conduct positive behavior in other health domains (Mata et al., 2009; Sniehotta et al., 2005).

Even if people do have a time-budget, e-bikes in particular might play a different role in such an equation than CBs, given that they require less energy and might not spend so much of the total “PA-time budget”. Further to this, active transport might have larger public health benefits than the typical infrequent higher intensity activities it replaces (Praznocy, 2012).

Some studies have assessed e-bikes and substitution effects. Castro et al. (2019) found physical activity (measured as MET minutes per week) from travel-related activities to be similar for e-bike and CB, implying that e-bikers does not necessarily reduce other PA. An intervention study found that for those starting to use an e-bike other physical activity was not significantly affected, that is indicating no substitution effect (Sundfør and Fyhri, 2017).

Effects on Travel Behavior

A person who buys an e-bike and takes up cycling will increase his or her physical activity levels, if this cycling does not substitute an evening gym class or a run around the park. So from a public health perspective increased cycling from e-bikes would in itself constitute a positive physical activity result.

But from a transport policy perspective, we are not really interested in generating more transport, but in replacing nonsustainable or passive transport with active and sustainable transport. From this perspective, replacing your regular trip on a conventional bicycle with an e-bike has a negative health effect, given no other adjustment. Only when people replace a passive mode of transport (i.e., car or public transport) will there be a positive health benefit, as the energy expenditure (indicated by METs) is higher when riding an e-bike compared to sitting still. So from the perspective of active mobility we are interested in knowing what travel modes e-bike replace. Rather than just looking at new trips, we are looking at travel mode share changes.

A review by Cairns et al. (2017) looking at e-bikes’ mode change effect found that proportion of car journeys being replaced by bike journeys ranged from 16% to 76%. A challenge with many studies looking at mode change is that they are either retrospective (i.e., people report previous travel behavior after they have purchased an e-bike) or cross-sectional (comparison of e-bike users with non-users at one time-point), thus giving little control over confounding factors. Kroesen (2017) found that e-bike users cycled 3.0 km a day, compared to 2.6 km a day for conventional cyclists, and also travel less with car than conventional cyclists.

A study using panel data from the Dutch Mobility survey, compared travel patterns for people who have bought an e-bike between two time periods ($N = 107$). The results of this study confirms previous findings: e-bike users reduce their share of conventional cycling, increase their total amount of cycling and reduce their amount of car driving (Sun et al., 2020).

In two studies conducted in Norway people using an e-bike were compared to a control group without access to an e-bike. In the first study, 60 participants were allowed to borrow the e-bikes for 2 or 4 weeks. The trial group increased their total bicycle use, and saw a twofold increase in cycling as share of all travel (Fyhri and Fearnley, 2015).

The large change in cycling previously observed in schemes and more hypothetical research designs was also replicated with actual customers. In the second study people who bought an e-bike ($N = 39$) were compare to a group wanting to buy one ($N = 142$) as well as broader comparison group ($N = 767$). People who purchased an e-bike increased their bicycle use from 2.1 to 9.2 km/day on average, representing a change in bike as share of all transport from 17% to 49%. An important point with this study is that the control groups also had an increase in their cycling activity during the study period (due to seasonal differences). But statistical tests showed that this increase was far lower than the observed changes for the people who had bought an e-bike.

Psychological Outcomes From Riding an E-Bike

Cycling to work on a conventional bike gives more positive affect and enjoyment compared to other modes of travel (Gatersleben and Uzzell, 2007; Rissel et al., 2016). One reason for this is that cycling happens outdoors and often in green space, which has been proven to be of benefit for mental restoration (Barton and Pretty, 2010). Combining physical activity and nature experiences might also work together to create a synergetic effect on health (Barton and Pretty, 2010). It has also been suggested that cycling to work improve vitality, cognitive performance and work capacity (Calogiuri et al., 2016). Other suggested benefits from active commuting are (1) Greater extent of commuting control and “arrival-time reliability”; (2) Sensory stimulation reaching enjoyable levels; (3) The mental effects of moderate intensity physical activity making one feel better; and (4) More likely for social interaction to occur (Wild and Woodward, 2019).

There are a few studies that have looked specifically at e-bikes and psychological outcomes. Page and Nilsson (2017) found that people who change their behavior to active commuting by e-bike report more positive affect, better physical health and more productive organizational behavior, compared with passive commuters. Some studies report enjoyment and positive experiences of the user (Fyhri and Fearnley, 2015; Plazier et al., 2017; Popovich et al., 2014), as well as favorable effects on mental well-being (Jones et al., 2016), and others conclude that due to the decreased amount of effort needed, it might be easier to concentrate (Theurel et al., 2012). However, few controlled studies have been performed and the evidence of psychological health benefits is inconclusive (Bourne et al., 2018).

Accident Risk From E-Bikes

In general, bicycles have a higher accident risk than most other modes of transport (Elvik and Vaa, 2005). Some have raised the concern that e-bikes are even more dangerous than conventional bikes, and that promoting them will lead to more accidents. When discussing this claim it is important to distinguish between a mere *exposure effect*, differences in *injury severity* and finally an *increased risk* and.

The exposure effect has to do with the question about whether people with e-bikes cycle more than people with CBs, and by this are more exposed to the risk of being a cyclist in traffic.

There have been some studies investigating if e-bikes have more accidents. Surprisingly few of these have made account for exposure differences. For example, an Israeli study went as far to advise against e-bikes based on an observed increase in accident figures (Siman-Tov et al., 2017). However, this study did not take into account possible changes in number of e-bikes in traffic during the study period. Also, as we have seen, people tend to ride longer when they switch from a conventional bike to an e-bike.

Another limitation of this study, as with many other studies, is that they are based on official injury data, or trauma registers. Even if these data have the advantage of following a strict protocol when an accident is defined, they only cover a small share of all accidents. Since, underreporting of bicycle accidents is well known, typically only 10% of bicycle traffic accidents are reported in official figures (Bjørnskau, 2005; Shinar et al., 2018), this implies that only the most serious accidents are included. Related to this, the question then is whether e-bikes lead to more *severe accidents*. Part of the reasoning behind this hypothesis is that they are heavier, and also that they can keep a higher speed. Most studies looking at injury severity conclude that there is no difference between e-bikes and conventional bike (Otte et al., 2014; Papoutsi et al., 2014; Weber et al., 2014).

Some early studies have indicated and increase in *accident risk* for e-bikes (Schepers et al., 2014), but later studies have not replicated this (Schepers et al., 2018; Weiss et al., 2018). A study, carried out in Norway ($N = 7752$), combined three data sets to compare conventional and electric bicycles, while at the same time controlling for age, gender and exposure (Fyhri et al., 2019). The study found an overall risk increase for e-bike users, but suggests that all of this risk can be attributed to female cyclists. In other words, women have a higher risk on e-bikes, whereas men do not. Men have a higher total risk. According to the study some of this risk increase can be attributed to experience with the bicycle. In other words, women are more likely to be inexperienced cyclists than men, and since they also are more likely to take up cycling with an e-bike, this will inevitably lead to more accidents. Future research should investigate if this is just a transitory effect, or if there is the risk difference is maintained even after e-bikes have been around for a while.

For conventional bicycles the health benefit of increased physical activity accumulated though cycling outweighs the risk of injuries (Andersen et al., 2018; De Hartog et al., 2010; WHO, 2019). Given that the overall risk lever for e-bikes is not higher than that of CBs, it seems safe to conclude that the same will hold for e-bikes, as well.

E-Bikes and Health Summarized

Physical activity is good for your health, and active travel can increase the total amount of physical activity you conduct during a day. The role the e-bike plays in this equation is not totally clear, but the evidence leans to the positive side. For the same pace and distance, an e-bike user will have less exercise than a conventional cyclist, due to the assistance of the electrical motor. Riding an e-bike will give different health gains for people who normally are inactive than people who are used to being active. Still, evidence shows that the e-bike provides physical activity at a moderate intensity level, regardless of baseline activity level.

The most important factor to consider when looking at the e-bike from an active mobility perspective is what transport mode it replaces. Second to that is the potential substitution effect. If people switch from a passive mode of transport to e-bikes this would imply an increase in physical activity levels. If, however, the e-bike replaces a conventional bike, physical activity will decrease. Evidence so far indicate that people with e-bike increase their amount of cycling compared to a situation where they do not have access to an e-bike. Longitudinal studies should be conducted to confirm this. Even if some of this cycling replaces former walking or conventional bike trips, the total time spent on active mobility is increased. Other physical activity is not significantly affected when starting to use an e-bike, implying no substitution effect.

The e-bike seems to cater for novice cyclists and increases the likelihood that these individuals take up and maintain cycling. It could also attract groups where the conventional bike is not seen as an option, such among elderly and sedentary individuals.

One challenge that then arises is that their lack of previous cycling experience might be a risk factor. Even if the balance of the evidence now points to no overall risk difference between e-bikes and CBs, there might be an elevated risk for novice riders, who often happen to be female. The health benefit of increased physical activity accumulated through cycling (with any bike) is normally considered as higher than the negative outcome from injuries.

References

- Ainsworth, B.E., Haskell, W.L., Herrmann, S.D., Meckes, N., Bassett Jr., D.R., Tudor-Locke, C., Greer, J.L., Vezina, J., Whitt-Glover, M.C., Leon, A.S., 2011. 2011 Compendium of Physical Activities: a second update of codes and MET values. *Med. Sci. Sports Exercise* 43, 1575–1581.
- Andersen, L.B., Riiser, A., Rutter, H., Goenka, S., Nordengen, S., Solbraa, A., 2018. Trends in cycling and cycle related injuries and a calculation of prevented morbidity and mortality. *J. Transport Health* 9, 217–225.
- Barton, J., Pretty, J., 2010. What is the best dose of nature and green exercise for improving mental health? A multi-study analysis. *Environ. Sci. Technol.* 44, 3947–3955.
- Bauman, A.E., 2004. Updating the evidence that physical activity is good for health: an epidemiological review 2000–2003. *J. Sci. Med. Sport* 7, 6–19.
- Berntsen, S., Malnes, L., Langaker, A., Bere, E., 2017. Physical activity when riding an electric assisted bicycle. *Int. J. Behav. Nutr. Phys. Activity* 14, 55.
- Bjørnskau, T., 2005. Sykkellulykker. Ulykkestyper, skadekonsekvenser og risikofaktorer. Transportøkonomisk Institutt, Oslo.
- Bourne, J.E., Sauchelli, S., Perry, R., Page, A., Leary, S., England, C., Cooper, A.R., 2018. Health benefits of electrically-assisted cycling: a systematic review. *Int. J. Behav. Nutr. Phys. Activity* 15, 116–1116.
- Börjesson, M., Eliasson, J., 2012. The value of time and external benefits in bicycle appraisal. *Transport Res. Part A: Policy Pract.* 46, 673–683.
- Cairns, S., Behrendt, F., Raffo, D., Beaumont, C., Kiefer, C., 2017. Electrically-assisted bikes: potential impacts on travel behaviour. *Transport Res. Part A: Policy Pract.* 103, 327–342.
- Calogiuri, G., Evensen, K., Weydahl, A., Andersson, K., Patil, G., Ihlebaek, C., Raanaas, R.K., 2016. Green exercise as a workplace intervention to reduce job stress. Results from a pilot study. *Work* 53, 99–111.
- Castro, A., Gaupp-Berhausen, M., Dons, E., Standaert, A., Laeremans, M., Clark, A., Anaya, E., Cole-Hunter, T., Avila-Palencia, I., Rojas-Rueda, D., Nieuwenhuijsen, M., Gerike, R., Panis, L.I., De Nazelle, A., Brand, C., Raser, E., Kahlmeier, S., Götschi, T., 2019. Physical activity of electric bicycle users compared to conventional bicycle users and non-cyclists: Insights based on health and transport data from an online survey in seven European cities. *Transport. Res. Interdisciplinary Perspect.* 100017.
- Celis-Morales, C.A., Lyall, D.M., Welsh, P., Anderson, J., Steell, L., Guo, Y., Maldonado, R., Mackay, D.F., Pell, J.P., Sattar, N., Gill, J.M.R., 2017. Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study. *BMJ* 357, j1456.
- CONEBI, 2017. European Bicycle Market 2017 edition - Industry and Market Profile (2016 statistics) [Online]. Available: <http://asociacionambe.es/wp-content/uploads/2014/12/European-Bicycle-Industry-and-Market-Profile-2017-with-2016-data.pdf> [Accessed].
- De Geus, B., De Smet, S., Nijs, J., Meeusen, R., 2007. Determining the intensity and energy expenditure during commuter cycling. *Br. J. Sports Med.* 41, 8–12.
- De Geus, B., Hendriksen, I., 2015. Cycling for transport, physical activity and health: what about Pedelecs? *Cycling Futures: From Research into Practice* 28, 17–31.
- De Geus, B., Joncheere, J., Meeusen, R., 2009. Commuter cycling: effect on physical performance in untrained men and women in Flanders: minimum dose to improve indexes of fitness. *Scand J Med Sci Sports* 19, 179–187.
- De Geus, B., Kempenaers, F., Lataire, P., Meeusen, R., 2013. Influence of electrically assisted cycling on physiological parameters in untrained subjects. *Eur. J. Sport Sci.* 13, 290–294.
- De Hartog, J.J., Boogaard, H., Nijland, H., Hoek, G., 2010. Do the health benefits of cycling outweigh the risks? *Environ. Health Perspect.* 118, 1109–1116.
- Elvik, R., Vaa, T., 2005. *The Handbook of Road Safety Measures*. Elsevier, Amsterdam; San Diego, CA.
- European Committee for Standardization, 2011. Cycles - Electrically power assisted cycles - EPAC Bicycles. [Online]. Available: https://standards.cen.eu/dyn/www/?p=204:110:0:::FSP_PROJECT,FSP_ORG_ID:37542,6314&cs=11DCF234E608CBEEA798ED6BD89F9CCE5 [Accessed].
- Fishman, E., Cherry, C., 2016. E-bikes in the mainstream: reviewing a decade of research. *Transport Rev.* 36, 72–91.
- Fyhri, A., Fearnley, N., 2015. Effects of e-bikes on bicycle use and mode share. *Transport. Res. Part D-Transport Environ.* 36, 45–52.
- Fyhri, A., Johansson, O.J., Bjørnskau, T., 2019. Gender differences in accident risk with e-bikes – survey data from Norway Accident. *Anal. Prevent.*
- Gatersleben, B., Uzzell, D., 2007. Affective appraisals of the daily commute - Comparing perceptions of drivers, cyclists, walkers, and users of public transport. *Environ. Behav.* 39, 416–431.
- Gjerset, A., 1992. *Idrettens treningslære* (eng. Sport Training Sciences). Universitetsforlaget, Oslo, Norway.
- Gojanovic, B., Welker, J., Iglesias, K., Daucourt, C., Gremion, G., 2011. Electric bicycles as a new active transportation modality to promote health. *Med. Sci. Sports Exercise* 43, 2204–2210.
- Haskell, W.L., Lee, I.M., Pate, R.R., Powell, K.E., Blair, S.N., Franklin, B.A., Macera, C.A., Heath, G.W., Thompson, P.D., Bauman, A., 2007. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Med. Sci. Sports Exerc.* 39, 1423–1434.
- Hendriksen, I.J., Zuiderveld, B., Kemper, H.C., Bezemer, P.D., 2000. Effect of commuter cycling on physical performance of male and female employees. *Med. Sci. Sports Exerc.* 32, 504–510.
- Höchsmann, C., Meister, S., Gehrig, D., Nussbaumer, M., Rossmeissl, A., Hanssen, H., Schmidt-Trucksäss, A., Schäfer, J., Gordon, E., Yanlei, L., 2018. Effect of e-bike versus bike commuting on cardiorespiratory fitness in overweight adults: a 4-week randomized pilot study. *Clin. J. Sport Med.* 28, 255–265.
- Jones, T., Harms, L., Heinen, E., 2016. Motives, perceptions and experiences of electric bicycle owners and implications for health, wellbeing and mobility. *J. Transport Geogr.* 53, 41–49.
- Kroesen, M., 2017. To what extent do e-bikes substitute travel by other modes? Evidence from the Netherlands. *Transport. Res. Part D: Transport Environ.* 53, 377–387.
- Langford, B.C., Cherry, C.R., Bassett Jr., D.R., Fitzhugh, E.C., Dhakal, N., 2017. Comparing physical activity of pedal-assist electric bikes with walking and conventional bicycles. *J. Transport Health* 6, 463–473.
- Lobben, S.E., Mildestvedt, T., Malnes, L., Berntsen, S., Bere, E., Tjelta, L.I., Kristoffersen, M., 2018. Bicycle usage among inactive adults provided with electrically assisted bicycles. *Acta Kinesiologiae Universitatis Tartuensis* 24, 60–73.
- Mata, J., Silva, M.N., Vieira, P.N., Carraca, E.V., Andrade, A.M., Coutinho, S.R., Sardinha, L.B., Teixeira, P.J., 2009. Motivational “spill-over” during weight control: increased self-determination and exercise intrinsic motivation predict eating self-regulation. *Health Psychol.* 28, 709–716.

- Moller, N.C., Ostergaard, L., Gade, J.R., Nielsen, J.L., Andersen, L.B., 2011. The effect on cardiorespiratory fitness after an 8-week period of commuter cycling—a randomized controlled study in adults. *Prev. Med.* 53, 172–177.
- Oja, P., Titze, S., Bauman, A., De Geus, B., Krenn, P., Reger-Nash, B., Kohlberger, T., 2011. Health benefits of cycling: a systematic review. *Scand J. Med. Sci. Sports* 21, 496–509.
- Otte, D., Facius, T., Mueller, C., 2014. Pedelecs im Unfallgeschehen und Vergleich zu konventionellen nicht motorisierten Zweirädern. *VKU Verkehrsunfall und Fahrzeugtechnik* 52.
- Page, N.C., Nilsson, V.O., 2017. Active commuting: workplace health promotion for improved employee well-being and organizational behavior. *Front. Psychol.* 7, 1994.
- Papoutsi, S., Martinolli, L., Braun, C.T., Exadaktylos, A.K., 2014. E-bike injuries: experience from an urban emergency department—A retrospective study from Switzerland. *Emergency Med. Int.* 2014.
- Peterman, J.E., Morris, K.L., Kram, R., Byrnes, W.C., 2016. Pedelecs as a physically active transportation mode. *Eur. J. Appl. Physiol.* 116, 1565–1573.
- Plazier, P.A., Weitkamp, G., Van Den Berg, A.E., 2017. Cycling was never so easy! An analysis of e-bike commuters' motives, travel behaviour and experiences using GPS-tracking and interviews. *J. Transport Geogr.* 65, 25–34.
- Popovich, N., Gordon, E., Shao, Z., Xing, Y., Wang, Y., Handy, S., 2014. Experiences of electric bicycle users in the Sacramento, California area. *Travel Behav. Soc.* 1, 37–44.
- Praznoczy, C., 2012. Les bénéfices et les risques de la pratique du vélo - Evaluation en Île-de-France. *Observatoire Régional de la Santé Île-de-France* 163.
- Rissel, C., Crane, M., Wena, L.M., Greaves, S., Standen, C., 2016. Satisfaction with transport and enjoyment of the commute by commuting mode in inner Sydney. *Health Promotion J. Australia* 27, 80–83.
- Schepers, J., Fishman, E., Den Hertog, P., Wolt, K.K., Schwab, A., 2014. The safety of electrically assisted bicycles compared to classic bicycles. *Accid. Anal. Prevent.* 73, 174–180.
- Schepers, J.P., Wolt, K.K., Fishman, E., 2018. The Safety of E-Bikes in The Netherlands. Discussion paper. International Transport Forum.
- Shinar, D., Valero-Mora, P., Van Strijp-Houtenbos, M., Haworth, N., Schramm, A., De Bruyne, G., Cavallo, V., Chliaoutakis, J., Dias, J., Ferraro, O.E., Fyhri, A., Sajatovic, A.H., Kuklane, K., Ledesma, R., Mascarelli, O., Morandi, A., Muser, M., Otte, D., Papadakaki, M., Sanmartin, J., Dulf, D., Saplioglu, M., Tzamalouka, G., 2018. Under-reporting bicycle accidents to police in the COST TU1101 international survey: Cross-country comparisons and associated factors. *Accid. Anal. Prevent.* 110, 177–186.
- Siman-Tov, M., Radomislensky, I., Group, I.T., Peleg, K., 2017. The casualties from electric bike and motorized scooter road accidents. *Traffic Injury Prevent.* 18, 318–323.
- Snihotta, F.F., Scholz, U., Schwarzer, R., 2005. Bridging the intention-behaviour gap: planning, self-efficacy, and action control in the adoption and maintenance of physical exercise. *Psychol. Health* 20, 143–160.
- Sun, Q., Feng, T., Kemperman, A., Spahn, A., 2020. Modal shift implications of e-bike use in the Netherlands: Moving towards sustainability? *Transport. Res. Part D: Transport Environ.* 78, 102202.
- Swain, D.P., Franklin, B.A., 2002. VO2 reserve and the minimal intensity for improving cardiorespiratory fitness. *Med Sci. Sports Exercise* 34.
- Sælensminde, K., 2004. Cost-benefit analyses of walking and cycling track networks taking into account insecurity, health effects and external costs of motorized traffic. *Transport Res. Part A: Policy Pract.* 38, 593–606.
- Theurel, J., Theurel, A., Lepers, R., 2012. Physiological and cognitive responses when riding an electrically assisted bicycle versus a classical bicycle. *Ergonomics* 55, 773–781.
- Thomson, H., Jepson, R., Hurley, F., Douglas, M., 2008. Assessing the unintended health impacts of road transport policies and interventions: translating research evidence for use in policy and practice. *BMC Public Health* 8, 339.
- Warburton, D.E.R., Bredin, S.S.D., 2017. Health benefits of physical activity: a systematic review of current systematic reviews. *Curr. Opin. Cardiol.* 32, 541–556.
- Weber, T., Scaramuzza, G., Schmitt, K.-U., 2014. Evaluation of e-bike accidents in Switzerland. *Accid. Anal. Prevent.* 73, 47–52.
- Weiss, R., Juhra, C., Wieskötter, B., Weiss, U., Jung, S., Raschke, M., 2018. How Probable is it That Seniors Using an E-Bike Will Have an Accident? - A New Health Care Topic, Also for Consulting Doctors. *Zeitschrift Fur Orthopädie Und Unfallchirurgie* 156, 78–84.
- WHO, 2019. Health economic assessment tool (HEAT) [Online]. Available: <https://www.heatwalkingcycling.org/#homepage> [Accessed].
- Wild, K., Woodward, A., 2019. Why are cyclists the happiest commuters? Health, pleasure and the e-bike. *J. Transport Health* 14, 100569.
- World Health Organization, 2016. Physical activity [Online]. Available from: <http://www.who.int/mediacentre/factsheets/fs385/en/> [Accessed].
- Åstrand, P.-O., Rodahl, K., 2003. *Textbook of Work Physiology: Physiological Bases of Exercise*. Human Kinetics, Champaign, IL.

Further Reading

- Cairns, S., Behrendt, F., Raffo, D., Beaumont, C., Kiefer, C., 2017. Electrically-assisted bikes: Potential impacts on travel behaviour. *Transport Res. Part A: Policy Pract.* 103, 327–342.
- Castro, A., Gaupp-Berhausen, M., Dons, E., Standaert, A., Laeremans, M., Clark, A., Anaya, E., Cole-Hunter, T., Avila-Palencia, I., Rojas-Rueda, D., Nieuwenhuijsen, M., Gerike, R., Panis, L.I., De Nazelle, A., Brand, C., Raser, E., Kahlmeier, S., Götschi, T., 2019. Physical activity of electric bicycle users compared to conventional bicycle users and non-cyclists: Insights based on health and transport data from an online survey in seven European cities. *Transport. Res. Interdisciplinary Perspect.* 100017.
- Elvik, R., Vaa, T., 2005. *The Handbook of Road Safety Measures*. Elsevier, Amsterdam; San Diego, CA.
- Fishman, E., Cherry, C., 2016. E-bikes in the mainstream: reviewing a decade of research. *Transport Rev.* 36, 72–91.
- Fyhri, A., Fearnley, N., 2015. Effects of e-bikes on bicycle use and mode share. *Transport. Res. Part D-Transport Environ.* 36, 45–52.
- Fyhri, A., Johansson, O.J., Bjørnskau, T., 2019. Gender differences in accident risk with e-bikes – survey data from Norway. *Accid. Anal. Prevent.*
- Sun, Q., Feng, T., Kemperman, A., Spahn, A., 2020. Modal shift implications of e-bike use in the Netherlands: Moving towards sustainability? *Transport. Res. Part D: Transport Environ.* 78, 102202.

INDEX

A

AASHTO Bicycle Design Guide, [2:120](#)
AASHTO green book, [4A:54](#)
Abnormal runway contact (ARC), [2:105](#)
Above ground level (AGL), [2:168](#)
Absolute product validity, [2:663](#)
Abutment, [2:593](#)
Accelerometers, [2:68](#)
Acceptable risk (AR), [2:3](#)
Acceptance, [7A:169](#)
Accessibility, [5:315](#)
Accessibility measures, [1:238](#)
Accessibility to airports, [5:211](#)
Accident, [1:64](#), [2:90](#), [2:127](#)
 causes of, [2:422](#)
 damage, [2:90](#)
 epidemiology analysis, [2:346](#)
 and fires, in tunnels, [2:714](#)
 injury, [2:90](#)
 location and time of year, [2:128](#)
 major, [2:90](#)
 serious, [2:90](#)
Accidental injuries, [7B:107](#)
Accident Investigation Board, [2:317](#)
Achieving sustainable mobility,
 typology for, [5:15](#)
 main agents, [5:16](#)
 main strategies, [5:15](#)
 sustainable mobility paths, [5:16](#)
AC motors, [3:428](#)
Acoustic emission-based detection,
 [2:462](#)
Active Railway Crossing, [4A:341](#)
Active safety systems, [6:57](#)
Active school transport, planning for,
 [6:433](#)
Active school travel in western Europe
 and the Anglosphere, [6:432](#)
Active Traffic Management (ATM),
 [4A:41](#), [4A:44](#), [4A:164](#)
Active Transit Priority, [4A:307](#), [4A:308](#)
Active transport, [7B:154](#)
 benefits of, [5:114](#)
 characteristics of, [5:141](#)
 and health benefits, [7B:170](#)

 policies and strategies for promoting,
 [5:144](#)
Activity-based models (ABM), [4B:8](#),
 [4B:188](#)
 alternate, [4B:9](#)
 hybrid, [4B:9](#)
 integrated models, [4B:10](#)
 rule-based, [4B:9](#)
 supplementary modules, [4B:10](#)
 survey data needs, [4B:10](#)
Activity-based travel demand models,
 [6:187](#)
Activity patterns, [1:132](#)
Activity space effects of electrified
 mobility tools, [5:151](#)
Actor-observer bias, [2:19](#)
Actuated Control Method (ACM),
 [4A:211](#)
Adaptive Control Strategies
 during the dispatching process,
 [4A:319](#)
 bus holding, [4A:319](#)
 bus stop skipping, [4A:320](#)
 bus substitution and insertion,
 [4A:320](#)
at the infrastructure level, [4A:317](#)
 bimodal perimeter control, [4A:319](#)
 intersections with multiple-lane
 approaches, [4A:317](#)
 intersections with single-lane
 approaches, [4A:317](#)
 perimeter control, [4A:318](#)
 perimeter control for dedicated bus
 lanes, [4A:318](#)
 presignals, [4A:317](#)
at the vehicle level, [4A:321](#)
 enroute guidance based on
 headways or timetables,
 [4A:321](#)
 enroute guidance based on signal
 timing information, [4A:321](#)
Adaptive Cruise Control (ACC), [2:174](#),
 [2:175](#), [2:181](#), [4B:153](#), [6:57](#),
 [6:307](#)
 system, [2:175](#)
 representation of, [2:174](#)

Adaptive driver assistance systems. *See*
 Advanced driver assistance
 systems (ADAS)
Adaptive Limited Look-Ahead
 Optimization of Traffic
 Signals-Decentralized
 (ALLONS-D), [4A:175](#)
Adaptive traffic control systems (ATCS),
 [4A:181](#)
Adaptive traffic signal control (ATSC),
 [4A:169](#)
Added value analysis, [1:541](#)
Additive random utility model (RUM),
 [1:128](#)
Administrative License Revocation/
 Suspension (ALR/ALS), [2:44](#)
Adoption and travel behavior change,
 [5:163](#)
Adoption motivators, [5:162](#)
Adoption rates, [5:161](#)
Advanced Air Mobility (AAM), [4A:369](#)
Advanced Airspace Operations, [4A:369](#)
Advanced Civil Speed Enforcement
 System, [2:464](#)
Advanced Cruise Control (ACC), [2:217](#)
Advanced driver assistance systems
 (ADAS), [2:202](#), [2:206](#), [2:237](#),
 [6:307](#)
Advanced driver support systems
 (ADAS), [7A:160](#)
Advanced planning and scheduling
 (APS), [3:76](#)
Advanced supply chain planning
 systems, [3:81](#)
 demand fulfillment and available to
 promise, [3:82](#)
 demand planning, [3:82](#)
 distribution planning, [3:82](#)
 hierarchical planning systems, [3:82](#)
 integral planning, [3:81](#)
 master planning, [3:82](#)
 production planning and scheduling,
 [3:82](#)
 strategic network planning, [3:82](#)
 transport planning, [3:82](#)
 true optimization, [3:81](#)

- Advanced Traveler Information System (ATIS), 4B:12, 4A:106, 289
- advisory information, 4A:107
- future directions and research needs, 4B:15
- information dissemination, 4A:106
- information generation, 4A:106
- multimodal trip planning, 4A:107
- planning for, 4B:14
 - modeling collective choices, 4B:15
 - modeling individual choices, 4B:14
- route guidance service systems, 4A:107
- technology of, 4B:12
 - predicting traffic states, 4B:14
 - reporting information, 4B:13
 - traffic monitoring, 4B:13
- traveler information, 4A:106
- trends in, 4A:107
- Advancing to a new mobility life stage, 5:217
- Adverse selection, 1:371
- Advertising as a road safety intervention, 7A:165
- Advisory speed, 2:622
- AEB Interurban, 2:620
- Aerobatic aircraft, 5:279
- Aerotropolis Phenomenon, 3:362
- Affective aspects refer to emotional
 - aspects associated with buying, 7A:83
- Africa, and the south east Asian nations (ASEAN), 1:397
- Age, 5:112
- Age-friendly transport system, 5:7
 - active travel, 5:8
 - community transport, 5:9
 - cycling, 5:8
 - governance, policy, and societal norms, 5:10
 - public buses, 5:9
 - taxis and paratransit, 5:10
 - train travel, 5:9
 - walking, 5:8
- Ageing, 7B:106
 - and driving performance, 2:234
 - and injury severity, 2:234
- Agent-based modeling simulation (ABMS), 3:82
- Agent-based models (ABM), 3:243
 - in transportation, 4B:68
- Agglomeration, 1:148
 - and coordination failure, 1:358
 - economies, 1:259, 1:356
- Aggregate demand models, 1:203
- Aggregate method of income elasticity calculation, 1:548
- Aggregate modal shares in different world regions, 5:301
- Aggregate mode choice models, 5:60
- Aggregate models, 3:233
- Aggregate willingness-to-pay, 1:193
- Aggression (AGG), 7A:121, 7A:211
- Aggressive drivers, 7A:124
- Aggressive driving, 2:17, 2:20, 7A:121, 7A:130
 - crash, and injury risk, 7A:123
 - increasing, 7A:123
 - prevalence of, 7A:122
- Aggressive speeders, 7A:130
- Agricultural commodity supply chain, 3:14
- Aimsun, 4B:69
- Air and HSR relationship, 5:224
- Airbags, 2:409
- Air Canada Flight 797 accident, 2:92
- Air Cargo, 3:355
 - accidents, 2:105
 - capacity, provision of, 2:102
 - economic development impacts, 3:357
 - empirical findings, 3:358
 - history, 3:356
 - national development, 3:358
 - regional development, 3:359
 - sectoral developments, 3:359
 - service, 2:100
 - specifics of, 5:258
 - supply chain, 3:356
- Air cargo airline business Models, 5:259
- Aircraft assembly, riveting for, 5:292
- Aircraft Belly-Hold Air Cargo Capacity, 2:102
- Aircraft belly-hold air cargo capacity, combination, 2:102
- Aircraft communications addressing and reporting system (ACARS), 2:92
- Aircraft final assembly line, 5:297
- Aircraft general description and parts, 5:273
- Aircraft manufacturing, 5:290
 - application, evolution of aluminum alloys for, 5:293
 - and assembly, rivetless structures for, 5:297
 - from blocks, 5:292
- Aircraft noise, 7B:56
- Aircraft noise standards, 7B:56
- Aircraft on Ground (AOG), 2:25
- Aircraft safety, 2:91
 - cabin safety, 2:93
 - fire, 2:92
 - high-lift systems, 2:91
 - human factors issues, 2:92
 - pilot fatigue, 2:92
 - stopping systems, 2:91
 - structural safety, 2:92
- Aircraft with fixed-wing surfaces
 - with a propulsion system, 5:273
 - without a propulsion system, 5:275
- Aircraft with mobile-wing surfaces (rotary wings), 5:276
- Air Drone, 3:375
- Airframe and Powerplant technicians (A&P), 2:27
- Air France Flight 4590, 2:95
- Air Freight, 3:369
 - contemporary airfreight services, 3:370
 - future challenges, 3:372
 - logistics industry, 3:361
 - complex and heterogeneous, 3:363
 - future challenges, 3:366
 - growth and change in, 3:361
 - marketing air freight services, 3:372
 - marketing airline airfreight services, 3:371
 - marketing airport airfreight services, 3:371
 - marketing, evolution of, 3:370
- Air-HSR modal integration
 - phenomenon, 5:224
- Airline, 2:101
 - costs and production, 1:393
 - demand, 1:392
 - management, 5:249
 - networks, 5:251
 - pricing, 1:393
 - privatization, 1:400
 - regulation, 5:263
- Airmail, 3:356
- Airplane and early air commerce, 5:193
- Airplane Collision Avoidance Systems
 - mid-air collision avoidance, 2:164
 - terrain collision awareness, 2:164
- Airplane icing, 2:352
- Air pollution, 1:56, 1:64, 5:110, 6:2, 7B:154
- Airport, 2:101
 - ancillary and city-center supplies, 5:236
 - capacity, 1:227
 - assessment, 4A:383
 - challenge, 1:228
 - costs, 5:230
 - safety, 2:94
 - airport terminal buildings, 2:94
 - aviation fuel handling, 2:95
 - bird strike, 2:95
 - deicing and anti-icing, 2:95
 - foreign object debris, 2:95
 - hangars and maintenance shops, 2:94
 - ramp operations, 2:94
 - runway incursions, 2:95
 - security
 - capabilities, 2:37
 - future, 2:39
 - terminal buildings, 2:94

- Airport Development Reference Manual (ADRM), 6:8
- Airport expansion, 5:232
- Airport integration by alliances and mergers, 5:231
- Airport management, 5:229
- governance and the objectives of, 5:229
- Airport network planning and its integration with the HSR system, 5:222
- empirical applications, 5:226
- theoretical models, 5:225
- Airport noise contours and accessibility rings, 1:108
- Airport, performance of, 5:233
- Airport pricing of infrastructure and concession businesses, 5:234
- Airport regulation, 5:234
- Airports as “natural monopolies”, 5:212
- Airports on the surrounding territory, impacts of, 5:211
- Airport Surface Detection Equipment—Model X (ASDE-X), 4A:375
- Air quality planning and policy, 6:236
- new concepts, methods and tools to provide new insights, 6:237
- active transport and planning role, 6:238
- exposome paradigm, 6:237
- modeling, 6:238
- “new normal”, opportunities and potential impacts, 6:238
- PASTA project, 6:237
- Air route development (ARD), 5:241, 5:243
- stakeholders’ engagement, 5:243
- steps of the ARP and ARD processes and their integration, 5:245
- Air route planning (ARP), 5:241
- demand-driven factors, 5:241, 5:242
- steps of the ARP and ARD processes and their integration, 5:245
- supply-driven factors, 5:242
- Air, Safe Transportation of Dangerous Goods by, 2:102
- Airspace Capacity Assessment, 4A:381
- Airspace Flow Program (AFP), 4A:366
- Airspace Operations, 4A:364
- advanced, 4A:369
- air traffic control, 4A:365
- emerging technologies, 4A:376
- unmanned aerial system traffic management, 4A:376
- unmanned aircraft systems, 4A:377
- urban air mobility, 4A:376
- integrated arrival/departure/surface metroplex traffic management, 4A:370
- integrated demand management, 4A:370
- national traffic flow management, 4A:366
- regional traffic flow management, 4A:366
- surface traffic operations, 4A:368
- terminal area operations, 4A:368
- traffic flow management, 4A:366
- Airspace Technology Demonstration-2 (ATD-2), 4A:370
- Airspace Users (AU), 4A:380
- Air traffic, 7B:38
- Air Traffic Control (ATC), 2:164, 2:353, 4A:365, 4A:380
- Air Traffic Controller (ATCO), 4A:380
- Air traffic control radar beacon system (ATCRBS) transponders, 2:164
- Air Traffic Control Tower (ATCT), 4A:368
- Air Traffic Flow and Capacity Management (ATFCM), 4A:380, 4A:387
- Air Traffic Flow Management (ATFM), 4A:380, 4A:384
- diversions, 4A:386
- ground delay, 4A:385
- holding pattern, 4A:386
- linear holding, 4A:385
- measures, 4A:384
- miles-in-trail, 4A:385
- minutes-in-trail, 4A:385
- rerouting, 4A:385
- slots, 4A:385
- Air Traffic Management (ATM), 3:376, 4A:380, 4A:365
- Air Traffic Services (ATS), 4A:380
- Air transport, 5:35, 5:40, 5:178, 6:7
- future of, 5:203
- air freight, slower but cheaper, 5:206
- beyond the barriers, 5:206
- big data, better decisions, 5:205
- drone revolution, 5:206
- improving footprint, enhancing sustainability, 5:204
- more people, more places—more integration, 5:203
- policy and the rules of the game, 5:205
- of hazardous materials, 2:305
- historical background, 5:263
- historical geographies of, 5:198
- system
- sub-systems of, 5:179
- sustainability of, 5:188
- Air vehicles, 5:269
- categories by types and roles, 5:278
- classification, 5:270
- Air vehicles (aircraft) with aerodynamic lift devices, 5:273
- Akaike’s information criterion (AIC), 4B:169
- Alcohol and drugs, 2:129
- Alcohol Consumption, 2:48
- Alcohol effects, on motor vehicle driving, 2:340
- Alcohol Ignition Interlocks, 2:48
- for all Convicted DUI Offenders, 2:45
- Alcohol Impairment, 2:40
- Alcohol Interlock Systems (AIS), 6:307
- Alcohol Monitoring, 2:47
- Alcohol Policy Information System (APIS), 2:42
- Alcohol Regulation, 2:42
- Alcohol Safety Action Program (ASAP), 2:50
- Alertness (AL), 7A:220
- Aleut International Association, 3:349
- All Ages and Abilities (AAA), 2:122
- All disasters are local, 2:240
- Pentagon, 2:241
- World Trade Center, 2:241
- All-door boarding, 5:330
- Allergy, 7B:107
- All-Purpose Incident Detection (APID) Algorithm, 4A:129
- All-terrain vehicles (ATV), 2:77
- All-way stop-controlled intersection (AWSC), 4A:251
- Alternatives for random utility, 5:60
- Amber light, 4A:178
- American Association of Automobiles Foundation for Traffic Safety (AAAFTS), 4B:88
- American Association of Motor Vehicle Administrators (AAMVA), 2:526
- American Association of Retired People (AARP), 4B:88
- American Association of State Highway and Transportation Officials (AASHTO), 2:118, 2:387, 2:623, 4B:118
- American Automobile Association (AAA), 2:393
- American Bureau of Shipping, 2:332
- American Council of Snowmobile Associations (ACSA), 2:82
- American National Standard for Multipurpose Off-Highway Utility Vehicles, 2:79
- American National Standard for Recreational Off-Highway Vehicles, 2:79
- American National Standards Institute (ANSI), 2:6
- American School Bus Council, 2:564

- American Society for Testing and Materials (ASTM), 4B:118
- American Society of Civil Engineers (ASCE), 2:146, 4B:118
- Americans with disabilities act (ADA), 1:275
- American Trucking Association, 3:232
- Ammonia (NH₃), 5:537
- Amphetamine, 2:224, 2:225
- Analysis of variance (ANOVA), 4B:169
- Analytical hierarchical process (AHP), 4B:83
- Analytical network process (ANP), 4B:84
- Analytic network process (ANP), 4A:396
- Anchorage Center (ZAN), 4A:364
- Ancillary institutions, 1:440
- Baltic Exchange, 1:441
- BIMCO, 1:441
- UNCTAD, 1:440
- Anderson-Darling (AD) test, 4B:169
- Anger and aggression in autonomous vehicles, 7A:127
- Anger and aggressive driving, strategies for reducing, 7A:126
- Anger and types of aggression, relationships between, 7A:126
- Animal Behavior, Monitoring of, 2:61
- Animal Crash, 2:54
- Awareness Among Public, 2:61
- Animal-Motor Vehicle Crash Databases, 2:58
- Animal-vehicle crashes (AVC), 2:53
- Annual average daily traffic (AADT), 2:284, 2:692, 4A:4
- Antecedents and consequences for speeding behaviour, 7A:131
- Anticorruption and humanitarian logistics, 3:141
- Anti-histamines, 2:226
- Anti-icing, 2:536
- Anti-lock braking for motorcycles, 6:307
- Anti-lock braking system (ABS), 2:448, 2:682, 6:57
- Antiskid braking systems, 2:91
- Application Programming Interface (API), 4A:289
- Applied traffic flow management (ATD-3), 4A:371
- Approach-avoidance task (AAT), 7A:50
- Approaches to risk-adjustment, 1:198
- Arc-based models, 4B:93
- Arctic, actors in, 3:349
- Arctic Athabaskan Council, 3:349
- Arctic Contaminants Action Program (ACAP), 3:349
- Arctic Council, 3:349
- Arctic geopolitics, 3:350
- Arctic Ice Regime Shipping System (AIRSS), 3:354
- Arctic marine environment, 3:350
- AIRSS, 3:354
- arctic weather, 3:352
- infrastructural limitations, 3:352
- The Polar Code, 3:353
- sea ice, 3:350
- shorter shipping routes, 3:352
- TP 13670E guidelines, 3:354
- TP14202E guidelines, 3:349
- UNCLOS, 3:354
- Arctic Monitoring and Assessment Program (AMAP), 3:349
- Arctic shipping, 3:349
- Arctic Waters Pollution Prevention Act, 3:354
- Arctic weather, 3:352
- Area Licensing Scheme (ALS), 4A:103, 4A:66, 4A:70
- Area Navigation (RNAV), 4A:380, 4A:373
- Armature voltage, 3:428
- Arrester bed, 2:600, 2:593
- Arterial Mobility, 4A:165
- Arterial Road Management, 4A:169
- concepts of, 4A:169
- future trends, 4A:176
- integrating emerging connected vehicle technologies, 4A:175
- signal coordination, 4A:170
- bandwidth, 4A:171
- complexities, 4A:171
- master clock, local clock, and green wave, 4A:170
- signal offset, 4A:171
- time-space diagram, 4A:170
- signal coordination frameworks and tools, 4A:172
- ALLONS-D, 4A:175
- MAXBAND, 4A:173
- OPAC, 4A:175
- PASSER V, 4A:173
- RHODES, 4A:174
- SCATS, 4A:174
- SCOOT, 4A:174
- SYNCHRO, 4A:175
- TRANSYT-7F, 4A:175
- UTOPIA, 4A:174
- Article 88b KVV, 2:574
- Articulated buses, 5:330
- Artificial intelligence (AI), 1:276, 2:182, 3:92, 6:167
- based methods, 1:24
- Artificially scarce goods, 1:371
- Ashford (United Kingdom), shared space in, 2:588
- Asian Freight Transport Policies, 3:32
- Asian Infrastructure Investment Bank (AIIB), 3:32, 3:502
- Asian logistics within the global economy, 3:144
- Asian logistics industry, 3:146
- Asian logistics performance, 3:145
- evolution, 3:144
- future, 3:148
- research findings, 3:145, 3:147
- third-party logistics development, 3:147
- Asset Management Rule, 4B:123
- Association of Oil Pipelines (AOPL), 3:469
- Assumption errors, 2:29
- Asthma, 7B:107
- Atmosphere, detection of incidents related to, 2:353
- Atmospheric issues in aviation, 2:93
- ice and snow, 2:93
- turbulence, 2:93
- volcanic ash, 2:93
- wind shear/microburst, 2:93
- Attention deficit hyperactivity disorder (ADHD), 2:19
- Attitudes and social norms, 2:19
- ATV fitted with a Quadbar, 2:80
- ATV Safety Institute (ASI), 2:79
- Audiences and their expectations, 7A:110
- Audio-tactile line marking (ATLM), 2:654
- Australia
- fatigue prevention in, 2:615
- intersection capacity of, 4A:250
- recreational boating in, 2:479
- death and causes of injury, 2:479
- road safety management in, 2:519
- Australia (CoastConnect), 4B:18
- Australian Civil Aviation Safety Authority, 2:30
- Australian Level Crossing Assessment Model (ALCAM), 4A:344
- Australian naturalistic driving study (ANDS), 7A:24
- Australian Rail Track Corporation (ARTC), 4A:344
- Australian Transport Safety Bureau (ATSB), 4A:342
- Austria (Go-Mobil), 4B:18
- AutoCAD, 4B:36
- Autogyro, 5:277
- Automated Driving
- history of, 2:181
- milestones in the development of, 2:181
- technology behind, 2:181
- Automated driving on highways, 2:183
- Automated driving system (ADS), 2:180, 2:207
- concept of operation, 2:207
- under development, 2:182
- Automated driving transport service (ADTS), 6:201

- Automated goods delivery vehicles, 2:182
- Automated goods transports (AGT), supply chain through, 3:409
- Automated guided vehicles (AGV), 3:79, 3:324, 3:409
- Automated shuttles and taxis, 2:182
- Automated stacking crane (ASC), 3:323, 3:328
- Automated Traffic Signal Performance Measures (ATSPM), 4A:182
- Automated transport systems
 - in closed environments, 3:408
 - in open environments, 3:409
- Automated truck platooning on highways, 2:183
- Automated vehicle (AV), 1:508, 2:180, 7A:161
 - automated driving systems under development, 2:182
 - automated Vehicles in Switzerland, Cost Structures for, 1:508
 - categorization of, 2:181
 - cost structures in an international comparison, 1:511
 - history of, 2:181
 - location, 4B:13
 - survey on people's intention to use automated vehicles, 1:511
 - technology, 2:181, 2:594, 6:198
- Automated Vehicle Identification (AVI), 4A:60, 4A:61
- Automated Vehicle Technology, 4A:182
- Automatic braking system, 2:91
- Automatic Dependent Surveillance-Broadcast (ADS-B), 4A:372
- Automatic Dependent Surveillance-Broadcast (ADS-B) transponders, 2:353
- Automatic Emergency Braking (AEB), 2:113, 2:175, 2:206
- Automatic identification and data capture (AIDC), 3:80
- Automatic identification system (AIS), 3:81
- Automatic incident detection (AID), 4A:128
 - function of, 4A:129
 - methods, 4A:129
 - data mining-based algorithms, 4A:130
 - deep learning, 4A:131
 - fuzzy logic-based algorithms, 4A:130
 - imbalance data problem, 4A:131
 - pattern recognition algorithms, 4A:129
 - statistical algorithms, 4A:129
 - time series and filtering algorithms, 4A:129
- traffic model and theoretical algorithms, 4A:130
- performance criteria for, 4A:131
- principle of, 4A:129
- Automatic Number Plate Recognition (ANPR), 1:135, 4A:103, 4A:60, 4A:61
- Automatic train operation (ATO) systems, 5:400
- Automatic Vehicle Location (AVL), 4A:221, 4A:309, 4A:321, 5:331
- Automation, 5:543
- Automation complacency, 2:340
- Automobile cities, 5:46
- Automobile inspection programs by state, 2:86
- Automobile Safety, 2:406
 - airbags, 2:409
 - child restraint systems, 2:411
 - collapsible steering columns, 2:413
 - crumple zones and safety cell, 2:413
 - head restraints, 2:410
 - inspection
 - data sources, 2:87
 - FARS Data, 2:87
 - future scope and research direction, 2:88
 - literature review, 2:85
 - NHTSA vehicle complaint data, 2:87
 - state laws and effectiveness measure, 2:86
 - vehicle inspection program, 2:86
 - laminated and tempered safety glass for windows, 2:412
 - other effective safety measures, 2:413
 - seat belts, 2:406
- Automobile sales, 6:32
- Automobiles Collision Avoidance Systems
 - ACC system, 2:175
 - blind spot detection, 2:175
 - cameras, 2:176
 - challenges, 2:177
 - cybersecurity, 2:178
 - future developments, 2:178
 - FWD, AEB and pedestrian AEB, 2:175
 - GPS, 2:176
 - history of, 2:173
 - human drivers and the effect of automation, 2:177
 - human drivers *versus* machines, 2:177
 - LDW and LK, 2:175
 - need for, 2:173
 - public perception, ethical challenges, and legal responsibility, 2:178
 - radars, lidars, and sonars, 2:176
 - rear cameras and parking assist, 2:176
 - rear cross traffic alert, 2:175
 - sensor faults and errors, 2:177
- sensors and state-awareness technologies, 2:176
- strategies of, 2:174
- system integration, 2:178
- V2V and V2X communication, 2:177
- 5G Automotive Association (5GAA), 2:184
- Autonomous cars, 7A:84
- Autonomous delivery vehicles, 6:322
- Autonomous Driving, advanced V2X applications for, 2:185
- Autonomous emergency braking (AEB), 2:618, 6:307
- Autonomous freight transport, future scenarios of, 3:409
- Autonomous ground vehicles (AGV), 3:375
- Autonomous vehicles (AV), 1:132, 1:149, 2:49, 5:323
 - adoption, 5:19
 - technologies, 5:19
- Autonomy, 7A:83
- Autonomy, in freight transport, 3:407
- Autopilot, 6:7
- Auto-regressive Integrated Moving-Average (ARIMA), 4A:129, 4A:166
- Auxiliary power unit (APU), 7A:183
- Average parking duration, 4B:113
- Average Variable Cost (AVC), 4A:76
- Average *versus* marginal cost and the impact of congestion, 1:16
- Aviation, 2:343
 - commercial, 2:90
 - corporate, 2:90
 - fuel handling, 2:95
 - general, 2:90
 - industry, 1:397
 - managing safety in, 2:96
 - military, 2:90
- Aviation and Transportation Security Act (ATSA), 2:34
- Aviation Maintenance Technicians (AMT), 2:27
- Aviation safety, 2:98
 - air cargo accidents, 2:105
 - aircraft belly-hold air cargo capacity, 2:102
 - cargo terminal operators, 2:101
 - consignees, 2:102
 - freighter aircraft air cargo hold capacity, 2:102
 - general sales agents, 2:101
 - historical air cargo safety occurrences, 2:103
 - historical trends in world air freight growth, 2:98
 - International air freight forwarders, 2:100
 - National Customs Authority, 2:101

Aviation safety (*Cont.*)
 ramp handling agent, 2:101
 safe transportation of dangerous
 goods by air, 2:102
 shippers, 2:100
 substantive safety, 2:103
 trucking firms, 2:100
 Aviation Safety Action Programs (ASAP),
 2:31
 Aviation security, 5:265
 Avoidance (AV), 7A:204
 Avoid Crashes With Animals, 2:59

B

Backhaul problem, 3:365
 BAC limits for driving, 2:42
 Balanced Scorecard (BSC), 3:16, 3:18
 Ballast water, 3:291
 Ballast Water Management Convention
 (BWMC), 3:291
 Baltic Dry Index (BDI), 3:268
 Bandwidth (BW), 4A:171, 4A:217
 Barnes Dance, 4A:352
 Barrier effect, 5:112
 Barrier effects of transportation
 infrastructure, 7B:74
 animal movement, importance of,
 7B:74
 impacts of increased mortality and
 reduced wildlife passage,
 7B:76
 roads, highways, and railroads as
 barriers, 7B:75
 Barriers hampering politicians' use of
 CBA when forming their
 judgments, 1:279
 Barriers to transport modes used by
 disabled and elderly people,
 5:87
 Barrier terminals, 2:598
 BART (bay area rapid transit) system,
 1:337
 Base freight rates, 3:249
 Basic Safety Message (BSM), 2:183
 Basic set of applications (BSA), 2:183
 Basic Transport Protocol (BTP), 2:183
 Basic Vickrey bottleneck mode, 1:146
 Battery electric vehicle (BEV), 7A:183,
 7B:148
 re-emergence and rise of, 6:67
 Bayes formula, 2:351
 Bayesian nonparametric statistics,
 4B:170
 Bayesian P values, 4B:171
 B2C supply chain, 3:12
 Behavioral demand model, 3:234
 Behavioral motivation, 7A:136
 Behavior change as a consequence of life
 course events, 7A:50

Beijing cycle expressway, 6:17
 Beijing Municipal Government (BMG),
 6:48
 Belgium (FlexiTec), 4B:19
 Belt and Road (B&R), 3:495
 Belt and Road Initiative (BRI), 3:349,
 3:436
 city clustering and transport system
 through, 3:501
 document, 3:495
 economic and transport corridors in,
 3:495
 funding for, 3:502
 transport infrastructure projects in,
 3:499
 Belytschko-Tsai type, 2:72
 Bendiness, 2:326
 Benefit-cost calculations, 1:111
 Benzodiazepines, 2:226
 Berkeley (UCB) Algorithm, 4A:129
 Bespoke chauffeur-driven services,
 3:204
 Beyond visual line-of-sight (BVLOS),
 3:377
 Bicing, 2:139
 Bicycle as a transportation option, 5:697
 Bicycle cities, 5:47
 Bicycle Collision Avoidance Systems
 environmental adaptation, 2:110
 feeling confident and secure, 2:111
 getting contact with and/or being
 localized, 2:111
 guidance, 2:111
 ITS improving the safety of cyclists,
 2:112
 methodology, 2:108
 multiple-comfort model, 2:109
 multiple-need model, 2:110
 result of safety impact assessment,
 2:113
 Bicycle freight, infrastructure for, 3:491
 Bicycle Helmets, 2:135
 Bicycle infrastructure
 and active transportation planning,
 5:699
 bicycling and safety, 2:122
 European Design approaches, 2:120
 forms of infrastructure, 2:118
 history of, 2:115
 other, 2:123
 stress classification system, 2:121
 Bicycle lanes, 2:118
 Bicycle routes, 2:120
 Bicycles
 bicycle-sharing systems, 2:139
 health impacts of cycling, 2:140
 policy recommendations, 2:142
 public BSS bicycles, 2:141
 SIN theory, 2:142
 value of cycling, 2:139

Bicycle sharing, history
 history
 coin-deposit bicycles, 6:25
 dockless and electric, 6:26
 free bicycles, 6:25
 marketing innovation, 6:25
 variations, 6:19
 access and affordances, 6:24
 actors, 6:24
 contract, 6:24
 governance, 6:25
 integration, 6:24
 operators, 6:21
 ownership, 6:21
 pricing, 6:23
 revenue, 6:23
 station-based (SBSS) schemes, 6:20
 Bicycle sharing, 6:19
 Bicycle sharing systems (BSS), 2:139,
 6:19
 occasional, 2:142
 and private bicycles, 2:140
 public, 2:141
 Bicycles (e-bikes), 5:701
 Bicycle to Vehicle communication
 (B2V), 2:112, 2:113
 Bicycling and safety, 2:122
 Bicyclist related accidents, 2:127
 Bicyclist safety, 2:545
 Bidirectional charging or vehicle-to-grid
 (V2G), 1:561
 Big Bayou Canot Bridge, 2:144
 Big data, 1:276, 4B:49, 6:1
 generated by non-PT Sources, 138
 from ICT Sources, 5:668
 mobile phones, 5:668
 social media and location-based
 services, 5:668
 for long-distance *versus* local public
 transport network planning,
 5:135
 for multimodal transport modeling,
 6:187
 for public transport planning, 5:134
 Big time chunks, 4B:99
 Bikability, 333
 Bike service, 2:277
 Bike share programs, 7B:160
 Bike sharing, 2:123, 4B:211, 7A:189,
 7B:165
 in cities, 4B:215
 designing, 4B:213
 helmets and, 7B:167, 7B:166
 multimodal urban bouquet
 4B:214
 operational management, 4B:213
 usage, 4B:214
 Bike-sharing systems (BSSs), 7B:161,
 7B:163, 7B:160
 BSS steps, 7B:161

- general framework used by health impact assessments on, 7B:163
 health benefits of BSSs, 7B:162
 twelve European BSSs included in the health impacts assessment, 7B:164
 users' profiles, 7B:166
 Bilevel optimization and network design, 4B:94
 Bi-level programming model, 4A:54
 Binding corporate rules (BCRs), 1:253
 Binomial distribution, 4B:171
 Biofuels, 7B:130
 Bird strike, 2:95
 Bivariate distribution, 4B:171
 Black carbon (BC), 7B:44
 Black ice, 2:534
 Blind spot detection (BSD), 2:181, 2:112, 2:175, 6:57
 system, 2:176
 Blitz attack, 2:159
 Blockchain Applications, in Logistics, 3:136, 3:137, 3:138
 anticorruption and humanitarian logistics, 3:141
 automation and smart contract, 3:141
 process and operations improvement, 3:140
 product provenance and traceability, 3:139
 trade finance and settlement, 3:141
 types of, 3:138
 Blockchain technology, 3:92
 Blood Alcohol Concentration (BAC), 2:520, 2:524, 2:40, 2:271, 2:612
 Boarding and alighting points, 5:137
 Boating, 2:478
 Boeing B747-400 freighter, 2:102
 Bogie, 3:426
 design, 3:431
 three-piece, 3:431
 Boiling liquid expanding vapor explosion (BLEVE), 2:305
 BOO (Build-Operate-Own) framework, 1:375
 Bootstrap methods, 4B:172
 Bordeaux tram, France, 6:32
 Border crossings and gauge changes, 3:443
 BOT (Build-Operate-Transfer) framework, 1:374
 Bottleneck
 current managements, 4A:140
 defined, 4A:134
 freeway network, 4A:138
 diverge, 4A:139
 merge, 4A:138
 weaving, 4A:139
 future management, 4A:142
 macroscopic mechanism of congestion, 4A:134
 model, 1:150, 1:157, 1:153
 modeling, 4A:136
 sag, 4A:136
 tunnel, 4A:136
 Boundary crash problem, 2:373
 Bounded acceleration (BA) capability, 4A:146
 Bounded rationality, 2:495
 Braess network with time-dependent demand, Equilibrium trajectories of, 1:604
 Braess paradox behavior, other systems that may exhibit, 1:604
 Braess paradox example topology, 1:602
 Brake system, task of, 3:426
 Braking systems, 2:91
 Brand image and the self, 7A:82
 Breakaway support, 2:593
 Break-bulk cargoes, 3:258
 Bridge parapet, 2:597, 2:593
 Bridge safety
 accidents not related to structural failures, 2:145
 analysis of fatal bridge crashes, 2:146
 fatalities on or under the US bridges, 2:147
 Maine bridge, 2:148
 nonfatal bridge crashes, 2:147
 results, 2:147
 structural failures, 2:145
 structural failures leading to vehicle occupant fatalities, 2:144
 British Broadcasting Corporation's (BBC) activities, 6:36
 British Medical Journal (BMJ), 2:6
 British Railways Board (BRB), 3:27
 British Road Safety Statement (BRSS), 2:525
 BRT/BRT-like systems, 5:328
 Buffalo CarShare, 2:150, 2:151
 Buffer measures, 4A:110, 4A:111
 Buffer time, 4A:116
 Buffer time index (BTI), 4A:116, 4A:110
 Build-develop-operate (BDO), 1:375
 Building Research Establishment Environmental Assessment Method (BREEAM), 3:20
 Build-lease-operate-transfer (BLOT), 1:375
 Build-operate-transfer (BOT), 1:375
 Build-own-operate (BOO), 1:375
 Build-own-operate-transfer (BOOT), 1:375
 Build-rent-own-transfer (BROT), 1:375
 Build-transfer-operate (BTO), 1:375
 Built-in Test Equipment (BITE), 2:25
 Bulgaria (FMS), 4B:19
 Bulk material supply chain, 3:12
 Bulk port, 3:299
 Bullwhip Effect, 3:130
 causes of, 3:131
 behavioral causes, 3:132
 delay structure, 3:131
 rational decision, 3:132
 measuring, 3:133
 anecdotal evidence, 3:133
 empirical evidence, 3:133
 macroeconomic shocks, 3:133
 taming the, 3:134
 Bump-and-rob carjacking, 2:158, 2:159
 Bundling, 3:250
 Bunker Fuel Cost, 3:337
 Burden of disease (BoD) approach, 7B:123, 7B:127
 Burden of disease in the context of transport policies, 7B:124
 Bureau of Public Roads (BPR), 4B:26, 4B:163
 Bureau of Transportation Statistics (BTS), 4B:87
 Bus, 1:245
 Bus adaptive control, future, 4A:322
 Bus cycle time and deadhead time, 5:328
 Bus holding, 4A:320
 Business aviation aircraft, typical final assembly line sequence of, 5:298
 Business-to-business (B2B) goods, 3:5
 Business-to-consumer (B2C), 3:5, 3:219
 Bus lanes along road segments, 6:255
 Bus Lane with Intermittent Priority (BLIP), 4A:311, 4A:312
 Bus priority, 6:254
 Bus priority measures, categories of, 6:255
 Bus rapid transit (BRT), 1:279, 4A:308, 6:121, 359
 Bus running time, stop time, and travel time, 5:327
 Bus sector, experience in, 1:308
 Bus service, 2:277
 Bus stop design, 6:255
 Bus stop skipping, 4A:320
 Bus substitution and insertion, 4A:320
 Bus transit operations, 5:326
 Bus transit service improvement strategies, 5:328
 Bus transit users' perception of service quality, 5:327
 Buy-build-operate (BBO), 1:375
 Bycyklen, 2:139
- ## C
- Cabin safety, 2:93
 Cadillac Cyclone XP-74, 2:173
 California Algorithms, 4A:129
 Call data records (CDRs), 6:136

- Camera Enforcement, 2:618
- Cameras, 2:176
- Canada
 - recreational boating in, 2:478
 - death and causes of injury, 2:479
 - road safety management in, 2:521
- Canada Labor Code, 3:354
- Canada Shipping Act 2001, 3:354
- Canadian Council of Motor Transport Administrators (CCMTA), 2:521, 2:522
- Canadian naturalistic driving study (CNDS), 7A:26
- Cannabis, 2:224, 2:225
- Canoeing, 2:480
- Capacity across transit modes, 5:350
- Capacity allocation, 1:382
- Capacity distribution function, 4A:145
- Capacity Drop, 4A:146, 4A:3
- Capacity Management, 4A:381
- Capacity of bandwidth (CBW), 4A:218
- Capacity planning and operational enhancements, 1:228
- Capital project planning, 5:354
- Capital stock, 1:449
 - capital stock estimates for transport infrastructure, 1:455
 - estimating initial value of capital stock, direct (synthetic) method, 1:454
- Carbon dioxide (CO₂), 1:72, 7B:44
 - emissions, 7B:22, 7B:129
 - from the transport sector, 1:73, 1:75
- Carbon dioxide (CO₂) emissions, 6:94
- Carbon emissions from road transport, policy instruments to reduce, 1:514
- Carbon fiber development for aircraft manufacturing, 5:293
- Carbon fiber reinforced plastic (CFRP), 2:63
- Carbon leakage, 1:461
- Carbon monoxide (CO), 5:267, 6:236
- Carbon Offsetting and Reduction Scheme for International Aviation (CORSIA), 5:267
- Carbon oxide (CO), 7B:44
- Carbon taxes, 1:515
- Car-free cities' implementation, prerequisites for, 7B:21
- Car-free city advocacy, early developments toward, 7B:18
- Car-free initiatives from around the world, 7B:18
- Cargo air transport aircraft (freighters and outsize cargo freight transporters), 5:283
- Cargo bike
 - construction types of, 3:490
 - technical evolution of, 3:489
- Cargo terminal operators (CTO), 2:101
- Cargo transportation, 5:607
 - bulk carrier, 5:608
 - container carrier, 5:609
 - gas carrier, 5:609
 - general cargo carrier, 5:610
 - Ro-Ro carrier, 5:611
 - tanker, 5:609
- Carjacking
 - blitz attack, 2:159
 - bump-and-rob, 2:159
 - definitional issues, 2:157
 - increased effort for, 2:161
 - increase risks for, 2:162
 - level of planning and motive, 2:158
 - method of acquisition, 2:158
 - normalcy illusion, 2:159
 - planning and motivation
 - classification for, 2:158
 - police impersonator, 2:160
 - prevalence of, 2:160
 - prevention activities, 2:161
 - reducing provocations, 2:162
 - reducing rewards for, 2:162
 - removing excuses, 2:162
 - situational crime prevention, 2:161
 - socioecological characteristics of, 2:160
 - street culture and, 2:160
 - trust violations, 2:159
 - typologies of, 158
- Carlson's model, 4A:36
- Car ownership, 1:92
- Car ownership, motives for, 5:384
- Car parking, 3:206, 6:113
- Car parking spaces, 6:2
- Car Park Occupancy Information System (COINS), 4A:289, 4A:292
- Carpool lanes, 4A:45
- Carpool service, 2:277
- Cars as symbols of success, status, and wealth, 7A:82
- Car sharing (CS), 6:142, 7A:187, 7A:188
 - background, 4B:217
 - data, 4B:217
 - demand, 4B:218
 - membership models, 4B:218
 - travel demand, 4B:219
 - further research, 4B:220
 - and the impact on new car registration, 6:225
 - empirical evidence from Italy, 6:226
 - impacts of, 6:144
 - implications for policy and planning, 6:144
 - methods, 4B:218
 - revealed data, 4B:218
 - safety, 2:151
 - stated preference, 4B:218
 - supply, 4B:220
 - surveys, 4B:218
 - travel surveys, 4B:218
 - types by, 6:143
 - business model, 6:143
 - operational mode, 6:143
- Car taxes, 1:245
- Car to Car Communication Consortium (C2C-CC), 2:185
- Car use, motives for, 5:384
- Catamaran ferry vessels, 2:290
- Categorization, 7A:108
- Cause codes, 2:332
- CAV stampede or creep, 5:310
- CBA and wider economic benefits of public transport in rural areas, 1:589
- C2C cycle, 3:115
- CCT (cool-white appearance), 7B:70
- Cellular-Based V2X (C-V2X), 2:184
- Census-based (CB) units, 2:368
- Census spatial units, 4B:24
- Central business center (CBD), 1:61
- Central business district (CBD), 4A:70, 4A:74, 4A:203, 4B:125
- Central Henan City Cluster, 3:501
- Centralized Assisted Parking Search (CAPS), 4A:289, 4A:292
- Centralized returns centers (CRC), 3:211
- Central processing unit (CPU)
 - microprocessors, 7A:14
- 21st century, transportation safety and security, 3:50
- Chain of causality, 2:7
- Changeable message signs, 4B:13
- Changing travel behaviors, 6:327
- Channel Filter Class (CFC), 2:69
- Charging connectors, 5:150
- Charles "Chuck" Poland Jr. Act, 2:566
- Chartered Institute of Highways and Transport (CHIT), 2:589
- Chartered Institute of Logistics & Transport (CILT), 3:19
- Chengdu-Chongqing City Cluster, 3:501
- Chicago Convention on International Civil Aviation (1944), 2:102
- Chicanes, 2:644
- Chicken-and-egg problem, 1:92
- Children, 2:14
 - Pedestrian Safety
 - cognitive capacity and skills, 2:417
 - countermeasures, 2:417
 - crash characteristics, 2:416
 - crash consequences, 2:416
 - crash contribution, 2:416
 - crash involvement, 2:415
 - crash recording, 2:415
- Traffic Education, 2:228

- Children's and youth's active travel as a planning concern, 6:432
- Children's independent mobility (CIM), 6:125
- factors influencing, 6:126
 - necessary to consider in planning, 6:125
 - need for change in planning perspectives, 6:129
 - social-ecological model of, 6:127
 - urban planning's role in, 6:127
- Child restraint systems (CRS), 2:411
- China, growing traffic safety challenge with e-bikes in, 2:446
- China-Pakistan Economic Corridor (CPEC), 3:495
- China Railway Corporation, 1:412
- China's Arctic Policy, 3:495
- China to Europe by railway, connecting, 3:500
- Chinese Railways, 3:431
- Chinese stakeholders, ports invested in, 3:497
- Chi-square distribution, 4B:172
- Choice experiment (CE), 3:234
- Choice experiments (CE), 1:216
- Choice model architecture and mathematical formulations, 4B:71
- case studies, 4B:75
 - ordered dependent variable models, 4B:72
 - generalized ordered logit model, 4B:75
 - ordered logit model, 4B:74
 - unordered dependent variable models, 4B:72
 - latent class multinomial logit model, 4B:73
 - mixed logit model, 4B:73
 - multinomial logit model, 4B:73
 - regret-based multinomial logit model, 4B:74
- Choice of policy instruments by a benevolent planner, 1:4
- Choice of policy instruments by the political process, 1:5
- Choosing to walk or bike, 5:335
- Chronic sleep disorders, 2:613
- Circadian rhythm, 2:612
- Citation and adjudication data, 2:560
- Citibike, 2:139
- Cities and travel modes, 5:46
- Cities, speed limits in, 2:632
- City Logistics, 1:22
- CityMobil and CityMobil2 projects, 6:200
- City-Port, 5:546
- City-scale or regional travel demand model systems, 4B:54
- City-Wide Coordinated Ramp Metering (CWCRM), 4A:2, 4A:21
- Civil Aircraft, 5:278
- Civil aircraft categories, 5:278
- Civil Aviation Authority (CAA), 2:31
- Civil Rights Act of 1964, 6:158
- Classic "aggregate demand" model, 1:127
- Classic Braess paradox, 1:602
- Classic mono-centric city model, 1:315
- extensions of, 1:317
 - asymmetric urban structure, 1:318
 - heterogeneous households in terms of income level, 1:317
 - mode choice, 1:317
 - polycentric city, 1:318
 - presence of traffic congestion, 1:317
- Class I facilities, 2:118
- Classification of Motor Vehicle Accidents, 2:6
- Class II facilities, 2:118
- Class III facilities, 2:120
- Claw back principle, 1:466
- Clean Air Act amendments of 1990, 6:175
- Clean vehicles, 5:17
- Clear of conflict", 2:167
- Clear zones, 2:596
- Climate change, 7B:41
- climate change and consequences, 7B:41
 - impact of the Covid-19 pandemic, 7B:41
- Climate change, 6:85
- Climate change and sea level rise, 3:303
- Climate change risk indexes (CCRI), 3:56
- Clinical trials supply chain, 3:15
- Closest points of approach (CPA), 2:165
- Cloud computing technology, 3:92
- Clustering of AVC, 2:54
- Coal seaborne trade, 3:264
- Coasting Trade Act, 3:354
- Cobb-Douglas cost function, 1:427
- Cobb-Douglas form, 1:256
- Cobb-Douglas function, 1:42
- Cocaine, 2:224, 2:225
- Cockpit decompression, 2:352
- Cognition, 5:114
- Cognitive bias, 2:19
- Cognitive distraction, 2:337
- Cognitive factors, 2:19
- Cognitive function in adults 7B:106
- Cognitive impairments, in older people, 2:430
- Cognitive restoration, 7B:104
- Coin-based-systems, 4A:356
- Cold ironing, 3:325
- Colectivos, 2:401
- Collaborative planning forecasting and replenishment (CPFR), 3:71
- Collaborative Trajectory Options Program (CTOP), 4A:370
- Collapsible steering columns, 2:413
- Collision avoidance and warning systems (CAWS), 2:173
- history of, 2:173
- Collision avoidance systems (CAS), 2:174
- airplane
 - mid-air collision avoidance, 2:164
 - terrain collision awareness, 2:164
 - automobiles
 - ACC system, 2:175
 - blind spot detection, 2:175
 - cameras, 2:176
 - challenges, 2:177
 - cybersecurity, 2:178
 - future developments, 2:178
 - FWD, AEB and pedestrian AEB, 2:175
 - GPS, 2:176
 - human drivers and the effect of automation, 2:177
 - human drivers *versus* machines, 2:177
 - LDW and LK, 2:175
 - need for, 2:173
 - public perception, ethical challenges, and legal responsibility, 178
 - radars, lidars, and sonars, 2:176
 - rear cameras and parking assist, 2:176
 - rear cross traffic alert, 2:175
 - sensor faults and errors, 2:177
 - sensors and state-awareness technologies, 2:176
 - strategies of, 2:174
 - system integration, 2:178
 - V2V and V2X communication, 2:177
 - strategies of, 2:174
- Collision during takeoff and landing (CTOL), 2:105
- Collision warning systems (CWS), 2:173
- passenger automobile with, 2:174
- Combination passenger airlines, 2:101
- Command-and-control instruments, 1:63
- Comma-separated value (CSV) format, 4B:165
- Commercial airline operator, principal types of, 6:9
- Commercial air transport, 6:7
- aircraft, 6:8
 - airlines, 6:9
 - airports, 6:8
 - airspace, 6:10
 - challenges, 6:10

- Commercial air transport aircraft (Airliners), 5:282
- Commercial aviation, 2:90
- Commercial Aviation Safety, 2:90
- Commercial logistics *versus* humanitarian logistics, 3:196
- Commercial motor vehicle (CMV), 2:201
- Commercial Vehicle Safety Alliance (CVSA), 201
- Common Safety Indicators (CSI), 2:468
- Communications Access for Land Mobiles (CALM), 2:183
- Communications Based Train Control (CBTC), 2:464
- Community severance, 5:112, 6:246
 - behavior of individuals and organizations, 6:249
 - causes and extent of the problem, 6:246
 - chain of potential wider effects of, 6:249
 - direct effects, 6:248
 - equity issues, 6:250
 - policies, 6:251
 - research, 6:251
 - wider effects, 6:249
 - on communities, 6:250
 - on individuals, 6:250
- Community transport, 6:314
 - challenges, 6:316
 - funding, 6:317
 - meeting demand, 6:317
 - operational issues, 6:317
 - delivery models and services offered, 6:314
 - evidence of benefits, 6:316
 - funding of services, 6:315
 - future perspective, 6:317, 6:318
 - drivers of growing demand, 6:317
 - responding to the climate emergency, 6:318
- Commuters, 1:187
- Commuting, 5:92
- Commuting and local labor markets, 1:300
 - shocks, 1:300
 - wages and commutes, 1:300
- Commuting costs, 1:156
- Commuting routes with natural features, 7B:107
- Commuting transport and the labor market, 1:100
- Comodality, 3:456
- Compact city, 5:17
- Company car, 1:152
- Company car benefits, 1:580
- Company-level cost functions, 1:42
- Comparative Farebox recovery ratios, 5:355
- Comparative risk assessment (CRA)
 - approach, 7B:125, 7B:127
 - framework applied to air pollution, 7B:126
- Compensatory effort, 2:216
- Competition, 1:383
- Competitive bidding, 1:309
- Competitive regulation, 1:368
- Competitive tendering, 1:367
- Competitive urbanism, 6:31
- Complementary and competing routes, 4B:56
- Complexity, 1:328
- Compliance with traffic legislation, 2:13
- Compulsory breath testing (CBT), 2:271
- Computable general equilibrium (CGE)
 - analysis, 1:520
 - economies of scale, 1:524
 - functional Forms, 1:522
 - model, 3:166
 - to calculate what CO2 tax, 1:73
 - structure of, 1:521
 - spatial CGE and transport, 1:523
- Computational challenges, 7B:157
- Computational effort, 2:216
- Computational Methods, 4B:2
 - deterministic, 4B:2
 - stochastic, 3
- Computer-aided dispatching (CAD)
 - system, 5:331
- Computer graphics interface (CGI), 7A:14
- Computer reservation system (CRS), 1:404, 5:250
- Computing Offsets, 4A:215
- Concept of Acceptable Risk, 2:3
- Concept of accessibility, 5:32
- Concept of cost, 1:13
- Concept of disruptive technologies, 6:310
- Concept of natural monopoly, 1:30
- Concept of service quality, 6:220
- Concept of "shared responsibility", 6:54
- Concept of transitions, 6:309
- Conceptual model for accessibility,
 - activity locations, and travel behaviour, 5:33
- Concession agreements, 1:374
- Concession period, 1:374
- Concrete barriers, 2:516
- Concurrent validity, 2:605
- Conduct disorder, 2:19
- Confidence interval (CI), 2:41, 4B:172
- Conflicts, 1:328
- Conflicts over space, 6:256
- Confrontive coping (CONF), 7A:204
- Congestion, 3:169, 4B:103, 6:206, 378
 - charge, 4A:72
 - and congestion pricing, 1:344
 - frequency of, 4A:119
 - pricing, 4A:74
 - user fee, 4A:81
 - values of time, 4A:80
 - welfare impact of, 4A:79
- Congestion charges, 1:139, 6:201, 6:206
 - effects of, 6:199
 - long-term effects, 6:201
 - short-term effects, 6:200
 - political support, 6:202
 - public support, 6:202
 - system, designing, 6:201
- Congestion externalities, 1:597
- Congestion on transport infrastructures, 134
- Congestion pricing, 4B:104, 142
- Congestion pricing and driving
 - restrictions, 517
- Congestion-related reactionary delay (CRRD), 1:378
- Congestion tax, 63
- Connected and Automated Driving, 2:184
 - advanced V2X applications, 2:185
 - V2X applications, 2:184
- Connected and automated driving systems (CADS), 5:20
- Connected and autonomous vehicles (CAV), 2:180, 4A:316, 4B:151, 6:57, 6:167, 6:217, 7B:140, 180
 - automated driving, 2:181
 - systems under development, 2:182
 - automated vehicle, 2:180
 - benefits of, 2:185
 - categorization of automated vehicles, 2:181
 - challenges and development trends, 2:186
 - demand impacts of, 4B:151
 - first- and second-order health implications, 7B:141
 - first-order impacts, 7B:140
 - impacts of, 4B:152
 - important to plan, 6:167
 - key policy and industry
 - recommendations, 6:171
 - key themes for CAVs Policy, 6:168
 - business models, 6:169
 - crisis and employment ethics, 6:168
 - customer readiness, acceptability, and trust, 6:169
 - cyber security and privacy, 6:169
 - infrastructure and land use, 6:168
 - integration, 6:168
 - legislation, 6:168
 - technology, 6:168
 - traffic congestion and travel
 - behaviour, 6:169
 - traffic safety, 6:169

- modeling traffic flow impacts of, 4B:153
- policy recommendation, 7B:143
- present benefits and risks to public health, 7B:140
- priority areas for planning and policy of connected and autonomous vehicles, 6:170
- projections about CAVs benefits and side effects, 6:167
- quantifying the impacts, 7B:143
- recommended CAV supporting policies, 7B:145
- second-order health impacts, 7B:142
- seven points of impact on transportation, 7B:141
- suggested workstream for quantifying CAVs health implications, 7B:144
- system control, 4B:154
- technology behind automated driving, 2:181
- Connected vehicle (CV), 2:180, 2:183
 - cellular-based V2X, 2:184
 - history of V2X, 2:183
 - path of DSRC, 2:183
 - technologies, 2:594, 4A:309
 - technology debate, 2:184
- Consensus, 3:136
- Consequences of Animal Crashes, 2:55
- Conservation of Arctic Flora and Fauna Working Group (CAFF), 3:349
- Conservation of Clean Air and Water in Europe (CONCAWE), 3:469
- Consignees, 2:102
- Consignor, 2:100
- Consolidation of research, 2:22
- Constant elasticity of substitution (CES), 1:522
- Constant elasticity of transformation (CET), 1:522
- Constant returns to scale (CRS), 1:521
- Construction Industry Research and Information Association (CIRIA), 3:203
- Construction supply chain, 3:13
- Construction Zones
 - future trends in work zones, 2:195
 - temporary traffic control devices, 2:190
 - temporary traffic control zone, 2:190
 - worker safety, 2:194
 - truck-mounted attenuators, 194
 - work-zone classification, 2:191
 - work-zone duration, 2:191
 - work-zone lane configuration, 2:191
 - work-zone safety and mobility impact assessment, 2:192
 - work-zone mobility impact assessment, 2:193
 - work-zone safety assessment, 2:192
 - work-zone temporary traffic control, 2:189
- Construct validity, 2:605
- Consumer Product Safety Commission (CPSC), 2:77
- Consumer protection, 5:266
- Consumer surplus, 1:64
- Consumer surplus in project evaluations, 1:239
- Consumer-to-consumer (C2C), 3:219
- Consumption Capital Asset Pricing Model, 1:198
- Contact sexual violence, 2:576
- Containerization, 3:323
- Containerization, 5:545, 5:551
- Containerization and intermodal transportation, 5:39
- Containerization and labor productivity, 5:552
- Containerization impact on port planning, development, and operations, 5:551
- Containerization's impact on maximum containership sizes, 5:552
- Containerization's impact on Ports, 5:551
- Container port, 3:299
- Container routing decision, 3:339
- Container (Liner) Shipping
 - asset management in, 3:250
 - challenges ahead, 3:255
 - companies
 - horizontal integration in, 3:252
 - operational strategies of, 3:251
 - vertical integration in, 3:254
 - demand-supply dynamics, 3:247
 - from infancy to a mature business, 3:247
- Container Terminal Automation
 - challenges, 3:324
 - future for, 3:325
 - slow pace of, 3:324
 - technologies, 3:323
- Container transshipment, 3:336
- Contemporary airfreight services, 3:370
- Contemporary airline regulation, major issues in, 5:265
- Contemporary geographies of air transport, 5:200
- Contemporary public transit governance, 5:350
- Continental United States (CONUS), 4A:364
- Contingent valuation (CV), 1:216, 3:234
- Contingent valuation method (CVM), 1:118, 1:119
- Contingent valuation studies of transportation noise, 1:109
- Continuous, intermittent, and impulsive noise, 7B:88
- Contract design, 1:302
 - assets, 1:311
 - franchising authority, 1:310
 - incentives, 1:311
 - length, 1:311
 - net or gross cost contracts, 1:310
 - size and scope, 1:310
 - treatment of staff, 1:311
 - vertical separation, 1:311
- Contracting, 1:332
- Contracting forms to buy transport infrastructure, 1:302
- Contracting forms to buy transport infrastructure, pros and cons of, 1:306
- Contraflow lanes, 4A:52
- Contrast, 2:361
- Contrast sensitivity, 2:333
- Controlled charging, 1:560
- Controlled flight into terrain (CFIT), 2:105
- Controlled observations, 7A:9
- Controller-area-network (CAN), 2:178
- Conventional motorized transport (cars, minibuses, buses, and trucks), 5:53
- Conventional transport appraisal methods of wider economic impacts (WEIs), 1:471
- Conventions or codes
 - allowable safe loadlines for ships under different conditions (LOADLINE), 1:437
 - avoidance of vessel collisions at sea (COLREGS), 1:437
 - ensuring the appropriate management of ships (ISM Code), 1:437
 - ensuring uniform standards of competence for seafarers (STCW), 437
 - reduction of marine pollution (MARPOL), 1:437
 - safety of life at sea (SOLAS), 1:437
 - security of ports (ISPS Code), 1:437
- Conway-Maxwell-Poisson model, 2:726
- Cooperative Adaptive Cruise Control (CACC), 2:185, 4B:153, 5:20
- Cooperative Awareness Message (CAM), 2:183
- Cooperative intelligent transport systems (C-ITS), 2:183
- Cooperative lane change (CLC), 2:185

- Coordinated Highways Action Response Team (CHART), 4A:124
- incident management, 4A:124
- traffic management, 4A:124
- traffic monitoring subsystem, 4A:124
- traveler information subsystem, 4A:124
- Coordinated phase, actuating, 4A:216
- Copulas, 4B:172
- Cordon schemes, 1:64
- Corona virus (COVID-19) pandemic, 4A:363
- Corporate Average Fuel Economy (CAFE) Standards, 2:296
- Corporate aviation, 2:90
- Corporate social responsibility (CSR), 2:428, 3:64
 - in logistics activities, 3:67
 - freight transport, 3:68
 - purchasing, 3:67
 - warehousing and terminals, 3:68
 - in logistics and SCM, 3:64
 - add-on or an integrated part, 3:67
 - collaboration and coordination, 3:66
 - interrelated and interdependent sustainability dimensions, 3:65
 - long-term perspective, 3:66
 - myriad of aspects, 3:65
 - numerous stakeholders, 3:66
 - voluntary basis, 3:66
- Corporatized ports, 4A:392
- Correlation analysis, 4B:5
- COSCO shipping ports, 3:498
- Cost allocation, 1:18
- Cost analysis in the railway sector, 1:425
- Cost-based *versus* incentive regulation, 5:236
 - investments, 5:237
 - pricing, 5:236
- Cost-benefit analysis (CBA), 1:114, 1:216, 1:242, 1:249, 1:278, 1:471, 6:55
 - CBA and CGE or LUTI, 1:545
 - CBA and EIA, 1:545
 - CBA and MCA, 1:545
 - CBA and Transport Models, 1:545
 - CBA as a source of bias in itself, 1:253
 - CBAs answer the right questions, 1:254
 - improving the robustness of CBAs through ex-post evaluations, 1:253
 - robustness of CBA in a future perspective, 1:254
 - for transport infrastructure investments, 1:72
- Cost calculation, 1:18
- Cost effectiveness analysis (CEA), 1:114
- Cost function, 1:15
 - of firm in industry, subadditive, 1:372
 - of shipping companies, 1:44
- Cost of Animal Crashes, 2:57
- Cost of transport time increases (CTTI), 1:321
- Cost of uncertainty (Risk), 1:196
- Cost of waiting, 1:196
- Cost overrun
 - concept of, 1:483
 - in Infrastructure Project, 1:304
 - measurement of, 1:485
 - potential ways of avoiding, 1:489
 - problems of, 1:484
 - reason for occurrence, 1:484
 - of transportation infrastructure projects, 1:483
 - worldwide, prevalence of, 1:486
- Cost-performance function, 4B:27
- Cost savings approach (CSA), 1:166
- Costs of Accidents
 - cost offsets and values of risk reduction, 2:198
 - data requirements for, 2:198
 - incidence and prevalence, 2:197
- Cost values usage for policy questions, 1:110
- CO₂ tax, 1:72, 1:74
- Countdown signals, 4A:353
- Countermeasure barriers, 7A:133
- Countermeasures to reduce speeding behaviour, 7A:132
- Countries Trade, 1:431
- Courier, express and parcel (CEP) services, 3:488
- Coventry University (United Kingdom)
 - shared space in, 2:589
- COVID-19 pandemic, 5:26, 5:130, 5:323, 5:701, 6:1, 6:123, 6:238, 6:368, 7A:191, 7B:3, 7B:96
 - impact in global development, 7B:5
- Cracking and roughness, 2:531
- Crane Technology, 3:327
 - gantry cranes, 3:328
 - quay cranes, 3:327
 - research trends, 3:333
- Crash, 2:381
 - and accident, 2:7
 - causes of, 2:256
 - contributory factors, overview of, 6:53
 - cushion, 2:600, 593
 - data, 2:88, 2:560
 - estimation of, 2:727
 - modeling, 2:726
 - underreporting and modeling issues, 2:728
 - underreporting of, 2:727
 - mechanics, 2:126
 - rate, 2:284, 2:323
- risk
 - current practice for estimating, 2:286
 - future of estimating, 2:288
- Crashes With Domestic Animals, 2:56
- Crash modification factor (CMF), 2:192, 2:322, 2:651, 6:55
- Crash Outcomes Data Evaluation System (CODES), 2:561
- Crash risk findings from naturalistic driving studies, 7A:116
- Crashworthiness, 2:63
- Creating livable and complete streets and communities, 5:337, 5:339
 - bicycle treatments, improving safety and comfort for cyclists along the street, 5:341
 - components, 5:338, 5:337
 - creating better crossings, 345
 - design principles for better crossings, 5:345
 - getting people across the street, safely and comfortably, 5:341
 - getting people along the street, safely and comfortably, 5:338
 - importance of human scale, 5:347
 - key tactical areas to achieve, 5:337
 - level of traffic stress: bicyclists need space and protection, 5:340
 - needs of bicyclists along the street, 5:339
 - providing access, connection, and convenience, 5:342
- Critical Infrastructure and Key Resources (CIKR), 3:51
- Crossbucks, 2:468
- Cross-elasticities between modes, important, reasons for, 1:201
- Cross-elasticity conventional wisdom, 1:204
- Cross-elasticity relationships, 1:202
- Crowding
 - main steps of cost estimation, 1:86
 - measuring the user cost of crowding, 1:85
 - modeling travel disutility, 1:85
 - in models of optimal transport supply, 1:87
 - physical and behavioral foundations, 1:84
 - survey-based data collection, 1:86
 - in a transport, 1:84
 - typical empirical estimates, 1:87
- Crowdshipping phenomenon, 6:323
- Crude oil seaborne trade, 3:270
- Cruise fleet, economies of scale, 5:595
- Cruise industry, 5:593
- Cruise Ports and destinations, 5:596
- Cruise Ships, 2:293

- accidents in the US waters, 2:293, 294
 - safety, 2:293
 - security, 2:294
 - sizes, 2:293
 - Cruising defined, 5:593
 - Cruising for parking, 1:160
 - Crumple zones and safety cell, 2:413
 - Crush protection devices (CPD), 2:79
 - Cultural attitudes, 1:131
 - Cumulative Curves, 4B:138
 - Curb, 2:593
 - delay, 2:432
 - extension, 2:644
 - radius reduction, 2:645
 - Current emissions from transport, 6:94
 - avoid-shift-improve framework, 6:97
 - improving progress toward sustainable mobility, 6:99
 - regime and structural change
 - Paris Agreement, 6:98
 - regime and structural change, 6:98
 - technological optimism, 6:98
 - visioning and backcasting approach, 6:96
 - Current transport systems affect human wellbeing, 6:369
 - Current transport systems affect public health, 6:368
 - matter of politics and power, 6:370
 - planning in the public domain 6:371
 - policy approach to a complex problem, 6:369
 - Customer relationship management (CRM), 3:91
 - Customer satisfaction, 6:220, 6:221
 - Customer satisfaction index (CSI), 6:221
 - Customer satisfaction surveys (CSS), 6:221
 - Customs and Border Protection (CBP), 2:292
 - Customs control, 2:101
 - Cut slope, 2:593
 - Cyber and Infrastructure Security Agency (CISA), 2:244
 - Cybersecurity, 2:178
 - Cybersecurity and Ports, 5:543
 - Cycle, 2:126
 - Cycle Hire, 2:139
 - Cycle length, 4A:214
 - Cycle logistics, development of, 3:491
 - Cycle superhighways and quiet ways of London, 5:662
 - Cycling
 - factors that influence, 5:698
 - health impacts of, 2:140
 - increasing importance, in practice and research, 1:414
 - modes, 5:53
 - policies, 6:241
 - value of, 2:139
 - Cycling as a service (CaaS) models, 6:24
 - Cycling attitudes, 7A:138
 - Cycling economics
 - costs and benefit, 1:415
 - Cycling policies, pitfalls in the evaluation of, 1:415
 - accessibility, 1:416
 - costs, 1:415
 - environmental effects, 1:416
 - health effect, 1:416
 - models, 1:415
 - safety effects, 1:417
 - wider economic effects, 1:416
 - Cycling satisfaction, 7A:139
 - extrinsic satisfaction sources, 7A:140
 - intrinsic satisfaction sources, 7A:141
 - Cycling subjective norms, 7A:138
 - Cycling to the future, 5:700
 - Cyclists and motorcyclists, hook turns for, 4A:207
 - Cypress Street Viaduct failure, 2:144
- ## D
- Dabbawallas, 3:481
 - Daily inventory of stressful events (DISE), 6:178
 - Dangerous goods, 2:102
 - Dangerous Goods Regulations (DGR), 2:103, 2:307
 - Daniel McFadden's nested logit model, 4B:126
 - Data analysis using qualitative methods, 7A:107
 - Data analytics, 4B:5
 - descriptive analytics, 4B:5
 - diagnostic analytics, 4B:5
 - predictive analytics, 4B:6
 - prescriptive analytics, 4B:7
 - technology, 3:92
 - Data challenges for policy makers, 6:137
 - complementarity between NEDF and traditional mobility data and modeling, 6:137
 - missing data and bias in NEDF, 6:137
 - skills and engagement, 6:138
 - Data collection and calculation of elasticities, 1:188
 - Data collection techniques, 7A:1
 - Data-driven machine learning, 4B:6
 - Data ecosystems, role of, 6:138
 - business models, 6:139
 - collection and access, 6:138
 - sharing, use, and privacy, 6:139
 - Data envelopment analysis (DEA), 3:319, 4A:396
 - Data fumes, 4B:49
 - Data generated by journey stage, 5:134
 - Data management, 7A:1
 - Data mining-based algorithms, 4A:130
 - Data, networks, and automation (DNA), 1:24
 - Data required for estimate burden of disease, 7B:124
 - Data types and analysis possibilities by journey stage, 5:136
 - on-board, 5:137
 - pre- and post-trip activities, 5:137
 - pre-trip planning, 5:136
 - at stops, 5:137
 - Days inventory held (DIH), 3:115
 - Days payable outstanding (DPO), 3:115
 - Days sales outstanding (DSO), 3:115
 - Daytime distribution and concentration of trips, 5:656
 - DBFO (Design-Build-Finance-Operate) framework, 1:375
 - DC traction machine, 3:428
 - Decarbonisation, 6:67
 - Decarbonizing Road Freight Transport, 3:395
 - indicators, 3:395, 3:396
 - limitations, 3:397
 - Deceleration rate to avoid a crash (DRAC), 2:665
 - Deceleration-to-safety time (DST), 2:665
 - Decentralized employment and excess commuting
 - efficient commuting patterns, 1:298
 - excess commuting productive 1:299
 - job search theory, 1:299
 - Decentralized employment and excess commuting, 1:298
 - Decentralized Environmental Notification Message (DENM), 2:183
 - Decision makers (DM), 4B:71
 - Decision-making trial and evaluation laboratory (DEMATEL) tool, 4A:396
 - Decisions on provided transport modes and the role of new technologies, 1:592
 - Decision theory, 4B:201
 - Decision tree, 4A:130
 - Dedicated all-cargo airlines, 2:101
 - Dedicated Bus Lane (DBL), 4A:311
 - Dedicated-purpose wagons, 3:431
 - Dedicated short range communication (DSRC), 1:24, 1:135, 2:183, 4A:103, 4A:60, 4A:61
 - Deduced cross-elasticities, 1:204
 - Deep belief network (DBN), 4A:131
 - Deep learning, 4A:131
 - Default maximum speed limit of 50 km/h, 2:636

- Define-Measure-Analysis-Improvement-Control (DMAIC), 3:91
- De Havilland DH106 Comet, 2:92
- Deicing, 2:95, 2:536
- Delamination, 2:530
- Delay
- assumption to use cumulative curve, 4A:259
 - defined, 4A:258
 - escalation, 4A:3
 - evaluation index, 4A:261
 - license plate matching, 4A:260
 - measure delay using cumulative curve, 4A:261
 - measuring vehicle run, 4A:260
 - probe vehicle, 4A:260
 - theoretical explanation by cumulative curve, 4A:258
- Delivery windows, 3:220
- Delta-V, 2:665
- Demand analysis in economics, 1:127
- Demand curve, 1:238
- Demand distortion, 3:130
- Demand for shipping - cargo interests, 1:440
- Demand management, 1:229
- administrative, 1:229
 - slot control or slot coordination, 1:229
 - traffic cap, 1:229
 - traffic diversion, 1:229
 - economic, 1:229
 - congestion pricing, 1:229
 - slot auction, 1:230
 - hybrid, 1:230
 - market-based capacity allocation instruments, 1:230
- Demand-responsive transit (DRT), 4B:17
- applications, 4B:18
 - Australia (CoastConnect), 4B:18
 - Austria (Go-Mobil), 4B:18
 - Belgium (FlexiTec), 4B:19
 - Bulgaria (FMS), 4B:19
 - Germany (Brgerbus), 4B:19
 - Ireland (Ring a Link & Local Link), 4B:19
 - Italy (ProntoBus), 4B:19
 - The Netherlands (RegioTaxi), 4B:19
 - Portugal (Médio Tejo), 4B:20
 - Slovenia (Sopotniki), 4B:20
 - Spain (Shotl & Castilla y León), 4B:20
 - Switzerland (Alpine Bus), 4B:20
 - United Kingdom (Suffolk Links), 4B:20
 - United States (ITNAmerica & MetroAccess), 4B:20
 - capacity, 4B:18
 - characteristics of, 4B:18
 - connectivity, 4B:18
 - evaluation, 4B:20
 - hours of operation, 4B:18
 - information and booking, 4B:18
- Demand responsive transport (DRT) systems, 6:87
- defining and characterizing, 6:88
 - delivery attributes, 6:88
 - demand responsive transport market and product position and potential, 6:90
 - development of DRT, 6:91
 - Phase 3: 2020 Onwards, 6:92
 - development of DRT, 6:91
 - evolution of DRT systems, 6:91
 - financial elements, 6:90
 - future implications, 6:92
 - organizational aspects, 6:89
- Demerit point system (DPS), 2:210, 2:267
- driver improvement measures, 2:212
 - effects of, 2:212
 - 3E model of, 2:212
 - implementation of, 2:213
 - underlining theories, 2:211
 - and variations, 2:210, 210
- Deming Plan-Do-Study-Act (PDSA) Cycle, 3:18
- Demographic drivers, 1:90
- Demographic factors, 2:18
- Density, defined, 3:467
- Department of Homeland Security (DHS), 2:243
- Department of Public Safety (DPS) Inspection and Citation Data, 2:88
- Departure time choice models, 4B:98
- background, 4B:98
 - data for empirical modeling, 4B:100
 - modeling techniques, 4B:99
 - policy evaluations, 4B:101
- Departure times and costs in the no-toll equilibrium, 1:155
- Departure times and costs in the optimal toll equilibrium, 1:156
- Departure times in the no-toll equilibrium, 1:154
- Departure times in the social optimum, 1:155
- De-polluting Commercial Vehicles in Cities, 6:283
- Depreciation, 1:453
- Deprivation, 5:112
- Deregulation, 1:366, 1:367
- Deregulation of regional airline market, 1:397
- Deregulation of regional airline markets
- Africa, 1:399
 - ASEAN, 1:399
 - China, 1:398
 - European Union, 1:398
 - India, 1:400
 - United States, 1:397
- Deregulation of the airline industry, 1:397
- Descriptive analytics, 4B:5
- Design-build-finance-operate (DBFO), 1:375
- Design-construct-manage-finance (DCMF), 1:375
- Design for Maintainability, 2:30
- Design Manual for Roads and Bridges (DMRB), 2:388
- Design of an Impact Attenuator Made of Composite Material, 2:70
- Design speed, 2:622
- Desired operating speed, 2:635
- Destination Management Organizations (DMOs), 5:244
- Detailed approach to qualitative research methods, 7A:39
- Detailed implementation plan (DIP), 3:283
- Detect and Avoid (DAA) system, 4A:377
- Determinants of trip costs, 1:151
- Deterministic computational methods, 4B:2
- Deterrence, 2:45, 211
- Deterrence mechanism, 7A:172
- certainty of punishment, 7A:172
 - general and specific deterrence, 7A:173
 - immediacy of punishment, 7A:172
 - punishment and punishment avoidance, 7A:173
 - severity of punishment, 7A:172
- Deterrence theory, 2:264, 7A:171
- criminology perspective, 7A:171
 - economic perspective, 7A:172
 - social learning perspective, 7A:172
- Detroit Metropolitan Area Transportation Study (DMATS), 4B:163
- Development of direct GHG emissions of global transport sector, 1:458
- Device-to-device (D2D) communications, 2:184
- Diagnostic analytics, 4B:5
- Diesel-powered locomotive, 3:424
- Differential speed limit (DSL), 2:622, 2:627
- Diffuse axonal injury (DAI), 2:349
- Digital Container Shipping Association (DCSA), 3:255
- Digital divide, 4B:49
- Digital footprints, 4B:49
- Digitalization, 3:223
- Digitalization, 5:543
- Digitalization and Data Sharing, 3:460

- Digital skills, 4B:49
- Digital technologies, 6:216
- Diminution of motor vehicle travel
possibility, 5:311
- Direct demand model, 4B:54
- Directed acyclic graph of randomized
and nonrandomized treatment
assignment, 1:285
- Direct rebound effect, 1:462
- Direct Torque Control (DTC), 3:428
- Dirigible airship, 5:192
- Dirty Dozen Maintenance Human
Factors, 2:29
- Disability prevalence disaggregated by
impairment for Great Britain,
7B:136
- Disabled people, crossing for, 4A:352
- Disabled travelers, 7B:135
- Disabled traveler statistics, 7B:136
- Disabled travelers' travel patterns
different to those of
nondisabled people, 7B:136
difficulties exist and persist, 7B:137
speculations about the future, 7B:138
technology to assist, 7B:138
- Disaggregate demand models, 3:233
- Disaster management and humanitarian
logistics, 3:195
- Disaster management cycle, 3:196
- Discharge capacity and ramp storage,
4A:28
- Discomfort budget, 1:84
- Discount rates in private investment,
1:196
- Discrete choice and the additive random
utility model, 1:238
- Discrete choice experiments (DCE),
1:118, 1:119
- Discrete choice models (DCM), 1:204,
1:237, 3:243, 4B:22, 4B:91
arc-based models, 4B:93
path-based models, 4B:92
- Discrete choices, 1:131, 1:128
- Discrete-event simulation (DES), 3:82
- Discrete method of income elasticity
calculation, 1:548
- Dislike of driving (DIS), 7A:204, 7A:220
- Displacement, 1:258
- Disruption, 2:719
characterizing, 2:719
coping with and mitigating, 2:721
agency and organizational
strategies, 2:721
community strategies, 2:722
individual strategies, 2:721
in transport infrastructure, causes of,
7B:36
- Disruptive technologies, 3:155
- Disruptive tool for transport
management, 6:376
- design of a new TDM strategy, 6:377
- Disruptive transportation, disruption to,
7A:191
- Distance Measuring Equipment (DME),
4A:380
- Distortions in other market, 1:65
- Distracted driving
countermeasures for, 7A:116
prevention and management of,
7A:116
- Distractions, types of, 7A:113
- Distributed ledger technology (DLT),
3:136
- Distribution, 1:20
- Distribution centers (DC), 3:21, 3:35
- Distribution from a single origin to
multiple destinations, 1:50
- Distribution of transport fuels between
transport modes and types of
transport, 1:459
- Ditch, 2:593
- Diverge bottleneck, 4A:139
- Divergent validity, 2:605
- Diversion curves, 4B:27
- 2-D mapping, 4B:38
- Dock, 2:139
- Dockland's light railway (DLR),
London, 6:31
- Dockless-bicycle-sharing programs, 6:47
- Dockless bikesharing system in
Singapore, 7B:162
- Dockless BSS (DBSS), 6:20
- Domino effect, 3:49
- Do not disturb while driving", 2:272
- Double-ended ferry, 2:291
- Downs-Thomson Paradox (DTP), 1:490
causal effects behind, 1:491
component effects, 1:493
cross-elasticities, 1:494
induced travel, 1:494
positive externalities in public
transport, 1:494
conditions really apply, 1:494
evidence of, 1:494
numerical illustration of, 1:493
- Drinking and Driving Legislation, 2:44
- Driver Advisory Systems (DAS), 4A:316
- Driver Alcohol Detection System for
Safety (DADSS), 2:49
- Driver and situation factors leading to
anger and aggression, 7A:125
- Driver and Vehicle Licensing (DVLA),
2:524
- Driver and Vehicle Standards (DVSA),
2:524
- Driver anger and appraisal tendencies,
7A:126
- Driver assistance systems (DAS), 2:178
- Driver behavior and performance
7A:59
- Driver behavior inventory (DBI),
7A:203, 7A:205
- Driver behavior monitoring, 6:428
- Driver behavior questionnaire, 7A:59
- Driver coping questionnaire, 7A:212
- Driver data, 2:560
- Driver distraction, 2:271, 2:337
- Driver distraction, 7A:113
- Driver education, 7A:119
- Driver Education Framework, 2:230
- Driver fatigue and drowsiness, 2:205
- Driver inattention, 7A:113
- Driver inattention, mechanisms of,
7A:114
- Driverless freight train operation, 3:433
- Drivers and barriers, 5:142
- Drivers at work
organizational safety climate among,
2:556
safety culture among, 2:557
- Drivers' hazard perception skill, 7A:151
- Driver state and mental workload, 2:216
driver state, 2:218
interpretation of measures, 2:219
mental workload during driving and
driver state, 2:216
in partly self-driving vehicles, 2:217
performance, 2:218
physiology, 2:219
self-reports, subjective measures,
2:218
- Driver stress inventory, 7A:204, 7A:212
- Driver training, 7A:158
- Driver visual search, 2:336
- Drive to decarbonize shipping, 5:535
- Driving aggression (AGG), 7A:204
- Driving and driver state, mental
workload, 2:216
- Driving automation technology (DAT),
3:408
- Driving behavior and skills, 7A:59
- Driving behavior questionnaire (DBQ),
7A:8
- Driving Education and Testing, 2:228
continuous education and training,
2:231
defensive and eco-driving courses,
2:232
first aid training, 2:231
problematically taught skills, 2:229
senior drivers, training for, 2:231
- Driving for work, 7A:221
- Driving performance, triggered
responses impact on, 7A:114
- Driving Regulation, 2:43
- Driving schools, 2:229
- Driving simulators, 7A:14
- Driving skill inventory, 7A:61
- Driving styles as a multidimensional
concept, 7A:65

- Driving under the influence (DUI), 2:42
 Driving while intoxicated (DWI), 2:45
 Driving-while-suspended (DWS), 2:47
 Drone, 5:22, 5:54, 5:311
 air *versus* ground drones, 3:379
 classification of, 3:374
 loading and unloading devices, 3:377
 loading and unloading infrastructure, 3:376
 logistics use cases, 3:375
 noise emissions, 3:379
 open issues, 3:380
 reach and bimodal approaches, 3:376
 regulations and their commercial implications, 3:376
 selected safety and security issues, 3:378
 Drowsiness, 2:205
 Drugs Use and the Risk of Road
 Accident, 2:221
 choice of estimator of risk, 2:222
 controlling for confounding factors, 2:222
 determining dose of drug, 2:221
 illicit and prescription drugs, 2:224
 illicit drugs, 2:225
 prescription drugs, 2:226
 publication bias, 223
 study quality assessment, 2:224
 Dry bulk sector, of shipping industry, 3:263
 coal seaborne trade, 3:264
 derived demand for, 3:263
 grains seaborne trade, 3:266
 iron ore seaborne trade, 3:263
 supply of, 3:268
 Dry ports, 3:344
 definition of, 3:344
 development matrix, 3:345
 in port development cycle, 3:345
 research agenda, 3:347
 Dual-process and dual-system models, 7A:50
 Durables supply chain, 3:14
 Durbin-Watson Test, 4B:173
 Dutch Algorithm, 4A:129
 Dynamic Algorithm, 4A:130
 Dynamic braking characteristics 3:429
 Dynamic congestion pricing, 1:150
 Dynamic message signs (DMS), 4A:332
 Dynamic ride-sharing (DRS), 5:21
 Dynamic Routes for Arrival in Weather (DRAW), 4A:371
 Dynamic Traffic Assignment (DTA), 4B:10, 4B:185
 Dynamic Traffic Flow Models, 4B:142
 car-following models, 4B:143
 numerical solution and traffic simulation, 4B:144
- E**
 E-activity, 4B:48
 E-bikes, 2:131, 2:124, 6:121, 7B:165, 7B:169
 accident risk from, 7B:172
 and cardiorespiratory fitness improvement, 7B:170
 effects on travel behavior, 7B:171
 and health, 7B:172
 intensity of physical activity from using e-bikes, 7B:170
 and micromobility, 2:131
 psychological outcomes from riding, 7B:172
 substitution effects, 7B:171
 usage, safety, and mode choice, 7A:184
 E-car, 7A:183
 Eco-approach and Departure (EAD), 4B:154
 Eco-driving, 5:712, 7A:169
 Eco (Green)-driving, 4B:153
 Eco-flying, 5:712
 Ecological research and practice, 7B:14
 Ecologic crisis, 7B:12
 E-Commerce, 3:172
 Economic activity, location, determined by, 1:356
 Economic characteristics of ports, 5:564
 Economic Deregulation, 3:26
 US freight railroads, 3:26
 Railroad Revitalization and Regulatory Reform Act, 3:26
 Staggers Rail Act of 1980, 3:26
 of US Truck Carriers, 3:26
 Household Goods Transportation Act of 1980, 3:26
 Motor Carrier Act of 1980, 3:26
 Economic driver, 1:91
 Economic evaluation of externalities, 6:193
 Economic growth and displacement, 1:259
 interregional transport investment, 1:259
 intracity transport investment, 1:259
 Economic impacts of transport improvements, 1:355
 Economic instruments for reducing carbon emissions from road and air transport, 1:459
 Economic properties of parking, 1:159
 Economic regulation, 3:24
 entry regulation, 3:25
 financial regulation, 3:25
 of ports, 5:565
 rate regulation, 3:24
 service regulation, 3:25
 of US freight railroads, 3:25
 of US truck carriers, 3:25
 Economics of urban freight transport, 1:326
 Economics theory, of road pricing, 4A:68
 Economies of scale (EOS), 1:51, 1:521
 and density, evidence from company-level cost functions, 1:44
 and scope, 1:18
 Economies of scope, 1:51
 Education, 5:256
 Effect of urban tolls and LEZ on congestion and pollution, 1:234
 LEZ, 1:234
 urban tolls, 1:234
 Efficacy of countermeasure research, 2:22
 Efficient transport facilitates trade, 1:433
 E-fulfillment, 3:220
 Egg, chicken, hen and cock, or family economics, 1:93
 Elasticities, 1:179, 1:201
 cross elasticities, 1:179
 deducing demand elasticities, 1:180
 demand elasticities, 1:179
 determining, 1:180
 diversion factors, 1:180
 elasticity value, 1:179
 evidence on diversion factors, 1:183
 evidence on elasticities of induced demand, 1:182
 evidence on rail elasticities, 1:182
 recent evidence on bus elasticities, 1:182
 recent evidence on road traffic demand elasticities, 1:181
 Elderly Driver Safety Issues
 ageing and driving performance, 2:234
 ageing and injury severity, 2:234
 aging of the world and some gender differences, 2:233
 automated vehicles, 2:238
 driving performance to crash incidence, 2:235
 giving up driving, 2:238
 licensing and medical approaches, 2:237
 road infrastructure implications, 2:236
 vehicle safety implications, 2:237
 Elderly road users, 2:15
 Electric bicycle (e-bike), 7A:183
 Electric bikes (E-bikes)
 safety challenge, in China, 2:446
 Electric cars, 1:510
 Electric locomotive, 3:424
 Electric mobility niches, emergence of, 6:64

- Electric mobility (EM) technologies, 6:64
- Electric traction systems, 3:427
- Electric vehicle (EV), 3:402, 5:147, 5:322, 6:65, 7A:84, 7B:148, 7B:154
- concept compared to the conventional vehicle concept, 5:149
- environmental and health impacts of, 7B:149
- innovation and diversification to circumvent the storage deficit, 6:65
- status and targets as of 2019 for key countries and regions, 7B:132
- usage, user experience and behavior in, 5:152
- Electric VTOL (eVTOL) aircraft, 4A:376, 5:702
- Electrodermal activity (EDA), 2:608
- Electroencephalogram (EEG), 2:216
- Electroencephalography (EEG), 2:608
- Electromobility, 7A:182
- Electronic control units (ECU), 2:174
- Electronic Fee Collection (EFC), 4A:60, 4A:61
- Electronic Flight Instruments System (EFIS), 2:351, 2:165
- Electronic observations, 7A:11
- Electronic Product Environmental Assessment Tool (EPEAT), 3:20
- Electronic Road Pricing (ERP), 4A:103, 4A:66, 4A:70
- Electronic stability control (ESC), 2:620, 6:57
- Electronic Toll Collection (ETC), 4A:60 2.0, 4A:66
- card, 4A:64
- case study in Japan, 4A:64
- components for, 4A:61
- effect on traffic on toll roads, 4A:62
- traffic management with, 4A:66
- Elevator traffic design, 5:687
- Elimination of Choice Expressing Reality (ELECTRE), 4B:85
- Embankment slopes, 2:596
- Embodied energy rebound effect, 1:174
- Emergency braking (EBR), 2:113
- Emergency logistics, drones usage in, 3:375
- Emergency Management Assistance Compact (EMAC), 2:245
- Emergency medical services (EMS), 2:317
- Emergency Operations Centers (EOC), 2:242
- Emergency Prevention, Preparedness and Response Working Group (EPPR), 3:349
- Emergency response systems
- all disasters are local, 2:240
- critical infrastructure, 2:244
- Emergency Operations Centers, 2:242
- Emergency Support Function-14, 2:244
- Emergency Support Function-1 Transportation, 2:244
- Incident Command System, 2:242
- mutual aid, 2:244
- National Incident Management System, 2:243
- National Response Framework, 2:243
- Pentagon, 2:241
- sector-specific agencies, 2:244
- transportation role in, 2:245
- World Trade Center, 2:241
- Emergency support function (ESF), 2:240, 5:128
- 14 cross sector business and infrastructure, 2:244
- 1 Transportation, 2:244
- Emergency Vehicle Priority/Preemption (EVP), 4A:221
- benefits of, 4A:221
- communication technologies for, 4A:222
- multiple preemption requests, 4A:225
- route-based strategy, 4A:225
- state of the art, 4A:225
- state of the practice, 4A:221
- systems of the real world, 4A:223
- transition out of, 4A:223
- Emergency vehicles (EV), 4A:221
- analytical contexts, 2:249
- analyzing, 2:248
- crash contributing factors, 2:251
- crash characteristics, 2:251
- driver characteristics, 2:252
- environmental characteristics, 2:252
- roadway characteristics, 2:251
- vehicle characteristics, 2:252
- crash data, 2:248
- crash statistics, 2:247
- insights, 2:253
- methodological overview, 2:249
- objective and scope, 2:248
- research framework, 2:252
- and traffic safety
- background, 2:247
- Emerging urban goods distribution trends, 6:322
- Emission and noise sources, 7B:88
- Emission control areas (ECA), 3:287, 7B:44
- Emission factors, 3:225
- Emissions Trading Systems (ETS), 3:296, 5:567
- Emotional factors, 2:20
- Emotion-focus (EF), 7A:204
- Empirical estimates, 1:9
- Empirical methods for measuring values of time, 1:69
- Employment centers, 6:2
- Employment effects of transport
- infrastructure
- econometric methodology, 1:362
- measure of accessibility, 1:361
- labor market flexibility, 1:362
- matching workers and job vacancies, 1:362
- role of distance
- transport mode choice, 1:362
- vacancies and competition for them, 1:362
- role of distance, 1:361
- results and policy implications, 1:364
- theoretical framework, 1:361
- Employment effects of transport
- infrastructure, 1:360
- Encouragement, 2:255, 2:256
- causes of crashes, 2:256
- insurance rewards, 2:259
- nudging, 2:257
- penalizing, 2:257
- rewarding and incentivizing, 2:258
- in road safety, 2:258
- safety management in companies, 2:258
- Endogeneity, sources of, 1:92
- Endogenous growth and new economic geography, 1:256, 1:259
- Energy Absorbing Deformation Mechanisms, 2:65
- Energy and Transport Planning, 6:214, 214
- Energy consumption by mode, 5:73
- air transport, 5:74
- land transport, 5:74
- maritime transport, 5:74
- Energy Efficiency Design Index (EEDI), 3:294
- Energy Efficiency Operational Indicator (EEOI), 3:294
- Energy Independence and Security Act (EISA), 2:297
- Energy Information Administration (EIA), 4B:87
- Energy-related carbon dioxide (CO₂) emissions, 6:225
- Energy Systems Integration (ESI), 3:402
- Energy tax, 1:72
- Energy transition, 5:543

- Enforcement and fines
 awareness raising/campaigning, [2:266](#)
 behavioral interventions, [2:265](#)
 demerit point system, [2:267](#)
 effective enforcement, [2:267](#)
 3E model of traffic safety, [2:263](#)
 fines, [2:266](#)
 law enforcement, [2:264](#)
 norms and traffic safety culture, [2:264](#)
 psychological learning theory, [2:264](#)
 reinforcement, [2:266](#)
 road safety work, [2:265](#)
 traffic law enforcement, [2:265](#)
- Enforcement countermeasures, [2:21](#)
- Enforcement models, [7A:173](#)
 deterrence production frontier, [7A:175](#)
 economic theory of choice under uncertainty, [7A:175](#)
 halo effect and enforcement coverage, [7A:174](#)
 halo effect, learning and violation rate, [7A:174](#)
 learning theory and enforcement scheduling, [7A:174](#)
 steady states of violation rates, [7A:173](#)
 time and distance halo, [7A:173](#)
- Enforcement of Impaired Driving Laws, [2:45](#)
- Engine-indicating and crew alerting system (EICAS), [2:92](#)
- Enhanced Automatic Train Control (E-ATC), [2:464](#)
- Enhanced ground proximity warning system (EGPWS), [2:164](#)
- Enhanced vision system (EVS), [2:353](#)
- Enhancing resilience, [7B:40](#)
 actions against flooding, [7B:41](#)
 recommended preventive actions to extreme weather events, [7B:41](#)
- Enterprise resource planning (ERP), [3:76](#)
 application menu, [3:78](#)
- Entrance metering at roundabouts, [4A:228](#)
- Environmental adaptation, [2:110](#)
- Environmental biodiversity, enriching, [7B:105](#)
- Environmental burden of childhood disease, [7B:127](#)
- Environmental countermeasures, [2:21](#)
- Environmental externalities, limitation of, [5:598](#)
- Environmental externality, [1:598](#)
- Environmental factors, [2:20](#)
- Environmental Impact Assessment (EIA), [6:150](#)
- Environmental impacts, [5:266](#)
- Environmental justice (EJ), [7B:81](#), [7B:82](#)
- Environmentally friendly operations, [4A:371](#)
- Environmental Management Schemes for Logistics and SCM, [3:20](#)
- Environmental noise management, [7B:88](#)
- Environmental outcomes, [1:582](#)
- Environmental protection, [5:603](#)
- Environmental Protection Agency (EPA), [2:296](#), [3:402](#), [4B:87](#)
- Environmental regulation, in ports, [5:566](#)
- Environmental Ship Index (ESI), [5:566](#)
- Environmental taxes, [5:255](#)
- Epidemiology, [2:269](#)
- Equipment Control Software (ECS), [3:324](#)
- Equity, [6:154](#)
- Equity as a goal of transport planning, [6:156](#)
- Equity as an impact of transport planning, [6:155](#)
- Equity considerations in transport planning, [6:154](#)
- Equivalent standard axle load (ESAL), [4A:88](#)
- Escalators, [5:690](#)
- E-scooter, [5:391](#)
 advantages for, [5:393](#)
 anticipated future impact, [5:397](#)
 and bike lanes, [5:396](#)
 challenges and opportunities, [5:394](#)
 first and last mile of, [5:394](#)
 operator's profile, [5:393](#)
 parking zones, [5:396](#)
- Essential life, [5:17](#)
- Establishment surveys, [4B:182](#)
- Estimating inter-modal cross-elasticities, [1:203](#)
- E-Tailing, [3:219](#)
 cross-border, [3:221](#)
 delivery options, [3:220](#)
 delivery windows, [3:220](#)
 e-fulfillment, [3:220](#)
 environmental impacts, [3:221](#)
 future developments in, [3:223](#)
- Ethylglucuronide (EtG), [2:48](#)
- Eurasia inhibitors, [3:449](#), [3:448](#)
- Eurasian railway network, [3:452](#)
- Eurasia rail freight
 border crossings and gauge changes, [3:443](#)
 cost and subsidies, [3:444](#)
 development and current market position, [3:437](#)
 enablers and inhibitors of, [3:436](#)
 environmental and political framework conditions, [3:443](#)
 infrastructure, [3:438](#)
 line speed *versus* travel speed, [3:443](#)
 policies and legislation, [3:444](#)
 political and macroeconomic conditions, [3:445](#)
 services, [3:454](#)
 train length, [3:443](#)
- Europe
 flash flood impact to roads in, [2:746](#)
 recreational boating in, [2:478](#)
- European Agreement, [2:614](#)
- European Aviation Safety Agency (EASA), [2:26](#), [2:30](#), [2:602](#)
- European Banking Association's Supply Chain Working Group, [3:115](#)
- European bicycle capitals, [2:115](#)
- European Cycling Federation, [2:141](#)
- European Design approaches, [2:120](#)
- European Environment Agency, [3:224](#)
- European HEMS and Air Ambulance Committee (EHAC), [2:319](#)
- European Mont Blanc Tunnel, [2:305](#)
- European Sea Ports Organization (ESPO), [5:566](#)
- European System of Accounts (ESA), [1:449](#)
- European Technology Platform called Alliance for Logistics Innovation through collaboration in Europe (ETP-ALICE), [3:458](#)
- European Union (EU), [3:402](#)
- European Union Emissions Trading Scheme, (EU ETS), [1:72](#), [75](#)
- European Union Freight Transport Policy, [3:30](#)
- European Union (EU) funds, [1:419](#)
- European White Paper on Transport in 1992, [3:280](#)
- Europe Container Terminals (ECT), [3:323](#)
- Europe-Freight Traffic, [1:410](#)
- Europe-Passenger Traffic, [1:410](#)
- Europe-The New Regulatory Bodies, [1:411](#)
- Euskadi Ta Askatasuna (ETA), [2:669](#)
- Evacuation, in humanitarian logistics, [3:197](#)
- Evacuation planning and transportation resilience, [2:276](#)
 all clear, return home, and back to normalcy, [2:278](#)
 friends and family, hotel, campsite, or public shelter, [2:277](#)
 new technologies, [2:278](#)
 resilient evacuation planning, [2:280](#)
 training and education, [2:279](#)
 walk, bike, drive, carpool, bus, or ride hailing services, [2:277](#)
- E-vehicle, [7A:183](#)
- Event characteristic and relationships to transportation systems, [5:129](#)

- Event Data Recorder (EDR), 2:348
- Event Logistics
- car parking, 3:206
 - illegal events, 3:205
 - public transport, 3:206
 - sustainability, 3:205
 - on tour perspectives, 3:204
 - traffic management, 3:202
 - transport economics and events, 3:205
 - transporting people, 3:203
 - transport logistics, 3:204
 - vehicles and their uses, 3:204
- Evidence implicating distraction as a traffic safety problem, 7A:115
- Evidence of transport investment's impact on the economy, 1:260
- Evolution of emergency and resilient transportation, 5:128
- Executive-business aircraft, 5:280
- Exhaust gas recirculation (EGR), 3:287
- Existing information platforms, 5:160
- Expect Departure Clearance Time (EDCT), 4A:370
- Expected time of arrival (ETA), 1:24
- Experimental Dynamic Tests, 2:71
- Explained sum of squares (ESS), 4B:169
- Explaining car use and willingness to change, 7A:84
- Exponential Smoothing Algorithm, 4A:129
- Ex post evaluation and the potential outcome framework for causal inference, 1:284
- Ex post evaluation under a nonignorable treatment assignment, 1:287
- difference-in differences, 1:288
 - instrumental variables, 1:287
 - regression discontinuity designs, 1:289
- Ex post evaluation via model-based adjustment for confounding, 1:285
- common support, 1:285
 - conditional independence, 1:285
 - doubly robust estimation, 1:287
 - key assumptions, 1:285
 - outcome regression, 1:286
 - propensity score methods, 1:286
 - stable unit treatment value assumption (SUTVA), 1:285
- Exposure
- annual average daily traffic, 2:284
 - current practice for estimating crash risk, 2:286
 - data, 2:88
 - defined, 2:282
 - future of estimating crash risk, 2:288
 - licensed drivers, 2:282
 - nonmotorized road users, 2:283
 - per capita, 2:282
 - registered vehicles, 2:283
 - risk and consequence, 2:132, 2:134
 - using exposure responsibly, 2:284
 - vehicle-kilometers traveled, 2:283
- Express lane, 4A:63
- Extended parallel process model [EPPM], 7A:167
- Extensions to mode choice models, 5:61
- External costs, 1:15, 1:53
- additional remarks, 1:59
 - affect, 1:56
 - and labor market interactions, 1:62
 - and land use, 1:61
 - of transportation, 1:58
 - and transport/land use interactions, 1:61
 - and travel mode choice, 1:60
- External environment, 5:252
- External factor and transit ridership, 4B:57
- External noise, 2:552
- External validity, 2:605
- Extra-logistics, drones usage in, 3:375
- Extra vulnerable road users, 2:429
- ## F
- Face validity, 2:604
- Facilitating infrastructure supply through predict and provide, 6:131
- Factors moderating driver distraction, 7A:114
- Fair Distribution, 1:274
- Fallingweight deflectometer (FWD), 4B:120
- False alarm rate (FAR), 4A:130
- Farebox recovery ratio, 1:267
- Fare card, 4A:327
- Fare Collection System (FCS), 4A:325
- Fare policies, 4A:325
- Fare structure, 4A:325
- differential structure, 4A:325
 - flat structure, 4A:325
- Fatal Analysis Reporting System (FARS), 2:9, 2:87, 2:146, 2:525, 2:368
- Fatal bridge crashes, analysis of, 2:146
- Fatality rates per population in the UN Regional Groups in 2013 and 2016, 6:52
- Fatigue, 2:319, 2:337
- causes of, 2:612
 - and crash likelihood, 2:611
 - defining, 2:612
 - diagnosing, 2:613
 - drivers most susceptible to, 2:613
 - educational initiatives, 2:615
 - engineering responses, 2:615
 - management, 2:31
 - preventing
 - Australia, 2:615
 - enforcement, 2:614
 - European Agreement, 2:614
 - United Kingdom, 2:615
 - The United States, 2:614
- preventing, 2:614
- Fatigue proneness (FP), 7A:204, 7A:221
- Fatigue Risk Management Systems (FRMS), 2:31
- Federal Aviation Administration (FAA), 2:27, 2:317, 2:602, 4A:363
- Federal Bureau of Investigation (FBI), 2:241
- Federal Communications Commission (FCC), 2:183
- Federal Emergency Management Agency (FEMA), 2:241
- Federal Express (FedEx), 2:99
- Federal Highway Administration (FHWA), 2:146, 2:243, 2:283, 2:525, 2:387, 2:488, 2:613, 2:200, 2:617, 4A:54, 4B:87, 4B:118
- Federal Motor Carrier Safety Administration (FMCSA), 2:525, 2:201, 3:382
- motor carrier safety performance data initiatives, 3:383
- Federal Motor Carrier Safety Regulations, 2:208
- Federal Motor Vehicle Safety Standards (FMVSS), 2:564
- Federal Railroad Administration (FRA), 2:468
- Federal Swedish Transport Administration, 2:523
- Federal Transit Administration (FTA), 4B:87
- Feebates, 1:516
- Fenghui GO, 3:92
- Ferry Vessels, 2:290
- accidents in the US waters, 2:291, 291
 - safety, 2:291
 - security, 2:292
 - types of, 2:290
- Fiber laminates, 5:296
- Fiber-reinforced polymer (FRP), 2:63
- Fidelity, 2:602, 603
- Field-Oriented Control (FOC), 3:428
- Fifth Assessment Report (AR5), 3:350
- Fill slope, 2:593
- Financial focus, 3:9
- Financial services providers (FSP), 3:117
- Financial supply chain management (FSCM), 3:114
- Financing and pricing of port infrastructure, 5:550
- Financing transport services in remote areas, 5:54

- Finite Difference Method (FDM), 2:348
- Finite element (FE)
modeling, 2:71
simulations, 2:64
- Fire accident rates, 2:714
- Firefighting Resources of Southern California Organized for Potential Emergencies (FIRESCOPE), 2:242
- Firms located relative to an export hub, 1:338
- First aid training, 2:231
- First attempts for alternative materials in aircraft manufacturing, 5:292
- First-best pricing, 1:95
- First born of jigs and tools in aircraft manufacturing, 5:292
- First-come-first-served (FCFS) policy, 4A:310
- First-degree price discrimination, 1:404
- First hard shoulder running scheme, 4A:42
- First-in-first-out (FIFO), 4A:139
- First Optimality Rule, 4A:77
- First-party logistics (1PL), 3:103
- Fixed *versus* variable costs, 1:13
- Fixing America's Surface Transportation (FAST) Act, 2:367, 4B:123
- Fixing America's Surface Transportation Act (FAST), 3:385
- Flashing lights, 2:468
- Fleet, 5:587
- Fleet safety culture, 2:203
- Flexible progression, 4A:219
- Flexible use of airspace (FUA), 4A:383, 4A:382
- Flight Management System (FMS), 2:92, 2:353, 2:165
- Flight Safety Foundation, 2:103
- Flight shaming, 3:366
- Flight simulator, 7A:14
- Floating car data (FCD), 4A:260
- Flow breakdown, 4A:143
capacity drop, 4A:146
characteristics of, 4A:143
bottlenecks and local disturbances, 4A:144
high-traffic load, 4A:144
necessary conditions for, 4A:143
future directions, 4A:150
long-term variations of, 4A:146
stochastic nature of, 4A:145
stop-and-go traffic, 4A:149
- Flower roundabout, 2:541
- Flow Management Positions (FMP), 4A:381
- Flown as booked, 3:370
- Focus groups, 7A:44
- Foot equivalent unit container (FEU), 3:500
- Foot ferry, 2:291
- Footprint, 2:296
- Force off, 4A:220
- Ford Model T, production of, 2:513
- Foreign object debris (FOD), 2:95
- Foresighted passengers, 5:235
- Forgiving devices and self-explaining roads, 2:654
- Formal role of CBA, 1:278
- Forms of infrastructure financing and the public-private partnerships, 1:374
- Forms of revenue subsidy payment, 6:351
- Forrester effect, 3:130
- Forward Collision Warning (FWD), 2:175
- Forward Concentration and Attention Learning (FOCAL) educational program, 7A:161
- Forward lighting systems, 2:364
- Forward logistics, 3:209
- Forward looking terrain alerts (FLTA), 2:355
- Forward progression, 4A:219
- Fossil fuel, 7B:129
driven cars, 1:462
vehicle, 1:560
- Fourth-party logistics (4PL), 3:103
- Four-wheeled ATV, 2:78
- Fragmentation *versus* harmonization of maritime regulatory framework, 5:604
- Franchising authority, 1:310
- Fratar adjustment for two-dimensional matrices, 4B:165
- Free bus pass in the United Kingdom, 6:60
- Freedom from unacceptable risk of harm, 2:3
- Freedom of choice, 5:36
- Free-floating bike system (FFBS), 4A:355
- Free-flow conditions, 2:622
- "Free flow" or satellite-based models, 1:468
- Free flow travel time (FFTT), 4A:258
- Free shipping, problem of, 3:173
- Freeway management, 4A:1
- Freeway Mobility, 4A:162
- Freeway network bottleneck, 4A:138
diverge, 4A:139
merge, 4A:138
weaving, 4A:139
- Freight, 5:410
- Freight bogie, types of, 3:431
- Freight cost components, 1:36
air *versus* sea, 1:36
key influential factors of costs, 1:38
- Freight Demand and Mode Choice, Modeling, 3:233
- behavioral demand model, 3:234
inventory-based model, 3:234
- Freighter Aircraft Air Cargo Hold Capacity, 2:102
- Freight generation, 1:9
- Freight Impacts, Reducing, 3:61
- Freight locomotives, classification of, 3:425
- Freight Mode Choice, decision factors, 3:231
- Freight multiple units (FMU), 3:420
- Freight Network Modeling
infrastructure networks and their operation, 3:158
network flow model, 3:157
new frontiers in research, 3:160
service network design, 3:159
vehicle routing, 3:160
- Freight railroad tactical planning, 5:465
- Freight Rates and Surcharges, 3:248
- Freight ton kilometers performed (FTK), 2:99
- Freight tonne-kilometers (FTK), 3:364
- Freight traffic at the world's largest airports, 5:2010, 42
- Freight traffic, management of, 5:452
focus on products suitable for railways, 5:452
incorporation of rail freight in the logistics chain, 5:453
increase of average freight speed, 5:453
issue of truck subsidiaries of rail freight companies, 5:453
- Freight Transport, 3:24
autonomy in, 3:407
behavioral research in, 3:242
decision-makers, 3:232
on environment, impact of, 3:58
GHG emissions, contribution of, 3:58
and logistics, 3:1, 3:3, 3:5
- Freight transportation, 5:22
- Freight transport models, 6:188
- Freight Transport Networks (FTN)
characteristics of, 3:53
resilience analysis, 3:56
resilience studies, 3:54
risk-based resilience of, 3:55
- Freight Transport Policy
in the European Union, United States and Asia, 3:30
Asian Freight Transport Policies, 3:32
European Union Freight Transport Policy, 3:30
United States Freight Transport Policy, 3:31
evolution of, 3:29
- Freight vehicles
emission models, 3:225

activity-based models, 3:225
 macroscopic models, 3:225
 microscopic models, 3:226
 Freight VITS, methods used to
 determine, 1:322
 Factor Costs, 1:322
 Freight wagon classification and design,
 3:430
 Frejo's model, 4A:37
 Frequency of Congestion (FOC),
 4A:110, 4A:112, 4A:119
 Friction function, 4B:37
 Friedman test, 7A:83
 Frontier function, 3:318
 Fuel cell electric vehicles (FCEVs),
 7B:148
 Fuel cells and hydrogen, 7B:130
 Fuel Economy Standards, 2:296
 mechanisms, 2:298
 safety effects
 vehicle age, 2:301
 vehicle size, 2:300
 vehicle weight, 2:298
 VMT rebound effect, 2:300
 safety effects, 2:298
 Fuel-efficiency improvements, 1:174
 Fuel system, in airplane safety threat,
 2:352
 Fuel tax, 1:139, 1:141, 1:515, 1:535,
 5:254
 contribute to reduction of vehicle
 CO₂ emissions, 1:515
 Fuel type by mode, 5:75
 Full Authority Digital Engine Control
 System (FADEC), 2:351
 Full container load (FCL), 3:238
 Full-truckload (FTL) carriers, 3:388
 Fuse plugs, 2:91
 Future of Maintenance and Inspection,
 2:32
 Futuristic Collaboration, 3:72
 Fuzzy rule based Bayesian reasoning
 (FuRBaR), 3:55
 Fuzzy Set Algorithms, 4A:130

G

Gains From Trade and Comparative
 Advantage, 1:258
 Gamification, 3:242
 Gamma distribution, 4B:174
 Gamma function, 4B:172
 Gantry crane operations, optimizing,
 3:331
 in container yards, 3:332
 at the hinterland interface, 3:332
 yard crane deployment, 3:331
 Gap acceptance, 2:432
 Gasoline price, 1:91
 Gas turbine locomotive, 3:424

Gates, 2:468
 Gate Turn Off Thyristors (GTO), 3:428
 Gaza's productivity, 1:258
 GDE self-evaluation and awareness
 skills, 7A:160
 Gender, 2:11, 5:112
 Gender and Rural Barriers of Mobility,
 7B:156
 Gender differences in children's travel,
 7A:194
 Gendered mobility, 6:173
 importance of activity models in
 improving, 6:179
 Gendered work, 6:174
 Gendered work and mobility, 6:174
 flexible schedules and
 telecommuting, 6:175
 gender pay gaps, family penalty, and
 commuting costs, 6:175
 needs of women in public transit,
 6:176
 private vehicle ownership and
 poverty, 6:178
 role of transit for female-headed
 minority households, 6:176
 telecommuting for climate action and
 family friendly policies, 6:175
 transportation policies that support
 women in poverty, 6:175
 women "opting-out" of careers or
 "pushed-out" due to policies,
 6:174
 womens's fear of victimization, 6:176
 Gender gaps, 7A:195
 Gender ratios, 5:658
 Gender split among private customers of
 various car-sharing providers
 in Europe around 2010, 5:660
 General Aggression Model, 2:18
 General aggression model for aggressive
 drivers, 7A:124
 General anger, 7A:121
 General aviation, 2:90
 General cargoes, 3:258
 General Data Protection Regulation
 (GDPR) guidance, 6:137, 138
 General driver stress scale, 7A:203
 General equilibrium (GE), 1:471
 General Estimates System (GES), 2:88,
 2:525
 Generalized cost for transport, 1:555
 Generalized cost function, 3:158
 Generalized cost of air transport, 5:212
 Generalized extreme value (GEV), 1:128
 Generalized heterogeneous data model
 (GHDM), 4B:128
 Generalized journey time (GJT), 1:180
 Generalized ordered logit model, 4B:75
 Generalized time, 1:557
 General Motor (GM), 4A:136

General Pedestrian Safety, 2:420
 causes of accidents, 2:422
 designing for, 2:425
 issues and potential solutions, 2:427
 need for, 2:420
 versus pedestrian security, 2:421
 protection zone for pedestrian, 2:424
 road user hierarchy, 2:424
 traffic devices for, 2:427
 General-purpose wagons, 3:431
 General Sales Agents (GSA), 2:101
 General Transit Feed Specification
 (GTFS), 4B:65
 3rd Generation Partnership Project
 (3GPP), 2:184
 Generic Formulations, 3:227
 optimal control, 3:229
 route and speed optimization, 3:228
 route optimization, 3:227
 speed optimization, 3:228
 Geneva Convention on Road
 Traffic, 2:2
 Geographical Information System
 (GIS), 6:237, 6:307
 Geographic distribution of AVCs, 2:54
 Geographic Information Systems (GIS),
 4B:34
 for accessibility modeling, 4B:37
 data integration and handling, 4B:34
 for four-step transportation modeling
 procedures, 4B:36
 LUTI models and, 4B:38
 new mobility modeling, 4B:40
 visualization for transportation
 modeling, 4B:38
 Geographic units, 2:368
 census-based units, 2:368
 developing new geographic units for
 macroscopic safety analysis,
 2:369
 political boundary units, 2:368
 postal units, 2:368
 traffic-based units, 2:368
 traffic analysis districts, 2:368
 traffic analysis zones, 2:368
 traffic safety analysis zones, 2:368
 Geographies of road pricing schemes,
 1:135
 Geography of cruise shipping, 5:594
 Geolocation, 1:164
 Geometric elements, of roundabouts,
 4A:232
 Geopolitics and pipelines, 3:469
 Germany (Bürgerbus), 4B:19
 Ghost driving, 2:751
 GIScience, 4B:40
 Giving up driving, 2:238
 Glare, 2:362, 7B:71
 Global distribution system (GDS),
 1:394, 1:406, 5:250, 251

- Global energy use in transport, 5:72
- Global Forum for Road Traffic Safety, 2:527
- Global harmonization, 4A:375
- Globalization, 3:144
- Global movement of goods, risk factors in the new world related to, 6:380
- general risk factors to consider, 6:381
- method for a business startup, 6:382
- tools to assist in international transportation planning for movement of goods
- project management, 6:382
- supply chain and transport process modeling, 6:384
- tools to assist in international transportation planning for movement of goods, 6:382
- transportation planning characteristics, 6:381
- Global Navigation Satellite System (GNSS), 2:353, 4A:60, 4A:61, 4A:72
- Global positioning system (GPS), 2:82, 2:174, 2:176, 4A:221, 6:136, 6:137, 7A:5, 7A:21
- Global Reporting Initiative (GRI), 3:65
- Global Shippers' Alliance (GSA), 1:440
- Global Status Report on Road Safety, 2:273
- Global stock of BE and PH cars in 2019, 6:68
- Global supply chain, 3:13
- Global transportation system, emergence of, 5:38
- Global warming, 1:457
- Goal programming, 4B:83
- Go/No-go Association Task (GNAT), 7A:50
- Good governance, 7B:4
- Good or expected progress of trips, 2:109
- Goods transport, 5:35
- Goods travel by air, 5:259
- Gotthard Tunnel, 2:305
- Governance for sustainable mobility policy, 6:133
- Governors Highway Safety Association (GHSA), 2:526, 2:382
- GPS technologies, 1:160, 1:518
- Graduated Driver Licensing (GDL), 2:230
- laws, 2:43
- for young drivers, 2:43
- Grafted decision tree, 4A:130
- Grains seaborne trade, 3:266
- Granger Causality, 3:359
- Graphical user interface (GUI) 4A:22
- Graphics processing unit (GPU) chips, 7A:14
- Gravity equations and discrete choice, 1:299
- amenities, 1:300
- logit model, 1:299
- Greater London Council, 4A:69
- Greater Mekong Subregion Economic Cooperation (GMSEC), 3:496
- Great Stone industrial park, 3:500
- Green award (GA), 5:566
- Green government, 5:16
- Green House Gas (GHG), 4A:45, 7B:44, 7B:129, 7B:148
- Greenhouse gas (GHG), 3:58, 3:287, 3:402
- Green House Gas (GHG) emissions, 1:72, 1:457, 1:496, 5:76, 5:110, 5:266, 6:67, 6:70, 6:225, 6:238
- air transport, 5:77
- maritime transport, 5:78
- road transport, 5:77
- Green light, 4A:178
- Green Light Optimal Speed Advisory (GLOSA), 2:184
- Green network that appears with the implementation of superblocks, 7B:29, 7B:30
- Green ports and automation, 3:325
- Green purchaser, 5:16
- Green replacement, 7B:107
- Green Routing, 3:224
- planning for, 3:226
- Greenspace, 7B:107
- Green time, 4A:214
- Green wave, 4A:169, 170
- Grenoble effect, 6:36
- Gross domestic product (GDP), 1:181, 1:256, 1:258, 1:392, 3:395
- GDP growth, 1:91
- GDP in valuing agglomeration, 1:473
- Gross fixed capital formation (GFCF), 1:454
- Gross vehicle weight rating (GVWR), 2:564
- Ground Drone, 3:375
- Ground proximity warning system (GPWS), 2:92, 2:164
- Growth of mobility on demand (MOD), 5:157
- Gruenspecht effect, 2:296
- Guardrail designs, 2:513
- Gwich' in Council International, 3:349
- H**
- Habit formation loop, 7A:137
- Habits, 7A:49
- habit measurement, 7A:56
- models of habit, 7A:55
- Habitual behavior, 7A:54
- Hackney cabs, 1:566
- Haddon Matrix Model, 2:346
- Haddon's matrix, 1:477
- Hägerstrand's theory, 4B:50
- Haito Goods, 3:92
- Harassment at bus rapid transit (BRT) stations, 7A:196
- Harbin-Changchun Metropolitan Area Cluster, 3:501
- Hard shoulder running, 4A:41
- applications, 4A:41
- impacts on traffic flow and road safety, 4A:43
- Hawk crossings, 4A:346
- Hazard monitoring (HM), 7A:204, 7A:220
- Hazardous Materials Regulations (HMR), 2:103
- Hazardous Materials Transport, 2:304
- challenges, 2:308
- concerning risk assessment, 2:308
- energy transition, 2:309
- low probability high impact, 2:308
- modes of, 2:304
- air, 2:305
- rail, 2:305
- road, 2:305
- water, 2:305
- rules and regulations, 2:306
- arrangements per transport mode, 2:307
- general arrangements, 2:306
- safety culture and human factors, 2:309
- terrorism, 2:309
- Hazard perception change across the life span, 7A:153
- Hazard perception, measured using a computer-based hazard perception test, 7A:151
- Hazard perception skill, improvement, 7A:153
- Hazard perception testing and training, 7A:155
- impact on crash risk of, 7A:155
- Hazard, types, sources effecting transportation, 5:128
- Head-on Crashes
- definition of, 2:311
- mitigation of, 2:314
- physics of, 2:311
- road conditions, 2:311
- severity of, 2:312
- traffic conditions, 2:312
- prevention of, 2:313
- infrastructure solutions, 2:313
- vehicle technology solutions, 2:314
- Head restraints, 2:410

- Heads-up display (HUD), 2:92
- Head-up display (HUD), 6:57
- Health, 7B:13
 - interrelated variables that influence health, 7B:13
- Health-adjusted life expectancies (HALE), 7B:123
- Health and development, 7B:4
- Health benefits, 7B:105
- Health impacts from noise pollution, 7B:54
- Health implications of a changing climate, 7B:98
- Health literacy, 7B:4
- Health loss paradox, 2:136
- Health promotion, 7B:4
- Health risks, 7B:106
- Healthy cities, 7B:4
- Healthy Life Year (HeaLY), 7B:123
- Heart rate variability (HRV), 2:608, 2:613
- Heavier-than-air aircraft (aerodynes), 5:272
- Heavy fuel oil (HFO), 3:286, 7B:44
- Heavy goods vehicles (HGV), 2:200
- Hedonic price and contingent valuation estimates, benefit transfers, 1:110
- Hedonic property value studies of transportation noise, 1:107
- Hegyi's model, 4A:37
- Helicopter emergency medical services (HEMS), 2:316
- Helicopter general description and parts, 5:276
- Helicopters, in emergency medical response, 2:316
 - future challenges of, 2:320
 - history, 2:316
 - safety record of, 2:317
 - balancing flight operations decision making, 2:319
 - developing technologies, 2:318
 - fatigue, 2:319
 - pushing the limits, 2:317
 - technical limitations and improvements, 2:318
- Helicopter (rotorcraft), 5:276
- Helmets, nonuse of, 2:445
- HERO-LIVE modules, 4A:22
 - activation/deactivation, 4A:23
 - ALINEA, 4A:23
 - critical occupancy, 4A:23
 - demand estimation, 4A:23
 - final ramp flow specification module, 4A:24
 - HERO coordinated ramp operation, 4A:24
 - implementation module, 4A:24
 - minimum queue control module, 4A:24
 - queue control module, 4A:24
 - queue estimation module, 4A:23
 - queue override module, 4A:24
 - waiting time
 - control module, 4A:24
 - estimation module, 4A:23
- HERO-LIVE ramp metering control algorithms, 4A:21
- Heterogeneity, 1:92, 3:237
- Heterogeneity in values of time and preferred arrival time, 1:148
- Heteroscedasticity, 4B:174
- Heuristic methods, in humanitarian logistics, 3:198
- Hierarchy of reverse logistics, 3:222
- High-beam headlamps, 2:364
- High-carbon transport systems (HCTS), 7B:96
 - health implications of, 7B:97
 - links between climate-change and, 7B:97
- High-income countries (HIC), 2:269
 - road traffic crashes, 2:270
 - behavioral risk factors, 2:270
- High-intensity Activated crossWalk (HAWK), 2:711
- High intensity activated crosswalk (Hawk crossing), 4A:351
- High-lift systems, 2:91
- Highly automatic vehicles (HAV), 2:148
- High Occupancy Algorithm (HIOCC), 4A:129
- High occupancy toll (HOT) lane, 1:64, 1:137, 4A:45
 - lane facilities in the United States, 4A:47
- High occupancy vehicle (HOV), 4B:98
- High-occupancy vehicle (HOV) lane, 1:137, 1:517, 4A:45
 - benefits and costs, 4A:50
 - classifications of, 4A:46
 - design, 4A:48
 - enforcement, 4A:50
 - implementation, 4A:48
 - lane facilities in the United States, 4A:47
 - objective of, 4A:45
 - overhead and roadside signs, 4A:48
 - past and current practices, 4A:46
 - pavement marking, 4A:49
 - performance monitoring, 4A:49
 - planning, 4A:48
 - types of, 4A:46
- High-speed crafts (HSC), 2:290
- High speed rail (HSR), 5:26, 6:261
 - extension of HSR throughout the world, 6:261
 - and improving cities' accessibility, 6:262
- infrastructure, 1:419
 - benefits of HSR, 1:420
 - cost of a HSR Line, 1:420
 - final assessment, 1:422
 - role in attracting businesses and tourists, 6:263
 - and urban projects, 6:263
- High-tech proposals, 5:309
- High-traffic load, 4A:144
- High-value transport intensive goods, 1:434
- High visibility enforcement programs (HVEs), 7A:11
- Highway advisory radio, 4B:13
- Highway Capacity Manual (HCM), 4A:143, 4A:134, 4A:247
- Highway Design and Operational Practices Related to Highway Safety, 2:513
- Highway-rail grade crossings (HRGX), 2:466, 2:751
- Highways, 1:64
- Highway Safety Improvement Program (HSIP), 2:525, 2:526
- Highway suicides, 2:658
 - influencing factors, 2:659
 - prevention, 2:659
 - statistics/incidence, 2:658
 - types of events, 2:659
- Highway transportation, 1:476
- Hire carriers, 3:232
- Historical Air Cargo Safety Occurrences, 2:103
- H1N1 flu virus, 5:267
- Hohhot, Bator, Erdos, and Yulin City Cluster, 3:501
- Home deliveries
 - within the broader urban logistics framework, 6:416
 - factors influencing home deliveries
 - economic factors, 6:414
 - environmental factors, 6:415
 - legal and regulatory factors, 6:415
 - political factors, 6:413
 - social factors, 6:414
 - technological factors, 6:415
 - factors influencing home deliveries, 6:413
 - measures for different stages of development with relevancy for, 6:416
 - and their impact on planning and policy, 6:413
 - typology of, 6:416
- Homeland Security Presidential Directive 5 (HSPD-5) 2:240
- Homo sapiens, 7B:13

- Hong Kong, electronic road pricing in, 4A:103
- Hook turns, 4A:203
 advantages of, 4A:209
 background, 4A:203
 for cyclists and motorcyclists, 4A:207
 impact on traffic safety, 4A:208
 limitations and disadvantages, 4A:211
 in Melbourne, 4A:205
 traffic management improvement, 4A:210
- Horizontal and vertical geometry, 2:322
 evaluating speeds, 2:329
 future research, 2:329
 safety impacts of horizontal curves, 2:323
 bendiness, 2:326
 curve radius, 2:323
 superelevation, 2:325
 transition curves, 2:324
 safety impacts of vertical alignment, 2:327
 grade, 2:327
 horizontal and vertical alignments, 2:328
- Horsepower limit, 3:295
- Hotbox detector (HBD), 4A:344
- Hotspots, 2:54
- Household Goods Transportation Act of 1980, 3:26
- Household surveys, 4B:181
- Household travel survey (HTS), 4B:10
- Housing prices, 1:162
- Hovercraft ferry vessels, 2:290
- Hub-and-spoke route structure, 1:50
- Hub and spoke systems, 5:554
- Hub and spoke systems and feeder networks, 5:555
- Human behavior in traffic, 7A:46
- Human body tolerance to physical force, 6:56
- Human drivers
 and the effect of automation, 2:177
versus machines, 2:177
- Human ecology, 7B:10, 7B:11
 holistic and systemic framework, 7B:12
- Human error in maintenance, 2:28
- Humane speeds, need for, 5:334
- Human factor
 definition of, 2:332
 in maintenance, 2:28
 in security and safety, 3:47
- Human Factors, in Transportation, 2:331
 cognitive characteristics and limitations, 2:335
 driver visual search, 2:336
 expectancy, 2:335
- information processing and attention, 2:335
 useful field of view, 2:336
- incidents and collisions, causes of, 2:332
- new developments in, 2:340
 automation, 2:340
 expert systems, 2:342
 feedback to operators, 2:342
 human factors guidelines, 2:341
 legalization of marijuana, 2:341
 naturalistic studies, 2:341
 tools that inform safety countermeasures, 2:342
 tools to inform planning, 2:342
- stressors, 2:337
 age, 2:339
 alcohol, 2:340
 distraction, 2:337
 fatigue, 2:337
 inexperience, 2:339
 marijuana, 2:340
 medical conditions, 2:339
- HUMANIST project, 2:109
- Humanitarian Logistics
 commercial logistics *versus*, 3:196
 and disaster management, 3:195
 disaster management cycle, 3:196
 evacuation in, 3:197
 exact and heuristic methods in, 3:198
 last mile distribution in, 3:198
 multicriteria optimization for, 3:199
 and SCM, 3:21
- Humanitarian supply chain, 3:13
- Humanitarian Transportation, 3:190
 chain, 3:192
 modal choice, 3:191
 new developments in, 3:192
 You Never Walk Alone, 3:191
- Hundred million vehicle miles (HMVM), 2:148, 2:627
- Hybrid bike-sharing system (HBSS), 4A:355
- Hybrid electric vehicle (HEV), 7A:183
- Hybrid locomotives, 3:424
- Hydraulic capsule pipelines in the 20th century, 5:650
- Hydrocarbons (HCs), 6:236
- Hydrofoil ferry vessels, 2:290
- Hydrogen, 5:537
- Hydrogen fuel cell (FC), 6:66
- Hyperconnected City Logistics (HCL), 1:23
- Illicit drugs, 2:225
- Illustrative cross-elasticities, 1:205
- Illustrative cross-elasticities implied by meta-model, 1:206
- Illustrative cross-elasticities of rail demand with respect to fuel cost, 1:207
- Impact Attenuator Definition, 2:67, 2:70
- Impact of transport investments on regional convergence, 5:4
- Impact on users of motorized modes of transport, 7B:66
- Impaired-Driving Laws, 2:44
- Imperfect pricing instruments, 1:96
- Implications and Consequences of Setting a Predefined AR Level for Road Traffic Safety, 2:3
- Implications for road safety going forward, 2:15
- Importance-performance analysis, 3:317
- Impression management, 7A:4
- Improved Documentation and Procedures, 2:30
- Incentive countermeasures, 2:21
- Incentives, 1:332, 1:335
 credibility, 1:334
 monetizing, 1:333
 monitoring operator's performance, 1:334
 objectives and targets for, 1:333
- Incident Action Plan (IAP), 2:242
- Incident clearance time, 4A:123
- Incident Commander (IC), 2:242
- Incident Command Post (ICP), 2:242
- Incident Command System (ICS), 2:241, 2:243, 2:242, 4A:125
 functions of, 2:242
- Incident detection, 4A:130
- Incident detection systems, airplanes
 atmosphere, detection of incidents related to, 2:353
 flight management system, 2:353
 runway incursion avoidance, 2:353
 terrain avoidance and warning system, 2:355
 traffic alert and collision avoidance system, 2:354
- Incident management, 4A:1
- Incidents and collisions, causes of, 2:332
- Inclement weather, 4A:159
- Inclusive transport, 7B:155
- Inclusive value, 1:129
- Income elasticity calculation, 1:548
- Income elasticity of air travel demand, 1:549
- Income elasticity of domestic air travel demand in China, 1:549
- Income elasticity of international air travel demand in China, 1:549

- Income or expenditures as a measure of economic status, 1:140
- Incremental Train Control System (ITCS), 2:464
- Independence from irrelevant alternatives (IIA), 1:129
- Independence of Irrelevant Alternatives (IIA), 4B:23
- In-Depth Traffic Accident Investigation, 2:346
- accident epidemiology analysis, 2:346
 - crash stage, 2:346
 - injury biomechanics investigation methods, 2:348
 - injury mechanisms and criterion, 2:349
 - postcrash stage, 2:346
 - precrash stage, 2:346
 - traffic accident reconstruction, 2:348
- India, intersection capacity of, 4A:249
- Indirect costs, 2:196
- Indirect method by modeling approaches (perpetual inventory method), 1:450
- Indirect method by modeling approaches (perpetual inventory method)
- basic idea, 1:450
 - gross capital stock and retirement distributions, 1:450
 - gamma distribution, 1:452
 - lognormal distribution, 1:451
 - normal distribution, 1:450
 - polynomial retirement function, 1:453
 - Weibull distributions, 1:452
 - Winfrey distributions, 1:451
- Individual factors, age, and gender, 2:129
- Individual rail vehicle classification, 5:491, 5:494
- elementary classification, 5:491
 - engineering vehicles, 5:494
 - power vehicles, 5:492
 - secondary classification, 5:491
 - trailer freight vehicles, 5:493
 - trailer passenger vehicles, 5:493, 5:495
- Indonesia, intersection capacity of, 4A:250
- Indoor transportation, 5:685
- advances in motor and variable speed drive technologies, 5:690
 - elevator traffic design, 5:687
 - escalators, 5:690
 - future developments, 5:693
 - intelligent and sophisticated control systems, 5:691
 - optimized exploitation of shaft space in buildings, 5:693
 - passenger conveyors, 5:690
 - space planning, 5:688
 - traction elevator, 5:685
- Induced private investment and land-use change, 1:357
- Induced traffic, 1:491
- Induced Travel, 4B:43
- case study-based analyses, 4B:44
 - econometric method, 4B:43
 - empirical findings, 4B:44
 - outstanding issues, 4B:45
 - transport model-based analyses, 4B:44
- Inductive logic programming (ILP), 4A:130
- Industrial zones to logistics zones, 3:36
- Industry sector, establishments by, 1:8
- Inequalities in transport modes and health, 5:114
- Inequality and Traffic Safety within country differences, 2:358
- future-impact of increasing autonomy, 2:358
 - international differences, 2:358
- Information and Communication Technology (ICT), 1:91, 3:98, 4A:60, 4B:46, 4B:226, 5:34, 5:35, 5:165, 6:184
- and accessibility, 5:34
 - and accessibility components, relationships between, 5:34
 - generation, 4B:227
 - modification, 4B:227
 - as modifiers of existing transport modes, 5:167
 - as a new transport mode, 5:165
 - for rail freight, 3:418
 - related activity concepts, 4B:48
 - substitution, 4B:226
 - and transport modes, 5:165
 - in the transport sector, 5:54
- Information for Drivers, 2:59
- Information frictions, 1:161
- Information programs, 1:517
- Information technology systems, 1:164
- Infrastructure, 2:127
- Infrastructure costs, 1:14
- Infrastructure-to-vehicle (I2V), 4A:61
- Injuries sustained by bicyclists, 2:130
- safety equipment, 2:130
- Injury, 5:111
- Injury surveillance data, 2:561
- Inland clearance depots (ICD), 3:344
- Inland Terminals (Dry Ports), 5:548
- Innovative supply chains, 3:14
- Innovative technology-based parking management solutions, 1:164
- Input output models (I/O), 1:541
- Inputs for public transport provision capital cost, 1:26
- cost of energy, 1:27
 - labor, 1:26
 - operation cost, 1:27
- Inputs for public transport provision, 1:26
- INRIX Global Traffic Scorecard indicator, 1:535
- Institute for Road Safety Research (SWOV), 2:522
- Institute of Transportation Engineers (ITE), 2:642, 4A:54
- Institutional complexities, 6:257
- Institutional weakness *vs.* growing scope for new policies, 6:121
- In-store *versus* On-line Shopping, 5:99
- Instrumental variables (IV), 4B:44
- Instrument Flight Rules (IFR), 4A:364
- Instrument flight rules (IFR), 2:319
- Instrument landing system (ILS), 2:353
- Insulated gate bipolar transistors (IGBT), 3:428
- Insurance, 2:151
- background of, 2:150
 - rewards, 2:259
- Insurance Institute for Highway Safety (IIHS), 2:152
- Integrated Choice and Latent Variable (ICLV) model, 7A:102
- model applications, 7A:104
 - model framework, 7A:104
- Integrated demand management (IDM), 4A:370
- Integrated Freeway and Arterial 4A:167
- Integrates arrival, departure and surface (IADS), 4A:370
- Integration of methods in observational field studies, 7A:11
- Integrators, 2:101
- Intelligent algorithms, 4A:294
- Intelligent and informational parking management, 4A:286
- Intelligent Parking Information System (IPIS), 4A:289
- Intelligent speed adaptation (ISA), 6:307
- Intelligent Speed Adaptation (ISA), 2:109, 2:271, 2:619
- Intelligent Transportation Systems (ITS), 2:108, 2:474, 2:654, 2:755
- in railroads, 2:474
- Intelligent transport system (ITS), 4A:60, 4A:106, 4A:162, 4A:289, 4A:2, 4B:49, 5:331, 5:711, 6:298
- applications, 6:300
 - direct impact, 5:167
 - and infrastructures, 6:1
 - main components, 6:302

- Intelligent transport system (ITS) (*Cont.*)
 automatic identification systems (AIS), 6:303
 automatic vehicle location systems (AVLS), 6:303
 Cartographic Databases and Geographic Information Systems (GIS), 6:303
 data collection for traffic flows or passengers through sensors and their related networks, 6:303
 electronic data interchange (EDI), 6:303
 telecommunication systems, 6:302
 in transport planning, 6:299
- Intentional Acts, 2:9
- Interactions between transport and urban development, 1:93
- Interchange merge control, 4A:140
- Intergenerational changes of mobility patterns among young adults, 5:219
- Intergovernmental Panel on Climate Change (IPCC), 3:350, 3:402, 7B:129
- Intermittent Bus Lane (IBL), 4A:311
- Intermodal freight transport (IFT), 5:577
 challenges for intermodal transport, 5:579
 definition of, 5:577
 different types and classification of, 5:578
 position of inland ports and terminals in inland waterway IFT, 5:579
- Intermodal loading unit (ILU), 3:456
- Intermodal Surface Transportation Efficiency Act, 6:173
- Intermodal Transport, 3:456
- Internal combustion engine (ICE), 6:64, 6:214, 7B:130, 7B:148
- Internal costs, 1:15
- Internal environment, 5:249
- Internal factor and transit ridership, 4B:57
- Internalization in second-best settings, 1:64
- International air freight forwarders, 2:100
- International Air Transport Association (IATA), 2:27, 2:103, 2:307, 4A:385
- International Air Transport Association's (IATA), 6:8
- International Association of Classification Societies (IACS), 3:352
- International Association of Ports and Harbors (IAPH), 5:566
- International Association of Snowmobile Administrators (IASA), 2:82
- International Association of Snowmobile Manufacturers Association (ISMA), 2:82
- International Civil Aviation Organization (ICAO), 2:30, 2:34, 2:90, 2:102, 2:353, 3:355, 4A:381
- International Committee on TCT (ICTCT), 2:663
- International Convention of Safety of Life at Sea (SOLAS), 2:294
- International Differences, 2:358
- International Energy Agency (IEA), 6:215
- International Energy Agency (IEA) scenario targets to 2040, compared to needed 2050 targets, 7B:132
- International Gateways, 3:37
 governance, 3:38
 logistics facilities of, 3:38
- International Longshore and Warehouse Union (ILWU), 3:324
- International Maritime Dangerous Goods Code (IMDG Code), 2:307
- International Maritime Organization (IMO), 2:307, 3:58, 3:81, 3:286, 3:294, 3:352, 5:566, 5:600
- International maritime trade, 5:508
- International migration, 1:90
- International Organisation for Standardization (ISO), 7A:113
- International Ship and Port Facility Security Code (ISPS Code), 3:302
- International Telecommunication Union (ITU), 4B:46
- International trade affect transport, 1:432
- International understanding, 2:8
- International Union of Railways (UIC), 6:261
- Internet data, 6:135
- Internet of Things (IOT), 1:23, 4A:289
- Interoperable Electronic Train Management System (I-ETMS), 2:464
- Interplay between parking and other markets, 1:162
- Interregional transport investment, 1:259
- Intersectionality, control for, 7A:196
- Intersection capacity
 capacity of roundabouts, 4A:255
 India, 4A:255
 Indonesia, 4A:256
 United States of America, 4A:255
- concept of, 4A:247
 Australia, 4A:250
 capacity of signalized intersection, 4A:248
 India, 4A:249
 Indonesia, 4A:250
 United Kingdom, 4A:251
 United States of America, 4A:249
 unsignalized intersection, 4A:251
 India, 4A:253
 Indonesia, 4A:254
 United States of America, 4A:252
- Intersection safety (INS), 2:113, 2:112
- Interstate Commerce Act of 1887, 3:25
- Interstate Commerce Commission (ICC), 3:25
- Intertemporal variation in values of time (VOTs), 1:170
- Intertemporal variation of valuations, 1:170
 longitudinal studies, 1:171
 other evidence, 1:172
 repeated cross-sectional VOT studies, 1:171
 transferability of cross-sectional models, 1:170
- Inter-vehicle communication (IVC) systems, 6:307
- Interventions to Improve the Quality of Maintenance and Inspection, 2:30
- Interventions to increase modal usage by older and disabled people, 5:90
- Intervention strategies related to stages of change, 7A:49
- Intervention strategies targeting selected psychological factors, 7A:48
- Interview protocols, 7A:40
- Interviews, 7A:40, 7A:42
- Intra- and intergenerational changes of travel, 5:218
- Intracity transport investment, 1:259
- Intra-logistics, drones usage in, 3:375
- Inuit Circumpolar Council, 3:349
- In-vehicle event data recorders (EDR), 6:307
- In-vehicle information systems (IVIS), 7A:161
- In-vehicle time (IVT), 1:122
- In-vehicle travel time (IVTT), 4B:194
- Inventory-based model, 3:234
- Inventory carrying costs (ICC), 3:115
- Investment and capital stock of transport infrastructure in Germany 2017, 1:455
- Investments in transport infrastructures, 5:1

Ireland (Ring a Link & Local Link), 4B:19
 Iron ore seaborne trade, 3:263
 Italy (ProntoBus), 4B:19
 Italy, tradition of limited traffic zone (ZTL), 1:137
 IT-based analytics capabilities, for logistics, 3:73
 Iterative proportional fitting (IPF), 4B:165, 4B:193
 I-35 westbound bridge, 2:145

J

Japan, electronic toll collection in, 4A:64
 Japan, flood impact to roads in, 2:746
 Jevons Paradox, 3:402
 Jihadist, 2:670
 Job Access, 4B:62
 Job Access Reverse Commute (JARC) program, 4B:63
 John Deere Crossover Gator utility vehicle, 2:78
 Jones Case, 3:385
 Junction type, 2:445
 Just-in-Time (JIT), 3:107
 continuous improvement and, 3:110
 inventory-transportation tradeoffs in, 3:109
 success, pros and cons, 3:107
 and sustainability, 3:111
 transportation in, 3:108
 contradiction, 3:108
 mode of, 3:109
 outbound *versus* inbound transportation, 3:109
 Toyota mixed-load system, 3:108
 warehouse, 3:108

K

Kaldor-Hicks Criterion, 1:190
 and random utility models, 1:193
 and standard transport appraisal, 1:193
 theoretical problems, 1:192
 Karolinska Sleepiness Scale, 2:218
 Kayaking, 2:480
 Key cargo airports, 5:261
 Key features and drivers of maritime regulation, 5:600
 Key performance indicators (KPI), 3:16
 Khazzoom-Brookes hypothesis, 3:402
 Knowledge-based errors, 2:29
 Kobotoke tunnel, 4A:148
 Kolmogorov-Smirnov test, 4B:175
 KoMoDo Project, 3:493
 Kruskal-Wallis test, 7A:83

L

Lack of safety on the world's roads, 7B:8

Laden trips
 average length of, 3:399
 average load on, 3:399
 Landlord port, 3:300, 4A:392, 5:560
 Land use, 1:149
 Land-use and transport challenges, future perspective, 6:109
 Land-use and transport feedback cycle, 6:109
 Land use and transport interaction (LUTI) models, 6:187, 187
 Land-use and transport planning, 6:107
 Land-use and transport systems, integration, 6:108
 Land use changes, 5:315
 Land use mix, 4A:303
 Land use-transportation integrated (LUTI), 4B:34, 4B:38
 Land-use transport interaction (LUTI) models, 1:93, 1:357
 Landward planning, 1:444
 Lane change test (LCT), 2:607
 Lane Departure Warning (LDW), 2:181, 2:175
 Lane departure warning system (LDWS), 6:57
 Lane Keeping Systems (LKS), 2:217
 Lane keeping warning devices (LDW), 6:307
 Lane narrowing, 2:645
 Lane Support, 2:620
 Lane-use control signals, 4A:56
 Lanzhou-Kathmandu South Asia Freight Rail (LKSAFR), 3:496
 Large truck crashes, critical issues for advanced driver assistance systems, 2:206
 automated driving systems, 2:207
 distraction from secondary tasks, 2:204
 driver fatigue and drowsiness, 2:205
 driver health and wellness, 2:205
 fleet safety culture, 2:203
 impairment from drugs and alcohol, 2:205
 large truck driver inattention, 2:204
 large truck roadside inspections, 2:208
 risk factors for, 2:202
 speeding, 2:206
 strong safety culture, 2:203
 Large truck roadside inspections, 2:208
 Large Truck Safety
 truck crash causation, 2:202
 truck size and weight, 2:201
 vehicle height, 2:201
 vehicle length, 2:201
 vehicle weight, 2:201
 vehicle width, 2:201
 Lasso method, 4B:110

Last-mile delivery, 3:188, 3:219
 Latent class multinomial logit model, 4B:73
 Latent demand, 4A:93
 empirical findings, 4B:44
 outstanding issues, 4B:45
 Latent demand, 1:186
 Latent errors, 2:256
 Lateral shift, 2:644
 Laws, 2:34, 2:80
 LDO (Lease-Develop-Operate) framework, 1:375
 Leading Pedestrian Interval (LPI), 4A:352
 Lead logistics providers (LLP), 3:103
 Lean system, 3:107
 Learning from Incidents, 2:31
 Lease contracts, 1:374
 Lease-develop-operate (LDO), 1:375
 Legislation and enforcement, 2:271
 Leisure, 5:255
 Lenoir link, 3:431
 LEONTIEF technology, 1:523
 Less container load (LCL), 3:238
 Less-than-truckload (LTL) carriers, 3:108, 3:388
 freight minimums and maximums, 3:393
 inventory control, 3:391
 selection and supply chain planning, 3:389
 operational level planning, 3:389
 strategic level planning, 3:389
 tactical level planning, 3:389
 and shippers, 3:390
 transit time, 3:392
 variability, 3:392
 Leveling the international airliner playing field, 5:195
 Level off, level off^r, 2:167
 Level of Service (LOS), 4A:17
 LE Vélo STAR, 2:139
 LEZ Versus Congestion Tolls, 1:235
 LHOVRA, 2:706
 Liberalization, 1:366
 Licensed drivers, 2:282
 License plate matching, 4A:260
 License Plate Recognition (LPR), 4A:289
 Licensing and car availability, 5:657
 Licensing and medical approaches, to elderly driver safety, 2:237
 Lidars, 2:176
 Life and Injuries, Value of
 altruistic *versus* individual values, 2:740
 fatal risk reductions, 2:737
 aggregating individual WTP, 2:739
 model operational, 2:738
 Life-cycle assessment (LCA), 4B:123
 Life cycle cost (LCC), 1:18, 1:306

- Life-cycle cost analysis (LCCA), 4B:121
- Life-Cycle Planning (LCP), 4B:123
- Life events-empirical work, 5:172
- Life jackets, 2:481
- Lifestyle changes, 1:91
- Light Detection and Ranging (LiDAR), 2:194
- Light emitting diode (LED), 2:362, 7B:69
- Lighter-than-air aircraft (aerostats), 5:270
- Lighting
 - lighting technologies, 2:362
 - light-emitting diodes, 2:362
 - roadway lighting, 2:364
 - lighting and safety, 2:364
 - roadway lighting practices, 2:365
 - vehicle lighting, 2:363
 - forward lighting, 2:363
 - visibility, 2:361
 - glare, 2:362
 - visual performance, 2:361
- Light pollution, 7B:68
- Light rail
 - common characteristics of, 6:33
 - promise of, 6:34
 - Land and Property Values, 6:36
 - non-user benefits, 6:35
 - strengths and limitations of, 6:35
 - user benefits, 6:34
 - renaissance, 6:33
 - social exclusion and gentrification, 6:37
 - sustainability, 6:37
 - transit around the world, 6:33
- Light rail investment, 6:31
- Light rail transit (LRT), 6:31
- Light rapid transit (LRT), 1:279
- Light trespass, 7B:70
- Light trucks, 2:296
- Likelihood ratio test, 4B:175
- Limited mobility, 5:85
- Limited-stop service (Express), 5:330
- Liner Service Networks, 3:249
- Liner shipping, 3:335
 - container routing decision and transit time, 3:339
 - number of ships and schedule tightness, 3:337
 - services
 - schedule design of, 3:340
 - scheduling of, 3:336
- Link performance function, 4B:27
- Liquefied biogas (LBG), 3:288
- Liquefied natural gas (LNG), 5:536
- Load factors, 3:163
- Loading bay management systems, 1:25
- Local area traffic management (LATM), 4A:268
 - blister island treatments, 4A:268
 - critical values, 4A:275
 - design solutions, 4A:269
 - good practice, 4A:269
 - through traffic, 4A:273
 - traffic problems in local areas, 4A:268
- Local clock, 4A:170
- Local distribution companies (LDC), 3:463
- Local freight demand, 3:170
- Local urban logistics innovations, 3:39
- Location Choice Models, 4B:22
 - addressing spatial autocorrelation in, 4B:23
 - choice sets for, 4B:24
 - discrete choice models, 4B:22
 - measurement units and modifiable areal unit problem, 4B:23
- Location referencing system (LRS), 4B:120
- Locomotive power traction systems, 3:427
- Logistics, 5:41
- Logistics and Supply Chain Management
 - activities, 3:8
 - extended, 3:9
 - agricultural commodity supply chain, 3:14
 - B2C supply chain, 3:12
 - bulk material supply chain, 3:12
 - clinical trials supply chain, 3:15
 - construction supply chain, 3:13
 - durables supply chain, 3:14
 - global supply chain, 3:13
 - humanitarian supply chain, 3:13
 - innovative supply chains, 3:14
 - military supply chain, 3:14
 - product supply chain, 3:10
 - promotional supply chain, 3:10
 - recovery supply chain, 3:13
 - recycling supply chain, 3:13
 - resource supply chain, 3:13
 - talent supply chain, 3:12
- Logistic service providers (LSP), 3:479
- Logistics Facilities, Planning Policies for, 3:39
 - local urban logistics innovations, 3:39
 - regional and metropolitan spatial planning policies, 3:39
- Logistics Information Systems, 3:76
 - advanced supply chain planning systems, 3:81
 - enterprise planning systems, 3:76
 - modeling and simulation, 3:82
 - tracking and tracing systems, 3:80
 - transport management systems, 3:77
 - warehouse management systems, 3:78
- Logistics networks, new theory of, 3:479
- fragmented networks and operations, 3:479
- low level of efficiency, 3:479
- service improvement and sustainability, 3:480
- Logistics Outsourcing, 3:102, 3:104
 - advantages of, 3:104
 - issues and influences, 3:106
 - risks of, 3:105
- Logistics Performance Index (LPI), 3:94, 3:151
 - delivering good quality services, 3:99
 - envelope of implementation, 3:100
 - guidelines, 3:95
 - logistics performance matters, 3:95
 - logistics performance persist, 3:95
 - methodology, 3:94
 - polymaking and business impact, 3:100
 - supply chain reliability and service quality, 3:98
 - supply chain resilience and sustainability, 3:99
- Logistics real estate, 3:37
- Logistics Service Performance
 - customer segments for, 3:90
 - evaluation of, 3:89
 - improvement of, 3:91
- Logistics Service Providers (LSP), 3:66, 3:102, 3:458, 3:460
- Logistics Systems
 - in developing countries, 3:153
 - comprehensive logistics strategies, 3:154
 - demand and benchmark logistics performance, 3:153
 - develop and maintain infrastructure, 3:154
 - exploit disruptive technologies, 3:155
 - regulatory environment, 3:154
 - skills development, 3:155
 - trade facilitation, 3:154
 - key bottlenecks in, 3:151
 - performance efficiency of, 3:150
- Logistics Zones, 3:36
 - development, local policies of, 3:36
 - industrial zones to logistics zones, 3:36
 - local policies of logistics zones development, 3:36
 - logistics real estate, 3:37
 - warehouses, distribution centers, and logistics zones, 3:36
- Logsums, 1:240
- London
 - congestion charge as a prototype, 5:378
 - congestion charging scheme, 1:136
 - congestion pricing scheme, 4A:72
 - road pricing in, 4A:72

Long distance passenger trains, 5:408
 Long distance travel (LDT), 6:272
 determinants of, 6:275
 macro drivers of growth, 6:274
 measurement and methods, 6:273
 policy issues, 6:275
 trends, 6:274
 Long John bike, 3:489
 Long-range solutions (LORA), 3:483
 Long-run equilibrium, of road transportation, 4A:84
 optimal capacity, 4A:84
 relaxation of assumptions, 4A:88
 Long-run *versus* short-run valuations, 1:102
 differences in valuations depending on the temporal scale, 1:104
 implications, 1:104
 Long short-term memory (LSTM), 4A:131
 Long-term effects of noise on health, 7B:91, 7B:92
 cardiovascular disease, 7B:91
 effects on cognitive development, 7B:91
 mental health, 7B:91
 Looping effects, 2:35
 Loss of control in flight (LOCI), 2:105
 Lost production, value of, 2:196
 Low- and middle-income countries (LMIC), 2:269
 road traffic crashes, 2:273
 legislation and enforcement, 2:273
 public education, advertisement, and mass media campaigns, 2:274
 road safety programs, 2:273
 Low- and middle-income countries (LMIC), 5:6
 Low-cost carriers (LCCs), 1:392
 Low emission zones (LEZ), 1:231
 Low-income countries (LIC), 2:269
 Low-level wind shear alert system (LLWSAS), 2:353
 Low-Pass Filtering, 4A:129
 Low-pressure pipelines and hypersonic transportation, 5:650
 Low range radio altimeter (LRRRA), 2:165
 Low Volume Incident Detection Algorithms, 4A:130
 LPS lamps, 7B:69
 Lyft, 4B:158

M

Machine learning (ML), 2:182, 6:167
 Machine vision-based inspection, 2:461
 Macrofundamental diagram (MFD), 4B:142
 Macroscopic fundamental diagram (MFD), 4A:319

Macroscopic Safety Analysis
 geographic units, 2:368
 census-based units, 2:368
 developing new geographic units for macroscopic safety analysis, 2:369
 political boundary units, 2:368
 postal units, 2:368
 macroscopic contributing factors to traffic crashes, 2:374
 climate factors, 2:376
 demographic factors in macroscopic safety studies, 2:375
 evaluating policies for safety, 2:377
 integration with transportation planning, 2:377
 land-use factors, 2:375
 recent research trends, 2:376
 simultaneous analysis of microscopic and macroscopic safety, 2:376
 socioeconomic factors, 2:375
 transportation factors, 2:374
 methodologies and issues, 2:370
 boundary crash problem, 2:373
 modifiable areal unit problem, 2:373
 spatial dependency, 2:372
 statistical modeling methods, 2:370
 traffic-based units, 2:368
 traffic analysis districts, 2:368
 traffic analysis zones, 2:368
 traffic safety analysis zones, 2:368
 Maggie's law, 2:612, 2:614
 Magnetic card, 4A:329
 Magnetic levitation transport systems, 5:650
 Maine Bridges, 2:148
 Main negative externalities in transport, 6:191
 Main port management models, 5:558
 Main road network in the Roman empire under emperor Hadrianus, 5:361
 Maintenance Environment Questionnaire (MEQ), 2:31
 Maintenance Event Decision Aid (MEDA), 2:31
 Maintenance personnel, 2:26
 Maintenance Repair and Overhaul (MRO), 2:25
 Major modes, energy, and space consumption, 5:308
 Major network structures of cargo airline or shipping lines, 1:41
 Manage bus priority planning, 6:257
 acceptance, 6:257
 collaboration, 6:258
 funding models, 6:258

guidelines and action plans, 6:258
 Management contracts, 1:374
 Managing finances of the rail activity, 5:448
 good understanding of costs and yields, 5:448
 increase of productivity, 5:448
 revenues/expenses ratio, 5:448
 segmentation of traffic, 5:449
 Man and Biosphere (MAB) Program, 7B:12
 Manual on Uniform Traffic Control Devices (MUTCD), 2:243, 2:388, 2:623, 2:648, 2:711, 4A:214
 Mapping maritime transport, 5:518
 Marco Polo program, 3:282
 Marginal costs of road transport, 1:51
 Marginal external cost (MEC), 1:58
 Marginal Opportunity Cost of Public Funds (MOC PF), 1:464
 Marginal private benefit (MPB), 1:58
 Marginal rate of substitution (MRS), 1:115
 Marginal social benefit (MSB), 1:58
 Marginal social cost (MSC)
 curve, 1:58
 pricing, 1:57
 Marginal vs average cost coverage under economies and diseconomies of scale, 1:372
 Marginal WTP (MWTP), 1:115
 Maricopa Association of Governments (MAG), 4A:223
 Marijuana, 2:340, 341
 Marine classification, 5:607
 Marine gas oil/marine diesel oil (MGO/MDO), 3:286
 Marine Liability Act, 3:354
 Marine spatial planning (MSP), 3:286
 Marine Transportation Security Act, 3:354
 Maritime Mobile Service Identify (MMSI), 3:81
 Maritime region as a framework, 5:524
 Maritime regulation
 key drivers of, 5:601
 key features of, 5:600
 three pillars of, 5:602
 Maritime route planning, 5:570
 commercial route planning, 5:571
 container shipping, 5:575
 cruise, 5:575
 dry bulk shipping, 5:574
 liner-shipping, 5:574
 market planning, 5:571
 navigational route planning, 5:573
 regulatory route planning, 5:572
 RoRo and RoPax shipping, 5:575
 service planning, 5:571

- Maritime route planning (*Cont.*)
 tanker shipping, 5:574
 tramp shipping, 5:573
 voyage planning, 5:572
- Maritime security (MARSEC), 3:302, 5:604
- Maritime shipping, 5:508
 Baltic freight index and Clarksons Average Earnings, 5:514
 container liner shipping, 5:514
 freight and timecharter rates, 5:515
 instability index, 5:513
 supply and demand dynamics, 5:515
 supply, demand, and freight rates, 5:513
 tramping, 5:513
- Maritime Silk Road (MSR), 3:495
- Maritime trade, significance of, 4A:390
- Maritime transport, 5:40, 5:508
- Maritime transport as a (spatial) network, 5:520
- Market-based approach, 1:197
- Market-based instruments, 1:63
- Market-based measures (MBM), 3:294
 Track, 3:296
- Market concentration, 5:511, 5:595
- Market cycles, 4A:394
- Market failure, 1:2, 1:56, 3:153
- Market failures in the transport sector, 1:3
- Market Imperfections and ôSmallö Changes, 1:357
- Marketing airline airfreight services, 3:371
- Market needs, 4A:393
- Market organization, 5:590
- Market places, 3:219
- Market power, 1:161
- Market rates, 1:199
- Market segmentation, 5:595
- Market selection, 5:252
- Market shares of rail, road, and air in France, 1:421
- Markov Chain Monte Carlo (MCMC), 4B:5, 4B:175
- Marshallian consumer's surplus, 1:239
- Maslow's theory, 6:39
- Master clock, 4A:170
- Material costs, 2:196
- Material-failure type accidents, 2:294
- Mathematical models, 6:184
- MATSim, 4B:69
- Maximum allowable operating pressure (MAOP), 3:465
- MAXimum BANDwidth (MAXBAND), 4A:173
- Maximum Likelihood Estimation (MLE), 4A:146
- Maximum size of containerships, development of, 5:553
- Maximum Sustainable Flow Rate, 4A:18
- MCDA models, 5:60
- McMaster Algorithm, 4A:129, 4A:130
- Mean sea level (MSL), 2:164
- Mean time to detection (MTTD), 4A:131
- Measures of crash proximity, 2:664
- Measures Suitable for Secondary Roads, 2:61
- Measures Targeted to Animals and Their Behavior, 2:59
- Measuring remoteness and inaccessibility, 5:52
- Media and the public policy process, relationship between, 6:400
- Media and the transport policy process, reciprocal relationship between, 6:401
- Media coverage, characteristics of, 6:404
- Median, 2:593
- Medical costs, 2:196
- Mega-cities, 3:179
- Megaproject, 1:470
- Megaprojects become a priority for policy, 1:470
- Mega-ships and diseconomies of scale in port, 5:552
- Mega transport projects (MTPs), 6:147
 complexity, 6:147
 cost, 6:147
 evaluation of sustainable development goals (SDG) remains under-developed, 6:150
 impacts, 6:147
 as independent and open "system of systems", 6:149
 participation practices need "opening up" to identify and mediate conflicts over impacts, 6:150
 path-dependent appraisal methods and narrow consideration of impacts, 6:149
 size, 6:147
 strategic Success Paradoxes, 6:149
 typical performance issues during delivery, 6:148
 diverse project cultures and rationales, 6:148
 misaligned and under-developed governance, 6:148
 strategic rent seeking behaviour, 6:148
- Melbourne
 hook turns in, 4A:205
 ramp metering in, 4A:8
 road network, 4A:203
 tram network, 4A:203
- Melbourne and Metropolitan Board of Works (MMBW), 4A:8
- Memory failures, 2:29
- Mental health, 7B:106
- Mental health factors, 2:19
- Mental workload, 2:216
 and driver state, 2:218
 in partly self-driving vehicles, 2:217
 during driving and driver state, 2:216
- Merge bottleneck, 4A:138
- Merkle tree hashing, 3:136
- Metabolic equivalent of the task, 7B:170
- Metal-oxide-semiconductor (MOS), 7A:14
- METANET model, 4A:35
- Metering signals, 4A:229
- MeToo! Movement, 2:577, 2:580
- METROPOLIS Model, 1:148
- Metropolitan Planning Organizations (MPO), 4B:36
- Metropolitan Transportation Authority (MTA), 2:241
- Microburst, 2:93
- Microeconomic theory, 1:67
- Micro-level aspects of the timetable and responses to disruptive, congestion-causing events, 1:380
- Microlevel Transport Simulation Models, 4B:69
- Micromobility, 2:131
- Micromobility, 5:391
- Micromobility vehicle (bicycle and e-scooter) sharing, 7A:187
- Microscopic models, 4B:142
- Microsimulation and Agent-Based Models in Transportation microlevel, 4B:68
- Microtransit, 6:101, 6:103, 6:105
 history and legality, 6:101
 market structure, 6:102
 micromobility markets, 6:105
 mobility as a service, 6:105
 regulation, 6:101
- Mid-air collision avoidance, 2:164
- Middle-aged road users, 2:15
- Middle-income countries (MIC), 2:273
- Migration, 1:258, 5:256
- Mileage (or Utilization) Tax, 1:518
- Miles-in-trail (MIT), 4A:385
- Military aviation, 2:90
- Military Ships, 5:613
- Military supply chain, 3:14
- Millennium development goals (MDGs), 7B:2
- Milvian Bridge, 2:145
- Minimum Legal Drinking Age (MLDA), 2:42
- Minimum parking requirements (MPRs), 1:162
- Ministry of Infrastructure and the Environment (MIE), 2:522
- Ministry of Transport, 1:4654

- Minutes-in-trail (MINIT), 4A:385
- Miracle on the Hudson, 2:53
- Misery Index (MI), 4A:111
- Mitigating transportation impacts on fish and wildlife, 7B:77
- not involving crossing structures, 7B:77
 - road-stream crossings, 7B:78
 - wildlife crossing structures, 7B:77
- Mitigating urban-related environmental hazards, 7B:104
- Mitigation, of head-on crashes, 2:314
- MITSIMLab, 4B:69
- Mixed lanes, 4A:63
- Mixed logit (ML), 3:233
- Mixed logit (ML) model, 4B:73
- Mixed recursive logit (MRL), 4B:93
- Mixed retailer, 3:219
- Mixed traffic, 2:573
- Mixed traffic train graph, heterogeneous services, 1:381
- Mixed traffic tram
- lane, 4A:339
 - operation, 4A:338
- Mobile network operator (MNO), 2:185
- Mobile phone network data (MND), 5:668
- Mobile phone usage and driving, 2:496
- Mobile traffic STOP sign, 2:571
- Mobility, 6:107
- Mobility analysis, framework for, 6:185
- Mobility as a Service (MaaS), 6:1, 6:12, 6:187, 6:318, 6:332
- anticipated positive implications for the current MaaS vision, 6:13
 - as evolution of transport system integration, 6:12
 - networked multi-actor MaaS innovation processes, 6:14
 - socio-technical convergence around the current MaaS vision, 6:14
 - toward responsible MaaS innovation and institutional transition, 6:16
 - unanticipated negative implications from the current MaaS vision, 6:15
- Mobility as a service (MaaS), 1:365
- Mobility-As-A-Service (MaaS), 4B:40, 4B:60
- Mobility as a Service (MaaS), 5:157, 5:35
- Mobility biographies
- understanding stability and change in, 5:173
- Mobility biographies, 5:171
- Mobility for older people during COVID-19, 7B:156
- Mobility for people with disabilities, 7B:156
- Mobility index, 4A:304
- Mobility in later life, importance of, 6:59
- Mobility justice (MJ), 7B:81, 7B:84
- Mobility resources, 5:121
- MobiTopp, 4B:69
- Modal split, 3:398
- Modal split in the presence of external costs, 1:60
- Modal Transport/Public Transport, 1:518
- Mode choice, 4B:187
- Mode choice of the firm, 5:57
- Modeling Studies, 1:323
- Modeling to assess the impact of MaaS and NMS
- data required to model MaaS and NMS, 6:335
 - demand modeling-key changes, 6:336
 - DeMAND agent-based model, 6:337
 - "mobility on demand laboratory environments" (MODLE), 6:336
 - local authorities' perspective, 6:334
 - requirements for demand models to model mobility, 6:337
 - stakeholders perspective, 6:334
 - transport operators' perspective, 6:335
- Modeling to assess the impact of MaaS and NMS, 6:333
- Modeling Transit Ridership, 4B:54
- Model Minimum Uniform Crash Criteria (MMUCC), 2:9, 2:382
- Model of age-friendly transport system, 5:7
- Model Predictive Control (MPC), 4A:37, 54
- Models and simulation for transport planning, 6:185
- limits and new trends of, 6:186
- Models of habit, 7A:55
- models based on other theories, 7A:56
 - models based on the norm activation model, 7A:55
 - models based on the theory of planned behavior, 7A:55
 - models based on the value belief norm theory, 7A:56
- Model Transit Ridership, 4B:53
- Modern roundabouts, 2:539
- Modern transport systems, role of paratransit in, 5:302
- Modes as the linchpin of the global economy, 5:38
- Mode Share Models, 4B:75
- Mode-specific aggregate origin-destination (OD) level, 4B:98
- Mode split, 4B:26
- Modifiable areal unit problem (MAUP), 2:371, 2:373
- Modifiable area unit problem (MAUP), 4B:23, 36
- Modifying spatial and temporal behavior, indirect impact, 5:168
- Mogridge's Hypothesis, 1:491
- Mohring Effect, 1:263
- empirical evidence on, 1:265
 - in the original framework, 1:264
 - in a simple framework, 1:263
- Mohring-Harwitz Theorem, 4A:86, 4A:88
- Monetary valuations of travel-related attributes, 1:102
- Monetary value of a gain (WTP)
- Monetizing safety preferences, 1:115
- Monica Algorithm, 4A:129
- Monitor vertical speed", 2:167
- Monocentric model, 1:297
- heterogeneous Workers, 1:298
 - Muth condition, 1:297
 - transportation cost, 1:298
- Monohull ferry, 2:291
- Monte Carlo Method (MCM), 2:348
- Monte Carlo simulation, 4B:4
- Mood improvements as a result of cycling, 7A:140
- Moral hazard, 1:371
- Mortality, 7B:106
- Mothers Against Drunk Driving (MADD), 2:49
- Motion Sickness Assessment Questionnaire (MSAQ), 2:606
- Motivational model for private car use, 5:385
- Motivations, 1:389
- Motor Carrier Act of 1980, 3:26, 3:382
- Motorcycle cities, 5:47
- Motorcycle crashes, factors contributing to, 7A:146
- Motorcycle safety, measures to improve, 7A:147
- Motorcycles lanes, exclusive, 2:447
- Motorcycle transport (and its facilitation by mobile phones), 5:53
- Motorcycling, diversity in, 7A:145
- motorcycle types, 7A:145
 - purposes of riding, 7A:145
 - rider demographics, 7A:146
 - riding locations, 7A:146
- Motorcyclist, 7A:144
- conflicting perspectives, 7A:144
- Motor vehicle crash reportability, 2:383
- challenge of defining reportability, 2:380
 - crash, 2:381
 - motor vehicle, 2:382
 - traffic crash, 2:384
 - narrowing the focus, 2:381
- Motor Vehicle Safety Act, 2:521

Motor Vehicle Traffic Crashes, 2:6
 Motor Vehicle Transport Act, 2:521
 Motorways of the Sea (MoS), 3:282
 concept, 3:282
 current situation of, 3:283
 Movement Numbering Scheme, 4A:216
 Moving Ahead for Progress in the 21st century (MAP-21) Act, 4B:123
 Moving-light-guide-system (MLGS), 4A:141
 Multi-criteria analysis (MCA), 6:150
 Multi-Criteria Decision Analysis (MCDA), 4B:79
 advantage of, 4B:79
 classifications of, 4B:82
 PROMETHEE, 4B:84
 stages in, 4B:80
 analyze the results and make recommendations, 4B:81
 assess the performance of alternatives, 4B:80
 assign weights for different criteria, 4B:81
 decision-making problem, 4B:80
 evaluating alternatives, 4B:80
 identify the project alternatives, 4B:80
 overall score of alternatives, 4B:81
 sensitivity analysis and final selection of alternative, 4B:82
 techniques for transportation-related decision-making, 4B:82
 analytical hierarchical process, 4B:83
 analytical network process, 4B:84
 ELECTRE, 4B:85
 goal programming, 4B:83
 PROMETHEE, 4B:84
 TOPSIS, 4B:85
 weighted global criterion method, 4B:82
 weighted sum model, 4B:83
 Multidimensional Driving Style Inventory (MDSI), 7A:65
 examples of studies using the, 7A:71
 and familial and social contexts, 7A:69
 MDSI and personality, 7A:68
 Multi-disciplinary approach for structuring modal choice determinants, 5:64
 Multi-Flight Common Route (MFCR), 4A:371
 Multimedia-based countermeasures, 22
 Multimodal behaviors, beneficial factors and barriers for
 psychographic characteristics, 5:123
 sociodemographic factors, 5:123
 Multimodal behaviors, beneficial factors and barriers for, 5:122

Multimodal behaviors, reasons and causes for, 5:121
 Multimodality as a facet of individual travel behavior, 5:119
 Multimodality as a property of the transport system, 5:118
 Multimodality as a transport policy strategy, 5:119
 Multimodality in transportation, 5:118
 Multimodality, measurements of, 5:120
 methods of capturing transport mode use, 5:120
 reference or observation periods, 5:121
 types of transport modes, 5:121
 variability of mode use, 5:121
 Multinomial logit (MNL), 3:233, 4B:9, 4B:77
 Multiperiod model, 1:117
 Multiple-criteria decision-making (MCDM), 1:514
 Multiple Discrete-Continuous Extreme Value (MDCEV), 4B:100, 4B:193
 Multiple-need model, 2:110
 environmental adaptation, 2:110
 feeling confident and secure, 2:111
 getting contact with and/or being localized, 2:111
 guidance, 2:111
 Multiple Priority Requests, 4A:310
 Multiregional or spatial input/output (MRIO) table, 3:165
 Muth condition, 2:297
 Mutual aid, 2:244
 Myopic passengers, 5:235

N

Narrow road, 4A:91
 National Academy of Sciences, Engineering, and Medicine (NASEM), 2:50, 3:32
 National Airspace System (NAS), 4A:363
 National Association of City Transportation Officials (NACTO), 2:120, 2:642
 National Automotive Sampling System (NASS), 2:525
 National Center for Atmospheric Research (NCAR), 2:93
 National Center for Missing and Exploited Children, 2:566
 National Center for Statistics and Analysis (NCSA), 2:410
 National Collision Database (NCDB), 2:521
 National Committee on Uniform Traffic Control Devices (NCUTCD), 2:630
 National Cooperative Highway Research Program (NCHRP), 2:341, 2:508, 2:642, 4A:193, 4B:118
 National Crime Victimization Survey (NCVS), 2:157
 National Customs Authority, 2:101
 National Disaster Preparedness Training Center, 2:280
 National Fire Academy, 2:241
 National Fire Protection Association (NFPA), 2:94
 National Freight Transport Models four-step model and beyond, 3:162
 modeling freight generation and distribution, 3:165
 modeling mode choice and route choice, 3:166
 National Highway System (NHS), 2:387
 National Highway System Designation Act (NHSDA), 2:625
 National Highway Traffic Safety Administration (NHTSA), 2:6, 2:49, 2:146, 2:152, 2:11, 2:525, 2:311, 2:404, 2:564, 2:572, 2:384, 3:402, 4B:87
 National Household Travel Survey (NHTS), 4B:10, 4B:87, 4B:223
 uses of the data, 4B:87
 National Incident Management System (NIMS), 2:240, 2:243, 4A:125, 5:129
 National Institute on Alcohol Abuse and Alcoholism (NIAAA), 2:42
 Nationalization, 3:27
 of railroads, 3:27
 Nationally determined contributions (NDCs), 1:72
 National maritime administrations, 1:437
 National Maximum Speed Law (NMSL), 2:625
 National motor vehicle crash causation study, 7A:124
 National Oceanic Atmospheric Administration (NOAA), 2:279
 National Operations Center (NOC), 2:242
 National/regional progress to date and targets, 7B:131
 National Response Framework (NRF), 2:240, 2:243
 National Road Safety Strategy (NRSS), 2:271, 2:519
 National Safety Code (NSC), 2:521
 National School Bus Glossy Yellow, 2:571
 National traffic flow management, 4A:366

- National Traffic Incident Management Coalition (NTIMC), 4A:123
- National Transportation Safety Board (NTSB), 2:28, 2:145, 2:525, 2:337, 2:624
- National Transport Development Policy Committee, 3:33
- National unified goals (NUG) 4A:123
- National Vehicle Crash Causation Study, 2:332
- National Weather Service (NWS) Doppler weather radars, 2:93
- Nations International Labor Organization (the ILO), 1:437
- Natural gas (NG), 5:536
- Naturalistic driving, 7A:21
- Naturalistic driving studies (NDS), 7A:21, 7A:124
- Australian naturalistic driving study (ANDS), 7A:24
- Canadian naturalistic driving study (CNDS), 7A:26
- case studies, 7A:23
- ethical issues related to, 7A:23
- European union, main research question, 7A:31
- future perspective, 7A:36
- limits and considerations of, 7A:23
- primary goal of the SHRP2 safety research program, 7A:34
- Shanghai naturalistic driving study (SHNDS), 7A:30
- strengths of, 7A:21
- United States, SHRP2 dataset useful for research, 7A:34
- Natural monopoly, 1:464
- econometrics of, 1:33
- formal definition of, 1:31
- regulating, 1:32
- in transports, 1:34
- Near Field Communication (NFC) technology, 4A:329
- Need-based theories, 7A:139
- Negative Binomial (NB), 2:726
- Neoclassical approach and its ethical problems, 1:216
- environmental goods, 1:217
- value of statistical life, 1:217
- value of time, 1:218
- Neoclassical approaches, 1:256
- Neoliberal urbanism, 6:31
- Nested Logit model, 1:92
- Nested recursive logit (NRL), 4B:93
- Net capital stock, 1:453
- The Netherlands
- road safety management in, 2:522
- The Netherlands (RegioTaxi), 4B:19
- Net or gross cost contracts, 1:310
- Network diversity and centralization of global maritime flows in 2004, 5:528
- Network flow model, 3:157
- NetworkManager, 4A:381
- Networks, 1:148
- Neurodevelopment in children, 7B:106
- New and Emerging Data Forms (NEDF), 6:135, 6:136, 6:137, 6:138
- opportunities and challenges for, 6:139
- New automobile industry, parking support technology for, 4A:287
- New Car Assessment Program (NCAP), 2:406, 2:525, 2:403
- New Delivery Concepts, 3:491
- New economic geography (NEG) regional models, 1:291
- theory, 1:256
- New information and communication technologies (NICT), 1:388
- New Jersey barrier, 2:516
- New offerings for policy makers, 6:136
- cross-sectoral modeling, 6:137
- emerging modeling and analysis methods, 6:136
- positively disrupting travel choice and behavior, 6:137
- transport network modeling, 6:136
- New public management, 1:365
- New South Wales (NSW), 4A:344
- New transport infrastructures, 1:91
- New transport modes role and technologies, 1:592
- New urban agenda, 7B:6
- New York City Transit (NYCT), 2:241
- NextGen Implementation, 4A:372
- automatic dependent surveillance-broadcast, 4A:372
- data communications, 4A:374
- global harmonization, 4A:375
- integrated technology, 4A:375
- operations, 4A:374
- optimal descent trajectories, 4A:374
- surface, 4A:375
- wind-optimal and user-preferred routes, 4A:374
- performance-based navigation, 4A:373
- weather integration, 4A:373
- NextGen Network Enabled Weather (NNEW), 4A:373
- NHTSA Alcohol and Highway Safety Reports, 2:50
- NHTSA vehicle complaint data, 2:87
- Niche port, 3:299
- Night vision goggles (NVG), 2:318
- Night vision system (NVS), 6:57
- Nitrogen oxides (NOx), 5:267, 6:214
- emissions, 3:287, 7B:154
- Nodal regions in the global container shipping network in 1996 and 2006, 5:525
- Noise, 1:64, 5:110, 7A:170
- annoyance, 7B:88
- cost valuation, 1:106
- emissions, and other environmental externalities, 5:232
- Noise Control Act of 1972, 5:266
- Noise exposure-response relation (ERR) curves, 7B:88
- Noise indicators and measures, 7B:53
- Noise mitigation measures, 7B:93
- Noise pollution, 7B:53
- Noise Pollution and Abatement Act of 1970, 5:266
- Noise pollution modeling, 7B:59
- Nominal Safety, 2:386
- conformance to guidelines, 2:387
- consequences of, 2:389
- types of safety, 2:386
- Non-aeronautical Services, 5:231
- Noncommunicable diseases (NCDs), 7B:4, 7B:106
- Noncontact sexual violence, 2:576
- Nondepletable condition, 1:373
- Non-destructive inspection (NDI), 2:26, 2:32
- Nonfatal bridge crashes, in the State of Maine, 2:147
- Nongovernmental organizations (NGOs), 5:600
- Non-governmental sector (NGO), 2:243
- Non-methane volatile organic compounds (NMVOCs), 5:267
- Nonmotorized road users, 2:283
- Nonoptimizing Behavior, 1:132
- Nonparametric Wilcoxon rank-sum test, 7A:83
- Non-pneumatic tires (NPT), 2:684
- Non-price factors of costs, air freight, 1:38
- Non-price factors of costs, shipping, 1:40
- Nonprofessional road users, 2:556, 2:558
- Nonrecurrent congestion, 4A:158
- causes of, 4A:158
- inclement weather, 4A:159
- special events, 4A:160
- traffic incidents, 4A:158
- work zones, 4A:160
- Nontechnical Skills Training, 2:30
- Nonverbal sexual abuse, 2:576
- Normalcy illusion, 2:159
- Northern Sea Route (NSR), 3:349
- Northwest Passage (NWP), 3:349
- No serious injury or loss of life, 2:633

Not-in-Traffic Surveillance System (NiTS), 2:384
 No traffic, 2:573
 Novel route overlap-based approach to solving (TNDFSP), 5:625, 5:627
 demand assignment, 5:631
 holistic network evaluation, 631
 objectives, 5:627
 route and network creation, 5:630
 Novice driver education and training, 7A:158
 common approaches to, 7A:158
 group-based education initiatives, 7A:159
 one-on-one professional driving lessons, 7A:158
 professional and lay supervision, 7A:158
 current and emerging driver education and training needs, 7A:160
 foundations of current wisdom in, 7A:159
 theoretical and emerging research into, 7A:159
 NPV-cost ratio, 1:250, 1:252
 Null hypothesis, 7A:81
 Numerical Simulations, 2:68

O

Objective/actual safety, 2:386
 Objective data on travel characteristics, disadvantage of, 1:131
 Objective risks, 2:494
 Obstacles to Bus Priority, 6:256
 Obstructive sleep apnea (OSA), 2:338
 Occupancy, 1:185
 Occupancy Time Method (OTM), 4A:254
 Ocean shipping, 3:60
 Ocular parameters, 2:613
 Off-highway vehicles (OHV), 2:77
 Off-hour deliveries (OHDs), 1:22
 Off-peak tickets, 4A:327
 Off-road vehicles (ORV), 2:77
 Offsets, 4A:215
 effects of vehicle in queue on, 4A:215
 fine-tuning, 4A:216
 Off-Street Parking
 intelligent and informational parking management, 4A:286
 management strategy, 4A:285
 parking charging research, 4A:286
 parking guidance information technology, 4A:287
 parking information service platform, 4A:287
 parking sharing research, 4A:286

 parking supply allocation, 4A:285
 parking support technology for new automobile industry, 4A:287
 Oil products seaborne trade, 3:274
 Ola, 4B:158
 OLA (Objective data, List of solutions/actions, and addressed Action Plans), 2:524
 Older People Pedestrian Safety
 built environment form and design, 2:431
 decline in health condition, 2:429
 cognitive impairments, 2:430
 physical impairments, 2:430
 sensory impairments, 2:429
 road crossing behavior, 2:432
 crossing the road, 2:433
 curb delay, 2:432
 explorative behavior at the crossing environment, 2:432
 gap judgment and acceptance, 2:432
 Older people safe as drivers, 5:7
 Olfactory repellents, 2:61
 Omitted variable bias, 2:726
 Omni-Channel Logistics
 definition, 3:184
 inventory management, 3:187
 challenges, 3:188, 3:189
 last-mile delivery, 3:188
 return logistics, 3:188
 network design, 3:186
 challenges, 3:187
 warehousing, 3:187
 and Omni-Channel Retailing, 3:184
 Omni-Channel Retailing, 3:184
 Omnichannel synergies, 3:211
 On- and Off-Street, 1:187
 On-board unit (OBU), 4A:72
 One Belt and One Road (OBOR), 3:495
 One-period Model, 1:115
 On-line auctions, 1:23
 Online transaction brokers, 1:163
 On-street parking, 2:645
 influencing mechanism of, 4A:278
 intelligent and informational parking management, 4A:280
 parking spaces scale and location layout, 4A:278
 parking supply and demand regulation strategies, 4A:280
 spatial design method of, 4A:278
 Open skies agreements between regional airline markets, 1:400
 Central Asia-China, 400
 China-US market, 1:400
 European Union-Morocco, 1:400
 Europe-United States, 1:400
 Open Street Map (OSM), 4B:40
 Operating expenses (OPEX), 3:261

Operating Instructions, 2:34
 Operating speed, 2:635, 2:622
 Operational design domain (ODD), 2:181, 6:198
 Operational strategies and approaches, 5:131
 Operation cost of public transport, 1:26
 contracts, 1:29
 efficiency, 1:29
 regulation, 1:29
 Operator risk preference and adverse effects, 1:335
 Opiates, 2:224, 2:225
 Opportunistic and acquisitive carjackings, 2:158
 Opportunistic and instrumental carjackings, 2:158
 Optical character recognition (OCR), 4A:71
 Optimal Capacity Rule, 4A:85
 Optimal descent trajectories (ODT), 4A:374
 Optimal Pricing Rule, 4A:77
 (Optimal) Producer Surplus and Subsidies, 1:246
 Optimized Policies for Adaptive Control (OPAC), 4A:175
 "Ordered GEV" model, 1:129, 1:130, 1:131
 Ordered logit model, 4B:74
 Order management system (OSM), 3:77
 Organic self-assessing documents, 7B:4
 Organizational Safety Culture, 2:555
 Organization and Human Resources, 5:252
 Organization for Economic Cooperation and Development (OECD), 3:231, 4A:52
 Organization of Petroleum Exporting Countries (OPEC), 2:625
 Organized and acquisitive carjackings, 2:158
 Organized and instrumental carjacking, 2:158
 Original design manufacturers (ODM), 3:145
 Original map of the Dwight D. Eisenhower System of Interstate and Defense Highways, USA, 5:362
 Origin and Destination (O&D) yield management system, 1:407
 Origin-destination commuting trips by dockless bikes (sample: "Mobike") in Beijing, 6:47
 Origin-Destination (OD) Demand, 4B:109
 multiple sources of data-based model, 4B:111

- estimation models, 4B:111
 - machine learning method, 4B:112
 - traffic data, 4B:111
 - survey-based models, 4B:109
 - traffic-count-data-based models, 4B:110
 - dynamic, 4B:111
 - optimization method, 4B:110
 - static, 4B:110
 - statistical method, 4B:110
 - Origin-destination matrices estimation, 6:427
 - Origin-destination matrix estimation (ODME), 4B:166
 - Outdoor Power Equipment Institute (OPEI), 2:79
 - Outdoor site-lighting performance (OSP) system, 7B:72
 - Out-of-vehicle time (OVT), 1:122
 - Out-of-vehicle travel time (OVTT), 4B:194
 - Output of public transportation, 1:27
 - density, 1:28
 - economies of scale, 1:28
 - scope, 1:28
 - Outsourcing, 3:102
 - Outsourcing Process and Determinants, 3:103
 - design phase, 3:104
 - operation phase, 3:104
 - selection phase, 3:103
 - strategy phase, 3:103
 - Own-mode shippers, 3:232
- P**
- Packaging Design
 - organizational considerations in, 3:123
 - requirements, 3:120
 - Packaging Logistics, 3:124
 - evaluating design alternatives, 3:122
 - Packaging system
 - design requirements and, 3:119
 - structure of, 3:121
 - Painted island, 2:487
 - Panel Study of Income Dynamics (PSID), 6:175
 - Paper ticket and cash, 4A:329
 - Paradigm shifts and the resurgence of public transport, cycling, and walking, 6:121
 - Parametric t-test or ANOVA, 7A:83
 - Parcel lockers, 1:23, 6:322
 - Parcel Size Distribution, 3:258
 - Pareto criterion, 1:190
 - definitions concerning, 1:190
 - history, 1:191
 - transport applications of, 1:191
 - Pareto efficiency, 4A:100
 - Paris Agreement, 5:566
 - Paris Convention (1919), 3:361
 - Park-and-ride schemes, 4B:116
 - Parking, 1:149
 - Parking accumulation, 4B:113
 - Parking and long-term decisions 1:163
 - Parking at the workplace, 6:115
 - Parking Behavior Analysis Models, 4B:114
 - factors that influence parking location choice, 4B:114
 - parking location and travel choices, 4B:115
 - parking location choice models, 4B:114
 - Parking charge at the workplace, 6:115
 - Parking demand and the generalized cost of transportation, 1:160
 - Parking Demand Forecasting Models, 4B:113
 - Parking Demand Management Strategies, 4B:115
 - park-and-ride schemes, 4B:116
 - parking pricing, 4B:115
 - parking supply-side strategies, 4B:115
 - Parking demand prediction theory, 4A:285
 - Parking dwell time, 1:185
 - Parking Guidance Information System (PGIS), 4A:287, 4A:289, 4A:292
 - Parking guidance information technology, 4A:287
 - Parking index, 4A:303
 - Parking information service platform, 4A:287
 - Parking Information Systems (PIS), 4A:289
 - benefits and advantages of, 4A:296
 - effective parking management, 4A:296
 - enhanced customer satisfaction, 4A:296
 - improved traffic efficiency, 4A:296
 - categories of, 4A:292
 - deployment, 4A:290
 - features and functions of, 4A:293
 - access control, automation, and emerging features, 4A:296
 - intelligent algorithms, booking and customer management, 4A:294
 - objective of, 4A:290
 - requirements of, 4A:292
 - system architecture, 4A:290
 - communication technologies, 4A:292
 - data collection and processing, 4A:292
 - detection/monitoring, 4A:291
 - information dissemination, 4A:292
 - Parking interactions, 1:65
 - Parking levy schemes, 6:3
 - Parking Lots
 - background, 2:392
 - existing parking challenges, 2:394
 - automation impacts on vehicle parking, 2:397
 - parking space cruising impacts, 2:395
 - vehicle parking and cyclists, 2:395
 - vehicle parking and older adults, 2:396
 - vehicle parking and pedestrians, 2:396
 - vehicle parking policies and regulations, 2:394
 - parking maneuvers, 2:393
 - Parking management, 4A:1, 6:114
 - Parking, mode, and activity alternatives, 1:186
 - Parking occupancy, 4B:113
 - Parking occupancy monitoring technologies, 4A:293
 - Parking Performance, 4A:301, 4A:302
 - Parking price elasticities, 1:185
 - Parking regulation and its political economy, 1:162
 - Parking rights, allocation of, 1:163
 - Parking search time, 4B:114
 - Parking space levies (PSL), 6:3
 - Parking space sensing, 1:164
 - Parking space, types of, 6:113
 - Parking supply allocation, 4A:285
 - Parking support technology for new automobile industry 4A:287
 - Parking turnover, 4B:113
 - Parking volume, 1:185
 - Partial charging like toll roads, 1:64
 - Partial least squares regression (PLSR), 4A:130
 - Participatory decision-support models, 6:188
 - Particularly Sensitive Sea Areas (PSSA), 3:291
 - Particulate emissions (PM), 4A:52
 - Particulate matter (PM), 1:21, 6:214, 7B:44
 - Part-time shoulder use, 4A:41
 - Passenger conveyors, 5:690
 - Passenger Demand Forecasting Handbook (PDFH), 1:174, 1:182, 1:183
 - Passenger environment surveys (PES), 6:221
 - Passenger traffic, 5:409
 - Passenger traffic at the world's largest airports, 2010, 5:42

- Passenger traffic, management of, 5:451
 advantages for frequent travelers and targeted segments of the market, 5:452
 competition from other modes, 5:451
 considering options for door-to-door transport, 5:452
 effects of elasticities, 5:452
 increase of traffic or profit, 5:452
 monitoring of passengers' exigencies, 5:451
 optimization of finances, 5:452
 Passenger transit, 5:349
 Passenger transportation, 5:611
 cruise ship, 5:612
 ferry, 5:612
 ocean liner, 5:611
 yacht, 5:612
 Passenger transport planning, evolution of, 6:118
 Passenger Van Safety
 accident statistics, 2:402
 factors affecting, 2:403
 driver's behaviors, 2:403
 driving skills, 2:403
 vehicle design, 2:403
 vehicle safety tests, 2:403
 vehicle's maintenance and inspection, 2:403
 general safety precautions, 2:404
 safety regulations, 2:404
 use, 2:401
 Passive acoustic systems, 6:306
 Passive alcohol sensors (PAS), 2:45
 Passive Railway Crossing, 4A:341
 Passive safety, 6:57
 Passive Transit Priority, 4A:307, 4A:308
 Past personal experiences and habits, influence of, 5:387
 Patching, 2:531
 Path-based models, 4B:92
 Path choice, 4A:331
 Path choice behavior, analyzing and predicting, 4B:90
 deterministic *versus* stochastic models, 4B:91
 preliminaries, 4B:91
 Pattern Recognition Algorithm (PATREG), 4A:129
 Pavement Management Approaches, 4B:118
 Pavement Management System (PMS), 4B:119
 data collection, 4B:119
 inventory information, 4B:119
 pavement condition evaluation, 4B:119
 traffic counts, 4B:120
 data management, 4B:120
 engineering and economic analysis, 4B:121
 future directions, 4B:122
 performance modeling and prediction, 4B:121
 reporting and feedback, 4B:122
 Pavement markings, 2:646, 4A:57
 Payment mechanism, of road pricing, 4A:71
 Payment methods, classification of, 4A:328
 Payment modes and their properties, 4A:328
 Payment structure, 1:375
 contractible costs, 1:376
 contractible profits, 1:375
 contractible revenues, 1:375
 Peak car in big cities, 6:327
 Pedestrian and cyclist detection system (PCDS), 2:113
 Pedestrian Automatic Emergency Braking, 2:175
 Pedestrian cities, 5:48
 Pedestrian crashes, 2:284
 Pedestrian Crossing (crosswalk), 4A:346
 classification of, 4A:346
 countdown signals, 4A:353
 disabled people, 4A:352
 signalized crossing, 4A:346, 4A:348
 hawk crossing, 4A:351
 pedestrian crossing time, 4A:348
 pedestrian-vehicle separation controls, 4A:351
 pelican crossing, 4A:349
 puffin crossing, 4A:349
 scramble crossings, 4A:352
 unsignalized, 4A:348
 crosswalks, 4A:346
 zebra crossing, 4A:347
 Pedestrian crossings, 2:487
 Pedestrian crossings, 5:113
 Pedestrian issues into policy, 5:321
 Pedestrianized zones, 6:2
 Pedestrian level of service (PLOS), 5:322
 Pedestrian Light CONTROLLED (PELICON), 2:711
 Pedestrian Light Controlled crossings (Pelican crossing), 4A:349
 Pedestrian, protection zone for, 2:424
 Pedestrian Safety, 2:544
 children
 cognitive capacity and skills, 2:417
 countermeasures, 2:417
 crash characteristics, 2:416
 crash consequences, 2:416
 crash contribution, 2:416
 crash involvement, 2:415
 crash recording, 2:415
 general, 2:420
 causes of accidents, 2:422
 designing for, 2:425
 issues and potential solutions, 2:427
 need for, 2:420
 versus pedestrian security, 2:421
 protection zone for pedestrian, 2:424
 road user hierarchy, 2:424
 traffic devices for, 2:427
 older people
 built environment form and design, 2:431
 cognitive impairments, 2:430
 crossing the road, 2:433
 curb delay, 2:432
 decline in health condition, 2:429
 explorative behavior at the crossing environment, 2:432
 gap judgment and acceptance, 2:432
 physical impairments, 2:430
 road crossing behavior, 2:432
 sensory impairments, 2:429
 at railroad grade crossings, 2:473
 visually impaired
 challenges to independent travel, 2:436
 historical perspective, 2:435
 modern travel environment, risk in, 2:436
 orientation and mobility, 2:435
 Pedestrian safety, 6:421
 Pedestrians and bicyclists, designing for needs of, 5:333
 Pedestrian space, planning for, 5:323
 Pedestrians' Quality Needs (PQN), 2:108
 Pedestrian travel, 5:52
 Pedestrian User Friendly Intelligent crossings (Puffin crossing), 4A:349
 Pedestrian-vehicle separation controls, 4A:351
 Penalty Charge Notice (PCN), 4A:73
 Penalty points, 2:210
 Pennsylvania rumble strip, 2:550
 Pentagon, 2:241
 People with Disability (PwD), 2:571
 Per capita, 2:282
 Perceived costs, 5:388
 Perceived General Health, 7B:106
 Perceived private cost (PPC), 1:58
 Perceived risks, 2:494
 Percentage of eye closure (PERCLOS), 2:613
 Perception, 5:112
 Perception of accessibility, 5:35
 Perception-response time (PRT), 2:650
 Perfect price discrimination, 1:404

- Performance-Based Navigation (PBN), 4A:372, 4A:373
- Performance-Based Parking
Management
building performance indicators, 4A:303
land use performance indicators, 4A:302
transportation performance indicators, 4A:304
- Performance-based planning, 4A:300
- Performance measures (PM)
frameworks, 3:17
and measures, 3:16
- Performance-shaping factors in maintenance, 2:29
- Perimeter control, 4A:318
bimodal, 4A:319
as a substitute for dedicated bus lanes, 4A:318
- Peripheral vision, 2:333
- Permissive right turn on red (RTOR) signal, 2:708
- Permitted Phase, 4A:185, 4A:196
definition of, 4A:196
phase, signal group and stage, 4A:196
implementation, 4A:197
operational and safety performance of, 4A:200
types of, 4A:196
- Perpetual inventory method (PIM), 1:449
- Personal costs, 1:14
- Personal Digital Assistants (PDA), 2:111
- Personal flotation devices (PFD), 2:481
- Personality and self-congruency theory, 7A:82
- Personality factors, 2:18
- Personal mobility devices (PMD), 2:420
- Personal Positioning Systems (PPS), 7B:138
- Personal rapid transit (PRT) system, 6:200
- Personal security, 7A:196
- Personal values, beliefs, and norms, 5:385
- Personal watercraft (PWC), 2:478, 2:480
- Person-related contributors to aggressive driving, 2:18
- Pesticide Exposure, 7B:107
- PEV Charging for the Energy System, Hazards of, 1:561
- Phone hotlines, 4B:13
- Photo/Video Traffic Enforcement, 2:439
best practices, 2:442
driver behavior, 2:440
history of, 2:439
photo enforcement work, 2:439
roadway crashes, 2:441
- Physical activity (PA), 5:109, 7B:105
- Physical activity (PA) from cycling, 7B:169
- Physical distraction, 2:337
- Physical impairments, in older people, 2:430
- Physical Internet (PI), 1:23
- Physical Internet (PI), 3:481
container development, 3:486
data, 3:483
definition, 3:482
digital tracking of containers in, 3:484
encapsulation layers in, 3:482
legal, 3:485
logistics networks by, 3:485
origins, 3:481
physical components, 3:482
process, 3:483
routing, storage, and open services, 3:485
simulation, 3:485
- Physics, of head-on crashes, 2:311
road conditions, 2:311
severity of, 2:312
traffic conditions, 2:312
- Pigouvian tax, 1:63, 4A:91
- Pilot fatigue, 2:92
- Piped Traffic Flow Models, 4B:145
heterogeneous traffic flow, 4B:147
LWR model, 4B:145
nonequilibrium Models, 4B:147
piped traffic flow models, 4B:148
- Pipeline, 3:463, 5:646
capacity, 3:465
economics, 3:467
engineering, 3:465
and geopolitics, 3:469
safety and integrity, 3:469
- Pipeline integrity management systems (PIMS), 3:469
- Pipeline safety management systems (PSMS), 3:469
- Pipeline transport, taxonomy of, 5:651
- Piper Alpha oil platform explosion, 2:309
- Place, 4B:229
- Plan-Do-Check-Act road safety strategy, 2:527
- Planned behavior, theory of, 7A:137
- Planning for rail transport, 6:161
capacity, 6:162
customer experience, 6:164
purpose, 6:161
resilience, 6:164
revenue and other benefits, 6:164
rolling stock, 6:163
route planning, 6:161
stations and terminals, 6:164
- Planning in tourism travel, factors affecting accessibility and, 6:40
- affordability, 6:40
inaccessibility, 6:41
integration, 6:40
land use factors, 6:40
mobility, 6:40
prioritization, 6:41
transport demand, 6:40
transport management, 6:41
transport network connectivity, 6:40
user information, 6:40
- Planning proposals and possible scenarios, 5:309
- Planning the railway activity, 5:449
business plans and master plans, 5:449
planning as an essential prerequisite for management, 5:449
project management, 5:449
- Planning time, 4A:116
- Planning time index (PTI), 4A:116, 4A:110
- Planning tourism travel
current trends in, 6:41
role of transport in, 6:39
- Planning transport infrastructure, need for a timely perspective on, 6:395
- Plateau car, 6:324
demand saturation, 6:324
demographic change, 6:327
evidence for, 6:324
innovative technologies and, 6:329
- Plateau effect, 2:361
- Plausible worst case scenarios, 2:278
- Pleasure, 7A:83
- Pleasure of cycling, 2:109
- Plowing, 2:535
- Plug-In Electric Vehicle Policy (PEV), 1:496
PEV policy instruments, categorization of, 1:497, 1:498, 1:499
principal market barriers, 1:497
- Plug-in-hybrid electric vehicle (PHEV), 7A:183, 7B:148
- Plus lanes, 4A:41
- Pneumatic capsule pipelines
cold war of, 5:647
during the Victorian Age, 5:646
- Pneumatic passenger railway, 5:649
- Point of sale (POS) data, 3:71
- Poisson gamma, 2:726, 371
- Poisson-lognormal model, 2:371
- Poisson regression models, 2:370
- Polar Amplification, 3:350
- Polar Cases
labor supply is independent of road pricing, 1:222
strict complementary between peak hour car travel and labor supply, 1:222

- POLARIS, 4B:69
- Polar Silk Road (PSR), 3:495, 3:499
- policy, 3:349
- Police Accredited Traffic Officers (PATO), 3:203
- Police Impersonator, 2:160
- Policies, 2:34
- Policies for obtaining social optimality with external costs, 1:63
- Policy evaluation, 1:318
- commuting subsidies, 1:320
- congestion pricing, 1:319
- infrastructure investment, 1:319
- Policy goals and pricing rules, 1:57
- Policy instruments for plug-in electric vehicles, 1:496
- Policy-maker's vision, 1:470
- Policy process for different transport policy, role of media in, 6:400
- Political acceptance, 1:544
- Political boundary units, 2:368
- Political climate, 5:253
- Political will, 7A:134
- Politics of mobility policy, 6:131
- Population growth, 1:20
- Portable collision avoidance system (PCAS), 2:355
- Portable concrete barriers, 2:517
- Portable digital assistants (PDA), 4B:47
- Port as a land use, 5:540
- Port as a regional hub, 5:542
- Port authorities, 3:308
- Port Authority Trans Hudson (PATH), 2:241
- Port cluster, 3:311
- components of, 3:311
- particularities of, 3:312, 3:313
- and tech cluster, 3:313
- Port community system (PCS), 3:310, 3:460
- Port competition and the need for port reforms, 5:549
- Port concessions as a regulatory tool, 5:566
- Port development, 4A:394
- Port Effectiveness Measurement, 3:316
- Port Efficiency
- empirical analysis in, 3:320
- measurement, 3:316, 3:317
- frontier function, 3:318
- Port, environmental and sustainability challenges facing, 3:303
- Port Hinterlands, 3:305
- broader supply chains, 3:306
- developing, 3:307
- port authorities, 3:308
- rail transport operators, 3:308
- shipping lines, 3:307
- terminal operators, 3:307
- emergence of port regionalization, 3:308
- shaping, 3:306
- Port investment, trends in, 1:446
- Port labor, 3:302, 5:549
- Port management, 4A:390, 5:557
- central theme of, 4A:393
- encompasses strategy and operations, 4A:390
- evolution before the 1980 decade, 5:557
- models, 5:559
- Port Ownership Styles, 4A:391
- Port performance indicators (PPI), 3:317, 4A:396, 4A:398
- aforementioned, 4A:399
- hierarchy of, 4A:398
- Port performance measurement (PPM), 4A:397
- methodology, 3:317
- port performance indicators, 4A:398
- process and methodology, 4A:397
- in South Korea, 4A:400
- DEMATEL and ANP to analyze PPIs' interdependent weights, 4A:400
- FER, 4A:400
- Port planning, 3:301
- Port planning, 5:547
- Port planning and development process, 1:445
- alternative design, 1:446
- development plan, 1:446
- environmental impact assessment and other regulations, 1:446
- existing demand, supply, and pricing, 1:445
- final development project, 1:446
- financial analysis and investment plan, 1:446
- performance indicators, 1:445
- traffic evolution, 1:445
- Port planning and investment, 1:443
- terminal planning, 1:443
- Port reforms, 5:550
- Port regionalization, 3:308
- Ports and climate change, 5:543
- Ports, and their economic characteristics, 5:563
- Ports before containerization, 5:548
- Port security, 3:302
- Ports, future challenges to, 5:542
- Ports, governance, 5:564
- Ports insertion into the global networks: governance implications, 5:560
- Port sites, changing character of, 5:541
- Port structure, 5:558
- Port technology, trends in, 1:446
- Port terminal operations, 5:547
- Port, the interface between land and sea, 5:546
- Portugal (Médio Tejo), 4B:20
- Positive attitude of politicians toward CBA, 1:280
- Positive externalities in transport 6:193
- Positive train control (PTC), 2:462, 2:475
- Positive transport policy for older people, 6:60
- Postal units, 2:368
- Post-COVID times and road safety, 7A:165
- Postencroachment time (PET), 2:664
- Post opening project evaluation (POPE), 1:254
- Pothole, 2:530
- maintenance, 2:532
- Powered Two- and Three-Wheeler (PTW), 2:443
- crashes, 2:444
- vehicle characteristics, 2:446
- weather conditions, 2:446
- global distribution of, 2:444
- interventions related to e-bikes, 2:448
- interventions to improve, 2:447
- personal protective equipment, 2:447
- road user's illegal behavior, 2:447
- road environment associated with, 2:445
- junction type, 2:445
- mixed traffic, 2:445
- pavement surface quality, 2:446
- road infrastructure, 2:445
- road user characteristics, 2:444
- age, 2:444
- attitudes and driving behavior, 2:444
- gender, 2:444
- lack of conspicuity, 2:445
- nonuse of helmets, 2:445
- safer roads, 2:447
- exclusive motorcycles lanes, 2:447
- road design, 2:447
- roadside hazards, 2:448
- speed limits and traffic calming, 2:448
- safer vehicles, 2:448
- antilock brake systems, 2:448
- headlight on, 2:448
- traffic safety challenges with, 2:444
- usage, 2:444
- Powered two-wheeler (PTW), 2:113, 7A:31, 7A:144
- Power law, 3:337
- Poynton (United Kingdom) shared space in, 2:588
- Practical implications, 2:135

- Predefined AR for the Road Transport System, 2:3
- Predictive analytics, 4B:6
- Predictive validity, 2:605
- Pregnancy outcomes and complications, 7B:105
- Premarshalling, 3:332
- Premature descent alert (PDA), 2:170
- Prescription drugs, 2:226
- Prescriptive analytics, 4B:7
- Presidential Policy Directive (PPD)-8, 2:244
- Presignals, 4A:317
 - at intersections with multiple-lane approaches, 4A:317
 - at intersections with single-lane approaches, 4A:317
- Pretertiary school campus traffic circulation, 568
 - active mobility, 2:569
 - carpooling, 2:569
 - interface management, 2:573
 - motives, 2:572
 - private car, 2:568
 - school bus, 2:571
 - walking school bus, 2:570
- Preventing Impaired Driving Offenders From Repeating, 2:46
- Price cap formula, 1:465
- Price change and transfer of surplus, 1:357
- Price discrimination
 - case of the Australian airline market, 1:405
 - new business model enhances price discrimination, 1:405
- Price discrimination, 1:404
- Price elasticity of demand, 5:251, 5:255
- Price regulation, 1:598
- Price structures, 1:186
- Pricing aeronautical services, 5:230
- Pricing of public transport services, 1:100
- Pricing only freight (but not passenger) transport, 1:99
- Pricing Public Transportation, 4A:90
 - base case, 4A:95
 - marginal cost of auto/private transport, 4A:97
 - marginal cost pricing
 - of both auto/private transport and bus/public transport, 4A:98
 - of bus/public transport, 4A:98
 - role of a transport fund, 4A:100
- Pricing *versus* supply management, 1:186
- Primary packaging, 3:121
- Principal component analysis (PCA), 4A:131
- 3D Printing (3DP), 1:25
 - concept of, 3:471
 - emergence of, 3:471
 - locations, 3:475
 - needs of consumers, 3:473
 - production, clients and distribution, 3:474
 - societal impact, 3:475
 - techniques, 3:472
 - technology, 5:169
 - on transport and logistics, 3:472
 - transport resistance, 3:474
 - transport volumes and technologies, 3:475
- Priority measures, types of, 6:254
- Private car, 2:568
- Private finance initiative (PFI), 1:386
- Private investor, 1:374
- Private motorized mobility, 6:214
- Private service port, 4A:393, 5:560
- Private transportation, excessive use of, 4A:74, 4A:76
- Private vehicles, 6:2
- Privatization, 3:27
- Privatization, 1:366
- Proactive manner, 2:634
- Probabilistic measure, 4A:111
- Probabilistic methods, 4A:118
- Probabilistic Neural Network (PNN), 4A:130
- Procedural noncompliance, 2:29
- Process validity, 2:663
- Producer surplus, 1:242
 - current trend of, 1:247
 - definition of, 1:242
 - and density, 1:247
 - in highways (driving and road use), 1:244
 - and mode choice, 1:246
 - in public transit supply, 1:245
- Production function and the SolowüSwann Model of Regional Growth, 1:256
- Production planning and scheduling (PP&S), 3:82
- Product Limit Method (PLM), 4A:146
- Product supply chain, 3:10
- Product type, 3:9
- Programme for European Traffic of Highest Efficiency and Unprecedented Safety (PROMETEUS), 2:174
- Programs, 2:34
- Progression Analysis and Signal System Evaluation Routine (PASSER), 4A:173
- Project Finance
 - benefits and risk, 6:292
 - definition of, 6:292
 - for innovative transport projects, 6:296
 - for transport projects
 - in practice, 6:295
 - in theory, 6:295
 - for transport projects, 6:294
- Promising CAV applications for buses and trucks, 5:310
- Propeller strike, 2:480
- Property damage only (PDO), 2:371
- Prospect theory, 5:60
- Protected bicycle phase, 4A:188, 4A:192
- Protected left turn phase, 4A:185, 4A:188
- Protected-only phase, 4A:185
- Protected pedestrian phase, 4A:186, 4A:192
- Protected Phase, 4A:185
 - definition of, 4A:185
 - implement, 4A:188
 - operational and safety performance of, 4A:192
 - types of, 4A:185
- Protected public transport phase, 4A:186, 4A:189
- Protected right-turn phase, 4A:186, 4A:189
- Protection of the Arctic Marine Environment (PAME), 3:349
- Proximity services (ProSe), 2:184
- P + R (Park-and-Ride) parking facilities, 4A:285, 4A:286
- Pseudo-Monte Carlo (PMC) simulation, 4B:73
- Psychological attachment, 7A:83
- Psychological-cognitive fidelity, 2:602
- Psychological countermeasures, 2:21
- Psychological determinants of modal choice within a multidisciplinary approach, 5:64
- Psychological learning theory, 2:264
- Psychological traffic calming, 2:587
- Psychomotor vigilance test (PVT), 2:614
- Public, 6:267
- Publication bias, 2:223
- Public authorities and private investors, relationship between, 1:374
- Public bike share (PBS), 6:19
- Public Bike-sharing Systems (PBSS), 4A:355
 - success factors for, 4A:358
 - information and education 4A:360
 - integration with other modes of transport, 4A:359
 - provision of infrastructure, 4A:359
 - service regulation, 4A:360
 - use of new technologies, 4A:360
- Public communication, 6:266
- Public consultation, 6:267
- Public decision making, 1:4

Public education, advertisement, and mass media campaigns, 2:272

Public engagement, 6:266

- advantages of, 6:267
- aims and objectives in transport planning seek to achieve, 6:267
- challenges of, 6:268
- limitations of, 6:268
- prospects and Research Agenda for, 6:270
- rationale for, 6:267
- successful, public engagement look like, 6:269
- in transport planning, 6:266
- used in transport planning to date, 6:269

Public finance *versus* private finance, 1:373

Public financing, 5:355

Public intervention in the financing of transport infrastructure, 1:371

- arguments, 1:371
- efficiency, 1:371
- equity, 1:371
- stabilization policy, 1:371

Public make-or-buy decision, 1:302

Public ownership *versus* private ownership, 1:377

Public policy decisions, 1:111

Public private partnership (PPP), 1:374, 1:385, 3:95

- manage, 1:386
- contract regulation, 1:387
- procurement organization, 1:386
- risk, 1:387
- in transport sector, 1:388
- in ports, 1:389
- in road transport, 1:388
- usages, 1:385

Public sector, 1:374

Public sector equality duty (PSED), 6:364

Public service ports, 4A:391, 5:559

Public shelter, 2:277

Public transit (PT), 1:64, 1:512

Public transit pricing, 1:267

Public transit subsidization, 1:269

Public transit with an SAV feeder (PTAV), 1:512

Public transport, 1:267, 1:365, 3:206, 6:254, 7B:38

Public Transportation, Sexual Violence in

- care, 2:576
- emerging issues and future debate, 2:581
- impact of, 2:581
- nature of, 2:576
- people most affected

- age and disability, 2:579
- ethnicity, 2:579
- gender, 2:577
- LGBTQI status, 2:578
- previously victimized, 2:579
- socioeconomic status, 2:580
- underreporting, 2:580
- prevention of, 2:581
- selection of international evidence, 2:580
- situational risky conditions for, 2:580

Public transport attractiveness, 1:492

Public transport authority (PTA), 1:332

Public Transport Fare Subsidy Scheme (PTFS), 4A:104

Public transport network planning, 6:388

- areas of low density and demand, 6:392
- issues in planning, 6:392
- objectives and purpose of network planning, 6:388
- objectives for public transport, 6:389
- principles of planning, 6:389
- role of economics and its impact on network planning, 6:389

Public transport networks, methods for designing, 5:625

- exact method approaches, 5:626
- existing methodologies, limitations of, 5:627
- heuristic and metaheuristic approaches, 5:626
- typical methodological procedures, 5:626

Public transport performance, 1:365

Public transport, procurement of, 1:308

Public transport provider, 5:17

Public transport, regulation of, 6:351

- main forms of regulation of, 6:352
- problems with regulation, 6:353
- qualitative regulation, 6:351
- quantitative regulation, 6:352

Public transport-related big data, characteristics of, 5:134

Public transport service quality in terms of passengers' perceptions, analysis of, 6:221

- new challenges, 6:222

Public transport (PT) systems planning, 6:198

Public transport (PT) systems planning and operations, 6:198

Public transport (PT) systems, usages of vehicle automation in, 6:199

Public transport with automated vehicles, planning for, 6:202

- potential impacts on urban mobility, 6:201
- transport modeling studies, 6:201

- travelers' behavior, 6:201

Public use microdata area (PUMA), 4B:193

Publishing in high impact scientific journals *versus* popular science magazines, 7A:110

Punctuality, 1:434

Pupillometry, 2:608

Purchasing power parities (PPP), 1:306

Purchasing power parities (PPP) raise funding, 1:306

Purchasing social responsibility (PSR), 3:67

Pure e-tailers, 3:219

Push for DB Contracting to Enhance Productivity, 1:305

Push-pull/pull push train formations, 5:497

P-value, 7A:81

Q

QR Code tickets, 4A:329

Quadbikes, 2:77

Qualitative research data, 7A:107

Quality-adjusted life year (QALY), 7B:123

Quality lapses in maintenance, 2:27

Quality-related factor, 1:465

Quantitative and qualitative research, General differences between, 7A:40

Quantitative risk assessment (QRA), 2:308

Quasi-Monte Carlo (QMC) simulation, 4B:73

Quay crane assignment problem (QCAP), 3:330

Quay crane (QC) operations, optimizing, 3:330

- quay crane assignment, 3:330
- quay crane scheduling, 3:330

Quay crane scheduling problem (QCSP), 3:330

Questionable research practices, 7A:84

Questionnaire research, 7A:82

Questionnaires, 1:216

Queue Jump Lane (QJL), 4A:311

Queue length, 4A:63, 4A:263

- as an evaluation index, 4A:266
- defined, 4A:263
- measure, 4A:267
- methodologies to measure, 4A:265
- in physical distance, 4A:264
- theoretical explanation of, 4A:263
- at traffic signal, 4A:266

R

Racial and ethnic differences in travel patterns, 7A:196

- Radars, 2:176
- Radial basis function neural network (RBFNN), 4A:130
- Radio Frequency Identification (RFID), 4A:289
- sensors, 4A:292
- technology, 4A:329
- Radio Frequency Identifications (RFID), 3:71
- Rafting, 2:480
- Rail and road costs by distance for a given payload, 5:411
- Rail cost function, 1:425, 1:426
- infrastructure provision costs, 1:429
- modeling approaches for railway costs, 1:426
- specification of, 1:427
- transport operations costs, 1:428
- Rail freight, 3:413
- commodities, 3:415
- fundamentals of, 415
- infrastructure and service provision, 3:417
- journey, components of, 5:411
- rail's share, 3:413
- in the 21st Century, 3:417
- changing nature of supply chains, 3:417
- competitive markets, 3:420
- information and communication technologies, 3:418
- innovation, 3:420
- operational efficiency improvements, 3:419
- sustainable, 3:418
- transport, 1:53
- transportation system, 5:465
- vehicles
- driverless freight train operation, 3:433
- Freight wagon classification and design, 3:430
- locomotive classification and design, 3:423
- locomotive power traction systems, 3:427
- tractive effort and dynamic braking characteristics, 3:429
- train configuration and distributed power, 3:432
- worldwide, 3:414
- Rail industry productivity, approaches to measure, 1:504
- Rail market share on the rail + air market in France, 1:421
- Rail reforms in Europe, empirical assessment of, 1:505
- Rail restructuring process, 5:486
- Railroad accident causes, 2:454
- collision, 2:456
- derailment, 2:454
- Railroad grade crossings, accidents at, 2:468
- Railroad Revitalization and Regulatory Reform Act (4-R Act), 3:26
- Railroad Safety, 2:466
- emerging technologies to improve, 2:474
- connected vehicles, 2:475
- intelligent transportation systems in railroads, 2:474
- positive train control, 2:475
- grade crossing elimination, 2:471
- grade separation, 2:472
- highway-rail grade crossings, safety at, 2:466
- accidents at railroad grade crossings, 2:468
- active control devices, 2:469
- crossbucks, 2:468
- flashing lights, 2:468
- gates, 2:468
- interconnection and preemption, 2:471
- pre-signal, 2:471
- traffic control devices at railroad grade crossing, 2:468
- human factor-relation strategies, 2:462
- engineer education and training, 2:464
- positive train control, 2:462
- modeling safety at, 2:472
- pedestrian detection, detection technologies for, 2:475
- pedestrian safety at railroad grade crossings, 2:473
- potential grade crossing safety contributing factors, 2:472
- risk mitigation strategies, 2:457
- acoustic emission-based detection, 2:462
- improved materials and manufacturing, 2:460
- machine vision-based inspection, 2:461
- manual inspection, 2:460
- rail inspection, 2:458
- track geometry measurement and analytics, 2:459
- statistics in the EU and Japan, 2:452
- statistics in the United States, 2:452
- trespassing and pedestrian safety, 2:472
- Rail Safety Improvement Act (RSIA), 2:462
- Rail transport, 5:406, 5:413
- debut: Stockton-Darlington, 5:413
- distances, travel times, and intermodal competition, 5:432
- first developments, 5:414
- geo-economic factors *vs.* economic history, 5:429
- journey by rail, 5:416
- local and the national, 5:414
- maturity phase, 5:415
- micro-level analysis: first/last-mile links, 5:441
- organization, 5:411
- overall geography of, 5:427
- physical geography, 5:431
- political factors, 5:433
- as a production of multiple factors, 5:429
- railway extension, 1820-2010 world and selected countries, 5:421, 5:423
- railway logics, 5:433
- regulation, standardization, and international agreements, 5:415
- resurrection of rail transport, 5:417
- revolution or continuity, 5:413
- technological and labor issues, 5:414
- territorial effects of, 5:437
- territorial equity effects, 5:439
- 20th century decline, 5:416
- urban systems, 5:432
- world: new constructions, 5:422
- world-railway density, 5:418, 5:420
- Rail transport, experience in, 1:308
- Rail transport externalities, 1:54
- Rail transport, of hazardous materials, 2:305
- Rail transport operators, 3:308
- Rail vehicle elementary classification, 5:491
- Railway company management, 5:479
- comparison of the models, 5:482
- integrated management model (vertical integration), 5:481
- management models, 5:480
- separated management model (vertical separation), 5:480
- Railway Crossing, 4A:341
- active, 4A:341
- design, 4A:343
- passive, 4A:341
- policy, 4A:343
- education and enforcement, 4A:344
- knowledge management, 4A:344
- risk management, 4A:344
- technology, 4A:344
- safety, 4A:342
- Railway diversification reframes terminal regulation, 5:455

- Railway infrastructure, management of, 5:450
 condition-based maintenance, 5:450
 deciding efficient values of infrastructure charges, 5:450
 level of outsourcing, 5:450
- Railway management, 5:444
 artificial intelligence, 5:446
 dealing with a heavy and rigid industry, 5:445
 environmental sensitivity, 5:446
 internal and external environment of railways, 5:446
 lifestyle, 5:446
 railways: engineering oriented or customer oriented, 5:445
 systems approach for railways, 5:446
- Railway noise, 7B:55
- Railway noise standards, 7B:56
- Railway station and terminal charging structure and most used criterias in pricing methods, 5:461
- Railway stations, 5:399
 functional specification of passenger railway station, 5:401
 geometric design, 5:402
 optimization models aim to, 5:404
 platform arrangement and track layout of, 5:401
 simulation of passenger flows in, 5:403
 timetabling, 5:402
- Railway suicides, 2:656
 influencing factors, 2:657
 prevention, 2:658
 statistics/incidence, 2:656
 types of events, 2:657
- Railway terminal pricing principles in Europe, 5:457
- Railway terminal regulation goals, 5:454
- Railway traffic management, 5:678
 examples of flexibility in planning, exploited by real time, 5:680
 mathematical models for, 5:681
 state of industry and application of, 5:681
 time scales and degrees of freedom of the different problems, 5:679
- Railway transport systems classification, 5:498, 5:497
 elementary classification of railway systems, 5:497
 secondary classification of railway systems, 5:497
 typical rolling stock characteristics per railway system, 5:499
- Raised crosswalk, 2:644
- Raised intersection, 2:644
- Raised medians, 2:645
- Ramp design, entry and exit, 4A:28
 design process, 4A:30
 performance-based design, 4A:29
 physical and ITS design elements, 4A:31
 storage and discharge capacity for design, 4A:31
- Ramp handling agent, 2:101
- Ramp metering, 4A:8
 application, 4A:10
 fine-grained spatiotemporal data, 4A:12
 performance measurement, 4A:10
 space resolution of data, 4A:12
 vehicle event data, 4A:12
- Ramp operations, 2:94
- Ramsey-equation parameters, 1:199
- Random Breath Testing (RBT), 2:42, 2:47, 2:45, 2:271
- Random regret, 5:60
- Random Utility Maximization (RUM), 4B:72
- Random utility models for mode choice, 5:59
- Random utility theory (RUT), 6:222
- Range-extended electric vehicles (REEV), 7A:183
- Range stress, 7A:169
- Rapid urbanization and transition to automobile Use, 6:120
- Rationale for state intervention in public transport markets, 6:349
 Mohring effect, 6:350
 participative requirement/transport poverty, 6:349
 reduced environmental impact of mobility, 6:350
 safety, 6:350
 use of land space, 6:350
- Real-life heterogeneity and its effects on the α , β , and γ parameters, 1:152
 further influences on value of time, 1:153
 income, 1:152
 trip purpose, 1:152
- Real-time Hierarchical Optimizing Distributed Effective System (RHODES), 4A:174
- Real-time transit information systems (RTTIS), 4A:331
 future research, 4A:336
 information demand/supply trade-off of, 4A:332
 social effects of, 4A:335
 on transit service, 4A:333
- Real-Time Transit Priority, 4A:307, 4A:309
- Reappraisal (RE), 7A:204
- Rear Cameras and Parking Assist, 2:176
- Rear Cross Traffic Alert (RCTA), 2:175
- Rebates, 1:516
- Rebound effect, 1:174, 1:461, 2:296, 3:402, 3:403
 direct effect, 1:174, 1:176
 empirical estimates of, 1:176
 estimating effect, 1:176
- Recent and projected costs of various electrofuels, low/medium/high cases, 7B:131
- Recovery supply chain, 3:13
- Recreational Boating Safety epidemiology of death and causes of injury, 2:479
 Australia, 2:479
 canoeing, kayaking, and rafting-related injuries, 2:480
 carbon monoxide poisoning, 2:480
 injuries, 2:480
 personal watercraft, 2:480
 propeller strike, 2:480
 small boats, 2:480
 Sweden, 2:479
 United States and Canada, 2:479
- exposure, 2:478
 Australia, 2:479
 Canada and Europe, 2:478
 locations and hazards, 2:479
 United States, 2:478
- risk factors, 480
 alcohol, 2:481
 excessive speed, 2:481
 improper lookout, 2:481
 injury and fatality risk factors, 2:481
 life jackets, 2:481
 operator inexperience, inattention, and poor judgment, 2:481
 overloading, 2:482
 rules violation, 2:482
 weather, 2:482
- strategies to improve safety, 2:482
 legislation and regulatory strategies, 2:482
 safe boating strategies, 2:482
 social psychological and behavioral strategies, 2:483
- Recreational Off-Highway Vehicle Association (ROHVA), 2:79
- Recreational off-road vehicles (ROV), 2:77
- Recreational travel, 4B:223
- Recreational utility vehicle (RUV), 2:77
- Recruitment, 7A:42
- Recursive crossnested logit (RCNL), 4B:93
- Recycling supply chain, 3:13
- Red Flag Acts, 2:487
- Red light, 4A:178
- Red signal violation, 2:706
- Reduced forms, 1:572

- auction *versus* negotiation, 1:575
 contract choice endogenous, 1:574
 contract choice exogenous, 1:575
- Reducing transport energy
 consumption, 5:78
 air and maritime transport, 5:79
 air transport, 5:80
 behavioral changes, 5:81
 land transport, 5:78
 maritime transport, 5:81
 regulations, 5:78
 road transport, 5:80
 technology, 5:80
- Redundant signs, 2:653
- Refined model DBB contracts, 1:303
- Reform of the public sector, 1:368
- Refuge Islands
 centerline rumble strips, 2:491
 median barrier on 13-m wide roads, 2:491
 medians with guardrails, 2:490
 pedestrian crossings, 2:487
 traffic medians, 2:488
 effect on mobility, 2:490
 median strip at intersections, 2:489
 median strip instead of two-way left turn lanes, 2:489
 type of median strip, 2:490
 width of the median strip, 2:489
- Regional concept of transport planning, 6:286
 defining a region, 6:286
 origin of regional problems and regional policy, 6:286
 role of transport in regional development, 6:287
- Regionalization, 2:369
- Regional (Supranational) Shippers' Associations, 1:440
- Regional traffic flow management, 4A:366
- Regional transport planning
 key themes of implementation of, 6:287
 delimitation, 6:287
 institutional reform and governance, 6:289
 integrated transport planning, 6:288
 planning techniques and goals, 6:288
- Regional transport planning, 6:286
- Registered vehicles, 2:283
- Regret-based multinomial logit model, 4B:74
- Regulations, 2:34
- Regulatory Asset Base (RAB), 1:466
- Regulatory reforms and competition in public transport, future of, 1:368
- Relationship to well-being
 spill-over effects on emotional well-being, 7A:180
 travel-related emotional well-being, 7A:180
 travel-related life satisfaction, 7A:180
- Relationship to well-being, 7A:180
- Relative product validity, 2:663
- Relative Risk of Being Involved in a Crash by BAC, 2:40
- Relaxation, 7A:83
- Relevance of the territorial scale in air transport, 5:208
- Relevant regulatory issues in the rail industry after the reforms, 5:486
 guarantee cost recovery, 5:487
 promote competition, 5:488
 quality regulation, 5:488
 regulate rail services, 5:486
- Reliability, 1:127
- Relocation of economic activity, estimation
 integration of broader approach with partial equilibrium appraisal methods, 1:474
 land Use-transport interaction (LUTI) models, 1:474
 spatial computable general equilibrium (SCGE) models, 1:473
 using general equilibrium and land use-transport interaction models, 1:473
- RELU-TRANS General Equilibrium model, 1:93
- Remaining service interval (RSI), 4B:123
- Remote areas, 5:51
- Reproducibility, 1:543
- Required-Times-of-Arrival (RTA), 4A:370
- Reservation-based Smart Parking System (RSPS), 4A:289, 4A:293
- Reservation systems, 1:23
- Reserved bus lanes, 5:328
- Reservoir Traffic Flow Models, 4B:148
- Residential Location Choice Models
 empirical choice modeling, 4B:126
 choice set formation, 4B:126
 data sources for estimation, 4B:128
 decision-makers, 4B:127
 decision rules and model form, 4B:128
 joint choices, 4B:127
 population strata, 4B:127
 further research, 4B:129
 implementation in transport models, 4B:129
 theoretical foundation, 4B:125
- Residential location in economy closed in population, 1:341
- Residential location in economy open to population, 1:339
- Residential parking, 6:116
- Residual sum of squares (RSS), 4B:169
- Resilience, 2:276, 3:53
- Resilience, 7B:35
- Resilience and related concepts, 7B:34
- Resilient evacuation planning, 2:280
- Resilient transport infrastructure, strategy for developing, 6:397
- Resolution advisory (RA), 2:164
- Resource supply chain, 3:13
- Response mechanism to noise exposure, 7B:88
- Response Surface Method (RSM), 2:348
- Responsible traveller, 5:17
- Restructuring process in the European rail industry, 1:503
 restructuring measures at horizontal dimension, 1:504
 restructuring measures at vertical dimension, 1:503
- Retail parking, 6:116
- Retrofit cities and change mode share reasons for transitions, 7B:19
- Retrofit cities and change mode share, 7B:19
- Retroreflectivity, 2:652
- Return logistics, 3:188
- Return on net assets (RONA), 3:115
- Returns, types of, 3:211
- Revealed preference (RP), 1:131, 1:216
 methods, 1:117
 risk, 2:3
 and stated preference methods, 3:234
 surveys, 4B:100
- Revenue Data, 2:88
- Revenue management, 5:252
- Revenue recycling, 1:139
- Reverse Commuting, 4B:62
- Reversed causality and simultaneous decisions, 1:92
- Reverse factoring (RF), 3:116
- Reverse Logistics (RL), 3:208, 3:221
 activities, 3:209
 background, 3:221
 capturing value, 3:222
 centralized returns centers, 3:211
 changing policies, 3:210
 costs, 3:221
 defined, 3:209
 digitalization, 3:223
 disposition decision-making, 3:212, 3:215
 disposition options, 3:212
 environmental issues, 3:222
 financial issues, 3:214
 food recalls, 3:213

- Reverse Logistics (RL) (*Cont.*)
 future developments in, 3:223
 lowering transaction risk, 3:210
 metrics, 3:216
 omnichannel synergies, 3:211
 retail return rates, 3:210
 secondary market flow, 3:215
 secondary markets, 3:214
 secondary market size, 3:215
 stock visibility and availability, 3:222
 strategic importance of drains, 3:209
 total US secondary market, 3:216
 types of returns, 3:211
 unsaleables, 3:214
- Reversible lanes, 4A:52
 advantages of, 4A:55
 control, 4A:56
 pavement markings, 4A:57
 physical separations, 4A:57
 signs and signals, 4A:56
 disadvantages of, 4A:56
 future research, 4A:58
 lane-use control signals for, 4A:57
 operational and safety issues of, 4A:56
 signs used in, 4A:57
 state-of-the art, 4A:54
 disordered traffic, 4A:55
 implementation, operation, and safety, 4A:55
 ordered traffic, 4A:54
 simulation and optimization, 4A:54, 4A:55
 warrants for, 4A:53
 AASHTO green book, 4A:54
 FHWA, 4A:54
 ITE, 4A:54
- Reversing social perceptions, 5:597
- Review evidence on cross-elasticities, 1:205
- RFid and radio wave, 6:306
- Rhino, 4B:36
- Ricardo's model, 1:258
- Ride hailing, 4B:158
 automation, 4B:160
 disruptiveness, 4B:159
 drivers, 2:700
 regulation, 4B:159
 service, 2:277
 travel demand, 4B:159
- Ride-pooling, 1:593
- RiderCourse ATV safety training, 2:79
- Rider Practices and Training, 2:78
- Ridership, 4A:331, 4B:53
- Ride-sharing, 593
- Ridesourcing, 7A:187, 7A:189
- 7 Rights, 3:1, 3:16
- Right-turn-on-reds (RTOR), 4A:200
- Rigid barriers, 2:598
- Risk, 2:36
 compensation hypothesis, 2:497
 taking, 2:13
- Risk-adjusted rates, 1:200
- Risk classification system (RCS), 2:615
- Risk perception and demand for risk
 mitigation in transport, relations between, 7A:76
- Risk perception and driving behavior, associations between, 7A:75
- Risk perception and sustainable transport mode use in urban areas, 7A:77
 active mode use in urban areas, 7A:77
 general mode use in urban areas, 7A:77
 risk perception and mode use on school trips, 7A:78
- Road Accident Risk and Drugs Use, 2:221
 choice of estimator of risk, 2:222
 controlling for confounding factors, 2:222
 determining dose of drug, 2:221
 illicit and prescription drugs, 2:224
 illicit drugs, 2:225
 prescription drugs, 2:226
 publication bias, 2:223
 study quality assessment, 2:224
- Road and travel in literature and popular culture, 5:307
- Road as metaphor and symbol, 5:307
- Road conditions
 and head-on crashes, 2:311
 various, 2:529
- Road crossing behavior, in older people, 2:432
 crossing the road, 2:433
 curb delay, 2:432
 explorative behavior at the crossing environment, 2:432
 gap judgment and acceptance, 2:432
- Road deaths
 change in, 2:526
 risk factors contribute to, 2:270
- Road design, 7A:133
- Road Diet, 2:645
 benefits of, 2:503
 ease of implementation, 2:504
 livability benefits, 2:504
 operational effects, 2:503
 safety benefits, 2:503
 complete streets, 2:503
 configuration, 2:500, 2:502
 costs, 2:504
 criteria for, 2:505
 history of, 2:500
 political context and impediments, 2:505
- Road electric vehicle and charging technology, 5:147
- Road environment, 6:56
- Road freight transport, 1:51
 determinants of, 3:397
 average fuel consumption, 3:399
 average length of laden trips, 3:399
 average load on laden trips, 3:399
 empty running, 3:399
 fuel CO₂ content, 3:400
 gross domestic product, 3:397
 modal split, 3:398
 value density, 3:398
 rebound effect in, 3:402, 3:403
 historical background, 3:402
 policy aspects, 3:405
- Road infrastructure and land-use, interaction between, 5:317
- Road infrastructure, characteristics of, 5:360
- Road infrastructure, impacts of, 5:367
- Road infrastructure implications, of crash patterns, 2:236
- Road infrastructure, magnitude of impacts and spillover effects of, 5:316
- Road infrastructure, management of, 5:365
- Road infrastructure, planning for, 5:363
- Road network, 1:211
 complex networks, 1:212
 dynamic models, 1:213
 vulnerability, 5:316
- Road pricing, 1:147, 4A:74
 economics theory of, 4A:68
 impact on environmental quality, case of London, 5:380
 in London, 4A:72
 on pollution, to evaluate the impact of, 5:379
 purpose and types of, 4A:70
 in Singapore, 4A:72
 in Stockholm, 4A:73
 studies, early, 4A:69
 technologies for, 4A:70
 ensuring compliance, 4A:71
 payment mechanism of, 4A:71
 unique identification of vehicles, 4A:71
 welfare impact of, 4A:78
- Road rage and aggressive driving, 7A:122
- Road, railway, and aircraft noise calculation methods, 7B:59
- Road safety advertising
 factors that underpin effective, 7A:165
 research priorities going forward, 7A:168
 role of, 7A:165
 step approach to message design and testing (SatMDT), 7A:166, 7A:168
- Road safety and macroscopic developments, 6:51
- Road Safety Audit (RSA)

- background, 2:508
- benefits and costs of, 2:509
- case studies, 2:511
 - active work zone safety audit, 2:511
 - road weather safety audit, 2:511
 - transit access safety audit, 2:511
- history and evolution of, 2:508
- in road project's life cycle, 2:510
- steps to perform, 2:509
- tools, 2:510
- Road safety, future perspectives, 6:57
- Road Safety Improvement Through Summer Road Maintenance, 2:532
 - pavement crack management, 2:532
 - pothole maintenance, 2:532
 - rut maintenance, 2:532
 - surface friction management, 2:533
- Road safety management, 6:54
 - in Australia, 2:519
 - in Canada, 2:521
 - in The Netherlands, 2:522
 - in Sweden, 2:523
 - in United Kingdom, 2:524
 - in United States, 2:525
- Road safety problem, 6:51
- Road safety programs, for LMIC, 2:273
- Road safety risk, 6:53
 - risk and exposure, definitions of, 6:53
 - risk assessment, confounding factors in, 6:54
 - risk assessment through traffic conflicts and surrogate safety measures, 6:54
- Road Safety Standards and Services Directorate (RSSS), 2:524
- Road Safety Strategic Partnership Group (RSSPG), 2:525
- Road Safety Strategy (RSS), 2:271, 2:522
- Road safety system, 6:52
- Road safety, three E's of, 2:255
- Roads and railroads impact on fish and wildlife, 7B:73
 - effects on habitat, 7B:73
 - road mortality, 7B:74
- Roads, Devolution, and Motoring Group (RDM), 2:524
- Road sector, preventive measures in, 2:557
- Roads frequently taken, 5:308
- Roadside, 2:593
 - hazards, 2:448
 - obstacles, 2:595
 - sensors, 2:693
- Roadside Safety Barriers (RSB), 2:513
- Roadside unit (RSU), 2:112
- Road striping
 - and lane, 2:648
 - types of, 2:649
- Road suicides, 2:659
- Road Traffic Crashes
 - behavioral risk factors, 2:270
 - encouraging safer road use, 2:271
 - legislation and enforcement, 2:271
 - public education, advertisement, and mass media campaigns, 2:272
 - road safety strategies, 2:271
 - epidemiology and scope, 2:269
 - low- and middle-income country contexts, 2:273
 - legislation and enforcement, 2:273
 - public education, advertisement, and mass media campaigns, 2:274
 - road safety programs, 2:273
- Road traffic crashes, 6:51
- Road traffic death, 7B:8
- Road traffic, fundamental diagram of, 4A:75
- Road traffic noise, 7B:54
- Road traffic noise standards, 7B:55
- Road traffic regulation, 3:206, 5:378
- Road traffic system, 2:347
- Road Transport, 3:232
- Road transportation
 - long-run equilibrium of, 4A:84
 - optimal capacity, 4A:84
 - relaxation of assumptions, 4A:88
 - short-run equilibrium of, 4A:83
- Road transport externalities, 1:54
- Road transport, future perspective, 5:306
- Road transport, of hazardous materials, 2:305
- Road transport planning at the urban scale, 5:373
 - evolving transport modes, 5:373
 - human-scaled access, 5:375
 - performance measures, 5:375
 - progressions, 5:375
 - re-boot, 5:376
- Road transport safety and health, 5:308
- Road Tunnels
 - accidents and fires in, 2:713
 - challenges in, 2:716
 - deviating provisions, 2:717
 - human factor, 2:716
 - organization and system degradation, 2:717
 - fire safety in, 2:715
 - scenario-based approach, 716
 - vehicles carrying dangerous goods, 2:716
 - recommendations, 2:717
- Road user, 6:55
 - children and seniors, 2:125
 - hierarchy, 2:424
 - the vehicle, and the infrastructure, 2:125
- Road User Licensing Insurance and Safety Division (RULIS), 2:524
- Road vehicle driver licensing
 - requirements, for visual acuity, 2:333
- Roadway, 2:593, 4A:52
 - clearance time, 4A:123
 - crashes, 2:441
 - data, 2:560
 - lighting, 2:364
 - practices, 2:365
 - and safety, 2:364
 - pavement conditions
 - combined maintenance strategies, 2:536
- Roadway system, 4B:27
- Road weather information systems (RWIS), 4B:13
- Road weather safety audit (RWSA), 2:511
- Robustness, 5:35
 - of CBA by planning and decision-making stages, 1:251
 - of CBA, contextual, 1:250
- Rollover protection systems (ROP), 2:79
- Root Mean Square Error (RMSE), 4B:195
- RoPax fast ferry, 2:291
- ROPO effect, 3:184
- Roundabouts, 2:645, 4A:227
 - in The Netherlands, 4A:228
 - safety of
 - bicyclist safety, 2:545
 - flower roundabout, 2:541
 - modern roundabouts, 2:539
 - pedestrian safety, 2:544
 - turbo roundabout, 2:543
 - signalized, 4A:227
- Route mapping of shared bikes vis-a-vis share of female employment, 5:661
- Royal National Institute of Blind People (RNIB), 7B:138
- RTI S Display Boards, 4A:331
- Rubbernecking, 4A:158
- Rule-of-a-Half, 1:240
- Rumble bars, 2:313
- Rumble strip, 2:313
 - designs and placements of, 2:549
 - on driver behavior and driver performance, 2:550
 - impact of, 2:551
 - traffic safety, 2:551
 - intention with, 2:549
 - negative effects of, 2:552
 - external noise, 2:552
 - maintenance, 2:553
 - safety and comfort for vulnerable road users, 2:552
 - surface, 2:552
- Run-flat tire, 2:683

Run-off-road crash, [2:593](#)
 Runway excursions (RE), [2:105](#)
 Runway Incursion Avoidance, [2:353](#)
 Rural Highways, Speed Limits on, [2:622](#)
 current practices in setting speed limits, [2:630](#)
 differential *vs.* uniform limits for trucks and buses, [2:627](#)
 relationship between speed and safety, [2:623](#)
 speed limit policy in the United States, [2:625](#)
 speed limits effects on actual speeds, [2:628](#)
 two-lane rural highways and speed limits, [2:629](#)

Rural logistics, drones usage in, [3:375](#)
 Rural tunnels, [2:713](#)
 Russian Association of Indigenous Peoples of the North, [3:349](#)
 Rutting, [2:530](#)

S

Saami Council, [3:349](#)
 Saccadic eye movement patterns, [2:608](#)
 Safe boating strategies, [2:482](#)
 Safe Crossing Guidance, [2:111](#)
 Safe driving, [2:229](#), [234](#)
 Safe operating speed, [2:635](#)
 Safe operations of civilian aircraft, [5:266](#)
 Safe Route Guidance, [2:111](#)
 Safer road use, encouraging, [2:271](#)
 legislation and enforcement, [2:271](#)
 public education, advertisement, and mass media campaigns, [2:272](#)
 road safety strategies, [2:271](#)
 Safer System Roads Infrastructure Programs (SSRIP), [2:520](#)
 Safe speed limit, [2:635](#)
 SafeStat, [3:384](#)
 Safe Systems Approach (SSA)
 principles of, [2:520](#)
 Safe Traffic Netherlands (VVN), [2:522](#)
 Safe transportation of dangerous goods by air, [2:102](#)
 Safety analysis, [6:428](#)
 Safety assurance (SA), [2:96](#)
 Safety barrier, [2:597](#), [593](#)
 Safety bicycle, [2:115](#)
 Safety Culture, [2:554](#)
 drivers at work, [2:556](#)
 measured, [2:555](#)
 nonprofessional road users, [2:556](#), [2:558](#)
 organizational, [2:555](#)
 preventive measures in road sector, [2:557](#)
 Safety Data Quality Management
 background, [2:560](#)

 data quality principles, [2:561](#)
 calculating data quality scores for each record, [2:562](#)
 data quality attributes, [2:561](#)
 data standards, data governance, and formal data quality management, [2:561](#)
 statistically valid expectations, [2:562](#)
 statistical methods of data quality control, [2:561](#)
 formal data quality management processes, [2:562](#)
 benefits of, [2:563](#)
 surface transportation safety data, [2:560](#)
 Safety Drop, [4A:3](#)
 Safety-in-numbers (SIN), [2:133](#), [2:141](#), [2:142](#)
 Safety management system (SMS), [2:96](#), [2:554](#)
 Safety margins, [2:109](#)
 Safety measurement system (SMS), [3:384](#)
 Safety Measures, [2:78](#)
 Safety of Life at Sea (SOLAS)
 Convention, 1974, [3:302](#)
 Safety performance functions (SPF), [2:192](#), [2:285](#)
 Safety record, of HEMS, [2:317](#)
 balancing flight operations decision making, [2:319](#)
 developing technologies, [2:318](#)
 fatigue, [2:319](#)
 pushing the limits, [2:317](#)
 technical limitations and improvements, [2:318](#)
 Safety risk management (SRM), [2:96](#)
 Safety zone, [2:596](#), [593](#)
 Safe walkable environments, [5:322](#)
 Sag, [4A:136](#)
 Sailing Speed Decision, [3:337](#)
 Sample size, [7A:41](#)
 SARS-CoV-2 global pandemic, [6:243](#)
 Satellite and aerial imagery, [6:136](#)
 Satisfaction with travel, [7A:177](#)
 combined travel modes, [7A:179](#)
 independent of travel mode [7A:178](#)
 nonmotorized travel modes, [7A:179](#)
 private motorized travel modes, [7A:178](#)
 public motorized travel modes, [7A:179](#)
 travel-related life satisfaction, [7A:180](#)
 Save LIVES Technical Package, [2:271](#)
 Scale, [4B:229](#)
 Scenarios for achieving CO2 reduction targets, [7B:132](#)
 Scheduled services, [1:80](#)

Scheduling and reduced-form approaches, [1:81](#)
 School bus
 emergency exits, [2:565](#)
 loading zone, [2:564](#)
 safety standards, [2:565](#)
 seat belts, [2:566](#)
 security concerns, [2:566](#)
 School campus traffic circulation, [2:568](#)
 pretertiary, [2:568](#)
 active mobility, [2:569](#)
 carpooling, [2:569](#)
 interface management, [2:573](#)
 motives, [2:572](#)
 private car, [2:568](#)
 school bus, [2:571](#)
 walking school bus, [2:570](#)
 tertiary, [2:573](#)
 mixed traffic, [2:573](#)
 no traffic, [2:573](#)
 segregated traffic, [2:573](#)
 School travel in other settings, [6:433](#)
 School travel planning (STP), [6:128](#)
 Schramm Case, [3:385](#)
 Science Say, [2:8](#)
 Scooter sharing, [7A:190](#)
 Scotland's transport appraisal guidance (STAG), [6:363](#)
 Scramble crossings, [4A:352](#)
 Screening and Brief Interventions, [2:47](#)
 Sea ice, [3:350](#)
 Seaport, [1:443](#)
 business models, [3:300](#)
 climate change and adaptation to sea level rise, [3:303](#)
 environmental and sustainability challenges facing ports, [3:303](#)
 labor, [3:302](#)
 marketing and competition between ports, [3:301](#)
 ownership, [3:300](#)
 planning, [3:301](#)
 port security, [3:302](#)
 typology and governance, [3:299](#)
 Seaports as clusters, [3:310](#)
 business ecosystem perspective, [3:313](#)
 clustering, [3:311](#)
 port cluster
 components of, [3:311](#)
 particularities of, [3:312](#), [3:313](#)
 ports in proximity, [3:314](#)
 synergies from clustering, [3:310](#)
 synergy services, relevance of, [3:313](#)
 Seat belt reminders, [2:620](#), [6:307](#)
 Seat belts, [2:406](#), [2:270](#)
 Seaward planning, [1:443](#)
 Secondary incidents, [4A:123](#)
 Secondary market, [3:214](#)
 flow, [3:215](#)
 relative to GDP, [3:216](#)

- size, 3:215
 - total US, 3:216
 - Secondary packaging, 3:121
 - Secondary surveillance radar (SSR), 2:164
 - Second-best pricing, 1:97, 98
 - Second-degree price discrimination, 1:404
 - Second-party logistics (2PL), 3:103
 - Second Strategic Highway Research Program (SHRP2), 7A:124
 - Second world war and reshaping postwar civil aviation, 5:194
 - Sector-specific agencies (SSA), 2:244
 - Security, 1:124, 2:34
 - Segmentation approaches as a basis for behavior change, 7A:51
 - Segmentation approaches in transport, 7A:52
 - Segregated cycle lane development, 6:47
 - Segregated traffic, 2:573
 - Selected predictions from the one-period model, 1:117
 - Selected shared mobility travel modes, classification of, 7A:188
 - Self-deception, 7A:4
 - Self-reports
 - in accidents, near accidents, and mileage, 7A:3
 - auto-tech-detect (ATD)-based Self-reports in traffic, 7A:5
 - coping with socially desirable responding in self-reports of driving, 7A:5
 - impression management and self-deception in self-reports, 7A:4
 - in memory and social desirability, 7A:4
 - in self-assessments, 7A:3
 - self-reports in driver skills and behavior, 7A:2
 - in traffic research, content and situation of, 7A:2
 - usage in traffic research, 7A:2
 - Semirigid systems, 2:598
 - Senior drivers, training for, 2:231
 - Seniors and traffic safety, 2:126
 - Sense of Place, 4B:229
 - citywide, 4B:232
 - quantifying, 4B:230
 - Sensor faults and errors, 2:177
 - Sensors, 6:426
 - Sensors and state-awareness technologies, 2:176
 - Sensory impairments, in older people, 2:429
 - Service coordination, 5:331
 - Service focus, 3:9
 - Service live of assets, 1:453
 - Service network design (SND), 5:464
 - basic SND and service-selection planning, 5:465
 - models, 5:465
 - for railroad planning, 5:465
 - static SND for integrated planning, 5:465
 - time-dependent SND for integrated planning, 5:465
 - Service organization, 5:589
 - Service provision, pricing, and subsidies, 1:589
 - Service Quality in Public Transport, Measurement of, 6:221
 - Service time, 4A:63
 - Service vessels, 5:612
 - 72nd session of the Marine Environment Protection Committee (MEPC 72), 3:294
 - Severe acute respiratory syndrome (SARS), 5:267
 - Severity of, head-on crashes, 2:312
 - Sevici, 2:139
 - Sexual harassment, 7A:196
 - Sexual Violence, in Public Transportation
 - care, 2:576
 - emerging issues and future debate, 2:581
 - impact of, 2:581
 - nature of, 2:576
 - people most affected
 - age and disability, 2:579
 - ethnicity, 2:579
 - gender, 577
 - LGBTQI status, 2:578
 - previously victimized, 2:579
 - socioeconomic status, 2:580
 - underreporting, 2:580
 - prevention of, 2:581
 - selection of international evidence, 2:580
 - situational risky conditions for, 2:580
 - Shanghai Containerized Freight Index, 3:249, 3:248
 - Shanghai Declaration on Promoting Health in 2030 Agenda, 7B:4
 - Shanghai naturalistic driving study (SHNDS), 7A:30
 - Shapiro-Wilk test, 7A:83
 - Shared automated vehicle (SAV), 4B:151, 6:198
 - Shared freight systems, 1:23
 - Shared mobility, 5:21, 5:108
 - business models of, 5:155
 - ecosystem, 2:151
 - on emergency evacuations, 7B:155
 - scheme, 5:17
 - services, 7A:187
 - Shared Space, 2:584
 - in Ashford, United Kingdom, 2:588
 - benefits of, 2:588
 - in Coventry, United Kingdom, 2:589
 - debates, 2:590
 - defined, 2:584
 - elements of, 2:587
 - history of, 2:585
 - in Poynton, United Kingdom, 2:588
 - primary goal of, 2:584
 - principles of, 2:586
 - traffic calming, 2:586
 - urban design, 2:586
 - risk, 2:587
 - road safety and crashes, 2:589
- Shared taxi, 5:303
- hybrid mode choice model, 5:304
 - mode characteristics, 5:303
 - motivate users, 5:303
 - perceptual indicators, 5:304
 - some basic indicators, 5:303
- Share of female Uber service riders 2017, 5:660
- Share of transport as a part of total direct GH emissions in EU28, 1:458
- Shifting equilibrium, 1:492
- Shift schedule, 2:613
- Shipbuilding industry, 5:617
 - competition and competitors, 5:619
 - demand and cycles, 5:617
 - price and backlog, 5:622
- Shipment consolidation (SCL), 3:109
- Shippers, 2:100
- Shippers, brokers, and third-party logistics, 3:385
- Shipping, 7B:44
 - assessment of shipping impact to atmospheric pollutants and human health, Shipping, 7B:46
- Shipping and natural environment, 3:286
 - emissions to air and impacts, 3:286
 - fuel alternatives, 3:288
 - GHG, 3:287
 - life cycle impact, 3:288
 - nitrogen oxides, 3:287
 - particles, 3:287
 - sulfur oxide emissions, 3:287
- emissions to water and impacts, 3:290
 - ballast water, 3:291
 - biofouling and use of hull coatings, 3:290
 - environmental effects of oil spills, 3:290
 - oil spills, 3:290
 - waste generated on board, 3:291
- noise, 3:291
- propulsion alternatives, 3:289
 - electricity/batteries, 3:289
 - fuel cells, 3:290
 - nuclear, 3:289
 - wind assisted, 3:289

- Shipping and natural environment (Cont.)
 - PSSA, 3:291
 - sustainable shipping in the future, 3:292
- Shipping and shipbuilding cycles, interrelation between, 5:623
- Shipping emissions, 7B:44
 - future perspectives, 7B:49
 - global and local regulatory framework, 7B:45
 - relative (%) impacts to PM2.5 and PNC in four mediterranean harbor cities, 7B:50
 - relative (%) shipping contribution to gaseous pollutants across Europe, 7B:49
 - relative (%) shipping contribution to PM2.5 and PM10, 7B:48
 - sulfur content limits in marine fuels according to IMO and EU legislation, 7B:45
- Shipping Industry
 - bulk freight market, 3:261
 - demand for sea transportation, 3:262
 - dry bulk sector of, 3:263
 - coal seaborne trade, 3:264
 - derived demand for, 3:263
 - grains seaborne trade, 3:266
 - iron ore seaborne trade, 3:263
 - supply of, 3:268
 - economic participants, 3:257
 - main sectors, 3:257
 - market equilibrium and shipping cycles, 3:262
 - supply of sea transportation, 3:261
 - wet bulk sector of, 3:270
 - crude oil seaborne trade, 3:270
 - oil products seaborne trade, 3:274
 - supply of, 3:275
- Shipping lines, 3:307
- Ship route, 3:335
- Shocks to the transport system, 7B:34
- Shopping mode choice, factors affecting, 5:101
 - empirical evidence, 5:101
 - probabilistic behavioral mode choice model, 5:103
- Shopping travel, modeling structure, 5:100
- Shopping travel patterns, 5:99
- Short bus lanes near intersections, 6:255
- Short Message Service (SMS) technology, 4A:329
- Short-run average fixed cost (SRAFC), 4A:83
- Short-run equilibrium, of road transportation, 4A:83
- Short-run *versus* long-run costs, 1:14
- Short Sea Shipping (SSS), 3:280
 - environmental sustainability and, 3:284
- Short-term effects of noise on health, 7B:88
 - short-term annoyance, 7B:88
 - sleep disturbance, 7B:90, 7B:89
- Shoulder, 2:593
- Shoulder rumble strip, 2:593
- SHRP 2 naturalistic driving study, 7A:8
- Side Area Safety
 - barrier terminals and transitions, 2:598
 - crash cushions and arrester beds, 2:600
 - motorcyclists and cyclists protection devices, 2:598
 - obstacles, 2:594
 - roadside hazards, 2:594
 - safety barriers and bridge parapets, 2:597
 - safety zones, 2:596
 - side slopes, 2:595
 - testing and approval of road restraint systems, 2:601
 - vehicle restraint systems, 2:597
- Side-by-side-vehicles (SSV), 2:77
- Side slopes, 2:595
- Sidewalk, 2:593
- SIDRA software, 4A:210
- Signal Coordination, 4A:169
 - on arterial, one-way and two-way, 4A:218
 - definition, 4A:213
 - elements, 4A:214
 - goals, 4A:213
 - in oversaturated conditions, 4A:220
- Signal detection theory, 7A:11
- Signal group, 4A:185, 4A:196
- Signalized crossing, 4A:346, 4A:348
 - hawk crossing, 4A:351
 - pedestrian crossing time, 4A:348
 - pedestrian-vehicle separation controls, 4A:351
 - pelican crossing, 4A:349
 - puffin crossing, 4A:349
 - scramble crossings, 4A:352
- Signalized Roundabouts, 4A:227
 - design methods, 4A:233
 - design parameters, 4A:232
 - central island for left turn vehicles, 4A:232
 - geometric elements, 4A:232
 - signal timing and phase plan, 4A:233
 - traffic volume of left-turning vehicles, 4A:232
- types of, 4A:228
 - entrance metering, 4A:228
 - entrance signalization, 4A:229
 - signalization of the circulatory lane, 4A:229
 - vehicular and pedestrian signals, 4A:229
- Signal offset, 4A:169, 4A:171
- Signal Plan
 - goals of designing, 4A:179
 - methods of designing, 4A:180
 - performance evaluation, 4A:182
- Signal Progression Schemes, 4A:219
 - alternate system, 4A:219
 - force off, 4A:220
 - offset reference point, 4A:220
 - simultaneous system, 4A:219
 - X-ple alternating, 4A:219
- Signal vehicle coupled control (SVCC), 4A:183
- Silent salesperson, 3:120
- Silk Road Economic Belt (SREB), 3:495
- Silk Road Fund (SRF), 3:497
- Silver Bridge collapse, 2:144
- SimMobility, 4B:69
- Simple integrated land-use orchestrator (SILO), 4B:69
- Simple model to explain the effect of low emission zones and congestion tolls
 - excess traffic generated by urban congestion, 1:232
 - extension of the model to incorporate pollution, 1:233
 - measures: congestion tolls *versus* LEZ, 1:233
 - overall assessment of congestion tolls and LEZ, 1:233
- Simple model to explain the effect of low emission zones and congestion tolls, 1:232
- Simply multiple-criteria analysis (MCA), 1:541
- Simulation methods, 4B:4
- Simulation of motor vehicles, 7A:15
- Simulation of Urban Mobility (SUMO), 4B:69
- Simulator fidelity, 7A:18
- Simulator sickness, 2:606
- Simulator sickness, 7A:18
- Simulator validity, 7A:17
- Simultaneous equation models, 4B:57
- Singapore Electronic Road Pricing Scheme (ERP), 1:135
- Singapore, road pricing in, 4A:72
- Single accidents, 2:127, 2:131, 2:132
- Single Band vs Multiband, 4A:218
- Single origin-destination system, 1:47
- Single-till *versus* dual-till regulation, 5:238
- Sinusoidal in-ground rumble strips, 2:550
- Sinus rumble strips, 2:550

- Situational crime prevention (SCP),
2:161
strategies, 2:161
- Sky glow, 7B:68
- Sleep deprivation, 2:612
- Sleep-related issues, 2:611
- Slope, 2:593
- Slope rate, 2:593
- Slovenia (Sopotniki), 4B:20
- Small-, and medium-sized businesses (SME), 3:141
- Smaller time chunks, 4B:99
- Small unmanned aircraft, 5:703
- Small unmanned aircraft system (sUAS), 5:703
- Small waterplane area twin hull (SWATH) ferry, 2:291
- Smart and new modes of transport provision and women, 5:660
- SmartBike, 4A:356
- Smart card, 4A:329
- Smart mobility based on interconnected mobility services, 5:123
- Smart parking trials, 4A:291
- Smartphone apps role as an enabler, 5:156
- Smartphones, 1:164, 7A:8
- Smart phone ticketing, 4A:329
- Smart roads, 6:304
data collection through automated passenger counting systems, 6:304
driver assistance and road safety technologies, 6:306
traffic data collection, 6:304
traffic sensors with the related networks, 6:304
- Smith System of Defensive Driving, 2:232
- Snow and Ice Control Materials, 2:537
- Snowmobile, 2:81
design, 2:82
injury and fatality, epidemiology of, 2:82
laws, 2:83
rider practices and training, 2:82
safety measures, 2:82
- Snowmobile Safety Certification Committee (SSCC), 2:82
- Sobriety Checkpoints, 2:45
- Social and distributional impact (SDI) appraisal of all new road projects, 6:155
- Social and distributional impact assessments, 6:362
case study
highways england approach to assessing equality impacts of new major highway schemes, 6:364
engagement and expertise, 6:366
practical challenges and guidance, 6:364
evidence, evaluation, and monitoring, 6:365
limitations of existing assessment framework, 6:365
spatial and temporal coverage for assessment, 6:365
principles of social impact assessment, 6:362
typical process for social and distributional impact assessment, 6:363
- Social cost benefit analysis (SCBA), 1:322
- Social costs, 1:15
- Social discount rates, 1:197
- Social exclusion, 7B:117
barriers to travel, 7B:119
nature of, 7B:117
relationship between social exclusion and health, 7B:118
role transport plays in the relationship between social exclusion and health
access, 7B:120
externalities of transport, 7B:120
travel as a form of physical activity, 7B:120
role transport plays in the relationship between social exclusion and health, 7B:119
social exclusion and the ability to travel, 7B:119
transport in the relationship between social exclusion and health, 7B:118
- Social exclusion, 1:275
- Social interaction and cohesion, 7B:105
- Social media polarization, 6:1
- Social networks, 5:112
- Social norms, 5:386
- Social Optimum (SO), 4A:77
- Social-political history of bicycles, 5:697
- Social time preference approach, 1:197
- Societal changes, 5:255
- Societal shocks, 7B:34
- Society of Automotive Engineers (SAE), 2:82, 181
- Sociodemographic and driving-related variables, 7A:66
- Socio-economic and territorial determinants, 5:529
cities, regions, and maritime flows, 5:529
the interconnection with other networks, 5:530
sea-land connectivity of European ports and cities in 2008, 5:533
urban centralization of the global maritime network in 1950 and 2010, 5:531
- Solow-Swann model, impact of transport infrastructure investment, 1:257
- Solow-Swann model of regional growth, 1:256
- Sonars, 2:176
- Sound Navigation Ranging, 2:176
- South Dakota 24/7 Sobriety Program, 2:47
- Southern Traffic Incident eXchange program, 4A:123
- South Korea, port performance measurement in, 4A:400
DEMATEL and ANP, 4A:400
FER, 4A:400
- Space, 4B:229
- Space planning, 5:688
- Space reservation systems, 1:163
- Space transportation
economics, 5:642
impact of small satellites, 5:641
inter-orbit transport, 5:644
modern launchers, 5:640
orbits and sub-orbits, 5:638
reusable systems, 5:642
- Space transportation, 5:638
- Spain (Shotl & Castilla y León), 4B:20
- Spatial autocorrelation, 2:372
- Spatial competition, 1:161
- Spatial computable general equilibrium (SCGE) models, 3:165
- Spatial determinants of demand for air transport, 5:209
- Spatial income elasticity of air travel demand, 1:552
- Spatial Mismatch, 4B:62
expanding, evolving, and challenging, 4B:63
hypothesis, emergence of, 4B:62
job access, 4B:65
measuring, 4B:64
other contexts, 4B:65
- Spatial resolution, 4B:55
- Spatial spillovers, 5:3
- Spatiotemporal Congested Areas (STCA), 4A:164
- Special purpose vehicle (SPV), 1:374
- Specialty Vehicle Institute of America, 2:79
- Specific energy absorption (SEA), 2:63
- Specifics in the transport market, 1:243
- Speed, 2:127, 2:129
break, 2:91
bumps, 2:643
cushions, 2:643
Enforcement, 2:618
governor, 2:206, 2:618

- Speed (*Cont.*)
- humps, 2:643
 - limit, 2:635, 622
 - limiter, 2:618
 - and safety, 2:641
- Speed and vehicle throughput efficiency, 5:334
- Speed Drop, 4A:3
- Speeding antecedents, 7A:131
- Speeding consequences, 7A:132
- Speeding countermeasures, 7A:133
- Speeding researcher, next steps for, 7A:134
- Speed Limit
- on Rural Highways, 2:622
 - current practices in setting speed limits, 2:630
 - differential *vs.* uniform limits for trucks and buses, 2:627
 - relationship between speed and safety, 2:623
 - speed limit policy in the United States, 2:625
 - speed limits impact, 2:628
 - two-lane rural highways and speed limits, 2:629
 - statutory, 2:622
 - on Urban Streets, 2:635
 - benefits of, 2:639
 - in cities, 2:632
 - operating speeds, 2:635
 - public perceptions and politics, 2:639
 - road function, 2:636
 - safe system approach and, 2:633
- Speed-reducing measures
- behavioral-based measures, 2:646
 - emerging technologies and measures, 2:646
 - active vertical deflections, 2:646
 - GPS-based speed governors, 2:646
 - mid-block traffic signals with speed detection, 2:647
 - engineering-based measures, 2:642
 - chicanes, 2:644
 - curb extension, 2:644
 - curb radius reduction, 2:645
 - lane narrowing, 2:645
 - lateral shift, 2:644
 - on-street parking, 2:645
 - other infrastructure improvements, 2:645
 - pavement markings, 2:646
 - raised crosswalk, 2:644
 - raised intersection, 2:644
 - raised medians, 2:645
 - road diet, 2:645
 - roundabouts, 2:645
 - speed bumps, 2:643
 - speed cushions, 2:643
 - speed humps, 2:643
 - speed tables, 2:643
 - surface treatment, 2:646
 - traffic calming devices with
 - horizontal deflections, 2:644
 - traffic calming measures with
 - vertical deflections, 2:642
- Speed tables, 2:643
- Speed zoning, 2:622
- Split Cycle Offset Optimization Technique (SCOOT), 4A:174
- Split Time, 4A:215
- Sport utility vehicles (SUV), 2:297
- Spring and Summer Road Conditions, 2:530
 - cracking and roughness, 2:531
 - low surface friction, 2:532
 - patching, 2:531
 - pavement surface distresses, 2:530
 - potholes and delamination, 2:530
 - rutting, 2:530
- Spring suspension, 3:426
- Staggers Act of 1980, 1:409
- Staggers Rail Act of 1980, 3:26
- Stakeholder Behavioral Analysis, 3:243, 3:244
 - agent-based models, 3:243
 - discrete choice models, 3:243
- Standard deviation of the lateral position (SDLP), 2:218
- Standard Normal Deviate (SND), 4A:129
- Standards, apply to new vehicles, 1:516
- Standing Advisory Committee of Trunk Road Assessment (SACTRA), 1:471
- Stand-up paddleboards (SUP), 2:478
- State-centered “transport equity” approach, 6:158
- Stated choice experiments (SCE), 3:243
- Stated preference (SP)
 - methods, 1:118
 - and revealed preference data, 1:81
 - survey, 4B:100
 - techniques, 1:67, 1:216, 6:221, 6:222
- Static models with a single road, 1:209
- Stationary Traffic Flow Models, 4B:141
- Station-based and dockless bicycle sharing information systems, 6:22
- Station-based bike system (SBBS), 355
- Statistical algorithms, 4A:129
- Statistical test, 7A:82
- Stena HSS class catamaran, 2:290
- Stick shaker, 2:91
- Stochastic computational methods, 4B:3
- Stockholm congestion tax scheme, 1:136
- Stockholm, road pricing in, 4A:73
- Stock keeping unit (SKU), 3:71, 3:79
- Stock visibility, 3:222
- Stop-and-go traffic, 4A:149
- Stopping systems, 2:91
- Strata tickets, 4A:327
- Strategically set prices, 1:304
- Strategic Highway Safety Plan (SHSP), 2:525, 2:526
- Strategic network planning (SNP), 3:82
- Strategic policy directions, 6:131
- Strategic road network (SRN), 2:524
- Streetcar, 4A:338
- Street culture and carjacking, 2:160
- Street livability, 5:333
- Street safety and livability, 5:334
- Streetscapes that encourage walking, 5:335
- Stress classification system, 2:121
- Stressors, 2:337
 - age, 2:339
 - alcohol, 2:340
 - distraction, 2:337
 - fatigue, 2:337
 - inexperience, 2:339
 - marijuana, 2:340
 - medical conditions, 2:339
- Stress reduction, 7B:104
- Stress vulnerabilities and performance effects, 7A:204
- Structural Engineers Association of California (SEAOC), 2:244
- Structural equation models (SEMs), 3:145, 7A:96
 - future SEM prospects, 7A:99
 - path diagram, 7A:97
 - SEM ambits of use, 7A:97
 - SEM use in travel behavior research, 7A:98
- Structural health monitoring (SHM), 2:32
- Structural models, 1:576
- Studded tires, 2:683
- Subjective/perceived safety, 2:386
- Substantive Safety, 2:103
- Substitution effects, 7B:171
- Substructures in the global maritime network, 5:527
- Subway systems, 5:471
 - automation, 5:472
 - crowding, 5:477
 - distance-based *versus* flat pricing, 5:476
 - electrification, 5:472
 - Mohring Effect, 5:476
 - network, 5:472
 - operator costs, 5:474
 - reliability, 5:477
 - rolling stock, 5:474
 - spatial and temporal demand patterns, 5:475
 - subsidies and value capture, 5:475
 - total ridership, 5:474

- Suicides
- highway, 2:658
 - influencing factors, 2:659
 - prevention, 2:659
 - statistics/incidence, 2:658
 - types of events, 2:659
- railway, 2:656
- influencing factors, 2:657
 - prevention, 2:658
 - statistics/incidence, 2:656
 - types of events, 2:657
- research, practice, and policy, 2:660
- road, 2:659
- worldwide, numbers of, 2:656
- Sulfur dioxide (SO₂), 7B:44
- Sulfur Emission Control Areas (SECA), 3:284
- Sulfur oxide emissions, 3:287
- Sunshine Skyway, 2:144
- Superblock, base for a functional and urbanism model
- air quality and urban noise in Barcelona. exposed population, 7B:32
 - bus and bicycle network, 7B:27
 - case of Barcelona, 7B:25
 - in Barcelona in a superblocks scenario, 7B:27
 - map for pedestrians and other uses, 7B:28
 - map of superblocks in Barcelona, 7B:26
 - networks scheme, current and future, based on superblocks, 7B:26
 - pedestrian network, 7B:28
 - public space of Barcelona dedicated to through traffic in the current situation, 7B:27
 - thermal simulation comparing example in current situation and with superblocks, 7B:33
- Superblocks, a tool that helps to mitigation and adaptation to climate change, 7B:32
- Superelevation, 2:325
- Super off-peak, 3:27
- Supervisory control and data acquisition (SCADA), 2:244
- Supplementary Licensing scheme, 4A:69
- Supply Chain
- applications, 3:10, 3:11, 3:12
 - categorizing, 3:42
 - demand risk, 3:42
 - environmental risk, 3:42
 - internal risks, 3:42
 - supply risk, 3:42
 - constructing resilient, 3:42
 - design, 3:43
 - drivers of, 3:42
 - engendering, 3:43
 - establishing collaborative, 3:44
 - management, 3:8
 - reliability and service quality, 3:98
 - resilience and sustainability, 3:99
 - risk reduction, 3:44
- Supply Chain Finance (SCF)
- definition of, 3:114
 - historical background of, 3:113
 - measuring the performance of, 3:115
 - relevance of, 3:114
 - solutions, 3:116
 - stakeholders, 3:117
 - value proposition of, 3:115
- Supply chain management (SCM), 3:77, 3:145
- Supply chain operations reference (SCOR), 3:17
- Supply Chains and Logistics, 3:72
- environmental sustainability and logistical constraints, 3:74
 - information and business analytics, 3:72
 - ordering pattern, 3:74
 - supply chain partners, 3:72
- Supply chains, complexity of, 1:328
- Supply-jointness, 1:373
- Supply-side structures
- alliances, 1:439
 - intercargos, 1:439
 - intertanko, 1:438
 - Shipowners Associations, 1:438
 - World Shipping Council (WSC), 1:439
- Supply-side structures, 1:438
- Supply *versus* demand equilibrium, 1:17
- Supporting activities (SA), 4A:399
- Support Vector Domain Description (SVDD) algorithm, 4A:131
- Support vector machine (SVM), 4A:130
- Supranational organizations, 1:438
- Surface effect ship (SES) ferry, 2:291
- Surface traffic operations, 4A:368
- Surface Transportation and Uniform Relocation Assistance Act (STURAA), 2:625
- Surface transportation safety data, 2:560
- citation and adjudication data, 2:560
 - crash data, 2:560
 - driver data, 2:560
 - injury surveillance data, 2:561
 - roadway data, 2:560
 - vehicle data, 2:560
- Surrogate Measures of Safety, 2:662
- concept of, 2:664
 - data collection methods, 2:666
 - definition of, 2:663
 - diagnose, 2:666
 - historical perspective, 2:663
 - operationalization of, 2:664
 - evasive action, 2:665
 - PET category, 2:665
 - severity indicators, 2:664
 - TTC, 2:665
 - perspectives, 2:667
 - properties of, 2:663
- Sustainability and health issues in transportation, 7A:1
- Sustainability challenge, 5:596
- Sustainability/Environmental Management Schemes for Logistics and SCM, 3:20
- Sustainability in transport
- strategies of transport mode to achieve sustainability, 5:711
- Sustainability in transport, 5:710
- Sustainable development (SD), 6:150
- Sustainable development goals (SDGs), 7B:3, 7B:2, 7B:10
- COVID-19 affecting all SDGs, 7B:7
 - decade of action (2020–2030), 7B:4
 - "Health in All Policies" (HiAP) approach, 7B:4
 - monitoring and reporting, 7B:3
 - related to urban health within a HiAP approach, 7B:5
 - research led by ISGLOBAL, at least 48 SDG targets, 7B:4
- Sustainable development, principles of, 2:523
- Sustainable development scenario (SDS), 6:215
- Sustainable Development Working Group (SDWG), 3:349
- Sustainable logistics toward logistics for sustainability, 3:69
- Sustainable mobility, 6:184
- Sustainable mobility of goods, 6:322
- Sustainable mobility paradigm (SMP), 6:132
- Sustainable mobility paths, 5:13
- Sustainable mobility policy perspective on electric mobility, 6:69
- Sustainable or Environmental Performance Measures, 3:18
- frameworks, 3:19
- Sustainable Road Safety, 2:523
- Sustainable transport, 7B:8
- Sustainable transport amid a growth agenda, 6:95
- Sweden
- recreational boating in death and causes of injury, 2:479
 - road safety management in, 2:523
- Swedish National Police Board, 2:524
- Swedish Occupational Fatigue Index (SOFI), 2:614
- Swedish Traffic Accident Data Acquisition (STRADA), 2:678, 2:561
- Swedish Traffic Safety Agency, 2:675

Swedish Vision Zero
 output, 2:678
 public road safety policy innovation, 2:675
 traditional approach to road safety, 2:676
 Switzerland (Alpine Bus), 4B:20
 Sydney Coordinated Area Traffic System (SCATS), 4A:174
 Symbolic aspects of the car and identity, 7A:81
 Synchro, 4A:175
 Synchromodality defined, 5:580
 Synchromodality projects including inland waterway, 5:582
 Synchromodal Transport, 3:458
 increasing the capacity of loading units, 3:459
 infrastructure and networks, 3:459
 synchronization of assets, 3:459
 Synchromodal transport, challenges and drivers for, 5:581
 Synchromodal transport in inland waterway sector, 5:585
 Synergies of EV charging and the electricity system
 decreasing electricity prices and system costs while increasing the efficiency of conventional power plants, 1:562
 increasing the degree of capacity utilization of grid infrastructure, 1:561
 integrating (local) provision of electricity from fluctuating renewable energy generation, 1:562
 user acceptance of load flexibility, 1:562
 Synergies of EV charging and the electricity system, 1:561
 Synergy services, 3:313
 Synthetic hydrocarbons, 7B:130
 Synthetic Population Generator (SPG), 4B:10
 System component failure ù powerplant (SCF-PP), 2:105
 System dynamics models (SD), 1:542
 System of automobility, 6:131
 System of systems engineering (SoSE), 2:187
 System perspective and the individual safety, 2:135
 System-Wide Information Management (SWIM), 4A:367

T

Tachograph, 2:614
 Tackle adverse health impacts by changing transport systems

changes in mobilities, 7B:99
 changes in the physical infrastructure of transport systems, 7B:99
 changes in transport planning and governance, 7B:100
 Tackling inequalities related to transport modes, 5:674
 Talent supply chain, 3:12
 Tanner equation, 4A:240
 Tardy trip measures, 4A:111, 4A:117
 Target speed, 2:635
 TASHA, 4B:69
 Task-focus (TF), 7A:204
 Taxation, 5:254
 Taxation of car use
 current taxation of car use, 1:535
 exploring implications of future evolutions for, 1:535
 autonomous vehicles, 1:537
 demographic and economic growth, 1:535
 developments in road pricing technologies, 1:538
 fuel efficiency and electric vehicles, 1:536
 shared mobility, 1:536
 Taxi automated vehicles (TAVs), 1:512
 Taxicabs, 6:101
 economics of taxi markets, 6:104
 labor, 6:104
 usage, 6:104
 history and legality, 6:101
 market structure, 6:102
 pre-booked, 6:102
 shared vehicles, 6:102
 street hail, 6:102
 regulation, 6:101
 regulatory frameworks, 6:103
 fares, 6:103
 limits to entry, 6:103
 public safety, 6:103
 vehicle requirements, 6:104
 Taxi drivers, 2:700
 Taxi-out time, 4A:368
 Taxis security
 background, 2:699
 collision type, 2:704
 drivers, 2:700
 age, 2:700
 fatigue, 2:701
 gender, 2:701
 income, 2:701
 speed and speeding violation, 2:701
 facts, 2:700
 recommendations, 2:704
 safer drivers, 2:704
 safer vehicles, 2:704
 safety of passengers, 2:705
 ride-hailing vehicle drivers, 2:700

road, 2:702
 horizontal element, 2:703
 section, 2:702
 surface, 2:703
 type, 2:702
 safety belt use, 2:703
 speed limit, 2:704
 taxi drivers, 2:700
 travel time, 2:703
 day of week, 2:703
 time of day, 2:703
 vehicle, 2:702
 color, 2:702
 condition, 2:702
 occupancy, 2:703
 Technical drivers, 1:91
 Technique for Order Preference by Similarity (TOPSIS), 4B:85
 Technological countermeasures, 2:21
 Technology and parking innovations, 1:163
 Technology enabled data for sustainable transport policy, 6:135
 Technology opportunities, 1:24
 Teleactivities, 4B:47, 47
 Telepresence, 4B:47
 Temporal and spatial modification of travel, 5:168
 fragmentation of activities, 5:168
 travel-based multitasking, 5:169
 Temporal resolution, 4B:55
 Temporary hard shoulder use, 4A:41
 Temporary traffic control (TTC), 2:189
 devices, 2:190
 zone, 2:190
 Terminal area operations, 4A:368
 Terminal handling charges (THC), 3:249
 Terminal Operating System (TOS), 3:324
 Terminal operators, 3:307
 Terminal Radar Approach Control (TRACON), 4A:368
 Terminal regulation
 charging patterns for railway terminal use across European countries, 5:460
 definition and scope, 5:454
 hot topics and future trends in, 5:460
 railway diversification reframes terminal regulation, 5:455
 railway market opening in Europe and its consequences on, 5:456
 railway terminal pricing principles in Europe, 5:457
 regulation goals, 5:454
 terminal management charging models across European countries, 5:457

- Terminological heterogeneity in research on multimodality, 5:119
- contrasting terminologies, 5:119
 - delimiting terminologies, 5:120
 - supplementary terminologies, 5:119
- Terrain avoidance and warning system (TAWS), 2:355, 2:164
- Terrain collision awareness, 2:164
- Territorial cohesion, 5:316
- Territoriality, 7A:83
- Territorial variation of spillover effects, 5:318
- Terrorism, 2:35, 2:309
- Terrorist Threat
- campaigns, 2:668
 - escalating violence, 2:669
 - security issues, 2:672
 - tactics, trends in, 2:671
- Tertiary packaging, 3:121
- Tertiary school campus traffic
- circulation, 2:573
 - mixed traffic, 2:573
 - no traffic, 2:573
 - segregated traffic, 2:573
- Testing Procedures, 2:68
- Texas Transportation Institute (TTI), 4A:129, 4A:173
- TFM Surface Data Integration (TSDI), 4A:369
- The Crow Manual, 2:120
- The father of the yellow school bus, 2:571
- Theoretical approaches to explain firm location, 1:291
- Theoretical perspectives of aggressive driving, 2:18
- Theoretical safety, 2:386
- Theories of behavior and behavioral change, 7A:46
- Theories of behavior change, 7A:48
- Theories that explain transport behavior as deliberate choices, 7A:46
- Theory of learned carelessness, 2:496
- Theory of mode choice in cities, 5:45
- Theory of planned behavior, 5:64
- Theory of planned behavior (TPB), 5:63
- empirical evidence in support of the TPB in general and in predicting individual car use, 5:65
 - extending the, 5:68
 - TPB for modeling the psychological factors mediating the impact of objective environmental characteristics on observable travel behaviour, 5:65
 - using the TPB for designing and evaluating behavioral change interventions, 5:66
- Theory of planned behaviour (TPB), 7A:167
- The Polar Code, 3:353
- The Purple Guide, 3:202
- The Road Traffic Act 1984, 3:206
- Things matters, 2:7
- Third party logistics (3PL), 3:145
- characteristics of, 3:85
 - development, 3:147
 - selection process, 3:86
 - importance of criteria, 3:87
 - selection criteria, 3:86
 - services by logistics activity, 3:86
- 85th percentile speed, 2:622
- Threat, 2:35
- Three-node network, 1:49
- Three-node system, 1:49
- Three-second rule, 4A:74
- Three Silk Roads, 3:495
- Three stage least squares (3SLS) method, 4B:57
- Thrill, 7A:83
- Thrill seeking (TS), 7A:204, 7A:220
- Throw-and-go method, 2:532
- Thrust reversers, 2:91
- Tilburg trial, 2:619
- Time-Based Flow Management (TBFM), 4A:370
- Timed transfer system (TTS), 5:331
- Time Series ARIMA Algorithm, 4A:129
- Time series deterministic computational method, 4B:2
- Time series income elasticity of air travel demand, 1:549
- Time-space diagram (TSD), 4A:170
- Timestamp, 3:136
- Time-to-collision (TTC), 2:664, 2:665
- Time To Green (TTG), 2:184
- Time-varying or endogenous travel time distributions, 1:80
- Tire Safety
- force and moment generation mechanism, 2:682
 - non-pneumatic tires, 2:684
 - run-flat tire, 2:683
 - studded tires, 2:683
- Toll plaza, 4A:60
- Tool ports, 4A:392, 5:559
- Tools to study resilience, 7B:39
- Top-20 cargo and passenger airports worldwide, 2009-18, 5:261
- Top-down processing, 2:336
- Toronto-Montreal data, 1:130
- Total cost, 1:15
- Total Differential Method (TDM), 2:348
- Total private costs (TPCs), 1:56
- Total social costs (TSCs), 1:56
- Total sum of squares (TSS), 4B:169
- Total Time Spent (TTS), 4A:33, 4A:54
- Tourism (general aviation) aircraft, 5:278
- Tourism and transport modes, 5:26
- Tourism for transport, 5:28
- Tourism travel planning, relevance of social networks and user-generated content for, 6:42
- Tourism travel planning, role of ICT for, 6:41
- Town Gates, 2:685
- correct decision-making in towns, 2:689
 - future uses of, 2:690
 - logic of action constraints, 2:687
 - logic of things inside, 2:686
- Toyota mixed-load system, 3:108
- Toyota Production System (TPS), 3:107
- Trace elements, 7B:44
- Tracking data, 6:135
- Traction elevator, 5:685
- Trade and transport costs developments, 1:434
- Trade facilitation, 2:101, 3:154
- Trade logistics, 3:150
- Trade-related freight demand, 3:170
- Traditional planning model for tourism travel, 6:39
- Traditional way of building transport infrastructure, 1:303
- Traffic Accident Commission (TAC), 2:521
- Traffic accident reconstruction, 2:348
- Traffic advisory (TA), 2:164
- Traffic alert and collision avoidance system (TCAS), 2:354, 2:164
- aural alerts in RA of, 2:167
 - conflict detection in, 2:166
 - II indication, 2:169
 - polar coordinate system, 2:166
 - structure of, 2:168
 - surveillance function of, 2:165
- Traffic analysis districts (TAD), 2:368
- Traffic analysis zone (TAZ), 2:368, 4B:24, 4B:26, 4B:34, 4B:193
- resolution, 4B:55
- Traffic assignment, 4B:26, 4B:27
- Traffic Aware Strategic Aircrew Requests (TASAR), 4A:371
- Traffic-based units, 2:368
- traffic analysis districts, 2:368
 - traffic analysis zones, 2:368
 - traffic safety analysis zones, 2:368
- Traffic calming, 2:448, 2:586, 2:644, 2:685
- Traffic capacity, 4A:63
- Traffic collision avoidance system (TCAS), 2:92
- Traffic collisions, 7A:171
- Traffic conditions and head-on crashes, 2:312

- Traffic conflict technique, 7A:12
- Traffic congestion, 4A:52, 4A:154
 - bottlenecks on freeways, 4A:155
 - imbalanced merging section, 4A:155
 - peak in traffic demand, 4A:154
 - sag and tunnel, 4A:156
 - weaving section, 4A:156
- Traffic conflict techniques (TCT), 2:663
- Traffic control devices (TCD)
 - at railroad grade crossing, 2:468
- Traffic control, types of, 4A:341
- Traffic-count-data-based models, 4B:110
 - dynamic, 4B:111
 - static, 4B:110
- Traffic crash, 2:384
- Traffic culture, 2:264
- Traffic data, 4B:111
- Traffic demand, 4B:109
- Traffic devices, for pedestrian safety, 2:427
- Traffic display indicator (TDI), 2:354
- Traffic flow, defined, 2:692
- Traffic Flow Management (TFM), 4A:365, 4A:366
- Traffic Flow Management System (TFMS), 4A:370
- Traffic Flow Measurements and Variable Definitions, 4B:140
- Traffic flow prediction and related problems, 4B:94
 - advantages and drawbacks, 4B:95
 - bilevel optimization and network design, 4B:94
 - traffic assignment and equilibrium, 4B:94
- Traffic Flow Volume and Safety
 - AADT and crash frequency, 2:696
 - AADT and crash severity, 2:697
 - data sources, 2:693
 - emerging methods, 2:693
 - roadside sensors, 2:693
 - traditional approach, 2:693
 - video vehicle detection, 2:693
 - less frequent traffic flow parameters, 2:694
 - precise crash time estimation and correction, 2:695
 - real-time traffic and crash frequency, 2:696
 - real-time traffic and crash severity, 2:697
 - real-time traffic data potential, 2:694
 - sum and average of traffic flow, 2:694
 - variance of traffic flow, 2:694
 - road segmentation, 2:694
 - safety indicators, 2:695
 - real-time crash risk, 2:695
- Traffic incident detection, 4A:128
- Traffic incident management (TIM), 4A:122
 - development, 4A:122, 4A:123
 - clearance, 4A:122
 - communication and information exchange, 4A:123
 - detection, 4A:122
 - incident command system, 4A:125
 - motorist information, 4A:122
 - quick clearance legislation, 4A:124
 - recovery, 4A:122
 - response, 4A:122
 - site management, 4A:122
 - system, 4A:124
 - traffic incident duration estimation, 4A:126
 - transportation management center, 4A:125
 - traveler information, 4A:126
 - verification, 4A:122
 - stages of, 4A:122
- Traffic incident management system (TIMS), 4A:124
 - automatic incident detection, 4A:124
 - dissemination of traffic and transport information to the public, 4A:124
 - provision of traffic information to stakeholders, 4A:124
 - traffic control and surveillance systems, 4A:124
- Traffic incidents, 4A:158
- Traffic Injury Research Foundation (TIRF), 2:6
- Traffic law enforcement, 2:265
- Traffic management, 3:202, 4A:1
- Traffic Management Centres (TMC), 4A:212
- Traffic Management Initiatives (TMI), 4A:365
- Traffic management with electronic toll collection, 4A:66
- Traffic medians, 2:488
 - effect on mobility, 2:490
 - median strip at intersections, 2:489
 - median strip instead of two-way left turn lanes, 2:489
 - type of median strip, 2:490
 - width of the median strip, 2:489
- Traffic model and theoretical algorithms, 4A:130
- Traffic Network Study Tool (TRANSYT), 4A:175
- Traffic operation centers (TOC), 4A:124
- Traffic oscillation, 4A:149
- Traffic psychology, 7A:9
- Traffic psychology, pitfalls of statistical methods in, 7A:81
- Traffic records, 2:560
- Traffic-related air pollution, 7B:153
- Traffic-related noise, 7B:154
- Traffic Risk, 1:117
- Traffic rules and regulations, 7A:171
- Traffic safety, 2:551, 7A:198
 - climate, 2:264
 - culture, 2:264, 7A:133
 - developing nations, 7A:199
 - effects of milled rumble strips, 2:552
 - 3E model of, 2:263
 - industrial world, 7A:198
- Traffic safety analysis zones (TSAZ), 2:368
- Traffic Signal
 - control modes, 4A:181
 - plan and its components, 4A:178
- Traffic Signal Coordination
 - bandwidth, 4A:217
 - computing offsets, 4A:215
 - coordinated phase, actuating, 4A:216
 - cycle length, 4A:214
 - green time, 4A:214
 - movement numbering scheme, 4A:216
 - offsets, 4A:215
 - effects of vehicle in queue on, 4A:215
 - fine-tuning, 4A:216
 - in oversaturated conditions, 4A:220
 - signal coordination
 - on arterial, one-way and two-way, 4A:218
 - single band vs multiband, 4A:218
 - split time, 4A:215
 - transition logic, 4A:217
- Traffic signal phase, 4A:185, 4A:196
- Traffic signal priority for buses, 6:256
- Traffic Signals and Safety
 - bicyclists at signalized intersections, 2:711
 - eliminating traffic signals, 2:712
 - low-volume periods, 2:712
 - new signal installations, 2:709
 - pavement friction, 2:707
 - pedestrians at signalized intersections, 2:710
 - permissive left turns, 2:707
 - permissive right turn on red, 2:708
 - queue presence, 2:707
 - red signal violation, 2:706
 - signal design and settings, 2:709
 - signalized crosswalks, 2:711
 - signal preemption, 2:711
 - signals visibility and readability, 2:709
- Traffic signs
 - and striping, 2:648
 - applications, 2:648
 - forgiving devices and self-explaining roads, 2:654
 - guide signs, 2:648

- human errors, 2:653
- inspection, 2:653
- redundant, 2:653
- regulatory signs, 2:648
- retroreflectivity, 2:652
- safety benefits, 2:651
- uniformity, 2:653
- warning signs, 2:648
- types of, 2:649
- Train/BRT, 1:246
- Train configuration and distributed power, 3:432
- Train ferry, 2:291
- Train length, 3:443
- Trainsets, definition and classification of, 5:495
- Trajectory Based Operation (TBO), 4A:370
- Tram Lane Configurations, 4A:338
- Tram Lanes and Road Rules, 4A:338
 - mixed traffic tram lane, 4A:339
 - tram lanes, 4A:339
 - tramways, 4A:338
- Tramway stop with level boarding, Toulouse, France, 6:32
- Trans-China Railway (TCR), 3:495
- Transcription, 7A:108
- Trans-European Transport Network (TEN-T), 3:460, 5:318
- Transferability of cross-sectional models
 - longitudinal studies, 1:171
- Transferability of cross-sectional models, 1:170
- Transfer boardings and alightings, 4B:56
- Transfer walking time, 1:123
- Transit captives, 2:576
- Transit cities, 5:47
- Transit Fare Collection
 - credit types, 4A:326
 - equipment, 4A:329
 - fare policies, 4A:325
 - fare structure, 4A:325
 - integration of fare payment, 4A:330
 - payment methods, 4A:327
 - special tickets and credits, 4A:327
 - ticketing technologies, 4A:329
- Transit finance, 5:354
- Transit Information Systems
 - impact on travel behavior, 4A:333
 - perception of safety and personal security, 4A:334
 - personal satisfaction with transit service, 4A:334
 - psychological effects of RTTIS on ridership, 4A:334
 - transit use, 4A:334
 - waiting time/travel time, 4A:333
- technological advancements in, 4A:335
- Transit investments, 1:89
- Transition curves, 2:324
- Transition from piston to gas turbine
 - airliners, 5:195
- Transition Logic, 4A:217
- Transit management, 5:353
- Transit mode shares, 5:350
- Transit on-board surveys, 4B:182
- Transit operations, 5:353
- Transit-oriented decisions, of
 - passengers, 4A:331
- Transit oriented development (TOD), 6:36
- Transit Priority, 4A:307
 - active, 4A:307, 4A:308
 - combined with other facilities, 4A:311
 - passive, 4A:307, 4A:308
 - real-time, 4A:307, 4A:309
 - signal coordination for transit, 4A:310
- Transit priority measures, 6:254
- Transit providers, 4A:331
- Transit Ridership Forecasting Models, 4B:55
 - complementary and competing routes, 4B:56
- emerging trends and future directions, 4B:60
- factors influencing transit ridership, 4B:57
 - accessibility of the transit system, 4B:58
 - accessibility provided by the transit system, 4B:58
 - direct demand models, 4B:58
 - external or exogenous factors, 4B:57
 - internal or endogenous factors, 4B:57
- interdependency between demand and supply, 4B:56
- model structure, 4B:58
- spatial resolution, 4B:55
- temporal resolution, 4B:55
- transfer boardings and alightings, 4B:56
- Transit signal priority (TSP), 4A:222, 4A:307, 5:328, 5:329
- Transparency, 1:543
- Trans-Polar Route (TSR), 3:349
- Transport, 5:92
- Transport access and health, 7B:111
 - access-related transport barriers to healthcare, 7B:112
 - direct and indirect relation between, 7B:112
 - equity of access to health care, 7B:114
- Transport Act (2000), 6:3
- Transport and climate change, links between, 7B:96
- Transport and rail market
 - from monopoly to competition, 5:447
- oligopolistic transport market, 5:447
- public or private railways, 5:448
- unified or separated railway, 5:447
- Transport and rail market, 5:447
- Transport and Road Safety (TARS)
 - research center, 2:80
- Transport as tourism, 5:27
- Transportation, 2:276
 - and health, 7B:6
 - conceptual model, 7B:17
 - perceived risk and objective risk, 2:494
 - risk judgments, 2:495
 - bounded rationality and the affect heuristic, 2:495
 - risk as feelings and perceived control, 2:496
 - risk perception and risk behavior, 2:496
 - learned carelessness, 2:496
 - mobile phone use while driving, 2:496
 - risk communication, 2:497
 - risk homeostasis and risk compensation, 2:497
- Transportation Asset Management (TAM), 4B:122
- Transportation Asset Management Plans (TAMP), 4B:118
- Transportation costs, comparative, 5:395
- Transportation equity, 1:271
 - and fair distribution, 1:274
 - future perspective, 1:276
 - and guiding moral principles, 1:274
 - importance of, 1:272
 - theorizing, 1:272
- Transportation Impact by Mode, 3:59
- Transportation improvements, 1:337, 1:338
 - before and after comparisons, 1:344
 - hedonic regression analysis, 1:344
 - market equilibrium simulations, 1:345
- Transportation Management
 - Associations (TMA), 4B:63
- Transportation management center (TMC), 4A:125
- Transportation mode inference, 6:429
- Transportation modeling and planning
 - software
 - activity-based models and the open-source revolution, 4B:166
 - data and features, 4B:164
 - core data elements, 4B:164
 - core software features, 4B:165
 - future directions for, 4B:167
 - historical overview, 4B:163
- Transportation modes and international mobility, 5:40

- Transportation modes in a globalized economy, 5:43
- Transportation network companies (TNC), 1:276, 4B:40, 5:702
 - and curbside management, 1:585
 - future of TNCs and vehicle automation, 1:586
 - impacts on travel, labor, and emissions, 1:584
 - possible impacts of shared automated vehicles, 1:586
 - relationship between SAVs and public transportation, 1:587
 - and social equity, 585
- Transportation performance indicators, 4A:304
- Transportation Research Board (TRB) Committee on Alcohol, 2:50
- Transportation Resilience, 3:53
- Transportation resilience and evacuation planning, 2:276
 - all clear, return home, and back to normalcy, 2:278
 - friends and family, hotel, campsite, or public shelter, 2:277
 - new technologies, 2:278
 - resilient evacuation planning, 2:280
 - training and education, 2:279
 - walk, bike, drive, carpool, bus, or ride haling services, 2:277
- Transportation role, in emergency response systems, 2:245
- Transportation Safety and Security, 3:47
 - human factor, 3:47
 - risks and vulnerabilities, 3:48
 - 21st century, 3:50
- Transportation safety, economics of, 1:476
 - commercial freight transportation, 1:478
 - commercial passenger transportation, 1:478
 - market fails, reason for, 1:478
 - bilateral crashes, 1:479
 - carrier myopia, 1:479
 - cognitive failures by individuals, 1:479
 - externalities, 1:479
 - imperfect competition, 1:480
 - imperfect information, 1:479
 - market interventions, 1:480
 - information provision, 1:481
 - insurance requirement, 480
 - liability, 1:480
 - safety regulation, 1:481
 - relative magnitude of the market failures, 1:480
 - socially desirable decisions
 - for commercial carriers, 1:477
 - by emergency responders, 1:478
 - by infrastructure providers, 1:478
 - by private users, 1:477
- Transportation security, 3:47
- Transportation simulations, microlevel in, 4B:67
- Transportation Simulators
 - eye tracking measures, 2:608
 - performance metrics, 2:607
 - simulator measures, 2:607
 - population and participants, 2:605
 - psychophysiology measures, 2:608
 - pupillometry, 2:608
 - simulation characteristics, fidelity, and validity, 2:602
 - measuring fidelity, 2:603
 - scenario design, 2:603
 - scenario validation, 2:604
 - simulator sickness, 2:606
 - subjective measures, 2:609
- Transportation statistics and databases
 - Akaike's information criterion, 4B:169
 - analysis of variance, 4B:169
 - Anderson-Darling (AD) test, 4B:169
 - autocorrelation in regression, 4B:170
 - Bayesian nonparametric statistics, 4B:170
 - Bayesian P values, 4B:171
 - binomial distribution, 4B:171
 - bivariate distributions, 4B:171
 - bootstrap methods, 4B:172
 - chi-square distribution, 4B:172
 - confidence interval, 4B:172
 - copulas, 4B:172
 - crosssectional data, 4B:178
 - degrees of freedom, 4B:172
 - dummy variables, 4B:173
 - Durbin-Watson Test, 4B:173
 - entropy, 4B:173
 - expected value, 4B:173
 - extreme value distributions, 4B:173
 - factor analysis, 4B:173
 - F distribution, 4B:173
 - gamma distribution, 4B:174
 - geometric and negative binomial distributions, 4B:174
 - heteroscedasticity, 4B:174
 - hypothesis testing, 4B:174
 - identifiability, 4B:175
 - Kolmogorov-Smirnov test, 4B:175
 - likelihood ratio test, 4B:175
 - lognormal distribution, 4B:175
 - Markov Chain Monte Carlo, 4B:175
 - mixture models, 4B:176
 - moving averages, 4B:176
 - multicollinearity, 4B:176
 - normal distribution, univariate, or multivariate, 4B:176
 - panel data, 4B:178
 - Poisson distribution, 4B:177
 - p values, 4B:177
 - revealed preference, 4B:178
 - sample size determination, 4B:177
 - sampling algorithms, 4B:177
 - stated preference data, 4B:179
 - student's t-tests, 4B:177
 - survival data, 4B:179
 - uniform distribution, 4B:178
 - Weibull distribution, 4B:178
- Transportation supply, 4B:103
- Transport behavior, observed changes in, 1:90
- Transport big data, 5:666
 - freight, 5:667
 - new modes and systems of mobility, 5:667
 - private transport, 5:666
 - public transport, 5:666
- Transport Canada (TC), 2:521
- Transport cost-benefit analysis on public decision making, impact of, 1:278
- Transport costs for firm location, empirical studies on the role of, 1:292
 - firms, types of, 1:293
 - geographical level of analysis, 1:293
 - measurement of transport cost, 1:293
 - transport infrastructure, types of, 1:294
- Transport demand management (TDM), 4B:131, 6:2, 6:374
 - assessment of TDM strategies, 6:376
 - effects of, 4B:132
 - long-term TDM strategies, 6:375
 - origin and evolution of, 6:374
 - planning TDM strategies for a transition time, 6:378
 - short-term TDM strategies, 6:375
 - state of the art of, 4B:131
 - better job-housing balance, 4B:134
 - congestion charging, 4B:133
 - distance-based charging, 4B:134
 - encouraging ride-sharing, 4B:134
 - flexible work schedule, 4B:133
 - mobility as a service, 4B:135
 - park and ride facilities, 4B:135
 - parking controls and pricing, 4B:134
 - promoting mixed-use development, 4B:132
 - public transport subsidization, 4B:135
 - school and tourist transport management programs, 4B:135
 - taxation, 4B:134
 - telecommunications and e-working, 4B:133
 - traffic information, 4B:133

- transit-oriented development, 4B:135
- urban design and site design, 4B:135
- vehicle quota system, 4B:134
- type of measure, 6:3
- Transport development in China, 6:48
- Transport economics and events, 3:205
- Transport-economy frictions, 1:260
- Transport external costs and labor market, 1:62
- Transport for tourism, 5:27
- Transport governance, 6:73
 - and concept of “system builders”, 6:75
 - future challenges, 6:75
 - meta-governor, 6:74
 - policy-situated governance, 6:74
 - strategy-tactics-operation framework (sto), 6:74
 - tender-based governance, 6:74
 - trusting partnership, 6:74
- Transport improvement in monocentric city, 1:342
- Transport improvement unlocking development, 1:358
- Transport infrastructure, 1:302
 - adapting the half-life concept to reassess, 6:396
 - development generate positive externalities on the economy, 1:372
 - development, taking stock of the relationship between durability and flexibility in, 6:396
 - effects on economic, 1:347
 - empirical evidence from microeconomic firm-level studies, 1:347, 1:348
 - identification issues and strategies, 1:349
 - industrial heterogeneity, 1:350
 - measuring the effect of transport improvements, 1:348
 - spatial heterogeneity, 1:352
 - investment, 1:258
 - investments and economic growth, 5:2
- Transport innovation, 5:36
- Transport innovation, older people and policy, 6:62
- Transport interventions
 - causal inference for ex post evaluation of, 1:284
 - classified as treatment variables, 1:284
- Transport investment, 1:356
 - and increasing productivity, 1:356
 - and long-run growth, 1:260
- Transport justice (TJ), 7B:81, 7B:83
- Transport logistics, 3:204
- Transport management systems (TMS), 3:76, 3:77
- Transport matters for equality, 5:671
- Transport mode, assessing inequalities associated with, 5:674, 5:675
- Transport mode choice, determinants of, 5:713
 - intercity mobility, 5:713
 - urban mobility, 5:713
- Transport mode choice for women, 5:657
- Transport model-based analyses, 4B:44
- Transport modeling and data management, 4B:1
- Transport modelling, 7A:1
- Transport modes, 5:1, 5:107
- Transport modes and health, inequalities in, 5:114
 - age, 5:115
 - distribution of impacts, 114
 - ethnicity, 5:115
 - gender and sexuality, 5:115
 - impairments, 5:115
 - socioeconomic position, 5:115
- Transport modes and remote areas, 5:51
- Transport modes, equality and the benefits and values of, 672
- Transport modes, risks and harms of, 672
 - climate change, 5:673
 - pollution and land use, 5:673
 - road traffic collisions, 5:672
- Transport modes, travel, and activities-related challenges, 5:34
- Transport modes used by disabled and elderly people, 5:85
- Transport network companies (TNCs), 6:199
- Transport network design problem (TNDP), 5:625
- Transport of passengers, 5:587
- Transport planning and policy
 - cross-sector working to develop shared planning and policy, 6:61
 - involving older people in policy-making, 6:34
 - transport policy, changing the vision of, 6:61
 - use of life stages in, 6:60
- Transport planning and reducing transport energy demand, 6:215
- Transport planning as a means to promote socio-economic equity, 6:157
- Transport planning in the global south, 6:118
 - adoption and development of the UTP in the global south, 6:118
- challenge for integration of mainstream approaches to UTP and theories, 6:123
- challenges and opportunities for, 6:120
- equity, segregation, and socioeconomic vulnerability, 6:122
- informality and innovation, 6:122
- paradigm shifts and the resurgence of public transport, cycling, and walking, 6:121
- Transport planning, policies and air quality, links between, 6:237
- Transport policy, 5:388, 7A:1
 - and governance, 6:73
 - and planning, accessibility tools for, 6:83
 - process, media influence on, 6:405
- Transport policy and planning, accessibility tools for
 - accessibility concept, 6:83
 - accessibility in planning practice and policy making, application of, 6:85
 - Dutch ABC location policy, 6:85
 - RIN guideline, Germany, 6:85
 - Switzerland, the agglomeration policy, 6:85
 - WebCAT in London, 6:85
- accessibility measures, 6:84
- accessibility tools, 6:84
- background and development over time, 6:83
- critical reflection for new horizons, 6:85
- implementation gap, 6:83
- Transport policy and planning, evaluation methods in, 6:230
- cost-benefit analysis, 6:230
- Environmental Impact Assessment (EIA), 6:233
- multicriteria analysis, 6:232
- participatory value evaluation (PVE), 6:231
- social impact assessment, 6:234
- willingness to allocate public budget experiments (WTAPB), 6:231
- Transport production and cost structure, 1:46
- Transport project financing, 6:292
- Transport risk perception, 7A:74
 - definition and measurement of, 7A:74
- Transport safety and security, 7A:1
- Transports interventions and outcomes, 1:283
- Transport system, 7A:1
 - disruptions, 7B:37
 - elements of, 2:126
- Transport time savings, 1:321

- Trans Siberian Railway (TSR), 3:436, 3:495
- Trans World Airlines (TWA), 2:99
- Travel-based multitasking, 4B:51
- Travel behaviour, 7A:54
- Travel characteristic, 1:131
- Travel choices and activity decisions, 4B:47
 - big data, 4B:49
 - digital divide and digital skills, 4B:49
 - future, 4B:51
 - intelligent transport systems, 4B:49
 - interactions with activity decisions, 4B:50
 - smart and mobile capabilities, 4B:48
 - tele-, online, virtual, digital, and e-activities, 4B:47
- Travel demand, 4B:159
- Travel demand management, 4B:38
 - policy, 4A:325
 - program, 6:173
- Travel demand models (TDM), 4B:8, 184
 - activity-based modeling, 4B:188
 - components, 4B:185
 - emerging issues, 4B:189
 - four-step modeling, 4B:187
 - model application, 4B:187
 - model validation, 4B:187
 - parameters, 4B:186
 - state of the practice, 4B:184
 - trip assignment methods, 4B:186
- Traveled way, 2:593
- Traveler implications and behavioral responses, 4B:103
 - change in temporary workplace location, 4B:105
 - changes in departure time/scheduling, 4B:104
 - changes in itineraries/stop-making, 4B:104
 - changes in route, 4B:103
 - changes in travel mode, 4B:105
 - outside-region residential, 4B:106
 - vehicle/equipment upgrade, 4B:105
 - within-region residential, 4B:106
- Traveler information stations, 4B:13
- Traveler Responses
 - to congestion
 - understanding and ability to predict, 4B:106
 - mapping of, 4B:104, 4B:106
- Traveling by automobile, risks for, 6:420
- Traveling by different modes, rail and air, risks for, 6:419
- Traveling electrons, 5:17
- Travel model calibration and validation
 - calibration and validation of, 4B:193
 - model calibration and validation, 4B:190
- sensitivity analysis of model
 - components and entire model system, 4B:196
 - sources of error and uncertainty in, 4B:192
- Travel pattern, 5:100
 - benefits and risks
 - environmental impacts, 4B:226
 - health and well-being, 4B:226
 - social and economic benefits, 4B:226
 - influencing factors
 - climate and atmospheric impacts, 4B:225
 - economic environment, 4B:225
 - government policies, 4B:225
 - human factors, 4B:224
 - social influences, 4B:224
 - recreation supply, 4B:225
 - temporal and spatial distribution, 4B:224
 - transport, 4B:225
 - travel mode choice characteristics, 4B:224
- Travel plans, 6:408
 - barriers to travel plan take up, 6:410
 - defining and characterizing, 6:408
 - emergence, 6:409
 - motivations for promoting or adopting travel plans, 6:409
 - stagnation, 6:410
 - transformation, 6:411
- Travel-related decisions, 1:102
- Travel surveys, 4B:181
 - establishment surveys, 4B:182
 - future of, 4B:182
 - household surveys, 4B:181
 - transit on-board surveys, 4B:182
- Travel time estimation, 6:427
- Travel Time Reliability (TTR), 4A:109, 4A:164
 - agreement among reliability measures, 4A:119
 - buffer measures, 4A:110
 - case study, 4A:112
 - classification of measures, 4A:109
 - congestion based measures, 4A:112
 - probabilistic measures, 4A:111
 - reliability performance measures, 4A:109
 - statistical range measures, 4A:110
 - tardy trip measures, 4A:111
- Travel time variability (TTV), 1:76, 1:80, 1:81, 1:82
 - application in demand prediction and cost-benefit analysis, 1:82
 - imperfect information or limited rationality, 1:80
 - slope model, 1:79
 - step model, 1:78
- and trip chains with multiple activities, 1:80
- Trends and drivers of transport demand, 1:90
 - demographic drivers, 1:90
 - economic driver, 1:91
 - gasoline price, 1:91
 - GDP growth, 1:91
 - information and communication technologies, 1:91
 - International migration, 1:90
 - lifestyle changes, 1:91
 - new transport infrastructures, 1:91
 - technical drivers, 1:91
 - urbanization, 1:91
- Trespassing, 2:472, 2:473
- Tri-Butyl-Tin (TBT), 3:291
- Triggered responses, 7A:113
- Trim-and-fill method, 2:223
- Trip assignment, 4B:188, 4B:186
- Trip-based models, advantages over, 4B:8
- Trip chaining analysis
 - in an era of big data, 4B:203
 - congestion, emissions, and public health, 4B:200
 - decision theory and models explaining trip chaining, 4B:201
 - factors affecting, 4B:202
 - transit use, and emerging mobility, 4B:203
 - trips to trip chains, 4B:200
- Trip chains, 5:656
- Trip (voyage) cost functions, 1:41
- Trip distances, 5:657
- Trip distribution, 4B:26, 4B:187
- Trip duration, 5:657
- Trip generation, 4B:26, 4B:187
- Triple bottom line (3BL), 3:64
- Trip purposes, 5:656
- Trisecting model of the transport system, 7B:35
- Truck crash causation, 2:202
- Trucking firms, 2:100
- Trucking industry, 1:51
 - carrier selection process, 3:384
 - carrier hiring, assessing negligence in, 3:384
 - legal battle, for assessing carriers, 3:385
 - minimal shipper, broker, and third-party logistics, 3:386
 - shippers, brokers, and third-party logistics, 3:385
 - in the United States, 3:382
- Truck-mounted attenuators (TMA), 2:194
- Truss bridges, 2:513
- Trust violation, 2:159

Tunnel, [4A:136](#)
 Turbo roundabout, [2:543](#), [4A:238](#)
 capacity calculation, [240](#)
 critical gap and follow-up time in, [4A:240](#)
 delays and level of service estimation, [4A:243](#)
 one and two-level, [4A:244](#)
 performance analysis of, [4A:238](#)
 Turbulence, [2:93](#)
 Tweel, [2:684](#)
 Twenty-foot equivalent units (TEU), [3:335](#), [4A:5](#)
 Two-way left-turn lane (TWLTL), [2:500](#)
 Two-way stop-controlled intersection (TWSC), [4A:251](#)

U

UAS Traffic Management (UTM), [4A:369](#)
 Uber (cars)
 Mobility as a Service (MaaS), [4B:158](#), [6:199](#)
 controversies, [1:566](#)
 drivers, [1:566](#)
 incumbent taxis, [1:566](#)
 regulation, [1:566](#)
 sustainable transport, [1:566](#)
 Ultralight aircraft, [5:279](#)
 Ultra Low Emission Zone (ULEZ), [4A:73](#)
 Ultra-safe high-risk industry (USHRI), [2:90](#)
 Uncertainties in CBA, major sources of, [1:251](#)
 Uncertainties in cost estimates, [1:252](#)
 Uncertainties in parameter values/unit prices, [1:252](#)
 Uncertainties in traffic forecasts, [1:253](#)
 Uncontrolled charging, [1:560](#)
 Uncontrolled observations field experiments (field studies), [7A:9](#)
 Unified theory of acceptance and use of technology (UTAUT), [7A:83](#)
 Uniform Crime Reporting (UCR), [3:50](#)
 Uniform pricing, [1:97](#)
 Uniform speed limit (USL), [2:627](#)
 Unilateral or asymmetrical externalities, [1:59](#)
 Uninterrupted growth and globalization, [5:594](#)
 United Kingdom
 fatigue prevention in, [2:615](#)
 road safety management in, [2:524](#)
 United Kingdom (Suffolk Links), [4B:20](#)
 United Kingdom, intersection capacity of, [4A:251](#)
 United Nation's Agenda for Sustainable Development 2030, [3:64](#)
 United Nations Convention on the Law of the Sea (UNCLOS), [3:349](#), [5:567](#), [354](#)
 United Nations Decade of Action for Road Safety (2011-20), [2:269](#)
 United Nations Economic and Social Council, [2:306](#)
 United Nations Economic and Social Council's Conference on Road Traffic, [2:2](#)
 United Nations Economic Commission for Europe (UNECE), [2:380](#)
 United Nations Framework Convention on Climate Change (UNFCCC), [5:267](#)
 United Nations International Maritime Organization (the IMO), [1:437](#)
 United Nations Road Safety Collaboration, [2:527](#)
 United Nations Security Council (UNSC), [3:349](#)
 United States
 analysis of fatal bridge crashes in, [2:146](#)
 bridges, fatalities on or under the, [2:147](#)
 fatigue prevention in, [2:614](#)
 flash flood impact to roads in, [2:747](#)
 recreational boating in, [2:478](#)
 death and causes of injury, [2:479](#)
 road safety management in, [2:525](#)
 structural failures leading to vehicle occupant fatalities in, [2:144](#)
 structural failures outside, [2:145](#)
 United States (ITNAmerica & MetroAccess), [4B:20](#)
 United States Census Bureau (USCB), [2:368](#)
 United States Freight Transport Policy, [3:31](#)
 United States Geological Survey (USGS), [2:279](#)
 United States National Highway Traffic Safety Administration, [2:173](#)
 United States of America, intersection capacity of, [4A:249](#)
 Unitized cargoes, [3:258](#)
 Unit load devices (ULD), [2:103](#), [3:371](#)
 Unit price contract (UPC), [1:303](#)
 Unmanned aerial system (UAS)
 traffic management, [4A:376](#)
 Unmanned aerial systems (UAS), [2:38](#)
 Unmanned Aerial Vehicle (UAV), [2:348](#), [3:407](#), [5:266](#)
 Unmanned Aircraft System (UAS), [4A:369](#), [4A:377](#)
 Unmanned aircraft systems traffic management (UTM), [5:703](#)
 Unmanned automated vehicles (UAVs), [6:280](#)
 Unpacking mental health among women
 gender and well-being studies, [6:179](#)
 Unpacking mental health among women, [6:178](#)
 Unpaid bill to society, [1:57](#)
 Unpriced equilibrium, [1:147](#)
 Unprotected road users, risks for, [6:420](#)
 Unraveling complexity in commuting and transport, [5:93](#)
 Unreliability index (Ulr), [4A:115](#), [4A:112](#)
 Unsignalized crosswalks, [4A:346](#)
 Unsustainable transport system, [5:14](#)
 Upper and Lower bound Method (ULM), [2:348](#)
 Urban Air Mobility (UAM), [4A:369](#), [4A:376](#), [5:702](#), [5:703](#)
 EASA Regulatory Developments, [5:707](#)
 FAA's Unmanned Aircraft Systems (UAS) Integration Pilot Program (IPP), [5:707](#)
 ICAO Regulatory Developments, [5:708](#)
 leveraging COVID-19 medical and sanitation use cases to build community awareness, [5:708](#)
 NASA's Advanced Air Mobility National Campaign, [5:707](#)
 potential barriers and challenges, [5:703](#), [704](#)
 Potential UAM market segments and service models, [5:706](#)
 Potential UAM use cases, [5:703](#)
 role of research, pilot demonstrations, and regulation, [5:703](#)
 understanding and guiding the potential UAM impacts, [5:706](#)
 urban air mobility maturity levels, [5:707](#)
 Urban congestion charging systems around the world, [6:206](#)
 Urban Consolidation Centers (UCCs), [1:22](#)
 Urban design, [2:586](#)
 Urban development, [1:89](#)
 Urban economies, [1:7](#)
 Urban environment and health, [7B:6](#)
 Urban Freight
 bicycles for, [3:493](#)
 challenges for, [3:488](#)
 flows, [3:170](#)
 E-Commerce, [3:172](#)
 local freight demand, [3:170](#)
 spatial organization of warehousing, [3:171](#)
 trade-related freight demand, [3:170](#)
 problem, [3:168](#)

- Urban Freight (*Cont.*)
 - air pollution and GHG emissions, 3:168
 - conflicts and safety, 3:169
 - congestion, 3:169
 - strategies for managing, 3:174
 - managing local freight demand, 3:175
 - managing trade-related freight, 3:174
 - and urban deliveries, 3:488
- Urban freight impacts require new policies, 6:278
- Urban freight policy, 6:278
 - access rules, 6:279
 - data analytics, 6:283
 - data collection, 6:283
 - electric quadricycle for deliveries, Amsterdam, the Netherlands, 6:283
 - encouraging silent delivery vehicles in Paris, France, 6:280
 - innovation in urban logistics, 6:280
 - low-emission zones, 6:279
 - negotiating with freight stakeholders, 6:280
 - policy assessment, 6:283
 - tolls to organize urban logistics, 6:279
 - urban freight modeling, 6:283
- Urban freight systems, challenges
 - confronting, 1:20, 1:21
 - achieving the sustainable development goals, 1:21
 - consolidation, 1:21
 - eCommerce, 1:21
 - health and well being, 1:21
 - population growth, 1:20
 - safety, 1:21
- Urban freight transport (UFT), 1:326, 1:327, 3:242
 - behavior change and stakeholder engagement in, 3:242
 - planning, 1:328
 - planning to foster sustainability, need for, 1:327
 - private and public coordination, 1:327
 - UFT externalities, 1:327
 - policies, 1:329
- Urban functionality, 7B:25
 - model of mobility and public space, 7B:25
 - superblock, 7B:25
- Urban goods distribution and transport planning, 6:321
- Urban greenspaces, 7B:103
- Urbanism disequilibrium model, 1:93
- Urbanization, 1:20, 1:91
- Urban logistics, 1:326
 - drones usage in, 3:376
 - planning, 6:281
- Urban Logistics and Freight Transport
 - complexity challenge, 3:180
 - urban logistics initiatives, 3:181
 - demand-side initiatives, 3:182
 - supply-side initiatives, 3:181
- Urban mobility, 6:2
- Urban mobility hierarchy, 5:300
- Urban Motorways and management, 4A:4
 - customer expectations, 4A:6
 - design and control of, 4A:7
 - in economy, 4A:5
 - in Melbourne, 4A:8
 - optimal outcomes, 4A:9
 - transport network, 4A:4
 - US experience, 4A:8
- Urban motorways and ramp meters, 4A:14, 4A:2
 - fundamental diagram, 4A:16
 - granularity of data, 4A:18
 - maximum flows, 4A:3
 - microdynamics, 4A:14
 - ramp metering and mixed findings, 4A:3
 - simplistic numerate, 4A:17
 - traffic control, importance of, 4A:3
- Urban noise pollution, 7B:57
- Urban planning research and practice, implications for, 6:110
- Urban policy measures, transferability of
 - case studies in urban transport policy transfer since 2011, 6:80
 - good practice guidelines, 6:77
 - EU Level, 6:78
 - global level, 6:77
 - research on urban transport policy transfer since 2011, 6:80
 - 10-step transferability process, 6:79
 - suggestions for further research, 6:81
- Urban policy measures, transferability of, 6:77
- Urban public transport, 5:712
- Urban public transport in cities, 6:254
- Urban recreational travel, 4B:223
- Urban regeneration strategies, 6:358
 - challenges and criticisms, 6:359
 - experiences around the world 6:358
 - sustainable mobility and regeneration, 6:356
- Urban regeneration strategies, 6:356
- Urban Streets, Speed Limits on, 2:635
 - benefits of, 2:639
 - in cities, 2:632
 - operating speeds, 2:635
 - public perceptions and politics, 2:639
 - road function, 2:636
 - safe system approach and, 2:633
- Urban Traffic Optimization by Integrated Automation (UTOPIA), 4A:174
- Urban transportation planners, 1:92
 - classical model, 1:92
- Urban Transportation Planning System (UTPS) software, 4B:163
- Urban transportation studies, 4B:26
- Urban Transportation Systems (UTS), 2:742
 - methods, 2:743
- Urban transport planning (UTP)
 - process, 6:118
 - adoption and development of the UTP in the global south, 6:118
- Urban Travel Forecasting
 - combined equilibrium models, 4B:29
 - combined mode and route choices, 4B:29
 - combined model of destination, mode, and route choice, 4B:31
 - four-step procedure with feedback, 4B:27
 - future research, 4B:33
 - problem setting, 4B:26
 - user equilibrium traffic assignment, 4B:28
- Urban tunnels, 2:713
- US Department of Transportation (USDOT), 2:642, 2:382, 2:384, 4B:87
- Useful field of view (UFOV), 2:336
- US Environmental Protection Agency (EPA), 7B:103
- Use of lightweight materials in aircraft manufacturing, 5:293
- User equilibrium (UE) traffic assignment, 4B:28
- US experience, ramp metering, 4A:8
- US Federal Highway Administration
 - Ramp Management and Control Handbook, 4A:8
- US Freight Railroads
 - economic deregulation, 3:26
 - Railroad Revitalization and Regulatory Reform Act, 3:26
 - Staggers Rail Act of 1980, 3:26
 - economic regulation, 3:25
- US Motor Carrier Act of 1935, 3:25
- US Nuclear Regulatory Commission, 2:280
- US Staggers Act, 1:409
- US Truck Carriers
 - economic deregulation, 3:26
 - Household Goods Transportation Act of 1980, 3:26
 - Motor Carrier Act of 1980, 3:26
 - economic regulation, 3:25
- US waters, ferry vessels accidents in, 2:291

Utilitarianism Versus Deontological Ethic, 1:218
 altruism and environmental goods, 1:219
 intrinsic or absolute value *versus* economic value, 1:218
 VSL, healthcare and scarce resources: standard essential, 1:218
 Utility Poles
 breakaway or energy absorbing poles, 2:734
 clear zone, 2:733
 contributing factors and statistical modeling, 2:732
 countermeasure selection, 2:736
 crashes, 2:731
 decrease the number of, 2:734
 in hazard locations, 2:735
 keep vehicles on the roadway, 2:734
 lateral offset, 2:734
 place utility lines underground, 2:734
 relocate, 2:734
 safety treatments, 2:733
 shield, 2:735
 Utility terrain vehicles (UTV), 2:77
 Utopia point, 4B:82

V

Validity and reliability, criterion for describing quality, 7A:109
 Valuation of CO₂, 1:278
 Valuation of travel time variability and travel information, 1:148
 Value added tax (VAT), 5:254
 Value-belief-norm theory, 5:385
 Value of a statistical life (VSL), 1:115, 1:278, 2:737
 Value of changes in travel time (VCTT), 1:67, 1:70
 Value of reliability (VOR), 3:236
 Value of risk reductions (VRR), 2:737
 Value of security, 1:124
 Value of time (VOT), 1:128, 1:278, 3:236
 in freight transport, 1:321, 3:236
 decision maker, 3:236
 future of, 3:239
 heterogeneity, 3:237
 obtaining data, 3:237
 reliability, 3:239
 Value of transfer, 1:123
 Value of travel time (VOTT), 1:165, 1:527, 4B:151
 CSA or inference from revealed preference data, 1:168
 data and estimation approach in national VTT studies, 1:528
 choice experiments, 1:530
 estimation models, 1:531
 overview over national studies, 1:528
 estimating VTT from SC data, 1:168
 estimation with revealed preference data, 1:533
 overview over some type of studies that estimate, 1:528
 selected challenges, 1:531, 1:532
 absolute size of the time saving (size effect), 1:531
 estimating VTT for business Trip, 1:532
 estimating VTT for cycling (and walking) for transportation, 1:532
 mode and user type effects, 1:532
 recruitment effects, 1:533
 Value of travel time savings (VOTTS), 2:738, 4B:100
 Value of walking and waiting times, 1:122
 Value-pricing in the United States, 1:137
 Valuing wider impacts and avoiding double counting, 1:472
 Van crash statistics, 2:402
 Van Miert Report, 3:282
 Vapor pressure, defined, 3:467
 Variability index, 4A:115
 Variable Message Sign (VMS), 4A:107, 4A:287, 4A:289
 Variable message signs, 4B:13
 Variables included in the cost function of Cargo Airlines, 1:43
 Variables included in the cost function of shipping companies, 1:43
 Variable Speed Limit (VSL), 4A:2
 control algorithms, 4A:38
 freeway traffic control measures, 4A:38
 induced capacity, 4A:35, 4A:36, 4A:37
 induced critical density, 4A:35, 4A:36, 4A:37
 induced free-flow speed, 4A:35, 4A:36
 macroscopic modeling of, 4A:34
 of Carlson et al., 4A:35
 of Frejo et al., 4A:36
 of Hegyi et al., 4A:34
 model comparison, 4A:37
 pollution-oriented algorithms, 4A:33
 safety-oriented, 4A:33
 for traffic efficiency improvement, 4A:33, 37
 discrete *versus* continuous, 4A:38
 local *versus* global, 4A:38
 optimization-based *versus* easy-to-implement, 4A:37
 Variance amplification, 3:130
 Vector-based and zoonotic infections, 7B:107
 Vehicle, 6:57
 Vehicle automation, 5:20
 need in cities, 6:200
 potential impacts on urban mobility
 planning for AVs as PT, 6:202
 transport modeling studies, 6:201
 travelers' behavior, 6:201
 potential impacts on urban mobility, 6:201
 in public transport, applications of, 6:199
 Vehicle automation in public transport (PT), 6:199
 systems, main usages of, 6:199
 Vehicle Certification (VCA), 2:524
 Vehicle data, 2:560
 Vehicle Design, 2:79
 Vehicle electrification, 7B:130
 Vehicle emissions, 7B:148
 Vehicle/fuel technology, 7B:129
 Vehicle identification, 4A:61
 Vehicle kilometers travelled (VKT), 2:282, 2:283, 4A:95, 4A:4
 Vehicle lighting, 2:363
 forward lighting, 2:363
 Vehicle miles traveled (VMT), 2:283, 2:296, 2:300
 Vehicle-miles-traveled (VMT), 4B:42, 4B:138, 4B:195
 Vehicle operating cost, 1:14, 4A:76
 Vehicle ownership models
 aggregated, 4B:206
 disaggregated, 4B:207
 new challenges, 4B:209
 objective and subjective factors, 4B:209
 objective factors, 4B:208
 subjective factors, 4B:208
 Vehicle parking
 and cyclists, 2:395
 and older adults, 2:396
 and pedestrians, 2:396
 policies and regulations, 2:394
 Vehicle Passive Safety, 2:64
 Vehicle restraint system, 2:597, 2:593
 Vehicle/road system, 2:109
 Vehicle routing, 3:160, 3:224
 Vehicle safety implications, of elderly driver crashes, 2:237
 Vehicle Sanction Laws, 2:47
 Vehicles and their uses, 3:204
 Vehicle's controller area network (CAN), 7A:31
 Vehicle scrappage schemes, 1:517
 Vehicle Star Rating (VSR), 2:80
 Vehicle taxes, 1:516
 Vehicle technology solutions, 2:314
 Vehicle-to-environment (V2X)
 communication, 2:177
 history of, 2:183

- Vehicle-to-infrastructure (V2I), 2:183, 2:177, 4A:61
- Vehicle-to-network (V2N), 2:183
- Vehicle-to-pedestrian (V2P), 2:183
- Vehicle to roadside communication (VRC), 4A:221
- Vehicle-to-vehicle (V2V) communication, 2:177
- Vehicle Trajectory, 4B:138
- Vehicle type, 2:129
- Vehicular and pedestrian signals, 4A:229
- Velo-city 2007 conference, 2:111
- Vendor managed inventory (VMI), 3:72
- Verbal sexual abuse, 2:576
- Vertical integration between airports and airlines' services, 5:210
- Vertical speed indicator and traffic display (VSI/TDI), 2:166
- Vertical take-off and landing (VTOL) capable aircraft, 5:702
- Vertical take-off and land (VTOL) rural air mobility, 5:703
- Very High Frequency Omnidirectional Range (VOR), 4A:381
- Vessel sharing agreement (VSA) 3:252
- Vessel size, scale increases in, 3:251
- ViaVan (vans), 6:199
- VicRoads, 4A:10, 4A:2
- control algorithms used, 4A:21
 - close-down sequence, 4A:25
 - dynamic activation and deactivation, 4A:24
 - faults and device failures, 4A:25
 - HERO-LIVE modules, 4A:22
 - HERO-LIVE suite of algorithms, 4A:22
 - incidents and events, 4A:25
 - manual operation, 4A:25
 - maximum cycle time, 4A:28
 - minimum cycle time, 4A:27
 - ramp meters, 4A:24
 - signal timings, 4A:26
 - start-up sequence, 4A:25
 - time of day activation, 4A:25
 - times of operation, 4A:24
 - guidelines, 4A:205
- Victoria line train graph, homogenous services, 1:381
- Video technology, 6:306
- Video vehicle detection, 2:693
- Vienna Convention on Road Signs and Signals, 2:648, 2:653
- Violation of independence, 7A:82
- "Virtual" garage operators, 1:163
- Viscosity, defined, 3:467
- Visibility
- glare, 2:362
 - and road condition, 2:130
 - visual performance, 2:361
- Vision Zero, 2:523, 2:554, 2:675
- VISSIM, 4B:69
- Visual acuity, 2:333
- Visual characteristics and limitations, 2:333
- closing velocity, 2:334
 - contrast sensitivity, 2:333
 - peripheral vision, 2:333
 - visual acuity, 2:333
- Visual distraction, 2:337
- Visual impacts of transport, 7B:61
- assess visual impacts, 7B:64
 - consequences of visual impacts, 7B:63
 - reduce or prevent visual impacts, 7B:65
 - subjectivity of visual impacts, 7B:64
- Visual intrusion-impact on users of non-motorized modes and non-users, 7B:61
- motorized vehicles, 7B:63
 - transport infrastructure, 7B:61
- Visual line-of-sight (VLOS), 3:377
- Visually Impaired Pedestrian Safety challenges to independent travel, 2:436
- historical perspective, 2:435
 - modern travel environment, risk in, 2:436
 - orientation and mobility, 2:435
- Volatile organic compounds (VOCs), 6:236, 7B:44
- Volcanic ash, 2:93
- Volpe National Transportation Systems Center, 2:525
- Volume-delay functions, 4B:69
- Voluntary local reviews (VLRs), 7B:4
- Volunteered geographic information (VGI), 4B:40
- Voting Theory, 1:194
- VRU Beacon system (VBS), 2:114, 2:113
- V's of big data, 5:665
- Vulnerability, 2:36, 36
- Vulnerability, 7B:35
- Vulnerable road, 7A:1
- Vulnerable road users (VRU), 2:183, 2:108, 7B:157
- Vulnerable users, 2:633
- W**
- Wagon, 3:423
- Waiting time, 1:122
- Walkability, 5:333, 6:340
- benefits of, 6:341
 - education and encouragement, 6:346
 - evaluation of walkability policy and planning, 6:346
 - measure the impact of walkability interventions in walking activity and experience, 6:347
 - measuring walkability, 6:347
 - measuring walking activity and experiences, 6:346
- index, 4A:304
- policies to assure pedestrian accessibility, 6:342
- policies to encourage a modal shift, 6:343
- policies to ensure pedestrians' rights, 6:342
- policies to promote walking as a health-enhancing activity
- 10 healthy streets indicators, 6:343
- policies to promote walking as a health-enhancing activity, 6:343
- policies to provide pedestrian safety and security, 6:342
- policy instruments, 6:344
- '6-Es' for walkability policy, 6:345
- pedestrian experience and behaviour, determinants of, 6:344
- policy objectives, 6:342
- regulations and fiscal measurements to improve walkability, 6:346
- urban and transport planning for walkability, engineering, 6:345
- Walkable pedestrian space, measuring, 5:321
- Walking times, 1:122
- Walking to school, 2:569
- WALRAS' Law, 1:522
- Wardrop Equilibrium, 4A:91
- Warehouse development in suburban areas, Atlanta, United States, 2018, 6:282
- Warehouse management systems (WMS), 3:80, 3:76, 3:78
- Warehouses, distribution centers, and logistics zones, 3:36
- Warehousing, spatial organization of, 3:171
- Warsaw Convention (1925), 3:361
- Washington Metropolitan Area Transit Authority (WMATA), 1:268
- Waste Electrical and Electronic Equipment (WEEE), 3:223
- Watts' hump, 2:643
- WAVE Short Message Protocol (WSMP), 2:183
- Ways of working and commuting
- concentration of employment in factories and offices, 5:94
 - desire for independence and flexibility, 5:95
 - home, field, farm, and workshop, 5:94
 - role of mass transit, 5:94
- Ways of working and commuting, 5:93

- W-beam guardrails, 2:516
- Weakening strict complementarity
between labor supply and
commuting, 1:224
spatial considerations, 1:225
two commuting modes, 1:224
two trip purposes, 1:224
weaker links between commuting and
labor supply, 1:224
- Weapons effect, 2:20
- Weather, Effects of
challenges, 2:748
impact of adverse, 2:744
international case studies, 2:746
flash flood impact to roads in
Europe, 2:746
flash flood impact to roads in the
United States, 2:747
flood impact to roads in Japan,
2:746
modeling methods, 2:746
- Weather-related transport system
disruptions, 7B:36
- Weaving, 4A:139
- Webber's air mattress theory, 1:337
- WebTAG guidance, 1:356
- Weigh-in-motion (WIM), 4A:62
- Weighted Average Capital Cost,
(WACC), 1:466
- Weighted average costs of capital
(WACC), 3:115
- Weighted global criterion method,
4B:82
- Weighted sum model, 4B:83
- Welch's t-test, 7A:83
- Welfare economics, 1:473
- Welfare impact, of congestion pricing,
4A:79
- Welfare loss of the fuel tax, 1:142
- Well-being, 5:108
- Werther effect, 2:659
- Western Extension Zone, 4A:72
- Wet bulk sector, of shipping industry,
3:270
crude oil seaborne trade, 3:270
oil products seaborne trade, 3:274
supply of, 3:275
- Whiplash injury, 2:410
- White Bike Plan, 2:139
- Wider economic impacts (WEIs), 1:471
- Wildlife warning reflectors, 2:61
- Wildlife whistles, 2:61
- William Alonso's bid-rent theory,
4B:125
- Willingness to accept (WTA), 1:166,
3:234
- Willingness to pay (WTP), 1:114, 1:166,
2:737, 3:234
- Windows, laminated and tempered
safety glass for, 2:412
- Wind shear alert systems (WSAS), 2:353
- Winner curse reversed, 1:467
- Winter road conditions, 2:535
- Winter Road Maintenance Guidelines,
2:537
determining snow and ice control
treatment, 2:537
determining the level of service need,
2:537
- Winter Road Maintenance Strategies,
2:535
abrasive application, 2:536
anti-icing, 2:536
deicing, 2:536
plowing, 2:535
- Winter Road Safety, 2:534
- Winter Road Surface Conditions and
Road Safety, 2:533
icy road surface, 2:534
loose snow-covered surface, 2:534
packed snow-covered surface, 2:534
wet pavement, 2:534
- Wireless Access in Vehicle Environments
(WAVE), 2:183
- Wireless connectivity capabilities
6:167
- Wireless Local Area Networks (WLAN),
2:183
- Wireless sensor networks (WSN), 6:307,
6:306
- Women and acceptance of new
transportation technology,
7A:197
- Women-only cars, 2:577
- Women-only waiting rooms, 2:577
- Women's complicated travel patterns,
7A:194
- Women's mode choice, 7A:195
- Women's role in "concerted cultivation"
of children, 6:174
- Woonerf zone, 2:574
- Woonerven zones, 2:570
- Word cloud of referenced works
7B:153
- Worker safety, 2:194
truck-mounted attenuators, 2:194
- Work-family conflict (WFC), 6:179
- Working capital management (WCM),
3:114
- Working Positions Management,
4A:380
- Workings of CBA, 1:249
- Workplace Parking Levy (WPL), 6:2,
6:115
impact of the Nottingham WPL, 6:5
limitations of the WPL and lessons
learned, 6:5
Nottingham WPL, 6:4
public acceptance of the WPL, 6:4
structure of the WPL Scheme, 6:5
- Workplace Parking Levy (England)
Regulations, 2009 No 2085,
6:3
- Work-zone classification, 2:191
work-zone duration, 2:191
work-zone lane configuration, 2:191
- Work zone road safety audit (WZRSA),
2:511
- Work-zone safety and mobility impact
assessment, 2:192
work-zone mobility impact
assessment, 2:193
work-zone safety assessment, 2:192
- Work-zone temporary traffic control,
2:190, 2:189
- World Air Freight Growth, historical
trends in, 2:98
- World Food Program (WFP), 3:21
- World Forum for Harmonization of
Vehicle Regulations, 2:527
- World Health Organization (WHO),
1:231, 1:476, 5:267, 7A:198,
7B:87, 7B:123, 178
fatalities from road crashes, 2:269
Healthy Cities, 6:370
- World Meteorological Organizations
(WMO), 3:350
- World's Largest Cargo Airports, 3:362
- World's largest cruise ships, 2:294
- World's major container ports and sea
routes, 2015, 5:41
- World total trade and maritime trade,
relationship, 5:509
- World Trade Center, 2:241
- World Trade Organization (WTO),
1:432
Trade Facilitation Agreement, 3:154
- World War II, 6:32
- Worldwide Air Travel Demand, 1:547
- Wrap-around addition (WAA), 1:375
- Wreck or Collision, 2:8
- Wrong-way driving (WWD)
contributing factors, 2:751
crashes, 2:757
safety countermeasures, 2:752
geometric design features, 2:754
Intelligent Transportation Systems,
2:755
signs, pavement markings, and
traffic signals, 2:752
- ## Y
- Yangtze River City Cluster, 3:501
- Yard crane (YC), 3:328
- Yard crane deployment problem
(YCDP), 3:331
- Yard crane scheduling problem (YCSP),
3:332
- Y-25 bogie, 3:431

Yellow Book, [2:513](#)

Yellow light, [4A:178](#)

Yield management

big data analytics, artificial
intelligence, and, [1:407](#)

conditions for application of, [1:406](#)

methods, [1:406](#)

short history of, [1:405](#)

Young adulthood, [5:216](#)

Youth, [2:14](#)

Z

Zebra Crossing, [4A:346](#), [4A:347](#)

Zero-carbon fuel options and issues,
[7B:130](#)

Zero-emission vehicle (ZEV)
[1:497](#)

Zero-Tolerance Laws for Underage
Drivers, [2:42](#)

ZIP Code, [2:368](#)

ZIP code tabulation areas (ZCTA)
[2:368](#)

Zombie, [2:424](#)